GPU-accelerated Applications

Hatem Ltaief
Principal Research Scientist, KAUST

Advanced School on High-Performance Computing and Applied AI for High-Resolution Regional Climate Modeling

September 8-19 2025

Top10 fastest supercomputers for HPL (Nov'24)

#	Site	Manufacturer	TOP10 Computer of the TOP500	Country	Cores	Rmax [Pflops]	Power [MW]
1	Lawrence Livermore National Laboratory	HPE	**El Capitan** HPE Cray EX255a, AMD EPYC 24C 1.8GHz, Instinct MI300A, Slingshot-11	USA	11,039,616	1,742	29.6
2	Oak Ridge National Laboratory	HPE		USA	9,066,176	1,353	24.6
3	Argonne National Laboratory	Intel		USA	4,742,808	1,012	38.7
4	Microsoft Azure	Microsoft		USA	1,123,200	561.2	
5	Eni S.p.A. Center for Computational Science	HPE		Italy	3,143,520	477.9	8.5
6	RIKEN Center for Computational Science	Fujitsu	**Fugaku** Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D	Japan	7,630,848	442.0	29.9
7	Swiss National Supercomputing Centre (CSCS)	HPE	**Alps** HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, GH200, Slingshot-11	Switzerland	2,121,600	434.9	7.1
8	EuroHPC / CSC	HPE	**LUMI** HPE Cray EX235a, AMD EPYC 64C 2.0GHz, Instinct MI250X, Slingshot-11	Finland	2,752,704	379.7	7.1
9	EuroHPC / CINECA	EVIDEN	**Leonardo** Atos BullSequana XH2000, Xeon 32C 2.6GHz, NVIDIA A100, HDR Infiniband	Italy	1,824,768	241.2	7.5
10	Lawrence Livermore National Laboratory	HPE	**Tuolumne** HPE Cray EX255a, AMD EPYC 24C 1.8GHz, Instinct MI300A, Slingshot-11	USA	1,161,216	208.1	3.4

Top10 fastest supercomputers for HPL-MxP (Nov’24)

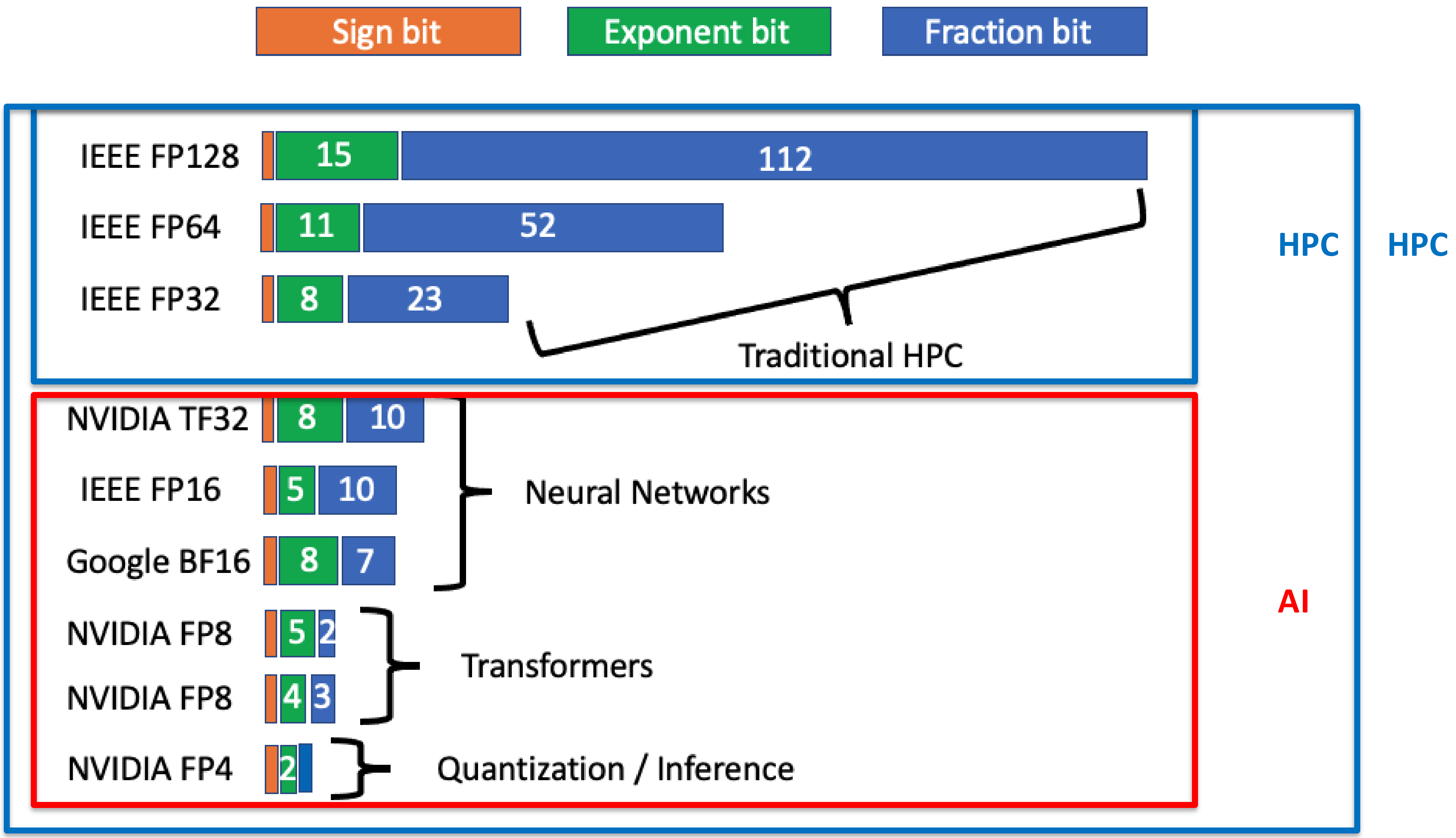
Rank	Site	Computer	Cores	HPL Rmax (Eflop/s)	TOP500 Rank	HPL-MxP (Eflop/s)	Speedup
1	DOE/SC/ANL **USA**	**Aurora**, HPE Cray EX, Intel Max 9470 52C, 2.4 GHz, Intel GPU MAX, Slingshot-11	8,159,232	1.012	3	11.6	11.5
2	DOE/SC/ORNL **USA**	**Frontier**, HPE Cray EX235a, AMD Zen-3 (Milan) 64C 2GHz, AMD MI250X, Slingshot-11	8,560,640	1.353	2	11.4	8.4
3	EuroHPC/CSC **Finland**	**LUMI**, HPE Cray EX235a, AMD Zen-3 (Milan) 64C 2GHz, AMD MI250X, Slingshot-11	2,752,704	0.380	8	2.35	6.2
4	RIKEN Center for Comput. Science, **Japan**	**Fugaku**, Fujitsu A64FX 48C 2.2GHz, Tofu D	7,630,848	0.442	6	2.00	4.5
5	EuroHPC/CINECA **Italy**	**Leonardo**, BullSequana XH2000, Xeon 8358 32C 2.6GHz, NVIDIA A100, QR NVIDIA HDR100 IB	1,824,768	0.241	9	1.80	7.5
6	CII, Institute of Science **Japan**	**TSUBAME 4**, HPE Cray XD665, AMD EPYC 9654 96C 2.4GHz, NVIDIA H100, Mellanox NDR200	172,800	0.025	47	0.64	25
7	NVIDIA **USA**	**Selene**, DGX SuperPOD, AMD EPYC 7742 64C 2.25 GHz, Mellanox HDR, NVIDIA A100	555,520	0.063	23	0.63	9.9
8	DOE/SC/LBNL/NERSC **USA**	**Perlmutter**, HPE Cray EX235n, AMD EPYC 7763 64C 2.45 GHz, Slingshot-10, NVIDIA A100	761,856	0.079	19	0.59	7.5
9	FZJ - Jülich **Germany**	**JUWELS Booster Module**, Bull Sequana XH2000 , AMD EPYC 24C 2.8GHz, HDR IB, NVIDIA A100	449,280	0.044	33	0.47	10
10	GENCI-CINES **France**	**Adastra**, HPE Cray EX235a, AMD EPYC 64C 2GHz, AMD 250X, Slingshot-11	319,072	0.046	30	0.30	6.5

Most energy efficient architectures

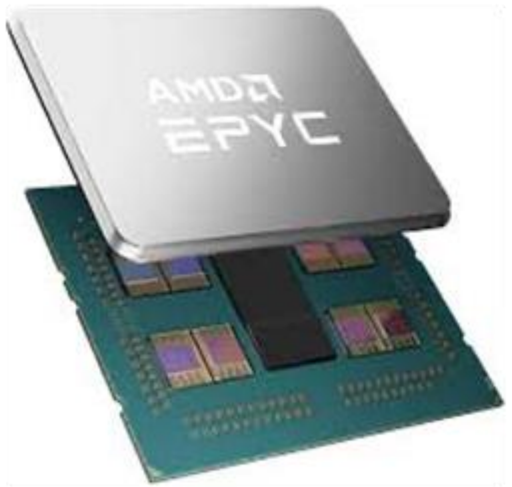
Computer	Host	Interconnect	Device	Rmax/ Power
JEDI, BullSequana XH3000	Grace Hopper Superchip 72C 3GHz	Quad-Rail NVIDIA InfiniBand NDR200	NVIDIA Grace	*72.73
ROMEO-2025, BullSequana XH3000	Grace Hopper Superchip 72C 3GHz	Quad-Rail NVIDIA InfiniBand NDR200	NVIDIA Grace	*70.91
Adastra 2, HPE Cray EX255a	AMD EPYC 24C 1.8GHz	Slingshot-11	AMD MI300A	*69.10
Isambard-AI phase 1, HPE Cray EX254n	NVIDIA Grace 72C 3.1GHz	Slingshot-11	NVIDIA Grace	*68.83
Capella, Lenovo ThinkSystem SD650 V3	AMD EPYC 9334 32C 2.7GHz	InfiniBand NDR200	NVIDIA H100	*68.05
JETI - JUPITER Exascale Transition Instrument, BullSequana XH3000	Grace Hopper Superchip 72C 3GHz	Quad-Rail NVIDIA InfiniBand NDR200	NVIDIA Grace	*67.96
Helios GPU, HPE Cray EX254n	NVIDIA Grace 72C 3.1GHz	Slingshot-11	NVIDIA Grace	*66.95
Henri, ThinkSystem SR670 V2	Intel Xeon Platinum 8362 32C 2.8GHz	Infiniband HDR	NVIDIA H100	65.40
HoreKa-Teal, ThinkSystem SD665-N V3	AMD EPYC 9354 32C 3.25GHz	Infiniband NDR200	NVIDIA Grace	62.96
rzAdams, HPE Cray EX255a	AMD EPYC 24C 1.8GHz	Slingshot-11	AMD MI300A	62.80

[Gflops/Watt]

Feeling like a kid in a candy store

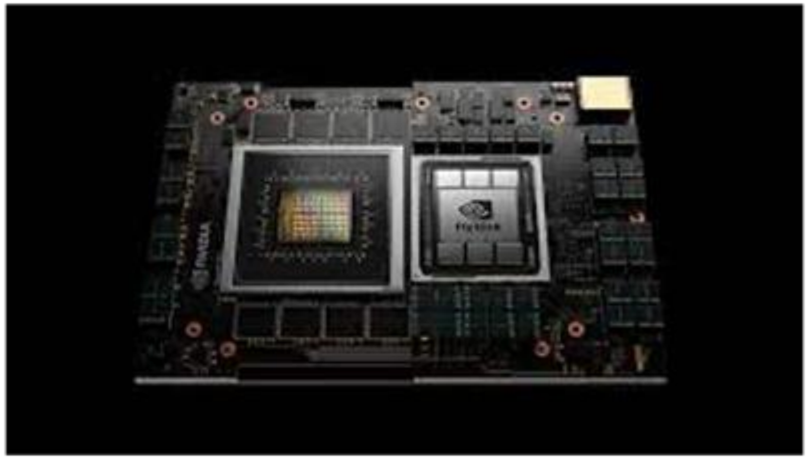


Feeling again like a kid in a candy store



AMD Epyc Genoa

High cache capacity
High memory bandwidth
x86 programming env
Memory-bound workloads



NVIDIA Grace Hopper

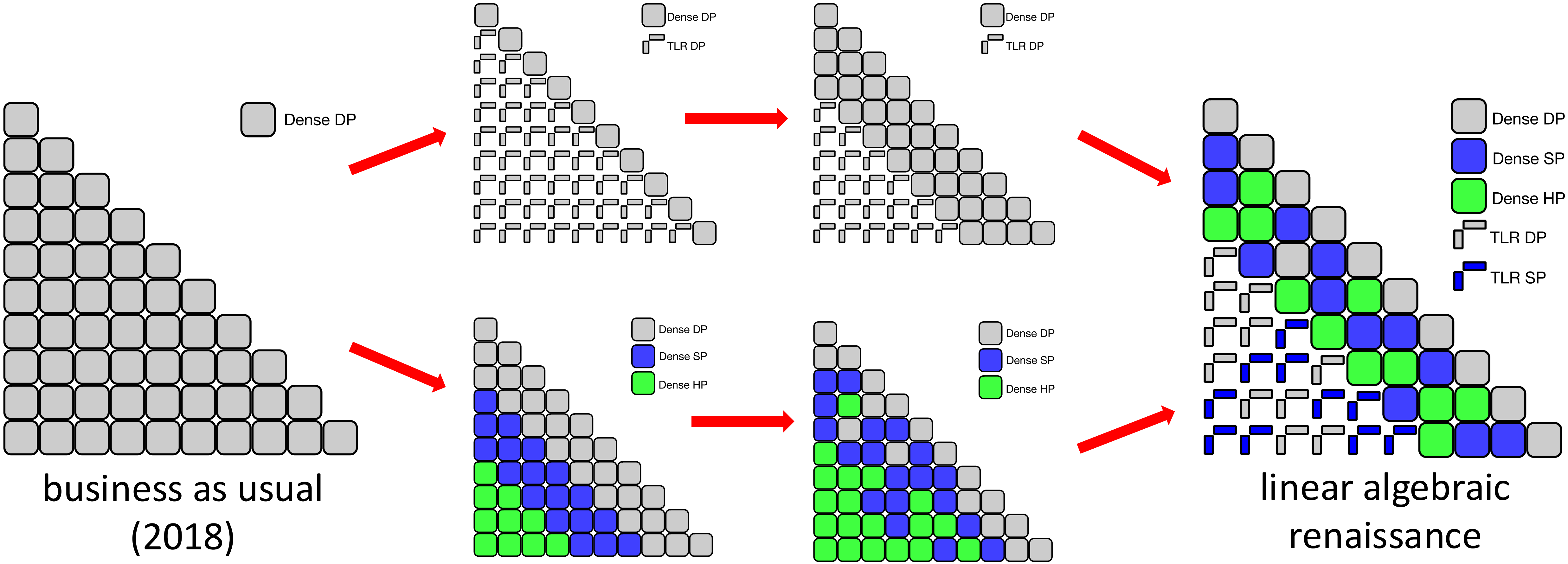
High speed CPU-GPU interconnect
Memory coherency
Support for mixed precisions
Compute-bound workloads

Cerebras CS-2
Graphcore IPU

AI-focused chip
Flat memory hierarchy
High SRAM bandwidth
Inference

HL, Y. Hong, A. Dabah, R. Alomairy, S. Abdulah, C. Goreczny, P. Gepner, M. Ravasi, D. Gratadour & D. Keyes, *Steering Customized AI Architectures for HPC Scientific Applications*, ISC23.

The journey toward linear algebra *Renaissance*



Don't over-solve: maintain just enough accuracy for the application purpose

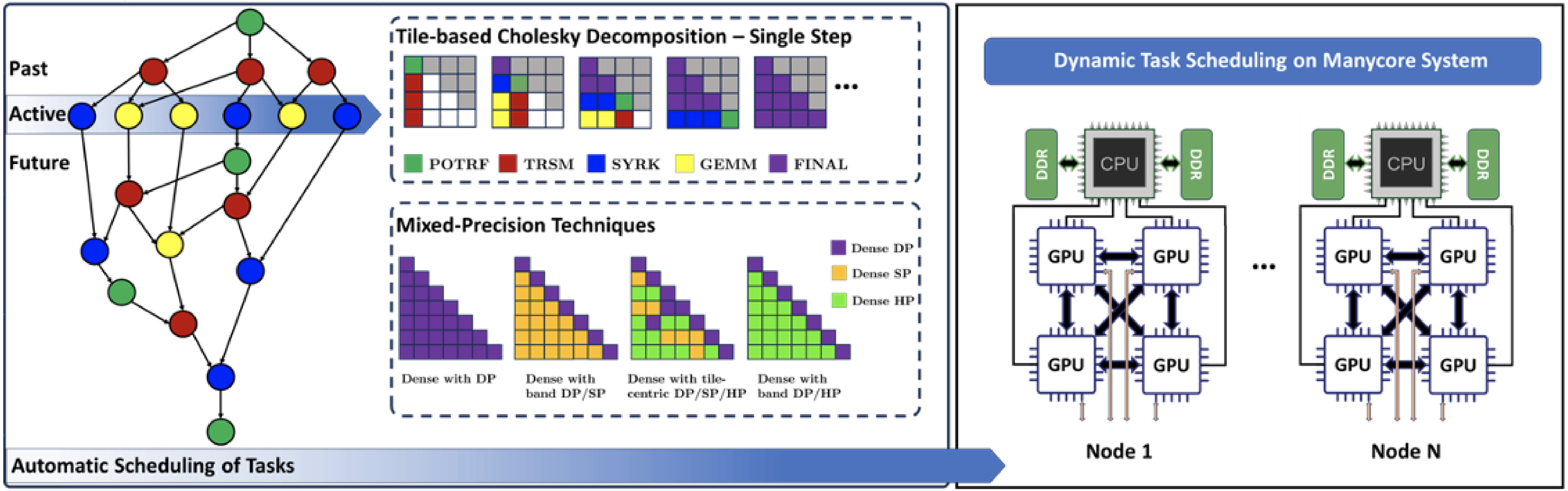
Don't over-store: no extra copies of the original matrix

S. Tomov, R. Nath, HL and J. Dongarra, *Dense linear algebra solvers for multicore with GPU accelerators*, IEEE IPDPS, 2010.

N. Higham and T. Mary, Mixed Precision Algorithms in Numerical Linear Algebra, Acta Numerica, 2022.

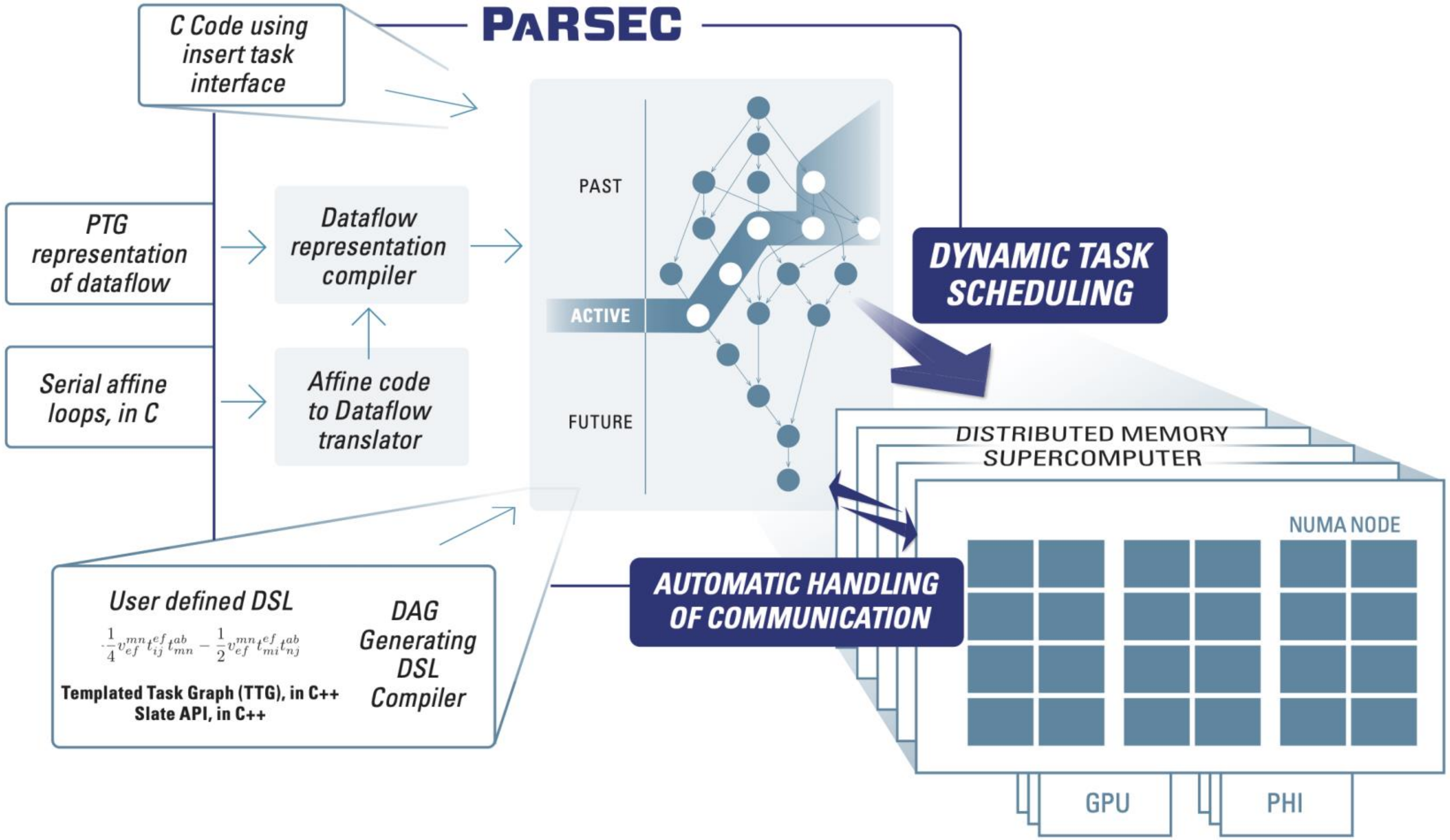
HL, M. Genton, D. Gratadour, D. Keyes and M. Ravasi, *Responsibly Reckless Matrix Algorithms for HPC Scientific Applications*, Computing in Science & Engineering, 2022.

Tile-Centric MxP Cholesky-based Solver



PaRSEC: a dynamic runtime system of the DOE ECP

DAG: directed acyclic graph
PTG: parameterized task graph
DSL: domain specific language



https://icl.utk.edu/parsec/

Applications developer: do not oversolve!


Hatem Ltaief
@HatemLtaief

Love it!

#QuoteOfTheDay

Dan Ernst @ernstdj · 9/8/24
The most efficient FLOP is the one you didn't use.

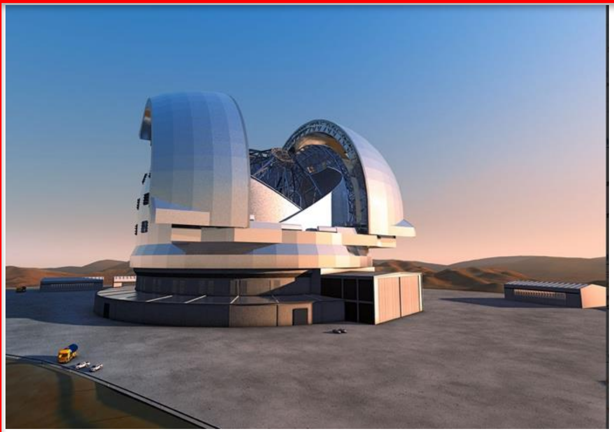
13:20 · 9/8/24 · **432** Views

1 8

Georg Hager @GeorgHager · 9/9/24
I opt for "The most efficient byte is the byte you don't transfer." Much more relevant performance and energy-wise IMO

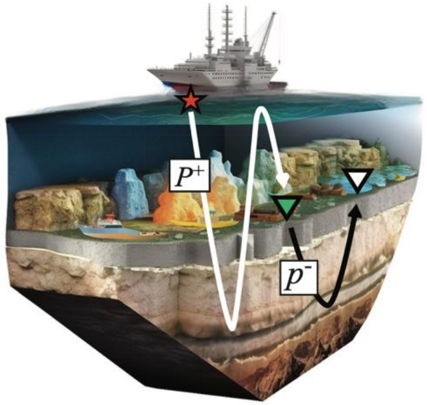
1 2 11 896

Challenging Scientific Applications



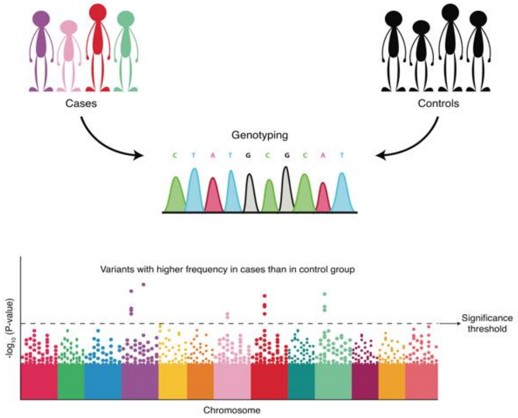
Computational Astronomy

Best Paper
PASC18



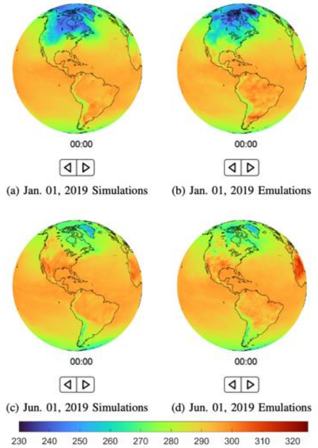
Seismic Processing

Gordon Bell Finalist
SC23



Computational Biology

Gordon Bell Finalist
SC24



Climate

Simulations / Emulations

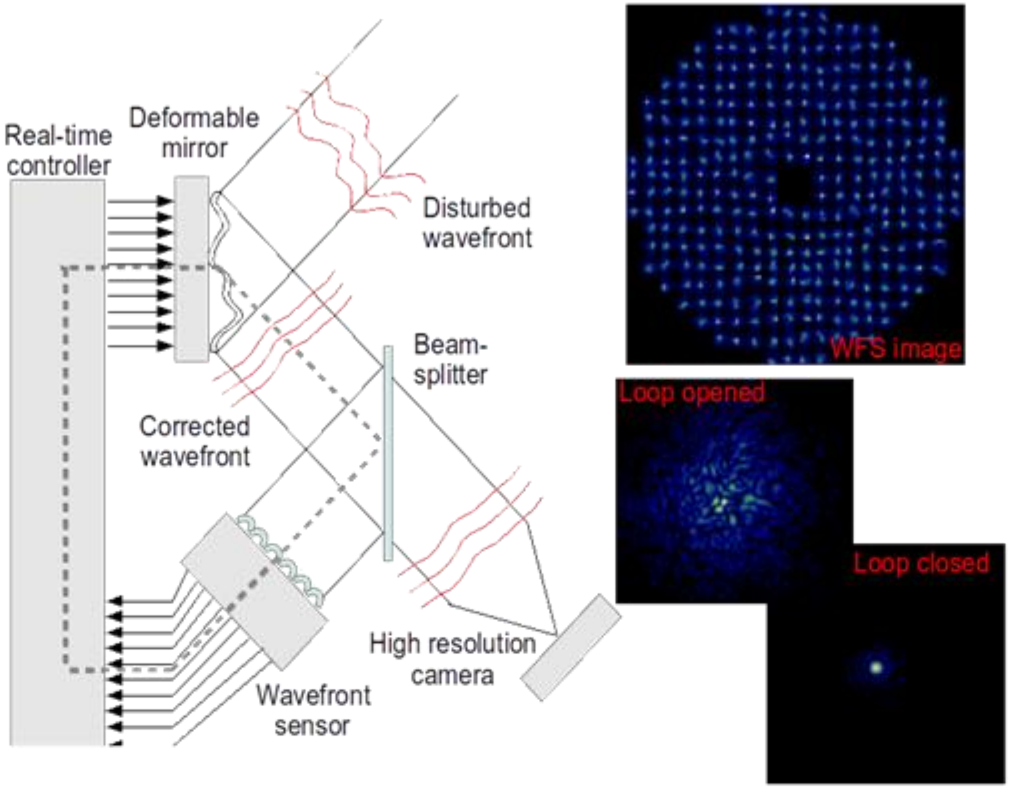
Gordon Bell Finalist SC22
Gordon Bell Winner SC24

