TACC
TEXAS ADVANCED COMPUTING CENTER

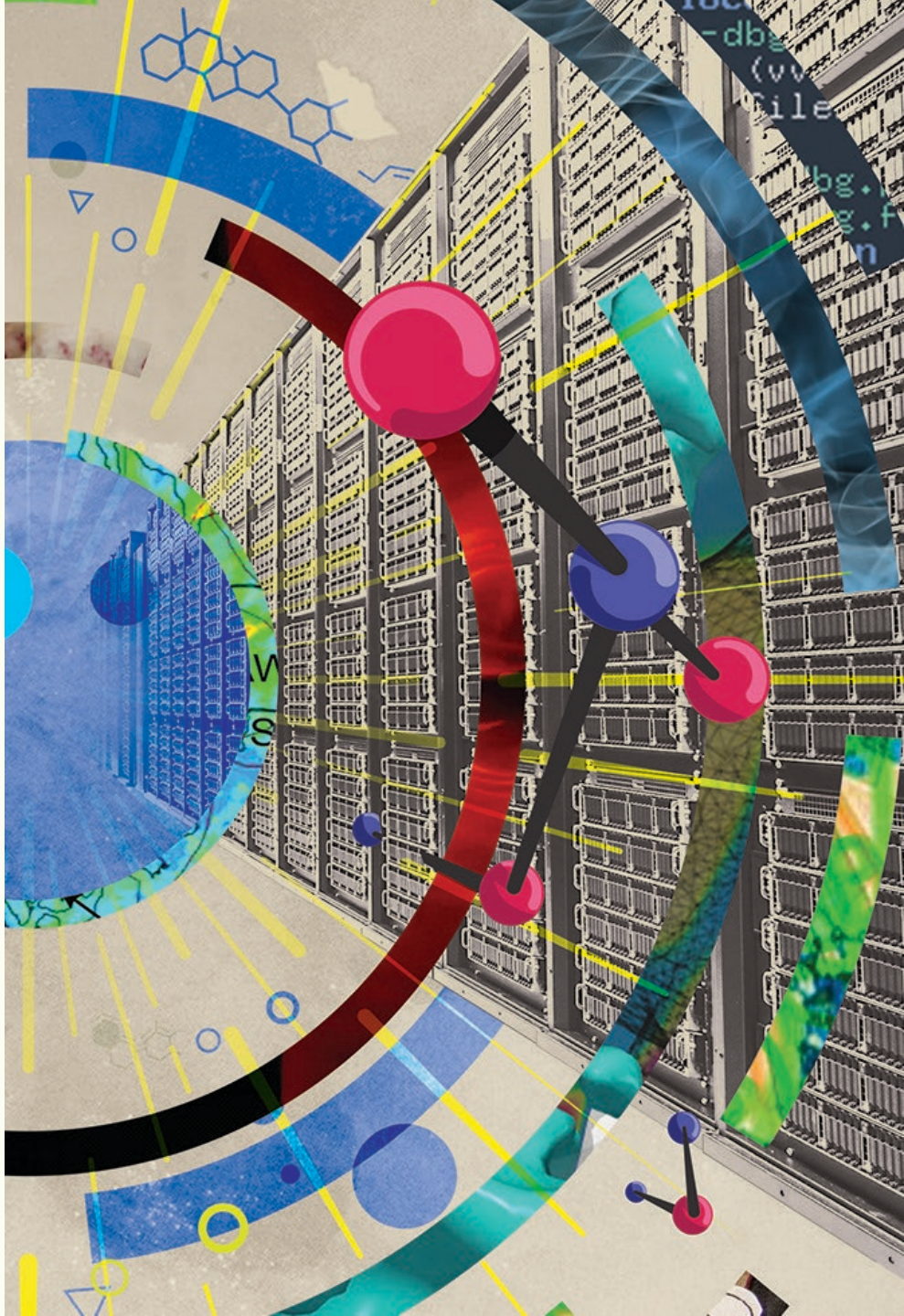
TEXAS
The University of Texas at Austin

Benchmarking for TACC systems and Experiences with Julia

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 multi 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)	8403/16800/470,400	RTX (Volta)	360
Lonestar6	AMD Epyc	600/1200/76,800	NV A100/NV H100	255/8
Stampede3	Xeon (Sapphire Rapids/ Skylake/Ice Lake)	2,024/4,048/150,080	Intel PVC/NV H100	80/96
Vista	**ARM (Grace)**	**827/1078/77,616**	**NV H100**	**576**
Horizon	*ARM (Grace/Vera)*	*6752/11,504/980,352 **	*NV B200*	*4,000*

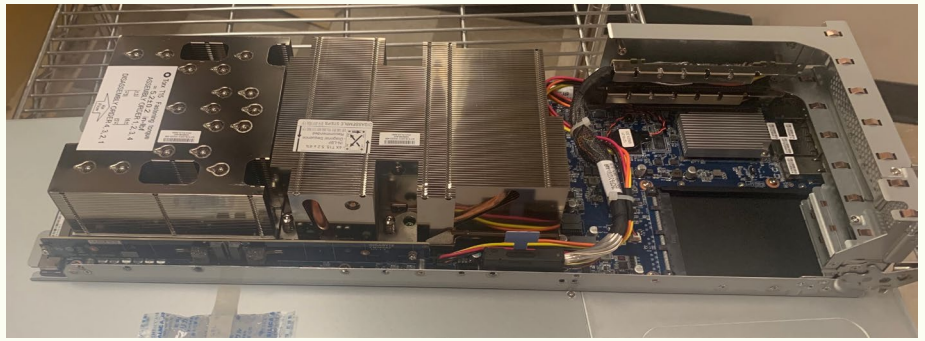
- *2,000 Grace Blackwell Blackwell nodes / 4,752 Vera Vera nodes*
- *Interconnect : CPU – NDR, GPU – XDR*

Vista

- New ML-centric resource
- Funded as a supplement to Frontera and by a UT AI initiative
- Vista is a bridge to Horizon
- Firsts for TACC:
  - First system with ARM as the primary CPU
  - First system with NVIDIA as the primary chip provider for both CPU and GPU
- System design was focused on supporting ML and scientific computing

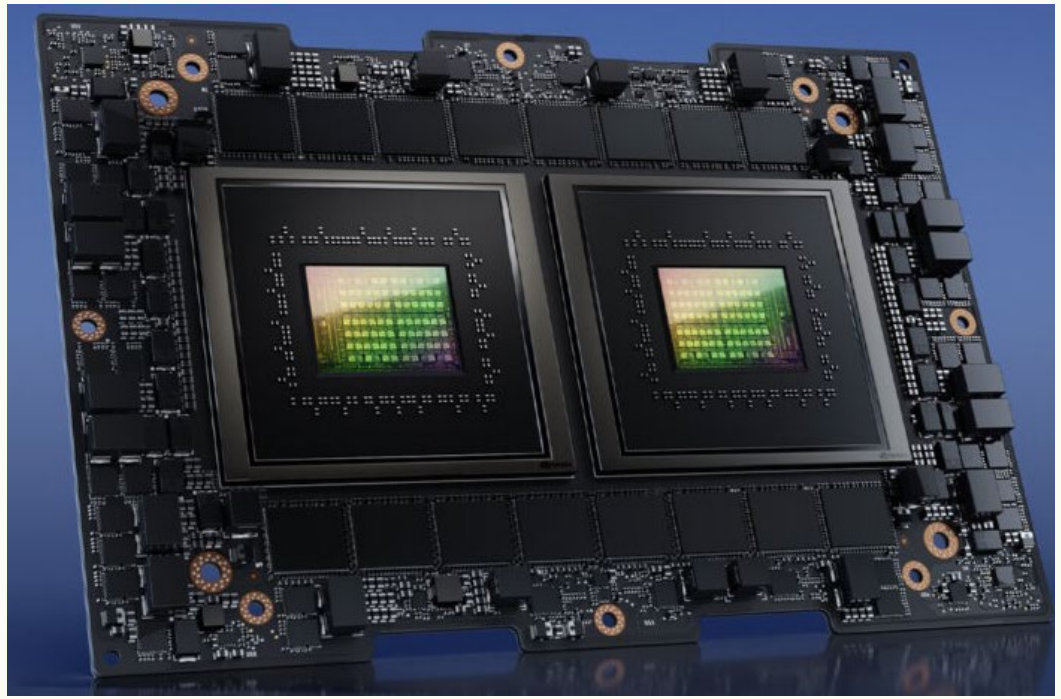
VISTA Hardware

- 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)

