For What Should the Bell Toll?

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

Thanks for making it an enriching experience!

Make it a referenceable resource, speakers. Please post your slides after any post-presentation revisions. (Five sets of slides are already at the workshop website.)

Please remember to respond to the ICERM survey

Aspirations

1) **to produce at least one article for the next ICERM quarterly that will have headline appeal to nonspecialists and convince the NSF of its wisdom in creating and sustaining the math institutes and ICERM in particular**
2) **to lead to papers by participants that incorporate ideas first heard or first understood or appreciated here**
3) **to lead to new collaborations that form this weekend by participants who have not previously worked together**

A follow-up to my ICERM talk of 8 September 2024

ICERM   Home   Programs

Modeling and Simulations in Fluids

Sep 7 - 8, 2024

Marking the 90th birthday
of Jim Glimm

- “Efficient computation through tuned approximation”
  - based on four Gordon Bell finalist nominations in three years
- Written up with Halle Bryant for ICERM quarterly newsletter
  - kicker: “Do linear algebra; see the world”
- Caught the attention of NVIDIA’s Tom Gibbs
  - invited for GPU Technology Conference 2025: “For what the Bell tolls”

CLIMATE CONSCIOUS COMPUTING: EXPLOITING DATA SPARSITY FOR MODEL PERFORMANCE AND SUSTAINABILITY

by David Keyes

A climate emulator exploiting data sparsity to produce unprecedented resolution in space and time – and minimize its own carbon footprint – was one of many applications featured at Modeling and Simulations in Fluids, a September Hot Topics Workshop. On behalf of a research team of twelve climate modelers, mathematicians, statisticians, and computer scientists, Professor David Keyes shared the method behind the model, which was recently nominated for the Association of Computer Machinery's (ACM) 2024 Gordon Bell Prize in Climate Modelling. Since 2022, Keyes has been a member of four research teams that have been finalists for the ACM Bell Prize and the recently instituted Bell Climate Prize. The prizes are awarded annually at the International Conference for High Performance Computing, Networking, Storage, and Analysis.

From 2011-2020, Keyes served as a member of ICERM's Board of Trustees. Currently, he is a Professor of Applied Mathematics and Computational Science and Director of the Extreme Computing Research Center at the King Abdullah University of Science and Technology (KAUST) in Saudi Arabia.

The Gordon Bell Prize bears the name of the founding head of the U.S. National Science Foundation's Computing and Information Science and Engineering (CISE) Directorate. A renowned computing systems engineer himself, Bell endowed the award in 1988 to challenge the high-performance computing community to deliver performance on real-world applications and reach beyond previous benchmarks concerned with supercomputer hardware design. Over the past two years, our team has competed for the prize with four different applications that promote the same "renaissance" in computational linear algebra, motivated by the need to compute with greater energy efficiency. Current exascale supercomputers draw upwards of 20 MegaWatts in continuous operation, suggesting an annual carbon footprint comparable to that of 20,000 average cars in the USA.

The highly scalable climate emulator described at ICERM offers faster, high-resolution simulations without the need for massive data storage. Instead of storing many simulations, it stores their statistical parameters, allowing it to generate an arbitrary number of emulations with the same statistics as the simulations. In the words of my frequent collaborator Marc Genton (KAUST), "PDE-based climate models are complex and can take weeks or months to run, even on the fastest supercomputers. They generate massive amounts of data that become nearly impossible to store, and this is increasingly a problem as climate scientists are constantly pushing for higher resolution." Genton has been designing the emulator's algorithm for nearly a decade, steering the overall reduction of computational time and storage requirements.

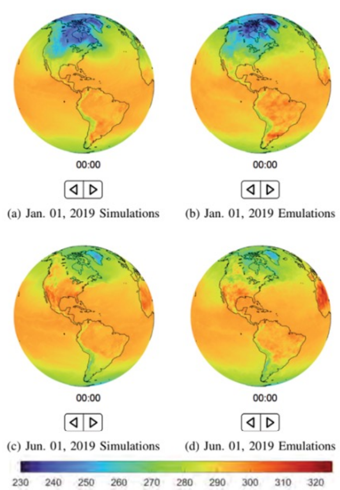


Photo courtesy of David Keyes. Figures show consistency between climate simulations and the less "expensive" climate emulator designed by Keyes and his collaborators.

During our process, our team ran our emulator on four of the world's top ten supercomputers: Frontier and Summit at Oak Ridge National Laboratory, ALPS at the Swiss National Computing Centre (CSCS), and Leonardo at the Italian CINECA. These runs demonstrated portability over four GPU processors: AMD's MI250X, and NVIDIA's V100, A100, and H100. By running on 9,025 nodes of Frontier, the climate emulator offers a remarkable resolution of 3.5 kilometers and can replicate local conditions on a timescale from days down to hours. The emulator uses a fast spherical harmonic transform (SHT) method that converts elements such as temperature, wind, and pressure into simple frequency amplitudes that describe how they change over time in more than 54 million locations around the globe. We have found that the keys to our performance accomplishments and to bringing down the carbon footprint of numerous applications are exploiting data sparsity in algebraic rank and floating point precision in fine-grained blocks of matrix operators. As I said during my ICERM talk, "Do Linear Algebra; see the world!" Or, better yet, "Do Linear Algebra; see many worlds."

Participants in ICERM's September 2024 Modeling and Simulations in Fluids Workshop

A vast number of computationally demanding applications are amenable to the use of our team's Hierarchical Computations on Manycore Architectures (HiCMA) software package, which has powered our climate emulator and our three additional Bell Prize finalist nominations. Specifically, HiCMA can be utilized for problems expressable as integral equations with smooth kernels or fractional differential operators, covariances in spatial distributions, Schur complements of elliptic discretizations, Hessians of PDE-constrained optimization, radial basis function discretizations, and many types of kernel matrices from Gaussian and other processes. Applications range from climate and weather modeling to seismic imaging, from wireless signal decoding to atmospheric turbulence mitigation in telescopy, to genome-wide association studies in adapting agriculture to climate stress. Indeed, HiCMA has been applied in all of these domains at KAUST.

I am encouraged by the supercomputing research community's growing interest in computations that are designed to reduce, not increase, floating point operations per second (flop/s), the rate by which methods have traditionally been evaluated. By strategically squeezing out operations unnecessary for obtaining a sufficiently accurate result, our team appeals to time to solution, reduction of memory footprint, and other metrics of merit that may be more appropriate going forward in supercomputing. Indeed, my talk emphasized using mathematics and computer architecture considerations to get the job done while minimizing computational requirements.

I believe that the late Gordon Bell would approve of the selection committee's decision to give podium attention to our data sparsity campaign at the last three ACM Supercomputing conferences. Indeed, he adapted the topics of "special prizes" several times in the early years as parallel computing evolved. There are traditionally six Gordon Bell Prize finalists per year from around the globe and three Climate Prize finalists. Our team from KAUST are the first finalists from the Middle East in the prize's thirty-seven year history. Perhaps this year will be the year! ■

ACM Gordon Bell Prize for Climate Modelling

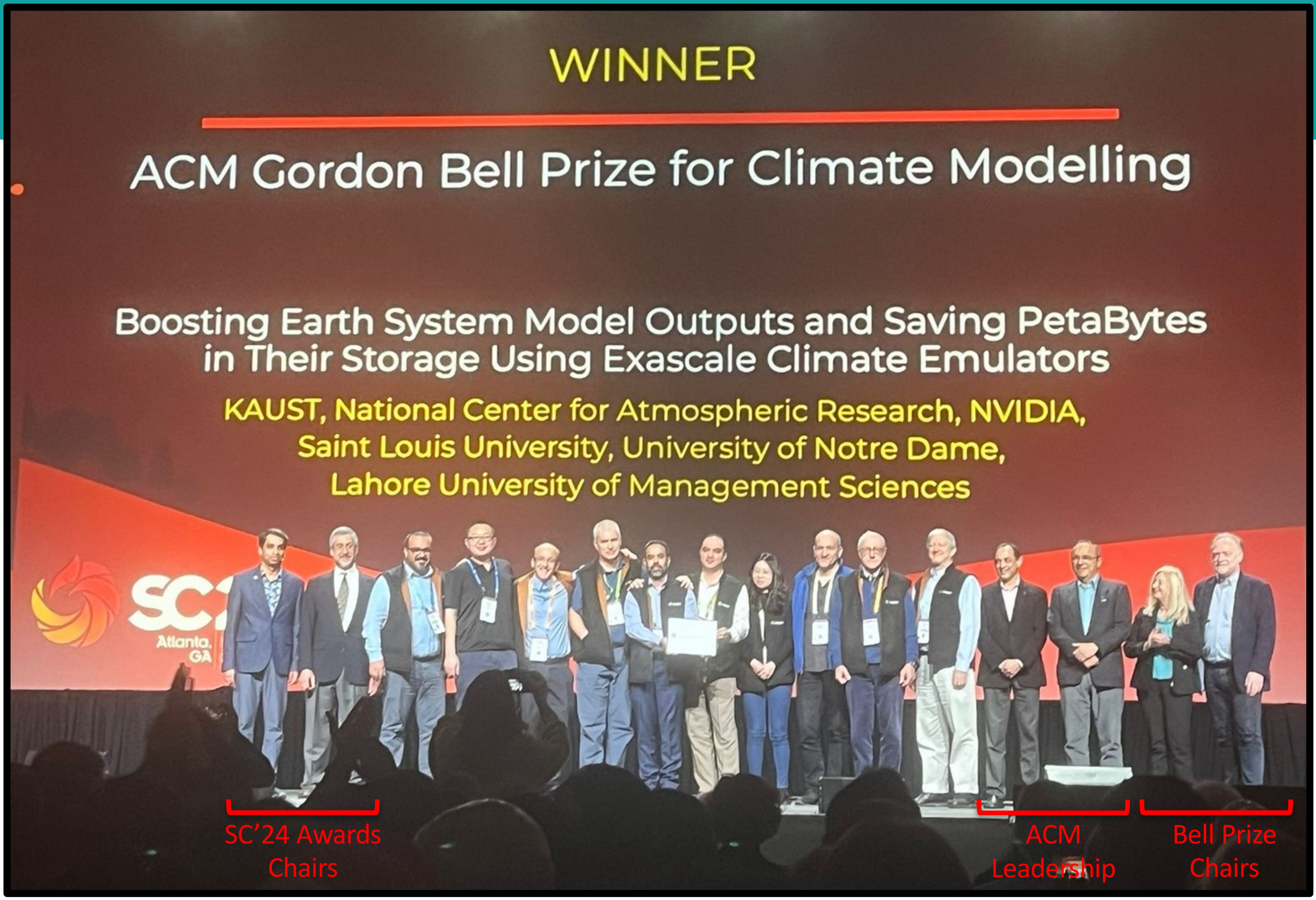
Innovations in applying high-performance computing to climate modelling applications

Award Recipients | Nominations | Committee Members

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


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
SC'24 Awards Chairs
ACM Leadership
Bell Prize Chairs

Nov 2024

GB Prize winners in the room this weekend

- Adams (2004)
- Gropp (1999)
- Hoefler (2019)
- Keyes (1999, 2024)
- Schulthess (2008, 2009)
- Smith (1999)
- Vuduc (2010)

Let me know if I missed anyone!

"The ACM Gordon Bell Prize for peak performance recognizes a high-performance computing (HPC) application that achieves outstanding speed and efficiency, **measured in floating-point operations per second (FLOPS)**, on a significant scientific or engineering problem."

This has been an inspiring metric for 37 years. Is this the most appropriate metric going forward?

Running on Frontier versus flying commercially

- Carbon footprint of a KiloWatt-hour is 0.5 kg CO2-equivalent
  - 10,000 kg CO2e hourly carbon footprint for a 20MW exaflop/s system (10 metric tons)
- Carbon footprint of one passenger-hour of commercial cruise Mach flight is about 0.25 metric tons CO2e
  - 1 hour of exaflop/s is roughly equivalent to 40 passenger-hours of flight

Flying these 39 passengers costs about the same *per hour* as Frontier

Carbon offset your next flight by efficient programming!

Better yet, please justify my flight here to give this talk ☺

Questions about the future of the GB Prize

- What should it measure?
- If operations per second, what kind of operations in the multi-precision GPU era?
- Given its potential for focusing attention, by what better metrics might modeling and simulation benefit?
- Can such benefit be extended to machine learning?
- How will it adapt to the hybrid classical-quantum era?
  - "flop equivalents"

Objectives for the next ~40 minutes

- Provide a quick update to last year's talk
- Present perspectives from first 37 years of the GB Prize
- Present Bill Gropp's involvement in a GB campaign (1999)
    - alluded to by Rich Vuduc yesterday
- Promote discussion of the curation of the prize among several of its sharers and answer:

"For what *should* the Bell toll?"

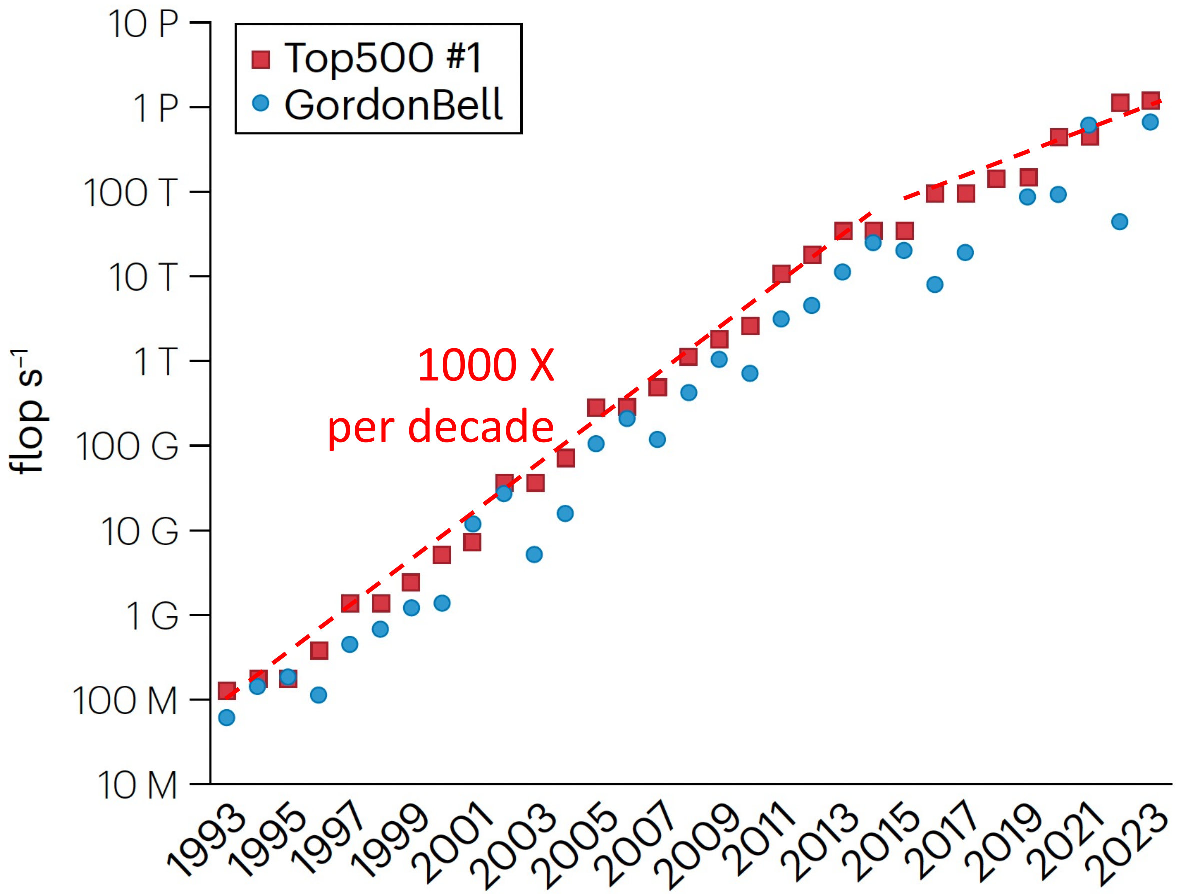
For what *has* the Bell tolled?

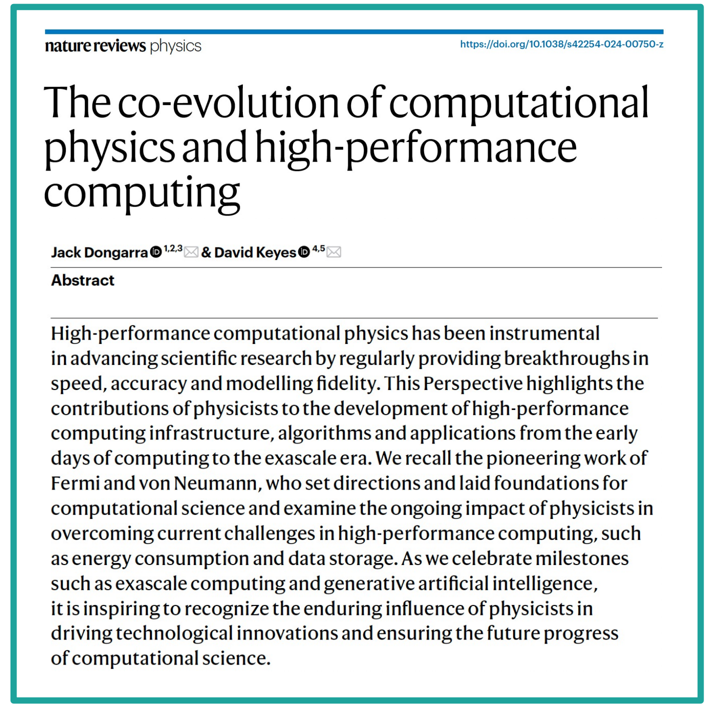
- *Scaling ("Karp Prize", 1987)*
- **Peak performance (1988)**
- Price/performance (1988)
- Compiler parallelization (1989)
- Speedup (1992)
- Special purpose machine (1995)
- Parallel language (2002)
- Lifetime achievement (2003)
- Algorithmic innovation (2008)
- Time to solution (2011)
- Scalability (2015)
- **+COVID modeling(2020-2022)**
- **+Climate modeling (2023-2032)**

GB Prize committees and Gordon, himself, tinkered with the prize over the years…

It is not irreverent of us to be creative with a metric that will induce positive behaviors for the future

Tracking of Top500 by GB Peak Prize winners




nature reviews physics
https://doi.org/10.1038/s42254-024-00750-z

The co-evolution of computational physics and high-performance computing

Jack Dongarra[1,2,3] & David Keyes[4,5]

Abstract

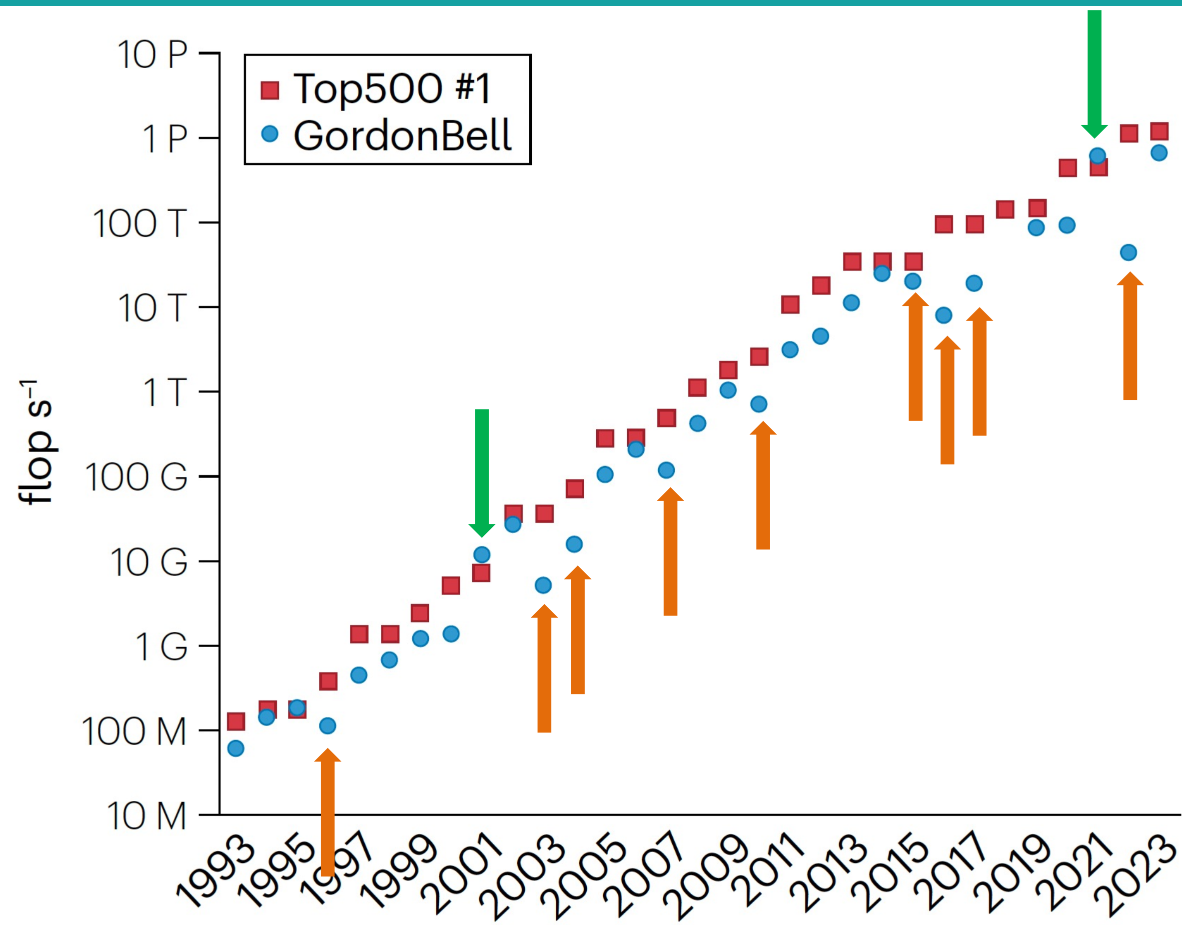
High-performance computational physics has been instrumental in advancing scientific research by regularly providing breakthroughs in speed, accuracy and modelling fidelity. This Perspective highlights the contributions of physicists to the development of high-performance computing infrastructure, algorithms and applications from the early days of computing to the exascale era. We recall the pioneering work of Fermi and von Neumann, who set directions and laid foundations for computational science and examine the ongoing impact of physicists in overcoming current challenges in high-performance computing, such as energy consumption and data storage. As we celebrate milestones such as exascale computing and generative artificial intelligence, it is inspiring to recognize the enduring influence of physicists in driving technological innovations and ensuring the future progress of computational science.

