Designing an HPC Cluster with Expert Help

2020 Webinar Series

advanced clustering technologies, inc.

About Us

- Advanced Clustering Technologies has been providing HPC solutions since 2001
- We have worked with hundreds of Universities and with government agencies, national labs and corporations
- We offer on-demand and on-premises solutions
- We can help with every phase of HPC acquisition from grant writing and RFP publishing through configuring and building and on to installation and deployment
- We typically build systems in the $50,000 ~ $1 million range

advanced clustering
technologies, inc.

Topics We Will Cover

- **Designing an HPC system**
  - Processors
  - Interconnect
  - Storage
  - GPUs

Topics We Will Cover

- **Preparing for the Future**
  - Datacenter readiness
  - Planning ahead

Topics We Will Cover

- **Wrapping Up**
  - Q & A
  - Resources
  - This webinar will be available on demand on our site at advancedclustering.com

Processors

Max 28 cores

VS

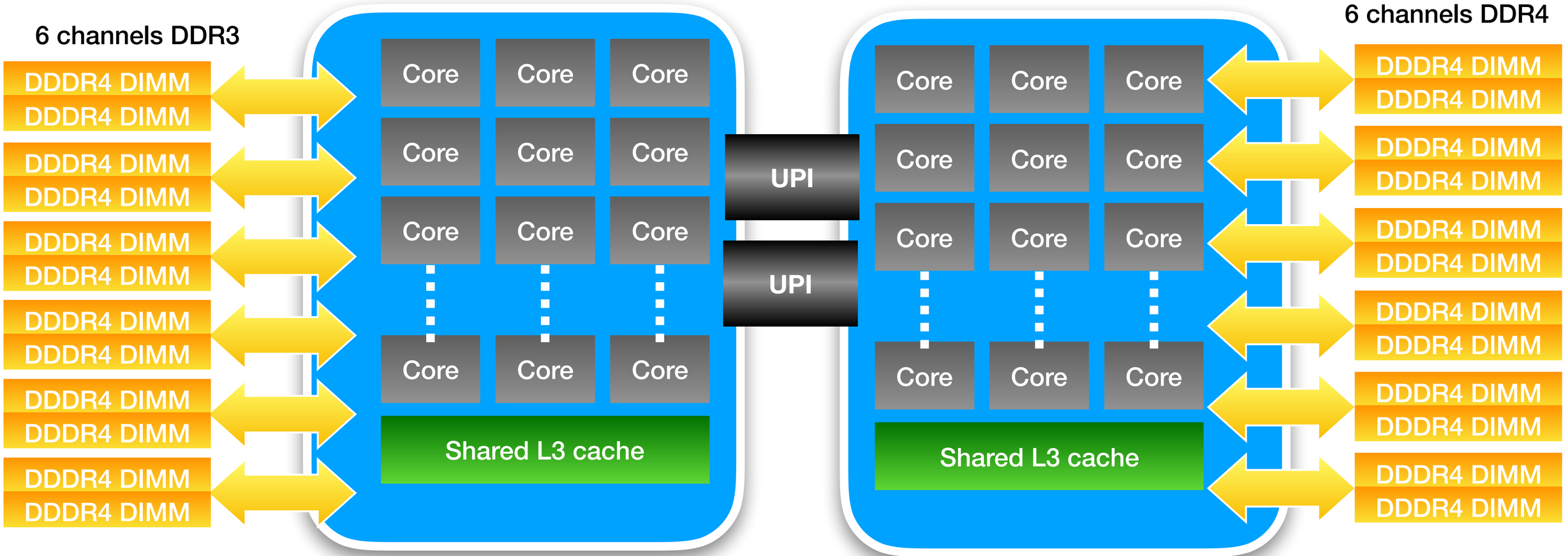
Max 64 cores

advanced clustering
technologies, inc.

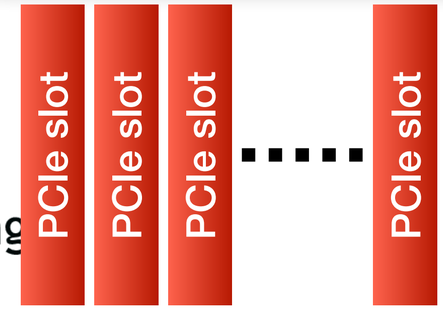
Intel Xeon CPUs

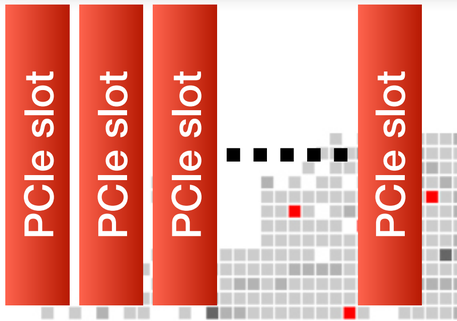
- Code named “Cascade Lake Refresh”
- 14nm process
- 4-28 core options
  - 50+ variants to choose from
- 6x DDR4 memory channels per socket, 2x DIMMs per channel
- 48 lanes of PCI-e Gen 3.0 bandwidth
- Up to 32x floating point ops per cycle
- Up to 38MB of L3 cache

Cascade Lake SP



48 lanes PCIe
Gen 3.0





48 lanes PCIe
Gen 3.0

advanced clustering
technologies, inc.

Xeon SKU Levels

Bronze 32XX

- Max 2 Sockets
- 6 channel DDR4 @ 2133
- 2x UPI links @ 9.6GT/s
- 16 DP FLOPs per cycle
- AVX 512 (1x 512b FMA)

Silver 42XX

- Max 2 Sockets
- 6 channel DDR4 @ 2400
- 2x UPI links @ 9.6GT/s
- 16 DP FLOPs per cycle
- AVX 512 (1x 512b FMA)
- Hyper Threading
- Turbo Boost

Gold 52XX

- Max 4 Sockets
- 6 channel DDR4 @ 2666
- 2x UPI links @ 10.4GT/s
- 16 DP FLOPs per cycle
- AVX 512 (1x 512b FMA)
- Hyper Threading
- Turbo Boost

Gold 62XX

- Max 4 Sockets
- 6 channel DDR4 @ 2933
- 3x UPI links @ 10.4GT/s
- 32 DP FLOPs per cycle
- AVX 512 (2x 512b FMA)
- Hyper Threading
- Turbo Boost

Platinum 82XX

- Max 8 Sockets
- 6 channel DDR4 @ 2933
- 3x UPI links @ 10.4GT/s
- 32 DP FLOPs per cycle
- AVX 512 (2x 512b FMA)
- Hyper Threading
- Turbo Boost

advanced clustering technologies, inc.

What is AVX?

AVX 1.0

- Launched with Sandy Bridge processors (E5-2600)
- 1x 256-bit basic FP
- 8x Dual precision floating point ops per cycle

AVX 2.0

- Launched with Haswell processors (E5-2600v3) and included in broadwell (E5-2600v4)
- 2x 256-bit FMA FP
- 16x Dual precision floating points ops per cycle

AVX 512

- Launched with Skylake processors
- Up to 2x 512-bit FMA FP
- Up to 32x Dual precision floating points ops per cycle

- 1x FMA on 32XX - 52XX (16x DP Flops/cycle)
- 2x FMA on 62XX - 82XX (32x DP Flops/cycle)

advanced clustering technologies, inc.

