Capítulo 4

Arquitectura de los multiprocesadores

4.1 Introducción a los multiprocesadores

Una opción para conseguir un sistema con más rendimiento es la utilización simultánea de varios procesadores. A estos sistemas se les denomina *sistemas multiprocesador*. Normalmente, cada procesador ejecuta diferentes programas simultáneamente, pero puede ejecutar las mismas instrucciones dependiendo del tipo de sistema. El principal motivo detrás del desarrollo de sistemas multiprocesador es incrementar la velocidad de operación. Parece obvio que debe producirse un aumento de la velocidad cuando más de un procesador opera de forma simultánea. El mejor incremento posible de velocidad sería aquél que fuese proporcional al numero de procesadores, lo que ocurriría en el caso en que todos los procesadores trabajasen de forma simultánea todo el tiempo sobre una aplicación y no hubiese computación adicional o transferencia de datos, o contención en los recursos en la implementación multiprocesador realizada. Tal incremento de velocidad no suele conseguirse ya que es muy improbable que consigamos que todos los procesadores realicen trabajo útil sobre un problema al mismo tiempo. Además, normalmente existe una sobrecarga sustancial en la comunicación entre procesadores. Sin embargo, es posible conseguir un considerable aumento en la velocidad (*speed-up*).

Podemos definir el *speedup* sobre p procesadores como

$$Speedup(p\,procesadores) \equiv \frac{Rendimiento(p\,procesadores)}{Rendimiento(1\,procesador)}$$

Para un único problema, el rendimiento de una máquina para ese problema es simplemente el recíproco del tiempo que tarda en completar el problema, así tenemos el siguiente caso:

$$Speedup(p\,procesadores) \equiv \frac{Tiempo(1\,procesador)}{Tiempo(p\,procesadores)}$$

En cuanto a los esquemas más utilizados en el diseño de computadores paralelos, los años 80 han visto aparecer dos tendencias: *multiprocesadores de memoria compartida* y *multicomputadores*.

La figura 4.1 muestra un esquema simplificado de estas arquitecturas. Los multicomputadores constan esencialmente de un conjunto de procesadores, cada uno conectado a su memoria local. Los procesadores se comunican mediante el paso de mensajes a través de una red de interconexión. La figura 4.1a

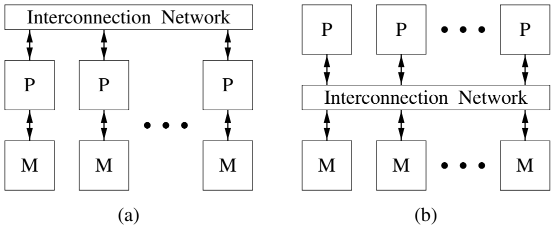


Figura 4.1: Representación esquemática de los computadores paralelos: (a) Un multicomputador. (b) Un multiprocesador de memoria compartida UMA. (M = memoria; P = procesador).

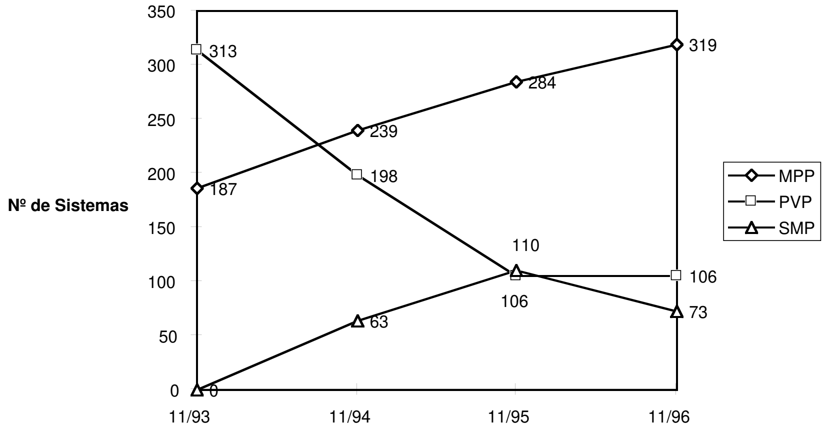


Figura 4.2: Tipos de sistemas usados en los 500 sistemas más rápidos existentes en el mundo (Obsérvese el descenso de los procesadores vectoriales paralelos).

muestra un esquema simple de esta arquitectura. En la figura 4.2 se muestra el número de sistemas existentes para diferentes tipos de arquitecturas paralelas.

La programación de multicomputadores no es una tarea sencilla. El programador debe de encargarse de la distribución de código y datos entre los procesadores de manera eficiente, realizar las llamadas que se encarguen de enviar los datos que necesiten otros procesadores. Por otro lado, los *multiprocesadores de memoria compartida* proporcionan un espacio de memoria único para todos los procesadores, simplificando la tarea de intercambiar datos entres los procesadores. El acceso a memoria compartida ha sido implementado, tradicionalmente, usando una red de interconexión entre los procesadores y la memoria. (Figura 4.1b). A esta arquitectura se le conoce tradicionalmente como arquitectura UMA (*Uniform Memory Access*).

Recientemente, los multiprocesadores de memoria compartida han seguido algunas de las tendencias establecidas previamente por los multicomputadores. En particular, la memoria ha sido distribuida físicamente entre los procesadores, reduciendo el tiempo de acceso a la memoria e incrementando la escalabilidad. A estos computadores paralelos se les denominan multiprocesadores de memoria compartida distribuida (DSM). Los accesos a la memoria remota se realizan a través de una red de interconexión, de una manera similar al caso de los multicomputadores. La diferencia principal entre los DSMs y los multicomputadores es que los mensajes se inician en respuesta a accesos a la memoria en vez de llamadas al sistema. Par reducir la latencia de la memoria, cada procesador tiene varios niveles de memoria caché. Esta arquitectura proporciona un tiempo no uniforme de acceso a la memoria (NUMA). De hecho, la mayor parte de la no uniformidad se debe a la diferencia entre los tiempos de acceso a las cachés y la memoria principal, más que el tiempo de acceso diferente entre las memorias locales y remotas. El principal problema que aparece en los DSMs es la coherencia de las cachés.

Aparte de las arquitecturas UMA y NUMA, existe una tercera posibilidad, la arquitectura COMA (Cache-Only Memory Architecture). En el modelo COMA, cada procesador tiene una parte de la memoria compartida, como en el modelo NUMA, pero cada parte es una memoria caché, es decir, toda la memoria compartida se comporta como una memoria caché. Por lo tanto, parte de la memoria caché puede tener datos que no han sido demandados por el procesador asociado. La arquitectura COMA permite la migración de datos hacia el procesador que realiza la petición, al igual que en el caso de cualquier multiprocesador con memoria caché. Dado que puede ser costoso que toda la memoria sea una caché, Hagwersen, Landin y Haridi (1992) describen un sistema COMA con una porción de memoria que actua como caché real, siendo el resto de la memoria lo que ellos denominan una memoria de atracción (los datos son atraidos a la memoria de atracción más cercana al procesador que los demandó).

El uso de redes de interconexión propias hace que los multicomputadores y los DSM sean bastante caros. Es por ello que se ha propuesto el uso de redes de estaciones de trabajo (NOWs) como una aproximación barata para construir computadores paralelos. Las NOWs intentan aprovecharse de los recientes avances que han tenido lugar en las redes de área local (LANs) con la aparición de redes de baja latencia como Fast Ethernet, ATM o Myrinet.

