Approche analytique pour l'étude des performances de serveurs multimédias multidisques en grappe

Hichem Kaddeche†‡, André-luc Beylot* et Monique Becker†‡

† Institut National des Télécommunications, Département informatique
9 rue Charles Fourier, 91011 Evry Cedex - France
Tél: (33 1) 60 76 47 42, Fax: (33 1) 60 76 47 80
E-mail : kaddeche@etna.int-evry.fr, mbecker@etna.int-evry.fr

‡ Laboratoire LIP6, Université de Paris 6
4 place jussieu, 75252 Paris Cedex - France

* Laboratoire PRISM, Université de Versailles St Quentin en Yvelines
Avenue des Etats-Unis, 78035 Versailles Cedex - France
E-mail : beylot@prism.uvsq.fr

Résumé

Les applications multimédias futures nécessitant des serveurs capables de fournir des séquences vidéo à la demande à plusieurs utilisateurs simultanés géographiquement éloignés sont nombreuses et variées. L'architecture multidisque en grappe (*cluster*) qui consiste en un ensemble de nœuds de stockage et de transmission reliés par un réseau d'interconnexion semble bien adaptée à ce type d'applications. Les fichiers vidéo sont fragmentés en blocs et distribués sur les différents disques.

Nous présentons dans cet article une approche et un modèle analytique pour l'étude de ces serveurs. Nous présentons en détail la modélisation des différents composants du système et le modèle simplifié final adopté. Le modèle permet de dimensionner rapidement les paramètres de fonctionnement. Il permet d'économiser des simulations coûteuses en temps et d'éviter les problèmes d'événements rares. Nous nous sommes intéressés à la qualité de transmission caractérisée par la probabilité de rupture momentanée dans la transmission des blocs. Une rupture momentanée de transmission se produit lorsque le bloc à envoyer n'est pas chargé dans les délais au niveau du nœud de transmission qui s'occupe de son envoi sur le réseau extérieur. Ce bloc ne sera pas envoyé sur le réseau extérieur ce qui engendrera une petite interruption au niveau de l'utilisateur. Le critère de performance étudié est alors la probabilité de retard pour un bloc. Cette probabilité dépend de la fonction de la répartition du temps de chargement d'un bloc : temps d'attente et d'extraction au niveau du disque et durée de traversée du réseau d'interconnexion. Dans le modèle simplifié final les disques sont modélisés par des files M/D/1 et le réseau d'interconnexion par un délai constant.

Le modèle proposé est validé par de longues simulations réalisées sur un modèle précis sous forme de réseau de files d'attente. L'utilisation du modèle pour le dimensionnement du système est alors présentée.

Mots clés : serveurs multimédias, architecture en grappe, modèles analytiques, évaluation de performances, disques, réseaux d'interconnexion.

Abstract

There are many different future multimedia applications requiring servers which are able to provide video sequences on-demand to simultaneous users through a high speed network. A convenient design for these servers seems to be a clustered architecture including a set of storage nodes (with a local disk array) and a set of delivery nodes, those two sets being interconnected by a switch. Video files are shared into blocks and distributed on the disks.

In this paper an analytical model is designed for the study of these servers. We present the detailed modeling of the different components of the system and a final simplified model. The model allows a quick dimensioning of the operating parameters. It allows to save expensive simulations and to avoid rare event problems. The performance evaluation study focuses on the quality of transmission characterized by the probability of happening of a break in the delivery of blocks. A brief transmission break occurs when the expected block is not loaded in time in the delivery node that should manage its transmission to the external network. If the block is late, it will not be transmitted and this will result in a short break at the user end. The performance criterion considered is the probability of delay for a block. This probability depends on the distribution of the loading time of a block : the response time of the disk plus the response time of the switch. In the final simplified model, the disks are modeled by M/D/1 queues and the switch is modeled by a constant delay.

The proposed model is validated through extensive simulations of an accurate queueing network model. The dimensioning of the system is then derived from the model results.

Key words : multimedia servers, clustered architecture, analytical models, performance evaluation, disk, switch.

1. Introduction

Les développements récents de l'informatique et des technologies multimédias ont donné naissance à de nombreuses nouvelles applications qui manipulent des données vidéo. La plupart de ces applications sont encore locales sur des machines isolées ou limitées à de petits réseaux locaux. Les applications multimédias à distance disponibles actuellement sont généralement de qualité limitée tant au niveau de l'image qu'au niveau de l'interactivité et du temps de réponse (les services du World Wide Web).

L'arrivée des réseaux à haut débit [Vetter 95] permettra de remédier aux limitations actuelles de transmission et rendra possible le développement de services multimédias distribués de qualité. La réalisation de tels services nécessite le développement de serveurs capables de stocker et de délivrer à la demande des données vidéo pour plusieurs utilisateurs simultanés. Les utilisateurs seront géographiquement distribués et connectés à distance au serveur à travers le réseau. Les charges visées pour ces systèmes sont de l'ordre de centaines voire de milliers d'utilisateurs simultanés. La problématique générale des systèmes multimédias avec ses deux aspects serveur et réseau est présentée dans [Vin 94].

