Statistics:
A spacious home for HPC

HPSC
HIGH-PERFORMANCE STATISTICAL COMPUTING

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

David Keyes
and the HiCMA group of KAUST's
Extreme Computing Research Center

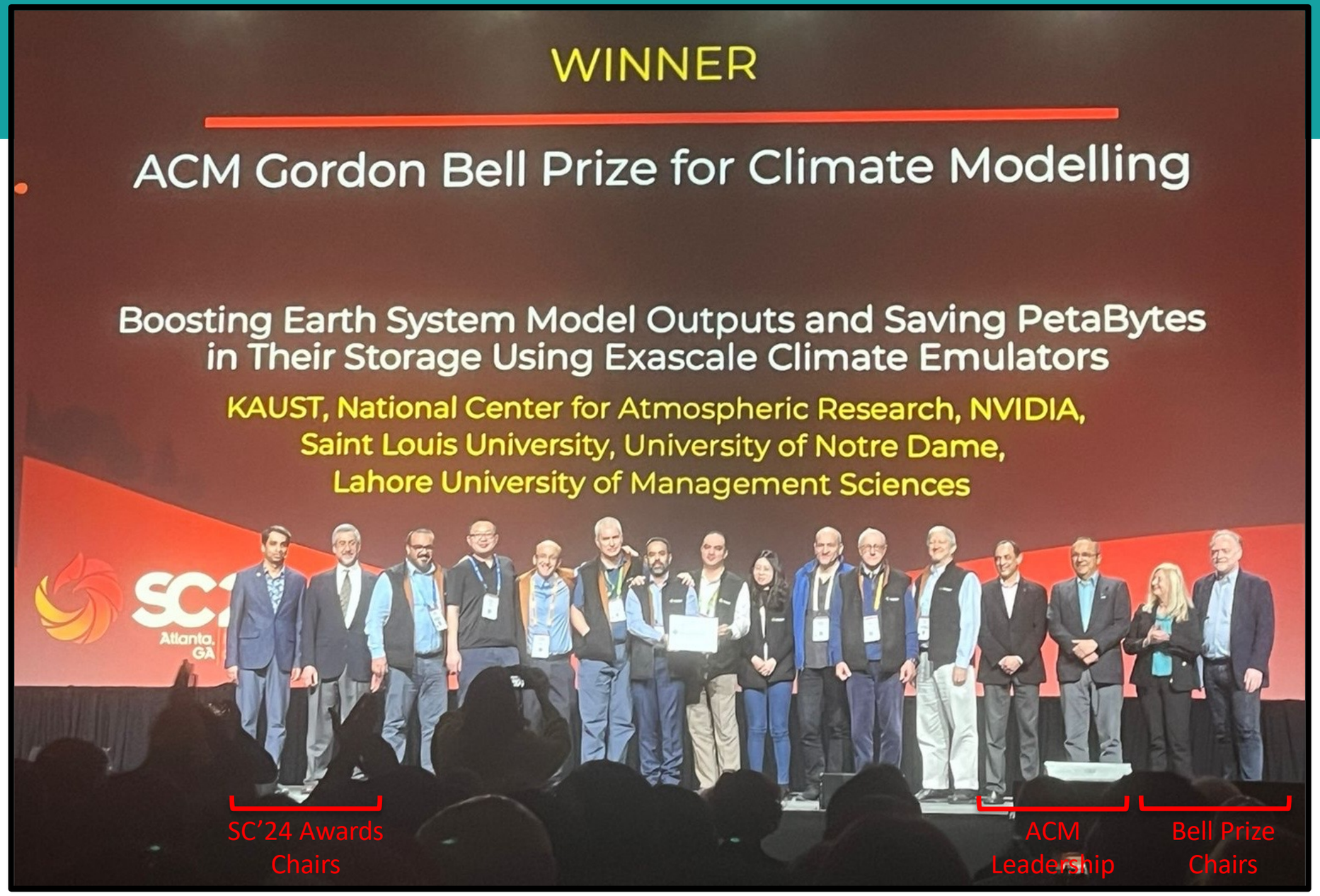
A “spacious home,” indeed!

Last week at SC’25, for the second year in a row, HPC’s top prize was awarded to a statistics application.

Many other statistics applications await their HPC re-implementation, often to accommodate bigger data.

Many other elements of the HPC ecosystem (hardware, software, algorithms) await greater fulfillment in statistics apps, often reusing existing solutions.

Nov
2025


WINNER
ACM Gordon Bell Prize for Climate Modelling
Boosting Earth System Model Outputs and Saving PetaBytes in Their Storage Using Exascale Climate Emulators
KAUST, National Center for Atmospheric Research, NVIDIA, Saint Louis University, University of Notre Dame, Lahore University of Management Sciences
Atlanta, GA
SC'24 Awards Chairs
ACM Leadership
Bell Prize Chairs

Nov 2024

Real-Time Bayesian Inference at Extreme Scale: A Digital Twin for Tsunami Early Warning Applied to the Cascadia Subduction Zone

Stefan Henneking
The University of Texas at Austin
Austin, TX, USA
stefan@oden.utexas.edu

Sreeram Venkat
The University of Texas at Austin
Austin, TX, USA
srvenkat@utexas.edu

Veselin Dobrev
Lawrence Livermore National Laboratory
Livermore, CA, USA
dobrev1@llnl.gov

John Camier
Lawrence Livermore National Laboratory
Livermore, CA, USA
camier1@llnl.gov

Tzanio Kolev
Lawrence Livermore National Laboratory
Livermore, CA, USA
kolev1@llnl.gov

Milinda Fernando
The University of Texas at Austin
Austin, TX, USA
milinda@oden.utexas.edu

Alice-Agnes Gabriel
University of California San Diego
San Diego, CA, USA
algabriel@ucsd.edu

Omar Ghattas
The University of Texas at Austin
Austin, TX, USA
omar@oden.utexas.edu

Abstract

We present a Bayesian inversion-based digital twin that employs acoustic pressure data from seafloor sensors, along with 3D coupled acoustic–gravity wave equations, to infer earthquake-induced spatiotemporal seafloor motion in real time and forecast tsunami propagation toward coastlines for early warning with quantified uncertainties. Our target is the Cascadia subduction zone, with one billion parameters. Computing the posterior mean alone would require 50 years on a 512 GPU machine. Instead, exploiting the shift invariance of the parameter-to-observable map and devising novel parallel algorithms, we induce a fast offline–online decomposition. The offline component requires just one adjoint wave propagation per sensor; using MFEM, we scale this part of the computation to the full El Capitan system (43,520 GPUs) with 92% weak parallel efficiency. Moreover, given real-time data, the online component exactly solves the Bayesian inverse and forecasting problems in 0.2 seconds on a modest GPU system, a ten-billion-fold speedup.

CCS Concepts

• **Mathematics of computing → Solvers; Mathematical software performance; Partial differential equations; Computation of transforms; Mesh generation; Discretization;** • **Computing methodologies → Massively parallel algorithms; Uncertainty quantification; Model verification and validation; Modeling methodologies; Real-time simulation; Data assimilation; Massively parallel and high-performance simulations; Scientific visualization;** • **Applied computing → Earth and atmospheric sciences; Mathematics and statistics.**

ACM ISBN 979-8-4007-1466-5/25/11
https://doi.org/10.1145/3712285.3771787

Keywords

Bayesian inverse problems, uncertainty quantification, digital twins, data assimilation, finite elements, real-time GPU supercomputing, tsunami early warning

ACM Reference Format:
Stefan Henneking, Sreeram Venkat, Veselin Dobrev, John Camier, Tzanio Kolev, Milinda Fernando, Alice-Agnes Gabriel, and Omar Ghattas. 2025. Real-Time Bayesian Inference at Extreme Scale: A Digital Twin for Tsunami Early Warning Applied to the Cascadia Subduction Zone. In *The International Conference for High Performance Computing, Networking, Storage and Analysis (SC '25), November 16–21, 2025, St Louis, MO, USA*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3712285.3771787

1 Justification for ACM Gordon Bell Prize

Fastest time-to-solution of a PDE-based Bayesian inverse problem with 1 billion parameters in 0.2 seconds, a ten-billion-fold speedup over SoA. Largest-to-date unstructured mesh FE simulation with 55.5 trillion DOF on 43,520 GPUs, with 92% weak and 79% strong parallel efficiencies in scaling over a 128× increase of GPUs on the full-scale *El Capitan* system.

2 Performance Attributes

Performance attribute	This submission
Category of achievement	Scalability, time-to-solution, peak performance
Type of method used	Bayesian inversion, FEM, real-time computing
Results reported based on	Whole application including I/O
Precision reported	Double precision
System scale	Results measured on full-scale system
Measurement mechanism	Timers, DOF throughput, FLOP count

2025

Boosting Earth System Model Outputs And Saving PetaBytes in Their Storage Using Exascale Climate Emulators

Sameh Abdulah[1,7], Allison H. Baker[2,8], George Bosilca[3,9], Qinglei Cao[4,10], Stefano Castruccio[5,11], Marc G. Genton[1,7], David E. Keyes[1,7], Zubair Khalid[1,6,12], Hatem Ltaief[1,7], Yan Song[1,7], Georgiy L. Stenchikov[1,7], and Ying Sun[1,7]

[1]Extreme Computing & Statistics & Earth Science, King Abdullah University of Science and Technology, KSA
[2]Computational and Information Sciences Lab, NSF National Center for Atmospheric Research, USA
[3]NVIDIA, USA
[4]Department of Computer Science, Saint Louis University, USA
[5]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, USA
[6] Department of Electrical Engineering, Lahore University of Management Sciences, Pakistan
[7]{*Firstname.Lastname*}*@kaust.edu.sa* [8]*abaker@ucar.edu* [9]*gbosilca@nvidia.com*
[10]*qinglei.cao@slu.edu* [11]*scastruc@nd.edu* [12]*zubair.khalid@lums.edu.pk*

***Abstract*—**
We present the design and scalable implementation of an exascale climate emulator for addressing the escalating computational and storage requirements of high-resolution Earth System Model simulations. We utilize the spherical harmonic transform to stochastically model spatio-temporal variations in climate data. This provides tunable spatio-temporal resolution and significantly improves the fidelity and granularity of climate emulation, achieving an ultra-high spatial resolution of 0.034°(∼3.5 km) in space. Our emulator, trained on 318 billion hourly temperature data points from a 35-year and 31 billion daily data points from an 83-year global simulation ensemble, generates statistically consistent climate emulations. We extend linear solver software to mixed-precision arithmetic GPUs, applying different precisions within a single solver to adapt to different correlation strengths. The `PaRSEC` runtime system supports efficient parallel matrix operations by optimizing the dynamic balance between computation, communication, and memory requirements. Our BLAS3-rich code is optimized for systems equipped with four different families and generations of GPUs, scaling well to achieve 0.976 EFlop/s on 9,025 nodes (36,100 AMD MI250X multi-chip module (MCM) GPUs) of Frontier (nearly full system), 0.739 EFlop/s on 1,936 nodes (7,744 Grace-Hopper Superchips (GH200)) of Alps, 0.243 EFlop/s on 1,024 nodes (4,096 A100 GPUs) of Leonardo, and 0.375 EFlop/s on 3,072 nodes (18,432 V100 GPUs) of Summit.

***Index Terms*—Dynamic runtime systems, High-performance computing, Mixed-precision computation, Spatio-temporal climate emulation, Spherical harmonic transform, Task-based programming models.**

I. JUSTIFICATION FOR THE GORDON BELL PRIZE

Exascale climate emulator developed using 318 billion hourly and 31 billion daily observations for generating climate emulations at ultra-high spatial resolution (0.034° ∼3.5 km).

Authors are listed alphabetically by their last names.

SC24, November 17-22, 2024, Atlanta, GA, USA
979-8-3503-5291-7/24/$31.00 ©2024 IEEE

Modeling climate data using spherical harmonics. Mixed-precision computations. `PaRSEC` dynamic runtime system. Running on 9,025 nodes on Frontier, 1,936 nodes on Alps, 1,024 nodes on Leonardo, and 3,072 nodes on Summit, with the hybrid Flop/s rates 0.976 EFlop/s, 0.739 EFlop/s, 0.243 EFlop/s, and 0.375 EFlop/s, respectively.

II. PERFORMANCE ATTRIBUTES

Problem size	54,486,360 spatial locations across the globe at a spatial resolution of 0.034° (∼3.5 km)
Category of achievement Type of method used	Scalability and peak performance Spherical Harmonic Transform (SHT) and Cholesky factorization
Results reported on basis of Precision reported System scale	Cholesky factorization Double and mixed-precision - 0.976 EFlop/s on 9,025 nodes of Frontier (36,100 AMD MI250X multi-chip module (MCM) GPUs) equivalent to 72,200 AMD Graphics Compute Dies (GCDs) - 0.739 EFlop/s on 1,936 nodes of Alps (7,744 NVIDIA Grace-Hopper Superchips (GH200)) - 0.243 EFlop/s on 1,024 nodes of Leonardo (4,096 NVIDIA A100 GPUs) - 0.375 EFlop/s on 3,072 nodes of Summit (18,432 NVIDIA V100 GPUs)
Measurement mechanism	Timers, Flops

III. OVERVIEW OF THE PROBLEM

Climate change, evident in rising temperatures, extreme weather events, sea-level rise, and ecosystem disruption, poses significant risks and urgently requires action due to intensified heatwaves, storms, droughts, floods, and biodiversity loss [1], [2]. We stand at a critical juncture where converging

2024

ACM Gordon Bell Prize for Climate Modelling

Innovations in applying high-performance computing to climate modelling applications

Award Recipients | **Nominations** | **Committee Members**

Home > ACM Gordon Bell Prize For Climate Modelling

About ACM Gordon Bell Prize for Climate Modelling

The Gordon Bell Prize for Climate Modelling will be awarded every year for ten years beginning in 2023 to recognize the contributions of climate scientists and software engineers. Nominations will be selected based on their impact and potential impact on the field of climate modelling, on related fields, and on wider society by applying high-performance computing to climate modelling applications. The award aims to recognize innovative parallel computing contributions toward solving the global climate crisis. Nominations will be selected based on the performance and innovation in their computational methods and their contributions toward improving climate modelling and our understanding of the Earth's climate system. Financial support for this $10,000 award is provided by Gordon Bell, a pioneer in high-performance and parallel computing.

(1934-2024)
Founding Director, US NSF Office of Computer and Information Sci & Eng

Motivation (1): data is no longer small

"Increasing amounts of data are being produced (e.g., by remote sensing instruments and numerical models), while techniques to handle millions of observations have historically lagged behind… Computational implementations that work with irregularly-spaced observations are still rare." - Dorit Hammerling, NCAR, July 2019

1M × 1M dense sym DP matrix requires 4 TB, $N^3 \sim 10^{18}$ Flops

Traditional approaches:

- *Global* low rank
- *Neglected (zeroed)* outer diagonals

Better HPC approaches:

- *Hierarchical* low rank
- *Reduced precision* outer diagonals

Motivation (2): energy cost is no longer small

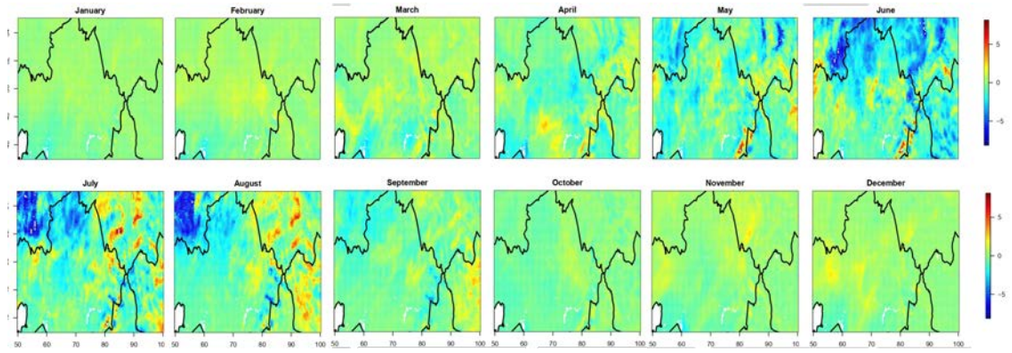
“Computational efficiency through tuned approximation”

- Solve with lowest possible energy (∫ power·dt) , not necessarily highest performance (in operations per sec)
- Satisfy application-worthy accuracy
- Squeeze out “easyflop/s” rather than racking them up

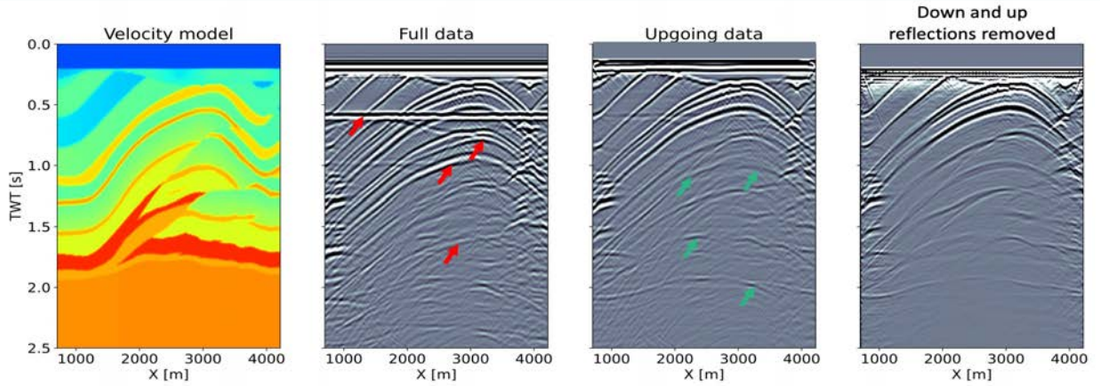
In 2024, performance “caught up” w/ efficiency (2 trends)

- Applications: increasing % of work tolerates low precision
- Architecture: going low in precision pays more than ever – “starring” FP8 and INT8 on Hopper

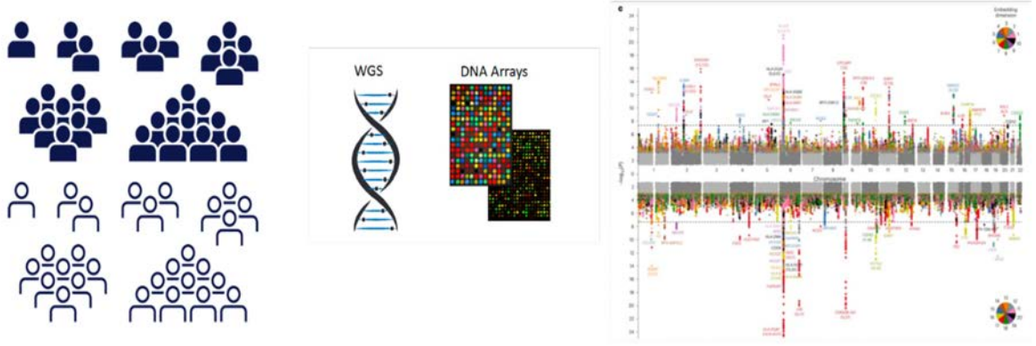
Our journey actually covers *four* Gordon Bell Prize finalist papers in the past three years



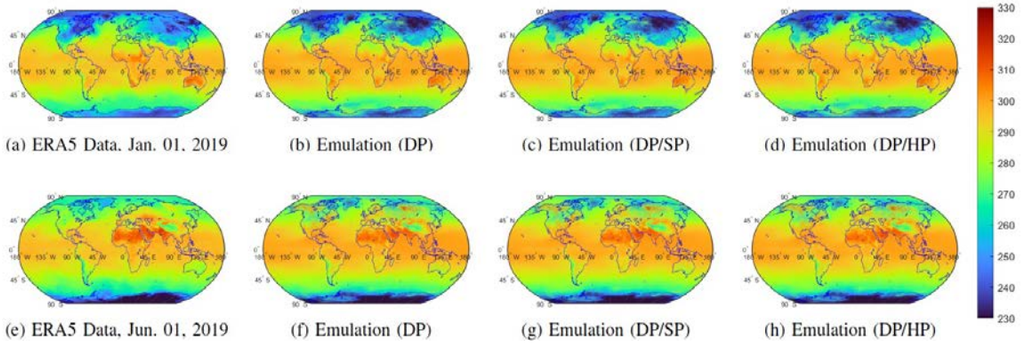
2022: Geospatial statistics



2023: Seismic processing

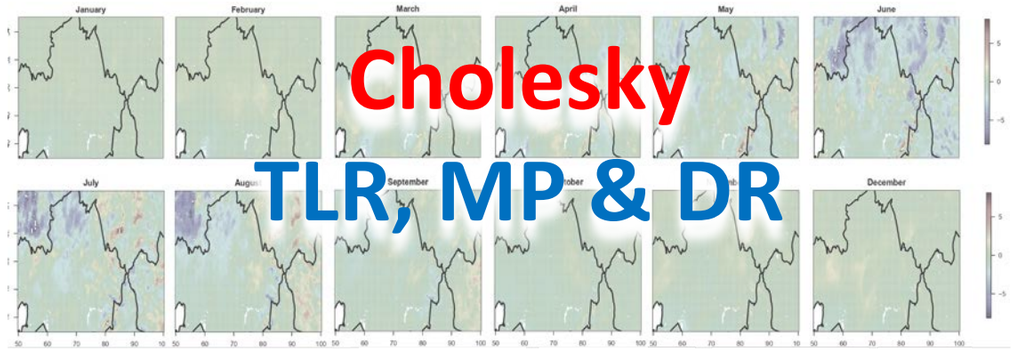


2024: Genomic association

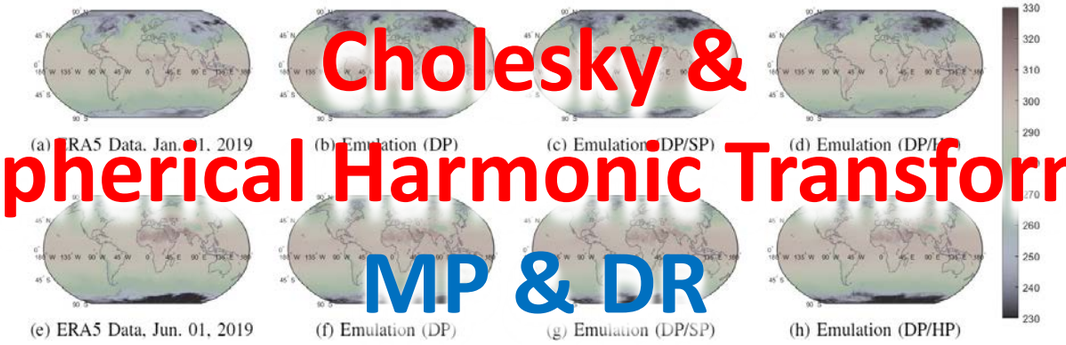


2024: Climate emulation

Algorithm: adaptive low rank and low precision substitutions for (default) dense double



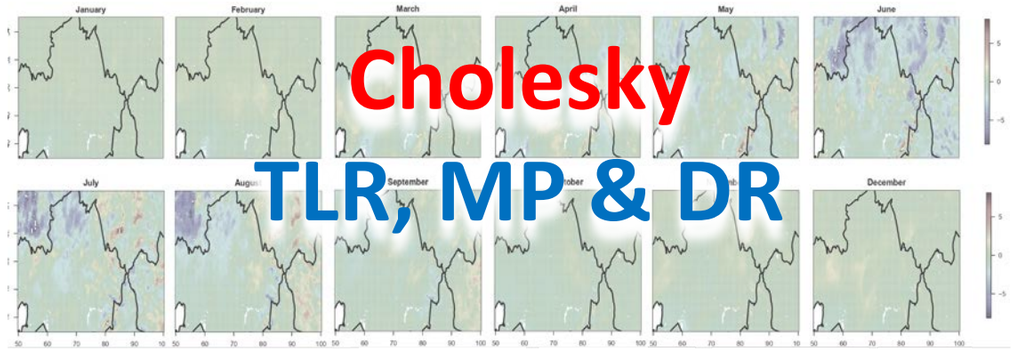
2022: Geospatial statistics



2024: Climate emulation

TLR = tile low rank, **MP** = mixed precision, **DR** = dynamic runtime system

Gordon Bell finalist “merits” and machines



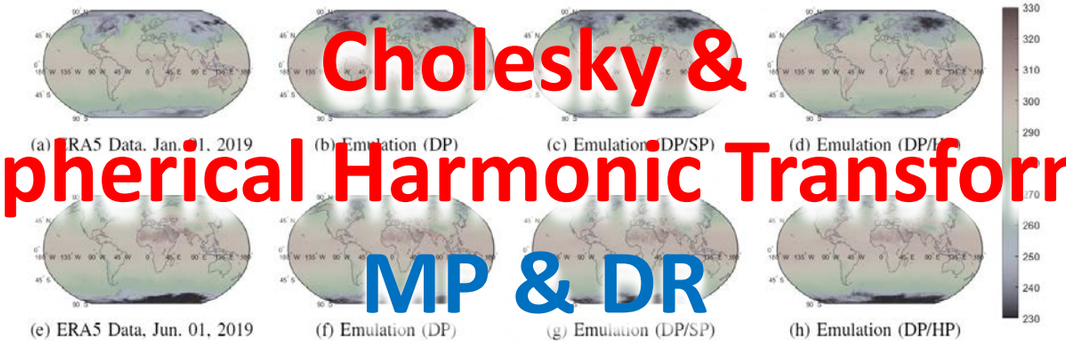
Cholesky

TLR, MP & DR

2022: Geospatial statistics

Time to solution @ Fugaku

ExaFlops @ Alps



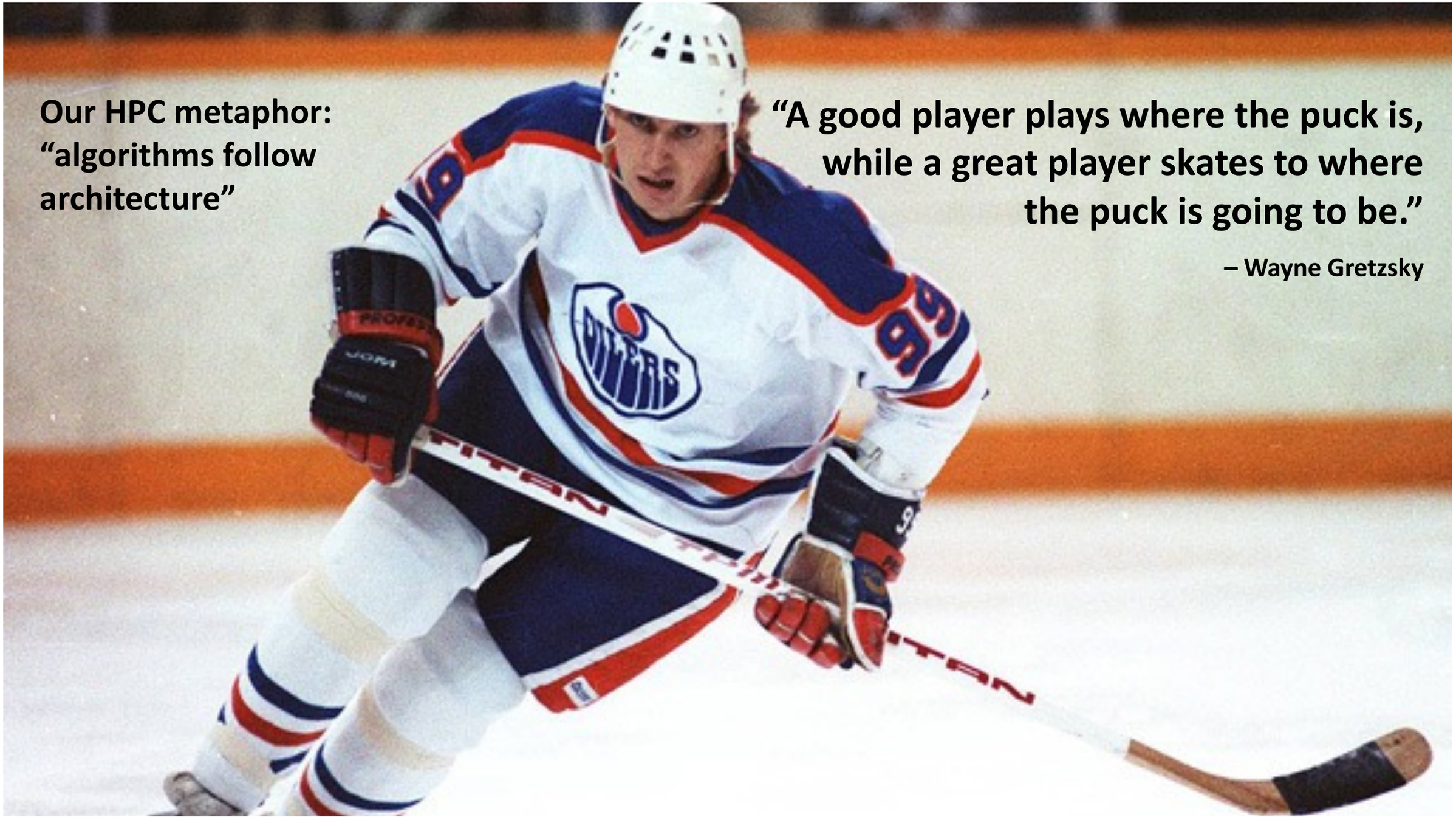
Cholesky &

Spherical Harmonic Transform

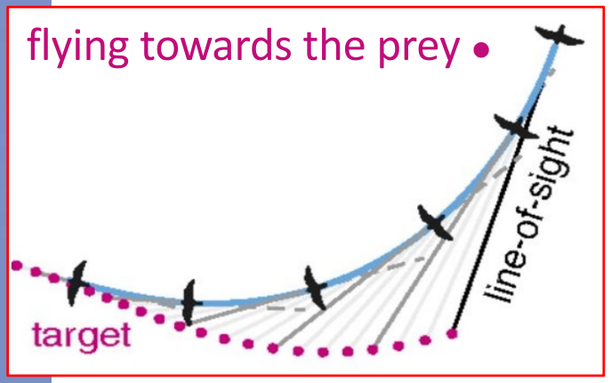
MP & DR

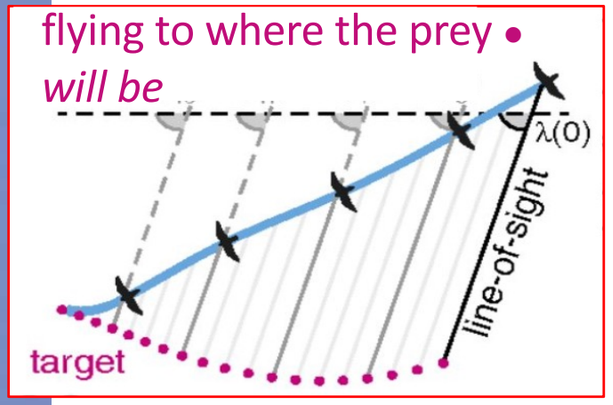
2024: Climate emulation

TLR = tile low rank, **MP** = mixed precision, **DR** = dynamic runtime system


Our HPC metaphor:
“algorithms follow architecture”
“A good player plays where the puck is, while a great player skates to where the puck is going to be.”
– Wayne Gretzsky

“I fly not to where the prey is, but to where it will be.”