Adaptive optics: outsmarting the atmospheric turbulence

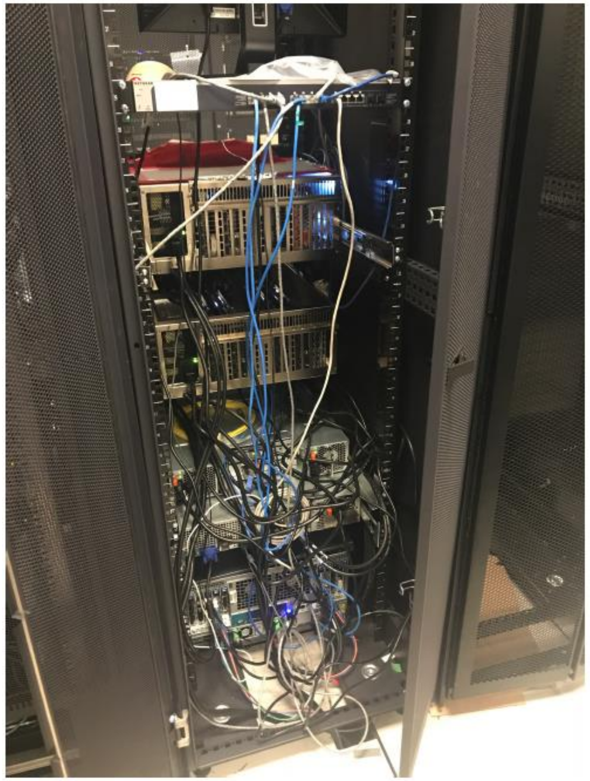
The sun observed with a compact camera

- Disturbs the trajectory of light rays
- Reduces astronomical images quality

Adaptive Optics: tomographic reconstructor, real-time controller



The Advanced Machinery

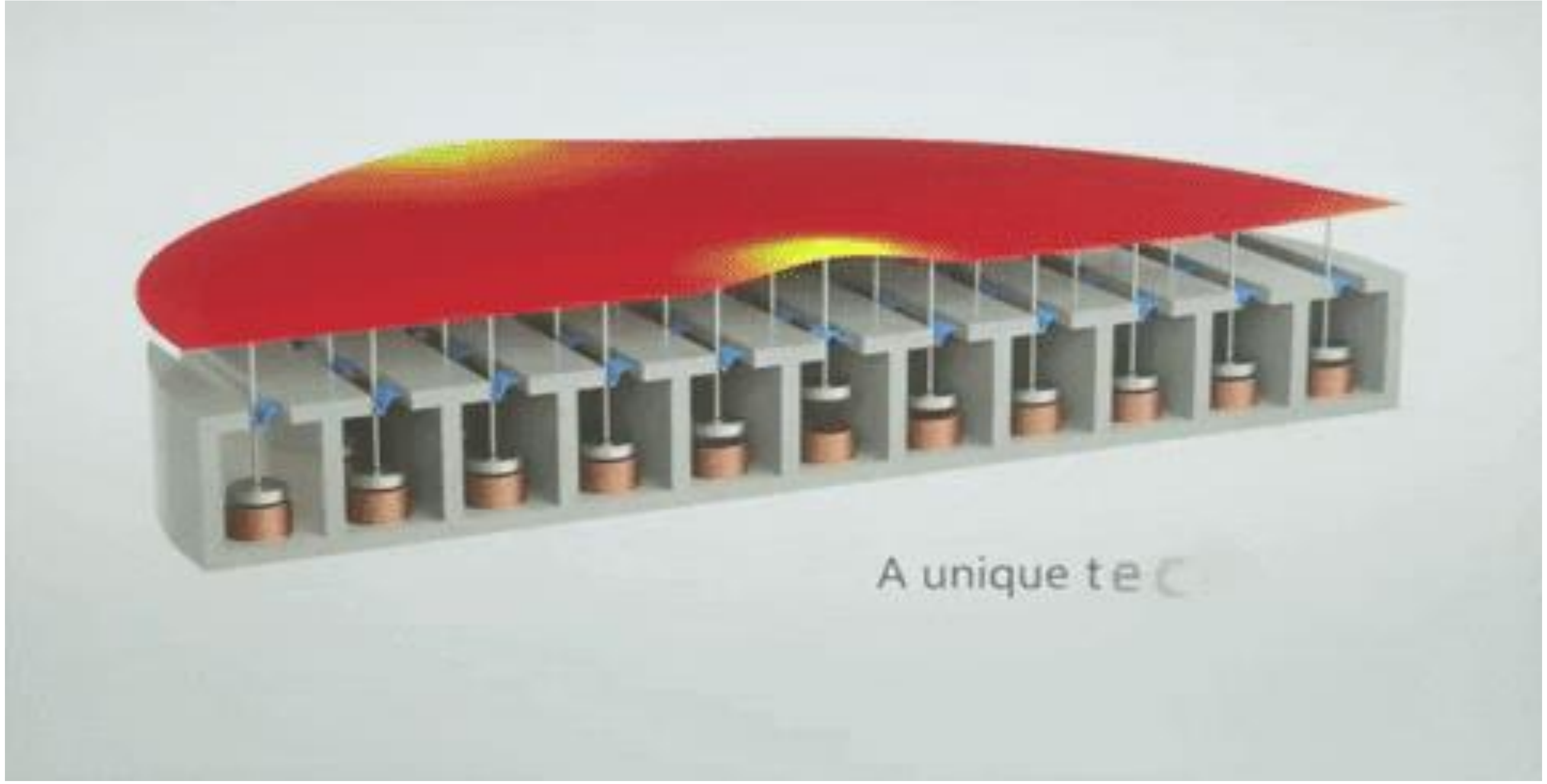


Operating at 14,000 feet

Surviving at 14,000 feet

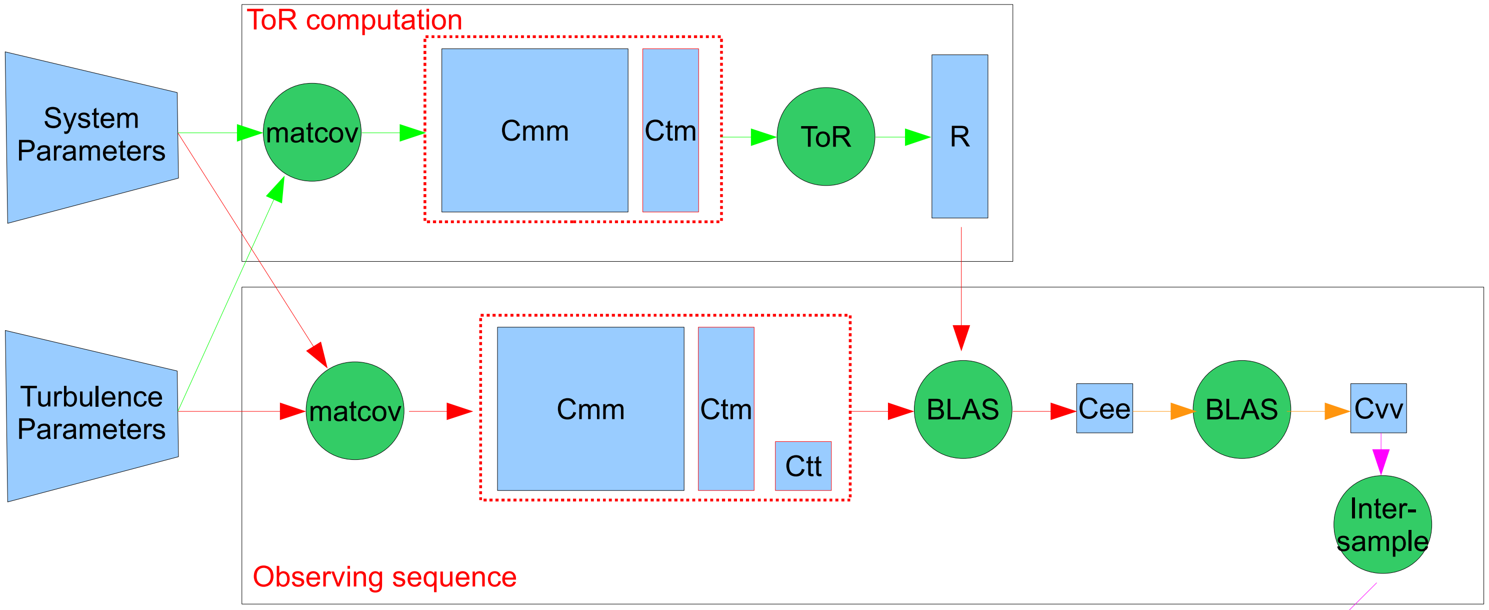
N. Doucet, HL, D. Gratadour and D. Keyes, *Mixed-Precision Tomographic Reconstructor Computations on Hardware Accelerators*, IEEE/ACM SC19 IA3 Workshop.
HL, D. Sukkari, O. Guyon, and D. Keyes, *Extreme Computing for Extreme Adaptive Optics: The Key to Finding Life Outside our Solar System*, ACM PASC18.
HL, J. Cranney, D. Gratadour, Y. Hong, L. Gatineau, and D. Keyes, *Meeting the real-time challenges of ground-based telescopes using low-rank matrix computations*, IEEE/ACM SC21.

Adaptive Optics: deformable mirrors



N. Doucet, HL, D. Gratadour and D. Keyes, *Mixed-Precision Tomographic Reconstructor Computations on Hardware Accelerators*, IEEE/ACM SC19 IA3 Workshop.
HL, D. Sukkari, O. Guyon, and D. Keyes, *Extreme Computing for Extreme Adaptive Optics: The Key to Finding Life Outside our Solar System*, ACM PASC18.
HL, J. Cranney, D. Gratadour, Y. Hong, L. Gatineau, and D. Keyes, *Meeting the real-time challenges of ground-based telescopes using low-rank matrix computations*, IEEE/ACM SC21.

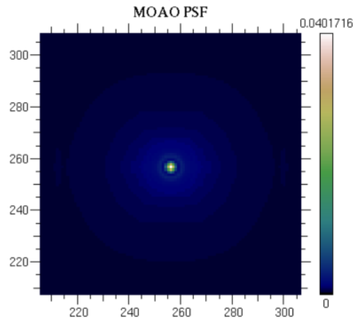
Adaptive Optics: deformable mirrors



$$R = C_{tm}.C_{mm}^{-1}$$

$$C_{ee} = C_{tt} - C_{tm}R^t - RC_{tm}^t + RC_{mm}R^t$$

$$C_{vv} = D^{\dagger}C_{ee}D^{\dagger t}$$

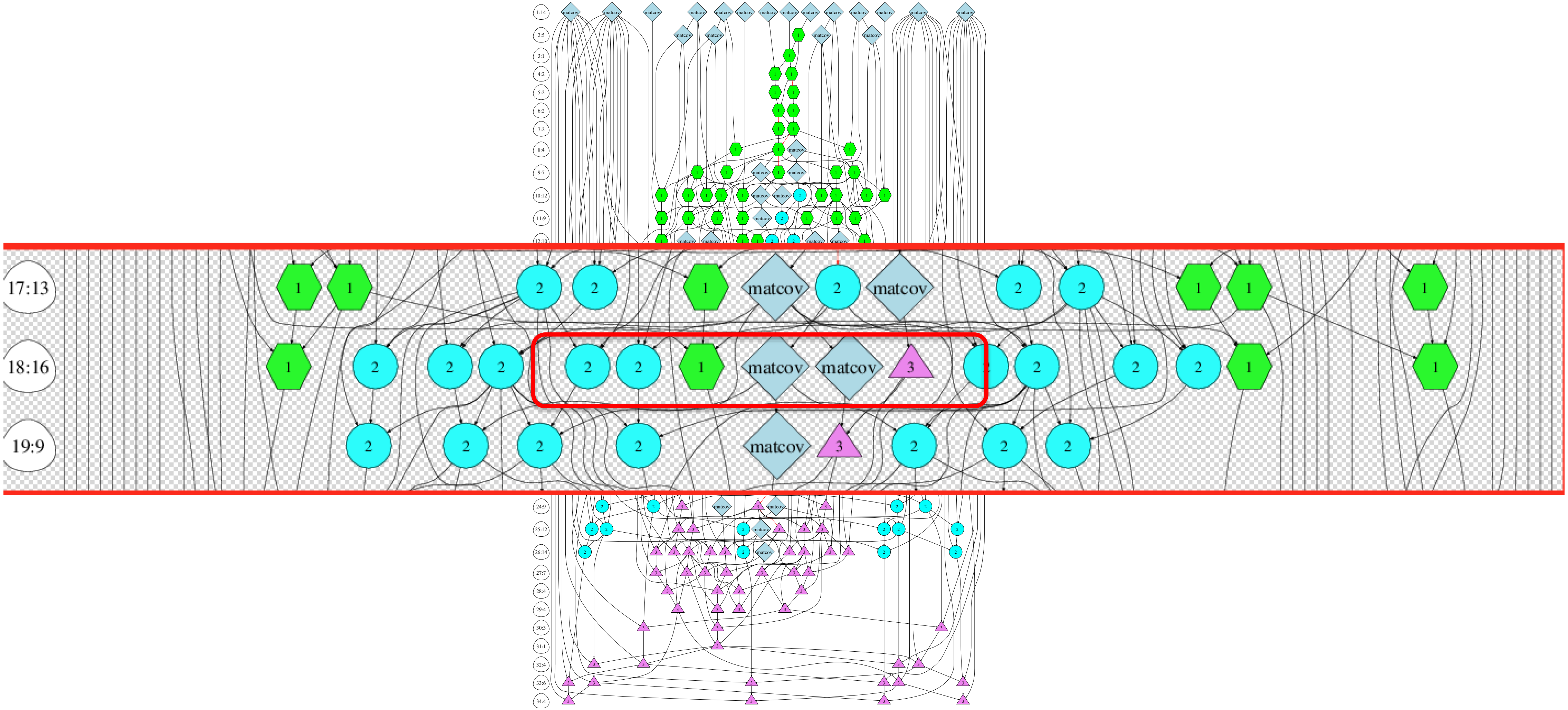


N. Doucet, HL, D. Gratadour and D. Keyes, *Mixed-Precision Tomographic Reconstructor Computations on Hardware Accelerators*, IEEE/ACM SC19 IA3 Workshop.

HL, D. Sukkari, O. Guyon, and D. Keyes, *Extreme Computing for Extreme Adaptive Optics: The Key to Finding Life Outside our Solar System*, ACM PASC18.

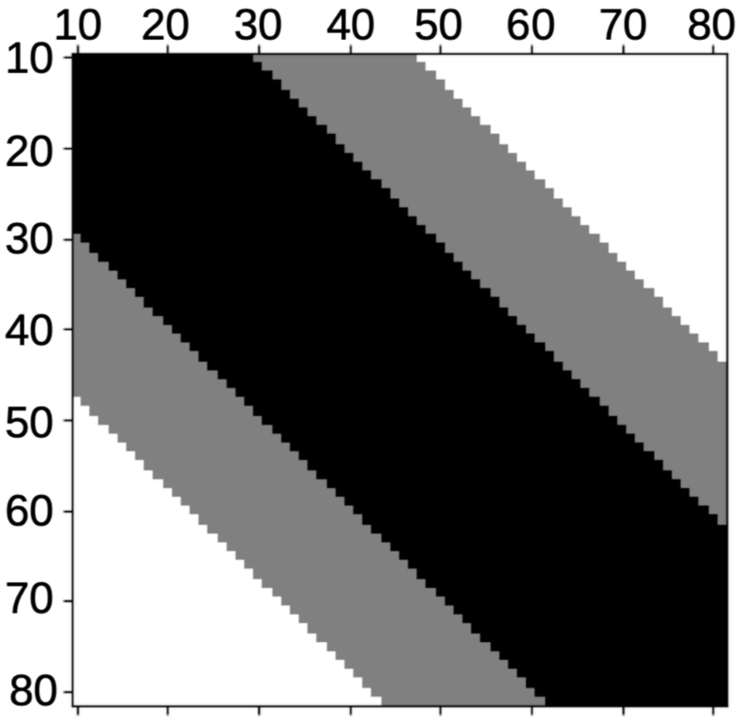
HL, J. Cranney, D. Gratadour, Y. Hong, L. Gatineau, and D. Keyes, *Meeting the real-time challenges of ground-based telescopes using low-rank matrix computations*, IEEE/ACM SC21.

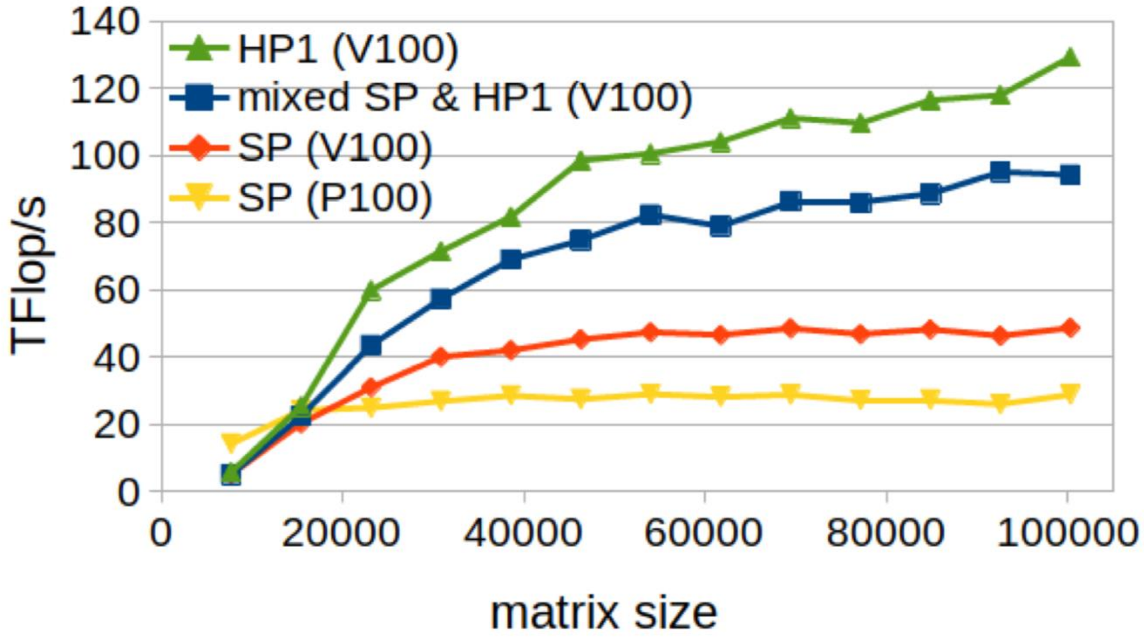
Asynchronous Many Tasks Execution



N. Doucet, HL, D. Gratadour and D. Keyes, *Mixed-Precision Tomographic Reconstructor Computations on Hardware Accelerators*, IEEE/ACM SC19 IA3 Workshop.
HL, D. Sukkari, O. Guyon, and D. Keyes, *Extreme Computing for Extreme Adaptive Optics: The Key to Finding Life Outside our Solar System*, ACM PASC18.
HL, J. Cranney, D. Gratadour, Y. Hong, L. Gatineau, and D. Keyes, *Meeting the real-time challenges of ground-based telescopes using low-rank matrix computations,* IEEE/ACM SC21.

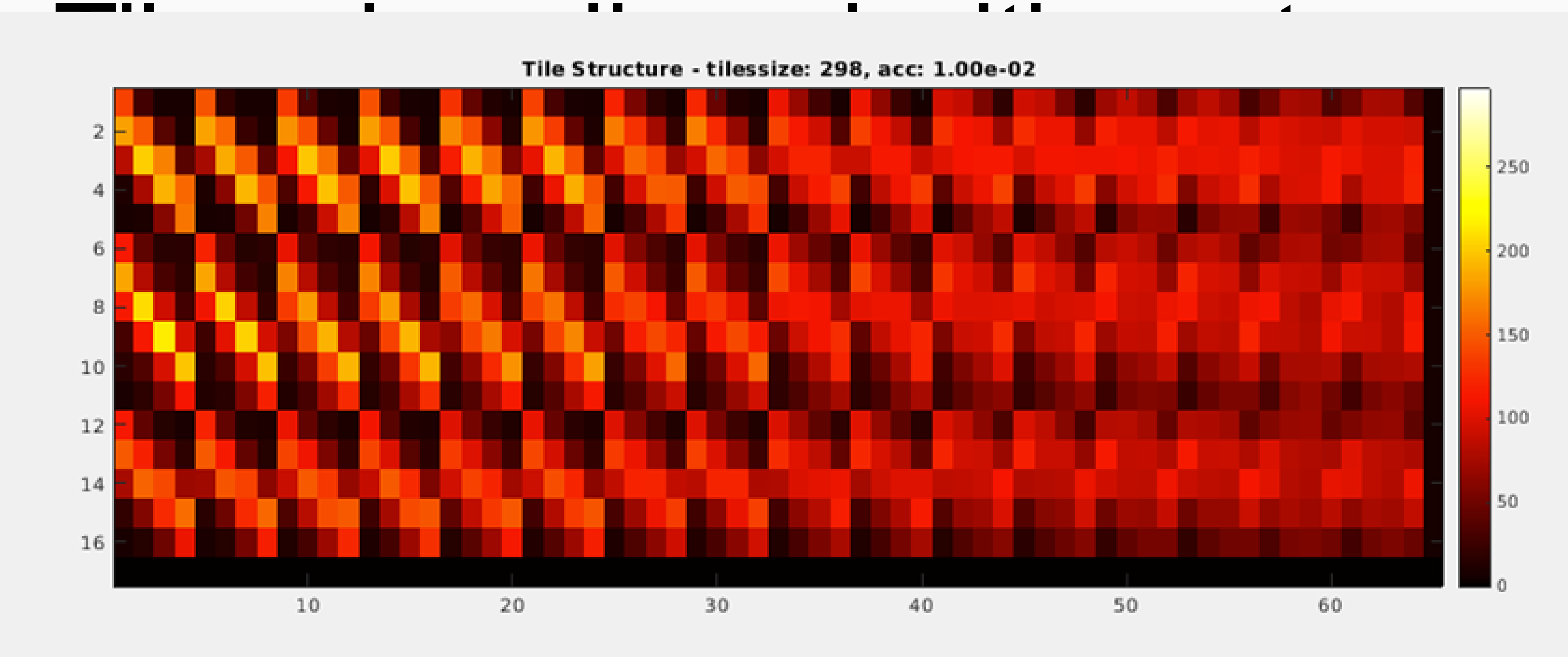
Adaptive Optics: tomographic reconstructor, real-time controller and deformable mirrors





N. Doucet, HL, D. Gratadour and D. Keyes, *Mixed-Precision Tomographic Reconstructor Computations on Hardware Accelerators*, IEEE/ACM SC19 IA3 Workshop.
HL, D. Sukkari, O. Guyon, and D. Keyes, *Extreme Computing for Extreme Adaptive Optics: The Key to Finding Life Outside our Solar System*, ACM PASC18.
HL, J. Cranney, D. Gratadour, Y. Hong, L. Gatineau, and D. Keyes, *Meeting the real-time challenges of ground-based telescopes using low-rank matrix computations,* IEEE/ACM SC21.

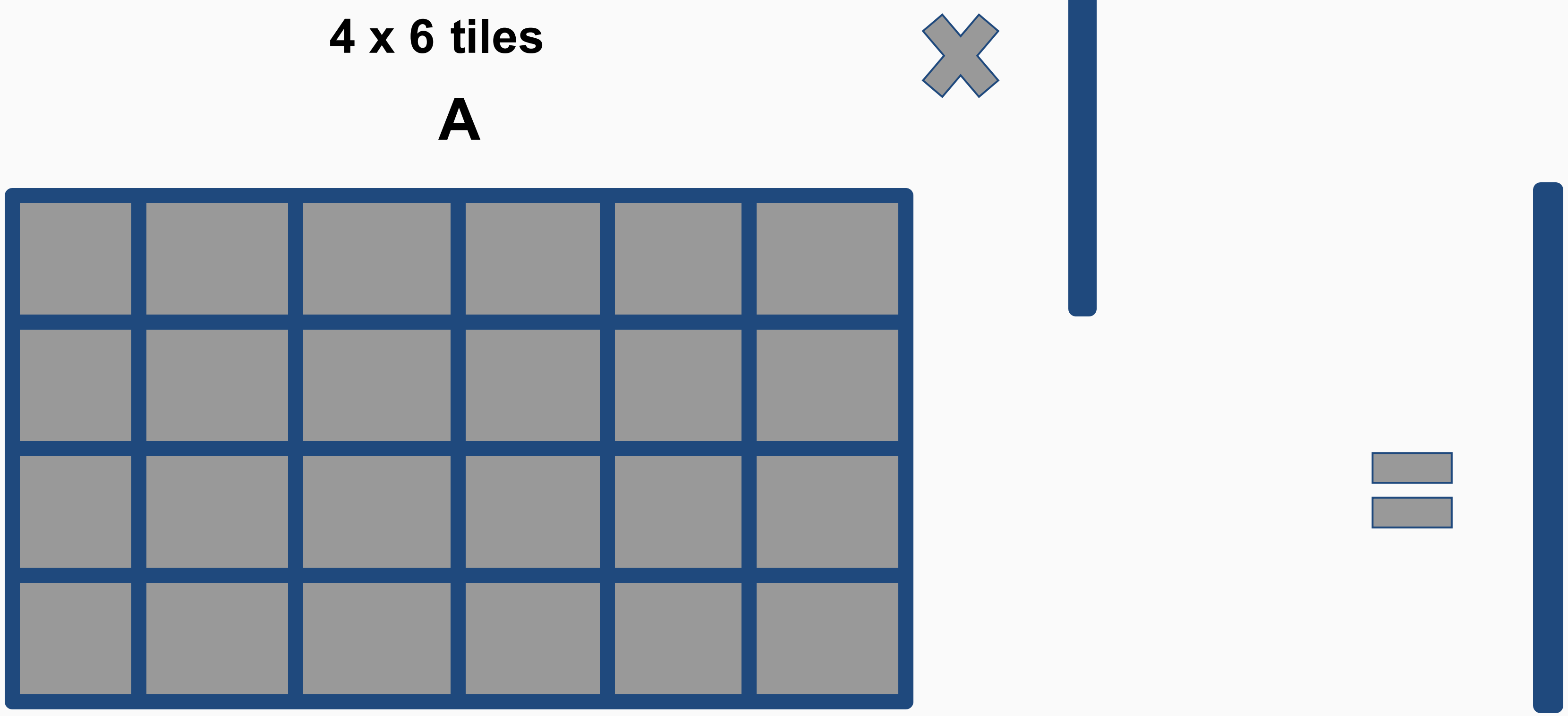
Splitting the matrix into tiles and looking at ranks



Accelerating RT Computations For Giant Telescopes

How to leverage data sparsity?

