



SCSI-DSDC

A SCSI Transport Layer Extension with
Separate Data and Control Paths for Scalable
Storage-Area-Network Architectures

Yitzhak Birk
Nafea Bishara

Agenda

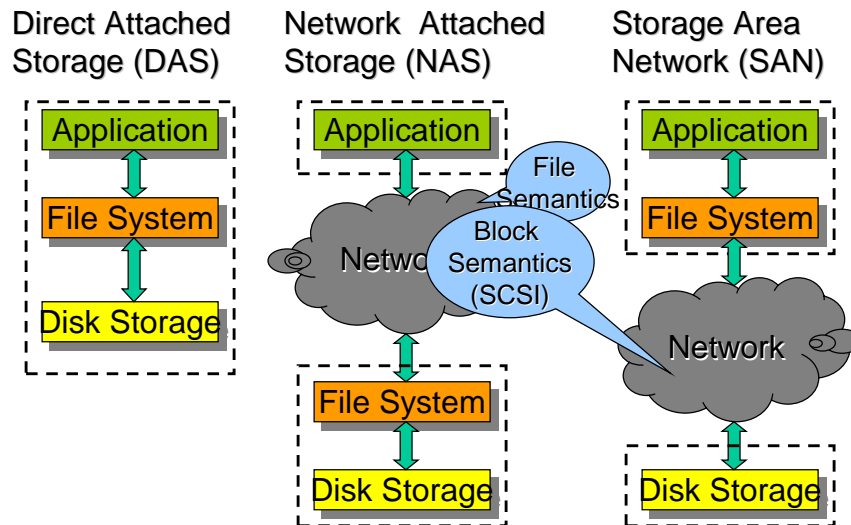
- Introduction to Storage Area Networks and Problem Statement
- SCSI Protocol and associated SCSI Transport Protocol
- Distributed and Split Data-Control extension to SCSI Transport Protocol
- Prototype and Performance results

Introduction to SAN and problem statement

Storage in the internet era

- Storage demand is increasing rapidly:
 - Traditional Internet and Enterprise application
 - Emerging of new killer applications:
 - Local Backups, Remote Backups and disaster recovery
 - Requirement from Storage subsystem:
 - Seamless Scalability: In performance and storage space
 - Multi-platform interoperability
 - Performance
 - Reliability and Availability
- ⇒ Dedicated Storage Sub-system independent of the clients or Application Servers

Various Enterprise Storage Models



Tsahi Birk and Nafea Bishara

5

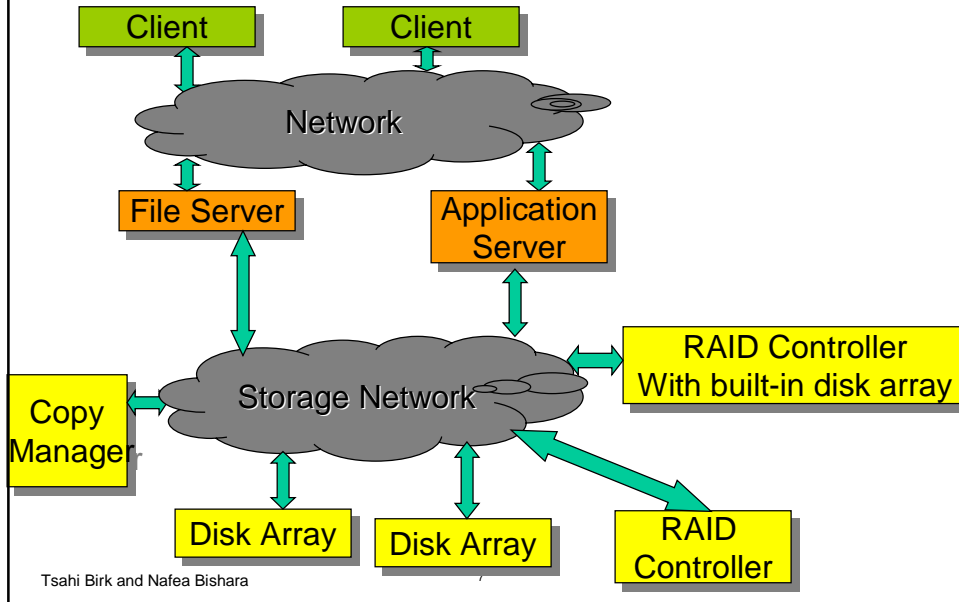
Storage Controller - The heart of SAN

- The storage Controller is a software/hardware entity that manages one or more storage-containing entities and provides:
 - A simple and abstract view of the managed devices
 - Making a large store from many small ones
 - Data striping (RAID)
 - For load balancing, throughput, fault-tolerance..
 - Manages spare disks
 - For seamless fail-over
 - Local and Remote mirroring
 - Data Caching
 - Access control (LUN masking and reservation)
 - LUN masking
- Controller hierarchy is supported as well

