DESS Images et Réseaux Option Réseaux

# LE CLUSTERING DANS LES ENTREPRISES

*Jean Yves Repetti*
*Laurent Cardona*

# Le clustering dans les entreprises

# 1. Introduction aux clusters

## 1) Histoire des clusters

Les clusters sont apparus au moment de la faillite du modèle de supercalculateur et alors que les processeurs devenaient de plus en plus rapides. L'enjeu a consisté, à partir de la fin des années 80, à mettre en place ce que les ingénieurs appelaient alors un "multi-ordinateur" (en 1987, l'université du Mississipi a commencé à travailler sur un cluster basé sur le Sun 4/110). C'est toutefois le projet Beowulf - l'utilisation d'un système d'exploitation Linux sur des PC communs - qui a véritablement lancé l'intérêt pour les clusters, ces grappes d'ordinateurs qui fournissent en commun un travail de calcul en parallèle sur un seul problème complexe. Le coût et la modularité d'un cluster le rendent moins onéreux qu'un superordinateur.

Marc Geoffroy

## 2) Problématique

Par définition, la productivité au sein d'une structure d'entreprise ou autre structure commerciale constitue une mesure de l'utilisation efficace des facteurs de production, c'est-à-dire de l'ensemble des moyens techniques, financiers et humains dont dispose cette entité.
Cette productivité doit être optimale et surtout ininterrompue, en effet l'arrêt même momentané d'un module du système d'informations peut paralyser le bon fonctionnement de l'entreprise pendant la période de remise en production du maillon manquant.

*Il suffit de prendre l'exemple d'un serveur de messagerie qui ne fonctionne plus pendant une demi-journée suite à un disque dur défectueux, le temps de le remplacer et de faire la restauration ; le service commercial ainsi que la direction ne peuvent plus répondre aux appels d'offres, le service clients ne peut pas faire le suivi des réclamations, le service technique ne peut plus passer de commandes. La structure est paralysée pour tous les échanges de courrier électronique, qui représente un pourcentage conséquent de la gestion des activités au sein d'une structure commerciale. Les résultats sont radicaux puisque l'entreprise cumule des contrats perdus, des bénéfices en moins, des heures de travail perdues, des sanctions pour l'équipe de commerciaux et pour le service informatique...*
Pour éviter ce genre de scénario catastrophe certains architectes ou administrateurs de systèmes d'informations décident d'implémenter un service de cluster sur les serveurs hébergeant les applications critiques : serveur de messagerie, serveur

ERP, serveur Web commerce électronique, serveur de base de données, serveur de fichiers ou autres.

# 3) Terminologie

**La technologie de clustering permet d'avoir une haute disponibilité des ressources publiées**. On utilise cette technologie pour avoir une disponibilité et stabilité des ressources proche de 100 %. Tolérance zéro pour les pannes matérielles ou logicielles. Il y a également une répartition des charges entre les nœuds d'un cluster.
**Un serveur de cluster est un groupe de serveurs gérant des ressources stockées sur des disques partagés.** Les nœuds et les disques sont connectés par un bus de liaison (SCSI ou Fibre Channel).
**Un serveur dans le cluster est appelé nœud dit node en anglais.**
**Les données publiques sont appelées ressources**, chaque disque du bus partagé représente un groupe de ressources ; pour publier un groupe de ressources accessible par les clients externes, il est nécessaire de créer un serveur virtuel en lui adressant une adresse IP virtuelle et un nom d'hôte.
Lorsqu'un client externe se connecte pour faire une requête sur les données, celle-ci transite par le serveur virtuel, qui fait office de « passerelle » entre les nœuds connecté aux disques partagés du cluster et le client, ainsi l'architecture du cluster est transparente du côté client. La connexion à un serveur virtuel se fait de manière tout à fait classique, par adresse IP ou nom d'hôte.
Par défaut chaque groupe de ressources est attribué à un nœud.
Dans le cas où le nœud a une défaillance quelconque, l'autre nœud prend en charge les groupes de ressources de son homologue, et répond aux requêtes distantes. C'est la phase de basculement entre les 2 nœuds, appelé **failover**, en conséquent la mise en place d'un cluster permet d'avoir une disponibilité des ressources proche de 100%.

**Haute disponibilité (Availability)** des ressources sur le cluster, celles-ci sont garanties disponibles à 99,9 % du temps.
Dans le cas où un des nœuds ne pourrait plus fournir des réponses aux requêtes des clients, alors les autres nœuds du cluster prennent le relais.
Ainsi la communication avec les clients et l'application hébergée ou autres ressources sur le cluster ne subit pas d'interruption ou une très courte interruption.

**Adaptabilité : (Scalability) :** il est possible d'ajouter un à plusieurs nœuds, ou d'ajouter des ressources physiques (disques, processeurs, mémoire vive) à un nœud du cluster. En effet, il est possible que de part les trop nombreuses requêtes sur le serveur que celui-ci soit en saturation au niveau de la charge processeur, mémoire ou autre, dans quel cas il est nécessaire d'ajouter des éléments, voir un autre nœud.

**Évolutivité** : Lorsque la charge totale excède les capacités des systèmes du cluster, d'autres systèmes peuvent lui être ajoutés. En architecture multiprocesseur, pour étendre les capacités du système, il faut dès le départ opter pour des serveurs haut de gamme coûteux autorisant l'ajout d'autres processeurs, de lecteurs et de la mémoire supplémentaires.

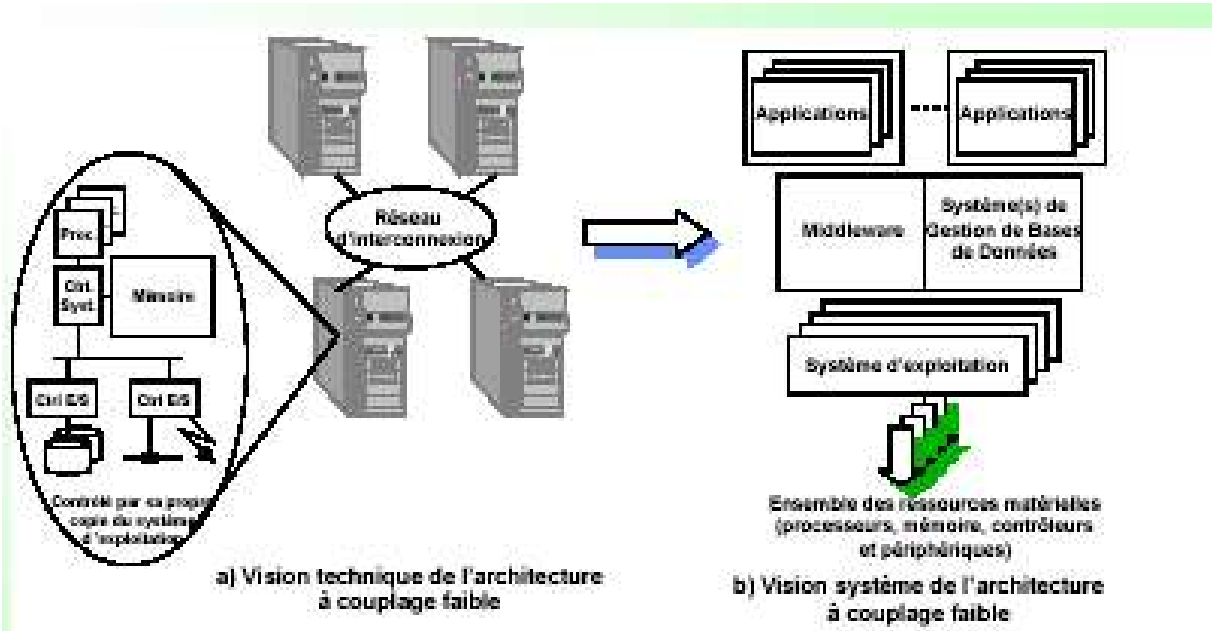
Avec la technologie du clustering, des systèmes monoprocesseurs standards plus petits (et moins onéreux) peuvent être ajoutés graduellement, afin de répondre à l'augmentation des besoins en puissance de traitement.

