TEXAS ADVANCED COMPUTING CENTER

WWW.TACC.UTEXAS.EDU

TEXAS
The University of Texas at Austin

TACC Site Updates and Benchmarking for HPC Systems

HPC Asia

Feb 19, 2025

PRESENTED BY:

Amit Ruhela

Texas Advanced Computing Center

Austin, Texas

Agenda

- TACC Overview
- Current System Architectures
- Benchmarking Methodology
- Forthcoming Systems
  - Performance Results

TACC IN A NUTSHELL

- Founded in 2001
- 190 Staff (~70 PhD)
- Operates the Frontera, Stampede3, Vista, Jetstream, and Chameleon systems for the National Science Foundation (NSF)
- Lonestar6 for our Texas academic and industry users.
- 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.

Texas Advanced Computing Center

Frontera

Lonestar6

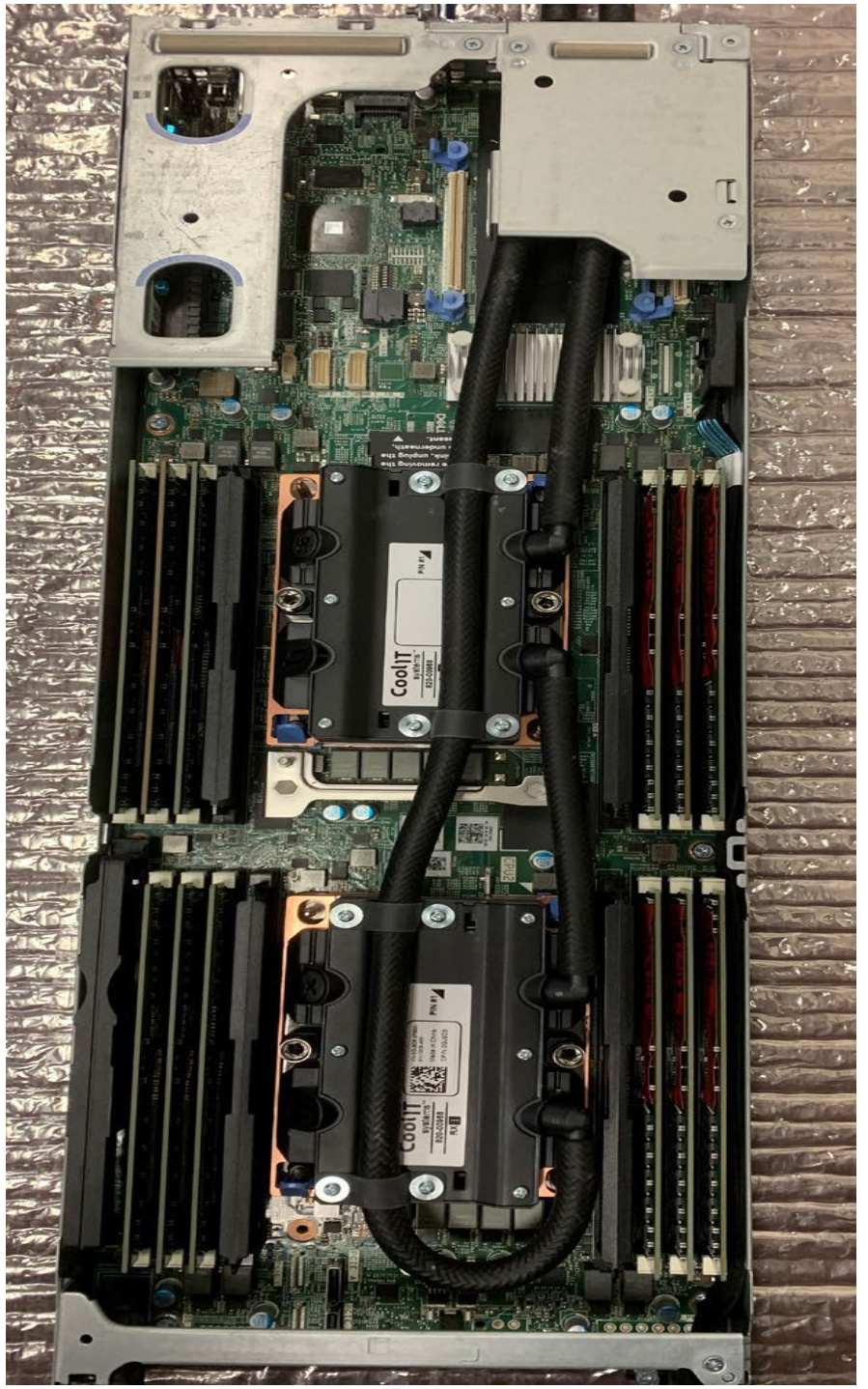
Frontera

Frontera Hardware Summary

- Compute nodes:
  - 8,008 Dell C6420 servers, dual-socket Intel 8280 28-core processors, 192GB, HDR100 IB
  - 396 Dell R640 servers, dual-socket Intel 8280 28-core processors, 192GB, HDR100 IB
