Tema 3. Multicomputadores tipo cluster

Introducción

- Introducción
- Características generales
- Tipos de clusters
- Modelos de Almacenamiento
- Redes para clusters

Tema 3. Multicomputadores tipo cluster

Introducción

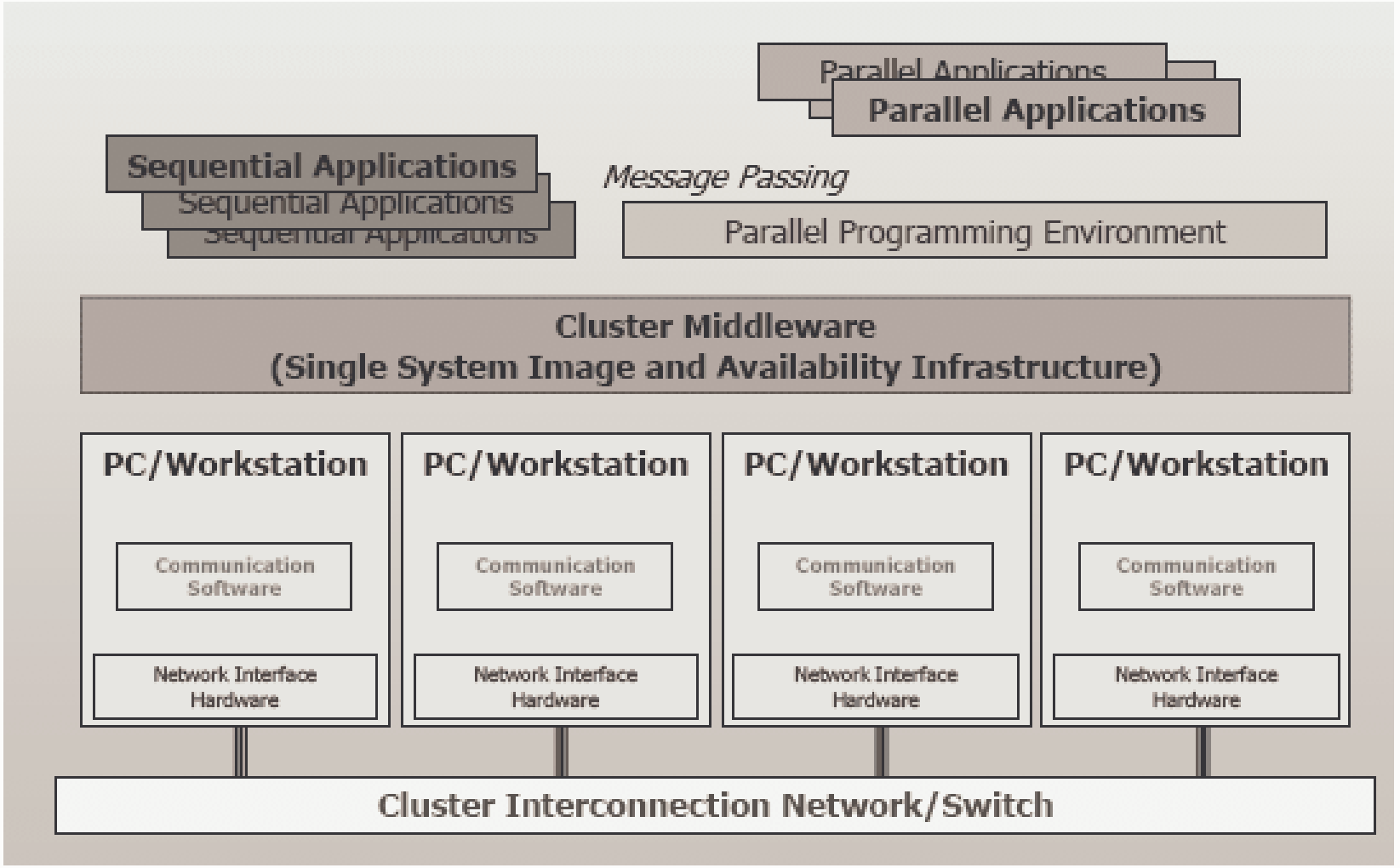
Un cluster es una tipo de arquitectura paralela distribuida que consiste de un conjunto de computadores independientes (y bajo coste en principio) interconectados operando de forma conjunta como un único
recurso computacional

Sin embargo, cada computador puede utilizarse de forma independiente o separada

Tema 3. Multicomputadores tipo cluster

Introducción

ARQUITECTURA DE UN CLUSTER



Tema 3. Multicomputadores tipo cluster

Características generales

COMPONENTES CLUSTER

- Múltiples nodos de computación
  - Los nodos pueden encontrarse encapsulados en un solo contenedor (típicamente interconectados por una SAN), o estar físicamente diferenciados (interconectados por una LAN).
  - Los nodos pueden ser PCs, estaciones de trabajo o SMPs.
- Diferentes sistemas operativos
  - UNIX, Linux, W2000, WXP, …
- Red de interconexión de altas prestaciones
  - Myrinet, Infiniband, Gigabit Ethernet, …
- Diferentes tipos de interfaz de red
  - En el bus E/S, en el bus de sistema, en el procesador.
- Protocolos rápidos de comunicación
  - Active messages, Fast messages, VIA, …

Tema 3. Multicomputadores tipo cluster

Características generales

COMPONENTES CLUSTER

- Middleware: SSI (Single System Image)
  - SSI ofrece al usuario una visión unificada de todos los recursos del sistema
  - SSI se define mediante hardware o software.
- Middleware: Disponibilidad
  - Infraestructura de alta disponibilidad, que incluye servicios de checkpointing, recuperación tras fallo, tolerancia a fallos, …
- SSI puede implementarse a tres niveles
  - Nivel hardware: Se ve el cluster como un sistema de memoria compartida distribuida (ej. Digital Memory Channel, DSM, …).
  - Nivel kernel SO: Ofrece alto rendimiento a las aplicaciones secuenciales y paralelas. Por ejemplo, gang-scheduling para programas paralelos, identificación de recursos sin uso, acceso global a recursos, migración de procesos (balanceo de carga), … (ej. Solaris MC, GLUnix, …).
  - Nivel aplicación o subsistema: Software de planificación y gestión de recursos (ej. LSF, CODINE, CONDOR, …), herramientas de gestión del sistema, sistemas de ficheros paralelos, …

