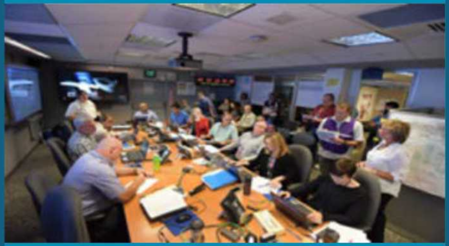
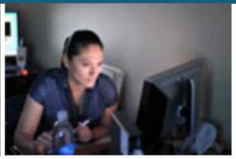
This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2019-6823C

VANGUARD

Vanguard Astra - Petascale ARM Platform for U.S. DOE/ASC Supercomputing





PRESENTED BY

Rob Hoekstra

Kevin Pedretti, Si Hammond, James Laros, Andrew Younge, Paul Lin, Courtney Vaughan

Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

SAND2019-XXXX C
Unclassified Unlimited Release

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.

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories

Vanguard Overview

Vanguard Program: Goals and Aims

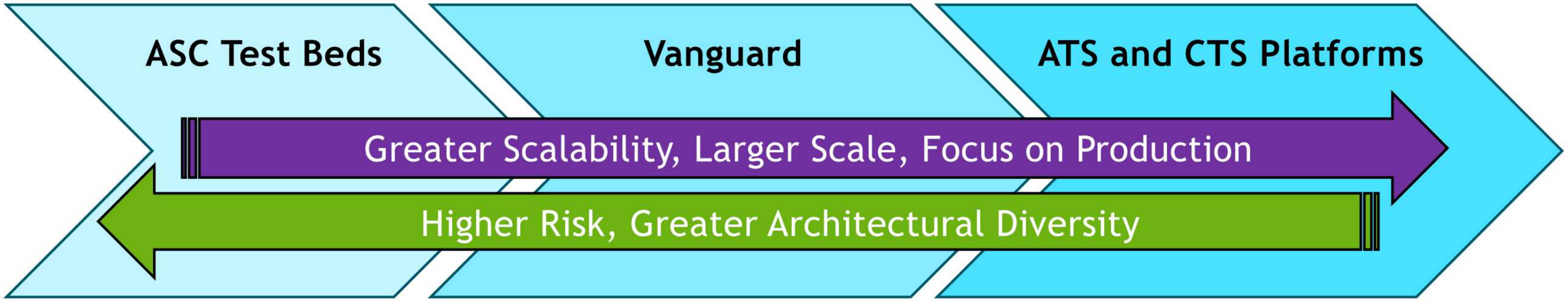
Prove viability of advanced technologies for NNSA/ASC integrated codes, at scale

- 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

Where Vanguard Fits in our Program Strategy


ASC Test Beds
Vanguard
ATS and CTS Platforms
Greater Scalability, Larger Scale, Focus on Production
Higher Risk, Greater Architectural Diversity

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

Sandia has a history with Arm as testbeds

2014

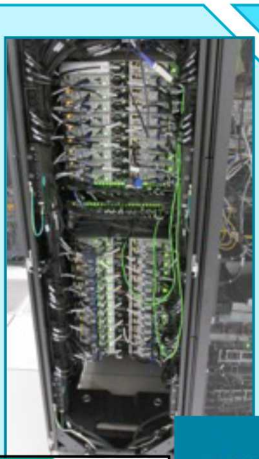
Hammer

Applied Micro
X-Gene-1
47 nodes

Sullivan

Cavium ThunderX1
32 nodes

2017



Mayer

Pre-GA Cavium
ThunderX2
47 nodes

2018

Vanguard/Astra

HPE Apollo 70
Cavium ThunderX2
2592 nodes

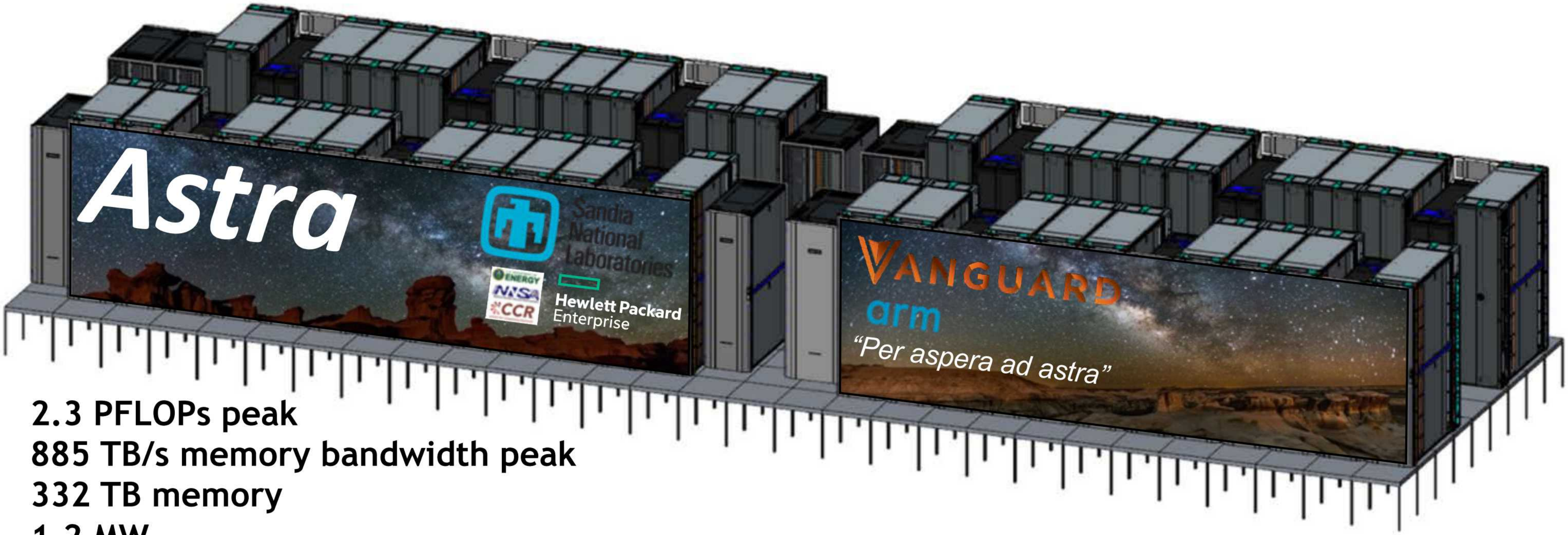
The World's First Petascale Arm Supercomputer

Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY NNSA

per aspera ad astra

through difficulties to the stars



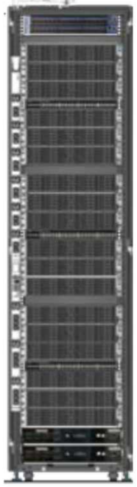
2.3 PFLOPs peak
885 TB/s memory bandwidth peak
332 TB memory
1.2 MW

Demonstrate viability of ARM for U.S. DOE Supercomputing

Vanguard-Astra System Packaging

HPE Apollo 70 Chassis: 4 nodes

HPE Apollo 70 Rack



18 chassis/rack

72 nodes/rack

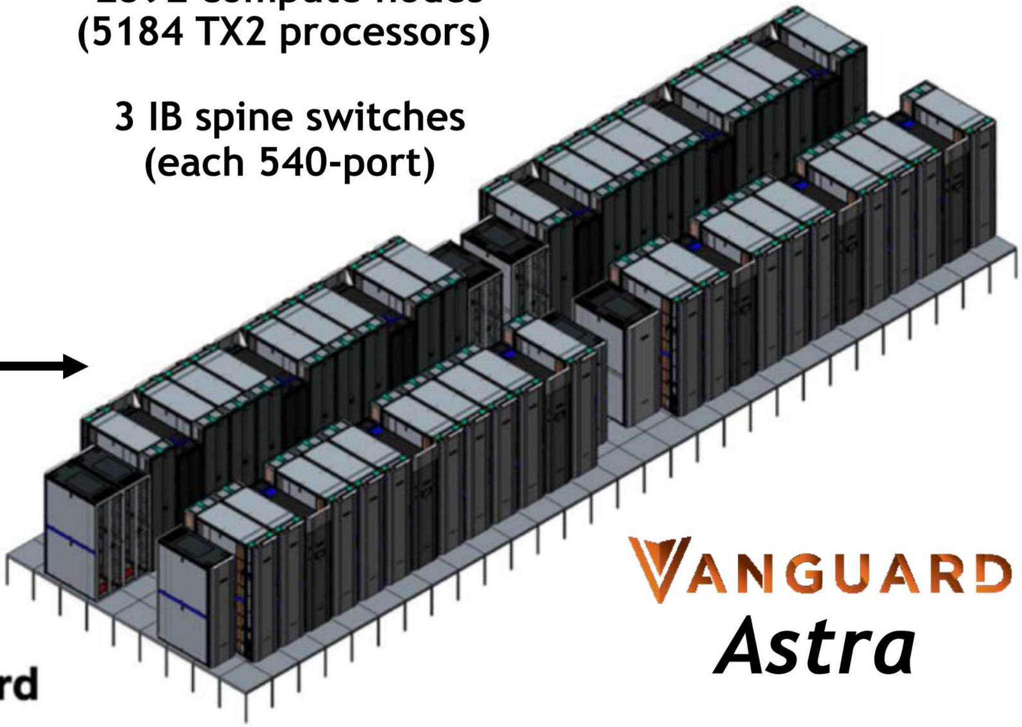
3 IB switches/rack
(one 36-port switch
per 6 chassis)

