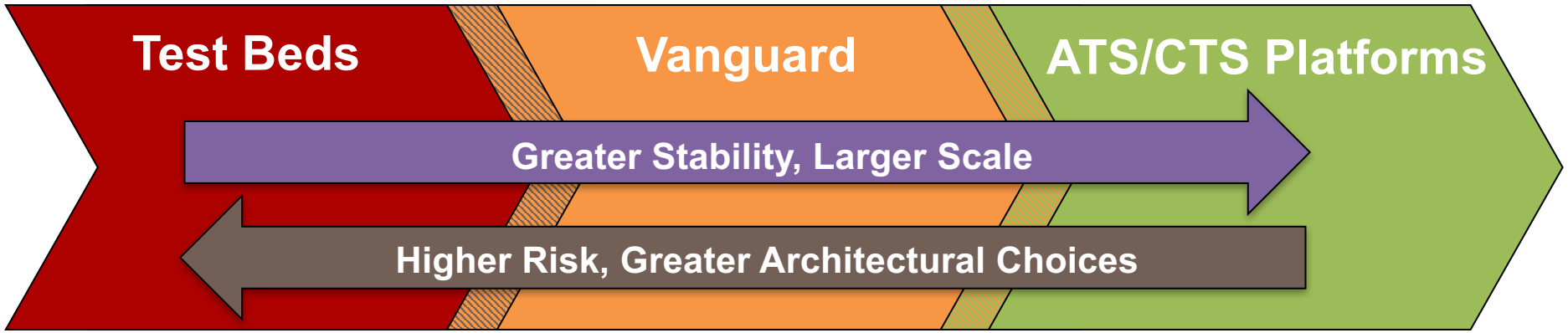
# I/O Research on Astra, the World's Largest ARM Supercomputer


*Presented by*

Matthew L. Curry, Ph.D.
Center for Computing Research
Sandia National Laboratories

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-NA-0003525.

# The Vanguard Program -- Where Astra Sits




- Spectrum of machines
  - Balance scale, breadth, user expectations
- Astra: The first machine in Sandia's Vanguard Program
  - Series of large-scale test platforms
  - Vet interesting technologies for "prime time"
- Many "firsts"/early showings for Sandia and HPE
  - First petascale aarch64 machine
  - First deployment of HPCM
  - First machine with HPE's Apollo 70 chassis

