UNIVERSITY of York

SC23

Denver, CO | i am hpc.

# Supercomputing 2023

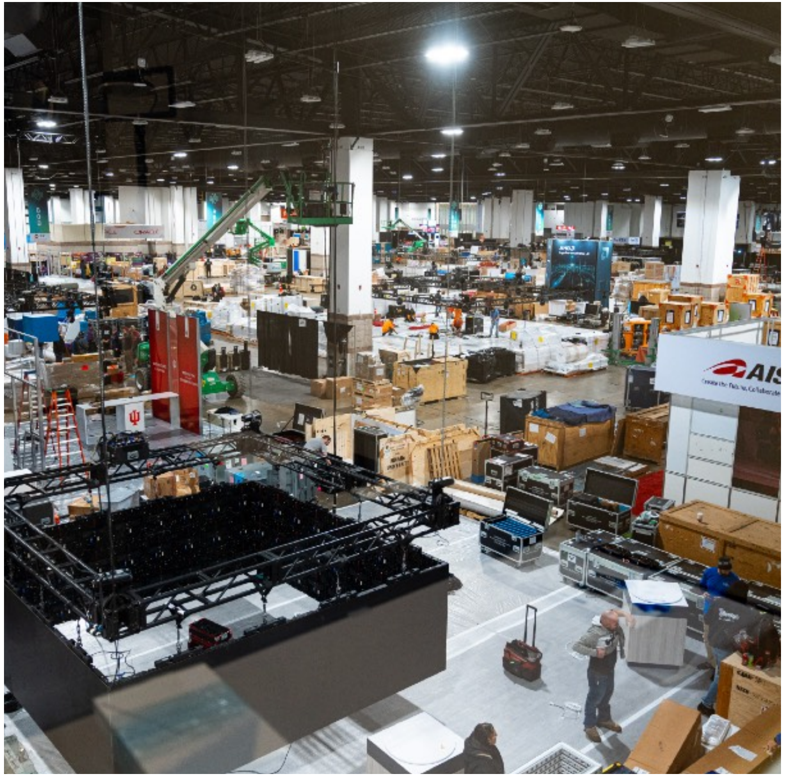
# Supercomputing

- SC is the premier international forum for HPC
- Includes:
  - Birds-of-a-Feather sessions (BoFs)
  - Panels
  - Technical Papers
  - Workshops
  - Tutorials
  - Exhibition

# Supercomputing

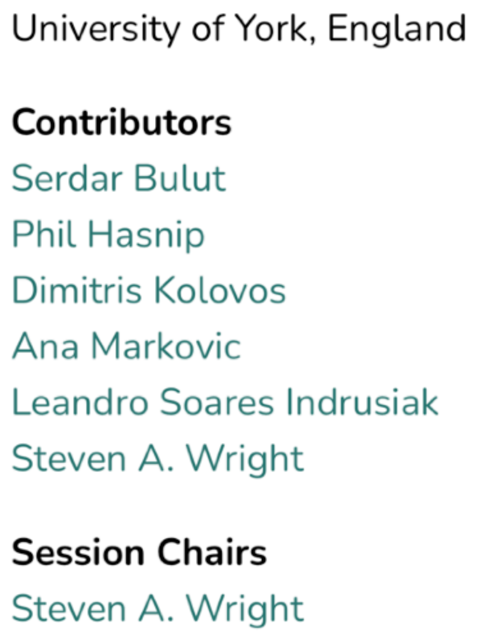
Highlights from 2023

- 14,000+ in-person attendees
  - The most *ever*
- 438 exhibitors
- Theme was *I am HPC*
- 14th PMBS Workshop



# Supercomputing

York at Supercomputing



University of York, England

**Contributors**

Serdar Bulut

Phil Hasnip

Dimitris Kolovos

Ana Markovic

Leandro Soares Indrusiak

Steven A. Wright

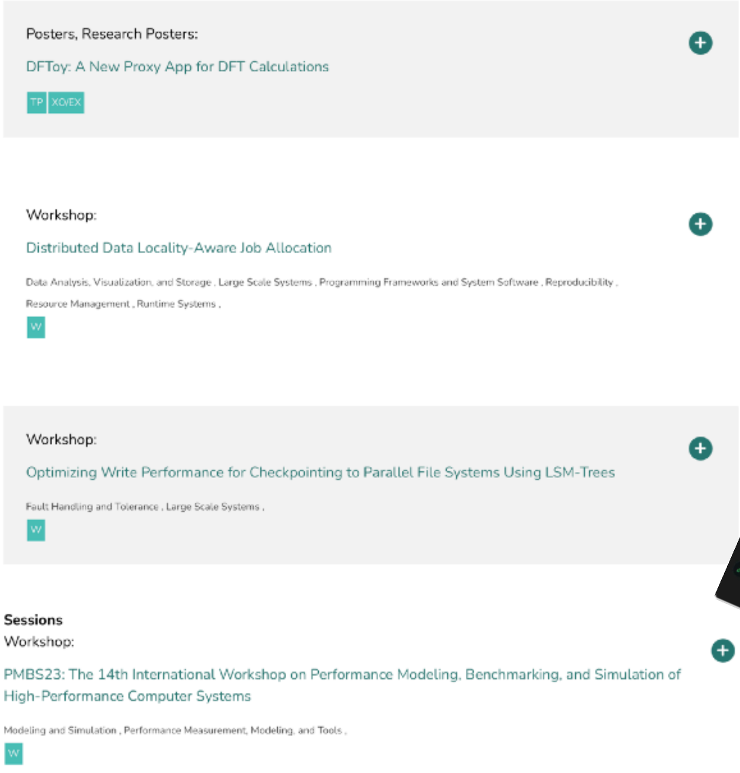
**Session Chairs**

Steven A. Wright

- (Probably…) The most representation from York at Supercomputing
  - 2 in-person attendees, 1 remote (… as far as I'm aware!)
- 1 Poster, 1 Workshop, 2 Workshop papers

# Supercomputing

## York at Supercomputing



Posters, Research Posters:

DFToy: A New Proxy App for DFT Calculations

TP XO/EX

Workshop:

Distributed Data Locality-Aware Job Allocation

Data Analysis, Visualization, and Storage , Large Scale Systems , Programming Frameworks and System Software , Reproducibility , Resource Management , Runtime Systems ,

W

Workshop:

Optimizing Write Performance for Checkpointing to Parallel File Systems Using LSM-Trees

Fault Handling and Tolerance , Large Scale Systems ,

W

Sessions

Workshop:

PMBS23: The 14th International Workshop on Performance Modeling, Benchmarking, and Simulation of High-Performance Computer Systems

