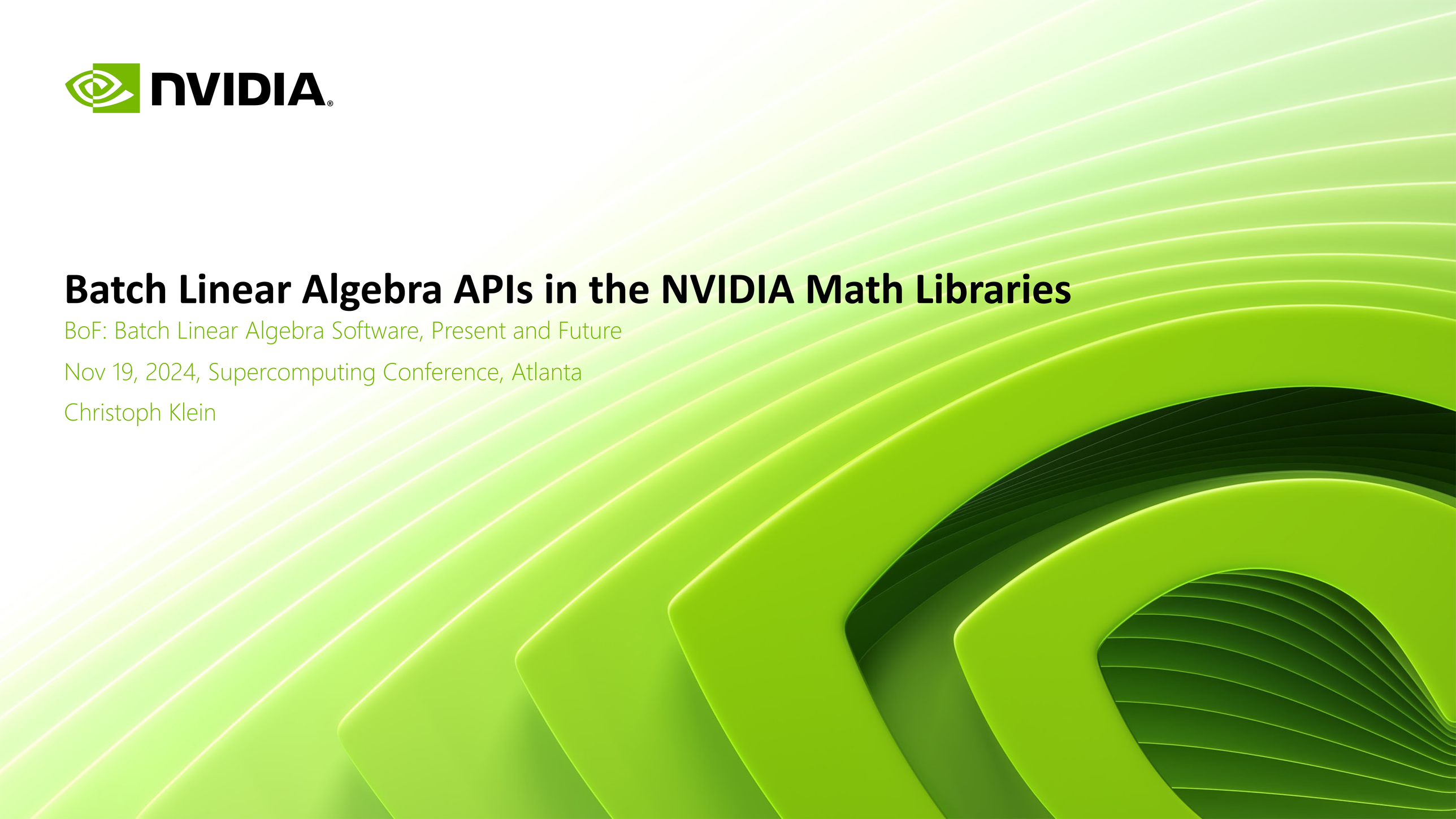


Batch Linear Algebra APIs in the NVIDIA Math Libraries

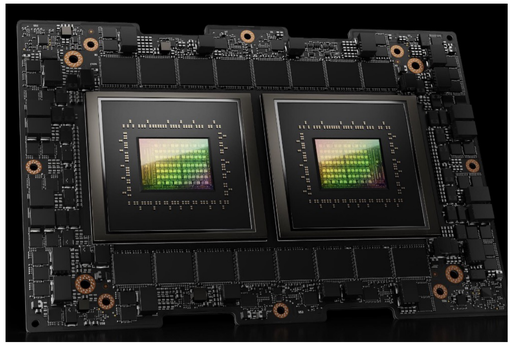
BoF: Batch Linear Algebra Software, Present and Future

