



Education

Continuous Available Commodity Storage

Rekha Singhal & Zia Saquib, CDAC, India

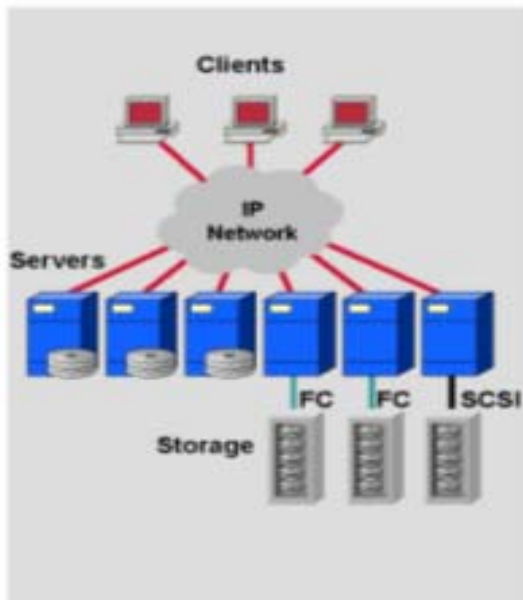
SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
 - ◆ Any slide or slides used must be reproduced without modification
 - ◆ The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.

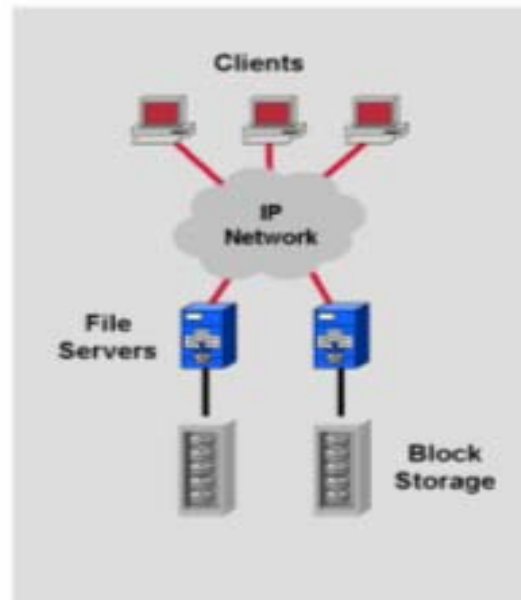
➤ Continuous Available Commodity Storage

- ◆ We shall discuss an alternative cost effective means of building a SAN using iSCSI on a TCP/IP network. We propose an architecture for which the main features are that the storage is inexpensive, continuously available, scalable to LAN/WAN, flexible and has high performance.
- ◆ The key components of the solution are the Controller and the Optimized IP Storage node which may be built using commodity hardware. These components may be placed together in a single node or in two different nodes giving rise to 2-layer and 3-layer architectures. We also aim to derive a theoretical framework for measuring performances of these architectures
- ◆ This session will appeal to Storage Developers and Managers, in SMI segment, who are entering into the area of IP storage networking.

