New ASC Magma System Overview and CTS-2 Update

Ian Karlin and Matt Leininger

November 14th, 2019

LLNL-PRES-798672

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

Outline

- ASC Magma “CTS-1+” Cluster
  - Technologies considered for Magma
  - Magma System Architecture & Timeline
  - Comparison to ASC Jade CTS-1
  - Performance Estimates

- CTS-2 Update
  - Current status & Timeline
  - Potential Architectures and Cost Trade-offs (Complex Landscape)
  - Current and Future ASC Workloads
  - Mapping workloads to Architectures (More challenging than past CTS)

Why Magma?

- Growing demand for capacity cycles across the ASC Program

- Decision Factors for Magma
  - Cost/performance & schedule
  - Compatibility with CTS-1 architecture (ease user transition)
  - Cluster integration/admin concerns
  - Liquid cooling solutions

- What technologies are viable for late 2019?
  - Intel Broadwell CPUs + Omni-Path Interconnect (End of Life)
  - Intel Skylake CPU’s + Omni-Path Interconnect
  - Marvell ARM64 ThunderX2 + Mellanox HDR
  - AMD Rome CPU’s + Mellanox HDR Interconnect
  - Intel Cascade Lake-SP + Omni-Path Interconnect
  - Intel Cascade Lake-AP + Omni-Path Interconnect

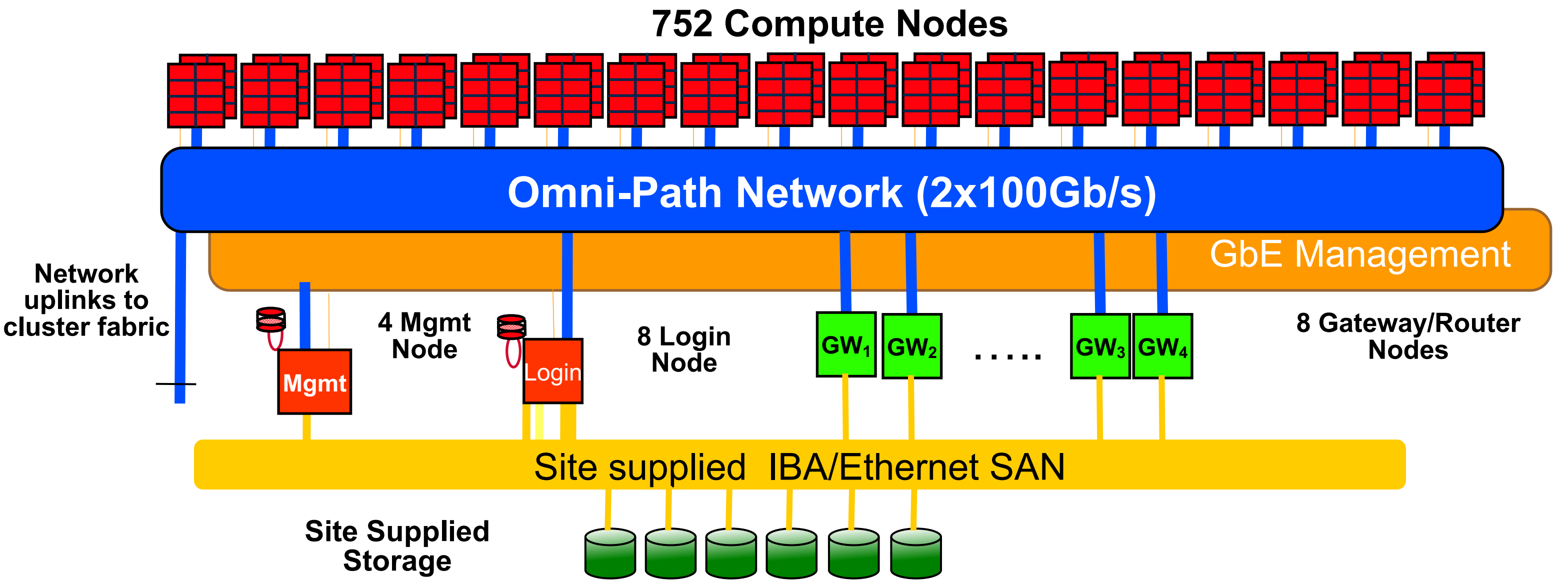
- Selected Intel Cascade Lake AP (CLX-AP) + Dual Omni-Path Interconnect

Magma is a Next-Generation CTS-1 System for the LLNL ASC Program

- 4 Scalable Units
- Intel Cascade Lake AP based nodes
- Intel Servers (4 node in 2U)
- CoolIT direct liquid cooling to CPUs and DIMMs – certified by Intel
- Dual-Rail Omni-Path Interconnect
- TOSS 3.5-x (same version as Jade)
- Delivered Nov 2019
- Operational Jan 2020