Tsahi Birk and Nafea Bishara

6

A Real World Storage Hierarchy Example



Storage Controller - Location in the SAN

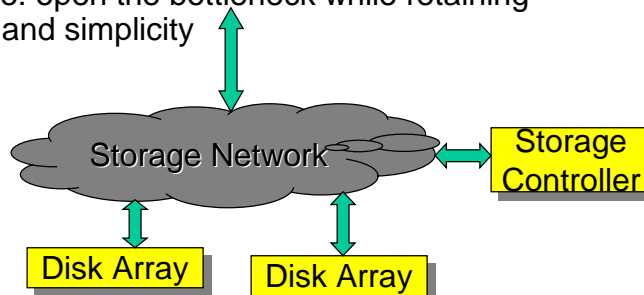
The SAN controller can be found:

- Internal to the host/server
 - Disadvantage: Does not allow sharing of the storage system between multiple hosts/servers
- In the same enclosure with the disk arrays/tape drives
 - Disadvantage: Controller capacity limited by the capacity of the enclosure
- A standalone entity, connected to the SAN, and manages other disk arrays and Controllers
 - Advantages:
 - Supports managing more than one enclosure.
 - Facilitates redundant controllers and makes backups simpler
 - Facilitates multi-vendor systems and interoperability

The industry trend is going toward standalone SAN controller

Problem Statement

- All control and data flows through the storage controller, making it a potential bottleneck.
 - Computational or Network bottleneck
 - Limits the overall storage area network performance
- The problem becomes more severe with 10Gigabit Networks and more that 1Gigabit HDD transfer rates
- The challenge: open the bottleneck while retaining compatibility and simplicity



Tsahi Birk and Nafea Bishara

Prior Art

- Many studies focused on distributed File-System implementations (NAS) or Network-Attached Storage/Secure Devices (NASD)
 - Relatively high-level semantics (files, objects)
 - Heavy requirement on storage devices and Controller, that need to handle files and objects, and even run part of application code
- Other studies focused on efficient data block transfers
 - Remote DMA (Over VIA, InfiniBand or over TCP/IP)
 - Parallel Transport Protocol
 - Focus on data moving and placement, but direct peer-to-peer architectures (no Controller)

