

Investigation of storage options for scientific computing on Grid and Cloud facilities

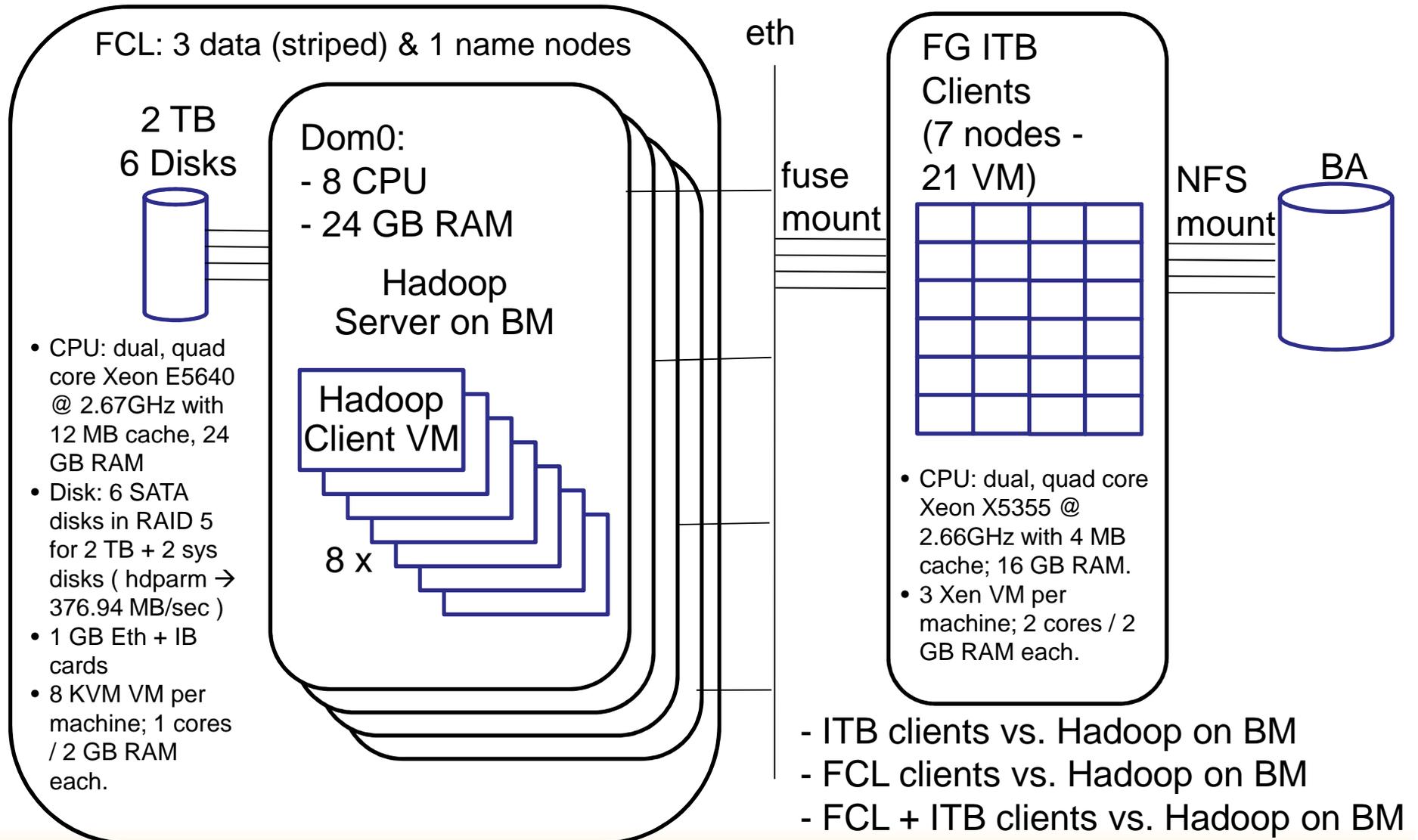
Overview

- Hadoop Test Bed
- Hadoop Evaluation
 - Standard benchmarks
 - Application-based benchmark
- Blue Arc Evaluation
 - Standard benchmarks
 - Application-based benchmark

Jun 29, 2011

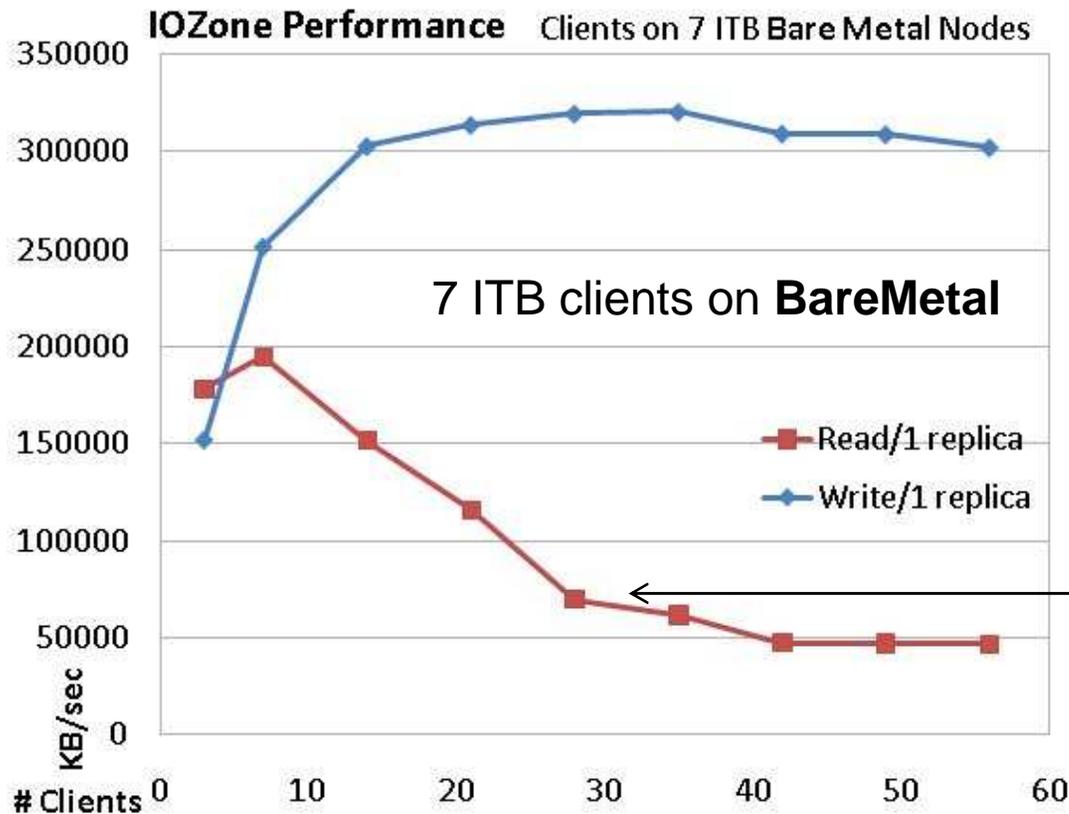
Gabriele Garzoglio & Ted Hesselroth
Computing Division, Fermilab

