MULTIPROCESADORES

Rubén Gran Tejero y Jesús Alastruey Benedé

Curso 2019-2020

3º curso del Grado en Ingeniería en Informática

Especialidad Ingeniería de Computadores

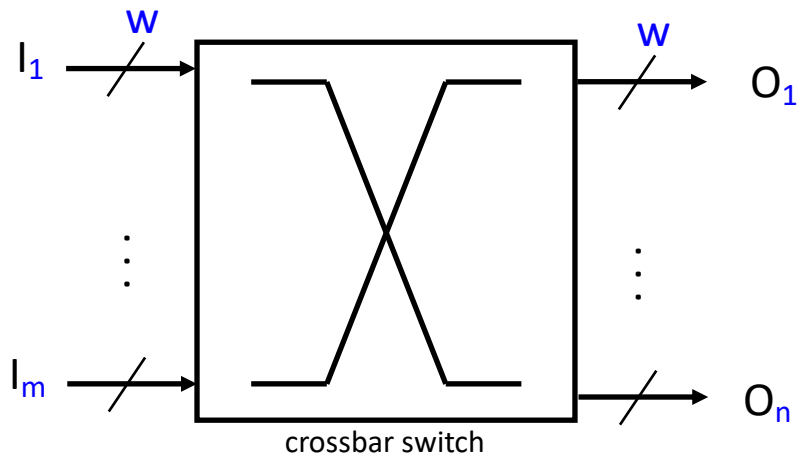
- Basada en la asignatura “Fundamentos de Arquitecturas Paralelas”, de Ingeniería informática, impartida desde el Curso 1995-96 hasta el 2012-13.
- Algunas transparencias tomadas de cursos del profesor José María LLabería, Facultad de Informática de Barcelona, UPC.

Redes de Interconexión Indirectas para Multiprocesadores

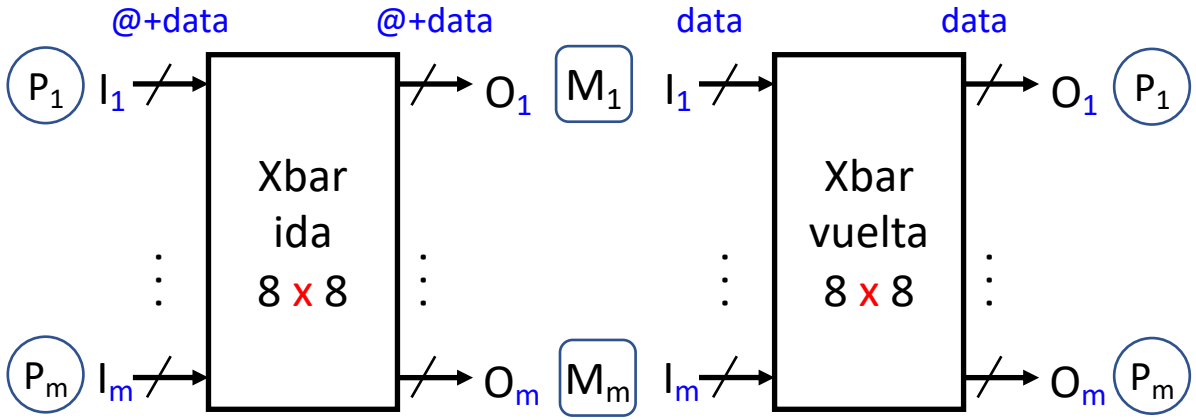
1. Crossbar
2. Multiples Buses
3. Redes multietapa Omega
4. Redes multietapa Clos

CROSSBAR SWITCH (CONMUTADOR)

- Red *indirecta* (indirecta = nodos-red-nodos)
  - Topología dinámica
  - m puertos de entrada y n puertos de salida, de anchura w bits
  - Permite conectar cualquier entrada con cualquier salida
  - Por tanto, no tiene degradación (*contention-free interconnect*)
- Coste medido en "puntos de cruce" proporcional a : m x n x w

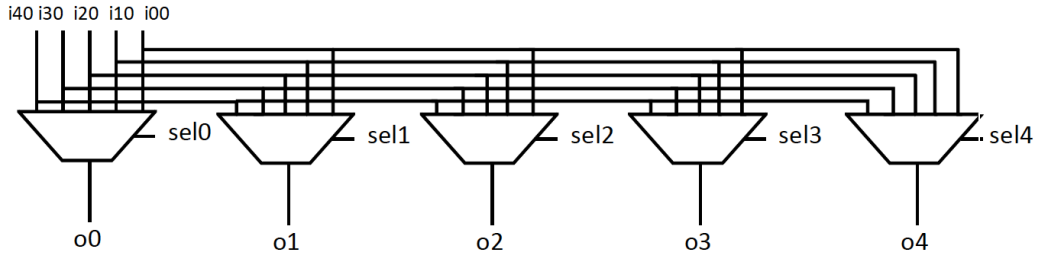


- Arbitraje: podemos pensar en un árbitro en cada salida, que selecciona una de las entradas que piden *esa* salida n x árbitro 2 de m
- Ejemplo de uso en un multi de memoria compartida sin caches



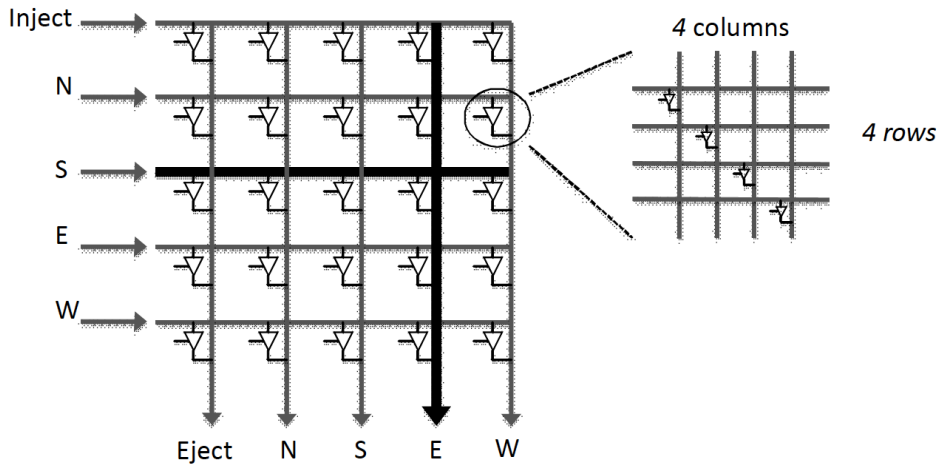
CROSSBAR SWITCH

- Implementación con multiplexores
  - Sun-Oracle T1-T5 (entre cache privadas y cache compartida)



Ejemplo xbar $5 \times 5 \times 1$ bit

- Implementación con CMOS pass gate, or tri-state CMOs buffer
  - Encaminadores de alta frecuencia



Ejemplo xbar $5 \times 5 \times 4$ bit (tb. para un router de una malla 2D, p.e.)