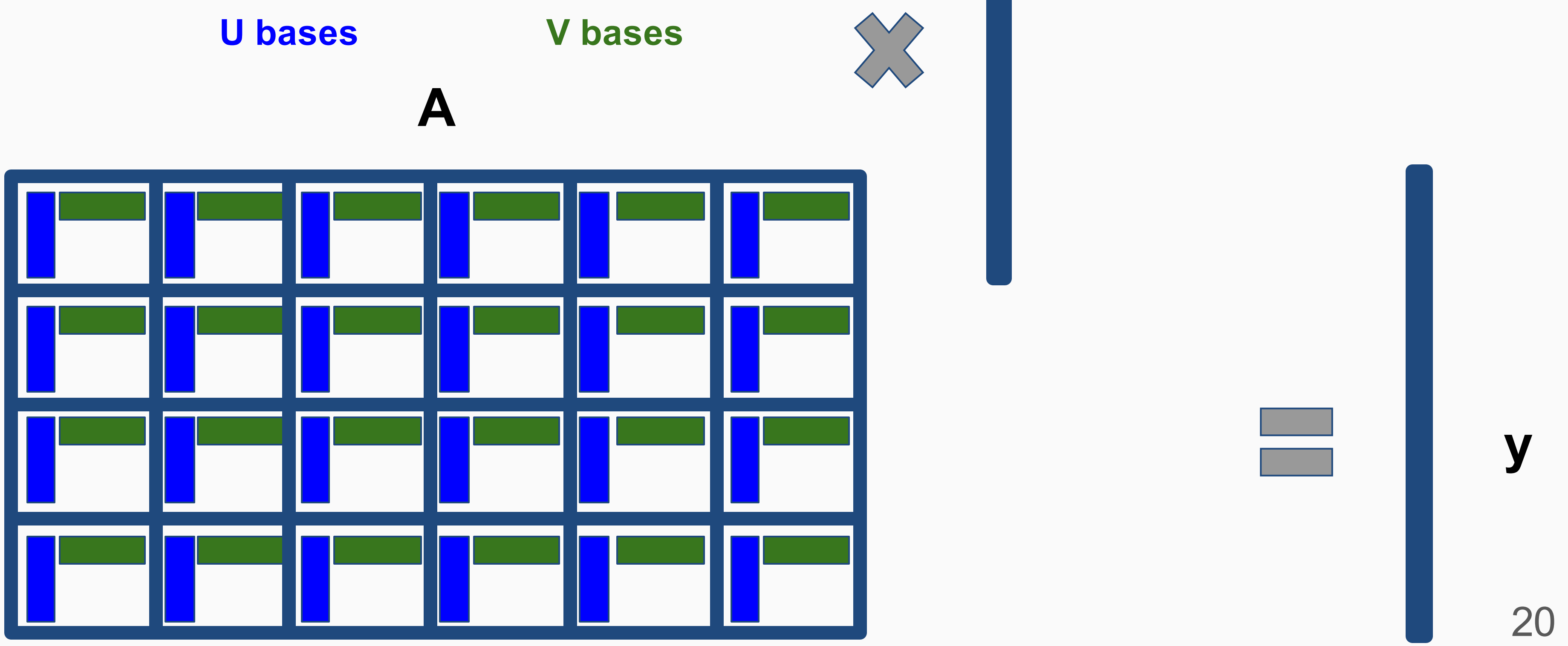
Tile Dense Matrix-Vector Multiplication



Accelerating RT Computations For Giant Telescopes

TLR-MVM

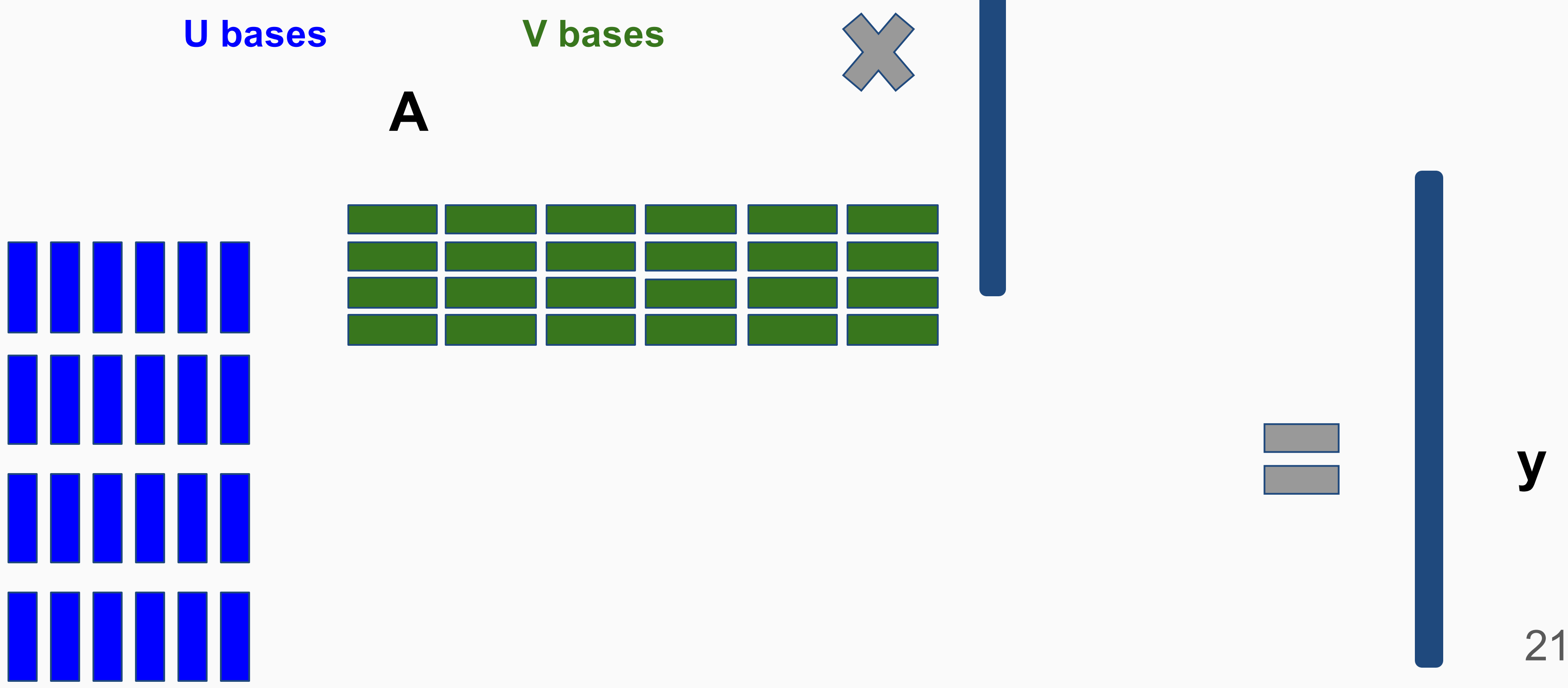
1) Compress
(SVD-like algorithms)



Accelerating RT Computations For Giant Telescopes

TLR-MVM

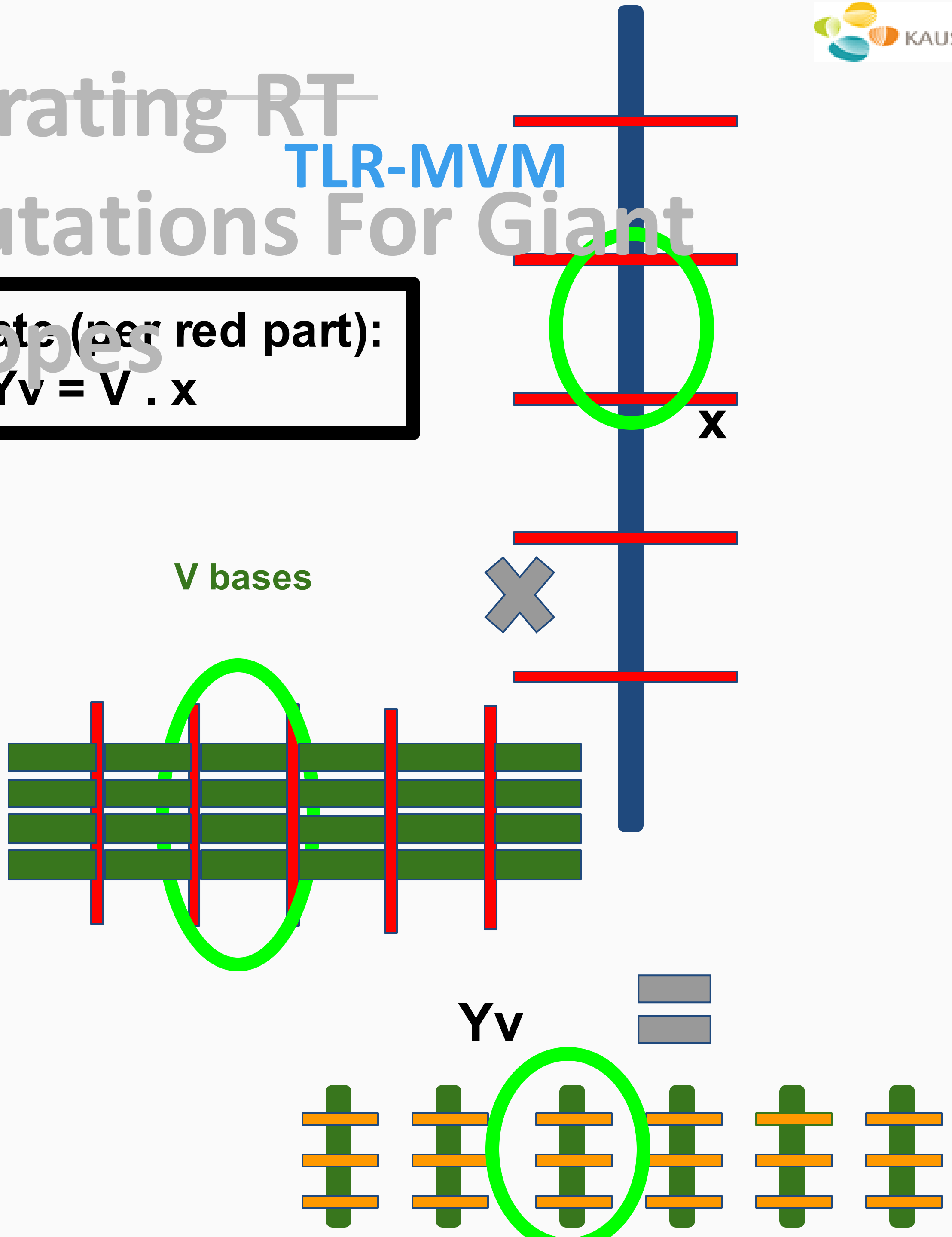
2) Stack the bases



Accelerating RT Computations For Giant Telescopes

TLR-MVM

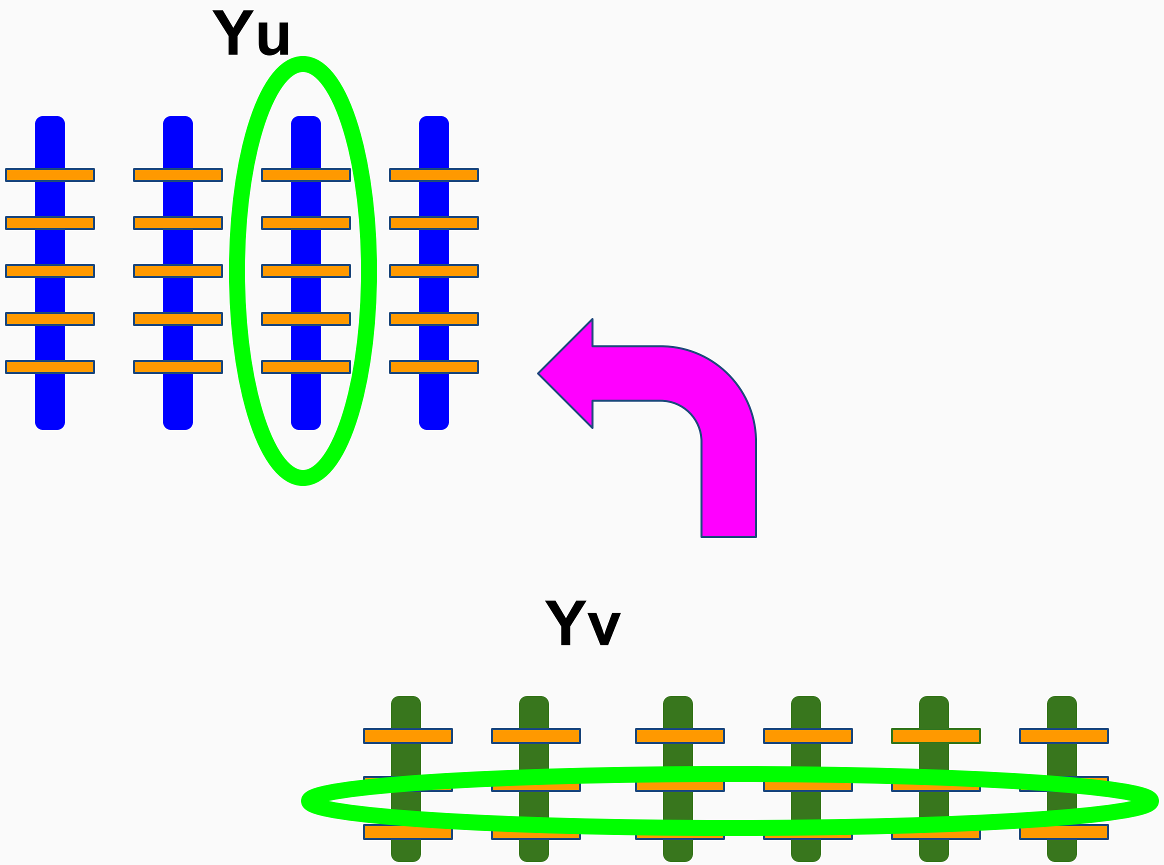
3) Calculate (per red part):
Yv = V . x



Accelerating RT Computations For Giant Telescopes

TLR-MVM

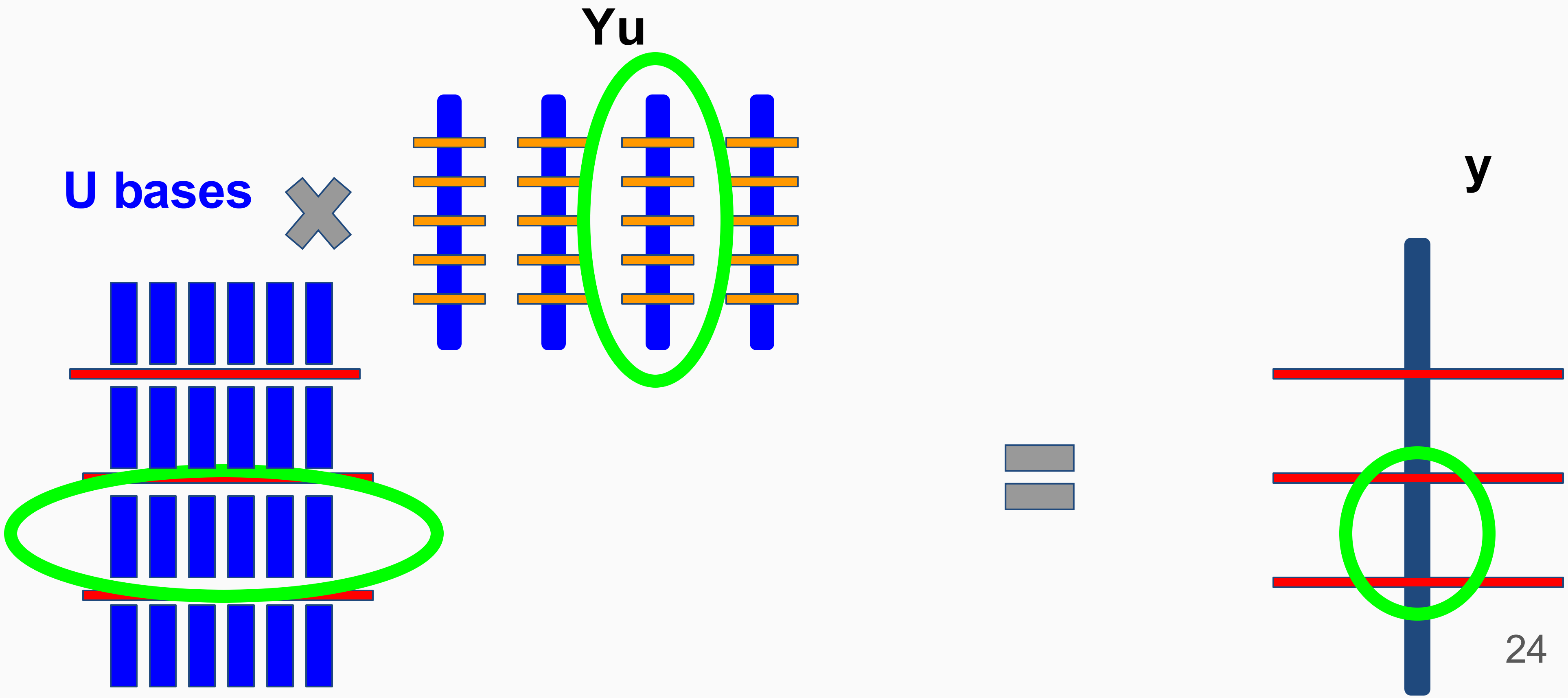
4) Translate
Yv (V bases) to Yu (U bases)



Accelerating RT Computations For Giant Telescopes

TLR-MVM

5) Calculate
y = U . Yu

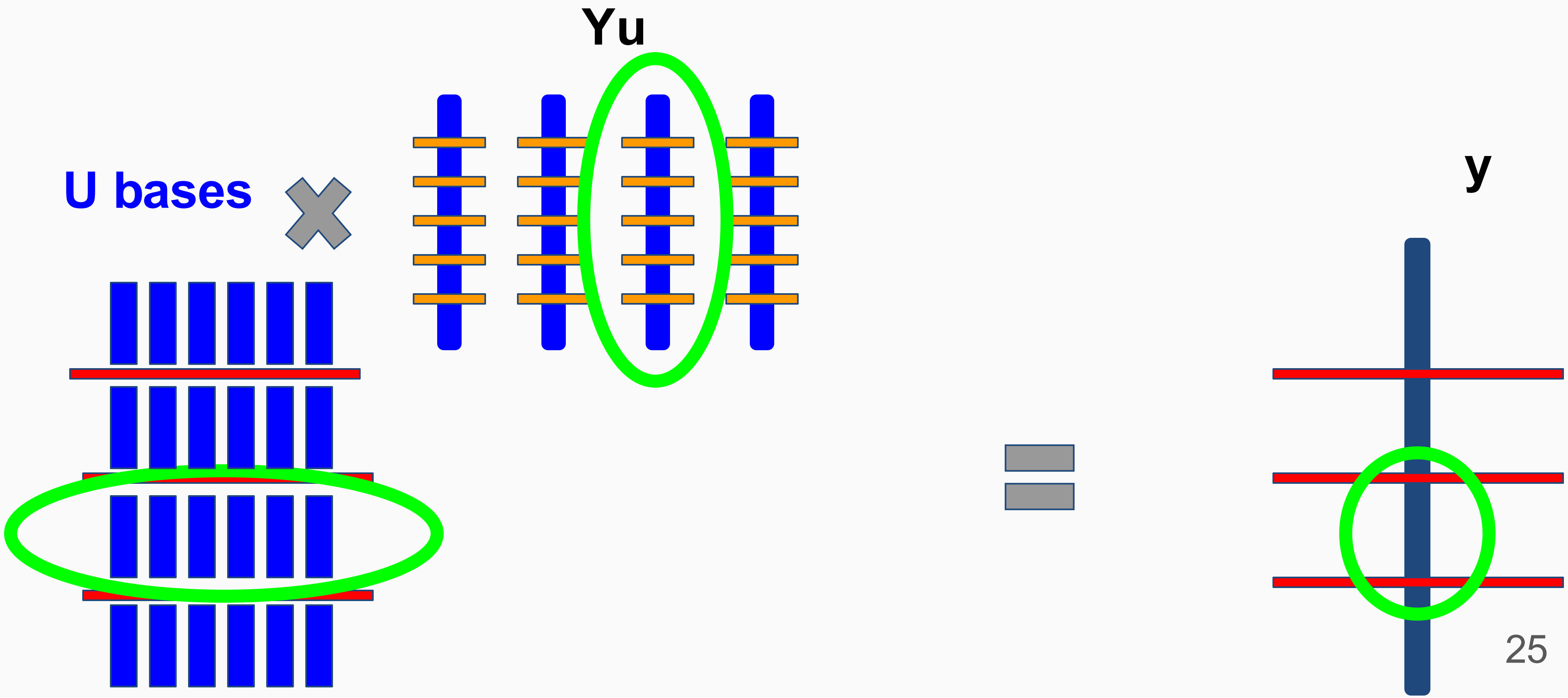


Accelerating RT Computations For Giant Telescopes

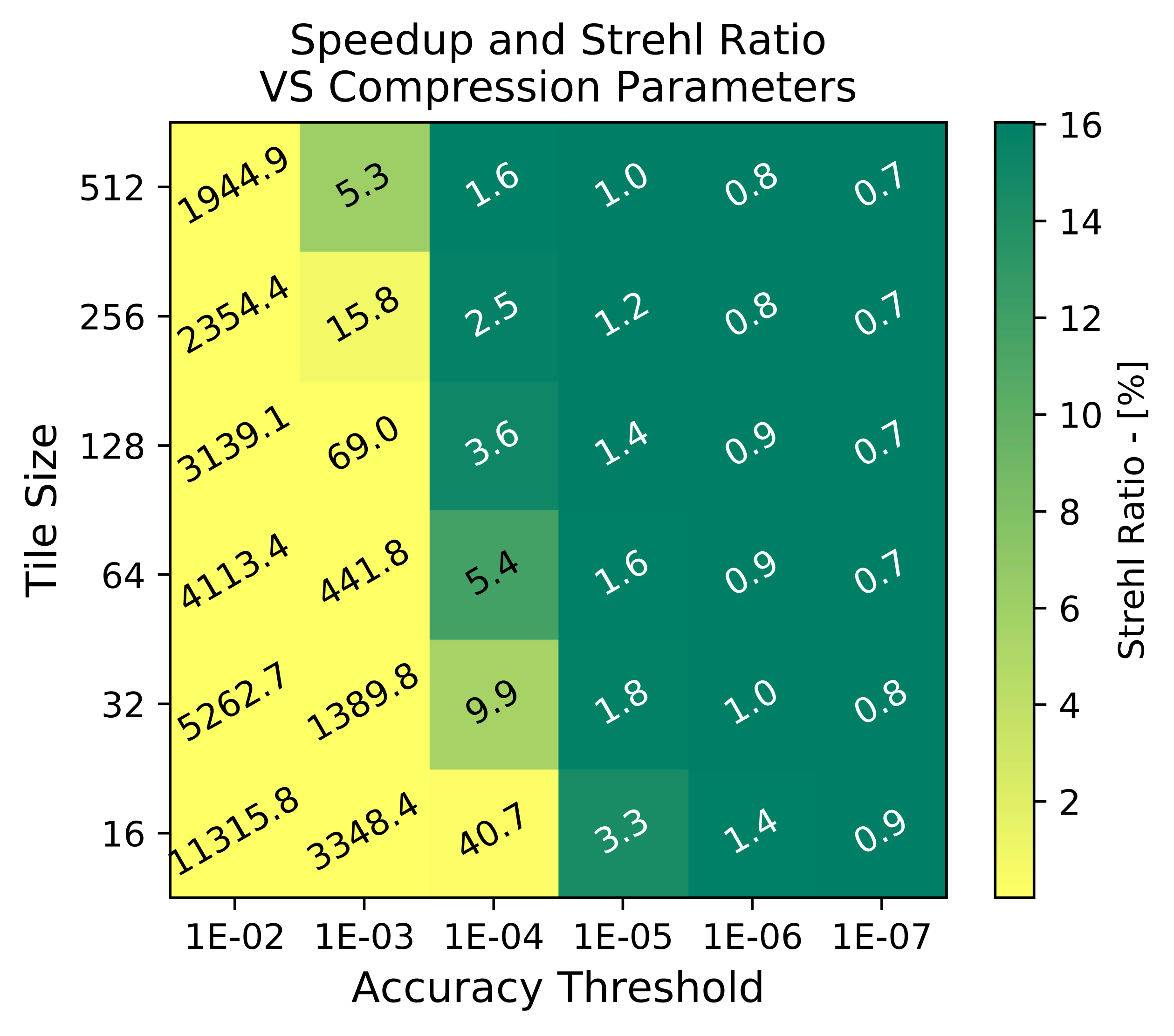
TLR-MVM

5) Calculate y = U . Yu

Rely on batch GEMV calls w/ variable sizes



Numerical Accuracy Assessment on MAVIS Datasets



Accelerating RT

Hardware / Software Specifications

Vendor	Intel	AMD		Fujitsu	NVIDIA	NEC
Family	Cascade Lake	EPYC Rome	Instinct	Primergy A64FX	Ampere GPU	SX-Aurora TSUBASA
Model	6248	7702	MI100	FX1000	A100	B300-8
Node(s)/Card(s)	1	1	1	16	1	8
Socket(s)	2	2	N/A	4	N/A	N/A
Cores	40	128	7680	48	6912	8
GHz	2.5	2.2	1.5	2.2	2.6	1.6
Memory Sustained BW	384GB DDR4 232GB/s	512GB DDR4 330GB/s	32GB HBM2 1.2TB/s	32GB HBM2 800GB/s	40GB HBM2e 1.5TB/s	48GB HBM2 1.5TB/s
LLC Sustained BW	27.5MB 1.1TB/s	**512MB** 4TB/s	8MB 3TB/s	32MB 3.6TB/s	40MB 4.8TB/s	16MB 2.1TB/s
Compiler	Intel compiler 19.1.0	GCC compiler 8.2.0		Fujitsu compiler 4.5.0	NVCC 11.0	NEC compiler 3.1.1
BLAS library	Intel MKL 2020	BLIS 3.0.0		Fujitsu SSL II	cuBLAS 11.0	NEC NLC 2.1.0
MPI library	OpenMPI 4.0.3	OpenMPI 3.1.2		Fujitsu MPI 4.0.1	NCCL 2.0	NEC MPI 2.13.0

