Ceph@UAB
Empowering Research

John-Paul Robinson
Research Computing
University of Alabama at Birmingham

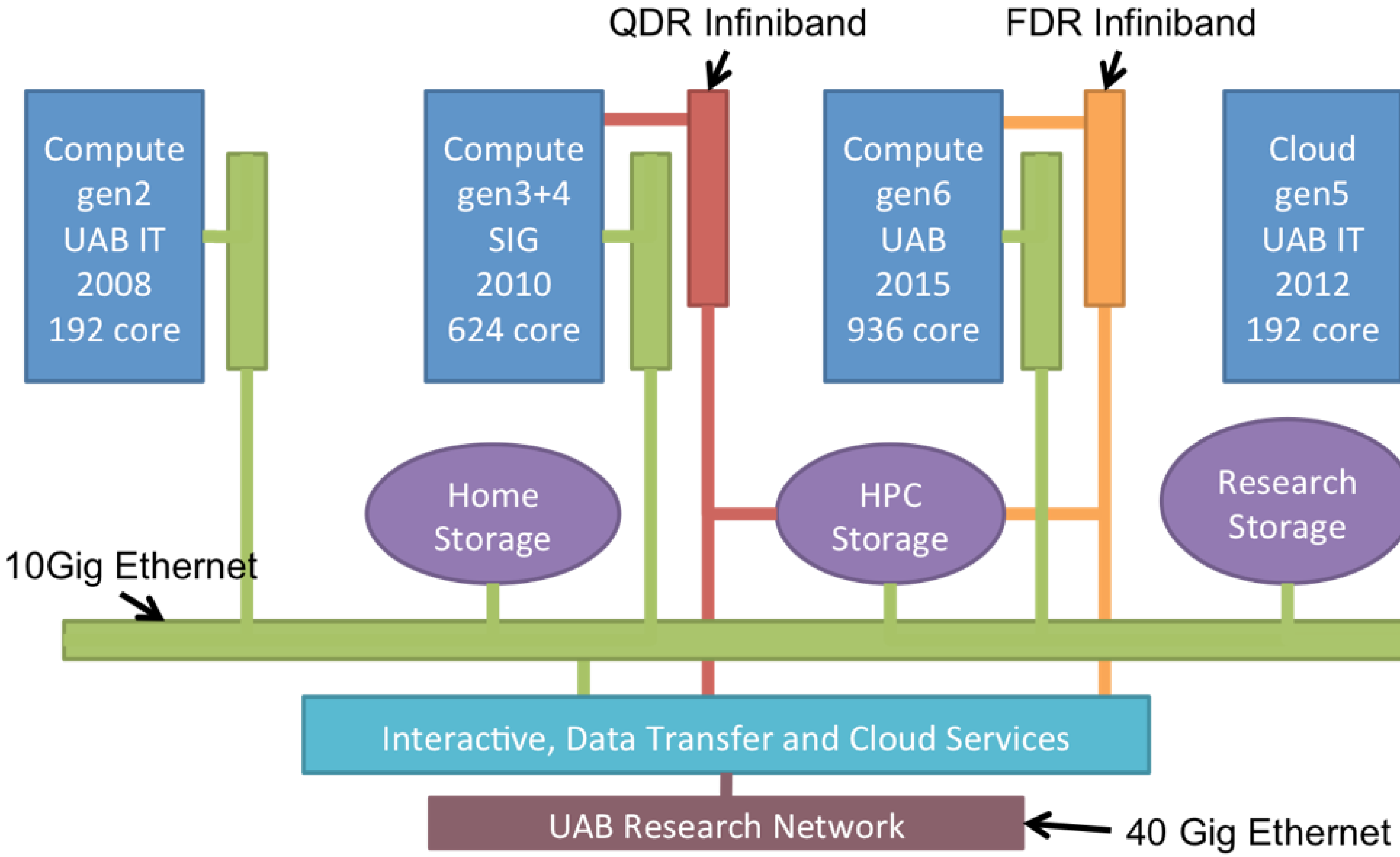
SC15 Ceph in HPC BoF
November 18, 2015

Hardware: Data Intensive Scientific Computing[1]

- 144 2.8TD drives
- 144 8GB DIMMS
- 24 8-core CPUs
- 12 Dell R720xd enclosures
- 12 dual-port Intel X510 10GigE cards
- 1 48-port S4820 10g switch
- Acquired 2012Q3

[1] Data Intensive Scientific Computing (DISC)

Research Computing System


QDR Infiniband
FDR Infiniband
Compute gen2 UAB IT 2008 192 core
Compute gen3+4 SIG 2010 624 core
Compute gen6 UAB 2015 936 core
Cloud gen5 UAB IT 2012 192 core
Home Storage
HPC Storage
Research Storage
10Gig Ethernet
Interactive, Data Transfer and Cloud Services
UAB Research Network
40 Gig Ethernet

