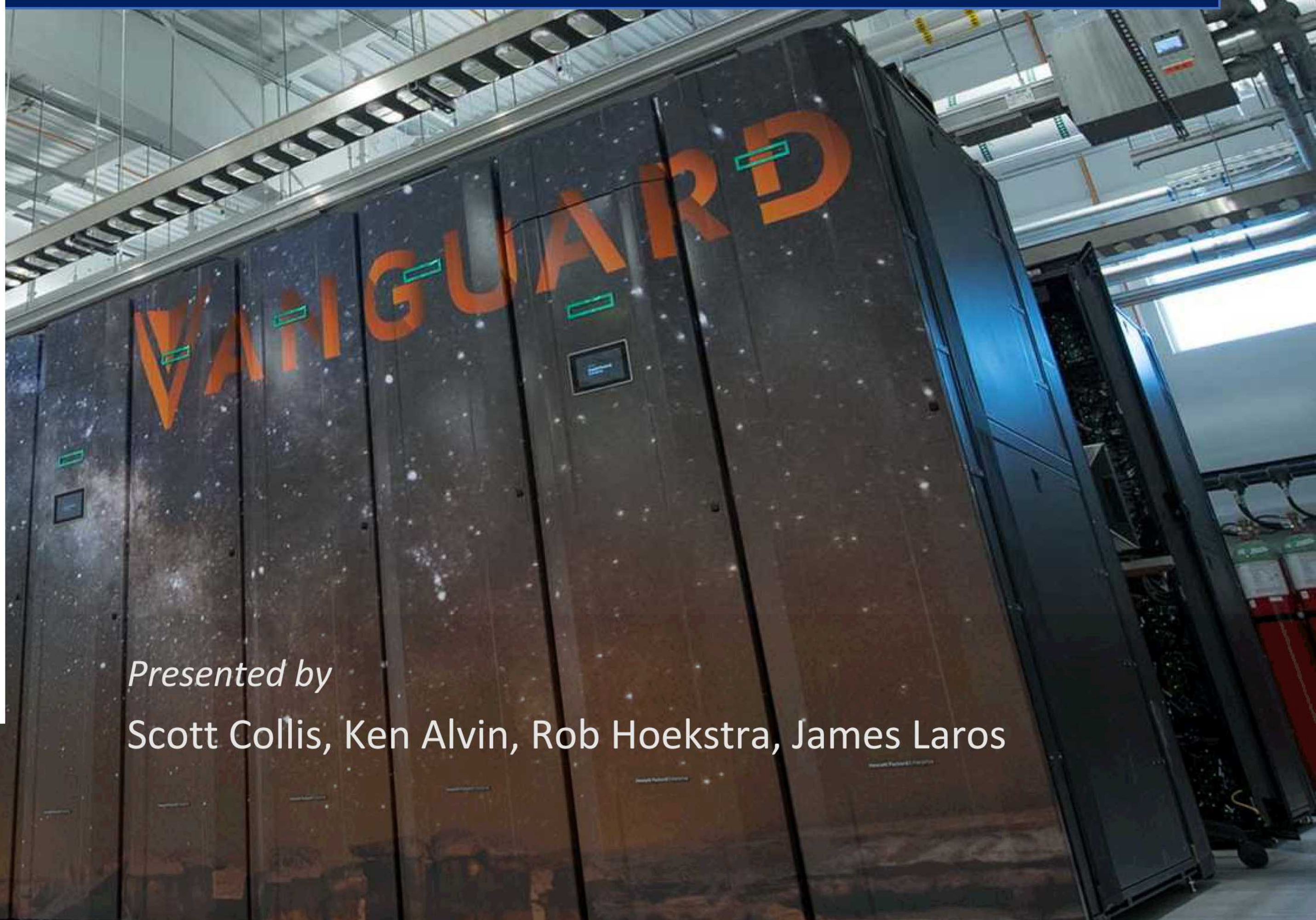
This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2019-1378C

ASTRA

"Per aspera ad astra"

RM SUPERCOMPUTER

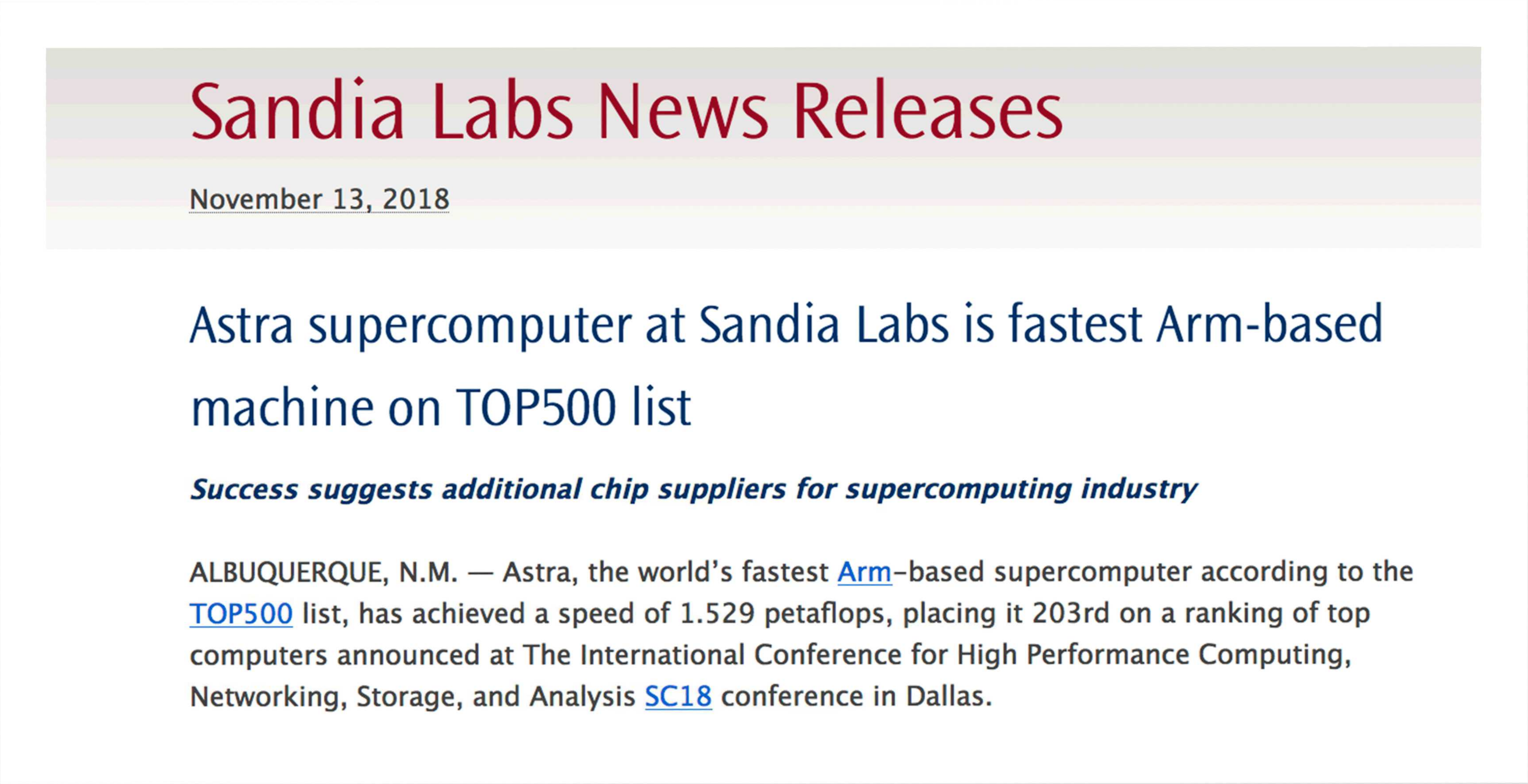


Presented by

Scott Collis, Ken Alvin, Rob Hoekstra, James Laros

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National inistration under contract DE-AC04-94AL85000.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.



