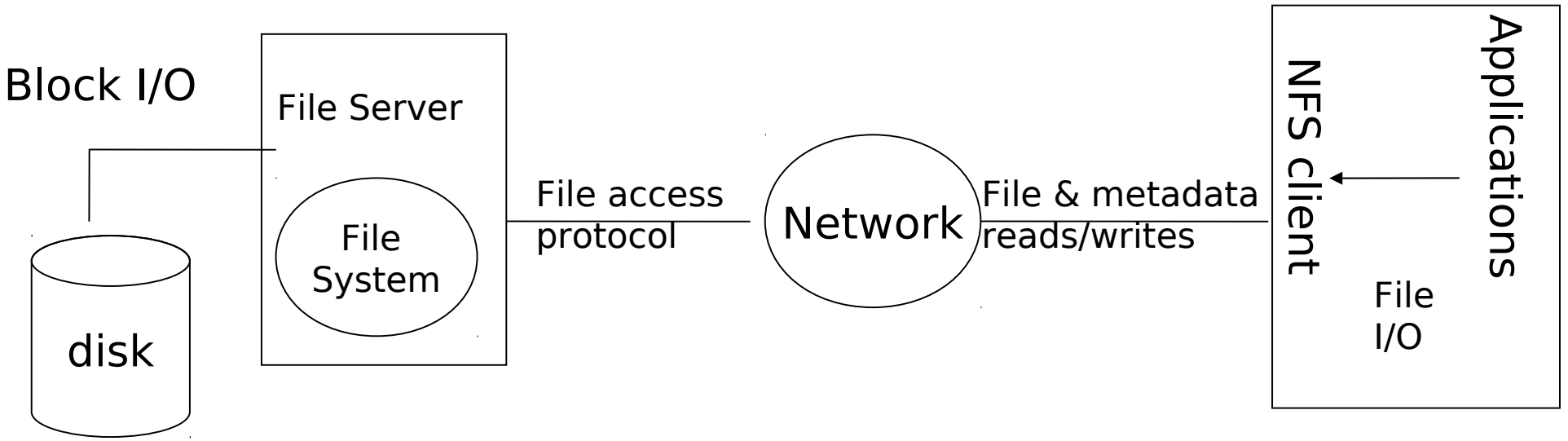
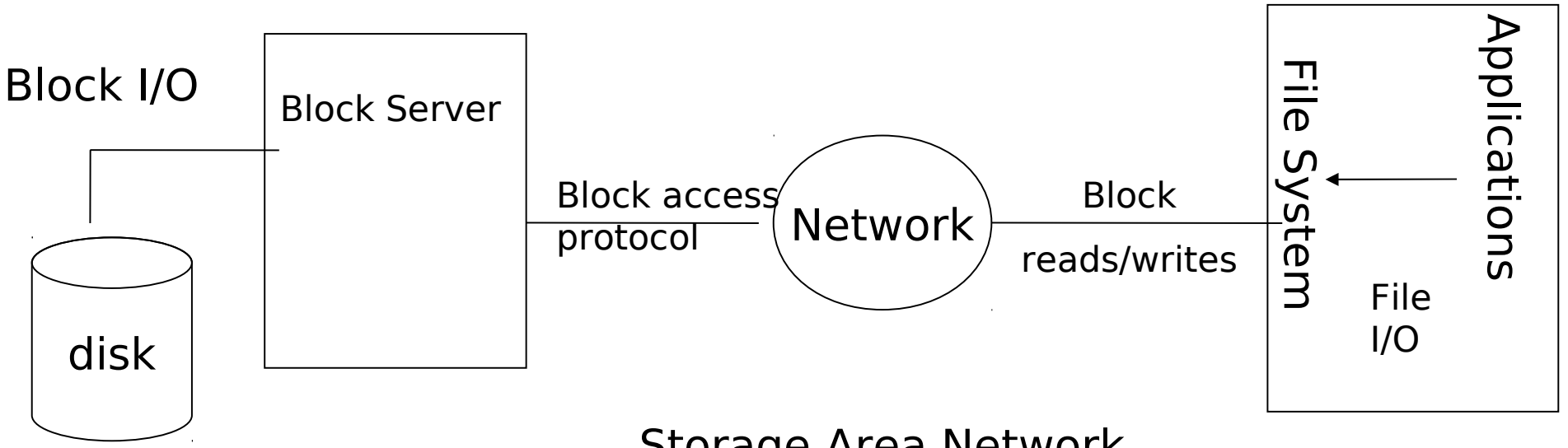
# Storage Systems