NVIDIA Grace Grace (GG)

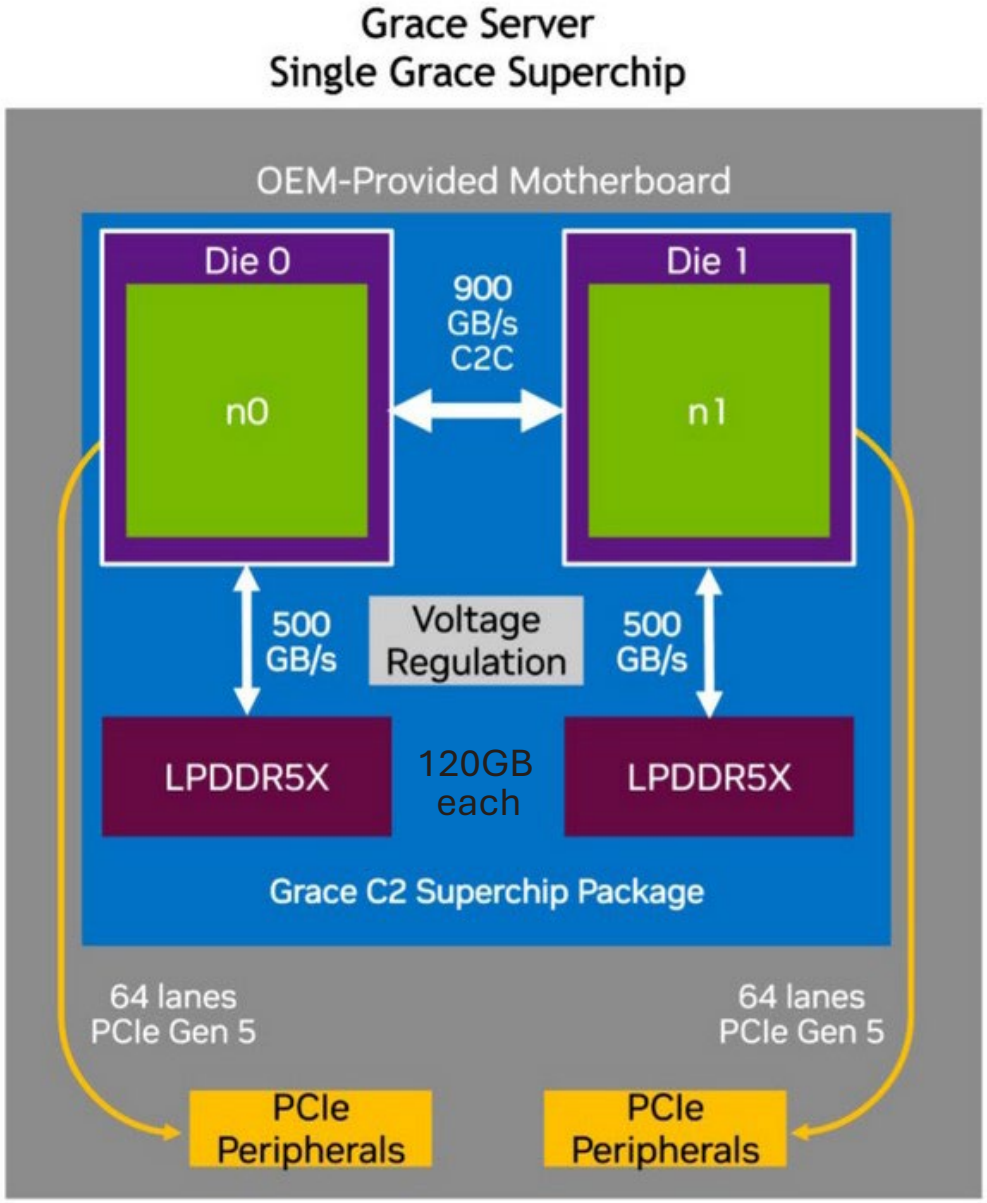


2
NUMA
Nodes

2
Compute
Dies

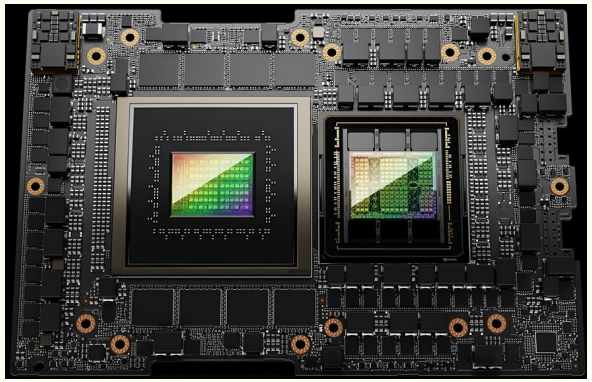
500
Watts
(CPU +
MEM)