Hadoop Test Bed: FCL Server on BM



Hadoop R/W BW – Clients on BM

- Testing access rates with different replica numbers.
- Clients access data via Fuse. Only semi-POSIX.
 - root app.: cannot write; untar: returned before data is available; chown: not all features supported; ...



Hadoop shows poor scalability in our configuration.

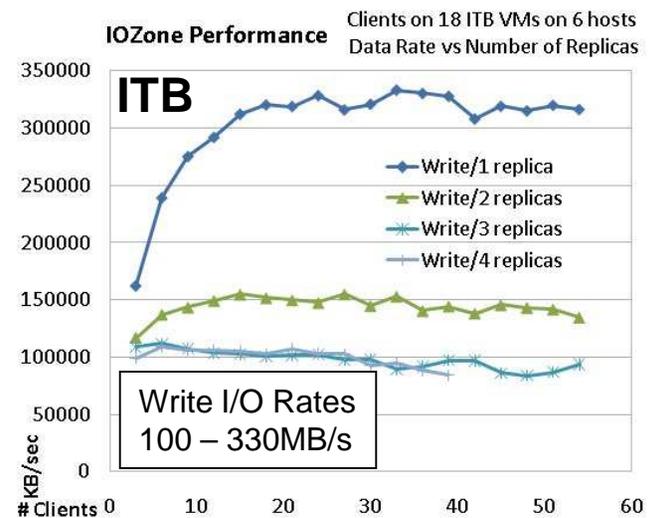
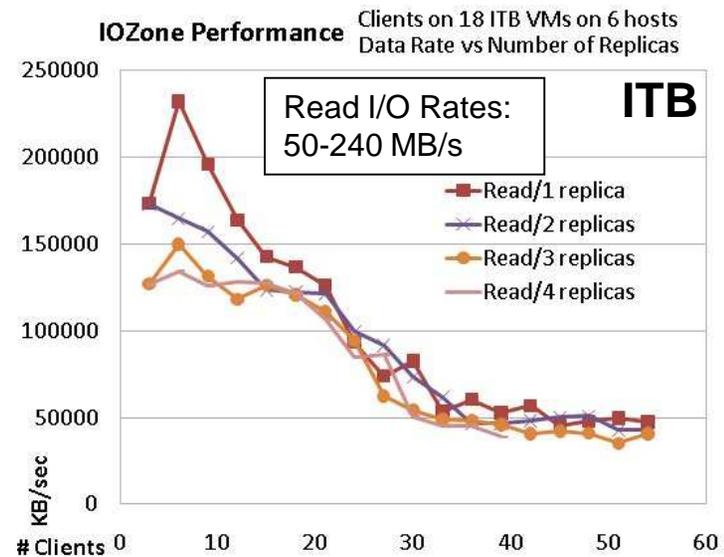
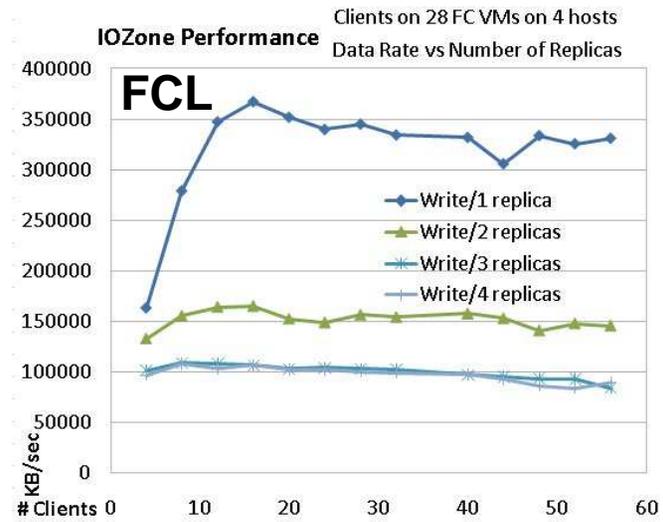
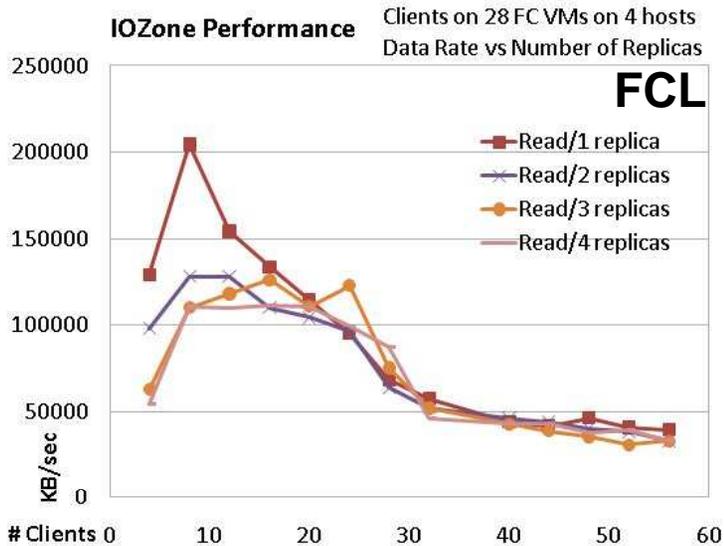
Hadoop 1 replica