# 4) Technologie

Sur le plan technique, le clustering consiste à mettre en grappe des serveurs qui partagent des périphériques communs en se répartissant la charge du traitement. Les unités de stockage doivent être commune, et il faut un bus de communication entre les différents serveurs.

Un cluster peut se limiter à une grappe d'ordinateurs personnels standards interconnectés par Ethernet.
Un cluster peut consister en systèmes SMP puissants interconnectés via un bus de communications et d'E/S à hautes performances.
Dans un cluster, la puissance de calcul globale peut être augmentée graduellement en ajoutant un autre système standard. Le cluster donnera à l'application client l'illusion de se trouver face à un seul serveur ou à un seul système, même si véritablement il y en a plusieurs.

## Le couplage lâche



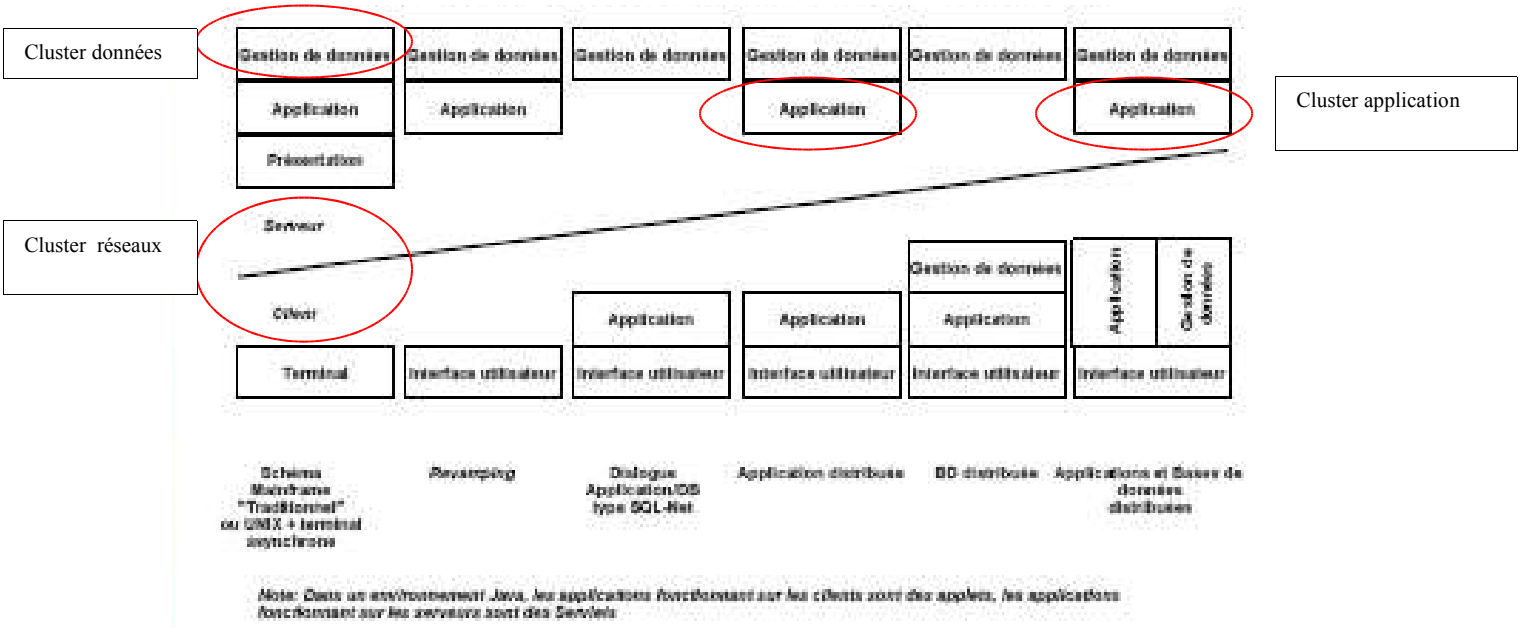

a) Vision technique de l'architecture à couplage faible

b) Vision système de l'architecture à couplage faible

- Le système est constitué par l'interconnexion, au moyen d'une technologie de réseau local rapide, d'un certain nombre de systèmes indépendants (appelés nœuds) , chacun de ces systèmes possède ses propres ressources (processeurs, mémoire, entrées-sorties) et fonctionne sous le contrôle de sa propre copie du système d'exploitation.
- C'est un couplage lâche car les nœuds ne partagent pas de mémoire (cas général).
- Différences par rapport à un système distribué : homogénéité des différents nœuds (fournisseur et système d'exploitation), proximité géographique et image système unique pour certaines ressources.
- Le mode naturel de communication des applications est le passage de messages.

# 2. Les différentes composantes du clustering

## 1) Les différentes architectures client serveur

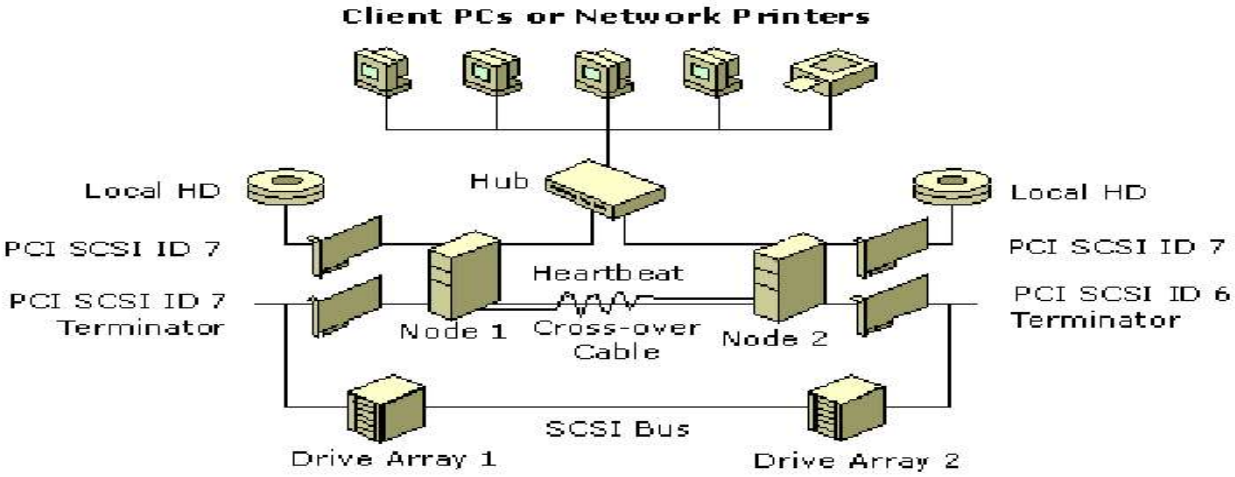



Par rapport aux différentes architectures client serveur répertoriées par le Gartner Group on s'aperçoit que les composantes suivantes peuvent être clusterisées :

- Data (technologies RAID)
- Operating System
- Application: attention le clustering n'est pas tolérance de pannes, les applications doivent être modifiées pour tenir compte de l'architecture cluster
- Network

# 3. Architecture materielle

## 1) Architecture globale




Un cluster est constitué de plusieurs nœuds. Ce sont des serveurs classiques ne nécessitant aucune particularité matérielle, malgré tout il est fortement conseillé d'estimer la charge des requêtes clients pour déterminer la configuration optimale, processeur(s) puissant(s), carte réseau, RAM, carte mère.
*Pour une plus grande stabilité du système il est nécessaire de posséder des configurations identiques sur les 2 nœuds à savoir même processeur(s), mémoire quantité RAM, accès disque SCSI ou IDE 7200tr/min ou 10000 tr/min, ainsi en cas de basculement les applications seront prises en charge de la même façon, et n'affecteront pas les temps entrée/sortie sur les ressources.*
Les applications serveur sont installées de façons identiques sur les deux nœuds du cluster.

