Unleashing Clustered Computing

@

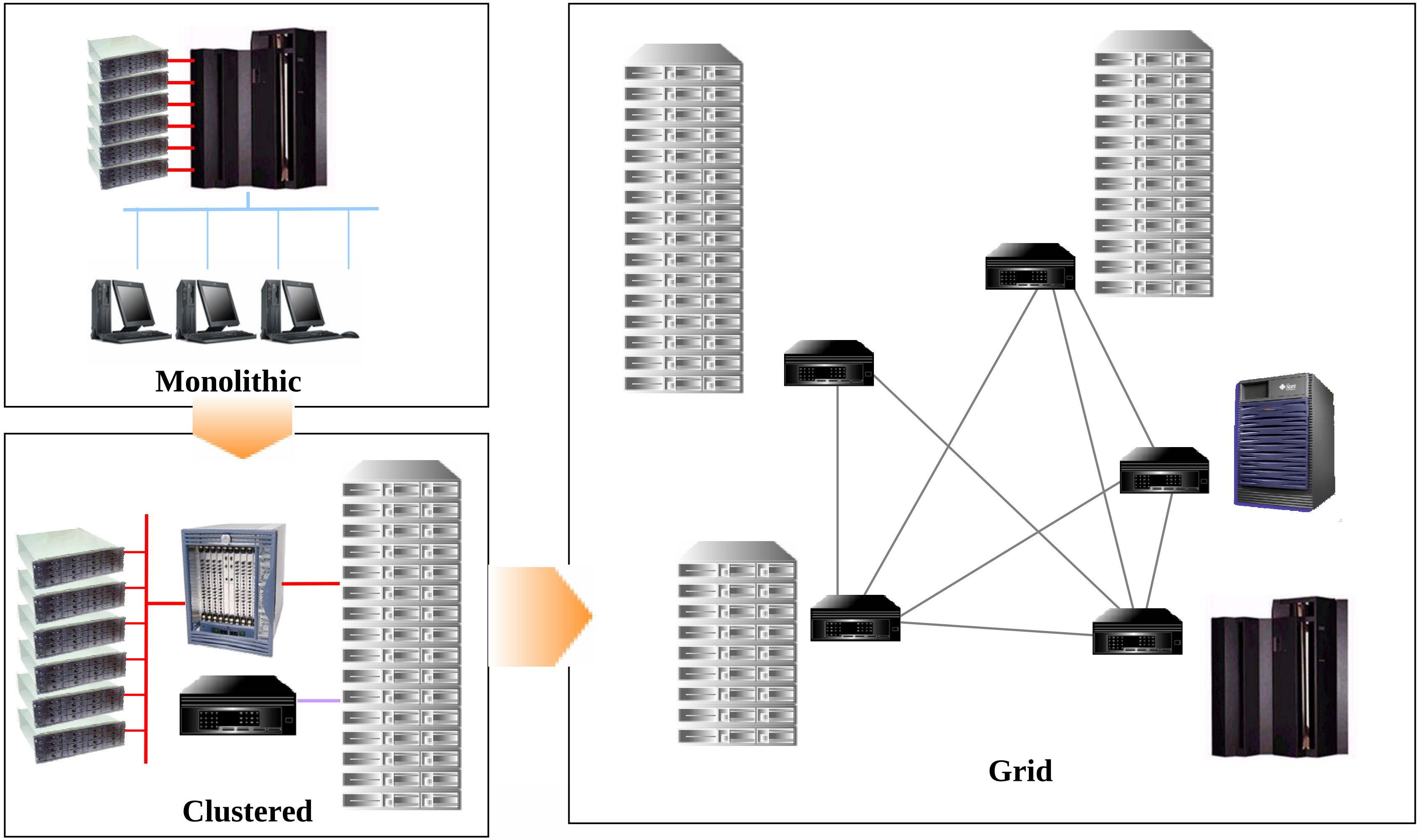
ECMWF-Workshop 2004

Gautham Sastri / Kolja Kuse

Company snapshot

- Founded in Nov/2002 by Gautham Sastri and Iain Findleton
  - Headquarters in Montreal, Canada
  - Offices in New York, Albuquerque, Munich & Reading (UK)
  - 20 employees worldwide
  - Exceptional team, with former employees of:
    - Sun, SGI, NEC, Cray, Sandia Nat’l Labs, etc.
- Well funded:
  - Entrepia Ventures (a division of the 5th largest investment bank in Japan)
  - Innovatech Montreal (a division of the Quebec pension fund)
- Has existing clients (Government, Oil & Gas, Health Sciences)
  - First customer ship in October 2003
- Several OEM relationships in place
- “Best Database Solution” award at LinuxWorld 2004

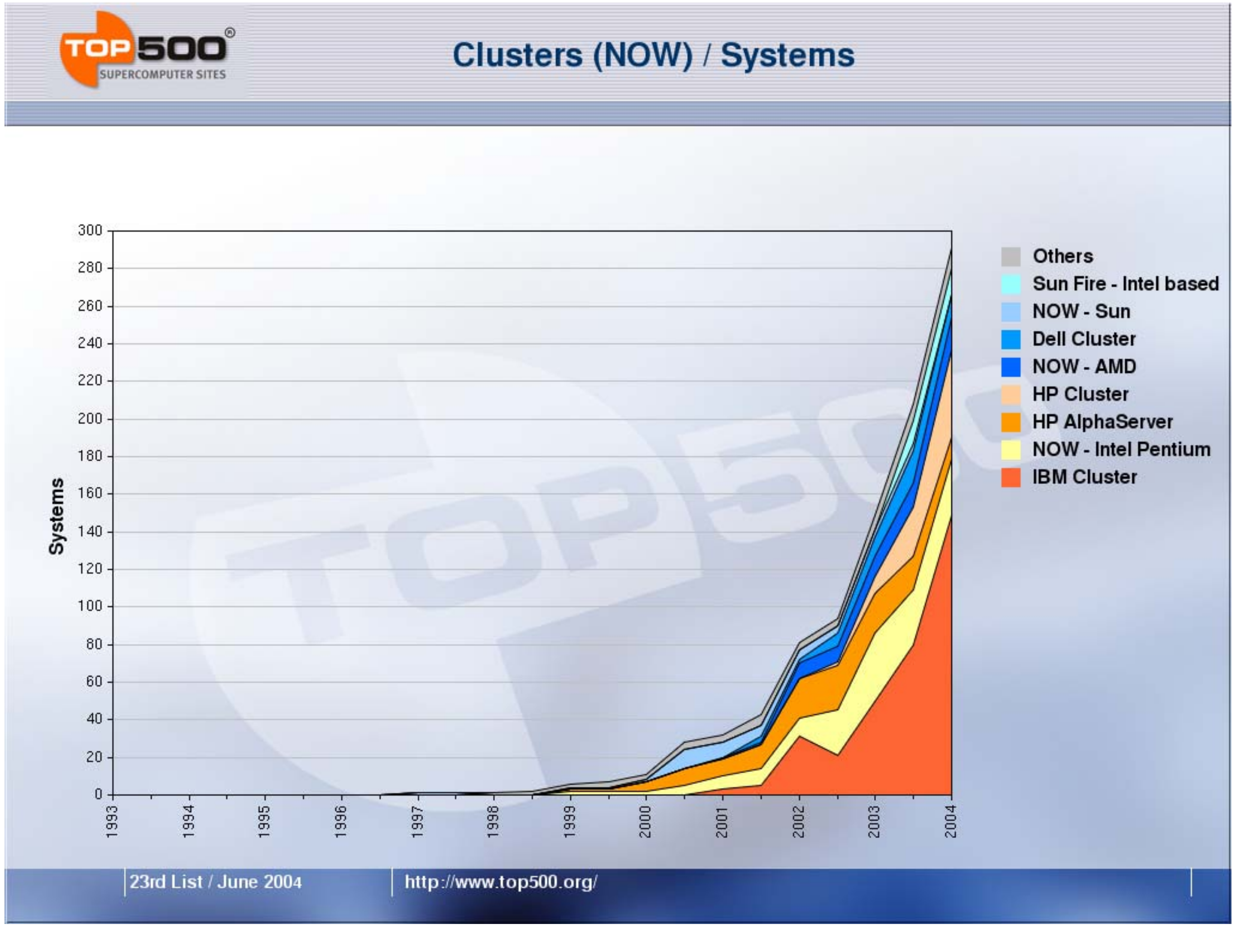
Trends in data processing architectures



Certain truths that are self-evident

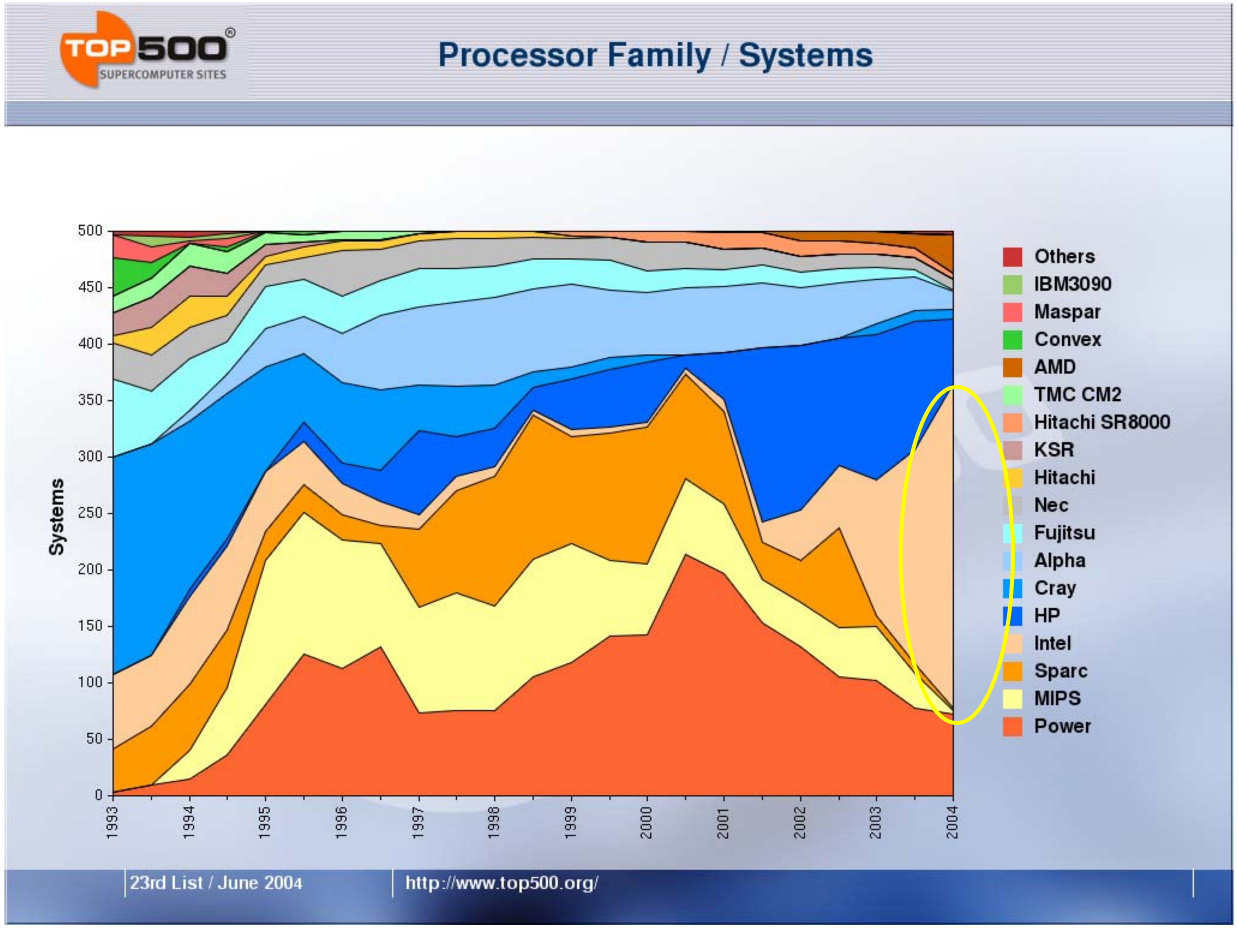
- Most computers are parallel computers
- CPUs are commodities ($450 for a 3.3 GHz CPU)
- Networks are commodities ($50 for a GigE port)
- Disk drives are commodities ($200 for a 250GB HDD)
- Scalable applications are *not commodities*
- *The value is in the integration (via software) of CPUs, networks and I/O to deliver scalable application bandwidth*

