Redes de alta performance

Universidad Tecnológica Nacional - FRBA

Autor: Gustavo Nudelman

Evolución

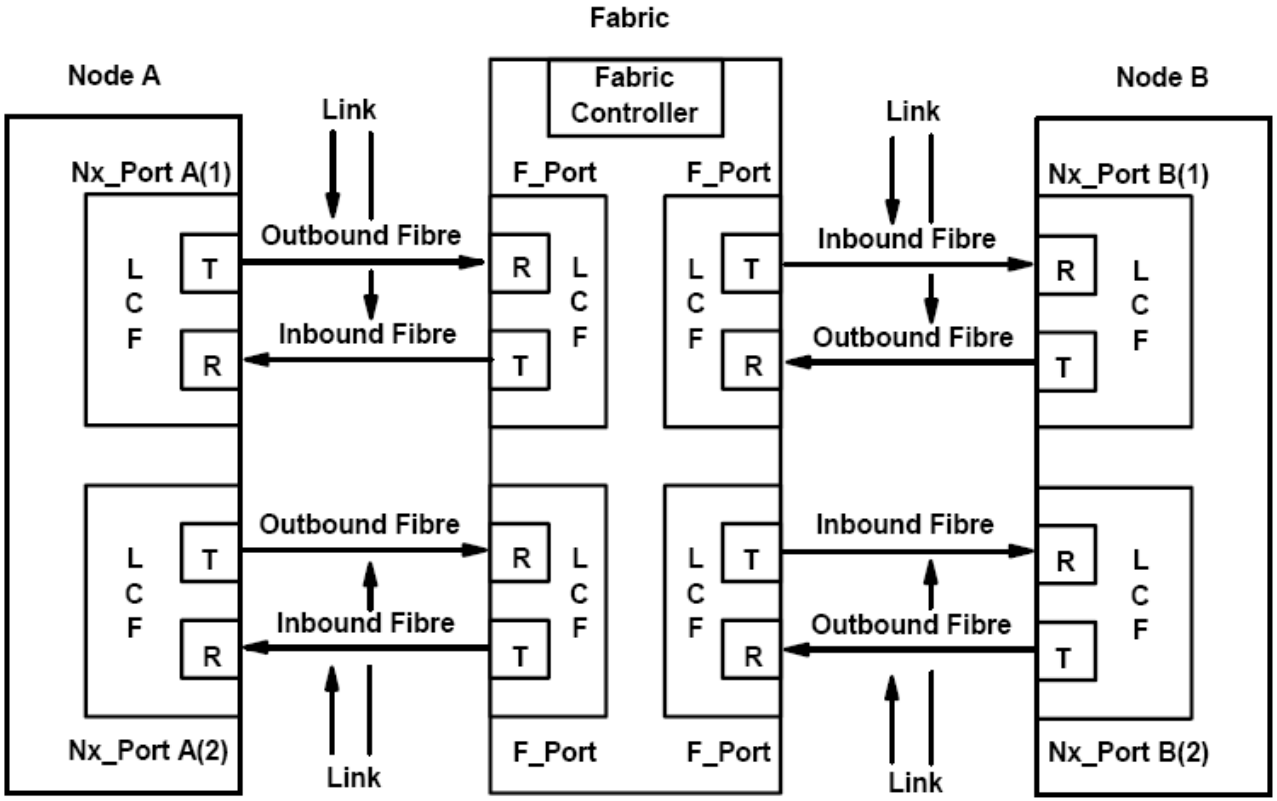
- A principios de los 90’ aparece la explosión de los negocios en internet – e comerce
- Los SLA que poseen estos negocios encausan una necesidad de HA/HP que nos lleva a sistemas de procesamiento distribuidos
- Necesidad de enlaces y redes mas veloces y seguras (en forma local y/o entre áreas metropolitanas)
- Necesidad de disminuir el overhead
- Se adoptan tecnologías existentes:
  - **Fibre Channel**
  - **Infiniband**

- La aplicación para la cual en sido adoptadas ha provocado que dichas tecnologías evoluciones a gran velocidad

Fibre Channel

- Comenzó en 1988 y se convirtió en un Standard ANSI en 1994
- Comenzó con enlaces de 1Gbps y actualmente provee velocidades de hasta 10 Gbps en una transmisión serie bidireccional
- Utilizada para redes Server-Storage (SAN) y topologías Server-Server (Cluster)
- FC No define el medio pero si la forma y/o topología de conexión por lo que no define distancias
- FC soporta una gran variedad de protocolos lo que la hace flexible a ser utilizada en diferentes tipos de redes.

Modelo Punto a punto switched - Fabric



•Transmisión serie punto a punto entre Nx_Ports

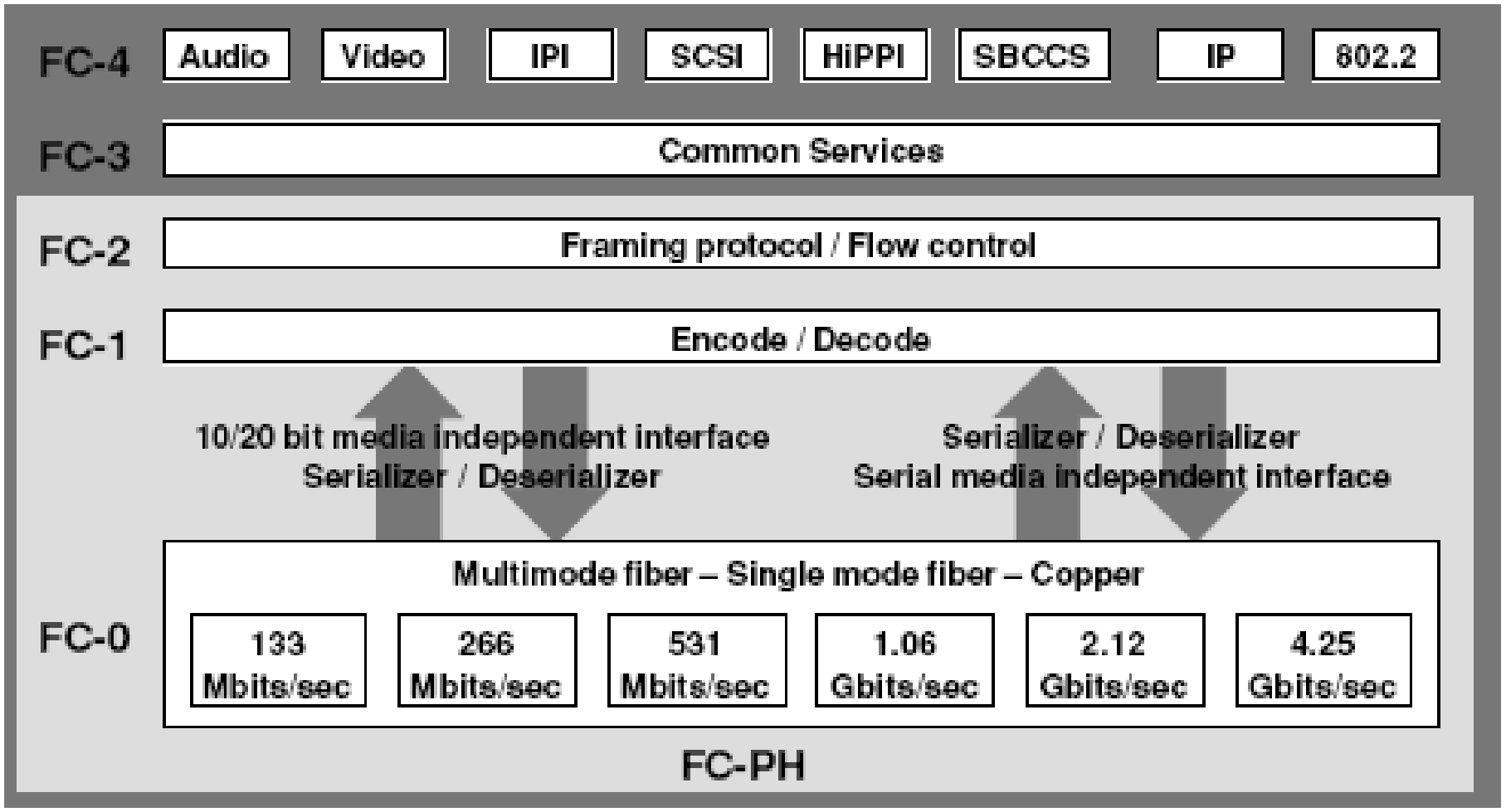
•El fabric rutea al F_Port que tiene enlazado el N_Port utilizando una vez establecida la conexión punto a punto. **La conexión es entre N_Ports**

