جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Right Precision, Right Place: Rethinking HPC Applications with Adaptive Mixed Precisions

Hatem Ltaief
Principal Research Scientist

King Abdullah University of Science and Technology

Workshop on Mixed- and Adaptive-Precision Numerics for Scientific Computing (MxP4S)

May 25 2026

Acknowledgments

Academic Collaborators:

KAUST: S. Abdulah, R. Alomairy (now at MIT), S. Bougoffa, M. G. Genton, Y. Hong (now at IU), D. E. Keyes, P. Nag, M. Ravasi (now at ShearWater), J. Ren, and Y. Sun

UTK: Q. Cao (now at SLU), G. Bosilca (now at NVIDIA), and J. Dongarra

Vendor Collaborators:

NVIDIA: R. Abdelkhalek, G. Paciucci, D. Ruau, and L. Slim

Cerebras: M. Jacquelin and L. Wilson

AMD: M. Gontier and M. Jaghoub

Resource Allocations:

Shaheen-2/3 and Ibex @ KAUST Supercomputing Lab, Saudi Arabia

HAWK @ HLRS, Germany (M. Resch)

Frontier @ ORNL, USA (US DOE)

Summit @ ORNL, USA (US DOE)

Fugaku @ Riken, Japan (S. Matsuoka)

Leonardo @ CINECA, Italy (G. Scipione)

Fromage @ NVIDIA, UK (F. Spiga)

AlgoDoers Team:

Valentin Le Fèvre

Ronald Kriemann

Vincent Etienne

Amik St-Cyr

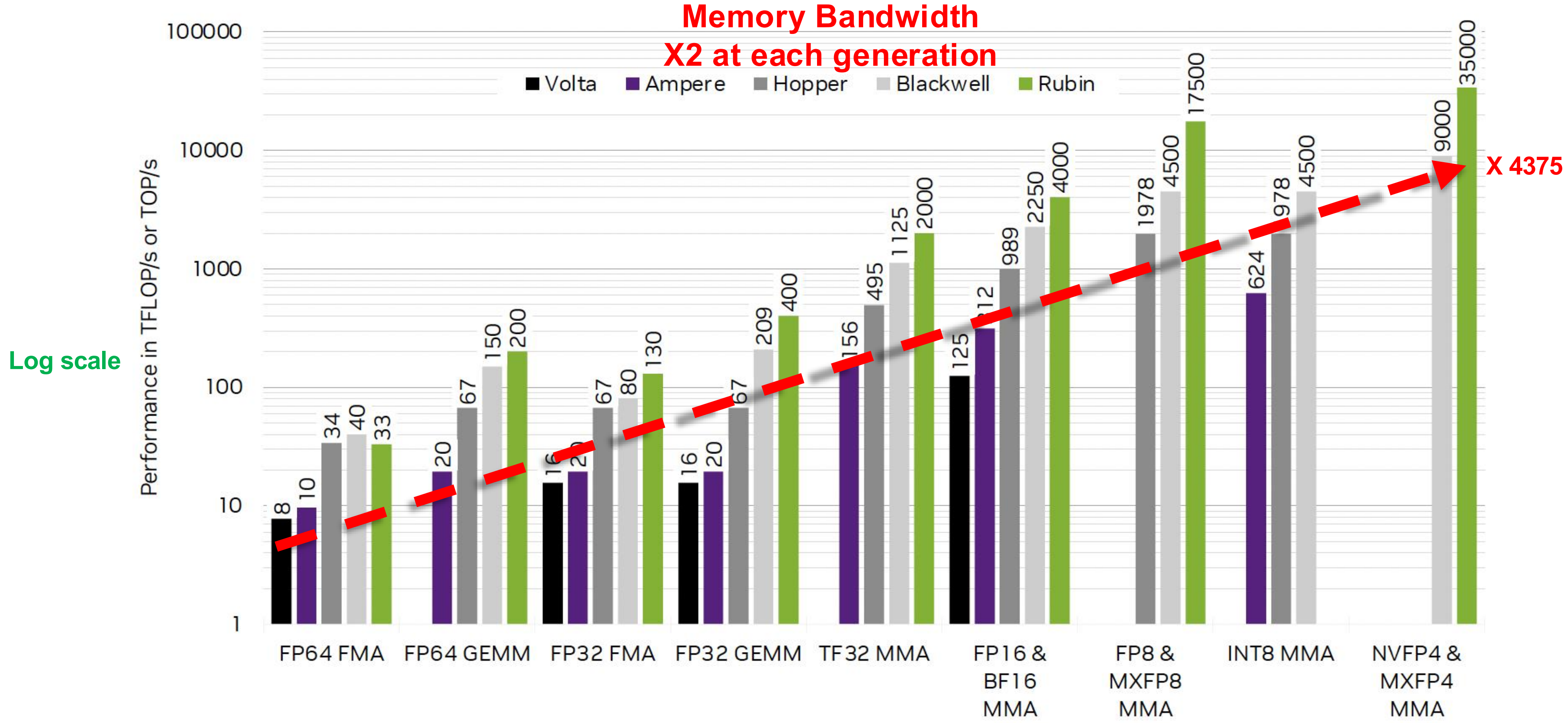
Ibtissem Kacem

ALGO DOERS

Top10 Fastest Supercomputers for HPL-MxP (Nov'25)

