le cnam

Clusters et machines massivement parallèles

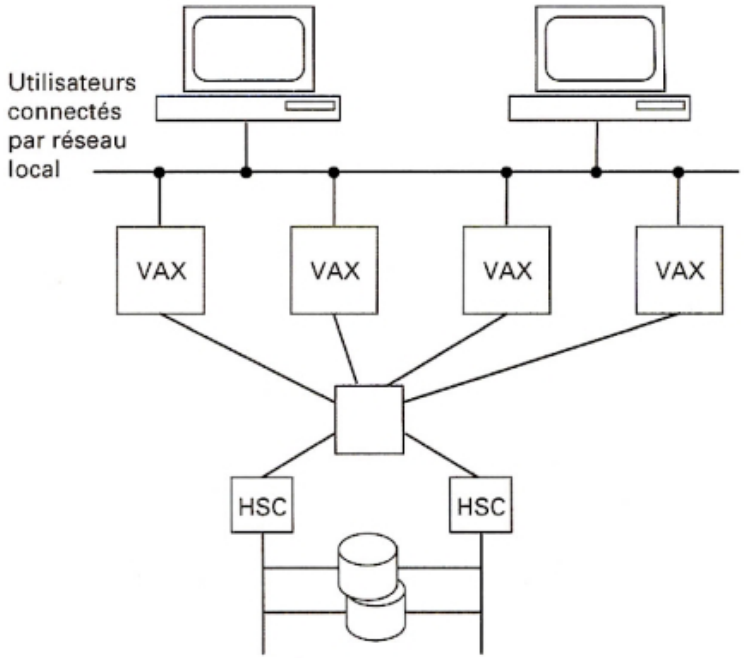
NSY 104

Introduction

- Qu'est ce qu'un *cluster* ?
  - Un cluster peut être défini comme un ensemble limité (de l'ordre de la dizaine)
    - de systèmes informatiques interconnectés
    - qui partagent des ressources de façon transparente
  - Transparence également vis-à-vis de l'utilisateur
  - Chacun de ces systèmes peut être considéré comme un système à part entière :
    - Il dispose d'un ensemble complet de ressources
      - Mémoire
      - Disques
      - processeurs...
  - Les systèmes qui composent un cluster sont appelés nœuds.
  - Les clusters sont aussi désignés comme des systèmes faiblement couplés
    - car ils ne partagent pas de mémoire.

Introduction

- Motivations initiales du développement
  - TANDEM – Guardian (fin 70')
    - Systèmes transactionnels à continuité de service
      - Reprise sur panne
  - DEC – VAX clusters (1983)
    - Augmentation des capacités de calcul sur des systèmes existants



HSC *Hierarchical Storage Controller*

Intérêts

- Le clusturing permet d'obtenir, à moindre coût, des systèmes disponibles et puissants
  - à partir des technologies des réseaux locaux (commutateurs)
  - et de composants du commerce
  - par opposition aux multiprocesseurs (MP)
- Ils sont donc plus aisément extensibles
  - Cout
  - Technologie
- Répondent à des besoins de
  - Puissance de calcul
  - Disponibilité
  - Tolérance aux fautes / pannes

Performances

- Définitions
  - Accélération
    - la capacité à exécuter une tâche déterminée en moins de temps possible
  - Accroissement
    - la capacité à traiter plus de tâches dans une période de temps déterminée

Performances

- Interconnexion se fait par une carte E/S
  - Contrôlée par logiciel
    - Latence
  - Contrairement aux MP
    - Bus interne
    - Contrôlé par le matériel
- Division de la mémoire
  - Un cluster a autant de mémoires que de nœuds
    - Elles sont indépendantes
  - Différent du MP qui peut utiliser toutes les mémoires
- Maintenance facilitée pour le cluster
  - Cette indépendance ente nœuds est un avantage
    - Vis-à-vis du MP
  - Nul besoin d'arrêter tout le cluster pour réparer / remplacer un nœud.

