Commodity Technology Systems 2 (CTS-2) Update

Matt Leininger
CTS-2 Technical Lead
Livermore Computing
& Tri-Lab CTS-2 Team

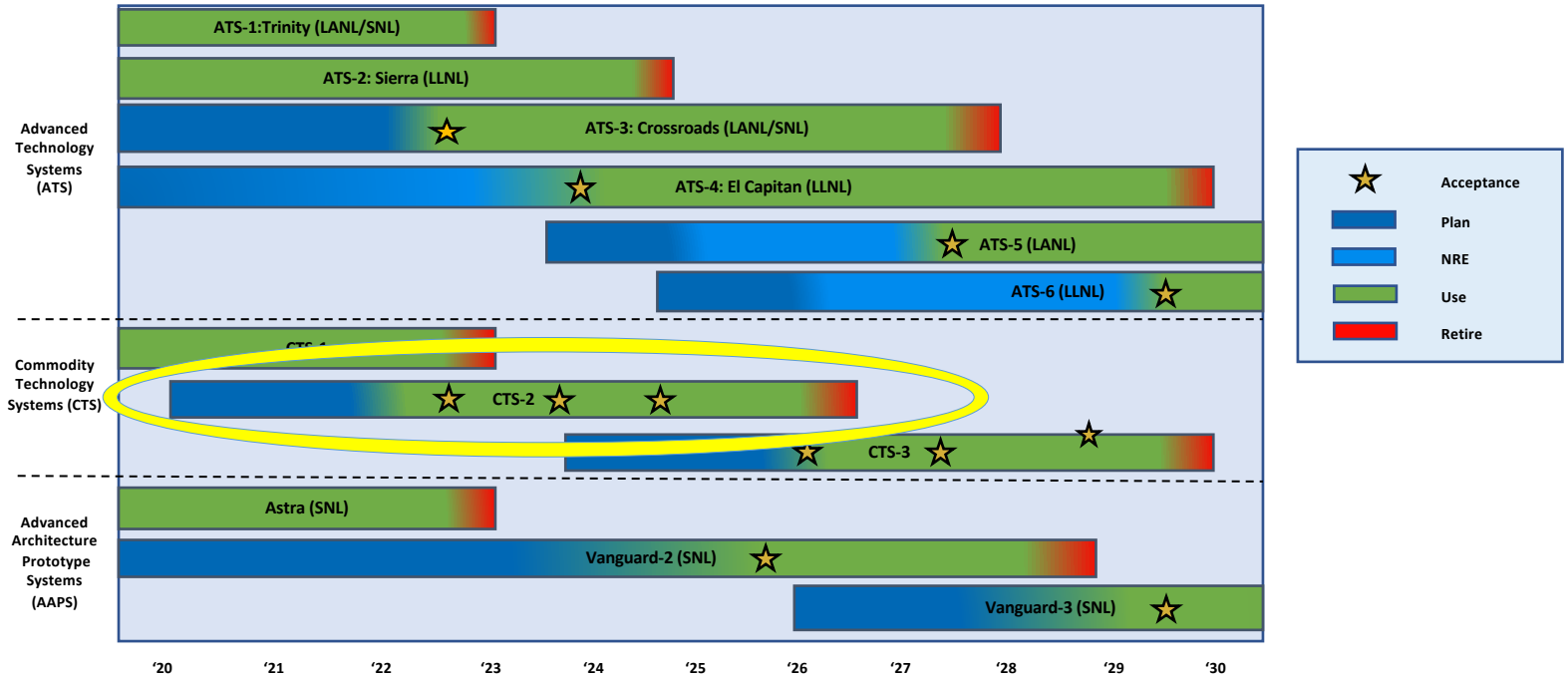
September 2022

LLNL-PRES-839624

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

ASC Platform Timeline



March 2022

Overview of NNSA Commodity Technology Systems (CTS)

What is CTS?

A Tri-lab (LLNL, LANL, Sandia) joint procurement to provide a common hardware platform for robust capacity computing platforms. CTS provides common simulation & operating environment when coupled to Tri-lab software stack (TOSS) & Tri-Lab Common Environment (TCE).

