Tema 3
Redes de Área de Sistema

Nicolás Calvo Cruz
Dpto. de Arquitectura y Tecnología de los Computadores
@ncalvocruz
ncalvocruz@ugr.es

Motivación

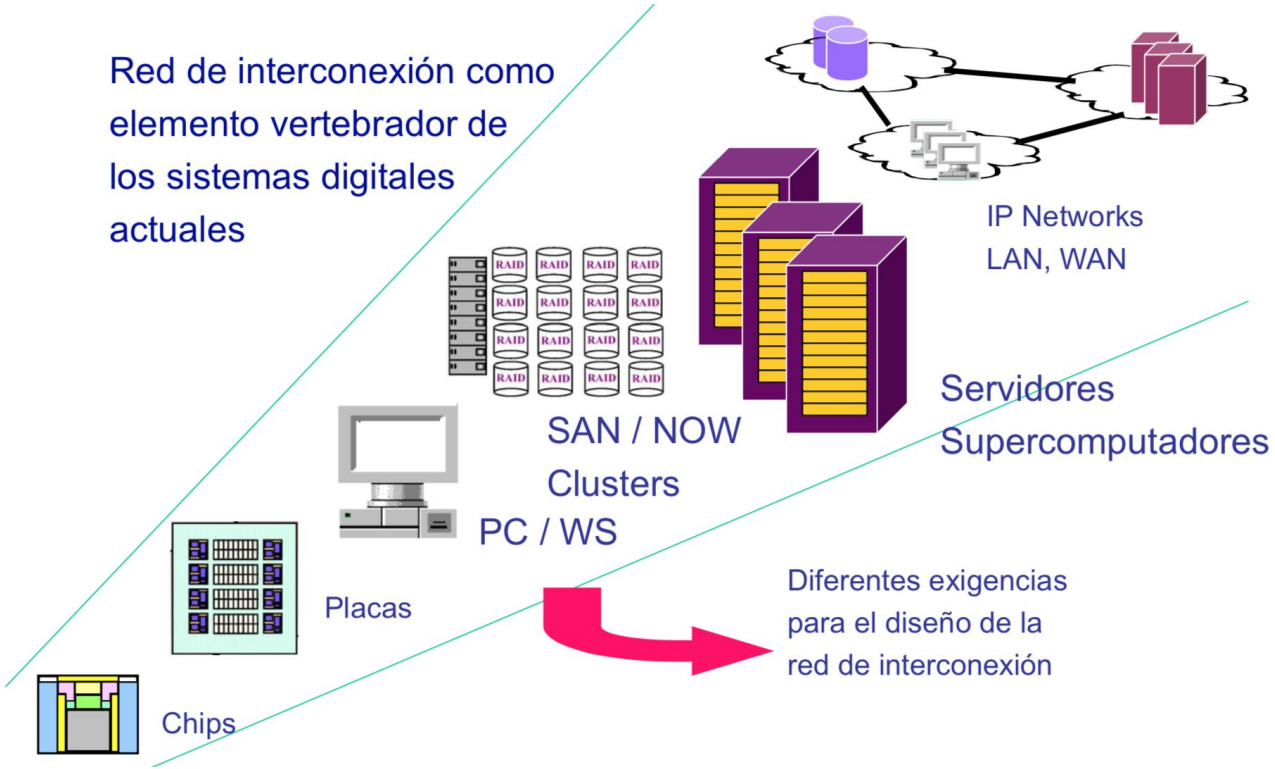
- Redes de **comunicación en computadores** paralelos.
- Hoy en día se están **sustituyendo** los **buses por redes** con conexiones **punto a punto** a todos los niveles:
  - Interno al chip
  - A nivel de tarjeta y placa
  - A nivel de chasis o caja
  - LAN y Router IP
- Conocer los algoritmos de **encaminamiento** y la **infraestructura** permite mejorar las **prestaciones**.

Objetivos

- **Distinguir** entre redes de **altas prestaciones y** redes **estándar**.
- **Conocer** la estructura general de **un conmutador**.
- **Estudiar** las **topologías** y nomenclaturas de las redes **de altas prestaciones**.
- Estudiar los **algoritmos de encaminamiento**.

Índice

1. Clasificación de Sistemas de Comunicación (I)

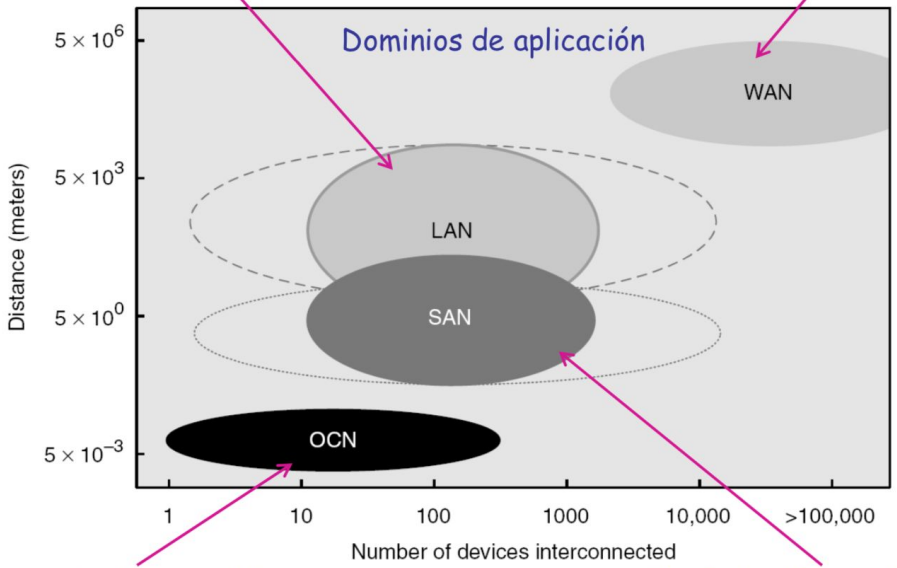


1. Clasificación de Sistemas de Comunicación (II)

Clases	Nº Nodos y Distancia	Utilización	Desarrollo	Ejemplos
Diseñadas a medida	pocos-cientos-miles decenas o cientos de metros	Multiprocesadores y Multicomputadores	Arquitecturas de Altas prestaciones	Cray's Aries Mellanox's Infiniband network interconnect Infiniband, Myricom
SAN: System Area Network	decenas-cientos-miles distancias desde decenas a cientos de metros	Conecta computadores en una sala/habitación	Redes a medida y LAN	
LAN	cientos-miles decenas de Km	Conecta computadores en un edificio o campus	Estaciones de Trabajo	Estándares: Fast Ethernet, Gigabit Ethernet
WAN	miles kilómetros	Conecta computadores a nivel mundial	Telecomunicaciones	Estándares: ATM

1. Clasificación de Sistemas de Comunicación (III)

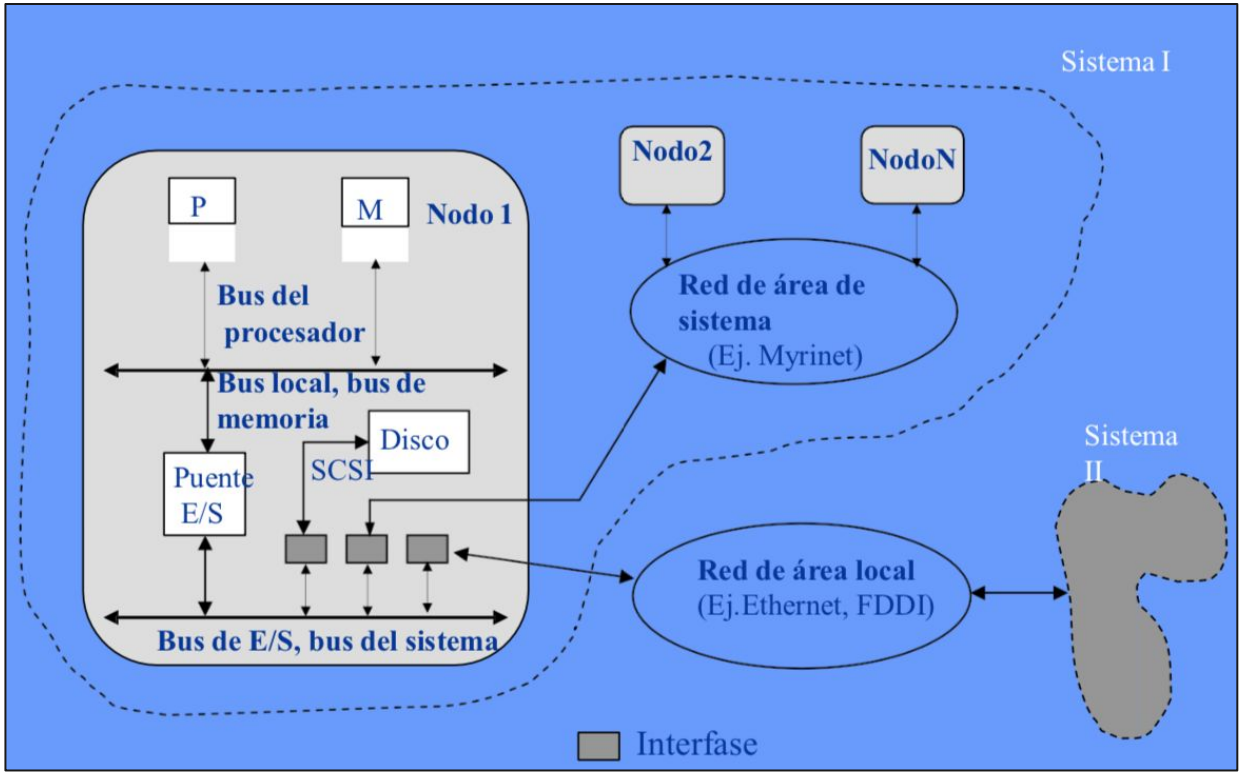
- Local Area Network (LAN). Ej: Ethernet (hasta 10 Gbps sobre una distancia de 40 Km). Tb Clusters de PCs
- Wide Area Network (WAN). Conexión a escala global. Ej: ATM



- Network-on-Chip (NoC, OCN). Ej: AMBA de ARM, SCCC de Intel, TILE-GX de Tilera
- System/Storage Area Network (SAN). Comunicación proc-mem-discos. Ej: Infiniband (hasta 200 Gbps sobre una distancia de 300 m)

