A Storage Menagerie: SANs, Fibre Channel, Replication and Networks

David L. Black
Distinguished Engineer

NANOG 51 Conference
Miami, FL – January 30, 2011

EMC² where information lives

© Copyright 2010 EMC Corporation. All rights reserved.

Storage Networking: NAS and SAN

- DAS: Direct Attached Storage (not networked)
- NAS: Network Attached Storage: File
  - Distributed filesystems: Serve files and directories
    - Shared access is default
  - NFS (Networked File System)
  - CIFS (Common Internet File System)
- SAN: Storage Area Network: Block
  - Distributed (logical) disks: Serve blocks
    - Non-shared access is default
    - Sharing requires clustering (application, filesystem or volume mgr.)
  - SCSI (Small Computer System Interface)-based
    - Examples: iSCSI, Fibre Channel (FC)
- Focus of this talk: SAN technology, protocols and networks

© Copyright 2010 EMC Corporation. All rights reserved.

SAN Storage Arrays: Overview

- Make logical disks out of physical disks
  - Array contains physical disks, servers access logical disks
- High reliability/availability:
  - Redundant hardware, server-to-storage multipathing
  - Disk Failures: Data mirroring, parity-based RAID
  - Internal Housekeeping (e.g., disk scrubbing)
    - Can often predict failures (e.g., replace drives before they fail)
  - Power Failures: UPS used, entire array may be battery-backed
    - Ride through power glitches, orderly shutdown on power failure
    - Battery-backed write-back cache, may dump to disk on power failure
- Extensive storage functionality
  - Slice, stripe, concatenate, thin provisioning, dedupe, auto-tier, etc.
  - Snapshot, clone, copy, remotely replicate, etc.

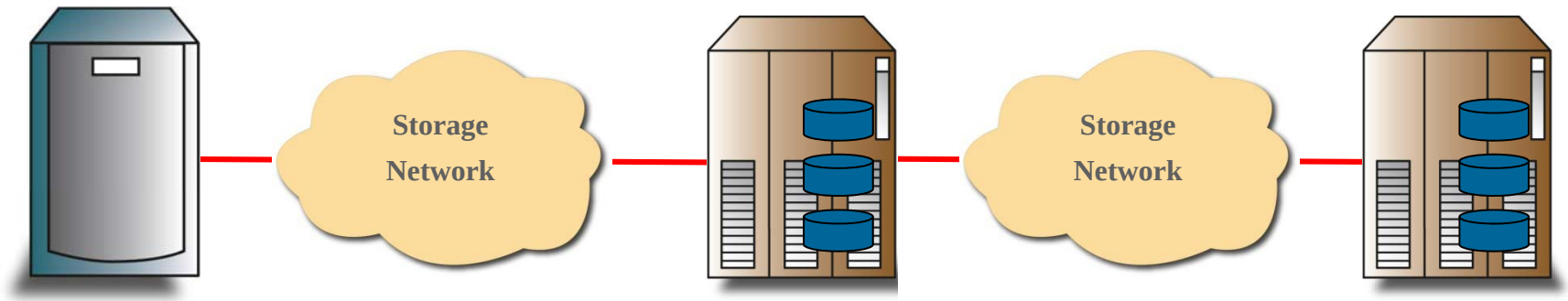
© Copyright 2010 EMC Corporation. All rights reserved.

SAN Storage Arrays: Examples

- Small SAN array: 3U rack-mount
  - 8 Fibre Channel ports (up to 8Gb/sec each)
  - Minimum: 4 disk drives, up to 2TB raw capacity/drive
  - Expands to 75 drives, up to 150TB raw capacity
  - Usable capacity depends on array configuration
- Very large SAN array: 10+ rack-sized cabinets
  - Up to 128 ports of 8Gb/sec Fibre Channel
  - 1Gb/sec and 10Gb/sec Ethernet also available
  - Up to 2400 drives, 2PB max usable capacity

© Copyright 2010 EMC Corporation. All rights reserved.

Storage Protocol Classes



Server to Storage Access

- SAN: Fibre Channel, iSCSI
- NAS: NFS, CIFS

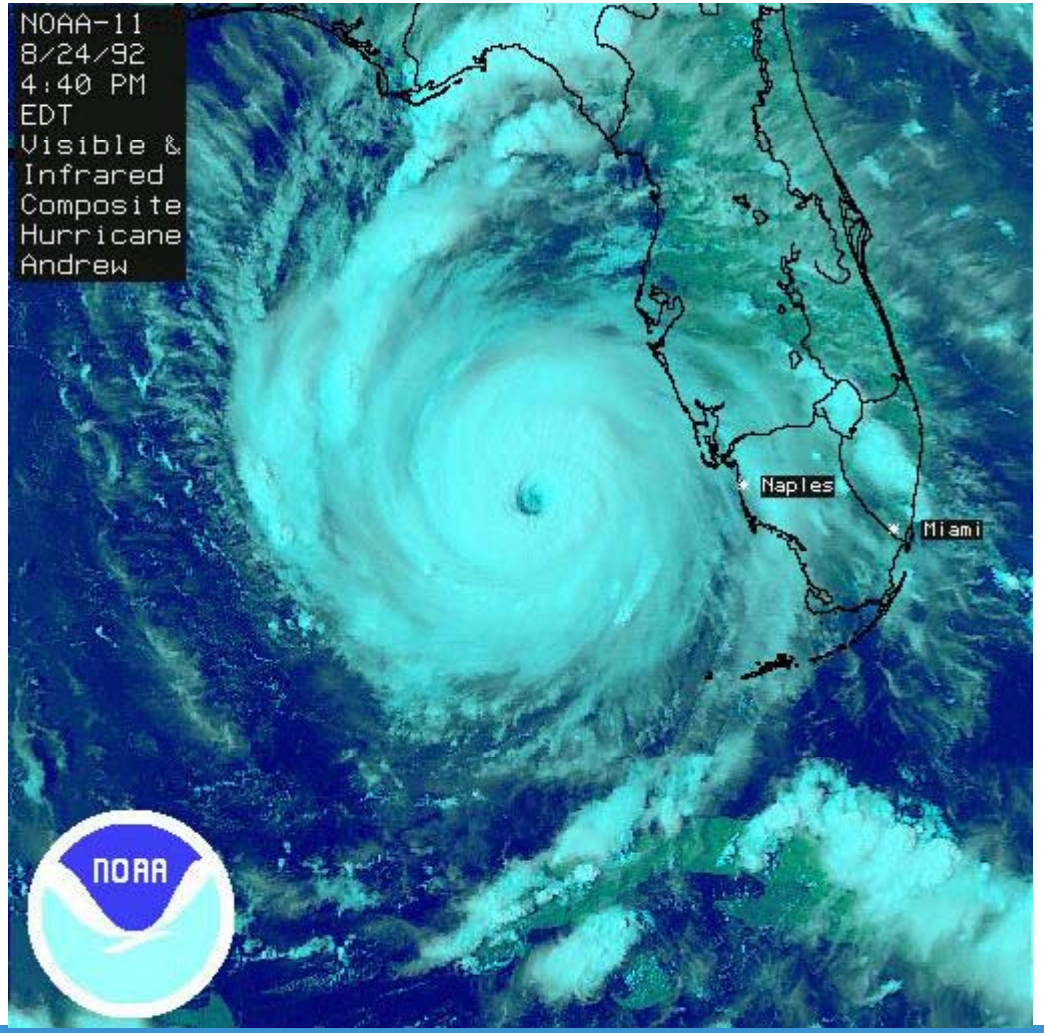
Storage Replication

- Array to Array, primarily SAN
- Often based on server to storage protocol

© Copyright 2010 EMC Corporation. All rights reserved.

