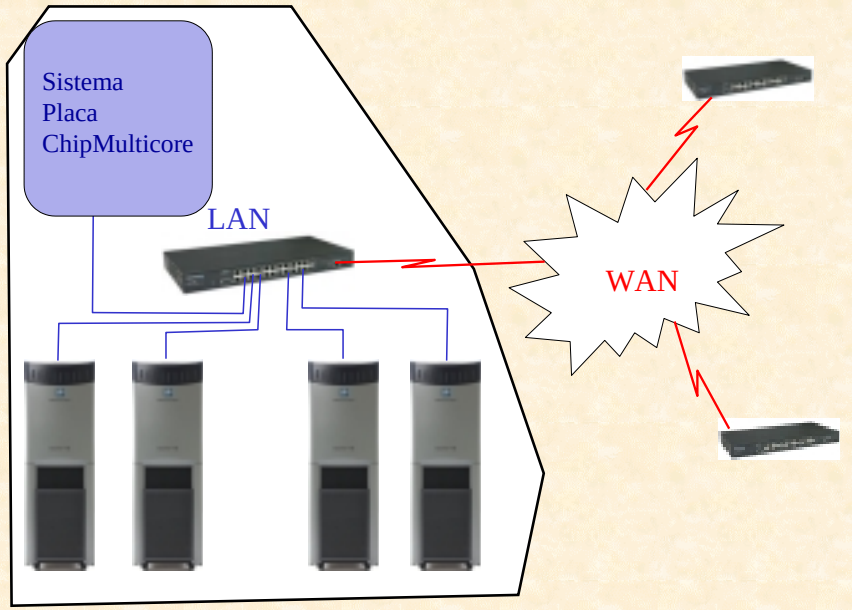
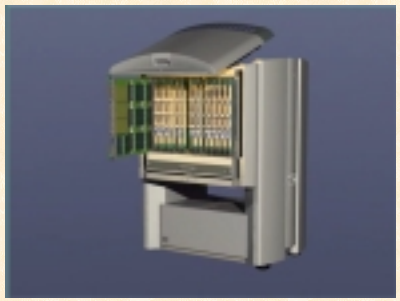
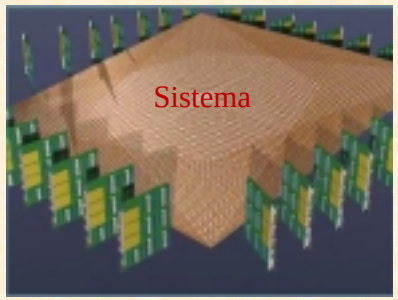
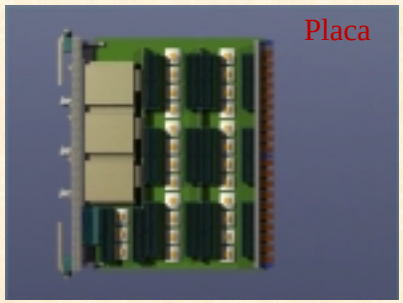
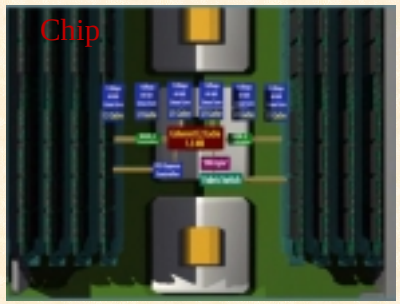






www.sicortex.com





- LAN/WAN Internet
  - Millones de nodos
  - # Nodos dinámico
  - Enlaces largos
  - Red irregular
  - Latencia alta
- Mutiprocesadores …
  - Cientos .. Miles
  - *Fijo*
  - Cortos
  - Regular
  - Baja

2 CONECTIVIDAD

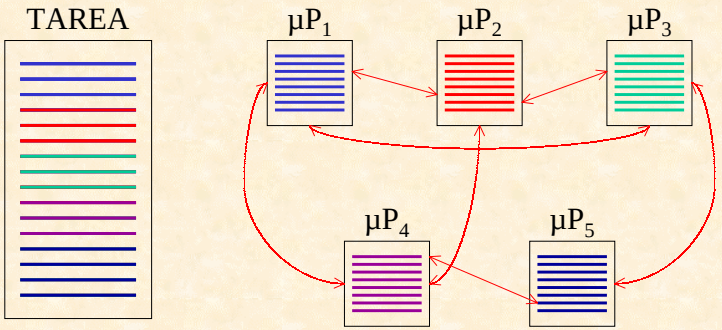
"Interconnection Networks. An engineering…" José Duato y …- 2003 [Capítulos 1 y 2]

"Principles and Practices of Interconnection ..." William James Dally y … - 2004 [Cap: 1,2,3,22]

- 1 Necesidad
- 2 Conceptos
  - 1 Clasificación de las redes
  - 2 Caracterización por Grafos
  - 3 Perfiles de comunicación
- 3 Redes de medio de transmisión compartido (Buses)
- 4 Redes directas (estáticas)
  - 1 Encaminamiento
  - 2 Array lineal, anillo, ..., hipercubo
- 5 Redes indirectas (dinámicas)
  - 1 Crossbar, redes multietapa (Ω)

QUEREMOS MÁS VELOCIDAD:

A menor Grano, mayor Grado

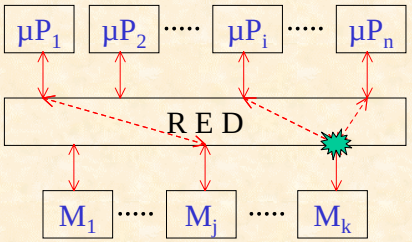


! AUMENTAN LAS NECESIDADES DE COMUNICACIÓN !

Comunicación Hw <===> Comunicación Sw

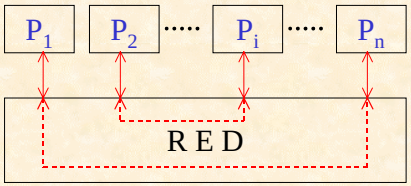
Memoria Común (Load/Store)

Comunicar µPi y Memoria



Paso Mensajes (Send/Receive)

Comunicar Pi con Pj



Es muy importante la Latencia y el Ancho de banda

www.euroben.nl/reports/web07/networks.html

Network	Bandwidth GB/s	Latency µs
Cray SeaStar2 (measured)	2.1	4.5
IBM (Infiniband) (measured)	1.2	4.5
SiCortex Kautz graph (stated)	2.0	1.0
SGI NumaLink (measured)	2.7	1.2
Infiniband (measured)	1.3	4.0
Infinipath (measured)	0.9	1.3
Myrinet 10-G (measured)	1.2	2.1
Quadrics QsNet^{II} (measured)	0.9	2.7
Gigabit Ethernet	0,1	29..120

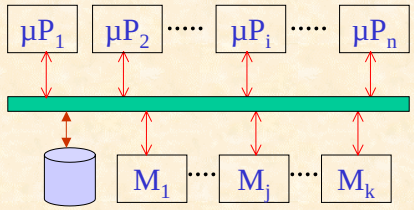
Coste * 50

¡ LA RED TIENE UNA IMPORTANCIA VITAL !