Les données vidéo consistent en une suite d'images qui n'ont de sens que lorsqu'elles sont présentées d'une façon continue contrairement aux données numériques classiques (texte, image) ou une simple continuité spatiale suffit. La conséquence de cette continuité temporelle est que ces données nécessitent des techniques différentes pour leur organisation et leur gestion. Un serveur multimédia doit assurer la distribution des flux vidéo à leur débit temps réel.

Contrairement aux serveurs mono-utilisateurs, la conception des serveurs multi-utilisateurs capables de gérer un grand nombre d'utilisateurs simultanés, est une opération complexe qui lance beaucoup de défis. Aux problèmes de volume de données, de hauts débits et de contraintes de continuité temporelle, s'ajoute la difficulté de la gestion des systèmes multi-utilisateurs. Le système doit être capable de servir le même fichier à différents utilisateurs sans dégradation de ses performances. Des solutions basées sur le parallélisme semblent apporter des réponses à un certain nombre de ces problèmes. La fragmentation des objets multimédias en plusieurs blocs et leur répartition sur plusieurs disques permet d'améliorer considérablement les performances en favorisant un meilleur équilibrage de la charge [Berson 94] [Haskin 95]. Les études des systèmes de stockage multidisques ont donné lieu à plusieurs propositions et à l'évaluation des performances de différents schémas [Livny 87] [Ganger 94].

Les applications futures pour ces serveurs multimédias sont nombreuses et variées. Elles concernent les services d'archives et d'information, la télévision à la demande, les films à la demande, l'enseignement à distance, les informations à la demande ... ce grand ensemble d'applications fait que ces serveurs sont d'une grande importance économique qui suscite l'intérêt d'industriels de l'informatique, des télécommunications et des loisirs. Cet enjeu économique est à la base des nombreux travaux de recherche qui sont menés actuellement dans le domaine.

Un certain nombre d'architectures ont été proposées pour les serveurs multimédias ces dernières années. Un état de l'art de la problématique et ses solutions est présenté dans [Gemmell 95]. Nous nous intéressons dans cette étude à une architecture qui semble bien adaptée pour ce genre de système [Damm 94] [Kaddeche 96]

[Tewari 96]. Il s'agit d'une architecture qui comporte un ensemble de nœuds de stockage et de transmission connectés par un réseau d'interconnexion. On parle alors de serveurs en grappe *(clustered servers*). L'étude que nous présentons traite des performances du système indépendamment du réseau extérieur.

Nous avons concentré principalement notre étude sur la qualité de la transmission qui dépend des ruptures momentanées dans la transmission des blocs. Nous avons estimé la probabilité pour qu'un bloc provoque une rupture (c'est à dire la probabilité pour qu'un bloc d'information arrive après sa date prévue de transmission). Nous présentons dans cet article une approche et un modèle analytique simplifié que nous validons par des simulations et exploitons pour le dimensionnement et l'étude de l'effet de certains paramètres internes de fonctionnement du système.

L'article est structuré comme suit. Le paragraphe 2 décrit l'architecture et le principe de fonctionnement du système. Dans le paragraphe 3, nous présentons le principe de l'approche analytique. Les paragraphes 4 et 5 présentent l'étude détaillée des disques et du réseau d'interconnexion. Le paragraphe 6 présente le modèle de simulation utilisé pour la validation de l'approche analytique. La validation et l'exploitation du modèle analytique sont alors présentées dans le paragraphe 7.

2. Architecture

Les séquences vidéo considérées sont numérisées et compressées suivant le format MPEG-2 (*Motion Picture Expert Group*) [ISO 93]. Chaque séquence produit un fichier informatique binaire classique. Le format de compression MPEG-2 utilisé induit un débit constant (*Constant Bit Rate , CBR*) de 0,5 Mo/s. La fonction du serveur est de stocker et de transmettre à la demande ces fichiers vidéo en respectant le débit requis.

Chaque fichier vidéo est fragmenté en blocs de même taille correspondant à des fragments de séquences de même durée. Les blocs sont distribués sur les différents disques suivant une politique statique de placement. Transmettre une séquence vidéo consistera alors à transmettre dans le bon ordre et d'une manière isochrone les différents blocs de la séquence en respectant le débit MPEG imposé.

L'architecture consiste en deux groupes de nœuds, nœuds de stockage et nœuds de transmission, connectés par un réseau d'interconnexion (RI). Les nœuds de transmission assurent la collecte des blocs et leur transmission sur le réseau extérieur à travers des processeurs d'interfaçage. Les nœuds de stockage assurent la gestion des disques ou des groupements des disques qui leur sont connectés. Le réseau d'interconnexion a pour fonction d'assurer la jonction entre les différents éléments de ces deux groupements et d'assurer ainsi l'acheminement des différentes entités échangées.