  - Costes en una implementación en silicio,
    - Área necesaria proporcional a $mxnxw^2$ superficie
    - Potencia dinámica proporcional a $(m+n)xw$ longitud cables
    - Potencia estática proporcional a $mxnxw$ nº puertas

CROSSBAR SWITCH

- Prestaciones relativas
  mediante una simulación estadística[1],
  sin tráfico real
  hipótesis de simulación sencillas:
  - Simulamos peticiones de procesadores a memorias
  - Procesadores -> Crossbar->Memorias (no melamos la vuelta)
  - Las peticiones de cada procesador a los módulos de memoria se modelan como “instancias independientes de variables aleatorias de distribución uniforme”
  - En cada ciclo de red se genera un vector de peticiones, se arbitra de forma aleatoria, se aceptan la que pasan y los módulos de memoria quedan libres
  - Los procesadores no guardan memoria del rechazo
  - En cuanto se sirve una petición se lanza otra (sig. Ciclo)

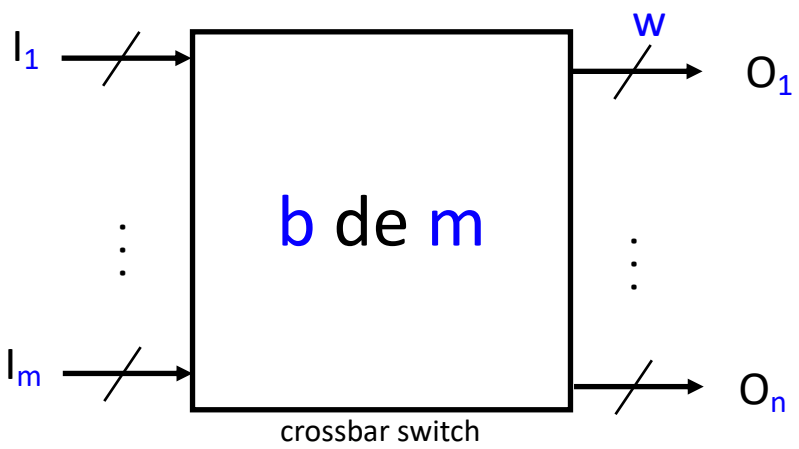
m=n	BW relativo pet.servidas/ciclo de red	%
2	1,5	75
4	2,62	65
8	4,95	62
16	9,59	59

- Nota histórica:
  Carnegie Mellon Univ., 1972
  C.mmp m=n=16, sin caches, PDP-11 (Digital Equipment Corporation)
  Ciclo de res = 1μs (1Mhz)

1) T.Lang, M.Valero, I. Alegre, “Bandwidth of Crossbar and Multiple-Bus Connections for Multiprocessors," IEEE Transactions on Computers, vol. 31, no. 12, pp. 1227-1234, Dec. 1982, doi:10.1109/TC.1982.1675947

MULTIPLES BUSES

- Red conmutada *indirecta*, topología dinámica:
  - m puertos de entrada y n puertos de salida, de anchura w bits
  - Permite conectar b entradas con b salidas, b < n
  - Por tanto, tiene degradación
- Coste medido en "puntos de cruce", proporcional a: (m+n)xbxw



- Arbitraje: podemos pensar en dos niveles:
  - 1º n x árbitros 1:m situados en cada memoria (hasta n peticiones)
  - 2º: 1 árbitro b:n seleccionando la memorias que pillan bus
- Ejemplo de uso en un multi de memoria mxn compartida sin caches

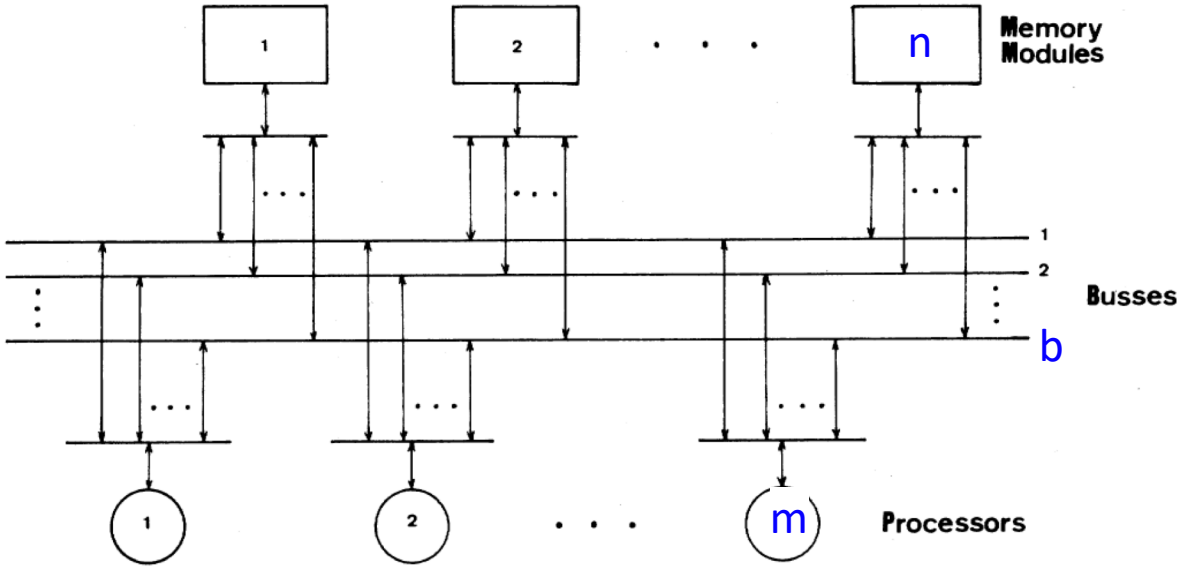


Fig. 2. Multiple-bus interconnection scheme.