Coûts et adaptations

- L'administration du cluster est plus couteuse (r/ MP)
  - On doit administrer tous les nœuds du cluster
  - Comme autant de machines indépendantes
- La chute des prix des composants utilisés dans les clusters a pourtant favorisé cette architecture, aux dépends des MP
  - Notamment concernant les prix de la mémoire
  - Et ce, malgré le désavantage des mémoires indépendantes des clusters
- On voit apparaître des MP permettant l'exécution d'OS différents sur différentes parties de la machine
- A l'opposé, on voit se former des clusters à partir de machines multiprocesseurs à mémoire partagée
  - Voire se délocaliser le stockage en dehors du cluster

Solution idéale ?

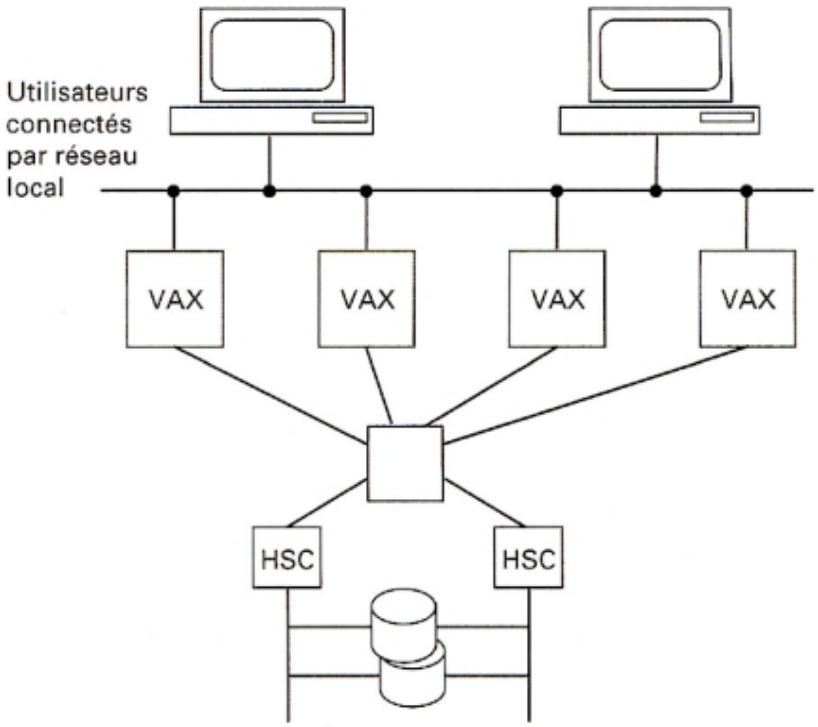
- Le cluster est
  - Peu couteux
  - Extensible
  - Facilement maintenable
- Il répond de manière appropriée aux besoins des fournisseurs de services en ligne
  - car ils gèrent de gros volumes de tâches indépendantes
    - Moteurs de recherche
    - Messageries

Architecture des 500 plus gros centres de calcul du monde (11/2010, http://top500.org)

Architecture	Count	Share	Rmax Sum (GF)	Rmax Peak (GF)	Processor Sum
Other	**2**	**0.40 %**	**94970**	**112947**	**17648**
MPP	**84**	**16.80 %**	**17936809**	**23875912**	**2851827**
Cluster	**414**	**82.80 %**	**25641314**	**40666452**	**3602852**
Totals	**500**	**100%**	**43673092.54**	**64655310.70**	**6472327**

DEC VAX Cluster

- Ressources clusterisées
  - Le système de fichiers
  - les files de travaux de traitement par lot (batch)
  - les files d'attente des travaux d'impression
  - le mécanisme de synchronisation
    - gestionnaire de verrous distribué
    - (*Distributed Lock Manager, DLM*)
  - Système d'interconnexion spécifique