Modeling and Simulation , Performance Measurement, Modeling, and Tools ,

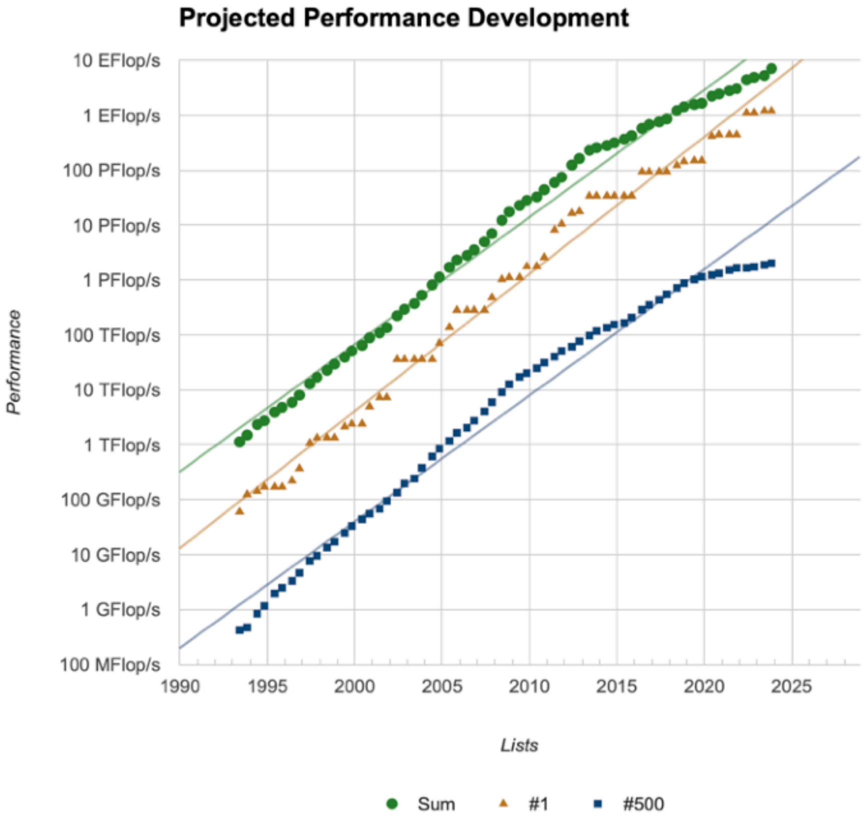
W

DFToy:
A new proxy app for DFT calculations
UNIVERSITY OF CAMBRIDGE
UNIVERSITY of York
SC23
Denver, CO
...tributed Data Locality-Aware Job Allocation
...Markovic
...kovic, Dimitris Kolovos, Leandro Soares Indrusiak
Optimizing Write Performance for Che...
Parallel File Systems Using LSM-Trees
Serdar Bulut, Steven A. Wright
University of York
14th International Workshop on
Performance Modeling, Benchmarking and Simulation
of High Performance Computer Systems
held in conjunction with SC23: The International Conference for High Performance Computing, Networking, Storage and Analysis

# The Top 500

- The Top 500 Supercomputers list is updated biannually, May (at ISC) and November (at SC)
- At SC, a Birds-of-a-Feather session announces the list, awards the top machines and summarises trends

# The Top 500



- **Frontier** is still the only *acknowledged* Exascale system (1.1 EFLOP/s)
- Europe now has ~~two~~ three Top ~~5~~ 10 systems!
  - LUMI (~~310~~ 380 PFLOP/s)
  - Leonardo (~~175~~ 240 PFLOP/s)
  - MareNostrum 5 (140 PFLOP/s)
- Top 10 systems contribute >50% the sum performance (~7 EFLOP/s)
  - We have about 10 very big supercomputers and 490 others!

# The Top 500

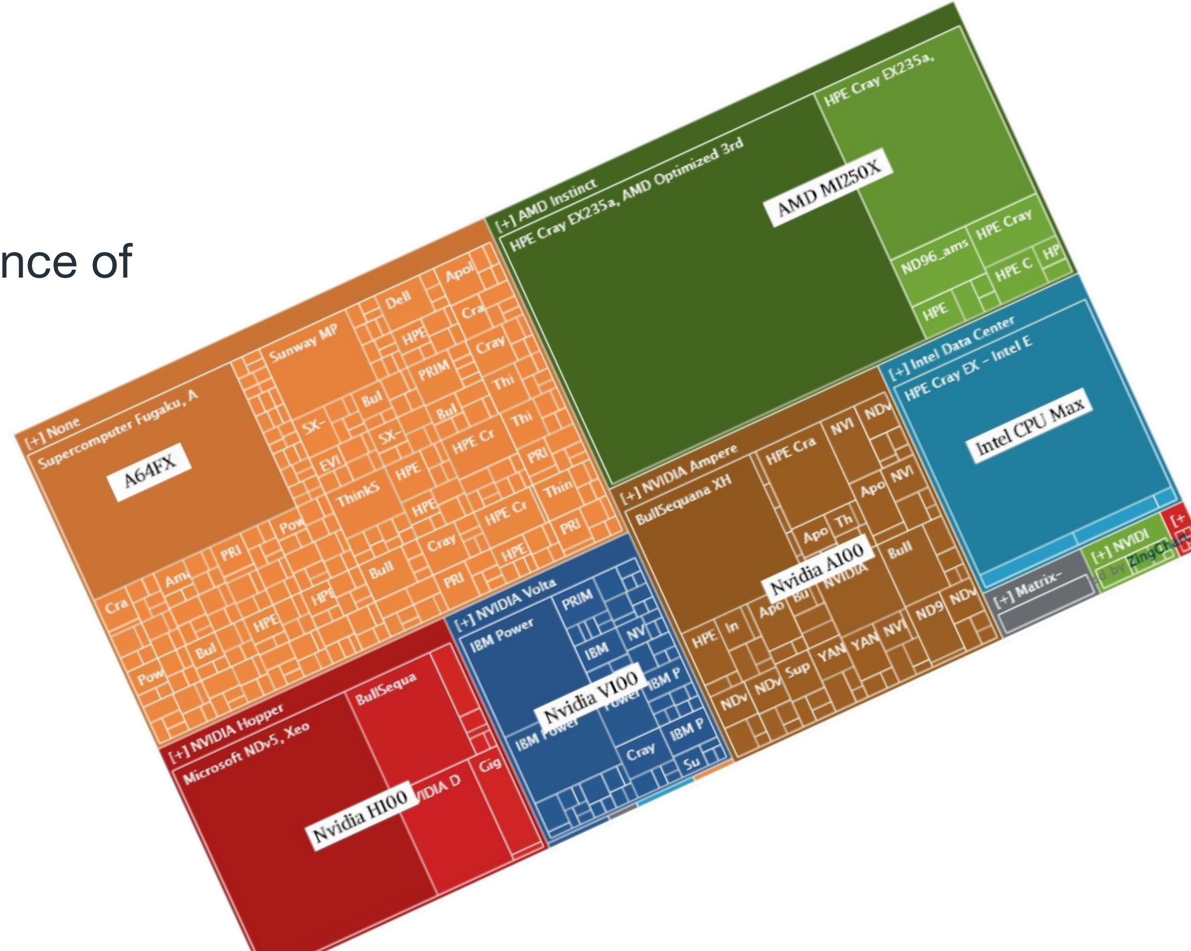
## Big Surprises

- Aurora was expected to unseat Frontier, with estimated peak ~2 EFLOP/s
  - However, only half the machine was benchmarked, achieving #2 with 585 PFLOP/s
- #3 system is an Microsoft Azure cloud instance with NVIDIA H100 GPUs, achieving 561 PFLOP/s