NPTEL Course
Jan 2012
(Lecture 07)

K. Gopinath
Indian Institute of Science




Network Attached Storage (NFS)




Storage Area Network

# NAS vs SAN

- Storage systems scaled to large sizes
  - Network Attached Storage (NAS) - NFS
  - Storage Area Networks (SAN)
    - Fibre Channel (FC)
    - Internet SCSI (iSCSI)
- Unit of Access
  - NAS – file level
  - SAN – block level
- Sharing
  - NAS – supports multiple clients
  - SAN – supports single client
    - If there are multiple clients, applications need to handle issues wrt concurrent accesses

# Modern Storage Protocol Stack

- PHY mostly same: fibre ($L_0$)
- Upper Layer Protocol (ULP): SCSI also same! ($L_n$)
- $L_1..L_{n-1}$ determine the perf
  - SONET
  - Fibre Channel (FC)
  - 10GEth
  - Infiniband
- Both “Telephone” or “Internet” models in storage
  - Telephone model: guaranteed service, preallocation/reservation of BW, etc (FC/Infiniband)
  - Internet model: statistical multiplexing (10GEth)

# Flow Control at Various Levels

- Link layer:
  - GigE MAC Control Sublayer (pause)
  - FC credit-based
  - Infiniband credit-based
- Transport layer: TCP
- Application: SCSI level