Tsahi Birk and Nafea Bishara

10

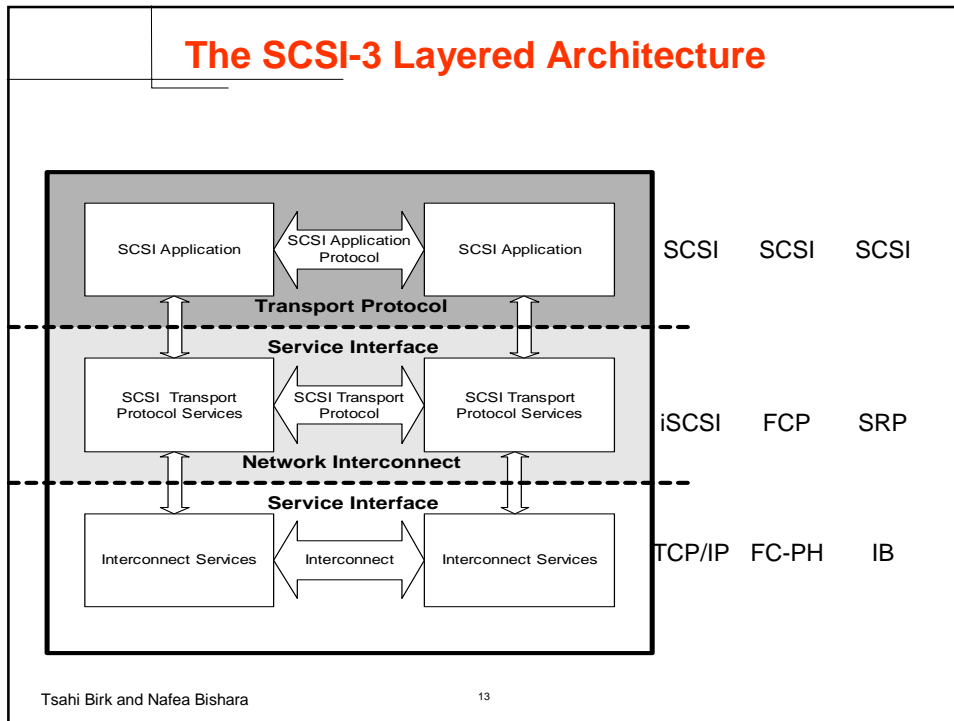
The Goal of Our Work

Define an architecture that would scale the SAN performance, relieving the controller bottleneck under the following constrains:

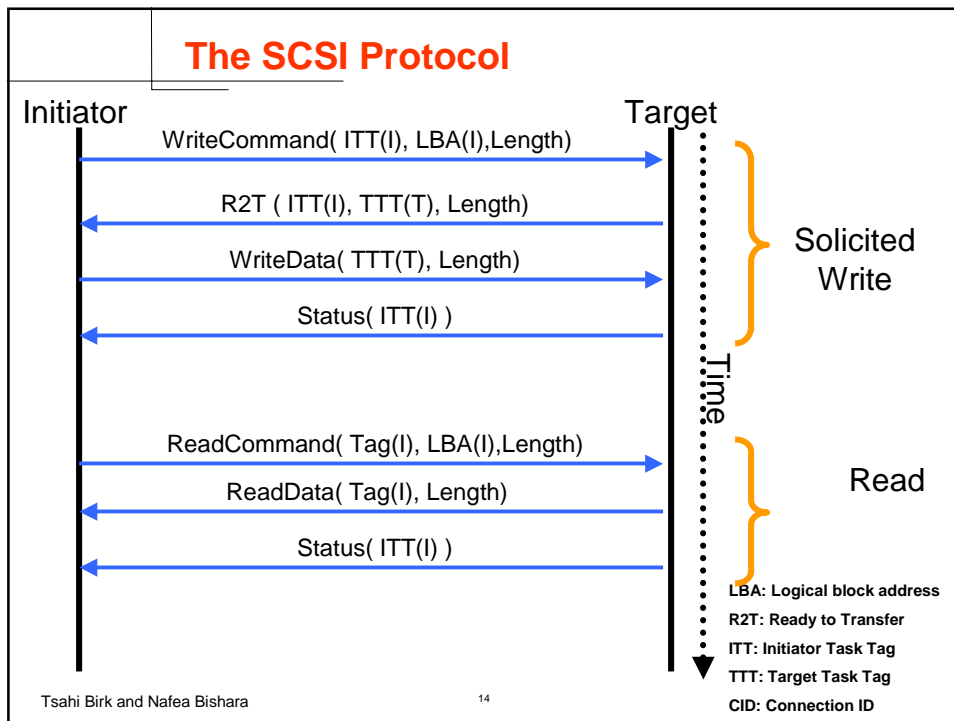
- Stay compliant with the SCSI-3 protocol (The de-facto block I/O protocol)
- Keep backward compatibility and comply with all SCSI/SAN software developed in the last two decades
 - support coexistence of devices supporting the new architecture and traditional devices in the same SAN