Magma 1st racks @ Penguin

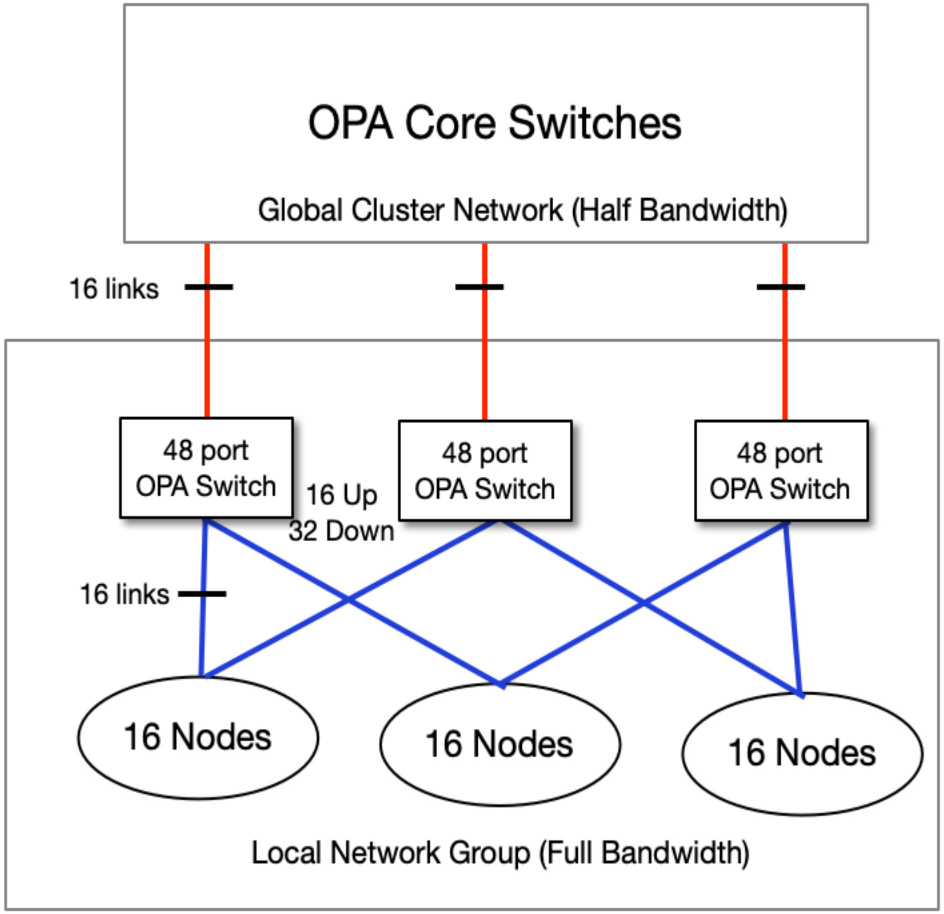
Magma Cluster Design



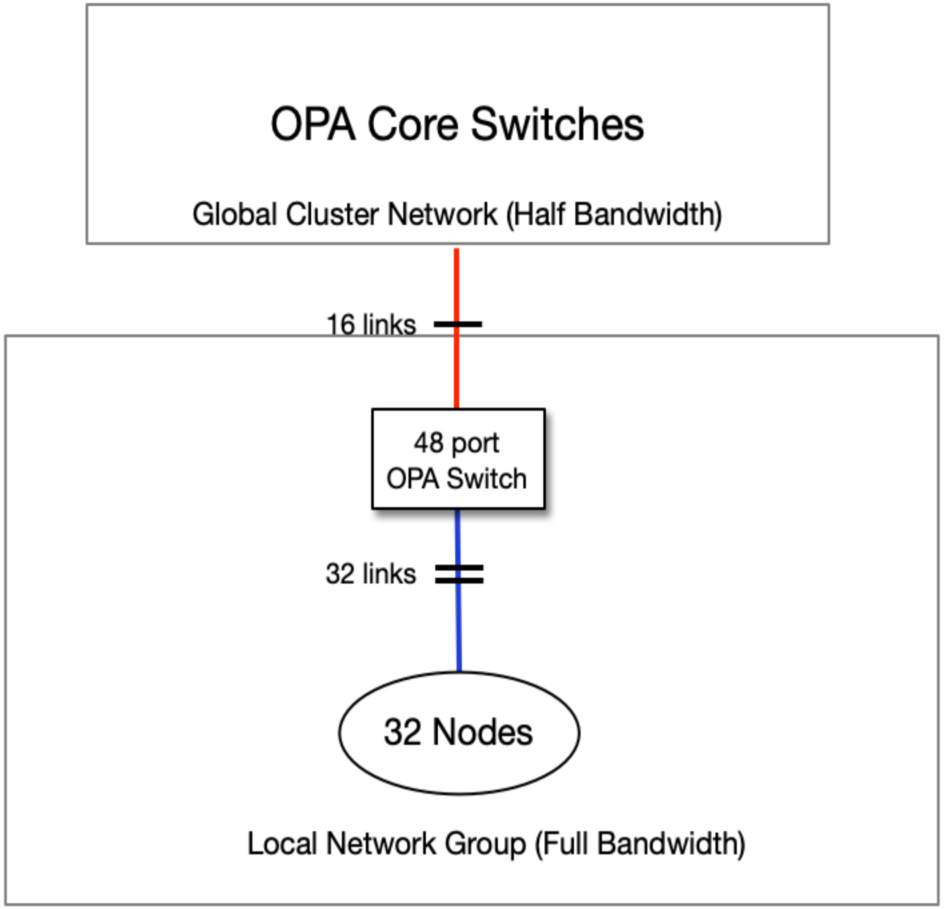
Magma Parameters (772 total nodes; 752 compute; 8 GW; 8 Login; 4 Mgmt)

- CLX-AP compute and login nodes
- CLX-SP gateway and management nodes
- Dual socket nodes; Total memory capacity 294 TB; 431 TB/s memory bandwidth
- 4 GB memory capacity per CPU core
- 5.6 PF/s theoretical peak FP64
- Over 73K cores

