Réorganisations et optimisations des supports de masses Aster

17 janvier 2003 - Aster - Côte d'Ivoire

rapport d'expertise

Éric Burghard - Assistant technique système et réseaux

Ce rapport technique propose différents schémas de réorganisation des supports de masses, plus adaptés aux fortes sollicitations d'Aster, et qui se conforment, au contraire de la configuration actuelle, à l'architecture de haute disponibilité qui a été choisie à l'origine du projet. Cette optimisation de bas niveau doit nécessairement s'effectuer avant toutes tentatives du même type au niveau des bases de données ou du progiciel, d'une part pour palier la faible résistance aux pannes de la configuration actuelle, inacceptable pour une application de comptabilité de l'état, et d'autre part pour disculper définitivement les systèmes dans le problème de performance du SI Aster en Côte d'Ivoire.

Reorg. disques 20030117

Table des matières

1. Configuration actuelle

1.1. Serveurs

Le trésor est équipé de deux serveurs IBM RS/6000 7026-H20 dotés de

- 4 processeurs PowerPC RS64 II, cadencés à 340Mhz
- 4Mo de cache de niveau 2
- 1Go de RAM par processeur

sur lesquels est installé le système Unix IBM AIX 4.3.2.1. Pour éviter les pannes critiques, la plupart des pièces indispensables au fonctionnement des serveurs ont été doublées. C'est le cas des alimentations, des interfaces réseaux, ou des interfaces disques.

1.1.1. Grappe de serveurs

A l'origine, il avait été décidé d'assurer la plus grande disponibilité possible de l'application Aster, en utilisant le deuxième serveur en cas de panne empêchant l'exécution normale du progiciel, sur le serveur principal.

Les serveurs coopérants entre eux sont regroupés dans ce qu'on appelle une grappe, et HACMP est l'outil informatique qui orchestre la reprise d'une application au sein d'une grappe de serveurs. Pour rendre ce basculement le plus transparent possible à l'utilisateur, il faut que le serveur de secours:

1. prenne l'identité réseau du serveur défectueux,
2. s'accapare de ses ressources disques qui lui sont désormais inutiles,
3. et relance tous les services que l'on désire reprendre (les bases de données).

1.1.2. Ressources disques

Chaque serveur dispose d'un accès matériel:

- exclusif à une baie de disques internes pourvue de 8 emplacements, et composée actuellement de 4 disques SCSI de 9Go, sur lesquels est installé le système d'exploitation et les outils Oracle,
- mutuel à deux baies (modèle 7133 D40) de disques, chacune pourvues de 16 emplacements, et composées actuellement de 6 disques SSA de 9Go. Chaque serveur possède deux contrôleurs SSA (advanced SerialRAID).

Dans la configuration actuelle, un seul serveur dispose de la totalité des ressources disques, ce qui limite l'utilisation conjointe des deux serveurs. L'installation des bases Aster dans la baie interne des serveurs, reste possible, mais empêche toute évolution du serveur puisque la

totalité de l'espace disque est alors utilisé. Cette installation anéanti surtout tout espoir de reprise en cas de panne, puisque l'état de fonctionnement de la baie interne est liée à celui du serveur.

1.2. Architecture de stockage en série (SSA)

Dans cette architecture, les disques sont placés en série le long d'une chaîne fermée par des contrôleur SSA au niveau du serveur. Les données des entrées/sorties peuvent transiter simultanément à haut débit (40 Mb/s) dans les deux sens de la boucle. Les données peuvent transiter de disque à disque ou de disque à contrôleur.

Cette architecture simple permet de réduire les coûts du disque par rapport à une architecture de type parallèle (SCSI, E-IDE), mais permet surtout de délocaliser, jusqu'à une dizaine de kilomètres, une baie de disques en la reliant simplement par fibre optique au serveur qui les exploite.

1.2.1. Résistance aux pannes

La configuration retenue à Abidjan (FIG. 1) a été de placer tous les disques dans deux boucles formées entre les quatre contrôleurs et les deux baies des deux serveurs. De la sorte, l'accès à la totalité des disques reste possible:

1. par un serveur même si une de ses deux interfaces tombe en panne: chaque contrôleur est doublé,
2. par le serveur de secours si les deux interfaces du deuxième serveur tombent en panne,
3. par le serveur de secours si le deuxième serveur tombe en panne.