1. Clasificación de Sistemas de Comunicación (IV)

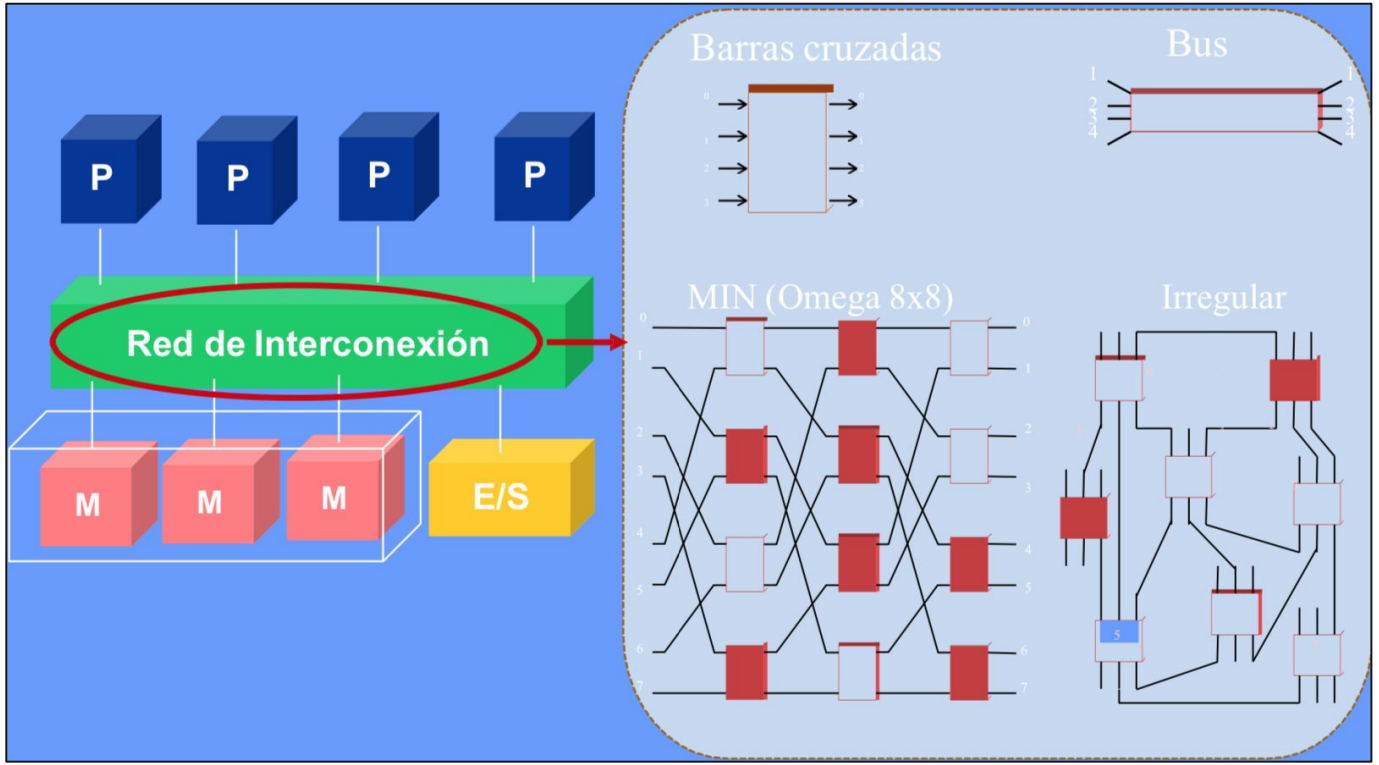
Se **relacionan:**



1. Clasificación de Sistemas de Comunicación (V)

En **HPC**:

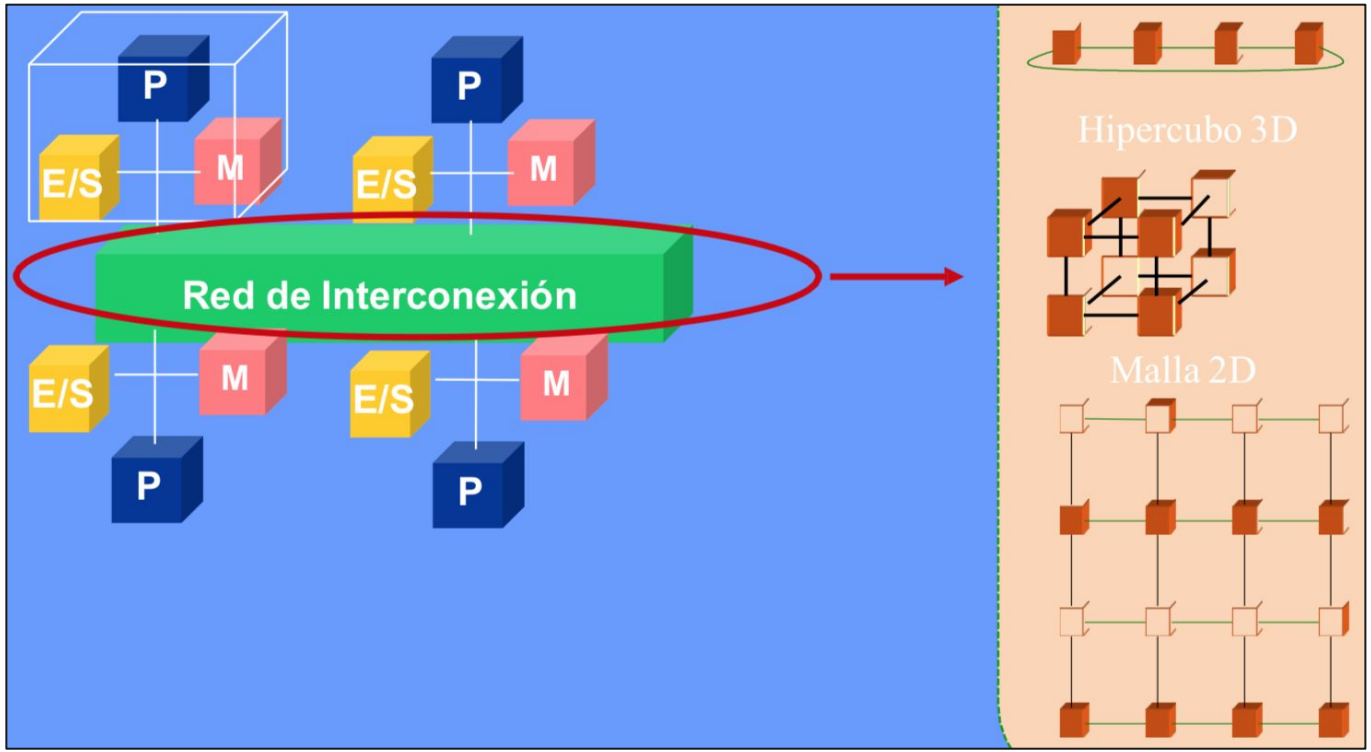
Multiprocesador de memoria centralizada



1. Clasificación de Sistemas de Comunicación (VI)

En **HPC**:

Multicomputador



2. Propiedades

1. Funcionalidad
2. Topología
3. Características:
    a. Diámetro de una red
    b. Ancho de la bisección
    c. Latencia
    d. Productividad
    e. Escalabilidad
    f. Grado de los nodos
    g. Niveles de servicio
    h. Calidad de servicio
    i. Alta disponibilidad
    j. Tolerancia a fallos
    k. Fiabilidad
    l. Remote Direct Memory Access

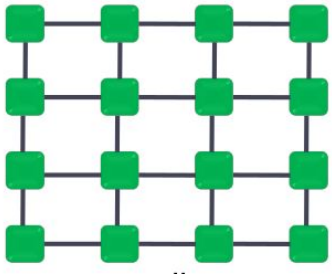
2. Propiedades: Funcionalidad

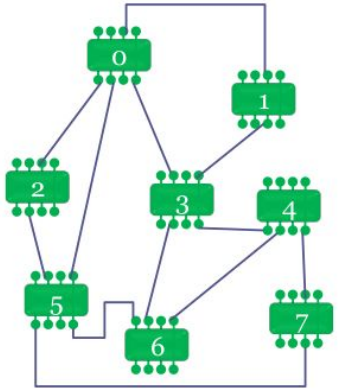
- **La Funcionalidad** de una red de comunicación **define su uso**.
- Según su funcionalidad encontramos:
  - Redes para **almacenamiento**
  - Redes para **computación**
  - Redes de **administración**

2. Propiedades: Topología (I)

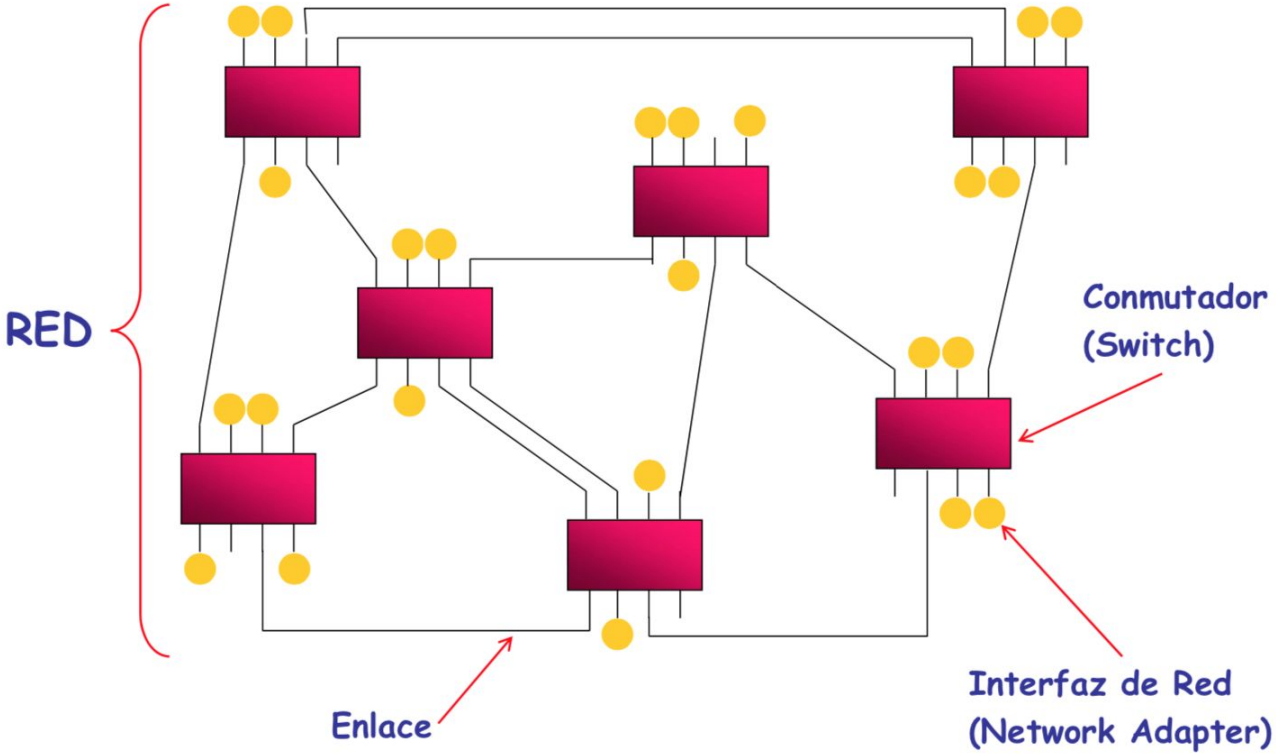
- La **topología** es la **estructura** de la interconexión física de la red.