4.2 Factores de diseño de las redes de interconexión

Las redes de interconexión juegan un papel importante en el rendimiento de los computadores paralelos modernos. Existen muchos factores que pueden afectar en la elección de un una red de interconexión apropiada para una plataforma de computación paralela determinada. Estos factores incluyen:

1. *Requisitos de rendimiento*. Los procesos que se ejecutan en procesadores diferentes se sincronizan y comunican a través de la red de interconexión. Estas operaciones se realizan normalmente mediante el paso de mensajes explícitos o mediante el acceso a variables compartidas. La *latencia* del mensaje es el tiempo que transcurre desde que se genera el mensaje en el nodo fuente hasta que este llega a su nodo destino. La latencia afecta directamente al tiempo en que el procesador está desocupado y al tiempo de acceso a direcciones presentes en memoria remota. Además, la red puede *saturarse* (puede ser imposible enviar los mensajes generados por los nodos) limitando el poder de computación efectivo de un computador paralelo. La máxima cantidad de información entregada por la red por unidad de tiempo se define como el *rendimiento* (throughput) de esa red.

2. *Escalabilidad*. Una arquitectura escalable implica que al incrementarse el número de procesadores, el ancho de banda de la memoria, la E/S y la red deben incrementarse proporcionalmente. En caso contrario los componentes cuyo ancho de banda no escale pueden convertirse en un cuello de botella para el resto del sistema, decrementando la eficiencia global.

3. *Expansibilidad incremental*. Es poco común que los clientes compren un computador paralelo con todos sus procesadores y memorias. Cuando el presupuesto lo permita, se pueden añadir más procesadores y memoria hasta que el sistema alcance su configuración máxima. En algunas redes de interconexión, el número de procesadores debe ser potencia de 2, lo que hace que sean difícil de expandir. En otros casos, la expansibilidad se realiza a costa de malgastar recursos. Por ejemplo, un red diseñada para un máximo de 1024 nodos puede tener muchos enlaces sin usar cuando la red se implementa con un tamaño menor. Las redes de interconexión deben de proporcionar una expansibilidad incremental, permitiendo añadir un pequeño número de nodos y minimizando los recursos malgastados.

4. *Particionabilidad*. Los computadores paralelos son frecuentemente usados de forma simultánea por varios usuarios. En este caso es deseable que el tráfico en la red producido por cada usuario no afecte al rendimiento de las otras aplicaciones. Esto puede conseguirse si la red puede separarse en subsistemas funcionales más pequeños. Esta propiedad también puede ser necesaria por razones de seguridad.

5. *Simplicidad*. Los diseños simples a menudo llevan a mayores frecuencias de reloj lo que implica un mayor rendimiento. Además, los usuarios aprecian aquellas redes que son fáciles de entender ya que este hecho facilita una mejor explotación de las mismas.

6. *Distancia*. Este factor puede llevar a implementaciones muy diferentes. En los multicomputadores y DSMs la red puede ocupar a lo máximo algunos armarios. La máxima distancia entre los nodos es pequeña lo que permite el uso de cables de cobre para transmitir las señales. En las NOWs, los enlaces tienen longitudes muy diferentes y algunos de ellos pueden ser muy largos lo que puede dar lugar a acoplamientos y ruidos electromagnéticos. El uso de enlaces de fibra óptica resuelve estos problemas, igualando el ancho de banda de enlaces cortos y largos.

7. *Restricciones físicas*. Una red de interconexión conecta procesadores, memorias, y/o dispositivos de E/S. Sería deseable que la red permitiese la conexión de un gran número de componentes a la vez que mantiene latencias de comunicación bajas. Al aumentar el número de componentes, el número de cables necesarios para interconectarlos aumenta. El encapsulamiento de estos componentes requiere normalmente cumplir ciertas restricciones físicas, como el control de la

temperatura de funcionamiento, limitaciones en la longitud de los cables, y limitaciones de espacio. Existen dos problemas principales en la implementación de redes grandes: la inclusión de cables en una área limitada, y el número de pines por chip (o tarjeta) dedicados a puertos de comunicación. En otras palabras, la complejidad de la conexión está limitada por la densidad de cable máxima y el número máximo de pines. La velocidad a la que una máquina puede funcionar está limitada por la longitud de los cables, y la mayor parte de la potencia consumida por un sistema se usa en la comunicación a través de los cables.

8. *Fiabilidad y reparabilidad*. Una red de interconexión debe de ser capaz de enviar información de forma fiable. Es posible diseñar redes de interconexión que continúen operando en presencia de un número limitado de fallos. Estas redes son capaces de enviar los mensajes a través de caminos alternativos cuando se detecta un fallo. Además de la fiabilidad, las redes de interconexión deben tener un diseño modular, permitiendo actualizaciones y reparaciones sin tener que detener todo el sistema. Los nodos también pueden fallar o ser eliminados de la red. Por ejemplo, en el caso de las NOWs, las estaciones pueden apagarse o encenderse. Este problema requiere el uso de algoritmos de reconfiguración que permitan una reconfiguración automática de la red.

9. *Trabajos esperados*. Los usuarios de una máquina de propósito general suelen tener necesidades muy diferentes. Si se conoce por anticipado la clase de aplicaciones que van a ser ejecutadas en el computador paralelo, es posible extraer información acerca de los patrones de comunicaciones habituales, tamaño de los mensajes, carga de la red, etc. Esta información se puede usar para optimizar algunos de los parámetros de diseño. Cuando esto no es posible, los parámetros de diseño deben seleccionarse para conseguir un diseño robusto que tenga un buen comportamiento ante un amplio rango de condiciones de tráfico.

10. *Restricciones de coste*. Finalmente, es evidente que la "mejor" red puede ser demasiado cara. La decisiones de diseño son a menudo decisiones entre coste y otros factores de diseño. Afortunadamente, el coste no es siempre proporcional al rendimiento. El uso de componentes disponibles en el mercado siempre que sea posible puede reducir considerablemente el coste total.

Para conseguir estos objetivos los diseñadores de redes deben de manipular tres variables independientes: *topología*, *encaminamiento* y *control de flujo*.

- La **topología** hace referencia al grafo de interconexión de la red $I = G(N, C)$ donde N son los nodos del grafo y C es el conjunto de enlaces unidireccionales o bidireccionales que los conectan. Si pensamos en un multiprocesador como en un problema de asignación de recursos, la topología es la primera forma de asignación de los mismos.

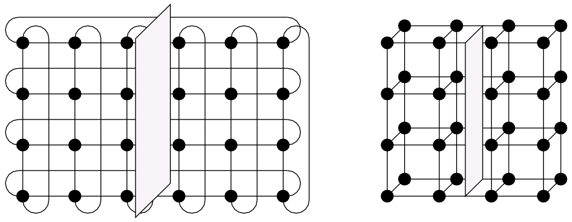
- El **encaminamiento** hace referencia al método que se usa para determinar el camino que sigue un mensaje desde el nodo origen al nodo destino. El encaminamiento se puede ver como una relación, $C \times N \times C$, que asigna al canal ocupado por la cabecera del mensaje y en función del nodo destino un conjunto de canales que pueden utilizarse a continuación para que el mensaje llegue a su destino. El encaminamiento es una forma dinámica de asignación de recursos. Dada un topología, un nodo actual, y un destino, la relación de encaminamiento determina como llevar un mensaje desde el nodo actual al nodo destino.

- El **control de flujo** hace referencia al método utilizado para regular el tráfico en la red. Es el encargado de evitar que los mensajes se entremezclen y controlar su avance para asegurar una progresión ordenada de los mismos a través de la red. Si dos mensajes quieren usar el mismo canal al mismo tiempo, el control de flujo determina (1) qué mensaje obtiene el canal y (2) qué pasa con el otro mensaje.

