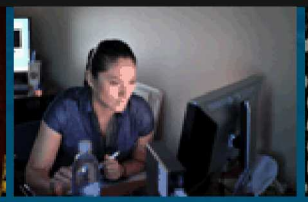
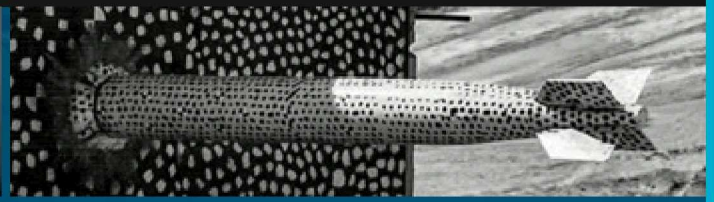
This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2018-7214C

VANGUARD





PRESENTED BY

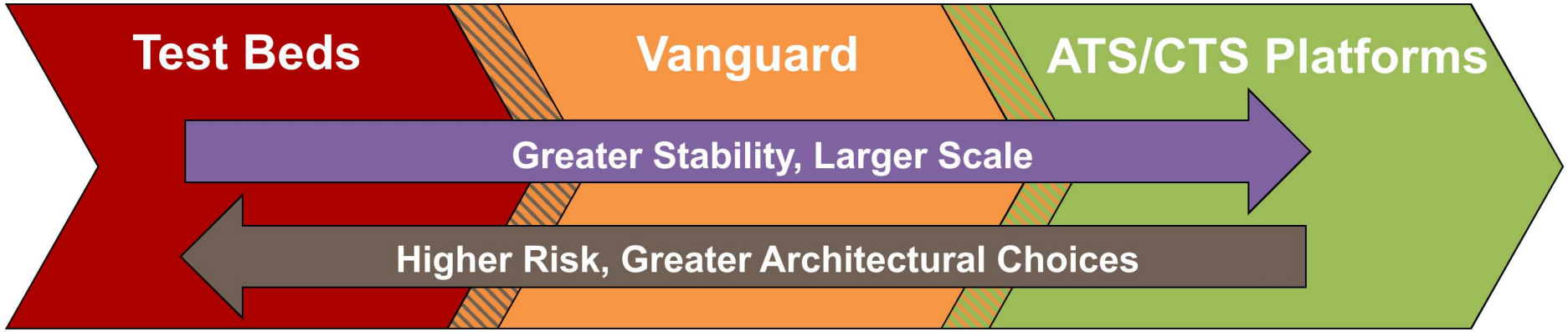
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Vanguard Program: Advanced Architecture 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