  - 16 Dell R840 servers, quad-socket Intel 8280 28-core processors, 384GB (L4 Cache), HDR100 IB, 6TB NVDIMM
  - 90 GRC GPU oil-immersion servers, four NVidia RTX5000 cards per node, FDR IB
- Mellanox HDR InfiniBand interconnect (200Gbps core, 100Gbps to nodes)
- Storage subsystems:
  - Four DataDirect Networks(DDN) 18K Exascaler storage arrays, 56PB storage, 300 GB/s bandwidth
  - 72 DDN IME flash servers, 3PB storage, 1.5TB/s bandwidth
  - Stockyard2 /work sitewide filesystem, 10PB, 80 GB/s bandwidth
  - Ranch archival subsystem, 100PB+ of tape capacity

Frontera Compute Nodes

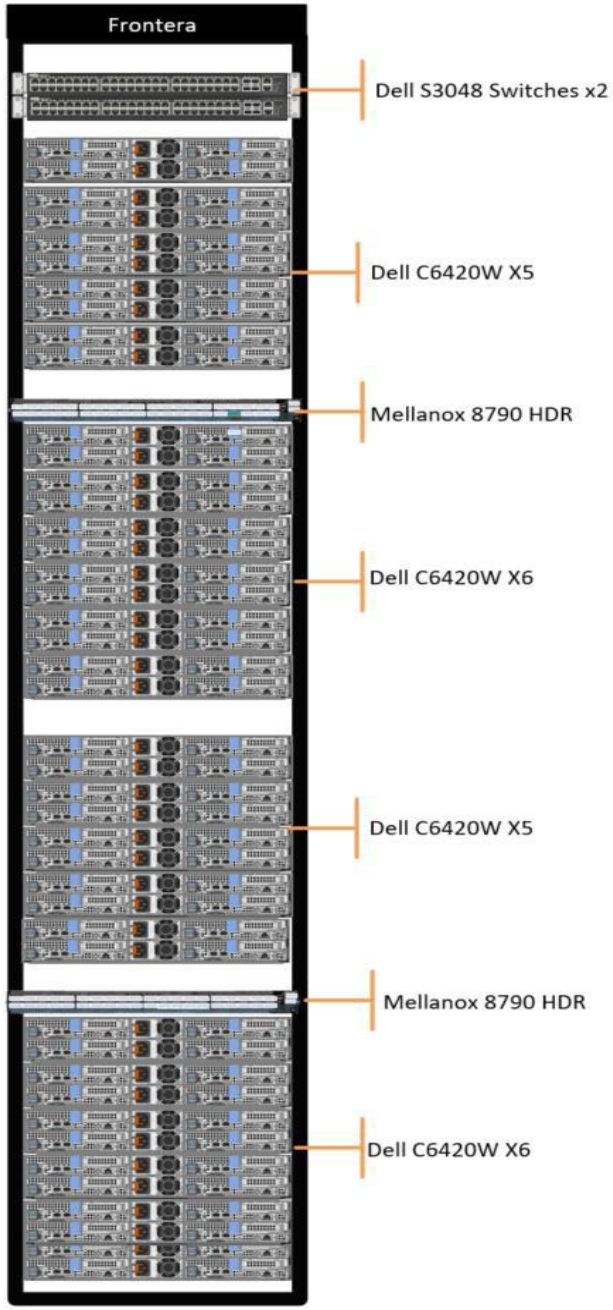
- 8,008 total compute nodes
  - Two Intel “Cascade Lake” (CLX) Xeon Platinum 8280 CPUs
  - 28 cores per socket, 205W TDP
  - Nominal frequency of 2.7GHz, 2.4 Tflops
  - 192 GB (12x16GB) 2933MHz DDR4
  - 240 GB SSD
  - Mellanox ConnectX-6 HDR100 InfiniBand (IB) PCIe card
  - Integrated iDRAC management
  - Cool-IT liquid cooling to the processors



