May 18th - 19th, 2016

Workshop on Batched, Reproducible, and Reduced Precision BLAS

Innovative Computing Laboratory
University of Tennessee
Knoxville, TN

Example of Cholesky's Efficient Implementations

J. Kurzak
P. Luszczek
M. Gates
H. Anzt

Scope

batched spotrf

precision

- IEEE double precision
- IEEE single precision
- relaxed single precision

← *machine learning*

size

- large (up to 500)
- small (up to 100)
- ultra small (~20)
- variable size
- fixed size

technology

- composite kernel
- monolithic kernel → *autotuning*

Motivation
machine learning

doubts

- Are they really so small?
- Is there really so many?
- Are they fixed size?

Alternating Least Squares

- Apache Mahout
- Spark MLlib
- GraphLab (Dato)
- Intel DAAL

Netflix Prize

- batch size: 17,000 and 500,000
- matrix size: 10 – 100
- uniform batch

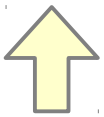
Techniques
kernel development

coding

- C++ templates (parametrization)
- #pragma unroll (low level unrolling)
- pyexpander (high level unrolling)

algorithmic

- LAPACK-style blocking
- PLASMA-style tiling
- lazy evaluation (left / top looking)



Texture Reads

- texture objects
- __ldg() intrinsic

Vector Types

- double2, float4, ...

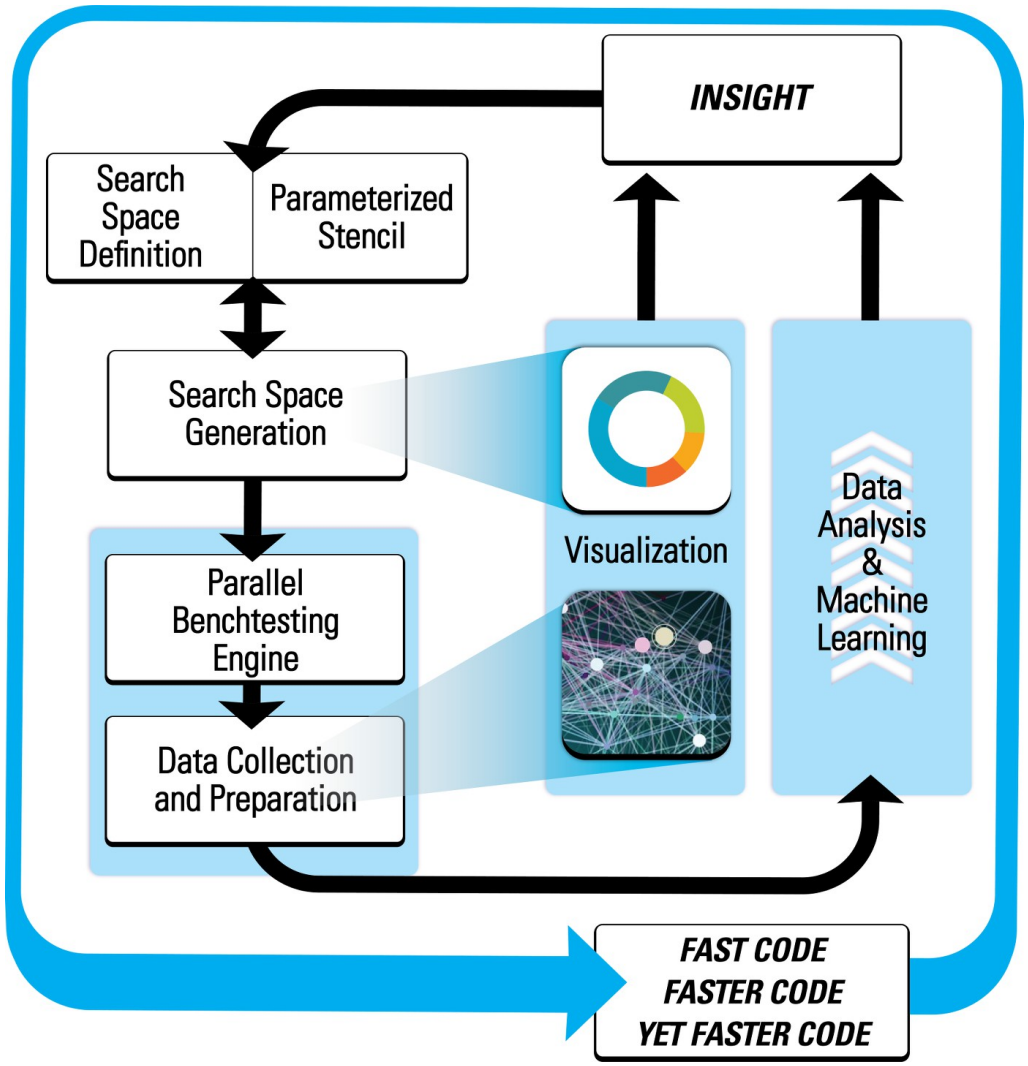
Tools

- nvprof
- nvdisasm

basically techniques for optimizing serial performance for memory efficiency and ILP

BEAST / BONSAI

autotuning

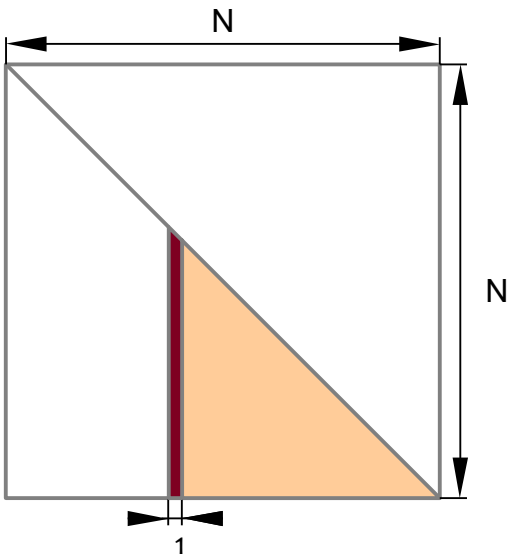


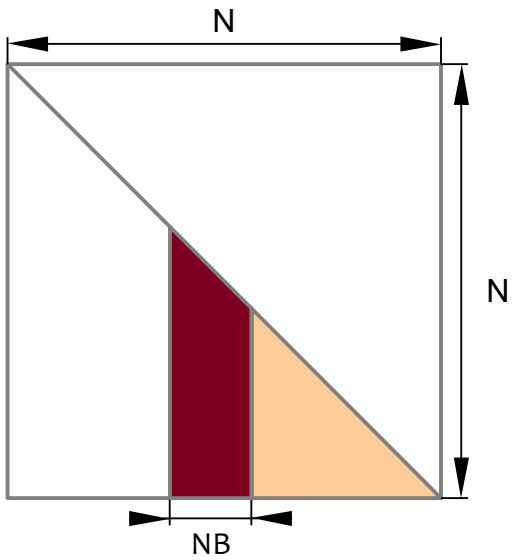
Thank you for your letters of collaboration!