Sandia Labs News Releases

November 13, 2018

Astra supercomputer at Sandia Labs is fastest Arm-based machine on TOP500 list

Success suggests additional chip suppliers for supercomputing industry

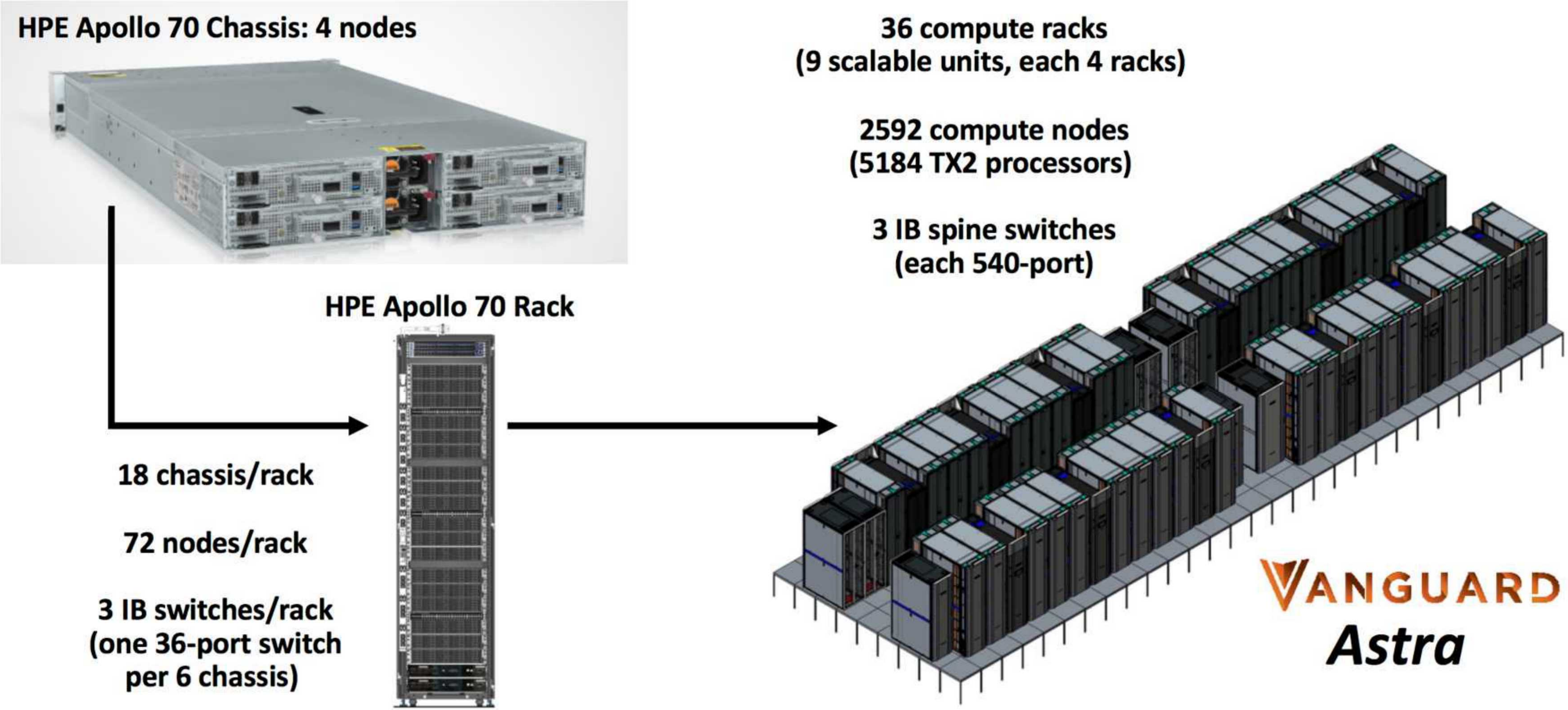
ALBUQUERQUE, N.M. — Astra, the world's fastest Arm–based supercomputer according to the TOP500 list, has achieved a speed of 1.529 petaflops, placing it 203rd on a ranking of top computers announced at The International Conference for High Performance Computing, Networking, Storage, and Analysis SC18 conference in Dallas.

TOP500 Lists

203	Sandia National Laboratories United States	**Astra** - Apollo 70, Cavium ThunderX2 CN9975-2000 28C 2GHz, 4xEDR Infiniband HPE	125,328	1,529.0	2,005.2	

36	203	**Astra** - Apollo 70, Cavium ThunderX2 CN9975-2000 28C 2GHz, 4xEDR Infiniband , HPE Sandia National Laboratories United States	125,328	1,529.0	66.94

Astra – the First Petscale Arm based Supercomputer



HPE Apollo 70 Chassis: 4 nodes

HPE Apollo 70 Rack

18 chassis/rack

72 nodes/rack

3 IB switches/rack
(one 36-port switch
per 6 chassis)

36 compute racks
(9 scalable units, each 4 racks)

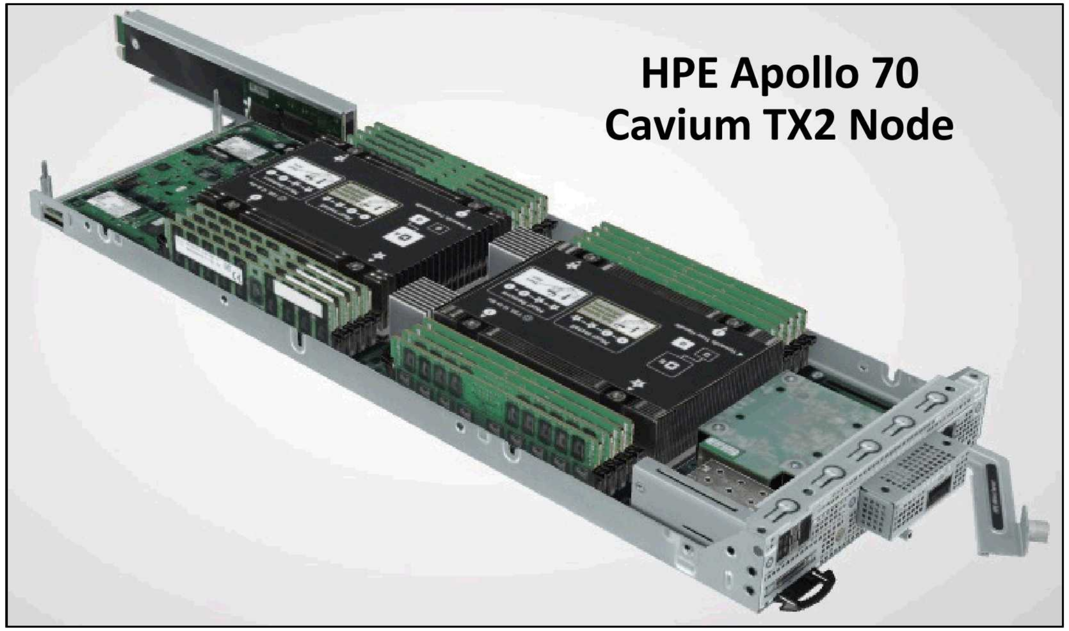
2592 compute nodes
(5184 TX2 processors)