TACC

Compute Rack Layout

- 91 total racks, 2,002 Dell C6420 chassis, 4 nodes per chassis
- 22 chassis / 88 nodes per rack in two groups
- Two 48-port GigE switches for data and management
- Two Mellanox QM8790 HDR 40-port switches
  - 44 nodes at 100Gbps, 18 uplinks at 200Gbps
- Fat-tree topology design with 11:9 oversubscription
- Liquid cooling manifolds
- Cool doors for supplemental cooling
- Rack cabling routed to allow for hot-swap of components



Software Stack

- Still running CentOS 7.9 kernel, considering big jump to Rocky 9 at some point…
- Multiple Mellanox OFED upgrades, now running 5.8.2, started with 4.5, 5.4.1 last meeting
- Lustre 2.12.9_ddn33 for servers and clients, backported patches applied, about to upgrade due to CVE
- Intel 19.1.1 Compiler and Intel MPI 19.0.7 (much improved over 19.0.5)
  - Intel OneAPI 2023 and 2024 compilers, libraries and tools all available as well
- MVAPICH 2.x advanced MOFED 5.0 MPI library
- Continuously adding additional 3rd party packages and upgrades; currently provisioning:
  - 1166 OS distro provided packages (1164 year 4, 1152 year 3, 1109 year 2, 881 year 1)
  - 1012 custom TACC maintained packages (882, 801, 759, 629)
  - 96 configuration files (90, 84, 78, 76)

Stampede

BACKGROUND: THE STAMPEDE SYSTEM LEGACY

- The NSF “Track2-funded” Stampede (2012-2017) and Stampede2 (2017-2023) systems delivered more than 21 million jobs to NSF users (and still counting)
- Workhorse x86 CPU-based systems for XSEDE and now ACCESS supporting hundreds of projects with thousands of users and diverse set of application
- Provided emergency response and priority for potential catastrophic events including hurricane tracking, storm surge modeling, severe storm prediction, earthquake response
- Both systems were in heavy demand and sustained >95% utilization during their life

Stampede3: System Design

- Reuse latest Stampede2 components, including all the Intel Ice Lake and most of the Skylake nodes along with OPA 100Gb fabric
- Add Intel Sapphire Rapids high-bandwidth memory nodes
- Add AI-focused GPU nodes with Intel Ponte Vecchio GPUs
- Install next generation 400Gb OPA fabric to all new nodes when available in 2025
- Replace filesystem with 12PB VAST storage subsystem with 50GB/s write, 400GB/s read bandwidth
- Access to Ranch tape archive with project quotas

Stampede3 : Compute Nodes

- 1064 Dell C6420 nodes with two Skylake 8160 processors
  - 48 cores at 2.1GHz, 192GB memory, 200GB SSD. 3.1TFlops
- 224 Dell R650 nodes with two Ice Lake 8380 processors
  - 80 cores at 2.3GHz, 256GB memory, 480GB SSD, 5.9TFlops
- 560 Dell C6520 nodes with dual Sapphire Rapids 9480 CPUs
  - 112 cores at 1.9GHz, 128GB HBM memory, 480GB SSD, 6.8TFlops
- 20 Dell XE9640 nodes with four Intel Data Center GPU Max 1550 accelerators (Ponte Vecchio), dual 8480 CPUs
  - 112 CPU cores, 512GB DRAM, 512GB HBM2e (GPU), 215TFlops
- 12+PFlops peak performance