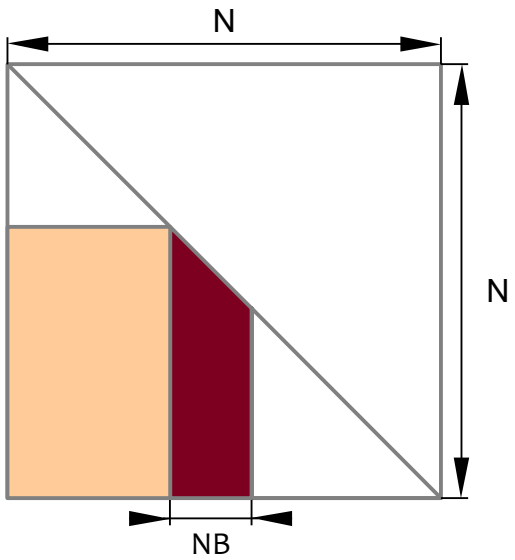
- NVIDIA
- Intel
- AMD

Cholesky

sportf







canonical

- BLAS 2
- memory bound

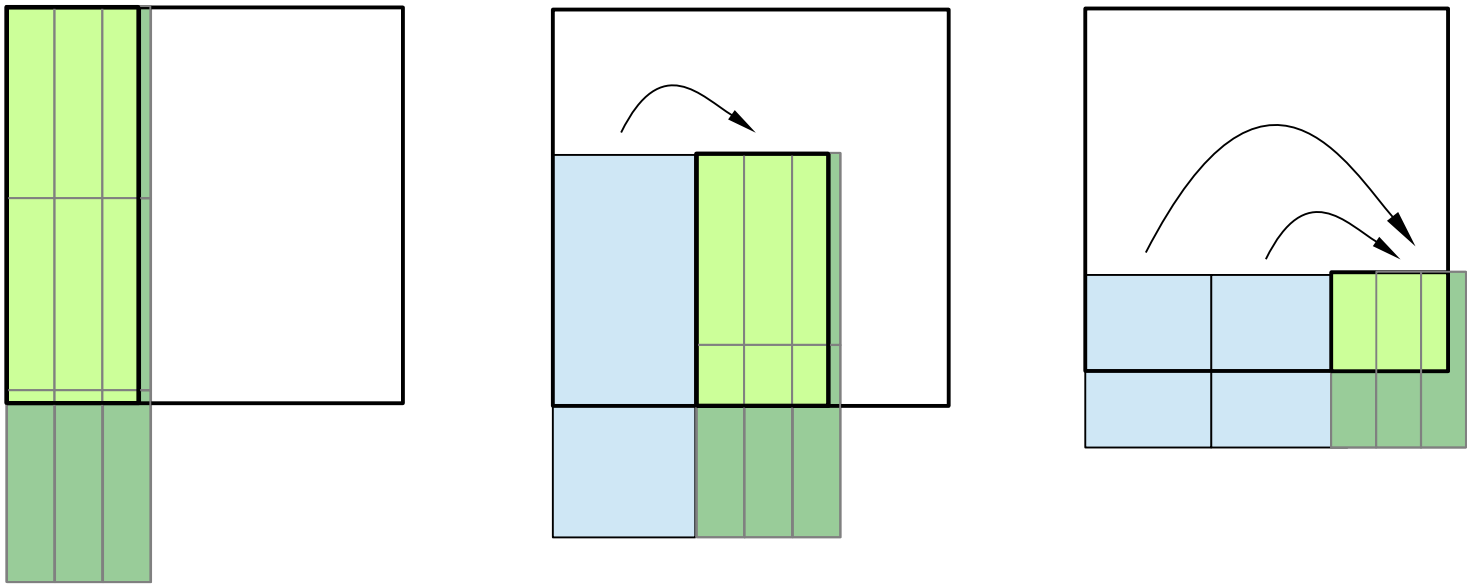
blocking (LAPACK)

- data locality
- register reuse
- surface to volume effect

lazy evaluation

- left-looking
- memory efficiency
- minimizing writes

Cholesky
spotrf

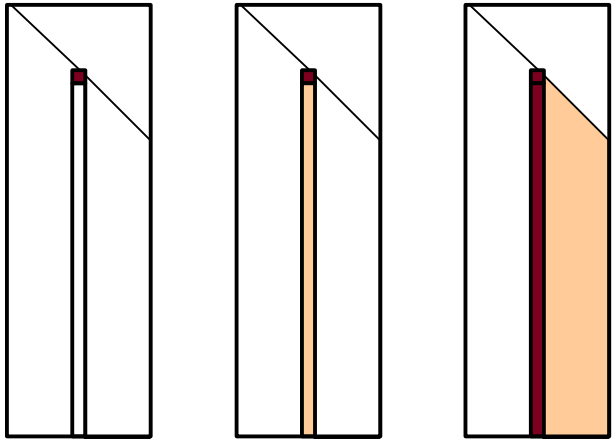


lazy evaluation / left-looking / “out of core”

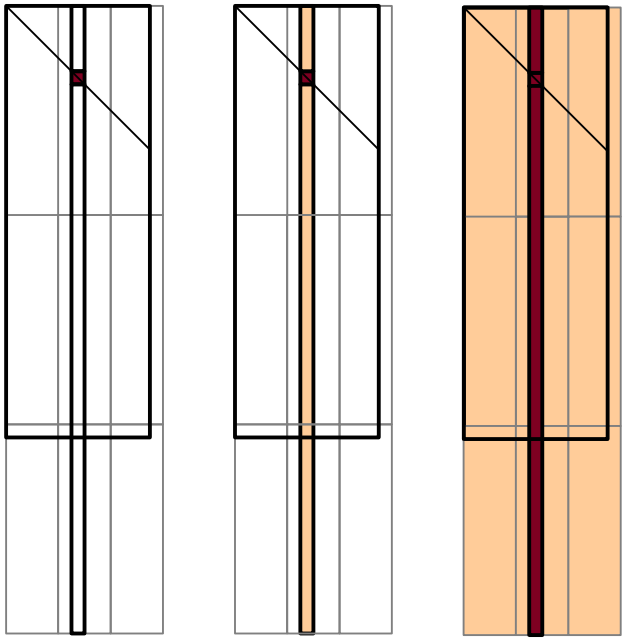
- maximizes data reuse
- minimizes writes