Tema 3. Multicomputadores tipo cluster

Características generales

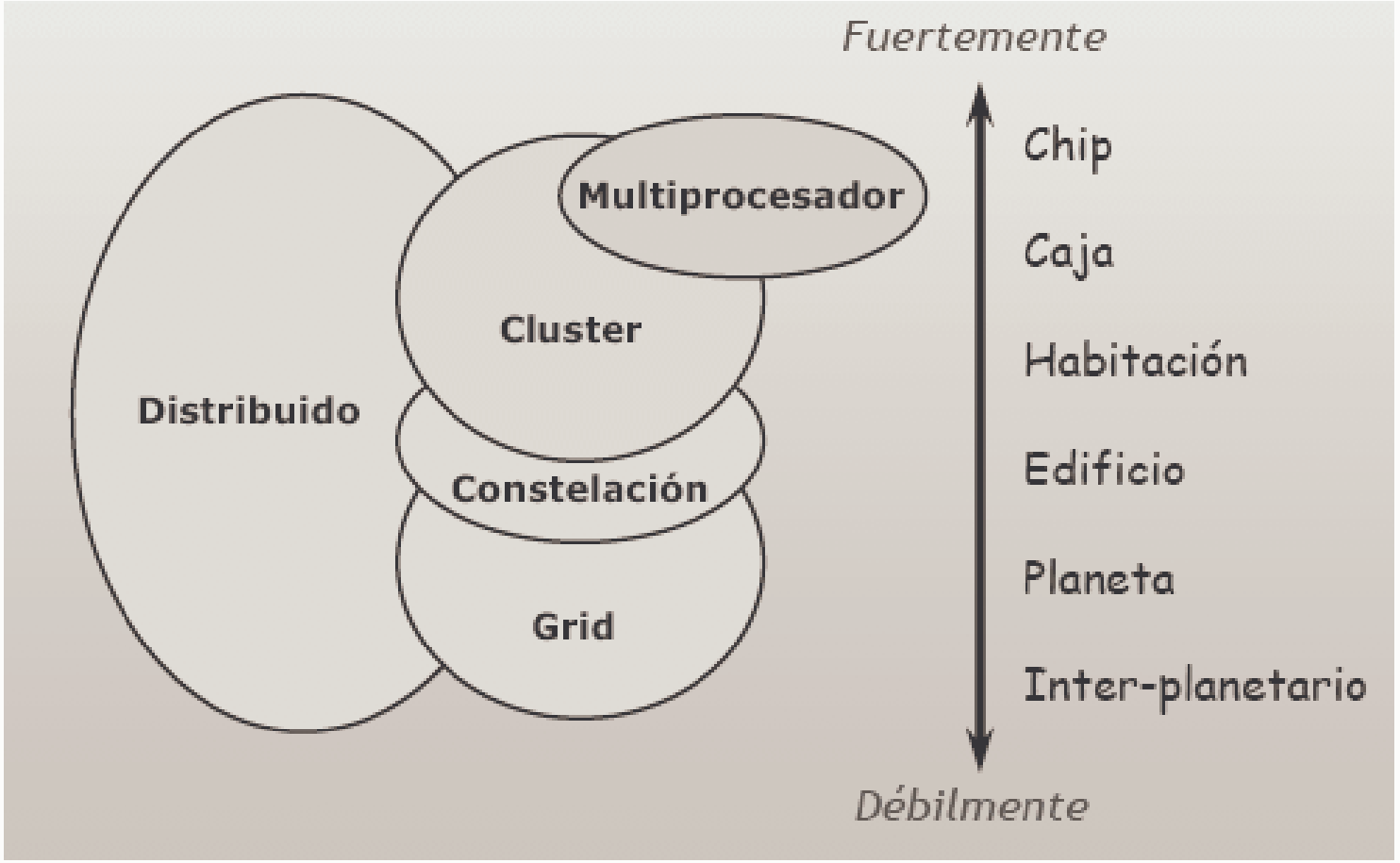
COMPONENTES CLUSTER

- Entornos y herramientas de programación paralela
  - Programación: MPI, PVM, OpenMP, DSM (Threadmarks, Linda), …
  - Depuradores paralelos: TotalView.
  - Análisis de rendimiento: VT (IBM SP), MPE, Pablo, Vampir, …
  - Administración: Parmon.
- Aplicaciones
  - Aplicaciones paralelas o distribuidas, secuenciales.

Tema 3. Multicomputadores tipo cluster

Características generales

ACOPLAMIENTO
DE LOS NODOS



Tema 3. Multicomputadores tipo cluster

Características generales

ACOPLAMIENTO
DEL SOFTWARE

- Integración que tengan todos los elementos software que existan en cada nodo

- Tipos:
  - Acoplamiento fuerte.
    - Software cuyos elementos se interrelacionan mucho unos con otros y posibilitan la mayoría de las funcionalidades del cluster de manera altamente cooperativa

    - Un Kernel distribuido entre los nodos (S.O. Distribuido)

    - Kernel en cada nodo que presentan todo el cluster como un sistema computador homogéneo con acceso a todos los recursos.

    - Sistema de **nombres únicos** y mapeo de todos los recursos físicos

Tema 3. Multicomputadores tipo cluster

Características generales

ACOPLAMIENTO
DEL SOFTWARE

- Tipos:
  - Acoplamiento medio.
    - No se necesita conocimiento exhaustivo de los recursos de otros nodos.
    - Sin embargo hay recursos que se siguen presentando de forma unificada aunque no pertenezcan a un solo nodo (Openmosix Capacidad de computación)
  - Acoplamiento débil.
    - Basados en aplicaciones construidas mediante bibliotecas: MPI, PVM, CORBA...
    - No existe capa de software que homogeneice el conjunto de los recursos. El cluster es visto como un conjunto de computadores a los cuales se tiene cierto acceso.

Tema 3. Multicomputadores tipo cluster

Características generales

MODELO DE GESTIÓN

- Centralizado:
  - Un nodo maestro para configurar el comportamiento de todo el sistema.
  - Este nodo es un punto crítico del sistema.
  - Facilita una mejor gestión del cluster

- Descentralizado:
  - Modelo distribuido donde cada nodo se administra y gestiona.
  - Pueden utilizar aplicaciones de alto nivel centralizadas para gestionar.
  - Información de configuración en archivos locales
  - Más tolerancia a fallos
  - Mayor dificultad en administración

Tema 3. Multicomputadores tipo cluster

Características generales

DIFERENCIAS ENTRE NODOS

- Homogéneo:
  - Todos los nodos se basan en la misma arquitectura y presentan recursos muy similares.
- Heterogéneo:
  - Cluster compuesto de nodos distintos en alguno de estos aspectos:
    - Tiempos de acceso distintos
    - Arquitecturas distintas
    - Sistemas Operativos distintos.
    - Rendimientos de los recursos distintos

Tema 3. Multicomputadores tipo cluster

Tipos clusters

DIFERENCIAS ENTRE NODOS

- Objetivos de los clusters:
  - •Mejorar rendimiento/abaratando coste
  - •Disminuir factores de riesgo del sistema
  - •Escalabilidad
- Clasificación:
  - • Alto rendimiento (High Performance)
  - • Alta disponibilidad (High Availability)
  - • Alta confiabilidad (High Confiability)

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alto rendimiento (High Performance)

  - Objetivo: mejorar el rendimiento en la obtención de la solución de un problema, en términos bien del tiempo de respuesta bien de su precisión
  - Aplicaciones: Generalmente estos problemas de computo suelen estar ligados a Cálculos matemáticos, Renderizaciones de gráficos, Compilación de programas, Compresión de datos, Descifrado de códigos, Rendimiento del sistema operativo...

  - Nivel implementación:

    - Librerías y bibliotecas de funciones. No implementan balanceo carga. El trabajo es repartido de forma manual.
    - Sistema operativo. Basan su funcionamiento en la compartición de los recursos y balanceo de carga dinámico.
    - Soluciónes Híbridas Ej: (PVM “procesos memoria compartida” + OpenMosix “Balanceo de carga dinámico”)

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alta disponibilidad (High Availability)
    - Más demandados por empresas para asegurar servicio a clientes
    - Objetivo: Ofrecer máxima disponibilidad de los servicios prestados por el cluster (24x7x365). Competencia para abaratar sistemas redundantes.
    - Intentan proporcionar fiabilidad, disponibilidad y servicios RAS.
        - Soluciones Hardware. Caras, Hardware redundante funcionando en paralelo y con sistemas de detección de fallos y recuperación.
        - Soluciones Software tipo Cluster.