Trends on the bleeding edge (1)



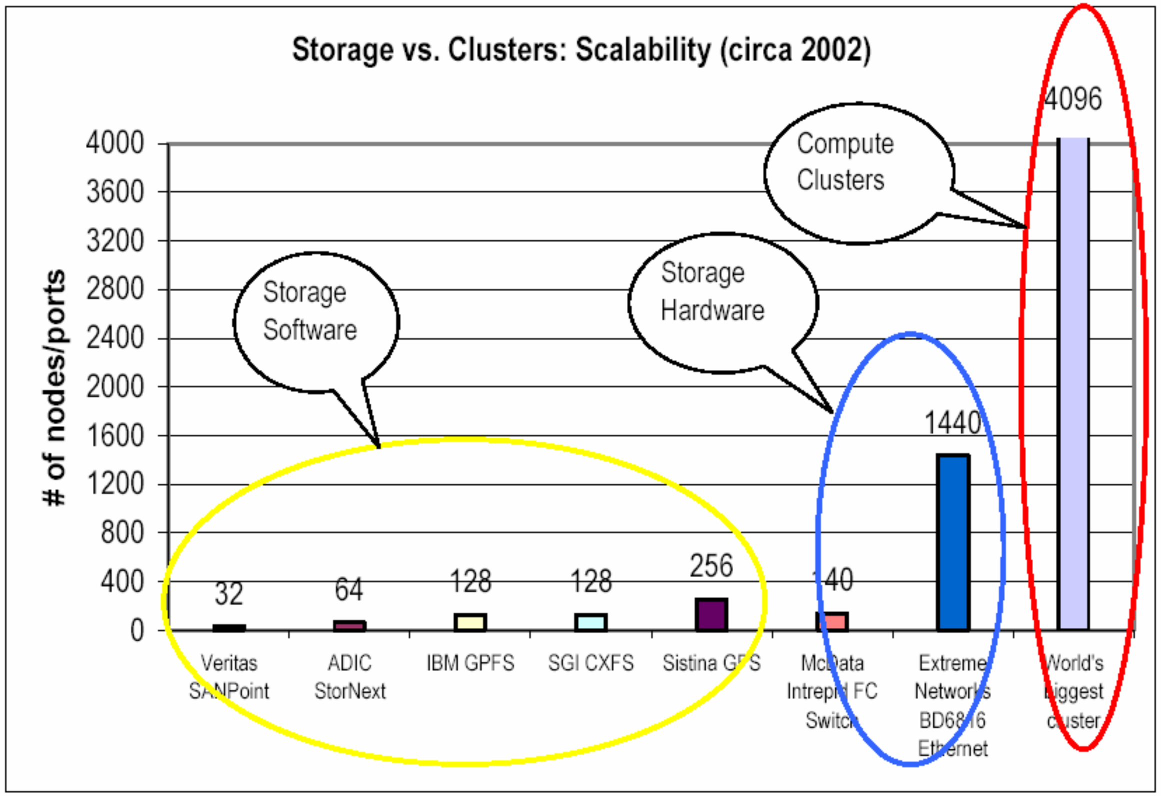
- Clusters are now running the biggest workloads that exist…

Trends on the bleeding edge (2)



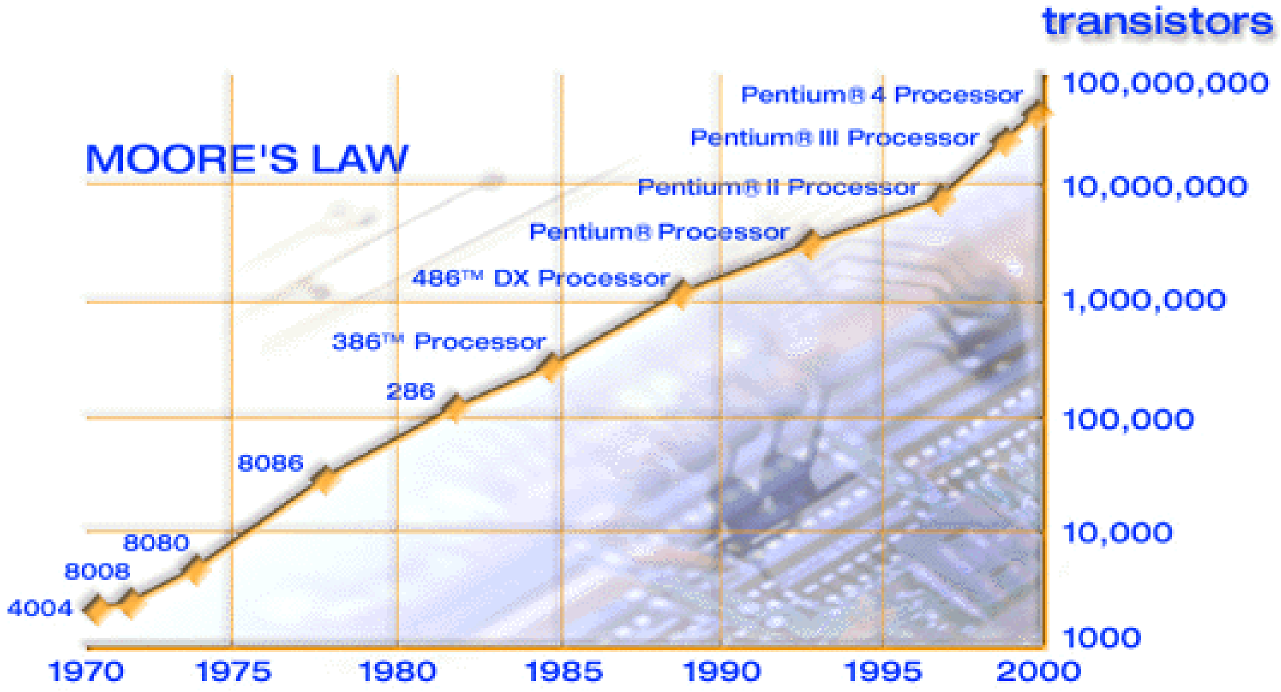
➢and cheap clusters are the ones gaining the most market share.

Some facts about I/O scaling



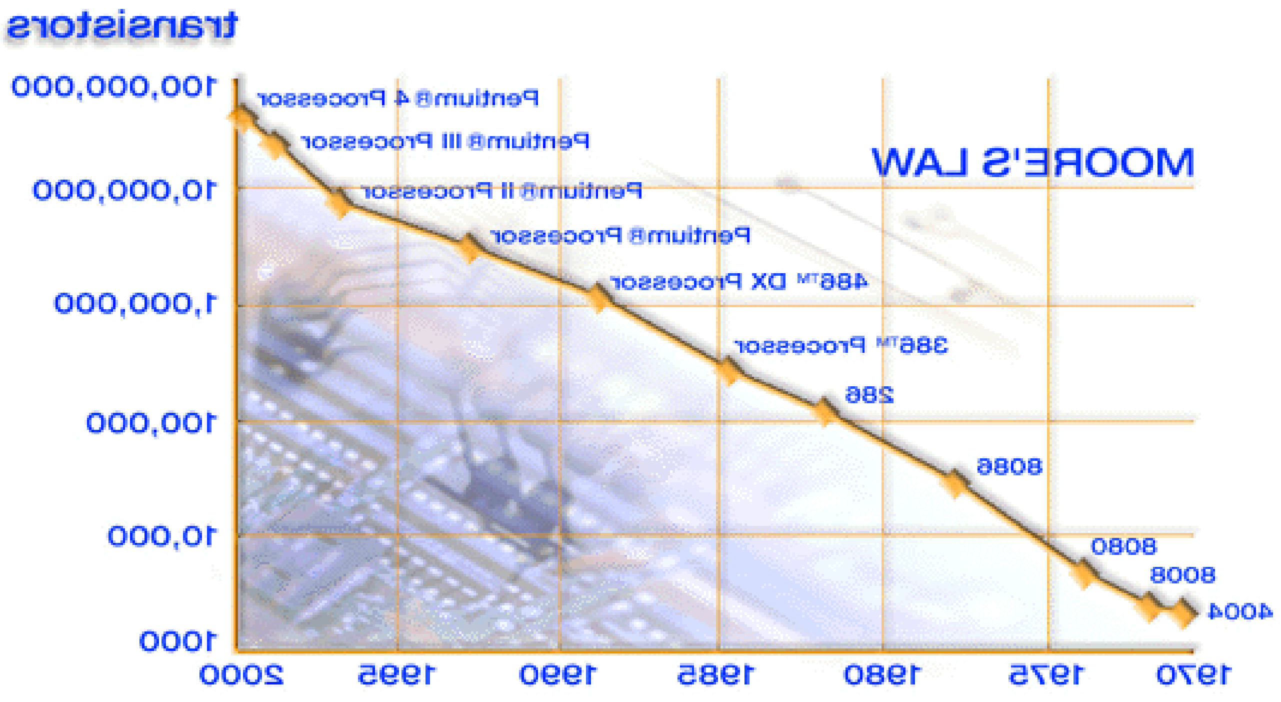
The true effect of Moore's Law (1)

Processors are getting faster and even faster…



The true effect of Moore’s Law (2)

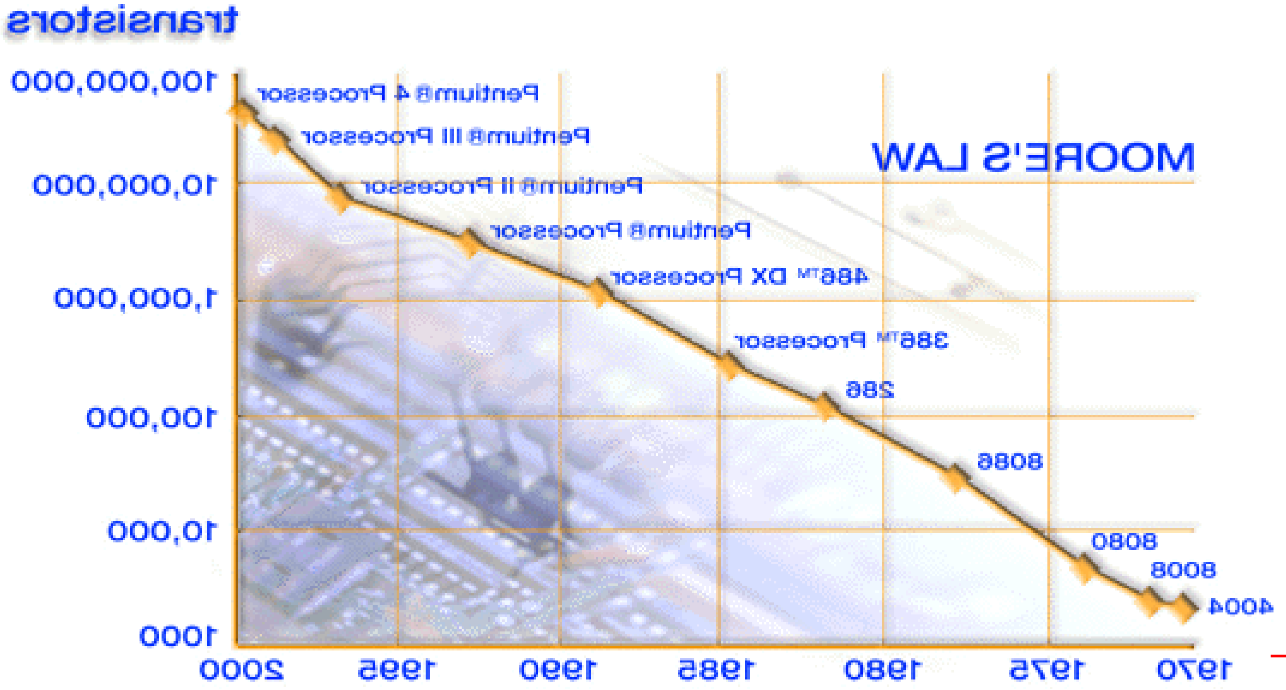
… while sustained application efficiency in relation to growing cluster size is plummeting . . .



