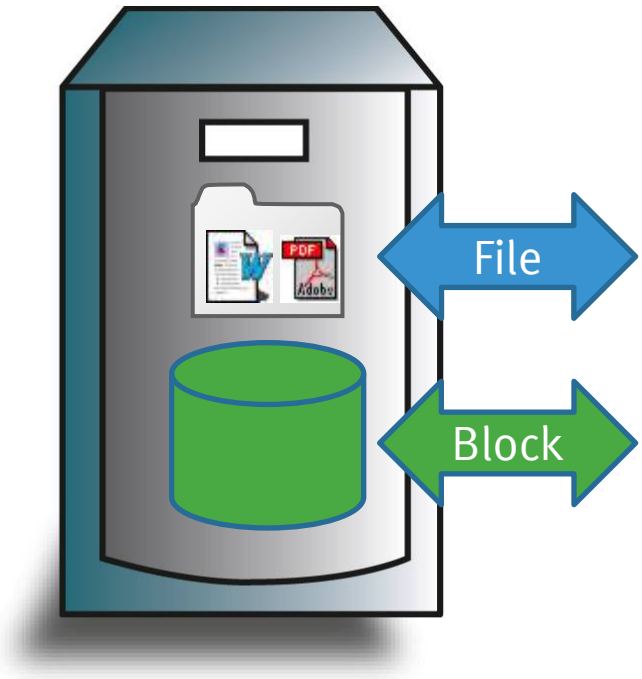
A Storage Menagerie: NAS, SAN & the IETF

IETF tsvarea meeting
Prague, CZ– March 30, 2011

© Copyright 2010 EMC Corporation. All rights reserved.

Storage Networking: NAS and SAN

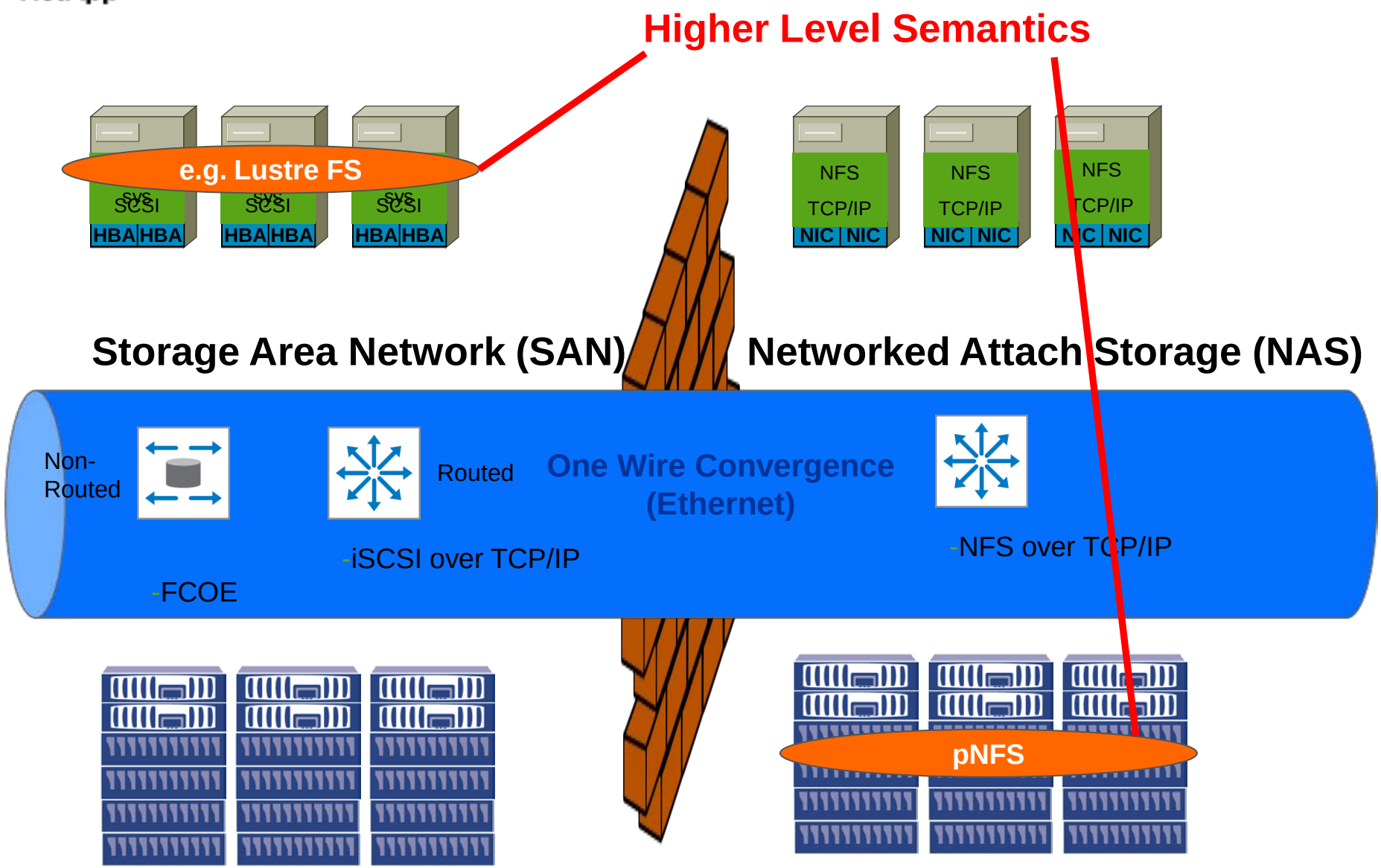


- NAS: Network Attached Storage: Remote Files
  - Distributed filesystems: Serve files and directories
  - NFS (Networked File System)
  - CIFS (Common Internet File System)
- SAN: Storage Area Network: Remote Disks [Blocks]
  - Distributed disks: Serve blocks
  - SCSI (Small Computer System Interface)-based
    - Examples: iSCSI, Fibre Channel (FC)

EMC² where information lives™

© Copyright 2010 EMC Corporation. All rights reserved.

