High Performance Linear Algebra

Hatem Ltaief
Senior Research Scientist
Extreme Computing Research Center
King Abdullah University of Science and Technology

4th International Workshop on Real-Time Control for Adaptive Optics

Observatoire de Paris
Dec 19-21 2016

Outline

Outline

Hardware Landscape: The TOP500 List

- Ranks the 500 fastest computers in the world (twice a year)
- High Performance LINPACK benchmark (HPL)
- Solves a dense system of linear equations $Ax = b$ using LU
- Compute intensive benchmark (Matrix-Matrix multiplication)
- Performance reported in Xflop/s: rate of execution, i.e., number of floating point operations per second

The Top500 List, November'16: http://www.top500.org

Rank	**Name**	**Vendor**	**Cores**	R_{max} **(Pflop/s)**	**Power (MW)**
1	TaihuLight	NRCPC Sunway SW26010	10,649,600	93	15,3
2	Tianhe-2	NUDT Intel Xeon E5	3,120,000	33,8	17,8
3	Titan	Cray AMD Opteron + NVIDIA K20x	560,640	17,6	8,21
4	Sequoia	IBM BG/Q	1,572,864	17,2	7,89
5	Cori	Cray Intel Xeon Phi	622,336	14,01	3,93
6	Oakforest	Fujitsu Intel Xeon Phi	556,104	13,55	2,72
7	K Computer	Fujitsu SPARC64	705,024	10,51	12,66
8	PizDaint	Cray Intel Xeon E5 + NVIDIA P100	206,720	9,78	1,31
9	Mira	IBM BG/Q	786,432	8,59	3,95
10	Trinity	Cray Intel Xeon E5	301,056	8,1	4,23

The Top500 List, November'16: http://www.top500.org

Rank	**Name**	**Vendor**	**Cores**	R_{max} **(Pflop/s)**	**Power (MW)**
1	TaihuLight	NRCPC Sunway SW26010	10,649,600	93	15,3
2	Tianhe-2	NUDT Intel Xeon E5	3,120,000	33,8	17,8
3	Titan	Cray AMD Opteron + NVIDIA K20x	560,640	17,6	8,21
4	Sequoia	IBM BG/Q	1,572,864	17,2	7,89
5	Cori	Cray Intel Xeon Phi	622,336	14,01	3,93
6	Oakforest	Fujitsu Intel Xeon Phi	556,104	13,55	2,72
7	K Computer	Fujitsu SPARC64	705,024	10,51	12,66
8	PizDaint	Cray Intel Xeon E5 + NVIDIA P100	206,720	9,78	1,31
9	Mira	IBM BG/Q	786,432	8,59	3,95
10	Trinity	Cray Intel Xeon E5	301,056	8,1	4,23
...	...	...	...	...	...
15	Shaheen 2	Cray Intel Xeon E5	196,608	5,53	2,83

The Top500 List, November'16: http://www.top500.org

Rank	Name	Vendor	Cores	R_{max} (Pflop/s)	Power (MW)
1	TaihuLight	NRCPC Sunway SW26010	10,649,600	93	15,3
2	Tianhe-2	NUDT Intel Xeon E5	3,120,000	33,8	17,8
3	Titan	Cray AMD Opteron + NVIDIA K20x	560,640	17,6	8,21
4	Sequoia	IBM BG/Q	1,572,864	17,2	7,89
5	Cori	Cray Intel Xeon Phi	622,336	14,01	3,93
6	Oakforest	Fujitsu Intel Xeon Phi	556,104	13,55	2,72
7	K Computer	Fujitsu SPARC64	705,024	10,51	12,66
8	PizDaint	Cray Intel Xeon E5 + NVIDIA P100	206,720	9,78	1,31
9	Mira	IBM BG/Q	786,432	8,59	3,95
10	Trinity	Cray Intel Xeon E5	301,056	8,1	4,23
...	...	...	...	...	...
15	Shaheen 2	Cray Intel Xeon E5	196,608	5,53	2,83

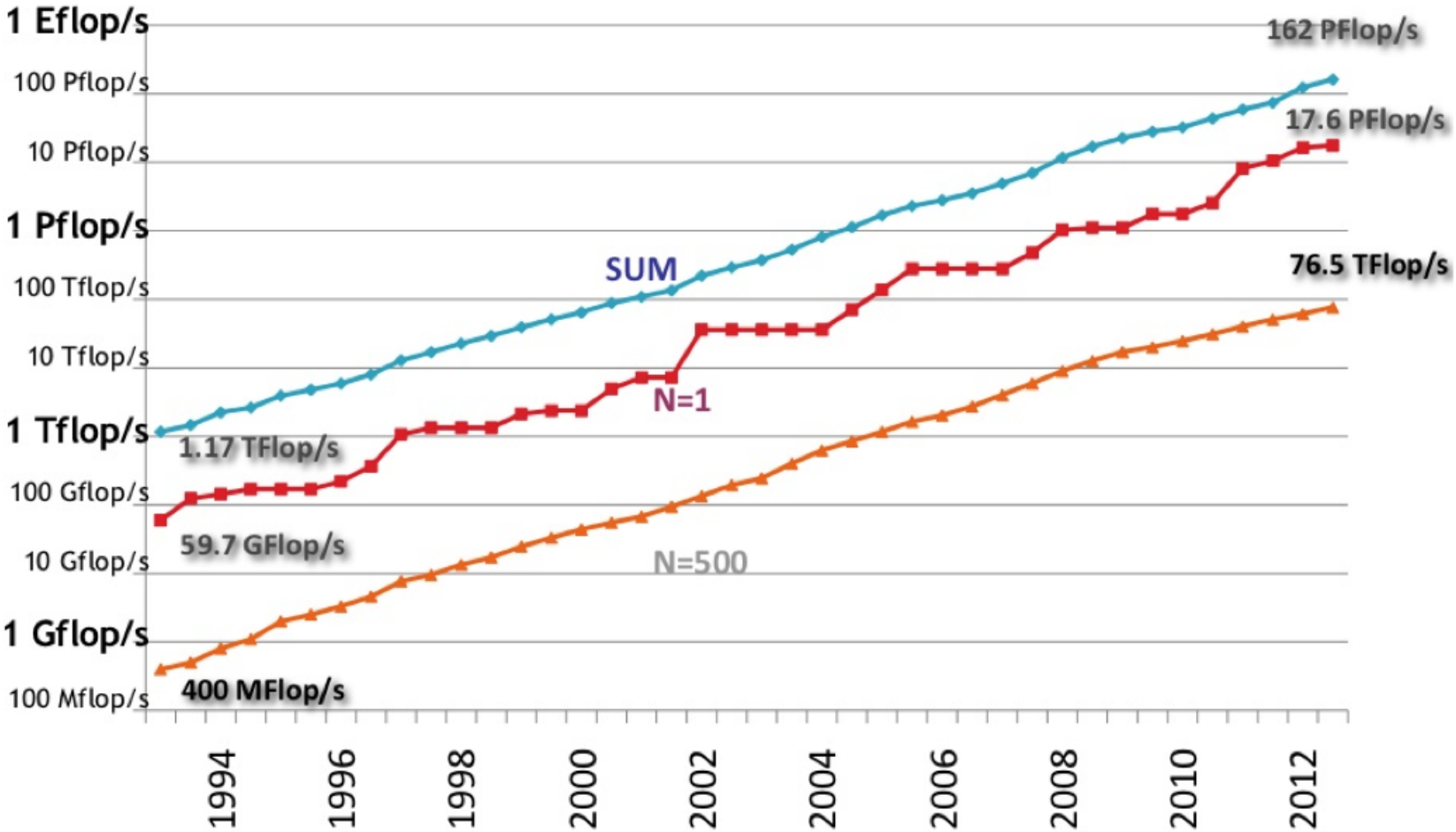
Ever increasing hardware complexity

The Top500 List, November'16: http://www.top500.org

Rank	Name	Vendor	Cores	R_{max} (Pflop/s)	Power (MW)
1	TaihuLight	NRCPC Sunway SW26010	10,649,600	93	15,3
2	Tianhe-2	NUDT Intel Xeon E5	3,120,000	33,8	17,8
3	Titan	Cray AMD Opteron + NVIDIA K20x	560,640	17,6	8,21
4	Sequoia	IBM BG/Q	1,572,864	17,2	7,89
5	Cori	Cray Intel Xeon Phi	622,336	14,01	3,93
6	Oakforest	Fujitsu Intel Xeon Phi	556,104	13,55	2,72
7	K Computer	Fujitsu SPARC64	705,024	10,51	12,66
8	PizDaint	Cray Intel Xeon E5 + NVIDIA P100	206,720	9,78	1,31
9	Mira	IBM BG/Q	786,432	8,59	3,95
10	Trinity	Cray Intel Xeon E5	301,056	8,1	4,23
...	...	...	...	...	...
15	Shaheen 2	Cray Intel Xeon E5	196,608	5,53	2,83