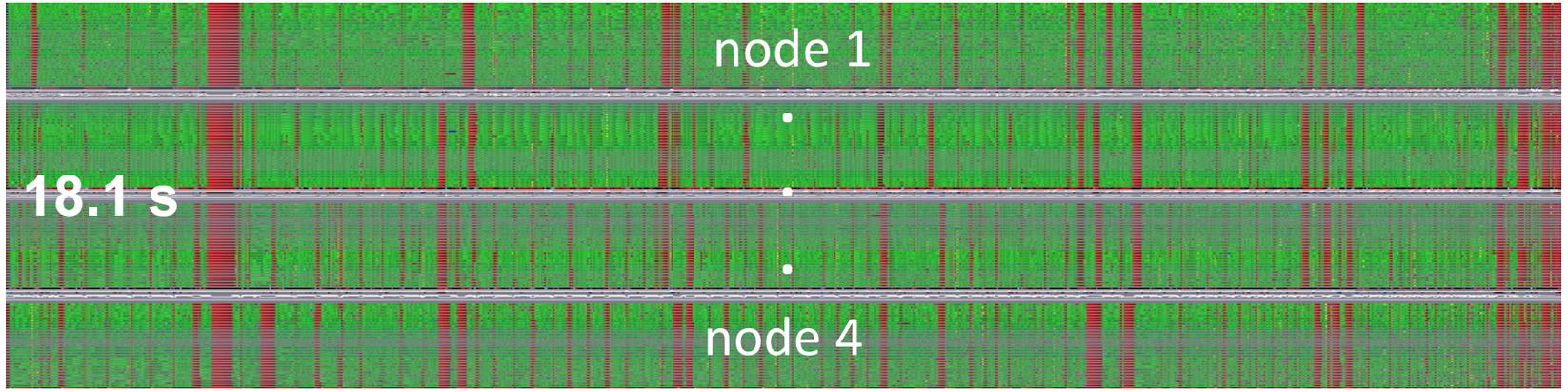


C. H. Brighton, et al.,
PNAS (2017)

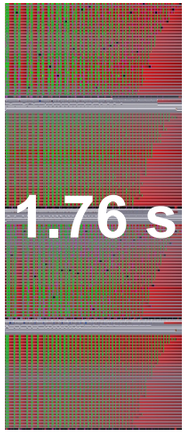
Our journey in tuned approximation began in 2018 with these time traces for tile low-rank (TLR) Cholesky

… for factorization of a dense 54K covariance matrix on four 32-core nodes of a Cray XC-40

Dense Tile-based Cholesky factorization (Chameleon)



Tile low rank (TLR) Cholesky factorization (HiCMA)



- TLR may score a lower percentage of peak (after squeezing out flops)
- TLR may have poorer load balance (a higher percentage of idle time (red) vs. computation (green))
- TLR may scale less efficiently (less able to cover data motion with computation)
- TLR is, however, **10X superior in time** for required application accuracy, at **about 65% of average power** compared to dense

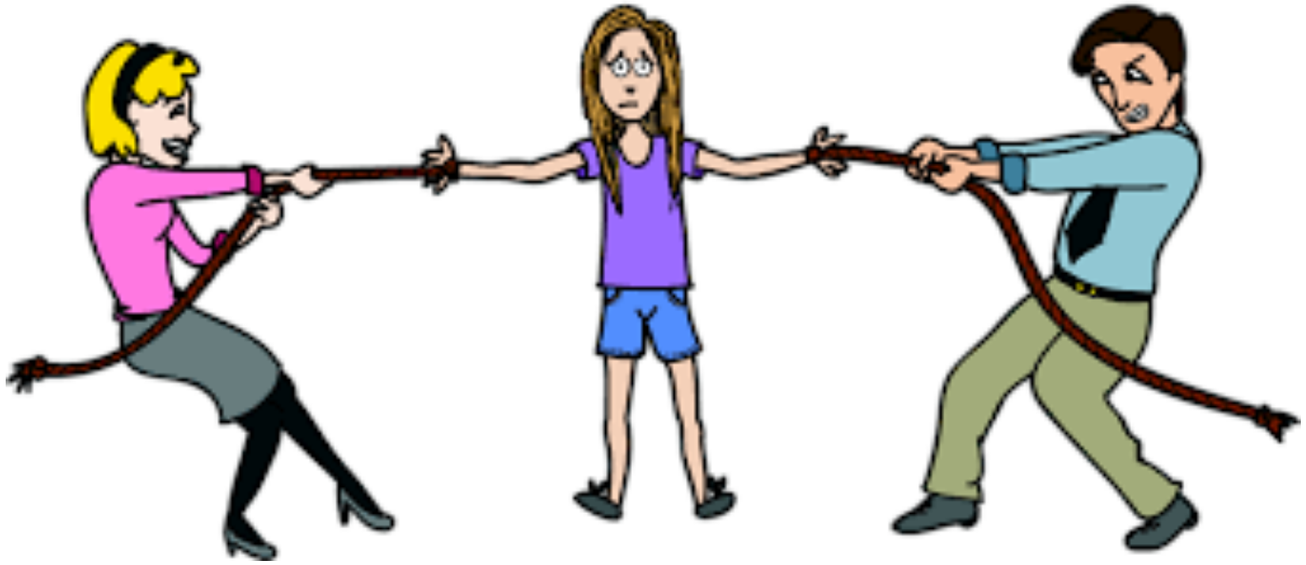
Akbudak, Ltaief, Mikhalev, Charara & K., *Exploiting Data Sparsity for Large-scale Matrix Computations,* Euro-Par 2018

Algorithmic philosophy

Algorithms must span a widening gulf …

adaptive algorithms

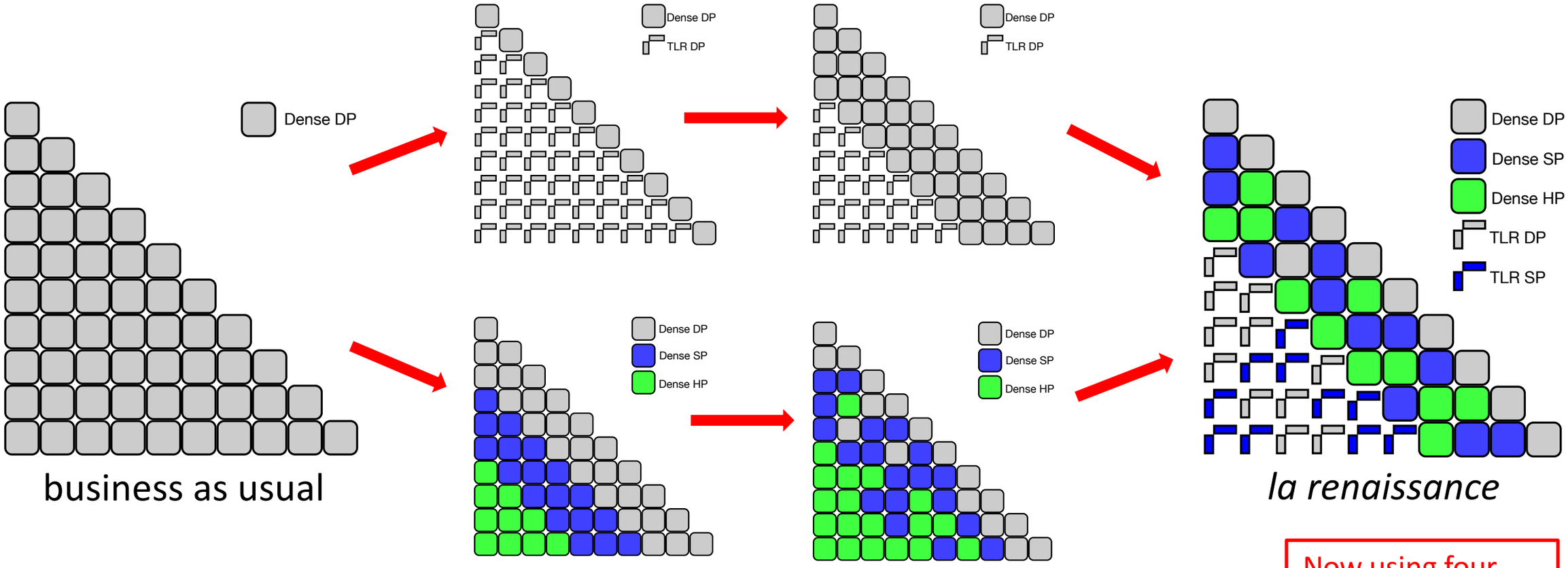
ambitious applications
increasingly dynamic and unstructured



austere architectures
increasingly optimized for uniformity

… a full employment program for algorithm developers

Computational efficiency through *tuned approximation*: a journey with *tile low rank* and *mixed precision*



business as usual

la renaissance

Don’t oversolve: maintain just enough accuracy for the application purpose

Economize on storage: no extra copies of the original matrix

Now using four precisions: FP64, FP32, FP16 & FP8

Linear algebraic “secret sauce”

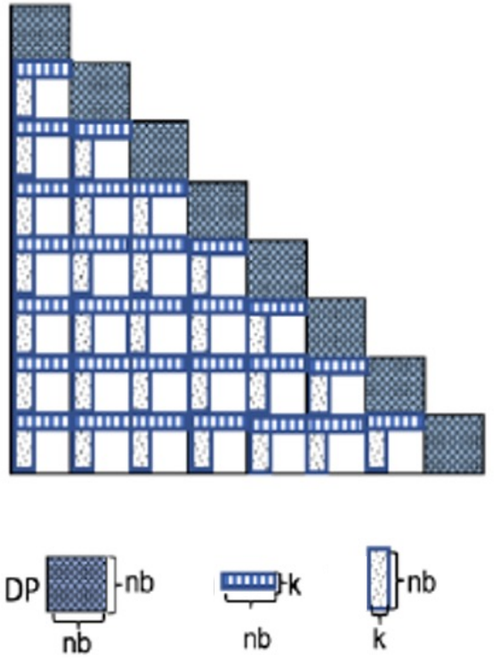
Where possible, without losing working accuracy:

- Replace default 64-bit IEEE standard double precision operatio with lower precisions
  - Save storage
  - Save data motion
  - Exploit special-purpose hardware optimized for low precision
- Replace default full rank blocks of discrete *linear operators* and/or discrete *field data* with lower rank blocks
  - Compress homomorphically (block-by-block, without decompression)
  - Save storage
  - Save data motion
  - Exploit special-purpose hardware optimized for BLAS3

Renaissance in numerical linear algebra (1): rank

Many formally dense matrices arising from

- **covariances** in statistics
- **Hessians** from PDE-constrained optimization (RSQP)
- **integral equations** with smooth Green's functions
- **Schur complements** within discretizations of PDEs
- **nonlocal operators** from fractional differential equations
- **radial basis functions** from unstructured meshing
- **kernel matrices** from GWAS & machine learning applications



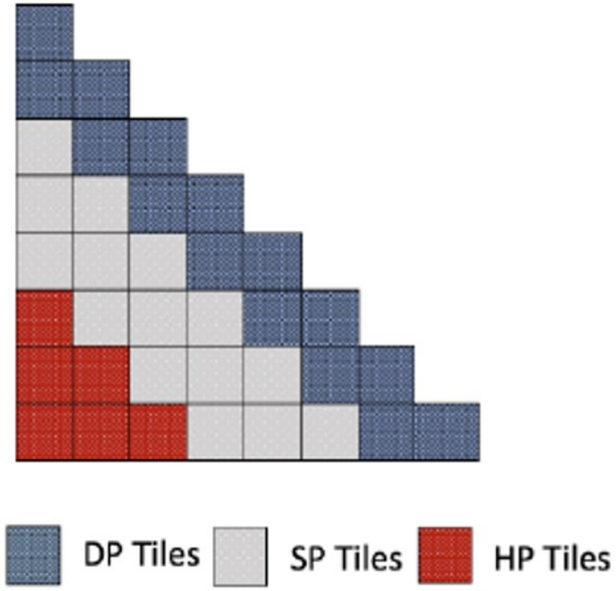
have exploitable low-rank structure in "most" their off-diagonal blocks (if well ordered, e.g., for $d > 1$ by Hilbert)

Renaissance in numerical linear algebra (2): precision

Many matrices arising in applications have blocks of **relatively small norm** and can be replaced with **reduced precision.**

Mixed precision algorithms have a long history, e.g., iterative refinement (1963, Wilkinson), where multiple copies of the matrix are kept in different precisions for different purposes.

There are many such new algorithms; see Higham & Mary, *Mixed precision algorithms in numerical linear algebra*, Acta Numerica (2022), Carson's EU Horizon project *inEXASCALE* (2023-)



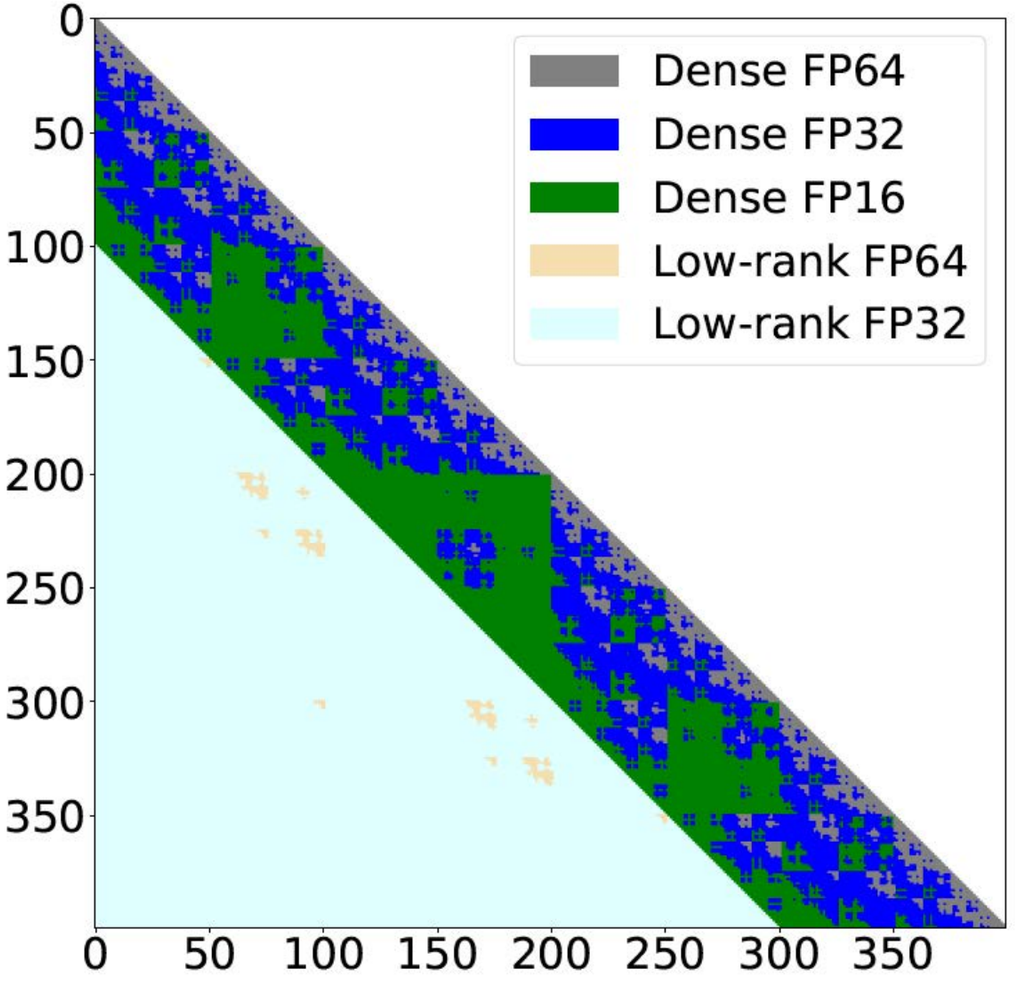
Renaissance in numerical linear algebra (3): combined

Moreover, these ideas can be combined, as in this 1M x 1M dense symmetric covariance matrix:

- Original in DP: 4 TB
- Replacement: 0.915 TB

Smaller workingsets mean larger problems fit in GPUs and last-level caches on CPUs, for data movement savings

- Also, net computational savings
- Data structures and programs are more complex



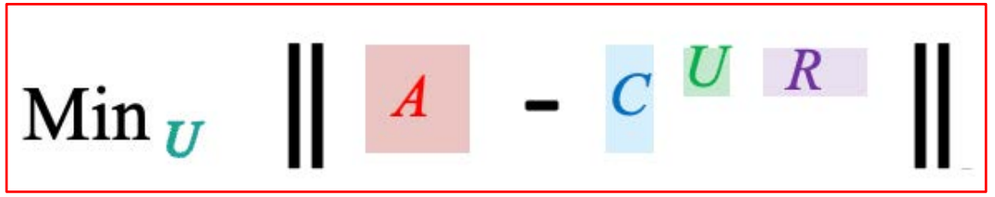
Block rank: a tuning knob

- Replace dense blocks with reduced rank representations, whether “born dense” or as arising during matrix operations
  - use high accuracy (high rank) to build “exact” solvers
  - use low accuracy (low rank) to build preconditioners
- Consider hardware parameters in tuning block sizes and maximum rank parameters, to complement mathematical considerations
  - e.g., cache sizes, warp sizes
- Select from already broad and ever broadening algorithmic menu to form low-rank blocks (next slide)
  - traditionally a flop-intensive vendor-optimized GEMM-based flat algorithm
- Implement in “batches” of leaf blocks
  - flattening trees in the case of hierarchical methods

Low-rank approximations for compressible tiles

Options for forming data sparse representations of the amenable off-diagonal blocks

- *standard SVD*: $O(n^3)$, too expensive, especially for repeated compressions after additive tile manipulations
- *randomized SVD* (Halko *et al*., 2011): $O(n^2 \log k)$ for rank k, requires only a small number of passes over the data, saving over the SVD in memory accesses as well as operations
- *adaptive cross approximation (ACA)* (Bebendorf, 2000): $O(k^2 n \log n)$, motivated by integral equation kernels
- *matrix skeletonization* (representing a matrix by a representative collection of row and columns), such as *CUR*, *sketching*, or *interpolatory decomposition*



Block precision: another tuning knob

- Consider 2-precision case, with machine epsilons (unit roundoffs) u_{high} and u_{low}, resp.
- Let $\| A \|_F$ be the Frobenius norm of the global matrix square matrix A, which is computable by streaming A through just once
- Let n_T be the number of tiles in each dimension of A
- Then any tile A_{ij} such that $n_T \| A_{ij} \|_F / \| A \|_F < u_{high} / u_{low}$ is stored in low precision; otherwise kept in high
- The mixed precision tiled matrix $\mathcal{A}$ thus formed satisfies

$$\| \mathcal{A} - A \|_F < u_{high} \| A \|_F$$

- Generalizes to multiple precisions
- Tiles can be converted dynamically at runtime

Higham & Mary, *Mixed Precision Algorithms in Numerical Linear Algebra* (2022), Acta Numerica, pp. 347-414

Example: covariance matrices from spatial statistics

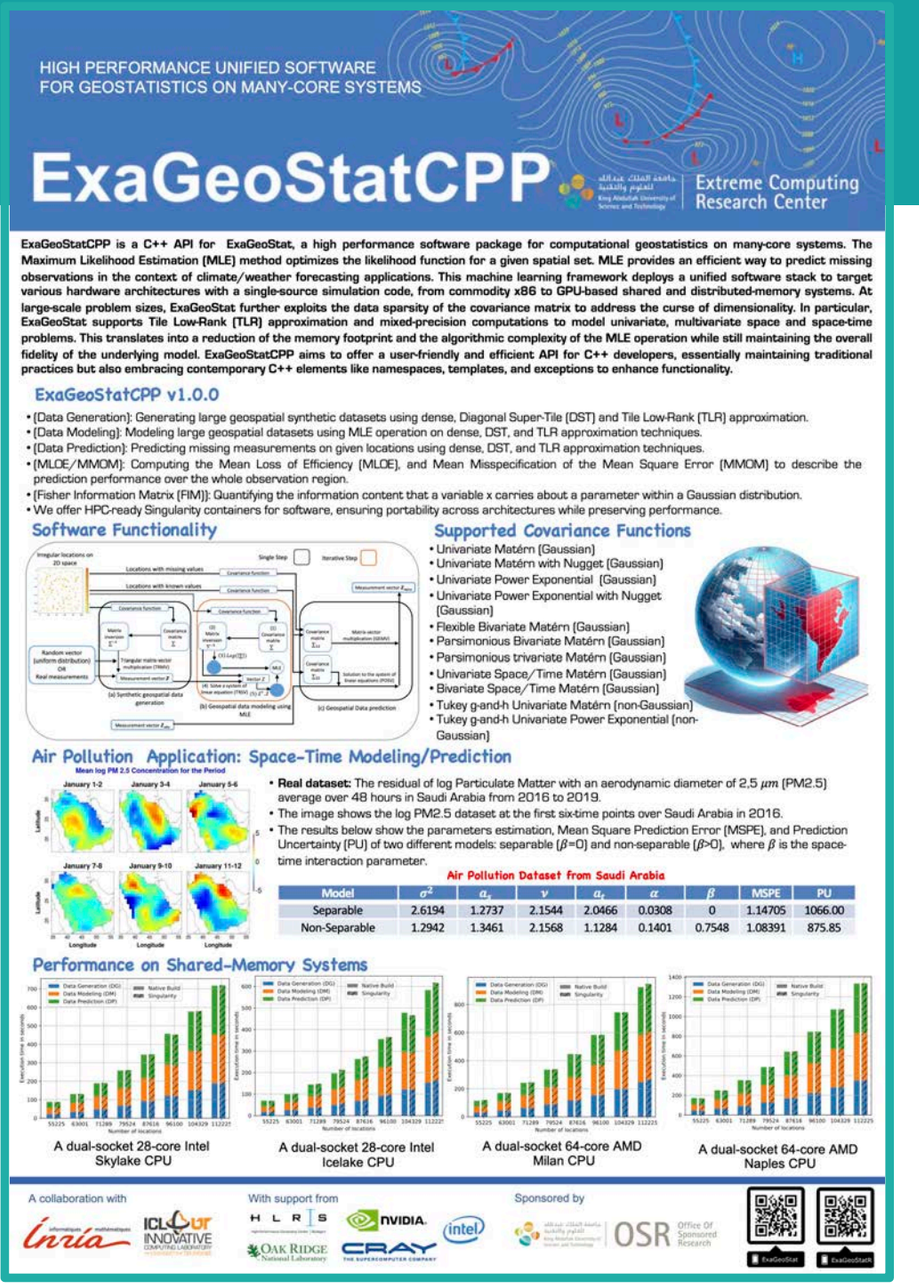
- Climate and weather applications have many measurements located regularly or irregularly in a region; prediction is needed at other locations
- Modeled as realization of Gaussian or Matérn spatial random field, with parameters to be fit
- Leads to evaluating, inside an optimization loop, the log-likelihood function involving a large dense (but data sparse) covariance matrix $\boldsymbol{\Sigma}$

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

- Apply inverse $\boldsymbol{\Sigma}^{-1}$ and determinant $|\boldsymbol{\Sigma}|$ with Cholesky

HPSC software

- Synthetic Dataset Generator
  - Generates large-scale geospatial datasets which can be used separately as benchmark datasets for other software packages
- Maximum Likelihood Estimator (MLE)
  - Evaluates the maximum likelihood function on large-scale geospatial datasets
  - Supports dense full machine precision, Tile Low-Rank (TLR) approximation, low-precision approximation accuracy, and now TLR-MP
- ExaGeoStat Predictor
  - Infers unknown measurements at new geospatial locations from the MLE model (kriging)



Covariance functions $\Sigma(\theta)$ supported in ExaGeoStat

Handful of parameters *with* physics, as opposed to trillions *without* physics ☺

Univariate Matern Kernel

$$C(r;\boldsymbol{\theta}) = \frac{\theta_1}{2^{\theta_3-1}\Gamma(\theta_3)}\left(\frac{r}{\theta_2}\right)^{\theta_3}\mathcal{K}_{\theta_3}\left(\frac{r}{\theta_2}\right)$$

(3 parameters to fit: variance, range, smoothness)

Space/Time Nonseparable Kernel

$$C(\mathbf{h},u) = \frac{\sigma^2}{a_t|u|^{2\alpha}+1}\mathcal{M}_\nu\left\{\frac{\|\mathbf{h}\|/a_s}{(a_t|u|^{2\alpha}+1)^{\beta/2}}\right\},$$

(6 parameters to fit, add: time-range, time-smoothness, and separability)

Multivariate Parsimonious Kernel

$$C_{ij}(\|\mathbf{h}\|;\boldsymbol{\theta}) = \frac{\rho_{ij}\sigma_{ii}\sigma_{jj}}{2^{\nu_{ij}-1}\Gamma(\nu_{ij})}\left(\frac{\|\mathbf{h}\|}{a}\right)^{\nu_{ij}}\mathcal{K}_{\nu_{ij}}\left(\frac{\|\mathbf{h}\|}{a}\right)$$

Tukey g-and-h Non-Gaussian Field with Kernel

$$\rho_Z(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(4\sqrt{2\nu}\frac{h}{\phi}\right)^{\nu}\mathcal{K}_\nu\left(4\sqrt{2\nu}\frac{h}{\phi}\right)$$

Multivariate Flexible Kernel

$$C(\mathbf{h};u) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)(a|u|^{2\alpha}+1)^{\delta+\beta d/2}}\left(\frac{c\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\beta/2}}\right)^{\nu}$$
$$\times K_\nu\left(\frac{c\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\beta/2}}\right), \quad (\mathbf{h};u)\in\mathbb{R}^d\times\mathbb{R},$$

Powered Exponential Kernel

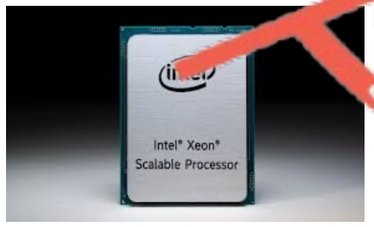
$$C(r;\boldsymbol{\theta}) = \theta_0\exp\left(\frac{-r^{\theta_2}}{\theta_1}\right)$$