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

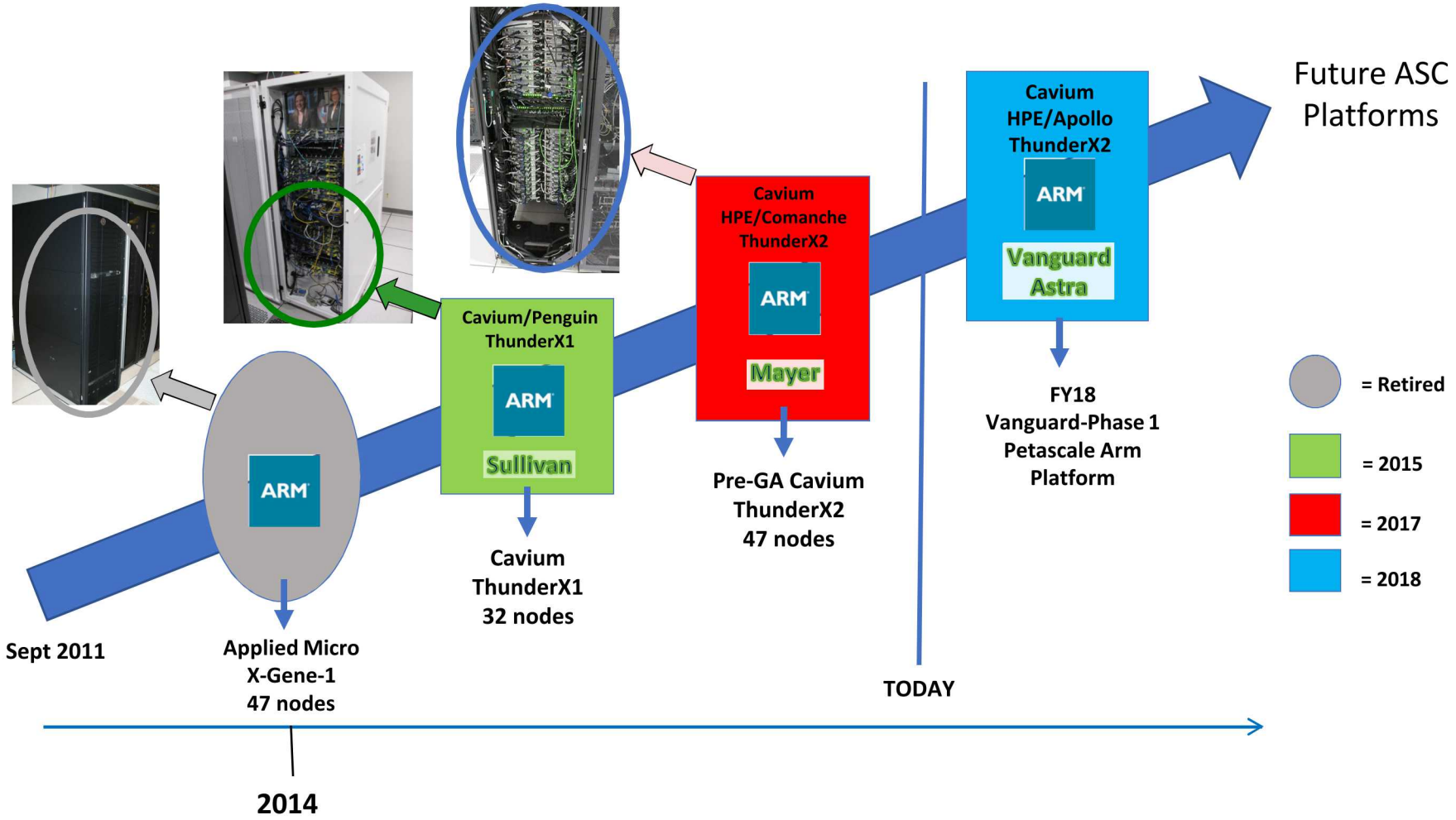
Vanguard

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

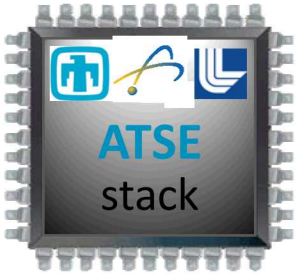
ATS/CTS Platforms

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

Sandia's NNSA/ASC ARM Platforms



Vanguard Tri-lab Software Effort

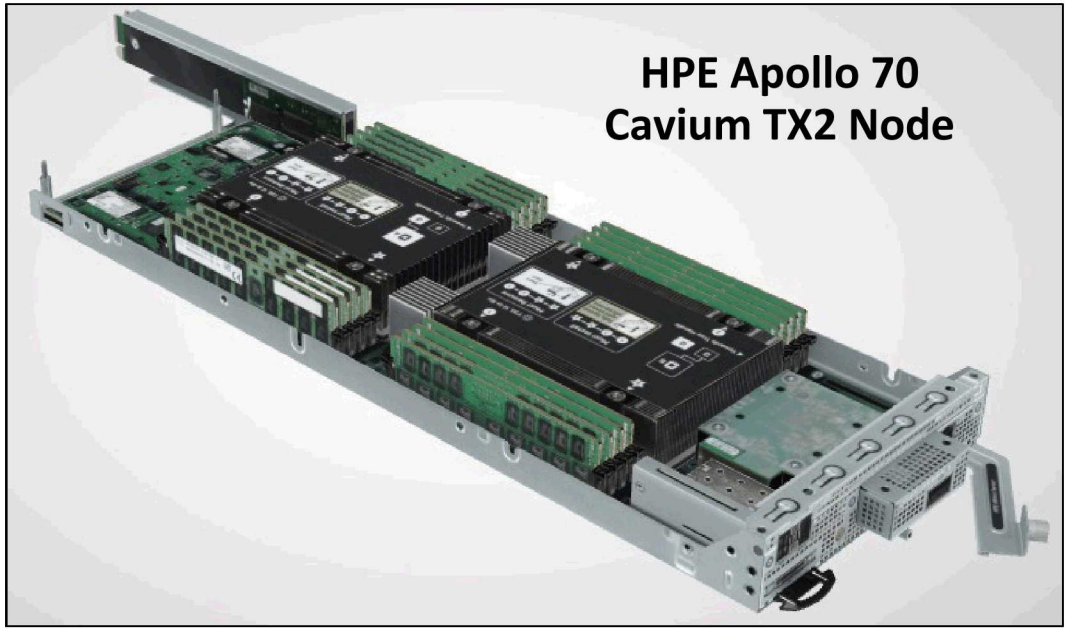


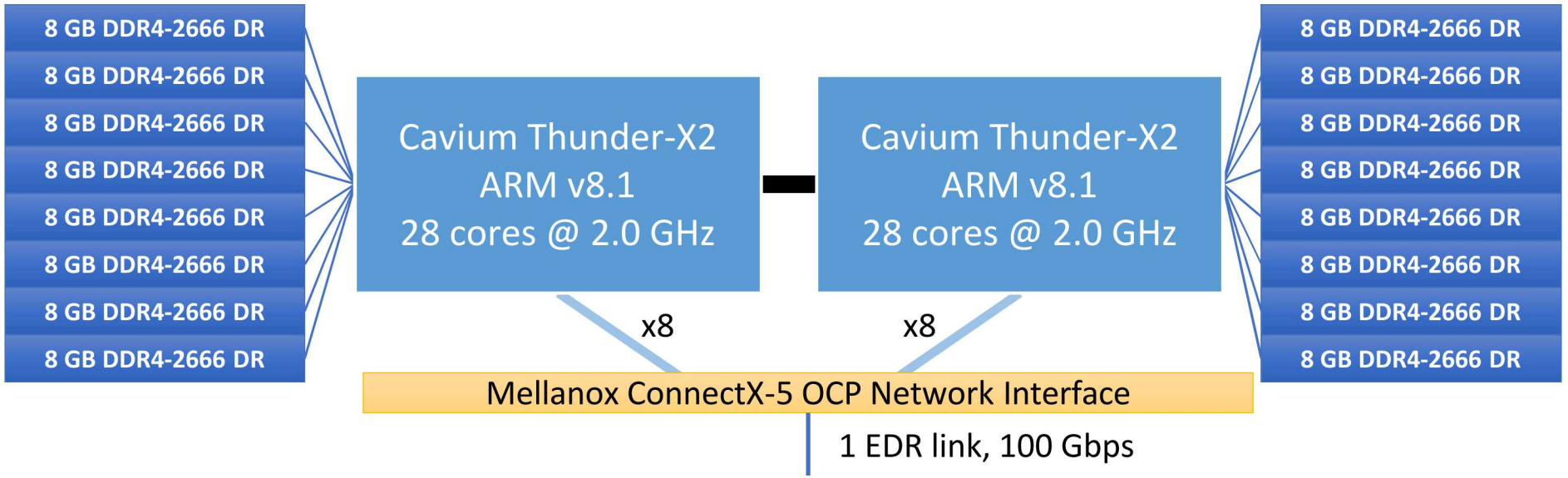
- **A**dvanced **T**ri-lab **S**oftware **E**nvironment
- Accelerate maturity of Arm ecosystem for ASC computing
  - Prove viability for NNSA integrated codes running at scale
  - Harden compilers, math libraries, tools, communication libraries
    - Heavily templated C++, Fortran 2003/2008, Gigabyte+ 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 (optimized Linux, network, filesystems, containers/VMs, ...)
  - Job scheduling and management (WLM, app launcher, user tools, ...)
  - System management (boot, system monitoring, image management, ...)
- Stack will be hardware agnostic
  - Initial focus Astra/Arm

