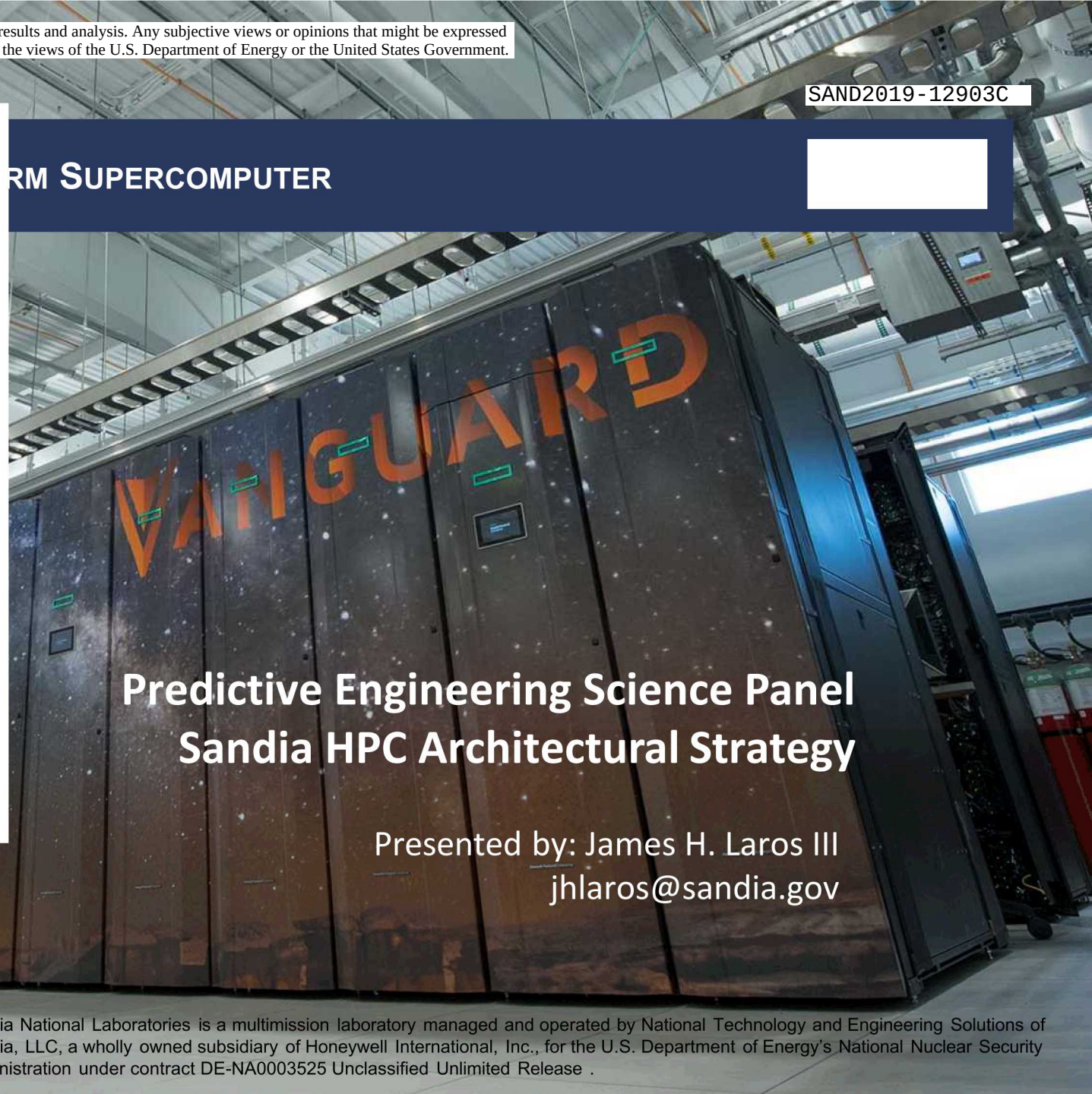
SAND2019-12903C

RM SUPERCOMPUTER



Predictive Engineering Science Panel
Sandia HPC Architectural Strategy

Presented by: James H. Laros III
jhlaros@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 .

Outline

- Brief Curriculum vitae
- Alliance for Computing at Extreme Scale (ACES)
- Advanced Architecture Testbed Program
- Advanced Technology Prototype Program
- Questions???

Curriculum vitae

ASCI Red

Cplant

Red Storm

Sandia has had a played a significant leadership and collaborative role in hardware architectures and system software for multiple decades

Alliance for Computing at Extreme Scale (ACES)