•Las conexiones punto a punto se mantienen hasta que un N_Port finalice la sesión o ocurra una excepción en el Fabric

Modelo Punto a punto switched – Fabric (2)

- Un Dominio FC lo forman dos o más Nodos interconectados a través de un Subsistema de Interconexión.
- Un enlace conecta dos N_ports mediante 2 hilos
- LOGIN/LOGOUT en la inicialización, cada N_Port establece una Sesión de LOGIN con todos lo demás, para intercambiar
- Cada Nx_Port obtiene información sobre las posibilidades de transmisión y recepción de tramas (Service Parameters), y los nombres de Nodo y de N_Port (Name e ID). A partir de entonces, y mientras dure la Sesión (hasta que alguno haga LOGOUT), se permiten las transferencias de los ULPs.
- La asignación de la dirección a cada N_Port lo realiza automáticamente el Fabric durante el FLOGGIN. a cada Nx_Port le es asignado un ID
- Durante el login se determina el máximo tamaño de frame que el switch fabric puede manejar y se almacena cada Nx_Port ID en tablas propias del switch
- El switch Fabric no interviene en control de flujo. lo hacen los N_Ports

Modelo de capas



Capa física FC0

- Líneas separadas para Tx y Rx (Full Duplex)
- Puede ser Fibra óptica o eléctrica (twisted pair)
- Alcance hasta 2Km con multimodo y 10Km con monomodo
- Responsable de transmitir los bit pero **no** de controlar el sincronismo

Capa FC1 (Codificación)

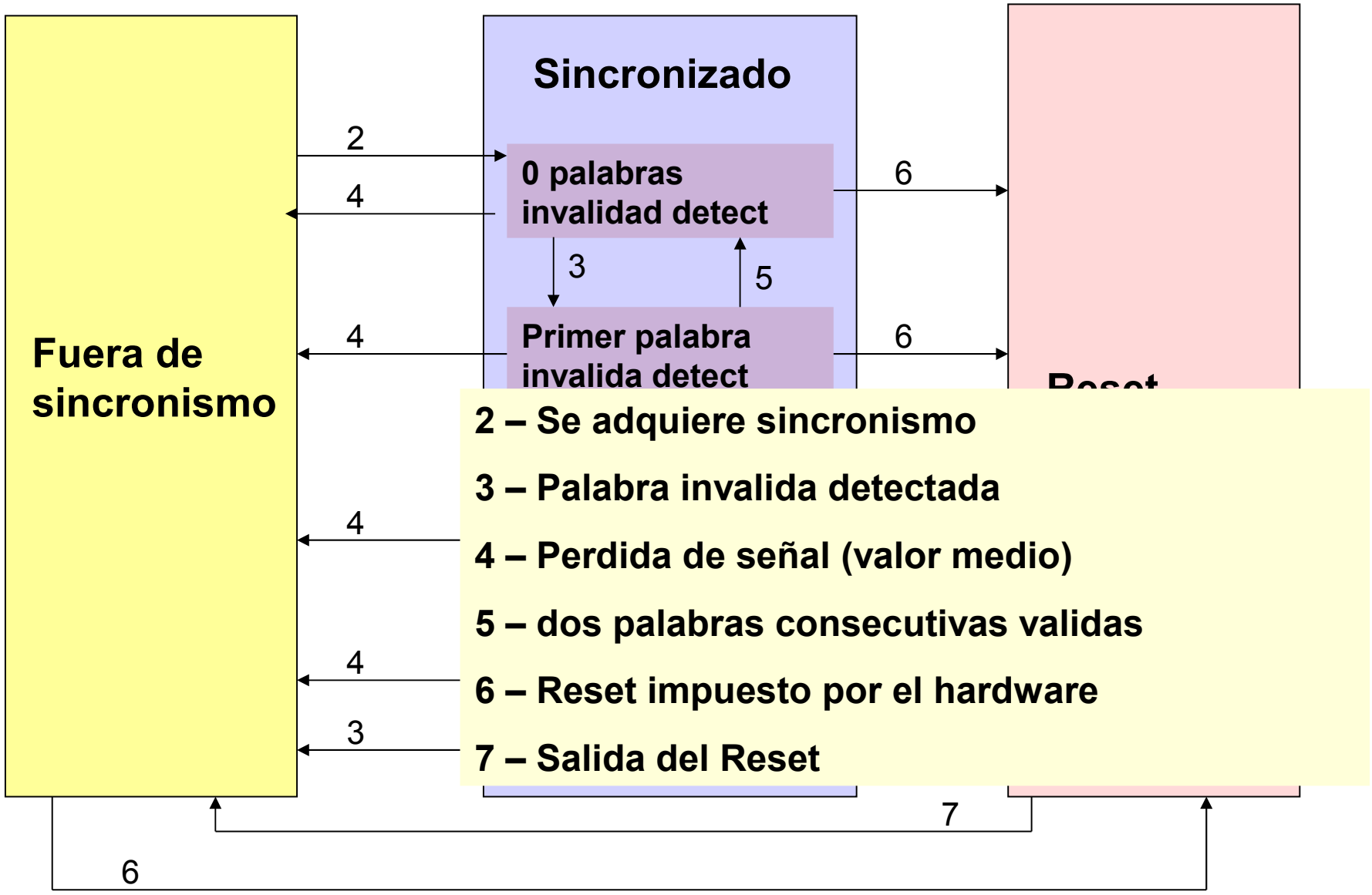
- Responsable de llevar el control del sincronismo – Esto lo realiza el firmware asociado a FC1 mediante una definida secuencia de estados.
- Como todo sistema de decodificación serie de alta velocidad se emplea NRZ
- A su vez se utiliza 8B/10B. Es decir: Se codifica cada caracter en 10 bits basado en tablas de conversión. 2^10=1024
- La cantidad de caracteres (10 bits) transmitidos debe ser múltiplo de 4 siendo esta la unidad de información de FC! Llamada “Word”
- Cada caracter puede codificarse de 2 formas diferentes. (512) →Aparece el concepto de disparidad
- Algunas combinaciones restantes se las utiliza como caracteres de control.
- Se logra no tener mas de 5 ceros o unos seguidos (ver RFC 06-085v3) – a diferencia del bit de stuffing, se asegura por codificación

