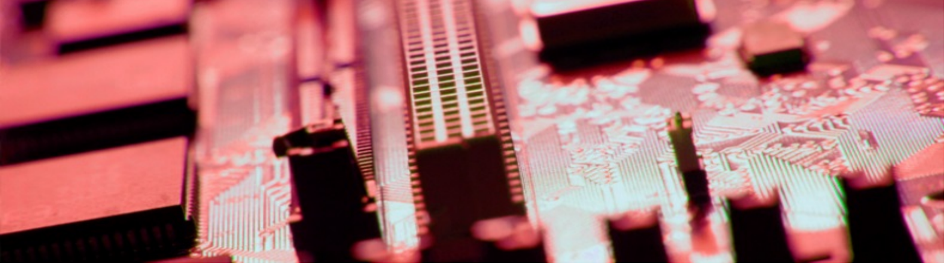


Linux Clusters Institute: Designing a Multipurpose Production Cluster

Henry Neeman, University of Oklahoma

Assistant Vice President, Information Technology - Research Strategy Advisor

Director, OU Supercomputing Center for Education & Research (OSCER)

Associate Professor, College of Engineering

Adjunct Faculty, School of Computer Science

Boomer @ U Oklahoma

874 Intel Xeon CPU chips/6992 cores

412 dual socket/oct core Sandy Bridge 2.0 GHz, 32 GB

23 dual socket/oct core Sandy Bridge 2.0 GHz, 64 GB

1 quad socket/oct core Westmere, 2.13 GHz, 1 TB

15,680 GB RAM

~360 TB global disk

QLogic Infiniband
(16.67 Gbps, ~1 microsec latency)