AVX frequency

1. Certain high-power instructions get run on the CPU (mostly Integer and FP AVX instructions)
2. These instructions cause the power control unit (PCU) to increase the voltage
3. With the increased voltage the CPU could now be drawing more power than TDP, so frequency is reduced to stay within TDP limits
4. 1ms after the last high power instruction has executed, the voltage will be reduced back to non-AVX levels

Highly optimized AVX workloads (like Linpack) will be more likely to operate at reduced frequency

AVX frequency

Model	6240
Cores	18
Threads	36
GHz	2.60 GHz
TDP	150 W

- When all cores are using non AVX — **2.6GHz** per core
- When all cores are using AVX 2.0 instructions — **2.0GHz** per core
- When all cores are using AVX 512 instructions — **1.6GHz** per core

Calculating TFLOPs

TFLOPs = Clock Frequency [TDP|AVX512] * Cores * FLOPs per cycle

- Example ***TDP*** Frequency: Cascade Lake Gold 6240
  - 2.6 (GHz) * 18 (Cores) * 16 (FLOPs) = 748.8 GFLOPs per socket
  - 1,497.6 GFLOPs per dual socket system

- Example ***AVX512*** Frequency: Cascade Lake Gold 6240
- 1.6 (GHz) * 16 (Cores) * 32 (FLOPs) = 819.2 GFLOPs per socket
  - 1,638.4 GFLOPs per dual socket system

What speed is my CPU?

It depends on the application. CPU frequency varies based on:

- Temperature of CPU
- Number of active compute cores
- Power consumption
- Use of high-power instructions

Heavy AVX usage = lower frequency to stay within TDP

AMD EPYC CPUs

- Code named “Rome"
- 7nm & 14nm process
- 8, 12, 16, 24, 32 and 64 cores options
- 8x DDR4 memory channels per socket, 2x DIMMs per channel
- 128 lanes of PCI-e Gen 4.0 bandwidth
- 16x floating point ops per cycle
- Support for PCI Gen 4
- Modular design
  - Multiple dies per CPU
  - CCD - Compute Core Die
  - IOD - I/O Die

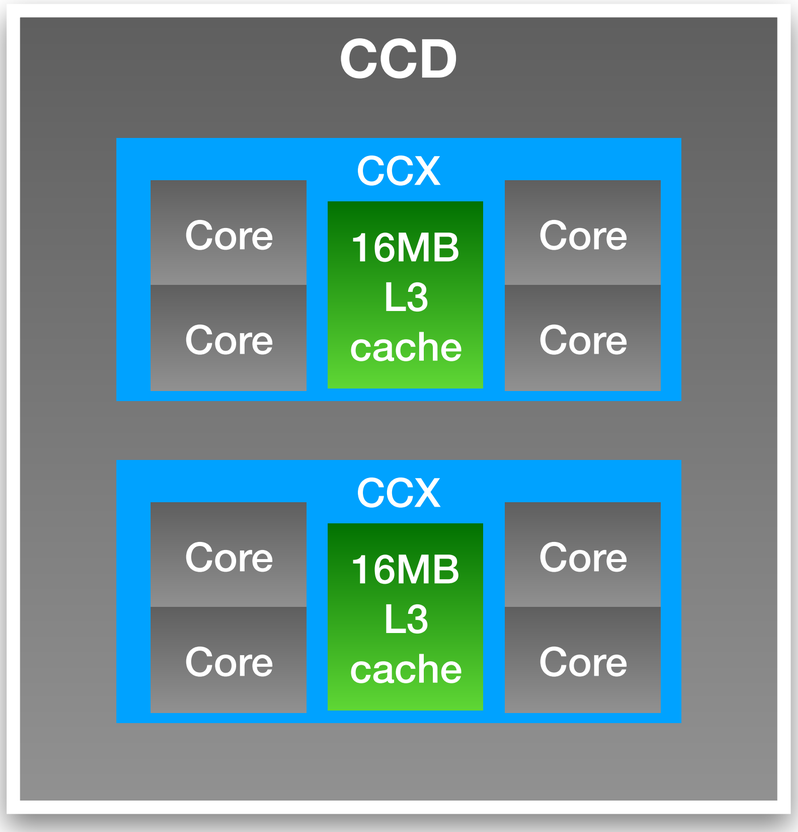
advanced clustering technologies, inc.

AMD EPYC “Rome”

COMPUTE CORE DIE (CCD)

- Each CCD consists of up to 2 cpu compute complexes (CCX) at 7nm
- Each CCX contains up to 4 compute cores and 16MB of shared L3 cache
- Multiple CCDs are added to the CPU to form various core count configurations
  - 64 core CPU has 8x CCD units

AMD EPYC “Rome”



COMPUTE CORE DIE (CCD)

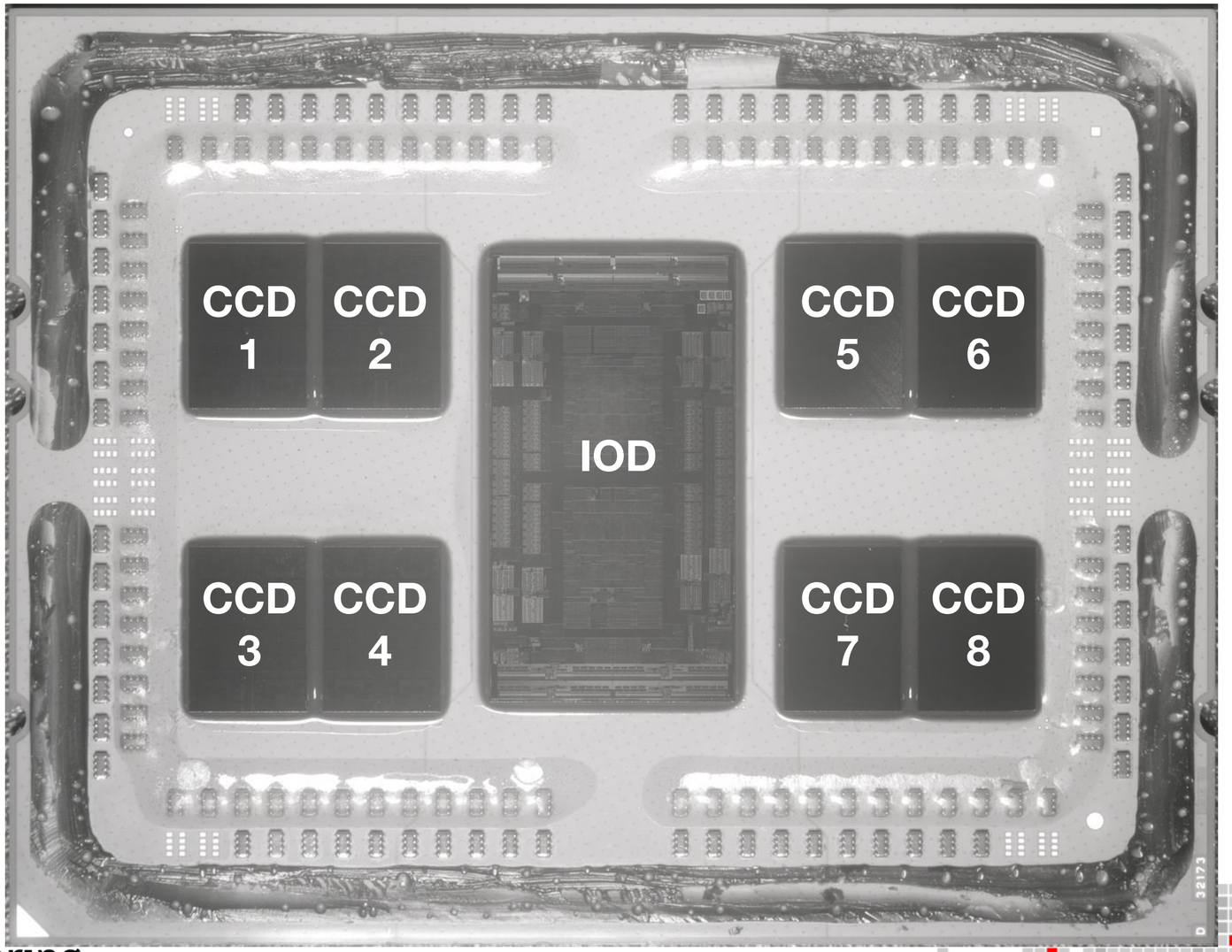
- Each CCD consists of up to 2 cpu compute complexes (CCX) at 7nm
- Each CCX contains up to 4 compute cores and 16MB of shared L3 cache
- Multiple CCDs are added to the CPU to form various core count configurations
  - 64 core CPU has 8x CCD units
- The CCD unit is based on the Zen2 architecture exactly same component as the desktop Ryzen processors

