The LLNL Near and Long Term Vision for Large-Scale Systems

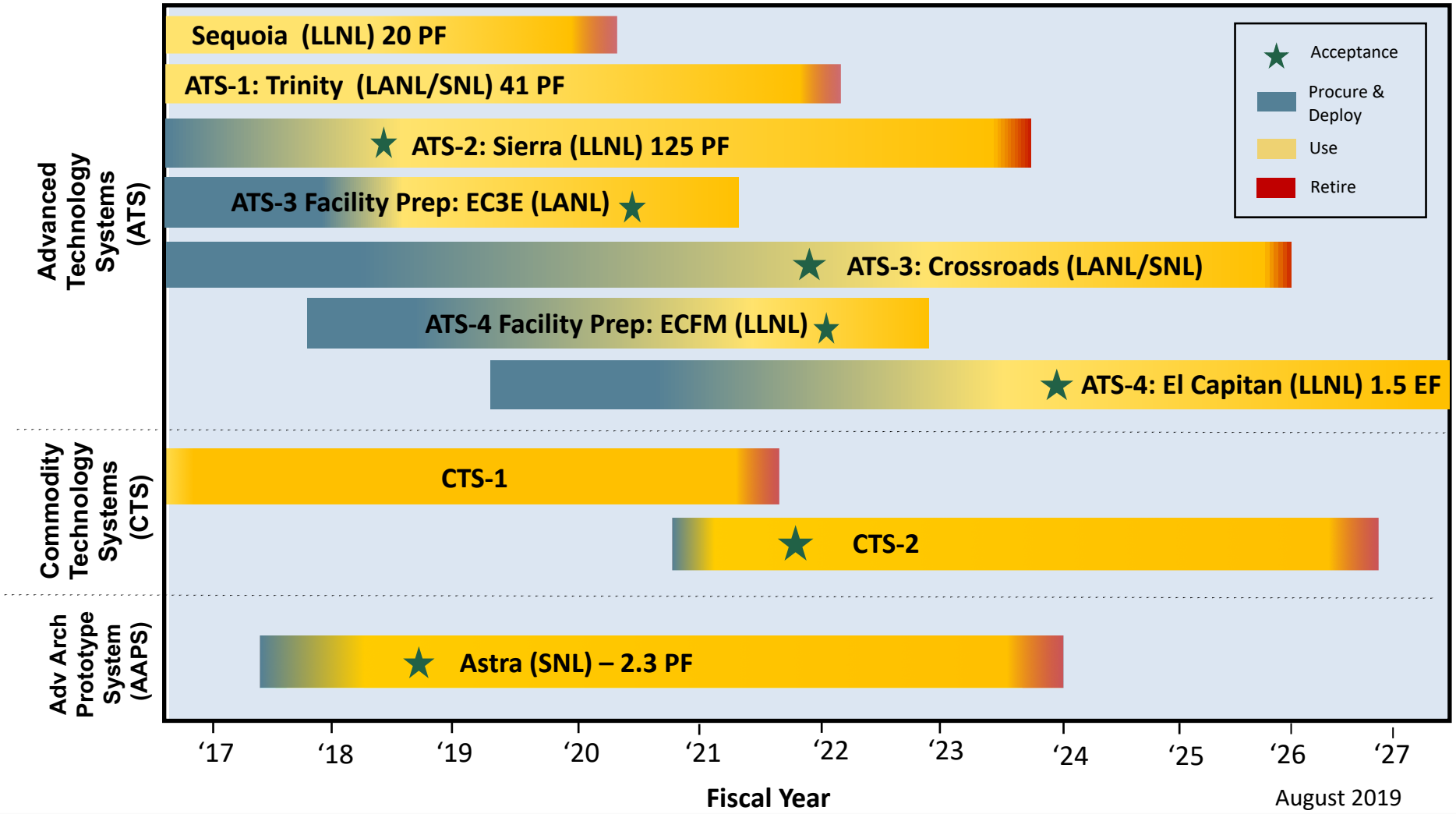
Bronis R. de Supinski
Chief Technology Officer for Livermore Computing

November 4, 2019

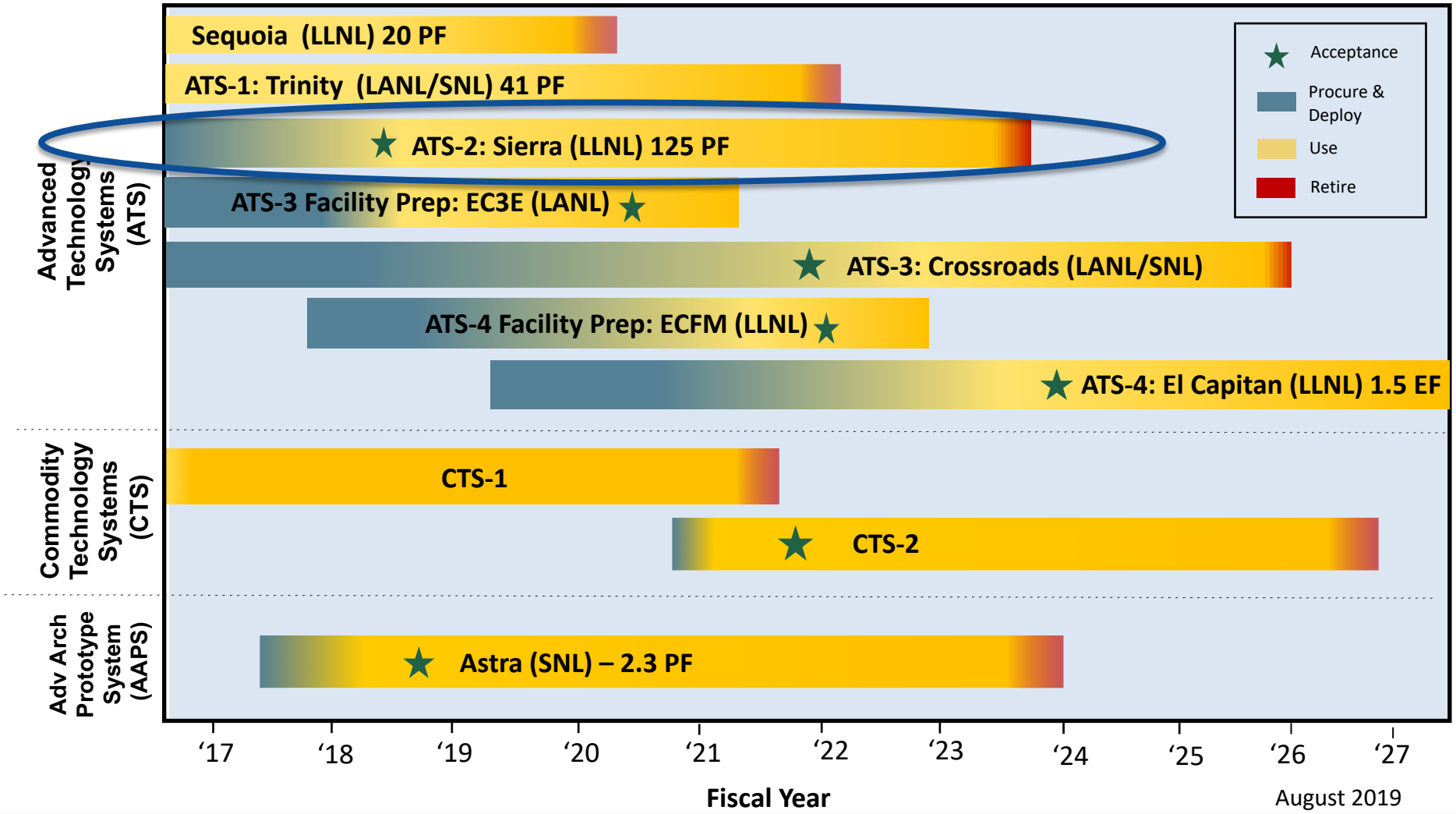
LLNL-PRES-787929
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