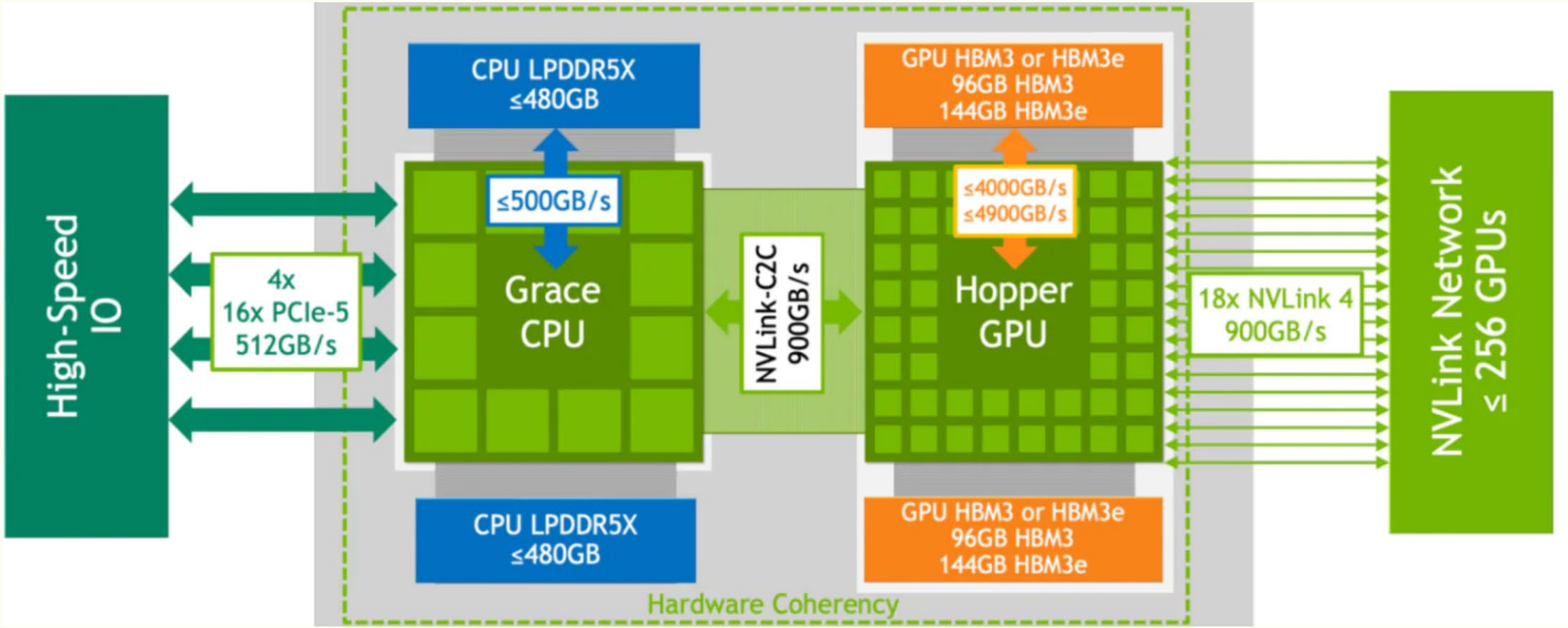
900
GB/s
worst-case
n to n



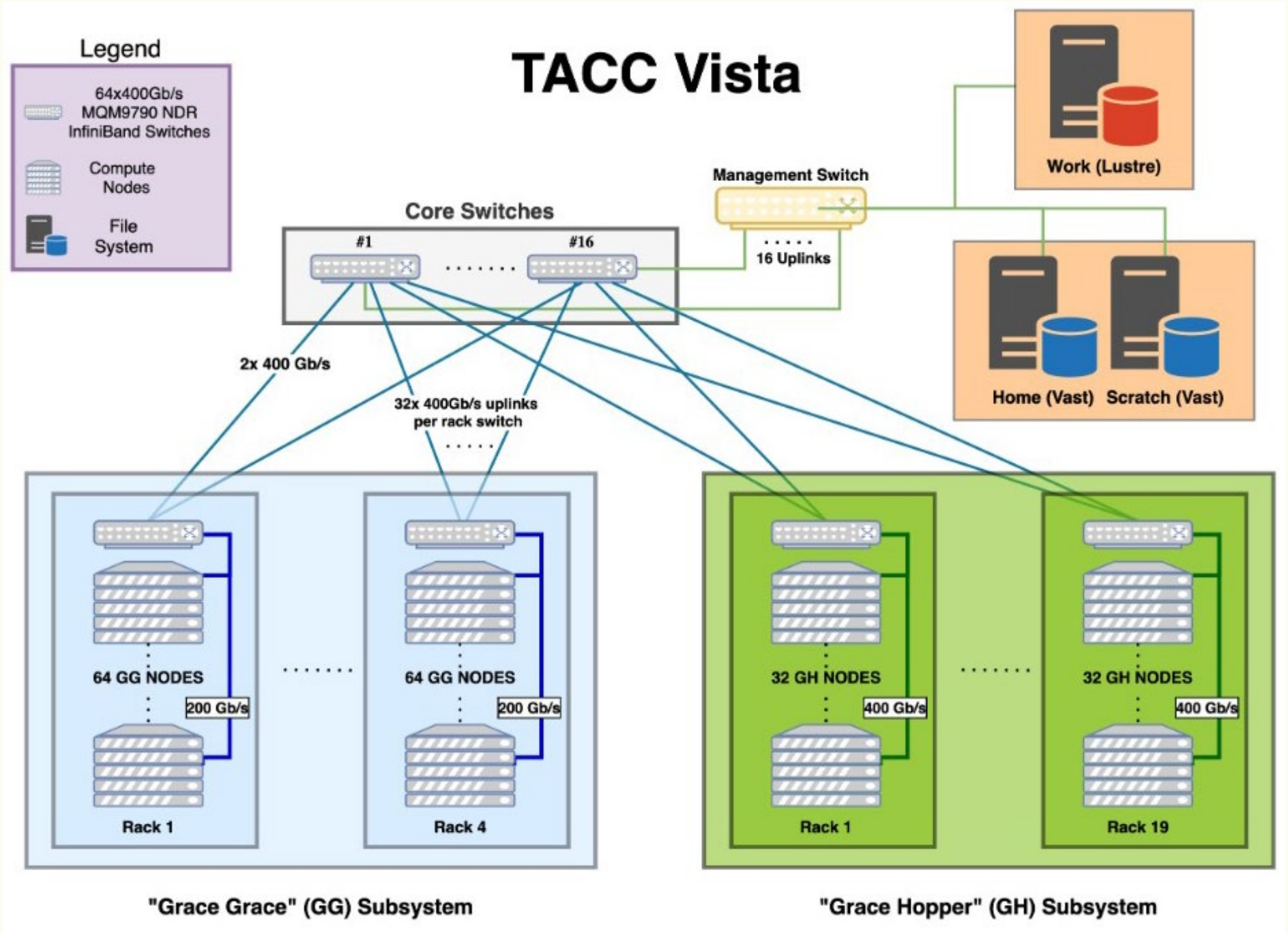
Courtesy : NVIDIA

NVIDIA Grace Hopper Superchip (GH200)





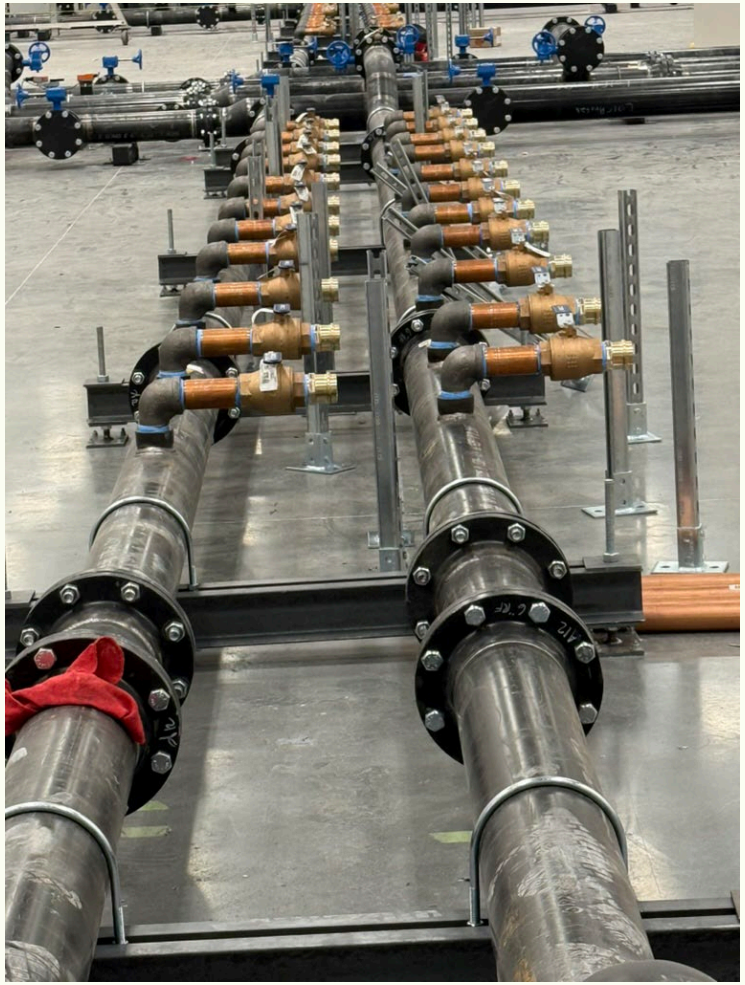
Courtesy : NVIDIA


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

Horizon Deployment

- Horizon will replace Frontera (39PF CPU, 2019)
  - 8,368 CPU compute nodes
  - 90 GPU (360GPUS) compute nodes
- Frontera used 6MW, Horizon will be double that.
- So this year has been about building a new home, which is now about complete
- Partnered with Sabey Datacenters in Round Rock, TX to build out a hall to our specs.

