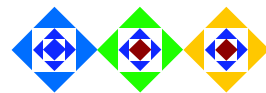


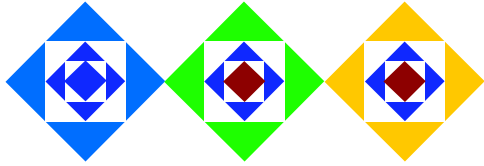
Design of the iSCSI Protocol

Kalman Meth

Julian Satran

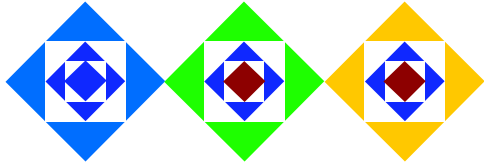
IBM Haifa Research Lab





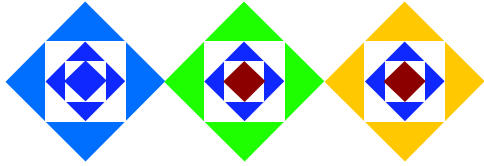
Overview

- What is iSCSI?
- Why TCP?
- Alternatives to TCP
- Drawbacks of TCP
- Data Transfer Model
- Data Placement
- Recovery



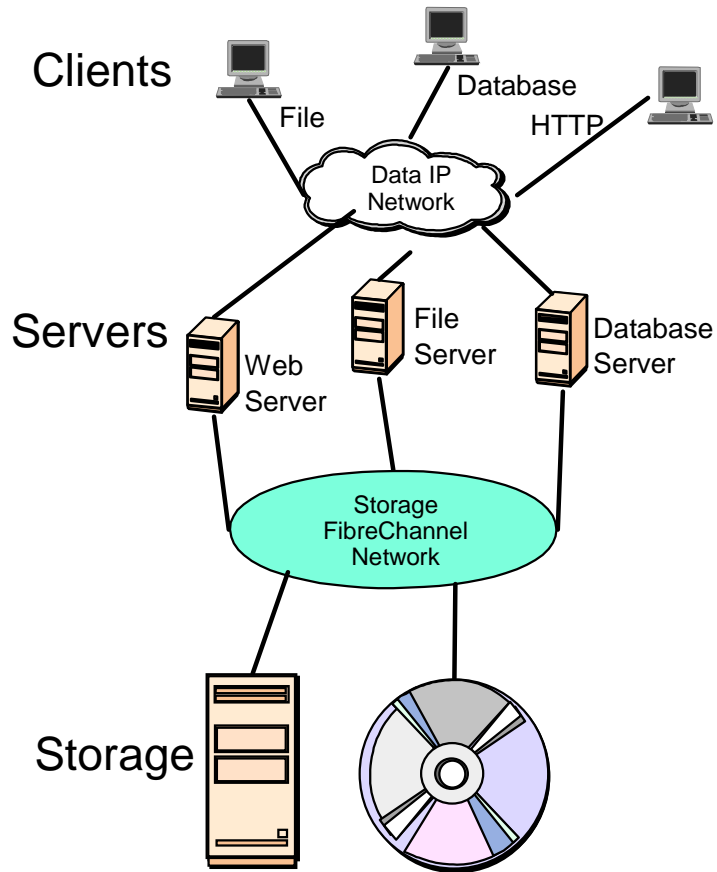
What is iSCSI?

- SCSI is a protocol for I/O devices such as disk, tape, CD ROM
- iSCSI = Internet SCSI = SCSI over TCP/IP
 - ▶ send SCSI commands over an IP network
- Related SCSI transport technologies
 - ▶ SCSI Fibre Channel Protocol (FCP)
 - ▶ Serial Storage Architecture (SSA)
 - ▶ Serial Bus Protocol (SBP)
 - ▶ SCSI over Infiniband?

