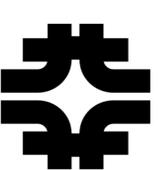




Investigation of Storage Options for Scientific Computing on Grid and Cloud Facilities



Gabriele Garzoglio (garzoglio@fnal.gov), Computing Sector, Fermilab, Batavia, IL

Introduction

- Set the scale:** measure storage metrics from running experiments to set the scale on expected bandwidth, typical file size, number of clients, etc.
 - http://home.fnal.gov/~garzoglio/storage/dzoro-sam-file-access.html
 - http://home.fnal.gov/~garzoglio/storage/cdf-sam-file-access-per-app-family.html
- Install storage solutions on FermiCloud testbed:** Lustre, BlueArc, Hadoop, OrangeFS
- Measure performance**
 - Run standard benchmarks on storage installations.
 - Study response of the technology to real-life (skim) applications access patterns (root-based)
 - Use HEPiX storage group infrastructure to characterize response to Intensity Frontier (IF) applications
- Fault tolerance:** simulate faults and study reactions
- Operations:** comment on potential operational issues. Clients on Virtual Machines: can we take advantage of the flexibility of cloud resources?

Data Access Tests

- IOZone** Writes 2GB file from each client and performs read/write tests.
 - Setup: 3-60clients on Bare Metal (BM) and 3-21 VM/nodes.
- Root-based applications**
 - Used off-line root-based framework (ana) of the Nova neutrino Intensity Frontier (IF) experiment. Ran a "skim job" that read a data file and discarded large fraction of events. Reads stressed storage access; writes proved CPU-bound
 - Setup: 3-60clients on Bare Metal and 3-21 VM/nodes.
- MDTest**
 - Different metadata FS operations on up to 50k files / dirs using different access patterns.
 - Setup: 21-504 clients on 21 VM.

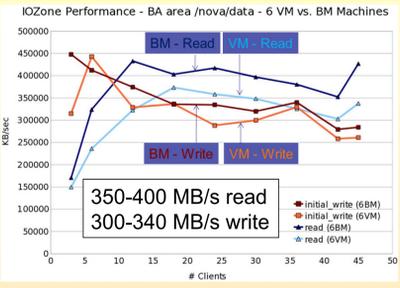
Conclusions

- Lustre on Bare Metal has the best performance as an external storage solution for the root skim application use case (fast read / little write). Consistent performance for general operations (tested via iozone)
 - Consider operational drawback of special kernel
 - On-board clients only via virtualization, but server VM allows only slow write.
- Hadoop, OrangeFS, and BlueArc have equivalent performance for the root skim use case.
- Hadoop has good operational properties (maintenance, fault tolerance) and a fast name server, but performance is not impressive.
- BlueArc at FNAL is a good alternative for general operations since it is a well known production quality solution.
- The results of the study support the growing deployment of Lustre at Fermilab, while maintaining the BlueArc infrastructure.

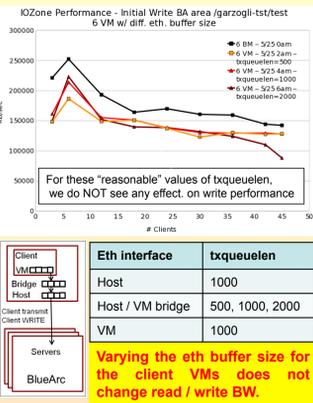
Storage	Benchmark	Read (MB/s)	Write (MB/s)	Notes
Lustre	IOZone	350	250 (70 on VM)	
	Root-based	12.6	-	
Hadoop	IOZone	50 - 240	80 - 300	Varies on replicas
	Root-based	7.9	-	
BlueArc	IOZone	300	330	Varies on conditions
	Root-based	8.4	-	
OrangeFS	IOZone	150-330	220-350	Varies on name nodes
	Root-based	8.1	-	

BlueArc

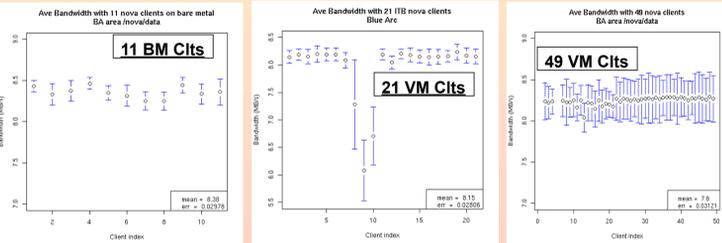
How well do VM clients perform vs. Bare Metal clients?



Do transmit ethernet buffer sizes (txqueuelen) affect write performance?



How well do VM clients perform vs. Bare Metal (BM) clients?



Root-app Read Rates:
21 Clts: 8.15 ± 0.03 MB/s (Lustre: 12.55 ± 0.06 MB/s, Hadoop: ~7.9 ± 0.1 MB/s)

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

IOZone

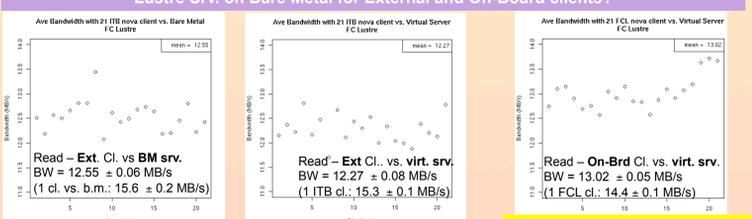
Root Benchmark

Lustre

How well does Lustre perform with servers on Bare Metal vs. VM?