32 of 37 GB Peak Prize winners are from physics

Field	Year
Cosmology	1992, 1996[a], 1997, 2012
Fluid dynamics	1993, 1996[a], 1999, 2007, 2010, 2013
Quantum chromodynamics	1995
Condensed matter physics	1998, 2005, 2006, 2008, 2009, 2011, 2019, 2023
Molecular dynamics	2000[a], 2014, 2020
Astrophysics	2000[a], 2001
Atmospheric dynamics	2002, 2016
Solid earth geophysics	2003, 2004, 2015, 2017
Quantum circuit	2021
Laser physics	2022

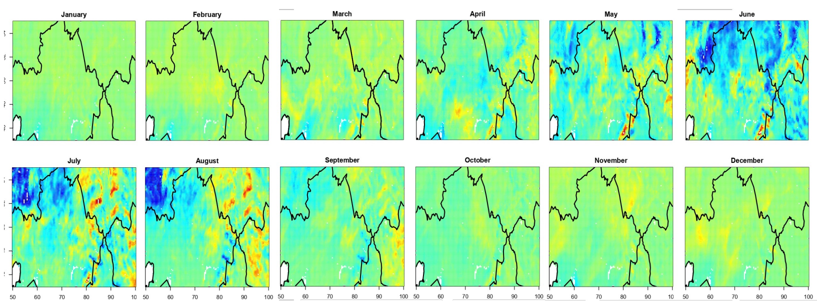
[a]Two teams tied for first place in 1996 and 2000.

Tracking of Top500 by GB Peak Prize winners

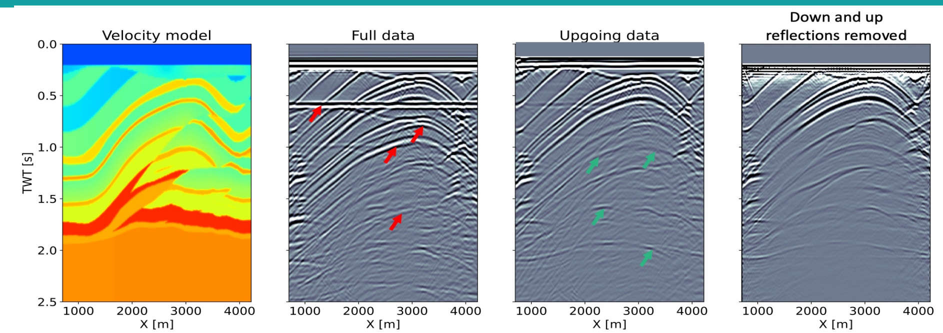


- Top500 #1 is non-decreasing
- GB Peak Performance has 9 times failed to exceed a GB precedent
- GB Peak Performance twice exceeded the Top500 HPL on a specialty machine
  - GRAPE
  - Anton
- Ratio of GB to Top500 #1 ranges from 9% to 160%, with a median of about 50%

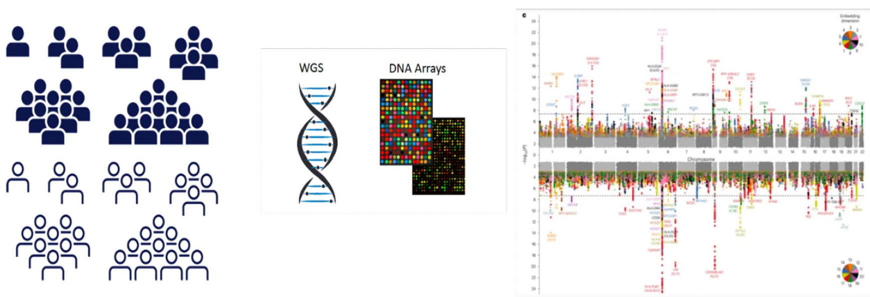
Our journey covers four Gordon Bell Prize finalist papers in the past three years



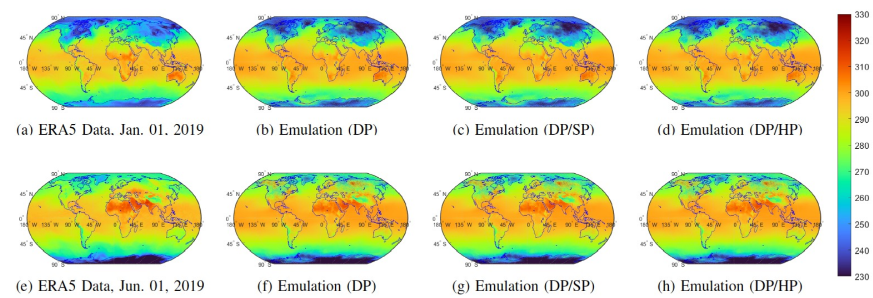
2022: Geospatial statistics



2023: Seismic processing

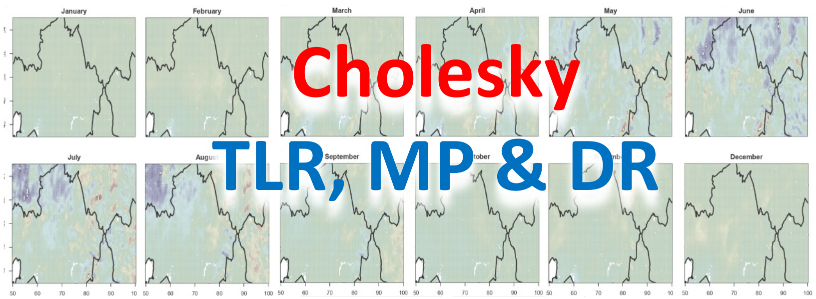


2024: Genomic association

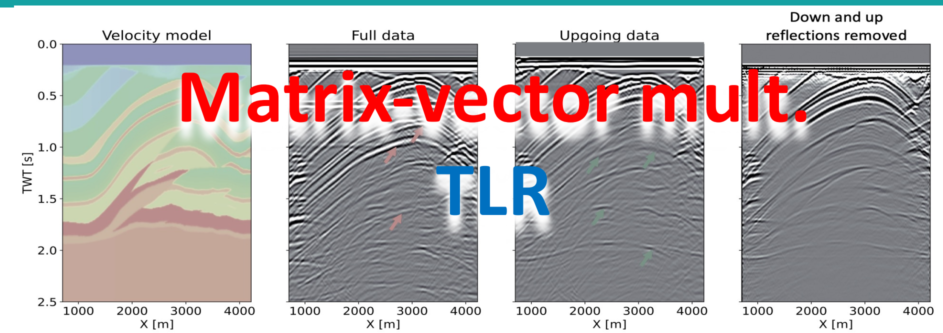


2024: Climate emulation

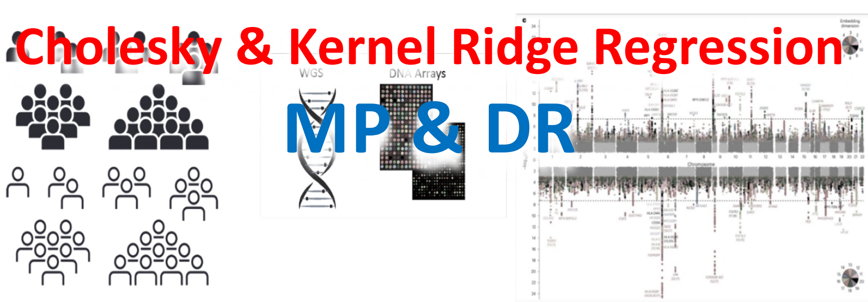
Algorithm: adaptive low rank and low precision substitutions for (default) dense double



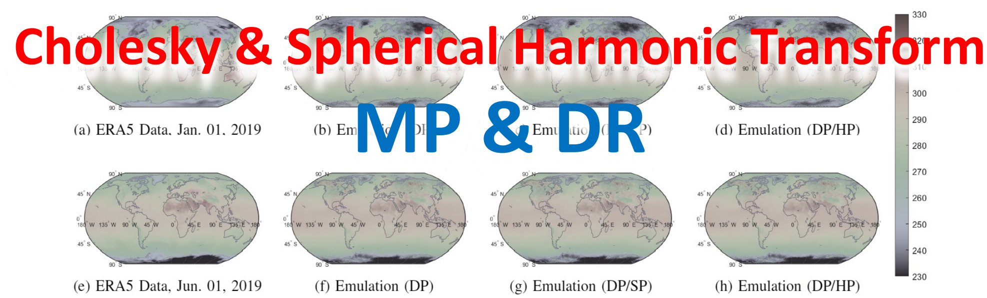
2022: Geospatial statistics



2023: Seismic processing



2024: Genomic association



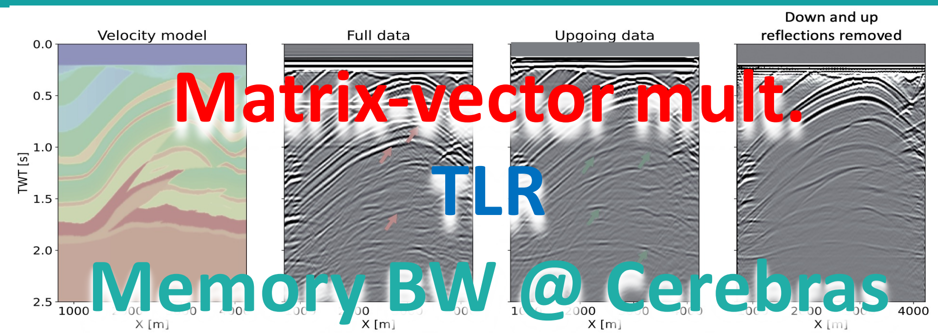
2024: Climate emulation

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

Gordon Bell finalist “merits” and machines

Cholesky
TLR, MP & DR
Time to solution @ Fugaku

2022: Geospatial statistics

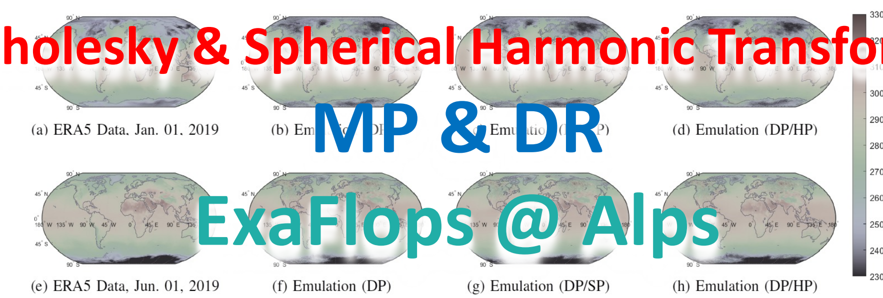


Matrix-vector mult.
TLR
Memory BW @ Cerebras

2023: Seismic processing

Cholesky & Kernel Ridge Regression
MP & DR
ExaOps @ Alps

2024: Genomic association



Cholesky & Spherical Harmonic Transform
MP & DR
ExaFlops @ Alps

2024: Climate emulation

One goal throughout:

Computational efficiency through tuned approximation

- not necessarily highest performance (ops per sec), but lowest energy to solution
- 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

A folk definition of insanity

“Insanity is doing the same thing over and over and expecting a different result.”

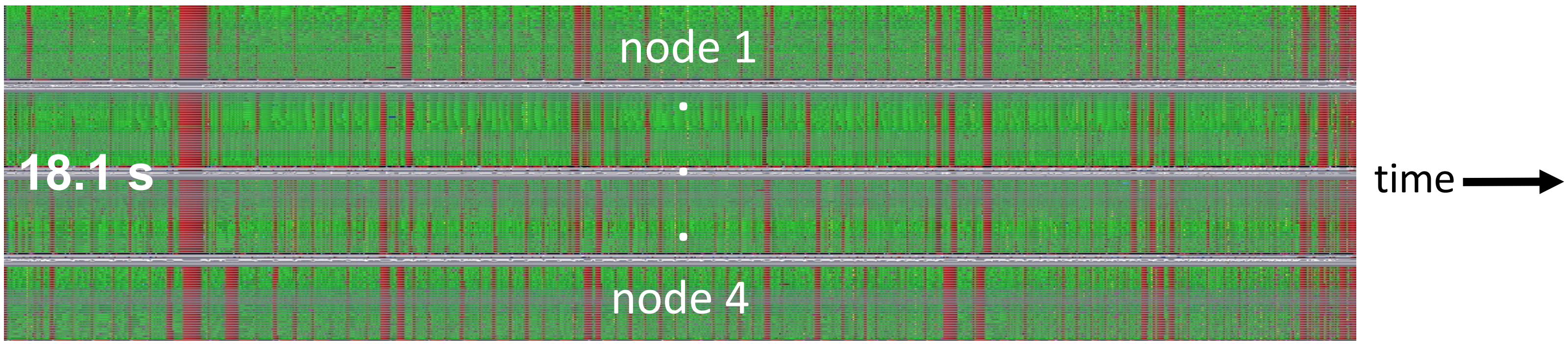
- commonly misattributed to Einstein, actually from the 1980s

Our “insane”, contrarian behavior eventually paid off ☺.

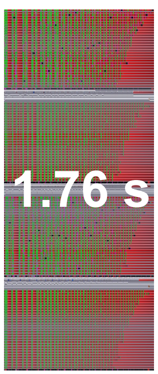
Our journey in tuned approximation began in 2018 with these time traces for tile low-rank (TLR) Cholesky

… 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

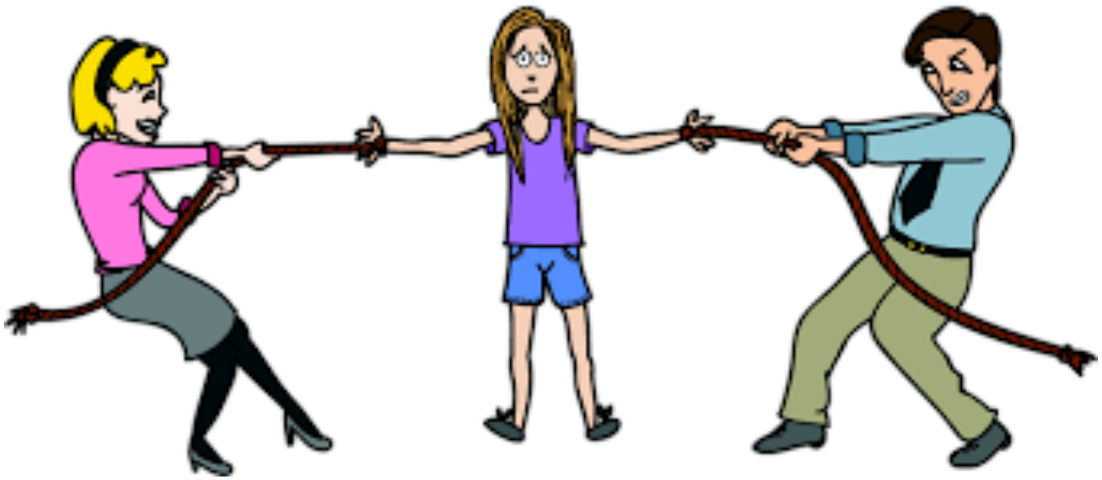
Akbudak, Ltaief, Mikhalev, Charara & K., *Exploiting Data Sparsity for Large-scale Matrix Computations,* Euro-Par 2018

Algorithmic philosophy

Algorithms must span a widening gulf …

adaptive algorithms

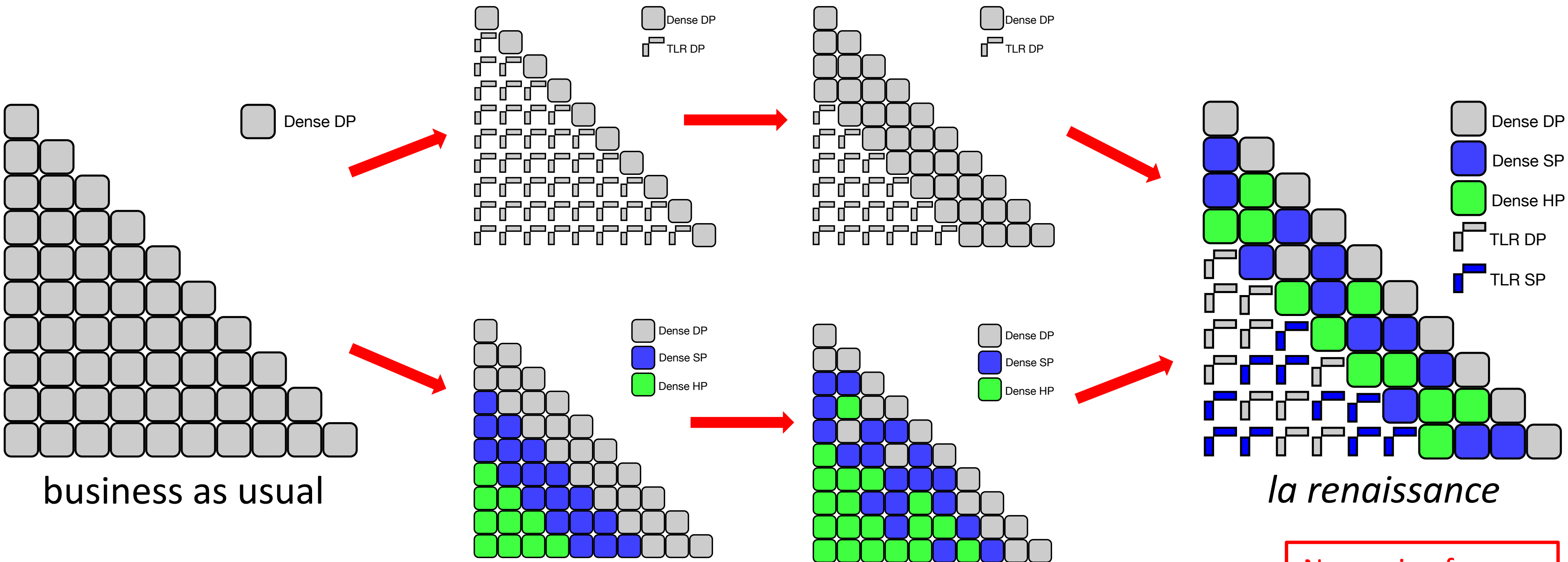
ambitious applications
increasingly dynamic and unstructured