## 2) Architecture disques

Les disques de cluster sont des disques durs partagés, chaque nœud peut accéder aux disques via le bus partagé.
Le stockage de toutes les ressources publiables, fichiers de données, files d'impression, applications, ressources, et services se font sur les disques partagés.
Il est nécessaire de partager les disques sur un bus, il y a deux méthodes d'implémentation pour le partage des disques sur un bus, la technologie **SCSI** et la technologie **Fibre Channel** sur un système SAN(Storage Area Network).

# 3) La chaîne SCSI

BOITIER SCSI EXTERNE
Modèle: DuraStor 6200SR Constructeur: Adaptec

La technologie SCSI est une technologie qui peut supporter jusqu'à 15 périphériques par contrôleur. Il est possible de connecter les périphériques de stockage à l'extérieur ou à l'intérieur du serveur pour constituer la chaîne de périphérique SCSI, il est conseillé dans le cas du clustering de connecter les disques partagés à l'extérieur dans un boîtier spécifique. (Voir photo)

Les avantages :

- Rapidité de transfert sur le bus pour les E/S
- Coût moins élevé par rapport à la fibre optique

Les inconvénients:

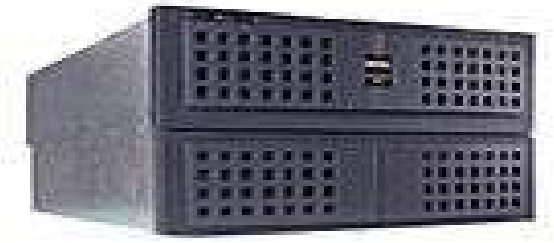
- Limitation à 2 nœuds par cluster
- Limitation par la distance des nœuds, une nappe SCSI est fragile et sa longueur doit être limité.

## 4) Fibre Channel

CONNECTIQUE FIBRE CHANNEL

La fibre permet de diffuser des signaux optique sur une très longue distance mais surtout à une vitesse de 200MB/sec et 400MB/sec en full duplex. Les signaux lumineux sont traduits en bits à trés grande vitesse par les cartes possédant un contrôleur de fibre optique.
La fibre channel est totalement insensible aux rayonnements électromagnétiques, supprimant ainsi tout problème de parasites.
La fibre channel a été conçue pour différentes technologies, les technologies ATM, SCSI, IP, IEEE 802.2 et autres sont supportées.
L'atténuation du signal est inférieure à celle d'un conducteur électrique. L'accès est supérieur à ce que l'on peut obtenir avec une nappe SCSI
Dans ce cas de figure les disques sont stockés dans un SAN (Storage Area Network), périphérique de stockage externe interfacé avec de la fibre optique.

BOITIER SAN - FIBRE CHANNEL
Modèle: FC4500 Stockage Fibre Channel - Constructeur: DELL



Les avantages :

- Une vitesse de communication supérieure entre les serveurs
- Permet de faire évoluer la montée en charge de son réseau
- 2, 3 ou 4 nœuds dans le cluster
- Pas de limitation de distance entre les nœuds et les disques partagés.

Les inconvénients:

- Coût élevé
- Déploiement technique complexe
- Coût de maintenance TCO élevé.
- Tous les nœuds doivent être équipés d'une carte avec un contrôleur de fibre optique, son coût n'est pas négligeable.

# 4. Architecture logicielle

# 1) Disques

Le quorum

Le quorum est un disque dans lequel est stocké toutes les informations concernant le paramétrage et la configuration du cluster à savoir les adresses IP des serveurs virtuels, leurs noms réseaux, les groupes de ressources, les fichiers log, retraçant les événements du cluster et autres informations portant sur la configuration du cluster.

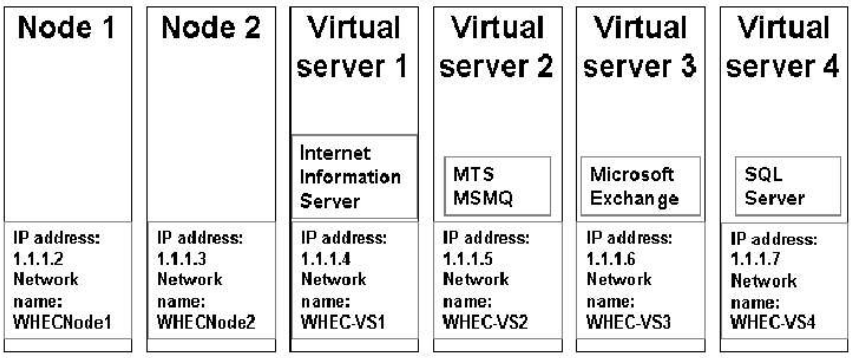
Le quorum est considéré comme le cerveau du cluster.

Il maintient une cohérence et un équilibre dans le cluster pour tous les nœuds. Il manage les données et joue le rôle d'arbitre au sein du cluster pour déterminer quel nœud contrôle le cluster à un instant T et il gère également l'attribution des groupes de ressources.

Sans quorum le cluster ne peut pas fonctionner et il est donc impératif de faire la sauvegarde du quorum régulièrement. Dans le cas où une sauvegarde du quorum est défaillante alors le cluster doit être entièrement restauré.

Le quorum est situé sur le bus partagé avec les autres disques partagés. Il est recommandé d'installer le quorum sur une partition de disques d'au moins 500 méga-octets.

# 2) Les serveurs virtuels




Chaque groupe de ressources est publié sur un serveur virtuel, lequel est accessible par les clients externes.

Le serveur virtuel est associé à une adresse IP et nom réseau unique.

Les clients accèdent au serveur virtuel de façon classique chemin UNC ou adresse IP. Dans le cas où un nœud est hors ligne, pour quelques raisons que ce soit, le basculement vers l'autre nœud du cluster est obsolète pour les clients extérieurs, et l'adresse du serveur virtuel et son accès n'est pas affecté, les utilisateurs continueront à se connecter aux ressources partagées par le même chemin, les clients ne sont en aucun cas en mesure de déterminer quel nœud répond à leur requête, car le chemin d'accès aux ressources reste inchangé. Tolérance de panne optimale.

# 3) Réseaux

Il y a deux types d'interfaces réseau sur un cluster, une interface connectée au réseau public et une interface connectée au réseau privé.

***Le réseau public***

Le réseau public est dédié à la communication entre les clients distants et le cluster. Un réseau public ne peut pas faire de communication nœud à nœud.
Dans ce réseau public tout client à la possibilité de se connecter à un serveur virtuel et d'utiliser ses ressources partagées.

***Le réseau privé***

Les nœuds de cluster ont un besoin permanent d'être en communication pour savoir si tous les nœuds sont en ligne.
Ce processus se fait via le réseau privé; le réseau privé est inexistant pour les clients distants. Il est implémenté pour des besoins techniques dans le cluster. Sur ce réseau privé transite des battements de cœur (heartbeat), il s'agit de datagrammes UDP envoyés d'un nœud à un autre pour savoir si l'un et l'autre sont en ligne.
Le cluster ne peut en aucun cas utiliser ce réseau pour un communication vers les clients distants.