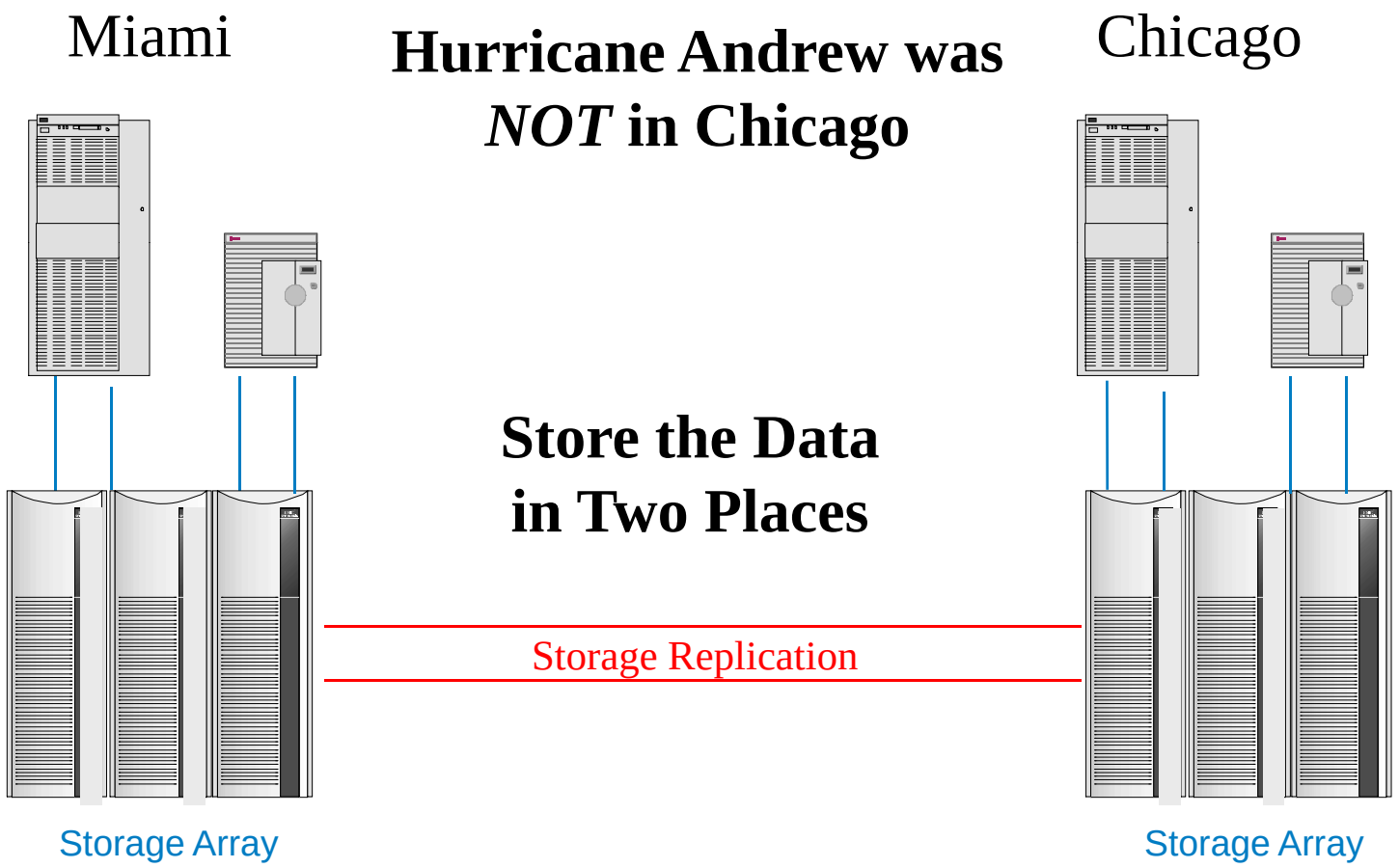
Why Storage Replication?

- Disasters Happen:
  - Power out
  - Phone lines down
  - Water everywhere
- The Systems are Down!
- The Network is Out!
- This is a problem …



© Copyright 2010 EMC Corporation. All rights reserved.

Storage Replication Rationale



© Copyright 2010 EMC Corporation. All rights reserved.

Storage Replication: 2 Types

- Synchronous Replication: Identical copy of data
  - Server writes not acknowledged until data replicated
  - Distance limited: Rule of thumb – 5ms round-trip or 100km (60mi)
  - Failure recovery: Incremental copy to resychronize
- Asynchronous Replication: Delayed consistent copy of data
  - Server writes acknowledged before data replicated
  - Used for higher latencies, longer distances (arbitrary distance ok)
  - Data consistency after failure: Manage replicated writes
- Replication often based on access protocol (e.g., FC, iSCSI)
  - Additional replication logic for error recovery, data consistency, etc.
  - Resulting replication protocol is usually vendor-specific

© Copyright 2010 EMC Corporation. All rights reserved.

Storage Replication and Network Failures

- Initial Reaction: Ignore network failure, hope it goes away
  - It often does ☺, but sometimes it doesn't ☹
  - Worst case: Initial indication of rolling site disaster ☹ ☹
- Failure goes away: Resume storage replication, catch up
  - Don't recopy all the data !! (bad protocol design)
  - Track what still needs to be replicated (e.g., bitmap, log)
- Site Disaster: Ensure remote data copy is consistent
  - Consistent data copy = Possible "power-fail" data image
    - Formal Requirement: Dependent write consistency
    - Database example: Data table write depends on transaction log write
  - All replication: Turn on/off consistently (with respect to server I/O)
    - May involve multiple arrays at each site, each with own network connections
  - Asynchronous replication: Organize replicated writes
    - Ordered log of writes: Ok at small to medium scale
    - Larger scale: Group writes, apply groups atomically at remote site

© Copyright 2010 EMC Corporation. All rights reserved.

Storage Replication: Network Sizing

- Characterize (measure) the I/O workload
  - Measuring I/O traffic beats guessing almost every time ☺
- Synchronous replication: Size network for worst case
  - I/O outruns network: Servers slow down ☹
  - Provide network capacity to handle peak I/O load
- Asynchronous replication: Size network to limit data loss
  - I/O outruns network: Data queue builds up in source array
    - No immediate impact on servers
    - Disaster: Data queue and data in flight are both lost
  - Limit data loss by limiting size of source data queue
    - RPO: Recovery Point Objective (typically measured in minutes/seconds)
    - Maximum RPO = Maximum amount of data that may be lost on disaster
  - Provision network capacity to limit source data queue size
    - Plus headroom for unexpected traffic peaks, anticipated growth
- Related term: RTO (Recovery Time Objective)
  - RTO = Wall clock time required to restore application(s) after disaster

© Copyright 2010 EMC Corporation. All rights reserved.

The SCSI Protocol Family: Foundation of SAN Storage

© Copyright 2010 EMC Corporation. All rights reserved.

SCSI ("scuzzy")

- SCSI = Small Computer System Interface
  - But used with computers of all sizes
- Client-server architecture (really master-slave)
  - Initiator (e.g., server) accesses target (storage)
  - Target is slaved to initiator, target does what it's told
- Target could be a disk drive
  - Embedded firmware, no admin interface
  - Resource-constrained by comparison to initiator
    - SCSI target controls resources, incl. data transfer (e.g., write)
- I/O performance rule of thumb: Milliseconds Matter
  - 5ms round-trip delay can cause visible performance issue