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*



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

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

Economize on storage: no extra copies of the original matrix

Linear algebraic “secret sauce”

Where possible, without losing working accuracy:

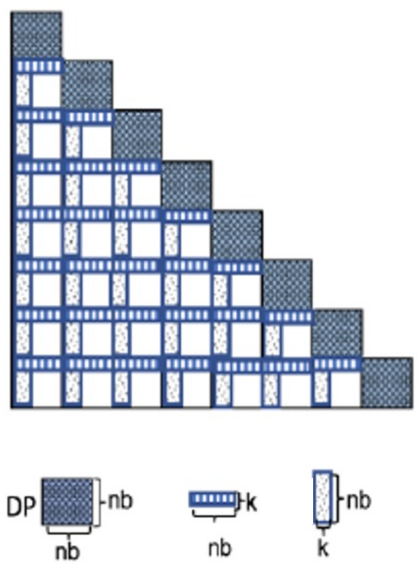
- Replace default 64-bit IEEE standard double precision operations with lower precisions
  - Save storage
  - Save data motion
  - Exploit special-purpose hardware optimized for low precision
- Replace default full rank blocks of discrete *linear operators* and/or discrete *field data* with lower rank blocks
  - 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

- **integral equations** with smooth Green's functions
- **covariances** in statistics
- **Schur complements** within discretizations of PDEs
- **Hessians** from PDE-constrained optimization
- **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)

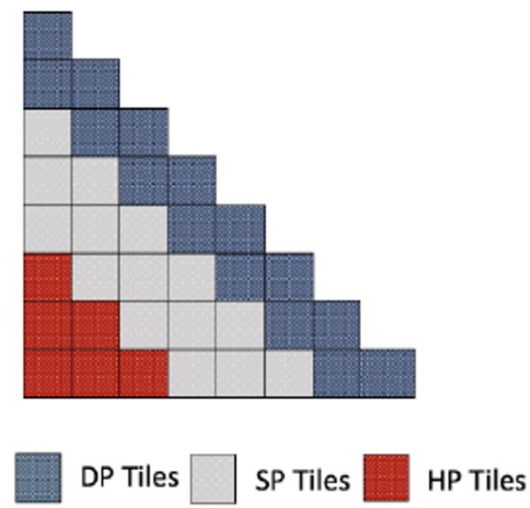


Renaissance in numerical linear algebra (2): precision

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-)



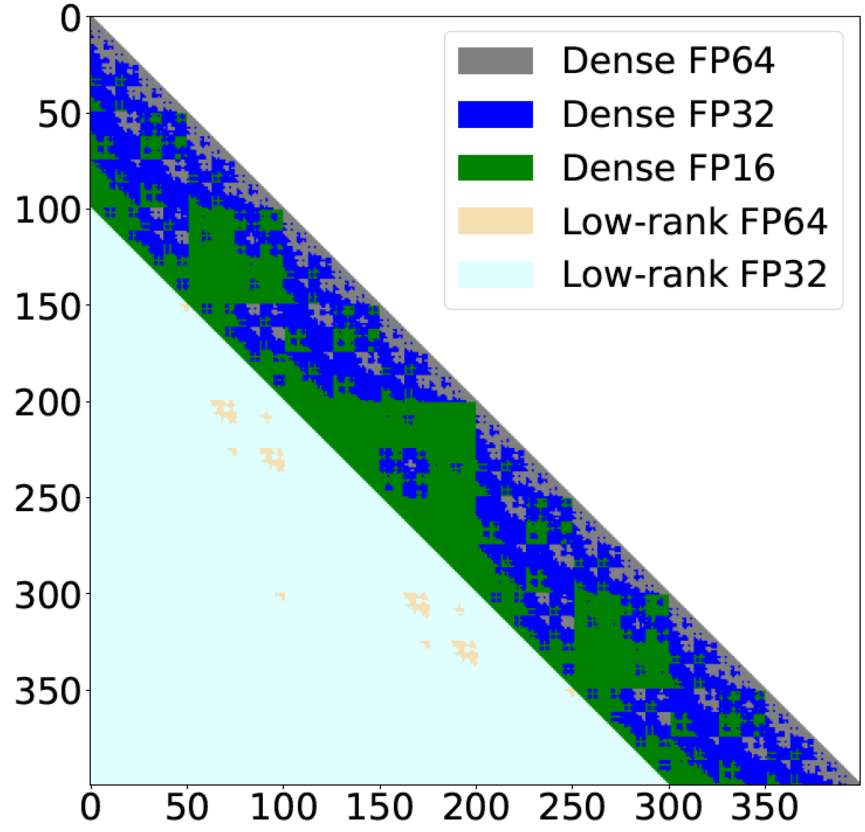
Renaissance in numerical linear algebra (3): combined

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



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

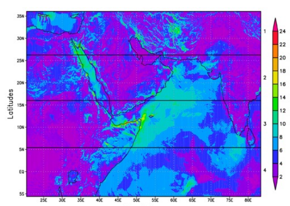
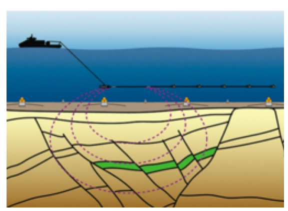
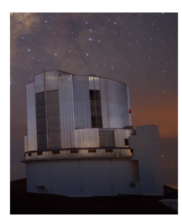
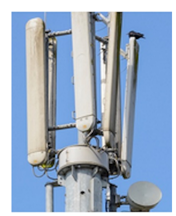
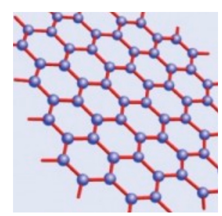
- *standard SVD*: $O(n^3)$, too expensive, especially for repeated compressions after additive tile manipulations
- *randomized SVD* (Halko *et al*., 2011): $O(n^2 \log k)$ for rank k, requires only a small number of passes over the data, saving over the SVD in memory accesses as well as operations
- *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*

$$\mathrm{Min}_U \; \| A - C\,U\,R \|$$

Application opportunities

With such new algorithms, today's HPC can extend many applications that possess

- memory capacity constraints (e.g., geospatial statistics, PDE-constrained optimization)
- power constraints (e.g., remote telescopes)
- real-time constraints (e.g., wireless communication)
- running time constraints (e.g., chemistry, materials, genome-wide associations)


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

Covariance functions $\Sigma(\boldsymbol{\theta})$ supported in ExaGeoStat

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

Univariate Matern Kernel

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

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

Space/Time Nonseparable Kernel

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

(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

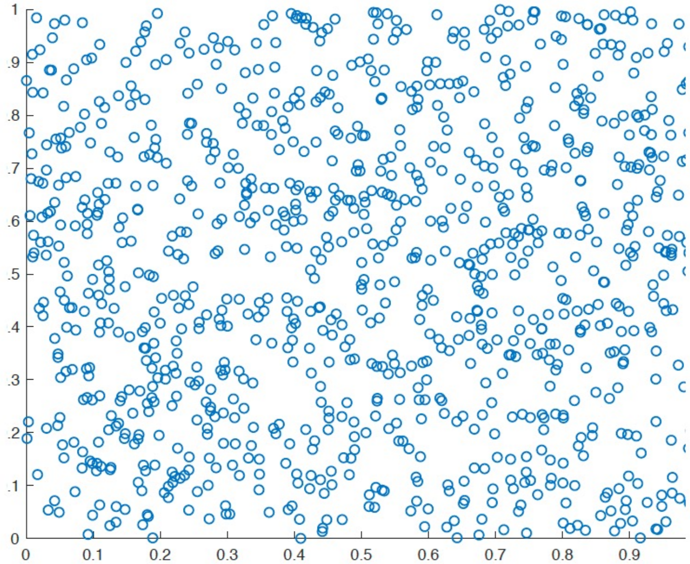
$$C(\mathbf{h};u) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)(a|u|^{2\alpha}+1)^{\delta+\beta d/2}}\left(\frac{c\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\beta/2}}\right)^{\nu} \times K_\nu\left(\frac{c\|\mathbf{h}\|}{(a|u|^{2\alpha}+1)^{\beta/2}}\right), \quad (\mathbf{h};u)\in\mathbb{R}^d\times\mathbb{R},$$

Powered Exponential Kernel

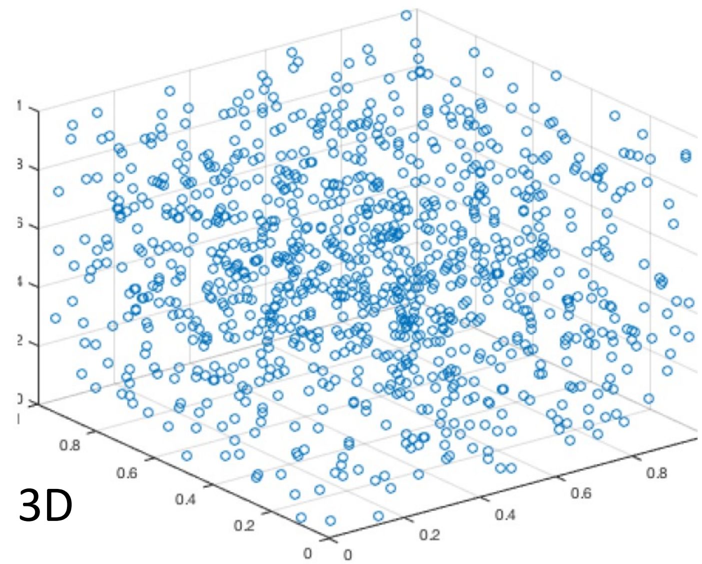
$$C(r;\boldsymbol{\theta}) = \theta_0\exp\left(\frac{-r^{\theta_2}}{\theta_1}\right)$$

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, $a_{ij} \sim \exp(-c|x_i - x_j|^2)$



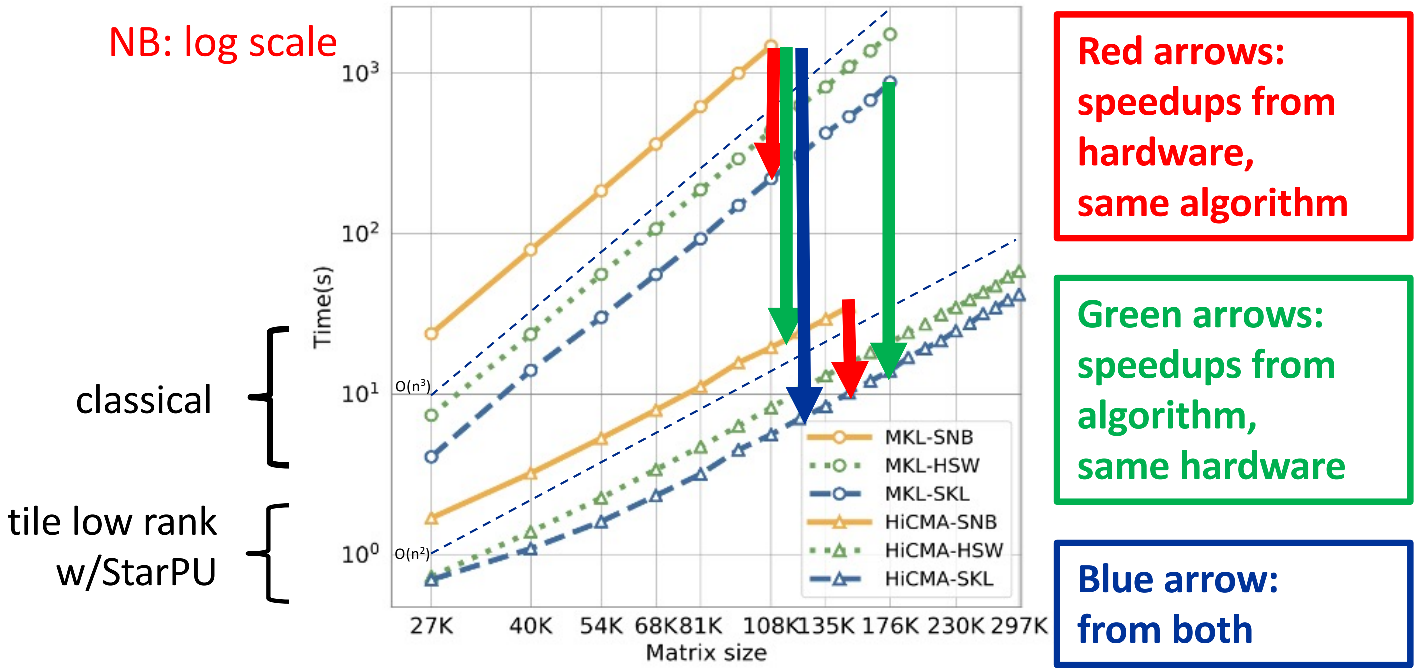
2D



3D

HiCMA TLR vs. Intel MKL on shared memory

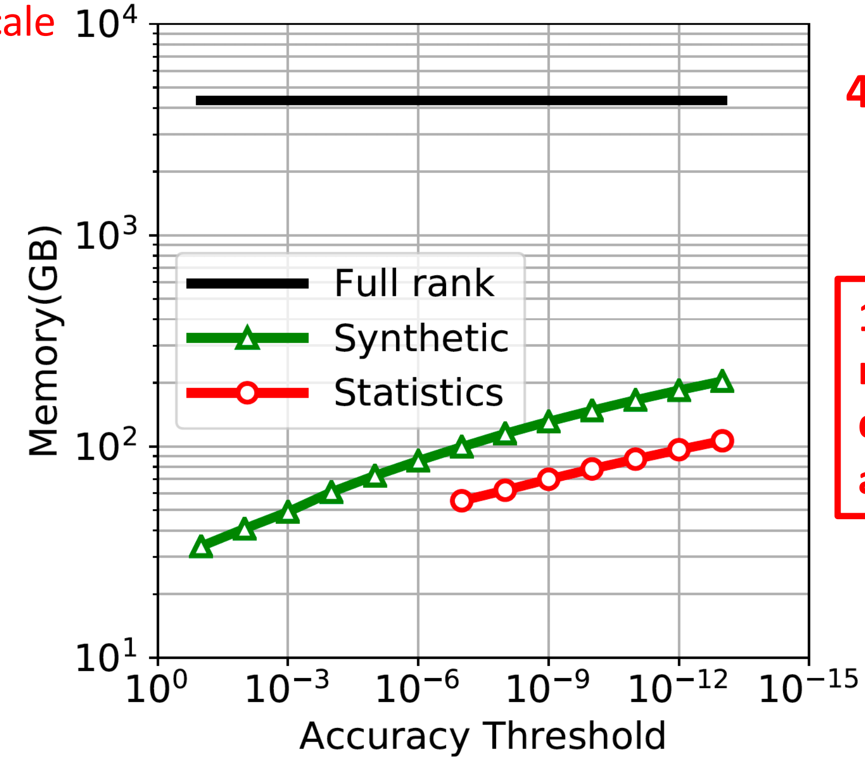
- Gaussian kernel to accuracy 1.0e-8 in each tile
- Three generations of Intel manycore (Sandy Bridge, Haswell, Skylake)
- Two generations of linear algebra (classical dense and tile low rank)



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



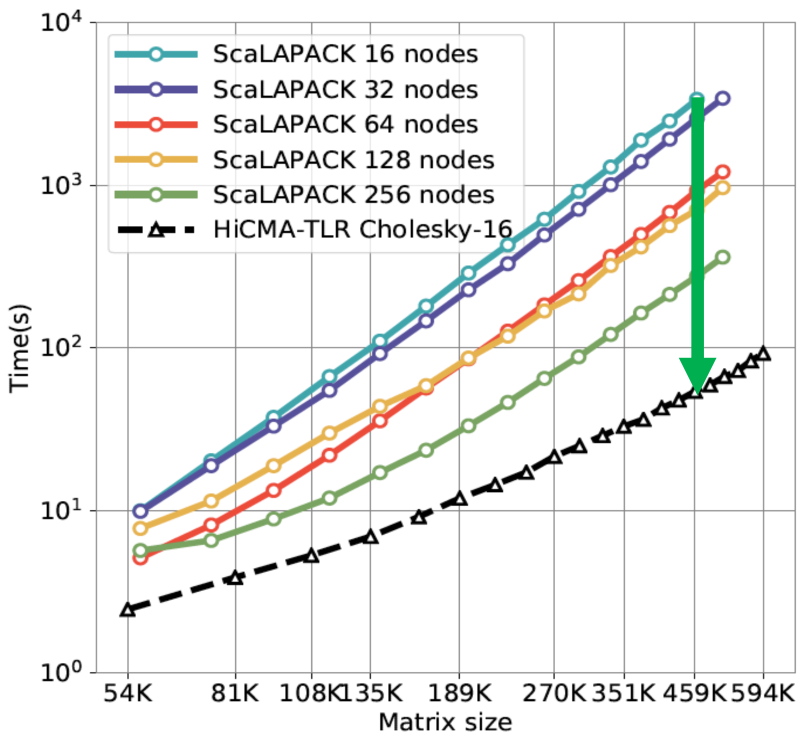
4 TB

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 0.6X	90
FP32		19.5	6[illegible]	[illegible]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

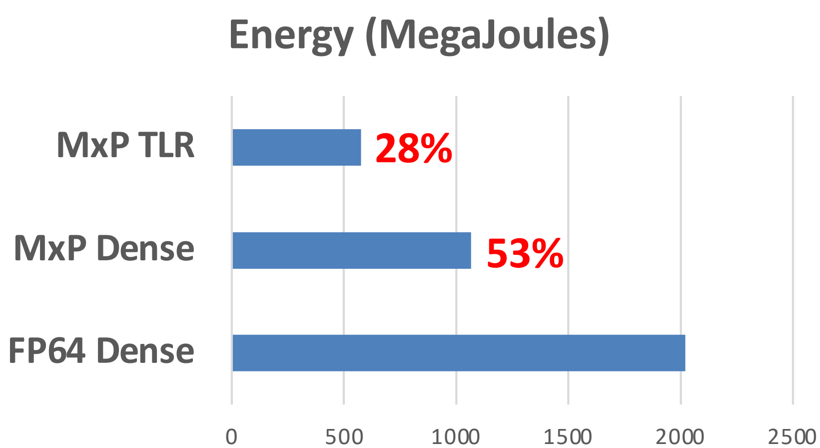
- Mathematical: (much) better than “no precision”
  - statisticians often approximate remote diagonals as *zero* after performing a diagonally clustered space-filling curve ordering (no error bounds available)
- Computational: faster time to solution
  - hence lower energy consumption and higher performance

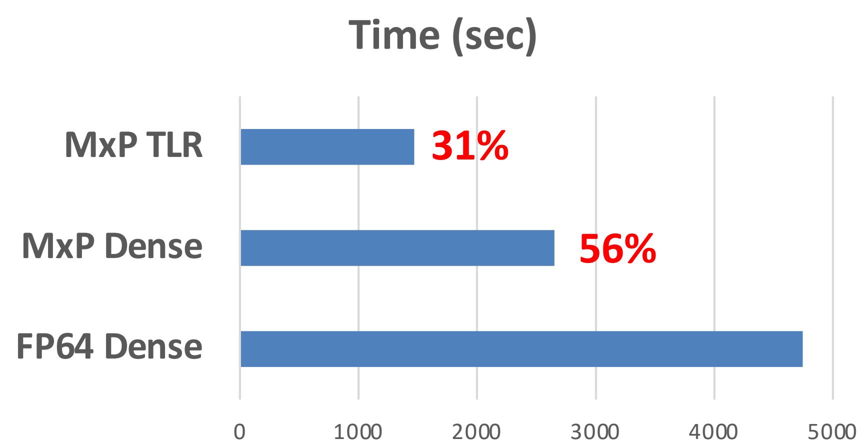
Peak Performance in TF/s	V100 NVLink	A100 NVLink	H100 SXM	B200
FP64	7.5	9.7	34	90
FP32		19.5	67	180
FP64 Tensor Core	15	19.5	67	40
FP/TF32 Tensor Core	8x	156	495	1125
FP16 Tensor Core	120	312 32x	990 30x	2250 225x
FP8/INT8 Tensor Core	-	624	1980	4500
FP4 Tensor Core	-	-	-	9000