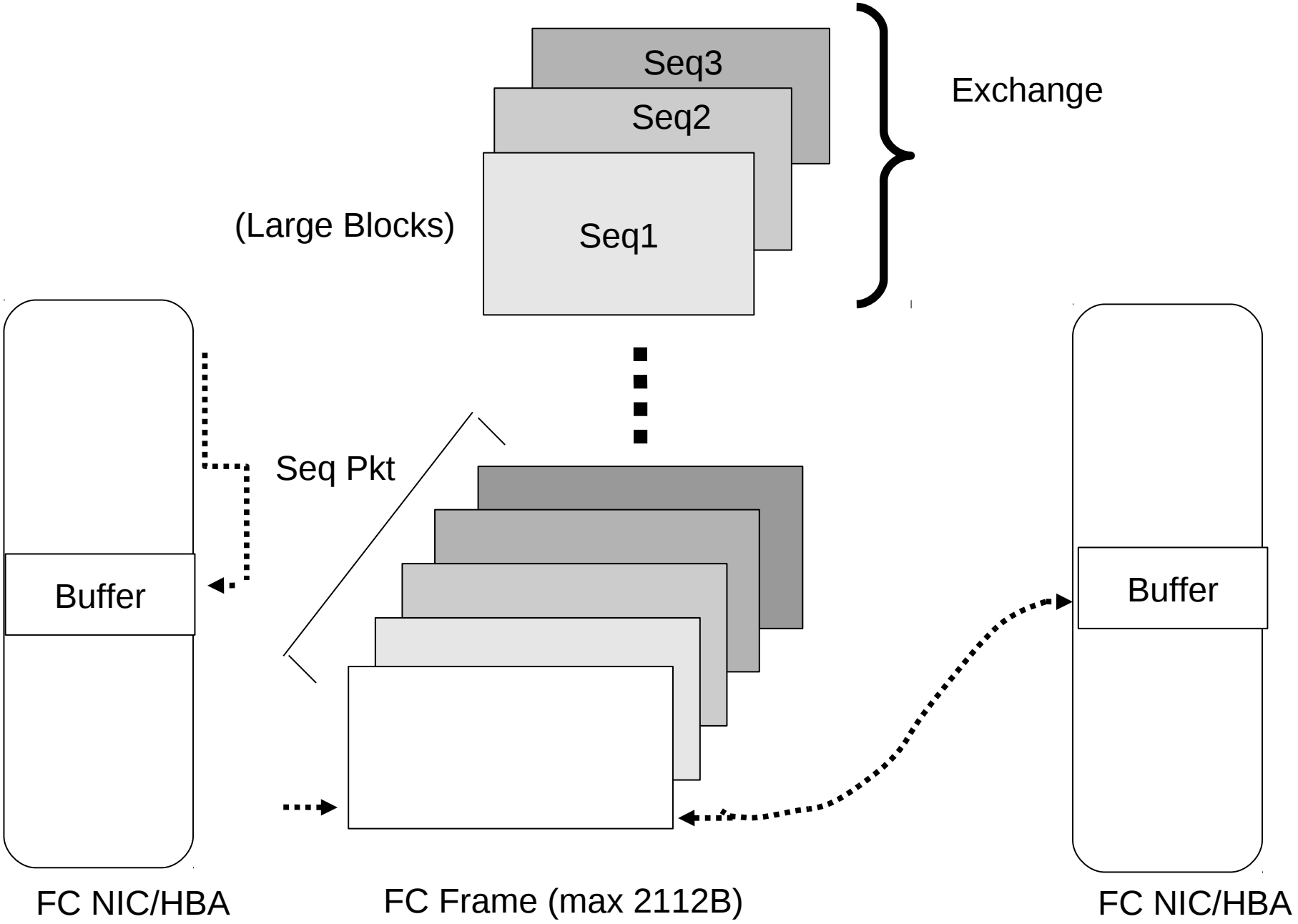
# SCSI and Fibre Channel

- SCSI: protocol to access data in a SAN
- SCSI: block level protocol, conventional flat cable used to connect disks to HBA
- SCSI alone unsuitable for SAN due to distance limitation
- Solution: Fibre Channel, iSCSI as the protocol to encapsulate SCSI