Quand une requête arrive au système, elle est affectée à un noeud de transmission qui gérera son service. Le noeud assurera la collecte et la transmission des différents blocs de la séquence demandée. L'envoi des blocs sur le réseau extérieur commence après une étape d'anticipation qui consiste à précharger simultanément A premiers blocs de la séquence considérée. Le nombre d'anticipation A est un paramètre interne de fonctionnement du système qui est fixé au préalable. Après cette étape d'anticipation commence l'étape de transmission. Les blocs sont envoyés sur le réseau extérieur d'une manière cyclique isochrone suivant une période égale à la période $\Delta_{émission}$ d'émission d'un bloc (taille du bloc / débit MPEG). La collecte des blocs se fait en parallèle suivant les mêmes cycles. A chaque envoi de bloc sur le réseau extérieur un ordre de chargement du bloc à extraire est envoyé au disque concerné. L'ordre d'extraction généré par le noeud de transmission traverse le réseau d'interconnexion pour atteindre le noeud de stockage puis le disque qui contient le bloc demandé. Une fois le bloc extrait, il est envoyé au noeud de transmission demandeur via le réseau d'interconnexion en empruntant le chemin inverse de l'ordre. Le bloc est alors stocké dans la mémoire du noeud de transmission jusqu'à l'arrivée de sa date d'échéance d'envoi sur le réseau extérieur. Si le bloc arrive après cette date d'échéance, il sera considéré comme perdu. Cela induit une petite interruption au niveau de l'utilisateur de durée égale à la durée d'émission du bloc. Ces cycles sont répétés jusqu'à la fin de la séquence ou la demande d'interruption par l'utilisateur.

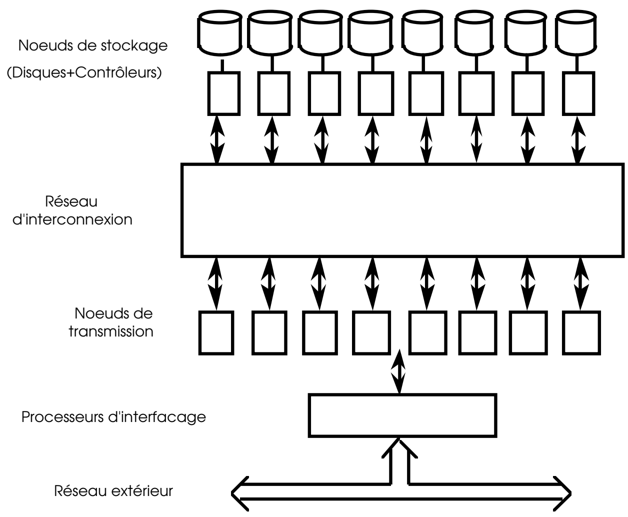

Figure 1. Architecture du serveur

Deux politiques de placement des blocs sont principalement utilisées : le placement rotatoire (*round-robin*) et le placement aléatoire (*random*). Dans le placement rotatoire les blocs successifs d'un même fichier sont placés sur des disques consécutifs. Dans le placement aléatoire, les blocs d'une séquence sont placés en utilisant une permutation aléatoire avec la contrainte de ne pas placer deux blocs successifs sur le même disque. Les avantages et inconvénients de ces deux politiques sont exposés dans [Tewari 96]. Pour cette étude nous considérons le placement aléatoire qui conduit à des résultats très comparables à ceux obtenus par un placement rotatoire pour le système étudié.

Dans la suite, le critère de performance (probabilité de retard) sera estimé en fonction des deux paramètres internes de fonctionnement : taille des blocs suivant laquelle seront fragmentés les fichiers et nombre de blocs d'anticipation (A).

3. Approche analytique

Nous présentons un modèle analytique simplifié qui permet d'étudier l'effet des deux paramètres internes de fonctionnement : la taille des blocs et le nombre d'anticipation. Les simulations du fonctionnement exact du système sont très coûteuses en temps de calcul. Le modèle proposé permettra de réaliser rapidement un premier dimensionnement des paramètres de fonctionnement. Ces valeurs peuvent être affinées par la suite par de longues simulations sur le modèle précis.

La symétrie dans l'architecture et le bon l'équilibrage de la charge entraînent une forte symétrie dans le fonctionnement du système. Les nœuds de stockage peuvent alors être considérés comme équivalents sur le plan opérationnel. Il en est de même pour les nœuds de transmission. Cette observation est très intéressante puisqu'elle permet de n'étudier les résultats que pour un couple de nœuds. Nous ferons une approximation d'indépendance et nous étudierons un couple isolé de nœuds en simplifiant les interactions avec les autres couples.

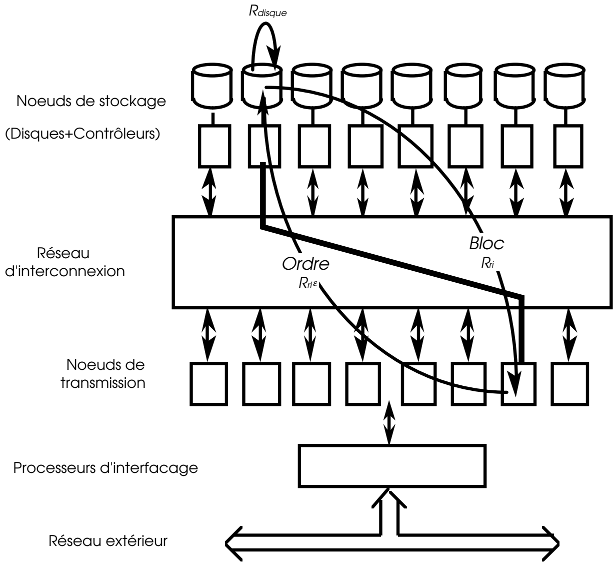

Figure 2. Cycle de transmission

Comme nous nous intéressons à la qualité de la transmission, nous allons essayer dans ce qui suit d'exprimer analytiquement la probabilité pour que la transmission d'un bloc cause une rupture momentanée de séquence. Pour ce faire nous allons analyser les différentes opérations impliquées dans la transmission d'une séquence.