MÚLTIPLES BUSES

- Prestaciones relativas
  mismas hipótesis que para el crossbar[1]
  m = n = 4, 8, 12, 16
  b = 1 … 16
  - b ≥ n/2 -> BW aumenta poco
  - b = n/2 + 1 -> 90% del BW del crossbar

1) T.Lang, M.Valero, I. Alegre, “Bandwidth of Crossbar and Multiple-Bus Connections for Multiprocessors," IEEE Transactions on Computers, vol. 31, no. 12, pp. 1227-1234, Dec. 1982, doi:10.1109/TC.1982.1675947

MÚLTIPLES BUSES: ON-CHIP COMMUNICATION ARCHITECTURES FOR EMBEDDED SYSTEMS BASED ON BUSES

Master (or Initiator)

- IP component that initiates a read or write data transfer

Slave (or Target)

- IP component that does not initiate transfers and only responds to incoming transfer requests

Arbiter

- Controls access to the shared bus
- Uses arbitration scheme to select master to grant access to bus

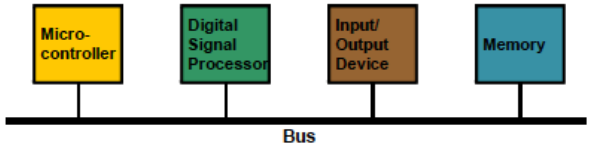
Decoder

- Determines the target for any transfer initiated by a master

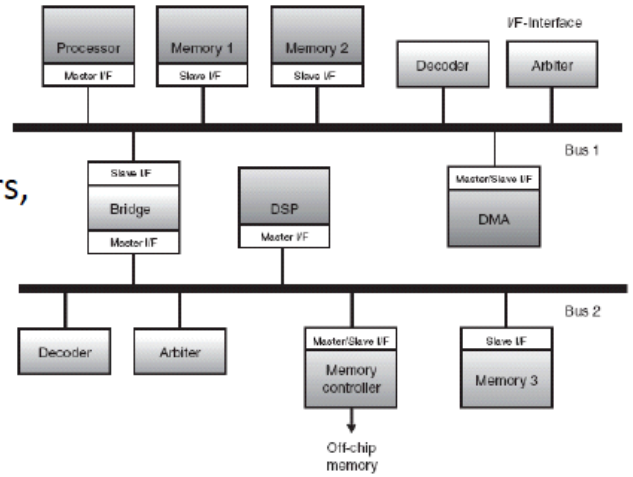
Bridge

- Connects two busses
- Acts as ***slave*** on one side and ***master*** on the other

2 processors, 1 bus

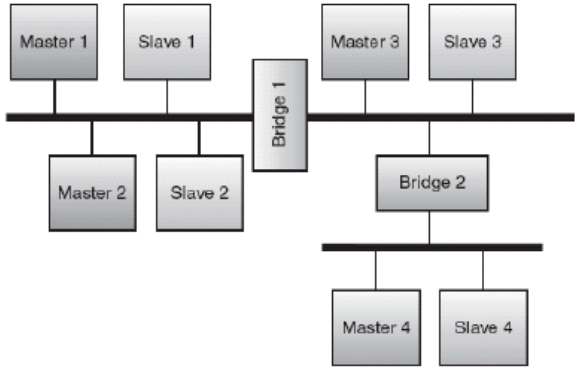


2 processors, 2 buses



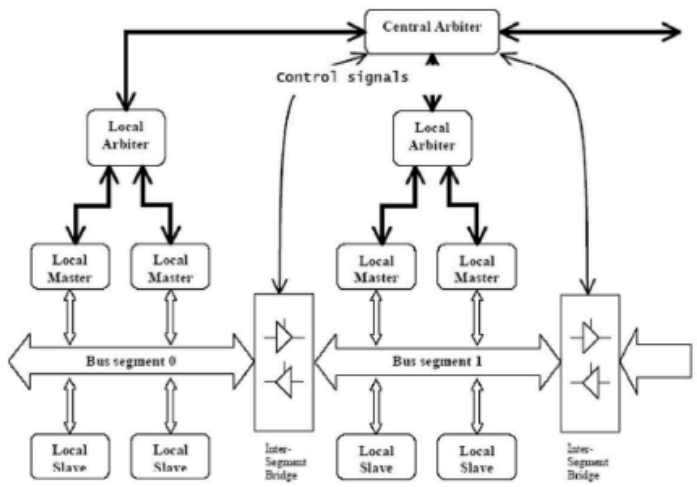
Hierarchical shared bus:

- improves system throughput
- multiple ongoing transfers on different buses



(phisically) Split bus:

- reduces impact of capacitance across two segments
- reduces contention and energy
- increases latency



MÚLTIPLES BUSES

- Tandem Computers, Inc. 1977 fault-tolerant multiprocessors. Tandem-16 NonStop.
  b= 2(x bus, y bus) ref[a]