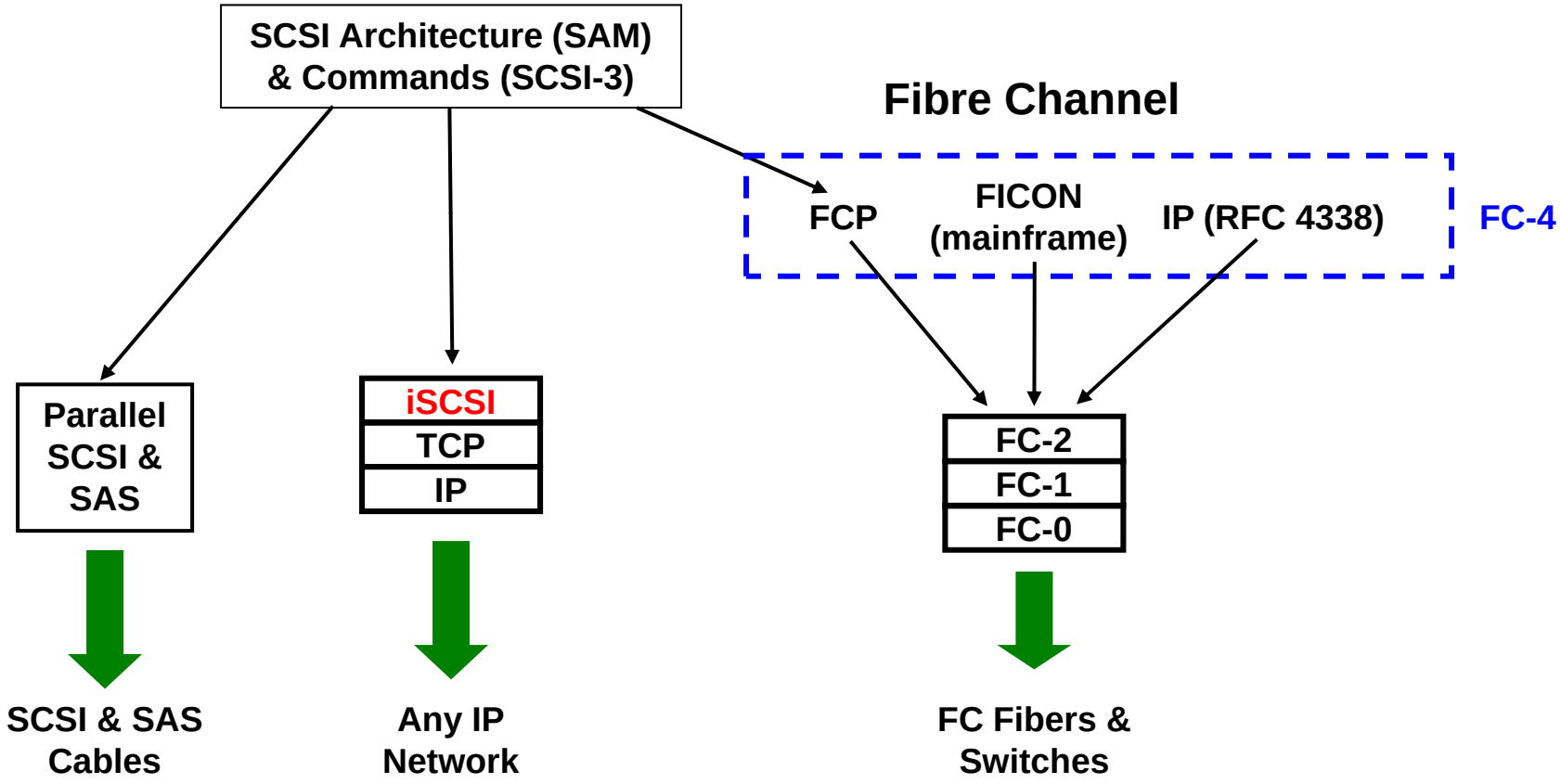
© Copyright 2010 EMC Corporation. All rights reserved.

SCSI Architecture

- SCSI Command Sets & Transports
  - SCSI Command sets: I/O functionality
  - SCSI Transports: Communicate commands and data
  - Same command sets used with all transports
- Important SCSI command sets
  - Common: SCSI Primary Commands (SPC)
    - All SCSI devices (e.g., INQUIRY, TEST UNIT READY)
  - Disk: SCSI Block Commands (SBC)
    - Concurrent block-addressed I/O commands for disks, storage arrays
  - Tape: SCSI Stream Commands (SSC)
    - Single stream of I/O commands to tape device
- SCSI Transport examples
  - FC: Fibre Channel (via SCSI Fibre Channel Protocol [FCP])
  - iSCSI: Internet SCSI
  - SAS: Serial Attached SCSI

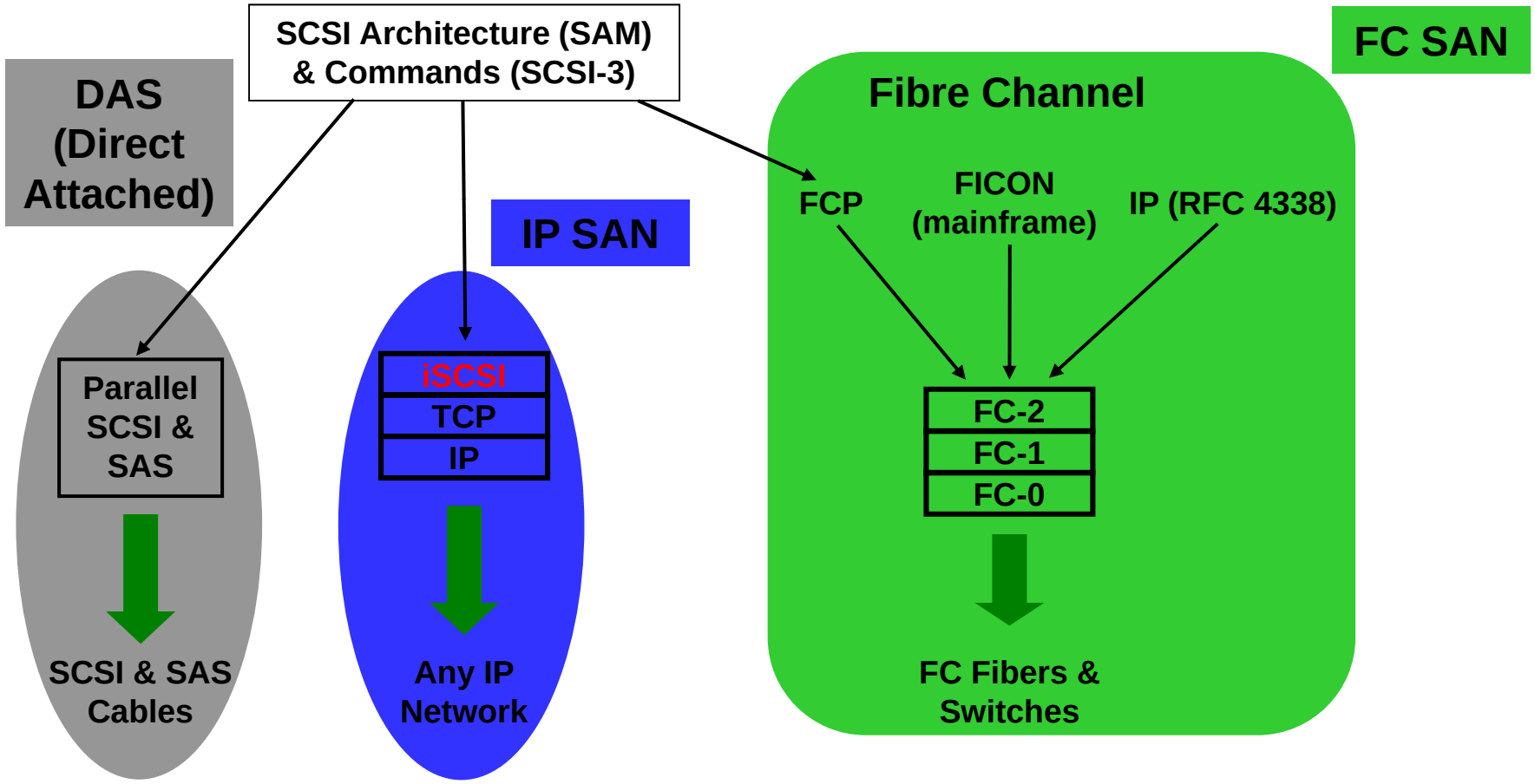
© Copyright 2010 EMC Corporation. All rights reserved.

