



CUSOLVER LIBRARY

DU-06709-001_v10.2 | February 2021



TABLE OF CONTENTS

Chapter 1. Introduction.....	1
1.1. cuSolverDN: Dense LAPACK.....	2
1.2. cuSolverSP: Sparse LAPACK.....	2
1.3. cuSolverRF: Refactorization.....	3
1.4. Naming Conventions.....	3
1.5. Asynchronous Execution.....	5
1.6. Library Property.....	5
1.7. high precision package.....	5
Chapter 2. Using the CUSOLVER API.....	6
2.1. General description.....	6
2.1.1. Thread Safety.....	6
2.1.2. Scalar Parameters.....	6
2.1.3. Parallelism with Streams.....	6
2.1.4. Link Third-party LAPACK Library.....	7
2.1.5. convention of info.....	7
2.1.6. usage of _bufferSize.....	7
2.2. cuSolver Types Reference.....	8
2.2.1. cuSolverDN Types.....	8
2.2.1.1. cusolverDnHandle_t.....	8
2.2.1.2. cublasFillMode_t.....	8
2.2.1.3. cublasOperation_t.....	8
2.2.1.4. cusolverEigType_t.....	8
2.2.1.5. cusolverEigMode_t.....	9
2.2.1.6. cusolverIIRSRefinement_t.....	9
2.2.1.7. cusolverDnIIRSParams_t.....	10
2.2.1.8. cusolverDnIIRSInfos_t.....	10
2.2.1.9. cusolverStatus_t.....	10
2.2.2. cuSolverSP Types.....	10
2.2.2.1. cusolverSpHandle_t.....	10
2.2.2.2. cusparseMatDescr_t.....	10
2.2.2.3. cusolverStatus_t.....	11
2.2.3. cuSolverRF Types.....	12
2.2.3.1. cusolverRfHandle_t.....	12
2.2.3.2. cusolverRfMatrixFormat_t.....	12
2.2.3.3. cusolverRfNumericBoostReport_t.....	12
2.2.3.4. cusolverRfResetValuesFastMode_t.....	12
2.2.3.5. cusolverRfFactorization_t.....	12
2.2.3.6. cusolverRfTriangularSolve_t.....	13
2.2.3.7. cusolverRfUnitDiagonal_t.....	13
2.2.3.8. cusolverStatus_t.....	13

2.3. cuSolver Formats Reference.....	13
2.3.1. Index Base Format.....	13
2.3.2. Vector (Dense) Format.....	14
2.3.3. Matrix (Dense) Format.....	14
2.3.4. Matrix (CSR) Format.....	14
2.3.5. Matrix (CSC) Format.....	15
2.4. cuSolverDN: dense LAPACK Function Reference.....	16
2.4.1. cuSolverDN Helper Function Reference.....	16
2.4.1.1. cusolverDnCreate().....	16
2.4.1.2. cusolverDnDestroy().....	17
2.4.1.3. cusolverDnSetStream().....	17
2.4.1.4. cusolverDnGetStream().....	17
2.4.1.5. cusolverDnCreateSyevjInfo().....	18
2.4.1.6. cusolverDnDestroySyevjInfo().....	18
2.4.1.7. cusolverDnXsyevjSetTolerance().....	18
2.4.1.8. cusolverDnXsyevjSetMaxSweeps().....	19
2.4.1.9. cusolverDnXsyevjSetSortEig().....	19
2.4.1.10. cusolverDnXsyevjGetResidual().....	19
2.4.1.11. cusolverDnXsyevjGetSweeps().....	20
2.4.1.12. cusolverDnCreateGesvdjInfo().....	20
2.4.1.13. cusolverDnDestroyGesvdjInfo().....	21
2.4.1.14. cusolverDnXgesvdjSetTolerance().....	21
2.4.1.15. cusolverDnXgesvdjSetMaxSweeps().....	21
2.4.1.16. cusolverDnXgesvdjSetSortEig().....	22
2.4.1.17. cusolverDnXgesvdjGetResidual().....	22
2.4.1.18. cusolverDnXgesvdjGetSweeps().....	22
2.4.1.19. cusolverDnIRSPParamsCreate().....	23
2.4.1.20. cusolverDnIRSPParamsDestroy().....	23
2.4.1.21. cusolverDnIRSPParamsSetRefinementSolver().....	24
2.4.1.22. cusolverDnIRSPParamsSetSolverMainPrecision().....	24
2.4.1.23. cusolverDnIRSPParamsSetSolverLowestPrecision().....	25
2.4.1.24. cusolverDnIRSPParamsSetSolverPrecisions().....	26
2.4.1.25. cusolverDnIRSPParamsSetTol().....	26
2.4.1.26. cusolverDnIRSPParamsSetTolInner().....	27
2.4.1.27. cusolverDnIRSPParamsSetMaxIters().....	28
2.4.1.28. cusolverDnIRSPParamsSetMaxItersInner().....	28
2.4.1.29. cusolverDnIRSPParamsGetMaxIters().....	29
2.4.1.30. cusolverDnIRSInfosCreate().....	29
2.4.1.31. cusolverDnIRSInfosDestroy().....	30
2.4.1.32. cusolverDnIRSInfosGetMaxIters().....	30
2.4.1.33. cusolverDnIRSInfosGetNIters().....	31
2.4.1.34. cusolverDnIRSInfosGetOuterNIters().....	31
2.4.1.35. cusolverDnIRSInfosRequestResidual().....	32

2.4.1.36. cusolverDnIRSInfosGetResidualHistory()	32
2.4.2. Dense Linear Solver Reference	34
2.4.2.1. cusolverDn<t>potrf()	34
2.4.2.2. cusolverDn<t>potrs()	37
2.4.2.3. cusolverDn<t>potri()	39
2.4.2.4. cusolverDn<t>getrf()	42
2.4.2.5. cusolverDn<t>getrs()	45
2.4.2.6. cusolverDn<t>geqrf()	47
2.4.2.7. cusolverDn<t>ormqr()	50
2.4.2.8. cusolverDn<t>orgqr()	54
2.4.2.9. cusolverDn<t>sytrf()	57
2.4.2.10. cusolverDn<t>potrfBatched()	60
2.4.2.11. cusolverDn<t>potrsBatched()	62
2.4.2.12. cusolverDn<t1><t2>gesv()	64
2.4.2.13. cusolverDnIRSXgesv()	70
2.4.3. Dense Eigenvalue Solver Reference	74
2.4.3.1. cusolverDn<t>gebrd()	74
2.4.3.2. cusolverDn<t>orgbr()	78
2.4.3.3. cusolverDn<t>sytrd()	82
2.4.3.4. cusolverDn<t>ormtr()	86
2.4.3.5. cusolverDn<t>orgtr()	90
2.4.3.6. cusolverDn<t>gesvd()	93
2.4.3.7. cusolverDn<t>gesvdj()	98
2.4.3.8. cusolverDn<t>gesvdjBatched()	103
2.4.3.9. cusolverDn<t>gesvdaStridedBatched()	108
2.4.3.10. cusolverDn<t>syevd()	114
2.4.3.11. cusolverDn<t>syevdx()	118
2.4.3.12. cusolverDn<t>sygvd()	124
2.4.3.13. cusolverDn<t>sygvdx()	129
2.4.3.14. cusolverDn<t>syevj()	135
2.4.3.15. cusolverDn<t>sygvj()	140
2.4.3.16. cusolverDn<t>syevjBatched()	146
2.5. cuSolverSP: sparse LAPACK Function Reference	150
2.5.1. Helper Function Reference	150
2.5.1.1. cusolverSpCreate()	151
2.5.1.2. cusolverSpDestroy()	151
2.5.1.3. cusolverSpSetStream()	151
2.5.1.4. cusolverSpXcsrissym()	152
2.5.2. High Level Function Reference	153
2.5.2.1. cusolverSp<t>csrslvlu()	154
2.5.2.2. cusolverSp<t>csrslvqr()	157
2.5.2.3. cusolverSp<t>csrslvchol()	160
2.5.2.4. cusolverSp<t>csrslvqr()	163

2.5.2.5. cusolverSp<t>csreigvsi()	167
2.5.2.6. cusolverSp<t>csreigs()	171
2.5.3. Low Level Function Reference	173
2.5.3.1. cusolverSpXcsrsmrcm()	173
2.5.3.2. cusolverSpXcsrsmmdq()	175
2.5.3.3. cusolverSpXcsrsmamd()	176
2.5.3.4. cusolverSpXcsrmetisnd()	178
2.5.3.5. cusolverSpXcsrzfd()	180
2.5.3.6. cusolverSpXcsrperm()	182
2.5.3.7. cusolverSpXcsrqrBatched()	184
2.6. cuSolverRF: Refactorization Reference	192
2.6.1. cusolverRfAccessBundledFactors()	192
2.6.2. cusolverRfAnalyze()	193
2.6.3. cusolverRfSetupDevice()	194
2.6.4. cusolverRfSetupHost()	196
2.6.5. cusolverRfCreate()	198
2.6.6. cusolverRfExtractBundledFactorsHost()	199
2.6.7. cusolverRfExtractSplitFactorsHost()	200
2.6.8. cusolverRfDestroy()	201
2.6.9. cusolverRfGetMatrixFormat()	201
2.6.10. cusolverRfGetNumericProperties()	202
2.6.11. cusolverRfGetNumericBoostReport()	202
2.6.12. cusolverRfGetResetValuesFastMode()	203
2.6.13. cusolverRfGet_Algs()	203
2.6.14. cusolverRfRefactor()	203
2.6.15. cusolverRfResetValues()	204
2.6.16. cusolverRfSetMatrixFormat()	205
2.6.17. cusolverRfSetNumericProperties()	206
2.6.18. cusolverRfSetResetValuesFastMode()	206
2.6.19. cusolverRfSetAlgs()	207
2.6.20. cusolverRfSolve()	207
2.6.21. cusolverRfBatchSetupHost()	209
2.6.22. cusolverRfBatchAnalyze()	211
2.6.23. cusolverRfBatchResetValues()	212
2.6.24. cusolverRfBatchRefactor()	213
2.6.25. cusolverRfBatchSolve()	214
2.6.26. cusolverRfBatchZeroPivot()	215
Chapter 3. Using the CUSOLVERMG API	217
3.1. General description	217
3.1.1. Thread Safety	217
3.1.2. Determinism	217
3.1.3. tile strategy	217
3.1.4. Global matrix versus local matrix	219

3.1.5. usage of _bufferSize.....	219
3.1.6. synchronization.....	220
3.1.7. context switch.....	220
3.1.8. NVLINK.....	220
3.2. cuSolverMG Types Reference.....	220
3.2.1. cuSolverMG Types.....	220
3.2.2. cusolverMgHandle_t.....	220
3.2.3. cusolverMgGridMapping_t.....	220
3.2.4. cudaLibMgGrid_t.....	221
3.2.5. cudaLibMgMatrixDesc_t.....	221
3.3. Helper Function Reference.....	221
3.3.1. cusolverMgCreate().....	221
3.3.2. cusolverMgDestroy().....	221
3.3.3. cusolverMgDeviceSelect().....	222
3.3.4. cusolverMgCreateDeviceGrid().....	222
3.3.5. cusolverMgDestroyGrid().....	223
3.3.6. cusolverMgCreateMatDescr().....	223
3.3.7. cusolverMgDestroyMatrixDesc().....	224
3.4. Dense Linear Solver Reference.....	224
3.4.1. cusolverMgGetrf().....	225
3.4.2. cusolverMgGetrs().....	227
3.5. Dense Eigenvalue Solver Reference.....	229
3.5.1. cusolverMgSyevd().....	230
Appendix A. cuSolverRF Examples.....	233
A.1. cuSolverRF In-memory Example.....	233
A.2. cuSolverRF-batch Example.....	237
Appendix B. CSR QR Batch Examples.....	241
B.1. Batched Sparse QR example 1.....	241
B.2. Batched Sparse QR example 2.....	245
Appendix C. QR Examples.....	251
C.1. QR Factorization Dense Linear Solver.....	251
C.2. orthogonalization.....	255
Appendix D. LU Examples.....	261
D.1. LU Factorization.....	261
Appendix E. Cholesky Examples.....	266
E.1. batched Cholesky Factorization.....	266
Appendix F. Examples of Dense Eigenvalue Solver.....	271
F.1. Standard Symmetric Dense Eigenvalue Solver.....	271
F.2. Standard Symmetric Dense Eigenvalue Solver.....	274
F.3. Generalized Symmetric-Definite Dense Eigenvalue Solver.....	277
F.4. Generalized Symmetric-Definite Dense Eigenvalue Solver.....	280
F.5. Standard Symmetric Dense Eigenvalue Solver (via Jacobi method).....	284
F.6. Generalized Symmetric-Definite Dense Eigenvalue Solver (via Jacobi method).....	288

F.7. batch eigenvalue solver for dense symmetric matrix.....	293
Appendix G. Examples of Singular Value Decomposition.....	299
G.1. SVD with singular vectors.....	299
G.2. SVD with singular vectors (via Jacobi method).....	303
G.3. batch dense SVD solver.....	308
G.4. SVD approximation.....	314
Appendix H. Examples of multiGPU eigenvalue solver.....	319
H.1. SYEVD of 1D Laplacian operator (example 1).....	320
H.2. SYEVD of 1D Laplacian operator (example 2).....	332
H.3. SYEVD of 1D Laplacian operator (example 3).....	337
Appendix I. Examples of multiGPU linear solver.....	342
I.1. GETRF and GETRS of 1D Laplacian operator (example 1).....	343
Appendix J. Acknowledgements.....	351
Appendix K. Bibliography.....	354

LIST OF FIGURES

Figure 1	Example of cusolveMG tiling for 3 Gpus	218
Figure 2	global matrix and local matrix	219

Chapter 1.

INTRODUCTION

The cuSolver library is a high-level package based on the cuBLAS and cuSPARSE libraries. It consists of two modules corresponding to two sets of API:

1. The cuSolver API on a single GPU
2. The cuSolverMG API on a single node multiGPU

Each of which can be used independently or in concert with other toolkit libraries. To simplify the notation, cuSolver denotes single GPU API and cuSolverMg denotes multiGPU API.

The intent of cuSolver is to provide useful LAPACK-like features, such as common matrix factorization and triangular solve routines for dense matrices, a sparse least-squares solver and an eigenvalue solver. In addition cuSolver provides a new refactorization library useful for solving sequences of matrices with a shared sparsity pattern.

cuSolver combines three separate components under a single umbrella. The first part of cuSolver is called cuSolverDN, and deals with dense matrix factorization and solve routines such as LU, QR, SVD and LDLT, as well as useful utilities such as matrix and vector permutations.

Next, cuSolverSP provides a new set of sparse routines based on a sparse QR factorization. Not all matrices have a good sparsity pattern for parallelism in factorization, so the cuSolverSP library also provides a CPU path to handle those sequential-like matrices. For those matrices with abundant parallelism, the GPU path will deliver higher performance. The library is designed to be called from C and C++.

The final part is cuSolverRF, a sparse re-factorization package that can provide very good performance when solving a sequence of matrices where only the coefficients are changed but the sparsity pattern remains the same.

The GPU path of the cuSolver library assumes data is already in the device memory. It is the responsibility of the developer to allocate memory and to copy data between GPU memory and CPU memory using standard CUDA runtime API routines, such as `cudaMalloc()`, `cudaFree()`, `cudaMemcpy()`, and `cudaMemcpyAsync()`.

cuSolverMg is GPU-accelerated ScaLAPACK. By now, cuSolverMg supports 1-D column block cyclic layout and provides symmetric eigenvalue solver.



The cuSolver library requires hardware with a CUDA compute capability (CC) of at least 2.0 or higher. Please see the *CUDA C++ Programming Guide* for a list of the compute capabilities corresponding to all NVIDIA GPUs.

1.1. cuSolverDN: Dense LAPACK

The cuSolverDN library was designed to solve dense linear systems of the form

$$Ax = b$$

where the coefficient matrix $A \in \mathbb{R}^{n \times n}$, right-hand-side vector $b \in \mathbb{R}^n$ and solution vector $x \in \mathbb{R}^n$

The cuSolverDN library provides QR factorization and LU with partial pivoting to handle a general matrix \mathbf{A} , which may be non-symmetric. Cholesky factorization is also provided for symmetric/Hermitian matrices. For symmetric indefinite matrices, we provide Bunch-Kaufman (LDL) factorization.

The cuSolverDN library also provides a helpful bidiagonalization routine and singular value decomposition (SVD).

The cuSolverDN library targets computationally-intensive and popular routines in LAPACK, and provides an API compatible with LAPACK. The user can accelerate these time-consuming routines with cuSolverDN and keep others in LAPACK without a major change to existing code.

1.2. cuSolverSP: Sparse LAPACK

The cuSolverSP library was mainly designed to solve sparse linear system

$$Ax = b$$

and the least-squares problem

$$x = \operatorname{argmin} \|A^* z - b\|$$

where sparse matrix $A \in \mathbb{R}^{m \times n}$, right-hand-side vector $b \in \mathbb{R}^m$ and solution vector $x \in \mathbb{R}^n$. For a linear system, we require $m=n$.

The core algorithm is based on sparse QR factorization. The matrix \mathbf{A} is accepted in CSR format. If matrix \mathbf{A} is symmetric/Hermitian, the user has to provide a full matrix, ie fill missing lower or upper part.

If matrix \mathbf{A} is symmetric positive definite and the user only needs to solve $Ax = b$, Cholesky factorization can work and the user only needs to provide the lower triangular part of \mathbf{A} .

On top of the linear and least-squares solvers, the **cuSolverSP** library provides a simple eigenvalue solver based on shift-inverse power method, and a function to count the number of eigenvalues contained in a box in the complex plane.

1.3. cuSolverRF: Refactorization

The cuSolverRF library was designed to accelerate solution of sets of linear systems by fast re-factorization when given new coefficients in the same sparsity pattern

$$A_i x_i = f_i$$

where a sequence of coefficient matrices $A_i \in R^{n \times n}$, right-hand-sides $f_i \in R^n$ and solutions $x_i \in R^n$ are given for $i=1, \dots, k$.

The cuSolverRF library is applicable when the sparsity pattern of the coefficient matrices A_i as well as the reordering to minimize fill-in and the pivoting used during the LU factorization remain the same across these linear systems. In that case, the first linear system ($i=1$) requires a full LU factorization, while the subsequent linear systems ($i=2, \dots, k$) require only the LU re-factorization. The later can be performed using the cuSolverRF library.

Notice that because the sparsity pattern of the coefficient matrices, the reordering and pivoting remain the same, the sparsity pattern of the resulting triangular factors L_i and U_i also remains the same. Therefore, the real difference between the full LU factorization and LU re-factorization is that the required memory is known ahead of time.

1.4. Naming Conventions

The cuSolverDN library functions are available for data types **float**, **double**, **cuComplex**, and **cuDoubleComplex**. The naming convention is as follows:

`cusolverDn<t><operation>`

where `<t>` can be **S**, **D**, **C**, **Z**, or **X**, corresponding to the data types **float**, **double**, **cuComplex**, **cuDoubleComplex**, and the generic type, respectively. `<operation>` can be Cholesky factorization (**potrf**), LU with partial pivoting (**getrf**), QR factorization (**geqrf**) and Bunch-Kaufman factorization (**sytrf**).

The cuSolverSP library functions are available for data types **float**, **double**, **cuComplex**, and **cuDoubleComplex**. The naming convention is as follows:

`cusolverSp[Host]<t>[<matrix data format>]<operation>[<output matrix data format>]<based on>`

where **cuSolverSp** is the GPU path and **cusolverSpHost** is the corresponding CPU path. `<t>` can be **S**, **D**, **C**, **Z**, or **X**, corresponding to the data types **float**, **double**, **cuComplex**, **cuDoubleComplex**, and the generic type, respectively.

The `<matrix data format>` is **csr**, compressed sparse row format.

The **<operation>** can be **ls**, **lsq**, **eig**, **eigs**, corresponding to linear solver, least-square solver, eigenvalue solver and number of eigenvalues in a box, respectively.

The **<output matrix data format>** can be **v** or **m**, corresponding to a vector or a matrix.

<based on> describes which algorithm is used. For example, **qr** (sparse QR factorization) is used in linear solver and least-square solver.

All of the functions have the return type **cusolverStatus_t** and are explained in more detail in the chapters that follow.

cuSolverSP API

routine	data format	operation	output format	based on
csrslsvlu	csr	linear solver (ls)	vector (v)	LU (lu) with partial pivoting
csrslsvqr	csr	linear solver (ls)	vector (v)	QR factorization (qr)
csrslsvchol	csr	linear solver (ls)	vector (v)	Cholesky factorization (chol)
csrslsqvqr	csr	least-square solver (lsq)	vector (v)	QR factorization (qr)
csreigvsi	csr	eigenvalue solver (eig)	vector (v)	shift-inverse
csreigs	csr	number of eigenvalues in a box (eigs)		
csrsymrcm	csr	Symmetric Reverse Cuthill-McKee (symrcm)		

The cuSolverRF library routines are available for data type **double**. Most of the routines follow the naming convention:

```
cusolverRf_<operation>_[[Host]](...)
```

where the trailing optional Host qualifier indicates the data is accessed on the host versus on the device, which is the default. The **<operation>** can be **Setup**, **Analyze**, **Refactor**, **Solve**, **ResetValues**, **AccessBundledFactors** and **ExtractSplitFactors**.

Finally, the return type of the cuSolverRF library routines is **cusolverStatus_t**.

1.5. Asynchronous Execution

The cuSolver library functions prefer to keep asynchronous execution as much as possible. Developers can always use the `cudaDeviceSynchronize()` function to ensure that the execution of a particular cuSolver library routine has completed.

A developer can also use the `cudaMemcpy()` routine to copy data from the device to the host and vice versa, using the `cudaMemcpyDeviceToHost` and `cudaMemcpyHostToDevice` parameters, respectively. In this case there is no need to add a call to `cudaDeviceSynchronize()` because the call to `cudaMemcpy()` with the above parameters is blocking and completes only when the results are ready on the host.

1.6. Library Property

The `libraryPropertyType` data type is an enumeration of library property types. (ie. CUDA version X.Y.Z would yield `MAJOR_VERSION=X`, `MINOR_VERSION=Y`, `PATCH_LEVEL=Z`)

```
typedef enum libraryPropertyType_t
{
    MAJOR_VERSION,
    MINOR_VERSION,
    PATCH_LEVEL
} libraryPropertyType;
```

The following code can show the version of cusolver library.

```
int major=-1,minor=-1,patch=-1;
cusolverGetProperty(MAJOR_VERSION, &major);
cusolverGetProperty(MINOR_VERSION, &minor);
cusolverGetProperty(PATCH_LEVEL, &patch);
printf("CUSOLVER Version (Major,Minor,PatchLevel): %d.%d.%d\n",
major,minor,patch);
```

1.7. high precision package

The `cusolver` library uses high precision for iterative refinement when necessary.

Chapter 2.

USING THE CUSOLVER API

2.1. General description

This chapter describes how to use the cuSolver library API. It is not a reference for the cuSolver API data types and functions; that is provided in subsequent chapters.

2.1.1. Thread Safety

The library is thread-safe, and its functions can be called from multiple host threads.

2.1.2. Scalar Parameters

In the cuSolver API, the scalar parameters can be passed by reference on the host.

2.1.3. Parallelism with Streams

If the application performs several small independent computations, or if it makes data transfers in parallel with the computation, then CUDA streams can be used to overlap these tasks.

The application can conceptually associate a stream with each task. To achieve the overlap of computation between the tasks, the developer should:

1. Create CUDA streams using the function `cudaStreamCreate()`, and
2. Set the stream to be used by each individual cuSolver library routine by calling, for example, `cusolverDnSetStream()`, just prior to calling the actual cuSolverDN routine.

The computations performed in separate streams would then be overlapped automatically on the GPU, when possible. This approach is especially useful when the computation performed by a single task is relatively small, and is not enough to fill the GPU with work, or when there is a data transfer that can be performed in parallel with the computation.

2.1.4. Link Third-party LAPACK Library

Starting with CUDA 10.1 update 2, NVIDIA LAPACK library `liblapack_static.a` is a subset of LAPACK and only contains GPU accelerated `stedc` and `bdsqr`. The user has to link `libcusolver_static.a` with `liblapack_static.a` in order to build the application successfully. Prior to CUDA 10.1 update 2, the user can replace `liblapack_static.a` with a third-party LAPACK library, for example, MKL. In CUDA 10.1 update 2, the third-party LAPACK library no longer affects the behavior of cusolver library, neither functionality nor performance. Furthermore the user cannot use `liblapack_static.a` as a standalone LAPACK library because it is only a subset of LAPACK.



The `liblapack_static.a` library, which is the binary of CLAPACK-3.2.1, is a new feature of CUDA 10.0.

- ▶ If you use `libcusolver_static.a`, then you must link with `liblapack_static.a` explicitly, otherwise the linker will report missing symbols. No conflict of symbols between `liblapack_static.a` and other third-party LAPACK library, you are free to link the latter to your application.
- ▶ The `liblapack_static.a` is built inside `libcusolver.so`. Hence, if you use `libcusolver.so`, then you don't need to specify a LAPACK library. The `libcusolver.so` will not pick up any routines from the third-party LAPACK library even you link the application with it.

2.1.5. convention of info

Each LAPACK routine returns an **info** which indicates the position of invalid parameter. If **info** = **-i**, then i-th parameter is invalid. To be consistent with base-1 in LAPACK, **cusolver** does not report invalid **handle** into **info**. Instead, **cusolver** returns **CUSOLVER_STATUS_NOT_INITIALIZED** for invalid **handle**.

2.1.6. usage of _bufferSize

There is no `cudaMalloc` inside **cuSolver** library, the user must allocate the device workspace explicitly. The routine **xyz_bufferSize** is to query the size of workspace of the routine **xyz**, for example **xyz** = **potrf**. To make the API simple, **xyz_bufferSize** follows almost the same signature of **xyz** even it only depends on some parameters, for example, device pointer is not used to decide the size of workspace. In most cases, **xyz_bufferSize** is called in the beginning before actual device data (pointing by a device pointer) is prepared or before the device pointer is allocated. In such case, the user can pass null pointer to **xyz_bufferSize** without breaking the functionality.

2.2. cuSolver Types Reference

2.2.1. cuSolverDN Types

The `float`, `double`, `cuComplex`, and `cuDoubleComplex` data types are supported. The first two are standard C data types, while the last two are exported from `cuComplex.h`. In addition, cuSolverDN uses some familiar types from cuBlas.

2.2.1.1. cusolverDnHandle_t

This is a pointer type to an opaque cuSolverDN context, which the user must initialize by calling `cusolverDnCreate()` prior to calling any other library function. An un-initialized Handle object will lead to unexpected behavior, including crashes of cuSolverDN. The handle created and returned by `cusolverDnCreate()` must be passed to every cuSolverDN function.

2.2.1.2. cublasFillMode_t

The type indicates which part (lower or upper) of the dense matrix was filled and consequently should be used by the function. Its values correspond to Fortran characters `'L'` or `'l'` (lower) and `'U'` or `'u'` (upper) that are often used as parameters to legacy BLAS implementations.

Value	Meaning
<code>CUBLAS_FILL_MODE_LOWER</code>	the lower part of the matrix is filled
<code>CUBLAS_FILL_MODE_UPPER</code>	the upper part of the matrix is filled

2.2.1.3. cublasOperation_t

The `cublasOperation_t` type indicates which operation needs to be performed with the dense matrix. Its values correspond to Fortran characters `'N'` or `'n'` (non-transpose), `'T'` or `'t'` (transpose) and `'C'` or `'c'` (conjugate transpose) that are often used as parameters to legacy BLAS implementations.

Value	Meaning
<code>CUBLAS_OP_N</code>	the non-transpose operation is selected
<code>CUBLAS_OP_T</code>	the transpose operation is selected
<code>CUBLAS_OP_C</code>	the conjugate transpose operation is selected

2.2.1.4. cusolverEigType_t

The `cusolverEigType_t` type indicates which type of eigenvalue solver is. Its values correspond to Fortran integer `1` ($A*x = \lambda*B*x$), `2` ($A*B*x = \lambda*x$), `3` ($B*A*x = \lambda*x$), used as parameters to legacy LAPACK implementations.

Value	Meaning
-------	---------

CUSOLVER_EIG_TYPE_1	$A*x = \lambda B*x$
CUSOLVER_EIG_TYPE_2	$A*B*x = \lambda x$
CUSOLVER_EIG_TYPE_3	$B*A*x = \lambda x$

2.2.1.5. cusolverEigMode_t

The **`cusolverEigMode_t`** type indicates whether or not eigenvectors are computed. Its values correspond to Fortran character 'N' (only eigenvalues are computed), 'V' (both eigenvalues and eigenvectors are computed) used as parameters to legacy LAPACK implementations.

Value	Meaning
CUSOLVER_EIG_MODE_NOVECTOR	only eigenvalues are computed
CUSOLVER_EIG_MODE_VECTOR	both eigenvalues and eigenvectors are computed

2.2.1.6. cusolverIRSRefinement_t

The **`cusolverIRSRefinement_t`** type indicates which solver type would be used for the specific cusolver function. Most of our experimentation shows that CUSOLVER_IRS_REFINE_GMRES is the best option.

More details about the refinement process can be found in Azzam Haidar, Stanimire Tomov, Jack Dongarra, and Nicholas J. Higham. 2018. Harnessing GPU tensor cores for fast FP16 arithmetic to speed up mixed-precision iterative refinement solvers. In Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC '18). IEEE Press, Piscataway, NJ, USA, Article 47, 11 pages.

Value	Meaning
CUSOLVER_IRS_REFINE_NOT_SET	Solver is not set. Default value.
CUSOLVER_IRS_REFINE_NONE	No solver
CUSOLVER_IRS_REFINE_CLASSICAL	Classical iterative refinement solver. Similar to the one used in LAPACK routines.
CUSOLVER_IRS_REFINE_GMRES	GMRES (Generalized Minimal Residual) based iterative refinement solver. In recent study, the GMRES method has drawn the scientific community attention for its ability to be used as refinement solver that outperforms the classical iterative refinement method.
CUSOLVER_IRS_REFINE_CLASSICAL_GMRES	Classical iterative refinement solver that uses the GMRES (Generalized Minimal Residual) internally to solve the correction equation at each iteration. We call the <i>classical refinement iteration</i> the outer iteration while the <i>GMRES</i> is called inner iteration. Note that if the tolerance of the inner GMRES is set very low, let say to machine precision, then the outer <i>classical refinement iteration</i> will performs only one iteration and thus this option will behaves like CUSOLVER_IRS_REFINE_GMRES.
CUSOLVER_IRS_REFINE_GMRES_GMRES	Similar to CUSOLVER_IRS_REFINE_CLASSICAL_GMRES which

is classical refinement process that uses GMRES to solve the inner correction system, here it is a GMRES (Generalized Minimal Residual) based iterative refinement solver that uses another GMRES internally to solve the preconditioned system.
--

2.2.1.7. cusolverDnIRSParams_t

This is a pointer type to an opaque `cusolverDnIRSParams_t` structure, which holds parameters for the iterative refinement linear solvers such as `cusolverDnXgesv()`. Use corresponding helper functions described below to either Create/Destroy this structure or Set/Get solver parameters.

2.2.1.8. cusolverDnIRSInfos_t

This is a pointer type to an opaque `cusolverDnIRSInfos_t` structure, which holds information about the performed call to an iterative refinement linear solver (e.g., `cusolverDnXgesv()`). Use corresponding helper functions described below to either Create/Destroy this structure or retrieve solve information.

2.2.1.9. cusolverStatus_t

This is the same as `cusolverStatus_t` in the sparse LAPACK section.

2.2.2. cuSolverSP Types

The `float`, `double`, `cuComplex`, and `cuDoubleComplex` data types are supported. The first two are standard C data types, while the last two are exported from `cuComplex.h`.

2.2.2.1. cusolverSpHandle_t

This is a pointer type to an opaque `cuSolverSP` context, which the user must initialize by calling `cusolverSpCreate()` prior to calling any other library function. An un-initialized Handle object will lead to unexpected behavior, including crashes of `cuSolverSP`. The handle created and returned by `cusolverSpCreate()` must be passed to every `cuSolverSP` function.

2.2.2.2. cusparseMatDescr_t

We have chosen to keep the same structure as exists in `cuSparse` to describe the shape and properties of a matrix. This enables calls to either `cuSparse` or `cuSolver` using the same matrix description.

```
typedef struct {
    cusparseMatrixType_t MatrixType;
    cusparseFillMode_t FillMode;
    cusparseDiagType_t DiagType;
    cusparseIndexBase_t IndexBase;
} cusparseMatDescr_t;
```

Please read documentation of CUSPARSE Library to understand each field of `cusparseMatDescr_t`.

2.2.2.3. cusolverStatus_t

This is a status type returned by the library functions and it can have the following values.

<code>CUSOLVER_STATUS_SUCCESS</code>	The operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INIT</code>	<p>The cuSolver library was not initialized. This is usually caused by the lack of a prior call, an error in the CUDA Runtime API called by the cuSolver routine, or an error in the hardware setup.</p> <p>To correct: call <code>cusolverCreate()</code> prior to the function call; and check that the hardware, an appropriate version of the driver, and the cuSolver library are correctly installed.</p>
<code>CUSOLVER_STATUS_ALLOC_FAIL</code>	<p>Resource allocation failed inside the cuSolver library. This is usually caused by a <code>cudaMalloc()</code> failure.</p> <p>To correct: prior to the function call, deallocate previously allocated memory as much as possible.</p>
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	<p>An unsupported value or parameter was passed to the function (a negative vector size, for example).</p> <p>To correct: ensure that all the parameters being passed have valid values.</p>
<code>CUSOLVER_STATUS_ARCH_MISMATCH</code>	<p>The function requires a feature absent from the device architecture; usually caused by the lack of support for atomic operations or double precision.</p> <p>To correct: compile and run the application on a device with compute capability 2.0 or above.</p>
<code>CUSOLVER_STATUS_EXECUTION_FAILED</code>	<p>The GPU program failed to execute. This is often caused by a launch failure of the kernel on the GPU, which can be caused by multiple reasons.</p> <p>To correct: check that the hardware, an appropriate version of the driver, and the cuSolver library are correctly installed.</p>
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	<p>An internal cuSolver operation failed. This error is usually caused by a <code>cudaMemcpyAsync()</code> failure.</p> <p>To correct: check that the hardware, an appropriate version of the driver, and the cuSolver library are correctly installed. Also, check that the memory passed as a parameter to the routine is not being deallocated prior to the routine's completion.</p>
<code>CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED</code>	<p>The matrix type is not supported by this function. This is usually caused by passing an invalid matrix descriptor to the function.</p> <p>To correct: check that the fields in <code>descrA</code> were set correctly.</p>

2.2.3. cuSolverRF Types

cuSolverRF only supports **double**.

2.2.3.1. cusolverRfHandle_t

The **`cusolverRfHandle_t`** is a pointer to an opaque data structure that contains the cuSolverRF library handle. The user must initialize the handle by calling **`cusolverRfCreate()`** prior to any other cuSolverRF library calls. The handle is passed to all other cuSolverRF library calls.

2.2.3.2. cusolverRfMatrixFormat_t

The **`cusolverRfMatrixFormat_t`** is an enum that indicates the input/output matrix format assumed by the **`cusolverRfSetupDevice()`**, **`cusolverRfSetupHost()`**, **`cusolverRfResetValues()`**, **`cusolverRfExtractBundledFactorsHost()`** and **`cusolverRfExtractSplitFactorsHost()`** routines.

Value	Meaning
<code>CUSOLVER_MATRIX_FORMAT_CSR</code>	matrix format CSR is assumed. (default)
<code>CUSOLVER_MATRIX_FORMAT_CSC</code>	matrix format CSC is assumed.

2.2.3.3. cusolverRfNumericBoostReport_t

The **`cusolverRfNumericBoostReport_t`** is an enum that indicates whether numeric boosting (of the pivot) was used during the **`cusolverRfRefactor()`** and **`cusolverRfSolve()`** routines. The numeric boosting is disabled by default.

Value	Meaning
<code>CUSOLVER_NUMERIC_BOOST_NOT_USED</code>	numeric boosting not used. (default)
<code>CUSOLVER_NUMERIC_BOOST_USED</code>	numeric boosting used.

2.2.3.4. cusolverRfResetValuesFastMode_t

The **`cusolverRfResetValuesFastMode_t`** is an enum that indicates the mode used for the **`cusolverRfResetValues()`** routine. The fast mode requires extra memory and is recommended only if very fast calls to **`cusolverRfResetValues()`** are needed.

Value	Meaning
<code>CUSOLVER_RESET_VALUES_FAST_MODE_OFF</code>	fast mode disabled. (default)
<code>CUSOLVER_RESET_VALUES_FAST_MODE_ON</code>	fast mode enabled.

2.2.3.5. cusolverRfFactorization_t

The **`cusolverRfFactorization_t`** is an enum that indicates which (internal) algorithm is used for refactorization in the **`cusolverRfRefactor()`** routine.

Value	Meaning
-------	---------

CUSOLVER_FACTORIZATION_ALG0	algorithm 0. (default)
CUSOLVER_FACTORIZATION_ALG1	algorithm 1.
CUSOLVER_FACTORIZATION_ALG2	algorithm 2. Domino-based scheme.

2.2.3.6. cusolverRfTriangularSolve_t

The **`cusolverRfTriangularSolve_t`** is an enum that indicates which (internal) algorithm is used for triangular solve in the **`cusolverRfSolve()`** routine.

Value	Meaning
CUSOLVER_TRIANGULAR_SOLVE_ALG0	algorithm 0.
CUSOLVER_TRIANGULAR_SOLVE_ALG1	algorithm 1. (default)
CUSOLVER_TRIANGULAR_SOLVE_ALG2	algorithm 2. Domino-based scheme.
CUSOLVER_TRIANGULAR_SOLVE_ALG3	algorithm 3. Domino-based scheme.

2.2.3.7. cusolverRfUnitDiagonal_t

The **`cusolverRfUnitDiagonal_t`** is an enum that indicates whether and where the unit diagonal is stored in the input/output triangular factors in the **`cusolverRfSetupDevice()`**, **`cusolverRfSetupHost()`** and **`cusolverRfExtractSplitFactorsHost()`** routines.

Value	Meaning
CUSOLVER_UNIT_DIAGONAL_STORED_L	unit diagonal is stored in lower triangular factor. (default)
CUSOLVER_UNIT_DIAGONAL_STORED_U	unit diagonal is stored in upper triangular factor.
CUSOLVER_UNIT_DIAGONAL_ASSUMED_L	unit diagonal is assumed in lower triangular factor.
CUSOLVER_UNIT_DIAGONAL_ASSUMED_U	unit diagonal is assumed in upper triangular factor.

2.2.3.8. cusolverStatus_t

The **`cusolverStatus_t`** is an enum that indicates success or failure of the cuSolverRF library call. It is returned by all the cuSolver library routines, and it uses the same enumerated values as the sparse and dense Lapack routines.

2.3. cuSolver Formats Reference

2.3.1. Index Base Format

The CSR or CSC format requires either zero-based or one-based index for a sparse matrix **A**. The GLU library supports only zero-based indexing. Otherwise, both one-based and zero-based indexing are supported in cuSolver.

2.3.2. Vector (Dense) Format

The vectors are assumed to be stored linearly in memory. For example, the vector

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

is represented as

$$(x_1 \ x_2 \ \dots \ x_n)$$

2.3.3. Matrix (Dense) Format

The dense matrices are assumed to be stored in column-major order in memory. The sub-matrix can be accessed using the leading dimension of the original matrix. For example, the $\mathbf{m} \times \mathbf{n}$ (sub-)matrix

$$\begin{pmatrix} a_{1,1} & \dots & a_{1,n} \\ a_{2,1} & \dots & a_{2,n} \\ \vdots & & \vdots \\ a_{m,1} & \dots & a_{m,n} \end{pmatrix}$$

is represented as

$$\begin{pmatrix} a_{1,1} & \dots & a_{1,n} \\ a_{2,1} & \dots & a_{2,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \dots & a_{m,n} \\ \vdots & \ddots & \vdots \\ a_{lda,1} & \dots & a_{lda,n} \end{pmatrix}$$

with its elements arranged linearly in memory as

$$(a_{1,1} \ a_{2,1} \ \dots \ a_{m,1} \ \dots \ a_{lda,1} \ \dots \ a_{1,n} \ a_{2,n} \ \dots \ a_{m,n} \ \dots \ a_{lda,n})$$

where $lda \geq m$ is the leading dimension of \mathbf{A} .

2.3.4. Matrix (CSR) Format

In CSR format the matrix is represented by the following parameters

parameter	type	size	Meaning
n	(int)		the number of rows (and columns) in the matrix.
nnz	(int)		the number of non-zero elements in the matrix.
csrRowPtr	(int *)	n+1	the array of offsets corresponding to the start of each row in the arrays csrColInd and csrVal . This array has also an extra entry at

			the end that stores the number of non-zero elements in the matrix.
<code>csrColInd</code>	<code>(int *)</code>	<code>nnz</code>	the array of column indices corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by row and by column within each row.
<code>csrVal</code>	<code>(S D C Z) *</code>	<code>nnz</code>	the array of values corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by row and by column within each row.

Note that in our CSR format sparse matrices are assumed to be stored in row-major order, in other words, the index arrays are first sorted by row indices and then within each row by column indices. Also it is assumed that each pair of row and column indices appears only once.

For example, the **4x4** matrix

$$A = \begin{pmatrix} 1.0 & 3.0 & 0.0 & 0.0 \\ 0.0 & 4.0 & 6.0 & 0.0 \\ 2.0 & 5.0 & 7.0 & 8.0 \\ 0.0 & 0.0 & 0.0 & 9.0 \end{pmatrix}$$

is represented as

`csrRowPtr` = (0 2 4 8 9)

`csrColInd` = (0 1 1 2 0 1 2 3 3)

`csrVal` = (1.0 3.0 4.0 6.0 2.0 5.0 7.0 8.0 9.0)

2.3.5. Matrix (CSC) Format

In CSC format the matrix is represented by the following parameters

parameter	type	size	Meaning
<code>n</code>	<code>(int)</code>		the number of rows (and columns) in the matrix.
<code>nnz</code>	<code>(int)</code>		the number of non-zero elements in the matrix.
<code>cscColPtr</code>	<code>(int *)</code>	<code>n+1</code>	the array of offsets corresponding to the start of each column in the arrays <code>cscRowInd</code> and <code>cscVal</code> . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix.
<code>cscRowInd</code>	<code>(int *)</code>	<code>nnz</code>	the array of row indices corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by column and by row within each column.
<code>cscVal</code>	<code>(S D C Z) *</code>	<code>nnz</code>	the array of values corresponding to the non-zero elements in the matrix. It is assumed

			that this array is sorted by column and by row within each column.
--	--	--	--

Note that in our CSC format sparse matrices are assumed to be stored in column-major order, in other words, the index arrays are first sorted by column indices and then within each column by row indices. Also it is assumed that each pair of row and column indices appears only once.

For example, the **4x4** matrix

$$A = \begin{pmatrix} 1.0 & 3.0 & 0.0 & 0.0 \\ 0.0 & 4.0 & 6.0 & 0.0 \\ 2.0 & 5.0 & 7.0 & 8.0 \\ 0.0 & 0.0 & 0.0 & 9.0 \end{pmatrix}$$

is represented as

`cscColPtr = (0 2 5 7 9)`

`cscRowInd = (0 2 0 1 2 1 2 2 3)`

`cscVal = (1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0)`

2.4. cuSolverDN: dense LAPACK Function Reference

This chapter describes the API of cuSolverDN, which provides a subset of dense LAPACK functions.

2.4.1. cuSolverDN Helper Function Reference

The cuSolverDN helper functions are described in this section.

2.4.1.1. cusolverDnCreate()

```
cusolverStatus_t
cusolverDnCreate(cusolverDnHandle_t *handle);
```

This function initializes the cuSolverDN library and creates a handle on the cuSolverDN context. It must be called before any other cuSolverDN API function is invoked. It allocates hardware resources necessary for accessing the GPU.

parameter	Memory	In/out	Meaning
handle	host	output	the pointer to the handle to the cuSolverDN context.

Status Returned

CUSOLVER_STATUS_SUCCESS	the initialization succeeded.
-------------------------	-------------------------------

CUSOLVER_STATUS_NOT_INITIALIZED	the CUDA Runtime initialization failed.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.

2.4.1.2. cusolverDnDestroy()

```
cusolverStatus_t
cusolverDnDestroy(cusolverDnHandle_t handle);
```

This function releases CPU-side resources used by the cuSolverDN library.

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.

Status Returned

CUSOLVER_STATUS_SUCCESS	the shutdown succeeded.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

2.4.1.3. cusolverDnSetStream()

```
cusolverStatus_t
cusolverDnSetStream(cusolverDnHandle_t handle, cudaStream_t streamId)
```

This function sets the stream to be used by the cuSolverDN library to execute its routines.

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
streamId	host	input	the stream to be used by the library.

Status Returned

CUSOLVER_STATUS_SUCCESS	the stream was set successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

2.4.1.4. cusolverDnGetStream()

```
cusolverStatus_t
cusolverDnGetStream(cusolverDnHandle_t handle, cudaStream_t *streamId)
```

This function sets the stream to be used by the cuSolverDN library to execute its routines.

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
streamId	host	output	the stream to be used by the library.

Status Returned

CUSOLVER_STATUS_SUCCESS	the stream was set successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

2.4.1.5. cusolverDnCreateSyevjInfo()

```
cusolverStatus_t
cusolverDnCreateSyevjInfo(
    syevjInfo_t *info);
```

This function creates and initializes the structure of **syevj**, **syevjBatched** and **sygvj** to default values.

parameter	Memory	In/out	Meaning
info	host	output	the pointer to the structure of syevj.

Status Returned

CUSOLVER_STATUS_SUCCESS	the structure was initialized successfully.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.

2.4.1.6. cusolverDnDestroySyevjInfo()

```
cusolverStatus_t
cusolverDnDestroySyevjInfo(
    syevjInfo_t info);
```

This function destroys and releases any memory required by the structure.

parameter	Memory	In/out	Meaning
info	host	input	the structure of syevj.

Status Returned

CUSOLVER_STATUS_SUCCESS	the resources are released successfully.
-------------------------	--

2.4.1.7. cusolverDnXsyevjSetTolerance()

```
cusolverStatus_t
cusolverDnXsyevjSetTolerance(
    syevjInfo_t info,
    double tolerance)
```

This function configures tolerance of **syevj**.

parameter	Memory	In/out	Meaning
info	host	in/out	the pointer to the structure of syevj.
tolerance	host	input	accuracy of numerical eigenvalues.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
-------------------------	---------------------------------------

2.4.1.8. cusolverDnXsyevjSetMaxSweeps()

```
cusolverStatus_t
cusolverDnXsyevjSetMaxSweeps(
    syevjInfo_t info,
    int max_sweeps)
```

This function configures maximum number of sweeps in **syevj**. The default value is 100.

parameter	Memory	In/out	Meaning
info	host	in/out	the pointer to the structure of syevj.
max_sweeps	host	input	maximum number of sweeps.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
-------------------------	---------------------------------------

2.4.1.9. cusolverDnXsyevjSetSortEig()

```
cusolverStatus_t
cusolverDnXsyevjSetSortEig(
    syevjInfo_t info,
    int sort_eig)
```

if **sort_eig** is zero, the eigenvalues are not sorted. This function only works for **syevjBatched**. **syevj** and **sygvj** always sort eigenvalues in ascending order. By default, eigenvalues are always sorted in ascending order.

parameter	Memory	In/out	Meaning
info	host	in/out	the pointer to the structure of syevj.
sort_eig	host	input	if sort_eig is zero, the eigenvalues are not sorted.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
-------------------------	---------------------------------------

2.4.1.10. cusolverDnXsyevjGetResidual()

```
cusolverStatus_t
cusolverDnXsyevjGetResidual(
    cusolverDnHandle_t handle,
    syevjInfo_t info,
    double *residual)
```

This function reports residual of **syevj** or **sygvj**. It does not support **syevjBatched**. If the user calls this function after **syevjBatched**, the error **CUSOLVER_STATUS_NOT_SUPPORTED** is returned.

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
info	host	input	the pointer to the structure of syevj.
residual	host	output	residual of syevj.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_SUPPORTED	does not support batched version

2.4.1.11. cusolverDnXsyevjGetSweeps()

```
cusolverStatus_t
cusolverDnXsyevjGetSweeps(
    cusolverDnHandle_t handle,
    syevjInfo_t info,
    int *executed_sweeps)
```

This function reports number of executed sweeps of **syevj** or **sygvj**. It does not support **syevjBatched**. If the user calls this function after **syevjBatched**, the error **CUSOLVER_STATUS_NOT_SUPPORTED** is returned.

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
info	host	input	the pointer to the structure of syevj.
executed_sweeps	host	output	number of executed sweeps.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_SUPPORTED	does not support batched version

2.4.1.12. cusolverDnCreateGesvdjInfo()

```
cusolverStatus_t
cusolverDnCreateGesvdjInfo(
    gesvdjInfo_t *info);
```

This function creates and initializes the structure of **gesvdj** and **gesvdjBatched** to default values.

parameter	Memory	In/out	Meaning
info	host	output	the pointer to the structure of gesvdj.

Status Returned

CUSOLVER_STATUS_SUCCESS	the structure was initialized successfully.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.

2.4.1.13. cusolverDnDestroyGesvdjInfo()

```
cusolverStatus_t
cusolverDnDestroyGesvdjInfo(
    gesvdjInfo_t info);
```

This function destroys and releases any memory required by the structure.

parameter	Memory	In/out	Meaning
info	host	input	the structure of gesvdj.

Status Returned

CUSOLVER_STATUS_SUCCESS	the resources are released successfully.
-------------------------	--

2.4.1.14. cusolverDnXgesvdjSetTolerance()

```
cusolverStatus_t
cusolverDnXgesvdjSetTolerance(
    gesvdjInfo_t info,
    double tolerance)
```

This function configures tolerance of **gesvdj**.

parameter	Memory	In/out	Meaning
info	host	in/out	the pointer to the structure of gesvdj.
tolerance	host	input	accuracy of numerical singular values.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
-------------------------	---------------------------------------

2.4.1.15. cusolverDnXgesvdjSetMaxSweeps()

```
cusolverStatus_t
cusolverDnXgesvdjSetMaxSweeps(
    gesvdjInfo_t info,
    int max_sweeps)
```

This function configures maximum number of sweeps in **gesvdj**. The default value is 100.

parameter	Memory	In/out	Meaning
info	host	in/out	the pointer to the structure of gesvdj.
max_sweeps	host	input	maximum number of sweeps.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
-------------------------	---------------------------------------

2.4.1.16. cusolverDnXgesvdjSetSortEig()

```
cusolverStatus_t
cusolverDnXgesvdjSetSortEig(
    gesvdjInfo_t info,
    int sort_svd)
```

if **sort_svd** is zero, the singular values are not sorted. This function only works for **gesvdjBatched**. **gesvdj** always sorts singular values in descending order. By default, singular values are always sorted in descending order.

parameter	Memory	In/out	Meaning
info	host	in/out	the pointer to the structure of gesvdj.
sort_svd	host	input	if sort_svd is zero, the singular values are not sorted.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
--------------------------------	---------------------------------------

2.4.1.17. cusolverDnXgesvdjGetResidual()

```
cusolverStatus_t
cusolverDnXgesvdjGetResidual(
    cusolverDnHandle_t handle,
    gesvdjInfo_t info,
    double *residual)
```

This function reports residual of **gesvdj**. It does not support **gesvdjBatched**. If the user calls this function after **gesvdjBatched**, the error **CUSOLVER_STATUS_NOT_SUPPORTED** is returned.

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
info	host	input	the pointer to the structure of gesvdj.
residual	host	output	residual of gesvdj.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_SUPPORTED	does not support batched version

2.4.1.18. cusolverDnXgesvdjGetSweeps()

```
cusolverStatus_t
cusolverDnXgesvdjGetSweeps(
    cusolverDnHandle_t handle,
    gesvdjInfo_t info,
    int *executed_sweeps)
```

This function reports number of executed sweeps of **gesvdj**. It does not support **gesvdjBatched**. If the user calls this function after **gesvdjBatched**, the error **CUSOLVER_STATUS_NOT_SUPPORTED** is returned.

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
info	host	input	the pointer to the structure of gesvdj.
executed_sweeps	host	output	number of executed sweeps.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_SUPPORTED	does not support batched version

2.4.1.19. cusolverDnIRSParamsCreate()

```
cusolverStatus_t
cusolverDnIRSParamsCreate(cusolverDnIRSParams_t *params);
```

This function creates and initializes the structure of parameters for an IRS solver such as the **cusolverDnIRSxgesv()** function to default values. For this release, this function is valid for only one call to an IRS solver, thus each new call to the IRS solver will requires its own Params structure. This restriction is going to be removed in future release and then if user want to reuse the same configuration to many call to an IRS solver it will allow him.

parameter	Memory	In/out	Meaning
params	host	output	Pointer to the cusolverDnIRSParams_t Params structure

Status Returned

CUSOLVER_STATUS_SUCCESS	The structure was created and initialized successfully.
CUSOLVER_STATUS_ALLOC_FAILED	The resources could not be allocated.

2.4.1.20. cusolverDnIRSParamsDestroy()

```
cusolverStatus_t
cusolverDnIRSParamsDestroy(cusolverDnIRSParams_t params);
```

This function destroys and releases any memory required by the Params structure. Since the Infos structure (see **cusolverDnIRSInfosCreate()** for more details) depends on the Params structure, this function cannot be called to destroy the Params structure if the Infos structure was not destroyed. For this release, this function is valid for only one call to an IRS solver, thus each call to an IRS solver should have its own Params and Infos structure. This restriction is going to be removed in future release and then if user want to reuse the same configuration to many call to an IRS solver it will allow him.

parameter	Memory	In/out	Meaning
-----------	--------	--------	---------

params	host	input	The cusolverDnIRSPParams_t Params structure
---------------	-------------	--------------	---

Status Returned

CUSOLVER_STATUS_SUCCESS	The resources are released successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.
CUSOLVER_STATUS_IRS_INFOS_NOT_DESTROYED	Not all the Infos structure associated with this Params structure have been destroyed yet.

2.4.1.21. cusolverDnIRSPParamsSetRefinementSolver()

```
cusolverStatus_t
cusolverDnIRSPParamsSetRefinementSolver(
    cusolverDnIRSPParams_t params,
    cusolverIRSRefinement_t solver);
```

This function sets the refinement solver to be used in the Iterative Refinement Solver functions such as the **cusolverDnIRSXgesv()** function. Details about values that can be set to and their meaning can be found in the **cusolverIRSRefinement_t** type section.

parameter	Memory	In/out	Meaning
params	host	in/out	The cusolverDnIRSPParams_t Params structure
solver	host	input	Type of the refinement solver to be used by the IRS solver such as cusolverDnIRSXgesv()

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.

2.4.1.22. cusolverDnIRSPParamsSetSolverMainPrecision()

```
cusolverStatus_t
cusolverDnIRSPParamsSetSolverMainPrecision(
    cusolverDnIRSPParams_t params,
    cudaDataType solver_main_precision);
```

This function sets the main precision (e.g., the INOUT data type) of the IRS solver. The value set here should be the same cuda datatype as the third argument on the call to the IRS solver. Note that, the user has to set the main precision before a first call to the IRS solver because it is NOT set by default with the Params creation. He can set it by either calling this function or **cusolverDnIRSPParamsSetSolverPrecisions()**. Possible values are described in the table of the corresponding IRS solver for example, see the description of the third argument of the **cusolverDnIRSXgesv()** IRS function.

parameter	Memory	In/out	Meaning
------------------	---------------	---------------	----------------

<code>params</code>	<code>host</code>	<code>in/out</code>	The <code>cusolverDnIRSPParams_t</code> Params structure
<code>solver_main_precision</code>	<code>host</code>	<code>input</code>	Allowed cuda datatype (for example <code>CUDA_R_FP64</code>). See the corresponding IRS solver for the table of supported precisions.

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	The operation completed successfully.
<code>CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE</code>	The Params structure was not created.

2.4.1.23. `cusolverDnIRSPParamsSetSolverLowestPrecision()`

```
cusolverStatus_t
cusolverDnIRSPParamsSetSolverLowestPrecision(
    cusolverDnIRSPParams_t params,
    cudaDataType lowest_precision_type);
```

This function sets the lowest precision that will be used by Iterative Refinement Solver. Note that, the user has to set the lowest precision before a first call to the IRS solver because it is NOT set by default with the Params creation. He can set it by either calling this function or `cusolverDnIRSPParamsSetSolverPrecisions()`. Usually this precision define the speedup that can be achieved. The ratio of the performance of the lowest precision over the `inout_data_type` precision define somehow the upper bound of the speedup that could be obtained. More precisely, it depends of many factors, but for large matrices sizes, it is the ratio of the matrix-matrix rank-k product (e.g., rank-k GEMM where k is around 256) that define the possible speedup. For instance, if the `inout` precision is real double precision FP64 and the lowest precision is FP32, then we can expect a speedup of at most 2X for large problem sizes. If the lowest precision was FP16, then we can expect 3X-4X. A reasonable strategy should take the number of right-hand sides and the size of the matrix as well as the convergence rate into account.

parameter	Memory	In/out	Meaning
<code>params</code>	<code>host</code>	<code>in/out</code>	The <code>cusolverDnIRSPParams_t</code> Params structure
<code>lowest_precision_type</code>	<code>host</code>	<code>input</code>	Allowed cuda datatype (for example <code>CUDA_R_FP16</code>). See the corresponding IRS solver for the table of supported precisions.

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	The operation completed successfully.
<code>CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE</code>	The Params structure was not created.

2.4.1.24. cusolverDnIRSParamsSetSolverPrecisions()

```
cusolverStatus_t
cusolverDnIRSParamsSetSolverPrecisions(
    cusolverDnIRSParams_t params,
    cudaDataType solver_main_precision,
    cudaDataType solver_lowest_precision );
```

This function set both, the main and the lowest precision of the IRS solver. It is a wrappers to both **`cusolverDnIRSParamsSetSolverMainPrecision()`** and **`cusolverDnIRSParamsSetSolverLowestPrecision()`**. Note that, the user has to set both the main and the lowest precision before a first call to the IRS solver because they are NOT set by default with the Params creation. He can set it by either calling this function or both functions **`cusolverDnIRSParamsSetSolverLowestPrecisions()`** and **`cusolverDnIRSParamsSetSolverMainPrecisions()`**.

parameter	Memory	In/out	Meaning
params	host	in/out	The cusolverDnIRSParams_t Params structure
solver_main_precision	host	input	Allowed cuda datatype (for example CUDA_R_FP64). See the corresponding IRS solver for the table of supported precisions.
solver_lowest_precision	host	input	Allowed cuda datatype (for example CUDA_R_FP16). See the corresponding IRS solver for the table of supported precisions.

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.

2.4.1.25. cusolverDnIRSParamsSetTol()

```
cusolverStatus_t
cusolverDnIRSParamsSetTol(
    cusolverDnIRSParams_t params,
    cudaDataType data_type,
    double val );
```

This function sets the tolerance for the refinement solver. By default it is set such that, all the RHS satisfy:

$$\mathbf{RNRM} < \mathbf{SQRT(N)} * \mathbf{XNRM} * \mathbf{ANRM} * \mathbf{EPS} * \mathbf{BWDMAX} \quad \text{where}$$

- ▶ **RNRM** is the infinity-norm of the residual
- ▶ **XNRM** is the infinity-norm of the solution
- ▶ **ANRM** is the infinity-operator-norm of the matrix A
- ▶ **EPS** is the machine epsilon that matches LAPACK ϵLAMCH('Epsilon')

The value **BWDMAX** is fixed to 1.0.

The user can use this function to change the tolerance to a lower or higher value. Our goal is to give the user as much control as we can such a way he can investigate and control every detail of the IRS solver.

parameter	Memory	In/out	Meaning
params	host	in/out	The cusolverDnIRSParams_t Params structure
data_type	host	input	cuda datatype of the inout_data_type
val	host	input	double precision value to which the refinement tolerance will be set internally.

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.

2.4.1.26. cusolverDnIRSParamsSetTolInner()

```
cusolverStatus_t
cusolverDnIRSParamsSetTolInner(
    cusolverDnIRSParams_t params,
    cudaDataType data_type,
    double val );
```

This function sets the tolerance for the inner refinement solver when the refinement solver consists of two level solvers (e.g., CUSOLVER_IRS_REFINE_CLASSICAL_GMRES or CUSOLVER_IRS_REFINE_GMRES_GMRES). It is not referenced in case of one level refinement solver such as CUSOLVER_IRS_REFINE_CLASSICAL or CUSOLVER_IRS_REFINE_GMRES. This function set the tolerance for the inner solver (e.g. the inner GMRES). For example, if the RefinementSolver was set to CUSOLVER_IRS_REFINE_CLASSICAL_GMRES setting this tolerance mean that the inner GMRES solver will converge to that tolerance at each outer iteration of the classical refinement solver. It is set to 1e-4 by default. Our goal is to give the user as much control as we can such a way he can investigate and control every detail of the IRS solver.

parameter	Memory	In/out	Meaning
params	host	in/out	The cusolverDnIRSParams_t Params structure
data_type	host	input	The cuda datatype (for example CUDA_R_FP64) of the inout data
val	host	input	Double precision real value to which the refinement tolerance will be set internally.

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.

2.4.1.27. cusolverDnIRSParamsSetMaxIters()

```
cusolverStatus_t
cusolverDnIRSParamsSetMaxIters(
    cusolverDnIRSParams_t params,
    int max_iters);
```

This function sets the total number of allowed refinement iterations after which solver will stop. Total means the maximum number of iterations allowed (e.g., outer and inner iterations when two level refinement solver is set). Default value is set to 50. Our goal is to give the user as much control as we can such a way he can investigate and control every detail of the IRS solver.

parameter	Memory	In/out	Meaning
params	host	in/out	The cusolverDnIRSParams_t Params structure
max_iters	host	input	Maximum total number of iterations allowed for the refinement solver

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.

2.4.1.28. cusolverDnIRSParamsSetMaxItersInner()

```
cusolverStatus_t
cusolverDnIRSParamsSetMaxItersInner(
    cusolverDnIRSParams_t params,
    cusolver_int_t maxiters_inner );
```

This function sets the maximal number of iterations allowed for the inner refinement solver. It is not referenced in case of one level refinement solver such as CUSOLVER_IRS_REFINE_CLASSICAL or CUSOLVER_IRS_REFINE_GMRES. The inner refinement solver will stop after reaching either the inner tolerance or the MaxItersInner value. By default, it is set to MaxIters (e.g., 50). Note that this value could not be larger than MaxIters since MaxIters is the total number of allowed iterations. Note that, if user call to set MaxIters after calling this function, the MaxIters has priority and will overwrite MaxItersInner to the minimum value of (MaxIters, MaxItersInner). Our goal is to give the user as much control as we can such a way he can investigate and control every detail of the IRS solver.

parameter	Memory	In/out	Meaning
params	host	in/out	The cusolverDnIRSParams_t Params structure
maxiters_inner	host	input	Maximum number of allowed inner iterations for the refinement solver. Meaningful when the refinement solver is a two levels solver such as CUSOLVER_IRS_REFINE_CLASSICAL_GMRES

			or CUSOLVER_IRS_REFINE_GMRES_GMRES. Value should be less or equal to MaxIters.
--	--	--	---

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.
CUSOLVER_STATUS_IRS_PARAMS_INVALID	if the value was larger than MaxIters.

2.4.1.29. cusolverDnIRSParamsGetMaxIters()

```
cusolverStatus_t
cusolverDnIRSParamsGetMaxIters(
    cusolverDnIRSParams_t params,
    cusolver_int_t *maxiters );
```

This function returns the maximal number of iterations MaxIters that is currently set within the current Params structure. Thus, it returns the value of the MaxIters parameter.

parameter	Memory	In/out	Meaning
params	host	in	The cusolverDnIRSParams_t Params structure
maxiters	host	output	The maximal number of iterations that is currently set

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.

2.4.1.30. cusolverDnIRSInfosCreate()

```
cusolverStatus_t
cusolverDnIRSInfosCreate(
    cusolverDnIRSParams_t params,
    cusolverDnIRSInfos_t* infos )
```

This function creates and initializes the Infos structure that will hold the refinement informations of an IRS solver. Such information includes the total number of iterations needed to converge (Niters), the outer number of iterations, and a pointer to the matrix of the convergence history residual norms. This function need to be called after the Params structure (see **cusolverDnIRSParamsCreate()**) has been created and before the call to an IRS solver such as **cusolverDnIRSXgesv()**. This function is valid for only one call to an IRS solver, since it hold info about that solve and thus each solve will requires its own Infos structure.

parameter	Memory	In/out	Meaning
params	host	in	The cusolverDnIRSParams_t Params structure

info	host	output	Pointer to the cusolverDnIRSInfos_t Infos structure
-------------	-------------	---------------	---

Status Returned

CUSOLVER_STATUS_SUCCESS	The structure was initialized successfully.
CUSOLVER_STATUS_ALLOC_FAILED	The resources could not be allocated.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.

2.4.1.31. cusolverDnIRSInfosDestroy()

```
cusolverStatus_t
cusolverDnIRSInfosDestroy(
    cusolverDnIRSPParams_t params,
    cusolverDnIRSInfos_t infos );
```

This function destroys and releases any memory required by the Infos structure. This function destroy all the informations (e.g., Nitters performed, OuterNitters performed, residual history etc) about a solver call, thus a user is supposed to call it once he is done from the informations he need.

parameter	Memory	In/out	Meaning
params	host	in/out	The cusolverDnIRSPParams_t Params structure
info	host	in/out	The cusolverDnIRSInfos_t Infos structure

Status Returned

CUSOLVER_STATUS_SUCCESS	the resources are released successfully.
CUSOLVER_STATUS_IRS_INFOS_NOT_INITIALIZED	The Infos structure was not created.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.

2.4.1.32. cusolverDnIRSInfosGetMaxIters()

```
cusolverStatus_t
cusolverDnIRSInfosGetMaxIters(
    cusolverDnIRSPParams_t params,
    cusolverDnIRSInfos_t infos,
    cusolver_int_t *maxiters );
```

This function returns the maximal number of iterations that is currently set within the current Params structure. This function is a duplicate of **cusolverDnIRSPParamsGetMaxIters()**.

parameter	Memory	In/out	Meaning
params	host	in	The cusolverDnIRSPParams_t Params structure
infos	host	in	The cusolverDnIRSInfos_t Infos structure

maxiters	host	output	The maximal number of iterations that is currently set
-----------------	-------------	---------------	--

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.
CUSOLVER_STATUS_IRS_INFOS_NOT_INITIALIZED	The Infos structure was not created.

2.4.1.33. cusolverDnIRSInfosGetNiters()

```
cusolverStatus_t cusolverDnIRSInfosGetNiters(
    cusolverDnIRSPParams_t params,
    cusolverDnIRSInfos_t infos,
    cusolver_int_t *niters );
```

This function returns the total number of iterations performed by the IRS solver. If it was negative it means that the IRS solver had numerical issues a fall back to a full precision solution most like happened. Please refer to the description of negative niters values in the corresponding IRS linear solver functions such as **cusolverDnXgesv()**.

parameter	Memory	In/out	Meaning
params	host	in	The cusolverDnIRSPParams_t Params structure
infos	host	in	The cusolverDnIRSInfos_t Infos structure
niters	host	output	The total number of iterations performed by the IRS solver

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.
CUSOLVER_STATUS_IRS_INFOS_NOT_INITIALIZED	The Infos structure was not created.