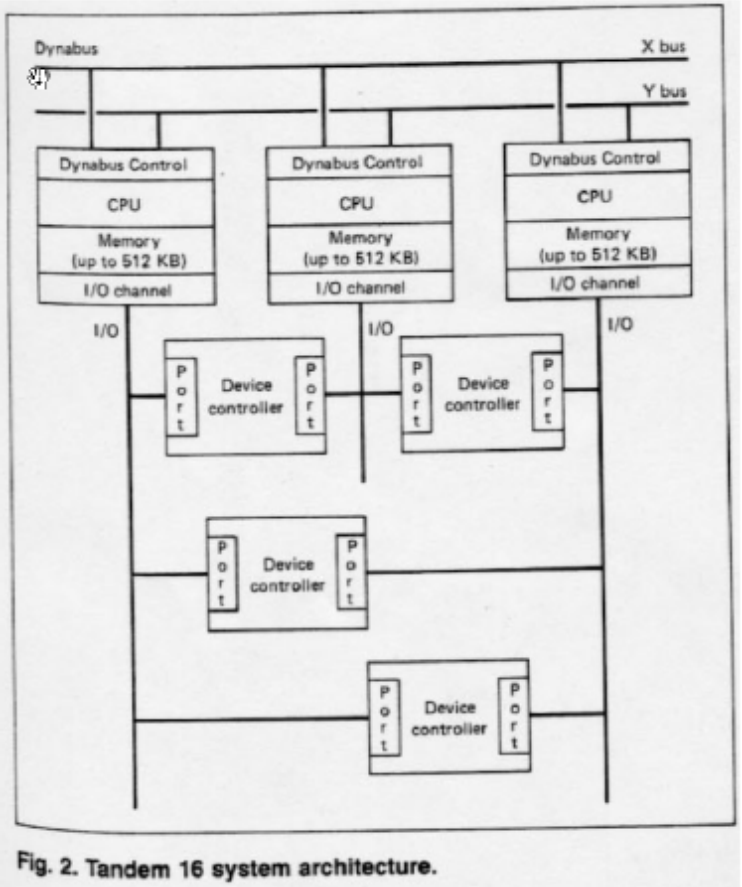


Fig. 2. Tandem 16 system architecture.

- Carnegie Mellon, 1977, CM*, Estructura jerárquica de múltiples buses, Posiblemente el primer multi de memoria compartida NUMA, En 1979, prototipo con 5 clusters, 10 módulos/cluster ref[b]

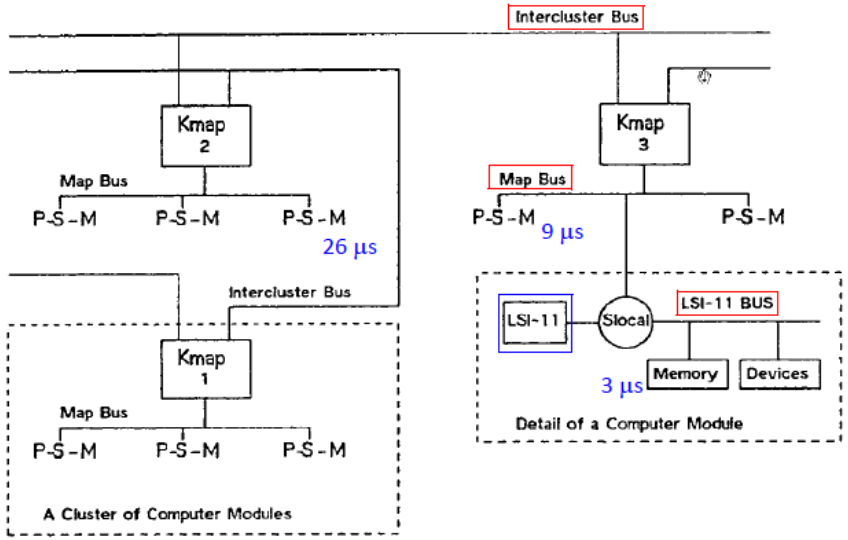


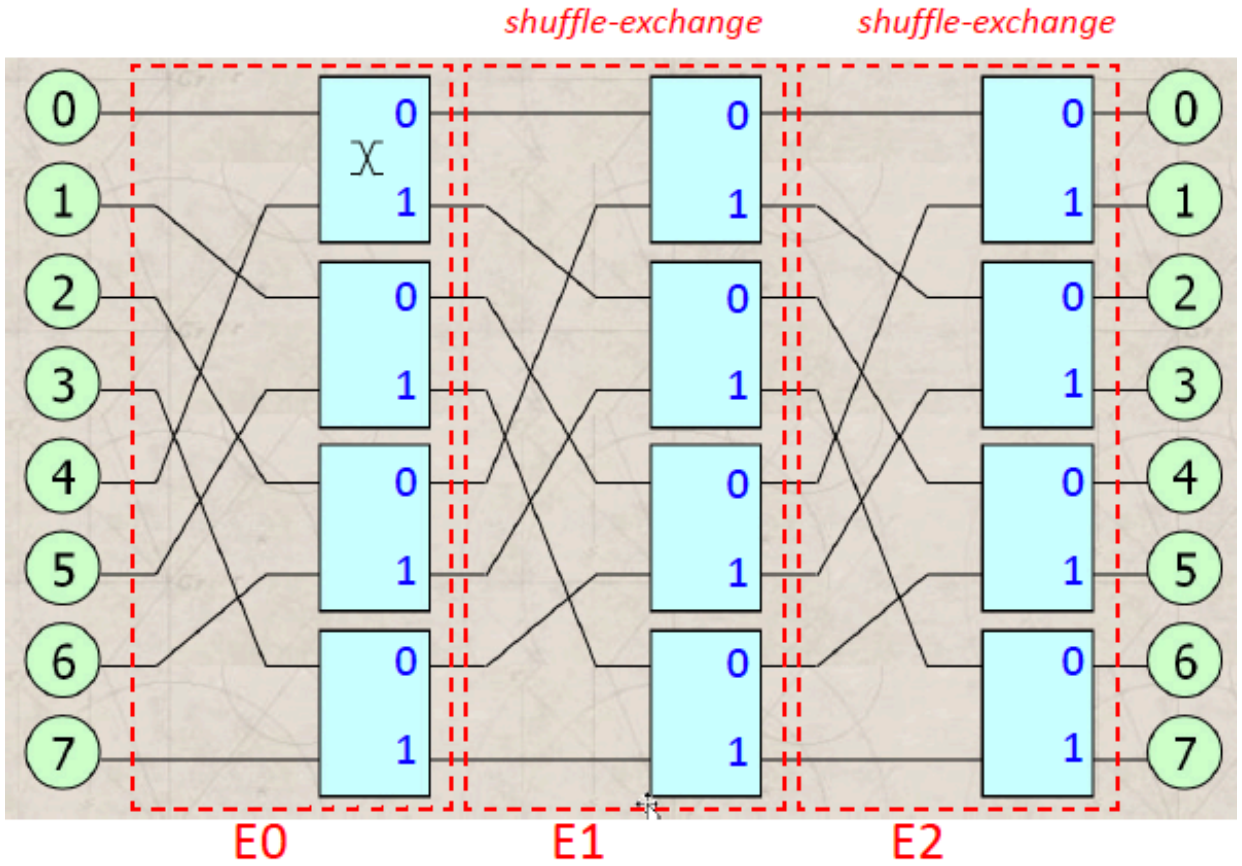
Figure 2.2 A Simple 3 Cluster Cm* System

a) Computer Structures: Principles and Examples. Daniel P. Sewiorek, C. Gordon Bell, and Allen Newell. Chapter 29. The Tandem 16: A Fault-Tolerant Computing System, pp. 470-480.
b) Fuller, Samuel H., "A collection of papers on Cm*: a multi-microprocessor computer system" (1977). Computer Sicence Department. Paper 2292.

