SAND2019-11171C

RM SUPERCOMPUTER





# xperiences Scaling a Production Arm Supercomputer to Petaflops and Beyond

Kevin Pedretti for the Astra Team

POCs: Jim Laros (Platform), Kevin Pedretti (Software Stack), Si Hammond (Apps)
jhlaros@sandia.gov, ktpedre@sandia.gov, sdhammo@sandia.gov

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and 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 Unclassified Unlimited Release .

# It Takes an Incredible Team…

- DOE Headquarters:
  - Thuc Hoang
  - Mark Anderson
- Sandia Procurement
- Sandia Facilities
- Incredible Sandia Team
- Colleagues at LLNL and LANL
  - Mike Lang
  - Rob Neely
  - Mike Collette
  - Alan Dayton
  - Trent D'Hooge
  - Todd Gamblin
  - Robin Goldstone
  - Anna Pietarila Graham
  - Sam Gutierrez
  - Steve Langer
  - Matt Leininger
  - Matt Legendre
  - Pat McCormick
  - David Nystrom
  - Howard Pritchard
  - Dave Rich
  - And loads more …

- HPE:
  - Mike V. and Nic Dube
  - Andy Warner
  - Erik Jacobson
  - John D'Arcy
  - Steve Cruso
  - Lori Gilbertson
  - Meredydd Evans
  - Cheng Liao
  - John Baron
  - Kevin Jameson
  - Tim Wilcox
  - Charles Hanna
  - Michael Craig
  - Patrick Raymond
  - And loads more …

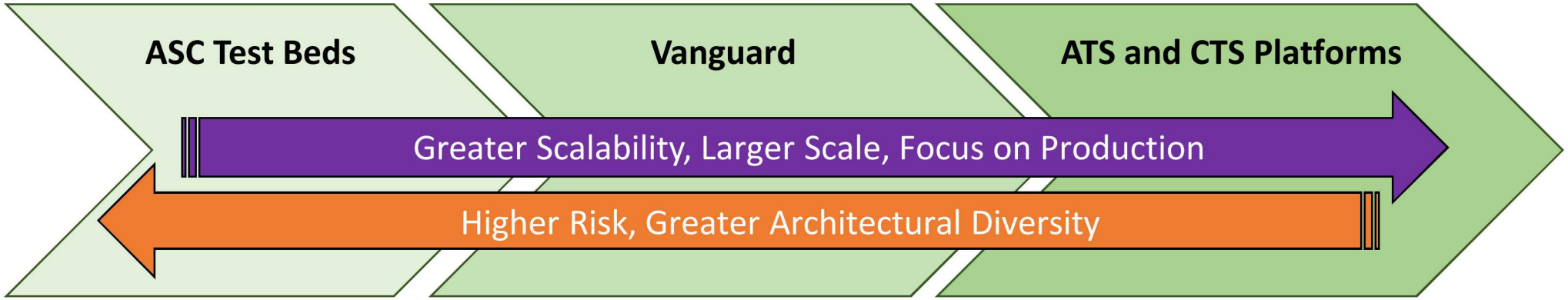
- Cavium/Marvel:
  - Giri Chukkapalli (now NVIDIA)
  - Todd Cunningham
  - Larry Wikelius
  - Raj Sharma Govindaiah
  - Kiet Tran
  - Joel James
  - And loads more…
- ARM:
  - ARM Research Team!
  - ARM Compiler Team!
  - ARM Math Libraries!
  - And loads more…

# It Takes an Incredible Team…

# Outline

- Astra Overview
- ATSE Software Stack
- Recent Application Results
- Conclusion – HPC on Arm, are we there yet?

# Vanguard Program: Advanced Technology Prototype Systems




**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, seek to increase technology and vendor choices
- **DOE/NNSA Tri-lab resource**

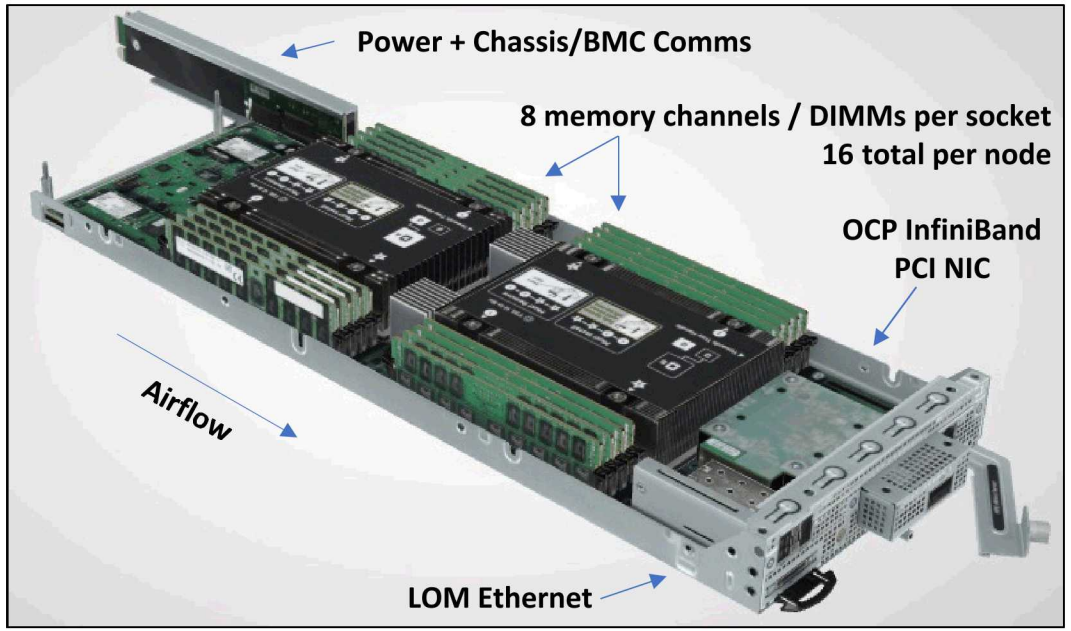
**Production Platforms**

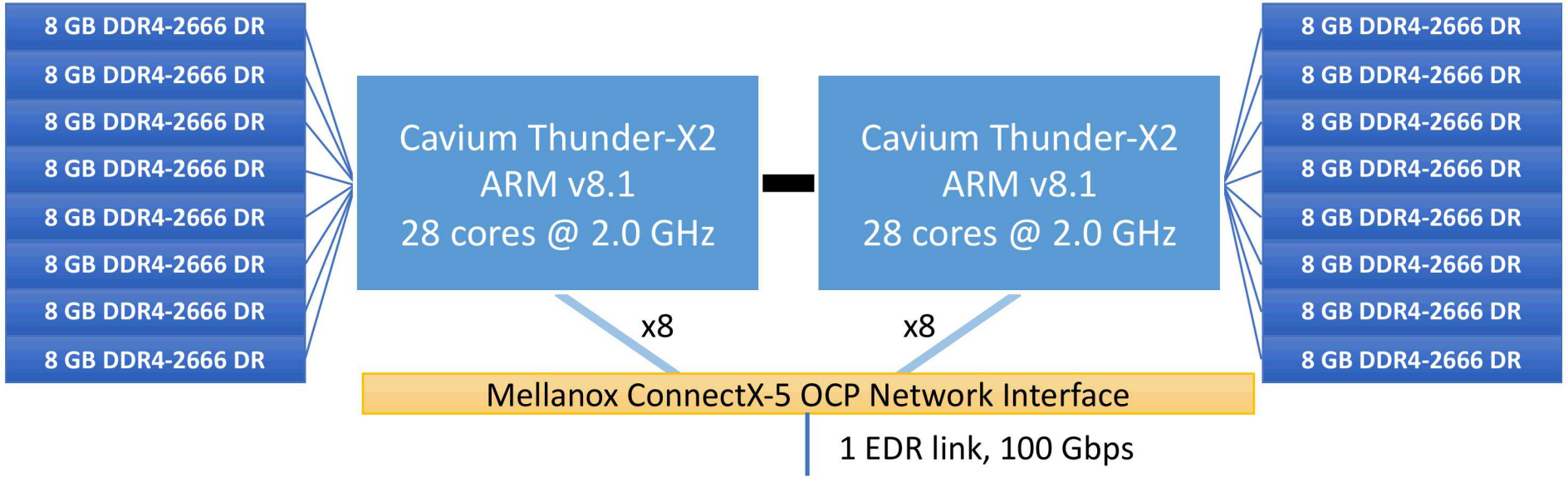
- Leadership-class systems (Petascale, Exascale, ...)
- Advanced technologies, sometimes first-of-kind
- Broad user-base
- Production use