3 IB spine switches
(each 540-port)

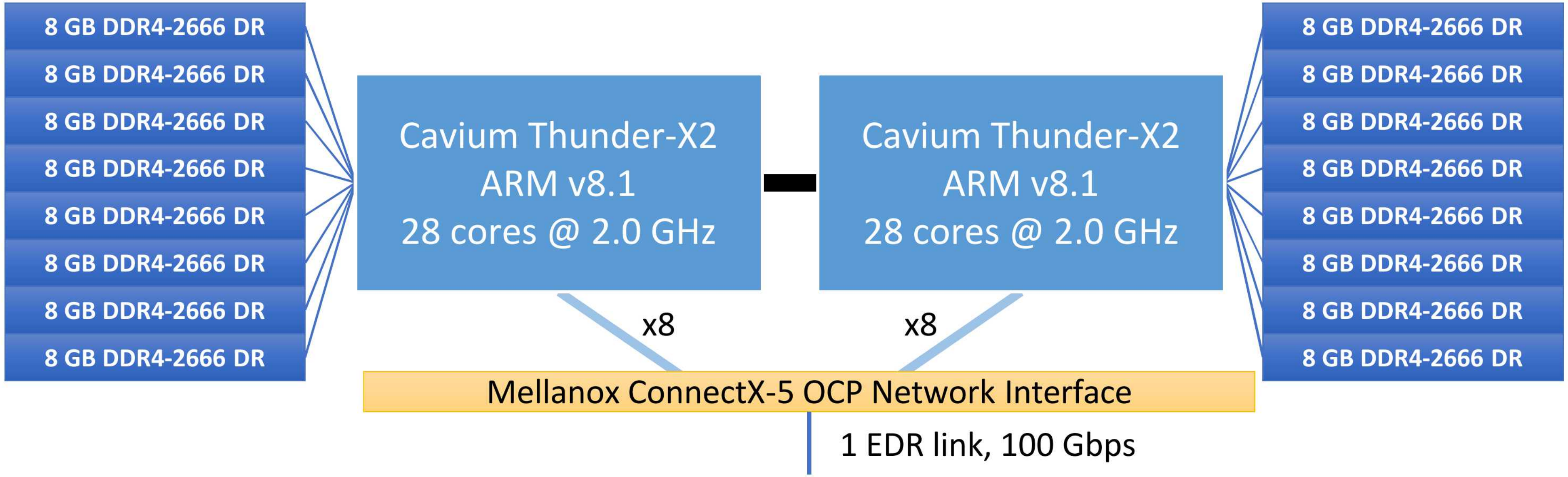
VANGUARD
Astra

Astra Architecture

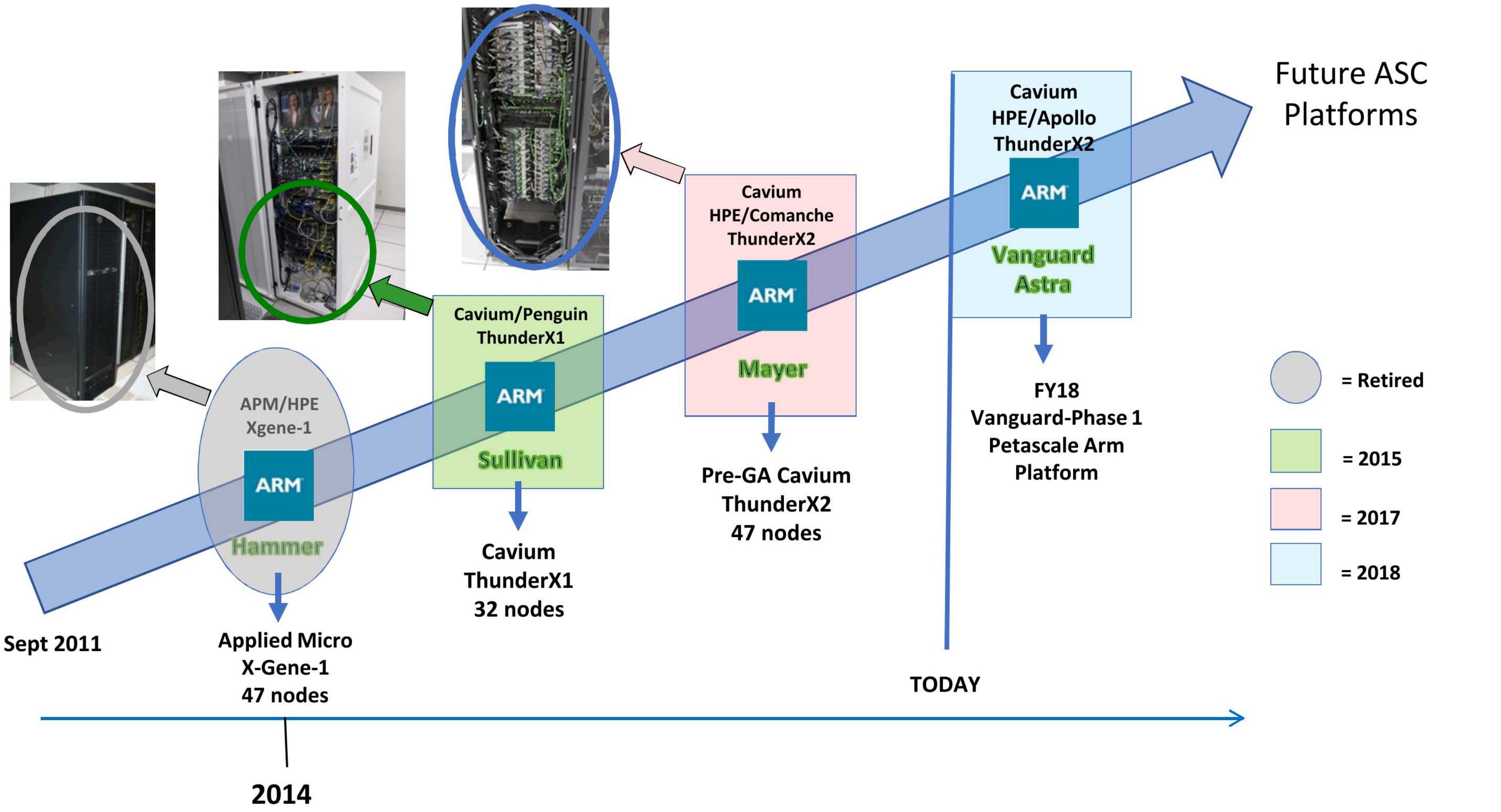
- **2,592** HPE Apollo 70 compute nodes
  - Cavium Thunder-X2 **Arm** SoC, 28 core, 2.0 GHz
  - 5,184 CPUs, 145,152 cores, 2.3 PFLOPs system peak
  - 128GB DDR Memory per node **(8 memory channels per socket)**
  - Aggregate capacity: 332 TB, Aggregate Bandwidth: 885 TB/s
- Mellanox IB EDR, ConnectX-5
- HPE Apollo 4520 All–flash storage, Lustre parallel file-system
  - Capacity: 403 TB (usable)
  - Bandwidth 244 GB/s



HPE Apollo 70 Cavium TX2 Node



Sandia’s NNSA/ASC ARM Platform Evolution



Sept 2011

APM/HPE Xgene-1

ARM

Hammer

Applied Micro X-Gene-1 47 nodes

2014

Cavium/Penguin ThunderX1

ARM

Sullivan

Cavium ThunderX1 32 nodes

Cavium HPE/Comanche ThunderX2

ARM

Mayer

Pre-GA Cavium ThunderX2 47 nodes

Cavium HPE/Apollo ThunderX2

ARM

Vanguard Astra

FY18 Vanguard-Phase 1 Petascale Arm Platform

TODAY

Future ASC Platforms

● = Retired

🟩 = 2015

🟥 = 2017

🟦 = 2018

ATSE: OpenHPC Evaluation

- Deployed OpenHPC on Mayer Arm-based testbed at Sandia
  - 47 compute nodes, dual-socket Cavium ThunderX2 28-core @ 2.0 GHZ
- Identified several gaps:
  - Focused on providing latest version of a given package
    - E.g., OpenHPC 1.3.5 moved to gcc7.3.0
      SNL validating with gcc7.2.0, need to use that version
  - Difficult to install multiple versions of a given package
  - Lacks architecture-optimized builds
    - HPL is 4.8x faster when compiled with an OpenBLAS targeting CaviumTX2 vs. OpenHPC OpenBLAS
  - Doesn’t support static linking (build recipes actively remove static libraries)
    - Many users like to build static binaries, ship single binary to classified
  - Hard to rebuild packages due to reliance on Open Build Service

Engaging with OpenHPC community to address gaps

Vanguard Program

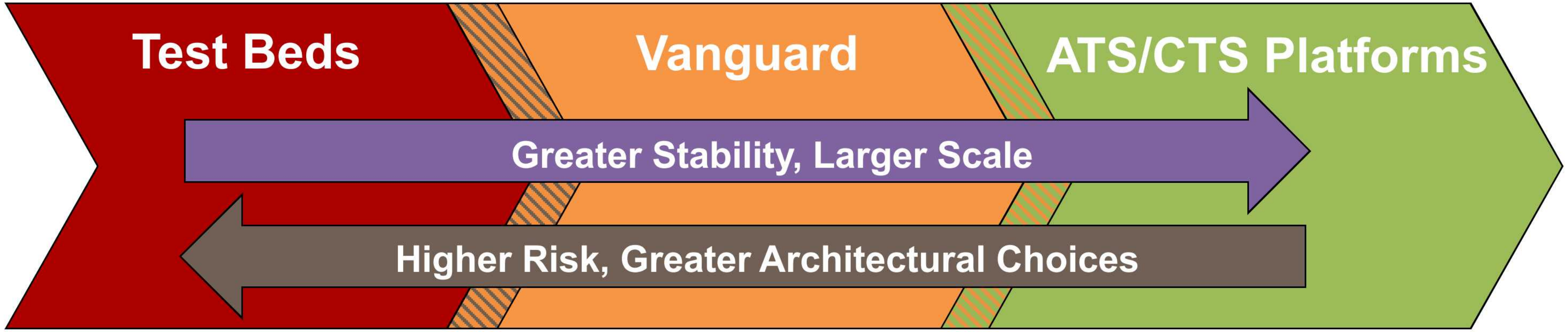
A proving ground for next-generation HPC technologies in support of the NNSA mission

Vanguard Program: Advanced Technology Prototype Systems

- **Prove viability of advanced technologies for NNSA integrated codes, at scale**
- Expand the HPC-ecosystem by developing emerging yet-to-be proven technologies
  - Is technology viable for future ATS/CTS platforms supporting ASC mission?
  - Increase technology AND integrator choices
- Buy down risk and increase technology and vendor choices for future NNSA production platforms
  - Ability to accept higher risk allows for more/faster technology advancement
  - Lowers/eliminates mission risk and significantly reduces investment
- Jointly address hardware and software technologies
- First Prototype platform targeting Arm Architecture

Success achieved through Tri-Lab involvement and collaboration

Where Vanguard Fits



Test Beds
Vanguard
ATS/CTS Platforms
Greater Stability, Larger Scale
Higher Risk, Greater Architectural Choices

Test Beds

- Small testbeds (~10-100 nodes)
- Breadth of architectures Key
- Brave users

Vanguard

- Larger-scale experimental systems
- Focused efforts to mature new technologies
- Broader user-base
- Not Production
- **Tri-lab resource but not for ATCC runs**

ATS/CTS Platforms

- Leadership-class systems (Petascale, Exascale, ...)
- Advanced technologies, sometimes first-of-kind
- Broad user-base
- PRODUCTION USE

Vanguard Program: Tri-Lab Software Effort (ATSE)

- Advanced Tri-lab Software Environment
- Accelerate maturity of ARM ecosystem for ASC computing
  - Prove viability for NNSA integrated codes running at scale
  - Harden compilers, math libraries, tools, communication libraries
    - Heavily templated C++, Fortran 2003/2008, Gigabyte+ binaries, long compiles
  - Optimize performance, verify expected results
- Build integrated software stack
  - Programming env (compilers, math libs, tools, MPI, OMP, SHMEM, I/O, ...)
  - Low-level OS (HPC-optimized Linux, network stack, filesystems, containers/VMs, ...)
  - Job scheduling and management (WLM, scalable app launcher, user tools, ...)
  - System management (OS image management, boot, system monitoring, ...)
- Leverage prototype aspect of system for scalable system software R&D

