



Computation Institute

Experiences with Ceph at the US ATLAS Midwest Tier 2 Center

Lincoln Bryant

Computation and Enrico Fermi Institutes
University of Chicago



THE UNIVERSITY OF
CHICAGO

efi.uchicago.edu
ci.uchicago.edu

About our facility



- ATLAS Midwest Tier 2 (MWT2)
 - 6800 job slots
 - 3.5 PB of dCache storage
 - 3 sites: University of Chicago, Indiana University, and University of Illinois
- University of Chicago ATLAS Tier 3
 - 330 job slots
 - 110 TB of XRootD storage
- University of Chicago Computing Cooperative (UC3)
 - Campus Grid
 - 500 dedicated job slots in our server room, more elsewhere on campus.
 - 50 TB of HDFS storage



Motivating Factors



- Started out as a plan to repurpose old worker nodes.
 - End of life hardware, decommissioned from production
 - Good testbed for a distributed filesystem designed for commodity hardware.
- HDFS was an option, but this was a good opportunity to evaluate something new.
 - CephFS is fully POSIX! Why not try it?



MWT2 Ceph Prototype



- Proof of concept with retired compute nodes.
- 10 storage hosts
 - Dual Opteron 285s @ 2.6GHz
 - 8GB RAM
 - 1Gbps NICs
 - 6 OSDs x 750GB storage
- 3 monitor hosts
 - Dual Opteron 275s @ 2.2GHz
 - 8GB RAM
 - 1Gbps NICs
- Overall, about 45TB raw



Lots of old disks!

MWT2 Ceph Prototype



- Installation was easy.
 - Up and running in just a few days.
 - Single `ceph.conf` file for the whole cluster, simple management.
 - Folks in the community have written both Puppet and Chef modules
- Performance was OK, considering the hardware.
 - Able to get around 80MB/s writes and 110MB/s reads through RADOS internal benchmark.
 - Performance could almost certainly be improved with better disks, SSD journals, and faster networking.



MWT2 Production Ceph



- Prototype considered a success – decided to buy some dedicated hardware.
- 2 Dell R510s
 - Dual Intel X5660s
 - 96 GB RAM
 - 2 PERC H800s, each with 2 MD1200 shelves
 - 2 SSDs in RAID 0
 - Total of 56 disks per host.
- 420TB raw, using 2x object replication