Cholesky

sportf panel



useful work

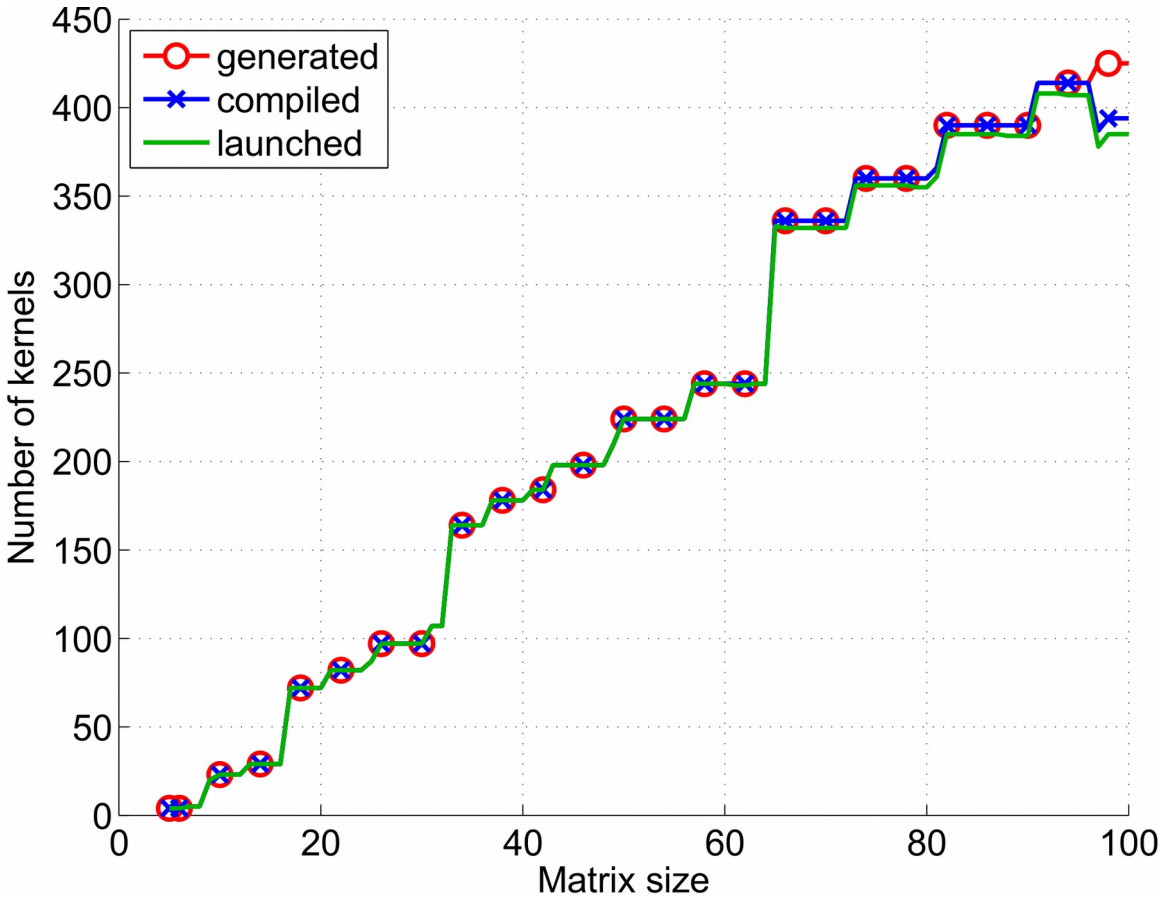


actual work

Use right-looking algorithms to maximize SIMT parallelism.

Do wasteful work, but minimum number of conditionals.

Cholesky
autotuning

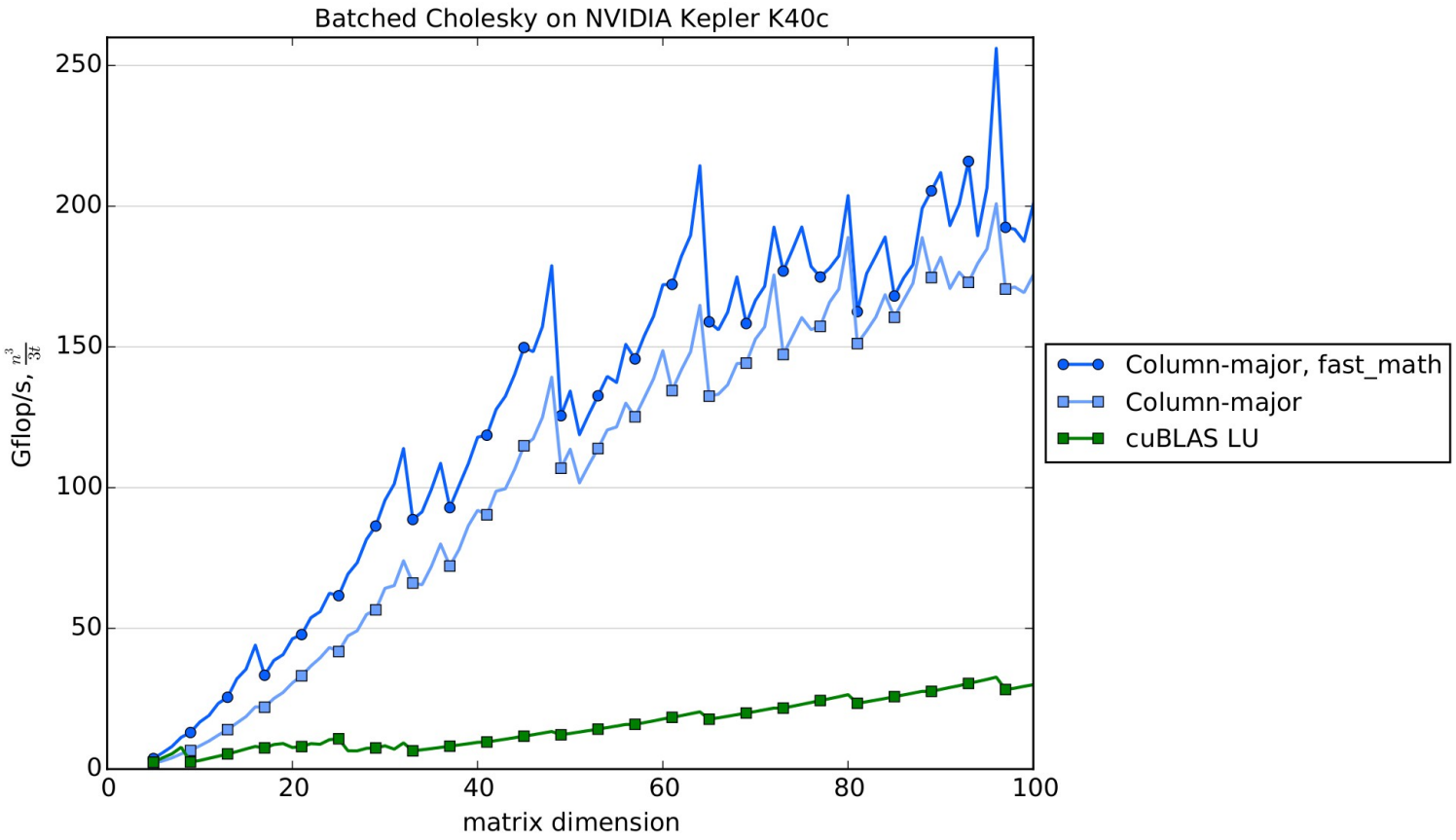


For each matrix size N tune:

- panel width (NB)
- thread block shape (blockDim.x, blockDim.y)
- *not an exhaustive sweep*

Kernels

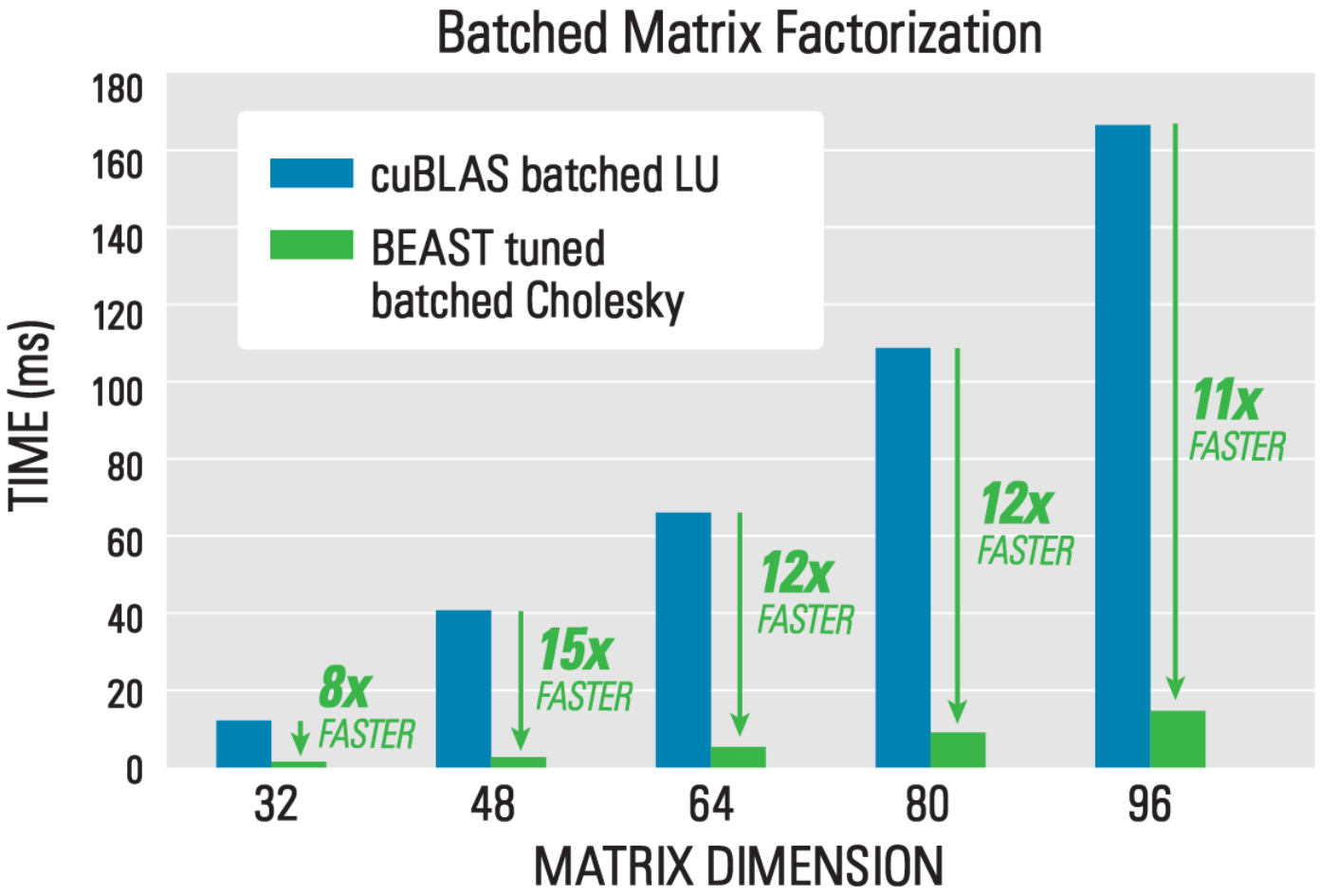
sposv_batched



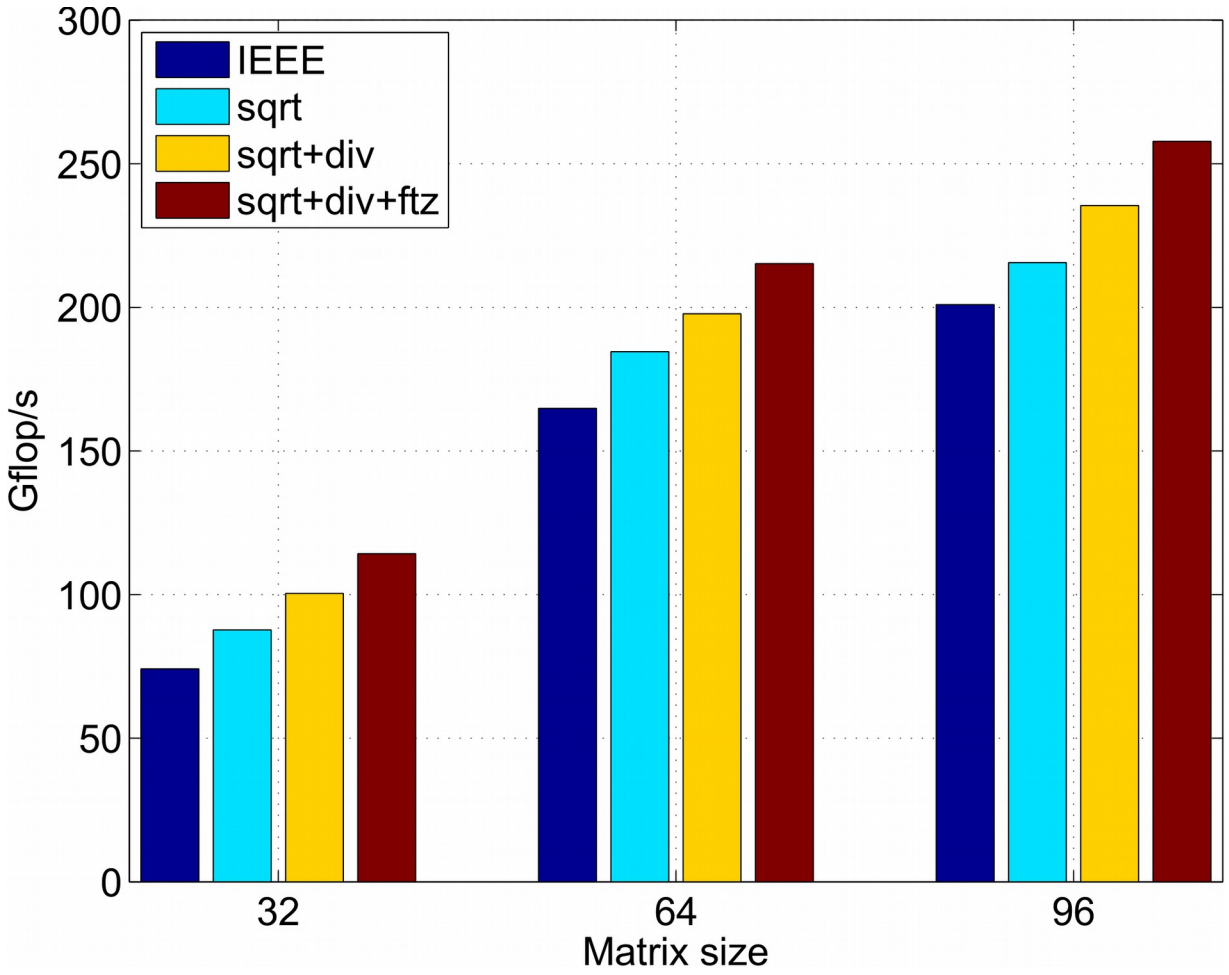
Implementation and Tuning of Batched Cholesky Factorization and Solve for NVIDIA GPUs
IEEE Transactions on Parallel and Distributed Systems
http://dx.doi.org/10.1109/TPDS.2015.2481890