CTS Advantages

- Reduce total cost of ownership associated with system procurements, integrations, and support.
- Rapid deployment of systems into production.
- Helps enable TCO reductions realized from TOSS.
- Supports ASC, DP & other computing programs.
- Strategy adaptive to new technology & workloads.
- Supports computing demands for HPC, data sciences, ML/AI, and visualization.

A Brief History of NNSA Commodity Systems

TLCC1 2007

- Tri-Lab Linux Capacity Clusters
- 1st joint procurement
- 2007-2010
- 11,000 CPU sockets
- 0.4 PetaFlops/sec

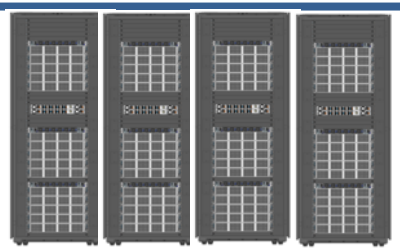
TLCC2 2011

- Tri-Lab Linux Capacity Clusters
- 2nd joint procurement
- 2011-2015
- 10,700 CPU sockets
- 1.8 PetaFlops/sec

CTS-1 2016

- Commodity Technology Systems
- 3rd joint procurement
- 2016-2021
- 21,000 CPU sockets
- 18 PetaFlops/sec

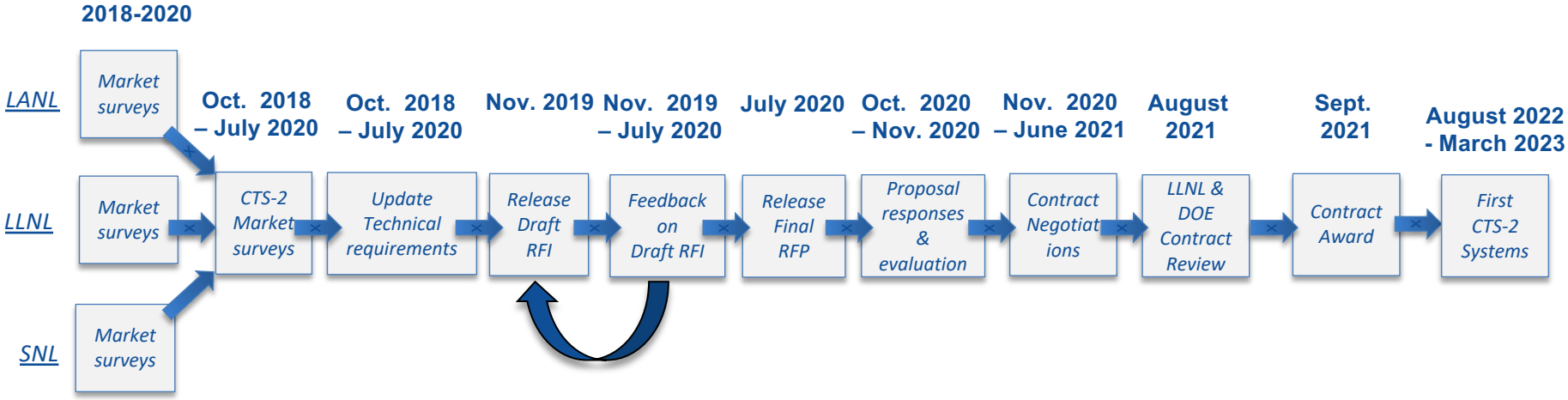
CTS-2 2022



- Commodity Technology Systems
- 4th joint procurement
- 2022-2025
- Likely > 21,000 CPU sockets
- Likely > 65 PetaFlops/sec

Commodity Platforms have successfully delivered to simulation programs at LLNL, Sandia, and LANL for 15 years and growing.