50-200 MB/s read
320 MB/s write

Lustre on Bare Metal srv. was
350 MB/s read
250 MB/s write

Aggregate Read BW decreases with number of clients.

Hadoop scalability – Clients on VM



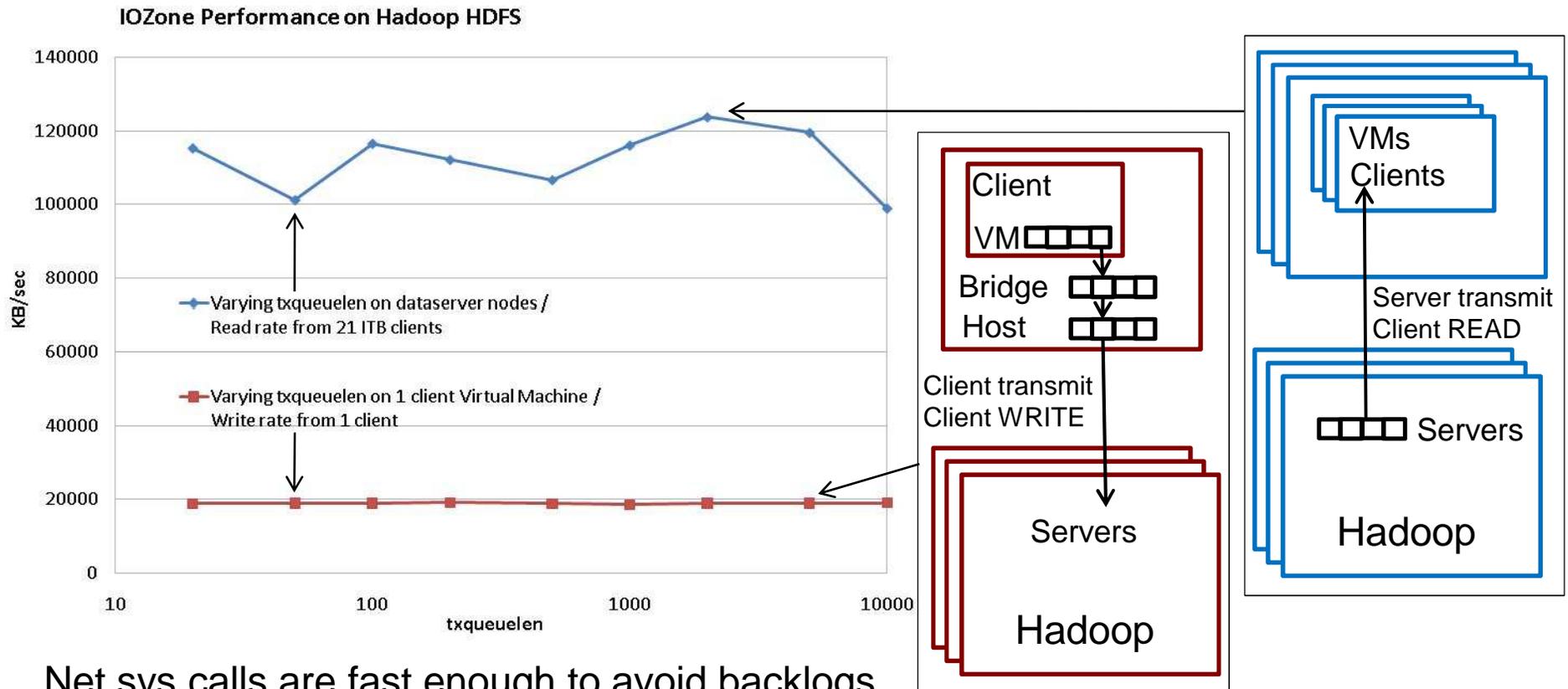
How does Hadoop scale for...

- Read / Write aggregate client BW vs. num processes
- from FCL and ITB
- with a different number of VMs and hosts

Hadoop shows poor scalability in our configuration.

Hadoop: Ethernet buffer tuning

Can we optimize transmit eth buffer size (txqueuelen ) ?
 At the client VM* for client writes AND at the servers for client reads ?



Net sys calls are fast enough to avoid backlogs from 21 hadoop clients for all txqueuelen

Varying the eth buffer size does not change read / write BW.

* txqueuelen on the physical machine AND network bridge at 1000

On-Board vs. Ext clnts vs. file repl. on B.M.