Human brain: 20 PetaFLOPS! (cf Kurzweil)

The Never Ending Race...



c/o J. Dongarra

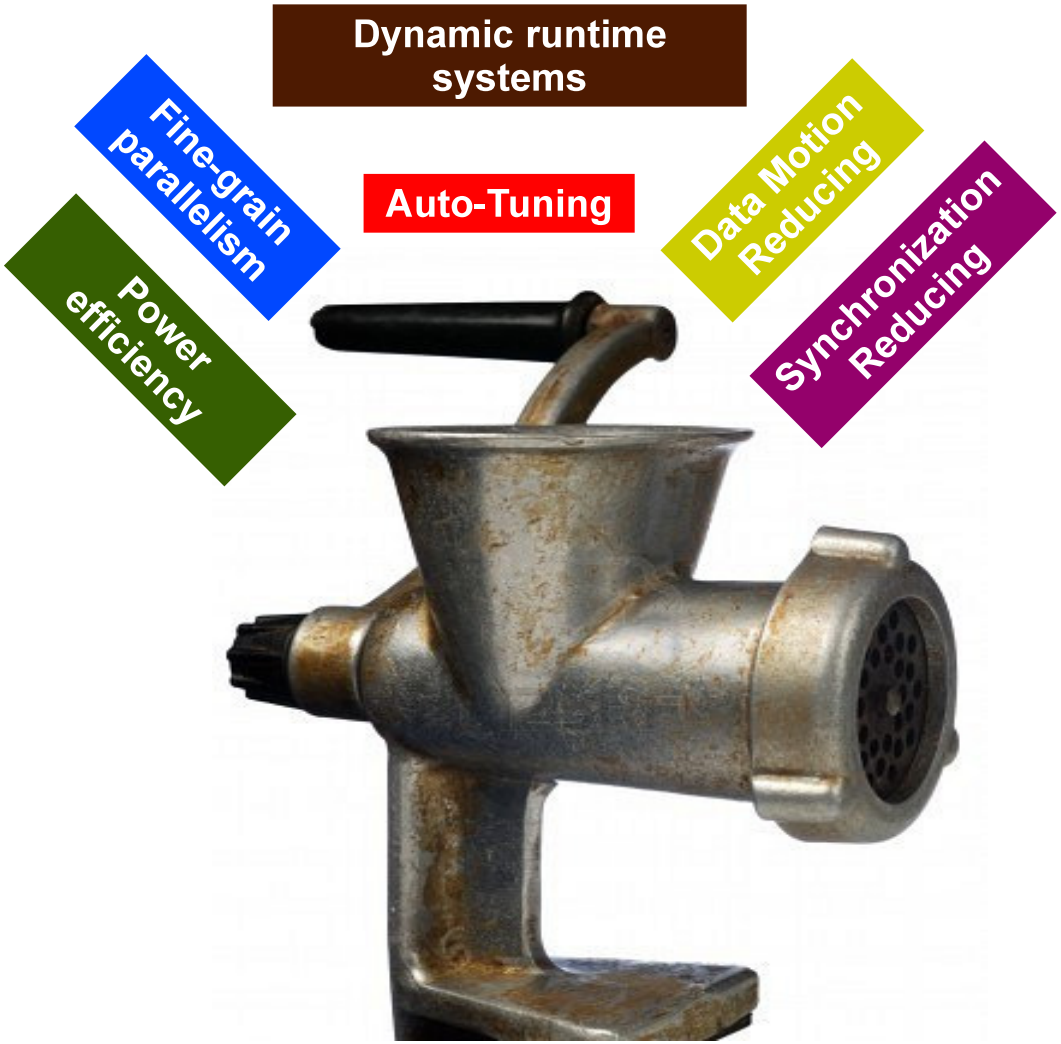
Just in case you are wondering whats beyond ExaFlops...

Mega, Giga, Tera, Peta, Exa, Zetta ...

10^{3}	kilo
10^{6}	mega
10^{9}	giga
10^{12}	tera
10^{15}	peta
10^{18}	exa
10^{21}	zetta

10^{24}	yotta
10^{27}	xona
10^{30}	weka
10^{33}	vunda
10^{36}	uda
10^{39}	treda
10^{42}	sorta
10^{45}	rinta
10^{48}	quexa
10^{51}	pepta
10^{54}	ocha
10^{57}	nena
10^{60}	minga
10^{63}	luma

Recipe for High Performance Linear Algebra



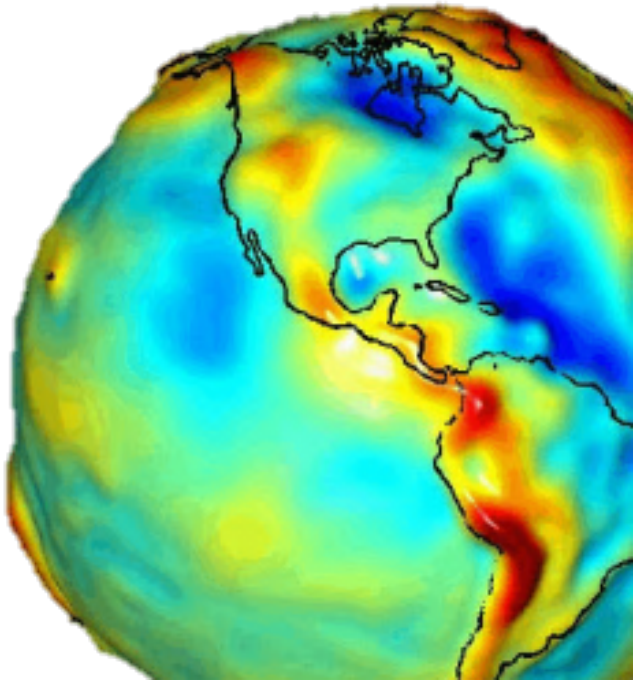
Outline

Geospatial Statistics

- Arising from multivariate large spatial data sets in climate/weather modeling to improve prediction for temperature, wind speeds, dust storm, etc.

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{Z}^T\Sigma^{-1}(\boldsymbol{\theta})\mathbf{Z} - \frac{1}{2}\log|\Sigma(\boldsymbol{\theta})|$$

(a) Problem Definition.



(b) Temperature prediction.

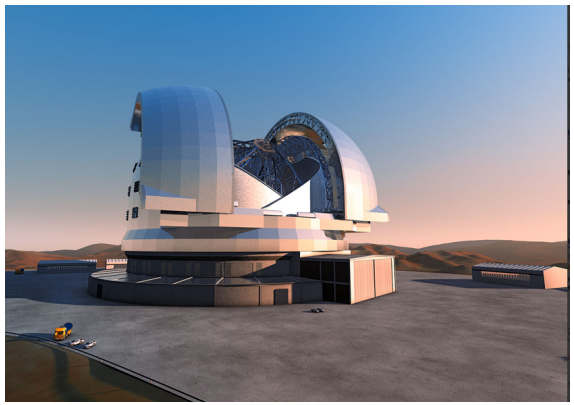
Figure: Modeling climate/weather forecasting.

Computational Ground-Based Astronomy

- Enhancing the observed image quality using MOAO by filtering out the noise coming from the adaptive optics instrumentation and the atmospheric turbulence.

$$R = C_{tm}.C_{mm}^{-1} \quad C_{ee} = C_{tt} - C_{tm}R^t - RC_{tm}^t + RC_{mm}R^t$$

(a) Problem Definition.



(b) The E-ELT.

