Performance Portability and Intel

Ian Karlin

November 20th, 2019

Lawrence Livermore National Laboratory

Magma is a Next-Generation CTS-1 System for the LLNL ASC Program and our third top 100 Intel Machine

- 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
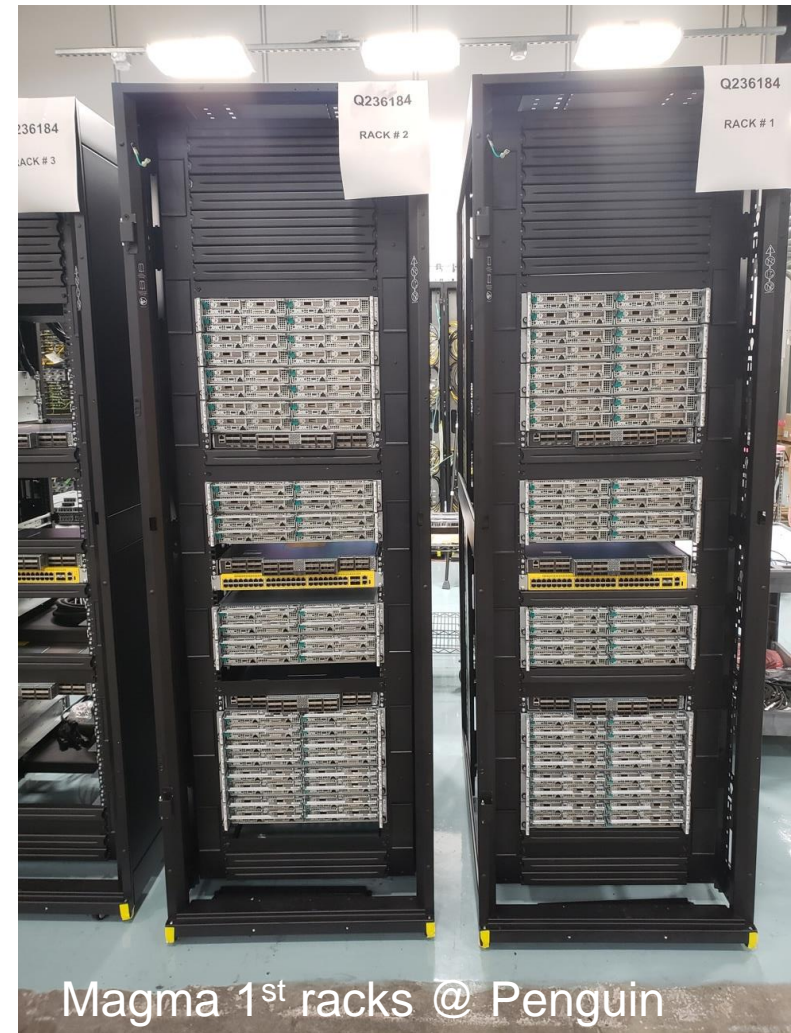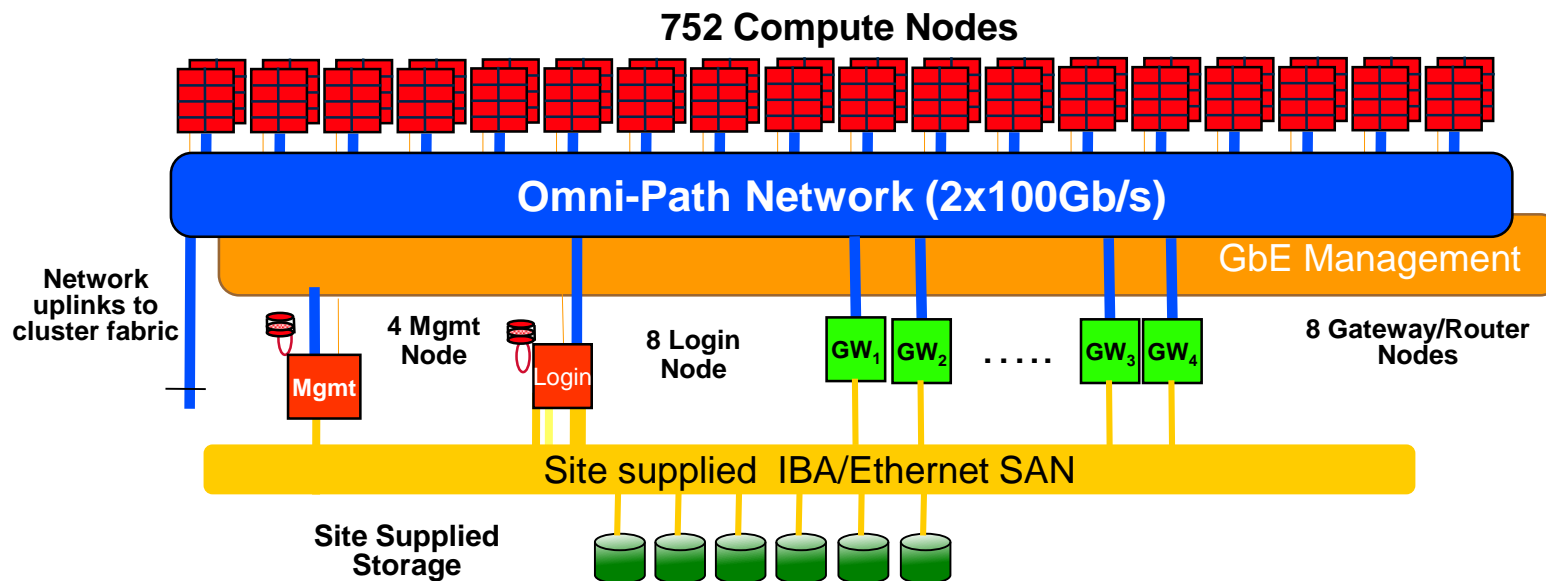