2.4.1.34. cusolverDnIRSInfosGetOuterNiters()

```
cusolverStatus_t
cusolverDnIRSInfosGetOuterNiters(
    cusolverDnIRSPParams_t params,
    cusolverDnIRSInfos_t infos,
    cusolver_int_t *outer_niters );
```

This function returns the number of iterations performed by outer refinement loop of the IRS solver. When RefinementSolver consists of one level solver such as CUSOLVER_IRS_REFINE_CLASSICAL or CUSOLVER_IRS_REFINE_GMRES, it is the same as Niters. When RefinementSolver consists of two levels solver such as CUSOLVER_IRS_REFINE_CLASSICAL_GMRES or CUSOLVER_IRS_REFINE_GMRES_GMRES, it is the number of the outer iteration. See description of cusolverIRSRefinementSolver_t type section for more details.

parameter	Memory	In/out	Meaning
params	host	in	The cusolverDnIRSParams_t Params structure
infos	host	in	The cusolverDnIRSInfos_t Infos structure
outer_niters	host	output	The number of iterations of the outer refinement loop of the IRS solver

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.
CUSOLVER_STATUS_IRS_INFOS_NOT_INITIALIZED	The Infos structure was not created.

2.4.1.35. cusolverDnIRSInfosRequestResidual()

```
cusolverStatus_t cusolverDnIRSInfosRequestResidual(
    cusolverDnIRSParams_t params,
    cusolverDnIRSInfos_t infos );
```

This function, once called, tell the IRS solver to store the convergence history of the refinement phase in a matrix, that could be accessed via a pointer returned by the **`cusolverDnIRSInfosGetResidualHistory()`** function.

parameter	Memory	In/out	Meaning
params	host	in	The cusolverDnIRSParams_t Params structure
infos	host	in	The cusolverDnIRSInfos_t Infos structure

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE	The Params structure was not created.
CUSOLVER_STATUS_IRS_INFOS_NOT_INITIALIZED	The Infos structure was not created.

2.4.1.36. cusolverDnIRSInfosGetResidualHistory()

```
cusolverStatus_t
cusolverDnIRSInfosGetResidualHistory(
    cusolverDnIRSParams_t params,
    cusolverDnIRSInfos_t infos,
    void **residual_history );
```

If the user called **`cusolverDnIRSInfosRequestResidual()`** before the call to the IRS solve function, this function return a pointer to the matrix of the convergence history residual norms. Precision of residual norms depends on the IRS input data type. If inout datatype has double precision (CUDA_R_FP64 or CUDA_C_FP64 inout datatype), this residual will be real double precision. Otherwise (CUDA_R_FP32 or CUDA_C_FP32 inout datatype) - residual will be with real single precision.

The residual history matrix consists of two columns (even for NRHS case) of `MaxIters+1` rows, thus a matrix of size `(MaxIters+1,2)`. Only the first **OuterNiters+1** rows contains the needed informations the other (e.g., `OuterNiters+2:Maxiters+1`) are garbage. On the first column, each row "*i*" specify the total number of iterations happened at this outer iteration "*i*" and on the second columns the residual norm corresponding to this outer iteration "*i*". Thus, the first row (e.g., outer iteration "*0*") consists of the initial residual (e.g., the residual before the refinement loop start) then the consecutive `OuterNiters` rows are the residual obtained at each outer iteration of the refinement loop. Note, it only consists of the history of the outer loop.

Thus, if the Refinementsolver was `CUSOLVER_IRS_REFINE_CLASSICAL` or `CUSOLVER_IRS_REFINE_GMRES`, then `OuterNiters=Niters` (`Niters` is the total number of iterations performed) and there is `Niters+1` rows of norms that correspond to the `Niters` outer iterations.

If the Refinementsolver was `CUSOLVER_IRS_REFINE_CLASSICAL_GMRES` or `CUSOLVER_IRS_REFINE_GMRES_GMRES`, then `OuterNiters <= Niters` corresponds to the outer iterations performed by the outer refinement loop. Thus there is `OuterNiters+1` residual norms where row "*i*" correspond to the outer iteration "*i*" and the first column specify the total number of iterations (outer and inner) that were performed while the second columns correspond to the residual norm at this stage.

For example, let say the user specify `CUSOLVER_IRS_REFINE_CLASSICAL_GMRES` as a Refinementsolver and let say it needed 3 outer iterations to converge and 4,3,3 inner iterations at each outer respectively. This consists of 10 total iterations. Thus on row 0 is for the first residual before the refinement start, so it has 0 iteration. On row 1 which correspond to the outer iteration 1, it will be shown 4 (4 is the total number of iterations that were performed till now) on row 2, it will be 7 and on row 3 it will be 10.

As summary, let define `ldh=Maxiters+1`, the leading dimension of the residual matrix. then `residual_history[i]` shows the total number of iterations performed at the outer iteration "*i*" and `residual_history[i+ldh]` correspond to its norm of the residual at this stage.

parameter	Memory	In/out	Meaning
<code>params</code>	host	in	The <code>cusolverDnIRSPParams_t</code> Params structure
<code>infos</code>	host	in	The <code>cusolverDnIRSInfos_t</code> Infos structure
<code>residual_history</code>	host	output	Returns a void pointer to the matrix of the convergence history residual norms. See the description above for the relation between the residual norm types and the inout datatype.

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	The operation completed successfully.
<code>CUSOLVER_STATUS_IRS_PARAMS_NOT_INITIALIZE</code>	The Params structure was not created.
<code>CUSOLVER_STATUS_IRS_INFOS_NOT_INITIALIZED</code>	The Infos structure was not created.

CUSOLVER_STATUS_INVALID_VALUE	This function was called without calling <code>cusolverDnIRSInfosRequestResidual()</code> in advance.
-------------------------------	---

2.4.2. Dense Linear Solver Reference

This chapter describes linear solver API of cuSolverDN, including Cholesky factorization, LU with partial pivoting, QR factorization and Bunch-Kaufman (LDLT) factorization.

2.4.2.1. `cusolverDn<t>potrf()`

These helper functions calculate the necessary size of work buffers.

```
cusolverStatus_t
cusolverDnSpotrf_bufferSize(cusolverDnHandle_t handle,
                           cublasFillMode_t uplo,
                           int n,
                           float *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnDpotrf_bufferSize(cusolverDnHandle_t handle,
                           cublasFillMode_t uplo,
                           int n,
                           double *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnCpotrf_bufferSize(cusolverDnHandle_t handle,
                           cublasFillMode_t uplo,
                           int n,
                           cuComplex *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnZpotrf_bufferSize(cusolverDnHandle_t handle,
                           cublasFillMode_t uplo,
                           int n,
                           cuDoubleComplex *A,
                           int lda,
                           int *Lwork);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSpotrf(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 float *A,
                 int lda,
                 float *Workspace,
                 int Lwork,
                 int *devInfo );

cusolverStatus_t
cusolverDnDpotrf(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 double *A,
                 int lda,
                 double *Workspace,
                 int Lwork,
                 int *devInfo );
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCpotrf(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 cuComplex *A,
                 int lda,
                 cuComplex *Workspace,
                 int Lwork,
                 int *devInfo );

cusolverStatus_t
cusolverDnZpotrf(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 cuDoubleComplex *A,
                 int lda,
                 cuDoubleComplex *Workspace,
                 int Lwork,
                 int *devInfo );
```

This function computes the Cholesky factorization of a Hermitian positive-definite matrix.

A is a **n×n** Hermitian matrix, only lower or upper part is meaningful. The input parameter **uplo** indicates which part of the matrix is used. The function would leave other part untouched.

If input parameter **uplo** is **CUBLAS_FILL_MODE_LOWER**, only lower triangular part of **A** is processed, and replaced by lower triangular Cholesky factor **L**.

$$A = L * L^H$$

If input parameter **uplo** is **CUBLAS_FILL_MODE_UPPER**, only upper triangular part of **A** is processed, and replaced by upper triangular Cholesky factor **U**.

$$A = U^H * U$$

The user has to provide working space which is pointed by input parameter **Workspace**. The input parameter **Lwork** is size of the working space, and it is returned by **potrf_bufferSize()**.

If Cholesky factorization failed, i.e. some leading minor of **A** is not positive definite, or equivalently some diagonal elements of **L** or **U** is not a real number. The output parameter **devInfo** would indicate smallest leading minor of **A** which is not positive definite.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

API of potrf

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
uplo	host	input	indicates if matrix A lower or upper part is stored, the other part is not referenced.
n	host	input	number of rows and columns of matrix A .
A	device	in/out	<type> array of dimension lda * n with lda is not less than max(1, n) .
lda	host	input	leading dimension of two-dimensional array used to store matrix A .
Workspace	device	in/out	working space, <type> array of size Lwork .
Lwork	host	input	size of Workspace , returned by potrf_bufferSize .
devInfo	device	output	if devInfo = 0, the Cholesky factorization is successful. if devInfo = -i , the i-th parameter is wrong (not counting handle). if devInfo = i , the leading minor of order i is not positive definite.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (n <0 or lda < max(1, n)).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.2.2. cusolverDn<t>potrs()

```

cusolverStatus_t
cusolverDnSpotrs(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 int nrhs,
                 const float *A,
                 int lda,
                 float *B,
                 int ldb,
                 int *devInfo);

cusolverStatus_t
cusolverDnDpotrs(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 int nrhs,
                 const double *A,
                 int lda,
                 double *B,
                 int ldb,
                 int *devInfo);

cusolverStatus_t
cusolverDnCpotrs(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 int nrhs,
                 const cuComplex *A,
                 int lda,
                 cuComplex *B,
                 int ldb,
                 int *devInfo);

cusolverStatus_t
cusolverDnZpotrs(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 int nrhs,
                 const cuDoubleComplex *A,
                 int lda,
                 cuDoubleComplex *B,
                 int ldb,
                 int *devInfo);

```

This function solves a system of linear equations

$$A * X = B$$

where **A** is a **n×n** Hermitian matrix, only lower or upper part is meaningful. The input parameter **uplo** indicates which part of the matrix is used. The function would leave other part untouched.

The user has to call **potrf** first to factorize matrix **A**. If input parameter **uplo** is **CUBLAS_FILL_MODE_LOWER**, **A** is lower triangular Cholesky factor **L** corresponding to

$A = L * L^H$. If input parameter **uplo** is **CUBLAS_FILL_MODE_UPPER**, **A** is upper triangular Cholesky factor **U** corresponding to $A = U^H * U$.

The operation is in-place, i.e. matrix **X** overwrites matrix **B** with the same leading dimension **ldb**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

API of potrs

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolveDN library context.
uplo	host	input	indicates if matrix A lower or upper part is stored, the other part is not referenced.
n	host	input	number of rows and columns of matrix A .
nrhs	host	input	number of columns of matrix X and B .
A	device	input	<type> array of dimension lda * n with lda is not less than max(1,n) . A is either lower cholesky factor L or upper Cholesky factor U .
lda	host	input	leading dimension of two-dimensional array used to store matrix A .
B	device	in/out	<type> array of dimension ldb * nrhs . ldb is not less than max(1,n) . As an input, B is right hand side matrix. As an output, B is the solution matrix.
devInfo	device	output	if devInfo = 0, the Cholesky factorization is successful. if devInfo = -i , the i-th parameter is wrong (not counting handle).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (n <0, nrhs <0, lda < max(1,n) or ldb < max(1,n)).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.2.3. cusolverDn<t>potri()

These helper functions calculate the necessary size of work buffers.

```
cusolverStatus_t
cusolverDnSpotri_bufferSize(cusolverDnHandle_t handle,
                           cublasFillMode_t uplo,
                           int n,
                           float *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnDpotri_bufferSize(cusolverDnHandle_t handle,
                           cublasFillMode_t uplo,
                           int n,
                           double *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnCpotri_bufferSize(cusolverDnHandle_t handle,
                           cublasFillMode_t uplo,
                           int n,
                           cuComplex *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnZpotri_bufferSize(cusolverDnHandle_t handle,
                           cublasFillMode_t uplo,
                           int n,
                           cuDoubleComplex *A,
                           int lda,
                           int *Lwork);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSpotri(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 float *A,
                 int lda,
                 float *Workspace,
                 int Lwork,
                 int *devInfo );

cusolverStatus_t
cusolverDnDpotri(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 double *A,
                 int lda,
                 double *Workspace,
                 int Lwork,
                 int *devInfo );
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCpotri(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 cuComplex *A,
                 int lda,
                 cuComplex *Workspace,
                 int Lwork,
                 int *devInfo );

cusolverStatus_t
cusolverDnZpotri(cusolverDnHandle_t handle,
                 cublasFillMode_t uplo,
                 int n,
                 cuDoubleComplex *A,
                 int lda,
                 cuDoubleComplex *Workspace,
                 int Lwork,
                 int *devInfo );
```

This function computes the inverse of a positive-definite matrix **A** using the Cholesky factorization

$$A = L * L^H = U^H * U$$

computed by **potrf()**.

A is a **n×n** matrix containing the triangular factor **L** or **U** computed by the Cholesky factorization. Only lower or upper part is meaningful and the input parameter **uplo** indicates which part of the matrix is used. The function would leave the other part untouched.

If the input parameter **uplo** is **CUBLAS_FILL_MODE_LOWER**, only lower triangular part of **A** is processed, and replaced by the lower triangular part of the inverse of **A**.

If the input parameter **uplo** is **CUBLAS_FILL_MODE_UPPER**, only upper triangular part of **A** is processed, and replaced by the upper triangular part of the inverse of **A**.

The user has to provide the working space which is pointed to by input parameter **Workspace**. The input parameter **Lwork** is the size of the working space, returned by **potri_bufferSize()**.

If the computation of the inverse fails, i.e. some leading minor of **L** or **U**, is null, the output parameter **devInfo** would indicate the smallest leading minor of **L** or **U** which is not positive definite.

If the output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting the handle).

API of potri

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.

uplo	host	input	indicates if matrix A lower or upper part is stored, the other part is not referenced.
n	host	input	number of rows and columns of matrix A .
A	device	in/out	<type> array of dimension lda * n where lda is not less than max(1, n) .
lda	host	input	leading dimension of two-dimensional array used to store matrix A .
Workspace	device	in/out	working space, <type> array of size Lwork .
Lwork	host	input	size of Workspace , returned by potri_bufferSize .
devInfo	device	output	if devInfo = 0, the computation of the inverse is successful. if devInfo = -i, the i-th parameter is wrong (not counting handle). if devInfo = i, the leading minor of order i is zero.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (n <0 or lda < max(1, n)).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.2.4. cusolverDn<t>getrf()

These helper functions calculate the size of work buffers needed.

```
cusolverStatus_t
cusolverDnSgetrf_bufferSize(cusolverDnHandle_t handle,
                           int m,
                           int n,
                           float *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnDgetrf_bufferSize(cusolverDnHandle_t handle,
                           int m,
                           int n,
                           double *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnCgetrf_bufferSize(cusolverDnHandle_t handle,
                           int m,
                           int n,
                           cuComplex *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnZgetrf_bufferSize(cusolverDnHandle_t handle,
                           int m,
                           int n,
                           cuDoubleComplex *A,
                           int lda,
                           int *Lwork );
```

The S and D data types are real single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSgetrf(cusolverDnHandle_t handle,
                int m,
                int n,
                float *A,
                int lda,
                float *Workspace,
                int *devIpivot,
                int *devInfo );

cusolverStatus_t
cusolverDnDgetrf(cusolverDnHandle_t handle,
                int m,
                int n,
                double *A,
                int lda,
                double *Workspace,
                int *devIpivot,
                int *devInfo );
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCgetrf(cusolverDnHandle_t handle,
                 int m,
                 int n,
                 cuComplex *A,
                 int lda,
                 cuComplex *Workspace,
                 int *devI piv,
                 int *devInfo );

cusolverStatus_t
cusolverDnZgetrf(cusolverDnHandle_t handle,
                 int m,
                 int n,
                 cuDoubleComplex *A,
                 int lda,
                 cuDoubleComplex *Workspace,
                 int *devI piv,
                 int *devInfo );
```

This function computes the LU factorization of a $m \times n$ matrix

$$P * A = L * U$$

where **A** is a $m \times n$ matrix, **P** is a permutation matrix, **L** is a lower triangular matrix with unit diagonal, and **U** is an upper triangular matrix.

The user has to provide working space which is pointed by input parameter **Workspace**. The input parameter **Lwork** is size of the working space, and it is returned by **getrf_bufferSize()**.

If LU factorization failed, i.e. matrix **A** (**U**) is singular, The output parameter **devInfo=i** indicates **U(i,i) = 0**.

If output parameter **devInfo = -i** (less than zero), the **i-th** parameter is wrong (not counting handle).

If **devI piv** is null, no pivoting is performed. The factorization is **A=L*U**, which is not numerically stable.

No matter LU factorization failed or not, the output parameter **devI piv** contains pivoting sequence, row **i** is interchanged with row **devI piv(i)**.

The user can combine **getrf** and **getrs** to complete a linear solver. Please refer to appendix D.1.

API of getrf

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
m	host	input	number of rows of matrix A .
n	host	input	number of columns of matrix A .
A	device	in/out	<type> array of dimension lda * n with lda is not less than max(1,m) .

lda	host	input	leading dimension of two-dimensional array used to store matrix A .
Workspace	device	in/out	working space, <type> array of size Lwork .
devI piv	device	output	array of size at least $\min(m,n)$, containing pivot indices.
devInfo	device	output	if devInfo = 0, the LU factorization is successful. if devInfo = -i, the i-th parameter is wrong (not counting handle). if devInfo = i, the $U(i,i) = 0$.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($m,n < 0$ or $lda < \max(1,m)$).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.2.5. cusolverDn<t>getrs()

```

cusolverStatus_t
cusolverDnSgetrs(cusolverDnHandle_t handle,
                 cublasOperation_t trans,
                 int n,
                 int nrhs,
                 const float *A,
                 int lda,
                 const int *devIpiv,
                 float *B,
                 int ldb,
                 int *devInfo );

cusolverStatus_t
cusolverDnDgetrs(cusolverDnHandle_t handle,
                 cublasOperation_t trans,
                 int n,
                 int nrhs,
                 const double *A,
                 int lda,
                 const int *devIpiv,
                 double *B,
                 int ldb,
                 int *devInfo );

cusolverStatus_t
cusolverDnCgetrs(cusolverDnHandle_t handle,
                 cublasOperation_t trans,
                 int n,
                 int nrhs,
                 const cuComplex *A,
                 int lda,
                 const int *devIpiv,
                 cuComplex *B,
                 int ldb,
                 int *devInfo );

cusolverStatus_t
cusolverDnZgetrs(cusolverDnHandle_t handle,
                 cublasOperation_t trans,
                 int n,
                 int nrhs,
                 const cuDoubleComplex *A,
                 int lda,
                 const int *devIpiv,
                 cuDoubleComplex *B,
                 int ldb,
                 int *devInfo );

```

This function solves a linear system of multiple right-hand sides

$$\text{op}(A) * X = B$$

where **A** is a **n×n** matrix, and was LU-factored by **getrf**, that is, lower triangular part of **A** is **L**, and upper triangular part (including diagonal elements) of **A** is **U**. **B** is a **n×nrhs** right-hand side matrix.

The input parameter **trans** is defined by

$$\text{op}(\mathbf{A}) = \begin{cases} \mathbf{A} & \text{if trans} == \text{CUBLAS_OP_N} \\ \mathbf{A}^T & \text{if trans} == \text{CUBLAS_OP_T} \\ \mathbf{A}^H & \text{if trans} == \text{CUBLAS_OP_C} \end{cases}$$

The input parameter **devI piv** is an output of **getrf**. It contains pivot indices, which are used to permute right-hand sides.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

The user can combine **getrf** and **getrs** to complete a linear solver. Please refer to appendix D.1.

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
trans	host	input	operation $\text{op}(\mathbf{A})$ that is non- or (conj.) transpose.
n	host	input	number of rows and columns of matrix A .
nrhs	host	input	number of right-hand sides.
A	device	input	<type> array of dimension $\text{lda} * n$ with lda is not less than $\max(1, n)$.
lda	host	input	leading dimension of two-dimensional array used to store matrix A .
devI piv	device	input	array of size at least n , containing pivot indices.
B	device	output	<type> array of dimension $\text{ldb} * \text{nrhs}$ with ldb is not less than $\max(1, n)$.
ldb	host	input	leading dimension of two-dimensional array used to store matrix B .
devInfo	device	output	if devInfo = 0, the operation is successful. if devInfo = -i , the i-th parameter is wrong (not counting handle).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($n < 0$ or $\text{lda} < \max(1, n)$ or $\text{ldb} < \max(1, n)$).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.2.6. cusolverDn<t>geqrf()

These helper functions calculate the size of work buffers needed.

```
cusolverStatus_t
cusolverDnSgeqrf_bufferSize(cusolverDnHandle_t handle,
                           int m,
                           int n,
                           float *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnDgeqrf_bufferSize(cusolverDnHandle_t handle,
                           int m,
                           int n,
                           double *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnCgeqrf_bufferSize(cusolverDnHandle_t handle,
                           int m,
                           int n,
                           cuComplex *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnZgeqrf_bufferSize(cusolverDnHandle_t handle,
                           int m,
                           int n,
                           cuDoubleComplex *A,
                           int lda,
                           int *Lwork );
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSgeqrf(cusolverDnHandle_t handle,
                int m,
                int n,
                float *A,
                int lda,
                float *TAU,
                float *Workspace,
                int Lwork,
                int *devInfo );

cusolverStatus_t
cusolverDnDgeqrf(cusolverDnHandle_t handle,
                int m,
                int n,
                double *A,
                int lda,
                double *TAU,
                double *Workspace,
                int Lwork,
                int *devInfo );
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCgeqrf(cusolverDnHandle_t handle,
                 int m,
                 int n,
                 cuComplex *A,
                 int lda,
                 cuComplex *TAU,
                 cuComplex *Workspace,
                 int Lwork,
                 int *devInfo );

cusolverStatus_t
cusolverDnZgeqrf(cusolverDnHandle_t handle,
                 int m,
                 int n,
                 cuDoubleComplex *A,
                 int lda,
                 cuDoubleComplex *TAU,
                 cuDoubleComplex *Workspace,
                 int Lwork,
                 int *devInfo );
```

This function computes the QR factorization of a $m \times n$ matrix

$$A = Q * R$$

where **A** is a $m \times n$ matrix, **Q** is a $m \times n$ matrix, and **R** is a $n \times n$ upper triangular matrix.

The user has to provide working space which is pointed by input parameter **Workspace**. The input parameter **Lwork** is size of the working space, and it is returned by **geqrf_bufferSize()**.

The matrix **R** is overwritten in upper triangular part of **A**, including diagonal elements.

The matrix **Q** is not formed explicitly, instead, a sequence of householder vectors are stored in lower triangular part of **A**. The leading nonzero element of householder vector is assumed to be 1 such that output parameter **TAU** contains the scaling factor τ . If **v** is original householder vector, **q** is the new householder vector corresponding to τ , satisfying the following relation

$$I - 2 * v * v^H = I - \tau * q * q^H$$

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

API of geqrf

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
m	host	input	number of rows of matrix A .
n	host	input	number of columns of matrix A .
A	device	in/out	<type> array of dimension lda * n with lda is not less than max(1, m) .

lda	host	input	leading dimension of two-dimensional array used to store matrix A .
TAU	device	output	<type> array of dimension at least $\min(m,n)$.
Workspace	device	in/out	working space, <type> array of size Lwork .
Lwork	host	input	size of working array Workspace .
devInfo	device	output	if devInfo = 0, the LU factorization is successful. if devInfo = -i, the i-th parameter is wrong (not counting handle).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($m,n < 0$ or $lda < \max(1,m)$).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.2.7. cusolverDn<t>ormqr()

These helper functions calculate the size of work buffers needed.

```
cusolverStatus_t
cusolverDnSormqr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasOperation_t trans,
    int m,
    int n,
    int k,
    const float *A,
    int lda,
    const float *tau,
    const float *C,
    int ldc,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnDormqr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasOperation_t trans,
    int m,
    int n,
    int k,
    const double *A,
    int lda,
    const double *tau,
    const double *C,
    int ldc,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnCunmqr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasOperation_t trans,
    int m,
    int n,
    int k,
    const cuComplex *A,
    int lda,
    const cuComplex *tau,
    const cuComplex *C,
    int ldc,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnZunmqr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasOperation_t trans,
    int m,
    int n,
    int k,
    const cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *tau,
    const cuDoubleComplex *C,
    int ldc,
    int *lwork);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSormqr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasOperation_t trans,
    int m,
    int n,
    int k,
    const float *A,
    int lda,
    const float *tau,
    float *C,
    int ldc,
    float *work,
    int lwork,
    int *devInfo);
```

```
cusolverStatus_t
cusolverDnDormqr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasOperation_t trans,
    int m,
    int n,
    int k,
    const double *A,
    int lda,
    const double *tau,
    double *C,
    int ldc,
    double *work,
    int lwork,
    int *devInfo);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCumqr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasOperation_t trans,
    int m,
    int n,
    int k,
    const cuComplex *A,
    int lda,
    const cuComplex *tau,
    cuComplex *C,
    int ldc,
    cuComplex *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnZunmqr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasOperation_t trans,
    int m,
    int n,
    int k,
    const cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *tau,
    cuDoubleComplex *C,
    int ldc,
    cuDoubleComplex *work,
    int lwork,
    int *devInfo);
```

This function overwrites $m \times n$ matrix **C** by

$$C = \begin{cases} \text{op}(Q) * C & \text{if side} == \text{CUBLAS_SIDE_LEFT} \\ C * \text{op}(Q) & \text{if side} == \text{CUBLAS_SIDE_RIGHT} \end{cases}$$

The operation of **Q** is defined by

$$\text{op}(Q) = \begin{cases} Q & \text{if transa} == \text{CUBLAS_OP_N} \\ Q^T & \text{if transa} == \text{CUBLAS_OP_T} \\ Q^H & \text{if transa} == \text{CUBLAS_OP_C} \end{cases}$$

Q is a unitary matrix formed by a sequence of elementary reflection vectors from QR factorization (**geqrf**) of **A**.

$$Q = H(1) H(2) \dots H(k)$$

Q is of order **m** if **side** = **CUBLAS_SIDE_LEFT** and of order **n** if **side** = **CUBLAS_SIDE_RIGHT**.

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is size of the working space, and it is returned by **geqrf_bufferSize()** or **ormqr_bufferSize()**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

The user can combine **geqrf**, **ormqr** and **trsm** to complete a linear solver or a least-square solver. Please refer to appendix C.1.

API of **ormqr**

parameter	Memory	In/out	Meaning
handle	host	input	Handle to the cuSolverDn library context.
side	host	input	Indicates if matrix Q is on the left or right of C .
trans	host	input	Operation op (Q) that is non- or (conj.) transpose.
m	host	input	Number of columns of matrix C .
n	host	input	Number of rows of matrix C .
k	host	input	Number of elementary reflections whose product defines the matrix Q .
A	device	in/out	<type> array of dimension lda * k with lda is not less than max(1,m) . The matrix A is from geqrf , so i-th column contains elementary reflection vector.
lda	host	input	Leading dimension of two-dimensional array used to store matrix A . if side is CUBLAS_SIDE_LEFT , lda \geq max(1,m) ; if side is CUBLAS_SIDE_RIGHT , lda \geq max(1,n) .
tau	device	output	<type> array of dimension at least min(m,n) . The vector tau is from geqrf , so tau(i) is the scalar of i-th elementary reflection vector.
C	device	in/out	<type> array of size ldc * n . On exit, C is overwritten by op(Q) * C .
ldc	host	input	Leading dimension of two-dimensional array of matrix C . ldc \geq max(1,m) .
work	device	in/out	Working space, <type> array of size lwork .
lwork	host	input	Size of working array work .
devInfo	device	output	If devInfo = 0, the ormqr is successful. If devInfo = -i , the i-th parameter is wrong (not counting handle).

Status Returned

CUSOLVER_STATUS_SUCCESS	The operation completed successfully.
--------------------------------	---------------------------------------

<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	The library was not initialized.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	Invalid parameters were passed (<code>m, n < 0</code> or wrong <code>lda</code> or <code>ldc</code>).
<code>CUSOLVER_STATUS_ARCH_MISMATCH</code>	The device only supports compute capability 2.0 and above.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	An internal operation failed.

2.4.2.8. `cusolverDn<t>orgqr()`

These helper functions calculate the size of work buffers needed.

```
cusolverStatus_t
cusolverDnSorgqr_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int k,
    const float *A,
    int lda,
    const float *tau,
    int *lwork);

cusolverStatus_t
cusolverDnDorgqr_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int k,
    const double *A,
    int lda,
    const double *tau,
    int *lwork);

cusolverStatus_t
cusolverDnCungqr_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int k,
    const cuComplex *A,
    int lda,
    const cuComplex *tau,
    int *lwork);

cusolverStatus_t
cusolverDnZungqr_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int k,
    const cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *tau,
    int *lwork);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSorgqr(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int k,
    float *A,
    int lda,
    const float *tau,
    float *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnDorgqr(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int k,
    double *A,
    int lda,
    const double *tau,
    double *work,
    int lwork,
    int *devInfo);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCungqr(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int k,
    cuComplex *A,
    int lda,
    const cuComplex *tau,
    cuComplex *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnZungqr(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int k,
    cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *tau,
    cuDoubleComplex *work,
    int lwork,
    int *devInfo);
```

This function overwrites $m \times n$ matrix **A** by

$$Q = H(1) * H(2) * \dots * H(k)$$

where **Q** is a unitary matrix formed by a sequence of elementary reflection vectors stored in **A**.

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is size of the working space, and it is returned by **orgqr_bufferSize()**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

The user can combine **geqrf**, **orgqr** to complete orthogonalization. Please refer to appendix C.2.

API of ormqr

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
m	host	input	number of rows of matrix Q . $m \geq 0$;
n	host	input	number of columns of matrix Q . $m \geq n \geq 0$;
k	host	input	number of elementary reflections whose product defines the matrix Q . $n \geq k \geq 0$;
A	device	in/out	<type> array of dimension $lda * n$ with lda is not less than $\max(1, m)$. i -th column of A contains elementary reflection vector.
lda	host	input	leading dimension of two-dimensional array used to store matrix A . $lda \geq \max(1, m)$.
tau	device	output	<type> array of dimension k . tau(i) is the scalar of i -th elementary reflection vector.
work	device	in/out	working space, <type> array of size lwork .
lwork	host	input	size of working array work .
devInfo	device	output	if info = 0, the orgqr is successful. if info = -i , the i -th parameter is wrong (not counting handle).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($m, n, k < 0$, $n > m$, $k > n$ or $lda < m$).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.

CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
--------------------------------	-------------------------------

2.4.2.9. cusolverDn<t>sytrf()

These helper functions calculate the size of the needed buffers.

```
cusolverStatus_t
cusolverDnSsytrf_bufferSize(cusolverDnHandle_t handle,
                           int n,
                           float *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnDsytrf_bufferSize(cusolverDnHandle_t handle,
                           int n,
                           double *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnCsytrf_bufferSize(cusolverDnHandle_t handle,
                           int n,
                           cuComplex *A,
                           int lda,
                           int *Lwork );

cusolverStatus_t
cusolverDnZsytrf_bufferSize(cusolverDnHandle_t handle,
                           int n,
                           cuDoubleComplex *A,
                           int lda,
                           int *Lwork );
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSsytrf(cusolverDnHandle_t handle,
                cublasFillMode_t uplo,
                int n,
                float *A,
                int lda,
                int *ipiv,
                float *work,
                int lwork,
                int *devInfo );

cusolverStatus_t
cusolverDnDsytrf(cusolverDnHandle_t handle,
                cublasFillMode_t uplo,
                int n,
                double *A,
                int lda,
                int *ipiv,
                double *work,
                int lwork,
                int *devInfo );
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCsyrtrf(cusolverDnHandle_t handle,
                  cublasFillMode_t uplo,
                  int n,
                  cuComplex *A,
                  int lda,
                  int *ipiv,
                  cuComplex *work,
                  int lwork,
                  int *devInfo );

cusolverStatus_t
cusolverDnZsyrtrf(cusolverDnHandle_t handle,
                  cublasFillMode_t uplo,
                  int n,
                  cuDoubleComplex *A,
                  int lda,
                  int *ipiv,
                  cuDoubleComplex *work,
                  int lwork,
                  int *devInfo );
```

This function computes the Bunch-Kaufman factorization of a $n \times n$ symmetric indefinite matrix

A is a $n \times n$ symmetric matrix, only lower or upper part is meaningful. The input parameter **uplo** which part of the matrix is used. The function would leave other part untouched.

If input parameter **uplo** is **CUBLAS_FILL_MODE_LOWER**, only lower triangular part of **A** is processed, and replaced by lower triangular factor **L** and block diagonal matrix **D**. Each block of **D** is either 1x1 or 2x2 block, depending on pivoting.

$$P^* A^* P^T = L^* D^* L^T$$

If input parameter **uplo** is **CUBLAS_FILL_MODE_UPPER**, only upper triangular part of **A** is processed, and replaced by upper triangular factor **U** and block diagonal matrix **D**.

$$P^* A^* P^T = U^* D^* U^T$$

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is size of the working space, and it is returned by **syrtrf_bufferSize()**.

If Bunch-Kaufman factorization failed, i.e. **A** is singular. The output parameter **devInfo** = **i** would indicate **D(i,i)=0**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

The output parameter **devIpiv** contains pivoting sequence. If **devIpiv(i) = k > 0**, **D(i,i)** is 1x1 block, and **i-th** row/column of **A** is interchanged with **k-th** row/column of **A**. If **uplo** is **CUBLAS_FILL_MODE_UPPER** and **devIpiv(i-1) = devIpiv(i) = -m < 0**, **D(i-1:i,i-1:i)** is a 2x2 block, and **(i-1)-th** row/column is interchanged

with m -th row/column. If `uplo` is `CUBLAS_FILL_MODE_LOWER` and `devI piv(i+1) = devI piv(i) = -m < 0`, `D(i:i+1,i:i+1)` is a 2x2 block, and $(i+1)$ -th row/column is interchanged with m -th row/column.

API of sytrf

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
uplo	host	input	indicates if matrix A lower or upper part is stored, the other part is not referenced.
n	host	input	number of rows and columns of matrix A .
A	device	in/out	<type> array of dimension <code>lda * n</code> with <code>lda</code> is not less than <code>max(1,n)</code> .
lda	host	input	leading dimension of two-dimensional array used to store matrix A .
ipiv	device	output	array of size at least <code>n</code> , containing pivot indices.
work	device	in/out	working space, <type> array of size <code>lwork</code> .
lwork	host	input	size of working space <code>work</code> .
devInfo	device	output	if <code>devInfo = 0</code> , the LU factorization is successful. if <code>devInfo = -i</code> , the i -th parameter is wrong (not counting handle). if <code>devInfo = i</code> , the <code>D(i,i) = 0</code> .

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed (<code>n < 0</code> or <code>lda < max(1,n)</code>).
<code>CUSOLVER_STATUS_ARCH_MISMATCH</code>	the device only supports compute capability 2.0 and above.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

2.4.2.10. cusolverDn<t>potrfBatched()

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSpotrfBatched(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    float *Aarray[],
    int lda,
    int *infoArray,
    int batchSize);

cusolverStatus_t
cusolverDnDpotrfBatched(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    double *Aarray[],
    int lda,
    int *infoArray,
    int batchSize);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCpotrfBatched(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    cuComplex *Aarray[],
    int lda,
    int *infoArray,
    int batchSize);

cusolverStatus_t
cusolverDnZpotrfBatched(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    cuDoubleComplex *Aarray[],
    int lda,
    int *infoArray,
    int batchSize);
```

This function computes the Cholesky factorization of a sequence of Hermitian positive-definite matrices.

Each **Aarray[i]** for **i=0,1,..., batchSize-1** is a **n×n** Hermitian matrix, only lower or upper part is meaningful. The input parameter **uplo** indicates which part of the matrix is used.

If input parameter **uplo** is **CUBLAS_FILL_MODE_LOWER**, only lower triangular part of **A** is processed, and replaced by lower triangular Cholesky factor **L**.

$$A = L * L^H$$

If input parameter **uplo** is **CUBLAS_FILL_MODE_UPPER**, only upper triangular part of **A** is processed, and replaced by upper triangular Cholesky factor **U**.

$$A = U^H * U$$

If Cholesky factorization failed, i.e. some leading minor of **A** is not positive definite, or equivalently some diagonal elements of **L** or **U** is not a real number. The output parameter **infoArray** would indicate smallest leading minor of **A** which is not positive definite.

infoArray is an integer array of size **batchsize**. If **potrfBatched** returns **CUSOLVER_STATUS_INVALID_VALUE**, **infoArray[0] = -i** (less than zero), meaning that the **i-th** parameter is wrong (not counting handle). If **potrfBatched** returns **CUSOLVER_STATUS_SUCCESS** but **infoArray[i] = k** is positive, then **i-th** matrix is not positive definite and the Cholesky factorization failed at row **k**.

Remark: the other part of **A** is used as a workspace. For example, if **uplo** is **CUBLAS_FILL_MODE_UPPER**, upper triangle of **A** contains cholesky factor **U** and lower triangle of **A** is destroyed after **potrfBatched**.

API of potrfBatched

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
uplo	host	input	indicates if lower or upper part is stored, the other part is used as a workspace.
n	host	input	number of rows and columns of matrix A .
Aarray	device	in/out	array of pointers to <type> array of dimension lda * n with lda is not less than max(1, n) .
lda	host	input	leading dimension of two-dimensional array used to store each matrix Aarray[i] .
infoArray	device	output	array of size batchSize . infoArray[i] contains information of factorization of Aarray[i] . if potrfBatched returns CUSOLVER_STATUS_INVALID_VALUE , infoArray[0] = -i (less than zero) means the i-th parameter is wrong (not counting handle). if potrfBatched returns CUSOLVER_STATUS_SUCCESS , infoArray[i] = 0 means the Cholesky factorization of i-th matrix is successful, and infoArray[i] = k means the leading submatrix of order k of i-th matrix is not positive definite.
batchSize	host	input	number of pointers in Aarray .

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed (<code>n<0</code> or <code>lda<max(1,n)</code> or <code>batchSize<1</code>).
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

2.4.2.11. `cusolverDn<t>potrsBatched()`

```
cusolverStatus_t
cusolverDnSpotrsBatched(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    int nrhs,
    float *Aarray[],
    int lda,
    float *Barray[],
    int ldb,
    int *info,
    int batchSize);
```

```
cusolverStatus_t
cusolverDnDpotrsBatched(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    int nrhs,
    double *Aarray[],
    int lda,
    double *Barray[],
    int ldb,
    int *info,
    int batchSize);
```

```
cusolverStatus_t
cusolverDnCpotrsBatched(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    int nrhs,
    cuComplex *Aarray[],
    int lda,
    cuComplex *Barray[],
    int ldb,
    int *info,
    int batchSize);
```

```
cusolverStatus_t
cusolverDnZpotrsBatched(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    int nrhs,
    cuDoubleComplex *Aarray[],
    int lda,
    cuDoubleComplex *Barray[],
    int ldb,
    int *info,
    int batchSize);
```

This function solves a sequence of linear systems

$$A[i] * X[i] = B[i]$$

where each **Aarray[i]** for $i=0,1,\dots, \text{batchSize}-1$ is a $n \times n$ Hermitian matrix, only lower or upper part is meaningful. The input parameter **uplo** indicates which part of the matrix is used.

The user has to call **potrfBatched** first to factorize matrix **Aarray[i]**. If input parameter **uplo** is **CUBLAS_FILL_MODE_LOWER**, **A** is lower triangular Cholesky factor **L** corresponding to $A = L * L^H$. If input parameter **uplo** is **CUBLAS_FILL_MODE_UPPER**, **A** is upper triangular Cholesky factor **U** corresponding to $A = U^H * U$.

The operation is in-place, i.e. matrix **X** overwrites matrix **B** with the same leading dimension **ldb**.

The output parameter **info** is a scalar. If **info** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

Remark 1: only **nrhs=1** is supported.

Remark 2: **infoArray** from **potrfBatched** indicates if the matrix is positive definite. **info** from **potrsBatched** only shows which input parameter is wrong (not counting handle).

Remark 3: the other part of **A** is used as a workspace. For example, if **uplo** is **CUBLAS_FILL_MODE_UPPER**, upper triangle of **A** contains cholesky factor **U** and lower triangle of **A** is destroyed after **potrsBatched**.

API of potrsBatched

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolveDN library context.
uplo	host	input	indicates if matrix A lower or upper part is stored.
n	host	input	number of rows and columns of matrix A .
nrhs	host	input	number of columns of matrix x and B .
Aarray	device	in/out	array of pointers to <type> array of dimension $\text{lda} * n$ with lda is not less than $\max(1, n)$. Aarray[i] is either lower cholesky factor L or upper Cholesky factor U .
lda	host	input	leading dimension of two-dimensional array used to store each matrix Aarray[i] .
Barray	device	in/out	array of pointers to <type> array of dimension $\text{ldb} * \text{nrhs}$. ldb is not less than $\max(1, n)$. As an input, Barray[i] is right hand side matrix. As an output, Barray[i] is the solution matrix.

ldb	host	input	leading dimension of two-dimensional array used to store each matrix Barray[i] .
info	device	output	if info = 0, all parameters are correct. if info = -i, the i-th parameter is wrong (not counting handle).
batchSize	host	input	number of pointers in Aarray .

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (n <0, nrhs <0, lda <max(1, n), ldb <max(1, n) or batchSize <0).
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.2.12. cusolverDn<t1><t2>gesv()

These functions are modelled after functions DSGESV and ZCGESV from LAPACK and compute the solution to a system of linear equations with multiple right hand sides using mixed precision iterative refinement

$$A \times X = B$$

Where **A** is **n-by-n** matrix and **X** and **B** are **n-by-nrhs** matrices.

Functions are designed to be as close to LAPACK drop-in replacements as possible. Parameters and behaviour are mostly the same as LAPACK counterparts. Description of LAPACK functions and differences from them are below.

<t1><t2>gesv() functions are designated by two floating point precisions - data type (full) precision and internal lower precision. cusolver<t1><t2>gesv() first attempts to factorize the matrix in lower precision and use this factorization within an iterative refinement procedure to obtain a solution with same normwise backward error as full precision. If the approach fails to converge, then the method switches to a full precision factorization and solve.

Note that in addition to the data type / lower floating point precision functions available in LAPACK - also functions with half precision as a lower precision are present. The following table specifies which precisions will be used for which interface function:

Supported combinations of floating point precisions for cusolver <t1><t2>gesv() functions

Interface function	Data type (matrix, rhs and solution)	Data floating point precision	Internal (compute) floating point precision
cusolverDnDDgesv	double	double	double
cusolverDnDSgesv	double	double	single
cusolverDnDHgesv	double	double	half

<code>cusolverDnSSgesv</code>	<code>float</code>	<code>single</code>	<code>single</code>
<code>cusolverDnSHgesv</code>	<code>float</code>	<code>single</code>	<code>half</code>
<code>cusolverDnZZgesv</code>	<code>cuDoubleComplex</code>	<code>double</code>	<code>double</code>
<code>cusolverDnZCgesv</code>	<code>cuDoubleComplex</code>	<code>double</code>	<code>single</code>
<code>cusolverDnZKgesv</code>	<code>cuDoubleComplex</code>	<code>double</code>	<code>half</code>
<code>cusolverDnCCgesv</code>	<code>cuComplex</code>	<code>single</code>	<code>single</code>
<code>cusolverDnCKgesv</code>	<code>cuComplex</code>	<code>single</code>	<code>half</code>

The iterative refinement process is stopped if

$$\text{ITER} > \text{ITERMAX}$$

or for all the RHS we have:

$$\text{RNRM} < \text{SQRT}(N) * \text{XNRM} * \text{ANRM} * \text{EPS} * \text{BWDMAX}$$

where

- ▶ ITER is the number of the current iteration in the iterative refinement process
- ▶ RNRM is the infinity-norm of the residual
- ▶ XNRM is the infinity-norm of the solution
- ▶ ANRM is the infinity-operator-norm of the matrix A
- ▶ EPS is the machine epsilon that matches LAPACK `<t1>LAMCH('Epsilon')`

The value ITERMAX and BWDMAX are fixed to 50 and 1.0 respectively.

Solve process results will be indicated by output parameter **info**, see parameter description.

User should provide a large enough workspace allocated on the device for the `<t1><t2>gesv()` functions. The amount of bytes required can be queried by the respective `<t1><t2>gesv_bufferSize()` functions.

cusolverDn<t1><t2>gesv_bufferSize() functions will return workspace buffer size in bytes required for corresponding cusolverDn<t1><t2>gesv() function.

```
cusolverStatus_t
cusolverDnDDgesv_bufferSize(
    cusolverHandle_t      handle,
    int                   n,
    int                   nrhs,
    double                *dA,
    int                   ldda,
    int                   *dipiv,
    double                *dB,
    int                   lddb,
    double                *dX,
    int                   lddx,
    void                  *dwork,
    size_t                *lwork_bytes);
```

```
cusolverStatus_t
cusolverDnDSgesv_bufferSize(
    cusolverHandle_t      handle,
    int                   n,
    int                   nrhs,
    double                *dA,
    int                   ldda,
    int                   *dipiv,
    double                *dB,
    int                   lddb,
    double                *dX,
    int                   lddx,
    void                  *dwork,
    size_t                *lwork_bytes);
```

```
cusolverStatus_t
cusolverDnDHgesv_bufferSize(
    cusolverHandle_t      handle,
    int                   n,
    int                   nrhs,
    double                *dA,
    int                   ldda,
    int                   *dipiv,
    double                *dB,
    int                   lddb,
    double                *dX,
    int                   lddx,
    void                  *dwork,
    size_t                *lwork_bytes);
```

```
cusolverStatus_t
cusolverDnSSgesv_bufferSize(
    cusolverHandle_t      handle,
    int                   n,
    int                   nrhs,
    float                 *dA,
    int                   ldda,
    int                   *dipiv,
    float                 *dB,
    int                   lddb,
    float                 *dX,
    int                   lddx,
    void                  *dwork,
    size_t                *lwork_bytes);
```

```
cusolverStatus_t
cusolverDnSHgesv_bufferSize(
    cusolverHandle_t      handle,
    int                   n,
    int                   nrhs,
```

Parameters of cusolverDn<T1><T2>gesv_bufferSize() functions

parameter	Memory	In/out	Meaning
handle	host	input	Handle to the cusolverDN library context.
n	host	input	Number of rows and columns of square matrix A . Should be non-negative.
nrhs	host	input	Number of right hand sides to solve. Should be non-negative. nrhs is limited to 1 if selected IRS solver is CUSOLVER_IRS_GMRES.
dA	device	in	Matrix A with size n-by-n . Can be NULL .
ldda	host	input	leading dimension of two-dimensional array used to store matrix A . ldda \geq n .
dipiv	device	None	Pivoting sequence. Not used and can be NULL .
dB	device	in	Set of right hand sides B of size n-by-nrhs . Can be NULL .
lddb	host	input	leading dimension of two-dimensional array used to store matrix of right hand sides B . lddb \geq n .
dX	device	in	Set of solution vectors X of size n-by-nrhs . Can be NULL .
lddx	host	input	leading dimension of two-dimensional array used to store matrix of solution vectors X . lddx \geq n .
dwork	device	none	Pointer to device workspace. Not used and can be NULL .

lwork_bytes	host	out	Pointer to a variable where required size of temporary workspace in bytes will be stored. Can't be NULL.
--------------------	-------------	------------	--

```
cusolverStatus_t cusolverDnZZgesv(
    cusolverDnHandle_t handle,
    int n,
    int nrhs,
    cuDoubleComplex * dA,
    int ldda,
    int * dipiv,
    cuDoubleComplex * dB,
    int lddb,
    cuDoubleComplex * dX,
    int lddx,
    void * dWorkspace,
    size_t lwork_bytes,
    int * iter,
    int * d_info);
```

```
cusolverStatus_t cusolverDnZCgesv(
    cusolverDnHandle_t handle,
    int n,
    int nrhs,
    cuDoubleComplex * dA,
    int ldda,
    int * dipiv,
    cuDoubleComplex * dB,
    int lddb,
    cuDoubleComplex * dX,
    int lddx,
    void * dWorkspace,
    size_t lwork_bytes,
    int * iter,
    int * d_info);
```

```
cusolverStatus_t cusolverDnZKgesv(
    cusolverDnHandle_t handle,
    int n,
    int nrhs,
    cuDoubleComplex * dA,
    int ldda,
    int * dipiv,
    cuDoubleComplex * dB,
    int lddb,
    cuDoubleComplex * dX,
    int lddx,
    void * dWorkspace,
    size_t lwork_bytes,
    int * iter,
    int * d_info);
```

```
cusolverStatus_t cusolverDnCCgesv(
    cusolverDnHandle_t handle,
    int n,
    int nrhs,
    cuComplex * dA,
    int ldda,
    int * dipiv,
    cuComplex * dB,
    int lddb,
    cuComplex * dX,
    int lddx,
    void * dWorkspace,
    size_t lwork_bytes,
    int * iter,
    int * d_info);
```

Parameters of cusolverDn<T1><T2>gesv() functions

parameter	Memory	In/out	Meaning
handle	host	input	Handle to the cusolverDN library context.
n	host	input	Number of rows and columns of square matrix A . Should be non-negative.
nrhs	host	input	Number of right hand sides to solve. Should be non-negative. nrhs is limited to 1 if selected IRS solver is CUSOLVER_IRS_GMRES.
dA	device	in/out	Matrix A with size n-by-n . Can't be NULL . On return - unchanged if solve process iterative refinement converged. If not - will contain full precision factorization of matrix A : $\mathbf{A} = \mathbf{P} * \mathbf{L} * \mathbf{U}$, where P - permutation matrix defined by vector ipiv , L and U - lower and upper triangular matrices.
ldda	host	input	leading dimension of two-dimensional array used to store matrix A . ldda \geq n .
dipiv	device	in/out	Vector that defines permutation matrix for factorization - row i was interchanged with row ipiv[i] If NULL then no pivoting is performed.
dB	device	in/out	Set of right hand sides B of size n-by-nrhs . Can't be NULL .
lddb	host	input	leading dimension of two-dimensional array used to store matrix of right hand sides B . lddb \geq n .
dX	device	in/out	Set of solution vectors x of size n-by-nrhs . Can't be NULL .
lddx	host	input	leading dimension of two-dimensional array used to store matrix of solution vectors x . ldx \geq n .
dWorkspace	device	in/out	Pointer to a workspace in device memory of size lwork_bytes .
lwork_bytes	host	in	Size of provided device workspace in bytes. Should be at least what was returned by <code>cusolverDn<T1><T2>gesv_bufferSize()</code> function
iter	host	output	If iter is <ul style="list-style-type: none"> ▶ <0 : iterative refinement has failed, full precision factorization has been performed. ▶ -1 : taking into account machine parameters, n, nrhs, it is determined a priori it is not worth working in lower precision

			<ul style="list-style-type: none"> ▶ -2 : overflow of an entry when moving from double to lower precision ▶ -3 : failure of gesv function ▶ -31: solver stopped the iterative refinement after reaching maximum allowed iterations ▶ >0 : iter is a number of iterations solver performed to reach convergence criteria
info	device	output	Status of the iterative refinement on the return. If 0 - solve was successful. If info = -i then i-th argument has is not valid. If info = i, then $\sigma(i, i)$ computed in full precision is exactly zero. The factorization has been completed, but the factor U is exactly singular, so the solution could not be computed.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed: <ul style="list-style-type: none"> ▶ $n < 0$ ▶ $ldda < \max(1, n)$ ▶ $lddb < \max(1, n)$ ▶ $lddx < \max(1, n)$ ▶ NULL where it's not allowed ▶ $nrhs$ is larger than allowed
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 7.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.2.13. cusolverDnIRSXgesv()

This function is designed to perform same functionality as **cusolverDn<T1><T2>gesv()** functions, but wrapped in a more generic and expert interface that gives user more control to parametrize the function as well as it provides more informations on output. See **cusolverDn<T1><T2>gesv()** description for detailed explanation of the algorithm functionality and behaviour. **cusolverDnIRSXgesv()** allows additional control of the solver parameters such as setting the - lowest floating point precision authorized to be used by the solver, refinement solver type, maximum allowed number of iterative solver iterations, tolerance of the refinement solver - through Xgesv parameters structure and helper functions. **cusolverDnIRSXgesv()** provides additional informations on the output such as the convergence history of the residual array and the number of iterations needed to converge.

Following table provides authorized values for lowest precision parameter for specified full data type. Note that if lowest precision matches full datatype, then full precision factorization will be used

Supported lower floating point precisions for factorization for provided full datatype

Data Type	Supported values for lowest precision in Xgesv parameters structure
CUDA_R_32F	CUDA_R_32F, CUDA_R_16F
CUDA_R_64F	CUDA_R_64F, CUDA_R_32F, CUDA_R_16F
CUDA_C_32F	CUDA_C_32F, CUDA_C_16F
CUDA_C_64F	CUDA_C_64F, CUDA_C_32F, CUDA_C_16F

Solve process results will be indicated by output parameter **info**, see parameter description.

User should provide large enough workspace allocated on device for the `cusolverDnIRSXgesv()` function. Amount of bytes required for the function can be retrieved by respective function `cusolverDnIRSXgesv_bufferSize()`

`cusolverDnIRSXgesv_bufferSize()` functions will return workspace buffer size in bytes required for corresponding `cusolverDnXgesv()` function with given parameters.

```
cusolverStatus_t
cusolverDnIRSXgesv_bufferSize(
    cusolverDnHandle_t      handle,
    cusolverDnIRSParams_t   params,
    cusolver_int_t          n,
    cusolver_int_t          nrhs,
    size_t                  * lwork_bytes);
```

Parameters of `cusolverDnIRSXgesv_bufferSize()` functions

parameter	Memory	In/out	Meaning
handle	host	input	Handle to the cusolverDn library context.
params	host	input	Xgesv solve parameters
n	host	input	Number of rows and columns of the square matrix A . Should be non-negative.
nrhs	host	input	Number of right hand sides to solve. Should be non-negative. Note that, nrhs is limited to 1 if the selected IRS refinement solver is CUSOLVER_IRS_REFINE_GMRES, CUSOLVER_IRS_REFINE_GMRES_GMRES, CUSOLVER_IRS_REFINE_CLASSICAL_GMRES.
lwork_bytes	host	out	Pointer to a variable, where the required size in bytes, of the workspace will be stored after a call to

			cusolverDnIRSXgesv_bufferSize. Can't be NULL.
--	--	--	---

```
cusolverStatus_t cusolverDnIRSXgesv(
    cusolverDnHandle_t      handle,
    cusolverDnIRSParams_t   gesv_irs_params,
    cusolverDnIRSInfos_t    gesv_irs_infos,
    cudaDataType             inout_data_type,
    int                      n,
    int                      nrhs,
    void*                    *dA,
    int                      *ldda,
    int                      *dipiv,
    void*                    *dB,
    int                      *lddb,
    void*                    *dX,
    int                      *lddx,
    void*                    *dWorkspace,
    size_t                   lwork_bytes,
    int                      *nitters,
    int                      *dinfo);
```

Parameters of cusolverDnIRSXgesv() functions

parameter	Memory	In/out	Meaning
handle	host	input	Handle to the cusolverDn library context.
gesv_irs_params	host	input	Solve parameters handle
gesv_irs_infos	host	in/out	Info structure parameter handle where information about performed solve will be stored.
inout_data_type	host	input	Datatype of Matrix, right hand side and solution
n	host	input	Number of rows and columns of square matrix A . Should be non-negative.
nrhs	host	input	Number of right hand sides to solve. Should be non-negative. Note that, nrhs is limited to 1 if the selected IRS refinement solver is CUSOLVER_IRS_REFINE_GMRES, CUSOLVER_IRS_REFINE_GMRES_GMRES, CUSOLVER_IRS_REFINE_CLASSICAL_GMRES. Number of right hand sides to solve. Should be non-negative.
dA	device	in	Matrix A with size n-by-n . Can't be NULL . On return - unchanged if the iterative refinement solver converged. If not - will contain full precision factorization of matrix A : $\mathbf{A} = \mathbf{P} * \mathbf{L} * \mathbf{U}$, where P - permutation matrix defined by vector ipiv , L and U - lower and upper triangular matrices.
ldda	host	input	leading dimension of two-dimensional array used to store matrix A . ldda >= n ..

dipiv	device	in/out	Vector that defines permutation matrix for factorization - row i was interchanged with row $ipiv[i]$ If NULL then no pivoting is performed.
dB	device	in	Set of right hand sides B of size n -by- $nrhs$. Can be NULL.
lddb	host	input	leading dimension of two-dimensional array used to store matrix of right hand sides B . $ldb \geq n$.
dX	device	in	Set of solution vectors x of size n -by- $nrhs$. Can be NULL.
lddx	host	input	leading dimension of two-dimensional array used to store matrix of solution vectors x . $ldx \geq n$.
dWorkspace	device	in/out	Pointer to a workspace in device memory of size <code>lwork_bytes</code> .
lwork_bytes	host	in	Size of device workspace. Should be at least what was returned by <code>cusolverDnIRSXgesv_bufferSize()</code> function
niters	host	output	<p>If iter is</p> <ul style="list-style-type: none"> ▶ <0 : iterative refinement has failed, full precision factorization has been performed. ▶ -1 : taking into account machine parameters, n, $nrhs$, it is a priori not worth working in lower precision ▶ -2 : overflow of an entry when moving from double to lower precision ▶ -3 : failure of gesv function ▶ $-maxiter$: solver stopped the iterative refinement after reaching maximum allowed iterations ▶ >0 : iter is a number of iterations solver performed to reach convergence criteria
dinfo	device	output	Status of the iterative refinement on the return. If 0 - solve was successful. If $dinfo = -i$ then i -th argument has is not valid. If $dinfo = i$, then $\tau(i,i)$ computed in full precision is exactly zero. The factorization has been completed, but the factor U is exactly singular, so the solution could not be computed.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed: <ul style="list-style-type: none"> ▶ <code>n < 0</code> ▶ <code>lda < max(1, n)</code> ▶ <code>ldb < max(1, n)</code> ▶ <code>ldx < max(1, n)</code> ▶ <code>NULL</code> where it's not allowed ▶ <code>nrhs</code> is larger than allowed
<code>CUSOLVER_STATUS_ARCH_MISMATCH</code>	the device only supports compute capability 7.0 and above.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

2.4.3. Dense Eigenvalue Solver Reference

This chapter describes eigenvalue solver API of cuSolverDN, including bidiagonalization and SVD.

2.4.3.1. `cusolverDn<t>gebrd()`

These helper functions calculate the size of work buffers needed.

```
cusolverStatus_t
cusolverDnSgebrd_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int *lwork );

cusolverStatus_t
cusolverDnDgebrd_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int *lwork );

cusolverStatus_t
cusolverDnCgebrd_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int *lwork );

cusolverStatus_t
cusolverDnZgebrd_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int *lwork );
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSgebrd(cusolverDnHandle_t handle,
                 int m,
                 int n,
                 float *A,
                 int lda,
                 float *D,
                 float *E,
                 float *TAUQ,
                 float *TAUP,
                 float *Work,
                 int Lwork,
                 int *devInfo );

cusolverStatus_t
cusolverDnDgebrd(cusolverDnHandle_t handle,
                 int m,
                 int n,
                 double *A,
                 int lda,
                 double *D,
                 double *E,
                 double *TAUQ,
                 double *TAUP,
                 double *Work,
                 int Lwork,
                 int *devInfo );
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCgebrd(cusolverDnHandle_t handle,
                 int m,
                 int n,
                 cuComplex *A,
                 int lda,
                 float *D,
                 float *E,
                 cuComplex *TAUQ,
                 cuComplex *TAUP,
                 cuComplex *Work,
                 int Lwork,
                 int *devInfo );

cusolverStatus_t
cusolverDnZgebrd(cusolverDnHandle_t handle,
                 int m,
                 int n,
                 cuDoubleComplex *A,
                 int lda,
                 double *D,
                 double *E,
                 cuDoubleComplex *TAUQ,
                 cuDoubleComplex *TAUP,
                 cuDoubleComplex *Work,
                 int Lwork,
                 int *devInfo );
```

This function reduces a general $m \times n$ matrix A to a real upper or lower bidiagonal form B by an orthogonal transformation: $Q^H * A * P = B$

If $m \geq n$, B is upper bidiagonal; if $m < n$, B is lower bidiagonal.

The matrix Q and P are overwritten into matrix A in the following sense:

if $m \geq n$, the diagonal and the first superdiagonal are overwritten with the upper bidiagonal matrix B ; the elements below the diagonal, with the array **TAUQ**, represent the orthogonal matrix Q as a product of elementary reflectors, and the elements above the first superdiagonal, with the array **TAUP**, represent the orthogonal matrix P as a product of elementary reflectors.

if $m < n$, the diagonal and the first subdiagonal are overwritten with the lower bidiagonal matrix B ; the elements below the first subdiagonal, with the array **TAUQ**, represent the orthogonal matrix Q as a product of elementary reflectors, and the elements above the diagonal, with the array **TAUP**, represent the orthogonal matrix P as a product of elementary reflectors.

The user has to provide working space which is pointed by input parameter **Work**. The input parameter **Lwork** is size of the working space, and it is returned by **gebrd_bufferSize()**.

If output parameter **devInfo** = $-i$ (less than zero), the i -th parameter is wrong (not counting handle).

Remark: **gebrd** only supports $m \geq n$.

API of gebrd

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
m	host	input	number of rows of matrix A .
n	host	input	number of columns of matrix A .
A	device	in/out	<type> array of dimension $lda * n$ with lda is not less than $\max(1, n)$.
lda	host	input	leading dimension of two-dimensional array used to store matrix A .
D	device	output	real array of dimension $\min(m, n)$. The diagonal elements of the bidiagonal matrix B : $D(i) = A(i, i)$.
E	device	output	real array of dimension $\min(m, n)$. The off-diagonal elements of the bidiagonal matrix B : if $m \geq n$, $E(i) = A(i, i+1)$ for $i = 1, 2, \dots, n-1$; if $m < n$, $E(i) = A(i+1, i)$ for $i = 1, 2, \dots, m-1$.
TAUQ	device	output	<type> array of dimension $\min(m, n)$. The scalar factors of the elementary reflectors which represent the orthogonal matrix Q .
TAUP	device	output	<type> array of dimension $\min(m, n)$. The scalar factors of the elementary reflectors which represent the orthogonal matrix P .

Work	device	in/out	working space, <type> array of size Lwork .
Lwork	host	input	size of Work , returned by gebrd_bufferSize .
devInfo	device	output	if devInfo = 0, the operation is successful. if devInfo = -i, the i-th parameter is wrong (not counting handle).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (m , n < 0, or lda < max (1, m)).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.3.2. cusolverDn<t>orgbr()

These helper functions calculate the size of work buffers needed.

```
cusolverStatus_t
cusolverDnSorgbr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    int m,
    int n,
    int k,
    const float *A,
    int lda,
    const float *tau,
    int *lwork);

cusolverStatus_t
cusolverDnDorgbr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    int m,
    int n,
    int k,
    const double *A,
    int lda,
    const double *tau,
    int *lwork);

cusolverStatus_t
cusolverDnCungbr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    int m,
    int n,
    int k,
    const cuComplex *A,
    int lda,
    const cuComplex *tau,
    int *lwork);

cusolverStatus_t
cusolverDnZungbr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    int m,
    int n,
    int k,
    const cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *tau,
    int *lwork);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSorgbr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    int m,
    int n,
    int k,
    float *A,
    int lda,
    const float *tau,
    float *work,
    int lwork,
    int *devInfo);
```

```
cusolverStatus_t
cusolverDnDorgbr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    int m,
    int n,
    int k,
    double *A,
    int lda,
    const double *tau,
    double *work,
    int lwork,
    int *devInfo);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCungbr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    int m,
    int n,
    int k,
    cuComplex *A,
    int lda,
    const cuComplex *tau,
    cuComplex *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnZungbr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    int m,
    int n,
    int k,
    cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *tau,
    cuDoubleComplex *work,
    int lwork,
    int *devInfo);
```

This function generates one of the unitary matrices **Q** or **P**H** determined by **gebrd** when reducing a matrix A to bidiagonal form: $Q^H * A * P = B$

Q and **P**H** are defined as products of elementary reflectors H(i) or G(i) respectively.

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is size of the working space, and it is returned by **orgbr_bufferSize()**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

API of orgbr

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
side	host	input	if side = CUBLAS_SIDE_LEFT, generate Q. if side = CUBLAS_SIDE_RIGHT, generate P**T.
m	host	input	number of rows of matrix Q or P**T.
n	host	input	if side = CUBLAS_SIDE_LEFT, m >= n >= min(m,k). if side = CUBLAS_SIDE_RIGHT, n >= m >= min(n,k).
k	host	input	if side = CUBLAS_SIDE_LEFT, the number of columns in the original m-by-k matrix reduced by gebrd. if side

			= CUBLAS_SIDE_RIGHT, the number of rows in the original k-by-n matrix reduced by <code>gebrd</code> .
A	device	in/out	<type> array of dimension <code>lda * n</code> On entry, the vectors which define the elementary reflectors, as returned by <code>gebrd</code> . On exit, the m-by-n matrix <code>Q</code> or <code>P**T</code> .
lda	host	input	leading dimension of two-dimensional array used to store matrix <code>A</code> . <code>lda</code> \geq <code>max(1,m)</code> ;
tau	device	output	<type> array of dimension <code>min(m,k)</code> if <code>side</code> is CUBLAS_SIDE_LEFT; of dimension <code>min(n,k)</code> if <code>side</code> is CUBLAS_SIDE_RIGHT; <code>tau(i)</code> must contain the scalar factor of the elementary reflector <code>H(i)</code> or <code>G(i)</code> , which determines <code>Q</code> or <code>P**T</code> , as returned by <code>gebrd</code> in its array argument <code>TAUQ</code> or <code>TAUP</code> .
work	device	in/out	working space, <type> array of size <code>lwork</code> .
lwork	host	input	size of working array <code>work</code> .
devInfo	device	output	if <code>info</code> = 0, the ormqr is successful. if <code>info</code> = -i, the i-th parameter is wrong (not counting handle).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (<code>m,n</code> <0 or wrong <code>lda</code>).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.3.3. cusolverDn<t>sytrd()

These helper functions calculate the size of work buffers needed.

```
cusolverStatus_t
cusolverDnSsytrd_bufferSize(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    const float *A,
    int lda,
    const float *d,
    const float *e,
    const float *tau,
    int *lwork);

cusolverStatus_t
cusolverDnDsytrd_bufferSize(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    const double *A,
    int lda,
    const double *d,
    const double *e,
    const double *tau,
    int *lwork);

cusolverStatus_t
cusolverDnChetrd_bufferSize(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    const cuComplex *A,
    int lda,
    const float *d,
    const float *e,
    const cuComplex *tau,
    int *lwork);

cusolverStatus_t
cusolverDnZhetrd_bufferSize(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    const cuDoubleComplex *A,
    int lda,
    const double *d,
    const double *e,
    const cuDoubleComplex *tau,
    int *lwork);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSsytrd(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    float *A,
    int lda,
    float *d,
    float *e,
    float *tau,
    float *work,
    int lwork,
    int *devInfo);
```

```
cusolverStatus_t
cusolverDnDsytrd(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    double *A,
    int lda,
    double *d,
    double *e,
    double *tau,
    double *work,
    int lwork,
    int *devInfo);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnChetrd(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    cuComplex *A,
    int lda,
    float *d,
    float *e,
    cuComplex *tau,
    cuComplex *work,
    int lwork,
    int *devInfo);
```

```
cusolverStatus_t CUDENSEAPI cusolverDnZhetrd(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    cuDoubleComplex *A,
    int lda,
    double *d,
    double *e,
    cuDoubleComplex *tau,
    cuDoubleComplex *work,
    int lwork,
    int *devInfo);
```

This function reduces a general symmetric (Hermitian) $n \times n$ matrix A to real symmetric tridiagonal form T by an orthogonal transformation: $Q^H * A * Q = T$

As an output, A contains T and householder reflection vectors. If `uplo = CUBLAS_FILL_MODE_UPPER`, the diagonal and first superdiagonal of A are overwritten by the corresponding elements of the tridiagonal matrix T , and the elements above the first superdiagonal, with the array `tau`, represent the orthogonal matrix Q as a product of elementary reflectors; If `uplo = CUBLAS_FILL_MODE_LOWER`, the diagonal and first subdiagonal of A are overwritten by the corresponding elements of the tridiagonal matrix T , and the elements below the first subdiagonal, with the array `tau`, represent the orthogonal matrix Q as a product of elementary reflectors.

