EMC²
where information lives®

iSCSI: Protocol and Functionality

David L. Black, Ph.D.

© Copyright 2006 EMC Corporation. All rights reserved.

Disclaimer – Roadmap Information

During this presentation EMC may refer to new features, capabilities, and possible release dates.

EMC (or appropriate division) makes no representation and undertakes no obligations with regard to product planning information, anticipated product characteristics, performance specifications, or anticipated release dates (collectively, "Roadmap Information"). Roadmap Information is provided by EMC as an accommodation to the recipient solely for purposes of discussion and without intending to be bound thereby.

© Copyright 2006 EMC Corporation. All rights reserved.

Session Goals

- Explain what iSCSI is
  - And the structure of the iSCSI protocol stack
- Explain how iSCSI provides storage access
  - And how it fits into storage and network infrastructure
- Explain how an iSCSI session is established
  - Plus cover security, boot, multipathing, etc.

NOTE: This is a technology session

Product specifics are covered in other sessions

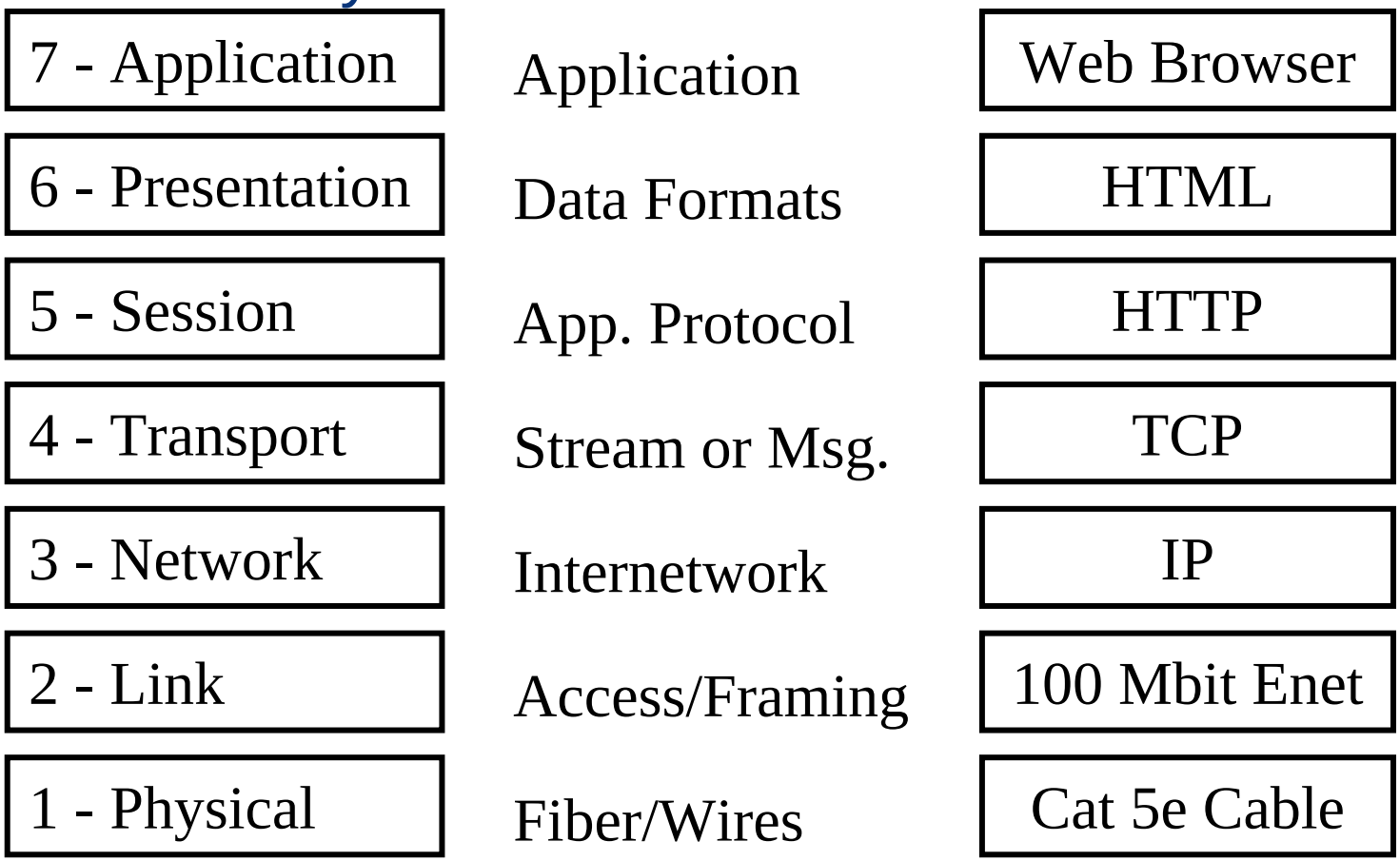
© Copyright 2006 EMC Corporation. All rights reserved.

Introduction

- What is iSCSI?
  - Internet Small Computer Systems Interface
  - SCSI storage access over TCP/IP networks
- Why is iSCSI interesting?
  - Reuse existing IP infrastructure and skills
  - IP protocols have better interoperability track records
  - Security can be better in IP networks
- iSCSI reuses networking and storage concepts
  - Next few slides: Review important basic concepts

© Copyright 2006 EMC Corporation. All rights reserved.

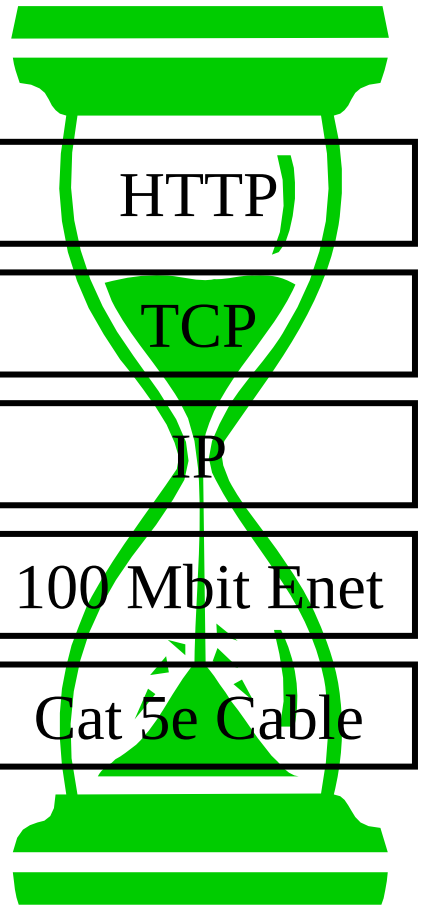
IP Network Layers



© Copyright 2006 EMC Corporation. All rights reserved.