AMD EPYC “Rome”

I/O DIE (IOD)

- Provides memory and I/O functionality (PCIe, SATA, USB, etc)
- All functions integrated no extra chipset needed
- 14nm process technology
- 8 memory channels at 3200MHz, 2 DIMMs per channel
- 128 lanes of PCIe Gen4 (64 used when in a dual socket configuration)

AMD EPYC “Rome”

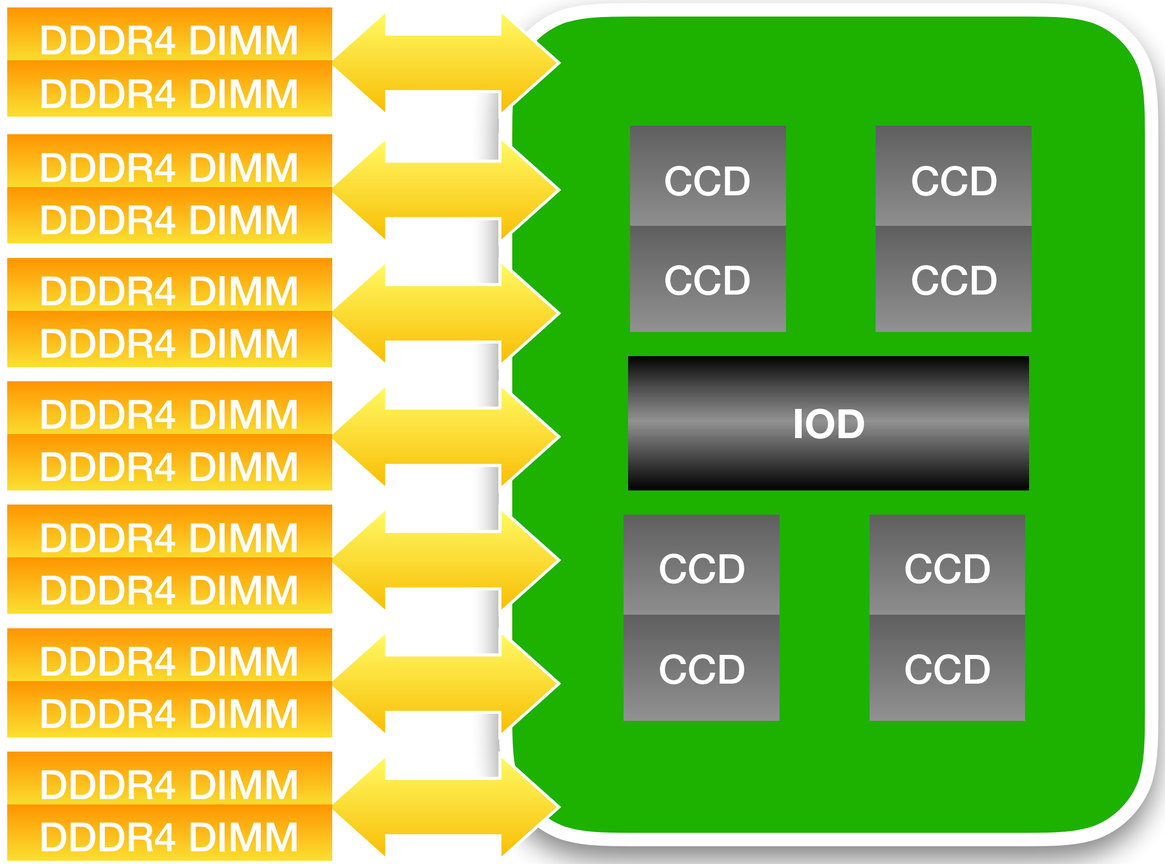


advanced clustering technologies, inc.

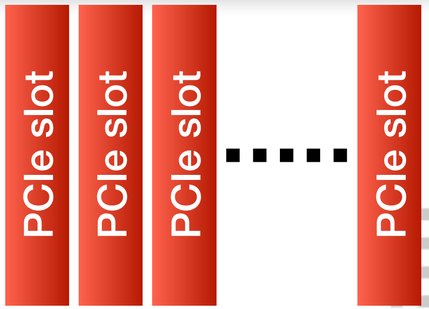
Single vs Dual socket

- Single socket systems are less expensive than dual socket systems without compromising performance
- Single socket offers better performance in box as there is no latency issue due to CPU intra-communications
- Single socket has reduced power and cooling requirements
- Single socket EPYC SKUs that end in “P” are only supported on one socket systems - available at lower costs than dual socket equivalents

EPYC 1 socket

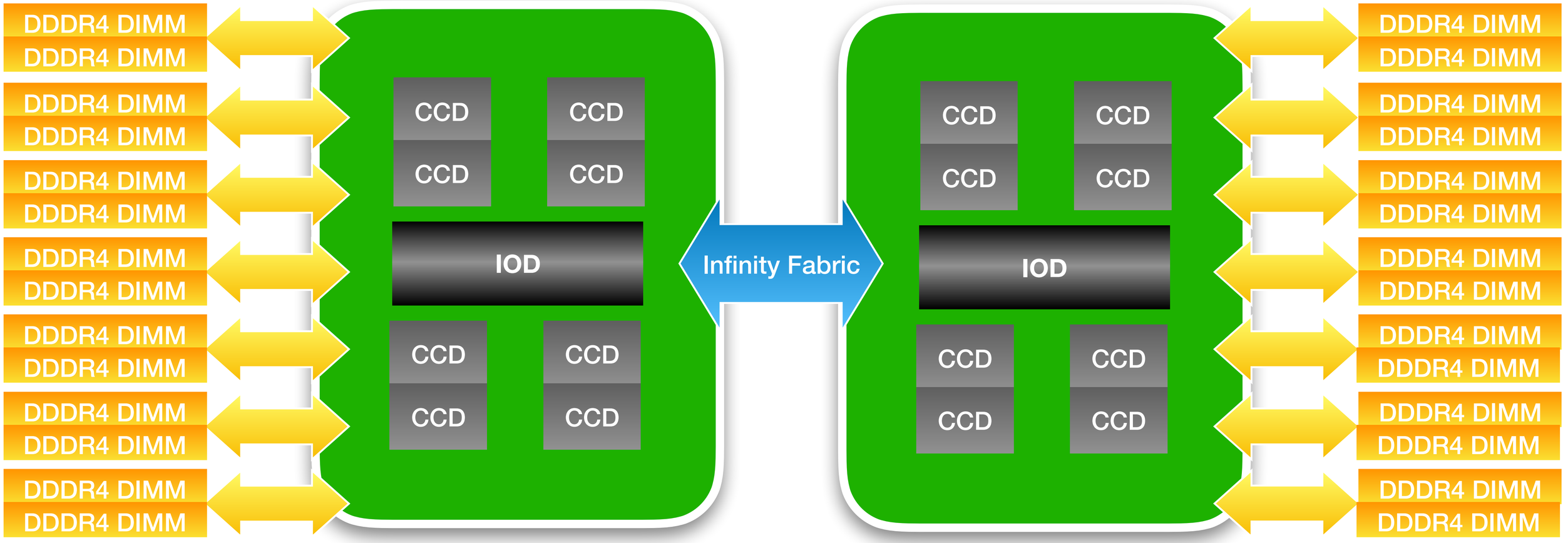


128 lanes PCIe
Gen 4.0



advanced clustering
technologies, inc.