The portable ExaGeoStat software stack

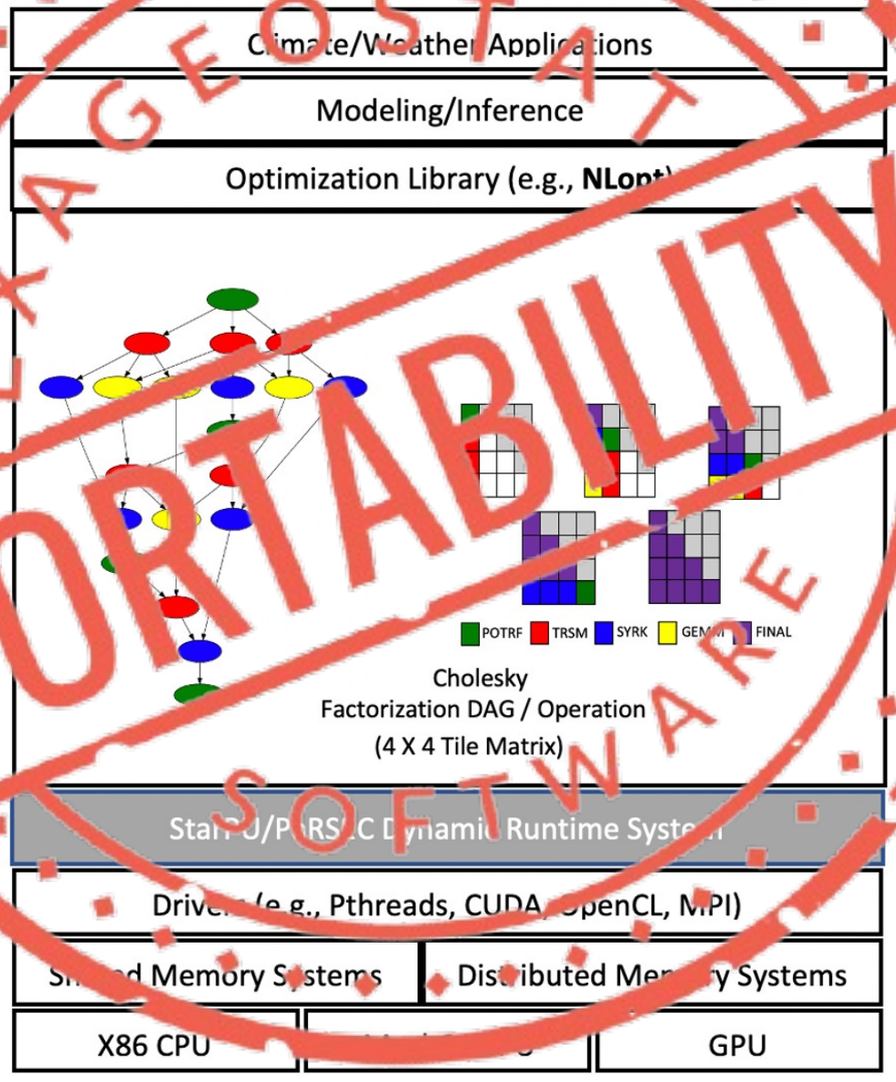
Fujitsu A64FX

NVIDIA V100

AMD EPYC



Intel X86



Fugaku

Summit

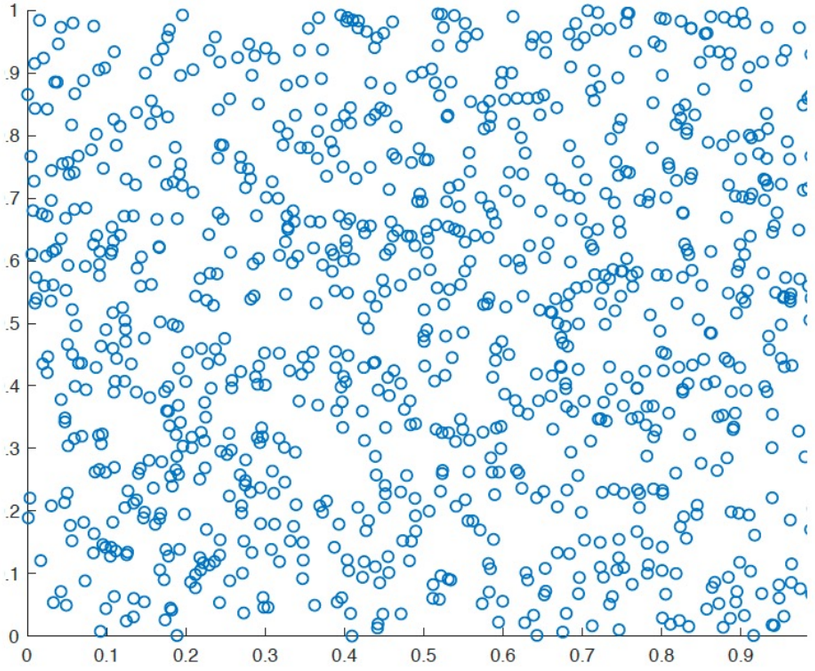
HAWK

Shaheen-2

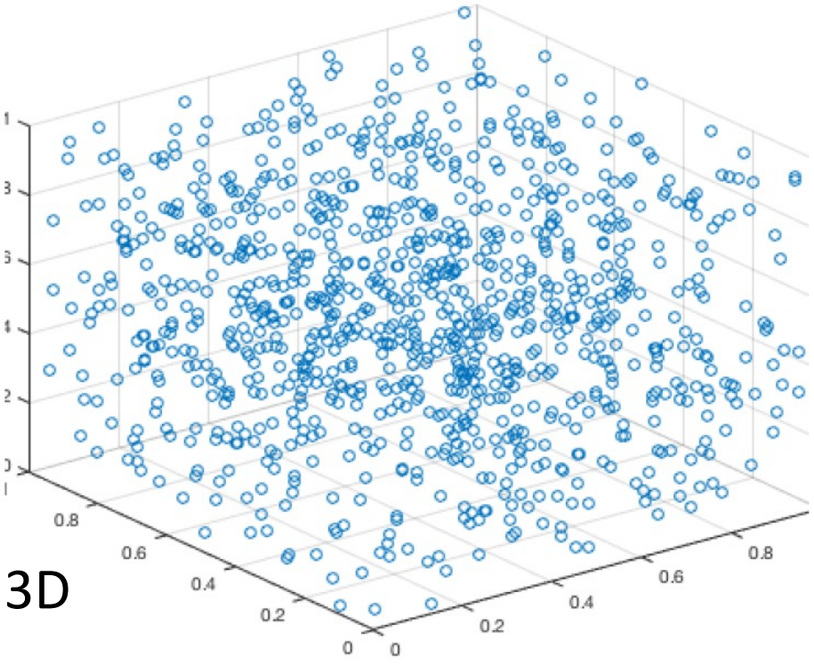
Synthetic scaling test

Random coordinate generation within the unit square or unit cube with Matérn kernel decay, each pair of points connected by square exponential decay, $a_{ij} \sim \exp(-c|x_i - x_j|^2)$

2D



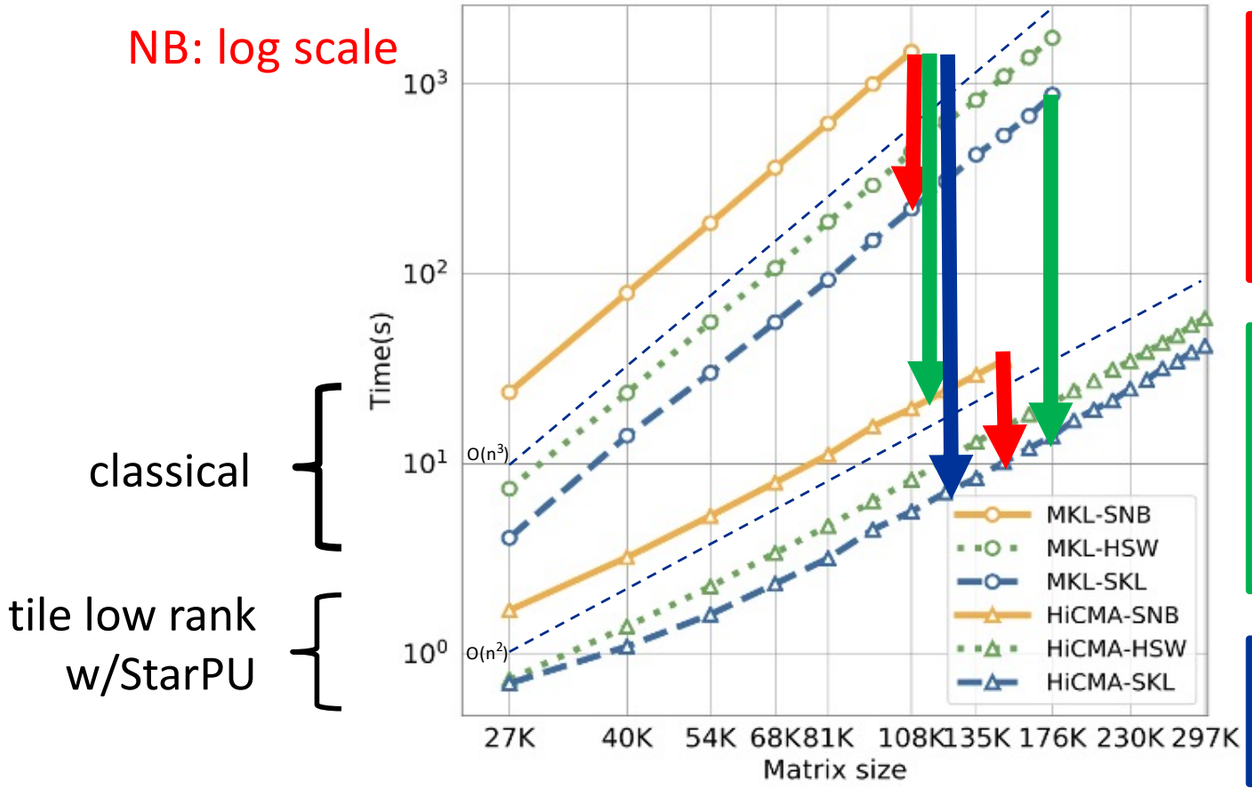
3D



HiCMA TLR vs. Intel MKL on shared memory

- Gaussian kernel to accuracy 1.0e-8 in each tile
- Three generations of Intel manycore (Sandy Bridge, Haswell, Skylake)
- Two generations of linear algebra (classical dense and tile low rank)

NB: log scale



Red arrows: speedups from hardware, same algorithm

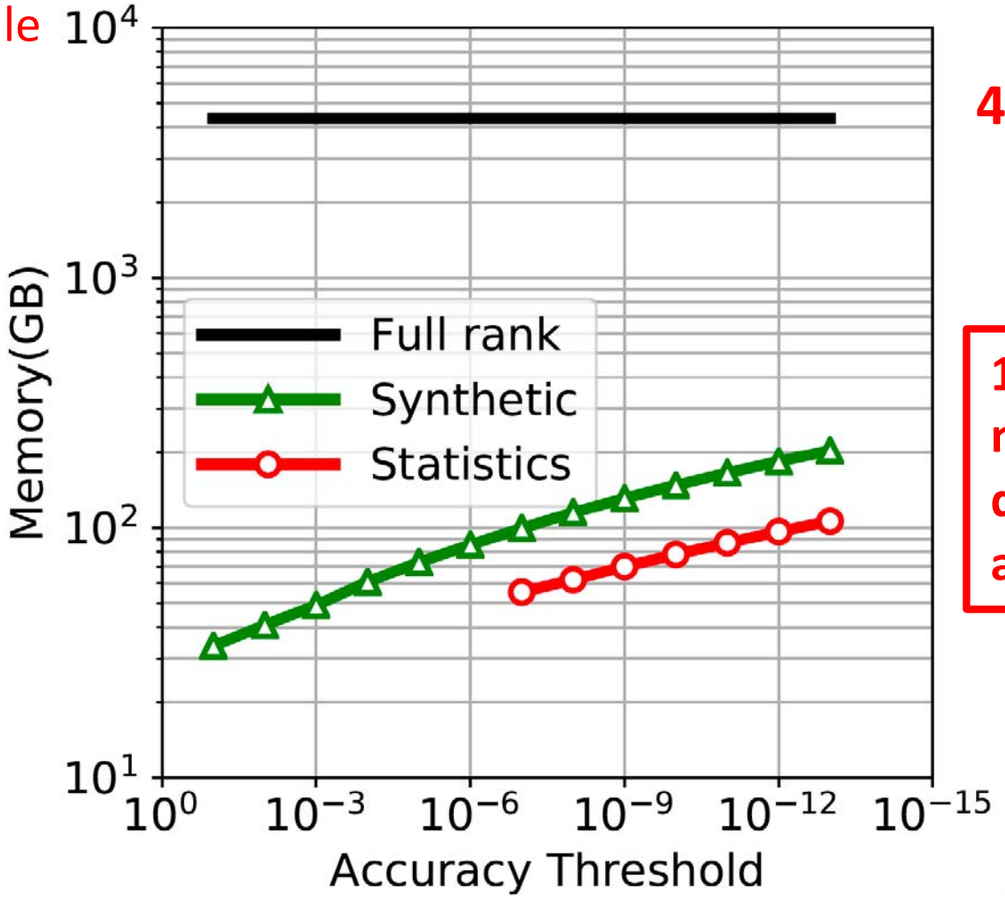
Green arrows: speedups from algorithm, same hardware

Blue arrow: from both

Akbudak, Ltaief, Mikhalev, Charara & K., *Exploiting Data Sparsity for Large-scale Matrix Computations,* Euro-Par 2018

Memory footprint for TLR fully DP matrix of size 1M

NB: log scale



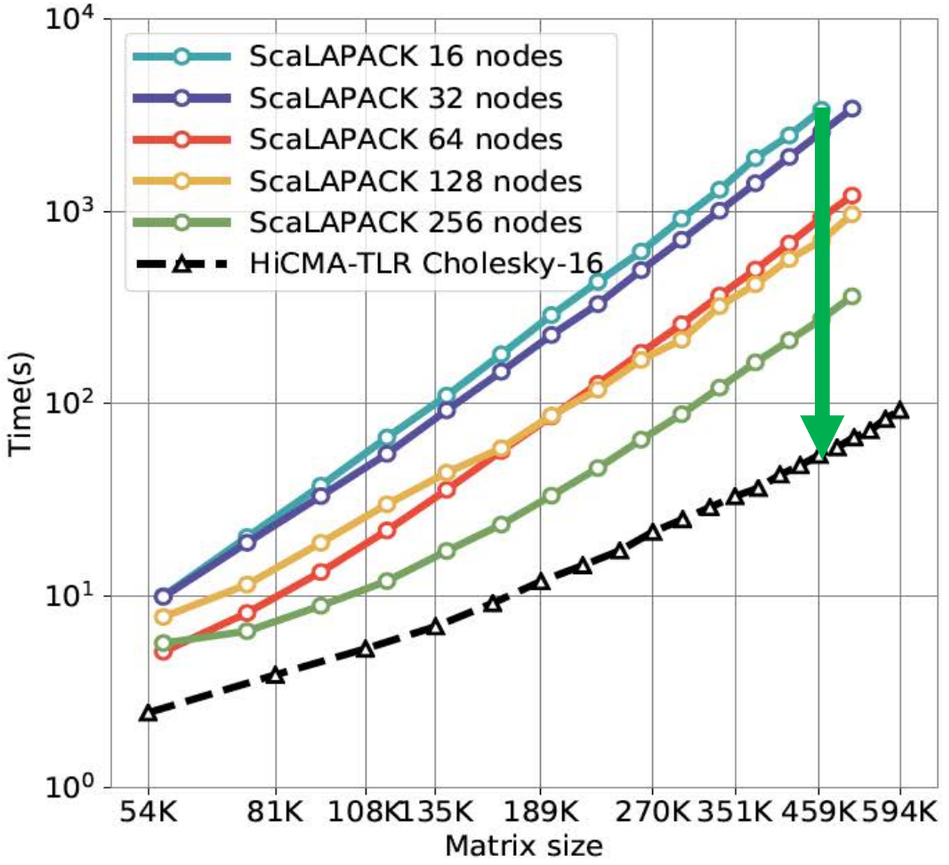
4 TB

1 to 2 orders of magnitude less, depending upon accuracy (x-axis)

Akbudak, Ltaief, Mikhalev, Charara & K., *Exploiting Data Sparsity for Large-scale Matrix Computations*, EuroPar 2018

HiCMA TLR vs. ScaLAPACK on distributed memory

NB: log scale



Green arrow: speedup from algorithm, same 16 nodes

Shaheen II at KAUST: a Cray XC40 system with 6,174 compute nodes, each of which has two 16-core Intel Haswell CPUs running at 2.30 GHz and 128 GB of DDR4 main memory

Akbudak, Ltaief, Mikhalev, Charara & K., *Exploiting Data Sparsity for Large-scale Matrix Computations,* Euro-Par 2018

Two motivations for mixed precision

- Mathematical: (much) better than “no precision”
  - statisticians often approximate remote diagonals as *zero* after performing a diagonally clustered space-filling curve ordering (no error bounds available)
- Computational: faster time to solution
  - hence lower energy consumption and higher performance

Peak Performance in TF/s	V100 NVLink	A100 NVLink	H100 SXM	B200
FP64	7.5	9.7	34	90
FP32		19.5	6	0
FP64 Tensor Core	15	19.5	67	40
FP/TF32 Tensor Core	8x	156 16x	495 16x	1125 125x
FP16 Tensor Core	120	312	989	2250
	rel. 2017	rel. 2020	rel. 2023	rel.2025

Two motivations for mixed precision

- Mathematical: (much) better than “no precision”
  - statisticians often approximate remote diagonals as *zero* after performing a diagonally clustered space-filling curve ordering (no error bounds available)
- Computational: faster time to solution
  - hence lower energy consumption and higher performance

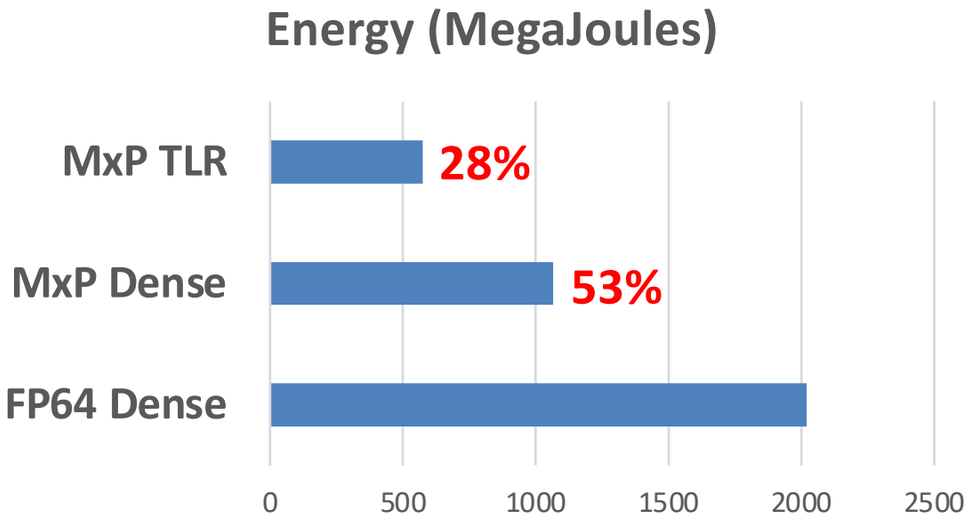
Peak Performance in TF/s	V100 NVLink	A100 NVLink	H100 SXM	B200
FP64	7.5	9.7	34	90
FP32		19.5	67	180
FP64 Tensor Core	15	19.5	67	40
FP/TF32 Tensor Core		156	495	1125
FP16 Tensor Core	120	312	990	2250
FP8/INT8 Tensor Core	-	624	1980	4500
FP4 Tensor Core	-	-	-	9000

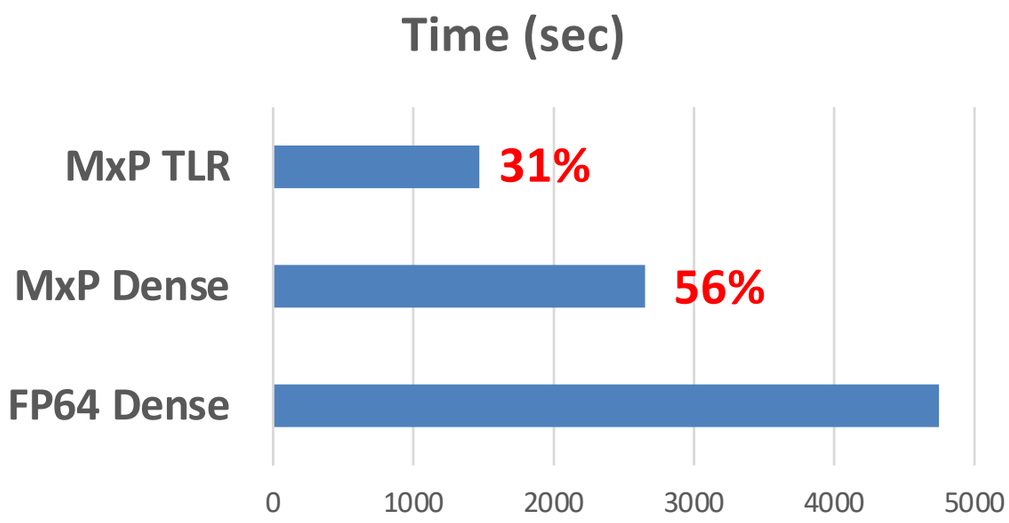
8x 32x 30x 225x

Energy and time savings

- Matérn 2D space kernel, matrix size 3.24M
- Solved to comparable accuracy by 3 algorithms
  - FP64 dense
  - adaptive mixed precision dense
  - adaptive mixed precision tile low rank

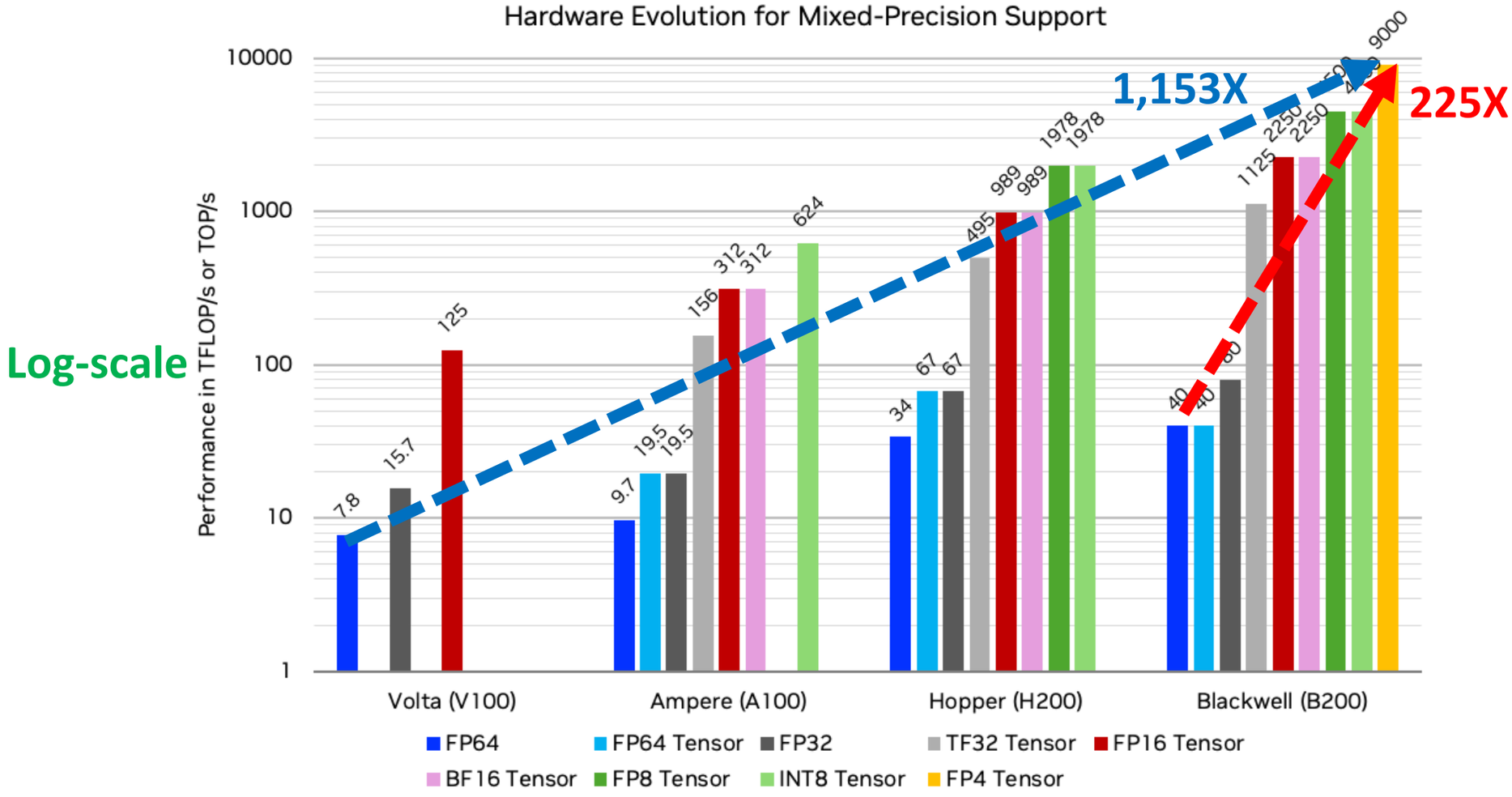
On 512 nodes of Shaheen III





Abdulah *et al.*, *Sustainably Modeling a Sustainable Future Climate*, Abu Dhabi Investment Authority, 2025

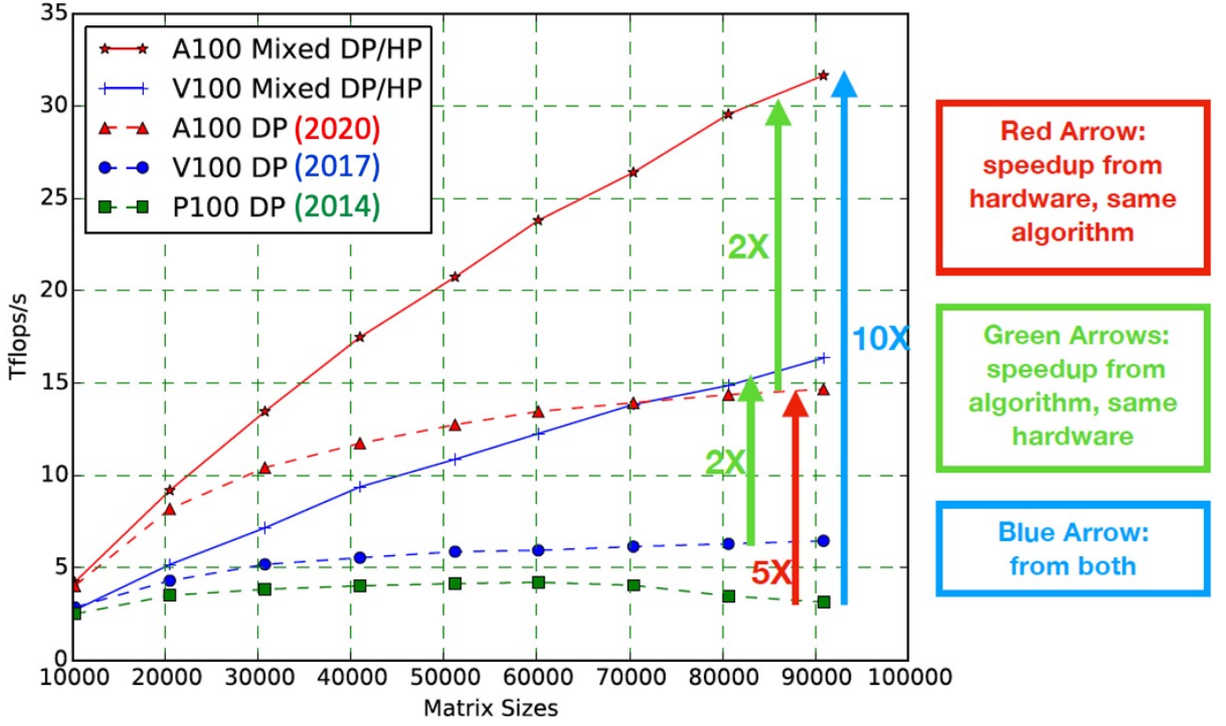
Peak performance of four generations of NVIDIA GPUs



c/o H. Bayraktar, NVIDIA

Mixed precision geospatial statistics on GPUs

- Gaussian kernel to accuracy 1.0e-9 in each tile
- Three generations of NVIDIA GPU (Pascal, Volta, Ampere)
- Two generations of linear algebra (double precision and mixed DP/HP)



Ltaief, Genton, Gratadour, K. & Ravasi, 2022, *Responsibly Reckless Matrix Algorithms for HPC Scientific Applications*, Computing in Science and Engineering

2022 Gordon Bell (regular)

II. PERFORMANCE ATTRIBUTES

Performance Attributes	Our submission
Problem Size	Nine million geospatial locations[1]
Category of achievement	Time-to-solution and scalability
Type of method used	Maximum Likelihood Estimation (MLE)
Results reported on basis of	Whole application
Precision reported	Double, single, and half precision
System scale	16K Fujitsu A64FX nodes of Fugaku[1]
Measurement mechanism	Timers; FLOPS; Performance modeling

Reshaping Geostatistical Modeling and Prediction for Extreme-Scale Environmental Applications

Qinglei Cao[2,6], Sameh Abdulah[1,5], Rabab Alomairy[1,5], Yu Pei[2,6], Pratik Nag[1,5], George Bosilca[2,7], Jack Dongarra[2,3,4,7], Marc G. Genton[1,5], David E. Keyes[1,5], Hatem Ltaief[1,5], and Ying Sun[1,5]

[1]Extreme Computing Research Center,
Division of Computer, Electrical and Mathematical Sciences and Engineering,
King Abdullah University of Science and Technology, Thuwal, KSA
[2]Innovative Computing Laboratory, University of Tennessee, Knoxville, TN USA
[3]The Oak Ridge National Laboratory, Oak Ridge, TN USA
[4]University of Manchester, Manchester, UK
[5]{*Firstname.Lastname*}*@kaust.edu.sa*
[6]{*qcao3, ypei2*}*@vols.utk.edu*
[7]{*bosilca, dongarra*}*@icl.utk.edu*

***Abstract*— We extend the capability of space-time geostatistical modeling using algebraic approximations, illustrating application-expected accuracy worthy of double precision from majority low-precision computations and low-rank matrix approximations. We exploit the mathematical structure of the dense covariance matrix whose inverse action and determinant are repeatedly required in Gaussian log-likelihood optimization. Geostatistics augments first-principles modeling approaches for the prediction of environmental phenomena given the availability of measurements at a large number of locations; however, traditional Cholesky-based approaches grow cubically in complexity, gating practical extension to continental and global datasets now available. We combine the linear algebraic contributions of mixed-precision and low-rank computations within a tile-based Cholesky solver with on-demand casting of precisions and dynamic runtime support from `PaRSEC` to orchestrate tasks and data movement. Our adaptive approach scales on various systems and leverages the Fujitsu A64FX nodes of `Fugaku` to achieve up to 12X performance speedup against the highly optimized dense Cholesky implementation.**

***Index Terms*—Space-Time Geospatial Statistics, Climate/Weather Prediction, Task-Based Programming Models, Dynamic Runtime Systems, Mixed-Precision Computations, Low-Rank Matrix Approximations, High Performance Computing.**

I. JUSTIFICATION FOR THE GORDON BELL PRIZE

Synergistic combination of mixed-precision computations and low-rank matrix approximations. Dynamic task-based runtime system and data movement. Scalability on 48,384 `Fugaku` nodes (2,322,432 cores) for maximum log-likelihood estimation (MLE). Performance speedup up to 12X over FP64 execution while attaining application-worthy accuracy. Incorporation into path-finding software framework for geostatistical applications.