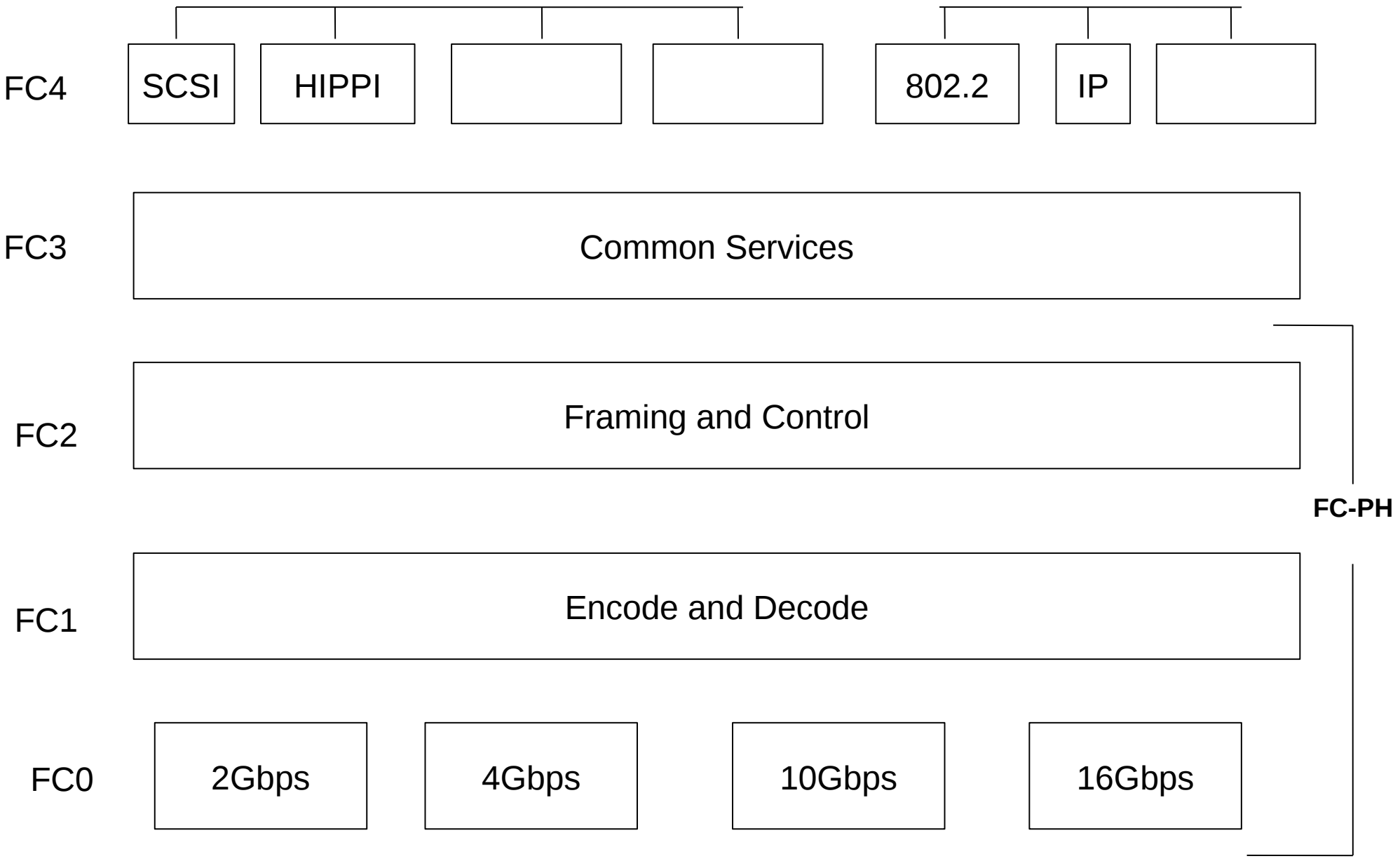
# Fibre Channel

- a frame-based protocol
  - Groups of frames in one direction: sequence
  - In both directions: exchange
- zero-copy send and receive (remote direct memory access [RDMA]) semantics
  - FC NIC keeps track of memory loc of buffer of receiver
    - Calculates new loc after each frame received successfully
    - No loss of synch with loss of a frame as cmds fully specified
  - Low host CPU utilization
  - Reduces memory requirements of Fibre Channel adapters for gigabit wire speeds.
- uses credit-based congestion control

