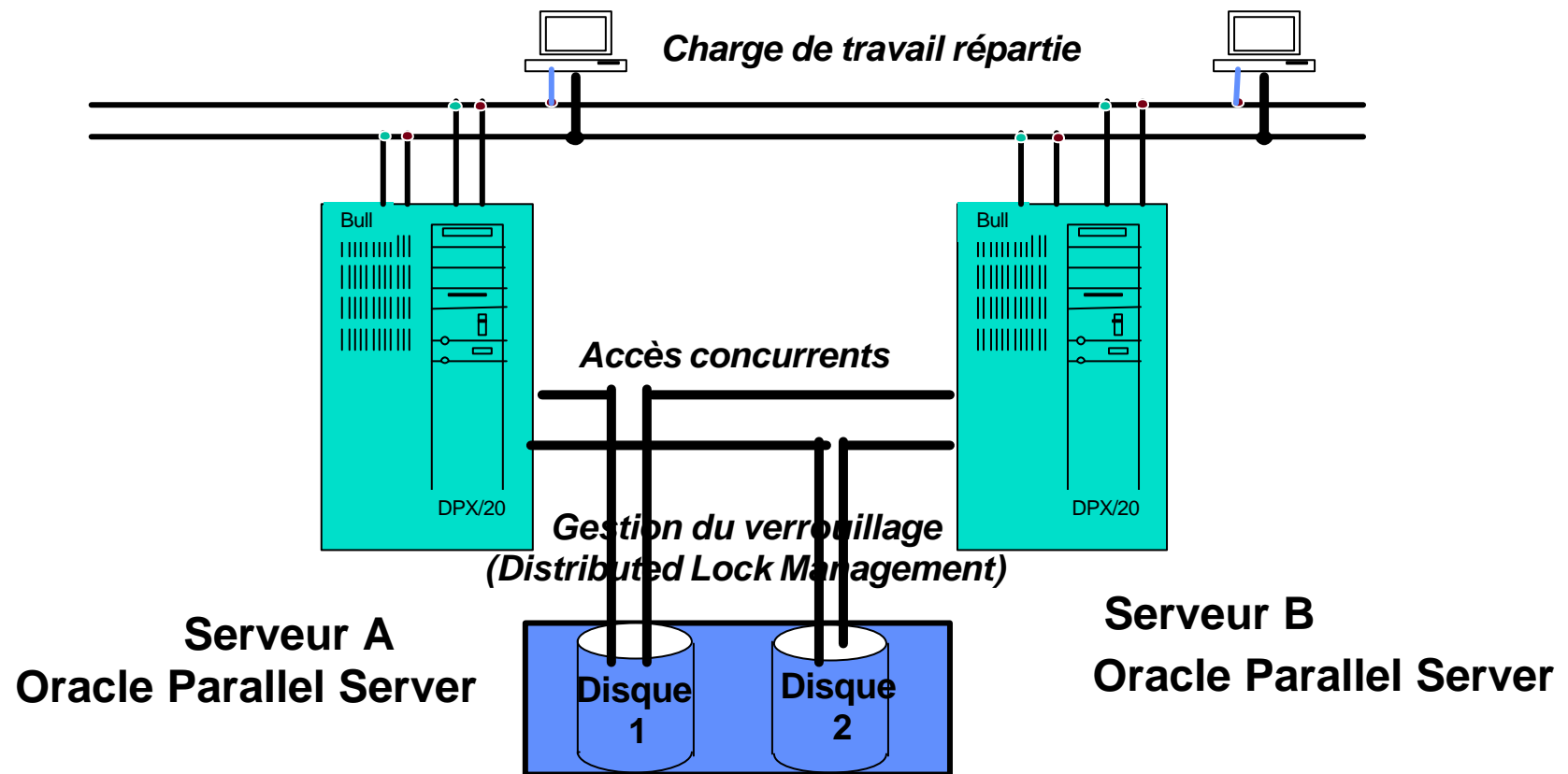


# Couplage lâche (clusters et MPP)

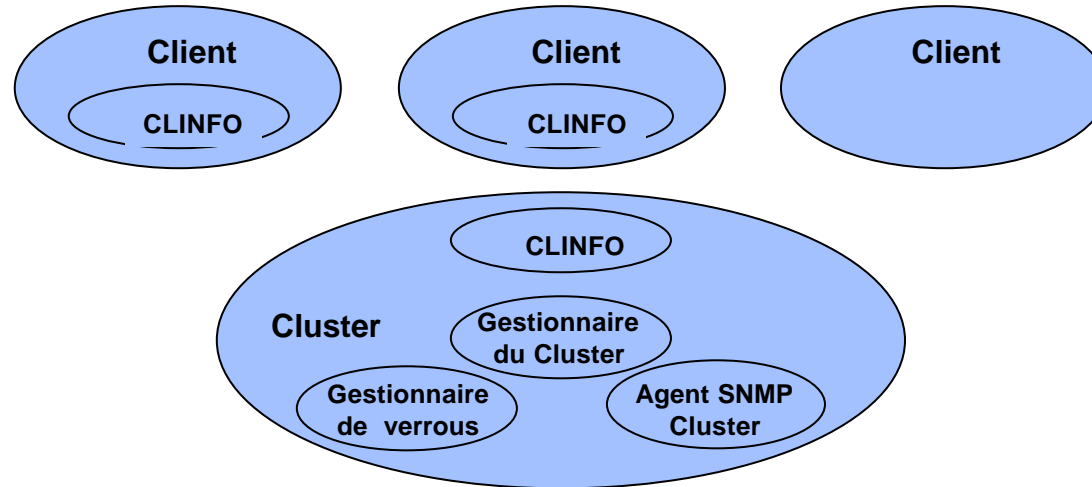