Suite à la phase d'anticipation la transmission de la séquence se fait d'une manière cyclique. Un cycle de transmission correspond à l'extraction d'un bloc et son envoi sur le réseau extérieur. Les opérations du cycle de transmission se produisent entre le noeud de transmission qui gère la séquence, le disque contenant le bloc à extraire et le noeud de stockage auquel est associé ce disque. Les opérations qui constituent le cycle sont au nombre de trois. La première opération consiste en la génération et l'envoi de l'ordre d'extraction du noeud de transmission vers le noeud de stockage adéquat. C'est l'étape de traversée aller du RI par l'ordre d'extraction. Elle engendre une durée de traversée $R_{ri\xi}$, temps de réponse du RI dans le sens aller ou sens ordre d'extraction. La deuxième étape est l'extraction du bloc du disque. Elle se déroule au niveau du disque et est composée d'une partie attente devant la ressource et une partie lecture du bloc. Cette étape engendre une durée globale R_{disque}, temps de réponse du disque. La troisième étape est l'envoi du bloc depuis le noeud de stockage vers le noeud de transmission demandeur. C'est l'étape de traversée retour du RI. Le bloc empruntera le chemin inverse de l'ordre d'extraction. Cette étape engendre une durée R_{ri}, temps de réponse du RI dans le sens retour ou sens bloc. La durée totale du cycle est alors la somme de ces trois v.a. temps de réponse : $R_{ri\xi}$, R_{disque} et R_{ri}.

Il y a une rupture momentanée de séquence quand le bloc n'arrive pas au noeud de transmission à la date prévue pour son envoi sur le réseau extérieur. Ceci se produit lorsque le cycle de transmission décrit ci-dessus dure plus que la période qui lui est allouée Δ_{cycle}. La durée allouée au cycle correspond à l'intervalle de temps qui sépare la date d'envoi de l'ordre d'extraction du bloc de la date prévue pour son envoi sur le réseau extérieur. Cette période n'est autre que le produit de la durée de transmission d'un bloc par le nombre d'anticipation :

$$\Delta_{cycle} = A \times \Delta_{émission} \qquad (1)$$

Cette période est constante pour une taille de bloc et un nombre d'anticipation donnés.

Ainsi il y a une rupture momentanée de séquence chaque fois que pour un bloc : $R_{ri\xi} + R_{disque} + R_{ri} > \Delta_{cycle}$. La probabilité π pour que la transmission d'un bloc occasionne une rupture momentanée de séquence s'écrit alors : $\pi = \Pr(R_{ri\xi} + R_{disque} + R_{ri} > \Delta_{cycle})$.

Sachant que la durée de traversée du RI est fortement liée à la taille de l'entité traitée et vue la différence de taille entre un ordre d'extraction (64 octets) et celle d'un bloc de donnée (+128 Koctets) la durée de traversée de l'ordre d'extraction $R_{ri\xi}$ sera négligée dans la suite de l'étude.

La probabilité de retard s'écrit alors :

$$\pi = \Pr(R_{disque} + R_{ri} > \Delta_{cycle}) \qquad (2)$$

Une résolution exacte du problème nécessite le calcul de la convolution des densités de probabilité des deux v.a. temps de réponse disque R_{disque} et temps de réponse RI (pour un bloc) R_{ri} en faisant l'approximation de l'indépendance de ces deux v.a.. Ce calcul est très compliqué. En effet les densités de probabilité des temps de réponse sont très difficiles à obtenir et particulièrement celle du RI.

Nous avons contourné cette difficulté en approchant la v.a. temps de réponse du RI par une constante égale à sa moyenne $\overline{R}_{ri}$. Le problème est alors ramené au calcul de la fonction de répartition du temps de réponse des disques. L'équation (2) devient :

$$\pi = \Pr(R_{disque} > \Delta_{cycle} - \overline{R}_{ri}) \qquad (3)$$

Il nous faudra donc obtenir une bonne approximation de la fonction de répartition du temps de réponse disque.

La résolution de cette équation ainsi que la justification et validation des hypothèses simplificatrices sont présentées dans l'étude détaillée qui suit.

4. Etude des disques

Présentons dans cette partie l'étude menée sur les composants disques de notre système. Nous commencerons par la présentation de l'architecture des disques, de leur fonctionnement et des différents modèles proposés dans la littérature. Nous présenterons ensuite la démarche adoptée pour notre modélisation, le modèle final simplifié, sa résolution et sa validation.

4.1. Architecture et fonctionnement

Rappelons l'architecture et le fonctionnement des disques afin de pouvoir les modéliser. Un disque est composé de plusieurs plateaux tournant continûment autour d'un axe, à une vitesse angulaire constante. Les données sont inscrites sur la surface des plateaux et on y accède par des têtes de lecture au moyen de bras alignés verticalement et pivotant autour d'un axe (Fig.3).

Sur un plateau les données sont organisées en cercles équidistants ayant l'axe du disque comme centre commun. L'ensemble des données placées à égale distance de l'axe forme une piste pour un plateau et un cylindre pour le disque (Fig.3). Sur une piste, les données sont groupées par secteurs. Tous les secteurs du disque ont la même taille. Le secteur est la quantité minimale d'octets transférés pour chaque lecture.