The ASC platform time line shapes the LLNL strategy



Sierra is the newest ASC ATS platform



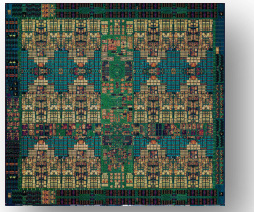
Sierra is LLNL's 125-petaflops ASC ATS system

Lawrence Livermore National Laboratory

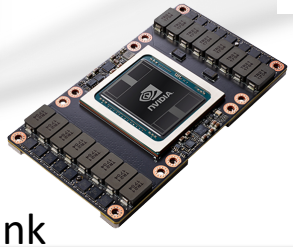
Sierra features a GPU-accelerated architecture

Components

IBM POWER9
- Gen2 NVLink

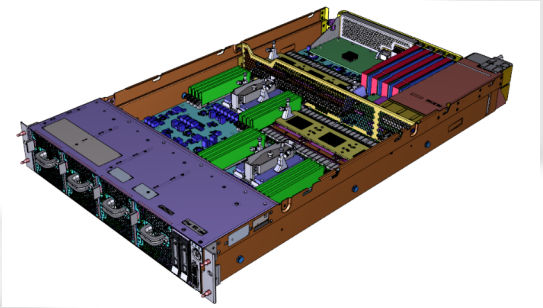


NVIDIA Volta
- 7 TFlop/s
- HBM2
- Gen2 NVLink



Compute Node

2 IBM POWER9 CPUs
4 NVIDIA Volta GPUs
NVMe-compatible PCIe 1.6 TB SSD
256 GiB DDR4
16 GiB Globally addressable HBM2 associated with each GPU
Coherent Shared Memory



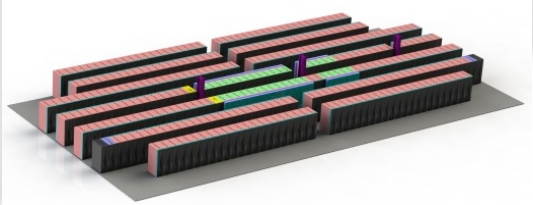
Mellanox Interconnect
Single Plane EDR InfiniBand
2 to 1 Tapered Fat Tree

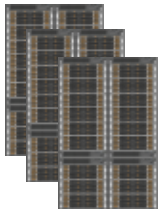
Compute Rack

Standard 19”
Warm water cooling

Compute System

4320 nodes
1.29 PB Memory
240 Compute Racks
125 PFLOPS
~12 MW





Spectrum Scale File System

154 PB usable storage
1.2 TB/s R/W bandwidth

Sierra architectural decisions reflect its planned UQ workload

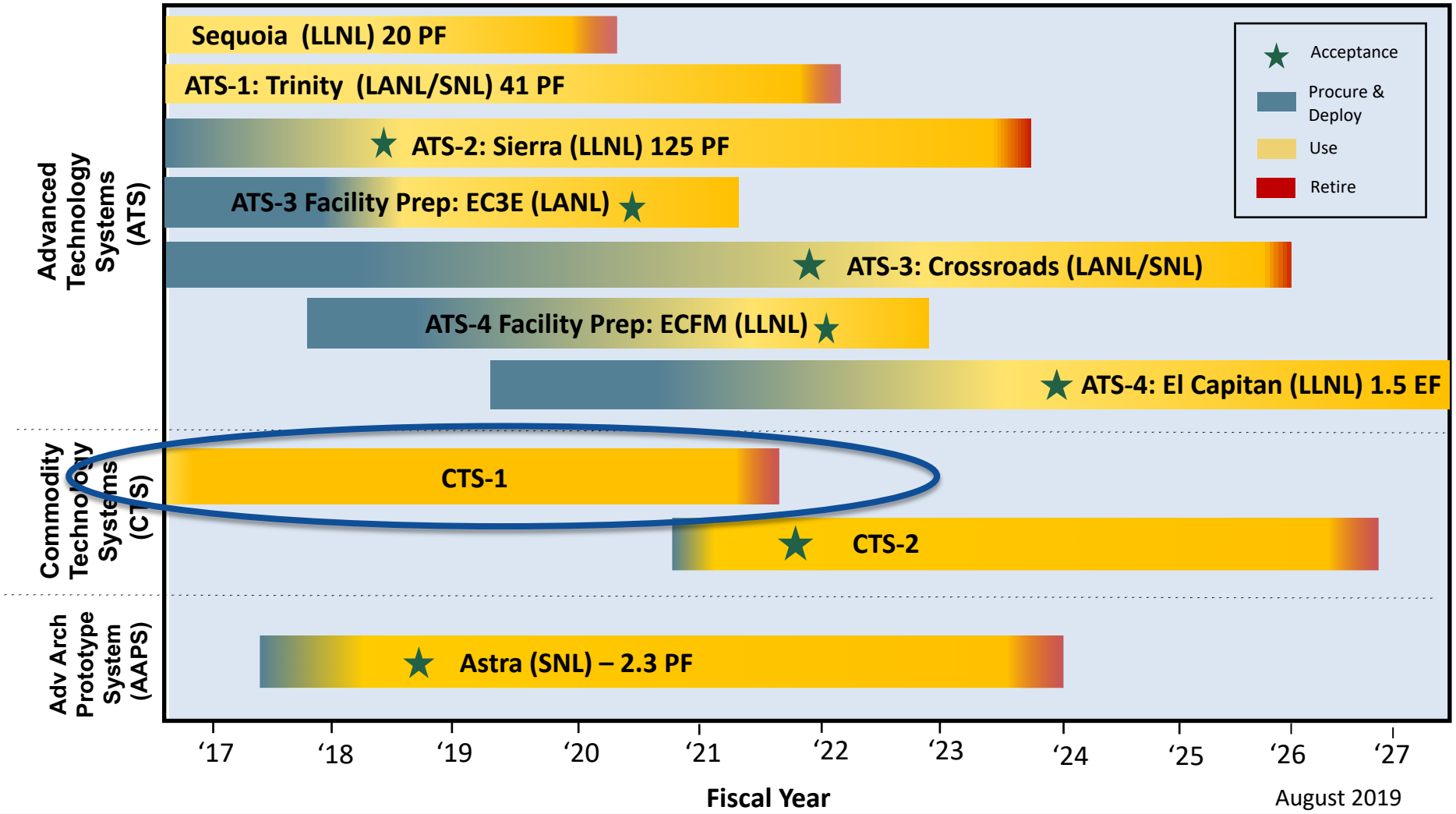
- Sierra is contrasted with ORNL's Summit system
  - Summit features 3 Voltas per Power9 (i.e., 6 GPUs per node)
  - Summit has a full bandwidth fat-tree
  - Summit has 2X main memory per node compared to Sierra
- Sierra's workload focuses on uncertainty quantification
  - Multiphysics ensemble calculations that stress throughput
  - Fit each physics package into $\leq$ 64 GiB memory per node
  - Aggregate memory footprint under total deployed on Sierra
  - Relatively low network demand, placed to minimize contention
- Sierra architectural decisions support this workload
  - Traded network and memory for compute nodes