## **Exemples**

# Clusters

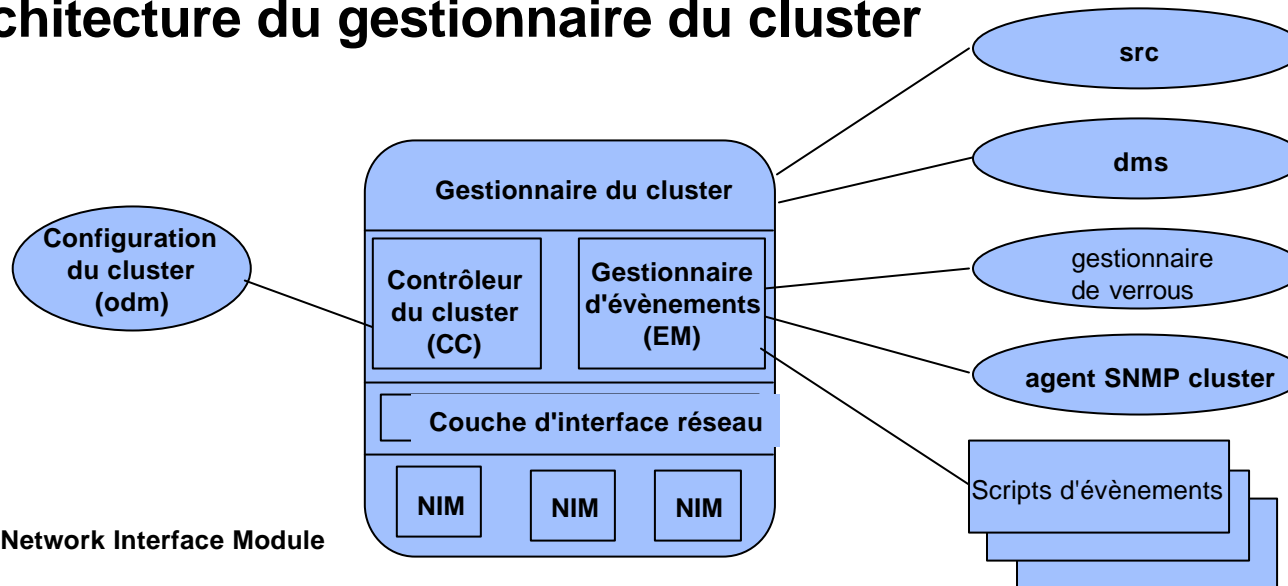
## ■ Cluster RS/6000 HACMP (High Availability - Cluster MultiProcessing)