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

Need for New Systems

1. Increases Workload demands
2. Next-gen Processor Architectures
3. New types of workloads

TACC Compute Hardware

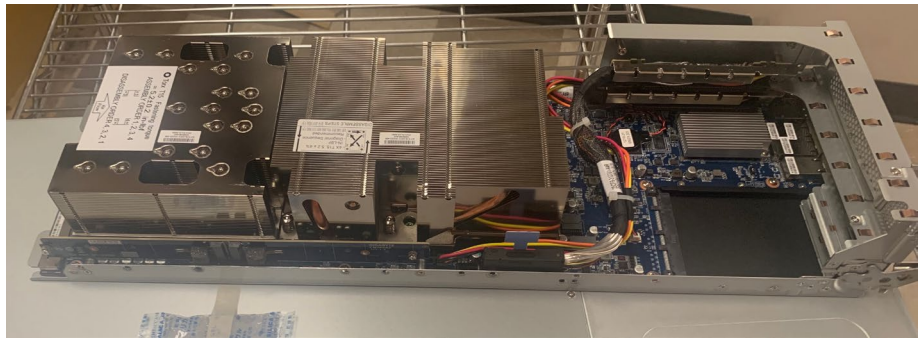
Resource	CPU type	#Nodes/Sockets/Cores	GPU Type	# GPUs
Frontera	Xeon (Cascade Lake)	8400/16800/470,400	RTX (Volta)	360
Lonestar-6	AMD Epyc	600/1200/76,800	NV A100	255
Stampede-3	Xeon (Sapphire Rapids)	2,024/4,048/150,080	Intel PVC	80
Vista	**ARM/Grace**	**840/1080/77,760**	**NV H100**	**600**
Horizon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

Vista

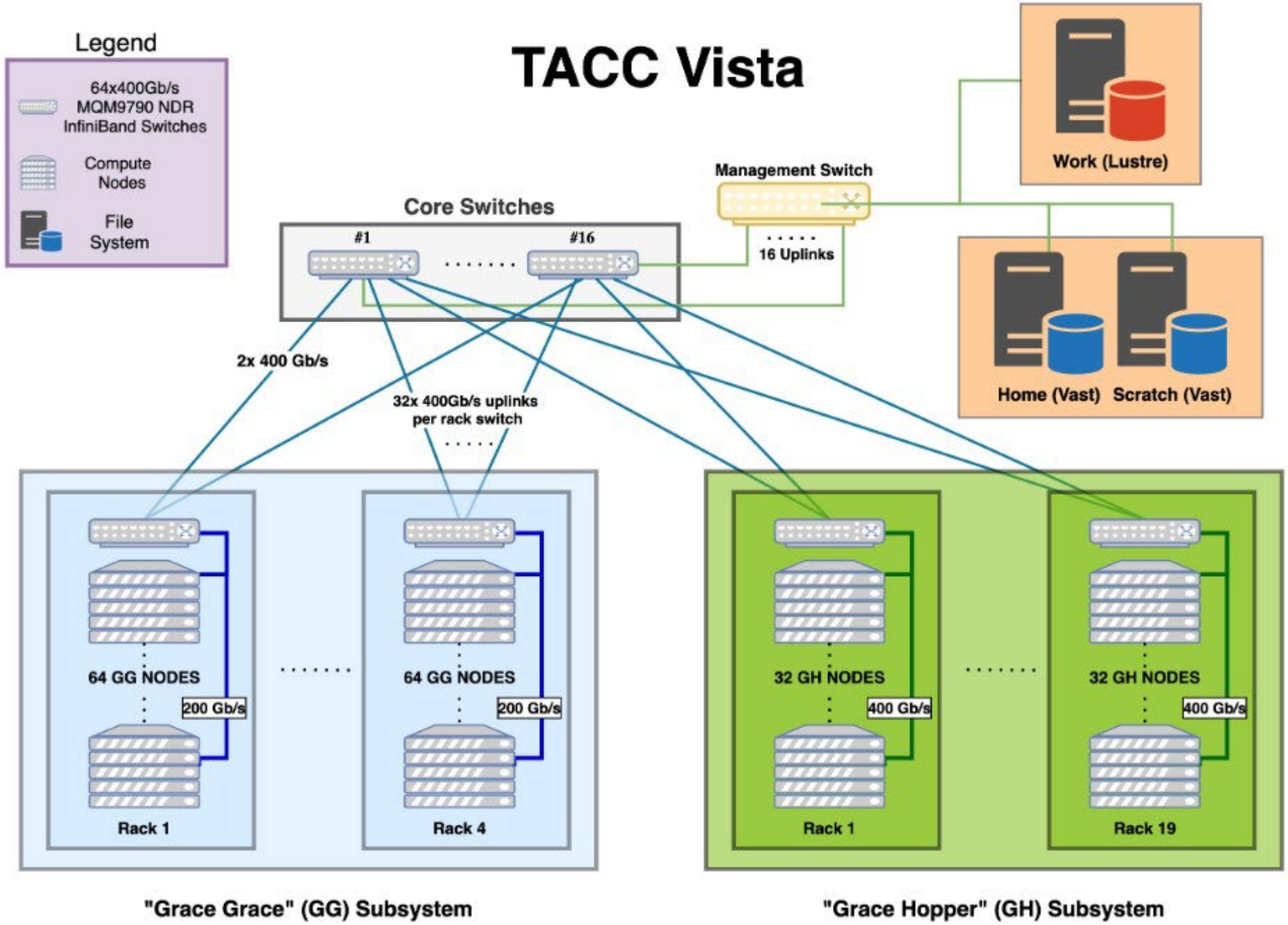
- Vista is a new AI-centric resource.
- Vista is half-funded as a supplement to Frontera, and half by UT-Austin AI initiatives.
- Vista is a bridge to Horizon.
- And Vista is a couple of firsts for TACC:
  - Our first system with an ARM as the primary CPU.
  - Our first system with NVIDIA as the primary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