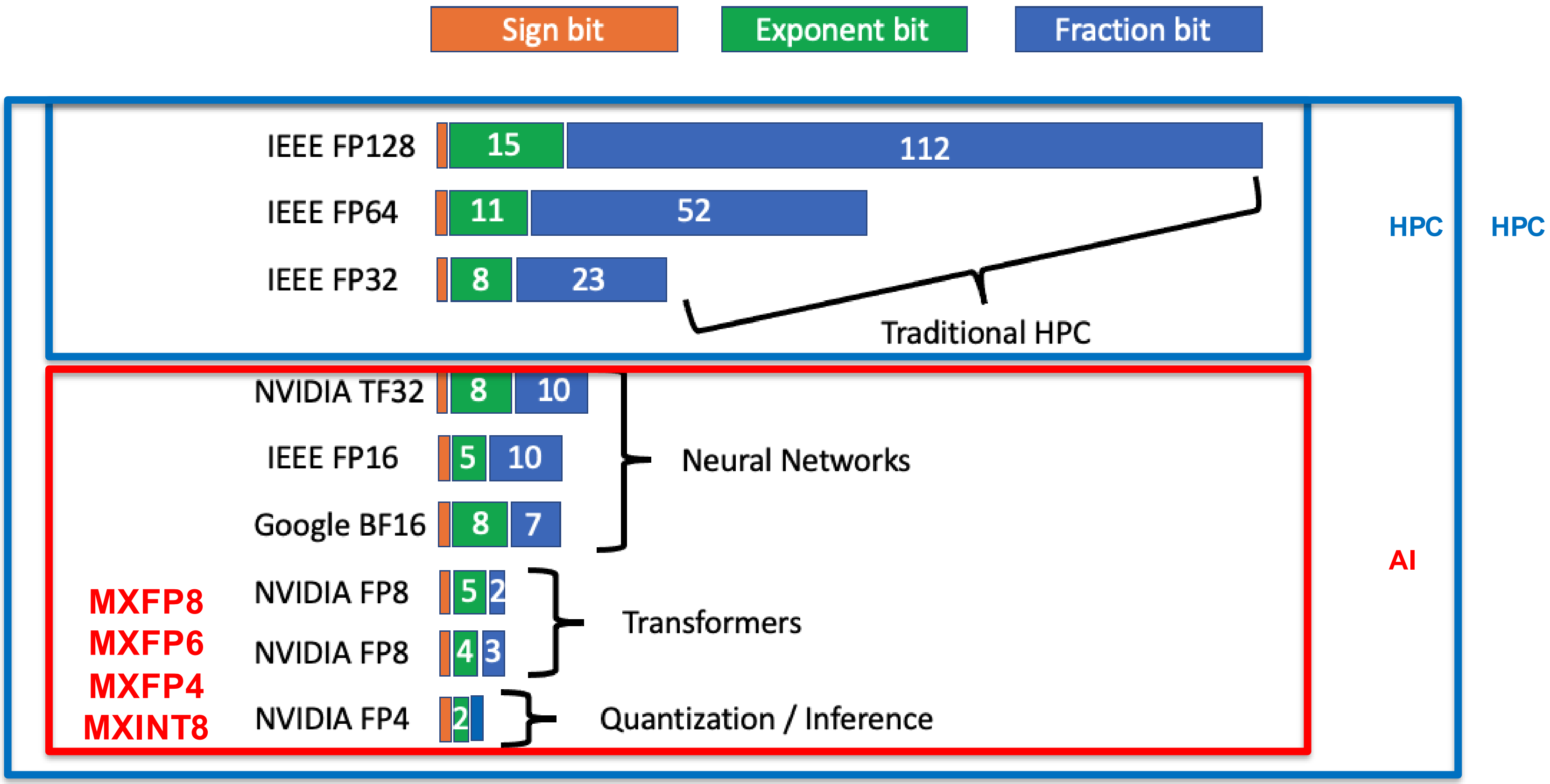
Rank	Site	Computer	Cores	HPL Rmax (Eflop/s)	TOP500 Rank	HPL-MxP (Eflop/s)	Speedup
1	DOE/SC/LLNL **USA**	**El Capitan**, HPE Cray 255a, AMD 4th Gen EPYC 24C 1.8 GHz, AMD Instinct MI300A, Slingshot-11	11,340,000	1.809	1	16.7	9.6
2	DOE/SC/ANL **USA**	**Aurora**, HPE Cray EX, Intel Max 9470 52C, 2.4 GHz, Intel GPU MAX, Slingshot-11	8,159,232	1.012	3	11.6	11.5
3	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
4	EuroHPC/FZJ **Germany**	**JUPITER Booster**, BullSequana XH3000, NVIDIA Grace 72C 3GHz, GH200, NVIDIA Infiniband NDR200	4,801,344	1.0	4	6.25	6.3
5	Softbank **Japan**	**CHIE-4**, NVIDIA DGX B200, Xeon Platinum 8570 56C 2.1 GHz, B200 SXM 180GB, Infiniband NDR400	662,256	0.135	17	3.3	24.4
6	AIST **Japan**	**ABCI 3.0**, HPE Cray XD670, Xeon Platinum 48C 2.1GHz, NVIDIA H200 SXM5 141 GB, Infiniband NDR200	479,232	0.145	16	2.36_8 1.2_{16}	16.3_8 8.3_{16}
7	EuroHPC/CSC **Finland**	**LUMI**, HPE Cray EX235a, AMD Zen-3 (Milan) 64C 2GHz, AMD MI250X, Slingshot-11	2,752,704	0.380	9	2.35	6.2
8	RIKEN Center for Compu-tational Science, **Japan**	**Fugaku**, Fujitsu A64FX 48C 2.2GHz, Tofu D	7,630,848	0.442	7	2.0	4.5
9	EuroHPC/CINECA **Italy**	**Leonardo**, BullSequana XH2000, Xeon Pl. 32C 2.6GHz, NVIDIA A100 SXM4 40 GB, QR NVIDIA HDR100 IB	1,824,768	0.241	10	1.8	7.6
10	KAIST **Saudi Arabia**	**Shaheen III**, HOE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200, Slingshot-11	574,464	0.123	18	0.7	5.7