CTS-2 Activities On Track for Deployments in CY2022

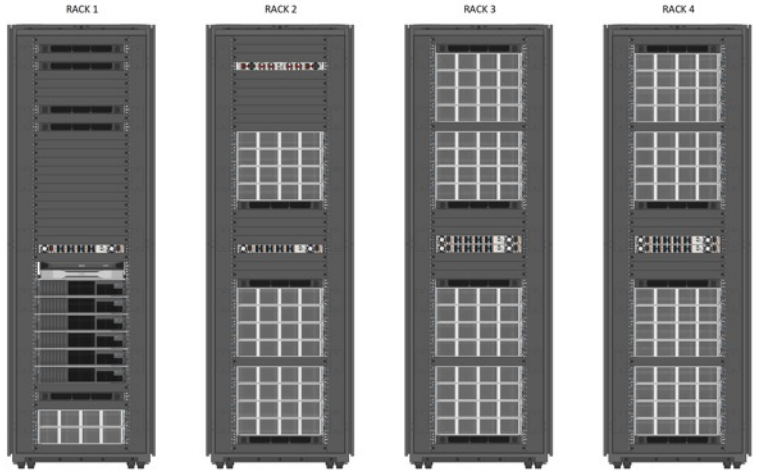


CTS-2 and Tri-Lab software teams (TOSS & TCE) work together throughout the CTS-2 procurement, deployment, and platform lifetimes.

CTS-2 Contract Awarded to Dell Technologies

- Compute Node Servers: Dell C6620 (4 nodes in 2U)
- Login/Mgmt/Gateway Servers: Dell 760 2U
- Intel Sapphire Rapids Xeon CPUs
- HPC Network: Cornelis Networks Omni-Path or Mellanox InfiniBand
- Liquid Cooling: CoolIT direct-to-chip liquid cooling
- Power Density: Up to ~66kW/compute rack
- Power: 480V 3 phase
- Options for GPUs: NVIDIA, AMD, & Intel

1 Scalable Unit CTS-2 Rack Layout

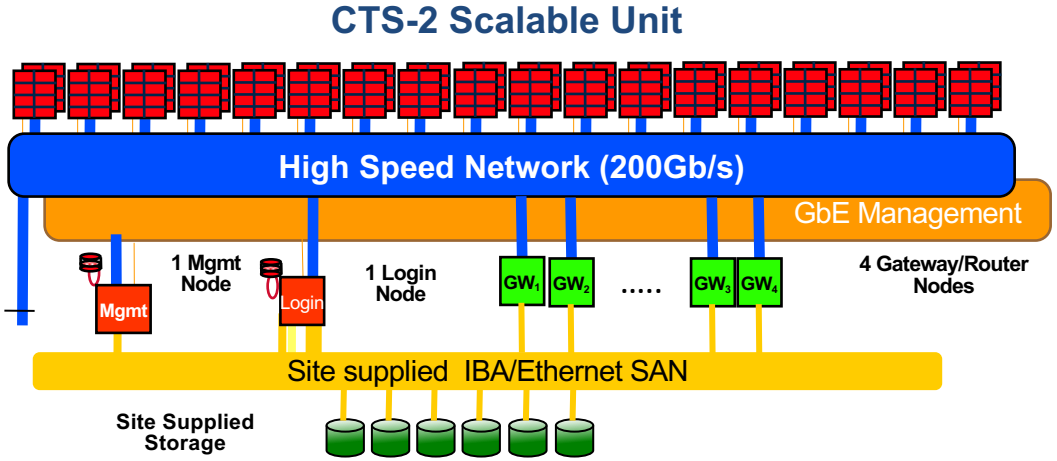


C6620 Compute Node

R760 Chassis

C6620 Chassis

First Dell CTS-2 Systems Arriving in Late 2022



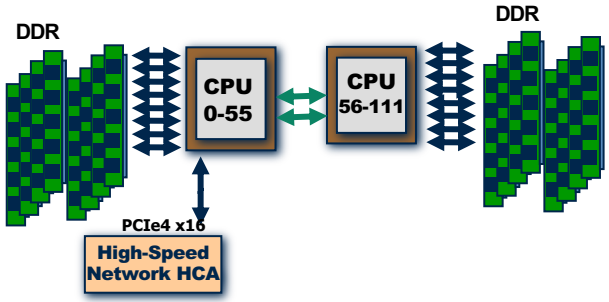
CTS-2 Scalable Unit (SU)			
Nodes	CPU Cores	Memory Capacity	Theoretical Peak
192	~21.5K	**DDR5: 49 TB** DDR5: 98 TB HBM2e: 25 TB	1.5 PF/s (FP64) 3.0 PF/s (FP32) 6.1 PF/s (FP16/BF16) 49 PF/s (BF16/new TMUL op)