4.2.1 Topología

Una topología se evalúa en términos de los siguientes cinco parámetros:

- **Ancho de la bisección**. El ancho de la bisección, B, es el mínimo número de canales que, al cortar, separa la red en dos partes iguales. La bisección del cableado, B_W, es el número de cables que cruzan esta división de la red. $B_W = BW$, donde W es el ancho de un canal en bits. Este parámetro nos da una cota inferior de la densidad del cableado. Desde el punto de vista del diseñador, B_W es un factor fijo, y el ancho de la bisección restrige la anchura de los canales a $W = \frac{B_W}{B}$. La figura 4.3 muestra un ejemplo del cálculo del ancho de la bisección.



(a) (b)

Figura 4.3: Ejemplos del cálculo del ancho de la bisección: (a) toro 2-D. (b) toro 3-D (no se muestran los enlaces de cierre).

- **Grado**. El grado de un nodo, d, es el número de canales incidentes a un nodo. El número de canales que llegan a un nodo es el *grado de entrada*, d_{in}. El número de canales que salen de un nodo es el *grado de salida*, d_{out}. El grado total es la suma, $d = d_{in} + d_{out}$. El grado afecta al coste de un nodo. El número de pins de cada nodo es Wd, y la complejidad de la lógica de control del router suele estar relacionada con d.

- **Diámetro**. El diámetro de una red es el más largo de los caminos más cortos entre dos nodos.

$$D = max_{i,j \in N}(min_{p \in P_{ij}} length(p)),$$

donde P_{ij} es el conjunto de caminos de i a j. El diámetro se utilizó hasta finales de los 80 como la principal figura de mérito de una red. De ahí la popularidad que tuvieron en esa época redes de bajo diámetro como los hipercubos.

- **Longitud de los cables**. La longitud de los cables en la red determina la velocidad a la que puede operar la red y la potencia disipada por los mismos. En la figura 4.4 se muestra un ejemplo de empaquetamiento.

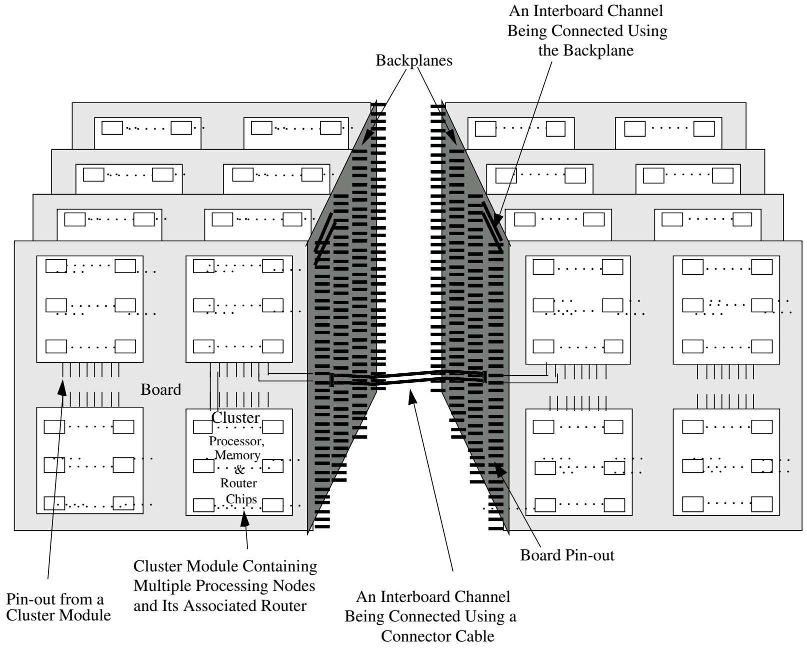


Figura 4.4: Ejemplo de empaquetamiento de un multicomputador.

- **Simetría**. Una red es simétrica si es isomorfa a ella misma independientemente del nodo que consideremos como origen. En una red simétrica la red *se vé* de la misma forma desde cualquier nodo. Este tipo de redes tienen la ventaja de que simplifican muchos de los problemas de manejo de recursos. Por ejemplo, dado un patrón de tráfico uniforme, en las redes simétricas se produce una carga uniforme de los canales de la red, cosa que no ocurre en las redes asimétricas.

4.2.2 Encaminamiento

El encaminamiento es el método usado por un mensaje para elegir un camino entre los canales de la red. El encaminamiento puede ser visto como un par (R, ρ), donde

$$R \subset C \times N \times C,$$

$$\rho : P(C) \times \alpha \longmapsto C.$$

La relación de encaminamiento R identifica los caminos permitidos que pueden ser usados por un mensaje para alcanzar su destino. Dada la posición actual del mensaje, C, y su nodo destino, N, R identifica un conjunto de canales permitidos, C, que pueden ser usados como siguiente paso en el camino.

La función ρ selecciona uno de los caminos de entre los permitidos. En cada paso del encaminamiento, ρ toma el conjunto de posibles canales siguientes, $P(C)$, alguna información adicional acerca del estado de la red, α, y elige un canal en concreto, C. La información adicional, α, puede ser constante, aleatoria, o basada en información sobre el tráfico de la red.

Los métodos de encaminamiento pueden clasificarse en *deterministas*, *inconscientes* (*oblivious*), o adaptativos. Con el encaminamiento determinista, el camino que sigue un mensaje depende únicamente de los nodos origen y destino. En este caso R es una función y α es constante (no se proporciona información adicional).

En un encaminamiento inconsciente se pueden elegir varios caminos a través de la red, pero no se utiliza información acerca del estado de la red para elegir un camino. El mensaje no es consciente del resto de tráfico en la red. Con un encaminamiento de este tipo, R es una relación (pueden existir varios caminos permisibles). Para que el encaminamiento sea inconsciente, α no puede contener información acerca del estado de la red. Puede ser aleatoria, una función de tiempo, o una función del contenido del mensaje.

El caso más general es el encaminamiento adaptativo, donde el router puede usar información acerca del estado de la red. Es este caso, α pude ser cualquier función.

4.2.3 Control de flujo

El control de flujo se refiere al método usado para regular el tráfico en una red. Determina cuando un mensaje o parte de un mensaje puede avanzar. También se hace cargo de la política de asignación de los recursos: buffers y canales; a las unidades de información: mensajes, paquetes y flits.

- **Mensaje**. Es la unidad lógica de comunicación. Dos objetos se comunican mediante el envio de un mensaje. Es la única unidad vista por los clientes de un servicio de red.

- **Paquete**. Un mensaje se divide en uno o más paquetes. Un paquete es la menor unidad que contiene información de encaminamiento. Si existe más de un paquete en un mensaje, cada paquete contiene además un número de secuencia que permite su reensamblaje.
- **Flit**. Un paquete se puede dividir a su vez en dígitos de control de flujo o *flits*, la menor unidad sobre la cual se puede realizar el control de flujo. Es decir, la comunicación de recursos, cables y buffers, se realizan en función de flits. En general, un flit no contiene información de encaminamiento. Únicamente el primer flit de un paquete sabe a donde se dirige. El resto de flits debe de seguir al flit cabecera para determinar su encaminamiento.

Una simple analogía ilustra la diferencia entre los paquetes y los flits. Los paquetes son como los automóviles. Dado que conocen a donde van, pueden entrelazarse libremente. Los flits, por otro lado, son como los vagones de un tren. Deben de seguir al flit cabecera para encontrar su destino. No se pueden mezclar con los flits de otros paquetes, o perderían su único contacto con su destino.

Otro aspecto importante que tiene que solucionar el control de flujo es el de bloqueo de un paquete. En este caso será necesario disponer de algún espacio tipo buffer para su almacenamiento temporal. Cuando ya no existe más espacio, el mecanismo de control de flujo detiene la transmisión de información. Cuando el paquete avanza y aparece más espacio de buffer disponible, la transmisión comienza de nuevo. Existe también la alternativa de eliminar el paquete cuando deja de existir espacio disponible o desviarse a través de otro canal.