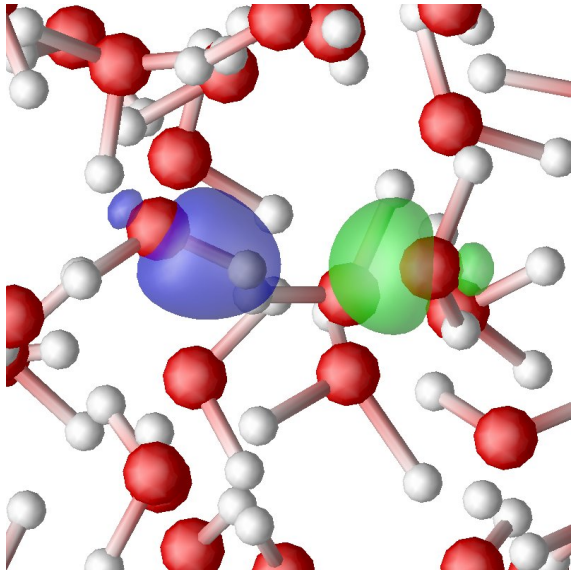
Figure: Finding new galaxies.

Computing the Eigenspectrum for Symmetric Hierarchical Low Rank Matrix

- Arising structural and vibrational analysis to problems in computational physics and chemistry like electronic and band structure calculations

$$(A - \lambda B)\, x = 0$$

(a) Problem Definition.



(b) Electronic structure.

Figure: Design of new materials.

Outline

Matrix Form

The Cholesky factorization of an $N \times N$ real symmetric, positive-definite matrix A has the form

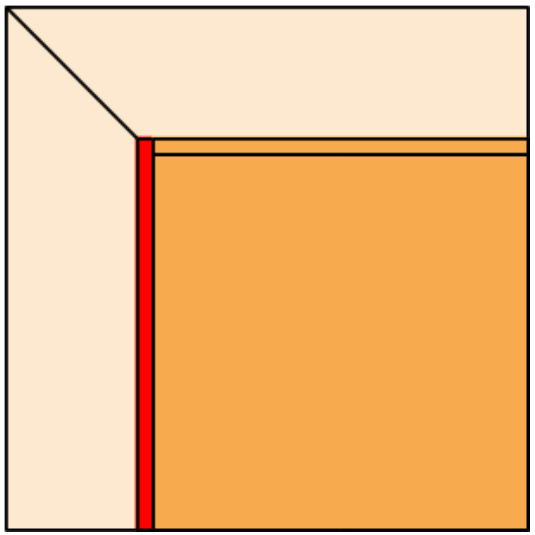
$$A = LL^T,$$

where L is an $N \times N$ real lower triangular matrix with positive diagonal elements.

A Look Back...

Software infrastructure and algorithmic design follow hardware evolution in time:

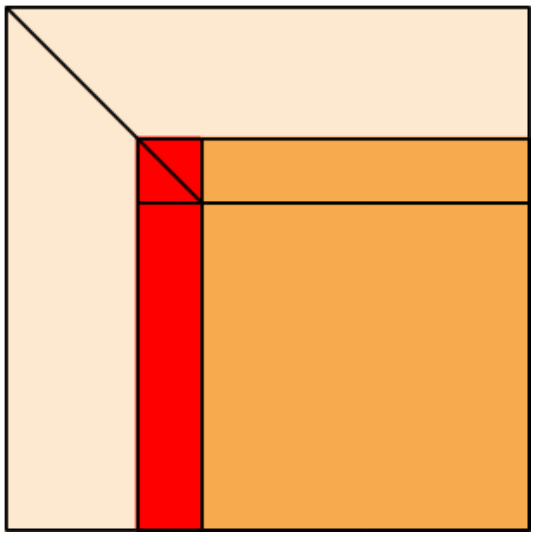
- 70's - LINPACK, vector operations:
  Level-1 BLAS operation



A Look Back...

Software infrastructure and algorithmic design follow hardware evolution in time:

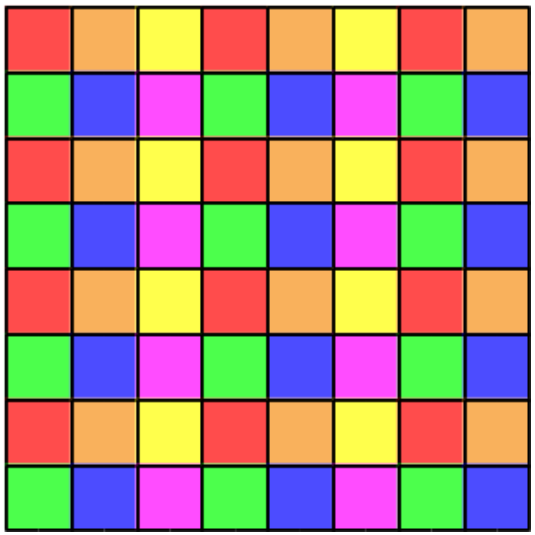
- 70's - LINPACK, vector operations:
  Level-1 BLAS operation
- 80's - LAPACK, block, cache-friendly:
  Level-3 BLAS operation



A Look Back...

Software infrastructure and algorithmic design follow hardware evolution in time:

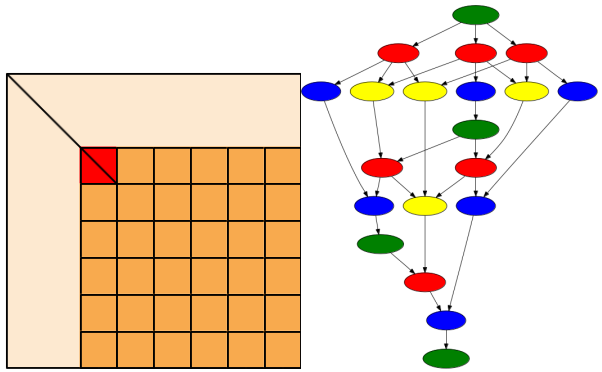
- 70's - LINPACK, vector operations:
  Level-1 BLAS operation
- 80's - LAPACK, block, cache-friendly:
  Level-3 BLAS operation
- 90's - ScaLAPACK, distributed memory:
  PBLAS Message passing



A Look Back...

Software infrastructure and algorithmic design follow hardware evolution in time:

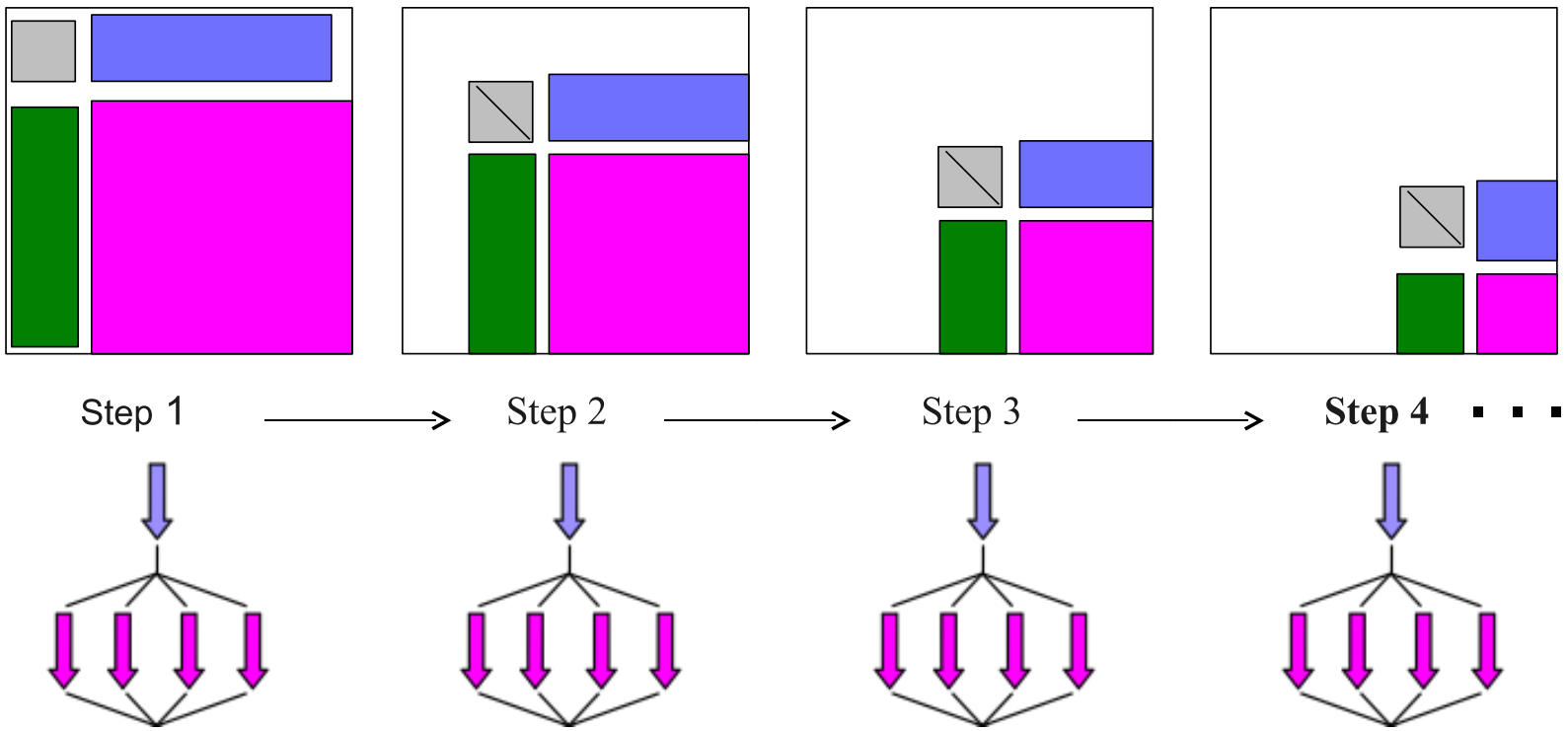
- 70's - LINPACK, vector operations:
  Level-1 BLAS operation
- 80's - LAPACK, block, cache-friendly:
  Level-3 BLAS operation
- 90's - ScaLAPACK, distributed memory:
  PBLAS Message passing
- 00's:
  - PLASMA, MAGMA, many x86/cuda cores friendly:
    DAG scheduler, tile data layout, some extra kernels