- System deliveries begin in August 2022
- Intel Sapphire Rapids CPUs (4th Generation Xeon-SP)
  - 56 cores/socket or 112 cores/node
  - **Default**: 256 GB DDR5
  - Alternative: 512 GB DDR5
  - Alternative: 128 GB HBM2e
- Cornelis 200 Gb/s High Speed Network
- Software Environment
  - TOSS4 based on RHEL 8.x
  - Tri-Lab Common Environment (TCE2)

Lawrence Livermore National Laboratory · Los Alamos National Laboratory

CTS-2 Base Node Architecture & Performance Estimates

CTS-2 Base Architecture



CPU + DDR5

- 2 socket 56 core 2.0+ GHz CPU
- 16x16 GB DDR5 DIMM no chipkill
- Total memory 256 GB DDR5
- ~2.2-2.4 GB/CPU core
- ~615 GB/s memory BW (theoretical)
- ~460 GB/s memory BW (actual)

Sapphire Rapids Processor	Overall Node FOM	HPCG	LAGHOS	Quicksilver	SNAP
DDR (relative to CTS-1)	4.9x	3.2x	6.2x	5.3x	6.5x

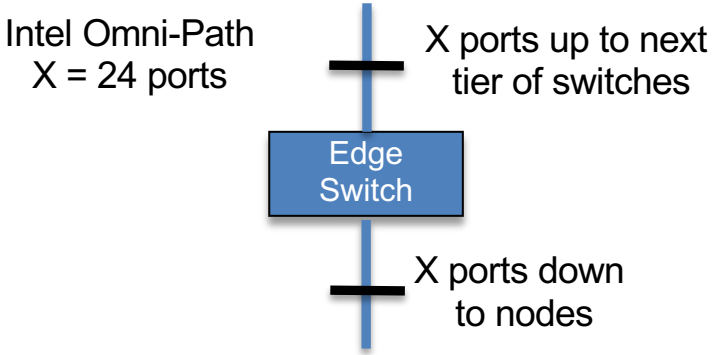
Sapphire Rapids Processor	Mini-app Node Performance	Power efficiency	Perf per $
DDR (relative to CTS-1)	4.9x	1.7x	~3x

Sapphire Rapids Processor	DGEMM (Est.)	STREAM (per Socket)
DDR	3.1 – 3.3 TF	230 GB/s

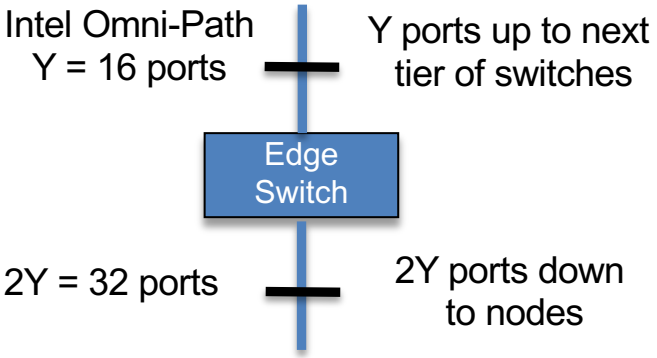
CTS-2 DDR will be available summer 2022

CTS-2 Architecture: Network Topology

Optional CTS-2 Configuration
Full Fat Tree Network



Default CTS-2 Configuration
2:1 Tapered Fat Tree Network

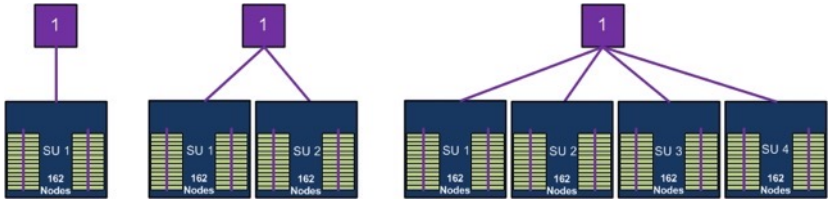


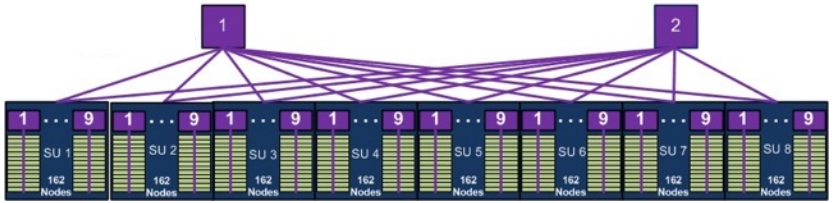
Maximize small/medium job throughput with 2:1 Tapered Network and more compute nodes!