The true effect of Moore’s Law (3)

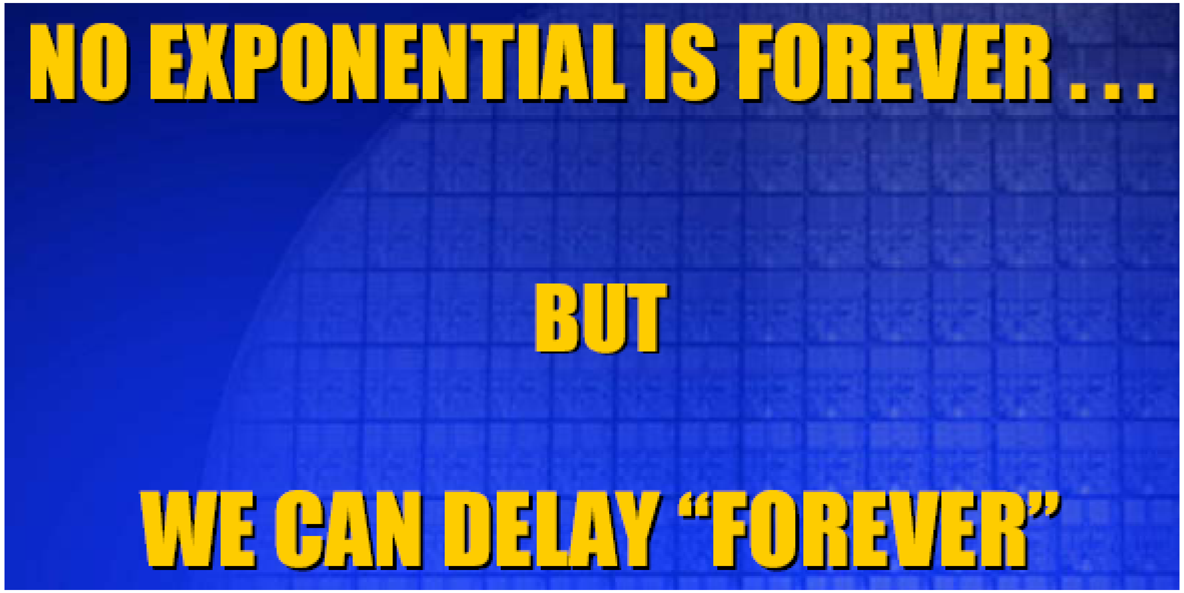
… this is also due to dramatic lack of I/O-capabilities.

Rel. Performance



Cluster size

Gordon Moore says:



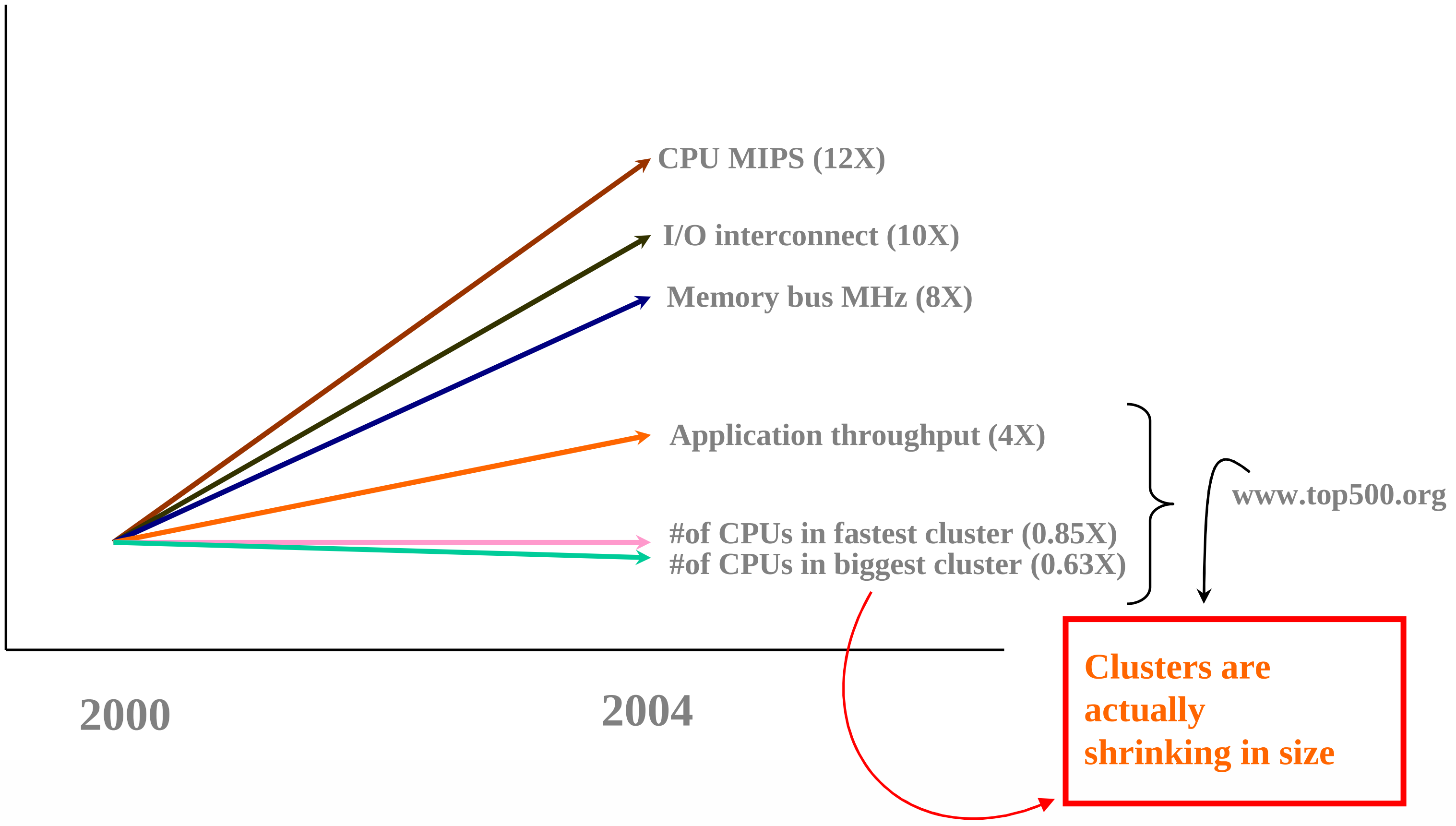
NO EXPONENTIAL IS FOREVER . . .

BUT

WE CAN DELAY "FOREVER"

“forever” cannot be delayed anymore…

Everything is getting faster . . . but applications are not scaling beyond a certain point any more . . .

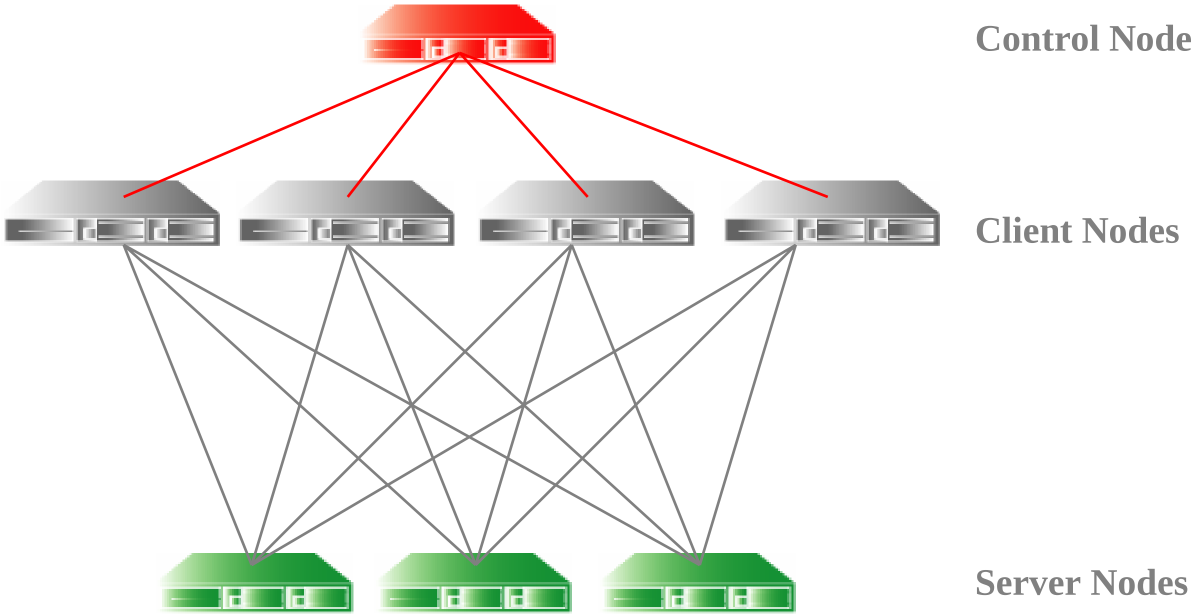


Clusters are actually shrinking in size

More Reality:

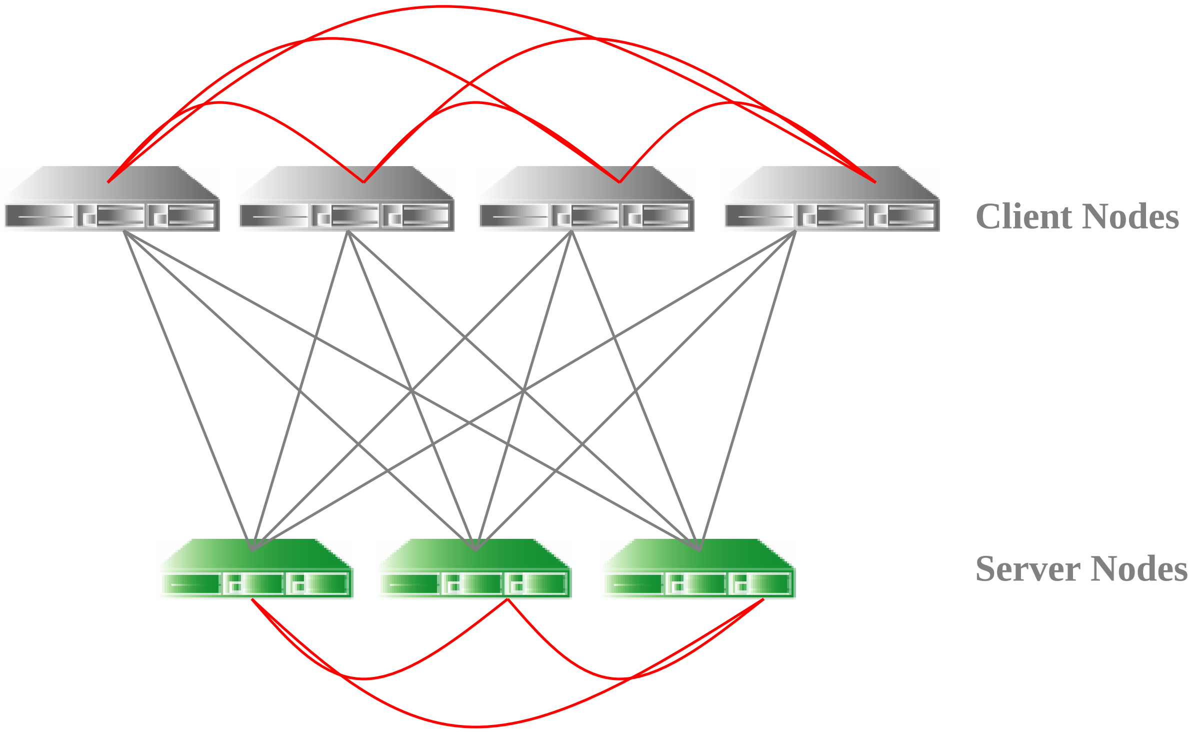
Today, “parallel” systems are not parallel in all respects

Examples of “parallel” Systems (1)