II. PERFORMANCE ATTRIBUTES

Performance Attributes	Our submission
Problem Size	Ten million geospatial locations
Category of achievement	Time-to-solution and scalability
Type of method used	Maximum Likelihood Estimation (MLE)
Results reported on basis of	Whole application
Precision reported	Double, single, and half precision
System scale	48,384 Fujitsu A64FX nodes of `Fugaku`
Measurement mechanism	Timers; Flops

III. OVERVIEW OF THE PROBLEM

Geostatistics is a means of modeling and predicting desired quantities directly from data. It is based on statistical assumptions and optimization of parameters and is often referred to as emulation, in contrast to simulation. It is complementary to first-principles modeling approaches rooted in conservation laws and typically expressed in PDEs. It may draw upon data from simulations and/or from observations. Alternative statistical approaches to predictions from first-principles methods, such as Monte Carlo sampling wrapped around simulations with a distribution of inputs, may be vastly more computationally expensive than sampling from an assumed parameterized distribution based on a much smaller number of simulations. Geostatistics is relied upon for economic and policy decisions for which billions of dollars or even lives are at stake, such as engineering safety margins into developments, mitigating hazardous air quality, siting fixed renewable energy resources, and estimating weather-dependent tourism demands. We consider herein evapotranspiration, important to agricultural irrigation and water resource management, as seen in Figure 1. Climate and weather predictions are among the principal workloads occupying supercomputers around the world and even minor improvements for regular production applications pay large dividends. A wide variety of such predictive codes have

GB’22 collaborators

KAUST Supercomputing Core Lab, HLRS-Stuttgart, Oak Ridge LCF, RIKEN, and:


Qinglei Cao


Yu Pei


George Boslica


Jack Dongarra


Rabab Alomairy


Pratik Nag

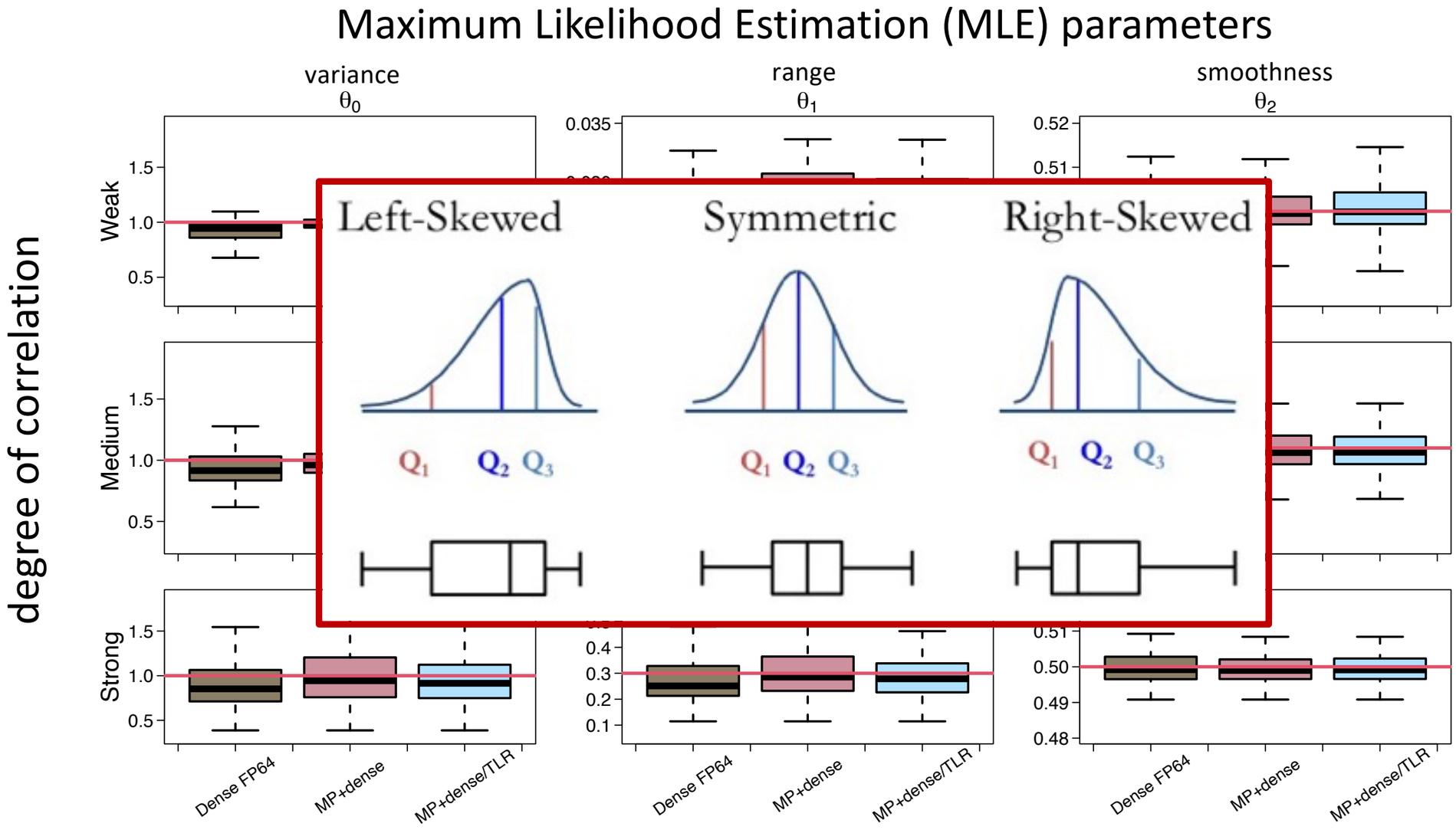

Sameh Abdulah


Hatem Ltaief


Ying Sun


Marc Genton

Accuracy on synthetic 2D space dataset



Accuracy on real 3D (2D space + time) dataset

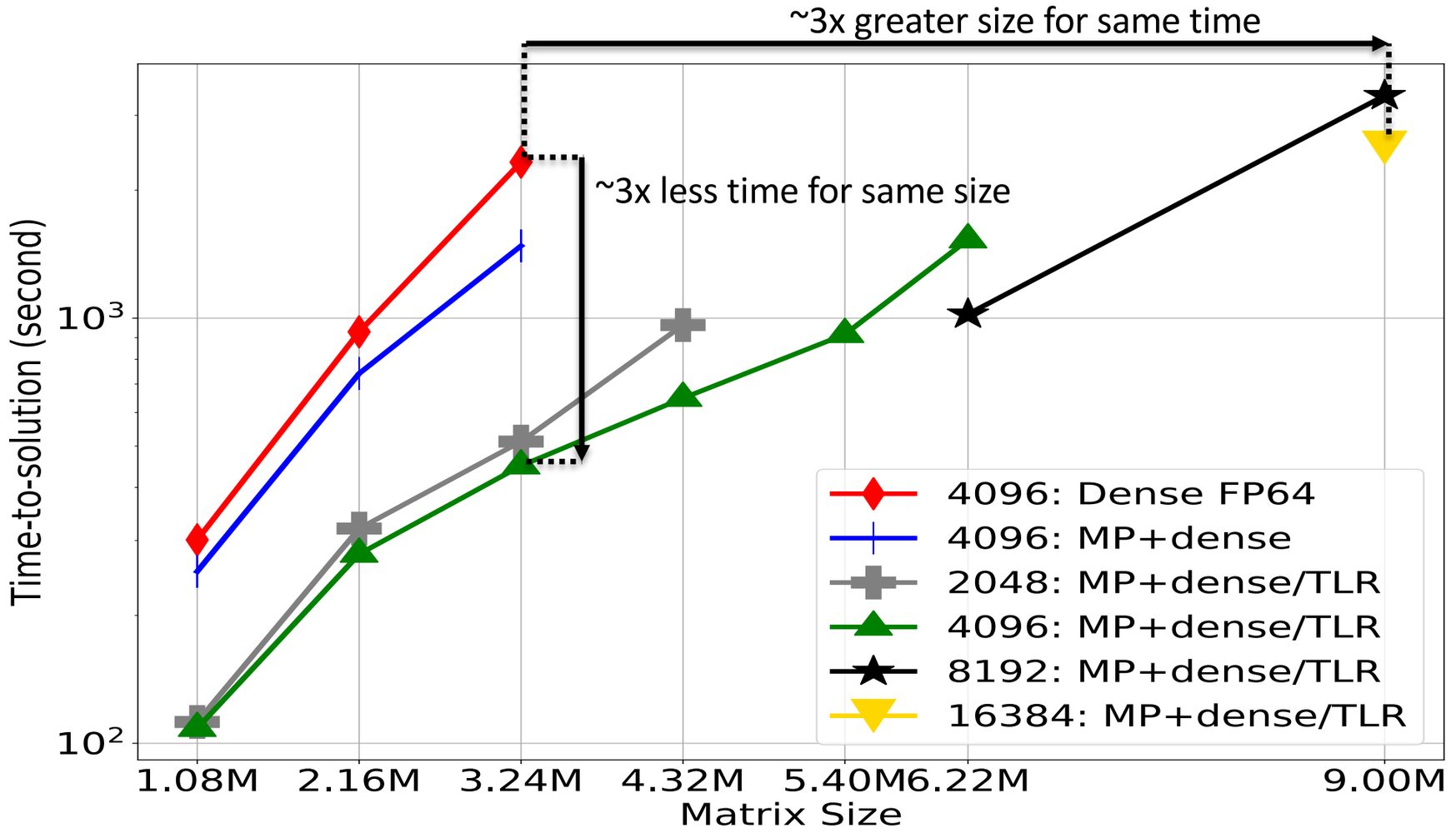
Variants	Variance (θ_0)	Range (θ_1)	Smoothness (θ_2)
Dense FP64	1.0087	3.7904	0.3164
MP+dense	0.9428	3.8795	0.3072
MP+dense/TLR	0.9247	3.7756	0.3068

Variants	Range-time (θ_3)	Smoothness-time (θ_4)	Nonsep-param (θ_5)
Dense FP64	0.0101	3.4890	0.1844
MP+dense	0.0102	3.4941	0.1860
MP+dense/TLR	0.0102	3.5858	0.1857

Variants	Log-Likelihood (llh)	MSPE
Dense FP64	-136675.1	0.9345
MP+dense	-136529.0	0.9348
MP+dense/TLR	-136541.8	0.9428

mean-square prediction error

Performance on up to 16K nodes of Fugaku



To be improved:

Pruning dynamic runtime system *PaRSEC* for Fugaku's small 32GB/node memory

2024 Gordon Bell (climate)

Problem size	54,486,360 spatial locations across the globe at a spatial resolution of 0.034° (~3.5 km)
Category of achievement Type of method used	Scalability and peak performance Spherical Harmonic Transform (SHT) and Cholesky factorization
Results reported on basis of Precision reported System scale	Cholesky factorization Double and mixed-precision - 0.976 EFlop/s on 9,025 nodes of Frontier (36,100 AMD MI250X multi-chip module (MCM) GPUs) equivalent to 72,200 AMD Graphics Compute Dies (GCDs) - 0.739 EFlop/s on 1,936 nodes of Alps (7,744 NVIDIA Grace-Hopper Superchips (GH200)) - 0.243 EFlop/s on 1,024 nodes of Leonardo (4,096 NVIDIA A100 GPUs) - 0.375 EFlop/s on 3,072 nodes of Summit (18,432 NVIDIA V100 GPUs)
Measurement mechanism	Timers, Flops

Boosting Earth System Model Outputs And Saving PetaBytes in Their Storage Using Exascale Climate Emulators

Sameh Abdulah[1,7], Allison H. Baker[2,8], George Bosilca[3,9], Qinglei Cao[4,10], Stefano Castruccio[5,11], Marc G. Genton[1,7], David E. Keyes[1,7], Zubair Khalid[1,6,12], Hatem Ltaief[1,7], Yan Song[1,7], Georgiy L. Stenchikov[1,7], and Ying Sun[1,7]

[1]Extreme Computing & Statistics & Earth Science, King Abdullah University of Science and Technology, KSA
[2]Computational and Information Sciences Lab, NSF National Center for Atmospheric Research, USA
[3]NVIDIA, USA
[4]Department of Computer Science, Saint Louis University, USA
[5]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, USA
[6] Department of Electrical Engineering, Lahore University of Management Sciences, Pakistan
[7]{*Firstname.Lastname*}*@kaust.edu.sa* [8]*abaker@ucar.edu* [9]*gbosilca@nvidia.com*
[10]*qinglei.cao@slu.edu* [11]*scastruc@nd.edu* [12]*zubair.khalid@lums.edu.pk*

***Abstract*—We present the design and scalable implementation of an exascale climate emulator for addressing the escalating computational and storage requirements of high-resolution Earth System Model simulations. We utilize the spherical harmonic transform to stochastically model spatio-temporal variations in climate data. This provides tunable spatio-temporal resolution and significantly improves the fidelity and granularity of climate emulation, achieving an ultra-high spatial resolution of 0.034°(~3.5 km) in space. Our emulator, trained on 318 billion hourly temperature data points from a 35-year and 31 billion daily data points from an 83-year global simulation ensemble, generates statistically consistent climate emulations. We extend linear solver software to mixed-precision arithmetic GPUs, applying different precisions within a single solver to adapt to different correlation strengths. The PaRSEC runtime system supports efficient parallel matrix operations by optimizing the dynamic balance between computation, communication, and memory requirements. Our BLAS3-rich code is optimized for systems equipped with four different families and generations of GPUs, scaling well to achieve 0.976 EFlop/s on 9,025 nodes (36,100 AMD MI250X multi-chip module (MCM) GPUs) of Frontier (nearly full system), 0.739 EFlop/s on 1,936 nodes (7,744 Grace-Hopper Superchips (GH200)) of Alps, 0.243 EFlop/s on 1,024 nodes (4,096 A100 GPUs) of Leonardo, and 0.375 EFlop/s on 3,072 nodes (18,432 V100 GPUs) of Summit.**

***Index Terms*—Dynamic runtime systems, High-performance computing, Mixed-precision computation, Spatio-temporal climate emulation, Spherical harmonic transform, Task-based programming models.**

I. JUSTIFICATION FOR THE GORDON BELL PRIZE

Exascale climate emulator developed using 318 billion hourly and 31 billion daily observations for generating climate emulations at ultra-high spatial resolution (0.034° ~3.5 km).

Authors are listed alphabetically by their last names.

Modeling climate data using spherical harmonics. Mixed-precision computations. PaRSEC dynamic runtime system. Running on 9,025 nodes on Frontier, 1,936 nodes on Alps, 1,024 nodes on Leonardo, and 3,072 nodes on Summit, with the hybrid Flop/s rates 0.976 EFlop/s, 0.739 EFlop/s, 0.243 EFlop/s, and 0.375 EFlop/s, respectively.

II. PERFORMANCE ATTRIBUTES

Problem size	54,486,360 spatial locations across the globe at a spatial resolution of 0.034° (~3.5 km)
Category of achievement Type of method used	Scalability and peak performance Spherical Harmonic Transform (SHT) and Cholesky factorization
Results reported on basis of Precision reported System scale	Cholesky factorization Double and mixed-precision - 0.976 EFlop/s on 9,025 nodes of Frontier (36,100 AMD MI250X multi-chip module (MCM) GPUs) equivalent to 72,200 AMD Graphics Compute Dies (GCDs) - 0.739 EFlop/s on 1,936 nodes of Alps (7,744 NVIDIA Grace-Hopper Superchips (GH200)) - 0.243 EFlop/s on 1,024 nodes of Leonardo (4,096 NVIDIA A100 GPUs) - 0.375 EFlop/s on 3,072 nodes of Summit (18,432 NVIDIA V100 GPUs)
Measurement mechanism	Timers, Flops

III. OVERVIEW OF THE PROBLEM

Climate change, evident in rising temperatures, extreme weather events, sea-level rise, and ecosystem disruption, poses significant risks and urgently requires action due to intensified heatwaves, storms, droughts, floods, and biodiversity loss [1], [2]. We stand at a critical juncture where converging

GB’24 climate prize collaborators

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


Allison Baker


George Boslica


Qinglei Cao


Stefano Castruccio


Gera Stenchikov


Sameh Abdulah


Marc Genton


Zubair Khalid


Hatem Ltaief


Yan Song


Ying Sun

Motivation – statistical alternative to ESMs

- Earth System Models (ESMs) are fundamental to the Intergovernmental Panel on Climate Change (IPCC) sixth assessment report (AR6)
  - climate statistics from ESMs based on PDEs require numerous runs
  - PDE simulations are inefficient (severely memory-bandwidth bound)
- The latest Coupled Model Intercomparison Project (CMIP6) is also storage intensive
  - more than 28 PetaBytes data from 45 participating organizations
- Simulations at “global storm-resolving” scales needed to understand how weather and extremes will be affected by climate change
  - compute and storage costs for ESMs escalate as climate community progresses toward ultra-high-resolution simulations

Enter climate *emulators*

- Climate emulators (CEs) are stochastic models parameterized a relatively small number of ESM runs
  - reproduce the statistics without massive ensemble averaging
- CEs quickly generate multiple emulations of the output of an ESM
- However, previous global CEs had not attained …
  - spatial resolution finer than 100 km
  - temporal resolution finer than daily

Contributions

- Developed and validated new climate emulator
  - emulates up to 54.5 million spatial locations across the globe with spatial resolution of 0.034° (3.5 km) at an hourly resolution for 35 years (1988-2022)

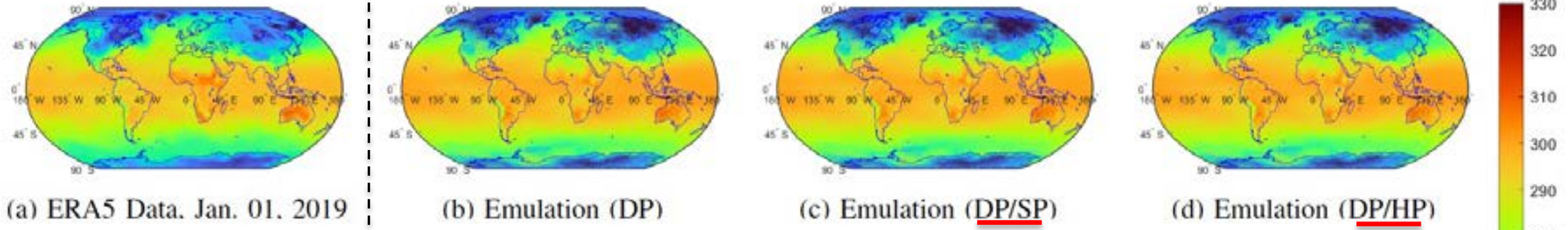
2.5 km yesterday (Hoefler)

- Addressed resolution limitations of existing emulators
  - compresses 2D data on sphere with fast SHTs
  - filters high frequency noise
  - democratizes climate realizations (workstations)
  - plays to architectural strengths (dense matrices)
  - lowers storage barrier

Climate emulation w/ Gaussian processes

Primary cycle-consuming routines for fitting the emulation model are tolerant of mostly lower precision (single and half)

January: ← data emulation →

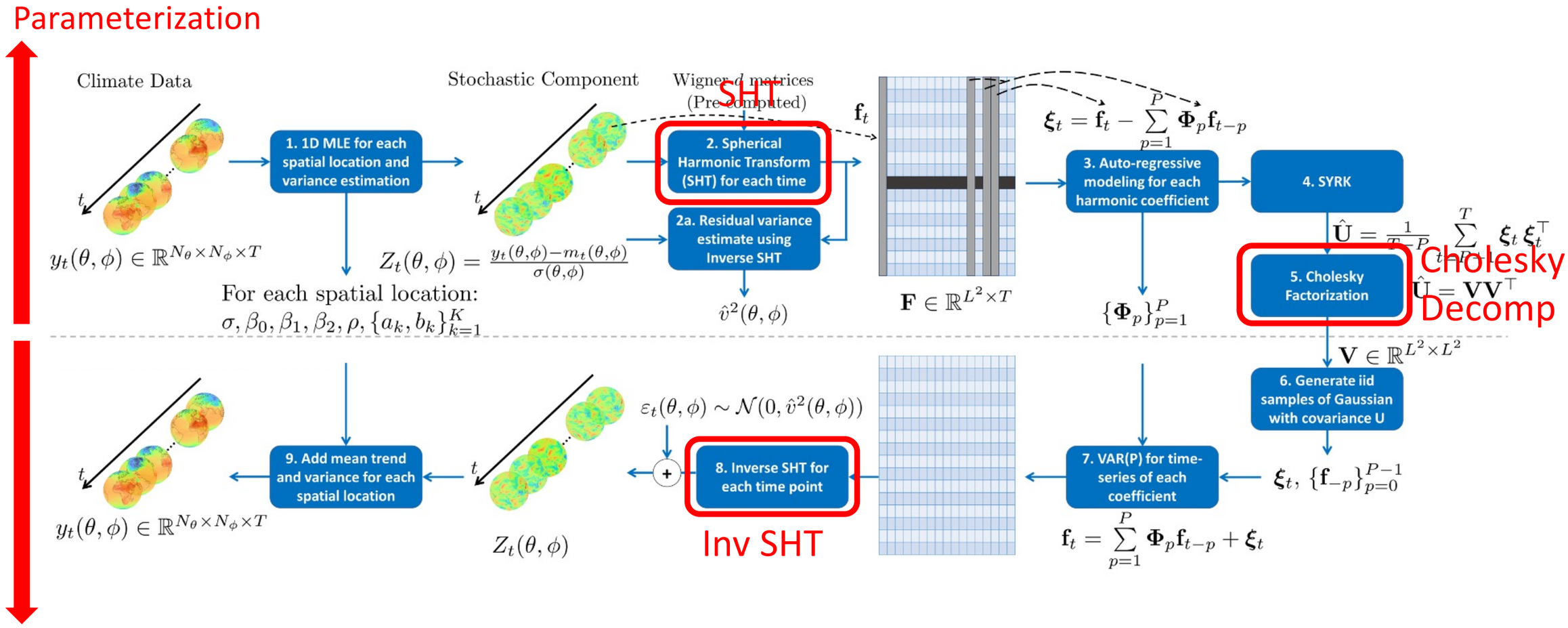


(a) ERA5 Data, Jan. 01, 2019 (b) Emulation (DP) (c) Emulation (DP/SP) (d) Emulation (DP/HP)

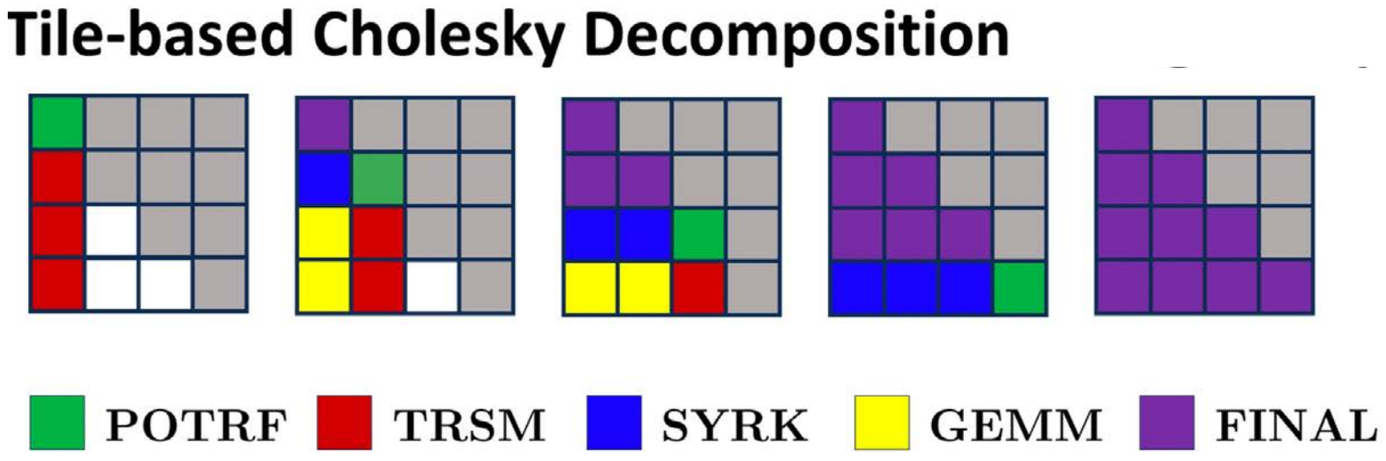
June:

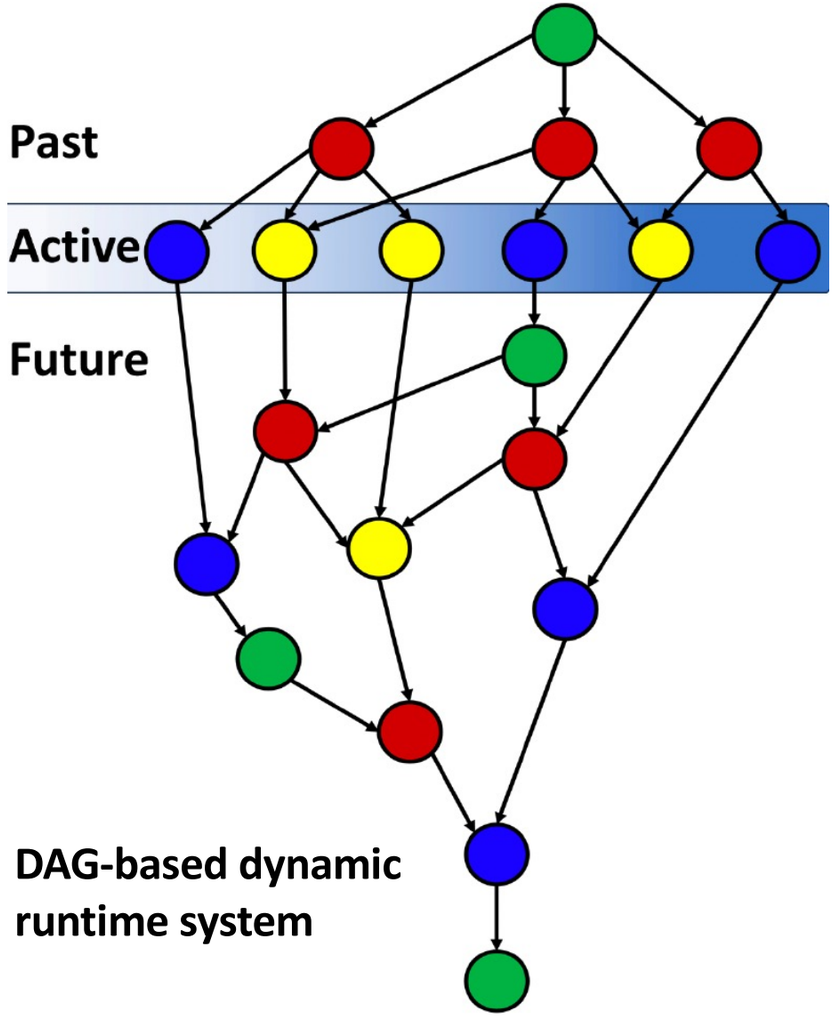
(e) ERA5 Data, Jun. 01, 2019 (f) Emulation (DP) (g) Emulation (DP/SP) (h) Emulation (DP/HP)

Algorithmic ingredients (2 stages, 2 major consumers)

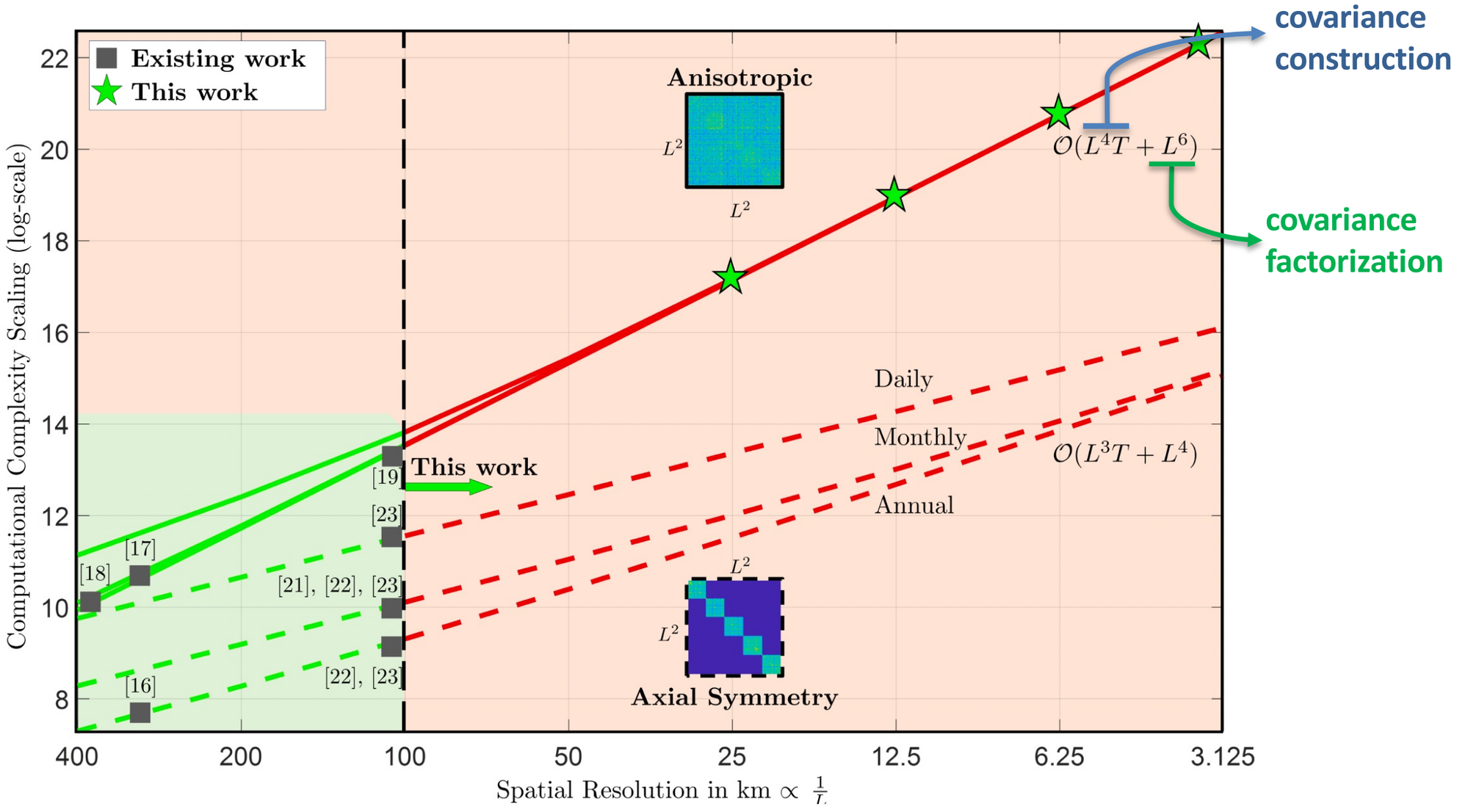


HPC ingredients (tiling and DAG dynamic runtime)





Expanding emulation resolution w/ memory austerity



Performance on four Top10 systems (eff. Pflops/s)

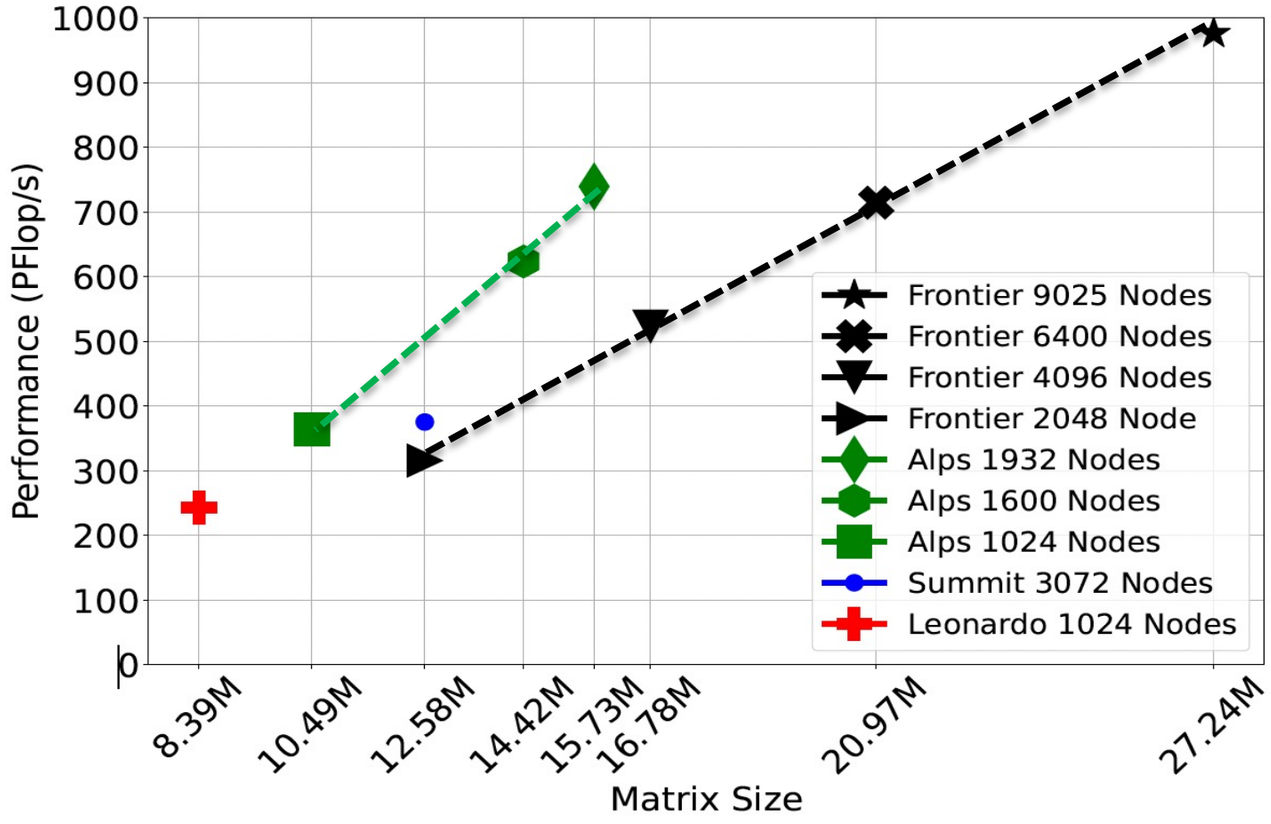


Fig. 8: Performance of largest runs on Summit, Leonardo, Alps, and Frontier; with additional run-up points on Alps and Frontier, all using the DP/HP precision variant.

2025 Gordon Bell

Performance attribute	This submission
Category of achievement	Scalability, time-to-solution, peak performance
Type of method used	Bayesian inversion, FEM, real-time computing
Results reported based on	Whole application including I/O
Precision reported	Double precision
System scale	Results measured on full-scale system
Measurement mechanism	Timers, DOF throughput, FLOP count

Real-Time Bayesian Inference at Extreme Scale: A Digital Twin for Tsunami Early Warning Applied to the Cascadia Subduction Zone

Stefan Henneking
The University of Texas at Austin
Austin, TX, USA
stefan@oden.utexas.edu

Sreeram Venkat
The University of Texas at Austin
Austin, TX, USA
srvenkat@utexas.edu

Veselin Dobrev
Lawrence Livermore National Laboratory
Livermore, CA, USA
dobrev1@llnl.gov

John Camier
Lawrence Livermore National Laboratory
Livermore, CA, USA
camier1@llnl.gov

Tzanio Kolev
Lawrence Livermore National Laboratory
Livermore, CA, USA
kolev1@llnl.gov

Milinda Fernando
The University of Texas at Austin
Austin, TX, USA
milinda@oden.utexas.edu

Alice-Agnes Gabriel
University of California San Diego
San Diego, CA, USA
algabriel@ucsd.edu

Omar Ghattas
The University of Texas at Austin
Austin, TX, USA
omar@oden.utexas.edu

Abstract

We present a Bayesian inversion-based digital twin that employs acoustic pressure data from seafloor sensors, along with 3D coupled acoustic–gravity wave equations, to infer earthquake-induced spatiotemporal seafloor motion in real time and forecast tsunami propagation toward coastlines for early warning with quantified uncertainties. Our target is the Cascadia subduction zone, with one billion parameters. Computing the posterior mean alone would require 50 years on a 512 GPU machine. Instead, exploiting the shift invariance of the parameter-to-observable map and devising novel parallel algorithms, we induce a fast offline–online decomposition. The offline component requires just one adjoint wave propagation per sensor; using MFEM, we scale this part of the computation to the full El Capitan system (43,520 GPUs) with 92% weak parallel efficiency. Moreover, given real-time data, the online component exactly solves the Bayesian inverse and forecasting problems in 0.2 seconds on a modest GPU system, a ten-billion-fold speedup.

CCS Concepts

• **Mathematics of computing** → **Solvers**; **Mathematical software performance**; **Partial differential equations**; **Computation of transforms**; **Mesh generation**; **Discretization**; • **Computing methodologies** → **Massively parallel algorithms**; **Uncertainty quantification**; **Model verification and validation**; **Modeling methodologies**; **Real-time simulation**; **Data assimilation**; **Massively parallel and high-performance simulations**; **Scientific visualization**; • **Applied computing** → **Earth and atmospheric sciences**; **Mathematics and statistics**.

SC '25, St Louis, MO, USA

ACM ISBN 979-8-4007-1466-5/25/11
https://doi.org/10.1145/3712285.3771787

Keywords

Bayesian inverse problems, uncertainty quantification, digital twins, data assimilation, finite elements, real-time GPU supercomputing, tsunami early warning

ACM Reference Format:
Stefan Henneking, Sreeram Venkat, Veselin Dobrev, John Camier, Tzanio Kolev, Milinda Fernando, Alice-Agnes Gabriel, and Omar Ghattas. 2025. Real-Time Bayesian Inference at Extreme Scale: A Digital Twin for Tsunami Early Warning Applied to the Cascadia Subduction Zone. In *The International Conference for High Performance Computing, Networking, Storage and Analysis (SC '25), November 16–21, 2025, St Louis, MO, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3712285.3771787

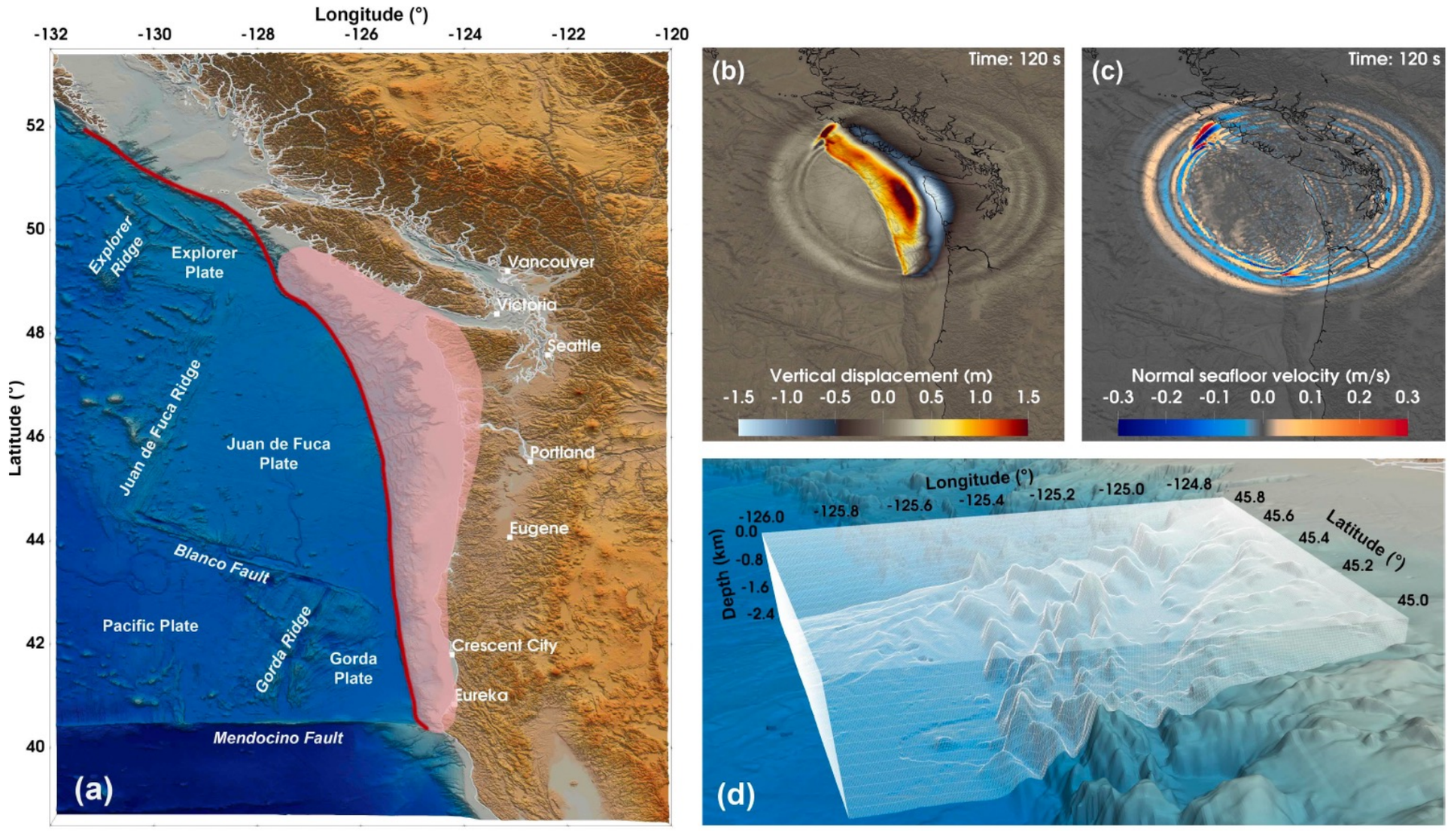
1 Justification for ACM Gordon Bell Prize

Fastest time-to-solution of a PDE-based Bayesian inverse problem with 1 billion parameters in 0.2 seconds, a ten-billion-fold speedup over SoA. Largest-to-date unstructured mesh FE simulation with 55.5 trillion DOF on 43,520 GPUs, with 92% weak and 79% strong parallel efficiencies in scaling over a 128× increase of GPUs on the full-scale *El Capitan* system.

2 Performance Attributes

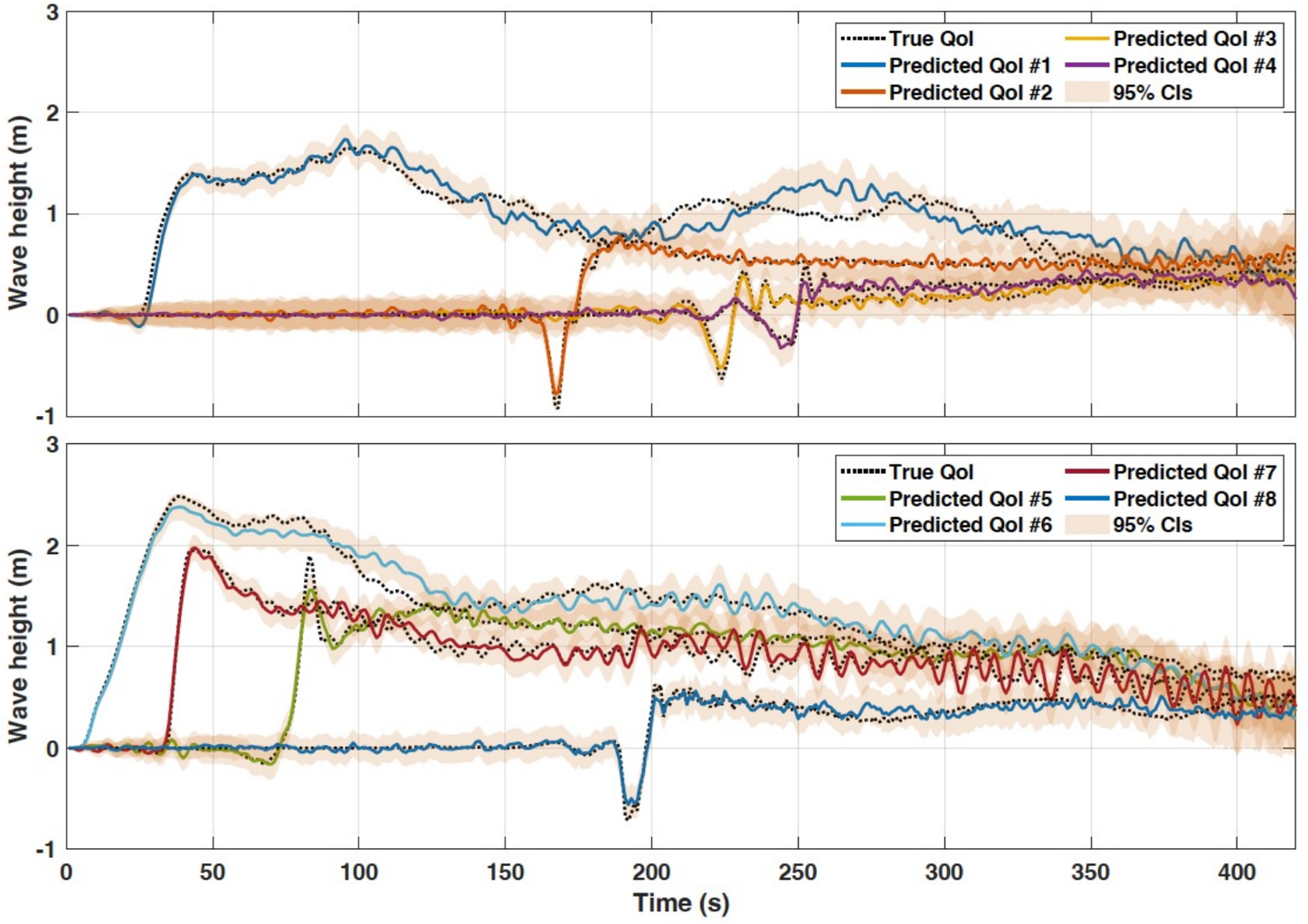
Performance attribute	This submission
Category of achievement	Scalability, time-to-solution, peak performance
Type of method used	Bayesian inversion, FEM, real-time computing
Results reported based on	Whole application including I/O
Precision reported	Double precision
System scale	Results measured on full-scale system
Measurement mechanism	Timers, DOF throughput, FLOP count

Physical situation being modeled

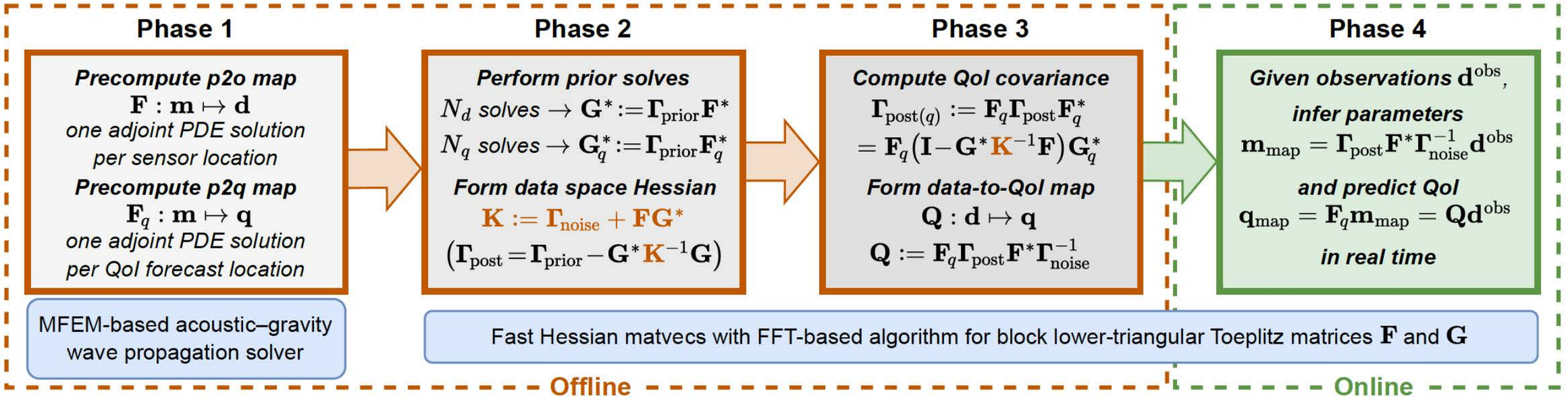


Prediction capability

- Eight quantities of interest (“QoIs”)
- Waveheights versus time over a 7-minute period from the initiative of the tsunami at sensitive coastal locations for people or infrastructure
- Comparing real-time prediction from surrogate to full fidelity model
- Showing confidence intervals (CIs)

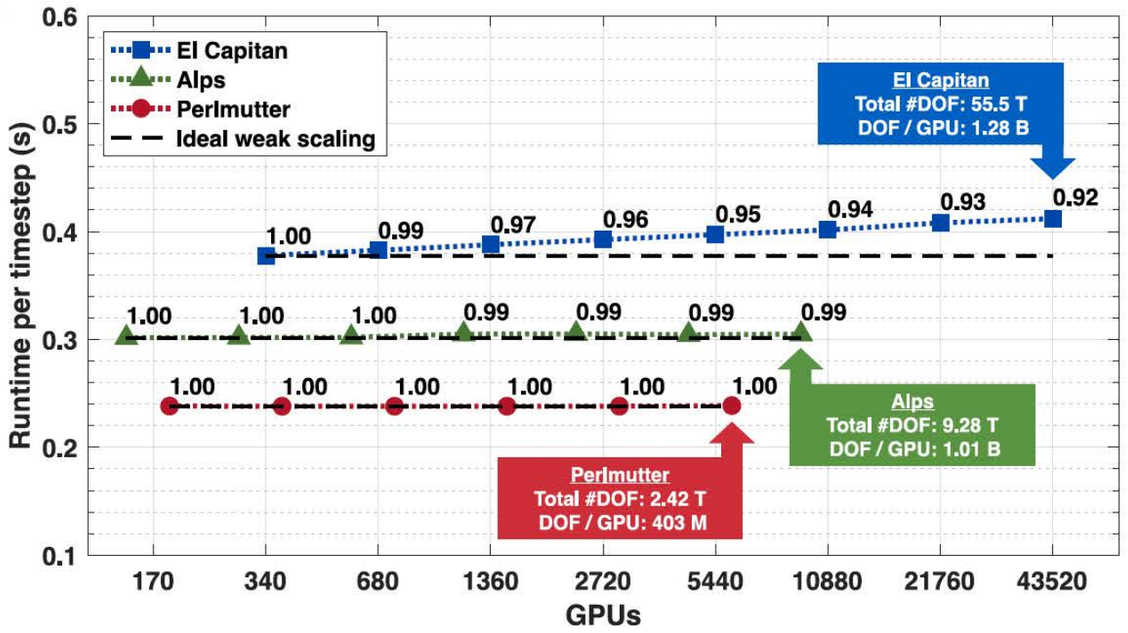


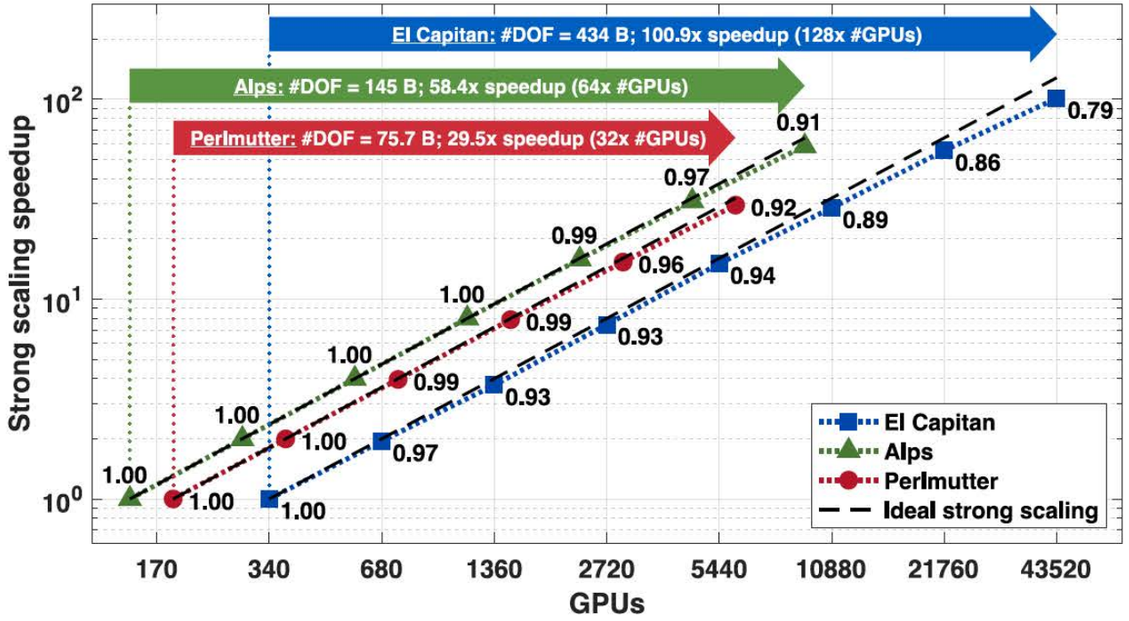
HPC considerations of the workflows



- Offline computations for high fidelity model require over 250K A100 GPU-hours
- Online computations for digtal twin require less than a second per set of observations

HPC scaling merits





Weak (left) and strong scaling (right)

- El Capitan (#1)
- ALPS (#7)
- Perlmutter (#30)

