IP Block Storage Protocols

David L. Black
EMC Corporation

IETF IP Storage Working Group
Co-Chair

April, 2001

IETF and the IP Storage Working Group

- IETF = Internet Engineering Task Force
  - Internet engineers: IP, TCP, UDP, etc.
  - "Rough Consensus and Running Code": Interoperability emphasis
  - Standards documents: RFCs (Request For Comments)
- IPS WG = IP Storage (ips) Working Group
  - Block storage over IP, based on existing protocols (SCSI, FC)
  - About 9 months old
  - Chairs: David L. Black (EMC), Elizabeth Rodriguez (Lucent)
- Getting Involved
  - Join the mailing list: `ips-request@ece.cmu.edu`
  - Read the drafts - linked to IPS WG charter on www.ietf.org
  - Come to our meetings

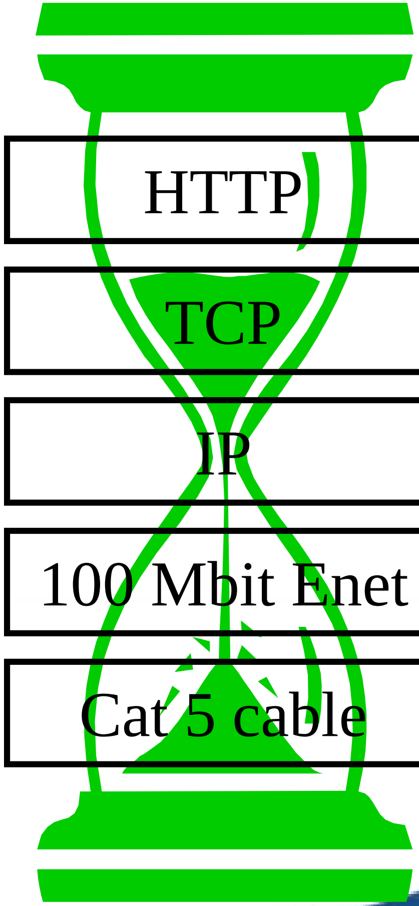
Outline

- Background
  - Network and Fibre Channel layers
  - SCSI concepts
- SCSI-based protocol
  - iSCSI
  - IP network with SCSI as a network service
- Fibre Channel-based protocols
  - FCIP and iFCP
  - FC fabric (or lookalike) using IP connectivity

IP Network Layers

7 - Application	Application	Web Browser
6 -Presentation	Data Formats	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 5 cable

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 5 cable

Fibre Channel Layers

FC-4 (ULP)	Upper Layer Protocols FCP (SCSI), VI, FICON, IP, etc.
🚫 FC-3	~~Common Services~~
FC-2	Frames and signaling protocols
FC-1	8b/10b coding and protocol
FC-0	Wire/Fibre and Transceivers