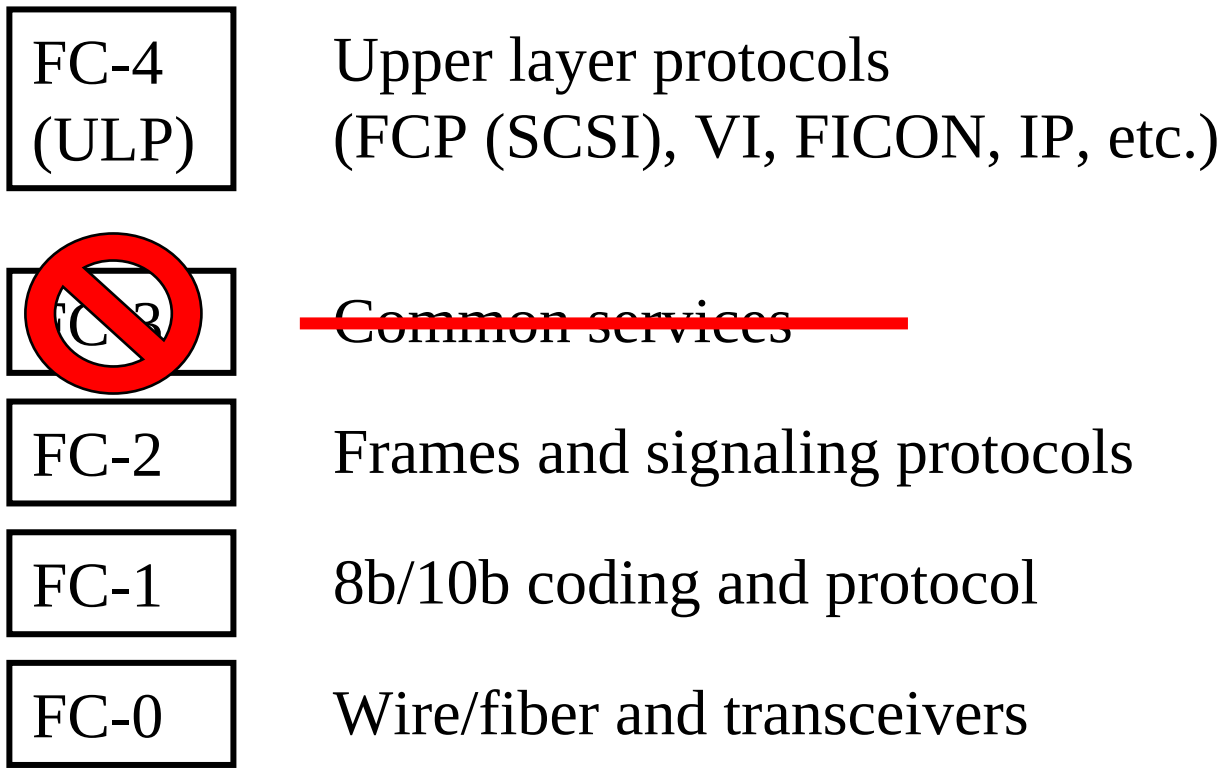
IP Network Layers - In Practice

7 - Application	Application	Web/HTML
5 - Session	App. Protocol	HTTP
4 - Transport	Stream or Msg.	TCP
3 - Network	Internetwork	IP
2 - Link	Access/Framing	100 Mbit Enet
1 - Physical	Fiber/Wires	Cat 5e Cable



© Copyright 2006 EMC Corporation. All rights reserved.

Fibre Channel Layers



© Copyright 2006 EMC Corporation. All rights reserved.

SCSI Concepts

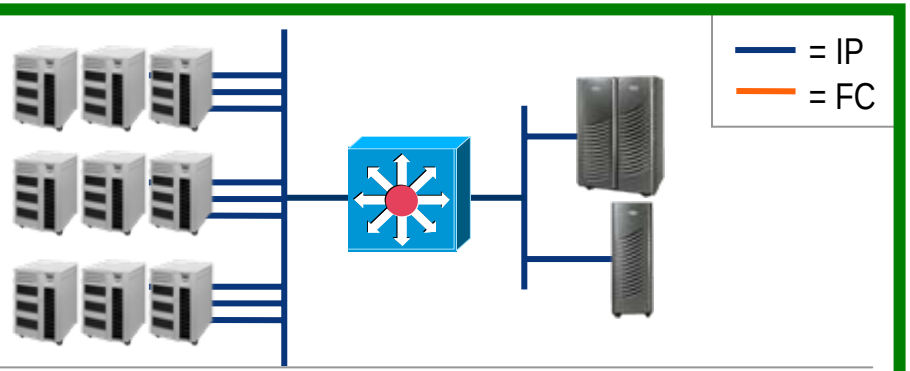
- *Initiator* connects to *Target*
  - Host connects to storage device
- *Target* exports *Logical Units*
  - Storage device exports volumes
- *Logical Units* have *Logical Unit Numbers* (LUNs)
  - Numbering is per target
  - Same LU may have different LUNs at different targets
- Active discovery
  - SCSI “Bus Walker” finds accessible targets

© Copyright 2006 EMC Corporation. All rights reserved.

IP Storage Network Scenarios and Protocols

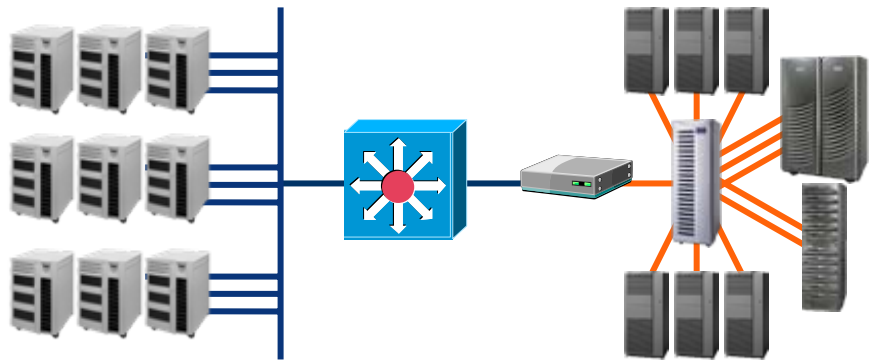
Native

- All Ethernet (no Fibre Channel)
- iSCSI protocol
- Standard Ethernet switches and routers



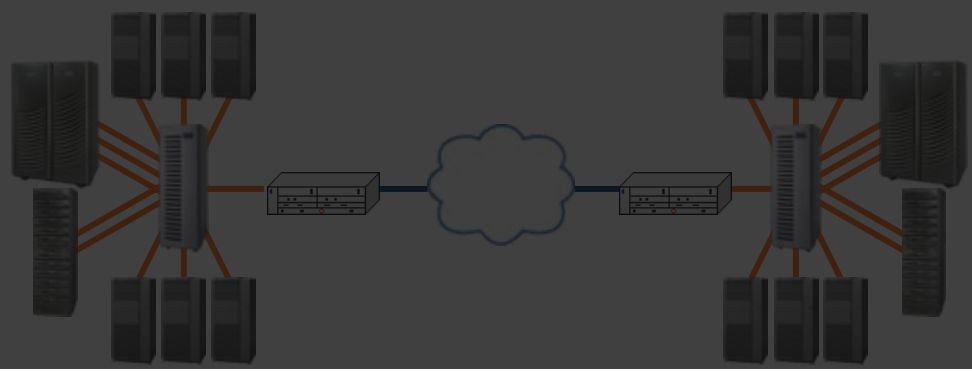
Bridging

- Servers Ethernet attached
- Storage Fibre Channel attached
- iSCSI protocol



Extension

- Servers and storage on SAN
- FCIP or iFCP protocol
- Host-to-storage or replication



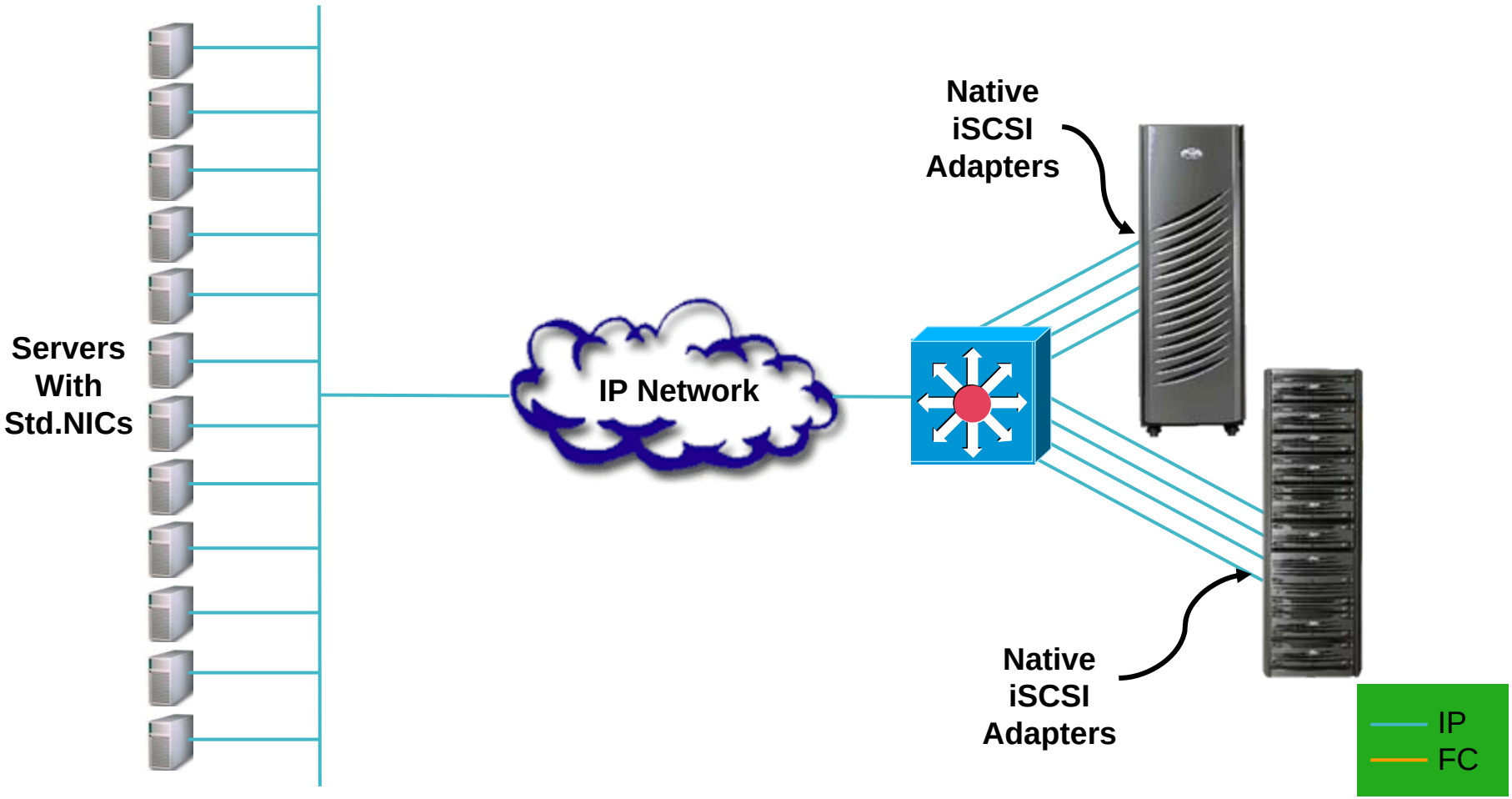
© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Overview