Two Worlds of Storage



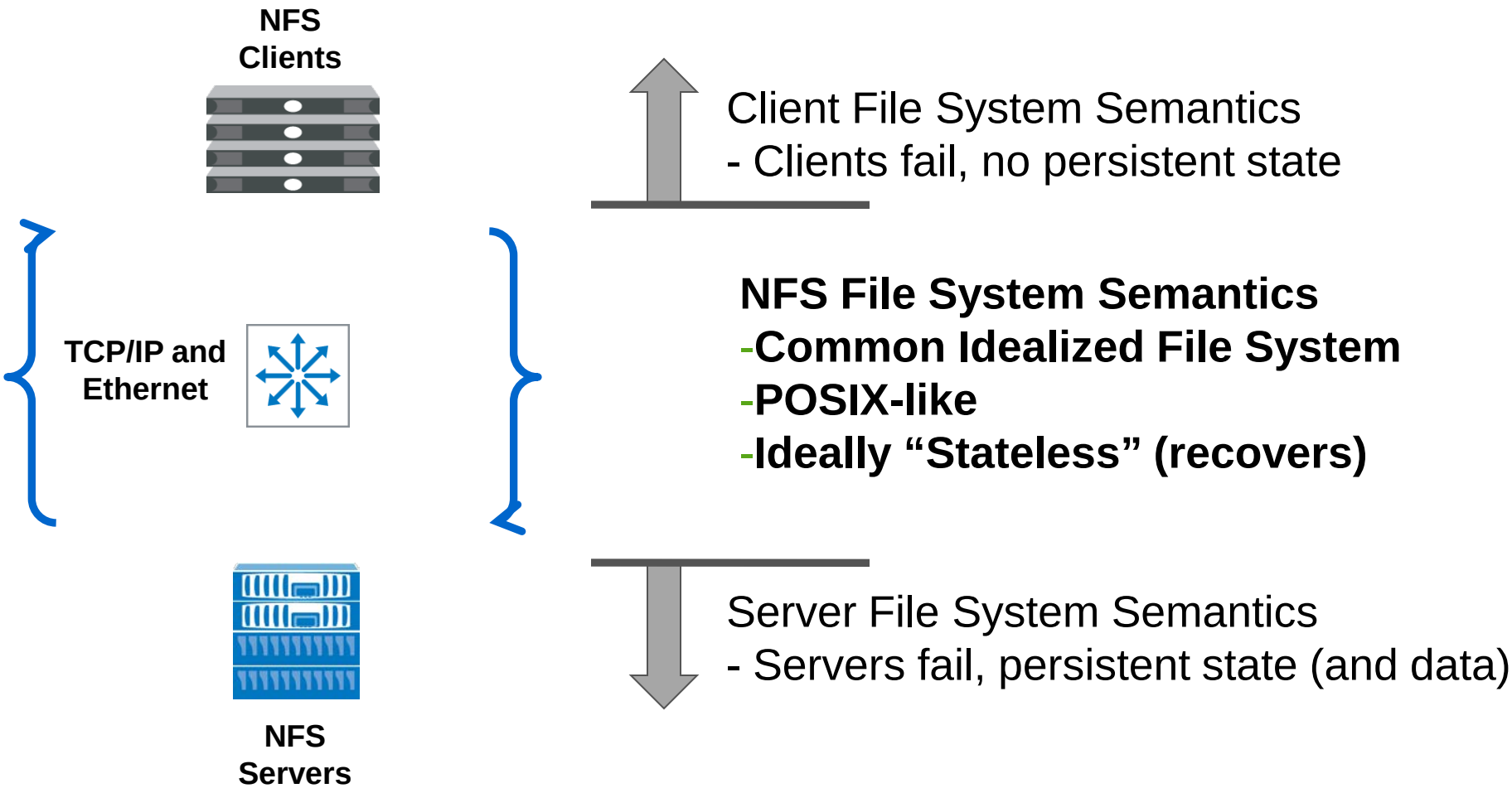
© 2011 NetApp. All rights reserved.

NAS: Network Attached Storage (Remote Files)

Brian Pawlowski
Co-chair NFSv4 WG

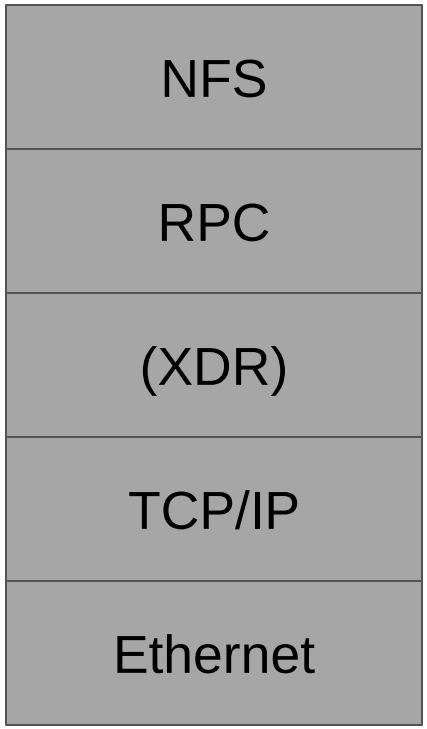
© Copyright 2010 EMC Corporation. All rights reserved.

Protocol vs. Implementation



© 2011 NetApp. All rights reserved.

Stack and Standards



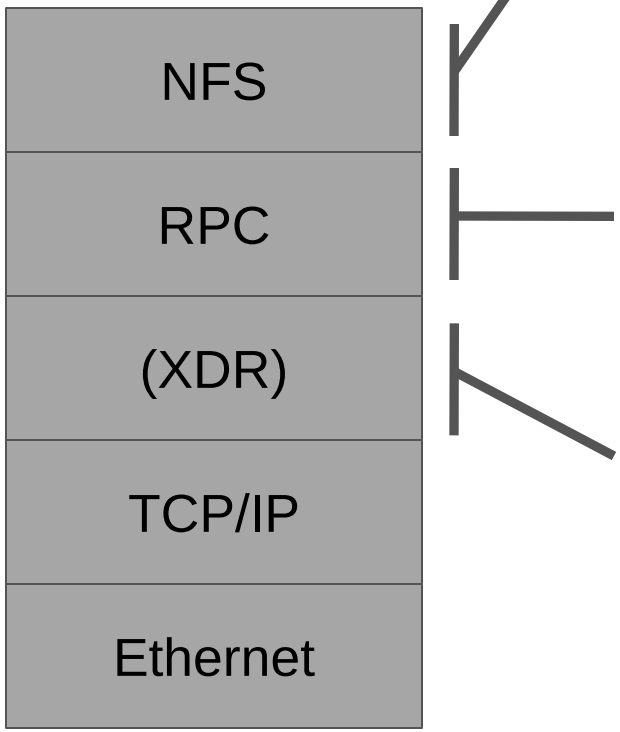
NFS
-RFC 5661 – NFS 4.1 (Proposed Standard)
-RFC 5662 – NFS 4.1 XDR (Proposed Standard
-RFC 5663 – Parallel NFS Block (Proposed Standard)
-RFC 5664 – Parallel NFS Object (Proposed Standard)

RPC
-RFC 5531 – V2 (Draft Standard)
-RFC 5403 – RPCSEC_GSS V2 (Proposed Standard)
-RFC 5665 – RPC IANA Considerations (Proposed Standard)

XDR
- RFC 5531 – (Standard)

© 2011 NetApp. All rights reserved.

Domains of Features



NFS (Network File System)
-Security Principal for Access Control
-Access Control List (ACL) operations
-Locking and Delegations using “leases”

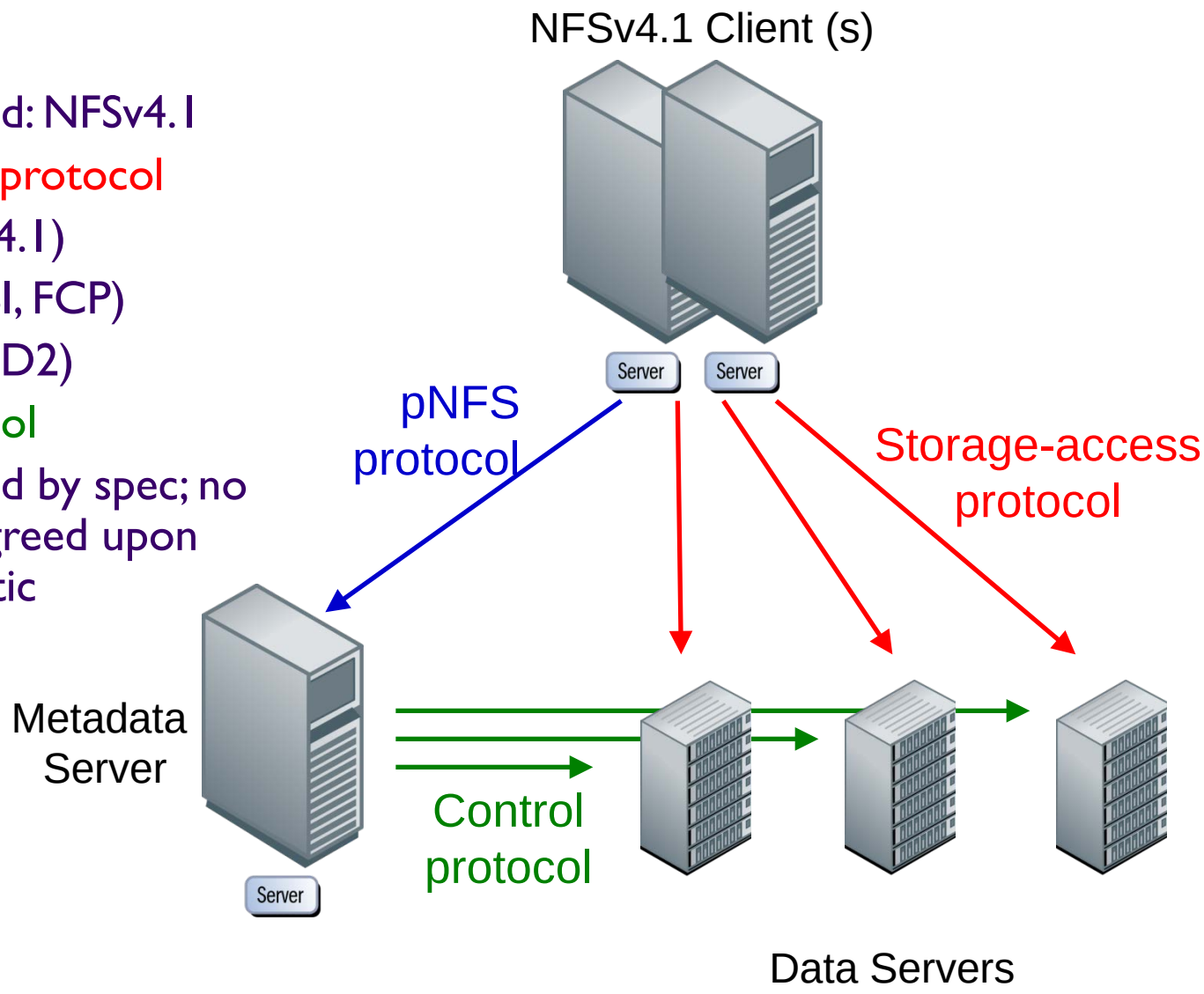
RPC (Remote Procedure Call)
-Security: Authentication Negotiation (flavors)