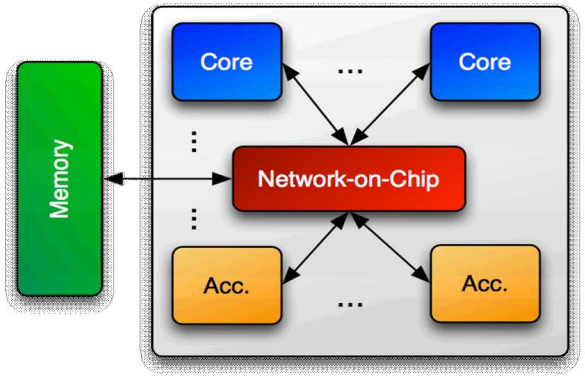
CROSSROADS

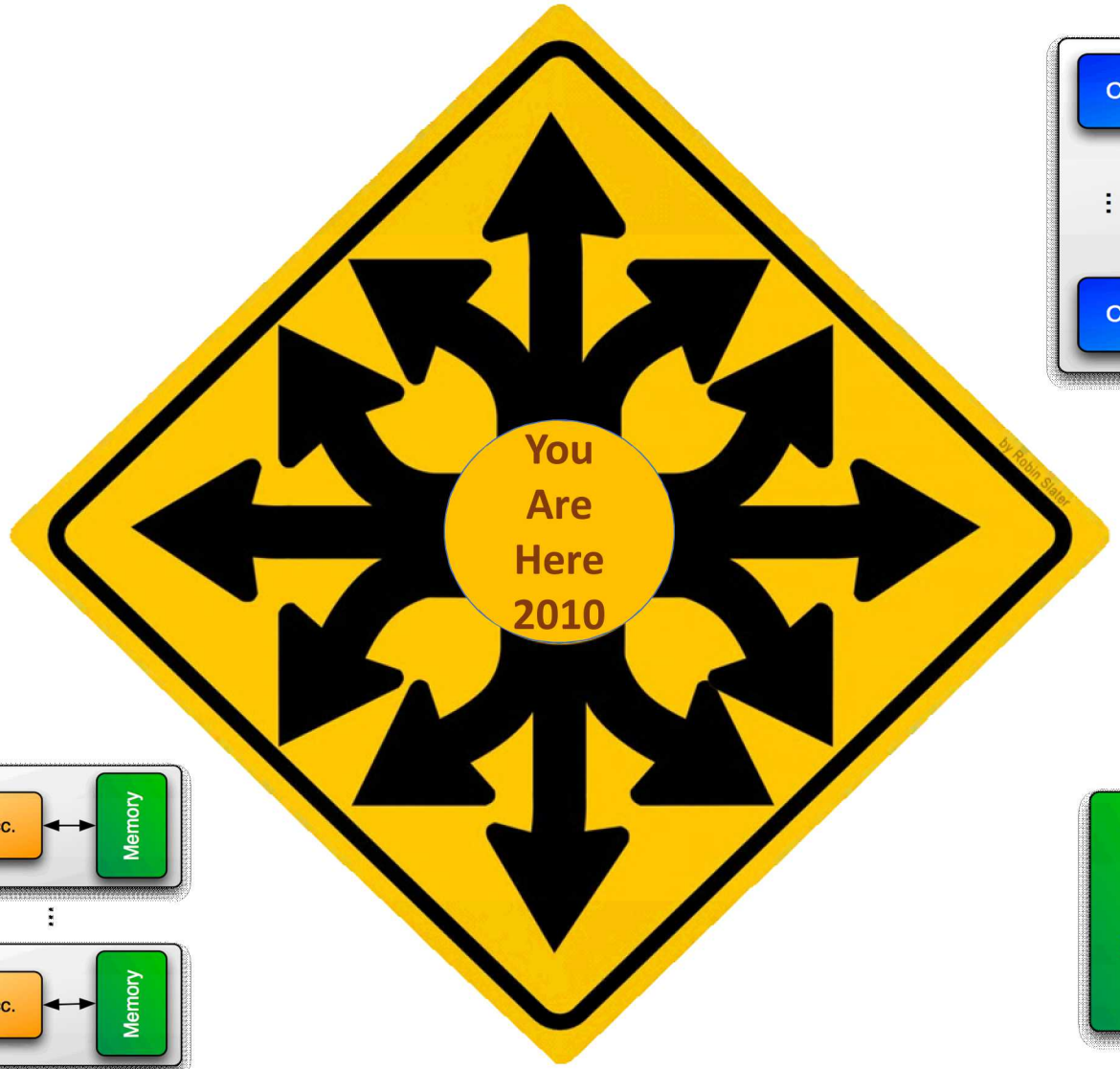
Sandia has partnered with Los Alamos Laboratory since 2008 to field Cielo, a capability class platform and Trinity, the first Advanced Technology Systems (ATS-1). ATS-3, Crossroads, is expected to be deployed in the 2021 timeframe.

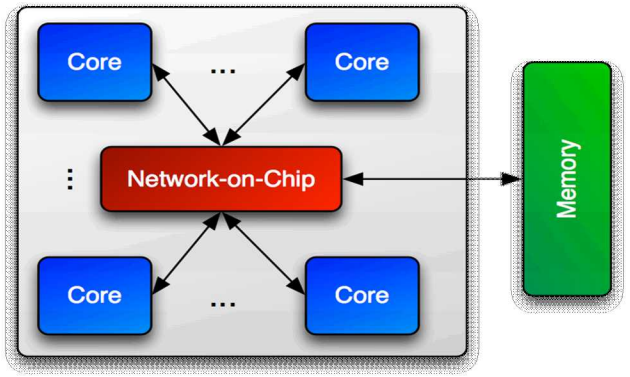
ASC

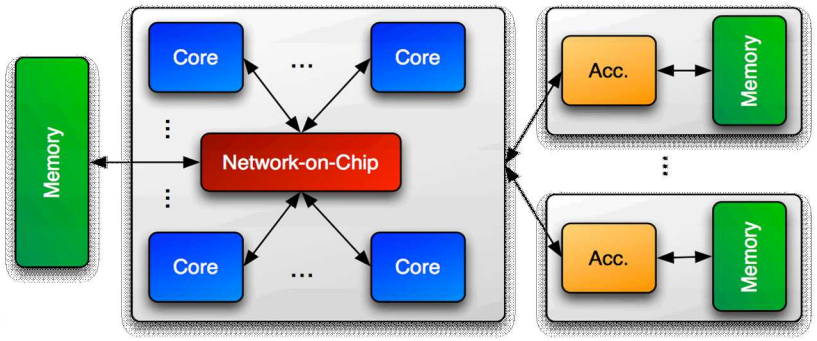
NNSA

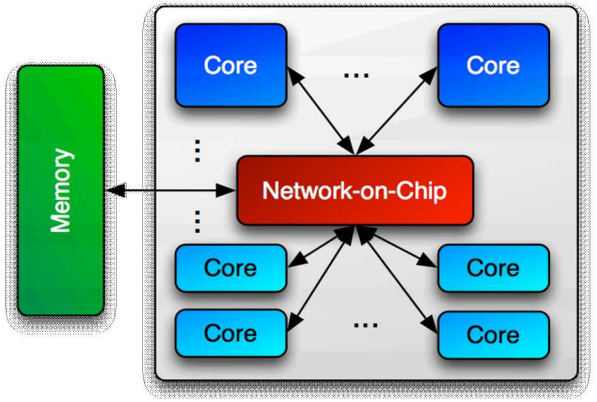
Advanced Architecture Testbed Program











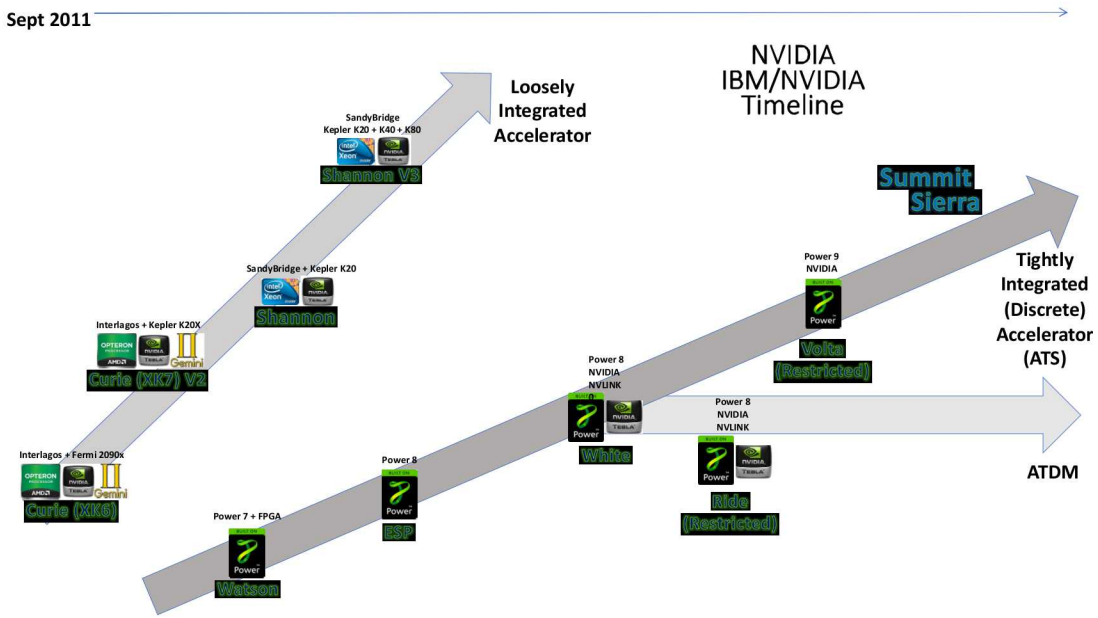
….to be a scout for future computer architectures