Accelerating RT

Hardware / Software Specifications

Vendor	Intel	AMD		Fujitsu	NVIDIA	NEC
Family	Cascade Lake	EPYC Rome	Instinct	Primergy A64FX	Ampere GPU	SX-Aurora TSUBASA
Model	6248	7702	MI100	FX1000	A100	B300-8
Node(s)/Card(s)	1	1	1	16	1	8
Socket(s)	2	2	N/A	4	N/A	N/A
Cores	40	128	7680	48	6912	8
GHz	2.5	2.2	1.5	2.2	2.6	1.6
Memory Sustained BW	384GB DDR4 232GB/s	512GB DDR4 330GB/s	32GB HBM2 1.2TB/s	32GB HBM2 800GB/s	40GB HBM2e 1.5TB/s	48GB HBM2 1.5TB/s
LLC Sustained BW	27.5MB 1.1TB/s	**512MB** 4TB/s	8MB 3TB/s	32MB 3.6TB/s	40MB 4.8TB/s	16MB 2.1TB/s
Compiler	Intel compiler 19.1.0	GCC compiler 8.2.0		Fujitsu compiler 4.5.0	NVCC 11.0	NEC compiler 3.1.1
BLAS library	Intel MKL 2020	BLIS 3.0.0		Fujitsu SSL II	cuBLAS 11.0	NEC NLC 2.1.0
MPI library	OpenMPI 4.0.3	OpenMPI 3.1.2		Fujitsu MPI 4.0.1	NCCL 2.0	NEC MPI 2.13.0

x86

MPI + OpenMP

Accelerating RT

Hardware / Software Specifications

Vendor	Intel	AMD		Fujitsu	NVIDIA	NEC
Family	Cascade Lake	EPYC Rome	Instinct	Primergy A64FX	Ampere GPU	SX-Aurora TSUBASA
Model	6248	7702	MI100	FX1000	A100	B300-8
Node(s)/Card(s)	1	1	1	16	1	8
Socket(s)	2	2	N/A	4	N/A	N/A
Cores	40	128	7680	48	6912	8
GHz	2.5	2.2	1.5	2.2	2.6	1.6
Memory Sustained BW	384GB DDR4 232GB/s	512GB DDR4 330GB/s	32GB HBM2 1.2TB/s	32GB HBM2 800GB/s	40GB HBM2e 1.5TB/s	48GB HBM2 1.5TB/s
LLC Sustained BW	27.5MB 1.1TB/s	**512MB** 4TB/s	8MB 3TB/s	32MB 3.6TB/s	40MB 4.8TB/s	16MB 2.1TB/s
Compiler	Intel compiler 19.1.0	GCC compiler 8.2.0		Fujitsu compiler 4.5.0	NVCC 11.0	NEC compiler 3.1.1
BLAS library	Intel MKL 2020	BLIS 3.0.0		Fujitsu SSL II	cuBLAS 11.0	NEC NLC 2.1.0
MPI library	OpenMPI 4.0.3	OpenMPI 3.1.2		Fujitsu MPI 4.0.1	NCCL 2.0	NEC MPI 2.13.0

ARM

MPI + OpenMP

Accelerating RT

Hardware / Software Specifications

Vendor	Intel	AMD		Fujitsu	NVIDIA	NEC
Family	Cascade Lake	EPYC Rome	Instinct	Primergy A64FX	Ampere GPU	SX-Aurora TSUBASA
Model	6248	7702	MI100	FX1000	A100	B300-8
Node(s)/Card(s)	1	1	1	16	1	8
Socket(s)	2	2	N/A	4	N/A	N/A
Cores	40	128	7680	48	6912	8
GHz	2.5	2.2	1.5	2.2	2.6	1.6
Memory Sustained BW	384GB DDR4 232GB/s	512GB DDR4 330GB/s	32GB HBM2 1.2TB/s	32GB HBM2 800GB/s	40GB HBM2e 1.5TB/s	48GB HBM2 1.5TB/s
LLC Sustained BW	27.5MB 1.1TB/s	**512MB** 4TB/s	8MB 3TB/s	32MB 3.6TB/s	40MB 4.8TB/s	16MB 2.1TB/s
Compiler	Intel compiler 19.1.0	GCC compiler 8.2.0		Fujitsu compiler 4.5.0	NVCC 11.0	NEC compiler 3.1.1
BLAS library	Intel MKL 2020	BLIS 3.0.0		Fujitsu SSL II	cuBLAS 11.0	NEC NLC 2.1.0
MPI library	OpenMPI 4.0.3	OpenMPI 3.1.2		Fujitsu MPI 4.0.1	NCCL 2.0	NEC MPI 2.13.0

Accelerators

ROCm / CUDA

Accelerating RT

Hardware / Software Specifications

Vendor	Intel	AMD		Fujitsu	NVIDIA	NEC
Family	Cascade Lake	EPYC Rome	Instinct	Primergy A64FX	Ampere GPU	SX-Aurora TSUBASA
Model	6248	7702	MI100	FX1000	A100	B300-8
Node(s)/Card(s)	1	1	1	16	1	8
Socket(s)	2	2	N/A	4	N/A	N/A
Cores	40	128	7680	48	6912	8
GHz	2.5	2.2	1.5	2.2	2.6	1.6
Memory Sustained BW	384GB DDR4 232GB/s	512GB DDR4 330GB/s	32GB HBM2 1.2TB/s	32GB HBM2 800GB/s	40GB HBM2e 1.5TB/s	48GB HBM2 1.5TB/s
LLC Sustained BW	27.5MB 1.1TB/s	**512MB** 4TB/s	8MB 3TB/s	32MB 3.6TB/s	40MB 4.8TB/s	16MB 2.1TB/s
Compiler	Intel compiler 19.1.0	GCC compiler 8.2.0		Fujitsu compiler 4.5.0	NVCC 11.0	NEC compiler 3.1.1
BLAS library	Intel MKL 2020	BLIS 3.0.0		Fujitsu SSL II	cuBLAS 11.0	NEC NLC 2.1.0
MPI library	OpenMPI 4.0.3	OpenMPI 3.1.2		Fujitsu MPI 4.0.1	NCCL 2.0	NEC MPI 2.13.0

Vector

MPI + OpenMP

Accelerating RT Computations For Giant

Hardware / Software Specifications

Vendor	Intel	AMD		Fujitsu	NVIDIA	NEC
Family	Cascade Lake	EPYC Rome	Instinct	Primergy A64FX	Ampere GPU	SX-Aurora TSUBASA
Model	6248	7702	MI100	FX1000	A100	B300-8
Node(s)/Card(s)	1	1	1	16	1	8
Socket(s)	2	2	N/A	4	N/A	N/A
Cores	40	128	7680	48	6912	8
GHz	2.5	2.2	1.5	2.2	2.6	1.6
Memory Sustained BW	384GB DDR4 232GB/s	512GB DDR4 330GB/s	32GB HBM2 1.2TB/s	32GB HBM2 800GB/s	40GB HBM2e 1.5TB/s	48GB HBM2 1.5TB/s
LLC Sustained BW	27.5MB 1.1TB/s	**512MB** 4TB/s	8MB 3TB/s	32MB 3.6TB/s	40MB 4.8TB/s	16MB 2.1TB/s
Compiler	Intel compiler 19.1.0	GCC compiler 8.2.0		Fujitsu compiler 4.5.0	NVCC 11.0	NEC compiler 3.1.1
BLAS library	Intel MKL 2020	BLIS 3.0.0		Fujitsu SSL II	cuBLAS 11.0	NEC NLC 2.1.0
MPI library	OpenMPI 4.0.3	OpenMPI 3.1.2		Fujitsu MPI 4.0.1	NCCL 2.0	NEC MPI 2.13.0

HBM

Accelerating RT Computations For Giant Telescopes

Hardware / Software Specifications

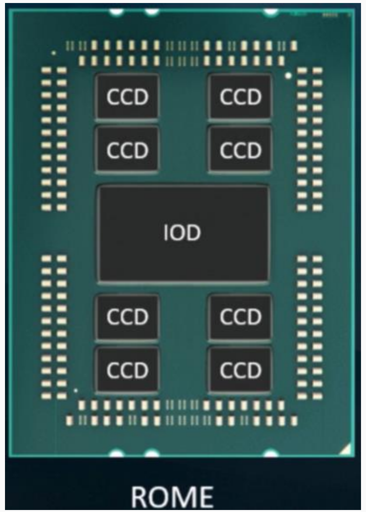
Vendor	Intel	AMD	Fujitsu	NEC	NVIDIA	Graphcore
Family	Cascade Lake	EPYC Milan	Primergy A64FX	SX-Aurora TSUBASA	Ampere GPU	IPU
Model	6248	7713	FX1000	B300-8	A100	Bow
Node(s)/Card(s)	1	1	16	8	1	1
Socket(s)	2	2	4	N/A	N/A	1
Cores	40	128	48	8	6912	1472
GHz	2.5	2.0	2.2	1.6	2.6	1.85
Memory Sustained BW	384GB DDR4 232GB/s	512GB DDR4 330GB/s	32GB HBM 800GB/s	48GB HBM2 1.5TB/s	40GB HBM2e 1.5TB/s	3.6GB 261TB/s
LLC Sustained BW	27.5MB 1.1TB/s	512MB 4TB/s	32MB 3.6TB/s	16MB 2.1TB/s	40MB 4.8TB/s	N/A
Compiler	Intel 19.1.0	GCC 7.5.0	Fujitsu 4.5.0	NEC 3.1.1	NVCC 11.0	POPLAR 2.6
BLAS library	Intel MKL 2020	BLIS 3.0.0	Fujitsu SSL II	NEC NLC 2.1.0	cuBLAS 11.0	N/A
MPI library	OpenMPI 4.0.3	OpenMPI 3.1.2	Fujitsu MPI 4.0.1	NEC MPI 2.13.0	NCCL 2.0	N/A

x86 - ARM - Vector
MPI + OpenMP

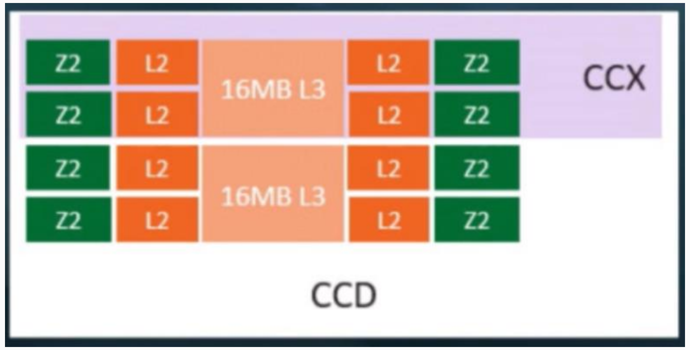
GPU
CUDA

Hardware / Software Specifications

Vendor	Intel	AMD		Fujitsu	NVIDIA	NEC
Family	Cascade Lake	EPYC Rome	Instinct	Primergy A64FX	Ampere GPU	SX-Aurora TSUBASA
Model	6248	7702	MI100	FX1000	A100	B300-8
Node(s)/Card(s)	1	1	1	16	1	8
Socket(s)	2	2	N/A	4	N/A	N/A
Cores	40	128	7680	48	6912	8
GHz	2.5	2.2	1.5	2.2	2.6	1.6
Memory Sustained BW	384GB DDR4 232GB/s	512GB DDR4 330GB/s	32GB HBM2 1.2TB/s	32GB HBM2 800GB/s	40GB HBM2e 1.5TB/s	48GB HBM2 1.5TB/s
LLC Sustained BW	27.5MB 1.1TB/s	**512MB** 4TB/s	8MB 3TB/s	32MB 3.6TB/s	40MB 4.8TB/s	16MB 2.1TB/s
Compiler	Intel compiler 19.1.0	GCC compiler 8.2.0		Fujitsu compiler 4.5.0	NVCC 11.0	NEC compiler 3.1.1
BLAS library	Intel MKL 2020	BLIS 3.0.0		Fujitsu SSL II	cuBLAS 11.0	NEC NLC 2.1.0
MPI library	OpenMPI 4.0.3	OpenMPI 3.1.2		Fujitsu MPI 4.0.1	NCCL 2.0	NEC MPI 2.13.0



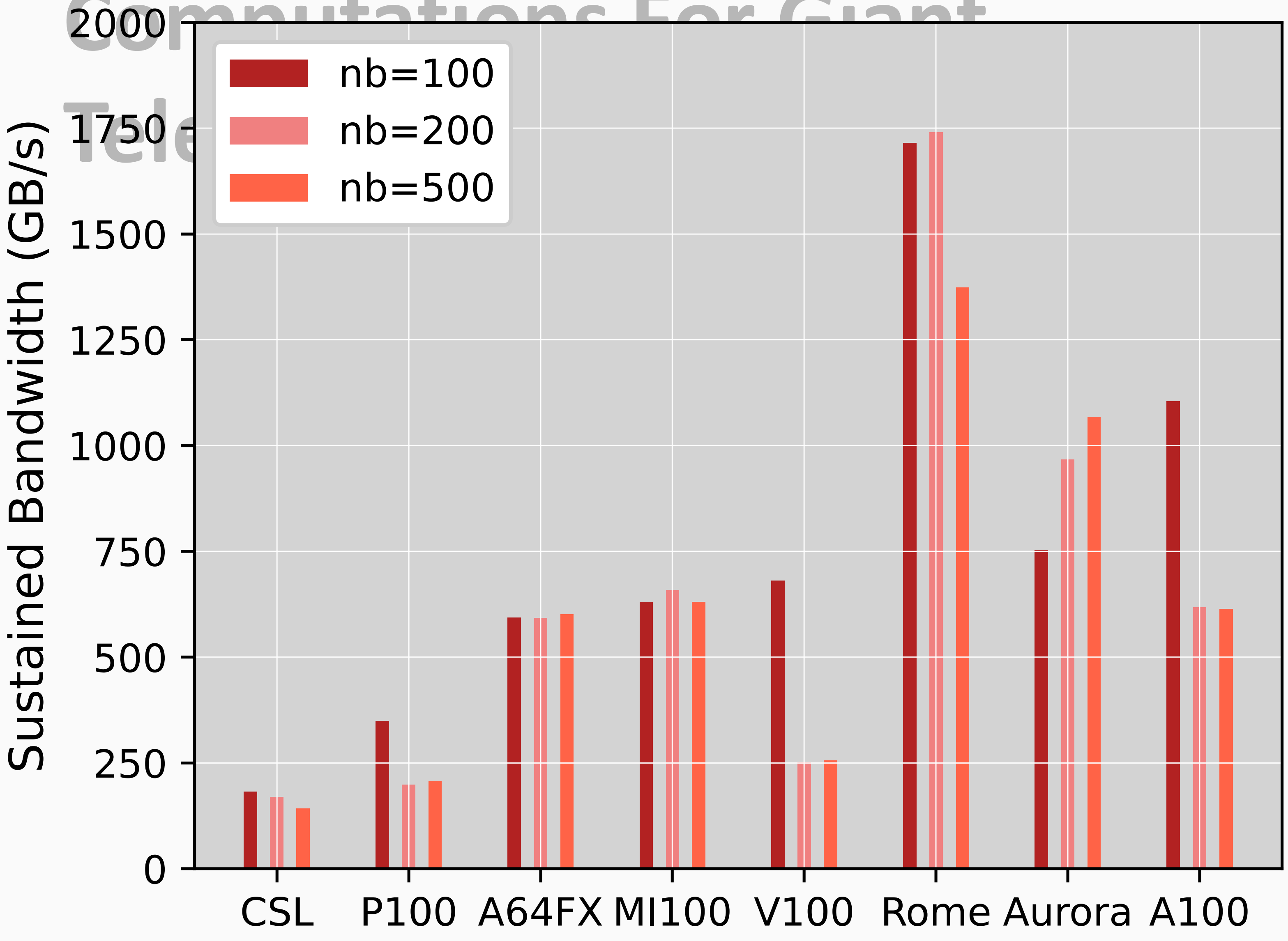
- IOD = “I/O Die” doing all the memory/PCI/other socket traffic
- CCD = “Core Compute Die”, a chiplet having compute cores only
- CCX = “Core Compute CompleX”, a set of cores sharing a L3 cache



Accelerating RT Computations For Giant Tele

Sustained Bandwidth on Synthetic Datasets

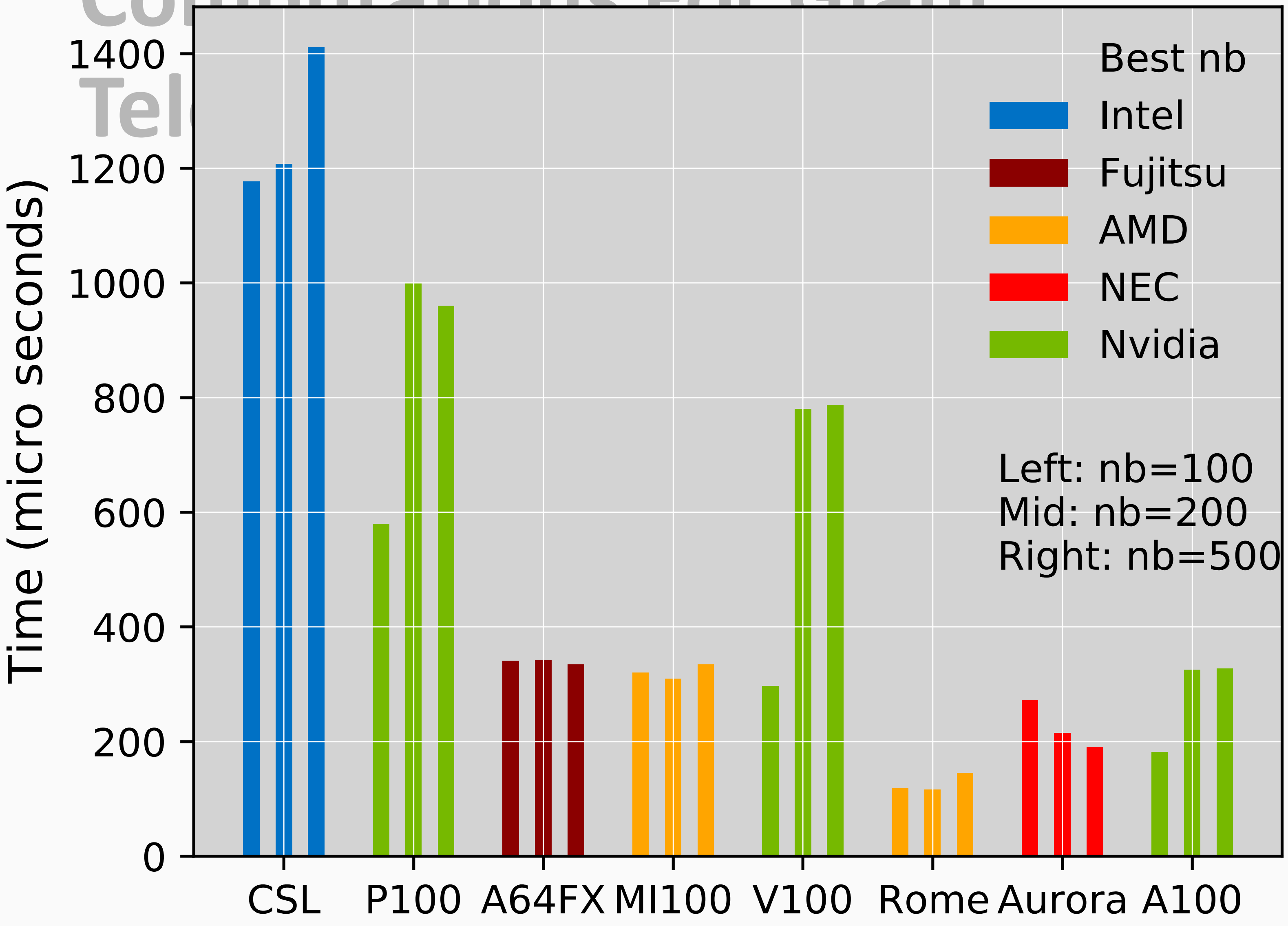
Rely on batch GEMV calls w/ constant sizes