Improve 0 to 60 time... Arm system arrival to useful work done

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





Astra Architecture

HPE Apollo 70 Chassis: 4 nodes

18 chassis/rack

72 nodes/rack

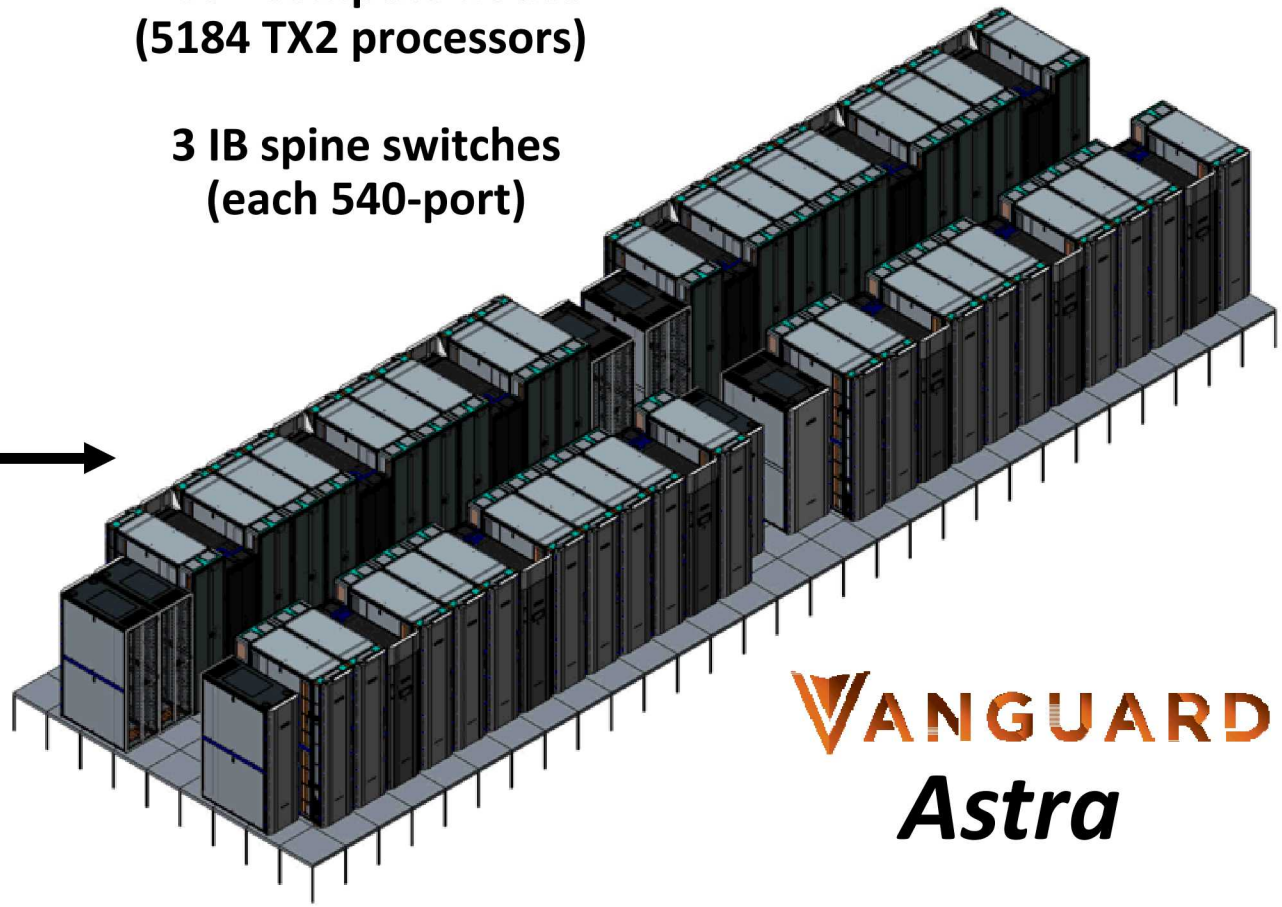
3 IB switches/rack
(one 36-port switch per 6 chassis)

HPE Apollo 70 Rack

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

2592 compute nodes
(5184 TX2 processors)