Tema 3. Multicomputadores tipo cluster

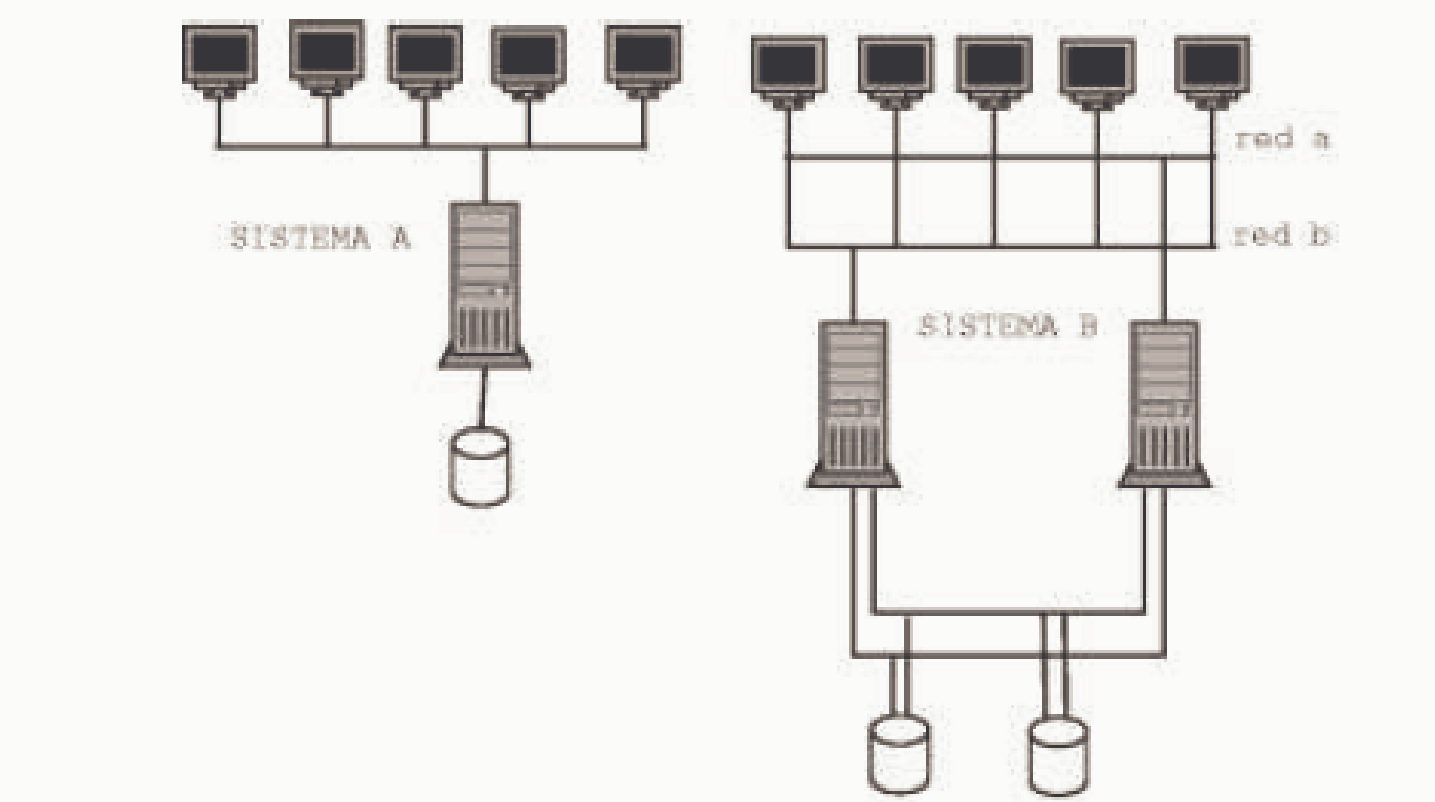
Tipos clusters

- Alta disponibilidad (High Availability)
    - Técnicas para proporcionar disponibilidad.
        - Basadas en redundancia sobre dispositivos críticos. El dispositivo redundante toma el control ante el fallos del dispositivo maestro.
            - Redundancia aislada. Existen dos posibilidades para dar una funcionalidad o servicio (procesadores, fuentes, imágenes de S.O...)
            - N-Redundancia. N equipos para proporcionar servicio. Mayor tolerancia a fallo.

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alta disponibilidad (High Availability)



Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alta disponibilidad (High Availability)
    - Modelo de funcionamiento de los dispositivos redundantes
        - Hot Standby. En cuanto un nodo maestro cae, el nodo esclavo ocupa su lugar. Si no ocurre fallo, nodo esclavo no realiza operaciones salvo backups (desperdicia recursos)
        - Carga mutua. Intentar dar más responsabilidades al nodo esclavo para rentabilizar su coste. Si al caer el maestro el nodo esclavo mantiene sus servicios les suma los servicios del maestro hay una pérdida de rendimiento. Lo más óptimo es que el esclavo abandone sus servicios para atender los que el maestro ha dejado de proporcionar. Más difícil de implementar.
        - Tolerante a fallos. Basados en N-Redundancia (Caen N-1 y sigue funcionando).
    - Técnicas basadas en reparación. Basadas en un sistema de diágnóstico de fallos y recuperación automática (Por ejemplo a partir de backup).

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alta disponibilidad (High Availability). Soluciones libres:
  - Linux HA. (http://linux-ha.org/)
    - The High-Availability Linux Project
    - Comenzo como lista de correo creada por Harald Milz para discutir como crear capacidades de alta disponibilidad en Linux.
    - Heartbeat (1998 Milz). Tecnología para la detección de inclusión o fallos de nodos en un cluster. Basado en envío periódico de paquetes interrogantes. Necesita gran cantidad de ancho de banda. SUSE Linux, Mandriva Linux, MSC Linux, Debian GNU/Linux, Ubuntu Linux, Red Flag Linux, and Gentoo Linux

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alta disponibilidad (High Availability). Soluciones libres:
  - Linux HA. (http://linux-ha.org/). Componentes:
    - Membership services (Miembros)
    - Communication Services (Comunicaciones)
    - Cluster management (Administración)
    - Resource (I/O) fencing (Tratamiento de nodos inciertos)
    - Resource Monitoring (Monitorización)
    - Storage Sharing/Replication Storage (Otros proyectos):
      - Storage Sharing (shared SCSI, FDDI, etc.)
      - Storage Replication
      - Application protocol (DNS, NIS, etc.)
      - rsync, etc.
      - DRBD, nbd, etc.