The user has to provide working space which is pointed by input parameter `work`. The input parameter `lwork` is size of the working space, and it is returned by `sytrd_bufferSize()`.

If output parameter `devInfo = -i` (less than zero), the `i-th` parameter is wrong (not counting handle).

API of sytrd

parameter	Memory	In/out	Meaning
<code>handle</code>	<code>host</code>	<code>input</code>	handle to the cuSolverDN library context.
<code>uplo</code>	<code>host</code>	<code>input</code>	specifies which part of A is stored. <code>uplo = CUBLAS_FILL_MODE_LOWER</code> : Lower triangle of A is stored. <code>uplo = CUBLAS_FILL_MODE_UPPER</code> : Upper triangle of A is stored.
<code>n</code>	<code>host</code>	<code>input</code>	number of rows (columns) of matrix A .
<code>A</code>	<code>device</code>	<code>in/out</code>	<type> array of dimension <code>lda * n</code> with <code>lda</code> is not less than <code>max(1, n)</code> . If <code>uplo = CUBLAS_FILL_MODE_UPPER</code> , the leading n -by- n upper triangular part of A contains the upper triangular part of the matrix A , and the strictly lower triangular part of A is not referenced. If <code>uplo = CUBLAS_FILL_MODE_LOWER</code> , the leading n -by- n lower triangular part of A contains the lower triangular part of the matrix A , and the strictly upper triangular part of A is not referenced. On exit, A is overwritten by T and householder reflection vectors.
<code>lda</code>	<code>host</code>	<code>input</code>	leading dimension of two-dimensional array used to store matrix A . <code>lda</code> \geq <code>max(1, n)</code> .
<code>D</code>	<code>device</code>	<code>output</code>	real array of dimension <code>n</code> . The diagonal elements of the tridiagonal matrix T : <code>D(i) = A(i, i)</code> .
<code>E</code>	<code>device</code>	<code>output</code>	real array of dimension <code>(n-1)</code> . The off-diagonal elements of the tridiagonal matrix T : if <code>uplo = CUBLAS_FILL_MODE_UPPER</code> ,

			$E(i) = A(i, i+1)$. if <code>uplo = CUBLAS_FILL_MODE_LOWER</code> $E(i) = A(i+1, i)$.
<code>tau</code>	<code>device</code>	<code>output</code>	<type> array of dimension $(n-1)$. The scalar factors of the elementary reflectors which represent the orthogonal matrix Q .
<code>work</code>	<code>device</code>	<code>in/out</code>	working space, <type> array of size <code>lwork</code> .
<code>lwork</code>	<code>host</code>	<code>input</code>	size of <code>work</code> , returned by <code>sytrd_bufferSize</code> .
<code>devInfo</code>	<code>device</code>	<code>output</code>	if <code>devInfo = 0</code> , the operation is successful. if <code>devInfo = -i</code> , the i -th parameter is wrong (not counting handle).

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed ($n < 0$, or $lda < \max(1, n)$, or <code>uplo</code> is not <code>CUBLAS_FILL_MODE_LOWER</code> or <code>CUBLAS_FILL_MODE_UPPER</code>).
<code>CUSOLVER_STATUS_ARCH_MISMATCH</code>	the device only supports compute capability 2.0 and above.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

2.4.3.4. `cusolverDn<t>ormtr()`

These helper functions calculate the size of work buffers needed.

```
cusolverStatus_t
cusolverDnSormtr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasFillMode_t uplo,
    cublasOperation_t trans,
    int m,
    int n,
    const float *A,
    int lda,
    const float *tau,
    const float *C,
    int ldc,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnDormtr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasFillMode_t uplo,
    cublasOperation_t trans,
    int m,
    int n,
    const double *A,
    int lda,
    const double *tau,
    const double *C,
    int ldc,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnCunmtr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasFillMode_t uplo,
    cublasOperation_t trans,
    int m,
    int n,
    const cuComplex *A,
    int lda,
    const cuComplex *tau,
    const cuComplex *C,
    int ldc,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnZunmtr_bufferSize(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasFillMode_t uplo,
    cublasOperation_t trans,
    int m,
    int n,
    const cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *tau,
    const cuDoubleComplex *C,
    int ldc,
    int *lwork);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSormtr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasFillMode_t uplo,
    cublasOperation_t trans,
    int m,
    int n,
    float *A,
    int lda,
    float *tau,
    float *C,
    int ldc,
    float *work,
    int lwork,
    int *devInfo);
```

```
cusolverStatus_t
cusolverDnDormtr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasFillMode_t uplo,
    cublasOperation_t trans,
    int m,
    int n,
    double *A,
    int lda,
    double *tau,
    double *C,
    int ldc,
    double *work,
    int lwork,
    int *devInfo);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCumtr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasFillMode_t uplo,
    cublasOperation_t trans,
    int m,
    int n,
    cuComplex *A,
    int lda,
    cuComplex *tau,
    cuComplex *C,
    int ldc,
    cuComplex *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnZumtr(
    cusolverDnHandle_t handle,
    cublasSideMode_t side,
    cublasFillMode_t uplo,
    cublasOperation_t trans,
    int m,
    int n,
    cuDoubleComplex *A,
    int lda,
    cuDoubleComplex *tau,
    cuDoubleComplex *C,
    int ldc,
    cuDoubleComplex *work,
    int lwork,
    int *devInfo);
```

This function overwrites $m \times n$ matrix **C** by

$$C = \begin{cases} \text{op}(Q) * C & \text{if side} == \text{CUBLAS_SIDE_LEFT} \\ C * \text{op}(Q) & \text{if side} == \text{CUBLAS_SIDE_RIGHT} \end{cases}$$

where **Q** is a unitary matrix formed by a sequence of elementary reflection vectors from **sytrd**.

The operation on **Q** is defined by

$$\text{op}(Q) = \begin{cases} Q & \text{if transa} == \text{CUBLAS_OP_N} \\ Q^T & \text{if transa} == \text{CUBLAS_OP_T} \\ Q^H & \text{if transa} == \text{CUBLAS_OP_C} \end{cases}$$

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is size of the working space, and it is returned by **ormtr_bufferSize()**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

API of ormtr

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
side	host	input	side = CUBLAS_SIDE_LEFT, apply Q or Q**T from the Left; side = CUBLAS_SIDE_RIGHT, apply Q or Q**T from the Right.
uplo	host	input	uplo = CUBLAS_FILL_MODE_LOWER: Lower triangle of A contains elementary reflectors from sytrd. uplo = CUBLAS_FILL_MODE_UPPER: Upper triangle of A contains elementary reflectors from sytrd.
trans	host	input	operation op(Q) that is non- or (conj.) transpose.
m	host	input	number of rows of matrix c.
n	host	input	number of columns of matrix c.
A	device	in/out	<type> array of dimension lda * m if side = CUBLAS_SIDE_LEFT; lda * n if side = CUBLAS_SIDE_RIGHT. The matrix A from sytrd contains the elementary reflectors.
lda	host	input	leading dimension of two-dimensional array used to store matrix A. if side is CUBLAS_SIDE_LEFT, lda >= max(1,m); if side is CUBLAS_SIDE_RIGHT, lda >= max(1,n).
tau	device	output	<type> array of dimension (m-1) if side is CUBLAS_SIDE_LEFT; of dimension (n-1) if side is CUBLAS_SIDE_RIGHT; The vector tau is from sytrd, so tau(i) is the scalar of i-th elementary reflection vector.
C	device	in/out	<type> array of size ldc * n. On exit, C is overwritten by op(Q) * C or C * op(Q).
ldc	host	input	leading dimension of two-dimensional array of matrix c. ldc >= max(1,m).
work	device	in/out	working space, <type> array of size lwork.
lwork	host	input	size of working array work.
devInfo	device	output	if devInfo = 0, the ormqr is successful. if devInfo = -i, the i-th parameter is wrong (not counting handle).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed (<code>m, n < 0</code> or wrong <code>lda</code> or <code>ldc</code>).
<code>CUSOLVER_STATUS_ARCH_MISMATCH</code>	the device only supports compute capability 2.0 and above.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

2.4.3.5. `cusolverDn<t>orgtr()`

These helper functions calculate the size of work buffers needed.

```
cusolverStatus_t
cusolverDnSorgtr_bufferSize(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    const float *A,
    int lda,
    const float *tau,
    int *lwork);

cusolverStatus_t
cusolverDnDorgtr_bufferSize(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    const double *A,
    int lda,
    const double *tau,
    int *lwork);

cusolverStatus_t
cusolverDnCungtr_bufferSize(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    const cuComplex *A,
    int lda,
    const cuComplex *tau,
    int *lwork);

cusolverStatus_t
cusolverDnZungtr_bufferSize(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    const cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *tau,
    int *lwork);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSorgtr(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    float *A,
    int lda,
    const float *tau,
    float *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnDorgtr(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    double *A,
    int lda,
    const double *tau,
    double *work,
    int lwork,
    int *devInfo);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCungtr(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    cuComplex *A,
    int lda,
    const cuComplex *tau,
    cuComplex *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnZungtr(
    cusolverDnHandle_t handle,
    cublasFillMode_t uplo,
    int n,
    cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *tau,
    cuDoubleComplex *work,
    int lwork,
    int *devInfo);
```

This function generates a unitary matrix Q which is defined as the product of $n-1$ elementary reflectors of order n , as returned by **sytrd**:

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is size of the working space, and it is returned by **orgtr_bufferSize()**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

API of orgtr

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
uplo	host	input	uplo = CUBLAS_FILL_MODE_LOWER: Lower triangle of A contains elementary reflectors from sytrd . uplo = CUBLAS_FILL_MODE_UPPER: Upper triangle of A contains elementary reflectors from sytrd .
n	host	input	number of rows (columns) of matrix Q .
A	device	in/out	<type> array of dimension lda * n On entry, matrix A from sytrd contains the elementary reflectors. On exit, matrix A contains the n-by-n orthogonal matrix Q .
lda	host	input	leading dimension of two-dimensional array used to store matrix A . lda >= max(1,n).
tau	device	output	<type> array of dimension (n-1) tau(i) is the scalar of i-th elementary reflection vector.
work	device	in/out	working space, <type> array of size lwork .
lwork	host	input	size of working array work .
devInfo	device	output	if devInfo = 0, the orgtr is successful. if devInfo = -i, the i-th parameter is wrong (not counting handle).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (n<0 or wrong lda).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.3.6. cusolverDn<t>gesvd()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverDnSgesvd_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int *lwork );

cusolverStatus_t
cusolverDnDgesvd_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int *lwork );

cusolverStatus_t
cusolverDnCgesvd_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int *lwork );

cusolverStatus_t
cusolverDnZgesvd_bufferSize(
    cusolverDnHandle_t handle,
    int m,
    int n,
    int *lwork );
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSgesvd (
    cusolverDnHandle_t handle,
    signed char jobu,
    signed char jobvt,
    int m,
    int n,
    float *A,
    int lda,
    float *S,
    float *U,
    int ldu,
    float *VT,
    int ldvt,
    float *work,
    int lwork,
    float *rwork,
    int *devInfo);
```

```
cusolverStatus_t
cusolverDnDgesvd (
    cusolverDnHandle_t handle,
    signed char jobu,
    signed char jobvt,
    int m,
    int n,
    double *A,
    int lda,
    double *S,
    double *U,
    int ldu,
    double *VT,
    int ldvt,
    double *work,
    int lwork,
    double *rwork,
    int *devInfo);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnGgesvd (
    cusolverDnHandle_t handle,
    signed char jobu,
    signed char jobvt,
    int m,
    int n,
    cuComplex *A,
    int lda,
    float *S,
    cuComplex *U,
    int ldu,
    cuComplex *VT,
    int ldvt,
    cuComplex *work,
    int lwork,
    float *rwork,
    int *devInfo);

cusolverStatus_t
cusolverDnZgesvd (
    cusolverDnHandle_t handle,
    signed char jobu,
    signed char jobvt,
    int m,
    int n,
    cuDoubleComplex *A,
    int lda,
    double *S,
    cuDoubleComplex *U,
    int ldu,
    cuDoubleComplex *VT,
    int ldvt,
    cuDoubleComplex *work,
    int lwork,
    double *rwork,
    int *devInfo);
```

This function computes the singular value decomposition (SVD) of a $m \times n$ matrix **A** and corresponding the left and/or right singular vectors. The SVD is written

$$A = U \Sigma V^H$$

where Σ is an $m \times n$ matrix which is zero except for its $\min(m, n)$ diagonal elements, **U** is an $m \times m$ unitary matrix, and **V** is an $n \times n$ unitary matrix. The diagonal elements of Σ are the singular values of **A**; they are real and non-negative, and are returned in descending order. The first $\min(m, n)$ columns of **U** and **V** are the left and right singular vectors of **A**.

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is size of the working space, and it is returned by **gesvd_bufferSize()**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle). if **bdsqr** did not converge, **devInfo** specifies how many superdiagonals of an intermediate bidiagonal form did not converge to zero.

The **rwork** is real array of dimension $(\min(m,n)-1)$. If **devInfo**>0 and **rwork** is not nil, **rwork** contains the unconverged superdiagonal elements of an upper bidiagonal matrix. This is slightly different from LAPACK which puts unconverged superdiagonal elements in **work** if type is **real**; in **rwork** if type is **complex**. **rwork** can be a NULL pointer if the user does not want the information from superdiagonal.

Appendix G.1 provides a simple example of **gesvd**.

Remark 1: **gesvd** only supports $m \geq n$.

Remark 2: the routine returns V^H , not **v**.

API of gesvd

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
jobu	host	input	specifies options for computing all or part of the matrix U : = 'A': all m columns of U are returned in array U ; = 'S': the first $\min(m,n)$ columns of U (the left singular vectors) are returned in the array U ; = 'O': the first $\min(m,n)$ columns of U (the left singular vectors) are overwritten on the array A ; = 'N': no columns of U (no left singular vectors) are computed.
jobvt	host	input	specifies options for computing all or part of the matrix V^{*T} : = 'A': all N rows of V^{*T} are returned in the array VT ; = 'S': the first $\min(m,n)$ rows of V^{*T} (the right singular vectors) are returned in the array VT ; = 'O': the first $\min(m,n)$ rows of V^{*T} (the right singular vectors) are overwritten on the array A ; = 'N': no rows of V^{*T} (no right singular vectors) are computed.
m	host	input	number of rows of matrix A .
n	host	input	number of columns of matrix A .
A	device	in/out	<type> array of dimension $lda * n$ with lda is not less than $\max(1, m)$. On exit, the contents of A are destroyed.
lda	host	input	leading dimension of two-dimensional array used to store matrix A .
S	device	output	real array of dimension $\min(m, n)$. The singular values of A , sorted so that $s(i) \geq s(i+1)$.
U	device	output	<type> array of dimension $ldu * m$ with ldu is not less than $\max(1, m)$. U contains the $m \times m$ unitary matrix U .
ldu	host	input	leading dimension of two-dimensional array used to store matrix U .

VT	device	output	<type> array of dimension $ldvt * n$ with $ldvt$ is not less than $\max(1, n)$. VT contains the $n \times n$ unitary matrix V^{**T} .
ldvt	host	input	leading dimension of two-dimensional array used to store matrix vt.
work	device	in/out	working space, <type> array of size lwork .
lwork	host	input	size of work , returned by gesvd_bufferSize .
rwork	device	input	real array of dimension $\min(m, n) - 1$. It contains the unconverged superdiagonal elements of an upper bidiagonal matrix if devInfo > 0.
devInfo	device	output	if devInfo = 0, the operation is successful. if devInfo = -i, the i-th parameter is wrong (not counting handle). if devInfo > 0, devInfo indicates how many superdiagonals of an intermediate bidiagonal form did not converge to zero.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($m, n < 0$ or $lda < \max(1, m)$ or $ldu < \max(1, m)$ or $ldvt < \max(1, n)$).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.3.7. cusolverDn<t>gesvdj()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverDnSgesvdj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int econ,
    int m,
    int n,
    const float *A,
    int lda,
    const float *S,
    const float *U,
    int ldu,
    const float *V,
    int ldv,
    int *lwork,
    gesvdjInfo_t params);
```

```
cusolverStatus_t
cusolverDnDgesvdj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int econ,
    int m,
    int n,
    const double *A,
    int lda,
    const double *S,
    const double *U,
    int ldu,
    const double *V,
    int ldv,
    int *lwork,
    gesvdjInfo_t params);
```

```
cusolverStatus_t
cusolverDnCgesvdj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int econ,
    int m,
    int n,
    const cuComplex *A,
    int lda,
    const float *S,
    const cuComplex *U,
    int ldu,
    const cuComplex *V,
    int ldv,
    int *lwork,
    gesvdjInfo_t params);
```

```
cusolverStatus_t
cusolverDnZgesvdj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int econ,
    int m,
    int n,
    const cuDoubleComplex *A,
    int lda,
    const double *S,
    const cuDoubleComplex *U,
    int ldu,
    const cuDoubleComplex *V,
    int ldv,
    int *lwork,
    gesvdjInfo_t params);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSgesvdj(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int econ,
    int m,
    int n,
    float *A,
    int lda,
    float *S,
    float *U,
    int ldu,
    float *V,
    int ldv,
    float *work,
    int lwork,
    int *info,
    gesvdjInfo_t params);
```

```
cusolverStatus_t
cusolverDnDgesvdj(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int econ,
    int m,
    int n,
    double *A,
    int lda,
    double *S,
    double *U,
    int ldu,
    double *V,
    int ldv,
    double *work,
    int lwork,
    int *info,
    gesvdjInfo_t params);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCgesvdj(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int econ,
    int m,
    int n,
    cuComplex *A,
    int lda,
    float *S,
    cuComplex *U,
    int ldu,
    cuComplex *V,
    int ldv,
    cuComplex *work,
    int lwork,
    int *info,
    gesvdjInfo_t params);

cusolverStatus_t
cusolverDnZgesvdj(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int econ,
    int m,
    int n,
    cuDoubleComplex *A,
    int lda,
    double *S,
    cuDoubleComplex *U,
    int ldu,
    cuDoubleComplex *V,
    int ldv,
    cuDoubleComplex *work,
    int lwork,
    int *info,
    gesvdjInfo_t params);
```

This function computes the singular value decomposition (SVD) of a $m \times n$ matrix \mathbf{A} and corresponding the left and/or right singular vectors. The SVD is written

$$\mathbf{A} = \mathbf{U} * \mathbf{\Sigma} * \mathbf{V}^H$$

where $\mathbf{\Sigma}$ is an $m \times n$ matrix which is zero except for its $\min(m, n)$ diagonal elements, \mathbf{U} is an $m \times m$ unitary matrix, and \mathbf{V} is an $n \times n$ unitary matrix. The diagonal elements of $\mathbf{\Sigma}$ are the singular values of \mathbf{A} ; they are real and non-negative, and are returned in descending order. The first $\min(m, n)$ columns of \mathbf{U} and \mathbf{V} are the left and right singular vectors of \mathbf{A} .

gesvdj has the same functionality as **gesvd**. The difference is that **gesvd** uses QR algorithm and **gesvdj** uses Jacobi method. The parallelism of Jacobi method gives GPU better performance on small and medium size matrices. Moreover the user can configure **gesvdj** to perform approximation up to certain accuracy.

gesvdj iteratively generates a sequence of unitary matrices to transform matrix **A** to the following form

$$U^H * A * V = S + E$$

where **S** is diagonal and diagonal of **E** is zero.

During the iterations, the Frobenius norm of **E** decreases monotonically. As **E** goes down to zero, **S** is the set of singular values. In practice, Jacobi method stops if

$$||E||_F \leq \text{eps} * ||A||_F$$

where **eps** is given tolerance.

gesvdj has two parameters to control the accuracy. First parameter is tolerance (**eps**). The default value is machine accuracy but The user can use function **cusolverDnXgesvdjSetTolerance** to set a priori tolerance. The second parameter is maximum number of sweeps which controls number of iterations of Jacobi method. The default value is 100 but the user can use function **cusolverDnXgesvdjSetMaxSweeps** to set a proper bound. The experimentis show 15 sweeps are good enough to converge to machine accuracy. **gesvdj** stops either tolerance is met or maximum number of sweeps is met.

Jacobi method has quadratic convergence, so the accuracy is not proportional to number of sweeps. To guarantee certain accuracy, the user should configure tolerance only.

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is the size of the working space, and it is returned by **gesvdj_bufferSize()**.

If output parameter **info** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle). If **info** = **min(m,n)+1**, **gesvdj** does not converge under given tolerance and maximum sweeps.

If the user sets an improper tolerance, **gesvdj** may not converge. For example, tolerance should not be smaller than machine accuracy.

Appendix G.2 provides a simple example of **gesvdj**.

Remark 1: **gesvdj** supports any combination of **m** and **n**.

Remark 2: the routine returns **v**, not V^H . This is different from **gesvd**.

API of gesvdj

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
jobz	host	input	specifies options to either compute singular value only or singular vectors as well: jobz = CUSOLVER_EIG_MODE_NOVECTOR : Compute singular values only; jobz = CUSOLVER_EIG_MODE_VECTOR : Compute singular values and singular vectors.
econ	host	input	econ = 1 for economy size for u and v .

m	host	input	number of rows of matrix A .
n	host	input	number of columns of matrix A .
A	device	in/out	<type> array of dimension lda * n with lda is not less than $\max(1, m)$. On exit, the contents of A are destroyed.
lda	host	input	leading dimension of two-dimensional array used to store matrix A .
S	device	output	real array of dimension $\min(m, n)$. The singular values of A , sorted so that $s(i) \geq s(i+1)$.
U	device	output	<type> array of dimension ldu * m if econ is zero. If econ is nonzero, the dimension is ldu * $\min(m, n)$. U contains the left singular vectors.
ldu	host	input	leading dimension of two-dimensional array used to store matrix U . ldu is not less than $\max(1, m)$.
V	device	output	<type> array of dimension ldv * n if econ is zero. If econ is nonzero, the dimension is ldv * $\min(m, n)$. V contains the right singular vectors.
ldv	host	input	leading dimension of two-dimensional array used to store matrix V . ldv is not less than $\max(1, n)$.
work	device	in/out	<type> array of size lwork , working space.
lwork	host	input	size of work , returned by gesvdj_bufferSize .
info	device	output	if info = 0, the operation is successful. if info = -i, the i-th parameter is wrong (not counting handle). if info = $\min(m, n) + 1$, gesvdj dose not converge under given tolerance and maximum sweeps.
params	host	in/out	structure filled with parameters of Jacobi algorithm and results of gesvdj .

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($m, n < 0$ or $lda < \max(1, m)$ or $ldu < \max(1, m)$ or $ldv < \max(1, n)$ or jobz is not CUSOLVER_EIG_MODE_NOVECTOR or CUSOLVER_EIG_MODE_VECTOR).
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.3.8. cusolverDn<t>gesvdjBatched()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverDnSgesvdjBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int m,
    int n,
    const float *A,
    int lda,
    const float *S,
    const float *U,
    int ldu,
    const float *V,
    int ldv,
    int *lwork,
    gesvdjInfo_t params,
    int batchSize);
```

```
cusolverStatus_t
cusolverDnDgesvdjBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int m,
    int n,
    const double *A,
    int lda,
    const double *S,
    const double *U,
    int ldu,
    const double *V,
    int ldv,
    int *lwork,
    gesvdjInfo_t params,
    int batchSize);
```

```
cusolverStatus_t
cusolverDnCgesvdjBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int m,
    int n,
    const cuComplex *A,
    int lda,
    const float *S,
    const cuComplex *U,
    int ldu,
    const cuComplex *V,
    int ldv,
    int *lwork,
    gesvdjInfo_t params,
    int batchSize);
```

```
cusolverStatus_t
cusolverDnZgesvdjBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int m,
    int n,
    const cuDoubleComplex *A,
    int lda,
    const double *S,
    const cuDoubleComplex *U,
    int ldu,
    const cuDoubleComplex *V,
    int ldv,
    int *lwork,
    gesvdjInfo_t params,
    int batchSize);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSgesvdjBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int m,
    int n,
    float *A,
    int lda,
    float *S,
    float *U,
    int ldu,
    float *V,
    int ldv,
    float *work,
    int lwork,
    int *info,
    gesvdjInfo_t params,
    int batchSize);
```

```
cusolverStatus_t
cusolverDnDgesvdjBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int m,
    int n,
    double *A,
    int lda,
    double *S,
    double *U,
    int ldu,
    double *V,
    int ldv,
    double *work,
    int lwork,
    int *info,
    gesvdjInfo_t params,
    int batchSize);
```


The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCgesvdjBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int m,
    int n,
    cuComplex *A,
    int lda,
    float *S,
    cuComplex *U,
    int ldu,
    cuComplex *V,
    int ldv,
    cuComplex *work,
    int lwork,
    int *info,
    gesvdjInfo_t params,
    int batchSize);

cusolverStatus_t
cusolverDnZgesvdjBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int m,
    int n,
    cuDoubleComplex *A,
    int lda,
    double *S,
    cuDoubleComplex *U,
    int ldu,
    cuDoubleComplex *V,
    int ldv,
    cuDoubleComplex *work,
    int lwork,
    int *info,
    gesvdjInfo_t params,
    int batchSize);
```

This function computes singular values and singular vectors of a sequence of general $\mathbf{m} \times \mathbf{n}$ matrices

$$A_j = U_j * \Sigma_j * V_j^H$$

where Σ_j is a real $\mathbf{m} \times \mathbf{n}$ diagonal matrix which is zero except for its $\min(\mathbf{m}, \mathbf{n})$ diagonal elements. U_j (left singular vectors) is a $\mathbf{m} \times \mathbf{m}$ unitary matrix and V_j (right singular vectors) is a $\mathbf{n} \times \mathbf{n}$ unitary matrix. The diagonal elements of Σ_j are the singular values of A_j in either descending order or non-sorting order.

gesvdjBatched performs **gesvdj** on each matrix. It requires that all matrices are of the same size \mathbf{m}, \mathbf{n} no greater than 32 and are packed in contiguous way,

$$A = (A_0 \ A_1 \ \dots)$$

Each matrix is column-major with leading dimension **lda**, so the formula for random access is $A_k(i,j) = A[i + lda*j + lda*n*k]$.

The parameter **s** also contains singular values of each matrix in contiguous way,

$$S = (S_0 \ S_1 \ \dots)$$

The formula for random access of **s** is $S_k(j) = S[j + \min(m,n)*k]$.

Except for tolerance and maximum sweeps, **gesvdjBatched** can either sort the singular values in descending order (default) or chose as-is (without sorting) by the function **cusolverDnXgesvdjSetSortEig**. If the user packs several tiny matrices into diagonal blocks of one matrix, non-sorting option can separate singular values of those tiny matrices.

gesvdjBatched cannot report residual and executed sweeps by function **cusolverDnXgesvdjGetResidual** and **cusolverDnXgesvdjGetSweeps**. Any call of the above two returns **CUSOLVER_STATUS_NOT_SUPPORTED**. The user needs to compute residual explicitly.

The user has to provide working space pointed by input parameter **work**. The input parameter **lwork** is the size of the working space, and it is returned by **gesvdjBatched_bufferSize()**.

The output parameter **info** is an integer array of size **batchSize**. If the function returns **CUSOLVER_STATUS_INVALID_VALUE**, the first element **info[0] = -i** (less than zero) indicates **i-th** parameter is wrong (not counting handle). Otherwise, if **info[i] = min(m,n)+1**, **gesvdjBatched** does not converge on **i-th** matrix under given tolerance and maximum sweeps.

Appendix G.3 provides a simple example of **gesvdjBatched**.

API of syevjBatched

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
jobz	host	input	specifies options to either compute singular value only or singular vectors as well: jobz = CUSOLVER_EIG_MODE_NOVECTOR : Compute singular values only; jobz = CUSOLVER_EIG_MODE_VECTOR : Compute singular values and singular vectors.
m	host	input	number of rows of matrix A_j . m is no greater than 32.
n	host	input	number of columns of matrix A_j . n is no greater than 32.
A	device	in/out	<type> array of dimension lda * n * batchSize with lda is not less than max(1,n) . on Exit: the contents of A_j are destroyed.
lda	host	input	leading dimension of two-dimensional array used to store matrix A_j .

S	device	output	a real array of dimension $\min(m,n) * \text{batchSize}$. It stores the singular values of A_j in descending order or non-sorting order.
U	device	output	<type> array of dimension $\text{ldu} * m * \text{batchSize}$. U_j contains the left singular vectors of A_j .
ldu	host	input	leading dimension of two-dimensional array used to store matrix U_j . ldu is not less than $\max(1,m)$.
V	device	output	<type> array of dimension $\text{ldv} * n * \text{batchSize}$. V_j contains the right singular vectors of A_j .
ldv	host	input	leading dimension of two-dimensional array used to store matrix V_j . ldv is not less than $\max(1,n)$.
work	device	in/out	<type> array of size lwork , working space.
lwork	host	input	size of work , returned by <code>gesvdjBatched_bufferSize</code> .
info	device	output	an integer array of dimension batchSize . If <code>CUSOLVER_STATUS_INVALID_VALUE</code> is returned, $\text{info}[0] = -i$ (less than zero) indicates i -th parameter is wrong (not counting handle). Otherwise, if $\text{info}[i] = 0$, the operation is successful. if $\text{info}[i] = \min(m,n)+1$, <code>gesvdjBatched</code> dose not converge on i -th matrix under given tolerance and maximum sweeps.
params	host	in/out	structure filled with parameters of Jacobi algorithm.
batchSize	host	input	number of matrices.

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed ($m,n < 0$ or $\text{lda} < \max(1,m)$ or $\text{ldu} < \max(1,m)$ or $\text{ldv} < \max(1,n)$ or jobz is not <code>CUSOLVER_EIG_MODE_NOVECTOR</code> or <code>CUSOLVER_EIG_MODE_VECTOR</code> , or $\text{batchSize} < 0$).
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

2.4.3.9. cusolverDn<t>gesvdaStridedBatched()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverDnSgesvdaStridedBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int rank,
    int m,
    int n,
    const float *A,
    int lda,
    long long int strideA,
    const float *S,
    long long int strideS,
    const float *U,
    int ldu,
    long long int strideU,
    const float *V,
    int ldv,
    long long int strideV,
    int *lwork,
    int batchSize);
```

```
cusolverStatus_t
cusolverDnDgesvdaStridedBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int rank,
    int m,
    int n,
    const double *A,
    int lda,
    long long int strideA,
    const double *S,
    long long int strideS,
    const double *U,
    int ldu,
    long long int strideU,
    const double *V,
    int ldv,
    long long int strideV,
    int *lwork,
    int batchSize);
```

```
cusolverStatus_t
cusolverDnCgesvdaStridedBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int rank,
    int m,
    int n,
    const cuComplex *A,
    int lda,
    long long int strideA,
    const float *S,
    long long int strideS,
    const cuComplex *U,
    int ldu,
    long long int strideU,
    const cuComplex *V,
    int ldv,
    long long int strideV,
    int *lwork,
    int batchSize);
```

```
cusolverStatus_t
cusolverDnZgesvdaStridedBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int rank,
    int m,
    int n,
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSgesvdaStridedBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int rank,
    int m,
    int n,
    const float *A,
    int lda,
    long long int strideA,
    float *S,
    long long int strideS,
    float *U,
    int ldu,
    long long int strideU,
    float *V,
    int ldv,
    long long int strideV,
    float *work,
    int lwork,
    int *info,
    double *h_R_nrmF,
    int batchSize);

cusolverStatus_t
cusolverDnDgesvdaStridedBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int rank,
    int m,
    int n,
    const double *A,
    int lda,
    long long int strideA,
    double *S,
    long long int strideS,
    double *U,
    int ldu,
    long long int strideU,
    double *V,
    int ldv,
    long long int strideV,
    double *work,
    int lwork,
    int *info,
    double *h_R_nrmF,
    int batchSize);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCgesvdaStridedBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int rank,
    int m,
    int n,
    const cuComplex *A,
    int lda,
    long long int strideA,
    float *S,
    long long int strideS,
    cuComplex *U,
    int ldu,
    long long int strideU,
    cuComplex *V,
    int ldv,
    long long int strideV,
    cuComplex *work,
    int lwork,
    int *info,
    double *h_R_nrmF,
    int batchSize);

cusolverStatus_t
cusolverDnZgesvdaStridedBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    int rank,
    int m,
    int n,
    const cuDoubleComplex *A,
    int lda,
    long long int strideA,
    double *S,
    long long int strideS,
    cuDoubleComplex *U,
    int ldu,
    long long int strideU,
    cuDoubleComplex *V,
    int ldv,
    long long int strideV,
    cuDoubleComplex *work,
    int lwork,
    int *info,
    double *h_R_nrmF,
    int batchSize);
```

This function **gesvda** (**a** stands for approximate) approximates the singular value decomposition of a tall skinny $m \times n$ matrix **A** and corresponding the left and right singular vectors. The economy form of SVD is written by

$$A = U * \Sigma * V^H$$

where Σ is an $n \times n$ matrix. U is an $m \times n$ unitary matrix, and V is an $n \times n$ unitary matrix. The diagonal elements of Σ are the singular values of A ; they are real and non-negative, and are returned in descending order. U and V are the left and right singular vectors of A .

gesvda computes eigenvalues of A^*T^*A to approximate singular values and singular vectors. It generates matrices U and V and transforms the matrix A to the following form

$$U^H * A * V = S + E$$

where S is diagonal and E depends on rounding errors. To certain conditions, U , V and S approximate singular values and singular vectors up to machine zero of single precision. In general, V is unitary, S is more accurate than U . If singular value is far from zero, then left singular vector U is accurate. In other words, the accuracy of singular values and left singular vectors depend on the distance between singular value and zero.

The input parameter **rank** decides the number of singular values and singular vectors are computed in parameter S , U and V .

The output parameter **h_RnrmF** computes Frobenius norm of residual.

$$A - U * S * V^H$$

if the parameter **rank** is equal n . Otherwise, **h_RnrmF** reports

$$\|U * S * V^H\| - \|S\|$$

in Frobenius norm sense. That is, how far U is from unitary.

gesvdaStridedBatched performs **gesvda** on each matrix. It requires that all matrices are of the same size m, n and are packed in contiguous way,

$$A = (A_0 \ A_1 \ \dots)$$

Each matrix is column-major with leading dimension **lda**, so the formula for random access is $A_k(i, j) = A[i + lda*j + strideA*k]$. Similarly, the formula for random access of S is $S_k(j) = S[j + StrideS*k]$, the formula for random access of U is $U_k(i, j) = U[i + ldu*j + strideU*k]$ and the formula for random access of V is $V_k(i, j) = V[i + ldv*j + strideV*k]$.

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is the size of the working space, and it is returned by **gesvdaStridedBatched_bufferSize()**.

The output parameter **info** is an integer array of size **batchSize**. If the function returns **CUSOLVER_STATUS_INVALID_VALUE**, the first element **info[0] = -i** (less than zero) indicates **i-th** parameter is wrong (not counting handle). Otherwise, if **info[i] = min(m, n) + 1**, **gesvdaStridedBatched** does not converge on **i-th** matrix under given tolerance.

Appendix G.4 provides a simple example of **gesvda**.

Remark 1: the routine returns V , not V^H . This is different from **gesvd**.

Remark 2: if the user is confident on the accuracy of singular values and singular vectors, for example, certain conditions hold (required singular value is far from zero), then the performance can be improved by passing null pointer to `h_RnormF`, i.e. no computation of residual norm.

API of gesvda

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
jobz	host	input	specifies options to either compute singular value only or singular vectors as well: <code>jobz = CUSOLVER_EIG_MODE_NOVECTOR</code> : Compute singular values only; <code>jobz = CUSOLVER_EIG_MODE_VECTOR</code> : Compute singular values and singular vectors.
rank	host	input	number of singular values (from largest to smallest).
m	host	input	number of rows of matrix A_j .
n	host	input	number of columns of matrix A_j .
A	device	input	<type> array of dimension <code>strideA * batchSize</code> with <code>lda</code> is not less than <code>max(1,m)</code> . A_j is of dimension <code>m * n</code> .
lda	host	input	leading dimension of two-dimensional array used to store matrix A_j .
strideA	host	input	value of type long long int that gives the address offset between <code>A[i]</code> and <code>A[i+1]</code> . <code>strideA</code> is not less than <code>lda*n</code> .
S	device	output	a real array of dimension <code>strideS*batchSize</code> . It stores the singular values of A_j in descending order. S_j is of dimension <code>rank * 1</code> .
strideS	host	input	value of type long long int that gives the address offset between <code>S[i]</code> and <code>S[i+1]</code> . <code>strideS</code> is not less than <code>rank</code> .
U	device	output	<type> array of dimension <code>strideU * batchSize</code> . U_j contains the left singular vectors of A_j . U_j is of dimension <code>m * rank</code> .
ldu	host	input	leading dimension of two-dimensional array used to store matrix U_j . <code>ldu</code> is not less than <code>max(1,m)</code> .
strideU	host	input	value of type long long int that gives the address offset between <code>U[i]</code> and <code>U[i+1]</code> . <code>strideU</code> is not less than <code>ldu*rank</code> .
V	device	output	<type> array of dimension <code>strideV * batchSize</code> . V_j contains the right singular vectors of A_j . V_j is of dimension <code>n * rank</code> .

ldv	host	input	leading dimension of two-dimensional array used to store matrix v_j . ldv is not less than $\max(1, n)$.
strideV	host	input	value of type long long int that gives the address offset between $v[i]$ and $v[i+1]$. strideV is not less than $ldv * rank$.
work	device	in/out	<type> array of size lwork , working space.
lwork	host	input	size of work , returned by <code>gesvdaStridedBatched_bufferSize</code> .
info	device	output	an integer array of dimension batchSize . If <code>CUSOLVER_STATUS_INVALID_VALUE</code> is returned, <code>info[0] = -i</code> (less than zero) indicates <i>i</i> -th parameter is wrong (not counting handle). Otherwise, if <code>info[i] = 0</code> , the operation is successful. if <code>info[i] = min(m,n)+1</code> , <code>gesvdaStridedBatched</code> dose not converge on <i>i</i> -th matrix.
h_RnormF	host	output	<double> array of size batchSize . <code>h_RnormF[i]</code> is norm of residual of <i>i</i> -th matrix.
batchSize	host	input	number of matrices. batchSize is not less than 1.

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed ($m, n < 0$ OR $lda < \max(1, m)$ OR $ldu < \max(1, m)$ OR $ldv < \max(1, n)$ OR $strideA < lda * n$ OR $strideS < rank$ OR $strideU < ldu * rank$ OR $strideV < ldv * rank$ OR $batchSize < 1$ OR <code>jobz</code> is not <code>CUSOLVER_EIG_MODE_NOVECTOR</code> OR <code>CUSOLVER_EIG_MODE_VECTOR</code>).
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

2.4.3.10. cusolverDn<t>syevd()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverDnSsyevd_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const float *A,
    int lda,
    const float *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnDsyevd_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const double *A,
    int lda,
    const double *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnCsyevd_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const cuComplex *A,
    int lda,
    const float *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnZsyevd_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const cuDoubleComplex *A,
    int lda,
    const double *W,
    int *lwork);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSsyevd(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    float *A,
    int lda,
    float *W,
    float *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnDsyevd(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    double *A,
    int lda,
    double *W,
    double *work,
    int lwork,
    int *devInfo);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCsyevd(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    cuComplex *A,
    int lda,
    float *W,
    cuComplex *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnZsyevd(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    cuDoubleComplex *A,
    int lda,
    double *W,
    cuDoubleComplex *work,
    int lwork,
    int *devInfo);
```

This function computes eigenvalues and eigenvectors of a symmetric (Hermitian) $n \times n$ matrix **A**. The standard symmetric eigenvalue problem is

$$A * V = V * \Lambda$$

where $\mathbf{\Lambda}$ is a real $n \times n$ diagonal matrix. \mathbf{V} is an $n \times n$ unitary matrix. The diagonal elements of $\mathbf{\Lambda}$ are the eigenvalues of \mathbf{A} in ascending order.

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is size of the working space, and it is returned by **syevd_bufferSize()**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle). If **devInfo** = **i** (greater than zero), **i** off-diagonal elements of an intermediate tridiagonal form did not converge to zero.

if **jobz** = CUSOLVER_EIG_MODE_VECTOR, **A** contains the orthonormal eigenvectors of the matrix **A**. The eigenvectors are computed by a divide and conquer algorithm.

Appendix F.1 provides a simple example of **syevd**.

API of syevd

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
jobz	host	input	specifies options to either compute eigenvalue only or compute eigen-pair: jobz = CUSOLVER_EIG_MODE_NOVECTOR : Compute eigenvalues only; jobz = CUSOLVER_EIG_MODE_VECTOR : Compute eigenvalues and eigenvectors.
uplo	host	input	specifies which part of A is stored. uplo = CUBLAS_FILL_MODE_LOWER: Lower triangle of A is stored. uplo = CUBLAS_FILL_MODE_UPPER: Upper triangle of A is stored.
n	host	input	number of rows (or columns) of matrix A .
A	device	in/out	<type> array of dimension lda * n with lda is not less than max(1,n). If uplo = CUBLAS_FILL_MODE_UPPER, the leading n-by-n upper triangular part of A contains the upper triangular part of the matrix A . If uplo = CUBLAS_FILL_MODE_LOWER, the leading n-by-n lower triangular part of A contains the lower triangular part of the matrix A . On exit, if jobz = CUSOLVER_EIG_MODE_VECTOR, and devInfo = 0, A contains the orthonormal eigenvectors of the matrix A . If jobz = CUSOLVER_EIG_MODE_NOVECTOR, the contents of A are destroyed.
lda	host	input	leading dimension of two-dimensional array used to store matrix A .
w	device	output	a real array of dimension n. The eigenvalue values of A , in ascending order ie, sorted so that w(i) <= w(i+1).
work	device	in/out	working space, <type> array of size lwork.

Lwork	host	input	size of work , returned by syevd_bufferSize .
devInfo	device	output	if devInfo = 0, the operation is successful. if devInfo = -i, the i-th parameter is wrong (not counting handle). if devInfo = i (> 0), devInfo indicates i off-diagonal elements of an intermediate tridiagonal form did not converge to zero;

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($n < 0$, or $lda < \max(1, n)$, or jobz is not CUSOLVER_EIG_MODE_NOVECTOR or CUSOLVER_EIG_MODE_VECTOR , or uplo is not CUBLAS_FILL_MODE_LOWER or CUBLAS_FILL_MODE_UPPER).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.3.11. cusolverDn<t>syevdx()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverDnSsyevdx_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    const float *A,
    int lda,
    float vl,
    float vu,
    int il,
    int iu,
    int *h_meig,
    const float *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnDsyevdx_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    const double *A,
    int lda,
    double vl,
    double vu,
    int il,
    int iu,
    int *h_meig,
    const double *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnCsyevdx_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    const cuComplex *A,
    int lda,
    float vl,
    float vu,
    int il,
    int iu,
    int *h_meig,
    const float *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnZsyevdx_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    const cuDoubleComplex *A,
    int lda,
    double vl,
    double vu,
    int il,
    int iu,
    int *h_meig,
    const double *W,
    int *lwork);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSsyevdx(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    float *A,
    int lda,
    float vl,
    float vu,
    int il,
    int iu,
    int *h_meig,
    float *W,
    float *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnDsyevdx(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    double *A,
    int lda,
    double vl,
    double vu,
    int il,
    int iu,
    int *h_meig,
    double *W,
    double *work,
    int lwork,
    int *devInfo);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCheevdx(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    cuComplex *A,
    int lda,
    float vl,
    float vu,
    int il,
    int iu,
    int *h_meig,
    float *W,
    cuComplex *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnZheevdx(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    cuDoubleComplex *A,
    int lda,
    double vl,
    double vu,
    int il,
    int iu,
    int *h_meig,
    double *W,
    cuDoubleComplex *work,
    int lwork,
    int *devInfo);
```

This function computes all or selection of the eigenvalues and optionally eigenvectors of a symmetric (Hermitian) $n \times n$ matrix **A**. The standard symmetric eigenvalue problem is

$$A * V = V * \Lambda$$

where Λ is a real $n \times h_meig$ diagonal matrix. **V** is an $n \times h_meig$ unitary matrix. **h_meig** is the number of eigenvalues/eigenvectors computed by the routine, **h_meig** is equal to **n** when the whole spectrum (e.g., **range** = **CUSOLVER_EIG_RANGE_ALL**) is requested. The diagonal elements of Λ are the eigenvalues of **A** in ascending order.

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is size of the working space, and it is returned by **syevdx_bufferSize()**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle). If **devInfo** = **i** (greater than zero), **i** off-diagonal elements of an intermediate tridiagonal form did not converge to zero.

if `jobz = CUSOLVER_EIG_MODE_VECTOR`, **A** contains the orthonormal eigenvectors of the matrix **A**. The eigenvectors are computed by a divide and conquer algorithm.

Appendix F.1 provides a simple example of `syevdx`.

API of `syevdx`

parameter	Memory	In/out	Meaning
<code>handle</code>	host	input	handle to the cuSolverDN library context.
<code>jobz</code>	host	input	specifies options to either compute eigenvalue only or compute eigen-pair: <code>jobz = CUSOLVER_EIG_MODE_NOVECTOR</code> : Compute eigenvalues only; <code>jobz = CUSOLVER_EIG_MODE_VECTOR</code> : Compute eigenvalues and eigenvectors.
<code>range</code>	host	input	specifies options to which selection of eigenvalues and optionally eigenvectors that need to be computed: <code>range = CUSOLVER_EIG_RANGE_ALL</code> : all eigenvalues/eigenvectors will be found, will becomes the classical <code>syevd/heeve</code> routine; <code>range = CUSOLVER_EIG_RANGE_V</code> : all eigenvalues/eigenvectors in the half-open interval $(vl, vu]$ will be found; <code>range = CUSOLVER_EIG_RANGE_I</code> : the <i>il</i> -th through <i>iu</i> -th eigenvalues/eigenvectors will be found;
<code>uplo</code>	host	input	specifies which part of A is stored. <code>uplo = CUBLAS_FILL_MODE_LOWER</code> : Lower triangle of A is stored. <code>uplo = CUBLAS_FILL_MODE_UPPER</code> : Upper triangle of A is stored.
<code>n</code>	host	input	number of rows (or columns) of matrix A .
A	device	in/out	<type> array of dimension <code>lda * n</code> with <code>lda</code> is not less than <code>max(1, n)</code> . If <code>uplo = CUBLAS_FILL_MODE_UPPER</code> , the leading <i>n</i> -by- <i>n</i> upper triangular part of A contains the upper triangular part of the matrix A . If <code>uplo = CUBLAS_FILL_MODE_LOWER</code> , the leading <i>n</i> -by- <i>n</i> lower triangular part of A contains the lower triangular part of the matrix A . On exit, if <code>jobz = CUSOLVER_EIG_MODE_VECTOR</code> , and <code>devInfo = 0</code> , A contains the orthonormal eigenvectors of the matrix A . If <code>jobz = CUSOLVER_EIG_MODE_NOVECTOR</code> , the contents of A are destroyed.
<code>lda</code>	host	input	leading dimension of two-dimensional array used to store matrix A . <code>lda</code> is not less than <code>max(1, n)</code> .
<code>vl, vu</code>	host	input	real values float or double for (C, S) or (Z, D) precision respectively. If <code>range = CUSOLVER_EIG_RANGE_V</code> , the lower and upper bounds of the

			interval to be searched for eigenvalues. $vl > vu$. Not referenced if <code>range = CUSOLVER_EIG_RANGE_ALL</code> or <code>range = CUSOLVER_EIG_RANGE_I</code> . Note that, if eigenvalues are very close to each other, it is well known that two different eigenvalues routines might find slightly different number of eigenvalues inside the same interval. This is due to the fact that different eigenvalue algorithms, or even same algorithm but different run might find eigenvalues within some rounding error close to the machine precision. Thus, if the user want to be sure not to miss any eigenvalue within the interval bound, we suggest that, the user subtract/add epsilon (machine precision) to the interval bound such as ($vl=vl-eps$, $vu=vu+eps$). this suggestion is valid for any selective routine from cuSolver or LAPACK.
<code>il,iu</code>	<code>host</code>	<code>input</code>	integer. If <code>range = CUSOLVER_EIG_RANGE_I</code> , the indices (in ascending order) of the smallest and largest eigenvalues to be returned. $1 \leq il \leq iu \leq n$, if $n > 0$; $il = 1$ and $iu = 0$ if $n = 0$. Not referenced if <code>range = CUSOLVER_EIG_RANGE_ALL</code> or <code>range = CUSOLVER_EIG_RANGE_V</code> .
<code>h_meig</code>	<code>host</code>	<code>output</code>	integer. The total number of eigenvalues found. $0 \leq h_meig \leq n$. If <code>range = CUSOLVER_EIG_RANGE_ALL</code> , $h_meig = n$, and if <code>range = CUSOLVER_EIG_RANGE_I</code> , $h_meig = iu-il+1$.
<code>w</code>	<code>device</code>	<code>output</code>	a real array of dimension n . The eigenvalue values of A , in ascending order ie, sorted so that $w(i) \leq w(i+1)$.
<code>work</code>	<code>device</code>	<code>in/out</code>	working space, <type> array of size <code>lwork</code> .
<code>lwork</code>	<code>host</code>	<code>input</code>	size of <code>work</code> , returned by <code>syevdx_bufferSize</code> .
<code>devInfo</code>	<code>device</code>	<code>output</code>	if <code>devInfo = 0</code> , the operation is successful. if <code>devInfo = -i</code> , the i -th parameter is wrong (not counting handle). if <code>devInfo = i</code> (> 0), <code>devInfo</code> indicates i off-diagonal elements of an intermediate tridiagonal form did not converge to zero;

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.

<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed (<code>n<0</code> , or <code>lda<max(1,n)</code> , or <code>jobz</code> is not <code>CUSOLVER_EIG_MODE_NOVECTOR</code> or <code>CUSOLVER_EIG_MODE_VECTOR</code> , or <code>range</code> is not <code>CUSOLVER_EIG_RANGE_ALL</code> or <code>CUSOLVER_EIG_RANGE_V</code> or <code>CUSOLVER_EIG_RANGE_I</code> , or <code>uplo</code> is not <code>CUBLAS_FILL_MODE_LOWER</code> or <code>CUBLAS_FILL_MODE_UPPER</code>).
<code>CUSOLVER_STATUS_ARCH_MISMATCH</code>	the device only supports compute capability 2.0 and above.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

2.4.3.12. cusolverDn<t>sygvd()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverDnSsygvd_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const float *A,
    int lda,
    const float *B,
    int ldb,
    const float *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnDsygvd_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const double *A,
    int lda,
    const double *B,
    int ldb,
    const double *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnCsygvd_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const cuComplex *A,
    int lda,
    const cuComplex *B,
    int ldb,
    const float *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnZsygvd_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *B,
    int ldb,
    const double *W,
    int *lwork);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSsygvd(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    float *A,
    int lda,
    float *B,
    int ldb,
    float *W,
    float *work,
    int lwork,
    int *devInfo);
```

```
cusolverStatus_t
cusolverDnDsygvd(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    double *A,
    int lda,
    double *B,
    int ldb,
    double *W,
    double *work,
    int lwork,
    int *devInfo);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnChegvd(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    cuComplex *A,
    int lda,
    cuComplex *B,
    int ldb,
    float *W,
    cuComplex *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnZhegvd(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    cuDoubleComplex *A,
    int lda,
    cuDoubleComplex *B,
    int ldb,
    double *W,
    cuDoubleComplex *work,
    int lwork,
    int *devInfo);
```

This function computes eigenvalues and eigenvectors of a symmetric (Hermitian) $\mathbf{n} \times \mathbf{n}$ matrix-pair (\mathbf{A}, \mathbf{B}) . The generalized symmetric-definite eigenvalue problem is

$$\text{eig}(\mathbf{A}, \mathbf{B}) = \begin{cases} \mathbf{A}^* \mathbf{V} = \mathbf{B}^* \mathbf{V}^* \boldsymbol{\Lambda} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_1} \\ \mathbf{A}^* \mathbf{B}^* \mathbf{V} = \mathbf{V}^* \boldsymbol{\Lambda} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_2} \\ \mathbf{B}^* \mathbf{A}^* \mathbf{V} = \mathbf{V}^* \boldsymbol{\Lambda} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_3} \end{cases}$$

where the matrix \mathbf{B} is positive definite. $\boldsymbol{\Lambda}$ is a real $\mathbf{n} \times \mathbf{n}$ diagonal matrix. The diagonal elements of $\boldsymbol{\Lambda}$ are the eigenvalues of (\mathbf{A}, \mathbf{B}) in ascending order. \mathbf{V} is an $\mathbf{n} \times \mathbf{n}$ orthogonal matrix. The eigenvectors are normalized as follows:

$$\begin{cases} \mathbf{V}^H \mathbf{B}^* \mathbf{V} = \mathbf{I} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_1}, \text{CUSOLVER_EIG_TYPE_2} \\ \mathbf{V}^H \text{inv}(\mathbf{B})^* \mathbf{V} = \mathbf{I} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_3} \end{cases}$$

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is size of the working space, and it is returned by **sygvdd_bufferSize()**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle). If **devInfo** = **i** ($i > 0$ and $i \leq n$) and **jobz** = CUSOLVER_EIG_MODE_NOVECTOR, **i** off-diagonal elements of an intermediate tridiagonal form did not converge to zero. If **devInfo** = **N + i** ($i > 0$), then the

leading minor of order **i** of **B** is not positive definite. The factorization of **B** could not be completed and no eigenvalues or eigenvectors were computed.

if **jobz** = CUSOLVER_EIG_MODE_VECTOR, **A** contains the orthogonal eigenvectors of the matrix **A**. The eigenvectors are computed by divide and conquer algorithm.

Appendix F.2 provides a simple example of **sygvd**.

API of **sygvd**

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
itype	host	input	Specifies the problem type to be solved: itype =CUSOLVER_EIG_TYPE_1: $A*x = (\text{lambda})*B*x$. itype =CUSOLVER_EIG_TYPE_2: $A*B*x = (\text{lambda})*x$. itype =CUSOLVER_EIG_TYPE_3: $B*A*x = (\text{lambda})*x$.
jobz	host	input	specifies options to either compute eigenvalue only or compute eigen-pair: jobz = CUSOLVER_EIG_MODE_NOVECTOR : Compute eigenvalues only; jobz = CUSOLVER_EIG_MODE_VECTOR : Compute eigenvalues and eigenvectors.
uplo	host	input	specifies which part of A and B are stored. uplo = CUBLAS_FILL_MODE_LOWER: Lower triangle of A and B are stored. uplo = CUBLAS_FILL_MODE_UPPER: Upper triangle of A and B are stored.
n	host	input	number of rows (or columns) of matrix A and B .
A	device	in/out	<type> array of dimension lda * n with lda is not less than $\max(1, n)$. If uplo = CUBLAS_FILL_MODE_UPPER, the leading n-by-n upper triangular part of A contains the upper triangular part of the matrix A . If uplo = CUBLAS_FILL_MODE_LOWER, the leading n-by-n lower triangular part of A contains the lower triangular part of the matrix A . On exit, if jobz = CUSOLVER_EIG_MODE_VECTOR, and devInfo = 0, A contains the orthonormal eigenvectors of the matrix A . If jobz = CUSOLVER_EIG_MODE_NOVECTOR, the contents of A are destroyed.
lda	host	input	leading dimension of two-dimensional array used to store matrix A . lda is not less than $\max(1, n)$.
B	device	in/out	<type> array of dimension ldb * n . If uplo = CUBLAS_FILL_MODE_UPPER, the leading n-by-n upper triangular part of B contains the upper triangular part of the matrix B . If uplo = CUBLAS_FILL_MODE_LOWER, the leading

			n-by-n lower triangular part of B contains the lower triangular part of the matrix B . On exit, if devInfo is less than n , B is overwritten by triangular factor U or L from the Cholesky factorization of B .
ldb	host	input	leading dimension of two-dimensional array used to store matrix B . ldb is not less than $\max(1, n)$.
W	device	output	a real array of dimension n . The eigenvalue values of A , sorted so that $W(i) \geq W(i+1)$.
work	device	in/out	working space, <type> array of size lwork .
lwork	host	input	size of work , returned by sygvd_bufferSize .
devInfo	device	output	if devInfo = 0, the operation is successful. if devInfo = -i, the i-th parameter is wrong (not counting handle). if devInfo = i (> 0), devInfo indicates either potrf or syevd is wrong.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (n <0, or lda < $\max(1, n)$, or ldb < $\max(1, n)$, or itype is not 1, 2 or 3, or jobz is not 'N' or 'V', or uplo is not CUBLAS_FILL_MODE_LOWER or CUBLAS_FILL_MODE_UPPER).
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.3.13. cusolverDn<t>sygvdx()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverDnSsygvdx_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    const float *A,
    int lda,
    const float *B,
    int ldb,
    float vl,
    float vu,
    int il,
    int iu,
    int *h_meig,
    const float *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnDsygvdx_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    const double *A,
    int lda,
    const double *B,
    int ldb,
    double vl,
    double vu,
    int il,
    int iu,
    int *h_meig,
    const double *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnCsygvdx_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    const cuComplex *A,
    int lda,
    const cuComplex *B,
    int ldb,
    float vl,
    float vu,
    int il,
    int iu,
    int *h_meig,
    const float *W,
    int *lwork);
```

```
cusolverStatus_t
cusolverDnZsygvdx_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    const cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *B,
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSsygvdx(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    float *A,
    int lda,
    float *B,
    int ldb,
    float vl,
    float vu,
    int il,
    int iu,
    int *h_meig,
    float *W,
    float *work,
    int lwork,
    int *devInfo);
```

```
cusolverStatus_t
cusolverDnDsygvdx(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    double *A,
    int lda,
    double *B,
    int ldb,
    double vl,
    double vu,
    int il,
    int iu,
    int *h_meig,
    double *W,
    double *work,
    int lwork,
    int *devInfo);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnChegvdx(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    cuComplex *A,
    int lda,
    cuComplex *B,
    int ldb,
    float vl,
    float vu,
    int il,
    int iu,
    int *h_meig,
    float *W,
    cuComplex *work,
    int lwork,
    int *devInfo);

cusolverStatus_t
cusolverDnZhegvdx(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cusolverEigRange_t range,
    cublasFillMode_t uplo,
    int n,
    cuDoubleComplex *A,
    int lda,
    cuDoubleComplex *B,
    int ldb,
    double vl,
    double vu,
    int il,
    int iu,
    int *h_meig,
    double *W,
    cuDoubleComplex *work,
    int lwork,
    int *devInfo);
```

This function computes all or selection of the eigenvalues and optionally eigenvectors of a symmetric (Hermitian) $n \times n$ matrix-pair (\mathbf{A}, \mathbf{B}) . The generalized symmetric-definite eigenvalue problem is

$$\text{eig}(\mathbf{A}, \mathbf{B}) = \begin{cases} \mathbf{A}^* \mathbf{V} = \mathbf{B}^* \mathbf{V}^* \boldsymbol{\Lambda} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_1} \\ \mathbf{A}^* \mathbf{B}^* \mathbf{V} = \mathbf{V}^* \boldsymbol{\Lambda} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_2} \\ \mathbf{B}^* \mathbf{A}^* \mathbf{V} = \mathbf{V}^* \boldsymbol{\Lambda} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_3} \end{cases}$$

where the matrix \mathbf{B} is positive definite. $\boldsymbol{\Lambda}$ is a real $n \times h_meig$ diagonal matrix. The diagonal elements of $\boldsymbol{\Lambda}$ are the eigenvalues of (\mathbf{A}, \mathbf{B}) in ascending order. \mathbf{V} is an $n \times h_meig$ orthogonal matrix. h_meig is the number of eigenvalues/eigenvectors

computed by the routine, **h_meig** is equal to **n** when the whole spectrum (e.g., **range** = **CUSOLVER_EIG_RANGE_ALL**) is requested. The eigenvectors are normalized as follows:

$$\begin{cases} V^H * B * V = I & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_1}, \text{CUSOLVER_EIG_TYPE_2} \\ V^H * \text{inv}(B) * V = I & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_3} \end{cases}$$

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is size of the working space, and it is returned by **sygvdx_bufferSize()**.

If output parameter **devInfo** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle). If **devInfo** = **i** ($i > 0$ and $i \leq n$) and **jobz** = **CUSOLVER_EIG_MODE_NOVECTOR**, **i** off-diagonal elements of an intermediate tridiagonal form did not converge to zero. If **devInfo** = **n + i** ($i > 0$), then the leading minor of order **i** of **B** is not positive definite. The factorization of **B** could not be completed and no eigenvalues or eigenvectors were computed.

if **jobz** = **CUSOLVER_EIG_MODE_VECTOR**, **A** contains the orthogonal eigenvectors of the matrix **A**. The eigenvectors are computed by divide and conquer algorithm.

Appendix F.2 provides a simple example of **sygvdx**.

API of sygvdx

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
itype	host	input	Specifies the problem type to be solved: itype = CUSOLVER_EIG_TYPE_1 : $A*x = (\text{lambda})*B*x$. itype = CUSOLVER_EIG_TYPE_2 : $A*B*x = (\text{lambda})*x$. itype = CUSOLVER_EIG_TYPE_3 : $B*A*x = (\text{lambda})*x$.
jobz	host	input	specifies options to either compute eigenvalue only or compute eigen-pair: jobz = CUSOLVER_EIG_MODE_NOVECTOR : Compute eigenvalues only; jobz = CUSOLVER_EIG_MODE_VECTOR : Compute eigenvalues and eigenvectors.
range	host	input	specifies options to which selection of eigenvalues and optionally eigenvectors that need to be computed: range = CUSOLVER_EIG_RANGE_ALL : all eigenvalues/eigenvectors will be found, will becomes the classical syevd/heevd routine; range = CUSOLVER_EIG_RANGE_V : all eigenvalues/eigenvectors in the half-open interval $(vl, vu]$ will be found; range = CUSOLVER_EIG_RANGE_I : the il -th through iu -th eigenvalues/eigenvectors will be found;
uplo	host	input	specifies which part of A and B are stored. uplo = CUBLAS_FILL_MODE_LOWER :

			Lower triangle of A and B are stored. uplo = CUBLAS_FILL_MODE_UPPER : Upper triangle of A and B are stored.
n	host	input	number of rows (or columns) of matrix A and B .
A	device	in/out	<type> array of dimension lda * n with lda is not less than max(1, n) . If uplo = CUBLAS_FILL_MODE_UPPER , the leading n-by-n upper triangular part of A contains the upper triangular part of the matrix A . If uplo = CUBLAS_FILL_MODE_LOWER , the leading n-by-n lower triangular part of A contains the lower triangular part of the matrix A . On exit, if jobz = CUSOLVER_EIG_MODE_VECTOR , and devInfo = 0, A contains the orthonormal eigenvectors of the matrix A . If jobz = CUSOLVER_EIG_MODE_NOVECTOR , the contents of A are destroyed.
lda	host	input	leading dimension of two-dimensional array used to store matrix A . lda is not less than max(1, n) .
B	device	in/out	<type> array of dimension ldb * n . If uplo = CUBLAS_FILL_MODE_UPPER , the leading n-by-n upper triangular part of B contains the upper triangular part of the matrix B . If uplo = CUBLAS_FILL_MODE_LOWER , the leading n-by-n lower triangular part of B contains the lower triangular part of the matrix B . On exit, if devInfo is less than n , B is overwritten by triangular factor U or L from the Cholesky factorization of B .
ldb	host	input	leading dimension of two-dimensional array used to store matrix B . ldb is not less than max(1, n) .
vl, vu	host	input	real values float or double for (C, S) or (Z, D) precision respectively. If range = CUSOLVER_EIG_RANGE_V , the lower and upper bounds of the interval to be searched for eigenvalues. vl > vu . Not referenced if range = CUSOLVER_EIG_RANGE_ALL or range = CUSOLVER_EIG_RANGE_I . Note that, if eigenvalues are very close to each other, it is well known that two different eigenvalues routines might find slightly different number of eigenvalues inside the same interval. This is due to the fact that different eigenvalue algorithms, or even same algorithm but different run might find eigenvalues within some rounding error close to the machine precision. Thus, if the user want to be sure not to miss any eigenvalue within the interval bound, we suggest that, the user

			subtract/add epsilon (machine precision) to the interval bound such as ($vl=vl-eps$, $vu=vu+eps$). this suggestion is valid for any selective routine from cuSolver or LAPACK.
<code>il,iu</code>	<code>host</code>	<code>input</code>	integer. If <code>range = CUSOLVER_EIG_RANGE_I</code> , the indices (in ascending order) of the smallest and largest eigenvalues to be returned. $1 \leq il \leq iu \leq n$, if $n > 0$; $il = 1$ and $iu = 0$ if $n = 0$. Not referenced if <code>range = CUSOLVER_EIG_RANGE_ALL</code> or <code>range = CUSOLVER_EIG_RANGE_V</code> .
<code>h_meig</code>	<code>host</code>	<code>output</code>	integer. The total number of eigenvalues found. $0 \leq h_meig \leq n$. If <code>range = CUSOLVER_EIG_RANGE_ALL</code> , $h_meig = n$, and if <code>range = CUSOLVER_EIG_RANGE_I</code> , $h_meig = iu-il+1$.
<code>W</code>	<code>device</code>	<code>output</code>	a real array of dimension n . The eigenvalue values of A , sorted so that $W(i) \geq W(i+1)$.
<code>work</code>	<code>device</code>	<code>in/out</code>	working space, <type> array of size <code>lwork</code> .
<code>lwork</code>	<code>host</code>	<code>input</code>	size of <code>work</code> , returned by <code>sygvdx_bufferSize</code> .
<code>devInfo</code>	<code>device</code>	<code>output</code>	if <code>devInfo = 0</code> , the operation is successful. if <code>devInfo = -i</code> , the i -th parameter is wrong (not counting handle). if <code>devInfo = i</code> (> 0), <code>devInfo</code> indicates either <code>potrf</code> or <code>syevd</code> is wrong.

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed ($n < 0$, or $lda < \max(1, n)$, or $ldb < \max(1, n)$, or <code>itype</code> is not <code>CUSOLVER_EIG_TYPE_1</code> or <code>CUSOLVER_EIG_TYPE_2</code> or <code>CUSOLVER_EIG_TYPE_3</code> or <code>jobz</code> is not <code>CUSOLVER_EIG_MODE_NOVECTOR</code> or <code>CUSOLVER_EIG_MODE_VECTORL</code> , or <code>range</code> is not <code>CUSOLVER_EIG_RANGE_ALL</code> or <code>CUSOLVER_EIG_RANGE_V</code> or <code>CUSOLVER_EIG_RANGE_I</code> , or <code>uplo</code> is not <code>CUBLAS_FILL_MODE_LOWER</code> or <code>CUBLAS_FILL_MODE_UPPER</code>).
<code>CUSOLVER_STATUS_ARCH_MISMATCH</code>	the device only supports compute capability 2.0 and above.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

2.4.3.14. cusolverDn<t>syevj()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverDnSsyevj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const float *A,
    int lda,
    const float *W,
    int *lwork,
    syevjInfo_t params);

cusolverStatus_t
cusolverDnDsyevj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const double *A,
    int lda,
    const double *W,
    int *lwork,
    syevjInfo_t params);

cusolverStatus_t
cusolverDnCsyevj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const cuComplex *A,
    int lda,
    const float *W,
    int *lwork,
    syevjInfo_t params);

cusolverStatus_t
cusolverDnZheevj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const cuDoubleComplex *A,
    int lda,
    const double *W,
    int *lwork,
    syevjInfo_t params);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSsyevj(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    float *A,
    int lda,
    float *W,
    float *work,
    int lwork,
    int *info,
    syevjInfo_t params);

cusolverStatus_t
cusolverDnDsyevj(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    double *A,
    int lda,
    double *W,
    double *work,
    int lwork,
    int *info,
    syevjInfo_t params);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCsyevj(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    cuComplex *A,
    int lda,
    float *W,
    cuComplex *work,
    int lwork,
    int *info,
    syevjInfo_t params);

cusolverStatus_t
cusolverDnZsyevj(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    cuDoubleComplex *A,
    int lda,
    double *W,
    cuDoubleComplex *work,
    int lwork,
    int *info,
    syevjInfo_t params);
```

This function computes eigenvalues and eigenvectors of a symmetric (Hermitian) $n \times n$ matrix **A**. The standard symmetric eigenvalue problem is

$$A^*Q = Q^*\Lambda$$

where Λ is a real $n \times n$ diagonal matrix. Q is an $n \times n$ unitary matrix. The diagonal elements of Λ are the eigenvalues of A in ascending order.

syevj has the same functionality as **syevd**. The difference is that **syevd** uses QR algorithm and **syevj** uses Jacobi method. The parallelism of Jacobi method gives GPU better performance on small and medium size matrices. Moreover the user can configure **syevj** to perform approximation up to certain accuracy.