***Le réseau mixte***

Il existe une autre configuration de réseau qui permet d'utiliser un réseau privé avec un réseau public, les trames UDP heartbeat transite sur le réseau public.
Les IP privé et public doivent être implémenter avec un masque de sous réseau identique. Cette implémentation n'est pas recommandée.

# 4) Répartition par éléments actifs du réseau

Lorsqu'il y a besoin, la machine d'entrée sur le réseau sera appelée Node Server (nœud serveur). Dans ce cas, cette machine se verra attribuer la charge de diviser la ou les tâches à travers tous les nodes du cluster en prenant garde bien sur à ne pas surcharger la machine réceptrice.
L'interconnexion de cet ensemble de nodes et de node server est appelé un Network Of Workstation (NOW, réseaux de stations de travail). En fait il s'agit de connecter tous les nodes du cluster par l'intermédiaire d'un élément fédérateur de niveau 2 comme un hub ou un switch (c'est plus cher mais meilleur !).
Enfin, l'arrivée des tâches, leur division, et leur répartition sur les nodes constitue notre Cluster.

# 5. Clusters Haute Disponibilité

Les clusters dit à haute disponibilité ont été créés pour prévenir contre les failles hardware et software d'une seule machine, et ceci afin de garder l'ensemble des services d'un système disponible du mieux possible. La redondance, le fonctionnement du cluster et l'assurance contre les pertes peuvent être garanties à 99,9 %.

# 1) Coût de l'indisponibilité

- **Coût moyen d'une heure d'indisponibilité du système** (Source Contingency Planning Research)

| Application | Secteur d'activité | Coût de l'indisponibilité |
|---|---|---|
| Courtage | Finance | $6,45 millions |
| Ventes par carte de paiement | Finance | $2,6 millions |
| Films à la demande | Loisirs | $150 000 |
| Téléachat | Distribution | $113 000 |
| Ventes sur catalogue | Distribution | $90 000 |
| Réservation aérienne | Transport | $89 500 |

Mesures permettant de mesurer les grandeurs et les caractéristiques de la haute disponibilité :

- **Mesures liées au concept de défaillance**
  - **MTTF, ou Temps moyen jusqu'à défaillance** (*Mean Time To Failure*)
  - **MTBF, ou Temps moyen entre défaillances** (*Mean Time Between Failures*)
  - **Note : si pas de redondance MTBF = MTTF + MTTRes**
- **Mesures liées au concept de maintenabilité**
  - **MTTRes, ou Temps moyen jusqu'à restauration (*Mean Time To Restore* ou *Mean Time to Recover*)**
  - **MTTRep, ou Temps moyen jusqu'à réparation (élément) (*Mean Time To Repair element*)**
- **Mesure liée au concept de disponibilité**
  - **At = MTTF/(MTTF+MTTRes)**

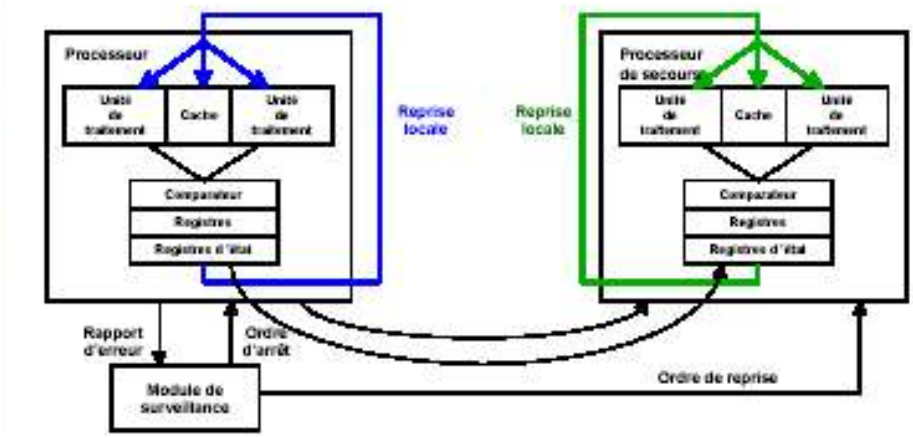
Des solutions peuvent être mises en place, au niveau matériel et au niveau logiciel :

# 2) Solutions apportées

**Au niveau matériel**

Concept de processeur de secours susceptible de venir remplacer un processeur défaillant,
Echange, en fonctionnement, des composants défaillants.

Adoption d'une stratégie pour les modules d'alimentation électrique et les ventilateurs.




Ces solutions sont souvent propriétaires :

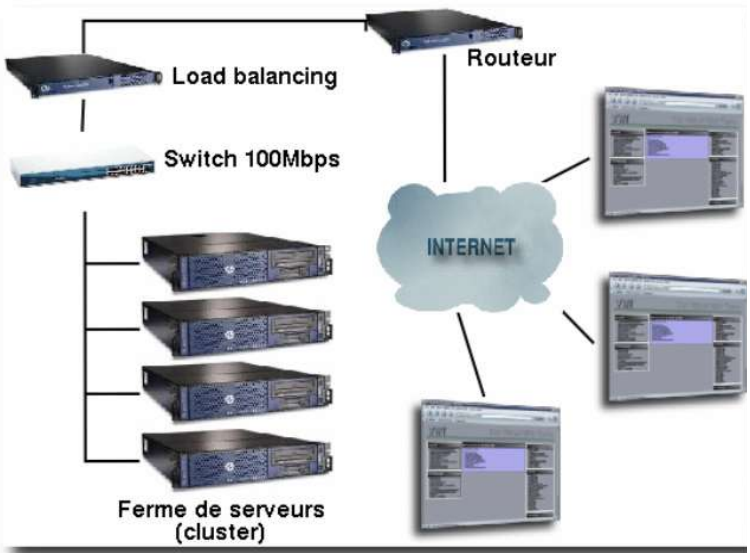
- Continuum Stratus
- Netraft 1800 de Sun
- IBM sur serveur S 390

**Au niveau logiciel**

- Microsoft :
  - Architecture Microsoft Cluster Server MSCS
- Linux :
  - Mission critical
  - Mosix
  - Beowulf
  - LVS

# 6. Clusters à répartition de charge

Les systèmes à répartition de charge permettent de distribuer l'exécution de processus systèmes ou réseaux à travers les nodes du cluster.
Le node server se voit ainsi attribuer la tâche de réceptionner le processus et de le répartir sur la machine adéquate. Cette dernière est en fait choisie car sa charge est faible et donc elle peut traiter le processus entrant de manière quasi instantanée. Elle peut aussi être choisie en fonction de sa spécialisation, c'est à dire qu'elle seule pourra traiter la demande sur l'ensemble des nodes du cluster.
Toutefois, même si les nodes du cluster n'utilisent pas les mêmes systèmes d'exportations et les mêmes entrées sorties, il existe tout de même une relation commune entre eux, matérialisée sous la forme d'une communication directe entre les machines ou à travers un node server contrôlant la charge de chaque workstation. Pour pouvoir répondre à ce besoin de communication, ce type de cluster utilise des algorithmes spécifiques permettant de distribuer la charge.
**Ce type de cluster est surtout largement utilisé dans le domaine du réseau et plus particulièrement sur les services lourds comme les serveurs WEB ou FTP**.
Ces systèmes requièrent des applications qui examinent la charge courante des nodes et déterminent quel node pourra résoudre de nouvelles requêtes. Ainsi, chaque machine se verra attribuer un processus et donc la qualité de service rendu s'en trouvera meilleure. De plus, il évite les surcharges que peut subir une seule machine destinée à répondre aux requêtes du réseau.