## ■ Architecture générale HACMP

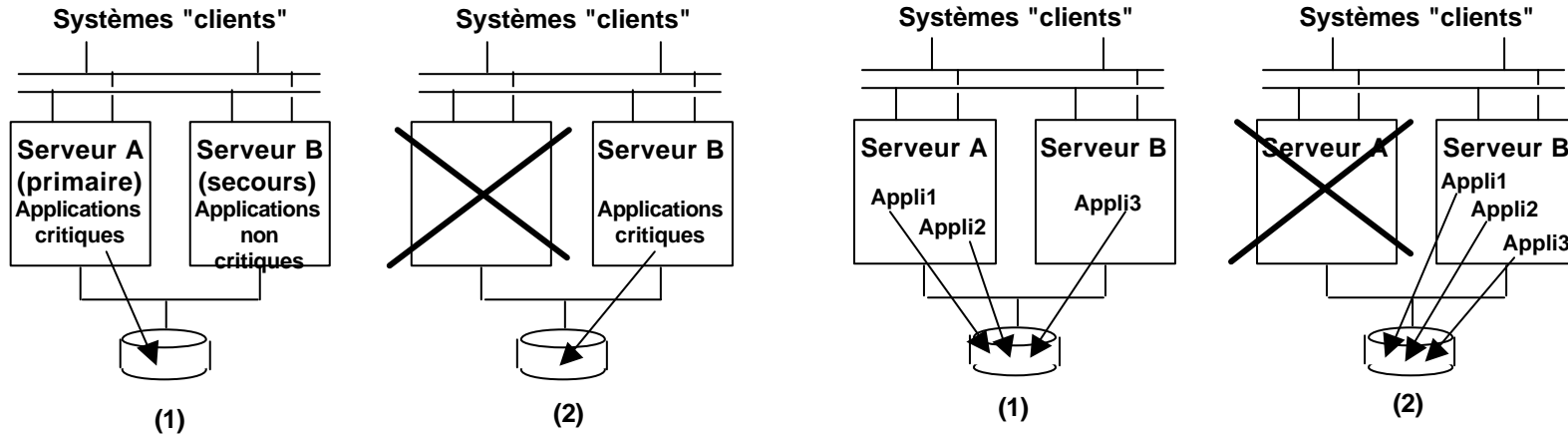


## ■ Architecture du gestionnaire du cluster



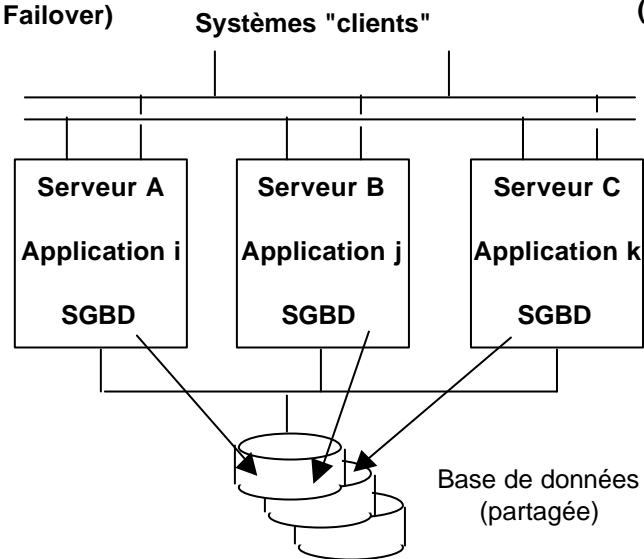
NIM : Network Interface Module

## ■ Modes de fonctionnement



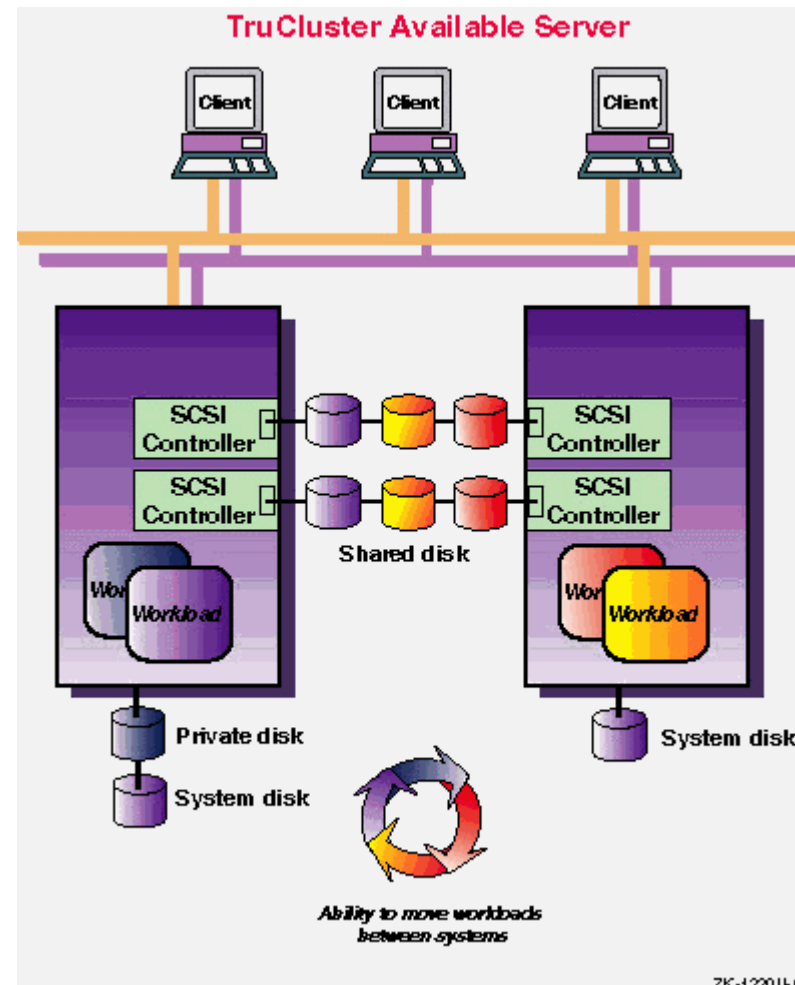
A) - Recouvrement simple  
 (Hot Standby ou Simple Failover)