How does it work?

syevj iteratively generates a sequence of unitary matrices to transform matrix A to the following form

$$V^H A V = W + E$$

where W is diagonal and E is symmetric without diagonal.

During the iterations, the Frobenius norm of E decreases monotonically. As E goes down to zero, W is the set of eigenvalues. In practice, Jacobi method stops if

$$\|E\|_F \leq \text{eps} * \|A\|_F$$

where **eps** is given tolerance.

syevj has two parameters to control the accuracy. First parameter is tolerance (**eps**). The default value is machine accuracy but The user can use function **cusolverDnXsyevjSetTolerance** to set a priori tolerance. The second parameter is maximum number of sweeps which controls number of iterations of Jacobi method. The default value is 100 but the user can use function **cusolverDnXsyevjSetMaxSweeps** to set a proper bound. The experimentis show 15 sweeps are good enough to converge to machine accuracy. **syevj** stops either tolerance is met or maximum number of sweeps is met.

Jacobi method has quadratic convergence, so the accuracy is not proportional to number of sweeps. To guarantee certain accuracy, the user should configure tolerance only.

After **syevj**, the user can query residual by function **cusolverDnXsyevjGetResidual** and number of executed sweeps by function **cusolverDnXsyevjGetSweeps**. However the user needs to be aware that residual is the Frobenius norm of E , not accuracy of individual eigenvalue, i.e.

$$\text{residual} = \|E\|_F = \|\Lambda - W\|_F$$

The same as **syevd**, the user has to provide working space pointed by input parameter **work**. The input parameter **lwork** is the size of the working space, and it is returned by **syevj_bufferSize()**.

If output parameter **info** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle). If **info** = **n+1**, **syevj** does not converge under given tolerance and maximum sweeps.

If the user sets an improper tolerance, **syevj** may not converge. For example, tolerance should not be smaller than machine accuracy.

if **jobz** = CUSOLVER_EIG_MODE_VECTOR, **A** contains the orthonormal eigenvectors **V**.

Appendix F.3 provides a simple example of **syevj**.

API of syevj

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
jobz	host	input	specifies options to either compute eigenvalue only or compute eigen-pair: jobz = CUSOLVER_EIG_MODE_NOVECTOR : Compute eigenvalues only; jobz = CUSOLVER_EIG_MODE_VECTOR : Compute eigenvalues and eigenvectors.
uplo	host	input	specifies which part of A is stored. uplo = CUBLAS_FILL_MODE_LOWER: Lower triangle of A is stored. uplo = CUBLAS_FILL_MODE_UPPER: Upper triangle of A is stored.
n	host	input	number of rows (or columns) of matrix A .
A	device	in/out	<type> array of dimension lda * n with lda is not less than max(1, n) . If uplo = CUBLAS_FILL_MODE_UPPER, the leading n-by-n upper triangular part of A contains the upper triangular part of the matrix A . If uplo = CUBLAS_FILL_MODE_LOWER, the leading n-by-n lower triangular part of A contains the lower triangular part of the matrix A . On exit, if jobz = CUSOLVER_EIG_MODE_VECTOR, and info = 0, A contains the orthonormal eigenvectors of the matrix A . If jobz = CUSOLVER_EIG_MODE_NOVECTOR, the contents of A are destroyed.
lda	host	input	leading dimension of two-dimensional array used to store matrix A .
W	device	output	a real array of dimension n . The eigenvalue values of A , in ascending order ie, sorted so that W(i) <= W(i+1) .
work	device	in/out	working space, <type> array of size lwork .
lwork	host	input	size of work , returned by syevj_bufferSize .
info	device	output	if info = 0, the operation is successful. if info = -i, the i-th parameter is wrong (not counting handle). if info = n +1, syevj dose not converge under given tolerance and maximum sweeps.
params	host	in/out	structure filled with parameters of Jacobi algorithm and results of syevj .

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($n < 0$, or $lda < \max(1, n)$, or <code>jobz</code> is not CUSOLVER_EIG_MODE_NOVECTOR or CUSOLVER_EIG_MODE_VECTOR, or <code>uplo</code> is not CUBLAS_FILL_MODE_LOWER or CUBLAS_FILL_MODE_UPPER).
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.3.15. cusolverDn<t>sygvj()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverDnSsygvj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const float *A,
    int lda,
    const float *B,
    int ldb,
    const float *W,
    int *lwork,
    syevjInfo_t params);
```

```
cusolverStatus_t
cusolverDnDsygvj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const double *A,
    int lda,
    const double *B,
    int ldb,
    const double *W,
    int *lwork,
    syevjInfo_t params);
```

```
cusolverStatus_t
cusolverDnCsygvj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const cuComplex *A,
    int lda,
    const cuComplex *B,
    int ldb,
    const float *W,
    int *lwork,
    syevjInfo_t params);
```

```
cusolverStatus_t
cusolverDnZsygvj_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const cuDoubleComplex *A,
    int lda,
    const cuDoubleComplex *B,
    int ldb,
    const double *W,
    int *lwork,
    syevjInfo_t params);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSsygvj(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    float *A,
    int lda,
    float *B,
    int ldb,
    float *W,
    float *work,
    int lwork,
    int *info,
    syevjInfo_t params);

cusolverStatus_t
cusolverDnDsygvj(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    double *A,
    int lda,
    double *B,
    int ldb,
    double *W,
    double *work,
    int lwork,
    int *info,
    syevjInfo_t params);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnChegvj(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    cuComplex *A,
    int lda,
    cuComplex *B,
    int ldb,
    float *W,
    cuComplex *work,
    int lwork,
    int *info,
    syevjInfo_t params);

cusolverStatus_t
cusolverDnZhegvj(
    cusolverDnHandle_t handle,
    cusolverEigType_t itype,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    cuDoubleComplex *A,
    int lda,
    cuDoubleComplex *B,
    int ldb,
    double *W,
    cuDoubleComplex *work,
    int lwork,
    int *info,
    syevjInfo_t params);
```

This function computes eigenvalues and eigenvectors of a symmetric (Hermitian) $\mathbf{n} \times \mathbf{n}$ matrix-pair (\mathbf{A}, \mathbf{B}) . The generalized symmetric-definite eigenvalue problem is

$$\text{eig}(\mathbf{A}, \mathbf{B}) = \begin{cases} \mathbf{A}^* \mathbf{V} = \mathbf{B}^* \mathbf{V} \mathbf{\Lambda} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_1} \\ \mathbf{A}^* \mathbf{B}^* \mathbf{V} = \mathbf{V}^* \mathbf{\Lambda} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_2} \\ \mathbf{B}^* \mathbf{A}^* \mathbf{V} = \mathbf{V}^* \mathbf{\Lambda} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_3} \end{cases}$$

where the matrix \mathbf{B} is positive definite. $\mathbf{\Lambda}$ is a real $\mathbf{n} \times \mathbf{n}$ diagonal matrix. The diagonal elements of $\mathbf{\Lambda}$ are the eigenvalues of (\mathbf{A}, \mathbf{B}) in ascending order. \mathbf{V} is an $\mathbf{n} \times \mathbf{n}$ orthogonal matrix. The eigenvectors are normalized as follows:

$$\begin{cases} \mathbf{V}^H \mathbf{B}^* \mathbf{V} = \mathbf{I} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_1}, \text{CUSOLVER_EIG_TYPE_2} \\ \mathbf{V}^H \text{inv}(\mathbf{B})^* \mathbf{V} = \mathbf{I} & \text{if } \text{itype} = \text{CUSOLVER_EIG_TYPE_3} \end{cases}$$

This function has the same functionality as **sygvd** except that **syevd** in **sygvd** is replaced by **syevj** in **sygvj**. Therefore, **sygvj** inherits properties of **syevj**, the user can use **cusolverDnXsyevjSetTolerance** and **cusolverDnXsyevjSetMaxSweeps** to configure tolerance and maximum sweeps.

However the meaning of residual is different from **syevj**. **sygvj** first computes Cholesky factorization of matrix **B**,

$$B = L * L^H$$

transform the problem to standard eigenvalue problem, then calls **syevj**.

For example, the standard eigenvalue problem of type I is

$$M * Q = Q * \Lambda$$

where matrix **M** is symmetric

$$M = L^{-1} * A * L^{-H}$$

The residual is the result of **syevj** on matrix **M**, not **A**.

The user has to provide working space which is pointed by input parameter **work**. The input parameter **lwork** is the size of the working space, and it is returned by **sygvj_bufferSize()**.

If output parameter **info** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle). If **info** = **i** ($i > 0$ and $i \leq n$), **B** is not positive definite, the factorization of **B** could not be completed and no eigenvalues or eigenvectors were computed. If **info** = **n+1**, **syevj** does not converge under given tolerance and maximum sweeps. In this case, the eigenvalues and eigenvectors are still computed because non-convergence comes from improper tolerance of maximum sweeps.

if **jobz** = CUSOLVER_EIG_MODE_VECTOR, **A** contains the orthogonal eigenvectors **V**.

Appendix F.4 provides a simple example of **sygvj**.

API of sygvj

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
itype	host	input	Specifies the problem type to be solved: itype =CUSOLVER_EIG_TYPE_1: $A*x = (\text{lambda})*B*x$. itype =CUSOLVER_EIG_TYPE_2: $A*B*x = (\text{lambda})*x$. itype =CUSOLVER_EIG_TYPE_3: $B*A*x = (\text{lambda})*x$.
jobz	host	input	specifies options to either compute eigenvalue only or compute eigen-pair: jobz = CUSOLVER_EIG_MODE_NOVECTOR : Compute eigenvalues only; jobz = CUSOLVER_EIG_MODE_VECTOR : Compute eigenvalues and eigenvectors.
uplo	host	input	specifies which part of A and B are stored. uplo = CUBLAS_FILL_MODE_LOWER: Lower triangle of A and B are stored. uplo = CUBLAS_FILL_MODE_UPPER: Upper triangle of A and B are stored.

n	host	input	number of rows (or columns) of matrix A and B .
A	device	in/out	<type> array of dimension lda * n with lda is not less than max(1, n) . If uplo = CUBLAS_FILL_MODE_UPPER , the leading n-by-n upper triangular part of A contains the upper triangular part of the matrix A . If uplo = CUBLAS_FILL_MODE_LOWER , the leading n-by-n lower triangular part of A contains the lower triangular part of the matrix A . On exit, if jobz = CUSOLVER_EIG_MODE_VECTOR , and info = 0, A contains the orthonormal eigenvectors of the matrix A . If jobz = CUSOLVER_EIG_MODE_NOVECTOR , the contents of A are destroyed.
lda	host	input	leading dimension of two-dimensional array used to store matrix A . lda is not less than max(1, n) .
B	device	in/out	<type> array of dimension ldb * n . If uplo = CUBLAS_FILL_MODE_UPPER , the leading n-by-n upper triangular part of B contains the upper triangular part of the matrix B . If uplo = CUBLAS_FILL_MODE_LOWER , the leading n-by-n lower triangular part of B contains the lower triangular part of the matrix B . On exit, if info is less than n , B is overwritten by triangular factor U or L from the Cholesky factorization of B .
ldb	host	input	leading dimension of two-dimensional array used to store matrix B . ldb is not less than max(1, n) .
W	device	output	a real array of dimension n . The eigenvalue values of A , sorted so that W(i) >= W(i+1) .
work	device	in/out	working space, <type> array of size lwork .
lwork	host	input	size of work , returned by sygvj_bufferSize .
info	device	output	if info = 0, the operation is successful. if info = -i, the i-th parameter is wrong (not counting handle). if info = i (> 0), info indicates either B is not positive definite or syevj (called by sygvj) does not converge.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($n < 0$, or $lda < \max(1, n)$, or $ldb < \max(1, n)$, or $itype$ is not 1, 2 or 3, or $jobz$ is not CUSOLVER_EIG_MODE_NOVECTOR or CUSOLVER_EIG_MODE_VECTOR, or $uplo$ is not CUBLAS_FILL_MODE_LOWER or CUBLAS_FILL_MODE_UPPER).
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.4.3.16. cusolverDn<t>syevjBatched()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverDnSsyevjBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const float *A,
    int lda,
    const float *W,
    int *lwork,
    syevjInfo_t params,
    int batchSize
);

cusolverStatus_t
cusolverDnDsyevjBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const double *A,
    int lda,
    const double *W,
    int *lwork,
    syevjInfo_t params,
    int batchSize
);

cusolverStatus_t
cusolverDnCsyevjBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const cuComplex *A,
    int lda,
    const float *W,
    int *lwork,
    syevjInfo_t params,
    int batchSize
);

cusolverStatus_t
cusolverDnZsyevjBatched_bufferSize(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    const cuDoubleComplex *A,
    int lda,
    const double *W,
    int *lwork,
    syevjInfo_t params,
    int batchSize
);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnSsyevjBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    float *A,
    int lda,
    float *W,
    float *work,
    int lwork,
    int *info,
    syevjInfo_t params,
    int batchSize
);
```

```
cusolverStatus_t
cusolverDnDsyevjBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    double *A,
    int lda,
    double *W,
    double *work,
    int lwork,
    int *info,
    syevjInfo_t params,
    int batchSize
);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverDnCheevjBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    cuComplex *A,
    int lda,
    float *W,
    cuComplex *work,
    int lwork,
    int *info,
    syevjInfo_t params,
    int batchSize
);

cusolverStatus_t
cusolverDnZheevjBatched(
    cusolverDnHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int n,
    cuDoubleComplex *A,
    int lda,
    double *W,
    cuDoubleComplex *work,
    int lwork,
    int *info,
    syevjInfo_t params,
    int batchSize
);
```

This function computes eigenvalues and eigenvectors of a sequence of symmetric (Hermitian) $\mathbf{n} \times \mathbf{n}$ matrices

$$A_j^* Q_j = Q_j^* \Lambda_j$$

where Λ_j is a real $\mathbf{n} \times \mathbf{n}$ diagonal matrix. Q_j is an $\mathbf{n} \times \mathbf{n}$ unitary matrix. The diagonal elements of Λ_j are the eigenvalues of A_j in either ascending order or non-sorting order.

syevjBatched performs **syevj** on each matrix. It requires that all matrices are of the same size \mathbf{n} and are packed in contiguous way,

$$A = (A_0 \ A_1 \ \dots)$$

Each matrix is column-major with leading dimension **lda**, so the formula for random access is $A_k(i,j) = A[i + lda*j + lda*n*k]$.

The parameter **w** also contains eigenvalues of each matrix in contiguous way,

$$W = (W_0 \ W_1 \ \dots)$$

The formula for random access of **w** is $W_k(j) = W[j + n*k]$.

Except for tolerance and maximum sweeps, **syevjBatched** can either sort the eigenvalues in ascending order (default) or chose as-is (without sorting) by the function **cusolverDnXsyevjSetSortEig**. If the user packs several tiny matrices into diagonal blocks of one matrix, non-sorting option can separate spectrum of those tiny matrices.

syevjBatched cannot report residual and executed sweeps by function **cusolverDnXsyevjGetResidual** and **cusolverDnXsyevjGetSweeps**. Any call of the above two returns **CUSOLVER_STATUS_NOT_SUPPORTED**. The user needs to compute residual explicitly.

The user has to provide working space pointed by input parameter **work**. The input parameter **lwork** is the size of the working space, and it is returned by **syevjBatched_bufferSize()**.

The output parameter **info** is an integer array of size **batchSize**. If the function returns **CUSOLVER_STATUS_INVALID_VALUE**, the first element **info[0] = -i** (less than zero) indicates **i-th** parameter is wrong (not counting handle). Otherwise, if **info[i] = n+1**, **syevjBatched** does not converge on **i-th** matrix under given tolerance and maximum sweeps.

if **jobz = CUSOLVER_EIG_MODE_VECTOR**, A_j contains the orthonormal eigenvectors V_j .

Appendix F.5 provides a simple example of **syevjBatched**.

API of syevjBatched

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverDN library context.
jobz	host	input	specifies options to either compute eigenvalue only or compute eigen-pair: jobz = CUSOLVER_EIG_MODE_NOVECTOR : Compute eigenvalues only; jobz = CUSOLVER_EIG_MODE_VECTOR : Compute eigenvalues and eigenvectors.
uplo	host	input	specifies which part of A_j is stored. uplo = CUBLAS_FILL_MODE_LOWER : Lower triangle of A_j is stored. uplo = CUBLAS_FILL_MODE_UPPER : Upper triangle of A_j is stored.
n	host	input	number of rows (or columns) of matrix each A_j .
A	device	in/out	<type> array of dimension $lda * n * batchSize$ with lda is not less than $\max(1, n)$. If uplo = CUBLAS_FILL_MODE_UPPER , the leading n -by- n upper triangular part of A_j contains the upper triangular part of the matrix A_j . If uplo = CUBLAS_FILL_MODE_LOWER , the leading n -by- n lower triangular part of A_j contains the lower triangular part of the matrix A_j . On exit, if jobz = CUSOLVER_EIG_MODE_VECTOR , and info[j] = 0 , A_j contains the orthonormal

			eigenvectors of the matrix A_j . If <code>jobz = CUSOLVER_EIG_MODE_NOVECTOR</code> , the contents of A_j are destroyed.
<code>lda</code>	<code>host</code>	<code>input</code>	leading dimension of two-dimensional array used to store matrix A_j .
<code>W</code>	<code>device</code>	<code>output</code>	a real array of dimension $n \times \text{batchSize}$. It stores the eigenvalues of A_j in ascending order or non-sorting order.
<code>work</code>	<code>device</code>	<code>in/out</code>	<type> array of size <code>lwork</code> , workspace.
<code>lwork</code>	<code>host</code>	<code>input</code>	size of <code>work</code> , returned by <code>syevjBatched_bufferSize</code> .
<code>info</code>	<code>device</code>	<code>output</code>	an integer array of dimension <code>batchSize</code> . If <code>CUSOLVER_STATUS_INVALID_VALUE</code> is returned, <code>info[0] = -i</code> (less than zero) indicates <i>i</i> -th parameter is wrong (not counting handle). Otherwise, if <code>info[i] = 0</code> , the operation is successful. if <code>info[i] = n+1</code> , <code>syevjBatched</code> dose not converge on <i>i</i> -th matrix under given tolerance and maximum sweeps.
<code>params</code>	<code>host</code>	<code>in/out</code>	structure filled with parameters of Jacobi algorithm.
<code>batchSize</code>	<code>host</code>	<code>input</code>	number of matrices.

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed ($n < 0$ or $lda < \max(1, n)$, or <code>jobz</code> is not <code>CUSOLVER_EIG_MODE_NOVECTOR</code> or <code>CUSOLVER_EIG_MODE_VECTOR</code> , or <code>uplo</code> is not <code>CUBLAS_FILL_MODE_LOWER</code> or <code>CUBLAS_FILL_MODE_UPPER</code>), or <code>batchSize < 0</code> .
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

2.5. cuSolverSP: sparse LAPACK Function Reference

This chapter describes the API of cuSolverSP, which provides a subset of LAPACK funtions for sparse matrices in CSR or CSC format.

2.5.1. Helper Function Reference

2.5.1.1. cusolverSpCreate()

```
cusolverStatus_t
cusolverSpCreate(cusolverSpHandle_t *handle)
```

This function initializes the cuSolverSP library and creates a handle on the cuSolver context. It must be called before any other cuSolverSP API function is invoked. It allocates hardware resources necessary for accessing the GPU.

Output

handle	the pointer to the handle to the cuSolverSP context.
---------------	--

Status Returned

CUSOLVER_STATUS_SUCCESS	the initialization succeeded.
CUSOLVER_STATUS_NOT_INITIALIZED	the CUDA Runtime initialization failed.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.

2.5.1.2. cusolverSpDestroy()

```
cusolverStatus_t
cusolverSpDestroy(cusolverSpHandle_t handle)
```

This function releases CPU-side resources used by the cuSolverSP library.

Input

handle	the handle to the cuSolverSP context.
---------------	---------------------------------------

Status Returned

CUSOLVER_STATUS_SUCCESS	the shutdown succeeded.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

2.5.1.3. cusolverSpSetStream()

```
cusolverStatus_t
cusolverSpSetStream(cusolverSpHandle_t handle, cudaStream_t streamId)
```

This function sets the stream to be used by the cuSolverSP library to execute its routines.

Input

handle	the handle to the cuSolverSP context.
streamId	the stream to be used by the library.

Status Returned

CUSOLVER_STATUS_SUCCESS	the stream was set successfully.
--------------------------------	----------------------------------

CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
---------------------------------	----------------------------------

2.5.1.4. cusolverSpXcsrissym()

```
cusolverStatus_t
cusolverSpXcsrissymHost(cusolverSpHandle_t handle,
                        int m,
                        int nnzA,
                        const cusparseMatDescr_t descrA,
                        const int *csrRowPtrA,
                        const int *csrEndPtrA,
                        const int *csrColIndA,
                        int *issym);
```

This function checks if **A** has symmetric pattern or not. The output parameter **issym** reports 1 if **A** is symmetric; otherwise, it reports 0.

The matrix **A** is an **m**×**m** sparse matrix that is defined in CSR storage format by the four arrays **csrValA**, **csrRowPtrA**, **csrEndPtrA** and **csrColIndA**.

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**.

The **csrsvlu** and **csrsvqr** do not accept non-general matrix. the user has to extend the matrix into its missing upper/lower part, otherwise the result is not expected. The user can use **csrissym** to check if the matrix has symmetric pattern or not.

Remark 1: only CPU path is provided.

Remark 2: the user has to check returned status to get valid information. The function converts **A** to CSC format and compare CSR and CSC format. If the CSC failed because of insufficient resources, **issym** is undefined, and this state can only be detected by the return status code.

Input

parameter	MemorySpace	description
handle	host	handle to the cuSolverSP library context.
m	host	number of rows and columns of matrix A .
nnzA	host	number of nonzeros of matrix A . It is the size of csrValA and csrColIndA .
descrA	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE .
csrRowPtrA	host	integer array of m elements that contains the start of every row.
csrEndPtrA	host	integer array of m elements that contains the end of the last row plus one.
csrColIndA	host	integer array of nnzA column indices of the nonzero elements of matrix A .

Output

parameter	MemorySpace	description
issym	host	1 if A is symmetric; 0 otherwise.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (m , nnzA ≤ 0), base index is not 0 or 1.
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED	the matrix type is not supported.

2.5.2. High Level Function Reference

This section describes high level API of cuSolverSP, including linear solver, least-square solver and eigenvalue solver. The high-level API is designed for ease-of-use, so it allocates any required memory under the hood automatically. If the host or GPU system memory is not enough, an error is returned.

2.5.2.1. cusolverSp<t>csrslsvlu()

```
cusolverStatus_t
cusolverSpScsrslsvlu[Host](cusolverSpHandle_t handle,
                           int n,
                           int nnzA,
                           const cusparseMatDescr_t descrA,
                           const float *csrValA,
                           const int *csrRowPtrA,
                           const int *csrColIndA,
                           const float *b,
                           float tol,
                           int reorder,
                           float *x,
                           int *singularity);
```

```
cusolverStatus_t
cusolverSpDcsrslsvlu[Host](cusolverSpHandle_t handle,
                           int n,
                           int nnzA,
                           const cusparseMatDescr_t descrA,
                           const double *csrValA,
                           const int *csrRowPtrA,
                           const int *csrColIndA,
                           const double *b,
                           double tol,
                           int reorder,
                           double *x,
                           int *singularity);
```

```
cusolverStatus_t
cusolverSpCcsrslsvlu[Host](cusolverSpHandle_t handle,
                           int n,
                           int nnzA,
                           const cusparseMatDescr_t descrA,
                           const cuComplex *csrValA,
                           const int *csrRowPtrA,
                           const int *csrColIndA,
                           const cuComplex *b,
                           float tol,
                           int reorder,
                           cuComplex *x,
                           int *singularity);
```

```
cusolverStatus_t
cusolverSpZcsrslsvlu[Host](cusolverSpHandle_t handle,
                           int n,
                           int nnzA,
                           const cusparseMatDescr_t descrA,
                           const cuDoubleComplex *csrValA,
                           const int *csrRowPtrA,
                           const int *csrColIndA,
                           const cuDoubleComplex *b,
                           double tol,
                           int reorder,
                           cuDoubleComplex *x,
                           int *singularity);
```

This function solves the linear system

$$A * x = b$$

A is an **n**×**n** sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA**, and **csrColIndA**. **b** is the right-hand-side vector of size **n**, and **x** is the solution vector of size **n**.

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**. If matrix **A** is symmetric/Hermitian and only lower/upper part is used or meaningful, the user has to extend the matrix into its missing upper/lower part, otherwise the result would be wrong.

The linear system is solved by sparse LU with partial pivoting,

$$P * A = L * U$$

cusolver library provides three reordering schemes, **symrcm**, **symamd**, and **csrmetisnd** to reduce zero fill-in which dramatically affects the performance of LU factorization. The input parameter **reorder** can enable **symrcm** (**symamd** or **csrmetisnd**) if **reorder** is 1 (2, or 3), otherwise, no reordering is performed.

If **reorder** is nonzero, **csrslsvlu** does

$$P * A * Q^T = L * U$$

where $Q = \text{symrcm}(A + A^T)$.

If **A** is singular under given tolerance (**max(tol, 0)**), then some diagonal elements of **U** is zero, i.e.

$$|U(j,j)| < \text{tol for some } j$$

The output parameter **singularity** is the smallest index of such **j**. If **A** is non-singular, **singularity** is -1. The index is base-0, independent of base index of **A**. For example, if 2nd column of **A** is the same as first column, then **A** is singular and **singularity** = 1 which means $U(1,1) \approx 0$.

Remark 1: **csrslsvlu** performs traditional LU with partial pivoting, the pivot of k-th column is determined dynamically based on the k-th column of intermediate matrix. **csrslsvlu** follows Gilbert and Peierls's algorithm [4] which uses depth-first-search and topological ordering to solve triangular system (Davis also describes this algorithm in detail in his book [1]). since cuda 10.1, **csrslsvlu** will incrementally reallocate the memory to store **L** and **U**. This feature can avoid over-estimate size from QR factorization. In some cases, zero fill-in of QR can be order of magnitude higher than LU.

Remark 2: only CPU (Host) path is provided.

Input

parameter	cusolverSp MemSpace	*Host MemSpace	description
handle	host	host	handle to the cuSolverSP library context.
n	host	host	number of rows and columns of matrix A .

nnzA	host	host	number of nonzeros of matrix A .
descrA	host	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE .
csrValA	device	host	<type> array of nnzA (= csrRowPtrA (n) - csrRowPtrA (0)) nonzero elements of matrix A .
csrRowPtrA	device	host	integer array of n + 1 elements that contains the start of every row and the end of the last row plus one.
csrColIndA	device	host	integer array of nnzA (= csrRowPtrA (n) - csrRowPtrA (0)) column indices of the nonzero elements of matrix A .
b	device	host	right hand side vector of size n .
tol	host	host	tolerance to decide if singular or not.
reorder	host	host	no ordering if reorder =0. Otherwise, symrcm , symamd , or csrmetisnd is used to reduce zero fill-in.

Output

parameter	cusolverSp MemSpace	*Host MemSpace	description
x	device	host	solution vector of size n , $x = \text{inv}(A) * b$.
singularity	host	host	-1 if A is invertible. Otherwise, first index j such that $u(j, j) \approx 0$

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (n , nnzA ≤ 0, base index is not 0 or 1, reorder is not 0, 1, 2, 3)
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED	the matrix type is not supported.

2.5.2.2. cusolverSp<t>csrslsvqr()

```

cusolverStatus_t
cusolverSpScsrslsvqr[Host](cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const float *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    const float *b,
    float tol,
    int reorder,
    float *x,
    int *singularity);

cusolverStatus_t
cusolverSpDcsrslsvqr[Host](cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const double *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    const double *b,
    double tol,
    int reorder,
    double *x,
    int *singularity);

cusolverStatus_t
cusolverSpCcsrslsvqr[Host](cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const cuComplex *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    const cuComplex *b,
    float tol,
    int reorder,
    cuComplex *x,
    int *singularity);

cusolverStatus_t
cusolverSpZcsrslsvqr[Host](cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const cuDoubleComplex *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    const cuDoubleComplex *b,
    double tol,
    int reorder,
    cuDoubleComplex *x,
    int *singularity);

```

This function solves the linear system

$$A * x = b$$

A is an **m**×**m** sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA**, and **csrColIndA**. **b** is the right-hand-side vector of size **m**, and **x** is the solution vector of size **m**.

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**. If matrix **A** is symmetric/Hermitian and only lower/upper part is used or meaningful, the user has to extend the matrix into its missing upper/lower part, otherwise the result would be wrong.

The linear system is solved by sparse QR factorization,

$$A = Q * R$$

If **A** is singular under given tolerance (**max(tol, 0)**), then some diagonal elements of **R** is zero, i.e.

$$|R(j,j)| < \text{tol for some } j$$

The output parameter **singularity** is the smallest index of such **j**. If **A** is non-singular, **singularity** is -1. The **singularity** is base-0, independent of base index of **A**. For example, if 2nd column of **A** is the same as first column, then **A** is singular and **singularity** = 1 which means **R(1,1) ≈ 0**.

cusolver library provides three reordering schemes, **symrcm**, **symamd**, and **csrmetisnd** to reduce zero fill-in which dramatically affects the performance of QR factorization. The input parameter **reorder** can enable **symrcm** (**symamd** or **csrmetisnd**) if **reorder** is 1 (2, or 3), otherwise, no reordering is performed.

Input

parameter	cusolverSp MemSpace	*Host MemSpace	description
handle	host	host	handle to the cuSolverSP library context.
m	host	host	number of rows and columns of matrix A .
nnz	host	host	number of nonzeros of matrix A .
descrA	host	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE .
csrValA	device	host	<type> array of nnz (= csrRowPtrA (m) - csrRowPtrA (0)) nonzero elements of matrix A .
csrRowPtrA	device	host	integer array of m + 1 elements that contains the start of every row and the end of the last row plus one.

csrColIndA	device	host	integer array of nnz ($= \text{csrRowPtrA}(\text{m}) - \text{csrRowPtrA}(0)$) column indices of the nonzero elements of matrix A .
b	device	host	right hand side vector of size m .
tol	host	host	tolerance to decide if singular or not.
reorder	host	host	no ordering if reorder =0. Otherwise, symrcm , symamd , or csrmetisnd is used to reduce zero fill-in.

Output

parameter	cusolverSp MemSpace	*Host MemSpace	description
x	device	host	solution vector of size m , $x = \text{inv}(\text{A}) * b$.
singularity	host	host	-1 if A is invertible. Otherwise, first index j such that $R(j, j) \approx 0$

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (m , nnz ≤ 0 , base index is not 0 or 1, reorder is not 0,1,2,3)
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED	the matrix type is not supported.

2.5.2.3. cusolverSp<t>csrsvchol()

```
cusolverStatus_t
cusolverSpScsrsvchol[Host](cusolverSpHandle_t handle,
                           int m,
                           int nnz,
                           const cusparseMatDescr_t descrA,
                           const float *csrVal,
                           const int *csrRowPtr,
                           const int *csrColInd,
                           const float *b,
                           float tol,
                           int reorder,
                           float *x,
                           int *singularity);
```

```
cusolverStatus_t
cusolverSpDcsrsvchol[Host](cusolverSpHandle_t handle,
                           int m,
                           int nnz,
                           const cusparseMatDescr_t descrA,
                           const double *csrVal,
                           const int *csrRowPtr,
                           const int *csrColInd,
                           const double *b,
                           double tol,
                           int reorder,
                           double *x,
                           int *singularity);
```

```
cusolverStatus_t
cusolverSpCcsrsvchol[Host](cusolverSpHandle_t handle,
                           int m,
                           int nnz,
                           const cusparseMatDescr_t descrA,
                           const cuComplex *csrVal,
                           const int *csrRowPtr,
                           const int *csrColInd,
                           const cuComplex *b,
                           float tol,
                           int reorder,
                           cuComplex *x,
                           int *singularity);
```

```
cusolverStatus_t
cusolverSpZcsrsvchol[Host](cusolverSpHandle_t handle,
                           int m,
                           int nnz,
                           const cusparseMatDescr_t descrA,
                           const cuDoubleComplex *csrVal,
                           const int *csrRowPtr,
                           const int *csrColInd,
                           const cuDoubleComplex *b,
                           double tol,
                           int reorder,
                           cuDoubleComplex *x,
                           int *singularity);
```


This function solves the linear system

$$A * x = b$$

A is an **m**×**m** symmetric postive definite sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA**, and **csrColIndA**. **b** is the right-hand-side vector of size **m**, and **x** is the solution vector of size **m**.

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL** and upper triangular part of **A** is ignored (if parameter **reorder** is zero). In other words, suppose input matrix **A** is decomposed as $A = L + D + U$, where **L** is lower triangular, **D** is diagonal and **U** is upper triangular. The function would ignore **U** and regard **A** as a symmetric matrix with the formula $A = L + D + L^H$. If parameter **reorder** is nonzero, the user has to extend **A** to a full matrix, otherwise the solution would be wrong.

The linear system is solved by sparse Cholesky factorization,

$$A = G * G^H$$

where **G** is the Cholesky factor, a lower triangular matrix.

The output parameter **singularity** has two meanings:

- ▶ If **A** is not postive definite, there exists some integer **k** such that **A**(0:k, 0:k) is not positive definite. **singularity** is the minimum of such **k**.
- ▶ If **A** is postive definite but near singular under tolerance (**max(tol, 0)**), i.e. there exists some integer **k** such that $G(k,k) \leq \text{tol}$. **singularity** is the minimum of such **k**.

singularity is base-0. If **A** is positive definite and not near singular under tolerance, **singularity** is -1. If the user wants to know if **A** is postive definite or not, **tol=0** is enough.

cusolver library provides three reordering schemes, **symrcm** **symamd**, and **csrmetisnd** to reduce zero fill-in which dramactically affects the performance of Cholesky factorization. The input parameter **reorder** can enable **symrcm** (**symamd** or **csrmetisnd**) if **reorder** is 1 (2, or 3), otherwise, no reordering is performed.

Remark 1: the function works for in-place (**x** and **b** point to the same memory block) and out-of-place.

Remark 2: the function only works on 32-bit index, if matrix **G** has large zero fill-in such that number of nonzeros is bigger than 2^{31} , then **CUSOLVER_STATUS_ALLOC_FAILED** is returned.

Input

parameter	cusolverSp MemSpace	*Host MemSpace	description
handle	host	host	handle to the cuSolverSP library context.
m	host	host	number of rows and columns of matrix A .
nnz	host	host	number of nonzeros of matrix A .

descrA	host	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE .
csrValA	device	host	<type> array of $nnz (= \text{csrRowPtrA}(m) - \text{csrRowPtrA}(0))$ nonzero elements of matrix A .
csrRowPtrA	device	host	integer array of $m + 1$ elements that contains the start of every row and the end of the last row plus one.
csrColIndA	device	host	integer array of $nnz (= \text{csrRowPtrA}(m) - \text{csrRowPtrA}(0))$ column indices of the nonzero elements of matrix A .
b	device	host	right hand side vector of size m .
tol	host	host	tolerance to decide singularity.
reorder	host	host	no ordering if reorder =0. Otherwise, symrcm , symamd , or csrmetisnd is used to reduce zero fill-in.

Output

parameter	cusolverSp MemSpace	*Host MemSpace	description
x	device	host	solution vector of size m , $x = \text{inv}(A) * b$.
singularity	host	host	-1 if A is symmetric postive definite.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($m, nnz \leq 0$, base index is not 0 or 1, reorder is not 0,1,2,3)
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED	the matrix type is not supported.

2.5.2.4. `cusolverSp<t>csrslsqvqr()`

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverSpScsrslsqvqr[Host](cusolverSpHandle_t handle,
                             int m,
                             int n,
                             int nnz,
                             const cusparseMatDescr_t descrA,
                             const float *csrValA,
                             const int *csrRowPtrA,
                             const int *csrColIndA,
                             const float *b,
                             float tol,
                             int *rankA,
                             float *x,
                             int *p,
                             float *min_norm);

cusolverStatus_t
cusolverSpDcsrslsqvqr[Host](cusolverSpHandle_t handle,
                             int m,
                             int n,
                             int nnz,
                             const cusparseMatDescr_t descrA,
                             const double *csrValA,
                             const int *csrRowPtrA,
                             const int *csrColIndA,
                             const double *b,
                             double tol,
                             int *rankA,
                             double *x,
                             int *p,
                             double *min_norm);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverSpCcsr1sqvqr[Host](cusolverSpHandle_t handle,
    int m,
    int n,
    int nnz,
    const cusparseMatDescr_t descrA,
    const cuComplex *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    const cuComplex *b,
    float tol,
    int *rankA,
    cuComplex *x,
    int *p,
    float *min_norm);

cusolverStatus_t
cusolverSpZcsr1sqvqr[Host](cusolverSpHandle_t handle,
    int m,
    int n,
    int nnz,
    const cusparseMatDescr_t descrA,
    const cuDoubleComplex *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    const cuDoubleComplex *b,
    double tol,
    int *rankA,
    cuDoubleComplex *x,
    int *p,
    double *min_norm);
```

This function solves the following least-square problem

$$x = \operatorname{argmin} \|A^* z - b\|$$

A is an **m**×**n** sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA**, and **csrColIndA**. **b** is the right-hand-side vector of size **m**, and **x** is the least-square solution vector of size **n**.

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**. If **A** is square, symmetric/Hermitian and only lower/upper part is used or meaningful, the user has to extend the matrix into its missing upper/lower part, otherwise the result is wrong.

This function only works if **m** is greater or equal to **n**, in other words, **A** is a tall matrix.

The least-square problem is solved by sparse QR factorization with column pivoting,

$$A^* P^T = Q^* R$$

If **A** is of full rank (i.e. all columns of **A** are linear independent), then matrix **P** is an identity. Suppose rank of **A** is **k**, less than **n**, the permutation matrix **P** reorders columns of **A** in the following sense:

$$A^* P^T = (A_1 \ A_2) = (Q_1 \ Q_2) \begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix}$$

where R_{11} and \mathbf{A} have the same rank, but R_{22} is almost zero, i.e. every column of A_2 is linear combination of A_1 .

The input parameter **tol** decides numerical rank. The absolute value of every entry in R_{22} is less than or equal to **tolerance=max(tol, 0)**.

The output parameter **rankA** denotes numerical rank of \mathbf{A} .

Suppose $y = P^*x$ and $c = Q^H*b$, the least square problem can be reformed by

$$\min ||A^*x - b|| = \min ||R^*y - c||$$

or in matrix form

$$\begin{pmatrix} R_{11} & R_{12} \\ & R_{22} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

The output parameter **min_norm** is $||c_2||$, which is minimum value of least-square problem.

If \mathbf{A} is not of full rank, above equation does not have a unique solution. The least-square problem is equivalent to

$$\begin{aligned} &\min ||y|| \\ &\text{subject to } R_{11}^*y_1 + R_{12}^*y_2 = c_1 \end{aligned}$$

Or equivalently another least-square problem

$$\min || \begin{pmatrix} R_{11} \setminus R_{12} \\ I \end{pmatrix}^* y_2 - \begin{pmatrix} R_{11} \setminus c_1 \\ 0 \end{pmatrix} ||$$

The output parameter \mathbf{x} is P^T*y , the solution of least-square problem.

The output parameter **p** is a vector of size **n**. It corresponds to a permutation matrix **P**. **p(i)=j** means **(P*x)(i) = x(j)**. If \mathbf{A} is of full rank, **p=0:n-1**.

Remark 1: **p** is always base 0, independent of base index of \mathbf{A} .

Remark 2: only CPU (Host) path is provided.

Input

parameter	cusolverSp MemSpace	*Host MemSpace	description
handle	host	host	handle to the cuSolver library context.
m	host	host	number of rows of matrix A .
n	host	host	number of columns of matrix A .
nnz	host	host	number of nonzeros of matrix A .
descrA	host	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are

			CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE.
csrValA	device	host	<type> array of $nnz (= \text{csrRowPtrA}(m) - \text{csrRowPtrA}(0))$ nonzero elements of matrix A .
csrRowPtrA	device	host	integer array of $m + 1$ elements that contains the start of every row and the end of the last row plus one.
csrColIndA	device	host	integer array of $nnz (= \text{csrRowPtrA}(m) - \text{csrRowPtrA}(0))$ column indices of the nonzero elements of matrix A .
b	device	host	right hand side vector of size m .
tol	host	host	tolerance to decide rank of A .

Output

parameter	cusolverSp MemSpace	*Host MemSpace	description
rankA	host	host	numerical rank of A .
x	device	host	solution vector of size n , $x = \text{pinv}(\mathbf{A}) * \mathbf{b}$.
p	device	host	a vector of size n , which represents the permutation matrix P satisfying $\mathbf{A} * \mathbf{P}^T = \mathbf{Q} * \mathbf{R}$.
min_norm	host	host	$ \mathbf{A} * \mathbf{x} - \mathbf{b} $, $\mathbf{x} = \text{pinv}(\mathbf{A}) * \mathbf{b}$.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($m, n, nnz \leq 0$), base index is not 0 or 1.
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED	the matrix type is not supported.

2.5.2.5. cusolverSp<t>csreigvsi()

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverSpScsreigvsi[Host](cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const float *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    float mu0,
    const float *x0,
    int maxite,
    float tol,
    float *mu,
    float *x);

cusolverStatus_t
cusolverSpDcsreigvsi[Host](cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const double *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    double mu0,
    const double *x0,
    int maxite,
    double tol,
    double *mu,
    double *x);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverSpCcsreigvsi[Host](cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const cuComplex *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    cuComplex mu0,
    const cuComplex *x0,
    int maxite,
    float tol,
    cuComplex *mu,
    cuComplex *x);

cusolverStatus_t
cusolverSpZcsreigvsi(cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const cuDoubleComplex *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    cuDoubleComplex mu0,
    const cuDoubleComplex *x0,
    int maxite,
    double tol,
    cuDoubleComplex *mu,
    cuDoubleComplex *x);
```

This function solves the simple eigenvalue problem $A * x = \lambda * x$ by shift-inverse method.

A is an $m \times m$ sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA**, and **csrColIndA**. The output parameter **x** is the approximated eigenvector of size **m**,

The following shift-inverse method corrects eigenpair step-by-step until convergence.

It accepts several parameters:

mu0 is an initial guess of eigenvalue. The shift-inverse method will converge to the eigenvalue **mu** nearest **mu0** if **mu** is a singleton. Otherwise, the shift-inverse method may not converge.

x0 is an initial eigenvector. If the user has no preference, just chose **x0** randomly. **x0** must be nonzero. It can be non-unit length.

tol is the tolerance to decide convergence. If **tol** is less than zero, it would be treated as zero.

maxite is maximum number of iterations. It is useful when shift-inverse method does not converge because the tolerance is too small or the desired eigenvalue is not a singleton.

Shift-Inverse Method


```

Given a initial guess of eigenvalue  $\mu_0$  and initial vector  $x_0$ 
 $x^{(0)} = x_0$  of unit length
for j = 0 : maxite
    solve  $(A - \mu_0 * I)$ 
    normalize  $x^{(k+1)}$  to unit length
    compute approx. eigenvalue  $\mu = x^H * A * x^{(k+1)}$ 
    if ||  $A - \mu$ 
endfor

```

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**. If **A** is symmetric/Hermitian and only lower/upper part is used or meaningful, the user has to extend the matrix into its missing upper/lower part, otherwise the result is wrong.

Remark 1: **[cu|h]solver[S|D]csreigvsi** only allows **mu0** as a real number. This works if **A** is symmetric. Otherwise, the non-real eigenvalue has a conjugate counterpart on the complex plan, and shift-inverse method would not converge to such eigenvalue even the eigenvalue is a singleton. The user has to extend **A** to complex number and call **[cu|h]solver[C|Z]csreigvsi** with **mu0** not on real axis.

Remark 2: the tolerance **tol** should not be smaller than $|\mu_0| * \text{eps}$, where **eps** is machine zero. Otherwise, shift-inverse may not converge because of small tolerance.

Input

parameter	cusolverSp MemSpace	*Host MemSpace	description
handle	host	host	handle to the cuSolver library context.
m	host	host	number of rows and columns of matrix A .
nnz	host	host	number of nonzeros of matrix A .
descrA	host	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE .
csrValA	device	host	<type> array of nnz (= csrRowPtrA(m) - csrRowPtrA(0)) nonzero elements of matrix A .
csrRowPtrA	device	host	integer array of m + 1 elements that contains the start of every row and the end of the last row plus one.
csrColIndA	device	host	integer array of nnz (= csrRowPtrA(m) - csrRowPtrA(0)) column indices of the nonzero elements of matrix A .
mu0	host	host	initial guess of eigenvalue.
x0	device	host	initial guess of eigenvector, a vector of size m.
maxite	host	host	maximum iterations in shift-inverse method.

<code>tol</code>	<code>host</code>	<code>host</code>	tolerance for convergence.
------------------	-------------------	-------------------	----------------------------

Output

parameter	cusolverSp MemSpace	*Host MemSpace	description
<code>mu</code>	<code>device</code>	<code>host</code>	approximated eigenvalue nearest <code>mu0</code> under tolerance.
<code>x</code>	<code>device</code>	<code>host</code>	approximated eigenvector of size <code>m</code> .

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_ALLOC_FAILED</code>	the resources could not be allocated.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed (<code>m, nnz <= 0</code>), base index is not 0 or 1.
<code>CUSOLVER_STATUS_ARCH_MISMATCH</code>	the device only supports compute capability 2.0 and above.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.
<code>CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED</code>	the matrix type is not supported.

2.5.2.6. cusolverSp<t>csreigs()

```

cusolverStatus_t
solv erspScsreigs[Host] (cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const float *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    cuComplex left_bottom_corner,
    cuComplex right_upper_corner,
    int *num_eigs);

cusolverStatus_t
cusolverSpDcsrreigs[Host] (cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const double *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    cuDoubleComplex left_bottom_corner,
    cuDoubleComplex right_upper_corner,
    int *num_eigs);

cusolverStatus_t
cusolverSpCcsrreigs[Host] (cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const cuComplex *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    cuComplex left_bottom_corner,
    cuComplex right_upper_corner,
    int *num_eigs);

cusolverStatus_t
cusolverSpZcsrreigs[Host] (cusolverSpHandle_t handle,
    int m,
    int nnz,
    const cusparseMatDescr_t descrA,
    const cuDoubleComplex *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    cuDoubleComplex left_bottom_corner,
    cuDoubleComplex right_upper_corner,
    int *num_eigs);

```

This function computes number of algebraic eigenvalues in a given box **B** by contour integral

$$\text{number of algebraic eigenvalues in box } B = \frac{1}{2 * \pi * \sqrt{-1}} \oint_C \frac{P'(z)}{P(z)} dz$$

where closed line **C** is boundary of the box **B** which is a rectangle specified by two points, one is left bottom corner (input parameter **left_bottom_corner**) and the other is right upper corner (input parameter **right_upper_corner**). $P(z) = \det(A - z \cdot I)$ is the characteristic polynomial of **A**.

A is an $m \times m$ sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA**, and **csrColIndA**.

The output parameter **num_eigs** is number of algebraic eigenvalues in the box **B**. This number may not be accurate due to several reasons:

1. the contour **C** is close to some eigenvalues or even passes through some eigenvalues.
2. the numerical integration is not accurate due to coarse grid size. The default resolution is 1200 grids along contour **C** uniformly.

Even though **csreigs** may not be accurate, it still can give the user some idea how many eigenvalues in a region where the resolution of disk theorem is bad. For example, standard 3-point stencil of finite difference of Laplacian operator is a tridiagonal matrix, and disk theorem would show "all eigenvalues are in the interval $[0, 4 \cdot N^2]$ " where N is number of grids. In this case, **csreigs** is useful for any interval inside $[0, 4 \cdot N^2]$.

Remark 1: if **A** is symmetric in real or hermitian in complex, all eigenvalues are real. The user still needs to specify a box, not an interval. The height of the box can be much smaller than the width.

Remark 2: only CPU (Host) path is provided.

Input

parameter	cusolverSp MemSpace	*Host MemSpace	description
handle	host	host	handle to the cuSolverSP library context.
m	host	host	number of rows and columns of matrix A .
nnz	host	host	number of nonzeros of matrix A .
descrA	host	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE .
csrValA	device	host	<type> array of $nnz (= \text{csrRowPtrA}(m) - \text{csrRowPtrA}(0))$ nonzero elements of matrix A .
csrRowPtrA	device	host	integer array of $m + 1$ elements that contains the start of every row and the end of the last row plus one.
csrColIndA	device	host	integer array of $nnz (= \text{csrRowPtrA}(m) - \text{csrRowPtrA}(0))$ column indices of the nonzero elements of matrix A .
left_bottom_corner	host	host	left bottom corner of the box.
right_upper_corner	host	host	right upper corner of the box.

Output

parameter	cusolverSp MemSpace	*Host MemSpace	description
num_eigs	host	host	number of algebraic eigenvalues in a box.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($m, nnz \leq 0$), base index is not 0 or 1.
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED	the matrix type is not supported.

2.5.3. Low Level Function Reference

This section describes low level API of cuSolverSP, including symrcm and batched QR.

2.5.3.1. cusolverSpXcsrsmrcm()

```
cusolverStatus_t
cusolverSpXcsrsmrcmHost(cusolverSpHandle_t handle,
    int n,
    int nnzA,
    const cusparseMatDescr_t descrA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    int *p);
```

This function implements Symmetric Reverse Cuthill-McKee permutation. It returns a permutation vector **p** such that **A(p,p)** would concentrate nonzeros to diagonal. This is equivalent to **symrcm** in MATLAB, however the result may not be the same because of different heuristics in the pseudoperipheral finder. The **cuSolverSP** library implements **symrcm** based on the following two papers:

E. Chuthill and J. McKee, reducing the bandwidth of sparse symmetric matrices, ACM '69 Proceedings of the 1969 24th national conference, Pages 157-172

Alan George, Joseph W. H. Liu, An Implementation of a Pseudoperipheral Node Finder, ACM Transactions on Mathematical Software (TOMS) Volume 5 Issue 3, Sept. 1979, Pages 284-295

The output parameter **p** is an integer array of **n** elements. It represents a permutation array and it indexed using the base-0 convention. The permutation array **p** corresponds to a permutation matrix **P**, and satisfies the following relation:

$$A(p,p) = P^* A^* P^T$$

A is an **n×n** sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA**, and **csrColIndA**.

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**. Internally **rcm** works on $A + A^T$, the user does not need to extend the matrix if the matrix is not symmetric.

Remark 1: only CPU (Host) path is provided.

Input

parameter	*Host MemSpace	description
handle	host	handle to the cuSolverSP library context.
n	host	number of rows and columns of matrix A .
nnzA	host	number of nonzeros of matrix A . It is the size of csrValA and csrColIndA .
descrA	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE .
csrRowPtrA	host	integer array of n+1 elements that contains the start of every row and the end of the last row plus one.
csrColIndA	host	integer array of nnzA column indices of the nonzero elements of matrix A .

Output

parameter	hsolver	description
p	host	permutation vector of size n.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (n , nnzA ≤ 0), base index is not 0 or 1.
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED	the matrix type is not supported.

2.5.3.2. cusolverSpXcsrSymmdq()

```
cusolverStatus_t
cusolverSpXcsrSymmdqHost(cusolverSpHandle_t handle,
    int n,
    int nnzA,
    const cusparseMatDescr_t descrA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    int *p);
```

This function implements Symmetric Minimum Degree Algorithm based on Quotient Graph. It returns a permutation vector **p** such that **A(p, p)** would have less zero fill-in during Cholesky factorization. The **cuSolverSP** library implements **symmdq** based on the following two papers:

Patrick R. Amestoy, Timothy A. Davis, Iain S. Duff, An Approximate Minimum Degree Ordering Algorithm, SIAM J. Matrix Analysis Applic. Vol 17, no 4, pp. 886-905, Dec. 1996.

Alan George, Joseph W. Liu, A Fast Implementation of the Minimum Degree Algorithm Using Quotient Graphs, ACM Transactions on Mathematical Software, Vol 6, No. 3, September 1980, page 337-358.

The output parameter **p** is an integer array of **n** elements. It represents a permutation array with base-0 index. The permutation array **p** corresponds to a permutation matrix **P**, and satisfies the following relation:

$$A(p,p) = P * A * P^T$$

A is an **n×n** sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA**, and **csrColIndA**.

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**. Internally **mdq** works on $A + A^T$, the user does not need to extend the matrix if the matrix is not symmetric.

Remark 1: only CPU (Host) path is provided.

Input

parameter	*Host MemSpace	description
handle	host	handle to the cuSolverSP library context.
n	host	number of rows and columns of matrix A .
nnzA	host	number of nonzeros of matrix A . It is the size of csrValA and csrColIndA .
descrA	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE .

csrRowPtrA	host	integer array of n+1 elements that contains the start of every row and the end of the last row plus one.
csrColIndA	host	integer array of nnzA column indices of the nonzero elements of matrix A .

Output

parameter	hsolver	description
p	host	permutation vector of size n .

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (n , nnzA ≤ 0), base index is not 0 or 1.
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED	the matrix type is not supported.

2.5.3.3. cusolverSpXcsrSymamd()

```
cusolverStatus_t
cusolverSpXcsrSymamdHost(cusolverSpHandle_t handle,
    int n,
    int nnzA,
    const cusparseMatDescr_t descrA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    int *p);
```

This function implements Symmetric Approximate Minimum Degree Algorithm based on Quotient Graph. It returns a permutation vector **p** such that **A(p, p)** would have less zero fill-in during Cholesky factorization. The **cuSolverSP** library implements **symamd** based on the following paper:

Patrick R. Amestoy, Timothy A. Davis, Iain S. Duff, An Approximate Minimum Degree Ordering Algorithm, SIAM J. Matrix Analysis Applic. Vol 17, no 4, pp. 886-905, Dec. 1996.

The output parameter **p** is an integer array of **n** elements. It represents a permutation array with base-0 index. The permutation array **p** corresponds to a permutation matrix **P**, and satisfies the following relation:

$$A(p,p) = P^* A^* P^T$$

A is an $n \times n$ sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA**, and **csrColIndA**.

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**. Internally **amd** works on $A + A^T$, the user does not need to extend the matrix if the matrix is not symmetric.

Remark 1: only CPU (Host) path is provided.

Input

parameter	*Host MemSpace	description
handle	host	handle to the cuSolverSP library context.
n	host	number of rows and columns of matrix A .
nnzA	host	number of nonzeros of matrix A . It is the size of csrValA and csrColIndA .
descrA	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE .
csrRowPtrA	host	integer array of $n+1$ elements that contains the start of every row and the end of the last row plus one.
csrColIndA	host	integer array of nnzA column indices of the nonzero elements of matrix A .

Output

parameter	hsolver	description
p	host	permutation vector of size n .

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($n, nnzA \leq 0$), base index is not 0 or 1.
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED	the matrix type is not supported.

2.5.3.4. cusolverSpXcsrmetisnd()

```
cusolverStatus_t
cusolverSpXcsrmetisndHost(
    cusolverSpHandle_t handle,
    int n,
    int nnzA,
    const cusparseMatDescr_t descrA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    const int64_t *options,
    int *p);
```

This function is a wrapper of **METIS_NodeND**. It returns a permutation vector **p** such that **A(p,p)** would have less zero fill-in during nested dissection. The **cuSolverSP** library links **libmetis_static.a** which is 64-bit metis-5.1.0 .

The parameter **options** is the configuration of **metis**. For those who do not have experiences of **metis**, set **options = NULL** for default setting.

The output parameter **p** is an integer array of **n** elements. It represents a permutation array with base-0 index. The permutation array **p** corresponds to a permutation matrix **P**, and satisfies the following relation:

$$A(p,p) = P^* A^* P^T$$

A is an **n×n** sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA**, and **csrColIndA**.

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**. Internally **csrmetisnd** works on $A + A^T$, the user does not need to extend the matrix if the matrix is not symmetric.

Remark 1: only CPU (Host) path is provided.

Input

parameter	*Host MemSpace	description
handle	host	handle to the cuSolverSP library context.
n	host	number of rows and columns of matrix A .
nnzA	host	number of nonzeros of matrix A . It is the size of csrValA and csrColIndA .
descrA	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE .
csrRowPtrA	host	integer array of n+1 elements that contains the start of every row and the end of the last row plus one.

<code>csrColIndA</code>	<code>host</code>	integer array of <code>nnzA</code> column indices of the nonzero elements of matrix <code>A</code> .
<code>options</code>	<code>host</code>	integer array to configure <code>metis</code> .

Output

parameter	*Host MemSpace	description
<code>p</code>	<code>host</code>	permutation vector of size <code>n</code> .

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_ALLOC_FAILED</code>	the resources could not be allocated.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed (<code>n</code> , <code>nnzA</code> ≤ 0), base index is not 0 or 1.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.
<code>CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED</code>	the matrix type is not supported.

2.5.3.5. cusolverSpXcsrzd()

```

cusolverStatus_t
cusolverSpScsrzfdHost(
    cusolverSpHandle_t handle,
    int n,
    int nnzA,
    const cusparseMatDescr_t descrA,
    const float *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    int *P,
    int *numnz)

cusolverStatus_t
cusolverSpDcsrzdHost(
    cusolverSpHandle_t handle,
    int n,
    int nnzA,
    const cusparseMatDescr_t descrA,
    const double *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    int *P,
    int *numnz)

cusolverStatus_t
cusolverSpCcsrzdHost(
    cusolverSpHandle_t handle,
    int n,
    int nnzA,
    const cusparseMatDescr_t descrA,
    const cuComplex *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    int *P,
    int *numnz)

cusolverStatus_t
cusolverSpZcsrzdHost(
    cusolverSpHandle_t handle,
    int n,
    int nnzA,
    const cusparseMatDescr_t descrA,
    const cuDoubleComplex *csrValA,
    const int *csrRowPtrA,
    const int *csrColIndA,
    int *P,
    int *numnz)

```

This function implements MC21, zero-free diagonal algorithm. It returns a permutation vector **p** such that **A(p, :)** has no zero diagonal.

A is an **n×n** sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA**, and **csrColIndA**. The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**.

The output parameter **p** is an integer array of **n** elements. It represents a permutation array with base-0 index. The permutation array **p** corresponds to a permutation matrix **P**, and satisfies the following relation:

$$A(p,:) = P * A$$

The output parameter **numnz** describes number of nonzero diagonal in permuted matrix **A(p, :)**. If **numnz** is less than **n**, matrix **A** has structural singularity.

Remark 1: only CPU (Host) path is provided.

Remark 2: this routine does not maximize diagonal value of permuted matrix. The user cannot expect this routine can make "LU without pivoting" stable.

Input

parameter	*Host MemSpace	description
handle	host	handle to the cuSolverSP library context.
n	host	number of rows and columns of matrix A .
nnzA	host	number of nonzeros of matrix A . It is the size of csrValA and csrColIndA .
descrA	host	the descriptor of matrix A . The supported matrix type is CUSPARSE_MATRIX_TYPE_GENERAL . Also, the supported index bases are CUSPARSE_INDEX_BASE_ZERO and CUSPARSE_INDEX_BASE_ONE .
csrValA	host	<type> array of nnzA (= csrRowPtrA (m) - csrRowPtrA (0)) nonzero elements of matrix A .
csrRowPtrA	host	integer array of n+1 elements that contains the start of every row and the end of the last row plus one.
csrColIndA	host	integer array of nnzA column indices of the nonzero elements of matrix A .

Output

parameter	*Host MemSpace	description
p	host	permutation vector of size n .
numnz	host	number of nonzeros on diagonal of permuted matrix.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed (n , nnzA ≤ 0), base index is not 0 or 1.
CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.

CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED	the matrix type is not supported.

2.5.3.6. cusolverSpXcsrperm()

```
cusolverStatus_t
cusolverSpXcsrperm_bufferSizeHost(cusolverSpHandle_t handle,
                                   int m,
                                   int n,
                                   int nnzA,
                                   const cusparseMatDescr_t descrA,
                                   int *csrRowPtrA,
                                   int *csrColIndA,
                                   const int *p,
                                   const int *q,
                                   size_t *bufferSizeInBytes);

cusolverStatus_t
cusolverSpXcsrpermHost(cusolverSpHandle_t handle,
                       int m,
                       int n,
                       int nnzA,
                       const cusparseMatDescr_t descrA,
                       int *csrRowPtrA,
                       int *csrColIndA,
                       const int *p,
                       const int *q,
                       int *map,
                       void *pBuffer);
```

Given a left permutation vector **p** which corresponds to permutation matrix **P** and a right permutation vector **q** which corresponds to permutation matrix **Q**, this function computes permutation of matrix **A** by

$$B = P * A * Q^T$$

A is an **m×n** sparse matrix that is defined in CSR storage format by the three arrays **csrValA**, **csrRowPtrA** and **csrColIndA**.

The operation is in-place, i.e. the matrix **A** is overwritten by **B**.

The permutation vector **p** and **q** are base 0. **p** performs row permutation while **q** performs column permutation. One can also use MATLAB command $B = A(p,q)$ to permute matrix **A**.

This function only computes sparsity pattern of **B**. The user can use parameter **map** to get **csrValB** as well. The parameter **map** is an input/output. If the user sets **map=0:1:(nnzA-1)** before calling **csrperm**, **csrValB=csrValA(map)**.

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**. If **A** is symmetric and only lower/upper part is provided, the user has to pass $A + A^T$ into this function.

This function requires a buffer size returned by `csrperm_bufferSize()`. The address of `pBuffer` must be a multiple of 128 bytes. If it is not, `CUSOLVER_STATUS_INVALID_VALUE` is returned.

For example, if matrix **A** is

$$A = \begin{pmatrix} 1.0 & 2.0 & 3.0 \\ 4.0 & 5.0 & 6.0 \\ 7.0 & 8.0 & 9.0 \end{pmatrix}$$

and left permutation vector **p** = (0, 2, 1), right permutation vector **q** = (2, 1, 0), then $P^*A^*Q^T$ is

$$P^*A^*Q^T = \begin{pmatrix} 3.0 & 2.0 & 1.0 \\ 9.0 & 8.0 & 7.0 \\ 6.0 & 5.0 & 4.0 \end{pmatrix}$$

Remark 1: only CPU (Host) path is provided.

Remark 2: the user can combine `csrsymrcm` and `csrperm` to get $P^*A^*P^T$ which has less zero fill-in during QR factorization.

Input

parameter	cusolverSp MemSpace	description
<code>handle</code>	host	handle to the cuSolver library context.
<code>m</code>	host	number of rows of matrix A .
<code>n</code>	host	number of columns of matrix A .
<code>nnzA</code>	host	number of nonzeros of matrix A . It is the size of <code>csrValA</code> and <code>csrColIndA</code> .
<code>descrA</code>	host	the descriptor of matrix A . The supported matrix type is <code>CUSPARSE_MATRIX_TYPE_GENERAL</code> . Also, the supported index bases are <code>CUSPARSE_INDEX_BASE_ZERO</code> and <code>CUSPARSE_INDEX_BASE_ONE</code> .
<code>csrRowPtrA</code>	host	integer array of <code>m+1</code> elements that contains the start of every row and end of last row plus one of matrix A .
<code>csrColIndA</code>	host	integer array of <code>nnzA</code> column indices of the nonzero elements of matrix A .
<code>p</code>	host	left permutation vector of size <code>m</code> .
<code>q</code>	host	right permutation vector of size <code>n</code> .
<code>map</code>	host	integer array of <code>nnzA</code> indices. If the user wants to get relationship between A and B , <code>map</code> must be set <code>0:1:(nnzA-1)</code> .
<code>pBuffer</code>	host	buffer allocated by the user, the size is returned by <code>csrperm_bufferSize()</code> .

Output

parameter	hsolver	description
<code>csrRowPtrA</code>	host	integer array of <code>m+1</code> elements that contains the start of every row and end of last row plus one of matrix <code>B</code> .
<code>csrColIndA</code>	host	integer array of <code>nnzA</code> column indices of the nonzero elements of matrix <code>B</code> .
<code>map</code>	host	integer array of <code>nnzA</code> indices that maps matrix <code>A</code> to matrix <code>B</code> .
<code>pBufferSizeInBytes</code>	host	number of bytes of the buffer.

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_ALLOC_FAILED</code>	the resources could not be allocated.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed (<code>m, n, nnzA <= 0</code>), base index is not 0 or 1.
<code>CUSOLVER_STATUS_ARCH_MISMATCH</code>	the device only supports compute capability 2.0 and above.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.
<code>CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED</code>	the matrix type is not supported.

2.5.3.7. `cusolverSpXcsrqrBatched()`

The create and destroy methods start and end the lifetime of a `csrqrInfo` object.

```
cusolverStatus_t
cusolverSpCreateCsrqrInfo(csrqrInfo_t *info);

cusolverStatus_t
cusolverSpDestroyCsrqrInfo(csrqrInfo_t info);
```


Analysis is the same for all data types, but each data type has a unique buffer size.

```
cusolverStatus_t
cusolverSpXcsrqrAnalysisBatched(cusolverSpHandle_t handle,
                                int m,
                                int n,
                                int nnzA,
                                const cusparseMatDescr_t descrA,
                                const int *csrRowPtrA,
                                const int *csrColIndA,
                                csrqrInfo_t info);

cusolverStatus_t
cusolverSpScsrqrBufferInfoBatched(cusolverSpHandle_t handle,
                                   int m,
                                   int n,
                                   int nnzA,
                                   const cusparseMatDescr_t descrA,
                                   const float *csrValA,
                                   const int *csrRowPtrA,
                                   const int *csrColIndA,
                                   int batchSize,
                                   csrqrInfo_t info,
                                   size_t *internalDataInBytes,
                                   size_t *workspaceInBytes);

cusolverStatus_t
cusolverSpDcsrqrBufferInfoBatched(cusolverSpHandle_t handle,
                                   int m,
                                   int n,
                                   int nnzA,
                                   const cusparseMatDescr_t descrA,
                                   const double *csrValA,
                                   const int *csrRowPtrA,
                                   const int *csrColIndA,
                                   int batchSize,
                                   csrqrInfo_t info,
                                   size_t *internalDataInBytes,
                                   size_t *workspaceInBytes);
```

Calculate buffer sizes for complex valued data types.

```
cusolverStatus_t
cusolverSpCcsrqrBufferInfoBatched(cusolverSpHandle_t handle,
                                   int m,
                                   int n,
                                   int nnzA,
                                   const cusparseMatDescr_t descrA,
                                   const cuComplex *csrValA,
                                   const int *csrRowPtrA,
                                   const int *csrColIndA,
                                   int batchSize,
                                   csrqrInfo_t info,
                                   size_t *internalDataInBytes,
                                   size_t *workspaceInBytes);

cusolverStatus_t
cusolverSpZcsrqrBufferInfoBatched(cusolverSpHandle_t handle,
                                   int m,
                                   int n,
                                   int nnzA,
                                   const cusparseMatDescr_t descrA,
                                   const cuDoubleComplex *csrValA,
                                   const int *csrRowPtrA,
                                   const int *csrColIndA,
                                   int batchSize,
                                   csrqrInfo_t info,
                                   size_t *internalDataInBytes,
                                   size_t *workspaceInBytes);
```

The S and D data types are real valued single and double precision, respectively.

```
cusolverStatus_t
cusolverSpScsrqrsvBatched(cusolverSpHandle_t handle,
                           int m,
                           int n,
                           int nnzA,
                           const cusparseMatDescr_t descrA,
                           const float *csrValA,
                           const int *csrRowPtrA,
                           const int *csrColIndA,
                           const float *b,
                           float *x,
                           int batchSize,
                           csrqInfo_t info,
                           void *pBuffer);

cusolverStatus_t
cusolverSpDcsrqrsvBatched(cusolverSpHandle_t handle,
                           int m,
                           int n,
                           int nnz,
                           const cusparseMatDescr_t descrA,
                           const double *csrValA,
                           const int *csrRowPtrA,
                           const int *csrColIndA,
                           const double *b,
                           double *x,
                           int batchSize,
                           csrqInfo_t info,
                           void *pBuffer);
```

The C and Z data types are complex valued single and double precision, respectively.

```
cusolverStatus_t
cusolverSpCcsrqrsvBatched(cusolverSpHandle_t handle,
                           int m,
                           int n,
                           int nnzA,
                           const cusparseMatDescr_t descrA,
                           const cuComplex *csrValA,
                           const int *csrRowPtrA,
                           const int *csrColIndA,
                           const cuComplex *b,
                           cuComplex *x,
                           int batchSize,
                           csrqrInfo_t info,
                           void *pBuffer);

cusolverStatus_t
cusolverSpZcsrqrsvBatched(cusolverSpHandle_t handle,
                           int m,
                           int n,
                           int nnzA,
                           const cusparseMatDescr_t descrA,
                           const cuDoubleComplex *csrValA,
                           const int *csrRowPtrA,
                           const int *csrColIndA,
                           const cuDoubleComplex *b,
                           cuDoubleComplex *x,
                           int batchSize,
                           csrqrInfo_t info,
                           void *pBuffer);
```

The batched sparse QR factorization is used to solve either a set of least-squares problems

$$x_j = \operatorname{argmin} \|A_j z - b_j\|, j = 1, 2, \dots, \text{batchSize}$$

or a set of linear systems

$$A_j x_j = b_j, j = 1, 2, \dots, \text{batchSize}$$

where each A_j is a $m \times n$ sparse matrix that is defined in CSR storage format by the four arrays **csrValA**, **csrRowPtrA** and **csrColIndA**.