Built to support 250KW/cabinet



September 2025
Underfloor cooling install

October 2025
Raised floor install

Horizon Racks

- As we are using the NVIDIA Blackwell “NVL-4” shelf, we aren’t limited to 72 GPUs per cabinet.
  - Also, they will weigh 6,000lbs with liquid and rear doors. (~2 Honda Civics, depending on your trim package).
- Horizon GPU racks will have 144 GPUs and come in at about 215KW per rack.
- Horizon CPU (Vera) racks will come a little later, but in roughly the same form factor – 72 dual-socket nodes per cabinet, >100KW per rack.
- 28 GPU Racks, 66 CPU racks, 34 storage, switching, management racks, 128 cabinets total (smaller than Stampede – but a lot more power).
- Estimating ~13MW

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

Open Call for scientific applications

Selecting few representative applications (20/140)

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

Holistic study of applications performance on variety of architectures

- Projects ran from 2022Q3 to 2025Q1
- Down selected to 11 projects for Application Performance Enhancement

Baseline performance and prediction

Characteristic Science Applications (CSA)

AWP-ODC

Athena-K

Changa

MILC

NAMD

PSDNS

Seissol

WESTPA

EPW

Enzo-E

MuST

CSA Applications

- (Earth) AWP-ODC: Seismic simulation for hazard management
  - Yifeng Cui; SDSC
- (Astrophysics) Athena++: Astrophysical Fluid Dynamics at Exascale
  - James Stone, Princeton University
- (Astrophysics) ChaNGa: Evolution of baryons and galaxies across the age of the Universe
  - Tom Quinn, University of Washington
- (Physics) MILC: Lattice QCD for flavor physics
  - Steve Gottlieb; Indiana University
- (Bio) NAMD: Molecular Mechanisms of Viral Infection
  - Emad Tajkhorshid, UIUC
- (CFD) PSDNS: Large-scale DNS
  - PK Yeung; GTech

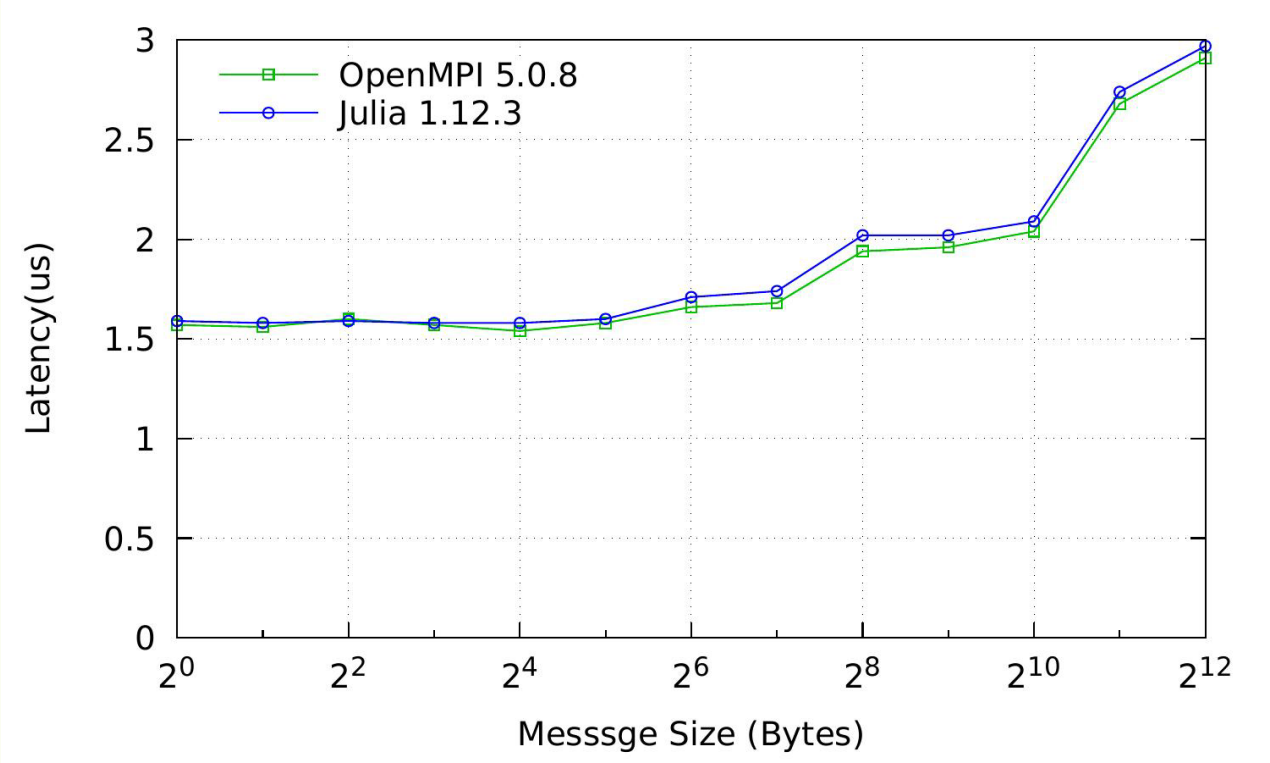
CSA Applications

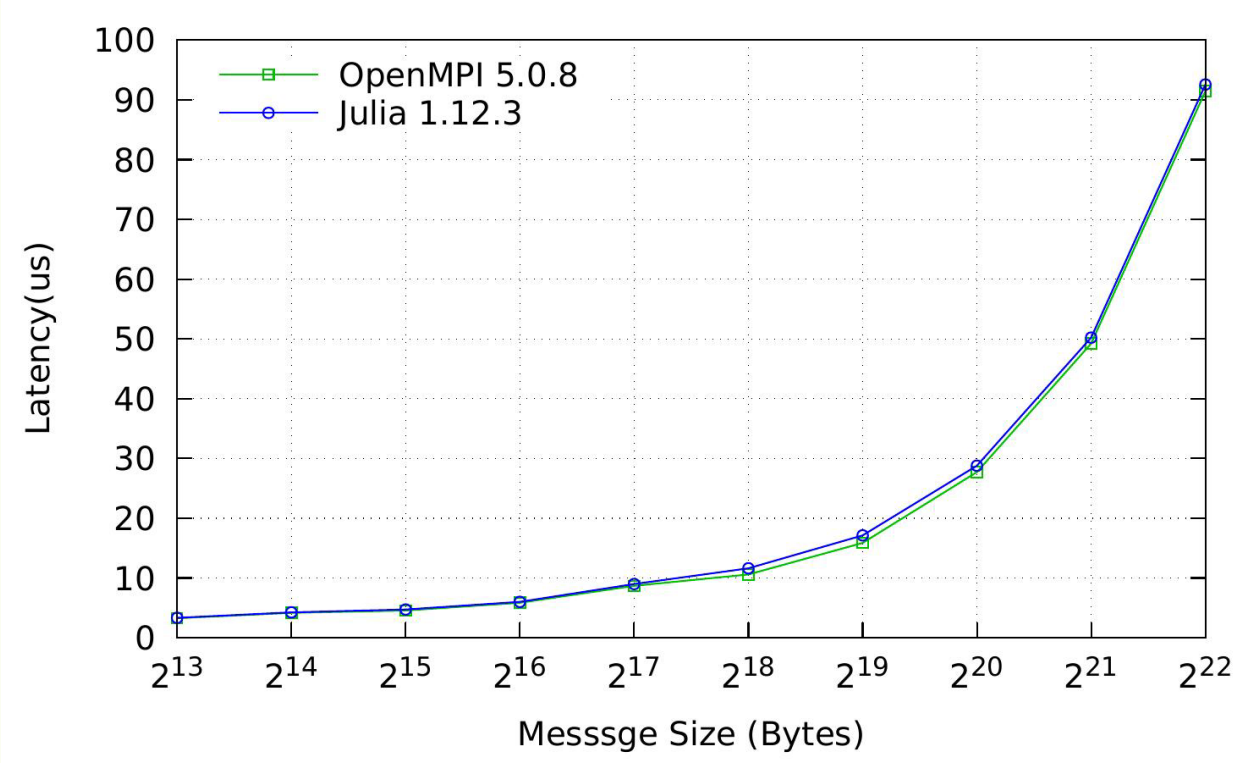
- (Earth) SeisSol: Off-fault inelastic processes and fluid effects in earthquake simulation
  - Alice Gabriel, UCSD
- (Bio) WESTPA: Application to Delta SARS-CoV-2 Spike Opening in Respiratory Aerosols
  - Lillian Chong; University of Pittsburgh
- (Chemistry) EPW: Quantum materials engineering at the exascale
  - Feliciano Giustino, Sabyasachi Tiwari; UT
- (ML) Enzo-E: Accelerating cosmological simulations of the first galaxies through deep learning
  - Mike Norman, UCSD
- (Material) MuST : Electron localization in materials
  - Yang Wang,PSC