Tema 3. Multicomputadores tipo cluster

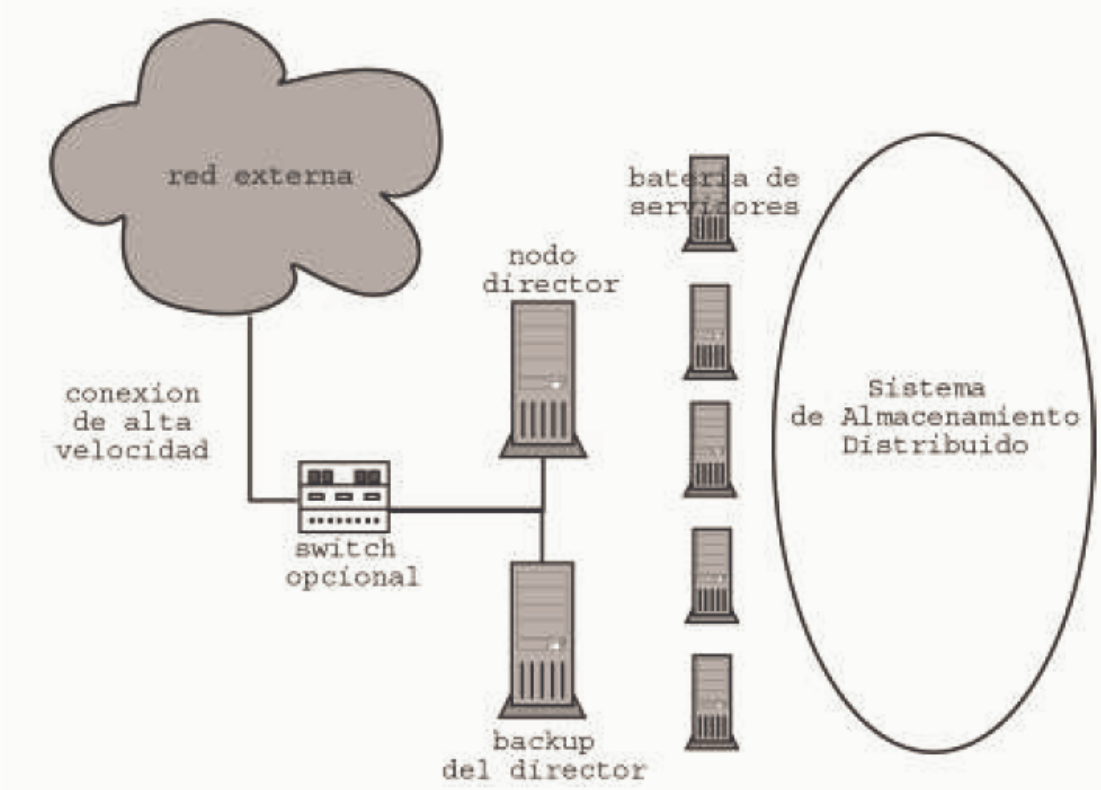
Tipos clusters

- Alta disponibilidad (High Availability). Soluciones libres:
  - Linux Virtual Server.
    - Objetivo: Proporcionar un entorno básico para construir un sistema que proporcione servicios con alta escalabilidad y disponibilidad basado en un conjunto de computadores tipo cluster.
    - Ofrecer todo el cluster externamente como un único servidor virtual.
    - Arquitectura
      - Nodo director recoge las peticiones externas y las reparte entre los nodos del cluster según el criterio de balanceo de carga adoptado.
      - Los servidores reales proporcionan los servicios habituales.
      - Repositorio común de datos. No es necesario, pero si facilita modificación de datos (supóngase un servidor web). Debe ser tolerante a fallos. Sistema de archivos compartidos con manejo de bloqueos (Linux: GFS, Coda, Intermezzo).

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alta disponibilidad (High Availability). Soluciones libres:
  - Linux Virtual Server.



Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alta disponibilidad (High Availability). Soluciones libres:
  - Linux Virtual Server.
    - La alta disponibilidad se obtiene incluyendo sistemas de monitorización de los distintos servicios proporcionados.
    - Si algún servicio falla, se borra de la lista de la que dispone el director.
    - Es posible utilizar un director de respaldo para la posible caída del director principal.
    - La detección de la caída del director principal o de uno de los servidores reales se puede realizar mediante software específico (Heartbeat)

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alta disponibilidad (High Availability). Soluciones libres:
  - Linux Virtual Server. Proceso.
    - Cliente pide servicio al director.
    - Si el puerto destino corresponde a un servicio virtual, director elige un servidor real para prestar el servicio (siguiendo el algoritmo de balanceo planificación).
    - Se incluye una entrada a una tabla para posteriores conexiones.
    - El servidor real, procesa las peticiones como si le viniera de un cliente y devuelve contestación.

Tema 3. Multicomputadores tipo cluster

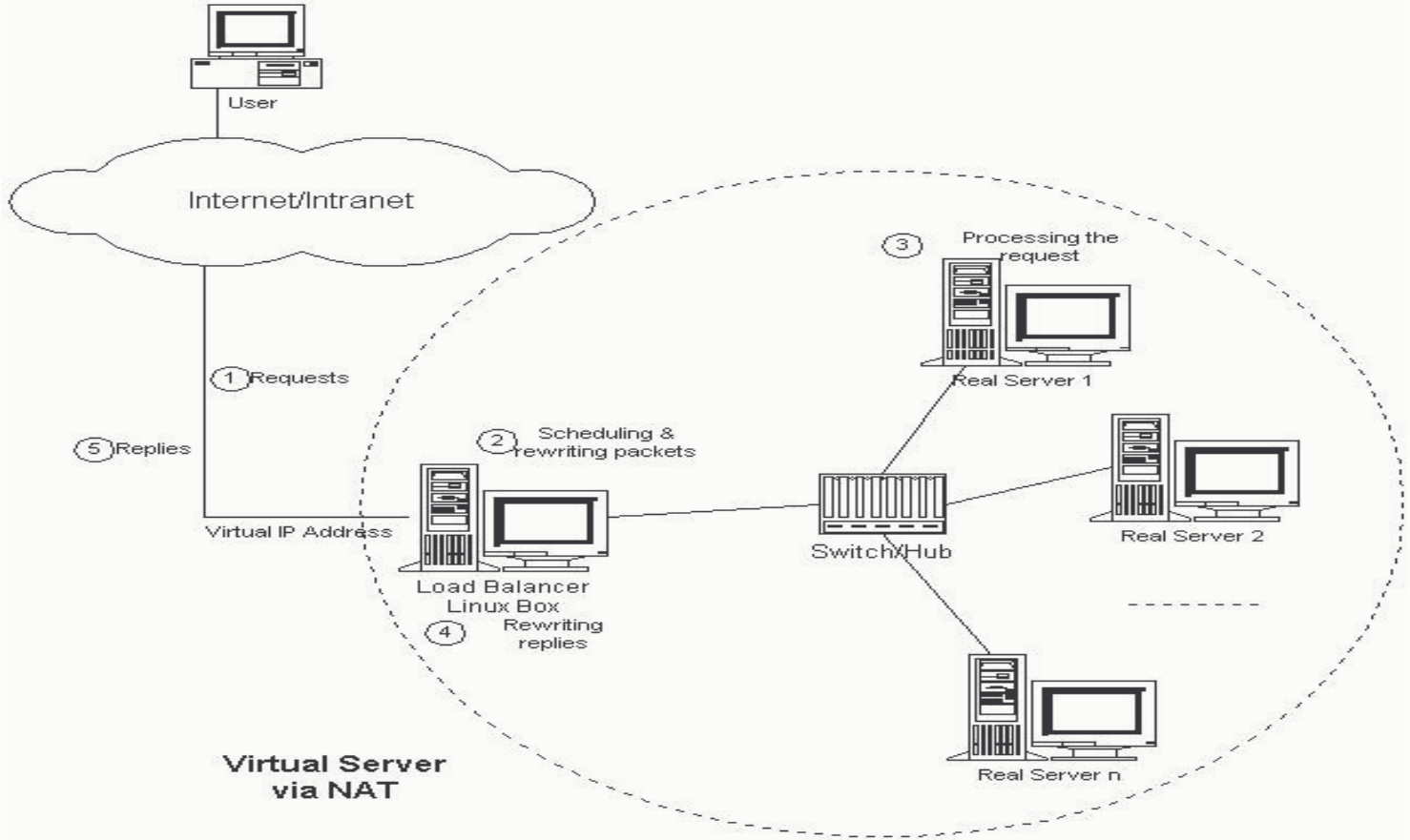
Tipos clusters

- Alta disponibilidad (High Availability). Soluciones libres:
    - Linux Virtual Server. Técnicas de balanceo de carga.
        - NAT (Network Address Translation).
        - IP Tunneling
        - Direct Routing

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- NAT (Network Address Translation).
  - Sustitución de direcciones IP en los paquetes.
  - Director actua como Gateway procesando paquetes IP.