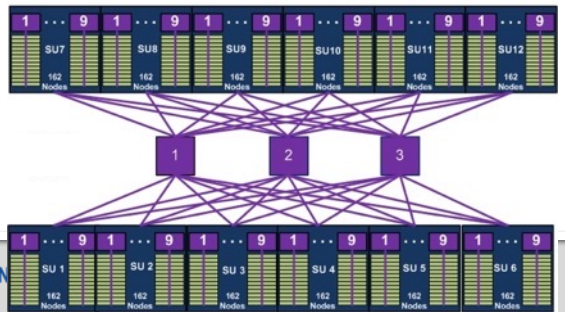
Scalable Units are the “Legos” that scale to Multi-SU platforms

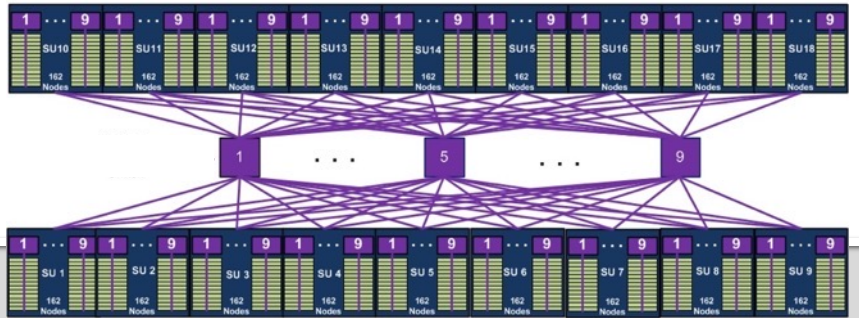
#SU’s	# Nodes	PFlops
1	192 - 200	1.45 – 1.5
2	384 - 400	2.9 – 3.0
4	768 - 800	5.8 - 6
6	1,152 – 1,200	8.7 – 9.0
8	1,536 – 1,600	11.6 – 12.0
12	2,304 – 2,400	17.3 – 18.0

One or more of each cluster size may be deployed at the Tri-labs









CTS-2 CoolIT Direct-to-Chip Liquid Cooling

CoolIT CHx750 CDU

CoolIT CHx80 CDU

Dell C6620 Compute Node

Evolution of Power & Cooling Requirements for Commodity Systems

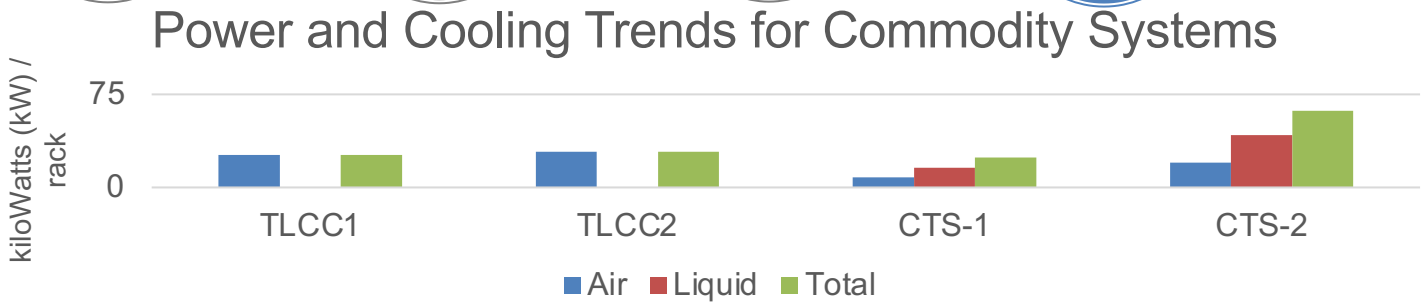
- 26 kW per rack
- 26 kW air cooled
- 208V power