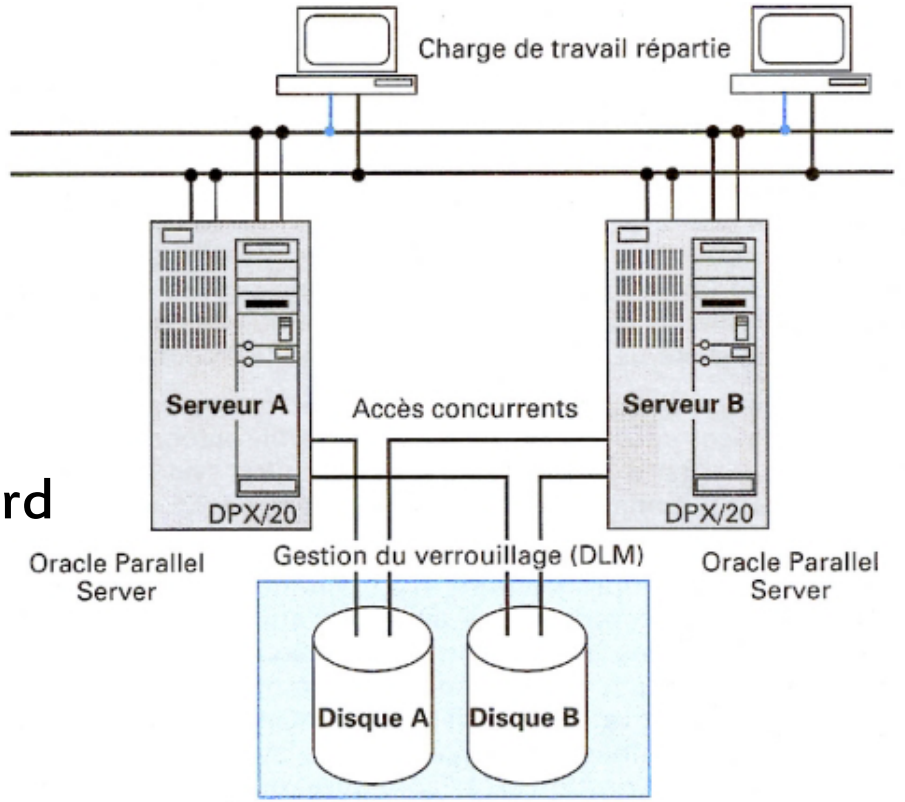
HSC *Hierarchical Storage Controller*

Bull Power Cluster

- Basé sur le logiciel HACMP
  - *High Availability Cluster Multi-Processing*
- Et des systèmes SMP Escala

- Ressources clusterisées
  - les unités de disques
  - les connexions réseau
  - le DLM

- Réseau d'interconnexion standard
  - Ethernet
  - FDDI



DLM *Distributed Lock Management*

Bull Power Cluster

- Ajouter Routage

Machines massivement parallèles (MPP)

- Une machine massivement parallèle (*Massively Parallel Processing*) possède un ensemble important (plusieurs centaines) de systèmes (nœuds) reliés par un réseau d'interconnexion spécialisé.

- Chaque nœud dispose de ses propres ressources
  - Processeurs
  - Mémoire
  - Contrôleurs E/S
  - Système d'exploitation.

Machines massivement parallèles (MPP)

- Doivent permettre de fournir une bande passante modulable
  - Idéalement, qui croit linéairement avec le nombre de nœuds
- Et une latence faible
  - Qui, idéalement, ne dépendrait pas du nombre de nœuds
- Leur construction physique spécialisée est adaptée à l'intégration d'un nombre important de nœuds
- Au début des MPP, les nœuds étaient monoprocesseurs
- Usages
  - Calcul intensif
  - Aide à la décision

Machines massivement parallèles (MPP)

- Les différences avec les clusters sont :
  - Le nombre maximal de nœuds,
    - une dizaine pour un cluster
    - plusieurs centaines pour les MPP
  - L'organisation physique des systèmes
    - conçue pour supporter un grand nombre de nœuds
    - et en faciliter l'ajout.
  - Le réseau d'interconnexion spécifique à haute performance
    - pour les clusters il s'agit plutôt de FDDI ou Ethernet
  - Les nœuds peuvent être des SMP
    - ils sont alors qualifiés de nœuds multi – processeurs.
- Contrainte majeure des MPP
  - Les applications sont développés spécifiquement pour ce type d'architecture.