# 7. Autres clusters

## 1) Clusters Scientifiques

Typiquement, il s'agit d'un système où l'ensemble des Nodes cumulent leur puissance de calcul pour arriver à des performances égales à celles qu'atteignent les supers calculateurs universitaires. En fait, il est considéré de l'extérieur comme étant une machine multiprocesseurs à part entière, spécialisée dans la résolution de problèmes scientifiques complexes.
Ce cluster utilise des applications spécialisées dans la paralèllisation de calcul à travers une couche de communication commune. En fait, même si TCP/IP représente le protocole réseaux du moment, il diffuse trop de paquets d'overhead, qui ne sont pas forcement nécessaires dans le cas d'un réseau fermé comme un cluster.
A la place, un administrateur pourra utiliser le Direct Memory Acces (DMA, similaire à celui utilisé par certains périphériques d'un PC) à travers ses nodes, qui fonctionne comme une forme de mémoire partagée accessible par l'ensemble des processeurs du système. Il pourra aussi utiliser un système de communication dit de low-overhead comme Message Passing Interface (MPI), qui est une API (Application Program Interface) pour développeurs d'applications de calculs parallèles.

## 2) Clusters de stockage

Ce type de système est comparable au cluster scientifique. Toutefois, ce n'est pas une puissance de calcul qui est recherchée ici, mais plutôt une puissance de stockage.
Les concepteurs de tels systèmes sont partis du constat que les entreprises utilisaient de plus en plus d'applications performantes utilisant des flux de données conséquents et donc nécessitant une capacité de stockage supérieure à celle d'un seul disque dur. Le système en clustering a pu heureusement contourner ce problème en offrant une vaste capacité de stockage virtuel.
En fait, physiquement, le fichier est découpé en bloc de taille raisonnable et stocké par morceau sur plusieurs disques. Virtuellement, on a l'impression que l'espace de stockage ne fait qu'un et que notre fichier est stocké en un seul morceau sur un "disque".
Il s'agit pour ce type de cluster d'utiliser le potentiel des systèmes dits de "stockage combiné", c'est à dire qu'il distribue les données par l'entremise de plusieurs disques répartis sur les nodes du cluster. Ainsi, tout utilisateur aura le loisir de travailler avec des fichiers de très grandes tailles, tout en minimisant les transferts (si la taille des blocs adoptée reste raisonnable).

# 8. Produits existants

## 1) Clusters propriétaires

Typiquement, il s'agit des systèmes proposés par les grands fournisseurs et acteurs mondiaux de l'informatique. Nous retrouvons IBM, SUN, Hewlet Packard, Compaq, Fujitsu, et bien d'autres encore. En fait, il s'agit, pour la plupart, de grands constructeurs de matériels informatiques (serveurs ou stations de travail puissantes), voir des sociétés développant des systèmes d'exploitation propriétaires de type Unix.
Ce sont ces sociétés qui les premières ont mis en place les systèmes en clustering dans les entreprises, en s'inspirant des recherches effectuées dans les grandes universités et laboratoires de recherche. En fait, c'est le meilleur moyen qu'elles ont pour proposer des systèmes complets clés en mains de leurs crus et surtout homogènes de bout en bout.
Toutefois, comme nous pouvons rapidement nous en douter, le coté propriétaire de ce type de système agit sur la non compatibilité des systèmes entre eux. Donc, toute mise à jour devra être effectuée avec des matériels de marque identique à celle de notre système. De plus, le coté spécifique de ces clusters aboutit à l'utilisation quasi obligatoire des prestations fournies par le constructeur.
Toutes ces petites choses font que le coût de tels clusters reste très élevé. En effet, le hardware et le software étant spécifiques à un seul fournisseur, ce dernier ne se

gène pas pour évaluer au plus haut la compétence vendue. Pour un cluster scientifique, il vous faut facilement compter dans les 100 000 € pour les premiers prix. Bien entendu ce chiffre ne représente rien au regard des architectures ultra puissantes coûtant quant à elles plusieurs millions d'euros (> à 1 000 000 €).

# 2) Clusters commerciaux

Il s'agit içi de systèmes proposés par des sociétés de prestations en informatique. En fait, les entreprises proposant la mise en place de clusters désirent rendre un service autour des technologies de clustering. Elles proposent donc une solution se voulant concurrente d'une autre mais performante et peu chère (en comparaison aux systèmes propriétaires).
Ces prestataires proposent, pour la plupart, des solutions employant l'existant informatique de leur client pour adapter leur système. De plus, pour rester concurrent sur le marché, ils utilisent souvent des distributions Linux pour appuyer leur développement et ceci afin de garder une certaine"généricité" quant au système proposé.
Les clusters proposés restent aussi dans un domaine très peu exploré par les grands constructeurs. En fait, ces sociétés de prestations proposent pour la plupart des systèmes pour les réseaux comme les firewalls (en système à Haute Disponibilité) ou encore les architectures de serveurs WEB (clusters à répartition de charge).
Toutefois, ce type de clustering ne s'affiche toujours pas comme étant la panacée. En effet, même si le milieu est concurrent, l'euphorie de ces dernières années a poussé ces sociétés à gonfler le prix de leurs prestations. De plus, au fur et à mesure que l'on avance dans cette ère technologique, les utilisateurs deviennent de plus en plus exigeant et donc les développements relatifs à ces demandes se trouvent de plus complexes et poussés. Tous ceci fait que les propositions commerciales de clustering restent encore des solutions assez chères pour une PME ou un particulier. Il faut compter en moyenne 50 000 € pour posséder un système correct

# 3) Clusters Microsoft

- **Présentation**

Depuis la version NT4 de son système d'exploitation (OS) Windows, Microsoft propose de mettre en place un cluster constitué de serveurs (Microsoft ! bien entendu !) pour répondre aux besoins croissants des entreprises en terme de messagerie électronique, de base de données et depuis quelques années de serveurs WEB ou FTP.
La firme de Redmond propose deux types de clustering :

**Le clustering de Service.** Il s'agit de réaliser des clusters d'application et de rendu de service. En fait Microsoft propose un cluster de Haute disponibilité, à tolérance aux fautes. Il permet de fournir une garantie et une qualité de service aux utilisateurs d'applications comme Microsft SQL Server™.

**Le clustering à répartition de charge.** Ici, Microsoft garanti une répartition de charge réseau sur des flux IP à travers un cluster constitué de 32 nodes au maximum. Typiquement, il s'agit de répartir les charges réseaux d'un serveur WEB, d'un serveur de média, etc.
Les solutions de clustering Microsoft reviennent chères. Dans un premier temps, l'utilisation d'OS et d'applications signés Microsoft coûtent un prix non négligeable. Il vous faut compter plusieurs milliers d'euros (€) pour les licences Microsoft Server 2000 (une licence par poste installé !!!), et autant pour les licences d'applications (comme les bases de données ou les applications bureautiques). Dans un deuxième temps, le matériel nécessaire à la mise en place de tels systèmes coûte cher compte tenue qu'il doit être performant et puissant. Les mini serveurs sont les machines les plus souvent utilisées pour recevoir les OS, et coûtent dans les 20 000 euros (€) pièces. Enfin, dans un troisième temps, le coté trop vulnérable et peu stable des outils Microsoft fait que les dépenses en prestations et protections atteignent des sommes devant être prise en compte (environ 40 000 € par an).

| Operating System | Edition | Network Load Balancing | Component Load Balancing | Server cluster |
|---|---|---|---|---|
| **Windows 2000** | | | | |
| | Advanced Server | 32 | 8 | 2 |
| | Datacenter Server | 32 | 8 | 4 |
| **Windows Server 2003** | | | | |
| | Enterprise Server | 32 | 8 | 8 |
| | Datacenter Server | 32 | 8 | 8 |