Hewlett Packard Enterprise

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

2592 compute nodes
(5184 TX2 processors)

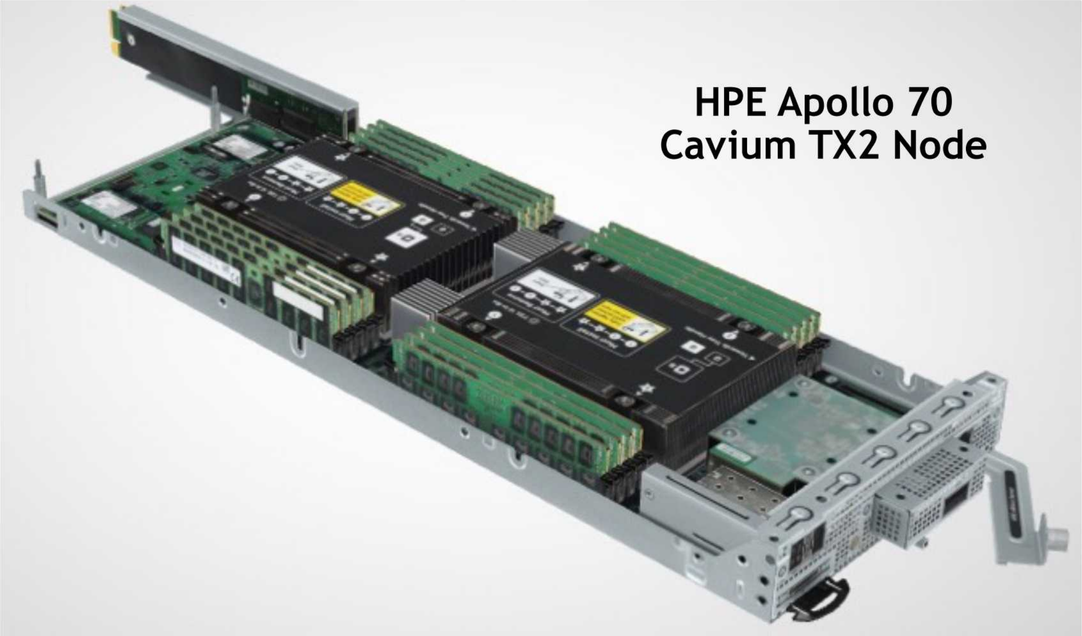
3 IB spine switches
(each 540-port)



VANGUARD
Astra

Vanguard-Astra Compute Node Building Block

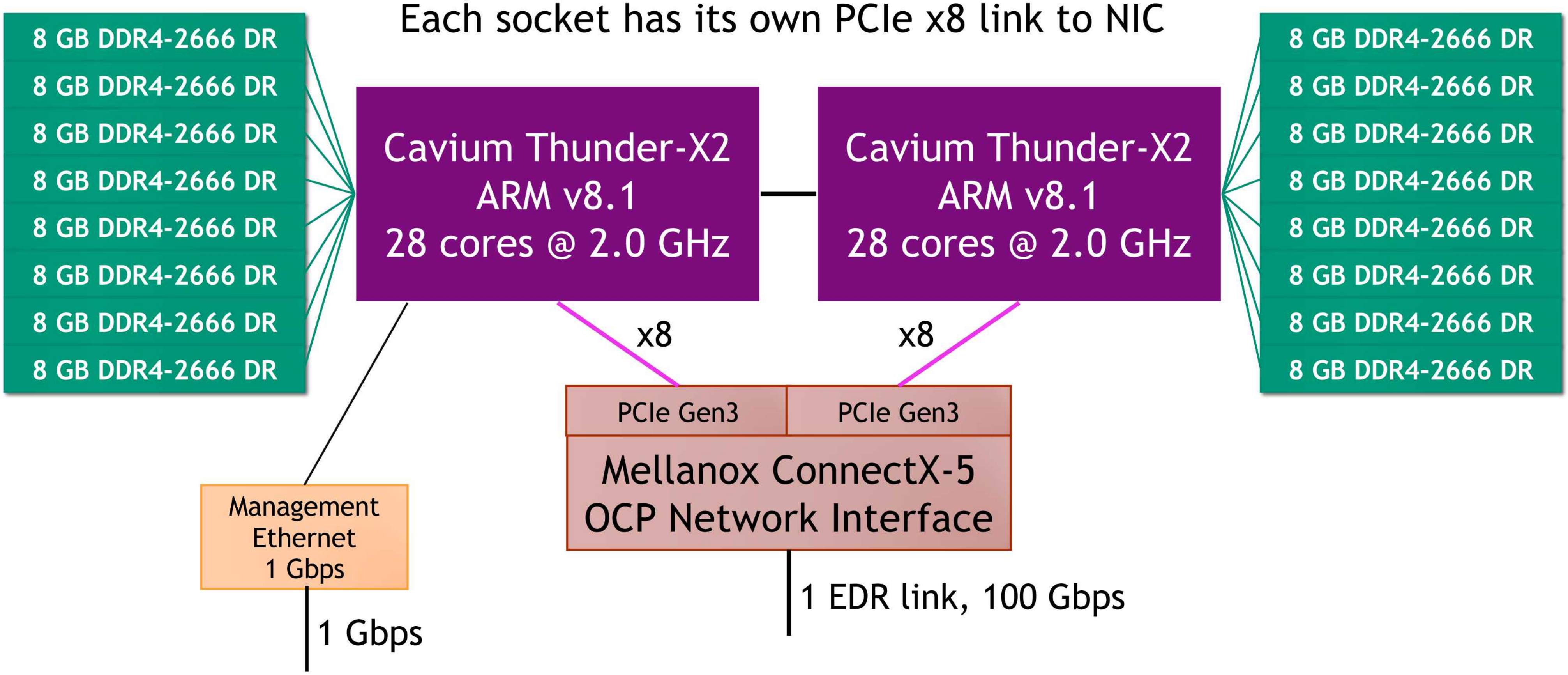
- Dual socket Cavium Thunder-X2
  - CN99xx
  - 28 cores @ 2.0 GHz
- 8 DDR4 controllers per socket
- One 8 GB DDR4-2666 dual-rank DIMM per controller
- Mellanox EDR InfiniBand ConnectX-5 VPI OCP
- Tri-Lab Operating System Stack based on RedHat 7.6+



HPE Apollo 70
Cavium TX2 Node

Vanguard-Astra Compute Node

8 DDR4 channels/socket, 1 DIMM/channel
Each socket has its own PCIe x8 link to NIC



8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR

Cavium Thunder-X2
ARM v8.1
28 cores @ 2.0 GHz

Cavium Thunder-X2
ARM v8.1
28 cores @ 2.0 GHz

8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR
8 GB DDR4-2666 DR

x8

x8

PCIe Gen3

PCIe Gen3

Mellanox ConnectX-5
OCP Network Interface

1 EDR link, 100 Gbps

Management
Ethernet
1 Gbps

1 Gbps

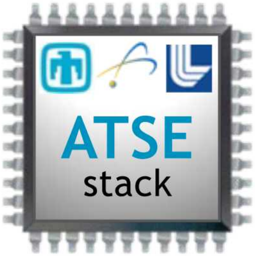
ATSE – Advanced Tri-lab Software Environment

Tri-Lab Software Effort for ARM

- Accelerate ARM ecosystem for DOE computing
  - Prove viability for ASC 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 environment (compilers, math libs, tools, MPI, OMP, I/O, ...)
  - Low-level OS (optimized Linux, network, filesystems, containers/VMs, ...)
  - Job scheduling and management (WLM, app launcher, user tools, ...)
  - System management (boot, system monitoring, image management, ...)