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

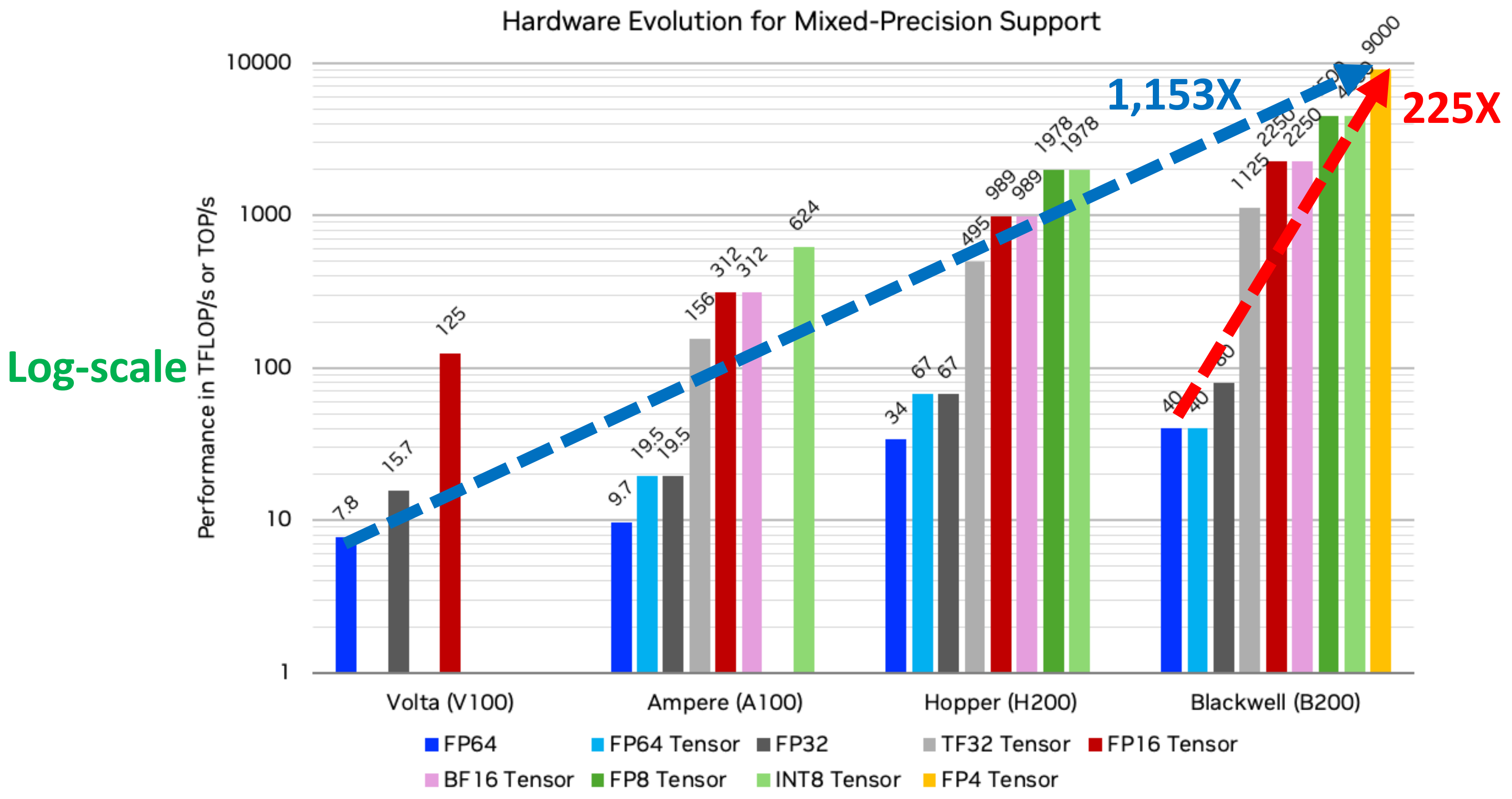
On 512 nodes of Shaheen III





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

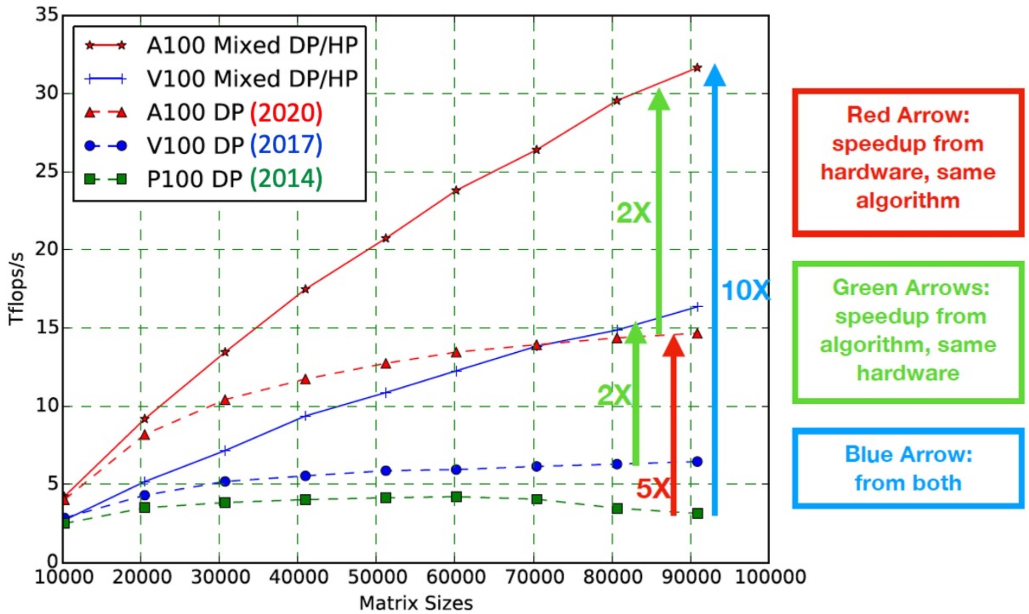
Peak performance of four generations of NVIDIA GPUs



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

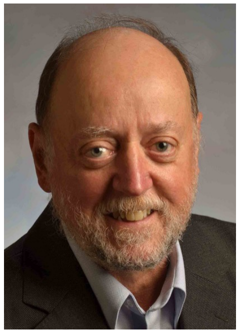
GB’22 collaborators

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

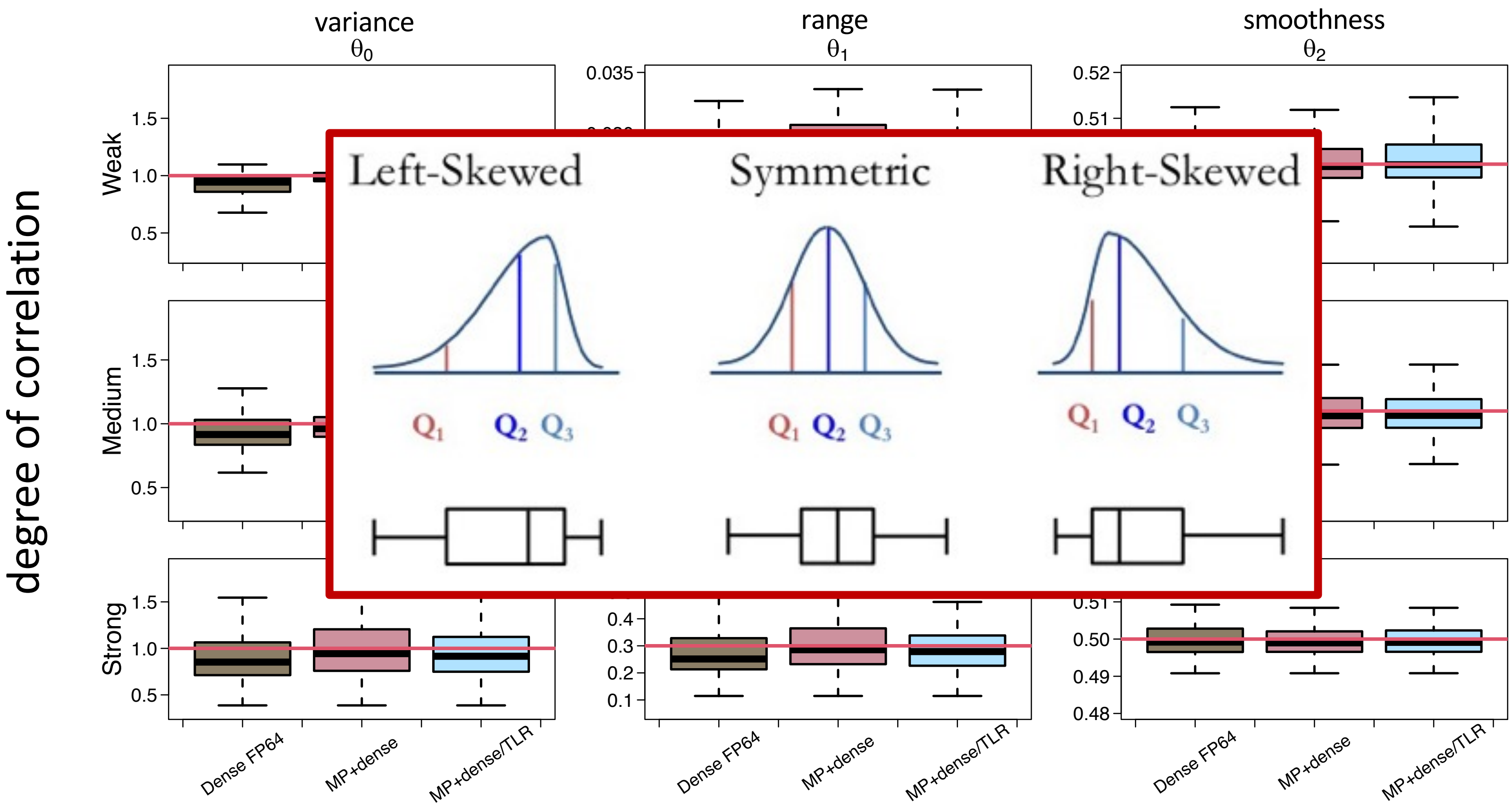
Hatem Ltaief

Ying Sun

Marc Genton

Accuracy on synthetic 2D space dataset

Maximum Likelihood Estimation (MLE) parameters



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

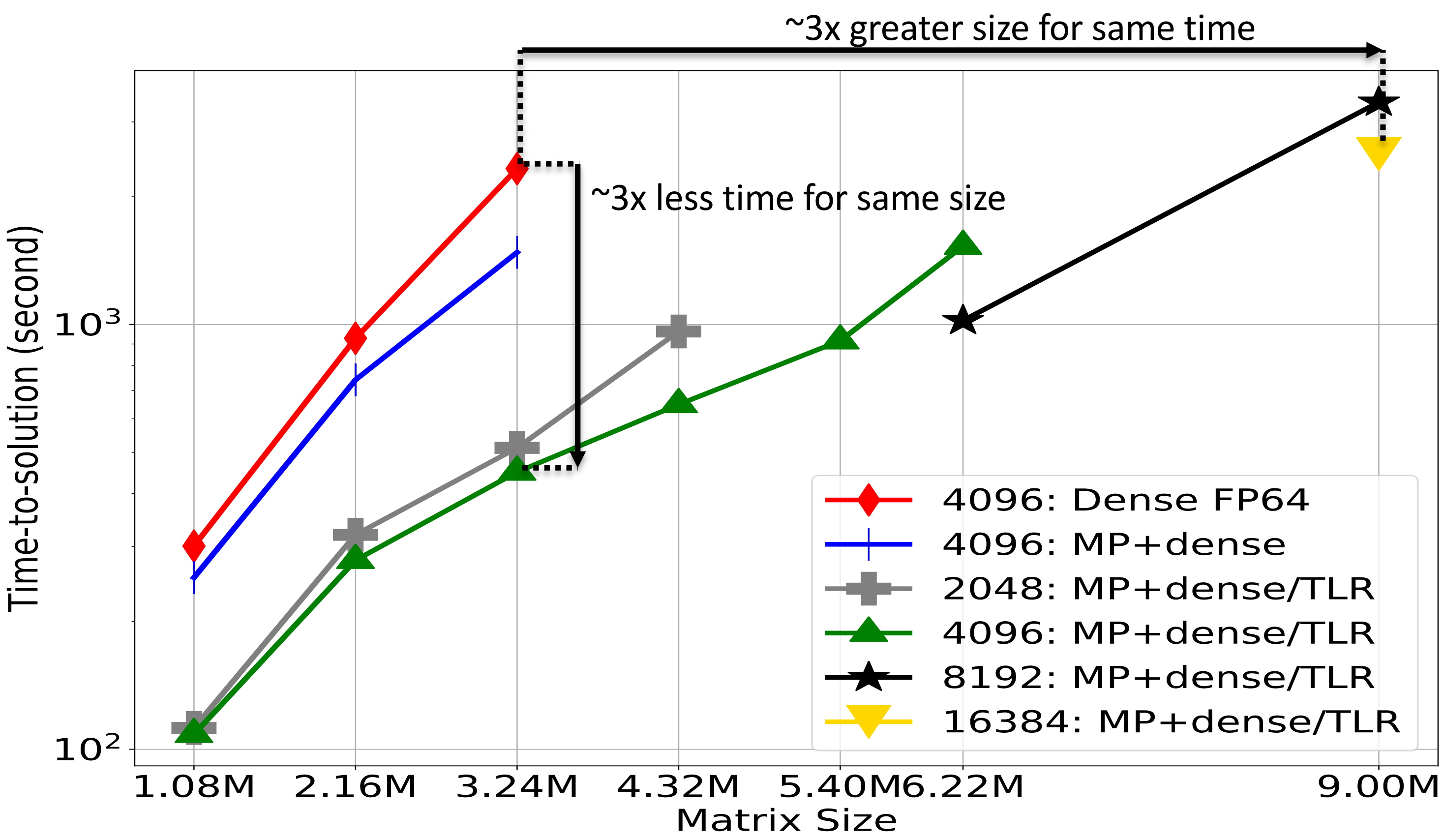
Variants	Variance (θ_0)	Range (θ_1)	Smoothness (θ_2)
Dense FP64	1.0087	3.7904	0.3164
MP+dense	0.9428	3.8795	0.3072
MP+dense/TLR	0.9247	3.7756	0.3068

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

2023 Gordon Bell (regular)

Performance Attributes	Our submission
Problem Size	Broadband 3D seismic dataset ($\sim 20k$ sources and receivers and frequencies up to $50Hz$)
Category of achievement	Sustained bandwidth Scalability
Type of method used	Algebraic compression
Results reported on basis of	Whole application (for GPU cluster) Main kernel (for Cerebras cluster)
Precision reported	Single precision complex
System scale	Up to 48 Cerebras CS-2 systems, i.e., $35,784,000$ processing elements[1]
Measurement mechanism	Timers; Memory accesses; Performance modeling

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

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

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

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

ABSTRACT

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

KEYWORDS

Seismic Processing, Low-Carbon Energy Applications, AI-optimized Architecture, Low-Rank Matrix Approximation, High Memory Bandwidth, Extreme Parallelism, Energy Efficiency.

ACM Reference Format:
Hatem Ltaief, Yuxi Hong, Leighton Wilson, Mathias Jacquelin, Matteo Ravasi, and David Keyes. 2023. Scaling the "Memory Wall" for Multi-Dimensional Seismic Processing with Algebraic Compression on Cerebras CS-2 Systems. In *The International Conference for High Performance Computing, Networking, Storage and Analysis (SC '23), November 12–17, 2023, Denver, CO, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581784.3627042

ACM ISBN 979-8-4007-0109-2/23/11.
https://doi.org/10.1145/3581784.3627042

1 JUSTIFICATION FOR THE GORDON BELL PRIZE

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

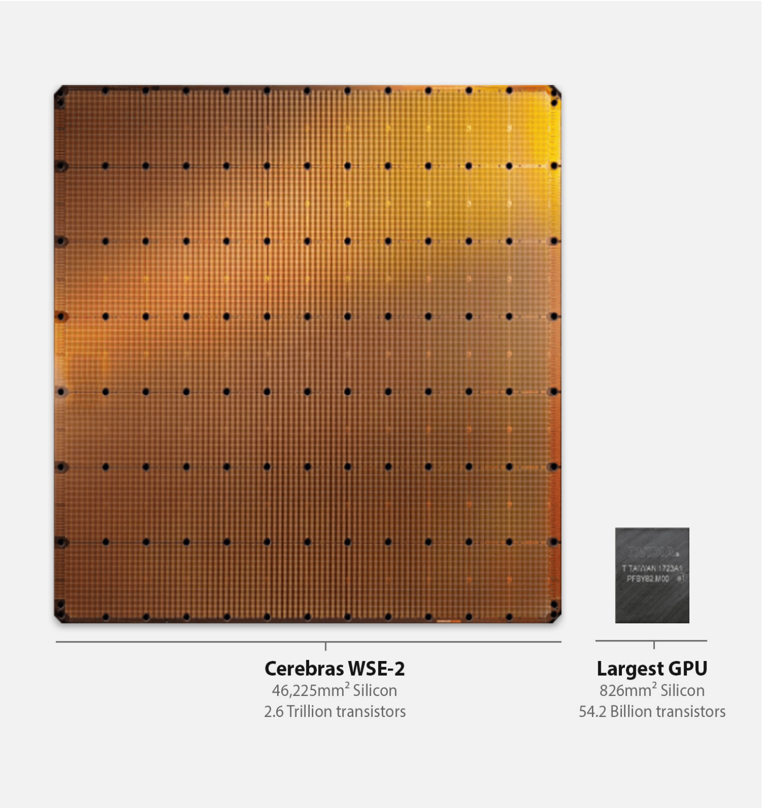
2 PERFORMANCE ATTRIBUTES

Performance Attributes	Our submission
Problem Size	Broadband 3D seismic dataset ($\sim 20k$ sources and receivers and frequencies up to 50Hz)
Category of achievement	Sustained bandwidth Scalability
Type of method used	Algebraic compression
Results reported on basis of	Whole application (for GPU cluster) Main kernel (for Cerebras cluster)
Precision reported	Single precision complex
System scale	Up to 48 Cerebras CS-2 systems, i.e., 35, 784, 000 processing elements
Measurement mechanism	Timers; Memory accesses; Sustained bandwidth; Performance modeling

3 OVERVIEW OF THE PROBLEM

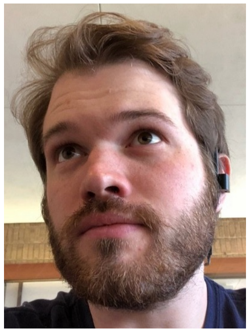
Reflection seismology is a remote sensing technique that utilizes reflected seismic waves to produce high-resolution images of the subsurface as well as estimates of the associated rock properties. While developed primarily to map anomalies corresponding to mineral or hydrocarbon deposits, it is now also being used for the

Cerebras CS-2 Wafer-Scale Engine (WSE)



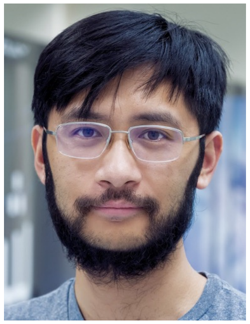
GB'23 collaborators

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



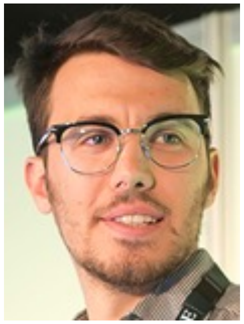
Leighton Wilson

Mathias Jacquelin



Yuxi Hong

Hatem Ltaief



Matteo Ravasi

TLR matvec for Multidimensional Deconvolution

Row dimension: # sources

Column dimension: # receivers

One such matrix for each independent frequency

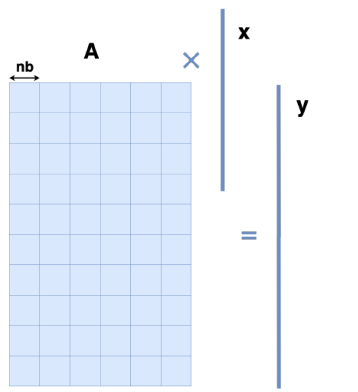


Fig. 2: Original dense MVM.