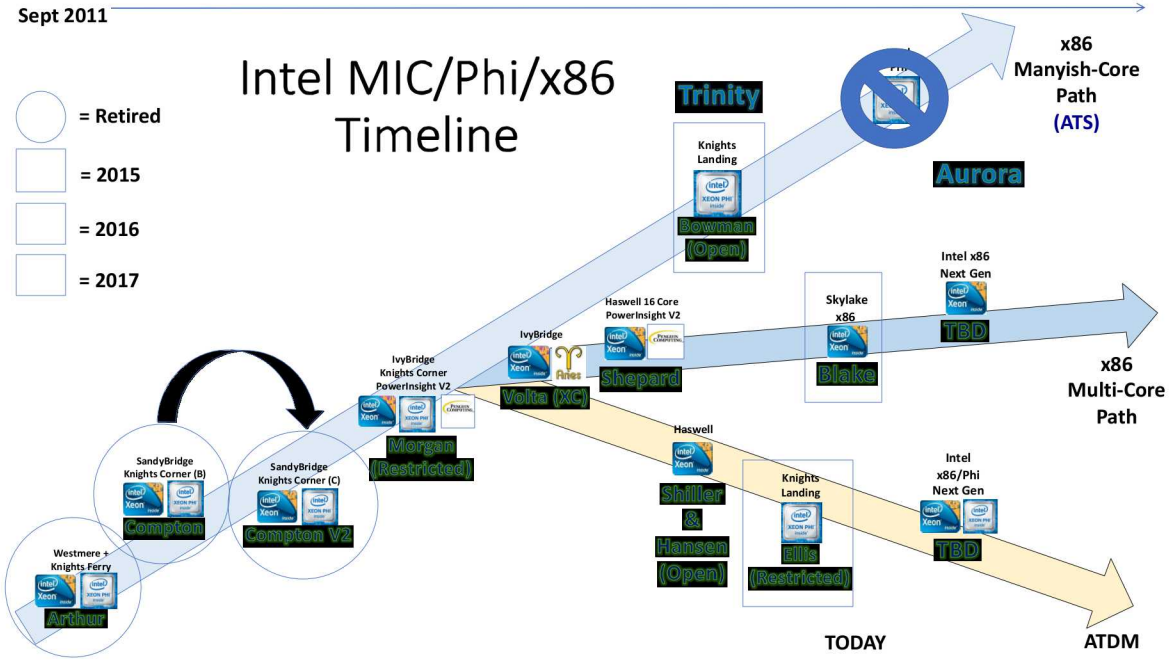
Advanced Architecture Testbed Program

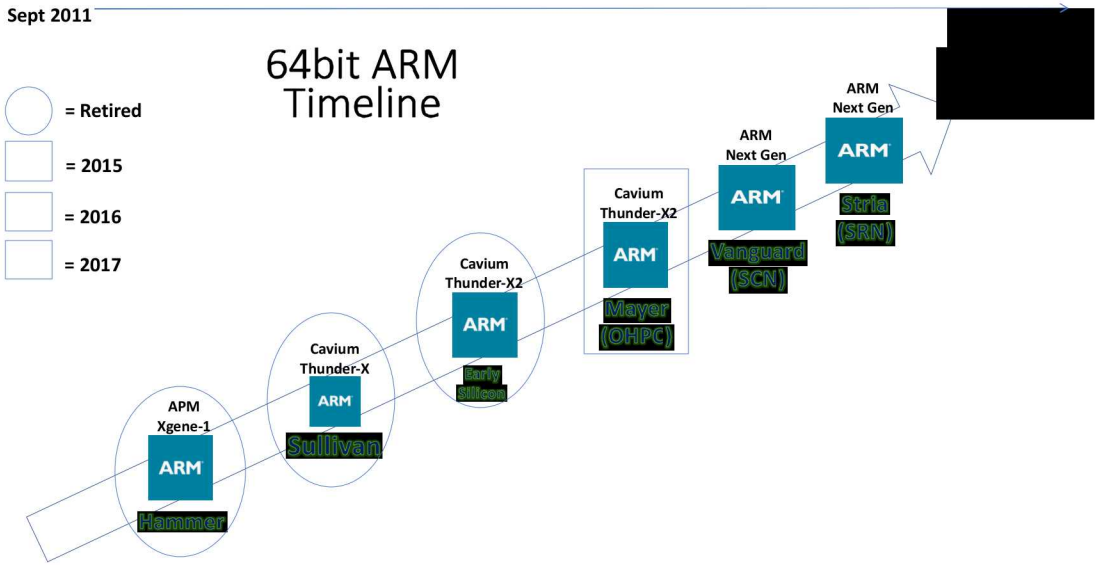
- **APPLICATION FOCUSED**
- Reduce impact on mission labs in a rapidly changing technology environment.
  - Significant **PRODUCTION** code rewrite/modification may be required
  - Ensure that when we make "the change" it is the right move for code longevity, porting efforts, performance *etc.*
  - Go through the pain up front so the transition for full codes is easier
  - Eliminate or reduce missteps