- 28 kW per rack
- 28 kW air cooled
- Fielded 1 liquid cooled solution
- 208V power

- 24 kW per rack
- 8 kW air cooled
- 16 kW liquid cooled
- 480V & 208V power

- Next generation
- 66 kW per rack
- All solutions liquid cooled
- 21 kW air cooled
- 45 kW liquid cooled
- 480V power

TLCC1 → TLCC2 → CTS-1 → CTS-2



Technology power density has reached a threshold where liquid cooling is a requirement for HPC platforms!

CTS-2 Architecture:
4th Generation Intel Sapphire Rapids Xeon CPUs

- Direct evolution of TLCC2 and CTS-1 CPUs
- Mature software solutions and ecosystem
- 2-socket servers with 56 CPU cores/socket
- 256 GB DDR5 per node / ~2.3 GB/core
- 8 channels DDR5/socket
- Same CPU as NNSA ASC Crossroads platform
- Smooth application transition from CTS-1 → CTS-2 → ATS-3 Crossroads
- BFloat16 and other new CPU features to support accelerated ML

Phase 1 Deployment Plan for CTS-2 Platforms

- First wave of LLNL Systems: 256 GB DDR5 Base Configuration

Lab	Name	# SU's	# Nodes	FP64 PetaFlops	Total Memory Capacity	Delivered to Lab	Est. General Availability	Program
LLNL	Mutt	1 SU	192	1.5 PF/s	49 TB	8/16/22	9/22	ASC (CZ)
LLNL	Whippet	< 1 SU	32	0.25 PF/s	8 TB	8/16/22	9/22	ASC (RZ)
LLNL	Poodle	< 1 SU	32	0.25 PF/s	8 TB	8/16/22	9/22	ASC (CZ)
LLNL	Hound	2 SU	384	3.0 PF/s	93 TB	9/27/2022	1QCY2023	ASC (RZ)
LLNL	Bengal	6 SU	1,152	9.1 PF/s	295 TB	9/27/2022	1QCY2023	ASC (SCF)
SNL	Amber	8 SU	1,536	12.1 PF/s	393 TB	11/8/2022	1QCY2023	ASC
LLNL	Dane	8 SU	1,536	12.1 PF/s	393 TB	12/20/2022	1QCY2023	ASC, Inst., PSAAP (CZ)
SNL	Stout	8 SU	1,536	12.1 PF/s	393 TB	1/31/2023	1QCY2023	Inst.

Additional CTS-2 systems deployment will depend on funding, power, cooling, and data center space permit.

CTS-2: Mutt, Whippet, & Poodle

Supply Chain Issues, Lease to Own, & Deployment Timeline

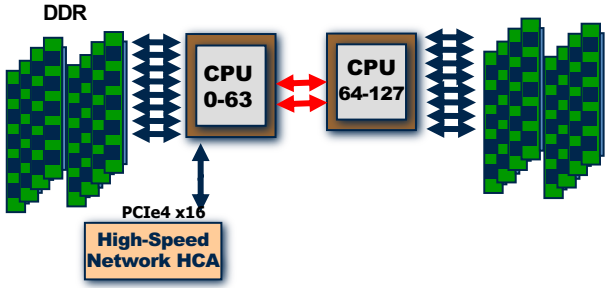
- Dell and CTS-2 components suppliers all report supply chain issues
  - Some components have 6-9 months lead time
  - These issues will likely continue through at least all of 2022, if not longer.
- Dell working with suppliers to mitigate risks as much as possible
  - NNSA supplying accurate forecasts for system orders will help.
  - Dell prioritizing CTS-2
- NNSA received a Defense Priority (DPAS) rating for CTS-2
  - Places CTS-2 systems on high priority list with system integrator and component suppliers.
  - CTS-2 will have a DX rating (perhaps DO for some systems)
  - El Capitan has a “DO” DPAS rating
  - DPAS will help, but the application of rating is on specific components & quantities – not entire systems