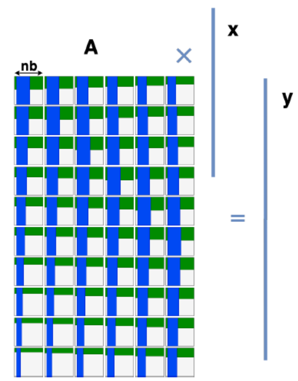


Fig. 3: Rank-compressed operator.

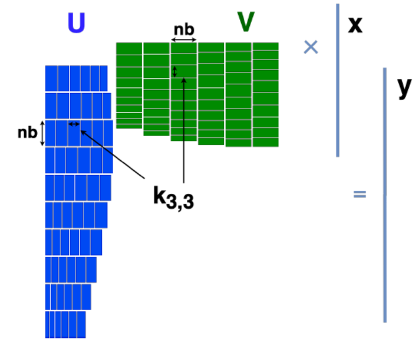


Fig. 4: Stacked bases U and V.

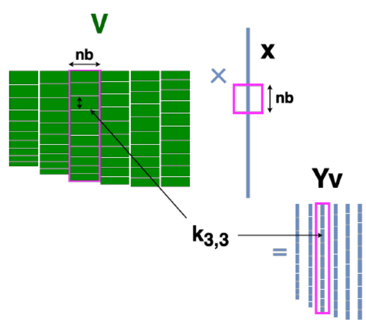


Fig. 5: V-batch stage of MVM.

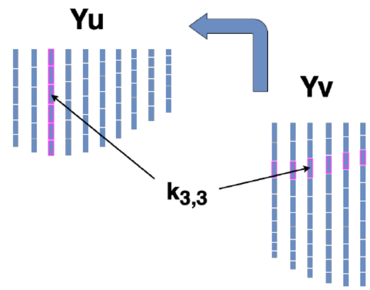


Fig. 6: Shuffle from V to U bases.

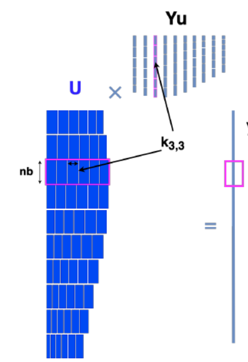
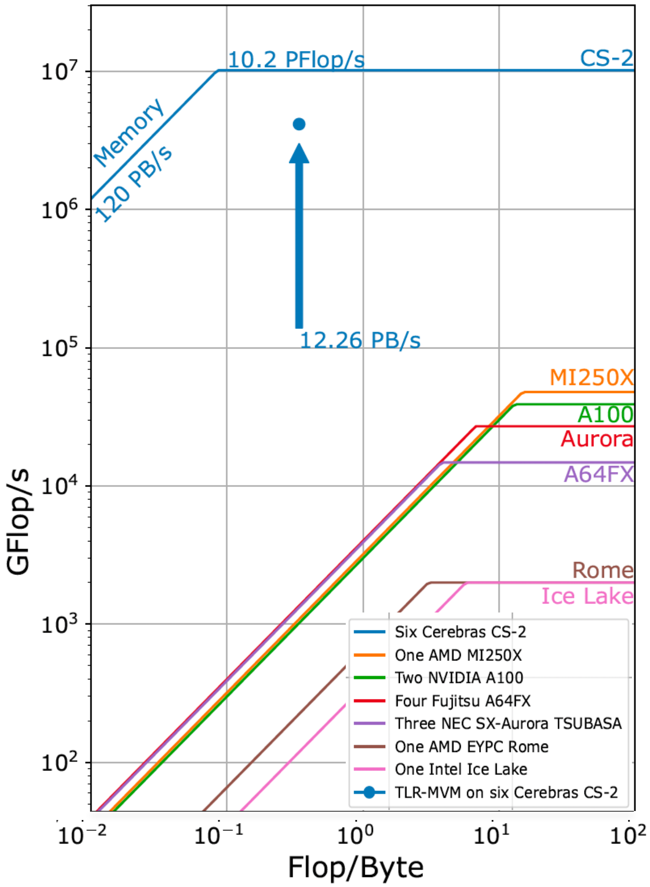


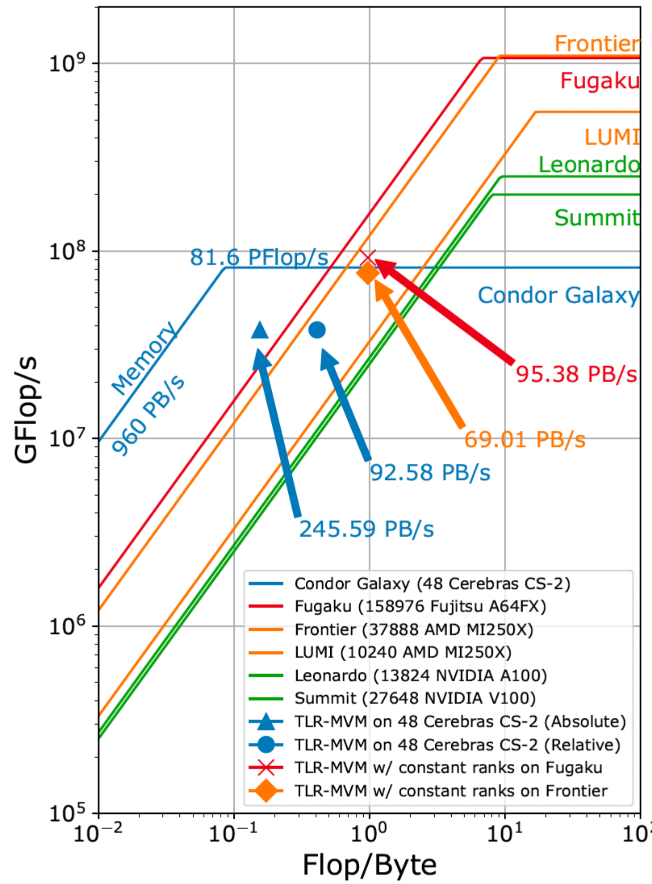
Fig. 7: U-batch of MVM.

2023 Gordon Bell submission

Roofline models of 6 CS-2's compared with other solutions



Roofline models of 48 CS-2's compared with top 5 (Nov 2023)



2024 Gordon Bell (regular)

Performance Attributes	Value
Problem Size	305K patients [real data, UK BioBank] 300K patients [synthetic data, msprime] 13M patients [synthetic data, random fill]
Category of achievement	Scalability, performance, time to solution
Type of method used	Kernel Ridge Regression
Results reported on basis of	Whole GWAS application: - mixed precision distance computation - mixed precision Cholesky factorization - mixed precision triangular solve
Precision reported	FP64, FP32, FP16, FP8, INT8
System scale	2/3 of `Summit` (18, 432 V100 GPUs) 1/3 of `Leonardo` (4, 096 A100 GPUs) full `Frontier` (36, 100 MI250X GPUs) 4/5 of `Alps` (8, 100 GH200 Superchips) Sustained 1.805 mixed precision ExaOp/s
Measurement mechanism	Timers, Flops

Toward Capturing Genetic Epistasis From Multivariate Genome-Wide Association Studies Using Mixed-Precision Kernel Ridge Regression

Hatem Ltaief[1,5], Rabab Alomairy[1,2,6], Qinglei Cao[3,7], Jie Ren[1,5], Lotfi Slim[4,8], Thorsten Kurth[4,9], Benedikt Dorschner[4,10], Salim Bougouffa[1,5], Rached Abdelkhalak[4,11], and David E. Keyes[1,5]

[1]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, KSA.
[2]Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA.
[3]Department of Computer Science, Saint Louis University, USA.
[4]NVIDIA, USA.
[5]{*Firstname.Lastname*}*@kaust.edu.sa*
[6]*rabab.alomairy@mit.edu* [7]*qinglei.cao@slu.edu* [8]*lslim@nvidia.com*
[9]*tkurth@nvidia.com* [10]*bdorschner@nvidia.com* [11]*rabdelkhalek@nvidia.com*

***Abstract*—We exploit the widening margin in tensor-core performance between [FP64/FP32/FP16/INT8,FP64/FP32/FP16/FP8/INT8] on NVIDIA [Ampere,Hopper] GPUs to boost the performance of output accuracy-preserving mixed-precision computation of Genome-Wide Association Studies (GWAS) of 305K patients from the UK BioBank, the largest-ever GWAS cohort studied for genetic epistasis using a multivariate approach. Tile-centric adaptive-precision linear algebraic techniques motivated by reducing data motion gain enhanced significance with low-precision GPU arithmetic. At the core of Kernel Ridge Regression (KRR) techniques for GWAS lie compute-bound cubic-complexity matrix operations that inhibit scaling to aspirational dimensions of the population, genotypes, and phenotypes. We accelerate KRR matrix generation by redesigning the computation for Euclidean distances to engage INT8 tensor cores while exploiting symmetry. We accelerate solution of the regularized KRR systems by deploying a new four-precision Cholesky-based solver, which, at 1.805 mixed-precision ExaOp/s on a nearly full Alps system, outperforms the state-of-the-art CPU-only REGENIE GWAS software by five orders of magnitude.**

***Index Terms*—Multivariate Genome-wide Association Studies, Kernel Ridge Regression, Nonlinear genotype-phenotype relationships, UK BioBank data, Tile-centric matrix computations, Mixed precision, Dynamic runtime system, GPU accelerators.**

I. JUSTIFICATION FOR THE GORDON BELL PRIZE

High-performance tile-centric matrix computations for Kernel Ridge Regression. End-to-end GWAS software supporting the largest-ever multivariate study of 305K patients from UK BioBank real datasets and 13M patients from synthetic datasets. Application-worthy FP32 accuracy using four precisions, including INT8 and FP8. Near-perfect weak-scaling on full-scale `Alps`, achieving 1.805 mixed precision ExaOp/s.

II. PERFORMANCE ATTRIBUTES

Performance Attributes	Value
Problem Size	305K patients [real data, UK BioBank] 300K patients [synthetic data, msprime] 13M patients [synthetic data, random fill]
Category of achievement	Scalability, performance, time to solution
Type of method used	Kernel Ridge Regression
Results reported on basis of	Whole GWAS application: - mixed precision distance computation - mixed precision Cholesky factorization - mixed precision triangular solve
Precision reported	FP64, FP32, FP16, FP8, INT8
System scale	2/3 of `Summit` (18, 432 V100 GPUs) 1/3 of `Leonardo` (4, 096 A100 GPUs) full `Frontier` (36, 100 MI250X GPUs) 4/5 of `Alps` (8, 100 GH200 Superchips) Sustained 1.805 mixed precision ExaOp/s
Measurement mechanism	Timers, Flops

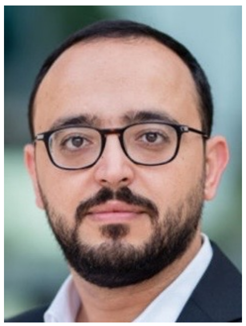
III. OVERVIEW OF THE PROBLEM

Genome-Wide Association Studies (GWAS) analyze DNA sequence variations spanning an entire genome (human or other) in order to identify genetic risk factors for diseases or other traits within a population. A main goal of GWAS is to use genetic factors to make predictions about individuals at risk and to identify the biological underpinnings of disease. This aids in the development of new diagnostic and therapeutic strategies [1].

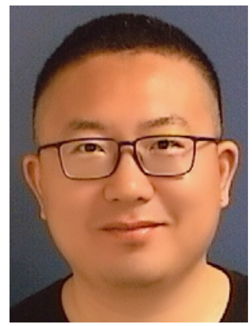
In a typical GWAS workflow, sketched in Fig. 1, several thousand to several million Single Nucleotide Polymorphisms (SNPs), the standard unit of genetic variation, are genotyped for large cohorts reaching into the millions of individuals. Extensive phenotypic information related to various traits or

GB’24 prize collaborators

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


Rached Abdelkhalek


Rabab Alomairy


Qinglei Cao

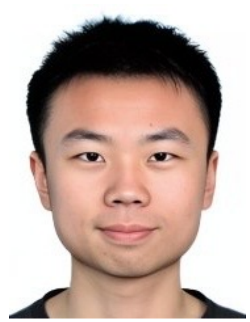

Benedikt Dorschner


Thorsten Kurth


Lotfi Slim

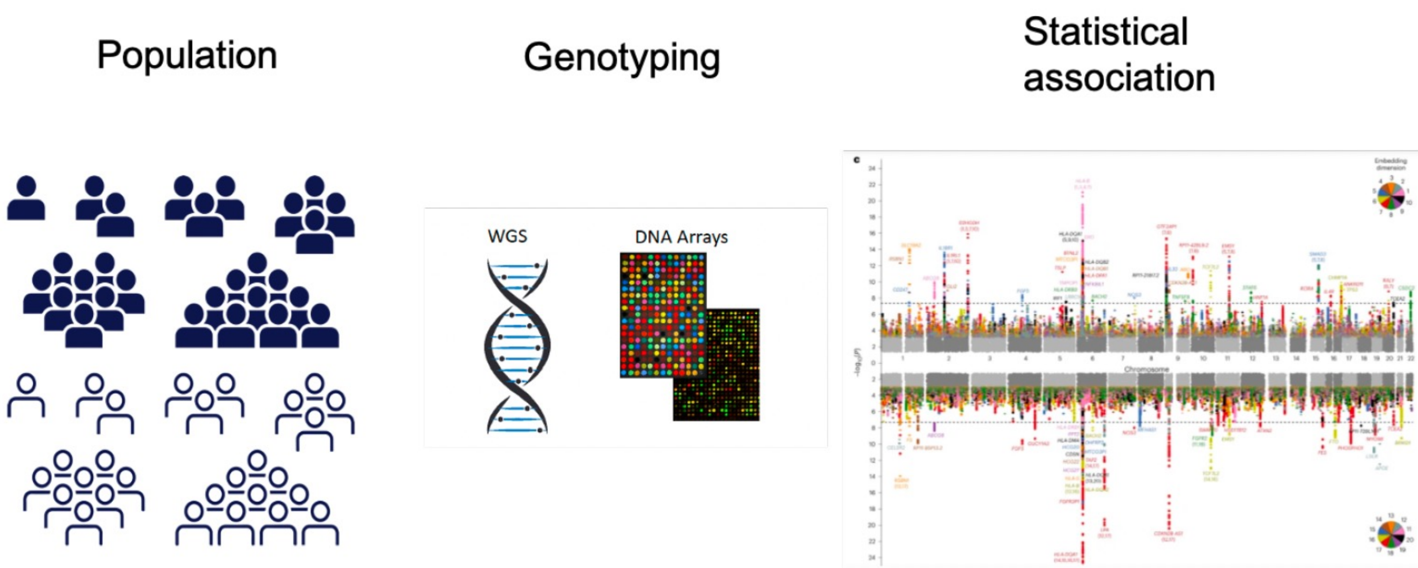

Salim Bougaffa


Hatem Ltaief


Jie Ren

Motivation for epistatic genome association studies

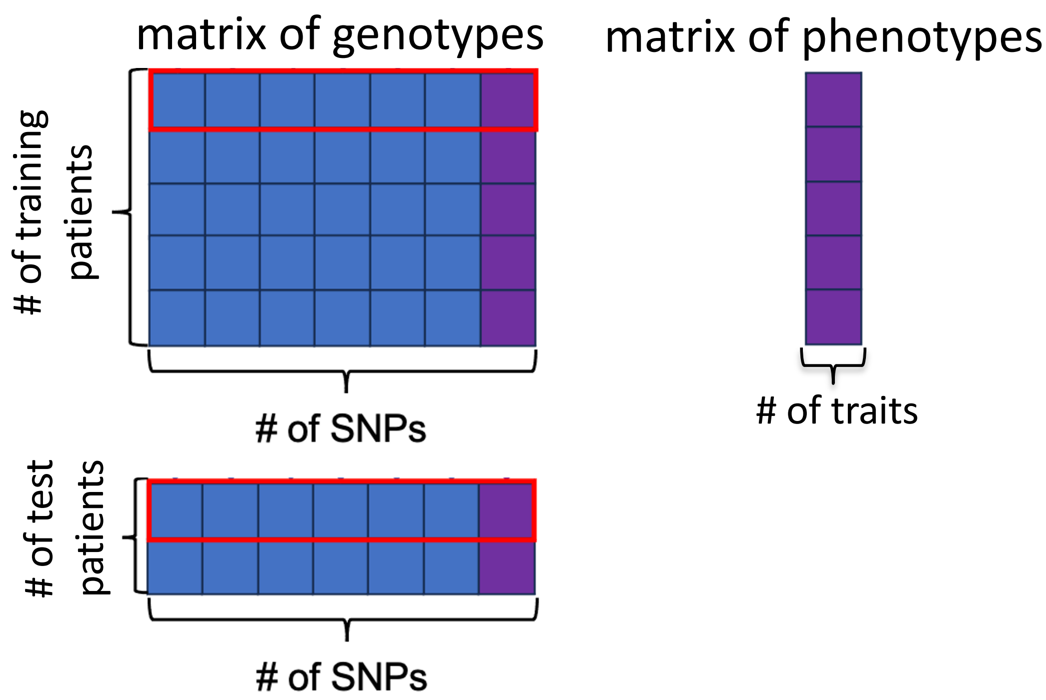
- Train statistical model on genotype/environmental-to-phenotype data
- Use to predict disease and other genetic/environmental characteristics
- Ridge regression is a linear association that considers individual SNPs
- Kernel ridge regression is correlates instances of multiple SNPs



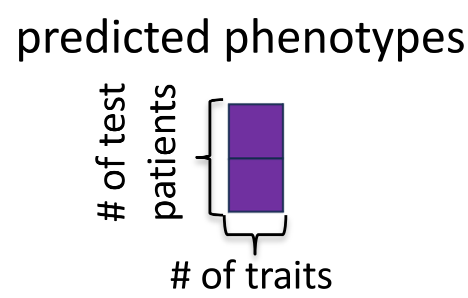
Genome Wide Association Studies (GWAS)

Inputs

integer real



Output



Euclidean distance between vectors is a GEMM!

Example: consider three 2D row vectors a, b, and c in matrix G:

$$G = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \\ c_1 & c_2 \end{pmatrix}$$

Let $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ and $a = (a_1^2 + a_2^2)^{1/2}$

etc.

We want all-pair square distances, w/symmetry

$$D = \begin{pmatrix} 0 & ||\mathbf{a}-\mathbf{b}||^2 & ||\mathbf{a}-\mathbf{c}||^2 \\ \text{sym} & 0 & ||\mathbf{b}-\mathbf{c}||^2 \\ & & 0 \end{pmatrix}$$

Now, $GG^T = \begin{pmatrix} a^2 & a_1b_1 + a_2b_2 & a_1c_1 + a_2c_2 \\ \text{sym} & b^2 & b_1c_1 + b_2c_2 \\ & & c^2 \end{pmatrix}$

Define $M = \begin{pmatrix} | & | & | \\ a^2 & b^2 & c^2 \\ | & | & | \end{pmatrix}$

Then $M + M^T = \begin{pmatrix} 2a^2 & a^2 + b^2 & a^2 + c^2 \\ \text{sym} & 2b^2 & b^2 + c^2 \\ & & 2c^2 \end{pmatrix}$

so, from $||\mathbf{a}-\mathbf{b}||^2 = a^2 + b^2 - 2\mathbf{a}\cdot\mathbf{b}$, etc.

$$\boxed{D = M + M^T - 2GG^T}$$

Kernel Ridge Regression is elementary linear algebra

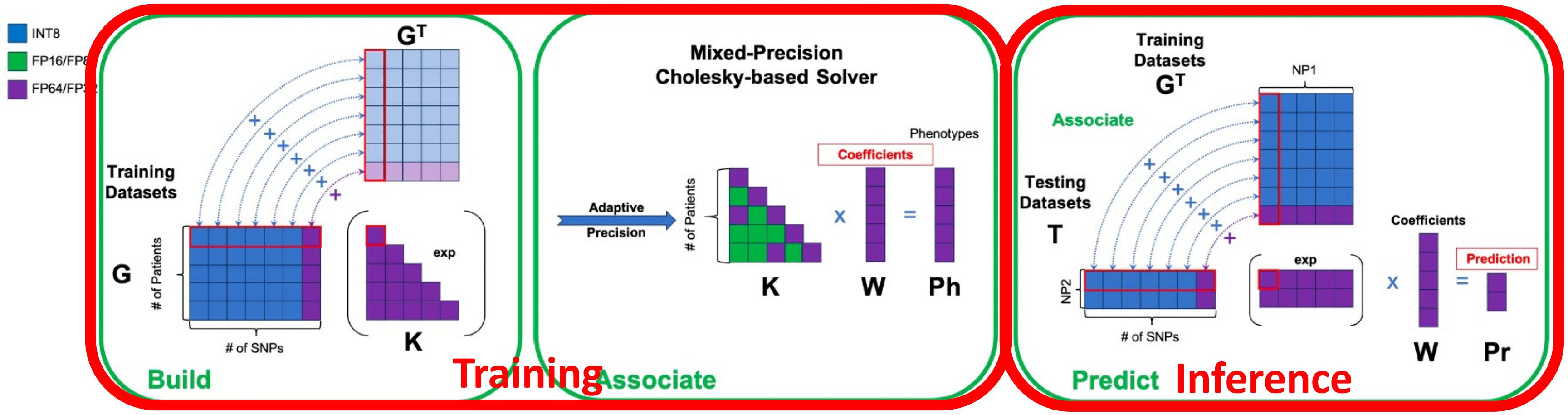


Fig. 3: Leveraging the INT8 / FP8 / FP16 / FP32 / FP64 multivariate KRR-based GWAS for genetic epistasis.