Capa FC1 (Sincronismo)

- **Sincronismo a Nivel bit** → Se logra cuando el Nx_Port o el Fx_Port logran sincronizar sus relojes internos con el flujo de bits

- **Sincronismo a nivel word**
  - Se consigue situando un Carácter 8.25 de la tabla al final de cada word
  - Se considera que se ha logrado la “Word sincronization” luego de 3 words consecutivas. La 3 ya se considera valida y se la releva a los niveles subsiguientes
  - El sincronismo a nivel caracter se controla en cada word y los errores detectados pueden ser:
    - Un carácter no permitido de las 1024 combinaciones “Invalid carácter”
    - El carácter especial 8.25 dentro de una Word y no en los limites
    - Carácter valido con Error de disparidad

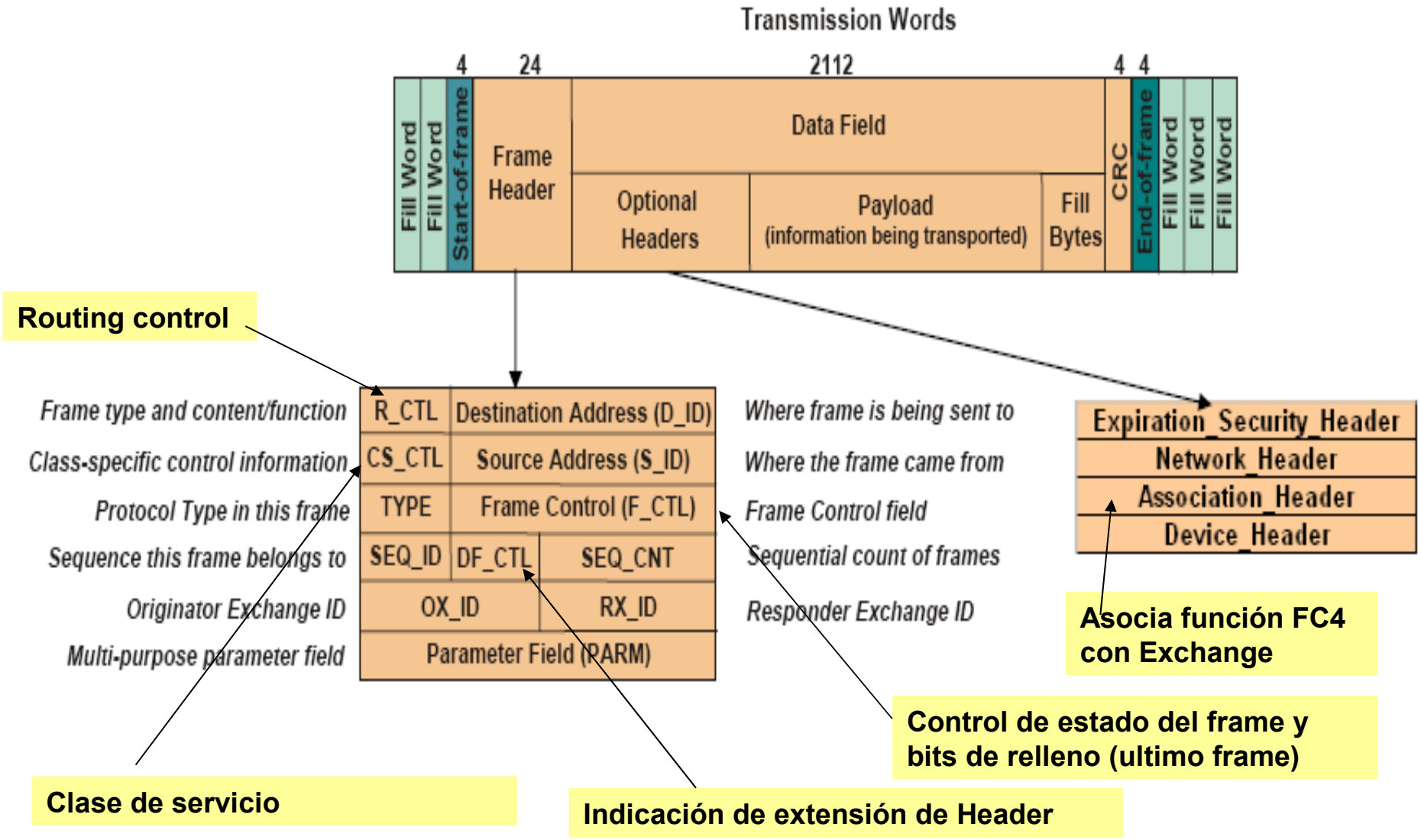
FC1 – Diagrama de estados de control de sincronismo



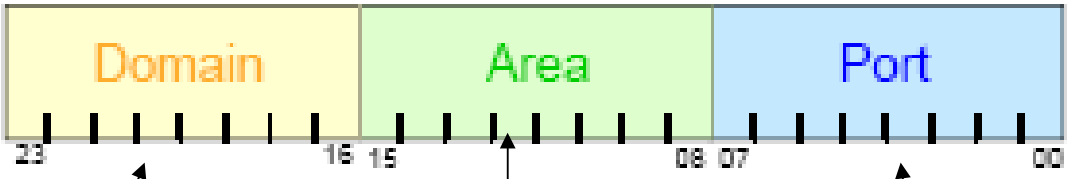
Capa FC2 (Frame)

- **Start of frame:**(4 bytes). Indica Clase de servicio, si hay que Establecer una conexión o Activar un circuito virtual y si es primera (Initiate) o sucesiva (Normal) Frame dentro de la Sequence. **Es decir: delimita y especifica el tipo de frame**
- **Header** (24 bytes, incluye el tipo de Frame (FT-0 -Link Control Frame- y FT-1 -Data Frame-), las direcciones de N_Ports origen y destino (S_ID, D_ID), OX_ID, RX_ID y SEQ_ID, Relative Offset para reensamblaje de Frames recibidas out-of-order, y SEQ_CNT (número de Frame dentro de la Sequence).. **E este nivel se hace el routing a niveles superiores, el ordenamiento**.
- **Data Field** (Payload), variable desde 0 hasta 2,112 bytes máximo, en incrementos de 4 bytes.
- **CRC** (4 bytes).
- **EOF**: End Of Frame (4 bytes). Indica si hay que Cerrar una conexión o Desactivar un circuito virtual, si es la última (Terminate) o no (Normal) Frame de la Sequence, si la Frame ha sido abortada por el transmisor (Abort), o si una entidad intermedia ha detectado un error en la Frame (Invalid).