The supported matrix type is **CUSPARSE_MATRIX_TYPE_GENERAL**. If **A** is symmetric and only lower/upper part is provided, the user has to pass $A + A^H$ into this function.

The prerequisite to use batched sparse QR has two-folds. First all matrices A_j must have the same sparsity pattern. Second, no column pivoting is used in least-square problem, so the solution is valid only if A_j is of full rank for all $j = 1, 2, \dots, \text{batchSize}$. All matrices have the same sparsity pattern, so only one copy of **csrRowPtrA** and **csrColIndA** is used. But the array **csrValA** stores coefficients of A_j one after another. In other words, **csrValA**[$k * \text{nnzA} : (k+1) * \text{nnzA}$] is the value of A_k .

The batched QR uses opaque data structure **csrqrInfo** to keep intermediate data, for example, matrix **Q** and matrix **R** of QR factorization. The user needs to create **csrqrInfo** first by **cusolverSpCreateCsrqrInfo** before any function in batched QR operation.

The **csrqrInfo** would not release internal data until **cusolverSpDestroyCsrqrInfo** is called.

There are three routines in batched sparse QR, **cusolverSpXcsrqrAnalysisBatched**, **cusolverSp[S|D|C|Z]csrqrBufferInfoBatched** and **cusolverSp[S|D|C|Z]csrqrsvBatched**.

First, **cusolverSpXcsrqrAnalysisBatched** is the analysis phase, used to analyze sparsity pattern of matrix **Q** and matrix **R** of QR factorization. Also parallelism is extracted during analysis phase. Once analysis phase is done, the size of working space to perform QR is known. However **cusolverSpXcsrqrAnalysisBatched** uses CPU to analyze the structure of matrix **A**, and this may consume a lot of memory. If host memory is not sufficient to finish the analysis, **CUSOLVER_STATUS_ALLOC_FAILED** is returned. The required memory for analysis is proportional to zero fill-in in QR factorization. The user may need to perform some kind of reordering to minimize zero fill-in, for example, **colamd** or **symrcm** in MATLAB. **cuSolverSP** library provides **symrcm** (**cusolverSpXcsrsymrcm**).

Second, the user needs to choose proper **batchSize** and to prepare working space for sparse QR. There are two memory blocks used in batched sparse QR. One is internal memory block used to store matrix **Q** and matrix **R**. The other is working space used to perform numerical factorization. The size of the former is proportional to **batchSize**, and the size is specified by returned parameter **internalDataInBytes** of **cusolverSp[S|D|C|Z]csrqrBufferInfoBatched**. while the size of the latter is almost independent of **batchSize**, and the size is specified by returned parameter **workspaceInBytes** of **cusolverSp[S|D|C|Z]csrqrBufferInfoBatched**. The internal memory block is allocated implicitly during first call of **cusolverSp[S|D|C|Z]csrqrsvBatched**. The user only needs to allocate working space for **cusolverSp[S|D|C|Z]csrqrsvBatched**.

Instead of trying all batched matrices, the user can find maximum **batchSize** by querying **cusolverSp[S|D|C|Z]csrqrBufferInfoBatched**. For example, the user can increase **batchSize** till summation of **internalDataInBytes** and **workspaceInBytes** is greater than size of available device memory.

Suppose that the user needs to perform 253 linear solvers and available device memory is 2GB. if **cusolverSp[S|D|C|Z]csrqrsvBatched** can only afford **batchSize** 100, the user has to call **cusolverSp[S|D|C|Z]csrqrsvBatched** three times to finish all. The user calls **cusolverSp[S|D|C|Z]csrqrBufferInfoBatched** with **batchSize** 100. The opaque **info** would remember this **batchSize** and any subsequent call of **cusolverSp[S|D|C|Z]csrqrsvBatched** cannot exceed this value. In this example, the first two calls of **cusolverSp[S|D|C|Z]csrqrsvBatched** will use **batchSize** 100, and last call of **cusolverSp[S|D|C|Z]csrqrsvBatched** will use **batchSize** 53.

Example: suppose that A_0, A_1, \dots, A_9 have the same sparsity pattern, the following code solves 10 linear systems $A_j x_j = b_j, j = 0, 2, \dots, 9$ by batched sparse QR.

```
// Suppose that A0, A1, ..., A9 are m x m sparse matrix represented by CSR
// format,
// Each matrix Aj has nonzero nnzA, and shares the same csrRowPtrA and
// csrColIndA.
// csrValA is aggregation of A0, A1, ..., A9.
int m ; // number of rows and columns of each Aj
int nnzA ; // number of nonzeros of each Aj
int *csrRowPtrA ; // each Aj has the same csrRowPtrA
int *csrColIndA ; // each Aj has the same csrColIndA
double *csrValA ; // aggregation of A0,A1,...,A9
const int batchSize = 10; // 10 linear systems

cusolverSpHandle_t handle; // handle to cusolver library
csrqrInfo_t info = NULL;
cusparsMatDescr_t descrA = NULL;
void *pBuffer = NULL; // working space for numerical factorization

// step 1: create a descriptor
cusparsCreateMatDescr(&descrA);
cusparsSetMatIndexBase(descrA, CUSPARSE_INDEX_BASE_ONE); // A is base-1
cusparsSetMatType(descrA, CUSPARSE_MATRIX_TYPE_GENERAL); // A is a general
// matrix

// step 2: create empty info structure
cusolverSpCreateCsrqrInfo(&info);

// step 3: symbolic analysis
cusolverSpXcsrqrAnalysisBatched(
    handle, m, m, nnzA,
    descrA, csrRowPtrA, csrColIndA, info);

// step 4: allocate working space for Aj*xj=bj
cusolverSpDcsrqrBufferInfoBatched(
    handle, m, m, nnzA,
    descrA,
    csrValA, csrRowPtrA, csrColIndA,
    batchSize,
    info,
    &internalDataInBytes,
    &workspaceInBytes);

cudaMalloc(&pBuffer, workspaceInBytes);

// step 5: solve Aj*xj = bj
cusolverSpDcsrqrsvBatched(
    handle, m, m, nnzA,
    descrA, csrValA, csrRowPtrA, csrColIndA,
    b,
    x,
    batchSize,
    info,
    pBuffer);

// step 7: destroy info
cusolverSpDestroyCsrqrInfo(info);
```

Please refer to Appendix B for detailed examples.

Remark 1: only GPU (device) path is provided.

Input

parameter	cusolverSp MemSpace	description
handle	host	handle to the cuSolverSP library context.
m	host	number of rows of each matrix A_j .
n	host	number of columns of each matrix A_j .
nnzA	host	number of nonzeros of each matrix A_j . It is the size <code>csrColIndA</code> .
descrA	host	the descriptor of each matrix A_j . The supported matrix type is <code>CUSPARSE_MATRIX_TYPE_GENERAL</code> . Also, the supported index bases are <code>CUSPARSE_INDEX_BASE_ZERO</code> and <code>CUSPARSE_INDEX_BASE_ONE</code> .
csrValA	device	<type> array of <code>nnzA*batchSize</code> nonzero elements of matrices A_0, A_1, \dots . All matrices are aggregated one after another.
csrRowPtrA	device	integer array of <code>m+1</code> elements that contains the start of every row and the end of the last row plus one.
csrColIndA	device	integer array of <code>nnzA</code> column indices of the nonzero elements of each matrix A_j .
b	device	<type> array of <code>m*batchSize</code> of right-hand-side vectors b_0, b_1, \dots . All vectors are aggregated one after another.
batchSize	host	number of systems to be solved.
info	host	opaque structure for QR factorization.
pBuffer	device	buffer allocated by the user, the size is returned by <code>cusolverSpXcsrqrBufferInfoBatched()</code> .

Output

parameter	cusolverSp MemSpace	description
x	device	<type> array of <code>m*batchSize</code> of solution vectors x_0, x_1, \dots . All vectors are aggregated one after another.
internalDataInBytes	host	number of bytes of the internal data.
workspaceInBytes	host	number of bytes of the buffer in numerical factorization.

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_ALLOC_FAILED</code>	the resources could not be allocated.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed (<code>m, n, nnzA <= 0</code>), base index is not 0 or 1.

CUSOLVER_STATUS_ARCH_MISMATCH	the device only supports compute capability 2.0 and above.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.
CUSOLVER_STATUS_MATRIX_TYPE_NOT_SUPPORTED	the matrix type is not supported.

2.6. cuSolverRF: Refactorization Reference

This chapter describes API of cuSolverRF, a library for fast refactorization.

2.6.1. cusolverRfAccessBundledFactors()

```
cusolverStatus_t
cusolverRfAccessBundledFactors(/* Input */
                               cusolverRfHandle_t handle,
                               /* Output (in the host memory) */
                               int* nnzM,
                               /* Output (in the device memory) */
                               int** Mp,
                               int** Mi,
                               double** Mx);
```

This routine allows direct access to the lower **L** and upper **U** triangular factors stored in the cuSolverRF library handle. The factors are compressed into a single matrix **M**= (**L**-**I**)+**U**, where the unitary diagonal of **L** is not stored. It is assumed that a prior call to the **cusolverRfRefactor()** was done in order to generate these triangular factors.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
nnzM	host	output	the number of non-zero elements of matrix M .
Mp	device	output	the array of offsets corresponding to the start of each row in the arrays Mi and Mx . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix M . The array size is n+1 .
Mi	device	output	the array of column indices corresponding to the non-zero elements in the matrix M . It is assumed that this array is sorted by row and by column within each row. The array size is nnzM .
Mx	device	output	the array of values corresponding to the non-zero elements in the matrix M . It is assumed that this array is sorted by row and by column within each row. The array size is nnzM .

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.

2.6.2. cusolverRfAnalyze()

```
cusolverStatus_t
cusolverRfAnalyze(cusolverRfHandle_t handle);
```

This routine performs the appropriate analysis of parallelism available in the LU re-factorization depending upon the algorithm chosen by the user.

$$A = L * U$$

It is assumed that a prior call to the **cusolverRfSetup[Host|Device] ()** was done in order to create internal data structures needed for the analysis.

This routine needs to be called only once for a single linear system

$$A_i x_i = f_i$$

parameter	MemSpace	In/out	Meaning
handle	host	in/out	the handle to the cuSolverRF library.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.
CUSOLVER_STATUS_ALLOC_FAILED	an allocation of memory failed.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.6.3. cusolverRfSetupDevice()

```
cusolverStatus_t
cusolverRfSetupDevice(/* Input (in the device memory) */
    int n,
    int nnzA,
    int* csrRowPtrA,
    int* csrColIndA,
    double* csrValA,
    int nnzL,
    int* csrRowPtrL,
    int* csrColIndL,
    double* csrValL,
    int nnzU,
    int* csrRowPtrU,
    int* csrColIndU,
    double* csrValU,
    int* P,
    int* Q,
    /* Output */
    cusolverRfHandle_t handle);
```

This routine assembles the internal data structures of the cuSolverRF library. It is often the first routine to be called after the call to the **cusolverRfCreate()** routine.

This routine accepts as input (on the device) the original matrix **A**, the lower (**L**) and upper (**U**) triangular factors, as well as the left (**P**) and the right (**Q**) permutations resulting from the full LU factorization of the first (**i=1**) linear system

$$A_i x_i = f_i$$

The permutations **P** and **Q** represent the final composition of all the left and right reorderings applied to the original matrix **A**, respectively. However, these permutations are often associated with partial pivoting and reordering to minimize fill-in, respectively.

This routine needs to be called only once for a single linear system

$$A_i x_i = f_i$$

parameter	MemSpace	In/out	Meaning
n	host	input	the number of rows (and columns) of matrix A .
nnzA	host	input	the number of non-zero elements of matrix A .
csrRowPtrA	device	input	the array of offsets corresponding to the start of each row in the arrays csrColIndA and csrValA . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix. The array size is n+1 .
csrColIndA	device	input	the array of column indices corresponding to the non-zero elements in the matrix. It

			is assumed that this array is sorted by row and by column within each row. The array size is nnzA .
csrValA	device	input	the array of values corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by row and by column within each row. The array size is nnzA .
nnzL	host	input	the number of non-zero elements of matrix L .
csrRowPtrL	device	input	the array of offsets corresponding to the start of each row in the arrays csrColIndL and csrValL . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix L . The array size is n+1 .
csrColIndL	device	input	the array of column indices corresponding to the non-zero elements in the matrix L . It is assumed that this array is sorted by row and by column within each row. The array size is nnzL .
csrValL	device	input	the array of values corresponding to the non-zero elements in the matrix L . It is assumed that this array is sorted by row and by column within each row. The array size is nnzL .
nnzU	host	input	the number of non-zero elements of matrix U .
csrRowPtrU	device	input	the array of offsets corresponding to the start of each row in the arrays csrColIndU and csrValU . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix U . The array size is n+1 .
csrColIndU	device	input	the array of column indices corresponding to the non-zero elements in the matrix U . It is assumed that this array is sorted by row and by column within each row. The array size is nnzU .
csrValU	device	input	the array of values corresponding to the non-zero elements in the matrix U . It is assumed that this array is sorted by row and by column within each row. The array size is nnzU .
P	device	input	the left permutation (often associated with pivoting). The array size is n .
Q	device	input	the right permutation (often associated with reordering). The array size is n .
handle	host	output	the handle to the GLU library.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	an unsupported value or parameter was passed.
CUSOLVER_STATUS_ALLOC_FAILED	an allocation of memory failed.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.6.4. cusolverRfSetupHost()

```
cusolverStatus_t
cusolverRfSetupHost(/* Input (in the host memory) */
    int n,
    int nnzA,
    int* h_csrRowPtrA,
    int* h_csrColIndA,
    double* h_csrValA,
    int nnzL,
    int* h_csrRowPtrL,
    int* h_csrColIndL,
    double* h_csrValL,
    int nnzU,
    int* h_csrRowPtrU,
    int* h_csrColIndU,
    double* h_csrValU,
    int* h_P,
    int* h_Q,
    /* Output */
    cusolverRfHandle_t handle);
```

This routine assembles the internal data structures of the cuSolverRF library. It is often the first routine to be called after the call to the **cusolverRfCreate()** routine.

This routine accepts as input (on the host) the original matrix **A**, the lower (**L**) and upper (**U**) triangular factors, as well as the left (**P**) and the right (**Q**) permutations resulting from the full LU factorization of the first (**i=1**) linear system

$$A_i x_i = f_i$$

The permutations **P** and **Q** represent the final composition of all the left and right reorderings applied to the original matrix **A**, respectively. However, these permutations are often associated with partial pivoting and reordering to minimize fill-in, respectively.

This routine needs to be called only once for a single linear system

$$A_i x_i = f_i$$

parameter	MemSpace	In/out	Meaning
n	host	input	the number of rows (and columns) of matrix A .

<code>nnzA</code>	<code>host</code>	<code>input</code>	the number of non-zero elements of matrix A .
<code>h_csrRowPtrA</code>	<code>host</code>	<code>input</code>	the array of offsets corresponding to the start of each row in the arrays <code>h_csrColIndA</code> and <code>h_csrValA</code> . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix. The array size is <code>n+1</code> .
<code>h_csrColIndA</code>	<code>host</code>	<code>input</code>	the array of column indices corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by row and by column within each row. The array size is <code>nnzA</code> .
<code>h_csrValA</code>	<code>host</code>	<code>input</code>	the array of values corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by row and by column within each row. The array size is <code>nnzA</code> .
<code>nnzL</code>	<code>host</code>	<code>input</code>	the number of non-zero elements of matrix L .
<code>h_csrRowPtrL</code>	<code>host</code>	<code>input</code>	the array of offsets corresponding to the start of each row in the arrays <code>h_csrColIndL</code> and <code>h_csrValL</code> . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix L . The array size is <code>n+1</code> .
<code>h_csrColIndL</code>	<code>host</code>	<code>input</code>	the array of column indices corresponding to the non-zero elements in the matrix L . It is assumed that this array is sorted by row and by column within each row. The array size is <code>nnzL</code> .
<code>h_csrValL</code>	<code>host</code>	<code>input</code>	the array of values corresponding to the non-zero elements in the matrix L . It is assumed that this array is sorted by row and by column within each row. The array size is <code>nnzL</code> .
<code>nnzU</code>	<code>host</code>	<code>input</code>	the number of non-zero elements of matrix U .
<code>h_csrRowPtrU</code>	<code>host</code>	<code>input</code>	the array of offsets corresponding to the start of each row in the arrays <code>h_csrColIndU</code> and <code>h_csrValU</code> . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix U . The array size is <code>n+1</code> .
<code>h_csrColIndU</code>	<code>host</code>	<code>input</code>	the array of column indices corresponding to the non-zero elements in the matrix U . It is assumed that this array is sorted by row and by column within each row. The array size is <code>nnzU</code> .

h_csrValU	host	input	the array of values corresponding to the non-zero elements in the matrix U . It is assumed that this array is sorted by row and by column within each row. The array size is $nnzU$.
h_P	host	input	the left permutation (often associated with pivoting). The array size in n .
h_Q	host	input	the right permutation (often associated with reordering). The array size in n .
handle	host	output	the handle to the cuSolverRF library.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	an unsupported value or parameter was passed.
CUSOLVER_STATUS_ALLOC_FAILED	an allocation of memory failed.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.6.5. cusolverRfCreate()

```
cusolverStatus_t cusolverRfCreate(cusolverRfHandle_t *handle);
```

This routine initializes the cuSolverRF library. It allocates required resources and must be called prior to any other cuSolverRF library routine.

parameter	MemSpace	In/out	Meaning
handle	host	output	the pointer to the cuSolverRF library handle.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	an allocation of memory failed.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.6.6. cusolverRfExtractBundledFactorsHost()

```
cusolverStatus_t
cusolverRfExtractBundledFactorsHost( /* Input */
    cusolverRfHandle_t handle,
    /* Output (in the host memory) */
    int* h_nnzM,
    int** h_Mp,
    int** h_Mi,
    double** h_Mx);
```

This routine extracts lower (**L**) and upper (**U**) triangular factors from the cuSolverRF library handle into the host memory. The factors are compressed into a single matrix **M** = (**L** - **I**) + **U**, where the unitary diagonal of (**L**) is not stored. It is assumed that a prior call to the `cusolverRfRefactor()` was done in order to generate these triangular factors.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
h_nnzM	host	output	the number of non-zero elements of matrix M .
h_Mp	host	output	the array of offsets corresponding to the start of each row in the arrays h_Mi and h_Mx . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix M . The array size is n+1 .
h_Mi	host	output	the array of column indices corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by row and by column within each row. The array size is h_nnzM .
h_Mx	host	output	the array of values corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by row and by column within each row. The array size is h_nnzM .

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_ALLOC_FAILED	an allocation of memory failed.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.

2.6.7. cusolverRfExtractSplitFactorsHost()

```
cusolverStatus_t
cusolverRfExtractSplitFactorsHost(/* Input */
                                  cusolverRfHandle_t handle,
                                  /* Output (in the host memory) */
                                  int* h_nnzL,
                                  int** h_Lp,
                                  int** h_Li,
                                  double** h_Lx,
                                  int* h_nnzU,
                                  int** h_Up,
                                  int** h_Ui,
                                  double** h_Ux);
```

This routine extracts lower (**L**) and upper (**U**) triangular factors from the cuSolverRF library handle into the host memory. It is assumed that a prior call to the **cusolverRfRefactor()** was done in order to generate these triangular factors.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
h_nnzL	host	output	the number of non-zero elements of matrix L .
h_Lp	host	output	the array of offsets corresponding to the start of each row in the arrays h_Li and h_Lx . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix L . The array size is n+1 .
h_Li	host	output	the array of column indices corresponding to the non-zero elements in the matrix L . It is assumed that this array is sorted by row and by column within each row. The array size is h_nnzL .
h_Lx	host	output	the array of values corresponding to the non-zero elements in the matrix L . It is assumed that this array is sorted by row and by column within each row. The array size is h_nnzL .
h_nnzU	host	output	the number of non-zero elements of matrix U .
h_Up	host	output	the array of offsets corresponding to the start of each row in the arrays h_Ui and h_Ux . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix U . The array size is n+1 .
h_Ui	host	output	the array of column indices corresponding to the non-zero elements in the matrix U . It is assumed that this array is sorted by

			row and by column within each row. The array size is <code>h_nnzU</code> .
<code>h_Ux</code>	<code>host</code>	<code>output</code>	the array of values corresponding to the non-zero elements in the matrix <code>U</code> . It is assumed that this array is sorted by row and by column within each row. The array size is <code>h_nnzU</code> .

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_ALLOC_FAILED</code>	an allocation of memory failed.
<code>CUSOLVER_STATUS_EXECUTION_FAILED</code>	a kernel failed to launch on the GPU.

2.6.8. `cusolverRfDestroy()`

```
cusolverStatus_t cusolverRfDestroy(cusolverRfHandle_t handle);
```

This routine shuts down the cuSolverRF library. It releases acquired resources and must be called after all the cuSolverRF library routines.

parameter	MemSpace	In/out	Meaning
<code>handle</code>	<code>host</code>	<code>input</code>	the cuSolverRF library handle.

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.

2.6.9. `cusolverRfGetMatrixFormat()`

```
cusolverStatus_t
cusolverRfGetMatrixFormat(cusolverRfHandle_t handle,
                          cusolverRfMatrixFormat_t *format,
                          cusolverRfUnitDiagonal_t *diag);
```

This routine gets the matrix format used in the `cusolverRfSetupDevice()`, `cusolverRfSetupHost()`, `cusolverRfResetValues()`, `cusolverRfExtractBundledFactorsHost()` and `cusolverRfExtractSplitFactorsHost()` routines.

parameter	MemSpace	In/out	Meaning
<code>handle</code>	<code>host</code>	<code>input</code>	the handle to the cuSolverRF library.
<code>format</code>	<code>host</code>	<code>output</code>	the enumerated matrix format type.
<code>diag</code>	<code>host</code>	<code>output</code>	the enumerated unit diagonal type.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

2.6.10. cusolverRfGetNumericProperties()

```
cusolverStatus_t
cusolverRfGetNumericProperties(cusolverRfHandle_t handle,
                              double *zero,
                              double *boost);
```

This routine gets the numeric values used for checking for "zero" pivot and for boosting it in the **`cusolverRfRefactor()`** and **`cusolverRfSolve()`** routines. The numeric boosting will be used only if **`boost > 0.0`**.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
zero	host	output	the value below which zero pivot is flagged.
boost	host	output	the value which is substituted for zero pivot (if the later is flagged).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

2.6.11. cusolverRfGetNumericBoostReport()

```
cusolverStatus_t
cusolverRfGetNumericBoostReport(cusolverRfHandle_t handle,
                                cusolverRfNumericBoostReport_t *report);
```

This routine gets the report whether numeric boosting was used in the **`cusolverRfRefactor()`** and **`cusolverRfSolve()`** routines.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
report	host	output	the enumerated boosting report type.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

2.6.12. cusolverRfGetResetValuesFastMode()

```
cusolverStatus_t
cusolverRfGetResetValuesFastMode(cusolverRfHandle_t handle,
                                rfResetValuesFastMode_t *fastMode);
```

This routine gets the mode used in the **`cusolverRfResetValues`** routine.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
fastMode	host	output	the enumerated mode type.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

2.6.13. cusolverRfGet_Algs()

```
cusolverStatus_t
cusolverRfGet_Algs(cusolverRfHandle_t handle,
                   cusolverRfFactorization_t* fact_alg,
                   cusolverRfTriangularSolve_t* solve_alg);
```

This routine gets the algorithm used for the refactorization in **`cusolverRfRefactor()`** and the triangular solve in **`cusolverRfSolve()`**.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
alg	host	output	the enumerated algorithm type.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

2.6.14. cusolverRfRefactor()

```
cusolverStatus_t cusolverRfRefactor(cusolverRfHandle_t handle);
```

This routine performs the LU re-factorization

$$A = L * U$$

exploring the available parallelism on the GPU. It is assumed that a prior call to the **`glu_analyze()`** was done in order to find the available parallelism.

This routine may be called multiple times, once for each of the linear systems

$$A_i x_i = f_i$$

parameter	Memory	In/out	Meaning
handle	host	in/out	the handle to the cuSolverRF library.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.
CUSOLVER_STATUS_ZERO_PIVOT	a zero pivot was encountered during the computation.

2.6.15. cusolverRfResetValues()

```
cusolverStatus_t
cusolverRfResetValues(/* Input (in the device memory) */
    int n,
    int nnzA,
    int* csrRowPtrA,
    int* csrColIndA,
    double* csrValA,
    int* P,
    int* Q,
    /* Output */
    cusolverRfHandle_t handle);
```

This routine updates internal data structures with the values of the new coefficient matrix. It is assumed that the arrays **csrRowPtrA**, **csrColIndA**, **P** and **Q** have not changed since the last call to the **cusolverRfSetup[Host|Device]** routine. This assumption reflects the fact that the sparsity pattern of coefficient matrices as well as reordering to minimize fill-in and pivoting remain the same in the set of linear systems

$$A_i x_i = f_i$$

This routine may be called multiple times, once for each of the linear systems

$$A_i x_i = f_i$$

parameter	MemSpace	In/out	Meaning
n	host	input	the number of rows (and columns) of matrix A .
nnzA	host	input	the number of non-zero elements of matrix A .
csrRowPtrA	device	input	the array of offsets corresponding to the start of each row in the arrays csrColIndA and csrValA . This array has also an extra entry at the end that stores

			the number of non-zero elements in the matrix. The array size is $n+1$.
csrColIndA	device	input	the array of column indices corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by row and by column within each row. The array size is nnzA .
csrValA	device	input	the array of values corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by row and by column within each row. The array size is nnzA .
P	device	input	the left permutation (often associated with pivoting). The array size is n .
Q	device	input	the right permutation (often associated with reordering). The array size is n .
handle	host	output	the handle to the cuSolverRF library.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	an unsupported value or parameter was passed.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.

2.6.16. cusolverRfSetMatrixFormat()

```
cusolverStatus_t
cusolverRfSetMatrixFormat(cusolverRfHandle_t handle,
                          gluMatrixFormat_t format,
                          gluUnitDiagonal_t diag);
```

This routine sets the matrix format used in the **cusolverRfSetupDevice()**, **cusolverRfSetupHost()**, **cusolverRfResetValues()**, **cusolverRfExtractBundledFactorsHost()** and **cusolverRfExtractSplitFactorsHost()** routines. It may be called once prior to **cusolverRfSetupDevice()** and **cusolverRfSetupHost()** routines.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
format	host	input	the enumerated matrix format type.
diag	host	input	the enumerated unit diagonal type.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

CUSOLVER_STATUS_INVALID_VALUE	an enumerated mode parameter is wrong.
-------------------------------	--

2.6.17. cusolverRfSetNumericProperties()

```
cusolverStatus_t
cusolverRfSetNumericProperties(cusolverRfHandle_t handle,
                             double zero,
                             double boost);
```

This routine sets the numeric values used for checking for "zero" pivot and for boosting it in the **cusolverRfRefactor()** and **cusolverRfSolve()** routines. It may be called multiple times prior to **cusolverRfRefactor()** and **cusolverRfSolve()** routines. The numeric boosting will be used only if **boost > 0.0**.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
zero	host	input	the value below which zero pivot is flagged.
boost	host	input	the value which is substituted for zero pivot (if the later is flagged).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

2.6.18. cusolverRfSetResetValuesFastMode()

```
cusolverStatus_t
cusolverRfSetResetValuesFastMode(cusolverRfHandle_t handle,
                                 gluResetValuesFastMode_t fastMode);
```

This routine sets the mode used in the **cusolverRfResetValues** routine. The fast mode requires extra memory and is recommended only if very fast calls to **cusolverRfResetValues()** are needed. It may be called once prior to **cusolverRfAnalyze()** routine.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
fastMode	host	input	the enumerated mode type.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	an enumerated mode parameter is wrong.

2.6.19. cusolverRfSetAlgs()

```
cusolverStatus_t
cusolverRfSetAlgs(cusolverRfHandle_t handle,
                  gluFactorization_t fact_alg,
                  gluTriangularSolve_t alg);
```

This routine sets the algorithm used for the refactorization in **`cusolverRfRefactor()`** and the triangular solve in **`cusolverRfSolve()`**. It may be called once prior to **`cusolverRfAnalyze()`** routine.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
alg	host	input	the enumerated algorithm type.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

2.6.20. cusolverRfSolve()

```
cusolverStatus_t
cusolverRfSolve(/* Input (in the device memory) */
               cusolverRfHandle_t handle,
               int *P,
               int *Q,
               int nrhs,
               double *Temp,
               int ldt,
               /* Input/Output (in the device memory) */
               double *XF,
               /* Input */
               int ldxf);
```

This routine performs the forward and backward solve with the lower $L \in R^{n \times n}$ and upper $U \in R^{n \times n}$ triangular factors resulting from the LU re-factorization

$$A = L * U$$

which is assumed to have been computed by a prior call to the **`cusolverRfRefactor()`** routine.

The routine can solve linear systems with multiple right-hand-sides (rhs),

$$AX = (LU)X = L(UX) = LY = F \text{ where } UX = Y$$

even though currently only a single rhs is supported.

This routine may be called multiple times, once for each of the linear systems

$$A_i x_i = f_i$$

parameter	MemSpace	In/out	Meaning
handle	host	output	the handle to the cuSolverRF library.
P	device	input	the left permutation (often associated with pivoting). The array size is n.
Q	device	input	the right permutation (often associated with reordering). The array size is n.
nrhs	host	input	the number right-hand-sides to be solved.
Temp	host	input	the dense matrix that contains temporary workspace (of size ldt*nrhs).
ldt	host	input	the leading dimension of dense matrix Temp (ldt >= n).
XF	host	in/out	the dense matrix that contains the right-hand-sides F and solutions x (of size ldxf*nrhs).
ldxf	host	input	the leading dimension of dense matrix XF (ldxf >= n).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	an unsupported value or parameter was passed.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.6.21. cusolverRfBatchSetupHost()

```
cusolverStatus_t
cusolverRfBatchSetupHost(/* Input (in the host memory) */
    int batchSize,
    int n,
    int nnzA,
    int* h_csrRowPtrA,
    int* h_csrColIndA,
    double *h_csrValA_array[],
    int nnzL,
    int* h_csrRowPtrL,
    int* h_csrColIndL,
    double *h_csrValL,
    int nnzU,
    int* h_csrRowPtrU,
    int* h_csrColIndU,
    double *h_csrValU,
    int* h_P,
    int* h_Q,
    /* Output */
    cusolverRfHandle_t handle);
```

This routine assembles the internal data structures of the cuSolverRF library for batched operation. It is called after the call to the **cusolverRfCreate()** routine, and before any other batched routines.

The batched operation assumes that the user has the following linear systems

$$A_j x_j = b_j, j = 1, 2, \dots, \text{batchSize}$$

where each matrix in the set $\{A_j\}$ has the same sparsity pattern, and quite similar such that factorization can be done by the same permutation **P** and **Q**. In other words, $A_j, j > 1$ is a small perturbation of A_1 .

This routine accepts as input (on the host) the original matrix **A** (sparsity pattern and batched values), the lower (**L**) and upper (**U**) triangular factors, as well as the left (**P**) and the right (**Q**) permutations resulting from the full LU factorization of the first (**i=1**) linear system

$$A_i x_i = f_i$$

The permutations **P** and **Q** represent the final composition of all the left and right reorderings applied to the original matrix **A**, respectively. However, these permutations are often associated with partial pivoting and reordering to minimize fill-in, respectively.

Remark 1: the matrices **A**, **L** and **U** must be CSR format and base-0.

Remark 2: to get best performance, **batchSize** should be multiple of 32 and greater or equal to 32. The algorithm is memory-bound, once bandwidth limit is reached, there is no room to improve performance by large **batchSize**. In practice, **batchSize** of 32 - 128 is often enough to obtain good performance, but in some cases larger **batchSize** might be beneficial.

This routine needs to be called only once for a single linear system

$$A_i x_i = f_i$$

parameter	MemSpace	In/out	Meaning
batchSize	host	input	the number of matrices in the batched mode.
n	host	input	the number of rows (and columns) of matrix A .
nnzA	host	input	the number of non-zero elements of matrix A .
h_csrRowPtrA	host	input	the array of offsets corresponding to the start of each row in the arrays h_csrColIndA and h_csrValA . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix. The array size is n + 1 .
h_csrColIndA	host	input	the array of column indices corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by row and by column within each row. The array size is nnzA .
h_csrValA_array	host	input	array of pointers of size batchSize , each pointer points to the array of values corresponding to the non-zero elements in the matrix.
nnzL	host	input	the number of non-zero elements of matrix L .
h_csrRowPtrL	host	input	the array of offsets corresponding to the start of each row in the arrays h_csrColIndL and h_csrValL . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix L . The array size is n + 1 .
h_csrColIndL	host	input	the array of column indices corresponding to the non-zero elements in the matrix L . It is assumed that this array is sorted by row and by column within each row. The array size is nnzL .
h_csrValL	host	input	the array of values corresponding to the non-zero elements in the matrix L . It is assumed that this array is sorted by row and by column within each row. The array size is nnzL .
nnzU	host	input	the number of non-zero elements of matrix U .
h_csrRowPtrU	host	input	the array of offsets corresponding to the start of each row in the arrays h_csrColIndU and h_csrValU . This

			array has also an extra entry at the end that stores the number of non-zero elements in the matrix u . The array size is $n+1$.
<code>h_csrColIndU</code>	host	input	the array of column indices corresponding to the non-zero elements in the matrix u . It is assumed that this array is sorted by row and by column within each row. The array size is $nnzU$.
<code>h_csrValU</code>	host	input	the array of values corresponding to the non-zero elements in the matrix u . It is assumed that this array is sorted by row and by column within each row. The array size is $nnzU$.
<code>h_P</code>	host	input	the left permutation (often associated with pivoting). The array size in n .
<code>h_Q</code>	host	input	the right permutation (often associated with reordering). The array size in n .
<code>handle</code>	host	output	the handle to the cuSolverRF library.

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_NOT_INITIALIZED</code>	the library was not initialized.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	an unsupported value or parameter was passed.
<code>CUSOLVER_STATUS_ALLOC_FAILED</code>	an allocation of memory failed.
<code>CUSOLVER_STATUS_EXECUTION_FAILED</code>	a kernel failed to launch on the GPU.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

2.6.22. cusolverRfBatchAnalyze()

```
cusolverStatus_t cusolverRfBatchAnalyze(cusolverRfHandle_t handle);
```

This routine performs the appropriate analysis of parallelism available in the batched LU re-factorization.

It is assumed that a prior call to the `cusolverRfBatchSetup[Host]()` was done in order to create internal data structures needed for the analysis.

This routine needs to be called only once for a single linear system

$$A_j x_j = b_j, j = 1, 2, \dots, \text{batchSize}$$

parameter	Memory	In/out	Meaning
<code>handle</code>	host	in/out	the handle to the cuSolverRF library.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.
CUSOLVER_STATUS_ALLOC_FAILED	an allocation of memory failed.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.6.23. cusolverRfBatchResetValues()

```
cusolverStatus_t
cusolverRfBatchResetValues(/* Input (in the device memory) */
    int batchSize,
    int n,
    int nnzA,
    int* csrRowPtrA,
    int* csrColIndA,
    double* csrValA_array[],
    int *P,
    int *Q,
    /* Output */
    cusolverRfHandle_t handle);
```

This routine updates internal data structures with the values of the new coefficient matrix. It is assumed that the arrays **csrRowPtrA**, **csrColIndA**, **P** and **Q** have not changed since the last call to the **cusolverRfbatch_setup_host** routine.

This assumption reflects the fact that the sparsity pattern of coefficient matrices as well as reordering to minimize fill-in and pivoting remain the same in the set of linear systems

$$A_j x_j = b_j, j = 1, 2, \dots, \text{batchSize}$$

The input parameter **csrValA_array** is an array of pointers on device memory. **csrValA_array(j)** points to matrix A_j which is also on device memory.

parameter	MemSpace	In/out	Meaning
batchSize	host	input	the number of matrices in batched mode.
n	host	input	the number of rows (and columns) of matrix A .
nnzA	host	input	the number of non-zero elements of matrix A .
csrRowPtrA	device	input	the array of offsets corresponding to the start of each row in the arrays csrColIndA and csrValA . This array has also an extra entry at the end that stores the number of non-zero elements in the matrix. The array size is n+1 .
csrColIndA	device	input	the array of column indices corresponding to the non-zero elements in the matrix. It is assumed that this array is sorted by row

			and by column within each row. The array size is nnzA .
csrValA_array	device	input	array of pointers of size batchSize , each pointer points to the array of values corresponding to the non-zero elements in the matrix.
P	device	input	the left permutation (often associated with pivoting). The array size in n .
Q	device	input	the right permutation (often associated with reordering). The array size in n .
handle	host	output	the handle to the cuSolverRF library.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	an unsupported value or parameter was passed.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.

2.6.24. cusolverRfBatchRefactor()

```
cusolverStatus_t cusolverRfBatchRefactor(cusolverRfHandle_t handle);
```

This routine performs the LU re-factorization

$$M_j = P^* A_j^* Q^T = L_j^* U_j$$

exploring the available parallelism on the GPU. It is assumed that a prior call to the **cusolverRfBatchAnalyze()** was done in order to find the available parallelism.

Remark: **cusolverRfBatchRefactor()** would not report any failure of LU refactorization. The user has to call **cusolverRfBatchZeroPivot()** to know which matrix failed the LU refactorization.

parameter	Memory	In/out	Meaning
handle	host	in/out	the handle to the cuSolverRF library.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.

2.6.25. cusolverRfBatchSolve()

```
cusolverStatus_t
cusolverRfBatchSolve(/* Input (in the device memory) */
                    cusolverRfHandle_t handle,
                    int *P,
                    int *Q,
                    int nrhs,
                    double *Temp,
                    int ldt,
                    /* Input/Output (in the device memory) */
                    double *XF_array[],
                    /* Input */
                    int ldxf);
```

To solve $A_j * x_j = b_j$, first we reform the equation by $M_j * Q * x_j = P * b_j$ where $M_j = P * A_j * Q^T$. Then do refactorization $M_j = L_j * U_j$ by **cusolverRfBatch_Refactor()**. Further **cusolverRfBatch_Solve()** takes over the remaining steps, including:

$$z_j = P * b_j$$

$$M_j * y_j = z_j$$

$$x_j = Q^T * y_j$$

The input parameter **XF_array** is an array of pointers on device memory. **XF_array(j)** points to matrix x_j which is also on device memory.

Remark 1: only a single rhs is supported.

Remark 2: no singularity is reported during backward solve. If some matrix A_j failed the refactorization and U_j has some zero diagonal, backward solve would compute NAN. The user has to call **cusolverRfBatch_Zero_Pivot** to check if refactorization is successful or not.

parameter	Memory	In/out	Meaning
handle	host	output	the handle to the cuSolverRF library.
P	device	input	the left permutation (often associated with pivoting). The array size in n.
Q	device	input	the right permutation (often associated with reordering). The array size in n.
nrhs	host	input	the number right-hand-sides to be solved.
Temp	host	input	the dense matrix that contains temporary workspace (of size ldt*nrhs).
ldt	host	input	the leading dimension of dense matrix Temp (ldt >= n).
XF_array	device	in/out	array of pointers of size batchSize, each pointer points to the dense matrix

			that contains the right-hand-sides \mathbf{F} and solutions \mathbf{x} (of size $\text{ldxf} \times \text{nrhs}$).
ldxf	host	input	the leading dimension of dense matrix \mathbf{xF} ($\text{ldxf} \geq n$).

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.
CUSOLVER_STATUS_INVALID_VALUE	an unsupported value or parameter was passed.
CUSOLVER_STATUS_EXECUTION_FAILED	a kernel failed to launch on the GPU.
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

2.6.26. cusolverRfBatchZeroPivot()

```
cusolverStatus_t
cusolverRfBatchZeroPivot(/* Input */
                        cusolverRfHandle_t handle
                        /* Output (in the host memory) */
                        int *position);
```

Although A_j is close to each other, it does not mean $M_j = P^* A_j^* Q^T = L_j^* U_j$ exists for every j . The user can query which matrix failed LU refactorization by checking corresponding value in **position** array. The input parameter **position** is an integer array of size **batchSize**.

The **j-th** component denotes the refactorization result of matrix A_j . If **position(j)** is -1, the LU refactorization of matrix A_j is successful. If **position(j)** is $k \geq 0$, matrix A_j is not LU factorizable and its matrix $U_j(j,j)$ is zero.

The return value of **cusolverRfBatch_Zero_Pivot** is **CUSOLVER_STATUS_ZERO_PIVOT** if there exists one A_j which failed LU refactorization. The user can redo LU factorization to get new permutation **P** and **Q** if error code **CUSOLVER_STATUS_ZERO_PIVOT** is returned.

parameter	MemSpace	In/out	Meaning
handle	host	input	the handle to the cuSolverRF library.
position	host	output	integer array of size batchSize . The value of position(j) reports singularity of matrix A_j , -1 if no structural/numerical zero, $k \geq 0$ if $A_j(k,k)$ is either structural zero or numerical zero.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_NOT_INITIALIZED	the library was not initialized.

CUSOLVER_STATUS_ZERO_PIVOT	a zero pivot was encountered during the computation.
----------------------------	--

Chapter 3.

USING THE CUSOLVERMG API

3.1. General description

This chapter describes how to use the cuSolverMG library API. It is not a reference for the cuSolverMG API data types and functions; that is provided in subsequent chapters.

3.1.1. Thread Safety

The library is thread-safe only if one cuSolverMG context per thread.

3.1.2. Determinism

Currently all cuSolverMG API routines from a given toolkit version, generate the same bit-wise results when the following conditions are respected :

- ▶ all GPUs participating to the computation have the same compute-capabilities and the same number of SMs.
- ▶ the tiles size is kept the same between run.
- ▶ number of logical GPUs is kept the same. The order of GPUs are not important because all have the same compute-capabilities.

3.1.3. tile strategy

The tiling strategy of cuSolverMG is compatible with ScaLAPACK. The current release only supports 1-D column block cyclic, column-major PACKED format.

Figure 1.a shows a partition of the matrix A of dimension M_A by N_A . Each column tile has T_A columns. There are seven columns tiles, labeled as 0,1,2,3,4,5,6, distributed into three GPUs in a **cyclic** way, i.e. each GPU takes one column tile in turn. For example, GPU 0 has column tile 0, 3, 6 (yellow tiles) and GPU 1 takes column tiles next to GPU 0 (blue tiles). Not all GPUs have the same number of tiles, in this example, GPU 0 has three tiles, others have only two tiles.

Figure 1.b shows two possible formats to store those column tiles locally in each GPU. Left side is called PACKED format and right side is UNPACKED format. PACKED format aggregates three column tiles in a contiguous memory block while UNPACKED format distributes these three column tiles into different memory blocks. The only difference between them is that PACKED format can have a big GEMM call instead of three GEMM calls in UNPACKED format. So theoretically speaking, PACKED format can deliver better performance than UNPACKED format. **cusolverMG** only supports PACKED format in the API. In order to achieve maximal performance, the user just needs to choose proper tile size **T_A** to partition the matrix, not too small, for example 256 or above is enough.

There is another parameter, called **LLD_A**, to control the leading dimension of the local matrix in each GPU. **LLD_A** must be greater or equal to **M_A**. The purpose of **LLD_A** is for better performance of GEMM. For small problem, GEMM is faster if **LLD_A** is power of 2. However for big problem, **LLD_A** does not show significant improvement. **cuSolverMG** only supports **LLD_A=M_A**.

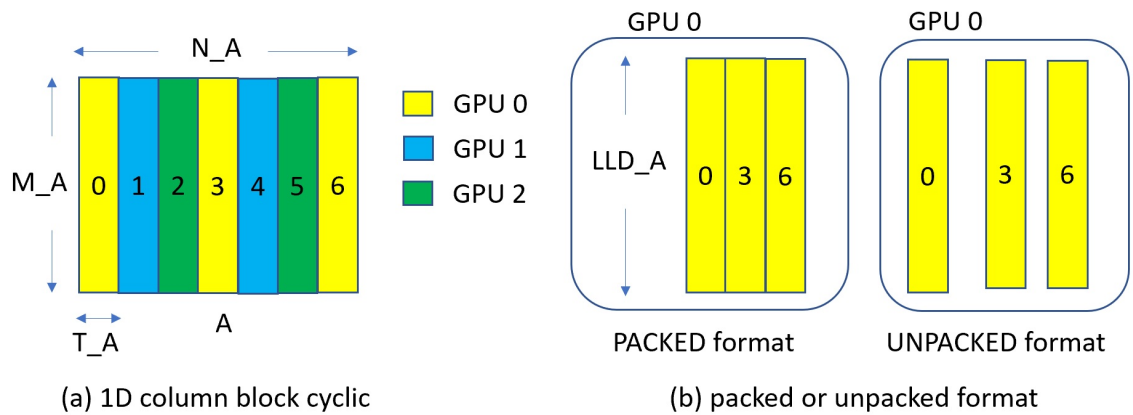


Figure 1 Example of **cusolverMG** tiling for 3 Gpus

The processing grid in **cuSolverMG** is a list of GPU IDs, similar to the process ID in **ScaLAPACK**. **cuSolverMG** only supports 1D column block cyclic, so only 1D grid is supported as well. Suppose **deviceId** is a list of GPU IDs, both **deviceId=1,1,1** and **deviceId=2,1,0** are valid. The former describes three logical devices are selected to run **cuSolverMG** routines, and all have the same physical ID, 0. The latter still uses three logical devices, but each has different physical ID. The current design only accepts 32 logical devices, that is, the length of **deviceId** is less or equal to 32. Figure 1 uses **deviceId=0,1,2**.

In practice, the matrix **A** is distributed into GPUs listed in **deviceId**. If the user chooses **deviceId=1,1,1**, all columns tile are located in GPU 1, this will limit the size of the problem because of memory capacity of one GPU. Besides, multiGPU routine adds extra overhead on data communication through off-chip bus, which has big performance impact if NVLINK is not supported or used. It would be faster to run on single GPU instead of running multiGPU version with devices of the same GPU ID.

3.1.4. Global matrix versus local matrix

To operate a submatrix of the matrix **A** is simple in dense linear algebra, just shift the pointer to the starting point of the submatrix relative to **A**. For example, `gesvd(10,10, A)` is SVD of **A**(0:9,0:9). `gesvd(10,10, A + 5 + 2*lda)` is SVD of 10-by-10 submatrix starting at **A**(5,2).

However it is not simple to operate on a submatrix of a distributed matrix because different starting point of the submatrix changes the distribution of the layout of that submatrix. **ScaLAPACK** introduces two parameters, **IA** and **JA**, to locate the submatrix. Figure 2 shows (global) matrix **A** of dimension **M_A** by **N_A**. The **sub(A)** is a **M** by **N** submatrix of **A**, starting at **IA** and **JA**. Please be aware that **IA** and **JA** are base-1.

Given a distributed matrix **A**, the user can compute eigenvalues of the submatrix **sub(A)** by either calling `syevd(A, IA, JA)` or gathering **sub(A)** to another distributed matrix **B** and calling `syevd(B, IB=1, JB=1)`.

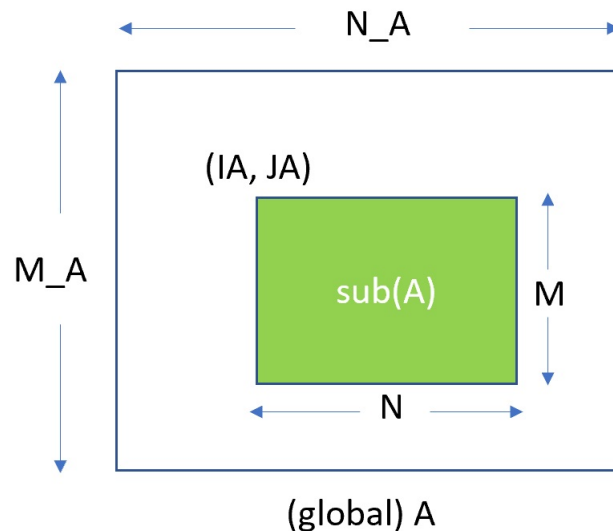


Figure 2 global matrix and local matrix

3.1.5. usage of _bufferSize

There is no `cudaMalloc` inside **cuSolverMG** library, the user must allocate the device workspace explicitly. The routine `xyz_bufferSize` is to query the size of workspace of the routine `xyz`, for example `xyz = syevd`. To make the API simple, `xyz_bufferSize` follows almost the same signature of `xyz` even it only depends on some parameters, for example, device pointer is not used to decide the size of workspace. In most cases, `xyz_bufferSize` is called in the beginning before actual device data (pointing by a device pointer) is prepared or before the device pointer is allocated. In such case, the user can pass null pointer to `xyz_bufferSize` without breaking the functionality.

`xyz_bufferSize` returns `bufferSize` for each device. The size is number of elements, not number of bytes.

3.1.6. synchronization

All routines are in synchronous (blocking call) manner. The data is ready after the routine. However the user has to prepare the distributed data before calling the routine. For example, if the user has multiple streams to setup the matrix, stream synchronization or device synchronization is necessary to guarantee distributed matrix is ready.

3.1.7. context switch

The user does not need to restore the device by `cudaSetDevice()` after each `cuSolverMG` call. All routines set the device back to what the caller has.

3.1.8. NVLINK

The peer-to-peer communication via NVLINK can dramatically reduce the overhead of data exchange among GPUs. `cuSolverMG` does not enable NVLINK implicitly, instead, it gives this option back to the user, not to interfere other libraries. The example code H.1 shows how to enable peer-to-peer communication.

3.2. cuSolverMG Types Reference

3.2.1. cuSolverMG Types

The `float`, `double`, `cuComplex`, and `cuDoubleComplex` data types are supported. The first two are standard C data types, while the last two are exported from `cuComplex.h`. In addition, `cuSolverMG` uses some familiar types from `cuBlas`.

3.2.2. cusolverMgHandle_t

This is a pointer type to an opaque `cuSolverMG` context, in which the user must initialize by calling `cusolverMgCreate()` prior to calling any other library function. An un-initialized handle object will lead to unexpected behavior, including crashes of `cuSolverMG`. The handle created and returned by `cusolverMgCreate()` must be passed to every `cuSolverMG` function.

3.2.3. cusolverMgGridMapping_t

The type indicates layout of grids.

Value	Meaning
<code>CUDALIBMG_GRID_MAPPING_ROW_MAJOR</code>	row-major ordering.
<code>CUDALIBMG_GRID_MAPPING_COL_MAJOR</code>	column-major ordering.

3.2.4. cudaLibMgGrid_t

opaque structure of the distributed grid.

3.2.5. cudaLibMgMatrixDesc_t

opaque structure of the distributed matrix descriptor.

3.3. Helper Function Reference

3.3.1. cusolverMgCreate()

```
cusolverStatus_t
cusolverMgCreate(cusolverMgHandle_t *handle)
```

This function initializes the cuSolverMG library and creates a handle on the cuSolverMG context. It must be called before any other cuSolverMG API function is invoked. It allocates hardware resources necessary for accessing the GPU.

Output

handle	the pointer to the handle to the cuSolverMG context.
---------------	--

Status Returned

CUSOLVER_STATUS_SUCCESS	the initialization succeeded.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.

3.3.2. cusolverMgDestroy()

```
cusolverStatus_t
cusolverMgDestroy(cusolverMgHandle_t handle)
```

This function releases CPU-side resources used by the cuSolverMG library.

Input

handle	the handle to the cuSolverMG context.
---------------	---------------------------------------

Status Returned

CUSOLVER_STATUS_SUCCESS	the shutdown succeeded.
-------------------------	-------------------------

3.3.3. cusolverMgDeviceSelect()

```
cusolverStatus_t
cusolverMgDeviceSelect(
    cusolverMgHandle_t handle,
    int nbDevices,
    int deviceId[] )
```

This function registers a subset of devices (GPUs) to **cuSolverMG** handle. Such subset of devices is used in subsequent API calls. The array **deviceId** contains a list of logical device ID. The term **logical** means repeated device ID are permitted. For example, suppose the user has only one GPU in the system, say device 0, if he sets **deviceId=0,0,0**, then **cuSolverMG** treats them as three independent GPUs, one stream each, so concurrent kernel launches still hold. The current design only supports up to 32 logical devices.

Input

handle	the pointer to the handle to the cuSolverMG context.
nbDevices	the number of logical devices
deviceId	an integer array of size nbDevices

Status Returned

CUSOLVER_STATUS_SUCCESS	the initialization succeeded.
CUSOLVER_STATUS_INVALID_VALUE	nbDevices must be greater than zero, and less or equal to 32.
CUSOLVER_STATUS_ALLOC_FAILED	the resources could not be allocated.
CUSOLVER_STATUS_INTERNAL_ERROR	internal error occurs when setting internal streams and events.

3.3.4. cusolverMgCreateDeviceGrid()

```
cusolverStatus_t
cusolverMgCreateDeviceGrid(
    cusolverMgGrid_t* grid,
    int32_t numRowsDevices,
    int32_t numColDevices,
    const int32_t deviceId[],
    cusolverMgGridMapping_t mapping)
```

This function setups grid of devices.

Only 1-D column block cyclic is supported, so **numRowDevices** must be equal to 1.

WARNING: **cusolverMgCreateDeviceGrid()** must be consistent with **cusolverMgDeviceSelect()**, i.e. **numColDevices** must be equal to **nbDevices** in **cusolverMgDeviceSelect()**.

parameter	Memory	In/out	Meaning
-----------	--------	--------	---------

grid	host	output	the pointer to the opaque structure.
numRowDevices	host	input	number of devices in the row.
numColDevices	host	input	number of devices in the column.
deviceId	host	input	integer array of size numColDevices , containing device IDs.
mapping	host	input	row-major or column-major ordering.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_INVALID_VALUE	numColDevices is not greater than 0. numRowDevices is not 1.

3.3.5. cusolverMgDestroyGrid()

```
cusolverStatus_t
cusolverMgDestroyGrid(
    cusolverMgGrid_t grid)
```

This function releases resources of a grid.

parameter	Memory	In/out	Meaning
grid	host	input/output	the pointer to the opaque structure.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
--------------------------------	---------------------------------------

3.3.6. cusolverMgCreateMatDescr()

```
cusolverStatus_t
cusolverMgCreateMatrixDesc(
    cusolverMgMatrixDesc_t * desc,
    int64_t numRows,
    int64_t numCols,
    int64_t rowBlockSize,
    int64_t colBlockSize,
    cudaDataType_t dataType,
    const cusolverMgGrid_t grid)
```

This function setups the matrix descriptor **desc**.

Only 1-D column block cyclic is supported, so **numRows** must be equal to **rowBlockSize**.

parameter	Memory	In/out	Meaning
desc	host	output	the matrix descriptor.
numRows	host	input	the number of rows of global A.
numCols	host	input	the number of columns of global A.

rowBlockSize	host	input	the number of rows per tile.
colBlockSize	host	input	the number of columns per tile.
dataType	host	input	data type of the matrix.
grid	host	input	the pointer to structure of grid.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_INVALID_VALUE	numRows , numCols , or rowBlockSize or colBlockSize is less than 0. numRows is not equal to rowBlockSize .

3.3.7. cusolverMgDestroyMatrixDesc()

```
cusolverStatus_t
cusolverMgDestroyMatrixDesc(
    cusolverMgMatrixDesc_t desc)
```

This function releases the matrix descriptor **desc**.

parameter	Memory	In/out	Meaning
desc	host	input/output	the matrix descriptor.

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
--------------------------------	---------------------------------------

3.4. Dense Linear Solver Reference

This chapter describes linear solver API of cuSolverMG, including LU with partial pivoting.

3.4.1. cusolverMgGetrf()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverMgGetrf_bufferSize(
    cusolverMgHandle_t handle,
    int M,
    int N,
    void *array_d_A[],
    int IA,
    int JA,
    cudaLibMgMatrixDesc_t descrA,
    int *array_d_IPIV[],
    cudaDataType_t computeType,
    int64_t *lwork);
```

```
cusolverStatus_t
cusolverMgGetrf(
    cusolverMgHandle_t handle,
    int M,
    int N,
    void *array_d_A[],
    int IA,
    int JA,
    cudaLibMgMatrixDesc_t descrA,
    int *array_d_IPIV[],
    cudaDataType_t computeType,
    void *array_d_work[],
    int64_t lwork,
    int *info );
```

This function computes the LU factorization of a $\mathbf{M} \times \mathbf{N}$ matrix

$$\mathbf{P} * \mathbf{A} = \mathbf{L} * \mathbf{U}$$

where \mathbf{A} is a $\mathbf{M} \times \mathbf{N}$ matrix, \mathbf{P} is a permutation matrix, \mathbf{L} is a lower triangular matrix with unit diagonal, and \mathbf{U} is an upper triangular matrix.

The user has to provide device working space in **array_d_work**. **array_d_work** is a host pointer array of dimension \mathbf{G} , where \mathbf{G} is number of devices. **array_d_work[j]** is a device pointer pointing to a device memory in j-th device. The data type of **array_d_work[j]** is **computeType**. The size of **array_d_work[j]** is **lwork** which is number of elements per device, returned by **cusolverMgGetrf_bufferSize()**.

If LU factorization failed, i.e. matrix \mathbf{A} (\mathbf{U}) is singular, The output parameter **info=i** indicates $\mathbf{U}(i, i) = 0$.

If output parameter **info = -i** (less than zero), the **i-th** parameter is wrong (not counting handle).

If **array_d_IPIV** is null, no pivoting is performed. The factorization is $\mathbf{A} = \mathbf{L} * \mathbf{U}$, which is not numerically stable.

array_d_IPIV must be consistent with **array_d_A**, i.e. **JA** is the first column of **sub(A)**, also the first column of **sub(IPIV)**.

No matter LU factorization failed or not, the output parameter **array_d_IPIV** contains pivoting sequence, row **i** is interchanged with row **array_d_IPIV(i)**.

The generic API has three different types, **dataTypeA** is data type of the matrix **A**, **computeType** is compute type of the operation and data type of the workspace (**array_d_work**) **descrA** contains **dataTypeA**, so there is no explicit parameter of **dataTypeA**. **cusolverMgGetrf** only supports the following four combinations.

Appendix I provides an example of **cusolverMgGetrf**.

valid combination of data type and compute type

DataTypeA	ComputeType	Meaning	
CUDA_R_32F	CUDA_R_32F	SGETRF	
CUDA_R_64F	CUDA_R_64F	DGETRF	
CUDA_C_32F	CUDA_C_32F	CGETRF	
CUDA_C_64F	CUDA_C_64F	ZGETRF	

Remark 1: tile size **TA** must be less or equal to 512.

API of getrf

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverMg library context.
M	host	input	number of rows of matrix sub(A) .
N	host	input	number of columns of matrix sub(A) .
array_d_A	host	in/out	a host pointer array of dimension G . It contains a distributed <type> array containing sub(A) of dimension M * N . On exit, sub(A) contains the factors L and U .
IA	host	input	The row index in the global array A indicating the first row of sub(A) .
JA	host	input	The column index in the global array A indicating the first column of sub(A) .
descrA	host	input	matrix descriptor for the distributed matrix A .
array_d_IPIV	host	output	a host pointer array of dimension G . it contains a distributed integer array containing sub(IPIV) of size min(M,N) . sub(IPIV) contains pivot indices.
computeType	host	input	Data type used for computation.
array_d_work	host	in/out	a host pointer array of dimension G . array_d_work[j] points to a device working space in j -th device, <type> array of size lwork .
lwork	host	input	size of array_d_work[j] , returned by cusolverMgGetrf_bufferSize . lwork denotes number of elements, not number of bytes.

info	host	output	if <code>info = 0</code> , the LU factorization is successful. if <code>info = -i</code> , the <i>i</i> -th parameter is wrong (not counting handle). if <code>info = i</code> , the $U(i,i) = 0$.
------	------	--------	---

Status Returned

CUSOLVER_STATUS_SUCCESS	the operation completed successfully.
CUSOLVER_STATUS_INVALID_VALUE	invalid parameters were passed ($M, N < 0$).
CUSOLVER_STATUS_INTERNAL_ERROR	an internal operation failed.

3.4.2. cusolverMgGetrs()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverMgGetrs_bufferSize(
    cusolverMgHandle_t handle,
    cublasOperation_t TRANS,
    int N,
    int NRHS,
    void *array_d_A[],
    int IA,
    int JA,
    cudaLibMgMatrixDesc_t descrA,
    int *array_d_IPIV[],
    void *array_d_B[],
    int IB,
    int JB,
    cudaLibMgMatrixDesc_t descrB,
    cudaDataType_t computeType,
    int64_t *lwork);
```

```
cusolverStatus_t
cusolverMgGetrs(
    cusolverMgHandle_t handle,
    cublasOperation_t TRANS,
    int N,
    int NRHS,
    void *array_d_A[],
    int IA,
    int JA,
    cudaLibMgMatrixDesc_t descrA,
    int *array_d_IPIV[],
    void *array_d_B[],
    int IB,
    int JB,
    cudaLibMgMatrixDesc_t descrB,
    cudaDataType_t computeType,
    void *array_d_work[],
    int64_t lwork,
    int *info );
```

This function solves a linear system of multiple right-hand sides

$$\text{op}(A) * X = B$$

where **A** is a **N**×**N** matrix, and was LU-factored by **getrf**, that is, lower triangular part of **A** is **L**, and upper triangular part (including diagonal elements) of **A** is **U**. **B** is a **N**×**NRHS** right-hand side matrix. The solution matrix **X** overwrites the right-hand-side matrix **B**.

The input parameter **TRANS** is defined by

$$\text{op}(\mathbf{A}) = \begin{cases} \mathbf{A} & \text{if TRANS} == \text{CUBLAS_OP_N} \\ \mathbf{A}^T & \text{if TRANS} == \text{CUBLAS_OP_T} \\ \mathbf{A}^H & \text{if TRANS} == \text{CUBLAS_OP_C} \end{cases}$$

The user has to provide device working space in **array_d_work**. **array_d_work** is a host pointer array of dimension **G**, where **G** is number of devices. **array_d_work[j]** is a device pointer pointing to a device memory in j-th device. The data type of **array_d_work[j]** is **computeType**. The size of **array_d_work[j]** is **lwork** which is number of elements per device, returned by **cusolverMgGetrs_bufferSize()**.

If **array_d_IPIV** is null, no pivoting is performed. Otherwise, **array_d_IPIV** is an output of **getrf**. It contains pivot indices, which are used to permute right-hand sides.

If output parameter **info** = **-i** (less than zero), the **i-th** parameter is wrong (not counting handle).

The generic API has three different types, **dataTypeA** is data type of the matrix **A**, **dataTypeB** is data type of the matrix **B**, and **computeType** is compute type of the operation and data type of the workspace (**array_d_work**) **descrA** contains **dataTypeA**, so there is no explicit parameter of **dataTypeA**. **descrB** contains **dataTypeB**, so there is no explicit parameter of **dataTypeB**. **cusolverMgGetrs** only supports the following four combinations.

valid combination of data type and compute type

DataTypeA	DataTypeB	ComputeType	Meaning
CUDA_R_32F	CUDA_R_32F	CUDA_R_32F	SGETRS
CUDA_R_64F	CUDA_R_64F	CUDA_R_64F	DGETRS
CUDA_C_32F	CUDA_C_32F	CUDA_C_32F	CGETRS
CUDA_C_64F	CUDA_C_64F	CUDA_C_64F	ZGETRS

Remark 1: tile size **TA** must be less or equal to 512.

Remark 2: only support **TRANS=CUBLAS_OP_N**.

Appendix I provides an example of **cusolverMgGetrs**.

API of gets

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverMG library context.
TRANS	host	input	operation op (A) that is non- or (conj.) transpose.
N	host	input	number of rows and columns of matrix sub (A) .
NRHS	host	input	number of columns of matrix sub (B) .

<code>array_d_A</code>	host	input	a host pointer array of dimension G . It contains a distributed <type> array containing <code>sub(A)</code> of dimension $M * N$. <code>sub(A)</code> contains the factors L and U .
<code>IA</code>	host	input	The row index in the global array A indicating the first row of <code>sub(A)</code> .
<code>JA</code>	host	input	The column index in the global array A indicating the first column of <code>sub(A)</code> .
<code>descrA</code>	host	input	matrix descriptor for the distributed matrix A .
<code>array_d_IPIV</code>	host	input	a host pointer array of dimension G . it contains a distributed integer array containing <code>sub(IPIV)</code> of dimension $\min(M, N)$. <code>sub(IPIV)</code> contains pivot indices.
<code>array_d_B</code>	host	in/out	a host pointer array of dimension G . It contains a distributed <type> array containing <code>sub(B)</code> of dimension $N * NRHS$.
<code>IB</code>	host	input	The row index in the global array B indicating the first row of <code>sub(B)</code> .
<code>JB</code>	host	input	The column index in the global array B indicating the first column of <code>sub(B)</code> .
<code>descrB</code>	host	input	matrix descriptor for the distributed matrix B .
<code>computeType</code>	host	input	Data type used for computation.
<code>array_d_work</code>	host	in/out	a host pointer array of dimension G . <code>array_d_work[j]</code> points to a device working space in j -th device, <type> array of size <code>lwork</code> .
<code>lwork</code>	host	input	size of <code>array_d_work[j]</code> , returned by <code>cusolverMgGetrs_bufferSize</code> . <code>lwork</code> denotes number of elements, not number of bytes.
<code>info</code>	host	output	if <code>info</code> = 0, the operation is successful. if <code>info</code> = $-i$, the i -th parameter is wrong (not counting handle).

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed ($N < 0$ or $NRHS < 0$).
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

3.5. Dense Eigenvalue Solver Reference

This chapter describes eigenvalue solver API of cuSolverMG.