EPYC 2 socket



64 lanes PCIe Gen 4.0

PCIe slot
PCIe slot
PCIe slot
PCIe slot

64 lanes PCIe Gen 4.0

PCIe slot
PCIe slot
PCIe slot
PCIe slot

advanced clustering technologies, inc.

EPYC SKUs

Uni-processor	Multi-processor		
7702P 64c 2.0GHz - 240W	**7H12** 64c 2.6GHz - 280W	**7502** 32c 2.5GHz - 180W	**7302P** 16c 2.8GHz - 155W
7502P 32c 2.5GHz - 200W	**7742** 64c 2.25GHz - 225W	**7452** 32c 2.35GHz - 155W	**7282** 16c 2.8GHz - 120W
7402P 24c 2.8GHz - 200W	**7662** 64c 2.0GHz - 225W	**7F72** 24C 3.2GHZ - 240W	**7272** 12c 2.9GHz - 120W
7302P 16c 2.8GHz - 180W	**7702P** 64c 2.0GHz - 200W	**7402** 24c 2.8GHz - 180W	**7F32** 8c 3.7GHz - 180W
7232P 8c 2.4GHz - 85W	**7642** 48c 2.3GHz - 225W	**7352** 24c 2.3GHz - 155W	**7262** 8c 3.2GHz - 155W
	7552 48c 2.2GHz - 200W	**7F52** 16c 3.5GHz - 240W	**7252** 8c 3.1GHz - 120W
	7542 32c 2.9GHz - 225W		

Calculating TFLOPs

TFLOPs = Clock Frequency * Cores * FLOPs per cycle

- Example: EPYC Rome 7502
  - 2.5 (GHz) * 32 (Cores) * 16 (FLOPs) = 1,280 GFLOPs per socket
  - 2,560 GFLOPs per dual socket system

EPYC has no AVX512 support

Unlike Intel - no frequency reductions when using AVX 2.0 instructions

Why choose Intel?

- Better per core performance
- AVX 512 instructions for applications that can use them
- More established eco-system
  - More system options and choices: Storage, GPU, etc
  - Intel compilers, MKL, most commercial code optimized for Intel
- Slightly less power consumption per node

Why choose AMD?

- More cores per system
  - If code scales well, AMD is a good choice
- Viable, cost effective single socket solution
- More memory channels, and faster memory support
- More and faster I/O expansion 128 lanes of PCIe Gen 4.0
- Fewer options, less confusion with performance numbers

Side-by-Side Comparison

	AMD	Intel
Process	7nm & 14 nm	14nm
Core Options	8, 12, 16, 24,32, 64	4-28
Memory	8x DDR4 channels/socket; 2x DIMMS/channel	6x DDR4 channels/socket; 2x DIMMS/channel
Bandwidth	128 lanes of PCI-e Gen 4.0	48 lanes of PCI-e Gen 3.0
Floating point ops per cycle	16x	32x

Sign up for a virtual 1-on-1 meeting

https://www.advancedclustering.com/meetup/

Interconnect

- Options for node-to-node Interconnect
  - Ethernet
  - InfiniBand

Which is right for you?

Simplified answer:

If no MPI codes, then stick with Ethernet

Ethernet

- Widely used everyone knows
- 1Gb / 10Gb pretty cost effective
- 25/50/100Gb available at somewhat reasonable costs
- Widest compatibility - easiest to use
- Every cluster will have Ethernet
- Downsides lower bandwidth - high latency
- Scalability: as your cluster grows beyond a single switch, the costs rise significantly as does the complexity of the network topology

InfiniBand

- InfiniBand is an RDMA interface (Remote Direct Memory Access) device
  - Does not operate like an Ethernet device
  - One host can directly read or write to another host's memory without involving the operating system
  - This OS bypass provides high bandwidth and low latency
- Supports multiple upper layer protocols (ULP): MPI, Storage (GPFS, Lustre, BeeGFS, NFSoRDMA)
- InfiniBand hardware from single vendor - Mellanox
  http://www.mellanox.com

InfiniBand

- Comes in various flavors - SDR, DDR, QDR, FDR, EDR and HDR
- All of the switching logic is run via software called a “subnet manager” (SM)
  - Subnet manager sweeps the network and discovers all the devices, and allows traffic to flow on the network
  - Subnet managers can be run in switch, or more cost effectively on a compute or head node of cluster

SDR*	10Gb
DDR*	20Gb
QDR*	40Gb
FDR*	56Gb
EDR	100Gb
HDR	100/200Gb

*** End of Life**

InfiniBand and PCIe

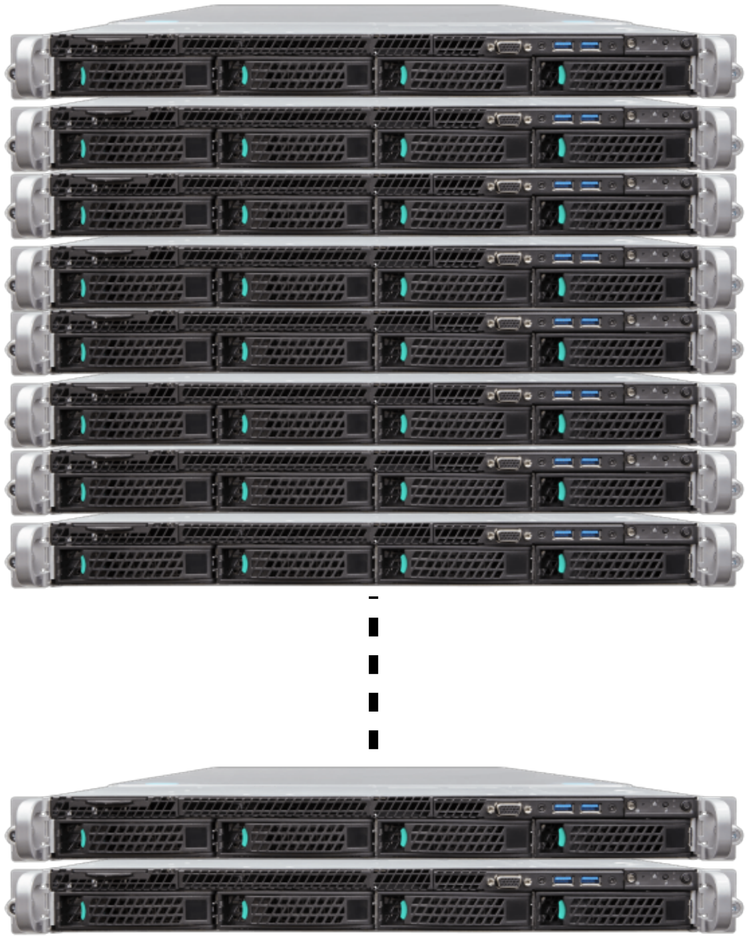
- Bus speed and PCI-e lanes are limiting factor of IB performance
- Servers usually only have 16x PCI-e slots
  - Intel systems only support PCI-e Gen 3
  - AMD can support PCI-e Gen 4