FP/TF/INT Performance of 5 Generations of NVIDIA GPUs

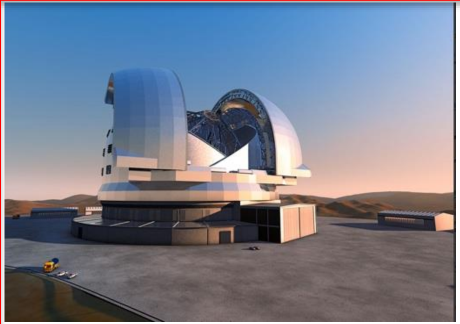


Slide courtesy
H. Bayraktar, NVIDIA

Feeling like a kid in a candy store

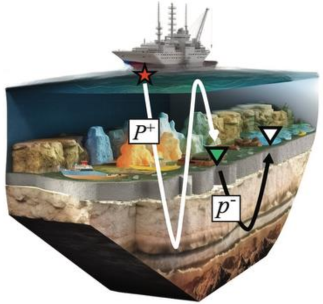


Four Challenging Scientific Applications



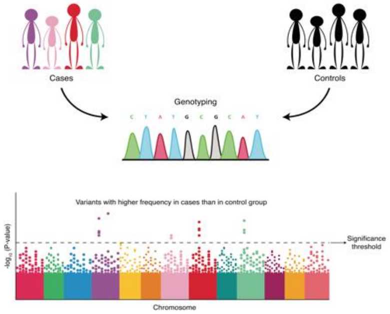
Computational Astronomy

Best Paper
PASC18



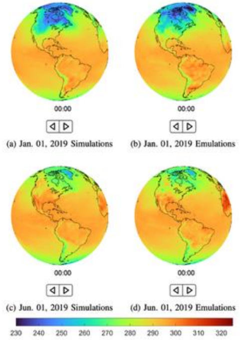
Seismic Processing

Gordon Bell Finalist
SC23



Computational Biology

Gordon Bell Finalist
SC24



Climate

Simulations / Emulations

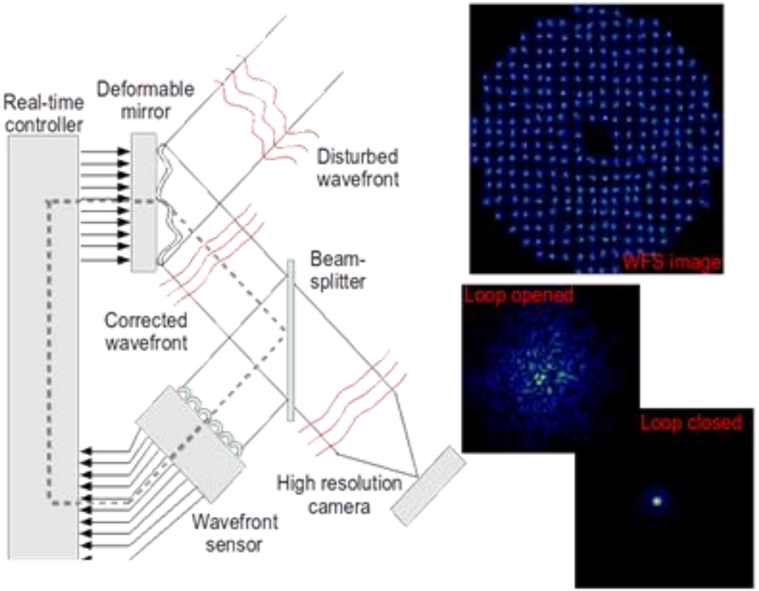
Gordon Bell Finalist SC22
Gordon Bell Winner SC24

Adaptive optics: outsmarting the atmospheric turbulence

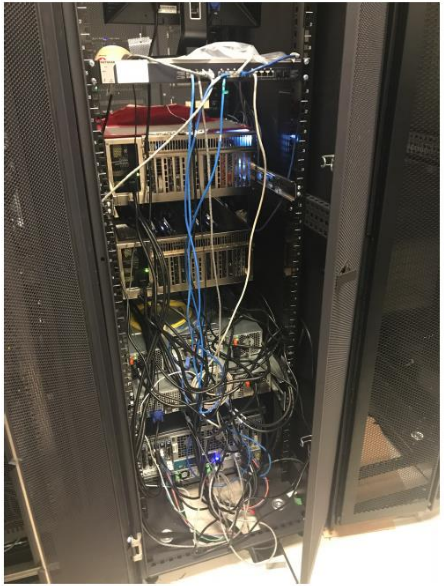
The sun observed with a compact camera

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

Adaptive Optics: tomographic reconstructor, real-time controller



The Advanced Machinery

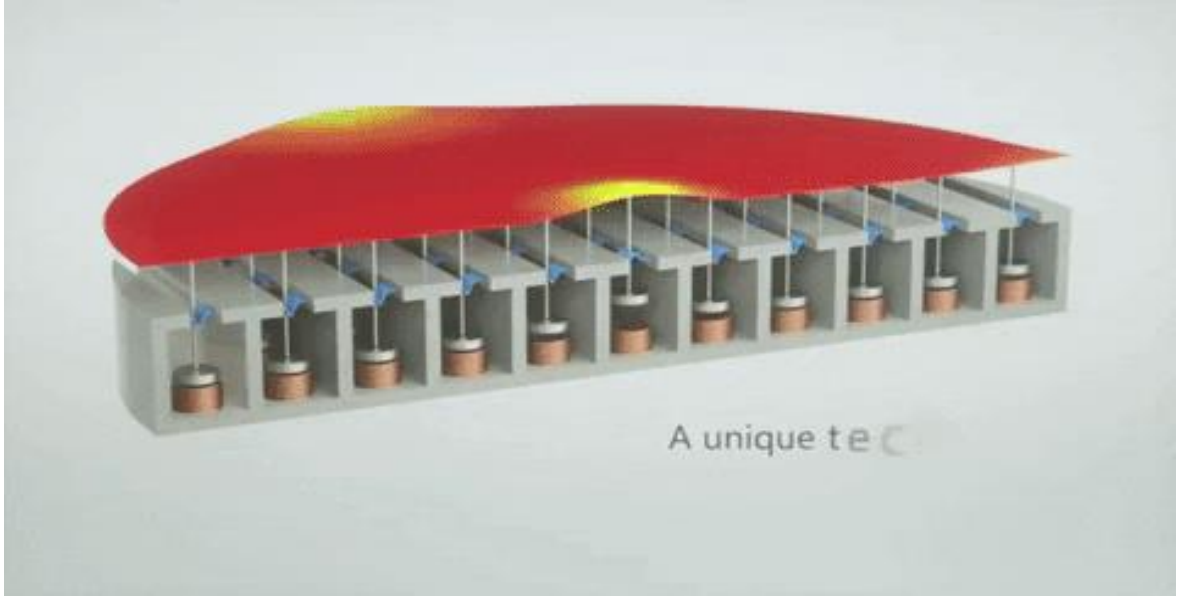


Operating at 14,000 feet

Surviving at 14,000 feet

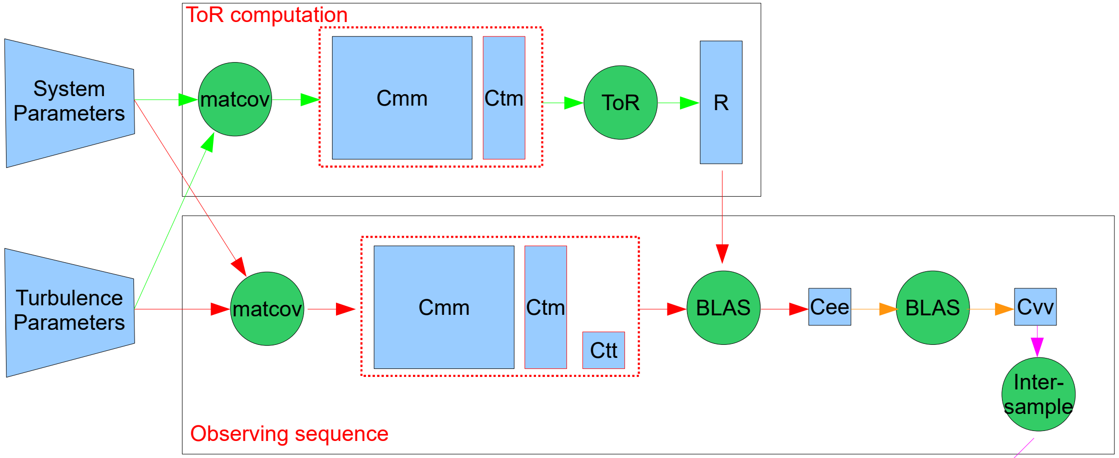
N. Doucet, HL, D. Gratadour and D. Keyes, *Mixed-Precision Tomographic Reconstructor Computations on Hardware Accelerators*, IEEE/ACM SC19 IA3 Workshop.
HL, D. Sukkari, O. Guyon, and D. Keyes, *Extreme Computing for Extreme Adaptive Optics: The Key to Finding Life Outside our Solar System*, ACM PASC18.
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.