Block Algorithms

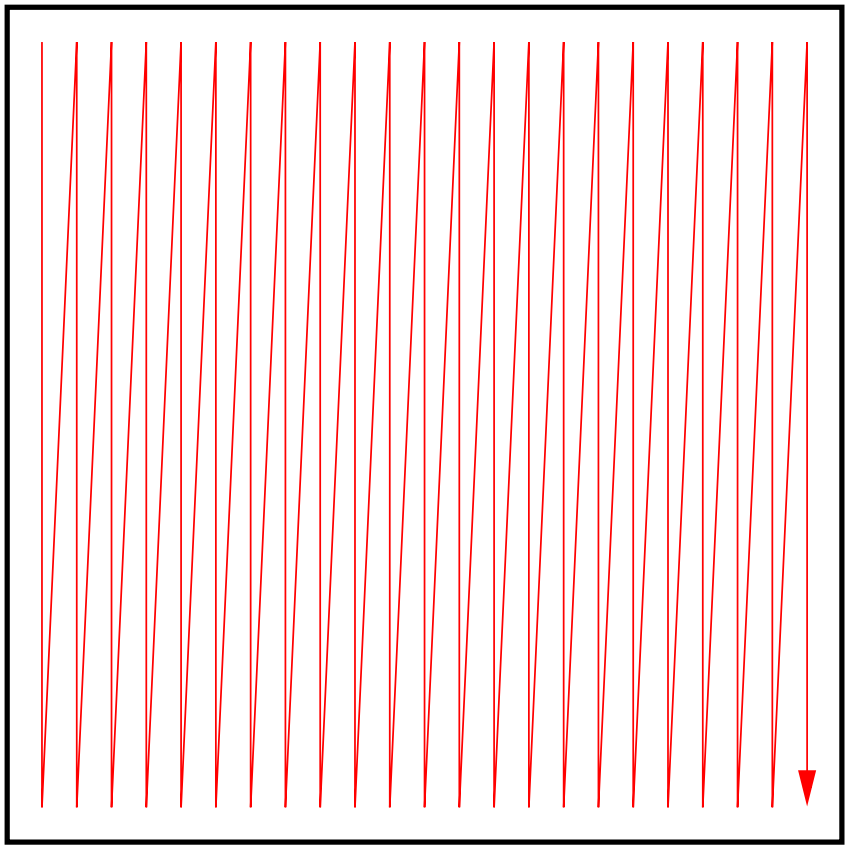
- Panel-Update Sequence
- Transformations are blocked/accumulated within the Panel (Level 2 BLAS)
- Transformations applied at once on the trailing submatrix (Level 3 BLAS)
- Parallelism hidden inside the BLAS
- Fork-join Model

Block Algorithms: Fork-Join Paradigm

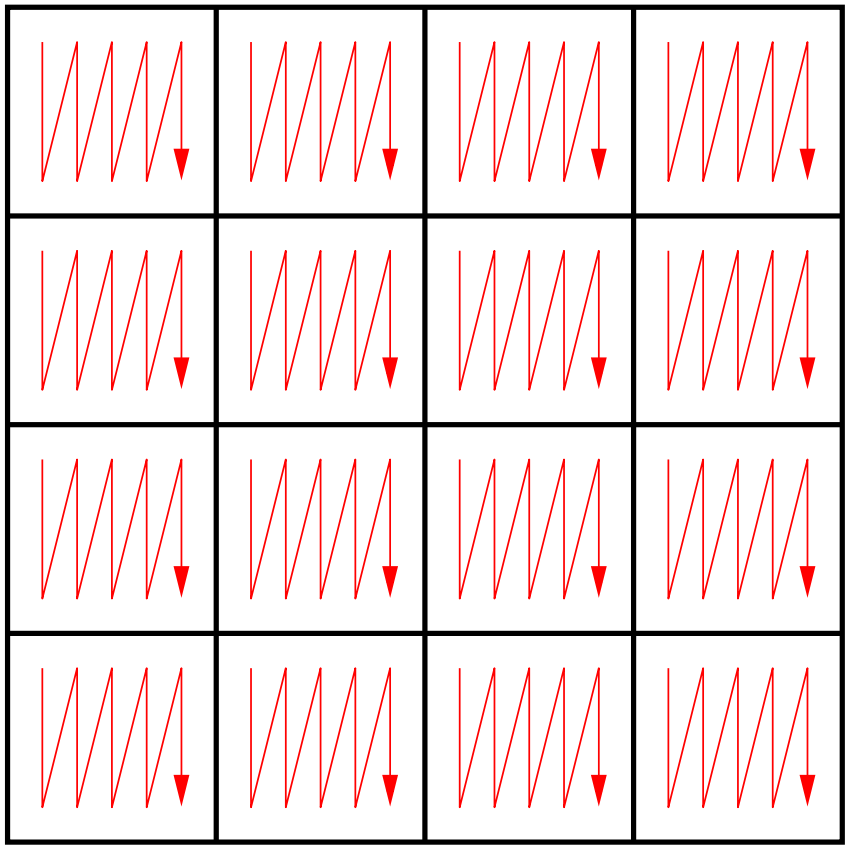


Tile Data Layout Format

LAPACK: column-major format



PLASMA: tile format



PLASMA: Tile Algorithms

PLASMA: Parallel Linear Algebra for Scalable Multi-core Architectures
$\Longrightarrow$ http://icl.cs.utk.edu/plasma/

- Parallelism is brought to the fore
- May require the redesign of linear algebra algorithms
- Tile data layout translation
- Remove unnecessary synchronization points between Panel-Update sequences
- DAG execution where nodes represent tasks and edges define dependencies between them
- Dynamic runtime system environment QUARK

MAGMA: x86 + MultiGPUs

MAGMA: Matrix Algebra on GPU and Multicore Architectures $\Longrightarrow$ http://icl.cs.utk.edu/magma/

- Lessons Learned from PLASMA!
- CUDA-based hybrid systems
- New high performance numerical kernels
- StarPU (INRIA, Bordeaux), OmpSs (BSC, Barcelona), PaRSEC (UTK, Knoxville)
- Both: x86 and GPUs $\Longrightarrow$ Hybrid Computations
- Similar to LAPACK in functionality