Advanced Tri-lab Software Environment (ATSE)

- Advanced Tri-lab Software Environment
  - Sandia leading development within DOE
  - Partnership across the ASC Labs and with HPE
  - Provide a user programming environment for Astra
    - Initial focus on ARM, have x86_64 port



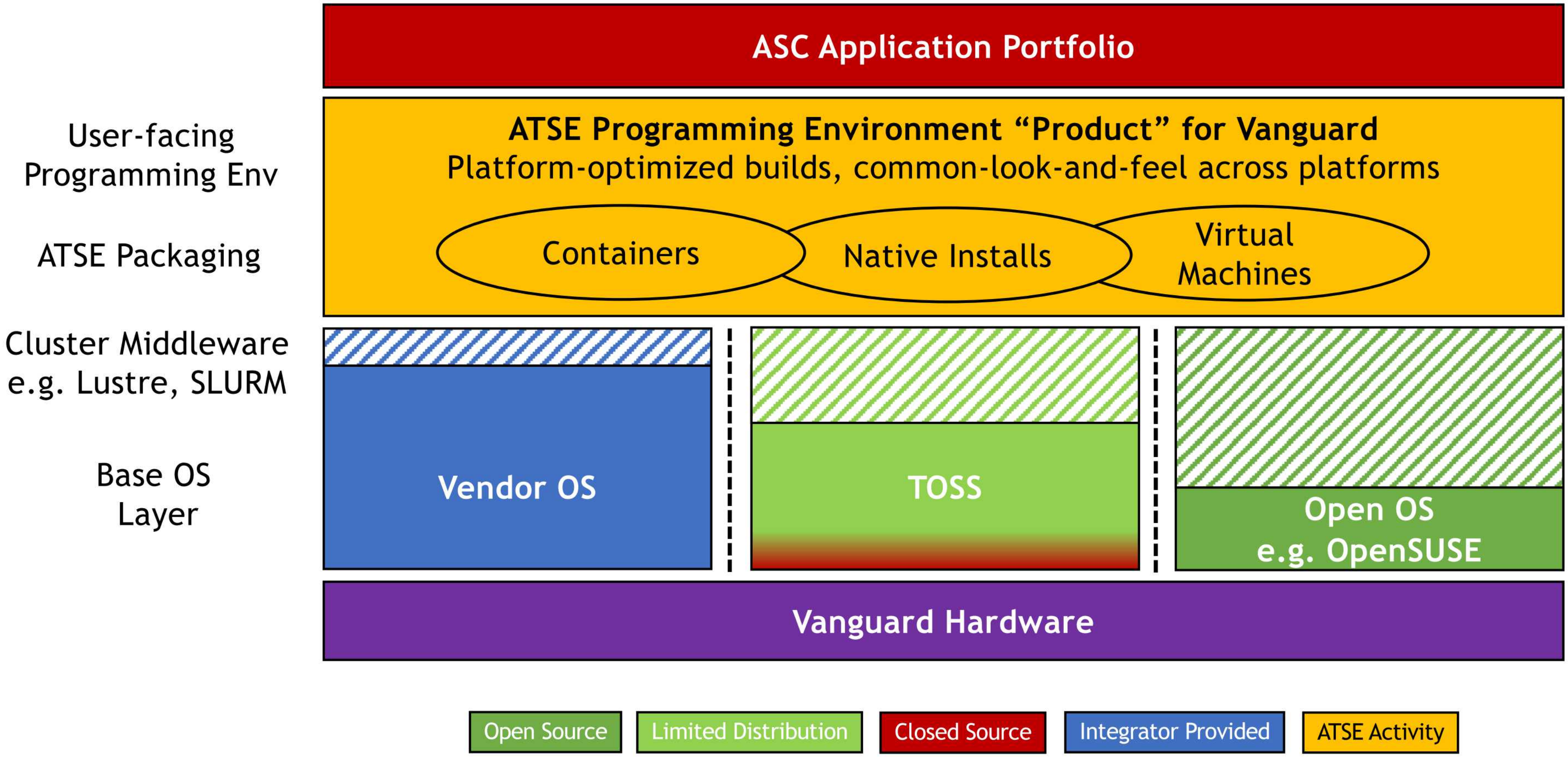
- Lasting value beyond Astra
  - Documented specification of:
    - Software components needed for HPC production applications
    - How they are configured (i.e., what features and capabilities are enabled) and interact
    - User interfaces and conventions
  - Reference implementation:
    - Deployable on multiple ASC systems and architectures with common look and feel
    - Tested against real workloads
    - Community inspired, focused and supported
    - Leveraging OpenHPC effort
    - Inform & improve vendor supplied software stack

ATSE is an integrated software environment for ASC workloads

ATSE R&D Efforts – Developing Next-Generation NNSA Workflows

- 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, …
- 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
- Resilience studies over Astra lifetime

ARM Tri-lab Software Environment (ATSE)



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

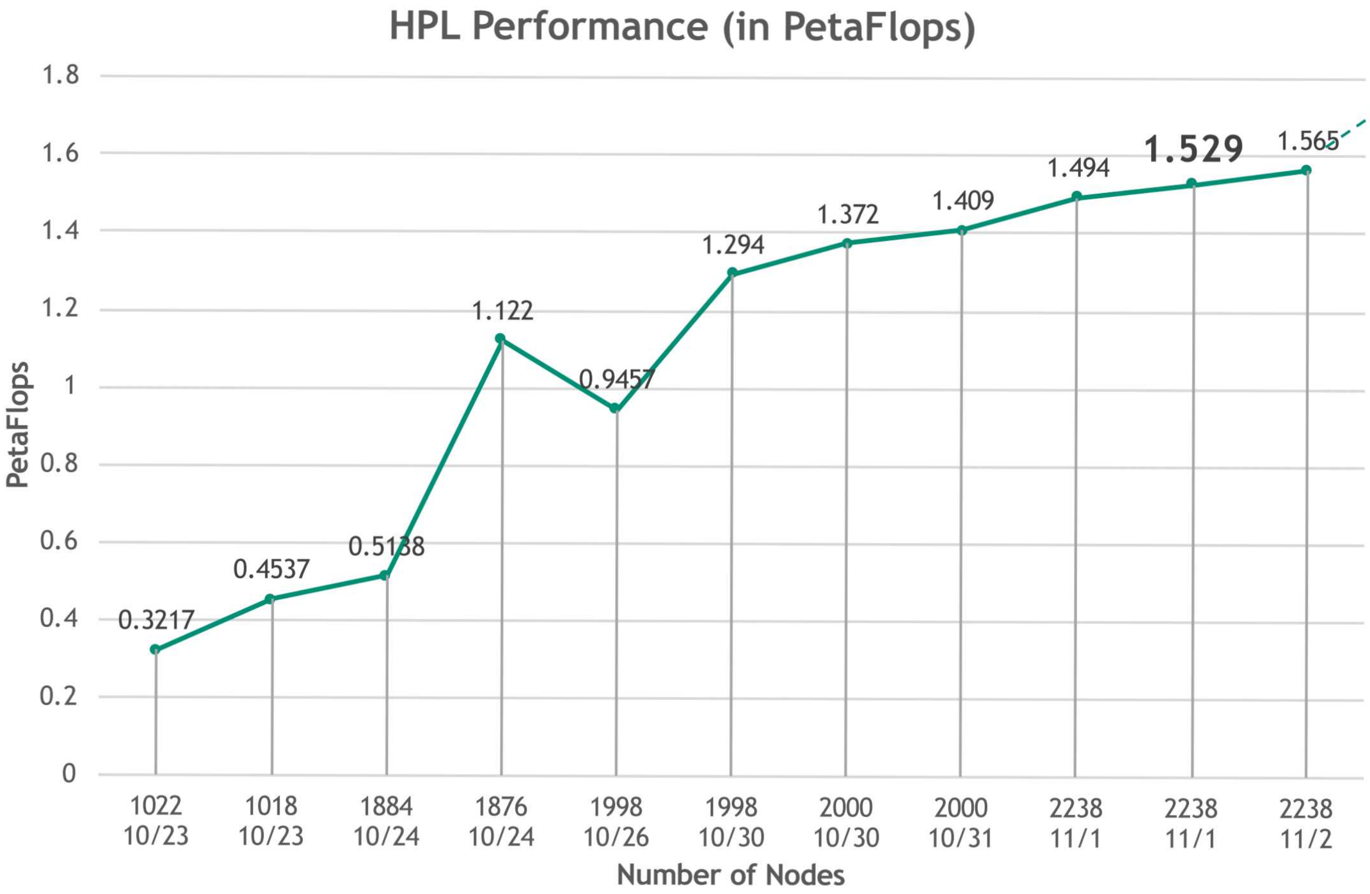
U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories

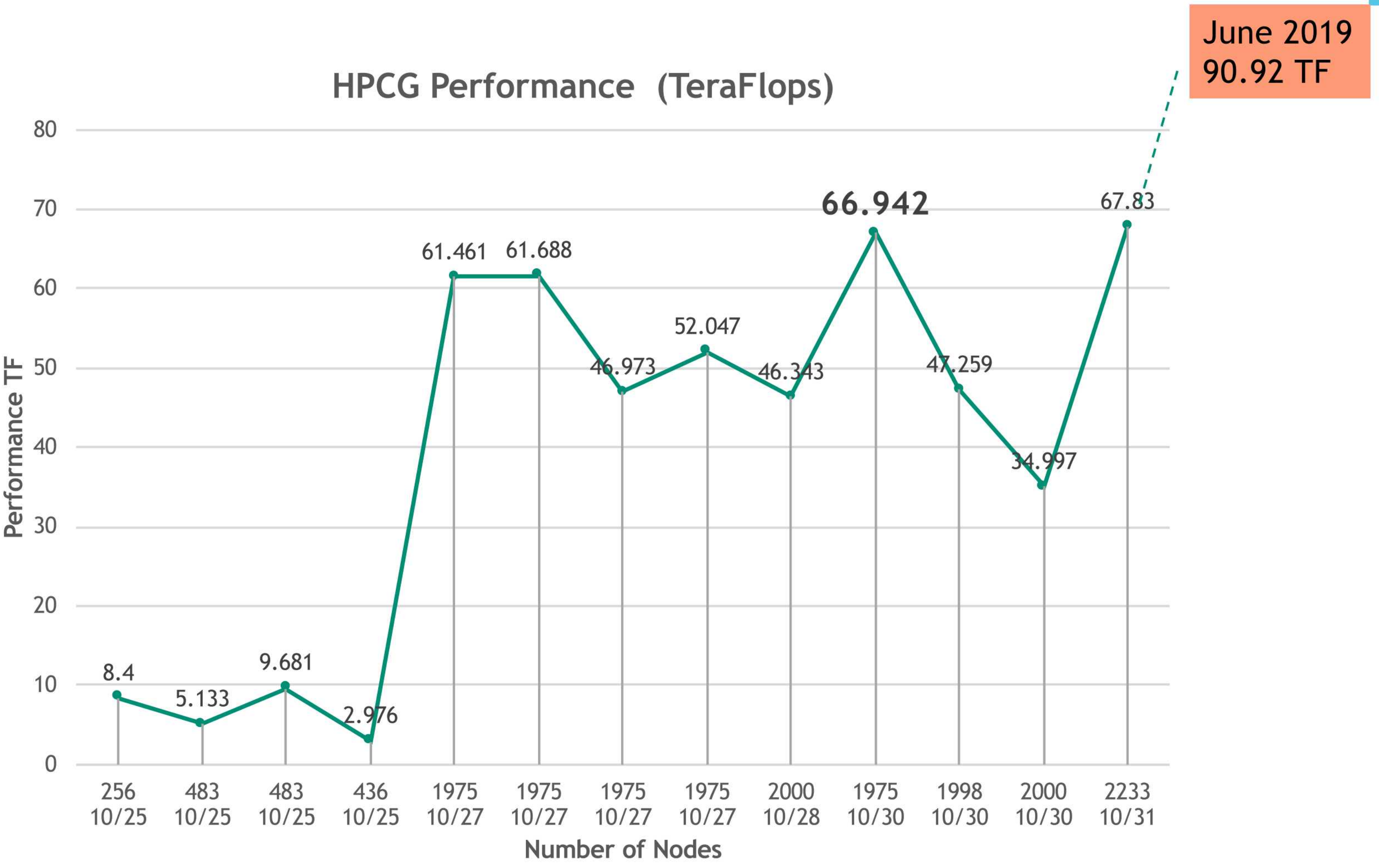
Moving Forward with Astra

HPL Benchmark

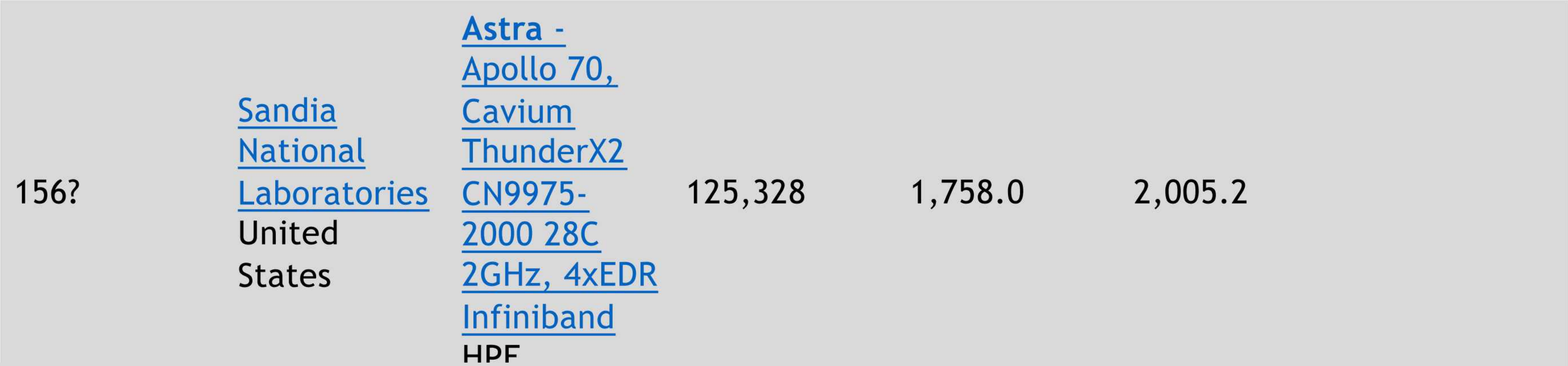


June 2019
1.758 PF

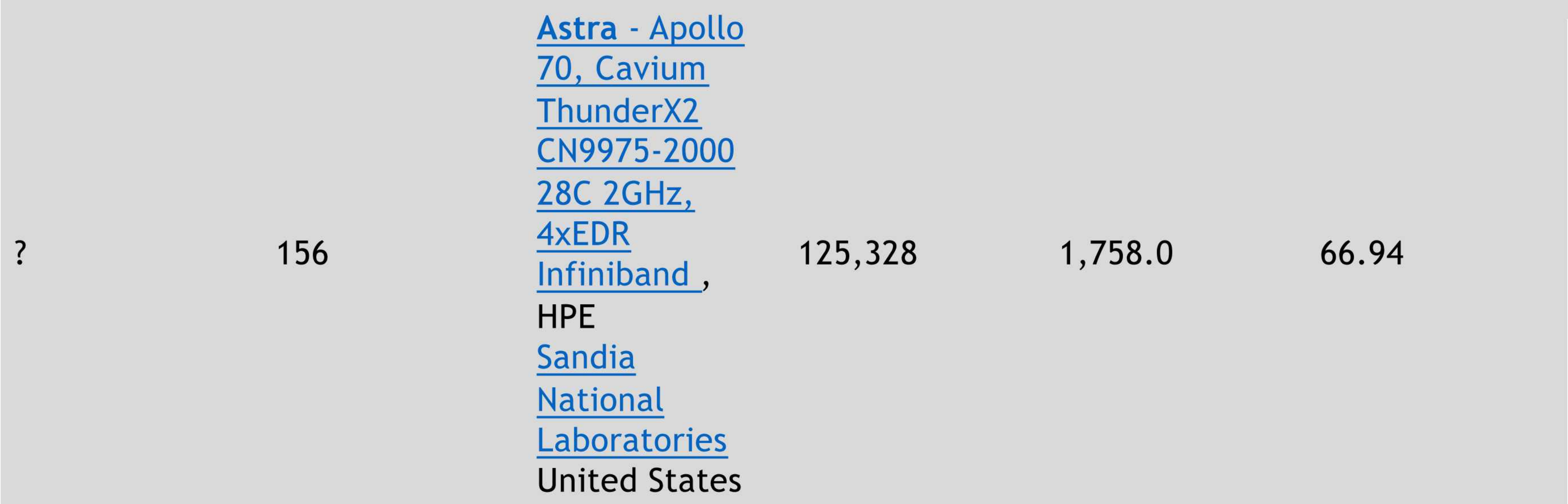
HPCG Benchmark



Latest Top500

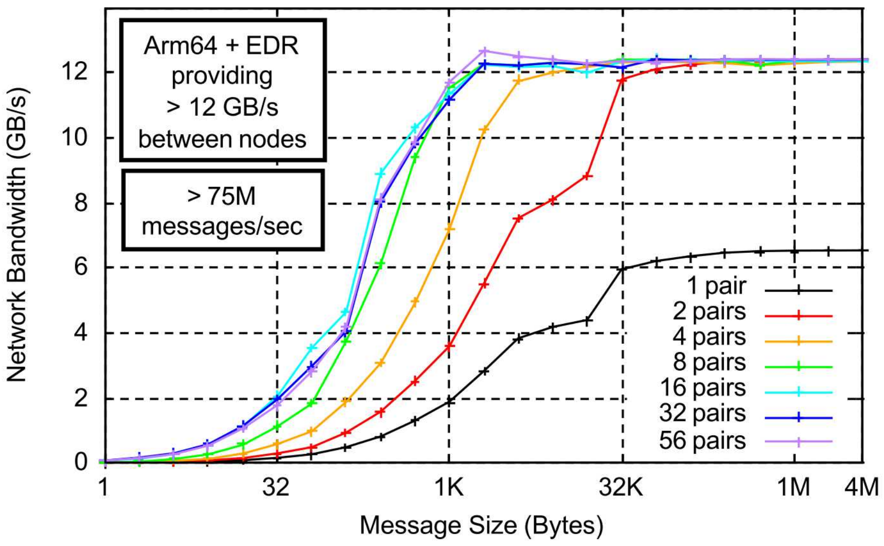


156?	Sandia National Laboratories United States	**Astra** - Apollo 70, Cavium ThunderX2 CN9975-2000 28C 2GHz, 4xEDR Infiniband HPE	125,328	1,758.0	2,005.2