- Internet Small Computer Systems Interface
- Provides storage access over TCP/IP networks
  - Maps SCSI functionality to TCP/IP protocol
  - Similar to mapping SCSI over Fibre Channel (FCP)
- Network Protocol
  - Peer to HTTP, NFS, FTP, Telnet, etc. (uses TCP)
- Can be used with existing IP & Ethernet networks
  - NICs, switches, routers, etc.

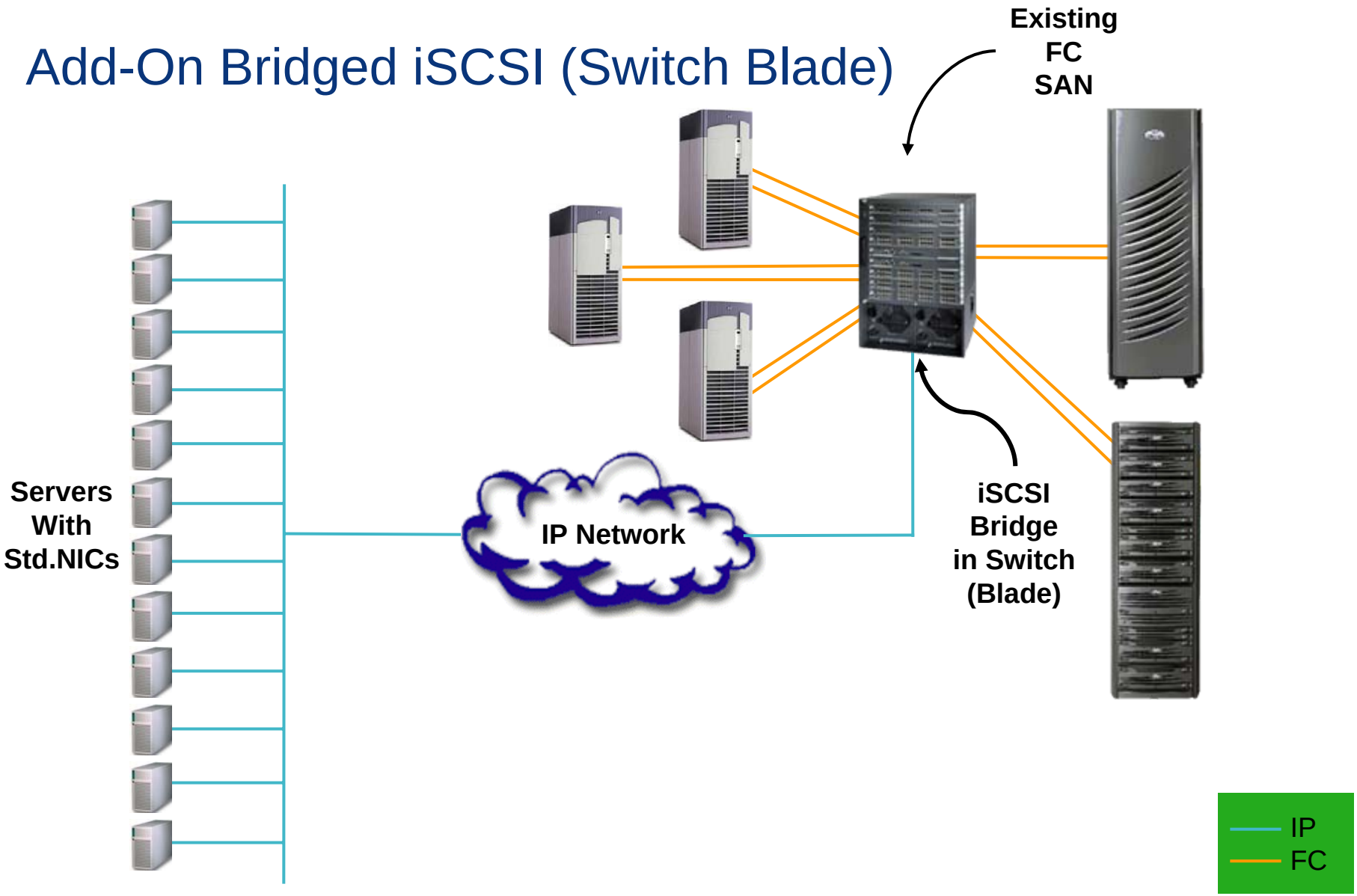
© Copyright 2006 EMC Corporation. All rights reserved.

Dedicated Native iSCSI



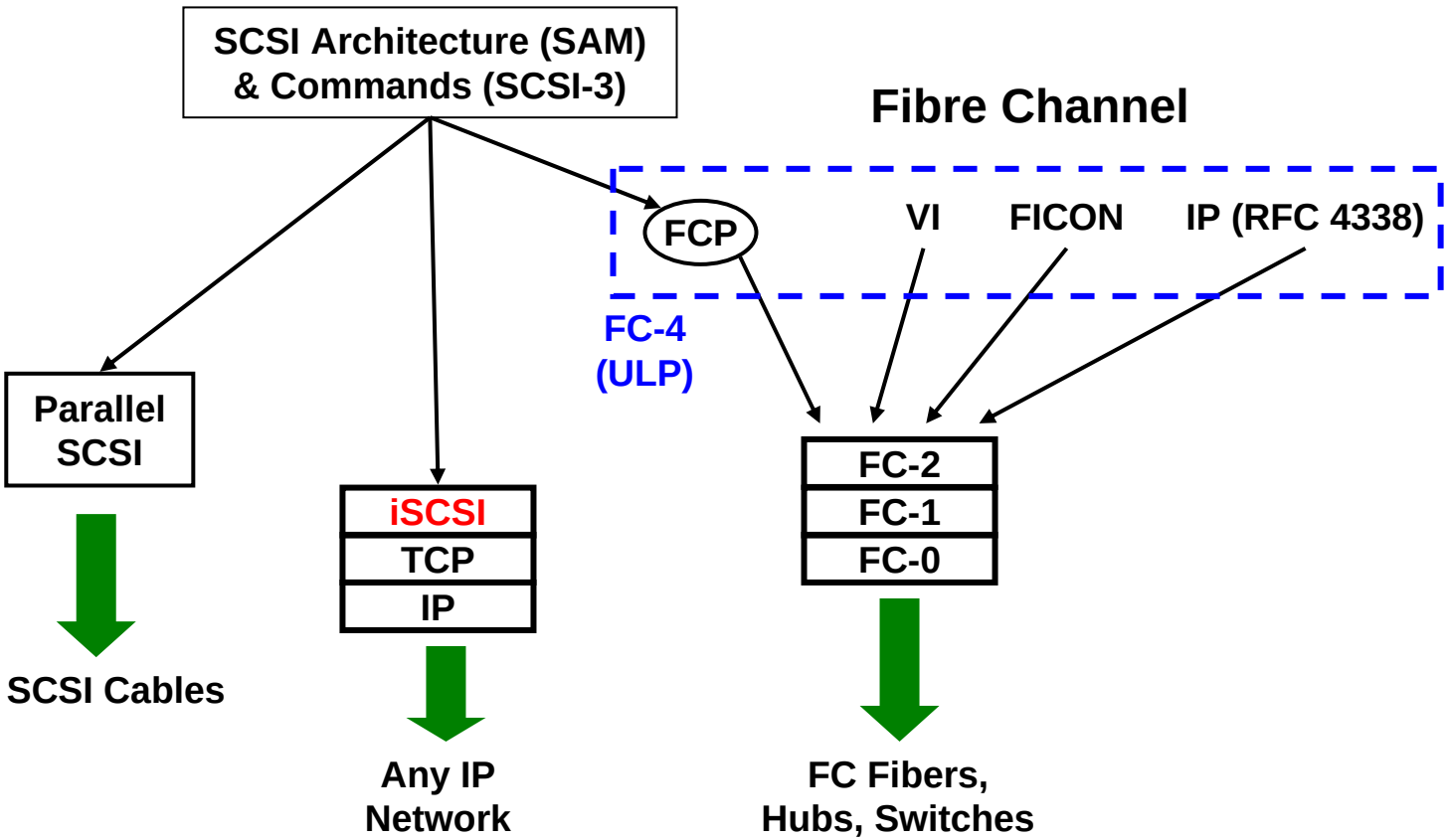
© Copyright 2006 EMC Corporation. All rights reserved.

