Commodity Technology Systems 2 (CTS-2) Update

June 2023

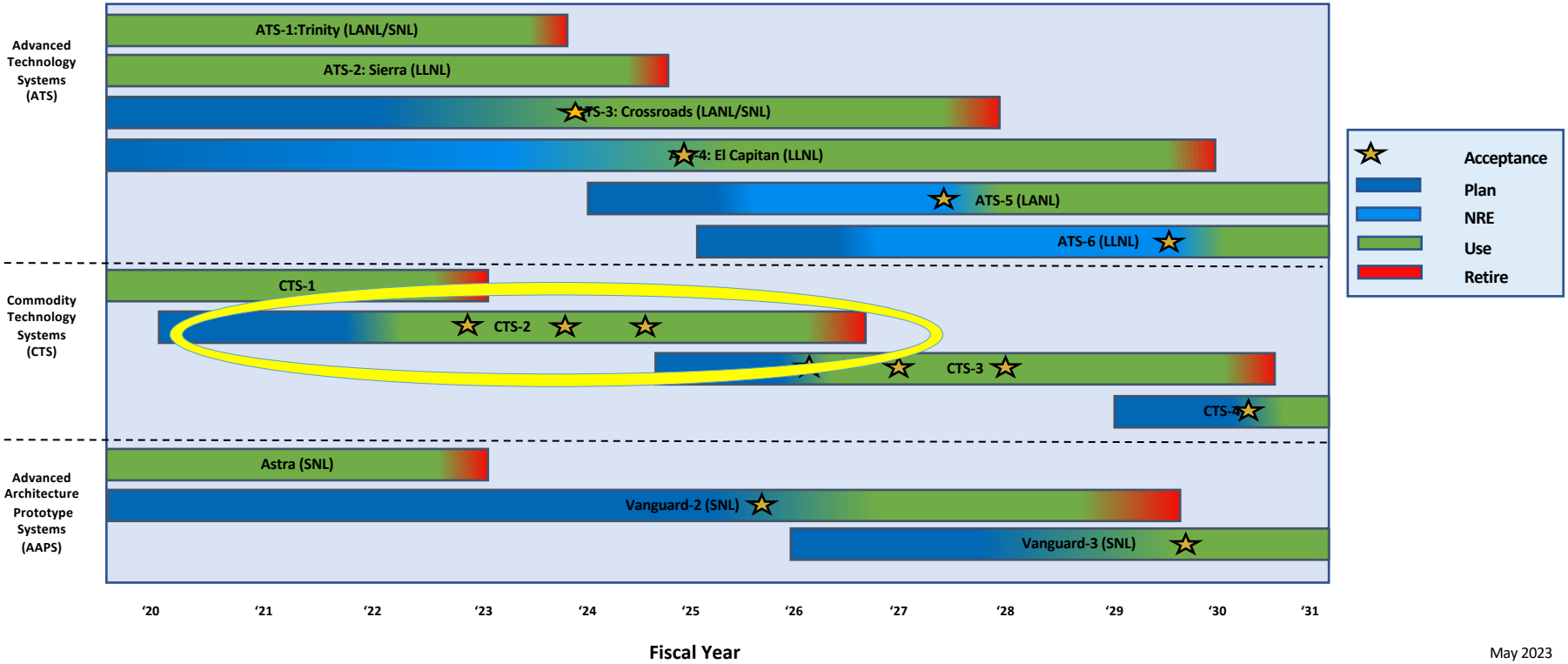
Matt Leininger
CTS-2 Technical Lead
Livermore Computing

LLNL-PRES-850876

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

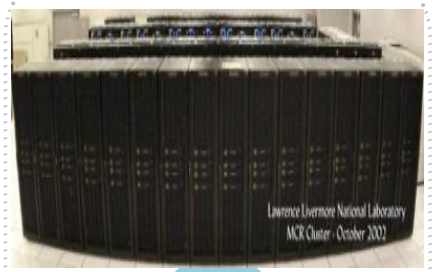


May 2023

Livermore Computing History with Linux Clusters

LLNL PCR

- Emperor & Adelie 2001
- Intel Pentium 4 Dual-socket
- 216 nodes (32-bit)
- Quadrics Elan3 Interconnect
- 1.5 TF/s