Improve 0 to 60 time... Vanguard-Astra arrival to useful work done

Artist Rendering

Construction Completed < 1 Year

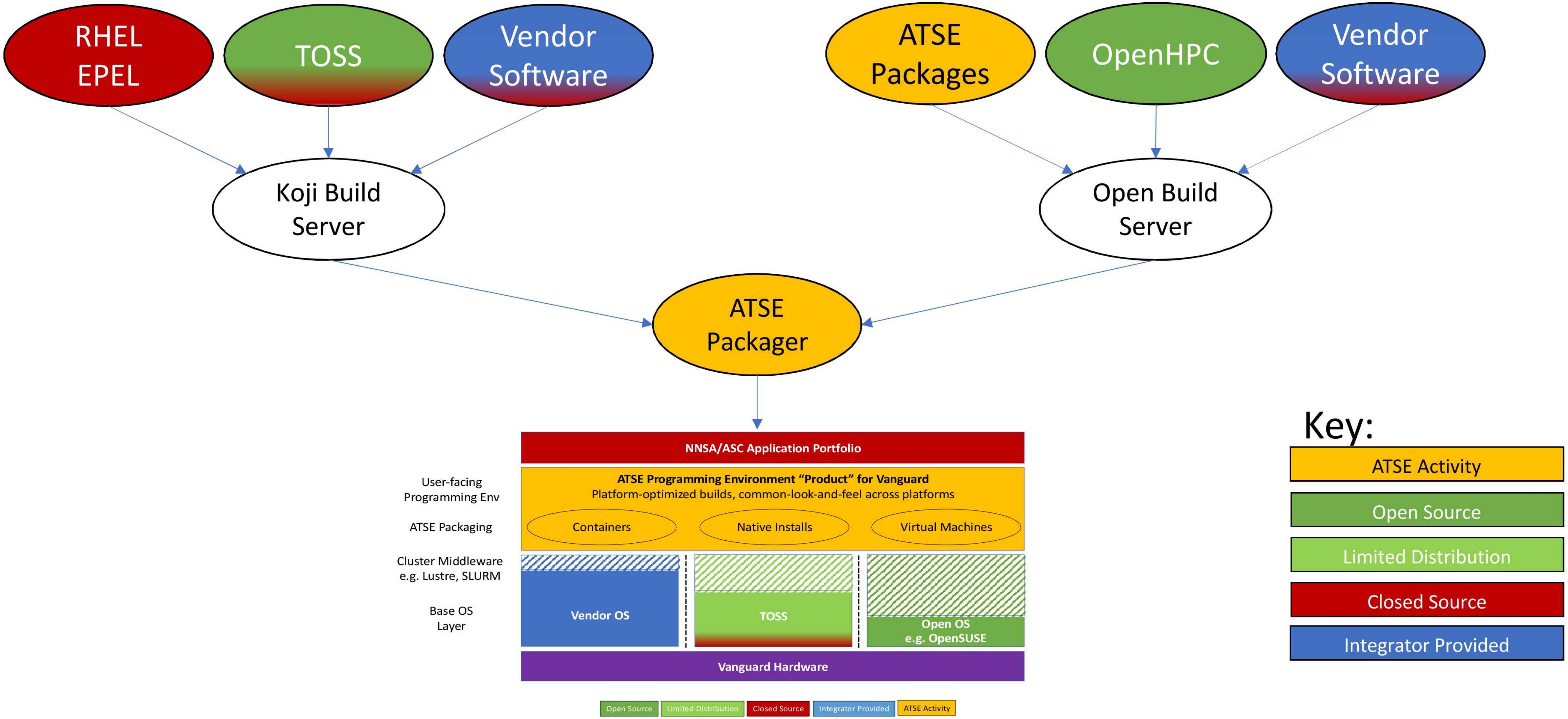
Celebrity Groundbreaking

Delivery and Integration Overlapped Constructic

ATSE: Goals

- Support NNSA mission applications, test against realistic input problems
- Be able to operate disconnected from the public internet
- Operate stand-alone or integrated with existing vendor infrastructure
- Support multiple processor architectures (arm64, x86_64, ppc64le, ...)
- Support multiple system architectures (Linux clusters, Cray, IBM, ...)
- Support multiple compiler toolchains (gcc, arm, intel, xlc, ...)
- Support multiple versions+configs of software packages with user selection
- Support both static and dynamic linking of applications
- Build static libraries with –fPIC
- Provide architecture-optimized packages
- Distribute container and virtual machine images of each ATSE release
- Source included, easy to rebuild and replace any non-proprietary package
- Provide open version of ATSE that is distributed publicly

ATSE: Pulling Components from Many Sources



RHEL EPEL

TOSS

Vendor Software

ATSE Packages

OpenHPC

Vendor Software

Koji Build Server

Open Build Server

ATSE Packager

NNSA/ASC Application Portfolio

User-facing Programming Env

ATSE Programming Environment "Product" for Vanguard

Platform-optimized builds, common-look-and-feel across platforms

ATSE Packaging

Containers

Native Installs

Virtual Machines

Cluster Middleware e.g. Lustre, SLURM

Base OS Layer

Vendor OS

TOSS

Open OS e.g. OpenSUSE

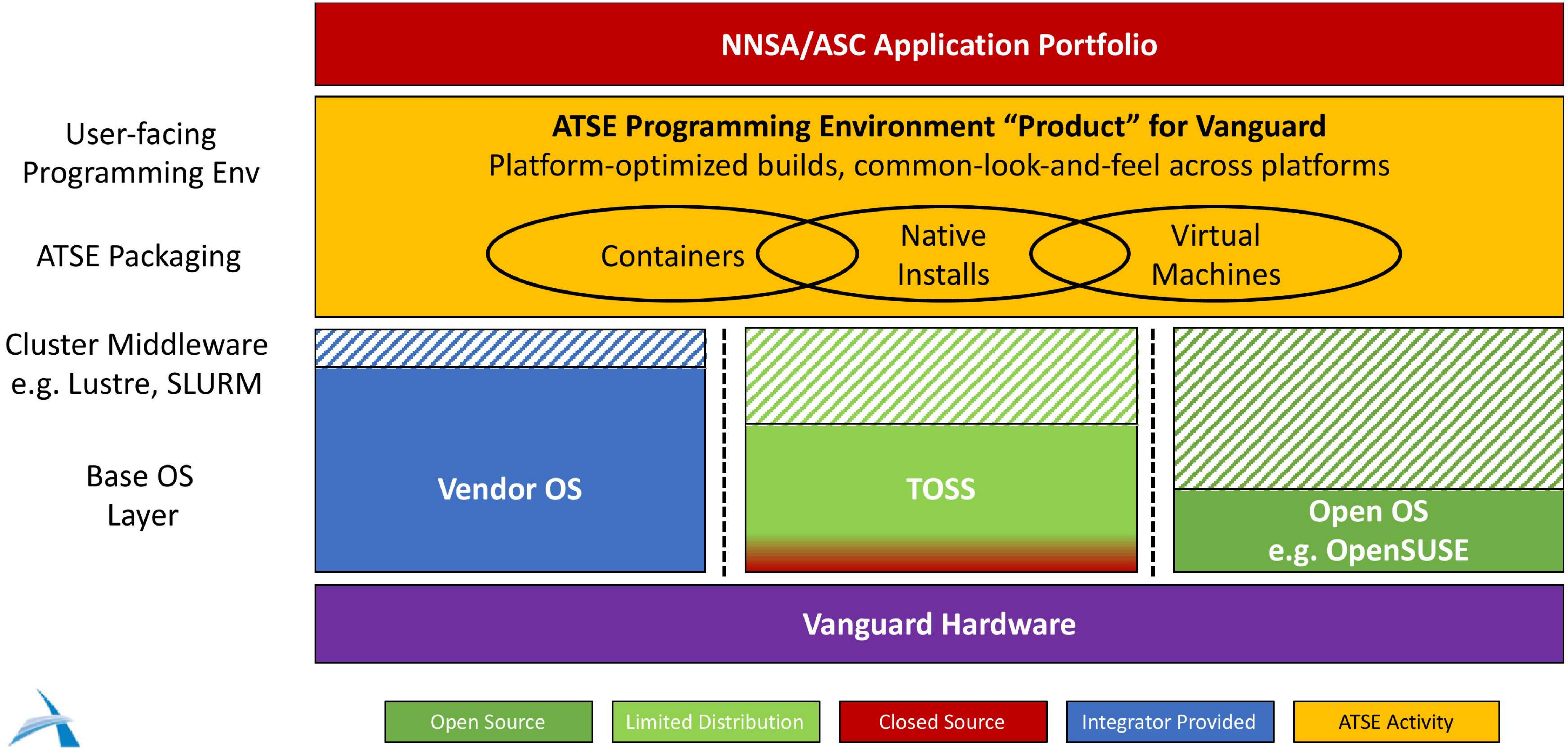
Vanguard Hardware

Open Source | Limited Distribution | Closed Source | Integrator Provided | ATSE Activity

Key:

- ATSE Activity
- Open Source
- Limited Distribution
- Closed Source
- Integrator Provided

ATSE: Integration with Multiple Base Operating Systems



User-facing Programming Env

ATSE Packaging

Cluster Middleware e.g. Lustre, SLURM

Base OS Layer

NNSA/ASC Application Portfolio

ATSE Programming Environment “Product” for Vanguard
Platform-optimized builds, common-look-and-feel across platforms