Nov 19, 2024, Supercomputing Conference, Atlanta

Christoph Klein

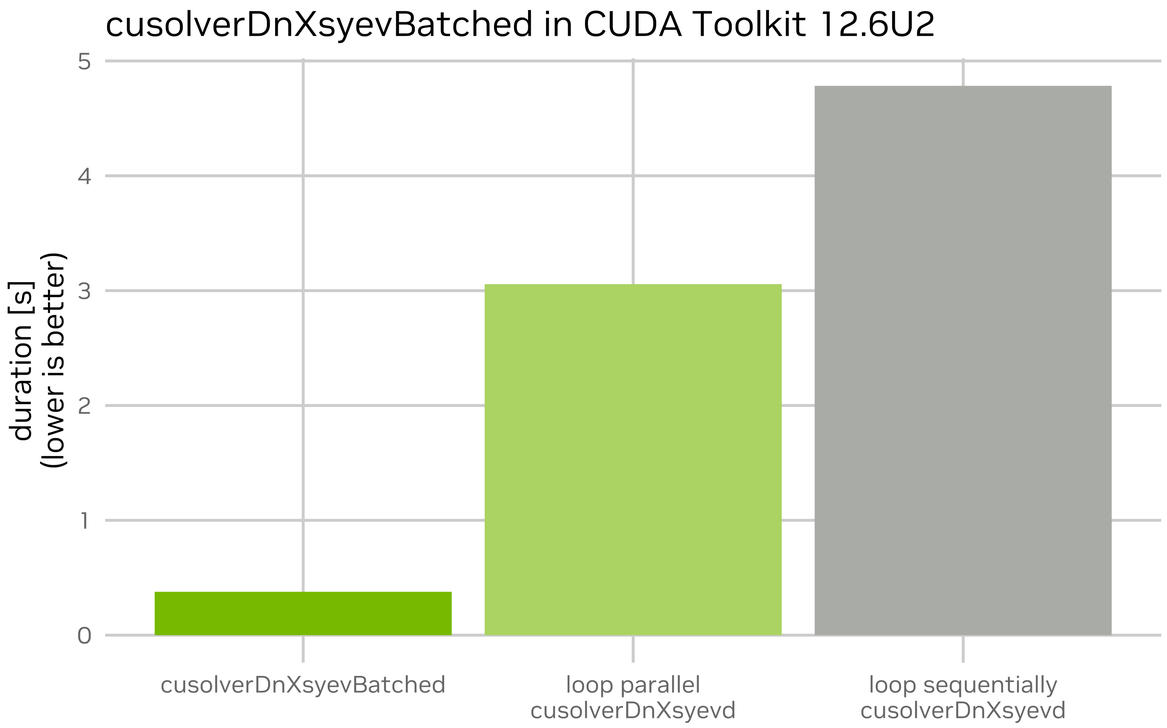
Overview

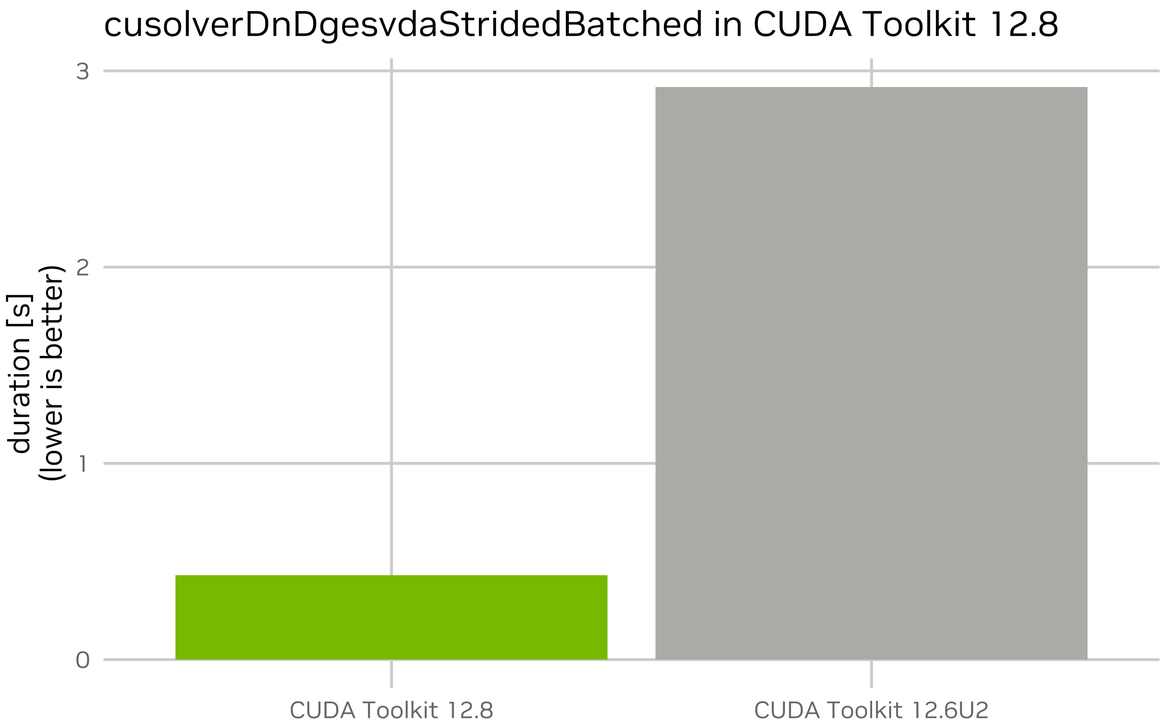
Batch Linear Algebra APIs in the NVIDIA Math Libraries