MCR

- M&IC Platform - 2002
- 1,152 nodes (32-bit)
- Intel Pentium 4 Dual-socket
- Quadrics Elan3 Interconnect
- 10 TF/s

Thunder

- M&IC Platform – 2004
- #2 on Top500 June 2004!
- 1,024 nodes (64-bit)
- Intel Itanium 2 Quad-socket
- Quadrics Elan4 Interconnect
- 20 TF/s

Peloton

- M&IC and ASC - 2005
- 80 – 1,152 nodes (64-bit)
- AMD Opteron Quad-socket
- InfiniBand DDR
- 120 TF/s (6 systems)

LLNL Leadership in Linux Clusters Enabled Future Platform Path for Commodity Technology Systems

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 Commodity Systems for NNSA

TLCC1

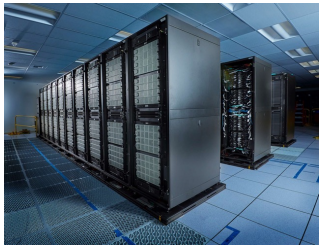
- Tri-Lab Linux Capacity Clusters
- 1st joint procurement
- 2007–2010
- 11,000 CPU sockets
- 0.4 PF/s

TLCC2

- Tri-Lab Linux Capacity Clusters
- 2nd joint procurement
- 2011–2015
- 10,700 CPU sockets
- 1.8 PF/s

CTS-1

- Commodity Technology Systems
- 3rd joint procurement
- 2016–2021
- 21,000 CPU sockets
- 18 PF/s



CTS-2

- Commodity Technology Systems
- 4th joint procurement
- 2022–2026
- Likely > 22,000 CPU sockets
- Likely > 70 PF/s

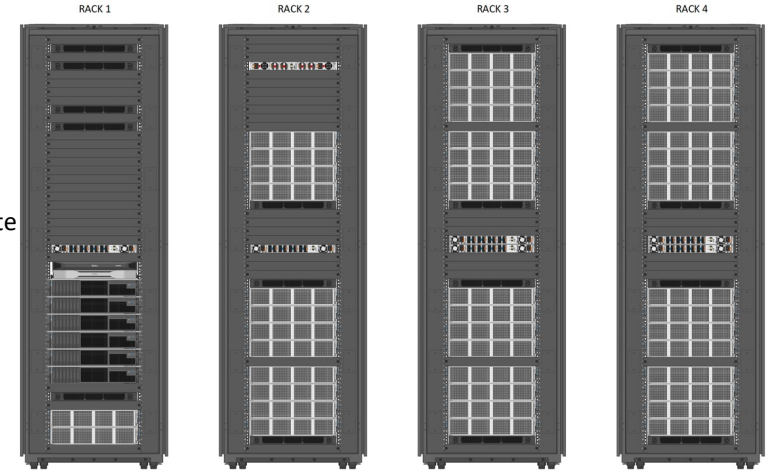
Commodity platforms have successfully delivered to ASC & Institutional simulation programs 16+ years

CTS-2 Contract Awarded to Dell Technologies

- Awarded September 2021
- Compute Node Servers: Dell C6620 (4 nodes in 2U)
- Login/Mgmt/Gateway Servers: Dell 760 2U
- Intel Sapphire Rapids Xeon CPUs
  - With 256 GB DDR5 memory
- HPC Network:
  - Cornelis Networks Omni-Path or Mellanox InfiniBand
- 66-70kW/rack; direct-to-chip liquid cooling; 480V 3-Phase
- Deliveries started in September 2022
- Options for GPUs, large memory, HBM, SSDs, etc.
- TOSS 4 based on RHEL 8.x will provide basis for common Tri-Lab environment