Julia Performance Results

MPI Benchmarks – P2P Latency

(OSU Benchmarks)



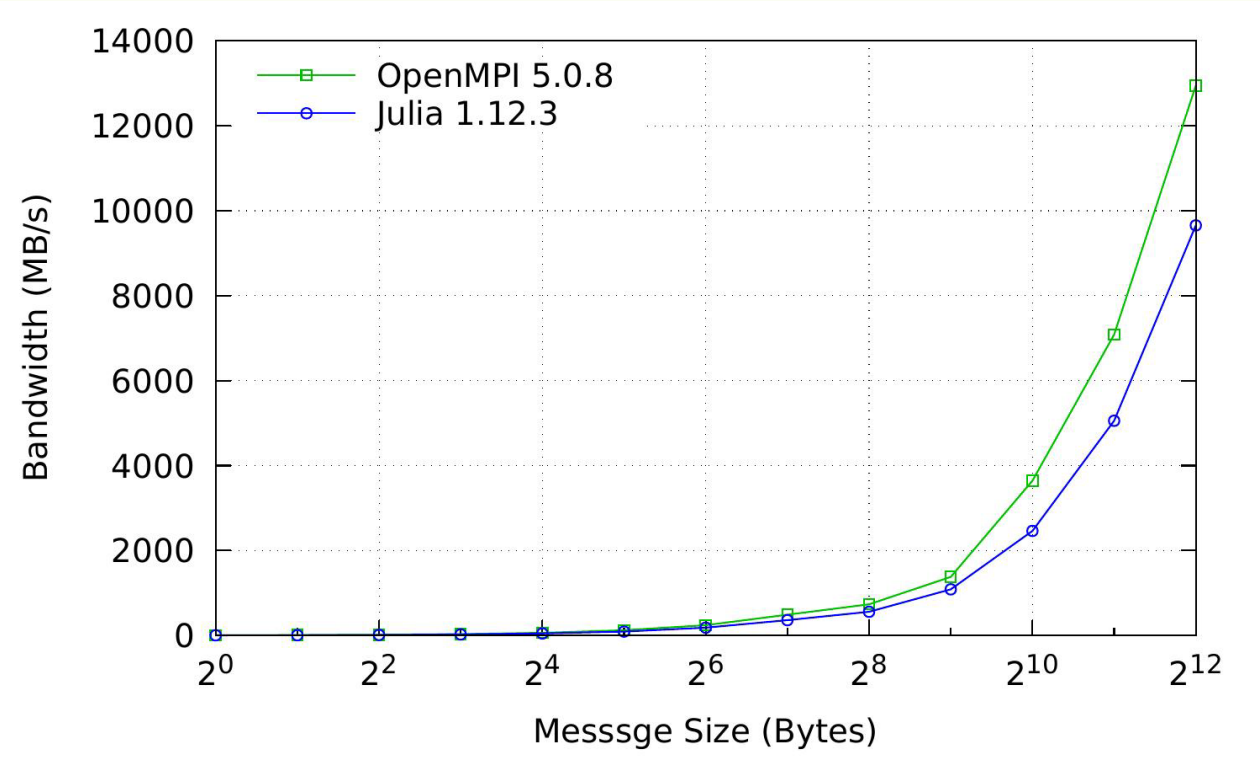


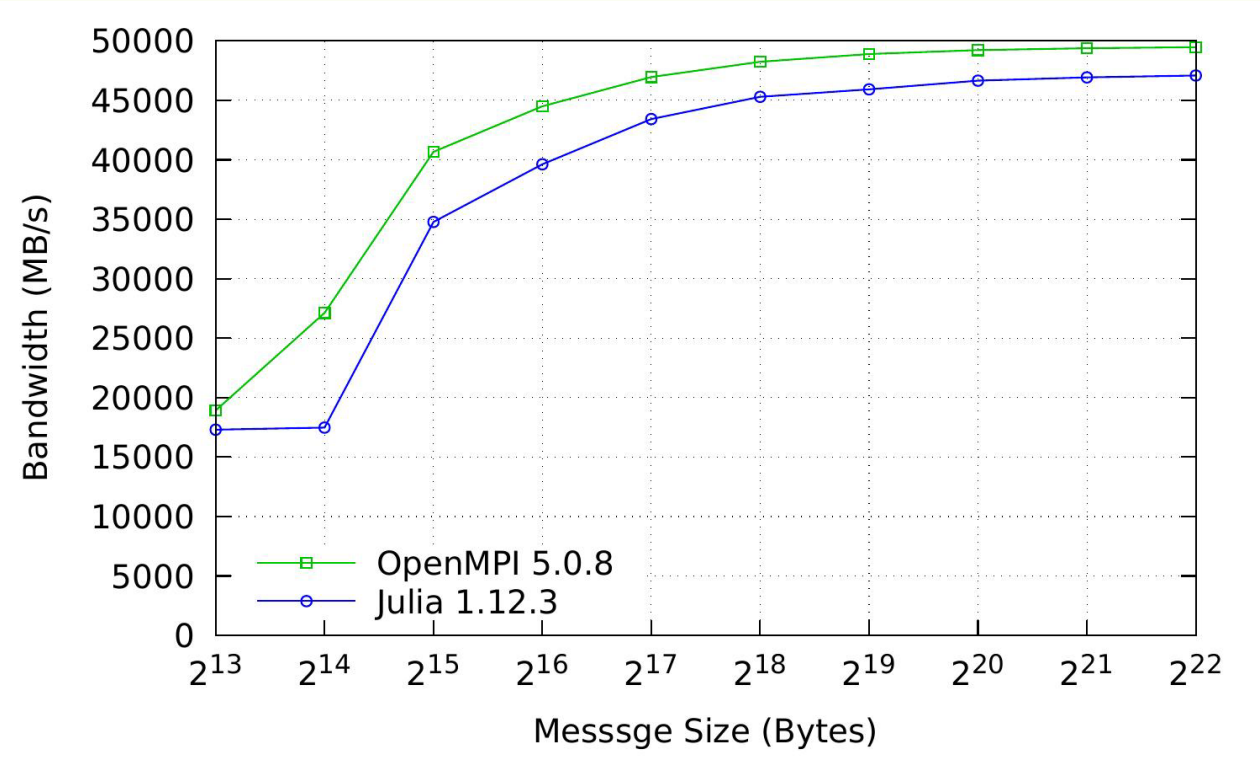
Observations (OSU Benchmarks compiled with Julia Wrappers vs compiled with OpenMPI):

1. The point-to-point message latency with Julia MPI wrappers is similar to OpenMPI.
2. The trend persist for all message sizes.