3.5.1. cusolverMgSyevd()

The helper functions below can calculate the sizes needed for pre-allocated buffer.

```
cusolverStatus_t
cusolverMgSyevd_bufferSize(
    cusolverMgHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int N,
    void *array_d_A[],
    int IA,
    int JA,
    cudaLibMgMatrixDesc_t descrA,
    void *W,
    cudaDataType_t dataTypeW,
    cudaDataType_t computeType,
    int64_t *lwork
);
```

```
cusolverStatus_t
cusolverMgSyevd(
    cusolverMgHandle_t handle,
    cusolverEigMode_t jobz,
    cublasFillMode_t uplo,
    int N,
    void *array_d_A[],
    int IA,
    int JA,
    cudaLibMgMatrixDesc_t descrA,
    void *W,
    cudaDataType_t dataTypeW,
    cudaDataType_t computeType,
    void *array_d_work[],
    int64_t lwork,
    int *info );
```

This function computes eigenvalues and eigenvectors of a symmetric (Hermitian) $N \times N$ matrix **A**. The standard symmetric eigenvalue problem is

$$A * V = V * \Lambda$$

where Λ is a real $N \times N$ diagonal matrix. V is an $N \times N$ unitary matrix. The diagonal elements of Λ are the eigenvalues of **A** in ascending order.

cusolverMgSyevd returns the eigenvalues in **W** and overwrites the eigenvectors in **A**. **W** is a host $1 \times N$ vector.

The generic API has three different types, **dataTypeA** is data type of the matrix **A**, **dataTypeW** is data type of the vector **W**, and **computeType** is compute type of the operation and data type of the workspace (**array_d_work**) **descrA** contains **dataTypeA**, so there is no explicit parameter of **dataTypeA**. **cusolverMgSyevd** only supports the following four combinations.

valid combination of data type and compute type

DataTypeA	DataTypeW	ComputeType	Meaning
CUDA_R_32F	CUDA_R_32F	CUDA_R_32F	SSYEVD
CUDA_R_64F	CUDA_R_64F	CUDA_R_64F	DSYEVD

CUDA_C_32F	CUDA_R_32F	CUDA_C_32F	CHEEVD
CUDA_C_64F	CUDA_R_64F	CUDA_C_64F	ZHEEVD

The user has to provide device working space in **array_d_work**. **array_d_work** is a host pointer array of dimension **G**, where **G** is number of devices. **array_d_work[j]** is a device pointer pointing to a device memory in j-th device. The data type of **array_d_work[j]** is **computeType**. The size of **array_d_work[j]** is **lwork** which is number of elements per device, returned by **cusolverMgSyevd_bufferSize()**.

array_d_A is also a host pointer array of dimension **G**. **array_d_A[j]** is a device pointer pointing to a device memory in j-th device. The data type of **array_d_A[j]** is **dataTypeA**. The size of **array_d_A[j]** is about **N*TA*(blocks per device)**. The user has to prepare **array_d_A** manually (please check the samples in Appendix H).

If output parameter **info = -i** (less than zero), the **i-th** parameter is wrong (not counting handle). If **info = i** (greater than zero), **i** off-diagonal elements of an intermediate tridiagonal form did not converge to zero.

if **jobz = CUSOLVER_EIG_MODE_VECTOR**, **A** contains the orthonormal eigenvectors of the matrix **A**. The eigenvectors are computed by a divide and conquer algorithm.

Remark 1: only **CUBLAS_FILL_MODE_LOWER** is supported, so the user has to prepare lower triangle of **A**.

Remark 2: only **IA=1** and **JA=1** are supported.

Remark 3: tile size **TA** must be less or equal to 1024. To achieve best performance, **TA** should be 256 or 512.

Appendix H provides three examples of **cusolverMgSyevd**.

API of **syevd**

parameter	Memory	In/out	Meaning
handle	host	input	handle to the cuSolverMG library context.
jobz	host	input	specifies options to either compute eigenvalue only or compute eigen-pair: jobz = CUSOLVER_EIG_MODE_NOVECTOR : Compute eigenvalues only; jobz = CUSOLVER_EIG_MODE_VECTOR : Compute eigenvalues and eigenvectors.
uplo	host	input	specifies which part of A is stored. uplo = CUBLAS_FILL_MODE_LOWER : Lower triangle of A is stored. uplo = CUBLAS_FILL_MODE_UPPER : Upper triangle of A is stored. Only CUBLAS_FILL_MODE_LOWER is supported.
N	host	input	number of rows (or columns) of matrix sub (A) .
array_d_A	host	in/out	a host pointer array of dimension G . It contains a distributed <type> array containing sub (A) of dimension N * N . If uplo = CUBLAS_FILL_MODE_UPPER , the leading N-by-N upper triangular part

			of <code>sub(A)</code> contains the upper triangular part of the matrix <code>sub(A)</code> . If <code>uplo = CUBLAS_FILL_MODE_LOWER</code> , the leading N-by-N lower triangular part of <code>sub(A)</code> contains the lower triangular part of the matrix <code>sub(A)</code> . On exit, if <code>jobz = CUSOLVER_EIG_MODE_VECTOR</code> , and <code>info = 0</code> , <code>sub(A)</code> contains the orthonormal eigenvectors of the matrix <code>sub(A)</code> . If <code>jobz = CUSOLVER_EIG_MODE_NOVECTOR</code> , the contents of <code>A</code> are destroyed.
<code>IA</code>	host	input	The row index in the global array <code>A</code> indicating the first row of <code>sub(A)</code> .
<code>JA</code>	host	input	The column index in the global array <code>A</code> indicating the first column of <code>sub(A)</code> .
<code>descrA</code>	host	input	matrix descriptor for the distributed matrix <code>A</code> .
<code>W</code>	host	output	a real array of dimension <code>N</code> . The eigenvalue values of <code>sub(A)</code> , in ascending order ie, sorted so that <code>W(i) <= W(i+1)</code> .
<code>dataTypeW</code>	host	input	Data type of the vector <code>W</code> .
<code>computeType</code>	host	input	Data type used for computation.
<code>array_d_work</code>	host	in/out	a host pointer array of dimension <code>G</code> . <code>array_d_work[j]</code> points to a device working space in <code>j</code> -th device, <type> array of size <code>lwork</code> .
<code>lwork</code>	host	input	size of <code>array_d_work[j]</code> , returned by <code>cusolverMgSyevd_bufferSize</code> . <code>lwork</code> denotes number of elements, not number of bytes.
<code>info</code>	host	output	if <code>info = 0</code> , the operation is successful. if <code>info = -i</code> , the <code>i</code> -th parameter is wrong (not counting handle). if <code>info = i (> 0)</code> , <code>info</code> indicates <code>i</code> off-diagonal elements of an intermediate tridiagonal form did not converge to zero;

Status Returned

<code>CUSOLVER_STATUS_SUCCESS</code>	the operation completed successfully.
<code>CUSOLVER_STATUS_INVALID_VALUE</code>	invalid parameters were passed (<code>N < 0</code> , or <code>lda < max(1, N)</code> , or <code>jobz</code> is not <code>CUSOLVER_EIG_MODE_NOVECTOR</code> or <code>CUSOLVER_EIG_MODE_VECTOR</code> , or <code>uplo</code> is not <code>CUBLAS_FILL_MODE_LOWER</code> , or <code>IA</code> and <code>JA</code> are not 1, or <code>N</code> is bigger than dimension of global <code>A</code> , or the combination of <code>dataType</code> and <code>computeType</code> is not valid.
<code>CUSOLVER_STATUS_INTERNAL_ERROR</code>	an internal operation failed.

Appendix A.

CUSOLVERRF EXAMPLES

A.1. cuSolverRF In-memory Example

This is an example in the C programming language of how to use the standard routines in the cuSolverRF library. We focus on solving the set of linear systems

$$A_i x_i = f_i$$

but we change the indexing from one- to zero-based to follow the C programming language. The example begins with the usual includes and main()

```
#include <stdio.h>
#include <stdlib.h>
#include <cuda_runtime.h>
#include "cusolverRf.h"

#define TEST_PASSED 0
#define TEST_FAILED 1

int main (void){
    /* matrix A */
    int n;
    int nnzA;
    int *Ap=NULL;
    int *Ai=NULL;
    double *Ax=NULL;
    int *d_Ap=NULL;
    int *d_Ai=NULL;
    double *d_rAx=NULL;
    /* matrices L and U */
    int nnzL, nnzU;
    int *Lp=NULL;
    int *Li=NULL;
    double* Lx=NULL;
    int *Up=NULL;
    int *Ui=NULL;
    double* Ux=NULL;
    /* reordering matrices */
    int *P=NULL;
    int *Q=NULL;
    int * d_P=NULL;
    int * d_Q=NULL;
    /* solution and rhs */
    int nrhs; // # of rhs for each system (currently only =1 is supported)
    double *d_X=NULL;
    double *d_T=NULL;
    /* cuda */
    cudaError_t cudaStatus;
    /* cuolverRf */
    cusolverRfHandle_t gH=NULL;
    cusolverStatus_t status;
    /* host sparse direct solver */
    /* ... */
    /* other variables */
    int tnnzL, tnnzU;
    int *tLp=NULL;
    int *tLi=NULL;
    double *tLx=NULL;
    int *tUp=NULL;
    int *tUi=NULL;
    double *tUx=NULL;
    double t1, t2;
```

Then we initialize the library.

```

/* ASSUMPTION: recall that we are solving a set of linear systems
   A_{i} x_{i} = f_{i} for i=0,...,k-1
   where the sparsity pattern of the coefficient matrices A_{i}
   as well as the reordering to minimize fill-in and the pivoting
   used during the LU factorization remain the same. */

/* Step 1: solve the first linear system (i=0) on the host,
   using host sparse direct solver, which involves
   full LU factorization and solve. */
/* ... */

/* Step 2: interface to the library by extracting the following
   information from the first solve:
   a) triangular factors L and U
   b) pivoting and reordering permutations P and Q
   c) also, allocate all the necessary memory */
/* ... */

/* Step 3: use the library to solve subsequent (i=1,...,k-1) linear systems
   a) the library setup (called only once) */
//create handle
status = cusolverRfCreate(&gH);
if (status != CUSOLVER_STATUS_SUCCESS){
    printf ("[cusolverRf status %d]\n",status);
    return TEST_FAILED;
}

//set fast mode
status = cusolverRfSetResetValuesFastMode(gH, GLU_RESET_VALUES_FAST_MODE_ON);
if (status != CUSOLVER_STATUS_SUCCESS){
    printf ("[cusolverRf status %d]\n",status);
    return TEST_FAILED;
}

```

Call refactorization and solve.

```

//assemble internal data structures (you should use the coefficient matrix A
//corresponding to the second (i=1) linear system in this call)
t1 = cusolver_test_seconds();
status = cusolverRfSetupHost(n, nnzA, Ap, Ai, Ax,
                             nnzL, Lp, Li, Lx, nnzU, Up, Ui, Ux, P, Q, gH);
cudaStatus = cudaDeviceSynchronize();
t2 = cusolver_test_seconds();
if ((status != CUSOLVER_STATUS_SUCCESS) || (cudaStatus != cudaSuccess)) {
    printf ("[cusolverRf status %d]\n",status);
    return TEST_FAILED;
}
printf("cusolverRfSetupHost time = %f (s)\n", t2-t1);

//analyze available parallelism
t1 = cusolver_test_seconds();
status = cusolverRfAnalyze(gH);
cudaStatus = cudaDeviceSynchronize();
t2 = cusolver_test_seconds();
if ((status != CUSOLVER_STATUS_SUCCESS) || (cudaStatus != cudaSuccess)) {
    printf ("[cusolverRf status %d]\n",status);
    return TEST_FAILED;
}
printf("cusolverRfAnalyze time = %f (s)\n", t2-t1);

/* b) The library subsequent (i=1,...,k-1) LU re-factorization
and solve (called multiple times). */
for (i=1; i<k; i++){
    //LU re-factorization
    t1 = cusolver_test_seconds();
    status = cusolverRfRefactor(gH);
    cudaStatus = cudaDeviceSynchronize();
    t2 = cusolver_test_seconds();
    if ((status != CUSOLVER_STATUS_SUCCESS) || (cudaStatus != cudaSuccess))
    {
        printf ("[cusolverRF status %d]\n",status);
        return TEST_FAILED;
    }
    printf("cuSolverReRefactor time = %f (s)\n", t2-t1);

    //forward and backward solve
    t1 = cusolver_test_seconds();
    status = cusolverRfSolve(gH, d_P, d_Q, nrhs, d_T, n, d_X, n);
    cudaStatus = cudaDeviceSynchronize();
    t2 = cusolver_test_seconds();
    if ((status != CUSOLVER_STATUS_SUCCESS) || (cudaStatus != cudaSuccess))
    {
        printf ("[cusolverRf status %d]\n",status);
        return TEST_FAILED;
    }
    printf("cusolverRfSolve time = %f (s)\n", t2-t1);
}

```

Extract the results and return.

```

        // extract the factors (if needed)
        status = cusolverRfExtractSplitFactorsHost(gH, &tnnzL, &tLp, &tLi,
&tLx,
                                                    &tnnzU, &tUp, &tUi, &tUx);
        if(status != CUSOLVER_STATUS_SUCCESS){
            printf ("[cusolverRf status %d]\n",status);
            return TEST_FAILED;
        }
        /*
        //print
        int row, j;
        printf("printing L\n");
        for (row=0; row<n; row++){
            for (j=tLp[row]; j<tLp[row+1]; j++){
                printf("%d,%d,%f\n",row,tLi[j],tLx[j]);
            }
        }
        printf("printing U\n");
        for (row=0; row<n; row++){
            for (j=tUp[row]; j<tUp[row+1]; j++){
                printf("%d,%d,%f\n",row,tUi[j],tUx[j]);
            }
        }
        */

        /* perform any other operations based on the solution */
        /* ... */

        /* check if done */
        /* ... */

        /* proceed to solve the next linear system */
        // update the coefficient matrix using reset values
        // (assuming that the new linear system, in other words,
        // new values are already on the GPU in the array d_rAx)
        t1 = cusolver_test_seconds();
        status = cusolverRfResetValues(n,nnzA,d_Ap,d_Ai,d_rAx,d_P,d_Q,gH);
        cudaStatus = cudaDeviceSynchronize();
        t2 = cusolver_test_seconds();
        if ((status != CUSOLVER_STATUS_SUCCESS) || (cudaStatus != cudaSuccess))
        {
            printf ("[cusolverRf status %d]\n",status);
            return TEST_FAILED;
        }
        printf("cusolverRfResetValues time = %f (s)\n", t2-t1);
    }

    /* free memory and exit */
    /* ... */
    return TEST_PASSED;
}

```

A.2. cuSolverRF-batch Example

This chapter provides an example in the C programming language of how to use the batched routines in the cuSolverRF library. We focus on solving the set of linear systems

$$A_i x_i = f_i$$

but we change the indexing from one- to zero-based to follow the C programming language. The first part is the usual includes and main definition

```
#include <stdio.h>
#include <stdlib.h>
#include <cuda_runtime.h>
#include "cusolverRf.h"

#define TEST_PASSED 0
#define TEST_FAILED 1

int main (void){
    /* matrix A */
    int batchSize;
    int n;
    int nnzA;
    int *Ap=NULL;
    int *Ai=NULL;
    //array of pointers to the values of each matrix in the batch (of size
    //batchSize) on the host
    double **Ax_array=NULL;
    //For example, if Ax_batch is the array (of size batchSize*nnzA) containing
    //the values of each matrix in the batch written contiguously one matrix
    //after another on the host, then Ax_array[j] = &Ax_batch[nnzA*j];
    //for j=0,...,batchSize-1.
    double *Ax_batch=NULL;
    int *d_Ap=NULL;
    int *d_Ai=NULL;
    //array of pointers to the values of each matrix in the batch (of size
    //batchSize) on the device
    double **d_Ax_array=NULL;
    //For example, if d_Ax_batch is the array (of size batchSize*nnzA)
    containing
    //the values of each matrix in the batch written contiguously one matrix
    //after another on the device, then d_Ax_array[j] = &d_Ax_batch[nnzA*j];
    //for j=0,...,batchSize-1.
    double *d_Ax_batch=NULL;
    /* matrices L and U */
    int nnzL, nnzU;
    int *Lp=NULL;
    int *Li=NULL;
    double* Lx=NULL;
    int *Up=NULL;
    int *Ui=NULL;
    double* Ux=NULL;
    /* reordering matrices */
    int *P=NULL;
    int *Q=NULL;
    int *d_P=NULL;
    int *d_Q=NULL;
```

Next we initialize the data needed and the create library handles

```

/* solution and rhs */
int nrhs; // # of rhs for each system (currently only =1 is supported)
//temporary storage (of size 2*batchSize*n*nrhs)
double *d_T=NULL;
//array (of size batchSize*n*nrhs) containing the values of each rhs in
//the batch written contiguously one rhs after another on the device
double **d_X_array=NULL;
//array (of size batchSize*n*nrhs) containing the values of each rhs in
//the batch written contiguously one rhs after another on the host
double **X_array=NULL;
/* cuda */
cudaError_t cudaStatus;
/* cusolverRf */
cusolverRfHandle_t gH=NULL;
cusolverStatus_t status;
/* host sparse direct solver */
...
/* other variables */
double t1, t2;

/* ASSUMPTION:
recall that we are solving a batch of linear systems
 $A_{\{j\}} x_{\{j\}} = f_{\{j\}}$  for  $j=0, \dots, \text{batchSize}-1$ 
where the sparsity pattern of the coefficient matrices  $A_{\{j\}}$ 
as well as the reordering to minimize fill-in and the pivoting
used during the LU factorization remain the same. */

/* Step 1: solve the first linear system (j=0) on the host,
using host sparse direct solver, which involves
full LU factorization and solve. */
/* ... */

/* Step 2: interface to the library by extracting the following
information from the first solve:
a) triangular factors L and U
b) pivoting and reordering permutations P and Q
c) also, allocate all the necessary memory */
/* ... */

/* Step 3: use the library to solve the remaining (j=1,...,batchSize-1)
linear systems.
a) the library setup (called only once) */
//create handle
status = cusolverRfcreate(&gH);
if (status != CUSOLVER_STATUS_SUCCESS){
    printf ("[cusolverRf status %d]\n",status);
    return TEST_FAILED;
}

```

We call the batch solve method and return.

```
//assemble internal data structures
t1 = cusolver_test_seconds();
status = cusolverRfBatchSetupHost(batchSize, n, nnzA, Ap, Ai, Ax_array,
                                nnzL, Lp, Li, Lx, nnzU, Up, Ui, Ux, P, Q,
gH);
cudaStatus = cudaDeviceSynchronize();
t2 = cusolver_test_seconds();
if ((status != CUSOLVER_STATUS_SUCCESS) || (cudaStatus != cudaSuccess)) {
    printf("[cusolverRf status %d]\n", status);
    return TEST_FAILED;
}
printf("cusolverRfBatchSetupHost time = %f (s)\n", t2-t1);

//analyze available parallelism
t1 = cusolver_test_seconds();
status = cusolverRfBatchAnalyze(gH);
cudaStatus = cudaDeviceSynchronize();
t2 = cusolver_test_seconds();
if ((status != CUSOLVER_STATUS_SUCCESS) || (cudaStatus != cudaSuccess)) {
    printf("[cusolverRf status %d]\n", status);
    return TEST_FAILED;
}
printf("cusolverRfBatchAnalyze time = %f (s)\n", t2-t1);

/* b) The library subsequent (j=1,...,batchSize-1) LU re-factorization
and solve (may be called multiple times). For the subsequent batches
the values can be reset using cusolverRfBatch_reset_values_routine. */
//LU re-factorization
t1 = cusolver_test_seconds();
status = cusolverRfBatchRefactor(gH);
cudaStatus = cudaDeviceSynchronize();
t2 = cusolver_test_seconds();
if ((status != CUSOLVER_STATUS_SUCCESS) || (cudaStatus != cudaSuccess)) {
    printf("[cusolverRf status %d]\n", status);
    return TEST_FAILED;
}
printf("cusolverRfBatchRefactor time = %f (s)\n", t2-t1);

//forward and backward solve
t1 = cusolver_test_seconds();
status = cusolverRfBatchSolve(gH, d_P, d_Q, nrhs, d_T, n, d_X_array, n);
cudaStatus = cudaDeviceSynchronize();
t2 = cusolver_test_seconds();
if ((status != CUSOLVER_STATUS_SUCCESS) || (cudaStatus != cudaSuccess)) {
    printf("[cusolverRf status %d]\n", status);
    return TEST_FAILED;
}
printf("cusolverRfBatchSolve time = %f (s)\n", t2-t1);

/* free memory and exit */
/* ... */
return TEST_PASSED;
}
```


Appendix B.

CSR QR BATCH EXAMPLES

B.1. Batched Sparse QR example 1

This chapter provides a simple example in the C programming language of how to use batched sparse QR to solve a set of linear systems

$$A_i x_i = b_i$$

All matrices A_i are small perturbations of

$$A = \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 2.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 3.0 & 0.0 \\ 0.1 & 0.1 & 0.1 & 4.0 \end{pmatrix}$$

All right-hand side vectors b_i are small perturbation of the Matlab vector 'ones(4,1)'.

We assume device memory is big enough to compute all matrices in one pass.

The usual includes and main definition

```

#include <stdio.h>
#include <stdlib.h>
#include <assert.h>

#include <cusolverSp.h>
#include <cuda_runtime_api.h>

int main(int argc, char*argv[])
{
    cusolverSpHandle_t cusolverH = NULL;
    // GPU does batch QR
    csrqrInfo_t info = NULL;
    cusparseMatDescr_t descrA = NULL;

    cusparseStatus_t cusparse_status = CUSPARSE_STATUS_SUCCESS;
    cusolverStatus_t cusolver_status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    cudaError_t cudaStat5 = cudaSuccess;

    // GPU does batch QR
    // d_A is CSR format, d_csrValA is of size nnzA*batchSize
    // d_x is a matrix of size batchSize * m
    // d_b is a matrix of size batchSize * m
    int *d_csrRowPtrA = NULL;
    int *d_csrColIndA = NULL;
    double *d_csrValA = NULL;
    double *d_b = NULL; // batchSize * m
    double *d_x = NULL; // batchSize * m

    size_t size_qr = 0;
    size_t size_internal = 0;
    void *buffer_qr = NULL; // working space for numerical factorization

    /*
    * A = | 1          |
    *      |          2          |
    *      |          3          |
    *      | 0.1  0.1  0.1  4 |
    *      CSR of A is based-1
    *
    * b = [1 1 1 1]
    */

```

Set up the library handle and data

```

const int m = 4 ;
const int nnzA = 7;
const int csrRowPtrA[m+1] = { 1, 2, 3, 4, 8};
const int csrColIndA[nnzA] = { 1, 2, 3, 1, 2, 3, 4};
const double csrValA[nnzA] = { 1.0, 2.0, 3.0, 0.1, 0.1, 0.1, 4.0};
const double b[m] = {1.0, 1.0, 1.0, 1.0};
const int batchSize = 17;

double *csrValABatch = (double*)malloc(sizeof(double)*nnzA*batchSize);
double *bBatch       = (double*)malloc(sizeof(double)*m*batchSize);
double *xBatch        = (double*)malloc(sizeof(double)*m*batchSize);
assert( NULL != csrValABatch );
assert( NULL != bBatch );
assert( NULL != xBatch );

// step 1: prepare Aj and bj on host
// Aj is a small perturbation of A
// bj is a small perturbation of b
// csrValABatch = [A0, A1, A2, ...]
// bBatch = [b0, b1, b2, ...]
for(int colidx = 0 ; colidx < nnzA ; colidx++){
    double Areg = csrValA[colidx];
    for (int batchId = 0 ; batchId < batchSize ; batchId++){
        double eps = ((double)((rand() % 100) + 1)) * 1.e-4;
        csrValABatch[batchId*nnzA + colidx] = Areg + eps;
    }
}

for(int j = 0 ; j < m ; j++){
    double breg = b[j];
    for (int batchId = 0 ; batchId < batchSize ; batchId++){
        double eps = ((double)((rand() % 100) + 1)) * 1.e-4;
        bBatch[batchId*m + j] = breg + eps;
    }
}

// step 2: create cusolver handle, qr info and matrix descriptor
cusolver_status = cusolverSpCreate(&cusolverH);
assert (cusolver_status == CUSOLVER_STATUS_SUCCESS);

cusparseset_status = cusparsesetCreateMatDescr(&descrA);
assert(cusparseset_status == CUSPARSE_STATUS_SUCCESS);

cusparsesetSetMatType(descrA, CUSPARSE_MATRIX_TYPE_GENERAL);
cusparsesetSetMatIndexBase(descrA, CUSPARSE_INDEX_BASE_ONE); // base-1

cusolver_status = cusolverSpCreateCsrqrInfo(&info);
assert(cusolver_status == CUSOLVER_STATUS_SUCCESS);

```

Call the solver

```

// step 3: copy Aj and bj to device
    cudaStat1 = cudaMalloc ((void**) &d_csrValA, sizeof(double) * nnzA *
batchSize);
    cudaStat2 = cudaMalloc ((void**) &d_csrColIndA, sizeof(int) * nnzA);
    cudaStat3 = cudaMalloc ((void**) &d_csrRowPtrA, sizeof(int) * (m+1));
    cudaStat4 = cudaMalloc ((void**) &d_b, sizeof(double) * m *
batchSize);
    cudaStat5 = cudaMalloc ((void**) &d_x, sizeof(double) * m *
batchSize);
    assert(cudaStat1 == cudaSuccess);
    assert(cudaStat2 == cudaSuccess);
    assert(cudaStat3 == cudaSuccess);
    assert(cudaStat4 == cudaSuccess);
    assert(cudaStat5 == cudaSuccess);

    cudaStat1 = cudaMemcpy(d_csrValA, csrValABatch, sizeof(double) * nnzA *
batchSize, cudaMemcpyHostToDevice);
    cudaStat2 = cudaMemcpy(d_csrColIndA, csrColIndA, sizeof(int) * nnzA,
cudaMemcpyHostToDevice);
    cudaStat3 = cudaMemcpy(d_csrRowPtrA, csrRowPtrA, sizeof(int) * (m+1),
cudaMemcpyHostToDevice);
    cudaStat4 = cudaMemcpy(d_b, bBatch, sizeof(double) * m * batchSize,
cudaMemcpyHostToDevice);
    assert(cudaStat1 == cudaSuccess);
    assert(cudaStat2 == cudaSuccess);
    assert(cudaStat3 == cudaSuccess);
    assert(cudaStat4 == cudaSuccess);

// step 4: symbolic analysis
    cusolver_status = cusolverSpXcsrqrAnalysisBatched(
        cusolverH, m, m, nnzA,
        descrA, d_csrRowPtrA, d_csrColIndA,
        info);
    assert(cusolver_status == CUSOLVER_STATUS_SUCCESS);

// step 5: prepare working space
    cusolver_status = cusolverSpDcsrqrBufferInfoBatched(
        cusolverH, m, m, nnzA,
        descrA, d_csrValA, d_csrRowPtrA, d_csrColIndA,
        batchSize,
        info,
        &size_internal,
        &size_qr);
    assert(cusolver_status == CUSOLVER_STATUS_SUCCESS);

    printf("numerical factorization needs internal data %lld bytes\n",
(long long) size_internal);
    printf("numerical factorization needs working space %lld bytes\n",
(long long) size_qr);

    cudaStat1 = cudaMalloc((void**) &buffer_qr, size_qr);
    assert(cudaStat1 == cudaSuccess);

```

Get results back

```

// step 6: numerical factorization
// assume device memory is big enough to compute all matrices.
cusolver_status = cusolverSpDcsrqrsvBatched(
    cusolverH, m, m, nnzA,
    descrA, d_csrValA, d_csrRowPtrA, d_csrColIndA,
    d_b, d_x,
    batchSize,
    info,
    buffer_qr);
assert(cusolver_status == CUSOLVER_STATUS_SUCCESS);

// step 7: check residual
// xBatch = [x0, x1, x2, ...]
cudaStat1 = cudaMemcpy(xBatch, d_x, sizeof(double)*m*batchSize,
    cudaMemcpyDeviceToHost);
assert(cudaStat1 == cudaSuccess);

const int baseA = (CUSPARSE_INDEX_BASE_ONE ==
    cusparseGetMatIndexBase(descrA)) ? 1:0 ;

for(int batchId = 0 ; batchId < batchSize; batchId++){
    // measure |bj - Aj*xj|
    double *csrValAj = csrValABatch + batchId * nnzA;
    double *xj = xBatch + batchId * m;
    double *bj = bBatch + batchId * m;
    // sup| bj - Aj*xj|
    double sup_res = 0;
    for(int row = 0 ; row < m ; row++){
        const int start = csrRowPtrA[row] - baseA;
        const int end = csrRowPtrA[row+1] - baseA;
        double Ax = 0.0; // Aj(row,:)*xj
        for(int colidx = start ; colidx < end ; colidx++){
            const int col = csrColIndA[colidx] - baseA;
            const double Areg = csrValAj[colidx];
            const double xreg = xj[col];
            Ax = Ax + Areg * xreg;
        }
        double r = bj[row] - Ax;
        sup_res = (sup_res > fabs(r)) ? sup_res : fabs(r);
    }
    printf("batchId %d: sup|bj - Aj*xj| = %E \n", batchId, sup_res);
}

for(int batchId = 0 ; batchId < batchSize; batchId++){
    double *xj = xBatch + batchId * m;
    for(int row = 0 ; row < m ; row++){
        printf("x%d[%d] = %E\n", batchId, row, xj[row]);
    }
    printf("\n");
}

return 0;
}

```

B.2. Batched Sparse QR example 2

This is the same as example 1 in appendix C except that we assume device memory is not enough, so we need to cut 17 matrices into several chunks and compute each chunk by batched sparse QR.

The usual includes and main definitions

```

#include <stdio.h>
#include <stdlib.h>
#include <assert.h>
#include <cusolverSp.h>
#include <cuda_runtime_api.h>

#define imin( x, y ) ((x)<(y)) ? (x) : (y)

int main(int argc, char*argv[])
{
    cusolverSpHandle_t cusolverH = NULL;
    // GPU does batch QR
    csrqrInfo_t info = NULL;
    cusparseMatDescr_t descrA = NULL;

    cusparseStatus_t cusparse_status = CUSPARSE_STATUS_SUCCESS;
    cusolverStatus_t cusolver_status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    cudaError_t cudaStat5 = cudaSuccess;

    // GPU does batch QR
    // d_A is CSR format, d_csrValA is of size nnzA*batchSize
    // d_x is a matrix of size batchSize * m
    // d_b is a matrix of size batchSize * m
    int *d_csrRowPtrA = NULL;
    int *d_csrColIndA = NULL;
    double *d_csrValA = NULL;
    double *d_b = NULL; // batchSize * m
    double *d_x = NULL; // batchSize * m

    size_t size_qr = 0;
    size_t size_internal = 0;
    void *buffer_qr = NULL; // working space for numerical factorization

    /*
    *   | 1      |
    * A = |      2      |
    *   |          3      |
    *   | 0.1  0.1  0.1  4 |
    *   CSR of A is based-1
    *
    * b = [1 1 1 1]
    */

```

Create the library handle

```

const int m = 4 ;
const int nnzA = 7;
const int csrRowPtrA[m+1] = { 1, 2, 3, 4, 8};
const int csrColIndA[nnzA] = { 1, 2, 3, 1, 2, 3, 4};
const double csrValA[nnzA] = { 1.0, 2.0, 3.0, 0.1, 0.1, 0.1, 4.0};
const double b[m] = {1.0, 1.0, 1.0, 1.0};
const int batchSize = 17;

double *csrValABatch = (double*)malloc(sizeof(double)*nnzA*batchSize);
double *bBatch       = (double*)malloc(sizeof(double)*m*batchSize);
double *xBatch       = (double*)malloc(sizeof(double)*m*batchSize);
assert( NULL != csrValABatch );
assert( NULL != bBatch );
assert( NULL != xBatch );

// step 1: prepare Aj and bj on host
// Aj is a small perturbation of A
// bj is a small perturbation of b
// csrValABatch = [A0, A1, A2, ...]
// bBatch = [b0, b1, b2, ...]
for(int colidx = 0 ; colidx < nnzA ; colidx++){
    double Areg = csrValA[colidx];
    for (int batchId = 0 ; batchId < batchSize ; batchId++){
        double eps = ((double)((rand() % 100) + 1)) * 1.e-4;
        csrValABatch[batchId*nnzA + colidx] = Areg + eps;
    }
}

for(int j = 0 ; j < m ; j++){
    double breg = b[j];
    for (int batchId = 0 ; batchId < batchSize ; batchId++){
        double eps = ((double)((rand() % 100) + 1)) * 1.e-4;
        bBatch[batchId*m + j] = breg + eps;
    }
}

// step 2: create cusolver handle, qr info and matrix descriptor
cusolver_status = cusolverSpCreate(&cusolverH);
assert (cusolver_status == CUSOLVER_STATUS_SUCCESS);

cusparsesetMatType(descrA, CUSPARSE_MATRIX_TYPE_GENERAL);
assert(cusparsesetMatType == CUSPARSE_STATUS_SUCCESS);

cusparsesetMatType(descrA, CUSPARSE_MATRIX_TYPE_GENERAL);
cusparsesetMatIndexBase(descrA, CUSPARSE_INDEX_BASE_ONE); // base-1

cusolver_status = cusolverSpCreateCsrqrInfo(&info);
assert(cusolver_status == CUSOLVER_STATUS_SUCCESS);

```

Set up the data

```

// step 3: copy Aj and bj to device
cudaStat1 = cudaMalloc ((void**) &d_csrValA, sizeof(double) * nnzA *
batchSize);
cudaStat2 = cudaMalloc ((void**) &d_csrColIndA, sizeof(int) * nnzA);
cudaStat3 = cudaMalloc ((void**) &d_csrRowPtrA, sizeof(int) * (m+1));
cudaStat4 = cudaMalloc ((void**) &d_b, sizeof(double) * m *
batchSize);
cudaStat5 = cudaMalloc ((void**) &d_x, sizeof(double) * m *
batchSize);
assert(cudaStat1 == cudaSuccess);
assert(cudaStat2 == cudaSuccess);
assert(cudaStat3 == cudaSuccess);
assert(cudaStat4 == cudaSuccess);
assert(cudaStat5 == cudaSuccess);

// don't copy csrValABatch and bBatch because device memory may be big enough
cudaStat1 = cudaMemcpy(d_csrColIndA, csrColIndA, sizeof(int) * nnzA,
cudaMemcpyHostToDevice);
cudaStat2 = cudaMemcpy(d_csrRowPtrA, csrRowPtrA, sizeof(int) * (m+1),
cudaMemcpyHostToDevice);
assert(cudaStat1 == cudaSuccess);
assert(cudaStat2 == cudaSuccess);

// step 4: symbolic analysis
cusolver_status = cusolverSpXcsrqrAnalysisBatched(
    cusolverH, m, m, nnzA,
    descrA, d_csrRowPtrA, d_csrColIndA,
    info);
assert(cusolver_status == CUSOLVER_STATUS_SUCCESS);

// step 5: find "proper" batchSize
// get available device memory
size_t free_mem = 0;
size_t total_mem = 0;
cudaStat1 = cudaMemGetInfo(&free_mem, &total_mem);
assert(cudaSuccess == cudaStat1);

int batchSizeMax = 2;
while(batchSizeMax < batchSize){
    printf("batchSizeMax = %d\n", batchSizeMax);
    cusolver_status = cusolverSpDcsrqrBufferInfoBatched(
        cusolverH, m, m, nnzA,
        // d_csrValA is don't care
        descrA, d_csrValA, d_csrRowPtrA, d_csrColIndA,
        batchSizeMax, // WARNING: use batchSizeMax
        info,
        &size_internal,
        &size_qr);
    assert(cusolver_status == CUSOLVER_STATUS_SUCCESS);

    if ( (size_internal + size_qr) > free_mem ){
        // current batchSizeMax exceeds hardware limit, so cut it by half.
        batchSizeMax /= 2; break;
    }
    batchSizeMax *= 2; // double batchSizeMax and try it again.
}
// correct batchSizeMax such that it is not greater than batchSize.
batchSizeMax = imin(batchSizeMax, batchSize);
printf("batchSizeMax = %d\n", batchSizeMax);

// Assume device memory is not big enough, and batchSizeMax = 2
batchSizeMax = 2;

```


Perform analysis and call solve

```

// step 6: prepare working space
// [necessary]
// Need to call cusolverDcsrqrBufferInfoBatched again with batchSizeMax
// to fix batchSize used in numerical factorization.
cusolver_status = cusolverSpDcsrqrBufferInfoBatched(
    cusolverH, m, m, nnzA,
    // d_csrValA is don't care
    descrA, d_csrValA, d_csrRowPtrA, d_csrColIndA,
    batchSizeMax, // WARNING: use batchSizeMax
    info,
    &size_internal,
    &size_qr);
assert(cusolver_status == CUSOLVER_STATUS_SUCCESS);

printf("numerical factorization needs internal data %lld bytes\n",
(long long)size_internal);
printf("numerical factorization needs working space %lld bytes\n",
(long long)size_qr);

cudaStat1 = cudaMalloc((void**)&buffer_qr, size_qr);
assert(cudaStat1 == cudaSuccess);

// step 7: solve  $A_j \cdot x_j = b_j$ 
for(int idx = 0 ; idx < batchSize; idx += batchSizeMax){
    // current batchSize 'cur_batchSize' is the batchSize used in numerical
    factorization
    const int cur_batchSize = imin(batchSizeMax, batchSize - idx);
    printf("current batchSize = %d\n", cur_batchSize);
    // copy part of  $A_j$  and  $b_j$  to device
    cudaStat1 = cudaMemcpy(d_csrValA, csrValABatch + idx*nnzA,
        sizeof(double) * nnzA * cur_batchSize, cudaMemcpyHostToDevice);
    cudaStat2 = cudaMemcpy(d_b, bBatch + idx*m,
        sizeof(double) * m * cur_batchSize, cudaMemcpyHostToDevice);
    assert(cudaStat1 == cudaSuccess);
    assert(cudaStat2 == cudaSuccess);
    // solve part of  $A_j \cdot x_j = b_j$ 
    cusolver_status = cusolverSpDcsrqrsvBatched(
        cusolverH, m, m, nnzA,
        descrA, d_csrValA, d_csrRowPtrA, d_csrColIndA,
        d_b, d_x,
        cur_batchSize, // WARNING: use current batchSize
        info,
        buffer_qr);
    assert(cusolver_status == CUSOLVER_STATUS_SUCCESS);
    // copy part of  $x_j$  back to host
    cudaStat1 = cudaMemcpy(xBatch + idx*m, d_x,
        sizeof(double) * m * cur_batchSize, cudaMemcpyDeviceToHost);
    assert(cudaStat1 == cudaSuccess);
}

```

Check results

```

// step 7: check residual
// xBatch = [x0, x1, x2, ...]
const int baseA = (CUSPARSE_INDEX_BASE_ONE ==
  cusparseGetMatIndexBase(descrA)) ? 1:0 ;

for(int batchId = 0 ; batchId < batchSize; batchId++){
  // measure |bj - Aj*xj|
  double *csrValAj = csrValABatch + batchId * nnzA;
  double *xj = xBatch + batchId * m;
  double *bj = bBatch + batchId * m;
  // sup| bj - Aj*xj|
  double sup_res = 0;
  for(int row = 0 ; row < m ; row++){
    const int start = csrRowPtrA[row] - baseA;
    const int end   = csrRowPtrA[row+1] - baseA;
    double Ax = 0.0; // Aj(row,:)*xj
    for(int colidx = start ; colidx < end ; colidx++){
      const int col = csrColIndA[colidx] - baseA;
      const double Areg = csrValAj[colidx];
      const double xreg = xj[col];
      Ax = Ax + Areg * xreg;
    }
    double r = bj[row] - Ax;
    sup_res = (sup_res > fabs(r)) ? sup_res : fabs(r);
  }
  printf("batchId %d: sup|bj - Aj*xj| = %E \n", batchId, sup_res);
}

for(int batchId = 0 ; batchId < batchSize; batchId++){
  double *xj = xBatch + batchId * m;
  for(int row = 0 ; row < m ; row++){
    printf("x%d[%d] = %E\n", batchId, row, xj[row]);
  }
  printf("\n");
}

return 0;
}

```

Appendix C.

QR EXAMPLES

C.1. QR Factorization Dense Linear Solver

This chapter provides a simple example in the C programming language of how to use a dense QR factorization to solve a linear system

$$Ax = b$$

A is a 3x3 dense matrix, nonsingular.

$$A = \begin{pmatrix} 1.0 & 2.0 & 3.0 \\ 4.0 & 5.0 & 6.0 \\ 2.0 & 1.0 & 1.0 \end{pmatrix}$$

The following code uses three steps:

Step 1: $A = Q^*R$ by `geqrf`.

Step 2: $B := Q^T B$ by `ormqr`.

Step 3: solve $R^*X = B$ by `trsm`.

The usual includes and main definition

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 * nvcc -c -I/usr/local/cuda/include ormqr_example.cpp
 * nvcc -o -fopenmp a.out ormqr_example.o -I/usr/local/cuda/lib64 -lcudart -
lcublas -lcusolver
 *
 */

#include <stdio.h>
#include <stdlib.h>
#include <assert.h>

#include <cuda_runtime.h>

#include <cublas_v2.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cublasHandle_t cublasH = NULL;
    cublasStatus_t cublas_status = CUBLAS_STATUS_SUCCESS;
    cusolverStatus_t cusolver_status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    const int m = 3;
    const int lda = m;
    const int ldb = m;
    const int nrhs = 1; // number of right hand side vectors
/*
 * | 1 2 3 |
 * A = | 4 5 6 |
 * | 2 1 1 |
 *
 * x = (1 1 1)'
 * b = (6 15 4)'
 */

```

Create the library handle and load the data

```

double A[lda*m] = { 1.0, 4.0, 2.0, 2.0, 5.0, 1.0, 3.0, 6.0, 1.0};
// double X[ldb*nrhs] = { 1.0, 1.0, 1.0}; // exact solution
double B[ldb*nrhs] = { 6.0, 15.0, 4.0};
double XC[ldb*nrhs]; // solution matrix from GPU

double *d_A = NULL; // linear memory of GPU
double *d_tau = NULL; // linear memory of GPU
double *d_B = NULL;
int *devInfo = NULL; // info in gpu (device copy)
double *d_work = NULL;
int lwork = 0;

int info_gpu = 0;

const double one = 1;

printf("A = (matlab base-1)\n");
printMatrix(m, m, A, lda, "A");
printf("=====\n");
printf("B = (matlab base-1)\n");
printMatrix(m, nrhs, B, ldb, "B");
printf("=====\n");

// step 1: create cusolver/cublas handle
cusolver_status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);

cublas_status = cublasCreate(&cublasH);
assert(CUBLAS_STATUS_SUCCESS == cublas_status);

// step 2: copy A and B to device
cudaStat1 = cudaMalloc ((void**)&d_A , sizeof(double) * lda * m);
cudaStat2 = cudaMalloc ((void**)&d_tau, sizeof(double) * m);
cudaStat3 = cudaMalloc ((void**)&d_B , sizeof(double) * ldb * nrhs);
cudaStat4 = cudaMalloc ((void**)&devInfo, sizeof(int));
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);
assert(cudaSuccess == cudaStat4);

cudaStat1 = cudaMemcpy(d_A, A, sizeof(double) * lda * m ,
cudaMemcpyHostToDevice);
cudaStat2 = cudaMemcpy(d_B, B, sizeof(double) * ldb * nrhs,
cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);

```

Call the solver

```

// step 3: query working space of geqrf and ormqr
cusolver_status = cusolverDnDgeqrf_bufferSize(
    cusolverH,
    m,
    m,
    d_A,
    lda,
    &lwork);
assert(cusolver_status == CUSOLVER_STATUS_SUCCESS);

cudaStat1 = cudaMalloc((void**) &d_work, sizeof(double)*lwork);
assert(cudaSuccess == cudaStat1);

// step 4: compute QR factorization
cusolver_status = cusolverDnDgeqrf(
    cusolverH,
    m,
    m,
    d_A,
    lda,
    d_tau,
    d_work,
    lwork,
    devInfo);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);
assert(cudaSuccess == cudaStat1);

// check if QR is good or not
cudaStat1 = cudaMemcpy(&info_gpu, devInfo, sizeof(int),
    cudaMemcpyDeviceToHost);
assert(cudaSuccess == cudaStat1);

printf("after geqrf: info_gpu = %d\n", info_gpu);
assert(0 == info_gpu);

// step 5: compute Q^T*B
cusolver_status = cusolverDnDormqr(
    cusolverH,
    CUBLAS_SIDE_LEFT,
    CUBLAS_OP_T,
    m,
    nrhs,
    m,
    d_A,
    lda,
    d_tau,
    d_B,
    ldb,
    d_work,
    lwork,
    devInfo);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);
assert(cudaSuccess == cudaStat1);

```

Check the results

```

    // check if QR is good or not
    cudaStat1 = cudaMemcpy(&info_gpu, devInfo, sizeof(int),
        cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);

    printf("after ormqr: info_gpu = %d\n", info_gpu);
    assert(0 == info_gpu);

// step 6: compute x = R \ Q^T*B

    cublas_status = cublasDtrsm(
        cublasH,
        CUBLAS_SIDE_LEFT,
        CUBLAS_FILL_MODE_UPPER,
        CUBLAS_OP_N,
        CUBLAS_DIAG_NON_UNIT,
        m,
        nrhs,
        &one,
        d_A,
        lda,
        d_B,
        ldb);
    cudaStat1 = cudaDeviceSynchronize();
    assert(CUBLAS_STATUS_SUCCESS == cublas_status);
    assert(cudaSuccess == cudaStat1);

    cudaStat1 = cudaMemcpy(XC, d_B, sizeof(double)*ldb*nrhs,
        cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);

    printf("X = (matlab base-1)\n");
    printMatrix(m, nrhs, XC, ldb, "X");

// free resources
    if (d_A) cudaFree(d_A);
    if (d_tau) cudaFree(d_tau);
    if (d_B) cudaFree(d_B);
    if (devInfo) cudaFree(devInfo);
    if (d_work) cudaFree(d_work);

    if (cublasH) cublasDestroy(cublasH);
    if (cusolverH) cusolverDnDestroy(cusolverH);

    cudaDeviceReset();

    return 0;
}

```

C.2. orthogonalization

This chapter provides a simple example in the C programming language of how to do orthogonalization by QR factorization.

$$A = Q * R$$

A is a 3x2 dense matrix,

$$A = \begin{pmatrix} 1.0 & 2.0 \\ 4.0 & 5.0 \\ 2.0 & 1.0 \end{pmatrix}$$

The following code uses three steps:

Step 1: $A = Q \cdot R$ by `geqrf`.

Step 2: form Q by `orgqr`.

Step 3: check if Q is unitary or not.

The usual includes and main definition

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 * nvcc -c -I/usr/local/cuda/include orgqr_example.cpp
 * g++ -fopenmp -o a.out orgqr_example.o -I/usr/local/cuda/lib64 -lcudart -
lcublas -lcusolver
 *
 */

#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <assert.h>

#include <cuda_runtime.h>

#include <cublas_v2.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cublasHandle_t cublasH = NULL;
    cublasStatus_t cublas_status = CUBLAS_STATUS_SUCCESS;
    cusolverStatus_t cusolver_status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    const int m = 3;
    const int n = 2;
    const int lda = m;

    /*
     *   A = | 1 2 |
     *       | 4 5 |
     *       | 2 1 |
     */

```


Create the library handle and load the data

```

double A[lda*n] = { 1.0, 4.0, 2.0, 2.0, 5.0, 1.0};
double Q[lda*n]; // orthonormal columns
double R[n*n]; // R = I - Q**T*Q

double *d_A = NULL;
double *d_tau = NULL;
int *devInfo = NULL;
double *d_work = NULL;

double *d_R = NULL;

int lwork_geqrf = 0;
int lwork_orgqr = 0;
int lwork = 0;

int info_gpu = 0;

const double h_one = 1;
const double h_minus_one = -1;

printf("A = (matlab base-1)\n");
printMatrix(m, n, A, lda, "A");
printf("=====\n");

// step 1: create cusolverDn/cublas handle
cusolver_status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);

cublas_status = cublasCreate(&cublasH);
assert(CUBLAS_STATUS_SUCCESS == cublas_status);

// step 2: copy A and B to device
cudaStat1 = cudaMalloc ((void**)&d_A , sizeof(double)*lda*n);
cudaStat2 = cudaMalloc ((void**)&d_tau, sizeof(double)*n);
cudaStat3 = cudaMalloc ((void**)&devInfo, sizeof(int));
cudaStat4 = cudaMalloc ((void**)&d_R , sizeof(double)*n*n);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);
assert(cudaSuccess == cudaStat4);

cudaStat1 = cudaMemcpy(d_A, A, sizeof(double)*lda*n,
cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);

```

Call the solver

```

// step 3: query working space of geqrf and orgqr
cusolver_status = cusolverDnDgeqrf_bufferSize(
    cusolverH,
    m,
    n,
    d_A,
    lda,
    &lwork_geqrf);
assert (cusolver_status == CUSOLVER_STATUS_SUCCESS);
cusolver_status = cusolverDnDorgqr_bufferSize(
    cusolverH,
    m,
    n,
    n,
    d_A,
    lda,
    &lwork_orgqr);
assert (cusolver_status == CUSOLVER_STATUS_SUCCESS);
// lwork = max(lwork_geqrf, lwork_orgqr)
lwork = (lwork_geqrf > lwork_orgqr)? lwork_geqrf : lwork_orgqr;

cudaStat1 = cudaMalloc((void**) &d_work, sizeof(double)*lwork);
assert(cudaSuccess == cudaStat1);

// step 4: compute QR factorization
cusolver_status = cusolverDnDgeqrf(
    cusolverH,
    m,
    n,
    d_A,
    lda,
    d_tau,
    d_work,
    lwork,
    devInfo);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);
assert(cudaSuccess == cudaStat1);

// check if QR is successful or not
cudaStat1 = cudaMemcpy(&info_gpu, devInfo, sizeof(int),
    cudaMemcpyDeviceToHost);
assert(cudaSuccess == cudaStat1);

printf("after geqrf: info_gpu = %d\n", info_gpu);
assert(0 == info_gpu);

// step 5: compute Q
cusolver_status = cusolverDnDorgqr(
    cusolverH,
    m,
    n,
    n,
    d_A,
    lda,
    d_tau,
    d_work,
    lwork,
    devInfo);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);
assert(cudaSuccess == cudaStat1);

```

Check the results

```

    // check if QR is good or not
    cudaStat1 = cudaMemcpy(&info_gpu, devInfo, sizeof(int),
        cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);

    printf("after orgqr: info_gpu = %d\n", info_gpu);
    assert(0 == info_gpu);

    cudaStat1 = cudaMemcpy(Q, d_A, sizeof(double)*lda*n,
        cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);

    printf("Q = (matlab base-1)\n");
    printMatrix(m, n, Q, lda, "Q");

// step 6: measure R = I - Q**T*Q
memset(R, 0, sizeof(double)*n*n);
for(int j = 0 ; j < n ; j++){
    R[j + n*j] = 1.0; // R(j,j)=1
}

cudaStat1 = cudaMemcpy(d_R, R, sizeof(double)*n*n, cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);

// R = -Q**T*Q + I
cublas_status = cublasDgemm_v2(
    cublasH,
    CUBLAS_OP_T, // Q**T
    CUBLAS_OP_N, // Q
    n, // number of rows of R
    n, // number of columns of R
    m, // number of columns of Q**T
    &h_minus_one, /* host pointer */
    d_A, // Q**T
    lda,
    d_A, // Q
    lda,
    &h_one, /* hostpointer */
    d_R,
    n);
assert(CUBLAS_STATUS_SUCCESS == cublas_status);

double dR_nrm2 = 0.0;
cublas_status = cublasDnrm2_v2(
    cublasH, n*n, d_R, 1, &dR_nrm2);
assert(CUBLAS_STATUS_SUCCESS == cublas_status);

printf("|I - Q**T*Q| = %E\n", dR_nrm2);

```

free resources

```
// free resources
if (d_A) cudaFree(d_A);
if (d_tau) cudaFree(d_tau);
if (devInfo) cudaFree(devInfo);
if (d_work) cudaFree(d_work);
if (d_R) cudaFree(d_R);

if (cublasH) cublasDestroy(cublasH);
if (cusolverH) cusolverDnDestroy(cusolverH);

cudaDeviceReset();

return 0;
}
```

Appendix D.

LU EXAMPLES

D.1. LU Factorization

This chapter provides a simple example in the C programming language of how to use a dense LU factorization to solve a linear system

$$Ax = b$$

A is a 3x3 dense matrix, nonsingular.

$$A = \begin{pmatrix} 1.0 & 2.0 & 3.0 \\ 4.0 & 5.0 & 6.0 \\ 7.0 & 8.0 & 10.0 \end{pmatrix}$$

The code uses `getrf` to do LU factorization and `getrs` to do backward and forward solve. The parameter `pivot_on` decides whether partial pivoting is performed or not.

...

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 * nvcc -c -I/usr/local/cuda/include getrf_example.cpp
 * g++ -fopenmp -o a.out getrf_example.o -I/usr/local/cuda/lib64 -lcusolver -lcudart
 */

#include <stdio.h>
#include <stdlib.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cudaStream_t stream = NULL;

    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    const int m = 3;
    const int lda = m;
    const int ldb = m;
    /*
     *   | 1 2 3 |
     *   A = | 4 5 6 |
     *       | 7 8 10 |
     *
     * without pivoting: A = L*U
     *   | 1 0 0 |   | 1 2 3 |
     *   L = | 4 1 0 |, U = | 0 -3 -6 |
     *       | 7 2 1 |   | 0 0 1 |
     *
     * with pivoting: P*A = L*U
     *   | 0 0 1 |
     *   P = | 1 0 0 |
     *       | 0 1 0 |
     *
     *   | 1      0      0 |   | 7 8      10 |
     *   L = | 0.1429 1      0 |, U = | 0 0.8571 1.5714 |
     *       | 0.5714 0.5    1 |   | 0 0      -0.5 |
     */

```

...

```

double A[lda*m] = { 1.0, 4.0, 7.0, 2.0, 5.0, 8.0, 3.0, 6.0, 10.0};
double B[m] = { 1.0, 2.0, 3.0 };
double X[m]; /* X = A\B */
double LU[lda*m]; /* L and U */
int Ipiv[m]; /* host copy of pivoting sequence */
int info = 0; /* host copy of error info */

double *d_A = NULL; /* device copy of A */
double *d_B = NULL; /* device copy of B */
int *d_Ipiv = NULL; /* pivoting sequence */
int *d_info = NULL; /* error info */
int lwork = 0; /* size of workspace */
double *d_work = NULL; /* device workspace for getrf */

const int pivot_on = 0;

printf("example of getrf \n");

if (pivot_on){
    printf("pivot is on : compute P*A = L*U \n");
}else{
    printf("pivot is off: compute A = L*U (not numerically stable)\n");
}

printf("A = (matlab base-1)\n");
printMatrix(m, m, A, lda, "A");
printf("====\n");

printf("B = (matlab base-1)\n");
printMatrix(m, 1, B, ldb, "B");
printf("====\n");

/* step 1: create cusolver handle, bind a stream */
status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaStreamCreateWithFlags(&stream, cudaStreamNonBlocking);
assert(cudaSuccess == cudaStat1);

status = cusolverDnSetStream(cusolverH, stream);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 2: copy A to device */
cudaStat1 = cudaMalloc ((void**)&d_A, sizeof(double) * lda * m);
cudaStat2 = cudaMalloc ((void**)&d_B, sizeof(double) * m);
cudaStat2 = cudaMalloc ((void**)&d_Ipiv, sizeof(int) * m);
cudaStat4 = cudaMalloc ((void**)&d_info, sizeof(int));
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);
assert(cudaSuccess == cudaStat4);

cudaStat1 = cudaMemcpy(d_A, A, sizeof(double)*lda*m,
cudaMemcpyHostToDevice);
cudaStat2 = cudaMemcpy(d_B, B, sizeof(double)*m, cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);

```

...

```

/* step 3: query working space of getrf */
status = cusolverDnDgetrf_bufferSize(
    cusolverH,
    m,
    m,
    d_A,
    lda,
    &lwork);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaMalloc((void**) &d_work, sizeof(double)*lwork);
assert(cudaSuccess == cudaStat1);

/* step 4: LU factorization */
if (pivot_on){
    status = cusolverDnDgetrf(
        cusolverH,
        m,
        m,
        d_A,
        lda,
        d_work,
        d_Ipiv,
        d_info);
}else{
    status = cusolverDnDgetrf(
        cusolverH,
        m,
        m,
        d_A,
        lda,
        d_work,
        NULL,
        d_info);
}
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == status);
assert(cudaSuccess == cudaStat1);

if (pivot_on){
    cudaStat1 = cudaMemcpy(Ipiv, d_Ipiv, sizeof(int)*m,
        cudaMemcpyDeviceToHost);
}
cudaStat2 = cudaMemcpy(LU, d_A, sizeof(double)*lda*m,
    cudaMemcpyDeviceToHost);
cudaStat3 = cudaMemcpy(&info, d_info, sizeof(int), cudaMemcpyDeviceToHost);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);

if (0 > info){
    printf("%d-th parameter is wrong \n", -info);
    exit(1);
}
if (pivot_on){
    printf("pivoting sequence, matlab base-1\n");
    for(int j = 0 ; j < m ; j++){
        printf("Ipiv(%d) = %d\n", j+1, Ipiv[j]);
    }
}
printf("L and U = (matlab base-1)\n");
printMatrix(m, m, LU, lda, "LU");
printf("=====\n");

```


...

```

/*
 * step 5: solve A*X = B
 *      B = | 1 |,   X = | -0.3333 |
 *           | 2 |,       | 0.6667 |
 *           | 3 |       | 0       |
 */
if (pivot_on){
    status = cusolverDnDgetrs(
        cusolverH,
        CUBLAS_OP_N,
        m,
        1, /* nrhs */
        d_A,
        lda,
        d_Ipiv,
        d_B,
        ldb,
        d_info);
} else {
    status = cusolverDnDgetrs(
        cusolverH,
        CUBLAS_OP_N,
        m,
        1, /* nrhs */
        d_A,
        lda,
        NULL,
        d_B,
        ldb,
        d_info);
}
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == status);
assert(cudaSuccess == cudaStat1);

cudaStat1 = cudaMemcpy(X, d_B, sizeof(double)*m, cudaMemcpyDeviceToHost);
assert(cudaSuccess == cudaStat1);

printf("X = (matlab base-1)\n");
printMatrix(m, 1, X, ldb, "X");
printf("====\n");

/* free resources */
if (d_A) cudaFree(d_A);
if (d_B) cudaFree(d_B);
if (d_Ipiv) cudaFree(d_Ipiv);
if (d_info) cudaFree(d_info);
if (d_work) cudaFree(d_work);

if (cusolverH) cusolverDnDestroy(cusolverH);
if (stream) cudaStreamDestroy(stream);

cudaDeviceReset();

return 0;
}

```

Appendix E.

CHOLSKY EXAMPLES

E.1. batched Cholesky Factorization

This chapter provides a simple example in the C programming language of how to use a batched dense Cholesky factorization to solve a sequence of linear systems

$$\mathbf{A}[i] * \mathbf{x}[i] = \mathbf{b}[i]$$

each $\mathbf{A}[i]$ is a 3x3 dense Hermitian matrix. In this example, there are two matrices, $\mathbf{A0}$ and $\mathbf{A1}$. $\mathbf{A0}$ is positive definite and $\mathbf{A1}$ is not.

The code uses `potrfBatched` to do Cholesky factorization and `potrsBatched` to do backward and forward solve. `potrfBatched` would report singularity on $\mathbf{A1}$.

```

...

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 * nvcc -c -I/usr/local/cuda/include batchchol_example.cpp
 * g++ -o a.out batchchol_example.o -L/usr/local/cuda/lib64 -lcusolver -
lcudart
 */

#include <stdio.h>
#include <stdlib.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t handle = NULL;
    cudaStream_t stream = NULL;

    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;

    const cublasFillMode_t uplo = CUBLAS_FILL_MODE_LOWER;
    const int batchSize = 2;
    const int nrhs = 1;
    const int m = 3;
    const int lda = m;
    const int ldb = m;

    /*
     *      | 1      2      3 |
     * A0 = | 2      5      5 | = L0 * L0**T
     *      | 3      5     12 |
     *
     *      | 1.0000      0      0 |
     * where L0 = | 2.0000      1.0000      0 |
     *             | 3.0000     -1.0000      1.4142 |
     *
     *      | 1      2      3 |
     * A1 = | 2      4      5 | is not s.p.d., failed at row 2
     *      | 3      5     12 |
     *
     */
}

```

...

```

double A0[lda*m] = { 1.0, 2.0, 3.0, 2.0, 5.0, 5.0, 3.0, 5.0, 12.0 };
double A1[lda*m] = { 1.0, 2.0, 3.0, 2.0, 4.0, 5.0, 3.0, 5.0, 12.0 };
double B0[m] = { 1.0, 1.0, 1.0 };
double X0[m]; /* X0 = A0\B0 */
int infoArray[batchSize]; /* host copy of error info */

double L0[lda*m]; /* cholesky factor of A0 */

double *Aarray[batchSize];
double *Barray[batchSize];

double **d_Aarray = NULL;
double **d_Barray = NULL;
int *d_infoArray = NULL;

printf("example of batched Cholesky \n");

printf("A0 = (matlab base-1)\n");
printMatrix(m, m, A0, lda, "A0");
printf("=====\n");

printf("A1 = (matlab base-1)\n");
printMatrix(m, m, A1, lda, "A1");
printf("=====\n");

printf("B0 = (matlab base-1)\n");
printMatrix(m, 1, B0, ldb, "B0");
printf("=====\n");

/* step 1: create cusolver handle, bind a stream */
status = cusolverDnCreate(&handle);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaStreamCreateWithFlags(&stream, cudaStreamNonBlocking);
assert(cudaSuccess == cudaStat1);

status = cusolverDnSetStream(handle, stream);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 2: copy A to device */
for(int j = 0 ; j < batchSize ; j++){
    cudaStat1 = cudaMalloc ((void**)&Aarray[j], sizeof(double) * lda * m);
    assert(cudaSuccess == cudaStat1);
    cudaStat2 = cudaMalloc ((void**)&Barray[j], sizeof(double) * ldb *
nrhs);
    assert(cudaSuccess == cudaStat2);
}
cudaStat1 = cudaMalloc ((void**)&d_infoArray, sizeof(int)*batchSize);
assert(cudaSuccess == cudaStat1);

cudaStat1 = cudaMalloc ((void**)&d_Aarray, sizeof(double*) * batchSize);
cudaStat2 = cudaMalloc ((void**)&d_Barray, sizeof(double*) * batchSize);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);

cudaStat1 = cudaMemcpy(Aarray[0], A0, sizeof(double) * lda * m,
cudaMemcpyHostToDevice);
cudaStat2 = cudaMemcpy(Aarray[1], A1, sizeof(double) * lda * m,
cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);

```

...

```

    cudaStat1 = cudaMemcpy(Barray[0], B0, sizeof(double) * m,
cudaMemcpyHostToDevice);
    cudaStat2 = cudaMemcpy(Barray[1], B0, sizeof(double) * m,
cudaMemcpyHostToDevice);
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);

    cudaStat1 = cudaMemcpy(d_Aarray, Aarray, sizeof(double)*batchSize,
cudaMemcpyHostToDevice);
    cudaStat2 = cudaMemcpy(d_Barray, Barray, sizeof(double)*batchSize,
cudaMemcpyHostToDevice);
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    cudaDeviceSynchronize();

/* step 3: Cholesky factorization */
    status = cusolverDnDpotrfBatched(
        handle,
        uplo,
        m,
        d_Aarray,
        lda,
        d_infoArray,
        batchSize);
    cudaStat1 = cudaDeviceSynchronize();
    assert(CUSOLVER_STATUS_SUCCESS == status);
    assert(cudaSuccess == cudaStat1);

    cudaStat1 = cudaMemcpy(infoArray, d_infoArray, sizeof(int)*batchSize,
cudaMemcpyDeviceToHost);
    cudaStat2 = cudaMemcpy(L0, Aarray[0], sizeof(double) * lda * m,
cudaMemcpyDeviceToHost);
    cudaDeviceSynchronize();
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);

    for(int j = 0 ; j < batchSize ; j++){
        printf("info[%d] = %d\n", j, infoArray[j]);
    }

    assert( 0 == infoArray[0] );
/* A1 is singular */
    assert( 2 == infoArray[1] );

    printf("L = (matlab base-1), upper triangle is don't care \n");
    printMatrix(m, m, L0, lda, "L0");
    printf("=====\n");

/*
 * step 4: solve A0*X0 = B0
 *      | 1 |      | 10.5 |
 *   B0 = | 1 |,   X0 = | -2.5 |
 *      | 1 |      | -1.5 |
 */
    status = cusolverDnDpotrsBatched(
        handle,
        uplo,
        m,
        nrhs, /* only support rhs = 1*/
        d_Aarray,
        lda,
        d_Barray,
        ldb,
        d_infoArray,
        batchSize);

```

...

```

    cudaStat1 = cudaDeviceSynchronize();
    assert(CUSOLVER_STATUS_SUCCESS == status);
    assert(cudaSuccess == cudaStat1);

    cudaStat1 = cudaMemcpy(infoArray, d_infoArray, sizeof(int),
        cudaMemcpyDeviceToHost);
    cudaStat2 = cudaMemcpy(X0 , Barray[0], sizeof(double)*m,
        cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    cudaDeviceSynchronize();

    printf("info = %d\n", infoArray[0]);
    assert( 0 == infoArray[0] );

    printf("X0 = (matlab base-1)\n");
    printMatrix(m, 1, X0, ldb, "X0");
    printf("=====\n");

/* free resources */
    if (d_Aarray ) cudaFree(d_Aarray);
    if (d_Barray ) cudaFree(d_Barray);
    if (d_infoArray ) cudaFree(d_infoArray);

    if (handle ) cusolverDnDestroy(handle);
    if (stream ) cudaStreamDestroy(stream);

    cudaDeviceReset();

    return 0;
}

```

Appendix F.

EXAMPLES OF DENSE EIGENVALUE SOLVER

F.1. Standard Symmetric Dense Eigenvalue Solver

This chapter provides a simple example in the C programming language of how to use `syevd` to compute the spectrum of a dense symmetric system by

$$Ax = \lambda x$$

where A is a 3x3 dense symmetric matrix

$$A = \begin{pmatrix} 3.5 & 0.5 & 0 \\ 0.5 & 3.5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

The following code uses **syevd** to compute eigenvalues and eigenvectors, then compare to exact eigenvalues {2,3,4}.

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *   nvcc -c -I/usr/local/cuda/include syevd_example.cpp
 *   g++ -o a.out syevd_example.o -L/usr/local/cuda/lib64 -lcudart -lcusolver
 */

#include <stdio.h>
#include <stdlib.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cusolverStatus_t cusolver_status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    const int m = 3;
    const int lda = m;
    /*
     *   | 3.5 0.5 0 |
     *   A = | 0.5 3.5 0 |
     *       | 0   0  2 |
     */
    double A[lda*m] = { 3.5, 0.5, 0, 0.5, 3.5, 0, 0, 0, 2.0};
    double lambda[m] = { 2.0, 3.0, 4.0};

    double V[lda*m]; // eigenvectors
    double W[m]; // eigenvalues

    double *d_A = NULL;
    double *d_W = NULL;
    int *devInfo = NULL;
    double *d_work = NULL;
    int lwork = 0;

    int info_gpu = 0;

    printf("A = (matlab base-1)\n");
    printMatrix(m, m, A, lda, "A");
    printf("=====\n");

```


call eigenvalue solver

```

// step 1: create cusolver/cublas handle
cusolver_status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);

// step 2: copy A and B to device
cudaStat1 = cudaMalloc ((void**)&d_A, sizeof(double) * lda * m);
cudaStat2 = cudaMalloc ((void**)&d_W, sizeof(double) * m);
cudaStat3 = cudaMalloc ((void**)&devInfo, sizeof(int));
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);

    cudaStat1 = cudaMemcpy(d_A, A, sizeof(double) * lda * m,
        cudaMemcpyHostToDevice);
    assert(cudaSuccess == cudaStat1);

// step 3: query working space of syevd
cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR; // compute eigenvalues
and eigenvectors.
cublasFillMode_t uplo = CUBLAS_FILL_MODE_LOWER;
cusolver_status = cusolverDnDsyevd_bufferSize(
    cusolverH,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_W,
    &lwork);
assert (cusolver_status == CUSOLVER_STATUS_SUCCESS);

    cudaStat1 = cudaMalloc((void**)&d_work, sizeof(double)*lwork);
    assert(cudaSuccess == cudaStat1);

// step 4: compute spectrum
cusolver_status = cusolverDnDsyevd(
    cusolverH,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_W,
    d_work,
    lwork,
    devInfo);
    cudaStat1 = cudaDeviceSynchronize();
    assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);
    assert(cudaSuccess == cudaStat1);

    cudaStat1 = cudaMemcpy(W, d_W, sizeof(double)*m, cudaMemcpyDeviceToHost);
    cudaStat2 = cudaMemcpy(V, d_A, sizeof(double)*lda*m,
        cudaMemcpyDeviceToHost);
    cudaStat3 = cudaMemcpy(&info_gpu, devInfo, sizeof(int),
        cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    assert(cudaSuccess == cudaStat3);

```

check the result

```
printf("after syevd: info_gpu = %d\n", info_gpu);
assert(0 == info_gpu);

printf("eigenvalue = (matlab base-1), ascending order\n");
for(int i = 0 ; i < m ; i++){
    printf("W[%d] = %E\n", i+1, W[i]);
}

printf("V = (matlab base-1)\n");
printMatrix(m, m, V, lda, "V");
printf("=====\n");

// step 4: check eigenvalues
double lambda_sup = 0;
for(int i = 0 ; i < m ; i++){
    double error = fabs( lambda[i] - W[i]);
    lambda_sup = (lambda_sup > error)? lambda_sup : error;
}
printf("|lambda - W| = %E\n", lambda_sup);

// free resources
if (d_A ) cudaFree(d_A);
if (d_W ) cudaFree(d_W);
if (devInfo) cudaFree(devInfo);
if (d_work ) cudaFree(d_work);

if (cusolverH) cusolverDnDestroy(cusolverH);

cudaDeviceReset();

return 0;
}
```

F.2. Standard Symmetric Dense Eigenvalue Solver

This chapter provides a simple example in the C programming language of how to use syevd to compute the spectrum of a dense symmetric system by

$$Ax = \lambda x$$

where A is a 3x3 dense symmetric matrix

$$A = \begin{pmatrix} 3.5 & 0.5 & 0 \\ 0.5 & 3.5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

The following code uses **syevd** to compute eigenvalues and eigenvectors, then compare to exact eigenvalues {2,3,4}.