Add-On Bridged iSCSI (Switch Blade)



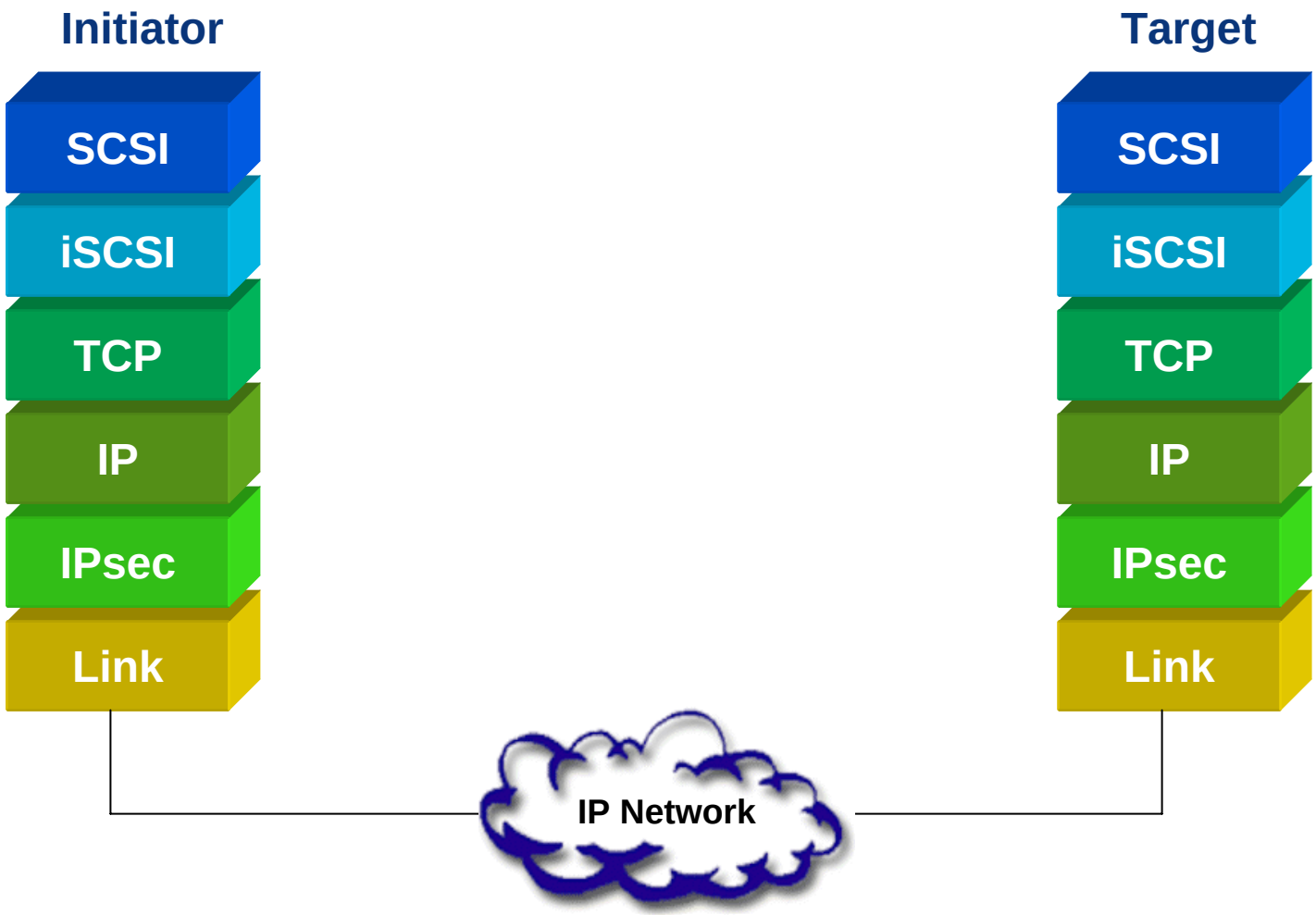
© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Relationship to Other SCSI Protocols



© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Protocol Stack



© Copyright 2006 EMC Corporation. All rights reserved.

Data Encapsulation Into Network Packets

Delivery of iSCSI Protocol Data Unit (PDU) for SCSI functionality (initiator, target, data read/write, etc.)

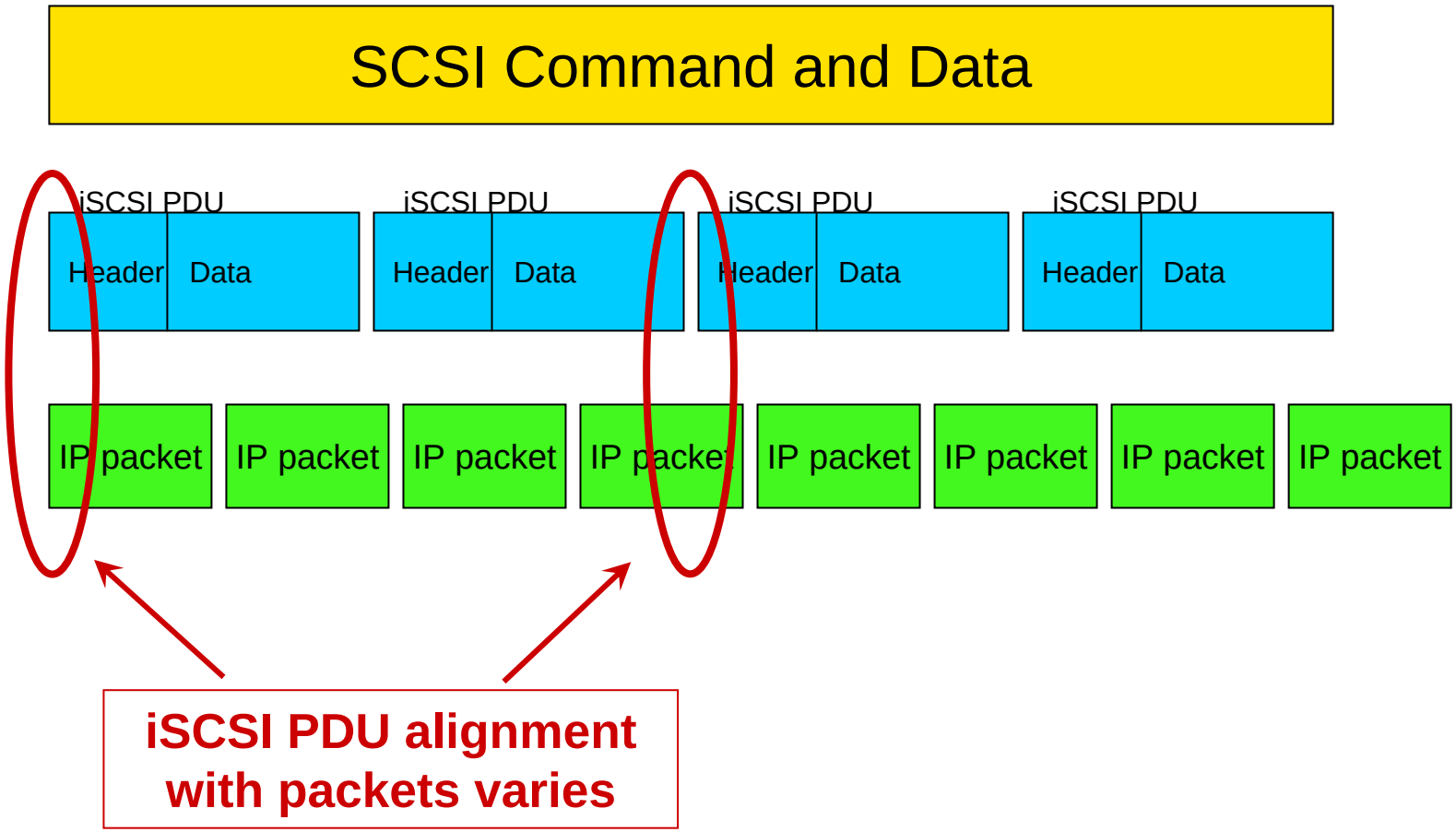
Reliable data transport and delivery (TCP Windows, ACKs, ordering, etc.) Also demux within node (port numbers)