RED MULTIETAPA OMEGA[1]

- Red conmutada *indirecta* (indirecta=nodos-red-nodos)
  - Familia: *multistage interconnection network*(MIN)
    - *Stage=shuffle-Exchange* baraja-intercambio
  - n puertos de entrada y n puertos de salida, de anchura w bits
  - Conecta cualquier entrada a cualquier salida, *pero* presenta degradación
- Puede verse como un crossbar cortado en varias etapas
  - Cada etapa tiene conexiones y crossbars pequeñitos k x k
  - El conexionado Ω siguen el patrón: *perfect shuffle*
- Costes:
  - Nº de etapas: $\log_k n$, cada una con n/k switches
  - Nº de conmutadores : $n/k \times \log_k n$

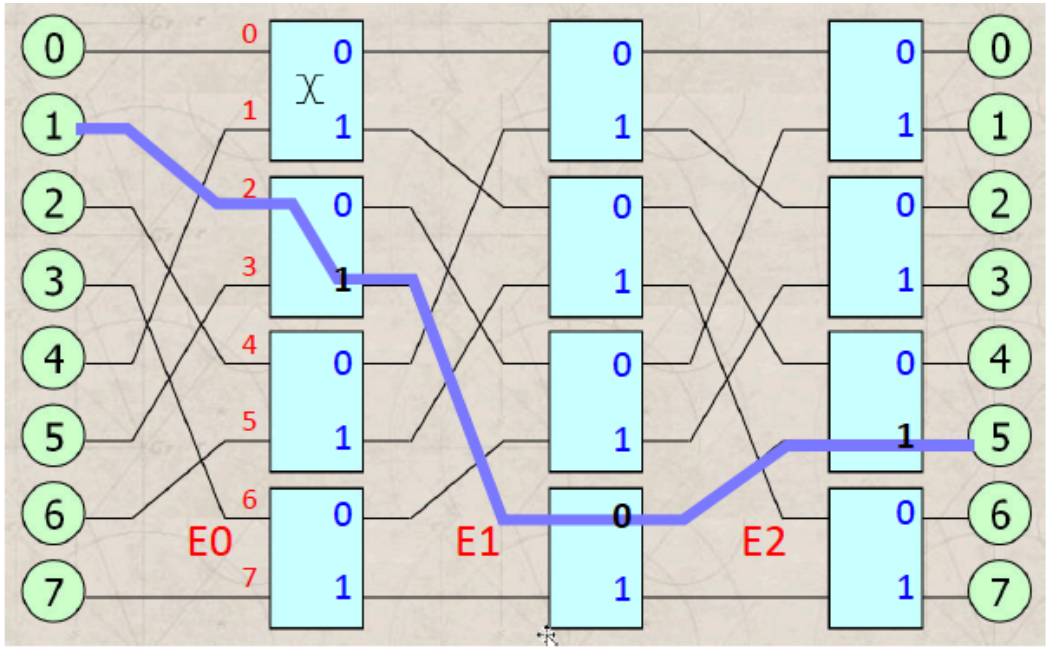
- Red Ω 8 x 8, con k = 2



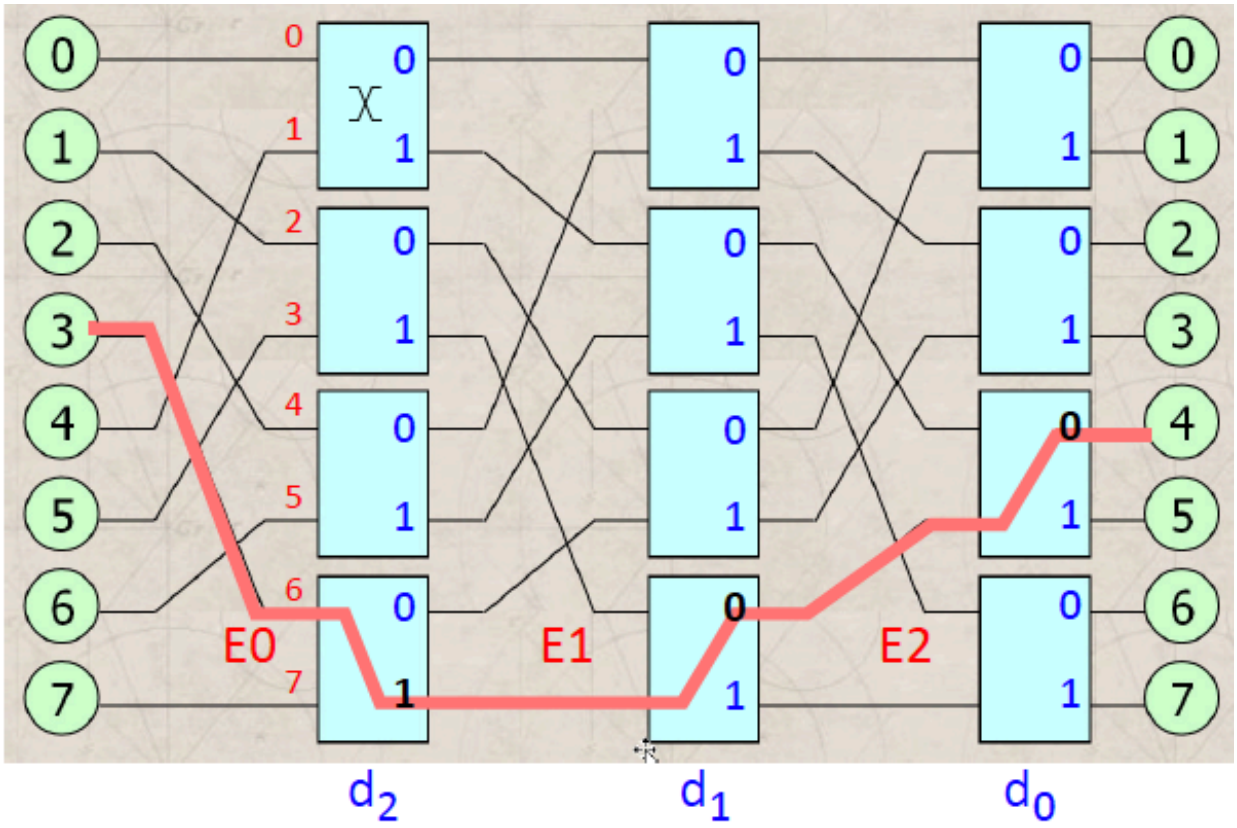
1) Duncan H. Lawrie (1975): "Access and Alignment of Data in an Array Processor", IEEE Transactions on Computers, Volume C-24, Number 12, pp. 1145– 1155, Dec. 1975.