Containers

Native Installs

Virtual Machines

Vendor OS

TOSS

Open OS
e.g. OpenSUSE

Vanguard Hardware

🟩 Open Source

🟩 Limited Distribution

🟥 Closed Source

🟦 Integrator Provided

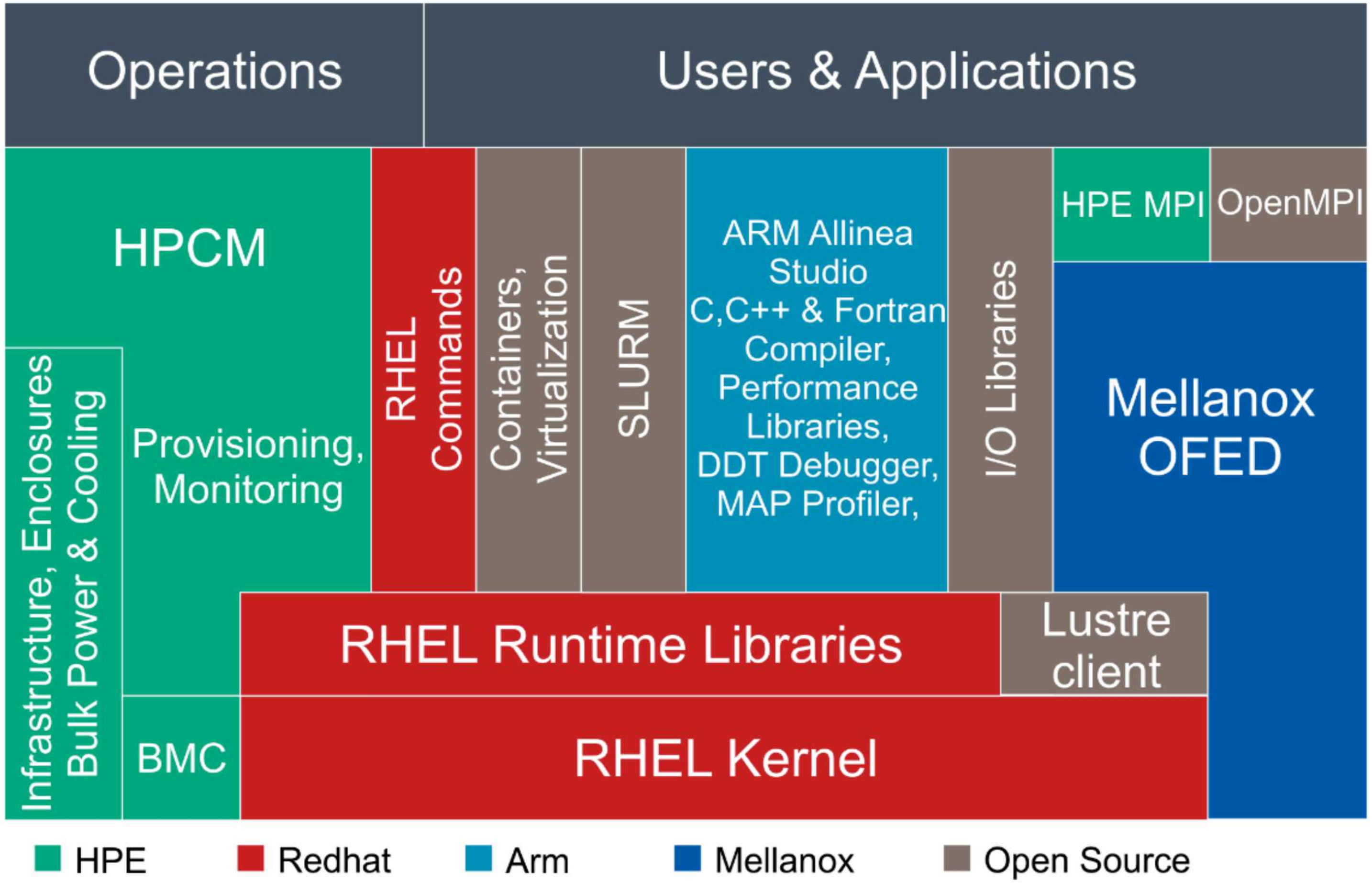
🟨 ATSE Activity

ATSE: Collaboration with HPE OLSS Effort

Hewlett Packard Enterprise

Open Leadership Software Stack (OLSS)

- HPE:
  - HPE MPI (+ XPMEM)
  - HPE Cluster Manager
- Arm:
  - Arm HPC Compilers
  - Arm Math Libraries
  - Allinea Tools
- Mellanox-OFED & HPC-X
- RedHat 7.x for aarch64



Operations

Users & Applications

HPCM

Infrastructure, Enclosures Bulk Power & Cooling

Provisioning, Monitoring

BMC

RHEL Commands

Containers, Virtualization

SLURM

ARM Allinea Studio C,C++ & Fortran Compiler, Performance Libraries, DDT Debugger, MAP Profiler,

I/O Libraries

HPE MPI

OpenMPI

Mellanox OFED

RHEL Runtime Libraries

Lustre client

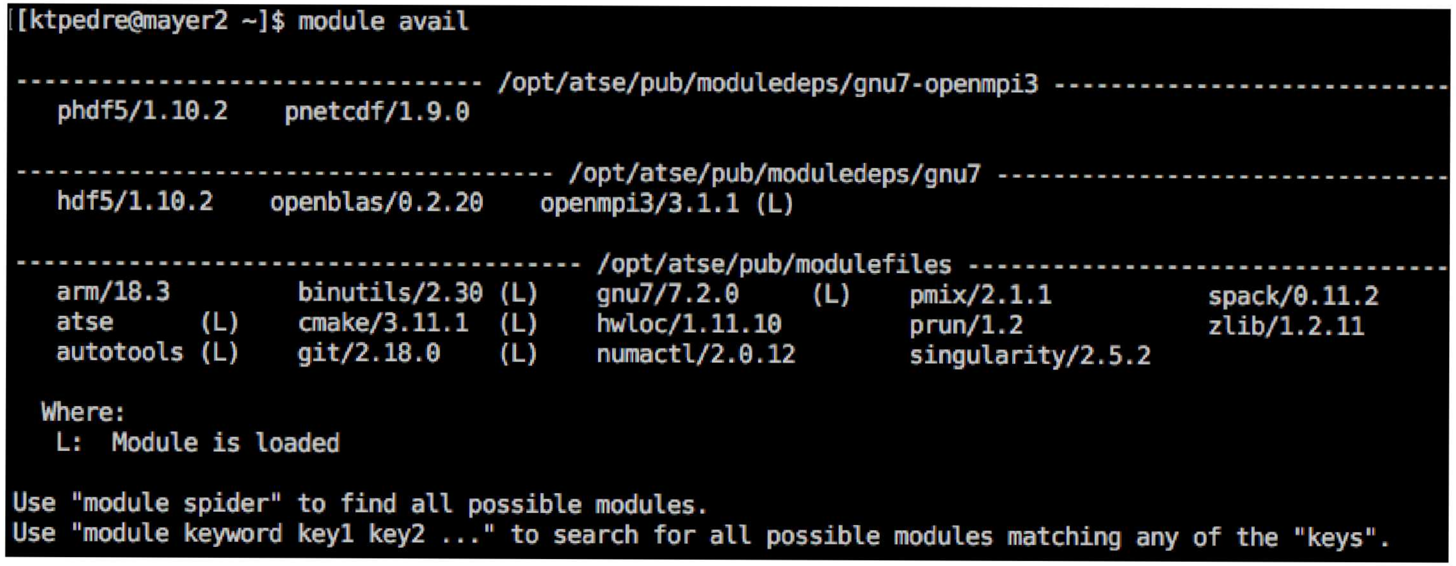
RHEL Kernel

🟩 HPE 🟥 Redhat 🟦 Arm 🟦 Mellanox 🟫 Open Source

ATSE: Deployed Beta Stack

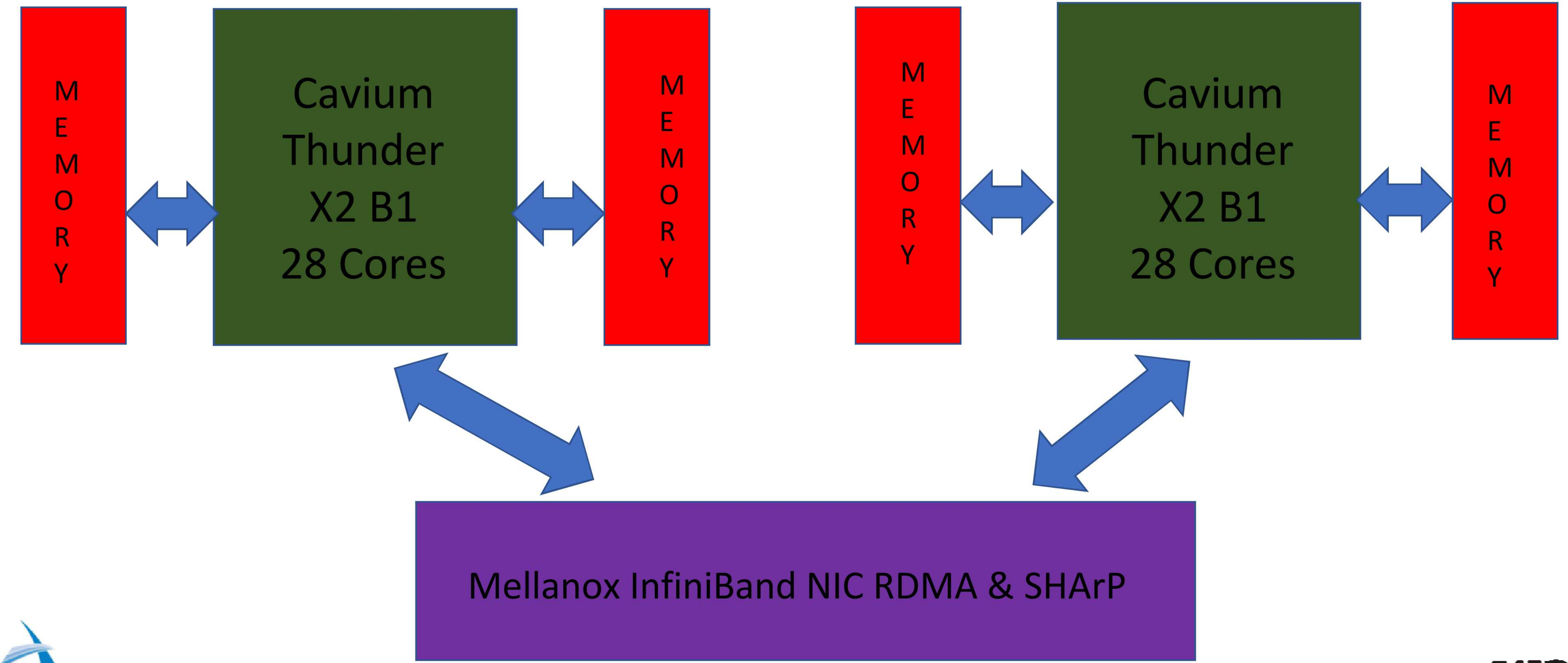
- Setup local Open Build Service (OBS) build farm at Sandia
- Built set of software packages needed for Astra milestone 1
  - When OpenHPC recipe was available, we tried to use it, modifying as necessary
  - Otherwise, we built a new build recipe in same style as OpenHPC
- Installed on Mayer testbed at Sandia, now using as the default user environment
- Tested with STREAM, HPL, HPCG, ASC mini-apps
- Compiler toolchain support
  - GNU compilers, 7.2.0
  - ARM HPC compilers

