TACC
TEXAS ADVANCED COMPUTING CENTER

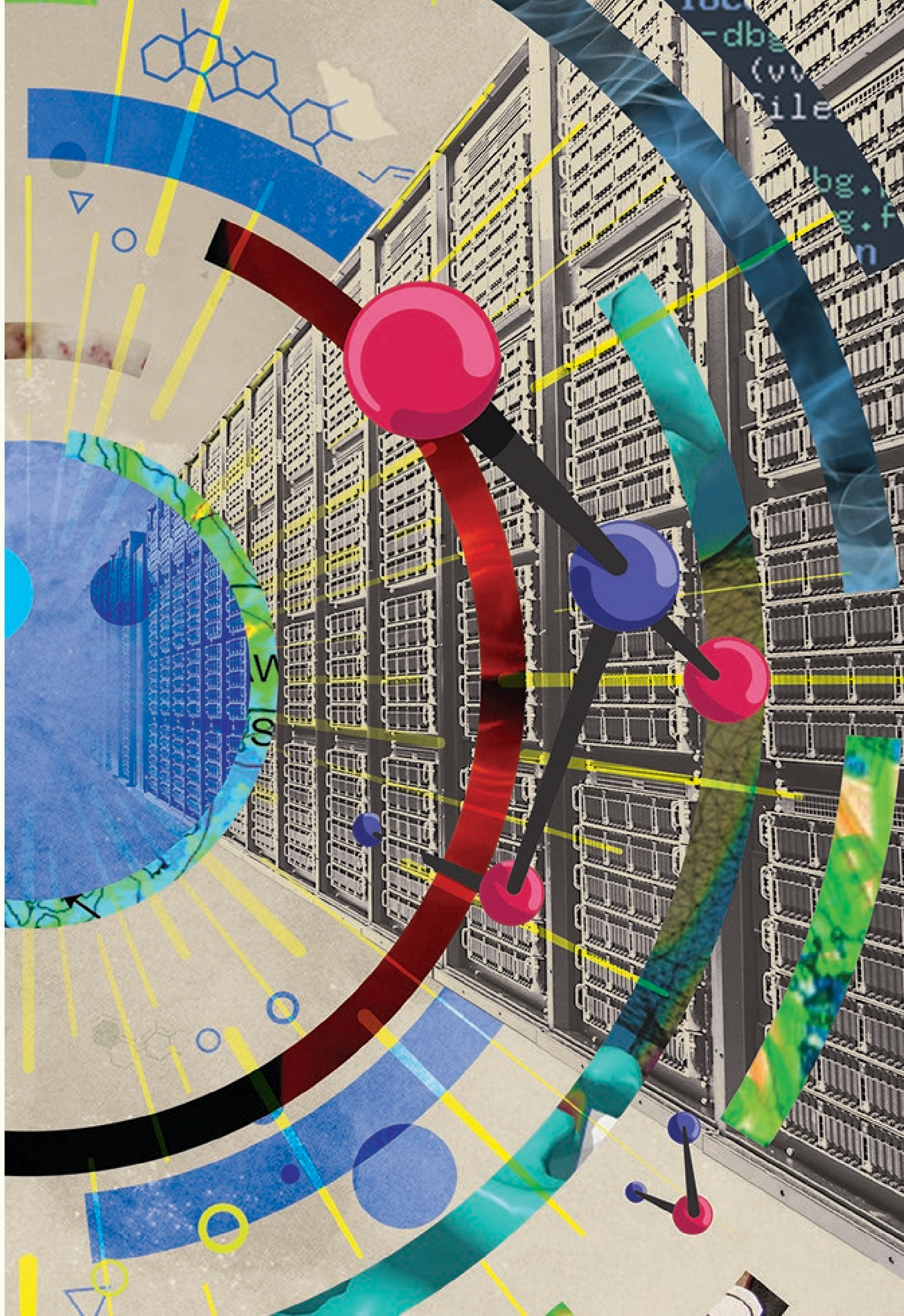
TEXAS
The University of Texas at Austin

Characteristic Science Applications

Amit Ruhela, PH.D.

Manager – HPC Tools, Texas Advanced Computing Center

Austin Texas



About TACC

The Texas Advanced Computing Center (TACC) at The University of Texas at Austin is the leading academic supercomputing center in the country.

TACC delivers world-class, innovative systems, tools, software, and expertise to researchers who seek to make an impact in the world, and advance discovery across disciplines.

TACC IN A NUTSHELL

- Founded in 2001 with a mission to enable discoveries that advance science and society through computing, collaboration, and education to ensure the power of advanced computing technologies is available to all.
- 190 Staff (~70 PhD)
- Facilitates Frontera, Stampede3, Lonestar6, Vista, Jetstream, and Chameleon systems for the National Science Foundation (NSF)
- Altogether, ~12k Nodes, ~1M CPU cores, ~1k GPUs
- About seven billion core hours over several million jobs per year
  - for 3,000 projects and ~40,000 users per year.
  - Frontera (60K) Lonestar6 (52K), and Stampede3 (90K) jobs per month