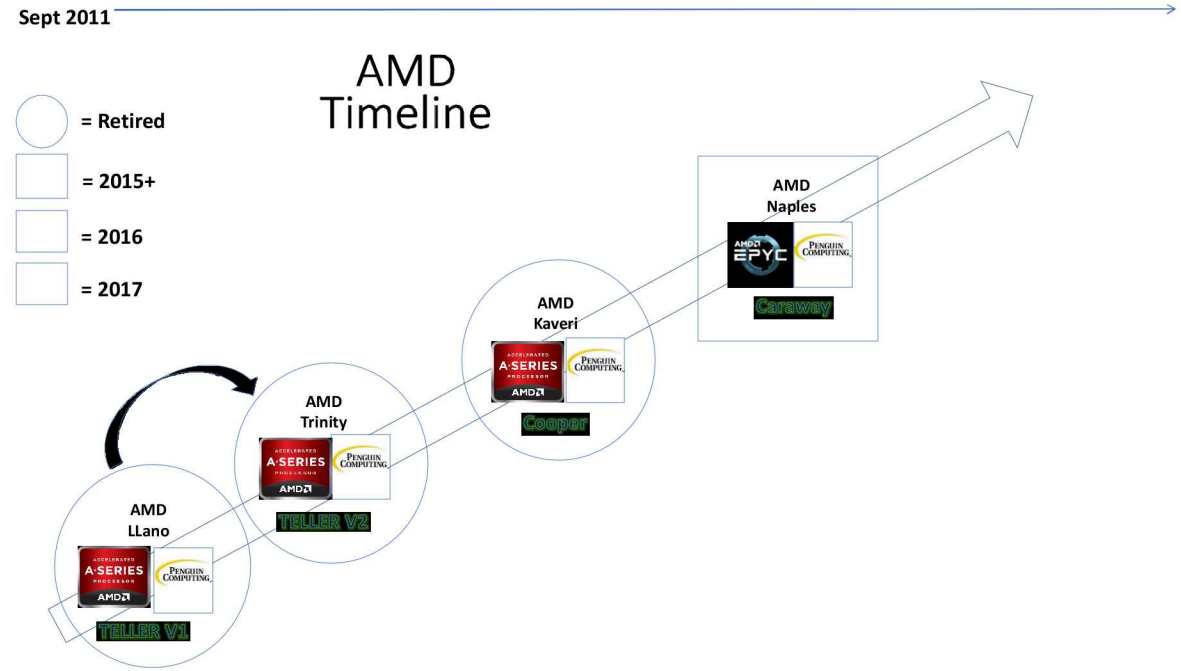
Mini-Apps

Production Codes

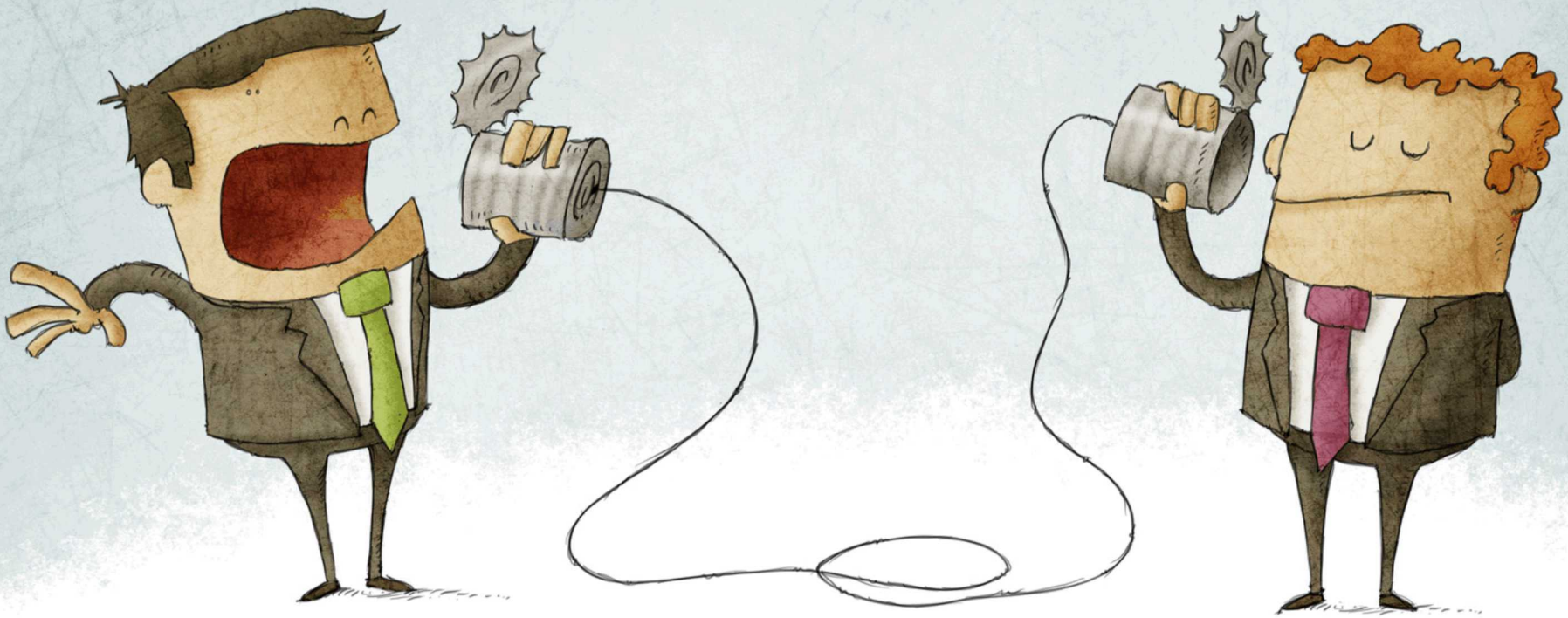

Sept 2011
NVIDIA
IBM/NVIDIA
Timeline
Loosely
Integrated
Accelerator
Tightly
Integrated
(Discrete)
Accelerator
(ATS)
SandyBridge
Kepler K20 + K40 + K80
Shannon V3
SandyBridge + Kepler K20
Shannon
Interlagos + Kepler K20X
Gemini
Curie (XK7) V2
Interlagos + Fermi 2090x
Gemini
Curie (XK6)
Power 7 + FPGA
Watson
Power 8
ESP
Power 8
NVIDIA
NVLINK
White
Power 8
NVIDIA
NVLINK
Ride
(Restricted)
Power 9
NVIDIA
Volta
(Restricted)
Summit
Sierra
ATDM


Sept 2011
Intel MIC/Phi/x86
Timeline
= Retired
= 2015
= 2016
= 2017
x86
Manyish-Core
Path
(ATS)
Aurora
Trinity
Knights
Landing
Bowman
(Open)
Intel x86
Next Gen
TBD
x86
Multi-Core
Path
Skylake
x86
Blake
Haswell 16 Core
PowerInsight V2
Shepard
IvyBridge
Volta (XC)
IvyBridge
Knights Corner
PowerInsight V2
Morgan
(Restricted)
SandyBridge
Knights Corner (B)
Compton
SandyBridge
Knights Corner (C)
Compton V2
Westmere +
Knights Ferry
Arthur
Haswell
Shiller
&
Hansen
(Open)
Knights
Landing
Ellis
(Restricted)
Intel
x86/Phi
Next Gen
TBD
TODAY
ATDM


Sept 2011
64bit ARM
Timeline
= Retired
= 2015
= 2016
= 2017
APM
Xgene-1
ARM
Hammer
Cavium
Thunder-X
ARM
Sullivan
Cavium
Thunder-X2
ARM
Early
Silicon
Cavium
Thunder-X2
ARM
Mayer
(OHPC)
ARM
Next Gen
ARM
Vanguard
(SCN)
ARM
Next Gen
ARM
Stria
(SRN)


Sept 2011
AMD
Timeline
= Retired
= 2015+
= 2016
= 2017
AMD
Naples
Caraway
AMD
Kaveri
Cooper
AMD
Trinity
TELLER V2
AMD
LLano
TELLER V1

- The Testbed program leverages Sandia's rich engineering and architecture history
- **The Testbed program acts as a conduit for MEANINGFUL conversation and co-design**
  - Not limited by duration of single procurement