Benchmarks



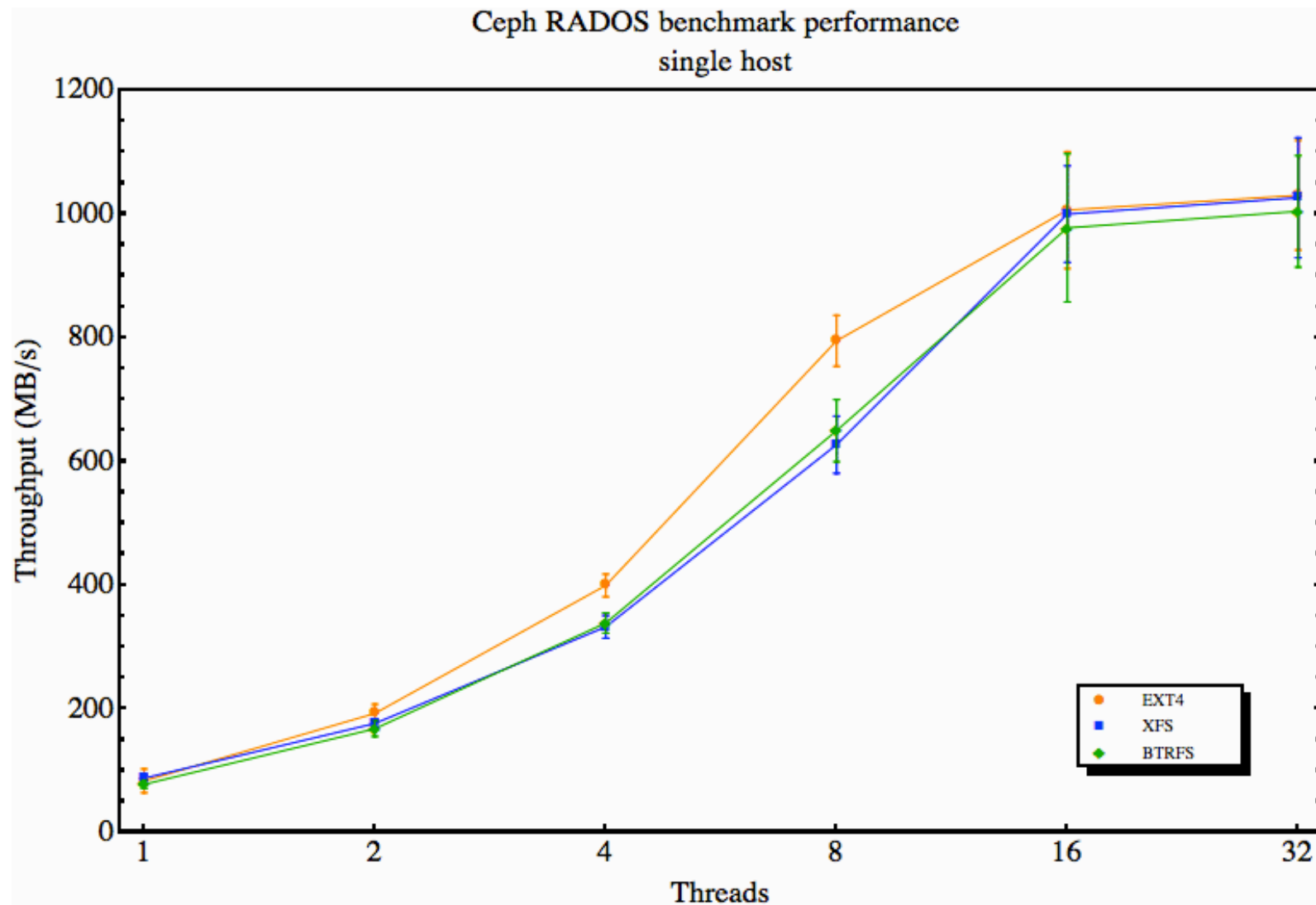
Ceph benchmark setup



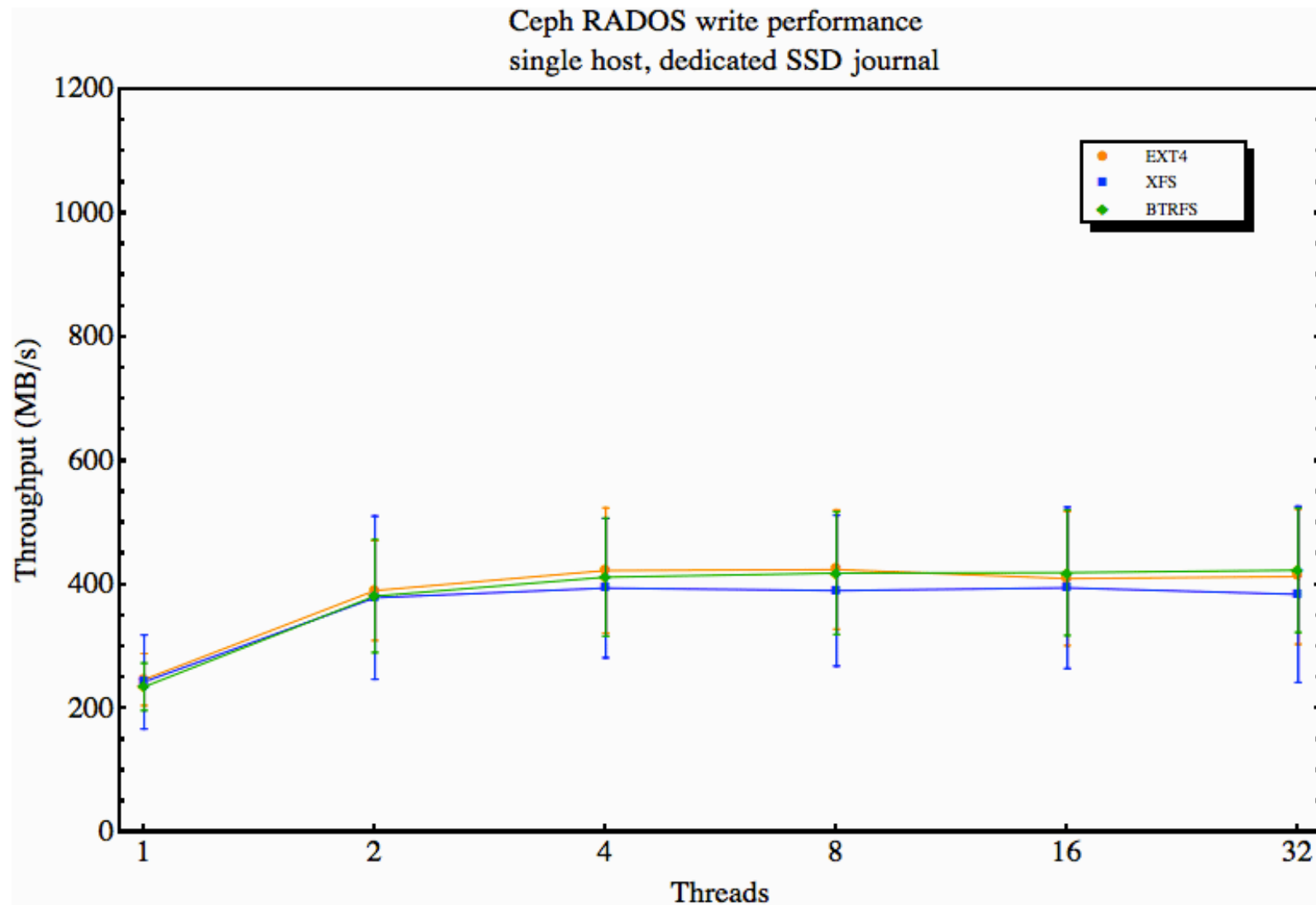
- After each benchmark, caches were flushed and disks were synced.
- These benchmarks were using only the new R510 hardware. The prototype machines weren't included in these tests.
- For the XRootD read tests, files were read by a 10Gbps connected host.
- For the XRootD write tests, files were copied from RAM disk to CephFS via a 10Gbps host.