# Data Transfer in FC



# SCSI and Fibre Channel

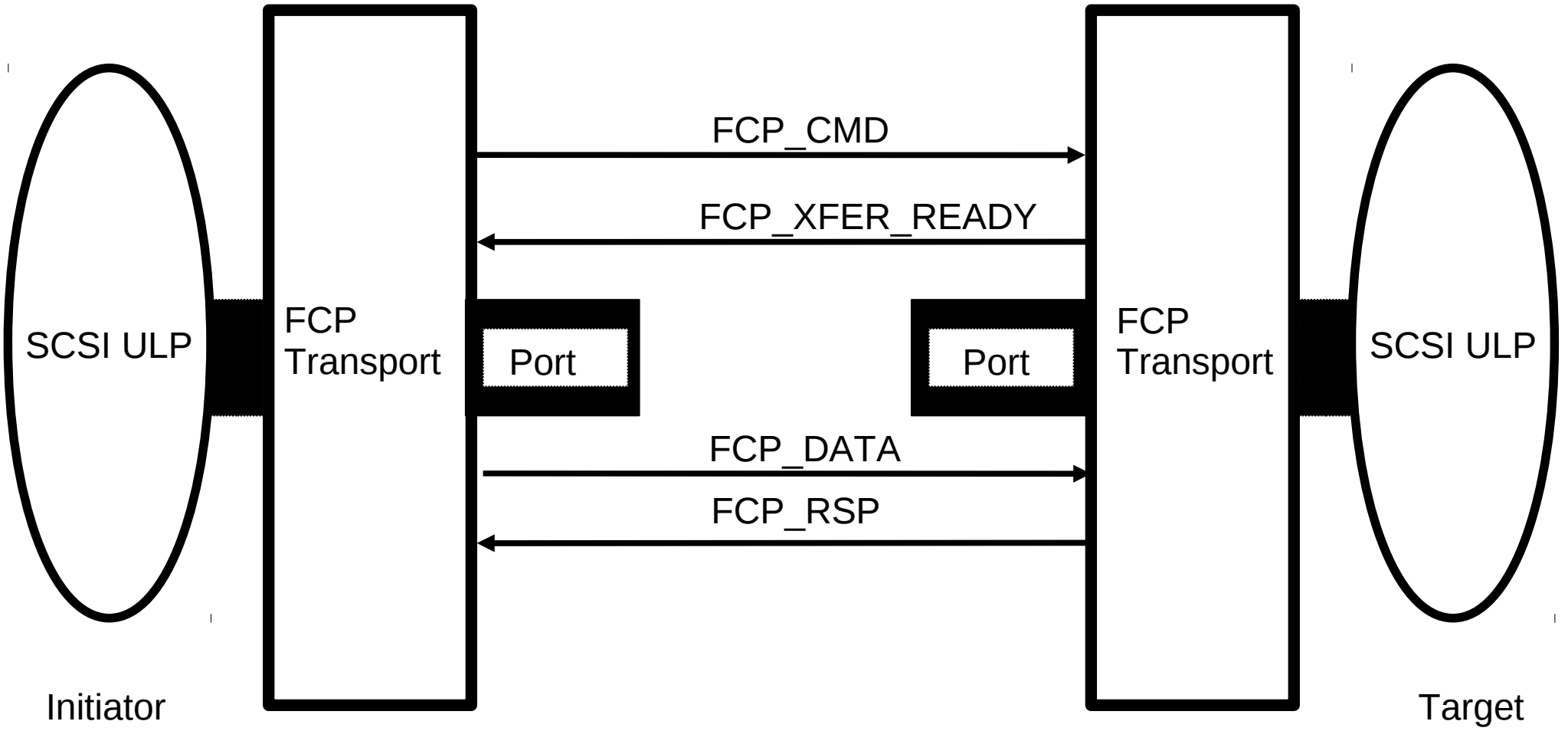


# FC Protocol for SCSI

- Defines ULP Mapping to Send SCSI Information
- Defines Data Information Units
  - FCP_CMND (unsolicited command) (eg. WriteCmd)
  - FCP_XFER_RDY (data descriptor) (eg. Ready2Transfer)
  - FCP_DATA (solicited data) (eg. WriteData)
  - FCP_RSP (command status) (eg. Status)
- Equates a SCSI IO Operation to an Exchange
- Equates the Associated SCSI Phases to Sequences

| SCSI function | FCP equivalent |
|---|---|
| I/O operation | Exchange |
| Req/Resp primitives | Sequence |
| Cmd service req | Unsolicited cmdIU(FCP_CMND) |
| Data delivery req | Data descr IU(FCP_XFER_RDY) |
| Data delivery action | Solicited data IU(FCP_DATA) |
| Cmd service response | Cmd status IU(FCP_RSP) |

# FC Protocol for SCSI (cont)



# Fibre Channel: Classes of Service

- *Class 1:* Dedicated connection
- *Class 2:* ACK based connectionless service
- *Class 3:* Connectionless and without ACKs
  - Most widely used
- *Class 4:* Virtual connection service; multiplexed
- *Class 6:* Reliable one to many multicast service

# FC-2 Transport Functions

- Flow Control
  - Buffer-to-Buffer Credit
    - Link Level
  - End-to-End Credit
    - Transport Level
  - (Also ULP Level but not at FC-2)
- Communication Models
  - Full Duplex
  - Half Duplex
- Block Management
- Data Reassembly
- Link Services
  - Basic Link Services
  - Extended Link Services
    - Login, Process Login, Discovery, ...

# FC-2 Transport Functions: Classes of Service

- Class 1
  - Supports EE Credit Flow Control
  - No BB Credit Flow Control
  - In Order Delivery Guaranteed
  - Guaranteed Max. Bandwidth Between Two Nodes
- Class 2
  - Supports EE Credit Flow Control
  - Supports BB Credit Flow Control
  - In Order Delivery Not Guaranteed
  - Allows for Better Use of Fabric Link Bandwidth
- Class 3
  - No EE Credit Flow Control
  - Supports BB Credit Flow Control
  - Requires ULP Level Flow Control
  - In Order Delivery Not Guaranteed
  - Allows for Better Use of Fabric Link Bandwidth
  - Added Performance Benefit of No ACKs
- Intermix: Unused Class 1 Bandwidth Used for Class 2 and 3