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *   nvcc -c -I/usr/local/cuda/include syevd_example.cpp
 *   g++ -o a.out syevd_example.o -L/usr/local/cuda/lib64 -lcudart -lcusolver
 */

#include <stdio.h>
#include <stdlib.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cusolverStatus_t cusolver_status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    const int m = 3;
    const int lda = m;
    /*
     *   | 3.5 0.5 0 |
     *   A = | 0.5 3.5 0 |
     *       | 0   0  2 |
     */
    double A[lda*m] = { 3.5, 0.5, 0, 0.5, 3.5, 0, 0, 0, 2.0};
    double lambda[m] = { 2.0, 3.0, 4.0};

    double V[lda*m]; // eigenvectors
    double W[m]; // eigenvalues

    double *d_A = NULL;
    double *d_W = NULL;
    int *devInfo = NULL;
    double *d_work = NULL;
    int lwork = 0;

    int info_gpu = 0;

    printf("A = (matlab base-1)\n");
    printMatrix(m, m, A, lda, "A");
    printf("=====\n");

```

call eigenvalue solver

```

// step 1: create cusolver/cublas handle
cusolver_status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);

// step 2: copy A and B to device
cudaStat1 = cudaMalloc ((void**)&d_A, sizeof(double) * lda * m);
cudaStat2 = cudaMalloc ((void**)&d_W, sizeof(double) * m);
cudaStat3 = cudaMalloc ((void**)&devInfo, sizeof(int));
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);

    cudaStat1 = cudaMemcpy(d_A, A, sizeof(double) * lda * m,
        cudaMemcpyHostToDevice);
    assert(cudaSuccess == cudaStat1);

// step 3: query working space of syevd
cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR; // compute eigenvalues
and eigenvectors.
cublasFillMode_t uplo = CUBLAS_FILL_MODE_LOWER;
cusolver_status = cusolverDnDsyevd_bufferSize(
    cusolverH,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_W,
    &lwork);
assert (cusolver_status == CUSOLVER_STATUS_SUCCESS);

    cudaStat1 = cudaMalloc((void**)&d_work, sizeof(double)*lwork);
    assert(cudaSuccess == cudaStat1);

// step 4: compute spectrum
cusolver_status = cusolverDnDsyevd(
    cusolverH,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_W,
    d_work,
    lwork,
    devInfo);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);
assert(cudaSuccess == cudaStat1);

    cudaStat1 = cudaMemcpy(W, d_W, sizeof(double)*m, cudaMemcpyDeviceToHost);
    cudaStat2 = cudaMemcpy(V, d_A, sizeof(double)*lda*m,
        cudaMemcpyDeviceToHost);
    cudaStat3 = cudaMemcpy(&info_gpu, devInfo, sizeof(int),
        cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    assert(cudaSuccess == cudaStat3);

```

check the result

```
printf("after syevd: info_gpu = %d\n", info_gpu);
assert(0 == info_gpu);

printf("eigenvalue = (matlab base-1), ascending order\n");
for(int i = 0 ; i < m ; i++){
    printf("W[%d] = %E\n", i+1, W[i]);
}

printf("V = (matlab base-1)\n");
printMatrix(m, m, V, lda, "V");
printf("=====\n");

// step 4: check eigenvalues
double lambda_sup = 0;
for(int i = 0 ; i < m ; i++){
    double error = fabs( lambda[i] - W[i]);
    lambda_sup = (lambda_sup > error)? lambda_sup : error;
}
printf("|lambda - W| = %E\n", lambda_sup);

// free resources
if (d_A ) cudaFree(d_A);
if (d_W ) cudaFree(d_W);
if (devInfo) cudaFree(devInfo);
if (d_work ) cudaFree(d_work);

if (cusolverH) cusolverDnDestroy(cusolverH);

cudaDeviceReset();

return 0;
}
```

F.3. Generalized Symmetric-Definite Dense Eigenvalue Solver

This chapter provides a simple example in the C programming language of how to use sygvd to compute spectrum of a pair of dense symmetric matrices (A,B) by

$$Ax = \lambda Bx$$

where A is a 3x3 dense symmetric matrix

$$A = \begin{pmatrix} 3.5 & 0.5 & 0 \\ 0.5 & 3.5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

and B is a 3x3 positive definite matrix

$$B = \begin{pmatrix} 10 & 2 & 3 \\ 2 & 10 & 5 \\ 3 & 5 & 10 \end{pmatrix}$$

The following code uses **sygvd** to compute eigenvalues and eigenvectors, then compare to exact eigenvalues {0.158660256604, 0.370751508101882, 0.6}.

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *   nvcc -c -I/usr/local/cuda/include sygvd_example.cpp
 *   g++ -o a.out sygvd_example.o -L/usr/local/cuda/lib64 -lcusolver
 */
#include <stdio.h>
#include <stdlib.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cusolverStatus_t cusolver_status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    const int m = 3;
    const int lda = m;

    /*
     *      | 3.5 0.5 0 |
     *   A = | 0.5 3.5 0 |
     *      | 0   0  2 |
     *
     *      | 10  2  3 |
     *   B = | 2   10  5 |
     *      | 3   5  10 |
     */
    double A[lda*m] = { 3.5, 0.5, 0, 0.5, 3.5, 0, 0, 0, 2.0};
    double B[lda*m] = { 10.0, 2.0, 3.0, 2.0, 10.0, 5.0, 3.0, 5.0, 10.0};
    double lambda[m] = { 0.158660256604, 0.370751508101882, 0.6};

    double V[lda*m]; // eigenvectors
    double W[m]; // eigenvalues

    double *d_A = NULL;
    double *d_B = NULL;
    double *d_W = NULL;
    int *devInfo = NULL;
    double *d_work = NULL;
    int lwork = 0;
    int info_gpu = 0;

    printf("A = (matlab base-1)\n");
    printMatrix(m, m, A, lda, "A");
    printf("=====\n");

    printf("B = (matlab base-1)\n");
    printMatrix(m, m, B, lda, "B");
    printf("=====\n");

```

call eigenvalue solver

```

// step 1: create cusolver/cublas handle
cusolver_status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);

// step 2: copy A and B to device
cudaStat1 = cudaMalloc ((void**)&d_A, sizeof(double) * lda * m);
cudaStat2 = cudaMalloc ((void**)&d_B, sizeof(double) * lda * m);
cudaStat3 = cudaMalloc ((void**)&d_W, sizeof(double) * m);
cudaStat4 = cudaMalloc ((void**)&devInfo, sizeof(int));
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);
assert(cudaSuccess == cudaStat4);

cudaStat1 = cudaMemcpy(d_A, A, sizeof(double) * lda * m,
cudaMemcpyHostToDevice);
cudaStat2 = cudaMemcpy(d_B, B, sizeof(double) * lda * m,
cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);

// step 3: query working space of sygv
cusolverEigType_t itype = CUSOLVER_EIG_TYPE_1; // A*x = (lambda)*B*x
cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR; // compute eigenvalues
and eigenvectors.
cublasFillMode_t uplo = CUBLAS_FILL_MODE_LOWER;
cusolver_status = cusolverDnDsygv_bufferSize(
    cusolverH,
    itype,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_B,
    lda,
    d_W,
    &lwork);
assert (cusolver_status == CUSOLVER_STATUS_SUCCESS);
cudaStat1 = cudaMalloc((void**)&d_work, sizeof(double)*lwork);
assert(cudaSuccess == cudaStat1);

// step 4: compute spectrum of (A,B)
cusolver_status = cusolverDnDsygv(
    cusolverH,
    itype,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_B,
    lda,
    d_W,
    d_work,
    lwork,
    devInfo);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);
assert(cudaSuccess == cudaStat1);

```

check the result

```

    cudaStat1 = cudaMemcpy(W, d_W, sizeof(double)*m, cudaMemcpyDeviceToHost);
    cudaStat2 = cudaMemcpy(V, d_A, sizeof(double)*lda*m,
cudaMemcpyDeviceToHost);
    cudaStat3 = cudaMemcpy(&info_gpu, devInfo, sizeof(int),
cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    assert(cudaSuccess == cudaStat3);

    printf("after sygvd: info_gpu = %d\n", info_gpu);
    assert(0 == info_gpu);

    printf("eigenvalue = (matlab base-1), ascending order\n");
    for(int i = 0 ; i < m ; i++){
        printf("W[%d] = %E\n", i+1, W[i]);
    }

    printf("V = (matlab base-1)\n");
    printMatrix(m, m, V, lda, "V");
    printf("=====\n");

// step 4: check eigenvalues
double lambda_sup = 0;
for(int i = 0 ; i < m ; i++){
    double error = fabs( lambda[i] - W[i]);
    lambda_sup = (lambda_sup > error)? lambda_sup : error;
}
printf("|lambda - W| = %E\n", lambda_sup);

// free resources
if (d_A ) cudaFree(d_A);
if (d_B ) cudaFree(d_B);
if (d_W ) cudaFree(d_W);
if (devInfo) cudaFree(devInfo);
if (d_work ) cudaFree(d_work);

if (cusolverH) cusolverDnDestroy(cusolverH);

cudaDeviceReset();

return 0;
}

```

F.4. Generalized Symmetric-Definite Dense Eigenvalue Solver

This chapter provides a simple example in the C programming language of how to use sygvd to compute spectrum of a pair of dense symmetric matrices (A,B) by

$$Ax = \lambda Bx$$

where A is a 3x3 dense symmetric matrix

$$A = \begin{pmatrix} 3.5 & 0.5 & 0 \\ 0.5 & 3.5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

and B is a 3x3 positive definite matrix

$$B = \begin{pmatrix} 10 & 2 & 3 \\ 2 & 10 & 5 \\ 3 & 5 & 10 \end{pmatrix}$$

The following code uses **sygvd** to compute eigenvalues and eigenvectors, then compare to exact eigenvalues {0.158660256604, 0.370751508101882, 0.6}.

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *   nvcc -c -I/usr/local/cuda/include sygvd_example.cpp
 *   g++ -o a.out sygvd_example.o -L/usr/local/cuda/lib64 -lcusolver
 */
#include <stdio.h>
#include <stdlib.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cusolverStatus_t cusolver_status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    const int m = 3;
    const int lda = m;

    /*
     *      | 3.5 0.5 0 |
     *   A = | 0.5 3.5 0 |
     *      | 0   0  2 |
     *
     *      | 10  2  3 |
     *   B = | 2   10  5 |
     *      | 3   5  10 |
     */
    double A[lda*m] = { 3.5, 0.5, 0, 0.5, 3.5, 0, 0, 0, 2.0};
    double B[lda*m] = { 10.0, 2.0, 3.0, 2.0, 10.0, 5.0, 3.0, 5.0, 10.0};
    double lambda[m] = { 0.158660256604, 0.370751508101882, 0.6};

    double V[lda*m]; // eigenvectors
    double W[m]; // eigenvalues

    double *d_A = NULL;
    double *d_B = NULL;
    double *d_W = NULL;
    int *devInfo = NULL;
    double *d_work = NULL;
    int lwork = 0;
    int info_gpu = 0;

    printf("A = (matlab base-1)\n");
    printMatrix(m, m, A, lda, "A");
    printf("=====\n");

    printf("B = (matlab base-1)\n");
    printMatrix(m, m, B, lda, "B");
    printf("=====\n");

```

call eigenvalue solver

```

// step 1: create cusolver/cublas handle
cusolver_status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);

// step 2: copy A and B to device
cudaStat1 = cudaMalloc ((void**)&d_A, sizeof(double) * lda * m);
cudaStat2 = cudaMalloc ((void**)&d_B, sizeof(double) * lda * m);
cudaStat3 = cudaMalloc ((void**)&d_W, sizeof(double) * m);
cudaStat4 = cudaMalloc ((void**)&devInfo, sizeof(int));
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);
assert(cudaSuccess == cudaStat4);

cudaStat1 = cudaMemcpy(d_A, A, sizeof(double) * lda * m,
cudaMemcpyHostToDevice);
cudaStat2 = cudaMemcpy(d_B, B, sizeof(double) * lda * m,
cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);

// step 3: query working space of sygv
cusolverEigType_t itype = CUSOLVER_EIG_TYPE_1; // A*x = (lambda)*B*x
cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR; // compute eigenvalues
and eigenvectors.
cublasFillMode_t uplo = CUBLAS_FILL_MODE_LOWER;
cusolver_status = cusolverDnDsygv_bufferSize(
    cusolverH,
    itype,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_B,
    lda,
    d_W,
    &lwork);
assert (cusolver_status == CUSOLVER_STATUS_SUCCESS);
cudaStat1 = cudaMalloc((void**)&d_work, sizeof(double)*lwork);
assert(cudaSuccess == cudaStat1);

// step 4: compute spectrum of (A,B)
cusolver_status = cusolverDnDsygv(
    cusolverH,
    itype,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_B,
    lda,
    d_W,
    d_work,
    lwork,
    devInfo);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);
assert(cudaSuccess == cudaStat1);

```

check the result

```

    cudaStat1 = cudaMemcpy(W, d_W, sizeof(double)*m, cudaMemcpyDeviceToHost);
    cudaStat2 = cudaMemcpy(V, d_A, sizeof(double)*lda*m,
cudaMemcpyDeviceToHost);
    cudaStat3 = cudaMemcpy(&info_gpu, devInfo, sizeof(int),
cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    assert(cudaSuccess == cudaStat3);

    printf("after sygvd: info_gpu = %d\n", info_gpu);
    assert(0 == info_gpu);

    printf("eigenvalue = (matlab base-1), ascending order\n");
    for(int i = 0 ; i < m ; i++){
        printf("W[%d] = %E\n", i+1, W[i]);
    }

    printf("V = (matlab base-1)\n");
    printMatrix(m, m, V, lda, "V");
    printf("=====\n");

// step 4: check eigenvalues
double lambda_sup = 0;
for(int i = 0 ; i < m ; i++){
    double error = fabs( lambda[i] - W[i]);
    lambda_sup = (lambda_sup > error)? lambda_sup : error;
}
printf("|lambda - W| = %E\n", lambda_sup);

// free resources
if (d_A ) cudaFree(d_A);
if (d_B ) cudaFree(d_B);
if (d_W ) cudaFree(d_W);
if (devInfo) cudaFree(devInfo);
if (d_work ) cudaFree(d_work);

if (cusolverH) cusolverDnDestroy(cusolverH);

cudaDeviceReset();

return 0;
}

```

F.5. Standard Symmetric Dense Eigenvalue Solver (via Jacobi method)

This chapter provides a simple example in the C programming language of how to use **syevj** to compute the spectrum of a dense symmetric system by

$$Ax = \lambda x$$

where A is a 3x3 dense symmetric matrix

$$A = \begin{pmatrix} 3.5 & 0.5 & 0 \\ 0.5 & 3.5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

The following code uses **syevj** to compute eigenvalues and eigenvectors, then compare to exact eigenvalues {2,3,4}.

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *   nvcc -c -I/usr/local/cuda/include syevj_example.cpp
 *   g++ -o syevj_example syevj_example.o -L/usr/local/cuda/lib64 -lcusolver -lcudart
 */
#include <stdio.h>
#include <stdlib.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cudaStream_t stream = NULL;
    syevjInfo_t syevj_params = NULL;

    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    const int m = 3;
    const int lda = m;
/*
 *   | 3.5 0.5 0 |
 *   A = | 0.5 3.5 0 |
 *       | 0   0  2 |
 *
 */
    double A[lda*m] = { 3.5, 0.5, 0, 0.5, 3.5, 0, 0, 0, 2.0};
    double lambda[m] = { 2.0, 3.0, 4.0};

    double V[lda*m]; /* eigenvectors */
    double W[m];      /* eigenvalues */

    double *d_A = NULL; /* device copy of A */
    double *d_W = NULL; /* eigenvalues */
    int *d_info = NULL; /* error info */
    int lwork = 0;      /* size of workspace */
    double *d_work = NULL; /* device workspace for syevj */
    int info = 0;       /* host copy of error info */

    /* configuration of syevj */
    const double tol = 1.e-7;
    const int max_sweeps = 15;
    const cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR; // compute
    eigenvectors.
    const cublasFillMode_t uplo = CUBLAS_FILL_MODE_LOWER;

```

configure parameters of syevj

```

/* numerical results of syevj */
double residual = 0;
int executed_sweeps = 0;

printf("example of syevj \n");
printf("tol = %E, default value is machine zero \n", tol);
printf("max. sweeps = %d, default value is 100\n", max_sweeps);

printf("A = (matlab base-1)\n");
printMatrix(m, m, A, lda, "A");
printf("=====\n");

/* step 1: create cusolver handle, bind a stream */
status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaStreamCreateWithFlags(&stream, cudaStreamNonBlocking);
assert(cudaSuccess == cudaStat1);

status = cusolverDnSetStream(cusolverH, stream);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 2: configuration of syevj */
status = cusolverDnCreateSyevjInfo(&syevj_params);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* default value of tolerance is machine zero */
status = cusolverDnXsyevjSetTolerance(
    syevj_params,
    tol);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* default value of max. sweeps is 100 */
status = cusolverDnXsyevjSetMaxSweeps(
    syevj_params,
    max_sweeps);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 3: copy A to device */
cudaStat1 = cudaMalloc ((void**)&d_A, sizeof(double) * lda * m);
cudaStat2 = cudaMalloc ((void**)&d_W, sizeof(double) * m);
cudaStat3 = cudaMalloc ((void**)&d_info, sizeof(int));
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);

cudaStat1 = cudaMemcpy(d_A, A, sizeof(double)*lda*m,
    cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);

```

call eigenvalue solver

```

/* step 4: query working space of syevj */
status = cusolverDnDsyejv_bufferSize(
    cusolverH,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_W,
    &lwork,
    syevj_params);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaMalloc((void**) &d_work, sizeof(double)*lwork);
assert(cudaSuccess == cudaStat1);

/* step 5: compute eigen-pair */
status = cusolverDnDsyejv(
    cusolverH,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_W,
    d_work,
    lwork,
    d_info,
    syevj_params);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == status);
assert(cudaSuccess == cudaStat1);

cudaStat1 = cudaMemcpy(W, d_W, sizeof(double)*m, cudaMemcpyDeviceToHost);
cudaStat2 = cudaMemcpy(V, d_A, sizeof(double)*lda*m,
    cudaMemcpyDeviceToHost);
cudaStat3 = cudaMemcpy(&info, d_info, sizeof(int), cudaMemcpyDeviceToHost);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);

if ( 0 == info ){
    printf("syevj converges \n");
}else if ( 0 > info ){
    printf("%d-th parameter is wrong \n", -info);
    exit(1);
}else{
    printf("WARNING: info = %d : syevj does not converge \n", info );
}

printf("Eigenvalue = (matlab base-1), ascending order\n");
for(int i = 0 ; i < m ; i++){
    printf("W[%d] = %E\n", i+1, W[i]);
}

printf("V = (matlab base-1)\n");
printMatrix(m, m, V, lda, "V");
printf("=====\n");

```

check the result

```

/* step 6: check eigenvalues */
double lambda_sup = 0;
for(int i = 0 ; i < m ; i++){
    double error = fabs( lambda[i] - W[i]);
    lambda_sup = (lambda_sup > error)? lambda_sup : error;
}
printf("|lambda - W| = %E\n", lambda_sup);

status = cusolverDnXsyevjGetSweeps(
    cusolverH,
    syevj_params,
    &executed_sweeps);
assert(CUSOLVER_STATUS_SUCCESS == status);

status = cusolverDnXsyevjGetResidual(
    cusolverH,
    syevj_params,
    &residual);
assert(CUSOLVER_STATUS_SUCCESS == status);

printf("residual |A - V*W*V**H|_F = %E \n", residual );
printf("number of executed sweeps = %d \n", executed_sweeps );

/* free resources */
if (d_A ) cudaFree(d_A);
if (d_W ) cudaFree(d_W);
if (d_info ) cudaFree(d_info);
if (d_work ) cudaFree(d_work);

if (cusolverH ) cusolverDnDestroy(cusolverH);
if (stream ) cudaStreamDestroy(stream);
if (syevj_params) cusolverDnDestroySyevjInfo(syevj_params);

cudaDeviceReset();

return 0;
}

```

F.6. Generalized Symmetric-Definite Dense Eigenvalue Solver (via Jacobi method)

This chapter provides a simple example in the C programming language of how to use **sygvj** to compute spectrum of a pair of dense symmetric matrices (A,B) by

$$Ax = \lambda Bx$$

where A is a 3x3 dense symmetric matrix

$$A = \begin{pmatrix} 3.5 & 0.5 & 0 \\ 0.5 & 3.5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

and B is a 3x3 positive definite matrix

$$B = \begin{pmatrix} 10 & 2 & 3 \\ 2 & 10 & 5 \\ 3 & 5 & 10 \end{pmatrix}$$

The following code uses **sygvj** to compute eigenvalues and eigenvectors.

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *   nvcc -c -I/usr/local/cuda/include sygvj_example.cpp
 *   g++ -o sygvj_example sygvj_example.o -L/usr/local/cuda/lib64 -lcusolver -
 *   lcudart
 */
#include <stdio.h>
#include <stdlib.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cudaStream_t stream = NULL;
    syevjInfo_t syevj_params = NULL;
    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    const int m = 3;
    const int lda = m;

    /*
     *      | 3.5 0.5 0 |
     *   A = | 0.5 3.5 0 |
     *      | 0   0  2 |
     *
     *      | 10  2  3 |
     *   B = | 2   10  5 |
     *      | 3   5  10 |
     */
    double A[lda*m] = { 3.5, 0.5, 0, 0.5, 3.5, 0, 0, 0, 2.0};
    double B[lda*m] = { 10.0, 2.0, 3.0, 2.0, 10.0, 5.0, 3.0, 5.0, 10.0};
    double lambda[m] = { 0.158660256604, 0.370751508101882, 0.6};

    double V[lda*m]; /* eigenvectors */
    double W[m];      /* eigenvalues */

    double *d_A = NULL; /* device copy of A */
    double *d_B = NULL; /* device copy of B */
    double *d_W = NULL; /* numerical eigenvalue */
    int *d_info = NULL; /* error info */
    int lwork = 0; /* size of workspace */
    double *d_work = NULL; /* device workspace for sygvj */
    int info = 0; /* host copy of error info */

```

configure parameters of Jacobi method

```

/* configuration of sygvj */
const double tol = 1.e-7;
const int max_sweeps = 15;
const cusolverEigType_t itype = CUSOLVER_EIG_TYPE_1; // A*x = (lambda)*B*x
const cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR; // compute
eigenvectors.
const cublasFillMode_t uplo = CUBLAS_FILL_MODE_LOWER;

/* numerical results of syevj */
double residual = 0;
int executed_sweeps = 0;

printf("example of sygvj \n");
printf("tol = %E, default value is machine zero \n", tol);
printf("max. sweeps = %d, default value is 100\n", max_sweeps);

printf("A = (matlab base-1)\n");
printMatrix(m, m, A, lda, "A");
printf("====\n");

printf("B = (matlab base-1)\n");
printMatrix(m, m, B, lda, "B");
printf("====\n");

/* step 1: create cusolver handle, bind a stream */
status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaStreamCreateWithFlags(&stream, cudaStreamNonBlocking);
assert(cudaSuccess == cudaStat1);

status = cusolverDnSetStream(cusolverH, stream);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 2: configuration of syevj */
status = cusolverDnCreateSyevjInfo(&syevj_params);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* default value of tolerance is machine zero */
status = cusolverDnXsyevjSetTolerance(
    syevj_params,
    tol);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* default value of max. sweeps is 100 */
status = cusolverDnXsyevjSetMaxSweeps(
    syevj_params,
    max_sweeps);
assert(CUSOLVER_STATUS_SUCCESS == status);

```

call eigenvalue solver

```

/* step 3: copy A and B to device */
    cudaStat1 = cudaMalloc ((void**) &d_A, sizeof(double) * lda * m);
    cudaStat2 = cudaMalloc ((void**) &d_B, sizeof(double) * lda * m);
    cudaStat3 = cudaMalloc ((void**) &d_W, sizeof(double) * m);
    cudaStat4 = cudaMalloc ((void**) &d_info, sizeof(int));
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    assert(cudaSuccess == cudaStat3);
    assert(cudaSuccess == cudaStat4);

    cudaStat1 = cudaMemcpy(d_A, A, sizeof(double) * lda * m,
        cudaMemcpyHostToDevice);
    cudaStat2 = cudaMemcpy(d_B, B, sizeof(double) * lda * m,
        cudaMemcpyHostToDevice);
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);

/* step 4: query working space of sygvj */
    status = cusolverDnDsygvj_bufferSize(
        cusolverH,
        itype,
        jobz,
        uplo,
        m,
        d_A,
        lda,
        d_B,
        lda, /* ldb */
        d_W,
        &lwork,
        syevj_params);
    assert(CUSOLVER_STATUS_SUCCESS == status);

    cudaStat1 = cudaMalloc((void**) &d_work, sizeof(double)*lwork);
    assert(cudaSuccess == cudaStat1);

/* step 5: compute spectrum of (A,B) */
    status = cusolverDnDsygvj(
        cusolverH,
        itype,
        jobz,
        uplo,
        m,
        d_A,
        lda,
        d_B,
        lda, /* ldb */
        d_W,
        d_work,
        lwork,
        d_info,
        syevj_params);
    cudaStat1 = cudaDeviceSynchronize();
    assert(CUSOLVER_STATUS_SUCCESS == status);
    assert(cudaSuccess == cudaStat1);

    cudaStat1 = cudaMemcpy(W, d_W, sizeof(double)*m, cudaMemcpyDeviceToHost);
    cudaStat2 = cudaMemcpy(V, d_A, sizeof(double)*lda*m,
        cudaMemcpyDeviceToHost);
    cudaStat3 = cudaMemcpy(&info, d_info, sizeof(int), cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    assert(cudaSuccess == cudaStat3);

```

check the result

```

    if ( 0 == info ){
        printf("sygvj converges \n");
    }else if ( 0 > info ){
        printf("Error: %d-th parameter is wrong \n", -info);
        exit(1);
    }else if ( m >= info ){
        printf("Error: leading minor of order %d of B is not positive definite \n", -info);
        exit(1);
    }else { /* info = m+1 */
        printf("WARNING: info = %d : sygvj does not converge \n", info );
    }

    printf("Eigenvalue = (matlab base-1), ascending order\n");
    for(int i = 0 ; i < m ; i++){
        printf("W[%d] = %E\n", i+1, W[i]);
    }

    printf("V = (matlab base-1)\n");
    printMatrix(m, m, V, lda, "V");
    printf("=====\n");

/* step 6: check eigenvalues */
    double lambda_sup = 0;
    for(int i = 0 ; i < m ; i++){
        double error = fabs( lambda[i] - W[i]);
        lambda_sup = (lambda_sup > error)? lambda_sup : error;
    }
    printf("|lambda - W| = %E\n", lambda_sup);

    status = cusolverDnXsyevjGetSweeps(
        cusolverH,
        syevj_params,
        &executed_sweeps);
    assert(CUSOLVER_STATUS_SUCCESS == status);

    status = cusolverDnXsyevjGetResidual(
        cusolverH,
        syevj_params,
        &residual);
    assert(CUSOLVER_STATUS_SUCCESS == status);

    printf("residual |M - V*W*V**H|_F = %E \n", residual );
    printf("number of executed sweeps = %d \n", executed_sweeps );

/* free resources */
    if (d_A ) cudaFree(d_A);
    if (d_B ) cudaFree(d_B);
    if (d_W ) cudaFree(d_W);
    if (d_info ) cudaFree(d_info);
    if (d_work ) cudaFree(d_work);
    if (cusolverH) cusolverDnDestroy(cusolverH);
    if (stream ) cudaStreamDestroy(stream);
    if (syevj_params) cusolverDnDestroySyevjInfo(syevj_params);

    cudaDeviceReset();
    return 0;
}

```

F.7. batch eigenvalue solver for dense symmetric matrix

This chapter provides a simple example in the C programming language of how to use `syevjBatched` to compute the spectrum of a sequence of dense symmetric matrices by

$$A_j x = \lambda x$$

where A_0 and A_1 are 3x3 dense symmetric matrices

$$A_0 = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$A_1 = \begin{pmatrix} 3 & 4 & 0 \\ 4 & 7 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

The following code uses **`syevjBatched`** to compute eigenvalues and eigenvectors

$$A_j = V_j^* W_j V_j^T$$

The user can disable/enable sorting by the function `cusolverDnXsyevjSetSortEig`.

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *   nvcc -c -I/usr/local/cuda/include batchsyevj_example.cpp
 *   g++ -o batchsyevj_example batchsyevj_example.o -L/usr/local/cuda/lib64 -
lcusolver -lcudart
 */
#include <stdio.h>
#include <stdlib.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cudaStream_t stream = NULL;
    syevjInfo_t syevj_params = NULL;

    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    const int m = 3; // 1<= m <= 32
    const int lda = m;
    const int batchSize = 2;

    /*
     *
     *   A0 = |  1  -1  0  |
     *         | -1   2  0  |
     *         |  0   0  0  |
     *
     *   A0 = V0 * W0 * V0**T
     *
     *   W0 = diag(0, 0.3820, 2.6180)
     *
     *   A1 = |  3   4  0  |
     *         |  4   7  0  |
     *         |  0   0  0  |
     *
     *   A1 = V1 * W1 * V1**T
     *
     *   W1 = diag(0, 0.5279, 9.4721)
     *
     */

```

setup matrices A0 and A1

```
double A[lda*m*batchSize]; /* A = [A0 ; A1] */
double V[lda*m*batchSize]; /* V = [V0 ; V1] */
double W[m*batchSize];     /* W = [W0 ; W1] */
int info[batchSize];       /* info = [info0 ; info1] */

double *d_A = NULL; /* lda-by-m-by-batchSize */
double *d_W = NULL; /* m-by-batchSize */
int* d_info = NULL; /* batchSize */
int lwork = 0; /* size of workspace */
double *d_work = NULL; /* device workspace for syevjBatched */

const double tol = 1.e-7;
const int max_sweeps = 15;
const int sort_eig = 0; /* don't sort eigenvalues */
const cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR; /* compute
eigenvectors */
const cublasFillMode_t uplo = CUBLAS_FILL_MODE_LOWER;

/* residual and executed_sweeps are not supported on syevjBatched */
double residual = 0;
int executed_sweeps = 0;

double *A0 = A;
double *A1 = A + lda*m;

/*
 *      | 1 -1 0 |
 *  A0 = | -1 2 0 |
 *      | 0 0 0 |
 *  A0 is column-major
 */
A0[0 + 0*lda] = 1.0;
A0[1 + 0*lda] = -1.0;
A0[2 + 0*lda] = 0.0;

A0[0 + 1*lda] = -1.0;
A0[1 + 1*lda] = 2.0;
A0[2 + 1*lda] = 0.0;

A0[0 + 2*lda] = 0.0;
A0[1 + 2*lda] = 0.0;
A0[2 + 2*lda] = 0.0;

/*
 *      | 3 4 0 |
 *  A1 = | 4 7 0 |
 *      | 0 0 0 |
 *  A1 is column-major
 */
A1[0 + 0*lda] = 3.0;
A1[1 + 0*lda] = 4.0;
A1[2 + 0*lda] = 0.0;

A1[0 + 1*lda] = 4.0;
A1[1 + 1*lda] = 7.0;
A1[2 + 1*lda] = 0.0;

A1[0 + 2*lda] = 0.0;
A1[1 + 2*lda] = 0.0;
A1[2 + 2*lda] = 0.0;
```

configure parameters of syevj

```

/* step 1: create cusolver handle, bind a stream */
status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaStreamCreateWithFlags(&stream, cudaStreamNonBlocking);
assert(cudaSuccess == cudaStat1);

status = cusolverDnSetStream(cusolverH, stream);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 2: configuration of syevj */
status = cusolverDnCreateSyevjInfo(&syevj_params);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* default value of tolerance is machine zero */
status = cusolverDnXsyevjSetTolerance(
    syevj_params,
    tol);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* default value of max. sweeps is 100 */
status = cusolverDnXsyevjSetMaxSweeps(
    syevj_params,
    max_sweeps);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* disable sorting */
status = cusolverDnXsyevjSetSortEig(
    syevj_params,
    sort_eig);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 3: copy A to device */
cudaStat1 = cudaMalloc ((void**)&d_A, sizeof(double) * lda * m *
batchSize);
cudaStat2 = cudaMalloc ((void**)&d_W, sizeof(double) * m * batchSize);
cudaStat3 = cudaMalloc ((void**)&d_info, sizeof(int) * batchSize);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);

cudaStat1 = cudaMemcpy(d_A, A, sizeof(double) * lda * m * batchSize,
cudaMemcpyHostToDevice);
cudaStat2 = cudaDeviceSynchronize();
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);

/* step 4: query working space of syevjBatched */
status = cusolverDnDsyevjBatched_bufferSize(
    cusolverH,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_W,
    &lwork,
    syevj_params,
    batchSize
);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaMalloc((void**)&d_work, sizeof(double)*lwork);
assert(cudaSuccess == cudaStat1);

```


call eigenvalue solver

```

/* step 5: compute spectrum of A0 and A1 */
status = cusolverDnDsyejvBatched(
    cusolverH,
    jobz,
    uplo,
    m,
    d_A,
    lda,
    d_W,
    d_work,
    lwork,
    d_info,
    syevj_params,
    batchSize
);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == status);
assert(cudaSuccess == cudaStat1);

    cudaStat1 = cudaMemcpy(V      , d_A      , sizeof(double) * lda * m * batchSize,
cudaMemcpyDeviceToHost);
    cudaStat2 = cudaMemcpy(W      , d_W      , sizeof(double) * m * batchSize      ,
cudaMemcpyDeviceToHost);
    cudaStat3 = cudaMemcpy(&info, d_info, sizeof(int) * batchSize      ,
cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    assert(cudaSuccess == cudaStat3);

    for(int i = 0 ; i < batchSize ; i++){
        if ( 0 == info[i] ){
            printf("matrix %d: syevj converges \n", i);
        }else if ( 0 > info[i] ){
/* only info[0] shows if some input parameter is wrong.
 * If so, the error is CUSOLVER_STATUS_INVALID_VALUE.
 */
            printf("Error: %d-th parameter is wrong \n", -info[i] );
            exit(1);
        }else { /* info = m+1 */
/* if info[i] is not zero, Jacobi method does not converge at i-th matrix. */
            printf("WARNING: matrix %d, info = %d : sygvj does not converge \n",
i, info[i] );
        }
    }

/* Step 6: show eigenvalues and eigenvectors */
double *W0 = W;
double *W1 = W + m;
printf("==== \n");
for(int i = 0 ; i < m ; i++){
    printf("W0[%d] = %f\n", i, W0[i]);
}
printf("==== \n");
for(int i = 0 ; i < m ; i++){
    printf("W1[%d] = %f\n", i, W1[i]);
}
printf("==== \n");

double *V0 = V;
double *V1 = V + lda*m;
printf("V0 = (matlab base-1)\n");
printMatrix(m, m, V0, lda, "V0");
printf("V1 = (matlab base-1)\n");
printMatrix(m, m, V1, lda, "V1");

```

cannot query residual and executed sweeps.

```

/*
 * The following two functions do not support batched version.
 * The error CUSOLVER_STATUS_NOT_SUPPORTED is returned.
 */
    status = cusolverDnXsyevjGetSweeps(
        cusolverH,
        syevj_params,
        &executed_sweeps);
    assert(CUSOLVER_STATUS_NOT_SUPPORTED == status);

    status = cusolverDnXsyevjGetResidual(
        cusolverH,
        syevj_params,
        &residual);
    assert(CUSOLVER_STATUS_NOT_SUPPORTED == status);

/* free resources */
    if (d_A ) cudaFree(d_A);
    if (d_W ) cudaFree(d_W);
    if (d_info ) cudaFree(d_info);
    if (d_work ) cudaFree(d_work);

    if (cusolverH) cusolverDnDestroy(cusolverH);
    if (stream ) cudaStreamDestroy(stream);
    if (syevj_params) cusolverDnDestroySyevjInfo(syevj_params);

    cudaDeviceReset();

    return 0;
}

```

Appendix G.

EXAMPLES OF SINGULAR VALUE DECOMPOSITION

G.1. SVD with singular vectors

This chapter provides a simple example in the C programming language of how to perform singular value decomposition.

$$A = U * \Sigma * V^H$$

A is a 3x2 dense matrix,

$$A = \begin{pmatrix} 1.0 & 2.0 \\ 4.0 & 5.0 \\ 2.0 & 1.0 \end{pmatrix}$$

The following code uses three steps:

Step 1: compute $A = U * S * V^T$

Step 2: check accuracy of singular value

Step 3: measure residual $A - U * S * V^T$

```

...

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 * nvcc -c -I/usr/local/cuda/include svd_example.cpp
 * g++ -o a.out svd_example.o -I/usr/local/cuda/lib64 -lcudart -lcublas -
lcusolver
 *
 */

#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cublas_v2.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %f\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cublasHandle_t cublasH = NULL;
    cublasStatus_t cublas_status = CUBLAS_STATUS_SUCCESS;
    cusolverStatus_t cusolver_status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    cudaError_t cudaStat5 = cudaSuccess;
    cudaError_t cudaStat6 = cudaSuccess;
    const int m = 3;
    const int n = 2;
    const int lda = m;
    /*
     *   | 1 2 |
     *   A = | 4 5 |
     *       | 2 1 |
     */
    double A[lda*n] = { 1.0, 4.0, 2.0, 2.0, 5.0, 1.0};
    double U[lda*m]; // m-by-m unitary matrix
    double VT[lda*n]; // n-by-n unitary matrix
    double S[n]; // singular value
    double S_exact[n] = {7.065283497082729, 1.040081297712078};

    double *d_A = NULL;
    double *d_S = NULL;
    double *d_U = NULL;
    double *d_VT = NULL;
    int *devInfo = NULL;
    double *d_work = NULL;
    double *d_rwork = NULL;
    double *d_W = NULL; // W = S*VT

    int lwork = 0;
    int info_gpu = 0;
    const double h_one = 1;
    const double h_minus_one = -1;

```

...

```

printf("A = (matlab base-1)\n");
printMatrix(m, n, A, lda, "A");
printf("=====\n");

// step 1: create cusolverDn/cublas handle
cusolver_status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);

cublas_status = cublasCreate(&cublasH);
assert(CUBLAS_STATUS_SUCCESS == cublas_status);

// step 2: copy A and B to device
cudaStat1 = cudaMalloc ((void**)&d_A , sizeof(double)*lda*n);
cudaStat2 = cudaMalloc ((void**)&d_S , sizeof(double)*n);
cudaStat3 = cudaMalloc ((void**)&d_U , sizeof(double)*lda*m);
cudaStat4 = cudaMalloc ((void**)&d_VT , sizeof(double)*lda*n);
cudaStat5 = cudaMalloc ((void**)&devInfo, sizeof(int));
cudaStat6 = cudaMalloc ((void**)&d_W , sizeof(double)*lda*n);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);
assert(cudaSuccess == cudaStat4);
assert(cudaSuccess == cudaStat5);
assert(cudaSuccess == cudaStat6);

cudaStat1 = cudaMemcpy(d_A, A, sizeof(double)*lda*n,
cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);

// step 3: query working space of SVD
cusolver_status = cusolverDnDgesvd_bufferSize(
    cusolverH,
    m,
    n,
    &lwork );
assert (cusolver_status == CUSOLVER_STATUS_SUCCESS);

cudaStat1 = cudaMalloc((void**)&d_work , sizeof(double)*lwork);
assert(cudaSuccess == cudaStat1);

// step 4: compute SVD
signed char jobu = 'A'; // all m columns of U
signed char jobvt = 'A'; // all n columns of VT
cusolver_status = cusolverDnDgesvd (
    cusolverH,
    jobu,
    jobvt,
    m,
    n,
    d_A,
    lda,
    d_S,
    d_U,
    lda, // ldu
    d_VT,
    lda, // ldvt,
    d_work,
    lwork,
    d_rwork,
    devInfo);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == cusolver_status);
assert(cudaSuccess == cudaStat1);

```

...

```

    cudaStat1 = cudaMemcpy(U , d_U , sizeof(double)*lda*m,
        cudaMemcpyDeviceToHost);
    cudaStat2 = cudaMemcpy(VT, d_VT, sizeof(double)*lda*n,
        cudaMemcpyDeviceToHost);
    cudaStat3 = cudaMemcpy(S , d_S , sizeof(double)*n ,
        cudaMemcpyDeviceToHost);
    cudaStat4 = cudaMemcpy(&info_gpu, devInfo, sizeof(int),
        cudaMemcpyDeviceToHost);
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    assert(cudaSuccess == cudaStat3);
    assert(cudaSuccess == cudaStat4);

    printf("after gesvd: info_gpu = %d\n", info_gpu);
    assert(0 == info_gpu);
    printf("=====\n");

    printf("S = (matlab base-1)\n");
    printMatrix(n, 1, S, lda, "S");
    printf("=====\n");

    printf("U = (matlab base-1)\n");
    printMatrix(m, m, U, lda, "U");
    printf("=====\n");

    printf("VT = (matlab base-1)\n");
    printMatrix(n, n, VT, lda, "VT");
    printf("=====\n");

// step 5: measure error of singular value
double ds_sup = 0;
for(int j = 0; j < n; j++){
    double err = fabs( S[j] - S_exact[j] );
    ds_sup = (ds_sup > err)? ds_sup : err;
}
printf("|S - S_exact| = %E \n", ds_sup);

// step 6: |A - U*S*VT|
// W = S*VT
cublas_status = cublasDdggmm(
    cublasH,
    CUBLAS_SIDE_LEFT,
    n,
    n,
    d_VT,
    lda,
    d_S,
    1,
    d_W,
    lda);
assert(CUBLAS_STATUS_SUCCESS == cublas_status);

```

...

```

// A := -U*W + A
cudaStat1 = cudaMemcpy(d_A, A, sizeof(double)*lda*n,
cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);
cublas_status = cublasDgemm_v2(
    cublasH,
    CUBLAS_OP_N, // U
    CUBLAS_OP_N, // W
    m, // number of rows of A
    n, // number of columns of A
    n, // number of columns of U
    &h_minus_one, /* host pointer */
    d_U, // U
    lda,
    d_W, // W
    lda,
    &h_one, /* hostpointer */
    d_A,
    lda);
assert(CUBLAS_STATUS_SUCCESS == cublas_status);

double dR_fro = 0.0;
cublas_status = cublasDnrm2_v2(
    cublasH, lda*n, d_A, 1, &dR_fro);
assert(CUBLAS_STATUS_SUCCESS == cublas_status);

printf("|A - U*S*VT| = %E \n", dR_fro);

// free resources
if (d_A) cudaFree(d_A);
if (d_S) cudaFree(d_S);
if (d_U) cudaFree(d_U);
if (d_VT) cudaFree(d_VT);
if (devInfo) cudaFree(devInfo);
if (d_work) cudaFree(d_work);
if (d_rwork) cudaFree(d_rwork);
if (d_W) cudaFree(d_W);

if (cublasH) cublasDestroy(cublasH);
if (cusolverH) cusolverDnDestroy(cusolverH);

cudaDeviceReset();

return 0;
}

```

G.2. SVD with singular vectors (via Jacobi method)

This chapter provides a simple example in the C programming language of how to perform singular value decomposition by **gesvdj**.

$$A = U * \Sigma * V^H$$

A is a 3x2 dense matrix,

$$A = \begin{pmatrix} 1.0 & 2.0 \\ 4.0 & 5.0 \\ 2.0 & 1.0 \end{pmatrix}$$

...

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 * nvcc -c -I/usr/local/cuda/include gesvdj_example.cpp
 * g++ -o gesvdj_example gesvdj_example.o -I/usr/local/cuda/lib64 -lcudart -lcusolver
 */
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %20.16E\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cudaStream_t stream = NULL;
    gesvdjInfo_t gesvdj_params = NULL;

    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    cudaError_t cudaStat5 = cudaSuccess;
    const int m = 3;
    const int n = 2;
    const int lda = m;
/*
 *   | 1 2 |
 *   A = | 4 5 |
 *       | 2 1 |
 */
    double A[lda*n] = { 1.0, 4.0, 2.0, 2.0, 5.0, 1.0};
    double U[lda*m]; /* m-by-m unitary matrix, left singular vectors */
    double V[lda*n]; /* n-by-n unitary matrix, right singular vectors */
    double S[n];      /* numerical singular value */
/* exact singular values */
    double S_exact[n] = {7.065283497082729, 1.040081297712078};
    double *d_A = NULL; /* device copy of A */
    double *d_S = NULL; /* singular values */
    double *d_U = NULL; /* left singular vectors */
    double *d_V = NULL; /* right singular vectors */
    int *d_info = NULL; /* error info */
    int lwork = 0; /* size of workspace */
    double *d_work = NULL; /* device workspace for gesvdj */
    int info = 0; /* host copy of error info */

```


...

```

/* configuration of gesvdj */
const double tol = 1.e-7;
const int max_sweeps = 15;
const cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR; // compute
eigenvectors.
const int econ = 0 ; /* econ = 1 for economy size */

/* numerical results of gesvdj */
double residual = 0;
int executed_sweeps = 0;

printf("example of gesvdj \n");
printf("tol = %E, default value is machine zero \n", tol);
printf("max. sweeps = %d, default value is 100\n", max_sweeps);
printf("econ = %d \n", econ);

printf("A = (matlab base-1)\n");
printMatrix(m, n, A, lda, "A");
printf("=====\n");

/* step 1: create cusolver handle, bind a stream */
status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaStreamCreateWithFlags(&stream, cudaStreamNonBlocking);
assert(cudaSuccess == cudaStat1);

status = cusolverDnSetStream(cusolverH, stream);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 2: configuration of gesvdj */
status = cusolverDnCreateGesvdjInfo(&gesvdj_params);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* default value of tolerance is machine zero */
status = cusolverDnXgesvdjSetTolerance(
    gesvdj_params,
    tol);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* default value of max. sweeps is 100 */
status = cusolverDnXgesvdjSetMaxSweeps(
    gesvdj_params,
    max_sweeps);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 3: copy A and B to device */
cudaStat1 = cudaMalloc ((void**) &d_A , sizeof(double)*lda*n);
cudaStat2 = cudaMalloc ((void**) &d_S , sizeof(double)*n);
cudaStat3 = cudaMalloc ((void**) &d_U , sizeof(double)*lda*m);
cudaStat4 = cudaMalloc ((void**) &d_V , sizeof(double)*lda*n);
cudaStat5 = cudaMalloc ((void**) &d_info, sizeof(int));
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);
assert(cudaSuccess == cudaStat4);
assert(cudaSuccess == cudaStat5);

cudaStat1 = cudaMemcpy(d_A, A, sizeof(double)*lda*n,
    cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);

```

...

```

/* step 4: query workspace of SVD */
status = cusolverDnDgesvdj_bufferSize(
    cusolverH,
    jobz, /* CUSOLVER_EIG_MODE_NOVECTOR: compute singular values only */
          /* CUSOLVER_EIG_MODE_VECTOR: compute singular value and singular
vectors */
    econ, /* econ = 1 for economy size */
    m,    /* nubmer of rows of A, 0 <= m */
    n,    /* number of columns of A, 0 <= n */
    d_A,  /* m-by-n */
    lda,  /* leading dimension of A */
    d_S,  /* min(m,n) */
          /* the singular values in descending order */
    d_U,  /* m-by-m if econ = 0 */
          /* m-by-min(m,n) if econ = 1 */
    lda,  /* leading dimension of U, ldu >= max(1,m) */
    d_V,  /* n-by-n if econ = 0 */
          /* n-by-min(m,n) if econ = 1 */
    lda,  /* leading dimension of V, ldv >= max(1,n) */
    &lwork,
    gesvdj_params);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaMalloc((void**)&d_work , sizeof(double)*lwork);
assert(cudaSuccess == cudaStat1);

/* step 5: compute SVD */
status = cusolverDnDgesvdj(
    cusolverH,
    jobz, /* CUSOLVER_EIG_MODE_NOVECTOR: compute singular values only */
          /* CUSOLVER_EIG_MODE_VECTOR: compute singular value and singular
vectors */
    econ, /* econ = 1 for economy size */
    m,    /* nubmer of rows of A, 0 <= m */
    n,    /* number of columns of A, 0 <= n */
    d_A,  /* m-by-n */
    lda,  /* leading dimension of A */
    d_S,  /* min(m,n) */
          /* the singular values in descending order */
    d_U,  /* m-by-m if econ = 0 */
          /* m-by-min(m,n) if econ = 1 */
    lda,  /* leading dimension of U, ldu >= max(1,m) */
    d_V,  /* n-by-n if econ = 0 */
          /* n-by-min(m,n) if econ = 1 */
    lda,  /* leading dimension of V, ldv >= max(1,n) */
    d_work,
    lwork,
    d_info,
    gesvdj_params);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == status);
assert(cudaSuccess == cudaStat1);

    cudaStat1 = cudaMemcpy(U, d_U, sizeof(double)*lda*m,
        cudaMemcpyDeviceToHost);
    cudaStat2 = cudaMemcpy(V, d_V, sizeof(double)*lda*n,
        cudaMemcpyDeviceToHost);
    cudaStat3 = cudaMemcpy(S, d_S, sizeof(double)*n ,
        cudaMemcpyDeviceToHost);
    cudaStat4 = cudaMemcpy(&info, d_info, sizeof(int), cudaMemcpyDeviceToHost);
    cudaStat5 = cudaDeviceSynchronize();
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    assert(cudaSuccess == cudaStat3);
    assert(cudaSuccess == cudaStat4);
    assert(cudaSuccess == cudaStat5);

```

...

```

    if ( 0 == info ){
        printf("gesvdj converges \n");
    }else if ( 0 > info ){
        printf("%d-th parameter is wrong \n", -info);
        exit(1);
    }else{
        printf("WARNING: info = %d : gesvdj does not converge \n", info );
    }

    printf("S = singular values (matlab base-1)\n");
    printMatrix(n, 1, S, lda, "S");
    printf("=====\n");

    printf("U = left singular vectors (matlab base-1)\n");
    printMatrix(m, m, U, lda, "U");
    printf("=====\n");

    printf("V = right singular vectors (matlab base-1)\n");
    printMatrix(n, n, V, lda, "V");
    printf("=====\n");

/* step 6: measure error of singular value */
double ds_sup = 0;
for(int j = 0; j < n; j++){
    double err = fabs( S[j] - S_exact[j] );
    ds_sup = (ds_sup > err)? ds_sup : err;
}
printf("|S - S_exact|_sup = %E \n", ds_sup);

status = cusolverDnXgesvdjGetSweeps(
    cusolverH,
    gesvdj_params,
    &executed_sweeps);
assert(CUSOLVER_STATUS_SUCCESS == status);

status = cusolverDnXgesvdjGetResidual(
    cusolverH,
    gesvdj_params,
    &residual);
assert(CUSOLVER_STATUS_SUCCESS == status);

printf("residual |A - U*S*V**H|_F = %E \n", residual );
printf("number of executed sweeps = %d \n", executed_sweeps );

/* free resources */
if (d_A ) cudaFree(d_A);
if (d_S ) cudaFree(d_S);
if (d_U ) cudaFree(d_U);
if (d_V ) cudaFree(d_V);
if (d_info) cudaFree(d_info);
if (d_work ) cudaFree(d_work);

if (cusolverH) cusolverDnDestroy(cusolverH);
if (stream ) cudaStreamDestroy(stream);
if (gesvdj_params) cusolverDnDestroyGesvdjInfo(gesvdj_params);

cudaDeviceReset();
return 0;
}

```

G.3. batch dense SVD solver

This chapter provides a simple example in the C programming language of how to use **gesvdjBatched** to compute the SVD of a sequence of dense matrices

$$A_j = U_j \Sigma_j V_j^H$$

where A0 and A1 are 3x2 dense matrices

$$A0 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \\ 0 & 0 \end{pmatrix}$$

$$A1 = \begin{pmatrix} 3 & 4 \\ 4 & 7 \\ 0 & 0 \end{pmatrix}$$

The following code uses **gesvdjBatched** to compute singular values and singular vectors.

The user can disable/enable sorting by the function `cusolverDnXgesvdjSetSortEig`.

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *   nvcc -c -I/usr/local/cuda/include gesvdjbatch_example.cpp
 *   g++ -o gesvdjbatch_example gesvdjbatch_example.o -L/usr/local/cuda/lib64 -
lcusolver -lcudart
 */
#include <stdio.h>
#include <stdlib.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const double*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            double Areg = A[row + col*lda];
            printf("%s(%d,%d) = %20.16E\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cudaStream_t stream = NULL;
    gesvdjInfo_t gesvdj_params = NULL;

    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    cudaError_t cudaStat5 = cudaSuccess;
    const int m = 3; /* 1 <= m <= 32 */
    const int n = 2; /* 1 <= n <= 32 */
    const int lda = m; /* lda >= m */
    const int ldu = m; /* ldu >= m */
    const int ldv = n; /* ldv >= n */
    const int batchSize = 2;
    const int minmn = (m < n)? m : n; /* min(m,n) */

    /*
     *      |  1  -1  |
     *   A0 = | -1   2  |
     *      |  0   0  |
     *
     *   A0 = U0 * S0 * V0**T
     *   S0 = diag(2.6180, 0.382)
     *
     *      |  3  4  |
     *   A1 = |  4  7  |
     *      |  0  0  |
     *
     *   A1 = U1 * S1 * V1**T
     *   S1 = diag(9.4721, 0.5279)
     */
}

```

setup matrices A0 and A1

```

double A[lda*n*batchSize]; /* A = [A0 ; A1] */
double U[ldu*m*batchSize]; /* U = [U0 ; U1] */
double V[ldv*n*batchSize]; /* V = [V0 ; V1] */
double S[minmn*batchSize]; /* S = [S0 ; S1] */
int info[batchSize]; /* info = [info0 ; info1] */

double *d_A = NULL; /* lda-by-n-by-batchSize */
double *d_U = NULL; /* ldu-by-m-by-batchSize */
double *d_V = NULL; /* ldv-by-n-by-batchSize */
double *d_S = NULL; /* minmn-by-batchSize */
int* d_info = NULL; /* batchSize */
int lwork = 0; /* size of workspace */
double *d_work = NULL; /* device workspace for gesvdjBatched */

const double tol = 1.e-7;
const int max_sweeps = 15;
const int sort_svd = 0; /* don't sort singular values */
const cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR; /* compute singular
vectors */

/* residual and executed_sweeps are not supported on gesvdjBatched */
double residual = 0;
int executed_sweeps = 0;

double *A0 = A;
double *A1 = A + lda*n; /* Aj is m-by-n */
/*
 *
 * A0 = | 1 -1 |
 *      | -1 2 |
 *      | 0 0 |
 * A0 is column-major
 */
A0[0 + 0*lda] = 1.0;
A0[1 + 0*lda] = -1.0;
A0[2 + 0*lda] = 0.0;

A0[0 + 1*lda] = -1.0;
A0[1 + 1*lda] = 2.0;
A0[2 + 1*lda] = 0.0;

/*
 *
 * A1 = | 3 4 |
 *      | 4 7 |
 *      | 0 0 |
 * A1 is column-major
 */
A1[0 + 0*lda] = 3.0;
A1[1 + 0*lda] = 4.0;
A1[2 + 0*lda] = 0.0;

A1[0 + 1*lda] = 4.0;
A1[1 + 1*lda] = 7.0;
A1[2 + 1*lda] = 0.0;

printf("example of gesvdjBatched \n");
printf("m = %d, n = %d \n", m, n);
printf("tol = %E, default value is machine zero \n", tol);
printf("max. sweeps = %d, default value is 100\n", max_sweeps);

printf("A0 = (matlab base-1)\n");
printMatrix(m, n, A0, lda, "A0");
printf("A1 = (matlab base-1)\n");
printMatrix(m, n, A1, lda, "A1");
printf("====\n");

```

configure parameters of gesvdj

```

/* step 1: create cusolver handle, bind a stream */
status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaStreamCreateWithFlags(&stream, cudaStreamNonBlocking);
assert(cudaSuccess == cudaStat1);

status = cusolverDnSetStream(cusolverH, stream);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 2: configuration of gesvdj */
status = cusolverDnCreateGesvdjInfo(&gesvdj_params);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* default value of tolerance is machine zero */
status = cusolverDnXgesvdjSetTolerance(
    gesvdj_params,
    tol);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* default value of max. sweeps is 100 */
status = cusolverDnXgesvdjSetMaxSweeps(
    gesvdj_params,
    max_sweeps);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* disable sorting */
status = cusolverDnXgesvdjSetSortEig(
    gesvdj_params,
    sort_svd);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 3: copy A to device */
cudaStat1 = cudaMalloc ((void**) &d_A , sizeof(double)*lda*n*batchSize);
cudaStat2 = cudaMalloc ((void**) &d_U , sizeof(double)*ldu*m*batchSize);
cudaStat3 = cudaMalloc ((void**) &d_V , sizeof(double)*ldv*n*batchSize);
cudaStat4 = cudaMalloc ((void**) &d_S , sizeof(double)*minmn*batchSize);
cudaStat5 = cudaMalloc ((void**) &d_info, sizeof(int )*batchSize);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);
assert(cudaSuccess == cudaStat4);
assert(cudaSuccess == cudaStat5);

cudaStat1 = cudaMemcpy(d_A, A, sizeof(double)*lda*n*batchSize,
    cudaMemcpyHostToDevice);
cudaStat2 = cudaDeviceSynchronize();
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);

```

call batched singular value solver

```

/* step 4: query working space of gesvdjBatched */
status = cusolverDnDgesvdjBatched_bufferSize(
    cusolverH,
    jobz,
    m,
    n,
    d_A,
    lda,
    d_S,
    d_U,
    ldu,
    d_V,
    ldv,
    &lwork,
    gesvdj_params,
    batchSize
);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat1 = cudaMalloc((void**) &d_work, sizeof(double)*lwork);
assert(cudaSuccess == cudaStat1);

/* step 5: compute singular values of A0 and A1 */
status = cusolverDnDgesvdjBatched(
    cusolverH,
    jobz,
    m,
    n,
    d_A,
    lda,
    d_S,
    d_U,
    ldu,
    d_V,
    ldv,
    d_work,
    lwork,
    d_info,
    gesvdj_params,
    batchSize
);
cudaStat1 = cudaDeviceSynchronize();
assert(CUSOLVER_STATUS_SUCCESS == status);
assert(cudaSuccess == cudaStat1);

cudaStat1 = cudaMemcpy(U, d_U, sizeof(double)*ldu*m*batchSize,
    cudaMemcpyDeviceToHost);
cudaStat2 = cudaMemcpy(V, d_V, sizeof(double)*ldv*n*batchSize,
    cudaMemcpyDeviceToHost);
cudaStat3 = cudaMemcpy(S, d_S, sizeof(double)*minm*n*batchSize,
    cudaMemcpyDeviceToHost);
cudaStat4 = cudaMemcpy(&info, d_info, sizeof(int) * batchSize,
    cudaMemcpyDeviceToHost);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);
assert(cudaSuccess == cudaStat4);

```


check the result

```

    for(int i = 0 ; i < batchSize ; i++){
        if ( 0 == info[i] ){
            printf("matrix %d: gesvdj converges \n", i);
        }else if ( 0 > info[i] ){
/* only info[0] shows if some input parameter is wrong.
 * If so, the error is CUSOLVER_STATUS_INVALID_VALUE.
 */
            printf("Error: %d-th parameter is wrong \n", -info[i] );
            exit(1);
        }else { /* info = m+1 */
/* if info[i] is not zero, Jacobi method does not converge at i-th matrix. */
            printf("WARNING: matrix %d, info = %d : gesvdj does not converge
\n", i, info[i] );
        }
    }

/* Step 6: show singular values and singular vectors */
double *S0 = S;
double *S1 = S + minmn;
printf("==== \n");
for(int i = 0 ; i < minmn ; i++){
    printf("S0(%d) = %20.16E\n", i+1, S0[i]);
}
printf("==== \n");
for(int i = 0 ; i < minmn ; i++){
    printf("S1(%d) = %20.16E\n", i+1, S1[i]);
}
printf("==== \n");

double *U0 = U;
double *U1 = U + ldu*m; /* Uj is m-by-m */
printf("U0 = (matlab base-1)\n");
printMatrix(m, m, U0, ldu, "U0");
printf("U1 = (matlab base-1)\n");
printMatrix(m, m, U1, ldu, "U1");

double *V0 = V;
double *V1 = V + ldv*n; /* Vj is n-by-n */
printf("V0 = (matlab base-1)\n");
printMatrix(n, n, V0, ldv, "V0");
printf("V1 = (matlab base-1)\n");
printMatrix(n, n, V1, ldv, "V1");

```

cannot query residual and executed sweeps

```

/*
 * The following two functions do not support batched version.
 * The error CUSOLVER_STATUS_NOT_SUPPORTED is returned.
 */
    status = cusolverDnXgesvdjGetSweeps(
        cusolverH,
        gesvdj_params,
        &executed_sweeps);
    assert(CUSOLVER_STATUS_NOT_SUPPORTED == status);

    status = cusolverDnXgesvdjGetResidual(
        cusolverH,
        gesvdj_params,
        &residual);
    assert(CUSOLVER_STATUS_NOT_SUPPORTED == status);

/* free resources */
    if (d_A) cudaFree(d_A);
    if (d_U) cudaFree(d_U);
    if (d_V) cudaFree(d_V);
    if (d_S) cudaFree(d_S);
    if (d_info) cudaFree(d_info);
    if (d_work) cudaFree(d_work);

    if (cusolverH) cusolverDnDestroy(cusolverH);
    if (stream) cudaStreamDestroy(stream);
    if (gesvdj_params) cusolverDnDestroyGesvdjInfo(gesvdj_params);

    cudaDeviceReset();

    return 0;
}

```

G.4. SVD approximation

This chapter provides a simple example in the C programming language of how to approximate singular value decomposition by **gesvdaStridedBatched**.

$$A = U * \Sigma * V^H$$

A0 and A1 are a 3x2 dense matrices,

$$A0 = \begin{pmatrix} 1.0 & 2.0 \\ 4.0 & 5.0 \\ 2.0 & 1.0 \end{pmatrix}$$

$$A1 = \begin{pmatrix} 10.0 & 9.0 \\ 8.0 & 7.0 \\ 6.0 & 5.0 \end{pmatrix}$$

```

...

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 * nvcc -c -I/usr/local/cuda/include gesvda_example.cpp
 * g++ -o gesvda_example gesvda_example.o -I/usr/local/cuda/lib64 -lcudart -lcusolver
 */

#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <assert.h>
#include <cuda_runtime.h>
#include <cusolverDn.h>

void printMatrix(int m, int n, const float*A, int lda, const char* name)
{
    for(int row = 0 ; row < m ; row++){
        for(int col = 0 ; col < n ; col++){
            float Areg = A[row + col*lda];
            printf("%s(%d,%d) = %20.16E\n", name, row+1, col+1, Areg);
        }
    }
}

int main(int argc, char*argv[])
{
    cusolverDnHandle_t cusolverH = NULL;
    cudaStream_t stream = NULL;

    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat1 = cudaSuccess;
    cudaError_t cudaStat2 = cudaSuccess;
    cudaError_t cudaStat3 = cudaSuccess;
    cudaError_t cudaStat4 = cudaSuccess;
    cudaError_t cudaStat5 = cudaSuccess;
    const int batchSize = 2;
    const int m = 3;
    const int n = 2;
    const int lda = m;
    const int ldu = m;
    const int ldv = n;
    const int rank = n;
    const long long int strideA = (long long int)lda*n;
    const long long int strideS = n;
    const long long int strideU = (long long int)ldu*n;
    const long long int strideV = (long long int)ldv*n;
    /*
     * A0 = | 1 2 |, A1 = | 10 9 |
     *      | 4 5 |      | 8 7 |
     *      | 2 1 |      | 6 5 |
     */
    float A[strideA*batchSize] = { 1.0, 4.0, 2.0, 2.0, 5.0, 1.0, 10.0, 8.0, 6.0,
    9.0, 7.0, 5.0};
    float U[strideU*batchSize]; /* left singular vectors */
    float V[strideV*batchSize]; /* right singular vectors */
    float S[strideS*batchSize]; /* numerical singular value */

    /* exact singular values */
    float S_exact[strideS*batchSize] = {7.065283497082729, 1.040081297712078,
    18.839649186929730, 0.260035600289472};

```

...

```

float *d_A = NULL; /* device copy of A */
float *d_S = NULL; /* singular values */
float *d_U = NULL; /* left singular vectors */
float *d_V = NULL; /* right singular vectors */
int *d_info = NULL; /* error info */
int lwork = 0; /* size of workspace */
float *d_work = NULL; /* device workspace for gesvda */
const cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR; // compute
eigenvectors.
double RnormF[batchSize]; /* residual norm */
int info[batchSize]; /* host copy of error info */

printf("example of gesvdaStridedBatched \n");
printf("A = (matlab base-1)\n");
printMatrix(m, n, A, lda, "A0");
printf("=====\n");
printMatrix(m, n, A + strideA, lda, "A1");
printf("=====\n");

/* step 1: create cusolver handle, bind a stream */
status = cusolverDnCreate(&cusolverH);
assert(CUSOLVER_STATUS_SUCCESS == status);
cudaStat1 = cudaStreamCreateWithFlags(&stream, cudaStreamNonBlocking);
assert(cudaSuccess == cudaStat1);
status = cusolverDnSetStream(cusolverH, stream);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* step 2: copy A to device */
cudaStat1 = cudaMalloc ((void**) &d_A, sizeof(float)*strideA*batchSize);
cudaStat2 = cudaMalloc ((void**) &d_S, sizeof(float)*strideS*batchSize);
cudaStat3 = cudaMalloc ((void**) &d_U, sizeof(float)*strideU*batchSize);
cudaStat4 = cudaMalloc ((void**) &d_V, sizeof(float)*strideV*batchSize);
cudaStat5 = cudaMalloc ((void**) &d_info, sizeof(int)*batchSize);
assert(cudaSuccess == cudaStat1);
assert(cudaSuccess == cudaStat2);
assert(cudaSuccess == cudaStat3);
assert(cudaSuccess == cudaStat4);
assert(cudaSuccess == cudaStat5);

cudaStat1 = cudaMemcpy(d_A, A, sizeof(float)*strideA*batchSize,
cudaMemcpyHostToDevice);
assert(cudaSuccess == cudaStat1);
cudaDeviceSynchronize(); /* sync with null stream */

/* step 3: query workspace of SVD */
status = cusolverDnSgesvdaStridedBatched_bufferSize(
cusolverH,
jobz, /* CUSOLVER_EIG_MODE_NOVECTOR: compute singular values only */
/* CUSOLVER_EIG_MODE_VECTOR: compute singular value and singular
vectors */
rank, /* number of singular values */
m, /* number of rows of Aj, 0 <= m */
n, /* number of columns of Aj, 0 <= n */
d_A, /* Aj is m-by-n */
lda, /* leading dimension of Aj */
strideA, /* >= lda*n */
d_S, /* Sj is rank-by-1, singular values in descending order */
strideS, /* >= rank */
d_U, /* Uj is m-by-rank */
ldu, /* leading dimension of Uj, ldu >= max(1,m) */
strideU, /* >= ldu*rank */
d_V, /* Vj is n-by-rank */
ldv, /* leading dimension of Vj, ldv >= max(1,n) */
strideV, /* >= ldv*rank */
&lwork,
batchSize /* number of matrices */
);
assert(CUSOLVER_STATUS_SUCCESS == status);

```

...

```

    cudaStat1 = cudaMalloc((void**)&d_work , sizeof(float)*lwork);
    assert(cudaSuccess == cudaStat1);

/* step 4: compute SVD */
    status = cusolverDnSgesvdaStridedBatched(
        cusolverH,
        jobz, /* CUSOLVER_EIG_MODE_NOVECTOR: compute singular values only */
              /* CUSOLVER_EIG_MODE_VECTOR: compute singular value and singular
vectors */
        rank, /* number of singular values */
        m,    /* nubmer of rows of Aj, 0 <= m */
        n,    /* number of columns of Aj, 0 <= n */
        d_A,  /* Aj is m-by-n */
        lda,  /* leading dimension of Aj */
        strideA, /* >= lda*n */
        d_S,  /* Sj is rank-by-1 */
              /* the singular values in descending order */
        strideS, /* >= rank */
        d_U,  /* Uj is m-by-rank */
        ldu,  /* leading dimension of Uj, ldu >= max(1,m) */
        strideU, /* >= ldu*rank */
        d_V,  /* Vj is n-by-rank */
        ldv,  /* leading dimension of Vj, ldv >= max(1,n) */
        strideV, /* >= ldv*rank */
        d_work,
        lwork,
        d_info,
        RnrmF,
        batchSize /* number of matrices */
    );
    cudaStat1 = cudaDeviceSynchronize();
    assert(CUSOLVER_STATUS_SUCCESS == status);
    assert(cudaSuccess == cudaStat1);

    cudaStat1 = cudaMemcpy(U, d_U, sizeof(float)*strideU*batchSize,
        cudaMemcpyDeviceToHost);
    cudaStat2 = cudaMemcpy(V, d_V, sizeof(float)*strideV*batchSize,
        cudaMemcpyDeviceToHost);
    cudaStat3 = cudaMemcpy(S, d_S, sizeof(float)*strideS*batchSize,
        cudaMemcpyDeviceToHost);
    cudaStat4 = cudaMemcpy(info, d_info, sizeof(int)*batchSize,
        cudaMemcpyDeviceToHost);
    cudaStat5 = cudaDeviceSynchronize();
    assert(cudaSuccess == cudaStat1);
    assert(cudaSuccess == cudaStat2);
    assert(cudaSuccess == cudaStat3);
    assert(cudaSuccess == cudaStat4);
    assert(cudaSuccess == cudaStat5);

    if ( 0 > info[0] ){
        printf("%d-th parameter is wrong \n", -info[0]);
        exit(1);
    }
    for(int idx = 0 ; idx < batchSize; idx++){
        if ( 0 == info[idx] ){
            printf("%d-th matrix, gesvda converges \n", idx );
        }else{
            printf("WARNING: info[%d] = %d : gesvda does not converge \n", idx,
info[idx] );
        }
    }

    printf("S = singular values (matlab base-1)\n");
    printf("U = left singular vectors (matlab base-1)\n");
    printf("V = right singular vectors (matlab base-1)\n\n");

```

...

```

printMatrix(rank, 1, S, n, "S0");
printf("=====\n");

printMatrix(m, rank, U, ldu, "U0");
printf("=====\n");

printMatrix(n, rank, V, ldv, "V0");
printf("=====\n");

float ds_sup = 0;
for(int j = 0; j < n; j++){
    float err = fabs( S[j] - S_exact[j] );
    ds_sup = (ds_sup > err)? ds_sup : err;
}
printf("|S0 - S0_exact|_sup = %E \n", ds_sup);

printf("residual |A0 - U0*S0*V0**H|_F = %E \n", RnormF[0] );

printMatrix(rank, 1, S + strideS, n, "S1");
printf("=====\n");

printMatrix(m, rank, U + strideU, ldu, "U1");
printf("=====\n");

printMatrix(n, rank, V + strideV, ldv, "V1");
printf("=====\n");

ds_sup = 0;
for(int j = 0; j < n; j++){
    float err = fabs( S[strideS + j] - S_exact[strideS + j] );
    ds_sup = (ds_sup > err)? ds_sup : err;
}
printf("|S1 - S1_exact|_sup = %E \n", ds_sup);

printf("residual |A1 - U1*S1*V1**H|_F = %E \n", RnormF[1] );

/* free resources */
if (d_A ) cudaFree(d_A);
if (d_S ) cudaFree(d_S);
if (d_U ) cudaFree(d_U);
if (d_V ) cudaFree(d_V);
if (d_info ) cudaFree(d_info);
if (d_work ) cudaFree(d_work);

if (cusolverH) cusolverDnDestroy(cusolverH);
if (stream ) cudaStreamDestroy(stream);

cudaDeviceReset();

return 0;
}

```

Appendix H.