Capa FC2 (Frame) (2)



Interpretación de D_ID / S_ID



Posibilidad de designar áreas dentro del switch

Similar para todo el switch (Hasta 239 switches)

Identificación del N_Port

Control de estado del frame (F_CTL)

Control Field	Word 2, Bits	Description	Reference
Exchange Context	23	0 = Originator of Exchange 1 = Responder of Exchange	See 16.4.
Sequence Context	22	0 = Sequence Initiator 1 = Sequence Recipient	See 16.4
First_Sequence	21	0 = Sequence other than first of Exchange 1 = first Sequence of Exchange	See 16.4
Last_Sequence	20	0 = Sequence other than last of Exchange 1 = last Sequence of Exchange	See 16.7
End_Sequence	19	0 = Data frame other than last of Sequence 1 = last Data frame of Sequence	See 9.7.6
End_Connection (Class 1or 6)	18	0 = Connection active 1 = End of Connection Pending (Class 1 or 6)	See 19.7.3
CS_CTL/Priority Enable	17	0 = Word 1, Bits 31-24 = CS_CTL 1 = Word 1, Bits 31-24 = Priority	See 9.5.2 See 9.5.3
Sequence Initiative	16	0 = hold Sequence initiative 1 = transfer Sequence initiative	See 16.6.3 See 16.6.4
X_ID reassigned - Obsolete	15		
Invalidate X_ID - Obsolete	14		
ACK_Form	13-12	00b = No assistance provided 01b = Ack_1 Required 10b = reserved 11b = Ack_0 Required	See 9.7.10
Data Compression – Obsolete	11		
Data Encryption - Obsolete	10		
Retransmitted Sequence	9	0 = Original Sequence transmission 1 = Sequence retransmission	See 20.7.2.3
Unidirectional Transmit (Class 1)	8	0 = Bi-directional transmission (Class 1) 1 = Unidirectional transmission (Class 1)	See 19.5.4
Continue Sequence Condition	7-6	Last Data frame - Sequence Initiator 00b = No information 01b = Sequence to follow-immediately 10b = Sequence to follow-soon 11b = Sequence to follow-delayed	See 16.6.5.5

FC2 - Tipos de frame

El campo R_CTL, en combinación con el campo type , definen el tipo de frame y consecuentemente el protocolo de nivel superior que llevan ciertos frames.

- Data Frames
  - a) Link_Data frames;
  - b) Device_Data frames
  - c) Video_Data frames.

- Link control frames (Campo type en 0ch)
  - a) Acknowledge (ACK) frames;
  - b) Link_Response (Busy and Reject) frames; and
  - c) Link_Control command frames.

En caso de “Link control frames”, el campo type se utiliza para códigos de error

Códigos para el campo “Type” para los Data frames.

Encoded Value in Word 2, bits 31-24	Description
00h to 03h	Reserved
04h	Obsolete
05h	IPv4, IPv6, and ARP over Fibre Channel (see RFC 2625, RFC 3831, RFC 4338 [a])
06h to 07h	Reserved
08h	Fibre Channel Protocol (see FCP-3)
09h	Obsolete
0Ah to 0Fh	Reserved - SCSI
10h	Reserved
11h to 13h	Obsolete
14h	Fibre Channel SATA Tunnelling Protocol (see FC-SATA)

[a] The IETF has published RFC 4338, which obsoletes both RFC 2625 and RFC 3831

FC2 (Frame, Sequence y Exchange)

Frame:

- Unidad mínima de transporte de información en FC2
- Cada frame conlleva un acuse de recibo

Sequence:

- es conceptualmente una unidad de nivel superior aunque pertenece a FC 2 y es la que ve el programador sin ser problema para este la segmentación de la misma en frames de tamaño previamente negociado
- Unidireccional entre ports
- Unidad de control de errores
- Cada sequence es única y esta identificada en un campo del frame header. Y cada frame dentro de la sequence posee su frame count. (relativo a la misma)

Exchange:

- Una o mas sequence **no concurrentes** (La concurrencia es a nivel exchange)
- Posee id's en el frame header: OX_ID y RX ID (creadas por el Originator y respomder)
- En general se dividen en: Comando – Datos - Status

FC2 (Sequence y exchanges)