B) - Recouvrement mutuel  
 (Mutual Takeover)

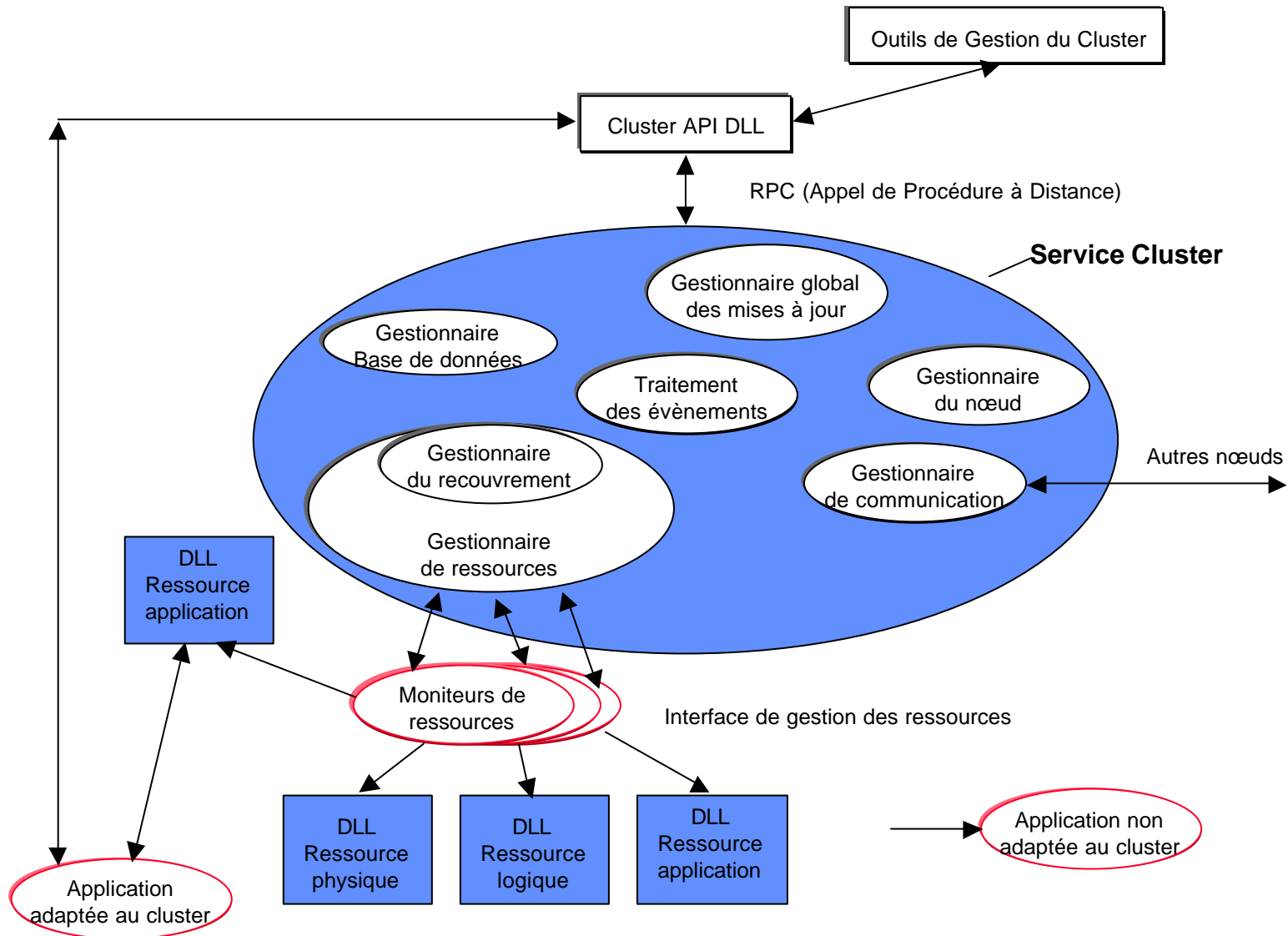


C) - Partage de charge  
 (Cluster MultiProcessing)