XDR (eXternal Data Representation
- Marshalling and unmarshalling

© 2011 NetApp. All rights reserved.

NFSv4.1 - Parallel NFS 101

- pNFS protocol
  - Standardized: NFSv4.1
- Storage-access protocol
  - Files (NFSv4.1)
  - Block (iSCSI, FCP)
  - Object (OSD2)
- Control protocol
  - Not covered by spec; no generally agreed upon characteristic



Source: SNIA Education

© 2011 NetApp. All rights reserved.

NFS future work and context

- NFS Version 4.2 (as a SMALL DELTA)
  - Small enhancements
  - Server side copy support
  - Space reservations
  - RPCSSEC GSS V3
- Data Center Concerns
  - Low latency in high bandwidth networks
- Expanding Use Cases
  - Large Streaming Data
  - Transactions
  - Storage for virtualized environments

© 2011 NetApp. All rights reserved.

SAN: Storage Area Networks (Remote Disks [Blocks])

David L. Black
Co-chair STORM WG

EMC² where information lives™

© Copyright 2010 EMC Corporation. All rights reserved.

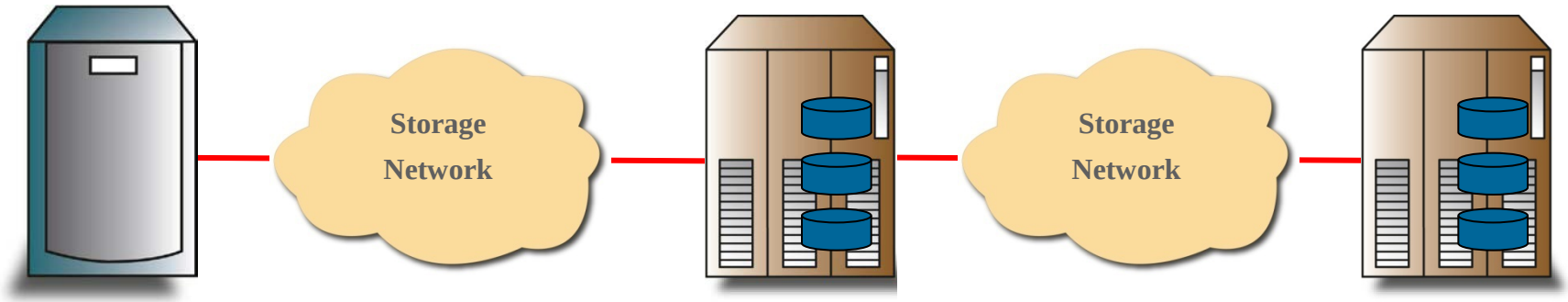
SAN Storage Arrays: Overview

- 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, failure prediction (e.g., disk scrubbing)
  - Power Failures: UPS is common, entire array may be battery-backed
- Extensive storage functionality
  - Slice, stripe, concatenate, thin provisioning, dedupe, auto-tier, etc.
  - Snapshot, clone, copy, remotely replicate, etc.

EMC² where information lives™

© Copyright 2010 EMC Corporation. All rights reserved.

Storage Protocol Classes



Server to Storage Access

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

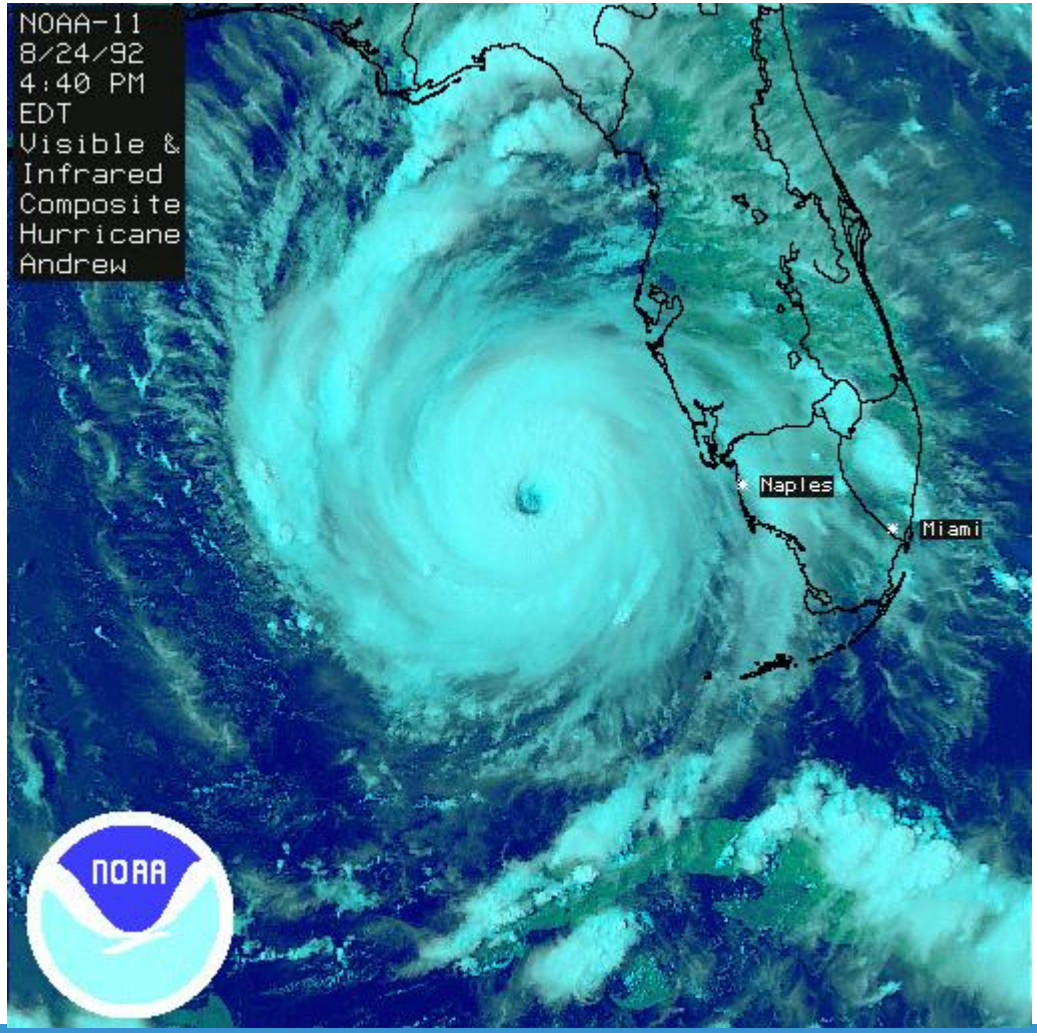
Storage Replication

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

© Copyright 2010 EMC Corporation. All rights reserved.

