El Capitan: The First NNSA Exascale System

Bronis R. de Supinski
Chief Technology Officer for Livermore Computing

August 23, 2023

LLNL-PRES-849945

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

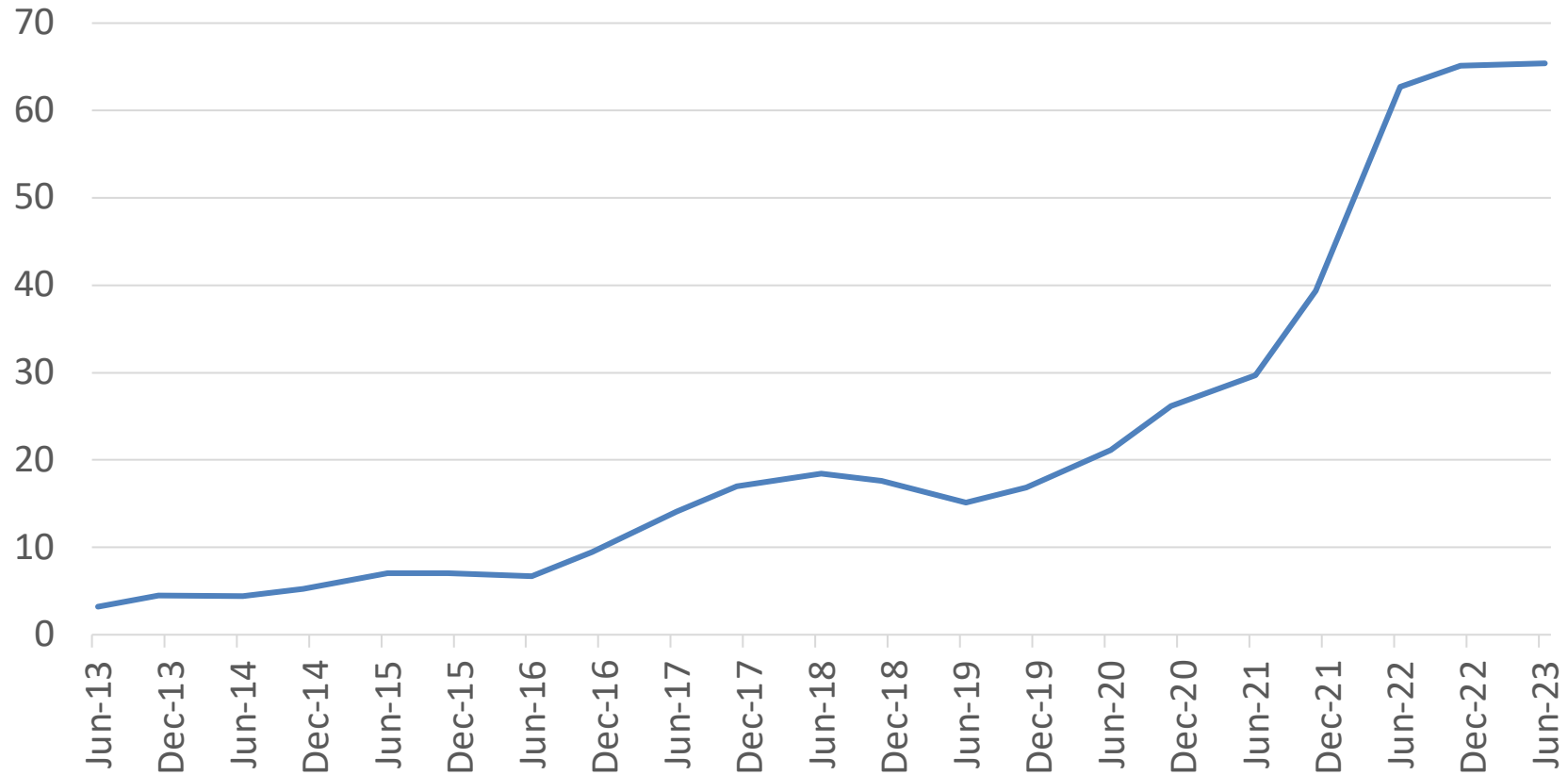
Lawrence Livermore National Laboratory

“Energy Efficiency” of large-scale systems has improved significantly over the last decade

EL CAPITAN
ADVANCED SIMULATION & COMPUTING

Green 500 Number 1

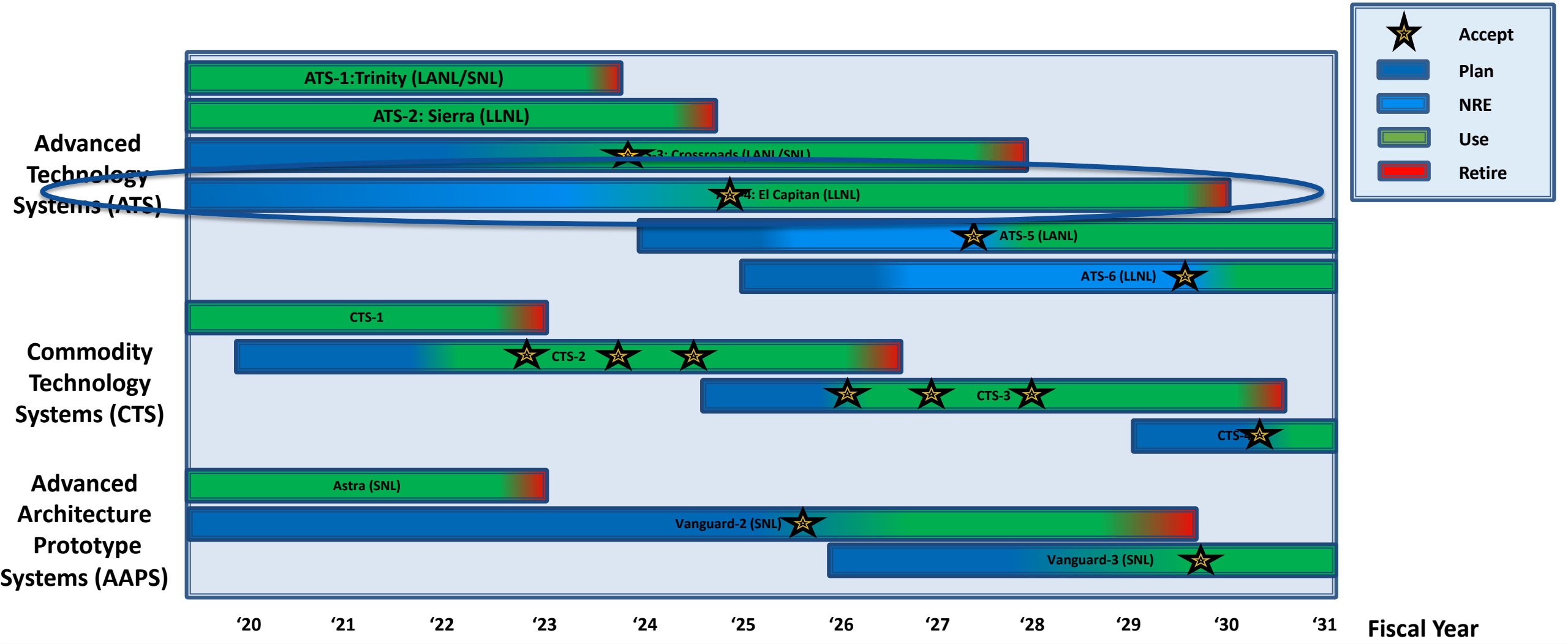
Gflops/W



Questions to ponder related to energy consumption and energy efficiency of large-scale systems

- What can be done to improve energy efficiency?
  - Where has academic research focused?
  - What has led to improvements historically?
  - Will those historic improvements continue?
- How should we measure energy efficiency?
  - Should we use Gflops/W?
    - Green500 uses HPL performance divided by energy to achieve that performance
  - What metrics apply to other types of systems?
  - What are the shortcomings of current approaches?
- How can we motivate improvements in energy efficiency?
  - What motivates users?
  - What motivates large-scale centers (i.e., system providers)?
- Would improvements in energy efficiency actually address community concerns?