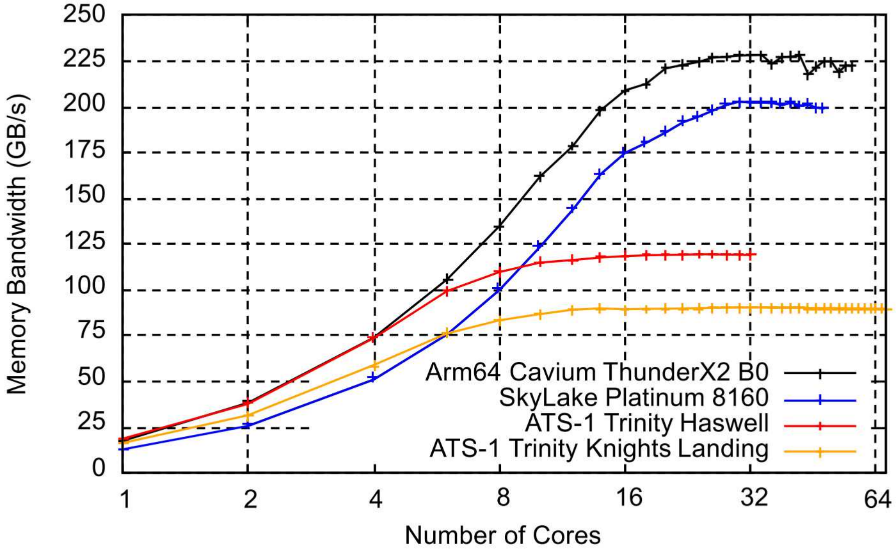
?	156	**Astra** - Apollo 70, Cavium ThunderX2 CN9975-2000 28C 2GHz, 4xEDR Infiniband , HPE Sandia National Laboratories United States	125,328	1,758.0	66.94

Astra Early Results



Arm64 + EDR providing > 12 GB/s between nodes

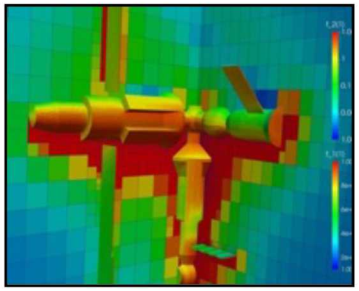
> 75M messages/sec

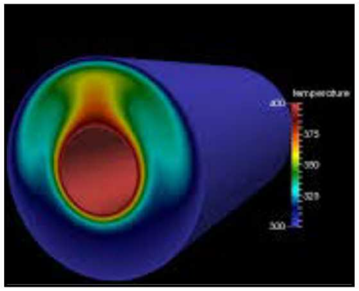


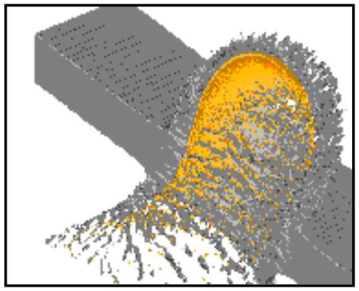
Early Results from Astra

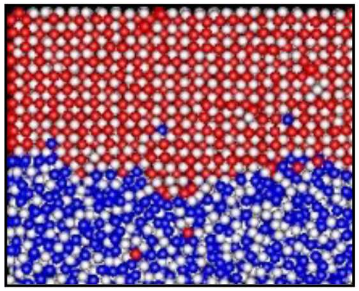
- System online for two weeks prior to data center completion
  - Top500 runs completed just 2 weeks later
- First Petascale ARM platform, designed for production workloads
  - HPL: 1.5 Pflops Rmax, 2 Pflops Rpeak on Top500
  - HPCG: 67 Tflops, 36th on Top500
- Already running application ports and many of our key frameworks

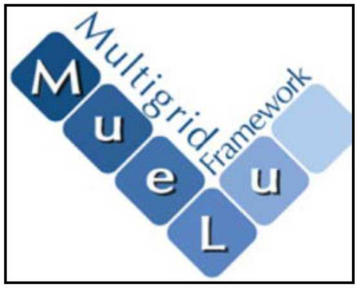
Baseline: Trinity ASC Platform (Current Production), dual-socket Haswell





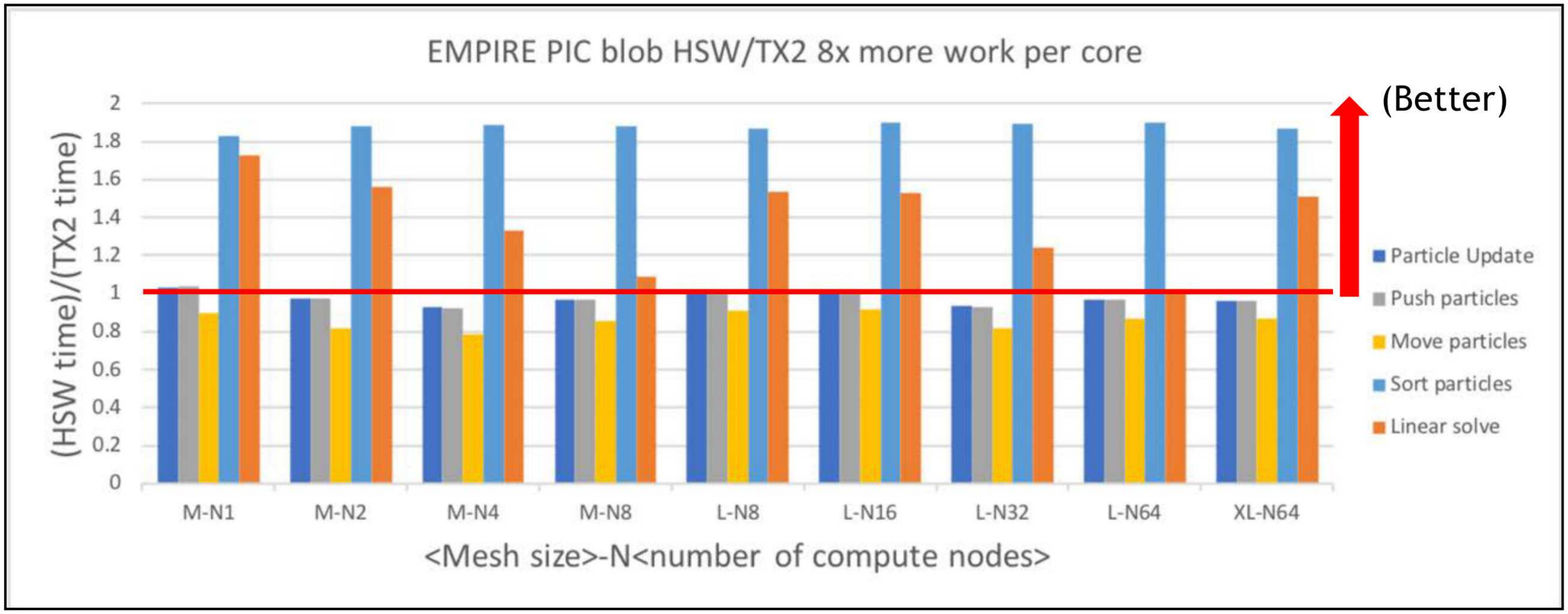






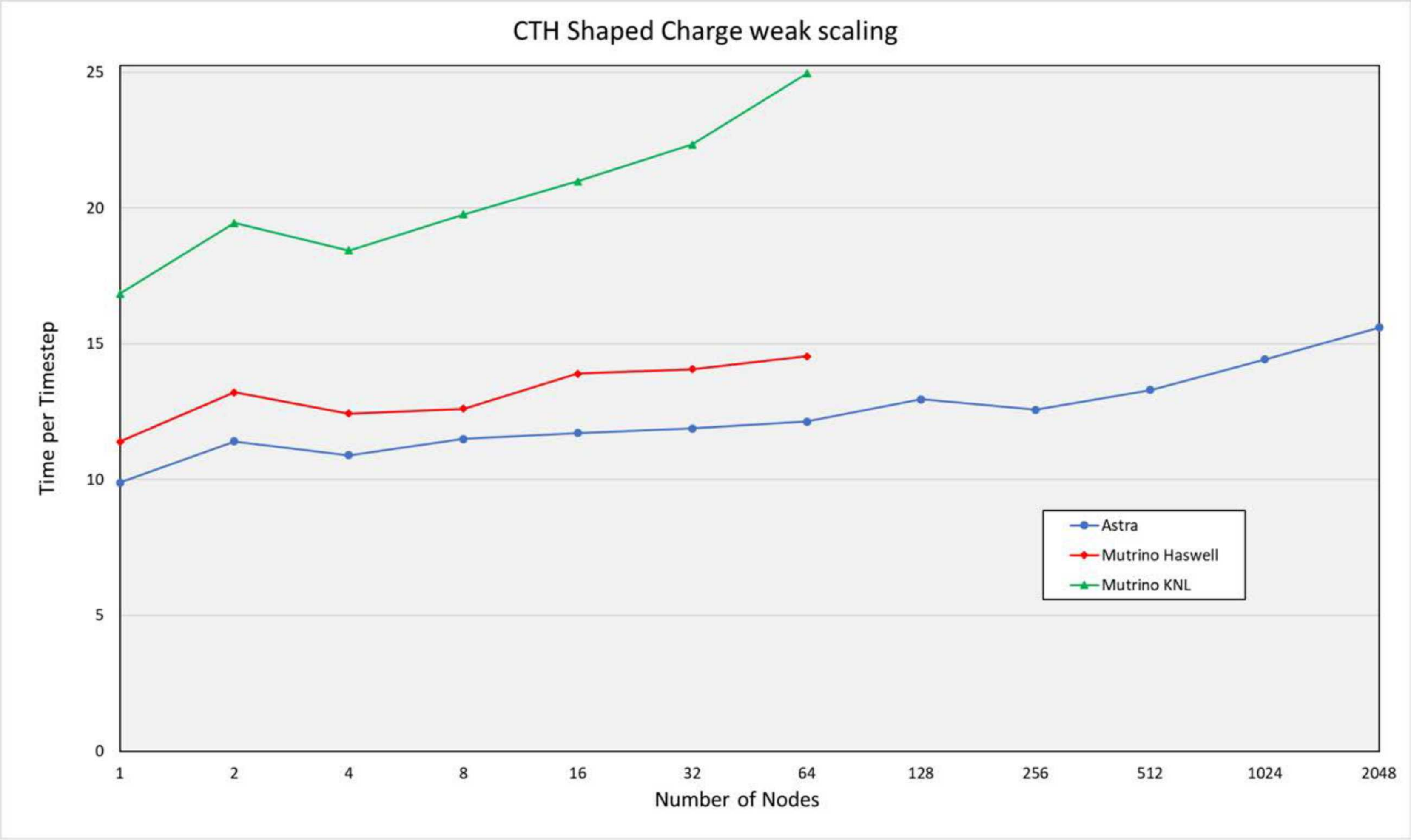
Monte Carlo	CFD Models	Hydrodynamics	Molecular Dynamics	Linear Solvers
1.62x	1.51x	1.33x	1.42x	2.03x