VISTA Hardware

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



- 240GB of LPDDR5X RAM (GG), 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


Legend
64x400Gb/s MQM9790 NDR InfiniBand Switches
Compute Nodes
File System
TACC Vista
Work (Lustre)
Management Switch
16 Uplinks
Core Switches
#1
#16
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

Software Configurations

1. Compilers : NVIDIA HPC SDK : 24.7 and 24.9 and GNU : 13.2 and 14.2
2. UCX 1.17
3. CUDA 11.8, 12.4, 12.5, 12.6
4. MPI
   a. OpenMPI 5.0.5
   b. MVAPICH-PLUS 4.0.0
   c. HPC-X (bundled in SDK, v4)
5. Profilers
   a. NVIDIA NSIGHT Systems and Compute
   b. GDB, Remora
6. Containers Runtime : *Apptainer and Charliecloud*
7. 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. . . 17 years ago.

- 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**

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 (CSA) Elements

CSAs are initiated with the following three elements

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

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

Skipping to the answer

- $TenX = \sum_{i=1}^{i=numApps} \Delta perf_i \times \frac{1}{numApps} \times W_i$
- Where $\Delta perf_i = \Delta T \times \Delta S \times \Delta E \times \Delta P$
- Easy, right?

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 $\Delta perf_i$

Some differences from normal benchmarks

- In a “normal” benchmark:
  - Fix the version of the code
  - Fix the input problem file
  - Fix the number of cores/processors/GPUs to compare
- We are not applying these constraints, because. . .
  - We *want* the software to improve.
  - We *want* to run the most relevant version of the problem, not the one from 5 years ago.
  - We want to compare the value of the entire system in delivering “10x”, not compare one node to another.

The model can deal with:

- Algorithmic Changes
- Code improvements
- New technologies, heterogeneity
- Workflows coupling codes (to a degree)
- Thick vs. thin nodes (i.e., comparing single socket CPU nodes to quad GPU nodes).
- Problems that run at odd sizes.
- Improvements in I/O performance.
- AI (training time improvements, network improvements).

Applications

1. Bio : NAMD
2. Bio : Amber
3. Earth: SeisSol
4. Earth: WRF
5. Earth: AWP-ODC
6. Materials : PARSEC
7. Materials : MuST
8. Materials : LAMMPS
9. CFD: ADCIRC
10. CFD: PSDNS
11. Physics : Milc