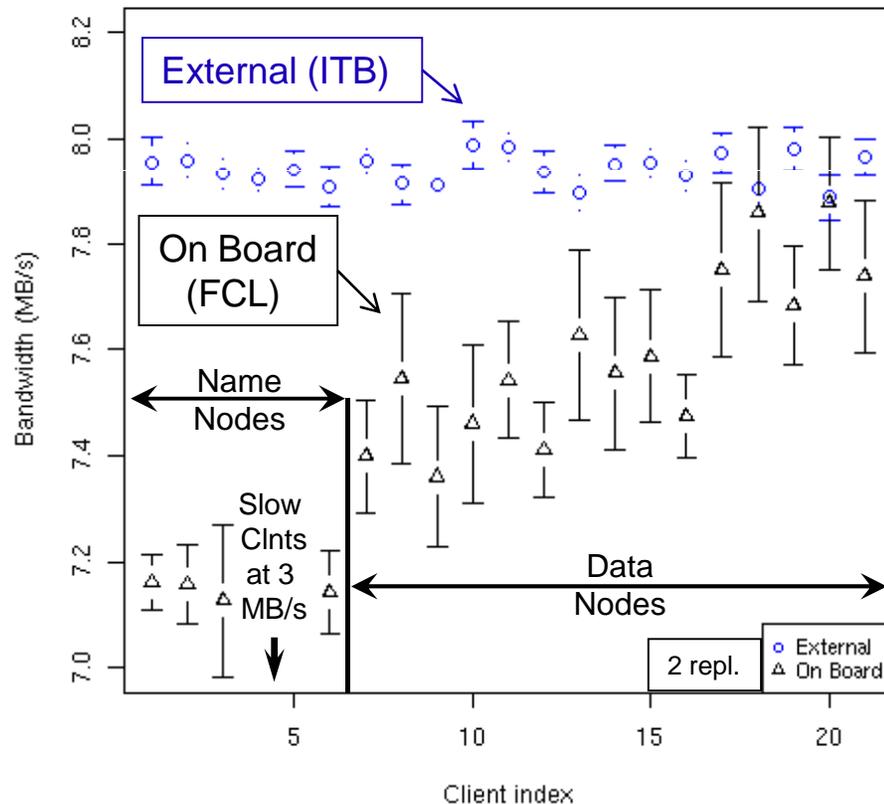
- How does read bw vary for on-board vs. external clients?

Ext (ITB) clients read ~5% faster than on-board (FCL) clients.

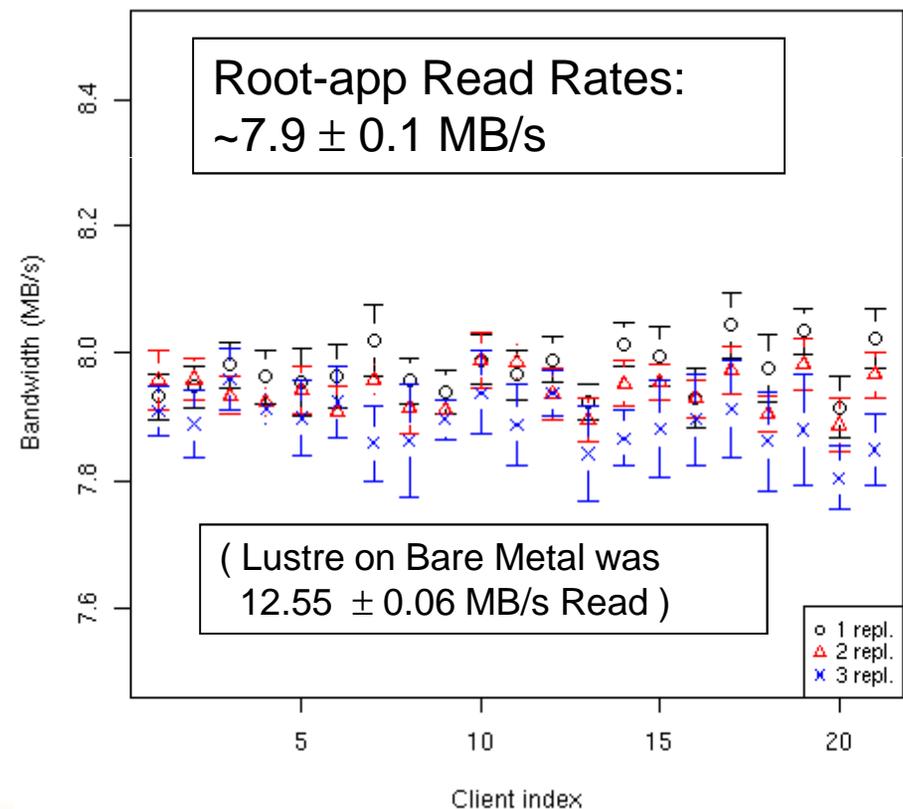
- How does read bw vary vs. number of replicas?

Number of replicas has minimal impact on read bandwidth.