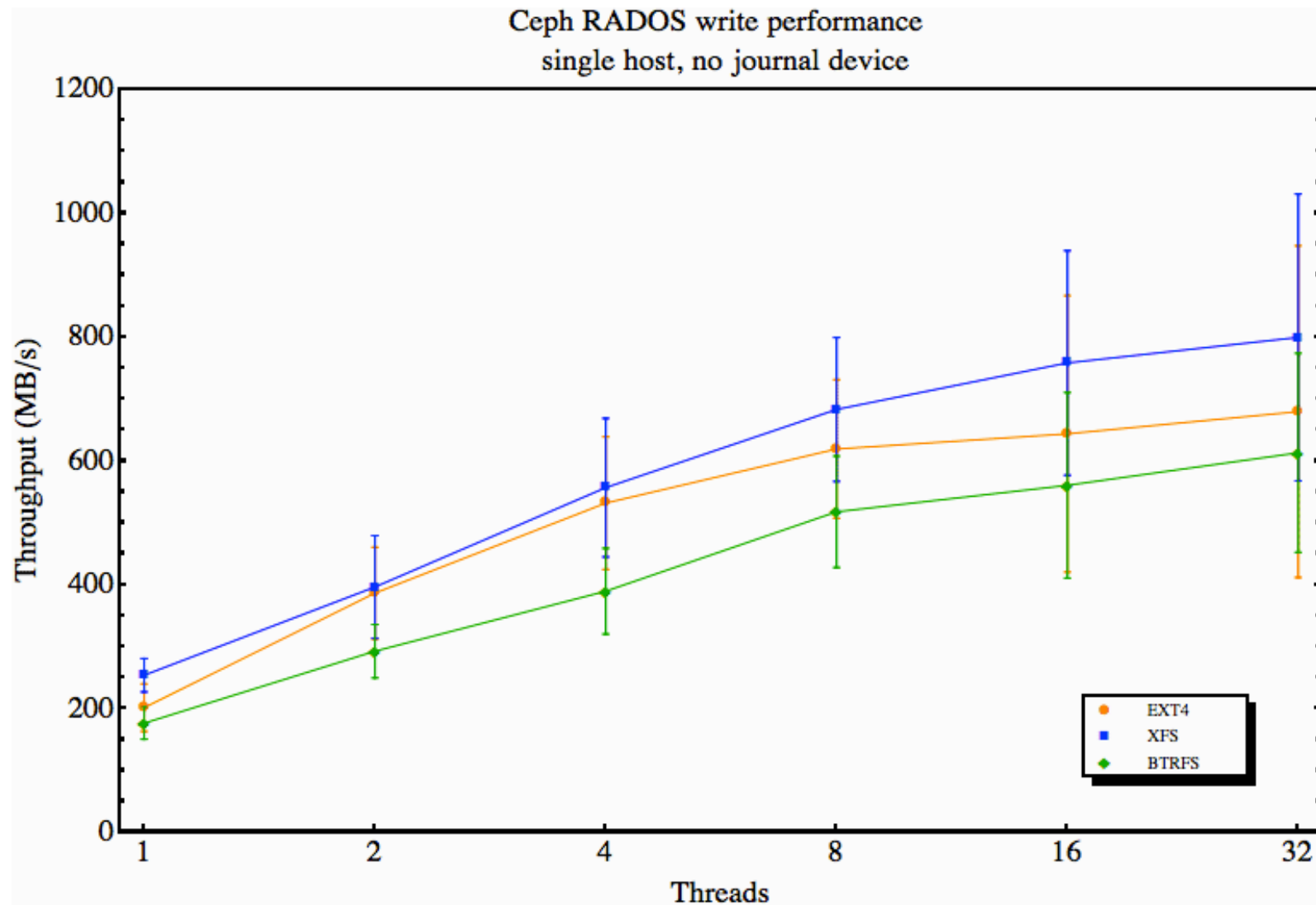
Ceph Reads – Internal Benchmarking



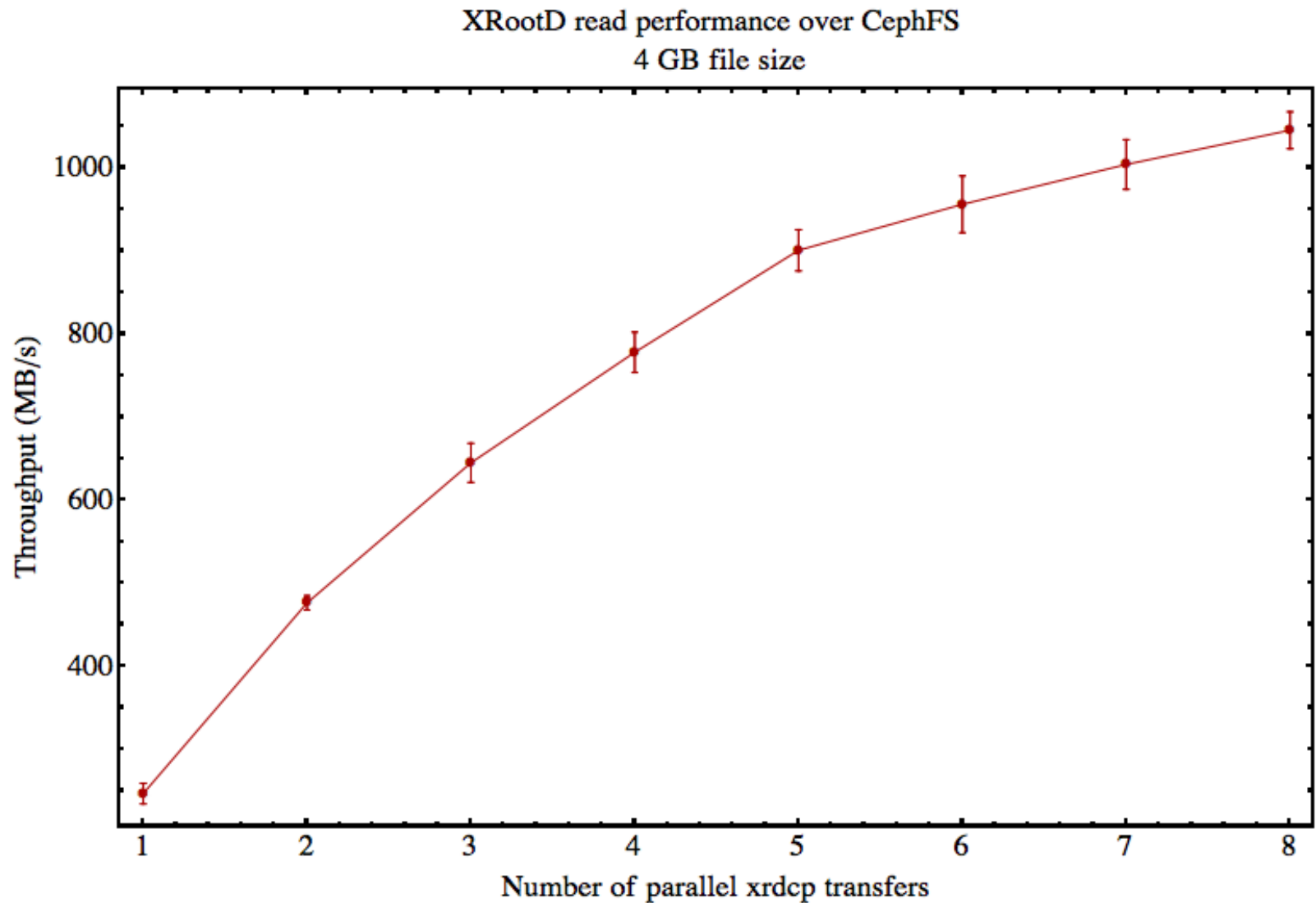