	GT/s per lane	BW on 16x slot
PCI-e Gen 1	2.5 GT/s	40Gb/s
PCI-e Gen 2	5 GT/s	80Gb/s
PCI-e Gen 3	8 GT/s	128Gb/s
PCI-e Gen 4	16 GT/s	256Gb/s

InfiniBand

- InfiniBand networks are highly scalable
- Smaller scale switches (36 or 40 port) can be connected with multiple links to grow your network
- Cabling is a challenge - large inflexible cables with QSFP connectors
- Copper cables limited to 2-3 meters on newest standards
- Fiber cables are available but much more expensive ($70 for copper cable, $500+ for fiber)

Under 36 node example

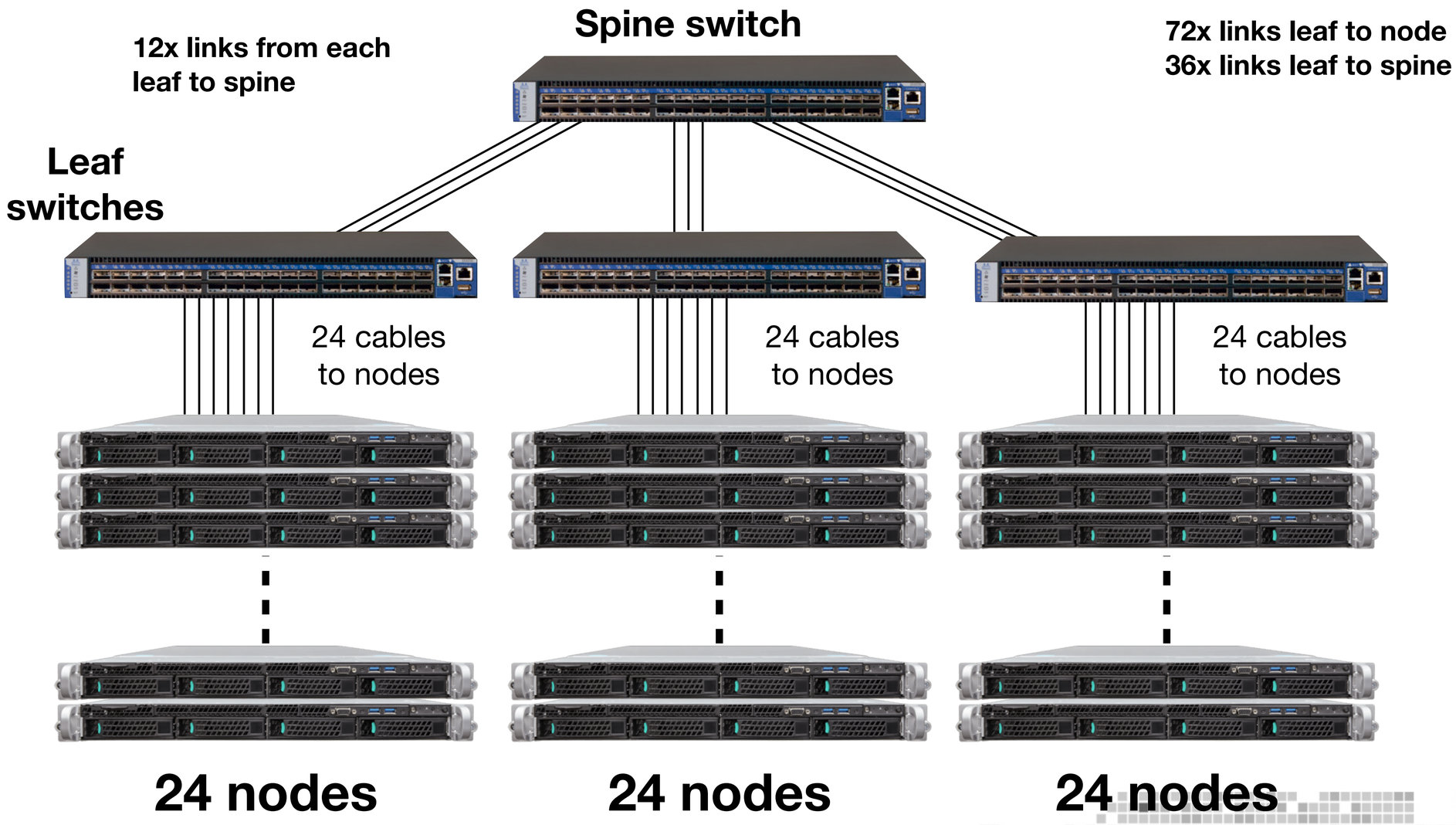


Up to 36 nodes

36 port InfiniBand Switch

advanced clustering technologies, inc.

72 nodes 2:1



advanced clustering technologies, inc.

Omni-Path

- In 2016 Intel decided to make it's own cluster interconnect called OmniPath
  - Similar in concept to InfiniBand uses 48 port switches
  - Runs at 100Gb
- Intel has stopped developing future versions of OmniPath only 100Gb equipment
- Omni-path may rise again; Intel recently sold it to Cornelis Networks

Why oversubscribe?

- Fewer switches, fewer cables = lower cost
- Most parallel applications don’t need 100Gb/s bandwidth per node - latency is more important, and very little impact caused by oversubscription
- Still can run 1:1 if you use all nodes on the same switch
    - Slurm has an interconnect plugin to facilitate this - users can request how many switches their job can span

Sign up for a virtual 1-on-1 meeting

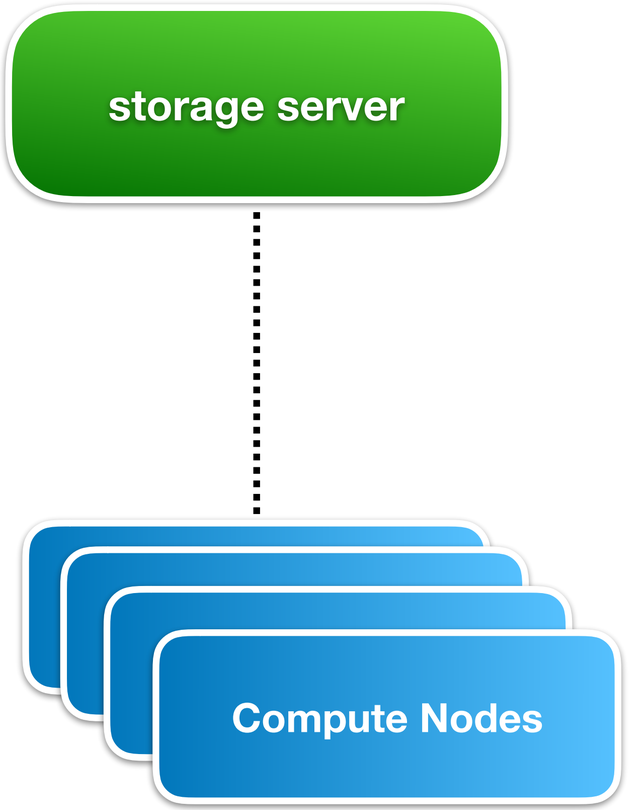
https://www.advancedclustering.com/meetup/

Storage

- Storage could be it’s own multi-hour long presentation
- 2 popular ways to accomplish cluster storage
  - NFS
  - Parallel filesystem