cuSOLVER	cuBLAS
POTRF/POTRS SVDJ (n<=32, m<=32) SVDA (performance improvements in CUDA Toolkit 12.8) SYEVJ SYEV (added in CUDA Toolkit 12.6U2) https://docs.nvidia.com/cuda/cusolver/index.html	GEMM (+grouped) GEMV TRSM GETRF/GETRS GEQRF/GELS GETRI https://docs.nvidia.com/cuda/cublas/
cuDSS	**NVPL BLAS**
Variably batched support will be available for sparse direct solver in upcoming 0.4.0 release • https://developer.nvidia.com/cudss	• GEMM  https://docs.nvidia.com/nvpl/_static/blas/index.html

cuSOLVER Batched Highlights

H200 double precision benchmarks on 1024 square matrices of size 512



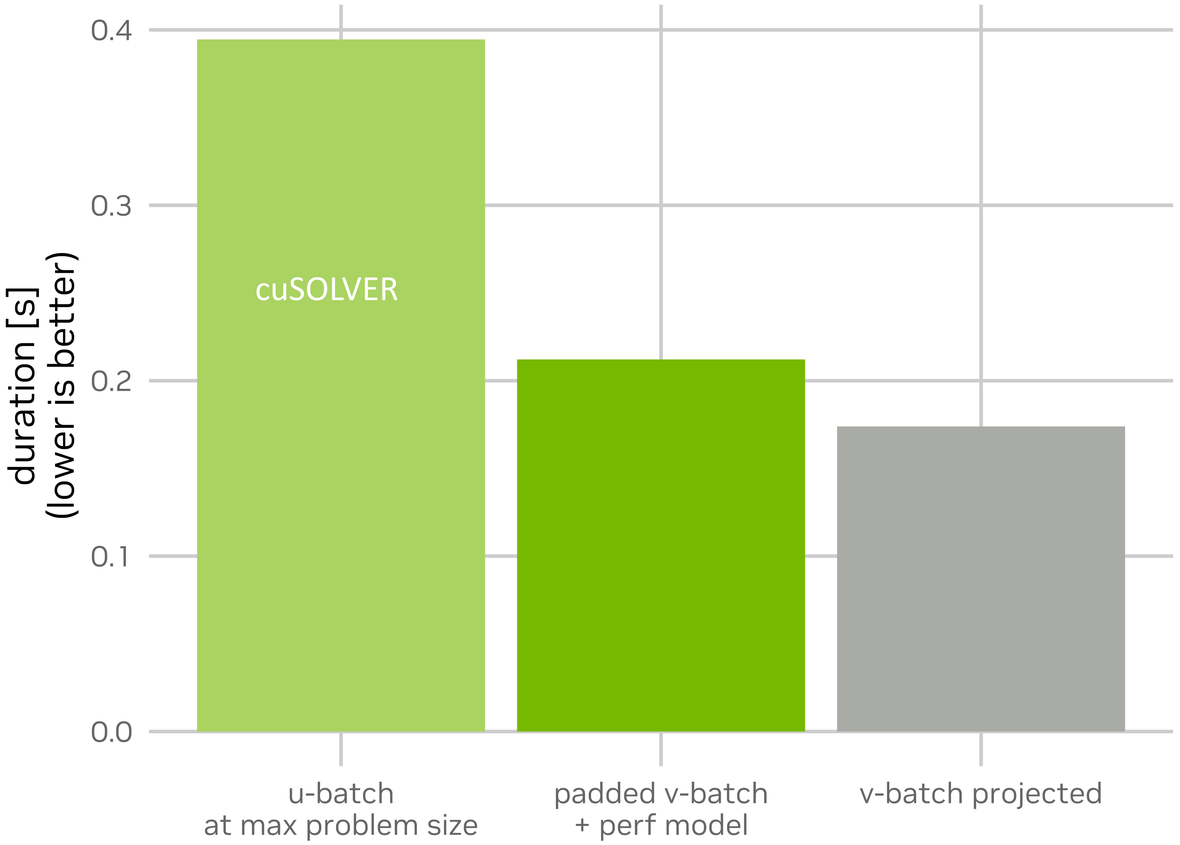


Outlook to variably batched approaches on GPUs

Variably batched support in cuDSS 0.4.0

```
cudssStatus_t
cudssMatrixCreateBatchCsr(cudssMatrix_t*       matrix,
                          int64_t              batchCount,
                          void*                nrows,
                          void*                ncols,
                          void*                nnz,