These tradeoffs improve Sierra's effectiveness by about 5%

Sierra system architecture details serve as the basis of two major LLNL systems

	Sierra	Lassen
Nodes	4,320	792
POWER9 processors per node	2	2
GV100 (Volta) GPUs per node	4	4
Node Peak (TFLOP/s)	29.1	29.1
System Peak (PFLOP/s)	125	23.0
Node Memory (GiB)	320	320
System Memory (PiB)	1.29	0.253
Interconnect	2x IB EDR	2x IB EDR
Off-Node Aggregate b/w (GB/s)	45.5	45.5
Compute racks	240	44
Network and Infrastructure racks	13	4
Storage Racks	24	4
Total racks	277	52
Peak Power (MW)	~12	~2.0

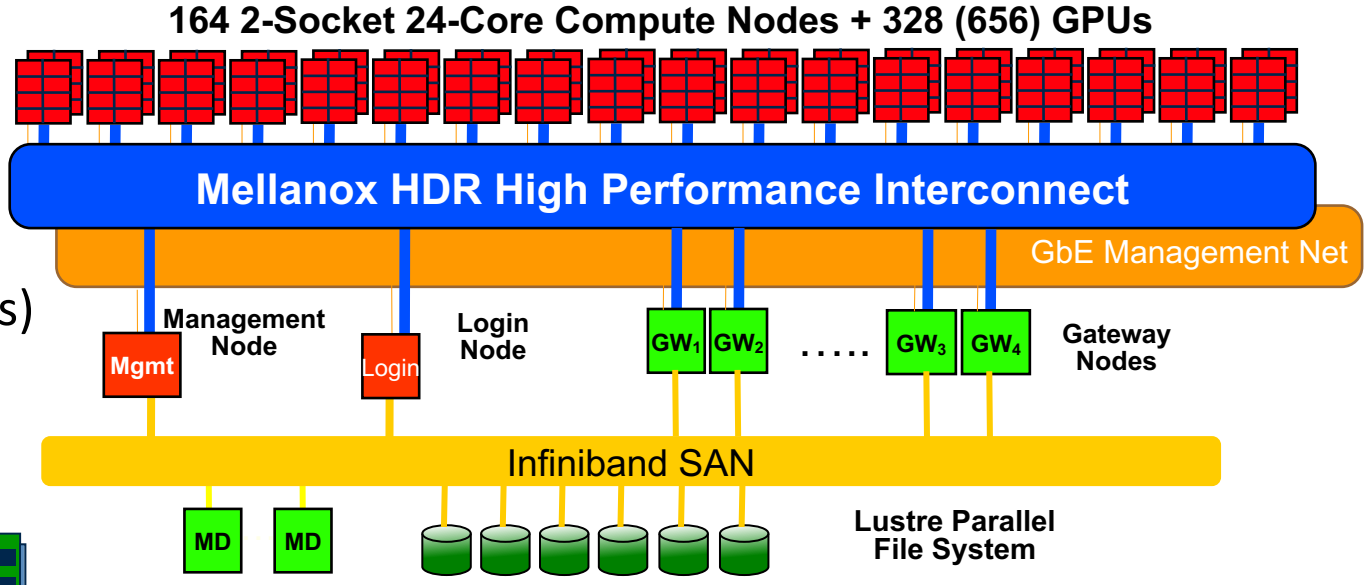
CTS1 is the current generation of ASC Commodity Technology Systems