EM Code (EMPIRE) on Astra



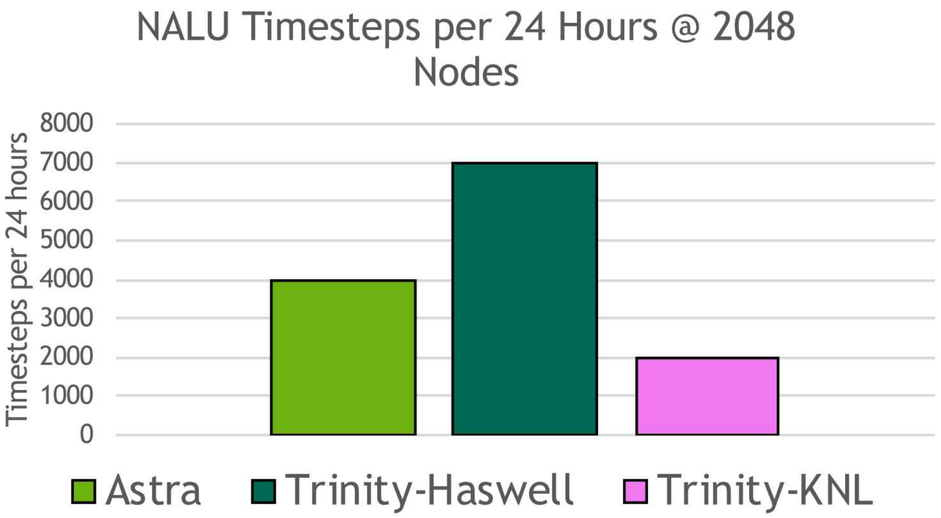
- TX2 node has ~2x memory bandwidth and 1.75x cores (56 vs. 32) of Trinity HSW node
- Strong scaling for medium mesh (1-8 nodes), strong scaling for large mesh (8-64 nodes)
- Sort and solve are more strongly bandwidth limited than particle push/move

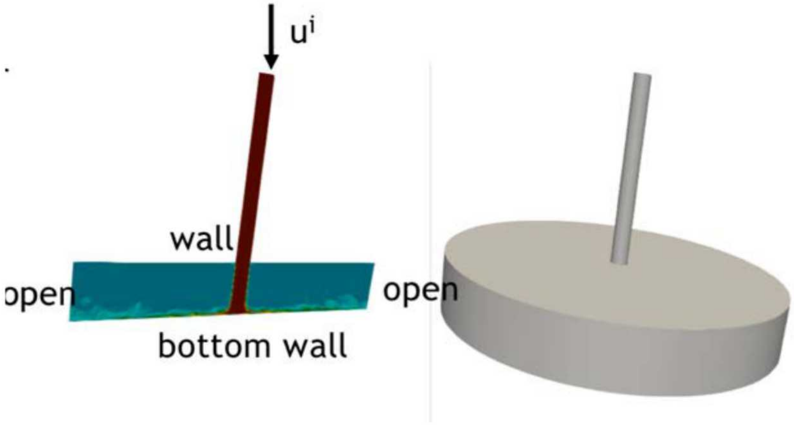
Hydrodynamics Code (CTH) on Astra



NALU CFD Simulation

- NALU – Large Scale CFD Simulation
  - Proxy for large-scale engineering code suite
  - Same mesh handling and I/O
  - Trilinos solvers using multi-grid libraries
- Results show strong solve kernel performance but slower assembly
  - Some routines do not scale well with increasing MPI rank counts (problem on Astra and KNL)





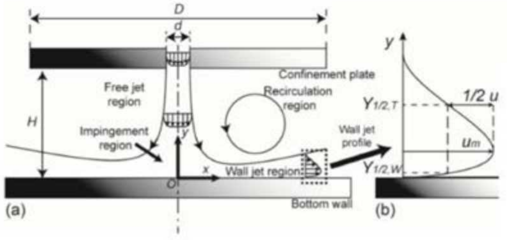
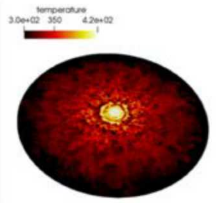


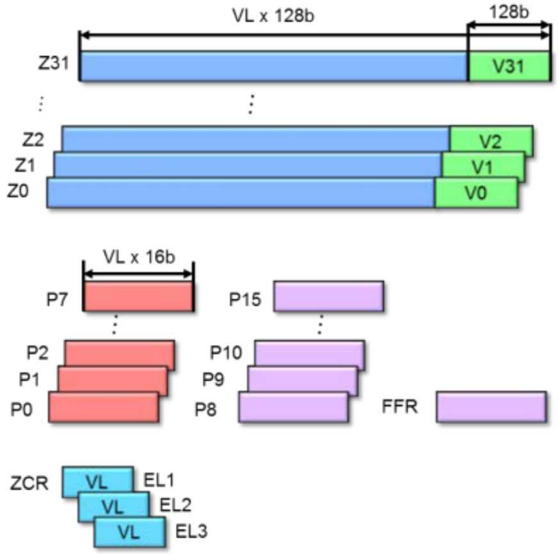
Figure 1. Schematic drawings of (a) the axisymmetric flow field formed by the impinging jet and, (b) the wall jet structure and nomenclature.



Moving Forward

SVE Enablement (Arm/Marvel)

- SVE work is underway
  - Using ArmIE (fast emulation) and RIKEN GEM5 Simulator
  - GCC and Arm toolchains
- Collaboration with RIKEN
  - Visited Sandia (participants from NNSA Labs, RIKEN)
  - Discussion of performance and simulation techniques
  - Deep-dive on SVE (GEM5)
- Short term plan
  - Use of SVE intrinsics for Kokkos-Kernels SIMD C++/data parallel types
  - Underpins number of key performance routines for Trilinos libraries
    - Seen large (6X) speedups for AVX512 on KNL and Skylake
    - Expect to see similar gains for SVE vector units
  - Critical performance enablement for Sandia production codes



Collaborations

- DOE (OoS ASCR/NNSA ASC)
  - ECP
  - Innovative Architectures
  - Algorithms
- Japan (MEXT/RIKEN,etc.)
  - SVE
  - Arm Architectural Modeling (GEM5/SST)
  - Algorithms
- UK (Univ. of Bristol)
  - Proxies/Benchmarks
  - Architectural Modeling
- France (CEA)
  - Algorithms
  - Proxies/Benchmarks
  - SysSW
- More…

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories

Extra Slides

It Takes an Incredible Team…

- DOE Headquarters:
  - Thuc Hoang
  - Mark Anderson
- Sandia Procurement
- Sandia Facilities
- Colleagues at LLNL and LANL
  - Trent D'Hooge
  - Mike Lang
  - Rob Neely
  - Dave Richards
- Incredible team at Sandia

- HPE:
  - Mike V. and Nic Dube
  - Andy Warner
  - John D'Arcy
  - Steve Cruso
  - Lori Gilbertson
  - Cheng Liao
  - John Baron
  - Kevin Jamieson
  - Tim Wilcox
  - Charles Hanna
  - Mike Craig
  - And loads more …

- Cavium/Marvel:
  - Giri Chukkapalli
  - Todd Cunningham
  - Larry Wikelius
  - Kiet Tran
  - Joel James
  - And loads more…
- ARM:
  - ARM Research Team!
  - ARM Compiler Team!
  - ARM Math Libraries!
  - And loads more…