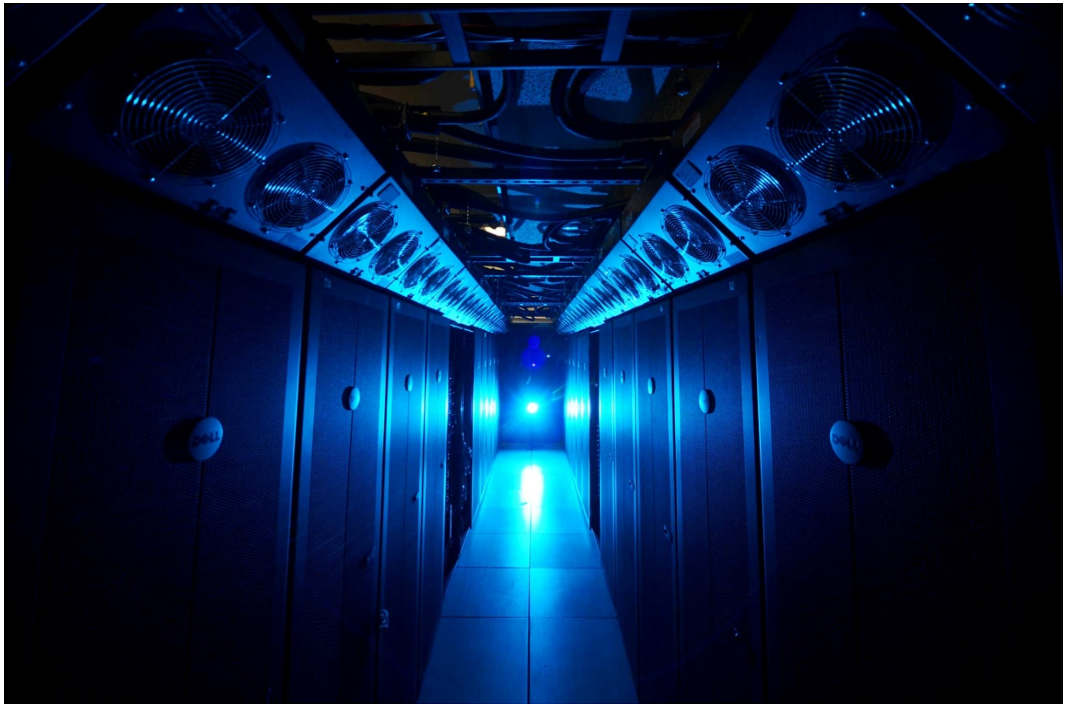
Dell Force10 Gigabit/10G Ethernet

CentOS 6

Peak speed: 111.6 TFLOPs*

*TFLOPs: trillion calculations per second

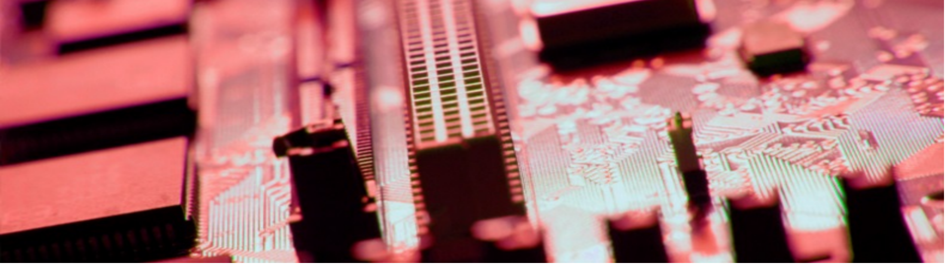
Photo: Jawanza Bassue



Just over 3x (300%) as fast as our 2008-12 supercomputer.

Just over 100x (10,000%) as fast as our first cluster supercomputer in 2002.

boomer.oscer.ou.edu



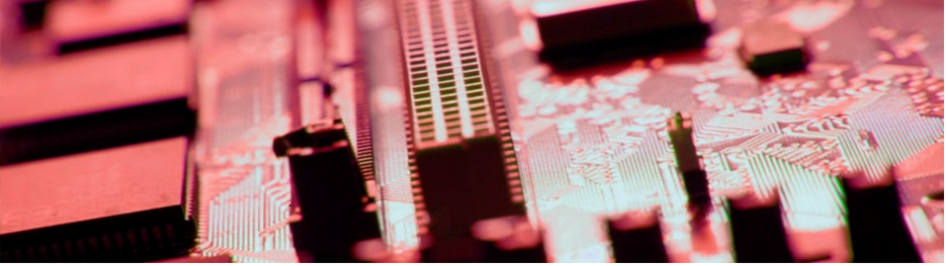
Boomer Design

Goals

- At least 3 times as fast as Sooner, Boomer’s predecessor
- ~2 GB RAM per CPU core
- Good balance of CPU, RAM, network and storage speeds
- Good balance of aggregate compute speed, RAM footprint and disk footprint
- Minimize the cost per subsystem to maximize overall system capability within a fixed budget amount
- A little bit of shared memory for those few who need it
- Reasonably manageable, to allow the OSCER operations team to devote a decent fraction of their time to needs of individual users
- Minimize single points of failure for robustness

Cluster Components

- Compute: Servers & Accelerators
- Storage
- Networks
- Software
- Racks



Compute:
Servers & Accelerators

Servers & Accelerators

- Compute nodes
- Accelerator nodes
- Accelerators
- Fat node
- Support nodes
- Diskfull nodes (storage)

Compute Nodes

- Chassis: Dell PowerEdge R620 rackmount, quantity 411
- CPU: dual Intel Xeon “Sandy Bridge” E5-2650, oct core, 2.0 GHz, memory speed 8.0 Gigatransfers/sec
- RAM: 32 GB (8 x 4 GB UDIMMs), 1333 MHz
  have 64 GB (8 x 8 GB RDIMMs), 1333 MHz
- Disk: single 250 GB, SATA, 7200 RPM, 2.5”
- Network cards
  - Infiniband: Intel (formerly QLogic) Quad Data Rate (40 Gbps) single port card (QLE7340-CK)
  - Ethernet: default network daughter card with quad GigE
- Power Supply: nonredundant 450 W
- Warranty: 3 year basic hardware replacement 10 x 5 x Next Business Day (except one with 3 year Gold 24 x 7 x 365)

http://i.dell.com/das/dih.ashx/673x448/sites/imagecontent/corporate/merchandizing/en/publishingimages/12g-image-gallery-poweredge-r620_3.jpg

