

TRINITY

SAND2017-9717PE

Trinity Architecture

Scott Hemmert

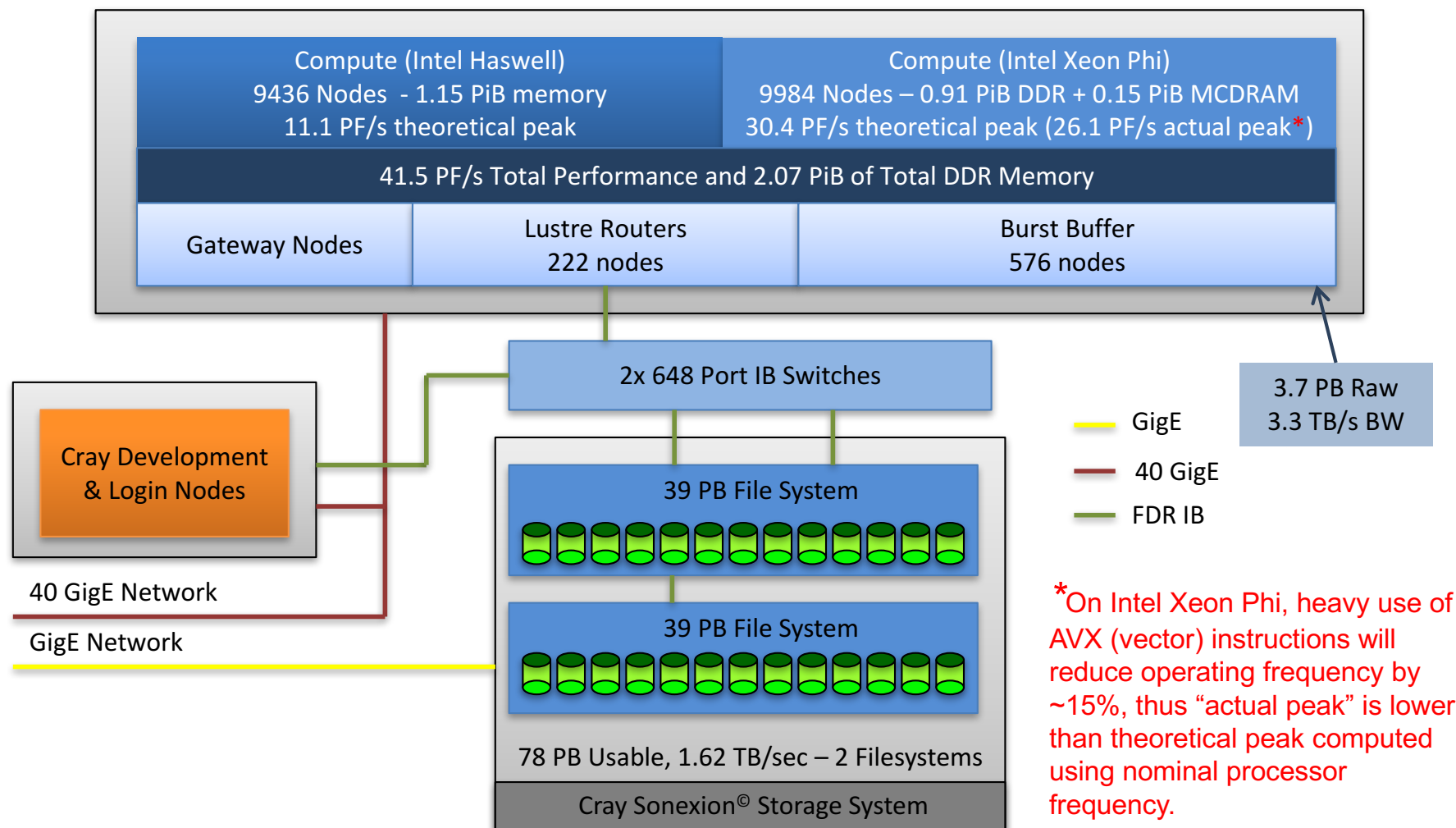


Alliance for Computing at Extreme Scale

Trinity Platform

- Trinity is a single system that contains both Intel Haswell and Xeon Phi (Knights Landing) processors
 - Haswell partition satisfies FY16 mission needs (well suited to existing codes).
 - KNL partition delivered in FY16 results in a system significantly more capable than current platforms and provides the application developers with an attractive next-generation target (and significant challenges)
 - Aries interconnect with the Dragonfly network topology
- Based on mature Cray XC40 architecture with Trinity introducing new architectural features
 - Intel Xeon Phi (Knights Landing) processors
 - Burst Buffer storage nodes
 - Advanced power management system software enhancements