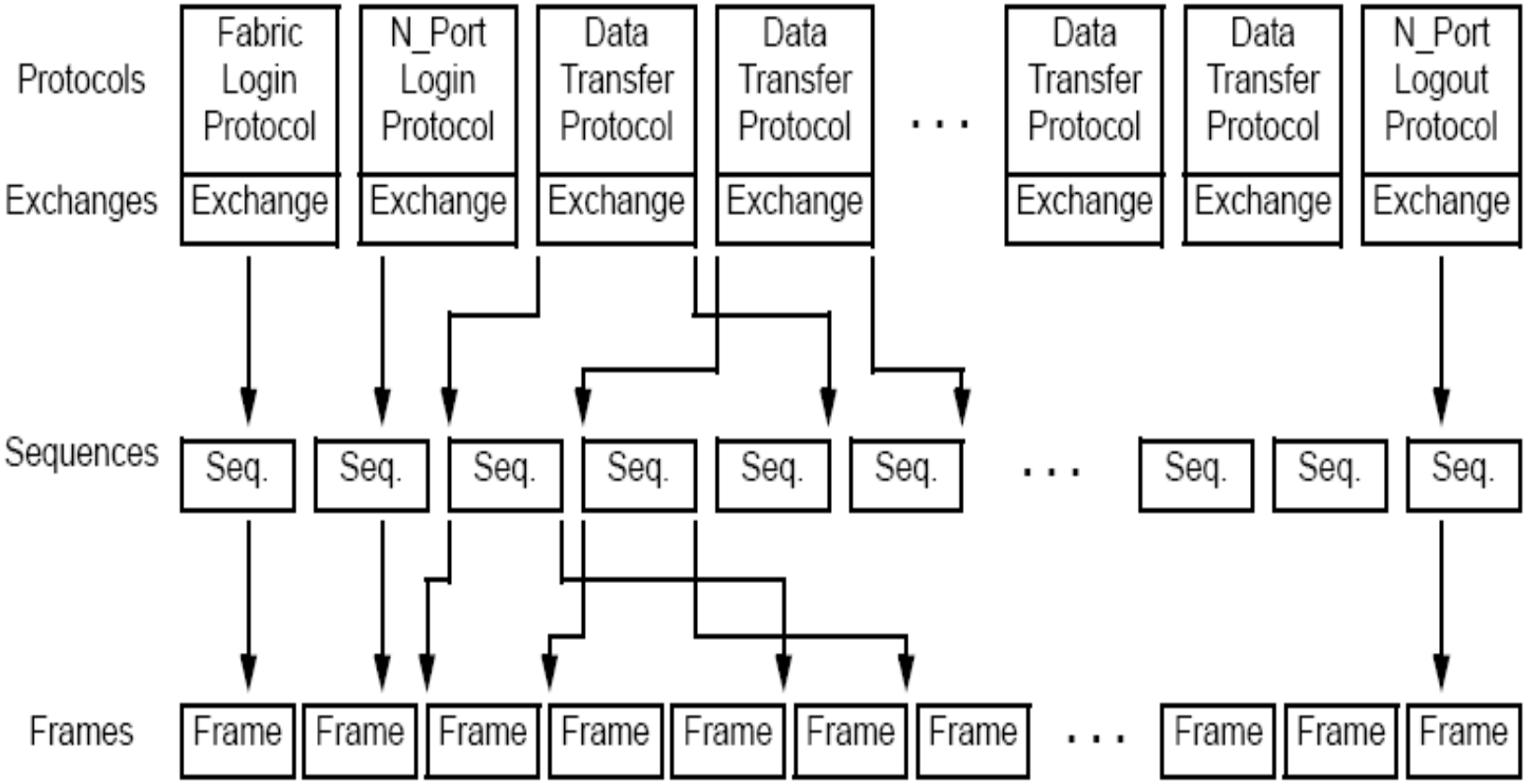
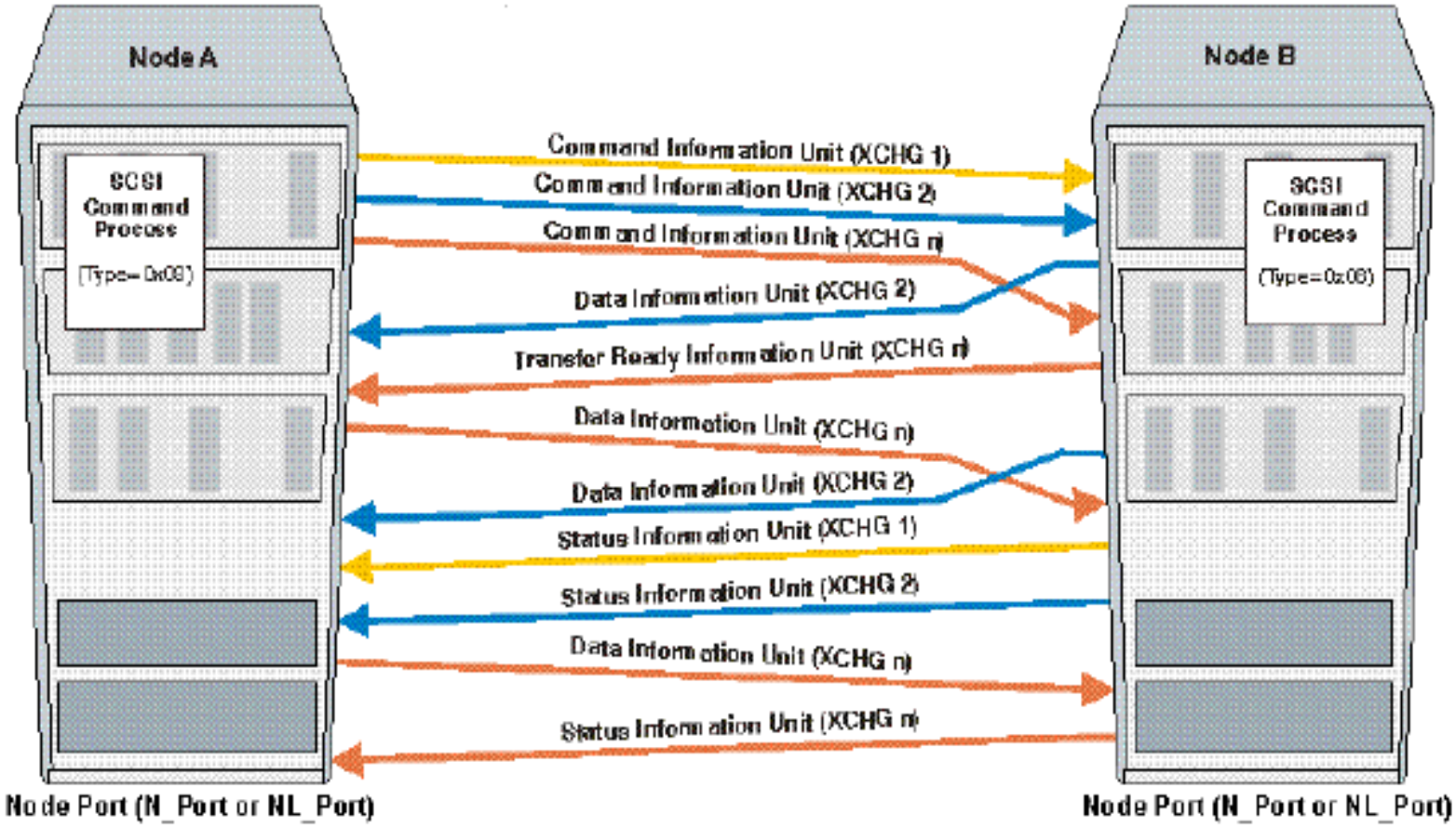


Figure 8 - FC-2 building block hierarchy

Sequence & exchange (2)

- Las funciones de envío de información de los protocolos de FC4 trabajan a nivel sequence, y es el protocolo quien se encarga de dividirla en frames.
- Un iniciador origina el Exchange y coloca un numero en OX_ID (originator Exchange)
- El responder asigna un RX_ID en los frames de respuesta
- Ambos memorizan los Exchanges ID´S y a partir de ahí el Exchange se identifica con el par en cada secuencia
- Los N_Ports llevan un traqueo de cada Exchange en forma de tablas identificado por el par OX_ID y RX_ID del Exchange.

FC2 (exchanges y optimización de BW)



Proceso de conexión

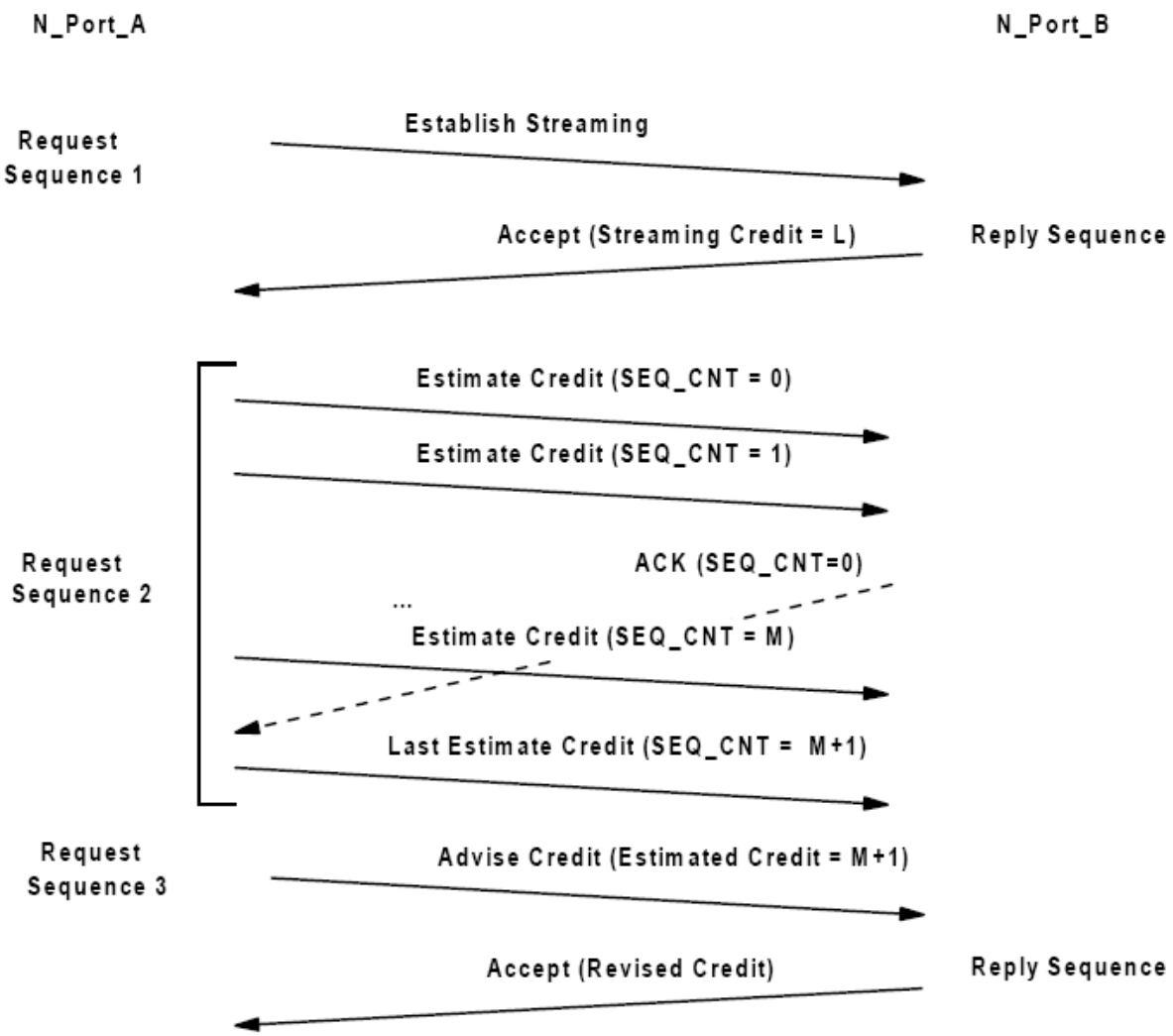
Existen 3 tipos de loggins en redes fabric:

- **Fabric Loggin (FLOGI)**
  - El N_Port se autentica en la red fabric y recibe una determinada dirección del switch.
  - Envía un FLOGI frame (determinado por el campo R_CTL) con el “Node name”, “Port Name” y algún “Service parameter”
  - El D_ID es 0xFFFFFE (Well known fabric address) y el FLOGI se envía con una source address 0x000000
  - El switch responde con un ACC “Accept” que contiene un Address valido
  - Se inicializa el crédito a nivel buffers
- **Port Loggin (PLOGI) (Habilitación del port)**
  - Una vez que se estableció el FLOGI-ACC se envia la dirección asignada al fabric port 0xFFFFFC (well known) para que este la almacene junto con otros datos como clase y tipo de transferencia
  - El switch informa que se ha incorporado un nodo, con toda la información de este a los demás ports.
- **Process Loggin (PRLI)**
  - Es directamente entre nodos transparente para el switch
  - En el caso de que el nodo solicitado este ocupado, el Fabric se encargará de responder con un frame “Busy” y con una “Reason Code” en el campo “Type” (Excepto clase 3 donde se descarta el frame) o “Stacked connect”
  - Si la conexión se establece, las responsabilidades de control de flujo quedan delegadas a los Nx_Ports

Negociación de ventana (Credit)

- Al igual que en TCP, existe una ventana para el aprovechamiento optimo del BW
- Esto tiene mayor peso en circuitos conmutados ya que la ventana se mantiene durante toda la conexión
- **Nx_Port End-to-end Credit):** Es el tamaño de ventana que se establece en una conexión de clase determinada (Cuantos frames sin ACK se pueden llegar a mandar)
- **Buffer to Buffer Credit** otorgado por cada Puerto al que tiene al otro lado del enlace
- Durante el Loggin se establecen algunos parámetros como el máximo tamaño de frame que se podrá utilizar. Esto es base para el calculo de ventana que tendrá lugar después al comenzar el streaming
- Comenzado el streaming de datos el proceso de establecimiento de ventana “Credit” se hace en 3 etapas
  - Se establece la sequence de streaming
  - En iniciador estima un valor “Credit”
  - El Iniciador “Sugiere” un valor Credit
- A partir de esto ambos extremos llevan contadores que se incrementan con cada frame transmitido y se decrementan con cada Ack recibido. Si no se reciben frames de tipo Ack, se deja de transmitir hasta no recibir el mismo.

Negociación de la ventana (Credit)



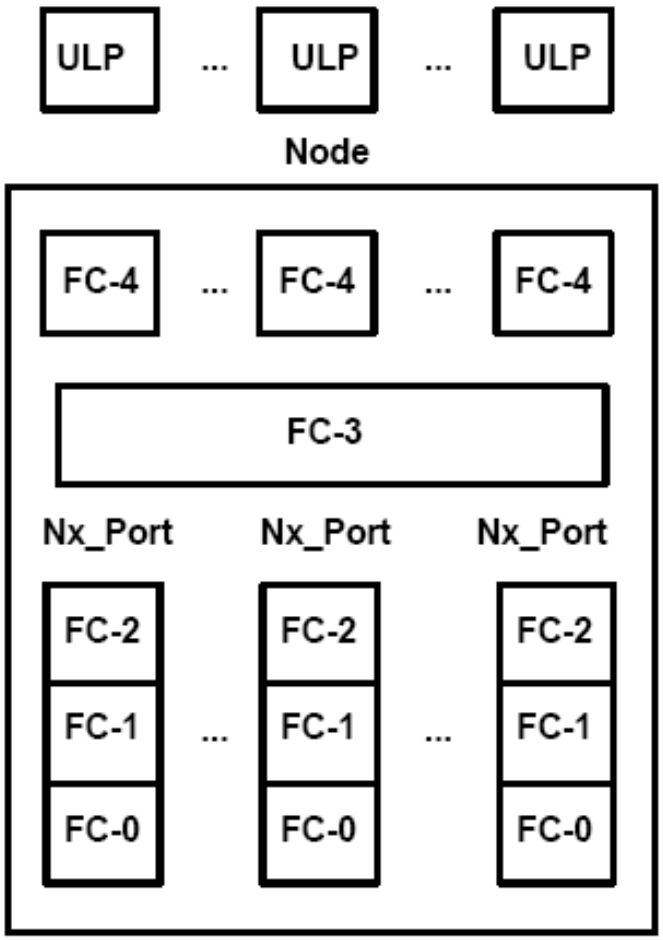
FC3 – Servicios generales

Name Server: Provee información de la base de datos que asocia nodos, puertos y protocolos de nivel superior asociados.

Alias Server: Asignación de un ID a un grupo de puertos (multicast). Aumento en la eficiencia de enrutado cuando varios puertos accedes a un mismo servicio.