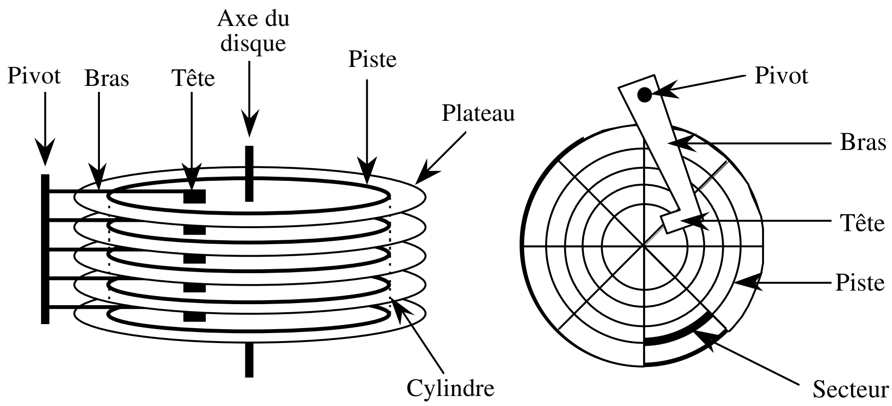

Figure 3. Architecture d'un disque

L'opération de lecture (ou d'écriture) de données se fait en trois étapes. La première étape consiste à déplacer la tête de lecture jusqu'à la bonne piste. C'est l'opération d'accès qui engendre une latence d'accès T_{seek} (*seek latency*). La deuxième étape est une opération de positionnement. Pour commencer la lecture il faut attendre que la donnée passe sous la tête. Cette étape engendre une latence de rotation T_{pos} (*rotation latency*). La troisième étape est l'opération de lecture proprement dite. Cette étape a une durée T_{lect} . La durée totale T_{tot} de l'opération de lecture d'un bloc sur un disque s'écrit alors comme la somme de ces trois durées, $T_{tot} = T_{seek} + T_{pos} + T_{lect}$. Notons que le disque ne peut faire qu'une lecture à la fois, en effet les différentes têtes disposent d'un seul canal qui est alloué à une seule tête à un instant donné.

4.2. Modèles de disques

Suivant le degré de précision recherché plusieurs modèles sont proposés dans la littérature. Notons qu'en raison de leur complexité la plupart des modèles qui se veulent précis sont uniquement étudiés par simulation. La modélisation du fonctionnement du disque en lecture consiste en la modélisation des trois étapes (accès, positionnement et lecture) décrites ci-dessus et qui constituent l'opération élémentaire de lecture d'un bloc.

La modélisation de l'opération d'accès consiste en la modélisation du mouvement de déplacement du bras entraînant les têtes de lecture. Le mouvement du bras et composé : d'une phase d'accélération, d'une phase uniforme si la vitesse maximale est atteinte et d'une phase de ralentissement. Suivant les caractéristiques du disque (accélération du bras et rayon du disque) et le nombre de cylindres parcourus, ces étapes sont plus ou moins importantes les unes par rapport aux autres. La durée T_{seek} de cette étape d'accès sera le cumul des durées de ces trois phases. Les phases accélérée et décéléré font apparaître un terme en $\sqrt{D}$, la phase uniforme fait apparaître un terme en D où D représente le nombre de cylindres parcourus pendant chacune de ces phases. Il existe dans la littérature principalement trois modèles précis. Si on désigne par D le nombre total de cylindres parcourus lors du déplacement, ces modèles se présentent comme suit : $T_{seek} = \alpha\sqrt{D} + \beta$ [Hennesy 90] [Scranton 83], $T_{seek} = \alpha\sqrt{D} + \beta D + \gamma$ [Lee Edward 93] et $T_{seek} = \begin{cases} \alpha_1\sqrt{D} + \beta_1 , & D < D_0 \\ \alpha_2 D + \beta_2 , & D \geq D_0 \end{cases}$ [Ruemmler 94]. Il existe aussi des modèles simples qui consistent à modéliser cette étape par un temps constant égal à la moyenne de cette opération observée sur plusieurs déplacements. Là aussi plusieurs méthodes sont proposées pour estimer cette moyenne.

La durée de l'opération de positionnement est généralement modélisée par une loi uniforme entre 0 et la durée d'une rotation complète du disque ou simplement par la moyenne de cette loi, soit la moitié de la durée d'une rotation complète.

L'opération de lecture a une durée T_{lec} constante qui dépend du débit de lecture D_{lec} du disque et de la taille du bloc t_{bloc}, $T_{lec} = t_{bloc} / D_{lec}$.

On voit que les modèles précis aboutissent à une partie variable (accès + positionnement) et une partie constante (lecture). Les modèles simples consistent à approcher la durée totale de l'opération par une constante, somme des moyennes des trois opérations.

4.3. Modélisation

Pour notre étude, vue l'importance des disques (goulets d'étranglement) dans les performances du système, nous avons tout d'abord cherché à utiliser pour le temps de service disque un modèle assez précis. Par souci de simplification de la résolution analytique nous avons ramené ensuite ce modèle à un modèle simple à temps de lecture constant. Nous présentons ci-dessous le modèle de départ et nous exposons la démarche de simplification entreprise tout en la justifiant.