- Se puede modelar mediante un **grafo** en el que los **vértices son conmutadores o interfaces de red** (NI) y las **aristas** son **conexiones.**

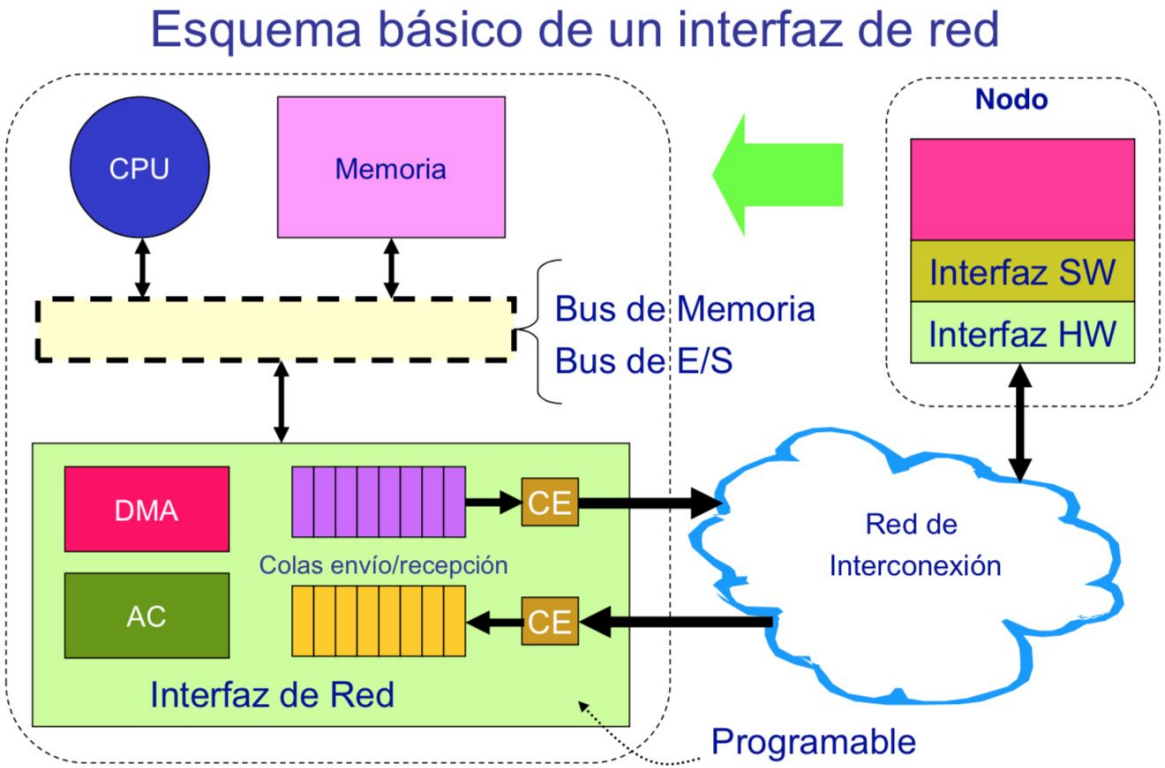




2. Propiedades: Topología (II)



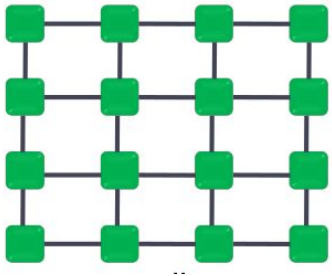
2. Propiedades: Topología (III)

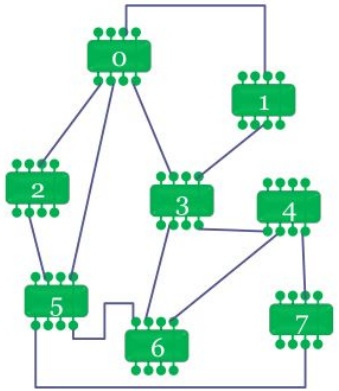


2. Propiedades: Topología (IV)

- **Clasificación** de redes **según la topología de los enlaces**:
  - **Regulares:** Son las redes en las que los enlaces siguen un patrón regular.

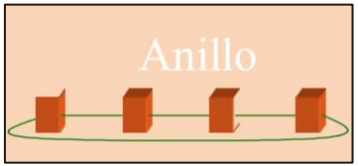
  - **Irregulares:** Son las redes en las que los enlaces no siguen un patrón regular.

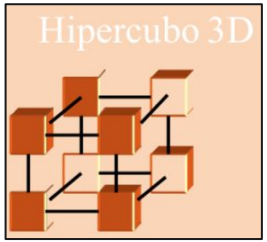


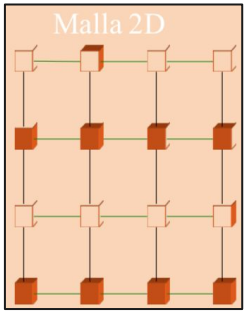


2. Propiedades: Topología (V)

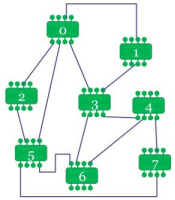
- Clasificación de redes **según la rigidez de los enlaces**:
  - **Estáticas (o Directas):** El conmutador (switch) está en la Interfaz de Red (NI) y los enlaces son rígidos, no se pueden adaptar. Se subdividen en:
    - Estrictamente ortogonales (anillos, mallas, toros, hipercubos…)
    - No ortogonales (árboles, estrella)



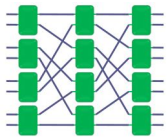


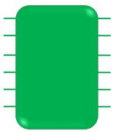


2. Propiedades: Topología (VI)



- Clasificación de redes **según la rigidez de los enlaces**:
  - **Dinámicas:** La topología no es rígida y se puede adaptar a las circunstancias. La adaptación se puede llevar a cabo con conmutadores:
    - Medio compartido (buses)
    - Medio no compartido (**o Indirectas**)
      - Irregulares
      - Barras Cruzadas
      - Multietapa (bloqueantes, no bloqueantes, reconfigurables)

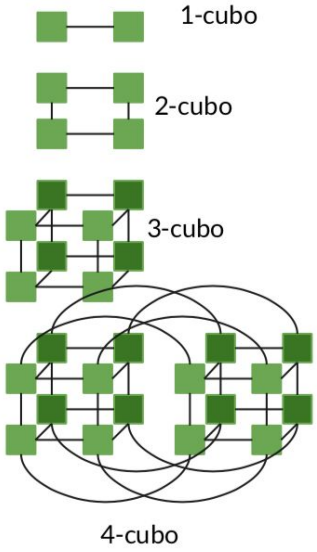


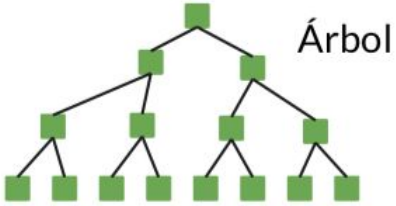


2. Propiedades: Topología (VII)

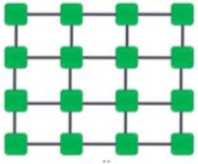
- Clasificación de redes **según la rigidez de los enlaces**:

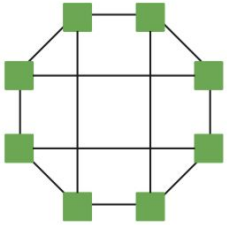
Estáticas o Directas





Malla





Anillo cordal de grado 3

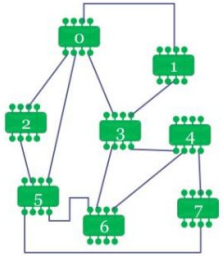
Dinámicas

Bus

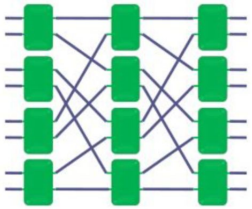
Medio Compartido

Barras cruzadas

Irregular



MIN (Omega 8x8)