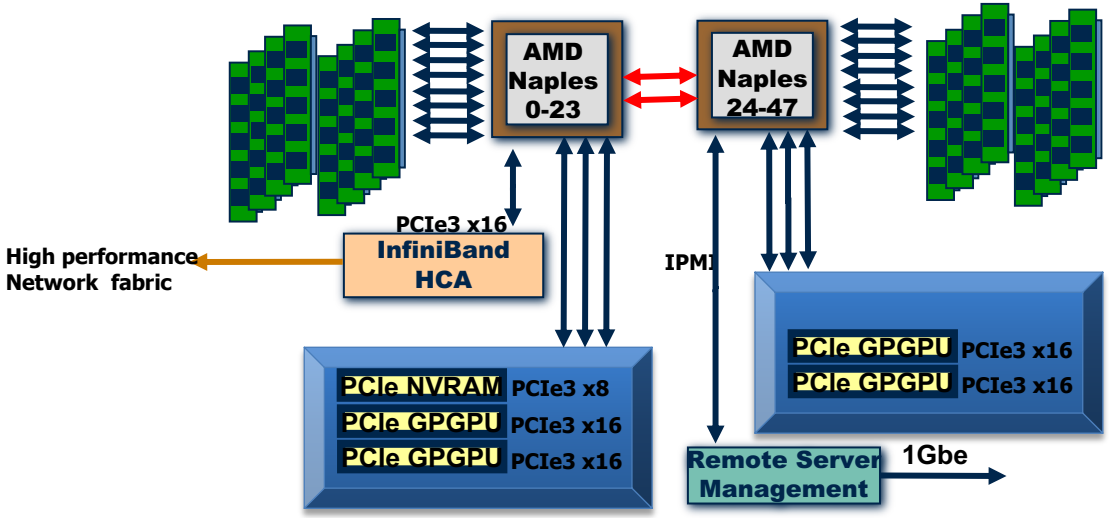


LLNL CTS1 systems include Jade and Quartz (Nos. 83 and 84 on June 2019 Top500)

Corona is a follow-on to Catalyst: AMD GPU cluster for HPC, ML, and Data Science

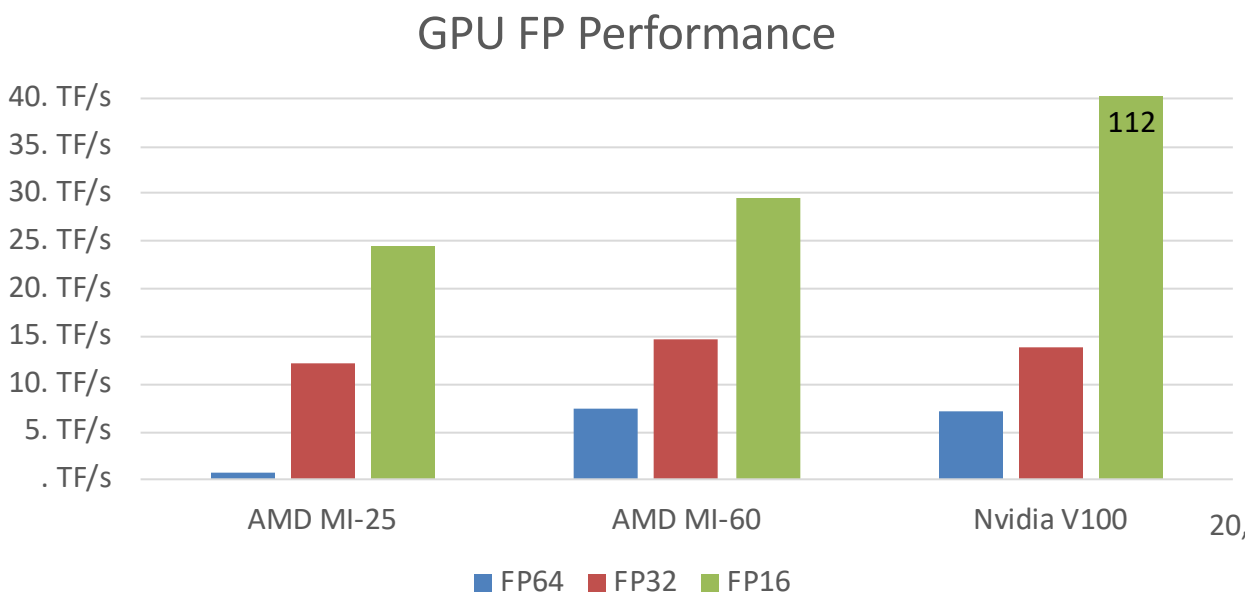
- Node architecture
  - AMD Naples 24-core 2.0 GHz
  - Memory: 256 GB; 5.3 GB/core
  - Memory BW: > 300 GB/s DDR
  - 1.6 TB NVMe (R/W: 3.35GB/s & 2.1 GB/s)
  - Mellanox HDR100
  - 4 GPU per compute node

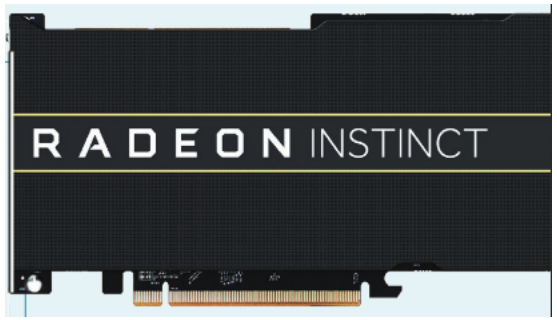


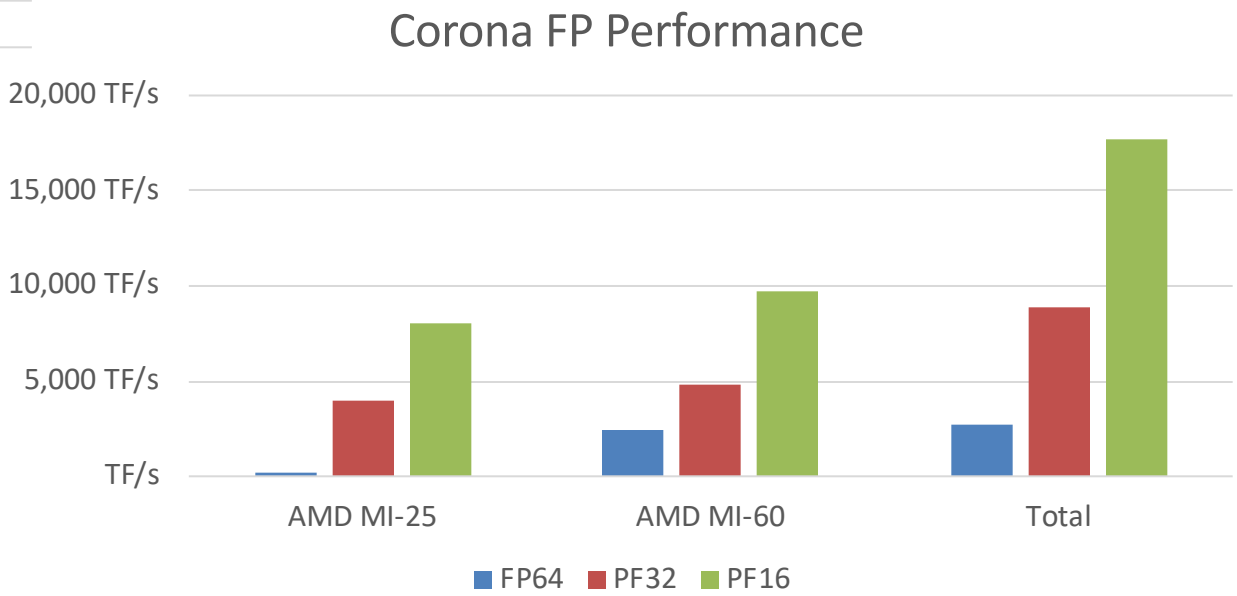


- System nodes
  - 82 (CPU + 4 MI-25 GPUs)
  - 82 (CPU + 4 MI-60 GPUs)
  - 4 Gateways
  - 1 Login
  - 1 Management