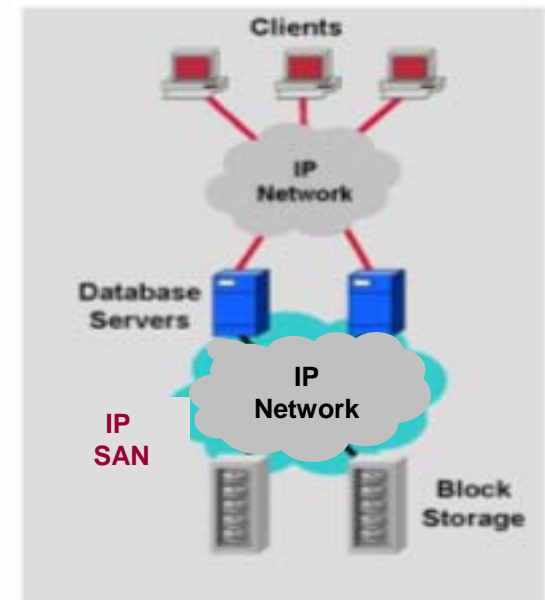
Storage Architectures



**Direct-Attached
Storage (DAS)**

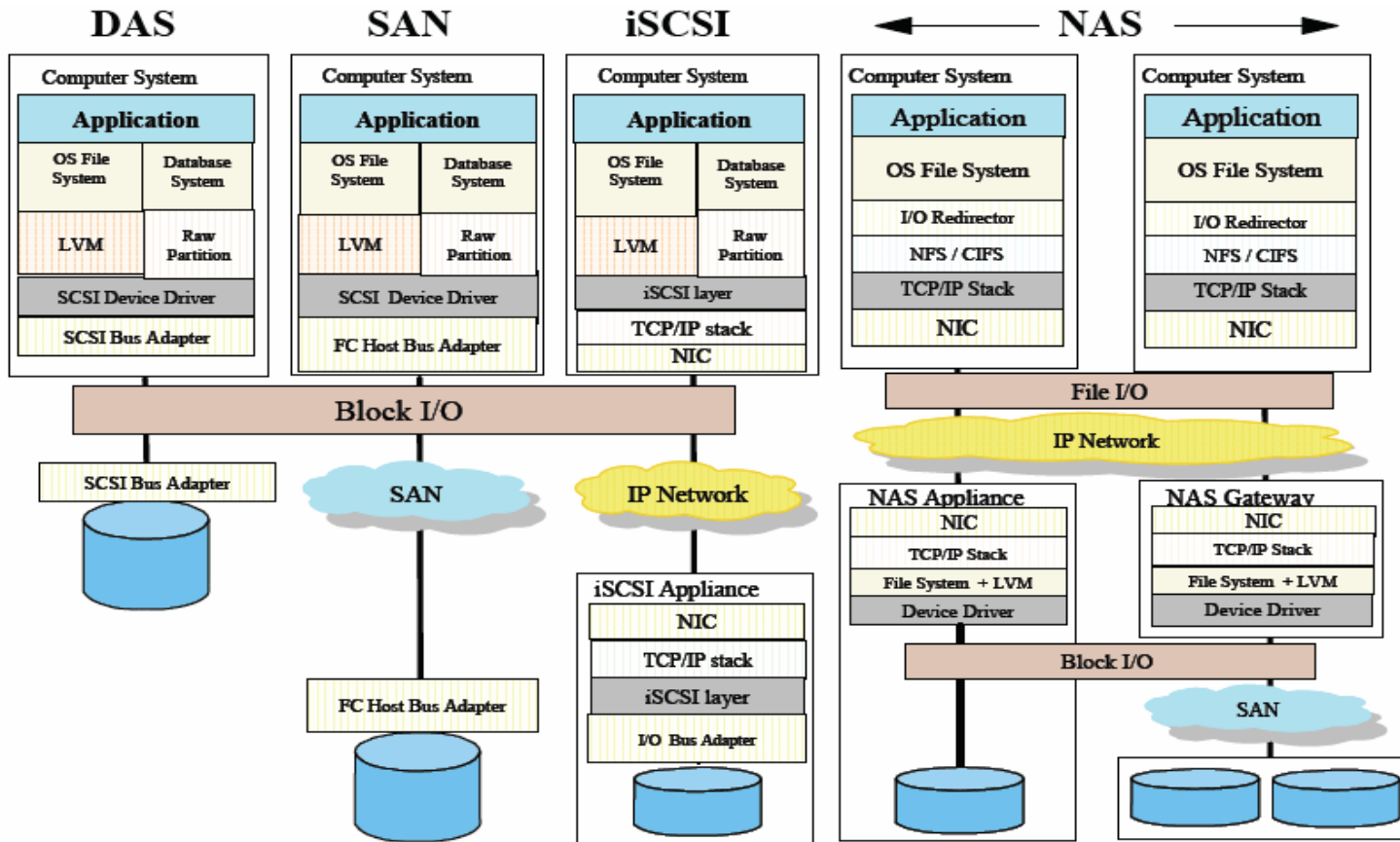


**Network-Attached
Storage (NAS)**