Systems at TACC

Frontera

Stampede3

Lonestar6

Vista

Need for New Systems

1. Next-gen processor architectures
2. Increases workload demands
3. Newer workloads

TACC Compute Hardware

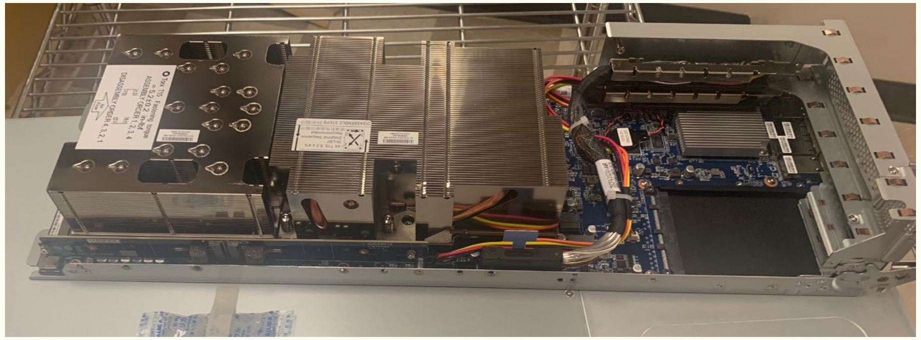
Resource	CPU type	#Nodes/Sockets/Cores	GPU Type	# GPUs
Frontera	Xeon (Cascade Lake)	8400/16800/470,400	RTX (Volta)	360
Lonestar-6	AMD Epyc	600/1200/76,800	NV A100	255
Stampede-3	Xeon (Sapphire Rapids)	2,024/4,048/150,080	Intel PVC	80
Vista	**ARM/Grace**	**840/1080/77,760**	**NV H100**	**600**
Horizon	*ARM*	Close to a million	Embargo	Thousands

Vista

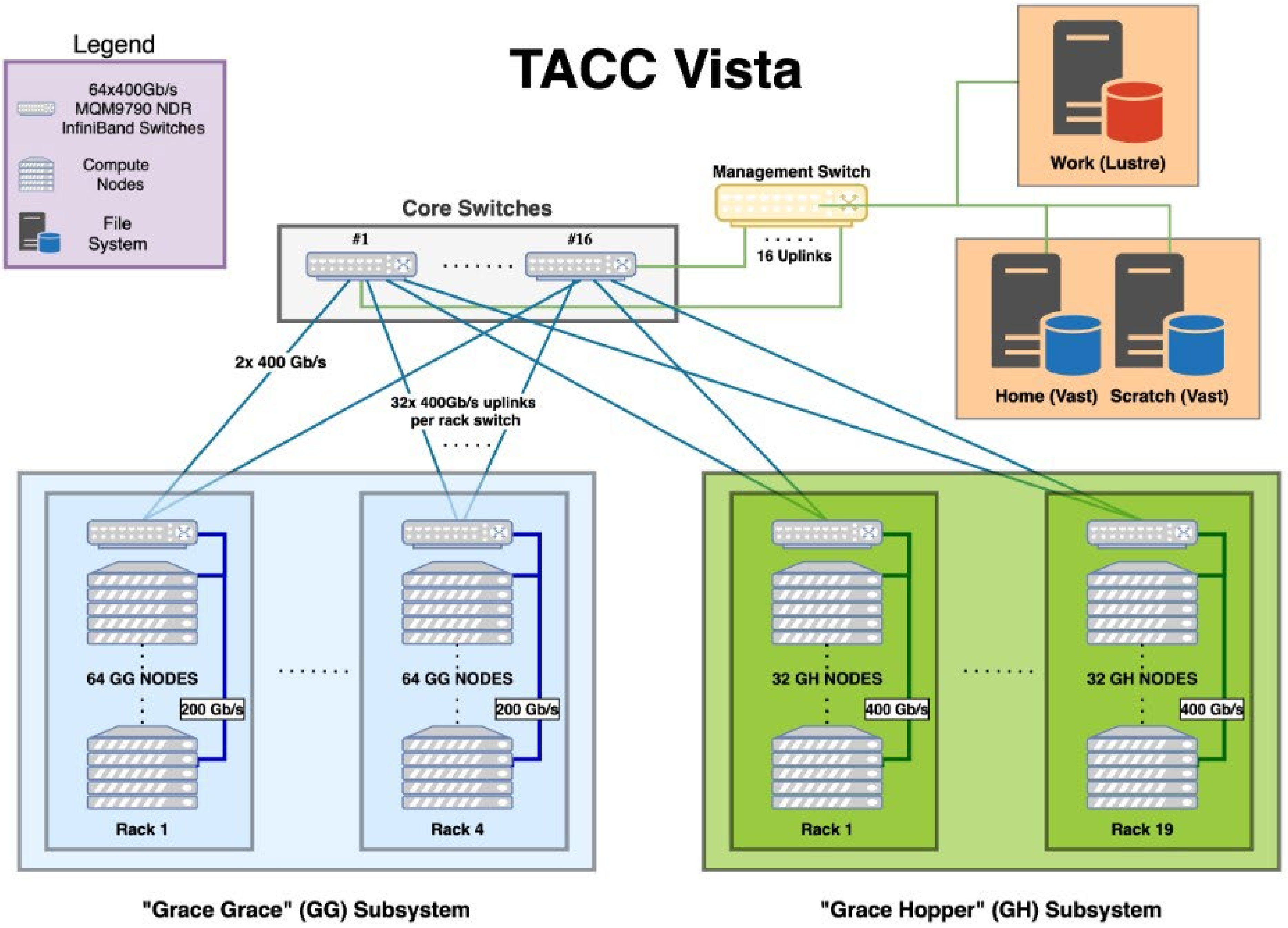
- Vista is a new AI-centric resource.
- Vista is half-funded as a supplement to Frontera, and half by UT-Austin AI initiative
- Vista is a bridge to Horizon.
- And Vista is a couple of firsts for TACC:
    - Our first system with an ARM as the primary CPU.
    - Our first system with NVIDIA as the primary chip (and interconnect) provider.
- While AI was in mind, we still have a strong scientific computing focus in how we v
  use it.

