



GlusterFS

Cluster File System

Z RESEARCH Inc.

Non-stop Storage

GlusterFS Cluster File System

GlusterFS is a Cluster File System that aggregates multiple storage bricks over InfiniBand RDMA into one large parallel network file system.



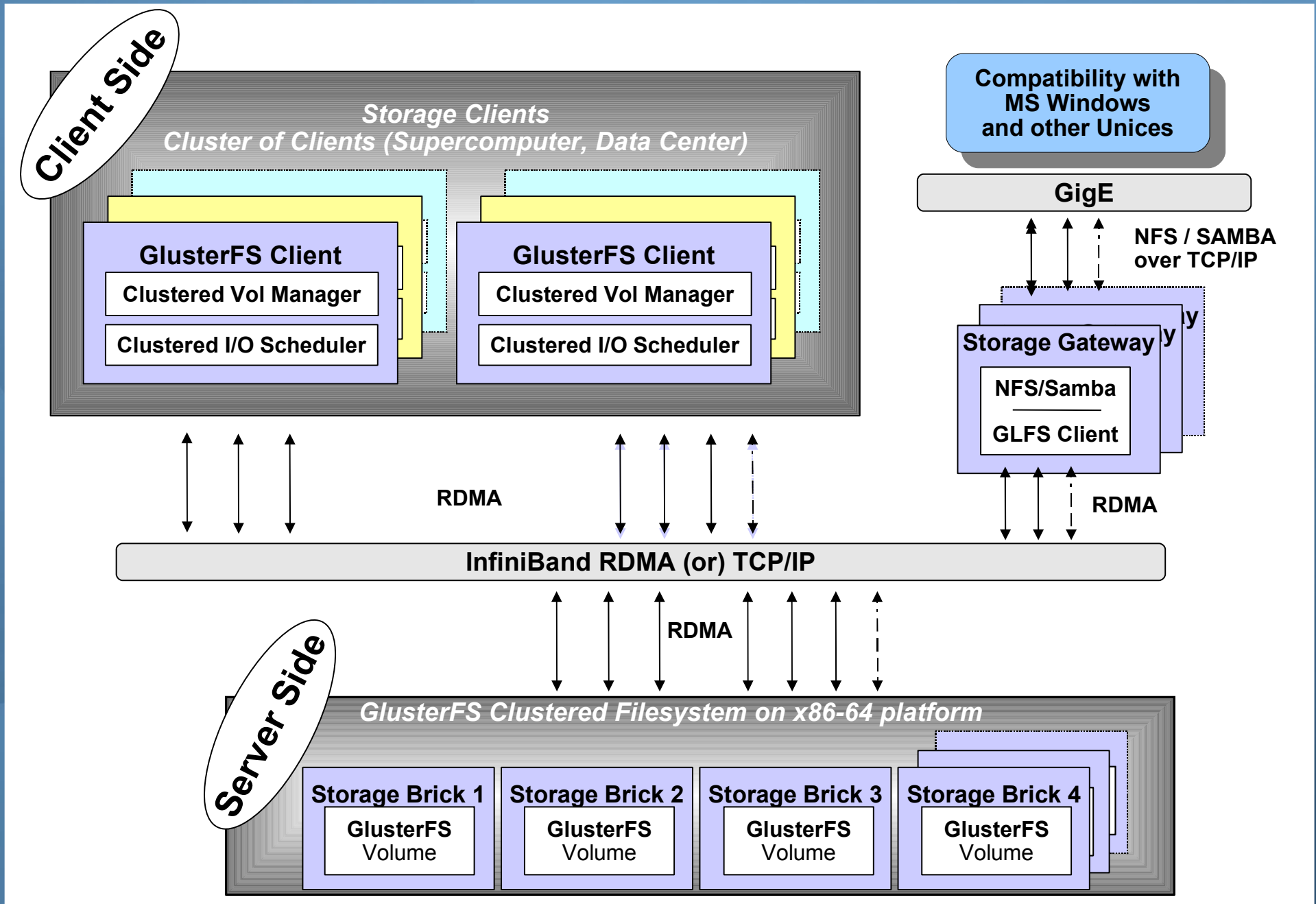
=

N x Performance & Capacity

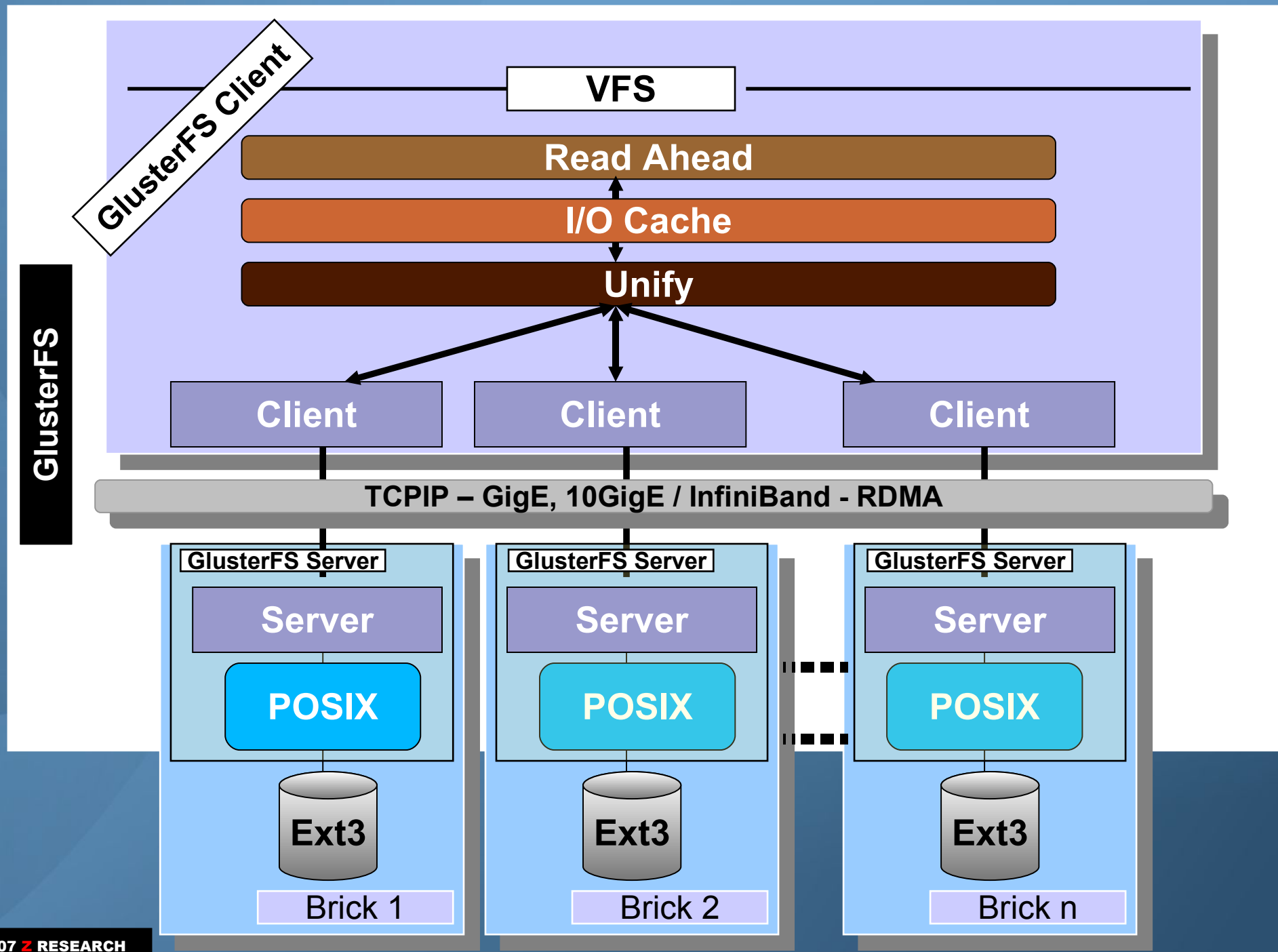
Key Design Considerations

- Capacity Scaling
 - Scalable beyond Peta Bytes
- I/O Throughput Scaling
 - Pluggable Clustered I/O Schedulers
 - Advantage of RDMA transport
- Reliability
 - Non Stop Storage
 - No Meta Data
- Ease of Manageability
 - Self Heal
 - NFS like Disk Layout
- Elegance in Design
 - Stackable Modules
 - Not tied to I/O Profiles or Hardware or OS