Magma Dual Rail Network Enables Large Local Groups



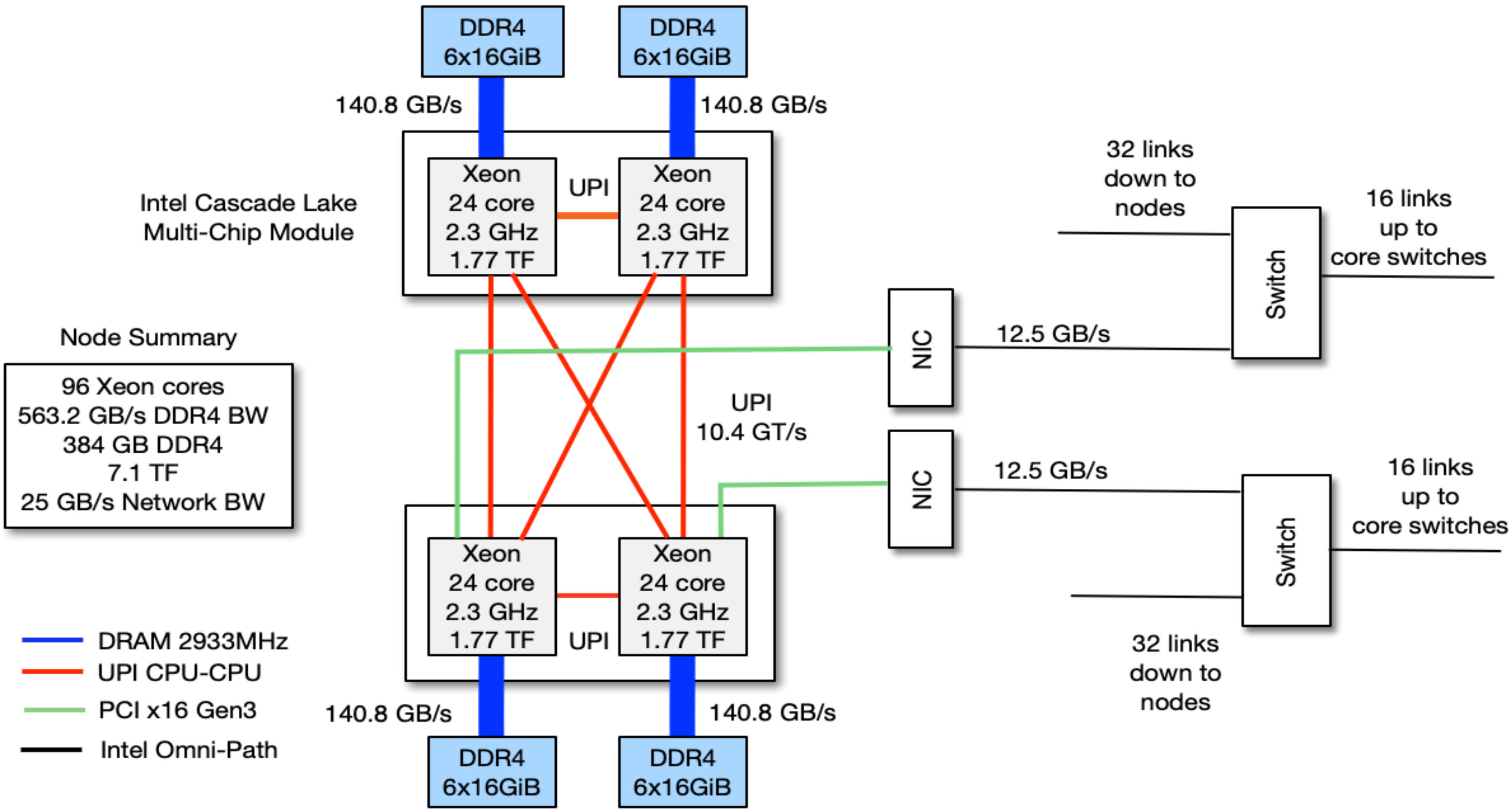
Magma Local Group:
48 nodes; 354 TF/s; 27 TB/s; 18.4 TB


OPA Core Switches
Global Cluster Network (Half Bandwidth)
16 links
48 port OPA Switch
32 links
32 Nodes
Local Network Group (Full Bandwidth)

CTS-1 Local Group:
32 nodes; 39 TF/s; 4.9 TB/s; 4 TB

Magma Node Level Architecture

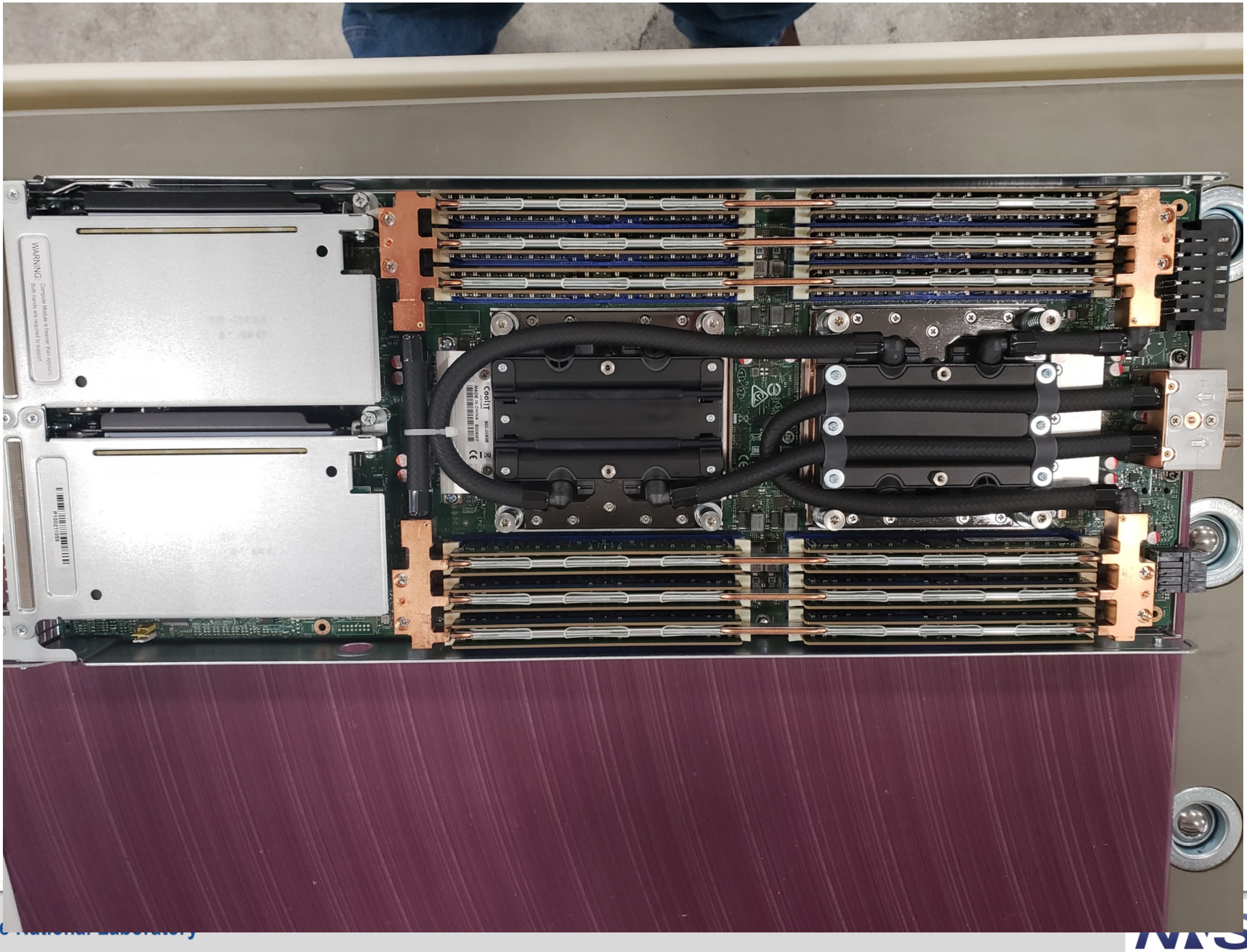
Magma Compute and Login Nodes:
Intel Xeon Cascade Lake Advanced Processor (CLX-AP)