ATSE Collaboration with HPE's HPC Software Stack

Hewlett Packard Enterprise

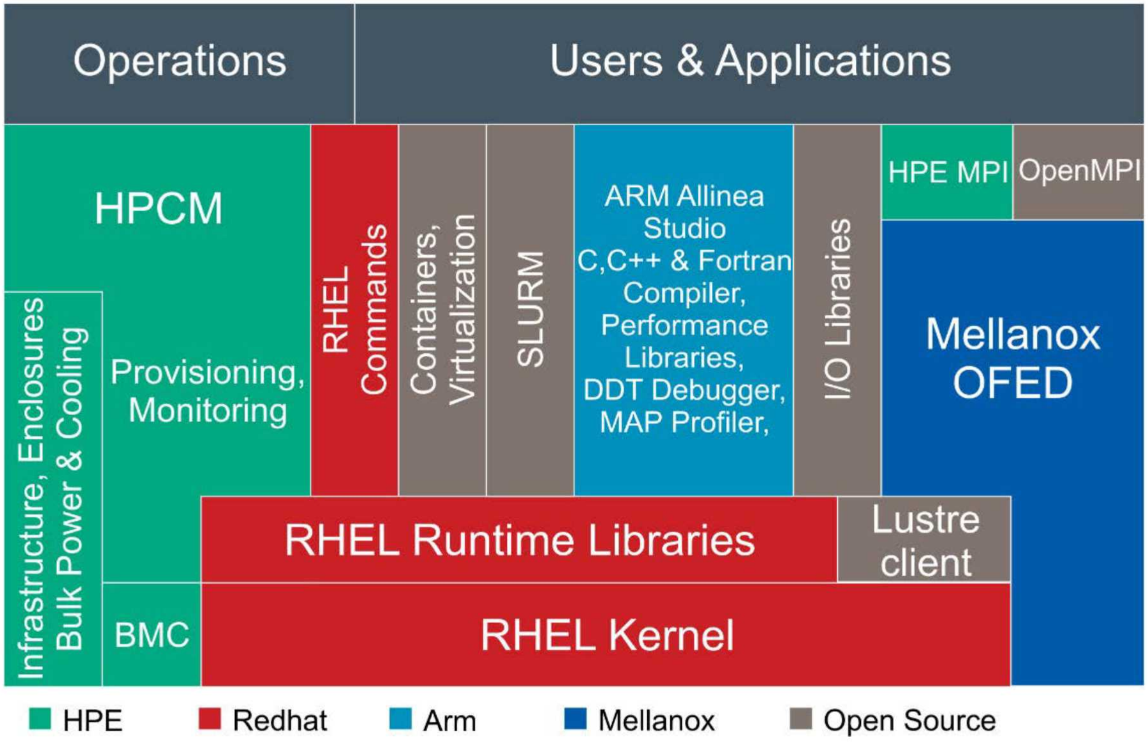
ATSE stack

HPE's HPC Software Stack

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

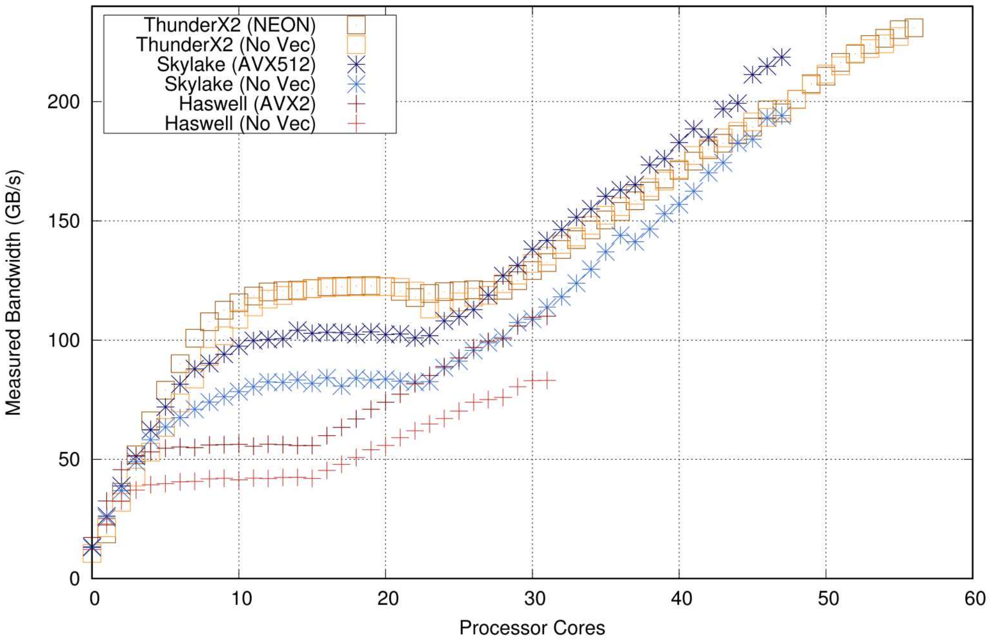
Lustre client

RHEL Kernel

🟩 HPE 🟥 Redhat 🟦 Arm 🟦 Mellanox ◼ Open Source

STREAM Triad Bandwidth

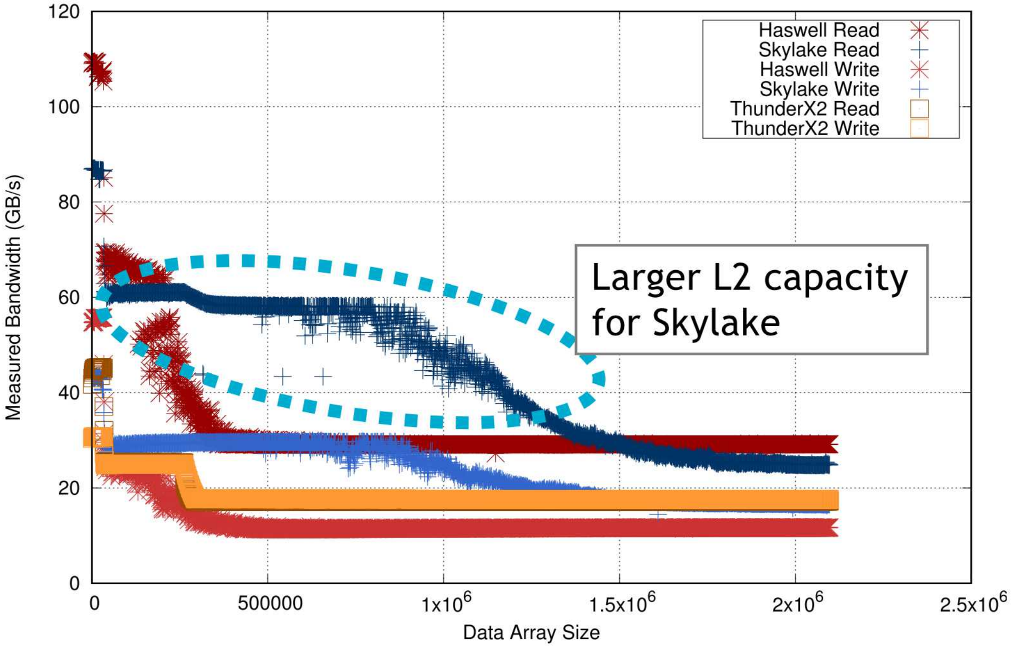
- ThunderX2 provides highest bandwidth of all processors
- Vectorization makes no discernable difference to performance at large core counts
  - Around 10% higher with NEON at smaller core counts (5 – 14)



Higher is better

Cache Performance

- Haswell has highest per-core bandwidth (read and write) at L1, slower at L2.
- Skylake redesigned cache sizes (larger L2, smaller L3) shows up in graph
  - Higher performance for certain work-set sizes (typical for unstructured codes)
- TX2 more uniform bandwidth at larger scale (see less asymmetry between read/write)



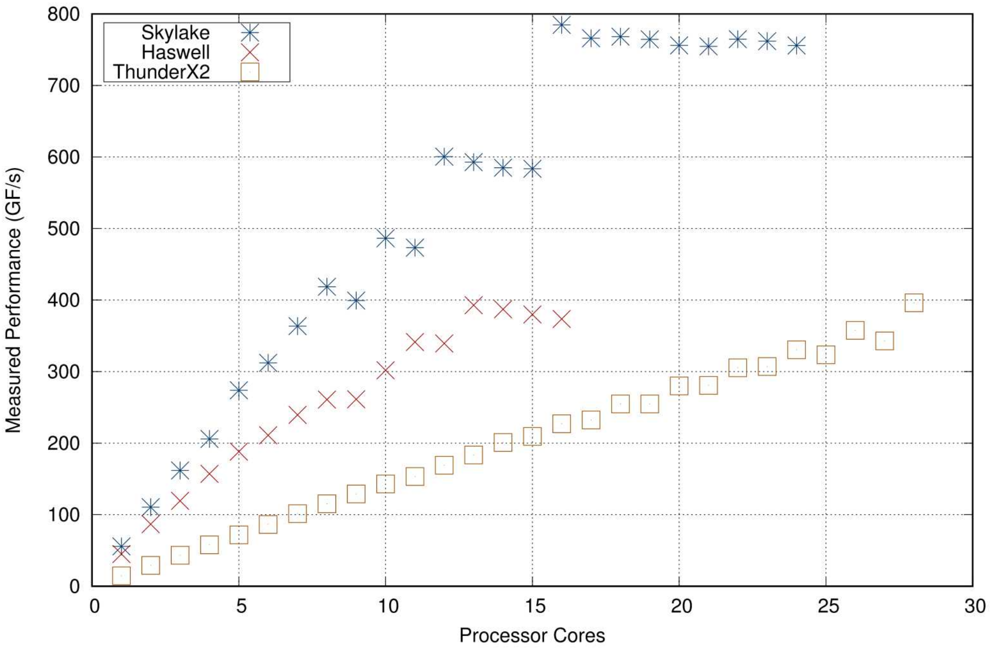
Larger L2 capacity for Skylake

Higher is better

DGEMM Compute Performance

- ThunderX2 has similar performance at scale to Haswell
  - Roughly twice as many cores (TX2)
  - Half the vector width (TX2 vs. HSW)