VISTA Hardware

- 256 Grace-Grace (GG) CPU nodes (144 cores(72+72), 3.1Ghz clock rate), 7.1 TF FP64 Performance
- 600 Grace-Hopper (GH) H100 nodes (1 CPU, 1 GPU).
  - 34 TF FP64
  - 67 TF FP64 Tensor Core
  - 990 TF FP16 Tensor Core
  - 1979 TF F8, Tensor Core

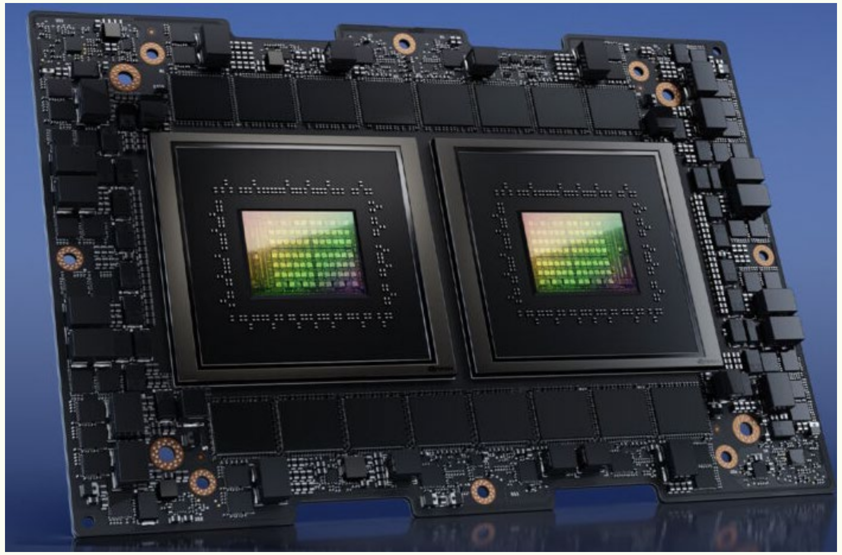


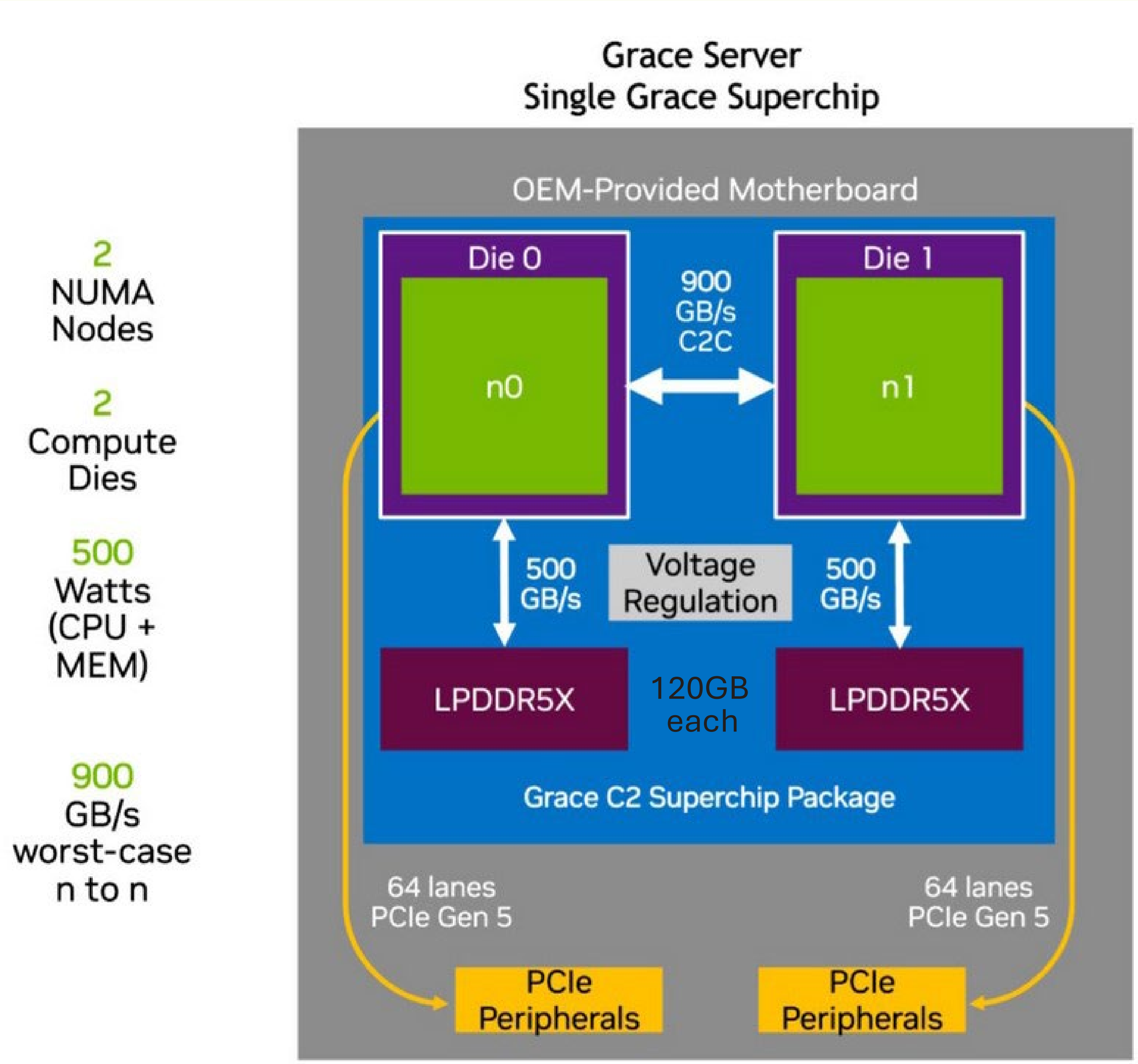
- Grace-Grace : 240GB of LPDDR5X RAM, 512 GB Local disk
- Grace-Hopper : 120GB of LPDDR5X RAM, 96GB HBM3(Hopper), 512 GB Local disk
- Network : Non-blocking NDR InfiniBand fat tree (200Gb/sec (GG) and 400Gb/sec (GH)).
- 15PB VAST Storage (shared 30PB storage pool with Stampede3).
- Rocky 9.3 (Blue Onyx)