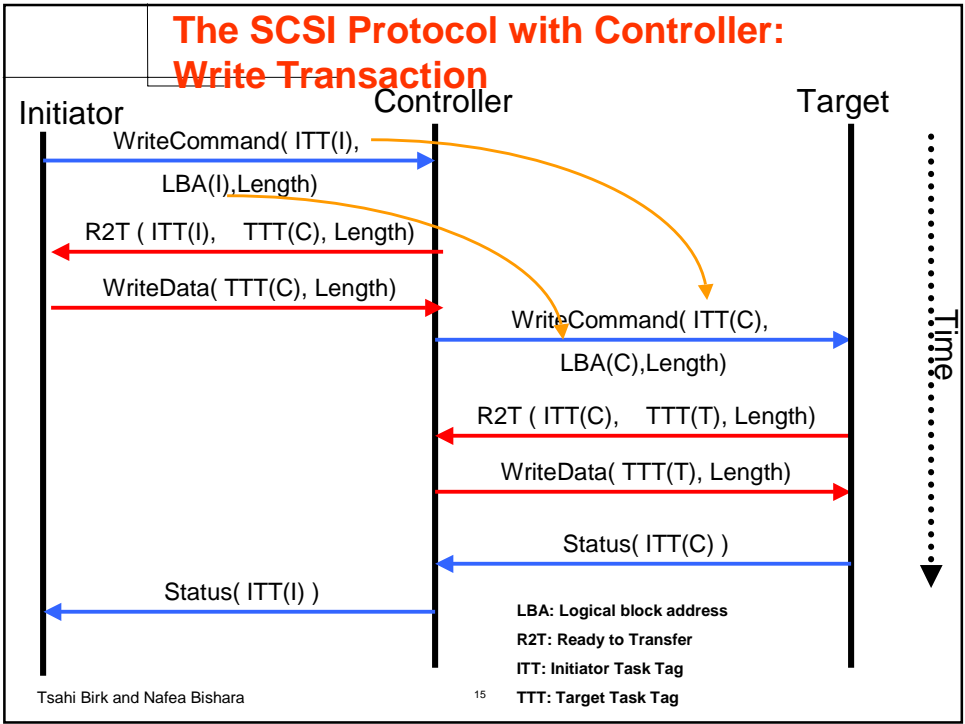
SCSI Protocol Suite

The SCSI-3 Layered Architecture



The SCSI Protocol





Distributed and Split Control-Data

Re-stating DSDC Objectives

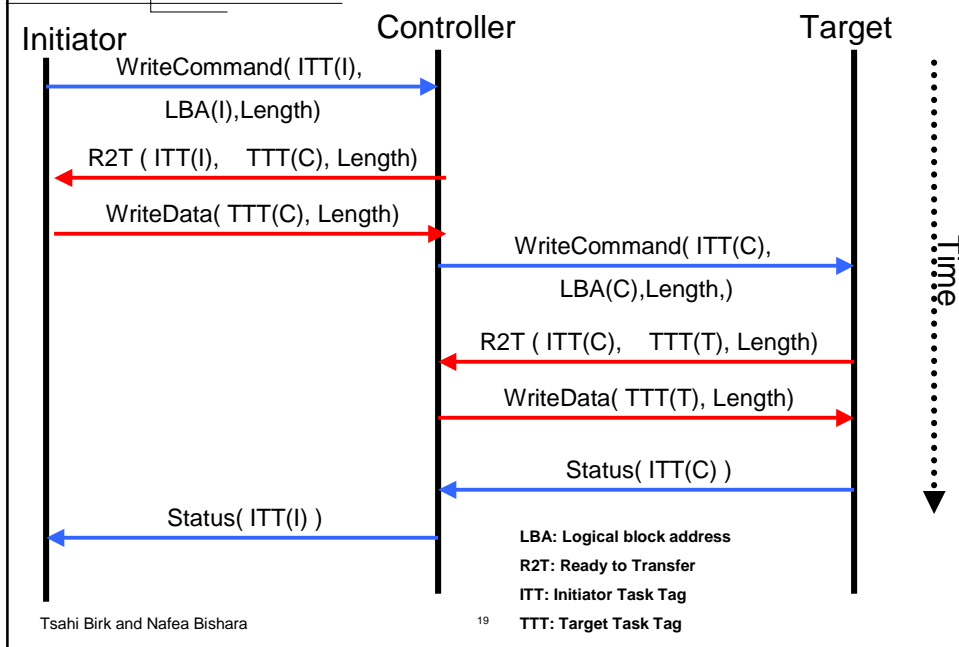
Define an architecture that would scale the SAN performance under the following constraints:

- Stay compliant with the SCSI model and application layer
 - To keep backward compatibility and the comply with all SCSI/SAN software developed in the last two decades
- Limit the changes to the SCSI transport protocol in the Initiator and Target
 - No change in the hardware and/or interconnect layer
 - No change to the SCSI application layer
 - Changes in the transport layer are limited to software/firmware changes.
- May require changes to the application layer in the controller
- Be generic enough to apply to (almost) any transport protocol