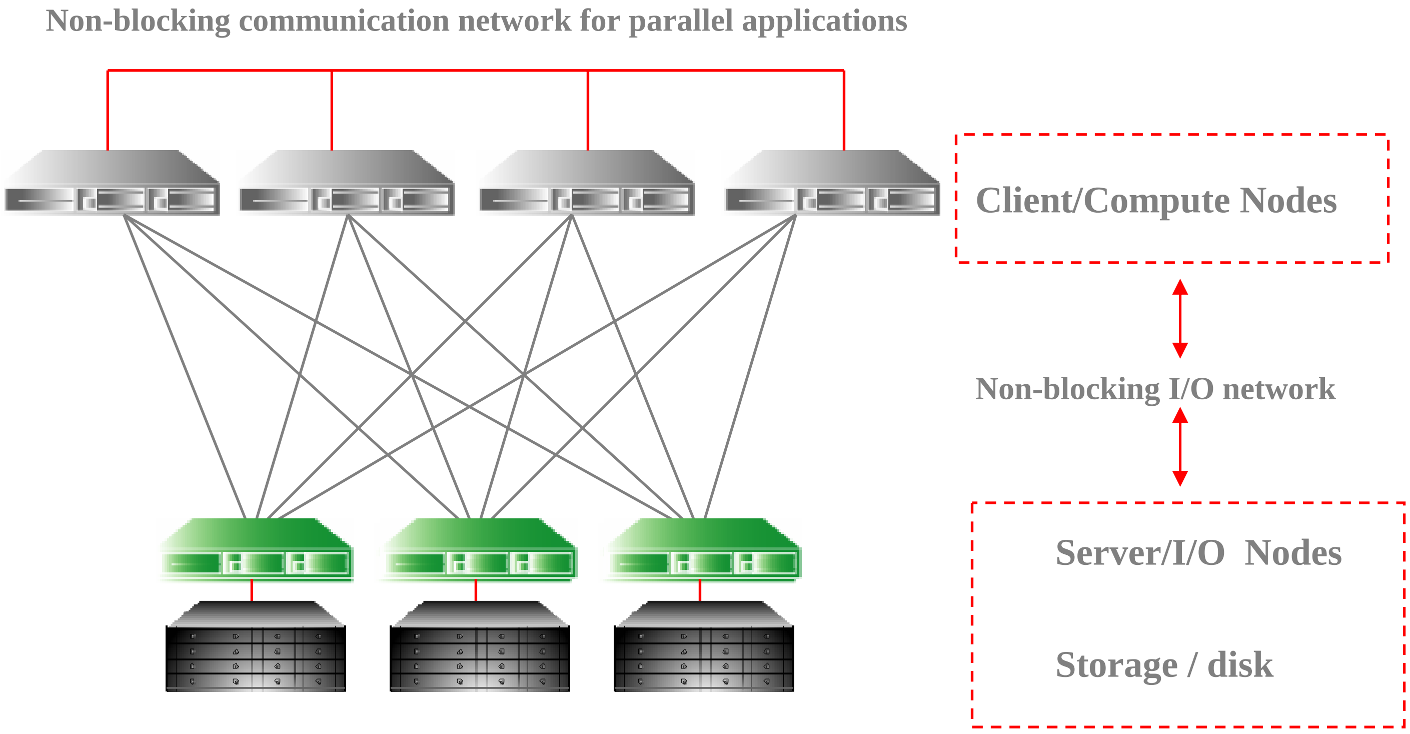
- ➢ *Centralized controller is the point of serialization that prevents scaling*
  - ▪Typically found in file system implementations such as Lustre, GFS, GPFS, etc.

Examples of “parallel” Systems (2)



- ***Inter-node communication between client nodes and/or between server nodes prevents scaling***
  - **Typically found in clustered database implementations**

An ideal parallel system looks like this:



- ➢ *No points of serialization, no N*N communications problems*
  - ➢Typically found in non-scalable SMP implementations

Terrascale's core technology: SASS

- Shared Access Scheduling System is a set of algorithms that provides cache coherence across thousands of application nodes regardless of geography. Key characteristics of SASS are:
  - Extremely low latency
  - "On demand" cache validation – eliminates unnecessary network traffic and/or broadcast storms
  - Extremely scalable
  - Enables clusters to behave like shared memory systems
- Sample applications
  - Massively parallel file systems
  - Massively parallel databases
  - Massively scalable RAID arrays
- *TerraGrid is only our first product based on SASS*

New Reality:

Why is SASS applicable to parallelize a file system?

Because networks have become very fast in bandwidth with latencies much lower, than we can find them in storage components

New Reality:

Why is SASS affordable?

Because these fast Networks have become extremely inexpensive, if you look at GigE and Infiniband for example . . . Leaving some margin for intelligent software . . .

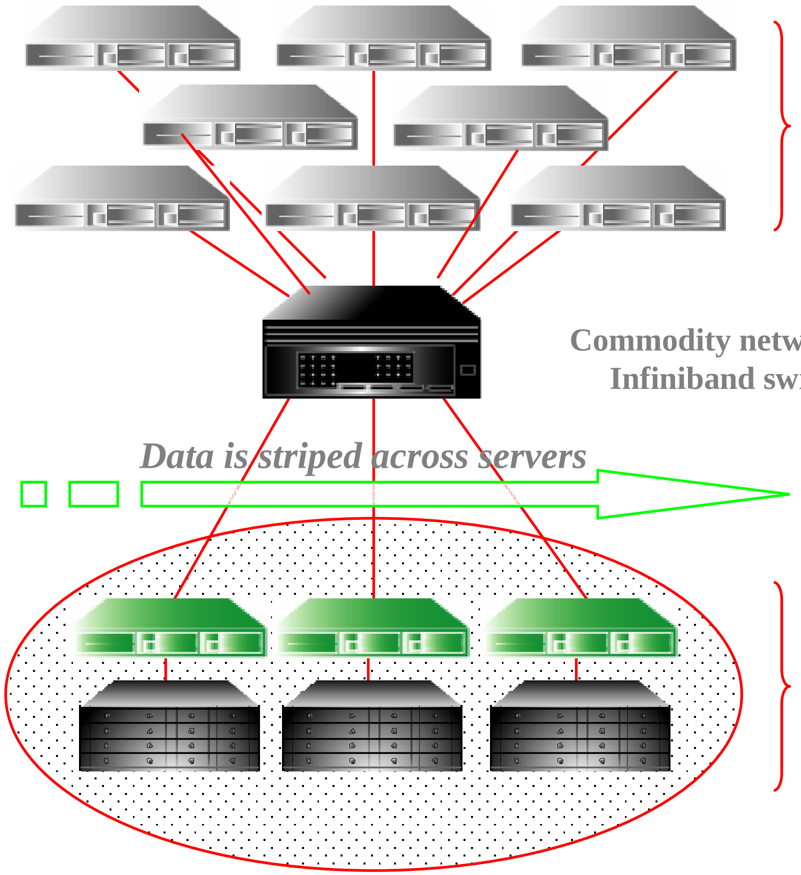
What is TerraGrid?

- TerraGrid is an *intelligent* software implementation of the iSCSI protocol stack that:
  - Is a block-level I/O platform that can scale linearly to 100s of Gbytes/sec of throughput and tens of Petabytes of capacity
  - Provides cache-coherence within the fabric
  - Fully harnesses the power of Linux file systems and utilities

- TerraGrid enables:
  - **Open-source "standalone" Linux file systems to be deployed as massively scalable global parallel file systems**
  - The acceleration of database engines to unprecedented levels of price/performance
  - The replacement of non-scalable proprietary/expensive RAID controllers with scalable, highly available I/O fabrics
  - OEMs and VARs to "roll their own" clustered NAS solutions

- *TerraGrid is NOT a file system or a clustered NAS box – instead, TerraGrid enables existing file systems and NAS solutions to achieve enhanced scaling and functionality with a global view on all data within a given "network"*

How TerraGrid works:



Client Nodes (100s) – *TerraGrid maintains cache coherence across clients*

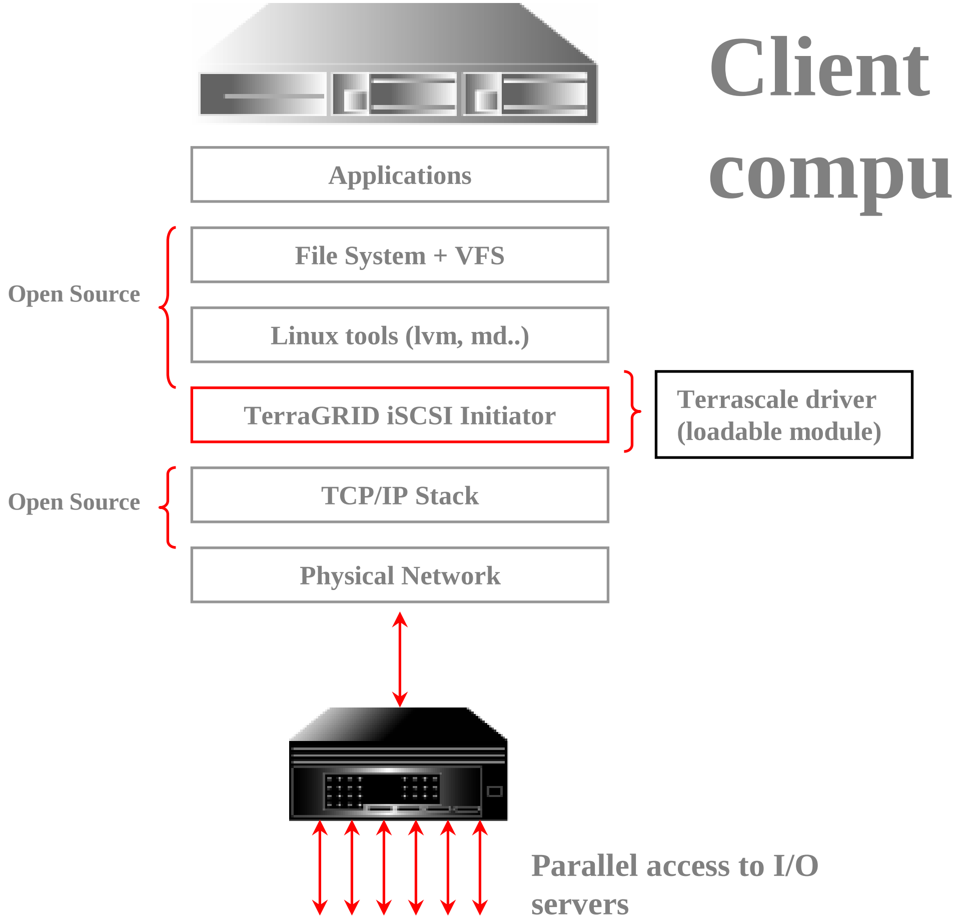
Commodity network, for ex. GigE or Infiniband switch (LAN/WAN)

Data is striped across servers

I/O Servers (10s) – *The aggregation of the I/O servers becomes a globally accessible virtual disk from the clients' perspective*

Unified, highly available global/parallel namespace

TerraGrid “Initiator”: Client Side



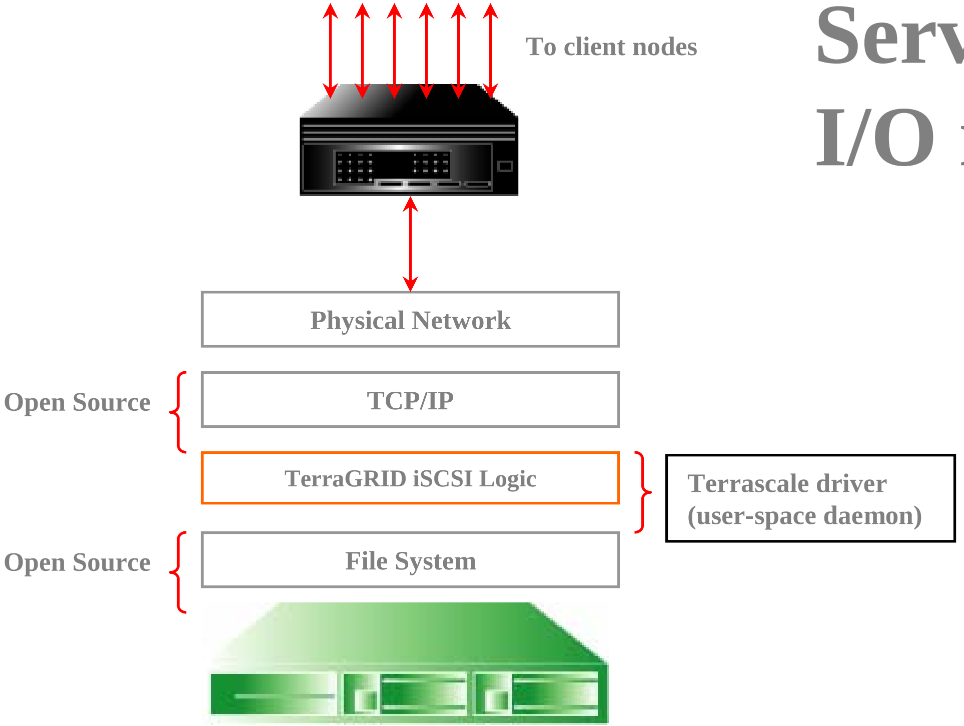
Client Side/ compute node

Parallel access to I/O servers

-Each client runs SW RAID0 to transmit all requests to multiple I/O servers
-High availability delivered transparently from the server side
-All I/O requests (block level, file level, file system level) are parallelized across I/O servers