Cholesky

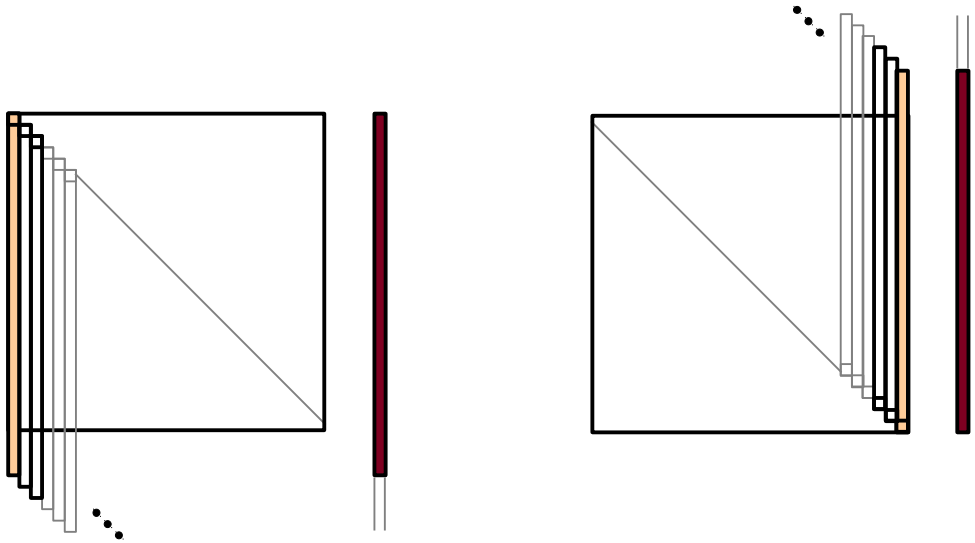
sportf



Cholesky
relaxing IEEE



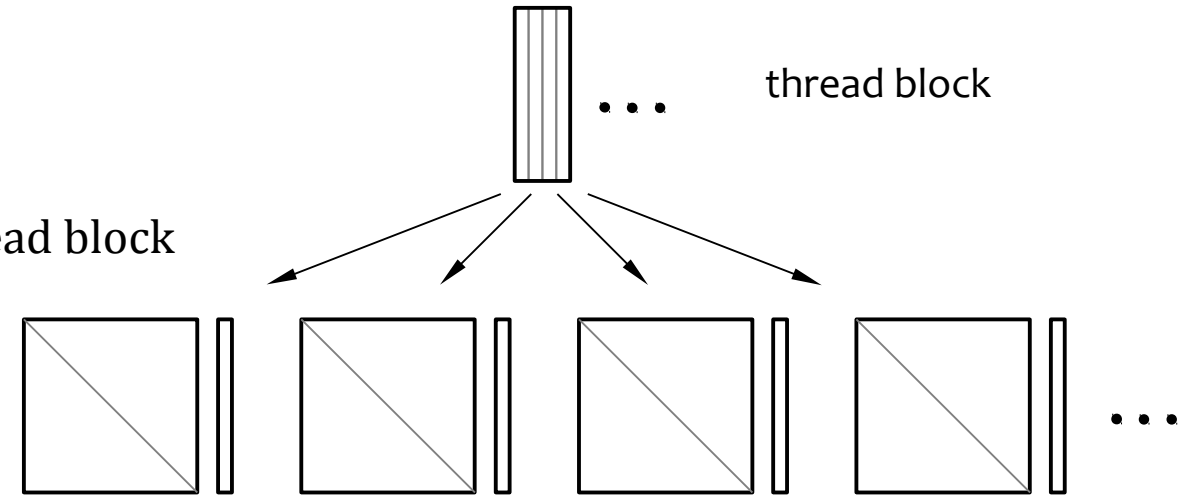
Cholesky
spotrs



Solve

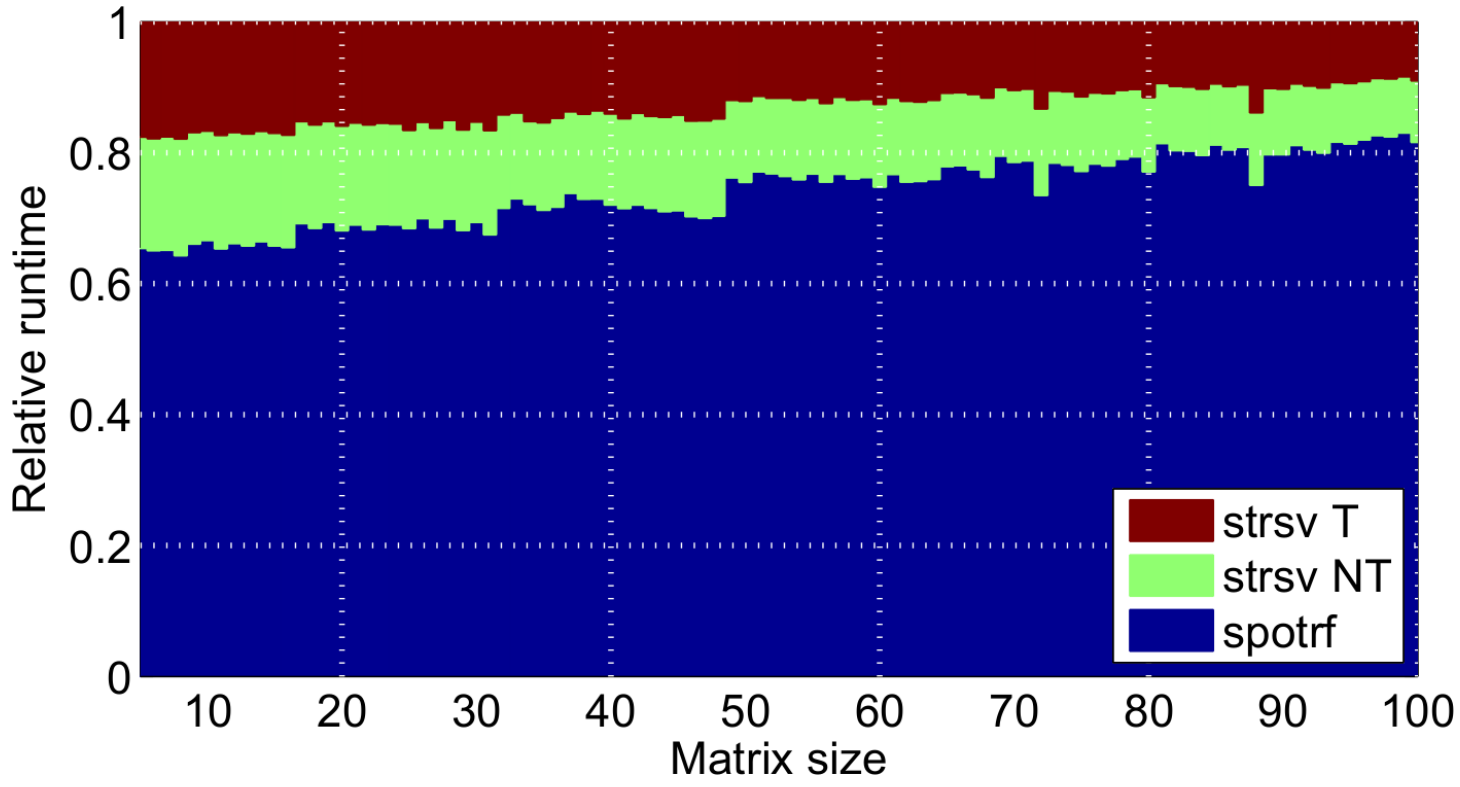
- single right hand side
- L in lower triangle
- L^T in upper triangle

- multiple solves in each thread block