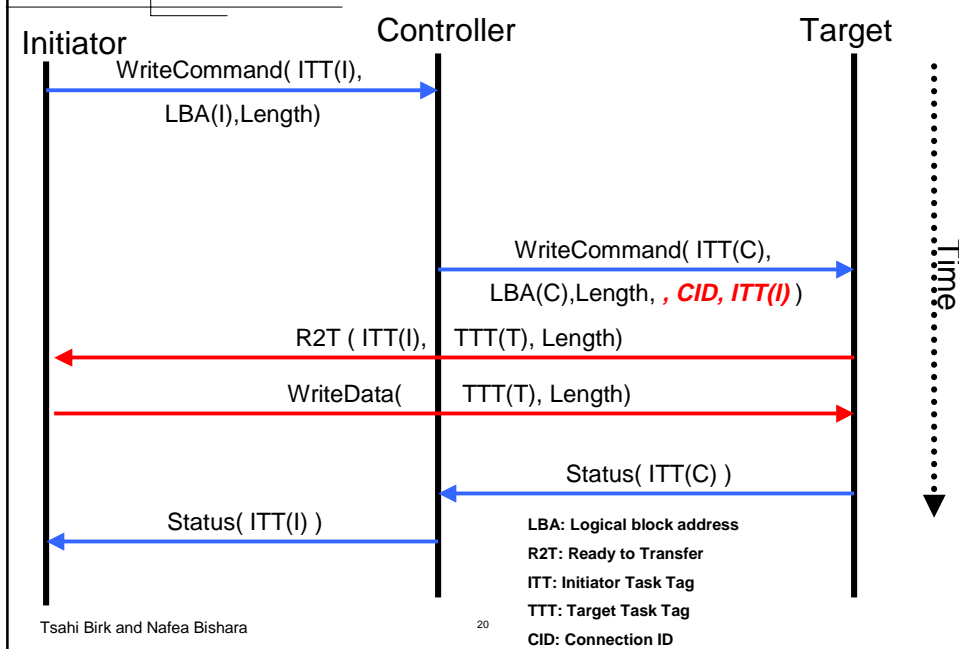
The two Principles for DSDC

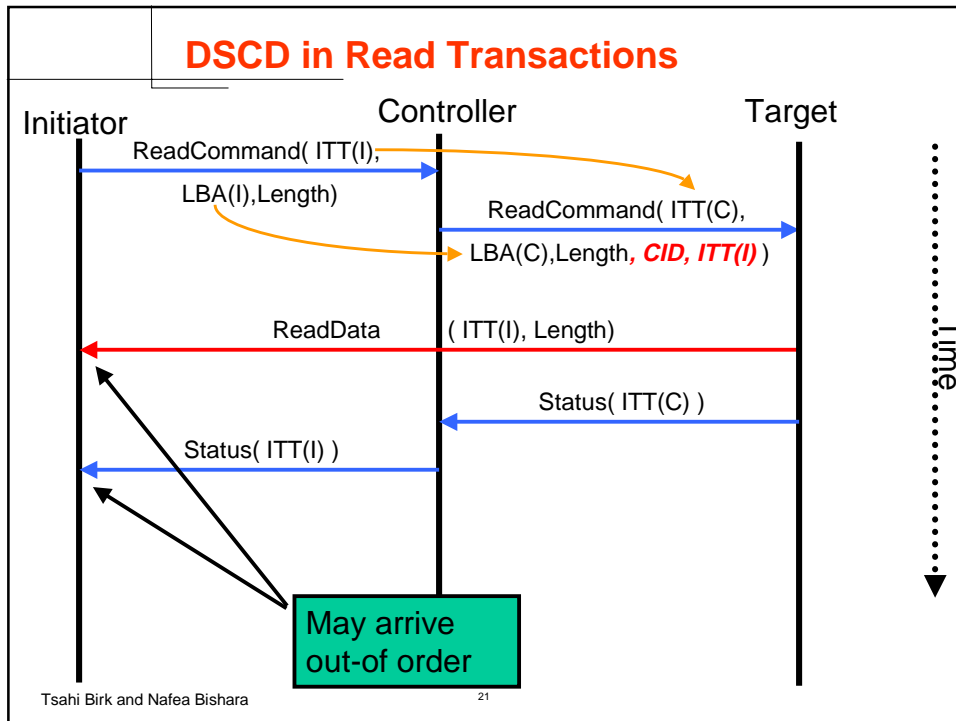
- Splitting control and data network connections
 - The key aspect to handle include: Ordering, Connection failure, Flow control....
- Direct inter-connect between Initiators and Targets for data transfer
 - Utilizing the inherent parallelism and full connectivity in Switched storage networks
 - Key challenges include: Authentication and Security, Abstraction of the target,....