Why Remote Replication?

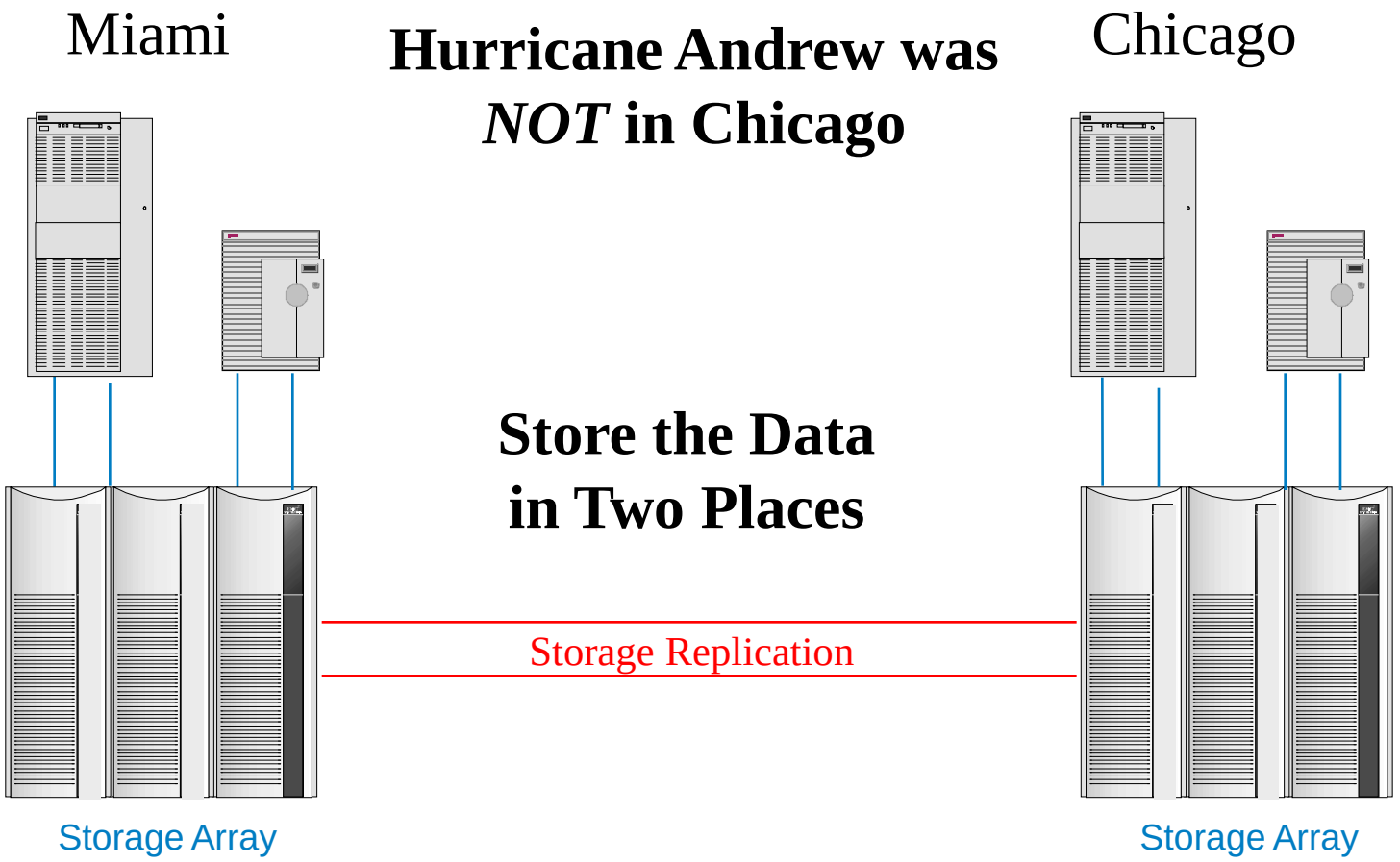
- Disasters Happen:
  - Power out
  - Phone lines down
  - Water everywhere
- The Systems are Down!
- The Network is Out!
- This is a problem …



EMC² where information lives

© Copyright 2010 EMC Corporation. All rights reserved.

Remote Replication Rationale



© Copyright 2010 EMC Corporation. All rights reserved.

Remote 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

EMC² where information lives™

© 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, e.g., data transfer for writes
- I/O performance rule of thumb: Milliseconds Matter
  - 5ms round-trip delay can cause visible performance issue

© Copyright 2010 EMC Corporation. All rights reserved.