High Level Trinity Configuration



Trinity Architecture Overview

Metric	Trinity		
Node Architecture	Complete Trinity System	Intel Haswell Processor Partition	Intel Xeon Phi Processor Partition
Memory Capacity	2.07 PiB	1.15 PiB	0.91 PiB
Memory BW	>6PB/sec	>1 PB/s	>1PB/s + >4PB/s
Theoretical Peak FLOPS	41.5 PF/s	11.1 PF/s	30.4 PF/s (26.1 PF/s)
Number of Nodes	19,420	9,436	9,984
Number of Cores	980,864	301,952	678,912
Number of Cabs (incl. I/O & BB)	110		
PFS Capacity (usable)	78 PB		
PFS Bandwidth (sustained)	1.45 TB/s		
BB Capacity (usable)	3.7 PB		
BB Bandwidth (sustained)	3.3 TB/s		

Integration into Final Trinity Configuration

[illegible]

Test Bed Systems

- Gadget – Software Development Testbed
- Application Regression Testbeds
 - Configuration
 - 100 Haswell Compute Nodes + 100 KNL Compute Nodes
 - 720 TB / 15 GB/s Sonexion 2000 Filesystem
 - 6 Burst Buffer Nodes
 - Trinitite
 - LANL Yellow Network
 - Mutrino
 - Sandia SRN Network

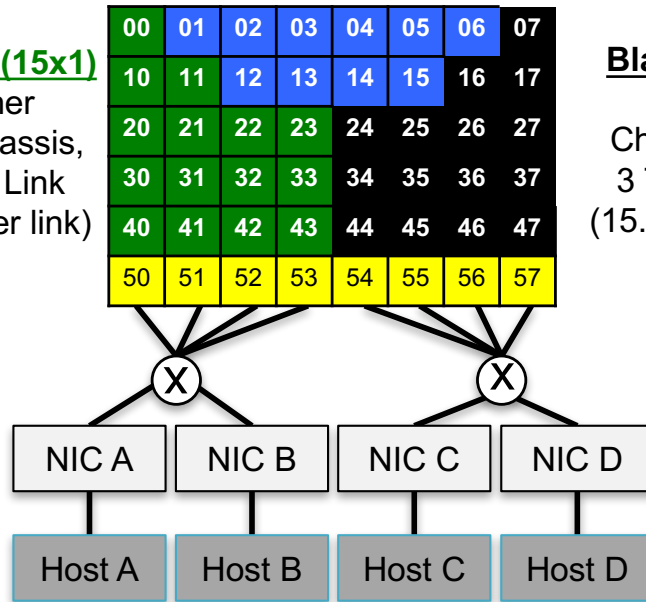
Cray Aries Interconnect

Blue Links (10x1)

To Other Groups,
10 Global Links
(4.7 GB/s per link)

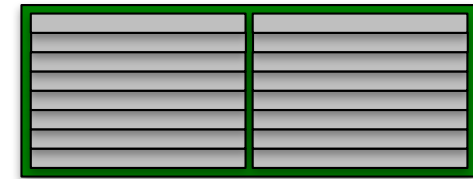
Green Links (15x1)

To 15 Other
Blades in Chassis,
1 Tile Each Link
(5.25 GB/s per link)



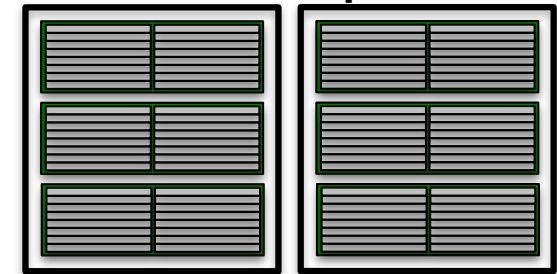
Black Links (5x3)