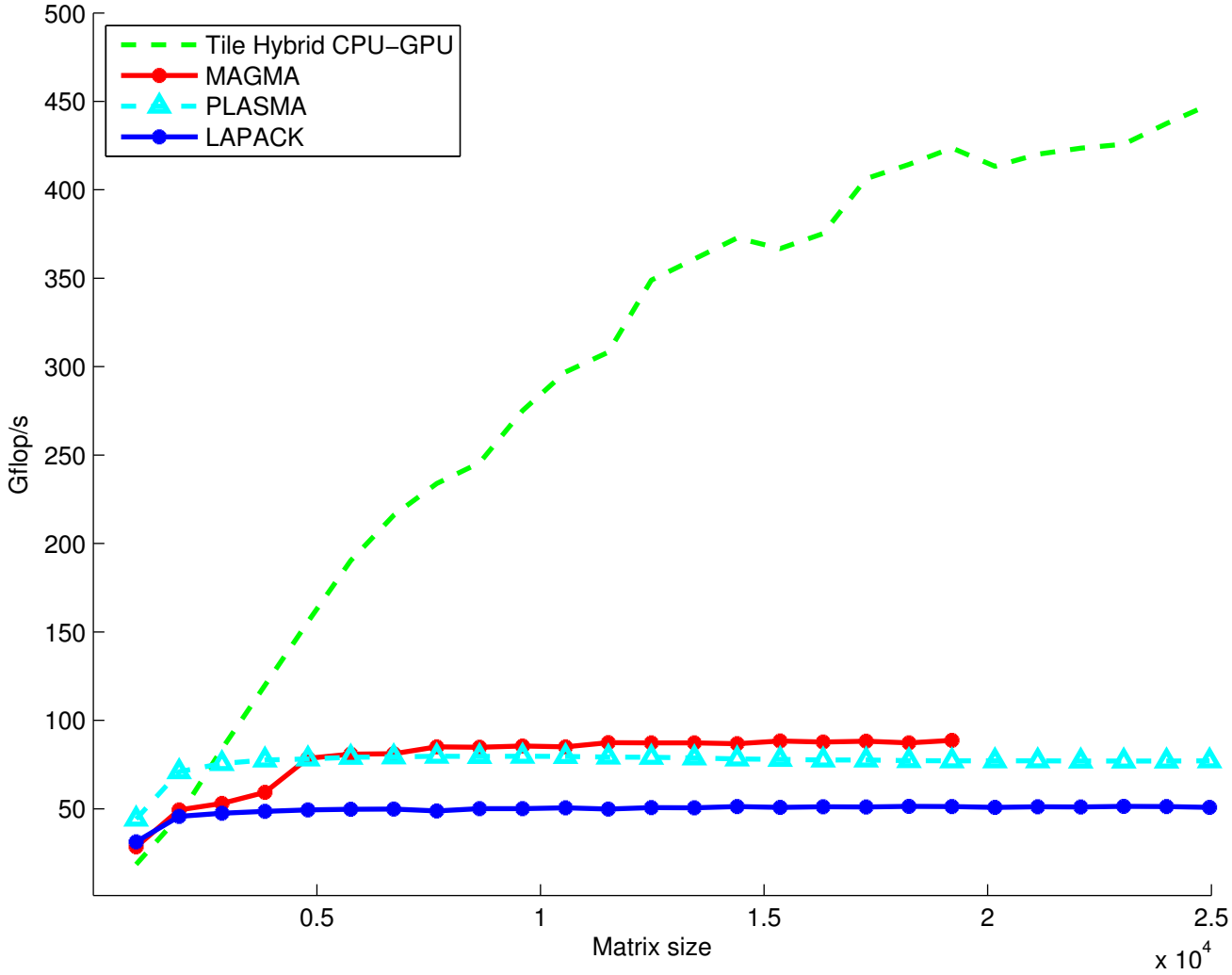
Dynamic Runtime System

- Conceptually similar to out-of-order processor scheduling

because it has:

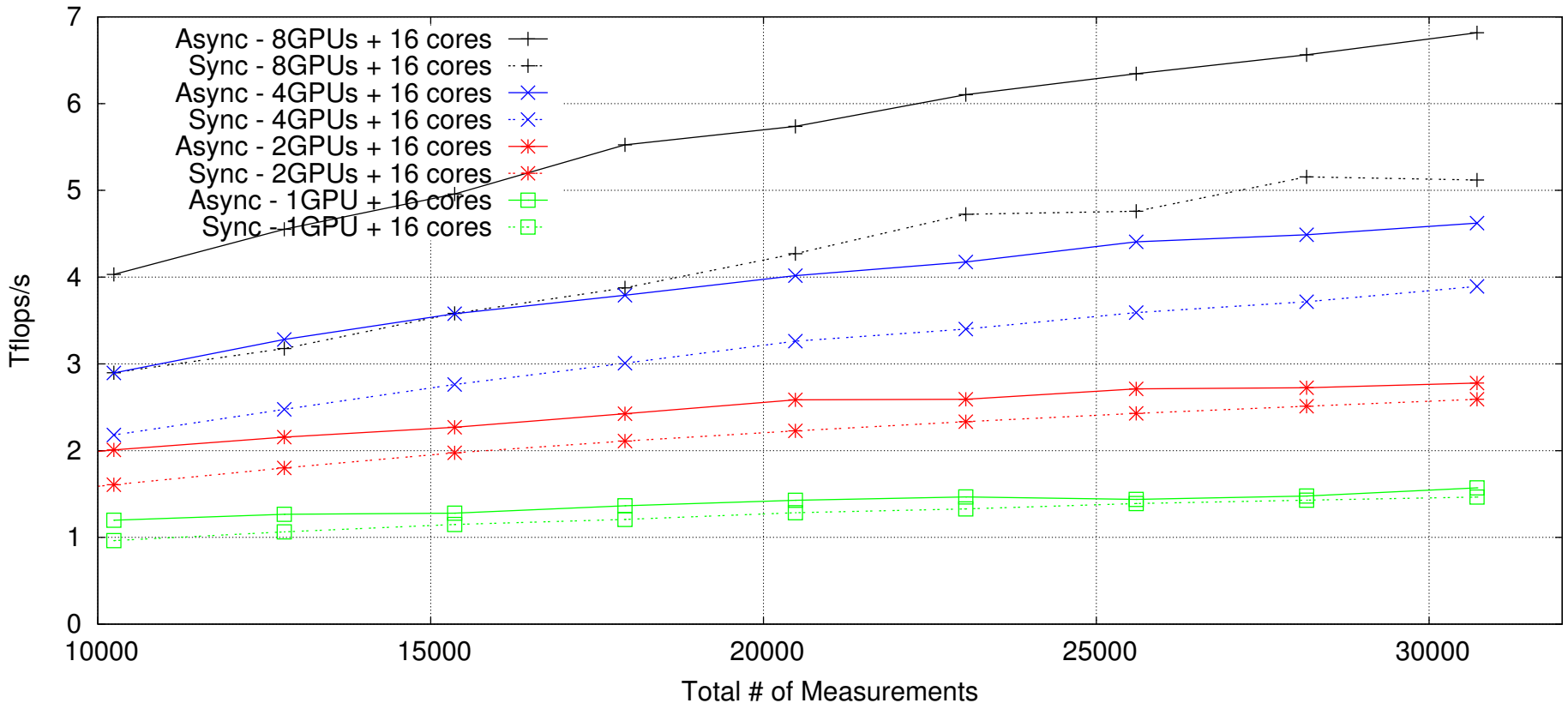
- Dynamic runtime DAG scheduler
- Out-of-order execution flow of fine-grained tasks
- Task scheduling as soon as dependencies are satisfied
- *Producer-Consumer*