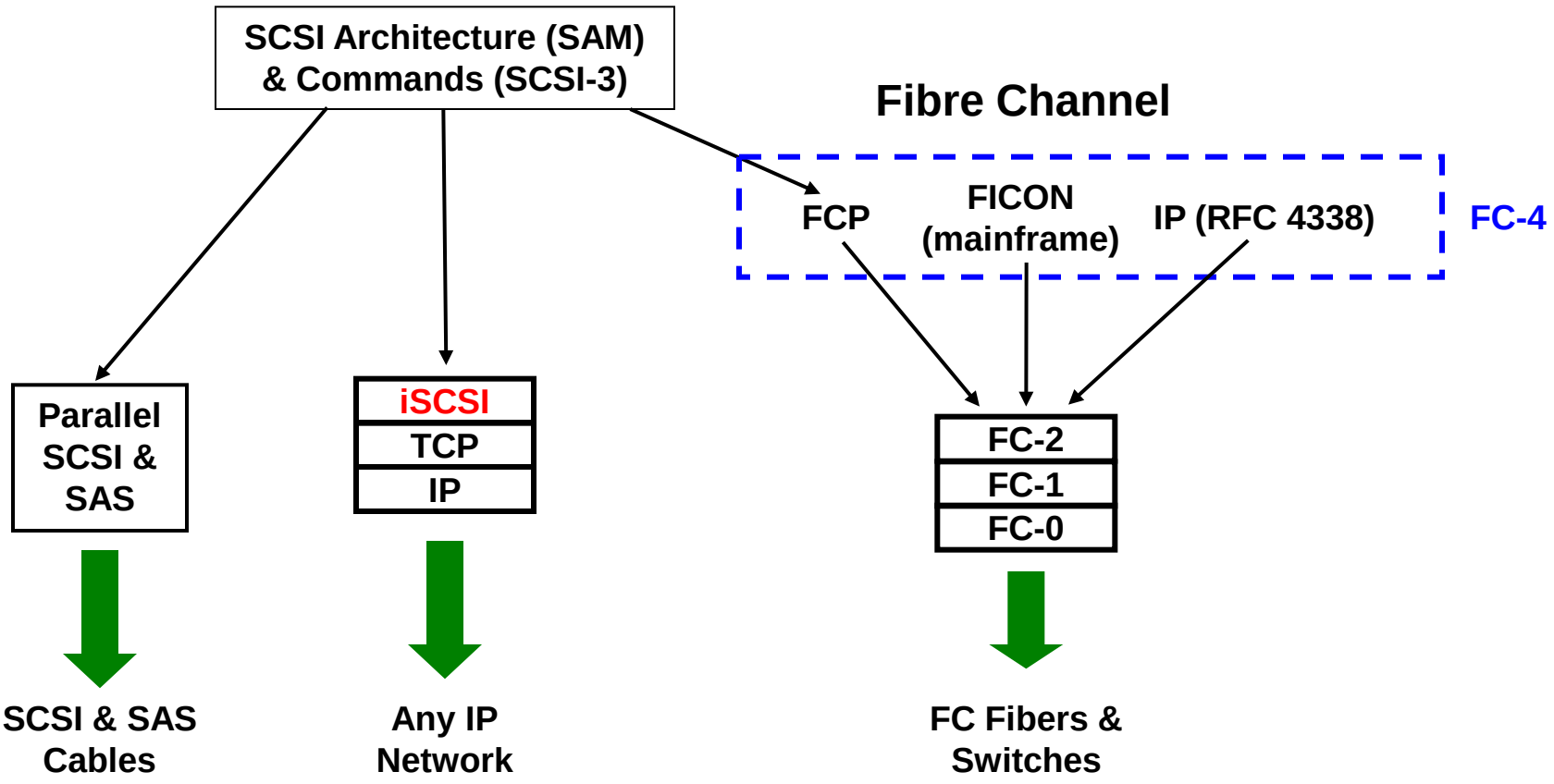
SCSI Architecture

- 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)
  - Disk: SCSI Block Commands (SBC)
  - Tape: SCSI Stream Commands (SSC)
- SCSI Transport examples
  - FC: Fibre Channel (via SCSI Fibre Channel Protocol [FCP])
  - iSCSI: Internet SCSI
  - SAS: Serial Attached SCSI
- Most SCSI functionality specified in T10 (e.g., commands)
  - T10 = SCSI standards organization (part of INCITS)

EMC² where information lives™

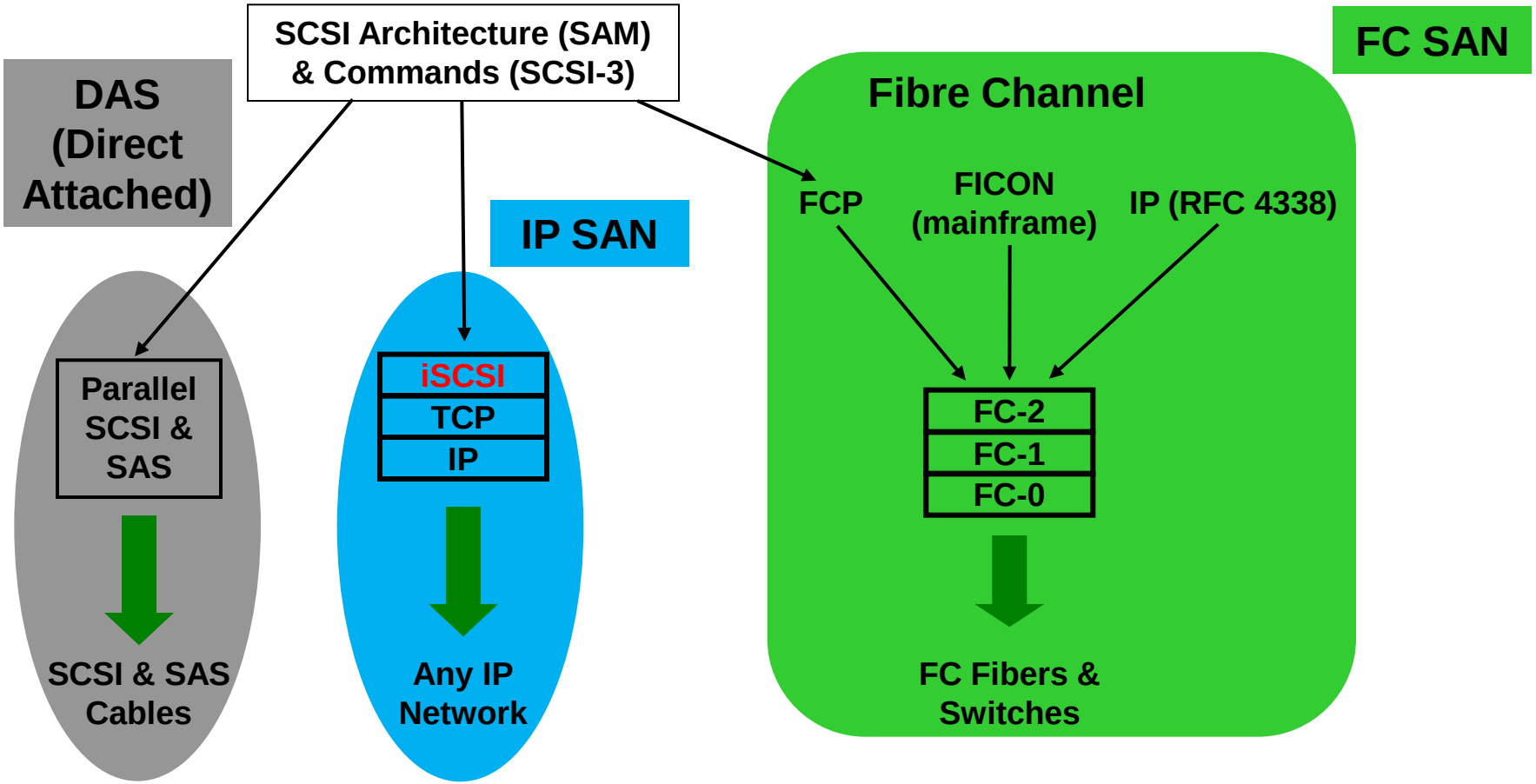
© Copyright 2010 EMC Corporation. All rights reserved.

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

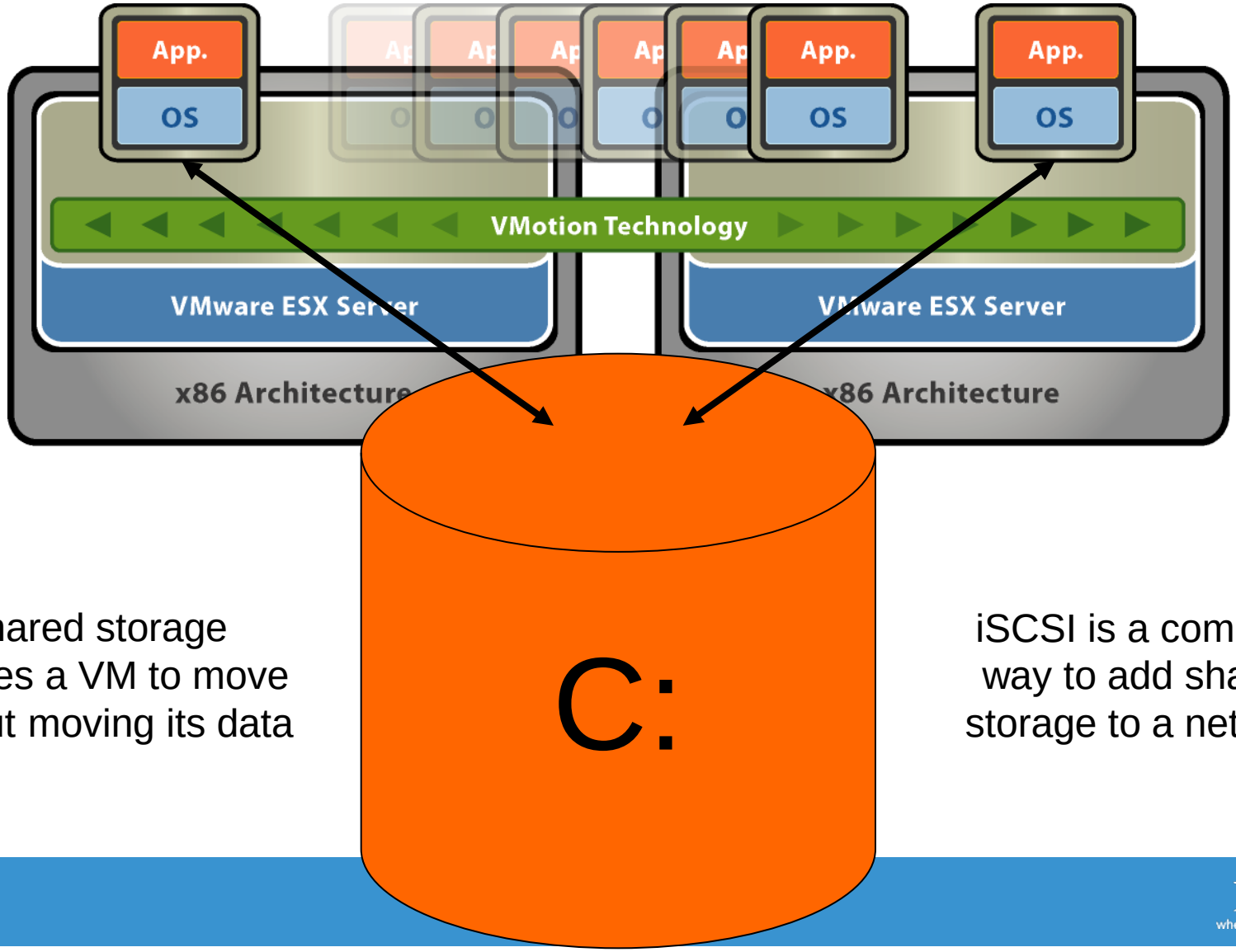
- SCSI over TCP/IP [RFC 3720 and friends]
  - TCP/IP encapsulation of commands, data transfer service (for read/write)
  - Communication session and connection setup
    - Multiple TCP/IP connections allowed in a single iSCSI session
  - Task management (e.g., abort command) & error reporting
- Typical usage: Within data center (1G & 10G Ethernet)
  - 1G Ethernet: Teamed links common when supported by OS
- 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
  - Data Center Bridging (DCB) Ethernet helps w/VLAN behavior
- ISCSI: Maintenance in STORM (STORage Maintenance) WG
  - Consolidate existing RFCs into one document
  - New draft adds a few new SCSI transport features (e.g., command priority)
- Most SCSI functionality is above the iSCSI level (see T10, not IETF)