Addition of AMD MI-60 GPUS significantly increases Corona's capability





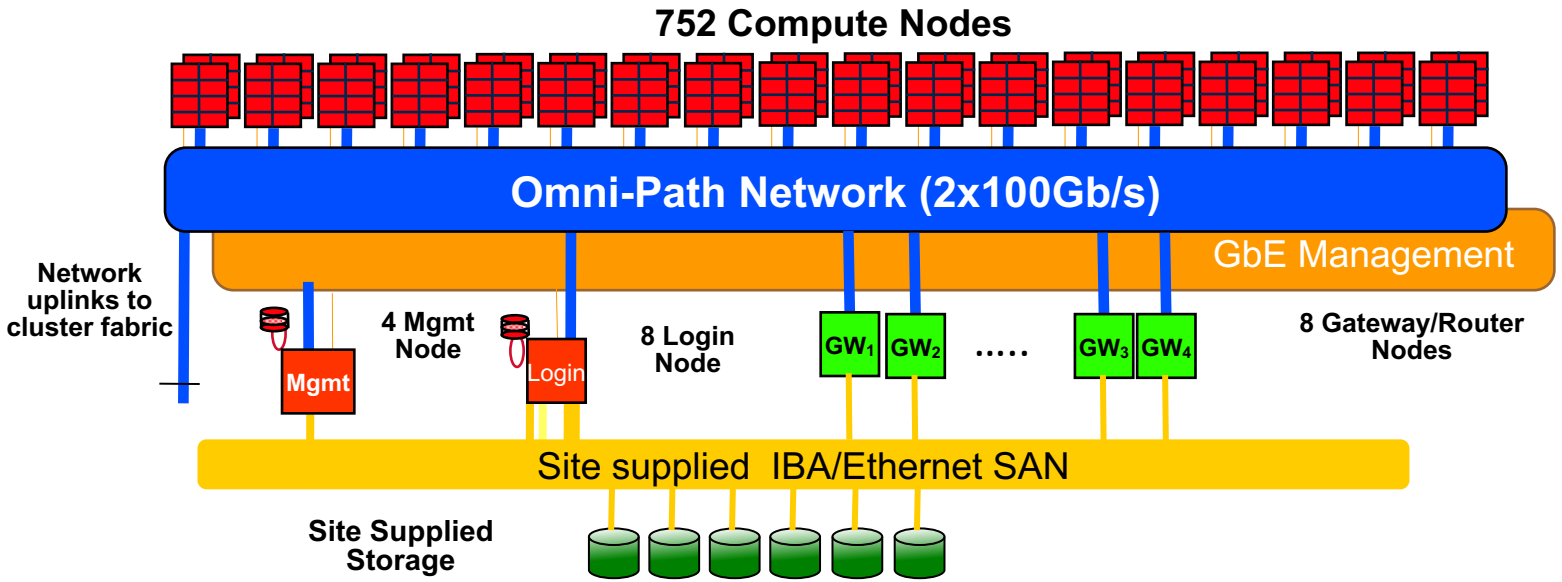


- Added 328 AMD MI-60 GPUs to Corona beginning in August 2019

Magma is a next-generation CTS1 system for the LLNL ASC Program

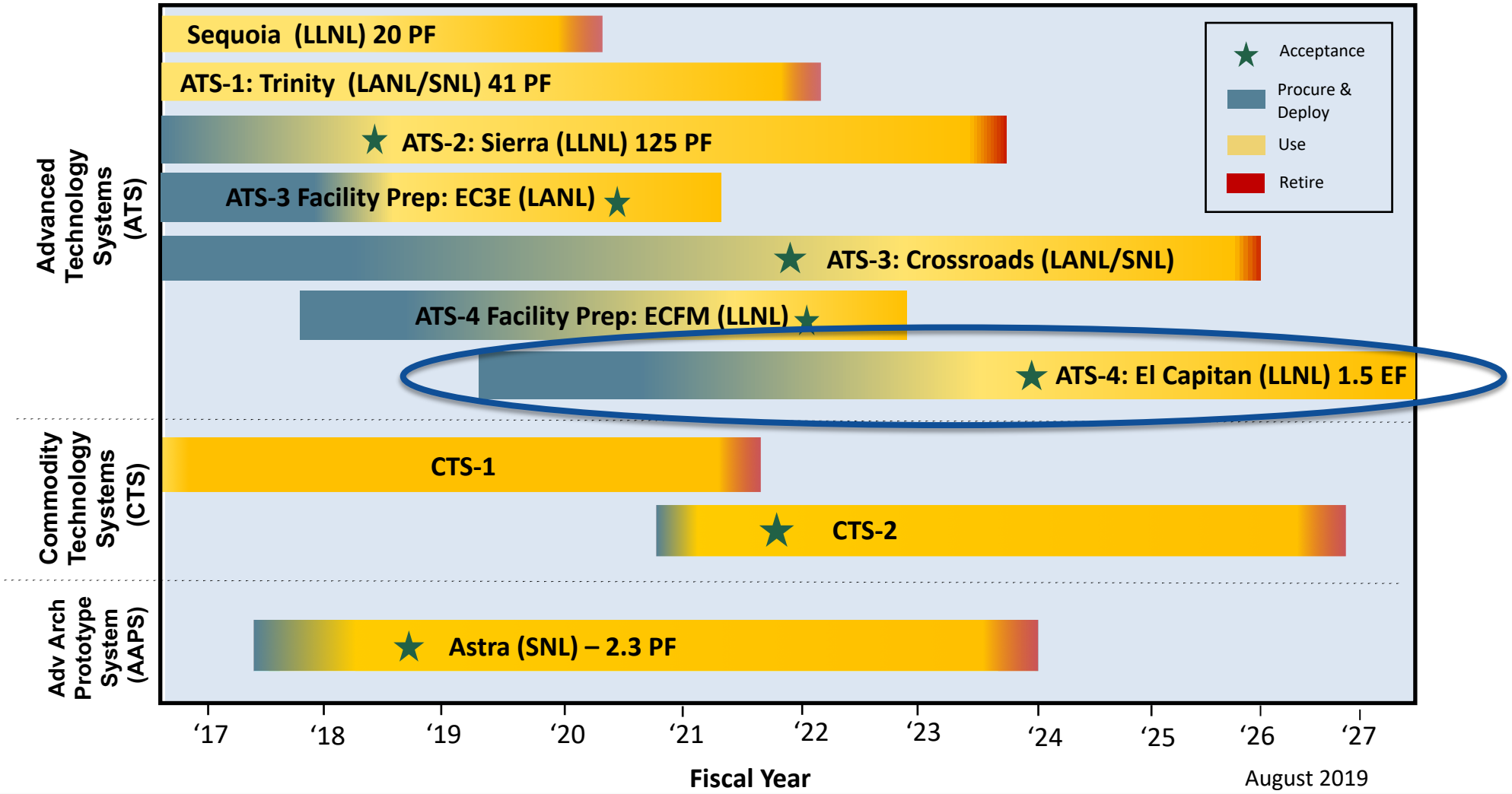
- 768 Nodes (4 Scalable Units)
- Intel Cascade Lake AP based nodes
- Intel Servers (4 node in 2U)
- CoolIT direct liquid cooling to CPUs and DIMMs
- Redundant CDUs
- Dual-Rail Omni-Path Interconnect
- 63.4kW per compute rack
  - ~25kW per compute rack for prior CTS1 systems
- Delivered November 2019, operational January 2020