TerraGrid “Target”: Server Side

Server Side/ I/O node

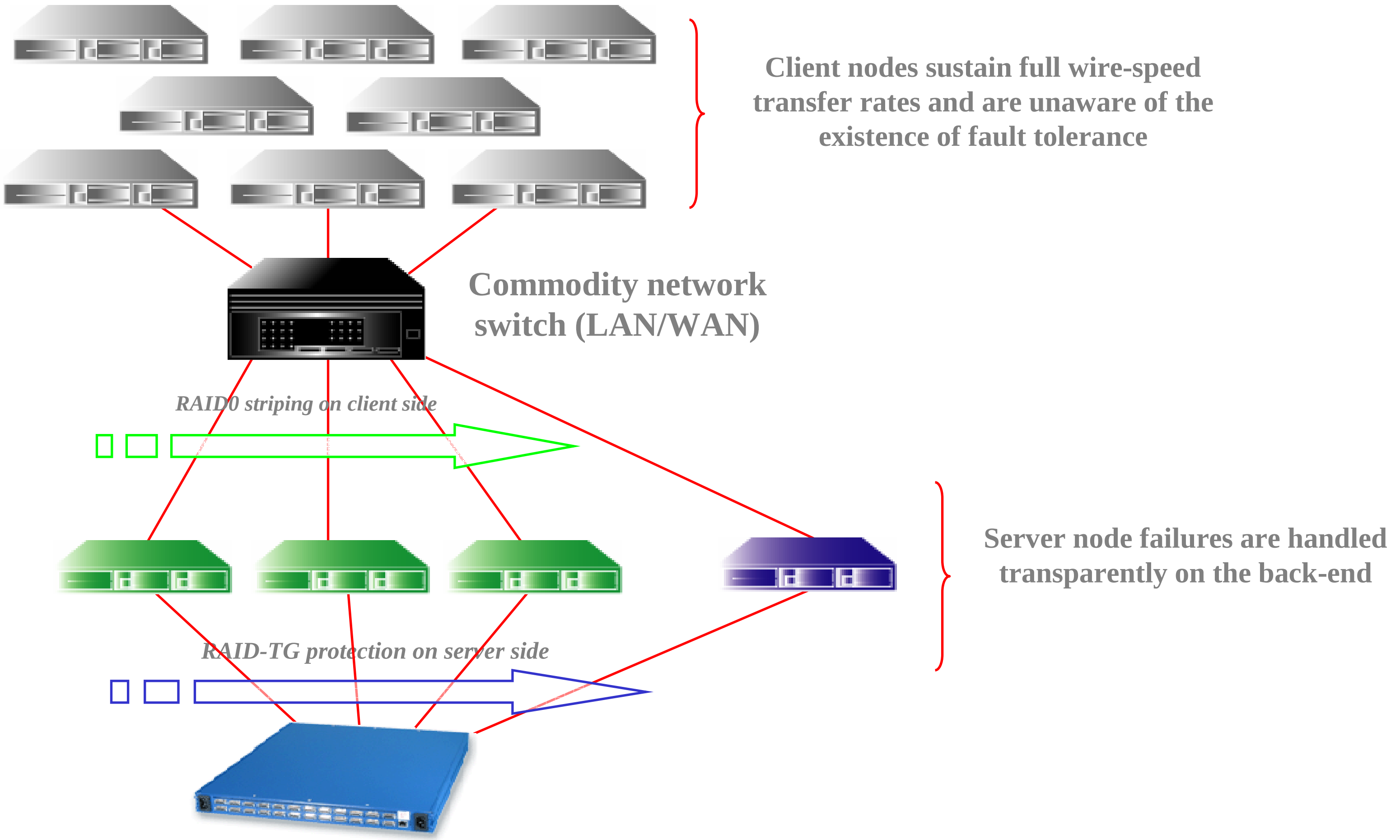


-Each server presents a file or set of files as block containers to the initiator pool

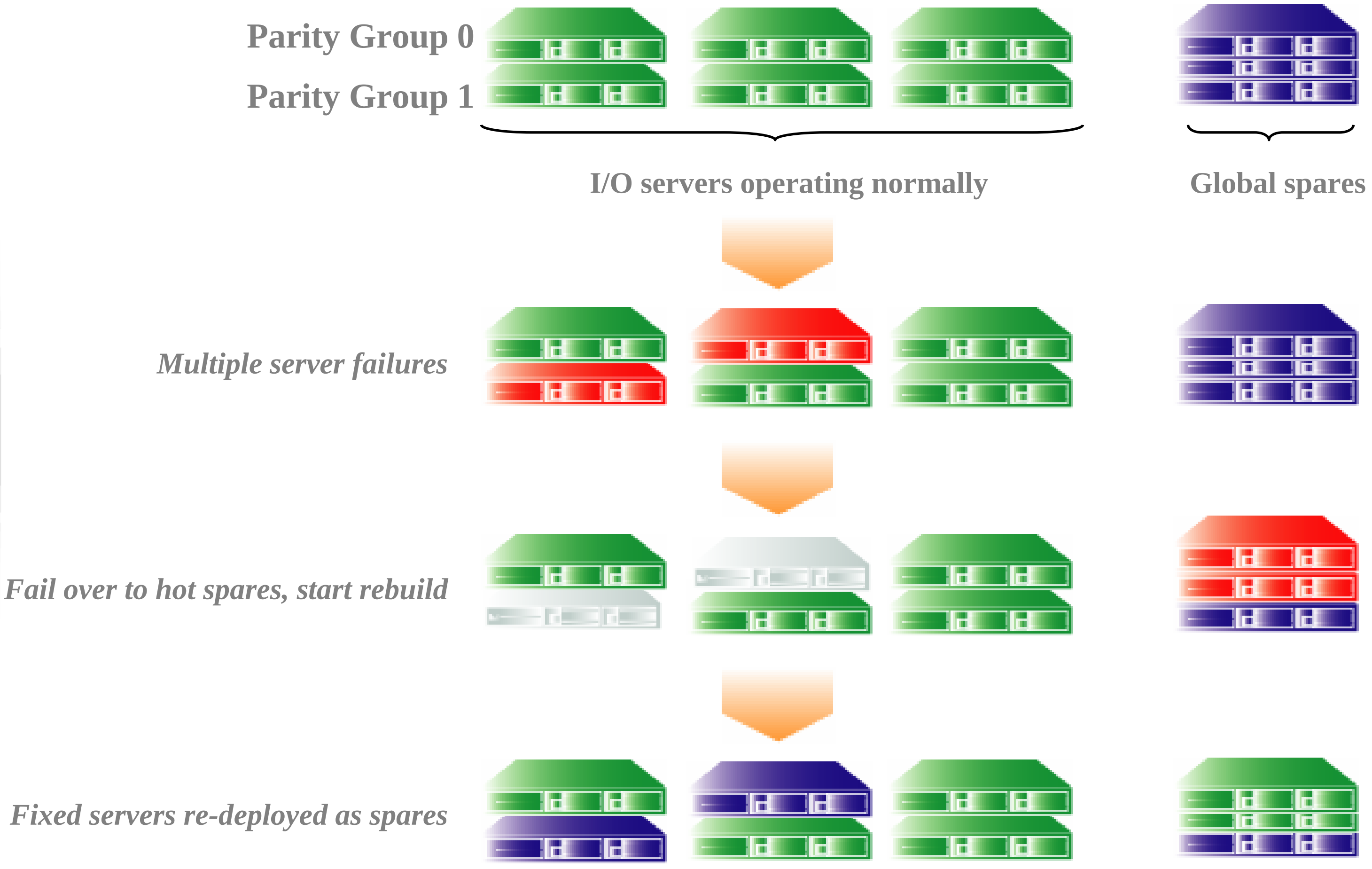
-Block container files reside on standard Linux file system (ext2, ext3, xfs)

-Multiple servers can fail without clients losing access to data

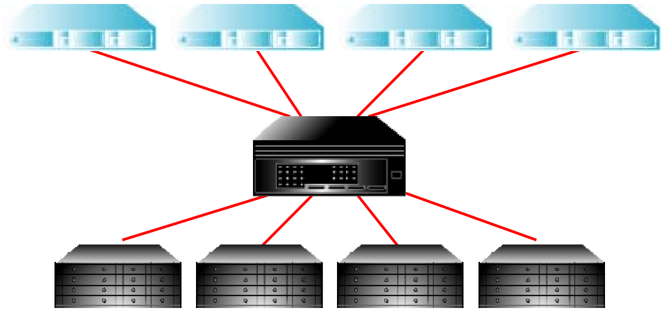
Introducing TerraGrid/HA


Client nodes sustain full wire-speed transfer rates and are unaware of the existence of fault tolerance
Commodity network switch (LAN/WAN)
RAID0 striping on client side
Server node failures are handled transparently on the back-end
RAID-TG protection on server side

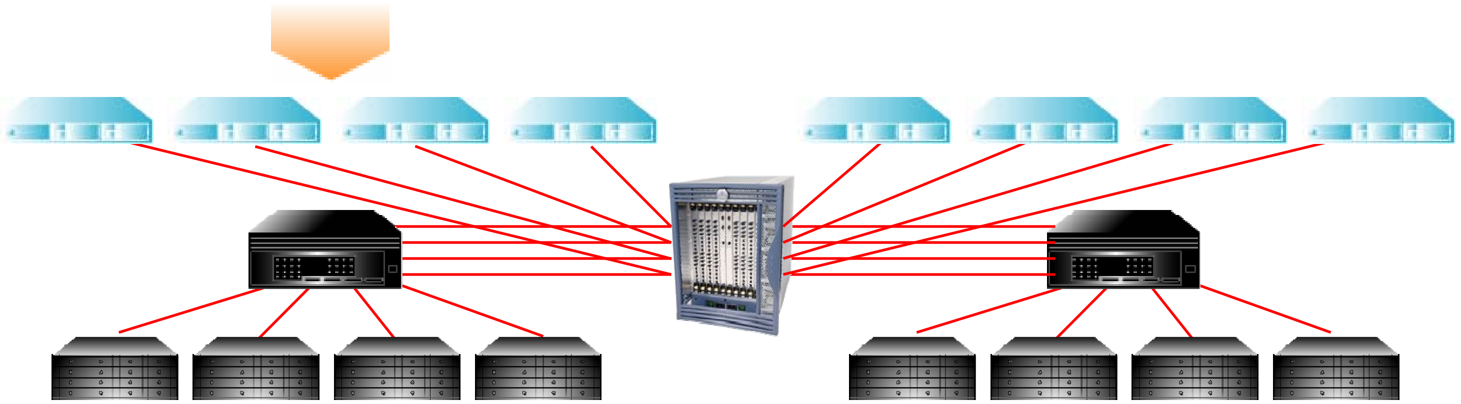
Fault Tolerance with TerraGrid/HA


Parity Group 0
Parity Group 1
I/O servers operating normally
Global spares
Multiple server failures
Fail over to hot spares, start rebuild
Fixed servers re-deployed as spares