EMC² where information lives™

© Copyright 2010 EMC Corporation. All rights reserved.

iSCSI Example: Live Virtual Machine Migration

Move running Virtual Machine across physical servers

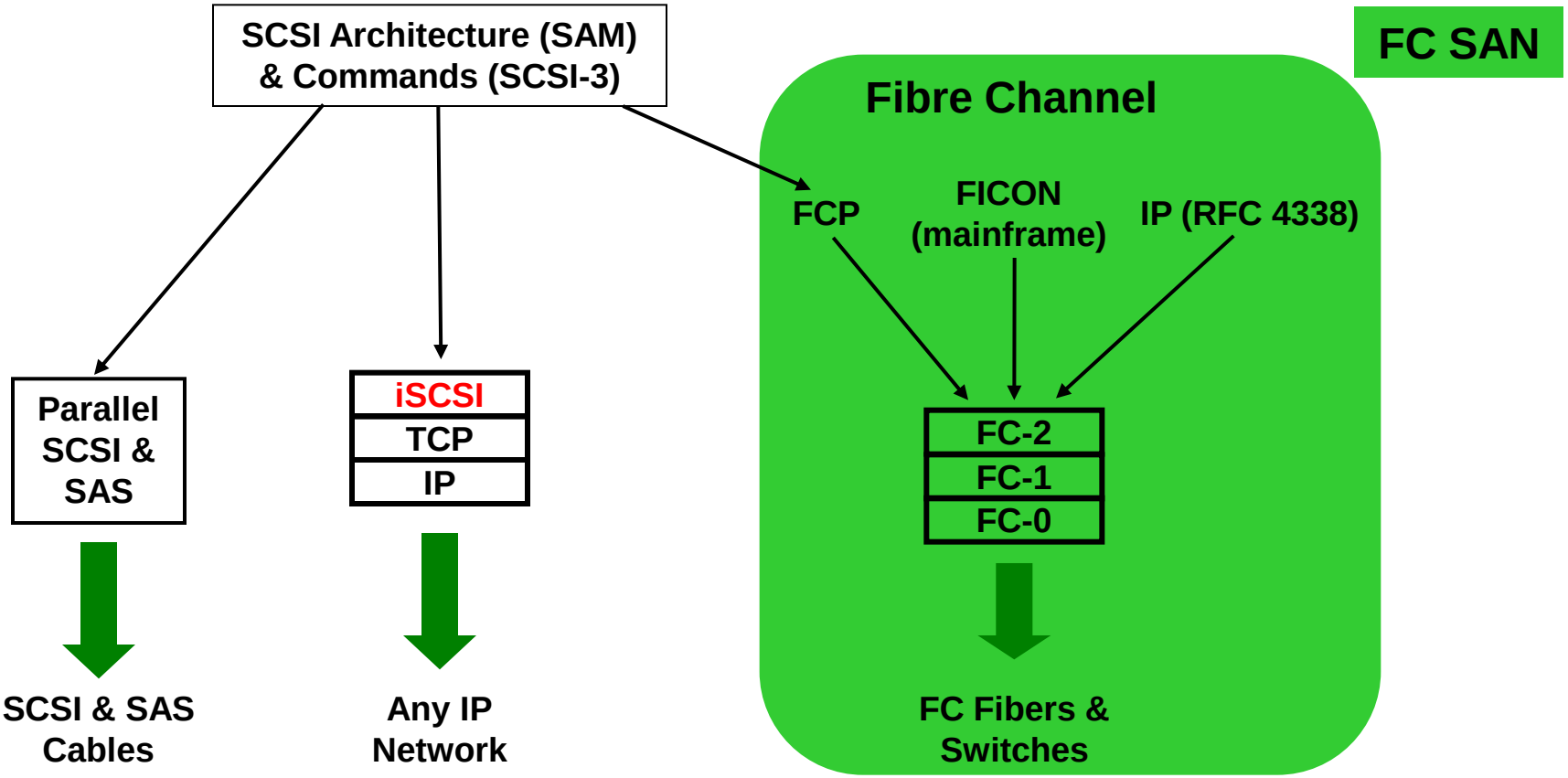


Shared storage enables a VM to move without moving its data

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.

Native Fibre Channel 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)
  - Buffer credit required to send (separate credit pool per direction)
  - FC link control operations return buffer credits to sender for reuse