4.3 Clasificación de las redes de interconexión según la topología

Entre otros criterios, las redes de interconexión se han clasificado en función del modo de funcionamiento (síncrono o asíncrono), y control de la red (centralizado, descentralizado, o distribuido). Hoy en día, los multicomputadores, multiprocesadores, y NOWs dominan el mercado de los computadores paralelos. Todas estas arquitecturas utilizan redes asíncronas con control distribuido. Por lo tanto, nos centraremos en otros criterios que son más significativos en la actualidad.

La figura 4.5 muestra una clasificación de las redes de interconexión conocidas en cuatro grupos en función principalmente de la topología de la red: redes de medio común, redes directas, redes indirectas, y redes híbridas. Para cada grupo, la figura muestra una jerarquía de clases, indicando alguna implementación real que utiliza dicha topología.

En las *redes de medio compartido*, el medio de transmisión está compartido por todos los dispositivos que tienen posibilidad de comunicación. Un enfoque alternativo consiste en tener enlaces punto a punto que conecta de forma directa cada elemento de comunicación con un subconjunto (normalmente reducido) de otros dispositivos existentes en la red. En este caso, la comunicación entre elementos de comunicación no vecinos requiere la transmisión de la información a través de varios dispositivos intermedios. A estas redes se les denominan *redes directas*. En vez de conectar de forma directa los elementos de comuncación, las *redes indirectas* los conectan mediante uno o más conmutadores. Si existen varios conmutadores, estos suelen estar conectados entre ellos mediante enlaces punto a punto. En este caso, cualquier comunicación entre distintos dispositivos requiere transmitir la información a través de uno o más conmutadores. Finalmente, es posible una aproximación *híbrida*.

Interconnection Networks

- Shared-Medium Networks
  - Local Area Networks
    - Contention Bus (Ethernet)
    - Token Bus (Arcnet)
    - Token Ring (FDDI Ring, IBM Token Ring)
  - Backplane Bus (Sun Gigaplane, DEC AlphaServer8X00, SGI PowerPath-2)
- Direct Networks (Router-Based Networks)
  - Strictly Orthogonal Topologies
    - Mesh
      - 2-D Mesh (Intel Paragon)
      - 3-D Mesh (MIT J-Machine)
    - Torus (k-ary n-cube)
      - 1-D Unidirectional Torus or Ring (KSR first-level ring)
      - 2-D Bidirectional Torus (Intel/CMU iWarp)
      - 3-D Bidirectional Torus (Cray T3D, Cray T3E)
    - Hypercube (Intel iPSC, nCUBE)
  - Other Topologies: Trees, Cube-Connected Cycles, de Bruijn Network, Star Graphs, etc.
- Indirect Networks (Switch-Based Networks)
  - Regular Topologies
    - Crossbar (Cray X/Y-MP, DEC GIGAswitch, Myrinet)
    - Multistage Interconnection Networks
      - Blocking Networks
        - Unidirectional MIN (NEC Cenju-3, IBM RP3)
        - Bidirectional MIN (IBM SP, TMC CM-5, Meiko CS-2)
      - Nonblocking Networks: Clos Network
  - Irregular Topologies (DEC Autonet, Myrinet, ServerNet)
- Hybrid Networks
  - Multiple-Backplane Buses (Sun XDBus)
  - Hierarchical Networks (Bridged LANs, KSR)
    - Cluster-Based Networks (Stanford DASH, HP/Convex Exemplar)
  - Other Hypergraph Topologies: Hyperbuses, Hypermeshes, etc.

Figura 4.5: Clasificación de las redes de interconexión. (1-D = unidimensional; 2-D = bidimensional; 3-D = tridimensional; CMU = Carnegie Mellon University; DASH = Directory Architecture for Shared-Memory; DEC = Digital Equipment Corp.; FDDI = Fiber Distributed Data Interface; HP = Hewlett-Packard; KSR = Kendall Square Research; MIN = Multistage Interconnection Network; MIT = Massachusetts Institute of Technology; SGI = Silicon Graphics Inc.; TMC = Thinking Machines Corp.)

4.3.1 Redes de medio compartido

La estructura de interconexión menos compleja es aquella en la que el medio de transmisión está compartido por todos los elementos de comunicación. En estas *redes de medio compartido*, sólo un dispositivo puede utilizar la red en un momento dado. Cada elemento conectado a la red tiene circuitos para manejar el paso de direcciones y datos. La red en si misma actúa, normalmente, como un elemento pasivo ya que no genera mensajes.

Un concepto importante es el de la *estrategia de* arbitraje que determina cómo se resuelven los conflictos de acceso al medio. Una característica de un medio compartido es la posibilidad de soportar un *broadcast* atómico en donde todos los dispositivos conectados al medio pueden monitorizar actividades y recibir la información que se está transmitiendo en el medio compartido. Esta propiedad es importante para un soporte eficiente de muchas aplicaciones que necesitan comunicaciones del tipo "uno a todos" (broadcast) o "uno a muchos" (multicast), como las barreras de sincronización y protocolos de coherencia de caché basados en *snoopy*. Debido al limitado ancho de banda, un medio compartido únicamente puede soportar un número limitado de dispositivos antes de convertirse en un cuello de botella.

Las redes de medio compartido pueden dividirse en dos grandes grupos: las redes de área local, usadas principalmente en la construcción de redes de ordenadores cuya distancia máxima no supera unos pocos kilómetros, y los buses usados en la comunicación interna de los uniprocesadores y multiprocesadores.

4.3.1.1 LANs de medio compartido

Es posible utilizar una LAN de alta velocidad como una columna vertebral que permita interconectar ordenadores para proporcionar un entorno de computación distribuido. Físicamente, una LAN usa hilos de cobre o fibra óptica como medio de transmisión. La topología utilizada puede ser un bus o un anillo. Debido a razones de implementación y rendimiento, no es práctico tener un control centralizado o algún mecanismo fijo de asignación de acceso que determine quién puede acceder al bus. Las tres alternativas principales de LANs basadas en un control distribuido se describen a continuación.

Bus de contención

El mecanismo de arbitraje del bus más popular es que todos los dispositivos compitan para tener acceso exclusivo al bus. Debido a la compartición del medio, todos los dispositivos pueden monitorizar el estado del bus y detectar colisiones. Aquí, el término "colisión" significa que dos o más dispositivos están usando el bus al mismo tiempo y sus datos colisionan. Cuando se detecta una colisión, los dispositivos causantes de la misma abortan la transmisión para intentarlo posteriormente. Entre las LANs que utilizan este mecanismo está la Ethernet que adopta el protocolo CSMA/CD (*Carrier-Sense Multiple Access with Collision Detection*). El ancho de banda de la Ethernet es 10 Mbps y la distancia máxima es de 250 metros (cable coaxial). Para romper la barrera de 10 Mbps ha aparecido Fast Ethernet que puede proporcionar un ancho de banda de 100 Mbps.

Token Bus

Una desventaja del bus de contención es su naturaleza no determinística ya que no puede garantizar cuanto se debe de esperar hasta ganar el acceso al bus. Por lo tanto, el bus de contención no es el idóneo para soportar aplicaciones de tiempo real. Para eleminar el compartamiento no determinístico, aparece un enfoque alternativo que implica pasar un testigo entre los dispositivos conectados a la red. El dispositivo que tiene el testigo tiene el acceso al bus. Cuando termina de transmitir sus datos, el testigo se pasa al siguiente dispositivo según algún esquema prefijado. Restringiendo el tiempo máximo que se puede estar en posesión del testigo se puede garantizar una cota superior al tiempo que un dispositivo debe esperar. Arcnet soporta token bus con un ancho de banda de 2.4 Mbps.

Token Ring

Una extensión natural al token bus es la de utilizar una estructura de anillo. El token ring de IBM proporciona anchos de banda de 4 y 16 Mbps sobre cable coaxial. El protocolo FDDI (*Fiber Distributed Data Interface*) es capaz de proporcionar un ancho de banda de 100 Mbps usando fibra óptica.