4.3.1. Modèle adopté

Chaque disque est modélisé par une file d'attente. Nous supposons que les arrivées d'ordres d'extraction sont poissonniennes. Cette hypothèse est justifiée par le grand nombre de requêtes émises et par le placement aléatoire des blocs sur les disques. Cette hypothèse sera validée par la suite.

Pour le temps de service, le modèle précis est le suivant :

$$T_{tot} = T_{seek} + T_{pos} + T_{lec} \quad \text{avec} \quad \begin{cases} T_{seek} = \alpha\sqrt{D} + \beta\,, \quad D > 0 \\ T_{pos} = unif[0,\gamma] \\ T_{lec} = \chi = t_{bloc} / D_{lec} \end{cases}$$

où D est le nombre de cylindres parcourus par la tête de lecture. α et β sont des constantes déterminées de sorte à retrouver T_{seek} minimal (déplacement d'un cylindre) et T_{seek} maximal (déplacement de tous les cylindres) donnés par le constructeur. γ est la durée d'une rotation entière du disque. Cette durée est directement déduite de la vitesse de rotation donnée par le constructeur. Les données techniques des disques utilisés sont résumées dans le tableau 1. Nous avons considéré les Deskstar3 d'IBM, disques d'actualité adaptés à ce type d'applications [IBM].

La distance parcourue D correspond à la différence entre la position du bras au moment de deux lectures consécutives. $D = \Delta X = |X_n - X_{n-1}|$ où X représente le numéro du cylindre sur lequel se trouve le bloc considéré. X est une variable discrète, mais vu le grand nombre de cylindre (~3000) nous allons relaxer cette contrainte afin de faciliter les calculs par la suite. Si l'on considère que la densité d'information est la même sur tout le disque, on peut alors prendre pour X une loi uniforme entre 0 et N, N+1 étant le nombre total de cylindres. Les ΔX successifs vont être corrélés. On peut néanmoins considérer qu'ils sont indépendants et supposer que X_n et X_{n-1} suivent deux lois uniformes indépendantes entre 0 et N.

Nous aboutissons ainsi à une file M/G/1 avec un service $B = T_{tot}$. Rappelons que le but de cette étape est de déterminer la densité de probabilité du temps de réponse des disques afin de calculer le critère de performance : probabilité de retard (cf. équation 3).

capacité	2 Go
rotation	5400 trs/min
D_{lec}	5,2 Mo/s
accès minimal	0,6 ms
accès maximal	17,4 ms
nb de cylindres	2870

Tableau 1. Données techniques des disques

4.3.2. Etude du temps de service

Dans ce qui suit étudions la densité de probabilité $b(x)$ du temps de service et ses premiers moments.

Remarquons tout d'abord que le temps d'accès peut se mettre sous la forme $T_{seek} = \Theta_s + \beta$ avec $\Theta_s = \alpha\sqrt{\Delta X}$.

Il vient alors : $f_{\Theta_s}(y) = \frac{2y}{\alpha^2} f_{\Delta X}\left(\left(\frac{y}{\alpha}\right)^2\right)$

Sachant que : $f_{X_n}(x) = \frac{1}{N} 1_{\{0 \le x \le N\}}$ on trouve : $f_{\Delta X}(x) = \frac{2(N-x)}{N^2}$

D'où : $f_{\Theta_s}(y) = \frac{2y}{\alpha^2}\left\{\frac{2}{N} - \frac{2}{N^2}\left(\frac{y}{\alpha}\right)^2\right\}$ et $F_{\Theta_s}(y) = \frac{2}{\alpha^2 N} y^2 - \frac{1}{\alpha^4 N^2} y^4$

On en déduit : $E[\theta_s] = \frac{8}{15}\alpha\sqrt{N}$, $Var[\theta_s] = \frac{11}{225}\alpha^2 N$

Pour le temps de positionnement T_{pos} , on a immédiatement : $f_{T_{pos}}(x) = \frac{1}{\gamma} \quad 1_{\{0 \le x \le \gamma\}}$

ainsi que les premiers moments : $E[T_{pos}] = \frac{\gamma}{2}$, $Var[T_{pos}] = \frac{\gamma^2}{12}$

On obtient alors les premiers moments de la distribution du temps de service :

$$E[B] = \frac{8}{15}\alpha\sqrt{N} + \beta + \frac{\gamma}{2} + \chi \qquad (4)$$

$$Var[B] = \frac{11}{225}\alpha^2 N + \frac{\gamma^2}{12} \qquad (5)$$

La densité de probabilité de $\Theta_s + T_{pos}$ se calcule par : $f_{\Theta_s + T_{pos}}(x) = f_{\Theta_s}(x) \otimes f_{T_{pos}}(x)$. Pour les valeurs numériques considérées, on a $\gamma \le \alpha\sqrt{N}$, il vient alors :

$$f_{\Theta_s + T_{pos}}(x) = \begin{cases} \frac{1}{\gamma}\left[\frac{2}{\alpha^2 N}x^2 - \frac{1}{\alpha^4 N^2}x^4\right] & , 0 \le x \le \gamma \\ \frac{1}{\gamma}\left[\left\{\frac{2}{\alpha^2 N}x^2 - \frac{1}{\alpha^4 N^2}x^4\right\} - \left\{\frac{2}{\alpha^2 N}(x-\gamma)^2 - \frac{1}{\alpha^4 N^2}(x-\gamma)^4\right\}\right] & , \gamma \le x \le \alpha\sqrt{N} \\ \frac{1}{\gamma}\left[1 - \left\{\frac{2}{\alpha^2 N}(x-\gamma)^2 - \frac{1}{\alpha^4 N^2}(x-\gamma)^4\right\}\right] & , \alpha\sqrt{N} \le x \le \gamma + \alpha\sqrt{N} \end{cases}$$