The Problem With Existing HA Schemes

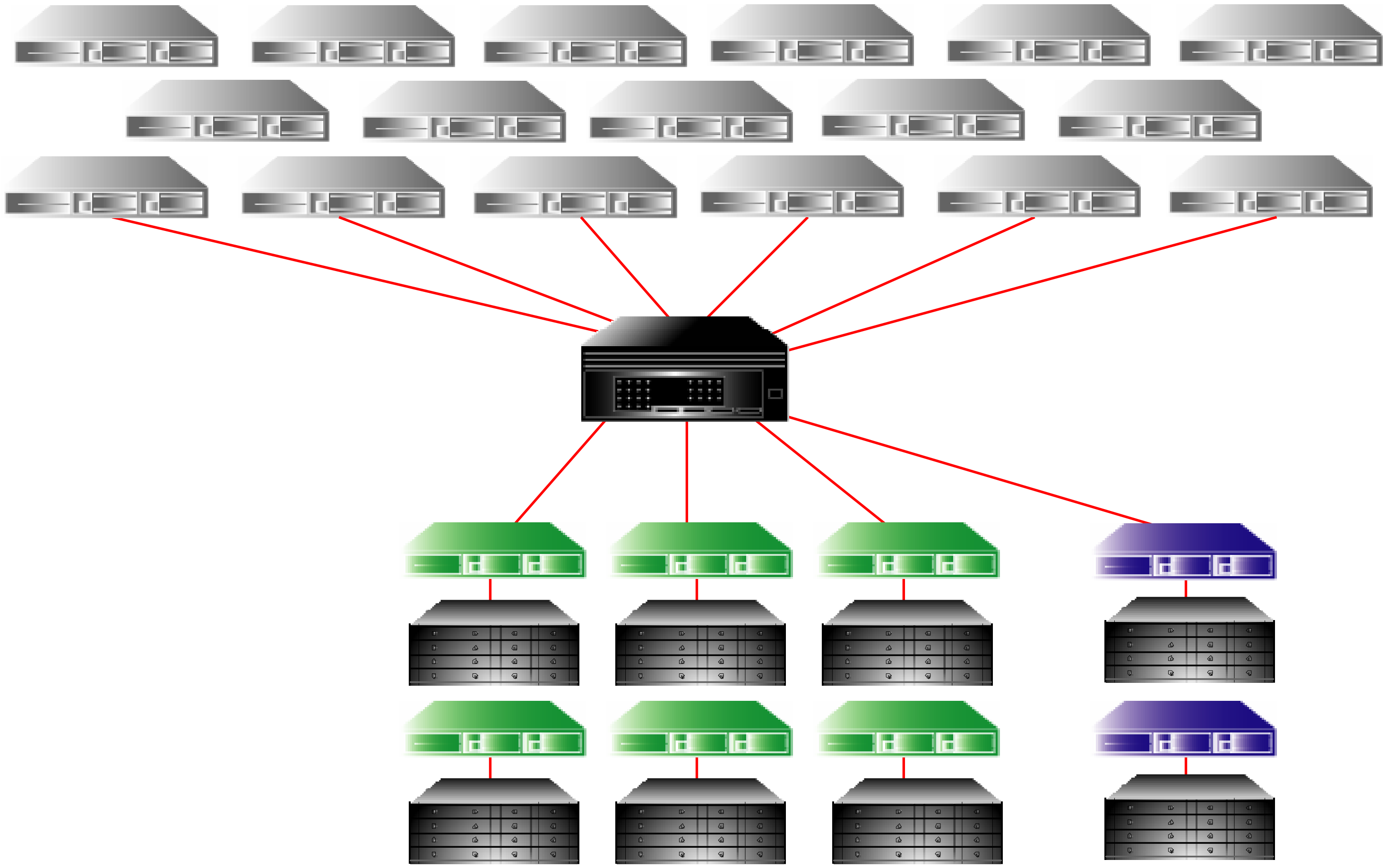


Traditional "monolithic" RAID controllers cost approximately $100,000 per Gbyte/sec of bandwidth (excluding disks). Typically, a maximum of 4-8 2Gbit FC connects to host are supported as well as 4-8 2 Gbit FC connects to disks. The fastest RAID controllers currently available deliver 1000-1400 Mbytes/sec.

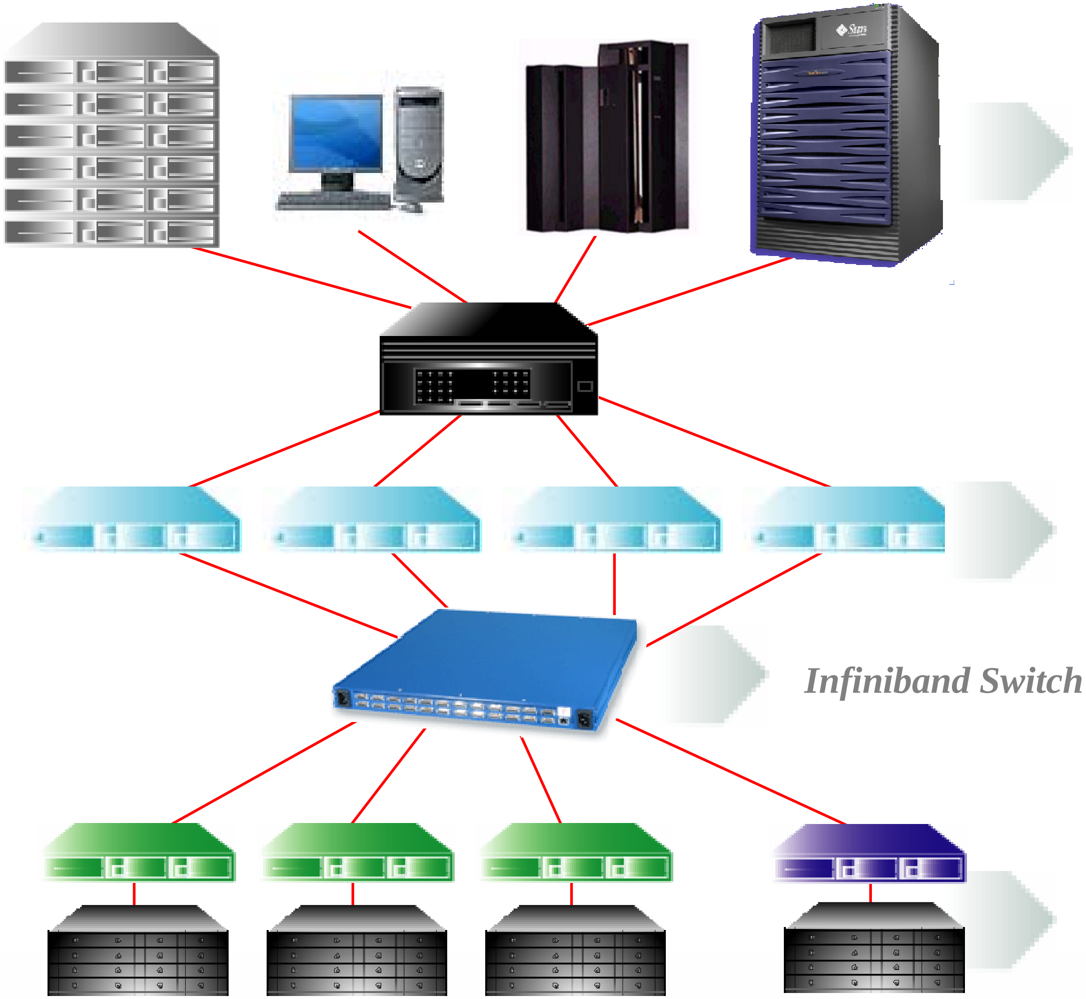


When additional capacity and/or bandwidth is required, it becomes necessary to install a FC switch. This adds cost, complexity and introduces I/O latencies.

Application: Hierarchical Cluster



Application: Scalable NAS



NFS/CIFS Clients

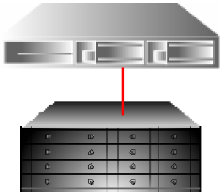
Clustered NAS Heads
100+MByte/sec each

Infiniband Switch

I/O Servers

TERRASCALE TECHNOLOGIES

Typical Database Deployments: *Monolithic*

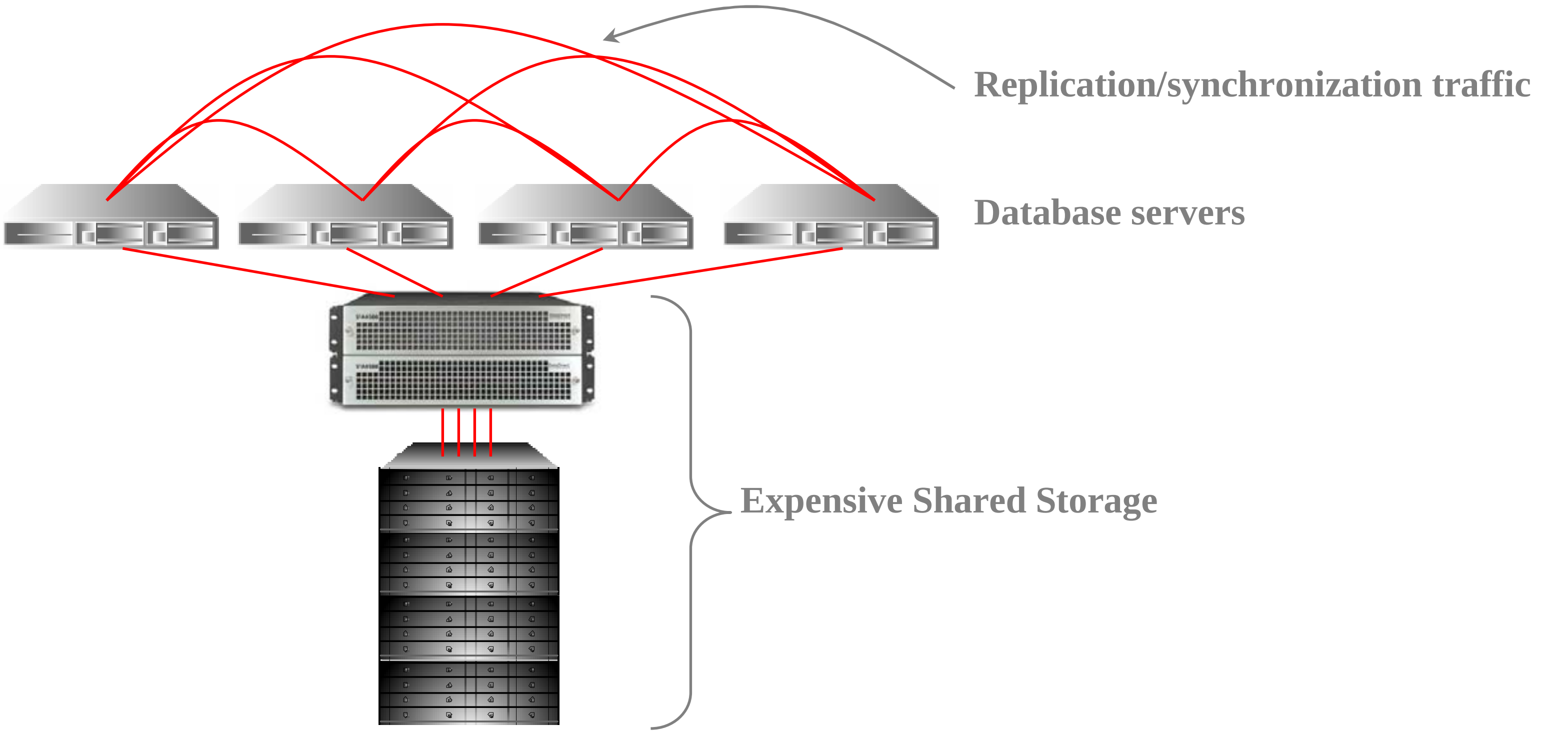


Database server

DAS Storage Array

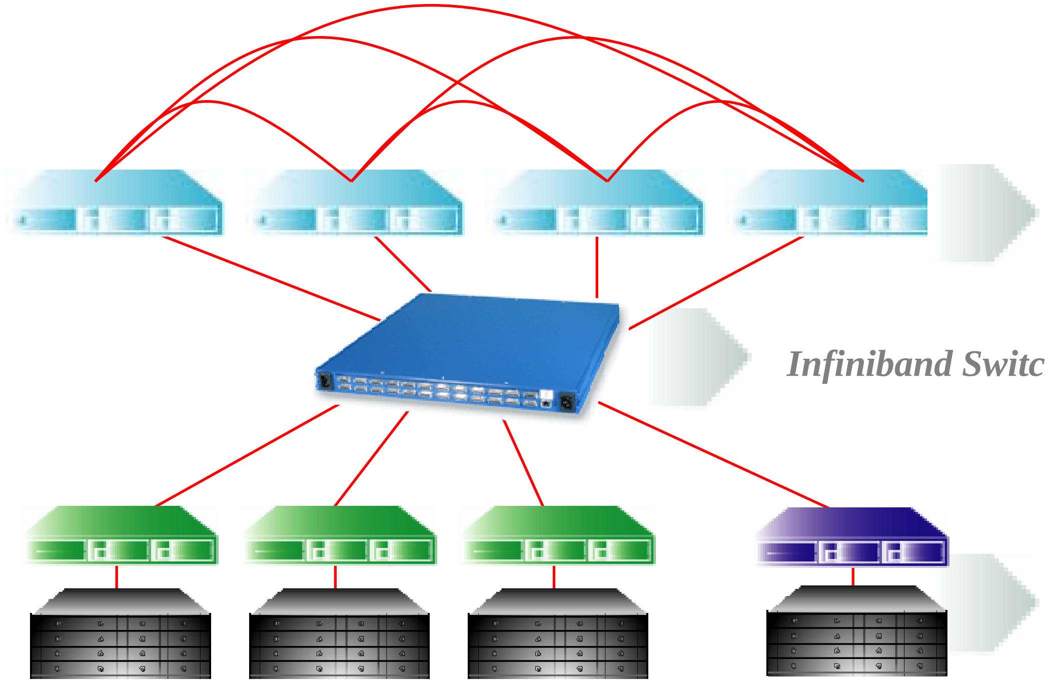
- ➢*Non-scalable:* Saturation occurs when server runs out of CPU/memory DAS box runs out of bandwidth
- ➢*No fault-tolerance:* Failure of server or storage results in database downtime
- ➢*Low performance:* Requests are issued serially to storage array. No parallel access to data
- ➢*Reference Performance:* ~20,000 TPM using *TWO* FC RAID arrays