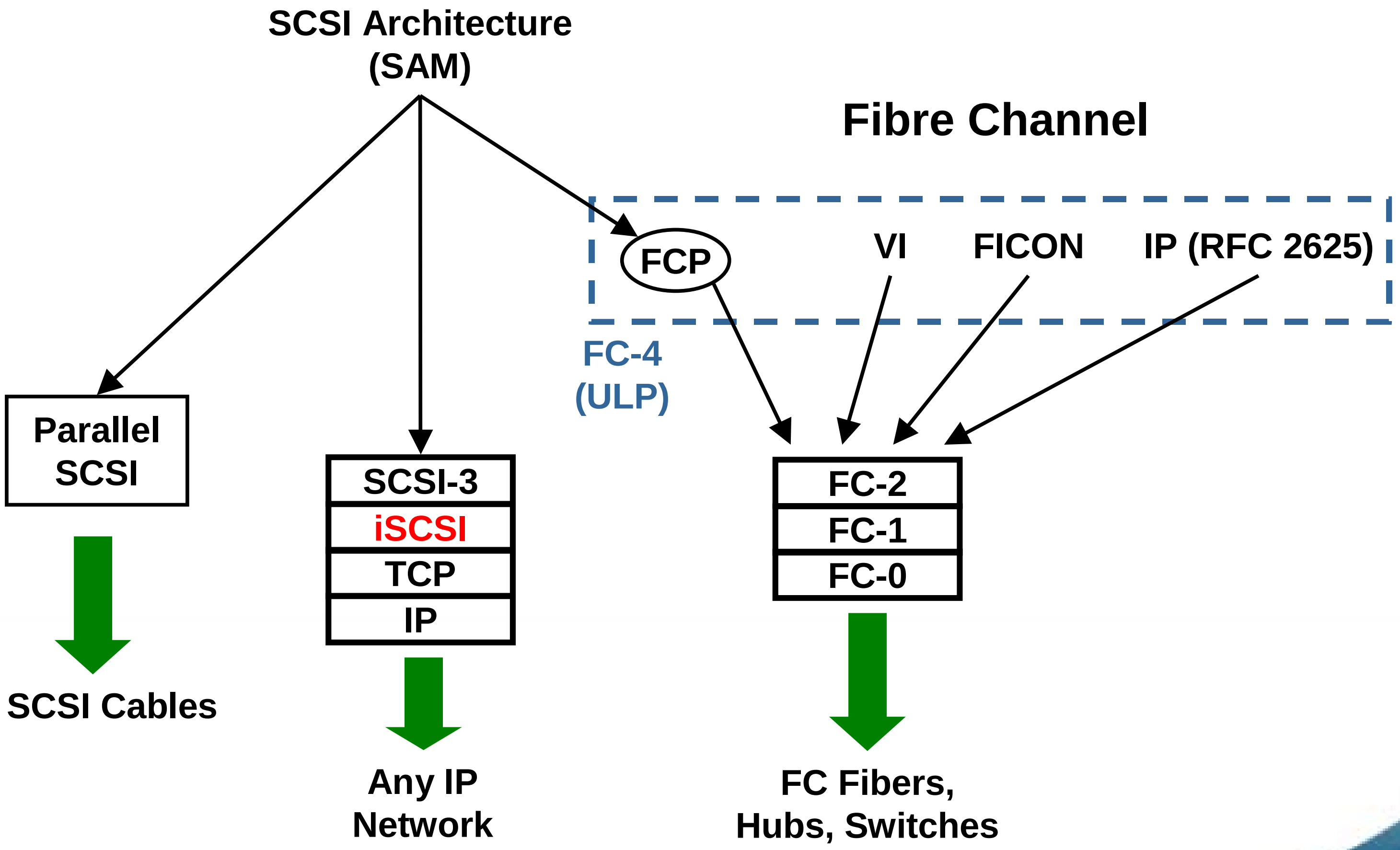
SCSI Concepts

- *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
  - "Bus walker" finds accessible targets

Outline

- Background
  - Network and Fibre Channel layers
  - SCSI concepts
- SCSI-based protocol
  - iSCSI
  - IP network with SCSI as a network service

iSCSI and Protocol Stacks



iSCSI Concepts

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

iSCSI Error Handling

- Sequence numbers detect missing things
  - Commands, responses, data blocks
  - Goal: Avoid SCSI retry if at all possible
    - Explicit iSCSI command retry can be used
- CRC: Work in progress
  - 32-bit CRC polynomial in current draft (not using IEEE CRC-32)
    - Defining a new 64 bit CRC considered and rejected
  - Separate CRCs computed over header and data
- Multiple Initiator support
  - AutoSense is mandatory
  - Auto Contingent Allegiance should be implemented

iSCSI Naming

- Rationale
  - Targets may share <IP address, TCP port>
  - Initiators and Targets may have multiple IP addresses
  - Unique names are important for third party commands
- Two types of globally unique names:
  - WWNs (EUI)
  - Reversed hostname (DNS) as naming authority
  - OUI and forward DNS being removed from draft
- New nameserver protocol: iSNS
  - Source of name to <IP address, TCP port> bindings
- World Wide Unique Identifiers: REJECTED
  - New global naming abstraction not needed