Higher is better

Accelerating RT Computations For Giant Tel

Time to Solution on Synthetic Datasets

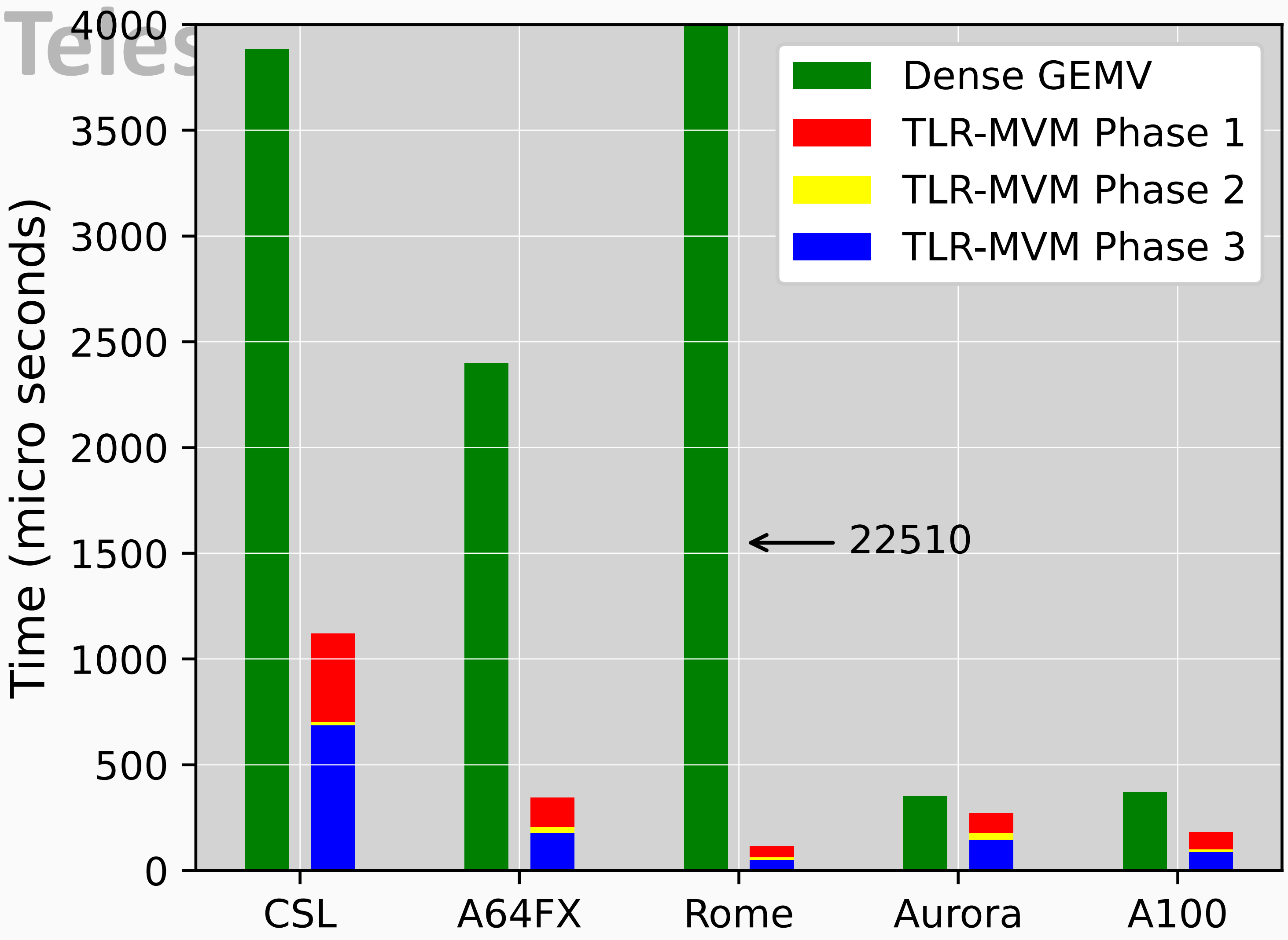


Rely on batch GEMV calls
w/ constant sizes

Lower is better

Accelerating RT Computations For Giant Teles

Dense Vs TLR MVM: Time Breakdown on Synthetic Datasets

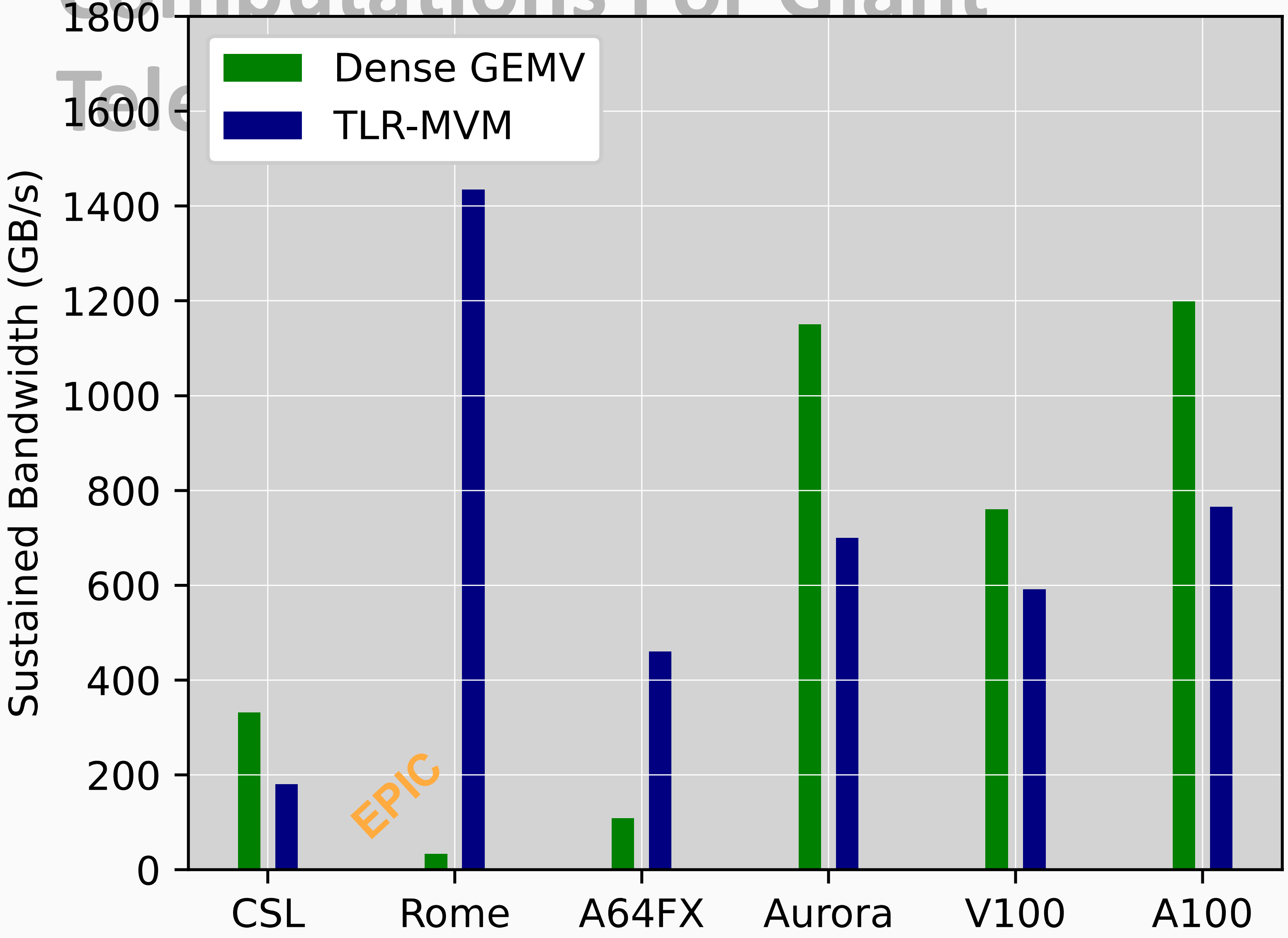


Rely on batch GEMV calls w/ constant sizes

Lower is better

Accelerating RT Computations For Giant Tele

Sustained Bandwidth on MAVIS Datasets

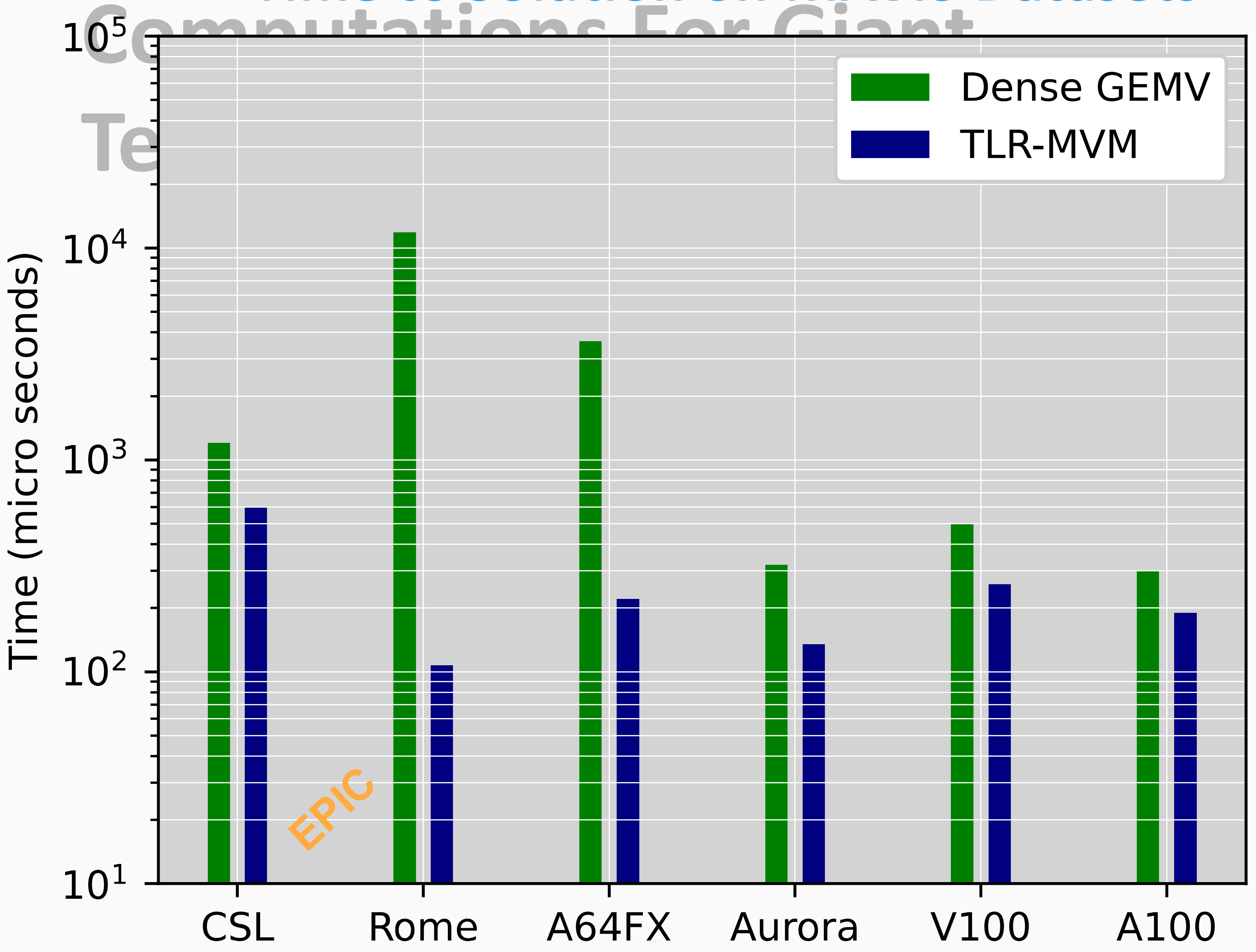


Rely on batch GEMV calls w/ variable sizes

Higher is better

Accelerating RT Computations For Giant Te

Time to Solution on MAVIS Datasets

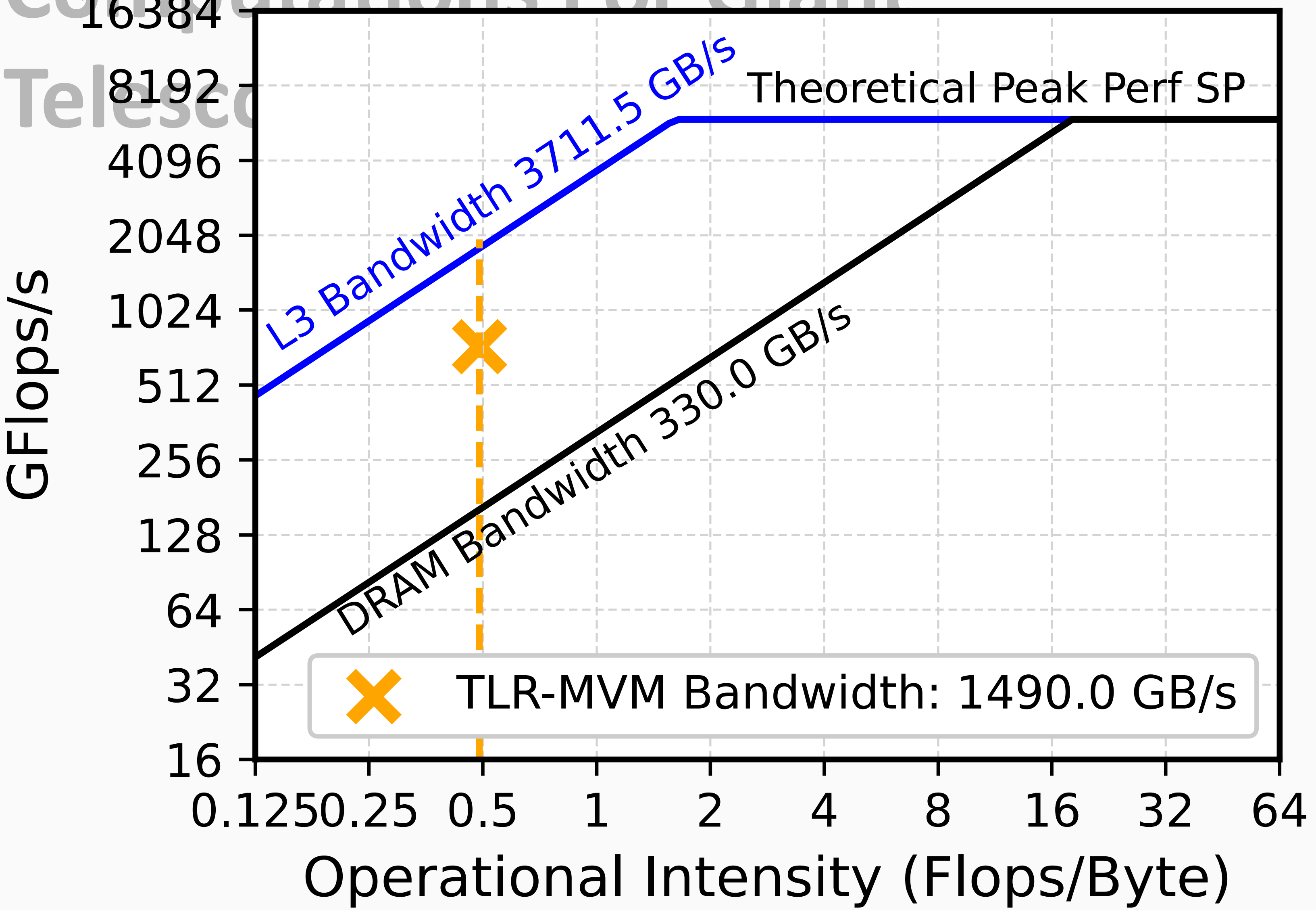


Rely on batch GEMV calls
w/ variable sizes

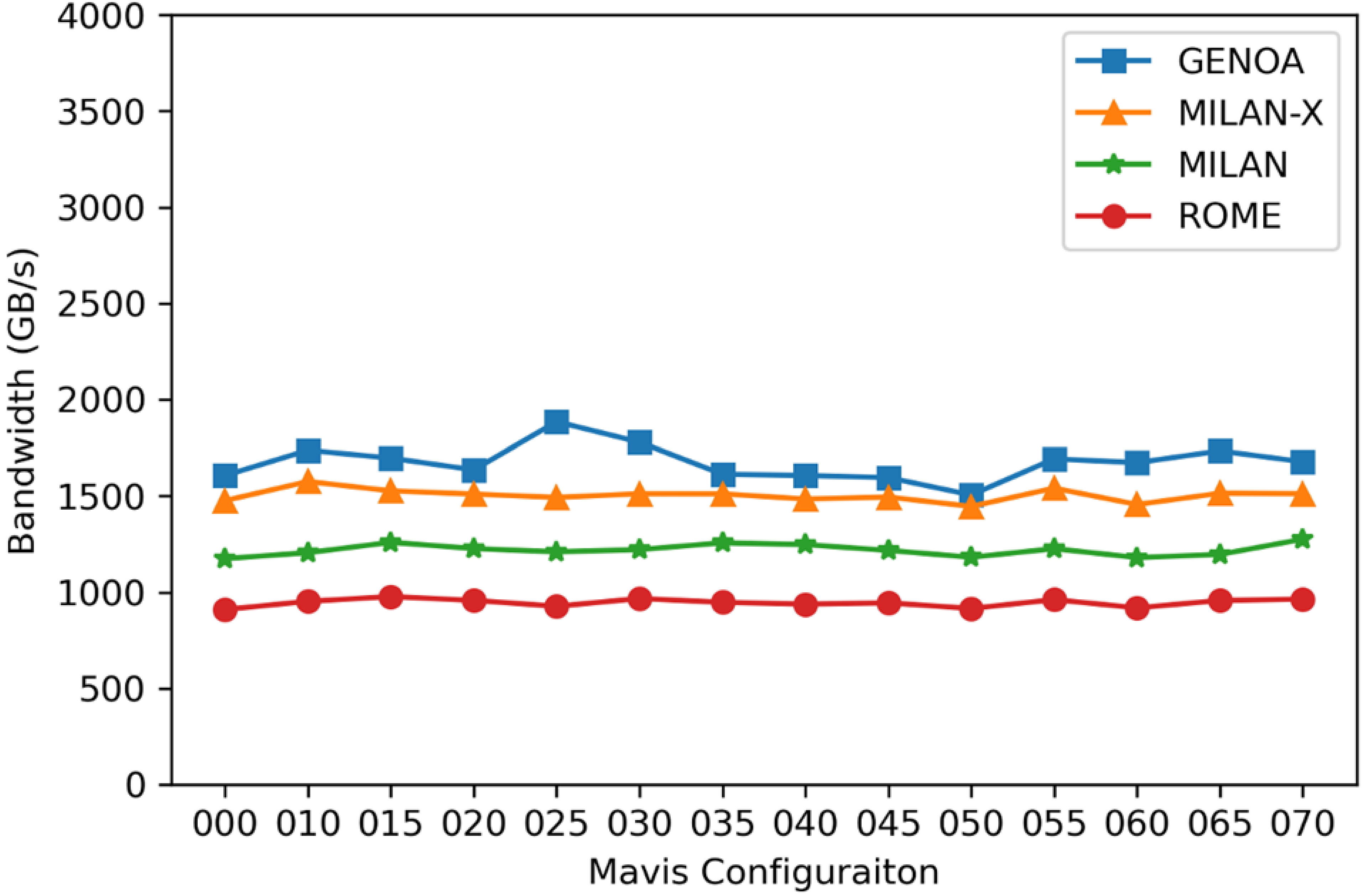
Lower is better

Accelerating RT Computations For Giant Telescopes

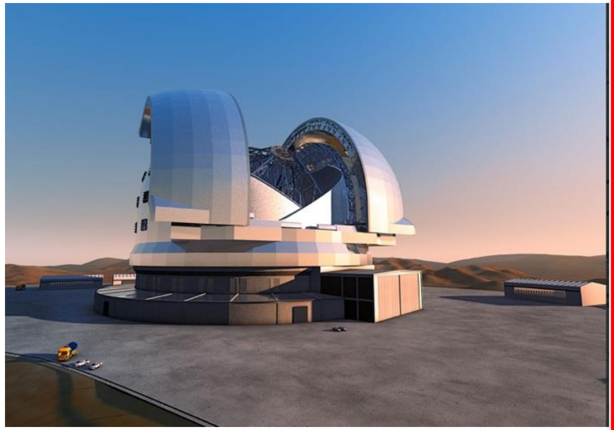
Roofline Performance Model



Performance Across AMD x86 Generations

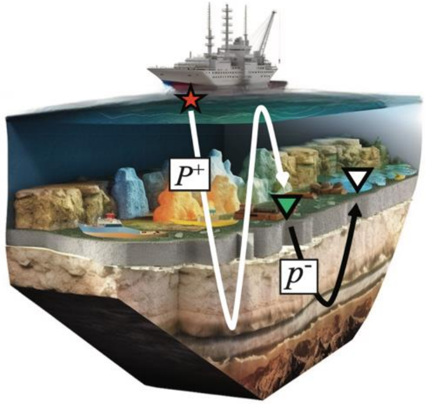


Challenging Scientific Applications



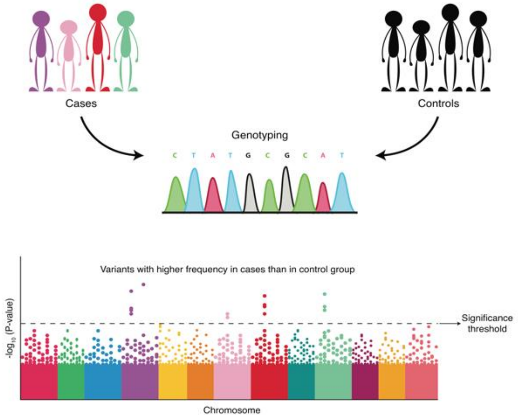
Computational Astronomy

Best Paper
PASC18



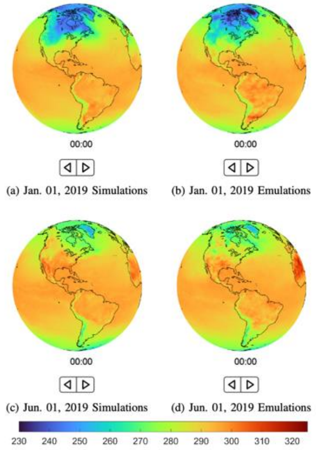
Seismic Processing

Gordon Bell Finalist
SC23



Computational Biology

Gordon Bell Finalist
SC24



Climate

Simulations / Emulations

Gordon Bell Finalist SC22
Gordon Bell Winner SC24

Multi-Dimensional Convolution operator

Scaling the "Memory Wall" for Multi-Dimensional Seismic Processing with Algebraic Compression on Cerebras CS-2 Systems

Hatem Ltaief
Yuxi Hong
Extreme Computing Research Center
Computer, Electrical and Mathematical Sciences & Engineering Division
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia
firstname.lastname@kaust.edu.sa

Leighton Wilson
Mathias Jacquelin
Cerebras Systems Inc.
Sunnyvale, California, United States
firstname.lastname@cerebras.net

Matteo Ravasi
David Keyes
Extreme Computing Research Center
Computer, Electrical and Mathematical Sciences & Engineering Division
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia
firstname.lastname@kaust.edu.sa

ABSTRACT

We exploit the high memory bandwidth of AI-customized Cerebras CS-2 systems for seismic processing. By leveraging low-rank matrix approximation, we fit memory-hungry seismic applications onto memory-austere SRAM wafer-scale hardware, thus addressing a challenge arising in many wave-equation-based algorithms that rely on Multi-Dimensional Convolution (MDC) operators. Exploiting sparsity inherent in seismic data in the frequency domain, we implement embarrassingly parallel tile low-rank matrix-vector multiplications (TLR-MVM), which account for most of the elapsed time in MDC operations, to successfully solve the Multi-Dimensional Deconvolution (MDD) inverse problem. By reducing memory footprint along with arithmetic complexity, we fit a standard seismic benchmark dataset into the small local memories of Cerebras processing elements. Deploying TLR-MVM execution onto 48 CS-2 systems in support of MDD gives a sustained memory bandwidth of 92.58PB/s on 35, 784, 000 processing elements, a significant milestone that highlights the capabilities of AI-customized architectures to enable a new generation of seismic algorithms that will empower multiple technologies of our low-carbon future.

November 12–17, 2023, Denver, CO, USA. ACM, New York, NY, USA, 12 pages.
https://doi.org/10.1145/3581784.3627042

1 JUSTIFICATION FOR THE GORDON BELL PRIZE

High-performance matrix-vector multiplication using low-rank approximation. Memory layout optimizations and batched executions on massively parallel Cerebras CS-2 systems. Leveraging AI-customized hardware capabilities for seismic applications for a low-carbon future. Application-worthy accuracy (FP32) with a sustained bandwidth of 92.58PB/s (for 48 CS-2s) would constitute the second-highest throughput from June'23 Top500.

2 PERFORMANCE ATTRIBUTES

Performance Attributes	Our submission
Problem Size	Broadband 3D seismic dataset ($\sim 20k$ sources and receivers and frequencies up to 50Hz)
Category of achievement	Sustained bandwidth

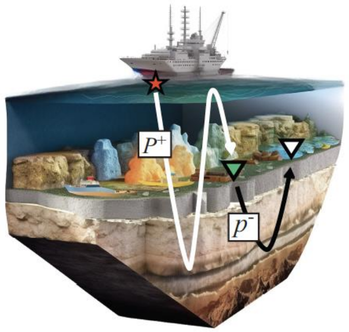
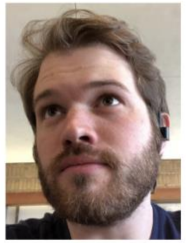


Figure 1: Schematic representation of the Multi-Dimensional Deconvolution problem. A red star indicates the source, a green triangle refers to the receiver, and the virtual source is represented by a white triangle.

48 Cerebras CS-2 systems, i.e., 35,784,000 processing elements

Group42 (Abu Dhabi), KAUST Supercomputing Core Lab and:

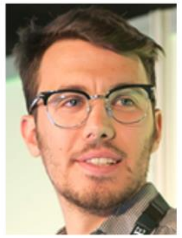
cerebras


Leighton Wilson


Mathias Jacquelin


Yuxi Hong


Hatem Ltaief


Matteo Ravasi


David Keyes

Multi-Dimensional Convolution operator

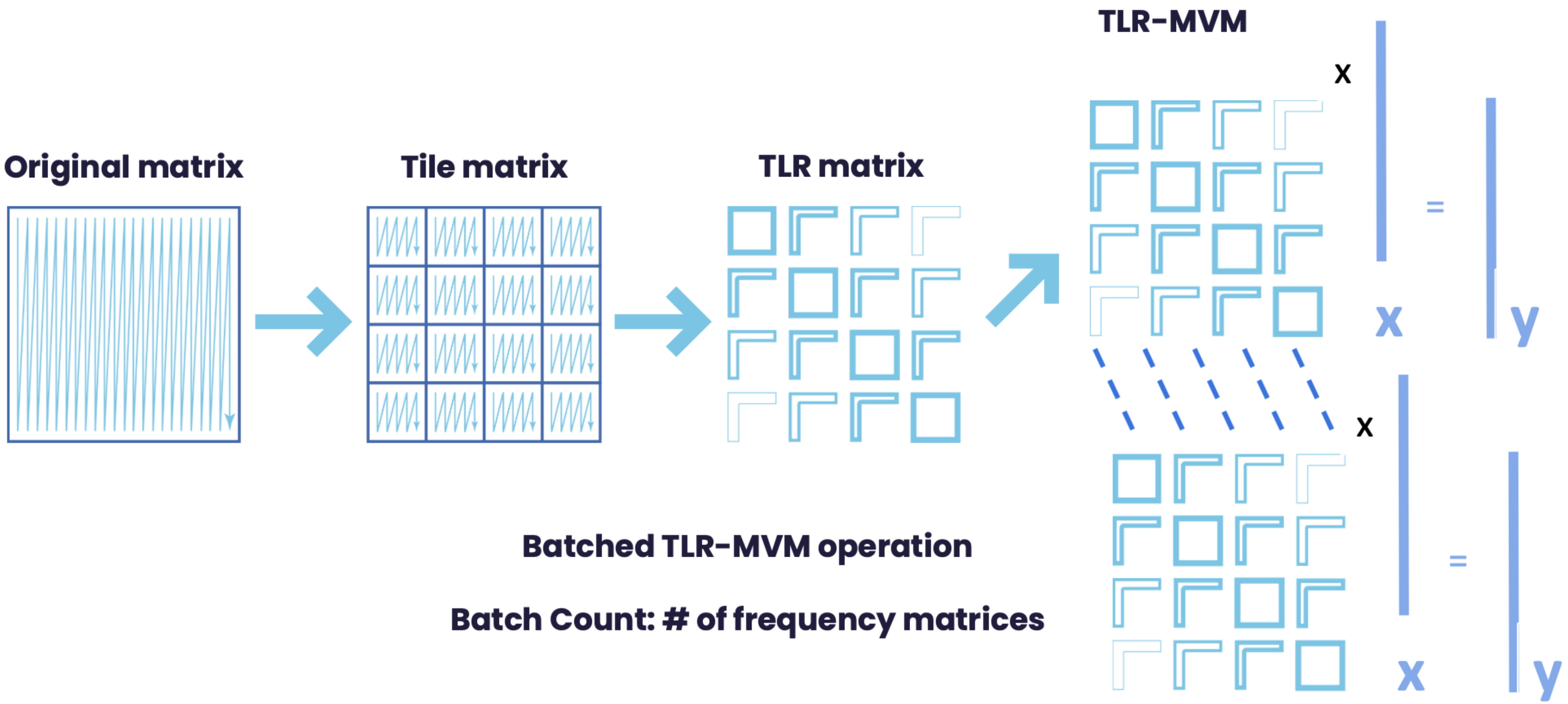
$$\mathbf{p}^{-} = \underbrace{\mathbf{P}^{+}}_{\underbrace{\mathbf{F}^{\mathrm{H}}}_{\text{Inverse FFT}} \; \underbrace{\mathbf{I}^{\mathrm{H}}_{f_{max}}}_{\text{Padding}} \; \underbrace{\widehat{\mathbf{P}}^{+}}_{\text{Batch matrix-vector multiplication}} \; \underbrace{\mathbf{I}_{f_{max}}}_{\text{Restriction}} \; \underbrace{\mathbf{F}}_{\text{FFT}}} \mathbf{r}$$

Our sweet spot!!

“multidimensional” refers the ability to track waves that are not simply vertical

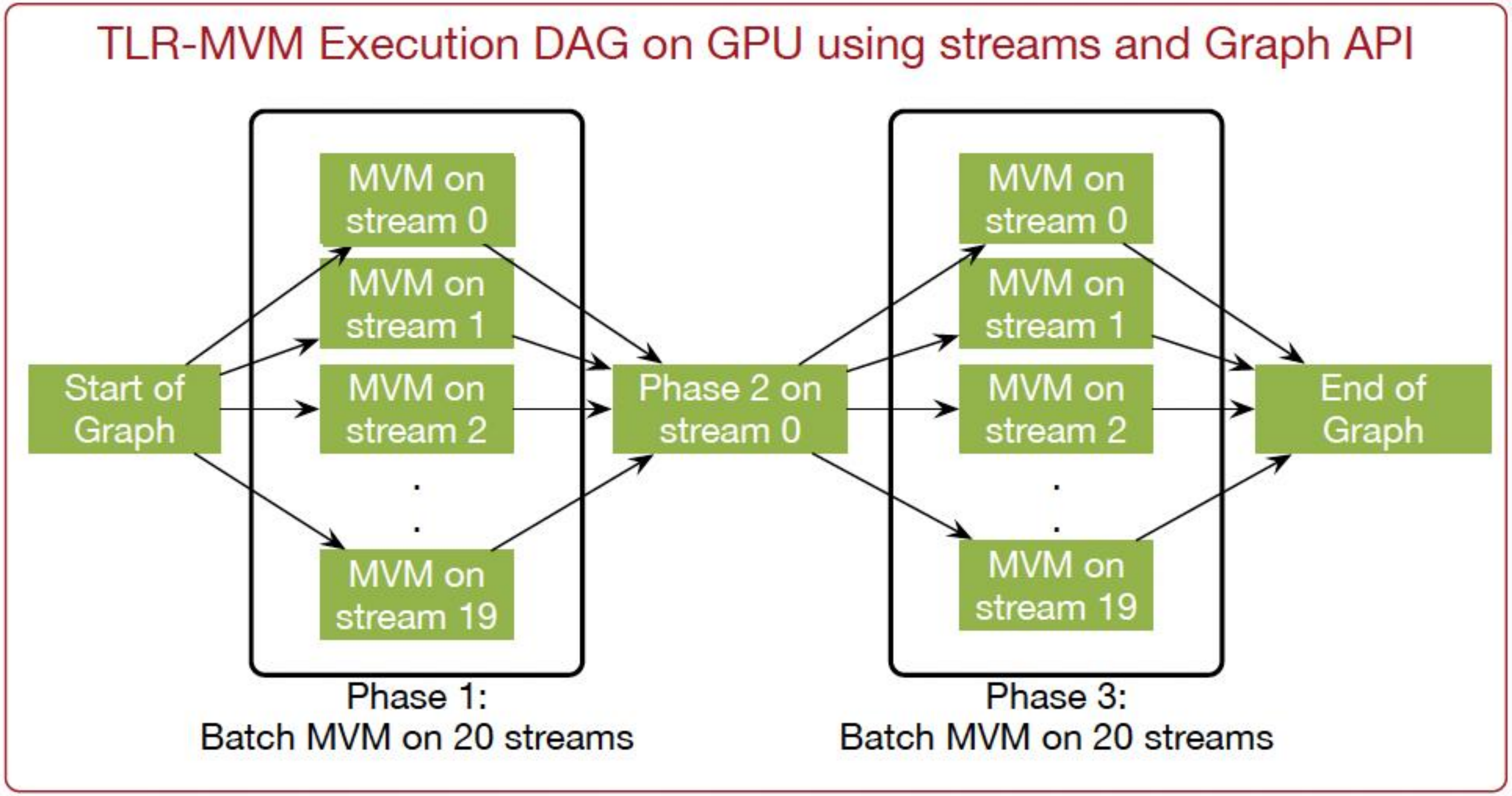
Ravasi and Vasconcelos, 2021

Compress to Impress: tile low-rank approximation



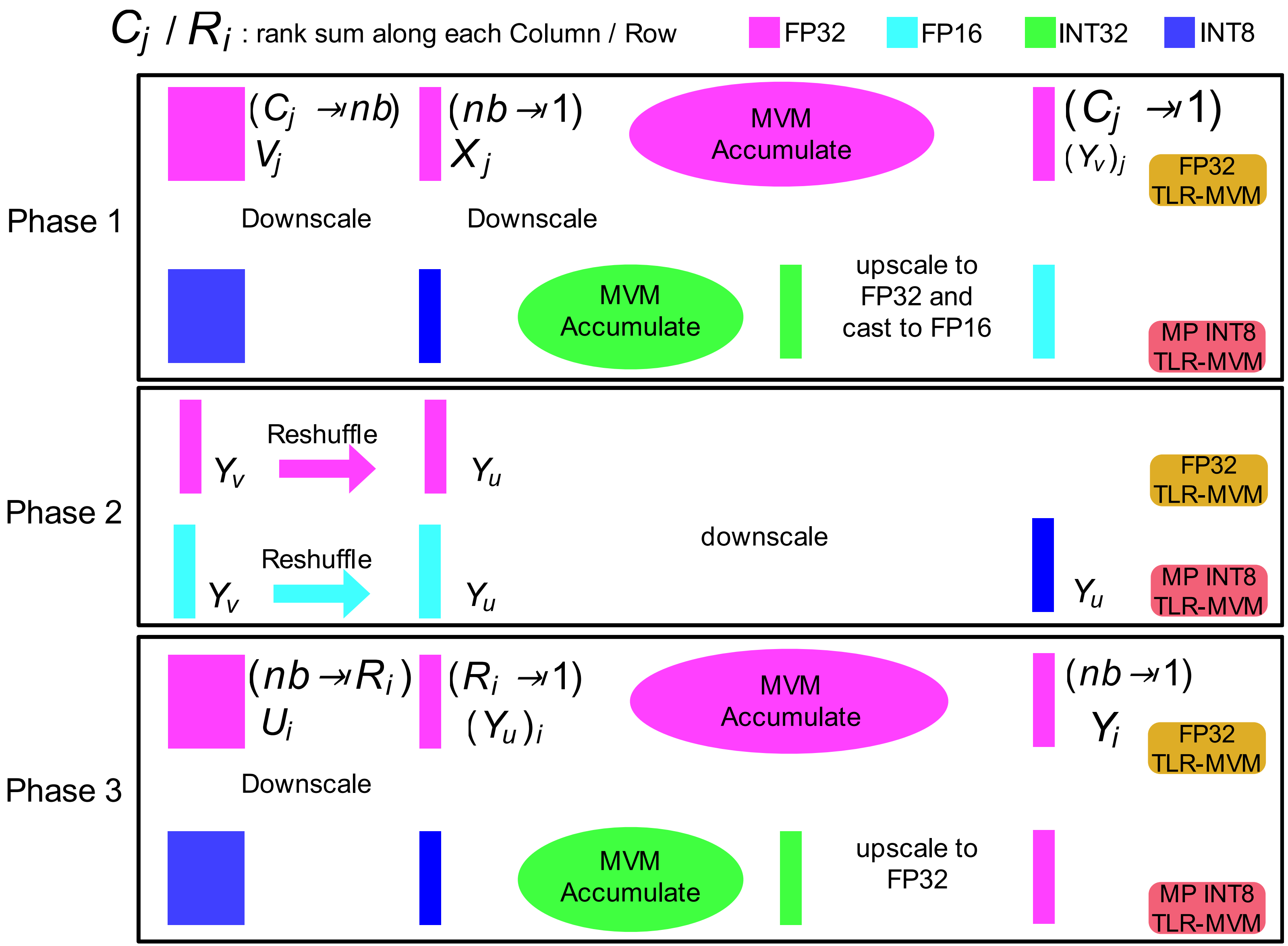
P. Amestoy, C. Ashcraft, O. Boiteau, A. Buttari, J.Y. L'Excellent, and C. Weisbecker, *Improving Multifrontal Methods by Means of Block Low-Rank Representations*, SIAM SISC, 2015.
K. Akbudak, HL, A. Mikhalev, A. Charara, A. Esposito, and D. Keyes, *Exploiting Data Sparsity for Large-Scale Matrix Computations*, ISC, 2017.
P. Amestoy, A. Buttari, J.Y. L'Excellent, and T. Mary, *Performance and Scalability of the Block Low-Rank Multifrontal Factorization on Multicore Architectures,* ACM TOMS, 2019.