To 5 Other
Chassis in Group,
3 Tiles Each Link
(15.75 GB/s per link)



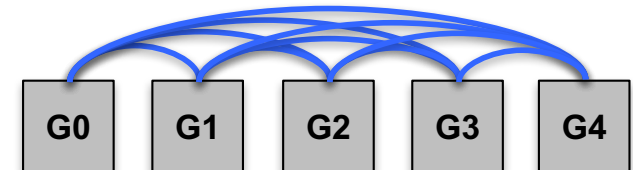
16 Blades Per Chassis
16 Aries, 64 Nodes
All-to-all Electrical Backplane

2. Group



6 Chassis Per Group
96 Aries, 384 Nodes
Electrical Cables, 2-D All-to-All

3. Global

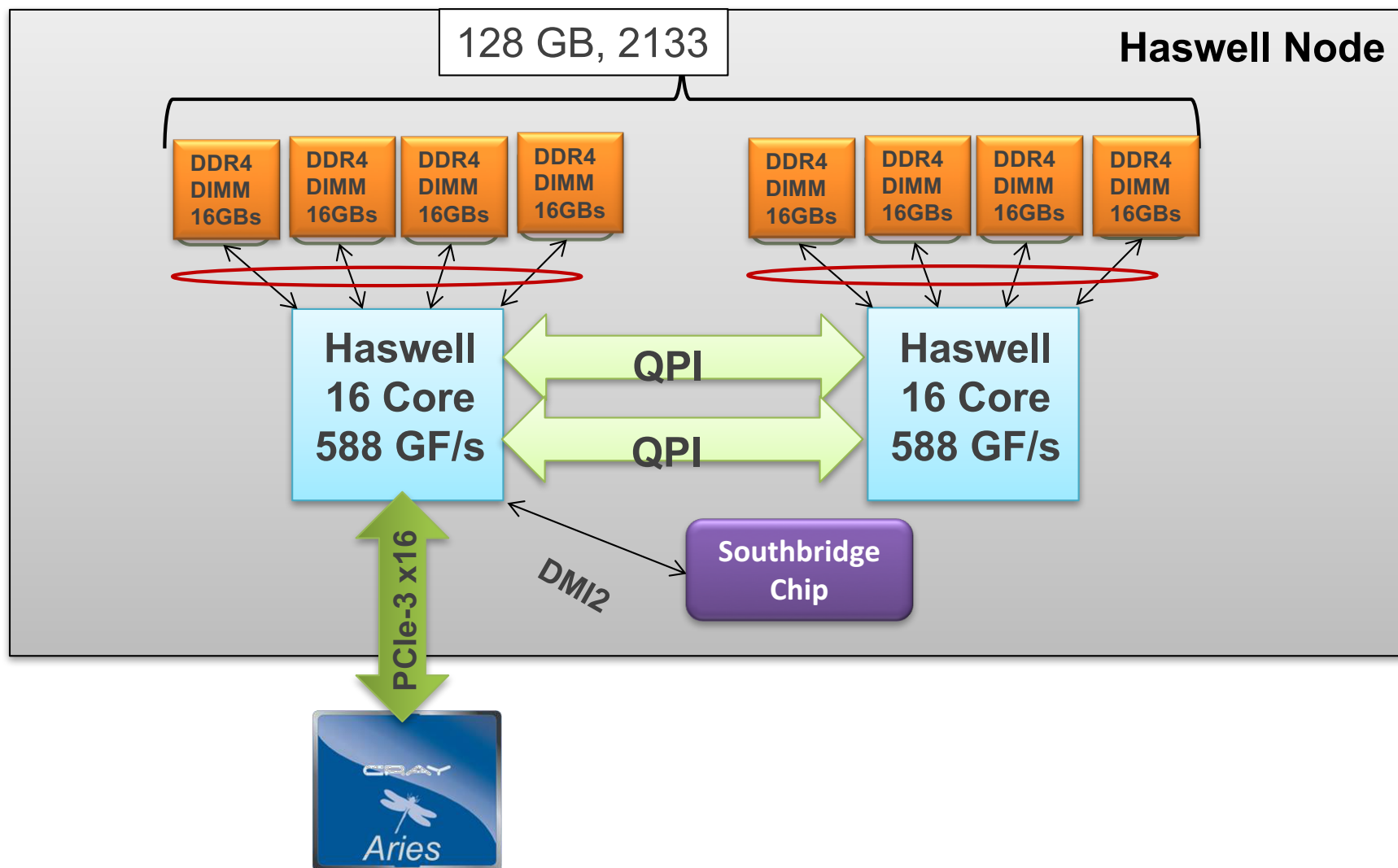


Up to 241 Groups
Up to 23136 Aries, 92544 Nodes
Optical Cables, All-to-All between Groups