C6620 Compute Node

1 Scalable Unit CTS-2 Rack Layout

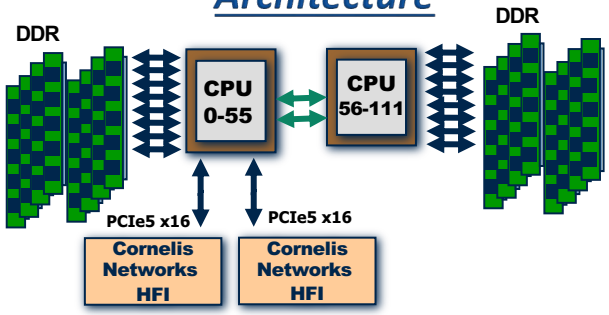


C6620 Chassis

R760 Chassis

CTS-2 Performance Results Show Substantial Improvements over CTS-1 (circa 2016)

CTS-2 Base Node Architecture



CPU + DDR5

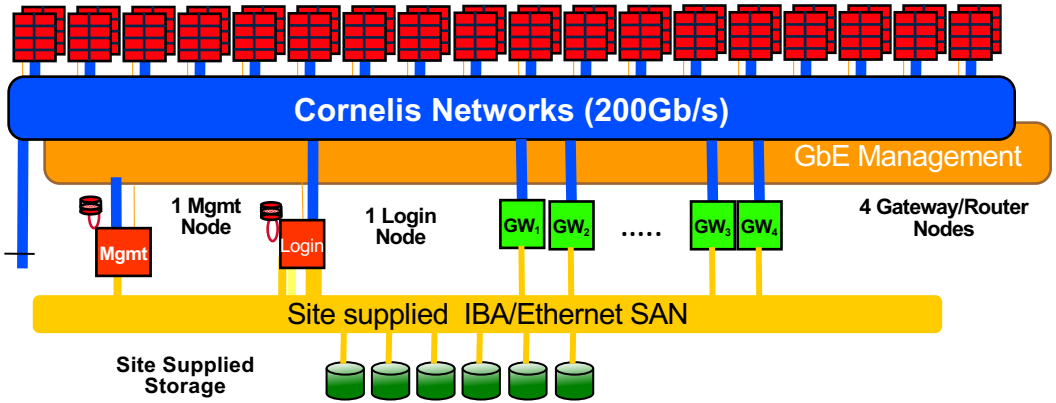
- 4th Gen Intel Xeon Scalable processor
- 2 socket x 56 cores
- 16x16 GB DDR5 1R DIMM
- 256 GB DDR5 total node memory
- 2.3 GB mem capacity/CPU core

4th Generation Intel Scalable Xeon - 56C	Overall Node FOM	HPCG (Linear Solver)	LAGHOS (Hydro)	Quicksilver (Monte Carlo)	SNAP (S_n Transport)
Speed up relative to Intel E5-2695v 18C	4.6x	3.5x	5.9x	5.4x	4.6x

4th Generation Intel Scalable Xeon - 56C	DGEMM (per socket)	STREAM (per Socket)	STREAM (single core)	HPL (per node)	HPL (128 nodes)
256GB DDR5 Single-rank DIMMs	2.8 TF/s	210 GB/s	21 GB/s	6.9 TF/s	866 TF/s

CTS-2 Scalable Unit Architecture

CTS-2 Scalable Unit (SU)