El Capitan, the next ASC ATS to be sited at LLNL will be the first NNSA exascale system



ASC Platform Strategy includes application code transition for all platforms

HPC exascale infrastructure demands are unprecedented and drove utility upgrades serving HPC at LLNL

Exascale Computing Facility Modernization (ECFM) Project

Status: Project is complete

Construction timeline: 3/2020 to 5/2022

Objective: enable LLNL to operate two exascale class systems simultaneously

ECFM Highlights

- No structural upgrades needed
  - Existing facility had ample square footage with 48,000 SF and structural integrity up to 625 lbs/SF
- Cooling scaled to 28,000 tons with new 18,000 ton cooling tower
  - Loop extended avoiding chillers
- Electrical supply upgraded from 45 MW to 85 MW
  - Capacity = 1771 Watts/SF
  - Dynamic monitoring and control systems to ensure seamless 24/7 HPC operations
  - Two electric utilities tied in parallel at 115kV

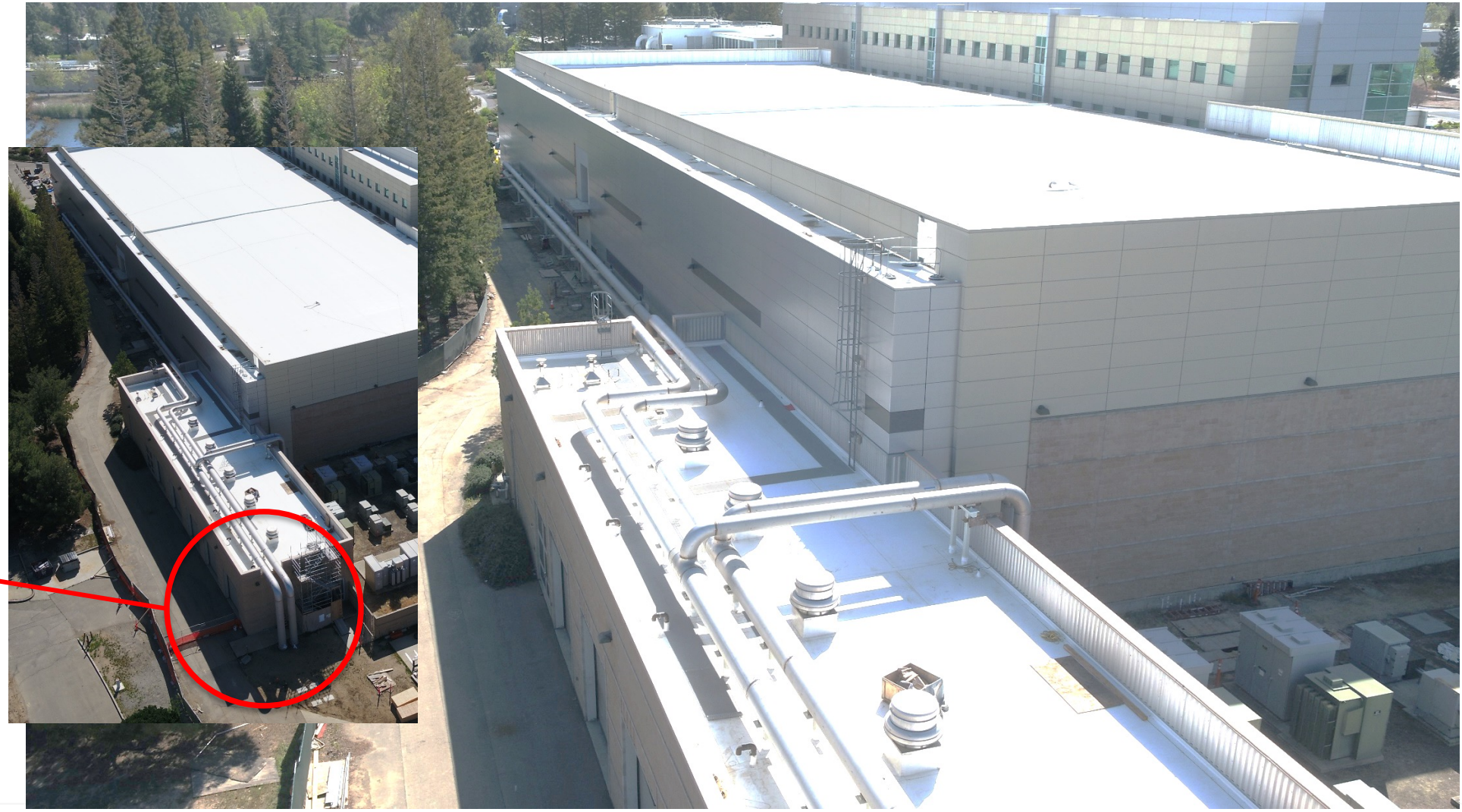
Exascale Computing Facility Modernization Project: 3/2020 – 5/2022

40MVA 115kV-13.8kV transformers provided for redundancy with tap changers to 60MVA