Ave Bandwidth with 21 ITB and 21 FCL nova clients
Hadoop Server on Bare Metal



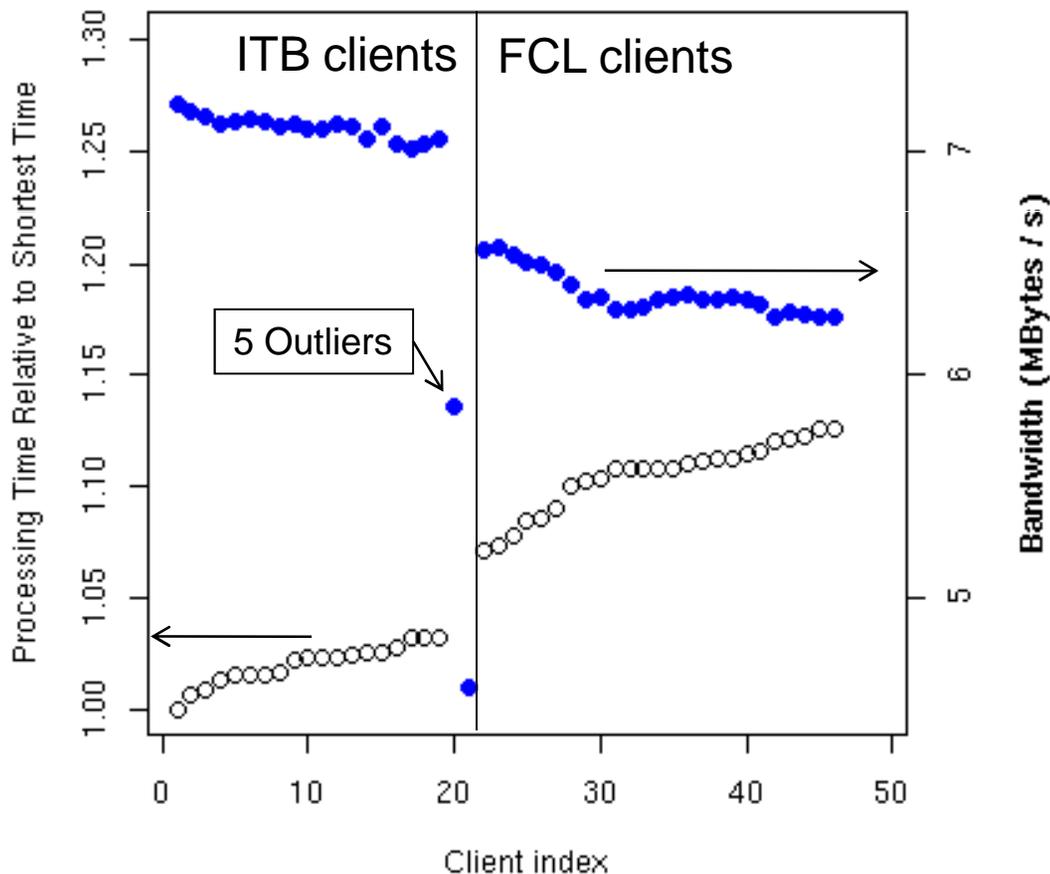
Ave Bandwidth with 21 ITB Nova Client for 1 to 3 Replicas
Hadoop Server on Bare Metal



49 Nova ITB / FCL clts vs. Hadoop

**49 clts (1 job / VM / core) saturate the bandwidth to the srv.
Is the distribution of the bandwidth fair?**

Relative Proc. Time and Bw w/ 49 nova clts vs. Bare Metal Hadoop



- Minimum processing time for 10 files (1.5 GB each) = 1117 s
- Client processing time ranges up to **233%** of min. time (113% w.o outliers)
- ITB clts (w/o 2 outliers):
 - Ave time = $102.0 \pm 0.2\%$
 - Ave bw = 7.11 ± 0.01 MB/s
- FCL clts (w/o 3 outliers):
 - Ave time = $110.5 \pm 0.3\%$
 - Ave bw = 6.37 ± 0.02 MB/s

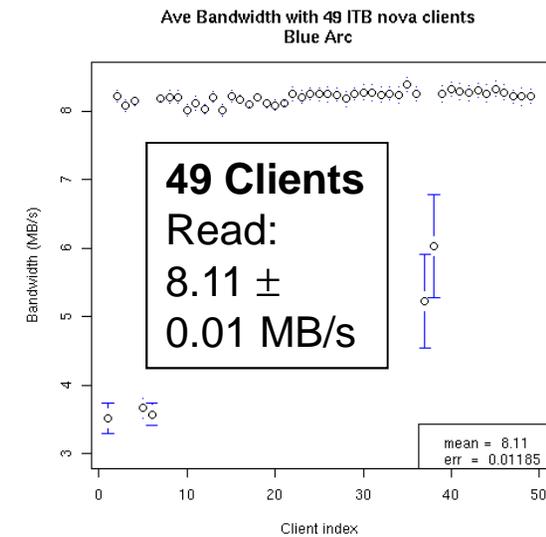
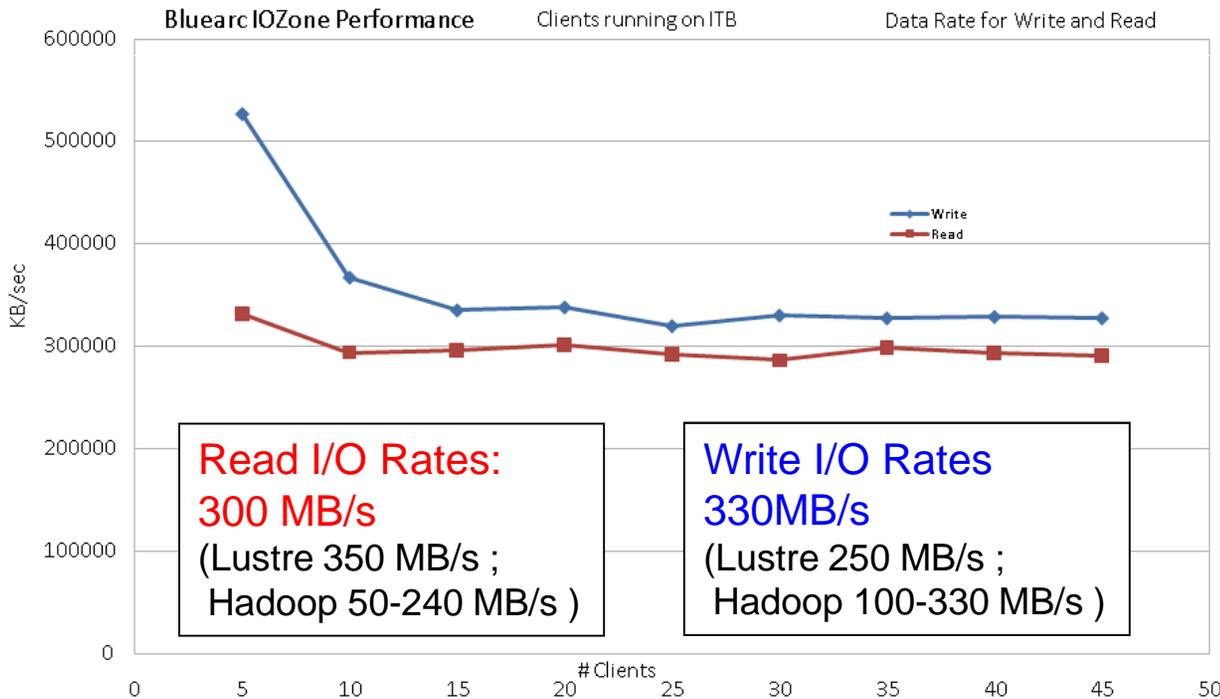
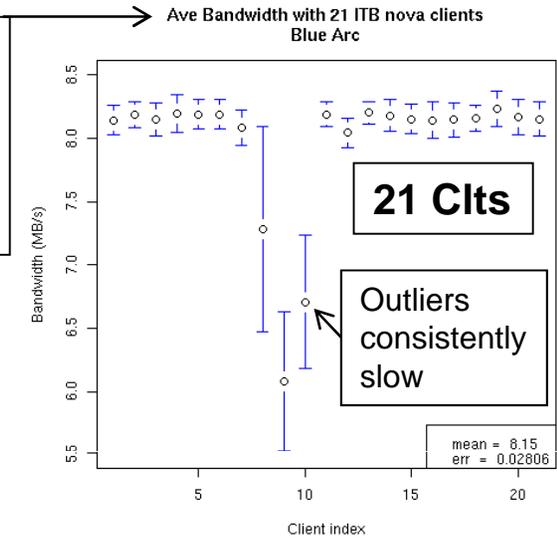
At saturation, ITB clnts read ~10% faster than FCL clnts (not observed in Lustre) (consistent w/ prev. slide).