State-of-the-art Performance Comparisons



H. Ibeid, D. Kaushik, D. Keyes and H. Ltaief, HIPC'11, India

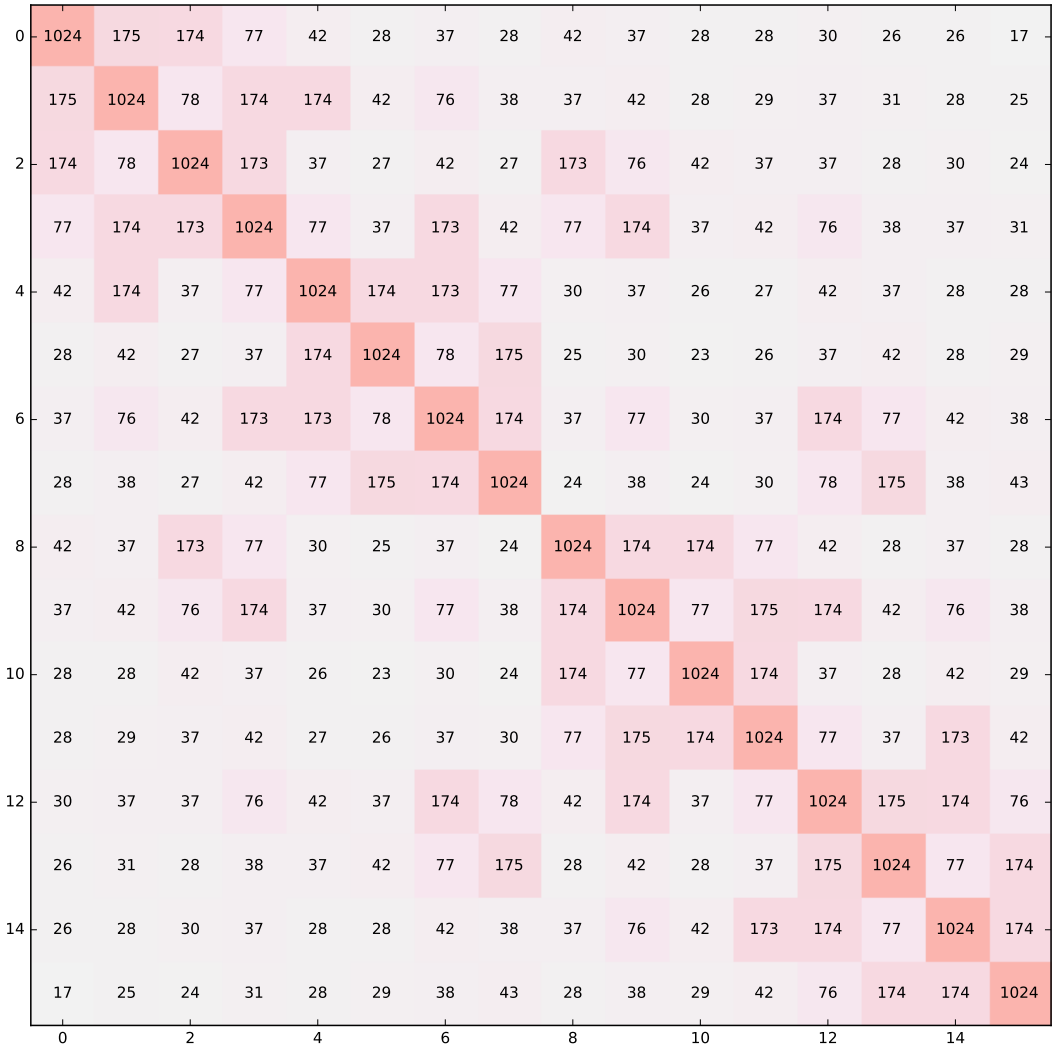
MOAO performance in Tflop/s on 16 SDB cores + 8 K40 GPUs



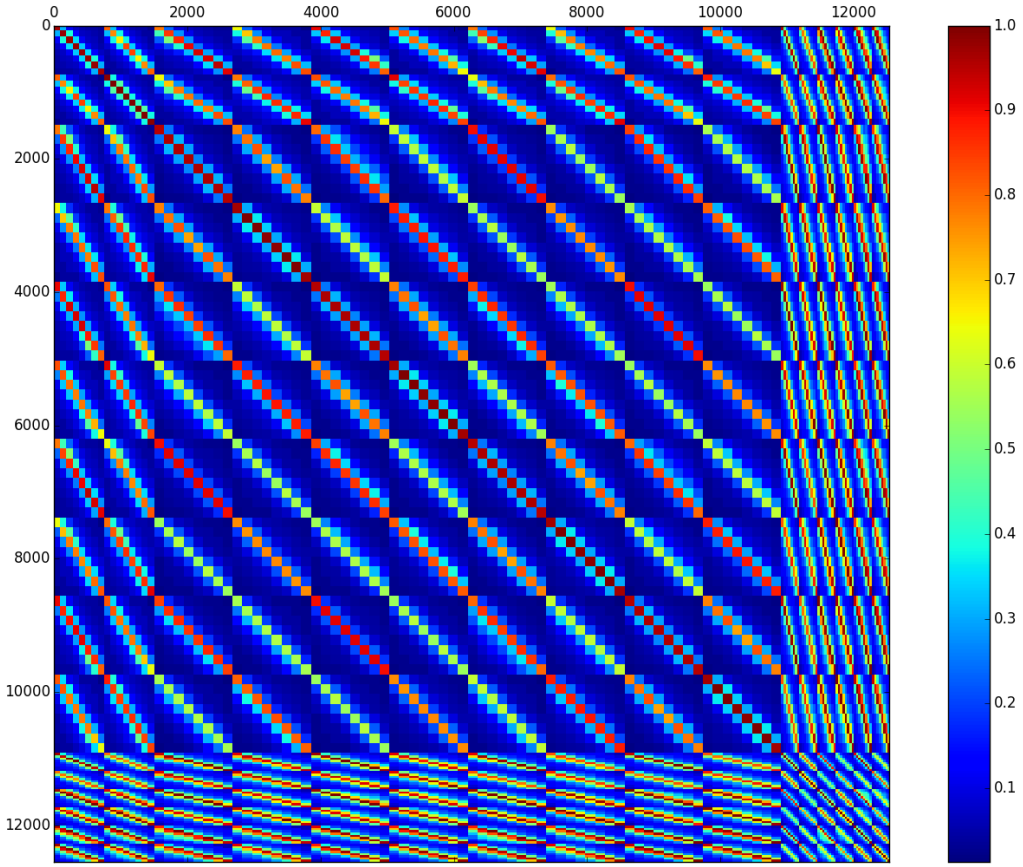
H. Ltaief, D. Gratadour, A. Charara and E. Gendron, PASC'16, Switzerland 6min to simulate 50 PSFs!