## ■ Exemple de configuration TruCluster Software de Compaq



## ■ Architecture Microsoft Cluster Server (MSCS)

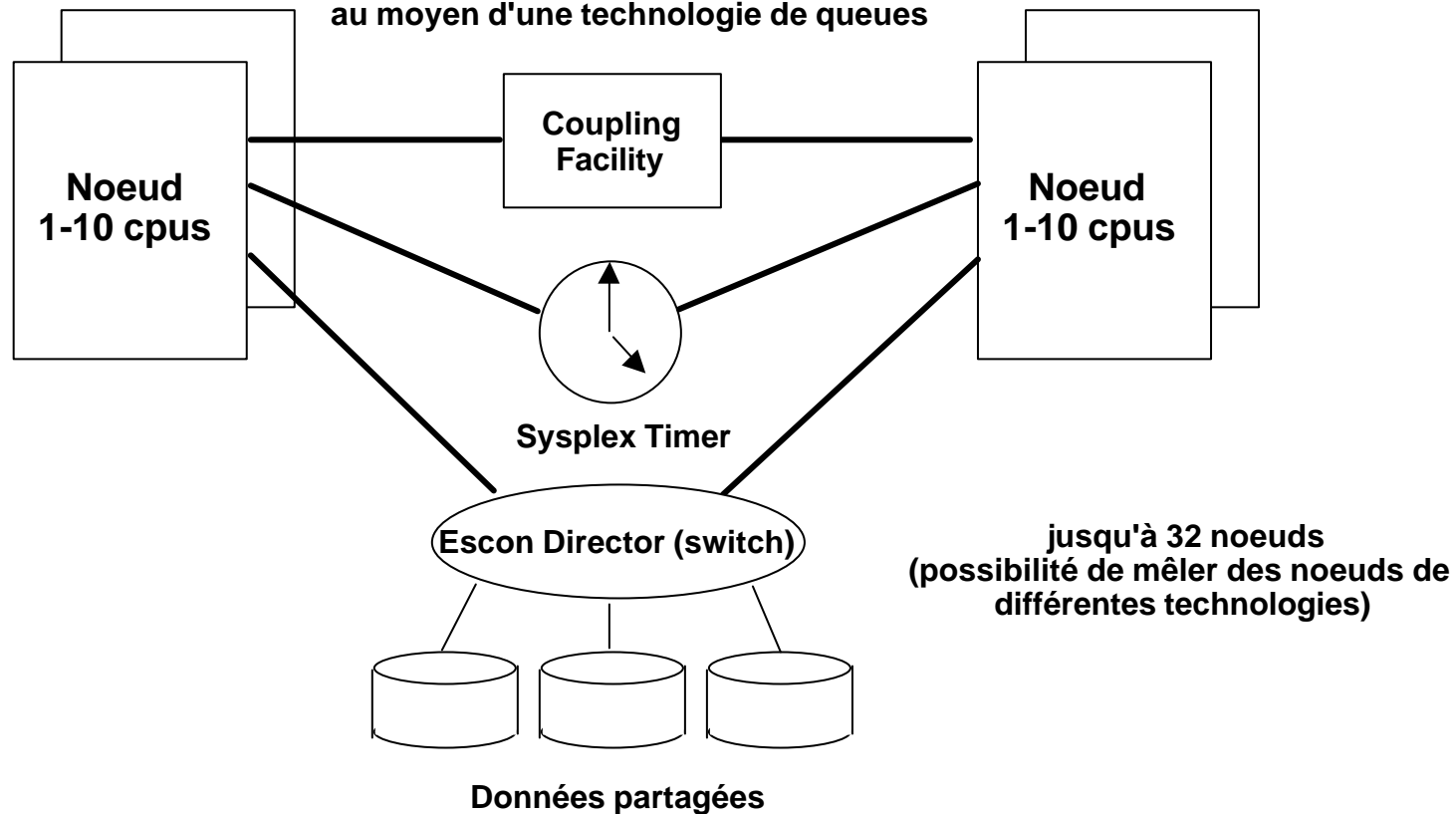


- **Technologie de clustering de serveurs NT - Nom de code Wolfpack MSCS (MicroSoft Cluster Service)**
- **Partenariat avec Compaq, Digital, HP, IBM, Intel, NCR et Tandem "Early Adopters"**
- **Planning**
  - **11/95 - 5/96 Open Process = Disclose interfaces and reviews**
  - **4Q96 SDK Release = Kit de développement et spécifications des APIs à destination des développeurs**
  - **4Q96 Bêta Test = test, par des développeurs, sur les systèmes des 6 "Early Adopters"**
  - **1H97 Phase 1 première version limitée à 2 noeuds (pour les 6 "Early Adopters")**
  - **2H97 Phase 1 Platform and Solution Expansion**
  - **Août 1998 seconde version et extension de la fonctionnalité**
  - **Windows 2000 Data Center**
    - **Jusqu'à 4 noeuds de 4 à 8 processeurs chacun**

## ■ Cluster à mémoire partagée IBM Sysplex

Données partagées au moyen de technologies de verrous et de caches

Distribution de la charge et communication par message  
au moyen d'une technologie de queues



## ■ Fonctionnalité IBM Sysplex

- Éléments de base de la technologie Coupling Facility
  - Cache Structure : mécanisme de cache et d'invalidation
  - List Structure : partage de données organisées en liste (e.g. implémentation de files d'attente, de status,...)
  - Lock Structure

### ■ Partage de données

- IMS
- DB2
- VSAM

#### Overhead

- Partage intensif de données
  - CMOS :  $17\% + 0.5\% \times (N-2)$
  - ECL :  $26\% + 0.5\% \times (N-2)$
- Sans partage  
~3%