Solicited Write Transaction in Traditional Controller



Solicited Write Transaction in DSDC Controller





- ### DSCD in READ Transactions: Complications
- Read Data and status run on different “connections”
 - Status: Target → Controller → Initiator
 - Data: Target → Initiator
 - May arrive in different order
 - Solution: The controller returns to the initiator the number of data transfers to expect per command (together with the status)
 - The initiator can use this to identify the end of the transaction
- Tsahi Birk and Nafea Bishara 22

Other SCSI Commands

- There are hundreds of other SCSI command
 - All but one (Unsolicited Write) do not transfer data
 - Unsolicited Write:
 - must go through the controller,
 - Solicited Write is the recommended command for SANs
- The handling of all commands except for Read and Solicited Write is unchanged by DSDC.

Other Issues in DSDC*

- Data Security and Authentication
 - We are not introducing any new threat
 - Simple extension to the authentication schemes showed in the Technical report
- Latencies
 - Are hardly affected and depend on the target vs. controller location relative to the initiator
 -
- RAID implementation in DSDC
 - Writing a Single block: Data traffic is cut by half on the Controller, and reduced by 25% in the network
 - Writing a Complete parity group: The DSDC does not save data transfers in the network nor through the controller.

(*) Discussed in details in the technical report in EE/Technion/Israel

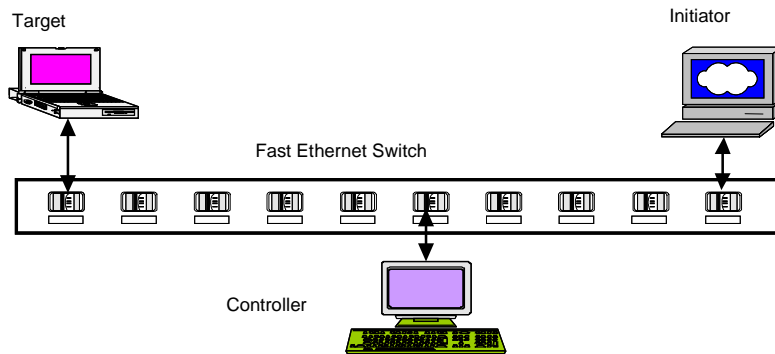
Other Issues in DSDC (Cont)

- Caching
 - Some controllers implement block caching in the controller itself
 - In DSDC, the caching can be implemented in the targets themselves
 - Normally, the targets are disk arrays with built-in cache
 - Smart caching schemes can be implemented by the DSDC controller to proactively fetch data for caching in its own memory.

Prototype Description

Testing Environment

- All computers running Linux Kernel 2.2.20
- iSCSI over TCP/IP over Ethernet
- The controller has two modes:
 - Traditional Controller mode
 - DSDC enabled Controller mode