Accelerator Nodes

http://i.crn.com/slideshows/2012/dell_poweredge/slide_1.jpg

Same as Compute Nodes, except:

- Intended to hold dual accelerator cards
  - 12 nodes have dual cards already
    - 9 x dual NVIDIA M2075
    - 3 x dual NVIDIA K20M
  - 12 nodes are about to have dual cards
    - 12 x dual Intel Xeon Phi 31S1P (ordered, coming shortly)
- Chassis: Dell PowerEdge R720 rackmount, quantity 24
- Disk: single 500 GB, SATA, 7200 RPM, 3.5”
- Power Supply: **redundant 1100 W**

Why Both Kinds?

- Accelerator nodes are very similar to compute nodes.
- But, they’re also a bit more expensive.
- When the number of such nodes is small (e.g., 24), that difference in price is acceptable, roughly equal to 1 node.
- But when the number of nodes is large (e.g., 435), that price difference can add up quickly – on Boomer, not having two kinds would have reduced the node count by ~7%.

Accelerators

- 2012: NVIDIA Tesla “Fermi” M2075, quantity 18
  - 448 compute cores (CUDA)
  - 6 GB GDDR5 video RAM
  - Peak compute speed 0.515 TFLOPs* double precision
- 2014: NVIDIA Tesla “Keppler” K20M, quantity 6
  - 2496 compute cores (CUDA)
  - 5 GB GDDR5 video RAM
  - Peak compute speed 1.17 TFLOPs* double precision
- 2014: Intel Xeon Phi 31S1P, quantity 24
  - 57 cores (souped-up Atom)
  - 8 GB GDDR5
  - Peak compute speed 1.003 TFLOPs*

* TFLOPs = trillions of floating point operations per second

http://videocardz.com/33058/nvidia-tesla-k20-to-be-based-on-kepler-gk110-gpu
https://encrypted-tbn3.gstatic.com/images?q=tbn:ANd9GcQq0SclRUJgPSi6TEvKK_OrZl5Rsir8Yy6_HO38ma7qop3q2Qk8lQ
http://www.servodata.com.pl/hardware/tesla/NVIDIA%20TESLA%20K20M%20by%20PNY%20Datasheet.pdf
http://www.colfax-intl.com/nd/xeonphi/31s1p-promo.aspx

Fat Node

- Chassis: Dell R910 rackmount, quantity 1
- CPU: dual Intel Xeon “Westmere” E7-4830, oct core, 2.13 GHz, 6.4 Gigatransfers/sec memory speed
- RAM: 1 TB (64 x 16 GB RDIMMs), 1066 MHz
- Disk: dual 300 GB SAS 10,000 RPM 6 Gbps, RAID1
- Network card: embedded quad GigE
- Power Supply: redundant quad 1100 W
- Warranty: 3 year, 24 x 7 x 365, 4 hour response

http://as.ideascp.com/cpwebsupport/images/products/Del_R910.jpg

Support Nodes

Some are the same as Compute Nodes, except:

- CPU: some have hex core E5-2620s (cheaper)

Some are the same as Accelerator Nodes, except:

- No accelerator cards
- CPU: some have hex core E5-2620s (cheaper)
- RAM: 64 GB (8 x 8 GB RDIMMs), 1333 MHz
- Disk: dual 500 GB, SATA, 7200 RPM, 3.5”, RAID1
- Power Supply: redundant 750 W
- Warranty: 3 year, 24 x 7 x 365, 4 hour response