# The Top 500

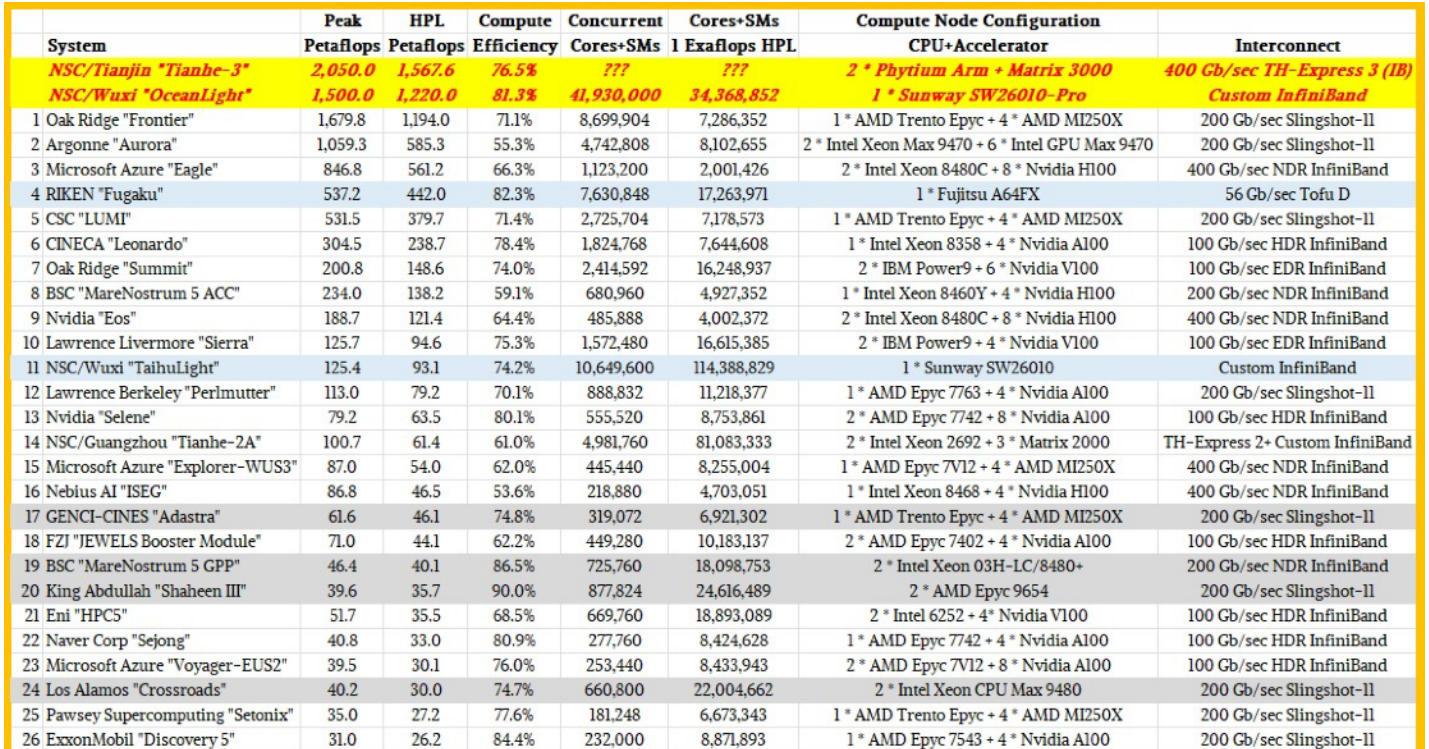
## Architectures

- A treemap of the Top 500 demonstrates the dominance of large systems, and of accelerators