Legend
64x400Gb/s MQM9790 NDR InfiniBand Switches
Compute Nodes
File System
TACC Vista
Work (Lustre)
Management Switch
Core Switches
#1
#16
16 Uplinks
Home (Vast)
Scratch (Vast)
2x 400 Gb/s
32x 400Gb/s uplinks per rack switch
64 GG NODES
200 Gb/s
Rack 1
64 GG NODES
200 Gb/s
Rack 4
32 GH NODES
400 Gb/s
Rack 1
32 GH NODES
400 Gb/s
Rack 19
"Grace Grace" (GG) Subsystem
"Grace Hopper" (GH) Subsystem

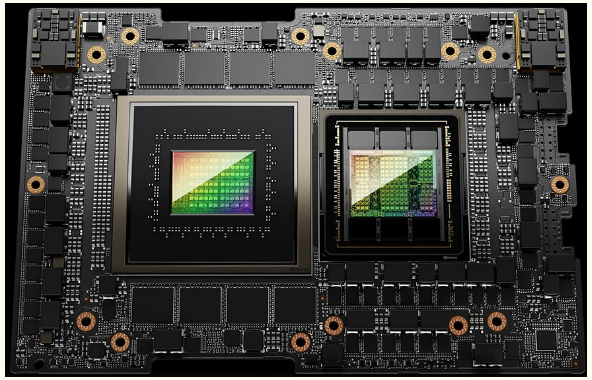
NVIDIA Grace Grace (GG)

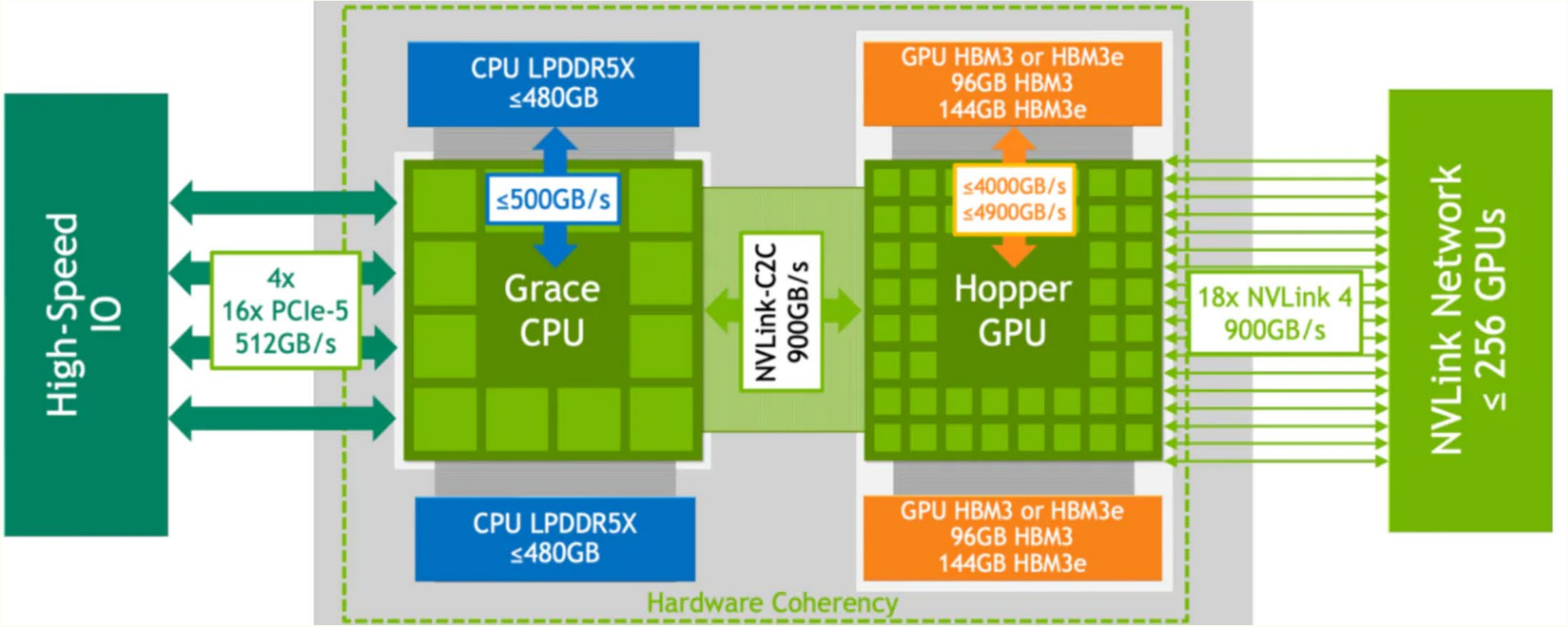




Courtesy : NVIDIA

NVIDIA Grace Hopper Superchip (GH200)