Arm Testbed Timeline

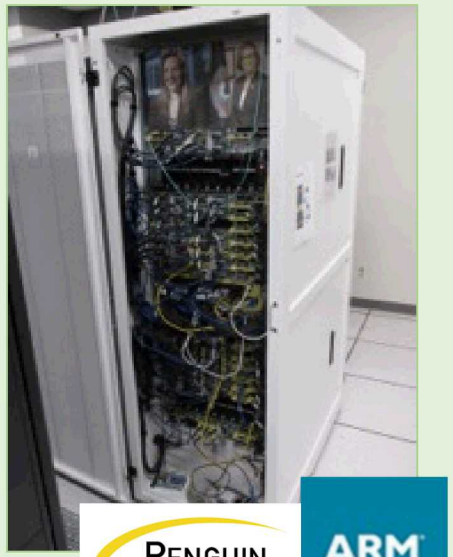
2014

Hewlett Packard Enterprise ARM

Hammer

Applied Micro
X-Gene-1
47 nodes

2015



Penguin Computing ARM

Sullivan

Cavium ThunderX1
32 nodes

2017

Hewlett Packard Enterprise ARM

Mayer

Pre-GA Cavium
ThunderX2
47 nodes

2018

Hewlett Packard Enterprise ARM

Vanguard/Astra/Stria

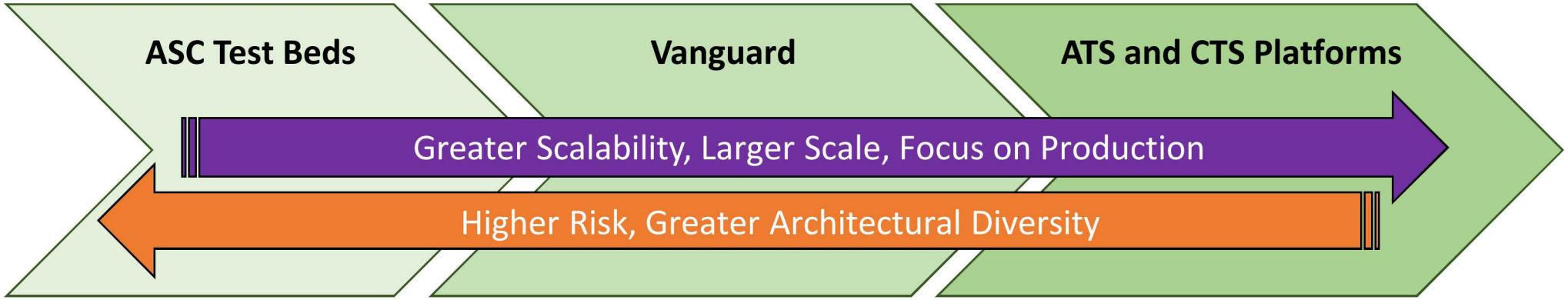
HPE Apollo 70
Cavium ThunderX2
2592 nodes

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 (ATSE)
- First Prototype platform targeting Arm Architecture