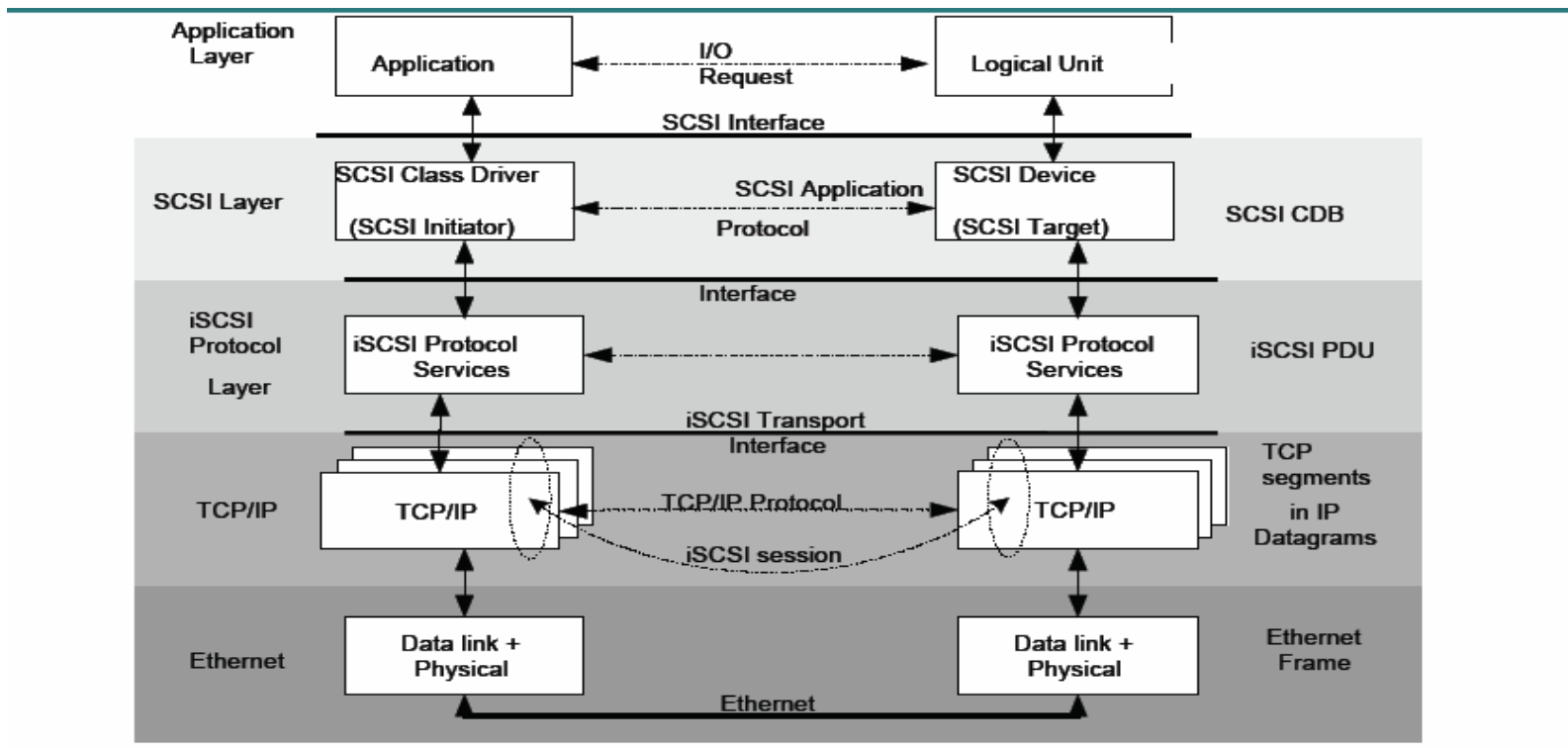


IP Storage Area Network

DAS, SAN, iSCSI and NAS



iSCSI Layered Model



Transparently encapsulates SCSI Command Descriptor Blocks (CDBs)