ITB and FCL clnts get the same share of the bw among themselves (within ~2%).

Blue Arc Evaluation

- Clients access is fully POSIX.
- FS mounted as NFS.
- Testing with FC volume: fairly lightly used.

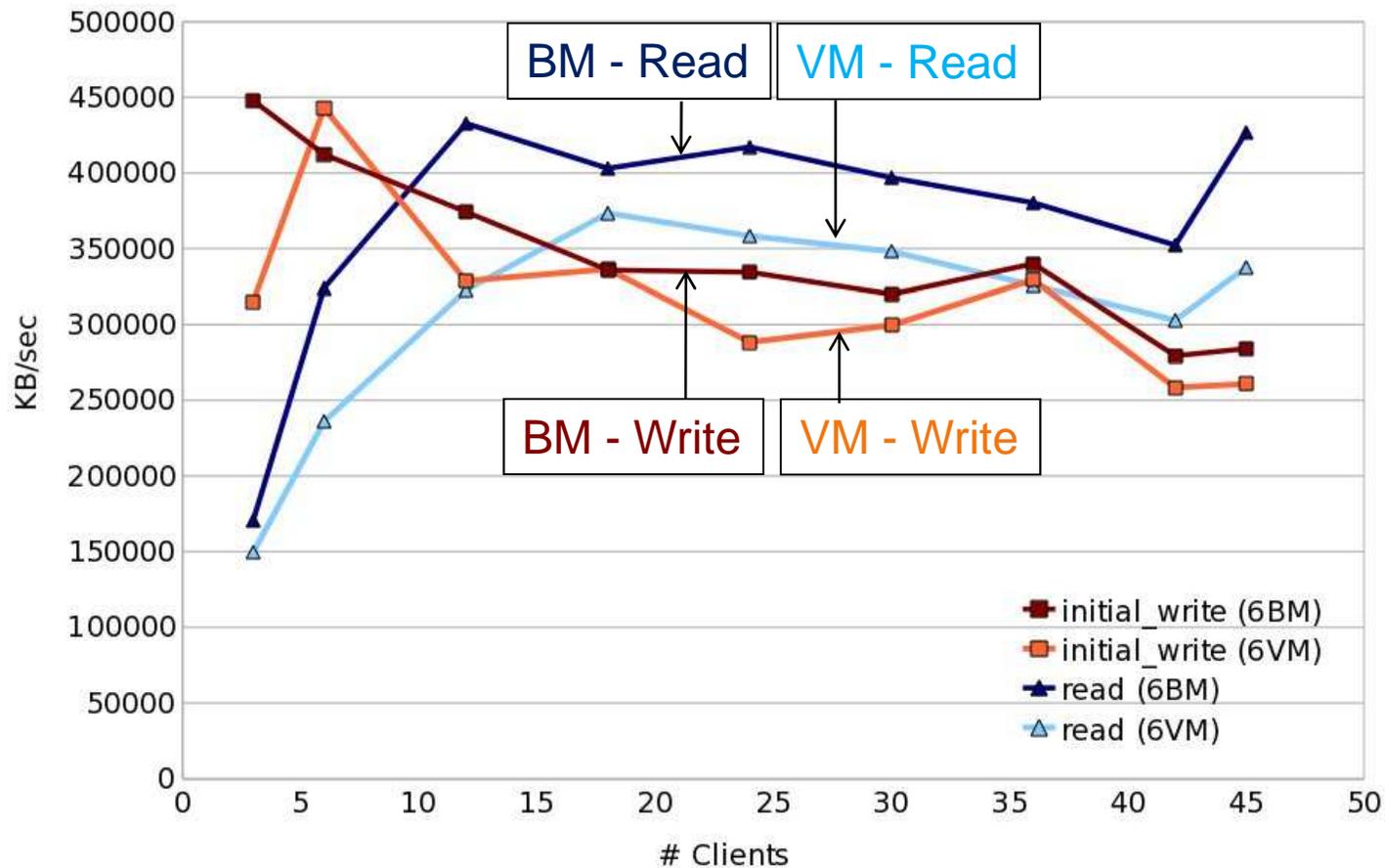
Root-app Read Rates:
 21 Clts: 8.15 ± 0.03 MB/s
 (Lustre: 12.55 ± 0.06 MB/s
 Hadoop: $\sim 7.9 \pm 0.1$ MB/s)