- CLASIFICACIÓN DE LAS REDES
  - MEDIO DE TRANSMISIÓN COMPARTIDO
  - DIRECTAS vs INDIRECTAS
  - TOTAL vs PARCIALMENTE CONECTADAS
- CARACTERIZACIÓN POR GRAFOS
  - GRADO Y DIÁMETRO
- PERFILES DE COMUNICACIÓN
  - 1 => 1; N => N; 1 => N; N => 1

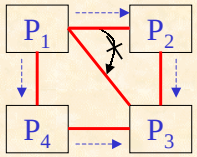
- Medio de Transmisión Compartido: Ponerse de acuerdo en su uso (maestro/esclavo, …)

Redes locales | Redes inalámbricas | Buses (Backplane)

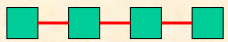


- Síncronos vs asíncronos
- Multiplexados
- Arbitraje del bus

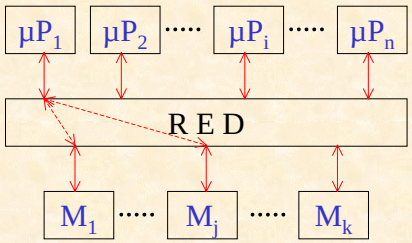
- Redes directas: Conexiones fijas entre los elementos (Pi, Pj) "invariables durante la ejecución"

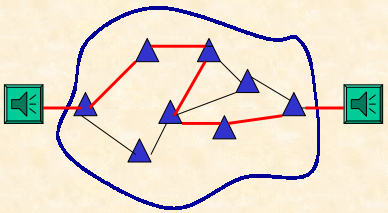


- Acoplamiento débil
- Amplio uso en multicomputadores
- Los propios Nodos encaminan
- Los caminos del origen al destino pueden ser distintos



- Redes indirectas: Conexiones varían entre los elementos (µPi, Mj) "variables durante la ejecución"

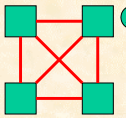




Red Telefónica

- Acoplamiento fuerte
- Amplio uso en multiprocesadores
- Encamina la propia red

- Totalmente conectadas:
  "Cada elemento tiene conexión directa con los demás"

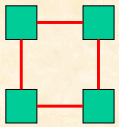


Latencia mínima (L_m)

Alto coste $O(n^2)$

No escalable

- Parcialmente conectadas:
  ¡ conexas !

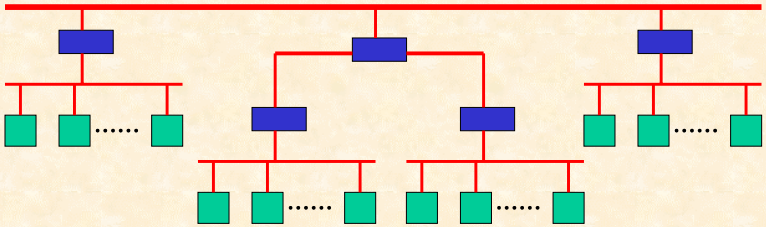


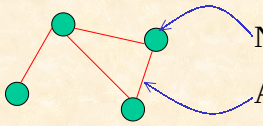
Mayor latencia ($2L_m$)

Menor coste $O(n)$

Encaminar más complejo

- Jerarquizadas: Aislar tráfico por "localidades"

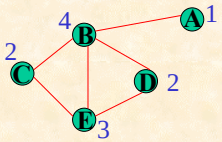




Nodos => µP y/o Bancos de Memoria

Aristas => Enlaces de comunicación

- Grado de un nodo: Líneas incidentes (Si unidireccionales Ge + Gs)



- Relacionado con el número de puertos E/S y, por lo tanto, con el coste
- Deseable constante y pequeño

- Grado de la red: El del nodo con mayor grado (4)
  - Deseable regularidad
  - Compromiso en el Grado

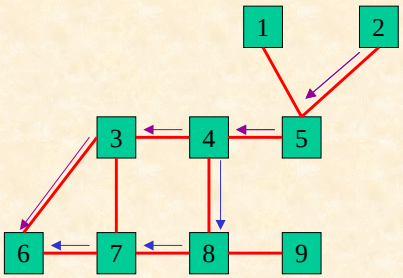
Más conectividad => Menor latencia
Mayor coste

Menor conectividad => Más latencia
Menor coste

- Diámetro de la red: Camino más distante de entre los mínimos que unen a dos nodos cualesquiera.

¿1, 2, 3, ...?

- Métrica => Número de saltos => 2



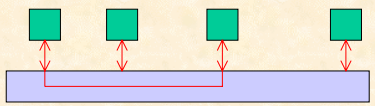
¿5? => 2, 5, 4, 8, 7, 6

4 => 2, 5, 4, 3, 6 más corto

- Relación directa con la latencia

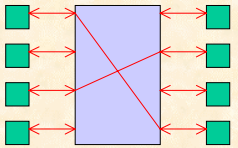
- Enlaces de comunicación establecidos concurrentemente.

1 => 1 Ventanilla única

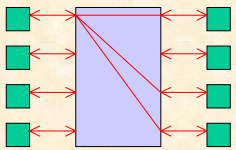


Bus Común

N => N Varias Ventanillas

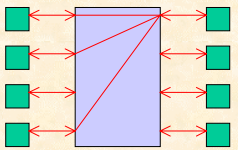


1 => N T.V. News

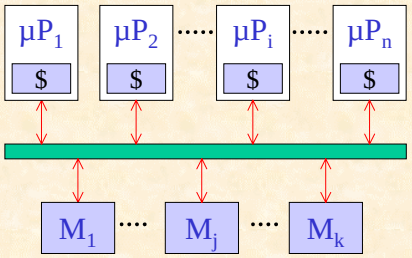


Difusión, Broadcast, Multicast

N => 1 Reducción



Máquinas CRCW

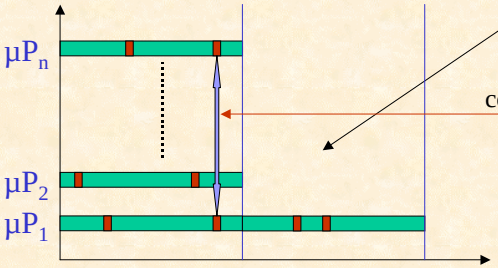


¿Cuántos Pi podré instalar?

Pentium 4 a 3,8GHz
Bus de 64 bits y 800MHz
¿Un único Pi satura el Bus?

$ ¡ Cachés !

≅ 98% Hit



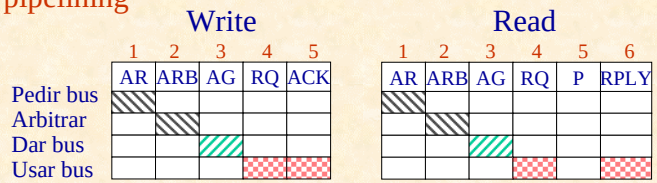
¡ Algunos problemas !

colisiones

Fallo costoso

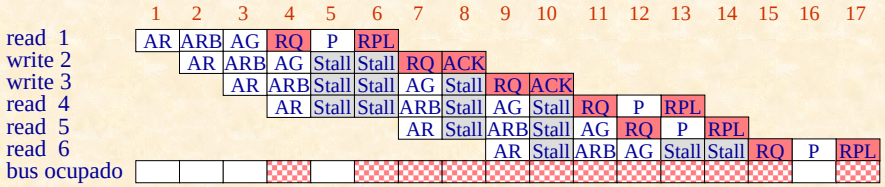
¿ Soluciones ?