### ■ Traitement parallèle

- Batch : BatchPipes/MVS
- OLTP : CICS, IMS TM
- Gestion des travaux : WLM (WorkLoad Manger)
  - Définition des objectifs de performance au niveau du cluster (concept de "service policy")
  - Reporting (Resource Management Facility)

### ■ Administration

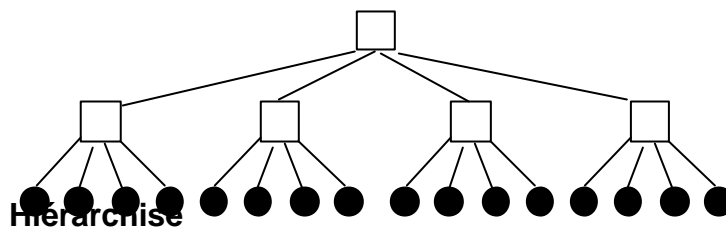
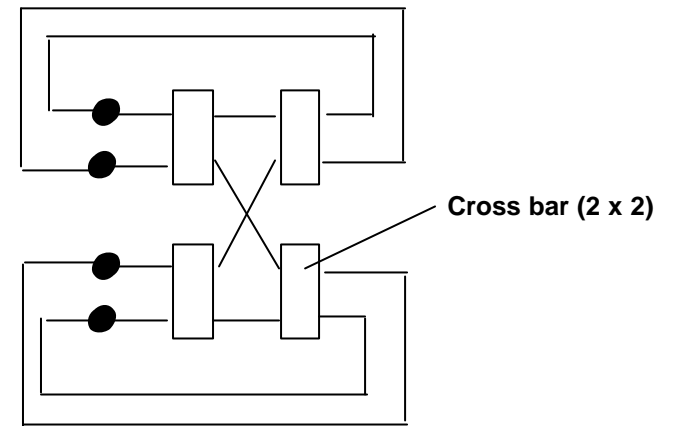
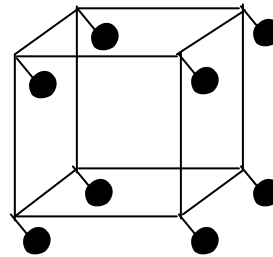
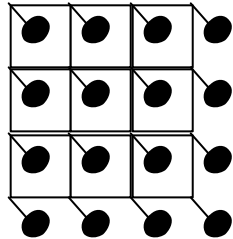
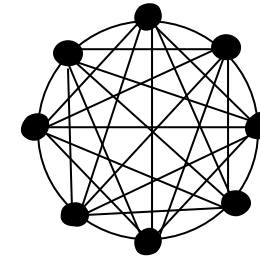
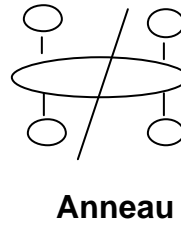
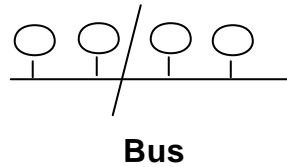
# Machines massivement parallèles (Massively Parallel Processing)

# MPP : Technologie d'interconnexion

- **La performance d'un MPP et sa scalabilité sont liées aux propriétés du réseau d'interconnexion**
- **Caractérisation du réseau d'interconnexion**
  - **Nombre de noeuds supportés**
  - **Bande passante**
    - **Bande passante totale bande passante d'un lien x nombre de liens (représentativité?)**
    - **"Bisection Bandwidth" (plus représentative)**
      - **Pour un réseau symétrique : bande passante observée sur la coupe "en deux" du système**
      - **Pour un réseau dissymétrique : bande passante minimum observée sur l'ensemble des coupes**
  - **Latence (tant au niveau matériel qu'au niveau logiciel)**
  - **Interface matériel/logiciel**
  - **Bloquant ou non-bloquant**
  - **Coût**
  - **Résistance aux défaillances**
  - **Standardisation?**

# MPP : Technologie d'interconnexion (2)

## Quelques topologies de réseaux d'interconnexion



# MPP : Technologie d'interconnexion (3)

## Quelques caractéristiques de réseaux d'interconnexion

Performances et coûts relatifs de divers interconnects pour 64 noeuds d'après [HEN94]

Critères	Bus	Anneau	Grille 2D	Hypercube (6)	Entièrement connecté
Bande passante totale	1	64	112	192	2016
Bisection	1	2	8	32	1024
Ports par switch	n.a.	3	5	7	64
Nombre total de liens	1	128	176	256	2080

[HEN94] John L. Hennessy, David A Patterson «Computer Organization and Design The Hardware Software Interface»  
Morgan Kaufmann San Mateo 1994

Caractéristiques de latence et bisection pour un réseau de type Hypercube de SGI [GAL97]  
(niveau matériel)

Nombre de noeuds	Latence moyenne (ns)	Bisection (GO/s)
8	118	6.4
16	156	12.8
64	274	51.2
256	344	205.0
512	371	410.0