3 IB spine switches
(each 540-port)



VANGUARD
Astra

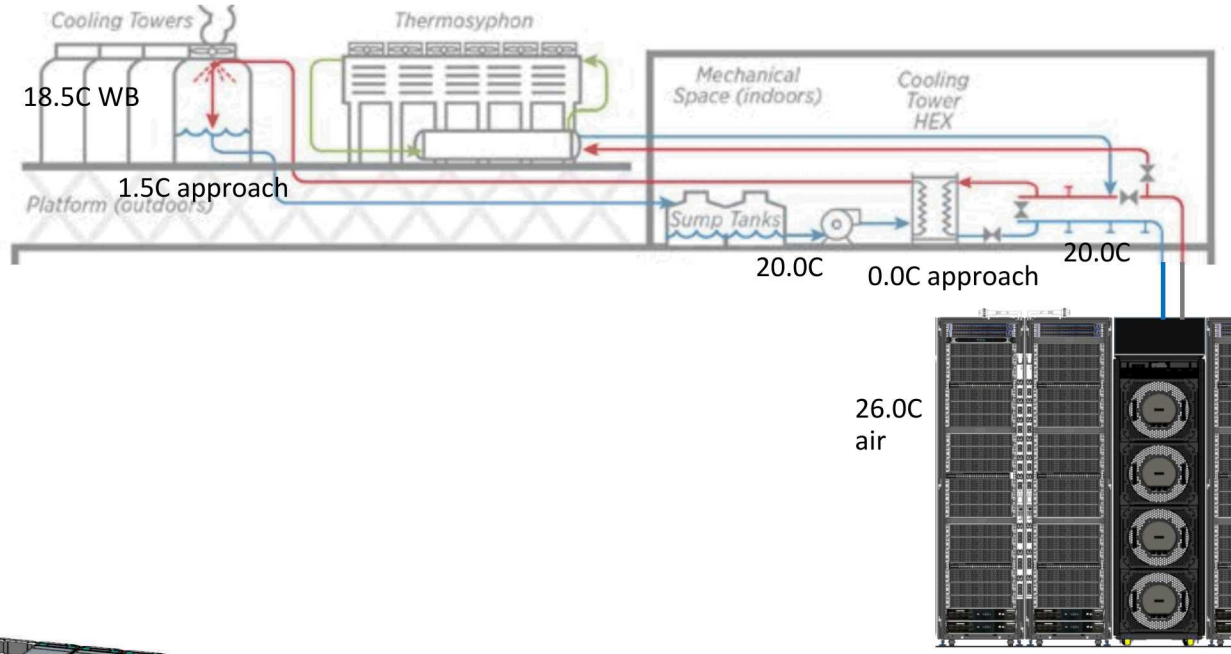
Astra Advanced Power and Cooling

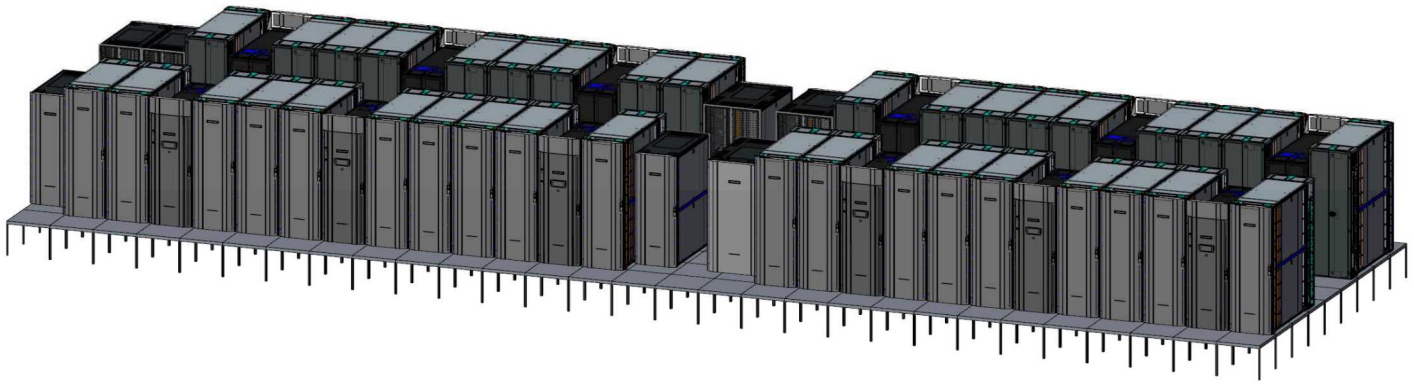
Extreme Efficiency:

- 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

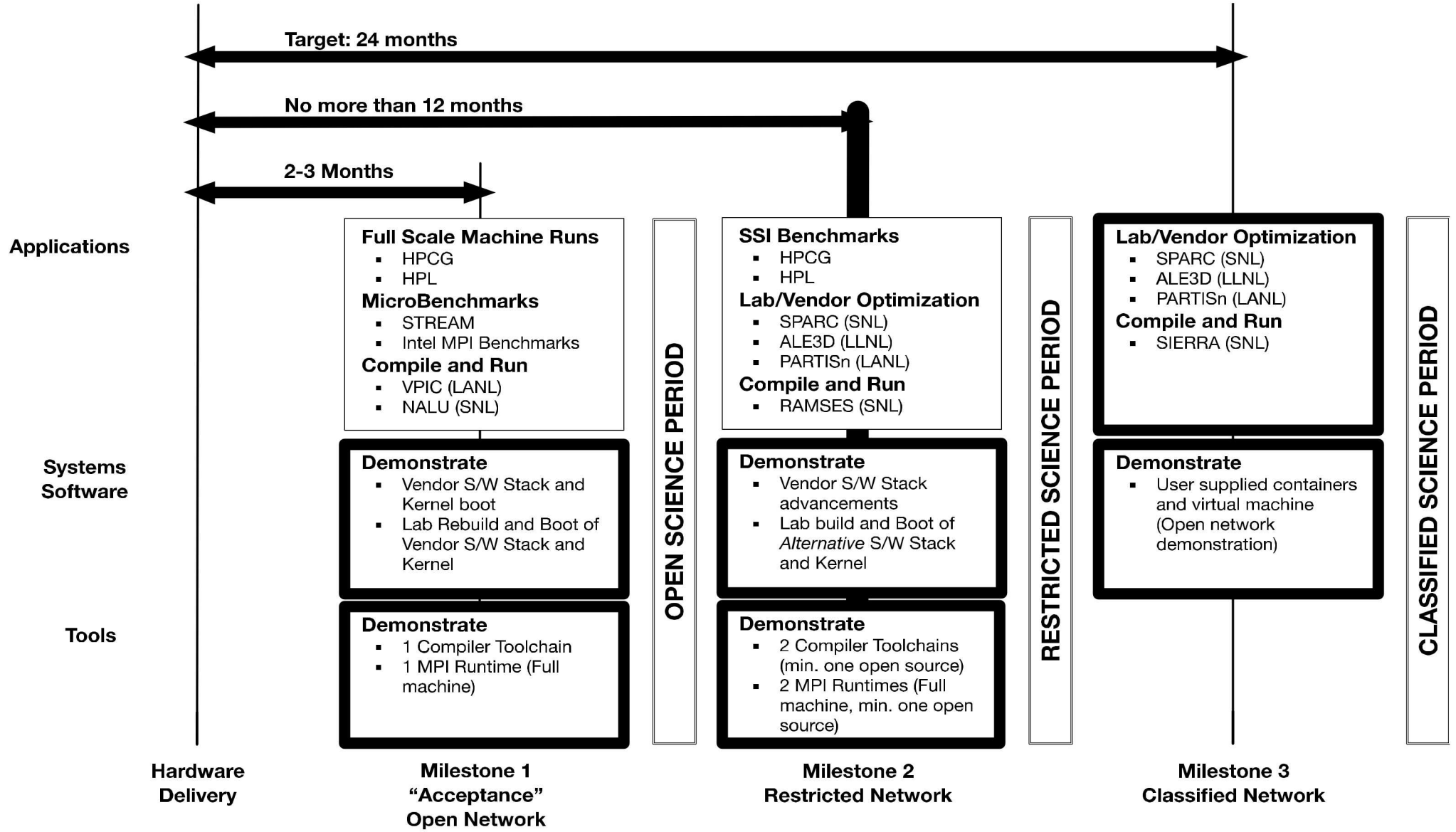
Efficient tower and HEX can take hottest 36 hours of the year of 18.5C wetbulb to make 20C water to the fan coils