GlusterFS Design



Stackable Design



Volume Layout



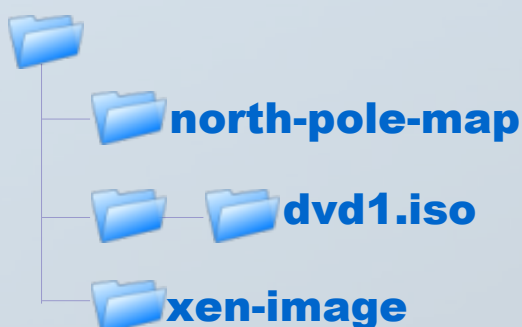
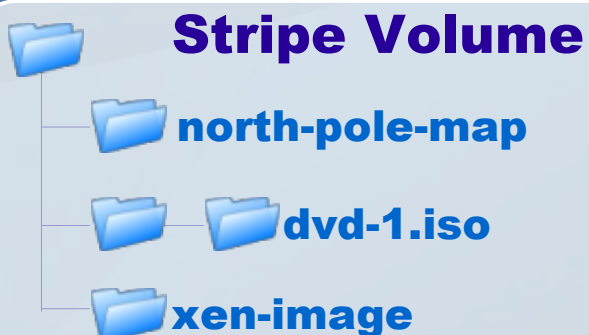
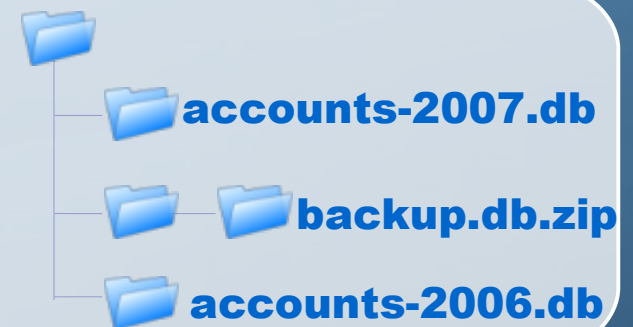
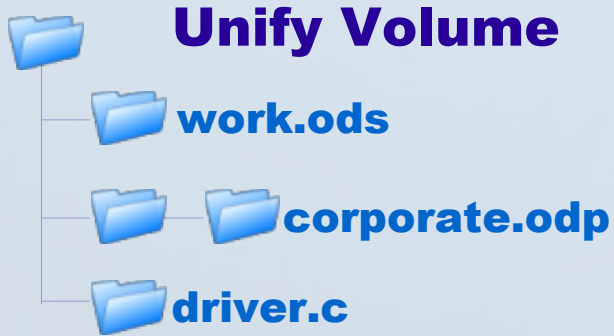
BRICK1



BRICK2



BRICK3



I/O Scheduling

- Adaptive least usage (ALU)
- NUFA
- Random
- Custom
- Round robin

volume bricks

type cluster/unify

subvolumes ss1c ss2c ss3c ss4c

option scheduler alu

option alu.limits.min-free-disk 60GB

option alu.limits.max-open-files 10000

option alu.order disk-usage:read-usage:write-usage:open-files-usage:disk-speed-usage

option alu.disk-usage.entry-threshold 2GB # Units in KB, MB and GB are allowed

option alu.disk-usage.exit-threshold 60MB # Units in KB, MB and GB are allowed

option alu.open-files-usage.entry-threshold 1024

option alu.open-files-usage.exit-threshold 32

option alu.stat-refresh.interval 10sec

end-volume

GlusterFS Benchmarks

Benchmark Environment

Method: Multiple 'dd' of varying blocks are read and written from multiple clients simultaneously.

GlusterFS Brick Configuration (16 bricks)

Processor - Dual Intel(R) Xeon(R) CPU 5160 @ 3.00GHz

RAM - 8GB FB-DIMM

Linux Kernel - 2.6.18-5+em64t+ofed111 (Debian)

Disk - SATA-II 500GB

HCA - Mellanox MHGS18-XT/S InfiniBand HCA

Client Configuration (64 clients)

RAM - 4GB DDR2 (533 Mhz)

Processor - Single Intel(R) Pentium(R) D CPU 3.40GHz

Linux Kernel - 2.6.18-5+em64t+ofed111 (Debian)