GPU Implementation w/ CUDA Graph

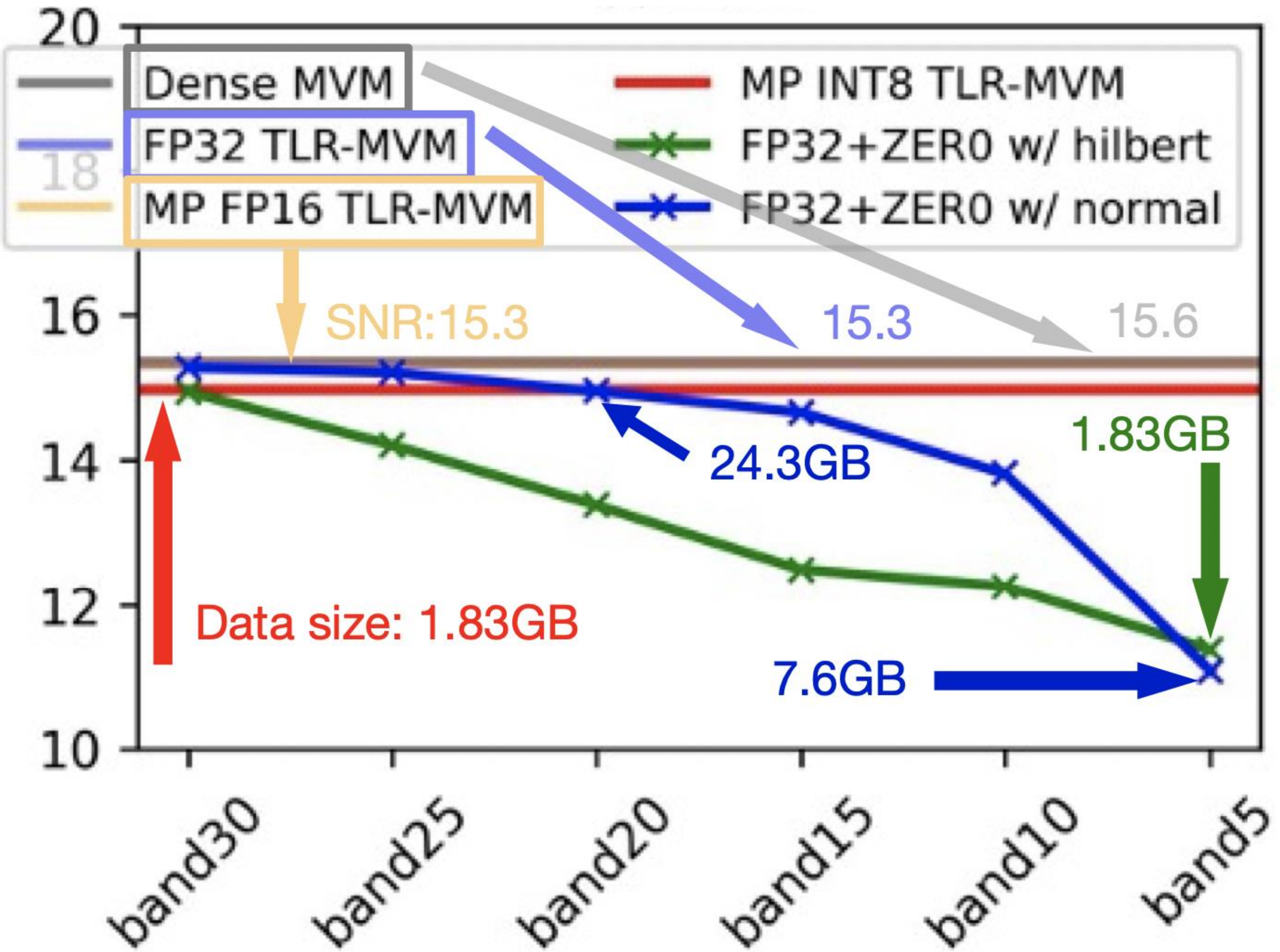


- **Streams to reach high occupancy on the GPU**
- **CUDA Graph to reduce kernel launch overheads**

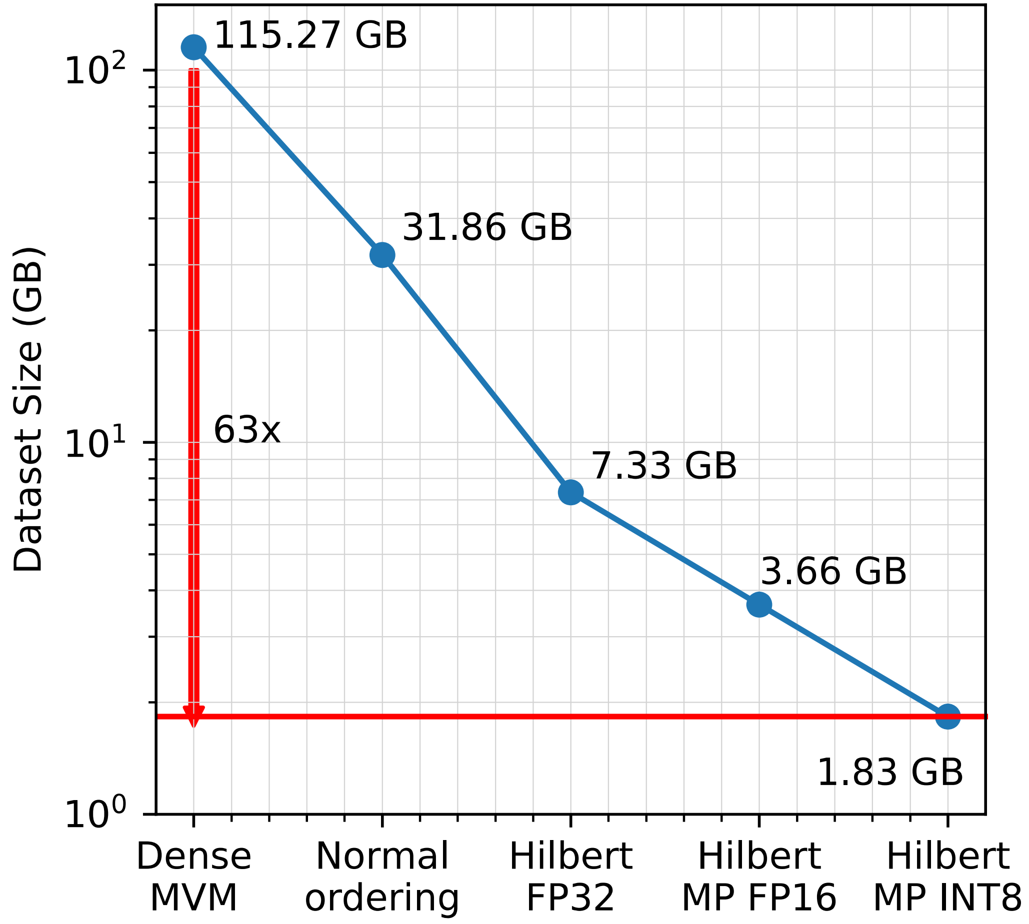
GPU Implementation w/ CUDA Graph + MxP



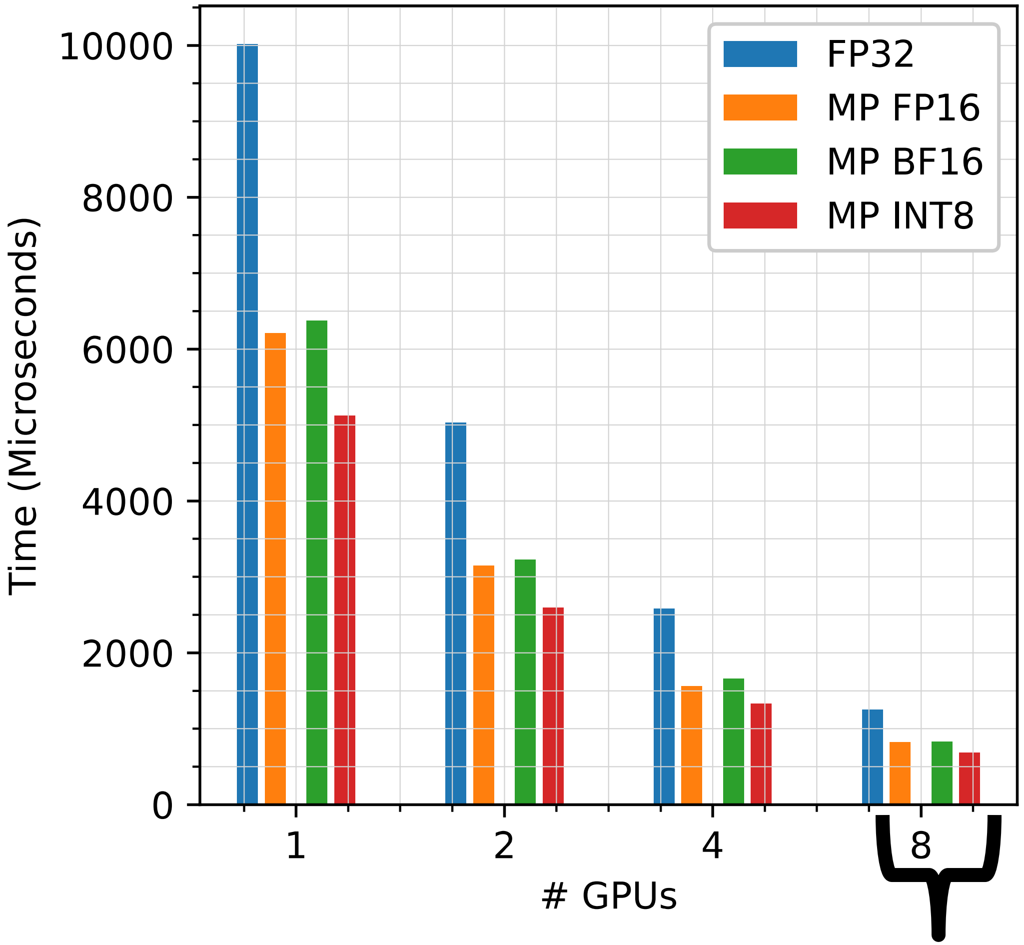
Accuracy Results



Performance Results



Memory footprint

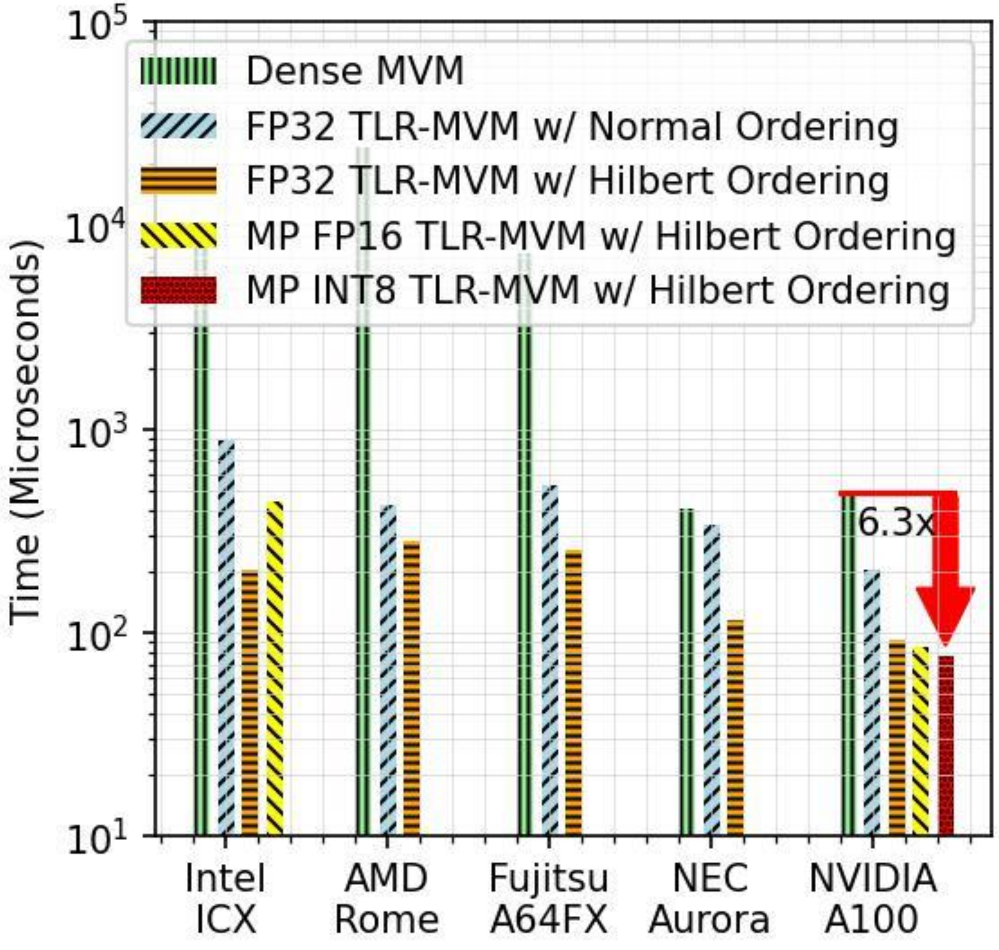


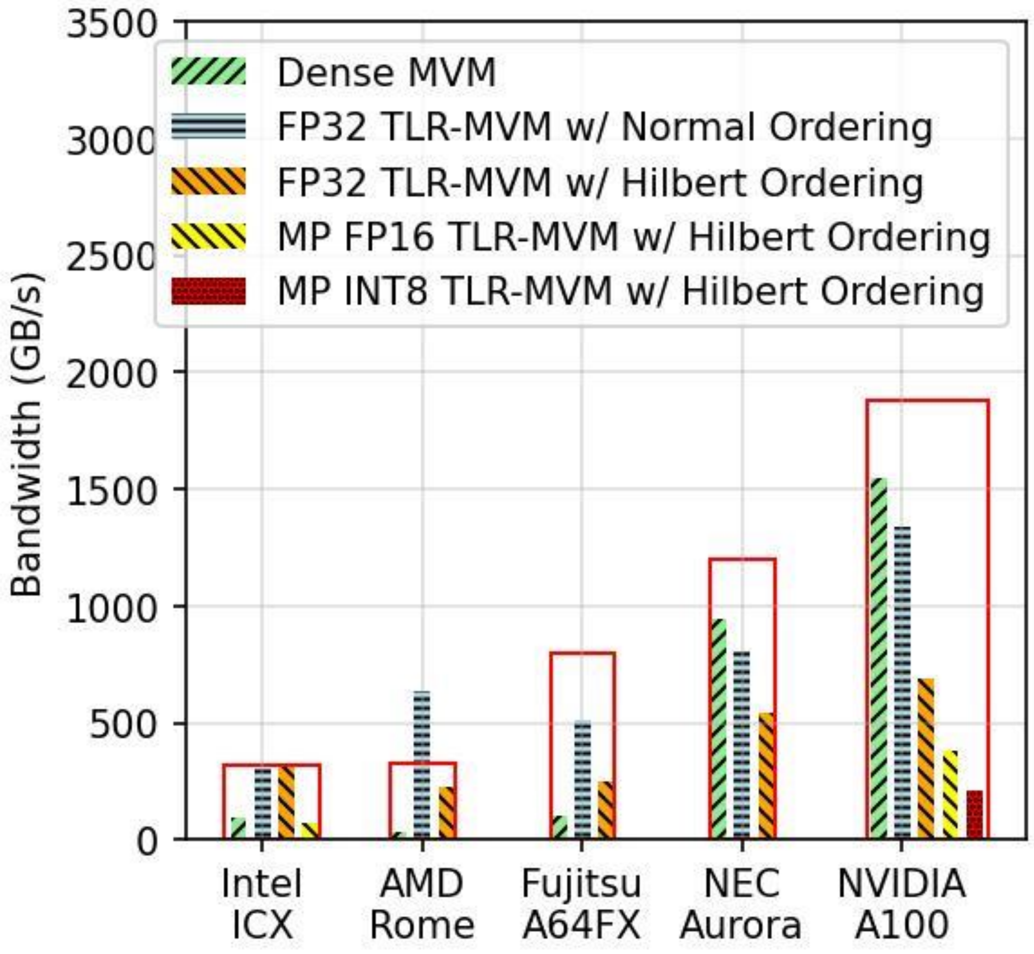
Scalability

Breaking the ms barrier!

TLR-MVM on NVIDIA A100 GPU

Performance Results





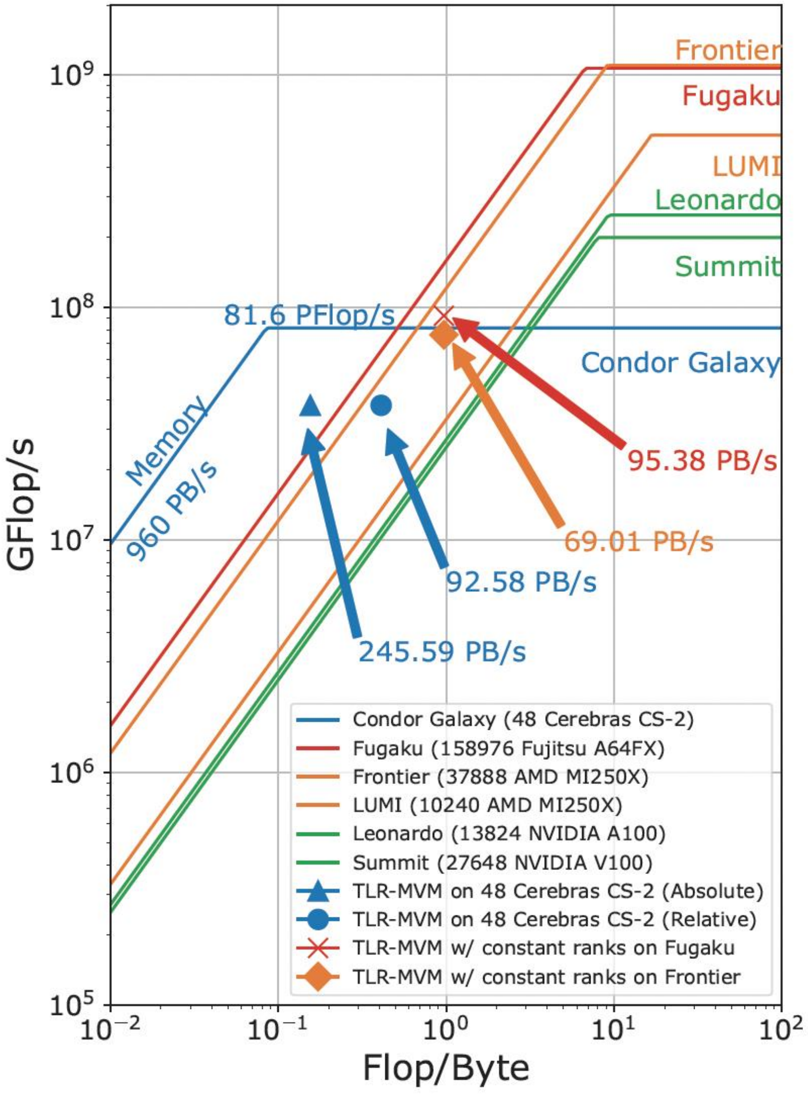
Comparisons against other chips

Scaling up on Cerebras CS-2 Wafer Scale

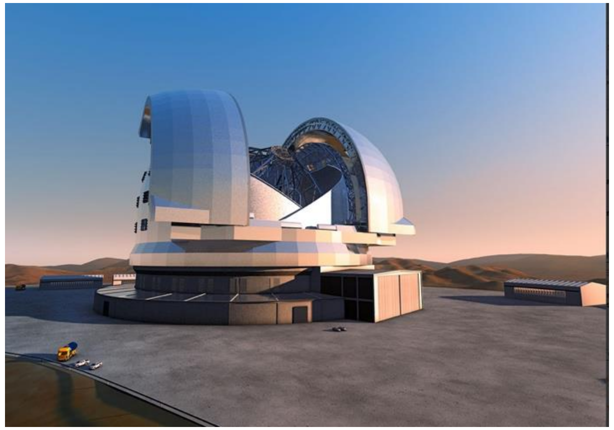
Strong scaling up to 48 CS-2 systems

Performance comparisons against the Top5 fastest Supercomputers

As per June 2023 Top500, 92.58 PB/s is:
- 2.3X > vs theo bw of Lumi #3
- 3X > vs theo bw of Leonardo #4 / Summit #5
- 35% > vs theo bw of Frontier #1
- close to est. sust. bw of Fugaku #2
- 3X > vs theo bw of Oceanlite

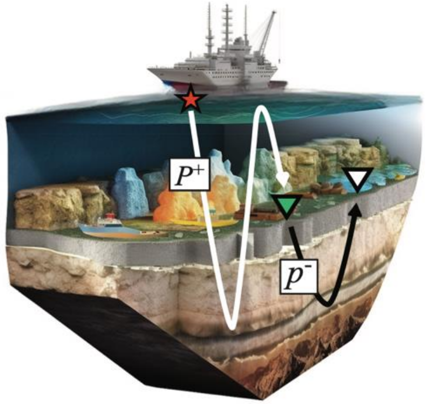


Challenging Scientific Applications



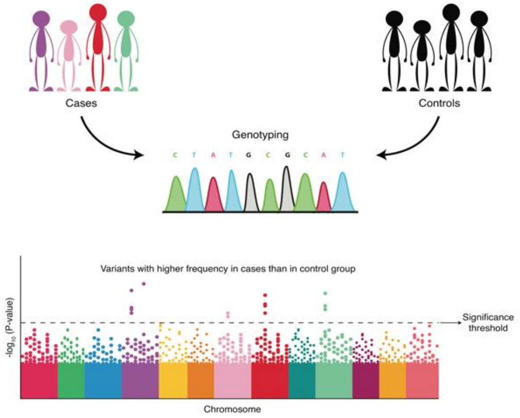
Computational Astronomy

Best Paper
PASC18



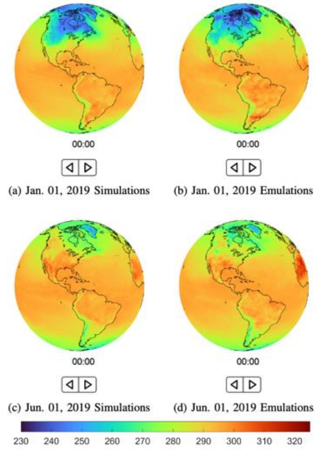
Seismic Processing

Gordon Bell Finalist
SC23



Computational Biology

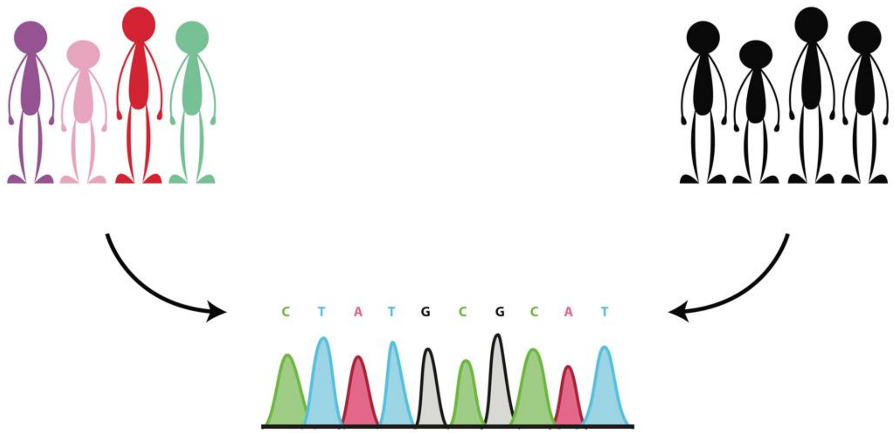
Gordon Bell Finalist
SC24

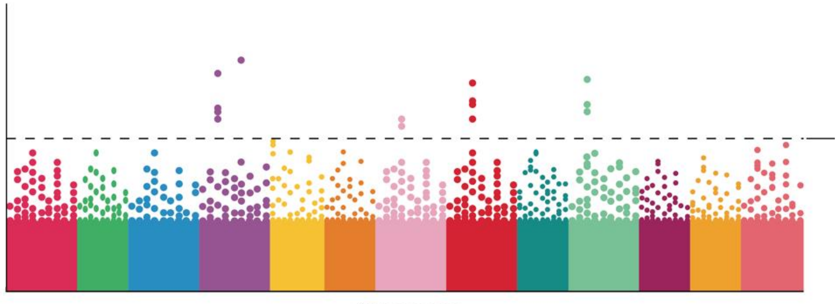


Climate
Simulations / Emulations

Gordon Bell Finalist SC22
Gordon Bell Winner SC24

Genome-Wide Association Studies (GWAS 101)





A training population (size N_{P1}) is sequenced for SNPs and characterized by additional environmental factors (generalized genotype size N_S)

The same population is characterized by prevalent traits, such as body features and diseases (phenotype size N_T)

Genotype variants are associated to phenotypic traits in "Manhattan plots" for potential diagnoses and (ultimately) genetic therapies

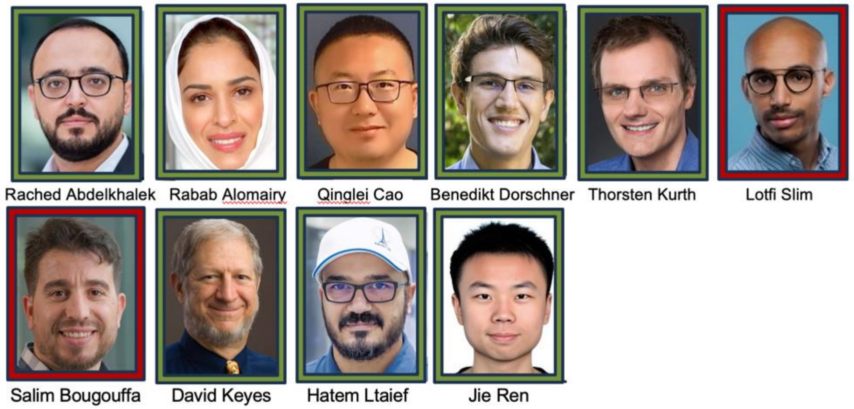
A test population (size N_{P2}) is sequenced and members' susceptibility to disease predicted

KAUST Supercomputing Core Lab, Oak Ridge LCF, CSCS Alps, CINECA Leonardo, and:

NVIDIA

SAINT LOUIS UNIVERSITY

Massachusetts Institute of Technology



Kernel Ridge Regression for Genetic Epistasis

Genetic **epistasis** occurs when the effect of one gene is influenced by one or more other genes, altering the expected genetic outcome.