Astra is the first Vanguard Platform

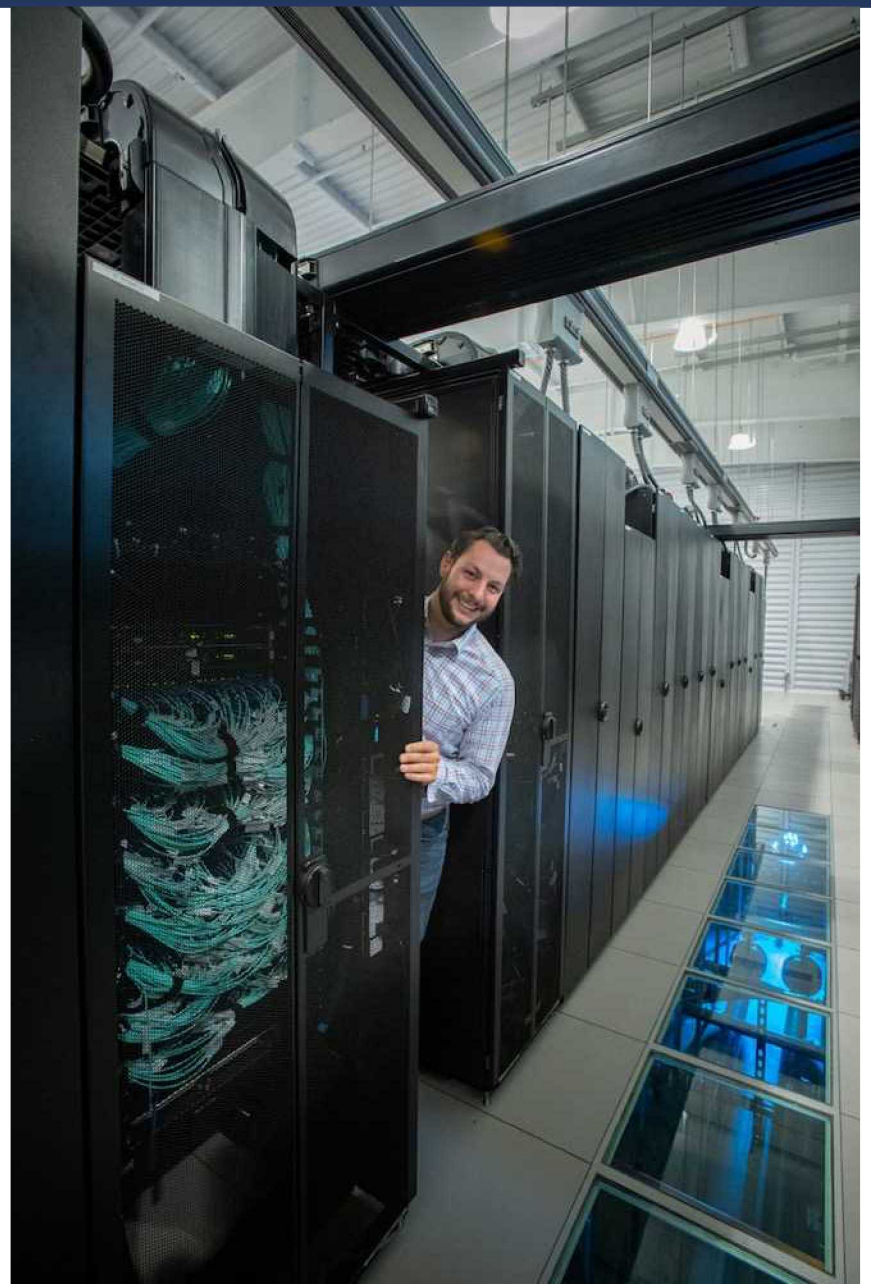
# Astra Node Architecture

- **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: 990 TB (usable)
  - Bandwidth 244 GB/s