# Machine Architecture

**HPE Apollo 70 Chassis: 4 nodes**

**18 chassis/rack**

**72 nodes/rack**

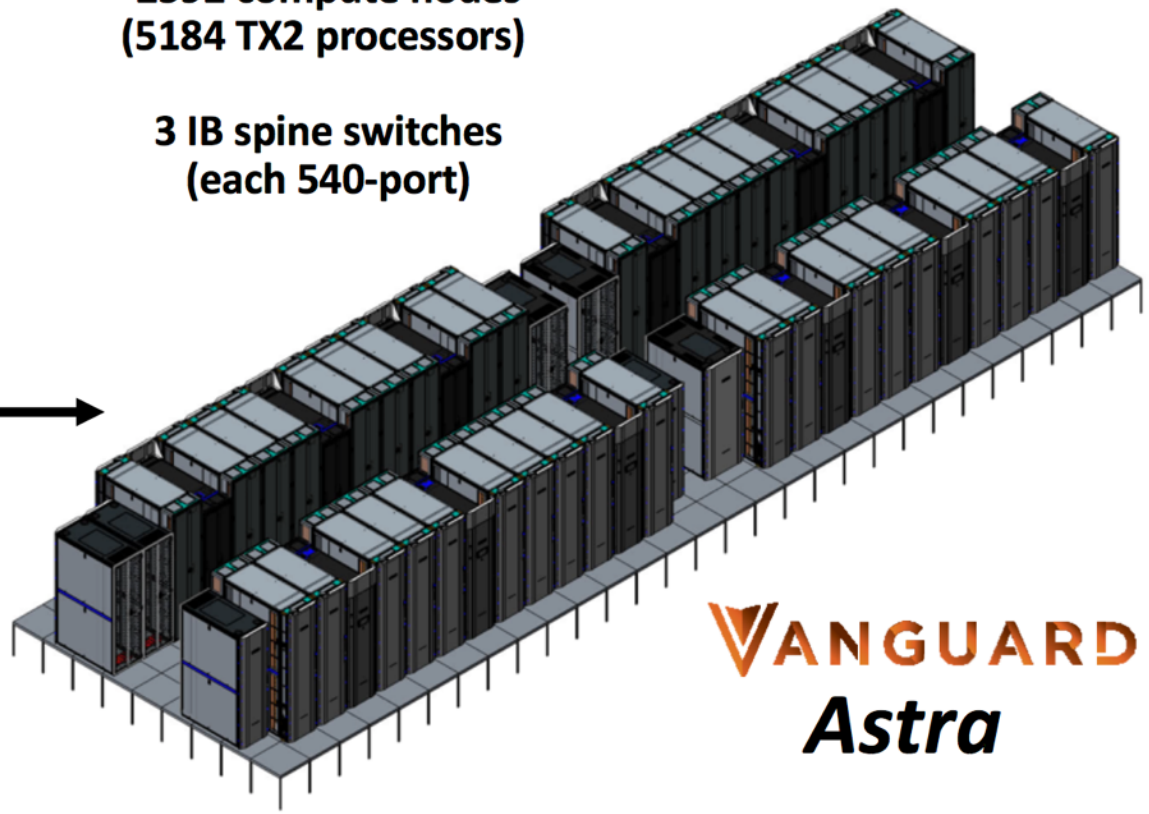
**3 IB switches/rack**
**(one 36-port switch per 6 chassis)**

**HPE Apollo 70 Rack**

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

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

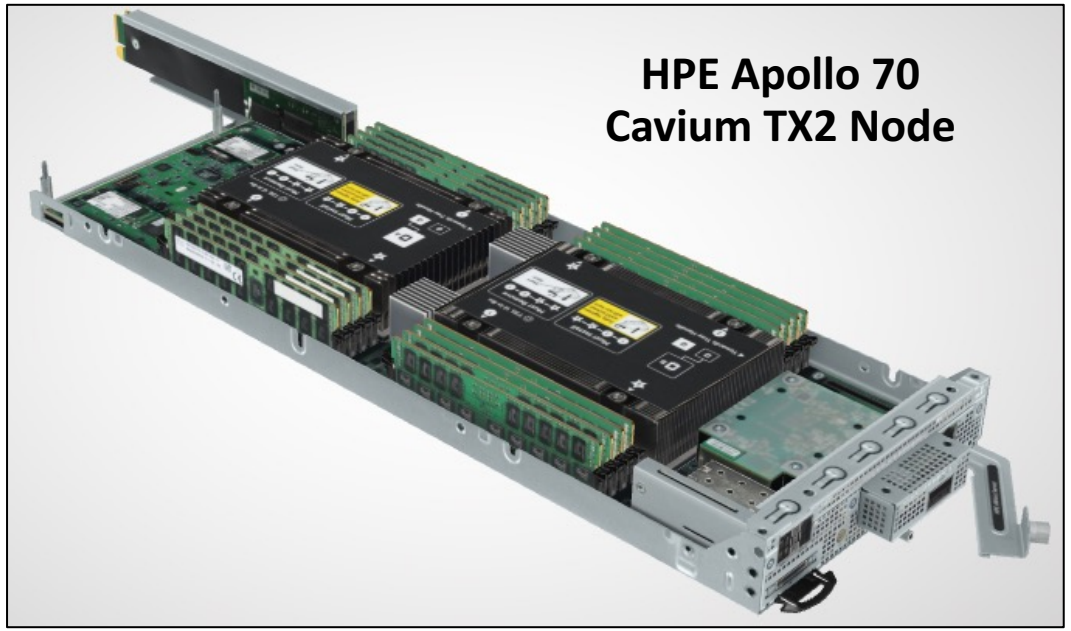
**3 IB spine switches**
**(each 540-port)**