Outline

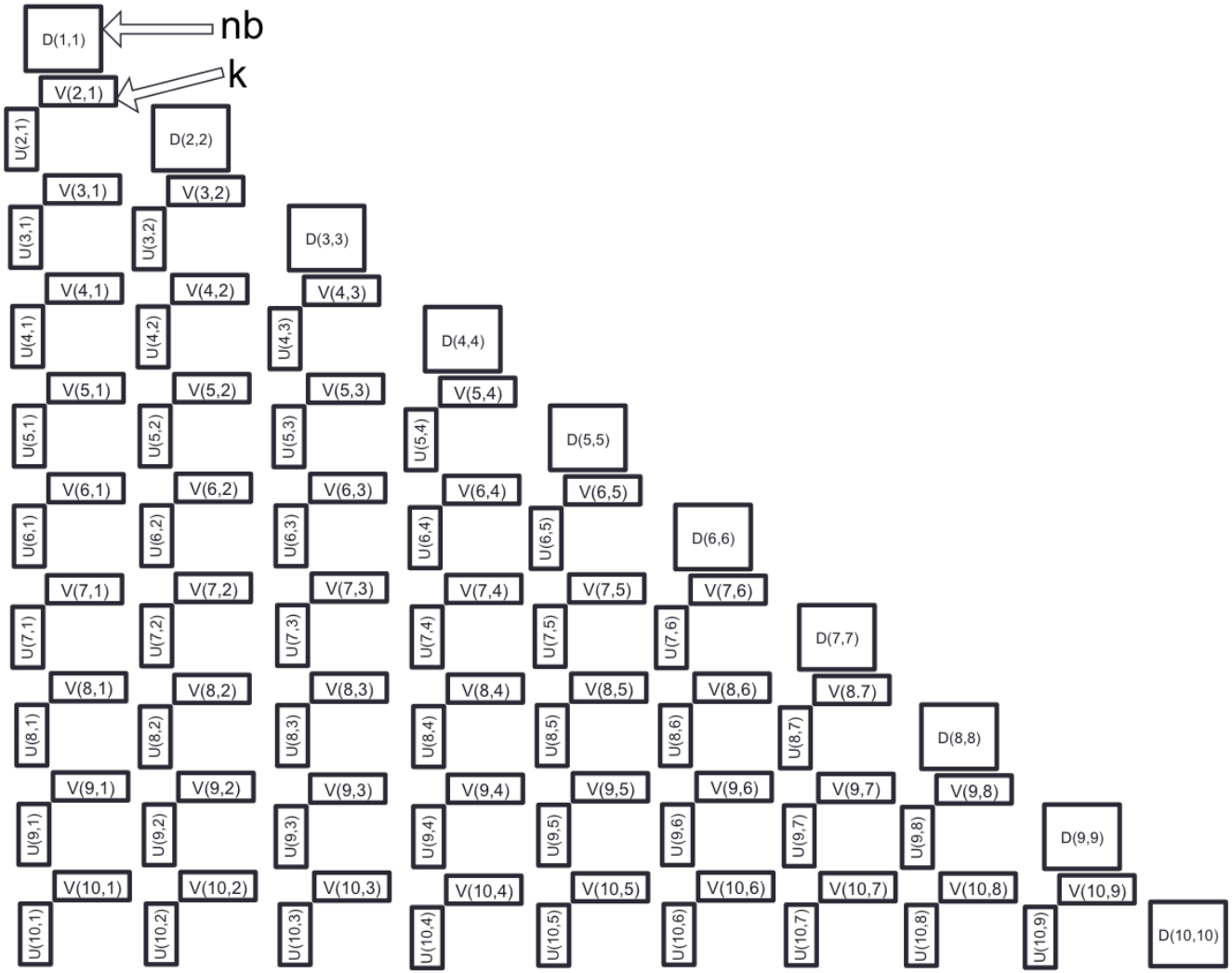
Geospatial Statistics



Computational Astronomy



HiCMA DPOTRF: batch algorithms, matrix structure and data-sparsity exploitation



HiCMA's methodology

1. Redesign the numerical algorithm so that it operates on the low rank data structure.
2. Formulate the new task-based low rank matrix computation using a DAG representation.
3. Employ a dynamic runtime system to ensure proper scheduling, asynchronous execution and load balancing.
4. Ensure a separation of concerns by abstracting the hardware complexity from users.

HiCMA DPOTRF for Climate Apps on KNL

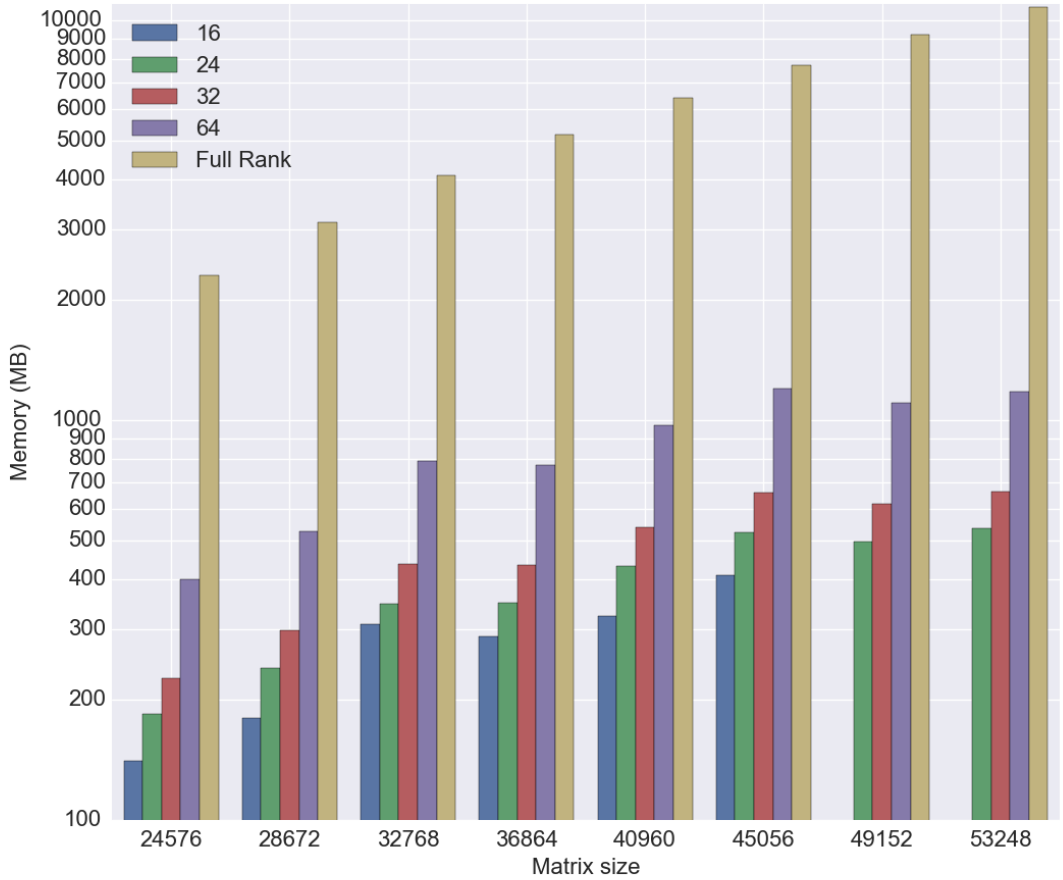


Figure: Memory footprint depending on the truncation rank.

HiCMA DPOTRF for Climate Apps on KNL

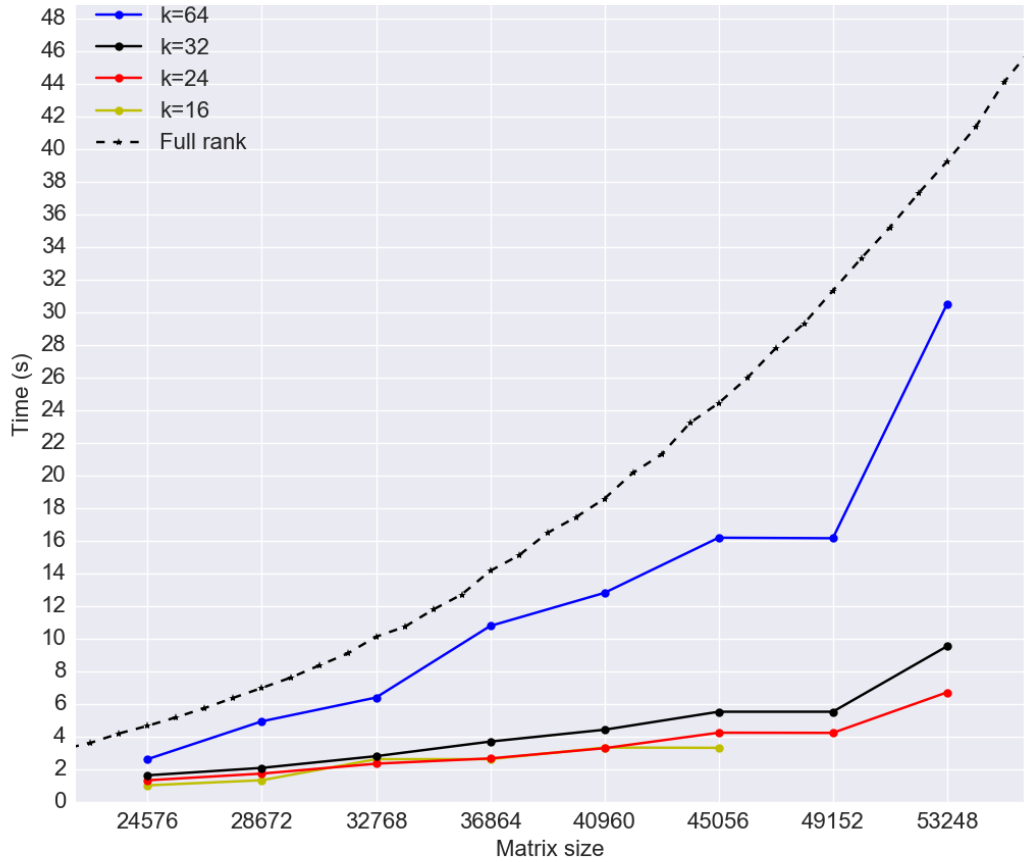


Figure: Performance in time (seconds) depending on the truncation rank.