Outline of presentation

Examples of High Performance Statistical Computing

- Gordon Bell campaigns of 2022, 2024*, 2025*
- Open source HPSC software

Twelve “universals” of HPC algorithms and software

- Things you *wouldn’t* consider in a proof of concept app
- Things you *must* consider in a high-performamce app

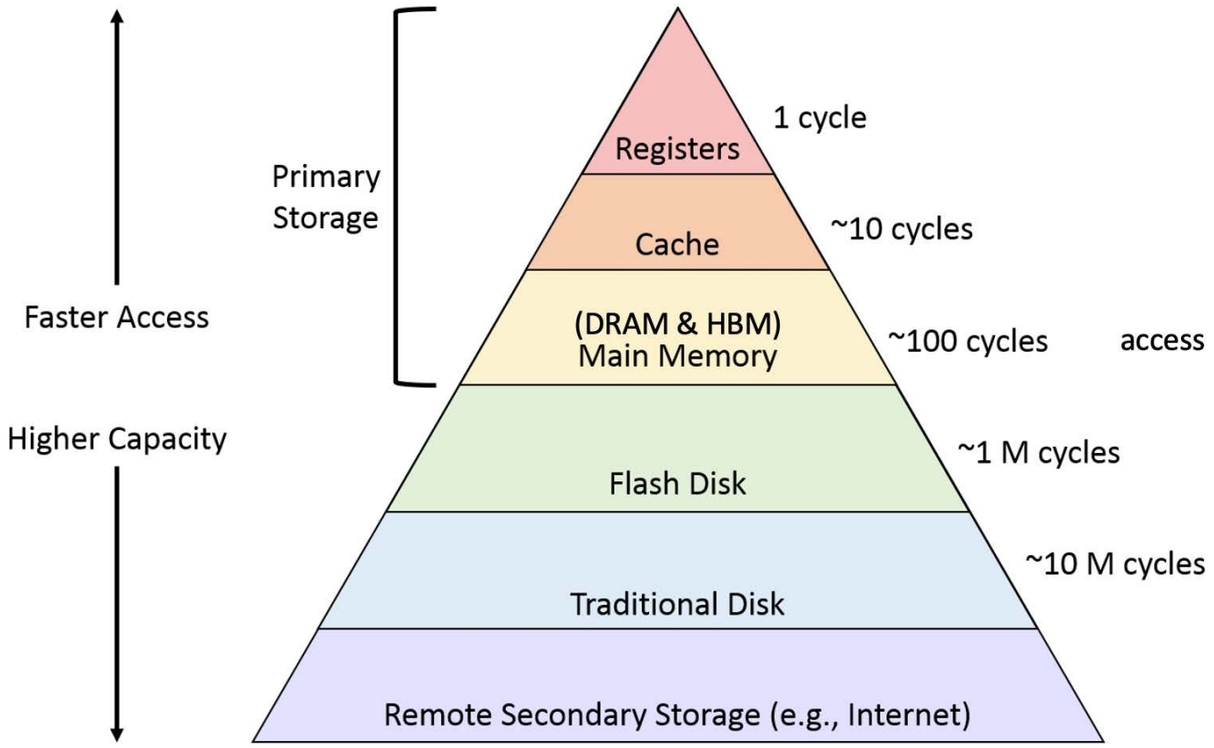
Twelve “elements” of the HPC ecosystem

- No need to start from scratch – HPSC can ride the HPC wave
- Will greatly enrich an *existing* HPC ecosystem

12 universals of HPC algorithms and software

- Reside high on the memory hierarchy
- Reduce synchrony in frequency and/or in span
- Reduce communication in number and/or volume of messages
- Employ dynamic scheduling and balancing
- Avoid over-resolving with respect to output accuracy requirements
- Reformulate applications before computing
- Exploit the “right to re-order”
- Exploit multiple hierarchical versions of the same system
- Exploit data sparsity to meet “curse of dimensionality” w/ “blessing of low rank”
- Take resilience into algorithms, relieving hardware and systems
- Code to specialized back-ends while presenting high-level APIs to users
- Consider multiple parallel programming models in one application

Reside high on the memory hierarchy



Operation	approximate energy cost
DP FMADD flop	100 pJ
DP DRAM read-to-register	5,000 pJ
DP word transmit-to-neighbor	7,500 pJ
DP word transmit-across-system	10,000 pJ

Remember that a *pico* (10^{-12}) of something done *exa* (10^{18}) times per second is a *mega* (10^{6})-somethings per second

- **100 pJ at 1 Eflop/s is 100 MW (for the flop/s only!)**
- **1 MW-year costs about $1M ($0.12/KW-hr × 8760 hr/yr)**
  - **We "use" 1.4 KW continuously, so 100MW is 71,000 people**

Reduce synchrony in frequency and/or in span

Frequency of inner products in preconditioned conjugate gradients can be reduced at the cost of greater arithmetic per synchronization and greater storage

Savings up to 2.5x on 3D hydrostatic ice sheet flow app

Table 1
Overview of the different preconditioned CG and CR variations. Column *flops* lists the number of flops ($\times N$) for AXPYS and dot-products. The *time* column has the time spent in global all-reduce communication (G), in the matrix–vector product (SPMV) and the preconditioner (PC). Column *#global synchronizations* has the number of global communication phases per iteration. The *memory* column counts the number of vectors that need to be kept in memory (excluding x and b).

	Flops	Time (excl, AXPYS, DOTS)	#Glob syncs	Memory
CG	10	2G + SpMV + PC	2	4
Chron/Gear-CG	12	G + SpMV + PC	1	5
CR	12	2G + SpMV + PC	2	5
Pipe-CG	20	max(G, SpMV + PC)	1	9
Pipe-CR	16	max(G, SpMV) + PC	1	7
Gropp-CG	14	max(G, SpMV) + max(G, PC)	2	6

NB: log scale

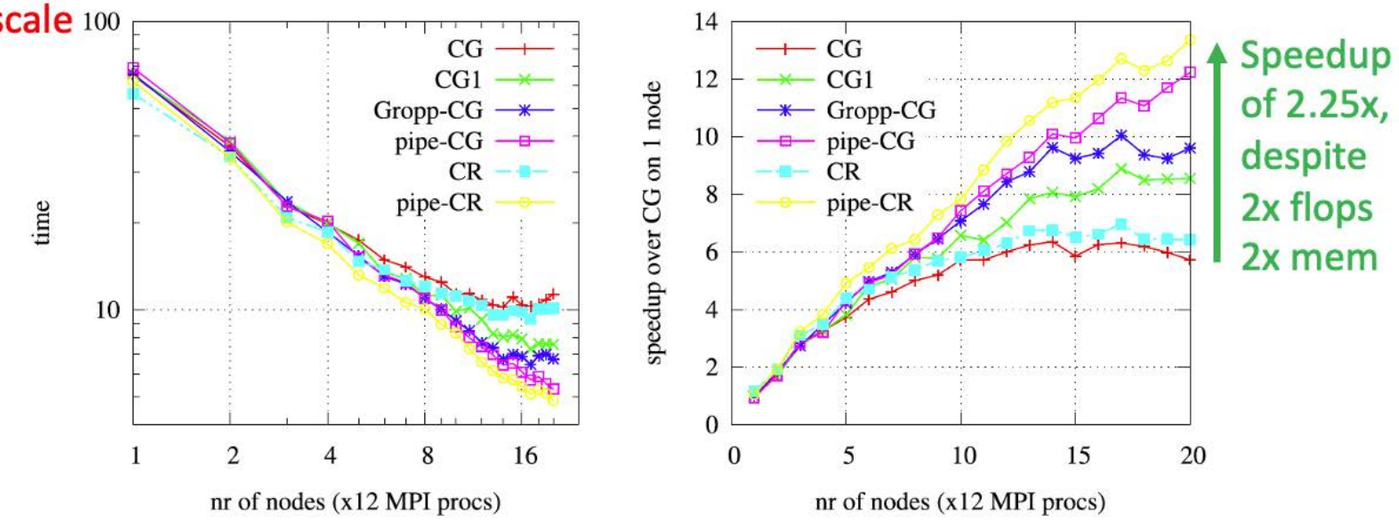


Fig. 3. Left: Time to solution for the 3*D* hydrostatic ice sheet flow simulation using $100 \times 100 \times 50$ *Q*1 finite elements. Right: Speedup as function of number of nodes over standard CG solver on a single node.

Reduce communication in number and/or volume of messages

Communication volume can be reduced at a cost of greater memory per node in so-called 2.5D matrix algorithms for LU, Cholesky, QR, Gram-Schmidt, and eigensolvers

Many implementations attain the theoretical lower bounds

SIAM J. MATRIX ANAL. & APPL.
Vol. 32, No. 3, pp. 866–901

MINIMIZING COMMUNICATION IN NUMERICAL LINEAR ALGEBRA*

GREY BALLARD†, JAMES DEMMEL‡, OLGA HOLTZ§, AND ODED SCHWARTZ¶

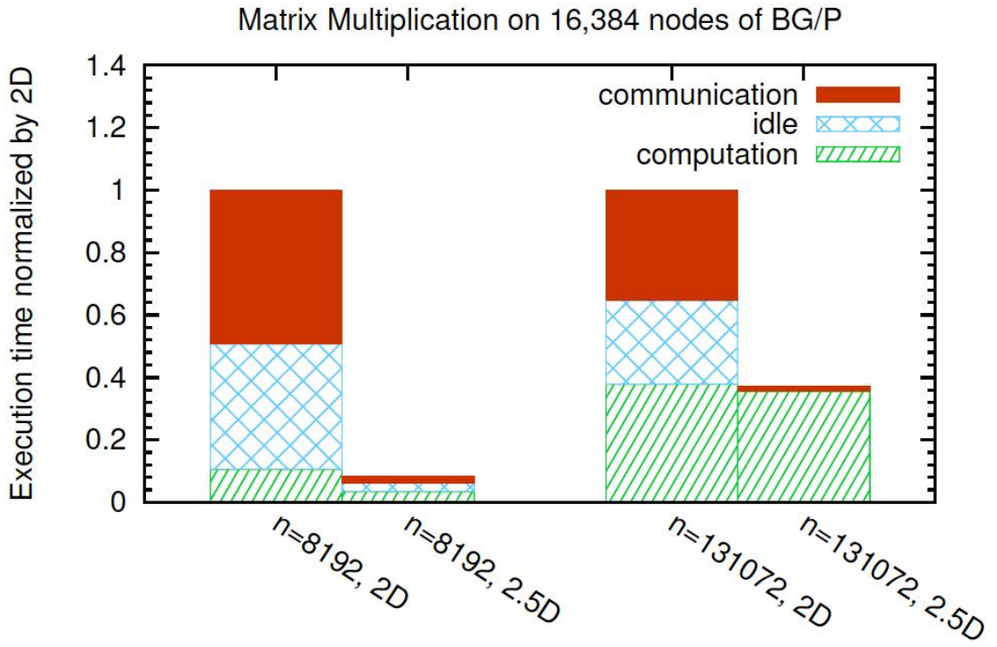
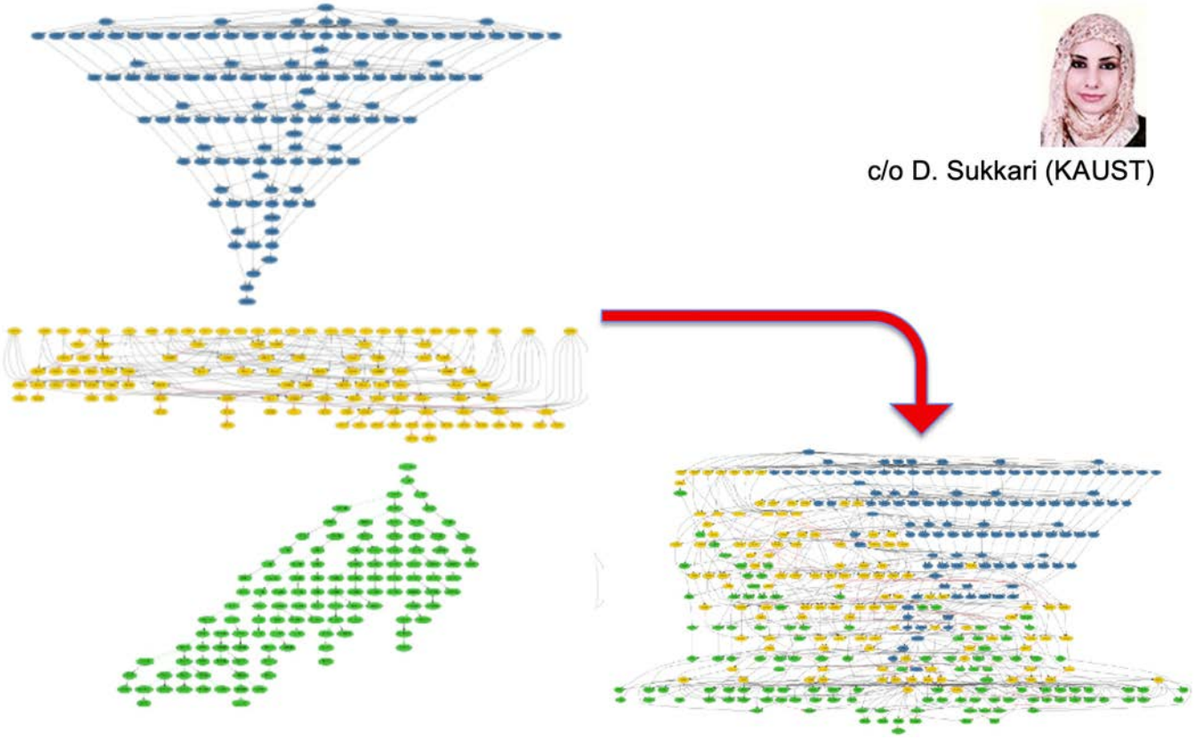


Figure 3.2. 2.5D matrix multiplication on BG/P, 16K nodes / 64K cores.

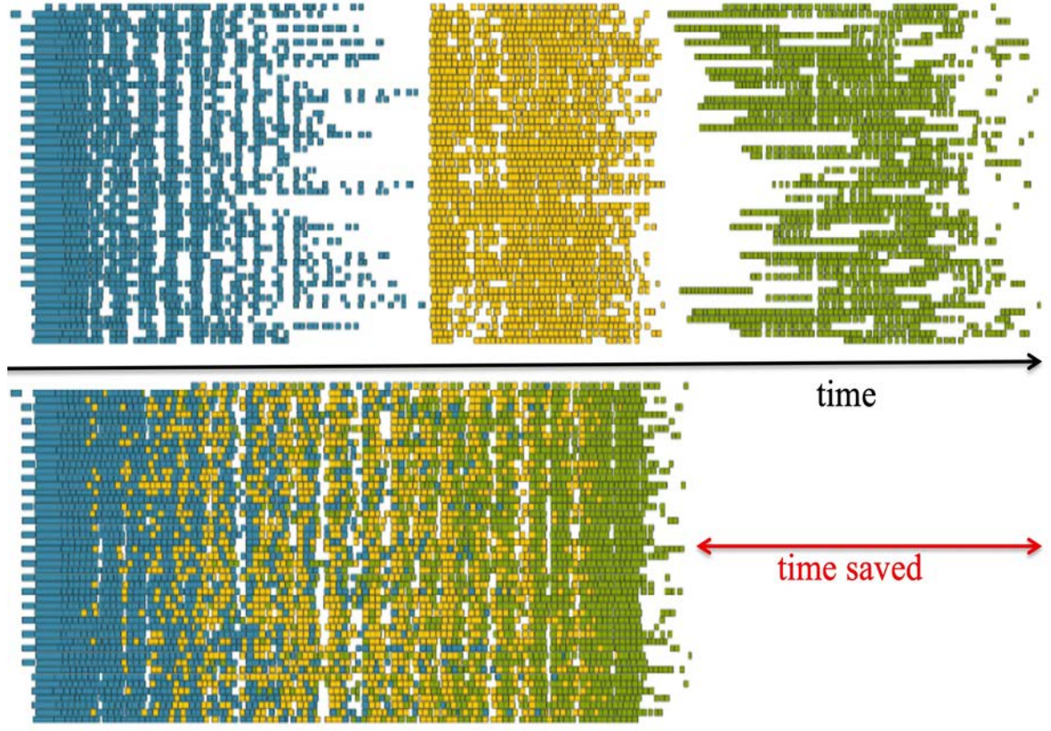
Employ dynamic scheduling and balancing

Loop orderings can be loosened and subroutine boundaries merged to find greater concurrency and adapt to dynamic load imbalance



c/o D. Sukkari (KAUST)

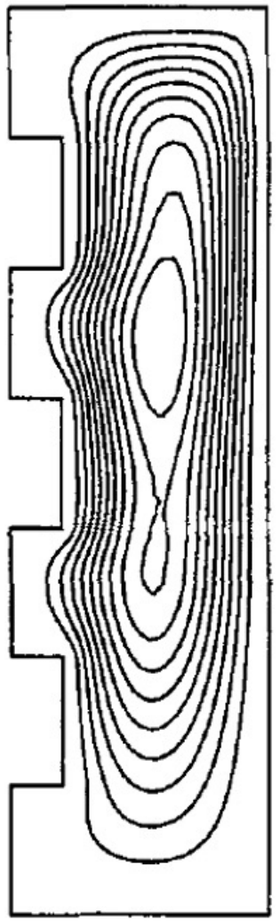
Significant runtime reductions can be achieved by squeezeing out idle time (comparisons for a generalized eigensolver)



Avoid over-resolving with respect to output accuracy requirements

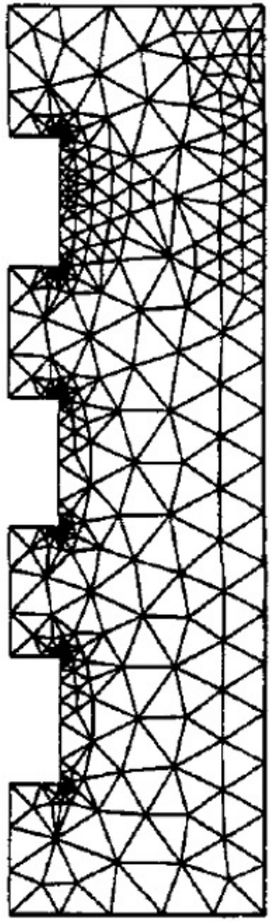
Sometimes, the output of interest from a computation is not a solution to high accuracy everywhere, but a *functional* of the solution to a *specified accuracy*, e.g.

- compute the convective heat flux across a fluid-solid boundary, obtainable without globally uniform accuracy
- use low fidelity surrogates in early inner iterations of “outer loop problems”



temperature contour

conservative mesh



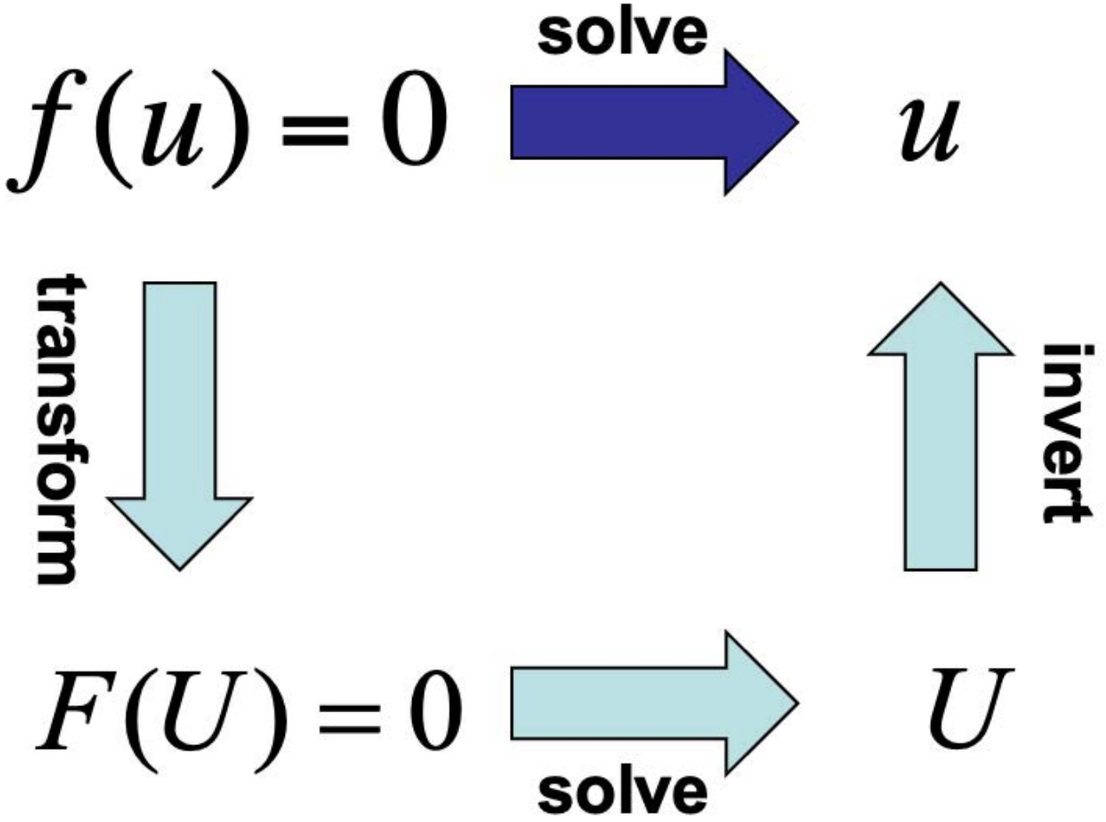
output bound mesh (flux to 1%)

Machiels, Peraire & Patera, *A posteriori FE Output Bounds for the Incompressible NS Equations*, (2001), J. Comp. Phys. **172**:401

Reformulate applications before computing

- It is often difficult to solve a problem directly, as posed, compared with
  - transforming problem to new space
  - solving problem in new space
  - transforming back to original space
- Caveat: there is a sometimes a *conservation of difficulty* in transforming back
- "Think. Then discretize."

  – Vladimir Rokhlin



Exploit the “right to re-order”

- Blocking to reduce indirection and reordering loops for temporal cache locality or shared memory concurrency can save 5x in unstructured CFD

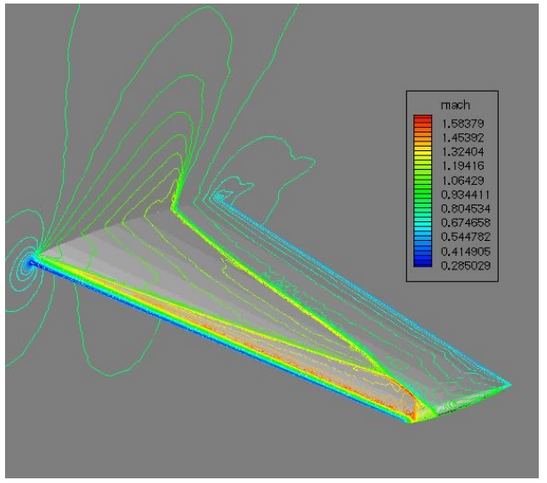


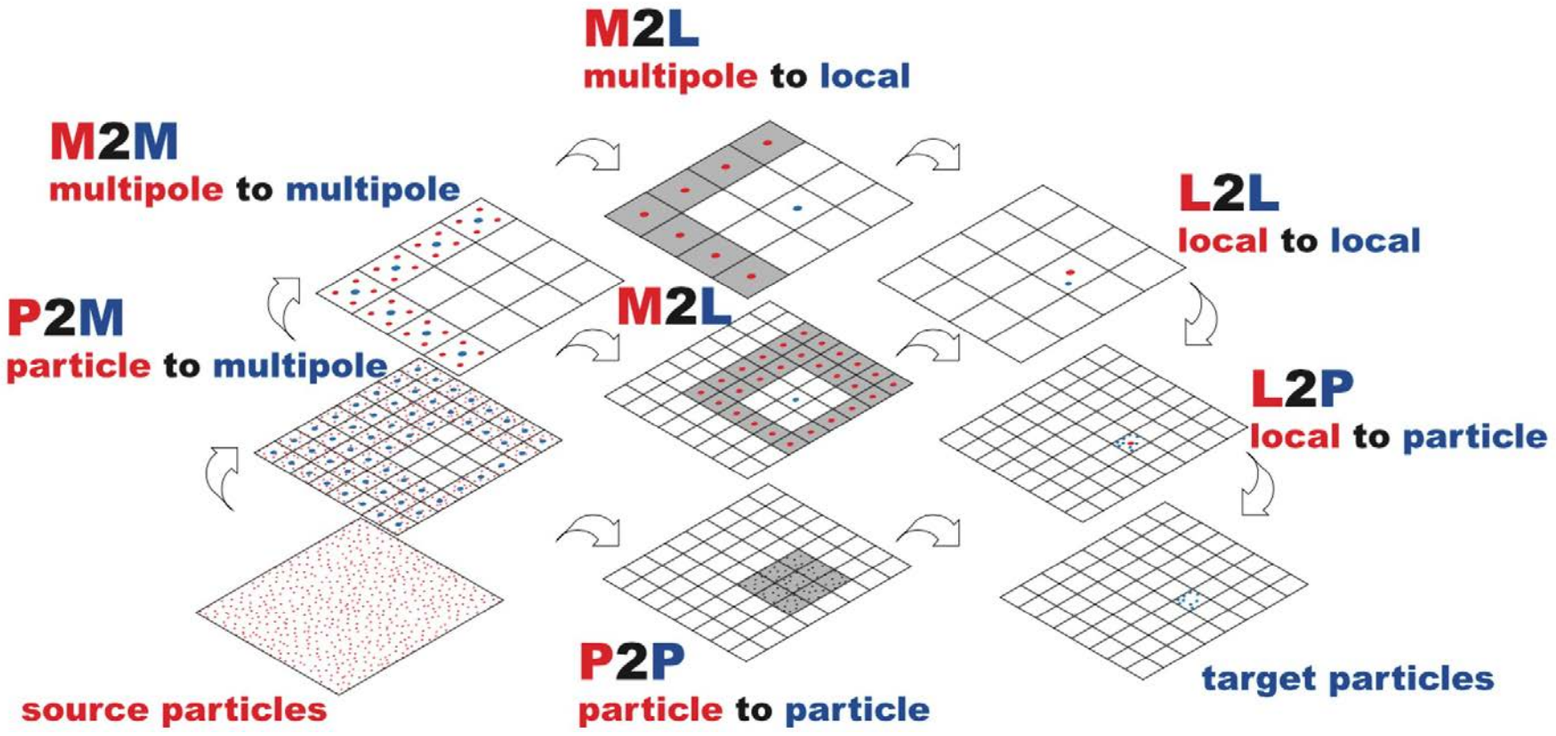
Table 3
Execution times for Euler flow over M6 wing for a fixed-size grid of 22,677 vertices (90,708 DOFs incompressible; 113,385 DOFs compressible)[a]

Enhancements			Results			
			Incompressible		Compressible	
Field interlacing	Structural blocking	Edge reordering	Time/step (s)	Ratio	Time/step (s)	Ratio
			83.6	–	140.0	–
×			36.1	2.31	57.5	2.44
×	×		29.0	2.88	43.1	3.25
		×	29.2	2.86	59.1	2.37
×		×	23.4	3.57	35.7	3.92
×	×	×	16.9	4.96	24.5	5.71

[a] The processor is a 250 MHz MIPS R10000. Activation of a layout enhancement is indicated by “×” in the corresponding column.

Exploit multiple hierarchical versions of a system

$O(N^2)$ complexity of computing forces of N interacting particles (e.g., gravitationally or electrostatically) can be reduced to $O(N)$ operations by local multipole expansion (**P2M**), coarsening (**M2M**), translation (**M2L**), refinement (**L2L**), and local re-expansion (**L2P**) at the target, plus small $O(n^2)$ **P2P** within a cell



Source cells in red, sample target cell in blue
Fast Multipole diagram c/o R. Yokota

Meet “curse of dimensionality” with “blessing of low rank”

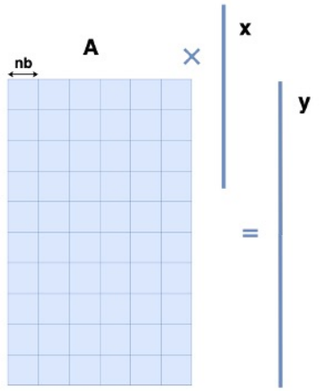


Fig. 2: Original dense MVM.

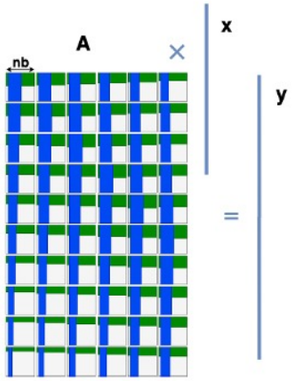


Fig. 3: Rank-compressed operator.