iSCSI Security Requirements/Goals

- Authentication: Who are you? Prove it!
  - Mutual Authentication: Initiator to Target **AND** vice-versa
- Integrity: Have these bits been tampered with?
  - Cryptographic integrity, not just checksum or CRC
  - Must be linked to authentication to prevent regeneration attack
- Authorization: What are you allowed to do?
  - iSCSI: Controls who can connect to which Target
  - LU, LUN and/or volume authorization is a SCSI issue, not iSCSI
- Confidentiality: Has this data been disclosed?
- MUST implement Authentication and Integrity

iSCSI Security

- Secure IP connection prior to iSCSI Login:
  - Integrity, authentication, and optionally confidentiality
  - Will use IPSec or TLS (SSL successor)
    - Have not decided which one, yet
- Inband authentication
  - SRP and Kerberos in current draft
  - Public key and Radius mechanisms will be added
  - Kerberos-based integrity checks (if Kerberos is used)
- Security work is still in progress
  - How does IPSec or TLS authentication relate to iSCSI names?
  - Which mechanisms MUST be implemented?

iSCSI and Framing

- iSCSI is a message-based protocol
  - Header indicates message length
- TCP is a byte-stream protocol
  - No message or record boundaries
  - Packet boundaries may not match iSCSI messages
- Suppose the network drops an iSCSI header
  - TCP will retransmit, eventually
  - But there's data in flight that iSCSI can't parse
    - How long is the missing message?
    - Where does all that data get buffered?

More iSCSI and Framing

- What if a header CRC check fails?
  - TCP won't retransmit
  - Where does the next header start?
    - Can we avoid closing the TCP connection?
- iSCSI includes a general "interface" to framing
  - Actual framing mechanisms are optional
  - Can be negotiated via iSCSI
- Possible Framing Mechanisms
  - TCP framing: draft-williams-tcpulpframe-01.txt
  - Markers: appendix of iSCSI draft
  - Word-stuffing: status and interest unclear

iSCSI Discovery and Boot

- Discovery - How does the "bus walker" find targets?
  - Static configuration (simple, small scale)
  - SLP for intermediate scale discovery
  - iSNS for larger scale and zoning support
    - SLP can discover iSNS server
- iSCSI bootstrap support: Mostly a variant of discovery
  - Have to discover the boot target (and the boot device)
  - Most practical boot issues are implementation (BIOS, etc.) rather than protocol specification issues.

Fibre Channel-Based Protocols

- FC Fabric (or lookalike) using IP connectivity
  - Little to no interest in Fibre Channel Arbitrated Loop (FC-AL)
- Tunnel
  - FCIP - IP network is transparent to FC fabric
- Gateway
  - iFCP - IP network implements FC fabric
- But first, some Fibre Channel background ...