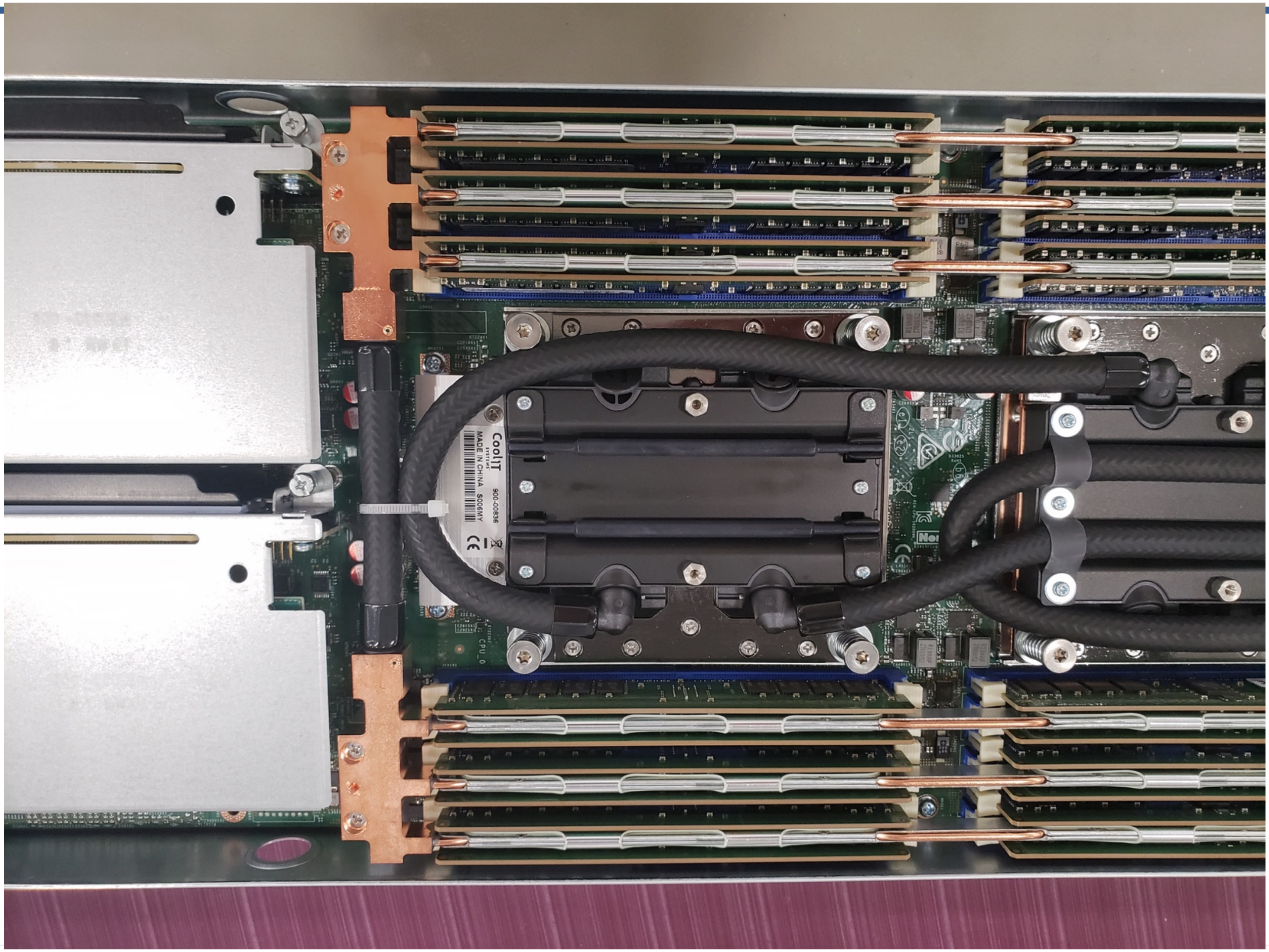
Node Summary

96 Xeon cores
563.2 GB/s DDR4 BW
384 GB DDR4
7.1 TF
25 GB/s Network BW

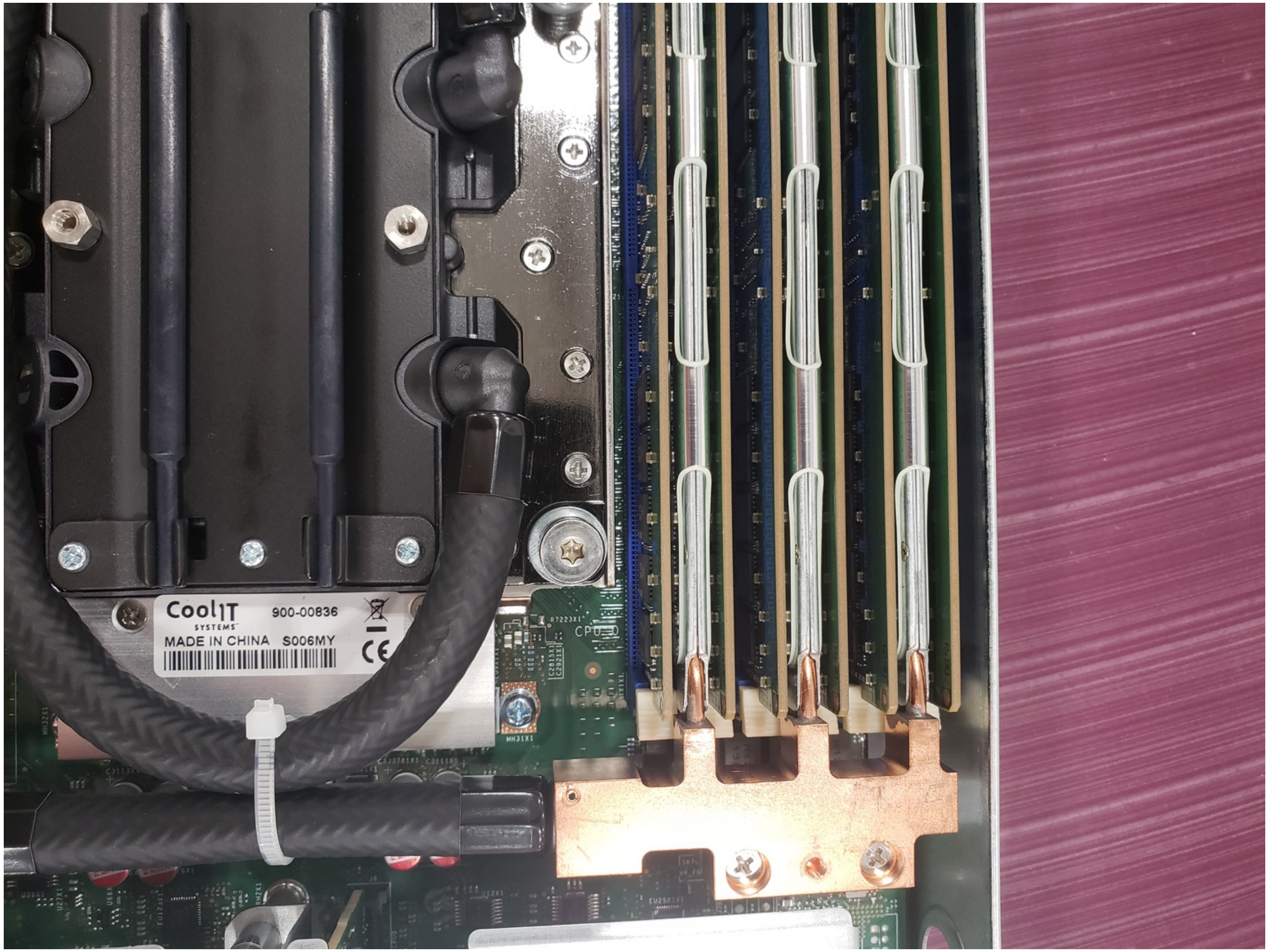
Magma Node Blade


WARNING
Compute Module is heavier than appears.
Both hands are required to support.
CoolIT