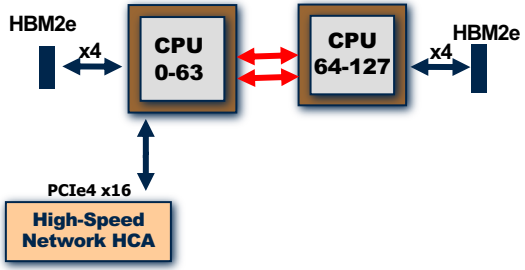
Supply Chain Issues, Lease to Own, & Deployment Timeline

- Ordering a new CTS-2 system takes time
  - 0-2 months for modifying build of materials for specific system and data center
  - 1-2 months for setting up lease to own (LTO)
  - Formally order system
  - 1-2 months for Dell to acquire parts (best case)
  - 1 month to build, test, and deliver to lab
  - At least 1 month to integrate, test, and stabilize system at Lab

- ***4 – 8*** months needed from time program decides to order system until system production

CTS-2 Market Inflection Point: CPU+DDR vs. CPU+HBM Price Parity





CPU + DDR5

- 2S ~64 core 2.0+ GHz CPU
- 16x32 GB DDR5 DIMM no chipkill
- Total memory 512 GB DDR5
- ~5+ GB/CPU core
- ~615 GB/s memory BW (theoretical)
- ~492 GB/s memory BW (actual?)

CPU + HBM

- 2S ~64 core 2.0+ GHz CPU
- 8x16 GB HBM2e no chipkill
- Total memory 128 GB HBM2e
- ~1.3 GB/ CPU core
- ~3,200 GB/s memory BW (theoretical)
- ~2,200 GB/s memory BW (actual?)

Is 3-4X the memory BW but ¼ the memory capacity a good trade-off?
Are ASC codes and capacity workloads ready and able to utilize the extra bandwidth?
Would at least a 30% performance improvement be worth ¼ of the memory capacity?

CTS-2 Next Steps

- CTS-2 platforms arriving in early to mid-2022
- More detailed information on CTS-2 will be available late summer 2021
- Early deployment experiences and performance in late summer 2022
- Questions?

Thanks to Tri-Lab CTS-2 Team!

- CTS-2 Tri-Lab Technical Leads
  - Matt Leininger (LLNL)
  - Trent D'Hooge (LLNL)
  - Randy Scott (SNL)
  - Steve Monk (SNL)
  - Amanda Bonnie (LANL)
  - Jim Lujan (LANL)
- Former members
  - Josip Loncaric (LANL)
  - Ian Karlin (LLNL)
  - Robin Goldstone (LLNL)
  - Ian Karlin (LLNL)
  - Jim Foraker (LLNL)
  - Jim Silva (LLNL)
  - Brian Ryjuin (LLNL)
  - Riyaz Haque (LLNL)
  - Clay England (LLNL)
  - Jeff Ogden (SNL)
  - Mike Glass (SNL)
  - Si Hammond (now NNSA HQ)
  - Doug Pase (SNL)
  - Phil Regier (SNL)
  - Brad Settlemyer (LANL)
  - Brett Holman (LANL)
  - Jesse Martinez (LANL)
  - Timothy Randles (LANL)
  - Galen Shipman (LANL)
  - Trent D'Hooge (LLNL)
  - Chris DePrater (LLNL)
  - Brandon Hong (LLNL)
  - Rigo Delgado (LLNL)
  - Lance Weems (LLNL)
  - Larry Baca (SNL)
  - Anthony Agelastos (SNL)
  - Jay Livesay (SNL)
  - Dave Martinez (SNL)
  - David Smith (SNL)
  - Eloy Romero Jr. (LANL)
  - Nathan Debardeleben (LANL)
  - Hai Ah Nam (LANL)
  - Howard Pritchard (LANL)
  - Mike Lang (LANL)
  - Gary Ward (LLNL)
  - Sofia Delgadillo-Marrufo (SNL)
  - Patrick Padilla (LANL)

Many additional subject matter experts provided reviews and feedback.

Questions?

CTS-2 Lead: Matt Leininger (matt@llnl.gov)

Cluster Integration Lead: Trent D'Hooge (dhooge1@llnl.gov)

CTS-2 Integration Lead: Jim Silva (silva50@llnl.gov)

Disclaimer
This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.