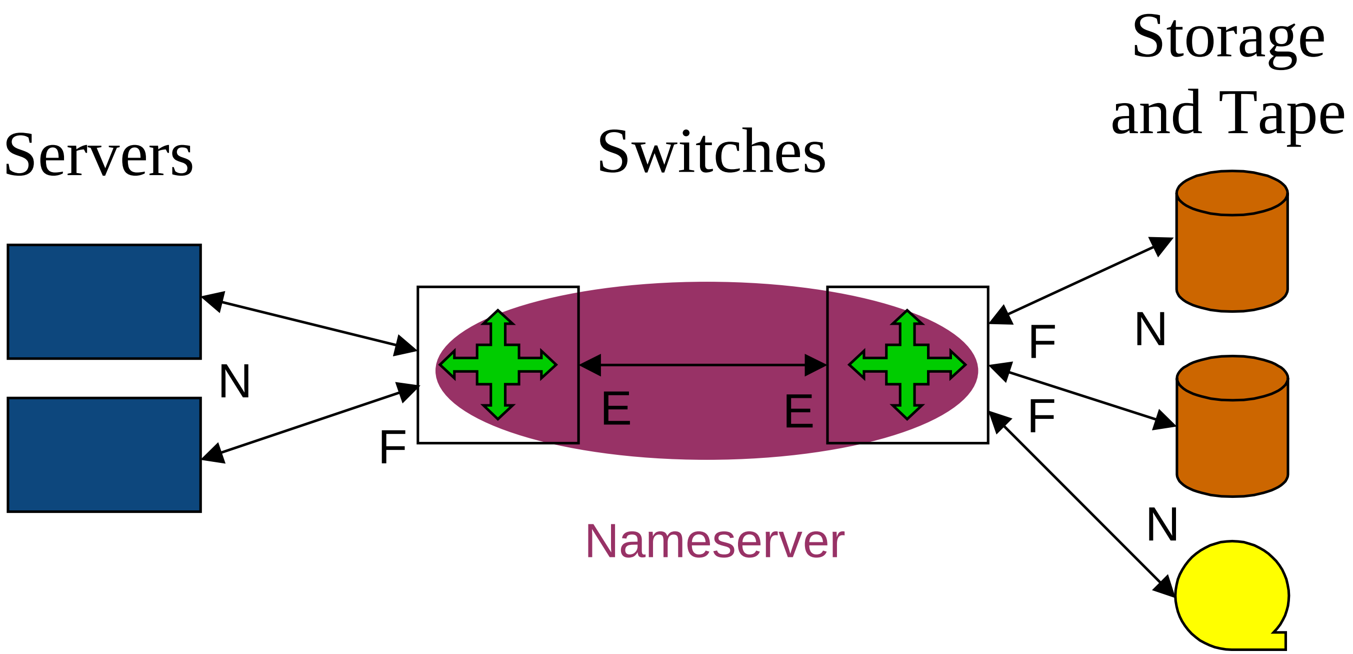
Fibre Channel Fabric Port Types

- Devices (host or server HBAs, storage targets)
  - ***N_Port***: "Node"
- Switch ports connected to devices
  - ***F_Port***: "Fabric"
  - N_Port must be connected to an F_Port and vice-versa
- Switch ports connected to switches
  - ***E_Port***: "Extension"
  - E_Port must be connected to an E_Port
- Additional port types exist (L, G, B, etc.)