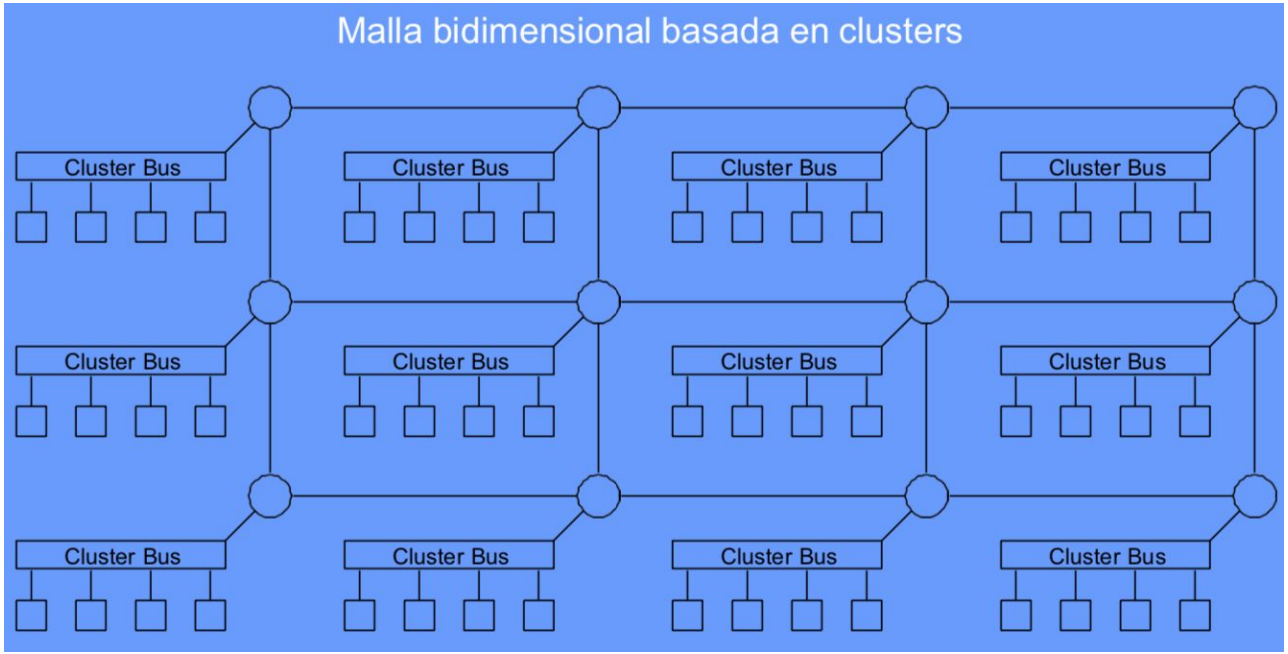


Indirectas

2. Propiedades: Topología (VIII)

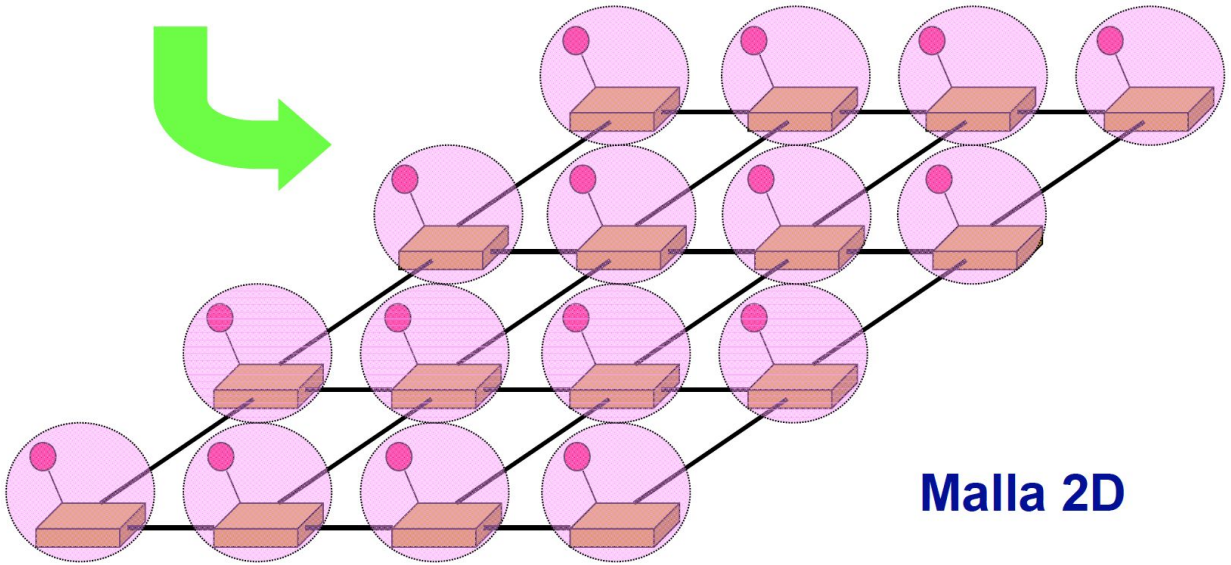
- Clasificación de redes **según la rigidez de los enlaces**:

Se puede mezclar:



2. Propiedades: Topología (IX)

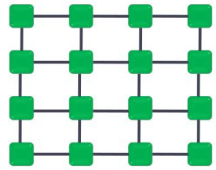
- Una red directa equivale a una red indirecta donde a cada conmutador se conecta un único nodo:

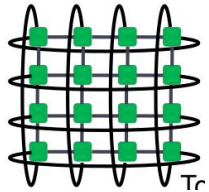


2.2 Topología: Redes estrictamente ortogonales (I)

- Se denomina red estática de dimensión n y $K_{n-1}*...*K_1*K_0$ a una red en la que los conmutadores tienen k nodos en la dimensión n-1, k nodos en la dimensión n-2,..., k nodos en la dimensión 1 y k nodos en la dimensión 0.

- Se dice que es estrictamente ortogonal si:
  - Cada conmutador tiene al menos un enlace en cada dimensión.
  - Cada enlace se mueve en una única dimensión.
  - Nodos representables en un espacio n-dimensional ortogonal.

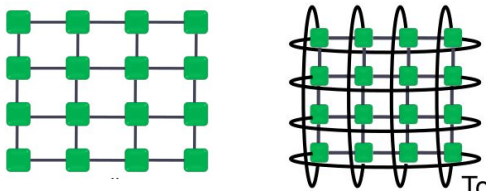




2.2 Topología: Redes estrictamente ortogonales (II)

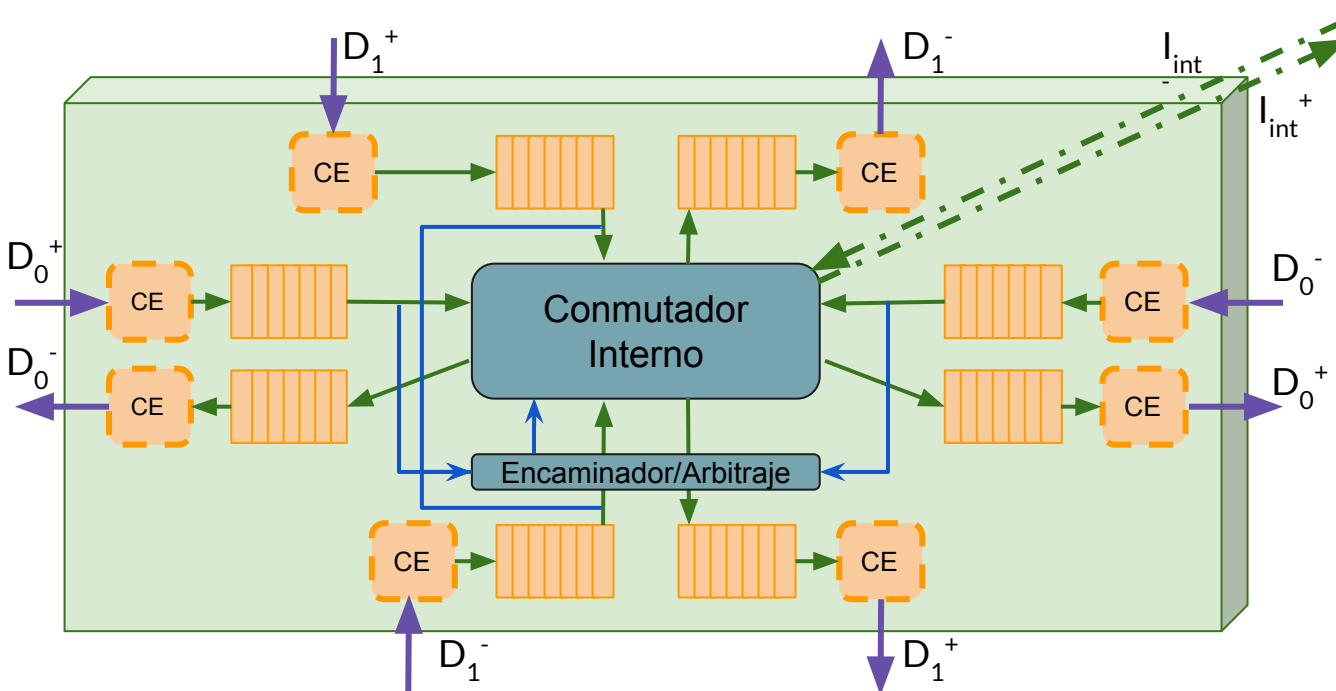
- Los nodos se numeran como puntos en un espacio vectorial en base k_i, según la dimensión i a la que pertenezca la numeración.
  - La identificación del nodo A será $a_{n-1}, \dots, a_1, a_0$

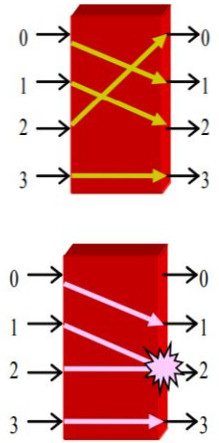
- Los enlaces unen nodos cuya distancia es 1. Llevarán un signo que indica el sentido del enlace, positivo, + o negativo, -.



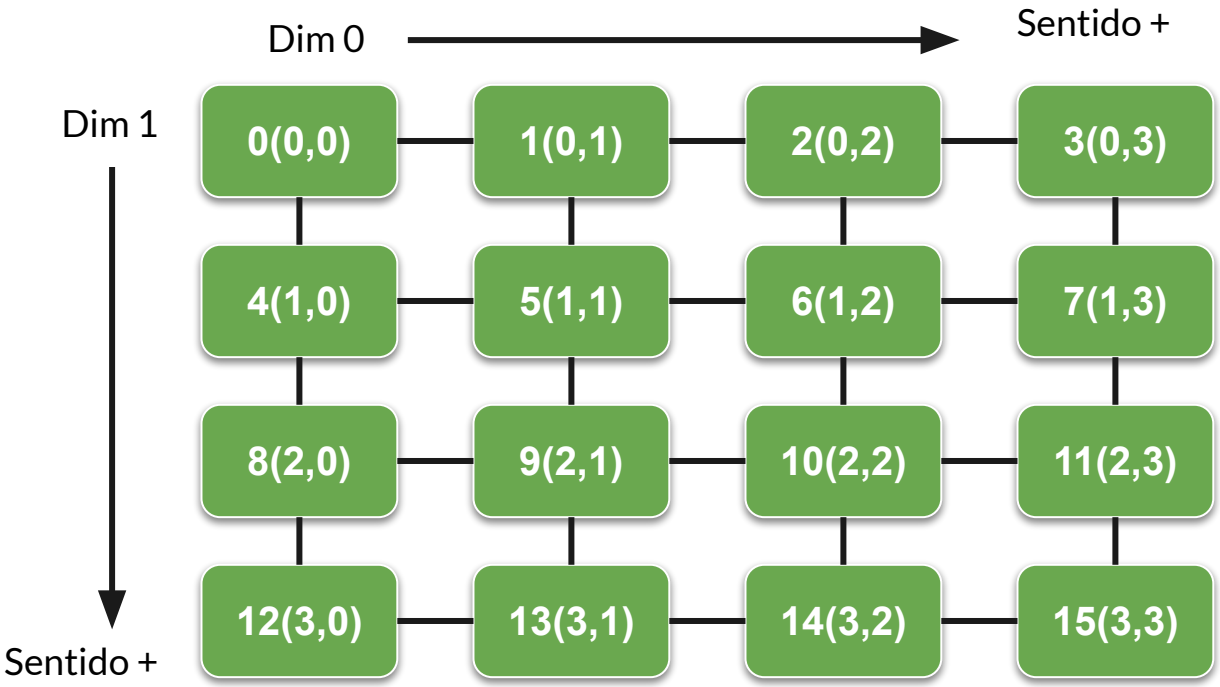
2.2 Topología: Redes estrictamente ortogonales (III)

- Conectores, conmutadores / switches.
- Ej: Conmutador para una red directa estrictamente ortogonal de dimensión 2





2.2 Topología: Redes estrictamente ortogonales (IV)



Malla de dim. 2 y base 4 (4x4 nodos)