Tema 3. Multicomputadores tipo cluster

Tipos clusters

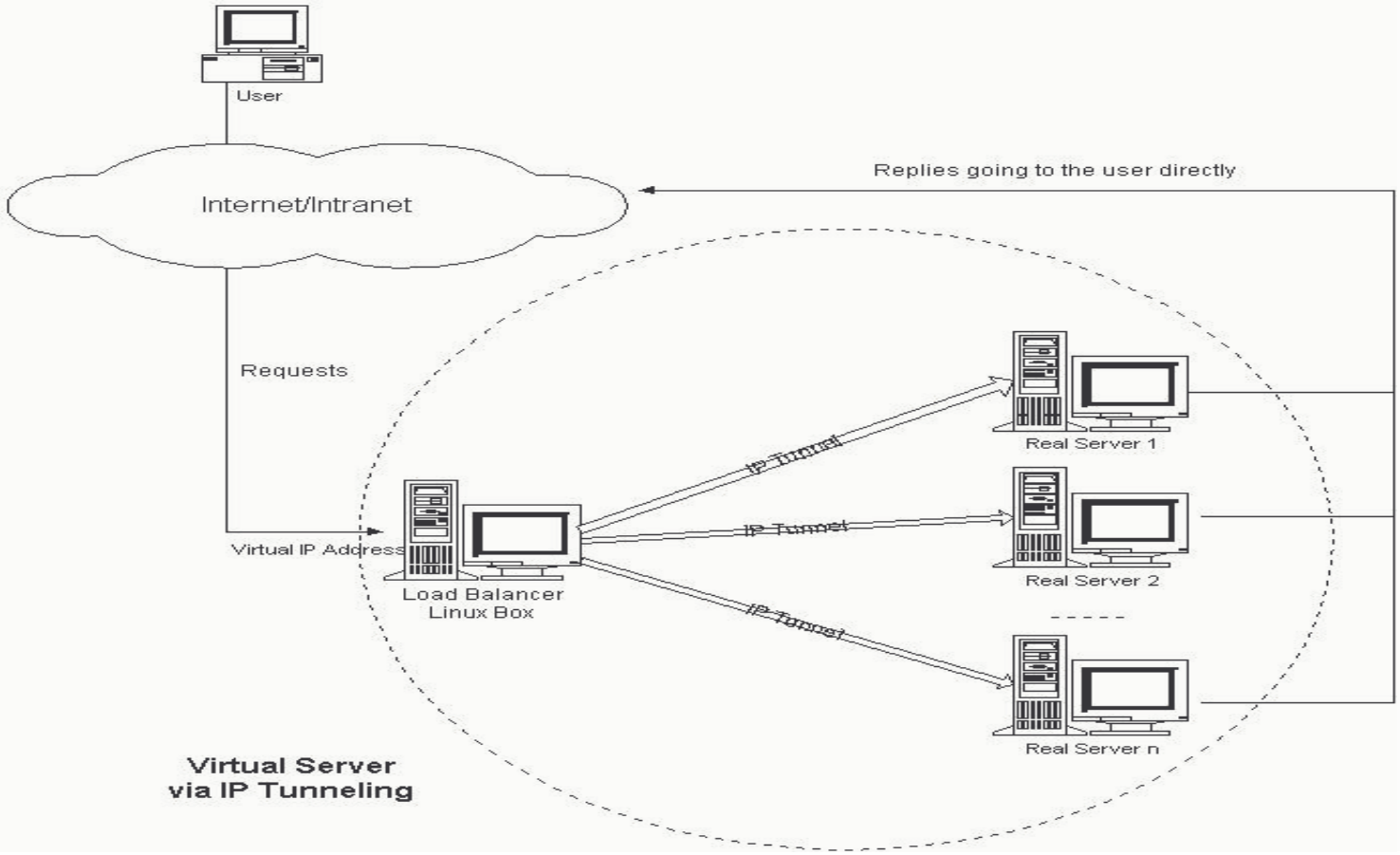
- NAT (Network Address Translation).
    - Ventajas:
        - Sencillez (No cambios en S.O.).
    - Desventajas:
        - Servidores reales en la misma red que el director.
        - Poca escalabilidad. Director debe reescribir todos los paquetes enviados por los servidores reales.
        - Solución-> Servidores reales envíen directamente los paquetes a los clientes y no al director.

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- IP Tunneling
  - Encapsula datagrama IP dentro de otro datagrama IP.
  - Director encapsula petición de los clientes y envia a servidores reales.

  Servidores reales ya pueden contestar directamente a clientes.



Tema 3. Multicomputadores tipo cluster

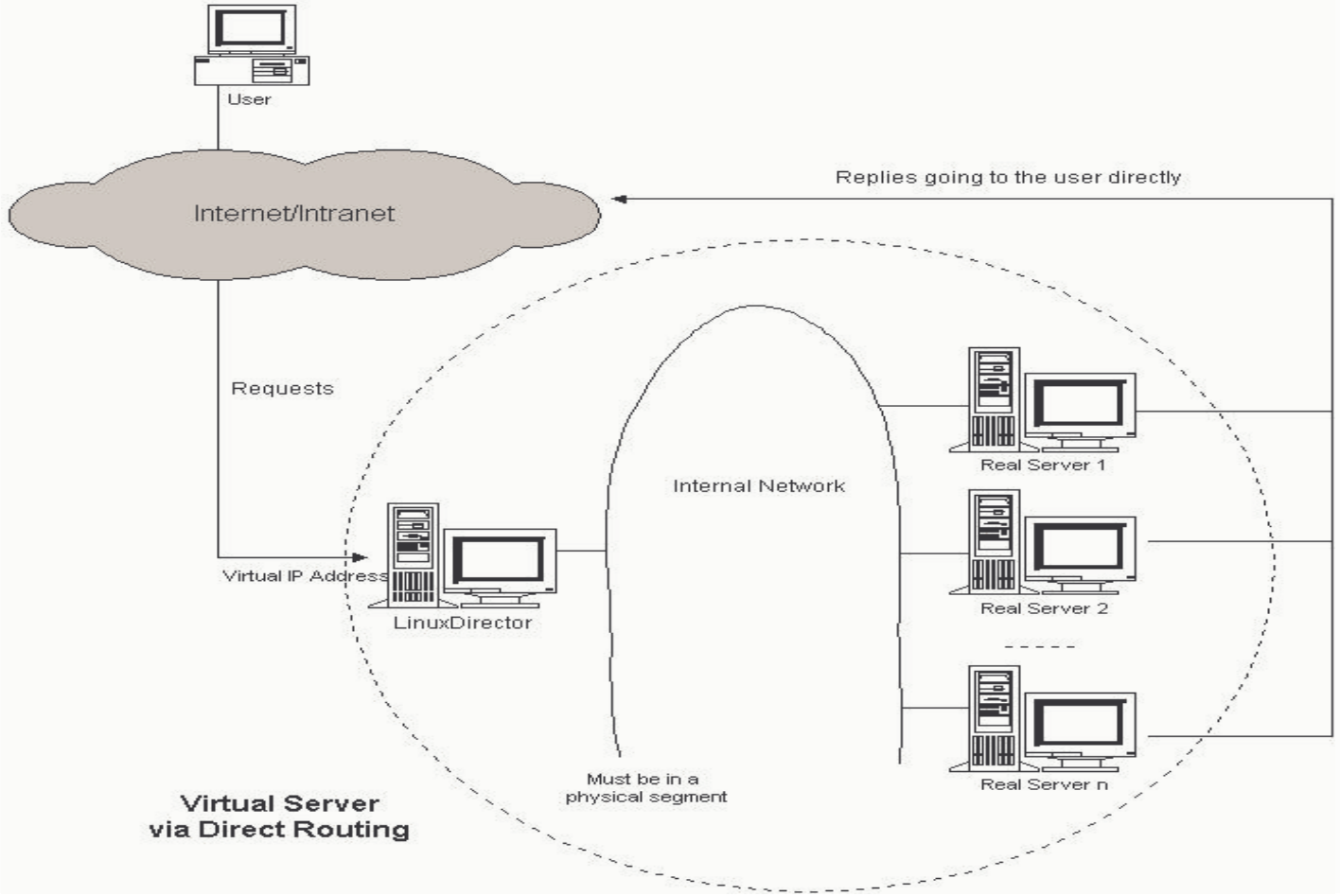
Tipos clusters

- IP Tunneling.
    - Ventajas:
        - Servidores reales no en la misma red que el director.
        - Mayor escalabilidad
    - Desventajas:
        - S.O. Deben soportar IP Tunneling y estar configurados.

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Direct Routing
  - Todos los servidores tienen la misma dirección IP
  - Director cambia la dirección MAC en los paquetes (Mantiene IP cliente)



Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Direct Routing.
    - Ventajas:
        - Mayor escalabilidad
        - No necesita encapsulación
    - Desventajas:
        - Servidores reales no deben responder a peticiones ARP.
        - Todos computadores misma red

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alta disponibilidad (High Availability). Windows (www.windowsclusters.org):
  - Servicio de Cluster
    - Windows Server 2003 Enterprise Edition/Datacenter Edition
    - 8 nodos en clúster
    - Proporciona alta disponibilidad y escalabilidad en aplicaciones críticas (Database, ERP, OLTP, file and print, e-mail)
  - Servicio de Balanceo de carga de red (NLB MAnager)
    - Todos Windows Server 2003.
    - Equilibra carga de tráfico IP entrante a través de los clústeres.
    - Usado para aplicaciones que escalan horizontalmente (Web server, Proxy,...)