La densité de probabilité du temps de service s'obtient par :

$$b(x) = f_{\Theta_s + T_{pos}}(x - \beta - \chi) \quad \text{pour} \quad \beta + \chi \le x \le \alpha\sqrt{N} + \beta + \gamma + \chi \qquad (6)$$

4.3.3. Résolution du modèle

Soient λ le taux d'arrivée, $b = E[B]$ le temps moyen de service et $\rho = \lambda b$ la charge de la file. Nous noterons par R et W respectivement le temps de réponse et le temps d'attente de la file. Pour alléger les équations on posera systématiquement $Y(x)$ la fonction de répartition de la variable aléatoire Y, $y(x)$ sa densité de probabilité et $Y^*(s)$ sa transformée de Laplace. On notera les différents moments de la distribution de la manière suivante : $y_i = E[Y^i]$.

On a alors (formule de Pollaczek-Kinchin) : $R^*(s) = B^*(s)W^*(s) = \dfrac{(1-\rho)sB^*(s)}{s-\lambda+\lambda B^*(s)}$

Cette transformée de Laplace est le plus souvent impossible à inverser formellement. C'est le cas, en l'occurrence pour le service décrit par le système (6). Pour obtenir la probabilité de retard nous allons approcher le temps de service par un temps de service déterministe égal à sa moyenne (cf. équation 4) et de là calculer la fonction de répartition du temps de réponse. La probabilité de retard est alors donnée par l'équation (3). Le modèle du disque est ainsi ramené à une file M/D/1. On validera l'approximation en comparant la fonction de répartition obtenue à celle donnée par simulation sur le modèle exact.

Dans le cas de la file M/D/1, on peut facilement calculer la valeur de la fonction de répartition du temps de réponse aux points kb. Pour cela, constatons que si la file est vide quand un client arrive, son temps de réponse sera égal à b. S'il y a k clients à son arrivée, son temps de réponse vaudra : $kb+\Omega$ où Ω est le temps résiduel de service.

Comme Ω est compris strictement entre 0 et b, le temps de réponse sera compris strictement entre kb et $(k+1)b$. Les arrivées étant poissonniennes, la distribution limite stationnaire aux instants d'arrivée sera la même que la distribution limite stationnaire à un instant d'arrivée. Par conséquent, si l'on note π_k, la probabilité limite stationnaire qu'il y ait k clients à l'arrivée du client, on aura :

$$\begin{cases} R((k+1)b) - R(kb) = \pi_k \ , \ k \geq 1 \\ R(b) = \pi_0 \end{cases}$$

D'où :

$$R(kb) = \sum_{j=0}^{k-1} \pi_j \quad , \ k \geq 1 \qquad (7)$$

Les π_k sont données par : $$\begin{cases} \pi_0 = 1-\rho \\ \pi_1 = (1-\rho)(e^{\rho}-1) \\ \pi_k = (1-\rho)\left\{\sum_{j=1}^{k-1}(-1)^{k-j} e^{j\rho}\left[\dfrac{(k\rho)^{k-j}}{(k-j)!} + \dfrac{(k\rho)^{k-j-1}}{(k-j-1)!}\right] + e^{k\rho}\right\}, \ k \geq 2 \end{cases}$$

Dans notre étude, nous aurons besoin de valeurs en des points qui ne sont pas forcément des multiples de b. Pour ce faire nous utilisons l'approximation :

$$Log\{R(kb+\upsilon)\} = \frac{1-\upsilon}{b} Log\{R(kb)\} + \frac{\upsilon}{b} Log\{R((k+1)b)\} \ , \quad 0 \leq \upsilon \leq b$$

Avant de passer à l'évaluation du modèle retenu, nous présentons les arguments qui justifient a priori la simplification effectuée.

4.3.4. Justification de la simplification

Pour bien utiliser un disque il faut choisir une taille de bloc assez grande car le débit global du disque croit avec la taille des blocs manipulés. Ceci est dû au fait que la durée de l'opération de lecture croit avec la taille du bloc alors que celles des opérations d'accès et de positionnement n'en dépendent pas. Par conséquent la proportion de la durée de l'opération de lecture p_{lec} dans le temps de service disque croit avec la taille des blocs. Lorsque la taille des blocs croit, le disque est donc mieux utilisé et le débit global augmente.

Le débit global du disque ($D_g = t_{bloc} / E[T_{seek} + T_{pos} + T_{lec}]$) s'écrit : $D_g = p_{lec} \times D_{lec}$ avec $p_{lec} = E[T_{lec}] / E[T_{seek} + T_{pos} + T_{lec}]$. En développant cette expression on trouve que p_{lec} s'écrit comme une fonction hyperbolique croissante de t_{bloc} :

$$p_{lec} = t_{bloc} / (t_{bloc} + v),\ v \text{ constante} \qquad (8)$$

p_{lec} admet 1 comme limite supérieure, ce qui correspond à un débit global maximum égal au débit de lecture D_{lec}.