Performance Results

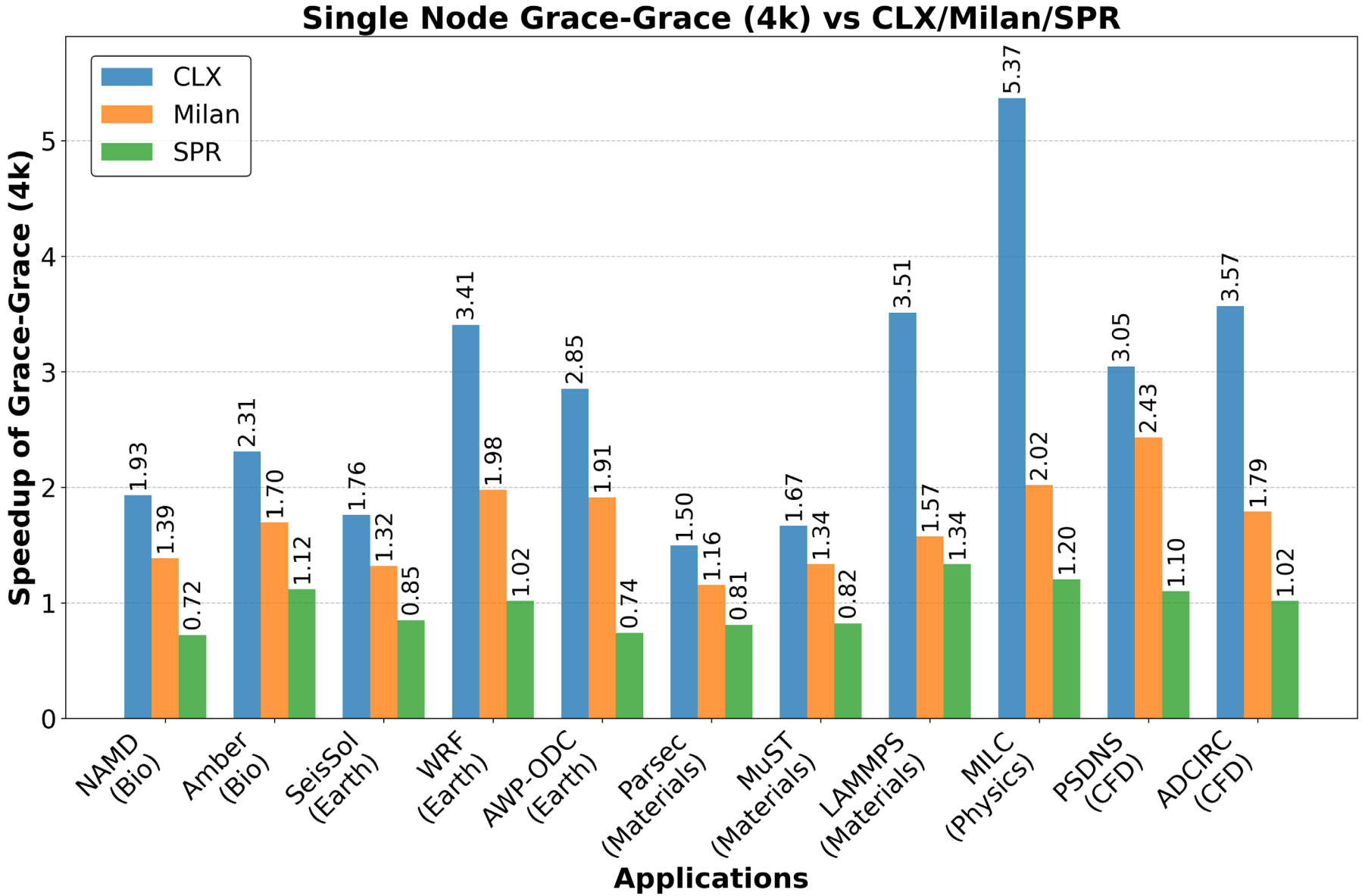
Single Node Comparison

Attribute	Cascade Lake	Milan	Sapphire Rapids	Grace	Grace vs Cascade Lake	Grace vs Milan	Grace vs Sapphire Rapids
Cores per Socket (core/socket)	28	64	56	72	2.6	1.13	1.3
Floating Point Peak Performance Per Socket (Gflops/s/socket)	2150	2509	3405	3917	1.8	1.6	1.2
Memory per socket (GiB/socket)	96	128	64	128	1.3	1.0	2.0
Memory Per Core (GiB/core)	3.4	2	1.1	1.8	0.5	0.9	1.6
Peak Memory Bandwidth per Socket (GB/s/socket)	140	195	1638	500	3.6	2.6	0.3
Sustained Memory Bandwidth per Socket (STREAM Triad GB/s)	111	152	700	445	4.0	2.9	0.64