BA IOZone for /nova/data – BM vs. VM

How well do VM clients perform vs. BM clients?

IOZone Performance - BA area /nova/data - 6 VM vs. BM Machines



BM Reads are (~10%) faster than VM Reads.

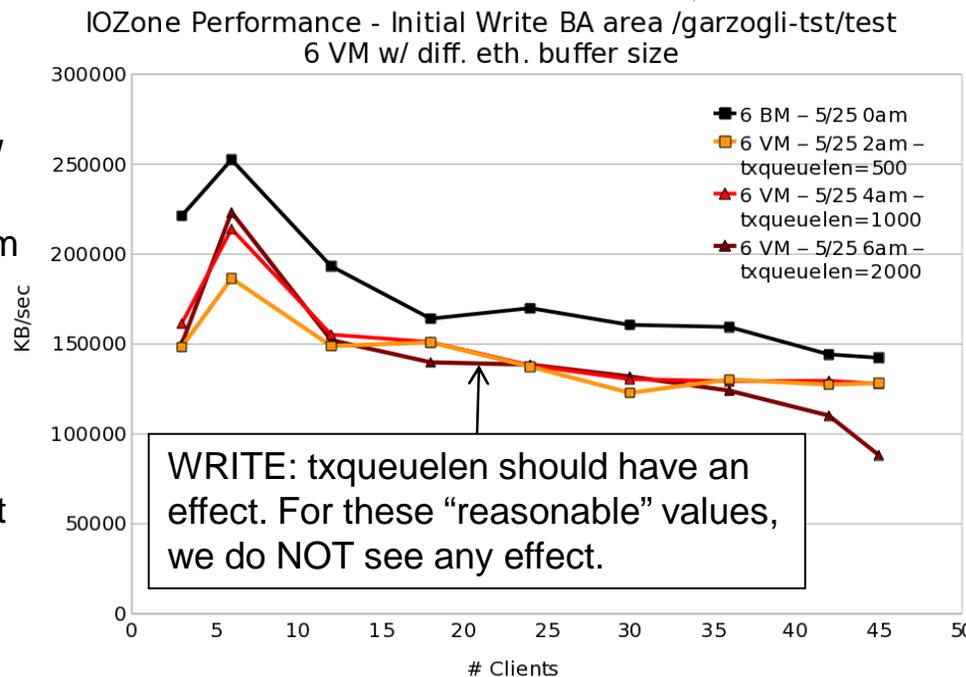
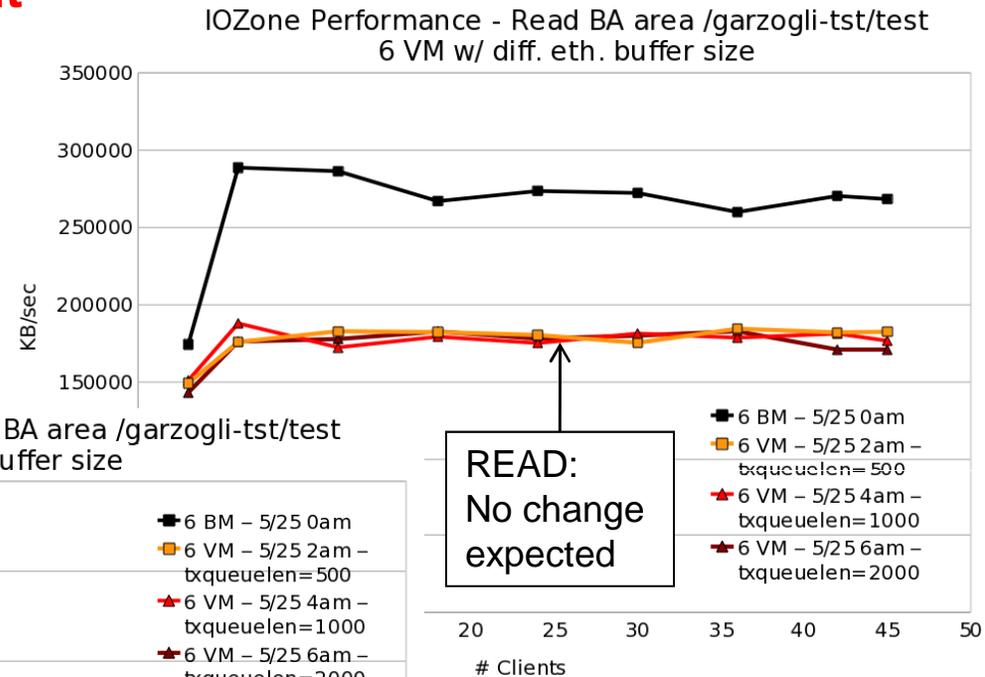
BM Writes are (~5%) faster than VM Writes.