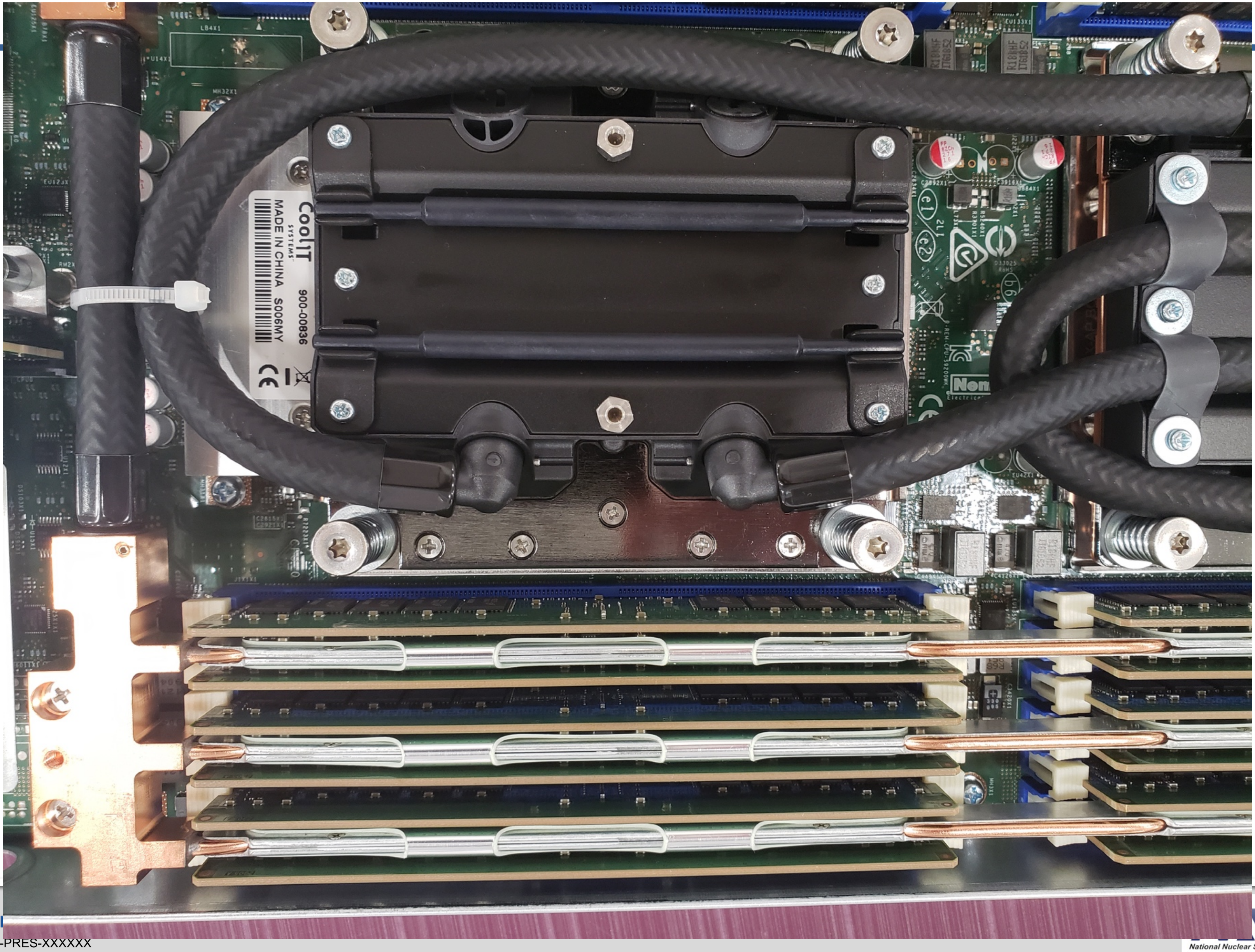
Magma Node Blade



Magma Node Blade



Magma Node Blade



Magma is a Significant Increase in ASC Capacity Computing

System	#Nodes	# Cores	Clock Rate (GHz)	PF/s
Magma	772	73.5K	2.3	5.3
Jade	2,688	96.8K	2.1	3.3

System	Memory Capacity (TB)	Memory Bandwidth (TB/s)	Network Injection BW (GB/s)	Network Bisection BW (TB/s)
Magma	294	431	25	9.6
Jade	344	413	12.5	16.8

Magma is nearly equivalent to another Jade system!

Questions on Magma?