Magma 1st racks @ Penguin

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!

Magma Cluster Design

752 Compute Nodes

Omni-Path Network (2x100Gb/s)

GbE Management

Network uplinks to cluster fabric

4 Mgmt Node

Mgmt

Login

8 Login Node

GW₁ GW₂ . . . . . GW₃ GW₄

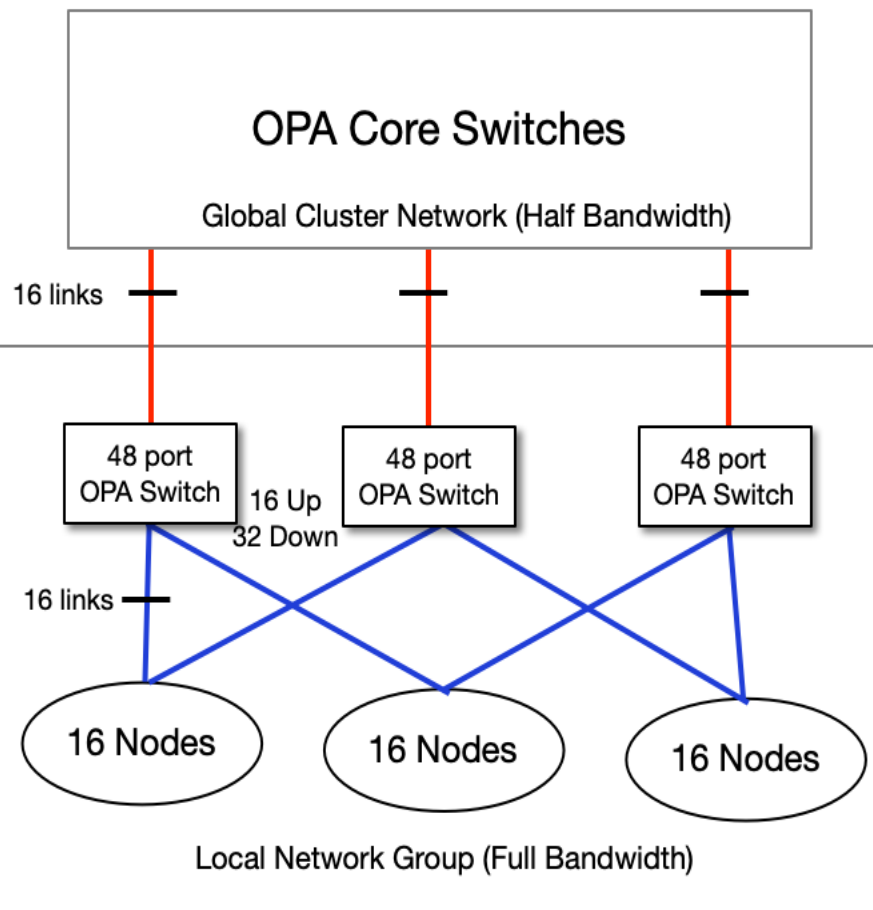8 Gateway/Router Nodes

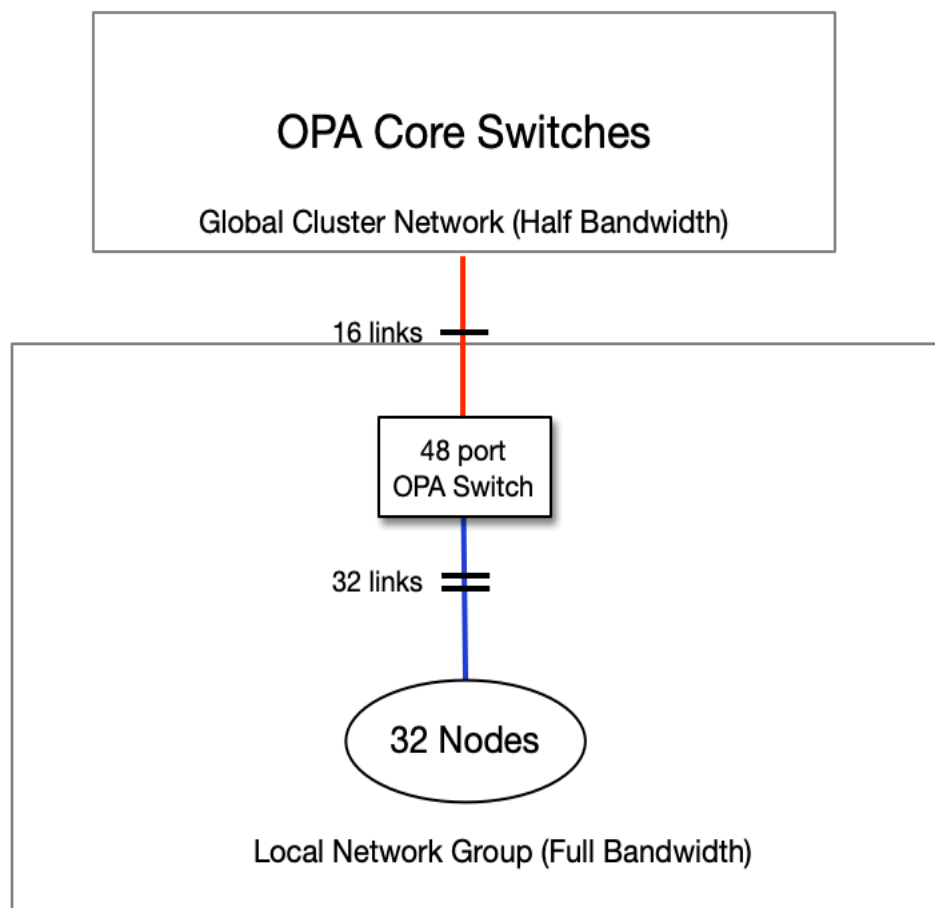Site supplied IBA/Ethernet SAN

Site Supplied Storage

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
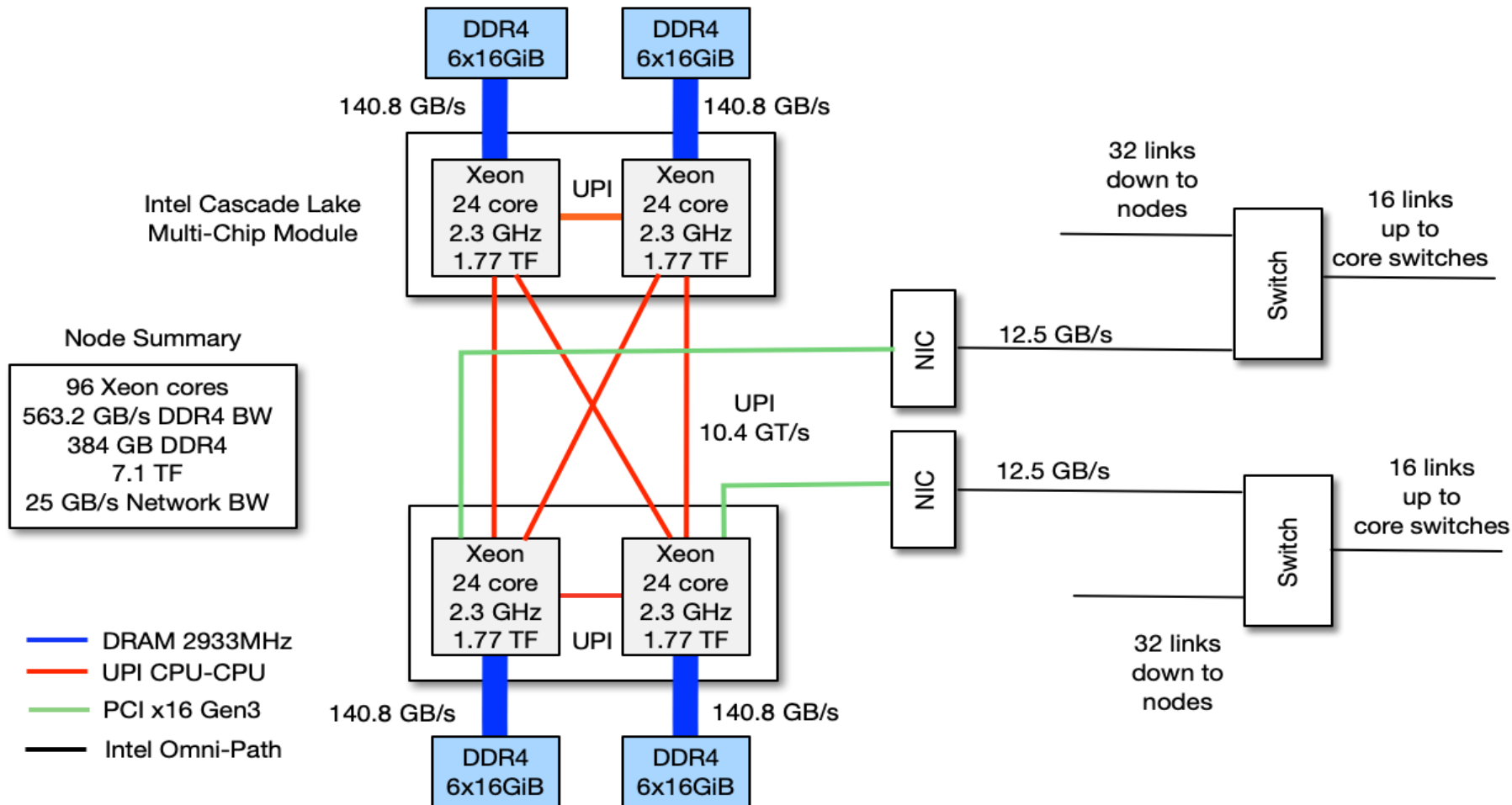
Magma Dual Rail Network Enables Large Local Groups

OPA Core Switches