- See strata in Intel MKL results, usually a result of matrix-size kernel optimization
  - ARM PL provides smoother performance results (essentially linear growth)

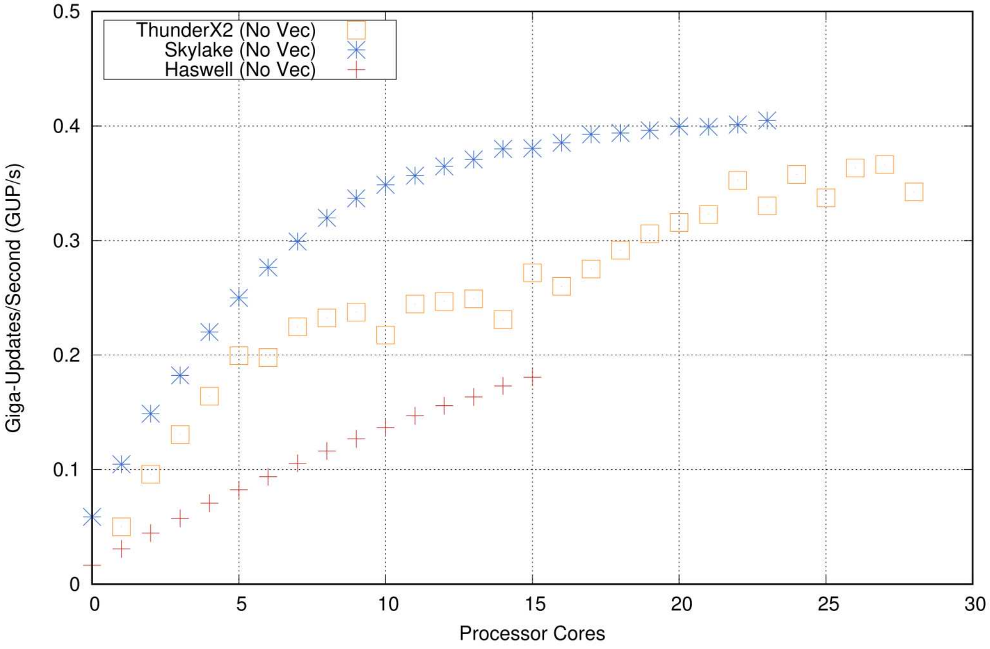


Higher is better

GUPS Random Access

- Running all processors in SMT-1 mode, SMT(>1) is usually better performance
  - Expect SMT2/4 on TX2 to give better numbers

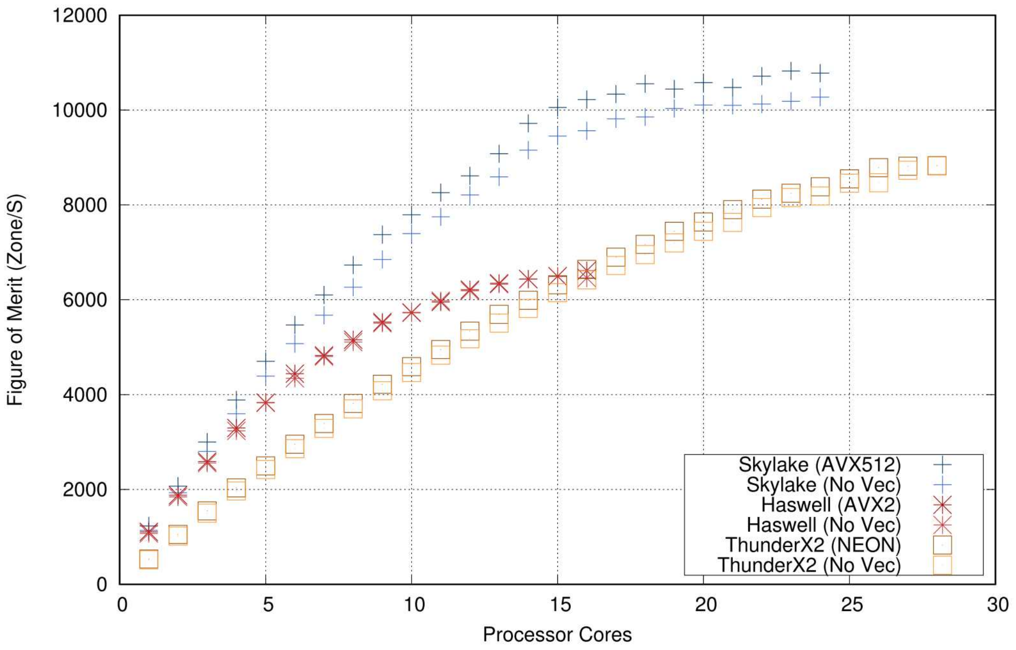
- Usually more cores gives higher performance (more load/store units driving requests).
  - Typical for TLB performance to be a limiter
  - Need to consider larger pages for future runs



Higher is better

LULESH Hydrodynamics Mini-App

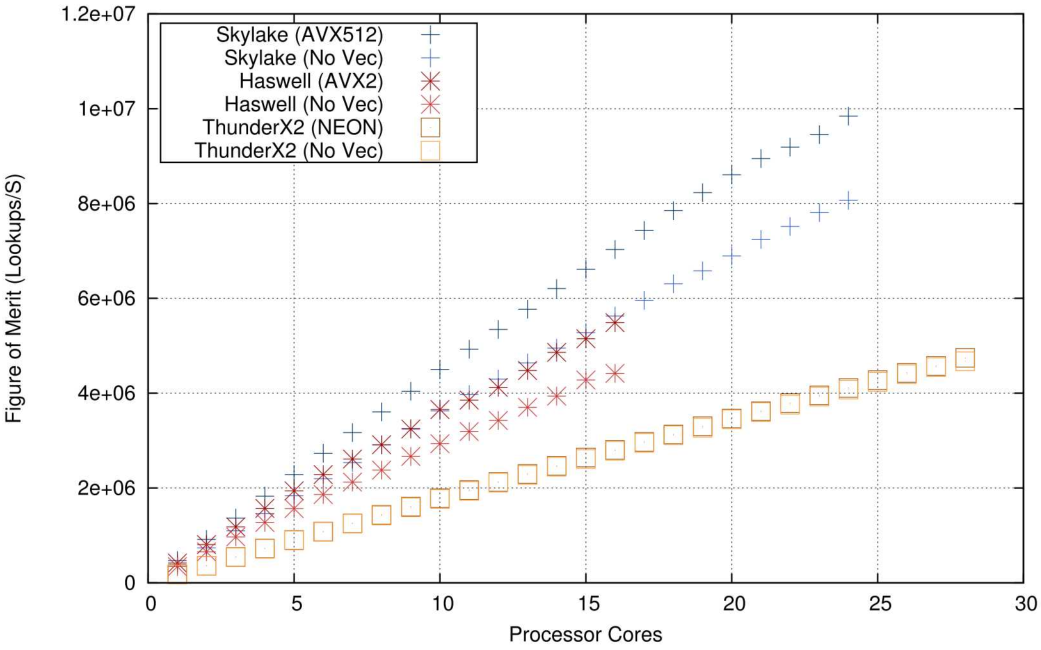
- Typically fairly intensive L2 accesses for unstructured mesh (although LULESH is regular structure in unstructured format)
- Expect slightly higher performance with SMT(>1) modes for all processesors



Higher is better

XSBench Cross-Section Lookup Mini-App

- Two level random-like access into memory, look-up in first table and then use indirection to reach second lookup
  - Means random access but is more like search so vectors can help
- See gain on Haswell and Skylake which both have vector-gather support
  - No support for gather in NEON
  - XSBench is mostly read-only (gather)



Higher is better

Containers on Astra

- Leverage containers and virtual machines on ARM
- Singularity Containers
  - ATSE container image
  - Working with Sylabs on full container solution
  - Support emerging ML/AI frameworks
  - Leverage remote builder, library, and secure signing services
  - Evaluate container scalability
- Linking with DOE Exascale “Supercontainers” project
- KVM Virtual Machine support
  - ARMv8.1 includes virtualization extensions, SR-IOV
  - Optimize and tune with libvirt for TX2