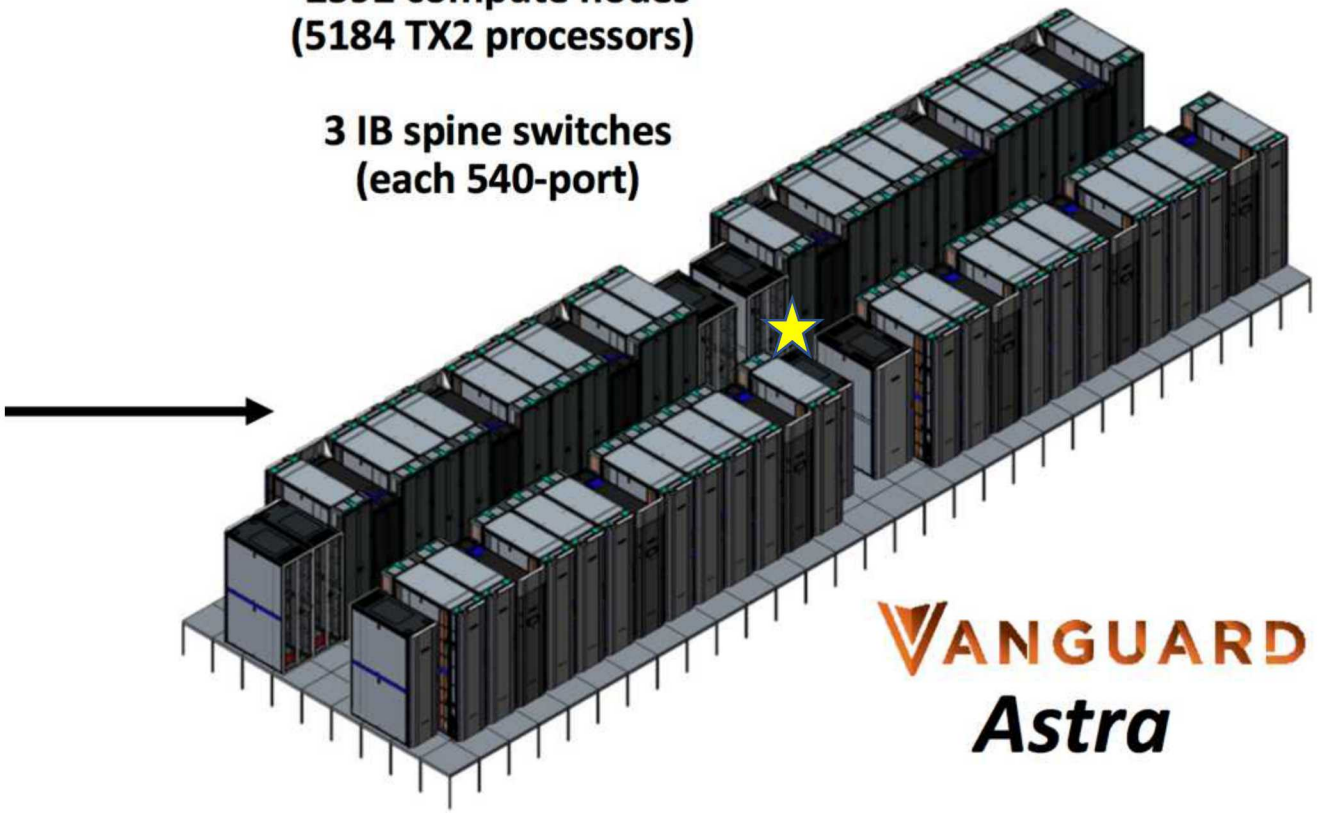
# Astra System Architecture



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

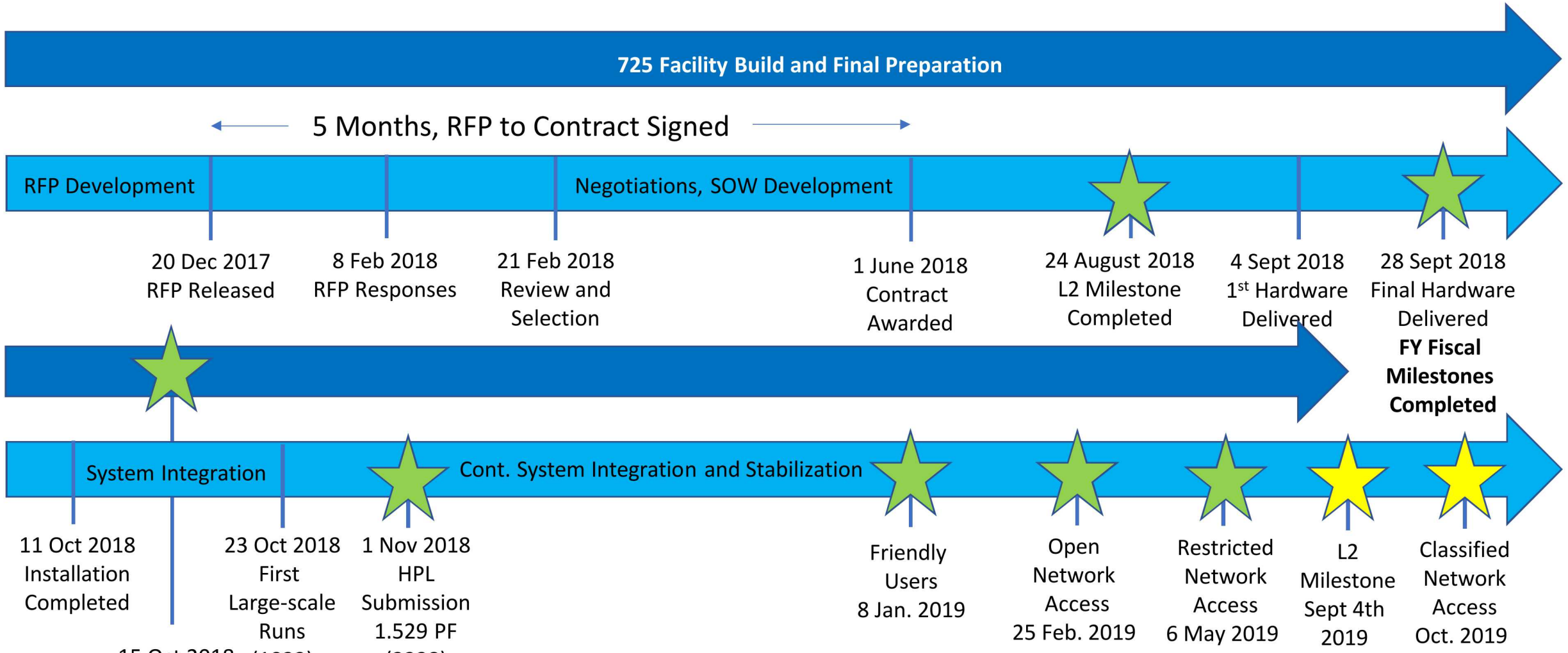
**2592 compute nodes**
**(5184 TX2 processors)**

**3 IB spine switches**
**(each 540-port)**



VANGUARD
Astra

# Vanguard-Astra: Timeline




June 2019 Top500:

HPL: #156 @ 1.758 PFlops/s on 2428 nodes (up from #204)

HPCG: #29 @ 90.92 TFlop/s on 2385 nodes (up from #38)

# Real-Time System Monitoring Has Been Key

- Tools: {BMC,PDU,Syslog,TX2MON} + TimescaleDB + Grafana

# Outline

- Astra Overview
- ATSE Software Stack
- Recent Application Results
- Conclusion – HPC on Arm, are we there yet?

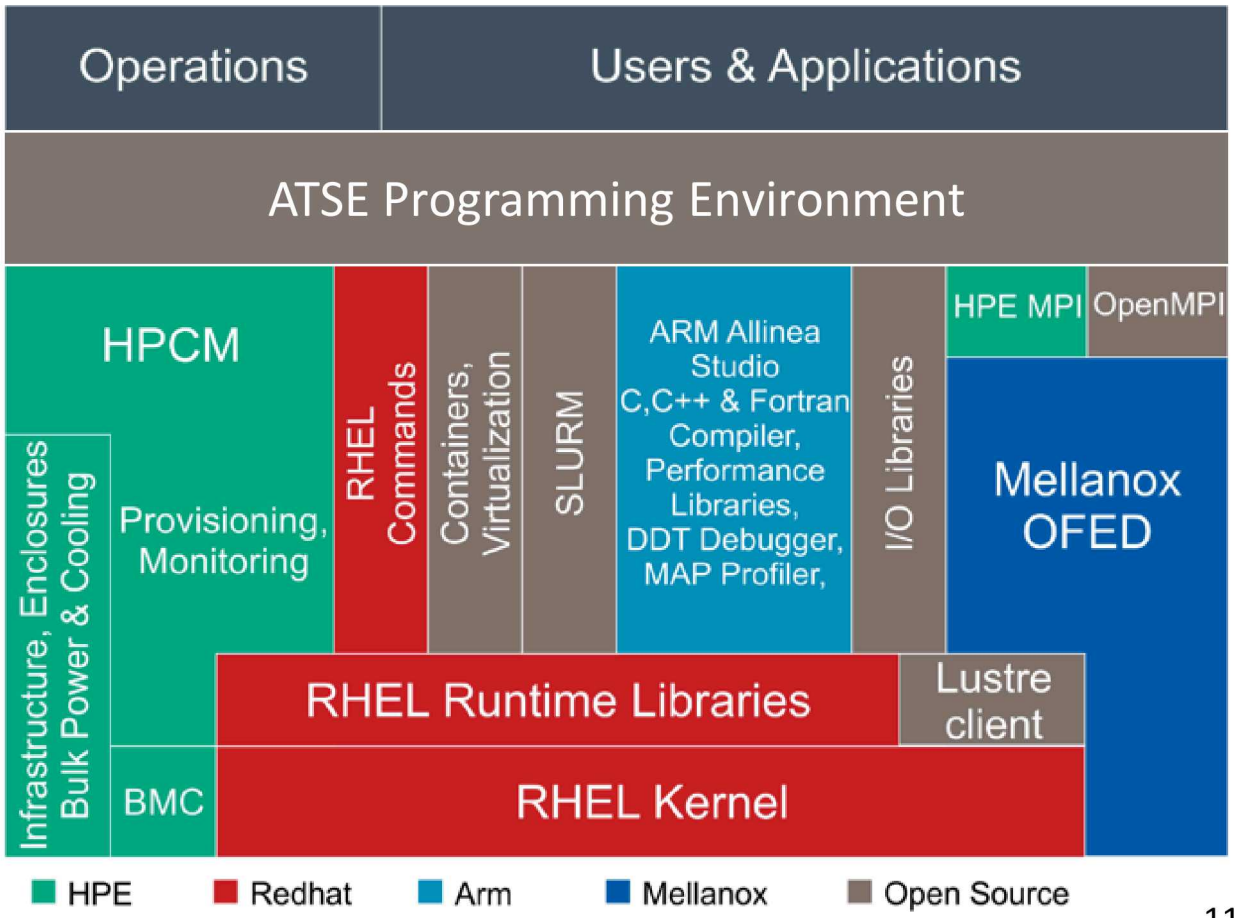
# Astra Collaboration with HPE's HPC Software Stack