$A=(a_1, a_0)$

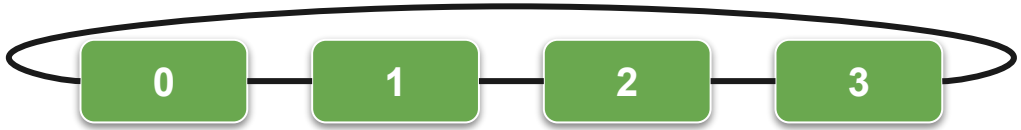
$a_1, a_0 \in \{0,1,2,3\}$

Dist. mínima, dm, entre F y D:

$$dm(F, D) = \sum_{i=0}^{n-1} |d_i - f_i|$$

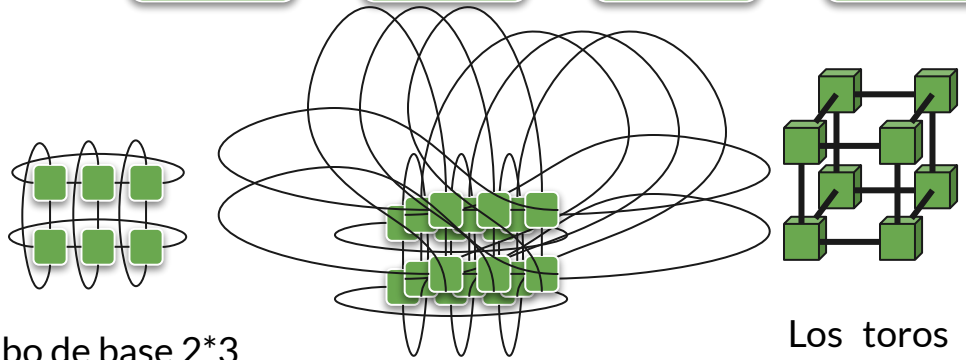
F = 2
D = 9
dm(2,9)=|1-2|+|2+0|=1+2=3

2.2 Topología: Redes estrictamente ortogonales (V)



Toro de dim 1 y base 4

Malla de dim 1 y base 4



Toro/Cubo de dim 3 y base 2
Hipercubo de dimensión 3

Cubo de base 2*3
2-cubo 2*3

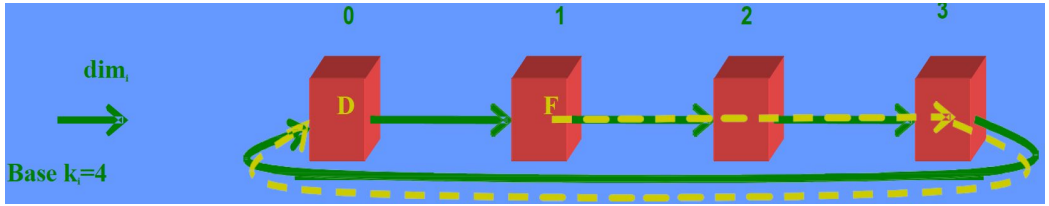
Cubo de base 3*2*3
3-cubo 3*2*3

Los toros son n-cubos k-arios, conectando k^n nodos en n dimensiones, con k nodos por dimensión

2.2 Topología: Redes estrictamente ortogonales (VI)

Distancia mínima en un toro unidireccional:

$$dm(F,D) = \sum_{i=0}^{n-1} (d_i - f_i) \bmod k$$



dm(1, 0) = (0 - 1) mod 4 = 3

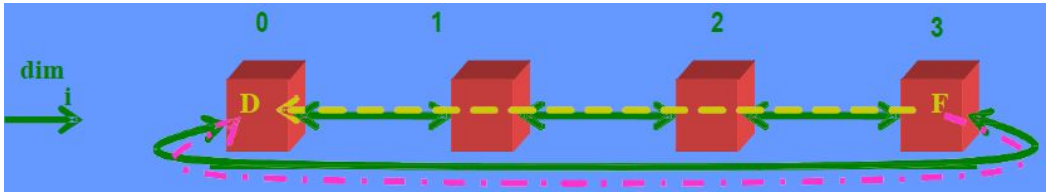
2.2 Topología: Redes estrictamente ortogonales (VII)

Distancia mínima en un toro bidireccional:

$$offset_i^+ = (d_i - f_i) \% k_i$$

$$offset_i^- = (f_i - d_i) \% k_i$$

$$Dist(F, D) = \sum_{i=0}^{n-1} min(offset^+_i, offset^-_i)$$



offset+(F, D)= (0-3) % 4 = 1
offset-(F, D) = (3-0) % 4 = 3

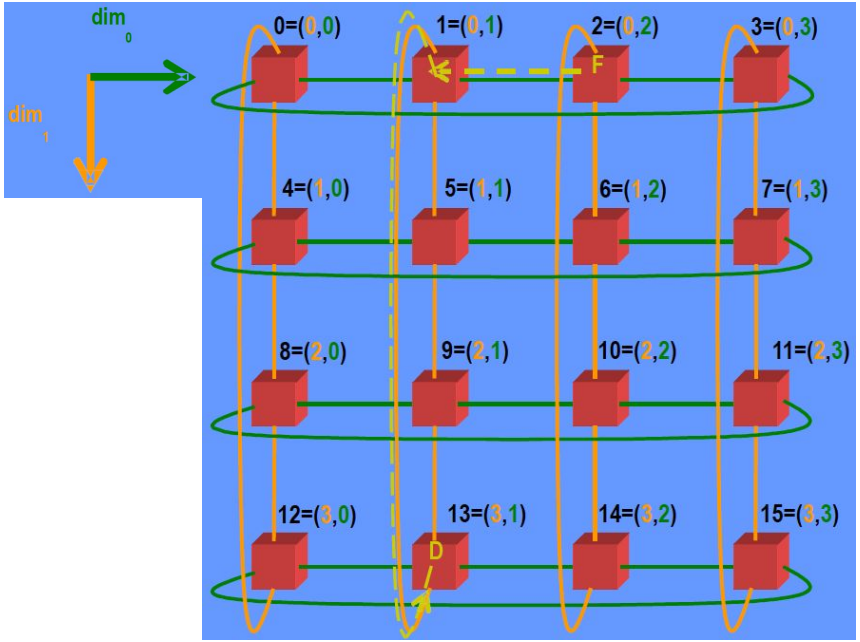
Dist(F, D) = min(1, 3) = 1

2.2 Topología: Redes estrictamente ortogonales (VIII)

Distancia mínima en un toro bidireccional:

2-cubo 4-ario:

Toro de dimensión 2
y base 4
(4x4 nodos)



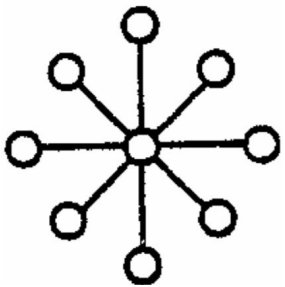
$A=(a_1, a_0)$

$a1, a0 \in \{0, 1, 2, 3\}$

dm(2, 13) = min(3, 1)+min(3, 1) = 2

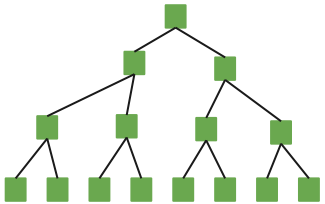
2.2 Topología: Redes no estrictamente ortogonales

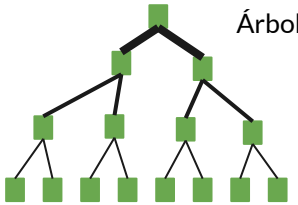
- No poseen un único enlace conectando cada dimensión.
- Incluye árboles y estrellas.



Estrella

Árbol

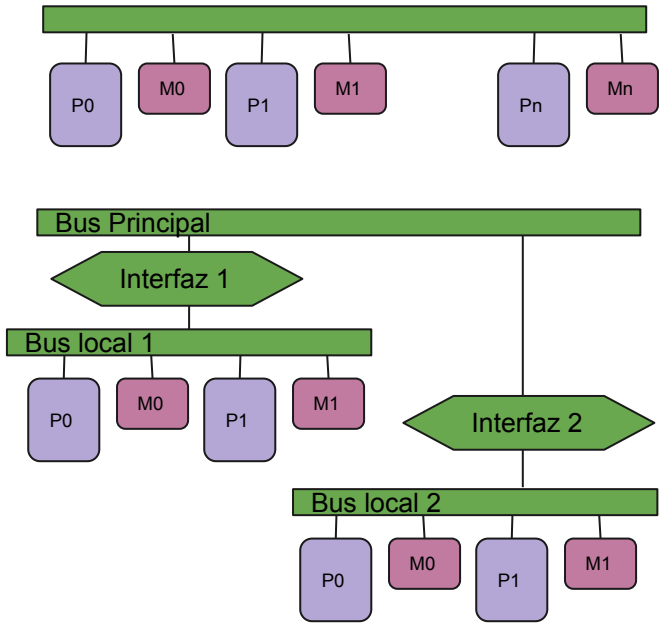




2.2 Topología: Redes dinámicas

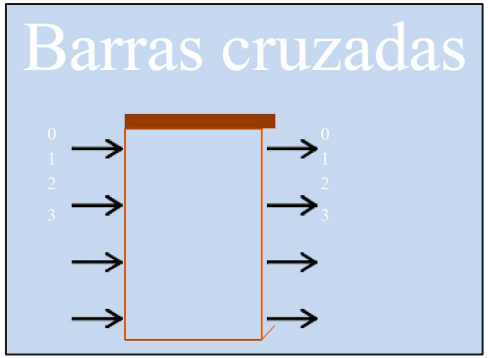
- Incluye los **buses**, las **redes de barras cruzadas** y las redes **multietapa**

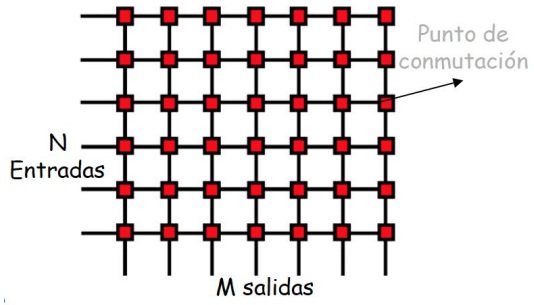
- **Buses:** Red que permite conectar un número variable de componentes que se acogen todos a las mismas normas.
  - Se llama de **medio compartido** porque en el medio sólo puede existir **una comunicación por instante**.
  - En caso de conflicto hay que decidir quién tiene prioridad, según las normas que se controla el árbitro del bus, que será la circuitería que decide quién y en qué orden acceden al bus.
  - Es la red con **menor coste**, y **peores prestaciones**
  - Se pueden incrementar sus prestaciones, estableciendo **jerarquías** de buses



2.2 Topología: Redes de medio no compartido (I)

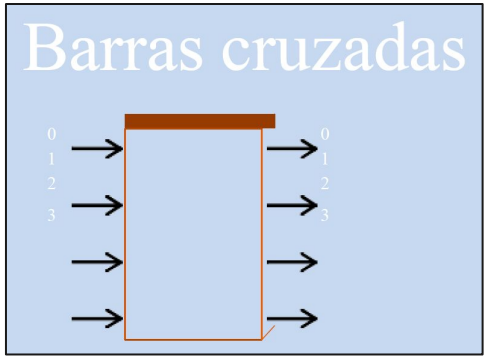
- **Redes de barras cruzadas (Crossbar Networks) de n entradas por m salidas (n*m)**
  - Todos los elementos están conectados mediante un **conmutador de líneas** cruzadas que permite **conexión entre todas las entradas y salidas** (totalmente conectada).
  - La conexión en cada cruce de líneas es **configurable**.
  - Se usa **para conectar procesadores y** módulos de **memoria**: no se puede acceder desde un procesador a dos módulos de memoria simultáneamente.