RED MULTIETAPA OMEGA

- Encaminamiento distribuido en cada etapa: bits de nodo destino ($d_2d_1d_0$) deciden puerto de salida en (E0 E1 E2)
- Red Ω 8x8, con k = 2. src=1, dst=5=101b



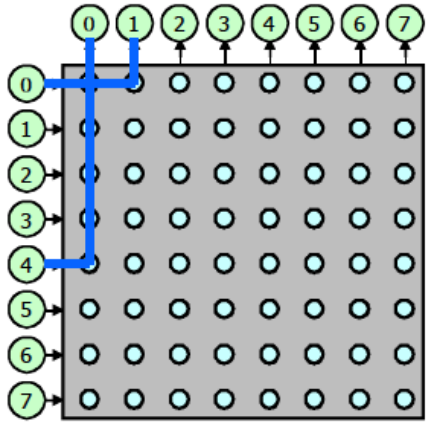
- Red Ω 8x8, con k = 2. src=3, dst=4=100b



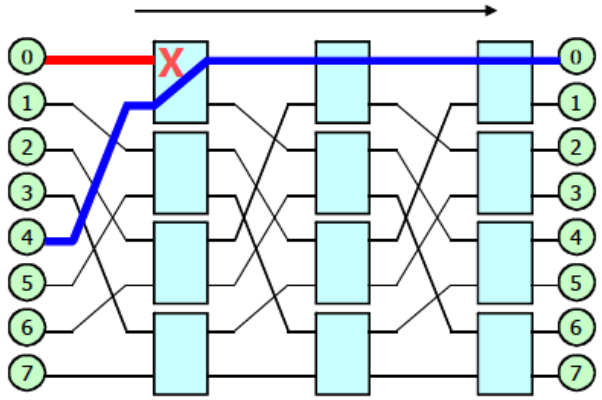
- src=1, dst=5 conflicto con src=3, dst=4 en puerto **0** del switch inferior de la etapa E1->*degradación*

RED MULTIETAPA OMEGA

- *Degradación*: los caminos desde diferentes orígenes a diferentes destinos comparten uno o mas enlaces. También se dice que es una red *bloqueante*.

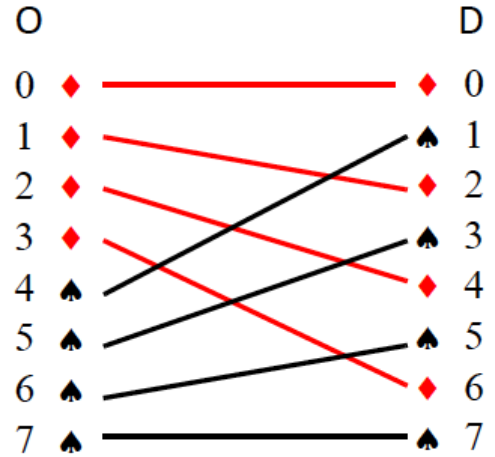


non-blocking topology



blocking topology

- Patrón de conexionado entre etapas: permutación
  - La permutación de la red Ω es la baraja perfecta (*perfect shuffle*)



PS(2)=4
PS(O) = ?