Hewlett Packard Enterprise

ATSE stack

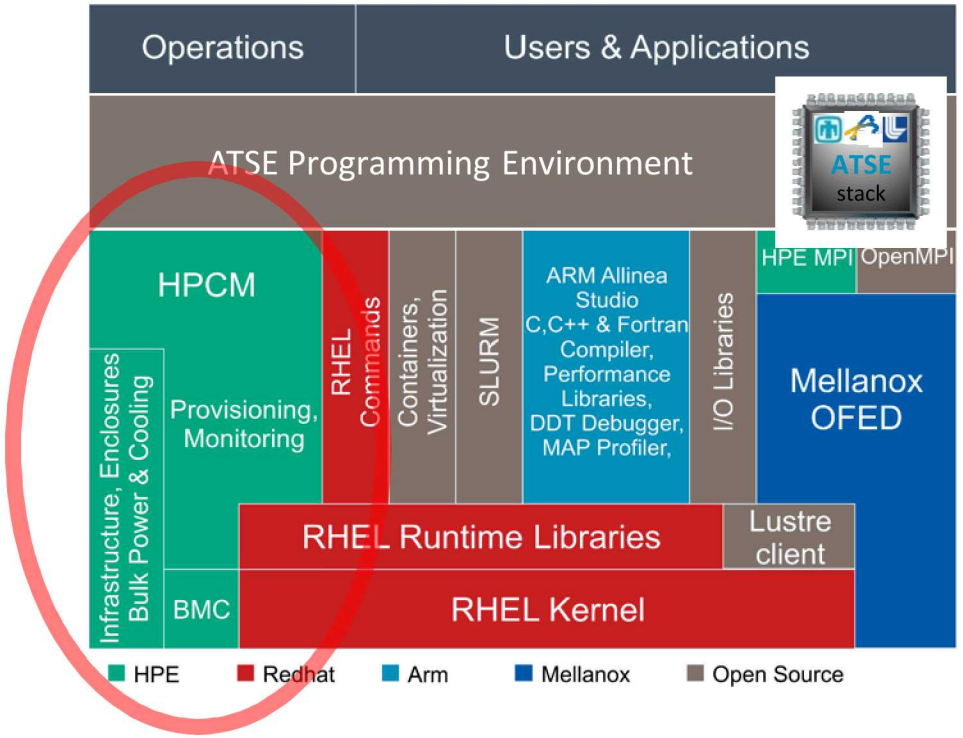
HPE's HPC Software Stack

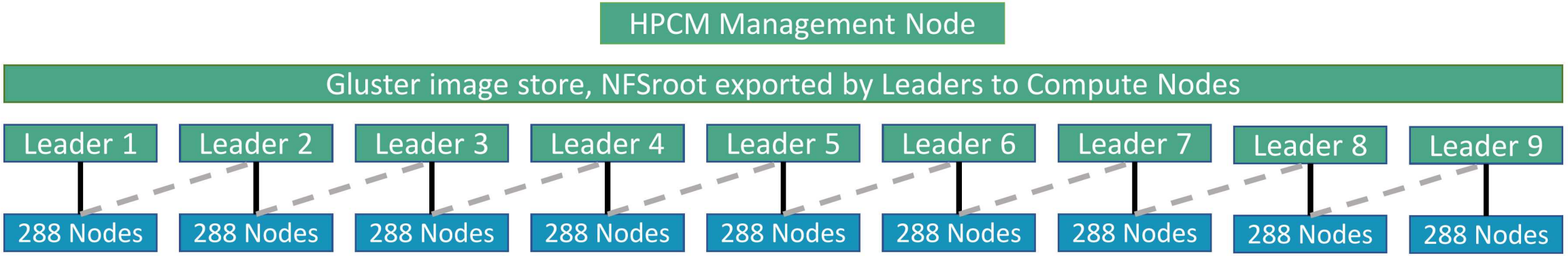
- HPE:
  - HPE Cluster Manager
  - HPE MPI (+ XPMEM)
- Arm:
  - Arm HPC Compilers
  - Arm Math Libraries
  - Allinea Tools
- Mellanox-OFED & HPC-X
- RedHat 7.x for aarch64



# HPCM Provides Scalable System Management for Astra

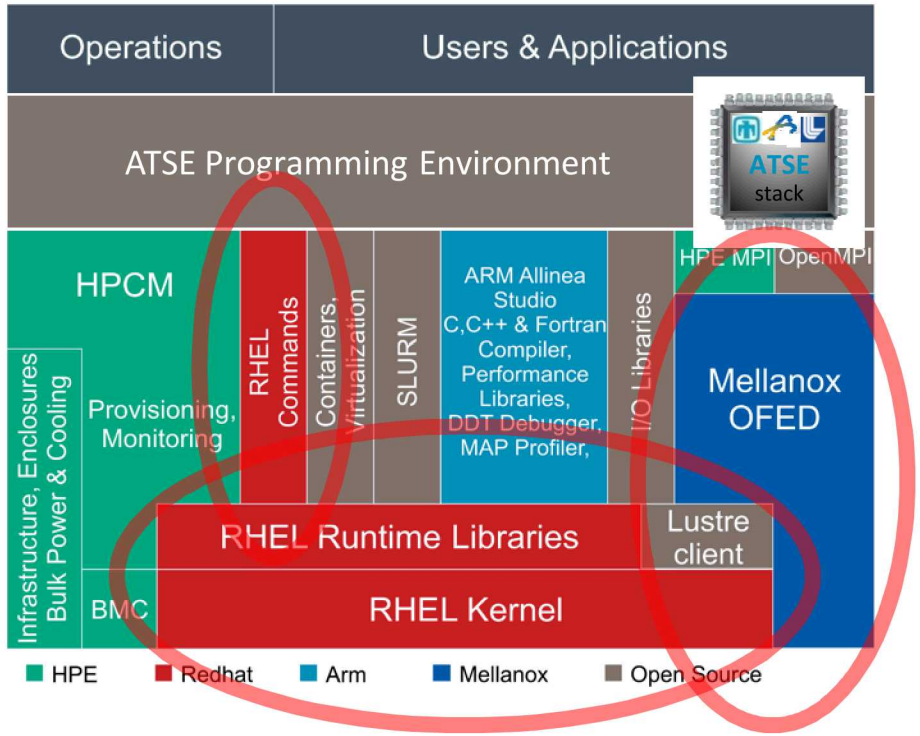
- HPCM: HPE Performance Cluster Manager
  - Merger of HPE CMU with SGI Icebox stack
  - New product at time of Astra deployment
- Collaboration resulted in new capabilities
  - Support for hierarchical leader nodes for non Icebox clusters (aka “Flat Clusters”)
    - **Demonstrated boot of 2592 nodes in < 10 min**
  - Resilient leader node failover
  - Scalable BIOS upgrades and configuration
  - Ability to deploy TOSS images (Tri-lab Operating System Stack)