Magma provides a high memory bandwidth design that addresses ASC application requirements



- Magma parameters 772 nodes (752 compute; 8 GW; 8 Login; 4 Mgmt)
  - CLX-AP compute and login nodes
  - CLX-SP gateway and management nodes
  - Dual socket nodes; Total memory capacity 295 TB; 430 TB/s memory bandwidth
  - 4 GB memory capacity per CPU core
  - 5.6 DP PF/s theoretical peak
  - Over 73K cores

El Capitan will be LLNL's next ASC ATS platform



The CORAL2 RFP targets exascale systems

EL CAPITAN
ADVANCED SIMULATION & COMPUTING

Energy consumption

- < 40MW, 20MW to 30MW desired

Reliability

- > 6 days without human action due to failure

Memory and storage

- 8PiB load/store memory; 50X file system capacity

Parallelism

- Must improve application performance 50X over Sequoia

El Capitan can meet all of these high-level requirements

Cray will deliver a highly capable GPU-accelerated system

EL CAPITAN
ADVANCED SIMULATION & COMPUTING

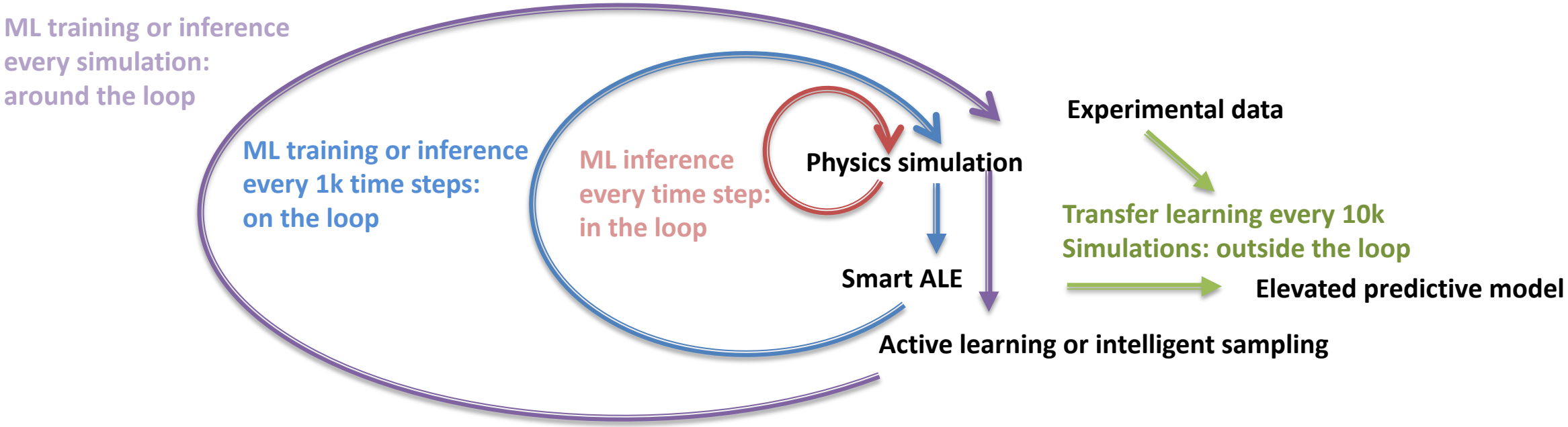
- El Capitan will meet its stockpile stewardship simulation mission
- System will feature:
  - Peak ≥ 1.5 DP exaflops
  - Peak power < 40 MW
    - Anticipating ~30MW
    - Facility will support 85MW total
  - Cray Slingshot interconnect

- Cray will provide several critical innovations
  - Cray and LLNL are working with ORNL jointly on non-recurring engineering (NRE) activities
  - Shasta software stack will feature greater flexibility under Cray's Compass program
  - El Capitan will include an innovative near node local storage solution

Late binding of the processor solution will ensure El Capitan provides the best possible value

Cognitive simulation wraps simulation in multiple layers of ML inference and training

- High precision scientific **simulation**
- Frequent machine learning **training**
- Potentially very high frequency **inference**



Cognitive simulation could drive continued exponential capability improvements

Does supporting cognitive simulation merit increased system-level heterogeneity?