➤ Performance

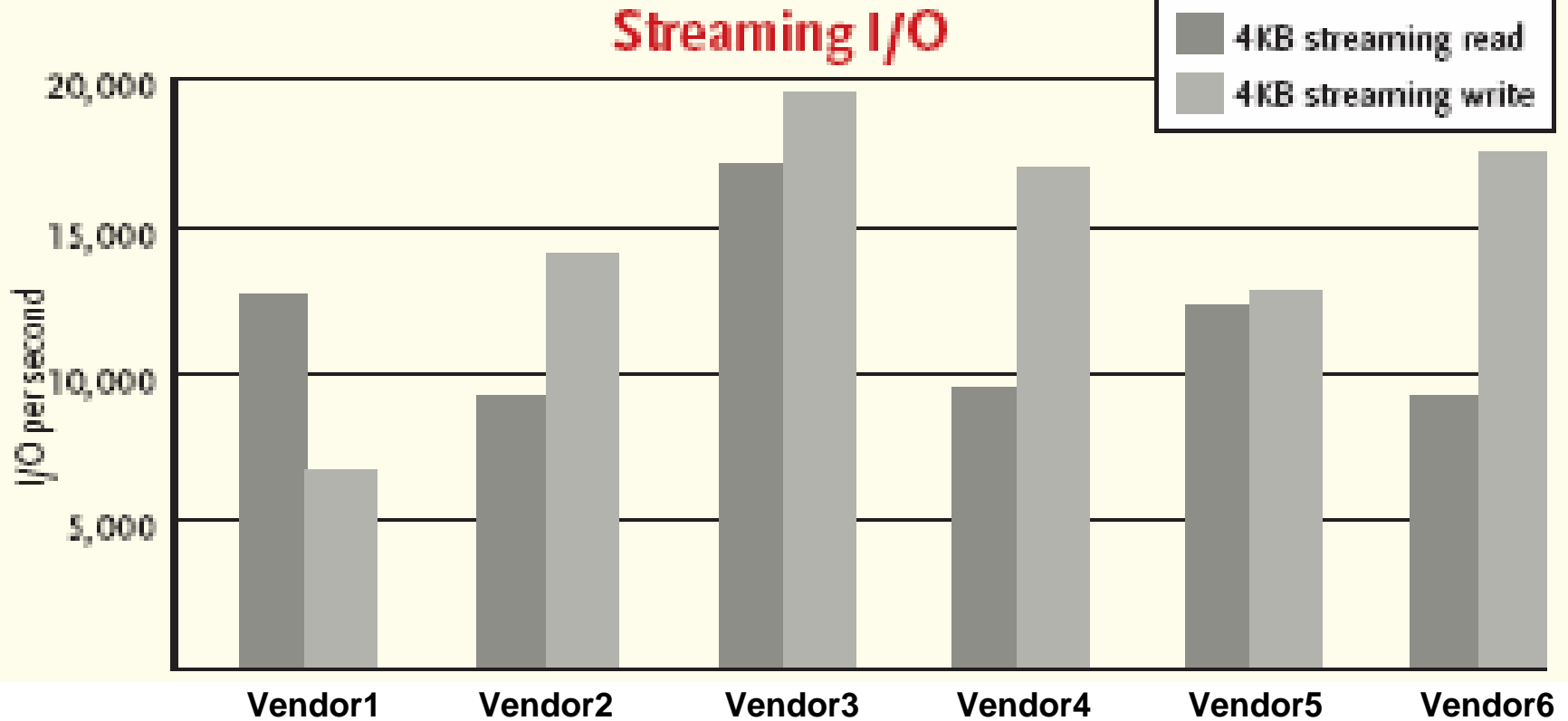
- ◆ Speed of Ethernet – 100 MBPS (now 1 GB to 10GB)
- ◆ Speed of PCI bus – 133 MHz (now 800 MHz PCI Express)
- ◆ Overhead of TCP/IP protocol
- ◆ Overhead of iSCSI protocol

➤ Security

- ◆ IPSec
- ◆ Challenge Handshake Authentication Protocol (CHAP)/Secure Remote Protocol (SRP)
- ◆ Encryption/Authentication

State of the Art

Reference: InfoWorld, October 3, 2005, Issue 40.



QOS parameters for IP SAN

➤ Cost

- ◆ Commodity hardware such as Ethernet , SATA hard disks and existing Servers/Desktops

➤ Performance

- ◆ Caching and Parallelism

➤ Scalability

- ◆ Clustering

➤ Availability

- ◆ Clustering

An Efficient Architecture

➤ Technology

- ◆ Caching, Parallelism, Distributed Computing and Clustering

➤ Intelligent Controller

- ◆ Block Virtualization
- ◆ Caching
- ◆ Redundant Array of Inexpensive Nodes (RAIN) cluster management software for addition/deletion of nodes in cluster

➤ Lazy Replication

➤ Tiered Storage

➤ Optimized IP Storage Array

Optimized IP Storage Array

➤ iSCSI Optimization

- ◆ Parameter negotiation
 - MaxRecvDataSegmentLength, ImmediateData, InitialR2T, MaxBurstLength, FirstBurstLength, MaxConnections
- ◆ Parallel iSCSI
 - Multiple physical network connections with an algorithm to execute out of order commands in parallel while exploiting operation semantics