EMC² where information lives™

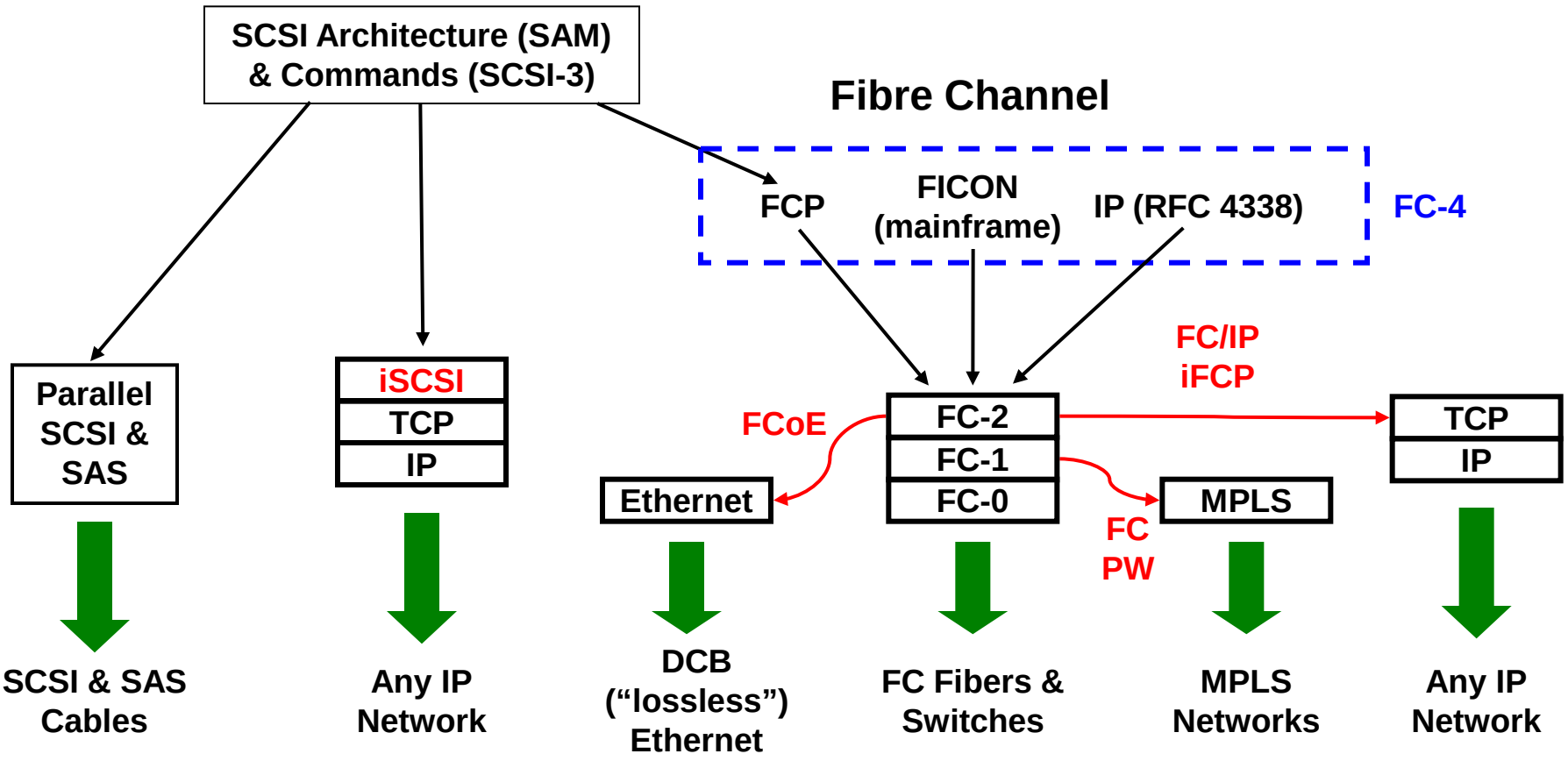
© Copyright 2010 EMC Corporation. All rights reserved.

FC timing and error recovery

- Strict timing requirements
  - R_A_TOV: Deliver FC frame or destroy it (typical: 10sec)
  - Timeout budget broken down into smaller per-link timeouts
- Heavyweight error recovery: No reliable FC transport protocol
  - Disk error recovery: Retry entire server I/O
    - 30sec and 60sec timeouts are typical.
  - Tape error recovery: Stop, figure out what happened and continue
    - Streaming tape drive stops streaming (ouch!).
- FC is *very* sensitive to drops and reordering
  - Congestion: Overprovision to avoid congestion-induced drops
  - Reordering: Needs to be avoided
    - FC receivers may reassemble a few reordered frames
    - More than a few reordered frames: FC receiver can't cope, drops them

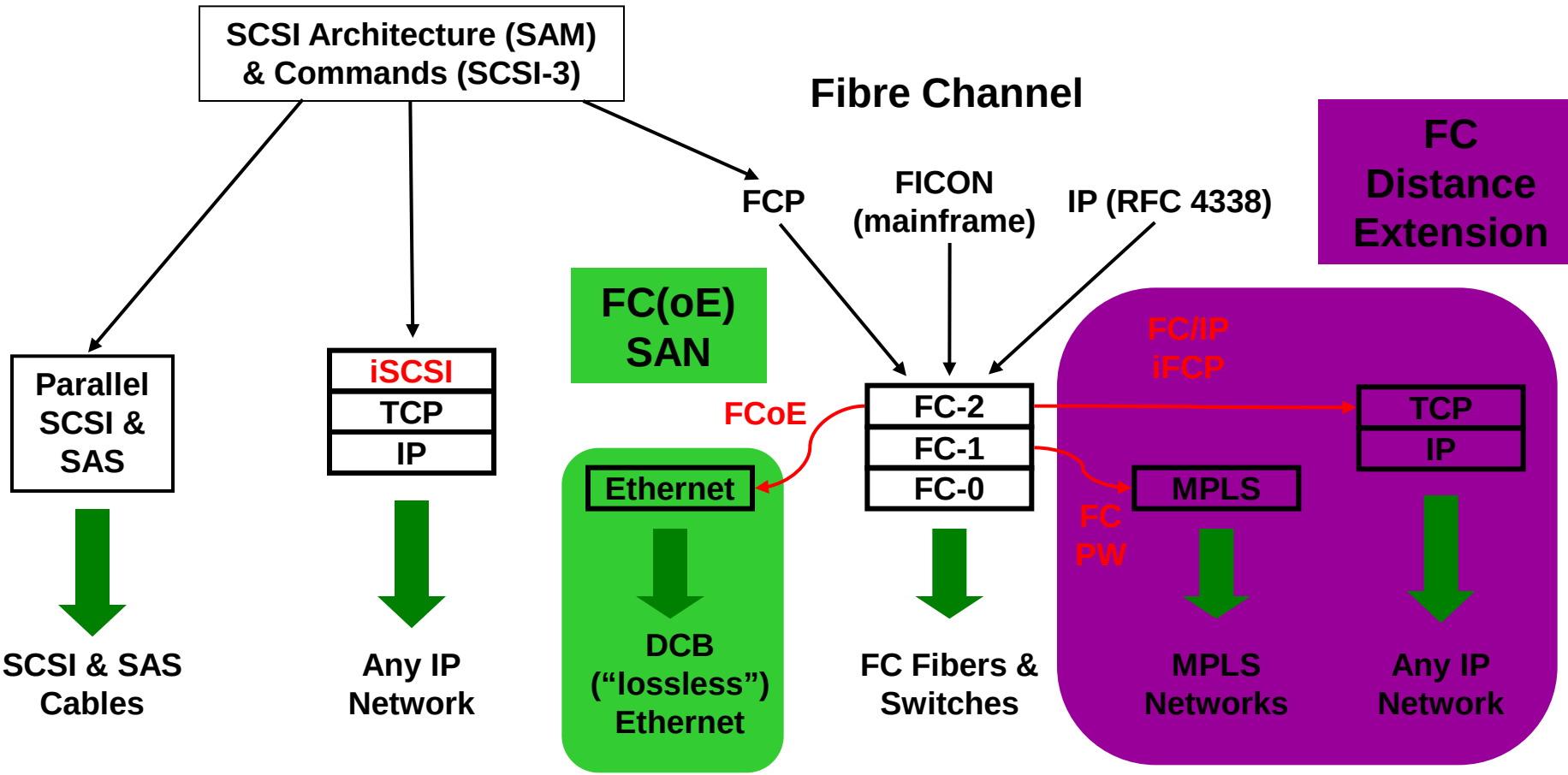
EMC²
where information lives™
© Copyright 2010 EMC Corporation. All rights reserved.