- Bus pipelining

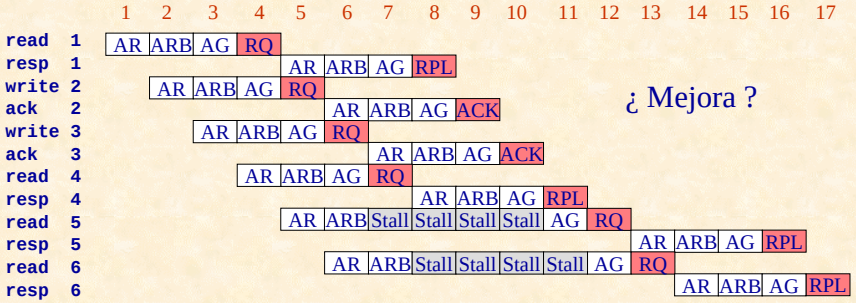


¿Cuántos ciclos 2W y 4R?

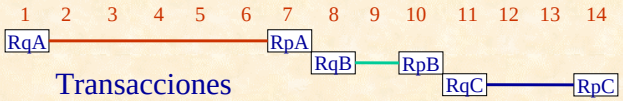
Con pipeline mejor



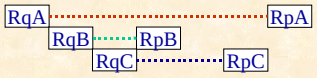
- Split transaction: Pipelining + Dividir la transacción en dos



¿ Mejora ?



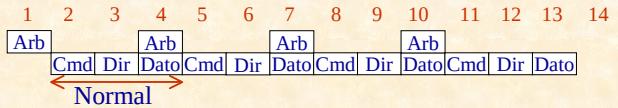
Transacciones variables: 1..6 ciclos

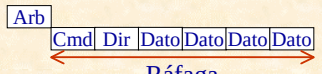


8 peticiones pendientes en SGI

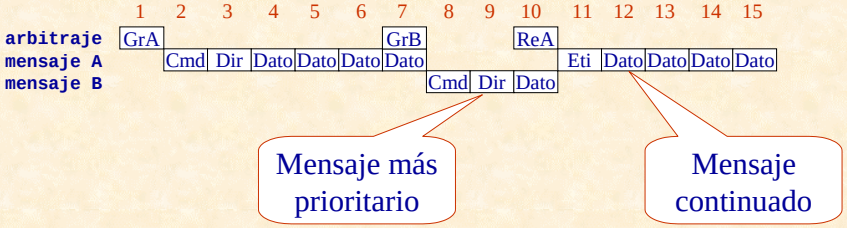
112 peticiones pendientes en SUN E 6000

- Modo ráfaga (Burst): Transacciones largas (línea de caché)





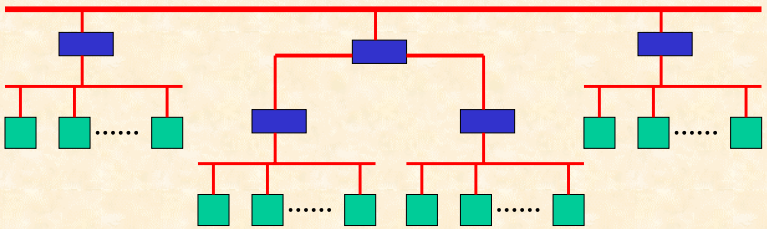
¿ Inconveniente ?



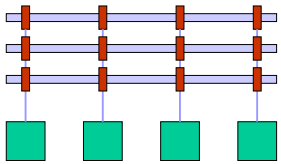
Mensaje más prioritario

Mensaje continuado

- Buses jerárquicos



- Buses múltiples

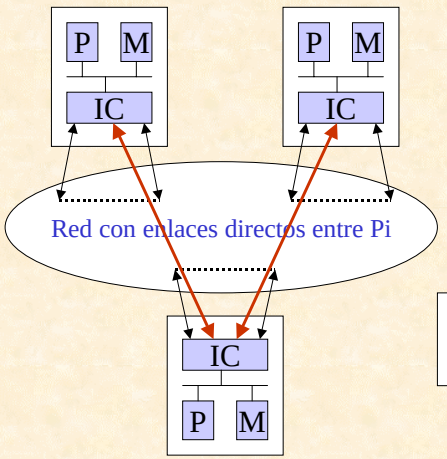


Concluyendo
Cachés (L1, L2 y L3)
Pipelining
Split Transaction
Modo ráfaga
Buses Jerárquicos
Buses Múltiples
Muy costoso + 32µP

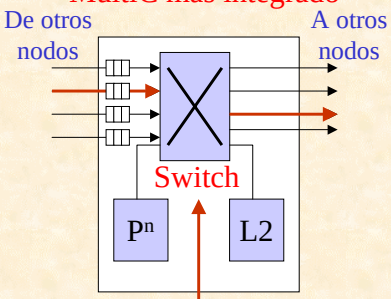
Difusión
Serialización

Frecuencia
Secuencial

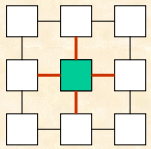
- Generalidades
- Encaminamiento
- Menor diámetro aumentando el grado
  - Array lineal
  - Anillo simple y de grado “n”
  - Conectividad total
- Compromiso grado vs diámetro y muchos nodos
  - Árbol, Fat Tree y Estrella
  - Mallas y Toroides
  - Hipercubo con y sin ciclo
- Tabla de parámetros