➤ OS optimization

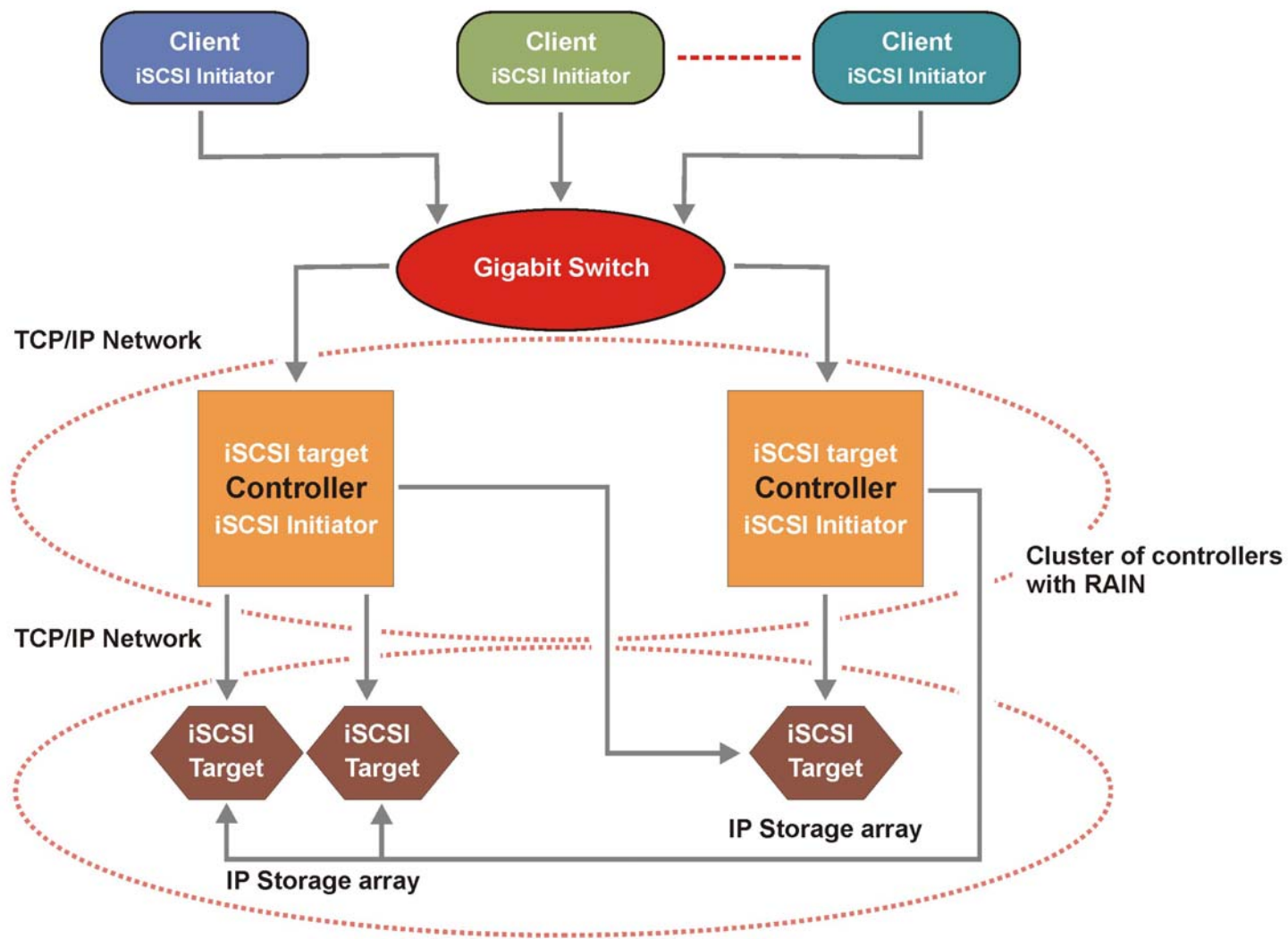
- ◆ Tweaking OS Parameters: Jumbo Frame, Checksum offloading to NIC, Zero-Copy in TCP/IP
- ◆ OS kernel Modification: disk and network based processing only