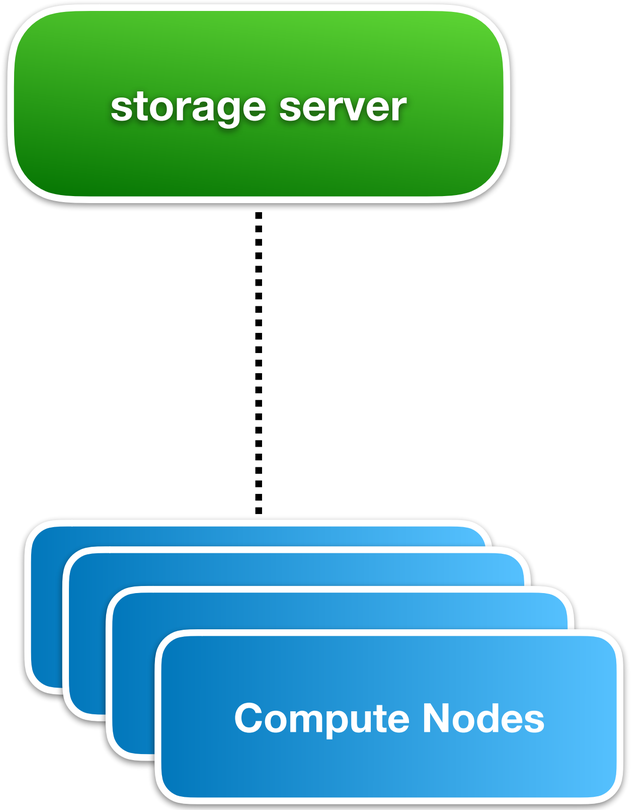
NFS server

- Typical setup in smaller scale clusters
- Easy to use, simple, understood by many
- Combination of:
  - Files are stored on a local storage device (usually RAID array). Using an underlying filesystem that are known and well tested: ext4, xfs, ZFS, etc
  - Network protocol to expose local files to other hosts



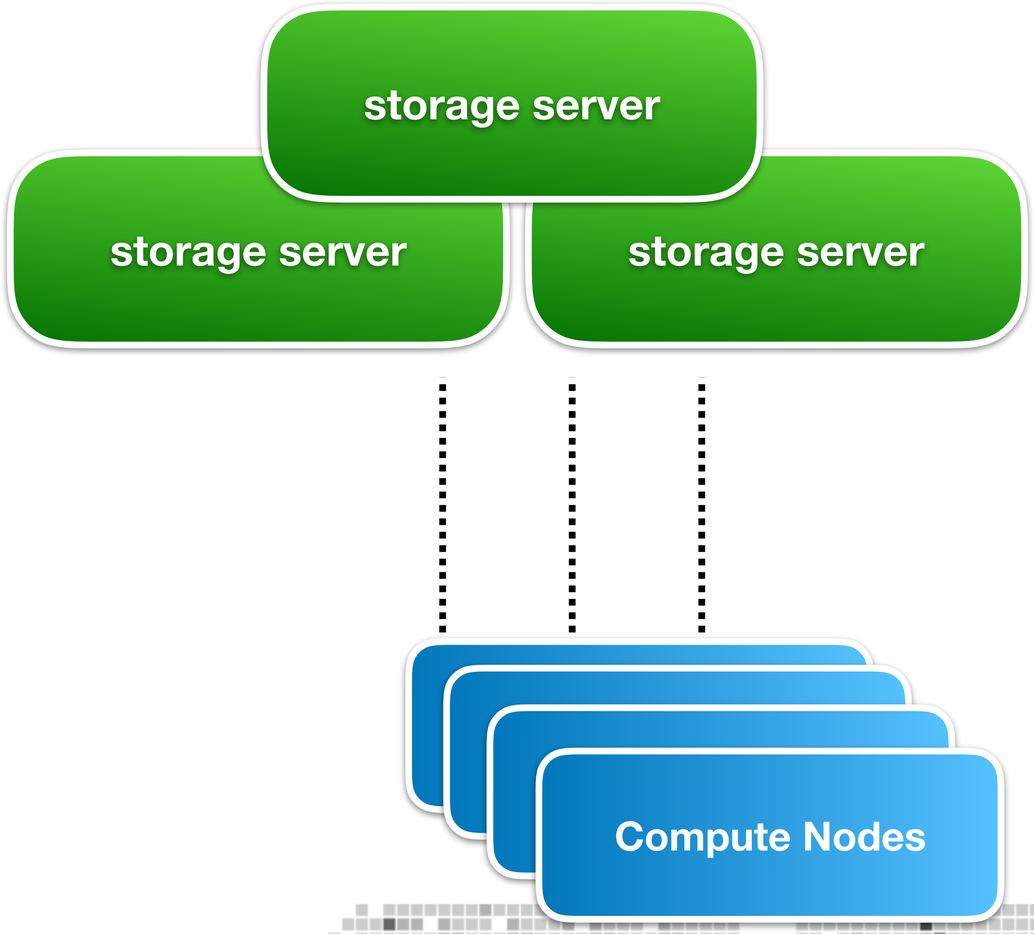
NFS server

- Challenge: Performance tied to one storage server
  - Limited capacity, space, IO operations, etc
- Can’t easily expand
  - Multiple NFS servers mean multiple namespaces
  - i.e. /home1 /home2 /home3, etc.



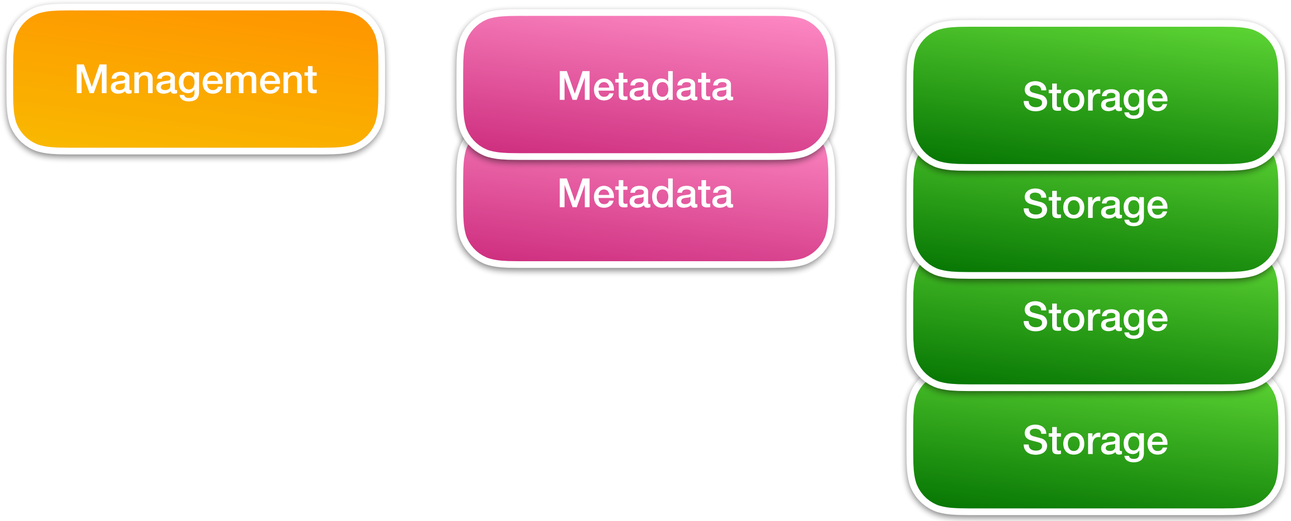
Parallel Filesystem

- Aggregate multiple storage servers into a single name space (filesystem)
  - I/O processed by multiple servers
  - Increased storage capacity
  - Easy scalability - need more performance, add more storage servers
  - Single filesystem name space
- Common options: Lustre, BeeGFS, GPFS