EL CAPITAN
ADVANCED SIMULATION & COMPUTING

ECFM team illustrates scale of 18,000 ton cooling towers

30” cooling loop extends from cooling towers into facility



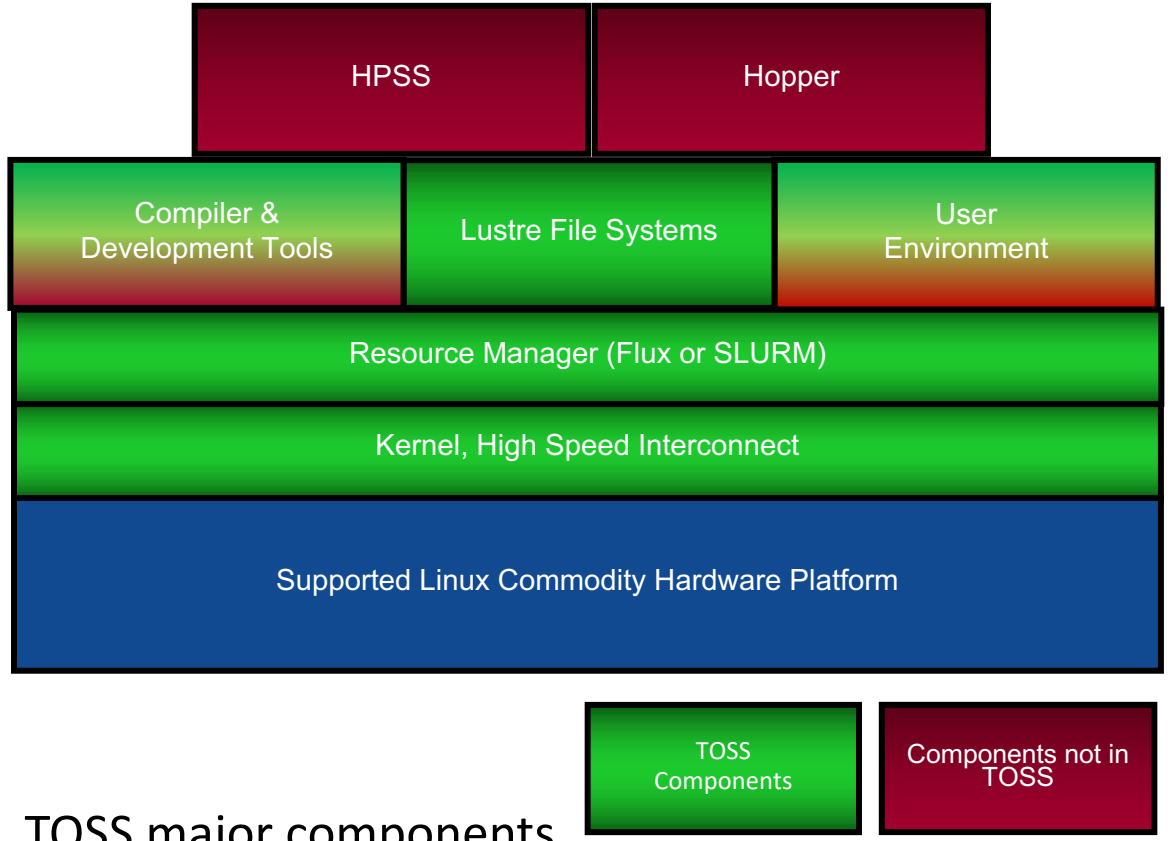
Cooling loop extends to (3) filter stations in facility to support future HPC systems as well as El Capitan

HPE (nee Cray) will deliver a highly capable AMD GPU-accelerated system

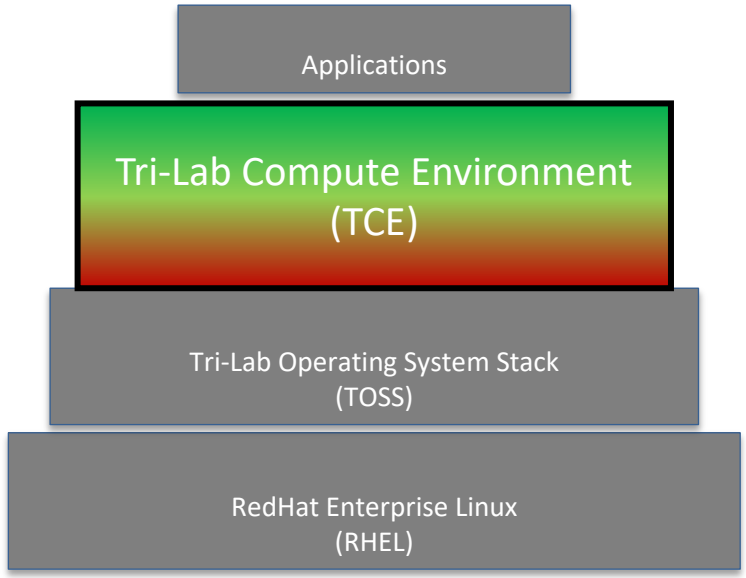
- El Capitan will meet its stockpile stewardship simulation mission
- System will feature:
  - Peak ≥ 2.0 DP exaflops
  - Peak power < 40 MW
    - Anticipating ~30MW
  - AMD MI300 APU - 3D chiplet design w/AMD CDNA 3 GPU, "Zen 4" CPU, cache memory and HBM chiplets
  - Slingshot interconnect
- HPE will provide several critical innovations
  - HPE and LLNL have worked with ORNL jointly on non-recurring engineering (NRE) activities
  - Will use TOSS software stack, enhanced with HPE software
  - El Capitan will include an innovative near node local storage solution

Late binding of the processor solution has ensured El Capitan provides the best possible value

El Capitan will be the first ATS to use TOSS and TCE in production



- TOSS major components
  - The OS – LLNL's Linux distribution based on RHEL
  - Resource Manager (SLURM or Flux)
  - Lustre



- The Tri-Lab Compute Environment (TCE) is an application development environment (DE)
  - Compilers (Intel, PGI, GNU, …)
  - MPI (MVAPICH, OpenMPI, …)
  - Debuggers (TotalView, Allinea)
  - Performance Tools

TOSS is a critical component of ASC's commodity Linux cluster strategy

What is it?

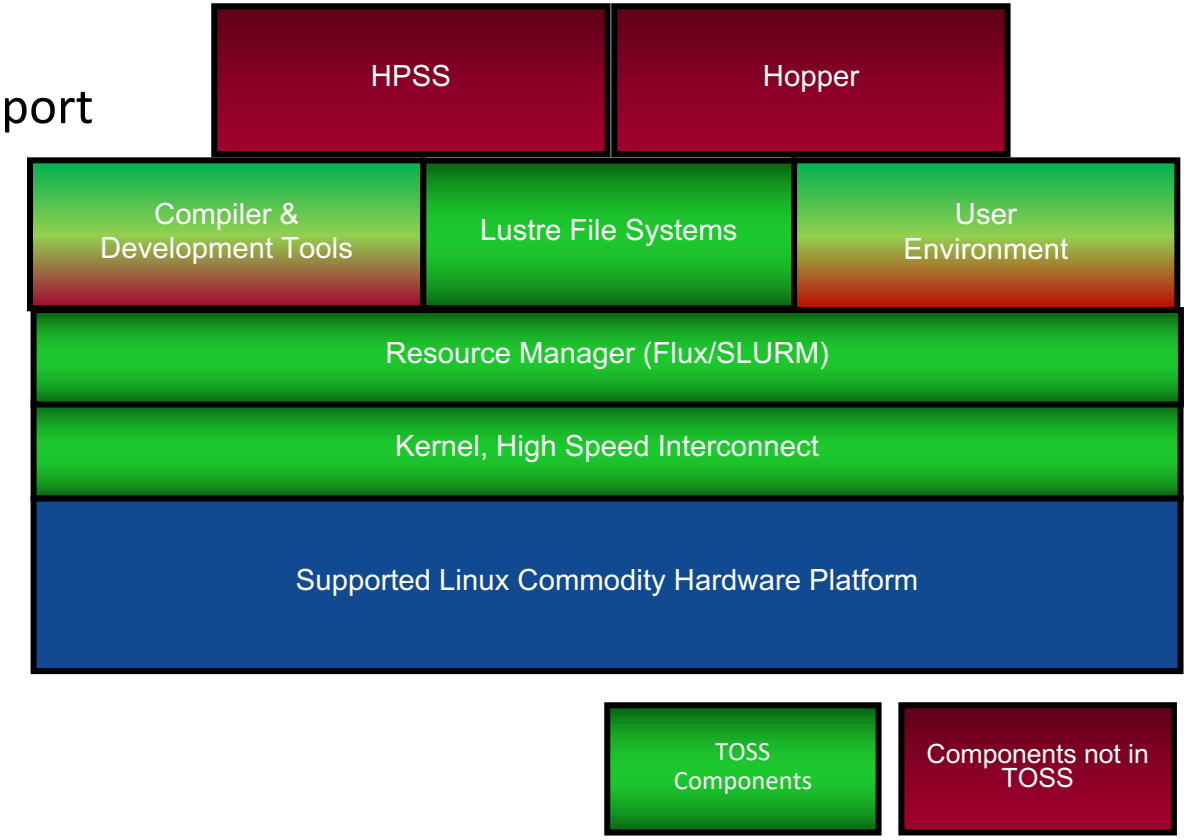
- A common operating system and computing environment for Tri-Lab Linux clusters
- A software stack for interconnected HPC clusters
- A methodology for building, quality assurance, integration, and configuration management

Why do it?