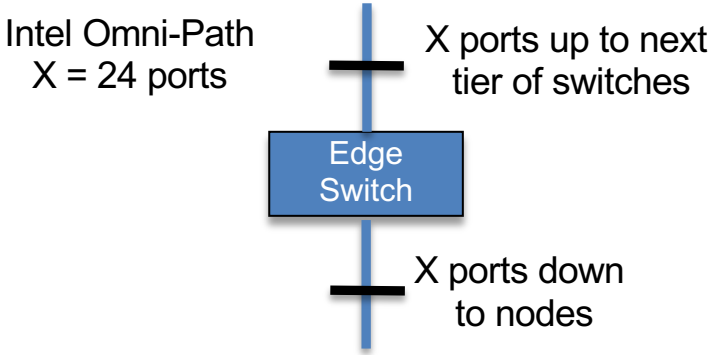


CTS-2 Scalable Unit (SU)			
Nodes	CPU Cores	Memory Capacity	Target Theoretical Peak (FP64)
196 Total 190 Compute	~21.5K	**DDR5: 49 TB** DDR5: 98 TB HBM2e: 25 TB	~1.4 PF/s

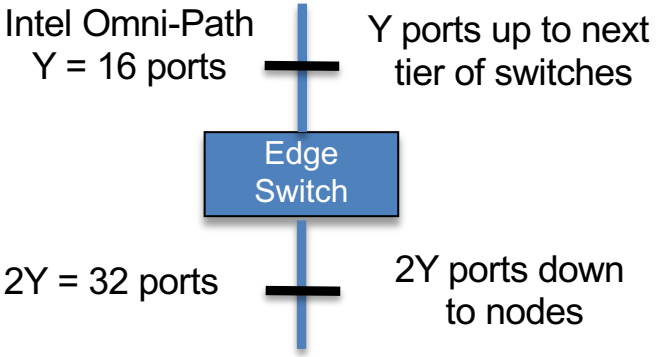
- Initial system delivered in September 2022
- Based on Dell C6620 & 760 servers
- 4th Generation Intel Xeon Scalable CPU
  - 56 cores/socket or 112 cores/node
  - **Default: 256 GB DDR5**
  - Alternative: 512 GB DDR5
  - Alternative: 128 GB HBM2e
- Cornelis Networks High Speed Network
  - Omni-Path 2x100 Gb/s
- CoolIT direct-to-chip liquid cooling
- 480V 3-phase power
- Software Environment
  - TOSS4 based on RHEL 8.x
  - Tri-Lab Common Environment (TCE2)

CTS-2 Architecture: Network Topology

Optional CTS-2 Configuration
Full Fat Tree Network



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

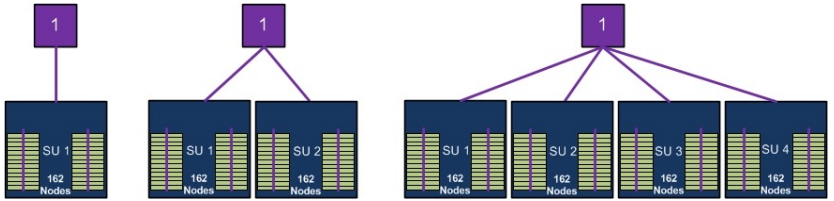


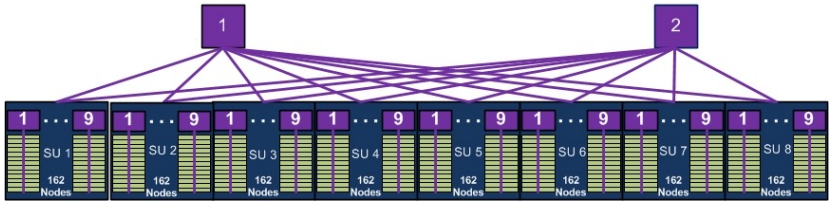
Maximize small/medium job throughput with 2:1 Tapered Network and more compute nodes!