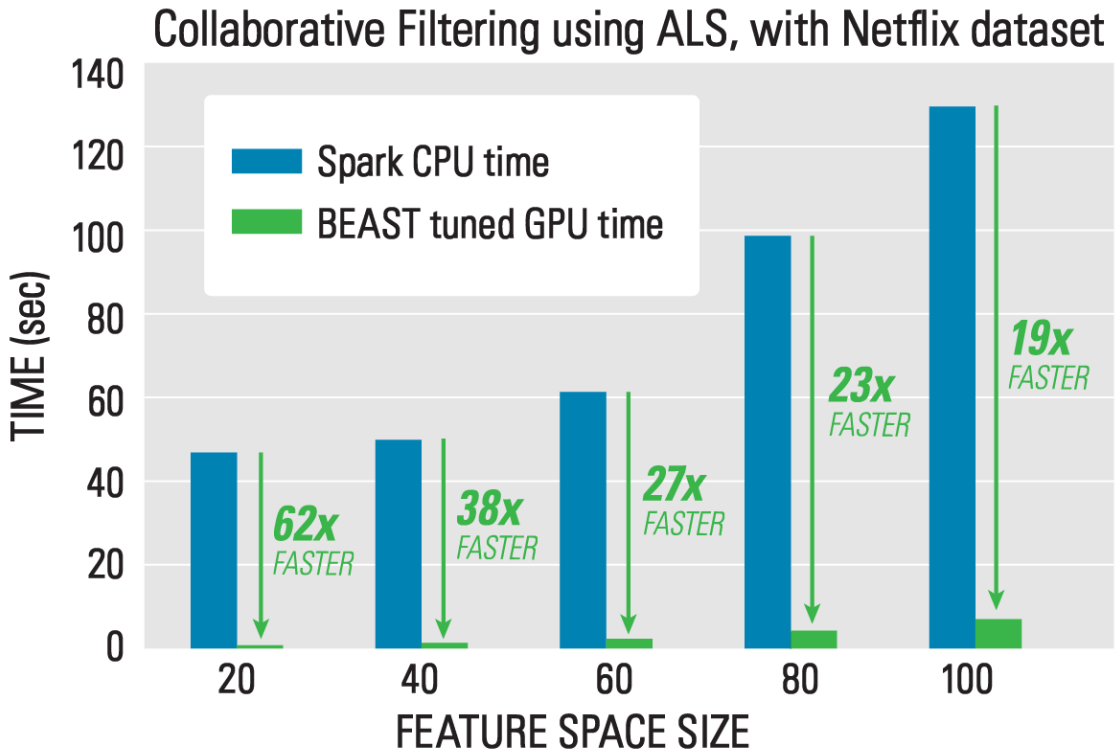
Cholesky

sposv



ALS

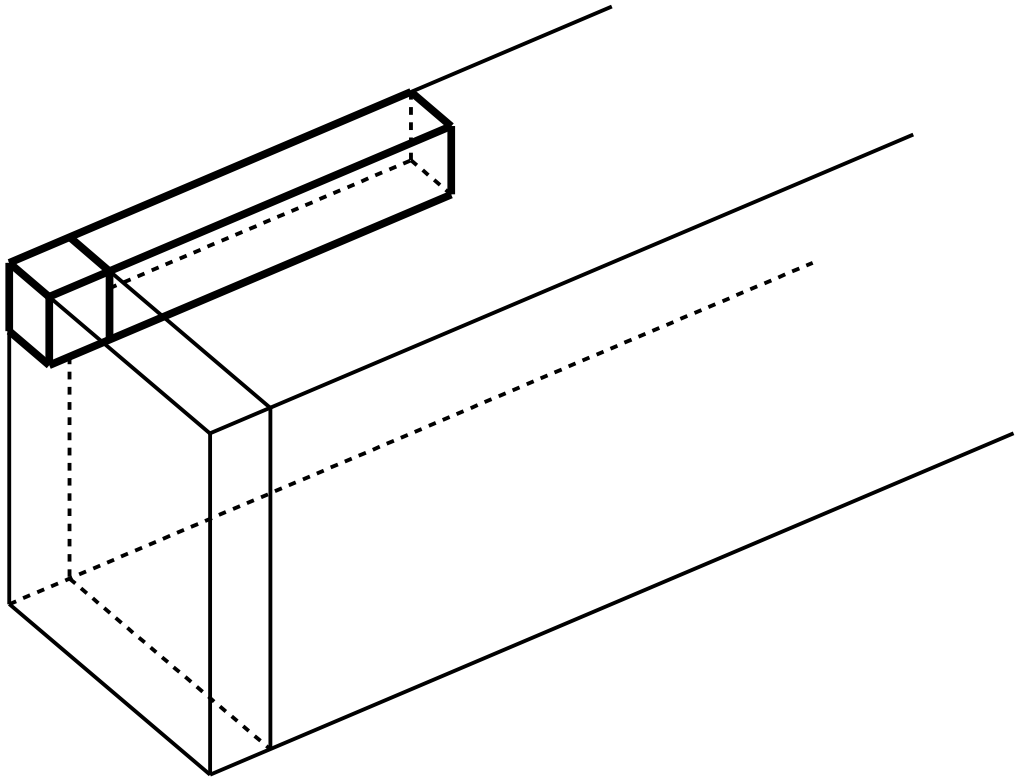
speedup over Spark



Accelerating Collaborative Filtering Using Concepts from High Performance Computing
2015 IEEE International Conference on Big Data
DOI: 10.1109/BigData.2015.7363811