Global Cluster Network (Half Bandwidth)

16 links

48 port OPA Switch — 16 Up 32 Down — 48 port OPA Switch — 48 port OPA Switch

16 links

16 Nodes — 16 Nodes — 16 Nodes

Local Network Group (Full Bandwidth)

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

Lawrence Livermore National Laboratory
LLNL-PRES-XXXXXX

Magma Node Level Architecture

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

DDR4 6x16GiB

140.8 GB/s   140.8 GB/s

Intel Cascade Lake Multi-Chip Module

Xeon 24 core 2.3 GHz 1.77 TF — UPI — Xeon 24 core 2.3 GHz 1.77 TF

32 links down to nodes

16 links up to core switches

Switch

NIC   12.5 GB/s

Node Summary

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

UPI 10.4 GT/s

NIC   12.5 GB/s

16 links up to core switches

Switch

Xeon 24 core 2.3 GHz 1.77 TF   UPI   Xeon 24 core 2.3 GHz 1.77 TF

32 links down to nodes

DRAM 2933MHz
UPI CPU-CPU
PCI x16 Gen3
Intel Omni-Path

140.8 GB/s   140.8 GB/s

DDR4 6x16GiB   DDR4 6x16GiB

The RAJA programming model enables porting to GPU machines while maintaining portability

- RAJA helps **insulate applications from technology disruption** and doesn't inhibit using new or platform-specific tools