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alta disponibilidad (High Availability). Windows:

Table 1: Server Cluster and NLB compared

Server Cluster	NLB
Used for databases, e-mail services, line of business (LOB) applications, and custom applications	Used for Web servers, firewalls, and Web services
Included with Windows Server 2003, Enterprise Edition, and Windows Server 2003, Datacenter Edition	Included with all four versions of Windows Server 2003
Provides high availability and server consolidation	Provides high availability and scalability
Can be deployed on a single network or geographically distributed	Generally deployed on a single network but can span multiple networks if properly configured
Supports clusters up to eight nodes	Supports clusters up to 32 nodes
Requires the use of shared or replicated storage	Doesn't require any special hardware or software; works "out of the box"

Tema 3. Multicomputadores tipo cluster

Tipos clusters

- Alta disponibilidad (High Availability). UNIX(2000):

Vendor	Operating System Software	Cluster Software
Compaq	Tru64 UNIX	TruCluster Server
Data General	DG/UX	DG/UX Clusters
Hewlett-Packard	HP/UX	MC/ServiceGuard
IBM	AIX	HACMP
Sequent	Dynix/ptx	ptx/CLUSTERS
Sun	Solaris	Sun Clusters

Tema 3. Multicomputadores tipo cluster

Modelo almacenamiento

- DAS (Directly Attached Storage)
  - Dispositivos de almacenamiento conectados a servidores directamente. Disco duro, array RAID, cinta, óptico,... en un
- NAS (Network Attached Storage)
  - Dispositivos de almacenamiento conectados en red y accesibles mediante protocolos de acceso a ficheros remotos (NFS, CIFS,...).
- SAN (Storage Area Networks)
  - Dispositivos de almacenamiento conectados en red propia (diferente a la red de datos)

Tema 3. Multicomputadores tipo cluster

Modelo almacenamiento

- DAS (Directly Attached Storage)
  - Disco duro, array RAID, cinta, óptico,... en un servidor que procesa las peticiones de los clientes.
  - Espacio de almacenamiento en compartimentos estanco.
  - Se genera redundancia innecesaria (copias de ficheros en distintos servidores)
  - Copias de seguridad colapsan la red de datos

Tema 3. Multicomputadores tipo cluster

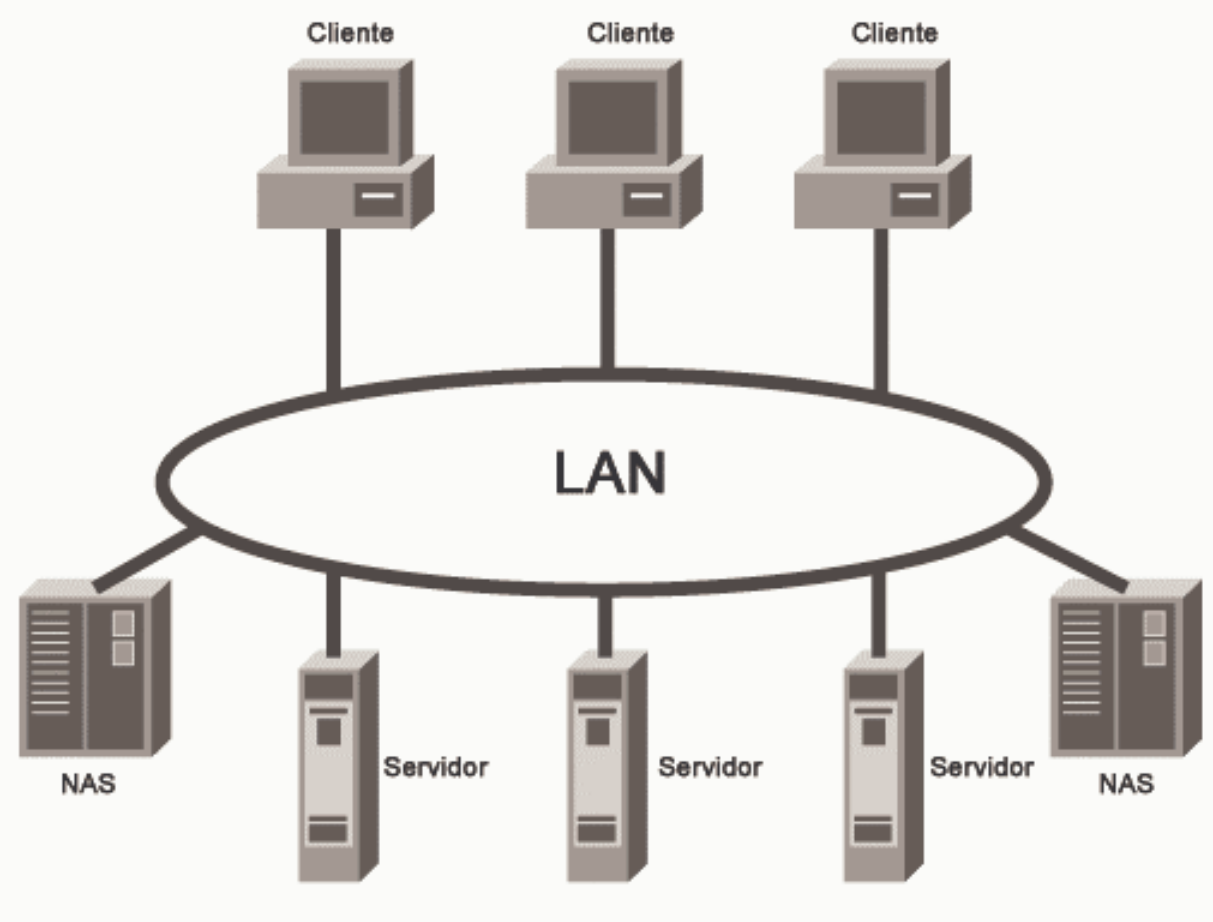
Modelo almacenamiento

- NAS (Network Attached Storage)
  - Acceso a datos en modo fichero.
  - Control del acceso al disco realizado por el propio dispositivo NAS (libera a los servidores).
  - Utiliza la red de datos (LAN o WAN), protocolo IP. Limitación en el escalado.
  - Solución económica y adecuada para compartir información (Servidores web, servidores de ficheros) con pocos requisitos en rendimiento.

Tema 3. Multicomputadores tipo cluster

Modelo almacenamiento

- NAS (Network Attached Storage)