Management Serices: Servicios especiales de administración

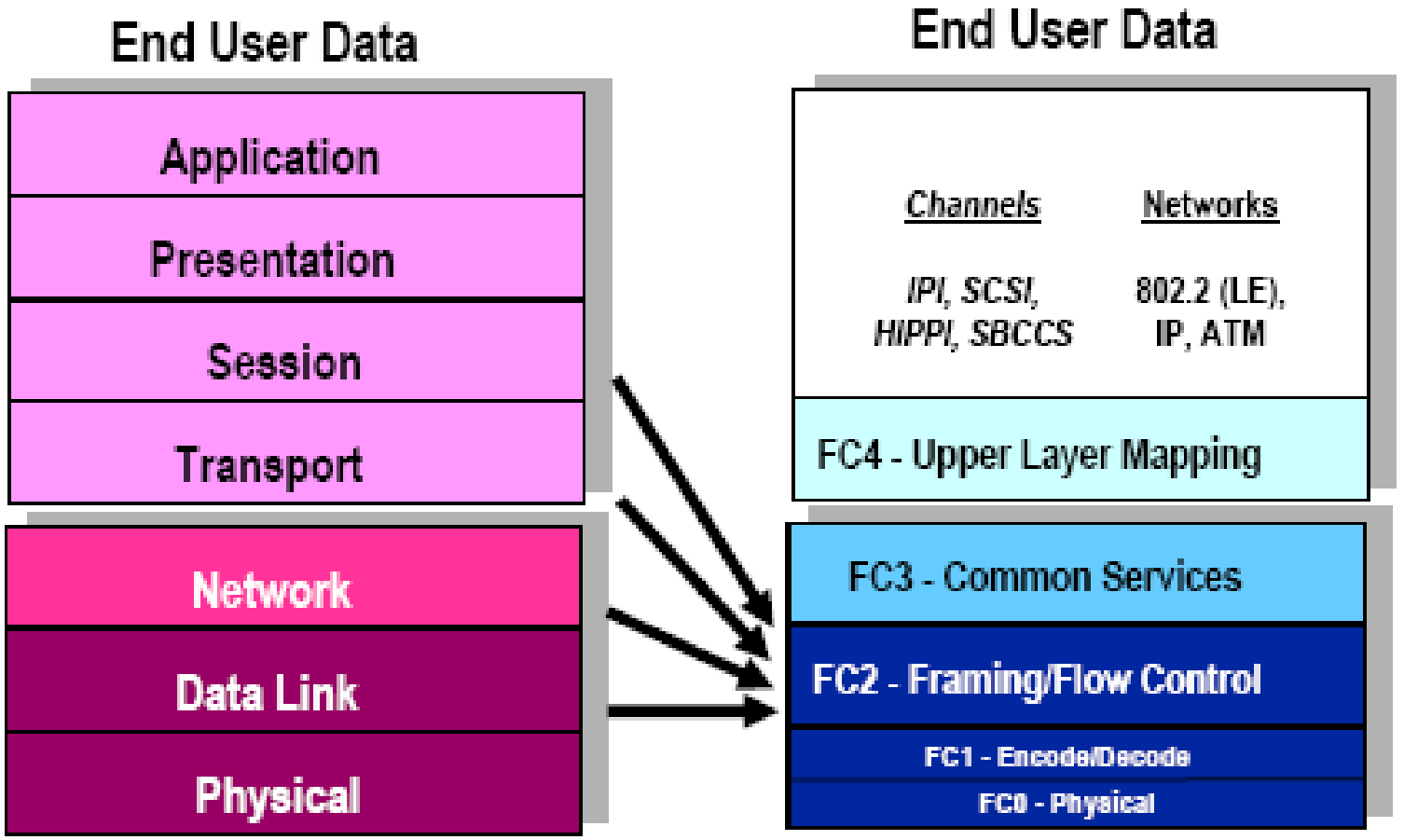
Key Server: Autentitacion para el management services y encriptación.



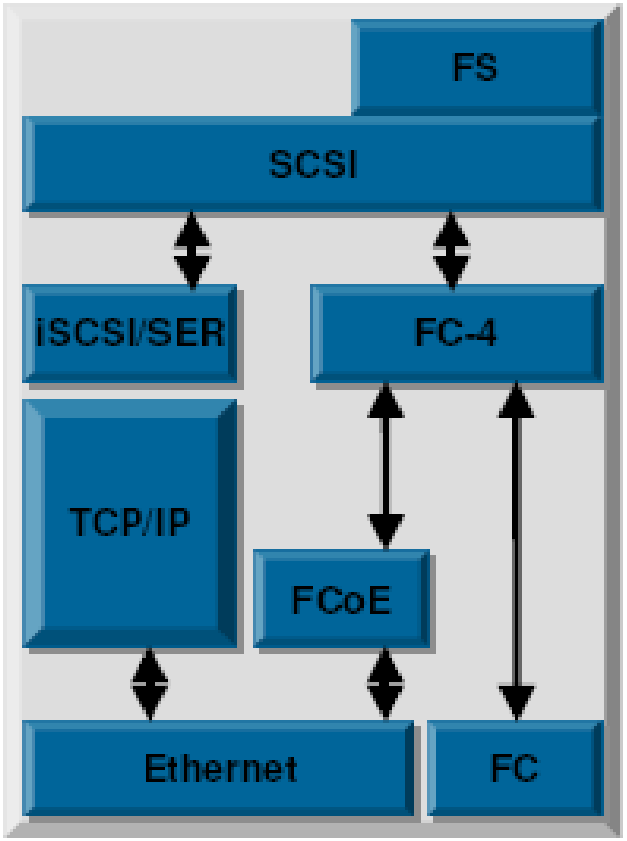
FC4 - ULP

- Small Computer System Interface (SCSI)
- Internet Protocol (IP)
- High Performance Parallel Interface (HIPPI)Framing Protocol
- Link Encapsulation (FC-LE)
- IEEE 802.2
- Asynchronous Transfer Mode - Adaption Layer 5 (ATM-AAL5)
- Intelligent Peripheral Interface - 3 (IPI-3) (disk and tape)
- Single Byte Command Code Sets (SBCCS)
- future ULPs...

Comparación con el modelo OSI

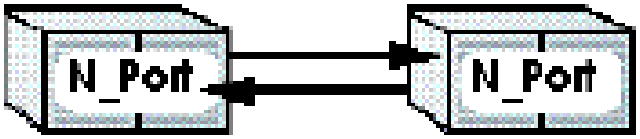


FCoE



Topologías

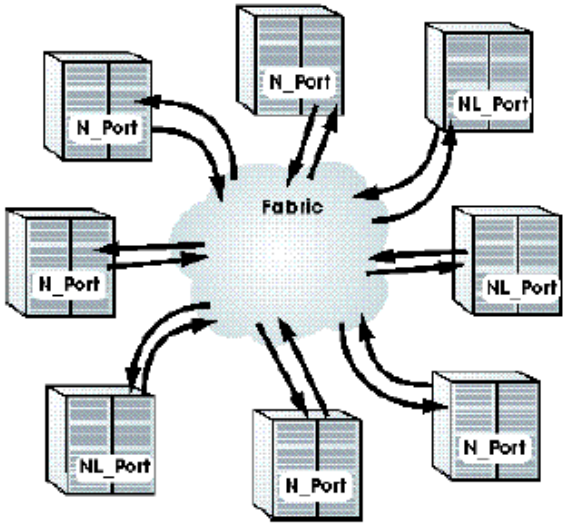
1) Point to point



- Se trata de un enlace entre 2 puertos que intercambian información
- Puede se también a través de un conmutador si este proporciona un circuito virtual permanente con BW fijo.
- En cualquiera de los casos, ambos nodos pueden hacer uso del 100% del BW disponible,

Topologías (2)

2) Switched (fabric)