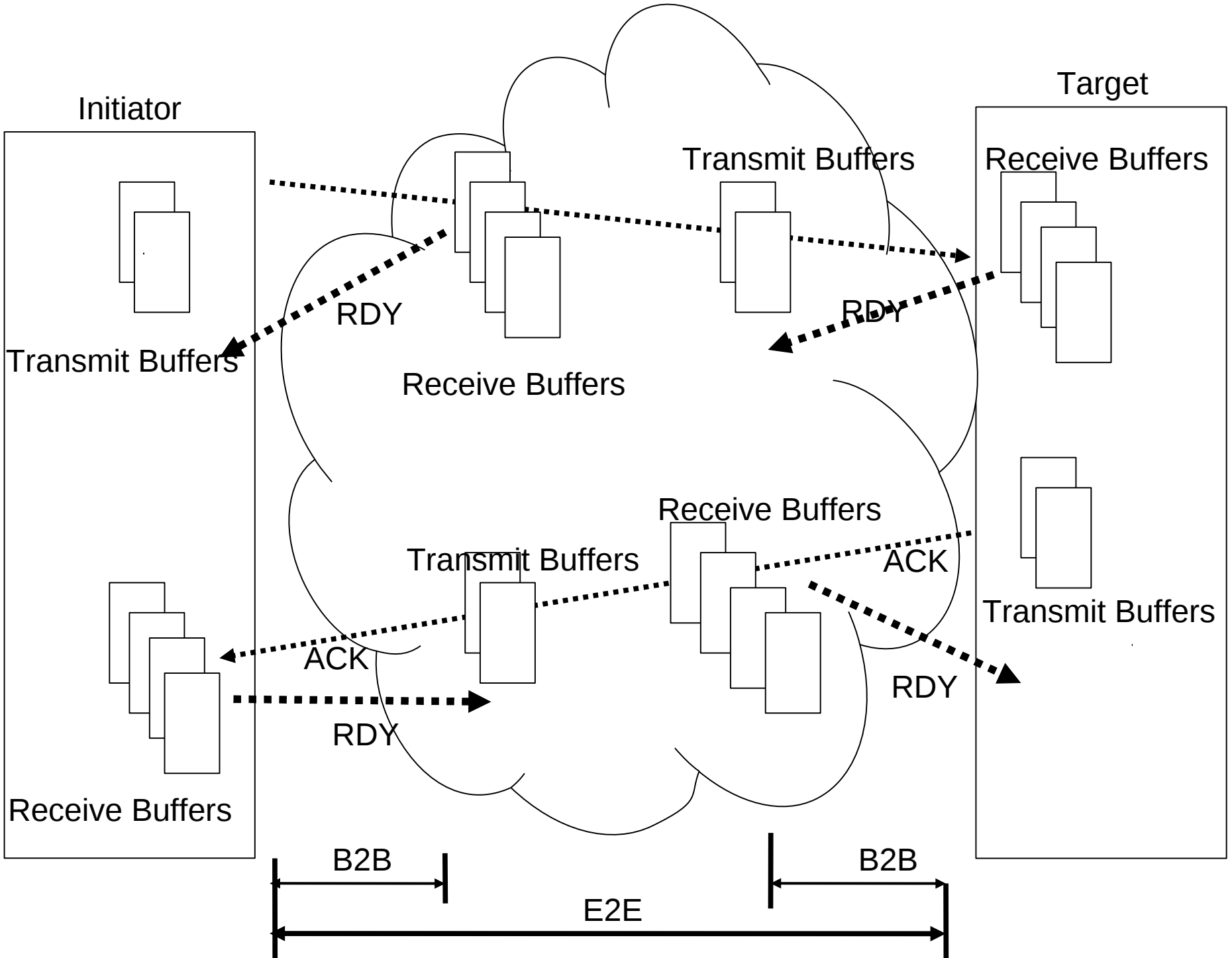
# FC Exception handling

- ULP Level
  - Task and Loop Management
- Transport Level
  - Sequence and Link Service Management
- Link Level
  - Link Management