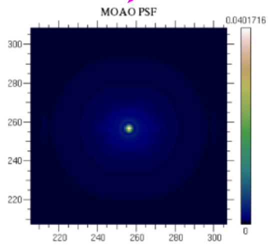
Adaptive Optics: deformable mirrors



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

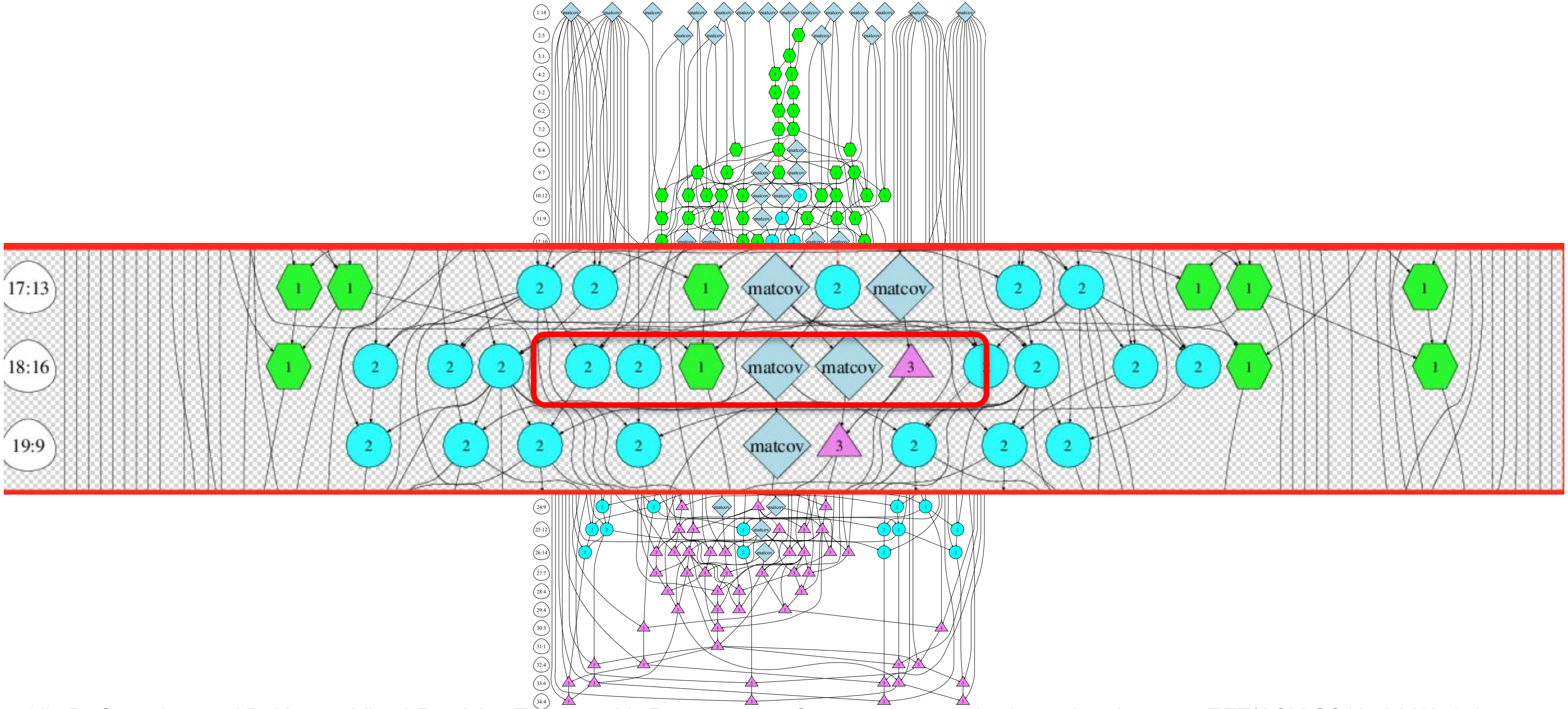
$$C_{ee} = C_{tt} - C_{tm}R^t - RC_{tm}^t + RC_{mm}R^t$$

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



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.

The *First* MxP Cholesky Factorization

Algorithm 1 Mixed-Precision Cholesky factorization.

```
for k=0..NT-1 do
    A[k][k] ← spotrf(A[k][k])
    for m=k+1..NT-1 do
        A[m][k] ← strsm(A[k][k], A[m][k])
        H[m][k] ← SP2HP(A[m][k])
    end for
    for n=k+1..NT-1 do
        A[n][n] ← ssyrk(A[n][k], A[n][n])
        for m=n+1..NT-1 do
            precision ← heuristic(m, n)
            if precision == float then
                A[m][n] ← sgemm(A[m][k], A[n][k], A[m][n])
            else if precision == half then
                if variant == 1 or 2 then
                    H[m][n] ← SP2HP(A[m][n])
                    H[m][n] ← hgemm(H[m][k], H[n][k], H[m][n])
                    A[m][n] ← SP2HP(H[m][n])
                else
                    A[m][n] ← hgemm(H[m][k], H[n][k], A[m][n])
                end if
            end if
        end for
    end for
end for
```

TABLE I
32/16-BIT MIXED PRECISION CUBLASGEMMEX SUPPORTED FUNCTION.

GEMM variants	A/B type	C type	Compute type	Alpha / Beta
HP3	16-bit	16-bit	16-bit	16-bit
HP2	16-bit	16-bit	32-bit	32-bit
HP1	16-bit	32-bit	32-bit	32-bit
SP	32-bit	32-bit	32-bit	32-bit

The first column will be used to specify the function's precision in the following.