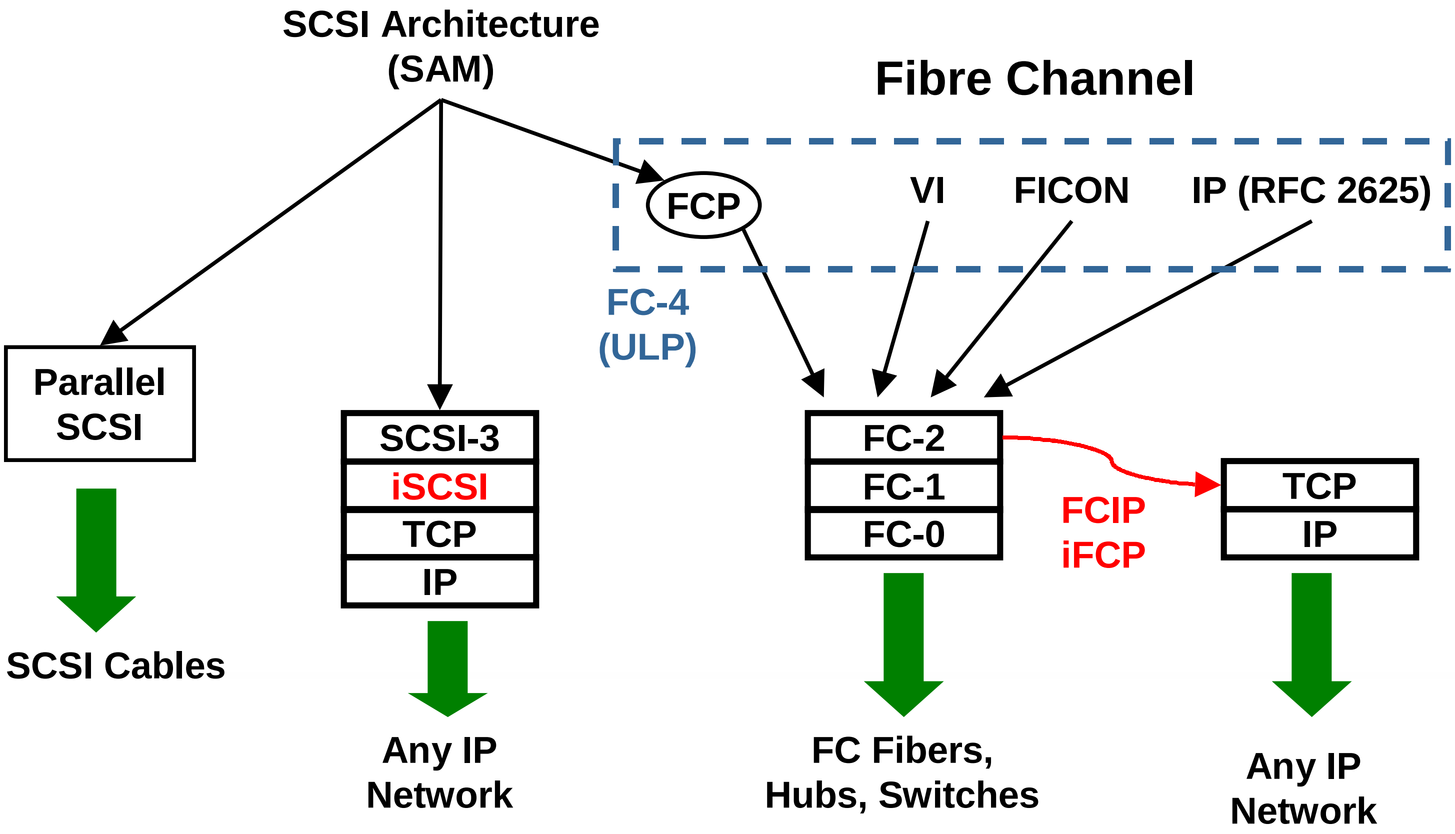
Fibre Channel Fabric Operation

- The "Fabric" is a visible first-class entity
- Fabric Login
  - Node logs into fabric as part of initialization
- Fabric Nameserver
  - Integrated into switches (fabric service)
  - Stores <name, address> records based on logins
    - <64-bit WWN, 24-bit S_ID/D_ID>
- Discovery: Node downloads nameserver info
  - Soft Zoning: limit info given to each node
- Extensive switch to switch communication