Tsahi Birk and Nafea Bishara

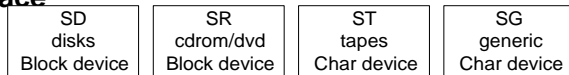
27

SCSI and iSCSI in Linux

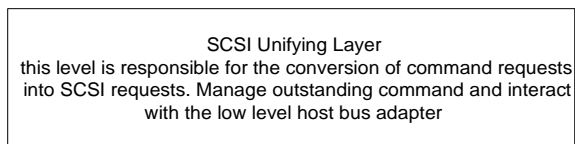
user space

kernel space

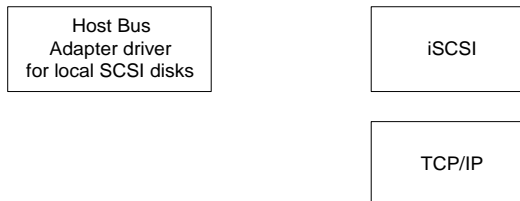
Upper Level



Mid Level

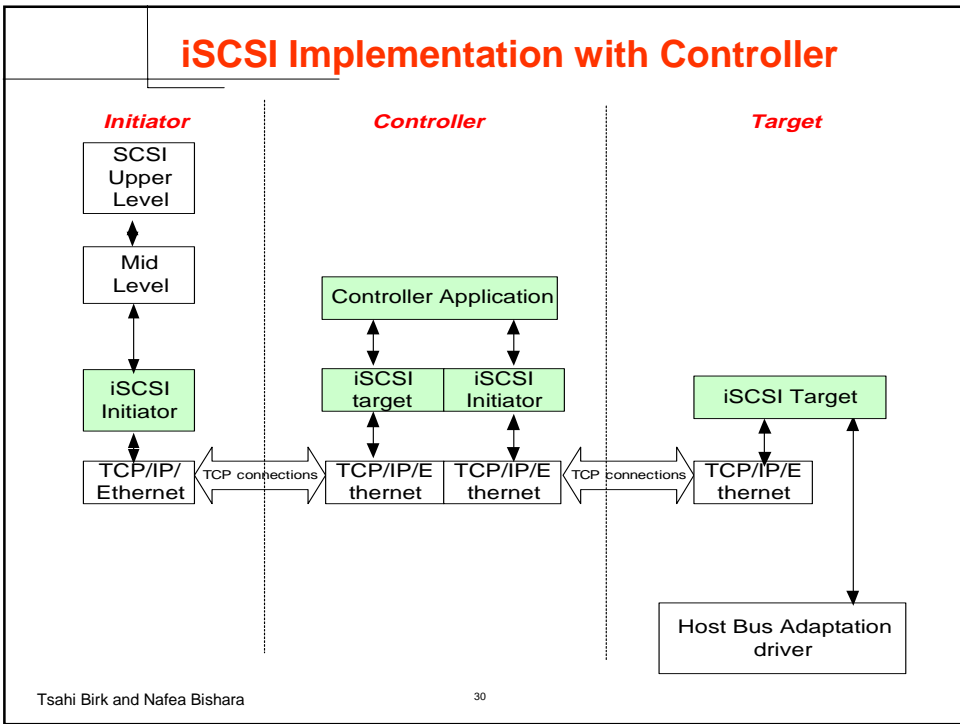
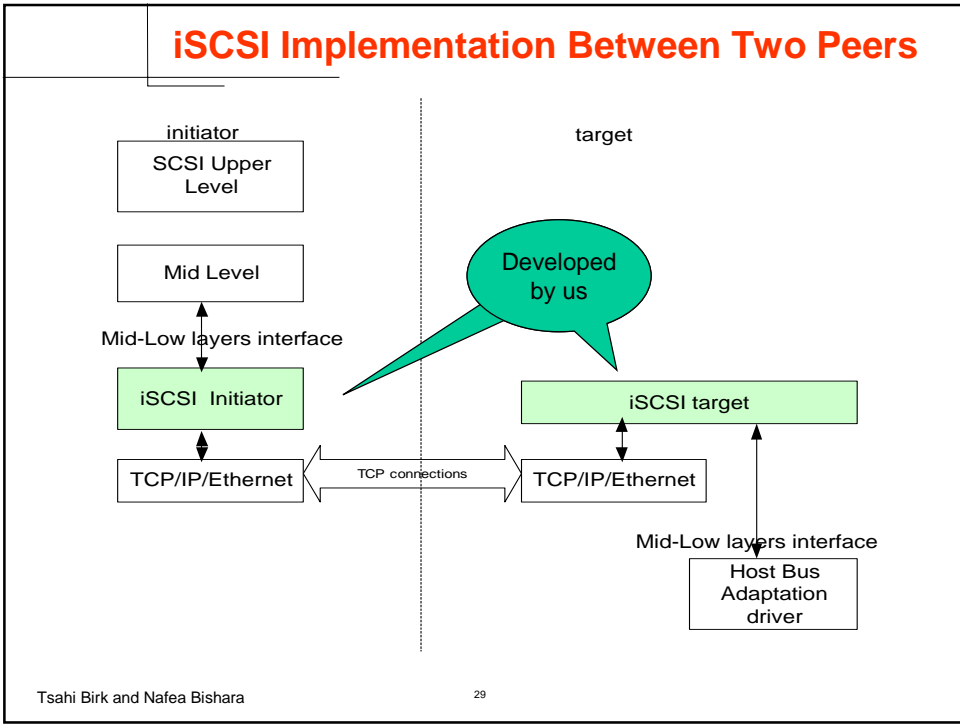