# TOSS Provides Robust Base OS for Astra / ARM

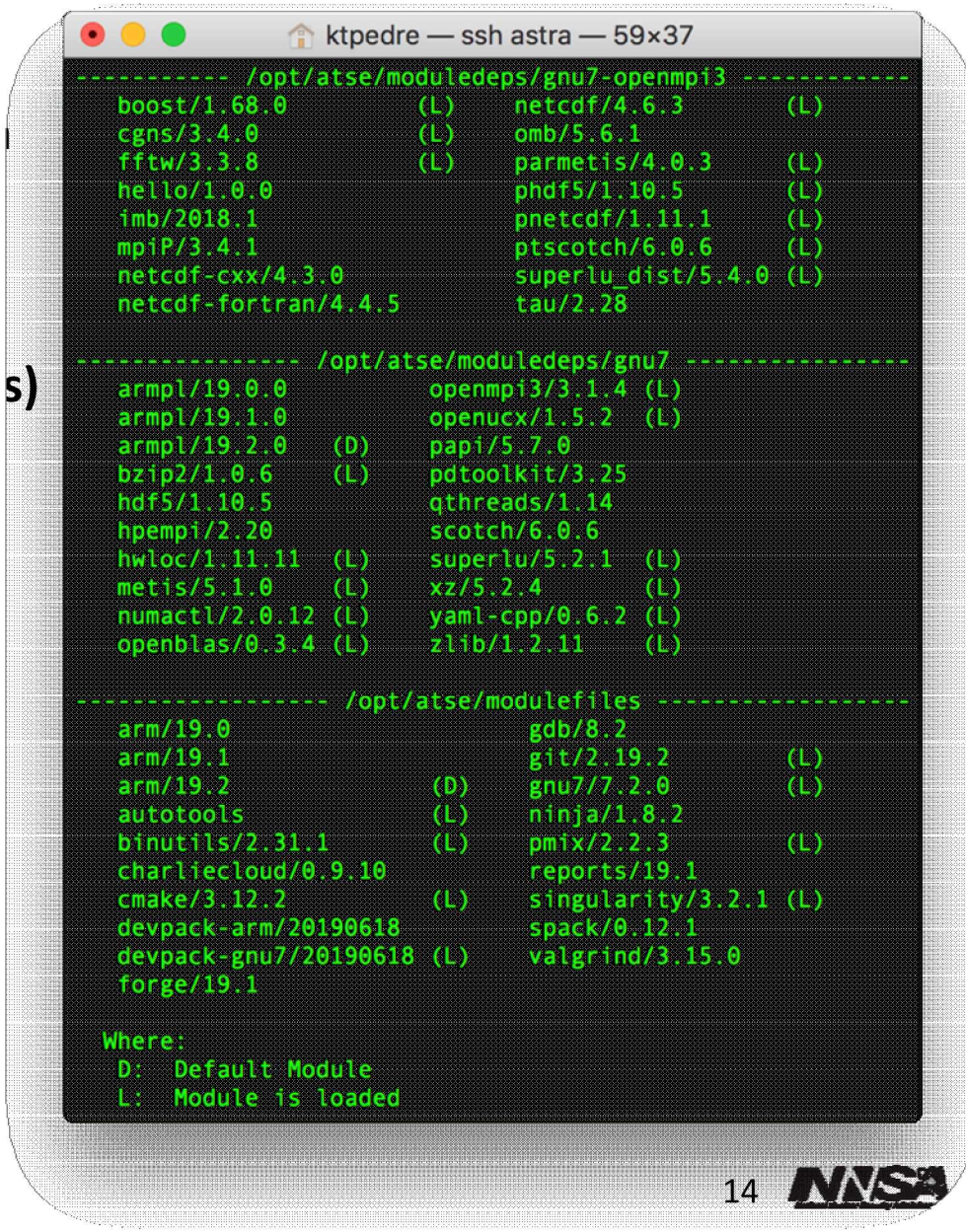
- TOSS: Tri-lab Operating System Stack (Lead: LLNL, LANL, SNL)
  - Targets commodity technology systems (model: vendors provide HW, labs provide SW)
  - Red Hat 7 based; x86_64, ppc64le, and aarch64
  - ~4K packages on all archs, 200+ specific to TOSS
  - Partnership with RedHat with direct support



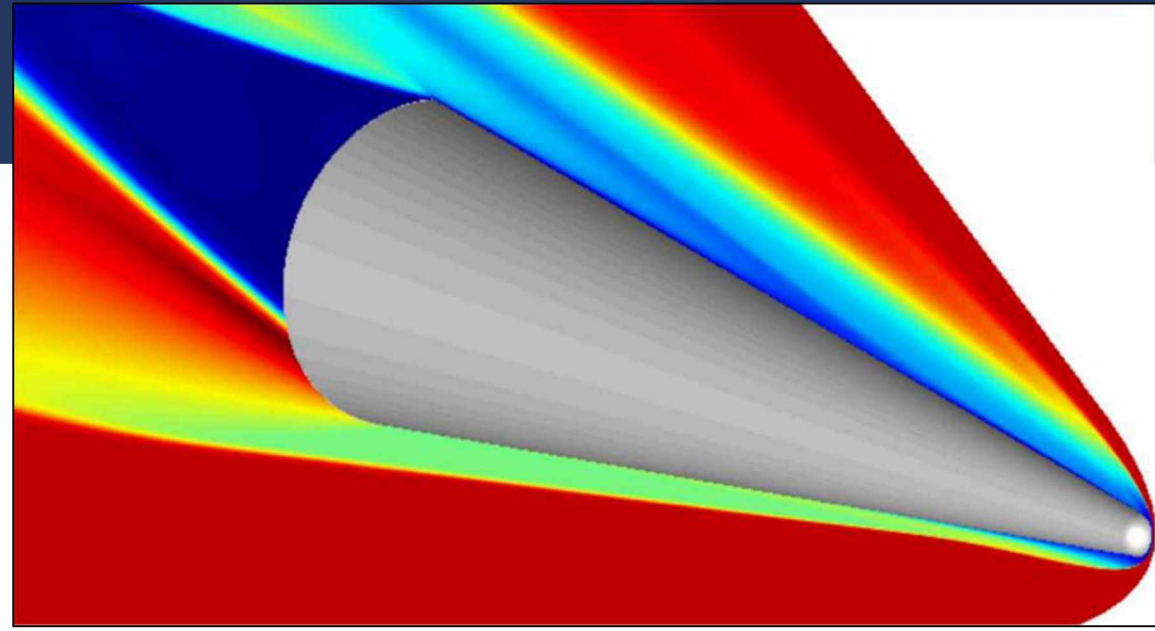

- Astra-related activities
  - Lustre enablement and bringup
  - Added support for Mellanox OFED InfiniBand stack, needed for advanced IB features
  - Debugged Linux Kernel issues on Arm, scale of Astra revealed bugs not previously seen
    - Kworker CPU hang – fix was in Linux upstream, but not in RedHat. Patch added to TOSS Linux kernel.
    - Sys_getdents64 oops – rare hang at job cleanup / exit.
      Actively debugging with RedHat + Marvell + HPE + Mellanox