- RAJA **insulates apps from variability** of programming model and architecture considerations
  - "Easy to do most things, possible to do really hard things"
    (apologies to Larry Wall)

- It's easy for one application to adopt features and/or optimizations developed in RAJA in the context of other applications (just grab a newer version)

- RAJA is reasonably **easy to grasp** for all application developers (esp. non CS experts)

- RAJA is **easy to integrate** with apps and **easy to adopt incrementally**

- RAJA **facilitates application flexibility** by promoting clean encapsulation

A SyCL Backend is under development for our codes that run on Aurora

SW4 runs on all the DOE systems

High resolution simulations of M_W 7.0 Hayward Fault earthquake

Cori-II (SW4): PGV = 2.1797349453 m/s

Sierra(SW4-RAJA): PGV = 2.1797349453 m/s
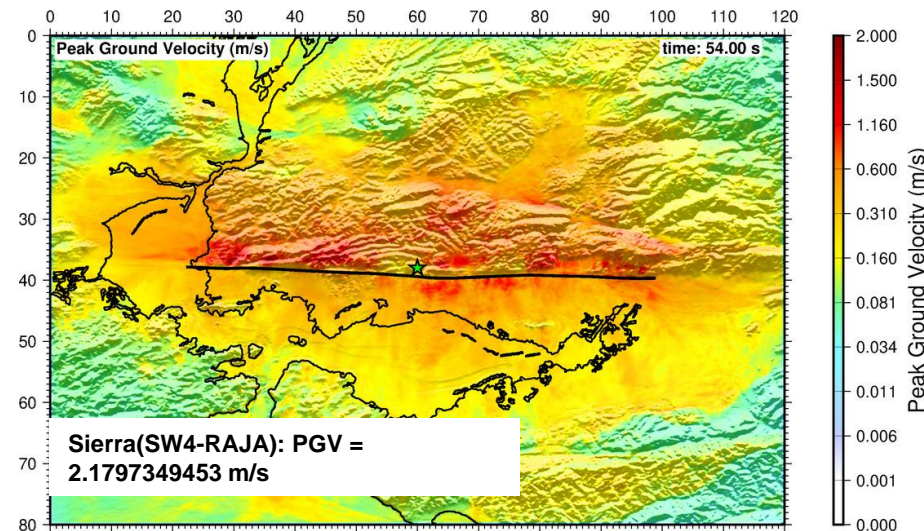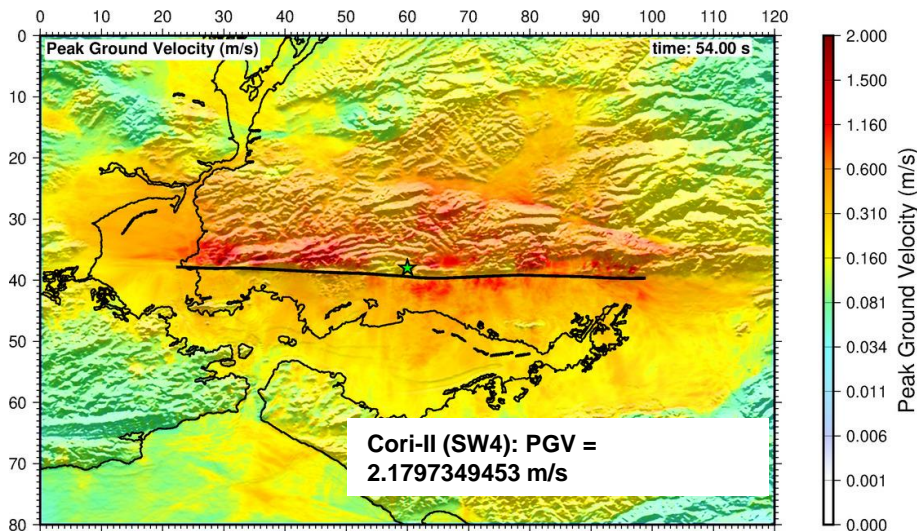
Two codes, different platforms: Peak Ground Velocity maps agree to machine precision!

LLNL has a significant investment in our codes running well on Intel hardware

- While we have no announced Intel GPU systems we want our codes to be able to run on them for procurement reasons

- We are making sure our codes continue to run well on our current Intel machines as we move onto GPU systems

- Performance Portability is key to us in the evolving ecosystem.

Lawrence Livermore
National Laboratory