Machines massivement parallèles (MPP)

- Concurrents des grands ordinateurs vectoriels
- Les SGBD ont été adaptés au traitement parallèle pour les grandes bases de données
- Toujours utilisés dans les systèmes d'aide à la décision

- Il est impératif que le réseau d'interconnexion soit efficace car une application MPP comprend :
  - Des phases de communication
    - données pour les processus qui s'exécutent en parallèle.
  - Des phases de calcul.

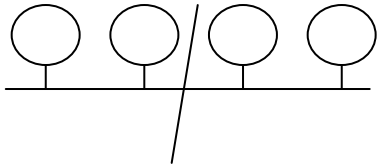
- La performance du réseau est donc fondamentale pour la performance des applications.

MPP et réseaux d'interconnexion

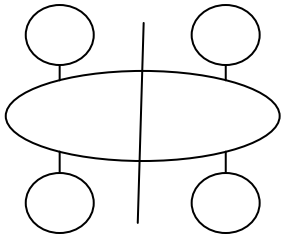
- Les performances sont liées à
  - La topologie du réseau.
  - Le nombre maximal de nœuds supportés.
  - La latence et la bande passante
  - La simplicité et la généralité de l'interface matériel – logiciel.
  - Le caractère bloquant ou non bloquant du réseau
  - Le coût
  - La résistance aux défaillances
    - en général les réseaux sont redondants.
    - En fonctionnement normal, deux réseaux sont utilisés, chacun s'occupe de 50% de la charge
    - En cas de défaillance de l'un des réseaux, la totalité du trafic est assurée par le réseau valide.
  - La conformité à un standard ou la possibilité d'en devenir un.

MPP et réseaux d'interconnexion

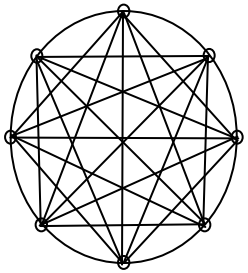
- Différentes topologies



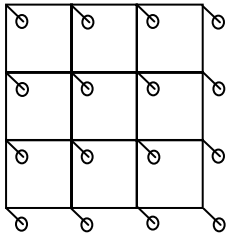
BUS



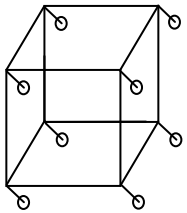
ANNEAU



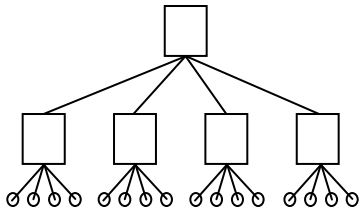
Complètement connecté



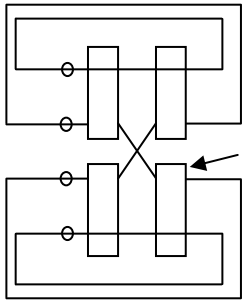
Grille 2D



Grille 3D



Arborescente



Réseau Oméga

MPP et réseaux d'interconnexion

- Caractéristiques
  - La latence
    - Le temps nécessaire à l'acheminement d'un message
      - depuis l'espace d'adressage d'un processus
      - jusqu'à l'espace d'un autre processus.
  - La bande passante
    - Le volume d'information transmis par unité de temps
      - Nombre de bits
    - La bande passante totale
      - il s'agit de la bande passante d'un lien multipliée par le nombre de liens existant dans le réseau d'interconnexion.
      - La bande passante totale suppose que les nœuds exploitent en totalité les différents liens et qu'il n'y ait aucun conflit sur ces liens.
    - La bande passante réellement utilisable
      - c'est la moyenne passant par les différents nœuds
      - la notion de *bissection bandswitch* est utilisée pour définir cette moyenne.