Ceph Writes – Internal Benchmark



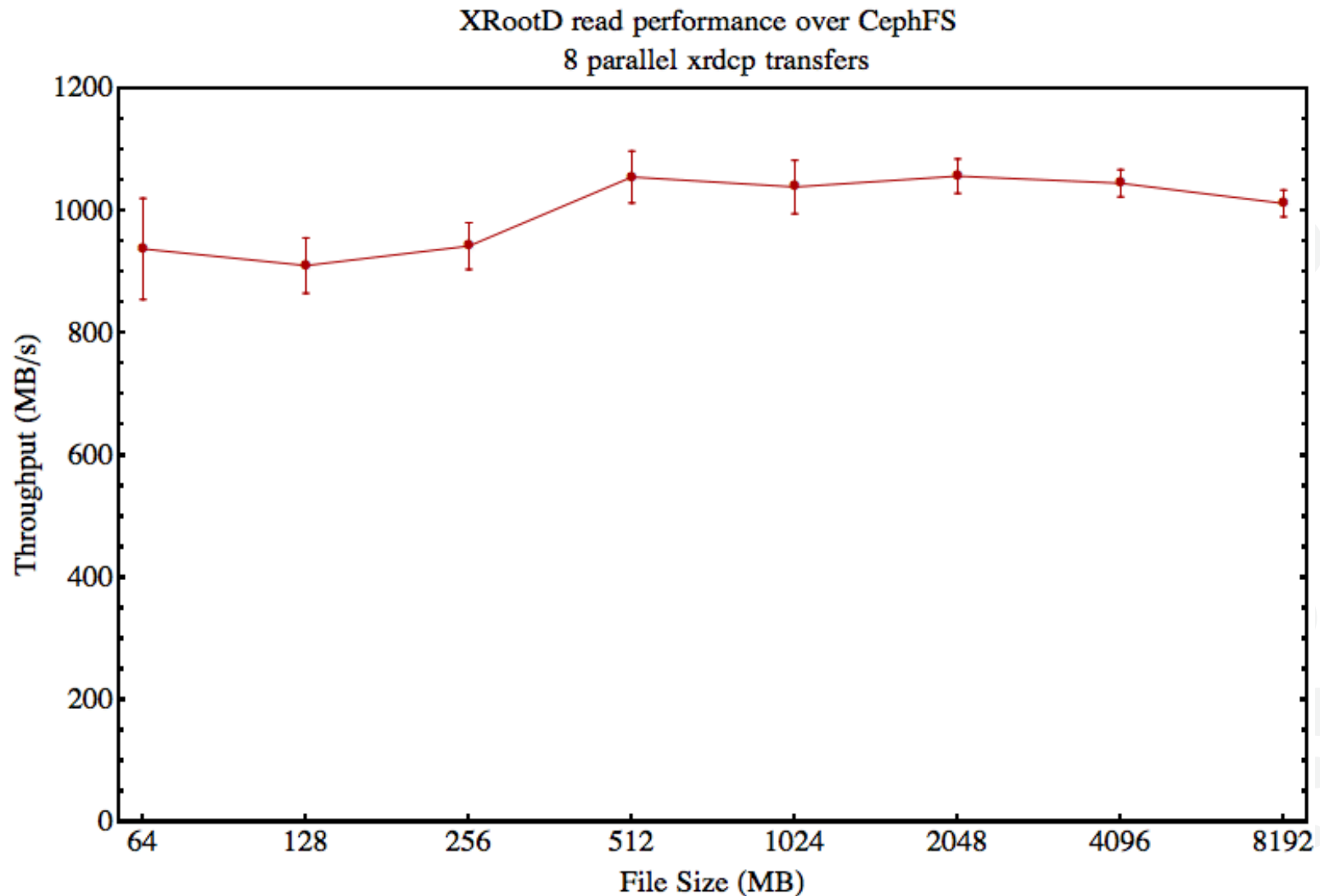
Ceph Writes (part 2) – Internal Benchmark