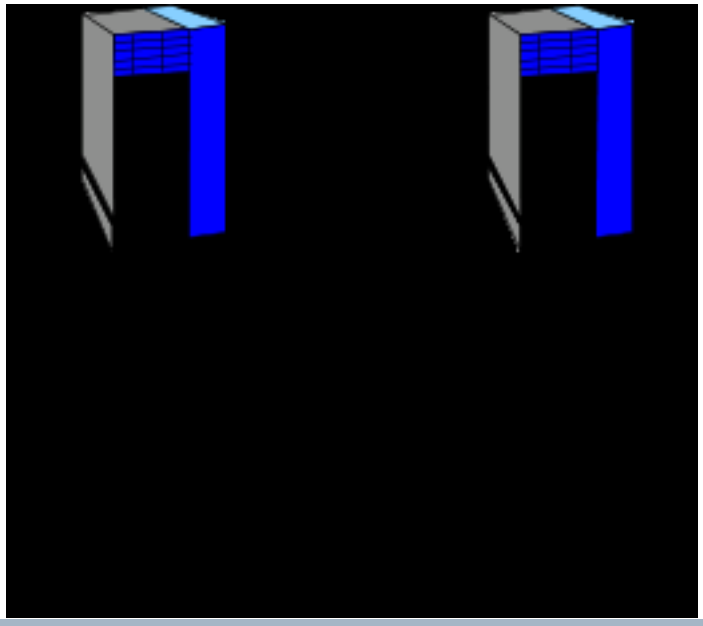

FIG. 1 : Configuration actuelle des disques partagés entre les deux serveurs

1.2.2. Contraintes matérielles

La configuration actuelle, orientée vers la haute disponibilité, limite les performances au niveau du débit disques/serveur, à cause de contraintes physiques, ou liées au matériel:

1. la redondance de données ne peut être gérée que de manière logicielle (voir),
2. la distance entre les deux baies, reliées par fibre optique, provoque une chute du débit disque/serveur qui passe de 40 Mb/s à 12 Mb/s.

On peut proposer d'autres configurations qui touchent à ces deux points de manière à s'affranchir des contraintes sous-jacentes, mais toujours au détriment de la disponibilité du système dans un environnement dégradé par des pannes. Les gains engendrés sont quantifiables dans l'absolu, mais ne peuvent être observables que si le système est actuellement au maximum de ses possibilités au niveau du débit disque/serveur.

1.2.3. Contraintes organisationnelles

Un moteur de base de données sollicite parfois de manière intensive les disques en écriture et/ou en lecture. En dehors des débits physiques bruts exposés à la section, tout ce qui touche à l'organisation des données au sein des disques peut aussi avoir une influence non négligeable sur les performances d'Aster.

Les améliorations que de telles optimisations engendreraient sur les performances de la base

de données sont cependant difficilement calculables puisqu'elles interviennent au sein de mécanismes non déterministes.

Plusieurs constats nous permettent tout de même de nous faire une idée du stress subi par les baies, et d'évoquer les points perfectibles:

1. 4 bases Aster peuvent tourner simultanément sur un même serveur à des degrés divers d'activité,
2. les données applicatives (formulaires, dépêches,...), les données des bases et leur indexes sont répartis sur le même groupe de disques. Autrement dit, la compétition pour l'accès aux données fait rage entre clients Aster, moteurs Oracle, et procédures d'exploitation qui se gênent mutuellement,
3. 40 clients en moyenne accèdent aux données applicatives, via le protocole de partage de fichier sur le réseau local.
4. aucune stratégie n'a été choisie pour placer les données fréquemment utilisées aux positions les plus efficaces dans la baie ou sur les disques,
5. les indexes (au contraire des données) ont été placés en mode miroir, alors qu'ils peuvent être régénérés à partir des données. Cette redondance est assuré au niveau logiciel (LVM), ce qui peux, selon les cas, doubler le débit des données au niveau des contrôleurs, et augmenter la charge processeur de 25%, tout en ralentissant les mises à jour des tables,

1.3. Optimisations possibles

Le serveur est confronté aujourd'hui à un problème de surcharge, et les utilisateurs à un problème de lenteur de l'application.

On peut doubler simplement la puissance de traitement en utilisant conjointement les deux serveurs, et en leur répartissant équitablement les bases Aster. Seulement dans la configuration actuelle un seul serveur accède de manière exclusive au seul groupe de volume qui englobe la totalité des disques externes. Il convient donc de leur attribuer à chacun une partie des disques SSA de manière à pouvoir utiliser les deux serveurs en parallèle vis-à-vis des ressources disques externes.