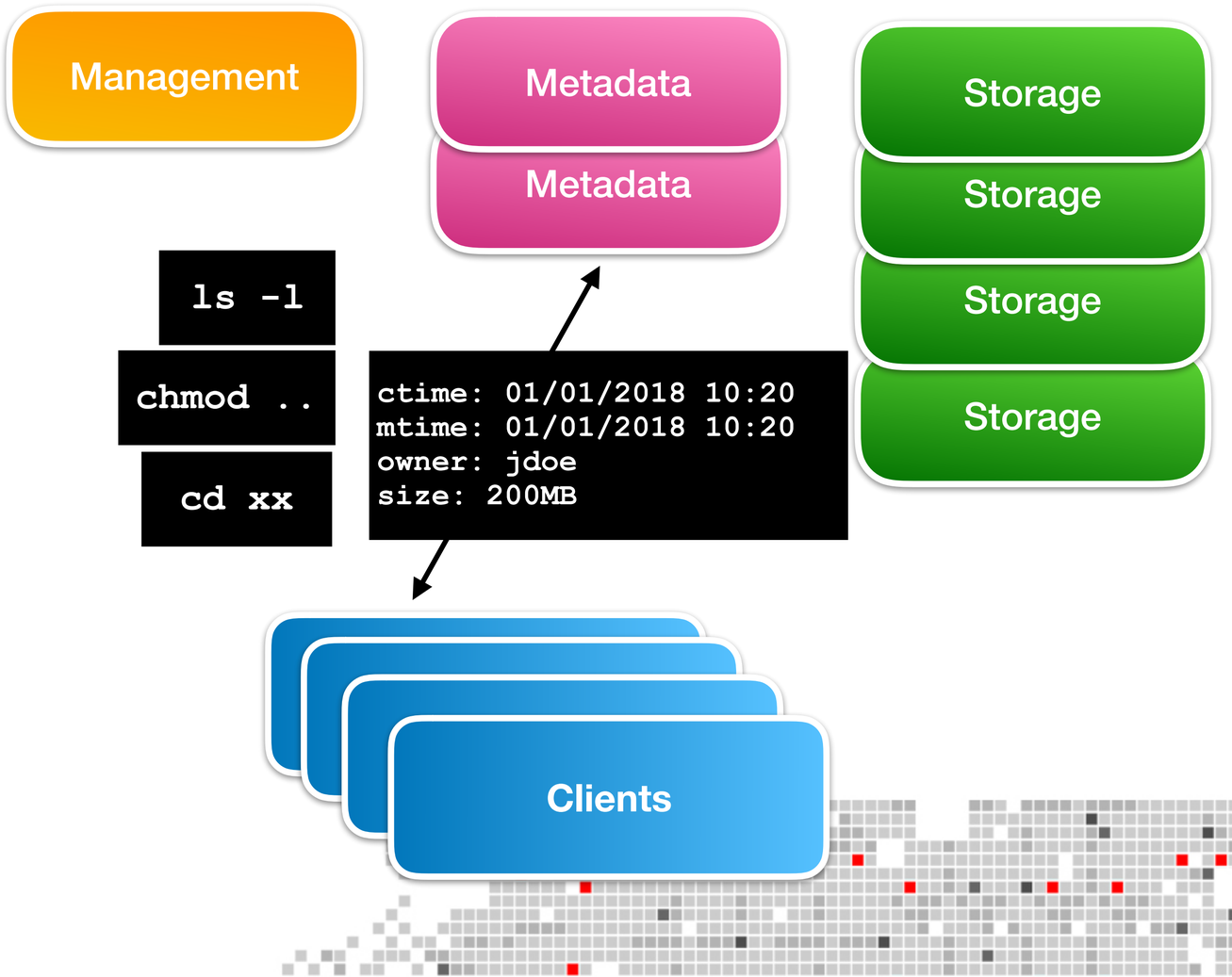
Filesystem Components

- **Services:**
  - Management Service
  - Storage Service
  - Metadata Service
  - Client Service (compute nodes)
- **Can run together, or on separate servers**



Metadata Service

- Stores information about the data
  - Directory entries
  - File and directory ownership / permissions
  - Size, update time, creation time
  - Location of chunks on storage
- Not involved in data access other than open / close



Storage Summary

- Parallel filesystems are great, but they are complicated
- Lots of components, requires skill and resources to manage effectively
- Barrier to entry is quite high, requires lots of hardware to be effective:
  - Dedicated meta data servers, with high IOPs
  - Multiple storage servers with lots of drives
- License / support costs (depending on filesystem)

GPUs

- Most popular options when looking at GPUs
  - NVIDIA Tesla A100
  - NVIDIA Tesla T4
  - NVIDIA Quadro RTX

	Tesla A100	Tesla T4	Quadro RTX 8000
Double precision floating point	9.7 TFLOPs	n/a	n/a
Single precision floating point	19.5 TFLOPs	8.1 TFLOPs	16.3 TFLOPs
Memory	40GB	16GB	48GB
Memory bandwidth	1,555GB/s	320GB/s	672GB/s
MSRP	$10,000	$2,500	$7,000