# The Top 500

## The “Unofficial” List



| | | Peak | HPL | Compute | Concurrent | Cores+SMs | Compute Node Configuration | |
|---|---|---|---|---|---|---|---|---|
| | System | Petaflops | Petaflops | Efficiency | Cores+SMs | 1 Exaflops HPL | CPU+Accelerator | Interconnect |
| | NSC/Tianjin "Tianhe-3" | 2,050.0 | 1,567.6 | 76.5% | ??? | ??? | 2 * Phytium Arm + Matrix 3000 | 400 Gb/sec TH-Express 3 (IB) |
| | NSC/Wuxi "OceanLight" | 1,500.0 | 1,220.0 | 81.3% | 41,930,000 | 34,368,852 | 1 * Sunway SW26010-Pro | Custom InfiniBand |
| 1 | Oak Ridge "Frontier" | 1,679.8 | 1,194.0 | 71.1% | 8,699,904 | 7,286,352 | 1 * AMD Trento Epyc + 4 * AMD MI250X | 200 Gb/sec Slingshot-11 |
| 2 | Argonne "Aurora" | 1,059.3 | 585.3 | 55.3% | 4,742,808 | 8,102,655 | 2 * Intel Xeon Max 9470 + 6 * Intel GPU Max 9470 | 200 Gb/sec Slingshot-11 |
| 3 | Microsoft Azure "Eagle" | 846.8 | 561.2 | 66.3% | 1,123,200 | 2,001,426 | 2 * Intel Xeon 8480C + 8 * Nvidia H100 | 400 Gb/sec NDR InfiniBand |
| 4 | RIKEN "Fugaku" | 537.2 | 442.0 | 82.3% | 7,630,848 | 17,263,971 | 1 * Fujitsu A64FX | 56 Gb/sec Tofu D |
| 5 | CSC "LUMI" | 531.5 | 379.7 | 71.4% | 2,725,704 | 7,178,573 | 1 * AMD Trento Epyc + 4 * AMD MI250X | 200 Gb/sec Slingshot-11 |
| 6 | CINECA "Leonardo" | 304.5 | 238.7 | 78.4% | 1,824,768 | 7,644,608 | 1 * Intel Xeon 8358 + 4 * Nvidia A100 | 100 Gb/sec HDR InfiniBand |
| 7 | Oak Ridge "Summit" | 200.8 | 148.6 | 74.0% | 2,414,592 | 16,248,937 | 2 * IBM Power9 + 6 * Nvidia V100 | 100 Gb/sec EDR InfiniBand |
| 8 | BSC "MareNostrum 5 ACC" | 234.0 | 138.2 | 59.1% | 680,960 | 4,927,352 | 1 * Intel Xeon 8460Y + 4 * Nvidia H100 | 200 Gb/sec NDR InfiniBand |
| 9 | Nvidia "Eos" | 188.7 | 121.4 | 64.4% | 485,888 | 4,002,372 | 2 * Intel Xeon 8480C + 8 * Nvidia H100 | 400 Gb/sec NDR InfiniBand |
| 10 | Lawrence Livermore "Sierra" | 125.7 | 94.6 | 75.3% | 1,572,480 | 16,615,385 | 2 * IBM Power9 + 4 * Nvidia V100 | 100 Gb/sec EDR InfiniBand |
| 11 | NSC/Wuxi "TaihuLight" | 125.4 | 93.1 | 74.2% | 10,649,600 | 114,388,829 | 1 * Sunway SW26010 | Custom InfiniBand |
| 12 | Lawrence Berkeley "Perlmutter" | 113.0 | 79.2 | 70.1% | 888,832 | 11,218,377 | 1 * AMD Epyc 7763 + 4 * Nvidia A100 | 200 Gb/sec Slingshot-11 |
| 13 | Nvidia "Selene" | 79.2 | 63.5 | 80.1% | 555,520 | 8,753,861 | 2 * AMD Epyc 7742 + 8 * Nvidia A100 | 100 Gb/sec HDR InfiniBand |
| 14 | NSC/Guangzhou "Tianhe-2A" | 100.7 | 61.4 | 61.0% | 4,981,760 | 81,083,333 | 2 * Intel Xeon 2692 + 3 * Matrix 2000 | TH-Express 2+ Custom InfiniBand |
| 15 | Microsoft Azure "Explorer-WUS3" | 87.0 | 54.0 | 62.0% | 445,440 | 8,255,004 | 1 * AMD Epyc 7V12 + 4 * AMD MI250X | 400 Gb/sec NDR InfiniBand |
| 16 | Nebius AI "ISEG" | 86.8 | 46.5 | 53.6% | 218,880 | 4,703,051 | 1 * Intel Xeon 8468 + 4 * Nvidia H100 | 400 Gb/sec NDR InfiniBand |
| 17 | GENCI-CINES "Adastra" | 61.6 | 46.1 | 74.8% | 319,072 | 6,921,302 | 1 * AMD Trento Epyc + 4 * AMD MI250X | 200 Gb/sec Slingshot-11 |
| 18 | FZJ "JEWELS Booster Module" | 71.0 | 44.1 | 62.2% | 449,280 | 10,183,137 | 2 * AMD Epyc 7402 + 4 * Nvidia A100 | 100 Gb/sec HDR InfiniBand |
| 19 | BSC "MareNostrum 5 GPP" | 46.4 | 40.1 | 86.5% | 725,760 | 18,098,753 | 2 * Intel Xeon 03H-LC/8480+ | 200 Gb/sec NDR InfiniBand |
| 20 | King Abdullah "Shaheen III" | 39.6 | 35.7 | 90.0% | 877,824 | 24,616,489 | 2 * AMD Epyc 9654 | 200 Gb/sec Slingshot-11 |
| 21 | Eni "HPC5" | 51.7 | 35.5 | 68.5% | 669,760 | 18,893,089 | 2 * Intel 6252 + 4* Nvidia V100 | 100 Gb/sec HDR InfiniBand |
| 22 | Naver Corp "Sejong" | 40.8 | 33.0 | 80.9% | 277,760 | 8,424,628 | 1 * AMD Epyc 7742 + 4 * Nvidia A100 | 100 Gb/sec HDR InfiniBand |
| 23 | Microsoft Azure "Voyager-EUS2" | 39.5 | 30.1 | 76.0% | 253,440 | 8,433,943 | 2 * AMD Epyc 7V12 + 8 * Nvidia A100 | 100 Gb/sec HDR InfiniBand |
| 24 | Los Alamos "Crossroads" | 40.2 | 30.0 | 74.7% | 660,800 | 22,004,662 | 2 * Intel Xeon CPU Max 9480 | 200 Gb/sec Slingshot-11 |
| 25 | Pawsey Supercomputing "Setonix" | 35.0 | 27.2 | 77.6% | 181,248 | 6,673,343 | 1 * AMD Trento Epyc + 4 * AMD MI250X | 200 Gb/sec Slingshot-11 |
| 26 | ExxonMobil "Discovery 5" | 31.0 | 26.2 | 84.4% | 232,000 | 8,871,893 | 1 * AMD Epyc 7543 + 4 * Nvidia A100 | 200 Gb/sec Slingshot-11 |

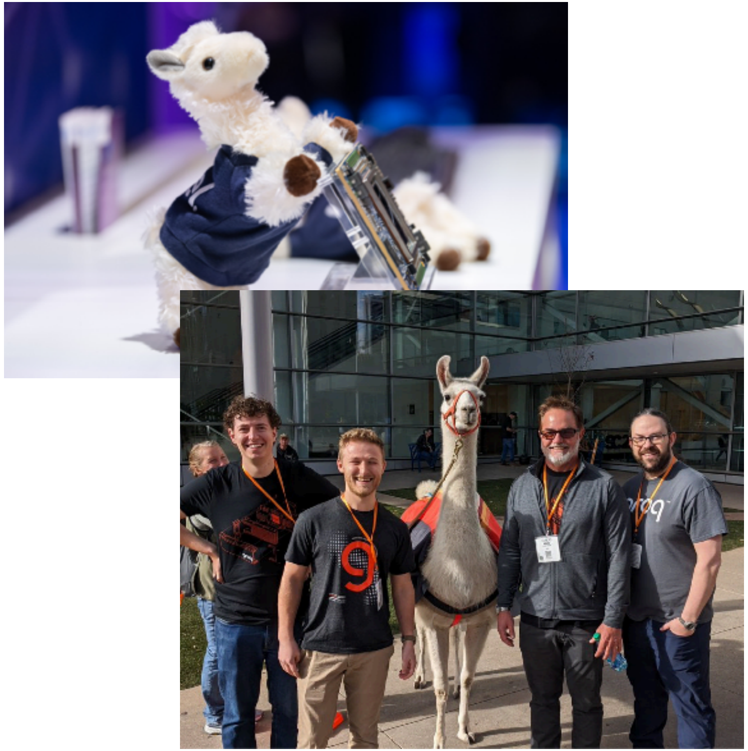
# HPC Systems in the UK

(…in the Top 100)