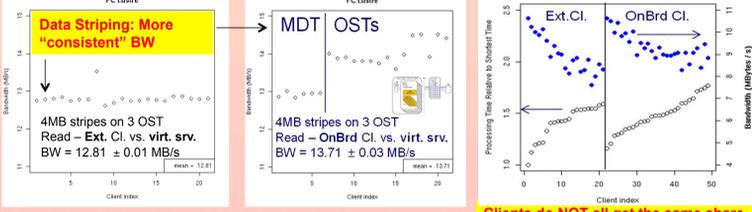


Read performance: how does Lustre Srv. on VM compare with Lustre Srv. on Bare Metal for External and On-Board clients?



Non-Striped Bare Metal (BM) Server: baseline for read (ext. cl.) Virtual Server is almost as fast as Bare Metal for read (ext. cl.)

Does Striping affect read BW for Ext. and On-Brd clients?

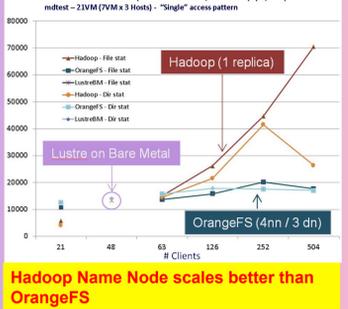


IOZone

Root Benchmark

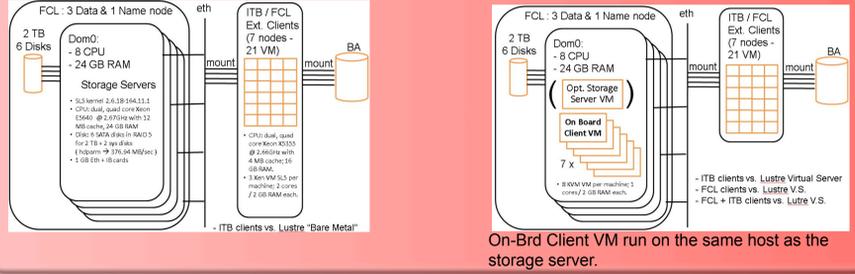
MetaData Comparison

How well do name nodes scale with number of clients?



Storage Testbed

"Bare Metal" Clients / Servers On-Board vs. External Clients



Hadoop

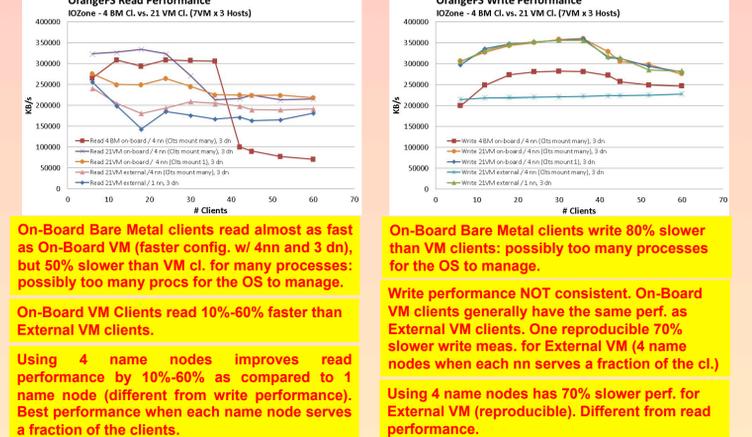
How well do VM clients perform vs. Bare Metal clients? Is there a difference for External vs. OnBoard clients? How does number of replica change performance?



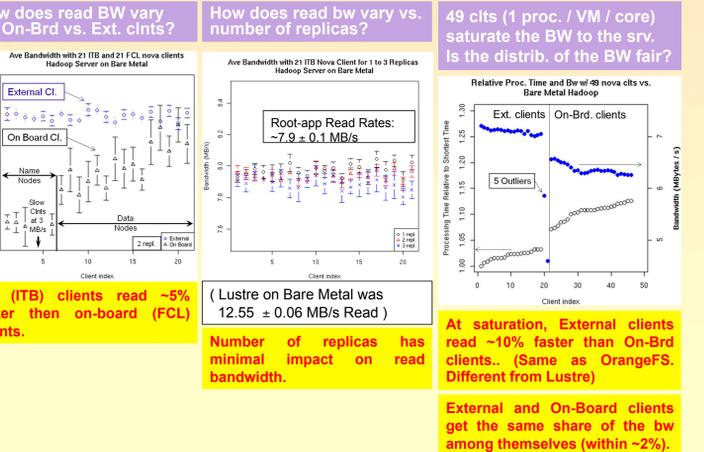
IOZone

OrangeFS

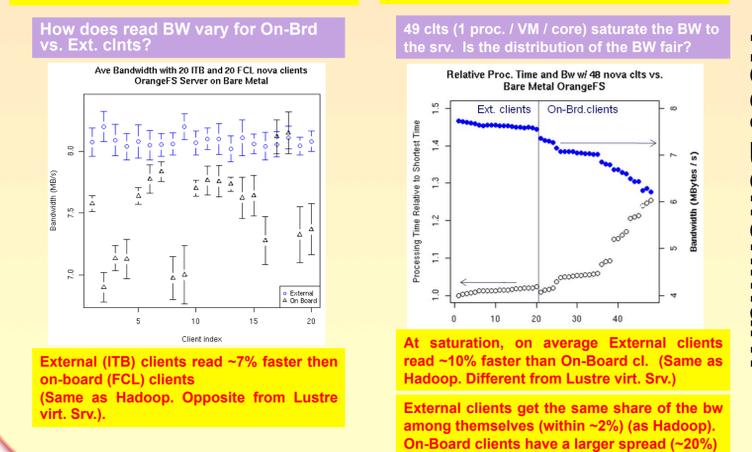
How well do VM clients perform vs. Bare Metal clients? Is there a difference for External vs. OnBoard clients? How does number of name nodes change performance?



IOZone



Root Benchmark



Root Benchmark