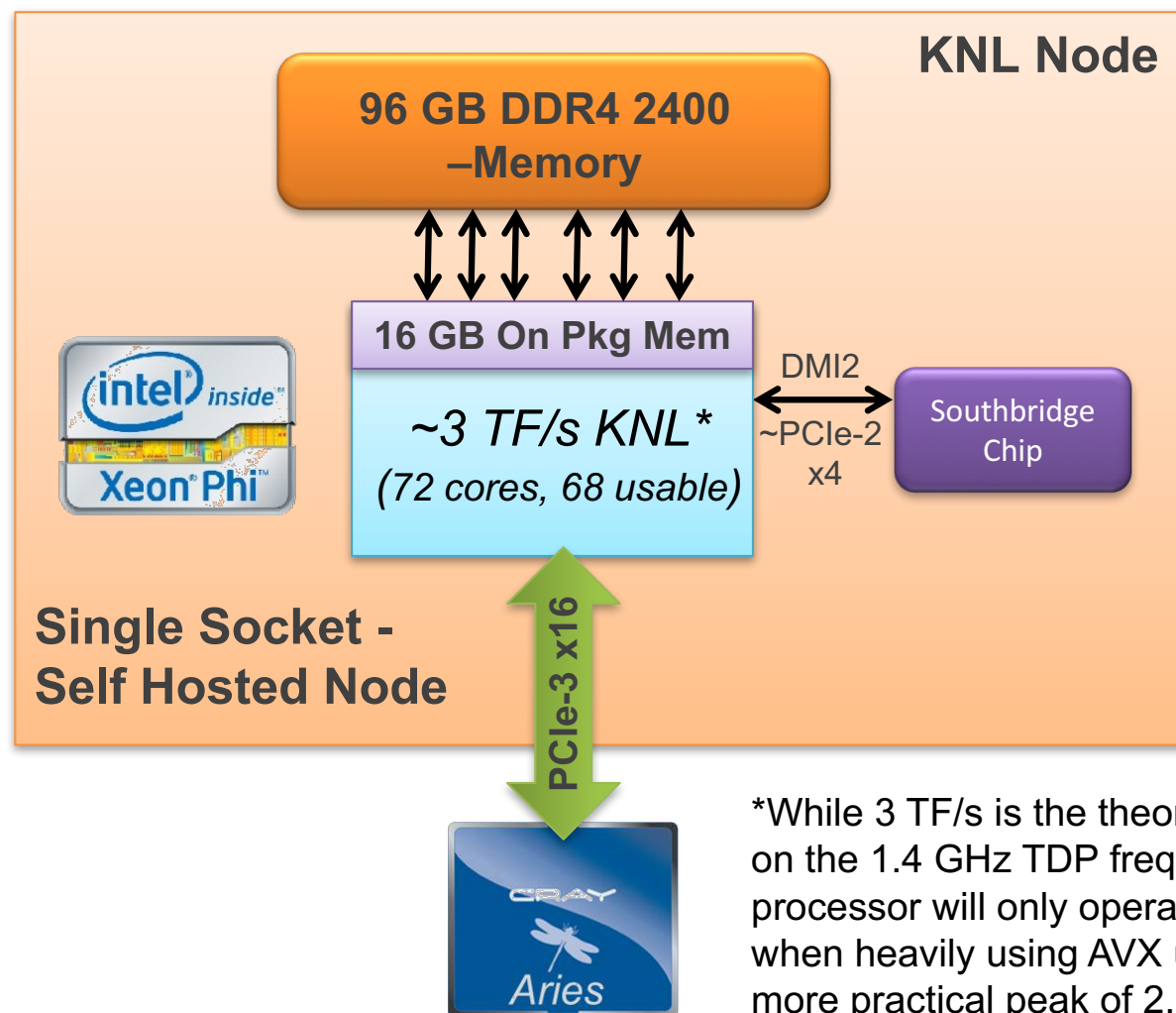
Gemini: 2 nodes, 62.9 GB/s routing bw
Aries 4 nodes, 204.5 GB/s routing bw