Fibre Channel Fabric Example



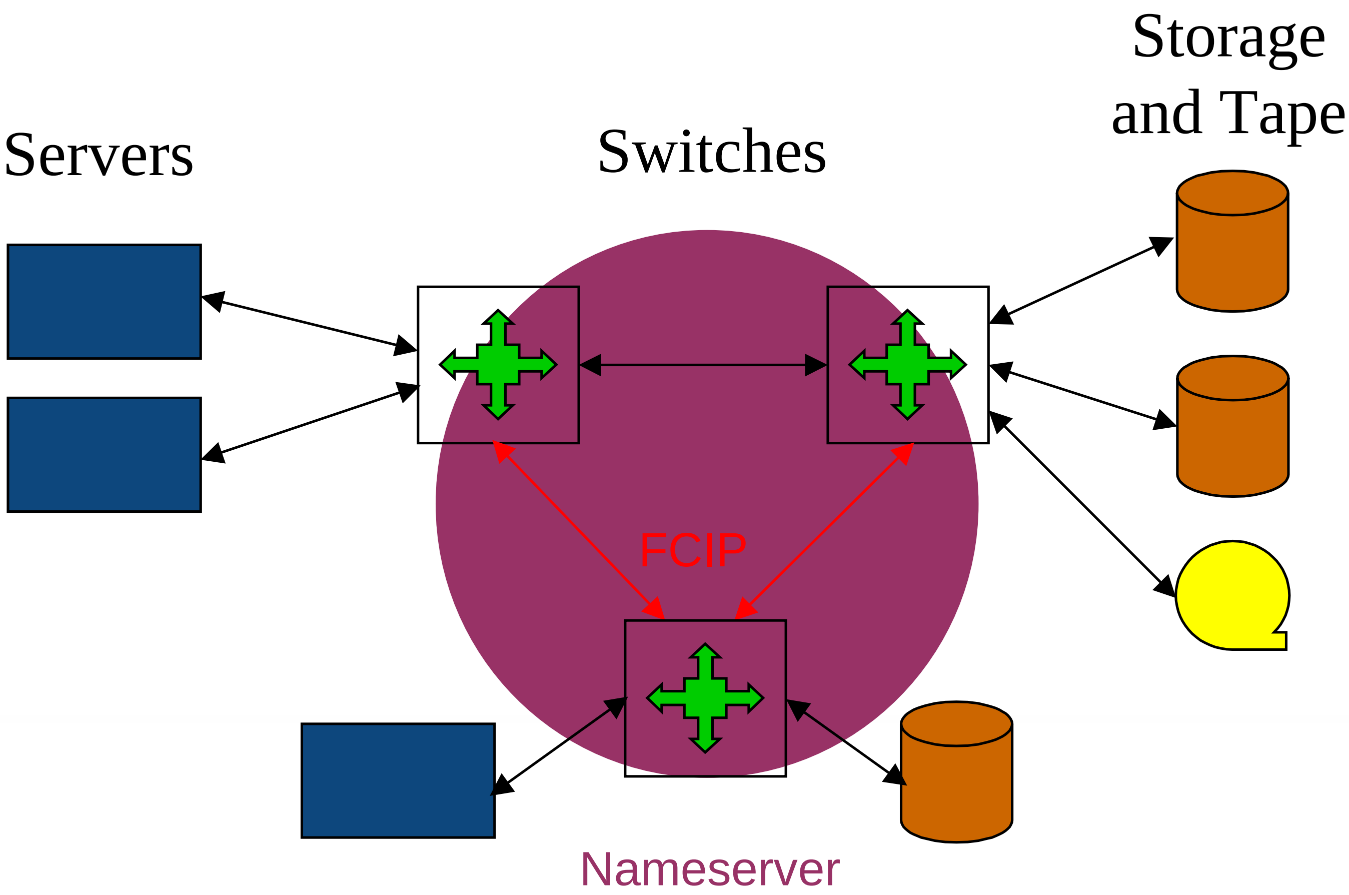
FCIP, iFCP and Protocol Stacks



FCIP: Fabric Interconnection Protocol

- Situation
  - Two Fibre Channel fabrics
  - Want to use IP to connect them and form a single fabric
- Solution: FCIP
  - Tunnel Fibre Channel through an IP network
  - Encapsulate FC-2 Frames in TCP/IP
- FCIP connection is transparent to switches
  - Looks like a pair of connected E_Ports
  - Fabric services run transparently
- iSNS can help with tunnel setup