The SCSI Protocol Family



© Copyright 2010 EMC Corporation. All rights reserved.

The SCSI Protocol Family and SANs



© Copyright 2010 EMC Corporation. All rights reserved.

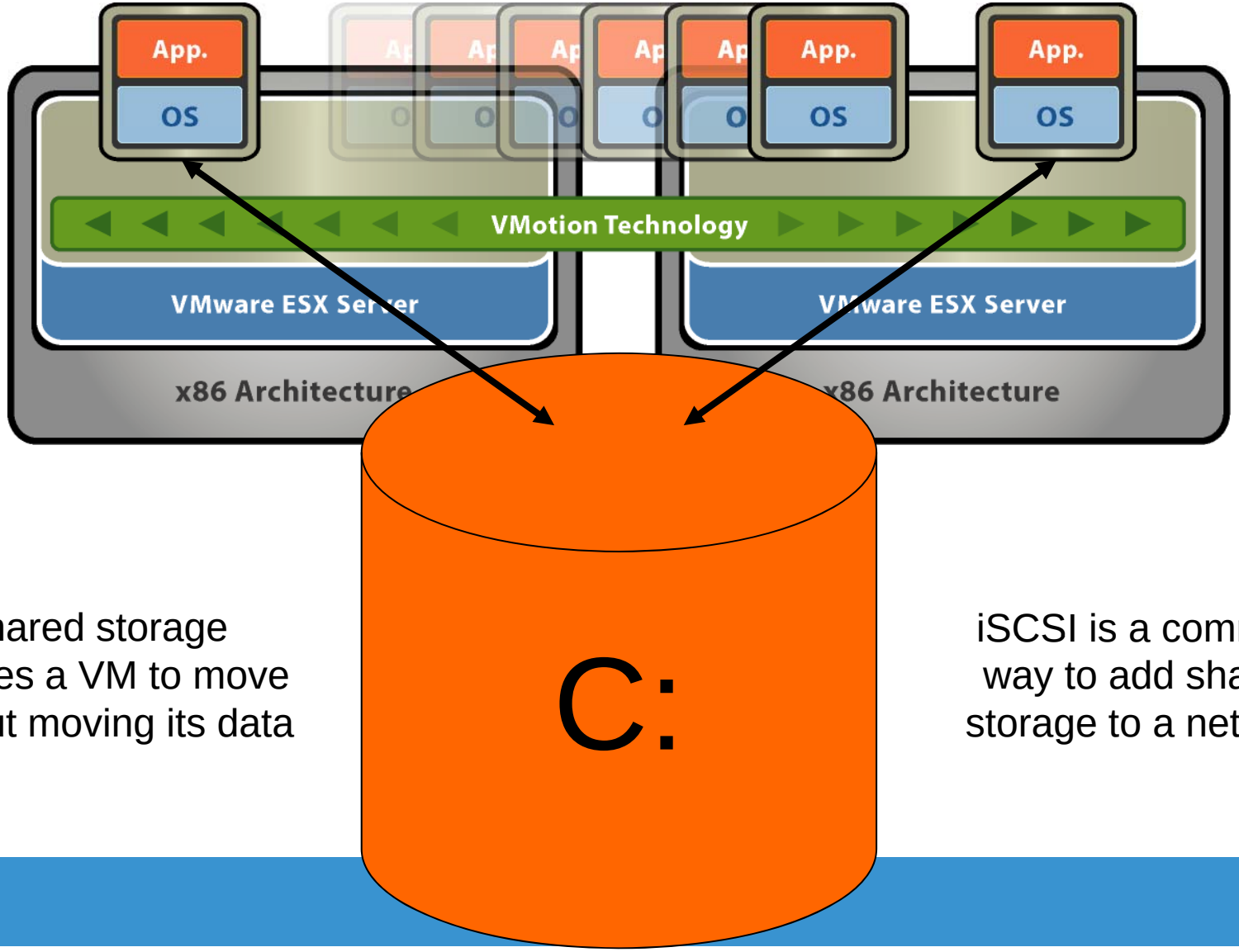
IP SAN: iSCSI

- SCSI over TCP/IP
  - TCP/IP encapsulation of commands
  - TCP/IP data transfer service (for read/write)
  - Communication session and connection setup
    - Multiple TCP/IP connections allowed in a single iSCSI session
  - Task management (e.g., abort command) & error reporting
- Typical usage: Within data center (1G & 10G Ethernet)
  - 1G Ethernet: Teamed links common when supported by O/S
- Separate LAN or VLAN recommended for iSCSI traffic
  - Isolation: Avoid interference with or from other traffic
  - Control: Deliver low latency, avoid spikes if other traffic spikes
  - Not strictly required, but helps in practice
- Data Center Bridging (DCB) Ethernet helps w/VLAN behavior
  - Reduces/avoids interference among VLANs
  - Use Ethernet Pause on links without VLANs

© Copyright 2010 EMC Corporation. All rights reserved.

iSCSI Example: Live Virtual Machine Migration

Move running Virtual Machine across physical servers

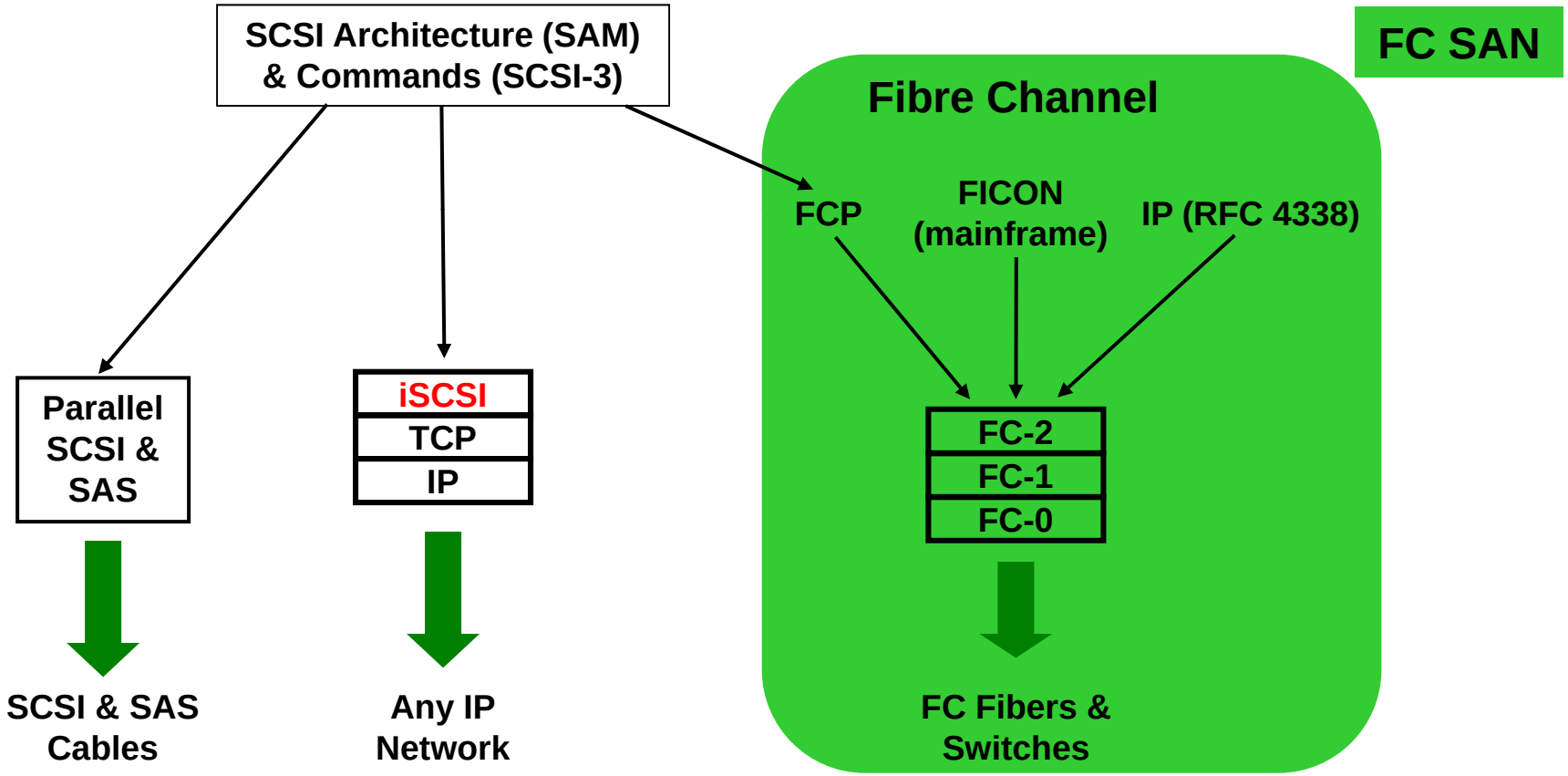


Shared storage enables a VM to move without moving its data

iSCSI is a common way to add shared storage to a network

© Copyright 2010 EMC Corporation. All rights reserved.