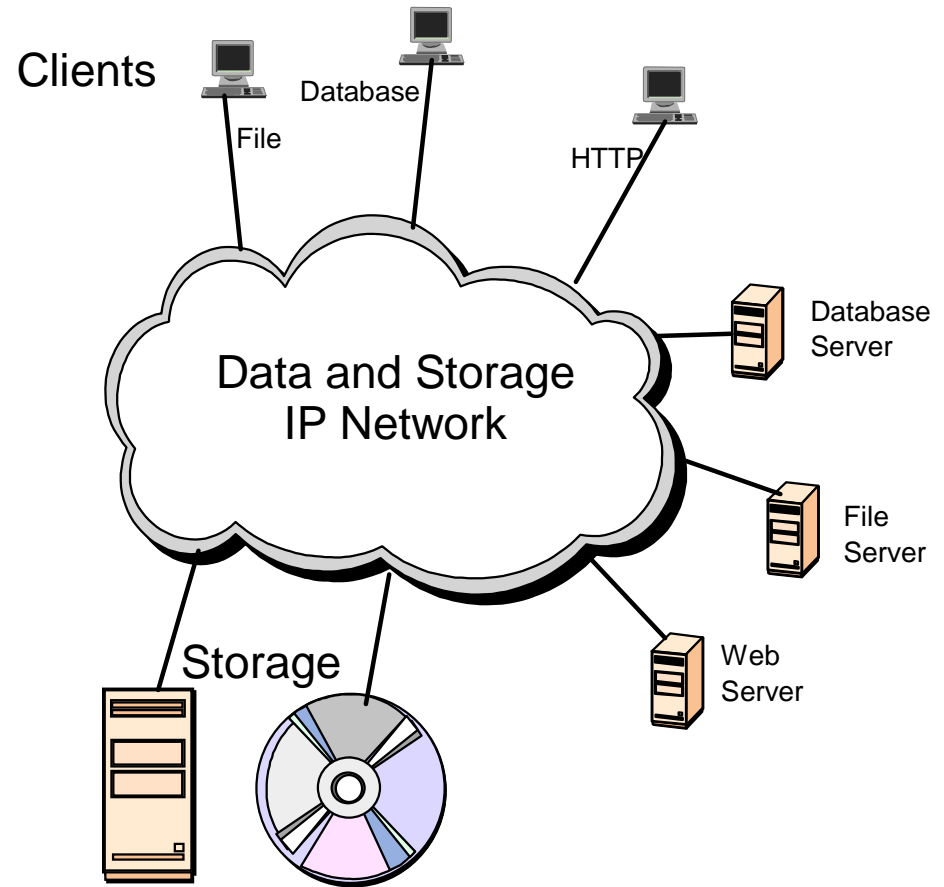


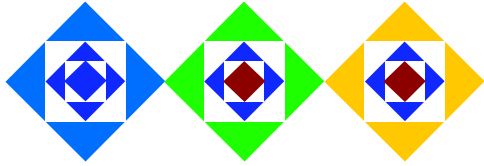
Classic SAN vs. iSCSI

Classic SAN



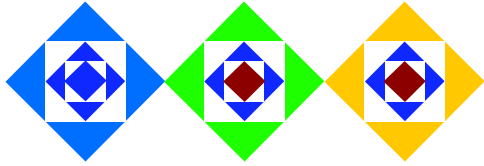
iSCSI





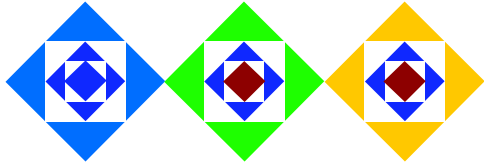
Layered Packet Format

Ethernet header (14)	IP header (20)	TCP header (20)	iSCSI header (48)	data ...
-------------------------	-------------------	--------------------	----------------------	----------



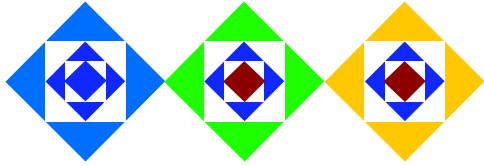
Why TCP?