Les problèmes de la section peuvent être traités, en même temps que cette réattribution de disques, de la manière suivante:

1. il faut définir au moins un groupe de volumes par serveur. Ce groupe de volume contiendra des systèmes de fichiers pour chacune des bases de données. La moitié des bases sera attribuée à tresor01, et l'autre à tresor02.
2. Il faut ensuite s'assurer que les données et les indexes se situent sur des disques différents du groupe de volume, grâce à un partitionnement en volumes logiques. De cette manière deux bases d'un même serveur ne rentrent pas en compétition pour l'accès aux données,

3. Les données applicatives destinés aux clients peuvent quand à elles se situer sur serveur NT.
4. de manière à s'affranchir du débit vers la baie distante plus faible que vers la baie locale, les bases destinées à être lancées sur un serveur seront installées dans la baie locale. Il faut ensuite placer les bases les plus actives en termes d'entrées/sorties sur les disques au milieu de la chaîne SSA,
5. enlever le mode miroir sur les indexes, et le mettre sur les données. En modifiant la topologie SSA, il est possible de mettre en place une redondance des données au niveau matériel réputée plus efficace (voir Annexe), mais au détriment de la disponibilité globale en environnement dégradé.

2. Configurations possibles

Quatre reconfigurations, qui découlent de trois orientations différentes, sont dès lors possibles:

1. on reste dans l'idée originelle qu'un seul serveur exécute les bases Aster et que l'autre sommeille en attendant une panne. Dans ce cas il faut au moins mettre les données en mode miroir pour rester cohérent sur cette idée de haute disponibilité,
2. on reste dans l'optique de haute disponibilité de l'application, mais on permet l'utilisation des deux serveurs simultanément sur des bases Aster:
    1. on optimise uniquement le placement des bases sur les disques en segmentant la baie en plusieurs groupes de volume, ou
    2. on redéfinie légèrement le branchement des serveurs et de baies de manière à lever la contrainte [enu:redondance] de la section pour gérer plus efficacement la redondance au niveau matériel,
3. on abandonne le concept de haute disponibilité, et on découple, les deux serveurs et les baies pour leur garantir une vitesse maximale d'accès aux disques avec une gestion de la redondance qui s'effectue au niveau matériel.

Les solutions 2a et 3 imposent une modification de la topologiela topologie concerne la façon dont sont branchés les baies et les contrôleurs. SSA, c'est à dire une reconfiguration matérielle, au contraire des autres solutions dont l'intervention se limite à des aspects logiciels et organisationnels.

2.1. Mise en place d'un miroir

Cette solution apporte par rapport à la configuration actuelle:

1. la mise en place effective d'une redondance des données,
2. la résistance aux pannes de disques et de baie.

2.1.1. Principes

On reste dans la configuration d'origine dans laquelle le serveur de secours attend la panne du serveur principal pour prendre possession des ressources disques et relancer les services. Un seul serveur à la fois accède donc à la totalité des 12 disques qui sont entièrement doublés, soit 6 disques utiles.

2.1.2. Proposition

Pour garantir la continuité de service même en cas de panne d'une baie, il faut s'assurer que chaque disque d'une baie donnée trouve sa doublure dans la baie distante.

Les recommandations 2 à 5 de la section (mise à part ce qui concerne la redondance gérée par le matériel) sont valables et doivent être appliquées.

2.1.3. Scénari de reprise

Dans la mesure ou le système est résistant à toutes les pannes référencées dans le tableau 2, l'indisponibilité du système en cas de panne, peut être de l'ordre de la minute si HACMP est correctement configuré, et de l'ordre de 15 minutes si la reprise est faite manuellement sur le serveur par un administrateur.

2.2. Mise en place de miroirs et parallélisme

Cette solution apporte par rapport à la configuration précédente la possibilité d'utiliser les deux serveurs en parallèle.

2.2.1. Principes

Au lieu d'affecter la totalité des disques à un seul serveur, ont créé deux groupes de volumes pour pouvoir en affecter un à chaque serveur. On dispose à ce moment de suffisamment d'espace pour installer les bases nécessaires. Chaque serveur devient principal pour ses propre bases et de secours pour les bases de l'autre serveur.

2.2.2. Proposition

Deux groupes de volumes seront définis, chacun avec la moitié des disques de chaque baie. Chaque groupe de volume contiendra 3 volumes logique: 2 volumes pour 2 ensembles de bases et un volume pour les index. On s'arrange au niveau de chaque volume logique pour que chaque disque trouve son miroir dans l'autre baie.

Chaque baie sera organisée de la façon suivante. L'ensemble des 6 disques seront regroupés en un seul groupe de volume composée par 3 volumes logiques

2.2.3. Scénari de reprise

De même que pour la solution précédente, HACMP peut être configuré pour accélérer et automatiser la reprise. Cette reconfiguration est légèrement plus complexe dans ce cas, puisque chaque serveur est a même de seconder sont homologue.

