

# Virtualization as a Vehicle for Linear Storage Scaling

*A Work in Progress*

Rob Farber (PNNL)

Evan Felix (PNNL)

David M. L. Brown Jr. (PNNL)

# Motivation: Linear Storage Scaling

- 2005 concept to wide-stripe Lustre
  - Wide-striping to achieve high streaming storage bandwidth
  - Avoid Lustre resource starvation (via separate virt. OS)
  - PNNL well positioned as MPP2 has high local disk storage bandwidth
    - Approximately 570 Fat nodes with 7 disks per node
    - Lot's of Lustre expertise and need (NWchem: a storage bw hog)
  - Motivation to examine other file-systems (GPFS, etc.)
    - Will storage bandwidth scale linearly as # devices increases?
    - Is this true for other file-systems GPFS,etc.?
- Keep storage and storage servers local
  - Avoid network traffic whenever possible
  - Active storage demonstrated serves don't need much CPU
  - Spoke with CFS (Peter) about preferential local Lustre allocation
- Action: test to verify a linear speed-up according to # disks

# Why Virtualization?

- Observation: Active Storage shows OSS/OST low CPU usage
- Compartmentalize File-system OS
  - **Consistency** (run a virtual file-system appliance everywhere)
    - Management, distribution, maintenance advantages
  - **Robustness**
    - Failover becomes an OS function (not vendor specific)
    - Migration of running OS possible (with appropriate network)
  - **Keep storage local**
    - Performance
    - **Allocate local wide-stripe temporary file-system on job start**
      - Scale wide-stripe bandwidth according to job size
      - Reduce Meta-data workload
    - **Minimize/eliminate contention** for shared FS servers
    - Teragrid?
  - **Monitoring** (profile guest OS, etc.)
  - **Power/Space Saving**
  - Others (device specification, QoS, etc.)

# References for Further Reading

- Potential foundation for Petascale computing
  - Farber, *Scientific Computing Columns*
    - June and July 2007
    - Shameless plug (give me article hits)
  - 2007 ACM paper: “*High-Performance Hypervisor Architectures: Virtualization in HPC Systems*”, Gavrilovska, et. al. at Georgia Tech
    - Nice paper Karsten!

# Delays

- Funding (as usual)
- IA64 Xen network driver had challenges
  - Didn't work on Itanium
  - Then had performance problems
- During this time: Evan Felix used half of MPP2 to demonstrate using wide-striping:
  - 80 GB/s sustained write
  - 136 GB/s sustained read
  - Good Job Evan!

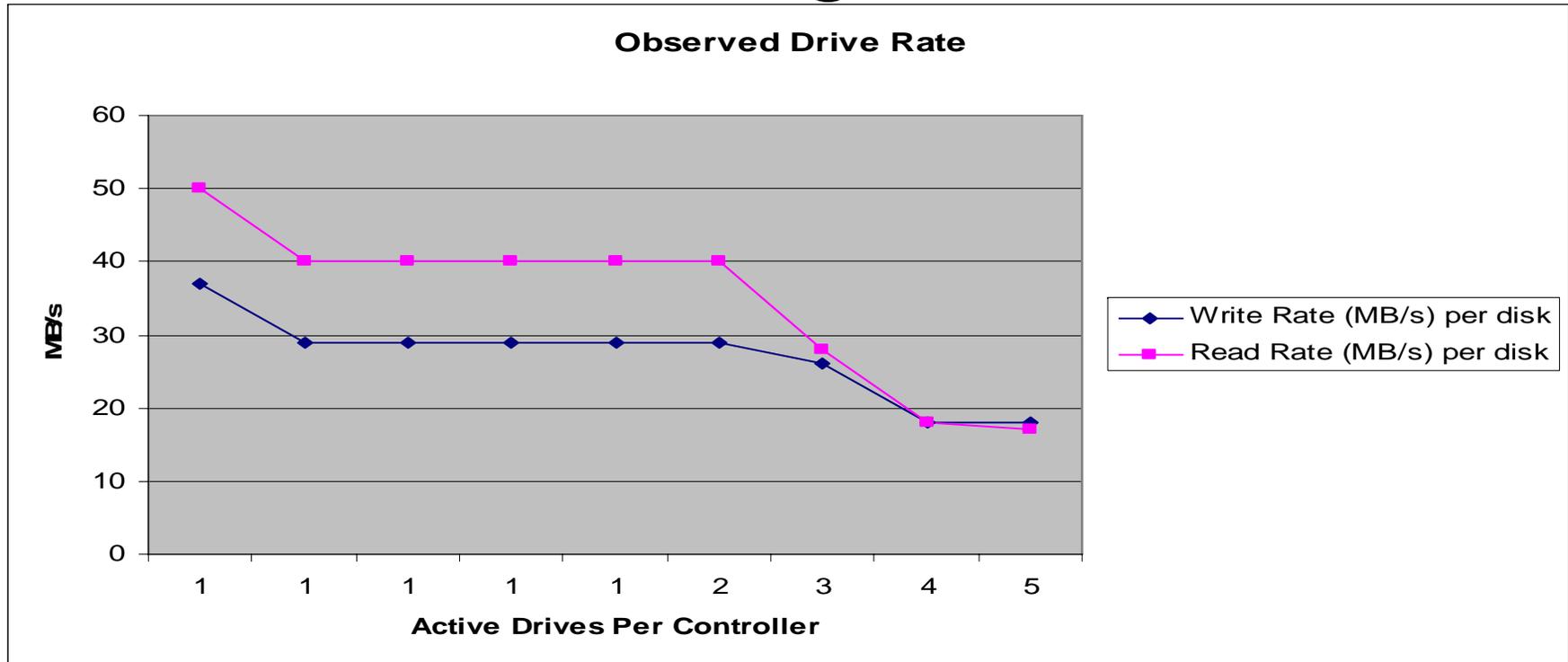
# 1st Test: Stability Problems

- It's hard to get all of MPP2
  - use scheduled downtime
- Had our first opportunity in Feb./Mar. 2007
  - Each drive was a separate domU OST
  - Each OSS was a separate domU instance
  - MDS was in a virtual machine
- Stability problems
  - Could only get a few OSS's connected
- No definite results on scaling behavior