Algorithm 2: Build the KRR matrix.

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

Algorithm 5: Kernel Matrix Definitions.

Function KERNELMATRIX(type, γ, p_1, p_2)
$N_S \leftarrow$ size(p_1)
if type == 'Gaussian' **then**
  return $e^{-\gamma \cdot \|p_1 - p_2\|^2}$
else if type == 'IBS' **then**
  return $\frac{p_1 \sim p_2}{N_S}$
end if

Kernel Ridge Regression vs standard Ridge Regression

$$\text{Pearson corr} := \rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \qquad \begin{cases} \mu_X = \mathbb{E}[X] \\ \sigma_X^2 = \mathbb{E}\left[(X-\mu_X)^2\right] \end{cases}$$

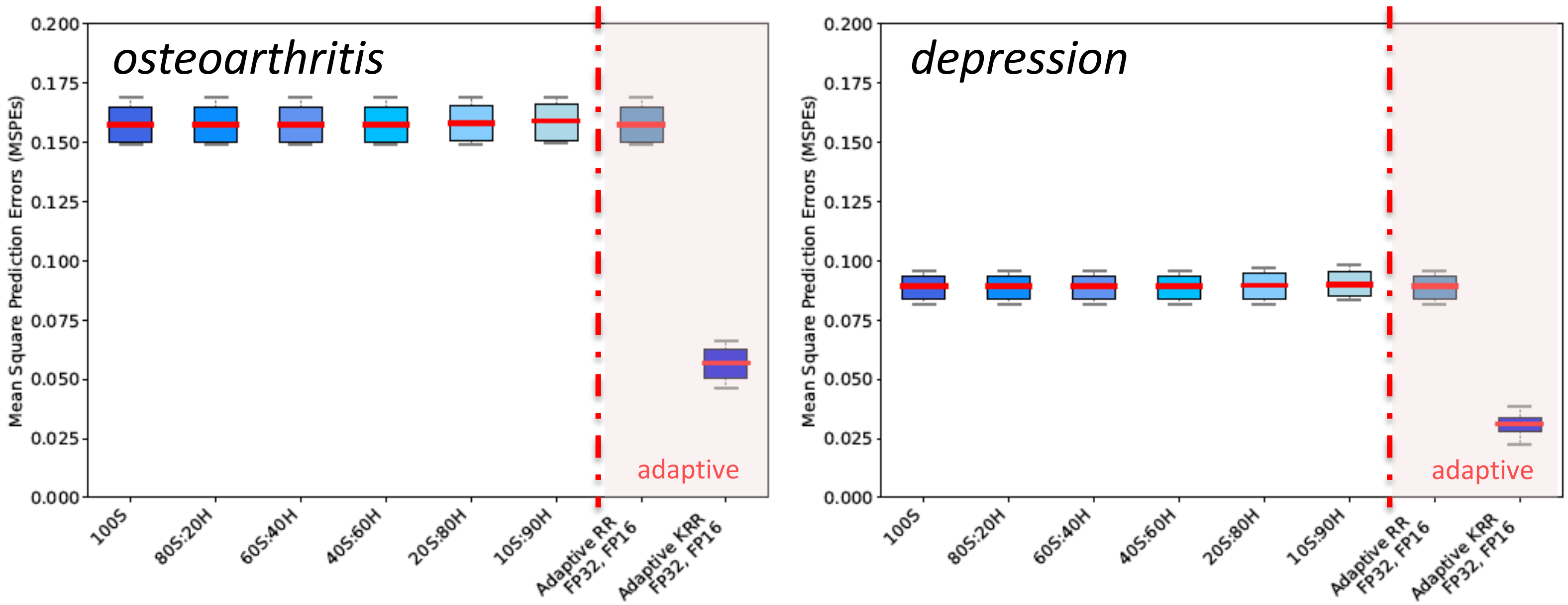
TABLE I: Comparing RR vs. KRR Pearson correlations.

Phenotypes	**RR**	**KRR**
Hypertension	0.2983	0.8071
Asthma	0.2517	0.8205
Allergic Rhinitis	0.2008	0.8652
Osteoarthritis	0.3189	0.8386
Depression	0.2041	0.8454

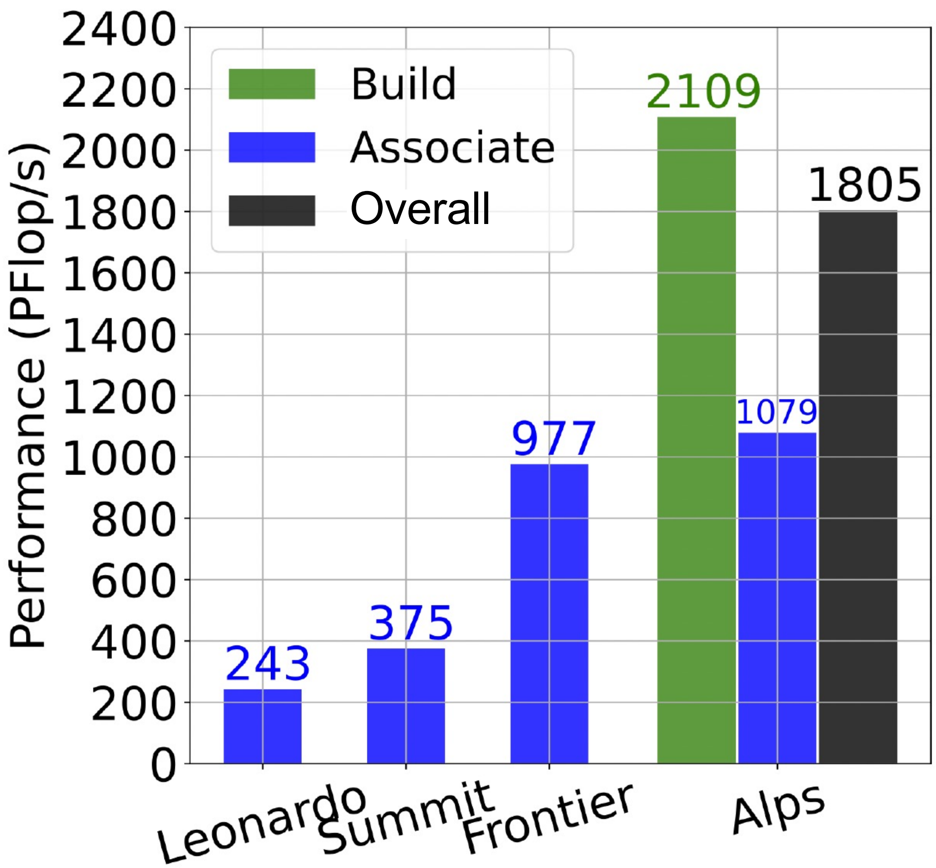
epistatic

Motivation for epistatic genome association studies

- Two representative phenotypes from the UK Biobank
- Box-and-whisker plots for 7 RR algorithms and KRR
- Increasing use of half precision in outer bands in first 5 plots
- Last 2 plots use adaptive precision, for RR and KRR, respectively



“Hero runs”: Leonardo, Summit, Frontier & Alps



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). 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.

Authors are listed alphabetically by their last names.

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

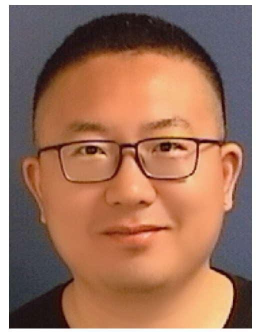
Climate change, evident in rising temperatures, extreme weather events, sea-level rise, and ecosystem disruption, poses significant risks and urgently requires action due to intensified heatwaves, storms, droughts, floods, and biodiversity loss [1], [2]. We stand at a critical juncture where converging

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

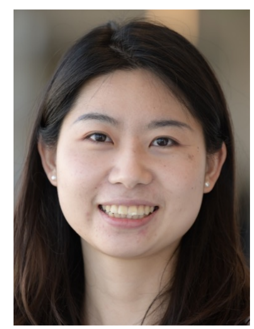

Gera Stenchikov


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

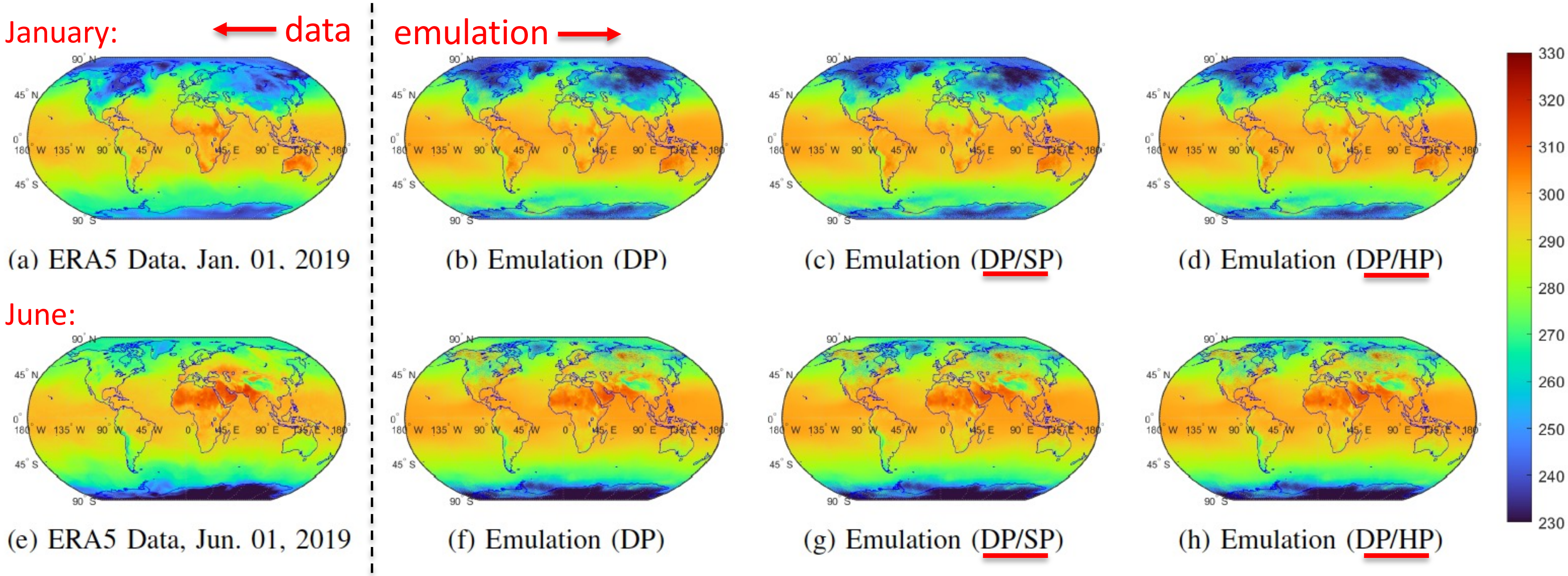
Contributions

- Developed and validated new climate emulator
  - emulates up to 54.5 million spatial locations across the globe with spatial resolution of 0.034° $\boxed{\text{(3.5 km)}}$ 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

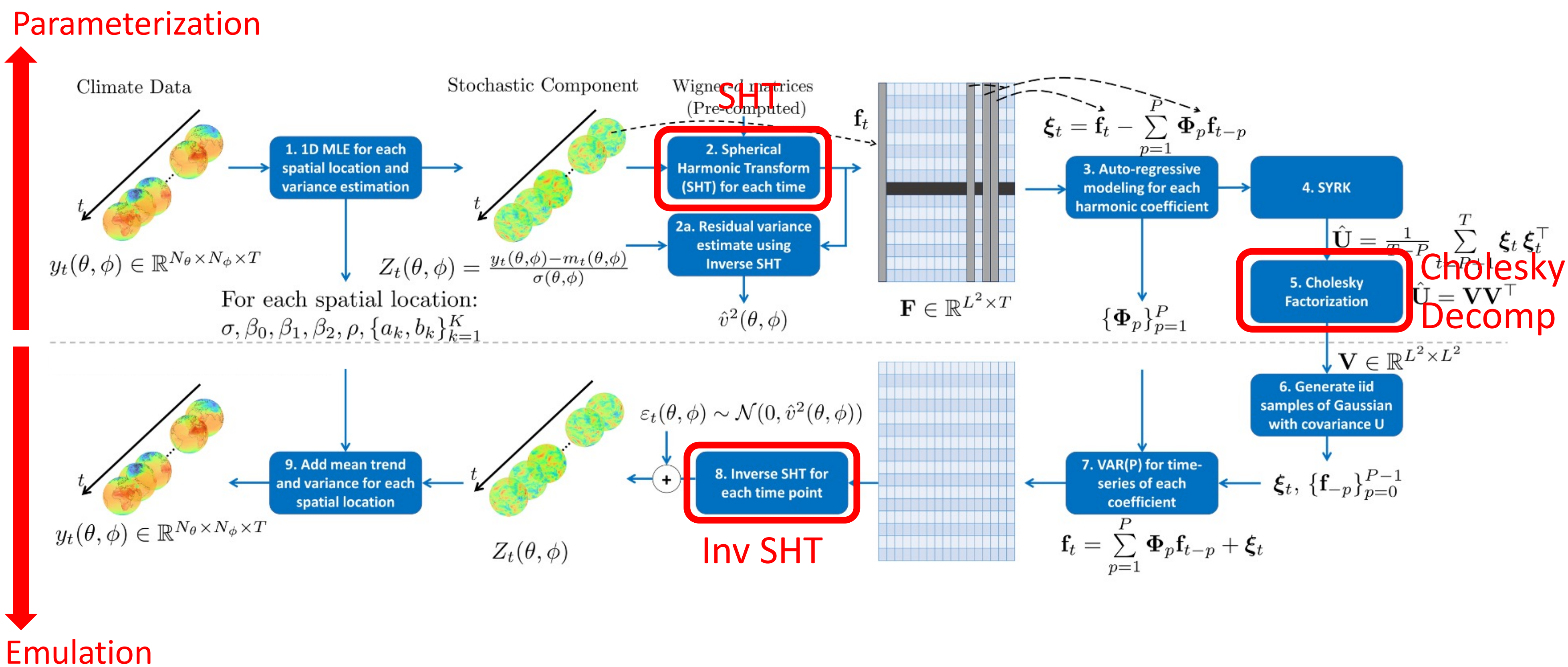
Primary cycle-consuming routines for fitting the emulation model are tolerant of mostly lower precision (single and half)



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

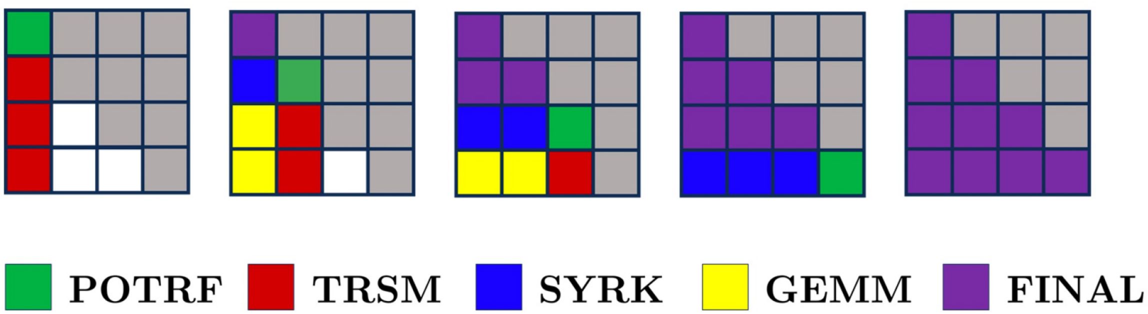
(e) ERA5 Data, Jun. 01, 2019 (f) Emulation (DP) (g) Emulation (DP/SP) (h) Emulation (DP/HP)

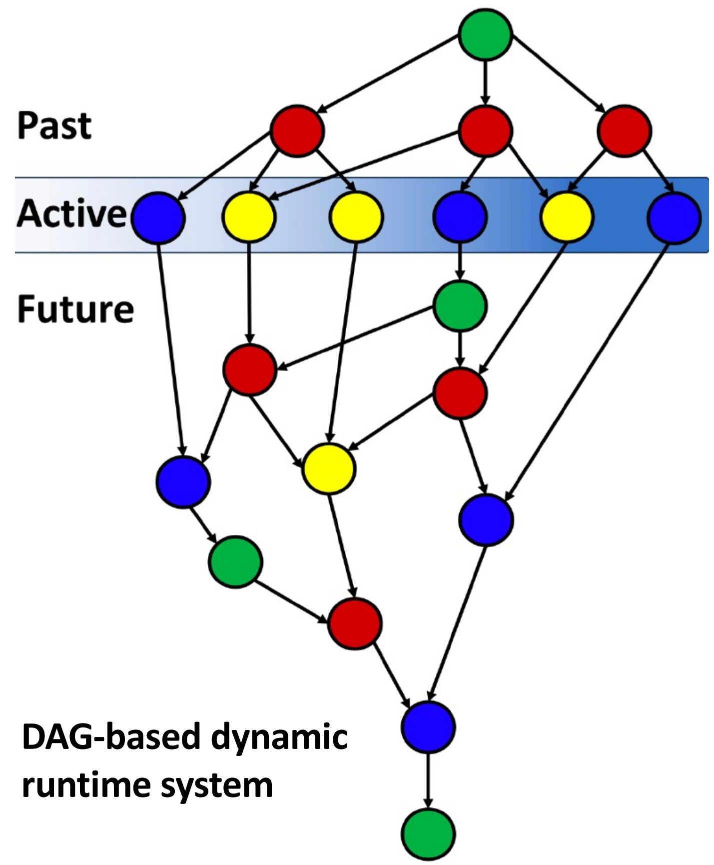
Algorithmic ingredients (2 stages, 2 major consumers)