- Reliable connection protocol
- Works over a variety of physical media
- Implemented on a wide variety of machines
- Field proven and scalable
- End-to-end connection model independent of the underlying network



SCSI over TCP Alternatives

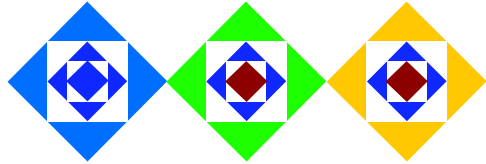
- SCSI over ...
 - ▶ Ethernet
 - ▶ IP
 - ▶ UDP
 - ▶ SCTP



Exploit features of TCP/IP

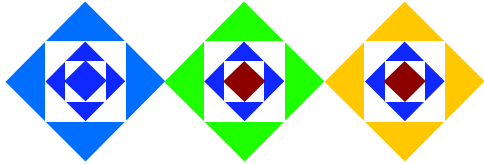
- TCP features
 - ▶ automatic acknowledgment
 - ▶ retransmission of lost and corrupted packets
 - ▶ guaranteed in-order delivery
 - ▶ congestion control

- IP-family features
 - ▶ IPSec (security)
 - ▶ SLP (discovery)
 - ▶ DHCP (configuration)



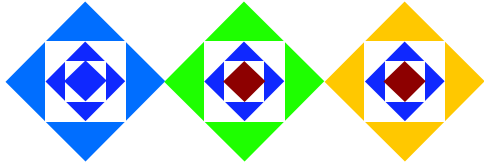
Drawbacks of using TCP

- Limited by TCP window size
 - ▶ cannot achieve maximum throughput on a single TCP connection
- Lost TCP packet causes delay in delivery of subsequent packets
 - ▶ if lose TCP packet, don't know where to find next iSCSI header(s)
- TCP checksum not sufficient for storage data integrity
- TCP usually entails multiple copying of data