The SCSI Protocol Family and Fibre Channel



© Copyright 2010 EMC Corporation. All rights reserved.

Fibre Channel: Introduction

- FC Communication: Based on frames
  - Frame Header: 24 bytes
  - Frame Payload: Up to 2112 bytes (holds at least 2kB of data)
- FC ports have types
  - N_Ports: Servers or Storage
  - F_Ports: Switch ports to communicate with N_Ports
  - E_Ports: Switch ports to communicate with switches
- FC ports and switches assembled into a “Fabric”
  - Frame routing: FSPF (FC Fabric adaptation of OSPF)
  - Fabric switch count limit: 239 (current practice: ~60)
  - Fabric supports server/storage port discovery & management
- Fabric scale: Up to thousands of ports

© Copyright 2010 EMC Corporation. All rights reserved.

Native FC Links

- SAN FC links: Always optical
  - FC disk drive interfaces are different (copper, no shared access)
- Link encoding: 8b/10b (Like 1Gig Ethernet)
  - Error detection, Embedded synchronization
  - Control vs. data word identification
- Links are always-on (IDLE control word)
- Speeds: 1, 2, 4, 8 Gbit/sec (single lane serial)
  - New: “16” Gbits/sec uses 64b/66b, not 8b/10b (32GFC is next)
  - Limited inter-switch use of 10Gbit/sec (also uses 64b/66b)
- Credit based flow control (not pause/resume or XON/XOFF)
  - Buffer credit required to send (separate credit pool per direction)
  - FC link control operations return credits to sender

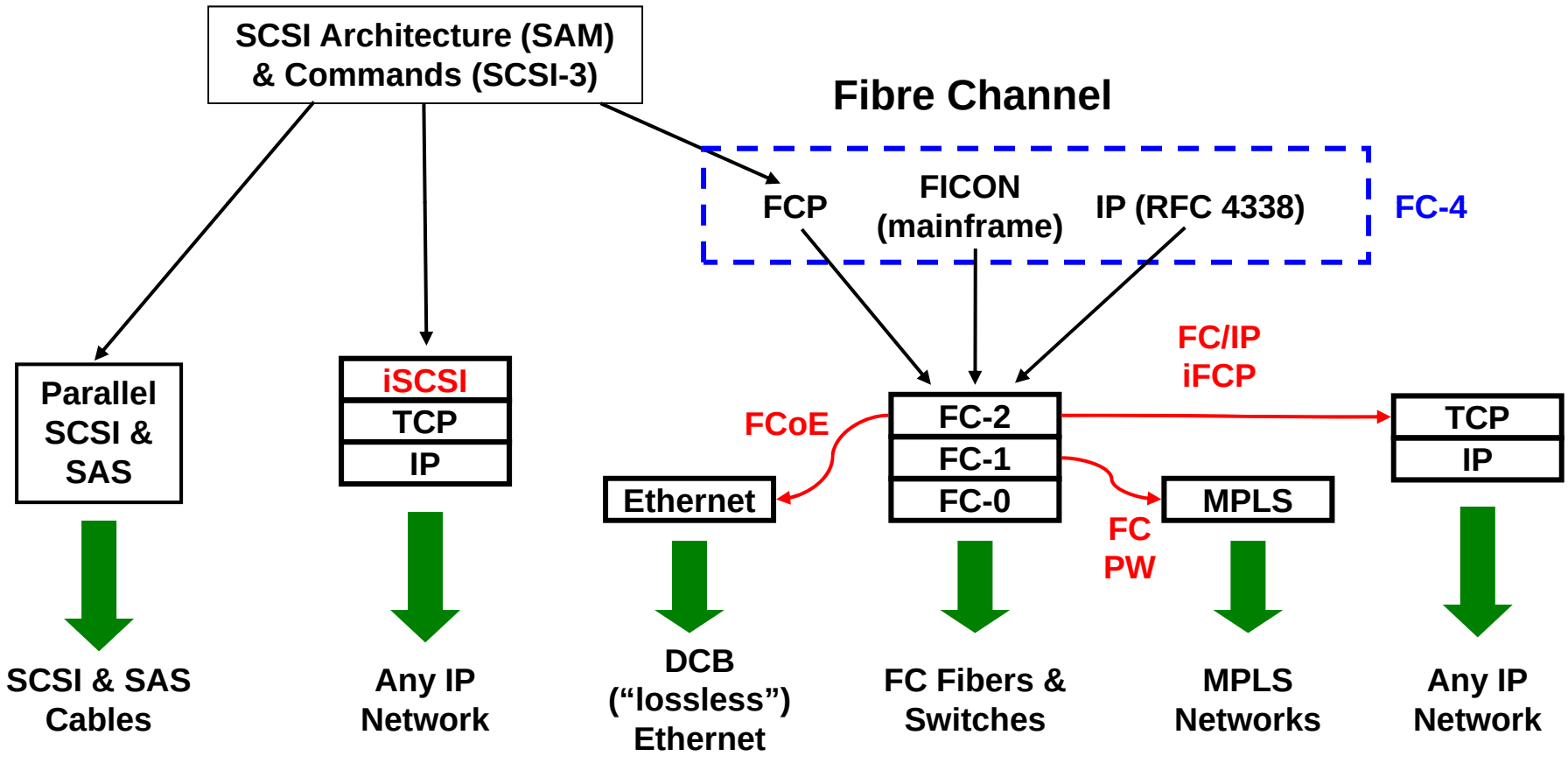
© Copyright 2010 EMC Corporation. All rights reserved.

FC timing and error recovery

- Strict timing requirements
  - R_A_TOV: Deliver FC frame or destroy it (typical: 10sec)
    - Failure to do so can corrupt data due to FC exchange ID reuse
  - Timeout budget broken down into smaller per-link timeouts
    - E_D_TOV, typical 1sec (round trip), also used to detect link failures
- Heavyweight error recovery: No reliable FC transport protocol
  - Disk: Retry entire server I/O (30sec or 60sec timeout is typical)
  - Tape: Stop, figure out what happened and continue from there
    - Streaming tape drive stops streaming (ouch!)
- FC is *very* drop-sensitive
  - Congestion: Overprovision to avoid congestion-induced drops
  - Errors: Rule of thumb: 10^{-15} BER (Bit Error Rate) or better
    - Native FC optics in data centers run at 10^{-18} BER or better
- Reminder: SCSI I/O is *very* latency sensitive