- #39 ARCHER2, still top UK system, 19.54 PF
  - AMD CPUs
- #41 Dawn, University of Cambridge, 19.46 PF
  - Xeon Sapphire Rapids + Xe-HPC Ponte Vecchio
- #79 Cambridge-1, 9.68 PF
  - AMD CPUs + NVIDIA A100

- Just before SC, University of Bristol announced Isambard-AI
  - £225m investment in NVIDIA GH200

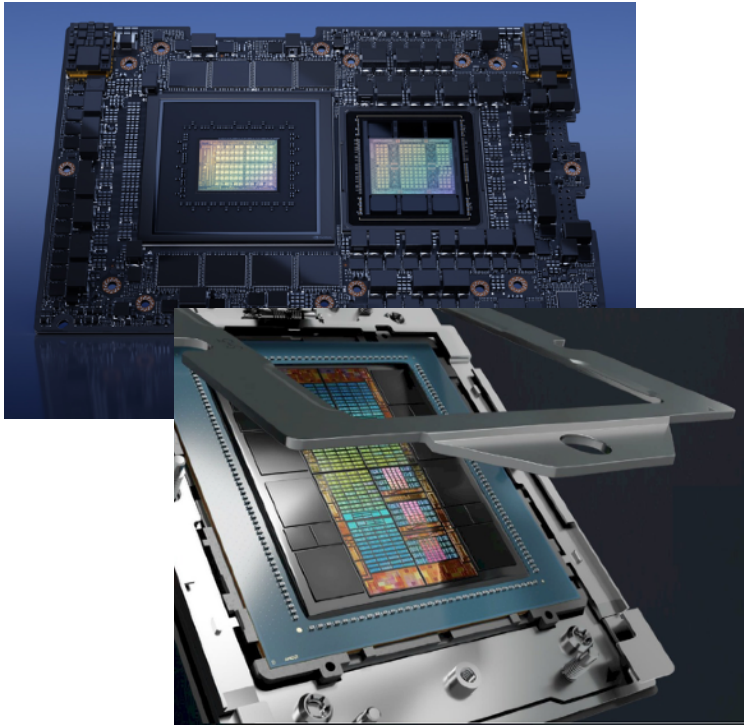
# Conference Themes



- SC always has a varied programme, but the “big” theme this year was:
  - LLMs! (at least according to exhibitors)
  - APUs (Accelerated Processing Units)
  - DAOS
  - HPSF

# New(?) Architectures

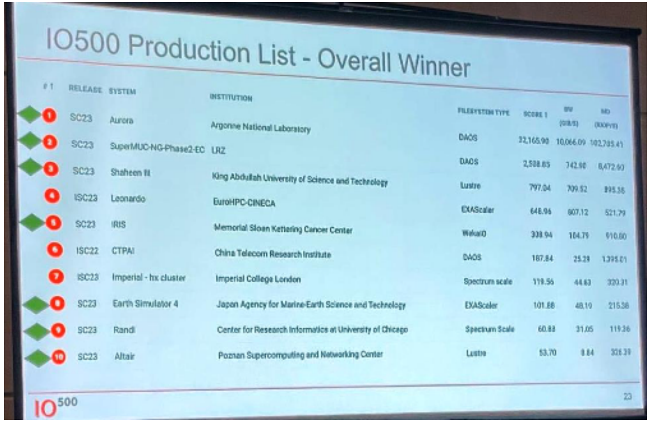
APUs and Superchips (and XPUs)



- APUs and Superchips combine a CPU and GPU on a single die
- Evolution of Summit/Sierra architecture
  - Essentially gain cache coherence and unified memory for CPU and GPU
  - CPU cores can handle things GPUs are bad at (I/O, divergence, etc)

- Intel abandoned their “XPU” in May
- AMD have MI300A in the works
- NVIDIA announced GH200 at SC

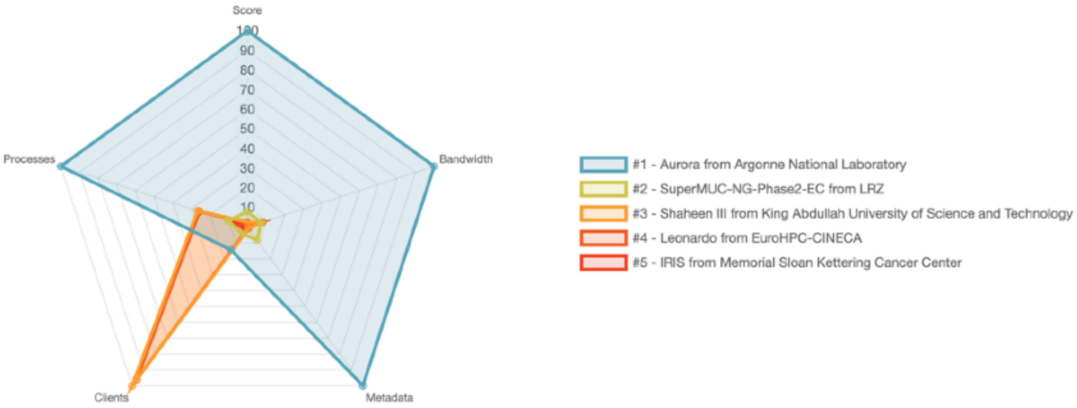
# DAOS



IO500 Production List - Overall Winner

| # ↑ | RELEASE | SYSTEM | INSTITUTION | FILESYSTEM TYPE | SCORE ↑ | BW (GIB/S) | MD (KIOPS) |
|---|---|---|---|---|---|---|---|
| 1 | SC23 | Aurora | Argonne National Laboratory | DAOS | 32,165.90 | 10,066.09 | 102,785.41 |
| 2 | SC23 | SuperMUC-NG-Phase2-EC | LRZ | DAOS | 2,508.85 | 742.90 | 8,472.60 |
| 3 | SC23 | Shaheen III | King Abdullah University of Science and Technology | Lustre | 797.04 | 709.52 | 895.35 |
| 4 | ISC23 | Leonardo | EuroHPC-CINECA | EXAScaler | 648.96 | 807.12 | 521.79 |
| 5 | SC23 | IRIS | Memorial Sloan Kettering Cancer Center | WekaIO | 338.94 | 104.79 | 910.60 |
| 6 | ISC22 | CTPAI | China Telecom Research Institute | DAOS | 187.84 | 25.28 | 1,395.01 |
| 7 | ISC23 | Imperial - hx cluster | Imperial College London | Spectrum scale | 119.56 | 44.63 | 320.31 |
| 8 | SC23 | Earth Simulator 4 | Japan Agency for Marine-Earth Science and Technology | EXAScaler | 101.88 | 48.19 | 215.38 |
| 9 | SC23 | Randi | Center for Research Informatics at University of Chicago | Spectrum Scale | 60.88 | 31.05 | 119.36 |
| 10 | SC23 | Altair | Poznan Supercomputing and Networking Center | Lustre | 53.70 | 8.84 | 326.39 |