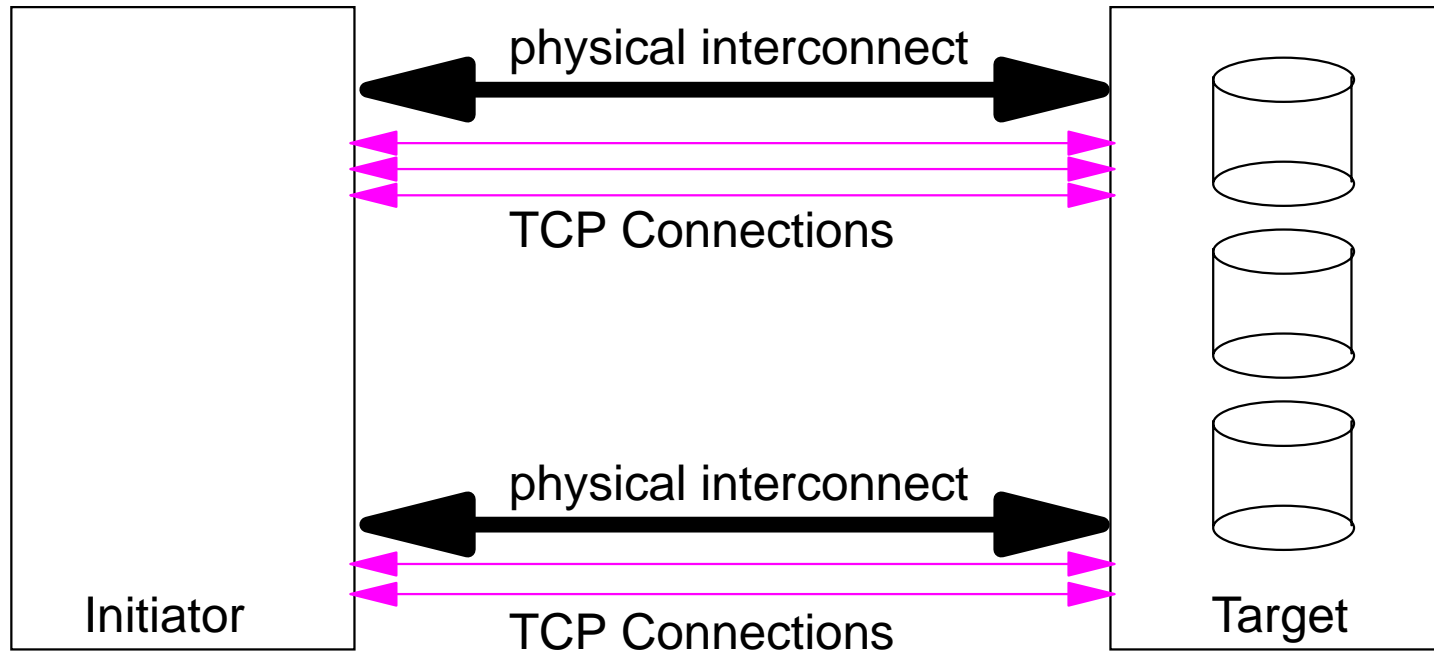


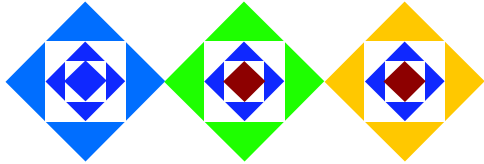
Sessions

- Collection of TCP connections between an Initiator and a Target
 - ▶ overcome bandwidth limitations imposed by TCP window size
 - ▶ utilize multiple CPUs in an SMP
- Connections of a session may traverse different physical interconnects
 - ▶ aggregate bandwidth from multiple interconnects
- Must now coordinate between multiple TCP connections



Sessions (cont.)

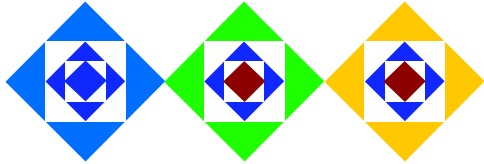




Data Transfer Model

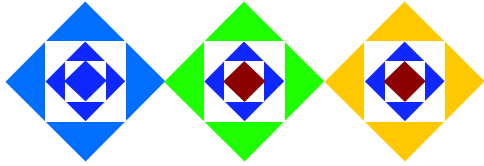
- Asymmetric
 - ▶ single control channel
 - ▶ multiple data channels
 - ▶ control channel used to transfer commands, status, task management

- Symmetric
 - ▶ all channels identical
 - ▶ send data and status over same channel as corresponding command



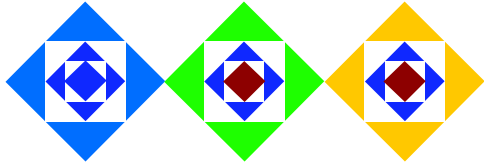
Data Transfer Model (*continued*)

- Advantages of Asymmetric model
 - ▶ no backlog of data to block control channel
 - ▶ Task Management operation can always be timely delivered
- Advantages of Symmetric model
 - ▶ iSCSI adapter can be self-contained
 - ▶ no need to transfer command between adapters
 - ▶ simpler software implementations