Update on Commodity Technology Systems (CTS-2) Procurement

Ian Karlin and Matt Leininger

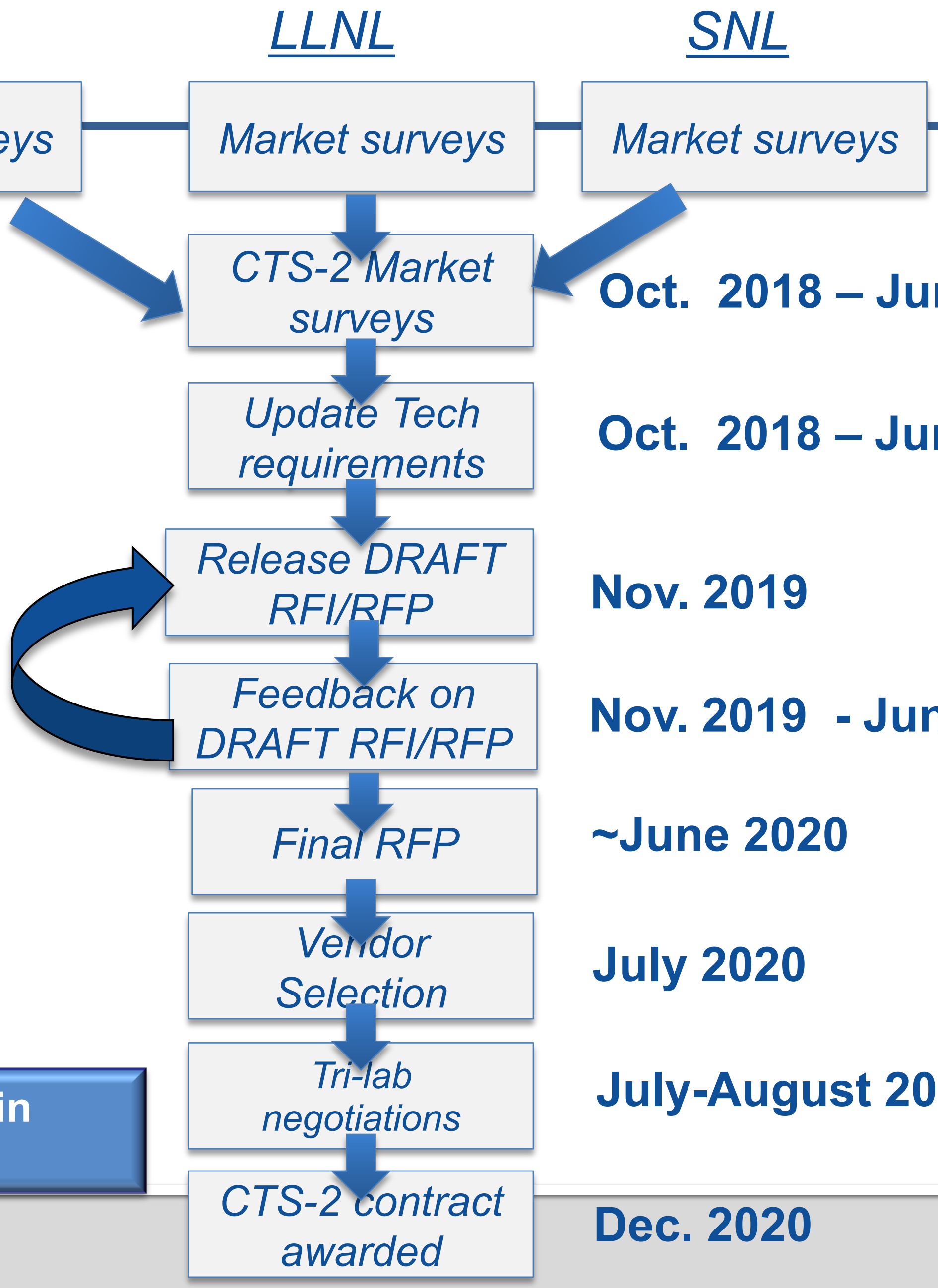
October 23, 2019

LLNL-PRES-798672

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

CTS-2 Activities Leading to RFP and Contract



LANL — *Market surveys*

LLNL — *Market surveys*

SNL — *Market surveys*

2018-2020

CTS-2 Market surveys — **Oct. 2018 – June 2020**

Update Tech requirements — **Oct. 2018 – June 2020**

Release DRAFT RFI/RFP — **Nov. 2019**

Feedback on DRAFT RFI/RFP — **Nov. 2019 - June 2020**