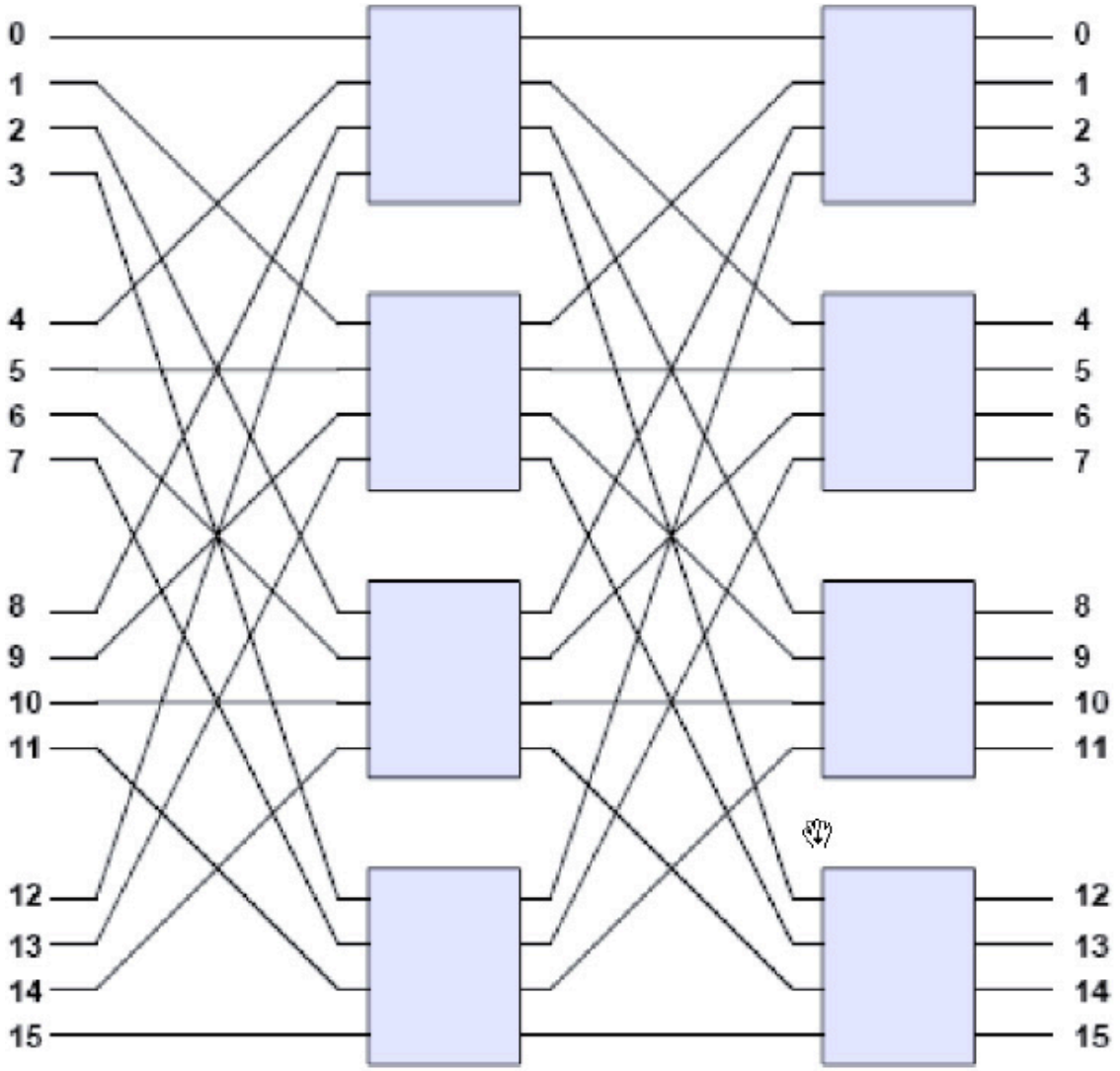
PS(o2,o1,o0)=

  - Hay muchas permutaciones de interés, por ejemplo la inversa de la baraja perfecta

RED MULTIETAPA OMEGA

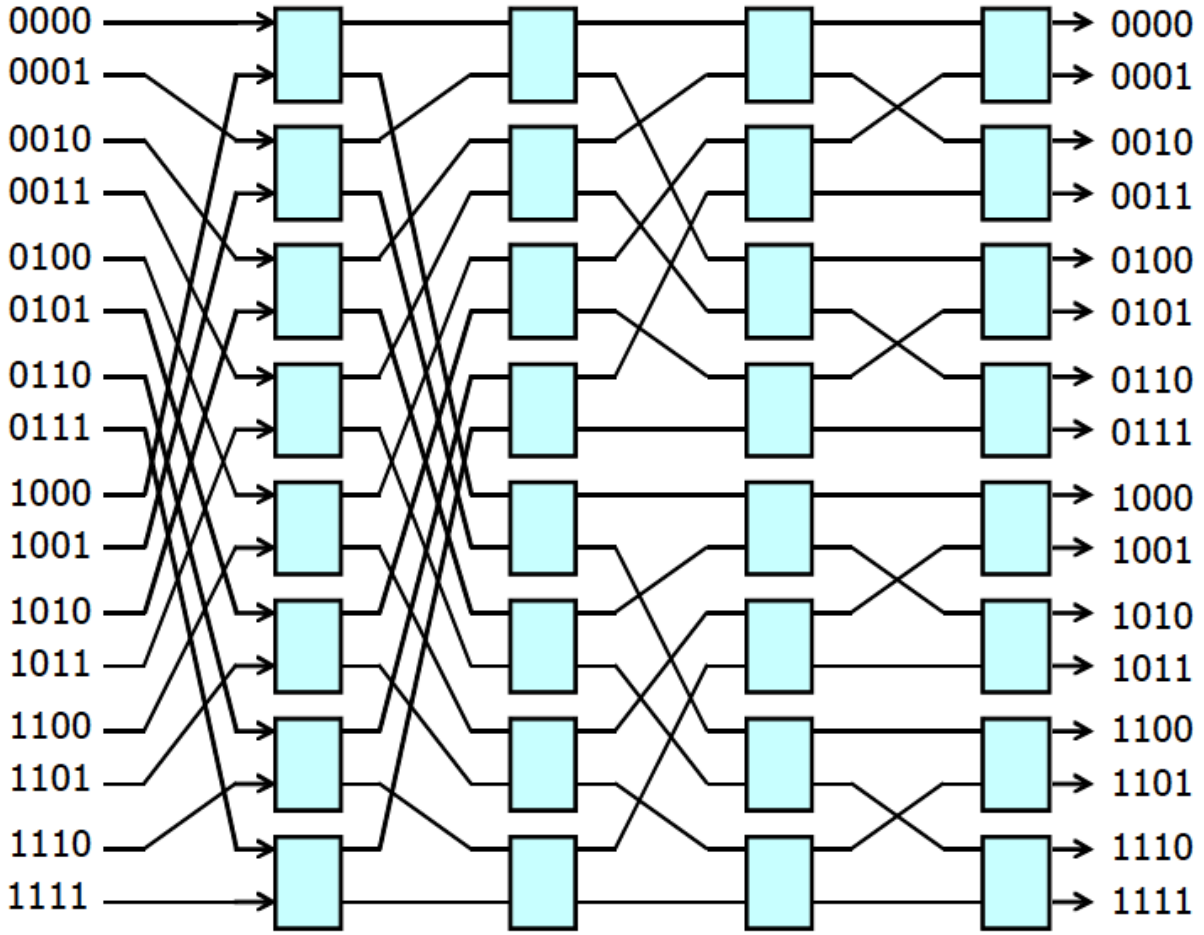
- Ejercicio 10.3.1: construir una red multietapa 8x8 con k=2, usando la baraja inversa en todas las etapas como patrón de conexión. Determinar el encaminamiento.

- Ejercicio 10.3.2: determinar la función de permutación del conexionado y el mecanismo de encaminamiento en la red Ω 16x16, k=4 que se muestra a continuación:



Red Multietapa Butterfly

- Otro miembro de la familia MIN:
  - Butterfly network: el patrón de conexionado, de izquierda a derecha, es intercambiar bit 0 y n-1 (E0); bit 0 y n-2 (E1); …etc.
  - Encaminamiento distribuido en cada etapa;
    - Bits de nodo destino (d2d1d0)
    - Deciden puerto de salida en (E2E1E0)
- Otro miembro de la familia MIN:
  - Baseline network[1]
  - Ejemplo 16 puertos, 4 etapas
- Ejercicio 10.3.3: Permutaciones de conexionado?, encaminamiento?