Courtesy : NVIDIA

Software Configurations (Latest configurations)

1. NVIDIA HPC SDK : 25.5
2. GNU : 13.2, 14.2, 15.1
3. UCX 1.18.1
4. CUDA 12.9
5. MPI : OpenMPI 5.0.8 MVAPICH-PLUS 4.0.0 HPC-X (bundled in SDK, v4)
6. Profilers
   a. NVIDIA NSIGHT Systems and Compute
   b. GDB, Remora
7. Containers Runtime : *Apptainer and Charliecloud*
8. Other modules :
   a) Gromacs, Lammps, Hypre, NAMD, NWCHEM, VASP, Trilinos
   b) Petsc, adios2, hdf5, phdf5, netcdf, pnetcdf, boost, eigen …

Benchmarking Methodology

Past Benchmarking : Frontera

Standard Benchmarking includes

1. MPI benchmarks
2. IO benchmarks
3. Stream
4. HPL
5. Key application workloads

From the solicitation:

- Use the SPP Benchmark + some microbenchmarks and reliability measures
- Target 2-3x Blue Waters (at 1/3 budget) --- 6-9x performance improvement per $ vs. 7 years ago.

The SPP was defined in 2006.

- Most of the codes still relevant (WRF,MILC, NWChem)
- Some are obsolete
- The *problem sizes* are no longer sufficient for measuring the full capabilities of the machine (though some still pushed us to ~5,000 nodes/250,000 cores).

Application Acceptance Tests

Application	Acceptance Threshold[s]	Frontera Time[s]	% over Threshold	Improvement over Blue Waters	Threshold Node[#]	Frontera Node[#]
AWP-ODC	335	326	1.03	**3.2**	1366	1366
CACTUS	1753	1433	1.22	**3.3**	2400	2400
MILC	1364	831	1.64	**9.5**	1296	1296
NAMD	62	60	1.03	**4.0**	2500	2500
NWChem	8053	6408	1.26	**3.8**	5000	1536
PPM	2540	2167	1.17	**3.6**	5000	4828
PSDNS	769	544	1.41	**2.8**	3235	2048
QMCPACK	916	332	2.76	**5.5**	2500	2500
RMG	2410	2307	1.04	**3.2**	700	686
VPIC	1170	981	1.19	**4.3**	4608	4096
WRF	749	635	1.18	**5.2**	4560	4200
Caffe	1203	1044	1.15	**3.2**	1024	1024

Average runtime improvement vs. Blue Waters: **4.3**

Characteristic Science Applications (CSA)

CSAs were initiated with the following three elements

- Application – science code or workflow
- Challenge problem – problem that cannot be readily solved today
- Figure of Merit (F.O.M.) – measure of performance of the application

The goal is to achieve an F.O.M. improvement of 10x

Performance of An App

We have essentially four factors in Application Performance:

- Did the runtime change? (An analog to Strong Scaling – run the same problem in less time).
- Did the problem size change? (An analog to Weak Scaling – run larger problems in fixed time)
- Did we use more or less of the total resource? (An analog to Throughput).
- Did the Physics change? (No good analog).

Note we aren’t *exactly* applying the scaling concepts from ”traditional” benchmarking – a strong scaling plot by definition looks at changes in node counts on a single homogeneous system, but the notion applies.

Performance of An App

We define $\Delta perf_i$, therefore, to be the product of four factors:

- ΔT – The Change in Runtime from Frontera to the new System.
- ΔS – The Change in problem size from Frontera to the new System
- ΔE – (Ensemble) The Change in the fraction of Frontera to the fraction of the new system used to achieve the benchmark.
- ΔP – The Change in physics in an enhanced model (what fraction of operations per datum is added).

$$\Delta perf_i = \Delta T \times \Delta S \times \Delta E \times \Delta P$$

The F.O.M. is a measurement defined for a specific application/workflow that leads to the desired $\Delta perf_i$

Benchmarking Methodology for Horizon

1. Open Call for scientific applications

2. Selecting few representative applications

3. Holistic study of applications performance on variety of architectures

4. Baseline performance and prediction

Characteristic Science Applications

20 applications selected from over 140 submissions

Covered areas:

- Astronomy and Astrophysics
- Biophysics and Biology
- Computational Fluid Dynamics
- Geodynamics and Earth Systems
- Materials Engineering
- Other

Projects ran from 2022Q3 to 2025Q1

Down selected to 11 projects for Application Performance Enhancement

Application Performance Enhancement

Characteristic Science Applications (CSA)