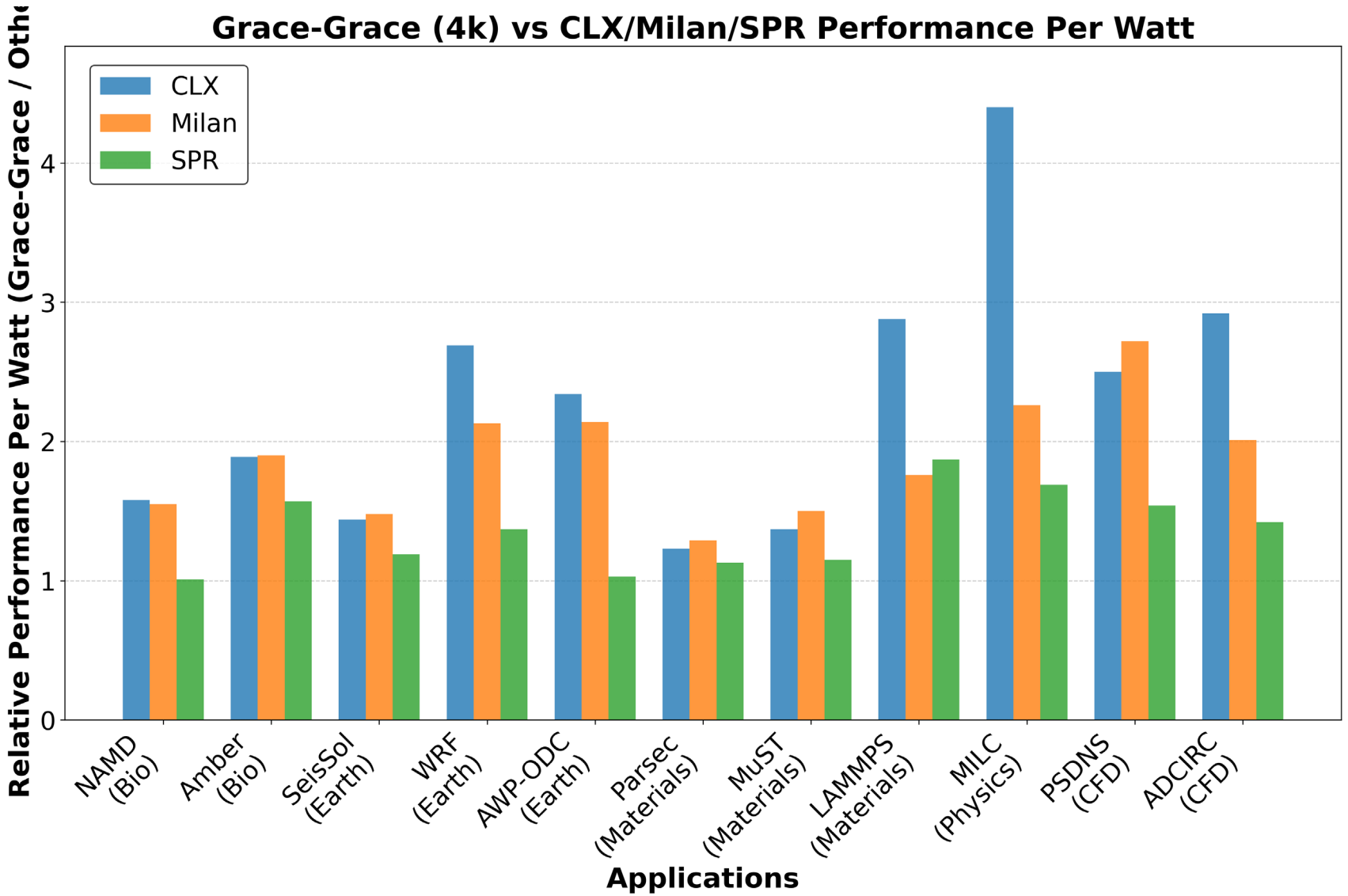
Single Node Experiments

Area	Application	CLX	Milan	SPR	Grace-Grace PageSize=4k	Grace-Grace PageSize=64K
Bio	NAMD	2.32	3.23	6.21	4.48	5.83
Bio	Amber	2.07	2.82	4.27	4.78	5.01
Earth	SeisSol	2075.76	1555.86	1000.52	1178.33	1190.59
Earth	WRF	1810.74	1052.27	541.34	552.6	531.49
Earth	AWP-ODC	328	220	84.97	114.95	114.543
Materials	Parsec	574.36	443.89	310.46	384	383.9
Materials	MuST	103.3	82.8	51	62	62
Materials	LAMMPS	960.6	2143.2	2526.4	3373.3	3451.4
Physics	MILC	3037.7	1142.2	681.6	566	565.7
CFD	PSDNS	968.495	772.988	350.132	318	247
CFD	ADCIRC	2152.01	1080.25	613.38	603.34	596.73

Speedup Comparisons



Energy Comparisons



Thanks! Questions?