MultiC más integrado

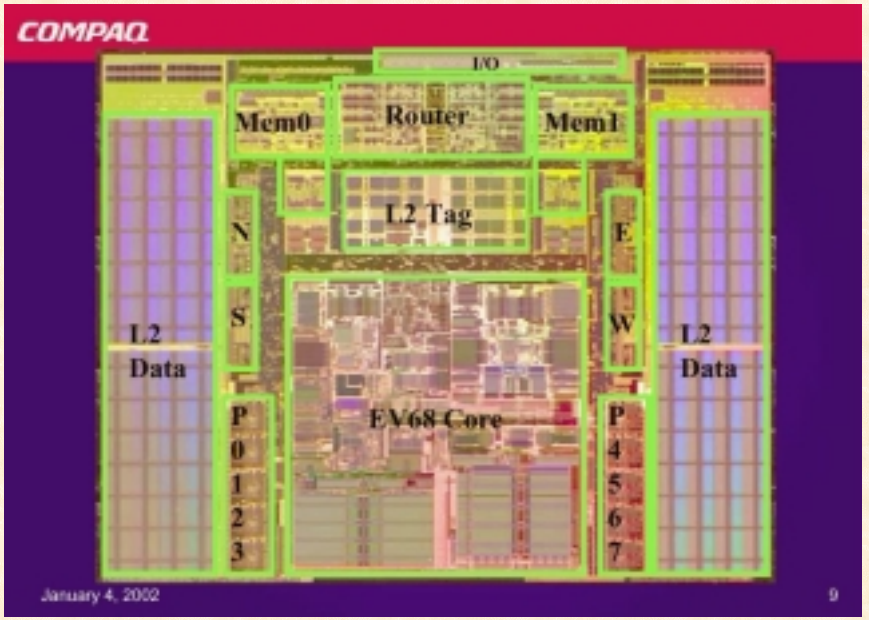


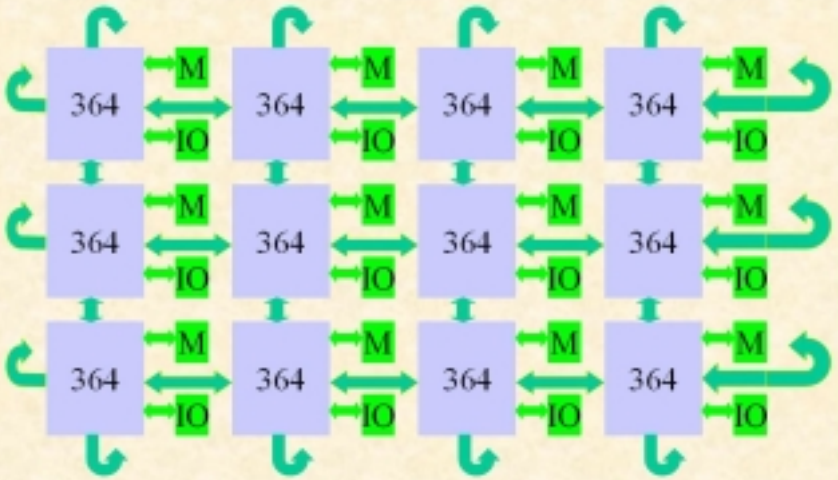
- Buffers
- Arbitraje
- Encamina.

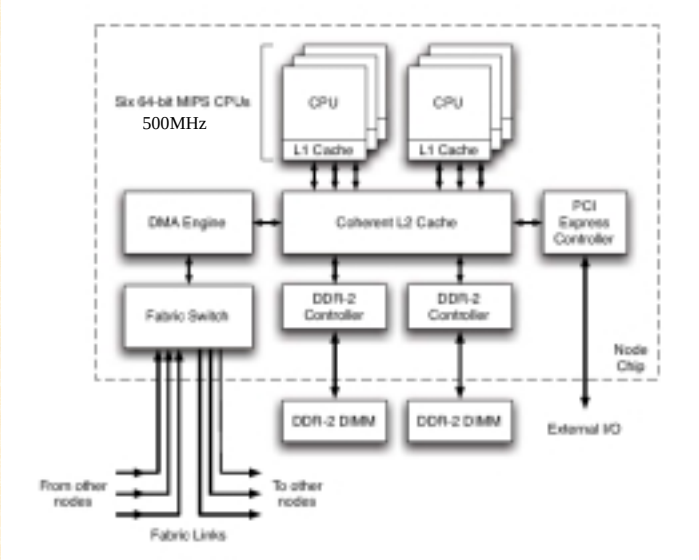


Nodos => PC’s o similares

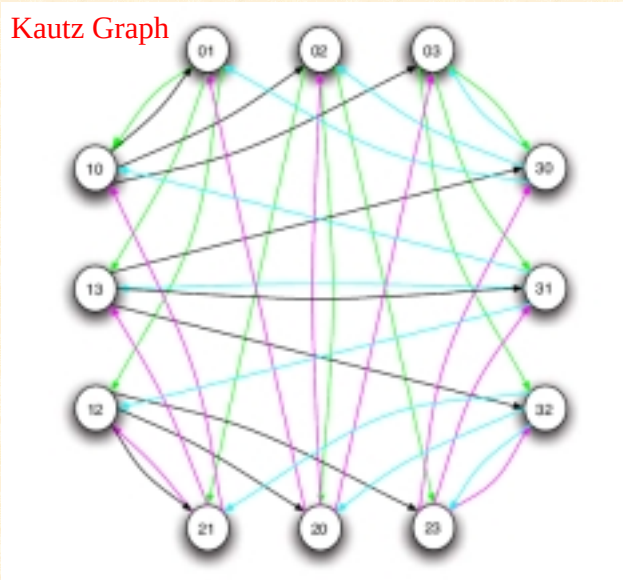
Ejemplos: Alpha 21364, SiCortex







www.sicortex.com

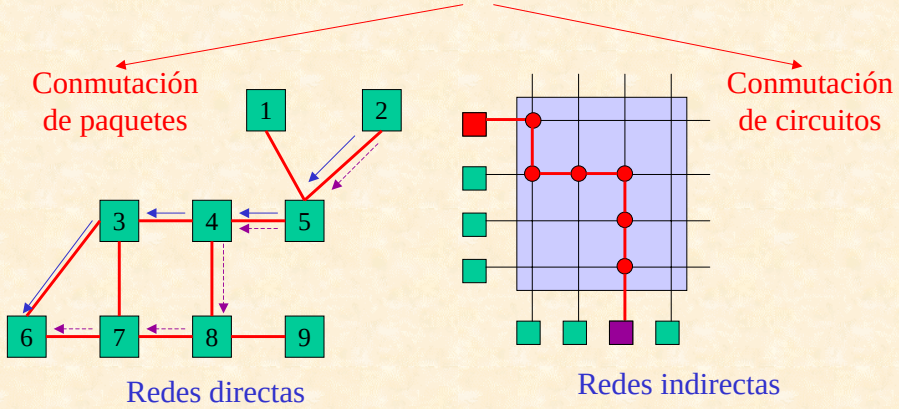


- Mecanismo Hw/Sw para que la información llegue del origen al destino.

Hay que distinguir entre:

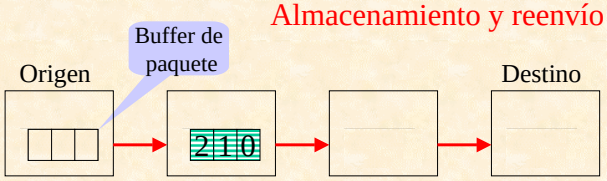
Algoritmo: Elección del camino y gestión de conflictos

Técnica: Modo de propagar la información



- En conmutación de paquetes veremos dos técnicas:

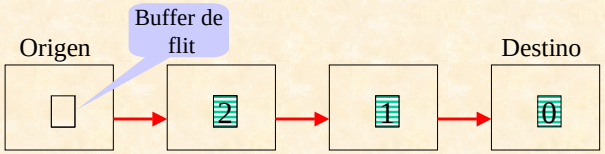
Almacenamiento y reenvío



Los mensajes se dividen en paquetes (64..1024bits) y se envían paquete a paquete

Elevada latencia (3*Tiempo trans. Paquete "Ttp")

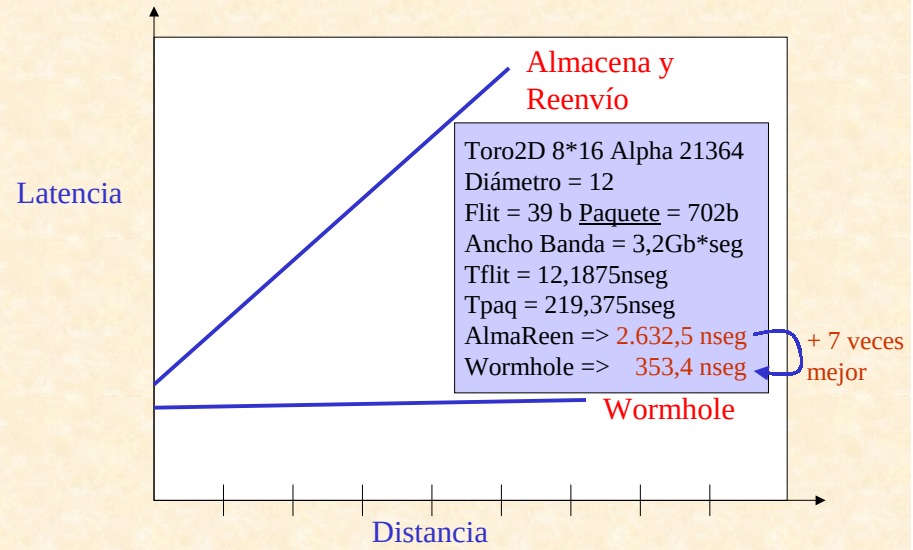
"Wormhole"