Note: results vary depending on the overall system conditions (net, storage, etc.)

BA IOZone – Eth buffers length on clients

Do we need to optimize transmit eth buffer size (txqueuelen) for the client VMs (writes) ?

Eth interface	txqueuelen
Host	1000
Host / VM bridge	500, 1000, 2000
VM	1000



Notes:

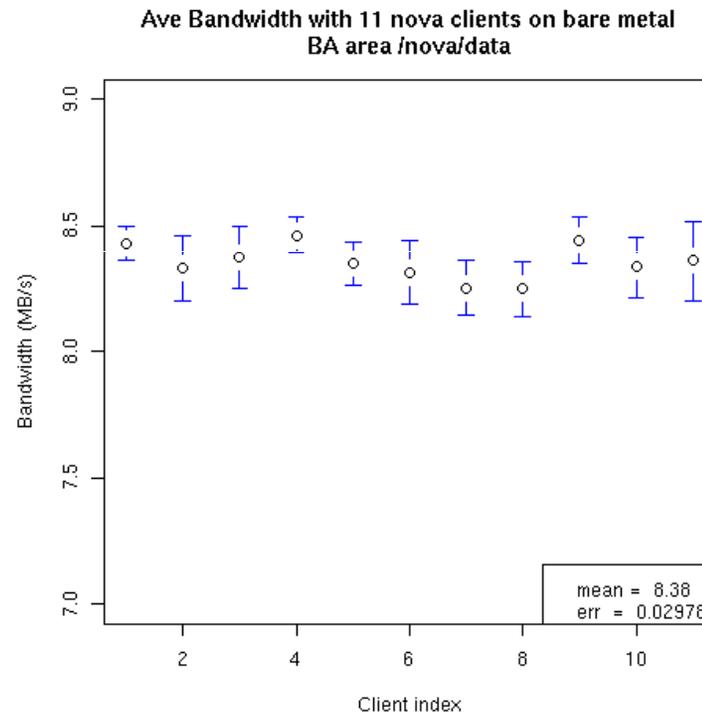
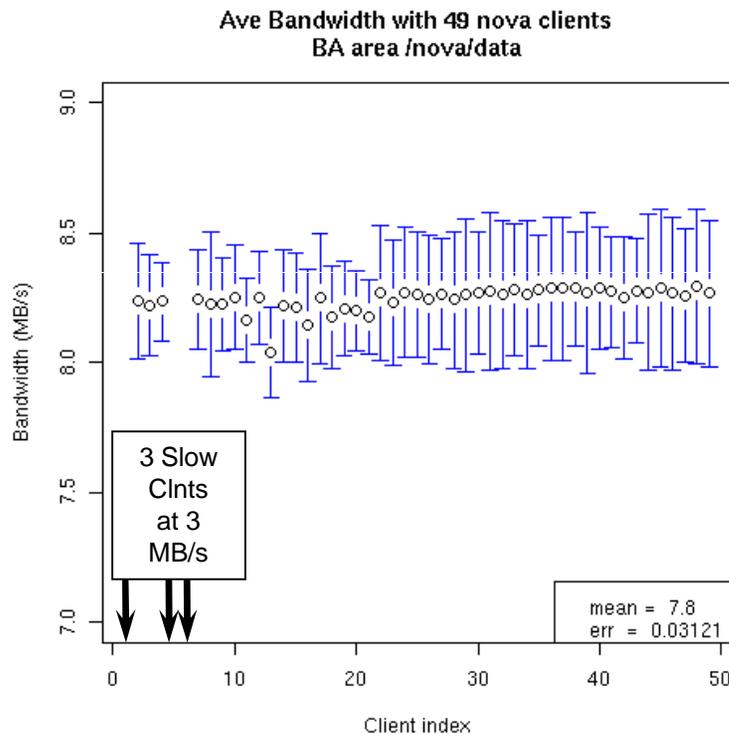
- Absolute BW varies with overall system conditions (net, storage, etc.)
- Meas. time: ~1h per line from midnight to 5 am

Varying the eth buffer size for the client VMs does not change read / write BW.

BA: VM vs. BM root-based clients

How well do VM clients perform vs. BM clients?

Read BW is the same on BM and VM.



Note: NOvA skimming app reads 50% of the events by design. On BA and Hadoop, clients transfer 50% of the file. On Lustre 85%, because the default read-ahead configuration is inadequate for this use case.

Results Summary

Storage	Benchmark	Read (MB/s)	Write (MB/s)	Notes
Lustre	IOZone	350	250 (70 on VM)	We'll attempt more tuning
	Root-based	12.6	-	
Hadoop	IOZone	50 - 240	100 - 330	Varies on num of replicas
	Root-based	7.9	-	
Blue Arc	IOZone	300	330	Varies on sys. conditions
	Root-based	8.4	-	
OrangeFS	IOZone	N/A	N/A	On Todo list
	Root-based	N/A	N/A	On Todo list

Conclusions

- Hadoop – IOZone: poor scalability for our configuration
 - Results on client VM vs BM are similar
 - Tuning eth transfer buffer size does NOT improve BW
 - Big variability depending on number of replicas
- Hadoop – Root-based benchmark
 - External clients are faster than on-board clients
 - At the rate of root-based application, number of replicas does not have an impact
 - Read 50% of the file when skimming 50% of the events (Lustre was 85% because of bad read-ahead config).
- Blue Arc – IOZone
 - BM I/O 5-10% faster than VM
 - Tuning eth transfer buffer size does NOT improve BW
- Blue Arc – Root-based benchmark:
 - At the rate of root-based app, BM & VM have same IO
 - Read 50% of the file when skimming 50% of the events.