Experiment 0: Emergency Storage

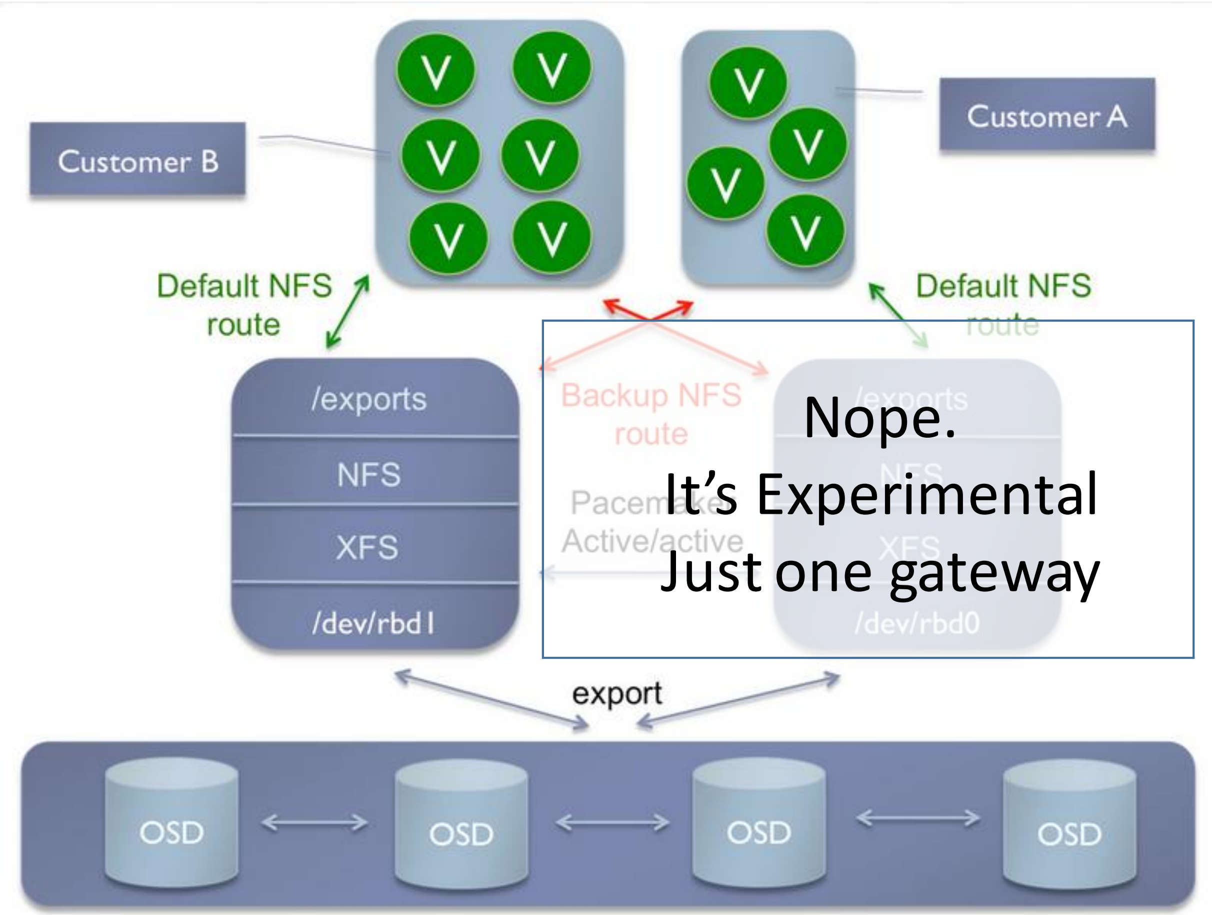
- Hardware arrival coincided with unexpected primary scratch file system failure
- Used Gluster to span across 4 nodes: 48 disks, no compute
- Worked very nicely and got us out of a pickle
- Returned to our mainline scratch a few weeks later
- Left 24 disks assigned via Gluster to our Galaxy science gateway, but no compute

Experiment 1: Ceph + OpenStack

- 2012Q4 learned Dell was working with OpenStack and a little gem called Ceph
- Ceph was exactly what we needed to solve our disk aggregation problem
- Crowbar/OpenStack/Hadoop is what we wanted to solve our compute needs
- PoC: 6 x Dell R720xd
  - 4 R720xd Ceph nodes: 44 osd -- leaving 64 cores and 384GB RAM untaxed
  - 2 R720xd Nova nodes: 32 cores and 192GB of RAM -- leaving 20 disks (56TB) unused
- Today in Ceph:
  - 77 osds: 77 up, 77 in
  - 65644 GB data
  - 128 TB used
  - 67200 GB / 193 TB avail (66% used)