- **Pourquoi choisir une solution Microsoft de clustering**

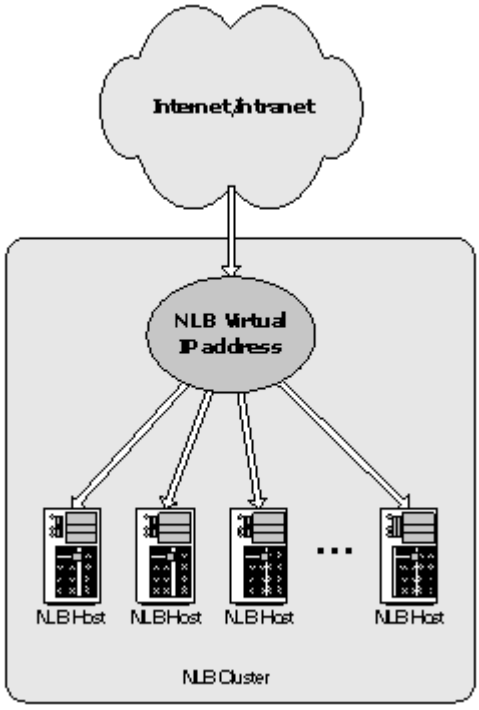
Microsoft a implémenté deux technologies de clustering sur ses serveurs Windows.

### Le service de cluster MSCS

Le service MSCS fournit une haute disponibilité pour les applications critiques, telles que les bases de données, les serveurs de messagerie, serveur de fichier et d'impression.

### Network Load Balancing

NLB permet d'équilibrer le trafic IP entrant.
A travers différentes règles établies les connexions entrantes sont réparties entre les différents nœuds du cluster, il peut y avoir jusqu'à 32 nœuds pour équilibrer la charge IP en mode Network Load Balancing.
Le service d'équilibrage de charge de réseau augmente la disponibilité et la montée en charge des applications serveur basées sur l'accès Internet, tels que des serveurs WEB, des serveurs médias streaming, serveur Windows Terminal serveur ou autres.




Il existe une troisième technologie de clustering implémentée sur les serveurs Application Center.

**Component Load Balancing - Application Center 2000**

Equilibrage de Composants, le service CLB est intégré à Application Center 2000 (ou versions antérieures), ce type de clustering permet de répartir la charge sur plusieurs nœuds du cluster, pour les applications basées sur la technologie des objets COM et COM+ , une mise a jour pour les objets WMI et la gestion du framework .NET est désormais disponible. On parle de clustering d'application dit clustering de puissance.

L'architecture CLB est souvent couplé à la l'architecture de cluster NLB, dans le cas de serveur WEB basé sur le commerce électronique.

**Les rôles:**

**NLB** - répartition des connexions IP et requêtes distantes, répartition de la charge et bande passante sur les nœuds, connexion au site web.
**CLB** - répartition des appels sur les modules d'applications hébergés sur le cluster, accès à l'application commerciale distribuée.

# 4) Approfondissement clusters haute disponibilité - MSCS

## LES RESSOURCES

Une ressource est une entité logicielle qui est publiée sur un serveur virtuel. Elle est ensuite automatiquement partagée et accessible par tous les clients disposant des droits d'accès sur celle-ci. Les ressources sont stockées physiquement sur les disques partagés. Une ressource peut être une adresse IP, un nom réseau,…

### Les groupes de ressources

Les ressources sont rassemblées par groupe, ces groupes sont stockés physiquement sur un seul et unique disque.
Un disque physique = un groupe de ressources. Toutes les ressources dans un groupe peuvent être gérées par les deux nœuds. Il est important de vérifier que le nœud sur lequel se trouve le groupe de ressources possède les capacités d'hébergement nécessaires pour supporter la charge des entrées et sorties.

### Le statut des ressources

en ligne, la ressource peut-être utilisée par un client où une autre ressource

hors ligne, la ressource ne peut pas être utilisée par un client car celle-ci n'est pas publiée

en ligne pending**,** la ressource est dans son processus de mise en ligne

hors ligne pending**,** la ressource est dans son processus de mises en hors ligne

### Les applications

Les applications aware possèdent un mode de mise en cluster

Exemple d'applications aware : SQL serveur, Exchange 2000 serveur, Back Office.

Les applications unaware, ces logiciels n'ont pas d'API permettant de gérer le service de cluster

### Les services

DFS, Distributed File System, système d'arborescence de fichiers distribués, concept inclus sur la suite des serveurs Windows 2000.Vous avez la possibilité de mettre en cluster votre arborescence de fichier pour accroître la tolérance de panne.

DHCP, Service d' attribution automatique d'adresses IP sur un réseau. Disponible uniquement sur la version Advanced Server.

WINS, pour la résolution des noms Netbios/IP.

## LA PHASE DE BASCULEMENT

### Le failover

Nous avons vu précédemment que toutes les ressources sont gérées par un unique nœud.

En cas de défaillance du nœud il y a un basculement automatique de prise en charge de toutes les ressources vers l'autre nœud du cluster, ce processus est appelé failover.

1-*Le client envoie une requête sur le serveur virtuel géré par le nœud 1.*

2-*Le nœud 1 ne peut pas répondre au client car il est hors service*

3-*Le nœud 1 n'émet plus de heartbeat sur le réseau privé ce qui entraîne le basculement des ressources du nœud 1 vers le nœud 2*

4-*Le nœud 2 prend le relais et publie les ressources du serveur virtuel 1 pour les clients externes*

5-*La requête client est alors correctement acheminé et la mise hors service du nœud 1 est obsolète pour le client.*

### Le failback

Après les diverses opérations de maintenance et/ou de remise à niveau sur le nœud hors ligne, la remise en production se fait grâce au procédé de failback.

Il est possible de paramétrer l'instant où le serveur sera remis en production dans le cluster. Il est préférable dans le cas d'un failback de restaurer le nœud durant les périodes creuses des entrées sorties sur le cluster, la nuit ou le matin avant l'arrivée des utilisateurs.

Le nœud reprend ensuite le management des groupes de ressources qui lui étaient attribués initialement.

1-*Le problème technique est résolu et le nœud est remis en production.*

2-*Les datagramme UDP, battements de cœurs sont de nouveaux générés sur le réseau entre les deux nœuds*

3-*Depuis le réseau privé, le nœud 2 est informé de la remise en production du nœud*

*Les groupes de ressources qui étaient initialement attribués au nœud 1 lui sont alors restitué.*

4-*Le nœud 1 synchronise les groupes de ressources et les publie via le serveur virtuel (dont l'IP n'a pas changé).*

5-*Les clients peuvent continuer à faire leur requête, et le failback reste obsolète du côté client, on peut note parfois en fonction des applications utilisées un micro-coupure de connexion au serveur.*

**INSTALLATION DU SERVICE DE CLUSTER**

Le service MSCS est un composant du serveur Windows 2000 inclus sur les versions Advanced serveur et Data Center serveur.

Pour créer le premier nœud du cluster, l'assistant poursuit en vous demandant de saisir plusieurs informations concernant ce premier noeud:

-Nom du cluster
-Nom du nœud
-Créer un compte Service de cluster
, administrateur du cluster
-Désigner des disques partagés sur un bus
-Paramétrage du disque Quorum
-Paramétrage des interfaces réseaux
Pour ajouter un nœud supplémentaire au cluster, l'assistant nous demande de saisir ces informations :
- Nom du cluster à rejoindre, l'intégration d'un nouveau nœud au cluster requiert de se connecter avec le compte Service de cluster crée au préalable.
-Nom du nœud.
-Paramétrage des interfaces Réseaux.
-Mise en production du cluster avec le compte Cluster service.