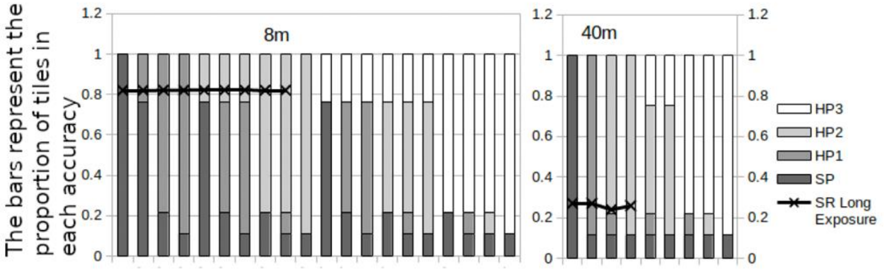
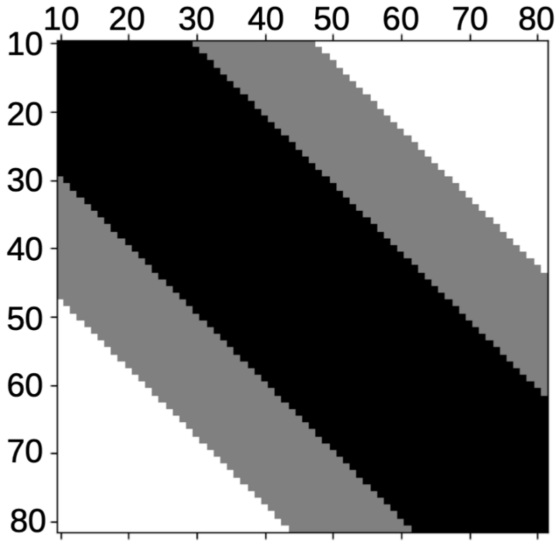
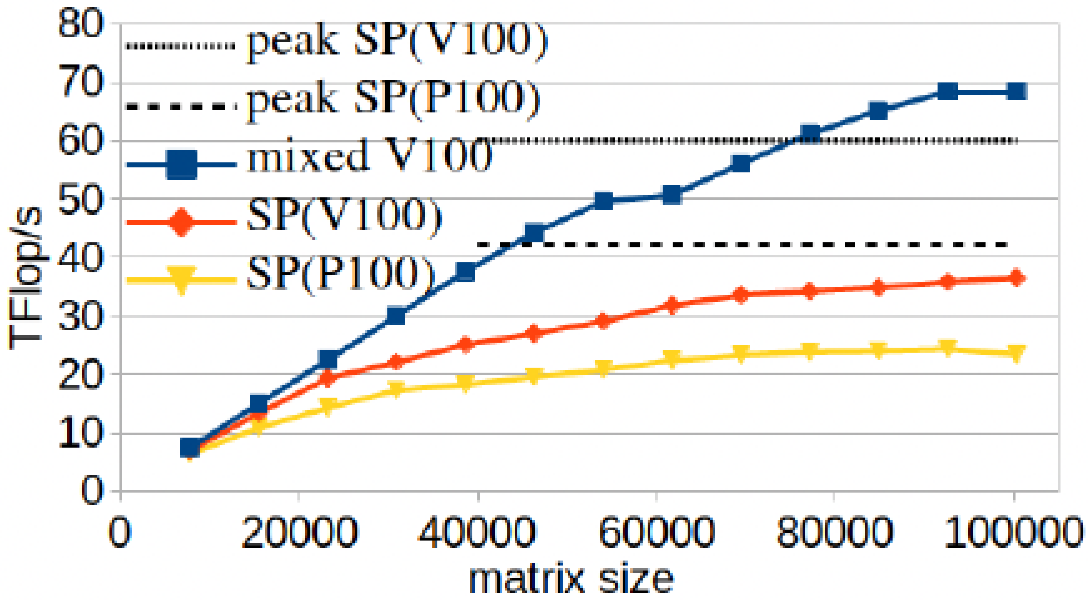


Fig. 5. Proportion of the different GEMM precision and associated ToR accuracy on a eight meter telescope with $17k$ measurements (left) and an forty meter telescope with $50k$ measurements (right).

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.

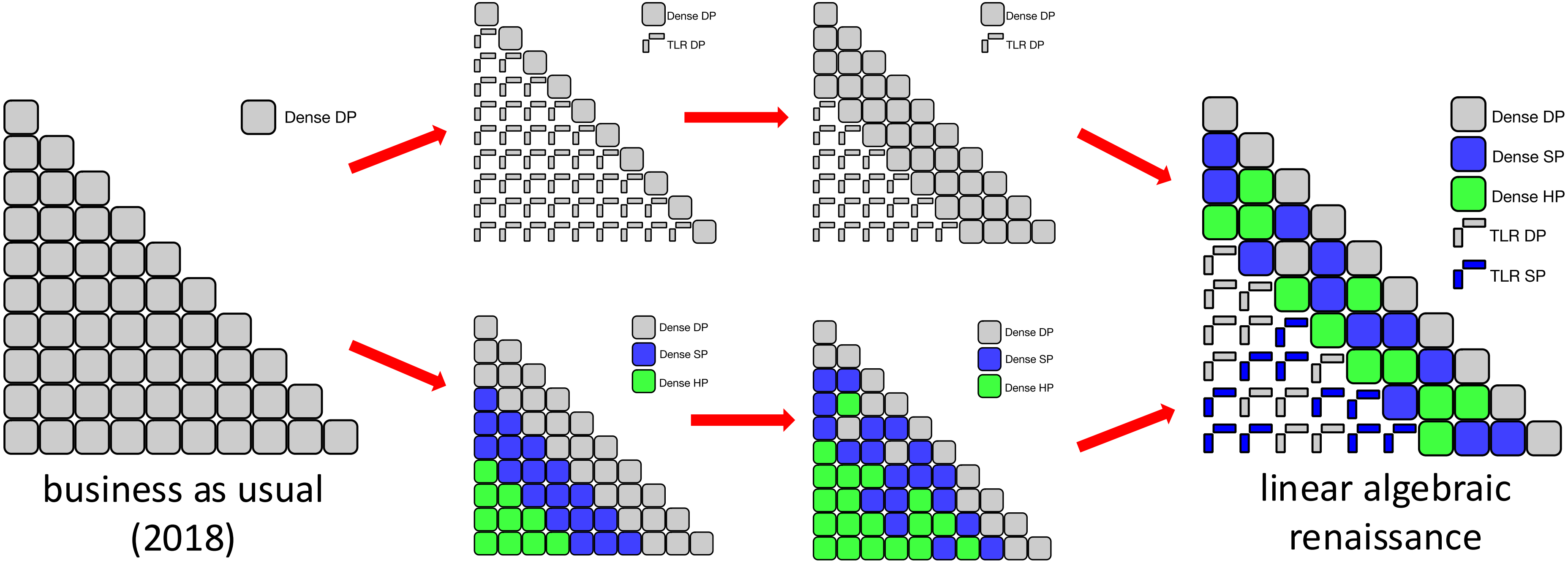
Adaptive Optics: tomographic reconstructor, real-time controller and deformable mirrors





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.