Cholesky

one thread per matrix



Pros

- zero synchronization
- zero load imbalance
- no shared memory

Cons

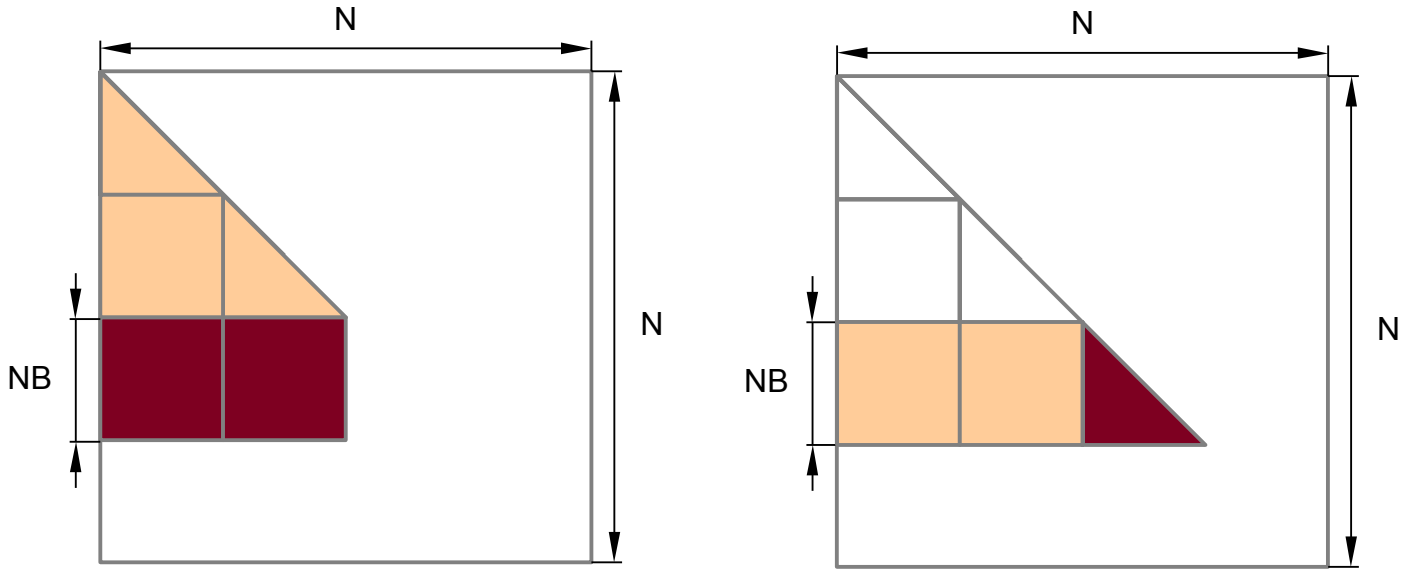
- no cache / shared memory reuse
- unthinkable on standard layout

  basically requires batch-major layout

- LAPACKE: column-major / row-major
- cuDNN: NCHW / NHWC

Cholesky

one thread per matrix



algorithmic

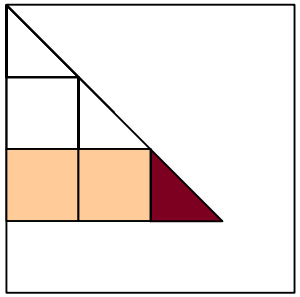
- PLASMA-style tiling
- the laziest evaluation (top-looking)

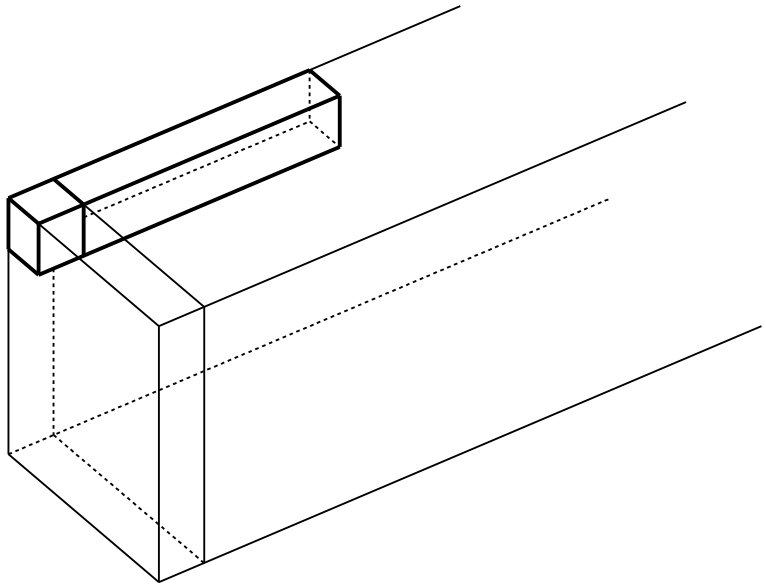
basically completely serial implementation from the standpoint of each thread