Tri-lab collaboration is integral to Vanguard

Vanguard Program: Advanced Technology Prototype Systems



Test Beds

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

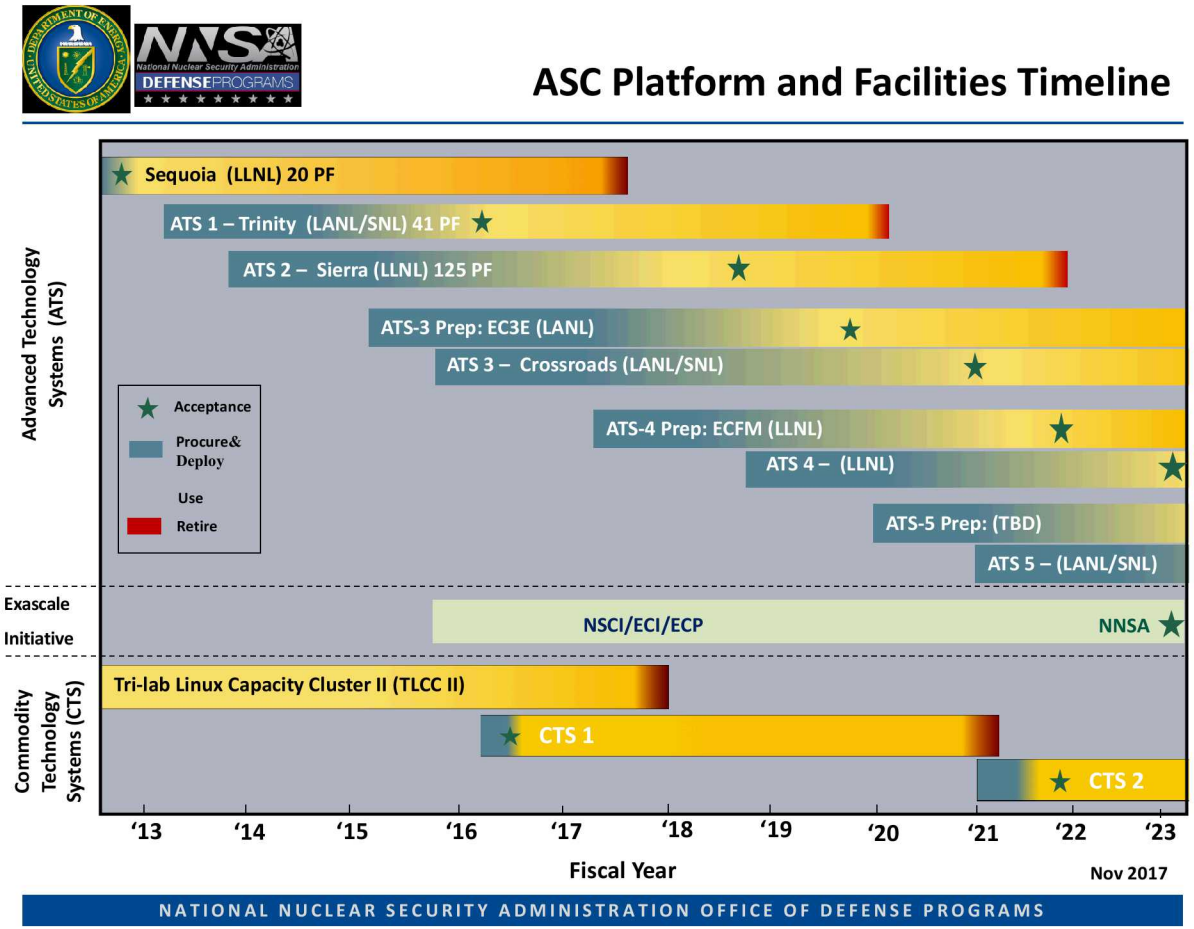
Vanguard

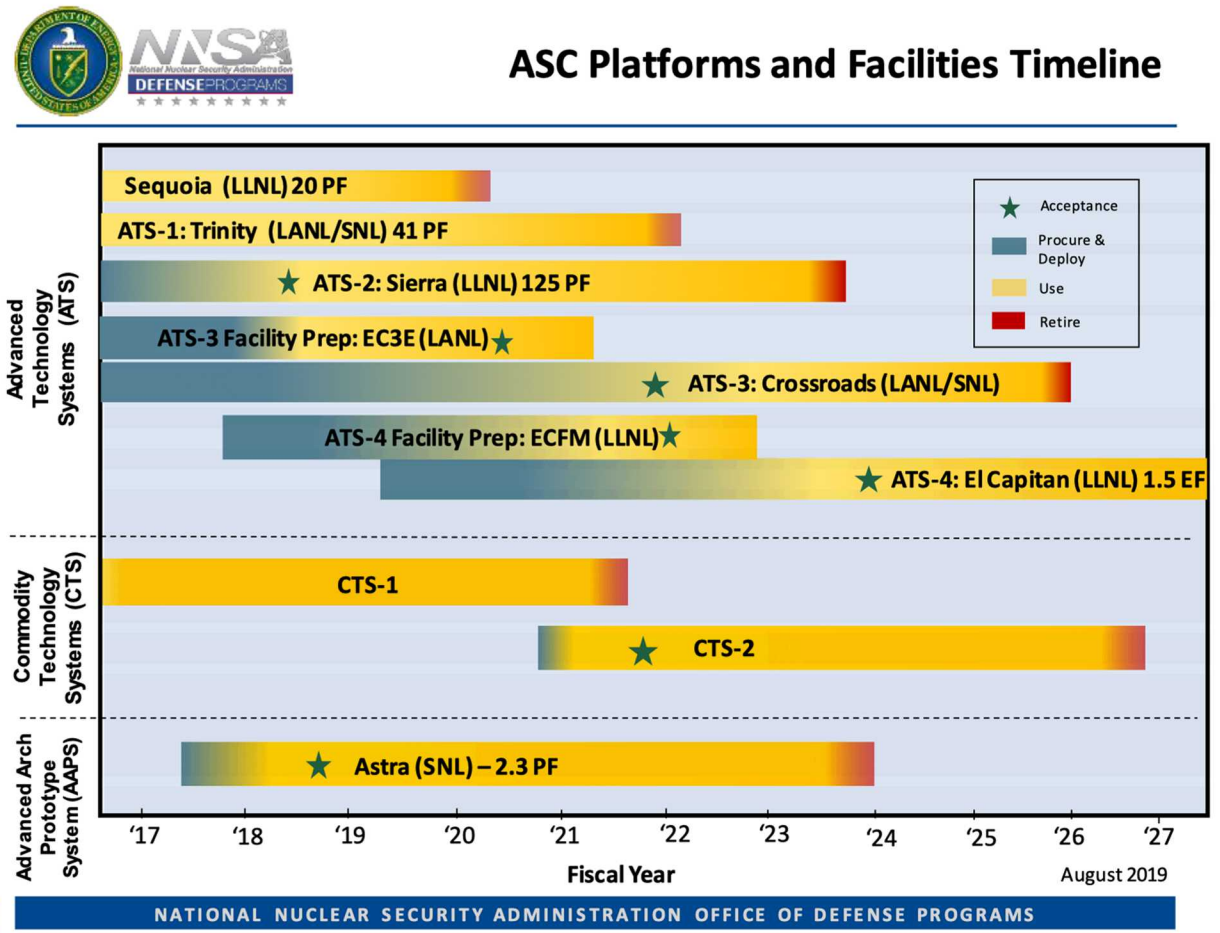
- Larger-scale experimental systems
- Focused efforts to mature new technologies
- Broader user-base
- Not production, seek to increase technology and vendor choices
- **DOE/NNSA Tri-lab resource**

Production Platforms

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

NNSA Platform Strategy Evolution





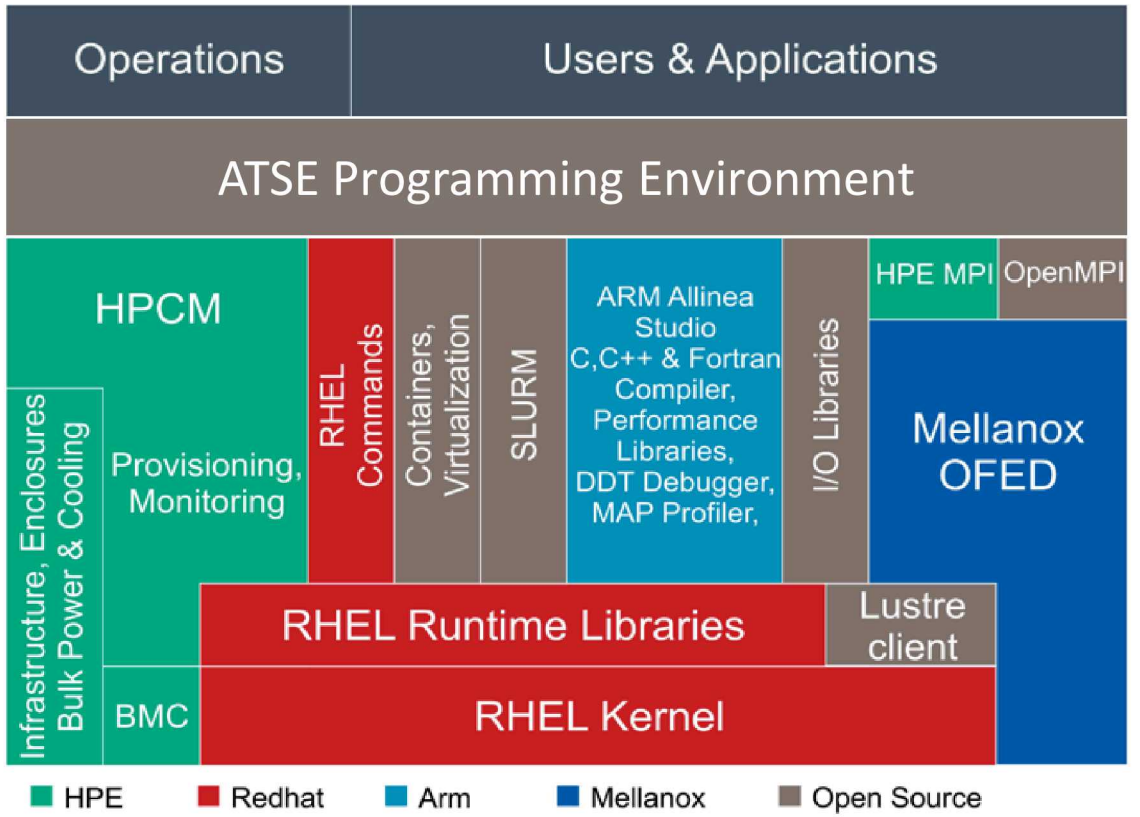
Addition of Advanced Architecture Prototype Systems

Positioning Future Vanguard Platform

- What is the right Risk profile?
- Balance R&D with production application support
- Should Vanguard platforms always end up in the classified environment?
- What is the right scale? Depends on:
  1. Technologies being investigated
  2. Software ecosystem considerations
  3. Budget
  4. Intersection of all

ATSE: an Integrated Software Environment for ASC Workloads

- Advanced Tri-lab Software Environment
  - User-facing programming environment co-developed with Astra
  - Provides a common set of libraries and tools used by ASC codes
  - Integrates with TOSS and the vendor software stack
- FY19 Accomplishments
  - Deployed TOSS + ATSE at transition to SRN (May'19)
  - Developed ATSE 1.2 with support for 2x compilers and 2x MPIs: {GNU7, ARM} x {OpenMPI3, HPE-MPI}
  - Packaged ATSE containers and tested up to 2048 nodes
  - Built Trilinos and several L2 milestone apps using ATSE
  - Obtained CNARS approval for SCN
- Future Directions
  - Migrate to Spack Stacks build (currently OpenHPC/RPM based)
  - Add support for SNL testbeds
  - Collaboration with RIKEN on McKernel