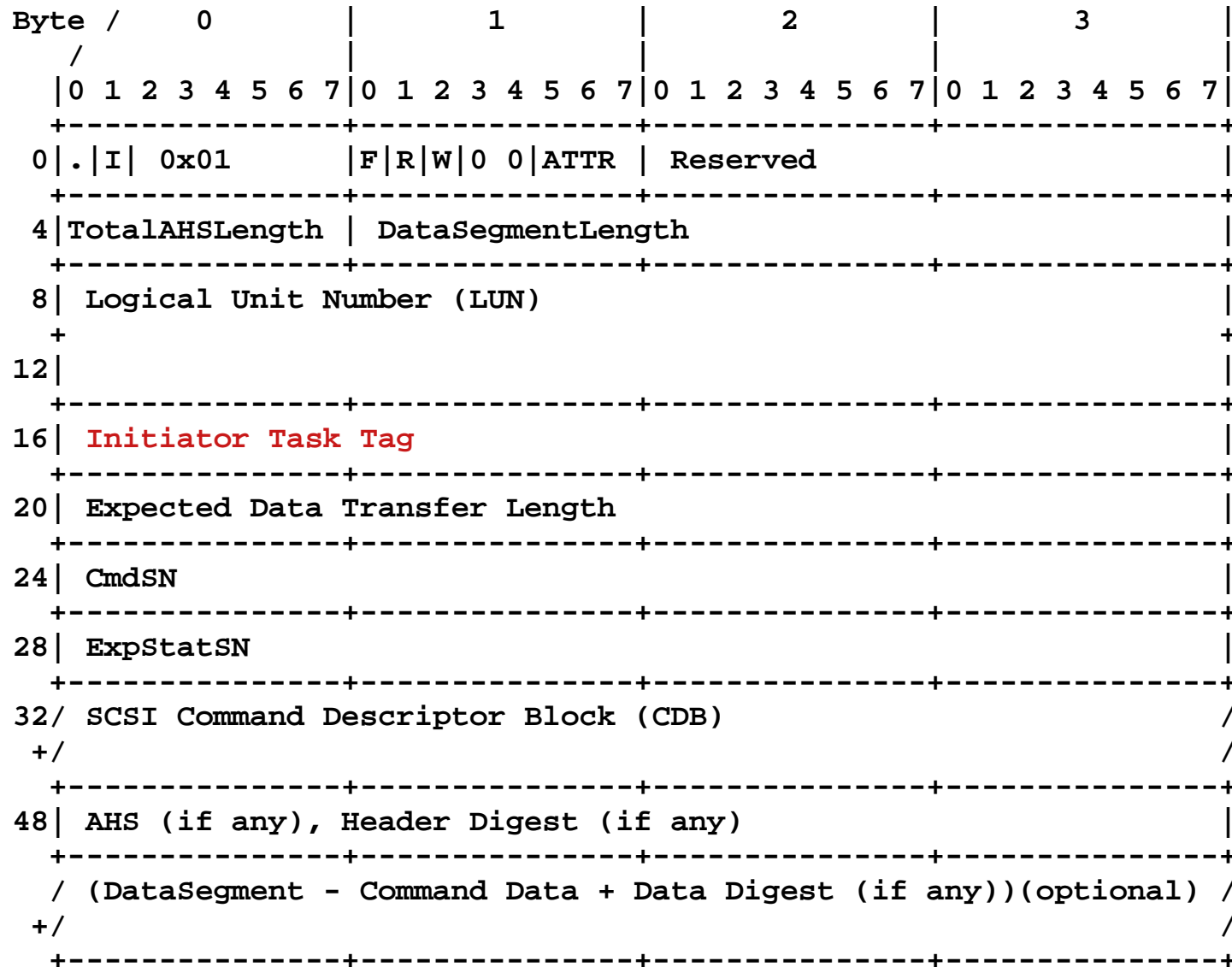


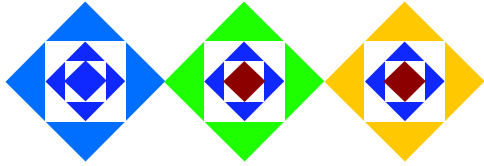
RDMA descriptors

- iSCSI Task Tags can be RDMA descriptors
 - ▶ used together with offset and length fields
- Initiator Task Tag
 - ▶ provided in Command PDUs
 - ▶ copied to Data-In PDUs
- Target Task Tag
 - ▶ provided on R2T
 - ▶ copied to Data-out PDUs