The journey toward linear algebra *Renaissance*

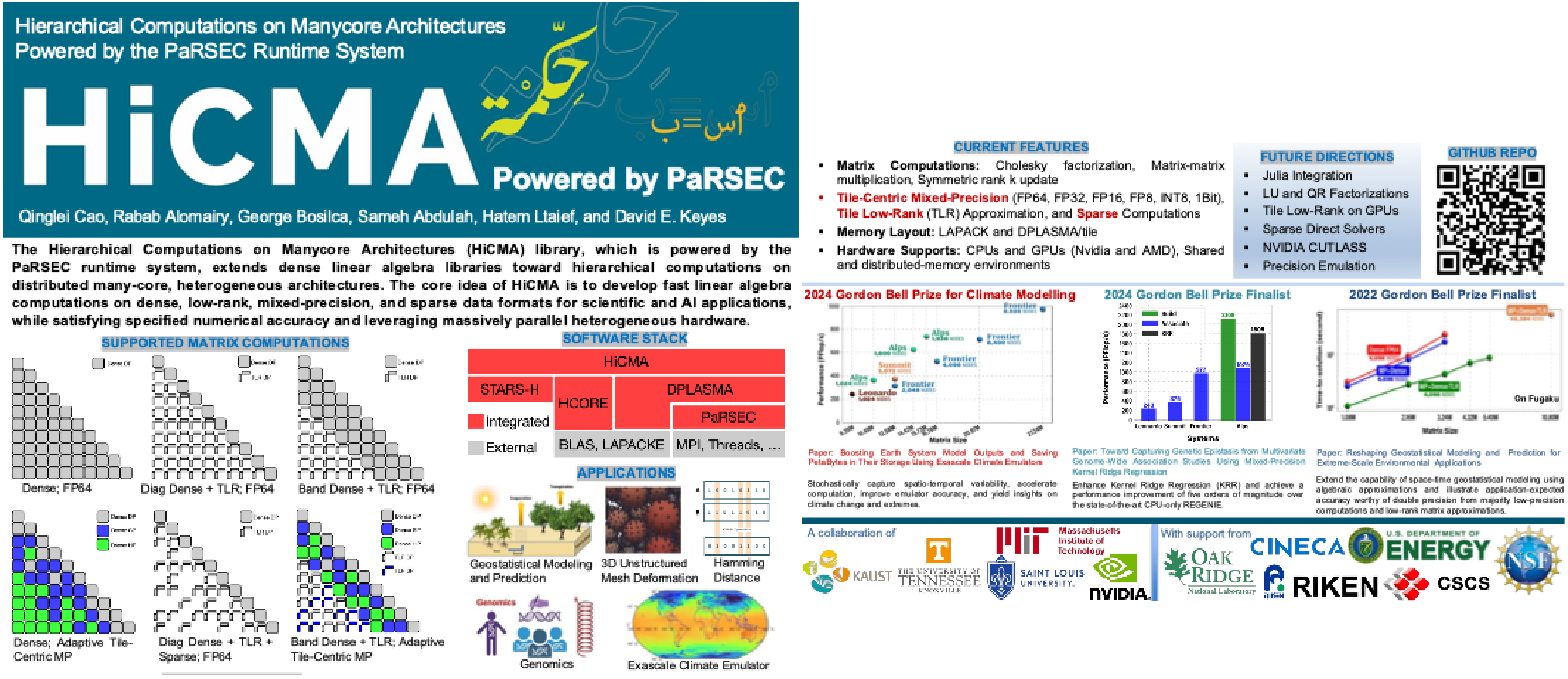


Don’t over-solve: maintain just enough accuracy for the application purpose

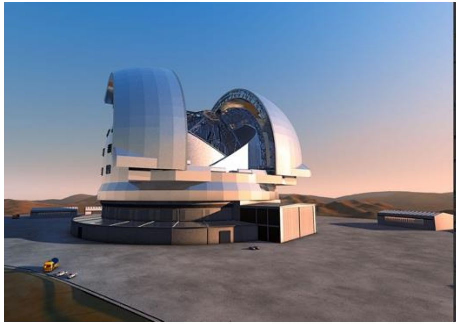
Don’t over-store: no extra copies of the original matrix

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.

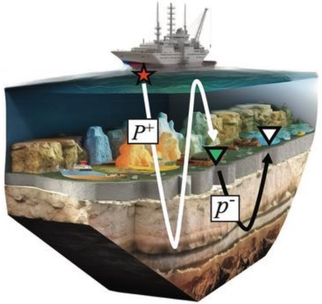
HiCMA Powered by PaRSEC: Released at SC25



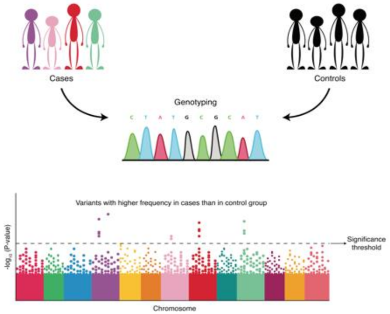
Four Challenging Scientific Applications



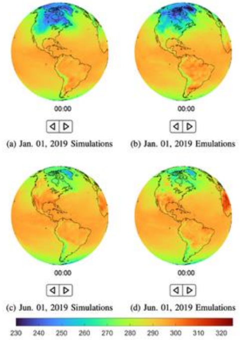
Computational Astronomy



Seismic Processing



Computational Biology



Climate
Simulations / Emulations

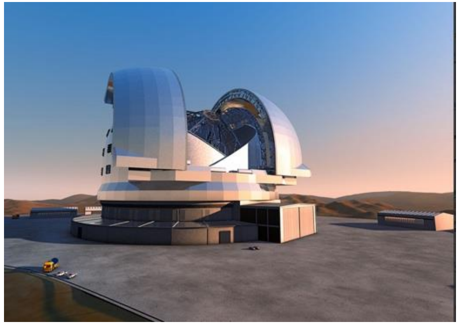
Best Paper
PASC18

Gordon Bell Finalist
SC23

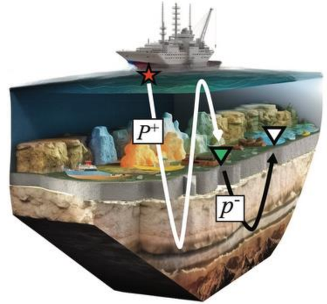
Gordon Bell Finalist
SC24

Gordon Bell Finalist SC22
Gordon Bell Winner SC24

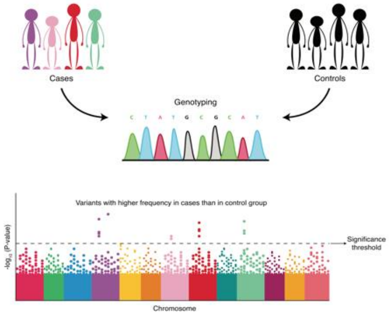
Four Challenging Scientific Applications



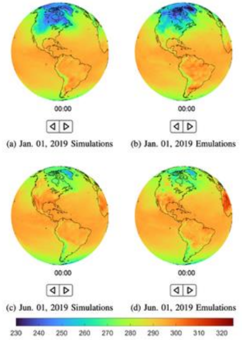
Computational Astronomy



Seismic Processing



Computational Biology



Climate
Simulations / Emulations

Best Paper
PASC18

Gordon Bell Finalist
SC23

Gordon Bell Finalist
SC24

Gordon Bell Finalist SC22
Gordon Bell Winner SC24

Challenges with ML-based Climate/Weather Forecasting Simulations

- Today, weather and climate data are usually huge!
  - A set of univariate/multivariate Z observations $\in \mathbb{R}^n$ at given n locations on one or more time slots.
  - Z observations could be temperature, precipitation, humidity, air pollution ... etc
- Maximum Likelihood Function: a statistical approach used to estimate the parameters of a model. This method is particularly useful in spatial analysis due to its ability to handle complex spatial dependencies and heterogeneity