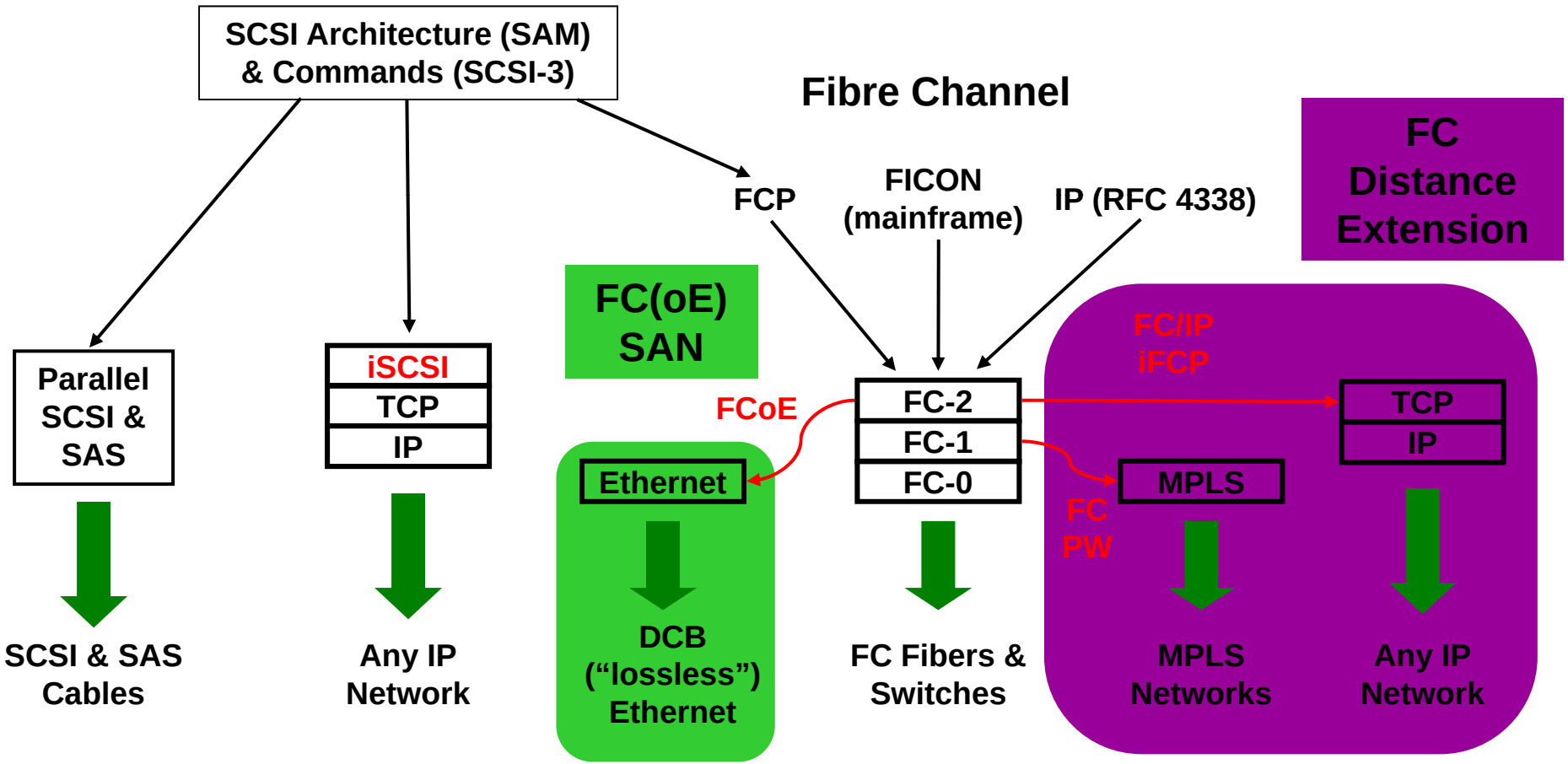
© Copyright 2010 EMC Corporation. All rights reserved.

The SCSI Protocol Family and Fibre Channel



© Copyright 2010 EMC Corporation. All rights reserved.

The SCSI Protocol Family and Fibre Channel



© Copyright 2010 EMC Corporation. All rights reserved.

FC Distance Extension

- Always between FC switches (E_Ports)
  - Primarily for storage replication (sync or async)
- Problem: Credit based flow control is distance sensitive
  - Solution: FC switches can provide additional buffering and credits

1. Simple: Extend native FC inter-switch link
   - Dark fiber or xWDM, no flow control (FC link control passed through)
2. Transparent: FC GFPT and FC pseudowire (pause/resume)
   - GFPT: (Asynchronous) Generic Framing Procedure – Transparent
   - FC PW (pseudowire): Extend FC over MPLS
3. Gateway-based: FC/IP and iFCP (TCP/IP)
   - TCP used for reliable delivery and flow control
   - Not transparent: Gateways used at extension endpoints

© Copyright 2010 EMC Corporation. All rights reserved.

FC Pseudowire (PW): FC over MPLS

- FC-PW: Based on FC-GFPT transparent extension protocol
- FC GFPT: Transport Generic Framing Protocol (Async)
  - Just send the 10b codes (from 8b/10b links)
  - Add on/off flow control (ASFC) to prevent WAN link "droop“
  - Used over SONET and similar telecom networks.
- FC-PW: Same basic design approach as FC-GFPT
  - Send 8b codes, use ASFC flow control
    - Original PW design used TCP-class flow control, replaced by ASFC
  - FC link control: separate packets
    - PW control word distinguishes control packets from data
  - IDLE suppression on WAN
  - Tight timeout for link initialization (R_T_TOV: 100ms rt)
- Notes: FC-PW is *new* and not currently specified for 16GFC
  - 16GFC (in development) uses 64b/66b encoding

© Copyright 2010 EMC Corporation. All rights reserved.

FC/IP and iFCP

- FC Switch to FC Switch extension via TCP/IP
  - E_D_TOV timeout (typically 1 sec rt) must be respected
  - Protocols include latency measurement functionality
- FC/IP: More common protocol
  - Only used for FC distance extension
- iFCP: More complex specification
  - FC distance extension: iFCP address transparent mode
  - iFCP not used for connection to servers or storage
    - iFCP address translation mode (specified for this usage): Deprecated (Historic)
- iSNS name server used to set up iFCP
  - iSNS can also be used with iSCSI
  - FC/IP has to be preconfigured
- iFCP is going away (being replaced by FC/IP in practice)

© Copyright 2010 EMC Corporation. All rights reserved.

FC/IP Network Customer Examples: Asynchronous Storage Replication

- Financial Customer A (USA):
  - ~2 PB (Peta Bytes !) of storage across 20 storage arrays (per site)
  - 5 x OC192 SONET = 50 Gb WAN @ 30 ms RTT (1000mi)
    - Network designed for › 70 % excess capacity
- Financial Customer B (USA):
  - ~5 PB of storage across 30 storage arrays (per site)
  - 2 x 10 Gb DWDM wavelengths @ 20ms RTT (700mi)
    - Network designed for › 50% excess capacity
- Financial Customer C (Europe):
  - ~0.7 PB (700 TB) across 9 arrays (per site)
  - 1 x 10 Gb IP @ 15ms RTT (500mi)
    - Current peak is 2 Gb/s of 6Gb available; WAN shared w/ tape
    - Network designed to support growth for 18 months