Final RFP — **~June 2020**

Vendor Selection — **July 2020**

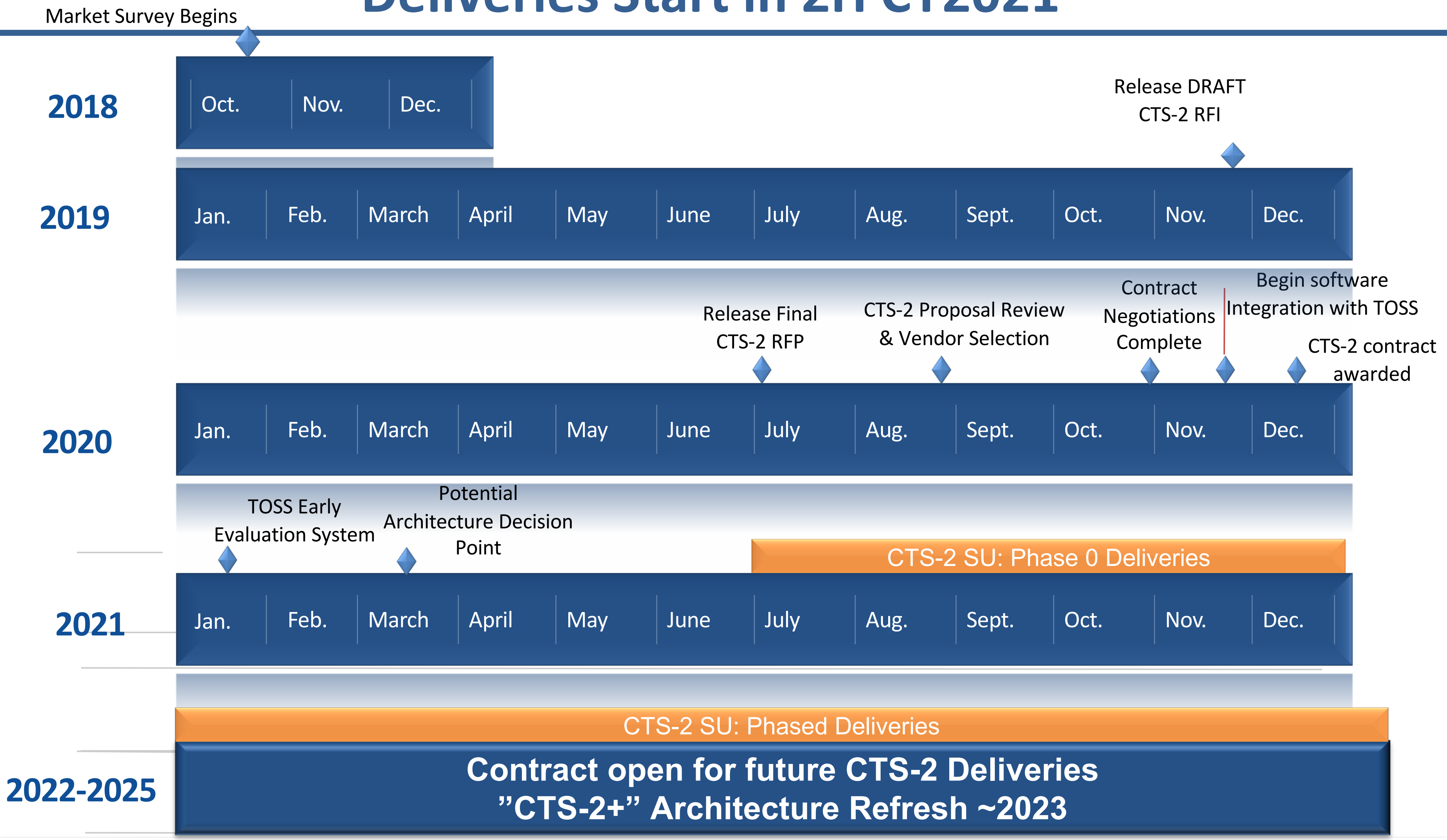
Tri-lab negotiations — **July-August 2020**

CTS-2 contract awarded — **Dec. 2020**

CTS-2 and TOSS teams continue to work together during CTS-2 deployment & lifetime support

Initial system deliveries in 2H CY2021

CTS-2 Procurement Timeline: Deliveries Start in 2H CY2021



Market Survey Begins

2018

Oct. Nov. Dec.

Release DRAFT CTS-2 RFI

2019

Jan. Feb. March April May June July Aug. Sept. Oct. Nov. Dec.