2.3. Mise en place d'un RAID5

2.3.1. Principes

Pour gérer la redondance au niveau matériel et rendre sa gestion la plus transparente possible, on attribue, à chaque serveur, la totalité des disques de sa baie locale. Ainsi sur les 6 disques de chaque baie:

- 5 disques sont placés en RAID5 (3 disques utiles, 1 disques de redondance, 1 disque de secours à chaud),
- 1 disque est placé en RAID0, et servira pour stocker les indexes des bases. Ceux-ci peuvent être régénérés, et n'ont donc pas besoin d'être redondant dans le système.

Soit au total, et par serveur, 27Go d'espace disque pour les bases de données, et 9 Go pour les indexes, ce qui représente plus d'espace que celui occupé actuellement par la totalité des bases.

2.3.2. Proposition

Dans cette configuration, chaque serveur garde l'accès par fibre optique à la baie distante, mais son contrôleur n'est plus secouru directement en cas de panne (FIG. 2). En revanche, puisque l'on a plus que 2 contrôleurs par boucle SSA, l'installation d'un système RAID5 est possible.

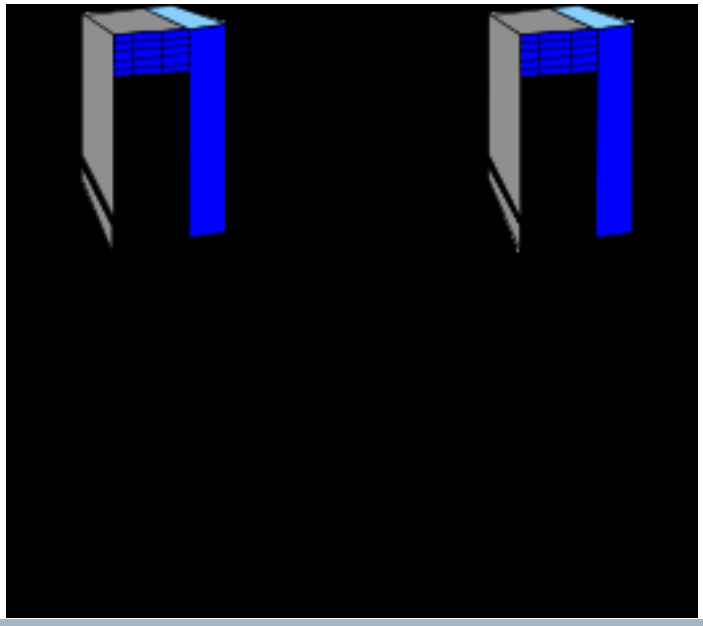

FIG. 2 : RAID 5 sans découplage des baies

L'accès aux données se fait à 104Mb/s sur les disques locaux, et 24 Mb/s sur les disques distants.

2.3.3. Scénari de reprise

- en cas de panne d'un contrôleur, il faut rebrancher les câbles du contrôleur défectueux sur le contrôleur de secours.
- en cas de panne d'un serveur, pour le peu que celui-ci reste alimenté en courant électrique, le second serveur peut reprendre l'exécution des bases affectées au serveur en panne, en s'appropriant ses ressources disques. Cette reprise peut être manuelle ou être configurée au niveau d'HACMP,
- en cas de panne sévère d'un serveur qui empêche l'alimentation des contrôleurs, il faut déplacer la baie sur le site du serveur de secours et la brancher sur le contrôleur de secours.

2.4. Mise en place d'un RAID5 et découplage

2.4.1. Principes

Cette configuration permet de disposer de deux serveurs autonomes possédant chacun un accès aux disques optimal en terme de débit, mais au sacrifice de la haute disponibilité

puisque dans cette configuration aucune procédure simple et rapide n'existe pour reprendre les données et les services d'un serveur tombé en panne.

2.4.2. Proposition

Dans cette configuration (FIG. 3), les deux baies externes sont désolidarisées et indépendantes de manière à :

- disposer d'un contrôleur de secours sans dépasser la limite de 2 contrôleurs par boucle pour pouvoir installer une grappe de disques RAID5,
- augmenter le débit entre le serveur et la baie locale, en supprimant la liaison optique.