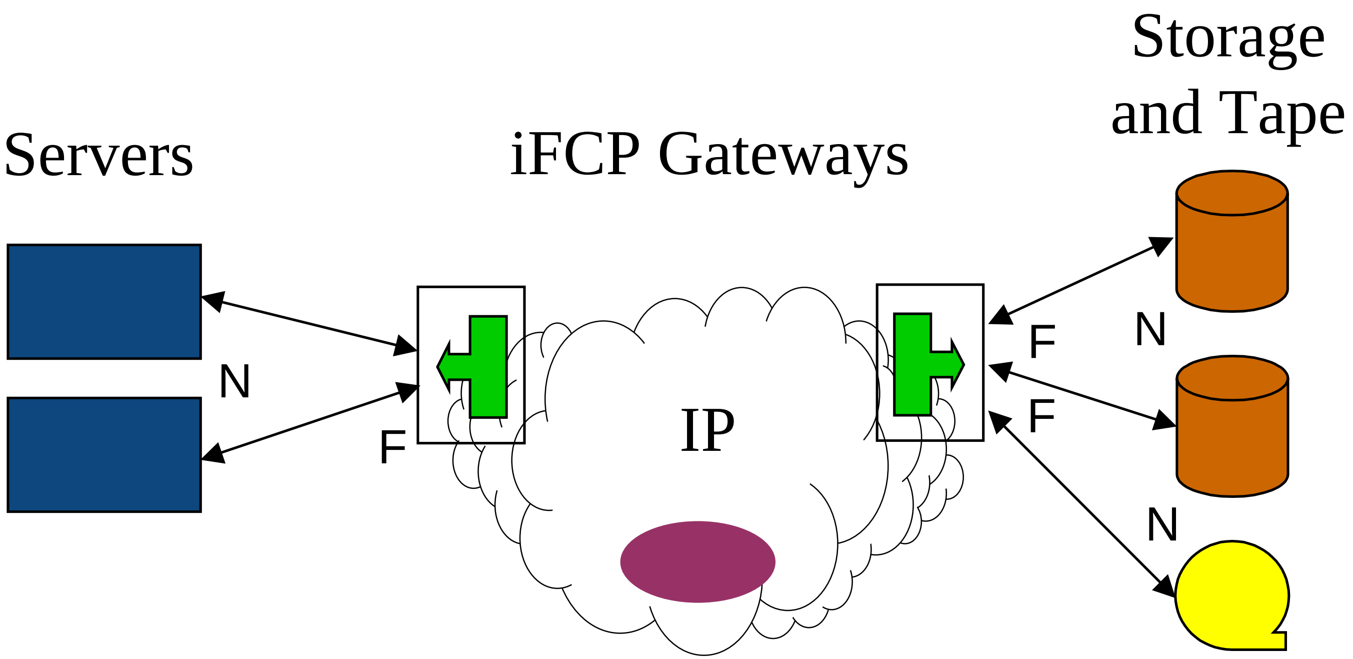
FCIP Fabric Example



iFCP: Fibre Channel Gateway Protocol

- Situation
  - Fibre channel devices (no switches)
  - Want to attach devices to an IP network
- Solution: iFCP gateways + iSNS
  - iFCP Gateway implements F_Port
    - N_port devices attach to gateway
    - Gateway implements interface to "fabric"`
  - iSNS server implements fabric nameserver
    - Gateway registers with iSNS on fabric login
  - Discovery info obtained from iSNS server by the gateway
    - Soft zoning implemented by iSNS server

iFCP Example



Servers

iFCP Gateways

Storage and Tape

iSNS Nameserver

FCIP and iFCP Addressing

- FCIP: No changes to Fibre Channel addressing
- iFCP: Gateway is an address translator
  - Translates 24-bit FC addresses (e.g., S_ID, D_ID) to
    - IP address of remote gateway +
    - 24-bit address of N_Port beyond gateway
  - Translation set up by two types of events
    - Information retrieval from iSNS
      - iSNS info includes IP address information
    - Encapsulated frame arrives from new source

FCIP/iFCP work in progress

- Both protocols encapsulate FC-2 frames
  - Common encapsulation format under development
  - Information in protocol headers will differ
- Robustness improvements to encapsulation
  - Header CRC
  - Re-sync to data stream after Header CRC failure
- Time outs
  - Prevent network delays from violating FC Time Out Values
  - Timestamp transmitted frames and discard stale ones
- Security - may follow iSCSI direction

iSCSI/FCIP/iFCP Status and Timetable

- Active work underway on all three protocols, e.g.,
  - Protocol header format changes
  - Determining required security mechanisms
- Next IETF IP Storage Working Group meetings:
  - April 30, May 1 in Nashua, NH (at T10 meetings)
    - Plus a discussion on structure of a SCSI MIB in May 2 T10 CAP meeting
  - Week of August 6 in London, UK (at IETF meetings)
- Current timetable: Finish main specifications in September
  - IPS WG charter will be revised over the summer
  - Milestones may change at that time

IP Block Storage Protocol Summary

- SCSI-based Protocol
  - iSCSI
  - IP network with SCSI as a network service
- Fibre Channel-Based Protocols
  - FC SAN (or lookalike) using IP connectivity
  - Tunnel
    - FCIP - IP network is transparent to FC fabric
  - Gateway
    - iFCP - IP network implements FC fabric
- iSNS nameserver for storage
  - Applies to all three protocols, required only by iFCP

Standards Organizations

- SCSI: T10
  - www.t10.org
- Fibre Channel: T11.3
  - 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
  - Co-chairs: *David Black (EMC),* Elizabeth Rodriguez (Lucent)
- Active coordination on overlapping matters