Disk - SATA-II 500GB

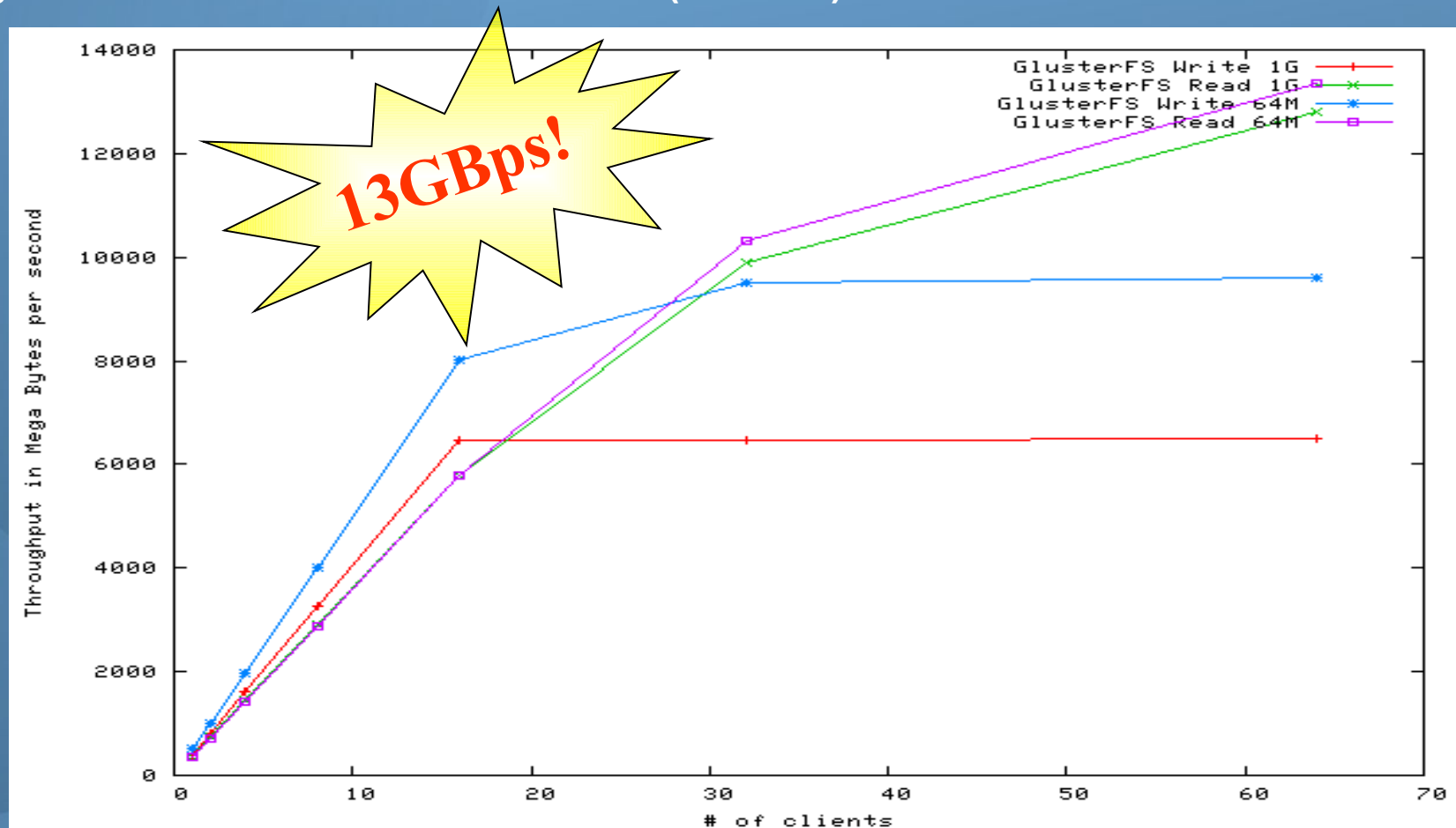
HCA - Mellanox MHGS18-XT/S InfiniBand HCA

Interconnect Switch: Voltaire port InfiniBand Switch (14U)