Provides IP "routing" capability so that packet can find its way through the network

Provides physical network capability (Cat 5, MAC, etc.)

© Copyright 2006 EMC Corporation. All rights reserved.

SCSI to iSCSI Mapping



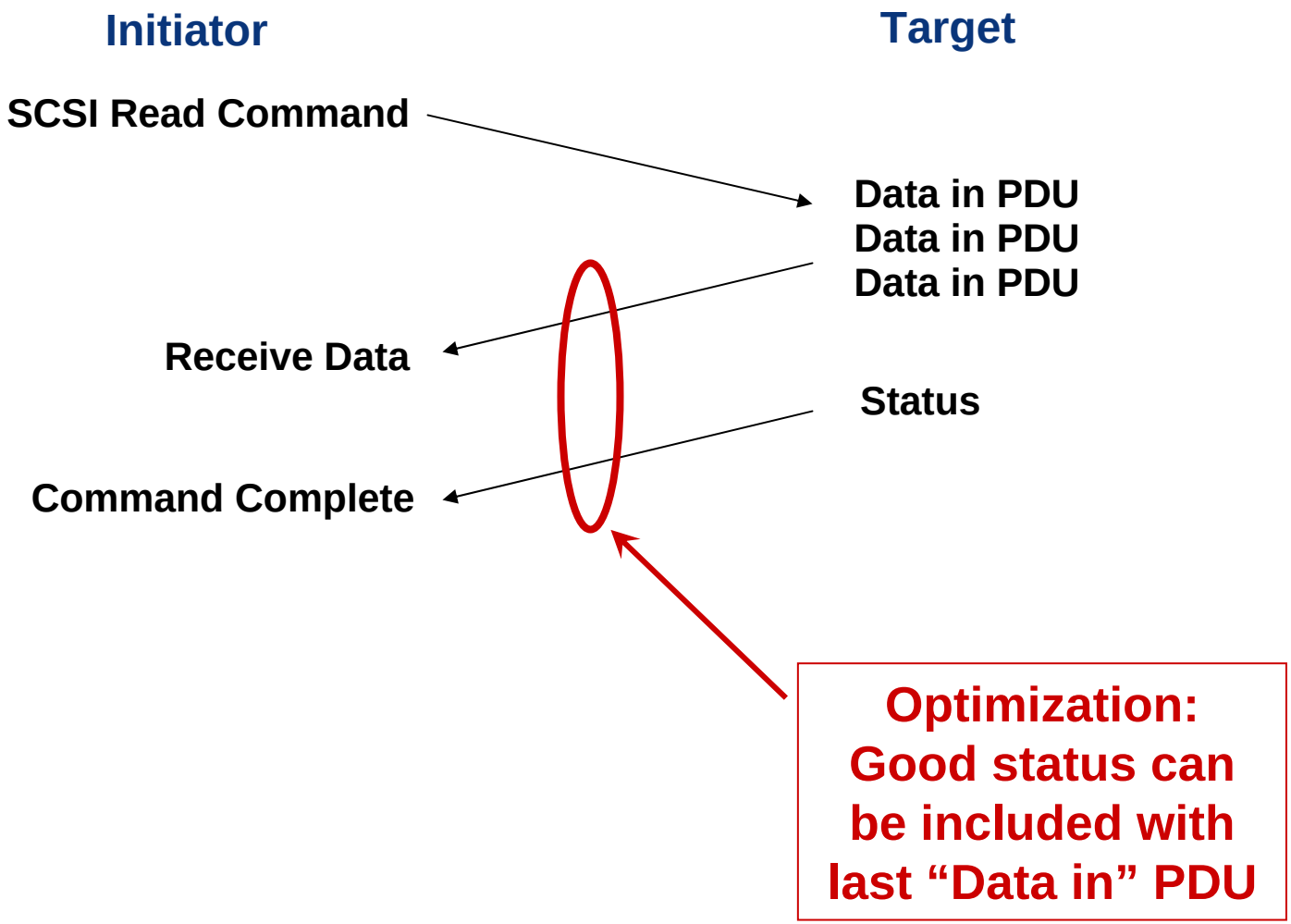
© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Concepts

- iSCSI Session: One Initiator and one Target
  - Multiple TCP connections allowed in a session
    - Exploit network parallelism
    - Error recovery possible across connections
- Most communication is based on SCSI
  - e.g., Ready to Transmit (R2T) for target flow control
- Important iSCSI additions to SCSI
  - Immediate and unsolicited data to avoid roundtrip
  - Login phase for connection setup
    - Text-based parameter negotiation
  - Explicit logout for clean teardown

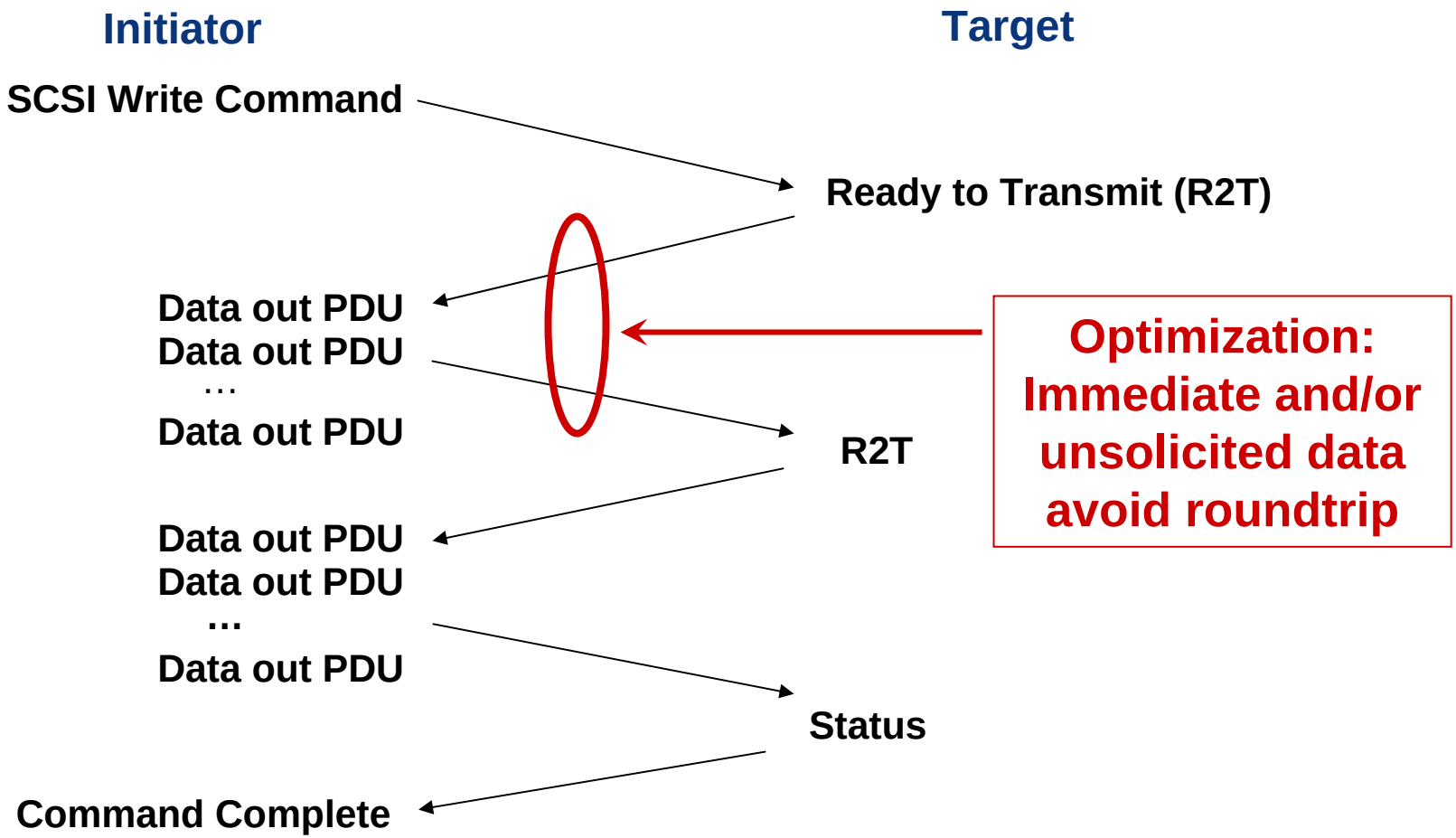
© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Read Example



© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Write Example



© Copyright 2006 EMC Corporation. All rights reserved.

Establishing an iSCSI Session

- Naming: Identify storage to access (target) *[What]*
  - Also identify initiator that wants to access storage
  - Naming is location-independent (unlike Fibre Channel)
- Discovery: Find storage to access *[Where]*
  - SCSI “Bus Walker” doesn’t scale to IP networks
- Login: Establish connections to storage *[How]*
  - Parameter negotiation prior to reads/writes
  - Login occurs on each TCP connection

© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Naming [What]

- Design rationale
  - Targets may share <IP address, TCP port>
  - Initiators and targets may have multiple IP addresses
  - Unique names are important for third-party commands
- iSCSI names: Globally unique
  - EUI-based (type of WWN)
    - eui.5006048dc7dfb1af
  - IQN: Reversed hostname (DNS) as naming authority
    - iqn.1991-05.com.microsoft:WindowsSystem1
  - NAA-based (more WWNs, including long WWNs)
    - naa.62004567BA64678D0123456789ABCDEF
- Intended usage: One iSCSI name per host
  - Regardless of the number of interfaces (NICs/HBAs)

© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Discovery [Where]

- SCSI discovery paradigm
  - "Bus Walker" looks for targets
  - Exhaustive search doesn't work in IP networks
- iSCSI discovery mechanisms
  - Small scale: Static configuration and SendTargets
    - Simple configuration mechanisms
  - Intermediate scale: SLP
    - Based on multicast or simple directory agent
  - Large scale: iSNS
    - Rich name service, similar to services provided by FC fabric
    - SLP can be used to discover iSNS Server

© Copyright 2006 EMC Corporation. All rights reserved.

Static Configuration and SendTargets

- Static Configuration: Tell initiator about target(s)
  - iSCSI target name and location (e.g., IP address, port)
    - iSCSI default TCP port: 3260
  - Simple mechanism, does not scale well
    - Especially if information is entered manually
- SendTargets command: Better scaling
  - Initiator issues SendTargets
  - Target responds with iSCSI names of targets
    - Also IP addresses and TCP ports if they differ
  - Moves most configuration from initiator to target
    - Only have to tell initiator an address of target system
    - Target provides the rest of the information

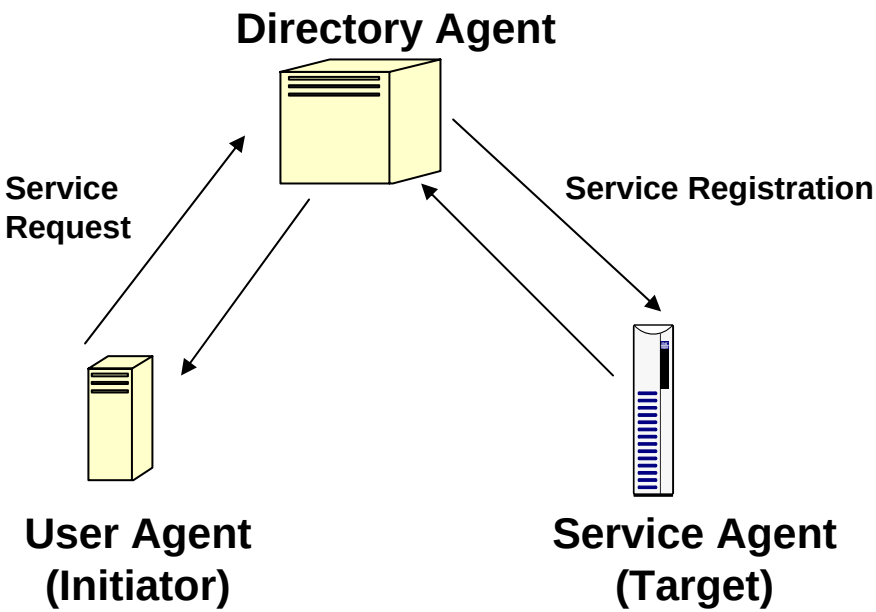
© Copyright 2006 EMC Corporation. All rights reserved.

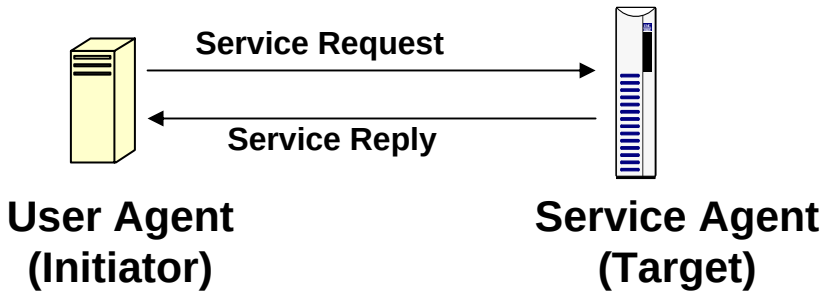
SLP (Service Location Protocol)

- Major SLP components
  - User Agent (UA) – Find services to use
  - Service Agent (SA) – Advertise services for use
  - Directory Agent (DA) – Connect users to services
- SLP function for iSCSI
  - Target advertises name:IP address:port
    - Either to DA in the network or on its own
  - Initiator contacts DA for target information
    - If no DA configured, use multicast to find targets
    - DA usage recommended if multicast is restricted
  - iSCSI template identifies iSCSI services in SLP

© Copyright 2006 EMC Corporation. All rights reserved.