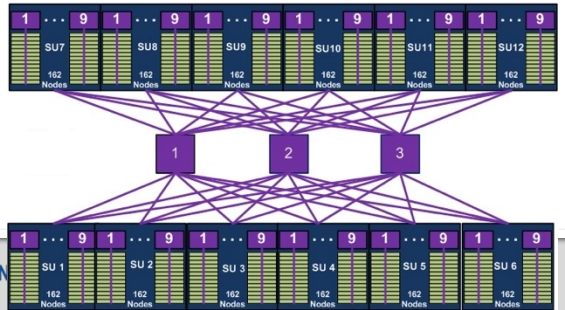
Scalable Units are the “Legos” that scale to Multi-SU platforms

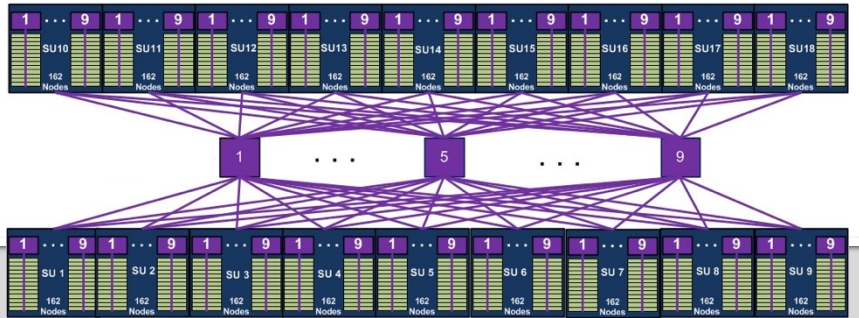
#SU’s	# Nodes	PFlops
1	192 - 200	1.45 – 1.5
2	384 - 400	2.9 – 3.0
4	768 - 800	5.8 - 6
6	1,152 – 1,200	8.7 – 9.0
8	1,536 – 1,600	11.6 – 12.0
12	2,304 – 2,400	17.3 – 18.0

Many different system sizes can be deployed depending on programmatic needs







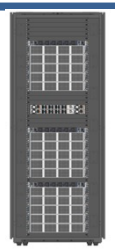


Evolution of Power & Cooling Requirements for Commodity Systems

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

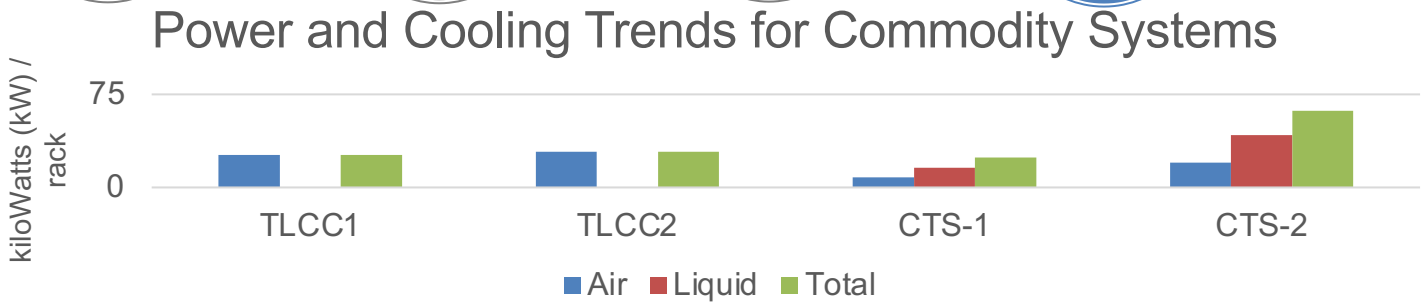
- 28 kW per rack
- 28 kW air cooled
- Fielded 1 liquid cooled solution
- 208V power

- 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 CoolIT Direct-to-Chip Liquid Cooling

CoolIT CHx750 CDU (Cooling Distribution Units)

CoolIT CHx80 CDU

Dell C6620 Compute Node

CTS-2 Deployments @ LLNL

- Mutt (TOSS Testbed)
- RZWhippet serial cluster
- Poodle (CZ) serial cluster
- 48 HBM nodes are being added July 2023

- ASC Bengal 6 SU
- Delivered; Power-on in July 2023
- LLNL Dane 8 SU
- Delivered; Power-on June 2023