Typical Database Deployments: *Clustered*



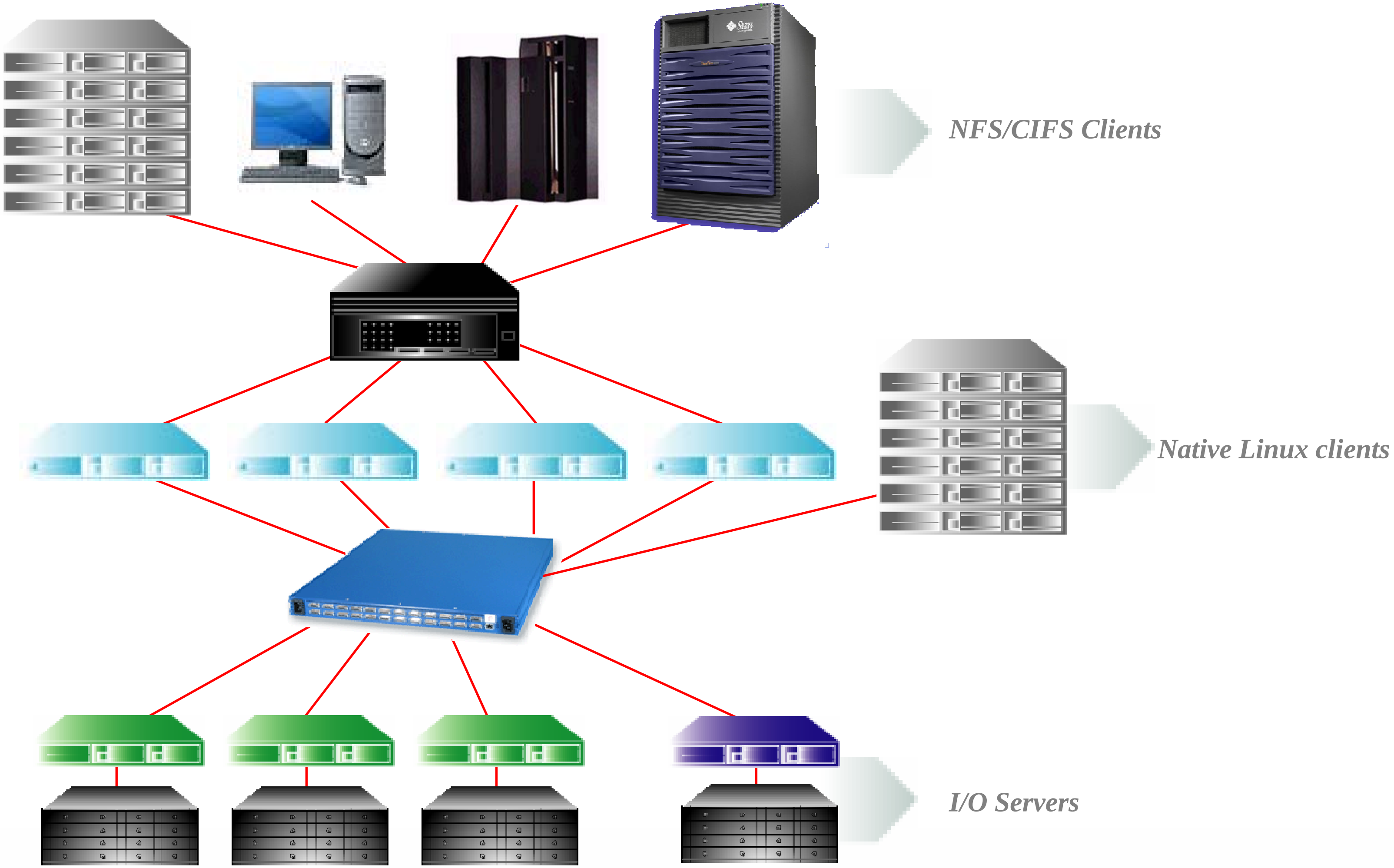
- ➢***Expensive:*** **Multi-ported FC controllers and/or FC switches required for shared storage**
- ➢***Low performance:*** **Requests from each server are issued serially to the FC HBA. FC operates at only 2 Gbits/sec**
- ➢***Limited Scalability:*** **Monolithic RAID controllers run out of IOPS very quickly**
- ➢***Multiple Fabrics:*** **Separate fabrics for I/O and inter-node replication**

Scalable TerraGrid Powered Databases


Low cost Clustered Database Engines (MySQL, Oracle10g, IBM DB2)
Infiniband Switch
Low cost I/O Servers

- ➢ *Parallel Access:* Multiple requests are issued to multiple I/O servers in parallel
- ➢*Fast:* 10 Gbit Infiniband vs. 2 Gbit FC
- ➢*High availability:* New RAID-TG algorithm provides low-cost HA with no performance penalty
- ➢*Radically altered $/TPM:* Target of ~30,000 TPM on $40,000 platform
- ➢*Low cost:* No expensive multi-ported FC storage, use of low-cost commodity components

Bringing it all together: Hybrid Deployment



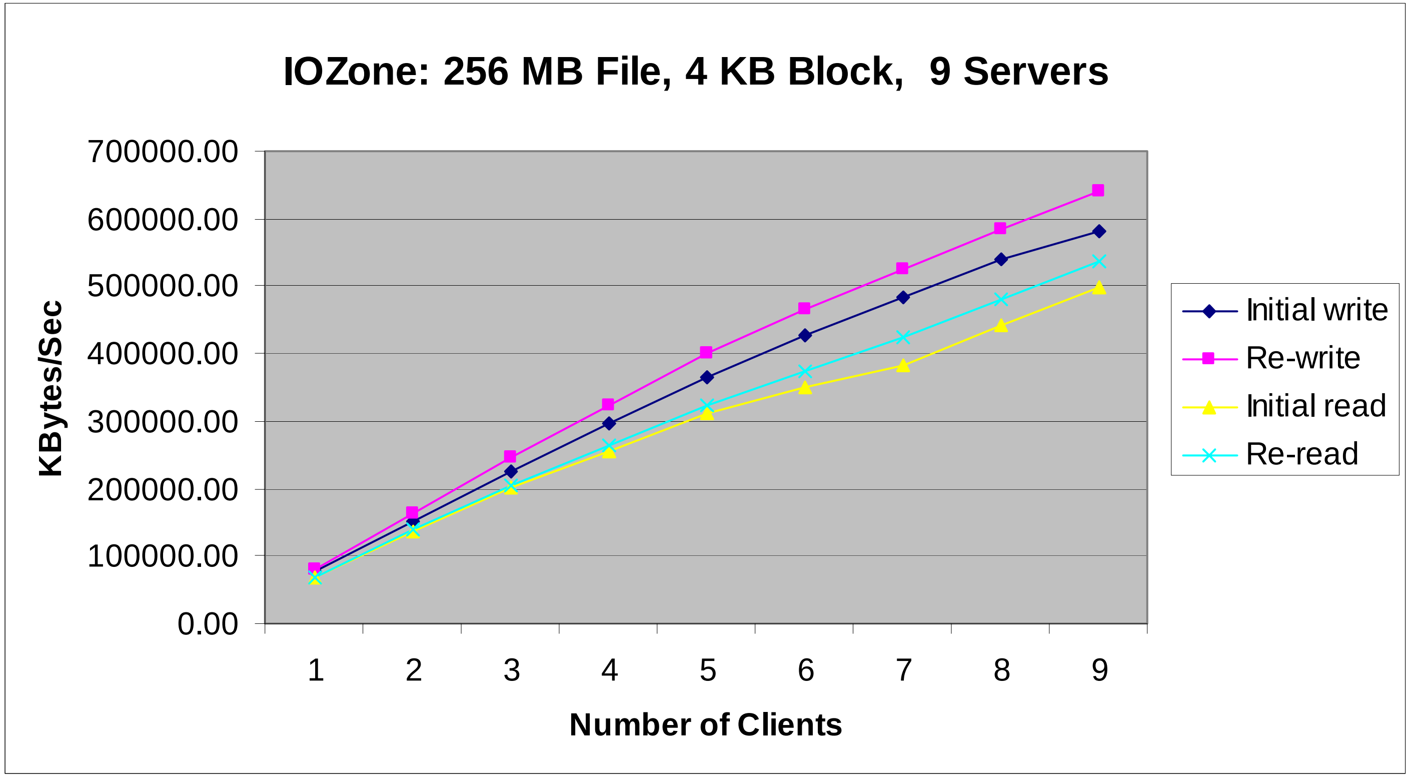
TerraGrid performance metrics

- File system performance (measured using open-source Linux/ext2)
  - 80 Mbytes/sec, 20,000 IOPS/sec per client over Gigabit Ethernet
  - 300 Mbytes/sec, 75,000 IOPS/sec per client over Infiniband
  - Scales linearly to 100s of clients
- Database performance
  - 30,000 TPM for $40,000
  - Single database engine with two I/O servers (15 SATA disks/server)
- RAID-TG performance
  - 110 Mbytes/sec, 27,500 IOPS/sec per client over Gigabit Ethernet
  - 400 Mbytes/sec, 100,000 IOPS/sec per client over Infiniband
  - Delivered performance with 24-port Infiniband switch: 4.8 Gbytes/sec, 1,200,000 IOPS/sec
  - Rebuild failed nodes at 100+ Mbytes/sec