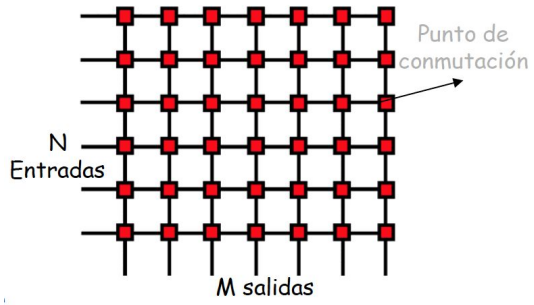




2.2 Topología: Redes de medio no compartido (II)

- **Redes de barras cruzadas (Crossbar Networks) de n entradas por m salidas (n*m)**
  - También se pueden usar **para conectar procesadores** entre sí.
  - Son redes **no bloqueantes:** El conjunto de entradas puede conectarse con el conjunto de salidas simultáneamente en cualquier permutación
  - Son fácilmente **escalables**, **aunque** muy **costosas** y solo son viables en dimensiones pequeñas





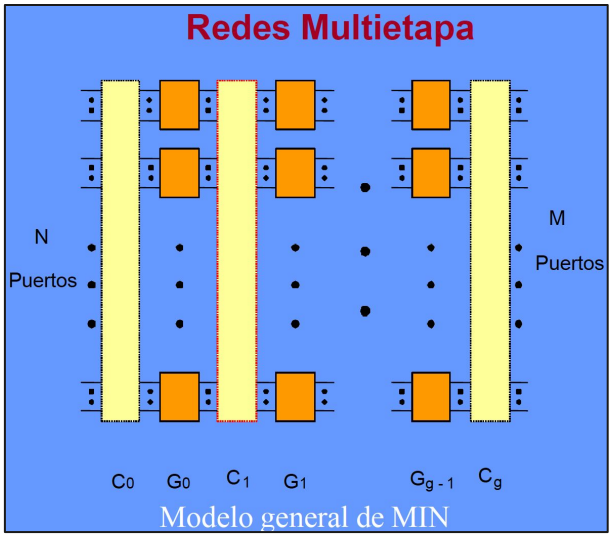
2.2 Topología: Redes de medio no compartido (III)

- **Redes multietapa (Redes MIN) de n entradas por m salidas (n*m):**
  - *Multistage Interconnection Network*
  - Están formadas por un conjunto de conmutadores y enlaces que se conectan siguiendo una **función de conexión** particular con ciertas propiedades.
  - La función de conexión debe **permitir conexiones desde todas las entradas a todas las salidas simultáneamente** sin repetir ninguna entrada ni salida. Se construyen con permutaciones de los bits que caracterizan cada entrada.

Estados de un conmutador 2x2:

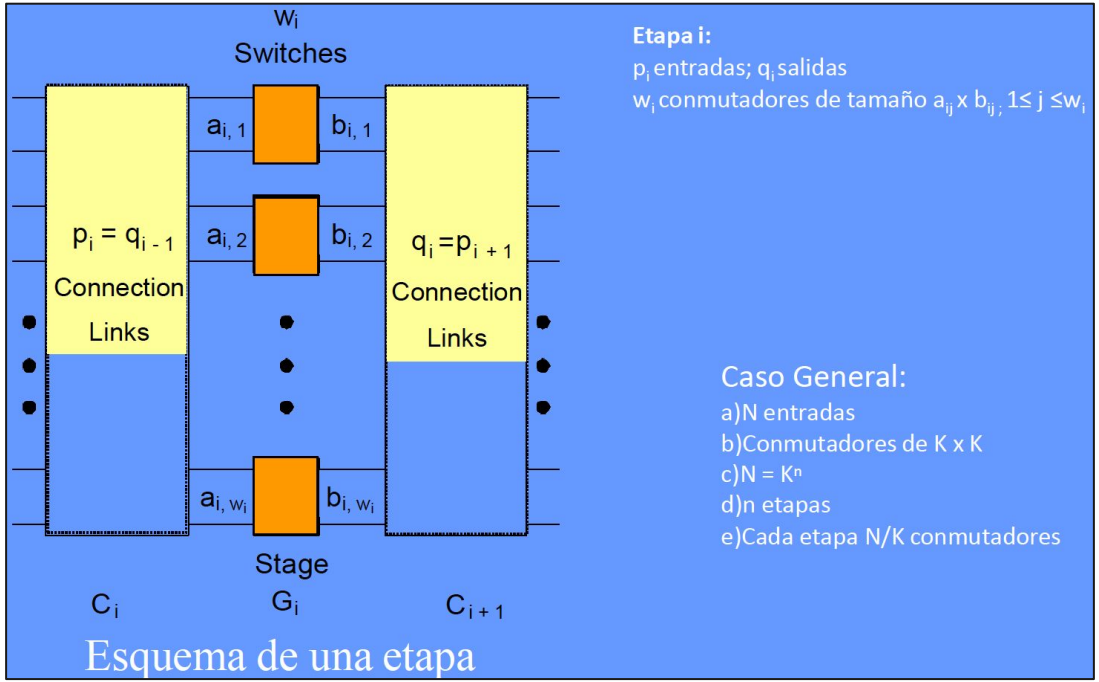
2.2 Topología: Redes de medio no compartido (IV)

- **Redes multietapa (Redes MIN) de n entradas por m salidas (n*m):**
  - Las funciones más conocidas es la de **baraje perfecto** (*perfect shuffle*) y la función de la permutación **mariposa** (*butterfly*)
  - Los conmutadores de la red permiten una conexión simultánea cruzada, paralela o de difusión.



2.2 Topología: Redes de medio no compartido (V)

- Redes multietapa (Redes MIN) de n entradas por m salidas (n*m):

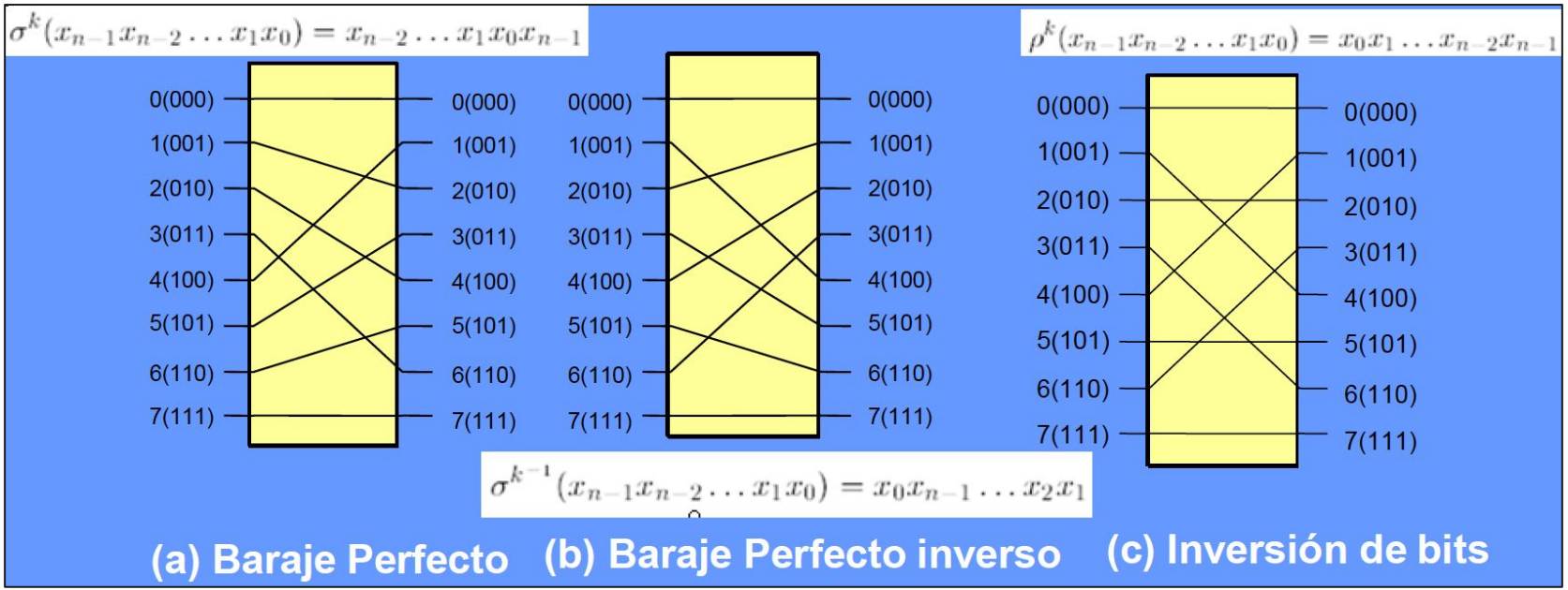


Esquema de una etapa

2.2 Topología: Redes de medio no compartido (VI)

- **Redes multietapa (Redes MIN) de n entradas por m salidas (n*m):**

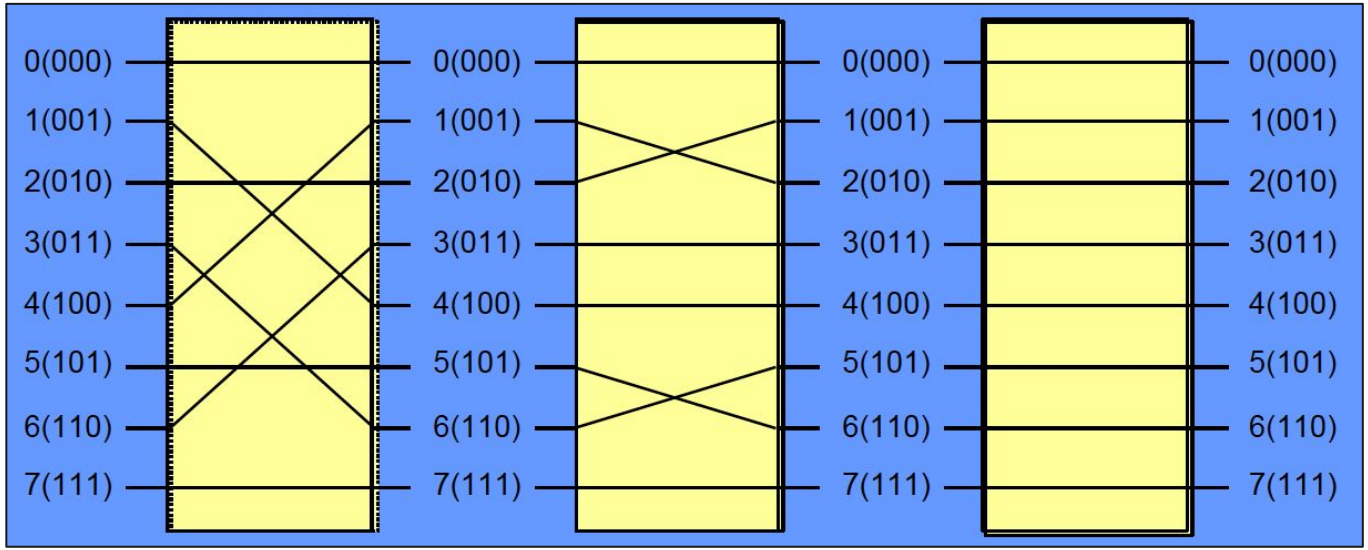
Esquemas de conexión:



2.2 Topología: Redes de medio no compartido (VII)

- **Redes multietapa (Redes MIN) de n entradas por m salidas (n*m):**

Esquemas de conexión:



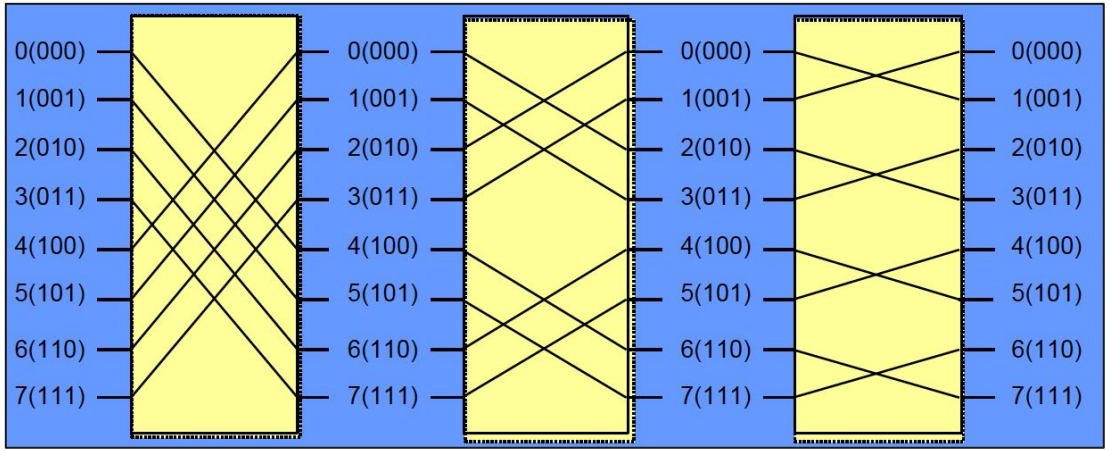
a) Mariposa segunda b) Mariposa segunda b) Mariposa cero

$$\beta_i^k(x_{n-1}\dots x_{i+1}x_ix_{i-1}\dots x_1x_0) = x_{n-1}\dots x_{i+1}x_0x_{i-1}\dots x_1x_i$$

2.2 Topología: Redes de medio no compartido (IX)

- **Redes multietapa (Redes MIN) de n entradas por m salidas (n*m):**

Esquemas de conexión:



a) Cubo segundo b) Cubo Primero c) Cubo Cero

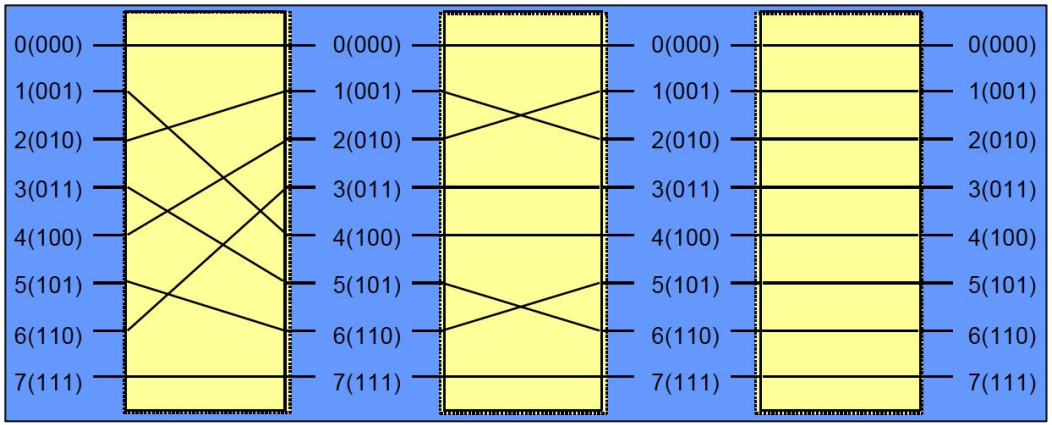
$$E_i(x_{n-1}\ldots x_{i+1}x_i x_{i-1}\ldots x_0) = x_{n-1}\ldots x_{i+1}\overline{x}_i x_{i-1}\ldots x_0$$

Conexión cubo

2.2 Topología: Redes de medio no compartido (X)

- **Redes multietapa (Redes MIN) de n entradas por m salidas (n*m):**

Esquemas de conexión:



a) Línea base segunda b) Línea base primera b) Línea base cero

$$\delta_i^k(x_{n-1}\dots x_{i+1}x_ix_{i-1}\dots x_1x_0) = x_{n-1}\dots x_{i+1}x_0x_ix_{i-1}\dots x_1$$

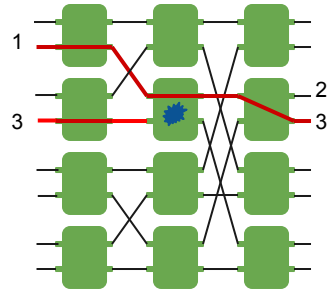
Conexión de línea base

2.2 Topología: Redes de medio no compartido (VI)

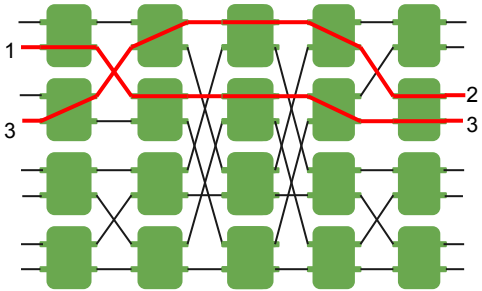
- Según disponibilidad de caminos:
    - Bloqueantes: Hay rutas que impiden la conmutación de otras
    - No bloqueantes: Todas las rutas se pueden simultanear
    - Reconfigurables

- Según el tipo de canales y conexiones:
    - Unidireccionales
    - Bidireccionales

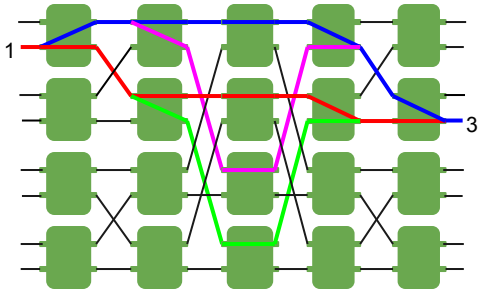
2.2 Topología: Redes de medio no compartido (VII)



Red de Mariposa

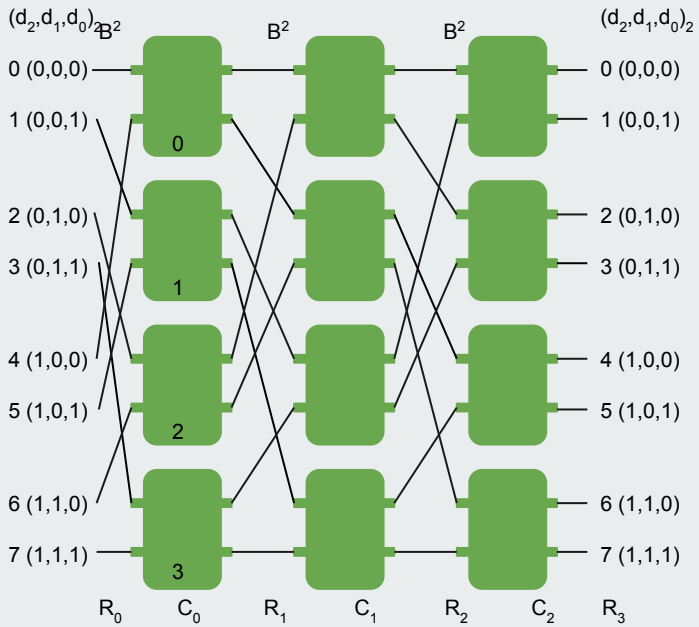


Red de Benes



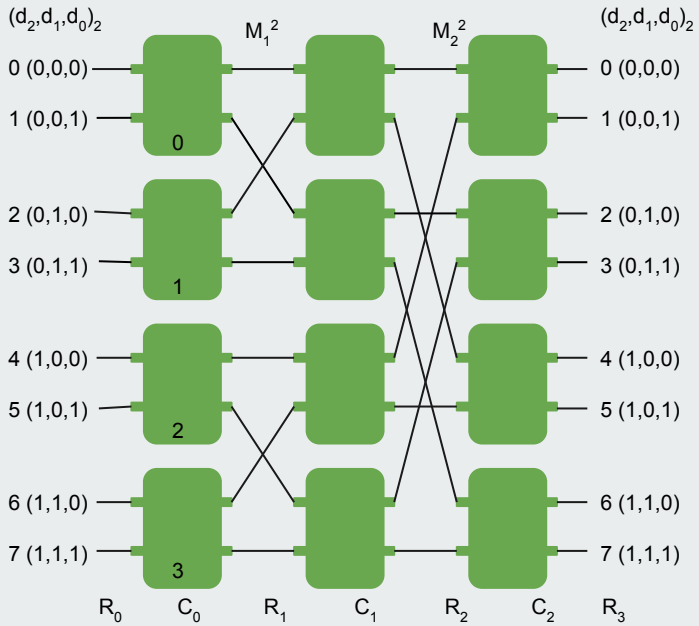
Red de Benes

2.2 Topología: Red omega



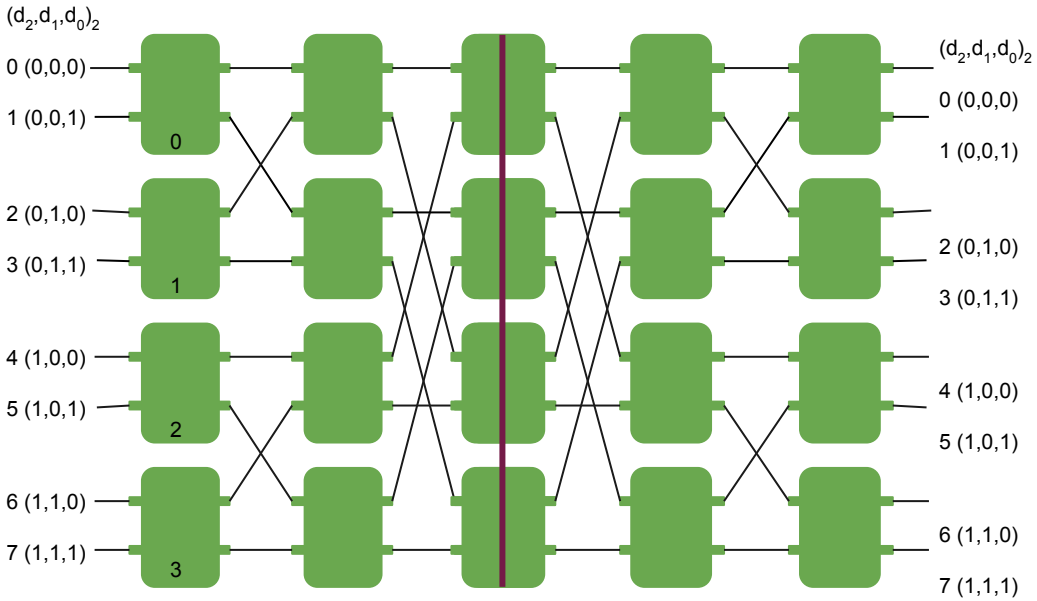
- Red de K^n*K^n ($8*8=2^3*2^3$), n = 3, k=2
  - Número de etapas, n, 3. Se nombran con una C y en base n, C_i
  - Conmutadores de k*k (2*2)
  - k^{n-1} conmutadores/etapa (2^2)
  - Subred R_i (i=0,...,n-1); baraje-k perfecto (baraje-2)
  - $B^k((f_{n-1},f_{n-2},...,f_1,f_0)_k)=(f_{n-2},...,f_1,f_0,f_{n-1})_k$
  - $B^k((f_2,f_1,f_0)_2)=(f_1,f_0,f_2)_2$