Tema 3. Multicomputadores tipo cluster

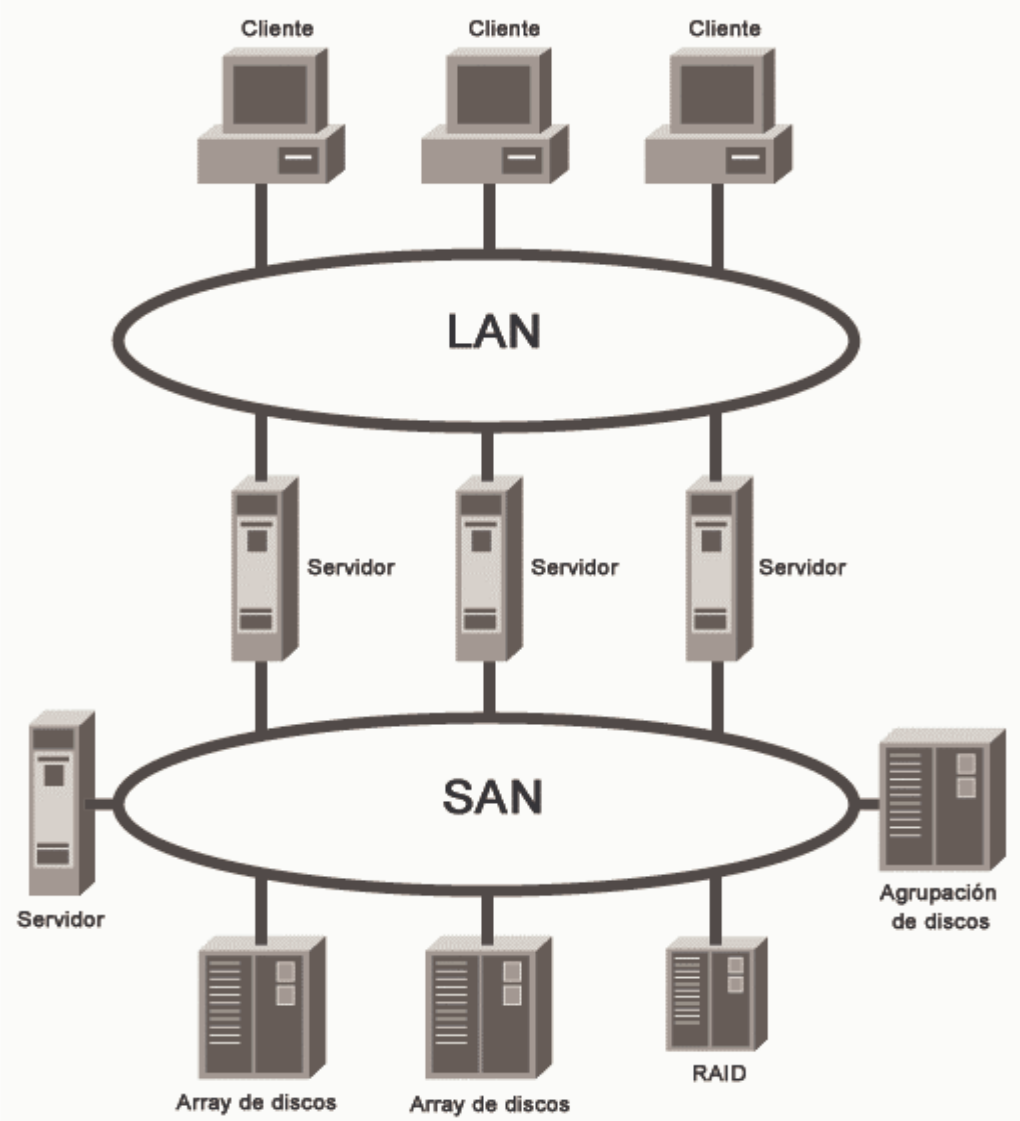
Modelo almacenamiento

- SAN (Storage Area Networks)
  - Red específica para el almacenamiento
  - Protocolos propios de comunicación (Fibrechannel, i-SCASI SCASI sobre IP,...)
  - Se ofrece acceso directo a disco, recuperación en modo bloque, no en fichero utilizado por NAS.
  - Solución más cara, escalable y adecuada cuando se espera un volumen de almacenamiento muy elevado o los servicios que hacen uso del almacenamiento tienen requisitos de rendimiento crítico (ERPs, Bases de Datos, etc.)

Tema 3. Multicomputadores tipo cluster

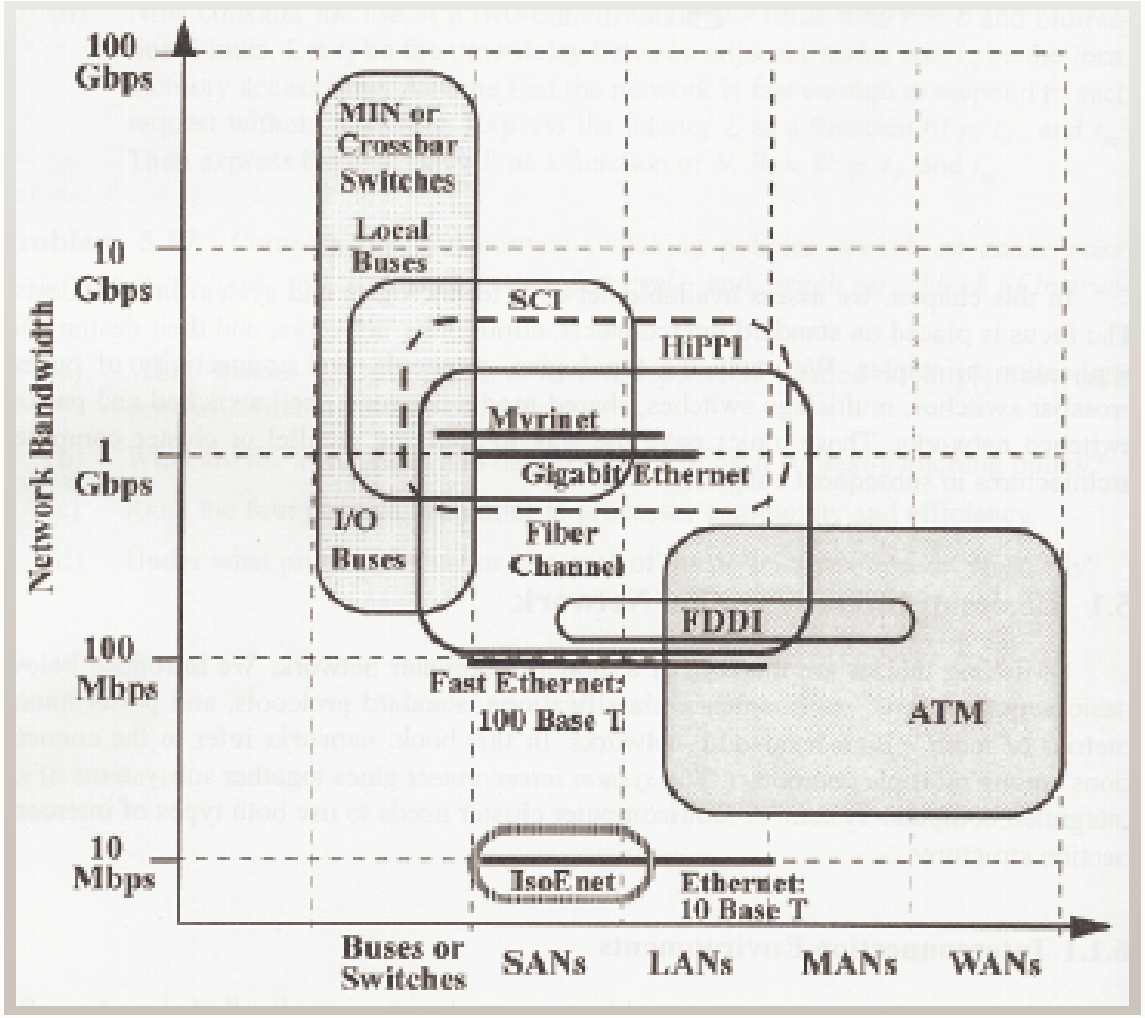
Modelo almacenamiento

- SAN (Storage Area Networks)