MPI Benchmarks – P2P Bandwidth

(OSU Benchmarks)



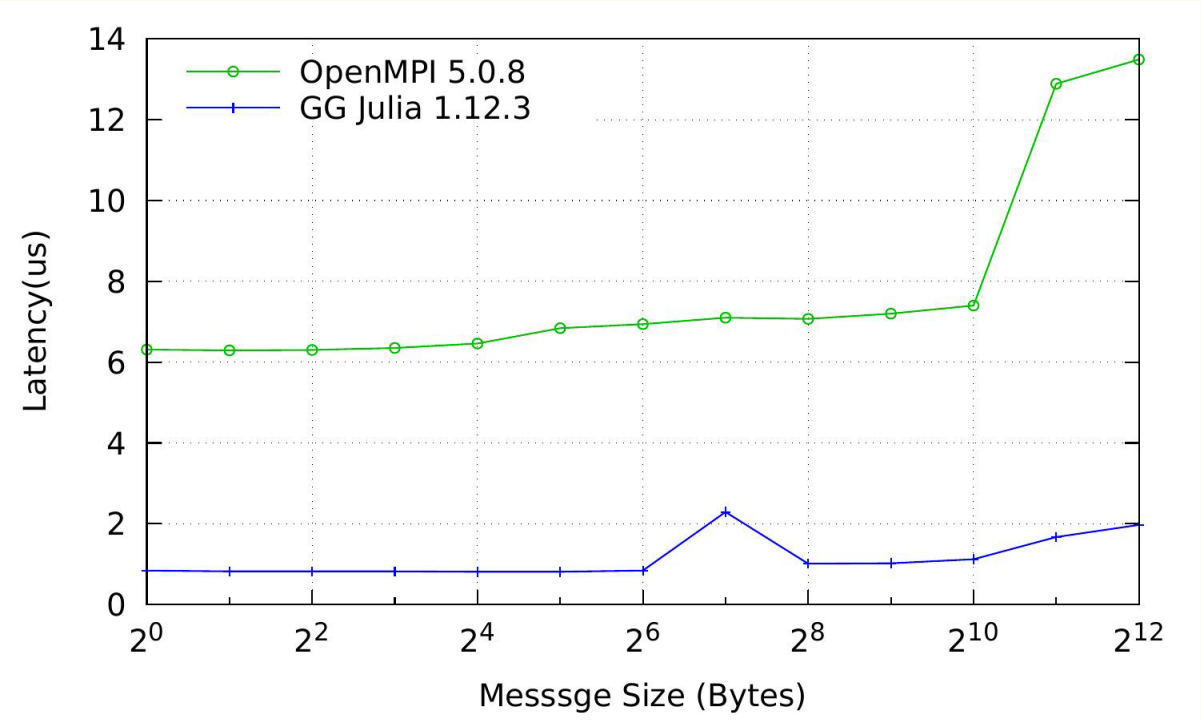


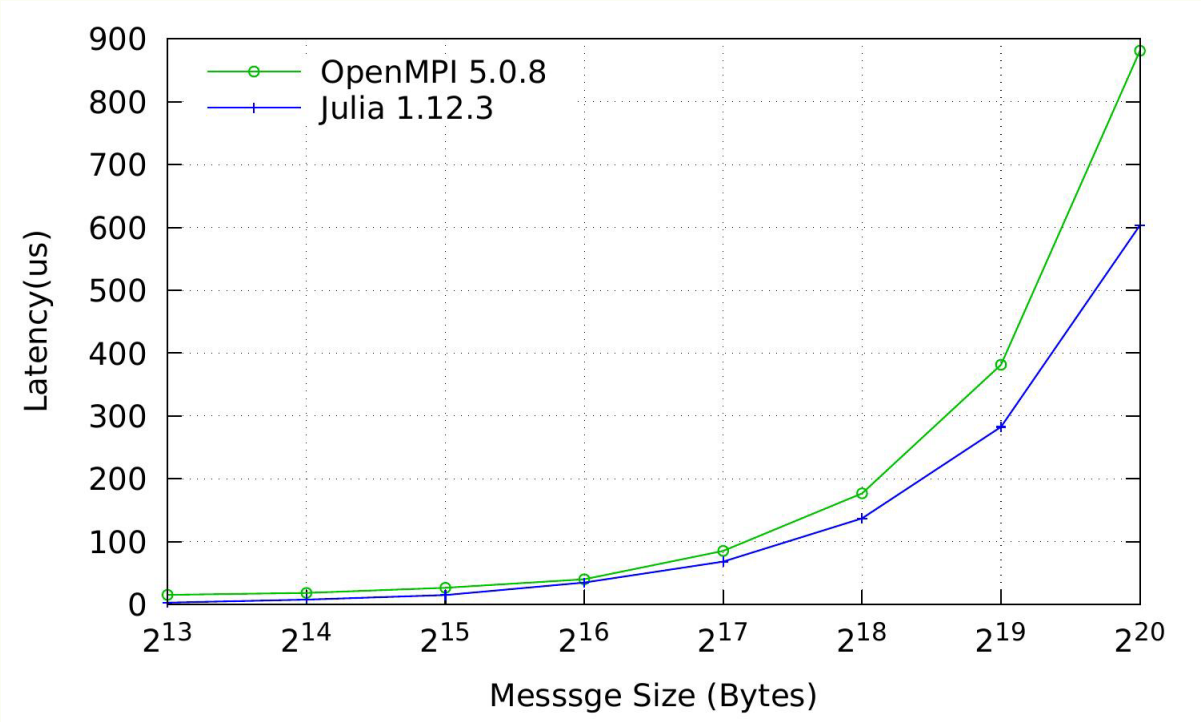
Observations :

1. Both Julia and Open MPI sufficiently saturate the interconnect bandwidth (NDR=400 Gb/s).
2. Julia indicate some overheads that needs investigation.

MPI Benchmarks – Collectives Broadcast

(OSU Benchmarks)