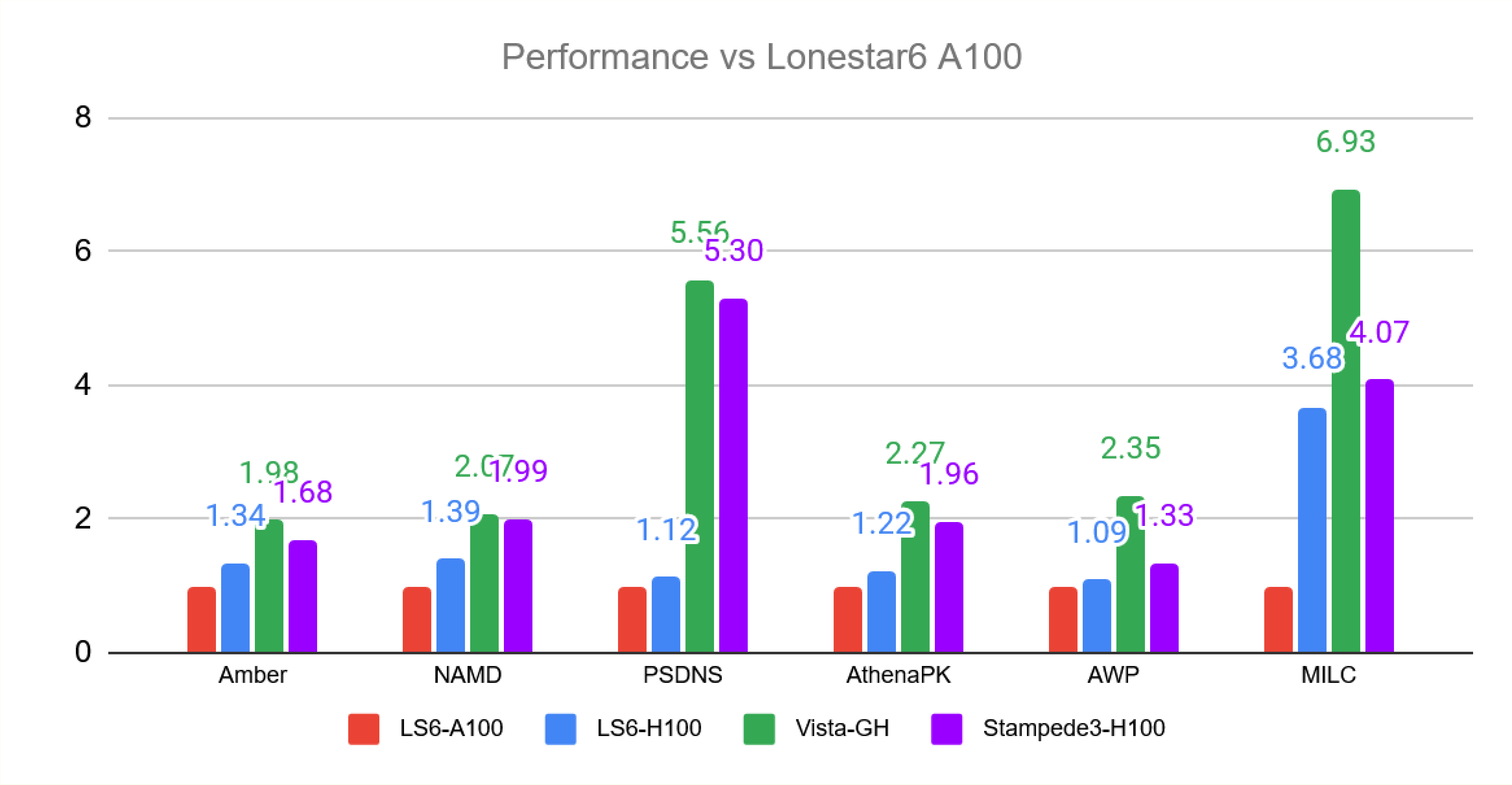
- AWP-ODC
- Athena-K
- Changa
- MILC
- NAMD
- PSDNS
- Seissol
- WESTPA
- EPW
- Enzo-E
- MuST