# FC vs IP

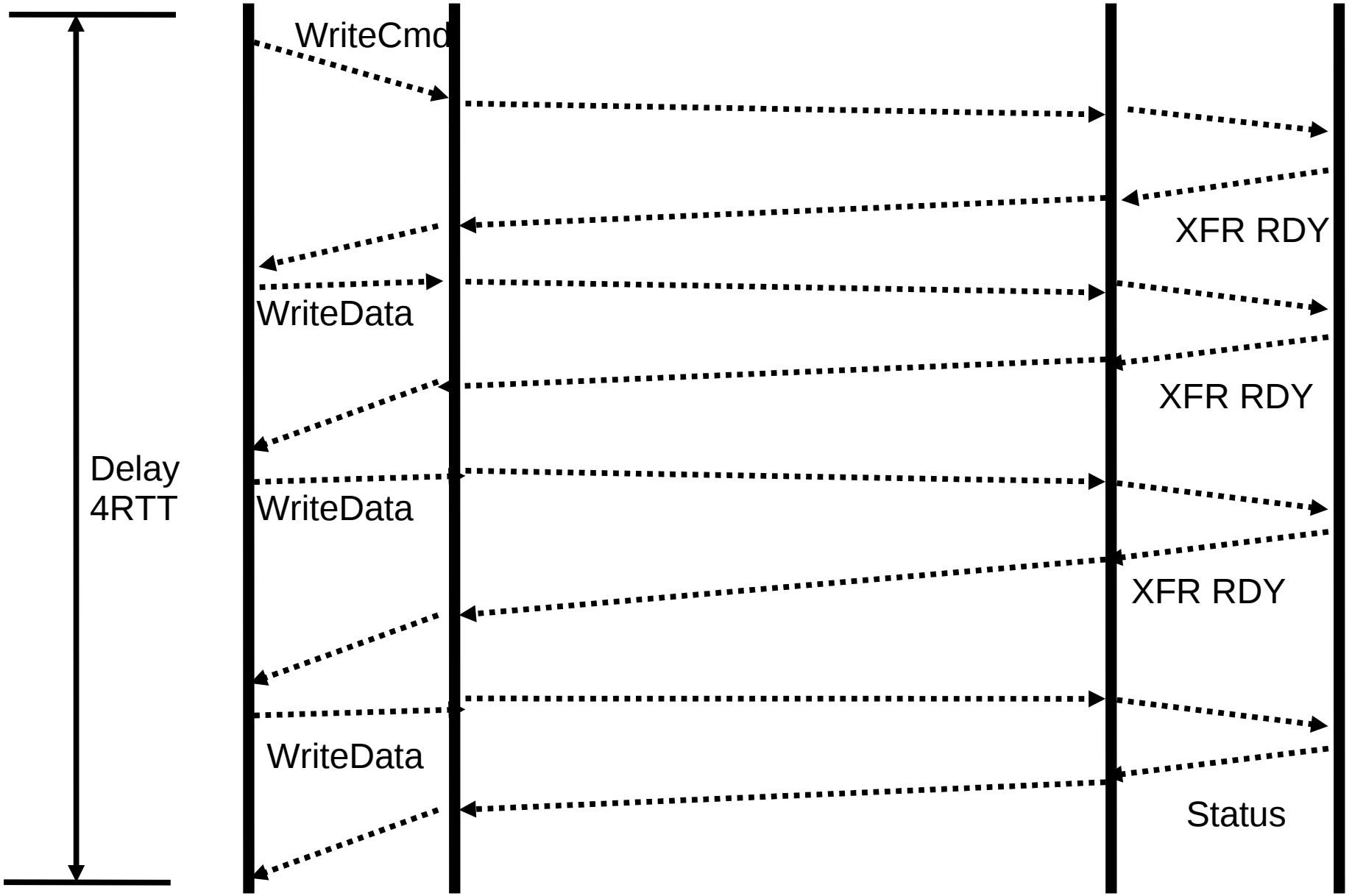
- speed of light in a vacuum is about 186,000 miles/sec (300,000 Km/s); the index of refraction of the typical single-mode fiber optic cable reduces that to about 100,000 miles/sec, or 100 miles/millisecond.
- Fibre Channel switches can tolerate a millisecond of delay without much performance degradation as its design is for use in a single data center
  - Add also latency of the switches, routers, or multiplexers in the transmission path but GbEth switches add only tens of microsecs, ASIC-based routers (on OC-48 links) only about 200 microsecs
  - With a round-trip delay of 10ms, FC maximum perf reduced to 13.5MBps for 64-credit switches (2112B x 64 /0.010s), and 3.4MBps for 16-credit switches.
  - Fibre Channel's credit-based flow control mechanism severely constrains throughput over distances that introduce more than about one millisecond of latency.