                          void**               rowStart,
                          void**               rowEnd,
                          void**               colIndices,
                          void**               values,
                          cudaDataType_t       indexType,
                          cudaDataType_t       valueType,
                          cudaMatrixType_t     mtype,
                          cudaMatrixViewType_t mview,
                          cudssIndexBase_t     indexBase)
```

Proof of concept: Uniform batch at maximum problem size VS variably batched eigenvalue solver

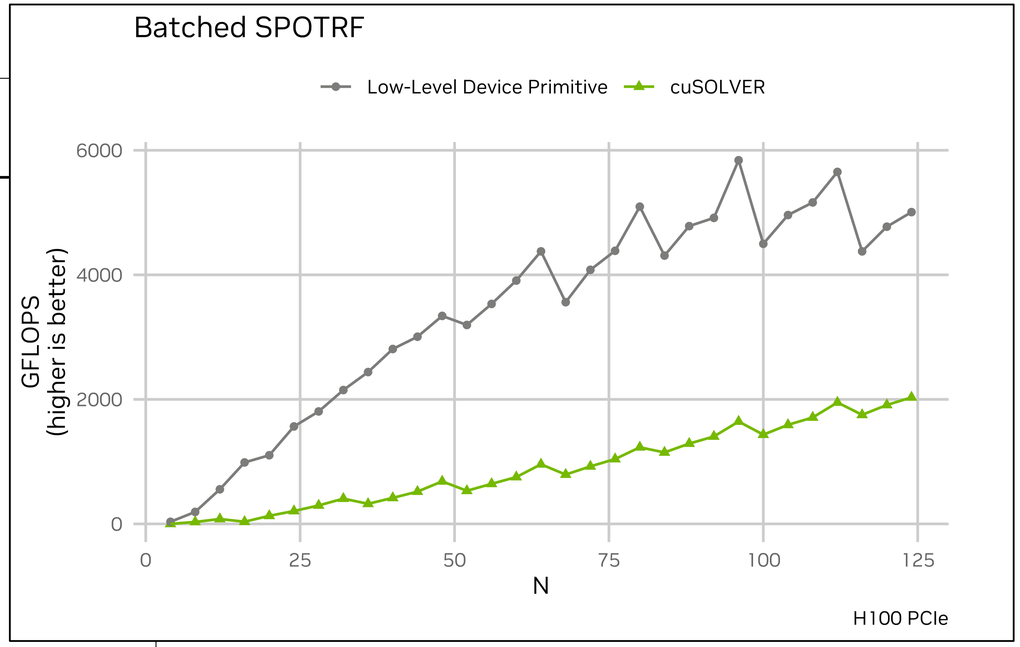


uniform matrix size distribution between 1 and 512

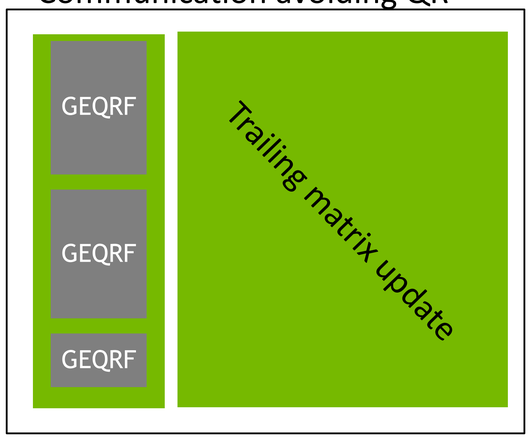
Redesigning panel factorizations on GPUs

Unifying batched and non-batched software stack

Current	Unified Design
Non-Batched APIs *latency optimized*	Non-Batched APIs
Batched APIs *throughput optimized*	Batched APIs
	Low-Level Primitives



Communication avoiding QR



Batched QR in panel

NVIDIA