ATSE Modules Interface, Mirrors OpenHPC



```
[[ktpedre@mayer2 ~]$ module avail

---------------------------------- /opt/atse/pub/moduledeps/gnu7-openmpi3 --------------------------
   phdf5/1.10.2    pnetcdf/1.9.0

-------------------------------------- /opt/atse/pub/moduledeps/gnu7 --------------------------------
   hdf5/1.10.2    openblas/0.2.20    openmpi3/3.1.1 (L)

---------------------------------------- /opt/atse/pub/modulefiles ----------------------------------
   arm/18.3             binutils/2.30 (L)    gnu7/7.2.0      (L)    pmix/2.1.1             spack/0.11.2
   atse      (L)        cmake/3.11.1  (L)    hwloc/1.11.10          prun/1.2               zlib/1.2.11
   autotools (L)        git/2.18.0    (L)    numactl/2.0.12         singularity/2.5.2

  Where:
   L:  Module is loaded

Use "module spider" to find all possible modules.
Use "module keyword key1 key2 ..." to search for all possible modules matching any of the "keys".
```

Astra-Network Communications Become More Important Due to the use of Multi-Core Sockets.


MEMORY
Cavium Thunder X2 B1 28 Cores
MEMORY
MEMORY
Cavium Thunder X2 B1 28 Cores
MEMORY
Mellanox InfiniBand NIC RDMA & SHArP

Astra-RDMA Network Improvements

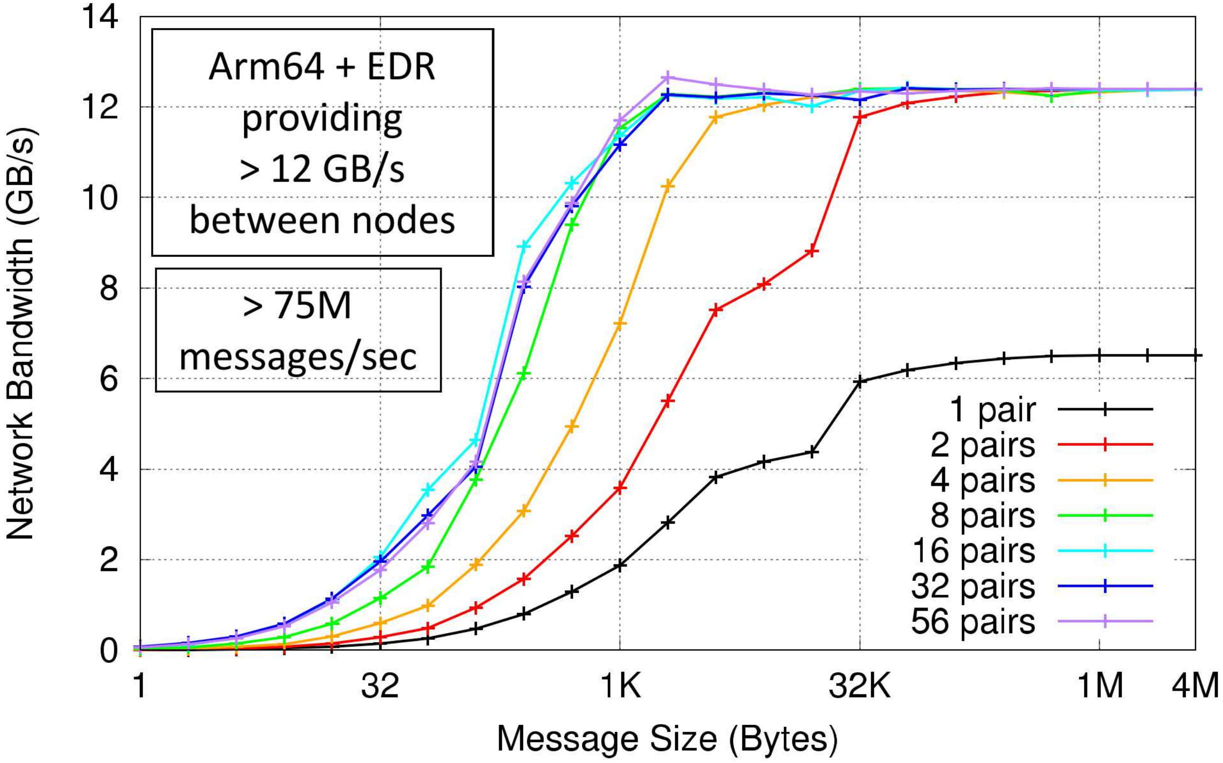
- New NIC interfaces increase memory bandwidth and reduce memory consumption
  - Hardware Collectives
  - Socket Direct allows cores from each socket to reach the RDMA fabric
  - UCX API

Socket Direct

Socket Direct feature enables a single NIC to be shared by multiple host processor sockets

- Share a single physical link to reduce cabling complexity and costs
- NIC arbitrates between host processors to ensure a fair level of service
- Required some complex O/S patches early on in test systems

OSU MPI Multi-Network Bandwidth



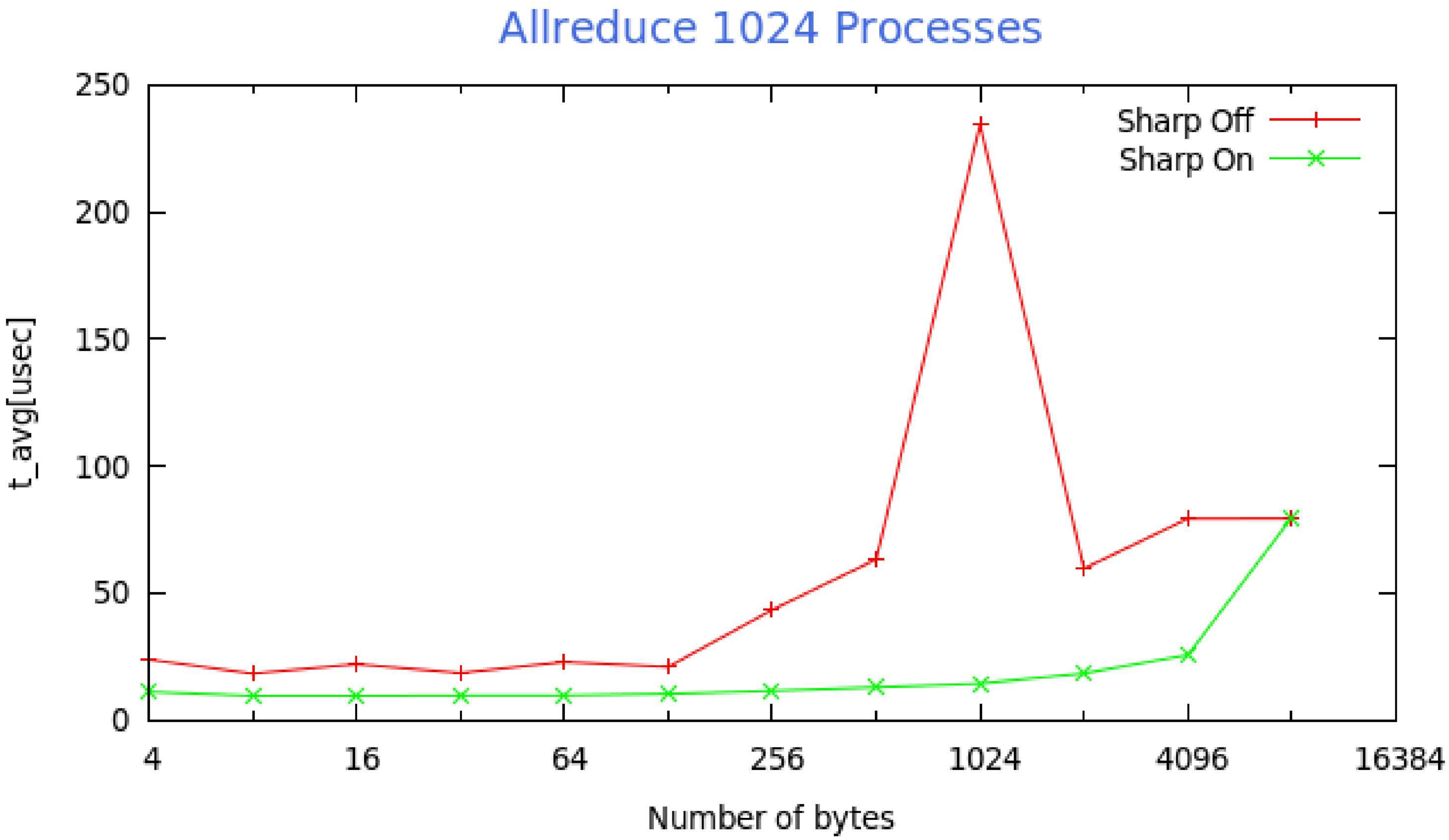
Arm64 + EDR providing > 12 GB/s between nodes

> 75M messages/sec

SHArP Hardware Collectives

- Reduce the amount of buffer space required for unexpected messages.
- Perform work, when possible, in the NIC
- Increase computational performance by off-loading work from the CPU scheduler
- Perform operations that are blocking computation in the NIC to help speed up the work

Astra Hardware Collective Performance



UCX API

- Astra is providing Sandia with experience using a large-scale implementation of UCX
  - ATSE OpenMPI is compiled to use the UCT UCX API
    - UCP abstacts multiple devices and transport layers
    - UCP provides message fragmentation and non-blocking operations.
    - UCP can use multiple connections for transport.