# 5) Clusters Linux

L'alternative, à tous ces clusters chers, est Linux. En effet, l'utilisation de l'OS au pingouin et l'ouverture d'esprit de sa communauté font de celui ci une solution très intéressante et très viable.
Ainsi, grâce à Linux, vous serez capable de mettre en place un cluster puissant et répondant à toutes les attentes que vous pouvez vous formuler. Il suffit de posséder quelques PC, une distribution Linux et quelques Logiciels permettant de réaliser la parallélisation entre les nodes du système. Une fois ceci fait, vous pouvez développer le cluster que vous désirez, et le proposer par exemple à votre société.

# 6) Orienter son choix sur Linux, Pourquoi ?

Sans conteste le choix de cette Opérating System pour l'élaboration d'un cluster reste le prix quasi nul de la plate forme et des applications qui vont autour. Bien entendu cette gratuité ne peut être mise en oeuvre que par la présence d'une communauté d'informaticiens passionnés de recherche, de solidarité et surtout doués de compétences.

**Le prix minimal**

Au regard des autres solutions de clustering proposées sur le marché, l'alternative Linux reste très inférieure en terme de coût. Ce que vous pouvez économiser dans l'utilisation d'une plate forme GNU/Linux et des applications périphériques, peut être réinjecté dans un matériel plus puissant (quoique vraiment pas nécessaire) pour accroître la puissance de votre cluster. Mais est-il raisonnable de surdimensionner vos besoins ?

- Il faut savoir que la mise en place d'un cluster peut s'effectuer à partir de la majorité des Unix quasi gratuit du marché (Linux, FreeBSD, OpenBSD...). Ceci s'oppose bien sur au prix démesuré des Unix propriétaires, voir des OS Microsoft.
- Ces distributions ont l'avantage de très bien fonctionner sur de simple PC bureautique
- Les logiciels de parallélisation (PVM, MPI...), permettant de transformer notre NOW (Network Of Workstation) en cluster, sont eux aussi sous licence GPL, donc gratuits. Linux et sa philosophie ont donc permis à une majorité de personnes de pouvoir avoir accès aux ressources informatiques puissantes, d'y contribuer et de s'y épanouir et ceci sans avoir à payer des licences hors de prix.

**Architecture recommandée**

Afin de construire un cluster performant, il est tout de même nécessaire de faire un choix hardware et software assez important. En effet, même s'il est tout à fait possible de monter des clusters Linux à partir de machines obsolètes et hétérogènes de bout en bout, les caractéristiques générales de votre système, mais surtout la puissance de votre cluster, s'en trouveront grandement affecté.
C'est pourquoi certains paramètres sont préférables, et comme toute chose, même si cela ne coûte pas cher à la base, la qualité et la rapidité ont un prix. Ainsi, avant de monter votre cluster, il vous faut bien réfléchir sur les points qui sont exposés dans la suite.

## Les projets de clusters sous Linux

Un journaliste informatique américain a dit qu'essayer de compter les projets de clustering sous Linux était l'équivalent d'essayer de compter le nombre de start-up dans la silicon valley".
En effet, contrairement à Windows NT, restreint par son environnement fermé, Linux offre un panel conséquent de systèmes en cluster, répondant aux différents besoins et usages. Dans cette partie je ne traiterais que des projets les plus connus de chaque domaine d'action du clustering.

## LVS

**Le premier projet de clustering dont je vais vous parler aborde les techniques de répartition de charge. En effet, Linux Virtual Server (LVS) a principalement été conçu afin d'apporter performance et disponibilité à des utilisateurs de serveurs réseaux (WEB, FTP...).**
Le principe consiste à interconnecter des serveurs existants, et à orchestrer la répartition de charge par un node server appelé un load-balanceur ("répartiteur de charge").

LVS est donc un projet créant un système à répartition de charge, à partir de traffic TCP/IP entre ces nodes. **Il est implémenté sous forme de patchs applicables au noyau Linux,** et permet alors à des applications réseaux comme les serveurs WEB de fonctionner sur des clusters acceptant plus de connections.
Si vous n'avez besoin que d'une solution simple et peu coûteuse, un cluster de station moyenne catégorie (à base de pentium II) équipé de beaucoup de RAM (256 Mo) peut vous procurer un système à répartition de charge tout à fait honorable.

## Beowulf

Lorsque vous demandez à un linuxien de vous parler de clustering sous Linux, sa première réponse sera, dans la plupart des cas, Beowulf. En fait, sous ce nom se présente un projet de clustering scientifique par le biais de l'OS au pingouin. Parrainé par la NASA depuis 1994, beowulf est devenu le cluster le plus connu du monde Linux.
Sous ce projet se cache une volonté très forte de vouloir concurrencer les supers calculateurs pour un coût nettement inférieur. Ainsi, ce système permet à un ensemble de nodes de fonctionner de concert tel un seul ordinateur. Pour cela il inclut les plus populaires API pour Linux (MPI et PVM) mais aussi des drivers réseaux très performants, etc. En fait, beowulf n'est ni plus ni moins qu'un package d'outils fonctionnant à travers le noyau Linux.
**De plus, pour monter un cluster Beowulf, les contraintes quant au matériel sont amenuisées. En effet, Beowulf permet de connecter des nodes hétérogènes et peu puissants à l'unité, à partir d'une grande variété de types de connexion**. En fait, beowulf permet à un ensemble de nodes de travailler tel un seul PC. Ainsi, les appels effectués vers la machine virtuelle seront en réalité exécutés en fonction de la puissance et de la disponibilité de chaque node. En terme clair, le système va constamment vérifier l'occupation des nodes et y répartir tous les

processus en cours. Mais attention, un seul programme ne pourra pas être divisé sur plusieurs nodes.
Beowulf sera sûrement le premier projet de clustering que vous regarderez (surtout si vos besoins s'orientent vers un clusters de calcul), et tout simplement car il se trouve être le plus documenté et le plus utilisé des clusters sous Linux. Plus qu'un simple paragraphe, c'est un exposé complet qu'il serait bien de faire. Beowulf est à tel point utilisé que l'on rencontre aujourd'hui de multitude de projets basés sur sa technologie.

### PVFS

Le monde du clustering se trouvait déjà vaste à l'arrivée de PVFS. Mais ce qui a fait de celui-ci un projet dont il faut parler est qu'il est l'un des premiers systèmes de fichiers virtuels et parallèles. En fait, il exploite tout le potentiel des systèmes dits de "stockage combiné". Ce projet est développé par le laboratoire de recherche de l'université de Clemson, et soutenu par la NASA.
**En fait plus qu'un cluster de stockage, PVFS se trouve être la solution idéale pour mettre en place un cluster d'entrées / sorties parallèles**. Grâce à ce projet, vous serez en mesure de distribuer des données par l'entremise de plusieurs disques répartis dans les nodes du cluster. Il permet entre autre de travailler sur des fichiers de très grandes tailles tout en minimisant les transferts.
PVFS fonctionne sur Linux depuis le noyau 2.2, et son interface d'entrée / sortie correspond exactement à celui du système Unix. Ceci à l'avantage de pouvoir employer les outils traditionnels sur les fichiers et dossiers sans avoir besoin de les recompiler.