Optimized IP Storage Array

➤ Multilevel Caching

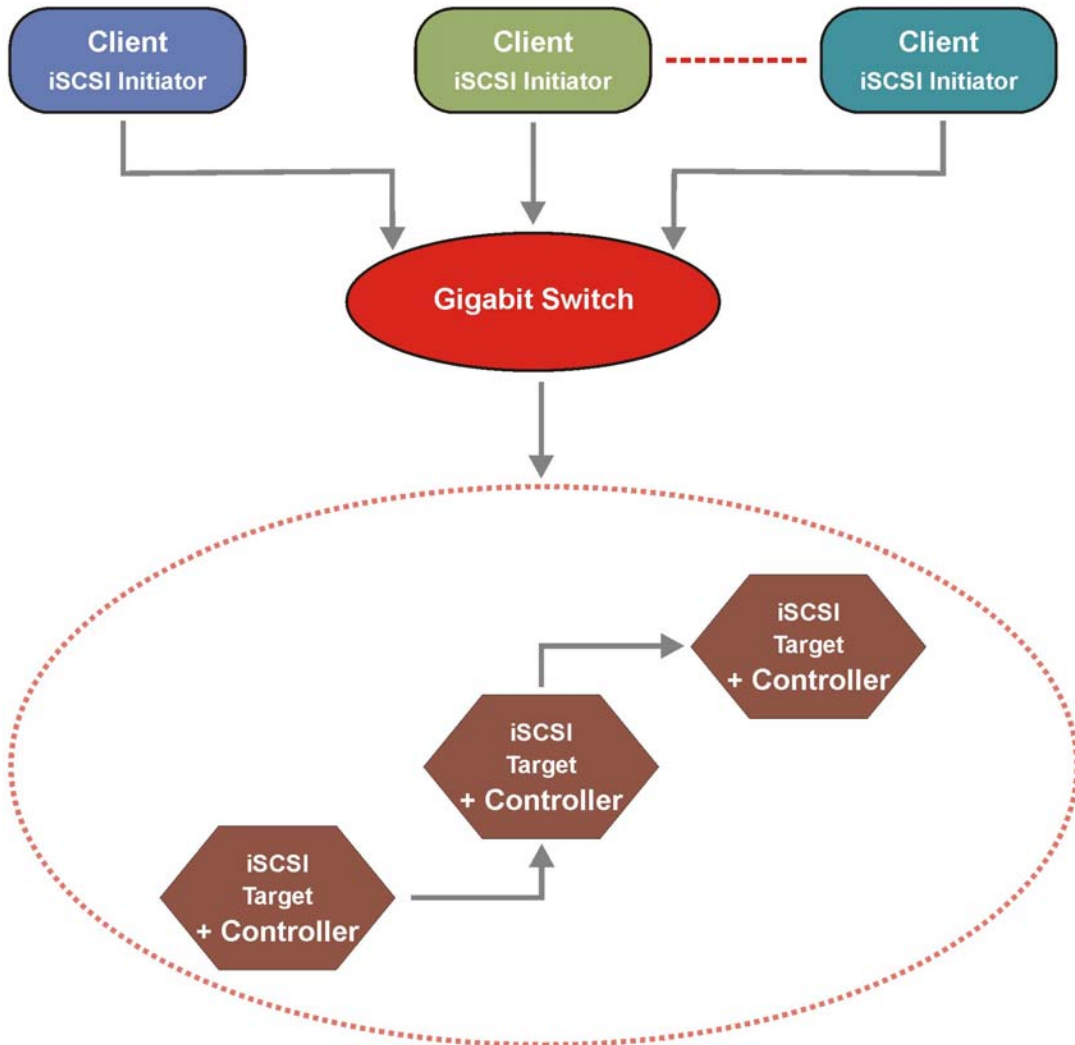
- ◆ Cache at initiator to avoid TCP/IP overheads
- ◆ Cache at target to avoid disk access delays and
- ◆ Cache at NIC card to overcome interconnect bus bottleneck

3 Layer Architecture



Set of IP storage Array i.e. commodity machines each running optimized iSCSI target

2 Layer Architecture



Cluster of commodity machines each running optimized iSCSI target and Controller module

- Adding a new optimized storage node leads to
 - ◆ Capacity expansion
 - ◆ Performance improvement through Load Balancing
 - ◆ Adding controller node in 3-layer architecture further adds to performance.
- Availability = $MTBF / (MTBF + MTTR)$
- RAIN makes sure that MTTR is very small due to fast switching of processing of a failed node to another functional node in the cluster.