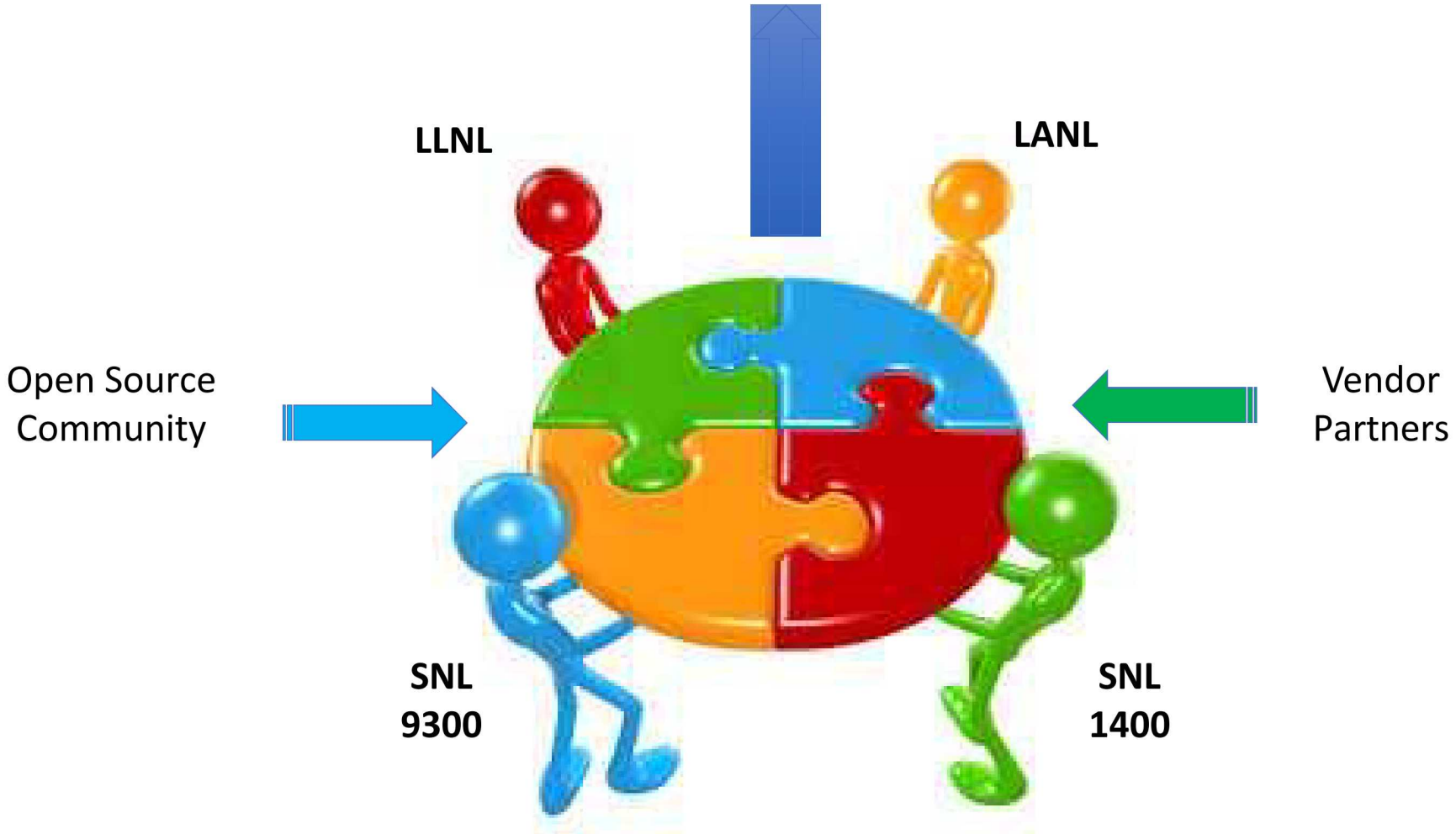


Projected power of the system by component										
	per constituent rack type (W)					total (kW)				
	wall	peak	nominal (linpack)	idle		racks	wall	peak	nominal (linpack)	idle
Node racks	39888	35993	33805	6761		36	1436.0	1295.8	1217.0	243.4
MCS300	10500	7400	7400	170		12	126.0	88.8	88.8	2.0
Network	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