[GAL97] Mike Galles « Spider: A High Speed Network Interconnect »  
IEEE Micro January/February 1997 pp34-39

# MPP : Technologie d'interconnexion (4)

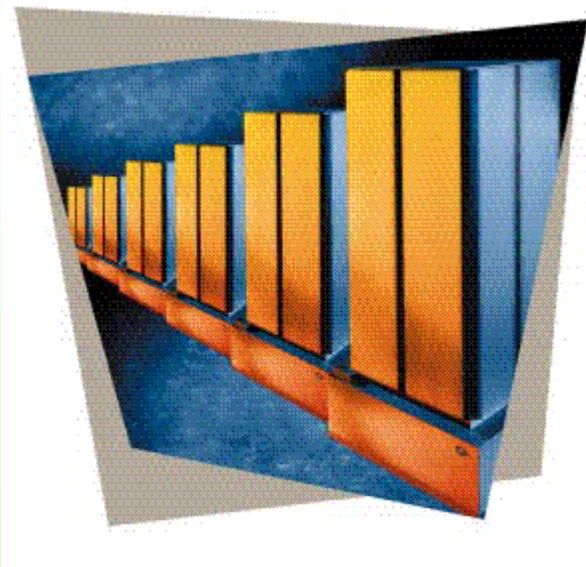
## *Quelques exemples de performance de l'interconnect au niveau "système"*

Comparaison des latences pour un message court d'un noeud à un autre [GIL97]

Environnement	Latence (µs)
Memory Channel DEC (HW)	2.9
Memory Channel DEC(base)	5.4
Memory Channel DEC MPI	6.9
Memory Channel DEC PVM	8.0
Cray T3d MPI	37
IBM SP2 MPI	40
IBM SP2 PVM	54
Intel Paragon	54
Alpha Server @200 UDP/FDDI	82
Alpha Server @300 TCP/IP/FDDI	165
Alpha Server @300 TCP/IP/Enet 10 Mb/s	190
CM5 PVM	190

[GIL97] Richard Gillet, Rochard Kaufmann « Using the Memory Channel Network »  
IEEE Micro January/February 1997 pp19-25

## IBM Scalable POWERparallel Systems SP2



RS/6000	RS/6000
RS/6000	RS/6000
RS/6000	RS/6000
RS/6000	RS/6000
RS/6000	RS/6000
RS/6000	RS/6000
RS/6000	RS/6000
RS/6000	RS/6000
High Perf. Switch	

- Jusqu'à 128 nœuds (8 armoires et 16 nœuds "thin", 8 nœuds "wide" ou 4 nœuds "high" au maximum par armoire) - 512 nœuds sur demande spéciale

- Nœuds fondés sur les systèmes RS/6000, configurés en tiroir et groupés en armoires (rack mount)

- Trois types de nœuds : High (SMP8 x 604 Escala), Wide et Thin (nœuds monoprocesseur)

- Réseau d'interconnexion (High Performance Switch) de type Oméga (cross bar 4 x 4)  
Débit d'un lien 40 Mo/s x 2

- Applications de calcul numérique intensif

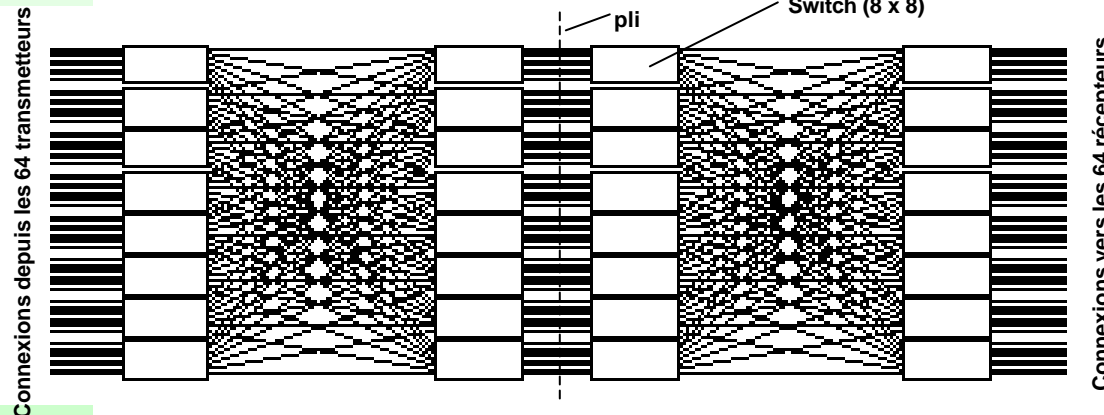
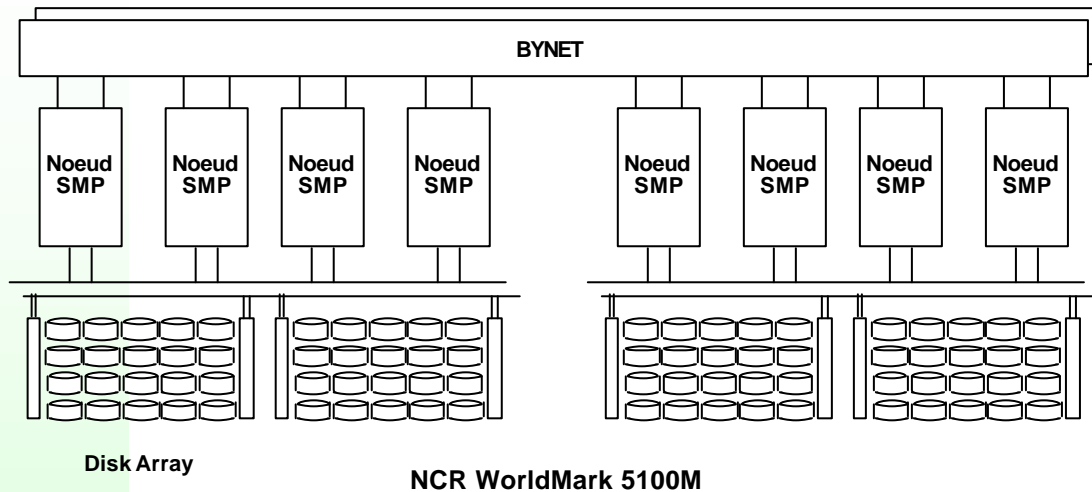
- Serveur de bases de données (Oracle Parallel Server) et DB2 Parallel Edition