- Reduce total cost of ownership and enable application portability
- Consistent source and software across architectures: X86, PowerPC, and ARM
- Install same software on all commodity hardware at the Tri-Labs

TOSS was designed as a software stack for HPC – large, interconnected clusters – but it will also run on desktops

- Major components
  - The OS – LLNL's Linux distribution based on RHEL
  - Leverages Red Hat's extensive QA testing and support
  - Includes Tri-Lab developed additions to support large-scale commodity clusters
  - Resource Manager (Flux, has been SLURM)
  - LDMS (monitoring)
  - Lustre
- Support infrastructure
  - Secure software build system
  - Bug tracking system
  - Software repository



TOSS adds system management tools, Lustre, user tools, hardware drivers, and more

Cluster Management Tools

- Pdsh – parallel remote shell
- Powerman – remote power management
- Conman – remote console management
- FreeIPMI – out-of-band systems management
- MUNGE – scalable authentication
- OMS/SMT – Infiniband diagnostics
- Whatsup – node up/down detection
- Genders – cluster configuration database
- Flux and SLURM – job scheduling
- Mrsh – remote shell with munge authentication
- Netroot – diskless boot support
- LDMS – lightweight runtime collection of high fidelity data

User Tools

- Compilers (PGI, Intel, GCC, clang)
- Debuggers (Totalview, Allinea)
- MPI libraries (MVAPICH, OpenMPI)
- I/O libraries (NetCDF, HDF5)
- Visualization & Graphics (Paraview, VisIt, mplayer, vlc)

Kernel Modules and Patches

- Lustre & ZFS
- Nvidia
- Network drivers (i40e, ixgbe)
- MSR-safe
- NFS support for > 16 groups
- Assorted bug fixes and enhancements

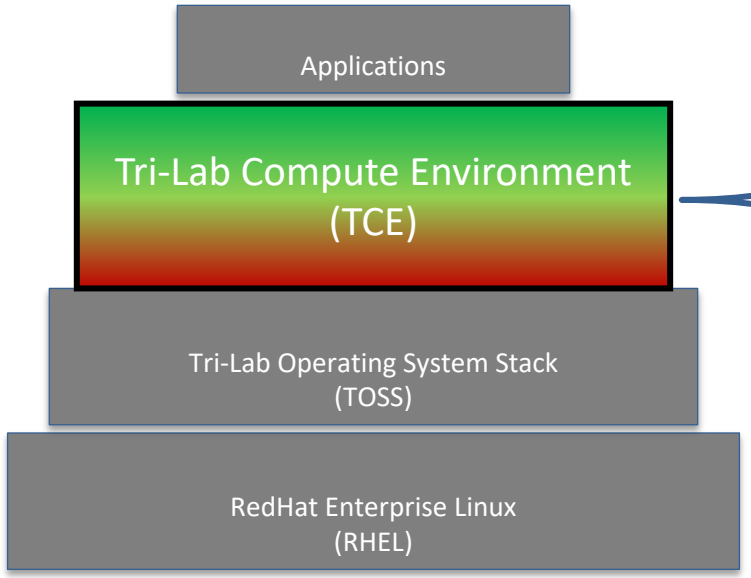
We use as much stock RHEL and EPEL software as we can

Creating TOSS (Intel/AMD example)

We down-select from over 40,000 RHEL Packages

TOSS supports Intel, AMD, ARM, PowerPC, accelerators, multiple interconnects, virtual environments, infrastructure systems, and Lustre file system servers

TCE: Tri-Lab Compute Environment



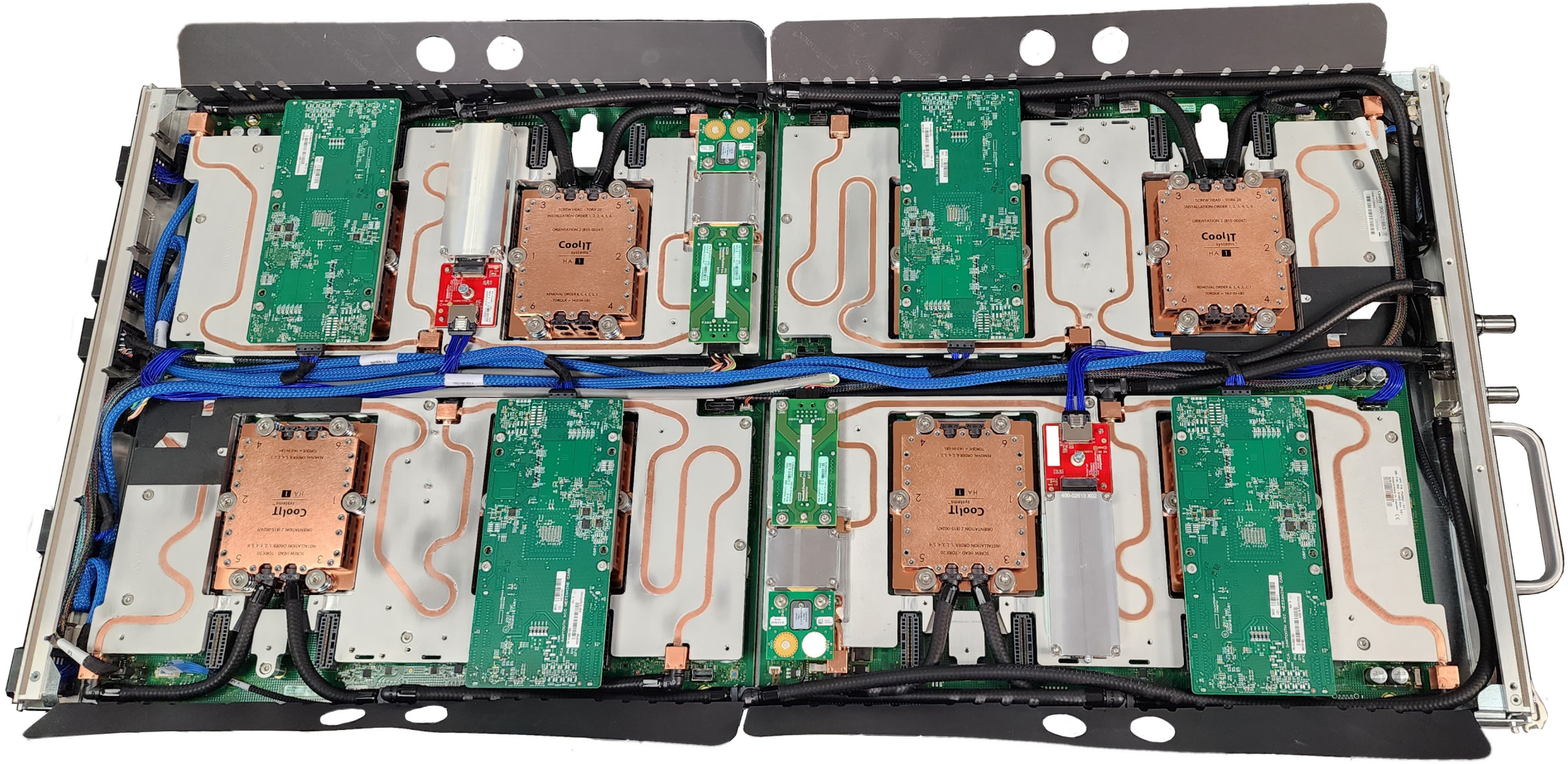
- TCE is an app development environment (DE)
  - Compilers (Intel, PGI, GNU, …)
  - MPI (MVAPICH, OpenMPI, …)
  - Debuggers (TotalView, Allinea)
  - Performance Tools
- Focuses on smooth user experience
  - Compiler wrappers for easy building
  - rpath avoids common user bugs
  - Libraries are integrated and designed to work together
- Expands on TOSS3's more limited DE
  - Quicker package update frequency
  - Adds many nice-to-have packages
  - Customizable for each cluster
  - Will include Cray Programming Environment, AMD compiler suite