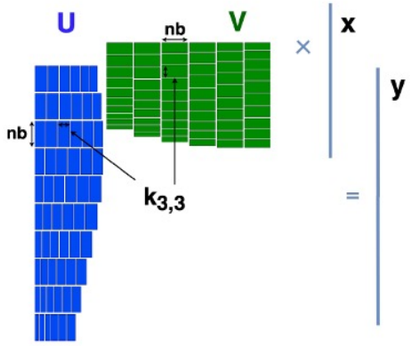


Fig. 4: Stacked bases U and V.

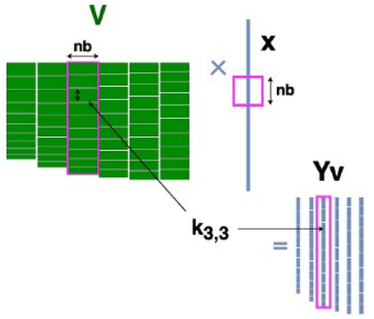


Fig. 5: V-batch stage of MVM.

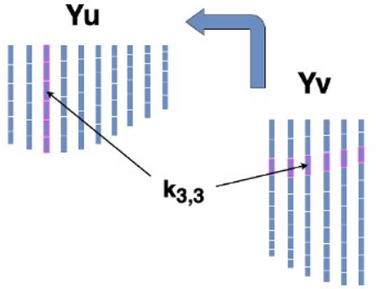


Fig. 6: Shuffle from V to U bases.

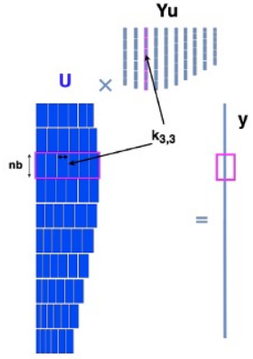
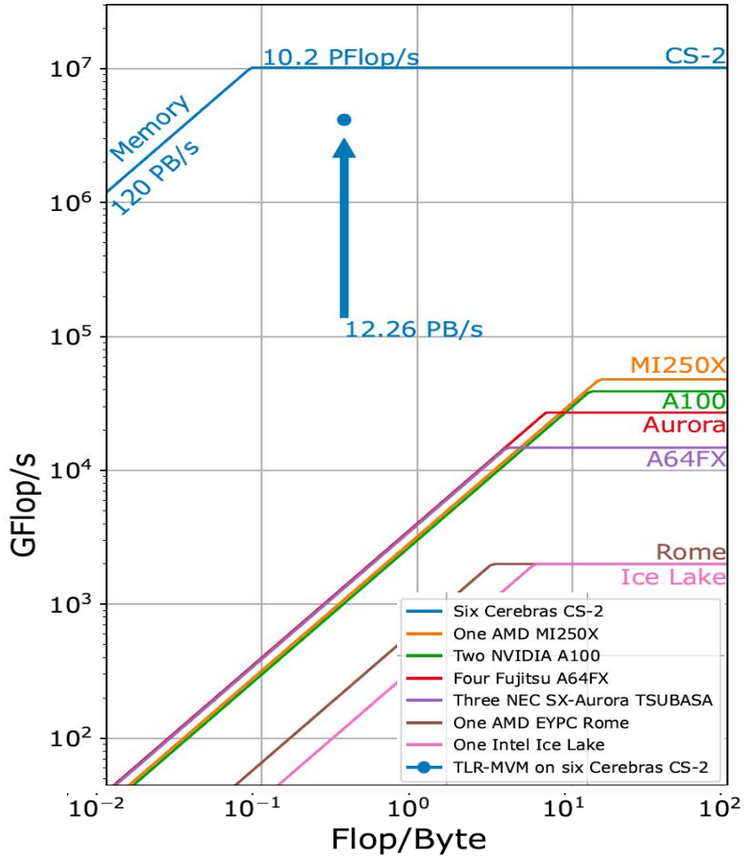


Fig. 7: U-batch of MVM.

Frequency domain source to receiver map in seismic processing application, with tile low-rank decomposition to save memory for all SRAM Cerebras CS-2 WFE



Roofline model showing CPUs and GPUs bandwidth-bound, where CS-2 WFE is near full compute potential

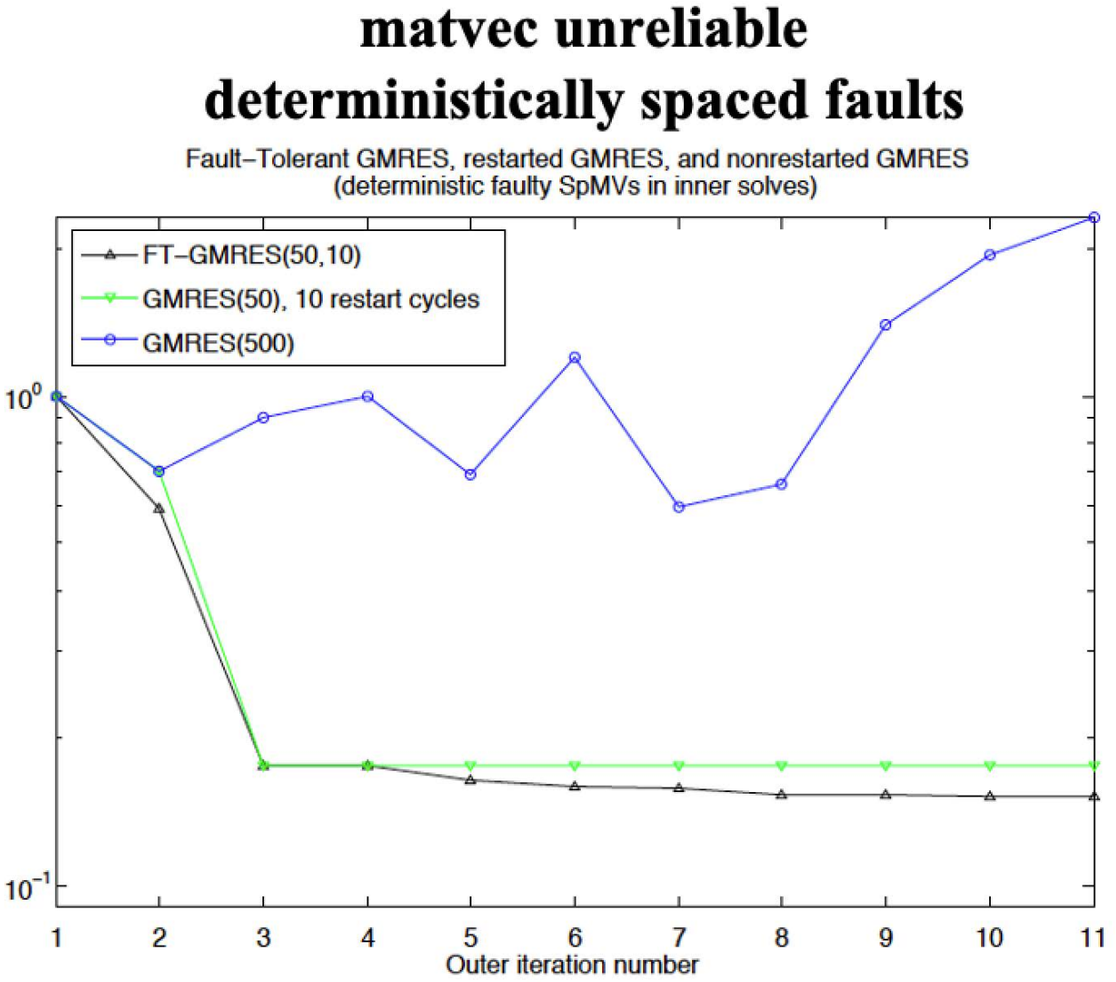
Take resilience into algorithms space

GMRES fails

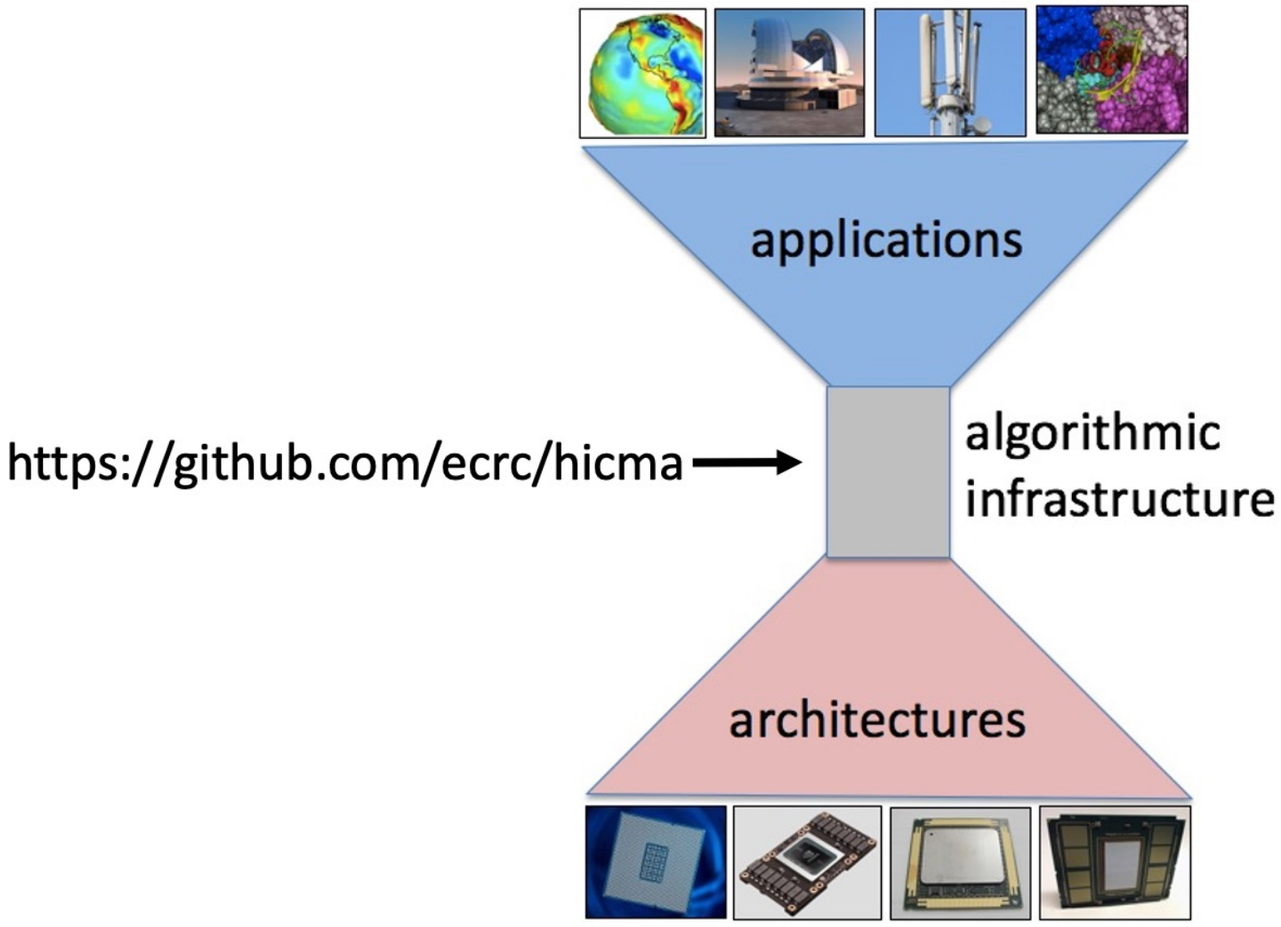
Restarted GMRES has resilience built in, through re-evaluation of the residual, like many iterative methods.

FT-GMRES, which is built on FGMRES, which already allows variations in preconditioning by storing two Krylov subspaces, is more robust.

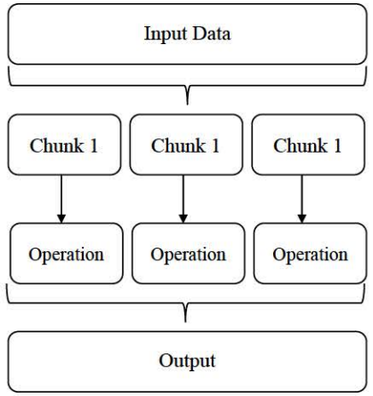
(Experiments from Sandia National Labs)



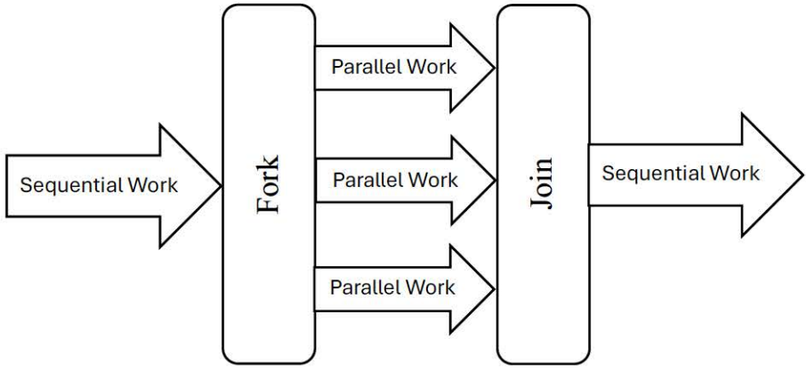
Code to specialized back-ends while presenting high-level APIs



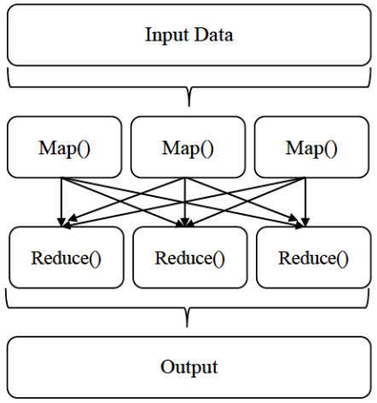
Consider multiple parallel programming models in one app



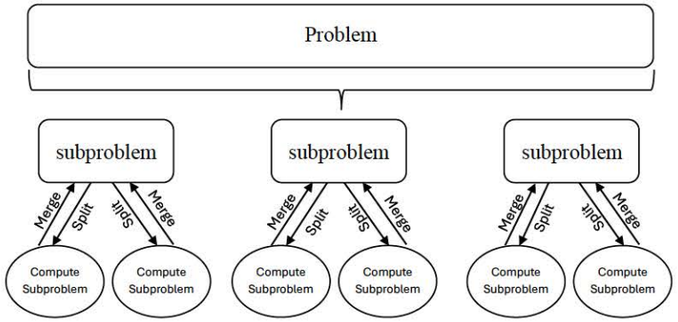
(a) Data Parallelism



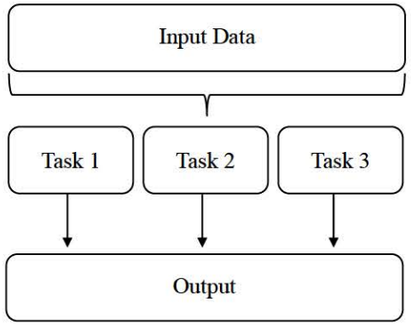
(b) Fork/Join



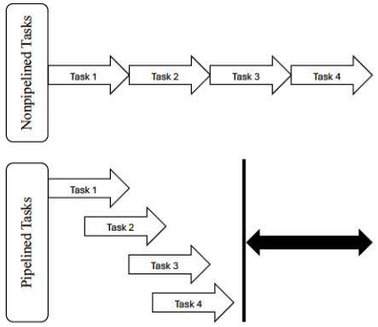
(c) Map/Reduce



(d) Divide and Conquer



(e) Task Parallelism



(f) Pipeline

For more on the HPC universals

siam | Society for Industrial and Applied Mathematics

CONFERENCES & EVENTS – SIAM CONFERENCES – PP26

SIAM CONFERENCES

SIAM Conference on Parallel Processing for Scientific Computing (PP26)

3 – 6 March 2026
Berlin

big trends

DOI:10.1145/3447737

BY DAVID KEYES

The Arab World Prepares the Exascale Workforce

THE ARAB WORLD is currently host to eight supercomputers in the Top500 globally, including the current #10 and a former #7. Hardware can become a honeypot for talent attraction—senior talent from abroad, and rising talent from within. Good return on investment from leading-edge hardware motivates forging collaborative ties to global supercomputing leaders, which leads to integration into the global campaigns that supercomputing excels in, such as predicting climate change and developing sustainable energy resources for its mitigation, positing properties of new materials and catalysis by design, repurposing already-certified drugs and discovering new ones, and big data analytics and machine learning applied to science and to society. While the petroleum industry has been the historical motivation for supercomputing in the Arab World, with its workloads of seismic imaging and reservoir modeling, the attraction today is universal.

However, it is not sufficient to install and boot supercomputers. Their purpose is performance, and their acquisition and operating costs are too high to use them any other way. In each phase of computation, the limiting resource must be identified and computation reorganized to push the bottleneck further away, ideally guided by a performance model. The soul of the machine is the software: the distributed shared memory data structures, the task graphs, the communication patterns. The software is generally not performance-portable; it must be re-tuned in each application-architecture context. As applications become more ambitious and architectures become more austere, algorithms and software must bridge the growing gap.

Hands-on opportunities to resolve this application-architecture tension are a lure to students who had not previously considered supercomputing careers. In some cases, they must surmount significant hurdles in mathematical or computational preparation to enlist, but enlist they do, and they often wind up in globally leading institutions upon graduation. The stories in this article grew up around a university-operated supercomputer, but they largely can be replicated with much less investment because the main challenges today are not in coordinating tens of thousands of nodes across a low-latency, high-bandwidth network. Rather, the challenges lie in extracting performance *from within* increasingly heterogeneous nodes. Furthermore, the cloud now provides high-performance computing (HPC) environments. It is estimated that the percentage of HPC jobs run in the public cloud nearly doubled from 10%–12% in 2018 to 20% in 2019.[4] Many supercomputers, including the currently #1-ranked Fugaku (featuring ARM-based Fujitsu A64fx) and #2-ranked Summit (featuring

Outline of presentation

Examples of High Performance Statistical Computing

- Gordon Bell campaigns of 2022, 2024*, 2025*
- Open source HPSC software

Twelve “universals” of HPC algorithms and software

- Things you *wouldn’t* consider in a proof of concept app
- Things you *must* consider in a high-perfornamce app

Twelve “elements” of the HPC ecosystem

- No need to start from scratch – HPSC can ride the HPC wave
- Will greatly enrich an *existing* ecosystem

12 elements of the HPC ecosystem

- Validation, verification, and uncertainty quantification
- Repeatability, replicability, and reproducibility
- Standardization of tools
- Pre-competitive industry roadmaps
- Discipline-wide benchmark problems
- Industry-wide performance benchmarks
- Virtuous cycle
- Vertical tool chain
- Real-time immersion
- Physical modeling and performance modeling
- Hourglass and reusability
- Convergence of the paradigms

Validation, verification, and uncertainty quantification

The Journal of Verification, Validation and Uncertainty Quantification disseminates original and applied research applied to: design of experiments; computational models; and analysis of experimental results. We encourage authors of papers that describe discipline-specific models and experiments to consider formulating their papers in two parts, the discipline specific part to be published in their home journals and the part describing the validation and uncertainty aspects of their work that would be published in the VV&UQ Journal.

Areas of interest include: Code verification; Solution verification; Validation; Uncertainty quantification; Model prediction; Model adequacy; Model accuracy; Predictive capacity; Model maturity; Phenomena identification and ranking table (PIRT); Design of experiments; Experimental uncertainty; Uncertainty in measurement; Model uncertainty; Model discrepancy; Sensitivity analysis; Model fidelity; Intended use; Context of use; Regulatory science; Aleatoric uncertainty; Epistemic uncertainty; Comparator; Quantification of margins and uncertainties (QMU); Fundamentals of probability; Applications of probability; Bayesian inference

Repeatability, replicability, and reproducibility

- Repeatability
  - A researcher can reliably repeat own computation.
- Replicability
  - An independent group can obtain the same result using the author's tools.
- Reproducibility
  - An independent group can obtain the same result using tools developed independently

Standardization of tools (common APIs for users)

Example: the Message Passing Interface (MPI, mpi-forum.org)

- Universal standard for distributing a computation over millions of memory domains
- Hardware-portable, multi-language communication library
- Allows flourishing of parallel application development (billions of dollars of investment today)
- MPI Forum first met April 1992, released MPI version 1.0 in June 1994
- Involved 80 people from 40 organizations (industry, academia, government labs) contributed by their organizations and funded centrally by ARPA and NSF
- In continual adaptation as hardware and applications evolve (currently MPI 5.0)

Industry roadmaps

- The ITRS roadmap
  - https://www.hpcwire.com/2016/07/28/transistors-wont-shrink-beyond-2021-says-final-itrs-report/?eid=328378666&bid=1482255
  - first published in 1998
- Predecessor, the NTRS, began in 1993
  - National Technology Roadmap for Semiconductors
  - 1 year after the US formed the NITRD
  - 24-year industry alliance stopped publishing in 2017
- Successor, the IRDS

Semiconductor device fabrication

MOSFET scaling (process nodes)

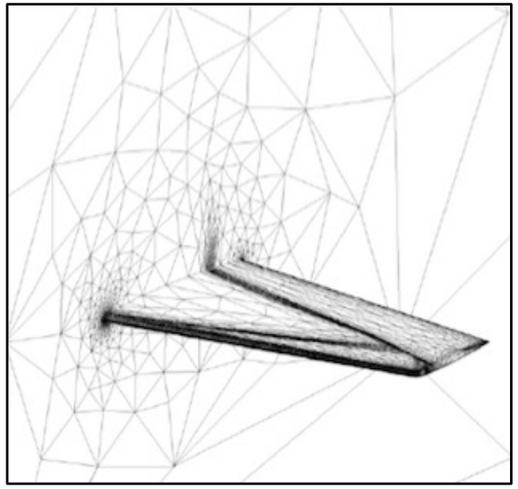
20 µm – 1968
10 µm – 1971
6 µm – 1974
3 µm – 1977
1.5 µm – 1981
1 µm – 1984
800 nm – 1987
600 nm – 1990
350 nm – 1993
250 nm – 1996
180 nm – 1999
130 nm – 2001
90 nm – 2003
65 nm – 2005
45 nm – 2007
32 nm – 2009
28 nm – 2010
22 nm – 2012
14 nm – 2014
10 nm – 2016
7 nm – 2018
5 nm – 2020
3 nm – 2022

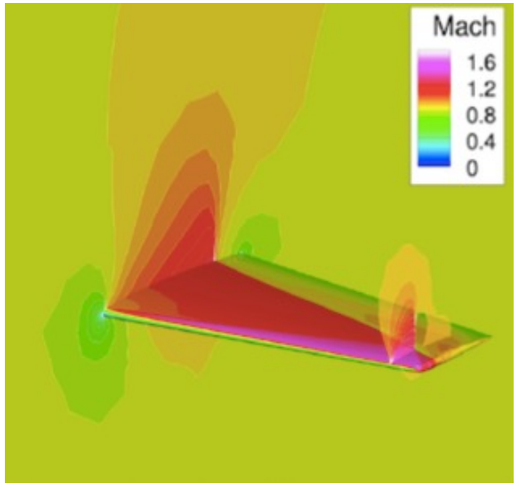
Future

2 nm ~ 2025
1 nm ~ 2027

Discipline-wide benchmark problems

Example: computational aerodynamics





- Standard problems on which anyone can demonstrate new discretization, algorithm, and implementation
  - Common in simulation, e.g., M6 wing
  - Becoming more common in data analytics, e.g., Netflix challenge
  - Now being used in scientific machine learning
- Are standard benchmarks established in statistics?