4.3.1.2 Bus del sistema (backplane bus)

Un *bus del sistema* es la estructura de interconexión más simple para los multiprocesadores basados en bus. Se usa normalmente para interconectar procesadores y módulos de memoria para proporcionar una arquitectura UMA. La figura 4.6 muestra una red con un bus único. Un bus de este tipo consta usualmente de 50 a 300 hilos físicamente realizados mediante una placa base.

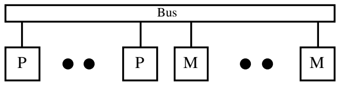


Figura 4.6: Una red con bus único. (M = memoria; P = procesador.)

Existen tres tipos de información en un bus de este tipo: datos, direcciones y señales de control. Las señales de control incluyen la señal de petición del bus y la señal de permiso de acceso al bus, entre muchas otras. Además del ancho de las líneas de datos, el máximo ancho de banda depende de la tecnología. El número de procesadores que pueden conectarse al bus depende de varios factores, como la velocidad del procesador, el ancho de banda del bus, la arquitectura caché y el comportamiento del programa.

4.3.2 Redes Directas

La escalabilidad es una característica importante en el diseño de sistemas multiprocesador. Los sistemas basados en buses no son escalables al convertirse el bus en un cuello de botella cuando se añaden más procesadores. Las *redes directas* o *redes punto a punto* son una arquitectura de red popular y que escalan bien incluso con un número elevado de procesadores. Las redes directas consisten en un conjunto de *nodos*, cada uno directamente conectado a un subconjunto (usualmente pequeño) de otros nodos en la red. En la figura 4.8 se muestran varios tipos de redes directas. Los correspondientes patrones de interconexión se estudiarán posteriormente. Cada nodo es un ordenador programable con su propio procesador, memoria local, y otros dispositivos. Estos nodos pueden tener diferentes capacidades funcionales. Por ejemplo, el conjunto de nodos puede contener procesadores vectoriales, procesadores gráficos, y procesadores de E/S. La figura 4.7 muestra la arquitectura de un nodo genérico. Un componente común en estos nodos es un *encaminador* (router) que se encarga de manejar la comunicación entre los nodos a través del envío y recepción de mensajes. Por esta razón, las redes directas también son conocidas como redes basadas en routers. Cada router tiene conexión directa con el router de sus vecinos. Normalmente, dos nodos vecinos están conectados por un par de canales unidireccionales en direcciones opuestas. También se puede utilizar un canal bidireccional para conectar dos nodos. Aunque la función del encaminador puede ser realizada por el procesador local, los encaminadores dedicados se han usado de forma habitual en multicomputadores de altas prestaciones, permitiendo el solapamiento de la computación y las comunicaciones dentro de cada nodo. Al aumentar el número de nodos en el sistema, el ancho de banda total de las comunicaciones, memoria y capacidad de procesamiento del sistema también aumenta. Es por esto que las redes directasson una arquitectura popular para construir computadores paralelos de gran escala.

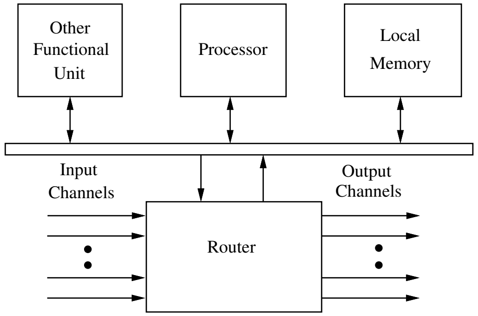


Figura 4.7: Arquitectura de un nodo genérico.

Cada encaminador soporta un número de canales de entrada y salida. Los canales *internos* o *puertos* conectan el procesador/memoria local al encaminador. Aunque normalmente existe únicamente un par de canales internos, algunos sistemas usan más canales internos para evitar el cuello de botella entre el procesador/memoria local y el router. Los canales *externos* se usan para comunicaciones entre los routers. Conectando los canales de entrada de un nodo a los canales de salida de otros nodos podemos definir la red directa. A no ser que se diga lo contrario, utilizaremos el término "canal" para referirnos a un canal externo. A dos nodos conectados directamente se les llaman *vecinos* o nodos *adyacentes*. Normalmente, cada nodo tiene un número fijo de canales de entrada y salida, cada canal de entrada está emparejado con el correspondiente canal de salida. A través de todas las conexiones entre estos canales, existen varias maneras de conectar los diferentes nodos. Obviamente, cada nodo de la red debe de poder alcanzar cualquier otro nodo.

4.3.3 Redes Indirectas

Las *redes indirectas* o *basadas en conmutadores* (switch) forman el tercer grupo de redes de interconexión. En lugar de proporcionar una conexión directa entre algunos nodos, la comunicación entre

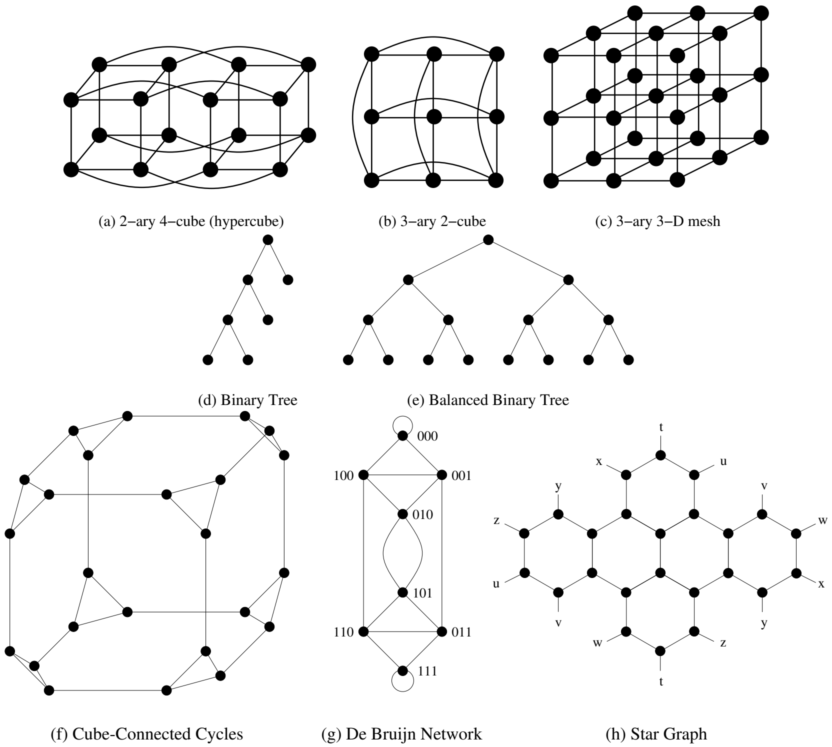


Figura 4.8: Algunas topologías propuestas para redes directas.

cualquier pareja de nodos se realiza a través de *conmutadores*. Cada nodo tiene un adaptador de red que se conecta a un conmutador. Cada conmutador consta de un conjunto de *puertos*. Cada puerto consta de un enlace de entrada y otro de salida. Un conjunto (posiblemente vacío) de puertos en cada conmutador están conectados a los procesadores o permanecen abiertos, mientras que el resto de puertos están conectados a puertos de otros conmutadores para proporcionar conectividad entre los procesadores. La interconexión de estos conmutadores define la topología de la red.

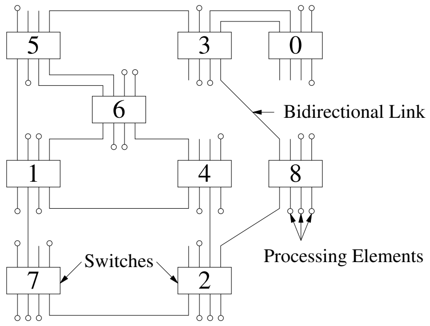


Figura 4.9: Un red conmutada con topología irregular.