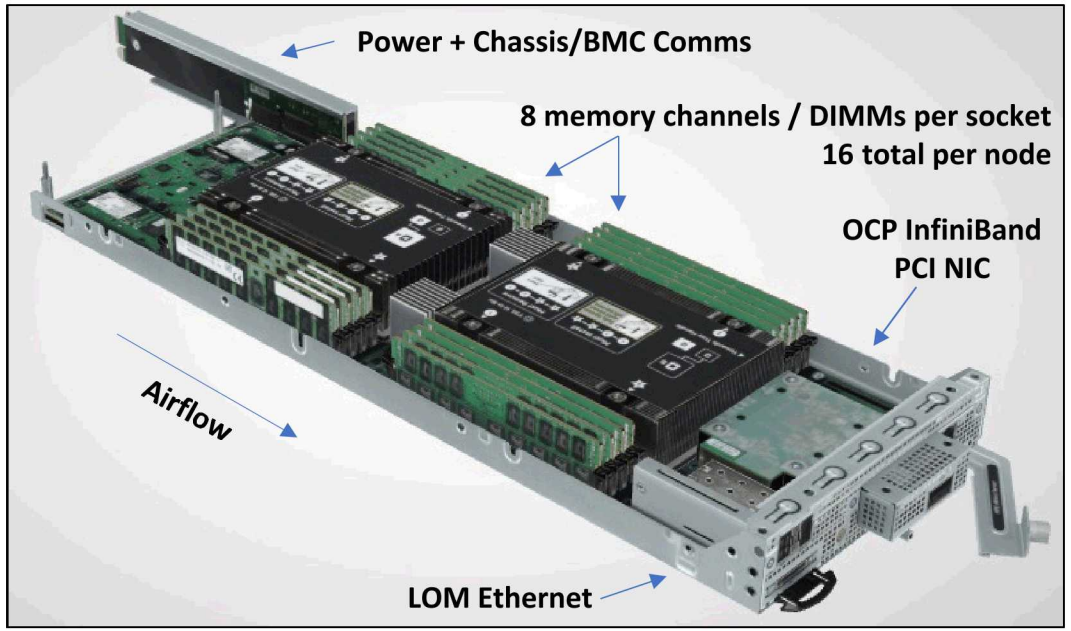


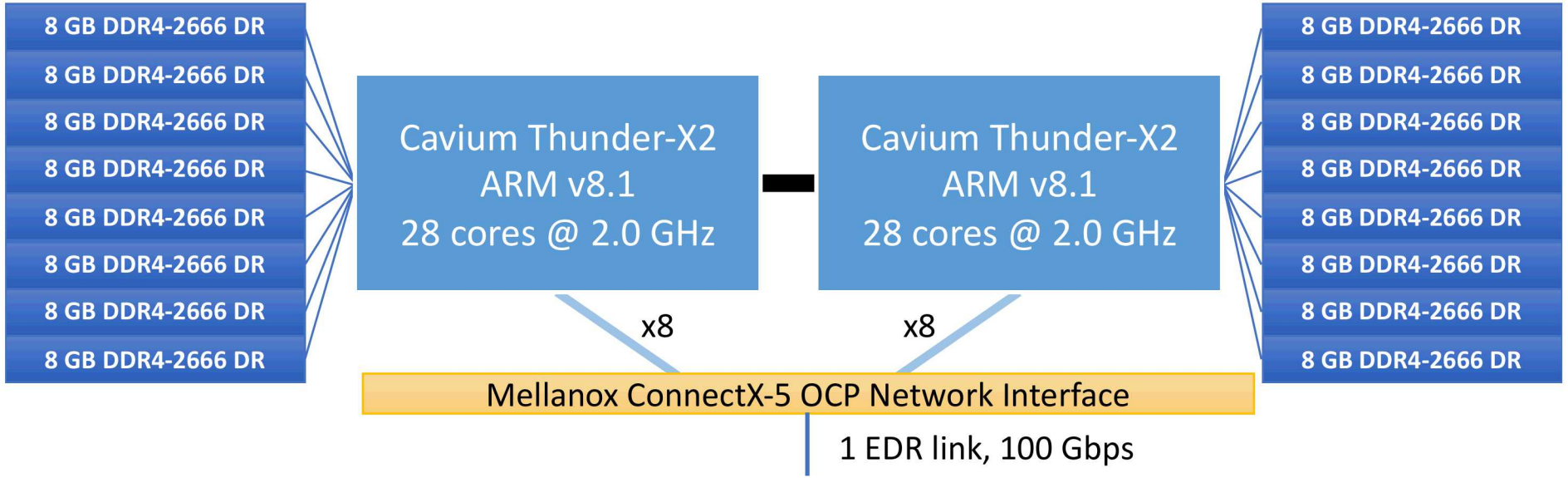
It Takes an Incredible Team…

Questions?

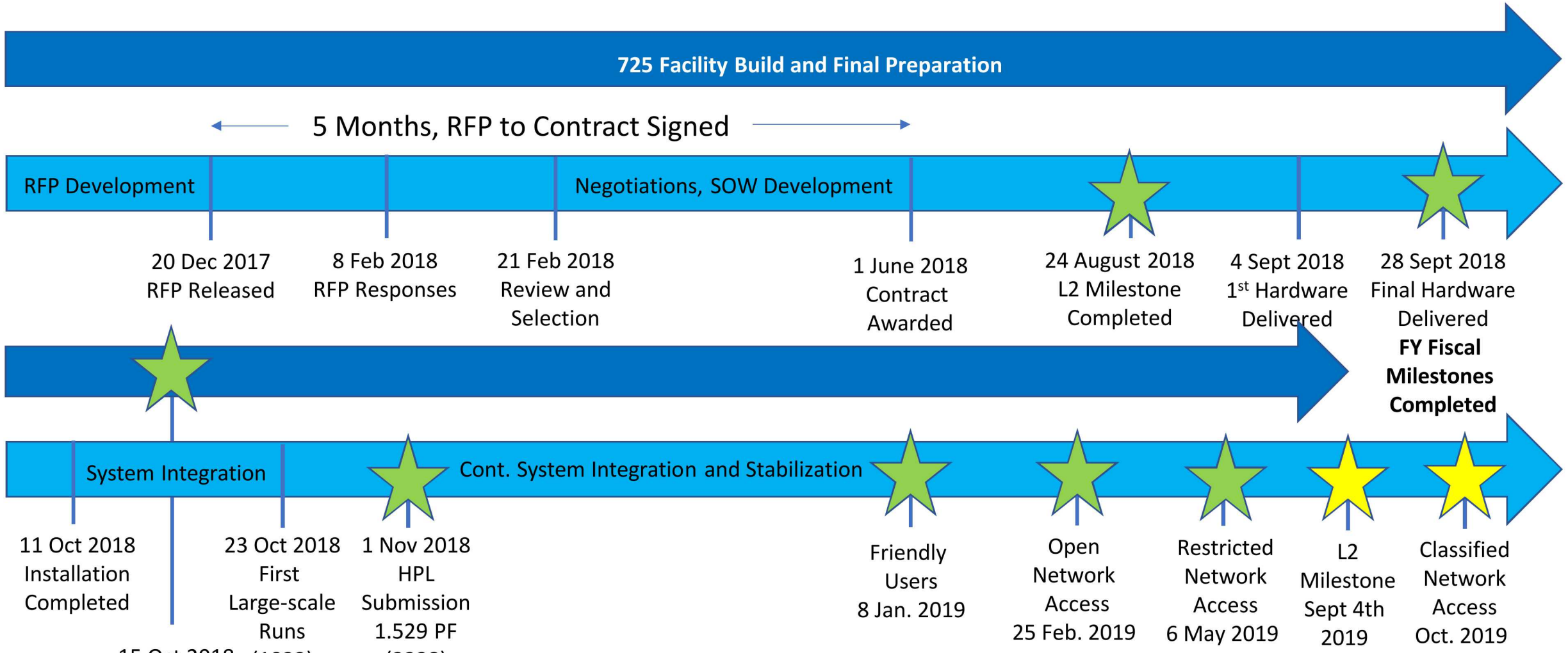
Astra Node Architecture

- **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





Vanguard-Astra: Timeline



June 2019 Top500:

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

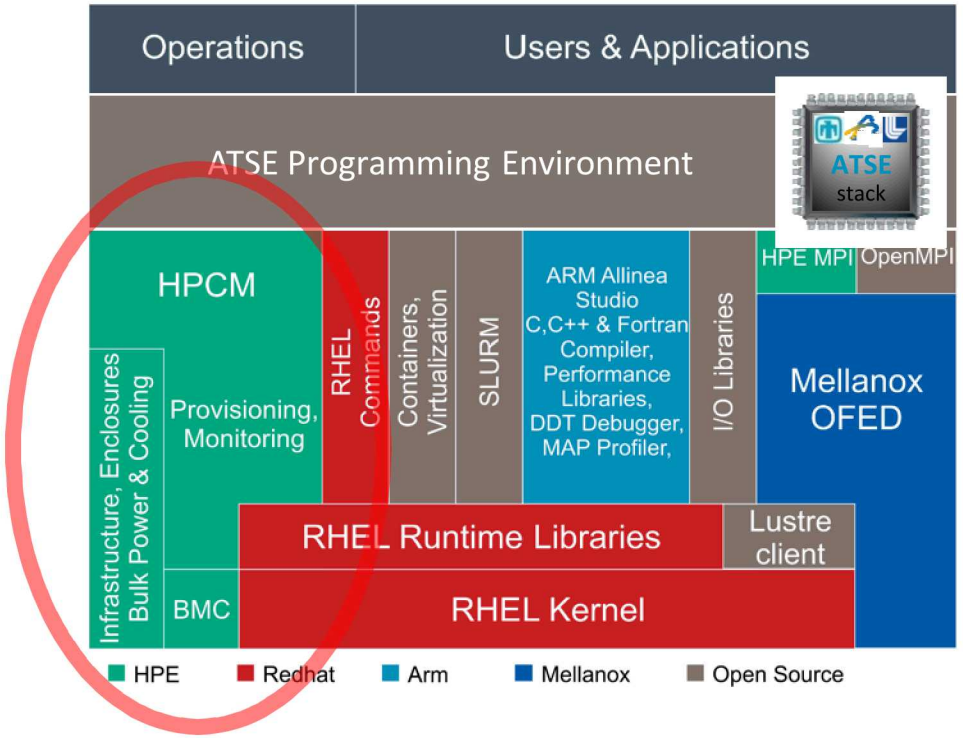
HPCG: #29 @ 90.92 TFlop/s on 2385 nodes (up from #38)

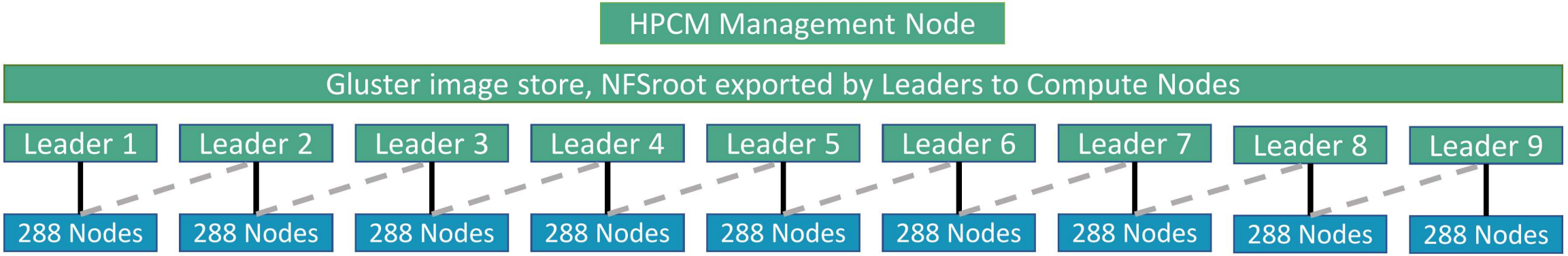
Real-Time System Monitoring Has Been Key

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

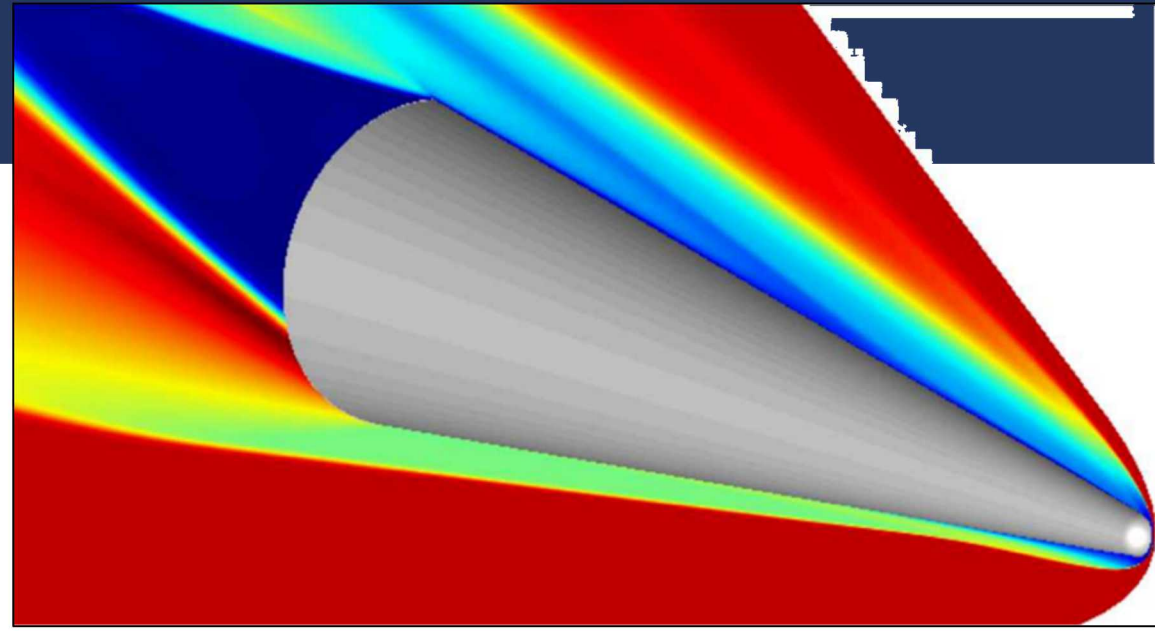
HPCM Provides Scalable System Management for Astra

- 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)





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)