SLP Structures





Directory Agent structure has better scalability

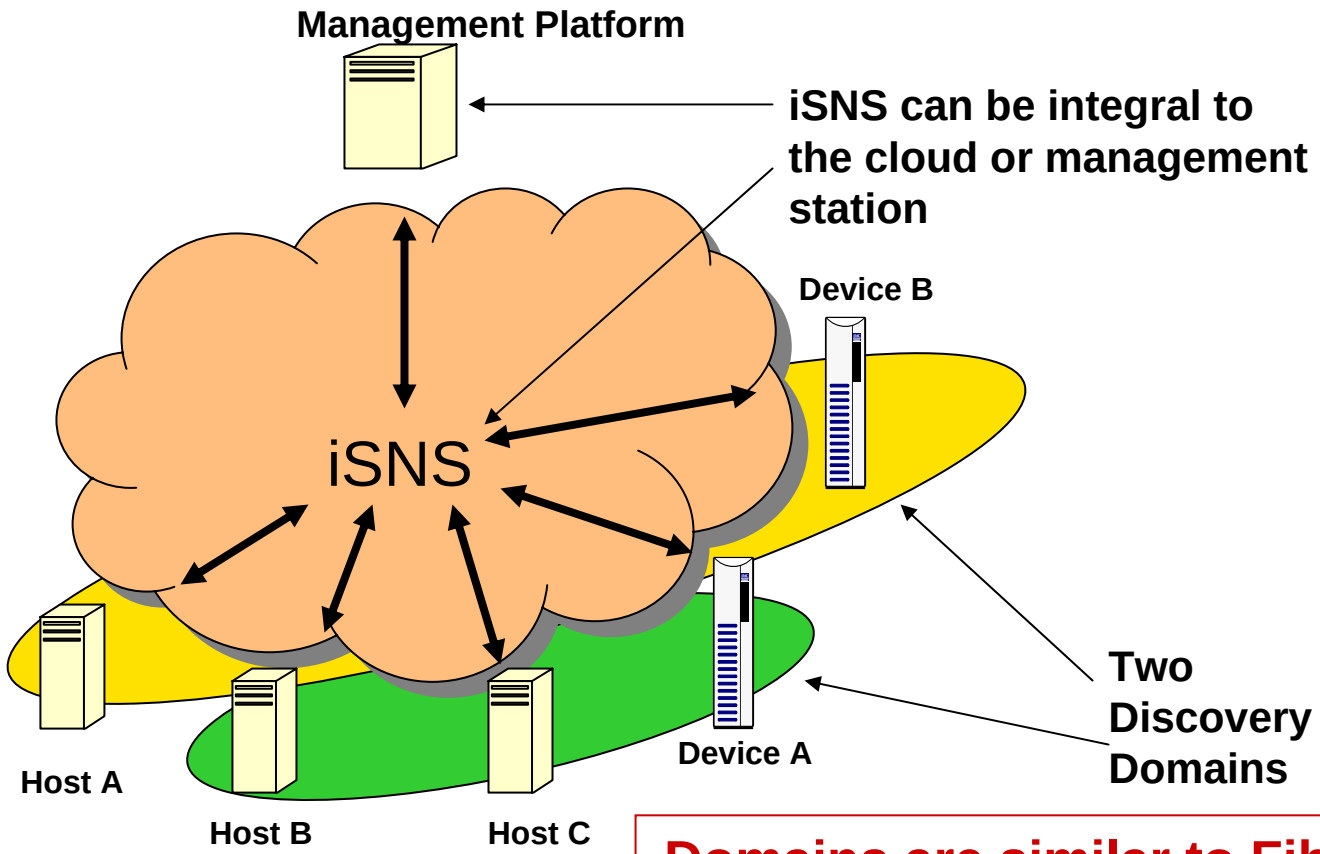
© Copyright 2006 EMC Corporation. All rights reserved.

iSNS (Storage Name Service)

- Modeled on Fibre Channel Nameserver
  - Discovery domains: Similar to soft FC zones
- Scalable discovery and configuration management
  - Asynchronous notification of changes
- Initiator retrieves all iSCSI target info from iSNS
  - Rich information repository (e.g., IPsec config info)
  - Enables more centralization of management

© Copyright 2006 EMC Corporation. All rights reserved.

iSNS Structure



Domains are similar to Fibre Channel zones, e.g., Host C will not discover Device B

© Copyright 2006 EMC Corporation. All rights reserved.

Login [How]

- Two types of login sessions
  - Discovery (SendTargets)
  - Normal (after any discovery mechanism)
- Normal login phases
  1. Security negotiation
  2. Operational parameter negotiation
  3. Full feature (perform I/O)
- Login uses text-based parameter negotiation
  - Syntax: key=value (or list of values)
  - Designed for extensibility

© Copyright 2006 EMC Corporation. All rights reserved.

Additional iSCSI Topics

- Security – Protect valuable data
- Error handling – Things will go wrong
- Implementation classes – NICs and HBAs
- Multipathing – Important HA mechanisms
- Boot – Yes, it can be done

© Copyright 2006 EMC Corporation. All rights reserved.

Security Properties

- Authentication: Who are you? Prove it!
  - Mutual Authentication: Initiator to Target AND vice-versa
- Integrity: Has this data been tampered with?
  - Cryptographic integrity, not just checksum or CRC
  - Linked to authentication to prevent regeneration attack
- Authorization: What are you allowed to do?
  - iSCSI: Who can connect to which Target
  - LUN masking & mapping handled by SCSI, not iSCSI
- Confidentiality: Has this data been disclosed?
- iSCSI: Usage is optional – subject to negotiation

© Copyright 2006 EMC Corporation. All rights reserved.

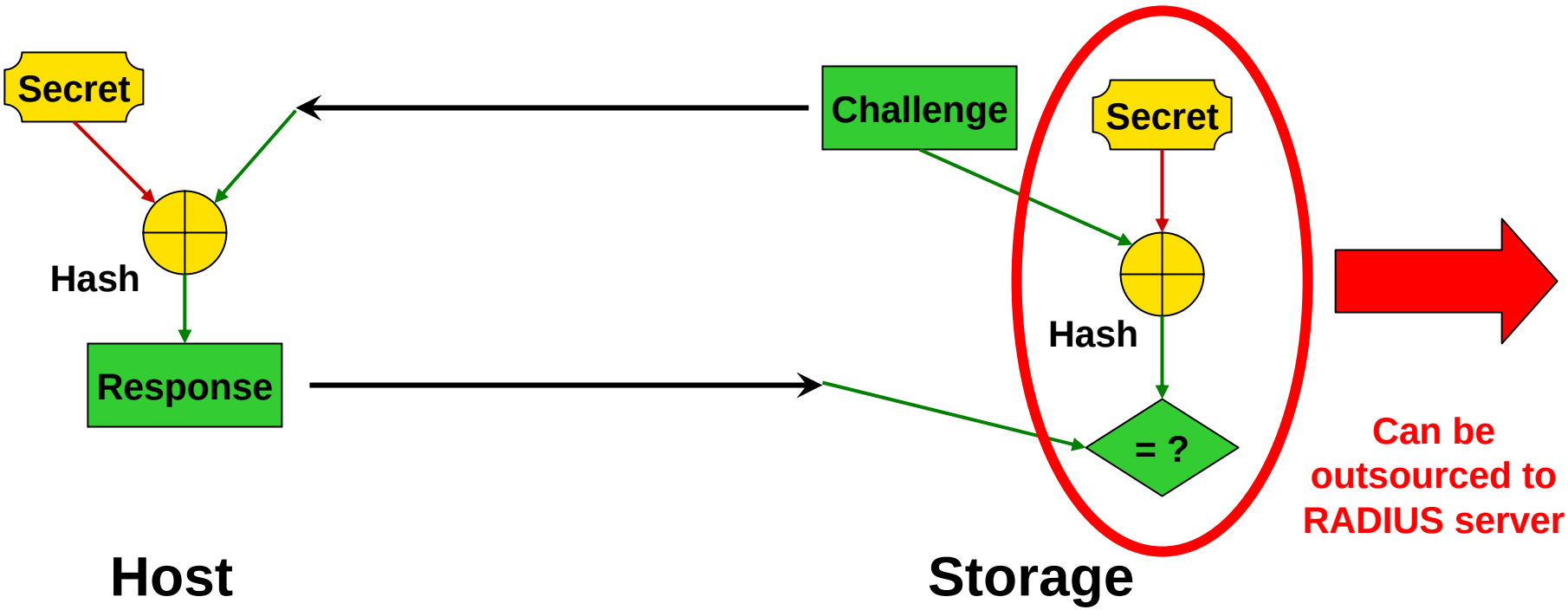
iSCSI Security: Protect valuable data

- Secure IP connection
  - Integrity, authentication, and confidentiality
  - Based on IKE and ESP (IPsec components)
- Extensive applied security requirements
  - Selection of Integrity (MAC) and encryption algorithms
  - Profile for usage of IKE authentication and key management
- Inband authentication (part of Login)
  - SRP, CHAP, Kerberos, and other mechanisms
  - CHAP with strong secrets is required
    - Can’t use passwords
  - iSCSI CHAP: Stronger than basic CHAP
    - When specification is followed

© Copyright 2006 EMC Corporation. All rights reserved.