Las redes conmutadas han evolucionado considerablemente con el tiempo. Se han propuesto un amplio rango de topologias, desde topologías regulares usadas en los procesadores array (matriciales) y multiprocesadores de memoria compartida UMA a las topologías irregulares utilizadas en la actualidad en los NOWs. Las topologías regulares tienen patrones de conexión entre los conmutadores regulares mientras que la topologias irregulares no siguen ningún patrón predefinido. En el tema siguiente se estudiarán más detenidamente las topologias regulares. La figura 4.9 muestra una red conmutada típica con topología irregular. Ambos tipos de redes pueden clasificarse además según el número de conmutadores que tiene que atravesar un mensaje para llegar a su destino. Aunque esta clasificación no es importante en el caso de topologías irregulares, puede significar una gran diferencia en el caso de redes regulares ya que algunas propiedades específicas se pueden derivar en función de este dato. Ejemplos de redes indirectas son las redes de barras cruzadas y las redes multietapa, que se estudiarán en detalle en el siguiente capítulo.

4.3.4 Redes Híbridas

En esta sección describiremos brevemente algunas topologías que no se encuadran en las vistas hasta ahora. En general, las redes híbridas combinan mecanismos de redes de medio compartido y redes directas o indirectas. Por tanto, incrementan el ancho de banda con respecto a las redes de medio compartido, y reducen la distancia entre nodos con respecto a la redes directas e indirectas. Existen varias aplicaciones concretas de las redes híbridas. Este es el caso de las *bridged LANs*. Sin embargo, para sistemas que necesitan muy alto rendimiento, las redes directas e indirectas consiguen una mejor escalabilidad que las redes híbridas ya que los enlaces punto a punto son más sencillos y más rápidos que los buses de medio compartido. La mayoría de los computadores paralelos de alto rendimiento

usan redes directas o indirectas. Recientemente las redes híbridas han ganado aceptación de nuevo. El uso de tecnología óptica permite la implementación de redes de alto rendimiento.

Se han propuesto redes híbridas para diferentes propósitos. En general, las redes híbridas pueden modelarse mediante hipergrafos, donde los vértices del hipergrafo representan un conjunto de nodos de procesamiento, y las aristas representan el conjunto de canales de comunicación y/o buses. Obsérvese que una arista en un hipergrafo puede conectar un número arbitrario de nodos. Cuando una arista conecta exactamente dos nodos entonces representa un canal punto a punto. En caso contrario representa un bus. En algunos diseños de redes, cada bus tiene un único nodo conductor. Ningún otro dispositivo puede utilizar ese bus. En ese caso, no existe necesidad de arbitraje. Sin embargo, todavía es posible la existencia de varios receptores en un tiempo dado, manteniéndose la capacidad de *broadcast* de los buses. Obviamente, cada nodo de la red debe poder controlar al menos un bus, por lo que el número de buses necesarios no puede ser menor que el número de nodos. En este caso, la topología de la red se puede modelar mediante un hipergrafo directo.

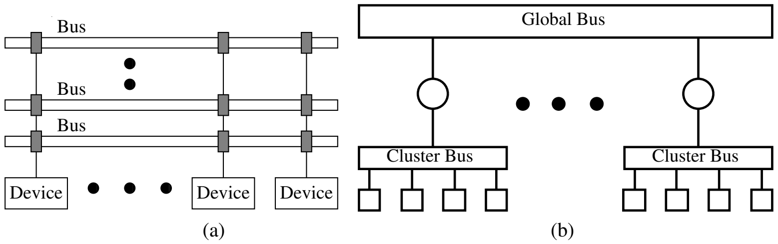


Figura 4.10: Redes Híbridas. (a) Una red multibus. (b) Una jerarquía de dos niveles de buses.

4.3.4.1 Redes multibus

Debido al ancho de banda limitado que proporcionan las redes de medio compartido, una red de medio compartido sólo es capaz de soportar un número pequeño de dispositivos, tiene una distancia limitada, y no es escalable. Algunos investigadores han estudiado como podría eliminarse ese cuello de botella. Una aproximación para incrementar el ancho de banda de la red se muestra en la figura 4.10a. Sin embargo, los problemas de cableado, y el coste del interface hacen que sea de poca utilidad para el diseño de multiprocesadores. Debido a las limitaciones eléctricas de la tecnología del encapsulado, es poco probable tener una red multibus con más de cuatro buses. Sin embargo, utilizando otras tecnologías de empaquetado como la multiplexación por división de la longitud de onda en fibra óptica hace posible integración de múltiples buses.

4.3.4.2 Redes jerárquicas

Otra aproximación para incrementar el ancho de banda de la red es la que se muestra en la figura 4.10b. Diferentes buses se interconectan mediante routers o puentes para transferir información de un lado a otro de la red. Estos routers o puentes pueden filtrar el tráfico de la red examinando la dirección destino de cada mensaje que le llegue. La red jerárquica permite expandir el área de la red y manejar más dispositivos, pero deja de ser una red de medio compartido. Esta aproximación se usa para interconectar varias LANs. Normalmente, el bus global tiene un mayor ancho de banda. En caso contrario, se convertiría en un cuello de botella. Esto se consigue utilizando una tecnología más rápida. Las redes jerárquicas han sido también propuestas como esquema de interconexión para los multiprocesadores de memoria compartida. También en este caso, el bus global puede convertirse en un cuello de botella.

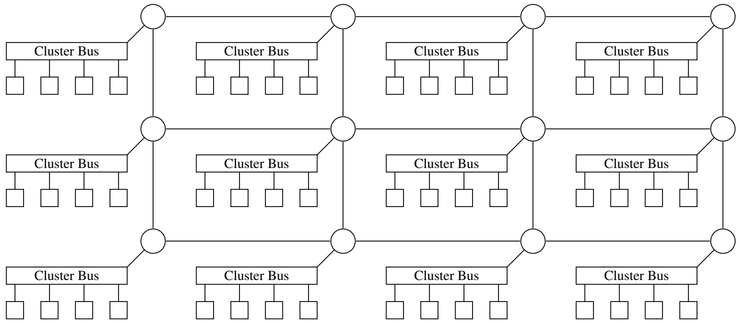


Figura 4.11: Malla bidimensional basada en clusters.

4.3.4.3 Redes basadas en clusters

Las redes basadas en clusters también tienen una estructura jerárquica. Incluso pueden considerarse como una subclase dentro de las redes jerárquicas. Estas redes combinan las ventajas de dos o más clases de redes a diferentes niveles en la jerarquía. Por ejemplo, es posible combinar las ventajas de los buses y los enlaces punto a punto usando buses en el nivel más bajo de la jerarquía para formar clusters, y una topología directa para conectar los cluster en el nivel superior. Este es el caso del computador paralelo DASH (*Stanford Directory Architecture for Shared-Memory*). La figura 4.11 muestra la arquitectura básica de este computador paralelo. En el nivel inferior cada cluster consta de cuatro procesadores conectados mediante un bus. En el nivel superior, una malla 2-D conecta los clusters. La capacidad *broadcast* del bus se usa a nivel de cluster para implementar un protocolo *snoopy* para mantener la coherencia de la caché. La red directa en el nivel superior elimina la limitación del ancho de banda de un bus, incrementando considerablemente la escalabilidad de la máquina.