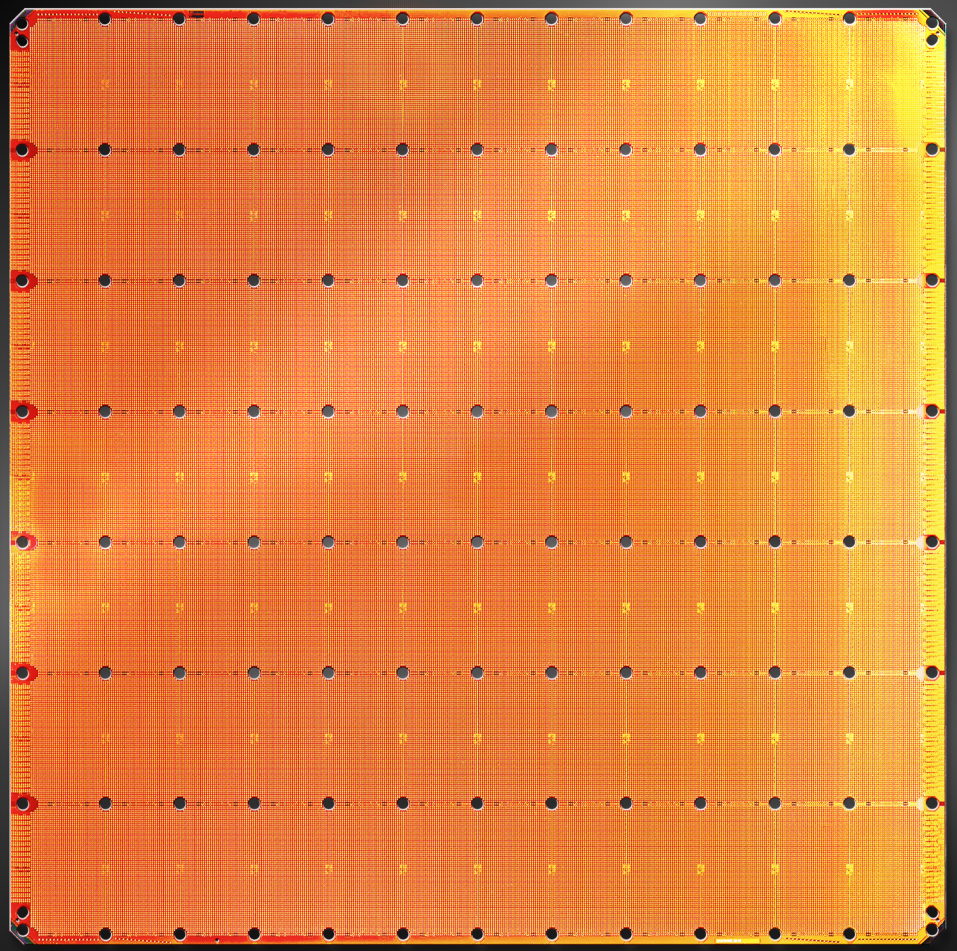
- Heterogeneity at the node-level is now common as GPU-based systems proliferate
- Wide range of new devices
- System-level heterogeneity involves a system with diverse node types **by design**
  - Deviates from the "cookie cutter" concept of system design most of us have
  - Allows non-integer ratios of diverse device types, which could yield a more cost-effective solution
  - Only rational if nodes (or resources) target different aspects of the overall workload
  - Already common in today's systems
    - Front-end nodes
    - Service nodes and management nodes
    - Range of node types in parallel file systems
- System-level heterogeneity raises significant challenges
  - How to partition work and to handle interactions across diverse nodes (or resources)?
  - How to determine the right mix of diverse resource types?
  - How to ensure system architecture and interconnect supports use of diverse resource types?
  - How to adapt applications to use heterogeneity at the system level? Likely requires more asynchrony

We will explore system-level heterogeneity by integrating machine learning resources into Lassen

- Lassen (unclassified Sierra) provides a unique opportunity
  - Number 10 system on Top500
  - Unused InfiniBand network switch ports
  - Highly capable on-node ML resource (GPUs)
- Projects will employ Center of Excellence (CoE) model to join industry hardware, low-level software and ML expertise with our expertise in applications, ML and system design
- First project: Cerebras CS-1
  - Wafer-scale integration supports unprecedented ANN training and inference response times
  - 1.2 Tb/s network bandwidth supports integration into Lassen to explore all levels of cognitive simulation
- Exploring other integrated testbeds
  - Will likely add at least one other to ensure we fully assess space