- Kernel ridge regression (KRR) further extends ridge regression by enabling modeling of complex, nonlinear interactions among SNPs and between SNPs and the trait.
- The objective function for KRR:

$$min_W ||P - KW||^2 + \lambda W^T KW$$

- Closed-form solution for KRR:

$$W = (\underbrace{K}_{} + \lambda I)^{-1} P$$

Build phase $K(p_i, p_j) = \exp(-\gamma \, || p_i - p_j||^2)$

Associate phase: Cholesky-based Solver

P: Phenotype (trait or disease status)
K: is the kernel matrix, often representing genetic relatedness (e.g., a genetic similarity matrix, such as the Genetic Relatedness Matrix, or GRM)
W: Coefficients that represents the contributions of individuals' genetic similarities to phenotype prediction.

G. Chávez, Y. Liu, P. Ghysels, X. S. Li and E. Rebrova, *Scalable and Memory-Efficient Kernel Ridge Regression*, IEEE IPDPS, 2020.

Generic KRR ALgorithms

Algorithm 1: Three-Phase Kernel Ridge Regression (KRR) for GWAS.

Input
N_{P1}: # of Patients in training set
N_{P2}: # of Patients in testing set
N_S: # of SNPs
N_{Ph}: # of Phenotypes
G: $N_{P1} \times N_S$ (Training genotype matrix)
P_h: $N_{P1} \times N_{Ph}$ (Training phenotype matrix)
T: $N_{P2} \times N_S$ (Testing genotype matrix)
γ: kernel bandwidth
α: regularization parameter
Output
K: $N_{P1} \times N_{P1}$ (KRR matrix)
W: $N_{P1} \times N_{Ph}$ (Weight matrix)
P_r: $N_{P2} \times N_{Ph}$ (Predictions)
Phase 1: $\textsc{Build}(\gamma, G, G, K)$
Phase 2: $\textsc{Associate}(\alpha, K, P_h, W)$
Phase 3: $\textsc{Predict}(\gamma, G, T, W, P_r)$

Algorithm 2: Build the KRR matrix.

Procedure $\textsc{Build}(\gamma, G_1, G_2, K)$
$N_{P1} \leftarrow \text{rowsize}(G_1)$
$N_{P2} \leftarrow \text{rowsize}(G_2)$
$K \leftarrow \text{zeros}(N_{P1}, N_{P2})$
for i in $\text{range}(1, N_{P1})$ **do**
  for j in $\text{range}(1, N_{P2})$ **do**
    $K[i,j] \leftarrow \textsc{KernelMatrix}(\text{type}, \gamma, G_1[i,:], G_2[j,:])$
  end for
end for

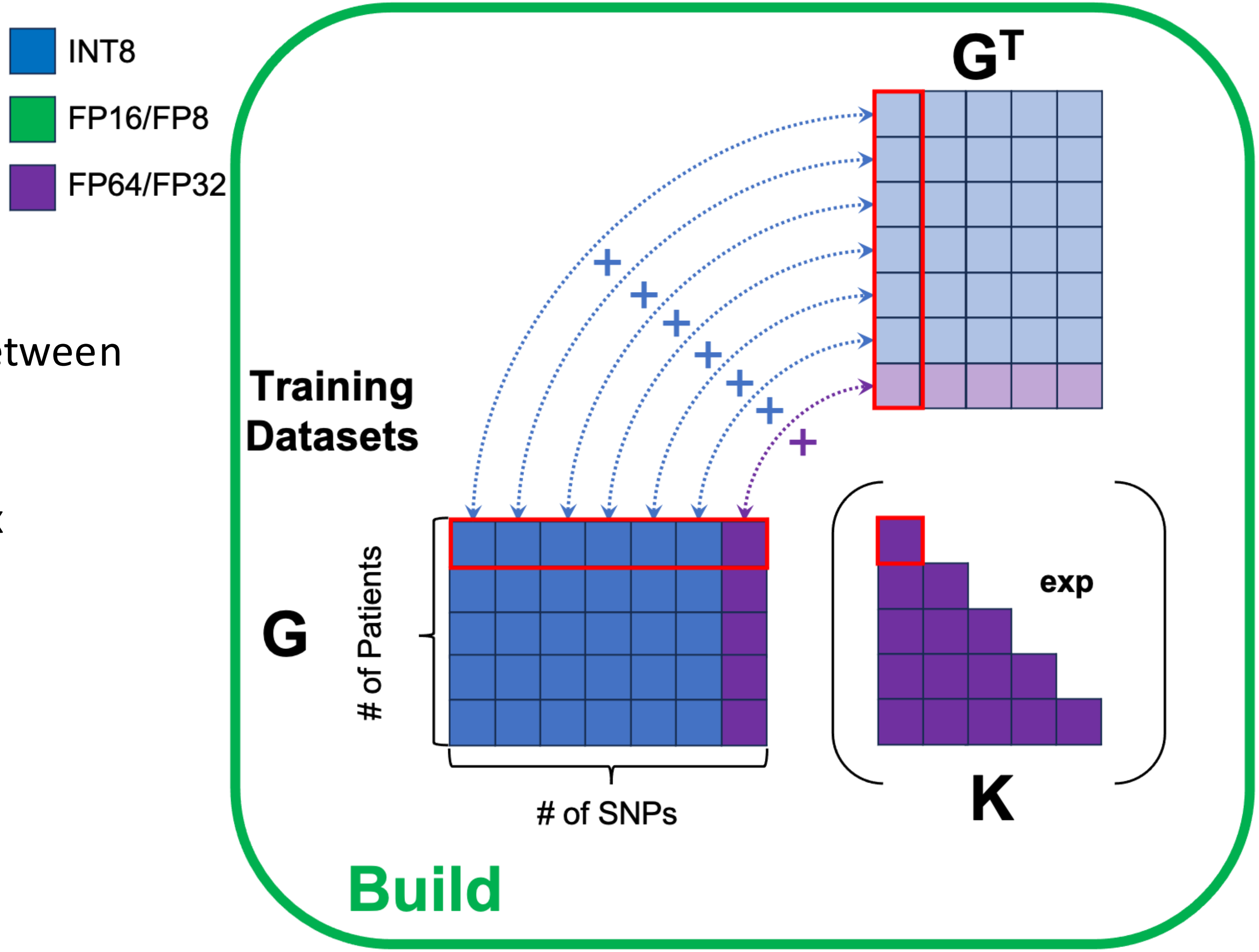
Algorithm 3: Associate genotype-phenotype.

Procedure $\textsc{Associate}(\alpha, K, P_h, W)$
Factorize the KRR matrix
$\tilde{K} \leftarrow \textsc{Factorize}(K + \alpha \cdot Id)$
Solve for W
$W \leftarrow \textsc{Solve}(\tilde{K}, P_h)$

Algorithm 4: Predict for a new cohort.

Procedure $\textsc{Predict}(\gamma, G, T, W, P_r)$
$N_{P1} \leftarrow \text{rowsize}(G)$
$N_{P2} \leftarrow \text{rowsize}(T)$
K: $N_{P2} \times N_{P1}$ (test-training kernel matrix)
$\textsc{Build}(\gamma, T, G, K)$
$P_r \leftarrow K \times W$

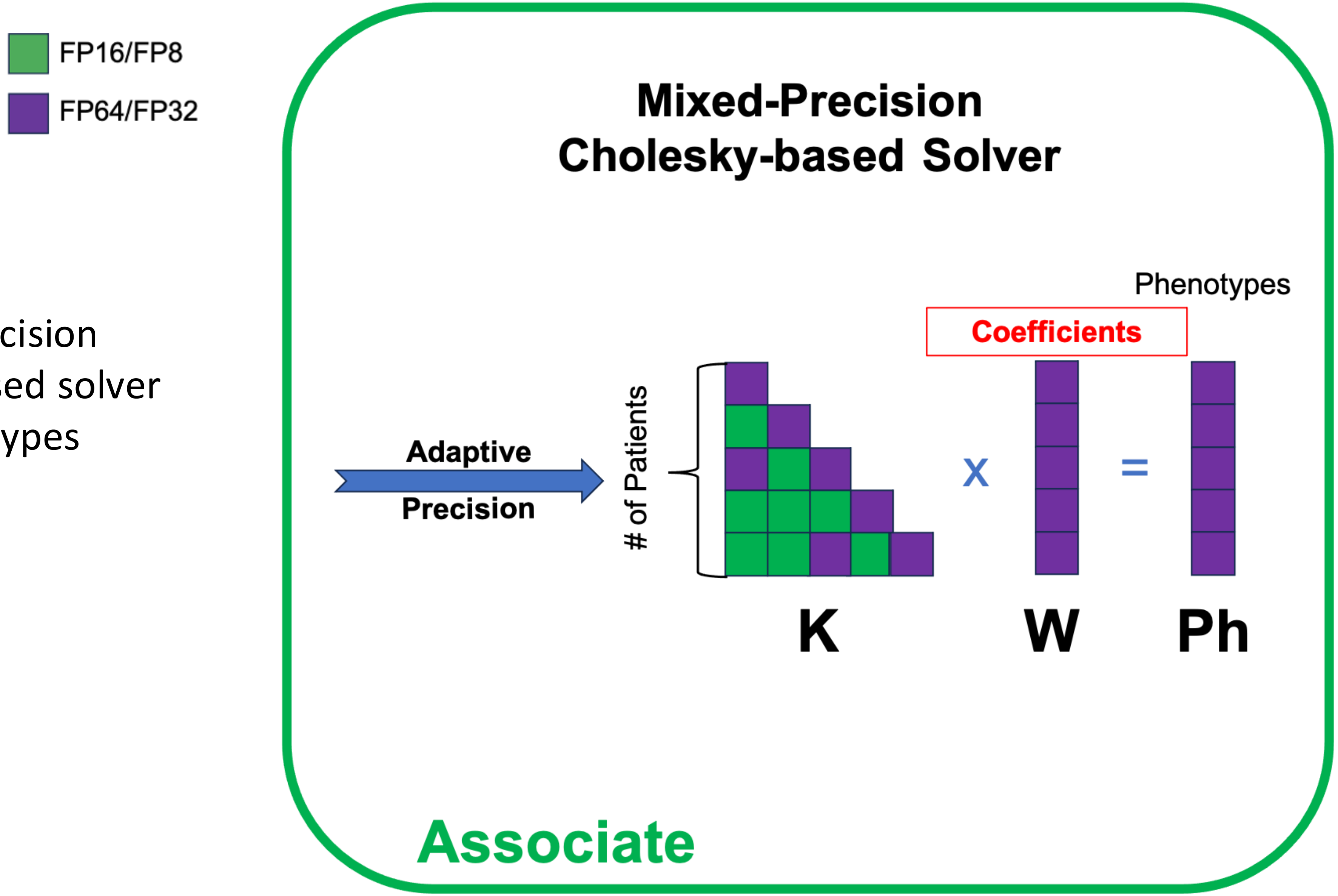
Build Phase



- Compute Euclidean distance between each pair of individual (slow)
- Exponent the results
- Generate the covariance matrix

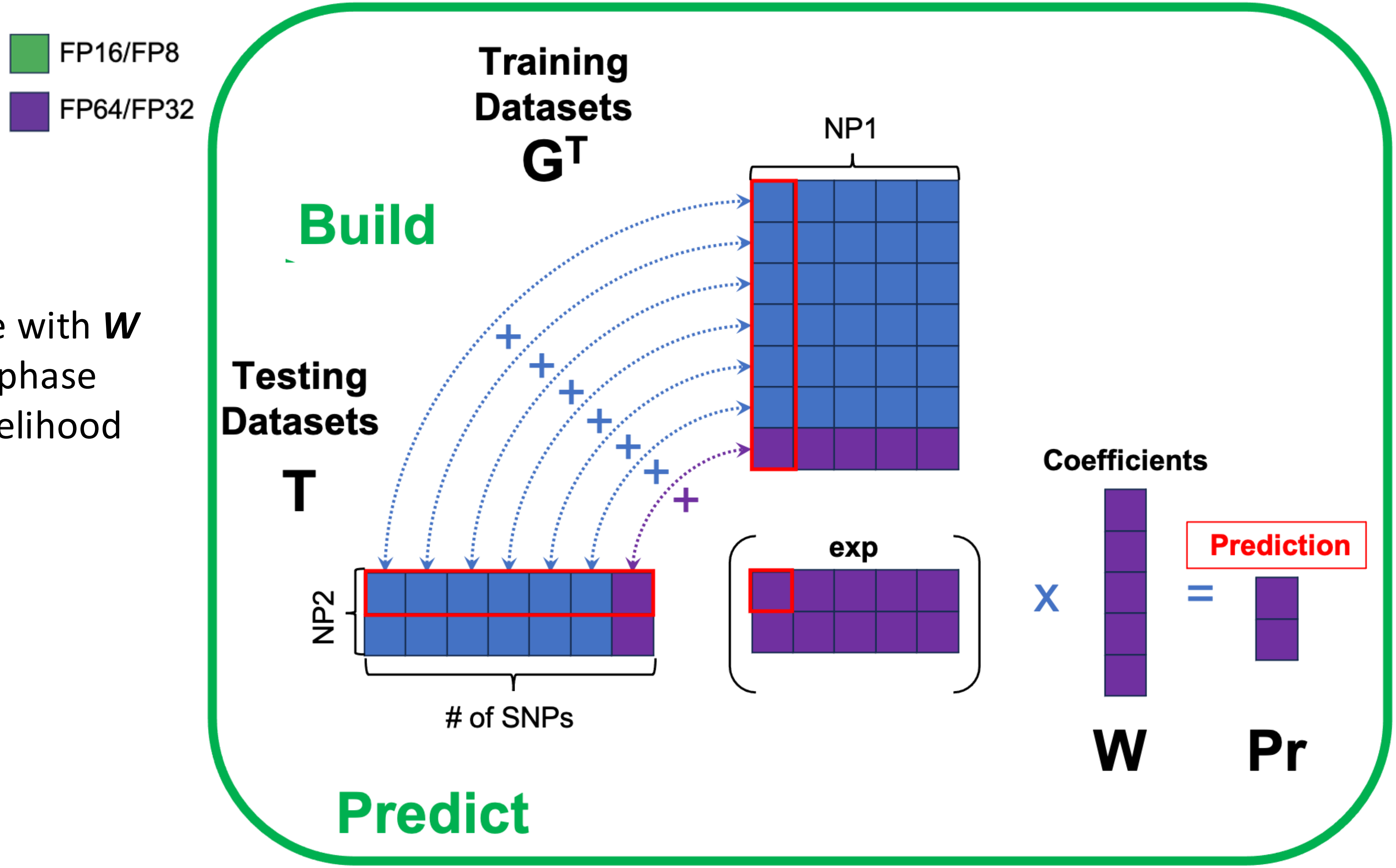
Associate Phase

- Perform adaptive precision
- Execute Cholesky-based solver against a list of phenotypes
- Save the *weights*



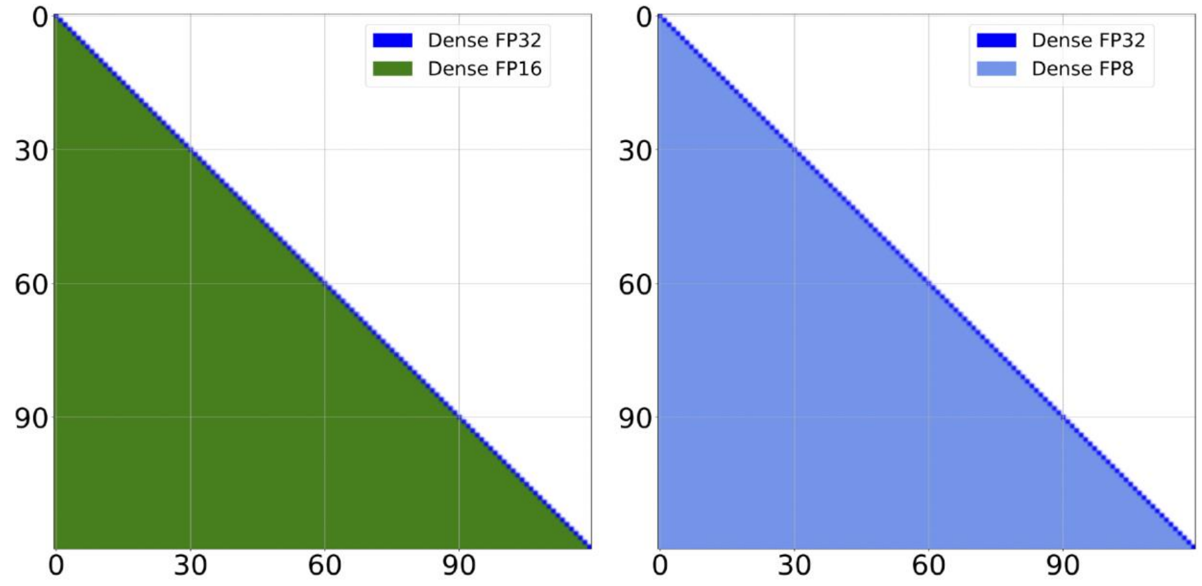
Predict Phase

- Perform inference with ***W***
- Rely on the build phase
- Determine the likelihood



GWAS surfing the AI wave w/ low precision arithmetic

- Precision assessment after constructing the kernel matrix for a real dataset comprising 305K patients and 43K genotypes (SNPs) after the Build phase.



(a) Activating FP16 with A100. (b) Activating FP8 with GH200.

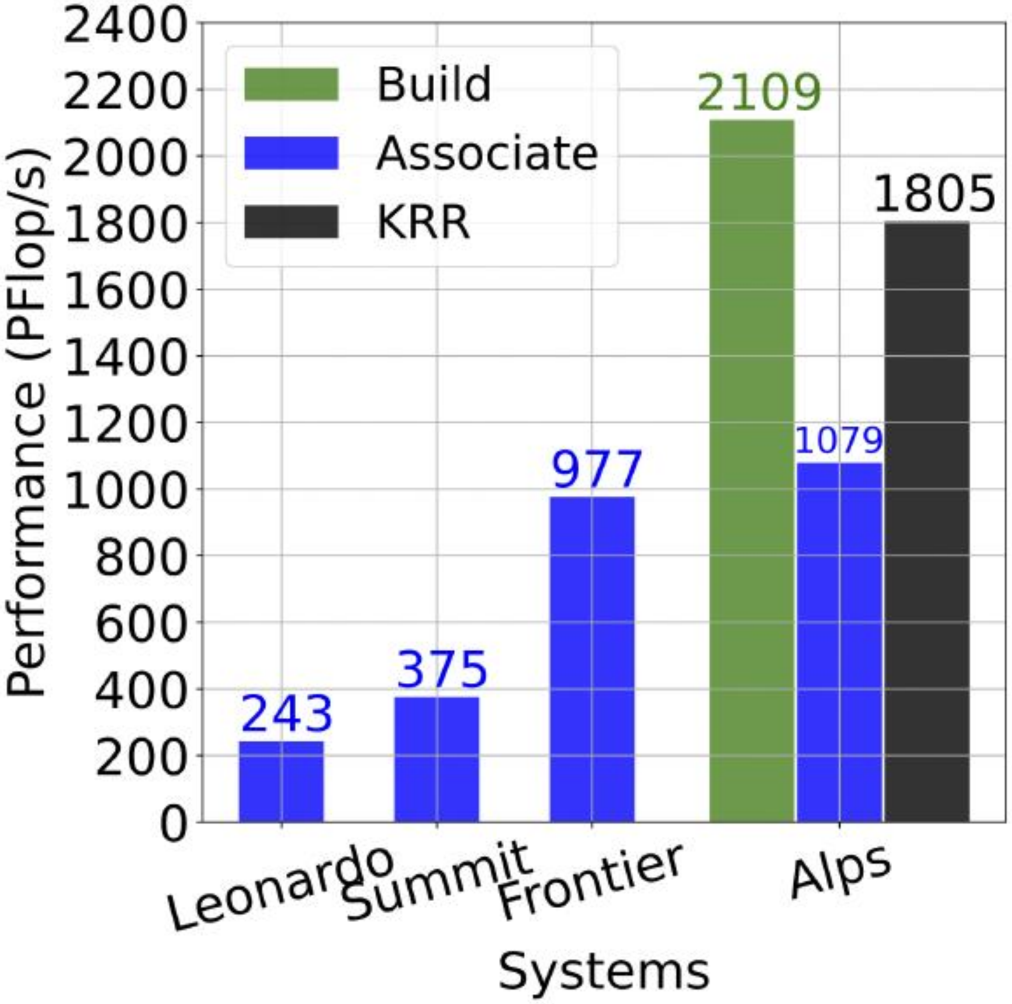
Precision heatmaps

$$\boxed{\|A_{ij}\|_F < u_{high} * \|A\|_F/(NT * u_{low})}$$ [2]

$$\boxed{\|\hat{A} - A\|_F \leq u_{high}\|A\|_F}$$

[2] Nick Higham and Theo Mary, *Mixed Precision Algorithms in Numerical Linear Algebra*, Acta Numerica, 2022.

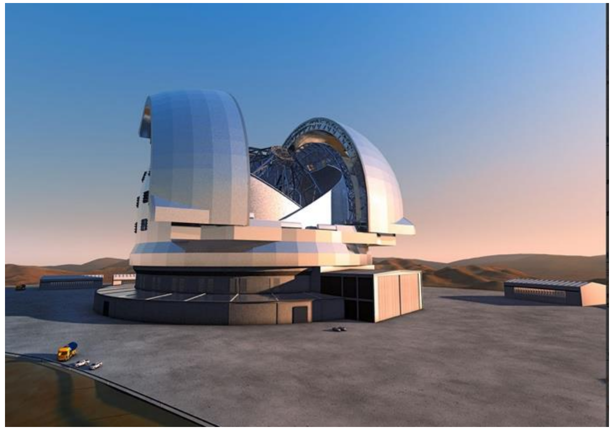
Hero run: Alps vs Frontier / Leonardo / Summit



Patients = 13.2M
SNPs = 20.6M

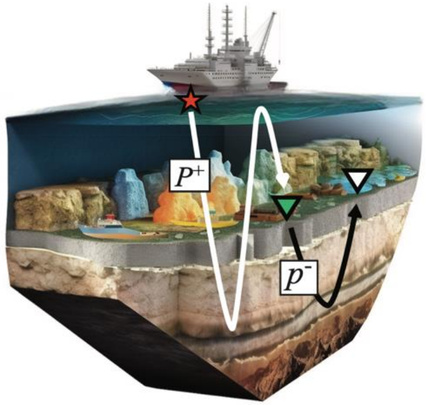
Five orders of magnitude faster than SOTA CPU-only REGENIE

Challenging Scientific Applications



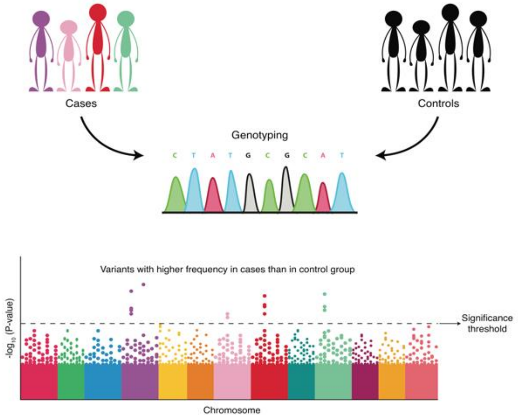
Computational Astronomy

Best Paper
PASC18



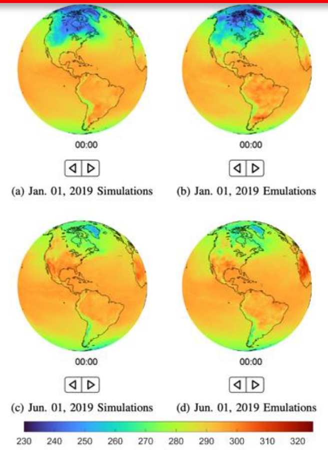
Seismic Processing

Gordon Bell Finalist
SC23



Computational Biology

Gordon Bell Finalist
SC24



Climate

Simulations / Emulations

Gordon Bell Finalist SC22
Gordon Bell Winner SC24

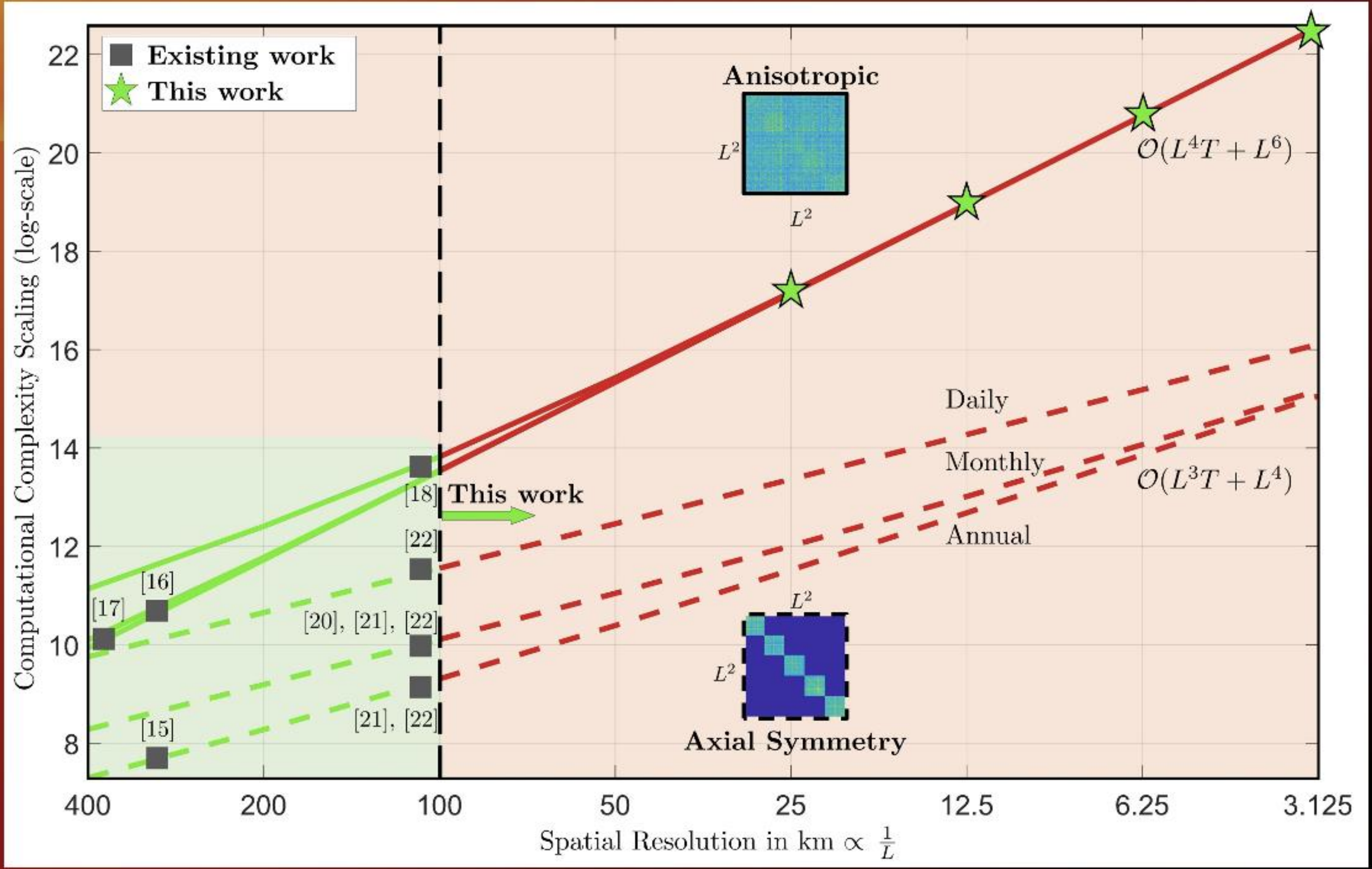
MOTIVATIONS & CHALLENGES

- Our understanding of the climate system relies on **state-of-the-art Earth System Models (ESMs)** based on PDEs and run on supercomputers (very few runs can be afforded!)
- ESMs play a fundamental role in the **Intergovernmental Panel on Climate Change (IPCC) sixth assessment report** (AR6) to forecast warming across various emission scenarios
- The latest **Coupled Model Intercomparison Project** (CMIP6) supports detailed comparisons of ESMs (generated ~28 Petabytes data from 45 modeling institutes)
- **Computational demands and petabyte-scale storage requirements/costs** for ESMs continue to escalate as the climate community progresses toward **ultra-high-resolution** simulations
- **Simulations at "global storm-resolving" scales** are needed to understand better how weather and extremes will be affected by climate change