Performance Results

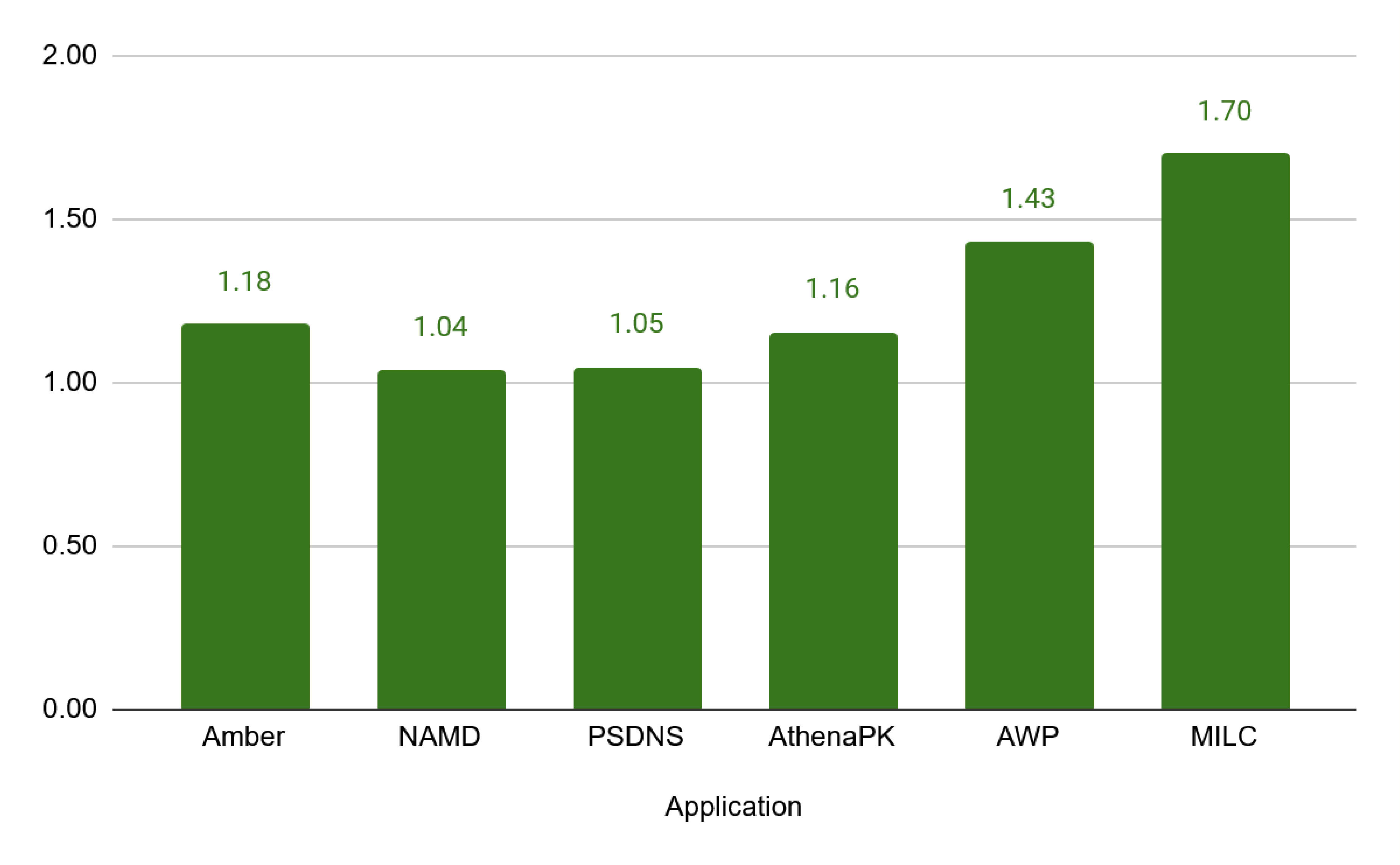
GPUs Comparison

Attribute	A100 PCIe Gen4	H100 PCIe Gen5	GH200	GH200 vs A100	GH200 vs H100
Performance (GFlops/s/card)	9700	25600	34,000	3.5	1.3
MBW (GB/s/card)	1555	2039	4,000	2.6	2.0
Capacity (GB/card)	40	80	96	2.4	1.2
CPU BW (GB/s)	32	64	450	14.1	7.0

Application Benchmarks (gpu)