TCE is the tools and development environment layer in the software ecosystem

MI300A at AMD: The AMD Austin Lab

AMD's commercially developed Instinct™ APU codenamed "MI300A"

MI300A at HPE: The HPE compute blade



AMD INSTINCT™ MI300: The world's first data center APU

- 4th Gen AMD Infinity Architecture:
  AMD CDNA™ 3 and EPYC™ CPU "Zen 4" together
  - CPU and GPU cores share a unified on-package pool of memory
- Groundbreaking 3D packaging
  - CPU | GPU | Cache | HBM
  - 24 Zen4 cores, 146B transistors, 128GB HBM3
- Designed for leadership memory bandwidth and application latency
- APU architecture designed for power savings
  - compared to discrete implementation

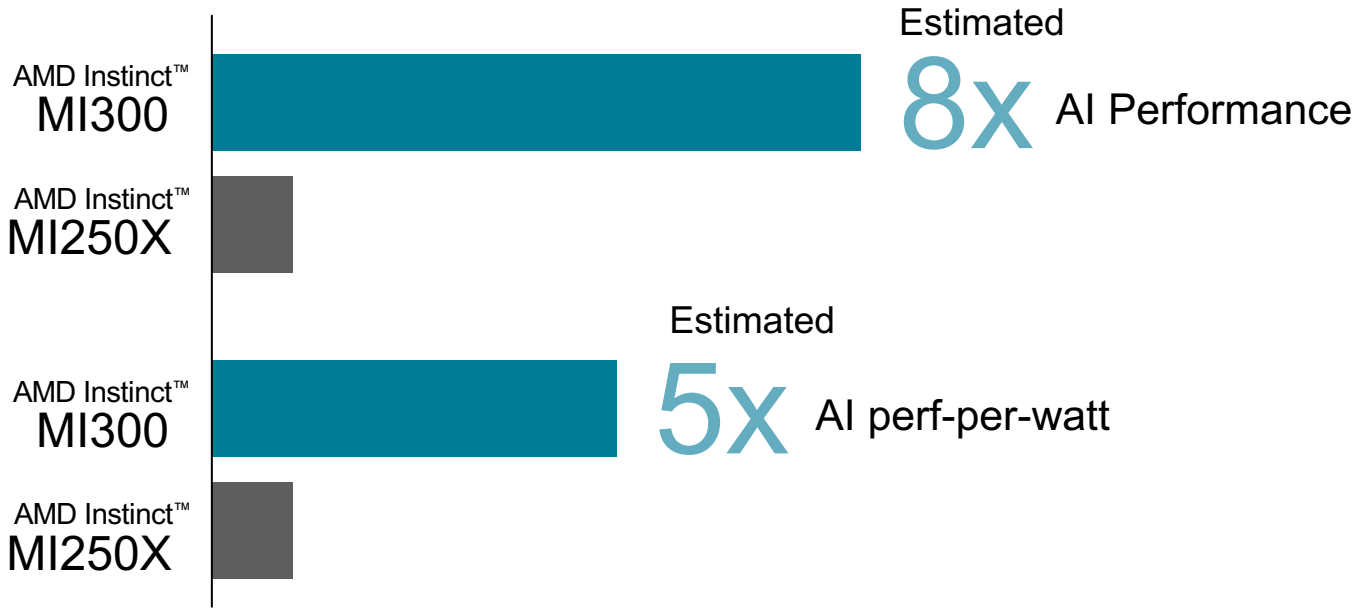
> 8X Expected AI Training Performance vs. MI250X

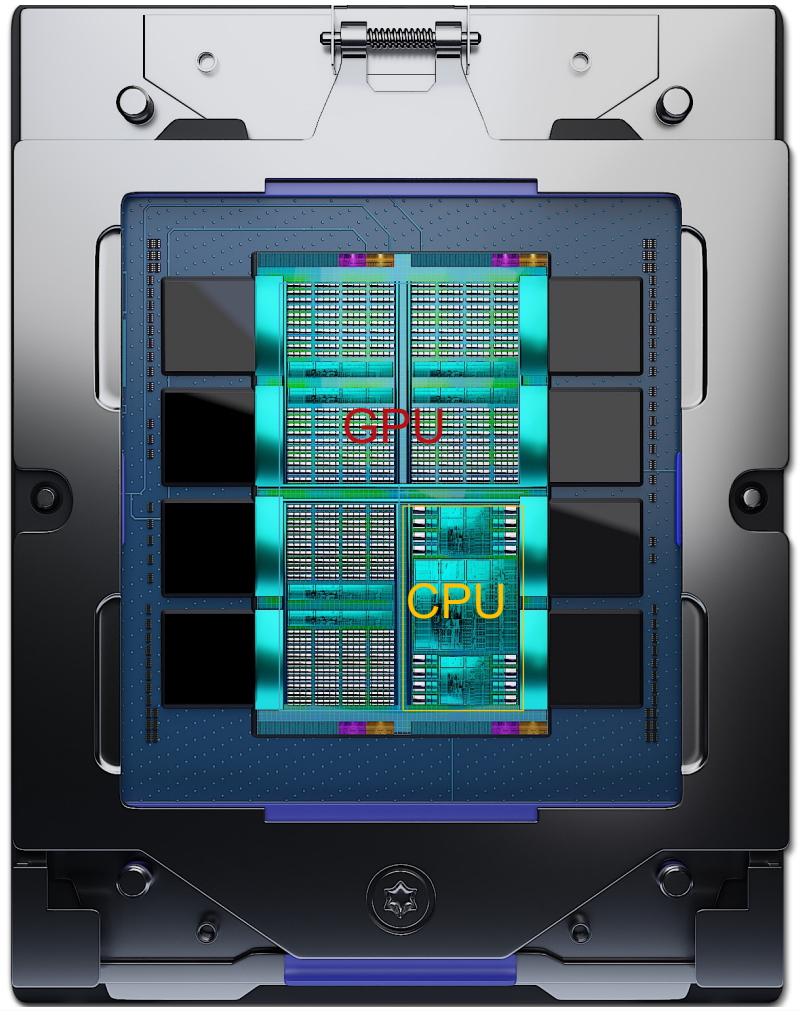
Preliminary data and projections, subject to change

Available 2023

MI300: Architectural innovation at the next level

- 5nm process technology with 3D stacking
- Next-gen Infinity Cache™ and 4th Gen Infinity Fabric base die
- New Math formats
- Unified memory APU architecture