2.2 Topología: Red mariposa

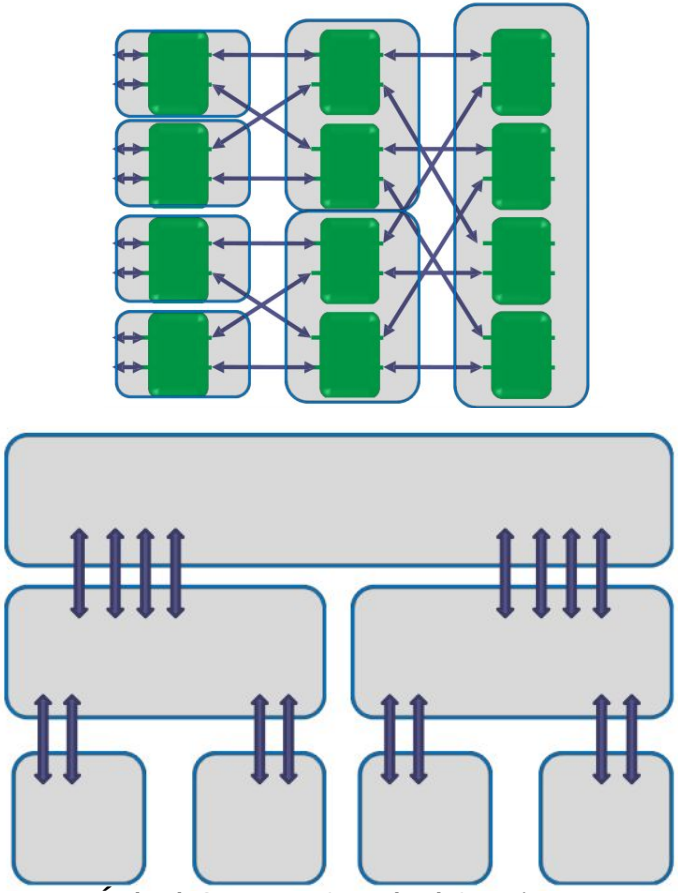


- Red de K^n*K^n ($8*8=2^3*2^3$), n = 3, k=2
  - Número de etapas, n, 3. Se nombran con una C y en base n, C_i
  - Conmutadores de k*k (2*2)
  - k^{n-1} conmutadores/etapa (2^2)
  - Subred R_i (i=0,...,n-1); Mariposa M_i^k
  - $M_i^k((f_{n-1},f_{n-2}, ..., f_i, ...,f_1,f_0)_k)=(f_{n-1},f_{n-2}, ..., f_0, ...,f_1,f_i)_k$
  - $M_2^2((f_2,f_1,f_0)_2)=(f_0,f_1,f_2)_2$
  - La red Mariposa puede ser unidireccional o bidireccional
- Mariposa bidireccional, equivale a una Red de benes reconfigurable y también a una red de árbol grueso

2.2 Topología: Red mariposa



Red de Benes 8*8



Árbol Grueso 3-Arbol 2-ario

2. Propiedades: Características

1. Diámetro: Máxima longitud (nº de enlaces atravesados) de entre los caminos óptimos (más cortos). **Diámetro grande** implica **poca** habilidad de **comunicación** entre nodos. Se buscan diámetros pequeños.

2. Ancho de bisección (b): **Mínimo de enlaces a cortar para dividir a la red en dos mitades similares** incluyendo el mismo número de conmutadores y de nodos. Si cada enlace es capaz de transportar *w* bits, el ancho de banda de la bisección será *bw*.

2. Propiedades: Características

3. Latencia: **Retraso máximo** producido por la comunicación d**e un mensaje** pequeño **entre dos nodos** cualesquiera de una red. También se le denomina contención y está relacionado con los tiempos de espera producidos durante el transporte de los datos.

4. Productividad: Nº **total de paquetes** de información que una red puede transportar **por unidad de tiempo**. Hay que tener cuidado con los puntos calientes (hot spot) que son los nodos o enlaces donde se concentra la mayor parte del tráfico de una red.

2. Propiedades: características

5. Escalabilidad: **Facilidad** con la que una red puede **extenderse** sin que sus prestaciones se vean afectadas.

6. Conectividad: Una red es **totalmente conectada** si existe una **conexión directa entre cualquier par** de nodos.

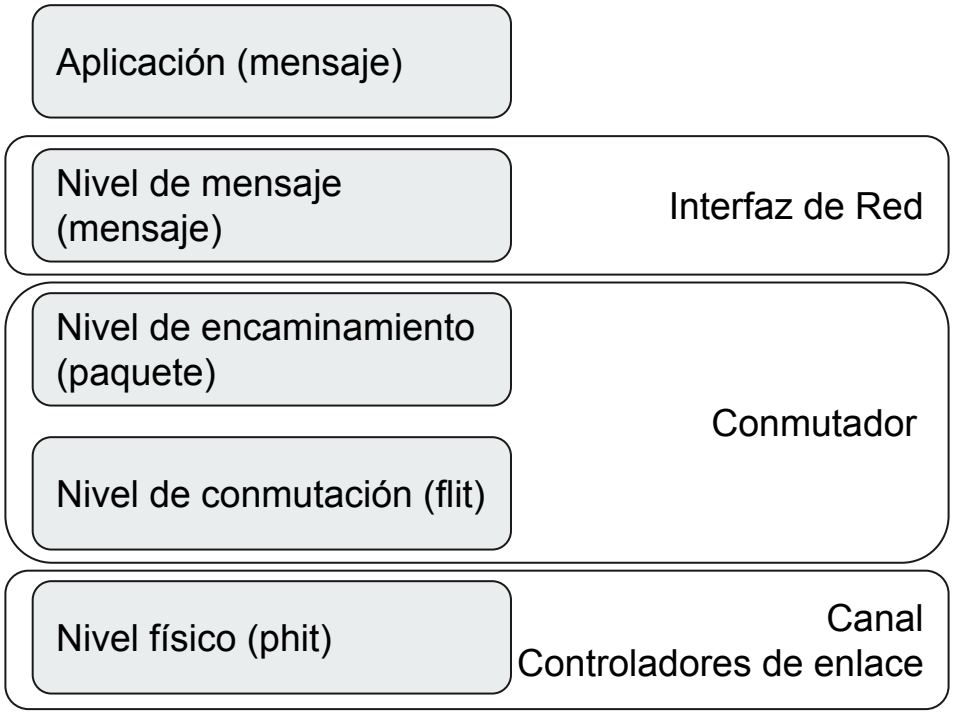
7. Grado de los nodos: **Número de enlaces** que tiene un nodo conectado con otros nodos. Se puede hablar de grado de entrada (GE), enlaces entrantes desde otros nodos y grado de salida (GS). En este caso el grado es GE+GS. Si el grado de todos los nodos es el mismo, la red es regular

2. Propiedades: características

8. Niveles de servicio: Nivel al que se trata la red:

- Físico
- De conmutación
- De encaminamiento
- De mensaje
- De aplicación

Según el nivel de servicio, la **unidad de comunicación** cambia.



2. Propiedades: características

9. Calidad del servicio (QoS):

- Definición 1: Efecto global de las **prestaciones** de un servicio que determinan el grado de **satisfacción** de quien la utiliza.
- Definición 2: Conjunto de **requisitos del servicio** que debe cumplir una red en el transporte de flujo.

La calidad del servicio se mejora realizando una buena **gestión de la congestión**, teniendo un bajo nivel de retardo, un alto rendimiento y un coste del servicio justo.

2. Propiedades: características

9. Calidad del servicio (QoS): Algunas de las medidas que se usan para determinar la calidad del servicio son:

- Disponibilidad: Tiempo mínimo para asegurar que una red estará en funcionamiento (99,99%)
- Ancho de banda: El mínimo ancho de banda que el operador garantiza (2Mbps)
- Pérdida de paquetes: El número máximo de paquetes perdidos, siempre que el volumen de comunicaciones no exceda de cierto valor (1,1%)
- Round Trip Delay: Retardo de ida y vuelta medio en los paquetes (80ms)
- Jitter: Fluctuación producida en el retardo de ida y vuelta (20ms) que retarda normalmente la formación de mensajes, por el retardo de los diferentes paquetes en los que se divide.

2. Propiedades: características

9. Calidad del servicio (QoS): Qué influye en la calidad del servicio
    - Algoritmos de encaminamiento
    - Controladores de enlace
    - Políticas de prioridades
    - Políticas de desbloqueos
    - Políticas de desestimación de paquetes
    - Cantidad y número de almacenamientos intermedios para unidades de transmisión
    - ...

2. Propiedades: características

10. Alta disponibilidad: Es una medida del **porcentaje del tiempo** que una **red** está **operativa** por mes o por año, o por día…

11. Tolerancia a Fallos: Es la **capacidad que tiene una red de seguir prestando** servicio a pesar de que alguna de sus partes no esté operativa. **Enrutamientos adaptativos.** Una red tolerante a fallos, tiene que ofrecer caminos alternativos entre cada par origen-destino.

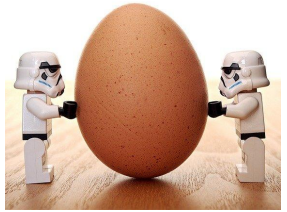
12. Fiabilidad: Es el **grado de seguridad** que podemos tener de que una transmisión llegará al destino bien, es decir, no se perderá, no llegarán los datos corruptos …

2. Propiedades: características

13. Remote Direct Memory Access (RDMA): Establece qué **tipo de acceso a memoria no local** se le permite a cada uno de los nodos que están conectados a la red, ya sea a través de un procesador o del mismo sistema de gestión de la memoria de cada nodo.

- Este acceso puede ser de lectura o de lectura/escritura, permitiendo en muchos casos la realización de operaciones atómicas.
- Si la red no permite este acceso, **la alternativa es el uso de mensajes**.
- Esta característica **permite**, por ejemplo, el **acceso desde una GPU a la memoria de una CPU**.

Trabajo para la próxima semana



Buscar información de la siguiente red HPC de un equipo Top500:

- Grupo 1: Tofu Interconnect D
- Grupo 2: Infiniband EDR
- Grupo 3: Sunway
- Grupo 4: TH Express-2
- Grupo 5: Cray/HPE
- Grupo 6: Mellanox HDR Infiniband
- Grupo 7: Dual-rail Mellanox EDR Infiniband

=> Un miembro del grupo presentará su trabajo la próxima semana. Habrá que entregar tanto la presentación como una transcripción de lo que se dice (incluyendo las fuentes).


Gracias.