Es posible realizar otras combinaciones. En lugar de combinar buses y redes directas, el multiprocesador HP/Convex Exemplar combina redes directas e indirectas. Este multiprocesador consta de crossbar no bloqueantes de tamaño 5×5 en en nivel inferior de la jerarquía, conectando cuatro bloques funcionales y un interface de E/S para formar un cluster o *hipernodo*. Cada bloque funcional consta de dos procesdores, dos bancos de memoria e interfaces. Estos hipernodos se conectan a un segundo nivel denominado *interconexión toroidal coherente* formada por múltiples anillos usando el Interface de Coherencia Escalable (SCI). Cada anillo conecta un bloque funcional de todos los hipernodos. En el nivel inferior de la jerarquía, el crossbar permite a todos los procesadores dentro del hipernodo acceder a los módulos de memoria entrelazada de ese hipernodo. En el nivel superior, los anillos implementan un protocolo de coherencia de la caché.

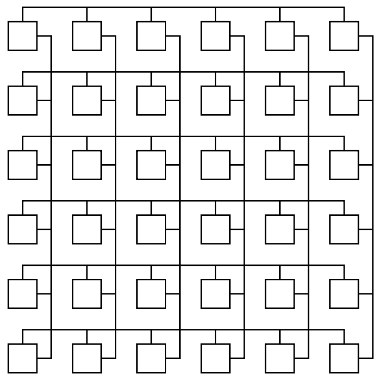


Figura 4.12: Una hipermalla bidimensional.

4.3.4.4 Otras topologías hipergrafo

Se han propuesto muchas otras topologías híbridas. Entre ellas, una clase particularmente interesante es la de *hipermallas*. Una hipermalla es una topología regular basada en un conjunto de nodos dispuestos en varias dimensiones. En lugar de existir conexiones directas entre los vecinos de cada dimensión, cada nodo está conectado a todos los nodos de la dimensión a través de un bus. Existen varias formas de implementar una hipermalla. La más directa consiste en conectar todos los nodos de cada dimensión a través de un bus compartido. La figura 4.12 muestra una hipermalla 2-D. En esta red, los buses están dispuestos en dos dimensiones. Cada nodo se conecta a un bus en cada dimensión. Esta topología fue propuesta por Wittie[1], y se le denomina *spanning-bus hypercube*. Esta topología tiene un diámetro muy pequeño, y la distancia media entre nodos escala muy bien en función del tamaño de la red. Sin embargo, el ancho de banda total no escala bien. Además, los frecuentes cambios en el maestro del bus causan una sobrecarga significante.

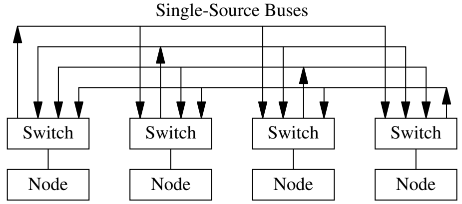


Figura 4.13: Una hipermalla unidimensional con conmutador crossbar distribuido.

Una implementación alternativa que elimina las restricciones señaladas arriba consiste en reemplazar el bus compartido que conecta los nodos a lo largo de una dimensión dada por un conjunto de tantos

[1] L. D. Wittie. *Communications structures for large networks of microcomputers*. IEEE Transactions on Computers, vol C-29, pp. 694-702, August 1980.

buses como nodos existan en esa dimensión. Esta es la aproximación propuesta en la *Distributed Crossbar Switch Hypermesh* (DCSH). La figura 4.13 muestra una dimensión de la red. Cada bus es controlado por un único nodo. Por tanto, no hay cambios en la pertenencia del bus. Además, el ancho de banda escala con el número de nodos. Los dos principales problemas, sin embargo, son el alto número de buses necesarios y el alto número de puertos de entrada y salida que necesita cada nodo.

4.4 Clasificación de las redes según la tecnología y distancia

Para finalizar con las taxonomías de redes, en la figura 4.14 se muestra otra clasificación de las redes que no atiende a su topología sino a la tecnlogía usada y a la distancia que cubren.

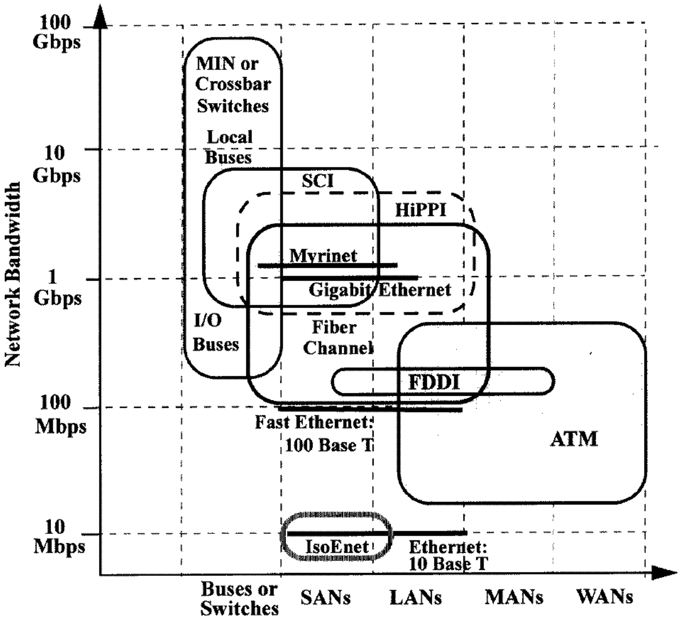


Figura 4.14: Redes de interconexión y tecnologías de redes para sistemas paralelos.

Una red de área de sistema (SAN) conecta procesadores, memorias y tarjetas de interfase con cables pequeños para una distancia corta. Como se ha comentado anteriormente, una LAN se localiza en un edificio, un campus, o una empresa. Una red de área metropolitana (MAN) cubre toda una ciuidad. Una WAN se utiliza para interconectar un número de redes pequeñas y puede ser extendida a grandes distancias.

En la figura 4.15 se definen un conjunto de términos para buses y redes que se usan ampliamente en la industria. Los pins de un procesador forman el *bus del procesador*. El bus que conecta el procesador

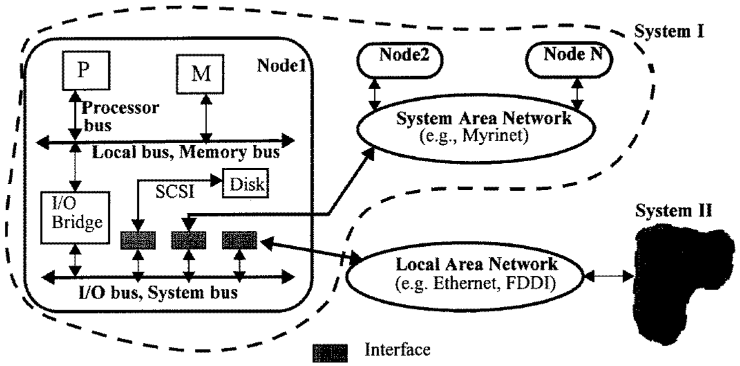


Figura 4.15: Ilustración de diferentes buses, SAN y LAN.

y los módulos de memoria se llama *bus local* o *bus de memoria*, cuya longitud es aproximandamente decenas de centímetros. El *bus del sistema* (también llamado bus E/S) proporciona slots para conectar los dispositvios de E/S, tales como discos, unidades de cinta, tarjetas de interfase de red, etc. Está conectado al bus local a través de un circuito puente de E/S. El bus del sistema se construye normalmente en la tarjeta madre o un backplane, limitandose así a 2 m. El término bus de E/S se usa a menudo para referirse al bus SCSI (small computer system interface). Dependiendo de la tecnología del cableado, un bus de E/S interno raramente excede los 3 m y un cable SCSI externo puede extenderse como mucho 20 m.