# Flow Control and Credits

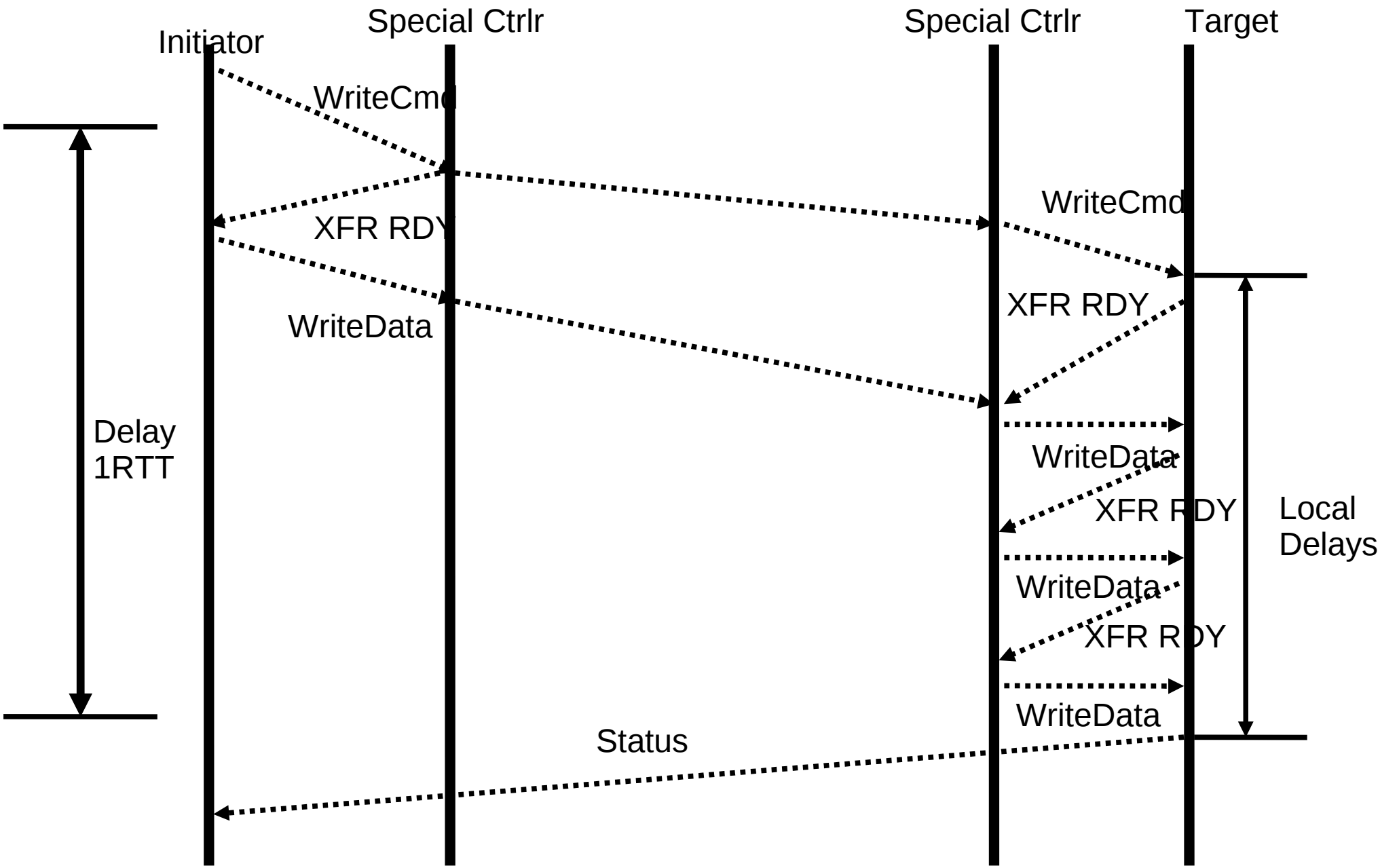


# Fibre Channel Without Opts

(A 3x512KB Write example: Not to scale)



# Fibre Channel With Optimizations



# Summary

- FC an effective protocol in the data center
- Not suitable for WAN
  - Need to encapsulate in other transport for higher BW