- In HPC I/O, DAOS shines
  - Aurora has fastest (production) I/O system in town
  - Many papers in workshops about DAOS performance



# HPSF

High Performance Software Foundation

- The Linux Foundation launched the HPSF
- Initial projects:
  - Spack, Kokkos, AMReX, WarpX, Trilinos, Apptainer, VTK-m, HPCToolkit, E4S, Charliecloud
- Membership:
  - AWS, HPE, Intel, NVIDIA, CEA, Kitware, Uni of Oregon, CIQ, Various DoE labs

- hpsfoundation.github.io

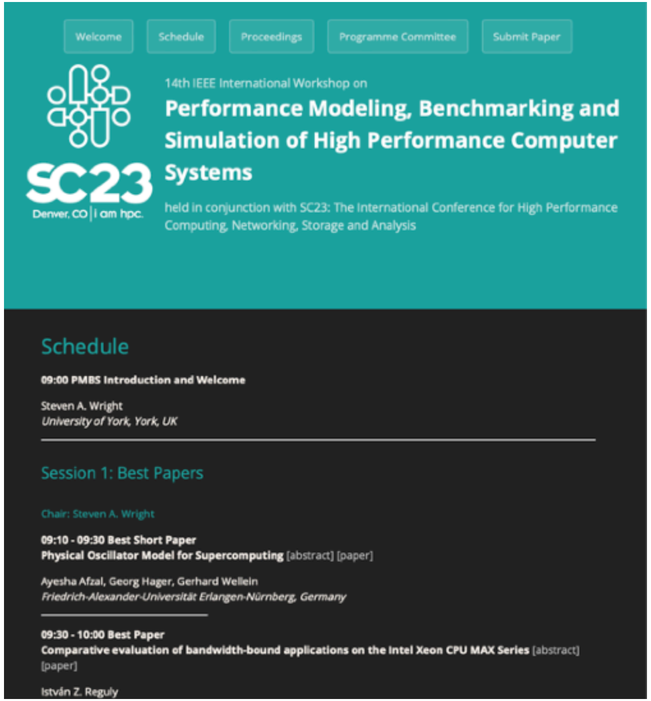
# Performance Modeling, Benchmarking and Simulation

- 14th Year of PMBS
- PMBS is concerned with the evaluation and comparison of HPC systems and applications primarily through:
  - Analytical performance modeling
  - Benchmarking and performance analysis
  - Use of advanced simulation techniques

# Performance Modeling, Benchmarking and Simulation

- Published 186 novel research papers at PMBS
- This year we accepted:
  - 10 full-length papers
  - 4 short paper

# Highlights



14th IEEE International Workshop on

**Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems**

held in conjunction with SC23: The International Conference for High Performance Computing, Networking, Storage and Analysis

Schedule

**09:00 PMBS Introduction and Welcome**

Steven A. Wright
*University of York, York, UK*

Session 1: Best Papers

Chair: Steven A. Wright

**09:10 - 09:30 Best Short Paper**
**Physical Oscillator Model for Supercomputing** [abstract] [paper]

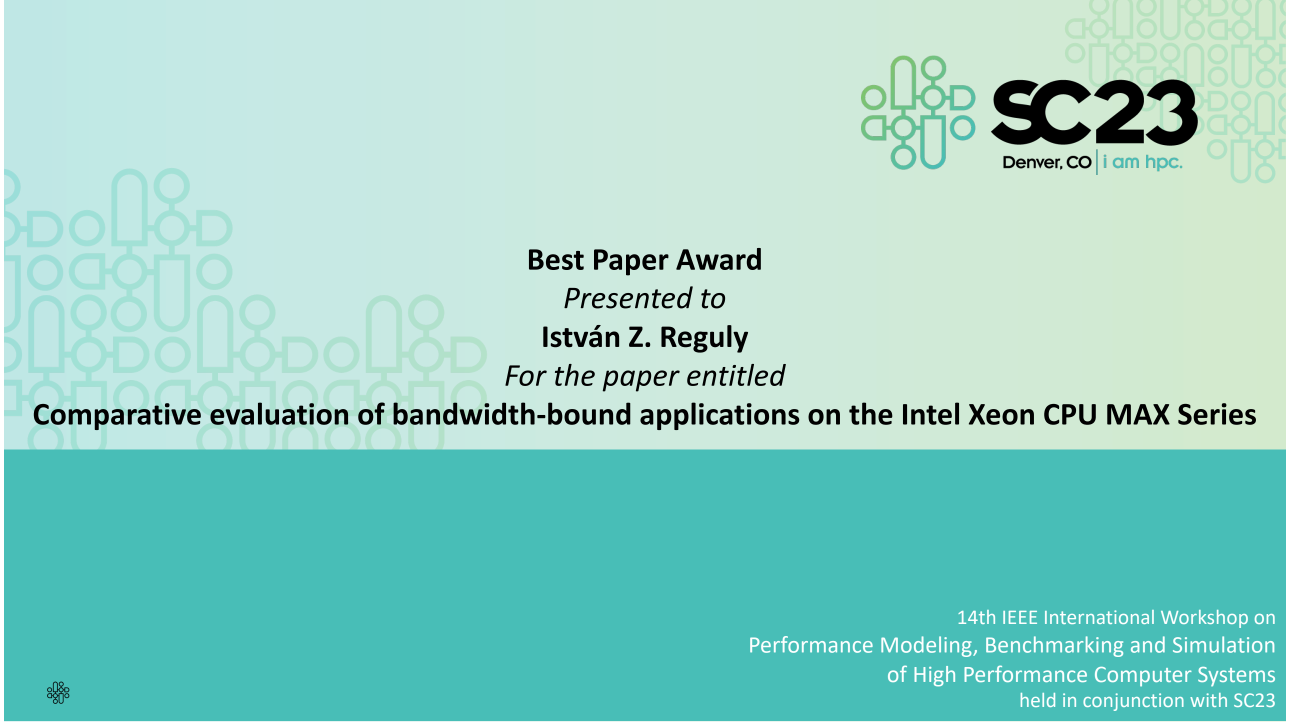
Ayesha Afzal, Georg Hager, Gerhard Wellein
*Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany*

**09:30 - 10:00 Best Paper**
**Comparative evaluation of bandwidth-bound applications on the Intel Xeon CPU MAX Series** [abstract] [paper]

István Z. Reguly

- Sessions:
  - Best Papers
  - Architecture Evaluations
  - Short Papers
  - Benchmarking
  - Scheduling
  - Performance Modeling

# Performance Modeling, Benchmarking and Simulation



SC23

Denver, CO | i am hpc.