Low Level



Tsahi Birk and Nafea Bishara

28



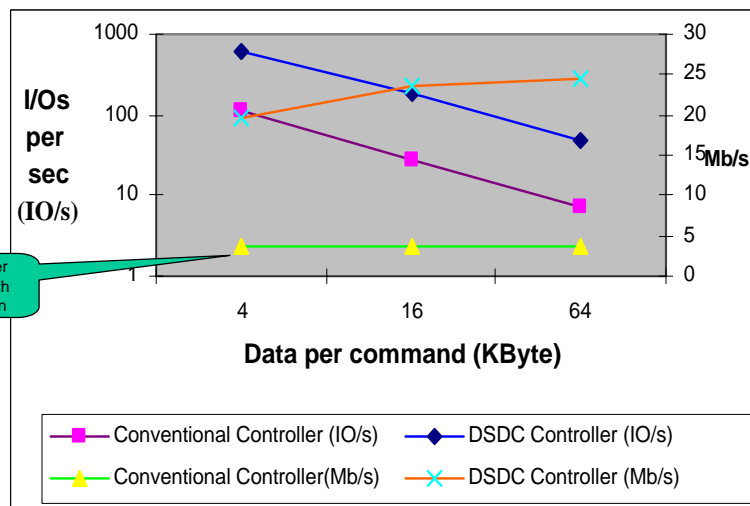
Performance Testing

- We worked with 1 Initiator, 1 Target and 1 Controller
 - Initiator and Target had 100Mbps Ethernet, to imitate a heavy load from initiators and a large group of targets
 - Controller had 10Mbps Ethernet
- We built our own SCSI testing utility that bypasses all file-system and buffers cache in Linux
- We test read and write sustained performance under various data lengths, in both modes: Traditional and DSDC-Enabled Controller

Tsahi Birk and Nafea Bishara

31

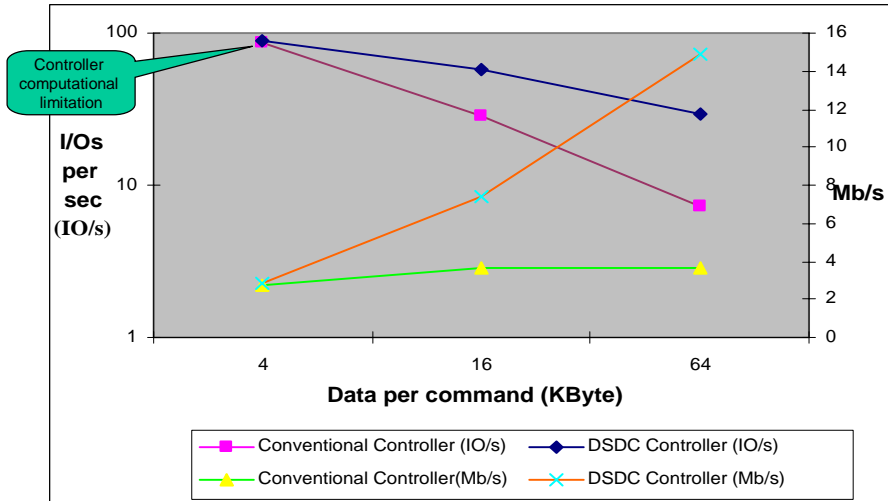
Performance Charts for READ



Tsahi Birk and Nafea Bishara

32

Performance Charts for WRITE



Tsahi Birk and Nafea Bishara

33

Summary: Advantages of DSDC

- Performance Scalability:
 - The total throughput of the SAN is not limited by the bandwidth of the connection between the controller and the SAN
- Centralized management
 - Since the controller is no longer the bandwidth bottleneck, it can handle more disk arrays, and saves the need to multiple controller in the SAN, simplifying management
- The changes are confined to the SCSI Transport layer in the Target/Initiator
 - The controller requires modification in the application layer
- The prototype demonstrate the correctness and the performance advantages

Tsahi Birk and Nafea Bishara

34

Thank You!