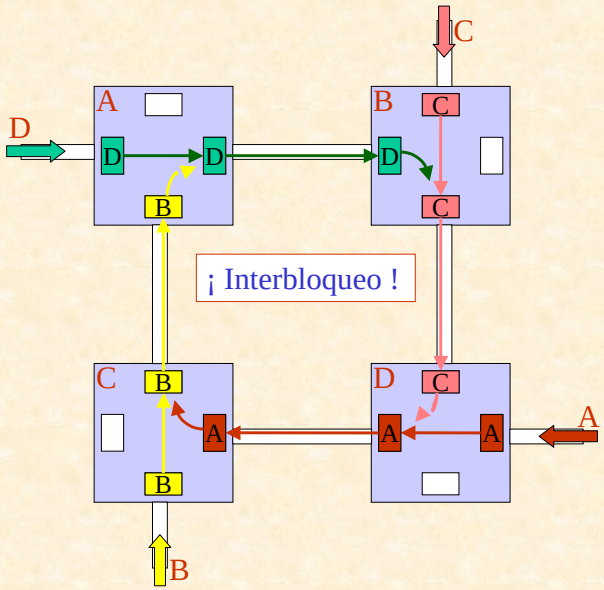


Los paquetes se dividen en flits (2..32 bits) y se envían flit a flit

Mejora la latencia (2*Tiempo trans. Flit + Ttp)

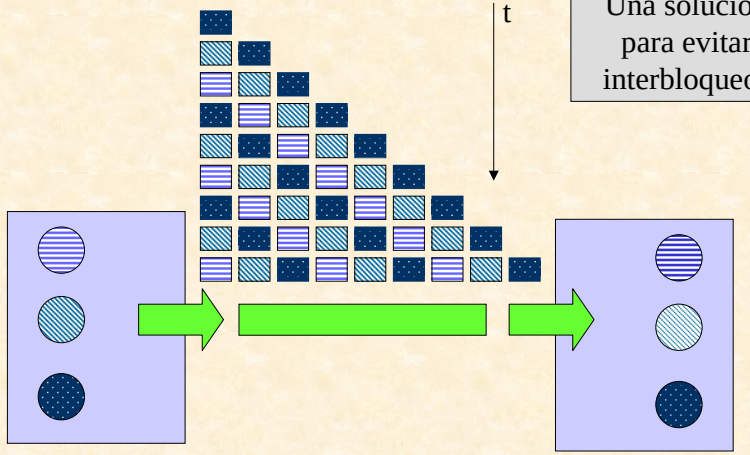
¿Similar a IP/ATM MPLS?



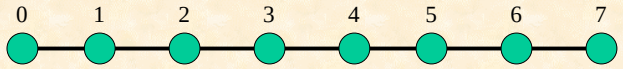


CANALES VIRTUALES

Una solución para evitar interbloqueos



ARRAY LINEAL



'N' nodos, 'N-1' enlaces

Grado de los nodos: 2, 1

Grado de la red: 2

Diámetro: N-1

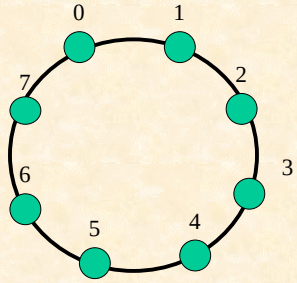
Escalable: SI, pero a costa de demasiado diámetro

Comentarios:

No es un bus

Ineficiente con 'N' grande

ANILLO (DE GRADO 2)



'N' nodos, 'N' enlaces

Grado de los nodos: 2

Grado de la red: 2

Diámetro: $\lfloor N/2 \rfloor$ (bidireccional)

Escalable: SI, pero a costa de demasiado diámetro

Comentarios:

Ineficiente con 'N' grande

ANILLO (DE GRADO 'n')

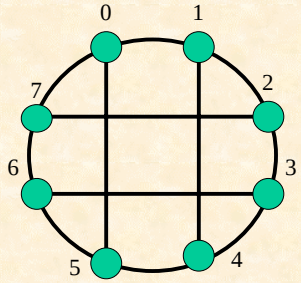
'N' nodos

Grado de los nodos: n

Grado de la red: n

Diámetro: ?????

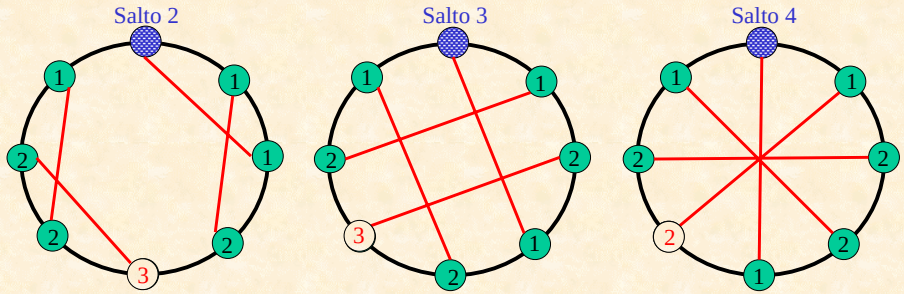
Escalable: SI, pero a costa de demasiado grado



Comentarios:

Unión de cada nodo 'N' con un nodo situado a 'x' enlaces de distancia

N = 8 n = 3



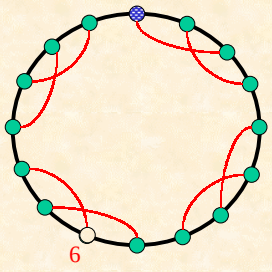
d = 3, $\overline{d}$ = 1,71

d = 3, $\overline{d}$ = 1,71

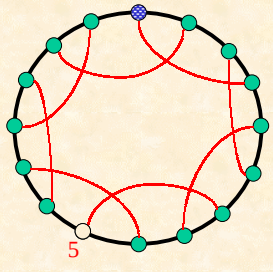
d = 2, $\overline{d}$ = 1,57

N = 16 n = 3

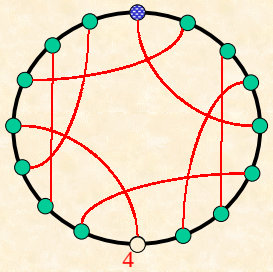
Salto 2



Salto 3



Salto 4



d = 6, $\overline{d}$ = 3,2

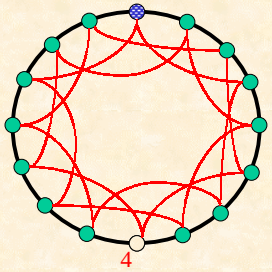
d = 5, $\overline{d}$ = 2,67

d = 4, $\overline{d}$ = 2,27

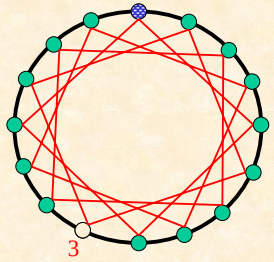
Salto 5 iguala y 7 y 8 empeoran

N = 16 n = 4

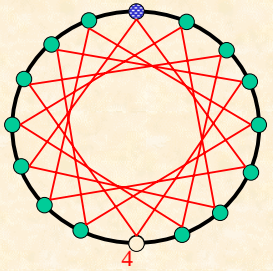
Salto 3



Salto 4



Salto 5



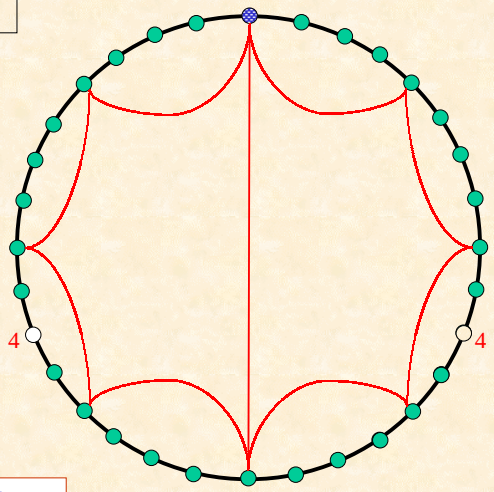
d = 4, $\overline{d}$ = 2,13

d = 3, $\overline{d}$ = 2

d = 4, $\overline{d}$ = 2,13

¿Cómo podría ser N=32 y n=5?

N = 32 n = 5



¿ Escalable ?

d = 4, $\overline{d}$ = ???

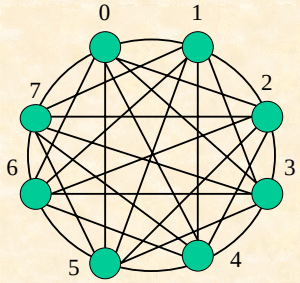
‘N’ nodos

Grado de los nodos: N-1

Grado de la red: N-1

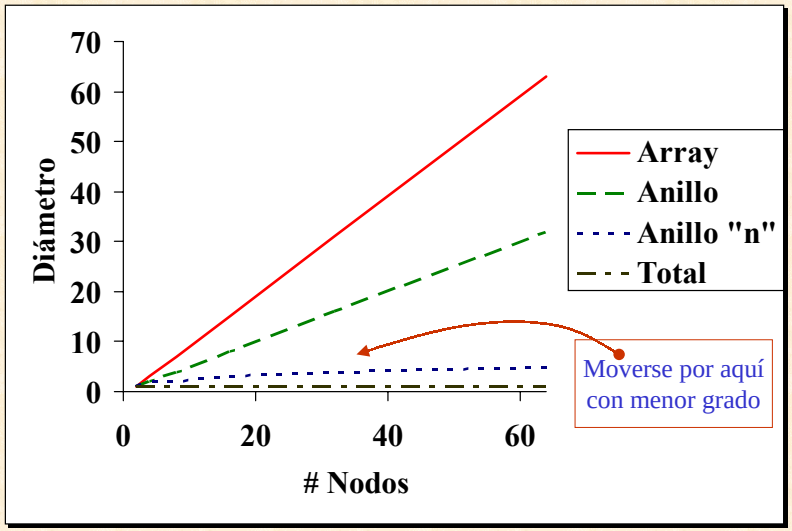
Diámetro: 1

Escalable: NO



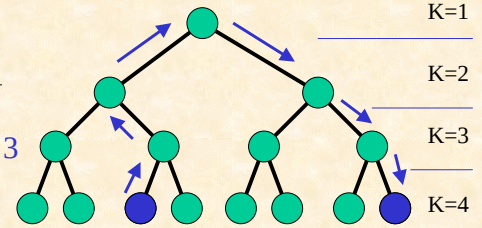
Comentarios:

Inviable con N algo grande



ÁRBOL BINARIO

'N=2^k-1' nodos
'k' dimensiones



Grado de los nodos: 1, 2, 3

Grado de la red: 3

Diámetro: 2*(K-1)

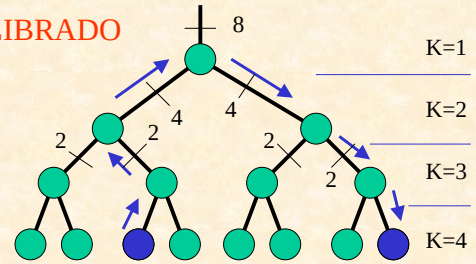
Escalable: SI

Comentarios:

Ineficiente con 'N' grande

ÁRBOL BINARIO EQUILIBRADO
"Fat Tree"

'N=2^k-1' nodos
'k' dimensiones



Grado de los nodos:
$1, 2^2, 2^3, 2^4, \ldots 2^K$

Grado de la red: 2^K

Diámetro: 2*(K-1)

Comentarios:

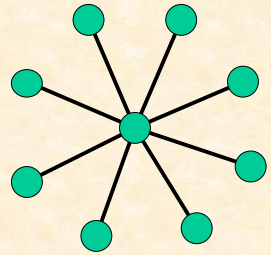
Escalable: NO

Grado muy grande en los nodos superiores

Grado de los nodos no equilibrado

ESTRELLA

'N' nodos, 'N-1' enlaces



Grado de los nodos: 1, N-1

Grado de la red: N-1

Diámetro: 2

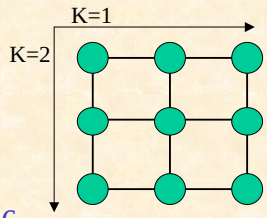
Escalable: SI, pero a costa de demasiado grado

Comentarios:

Arbol no binario de 2 niveles

Grado excesivo en el nodo central

MALLAS (2D y 3D)



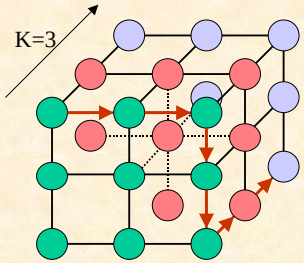
'N' nodos ('n' por dimensión), 'k' dimensiones, $N=n^k$

Grado de los nodos: 2, 3, 4, 5, 6

Grado de la red: 2*k

Diámetro: K*(n-1)

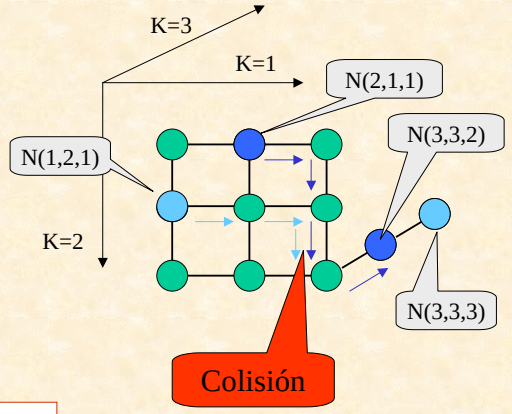
Escalable: SI, pero a costa de grado elevado o demasiado diámetro



M3D 512 $_{8*8*8}$ => D=21

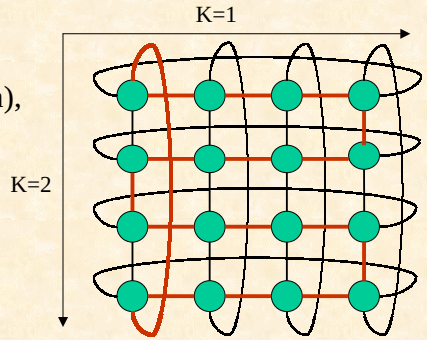
¿Escalabilidad cuadrática o cúbica?