Outline

New data structures needed

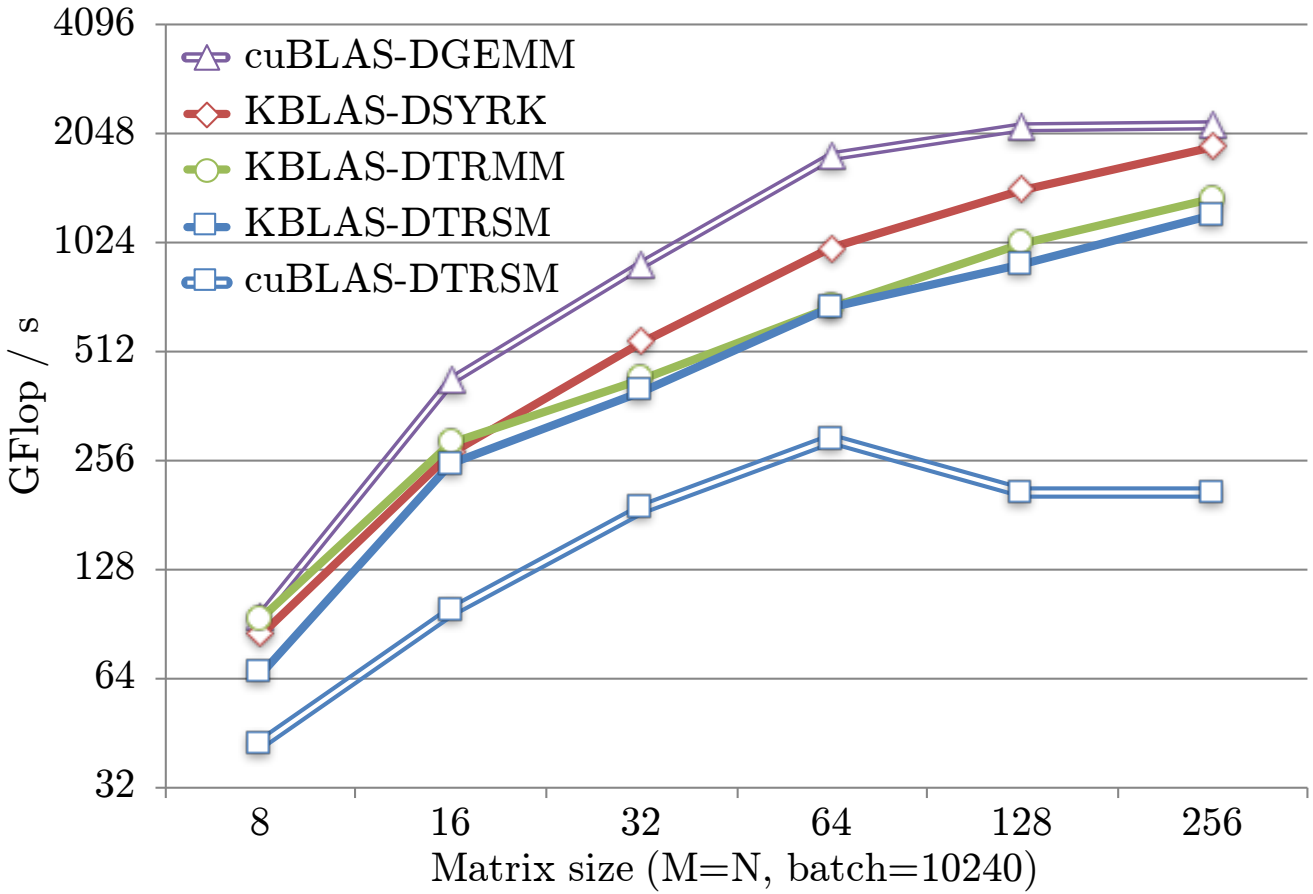
- Redesign the algorithm from tile-centric to kernel-centric
- Increase hardware occupancy
- Refactor the code around batch BLAS (incomplete vendor support)
- Use contiguous memory buffers
- Expose opportunities for HW/SW prefetching

Batch BLAS Operations: KBLAS

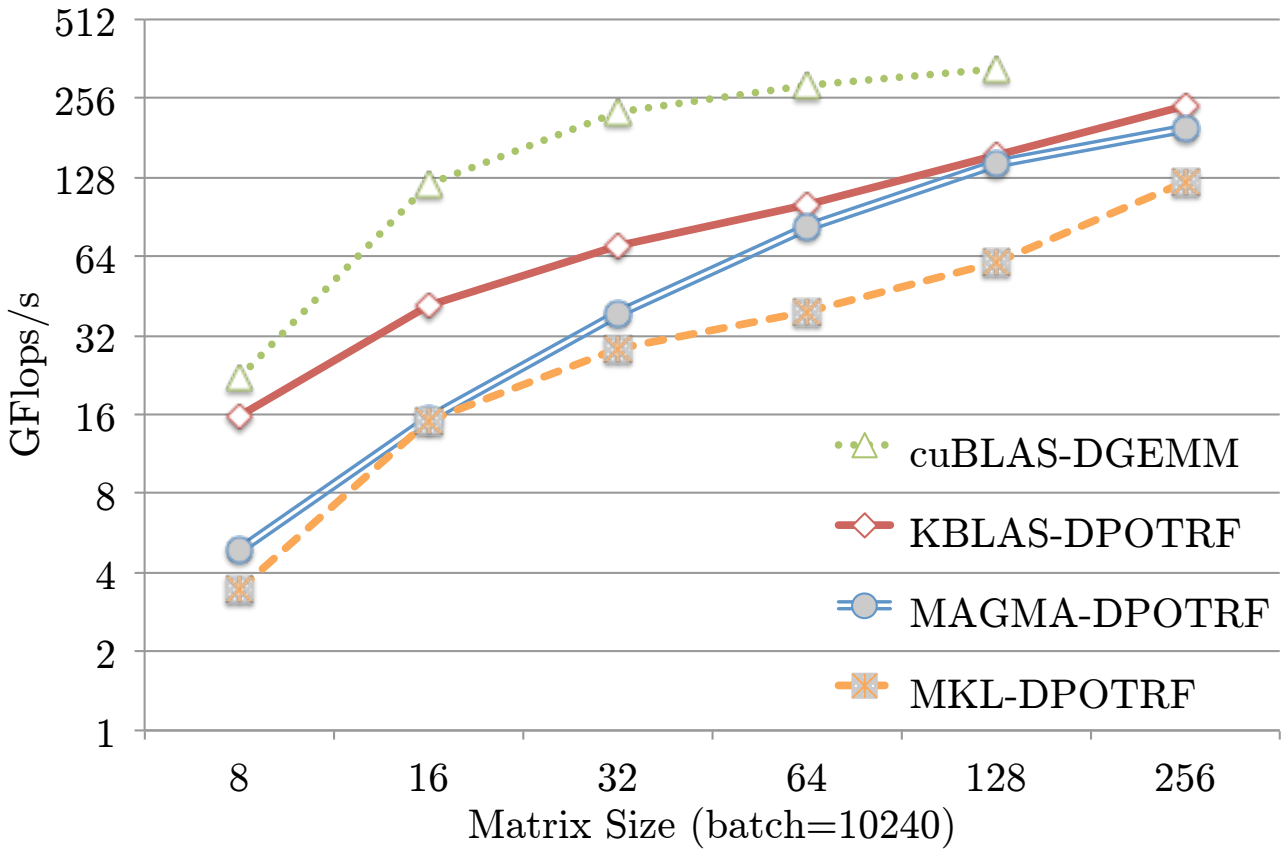
$\mathcal{H}$-Matrix computations:

- Level 3 BLAS: SYRK, TRSM, GEMM
- Factorization: POTRF
- Solve: POTRS, POSV
- Inversion: LAUUM, TRTRI, POTRI, POTI
- Truncation: QR, SVD

KBLAS Performance Results: Level 3 BLAS on NVIDIA P100



KBLAS Performance Results: Factorization on NVIDIA K40



Students/Collaborators/Vendors

- Extreme Computing Research Center @ KAUST: **S. Abdullah, K. Akbudak, W. Boukaram, A. Charara, G. Chávez, M. Genton, D. Keyes, A. Mikhalev, D. Sukkari, G. Turkiyyah and Y. Sun**
- L'Observatoire de Paris, LESIA: **R. Dembet, N. Doucet, E. Gendron, D. Gratadour, A. Sevin and F. Vidal**
- Innovative Computing Laboratory @ UTK: **PLASMA/MAGMA/PaRSEC Teams**
- INRIA/INP/LaBRI Bordeaux, France: **Runtime/HiePACS Teams**
- Max-Planck Institute @ Leipzig, Germany: **R. Kriemann**
- Barcelona Supercomputing Center, Spain: **OmpSs Team**
- KAUST Supercomputing Lab and IT Research Computing support
- NVIDIA GPU Research Center
- Intel Parallel Computing Center
- Cray Center of Excellence