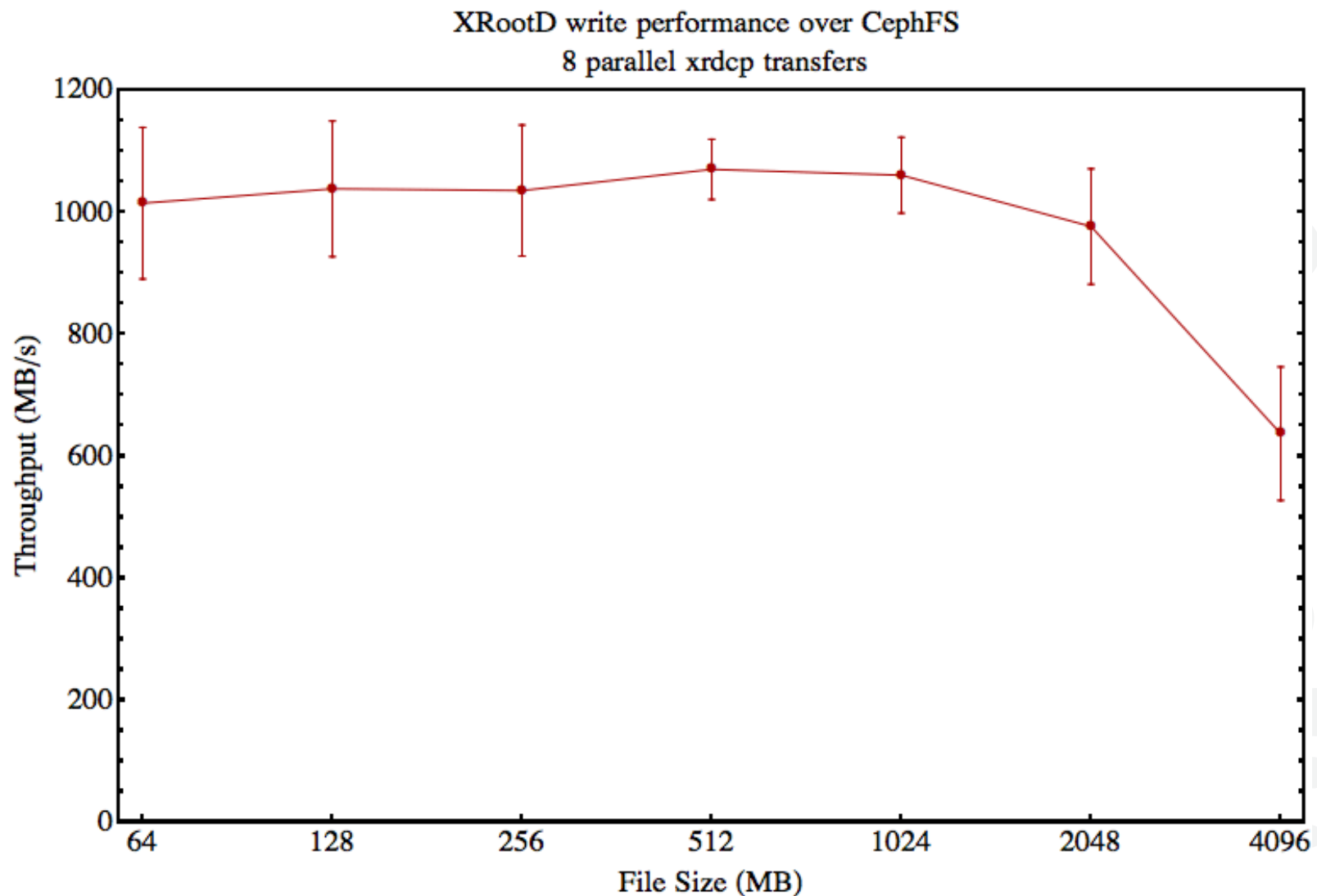
Ceph Reads – XRootD, variable # of threads



Ceph Reads – XRootD, variable file sizes



Ceph Writes – XRootD, variable file sizes



Benchmark Summary



- Two SSDs in RAID 0 were definitely not sufficient for 56 disks
 - Common wisdom is to use 1 SSD per 5-6 spinners.
- Very little performance variation in reading files of various sizes.
- Ceph performance generally increases as the workload becomes more parallel
 - We are probably bottlenecked by single 10 Gbps links. Need to set up 20 Gbps bonds.