Industry-wide performance benchmarks

top500.org

top500.org

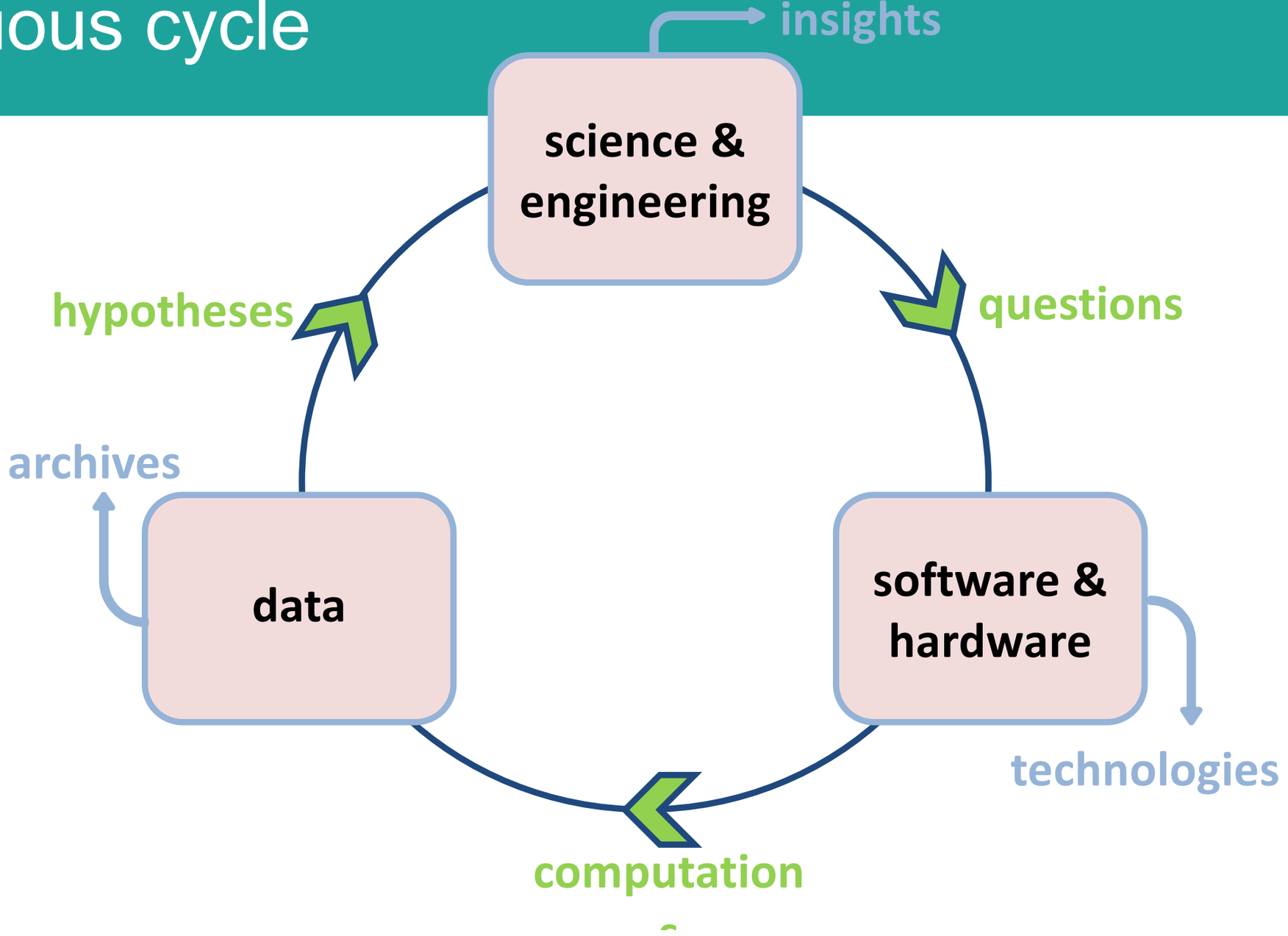
graph500.org

top500.org

benchcouncil.org/hpcai500

hpgmg.org

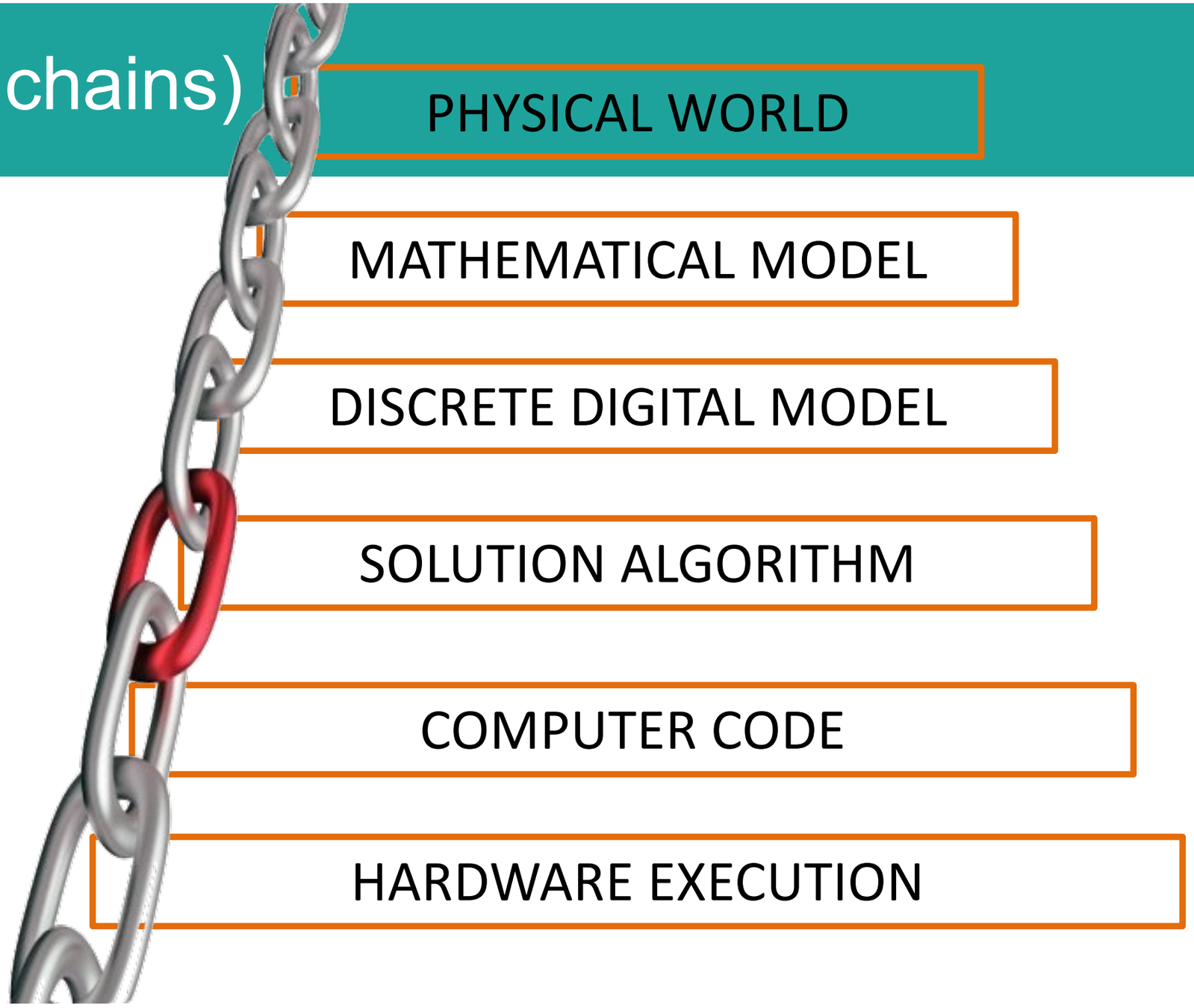
Virtuous cycle



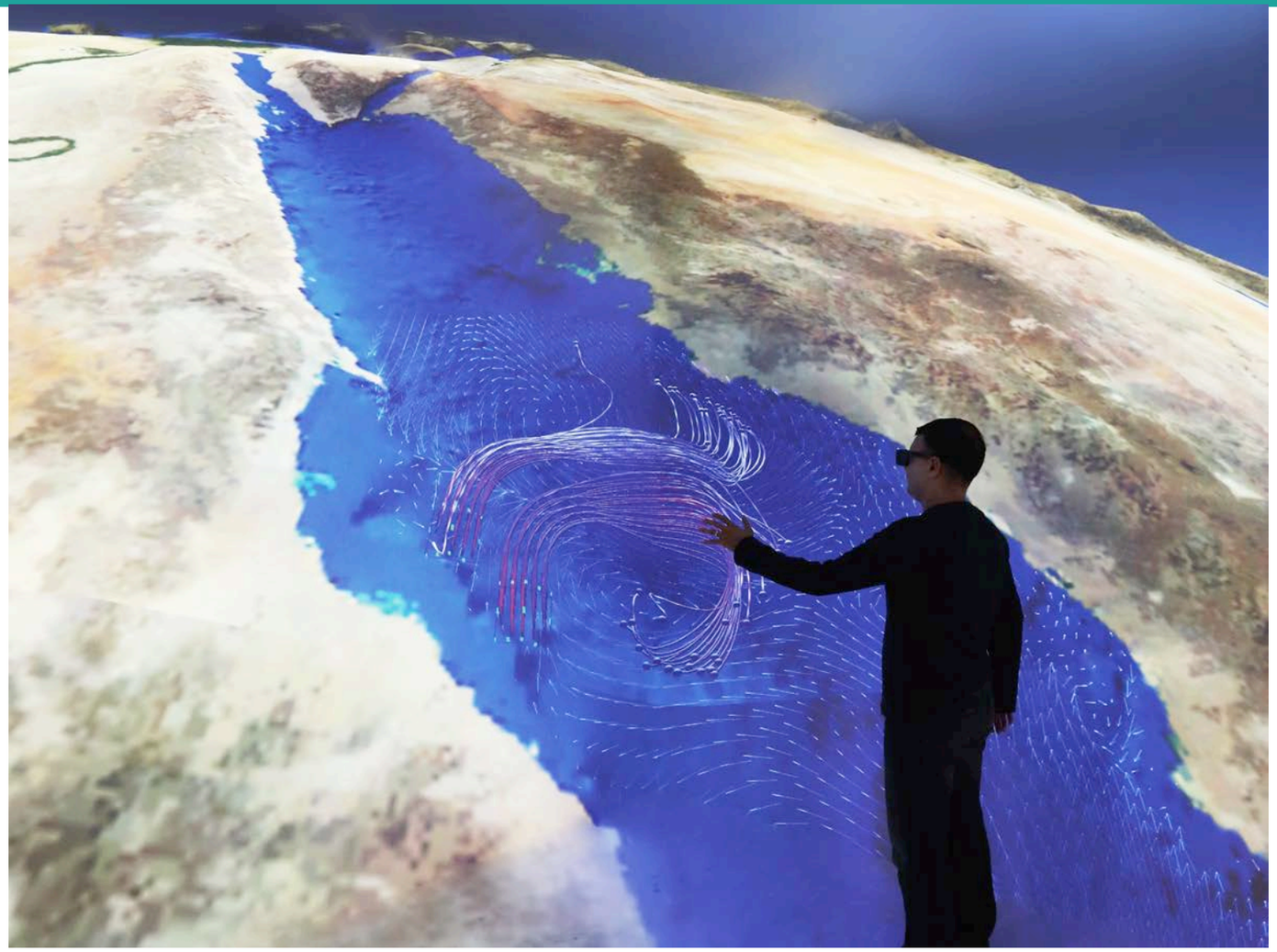
Verticals (tool chains)

Examples:

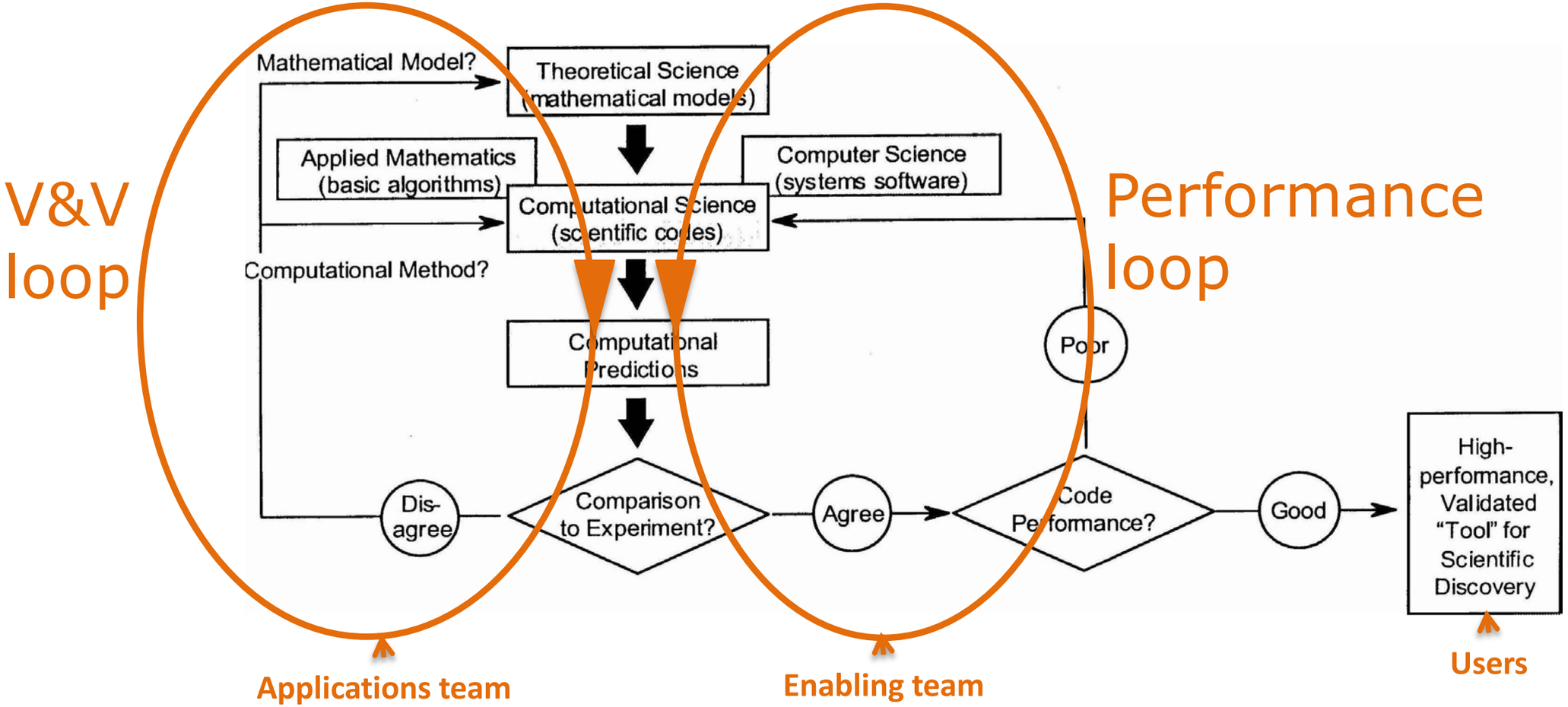
- oil & gas
- finance
- pharma
- aero
- auto
- agriculture
- mining
- wireless
- satellites
- etc.



Real-time immersion (human computer interaction)



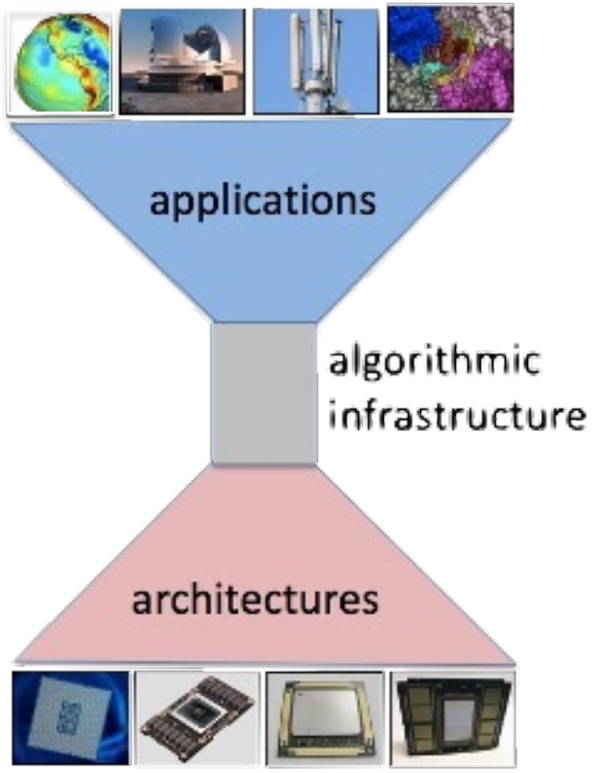
Physical modeling and performance modeling



Credit: T. Dunning, UIUC/NNSA, SciDAC report (2000)

Hourglass and reusability

Originating from the seven-layer internet protocol)



Convergence of the paradigms

	to **Simulation**	to **Analytics**	to **Learning**
Simulation provides	–		
Data Analytics provides		–	
Machine Learning provides			–

Improvements to any one paradigm improve the other two

Convergence of the paradigms

	to **Simulation**	to **Analytics**	to **Learning**
Simulation provides	–		
Data Analytics provides	**Steering in high dimensional parameter space; *In situ* processing**	–	
Machine Learning provides	**Replacement of models with learned functions; Smart data compression**		–

Improvements to any one paradigm improve the other two

Convergence of the paradigms

	to **Simulation**	to **Analytics**	to **Learning**
Simulation provides	–	**Physics-based "regularization"**	
Data Analytics provides	**Steering in high dimensional parameter space; *In situ* processing**	–	
Machine Learning provides	**Replacement of models with learned functions; Smart data compression**	**Detection and classification; Imputation of missing data**	–

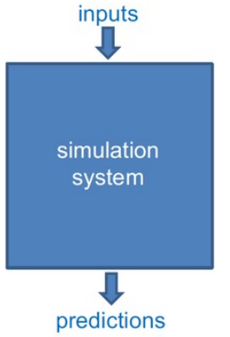
Improvements to any one paradigm improve the other two

Convergence of the paradigms

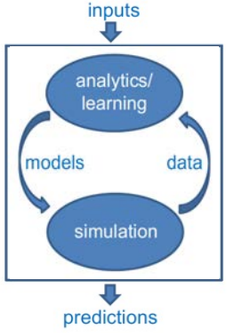
	to **Simulation**	to **Analytics**	to **Learning**
Simulation provides	–	Physics-based "regularization"	**Data for training (augmenting real-world data)**
Data Analytics provides	Steering in high dimensional parameter space; *In situ* processing	–	**Cleaned feature vectors for training**
Machine Learning provides	Replacement of models with learned functions; Smart data compression	Detection and classification; Imputation of missing data	–

Improvements to any one paradigm improve the other two

Traditional simulation systems produce data from models. They do not learn from new data.



Data analytics and machine learning produce models from data. They are part of a virtuous cycle.



Outline of presentation

Examples of High Performance Statistical Computing

- Gordon Bell campaigns of 2022, 2024*, 2025*
- Open source HPSC software

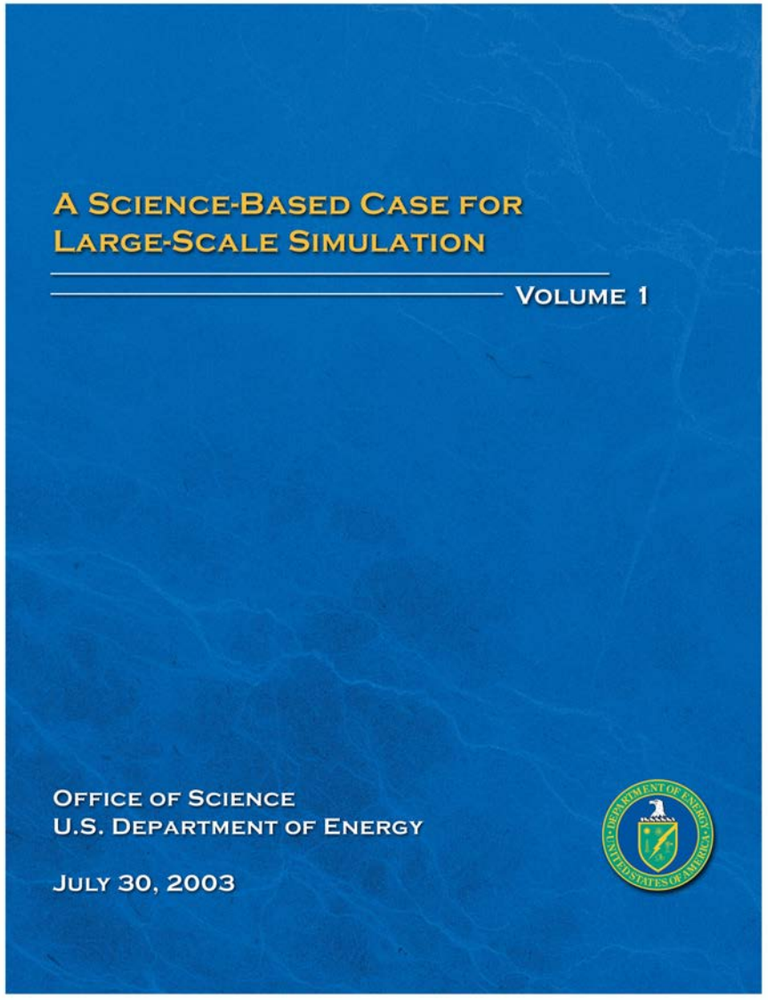
Twelve “universals” of HPC algorithms and software

- Things you *wouldn’t* consider in a proof of concept app
- Things you *must* consider in a high-perfornamce app

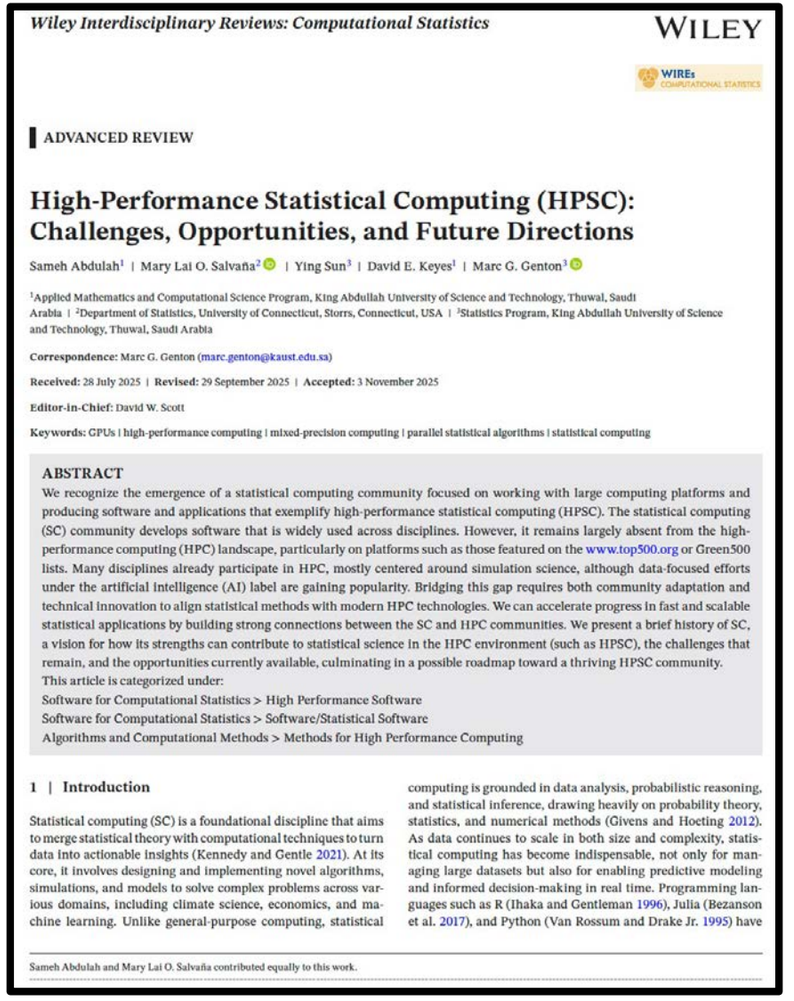
Twelve “elements” of the HPC ecosystem

- No need to start from scratch – HPSC can ride the HPC wave
- Will greatly enrich an *existing* ecosystem

Need a motivational manifesto for stakeholders



- For HPC (2003)
- 2 volumes, 365 pages
- 315 research contributors in mathematics, computer science, and science & engineering applications



Wiley Interdisciplinary Reviews: Computational Statistics WILEY

WIREs COMPUTATIONAL STATISTICS

ADVANCED REVIEW

High-Performance Statistical Computing (HPSC): Challenges, Opportunities, and Future Directions

Sameh Abdulah[1] | Mary Lai O. Salvaña[2] | Ying Sun[3] | David E. Keyes[1] | Marc G. Genton[3]

[1]Applied Mathematics and Computational Science Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia | [2]Department of Statistics, University of Connecticut, Storrs, Connecticut, USA | [3]Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Correspondence: Marc G. Genton (marc.genton@kaust.edu.sa)

Received: 28 July 2025 | **Revised:** 29 September 2025 | **Accepted:** 3 November 2025

Editor-in-Chief: David W. Scott

Keywords: GPUs | high-performance computing | mixed-precision computing | parallel statistical algorithms | statistical computing

ABSTRACT

We recognize the emergence of a statistical computing community focused on working with large computing platforms and producing software and applications that exemplify high-performance statistical computing (HPSC). The statistical computing (SC) community develops software that is widely used across disciplines. However, it remains largely absent from the high-performance computing (HPC) landscape, particularly on platforms such as those featured on the www.top500.org or Green500 lists. Many disciplines already participate in HPC, mostly centered around simulation science, although data-focused efforts under the artificial intelligence (AI) label are gaining popularity. Bridging this gap requires both community adaptation and technical innovation to align statistical methods with modern HPC technologies. We can accelerate progress in fast and scalable statistical applications by building strong connections between the SC and HPC communities. We present a brief history of SC, a vision for how its strengths can contribute to statistical science in the HPC environment (such as HPSC), the challenges that remain, and the opportunities currently available, culminating in a possible roadmap toward a thriving HPSC community.

This article is categorized under:

Software for Computational Statistics > High Performance Software

Software for Computational Statistics > Software/Statistical Software

Algorithms and Computational Methods > Methods for High Performance Computing

1 | Introduction

Statistical computing (SC) is a foundational discipline that aims to merge statistical theory with computational techniques to turn data into actionable insights (Kennedy and Gentle 2021). At its core, it involves designing and implementing novel algorithms, simulations, and models to solve complex problems across various domains, including climate science, economics, and machine learning. Unlike general-purpose computing, statistical computing is grounded in data analysis, probabilistic reasoning, and statistical inference, drawing heavily on probability theory, statistics, and numerical methods (Givens and Hoeting 2012). As data continues to scale in both size and complexity, statistical computing has become indispensable, not only for managing large datasets but also for enabling predictive modeling and informed decision-making in real time. Programming languages such as R (Ihaka and Gentleman 1996), Julia (Bezanson et al. 2017), and Python (Van Rossum and Drake Jr. 1995) have

Sameh Abdulah and Mary Lai O. Salvaña contributed equally to this work.

- For HPSC (2025)
- 25 pages
- ~300 references

HPSC rides the HPC wave

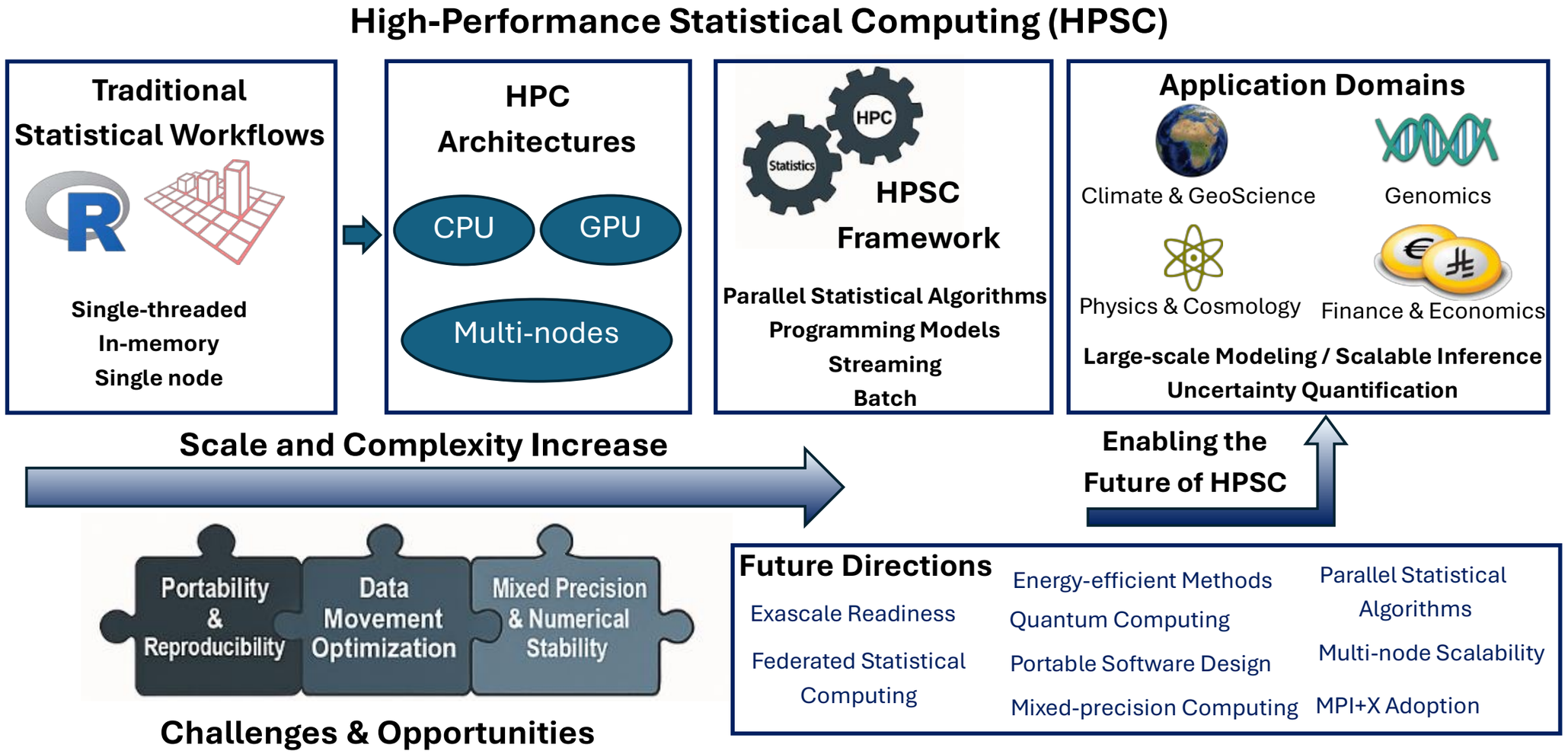


Fig 4 from Abdulah et al. (2025)

Sustainable supercomputing – two meanings

Computing sustainably

- computing no more than necessary for a given science or engineering target

Computing to *support* sustainability, e.g.,

- clean energy
- affordable energy

adopted 2015

Alternative conclusion

"Do linear algebra;
see the world!"

Earl Blossom, 1891-1970

For follow-up

1) *Parallel Approximation of the Maximum Likelihood Estimation for the Prediction of Large-Scale Geostatistics Simulations,* S. Abdulah, H. Ltaief, Y. Sun, M. G. Genton & D. Keyes, 2018, IEEE International Conference on Cluster Computing (CLUSTER), 2018, pp. 98-108, doi: 10.1109/CLUSTER.2018.00089.
2) *Hierarchical Algorithms on Hierarchical Architectures*, D. Keyes, H. Ltaief & G. Turkiyyah, 2020, Philosophical Transactions of the Royal Society, Series A 378:20190055, doi 10.1098/rsta.2019.0055
3) *Responsibly Reckless Matrix Algorithms for HPC Scientific Applications*, H. Ltaief, M. G. Genton, D. Gratadour, D. Keyes & M. Ravasi, 2022, Computing in Science and Engineering, doi 10.1109/MCSE.2022.3215477.
4) *Reshaping Geostatistical Modeling and Prediction for Extreme-Scale Environmental Applications*, Q. Cao, S. Abdulah, R. Alomairy, Y. Pei, P. Nag, G. Bosilca, J. Dongarra, M. G. Genton, D. E. Keyes, H. Ltaief & Y. Sun, 2022, in proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'22), IEEE Computer Society (ACM Gordon Bell Finalist), doi 10.1109/SC41404.2022.00007.
5) *Mixed Precision Algorithms in Numerical Linear Algebra*, 2022, N. J. Higham & T. Mary, Acta Numerica, pp. 347—414, doi:10.1017/S0962492922000022.
6) *Boosting Earth System Model Outputs and Saving PetaBytes in their Storage using Exascale Climate Emulators,* S. Abdulah, A. Baker, G. Bosilca, Q. Cao, S. Castruccio, M. G. Genton, D. E. Keyes, Z. Khalid, H. Ltaief, Y. Song, G. Stenchikov & Y. Sun, 2024, in Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'24), ACM (Gordon Bell Climate Prize).
7) *Real-Time Bayesian Inference at Extreme Scale: A Digital Twin for Tsunami Early Warning Applied to the Cascadia Subduction Zone,* S. Henneking, S. Venkat, V. Dobrev, J. Camier, T. Kolev, M. Fernando, A.-A. Gabriel, O. Ghattas, 2025, in Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'25), ACM (Gordon Bell Prize).


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