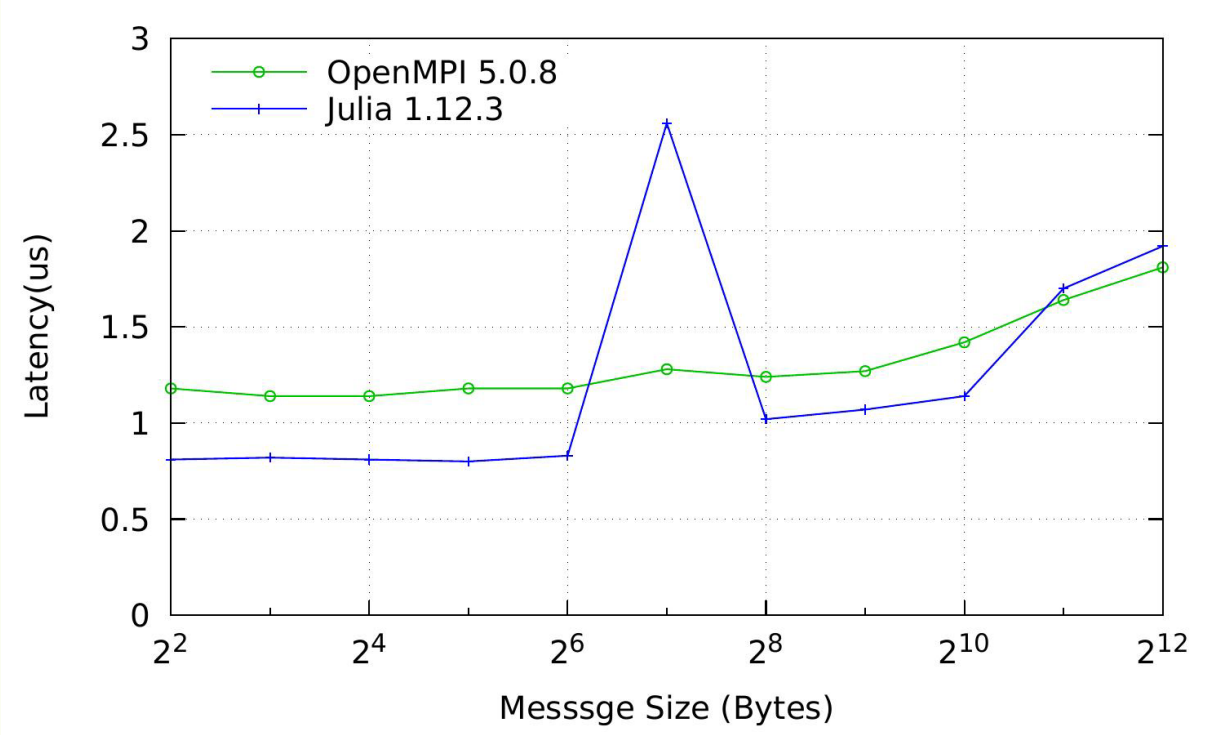


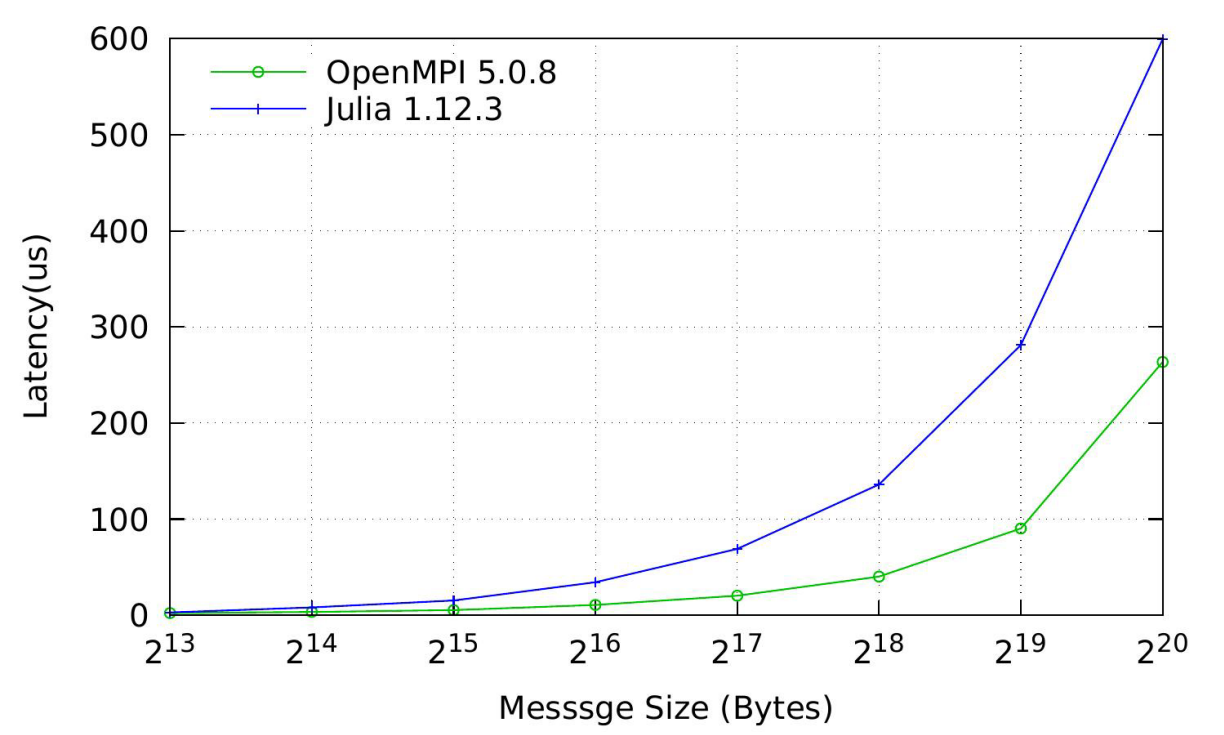
Observation :

- Broadcast communication in the Julia implementation demonstrates strong performance for both small and large message sizes..

MPI Benchmarks – Collectives Reduce

(OSU Benchmarks)