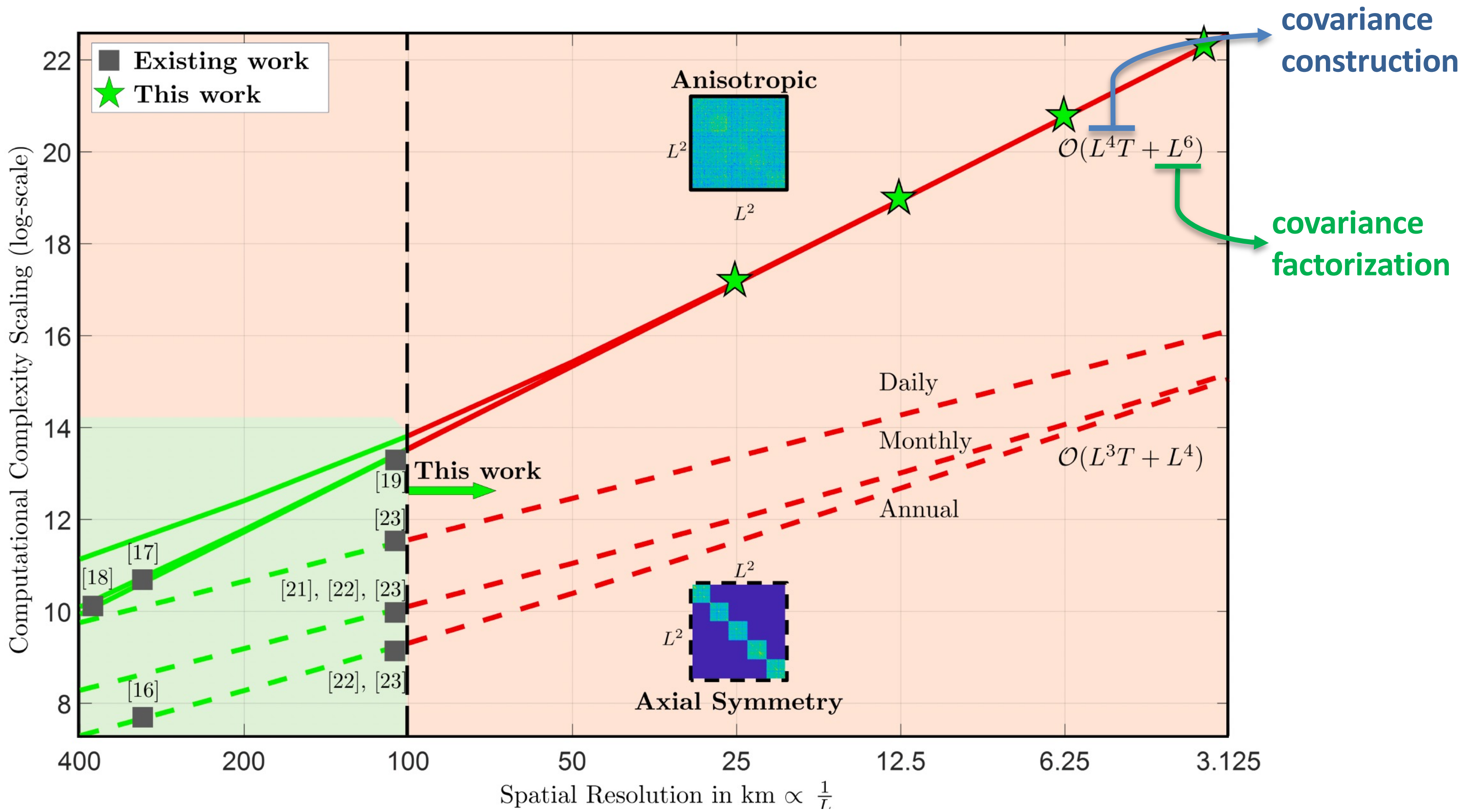
HPC ingredients (tiling and DAG dynamic runtime)

Tile-based Cholesky Decomposition





Expanding emulation resolution w/ memory austerity



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

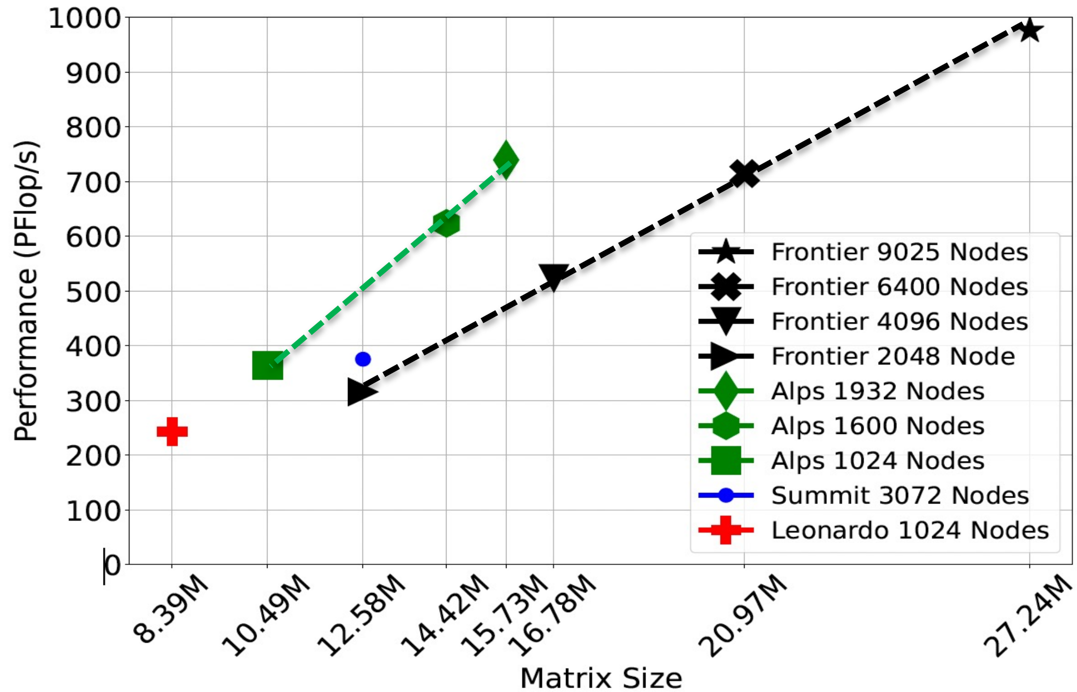


Fig. 8: Performance of largest runs on Summit, Leonardo, Alps, and Frontier; with additional run-up points on Alps and Frontier, all using the DP/HP precision variant.

Performance on three generations of NVIDIA hardware

- Exploits adaptive precision
  - PaRSEC dynamic runtime system to equi-distribute load adaptively
- Distributed memory over multi-GPU nodes
  - excellent weak scaling efficiency
  - up to 72% strong scaling efficiency on 12,288 V100 GPUs on Summit ("Summit's last stand" – decommissioned shortly afterwards)
- Impressive per node performance compared to PDE-based codes
  - 0.375 EFlop/s on 3,072 nodes Summit (V100)
  - 0.243 EFlop/s on 1,024 nodes of Leonard (A100)
  - 0.739 EFlop/s on 1,936 nodes of Alps (GH200)

GPU comparison

- For largest problem addressable on 1024 nodes (memory saturating)
  - 4 GPUs each for Leonardo, Alps, and Frontier
  - 6 GPUs per node for Summit

- GH200 is 72% faster than MI250X and 64% faster than A100

System	ORNL Summit	CINECA Leonardo	CSCS Alps	ORNL Frontier
Vendor	NVIDIA			AMD
GPU	V100	A100	GH200	MI250X
# GPUs	6,144	4,096	4,096	4,096
Matrix Size	6.20M	8.39M	10.49M	8.39M
Performance (PFlop/s)	153.6	243.1	384.2	223.7
TFlop/s/GPU	25	57.2	93.8	54.6

25 years ago… NASA's first Gordon Bell Prize

Achieving High Sustained Performance in an Unstructured Mesh CFD Application

W. K. Anderson,* W. D. Gropp,† D. K. Kaushik,‡ D. E. Keyes,§ and B. F. Smith¶

1 Overview

Many applications of economic and national security importance require the solution of nonlinear partial differential equations (PDEs). In many cases, PDEs possess a wide range of time scales—some (e.g., acoustic) faster than the phenomena of prime interest (e.g., convective), suggesting the need for implicit methods. In addition, many applications are geometrically complex and possess a wide range of length scales, requiring an unstructured mesh to adequately resolve the problem without requiring an excessive number of mesh points and to accomplish mesh generation and adaptation (almost) automatically. The best algorithms for solving nonlinear implicit problems are often Newton methods, which themselves require the solution of very large, sparse linear systems. The best algorithms for these sparse linear problems, particularly at very large sizes, are often preconditioned iterative methods. This nested hierarchy of tunable algorithms has proved effective in solving complex problems in areas such as aerodynamics, combustion, radiation transport, and global circulation. Typically, for

*Fluid Mechanics and Acoustics Division, NASA Langley Research Center, Hampton, VA 23682, w.k.anderson@larc.nasa.gov.

†Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, gropp@mcs.anl.gov. This work was supported in part by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, U.S. Department of Energy, under Contract W-31-109-Eng-38.

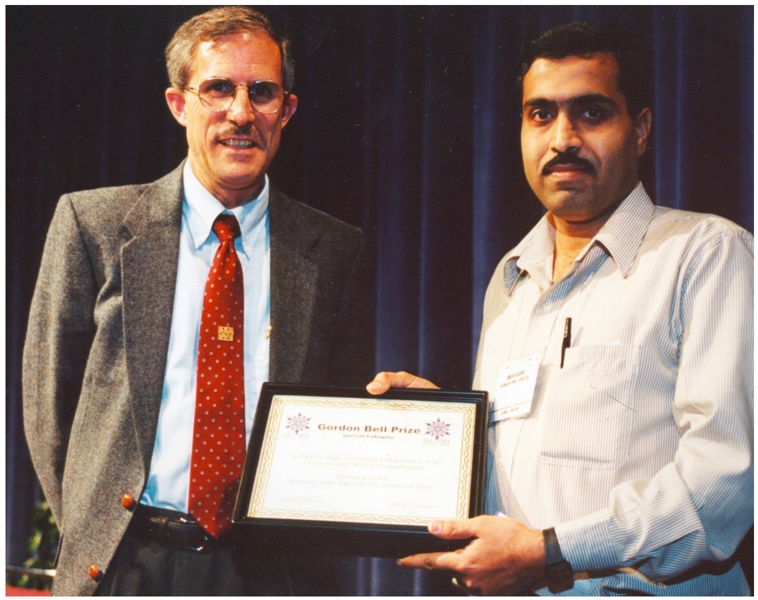
‡Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439 and Computer Science Department, Old Dominion University, Norfolk, VA 23529, kaushik@cs.odu.edu. This work was supported by a GAANN Fellowship from the U.S. Department of Education and by Argonne National Laboratory under contract 983572401.

§Mathematics & Statistics Department, Old Dominion University, Norfolk, VA 23529, ISCR, Lawrence Livermore National Laboratory, Livermore, CA 94551, and ICASE, NASA Langley Research Center, Hampton, VA 23681, keyes@icase.edu. This work was supported by the National Science Foundation under grant ECS-9527169, by NASA under contracts NAS1-19480 and NAS1-97046, by Argonne National Laboratory under contract 982232402, and by Lawrence Livermore National Laboratory under subcontract B347882.

¶Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, bsmith@mcs.anl.gov. This work was supported in part by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, U.S. Department of Energy, under Contract W-31-109-Eng-38.

FUN3D
Fully Unstructured Navier-Stokes

PETSc



Kaushik

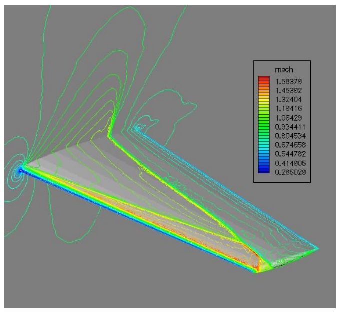
Anderson

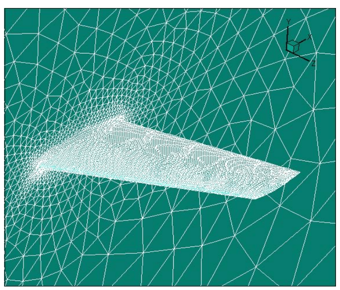
Gropp

Smith

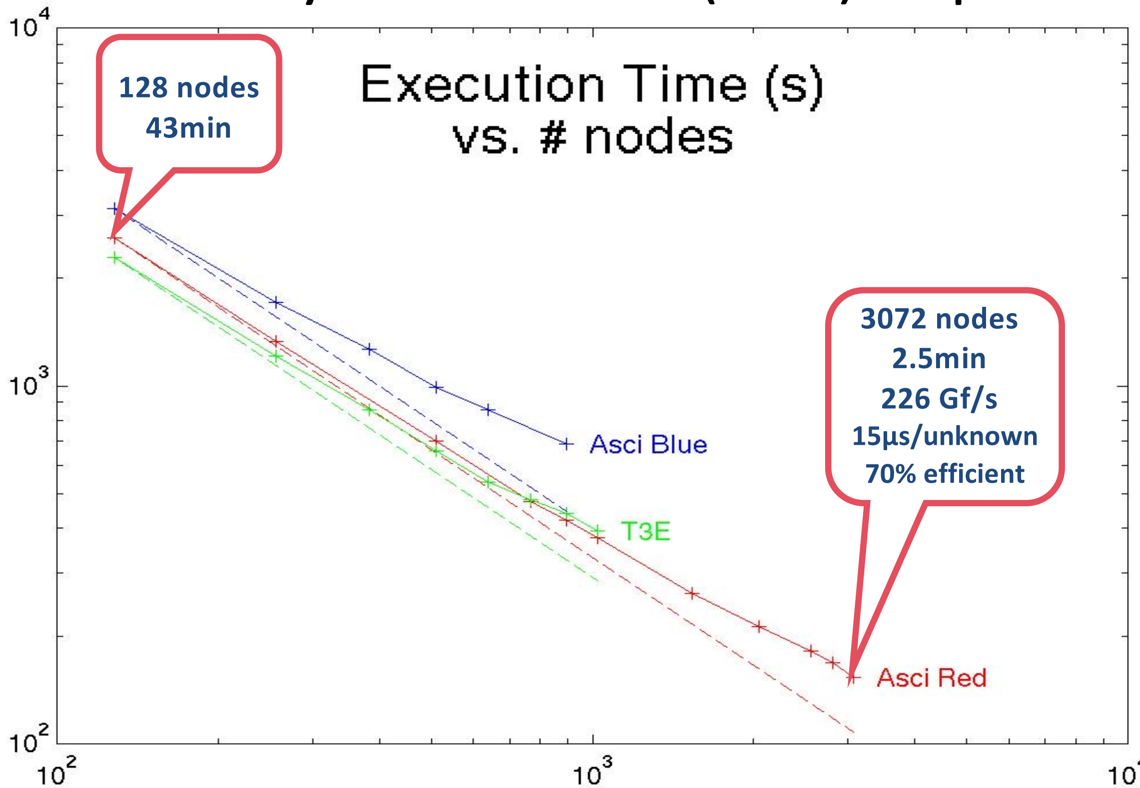
PETSc-FUN3D strong scaling (1999)

Jacobian-free Newton-Krylov-Schwarz (NKS) implicit solver





M6 wing with 11M dof



Hardware environment

- Scaling demonstrated on 3 Top 10 systems
  - Intel ASCI Red, Cray T3E, and IBM ASCI Blue
  - different network topologies
  - scaling efficiency in flop/s
  - scaling efficiency in execution time
- Per-processor efficiency evaluated on 14 leading scientific computing nodes
  - MIPS R10000, 2 DEC Alpha, 3 Sparc, 4 IBM Power, 4 Intel Pentium
  - different ISAs and memory systems
  - percentage of peak ranged from 7.6% (DEC) to 25.4% (MIPS)

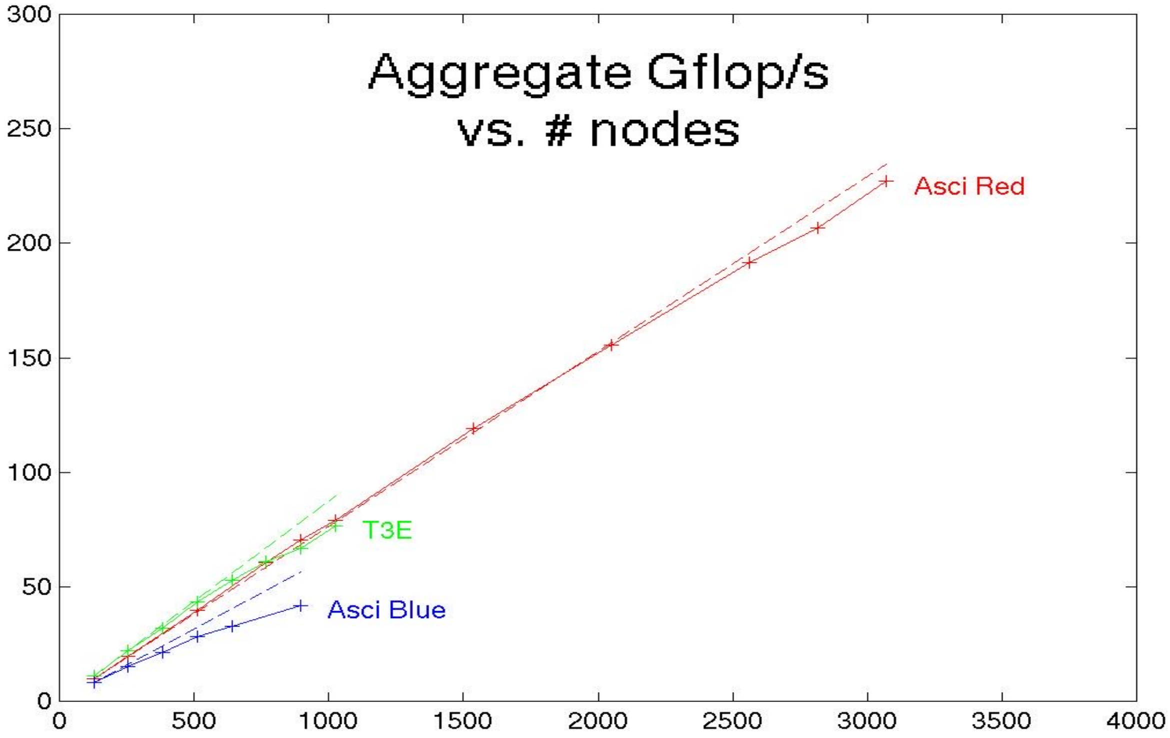
Features of this submission

- Based on "legacy" (but contemporary) CFD application with significant F77 code reuse
- Portable, message-passing library-based parallelization, run on NT boxes through Tflop/s ASCI platforms
- Simple multithreaded extension (for ASCI Red)
- Sparse, unstructured data, implying memory indirection with only modest reuse - nothing in this category has ever advanced to Bell finalist round
- Wide applicability to other implicitly discretized multiple-scale PDE workloads - of interagency, interdisciplinary interest
- Extensive profiling has led to follow-on algorithmic research

Four primary PDE solution kernels

- Vertex-based loops
  - state vector and auxiliary vector updates
- Edge-based “stencil op” loops
  - residual evaluation
  - approximate Jacobian evaluation
  - Jacobian-vector product (often replaced with matrix-free form, involving residual evaluation)
- Sparse, narrow-band recurrences
  - approximate factorization and back substitution
- Vector inner products and norms
  - orthogonalization/conjugation
  - convergence progress and stability checks

Fixed-size parallel scaling results (flop/s)



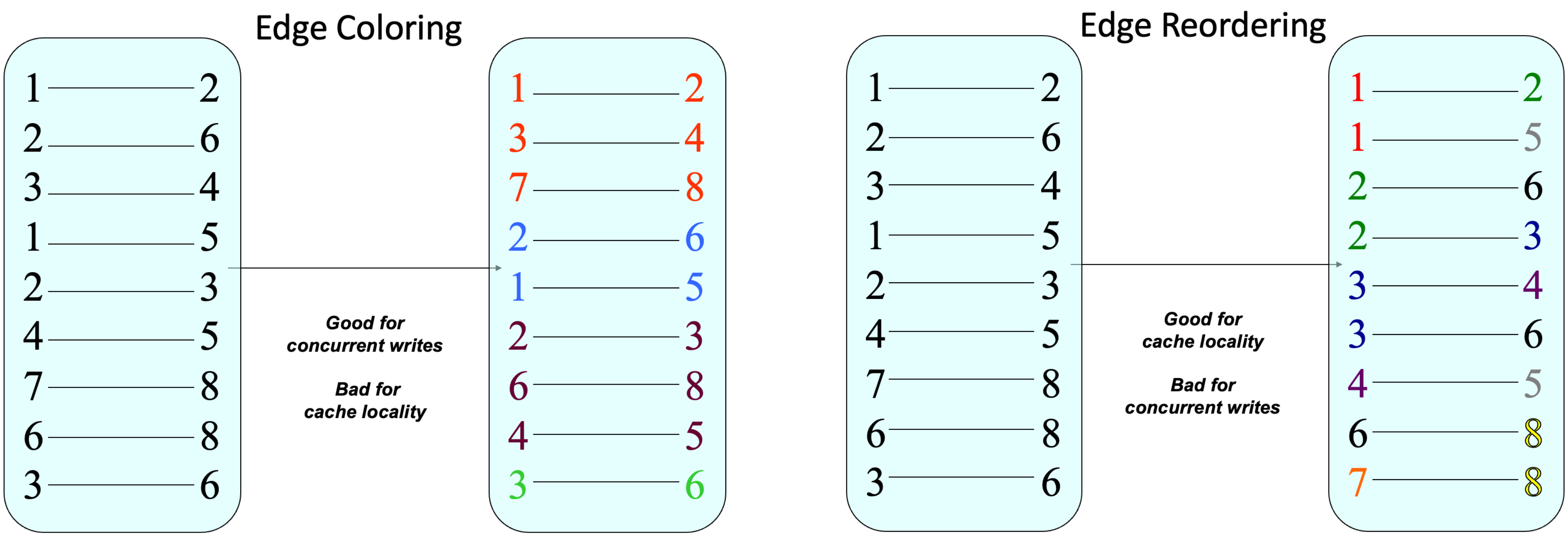
Excerpts from 1999 GB Special Prize paper

"Although these algorithms are efficient (in the sense of using relatively few floating-point operations to arrive at the final result), they do not necessarily achieve the absolute flops-per-second (flop/s) ratings that less efficient or less versatile algorithms may."