Diskfull Nodes

Same as R720 Support Nodes, except:

- Chassis: R720xd rackmount
- CPU: some have hex core E5-2620s (cheaper)
- Disk: 12 x 3 TB Near-line SAS, 7200 RPM, 3.5” (19 TB useable)

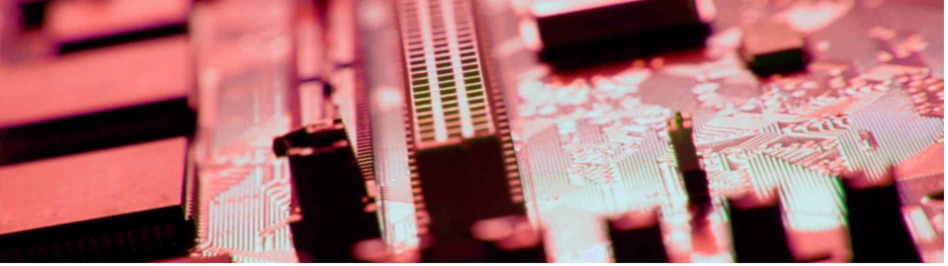
Support & Diskfull Node List

Support

- 2 x Infiniband subnet manager
- 2 x Data handler
- 2 x cluster archive
- 2 x external archive
- 3 x login
- 3 x virtualization
- 1 x web/smtp
- 1 x administrator
- 1 x backup non-diskfull

Diskfull

- 1 x backup diskfull
- 6 x home/slow scratch
- 1 x Red Hat Satellite
- 1 x Condor
- 1 x high energy physics
- 2 x spare



Storage

Storage

- Home
- Scratch
  - Slow scratch
  - Fast scratch
- Backups

Home vs Scratch

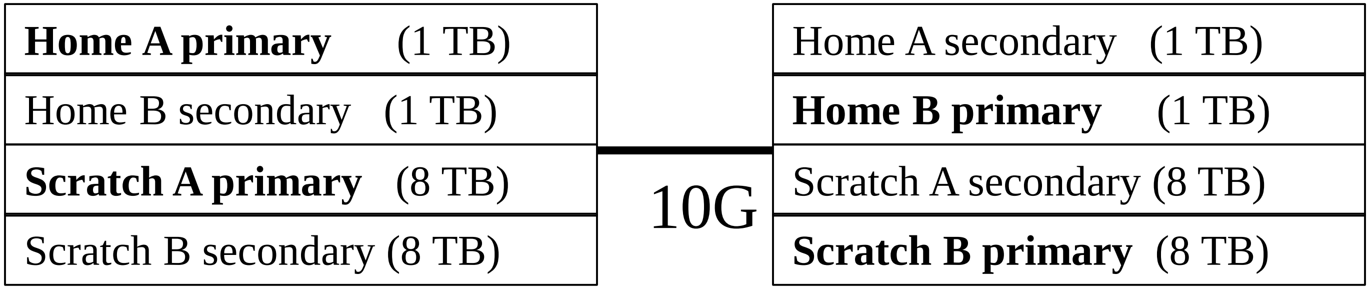
- Home: intended for application software, small input files etc – not where the bulk of a big data project would live.
  - Can be slow
  - Governed by quota limits (10 GB for most users)
  - Must be backed up nightly
- Scratch: intended for large data files, typically outputs of computational experiments
  - Some can be slow, others must be fast
  - No quota limits, just physical filesystem size
  - Not possible to do backups – too big, takes too long

Storage Reliability is the Problem

- Users don’t care about subsystems; they care about amenities.
  - Can I read the files that I’ve already created?
  - Can I write new files?
- In 2007, we did an informal study of the “TeraGrid,” a national supercomputing metacenter that then was distributed among 11 resource provider sites coast to coast.
  - Repeated the study in 2013 for XSEDE (7 resource providers).
  - **More than half of all cluster supercomputer failures (whose causes were identified in outage reports) were due to storage subsystem failures**.
- If the storage is offline, the cluster is offline.
- Conclusion: Storage reliability is crucial to productivity in computing- and data-intensive research.