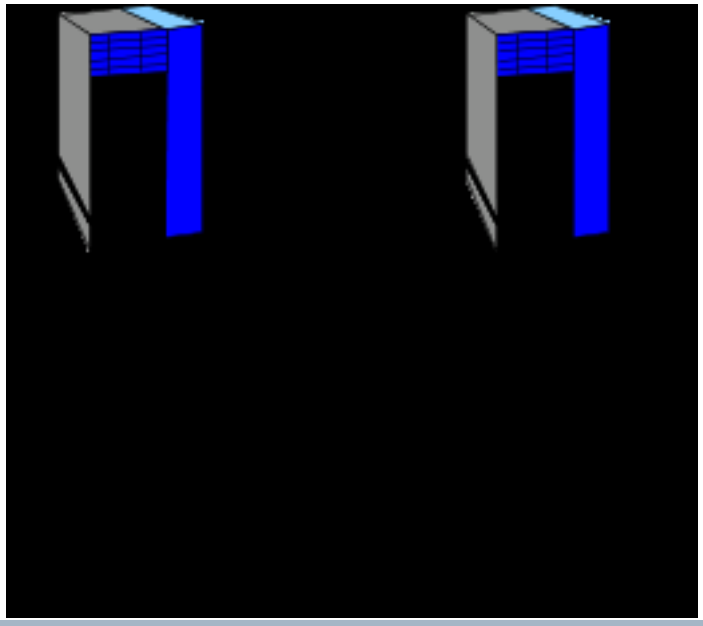

FIG. 3 : RAID5 avec baies découplées

Le débit entre un serveur et sa baie locale passe à 160Mb/s en lecture/écriture et devient optimal dans l'absolu.

2.4.3. Scénari de reprise

- en cas de panne d'un contrôleur, le contrôleur de secours prend automatiquement le relais,
- en cas de panne d'un serveur, il faut déplacer la baie du serveur en panne et l'installer sur le site du serveur de secours,

3. Classification RAID

La classification RAID (Redundancy Array of Independant Disk) permet de caractériser le niveau de redondance des données au niveau d'une grappe de N disques utiles:

- le niveau RAID0, correspond à une grappe sans redondances et donc sans tolérance aux pannes. Ce niveau nécessite N disques,
- le niveau RAID1, correspond à une grappe dans laquelle chaque disque est cloné; il garantit l'intégrité des données lors de la panne de la moitié des disques, pour le peu qu'un disque et son clone ne tombent pas simultanément en panne. Ce niveau nécessite 2*N disques.
- le niveau RAID5 garantit l'intégrité des données lors de la panne d'un disque de la grappe; les données qui se trouvaient sur le disque défectueux peuvent être récupérées à partir des autres disques en état de fonctionnement. Ce niveau nécessite N+1 disques. Cependant si un deuxième disque tombe en panne avant le remplacement du premier, les données peuvent être corrompues. Pour minimiser les chances de double panne, un disque non utilisé peut être placé dans la grappe de manière à secourir immédiatement un disque défectueux. Dans ce cas ce niveau nécessite N+2 disques.

Les contrôleurs SSA des serveurs unix permettent de mettre en place une grappe RAID5 si et seulement si la boucle SSA ne contient pas plus de 2 contrôleurs. Préalablement à la mise ne place de grappes RAID5, cette contrainte nécessite de changer la topologie actuelle des deux boucles SSA qui contiennent chacune 4 contrôleurs (FIG. 1).

Les avantages d'une configuration RAID5 concernent:

1. la mise en place d'une redondance des données transparente et efficace, puisqu'elle est gérée au niveau matériel,
2. la résistance aux pannes disques.

4. Tableaux comparatifs

	Configurations				
Caractéristiques	Origine	Mirroirs	Miroirs x2	RAID5	RAID5 x2
Débit baies locales Mb/s	104	104	104	104	160
Débit baies distantes Mb/s	24	24	24	24	-
Type de miroir	logiciel	logiciel	logiciel	matériel	matériel
Miroir actif	non	oui	oui	oui	oui

Serveurs autonomes	non	non	oui	oui	oui
Type de reprise	auto	auto	auto	auto	manuelle
Espace disque Go	108	54	2*27	2*36	2*36
Disques utiles	12	6	2*3	2*4	2*4

Table 1: Comparaisons des différentes configurations

	Configurations				
Tolérances	Origine	Miroirs	Miroirs x2	RAID5	RAID5 x2
panne de contrôleur	oui	oui	oui	non	oui
indisponibilité	0	0	0	4h	0
panne de serveur	non	oui	oui	oui	non
indisponibilité	4h	5min	5min	5min	4h
panne de disques	non	6	6	1	1
indisponibilité	1j	0	0	0	0
panne de baie	non	oui	oui	non	non
indisponibilité	1j	0	0	1j	1j

Table 2: Tolérances aux pannes des différentes configurations et temps d'indisponibilité en cas de panne

5. Références

- Site IBM
- Ressources RS/6000
- Documentation des baies et contrôleurs SSA IBM
- Ressources administration Aix

Éric BURGHARD