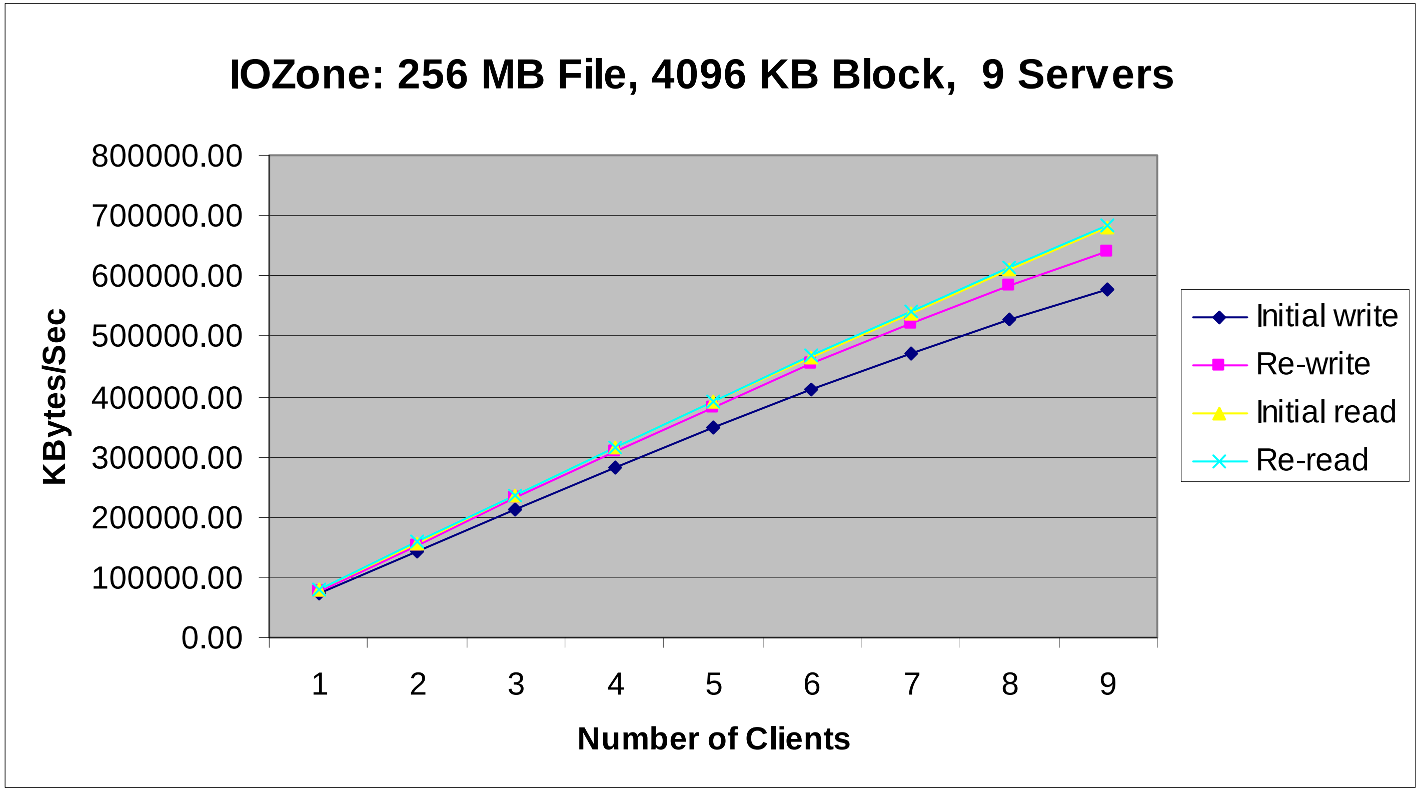
TerraGrid benchmarks: Overview

- Application Nodes (Clients)
  - 9 x IBM x335 dual-CPU, 2.0 GHz Xeon, 512MB RAM
  - Onboard Broadcom Gigabit Ethernet NIC
  - Each initator mounted an *ext2* file system on TerraGrid – all initiators see unified namespace
- I/O Servers
  - 9 x IBM x335 dual-CPU, 2.4 GHz Xeon, 512MB RAM
  - Onboard Broadcom Gigabit Ethernet NIC
  - Single, internal U320 10K RPM SCSI HDD
- Network Switch
  - 32-port Extreme Networks 7i Gigabit Ethernet Switch
- Benchmark
  - The *iozone* benchmark was used to collect the performance data presented herein
  - All data is based on *single stream* I/O (one file per initiator)

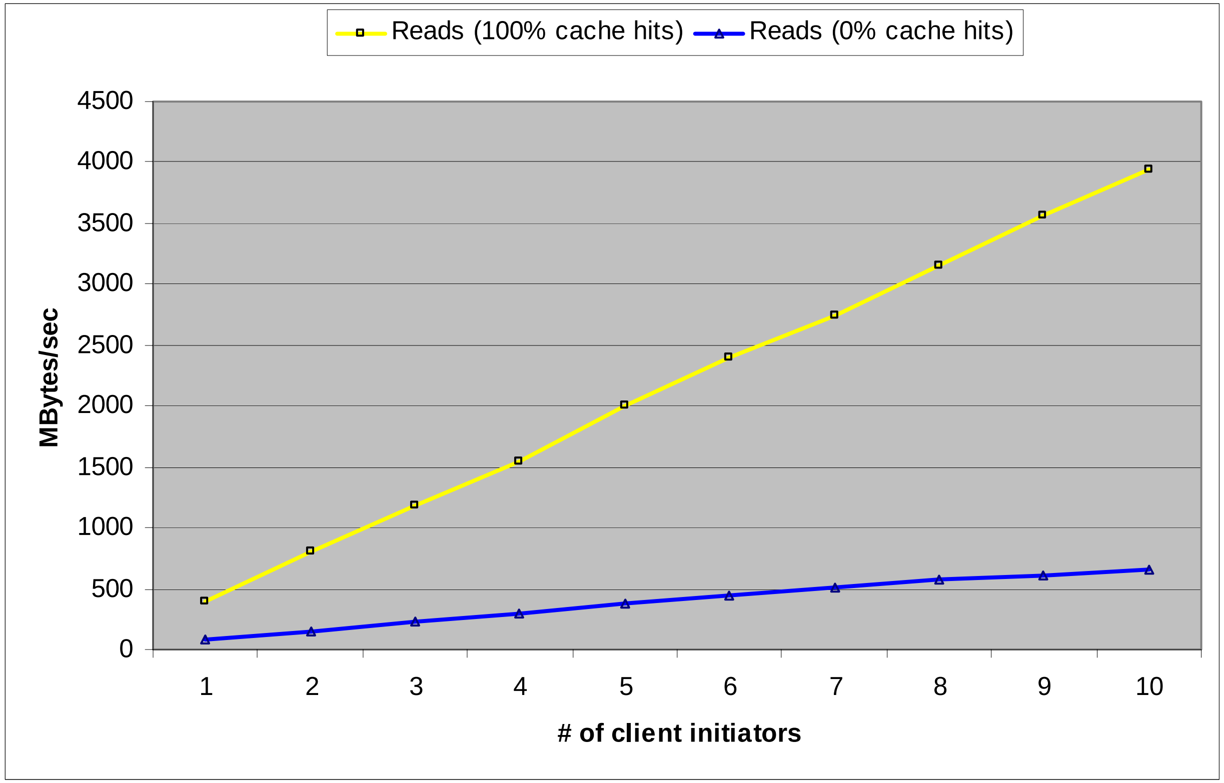
TerraGrid performance scaling (1)



TerraGrid performance scaling (2)

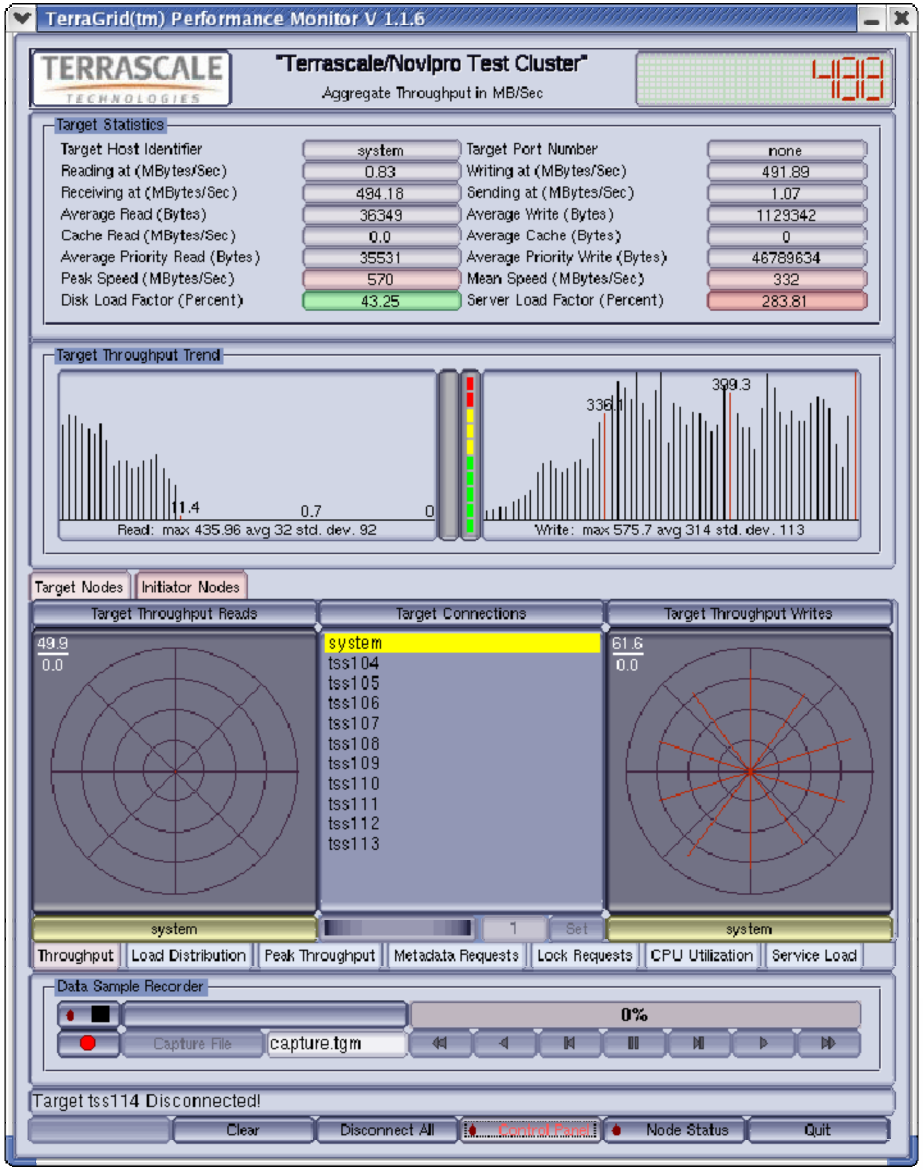


TerraGrid performance scaling (5)



- 100% cache hits were achieved using initiator smart buffering logic
- Smart buffers are cache-coherent with other initiators at block-level granularity

tgmon: TerraGrid Performance Monitor



TerraGrid(tm) Performance Monitor V 1.1.6

"Terrascale/Novlpro Test Cluster"

Aggregate Throughput in MB/Sec

Target Statistics

Target Host Identifier	system	Target Port Number	none
Reading at (MBytes/Sec)	0.83	Writing at (MBytes/Sec)	491.89
Receiving at (MBytes/Sec)	494.18	Sending at (MBytes/Sec)	1.07
Average Read (Bytes)	36349	Average Write (Bytes)	1129342
Cache Read (MBytes/Sec)	0.0	Average Cache (Bytes)	0
Average Priority Read (Bytes)	35531	Average Priority Write (Bytes)	46789634
Peak Speed (MBytes/Sec)	570	Mean Speed (MBytes/Sec)	332
Disk Load Factor (Percent)	43.25	Server Load Factor (Percent)	283.81

Target Throughput Trend

Read: max 435.96 avg 32 std. dev. 92

Write: max 575.7 avg 314 std. dev. 113

Target Connections

system
tss104
tss105
tss106
tss107
tss108
tss109
tss110
tss111
tss112
tss113

Data Sample Recorder

Target tss114 Disconnected!

TerraGrid Roadmap

- Working with fabric vendors to produce mutual certifications
  - Recently did demos at conferences with Voltaire, more vendors to be announced shortly, Linux-World Award “Best Database Product”
- Terrascale NAS building blocks
  - Terrascale is working with blade & IB vendors to test and certify reference NAS platforms
  - Objective is to enable OEMs/VARs/Clients to build their own scalable NAS solutions while preserving commodity hardware pricing
  - 100Ks of IOPS/sec, Gbytes/sec of throughput
- Building relationship to HW-vendors
  - Intel- & Opteron-Blade vendors: Verrari, Angstrom MS, Western Scientific
  - work with Sun and HP in projects
  - working on strategic OEM relationships

TerraGrid Roadmap

- Terrascale DB reference platforms
  - Publish performance metrics for various databases (vendor certified) running on different TerraGrid configurations & H/W platforms
- SC2004 StorCloud Challenge
  - Partnership with Sandia National Laboratories and the ASCI program in Pittsburgh, and some other surprise at SC2004
- Add more features to file system
  - Snapshots
  - Grid deployments
  - Self-healing
- Add more base file systems in addition to the ext2
  - Reiser4
  - XFS
  - port the server-side user space deamon to Solaris and other Unix-Servers

At a glance: TerraGrid Benefits

- Storage networking is consolidated onto existing network fabric
  - Substantial cost reduction and enhanced functionality
- Linear scaling of capacity and bandwidth
  - Seamless growth to Petabytes of capacity and 100s of GB/sec of throughput
- Unified namespace achieved using standard an Open Source Linux file system
  - Massively parallel block, file and file system access
  - Deployable on LANs (clusters) and WANs (GRID) and Low Latency Fabrics
- Extreme availability
  - Multiple server failures will not result in data loss – dial in the desired MTBF
  - Extremely fast rebuild rates
- Multiple deployment models
- Co-exists with existing SAN and SRM software
- Non-disruptive deployment model
- Commodity hardware and software pricing
- capability to make ANY existing file system really scale out