Performance Benchmarks for SAN

➤ Throughput (MB/sec)

- ◆ Number of bytes accessed per second or, the number of I/O operations (transactions) which includes both read and write operations, carried out per second.

➤ Response Time (ms)

- ◆ The response time is the time taken to finish a storage operation i.e. time lag between the submission of a block access request and getting its result back. It includes queuing time and latency as well

Theoretical Performance

➤ Assumptions

- ◆ NVRAM, so read and write in cache only.
- ◆ Write back strategy
- ◆ Same bandwidth, 'B', between all nodes in the cluster
- ◆ Same number of sequences for both read and write requests, 'N' which is the number of sequences for data transfer as part of iSCSI protocol

Theoretical Performance

$$\blacktriangleright \text{Response Time} = N * C + N * [P * T_{\text{cacheaccess}} + (1 - P) * T_{\text{clusterdataaccess}}] + N * \text{RTT} + S/B$$

Where,

- ◆ C is the system dependent time taken for data encapsulation.
- ◆ $T_{\text{cacheaccess}}$ is time to read from cache at controller.
- ◆ T_{HDaccess} is time to read data of size S from hard disk.
- ◆ S is the size of data to read or write from network storage
- ◆ b_{max} is maximum data that can be received at target/initiator.
- ◆ RTT is the round trip time.
- ◆ P is the probability to find data in cache (Hit ratio).

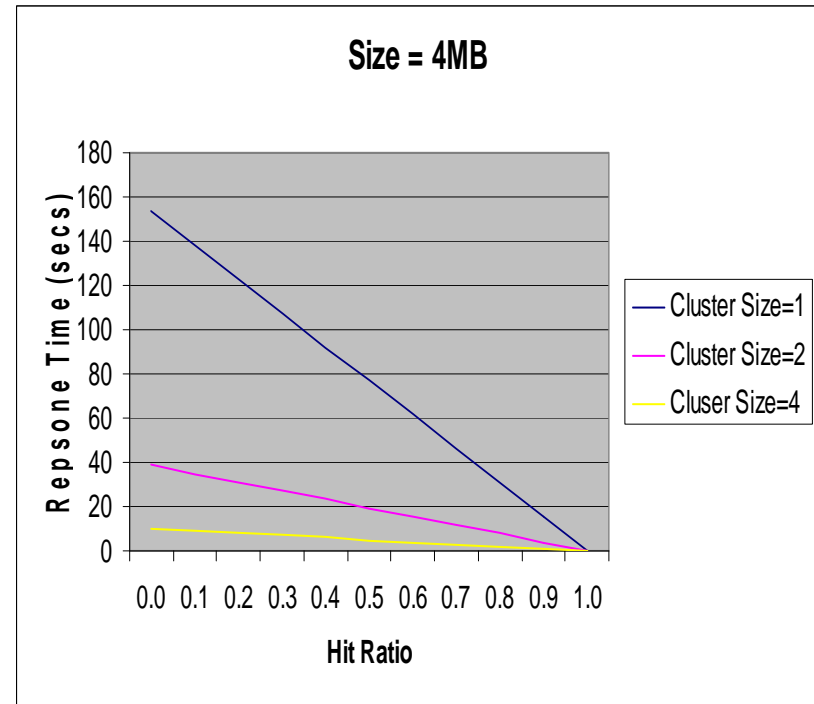
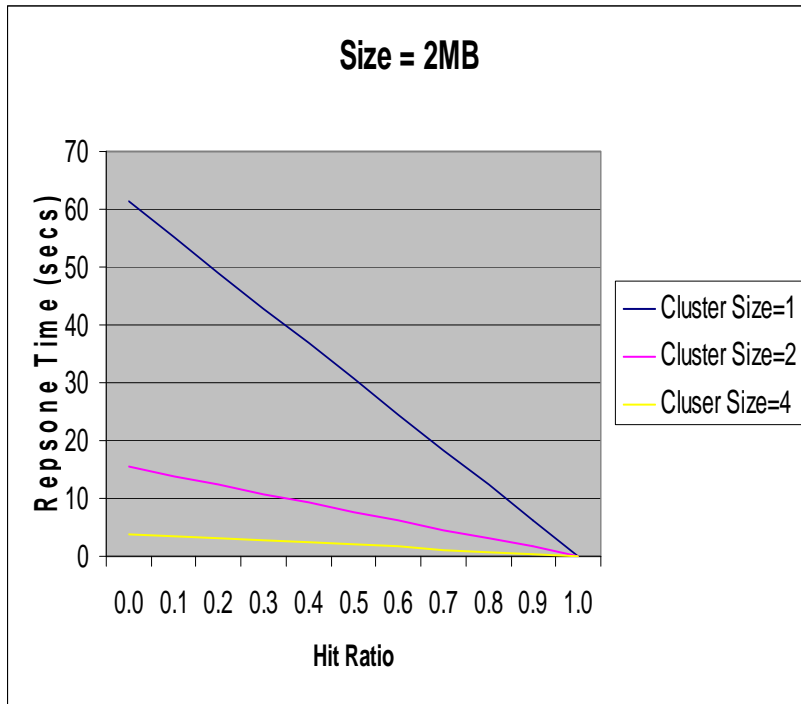
Theoretical Performance