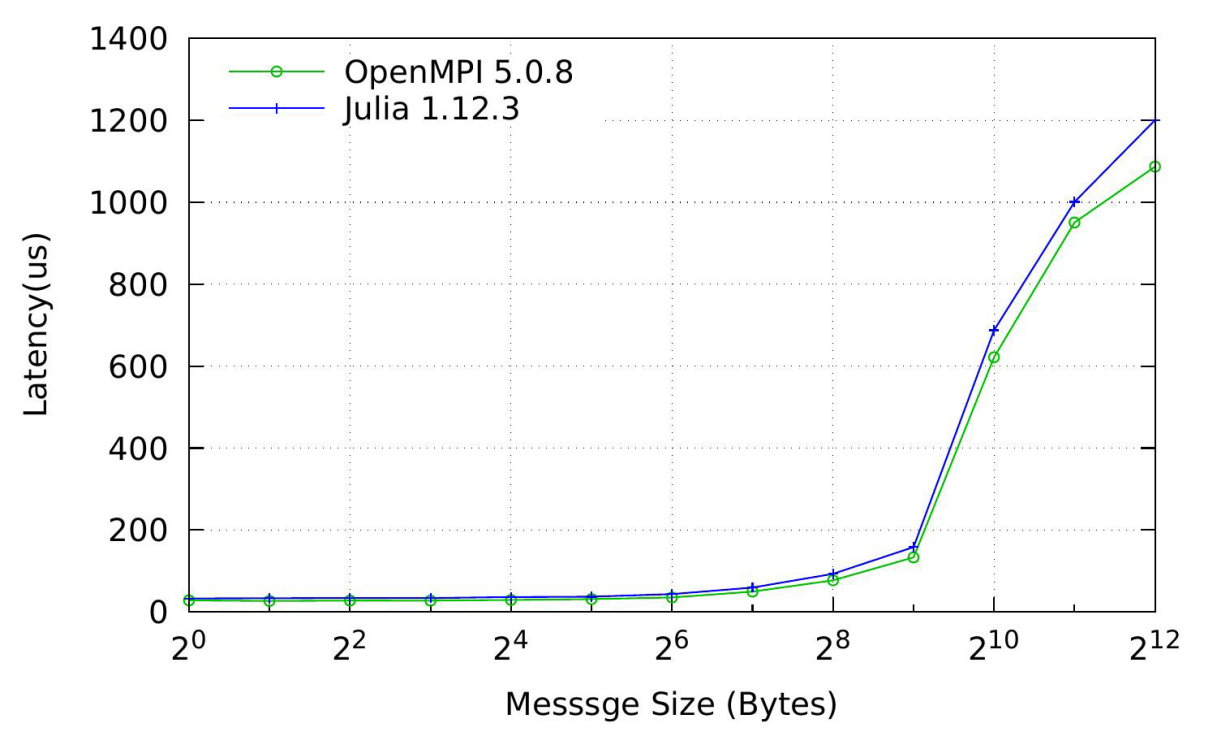


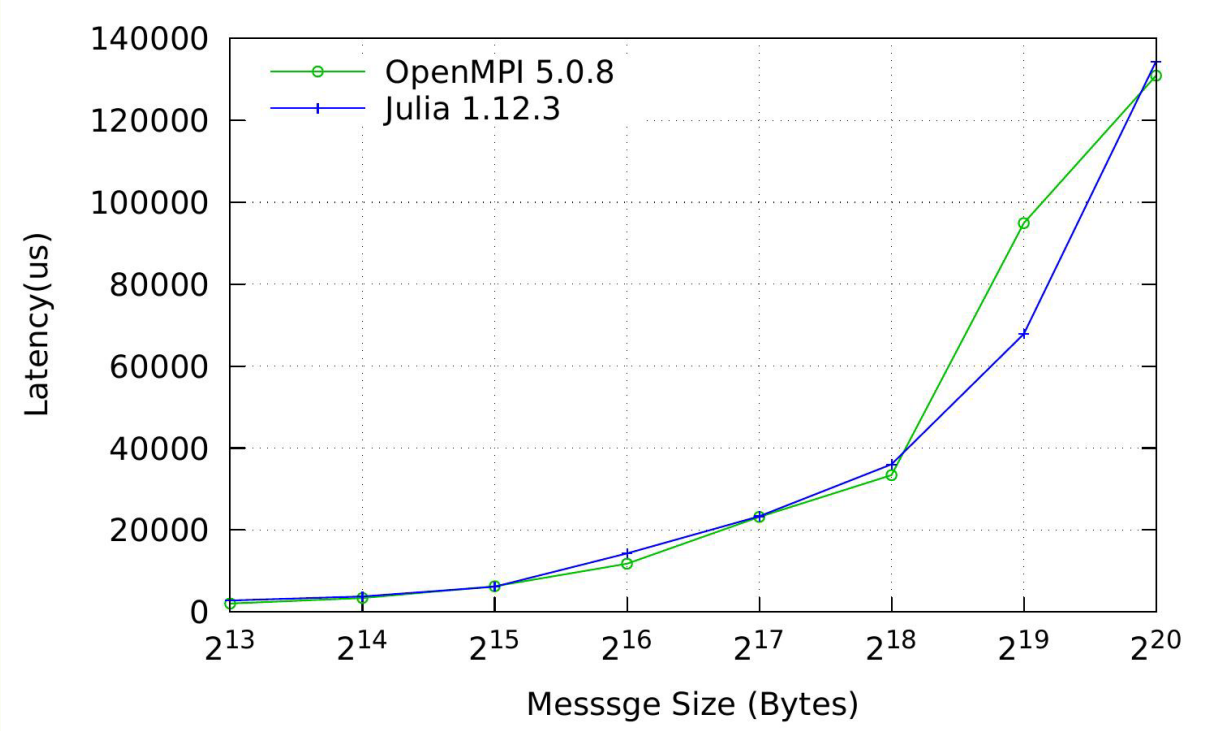
Observation :

- While the Julia implementation performed well for small message sizes, OpenMPI showed better performance for larger message sizes.

MPI Benchmarks – Collectives Alltoall

(OSU Benchmarks)

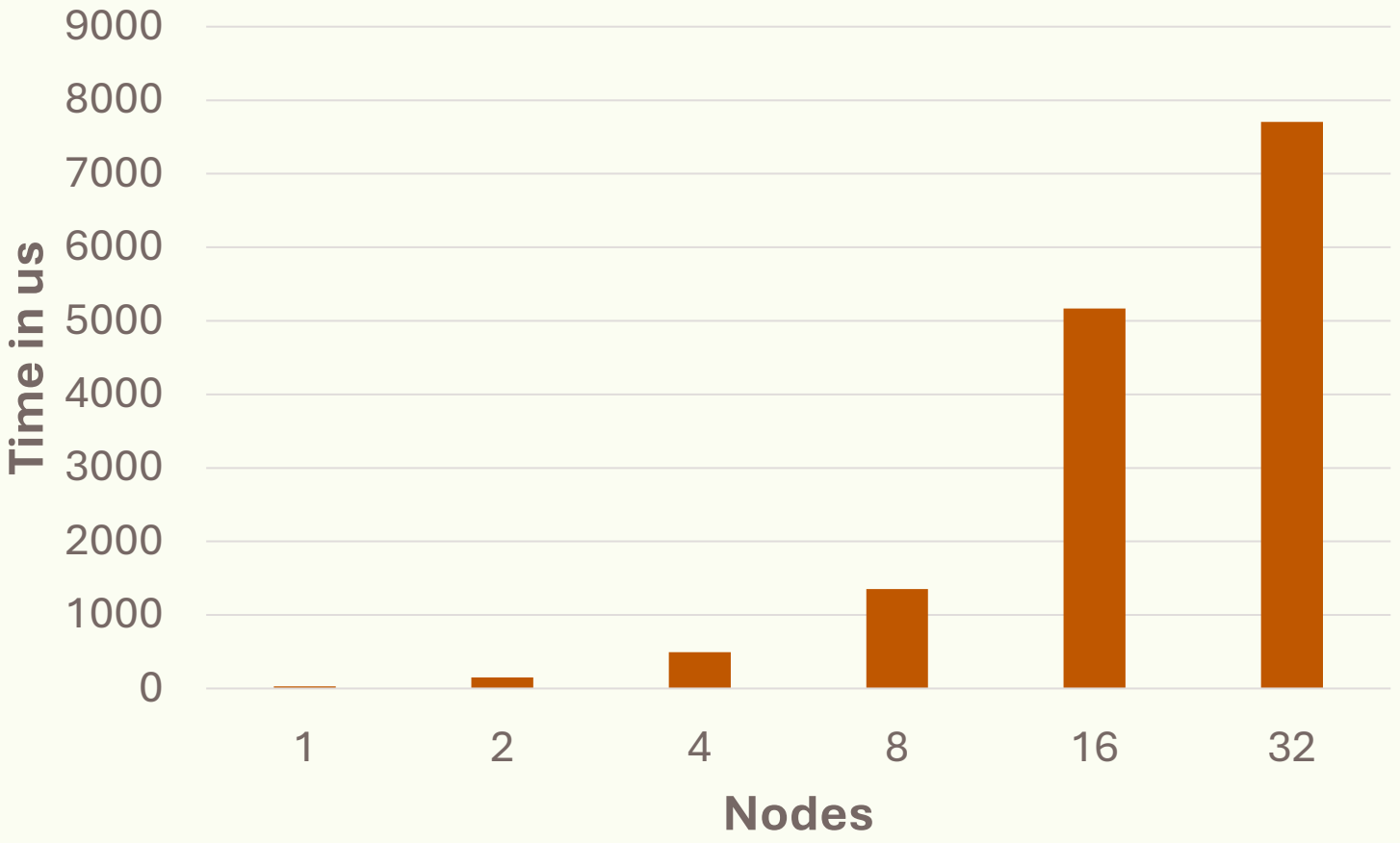




Observation :

- No overheads seen with Julia wrappers for Alltoall communication.

Alltoall – Scaling Performance



(MPIBenchmarks.jl)

Message Size = 64 Bytes
Each nodes has 144 Cores

1N=144 Ranks
32N=4608 Ranks

Observation :

- Latency increase with node counts is expected for the dense alltoall communication.

Concluding Remarks

1. Seamless installation and setup on Vista Machine
   - Performance tuning for shared setup
2. Version maintenance for both CPU and GPU based libraries
3. Guiding users through trainings and best practices
4. Benchmarking suite for High Performance Computing domain.

THANK YOU FOR ATTENDING.

E-mail:
aruhela@tacc.utexas.edu

Connect with us: /**taccutexas**