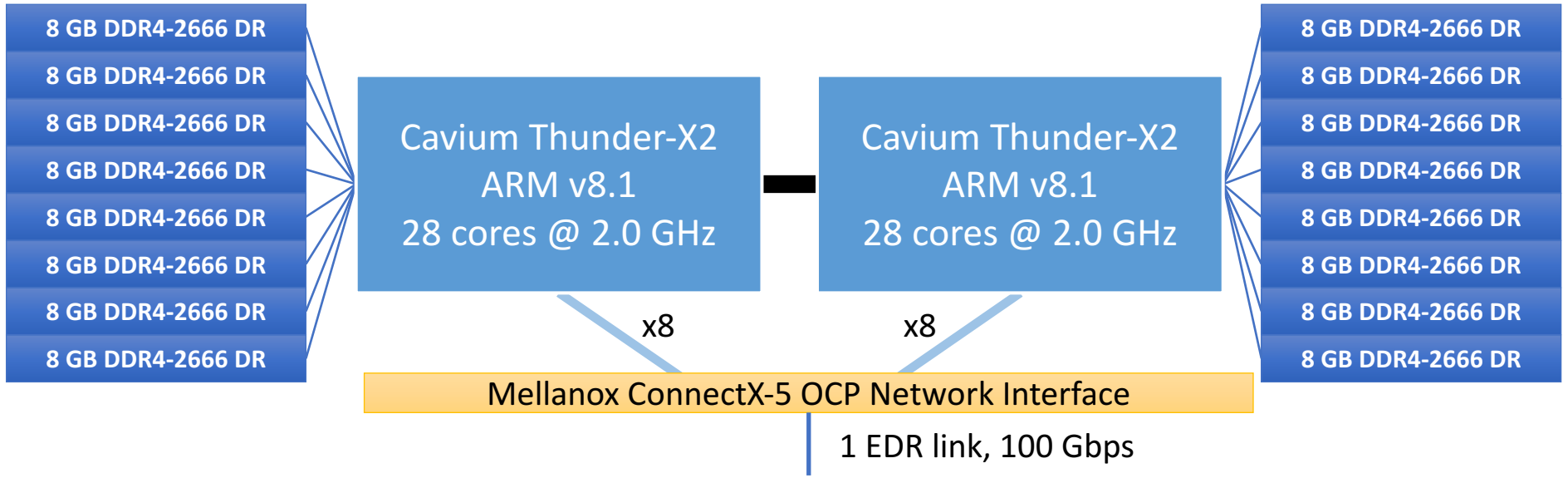


VANGUARD
Astra

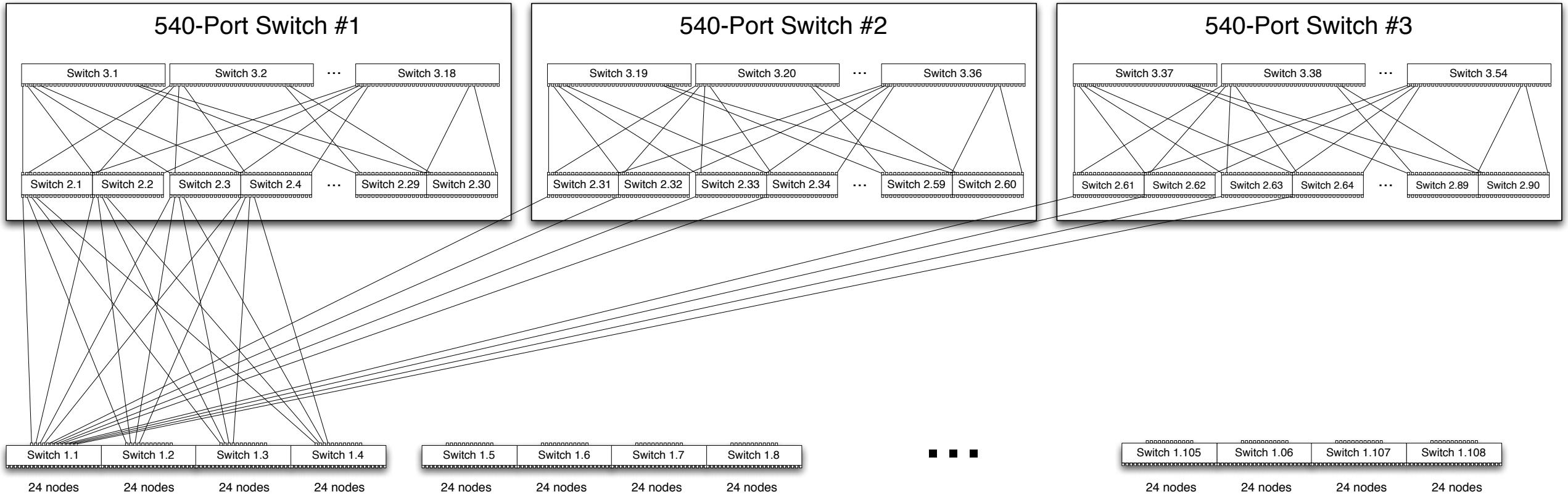
# Node Architecture

- **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 (100 Gb)