Aries has advanced adaptive routing

Trinity Haswell Compute Node



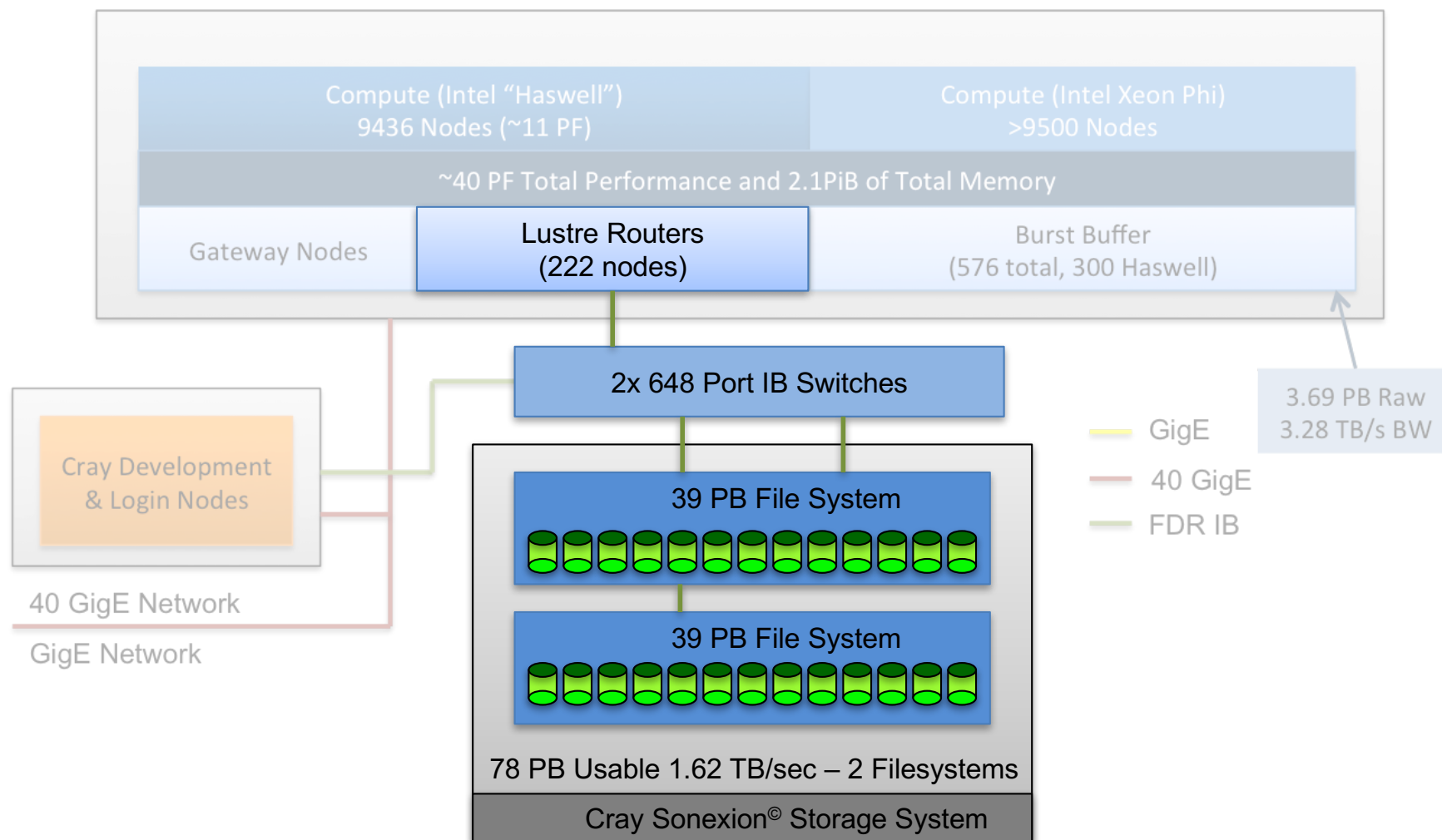
Trinity KNL Compute Node



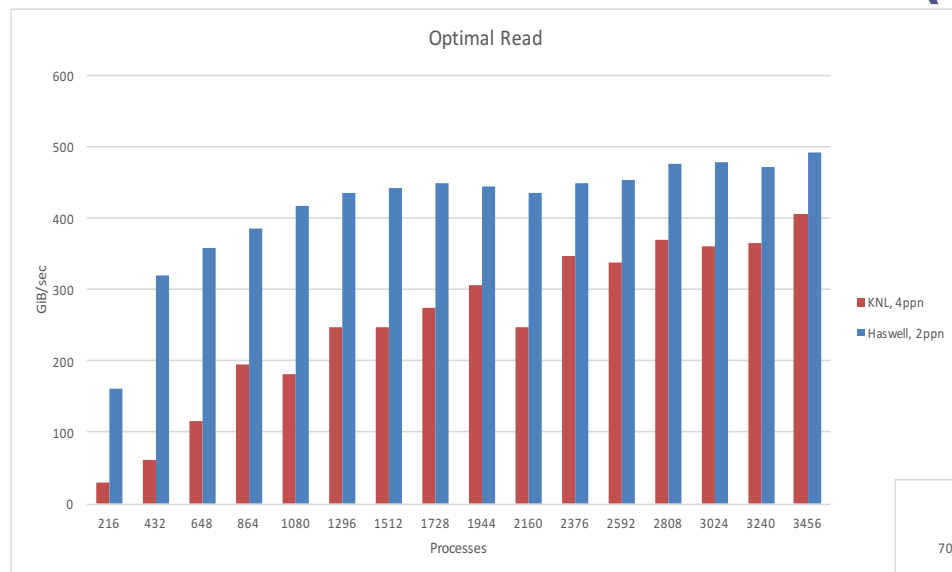
Compute Node specifications

	Haswell	Knights Landing
Memory Capacity (DDR)	2x64=128 GiB	96 GiB
Memory Bandwidth (DDR)	136.5 GB/s	115.2 GB/s
# of sockets per node	2	1
# of cores	2x16=32	64 to 72
Core frequency (GHz)	2.3	TBD