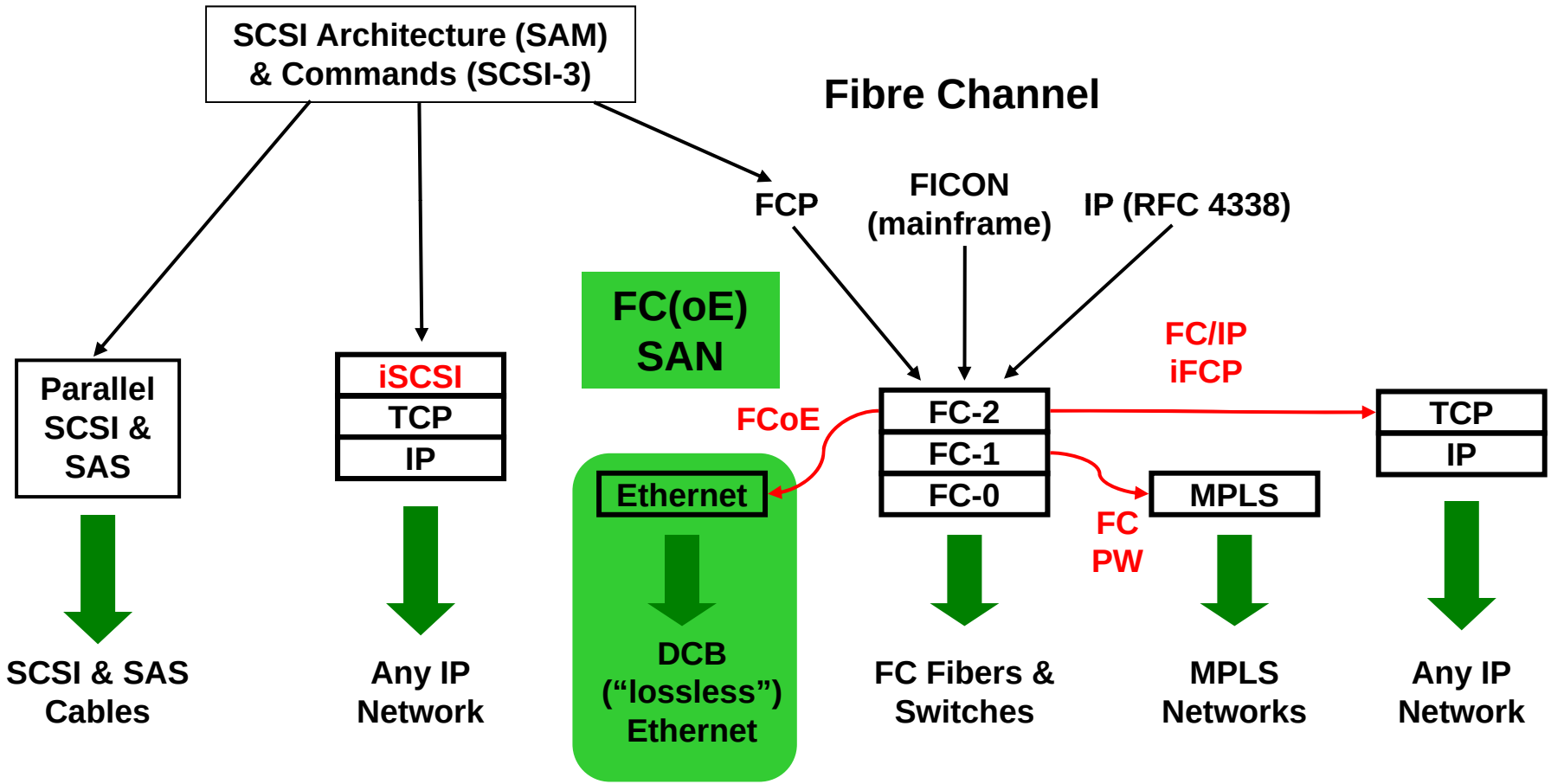
© Copyright 2010 EMC Corporation. All rights reserved.

FC/IP Network Customer Examples: Asynchronous Replication Details

- Typical Data Center connectivity:
  - Fibre Channel: Array to FC switch (FC interface)
  - Ethernet: FC switch (FC/IP interface) to WAN equipment
    - "I" in IP: Internetwork
    - FC switches don't need WAN interfaces ☺
- Customer Examples: two concurrent replication legs
  - Synchronous leg (max 200km/125mi) to bunker site
  - Asynchronous leg (longer distance) to recovery site
  - Descriptions on previous slide: Asynchronous leg
- Primary site disaster: Recover latest data from bunker
  - No data loss (Recovery Point Objective = zero)

© Copyright 2010 EMC Corporation. All rights reserved.

The SCSI Protocol Family and Fibre Channel



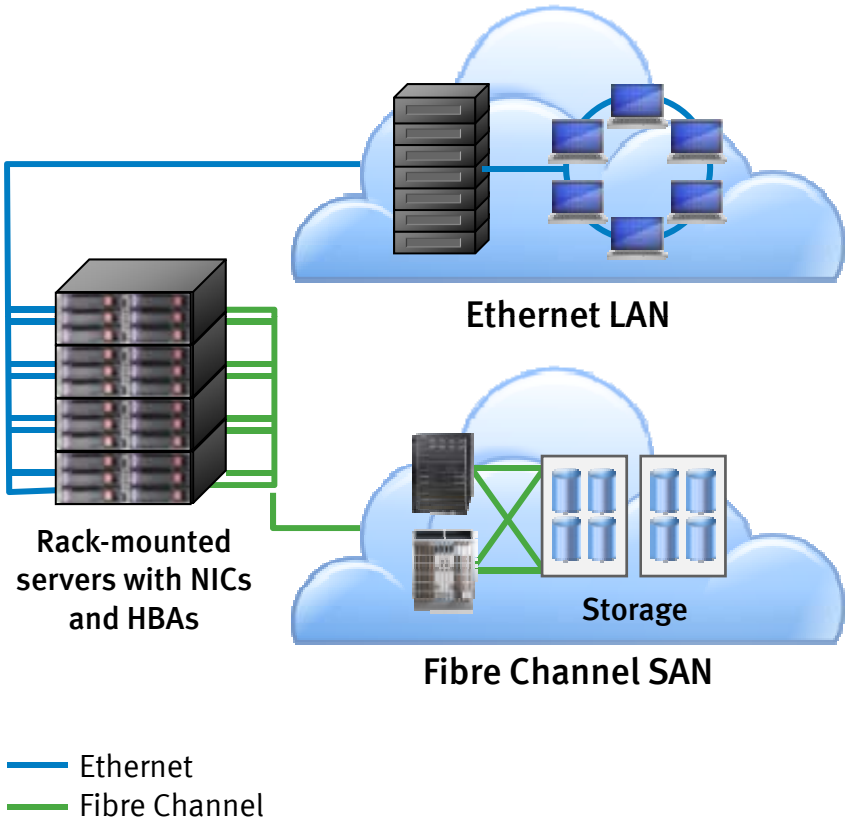
© Copyright 2010 EMC Corporation. All rights reserved.

FCoE: Fiber Channel over Ethernet

- Use Ethernet for FC instead of optical FC links
  - Encapsulate FC frames in Ethernet frames (no TCP/IP)
    - Requires at least baby jumbo Ethernet frames (2.5k)
  - Requires “lossless” (DCB) Ethernet and dedicated VLAN
    - Should dedicate bandwidth to VLAN – avoid drops and delays
- FIP (FCoE Initialization Protocol): Uses Ethernet multicast
  - Ethernet bridges are transparent: Potential link has › 2 ends !!
  - FIP discovers virtual ports, creates virtual links over Ethernet
- FCoE is a Data Center technology:
  - Typically server to storage, leverages FC discovery/management
  - Can also be used to interconnect FC/FCoE switches
- FCoE: Not appropriate for WAN
  - Need DCB (“lossless”) Ethernet WAN service
  - FIP use of multicast does not scale well to WAN

© Copyright 2010 EMC Corporation. All rights reserved.