# ATSE is an Integrated Software Environment for ASC Workloads

- Advanced Tri-lab Software Environment
  - User-facing programming environment co-developed with As
  - Provides a common set of libraries and tools used by ASC coc
  - Integrates with TOSS and the vendor software stack
  - Derived from OpenHPC package recipes, similar look and feel **(add uarch optimizations, static libraries, -fPIC, add missing** s)
- FY19 Accomplishments
  - Deployed TOSS + ATSE at transition to SRN (May'19)
  - Developed ATSE 1.2 with support for 2x compilers and 2x MP {GNU7, ARM} x {OpenMPI3, HPE-MPI}
  - Built Trilinos and many ASC apps with ATSE
  - Packaged ATSE containers and tested up to 2048 nodes
- Future Directions
  - Migrate to Spack Stacks build
  - Add support for SNL adv. arch testbeds
  - Collaboration with RIKEN on McKernel



# Containerized SPARC HIFiRE-1 on Astra



**In job script:**

```
mpirun  \
        --map-by core \
        --bind-to core \
        singularity exec atse-astra-1.2.1.simg
        container_startup.sh
```

**container_startup.sh**

```
#!/bin/bash
module purge
module load devpack-gnu7
./sparc
```

**Early Results: SPARC on Astra, 56 MPI processes per node**

| Nodes | Trials | Native (seconds) | Container (seconds) | % Diff vs. Native |
|---|---|---|---|---|
| 128 | 2 | 8164 | 8169 | + 0.1% |
| 256 | 3 | 4473 | 4505 | + 0.7% |
| 512 | 3 | 2634 | 2636 | + 0.1% |
| 1024 | 1* | 1827 | 1762 | - 3.6% |
| 2048 | 2 | 1412 | 1429 | + 1.2 % |

**Points:**

- Supporting SPARC containerized build & deployment on Astra
- Enables easy test of new or old ATSE software stacks
- Near-native performance using a container
- Testing HIFiRE-1 Experiment (MacLean et al. 2008)

# Outline

- Astra Overview
- ATSE Software Stack
- Recent Application Results
- Conclusion – HPC on Arm, are we there yet?

# Application Porting Summary

- **Applications ported during open and restricted phases:**
  - **SNL: SPARC, EMPIRE, SPARTA, Xyce, NALU, HOMME-X, LAMMPS, CTH, Zapotec**
  - **LANL: FLAG, PARTISN, VPIC**
  - **LLNL: ALE3D, Ares, PF3D**

- Utilized ATSE provided software stack and modules
  - Early work on ATSE using testbeds helped to iron out some initial issues
- Performance results vary, in some cases Trinity Haswell/CTS-1 are faster, others are slower
- Astra shows good scalability out to 2,048 nodes
- Early indications are that still room for improvement in compilers and math libraries (subject of continuing Astra projects)

# Peak System Performance

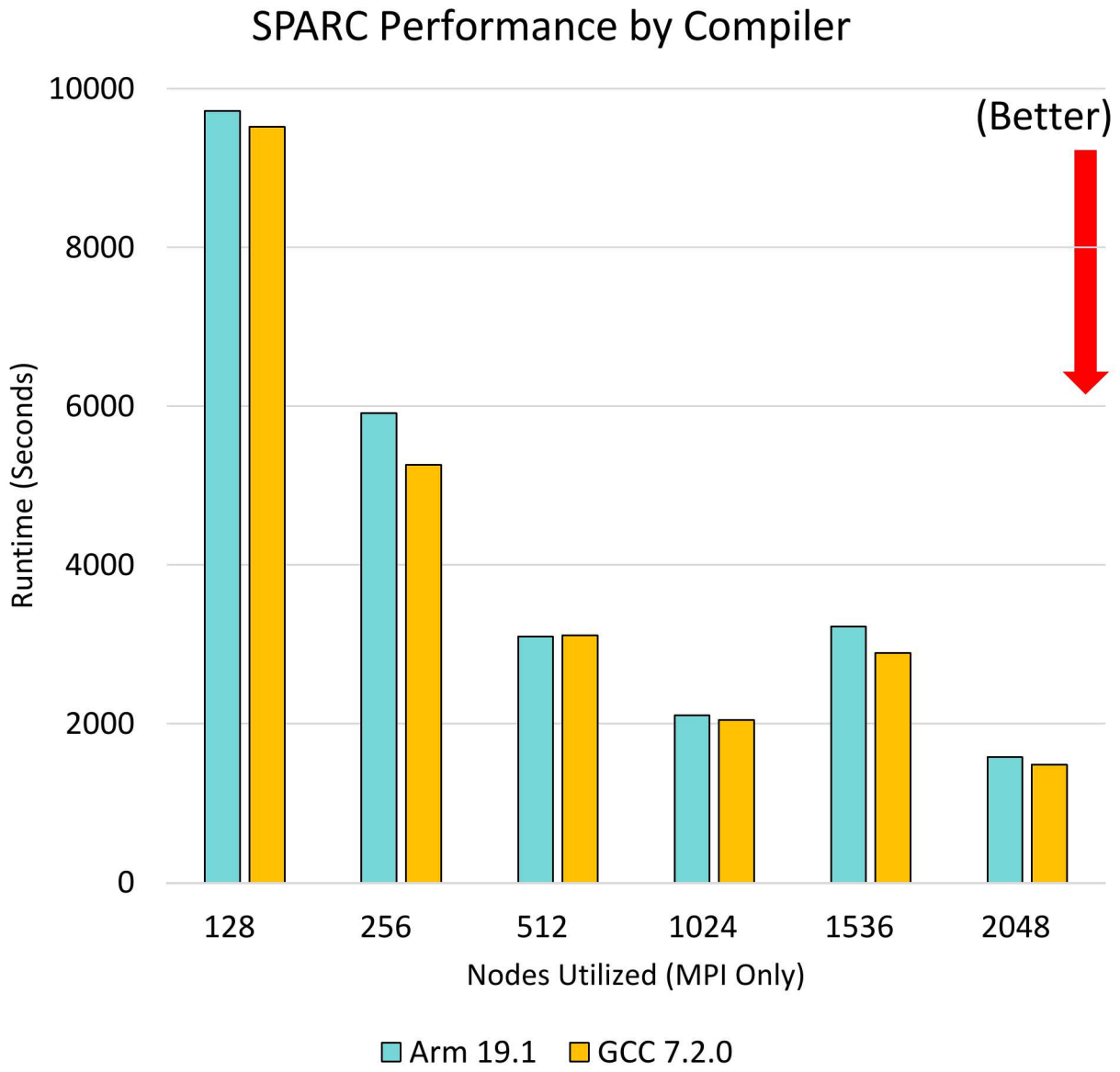
| | | CTS1 | Trinity | | Sierra | | Astra |
|---|---|---|---|---|---|---|---|
| | | Broadwell | Haswell | KNL | POWER9 | V100 GPU | ThunderX2 |
| LINPACK FLOP Rates (per Node) | Perf | 1.09 TF/s | ~0.86 TF/s | ~2.06 TF/s | ~1 TF/s | ~21.91 TF/s | ~0.71 TF/s |
| | Rel | 1.00X | 0.79X | 1.89X | 0.91X | 20.01X | 0.65X |
| Memory Bandwidth (STREAM) (per Node) | Perf | ~136 GB/s | ~120 GB/s | ~90 GB/s / ~350 GB/s | ~270GB/s | ~850 GB/s x 4 = ~3.4 TB/s | ~250 GB/s |
| | Rel | 1.00X | 0.88X | 0.66X / 2.57X | 1.99X | 25.00X | 1.84X |
| Power (Max TDP, per Node) | Watts | 120W x 2 = 240W | 135W x 2 = 270W | ~250W | 190W x 2 = 380W | ~300W x 4 = 1.2kW | ~180W x 2 = 360W |
| | Rel | 1.00X | 1.13X | 1.04X | 1.58X | 5.00X | 1.50X |