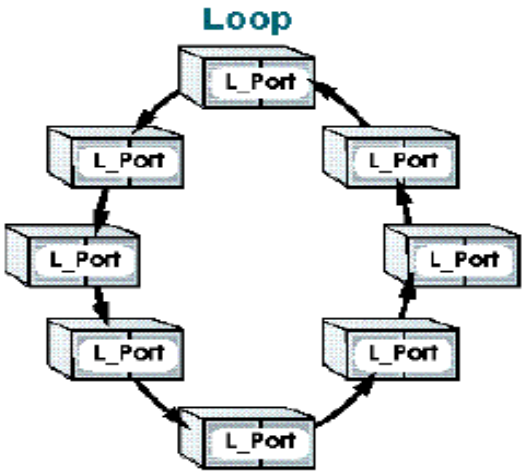


- Los frames de redirigen al nodo correspondiente (no es tarea del nodo)
- El switch redirige los frames, analizando su header. Esto lo hace según el modo o clase que se halla seleccionado
- En una red fabric se puede trabajar hasta con 2`24 ports

Topologías (3)

3) Loop

Cada nodo pide por el control del loop disponiendo el 100% del BW en una comunicación bidireccional por un determinado time slot



• Pueden conectar hasta 127 ports

•Utilizan los llamados ***port bypass circuit*** que se instalan en los propios nodos o funcionan externamente como hubs. (Estos últimos) evitan la caída del loop por un nodo

•Se mantiene un registro de loop (acerca de los nodos activos y es el bypass circuit quien se encarga de actualizarlo.

•Cada nodo pide por el control del loop con un mensaje llamado “primitiva” hacia el nodo consecutivo. Y si la primitiva no es para este, el mismo la releva al siguiente.

•Se utilizan en general transferencias no orientadas a la conexión con ack (class 3)

•Ante 2 attempts tiene prioridad la dirección de menor valor

•Utilizado internamente en enclosures y ventajoso para “hot swap”

Clases de servicio

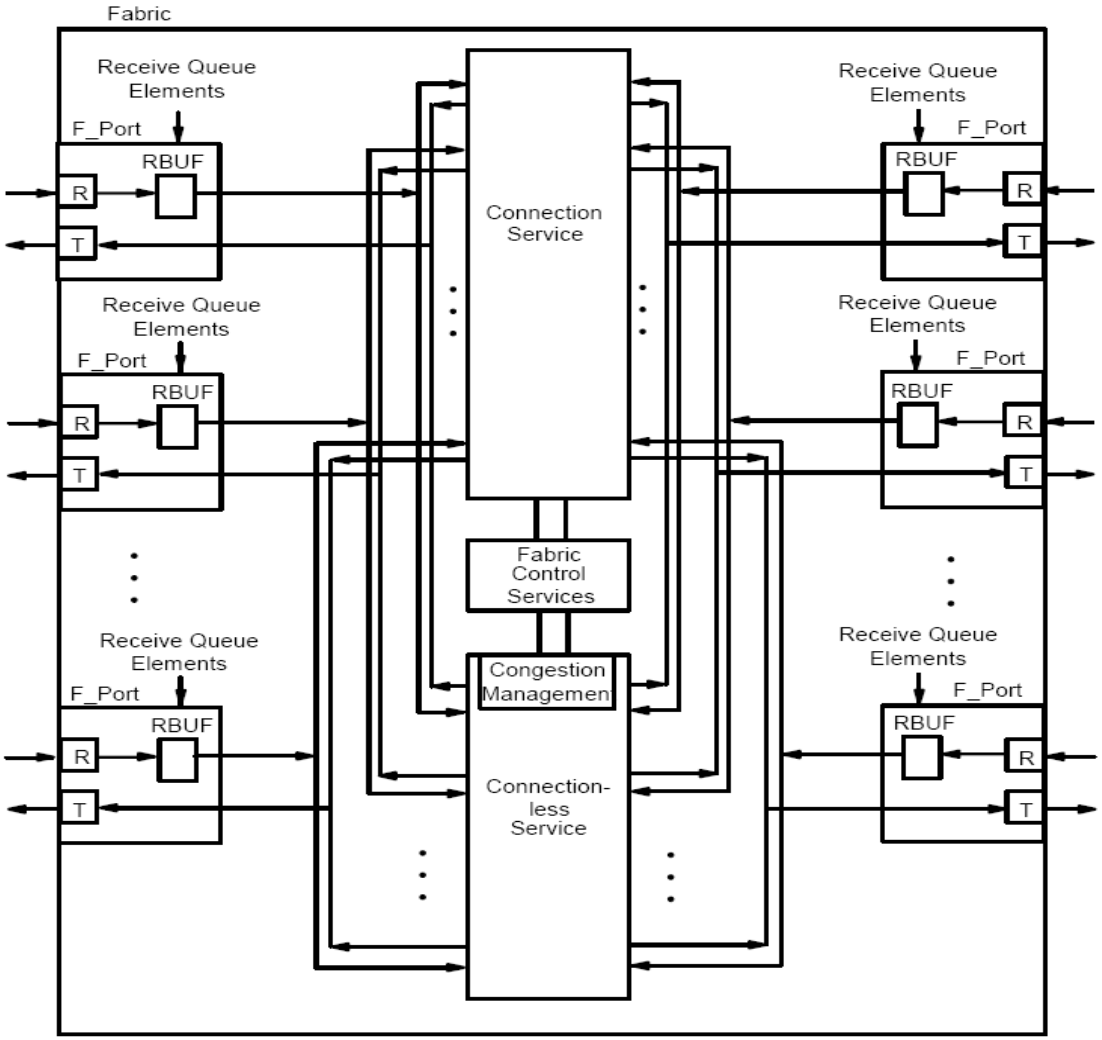
Los switches de fibre channel pueden ser configurados para operar en lo que se denomina “clases”

- **Class 1: Acknowledged connection service**
  - El switch crea un circuito virtual con con full BW – End to End path
  - Solo existe overhead en el inicio y fin de conexion → eficiencia para transferencias de datos masivos
  - Posee acuse de recibo
  - Posee servicios propios de la clase como *Camp on* y *Stacked connect* de manera de monitorear cambios de estado y encolar pedidos de conexión.
  - Permite establecer conexiones unidireccionales para recibir datos de dos puertos diferentes y buffers para adaptar diferentes velocidades

Clases de servicio (2)

- **Class 2: Acknowledged connectionless service**
  - Cada frame se rutea de forma independiente
  - Como class 1 .. Es de entrega garantizada
  - A diferencia de class 1, el path entre los nodos no es dedicado – multiplexado
  - Si el nodo destino no esta disponible, el switch devuelve un frame “busy”
- **Class 3: Unacknowledged connectionless service**
  - Similar a Class 2 pero si acuse de recibo
  - Se utiliza en topologías loop
  - Utilizada por protocolos de nivel superior (ULP) donde realizan el control de flujo a dicho nivel:
- **Class 4: Fractional Bandwidth acknowledged connection oriented service**
  - Orientado a la conexion con reserva de BW → Calidad de servicio con BW garantizado
  - Cada nodo puede pedir hasta 256 conexiones concurrentes
  - Un Circuito virtual por cada dirección
  - En router se encarga de realizar buffering de manera de no exceder el BW solicitado
  - Posee control de flujo end-end pero sin embargo se lo utiliza para streaming debido a su BW garantizado.

FC - Switch


Fabric
Receive Queue Elements
F_Port
RBUF
R
T
Connection Service
Fabric Control Services
Congestion Managemen
Connection-less Service