# 2<sup>nd</sup> Test: 32-node Lustre Test: Linear Scaling

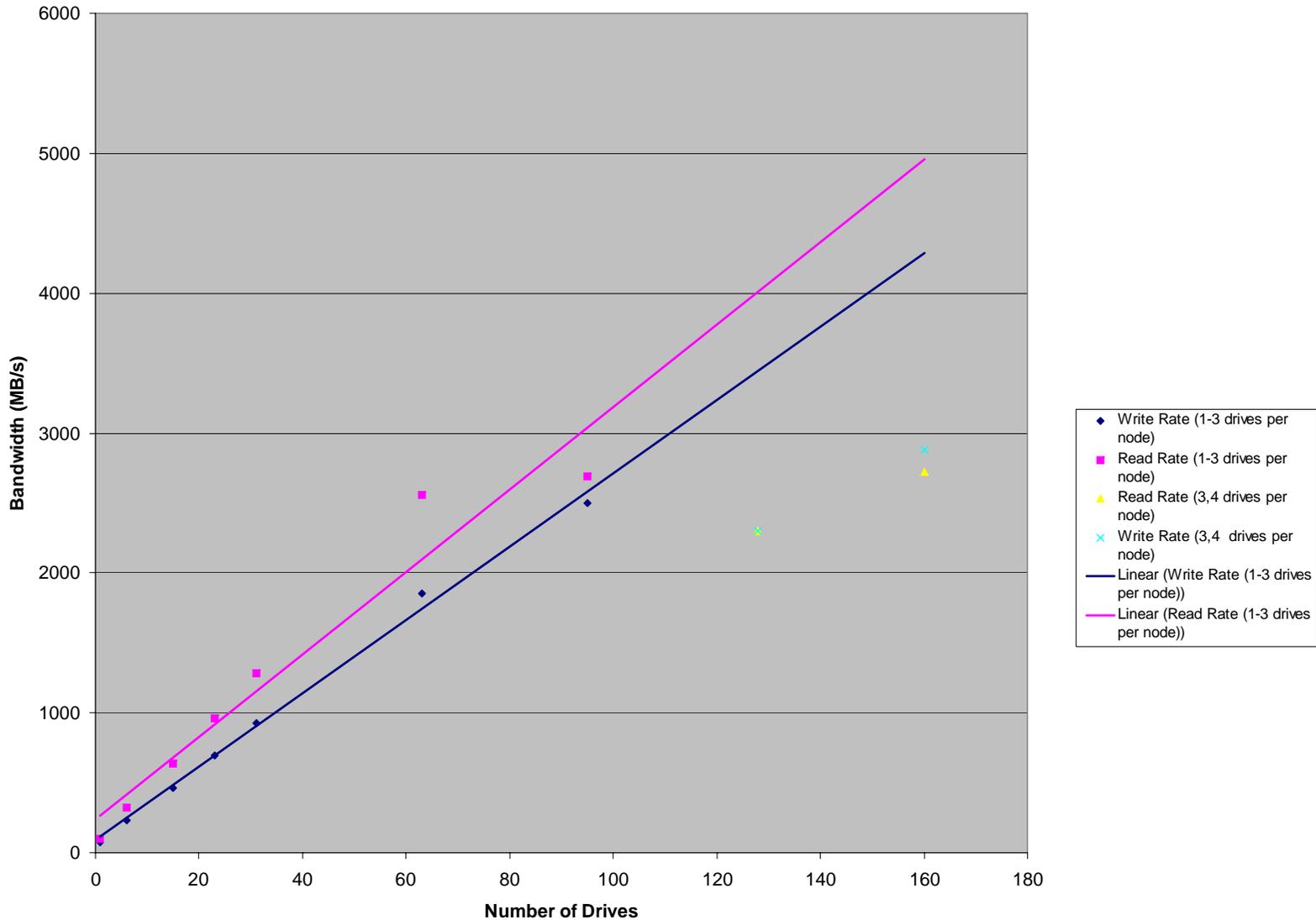
- DomU client
- Dom0 OSS with 7 OST's
  - Host OS manages the SCSI controller
  - Fewer virtualized systems (better stability?)
- Native MDS
- Use only local files
- If linear scaling will occur, this should be the setup.

# Linear Scaling Observed



- 1-3 drives per node -> linear scaling
- 4+ drives per node
  - 100% CPU usage observed
  - Performance degraded Our guess as to reason:
  - Hypothesis: CPU copy from host to client is limiting performance

# Scaling by Number of Drives



# Need to Eliminate Data Movement

- Need shared space between Host and Guest OS:
  - IB VMM-bypass paper demonstrated high performance
  - IOMMU and VT-d: Map host and guest pages together?

# Next Steps

- Need more machine time
- New Technology
  - Verify IOMMU or VT-D performance increase
- Test GPFS direct-connect scaling
- Other file-systems

# Thanks

- Supported by Petascale Data Storage Institute (PDSI)
  - Thanks Gary, Fred and others!
- EMSL (Environmental Molecular Scientific Laboratory) at PNNL
  - This research was performed in part using the Molecular Science Computing Facility (MSCF) in the William R. Wiley Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research and located at the Pacific Northwest National Laboratory, operated for the Department of Energy by Battelle.