**Best Paper Award**

*Presented to*

**István Z. Reguly**

*For the paper entitled*

**Comparative evaluation of bandwidth-bound applications on the Intel Xeon CPU MAX Series**

14th IEEE International Workshop on
Performance Modeling, Benchmarking and Simulation
of High Performance Computer Systems
held in conjunction with SC23

# Best Paper

Comparative evaluation of Intel Xeon CPU MAX

- Intel Xeon CPU MAX is a “fat” x86 CPU architecture with on-chip High Bandwidth Memory (HBM)

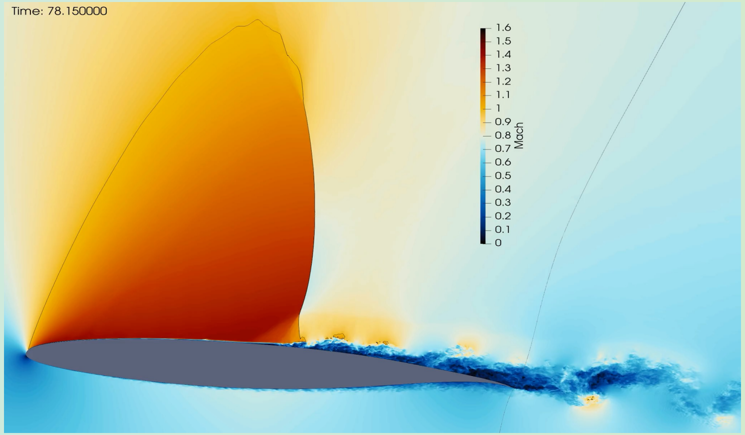
- Xeon CPU MAX 9480
  - 56 cores (1.9-2.6 GHz)
  - 64 GB HBM2e
  - 4 NUMA regions
  - Dual socket

# Best Paper

Comparative evaluation of Intel Xeon CPU MAX

- The Competition:
  - AMD EPYC 7V73X
    - 60 cores (2.2-3.5 GHz)
    - 768 MB L3, 448 GB DDR4
  - Intel Xeon Platinum 8360Y
    - 36 cores (2.4-2.8 GHz)
    - 512 GB DDR4

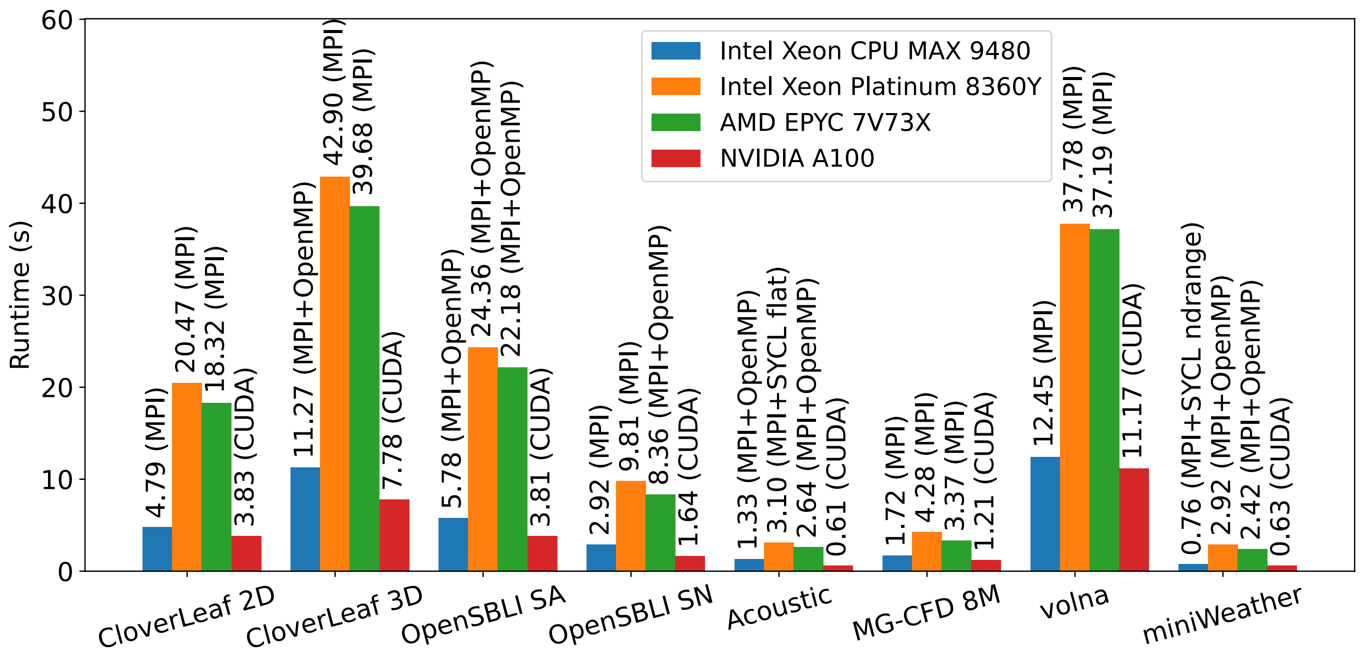
# Applications



- Test suite (mostly) based on OPS/OP2 DSL apps
  - Structured mesh stencil codes (varying computational intensity)
  - Unstructured mesh codes
  - Test harness to streamline compilation & runs: https://github.com/reguly/tests
- CloverLeaf 2D/3D – low order + lots of small boundary loops (DP)
- Acoustic – high order, cache-intensive (SP)
- OpenSBLI – more data movement (SA), more recompute (SN) versions (DP)
- miniWeather – atmospheric dynamics, low order (DP)
- MG-CFD – lots of indirect accesses, data races (DP)
- Volna – fewer computations with indirections/races (SP)
- +miniBUDE – compute/latency intensive (SP)