Tema 2. Sistemas de comunicación en computadores paralelos

Redes para clusters



Tema 2. Sistemas de comunicación en computadores paralelos

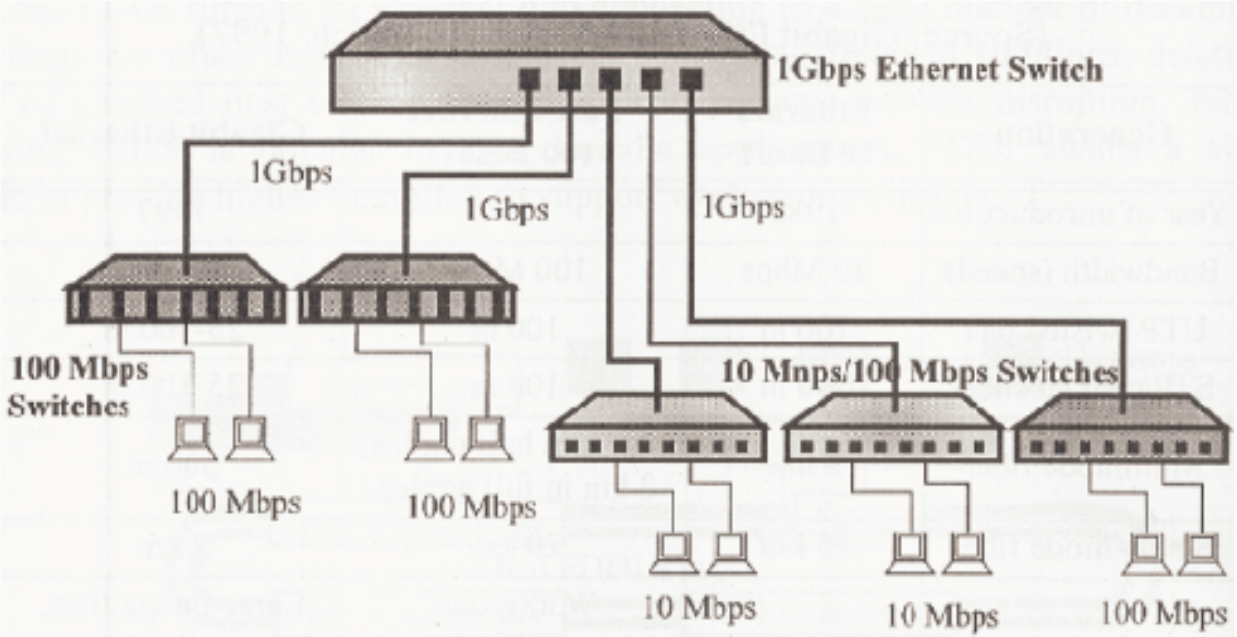
Redes para clusters

- Redes Ethernet
  - Ethernet
    - LAN introducida en 1982 y más utilizada en los años 90
    - Ancho de banda a 10 Mbits/seg
    - No muy adecuada para clusters debido a su bajo ancho de banda
    - Basada en el concepto de dominios de colisión
  - Fast Ethernet
    - LAN introducida en 1994 y más utilizada actualmente
    - Ancho de banda a 100 Mbits/seg
    - Mediante el uso de conmutadores y hubs se pueden definir varios dominios de colisión separados
    - El S.O. interviene en la introducción y extracción de los mensajes, vía interrupciones
    - La latencia aplicación-aplicación depende del driver y del API
      - TCP/IP aprox. 150μs
      - U-Net, MVIA aprox. 50 μs

Tema 2. Sistemas de comunicación en computadores paralelos

Redes para clusters

- Redes Ethernet
  - Gigabit Ethernet
    - LAN introducida en 1998
    - Ancho de banda a 1-10 Gbits/seg
    - Basado en conmutadores punto-a-punto rápidos



Tema 2. Sistemas de comunicación en computadores paralelos

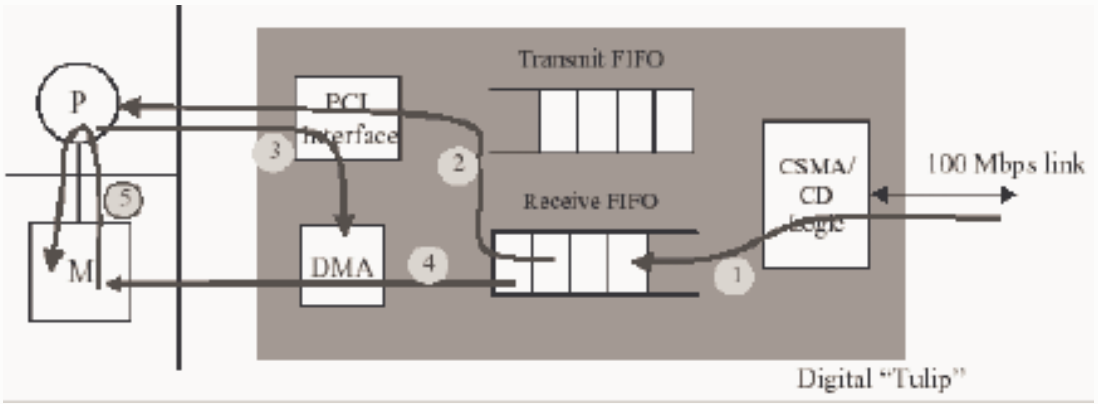
Redes para clusters

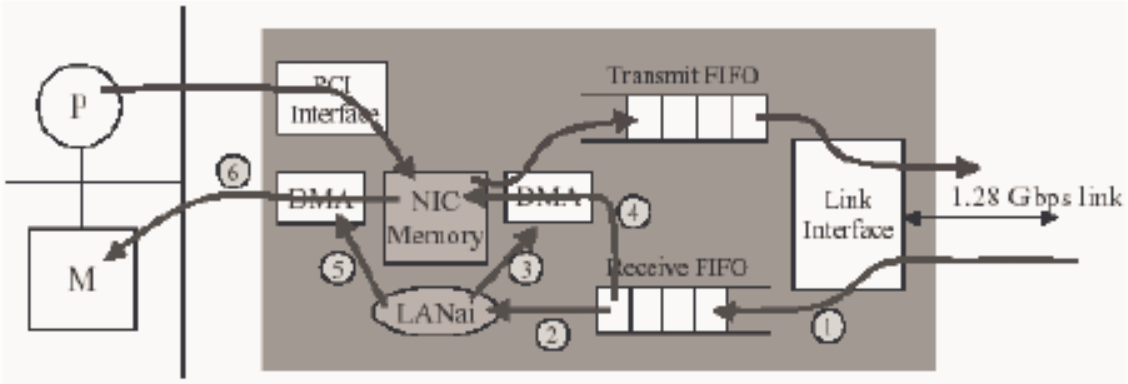
- Myrinet
  - Red diseñada por Myricom
  - Válida para LAN y SAN (System-Area Network)
  - Ancho de banda de 1,28 Gbits/seg, con conmutación wormhole
  - La responsabilidad se reparte entre el procesador del nodo y LANai
    - LANai es un dispositivo de comunicación programable que ofrece un interfaz a Myrinet.
    - LANai se encarga de las interrupciones originadas por los mensajes
    - DMA directa entre cola FIFO y memoria NIC, y entre memoria NIC y memoria del nodo
    - La memoria NIC puede asignarse al espacio lógico de los procesos
    - La latencia aplicación-aplicación es de aprox. 9µs.

Tema 2. Sistemas de comunicación en computadores paralelos

Redes para clusters

- Myrinet





Tema 2. Sistemas de comunicación en computadores paralelos

Redes para clusters

- InfiniBand
  - Nuevo estándar que define un nuevo sistema de interconexión a alta velocidad punto a punto basado en switches.
  - Diseñado para conectar los nodos de procesado y los dispositivos de E/S, para formar una red de área de sistema (SAN)
  - Rompe con el modelo de E/S basada en transacciones locales a través de buses, y apuesta por un modelo basado en el paso remoto de mensajes a través de canales.
  - Arquitectura independiente del sistema operativo y del procesador del equipo.
  - Soportada por un consorcio de las empresas más importantes en el campo: IBM, Sun, HP-Compaq, Intel, Microsoft, Dell, etc
  - Características de la red Infiniband
    - No hay bus E/S
    - Todos los sistemas se interconectan mediante adaptadores HCA o TCA
    - La red permite múltiples transferencias de datos paquetizados
    - Permite RDMA (Remote Memory Access Read or Write)
    - Implica modificaciones en el software del sistema

Tema 2. Sistemas de comunicación en computadores paralelos

Redes para clusters

- InfiniBand