$$\text{Response Time} = (S/b_{\max} * (I-P)/C_n) \\ [S/B + T_{HD\text{access}}/C_n + S/b_{\max}(C + RTT)] + \\ S/b_{\max} * P * T_{\text{cacheaccess}} + S/b_{\max} * (C + RTT) + S/B$$

Where,

- N is the number of sequences for data transfer as part of iSCSI protocol
- C is the system dependent time taken for data encapsulation.
- $T_{\text{cacheaccess}}$ is time to read from cache at controller.
- $T_{HD\text{access}}$ is time to read data of size S from hard disk.
- RTT is the round trip time delay in the network between any two nodes.
- B is the bandwidth provided by the underlying TCP/IP network.
- C_n is the number of nodes in the cluster.
- S is the size of data to read or write from network storage
- b_{\max} is maximum data that can be received at target/initiator.
- P is the probability to find data in cache (Hit ratio).

Theoretical Performance



Infiniband iSCSI Storage

- Add InfiniBand support to existing iSCSI based product lines
- Leverage InfiniBand's high Performance
- Leverage iSCSI management tools

Infiniband based iSCSI

- **Data Path – Over InfiniBand Verbs/CMA**
 - ◆ High Bandwidth, Low Latency, RDMA
- **Control Path – Using iSCSI headers**
 - ◆ Discovery, naming, security, error-recovery, booting, etc.
- **Leverages the wide adoption of iSCSI**
 - ◆ OS code and storage products
 - ◆ Management tools and standard interfaces
 - ◆ Standardization, Testing and Protocol maturity
 - ◆ End-user training

Infiniband based iSCSI

- InfiniBand solutions leverage the iSER (iSCSI RDMA) protocol.
- iSER brings significantly greater performance to iSCSI and leverages the protocol's existing comprehensive management capabilities, allowing heterogeneous storage environments to utilize a single protocol and management infrastructure.
- iSER over 10 Gbps (SDR) and 20 Gbps (DDR) InfiniBand delivers markedly better performance than Fibre Channel at significantly lower costs.

Strategies for IPSAN

- High Performance Computing Environments
- Availability – SLA Guarantees by HA Levels
- Cost - Total Cost of Ownership (TCO)
- Scalability – Virtualization, Clustering
- Interoperability
- Manageability (Virtualization, Self-Configuring, Self-Optimizing with Workload & Self-Healing ...)
- Security – IPsec6 vs. CIM based

Related Work in Literature

➤ Literature

- ◆ Caching
 - X. He, Q. Yang, and M. Zhang, "A Caching Strategy to Improve iSCSI Performance", Proceedings of the 27th Annual IEEE Conference on Local Computer Networks , 2002.Parallel iSCSI by Qing Yang
- ◆ Parallel iSCSI
 - Q. Yang"On Performance of parallel iSCSI Protocol for Networked Storage System", Proceedings of 20th International Conference on Advanced Information Networking and Applications, IEEE 2006.

Refer to Other Tutorials



**Check out SNIA Tutorial:
IP Storage, 2006**

- Please send any questions or comments on this presentation to SNIA: trackstorage@snia.org

**Many thanks to the following individuals
for their contributions to this tutorial.**

- SNIA Education Committee

**Rekha Singhal
Zia Saquib**

**Brandy Barton
Nancy Clay**

Rob Peglar