Release Final CTS-2 RFP

CTS-2 Proposal Review & Vendor Selection

Contract Negotiations Complete

Begin software Integration with TOSS

CTS-2 contract awarded

2020

Jan. Feb. March April May June July Aug. Sept. Oct. Nov. Dec.

TOSS Early Evaluation System

Potential Architecture Decision Point

CTS-2 SU: Phase 0 Deliveries

2021

Jan. Feb. March April May June July Aug. Sept. Oct. Nov. Dec.

CTS-2 SU: Phased Deliveries

2022-2025

Contract open for future CTS-2 Deliveries
”CTS-2+” Architecture Refresh ~2023

Multiple CPU Architectures are Viable for CTS-2

- Intel Xeon CPUs
  - Direct evolution of TLCC2 and CTS-1 CPUs
  - Mature software solutions and ecosystem
- AMD Epyc CPUs
  - Emerging as strong competitor to Intel Xeon
  - Many x86_64 tools already work
  - Compilers are still maturing with respect to performance
- Marvell ThunderX or Fujitsu (ARM64)
  - Also emerging as a strong competitor
  - Tools, compilers, and other software is still maturing
  - Learn from SNL Astra and LLNL ARM testbed experiences
- Power10
  - Typically supports very large memory per core!

Processor architecture & software readiness will be one key aspect of any CTS-2 selection

Data that will help us spend our money more effectively.

- What fraction of the workload today is 2D vs. 3D?
  - This is not number of jobs, but rather cycles used
  - Are there node counts below, which all jobs are 2D and above which all are 3D?
- Where do you think you are going in the future?
  - Past data shows job sizes are using about 2x the compute 4 years later. Is this 2x larger jobs or more strong scaled workloads?
- Are you more concerned about improving time to solution, total throughput, or something else?
  - Different designs will have different tradeoffs for each, some of which is workload dependent.
- Do you have a good understanding of typical zone counts or other parameters that would help us out?

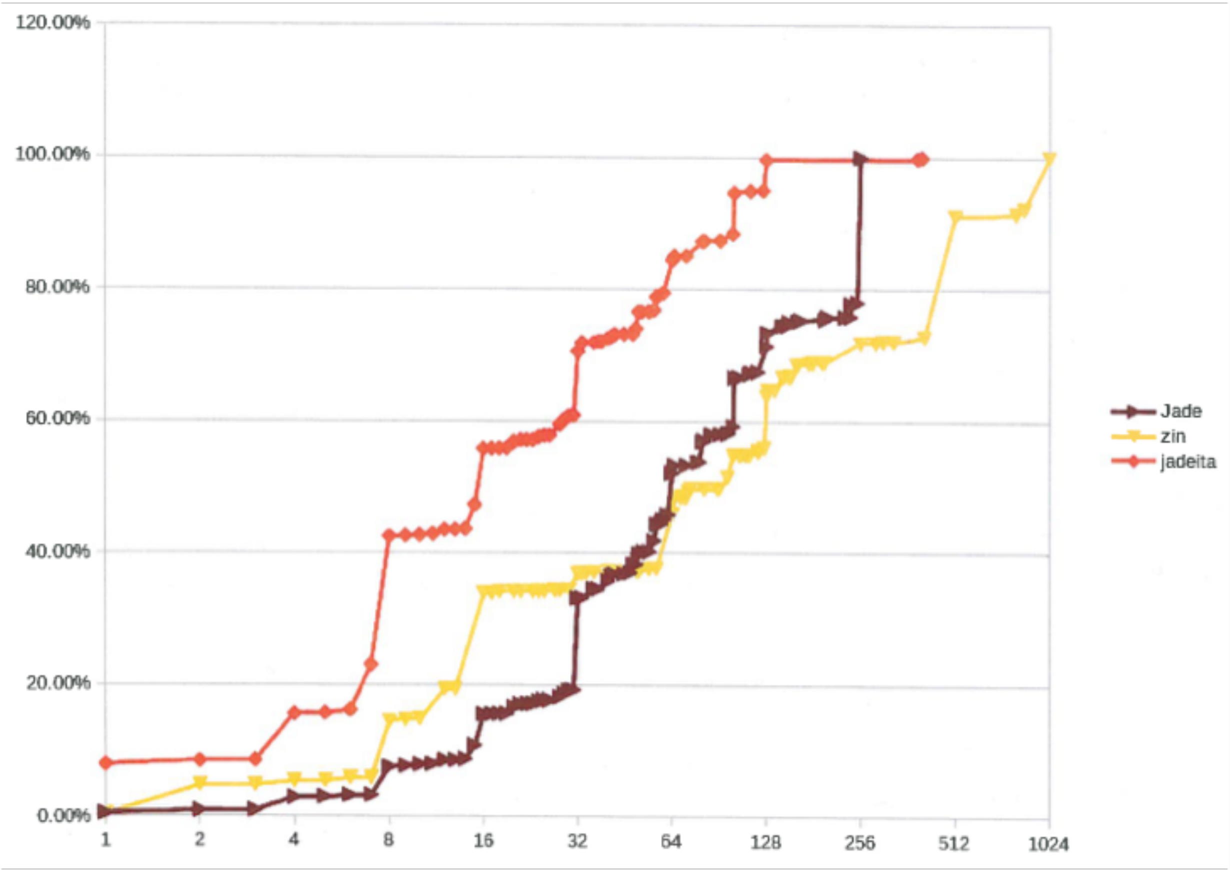
Next Steps

- Our goal is to provide the ASC program with sufficient data to make an informed decision on CTS-2 architectures
- Looking for your feedback
- Willing to meet with smaller groups/code teams/users
- We need to begin collecting data soon
- We can leverage LC data on job sizes, but need more
- Matt Leininger (matt@llnl.gov)
- Ian Karlin (karlin1@llnl.gov)

What are your biggest concerns for CTS-2?

Backup Slides

Machines are used differently are the workloads different?



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.

Lawrence Livermore
National Laboratory