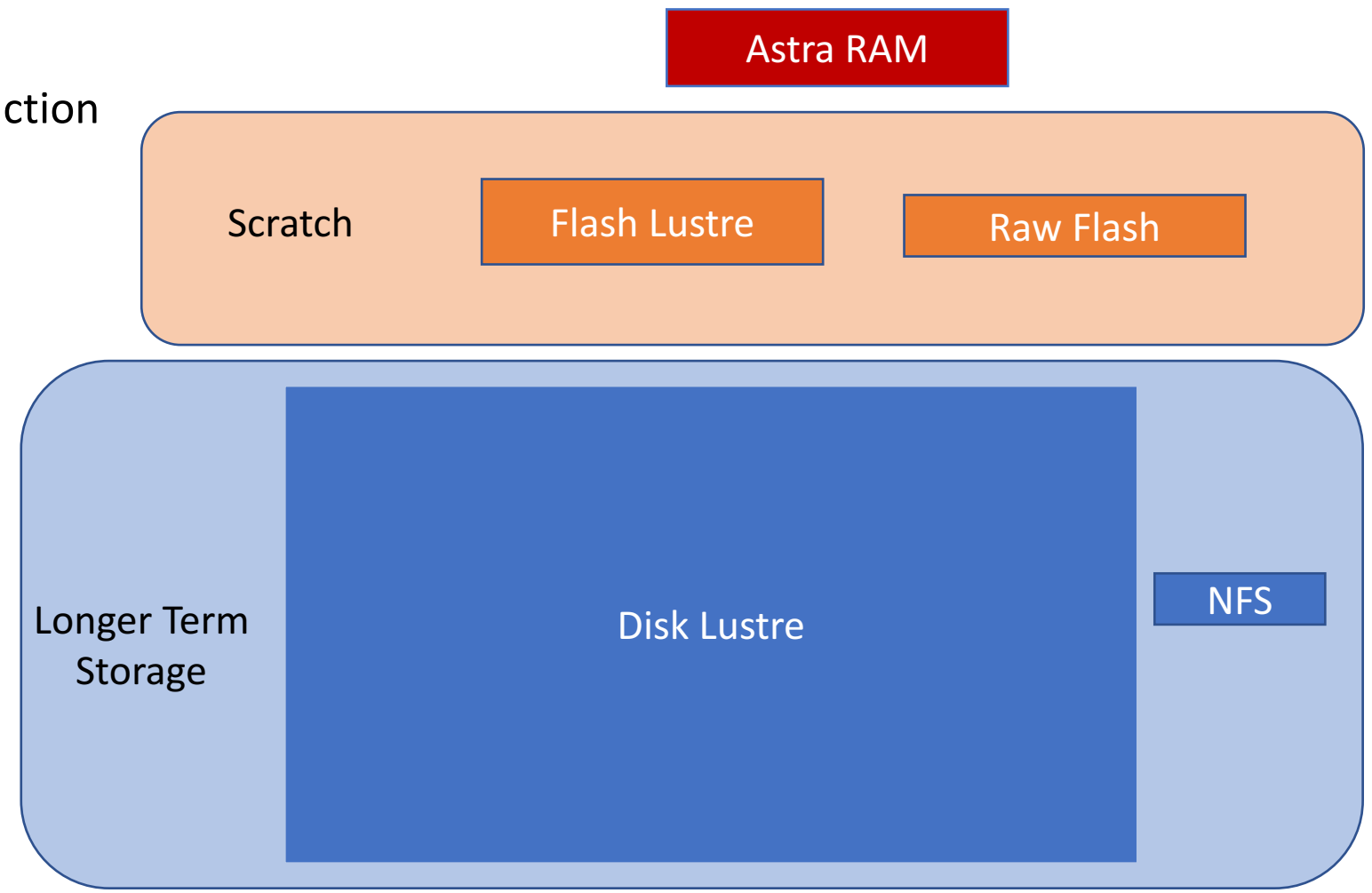
# Network Topology




- Mellanox Switch-IB2 EDR, Radix 36 switches, 3 level fat tree, 2:1 taper at L1 for compute nodes, SHArP
- L1 switches have 4 uplinks to each core switch
- 108 L1 switches * 24 nodes/switch = 2592 compute nodes