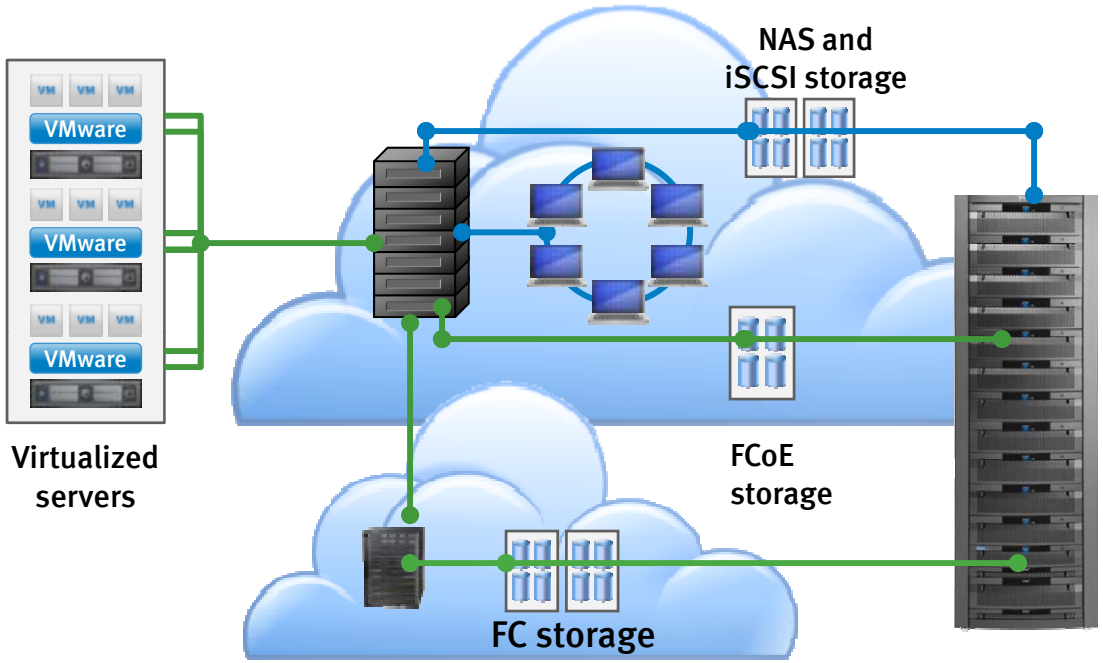
FCoE: Before



- Two separate networks
  - Ethernet LAN—server network connectivity
  - Fibre Channel SAN—server-to-storage access
- Ethernet and Fibre Channel networks are distinct technologies with different…
  - Terminology
  - Protocols
  - Management tools
  - Cabling
  - Interface cards
- Multiple networks require…
  - Unique equipment (cabling, switching infrastructure, adapters)
  - More people to manage and support two different technologies

© Copyright 2010 EMC Corporation. All rights reserved.

FCoE: After



- 10 Gigabit Ethernet (10 GbE) emerging as a converged network technology
  - Ideal for large enterprises and service providers
  - Many storage applications require › 1 Gigabit Ethernet bandwidth
  - Multiple storage protocols, e.g., iSCSI and FCoE
- Deployable today with FCoE
  - And other storage protocols
- Transition over time to converged networks & storage
  - Not all-at-once

© Copyright 2010 EMC Corporation. All rights reserved.

DCB Ethernet and Converged Networks

- DCB (Data Center Bridging) Ethernet consists of:
  - PFC (Priority based Flow Control): Pause per traffic class
  - ETS (Enhanced Transmission Selection): Bandwidth per traffic class
  - CN (Congestion Notification): Source backpressure at layer 2
  - DCBX (Data Center Bridging Capability Exchange): Negotiate usage
- FCoE *requires* DCB Ethernet
  - DCB eliminates packet drops
  - FC (and hence FCoE) is *very* drop-sensitive
- DCB: Useful for iSCSI and other TCP-based protocols
  - Avoids interference across traffic classes (e.g., consistent RTT)
  - Removes packet drops that would cause TCP backoff
- Convergence: Additional network management required
  - E.g., Configure bandwidth and priorities for traffic classes

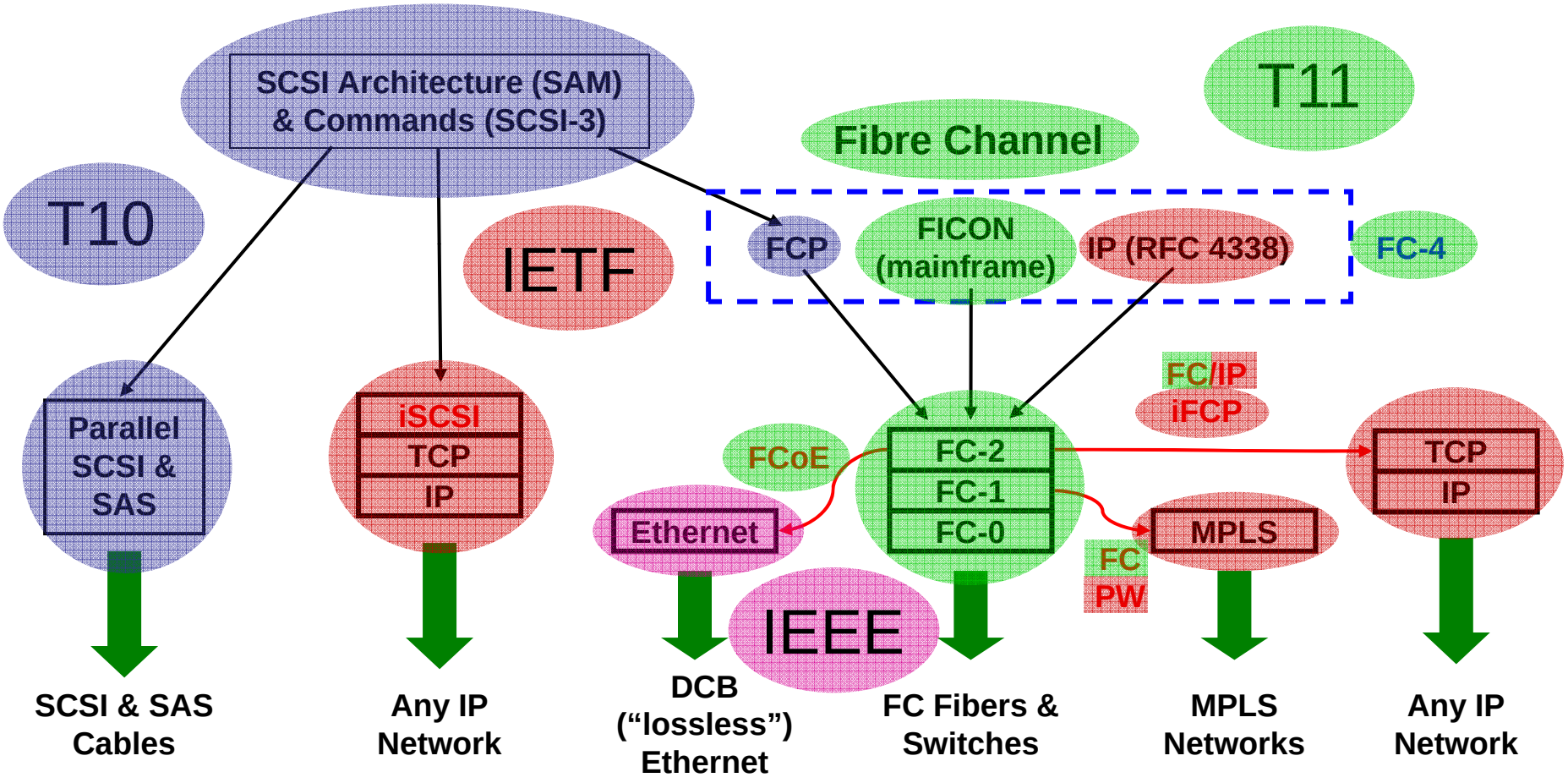
© Copyright 2010 EMC Corporation. All rights reserved.

Conclusion

- Storage Area Network functionality
  - Server to storage access
  - Storage replication
- SAN protocols (access, and basis for replication)
  - IP SAN: iSCSI
  - FC SAN: FC (and FCoE)
- Types of FC distance extension
  - Simple: Dark fiber and WDM
  - Transparent: FC GFPT and FC PW
  - Gateway: FC/IP and iFCP
- Storage can generate a lot of network traffic

© Copyright 2010 EMC Corporation. All rights reserved.

The SCSI Protocol Family and Standards Orgs.



© Copyright 2010 EMC Corporation. All rights reserved.

THANK YOU

© Copyright 2010 EMC Corporation. All rights reserved.