no parallelization or vectorization considerations

Cholesky

one thread per matrix

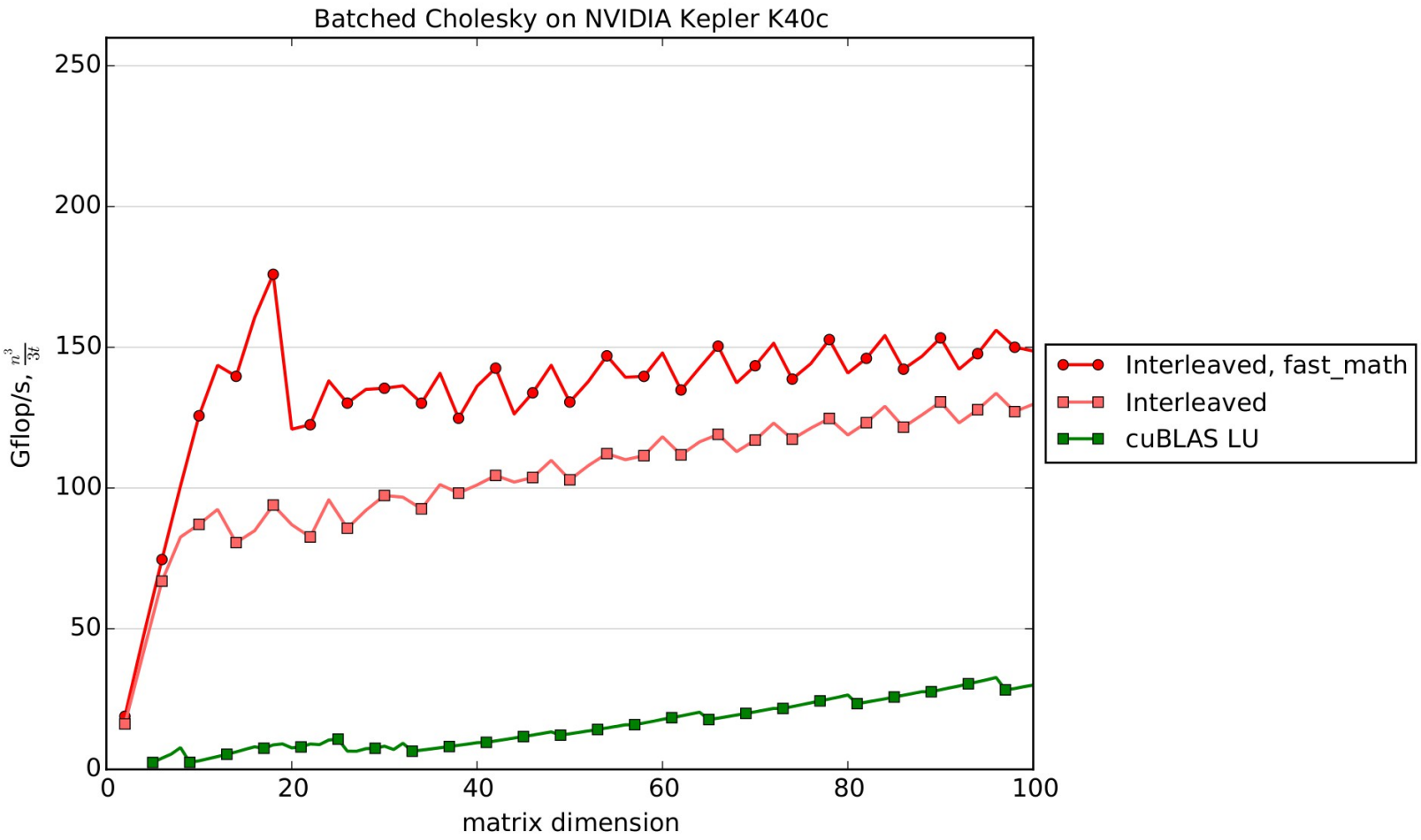




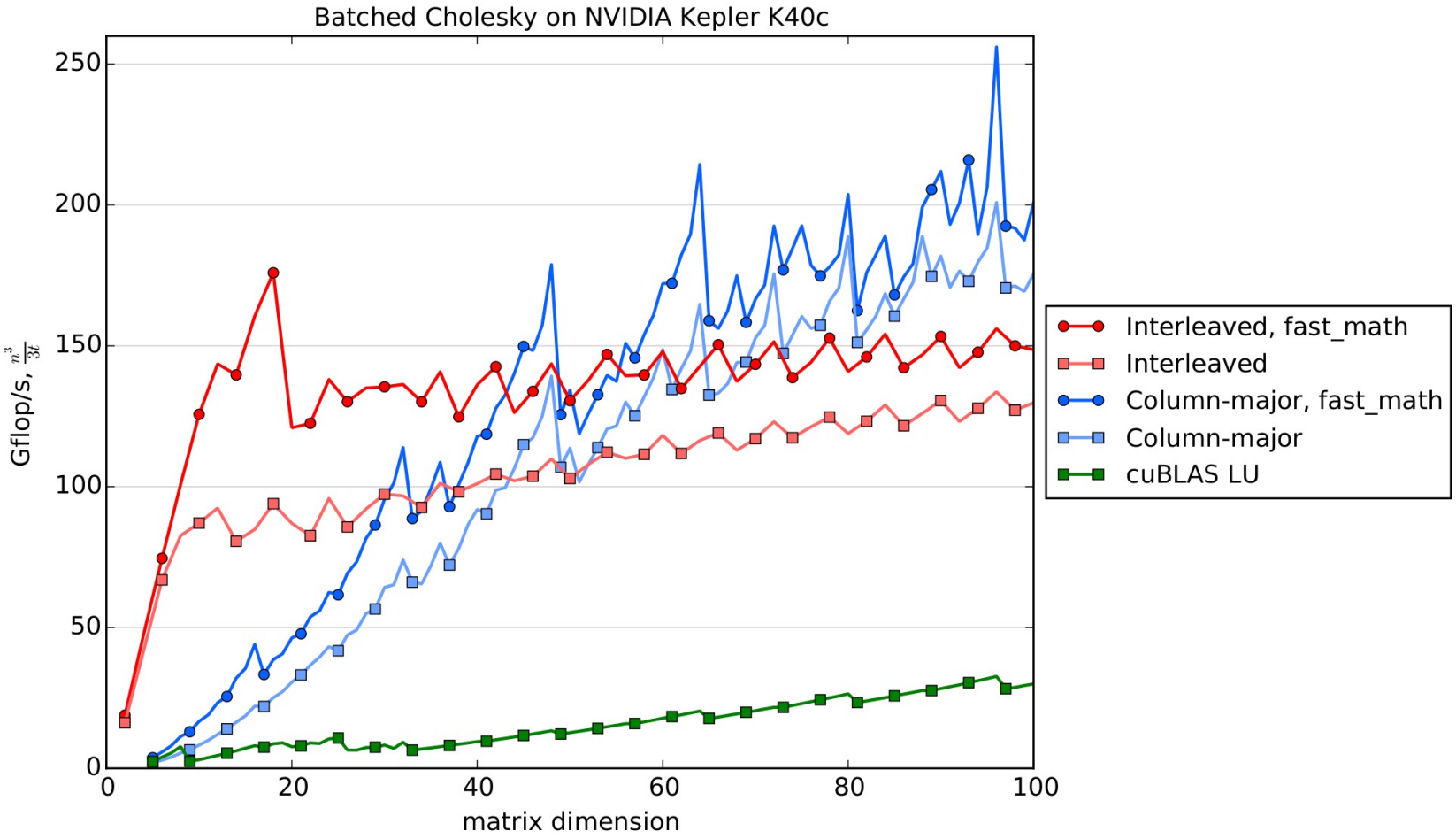
tuning parameters

- *right-looking, left-looking, top-looking*
- *thread block length (blockDim.x)*
- *tile size (NB)*
- *unrolling tile operations of the full factorization*

Cholesky
performance



Cholesky
performance



Conclusions

- For batched on GPUs you have to write specialized routines.
- We know how (common DLA wisdom applies).
- Autotuning works like a charm.

- For the most part on CPUs MKL+OpenMP gets the job done.

- Unorthodox layouts?
- Layout translation?
- On the fly?


ICL UT
INNOVATIVE
COMPUTING LABORATORY
THE UNIVERSITY of TENNESSEE