Experiment 2: Research Storage

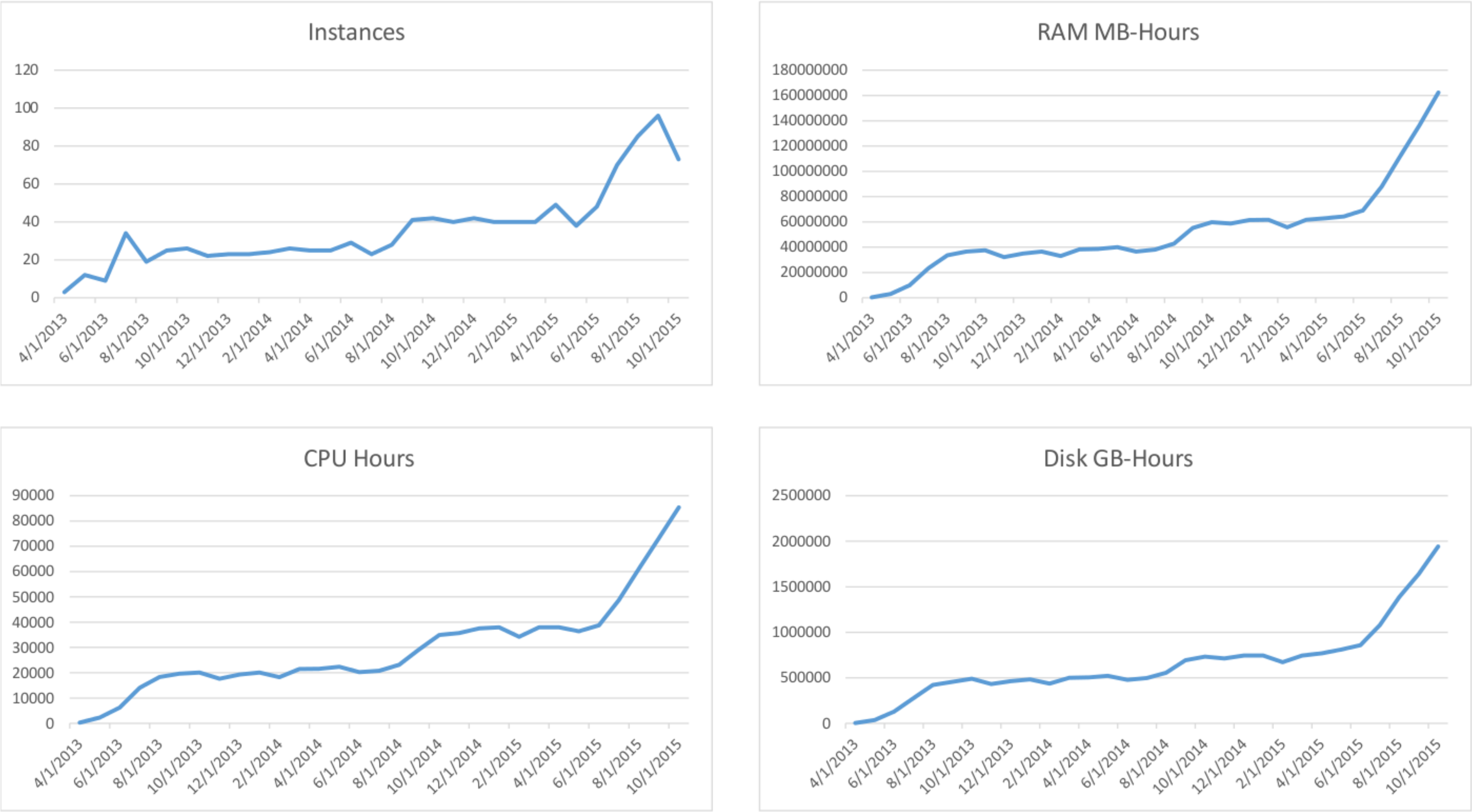
- Thin provision RBD devices
- RBD-to-NFS gateway using rbd and nfsd kernel module (Thanks Sebastien Hann!)
- Expose research storage drives to HPC cluster
- “First Terabyte Free”
- Today:
  - 259 containers
  - 268 TB allocated
  - 46 TB used
  - Almost 6:1 over provisioning
- Those who use it fill it!



Nope.
It’s Experimental
Just one gateway

Experiment N: Developers Love Freedom

Research Computing OpenStack Usage April 2013 to October 2015



Instances

RAM MB-Hours

CPU Hours

Disk GB-Hours

The Lessons

- Early adoption means fixing things
    - Network drivers, to get clean 10G performance
    - Gdisk and ceph-disk-prepare, partition bug cost 30% storage
- Don’t provision too many things into one container
    - a.k.a. know how you pack your containers so you can safely repack them
    - Leads to underutilized compute
    - Semi-static provisioning and using storage (data is not easy to move)
- Have enough hardware
    - Set aside enough to treat as single function black box storage (production)
    - HPC prod and dev tend to sit on top of each other
    - Without the dynamic provisioning context of an SGE or SLURM job this is more difficult to manage
- Storage acquisition is not the problem
    - It's easy and cheap to buy disks
    - IT processes tend to be the problem: how deployed, how maintained, how charged
- Moving forward
    - Buying more storage: all-in-wonder or developer toolbox?
    - Ceph dynamic provisioning model is crucial
    - OpenStack is central to provisioning compute, network, and storage
    - Want better use of compute