Guidance figures, used peak values for benchmarks and TDP

GCC/Arm Compiler Comparison

# SPARC CFD Simulation Code

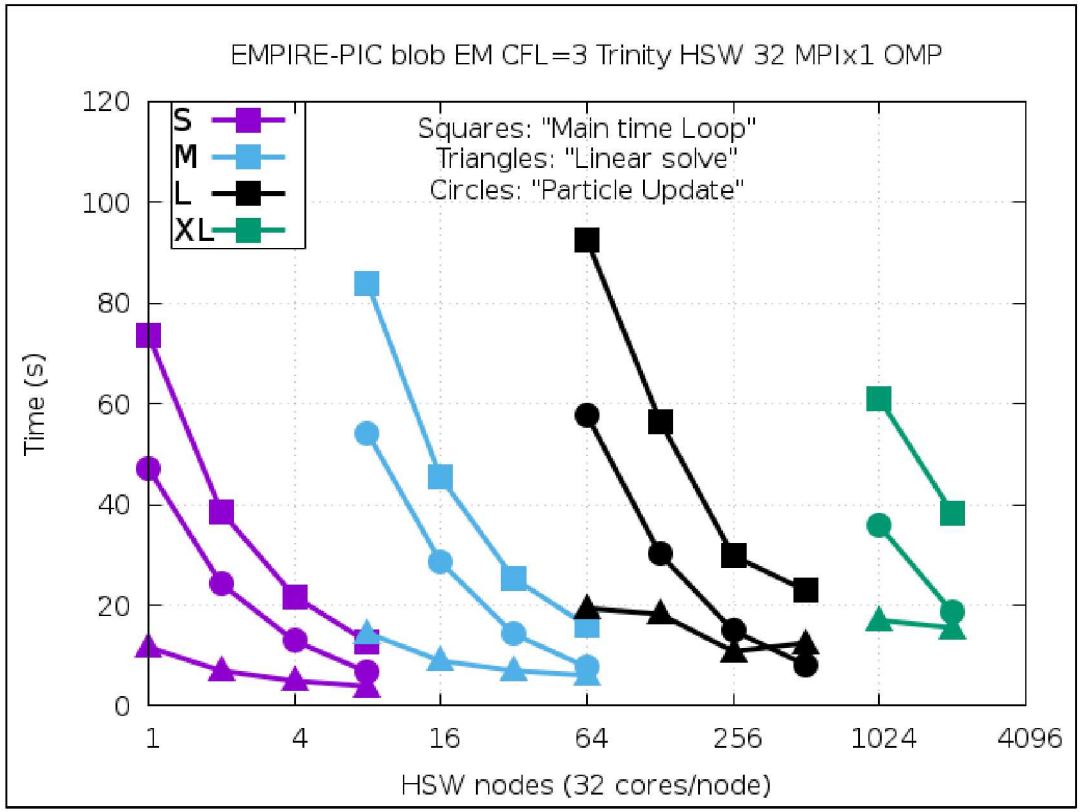
- SPARC is Sandia's latest CFD modeling code
  - Developed under NNSA ATDM Program
  - Written to be threaded and vectorized
  - Uses Kokkos programming abstractions
  - Approximately 2-3M lines of code for binary (including Trilinos packages, mostly C++, tiny bit of Fortran)
- Mixture of assembly and solve phases
- Successfully compiles with GCC and Arm HPC compilers on Astra
- Results show performance with Arm HPC compiler varies from 0.5% faster than GCC to 10% slower
  - This seems to be consistent across our code portfolio at present



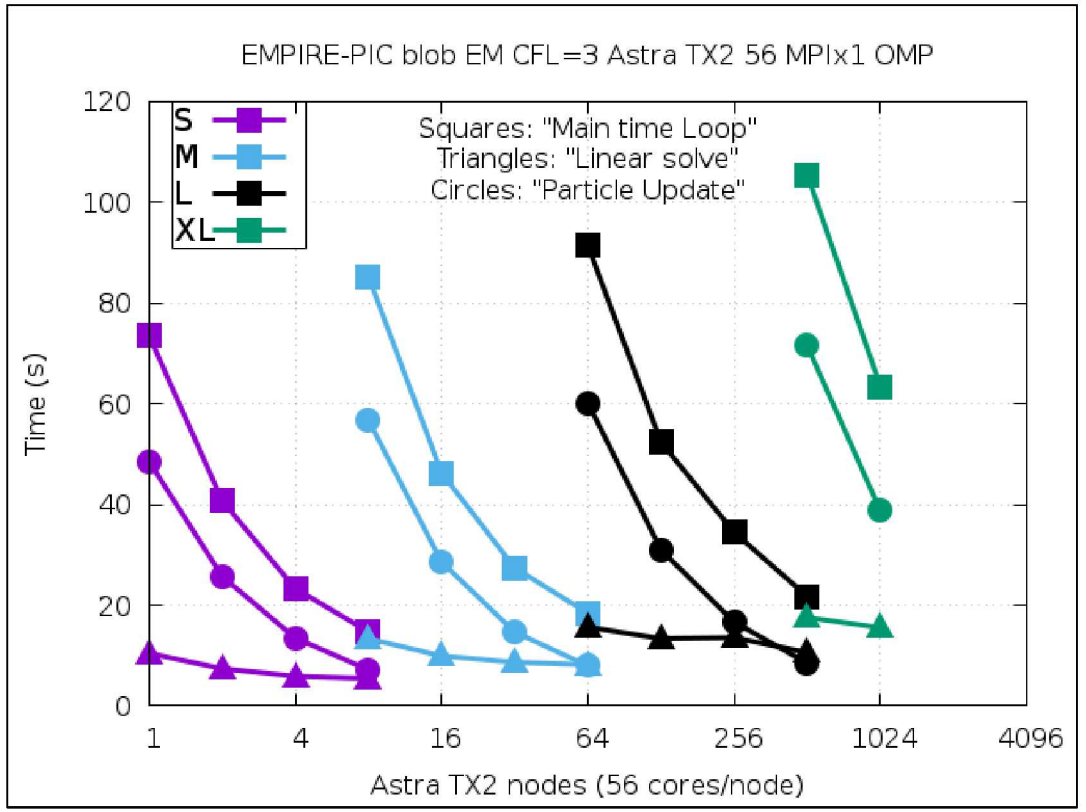
**Similar Performance TX2 vs. HSW**

# EMPIRE - PIC

(Better)