MALLAS (encaminamiento: ordenado por direcciones)



¿ Interbloqueos ?

TOROIDES (2D y 3D)



'N' nodos ('n' por dimensión), 'k' dimensiones, $N=n^k$

Grado de los nodos: 2*k

Grado de la red: 2*k

Diámetro: $K * \lfloor n/2 \rfloor$

Escalable: SI, pero a costa de elevado grado o mucho diámetro

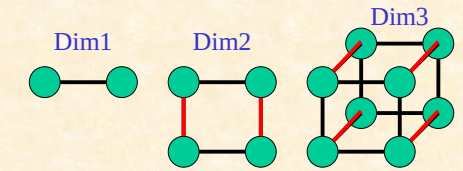
T3D 512 $_{8*8*8}$ => D=12

Comentarios:

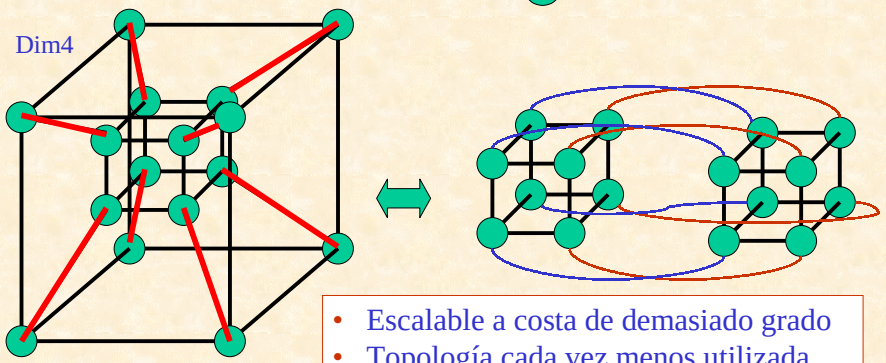
Combina características de malla y anillo

¡ Anillo embebido !

HIPERCUBO 'N=2^k' nodos, 'k' dimensiones = $\log_2$ N

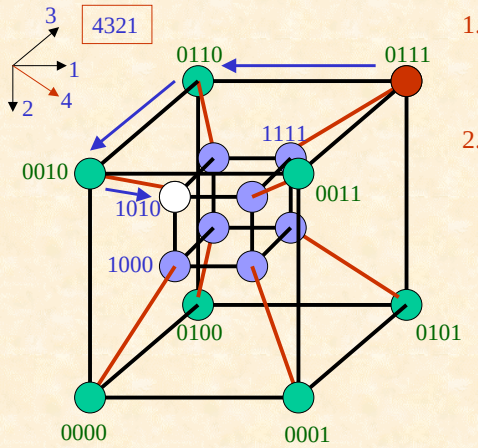


- Diámetro = $\log_2$ N
- Grado = $\log_2$ N
- Fácil encaminar



- Escalable a costa de demasiado grado
- Topología cada vez menos utilizada

Encaminamiento en HIPERCUBO (Sea N=16)



1. Numerar nodos en binario. Nodos adyacentes difieren en un bit (el asociado a la dirección que les une)
2. Enviar mensaje por el enlace asociado a la menor dirección donde no coinciden bit del nodo actual y bit del nodo destino

¿ Realizar ORX ?

0111 ORX 1010 = 1101

Nodo actual	0111	0110	0010	1010
Nodo destino	1010	1010	1010	1010

HIPERCUBO CON CICLOS

'N=k*2ᵏ' nodos, 'k' dimensiones

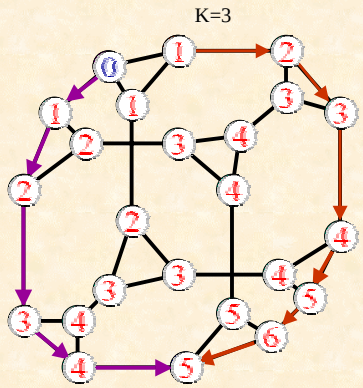
Grado de los nodos: 3

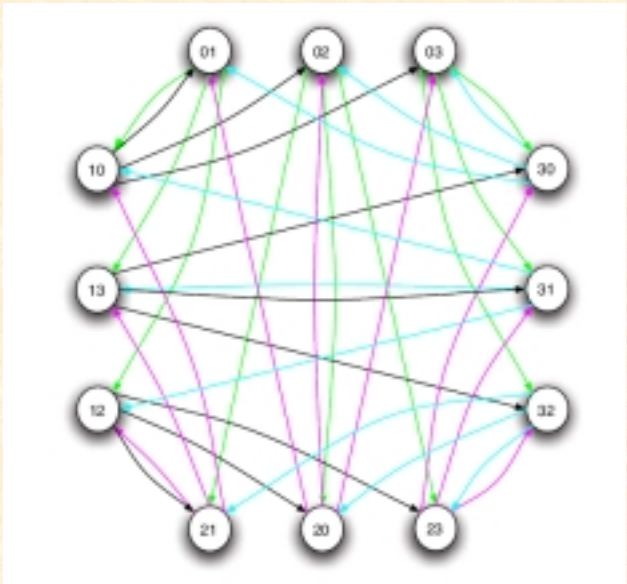
Grado de la red: 3

Diámetro: 2*k - 1 + ⌊k/2⌋

Escalable: Más escalable que el hipercubo sin ciclos

Comentarios:
Mejora el grado, pero empeora el diámetro





- ¿Cómo conectar unos 512 nodos?

Topología	Diámetro	Grado
M3D 8*8*8	21	6*
T3D 8*8*8	12	6
Hipercubo 9	9	9

384 N	HiperCiclo 6	14	3	T3D 8*8*6	11	6
896 N	HiperCiclo 7	16	3	T3D 10*10*9	14	6

972 N	Grafo Kautz	6	3

5832 núcleos

Topología	Nº de nodos	Grado	Diámetro
Array lineal	N	2	N-1
Anillo	N	2	$\lfloor N/2 \rfloor$
Anillo de grado 'n'	N	$n=\log_2 N$	n-1
Árbol binario	2^K-1	3	2*(K-1)
Árbol binario equilibrado	2^K-1	2^K	2*(K-1)
Estrella	N	N-1	2
Malla	n^K	2*K	K*(n-1)
Toroide	n^K	2*K	K*$\lfloor n/2 \rfloor$
Hipercubo	2^K	K	K
Hipercubo con ciclos	$K*2^K$	3	2*K - 1 +$\lfloor K/2 \rfloor$

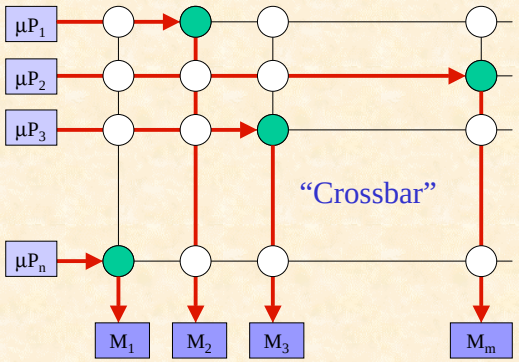
MIMD

HWANG (1993) IDENTIFICA TRES GENERACIONES:

- 1983-1987 Hipercubo con *Encaminamiento* Sw
- 1988-1992 Malla con *Encaminamiento* Hw (Sw de grano medio)
- 1993-1997 µP y *comunicaciones* en el mismo chip (grano fino)
- ¿2007? Multiprocessor systems-on-chips (MPSoCs) Niagara
  Hoy 4 núcleos .. 64 en 2010 .. ¿Se llegará a 1.000?