Astra Acceptance and Milestones



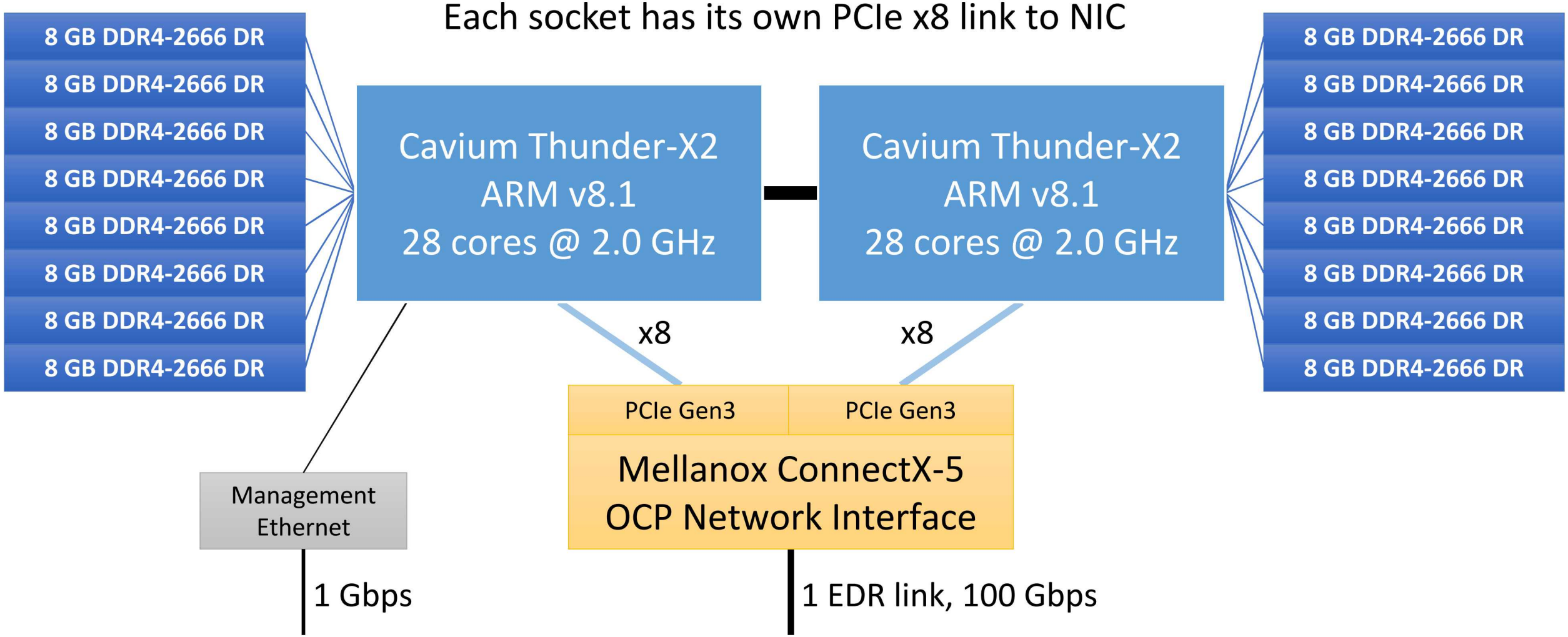
Vanguard Collaboration



Backup

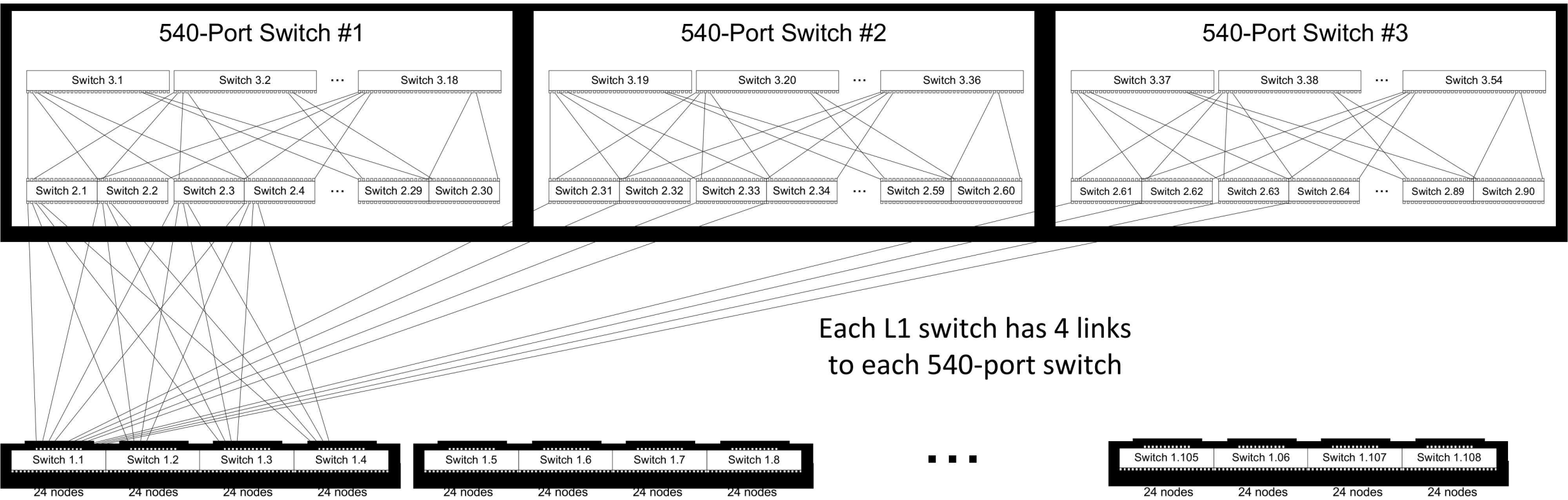
Vanguard-Astra Compute Node

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



Network Topology Visualization

Mellanox Switch-IB2 EDR, Radix 36 switches, 3 level fat tree, 2:1 taper at L1, SHARP

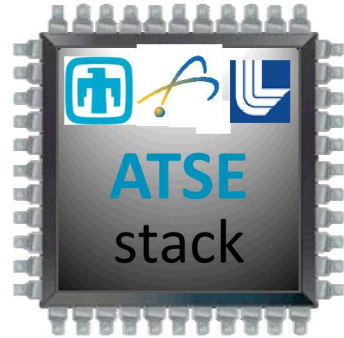


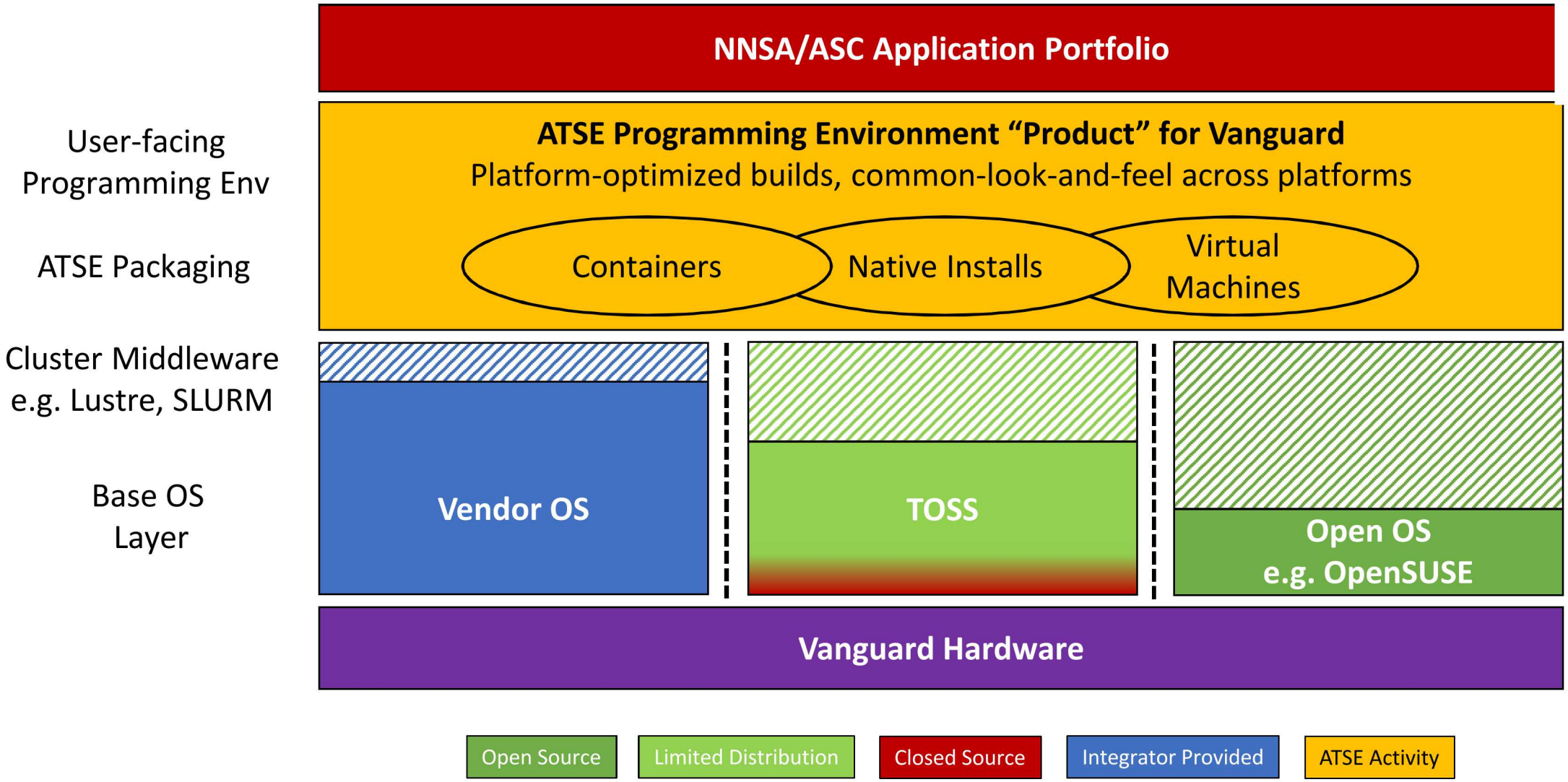
108 L1 switches * 24 nodes/switch = 2592 compute nodes