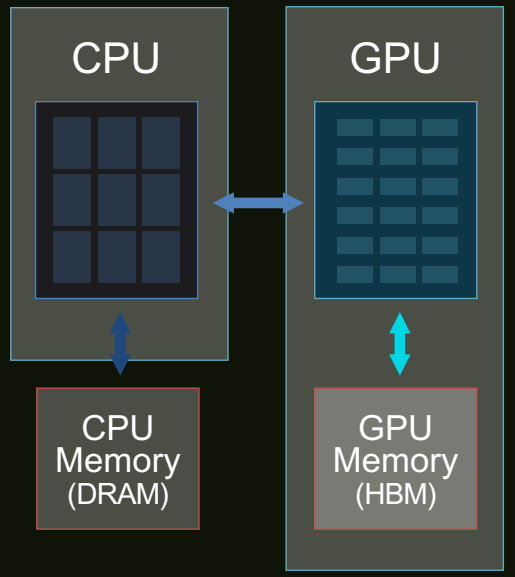
3D CPU+GPU integration for next-level efficiency

EL CAPITAN
ADVANCED SIMULATION & COMPUTING

AMD CDNA™ 2 Coherent Memory Architecture → **AMD CDNA™ 3 Unified Memory APU Architecture**

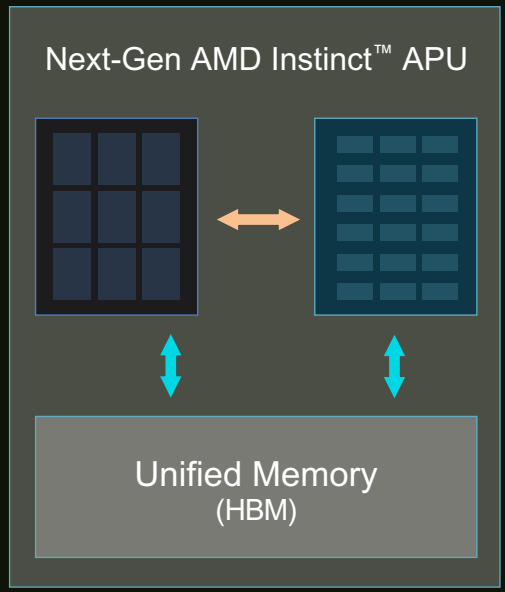
AMD Instinct™ MI250 Accelerator

- Simplifies programming
- Low overhead 3rd Gen Infinity interconnect
- Industry standard modular design



AMD Instinct™ MI300 Accelerator

- Eliminates redundant memory copies
- High bandwidth, low latency communication
- Low TCO with unified memory APU package



Hewlett Packard Enterprise

Rabbit Program

Near Node Local Storage

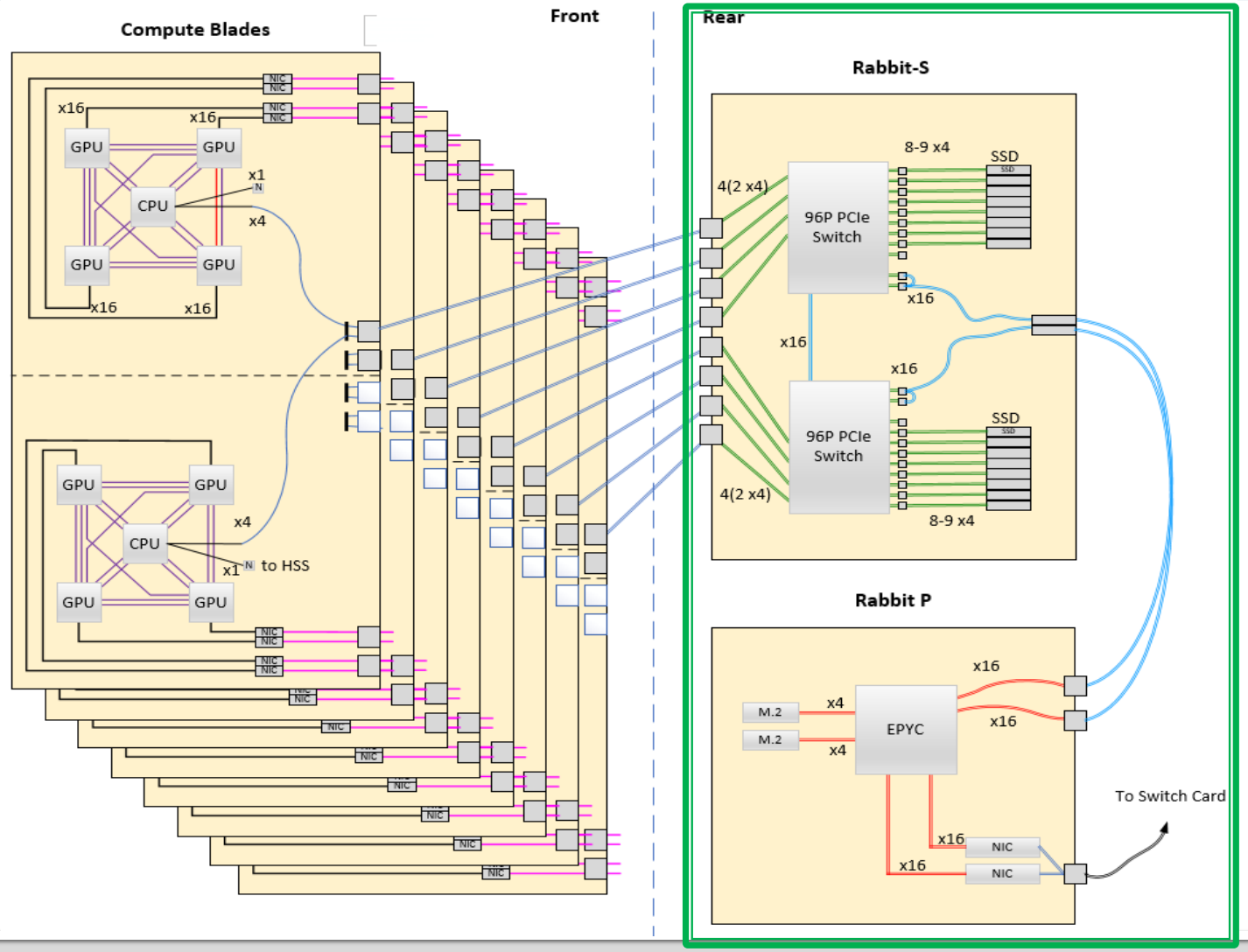
Near-node local storage was a key aspect in El Capitan selection

- El Capitan will deploy one Rabbit module for every compute chassis
- Rabbit modules will:
  - Reduce system interference
  - Enable efficient defensive I/O
  - Likely serve as OS file cache
  - Possibly support more efficient input
    - Particularly for ML training
    - Stage-in of restart files is more complex
- Rabbit modules are one of HPE's essential innovations
  - Many funded under non-recurring engineering (NRE) contract, joint with Oak Ridge National Laboratory
  - Opportunities for other sites to deploy Rabbit modules, extend NRE directions