### Linux HA-Project

Ici il s'agit de parler d'un projet de clustering voulant fournir un système à haute disponibilité. En fait, ce projet est à peine arrivé à maturité et les grands constructeurs commencent à s'en servir dans le but de proposer des solutions moins chères à base de Linux.
Toutefois, de plus en plus de personnes s'intéressent à ce type de système, et Linux HA (Hihg Avaibility) vient à être de plus en plus étudié comme une solution adéquate. C'est pourquoi les développeurs du projet se sont attelés à essayer de mettre leur système le plus portable et le plus compatible avec les autres systèmes qui soient. Par exemple, il arrive aujourd'hui à travailler en collaboration avec des serveurs LVS. Certaines parties du développement de Linux HA ont aussi été reprise dans certaines distributions (Mandrake) comme éléments à part entière du système.
Ainsi, le projet est constitué d'applications capables de maintenir un heartbeat entre des nodes d'un cluster. Dans le cas où ce signal venait à disparaître, une application prend en charge l'usurpation d'identité de la machine incriminée pour la redonner à un node redondant. Le node défectueux peut être remplacé en quelques millisecondes grâce à cette méthode. Mais pour bien dimensionner son système à haute disponibilité, le principal inconvénient vient dans le fait qu'il faut bien dimensionner son signal de heartbeat, ceci afin d'éviter de trop longs moments d'inactivitée d'un service en cas de panne du serveur, ou bien d'éviter de polluer le réseaux par des signaux intempestifs apparaissant trop souvent.
Malgré le récent engouement pour ce projet, Linux HA reste encore très peu documenté.

## Alinka

Cette société française, fondée en 1999, propose des logiciels dans l'installation et l'administration de clusters sous Linux. En fait il s'agit de pouvoir installer une application de manière à ce qu'elle se répartisse automatiquement à travers les nodes du cluster, de pouvoir mettre à jour le cluster de manière simplifiée, d'ajouter ou de retirer des nodes sans problèmes, mais aussi de pouvoir observer les charges processeurs ou réseaux tout au long d'un calcul ou d'une répartition.
**Le deuxième logiciel proposé, Alinka Orange, est destiné aux clusters orientés sur les services réseaux comme les clusters Haute Disponibilité ou les clusters à répartitin de charge. Il est destiné à pouvoir gérer des clusters destinés typiquement aux hébergeurs de sites, en dupliquant les services sur les nodes, et en répartissant les charges utilisateurs entres eux.**
Malgré l'aspect commercial de Alinka, cette société propose tout de même une diffusion du cœur des applications Alinka. Ceci vient du fait que les logiciels fonctionnent grâce à des applications Open Source comme PVM, MPI, MOSIX... De plus, le cœur même des **applications Alinka fonctionne avec une base de données PostgreSQL**, et ceci afin d'offrir une très grande évolutivité au logiciel. L'entreprise compte sur la communauté linuxienne pour que son produit évolue, et donc met à disposition des laboratoires et des particuliers, les sources des ses développements.
Pourquoi vendent-ils leur solution me demanderez vous. Tout simplement car Linux reste encore trop académique et la carence en offres commerciales dans le domaine de l'administration des clusters fait foi. Ainsi, l'objectif de Alinka est de rendre accessible ces technologies à d'autres secteurs d'activités, en leur fournissant des outils de qualités répondant aux contraintes industrielles d'une part, et d'autres part en leur apportant un support technique à des solutions commerciales packagées et conviviales, pour lesquels elles auront un interlocuteur identifié et non une communauté.

## MOSIX

Développé à l'origine pour le monde Unix et les ordinateurs spécialisés (comme les VAX 780), MOSIX a su se recentrer sur le monde du PC en 1992. Depuis cette date, les développeurs de MOSIX sont passés d'une plate forme de développement BSD à une plate forme Linux.
Le but premier de l'équipe de développement de MOSIX (basé à l'université hébraïque de Jerusalem) est de fournir un système en clustering, à travers Linux, agissant comme une simple machine vu de l'extérieur (c'est à dire vu des utilisateurs et des processus).
L'idée est fort simple et les initiales de MOSIX permettent de résumer très rapidement le fondement de ce projet de clustering. Ainsi, MOSIX signifie :
*Multicomputer Operating System for UnIX*
**MOSIX va donc permettre de distribuer des tâches et des processus à travers un ensemble de machines en réseaux et spécialisées dans l'exécution parallèle de processus**.

**MOSIX est donc un cluster à répartition de charges entre processus**. En effet, chaque application pourra être migrée entre les nodes afin de tirer avantage de la

meilleure ressource disponible. **Son idée principale consiste à répartir sur plusieurs machines, non pas le calcul, mais le multitâche.**

MOSIX peut répondre à plusieurs types de clusters en fonction du matériel et de sa disponibilité.
Le premier type de clusters appelé un "**single pool**" ;il s'agit d'interconnecter en tant que nodes, des serveurs et des stations de travail. En tant que workstation c'est un avantage car on peut tirer parti de toutes les capacités d'un serveur, mais cela peut aussi être un inconvénient dans la mesure où le PC servira lui aussi de machine de traitement.
Le deuxième type de clusters est dit "**server pool**". Là, il s'agit de n'interconnecter que les serveurs au sein du cluster. Ici, les stations de travail ne font plus partie du cluster et donc ne sont plus utilisées pour exécuter les tâches réparties sur le système. Toutefois, cette méthode à le désavantage de devoir se connecter sur un des serveurs afin de pouvoir utiliser le cluster.
Le troisième type de cluster est nommé "**adaptive pool**". Ici, il s'agit de reprendre la technique du single pool et de n'autoriser les workstations à pénétrer sur le système en tant que nodes, qu'à des moments bien précis (par l'intermédiaire la crontab par exemple). En fait le cluster réagira en tant que server pool à des moments, et en tant que single pool à d'autres. Cette technique à l'avantage de pouvoir utiliser les stations de travail au sein du cluster lorsque l'utilisateur ne l'utilise pas (typiquement la nuit, ou lorsqu'il part en pause, par l'intermédiaire de son économiseur d'écran par exemple).
Le dernier type de cluster est appelé "**Half-duplex pool**". En fait, c'est un single pool "intelligent". Les workstations peuvent envoyer des processus au serveurs du cluster mais celles-ci ne s'occuperont que de leurs propres processus. On dit que les stations de travail ne font partie du cluster que pour exécuter leurs propres processus, et ne se soucient donc guère des processus extérieurs.
Ainsi, MOSIX affiche une grande variété de solutions permettant de l'utiliser dans diverses architectures déjà mises en place..

# 7) Conclusion

### Choix du système d'exploitation

Pour les serveurs d'applications, l'architecture clustérisé est souvent une extension de l'architecture de base ( celle sur laquelle à était développé l'application), le choix de l'os cluster à déjà était fait dans les premières étapes du développement de l'application.

### Choix du type de cluster

Le type de cluster le plus utilisé concerne la haute disponibilité (service 24/24 et 7 / 7) ; l'équilibrage de charge concerne le plus souvent les serveurs d'applications web.
Dans une prise de décision le facteur coût a une importance prépondérante, ainsi la répartition de charge peut se faire par des éléments actifs de réseaux, ou des composants logiciels réseaux (DNS, MX,…) sans passer par des clusters.

# 8) Bibliographies et sites Internet

Architecture de l'ordinateur – Andrew Tanenbaum – InterEditions

Client serveur – S. Miranda, A. Ruols - Eyrolles

http://clusters.top500.org/
The Linux Clustering Information Center. Site centralisant des liens et des informations en ce qui concerne le clustering sous linux.

http://www.sun.com/clusters/
Le cluster de SUN.

http://hp-linux.cern.ch/
Projet interressant de clustering à partir d'HP-UX et de stations HP.

http://www.microsoft.com/windows2000/technologies/clustering/default.asp
Une page de documentation sur les clusters à base de Windows 2000 Server.

http://www.linuxvirtualserver.org/
Le site officiel de LVS.

http://www.beowulf.org/
Le site officiel de Beowulf

http://www.mosix.org/
Le site officiel de MOSIX