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 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
  - FC link control: separate packets
  - 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
  - IETF Last Call recently completed (draft-ietf-pwe3-fc-encap-15)
  - 16GFC (in development) uses 64b/66b encoding

EMC² where information lives™

© 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 (RFC 3821 & RFC 3643)
  - Only used for FC distance extension
- iFCP: More complex specification (RFC 4172 & RFC 3643)
  - FC distance extension: iFCP address transparent mode
  - iFCP not used for connection to servers or storage
- iFCP is going away (being replaced by FC/IP in practice)
  - iFCP update (remove unused translation mode): RFC 6172 (storm WG)

EMC² where information lives™

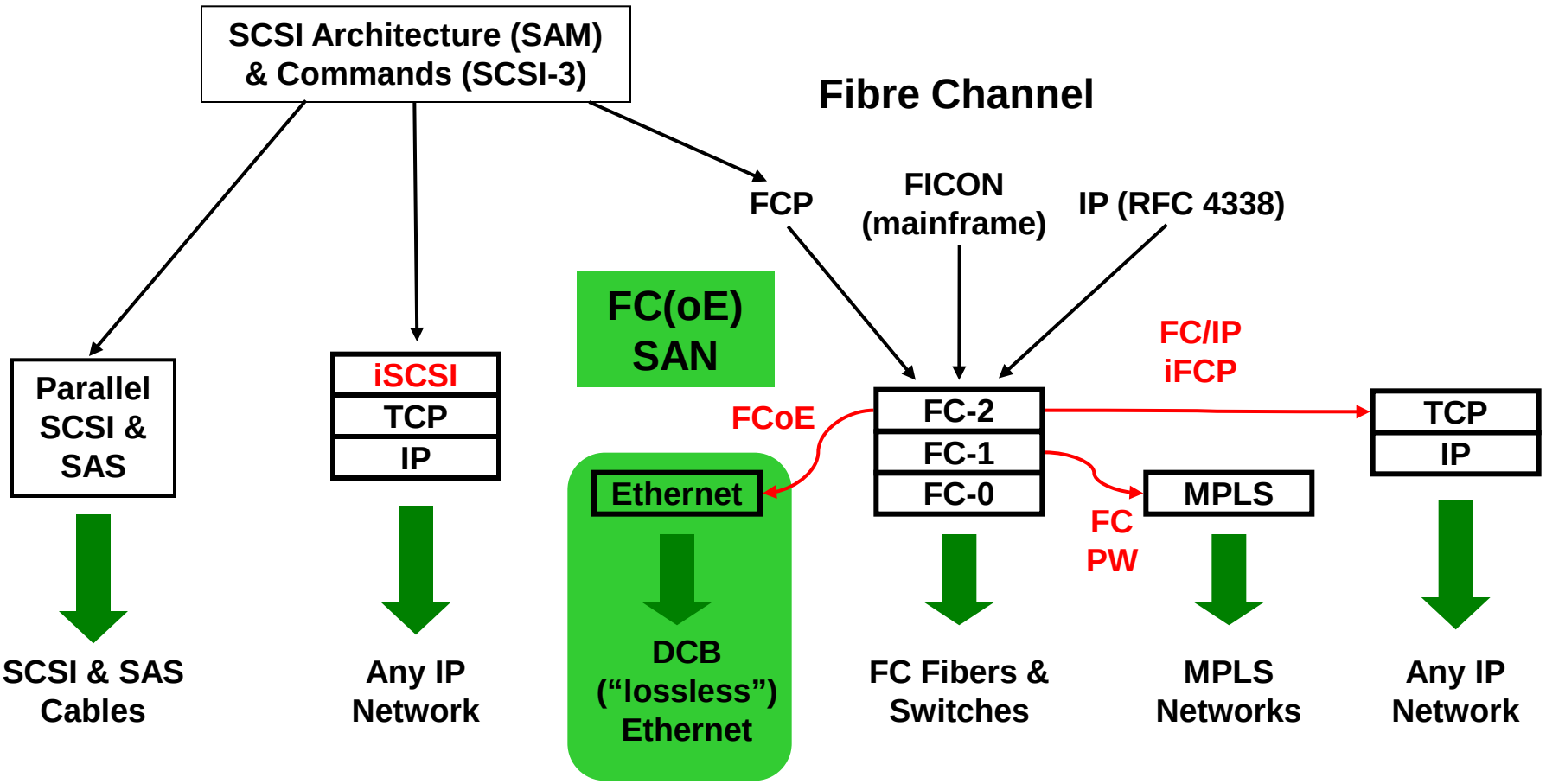
© Copyright 2010 EMC Corporation. All rights reserved.

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.

The SCSI Protocol Family and Fibre Channel



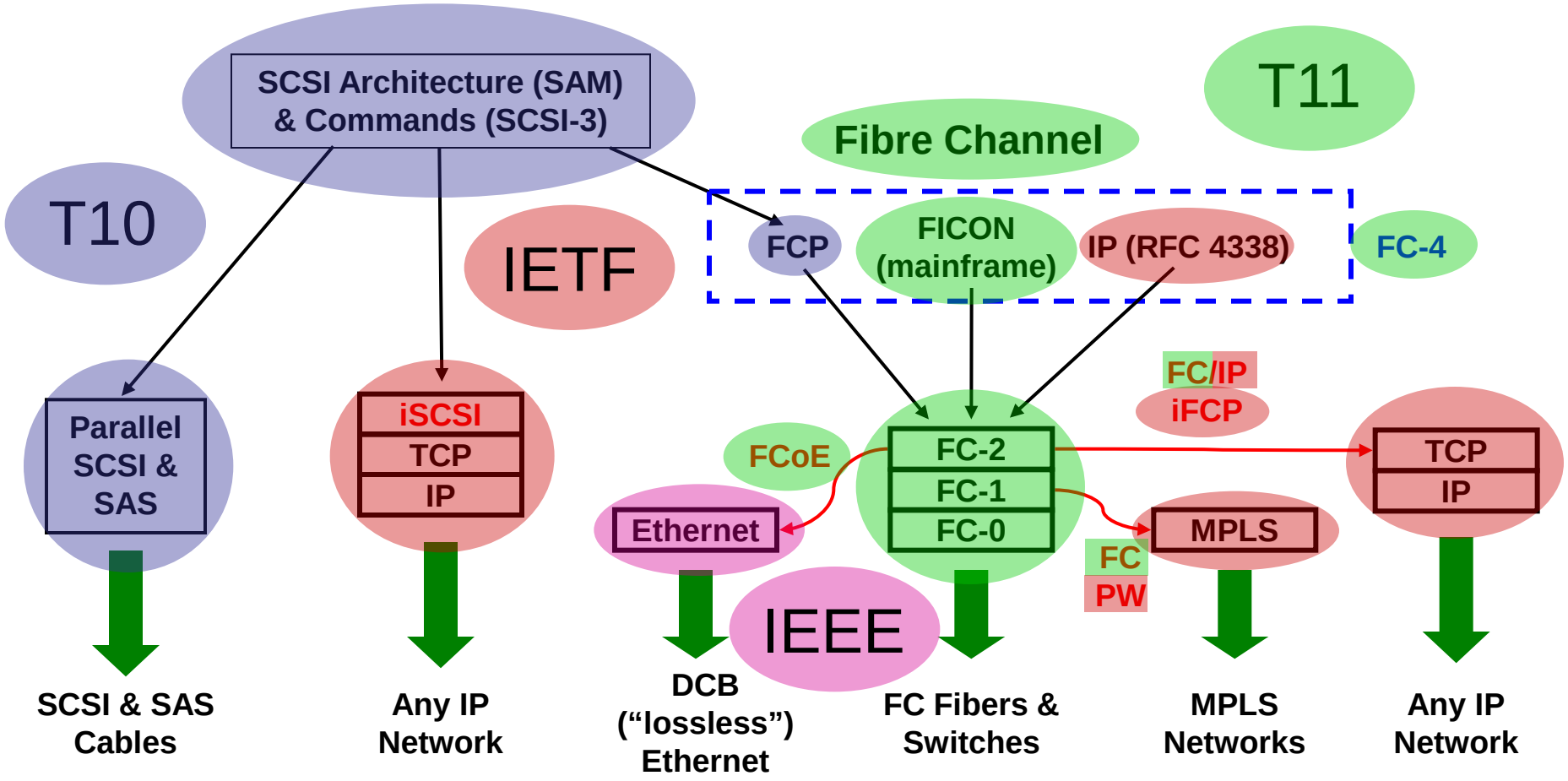
© Copyright 2010 EMC Corporation. All rights reserved.

FCoE: Fibre 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.

The SCSI Protocol Family and Standards Orgs.



© Copyright 2010 EMC Corporation. All rights reserved.

THANK YOU

© Copyright 2010 EMC Corporation. All rights reserved.