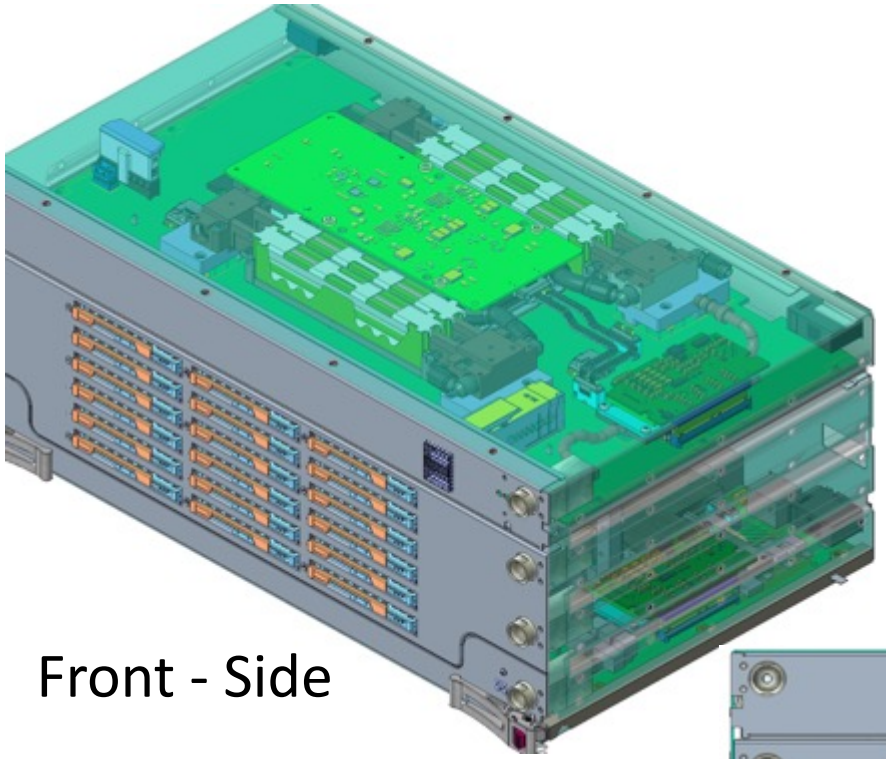
We will deploy other future heterogeneous system architectures with data analysis nodes

Rabbit modules are a 4U near node local storage solution

- All in one solution: Rabbit- 4U
  - Houses 18 SSD's (16+ 2 spares) that attach to Rabbit-S board
  - Locates Storage Processor (AMD Epyc CPU) on Rabbit-P board
- Compute blades direct attached to Rabbit-S through bulkhead cables
- Rabbit-S to Rabbit-P board connections are internal (no external cables)
- Deployed in LLNL EAS3s