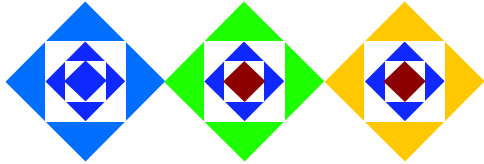
SCSI Command PDU





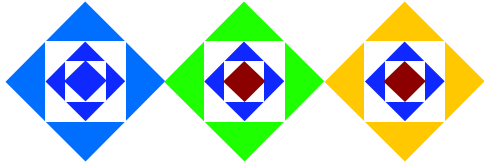
Data-in PDU

Byte /	0	1	2	3
/				
	0 1 2 3 4 5 6 7	0 1 2 3 4 5 6 7	0 1 2 3 4 5 6 7	0 1 2 3 4 5 6 7
0	. . 0x25	F A 0 0 0 O U S	Reserved	Status or Rsvd
4	TotalAHSLength		DataSegmentLength	
8	LUN or Reserved			
12				
16	Initiator Task Tag			
20	Target Transfer Tag or 0xffffffff			
24	StatSN or Reserved			
28	ExpCmdSN			
32	MaxCmdSN			
36	DataSN			
40	Buffer Offset			
44	Residual Count			
48				



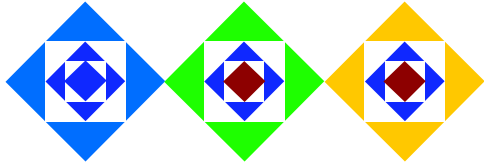
Out of Order Data Placement

- TCP delivers data in order
- If packet is dropped (at 10 Gb) or have digest error, have big data backlog
 - ▶ either store data on adapter (100s of MBs)
 - ▶ save data in temporary host memory and copy
 - ▶ drop data after missing packet
- Use markers (or framing) to find next iSCSI PDU
- Place data (from next PDU) in memory
 - ▶ don't yet inform application of data arrival
 - ▶ preserve TCP ordering semantics



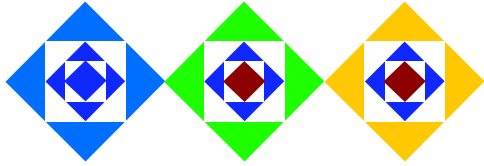
Framing

- Can we tell from TCP packet where next iSCSI/ULP (Upper Level Protocol) packet begins?
- Needed when a packet is dropped or corrupted to jump to next iSCSI/ULP packet
- IETF Working Group looking into problem
- No agreed upon mechanism yet



iSCSI Recovery

- Main reasons for iSCSI-level recovery
 - ▶ TCP connections occasionally break
 - maintain session across new connection
 - ▶ Digest errors
- Critical for long distance and tape operations
 - ▶ do not want to restart a large data transfer due to a transient TCP problem
- Levels of Recovery
 - ▶ Session Recovery (required)
 - ▶ Connection Recovery
 - ▶ Recovery within connection
 - ▶ Recovery within command



Summary

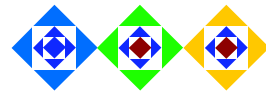
- iSCSI leverages existing features of TCP and the IP family of protocols
- iSCSI was designed with features to overcome TCP limitations
 - ▶ sessions with multiple connections
 - ▶ CRC digests
 - ▶ possible out of order data placement
- Multiple recovery options for different environments

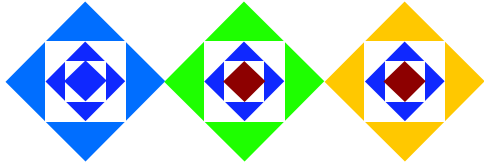
Design of the iSCSI Protocol

Kalman Meth

Julian Satran

IBM Haifa Research Lab





Separate or Common Network