Trinity HSW 32 MPI x 1 OMP




Astra TX2 56 MPI x 1 OMP

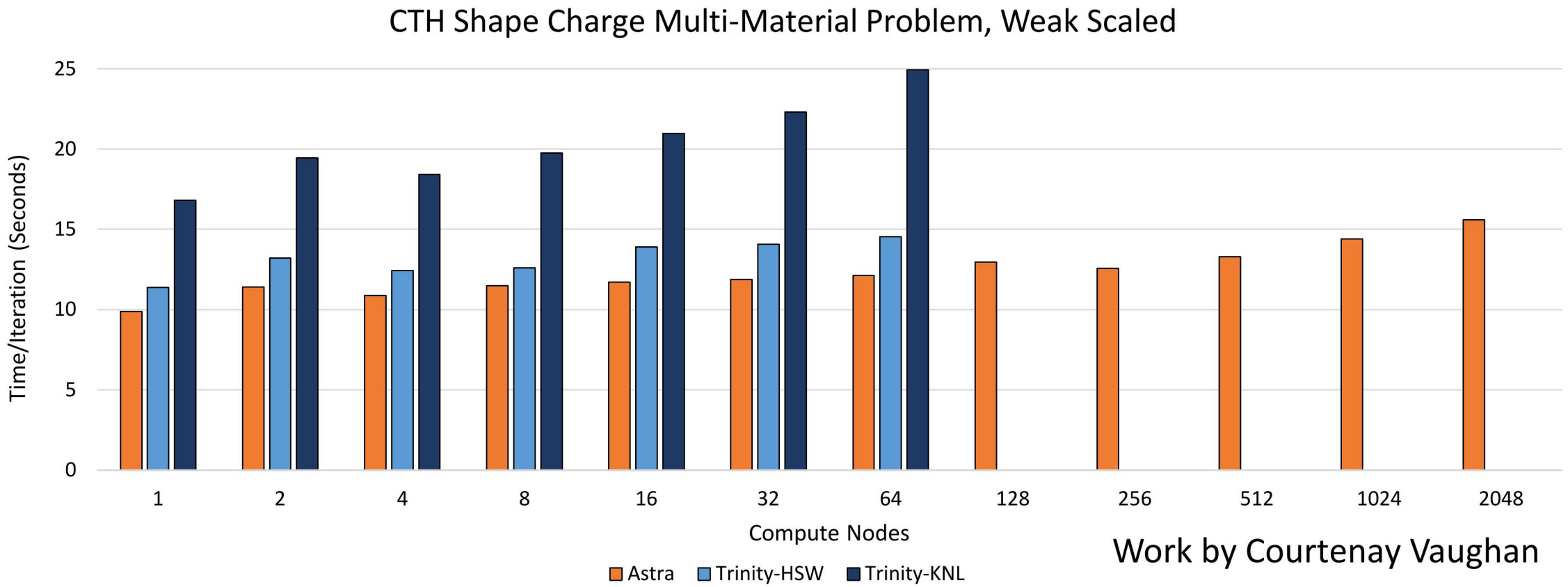
- Similar performance of Trinity-Haswell and Astra (MPI Only, performance is within 10% except for XL blob meshes which were run on fewer nodes for Astra)
- Similar scaling behavior between platforms

Work by Paul Lin and EMPIRE Team

**Fortran Compiler Stress Test**

# CTH (Hydrodynamics, Fortran)

(Better)

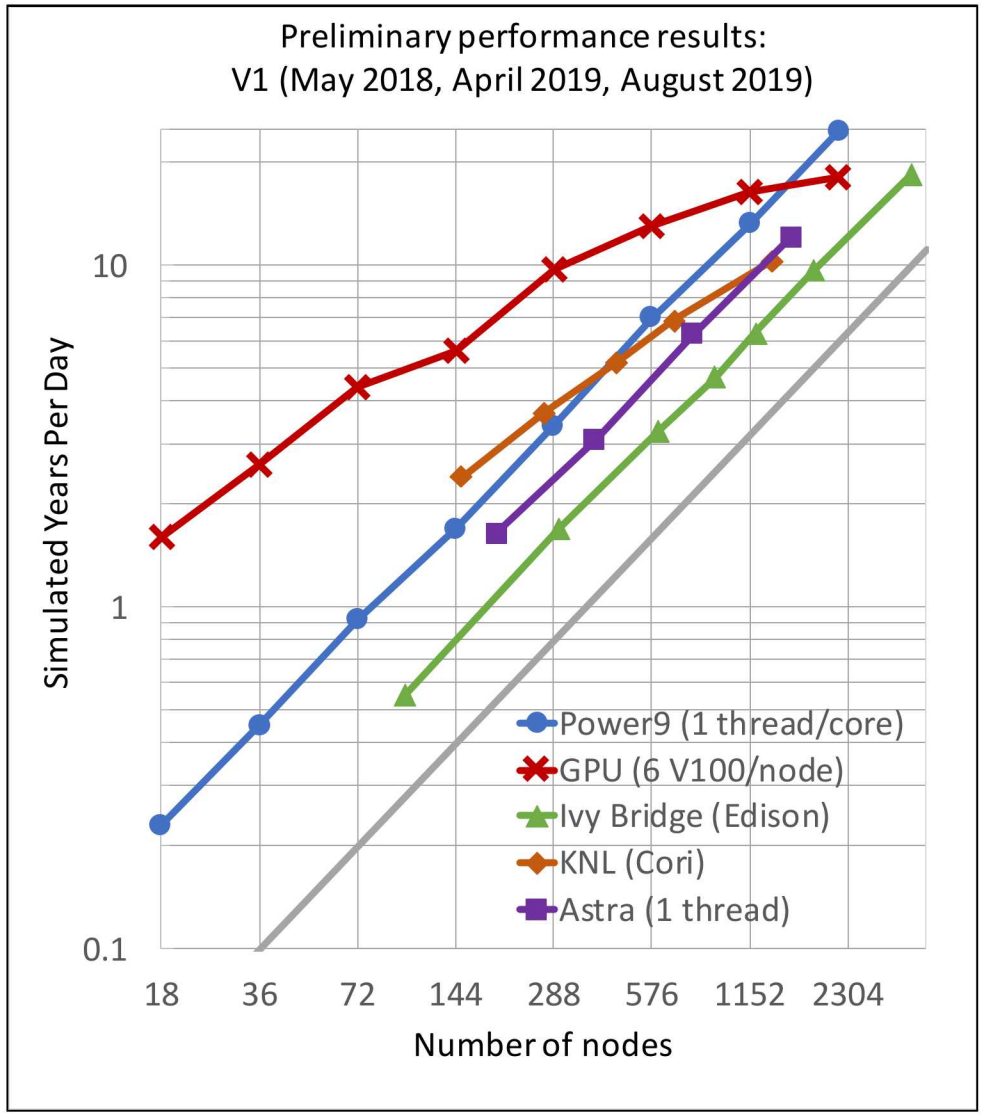



Work by Courtenay Vaughan

- CTH uses significant number of Fortran features (mixture from FORTRAN-IV to Fortran-90)
  - Large complex code which is extremely well trusted by analysts, known to be challenging with Fortran compilers on new platforms
- Successfully compiles with Arm Flang (used for these results) and ATSE-GCC installs

# HOMME (Climate)

(Better)




- Climate modeling code which is partially developed at Sandia (ASCR)
  - Known to drive components and third parties libraries very hard (frequently the first to find issues during porting)
  - Strong driver for improvements in Trillinos solver libraries across DOE platforms
- Good scalability (want to see near straight lines if possible)
- Recent SMT-2 results are around 10% faster

Work by Oksana Guba and HOMME Team

# Outline

- Astra Overview
- ATSE Software Stack
- Recent Application Results
- Conclusion – HPC on Arm, are we there yet?

# HPC on Arm, are we there yet? … Yes

- Basic HPC components supported and demonstrated @ scale
  - InfiniBand, UCX, MPI, Lustre, Linux, SLURM, …
- Compilers and math libraries work sufficiently well to get codes running
- Performance competitive with leading alternatives
- Offerings from a range of integrators

# HPC on Arm, are we there yet? … Yes

- Basic HPC components supported and demonstrated @ scale
  - InfiniBand, UCX, MPI, Lustre, Linux, SLURM, …
- Compilers and math libraries work sufficiently well to get codes running
- Performance competitive with leading alternatives
- Offerings from a range of integrators

- SVE not proven yet, lack of accelerator options (changing)
- Performance not tuned in many packages / kernels yet
  - Need threaded and vectorized versions of kernels
- Still need work on profilers, debuggers, and memory correctness
- Lacking of standards for performance counters + power/energy

# Questions?

*Exceptional Service in the National Interest*