Storage: Home & Slow Scratch

- 3 pairs of diskfull nodes, pairs connected at 10G, mirrored live
- Each diskfull node (12 x 3 TB) is split into multiple partitions:
  - OS: 2 drives in RAID1 (mirrored) configuration
  - 10 drives in RAID6 plus hot spare (~19 TB useable)
    - Home: 1 TB per server = 6 TB total useable space
    - Slow scratch: 8 TB per server = 48 TB total
- Home and slow scratch are mirrored on physically connected pairs of diskfull nodes.



Home A primary (1 TB)		Home A secondary (1 TB)
Home B secondary (1 TB)		**Home B primary** (1 TB)
Scratch A primary (8 TB)	10G	Scratch A secondary (8 TB)
Scratch B secondary (8 TB)		**Scratch B primary** (8 TB)

Fast Scratch

- Panasas ActiveStor11
- 6 “shelves” (enclosures)
- 20 x 2 TB SATA 7200 RPM per shelf
- Peak speed
  - Per shelf: 950 MB/sec write, 1150 MB/sec read
  - Total: 5700 MB/sec write, 6900 MB/sec read
  - NOTE: 6900 MB/sec = 55.2 Gbps

- Bidirectional RAID: RAID6 within each shelf, plus equivalent of RAID6 among the shelves
  - 240 TB raw = ~195 TB useable
- Hidden daily sync’ed copy of all home partitions (up to 6 TB)
  - “Bit rot” (unrecoverable read error) protection

http://www.computing.co.uk/IMG/384/181384/panasas-activestor-11-rack-product-shot-370x599.jpg?1308561120

Why Slow and Fast Scratch?

- Fast scratch: for users who do small numbers of large I/O transactions (almost exclusively meteorology researchers)
  - Example: Every 5 minutes, every process dumps 100 MB.
  - Performance limitation: disk bandwidth (5+ GB/sec)
- Slow scratch: for users who do large numbers of small I/O transactions (everyone else)
  - Example: Every microsecond, some process will write 1 KB.
  - Performance limitation: disk latency (~1 millisecond)
    - But actually, the RAM in the diskfull server acts as a cache for writes, so the real latency is Ethernet latency, which tends to be more like 10 microseconds – factor of 100 speedup!
    - Once the transaction has committed to the diskfull server’s RAM, the data can migrate to disk at their own rate, but that migration time isn’t experienced by the user’s application.

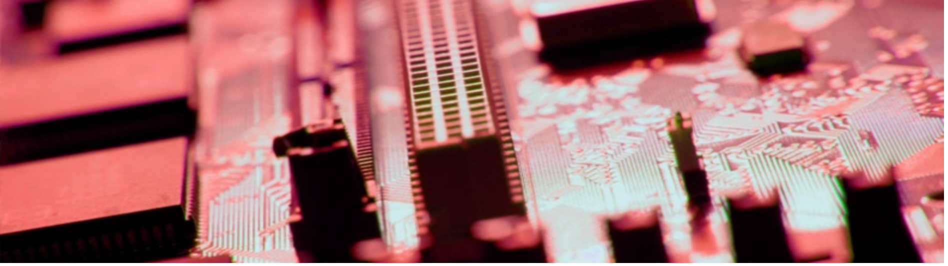
Backups

- Dell TL4000 tape library
  - dual LTO-5 tape drives
  - 48 tape cartridges = 72 TB raw capacity (more than enough)
- Software: Bacula (free open source)
- Schedule
  - Nightly incremental (all new and changed files)
  - Weekly full dumps (all files)
  - All files are retained for 2 weeks.
  - Full dump is sent offsite every few months.
- Home? YES, even though mirrored live and rsync'ed nightly
  - Need to be able to restore files that shouldn't have been deleted.
- Scratch? NO
  - Slow scratch is mirrored live.
  - Fast scratch is bidirectional RAID, so chance of failure is tiny.