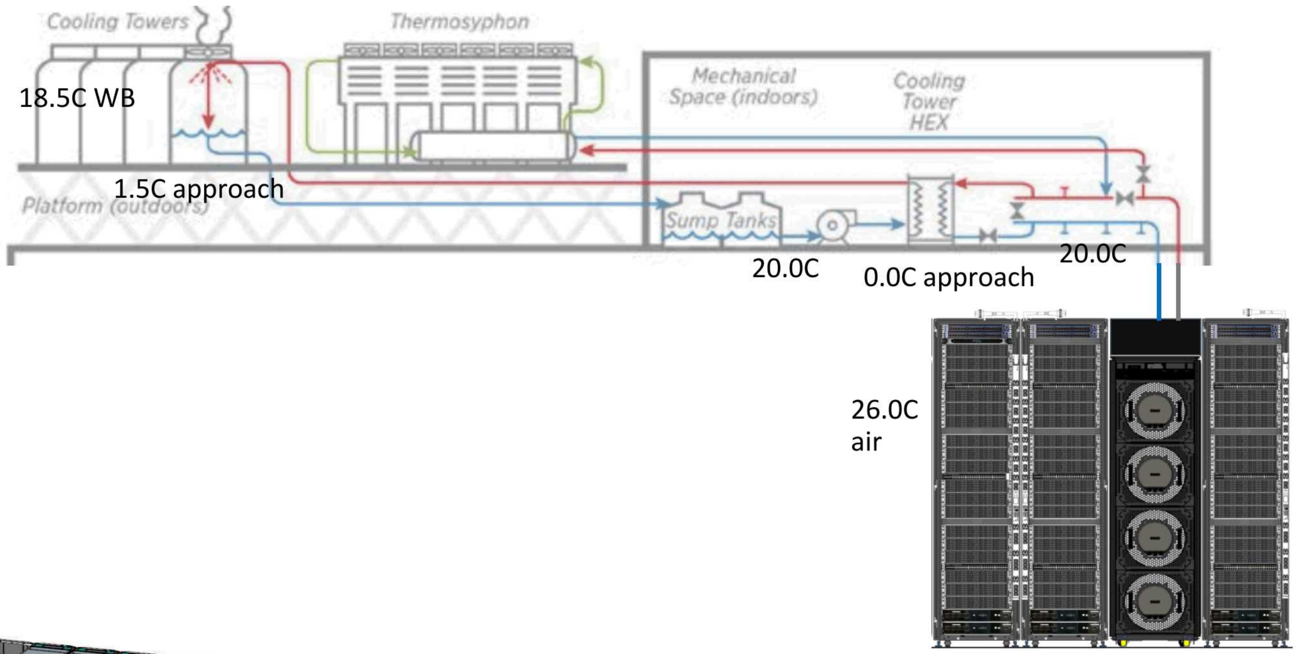
Astra Advanced Power and Cooling

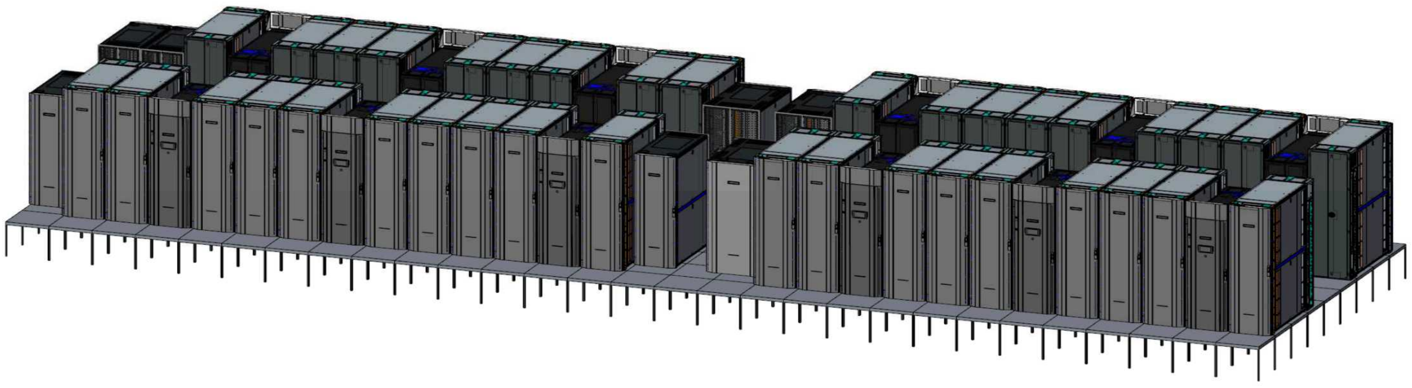
Extreme Efficiency:

- Total 1.2 MW in the 36 compute racks are cooled by only 12 fan coils
- These coils are cooled without compressors year round. No evaporative water at all almost 6000 hours a year
- 99% of the compute racks heat never leaves the cabinet, yet the system doesn't require the internal plumbing of liquid disconnects and cold plates running across all CPUs and DIMMs

Sandia Thermosyphon Cooler Hybrid System for Water Savings

Efficient tower and HEX can take hottest 36 hours of the year of 18.5C wetbulb to make 20C water to the fan coils

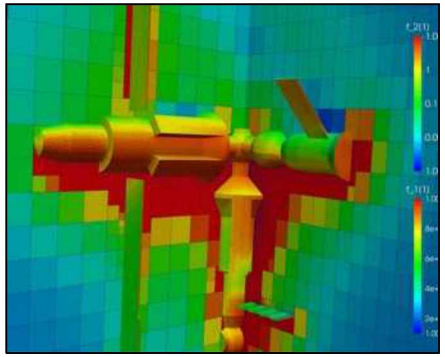


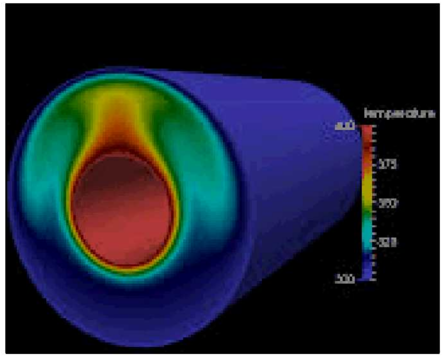


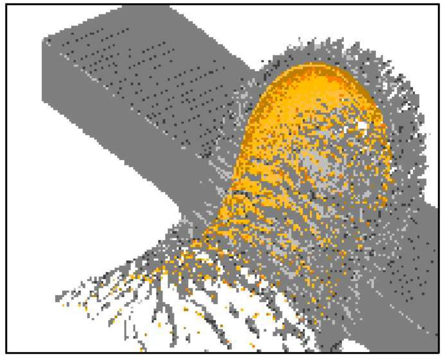
Projected power of the system by component										
	per constituent rack type (W)					total (kW)				
	wall	peak	nominal (linpack)	idle		racks	wall	peak	nominal (linpack)	idle
Node racks	39888	35993	33805	6761		36	1436.0	1295.8	1217.0	243.4
MCS300	10500	7400	7400	170		12	126.0	88.8	88.8	2.0
Network	12624	10023	9021	9021		3	37.9	30.1	27.1	27.1
Storage	11520	10000	10000	1000		2	23.0	20.0	20.0	2.0
utility	8640	5625	4500	450		1	8.6	5.6	4.5	0.5
							1631.5	1440.3	**1357.3**	274.9

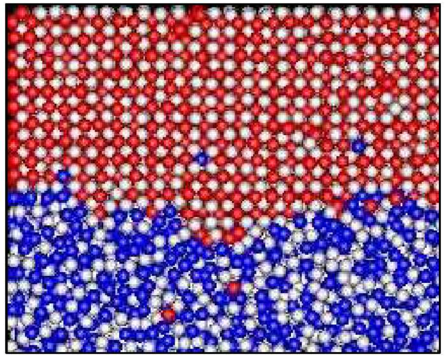
Early Results from Astra

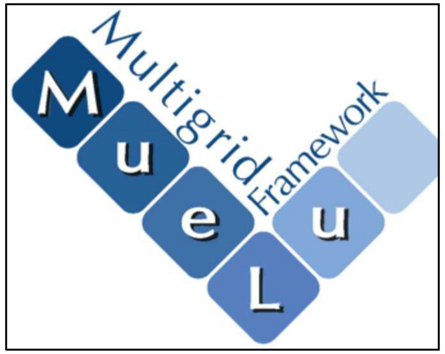
Baseline: Trinity ASC Platform (Current Production), dual-socket Haswell