# of memory channels	2x4=8	6
Memory Technology	2133 MHz DDR4	2400 MHz DDR4
Threads per core	2	4
Vector units & width (per core)	2x256 AVX2	2x512 AVX-512F
On-chip MCDRAM		16 GB @ > 500 GB/s

File System Configuration

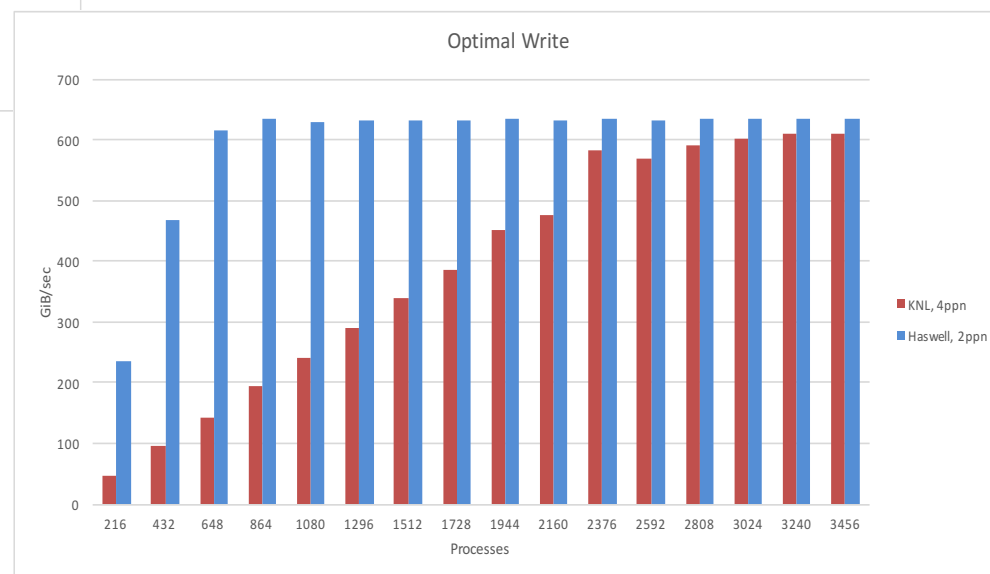


Optimal Read/Write KNL, Haswell Comparison Based on Streams/OST (216 OSTs per FS)