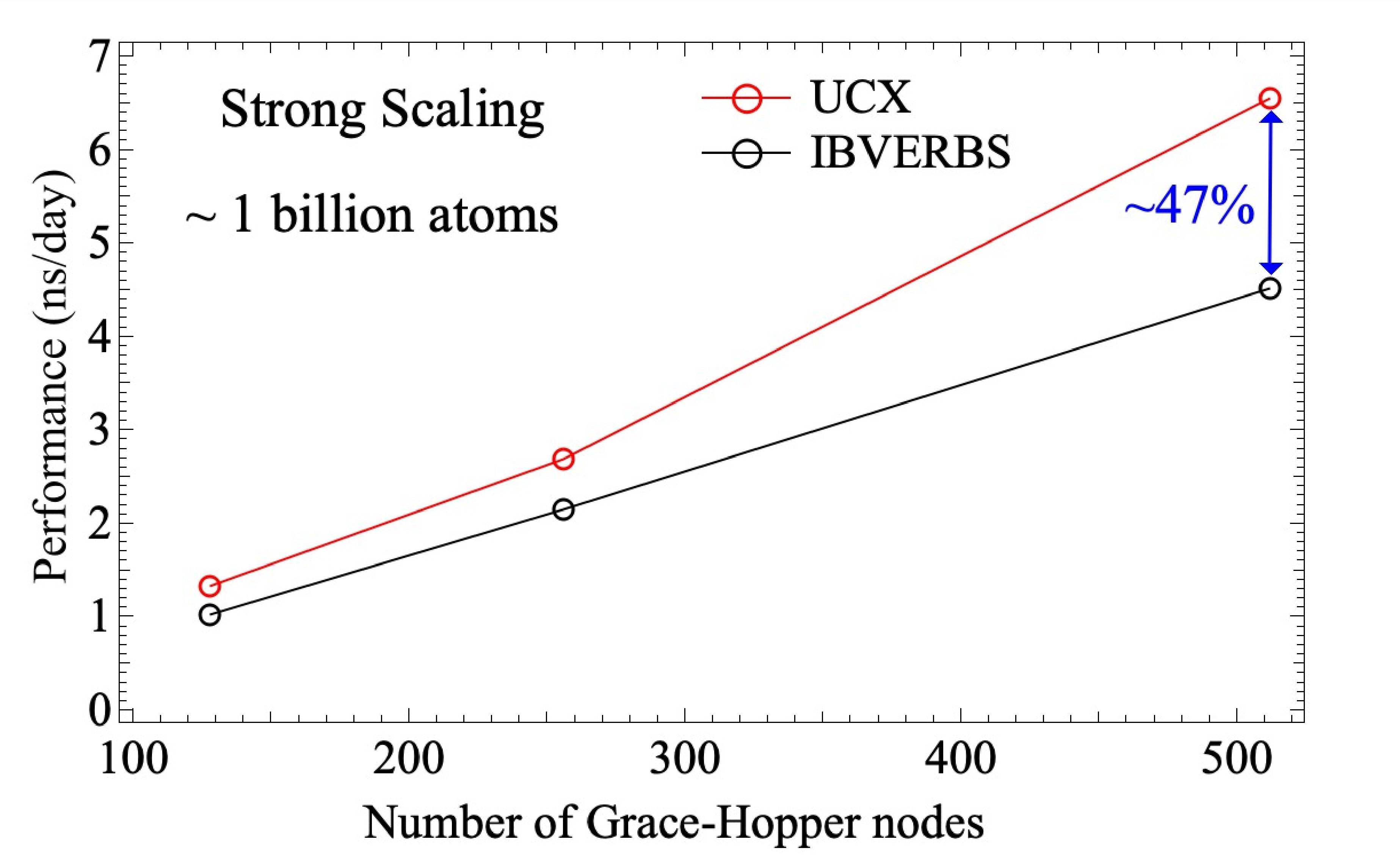


H100 Performance on Vista and Stampede3 systems

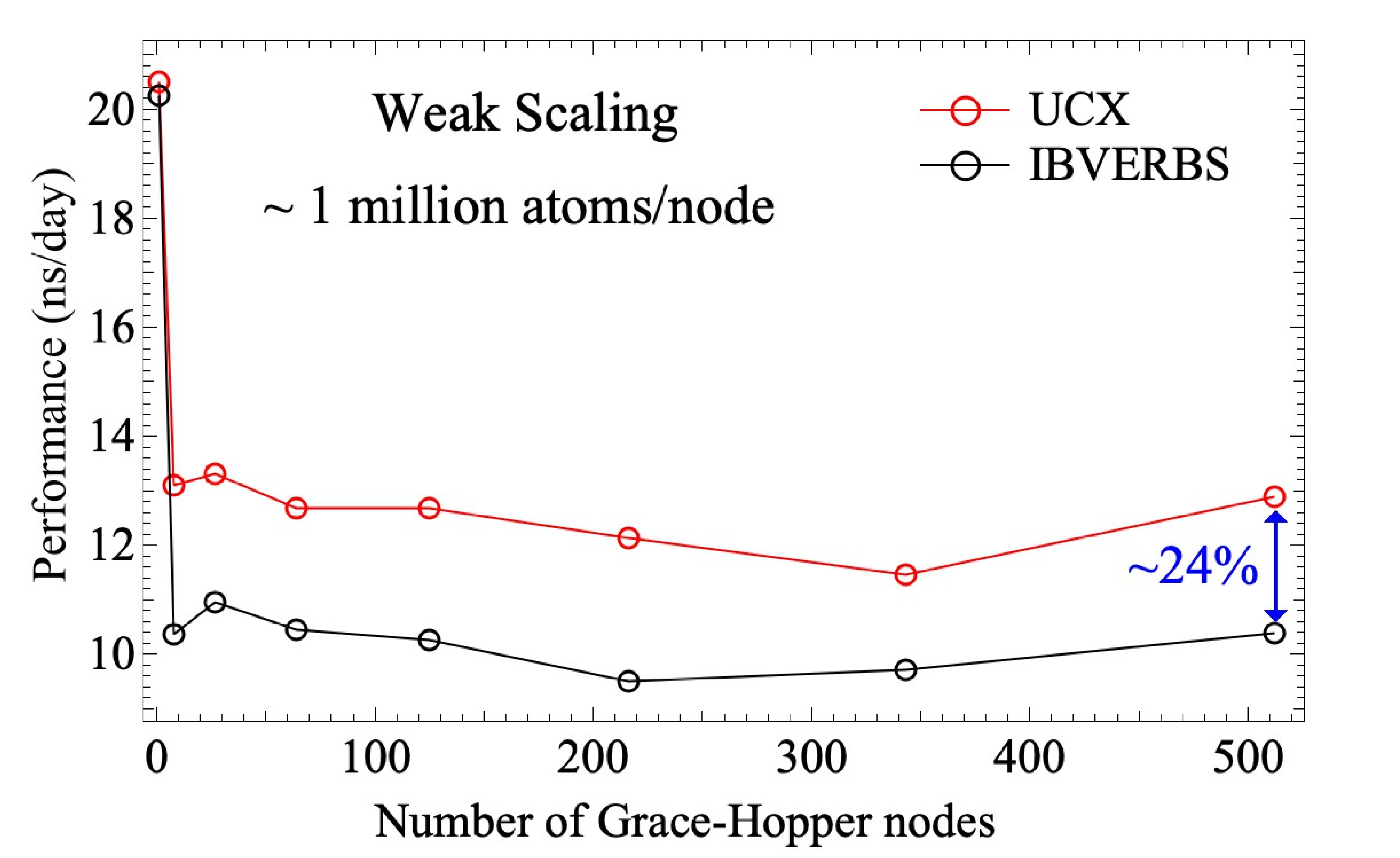


Application Performance at Scale

NAMD : Strong Scaling



NAMD : Weak Scaling



THANK YOU FOR ATTENDING.

E-mail:
aruhela@tacc.utexas.edu

Connect with us: /**taccutexas**