# Astra's Storage

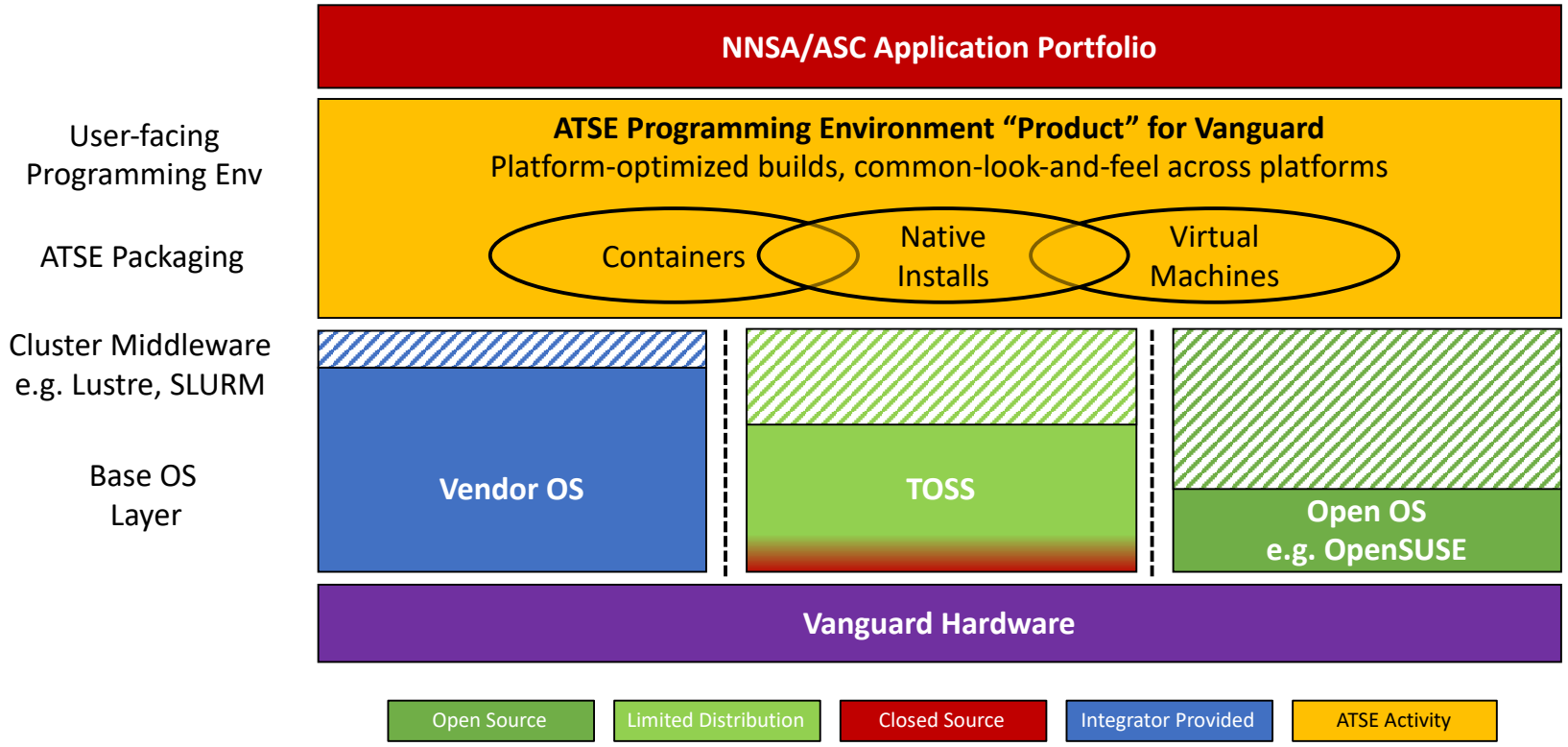
- Flash-based Lustre (403 TB @ ~240 GB/s)
  - On-machine
  - NVMe
  - 40xOSS, 2xMDS, ZFS
  - Imminent resize expected for production
- Raw flash (288 TB @ ~360 GB/s)
  - Experimental
  - In compute section
  - 72 nodes x 2 slots
- Disk-based Lustre (5.6 PB @ ~30 GB/s)
  - Off-machine, enterprise model
  - Three LNET routers
- Projects and home (120 TB @ ~1GB/s)
  - NFS over 10 GigE



# I/O R&D

- Alternative File System Testing
  - BeeGFS underway, others as time permits
  - Hopes for leaving file systems available for production
- Lustre over NVMe Evaluation
  - Feature is relatively new – Compare and contrast with others, as well as theoretical expectation
  - Requires some I/O characterization
  - Fitness for purpose – Burst Buffer
- Support software
  - Migration between tiers is still proprietary/product-specific
    - Burst buffers (Datawarp, IME)
    - HSM
    - BeeOND, LOD
  - Node-local storage is still uncommon/new in DOE machines
    - Subset model is unique

# ATSE – Advanced Technology Software Environment




- A Test Vehicle for System Software
- R&D task of figuring out what works and what doesn't on ARM
  - Surprising and interesting results already from non-I/O and I/O-adjacent software
- Platform for introducing interesting software stack modifications
  - Infrastructure is huge, but now we have it

# Other Projects

- Application porting and measurement
  - Some applications do not yet leverage burst tiers effectively
- Continuing focus on production readiness
  - End of open science period by May
  - Production environment soon after
  - Tension with possible research
- Analytical evaluation of resilience
  - Leverage understanding of how users will tend to use flash tier
- Workflows leveraging containers and virtual machines

# Conclusions

- Astra, and future Vanguard machines, are an important resource for I/O research
  - A historically disruptive research subject
  - Semi-production with friendly users
  - Touchpoint for deep industry/academic interaction
    - Yes, even foreign collaborators (during particular periods)
- Astra's research is skewed more toward production
  - Future machines may include room for more basic research


*Exceptional Service in the National Interest*

mlcurry@sandia.gov