# Comparison of best parallelizations

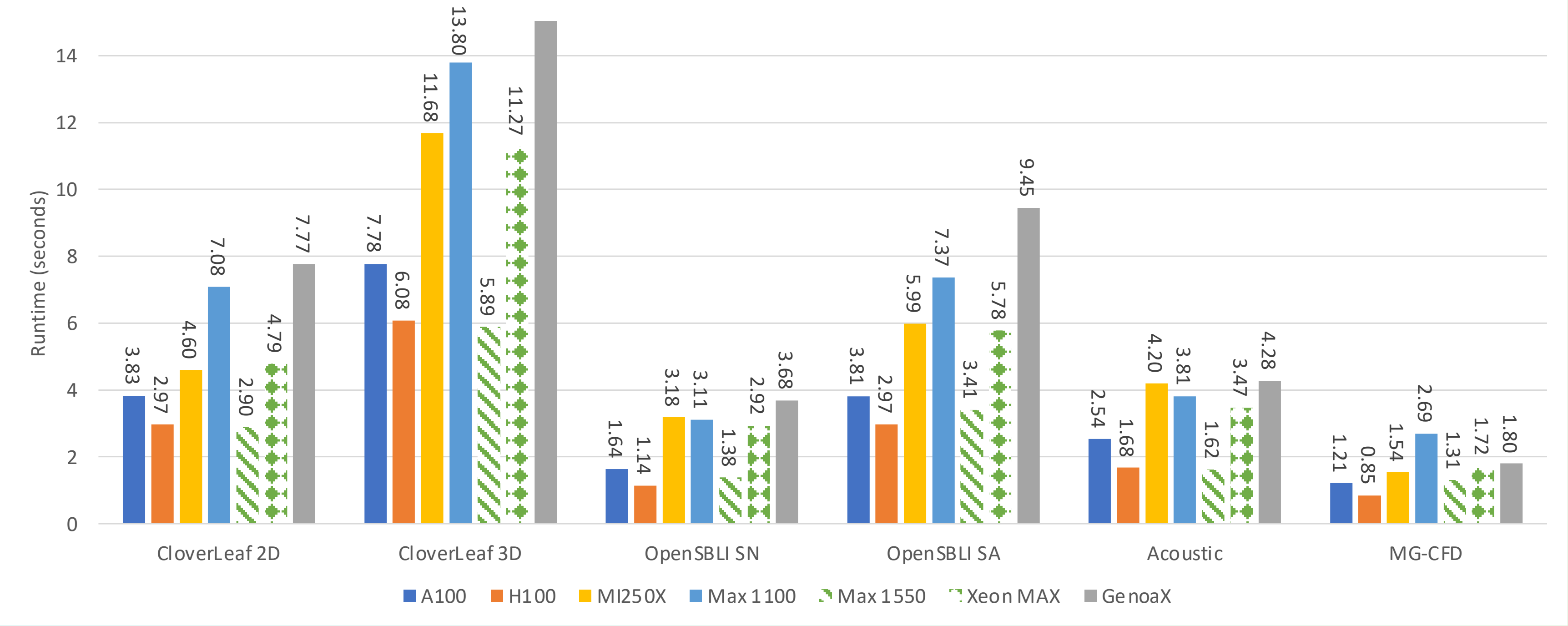
- CloverLeaf – most BW bound. 3.5-4.3x
- OpenSBLI SN/Acoustic – cache & latency. 2-3.3x
- MG-CFD/volna - lantecy. 2-3x
- miniBUDE – compute, latency 1.36-1.8x
- Vs. A100: 1.1-2.2x slower
  - No MPI comms on GPU



Speedup relative to Intel Xeon Platinum 8360Y and AMD EPYC 7V73X

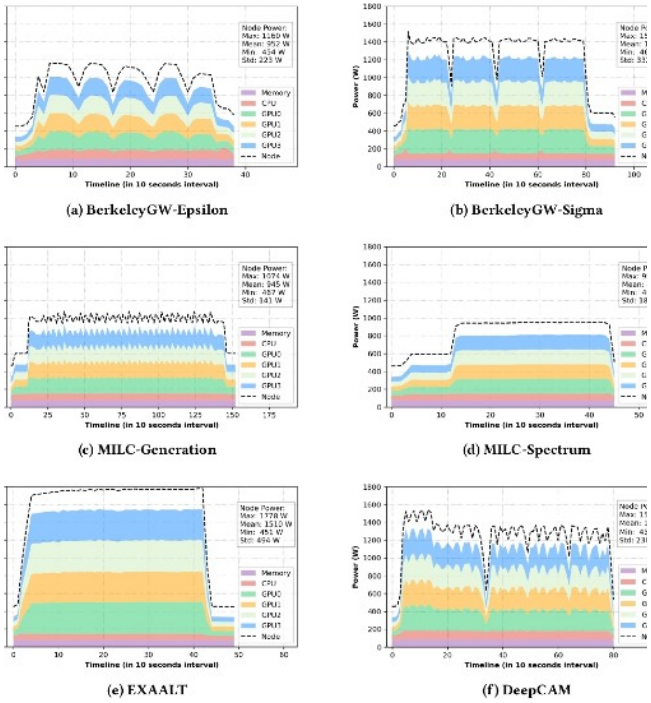
| | CloverLeaf 2D | CloverLeaf 3D | OpenSBLI SA | OpenSBLI SN | Acoustic | MG-CFD 8M | volna | miniWeather | miniBUDE |
|---|---|---|---|---|---|---|---|---|---|
| 8360Y | 4.27 | 3.81 | 4.21 | 3.36 | 2.33 | 2.49 | 3.03 | 3.82 | 1.88 |
| 7V73X | 3.82 | 3.52 | 3.83 | 2.86 | 1.98 | 1.95 | 2.99 | 3.17 | 1.36 |

# Comparison to more CPUs & GPUs



# Other Notable Papers

## Power Analysis of NERSC Workloads



(a) BerkeleyGW-Epsilon

(b) BerkeleyGW-Sigma

(c) MILC-Generation

(d) MILC-Spectrum

(e) EXAALT

(f) DeepCAM

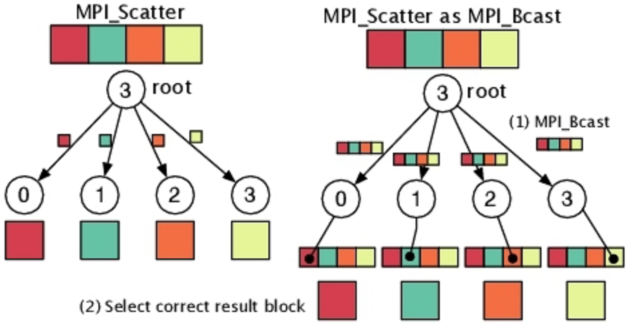
- Paper analyses the power characteristics of NERSC production workloads
  - Large gap between average and peak power usage
  - Large swing in power during application (with CPU/GPU applications)

- Z. Zhao, et al. 2023. Power Analysis of NERSC Production Workloads., 10.1145/3624062.3624200

# Other Notable Papers

Verifying Performance Guidelines for MPI Collectives

- Paper analyses performance guidelines for MPI collectives
  - Propose a benchmarking tool to test performance guidelines (e.g. MPI_Scatter ≤ MPI_Bcast)
  - Demonstrate that in many cases, MPI libraries require optimisation (because they fail some tests!)



- S. Hunold. 2023. Verifying Performance Guidelines for MPI Collectives at Scale., 10.1145/3624062.3625532