EXAMPLES OF MULTIGPU EIGENVALUE SOLVER

This chapter provides three examples to perform multiGPU symmetric eigenvalue solver. The difference among them is how to generate the testing matrix. The testing matrix is a tridiagonal matrix, from standard 3-point stencil of Laplacian operator with Dirichlet boundary condition, so each row has $(-1, 2, -1)$ signature.

The spectrum has analytic formula, we can check the accuracy of eigenvalues easily. The user can change the dimension of the matrix to measure the performance of eigenvalue solver.

The example code enables peer-to-peer access to take advantage of NVLINK. The user can check the performance by on/off peer-to-peer access.

The procedures of these three examples are 1) to prepare a tridiagonal matrix in distributed sense, 2) to query size of the workspace and to allocate the workspace for each device, 3) to compute eigenvalues and eigenvectors, and 4) to check accuracy of eigenvalues.

The example 1 allocates distributed matrix by calling **createMat**. It generates the matrix on host memory and copies it to distributed device memory via **memcpyH2D**.

The example 2 allocates distributed matrix manually, generates the matrix on host memory and copies it to distributed device memory manually. This example is for the users who are familiar with data layout of ScaLAPACK.

The example 3 allocates distributed matrix by calling **createMat** and generates the matrix element-by-element on distributed matrix via **memcpyH2D**. The user needs not to know the data layout of ScaLAPACK. It is useful when the matrix is sparse.

H.1. SYEVD of 1D Laplacian operator (example 1)

...

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *      nvcc -ccbin gcc -I/usr/local/cuda/include -c main.cpp -o main.o
 *      nvcc -cudart static main.o -lcusolverMg
 */
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <assert.h>
#include <cuda_runtime.h>
#include "cusolverMg.h"
#include "util.hxx"

// #define SHOW_FORMAT

#ifndef IDX2F
#define IDX2F(i,j,lda) (((j)-1)*((size_t)lda))+((i)-1)
#endif /* IDX2F */

#ifndef IDX1F
#define IDX1F(i) ((i)-1)
#endif /* IDX1F */

static void print_matrix(
    int m,
    int n,
    const double *A,
    int lda,
    const char* name)
{
    printf("%s = matlab base-1, %d-by-%d matrix\n", name, m, n);
    for(int row = 1 ; row <= m ; row++){
        for(int col = 1 ; col <= n ; col++){
            double Aij = A[IDX2F(row, col, lda)];
            printf("%s(%d,%d) = %20.16E\n", name, row, col, Aij );
        }
    }
}

static void gen_1d_laplacian(
    int N,
    double *A,
    int lda)
{
    memset(A, 0, sizeof(double)*lda*N);
    for(int J = 1 ; J <= N; J++){
        /* A(J,J) = 2 */
        A[ IDX2F( J, J, lda ) ] = 2.0;
        if ( (J-1) >= 1 ){
            /* A(J, J-1) = -1 */
            A[ IDX2F( J, J-1, lda ) ] = -1.0;
        }
        if ( (J+1) <= N ){
            /* A(J, J+1) = -1 */
            A[ IDX2F( J, J+1, lda ) ] = -1.0;
        }
    }
}

```


...

```

int main( int argc, char* argv[])
{
    cusolverMgHandle_t handle = NULL;
    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat = cudaSuccess;

    /* maximum number of GPUs */
    const int MAX_NUM_DEVICES = 16;

    int nbGpus = 0;
    int deviceList[MAX_NUM_DEVICES];

    const int N    = 2111;
    const int IA   = 1;
    const int JA   = 1;
    const int T_A  = 256; /* tile size */
    const int lda  = N;

    double *A = NULL; /* A is N-by-N */
    double *D = NULL; /* D is 1-by-N */
    int info = 0;

    cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR;

    cudaLibMgMatrixDesc_t descrA;
    cudaLibMgGrid_t gridA;
    cusolverMgGridMapping_t mapping = CUDALIBMG_GRID_MAPPING_COL_MAJOR;

    double **array_d_A = NULL;

    int64_t lwork = 0 ; /* workspace: number of elements per device */
    double **array_d_work = NULL;

    printf("test 1D Laplacian of order %d\n", N);

    printf("step 1: create Mg handle and select devices \n");
    status = cusolverMgCreate(&handle);
    assert(CUSOLVER_STATUS_SUCCESS == status);

    cudaStat = cudaGetDeviceCount( &nbGpus );
    assert( cudaSuccess == cudaStat );

    nbGpus = (nbGpus < MAX_NUM_DEVICES)? nbGpus : MAX_NUM_DEVICES;
    printf("\tthere are %d GPUs \n", nbGpus);
    for(int j = 0 ; j < nbGpus ; j++){
        deviceList[j] = j;
        cudaDeviceProp prop;
        cudaGetDeviceProperties(&prop, j);
        printf("\tdevice %d, %s, cc %d.%d \n",j, prop.name, prop.major,
prop.minor);
    }

    status = cusolverMgDeviceSelect(
        handle,
        nbGpus,
        deviceList);
    assert(CUSOLVER_STATUS_SUCCESS == status);

    printf("step 2: Enable peer access.\n");
    assert( 0 == enablePeerAccess( nbGpus, deviceList ) );

```

...

```

printf("step 3: allocate host memory A \n");
A = (double *)malloc (sizeof(double)*lda*N);
D = (double *)malloc (sizeof(double)*N);
assert( NULL != A );
assert( NULL != D );

printf("step 4: prepare 1D Laplacian \n");
gen_1d_laplacian(
    N,
    &A[ IDX2F( IA, JA, lda ) ],
    lda
);

#ifdef SHOW_FORMAT
    print_matrix( N, N, A, lda, "A");
#endif

printf("step 5: create matrix descriptors for A and D \n");

status = cusolverMgCreateDeviceGrid(&gridA, 1, nbGpus, deviceList,
mapping );
assert(CUSOLVER_STATUS_SUCCESS == status);
/* (global) A is N-by-N */
status = cusolverMgCreateMatrixDesc(
    &descrA,
    N, /* nubmer of rows of (global) A */
    N, /* number of columns of (global) A */
    N, /* number or rows in a tile */
    T_A, /* number of columns in a tile */
    CUDA_R_64F,
    gridA );
assert(CUSOLVER_STATUS_SUCCESS == status);

printf("step 6: allocate distributed matrices A and D \n");

array_d_A = (double**)malloc(sizeof(double*)*nbGpus);
assert(NULL != array_d_A);
/* A := 0 */
createMat<double>(
    nbGpus,
    deviceList,
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    lda, /* leading dimension of local A */
    array_d_A
);

printf("step 7: prepare data on devices \n");
memcpyH2D<double>(
    nbGpus,
    deviceList,
    N,
    N,
/* input */
    A,
    lda,
/* output */
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    lda, /* leading dimension of local A */
    array_d_A, /* host pointer array of dimension nbGpus */
    IA,
    JA
);

```

...

```

printf("step 8: allocate workspace space \n");
status = cusolverMgSyevd_bufferSize(
    handle,
    (cusolverEigMode_t)jobz,
    CUBLAS_FILL_MODE_LOWER, /* only support lower mode */
    N,
    (void**)array_d_A,
    IA, /* base-1 */
    JA, /* base-1 */
    descrA,
    (void*)D,
    CUDA_R_64F,
    CUDA_R_64F,
    &lwork);
assert(CUSOLVER_STATUS_SUCCESS == status);

printf("\tallocate device workspace, lwork = %lld \n", (long long)lwork);
array_d_work = (double**)malloc(sizeof(double)*nbGpus);
assert(NULL != array_d_work);
/* array_d_work[j] points to device workspace of device j */
workspaceAlloc(
    nbGpus,
    deviceList,
    sizeof(double)*lwork, /* number of bytes per device */
    (void**)array_d_work
);

/* sync all devices */
cudaStat = cudaDeviceSynchronize();
assert(cudaSuccess == cudaStat);

printf("step 9: compute eigenvalues and eigenvectors \n");
status = cusolverMgSyevd(
    handle,
    (cusolverEigMode_t)jobz,
    CUBLAS_FILL_MODE_LOWER, /* only support lower mode */
    N,
    (void**)array_d_A, /* exit: eigenvectors */
    IA,
    JA,
    descrA,
    (void**)D, /* exit: eigenvalues */
    CUDA_R_64F,
    CUDA_R_64F,
    (void**)array_d_work,
    lwork,
    &info /* host */
);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* sync all devices */
cudaStat = cudaDeviceSynchronize();
assert(cudaSuccess == cudaStat);

/* check if SYEVD converges */
assert(0 == info);

```

...

```

printf("step 10: copy eigenvectors to A and eigenvalues to D\n");

memcpyD2H<double>(
    nbGpus,
    deviceList,
    N,
    N,
/* input */
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    lda, /* leading dimension of local A */
    array_d_A,
    IA,
    JA,
/* output */
    A, /* N-y-N eigenvectors */
    lda
);

#ifdef SHOW_FORMAT
printf("eigenvalue D = \n");
/* D is 1-by-N */
print_matrix(1, N, D, 1, "D");
#endif

printf("step 11: verify eigenvalues \n");
printf("    lambda(k) = 4 * sin(pi/2 * k/(N+1))^2 for k = 1:N \n");
double max_err_D = 0;
for(int k = 1; k <= N ; k++){
    const double pi = 4*atan(1.0);
    const double h = 1.0/((double)N+1);
    const double factor = sin(pi/2.0 * ((double)k)*h);
    const double lambda = 4.0*factor*factor;
    const double err = fabs(D[IDX1F(k)] - lambda);
    max_err_D = (max_err_D > err)? max_err_D : err;
//    printf("k = %d, D = %E, lambda = %E, err = %E\n", k, D[IDX1F(k)],
lambda, err);
}
printf("\n|D - lambda|_inf = %E\n\n", max_err_D);

printf("step 12: free resources \n");
destroyMat(
    nbGpus,
    deviceList,
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    (void**)array_d_A );

workspaceFree( nbGpus, deviceList, (void**)array_d_work );

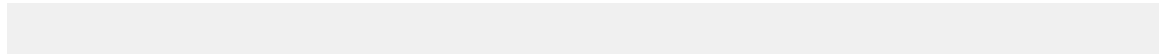
if (NULL != A) free(A);
if (NULL != D) free(D);

if (NULL != array_d_A ) free(array_d_A);
if (NULL != array_d_work) free(array_d_work);

return 0;
}

```

...



...

```

/* util.hxx */

/*
 * nbGpus : (int) number of gpus in deviceList array.
 * deviceList : (*int) list of device ids.
 *
 * The function restores the input device before leaving.
 */
static int enablePeerAccess (const int nbGpus, const int *deviceList )
{
    int currentDevice = 0;
    cudaGetDevice( &currentDevice );

    /* Remark: access granted by this cudaDeviceEnablePeerAccess is
    unidirectional */
    /* Rows and columns represents a connectivity matrix between GPUs in the
    system */
    for(int row=0; row < nbGpus; row++) {
        cudaSetDevice(row);
        for(int col=0; col < nbGpus; col++) {
            if( row != col ){
                cudaError_t cudaStat1 = cudaSuccess;
                cudaError_t cudaStat2 = cudaSuccess;
                int canAccessPeer = 0;
                cudaStat1 = cudaDeviceCanAccessPeer( &canAccessPeer, row, col );
                if ( canAccessPeer ){
                    printf("\t Enable peer access from gpu %d to gpu %d\n", row,
col );
                    cudaStat2 = cudaDeviceEnablePeerAccess( col, 0 );
                }
                assert(cudaStat1 == cudaSuccess);
                assert(cudaStat2 == cudaSuccess);
            }
        }
        cudaSetDevice( currentDevice );
        return 0;
    }
}

static int workspaceFree(
    int num_devices,
    const int *deviceIdA, /* <int> dimension num_devices */
    void **array_d_work /* <t> num_devices, host array */
    j */
    )
{
    int currentDev = 0; /* record current device ID */
    cudaGetDevice( &currentDev );

    for(int idx = 0 ; idx < num_devices ; idx++){
        int deviceId = deviceIdA[idx];
        /* WARNING: we need to set device before any runtime API */
        cudaSetDevice( deviceId );

        if (NULL != array_d_work[idx]){
            cudaFree(array_d_work[idx]);
        }
    }
    cudaSetDevice(currentDev);
    return 0;
}

```

...

```

static int workspaceAlloc(
    int num_devices,
    const int *deviceIdA, /* <int> dimension num_devices */
    size_t sizeInBytes, /* number of bytes per device */
    void **array_d_work /* <t> num_devices, host array */
    /* array_d_work[j] points to device workspace of device
    j */
)
{
    cudaError_t cudaStat1 = cudaSuccess;

    int currentDev = 0; /* record current device ID */
    cudaGetDevice( &currentDev );

    memset(array_d_work, 0, sizeof(void*)*num_devices);
    for(int idx = 0 ; idx < num_devices ; idx++){
        int deviceId = deviceIdA[idx];
/* WARNING: we need to set device before any runtime API */
        cudaSetDevice( deviceId );

        void *d_workspace = NULL;

        cudaStat1 = cudaMalloc(&d_workspace, sizeInBytes);
        assert( cudaSuccess == cudaStat1 );
        array_d_work[idx] = d_workspace;
    }
    cudaSetDevice(currentDev);
    return 0;
}

/* create a empty matrix A with A := 0 */
template <typename T_ELEM>
int createMat(
    int num_devices,
    const int *deviceIdA, /* <int> dimension num_devices */
    int N_A, /* number of columns of global A */
    int T_A, /* number of columns per column tile */
    int LLD_A, /* leading dimension of local A */
    T_ELEM **array_d_A /* host pointer array of dimension num_devices */
)
{
    cudaError_t cudaStat1 = cudaSuccess;
    int currentDev = 0; /* record current device id */
    cudaGetDevice( &currentDev );
    cudaDeviceSynchronize();
    const int A_num_blks = ( N_A + T_A - 1 ) / T_A;
    const int max_A_num_blks_per_device = (A_num_blks + num_devices-1)/
num_devices;
/* Allocate base pointers */
    memset(array_d_A, 0, sizeof(T_ELEM*) * num_devices);
    for( int p = 0 ; p < num_devices ; p++){
        cudaStat1 = cudaSetDevice(deviceIdA[p]);
        assert(cudaSuccess == cudaStat1);
/* Allocate max_A_num_blks_per_device blocks per device */
        cudaStat1 =
        cudaMalloc( &(array_d_A[p]), sizeof(T_ELEM)*LLD_A*T_A*max_A_num_blks_per_device );
        assert(cudaSuccess == cudaStat1);
/* A := 0 */
        cudaStat1 = cudaMemset( array_d_A[p],
0, sizeof(T_ELEM)*LLD_A*T_A*max_A_num_blks_per_device );
        assert(cudaSuccess == cudaStat1);
    }
    cudaDeviceSynchronize();
    cudaSetDevice(currentDev);
    return 0;
}

```

...

```

static int destroyMat (
    int num_devices,
    const int *deviceIdA, /* <int> dimension num devices */
    int N_A, /* number of columns of global A */
    int T_A, /* number of columns per column tile */
    void **array_d_A) /* host pointer array of dimension num_devices */
{
    cudaError_t cudaStat = cudaSuccess;

    int currentDev = 0; /* record current device id */
    cudaGetDevice( &currentDev );

    const int num_blks = (N_A + T_A - 1) / T_A;
    for( int p = 0 ; p < num_devices ; p++){
        cudaStat = cudaSetDevice(deviceIdA[p]);
        assert(cudaSuccess == cudaStat);

        if ( NULL != array_d_A[p] ){
            cudaFree( array_d_A[p] );
        }
    }
    memset(array_d_A, 0, sizeof(void*)*num_devices);
    cudaSetDevice(currentDev);
    return 0;
}

template <typename T_ELEM>
static int mat_pack2unpack(
    int num_devices,
    int N_A, /* number of columns of global A */
    int T_A, /* number of columns per column tile */
    int LLD_A, /* leading dimension of local A */
    T_ELEM **array_d_A_packed, /* host pointer array of dimension num_devices */
    /* output */
    T_ELEM **array_d_A_unpacked /* host pointer array of dimension num_blks */
)
{
    const int num_blks = ( N_A + T_A - 1 ) / T_A;

    for(int p_a = 0 ; p_a < num_devices ; p_a++){
        T_ELEM *d_A = array_d_A_packed[p_a];
        int nz_blks = 0;
        for(int JA_blk_id = p_a ; JA_blk_id < num_blks ; JA_blk_id+=num_devices)
        {
            array_d_A_unpacked[JA_blk_id] = d_A + (size_t)LLD_A * T_A *
nz_blks ;
            nz_blks++;
        }
    }
    return 0;
}

```


...

```

/*
 * A(IA:IA+M-1, JA:JA+N-1) := B(1:M, 1:N)
 */
template <typename T_ELEM>
static int memcpyH2D(
    int num_devices,
    const int *deviceIdA, /* <int> dimension num_devices */
    int M, /* number of rows in local A, B */
    int N, /* number of columns in local A, B */
/* input */
    const T_ELEM *h_B, /* host array, h_B is M-by-N with leading dimension ldb
 */
    int ldb,
/* output */
    int N_A, /* number of columns of global A */
    int T_A, /* number of columns per column tile */
    int LLD_A, /* leading dimension of local A */
    T_ELEM **array_d_A_packed, /* host pointer array of dimension num_devices */
    int IA, /* base-1 */
    int JA /* base-1 */
)
{
    cudaError_t cudaStat1 = cudaSuccess;

    int currentDev = 0; /* record current device id */

/* Quick return if possible */
    if ( (0 >= M) || (0 >= N) ){
        return 0;
    }

/* consistent checking */
    if ( ldb < M ){
        return 1;
    }

    cudaGetDevice( &currentDev );
    cudaDeviceSynchronize();

    const int num_blks = ( N_A + T_A - 1 ) / T_A;

    T_ELEM **array_d_A_unpacked = (T_ELEM**)malloc(sizeof(T_ELEM*)*num_blks);
    assert(NULL != array_d_A_unpacked);

    mat_pack2unpack<T_ELEM>(
        num_devices,
        N_A, /* number of columns of global A */
        T_A, /* number of columns per column tile */
        LLD_A, /* leading dimension of local A */
        array_d_A_packed, /* host pointer array of size num_devices */
/* output */
        array_d_A_unpacked /* host pointer array of size num_blks */
    );

/* region of interest is A(IA:IA+N-1, JA:JA+N-1) */
    const int N_hat = (JA-1) + N; /* JA is base-1 */

    const int JA_start_blk_id = (JA-1)/T_A;
    const int JA_end_blk_id = (N_hat-1)/T_A;

```

...

```

    for(int p_a = 0 ; p_a < num_devices ; p_a++){
/* region of interest: JA_start_blk_id:JA_end_blk_id */
        for(int JA_blk_id = p_a; JA_blk_id <= JA_end_blk_id ; JA_blk_id
+=num_devices){
            if ( JA_blk_id < JA_start_blk_id ) { continue; }
/*
 * process column block of A
 *   A(A_start_row:M_A, A_start_col : (A_start_col + IT_A-1) )
 */
            const int IBX_A = (1 + JA_blk_id*T_A); /* base-1 */
            const int A_start_col = imax( JA, IBX_A ); /* base-1 */
            const int A_start_row = IA; /* base-1 */

            const int bdd = imin( N_hat, (IBX_A + T_A -1) );
            const int IT_A = imin( T_A, (bdd - A_start_col + 1) );

            const int loc_A_start_row = A_start_row; /* base-1 */
            const int loc_A_start_col = (A_start_col-IBX_A)+1; /* base-1 */

            T_ELEM *d_A = array_d_A_unpacked[JA_blk_id] +
IDX2F( loc_A_start_row, loc_A_start_col, LLD_A );
            const T_ELEM *h_A = h_B + IDX2F( A_start_row - IA + 1, A_start_col -
JA + 1, ldb );

            cudaStat1 = cudaMemcpy2D(
                d_A, /* dst */
                (size_t)LLD_A * sizeof(T_ELEM),
                h_A, /* src */
                (size_t)ldb * sizeof(T_ELEM),
                (size_t)M * sizeof(T_ELEM),
                (size_t)IT_A,
                cudaMemcpyHostToDevice
            );
            assert( cudaSuccess == cudaStat1 );
        } /* for each tile per device */
    } /* for each device */
    cudaDeviceSynchronize();
    cudaSetDevice(currentDev);

    if ( NULL != array_d_A_unpacked ) { free(array_d_A_unpacked); }
    return 0;
}

/*
 * B(1:M, 1:N) := A(IA:IA+M-1, JA:JA+N-1)
 */
template <typename T_ELEM>
static int memcpyD2H(
    int num_devices,
    const int *deviceIdA, /* <int> dimension num_devices */
    int M, /* number of rows in local A, B */
    int N, /* number of columns in local A, B */
    /* input */
    int N_A, /* number of columns of global A */
    int T_A, /* number of columns per column tile */
    int LLD_A, /* leading dimension of local A */
    T_ELEM **array_d_A_packed, /* host pointer array of dimension num_devices */
    int IA, /* base-1 */
    int JA, /* base-1 */
    /* output */
    T_ELEM *h_B, /* host array, h_B is M-by-N with leading dimension ldb */
    int ldb
)
{

```

...

```

    cudaError_t cudaStat1 = cudaSuccess;
    int currentDev = 0; /* record current device id */

/* Quick return if possible */
    if ( (0 >= M) || (0 >= N) ){
        return 0;
    }
/* consistent checking */
    if ( ldb < M ){
        return 1;
    }
    cudaGetDevice( &currentDev );
    cudaDeviceSynchronize();

    const int num_blks = ( N_A + T_A - 1 ) / T_A;
    T_ELEM **array_d_A_unpacked = (T_ELEM**)malloc(sizeof(T_ELEM)*num_blks);
    assert(NULL != array_d_A_unpacked);

    mat_pack2unpack<T_ELEM>(
        num_devices,
        N_A, /* number of columns of global A */
        T_A, /* number of columns per column tile */
        LLD_A, /* leading dimension of local A */
        array_d_A_packed, /* host pointer array of size num_devices */
        array_d_A_unpacked /* host pointer array of size num_blks */
    );
/* region of interest is A(IA:IA+N-1, JA:JA+N-1) */
    const int N_hat = (JA-1) + N; /* JA is base-1 */
    const int JA_start_blk_id = (JA-1)/T_A;
    const int JA_end_blk_id = (N_hat-1)/T_A;
    for(int p_a = 0 ; p_a < num_devices ; p_a++){
/* region of interest: JA_start_blk_id:JA_end_blk_id */
        for(int JA_blk_id = p_a; JA_blk_id <= JA_end_blk_id ; JA_blk_id
+=num_devices){
            if ( JA_blk_id < JA_start_blk_id ) { continue; }
/* process column block, A(A_start_row:M_A, A_start_col : (A_start_col +
IT_A-1) ) */
            const int IBX_A = (1 + JA_blk_id*T_A); /* base-1 */
            const int A_start_col = imax( JA, IBX_A ); /* base-1 */
            const int A_start_row = IA; /* base-1 */
            const int bdd = imin( N_hat, (IBX_A + T_A -1) );
            const int IT_A = imin( T_A, (bdd - A_start_col + 1) );
            const int loc_A_start_row = A_start_row; /* base-1 */
            const int loc_A_start_col = (A_start_col-IBX_A)+1; /* base-1 */
            const T_ELEM *d_A = array_d_A_unpacked[JA_blk_id] +
IDX2F( loc_A_start_row, loc_A_start_col, LLD_A );
            T_ELEM *h_A = h_B + IDX2F( A_start_row - IA + 1, A_start_col - JA +
1, ldb );
            cudaStat1 = cudaMemcpy2D(
                h_A, /* dst */
                (size_t)ldb * sizeof(T_ELEM),
                d_A, /* src */
                (size_t)LLD_A * sizeof(T_ELEM),
                (size_t)M * sizeof(T_ELEM),
                (size_t)IT_A,
                cudaMemcpyDeviceToHost
            );
            assert( cudaSuccess == cudaStat1 );
        } /* for each tile per device */
    } /* for each device */
    cudaDeviceSynchronize();
    cudaSetDevice(currentDev);
    if ( NULL != array_d_A_unpacked ) { free(array_d_A_unpacked); }
    return 0;
}

```

H.2. SYEVD of 1D Laplacian operator (example 2)

...

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *      nvcc -ccbin gcc -I/usr/local/cuda/include -c main.cpp -o main.o
 *      nvcc -cudart static main.o -lcusolverMg
 */
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <assert.h>
#include <cuda_runtime.h>
#include "cusolverMg.h"
#include "util.hxx"

// #define SHOW_FORMAT

#ifndef IDX2F
#define IDX2F(i,j,lda) (((j)-1)*((size_t)lda))+((i)-1)
#endif /* IDX2F */

#ifndef IDX1F
#define IDX1F(i) ((i)-1)
#endif /* IDX1F */

#define imin(x,y) ((x) < (y)) ? (x) : (y)

static void print_matrix(
    int m,
    int n,
    const double *A,
    int lda,
    const char* name)
{
    printf("%s = matlab base-1, %d-by-%d matrix\n", name, m, n);
    for(int row = 1 ; row <= m ; row++){
        for(int col = 1 ; col <= n ; col++){
            double Aij = A[IDX2F(row, col, lda)];
            printf("%s(%d,%d) = %20.16E\n", name, row, col, Aij );
        }
    }
}

static void gen_1d_laplacian(
    int N,
    double *A,
    int lda)
{
    memset(A, 0, sizeof(double)*lda*N);
    for(int J = 1 ; J <= N; J++){
        /* A(J,J) = 2 */
        A[ IDX2F( J, J, lda ) ] = 2.0;
        if ( (J-1) >= 1 ){
            /* A(J, J-1) = -1 */
            A[ IDX2F( J, J-1, lda ) ] = -1.0;
        }
        if ( (J+1) <= N ){
            /* A(J, J+1) = -1 */
            A[ IDX2F( J, J+1, lda ) ] = -1.0;
        }
    }
}

```

...

```

int main( int argc, char* argv[])
{
    cusolverMgHandle_t handle = NULL;
    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat = cudaSuccess;
/* maximum number of GPUs */
    const int MAX_NUM_DEVICES = 16;

    int nbGpus = 0;
    int deviceList[MAX_NUM_DEVICES];

    const int N    = 2111;
    const int IA   = 1;
    const int JA   = 1;
    const int TA   = 256; /* tile size */
    const int lda  = N;

    double *A = NULL; /* A is N-by-N */
    double *D = NULL; /* D is 1-by-N */
    int info = 0;

    cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR;

    cudaLibMgMatrixDesc_t descrA;
    cudaLibMgGrid_t gridA;
    cusolverMgGridMapping_t mapping = CUDALIBMG_GRID_MAPPING_COL_MAJOR;

    double **array_d_A = NULL;

    int64_t lwork = 0 ; /* workspace: number of elements per device */
    double **array_d_work = NULL;

    printf("test 1D Laplacian of order %d\n", N);

    printf("step 1: create Mg handle and select devices \n");
    status = cusolverMgCreate(&handle);
    assert(CUSOLVER_STATUS_SUCCESS == status);

    cudaStat = cudaGetDeviceCount( &nbGpus );
    assert( cudaSuccess == cudaStat );

    nbGpus = (nbGpus < MAX_NUM_DEVICES)? nbGpus : MAX_NUM_DEVICES;
    printf("\tthere are %d GPUs \n", nbGpus);
    for(int j = 0 ; j < nbGpus ; j++){
        deviceList[j] = j;
        cudaDeviceProp prop;
        cudaGetDeviceProperties(&prop, j);
        printf("\tdevice %d, %s, cc %d.%d \n",j, prop.name, prop.major,
prop.minor);
    }

    status = cusolverMgDeviceSelect(
        handle,
        nbGpus,
        deviceList);
    assert(CUSOLVER_STATUS_SUCCESS == status);

    printf("step 2: Enable peer access \n");
    assert( 0 == enablePeerAccess( nbGpus, deviceList ) );

```

...

```

printf("step 3: allocate host memory A \n");
A = (double *)malloc (sizeof(double)*lda*N);
D = (double *)malloc (sizeof(double)*N);
assert( NULL != A );
assert( NULL != D );

printf("step 4: prepare 1D Laplacian \n");
gen_1d_laplacian(
    N,
    &A[ IDX2F( IA, JA, lda ) ],
    lda
);

#ifdef SHOW_FORMAT
    print_matrix( N, N, A, lda, "A");
#endif

printf("step 5: create matrix descriptors for A and D \n");

status = cusolverMgCreateDeviceGrid(&gridA, 1, nbGpus, deviceList,
mapping );
assert(CUSOLVER_STATUS_SUCCESS == status);
/* (global) A is N-by-N */
status = cusolverMgCreateMatrixDesc(
    &descrA,
    N, /* nubmer of rows of (global) A */
    N, /* number of columns of (global) A */
    N, /* number or rows in a tile */
    T_A, /* number of columns in a tile */
    CUDA_R_64F,
    gridA );
assert(CUSOLVER_STATUS_SUCCESS == status);

printf("step 6: allocate distributed matrices A and D \n");

array_d_A = (double**) malloc (sizeof(double*) * nbGpus );
assert( NULL != array_d_A );

const int A_num_blks = ( N + T_A - 1 ) / T_A;
const int blks_per_device = (A_num_blks + nbGpus-1)/nbGpus;

for( int p = 0 ; p < nbGpus ; p++){
    cudaSetDevice(deviceList[p]);
    cudaStat =
    cudaMalloc( &(array_d_A[p]), sizeof(double)*lda*T_A*blks_per_device );
    assert(cudaSuccess == cudaStat);
}

printf("step 7: prepare data on devices \n");
/* The following setting only works for IA = JA = 1 */
for( int k = 0 ; k < A_num_blks ; k++){
/* k = ibx * nbGpus + p */
    const int p = (k % nbGpus);
    const int ibx = (k / nbGpus);
    double *h_Ak = A + (size_t)lda*T_A*k;
    double *d_Ak = array_d_A[p] + (size_t)lda*T_A*ibx;
    const int width = imin( T_A, (N - T_A*k) );
    cudaStat = cudaMemcpy(d_Ak, h_Ak, sizeof(double)*lda*width,
    cudaMemcpyHostToDevice);
    assert(cudaSuccess == cudaStat);
}
/* sync all devices */
cudaStat = cudaDeviceSynchronize();
assert(cudaSuccess == cudaStat);

```

...

```

printf("step 8: allocate workspace space \n");
status = cusolverMgSyevd_bufferSize(
    handle,
    (cusolverEigMode_t)jobz,
    CUBLAS_FILL_MODE_LOWER, /* only support lower mode */
    N,
    (void**)array_d_A,
    IA, /* base-1 */
    JA, /* base-1 */
    descrA,
    (void*)D,
    CUDA_R_64F,
    CUDA_R_64F,
    &lwork);
assert(CUSOLVER_STATUS_SUCCESS == status);

printf("\tallocate device workspace, lwork = %lld \n", (long long)lwork);
array_d_work = (double**)malloc(sizeof(double)*nbGpus);
assert(NULL != array_d_work);
/* array_d_work[j] points to device workspace of device j */
workspaceAlloc(
    nbGpus,
    deviceList,
    sizeof(double)*lwork, /* number of bytes per device */
    (void**)array_d_work
);

/* sync all devices */
cudaStat = cudaDeviceSynchronize();
assert(cudaSuccess == cudaStat);

printf("step 9: compute eigenvalues and eigenvectors \n");
status = cusolverMgSyevd(
    handle,
    (cusolverEigMode_t)jobz,
    CUBLAS_FILL_MODE_LOWER, /* only support lower mode */
    N,
    (void**)array_d_A, /* exit: eigenvectors */
    IA,
    JA,
    descrA,
    (void*)D, /* exit: eigenvalues */
    CUDA_R_64F,
    CUDA_R_64F,
    (void**)array_d_work,
    lwork,
    &info /* host */
);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* sync all devices */
cudaStat = cudaDeviceSynchronize();
assert(cudaSuccess == cudaStat);

/* check if SYEVD converges */
assert(0 == info);

```

...

```

printf("step 10: copy eigenvectors to A and eigenvalues to D\n");

memcpyD2H<double>(
    nbGpus,
    deviceList,
    N,
    N,
/* input */
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    lda, /* leading dimension of local A */
    array_d_A,
    IA,
    JA,
/* output */
    A, /* N-y-N eigenvectors */
    lda
);

#ifdef SHOW_FORMAT
printf("eigenvalue D = \n");
/* D is 1-by-N */
print_matrix(1, N, D, 1, "D");
#endif

printf("step 11: verify eigenvalues \n");
printf("    lambda(k) = 4 * sin(pi/2 * k/(N+1))^2 for k = 1:N \n");
double max_err_D = 0;
for(int k = 1; k <= N ; k++){
    const double pi = 4*atan(1.0);
    const double h = 1.0/((double)N+1);
    const double factor = sin(pi/2.0 * ((double)k)*h);
    const double lambda = 4.0*factor*factor;
    const double err = fabs(D[IDX1F(k)] - lambda);
    max_err_D = (max_err_D > err)? max_err_D : err;
//    printf("k = %d, D = %E, lambda = %E, err = %E\n", k, D[IDX1F(k)],
lambda, err);
}
printf("\n|D - lambda|_inf = %E\n\n", max_err_D);

printf("step 12: free resources \n");
destroyMat(
    nbGpus,
    deviceList,
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    (void**)array_d_A );

workspaceFree( nbGpus, deviceList, (void**)array_d_work );

if (NULL != A) free(A);
if (NULL != D) free(D);

if (NULL != array_d_A ) free(array_d_A);
if (NULL != array_d_work) free(array_d_work);

return 0;
}

```


H.3. SYEVD of 1D Laplacian operator (example 3)

...

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *      nvcc -ccbin gcc -I/usr/local/cuda/include -c main.cpp -o main.o
 *      nvcc -cudart static main.o -lcusolverMg
 */
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <assert.h>
#include <cuda_runtime.h>
#include "cusolverMg.h"
#include "util.hxx"

// #define SHOW_FORMAT

#ifndef IDX2F
#define IDX2F(i,j,lda) (((j)-1)*((size_t)lda))+((i)-1)
#endif /* IDX2F */
#ifndef IDX1F
#define IDX1F(i) ((i)-1)
#endif /* IDX1F */

static void print_matrix(
    int m,
    int n,
    const double *A,
    int lda,
    const char* name)
{
    printf("%s = matlab base-1, %d-by-%d matrix\n", name, m, n);
    for(int row = 1 ; row <= m ; row++){
        for(int col = 1 ; col <= n ; col++){
            double Aij = A[IDX2F(row, col, lda)];
            printf("%s(%d,%d) = %20.16E\n", name, row, col, Aij );
        }
    }
}

/* the caller must set A = 0 */
static void gen_1d_laplacian(
    cusolverMgHandle_t handle,
    int nbGpus,
    const int *deviceList,
    int N,          /* number of columns of global A */
    int T_A,        /* number of columns per column tile */
    int LLD_A,      /* leading dimension of local A */
    double **array_d_A /* host pointer array of dimension nbGpus */
)
{
    double two = 2.0;
    double minus_one = -1.0;
    for(int J = 1 ; J <= N; J++){
        /* A(J,J) = 2 */
        memcpyH2D<double>(nbGpus, deviceList, 1, 1, &two, 1, N, T_A, LLD_A,
            array_d_A, J, J);
        if ( (J-1) >= 1 ){
            /* A(J, J-1) = -1 */
            memcpyH2D<double>(nbGpus, deviceList, 1, 1, &minus_one, 1, N, T_A,
                LLD_A, array_d_A, J, J-1);
        }
        if ( (J+1) <= N ){
            /* A(J, J+1) = -1 */
            memcpyH2D<double>(nbGpus, deviceList, 1, 1, &minus_one, 1, N, T_A,
                LLD_A, array_d_A, J, J+1);
        }
    }
}

```

...

```

int main( int argc, char* argv[])
{
    cusolverMgHandle_t handle = NULL;
    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat = cudaSuccess;

    /* maximum number of GPUs */
    const int MAX_NUM_DEVICES = 16;

    int nbGpus = 0;
    int deviceList[MAX_NUM_DEVICES];

    const int N    = 2111;
    const int IA   = 1;
    const int JA   = 1;
    const int T_A  = 256; /* tile size */
    const int lda  = N;

    double *A = NULL; /* A is N-by-N */
    double *D = NULL; /* D is 1-by-N */
    int info = 0;

    cusolverEigMode_t jobz = CUSOLVER_EIG_MODE_VECTOR;

    cudaLibMgMatrixDesc_t descrA;
    cudaLibMgGrid_t gridA;
    cusolverMgGridMapping_t mapping = CUDALIBMG_GRID_MAPPING_COL_MAJOR;

    double **array_d_A = NULL;

    int64_t lwork = 0 ; /* workspace: number of elements per device */
    double **array_d_work = NULL;

    printf("test 1D Laplacian of order %d\n", N);

    printf("step 1: create Mg handle and select devices \n");
    status = cusolverMgCreate(&handle);
    assert(CUSOLVER_STATUS_SUCCESS == status);

    cudaStat = cudaGetDeviceCount( &nbGpus );
    assert( cudaSuccess == cudaStat );

    nbGpus = (nbGpus < MAX_NUM_DEVICES)? nbGpus : MAX_NUM_DEVICES;
    printf("\tthere are %d GPUs \n", nbGpus);
    for(int j = 0 ; j < nbGpus ; j++){
        deviceList[j] = j;
        cudaDeviceProp prop;
        cudaGetDeviceProperties(&prop, j);
        printf("\tdevice %d, %s, cc %d.%d \n",j, prop.name, prop.major,
prop.minor);
    }

    status = cusolverMgDeviceSelect(
        handle,
        nbGpus,
        deviceList);
    assert(CUSOLVER_STATUS_SUCCESS == status);

    printf("step 2: Enable peer access.\n");
    assert( 0 == enablePeerAccess( nbGpus, deviceList ) );

```

...

```

printf("step 3: allocate host memory A \n");
A = (double *)malloc (sizeof(double)*lda*N);
D = (double *)malloc (sizeof(double)*N);
assert( NULL != A );
assert( NULL != D );

printf("step 4: create matrix descriptors for A and D \n");
status = cusolverMgCreateDeviceGrid(&gridA, 1, nbGpus, deviceList,
mapping );
assert(CUSOLVER_STATUS_SUCCESS == status);
/* (global) A is N-by-N */
status = cusolverMgCreateMatrixDesc(
    &descrA,
    N, /* number of rows of (global) A */
    N, /* number of columns of (global) A */
    N, /* number of rows in a tile */
    T_A, /* number of columns in a tile */
    CUDA_R_64F,
    gridA );
assert(CUSOLVER_STATUS_SUCCESS == status);

printf("step 5: allocate distributed matrices A and D, A = 0 and D = 0 \n");

array_d_A = (double**)malloc(sizeof(double*)*nbGpus);
assert(NULL != array_d_A);

/* A := 0 */
createMat<double>(
    nbGpus,
    deviceList,
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    lda, /* leading dimension of local A */
    array_d_A );

printf("step 6: prepare 1D Laplacian on devices \n");
gen_1d_laplacian(
    handle,
    nbGpus,
    deviceList,
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    lda, /* leading dimension of local A */
    array_d_A );

printf("step 7: allocate workspace space \n");
status = cusolverMgSyevd_bufferSize(
    handle,
    (cusolverEigMode_t)jobz,
    CUBLAS_FILL_MODE_LOWER, /* only support lower mode */
    N,
    (void**)array_d_A,
    IA, /* base-1 */
    JA, /* base-1 */
    descrA,
    (void*)D,
    CUDA_R_64F,
    CUDA_R_64F,
    &lwork);
assert(CUSOLVER_STATUS_SUCCESS == status);

```

...

```

printf("\tallocate device workspace, lwork = %lld \n", (long long)lwork);
array_d_work = (double**)malloc(sizeof(double*)*nbGpus);
assert( NULL != array_d_work);
/* array_d_work[j] points to device workspace of device j */
workspaceAlloc(
    nbGpus,
    deviceList,
    sizeof(double)*lwork, /* number of bytes per device */
    (void**)array_d_work
);

/* sync all devices */
cudaStat = cudaDeviceSynchronize();
assert(cudaSuccess == cudaStat);

printf("step 8: compute eigenvalues and eigenvectors \n");
status = cusolverMgSyevd(
    handle,
    (cusolverEigMode_t)jobz,
    CUBLAS_FILL_MODE_LOWER, /* only support lower mode */
    N,
    (void**)array_d_A, /* exit: eigenvectors */
    IA,
    JA,
    descrA,
    (void*)D, /* exit: eigenvalues */
    CUDA_R_64F,
    CUDA_R_64F,
    (void**)array_d_work,
    lwork,
    &info /* host */
);
assert(CUSOLVER_STATUS_SUCCESS == status);

/* sync all devices */
cudaStat = cudaDeviceSynchronize();
assert(cudaSuccess == cudaStat);

/* check if SYEVD converges */
assert(0 == info);

printf("step 9: copy eigenvectors to A and eigenvalues to D\n");
memcpyD2H<double>(
    nbGpus,
    deviceList,
    N,
    N,
/* input */
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    lda, /* leading dimension of local A */
    array_d_A,
    IA,
    JA,
/* output */
    A, /* N-y-N eigenvectors */
    lda
);
#ifdef SHOW_FORMAT
printf("eigenvalue D = \n");
/* D is 1-by-N */
print_matrix(1, N, D, 1, "D");
#endif

```

...

```

printf("step 10: verify eigenvalues \n");
printf("    lambda(k) = 4 * sin(pi/2 * k/(N+1))^2 for k = 1:N \n");
double max_err_D = 0;
for(int k = 1; k <= N ; k++){
    const double pi = 4*atan(1.0);
    const double h = 1.0/((double)N+1);
    const double factor = sin(pi/2.0 * ((double)k)*h);
    const double lambda = 4.0*factor*factor;
    const double err = fabs(D[IDX1F(k)] - lambda);
    max_err_D = (max_err_D > err)? max_err_D : err;
//    printf("k = %d, D = %E, lambda = %E, err = %E\n", k, D[IDX1F(k)],
lambda, err);
}
printf("\n|D - lambda|_inf = %E\n\n", max_err_D);

printf("step 11: free resources \n");
destroyMat(
    nbGpus,
    deviceList,
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    (void**)array_d_A );

workspaceFree( nbGpus, deviceList, (void**)array_d_work );

if (NULL != A) free(A);
if (NULL != D) free(D);

if (NULL != array_d_A ) free(array_d_A);
if (NULL != array_d_work) free(array_d_work);

return 0;
}

```

Appendix I.

EXAMPLES OF MULTIGPU LINEAR SOLVER

This chapter provides examples to perform multiGPU linear solver.

The example code enables peer-to-peer access to take advantage of NVLINK. The user can check the performance by on/off peer-to-peer access.

The example 1 solves linear system by LU with partial pivoting (**getrf** and **getrs**). It allocates distributed matrix by calling **createMat**. Then generates the matrix on host memory and copies it to distributed device memory via **memcpyH2D**.

I.1. GETRF and GETRS of 1D Laplacian operator (example 1)

Please refer H.1 for util.hxx.

```

/*
 * How to compile (assume cuda is installed at /usr/local/cuda/)
 *      nvcc -ccbin gcc -I/usr/local/cuda/include -c main.cpp -o main.o
 *      nvcc -cudart static main.o -lcusolverMg
 */
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <assert.h>
#include <cuda_runtime.h>
#include "cusolverMg.h"
#include "util.hxx"

// #define SHOW_FORMAT

#ifndef IDX2F
#define IDX2F(i,j,lda) (((j)-1)*((size_t)lda))+((i)-1)
#endif /* IDX2F */

#ifndef IDX1F
#define IDX1F(i) ((i)-1)
#endif /* IDX1F */

static void print_matrix(
    int m,
    int n,
    const double *A,
    int lda,
    const char* name)
{
    printf("%s = matlab base-1, %d-by-%d matrix\n", name, m, n);
    for(int row = 1 ; row <= m ; row++){
        for(int col = 1 ; col <= n ; col++){
            double Aij = A[IDX2F(row, col, lda)];
            printf("%s(%d,%d) = %20.16E\n", name, row, col, Aij );
        }
    }
}

/* compute ||x||_inf */
static double vec_nrm_inf(
    int n,
    const double *x)
{
    double max_nrm = 0;
    for(int row = 1; row <= n ; row++){
        double xi = x[ IDX1F(row) ];
        max_nrm = ( max_nrm > fabs(xi) )? max_nrm : fabs(xi);
    }
    return max_nrm;
}

```

...

```

/* A is 1D laplacian, return A(N:-1:1, :) */
static void gen_ld_laplacian_perm(
    int N,
    double *A,
    int lda)
{
    memset(A, 0, sizeof(double)*lda*N);
    for(int J = 1 ; J <= N; J++){
        A[ IDX2F( N-J+1, J, lda ) ] = 2.0;
        if ( (J-1) >= 1 ){
            A[ IDX2F( N-J+1, J-1, lda ) ] = -1.0;
        }
        if ( (J+1) <= N ){
            A[ IDX2F( N-J+1, J+1, lda ) ] = -1.0;
        }
    }
}

int main( int argc, char* argv[])
{
    cusolverMgHandle_t handle = NULL;
    cusolverStatus_t status = CUSOLVER_STATUS_SUCCESS;
    cudaError_t cudaStat = cudaSuccess;
/* maximum number of GPUs */
    const int MAX_NUM_DEVICES = 16;

    int nbGpus = 0;
    int deviceList[MAX_NUM_DEVICES];

    const int N    = 611;
    const int IA   = 1;
    const int JA   = 1;
    const int T_A  = 256; /* tile size of A */
    const int lda  = N;

    const int IB   = 1;
    const int JB   = 1;
    const int T_B  = 100; /* tile size of B */
    const int ldb  = N;

    double *A = NULL; /* A is N-by-N */
    double *B = NULL; /* B is N-by-1, right-hand-side vector */
    double *X = NULL; /* X is N-by-1, solution vector */
    int *IPIV = NULL; /* IPIV is 1-by-N, pivoting sequence */
    int info = 0;

    cudaLibMgMatrixDesc_t descrA;
    cudaLibMgMatrixDesc_t descrB;
    cudaLibMgGrid_t gridA;
    cudaLibMgGrid_t gridB;
    cusolverMgGridMapping_t mapping = CUDALIBMG_GRID_MAPPING_COL_MAJOR;

    double **array_d_A = NULL;
    double **array_d_B = NULL;
    int **array_d_IPIV = NULL;

    int64_t lwork_getrf = 0 ;
    int64_t lwork_getrs = 0 ;
    int64_t lwork = 0 ; /* workspace: number of elements per device */
    double **array_d_work = NULL;

    printf("test permuted 1D Laplacian of order %d\n", N);

```


...

```

printf("step 1: create Mg handle and select devices \n");
status = cusolverMgCreate(&handle);
assert(CUSOLVER_STATUS_SUCCESS == status);

cudaStat = cudaGetDeviceCount( &nbGpus );
assert( cudaSuccess == cudaStat );

nbGpus = (nbGpus < MAX_NUM_DEVICES)? nbGpus : MAX_NUM_DEVICES;
printf("\tthere are %d GPUs \n", nbGpus);
for(int j = 0 ; j < nbGpus ; j++){
    deviceList[j] = j;
    cudaDeviceProp prop;
    cudaGetDeviceProperties(&prop, j);
    printf("\tdevice %d, %s, cc %d.%d \n",j, prop.name, prop.major,
prop.minor);
}

status = cusolverMgDeviceSelect(
    handle,
    nbGpus,
    deviceList);
assert(CUSOLVER_STATUS_SUCCESS == status);

printf("step 2: Enable peer access.\n");
assert( 0 == enablePeerAccess( nbGpus, deviceList ) );

printf("step 3: allocate host memory A \n");
A = (double *)malloc (sizeof(double)*lda*N);
B = (double *)malloc (sizeof(double)*ldb*1);
X = (double *)malloc (sizeof(double)*ldb*1);
IPIV = (int *)malloc (sizeof(int)*N);
assert( NULL != A );
assert( NULL != B );
assert( NULL != X );
assert( NULL != IPIV );

/* permute 1D Laplacian to enable pivoting */
printf("step 4: prepare permuted 1D Laplacian for A and B = ones(N,1) \n");
gen_ld_laplacian_perm(
    N,
    &A[ IDX2F( IA, JA, lda ) ],
    lda
);
#ifdef SHOW_FORMAT
    print_matrix( N, N, A, lda, "A");
#endif
/* B = ones(N,1) */
for(int row = 1 ; row <= N ; row++){
    B[IDX1F(row)] = 1.0;
}
printf("step 5: create matrix descriptors for A and B \n");
status = cusolverMgCreateDeviceGrid(&gridA, 1, nbGpus, deviceList,
mapping );
assert(CUSOLVER_STATUS_SUCCESS == status);
status = cusolverMgCreateDeviceGrid(&gridB, 1, nbGpus, deviceList,
mapping );
assert(CUSOLVER_STATUS_SUCCESS == status);
/* (global) A is N-by-N */
status = cusolverMgCreateMatrixDesc(
    &descrA,
    N, /* number of rows of (global) A */
    N, /* number of columns of (global) A */
    N, /* number of rows in a tile */
    T_A, /* number of columns in a tile */
    CUDA_R_64F,
    gridA );
assert(CUSOLVER_STATUS_SUCCESS == status);

```

...

```

/* (global) B is N-by-1 */
status = cusolverMgCreateMatrixDesc(
    &descrB,
    N, /* number of rows of (global) B */
    1, /* number of columns of (global) B */
    N, /* number of rows in a tile */
    T_B, /* number of columns in a tile */
    CUDA_R_64F,
    gridB );
assert(CUSOLVER_STATUS_SUCCESS == status);

printf("step 6: allocate distributed matrices A, B and IPIV \n");
array_d_A = (double**)malloc(sizeof(double*)*nbGpus);
assert(NULL != array_d_A);
array_d_B = (double**)malloc(sizeof(double*)*nbGpus);
assert(NULL != array_d_B);
array_d_IPIV = (int**)malloc(sizeof(int*)*nbGpus);
assert(NULL != array_d_IPIV);

/* A := 0 */
createMat<double>(
    nbGpus,
    deviceList,
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    lda, /* leading dimension of local A */
    array_d_A
);
/* B := 0 */
createMat<double>(
    nbGpus,
    deviceList,
    1, /* number of columns of global B */
    T_B, /* number of columns per column tile */
    ldb, /* leading dimension of local B */
    array_d_B
);
/* IPIV := 0, IPIV is consistent with A */
createMat<int>(
    nbGpus,
    deviceList,
    N, /* number of columns of global IPIV */
    T_A, /* number of columns per column tile */
    1, /* leading dimension of local IPIV */
    array_d_IPIV
);
printf("step 7: prepare data on devices \n");
/* distribute A to array_d_A */
memcpyH2D<double>(
    nbGpus,
    deviceList,
    N,
    N,
/* input */
    A,
    lda,
/* output */
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    lda, /* leading dimension of local A */
    array_d_A, /* host pointer array of dimension nbGpus */
    IA,
    JA
);

```

...

```

/* distribute B to array_d_B */
memcpyH2D<double>(
    nbGpus,
    deviceList,
    N,
    1,
/* input */
    B,
    ldb,
/* output */
    1, /* number of columns of global B */
    T_B, /* number of columns per column tile */
    ldb, /* leading dimension of local B */
    array_d_B, /* host pointer array of dimension nbGpus */
    IB,
    JB
);

printf("step 8: allocate workspace space \n");
status = cusolverMgGetrf_bufferSize(
    handle,
    N,
    N,
    (void**)array_d_A,
    IA, /* base-1 */
    JA, /* base-1 */
    descrA,
    array_d_IPIV,
    CUDA_R_64F,
    &lwork_getrf);
assert(CUSOLVER_STATUS_SUCCESS == status);

status = cusolverMgGetrs_bufferSize(
    handle,
    CUBLAS_OP_N,
    N,
    1, /* NRHS */
    (void**)array_d_A,
    IA,
    JA,
    descrA,
    array_d_IPIV,
    (void**)array_d_B,
    IB,
    JB,
    descrB,
    CUDA_R_64F,
    &lwork_getrs);
assert(CUSOLVER_STATUS_SUCCESS == status);

lwork = (lwork_getrf > lwork_getrs)? lwork_getrf : lwork_getrs;
printf("\tallocate device workspace, lwork = %lld \n", (long long)lwork);
array_d_work = (double**)malloc(sizeof(double)*nbGpus);
assert( NULL != array_d_work);

/* array_d_work[j] points to device workspace of device j */
workspaceAlloc(
    nbGpus,
    deviceList,
    sizeof(double)*lwork, /* number of bytes per device */
    (void**)array_d_work
);
cudaStat = cudaDeviceSynchronize(); /* sync all devices */
assert(cudaSuccess == cudaStat);

```

...

```

printf("step 9: solve A*X = B by GETRF and GETRS \n");
status = cusolverMgGetrf(
    handle,
    N,
    N,
    (void**)array_d_A,
    IA,
    JA,
    descrA,
    array_d_IPIV,
    CUDA_R_64F,
    (void**)array_d_work,
    lwork,
    &info /* host */
);
assert(CUSOLVER_STATUS_SUCCESS == status);
cudaStat = cudaDeviceSynchronize(); /* sync all devices */
assert(cudaSuccess == cudaStat);
assert(0 == info); /* check if A is singular */

status = cusolverMgGetrs(
    handle,
    CUBLAS_OP_N,
    N,
    1, /* NRHS */
    (void**)array_d_A,
    IA,
    JA,
    descrA,
    array_d_IPIV,
    (void**)array_d_B,
    IB,
    JB,
    descrB,
    CUDA_R_64F,
    (void**)array_d_work,
    lwork,
    &info /* host */
);

assert(CUSOLVER_STATUS_SUCCESS == status);
cudaStat = cudaDeviceSynchronize(); /* sync all devices */
assert(cudaSuccess == cudaStat);
assert(0 == info); /* check if parameters are valid */

printf("step 10: retrieve IPIV and solution vector X \n");
memcpyD2H<double>(
    nbGpus,
    deviceList,
    N,
    1,
/* input */
    1, /* number of columns of global B */
    T_B, /* number of columns per column tile */
    ldb, /* leading dimension of local B */
    array_d_B,
    IB,
    JB,
/* output */
    X, /* N-by-1 */
    ldb
);

```

...

```

/* IPIV is consistent with A, use JA and T_A */
memcpyD2H<int>(
    nbGpus,
    deviceList,
    1,
    N,
/* input */
    N, /* number of columns of global IPIV */
    T_A, /* number of columns per column tile */
    1, /* leading dimension of local IPIV */
    array_d_IPIV,
    1,
    JA,
/* output */
    IPIV, /* 1-by-N */
    1
);

#ifdef SHOW_FORMAT
/* X is N-by-1 */
print_matrix(N, 1, X, ldb, "X");
#endif

#ifdef SHOW_FORMAT
/* IPIV is 1-by-N */
printf("IPIV = matlab base-1, 1-by-%d matrix\n", N);
for(int row = 1; row <= N ; row++){
    printf("IPIV(%d) = %d \n", row, IPIV[ IDX1F(row) ]);
}
#endif

printf("step 11: measure residual error |b - A*x| \n");
double max_err = 0;
for(int row = 1; row <= N ; row++){
    double sum = 0.0;
    for(int col = 1; col <= N ; col++){
        double Aij = A[ IDX2F( row, col, lda ) ];
        double xj = X[ IDX1F(col) ];
        sum += Aij*xj;
    }
    double bi = B[ IDX1F(row) ];
    double err = fabs( bi - sum );

    max_err = ( max_err > err )? max_err : err;
}
double x_nrm_inf = vec_nrm_inf(N, X);
double b_nrm_inf = vec_nrm_inf(N, B);
double A_nrm_inf = 4.0;
double rel_err = max_err / (A_nrm_inf * x_nrm_inf + b_nrm_inf);
printf("\n|b - A*x|_inf = %E\n", max_err);
printf("|x|_inf = %E\n", x_nrm_inf);
printf("|b|_inf = %E\n", b_nrm_inf);
printf("|A|_inf = %E\n", A_nrm_inf);
/* relative error is around machine zero */
/* the user can use |b - A*x| / (N*|A|*|x|+|b|) as well */
printf("|b - A*x| / (|A|*|x|+|b|) = %E\n\n", rel_err);

```

...

```

printf("step 12: free resources \n");
destroyMat(
    nbGpus,
    deviceList,
    N, /* number of columns of global A */
    T_A, /* number of columns per column tile */
    (void**)array_d_A );
destroyMat(
    nbGpus,
    deviceList,
    1, /* number of columns of global B */
    T_B, /* number of columns per column tile */
    (void**)array_d_B );
destroyMat(
    nbGpus,
    deviceList,
    N, /* number of columns of global IPIV */
    T_A, /* number of columns per column tile */
    (void**)array_d_IPIV );

workspaceFree( nbGpus, deviceList, (void**)array_d_work );

if (NULL != A) free(A);
if (NULL != B) free(B);
if (NULL != X) free(X);
if (NULL != IPIV) free(IPIV);

if (NULL != array_d_A ) free(array_d_A);
if (NULL != array_d_B ) free(array_d_B);
if (NULL != array_d_IPIV) free(array_d_IPIV);
if (NULL != array_d_work) free(array_d_work);

return 0;
}

```

Appendix J.

ACKNOWLEDGEMENTS

NVIDIA would like to thank the following individuals and institutions for their contributions:

- ▶ CPU LAPACK routines from netlib, CLAPACK-3.2.1 (<http://www.netlib.org/clapack/>)

The following is license of CLAPACK-3.2.1.

Copyright (c) 1992-2008 The University of Tennessee. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer listed in this license in the documentation and/or other materials provided with the distribution.
- Neither the name of the copyright holders nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- ▶ METIS-5.1.0 (<http://glaros.dtc.umn.edu/gkhome/metis/metis/overview>)

The following is license of METIS (Apache 2.0 license).

Copyright 1995-2013, Regents of the University of Minnesota

Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at

<http://www.apache.org/licenses/LICENSE-2.0>

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- ▶ QD (A C++/fortran-90 double-double and quad-double package) (<http://crd-legacy.lbl.gov/~dhbailey/mpdist/>)

The following is license of QD (modified BSD license).

Copyright (c) 2003-2009, The Regents of the University of California, through Lawrence Berkeley National Laboratory (subject to receipt of any required approvals from U.S. Dept. of Energy) All rights reserved.

1. Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

(1) Redistributions of source code must retain the copyright notice, this list of conditions and the following disclaimer.

(2) Redistributions in binary form must reproduce the copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

(3) Neither the name of the University of California, Lawrence Berkeley National Laboratory, U.S. Dept. of Energy nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

2. THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

3. You are under no obligation whatsoever to provide any bug fixes, patches, or upgrades to the features, functionality or performance of the source code ("Enhancements") to anyone; however, if you choose to make your Enhancements

available either publicly, or directly to Lawrence Berkeley National Laboratory, without imposing a separate written license agreement for such Enhancements, then you hereby grant the following license: a non-exclusive, royalty-free perpetual license to install, use, modify, prepare derivative works, incorporate into other computer software, distribute, and sublicense such enhancements or derivative works thereof, in binary and source code form.

Appendix K.

BIBLIOGRAPHY

- [1] Timothy A. Davis, Direct Methods for sparse Linear Systems, siam 2006.
- [2] E. Chuthill and J. McKee, reducing the bandwidth of sparse symmetric matrices, ACM '69 Proceedings of the 1969 24th national conference, Pages 157-172.
- [3] Alan George, Joseph W. H. Liu, An Implementation of a Pseudoperipheral Node Finder, ACM Transactions on Mathematical Software (TOMS) Volume 5 Issue 3, Sept. 1979 Pages 284-295.
- [4] J. R. Gilbert and T. Peierls, Sparse partial pivoting in time proportional to arithmetic operations, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 862-874.
- [5] Alan George and Esmond Ng, An Implementation of Gaussian Elimination with Partial Pivoting for Sparse Systems, SIAM J. Sci. and Stat. Comput., 6(2), 390-409.
- [6] Alan George and Esmond Ng, Symbolic Factorization for Sparse Gaussian Elimination with Partial Pivoting, SIAM J. Sci. and Stat. Comput., 8(6), 877-898.
- [7] John R. Gilbert, Xiaoye S. Li, Esmond G. Ng, Barry W. Peyton, Computing Row and Column Counts for Sparse QR and LU Factorization, BIT 2001, Vol. 41, No. 4, pp. 693-711.
- [8] Patrick R. Amestoy, Timothy A. Davis, Iain S. Duff, An Approximate Minimum Degree Ordering Algorithm, SIAM J. Matrix Analysis Applic. Vol 17, no 4, pp. 886-905, Dec. 1996.
- [9] Alan George, Joseph W. Liu, A Fast Implementation of the Minimum Degree Algorithm Using Quotient Graphs, ACM Transactions on Mathematical Software, Vol 6, No. 3, September 1980, page 337-358.
- [10] Alan George, Joseph W. Liu, Computer Solution of Large Sparse Positive Definite Systems, Englewood Cliffs, New Jersey: Prentice-Hall, 1981.
- [11] Iain S. Duff, ALGORITHM 575 Permutations for a Zero-Free Diagonal, ACM Transactions on Mathematical Software, Vol 7, No 3, September 1981, Page 387-390
- [12] Iain S. Duff and Jacko Koster, On algorithms for permuting large entries to the diagonal of a sparse matrix, SIAM Journal on Matrix Analysis and Applications, 2001, Vol. 22, No. 4 : pp. 973-996

- [13] "A Fast and Highly Quality Multilevel Scheme for Partitioning Irregular Graphs". George Karypis and Vipin Kumar. SIAM Journal on Scientific Computing, Vol. 20, No. 1, pp. 359-392, 1999.

Notice

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication of otherwise under any patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all other information previously supplied. NVIDIA Corporation products are not authorized as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

Trademarks

NVIDIA and the NVIDIA logo are trademarks or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2014-2021 NVIDIA Corporation. All rights reserved.