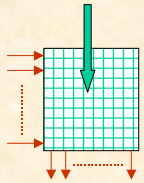
¿Conexión interna?

Simil con intracluster



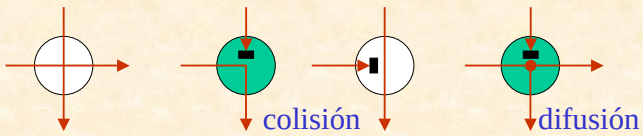
Perfil N*M

O (N^2)



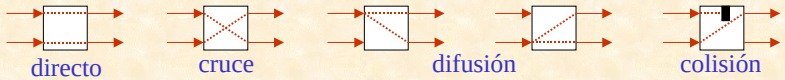
Muchas patas

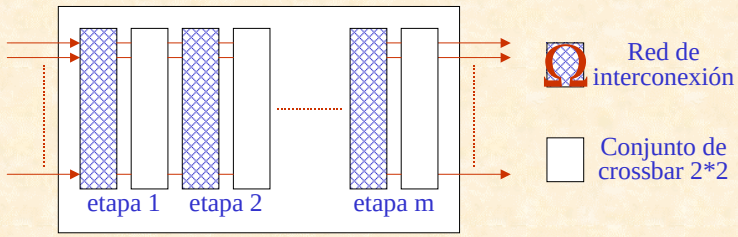
Funcionalidad de los conmutadores simples:



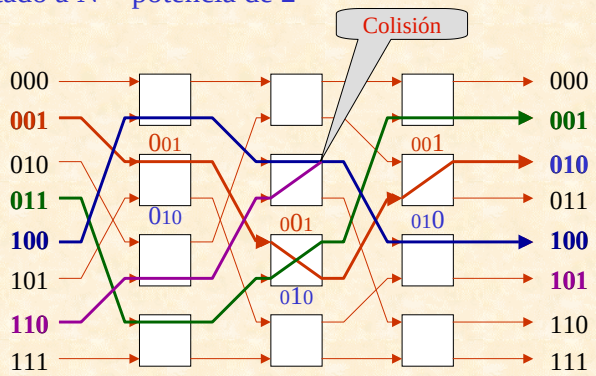
- O (64)
- Perfil 8*8 ¿ Reducir O(N^2) a costa de … ?
- Latencia 1

Usar sólo crossbar 2*2





- Red de interconexión "perfect Suffle"
- Limitado a N = potencia de 2



¿Encaminamiento?

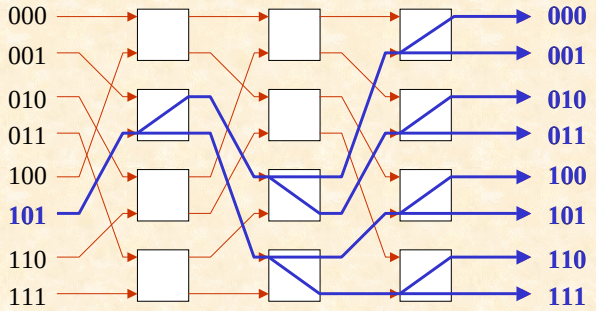
Sea de 001 a 010

Bit igual => directo

¿Mejorable?

Bit distinto => cruce

¿ Latencia y O( ) ?



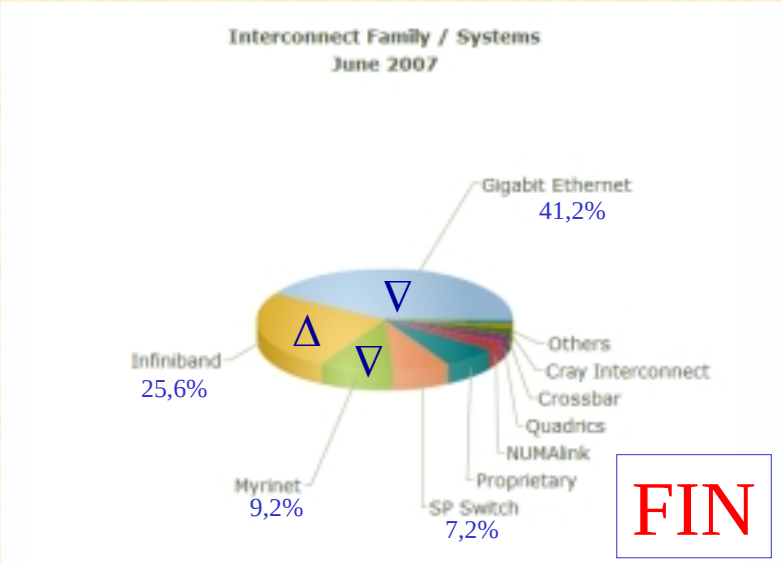
¡ Permite difusión !

	Bus	Multietapa	Crossbar
Latencia	Cte.	Log_2N	Cte.
Complejidad Conmuta.	N	$2\ N\ log_2N$	N^2
Perfil de Comunicación	1➔1	N➔N (*)	N➔N

- BUS — Barato y limitado 2..32
- CROSSBAR — Más caro. Bueno para N moderado. Mayor ancho de banda y fácil encaminar
- MULTIETAPA — Compromiso entre Bus y Crossbar

#NODOS	TIPO DE RED	SUPERCOMPUTADOR
..32	Crossbar	Bull NovaScale
..248	Red Clos	C-DAC PARAM Padma
..64	Crossbar	Cray Inc. X1
..30508	Toro 3D	Cray Inc. XT3 y XT4
..8192	Toro 3D	Cray Inc. XMT
..128	Crossbar	Fujitsu/Siemens M9000 Series
..32	Crossbar	Fujitsu/Siemens PRIMEQUEST 500
..64	Crossbar	Hitachi BladeSymphony
..256	Crossbar multidim.	Hitachi SR 11000
..64	Crossbar	HP Integrity SuperDome
..96*2	Red Ω	IBM eServer p575
..221184*4	Toro 3D y árbol	IBM BlueGene/L&P
..960	Crossbar	Liquid Computing LiquidIQ system
..32	Crossbar	NEC Express5800/1000 series
..4096	Crossbar multidim.	NEC SX-8 series
..512	Fat tree	SGI Altix 4000 series
..5832	Kautz graph	SiCortex SC series
..128	Crossbar	Sun M9000

intercluster


Interconnect Family / Systems
June 2007
Gigabit Ethernet
41,2%
Infiniband
25,6%
Myrinet
9,2%
SP Switch
7,2%
Proprietary
NUMAlink
Quadrics
Crossbar
Cray Interconnect
Others

FIN