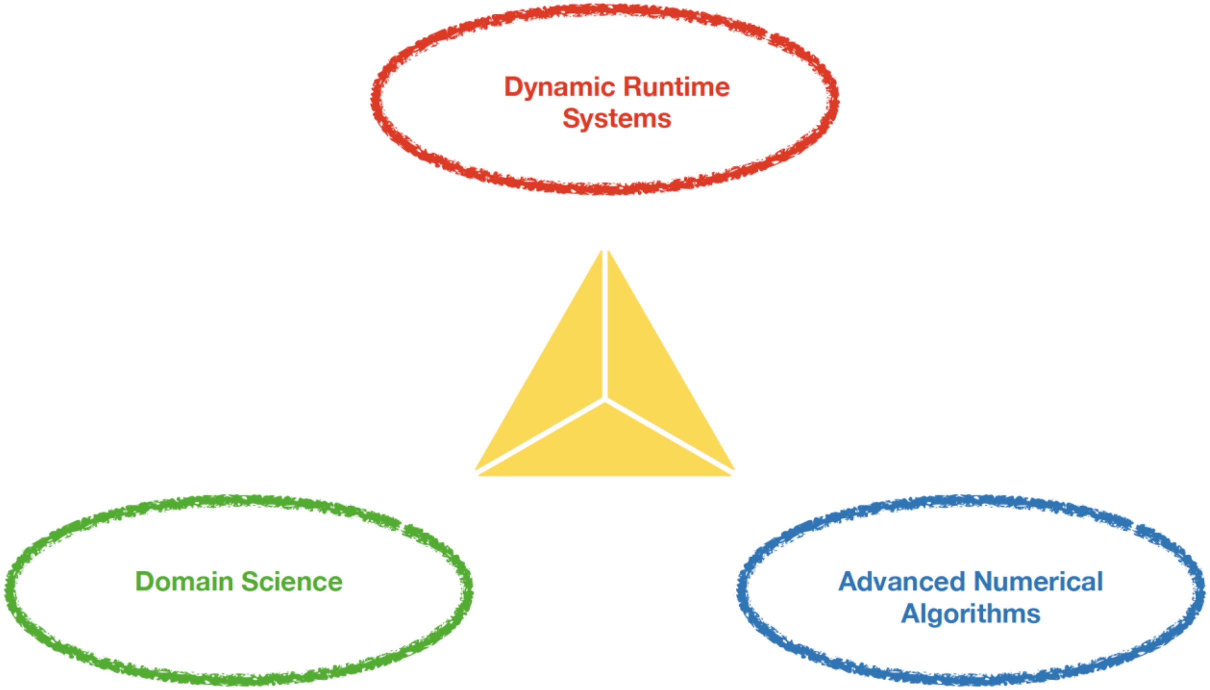
$$\boldsymbol{\Sigma}(\boldsymbol{\theta})$$

  - Prohibitive computational Cost and memory requirements:

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

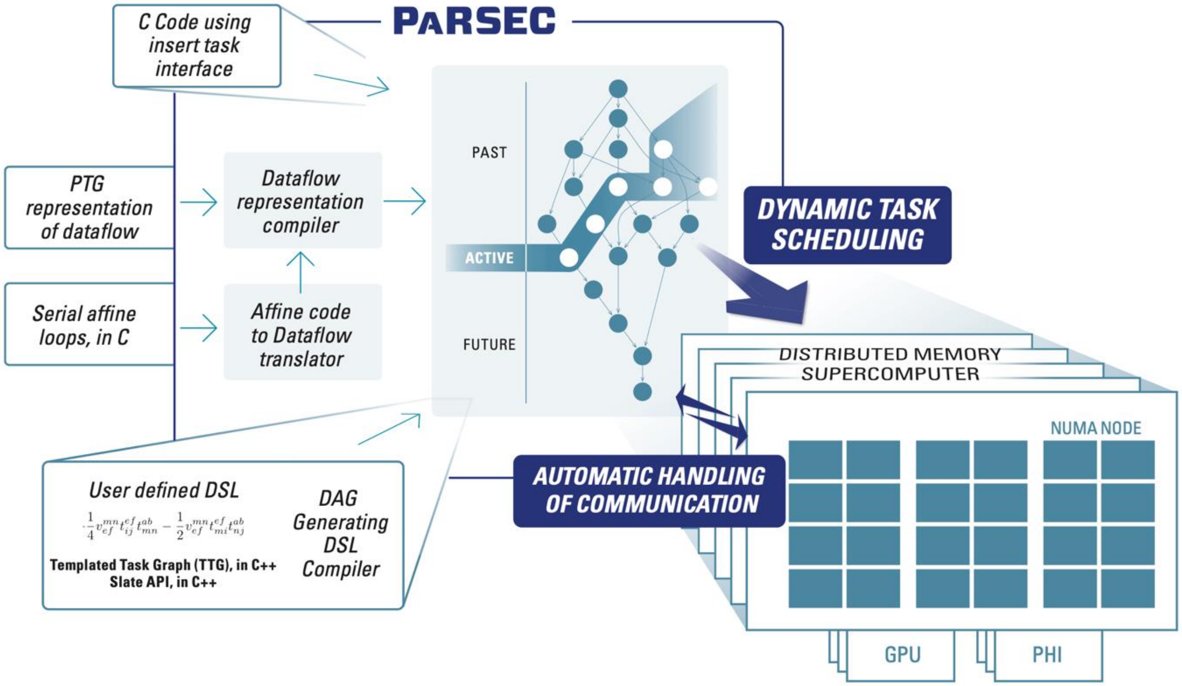
  - The linear solver and log-determinant involving floating point operations on n-by-n covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ with $O(n^3)$ complexity and $O(n^2)$ memory footprint
  - For instance: 10^6 locations require 8TB Memory!

Key Approach Based on a Separation of Concerns



PaRSEC: a dynamic runtime system of the DOE ECP

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

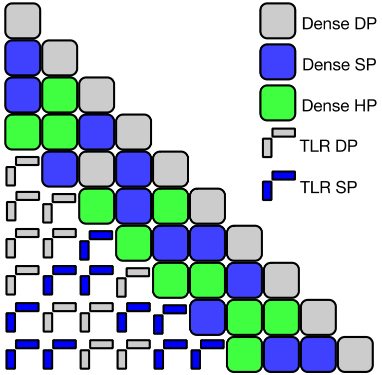


https://icl.utk.edu/parsec/

Elapsed Time and Energy Consumption on Shaheen-III

	Problem	Time (s)	Time (s) SC		Time (s) MC		Time (s) WC	
# Nodes	Matrix Size	**Dense FP64**	**Dense+MP**	**TLR+MP**	**Dense+MP**	**TLR+MP**	**Dense+MP**	**TLR+MP**
256	2.16M	2859	2847	1036	2180	921	1600	891
512	3.24M	4745	4742	1791	3009	1579	2649	1470
1024	4.32M	4910	4907	2213	4361	1906	3191	1767
2048	6.22M	6202	6207	3890	5575	3206	4962	2456

2020 MJ **1065 MJ** **575 MJ**



linear algebraic renaissance (2022)

Sustainably Modeling a Sustainable Future Climate, R. Alomairy, S. Abdulah, Q. Cao, M. G. Genton, D. E. Keyes, HL, IEEE HPEC, 2025.