Vanguard-Astra Infrastructure

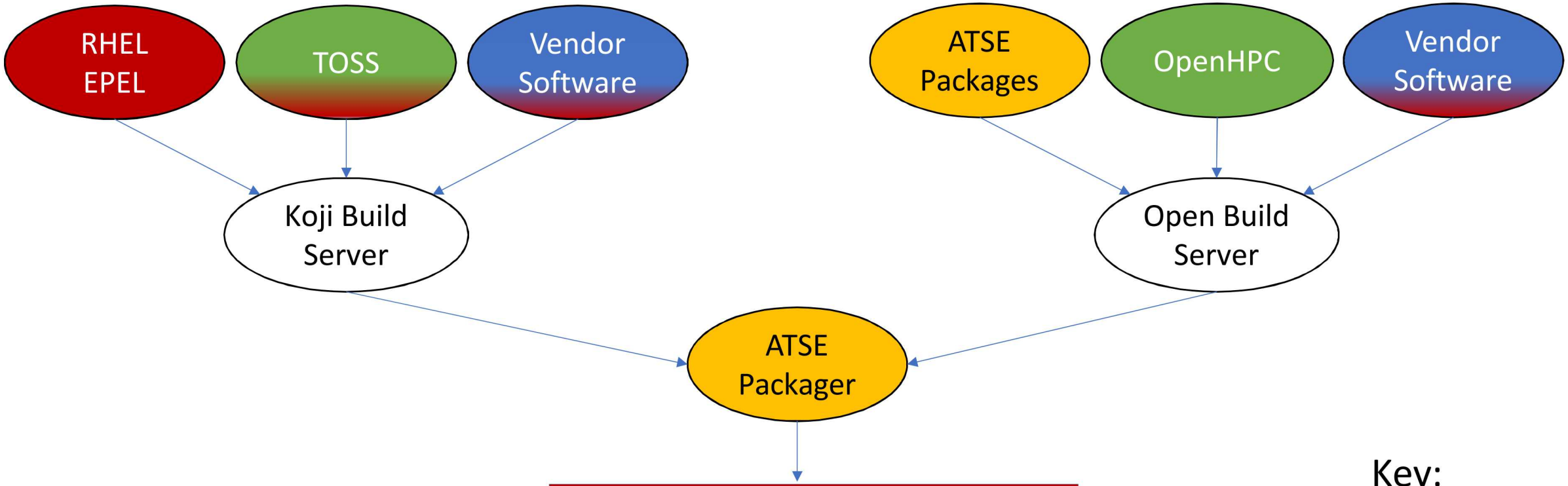
Login & Service Nodes	Four login/compilation nodes Three Lustre routers to connect to external Sandia filesystem(s) Two general service nodes
Interconnect	EDR InfiniBand in fat tree topology 2:1 oversubscribed for compute nodes 1:1 full bandwidth for in-platform Lustre storage
System Management	Dual HA management nodes running HPE Performance Software – Cluster Manager (HPCM) Ethernet management network, connects to all nodes One boot server per scalable unit (288 nodes)
In-platform Storage	All-flash Lustre storage system 403 TB usable capacity 244 GB/s throughput

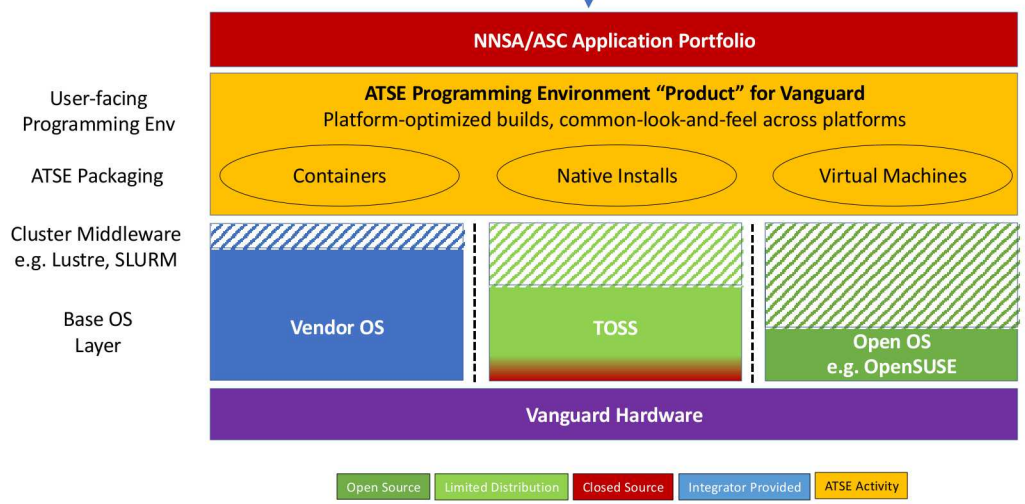
Vanguard-Astra Software Stack





Integrate Components from Many Sources





Key:

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

Close Collaboration with HPE Open Leadership Software Stack (OLSS) Effort

Hewlett Packard Enterprise

- 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