http://i.dell.com/images/global/products/pvaul/pvaul_tl4000_145x161.jpg

Scratch Autopurging

- Any file in a scratch partition over 2 weeks old is “deleted.”
  - Actually it’s moved to a hidden secret directory.
- While a file is in a scratch hidden secret directory, it can be recovered on request (OSCER staff have to move it back).
- Any file in a scratch hidden secret directory for over 1 week is really deleted and then never recoverable.
- We send weekly notices, to all users who have scratch files, about what those files’ fate will be, including the timeline.
- This applies to all scratch partitions, both slow and fast.



Networks

Networks

- Infiniband
- Ethernet

Why Both Infiniband & Ethernet?

- Infiniband: very low latency (~1.5 microseconds)
    - Good for many very small messages, which is how most parallel computing works.
- Ethernet: higher latency (~10 microseconds)
    - Better latency than disk (1-10 milliseconds)
    - Cost: Much cheaper than Infiniband.
    - Separation: Keeps the I/O traffic separate from the message passing traffic – it's hard to optimize a network for both.
    - Failover: If the Infiniband fails, then the message passing can go over Ethernet, though substantially slower – a slowdown is better than not getting any work done at all.

Infiniband

- Intel (formerly QLogic) Quad Data Rate (40 Gbps peak)
- Leaf-and-core design
  - Each top-of-rack leaf switch (model 12200) is single connected to each of up to 24 compute/accelerator nodes.
- Clos network: Each switch is a small crossbar (36 ports).
- 2.4:1 oversubscribed, 16.67 Gbps peak as configured
  - Each leaf switch has 10 uplinks to the core switch.
  - If we had 12 uplinks and 24 nodes per leaf switch, that'd be 2:1 oversubscription.
  - If we had 18 uplinks and 18 nodes per leaf switch, that'd be full bandwidth.
- Single core switch: 12800-120 (up to 216 uplinks)

Why 2.4:1 Oversubscribed?

- Infiniband is expensive.
- Each component is expensive:
    - Core switch
    - Leaf switches
    - Cables (copper is expensive, fiber is very expensive)
    - Cards
- Issues governed by oversubscription rate:
    - Bandwidth (not important)
    - Injection rate (messages per second – very important)
- We found that 2.4:1 was the best balance of cost vs performance: an extra 20% injection rate (at 2:1) wouldn’t solve any meaningful problems but would cost quite a lot.