Projects utilizing Ceph



FAXbox: from the grid to your laptop



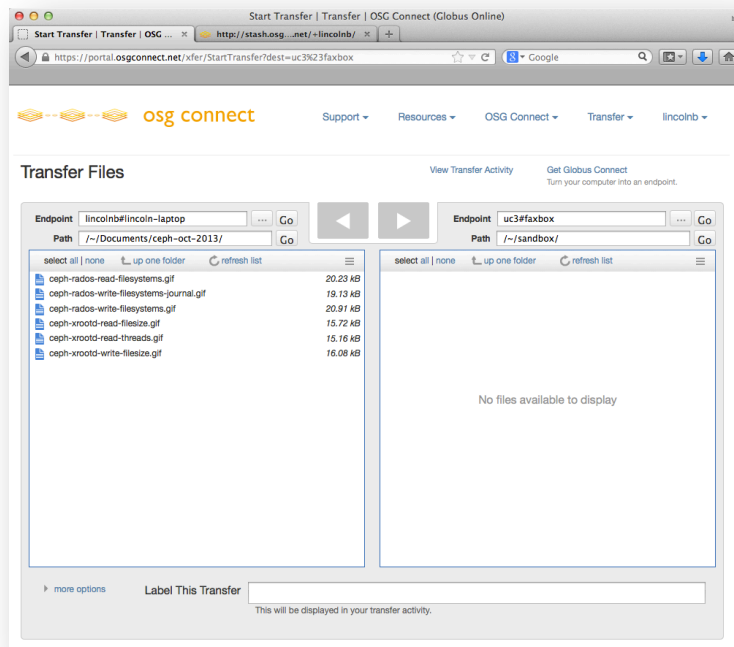
- Thanks to high speed WAN connections and Federated ATLAS XRootD (FAX), jobs no longer have to be co-located with data.
 - Great for Tier 3 users! But where does the output go?
- Solution: FAXbox
 - General purpose storage for analysis computing workflows
 - Data can be accessed via XRootD or Globus Online
 - Backed by Ceph.



Stash: Data served up however you like.



- Generalization of the FAXbox service for those outside of the HEP community.
- Move data to/from the grid with Globus Online, HTTP, Chirp/Parrot, XRootD, etc.



OpenStack and Ceph at MWT2



- We have a lot of hypervisors in the MWT2 data center. (T2, T3, Campus Grid, CI Connect, etc.)
- Idea: Use Ceph to back our virtual machine infrastructure.
 - Already an industry-proven solution.
 - Shared storage allows for live migration
 - VM images live in the Ceph pool
 - Access to the object store via radosgw
- Prototype is running. Plans for full deployment forthcoming.





Looking forward and conclusions



Some words of warning



- Metadata servers do not currently scale
 - Only one MDS is supported at this time, no redundancy
- Quotas aren't currently supported.
- Using a bleeding edge kernel (3.12 RC1+) is essentially required for machines mounting CephFS.
 - Hopefully EL7 will support kernel 3.12.



Future plans / open questions



- Our campus grid has 50 TB HDFS cluster that is looking pretty lonely. Why not convert it to Ceph and merge it into our pool?
- How much work would it be to add a BeStMan gateway to our Ceph cluster?
- Ceph replicates objects in the event of a disk failure, but what are the best practices to minimize the time in the danger zone?
- Has anyone plugged Ceph into various monitoring/reporting frameworks? Nagios, Cacti, Ganglia etc?



Conclusions



- Ceph is most performant under highly parallel workloads.
- Since Ceph is POSIX-compliant, we've had no trouble putting arbitrary transfer mechanisms on top.
- MWT2 will definitely continue to pursue Ceph for future projects.



Questions?



(Extra slides)



CephFS - Kernel bugs!



- Ran into a problem with files disappearing/reappearing. Should be fixed for kernel 3.12
 - We compiled our own kernel with the appropriate patches cherry picked out of git.