Cognitive simulation could drive continued exponential capability improvements

Cerebras Wafer Scale Engine



Cerebras WSE

1.2 Trillion Transistors

46,225 mm² Silicon

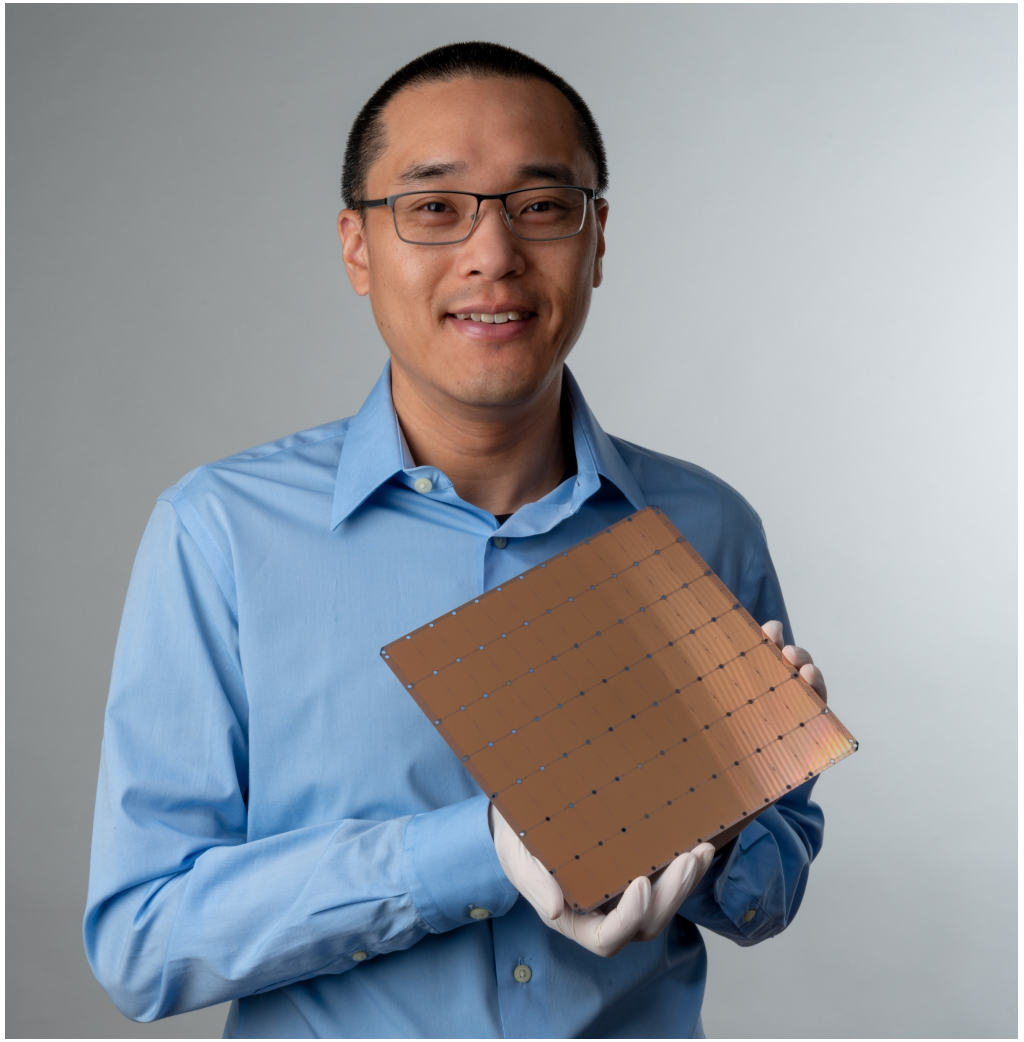
Largest GPU

21.1 Billion Transistors

815 mm² Silicon

Cerebras CS-1 features the largest chip ever built

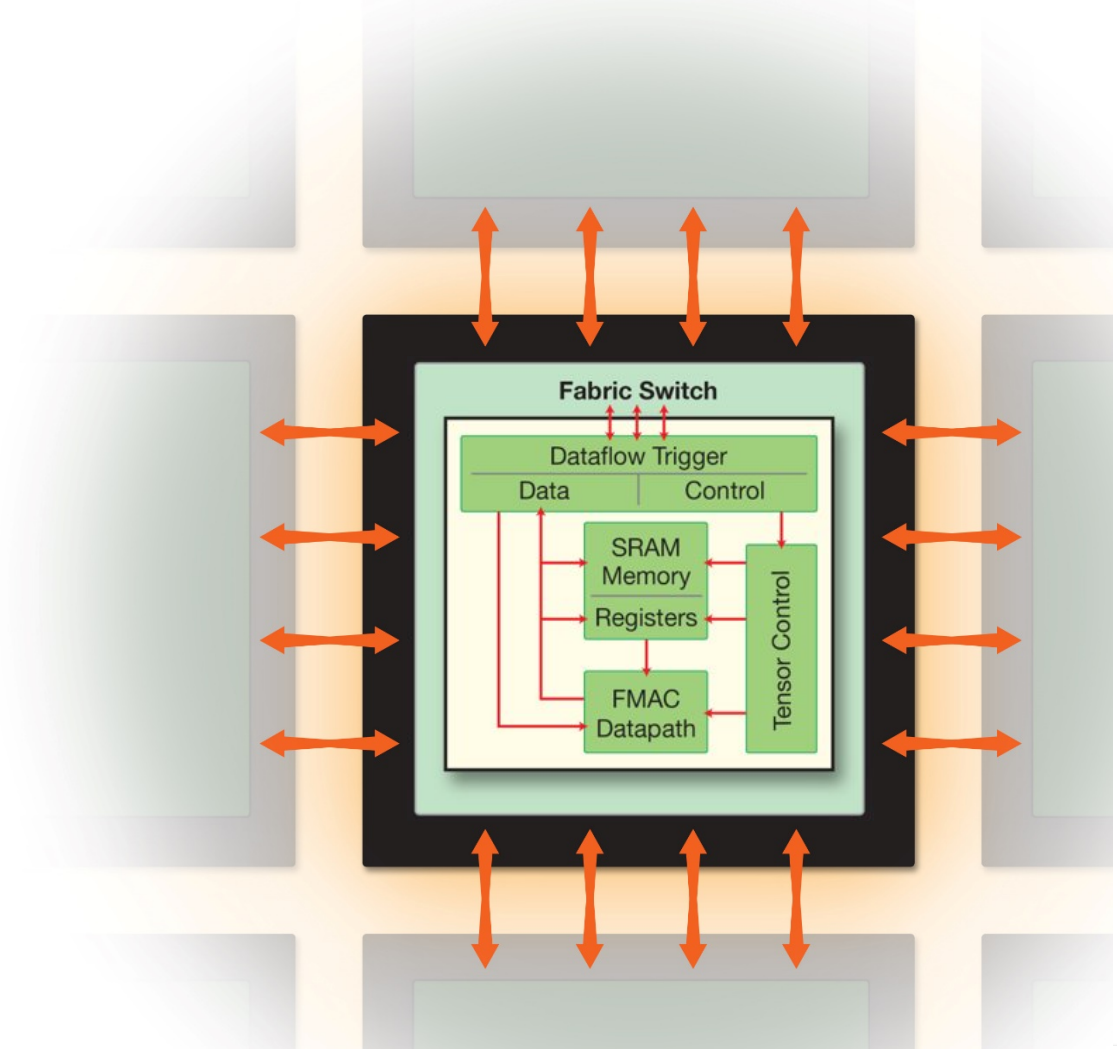
- 46,225 mm^2 silicon
- 1.2 trillion transistors
- 400,000 AI optimized cores
- 18 Gigabytes of On-chip Memory
- 9 PByte/s memory bandwidth
- 100 Pbit/s fabric bandwidth
- TSMC 16nm process



CS-1 flexible cores are optimized for tensor operations, the key to supporting rapidly evolving NN architectures

- Fully programmable compute core
- Full array of general instructions with ML extensions
- Flexible general ops for control processing
  - e.g. arithmetic, logical, load/store, branch
- Optimized tensor ops for data processing
  - Tensors as first class operands
  - e.g. fmac [z] = [z], [w], a

```
                3D   3D 2D scalar
```



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
- CTS1 systems
  - Support critical architectural diversity
  - Include highly capable new systems
- CTS2 systems
  - Will continue the Linux cluster tradition
  - GPU-accelerated systems are likely
- LLNL will explore system-level heterogeneity
  - Cerebras CS-1 system integrated into Lassen will support cognitive simulation
  - If Lassen experiments are successful could incorporate approach into El Capitan
  - Expect system-level heterogeneity models to guide aspects of LLNL's next ATS procurement

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

Lawrence Livermore National Laboratory