Can we do more? Microscaling coming to rescue

Open Compute Project • OCP Microscaling Formats (MX) Specification

OPEN Compute Project

OCP Microscaling Formats (MX) Specification

Version 1.0

1. License

Contributions to this Specification are made under the terms and conditions set forth in Open Web Foundation Modified Contributor License Agreement ("OWF CLA 1.0") ("Contribution License") by:

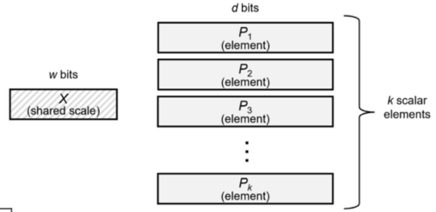
AMD, Arm, Intel, Meta, Microsoft, NVIDIA, Qualcomm

Usage of this Specification is governed by the terms and conditions set forth in **Open Web Foundation Modified Final Specification Agreement ("OWFa 1.0") ("Specification License").**

5.1 Microscaling (MX)

An MX-compliant format is characterized by three components:

- Scale (X) data type / encoding
- Private elements (P_i) data type / encoding
- Scaling block size (k)

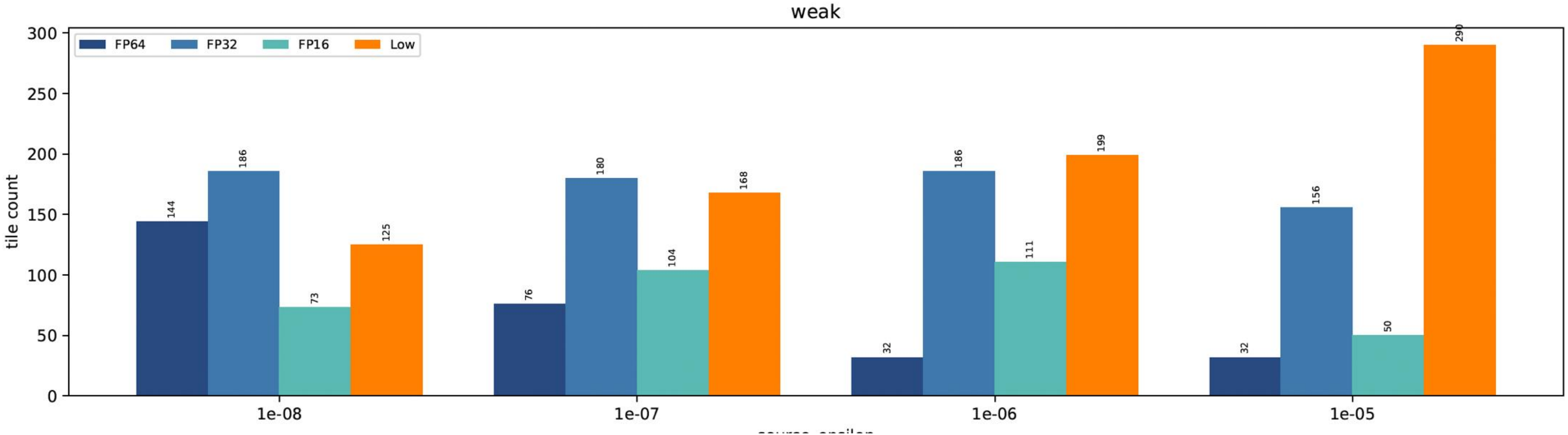


Format Name	Element Data Type	Element Bits (d)	Scaling Block Size (k)	Scale Data Type	Scale Bits (w)
MXFP8	FP8 (E5M2)	8	32	E8M0	8
	FP8 (E4M3)				
MXFP6	FP6 (E3M2)	6	32	E8M0	8
	FP6 (E2M3)				
MXFP4	FP4 (E2M1)	4	32	E8M0	8
MXINT8	INT8	8	32	E8M0	8

Table 1. Format names and parameters of concrete MX-compliant formats.

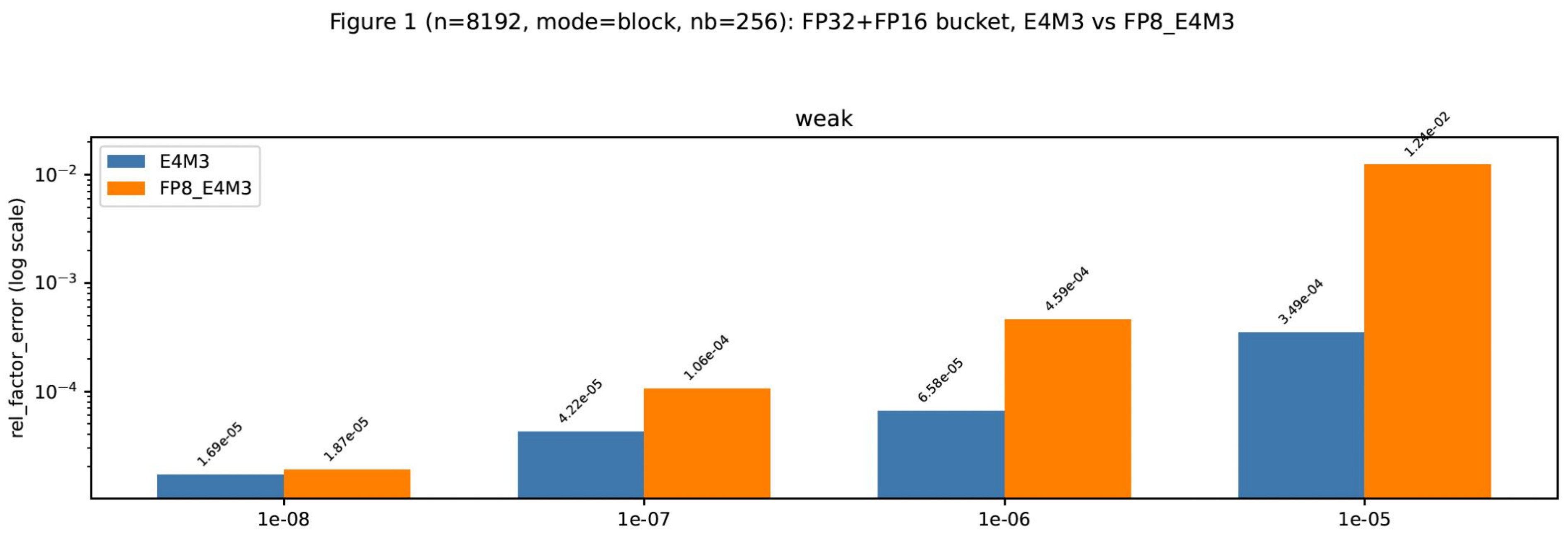
Can we do more? Microscaling coming to rescue

Figure 2 (n=8192, mode=block, nb=256): tile counts for FP32+FP16 bucket (FP64, FP32, FP16, Low)



w/ A. Abdurakhmanov / S. Fahmy / D. Keyes @KAUST and T. Mary @Sorbonne

Can we do more? Microscaling coming to rescue



w/ A. Abdurakhmanov / S. Fahmy / D. Keyes @KAUST and T. Mary @Sorbonne

Do Linear Algebra, See the World!

Thanks,

QUESTIONS?