CTS-2 Phase 1 Deployment Plan
Sept. 2022 - April 2023

Lab	System Name	System Size	# Nodes	FP64 PetaFlops	Total Memory Capacity	Program
LLNL	Mutt	~1 SU	145	1.5 PF/s	49 TB	ASC (Testbed)
LLNL	RZWhippet	1 rack	41	0.25 PF/s	8 TB	ASC
LLNL	Poodle	1 rack	41	0.25 PF/s	8 TB	ASC
LLNL	RZHound	2 SU	386	3.0 PF/s	93 TB	ASC
LLNL	Bengal	6 SU	1,170	9.1 PF/s	295 TB	ASC
SNL	Amber	8 SU	1,544	12.1 PF/s	393 TB	ASC
LLNL	Dane	8 SU	1,544	12.1 PF/s	393 TB	ASC, Inst., PSAAP
SNL	Stout	8 SU	1,544	12.1 PF/s	393 TB	Inst.

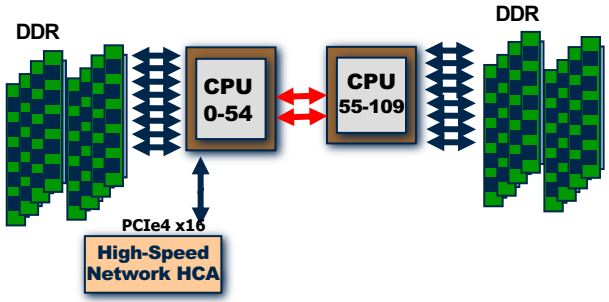
First Wave of CTS-2 Platforms will be Production in 2023 (late summer)

CTS-2 May-Dec. 2023 Deployment Plan

Lab	System Name	System Size	# Nodes	FP64 PetaFlops	Total Memory Capacity	Program
LLNL	Mutt	+48 HBM nodes	193	1.4 PF/s	37 TB (DDR5) 6 TB (HBM2e)	ASC (Testbed)
UoRochester / LLE	-	2 SU	396	2.8 PF/s	203 TB	Lab for Laser Energetics
Naval Nuclear Lab	-	2 SU	396	2.8 PF/s	203 TB	NNL Programs
Sandia	-	4 SU	344+	2.8 PF/s	88 TB	Institutional
Sandia	-	-	50	12 PF/s (GPU)	16 TB (HBM3)	Institutional
Idaho National Lab	-	2 SU	396	2.8 PF/s	102 TB	INL Programs
KCNSC	-	1 SU	198	0.9 PF/s	51 TB	KCNSC (outside CTS-2)
LLNL	-	1 SU	198	0.9 PF/s	51 TB	GS
LLNL	-	1 SU	198	1.4 PF/s	51 TB	GS
Sandia	-	6 SU	1,170	9.1 PF/s	295 TB	GS
LLNL	-	~3 SU	594	4.2 PF/s	153 TB	GS

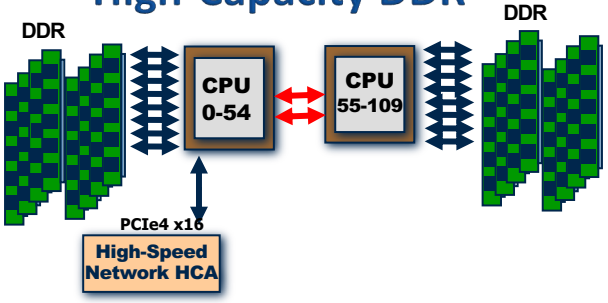
CTS-2 Memory Options: Traditional DDR or High-Bandwidth Memory

CTS-2 Base Architecture



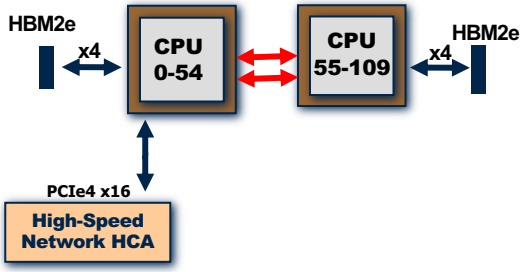
CPU + DDR5
- 2S 56 core 2.0+ GHz CPU
- 16x16 GB DDR5 DIMM no chipkill
- Total memory 256 GB DDR5
- ~2.2-2.4 GB/CPU core
- 615 GB/s memory BW (theoretical)
- 440 GB/s memory BW (actual)