CLIMATE EMULATORS

- Climate emulators are **stochastic models designed to mimic ESM behavior** using simulated data from a few runs of the ESM
- Climate emulators can **quickly generate multiple emulations** of the output of an ESM, which is **crucial for quantifying the uncertainties** in future climate projections
- Climate emulators are designed to **complement and boost the usefulness of ESMs**
- Current global climate emulators **have not yet attained a spatial resolution finer than 100 km**
- Current global climate emulators **have not yet attained a temporal resolution finer than daily (and only then with an assumption of axial symmetry)**
- Current global climate emulators **without assumption of axial symmetry have not yet attained a temporal resolution finer than annually**

CLIMATE EMULATORS: EXISTING WORK & CHALLENGES



SUMMARY OF CONTRIBUTIONS (1/2)

- Developed and validated **our own climate emulator** (Song, Khalid, Genton, 2024, JASA)
- Our **Exascale Climate Emulator surpasses** existing climate emulators **by a factor of 245,280X** (28X in space and 8,760X in time)
- Addresses limitations of existing emulators – **spherical harmonic transform (SHT) to model anisotropic interactions**
- Our Exascale Climate Emulator can emulate up to **54.5 million spatial locations** across the globe with an **ultra-high spatial resolution of up to 0.034° (3.5 km) at an hourly resolution.** This equates to **477 billion data points** for a single year of emulation
- Virtually **saving infinite number of Petabytes of storage and cost** via on-demand climate data emulations
- **Breaking the ultra-high-resolution barrier** of climate emulators for **advancing climate research and policy making**

KAUST Shaheen III

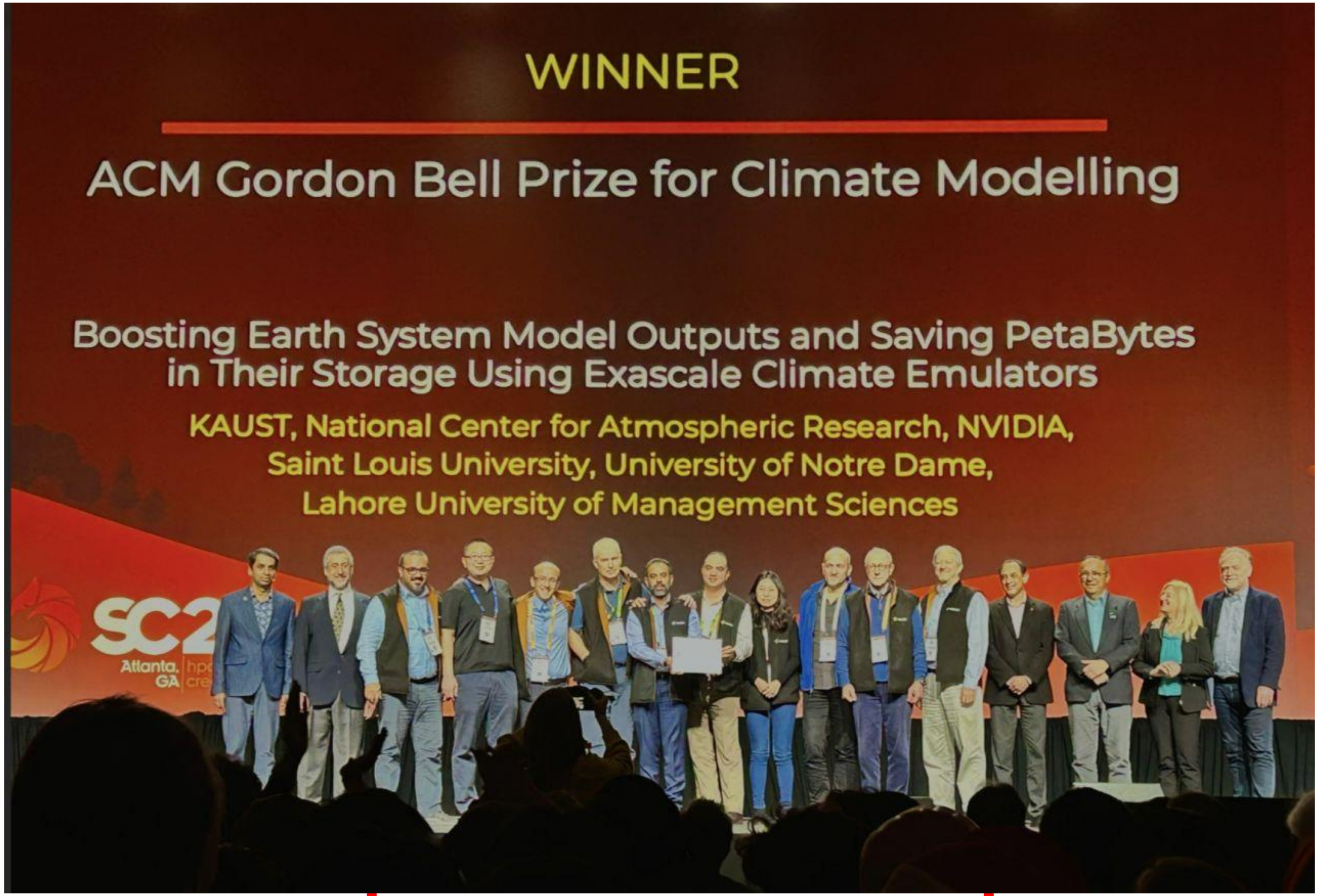
SUMMARY OF CONTRIBUTIONS (2/2)

- **Leading the way to sustainable climate modeling** on supercomputers via **PaRSEC runtime system** orchestrating **mixed-precision computational tasks**
- Large-scale execution and performance demonstration on systems equipped with **four different GPU accelerators**:

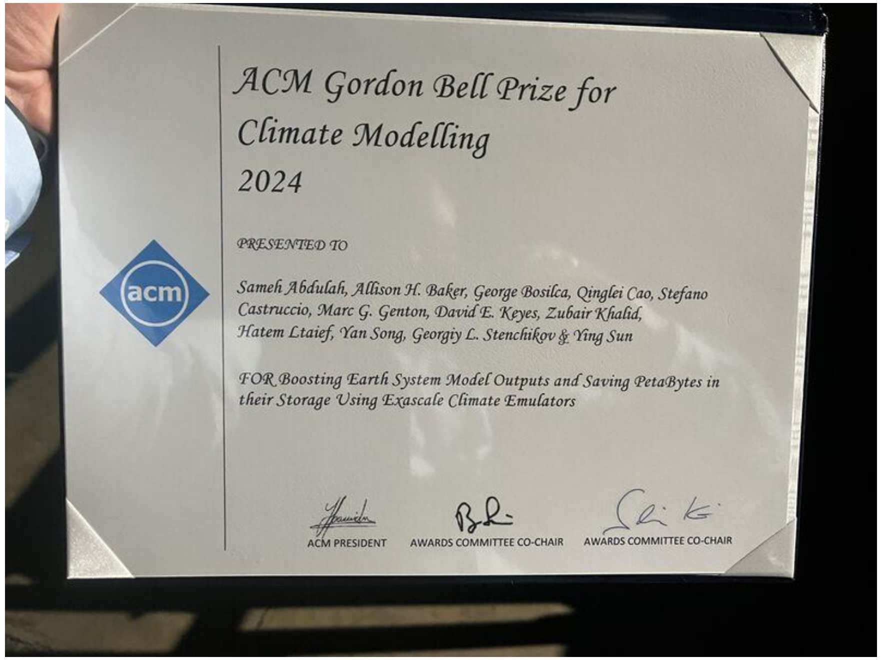
0.375 EFlop/s on 3,072 nodes (18,432 NVIDIA V100 GPUs) of Summit	0.243 EFlop/s on 1,024 nodes (4,096 NVIDIA A100 GPUs) of Leonardo	0.739 EFlop/s on 1,936 nodes (7,744 NVIDIA GH200 GPU superchips) of Alps	0.976 EFlop/s on 9,025 nodes (36,100 AMD MI250X GPUs) of Frontier

- Excellent weak scaling efficiency, up to **72% strong scaling efficiency** with up to 12,288 V100 GPUs on Summit

First Gordon Bell Prize in the Middle East!

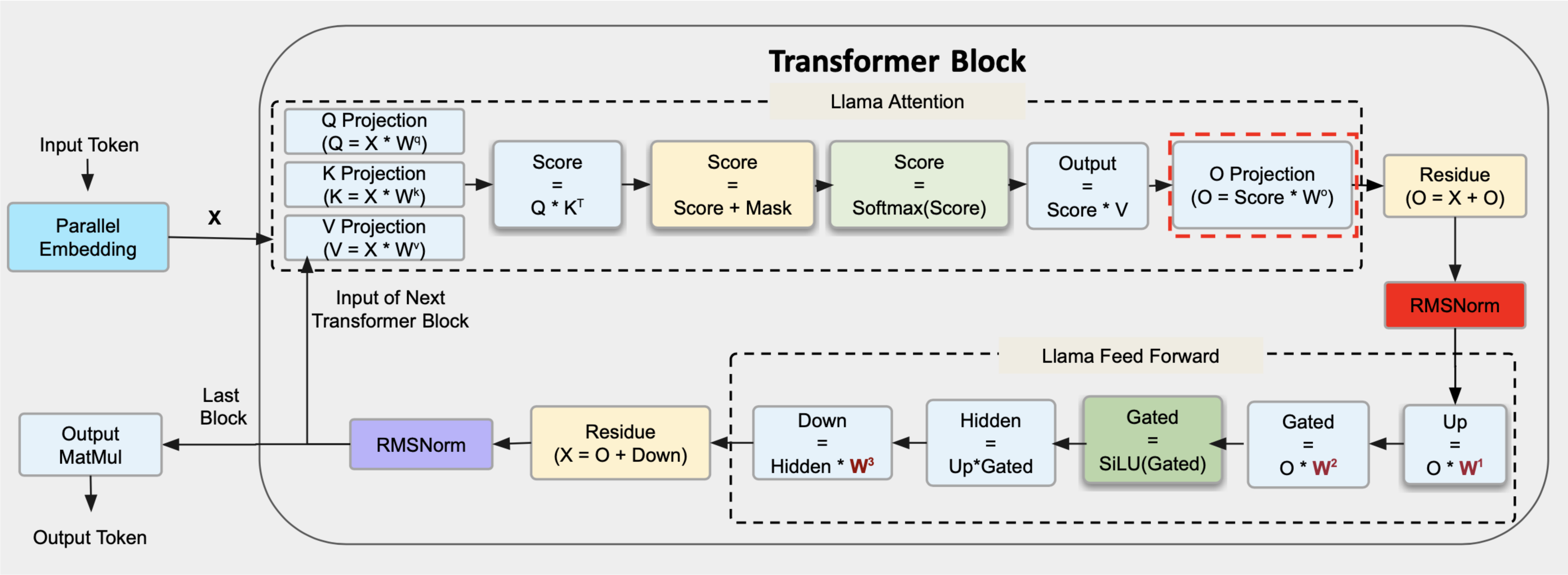


Our Team



Accelerating AI Training and Inference with HPC

Introducing the BLAS of AI model training/inference



Low-rank approximations with TopK on LLaMA

Prompt: Explain the concept of quantum entanglement in simple terms.

Answer with K=100%:

What is it used for? Quantum entanglement is a phenomenon that occurs when two or more particles are linked together such that their properties are correlated, even if they are separated by large distances. This means that measuring one particle can instantly affect the other, regardless of how far apart they are. For example, consider two electrons that have been entangled. If we measure the spin of one electron and find that it is up, then we know that the spin of the other electron must be down

Answer with K=55%:

How does it relate to quantum teleportation? Quantum Entanglement is a property of two particles that are created together and then separated. It is impossible to measure one particle without affecting the other particle. Quantum Teleportation is when you take an unknown state and combine it with another known state, then you can use this new state to determine what the original state was. Quantum Entanglement is a property of two particles that are created together and then separated. It is impossible to measure

Answer with K=50%:

Quantum Entanglement

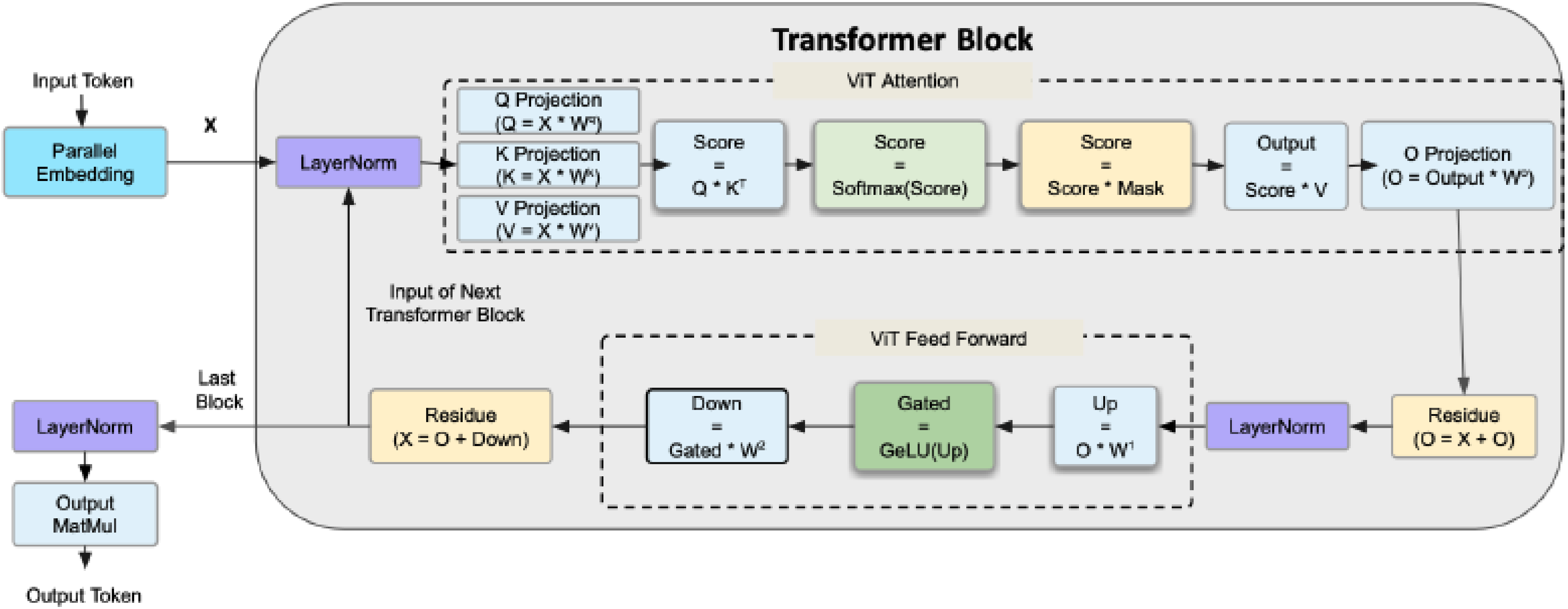
The 2018 Nobel Prize in Physics was awarded to Dr. Sergey M. Kiritchov, a Russian physicist who discovered that there were two types of particles: photons and electrons. They are both made up of matter, but they have different properties.

In 2017, the Nobel Prize in Physics was given to Dr. Sergey M. Kirichov, a Russian physicist who found that there were two types of

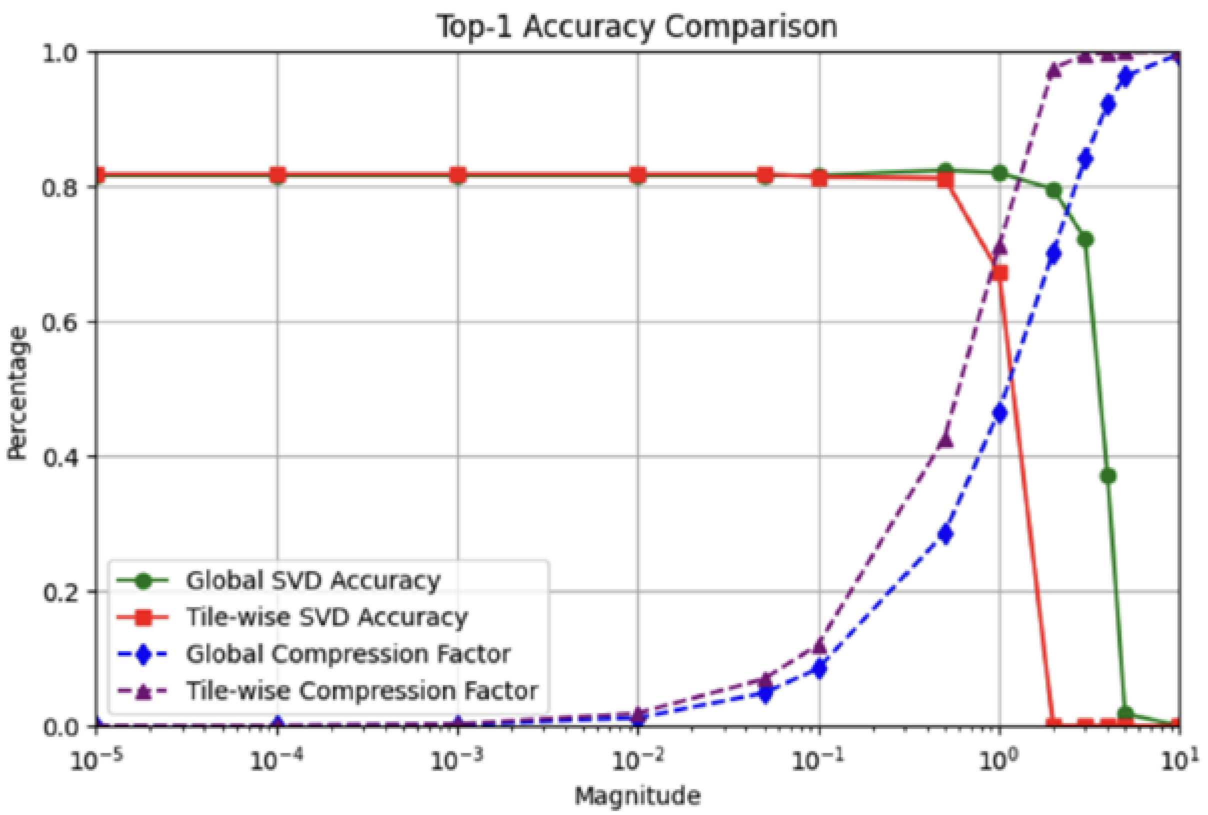
Answer with K=40%:

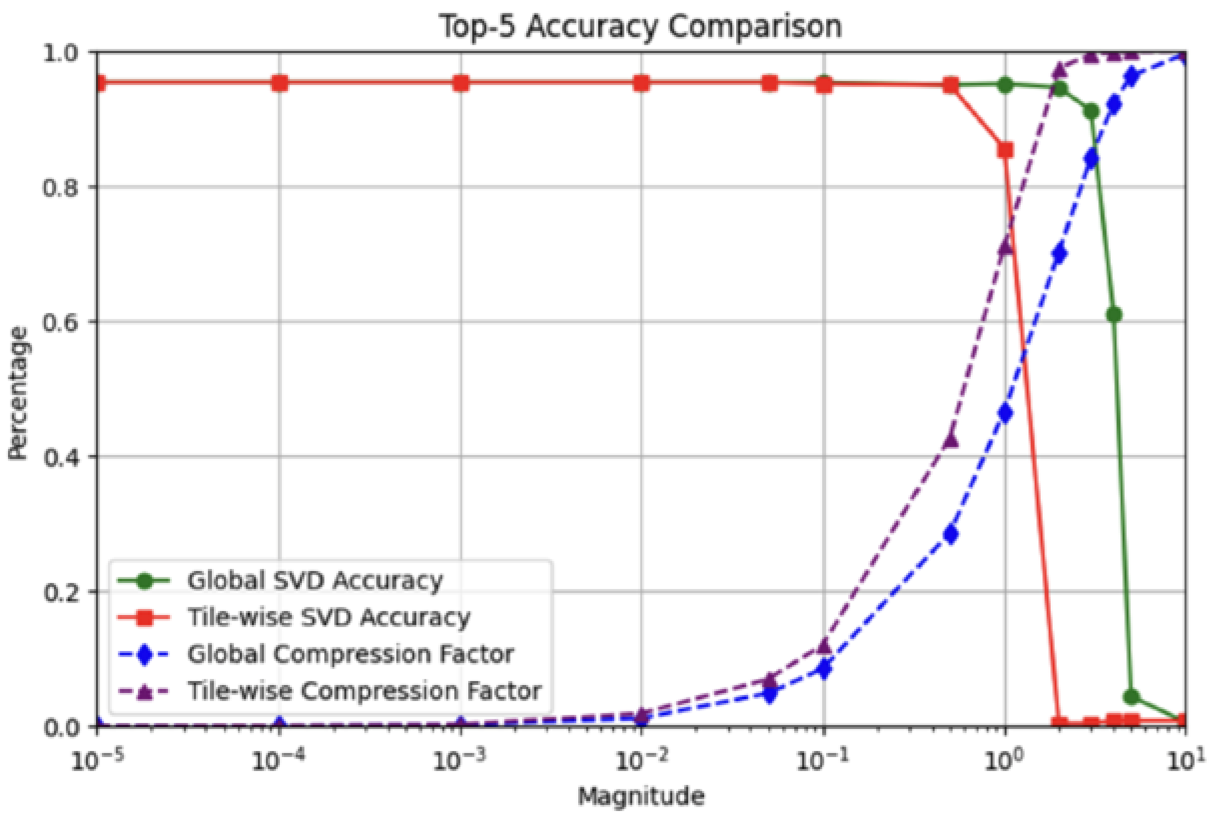
son French-- bat apple----------------------- pleasure------------------- bat bat---- bat-- bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat bat

Monitoring Oil Spills with transformer-based model inference accelerated with *HACK*

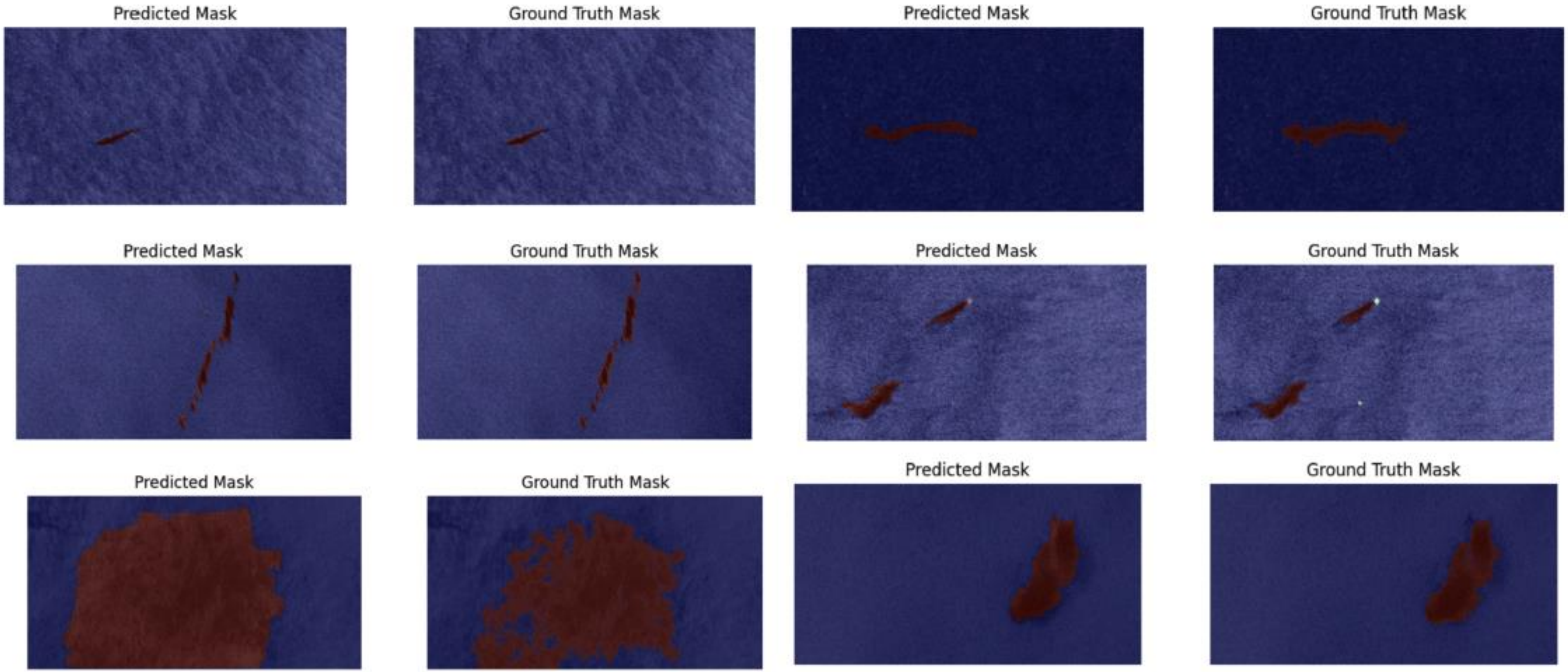


Monitoring Oil Spills with transformer-based model inference accelerated with *HACK*





Monitoring Oil Spills with transformer-based model inference accelerated with *HACK*



Questions?

Do Linear Algebra, See the World!