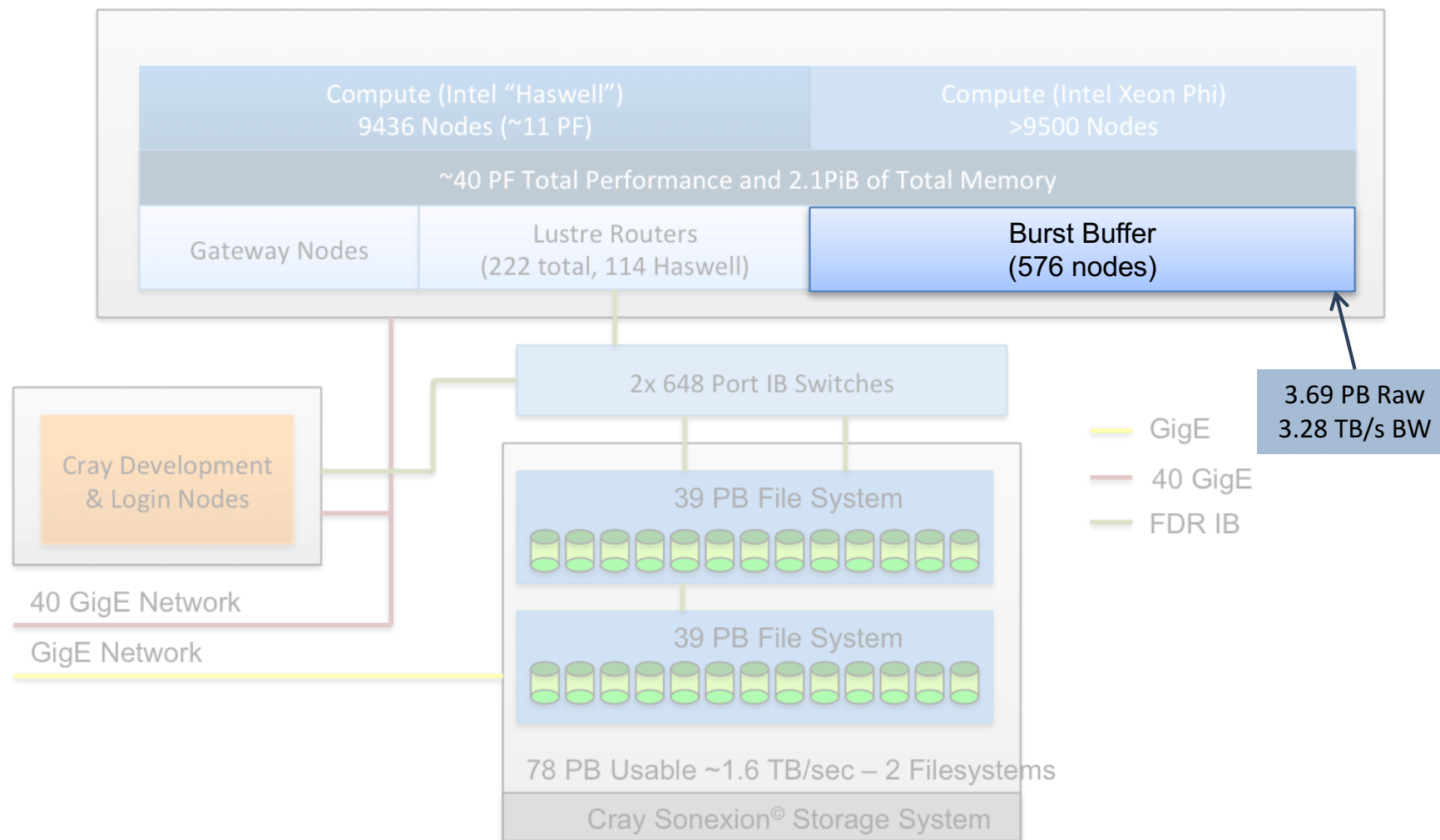


Optimal read: ~500 GiB/s
Sustained read: ~300 GiB/s

Optimal write: ~600 GiB/s
Sustained write: ~500 GiB/s

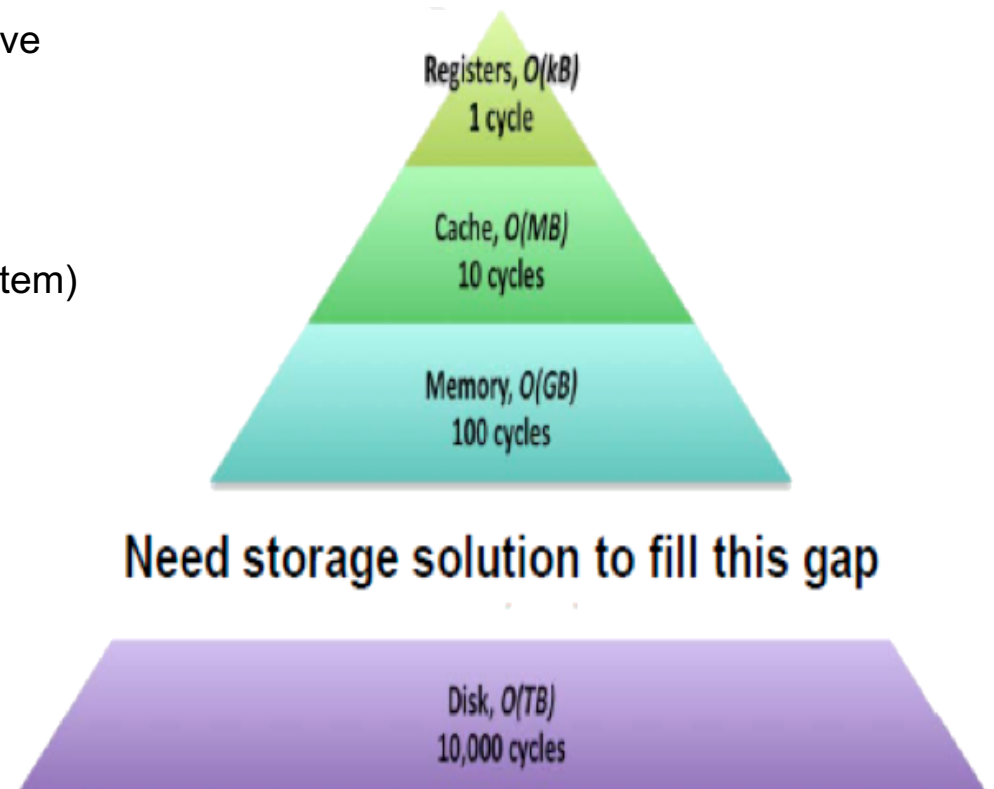


Burst Buffer Configuration



Burst Buffers will Improve Productivity and Enable Memory Hierarchy Research

- Technology Drivers:
 - Solid State Disk (SSD) cost decreasing
 - Lower cost of bandwidth than hard disk drive
- Trinity Operational Plans:
 - SSD based 3 PB Burst Buffer
 - 3.28 TB/Sec (2x speed of Parallel File System)
- Burst Buffer will improve operational efficiency by reducing defensive IO time
- Burst Buffer fills a gap in the Memory and Storage Hierarchy and enables research into related programming models





DataWarp Details

- DataWarp nodes built from Cray service nodes
 - 16-core Intel Sandy Bridge with 64 GiB memory
 - Two Intel P3608 SSD cards (4 TB per card)
 - Capacity overprovisioned to get to 10 drive writes per day endurance (standard is 3 DWPD)
- Usage modes
 - Striped scratch
 - Striped private
 - Paging (possible future mode)
 - Cache (possible future mode)
- Integrated with workload manager
 - Stage in/Stage out (single job lifetime)
 - Persistent allocations (accessible by multiple jobs)

Trinity Installation