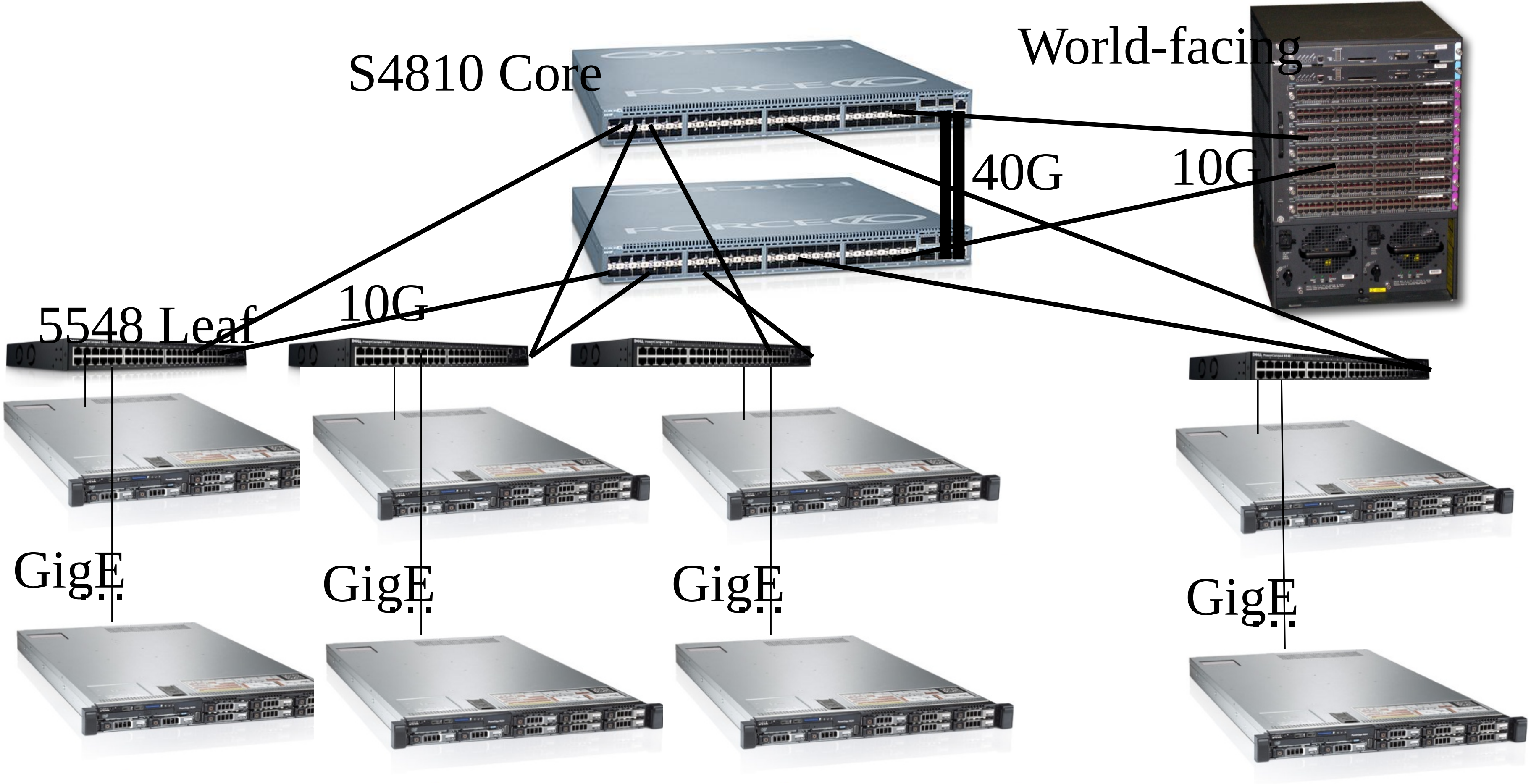
Ethernet

- I/O network
- Management network

Ethernet I/O Network

- Dell PowerConnect 5548 top-of-rack leaf switch, quantity 23
  - 48 x GigE ports, 4 x 10G uplink ports
  - Each 5548 leaf switch has up to 24 compute and/or accelerator nodes connected to it.
- Dell Force10 S4810 core switches, quantity 2
  - 48 x 10G ports, 4 x 40G ports
  - Each 5548 leaf switch is uplinked at single 10G to each of the two S4810 core switches (20G total).
    - 2 x 10G uplinks for 24 GigE node links => 1.2:1 oversubscribed
  - The S4810 switches are crosslinked at 2 x 40G using layer 2 stacking: the two S4810 switches pretend to be one big switch.

Ethernet I/O Network



http://i.dell.com/das/xa.ashx/global-site-design%20WEB/038ddbbc-c153-ae6b-2f5d-4991cc3aae96/1/OriginalJPG?id=Dell/Product_Images/Dell_Enterprise_Products/Networking_Products/Force10/s4810/hero/networking-force-10-s4810-hero-504x350.psd
http://i.dell.com/images/global/products/pwcnt/pwcnt_highlights/powerconnect-5548-overview-5.jpg
http://www.vistacomputer.net/ebay/images/6509_Config_2-SUP1_7-6348.jpg

Why GigE to Nodes, 10G to Core?