1) PowerPoint de: *Interconnection Networks. Computer Architecture: A Quantitative Approach* 4th Ed., App. E. T. Pinkston and José Duato + J. Flich, UPV.http://ceng.usc.edu/smart/**slides**/appendixE.html

HISTORIA. MULTIS DE MEMORIA COMPARTIDA, ESCALABLES, BASADOS EN MIN

Compañía/Univ.	Nombre	NºCPUs	Año
New York Univ.	NYU Ultracomputer	4096	1980-3
IBM	RP3	512	1985
Illinois University	Cedar	32	1986
BBN laboratories	Butterfly switches 4x4	256	1986

COMPARACIÓN COSTE-PRESTACIONES CROSSBAR VS. RED DELTA[1]

- Δ similar a la red Ω
  - Probabilidad de aceptación de una petición, suponiendo distribución uniforme en cada ciclo de red.
  - Bw_{red}= nº CPUs x Prob x w bits/ciclo

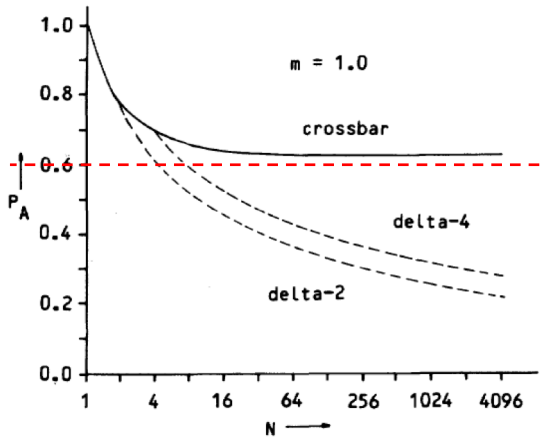


Fig. 11. Probability of acceptance of $N \times N$ networks.

- Relación calidad-precio: Probabilidad anterior/ Coste de la redes crossbar y delta-2 en función de una estimación del número de puertas. Se normaliza a la relación calidad-precio del crossbar 1x1

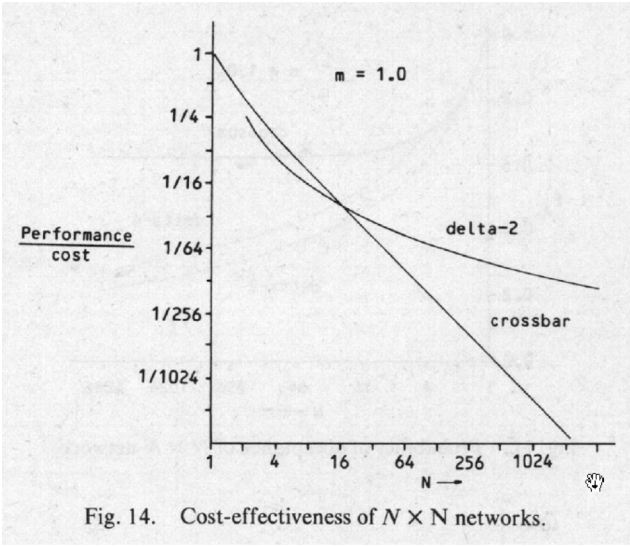


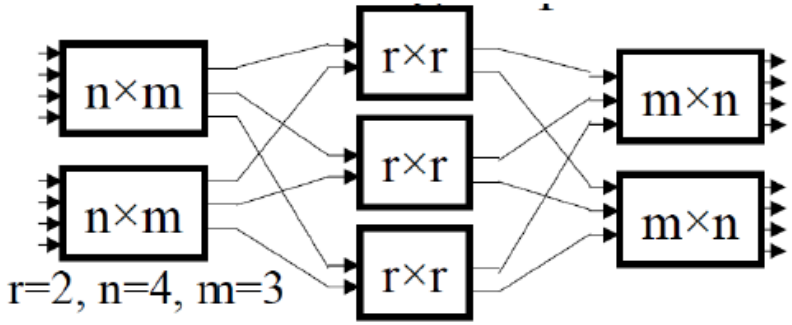
Fig. 14. Cost-effectiveness of $N \times N$ networks.

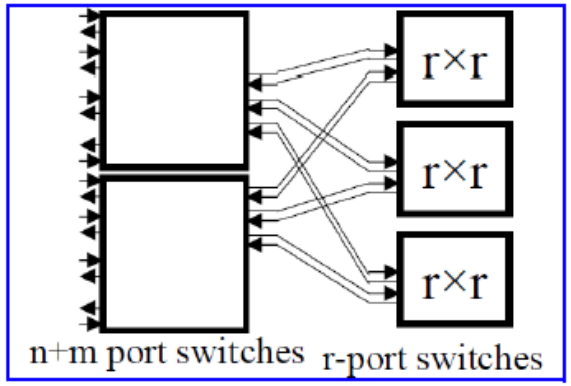
1) J. Patel. "Performance of Processor-Memory Interconnections for Multiprocessors," IEEE Transactions on Computers, vol.C-30, no.10, pp.771,780, Oct. 1981.

MULTISTAGE CLOS NETWORKS[1]

- Multi-stage circuit switching network proposed by Charles Clos in 1953 for telephone swithching systems
  - Can have any odd number of stages, e.g., 5
- Allows forming a large switch form smaller switches
- The number of cross-points is reduces –> Lower cost
- 3-Stage Clos (n,m,r)
  - Ingress (r nxm), middle (m rxr), egress (r mxn)
- Non-blocking:
  - Strict-sense if $m \geq 2n-1$
  - Rearrageably if $m \geq n$

- Folded: Merge input and output into one switch = Fat-tree





1)Data Center Network Topologies. Prof. Raj Jain. CSE570S: Recent Advances in Networking - Data Center Virtualization, SDN, Big Data, Cloud Computing, Internet of Things (Fall 2013).http://www.cse.wustl.edu/%7Ejain/cse570-13/