Say No to GeForce

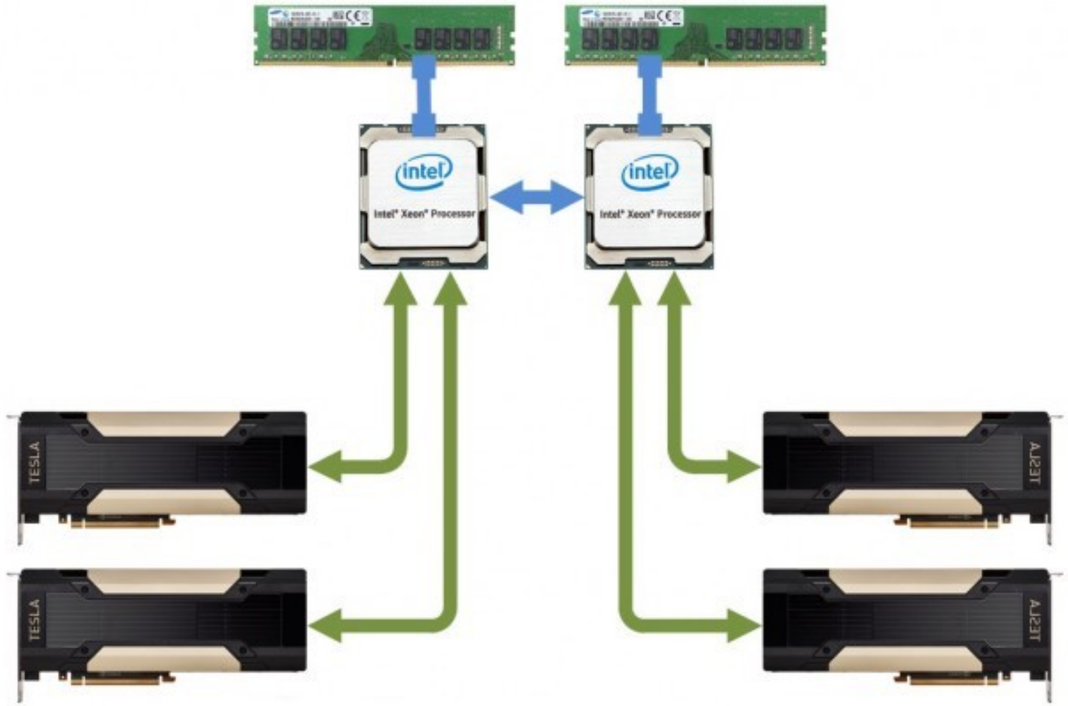
NVIDIA GeForce is for Gaming Only

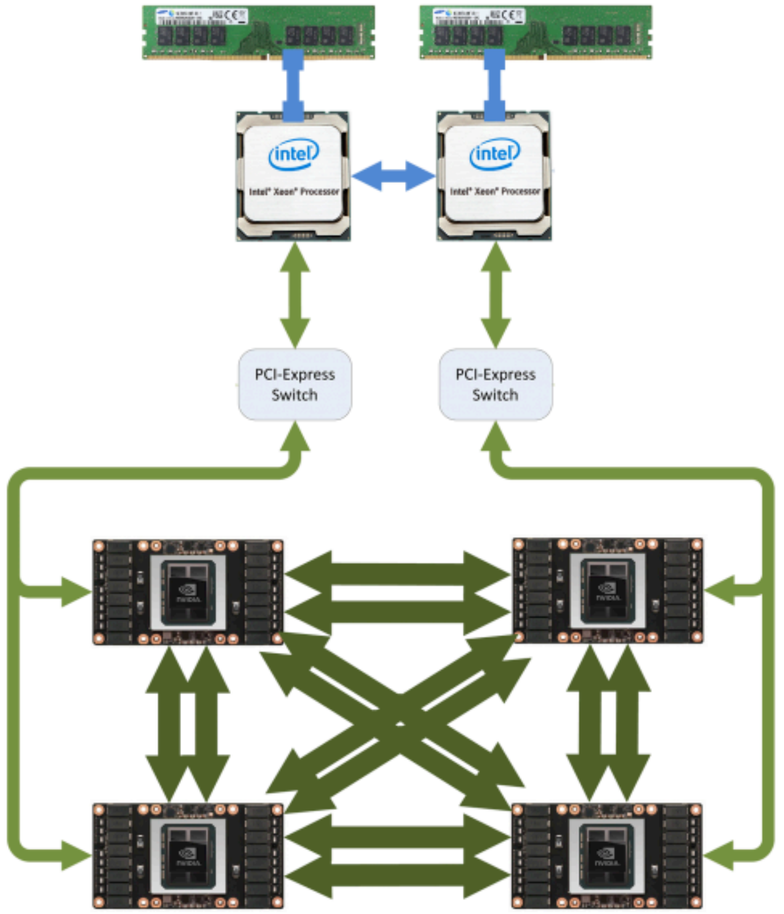
- No ECC memory, no GPU direct RDMA, no management features
- Geforce cooling is not optimized for rack mount enclosures
- Geforce delivers less performance when used with professional applications
- Data center deployment is not covered by GeForce EULA or warranty
- If you don’t want to spend money on Teslas, you can use Quadro RTX cards, which have passively cooled options.

Tesla A100

advanced clustering
technologies, inc.

PCIe vs. SXM with NVLink





advanced clustering technologies, inc.

Maximize Your Budget

8 x Single GPU Node

Model: ACTserv x2280c

- **Base system:** 2U Dual Xeon Server with support for 8x GPUs
- **Processor(s):** 2x Intel 16-Core Xeon Gold 6226R 2.9GHz - 150W
- **Memory:** 384GB - 12x 32GB DDR4 2933MHz
- **Drive configuration:** 8x 2.5" SATA hotswap drive bays
- **Storage 2.5" SATA drive bays:** 960GB Value Data Center SATA 2.5" Solid State Drive
- **GPU expansion slots:** NVIDIA Tesla A100 PCI-E 40GB GDDR5 Passive Single GPU
- **Networking:** 2x RJ45 10Gb ethernet ports
- **Management:** Remote iKVM with in-band management
- **Power supply:** Dual 2200W redundant power supply (requires 208/240V)
- **Power cables:** 2x 3ft Power Cable (C19 / C20) - Black
- **Warranty:** 3 year standard warranty

Maximize Your Budget

8 x Single GPU Node

Model: ACTserv x2280c

- **Base system:** 2U Dual Xeon Server with support for 8x GPUs
- **Processor(s):** 2x Intel 16-Core Xeon Gold 6226R 2.9GHz - 150W
- **Memory:** 384GB - 12x 32GB DDR4 2933MHz
- **Drive configuration:** 8x 2.5" SATA hotswap drive bays
- **Storage 2.5" SATA drive bays:** 960GB Value Data Center SATA 2.5" Solid State Drive
- **GPU expansion slots:** NVIDIA Tesla A100 PCI-E 40GB GDDR5 Passive Single GPU
- **Networking:** 2x RJ45 10Gb ethernet ports
- **Management:** Remote iKVM with in-band management
- **Power supply:** Dual 2200W redundant power supply (requires 208/240V)
- **Power cables:** 2x 3ft Power Cable (C19 / C20) - Black
- **Warranty:** 3 year standard warranty

$17,267.50 per node

TOTAL PRICE: $138,140.00

1 x 8 GPU Node

Model: ACTserv x2280c

- **Base system:** 2U Dual Xeon Server with support for 8x GPUs
- **Processor(s):** 2x Intel 16-Core Xeon Gold 6226R 2.9GHz - 150W
- **Memory:** 384GB - 12x 32GB DDR4 2933MHz
- **Drive configuration:** 8x 2.5" SATA hotswap drive bays
- **Storage 2.5" SATA drive bays:** 960GB Value Data Center SATA 2.5" Solid State Drive
- **GPU expansion slots:** 8x NVIDIA Tesla A100 PCI-E 40GB GDDR5 Passive GPU
- **Networking:** 2x RJ45 10Gb ethernet ports
- **Management:** Remote iKVM with in-band management
- **Power supply:** Dual 2200W redundant power supply (requires 208/240V)
- **Power cables:** 2x 3ft Power Cable (C19 / C20) - Black
- **Warranty:** 3 year standard warranty

$73,005.00 per node

TOTAL PRICE: $73,005.00

advanced clustering
technologies, inc.

GPU Configuration Extra Points

- Compute nodes that can support GPUs are different than standard compute nodes
- Most of the time you can’t just add a GPU to an existing node
- Modern NVIDIA GPUs are physically large, and require up to 300W per card - server needs the space, cooling, and power to support cards
- Systems available with 8+ GPUs per node - Nodes can draw over 3000W

Sign up for a virtual 1-on-1 meeting

https://www.advancedclustering.com/meetup/

Things to Consider

- Involve your datacenter team from the beginning
- HPC systems are not like enterprise IT - goal is to be 100% utilized with high-end CPUs, networks, etc.
- Questions for datacenter:
  - What type of power is available?
  - What is the Heat density - power/cooling limits per rack?
  - Is there a loading dock?

What's next?

- More segmentation - Intel launching 2 server Xeon platforms in 2021; AMD is on target to release followup CPU "Milan" in 2021
- Power per node is increasing
- Air cooling higher end CPUs will be impossible
  - Direct to chip liquid cooling is coming
- Rack systems are getting deeper

Power Consumption

- Most common node form-factor is a 2U server with 4 nodes
- Each node is around 500W for Intel, 600 for AMD
- So a 2U system can be between 2,000 - 2,400W
- A standard 120V 15A circuit in the wall can only provide 1800W - according to electrical code you should only use 80% of that
- Lots of data centers use 208V 30A circuits which provide about 5000W of power.
- A high-end GPU node can easily draw 3000W

Resources

- HPC pricing guide
  https://www.advancedclustering.com/act_slider/hpc-pricing-guide/
- Grant white paper
  https://www.advancedclustering.com/grant-writing-white-paper/
- RFP writing guide
  https://www.advancedclustering.com/white-paper-writing-rfp-hpc-equipment-purchases/

Cluster Management

- Advanced Clustering's ClusterVisor is our cluster management software tool, which enables you to manage hardware and the operating system plus software and networking from a single GUI
- Manage your cluster from a web interface or from the command line
- Make updates that are easily implemented cluster-wide
- Visualize your entire cluster with our built-in rack diagram
- Learn all about ClusterVisor in the webinar devoted to its features and benefits

Sign up for a virtual 1-on-1 meeting

https://www.advancedclustering.com/meetup/