- GigE to the nodes, 10G to the core:
  - Dell offered an all-10G solution, but it was hugely expensive:
    - Many expensive 10G switches vs many cheap GigE switches
    - Many expensive 10G transceivers vs no GigE transceivers
    - Many expensive 10G cables vs many cheap Cat6 Ethernet cables
  - All of the storage collectively can only move a peak of roughly 8 GB/sec (which is 64 Gbps) – but the compute/accelerator nodes attached to the 5548 leaf switches are uplinked to the core at an aggregate of 360 Gbps, and the core switches are crosslinked at 80 Gbps.
  - Leaf-to-core (24 x GigE vs 2 x 10G) is only a 17% loss of bandwidth, at a significant reduction in cost.

Why Else This Configuration?

- Dual S4810 core switches: A switch failure reduces bandwidth but not accessibility.
  - Exception: Because of cost, fast scratch (Panasas) shelves are only single connected, all to the same core switch, so loss of that core switch would shut down Panasas until its 10G cables were physically moved to the other core switch (where 6 ports are reserved for it).
    - This is really only a problem during an ice storm – especially during Christmas break.

- Each core switch connected to the world-facing switch at 10G: robustness in case of a core switch loss, plus we already had 2 world-facing 10G transceivers on hand (they're expensive).

Ethernet Management Network

- Each 5548 top-of-rack leaf switch has 48 GigE ports.
- But, compute/accelerator nodes only consume 24 of those 48 GigE ports.
- So, the management network piggybacks on the exact same switches as the Ethernet I/O network:
  - A single 5548 switch is cheaper than a pair of 5524 switches.
  - Using a shared 5548 switch, only dual 10G uplinks per rack are needed, not quad – and 10G transceivers, both on the 5548 leaf switches and on the S4810 core switches, while individually quite affordable, add up quickly, especially for 23 top-of-rack switches with dual uplinks each.
  - Management traffic is minimal, so this doesn't conflict with I/O traffic, which is substantial.

Software

- OS: CentOS 6
- Compiler families
  - Intel
  - Portland Group
  - Numerical Algorithms Group
  - GNU (free)
- Scheduler: IBM Platform HPC Enterprise
- Parallel Debugger: TotalView
- Lots and lots of science and engineering applications
  - Mostly free open source
  - A good bit of homebrew
  - Modest commercial

Racks

- 21 racks total
- Power Distribution Units: metered – can monitor power draw

To Learn More About OSCER

http://www.oscer.ou.edu/

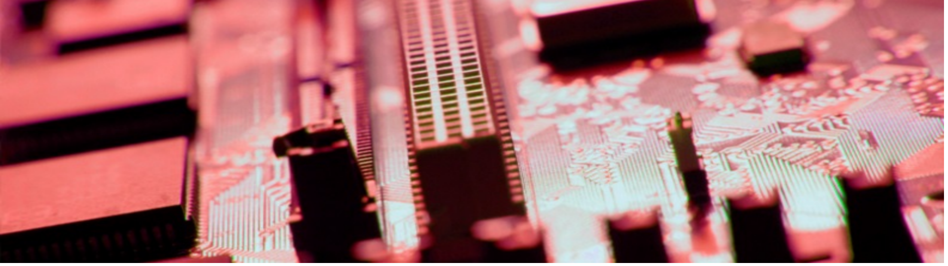
Acknowledgements

OSCER Operations Team: Brandon George, David Akin, Brett Zimmerman, Joshua Alexander

OU CIO/VPIT Loretta Early, Asst VPIT Eddie Huebsch

OU VPR Kelvin Droegemeier

OU Dean of University Libraries Rick Luce



Thanks for your attention!

Questions?

www.oscer.ou.edu