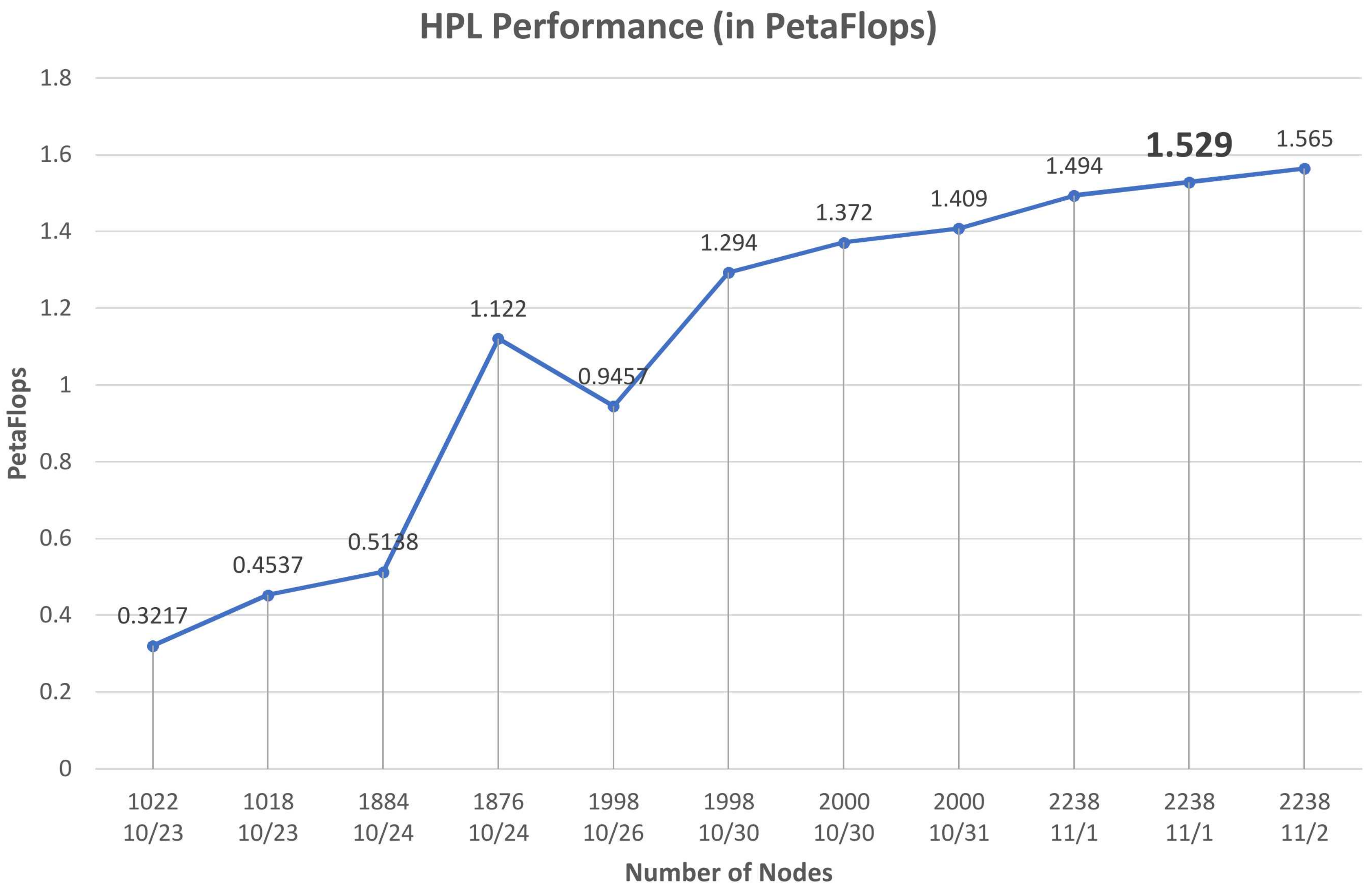




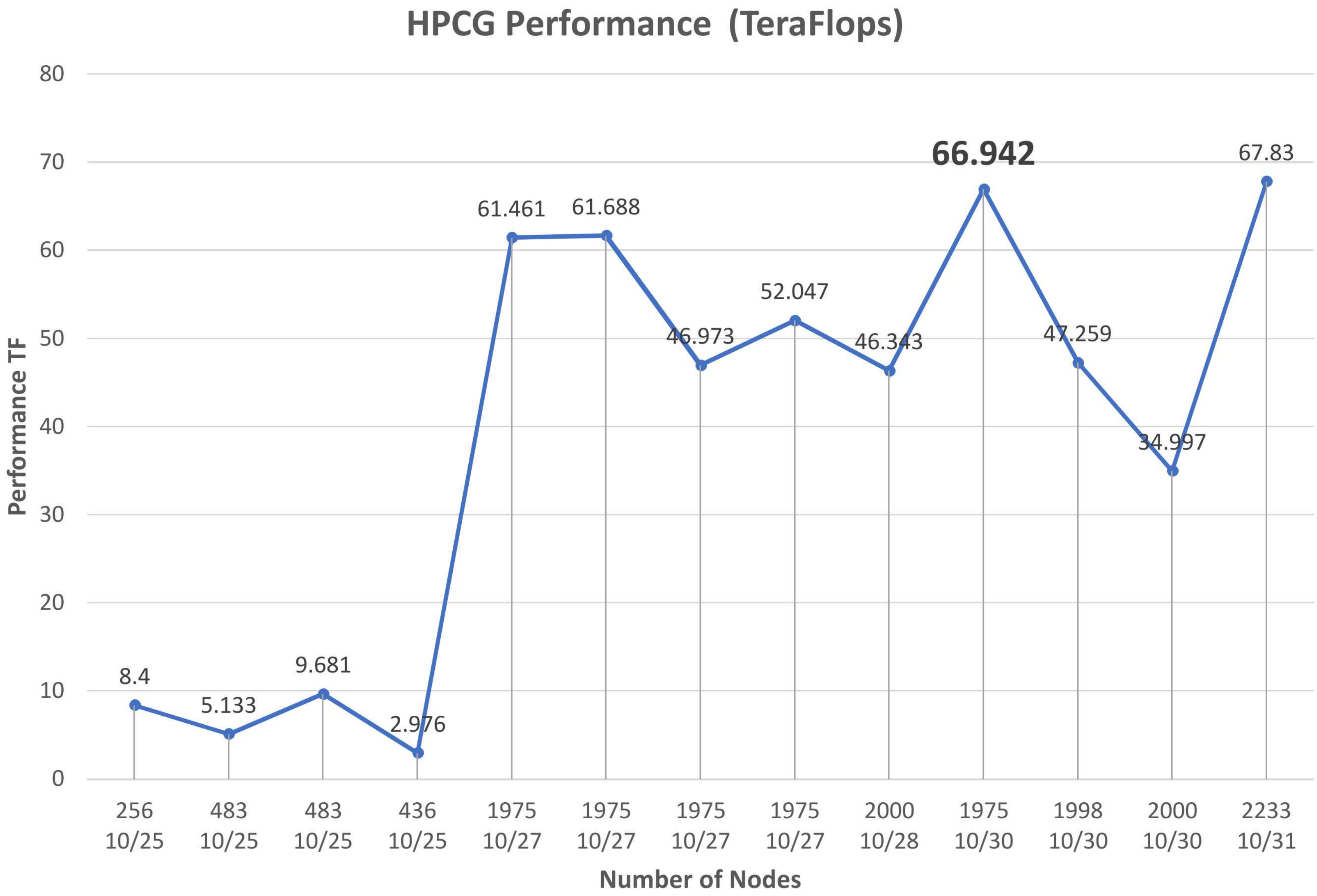


Monte Carlo	CFD Models	Hydrodynamics	Molecular Dynamics	Linear Solvers
1.60X	1.45X	1.30X	1.42X	1.87X

Initial Large Scale Testing and Benchmarks (HPL)



Initial Large Scale Testing and Benchmarks (HPCG)



ATSE: R&D Efforts

- Workflows leveraging containers and virtual machines
  - Support for machine learning frameworks
  - ARMv8.1 includes new virtualization extensions, SR-IOV
- Evaluating parallel filesystems + I/O systems @ scale
  - GlusterFS, Ceph, BeeGFS, Sandia Data Warehouse, …
- Resilience studies over Astra lifetime
- Improved MPI thread support, matching acceleration
- OS optimizations for HPC @ scale
  - Exploring spectrum from stock distro Linux kernel to HPC-tuned Linux kernels to non-Linux lightweight kernels and multi-kernels
  - Arm-specific optimizations

ATSE: Next Steps

- Continue adding and optimizing packages and build test framework
- Package container and VM images
  - Lab-internal version, hosted on Sandia Gitlab Docker registry
  - Externally distributable versions, stripping out proprietary components
- Submit ”trial-run” patches filling gaps to OpenHPC community
- Explore Spack build and packaging
- Continue collaboration with HPE OLSS team, first external ATSE customer

- LANL: CharlieCloud/Bee support, LLVM compiler work, application readiness
- LLNL: TOSS, “Spack Stacks” ATSE build, application readiness

Questions?

ASTRA
"Per aspera ad astra"
Sandia National Laboratories
NNSA
Advanced Simulation & Computing
Hewlett Packard Enterprise
WESTWIND
U.S. DEPARTMENT OF ENERGY
Owens Imagery
VANGUARD

Exceptional Service in the National Interest