MPP et réseaux d'interconnexion

- La bissection bandswitch :
  - Pour un réseau symétrique, c'est la bande passante observée sur une coupe en deux du réseau d'interconnexion.
  - Pour un réseau dissymétrique, c'est la bande passante minimale observée sur l'ensemble de coupes du réseau.

- Un réseau d'interconnexion parfait
  - Une latence constante (indépendante du nombre de nœuds).
    - Le temps nécessaire à l'acheminement d'un message entre deux nœuds quelconques est indépendant du nombre de nœuds du système.
  - Une bissection croissant linéairement avec le nombre de nœuds.
    - L'ajout d'un nœud dans le système apporte alors une contribution constante à la bande passante.

- Pour les clusters, le réseau est souvent fondé sur les technologies de réseau local,
  - la latence augmente avec le nombre de nœuds car il y a contention
  - le débit global est constant.

- Dans le cas des MPP, le réseau d'interconnexion cherche à se rapprocher des caractéristiques idéales
  - latence constante
  - débit dépendant linéairement du nombre de nœuds.

MPP et réseaux d'interconnexion

- Un réseau complètement connecté possède une topologie telle que tout nœud possède un lien dédié avec tous les autres nœuds du système.

- Le tableau met en évidence la différence importante qui existe entre
  - la bande passante totale
  - la bissection.

- Le choix d'une topologie résulte d'un compromis entre sa bande passante et son coût.

Critères		Bus	Anneau	Grille 2D	Hypercube	Complètement connecté
Bande Passante	Bande Passante Totale	1	64	112	192	2016
	Bissection	1	2	8	32	1024
Coût	Port par Switch	Pas applicable	3	5	7	64
	Nombre total de liens	1	128	176	256	2080

MPP et réseaux d'interconnexion

Caractéristiques du réseau d'interconnexion Spider de SGI (topologie HyperCube)

- Ce réseau présente les caractéristiques idéales en termes
  - de bande passante
  - de croissance linéaire de la bissection en fonction du nombre de nœuds.
- La latence augmente de façon sensible avec le nombre de nœuds connectés.

- Pour l'essentiel, la latence est du au logiciel et non au matériel.

Nombre de nœuds	Latence moyenne (ns)	Bissection Go/s
8	118	6.4
16	156	12.8
64	274	51.2
256	344	205.5
512	371	410.0

Synthèse

Caractéristiques visibles par l'utilisateur	SMP	Cluster	MPP
Accélération/accroissement	**Accroissement** et accélération	Accroissement puis accélération	**Accélération** puis accroissement
Haute disponibilité	Typiquement non	Oui	Possible (mais n'est généralement pas un objectif)
Configurations importantes (100 processeurs et au-delà)	N'existe pas sur technologie de commodité Nécessite des technologies spécifiques	Possible (dépend des propriétés de l'interconnect)	Oui
Image système unique	Complète (par définition)	Limitée aux ressources clusterisées	Limitée
Partage	Tout y compris le système d'exploitation et la mémoire cohérente	N'existe pas ou limité (typiquement les disques et les connexions réseau)	Pas ou limité
Programmation	Processus unique ou processus multiples (threads)	Programmation spécifique requise pour exploiter le parallélisme	Programmation spécifique requise pour exploiter le parallélisme (élément plus crucial que pour les clusters)
Flexibilité pour l'intégration de nœuds de générations différentes dans le système	Pas applicable	Oui	Typiquement non
Types d'application	Transactionnel Aide à la décision	Aide à la décision Solutions à haute disponibilité	Applications de calcul intensif Aide à la décision Solutions à continuité de service
Exemples	Bull Escala DEC AlphaServer 8400 HP T500/K400 Sequent Symmetry SGI	BullPowerCluster DEC Clusters IBM HACMP Pyramid/SNI Reliant Sequent ptx/Clusters Tandem Himalaya	IBM SP2 ICL Gold Rush Parsytec Pyramid/SNI RM1000 Unisys Opus

René J. CHEVANCE