"This level of performance (in excess of 100 Gflop/s) is well above what is commonly considered achievable for sparse-matrix and unstructured mesh computations and requires a combination of scalable algorithms and data structure optimizations, as well as powerful, tightly networked computers."

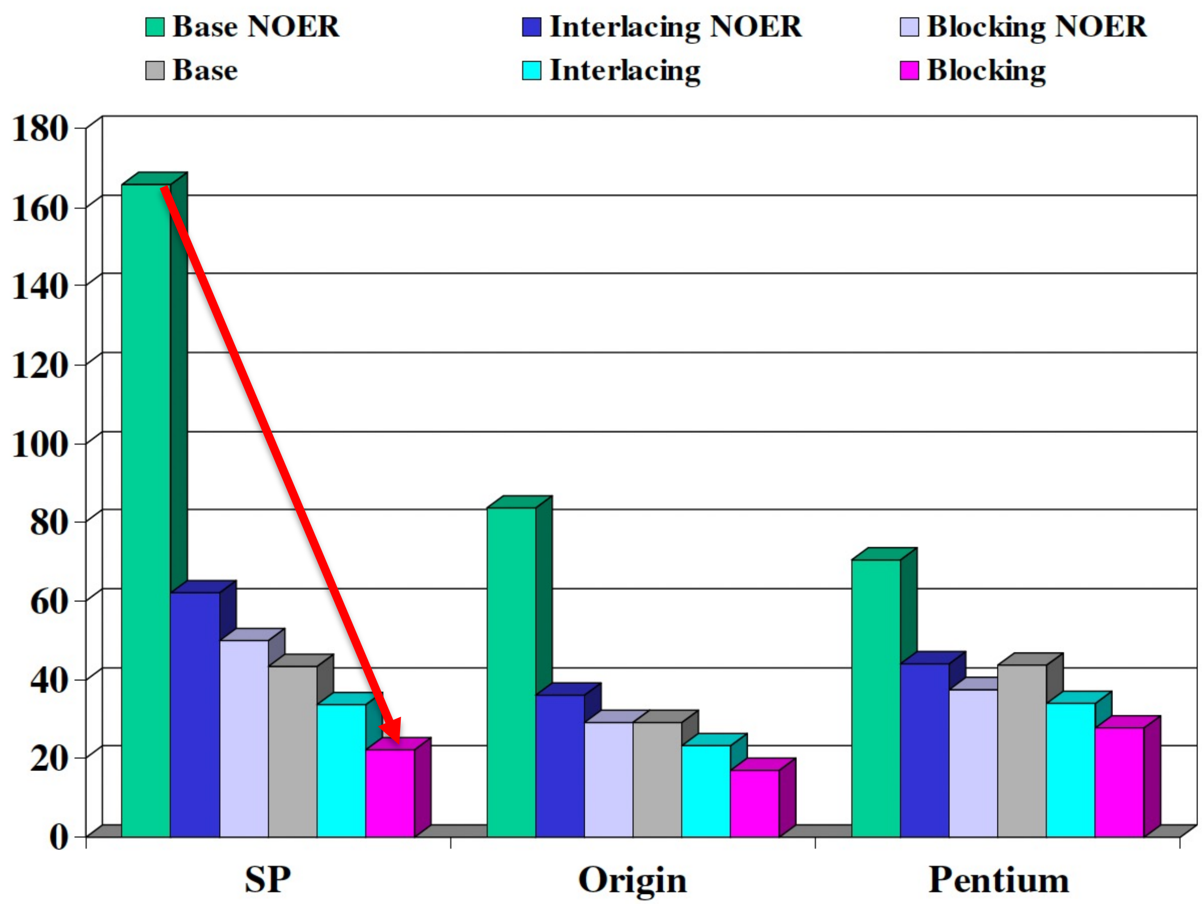
"High sustained scalable performance has been demonstrated on simulations that use implicit algorithms of choice for unstructured PDEs. In the history of the peak performance Bell Prize competition, PDE-based computations have led [...] in 1988, 1989, 1990, and 1996. All of these leading entries have been obtained on vector or SIMD architectures, and all were based on structured meshes."

FUN3D coloring slide

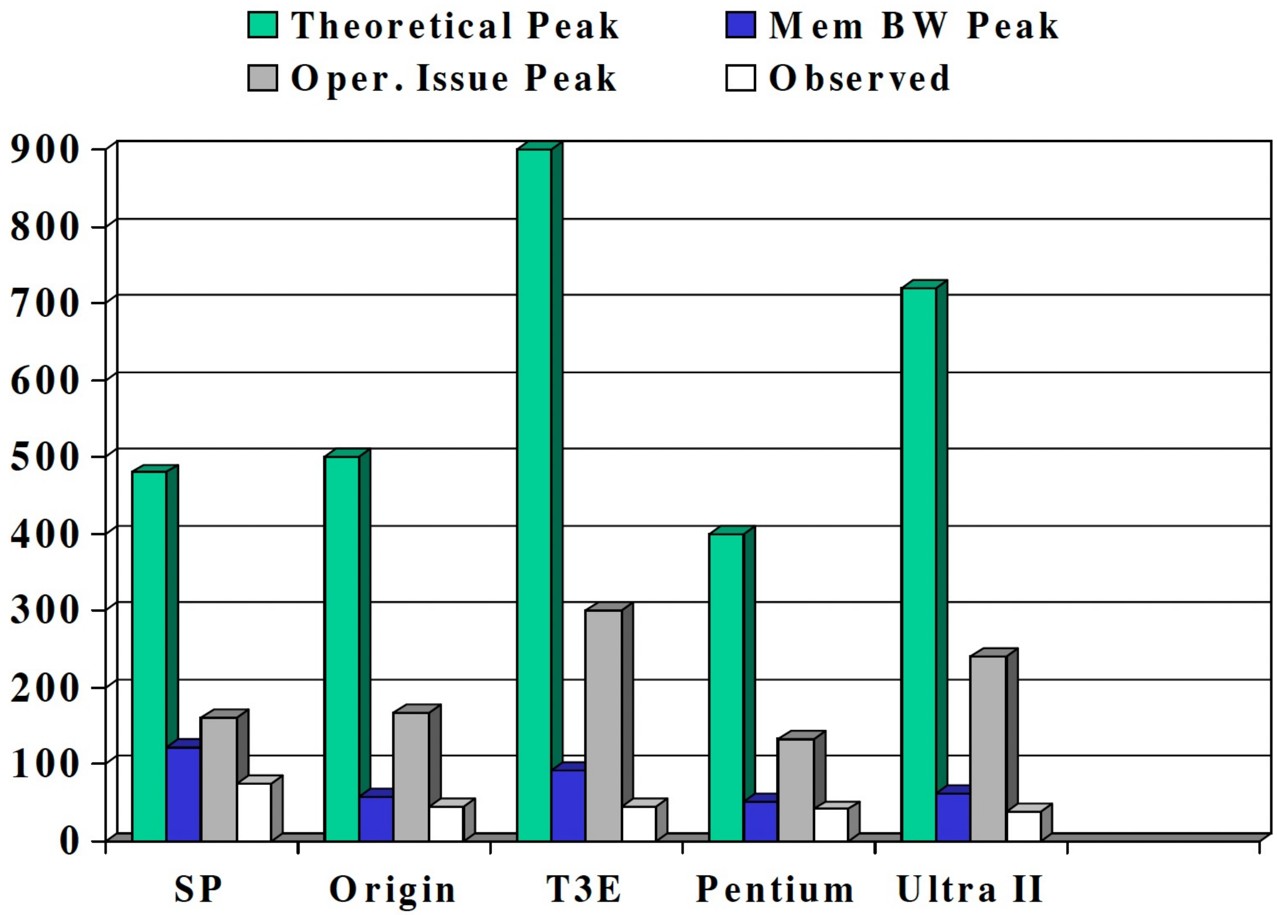


Explored many code optimizations in this unstructured, integer-instruction laden code

Up to 8-fold single node performance boost



Early "Roofline Model" analysis



For what the Bell *should* toll

(prereq) • Best algorithm for scientific outcome of interest

- Followed by best implementation for each stage

(1999) • High percentage of peak of the limiting resource for each phase of the algorithm, whether…

- flop/s
- byte/s
- floats/byte
- reductions

need roofline analysis, one per phase, ideally based on HPL Rmax and HPCG, rather than theoretical flops and bandwidth

(2025) • "Science per Joule", based on what is the state of the art for the relevant application

Statistics – the original ML

The "regression" form of machine learning is "under control" with linear algebra advances.

The "classification" form and generative form of ML is out of control.

ADVANCED REVIEW
Wiley Interdisciplinary Reviews: Computational Statistics

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 23955-6900, Saudi Arabia
[2]Department of Statistics, University of Connecticut, Storrs, CT 06269-4120, USA
[3]Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

Correspondence
Marc G. Genton, Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia
Email: marc.genton@kaust.edu.sa

Funding information
King Abdullah University of Science and Technology

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.

KEYWORDS
GPUs, high-performance computing, mixed-precision computing, parallel statistical algorithms, statistical computing

Sep 2025

Enter the QPUs

"Our analysis reveals a significant overlap between the technological capabilities [of quantum computing] and the algorithmic requirements in [materials science, quantum chemistry, and high-energy physics]. We anticipate that the execution time of large-scale quantum workflows will become a major performance parameter and propose a simple metric, the **Sustained Quantum System Performance (SQSP)**, to compare system-level performance and throughput for a heterogeneous workload."

QUANTUM @ NERSC

Report No. LBNL-2001699

Quantum Computing Technology Roadmaps and Capability Assessment for Scientific Computing

An analysis of use cases from the NERSC workload

Daan Camps Ermal Rrapaj Katherine Klymko Hyeongjin Kim Kevin Gott
Siva Darbha Jan Balewski Brian Austin Nicholas J. Wright

National Energy Research Scientific Computing Center (NERSC)
Lawrence Berkeley National Laboratory (LBNL)
Berkeley, CA, USA

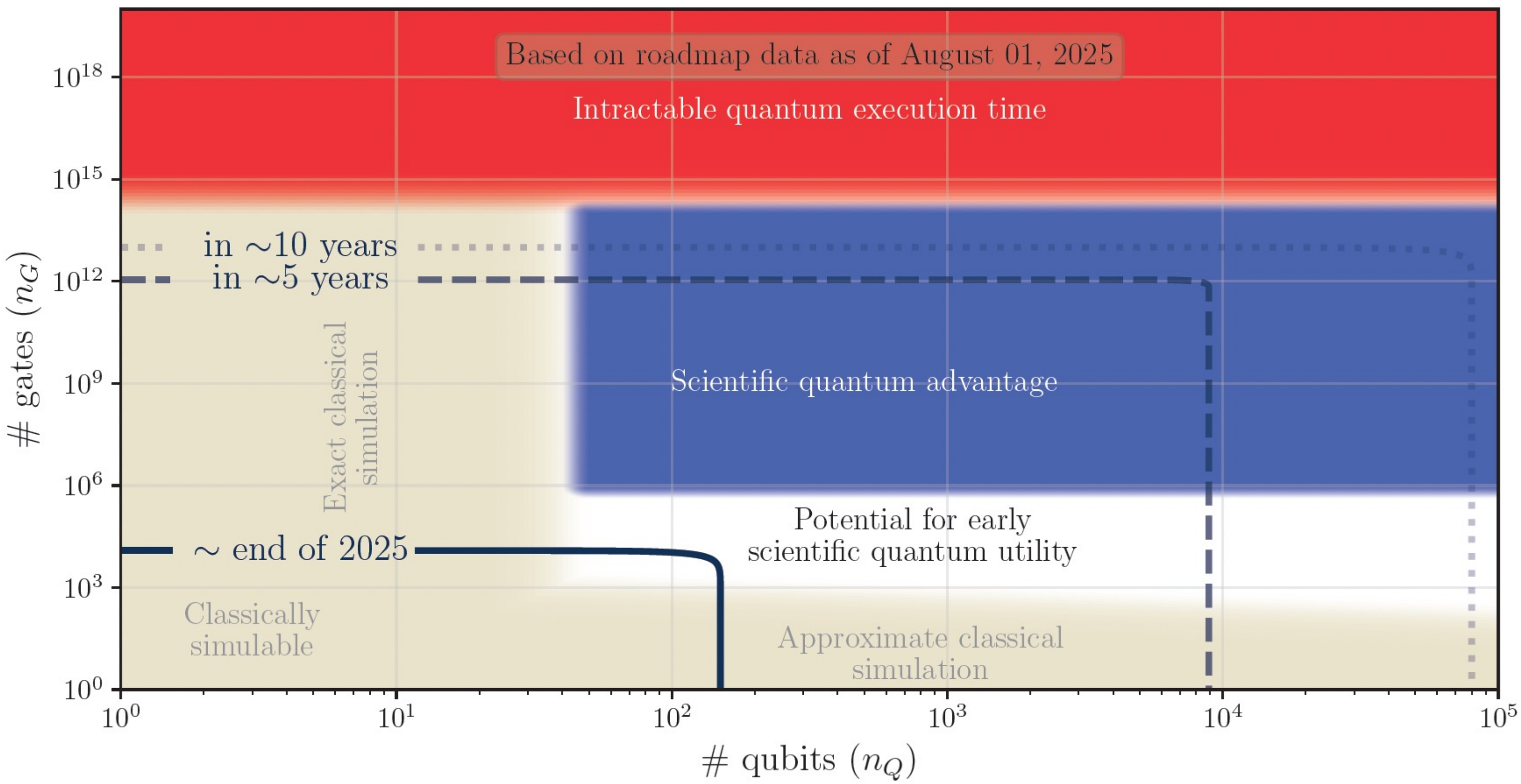
September 11, 2025

Abstract

The National Energy Research Scientific Computing Center (NERSC), as the high-performance computing (HPC) facility for the Department of Energy's Office of Science, recognizes the essential role of quantum computing in its future mission. In this report, we analyze the NERSC workload and identify materials science, quantum chemistry, and high-energy physics as the science domains and application areas that stand to benefit most from quantum computers. These domains jointly make up over 50% of the current NERSC production workload, which is illustrative of the impact quantum computing could have on NERSC's mission going forward. We perform an extensive literature review and determine the quantum resources required to solve classically intractable problems within these science domains. This review also shows that the quantum resources required have consistently decreased over time due to algorithmic improvements and a deeper understanding of the problems. At the same time, public technology roadmaps from a collection of ten quantum computing companies predict a dramatic increase in capabilities over the next five to ten years. Our analysis reveals a significant overlap emerging in this time frame between the technological capabilities and the algorithmic requirements in these three scientific domains. We anticipate that the execution time of large-scale quantum workflows will become a major performance parameter and propose a simple metric, the Sustained Quantum System Performance (SQSP), to compare system-level performance and throughput for a heterogeneous workload.

Sep 2025

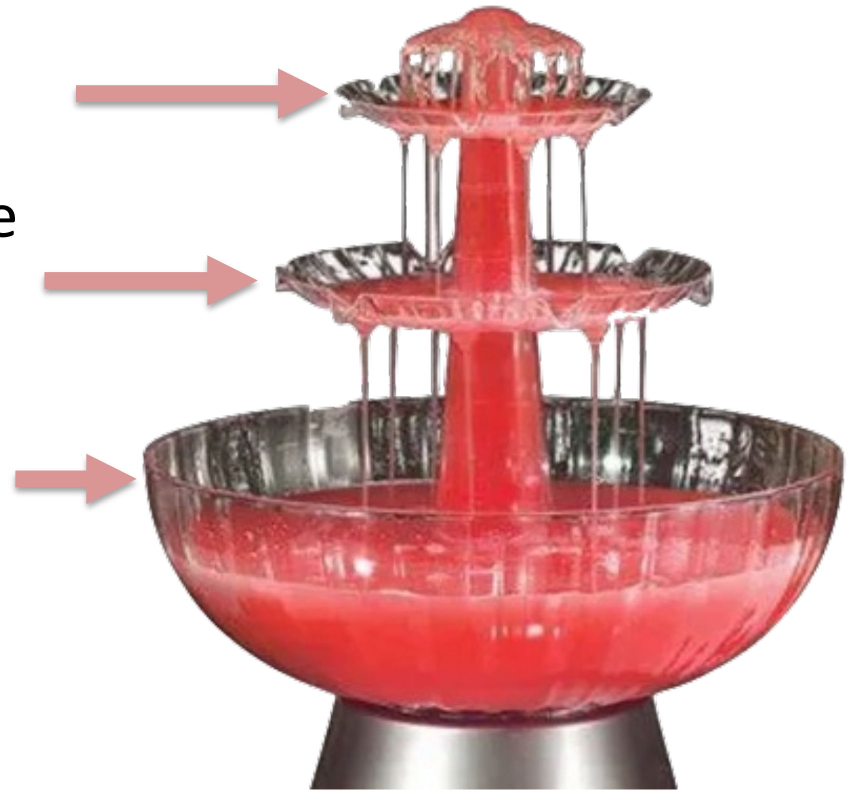
Enter the QPUs – hybrid classical-quantum computing



Sep 2025

A cascadic philosophy for future HPC

- QPUs will be used everywhere they have advantage
- GPUs will be used for everything else with high arithmetic intensity and structured memory accesses
- CPUs will take up the low arithmetic intensity unstructured remainder
- Efficiently assigning the resources of a QPU-GPU-CPU hybrid supercomputer to multiple jobs will be a rich programming challenge



Allusion

For *Whom* the Bell Tolls

John Donne (1572 - 1631)

No man is an island,
Entire of itself.
Each is a piece of the continent,
A part of the main…

How prophetic of of High Performance Computing
four centuries later!

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) *Scaling the "Memory Wall" for Multi-Dimensional Seismic Processing with Algebraic Compression on Cerebras CS-2 Systems,* H. Ltaief, Y. Hong, L. Wilson, M. Jacquelin, M. Ravasi, & David Keyes, 2023, in proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'22), IEEE Computer Society (ACM Gordon Bell Finalist), doi 10.1145/3581784.362704.
7) *Toward Capturing Complex Genetic Architectures from Genome-Wide Association Studies Using Mixed-Precision Kernel Ridge Regression,* H. Ltaief, R. Alomairy, J. Ren, Q. Cao, L. Slim, S. Bougouffa, R. Abdelkhalek, D. Ruau, & D. Keyes, 2024, to appear in Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'24), ACM (Gordon Bell Finalist).
8) *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, to appear in Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'24), ACM (Gordon Bell Finalist).


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

Final question –

Paraphrasing (by three characters) Samuel B. Morse’s first telegraph transmission from Washington, DC to Alfred Vail, on the other side of the test in Baltimore:

W H A T H A T H G R O P P W R O U G H T

.-- .... .- - .... .- - .... --. .-. --- .--. .--. .-- .-. --- ..- --. .... -

..--..

?