CTS-2 Architecture: High-Capacity DDR



CPU + DDR5
- 2S 56 core 2.0+ GHz CPU
- 16x32 GB DDR5 DIMM no chipkill
- Total memory 512 GB DDR5
- ~4.4-4.8 GB/CPU core
- 615 GB/s memory BW (theoretical)
- 500 GB/s memory BW (actual)

CTS-2 Architecture: High-Bandwidth Memory

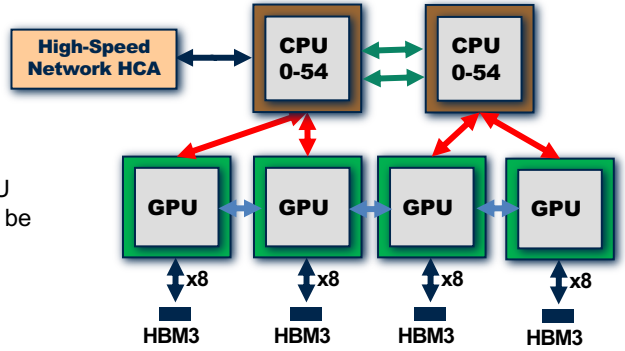


CPU + HBM
- 2S 56 core 2.0+ GHz CPU
- 8x16 GB HBM2e no chipkill
- Total memory 128 GB HBM2e
- ~1.3 GB/ CPU core
- 3,200 GB/s memory BW (theoretical)
- 2,000 GB/s memory BW (actual)

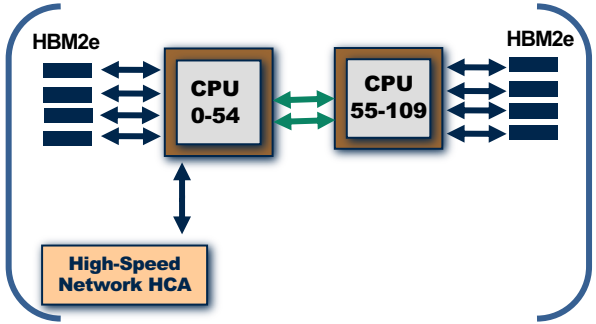
Is 3-4X increase in memory BW but ½-¼ the memory capacity a good trade-off?
Are ASC codes and capacity workloads ready and able to utilize the extra bandwidth?
Would at least a 30% performance improvement be worth ½-¼ of the memory capacity?

CTS-2 Architecture Performance Tradeoffs: GPUs vs. CPU+HBM

Unified CPU/GPU memory may not be available.



= ~10x



Processor	DDR5 Capacity	HBM Capacity	GB/core	Theoretical Memory Bandwidth	Actual Memory Bandwidth	Theoretical Peak (FP64)
CPU+GPU	256 GB	320 GB	-	16 TB/s	14 TB/s	***240 TF/s***
CPU+HBM	**-**	***1,280 GB***	***1.1-1.2***	***32 TB/s***	***20 TB/s***	72 TF/s

For given cost, CPU+HBM provides more HBM memory capacity & performance.

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
  - 2-3 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
  - 6-15 months for facilities preparation

- ***6-15*** months needed from time program decides to order system until system production

CTS-2 Summary

- First “large” LLNL CTS-2 platforms:
  - Target production late summer/early fall
  - Currently being integrated & tested or waiting for facilities power
- Future CTS-2 purchases could include:
  - Base CPU + DDR (solid foundation for entire LLNL application portfolio)
  - CPU + HBM (good for memory bandwidth bound codes)
  - CPU + GPU (no unified memory; HPC+AI/ML)
- Need input from LLNL programs & User community

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)


Lawrence Livermore
National Laboratory