CHAP Authentication Protocol

- Based on shared secret, random challenge
  - Uses a secure (one-way) hash, usually MD5
  - One-way hash: Computationally infeasible to invert



© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Error Detection

- Sequence numbers detect missing things
  - Commands, responses, data blocks
  - Goal: Avoid SCSI retry if at all possible
  - Command sequencing also used for flow control
    - Sliding window of commands target will accept
    - Data flow control: R2T (Ready to Transmit) mechanism
- Optional digests improve communication integrity
  - In addition to TCP checksum and Ethernet CRC
  - New 32-bit CRC polynomial (not the Ethernet CRC-32)
  - Separate CRCs computed over header and data
    - Allows an iSCSI proxy (e.g., router) to preserve data CRC

© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Error Recovery: Three Levels

- Error recovery level 0: Session recovery
  - Basic recovery mechanism that always works
  - Recover by session restart (close all TCP connections)
- Error recovery level 1: Digest failure recovery
  - Recover from digest failure without session restart
  - Recover by reissuing commands, data and/or status on same connection
- Error recovery level 2: Connection recovery
  - Open new TCP connection to replace failed connection
  - New connection picks up at point where old one failed
- Error recovery level negotiated during login

© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Implementation Classes

- NIC: iSCSI driver in software, standard NIC
  - Utilizes operating system TCP/IP stack
  - Link aggregation is below iSCSI driver
  - Digests and IPsec handled by software
  - Higher CPU utilization (but not prohibitive)
- HBA: Offload both TCP/IP and iSCSI
  - Appears as a SCSI controller to the operating system
  - Digests and IPsec handled by hardware
  - Lower CPU utilization due to full offload
  - Harder to support link aggregation and iSCSI sessions that span multiple HBAs

© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Multipathing Mechanisms

- Ethernet trunking
  - Link layer (2), below TCP, transparent to iSCSI
- Multiple TCP connections
  - In a single iSCSI session (layer 5)
  - Same or different hardware (Ethernet) ports
  - Difficult when TCP and iSCSI are offloaded
- Multiple iSCSI sessions
  - Multipathing software (e.g., PowerPath) above iSCSI
  - Same or different hardware (e.g., Ethernet) ports
- iSCSI also supports HTTP-style redirects
  - Target has been temporarily or permanently moved

© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Boot

- Have to discover the boot target
  - Can use DHCP (root path option) for this
  - Boot is usually from LUN zero
- Boot requires early access to system volume
  - Must be available prior to operating system running
  - iSCSI protocol can support booting
- NIC and iSCSI software driver: Have to modify OS
  - PXE (DHCP + TFTP) can download modified OS image
  - Int 13 BIOS boot: Need iSCSI driver in system BIOS
- HBA: No OS modifications needed
  - Int 13 BIOS boot: iSCSI can be in HBA card BIOS

© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI Status

- iSCSI protocol specification: Done
  - IETF RFC 3720 – April 2004
- iSCSI ancillary documents: Done
  - Naming, discovery, management - most published as RFCs
  - Final MIB documents (iSCSI, iSNS) to be published during 2006
- New protocol document: iSCSI Implementers Guide
  - Clarification of issues that have arisen in implementations
- New work area: iSER
  - iSER = iSCSI extensions for RDMA
    - RDMA = Remote Direct Memory Access (remote DMA)
  - Extend iSCSI to exploit RDMA
    - IP RDMA – IETF Remote Direct Data Placement (rddp) WG

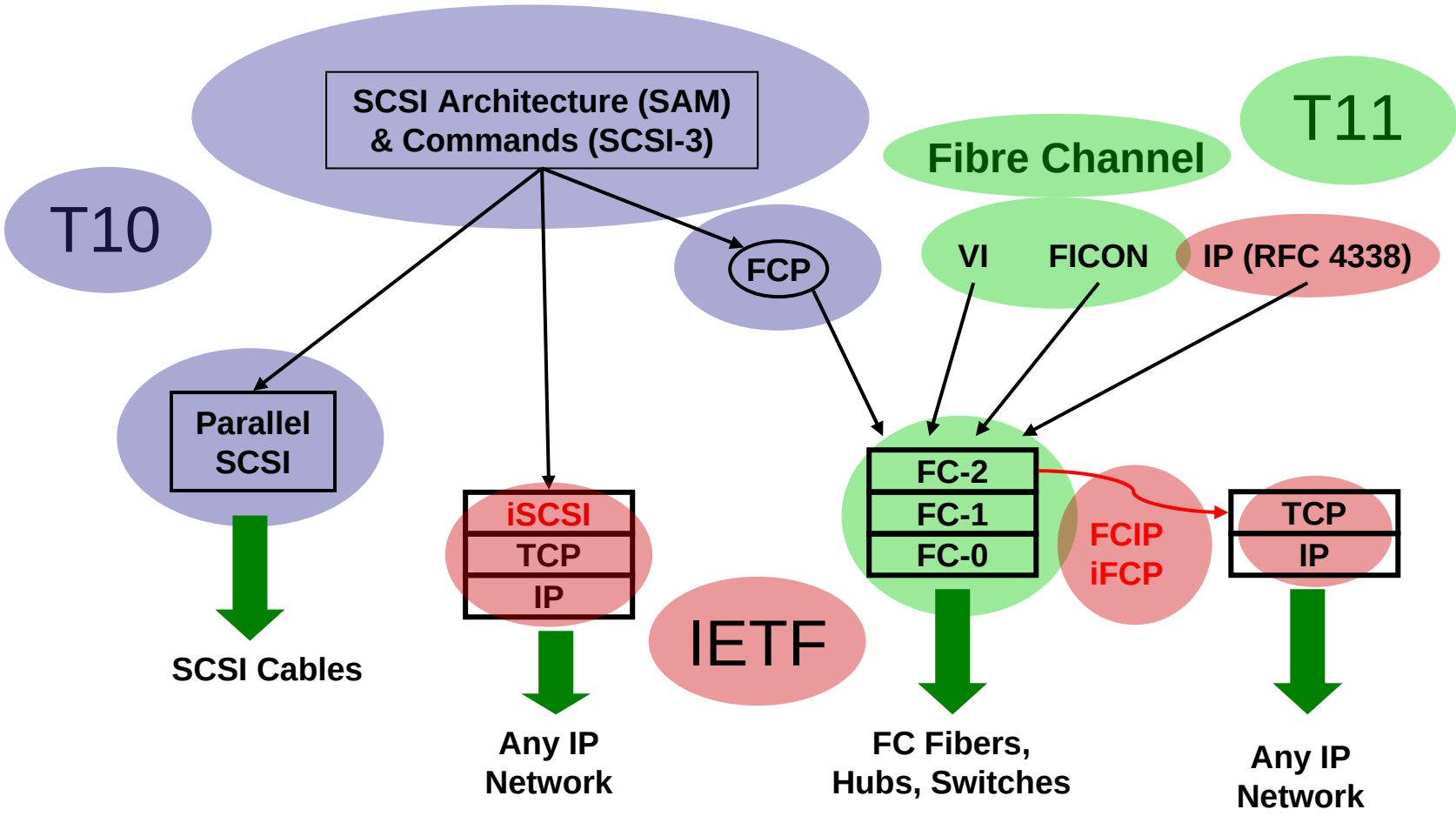
© Copyright 2006 EMC Corporation. All rights reserved.

iSCSI: Summary and Conclusion

- iSCSI: SCSI storage access over TCP/IP networks
  - Protocol stack: SCSI, iSCSI, TCP, IP (& IPsec), Ethernet
  - Works over any IP network, not just Ethernet
- iSCSI transports SCSI commands and data
  - Native iSCSI storage access
  - Bridged access to Fibre Channel storage
- iSCSI session establishment
  - Target naming (multiple formats) *[What]*
  - Target discovery (multiple mechanisms) *[Where]*
  - Login negotiation (multiple parameters) *[How]*
  - Followed by: Full feature phase (e.g., reads and writes)

© Copyright 2006 EMC Corporation. All rights reserved.

SCSI Protocols and Standards Organizations



© Copyright 2006 EMC Corporation. All rights reserved.

Standards Organizations

- SCSI: T10
  - www.t10.org
- Fibre Channel: T11
  - www.t11.org
- IETF IP Storage Working Group
  - http://www.ietf.org/html.charters/ips-charter.html
    - Latest versions of drafts are linked to that page
  - Chair: David L.Black (EMC)
- Active coordination on overlapping matters

© Copyright 2006 EMC Corporation. All rights reserved.


EMC²
where information lives®