Muchos multiprocesadores y multicomputadores dividen su hardware en un número de nodos. La red que conecta esos nodos para formar un sólo sistema se denomina SAN. A difierencia, de la LAN que se utiliza para conectar diferentes sistemas. Una SAN se expande de 3 a 25 m. Una LAN de 25 a 500 m o 2 Km y una MAN 25 Km o más.

4.5 Rendimiento de las redes de interconexión

Vamos a suponer que tenemos una red con dos computadores. Por tanto, si las dos máqinas se quieren comunicar, tendrán que hacerlo mediante el envío y recepción de mensajes.

Así, los parámetros que influyen en el rendimiento son:

- *Ancho de banda*. Máxima velocidad (Mb/s) a la cual la red puede propagar información, una vez que el mensaje entra en la red
- *Tiempo de escape*. Tiempo para que el primer bit del mensaje llegue al receptor, incluyendo los retardos debidos a repetidores u otro hardware de la red. Puede ser ms para una WAN o ns para un MPP.

- *Tiempo de transmisión*. Tiempo para que el mensaje pase a través de la red (no incluye el tiempo de escape) y es igual al tamaño del mensaje dividido por el ancho de banda. Se asume que no hay otros mensajes en la red.
- *Latencia de transporte*. La suma del tiempo de escape y el tiempo de transmisión. Es el tiempo que el mensaje paso en la red de interconexión, sin incluir los overheads de inyectar el mensaje en la red ni de salir cuando llega.
- *Overhead del emiso*r. Tiempo del procesador para inyectar el mensaje en la red de interconexión, incluyendo tanto las componentes software como hardware. Una vez que el procesador está libre, cualquier retardo se considera parte de la lantencia de transporte.
- *Overhead del receptor*. Tiempo para que el procesador coja el mensaje de la red de interconexión, incluyendo tanto las componentes software como hardware.

La latencia total de un mensaje se puede expresar como:

$$\begin{gathered} Latencia\ total = Overhead\ del\ emisor + Tiempo\ de\ escape \\ + \frac{Tamano\ del\ mensaje}{Ancho\ de\ banda} + Overhead\ del\ receptor \end{gathered}$$

En algunas aplicaciones y redes, los overheads dominan la latencia total del mensaje.

Se puede simplificar la ecuación combinando los overhead del emisor, receptor y tiempo de escape en un sólo término llamado *Overhead*:

$$Latencia\ total = Overhead + \frac{Tamano\ del\ mensaje}{Ancho\ de\ banda}$$

Podemos usar esta fórmula para calcular el ancho de banda efectivo comunicado por la red cuando varía el tamaño del mensaje:

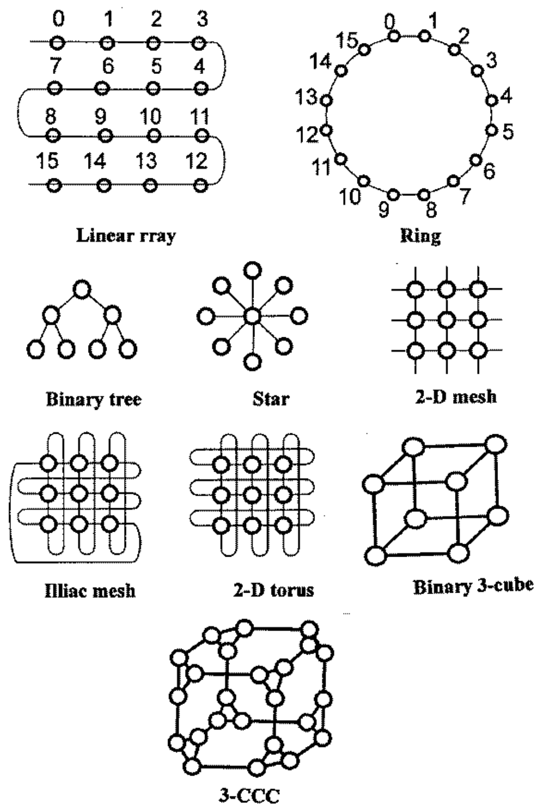
$$Ancho\ de\ banda\ efectivo = \frac{Tamano\ del\ mensaje}{Latencia\ total}$$

Otro tipo de métricas de rendimiento, para las redes directas o indirectas, es decir, casos donde se supone que hay conectados más de dos nodos son :

- *Ancho de banda por puerto*. Máximo número de bits (o bytes) que se pueden transmitir por segundo desde cualquier puerto a otro puerto de la red. En una red simétrica este acho de banda es independiente de la localización del puerto.
- *Ancho de banda agregado*. Máximo número de bits (o bytes) que se pueden transmitir desde una mitad de los nodos a la otra mitad de los nodos por segundo.
- *Ancho de banda de la bisección*. Máximo número de bits (o bytes) por segundo que cruzan todos los cables de la bisección que divide la red en dos partes iguales. Si b es el número de líneas que cruzan el plano de la bisección y w el número de cables por línea (ancho del canal). Si cada cable transmite r b/s, el ancho de banda de la bisección se define como B=bwr b/s.

Ejercicios

1. Completar la tabla rellenado los valores que faltan a partir del tamaño de la red, para las redes que aparecen en la figura.



Tipo de red	Grado	Diametro	Ancho de bisección	No. de links	Simetría	Tamaño de la red
Array lineal						N nodos
Anillo						N nodos
Complet. conec.						N nodos
Arbol binario						h=$\lceil \log_2 N \rceil$
Estrella						N nodos
Malla 2D						$r \times r$ para r=$\sqrt{N}$
Malla Illiac						$r = \sqrt{N}$
Toro 2D						$toro\ r \times r\ r = \sqrt{N}$
Hipercubo						$N = 2^n$ nodos
CCC						$N = k2^k$ nodos $k \geq 3$
k-ary n-cube						$N = k^n$ nodos

2. Diseñar una red directa para un multicomputador con 64 nodos usando un toro de tres dimensiones, un hipercubo binario de seis dimensiones y un cubo de ciclos conectados (CCC) con un diámetro mínimo.

(a) Sea d el grado de un nodo, D el diámetro de la red, y L el número total de líneas de la red. La calidad de la red se mide por $(d \times D \times L)^{-1}$. Categoriza estas tres arquitecturas de acuerdo con la calidad.

(b) La distancia internodo media es el número medio de hops (líneas) a lo largo del camino más corto entre un nodo y otro. La media se calcula para todos los pares (fuentes y destino) . Ordenar las tres arquitecturas de acuerdo con su distancia media internodo, asumiendo que la probabilidad de que un nodo envie un mensaje a otros nodos con distancia i es $(D - i + 1)/\sum_{k=1}^{D} k$.

3. Asume una red con un ancho de banda de 10 Mb/s tiene un overhead de emisión de 230 μs y un overhead de recepción de 270 μs. Supón que las dos máquinas están a 100 metros y una quiere enviar un mensaje de 1000 bytes a la otra (incluyendo la cabecera).

(a) Calcular la latencia total para enviar el mensaje de una máquina a la otra.

(b) Realizar los mismos cálculos que en el apartado (a) asumiendo ahora que las máquinas están a 1000 km.

4. Asume las siguientes redes de 64 nodos: Bus, totalmente conectada, anillo y toro 2D. Suponemos que realizan comunicación todos con todos: cada nodo hace una transferencia a cada nodo. Vamos a simplificar el costo de comunicación de forma que se utiliza una unidad de tiempo para pasar de un conmutador a otro, y no hay costo dentro ni fuera del procesador.

(a) Asume que cada link de cada interconexión tiene la misma velocidad y que un nodo puede enviar tantos mensajes como quiera a la vez. ¿Cuánto tiempo necesitan para completar la comunicación? (Ancho de banda de la bisección).

(b) Realiza los mismos cálculos que en el apartado (a) pero suponiendo que cada nodo sólo puede repartir un mensaje en cada unidad de tiempo.