- "LAN consolidation"

# NCR/Teradata Worldmark 5200

- **Architecture (matérielle et logicielle) de type Share Nothing fondée sur x86 et un interconnect intelligent (YNET).  
Premier système livré en 1984**
- **Originellement, architecture logicielle spécifique : TOS Teradata Operating System (16 bits) et Teradata DBMS**
- **Systèmes destinés essentiellement aux applications d'aide à la décision et supportant des bases de données > 1 TB**
- **Actuellement, utilisation de solutions standards avec UNIX et le DBMS Teradata (qui est aussi porté sur NT)**
- **Architecture matérielle fondée sur un "building block" de type SMP
  - **nœud quadriprocesseur : 4 Xeon@450Mhz****
- **Supporte jusqu'à 512 nœuds, soit un maximum de 2048 processeurs**
- **Nouvel interconnect : BYNET**

# NCR/Teradata (2)



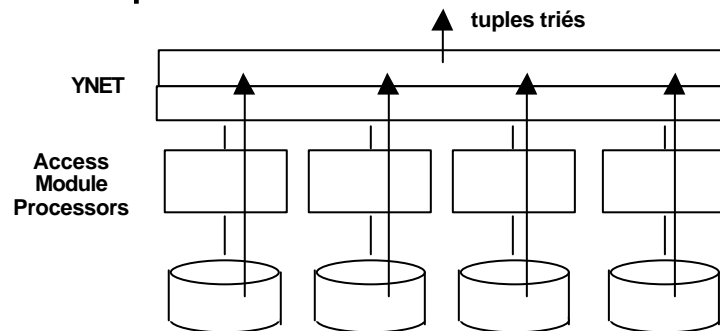
BYNET - Structure d'un BYNET redondant pour une configuration à 64 nœuds  
 Chaque nœuds a deux chemins BYNET vers chacun des autres nœuds du système  
 (Source NCR)

## • BYNET

- Topologie de type BANYAN
- Pour une liaison logique, deux types de liaisons :
  - parallèle 8 bits (10.4 MB/s)  
"Forward Channel"
  - sérielle (10.4 Mb/s)  
"Back Channel"
- Carte contrôleur MCA, 2 liens par carte, 2 contrôleurs par nœud.
- Concept de channel program
- Support "monocast", "broadcast" et "multicast"
- Deux types de protocole :
  - Basic Link Manager (avec acquittement du type "2 Phase Commit")
  - BYNET Low Latency Interface (service type datagramme)
- Bissection =  $20.8 \text{ MB/s} \times \text{nombre\_de\_noeuds}$
- Latence
  - Switch = 673 ns
  - Application/Application = 300  $\mu\text{s}$
- Services de base Message, Channel, Group, Global Semaphore, Merge et Configuration qui permettent l'implémentation de :
  - la distribution des données
  - contrôle des étapes de traitement parallèle
  - status et configuration

## ■ Évolution du réseau d'interconnexion

- YNET réseau arborescent permettant de réaliser le tri des tuples



Le concept de réseau intelligent YNET correspondait à un certain état de la technologie :
 

- processeurs lents
- petites mémoires

- BYNET ne "trie" plus, abstraction pour le parallélisme VPROC (Virtual Process) et tri par VPROC