GlusterFS version 1.3.pre0-BENKI

Aggregated Bandwidth

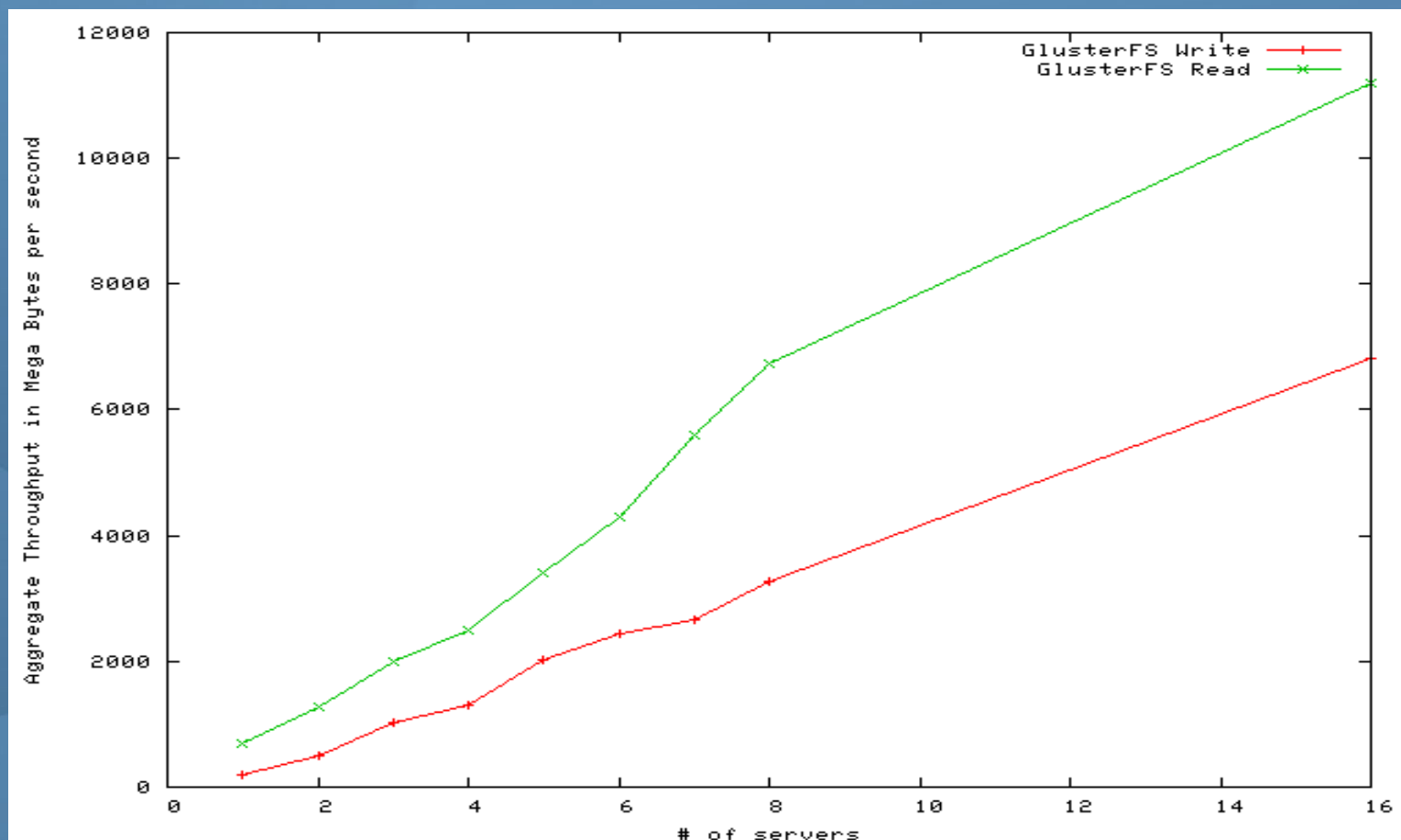
Aggregated I/O Benchmark on 16 bricks(servers) and 64 clients over IB Verbs transport.



- Peak aggregated read throughput -13GBps.
- After a particular threshold, write performance plateaus because of disk I/O bottleneck.
- System memory greater than the peak load will ensure best possible performance.
- ib-verbs transport driver is about 30% faster than ib-sdp transport driver.

Scalability

Performance improves when the number of bricks are increased



Throughput increases with corresponding increased in servers from 1 to 16

Hardware Platform Example

Storage Building Block

- Intel SE5000PSL (Star Lake) baseboard
 - ✓ Uses 2 (or 1) dual core LV Intel® Xeon® (Woodcrest) processors
 - ✓ Uses 1 Intel SRC SAS144e (Boiler Bay) RAID card
 - ✓ Two 2 ½” boot HDDs or boot from DOM
- 2U form factor, 12 hot-swap HDDs
 - 3 ½” SATA 3.0Gbps (7.2k or 10k RPM)
 - SAS (10k or 15k RPM)
- SES compliant enclosure management FW
- External SAS JBOD expansion

Intel EPSD
McKay Creek

Add-in Card Options

- Infiniband
- 10 GbE
- Fibre Channel



<http://www.gluster.org>

<http://www.zresearch.com>

Thank You!

Hitesh Chellani
Tel: 510-354-6801
hitesh@zresearch.com