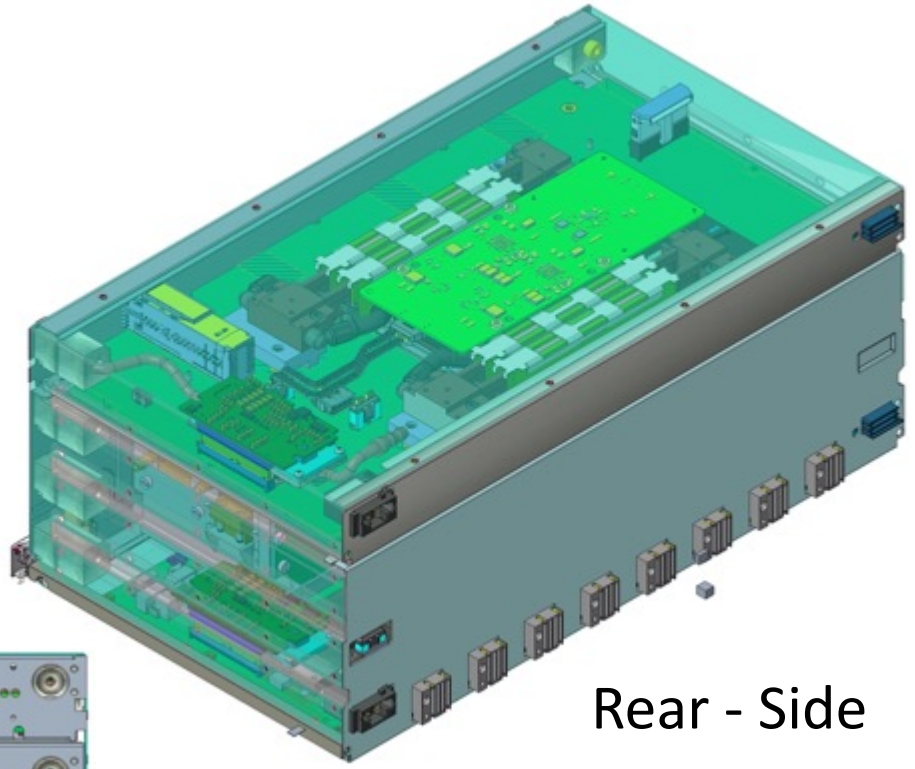


Rabbit 4U design provides easy access to SSDs



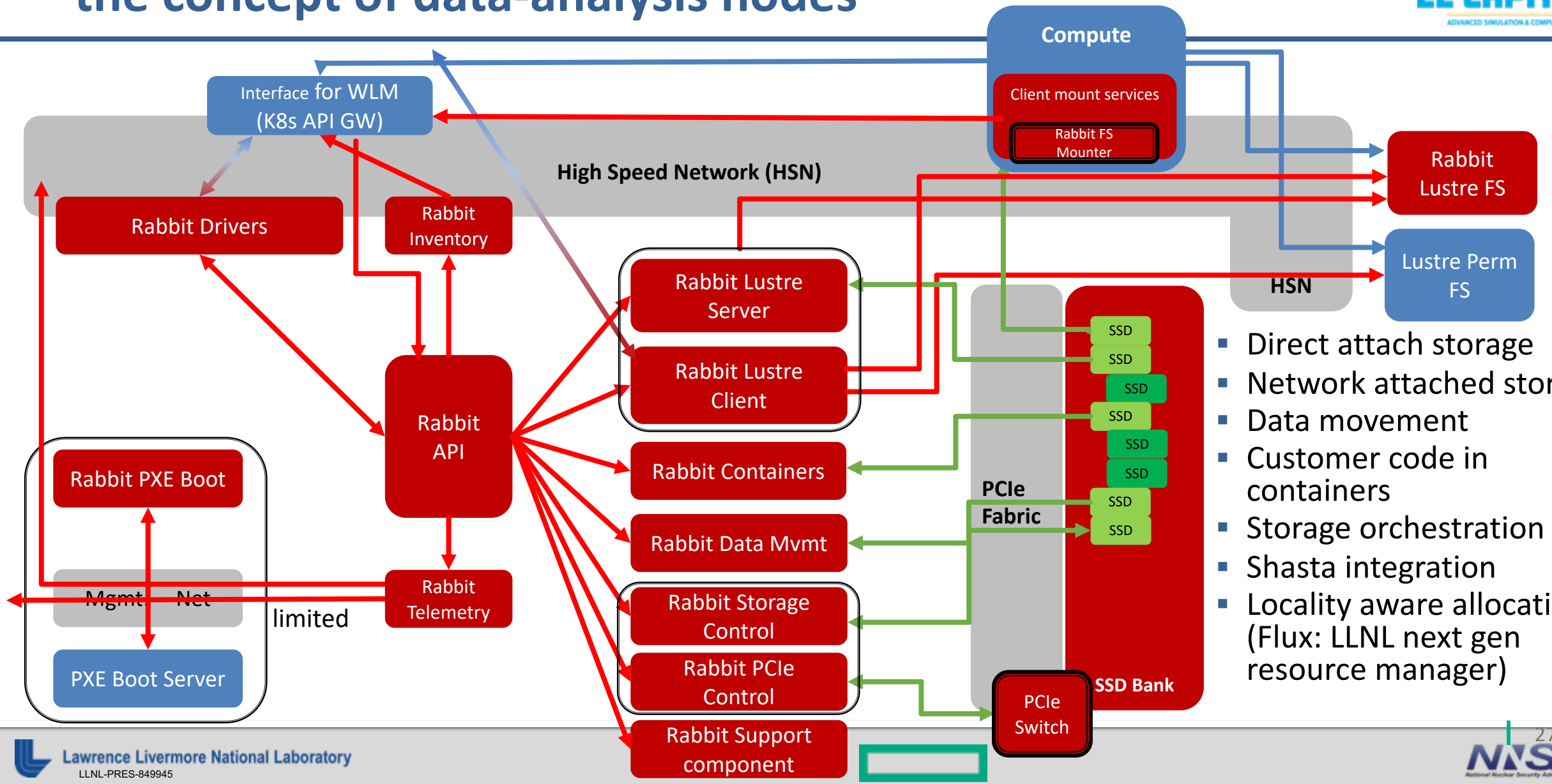
Front - Side

Front



Rear - Side

Rabbit software will enable many use cases including the concept of data-analysis nodes

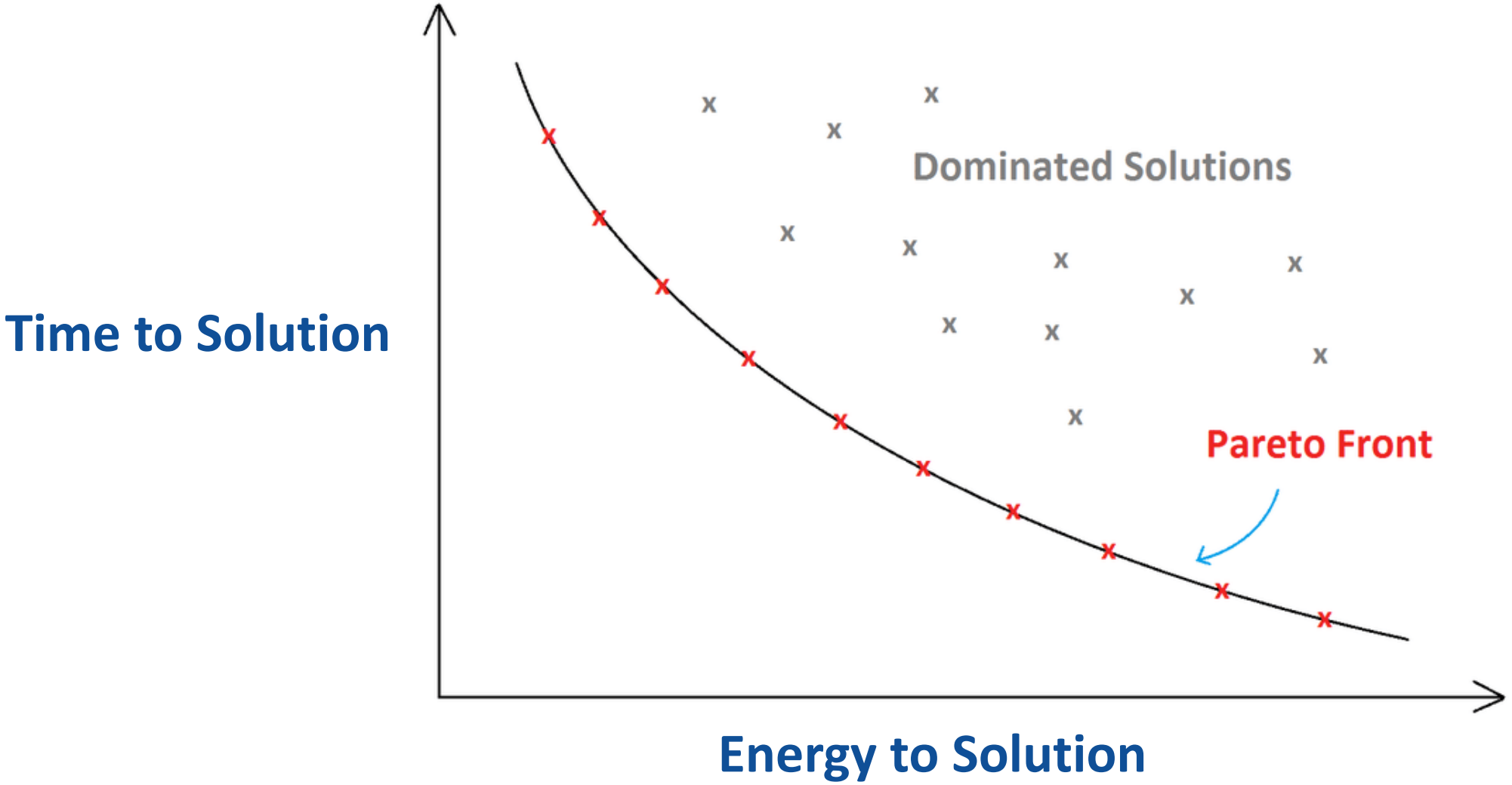


- Direct attach storage
- Network attached storage
- Data movement
- Customer code in containers
- Storage orchestration
- Shasta integration
- Locality aware allocations (Flux: LLNL next gen resource manager)

Questions to ponder related to energy consumption and energy efficiency of large-scale systems

- What can be done to improve energy efficiency?
  - Where has academic research focused?
  - What has led to improvements historically?
  - Will those historic improvements continue?
- How should we measure energy efficiency?
  - Should we use Gflops/W?
    - Green500 uses HPL performance divided by energy to achieve that performance
  - What metrics apply to other types of systems?
  - What are the shortcomings of current approaches?
- How can we motivate improvements in energy efficiency?
  - What motivates users?
  - What motivates large-scale centers (i.e., system providers)?
- Would improvements in energy efficiency actually address community concerns?

Users are likely to work towards Pareto optimal executions: Will that provide the desired result?



Questions to ponder related to energy consumption and energy efficiency of large-scale systems

- What can be done to improve energy efficiency?
  - Where has academic research focused?
  - What has led to improvements historically?
  - Will those historic improvements continue?
- How should we measure energy efficiency?
  - Should we use Gflops/W?
    - Green500 uses HPL performance divided by energy to achieve that performance
  - What metrics apply to other types of systems?
  - What are the shortcomings of current approaches?
- How can we motivate improvements in energy efficiency?
  - What motivates users?
  - What motivates large-scale centers (i.e., system providers)?
- Would improvements in energy efficiency actually address community concerns?

LLNL's platform strategy builds on our successful history to continue to meet ASC's mission

- El Capitan will continue LLNL's GPU-accelerated era that Sierra began
- TOSS will provide commonal system software stack across all LLNL systems
- LLNL's Flux resource manager will be critical for both and will support heterogeneous system architectures

- El Capitan will continue the overall trend towards more energy-efficient hardware
  - Frontier achieves 52.592 Glops/W (so < 20 MW/HPL Exaflop)
  - Expect El Capitan hardware to imcrease energy efficiency for a variety of reason
- LLNL is actively exploring ways to improve energy efficiency in practice

LLNL's strategy ensures that we deliver the best value possible for our mission