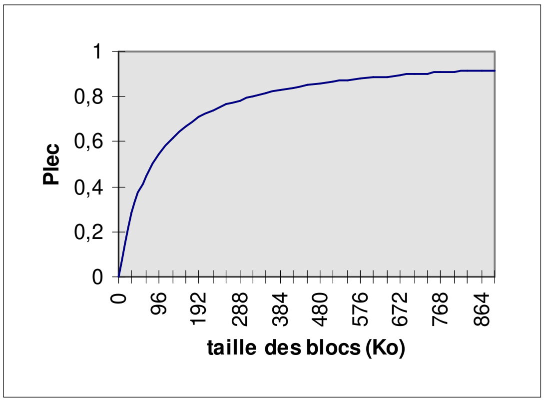

Figure 4. p_{lec} fonction de la taille des blocs

On voit sur la figure 4 que la proportion de lecture devient intéressante pour les tailles de bloc supérieures à 128 Ko qui engendre une proportion de lecture de 0,6. Dans la suite de l'étude nous nous intéresserons à des tailles supérieures à ce seuil.

Pour évaluer l'effet de la partie variable dans le temps total de service du disque, nous avons tracé son coefficient de variation $\sqrt{Var[B]} / E[B]$ en fonction de la taille des blocs. Cette variance a été calculée grâce aux équations (4) et (5). Nous constatons sur la figure 5 que ce coefficient est décroissant avec la taille des blocs et est inférieur à 0,12 pour les tailles qui nous intéressent (supérieures à 128 Ko). Ceci appuie l'hypothèse simplificatrice qui consiste à prendre le temps de service constant.

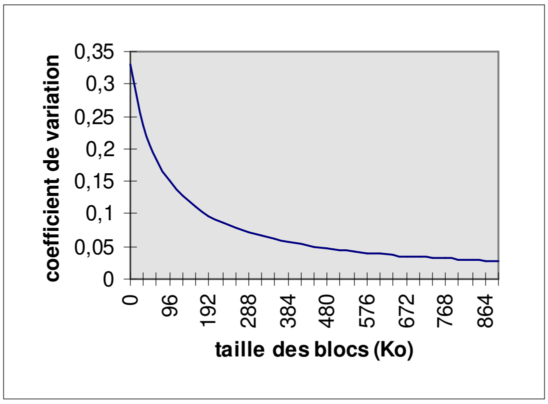

Figure 5. Coefficient de variation de B en fonction de la taille des blocs

Dans le même souci de justification de l'approximation effectuée, nous présentons ci dessous un tableau comparatif des quatre premiers moments du temps de réponse des disques pour chacun des deux modèles : modèle précis et modèle simplifié déterministe sous l'hypothèse d'arrivées poissonniennes. Les moments de la distribution du temps de réponse sont obtenus en dérivant successivement la transformée de Laplace suivante (formule de Takacs) : $\begin{cases} w_k = \dfrac{\lambda}{1-\rho} \displaystyle\sum_{j=1}^{k} \binom{k}{j} \dfrac{b_{j+1}}{j+1} w_{k-j} \\ w_0 = 1 \end{cases}$

Les différents moments du temps de réponse valent alors : $r_k = \sum_{j=0}^{k} \binom{k}{j} b_j w_{k-j}$

Pour un temps de service déterministe, nous avons : $b_k = b_1^k = b^k$

Pour le modèle précis, notons v_k les moments de la distribution de $T_{pos} + \theta_s$, γ_k ceux de T_{pos} et η_k ceux de θ_s . Il vient :

$$b_k = \sum_{j=0}^{k} \binom{k}{j} v_j (\beta + \chi)^{k-j} \quad \text{où} \quad v_k = \sum_{j=0}^{k} \binom{k}{j} \gamma_j \eta_{k-j} \quad \text{avec} \quad \gamma_k = \frac{\gamma^k}{k+1} \quad \text{et} \quad \eta_k = 4\left\{\frac{1}{k+2} - \frac{1}{k+4}\right\}\left(\alpha\sqrt{N}\right)^k$$

On voit sur le tableau 2 que les moments pondérés du temps de réponse pour le modèle précis et pour le modèle déterministe sont assez proches ce qui vient encore appuyer l'approximation.

	$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
moment	Précis	Détermi.	Précis	Détermi.	Précis	Détermi.
1	1.507	1.500	2.183	2.167	5.562	5.500
2	2.889	2.833	6.988	6.833	54.894	53.500
3	7.275	7.000	31.774	30.555	806.955	775.000
4	23.713	22.367	191.433	181.004	15813.0	14695.3

Tableau 2. Premiers "moments" r_k / b^k du temps de réponse

4.5. Evaluation du modèle simplifié

Nous passons maintenant à l'évaluation du modèle simplifié en comparant le temps de réponse qu'il engendre au temps de réponse obtenu par le modèle précis. Pour ce faire nous comparons la fonction de répartition du temps de réponse calculé analytiquement pour M/D/1 (cf. équation 7) à celui obtenu par simulation sur le modèle précis avec arrivées poissonniennes.

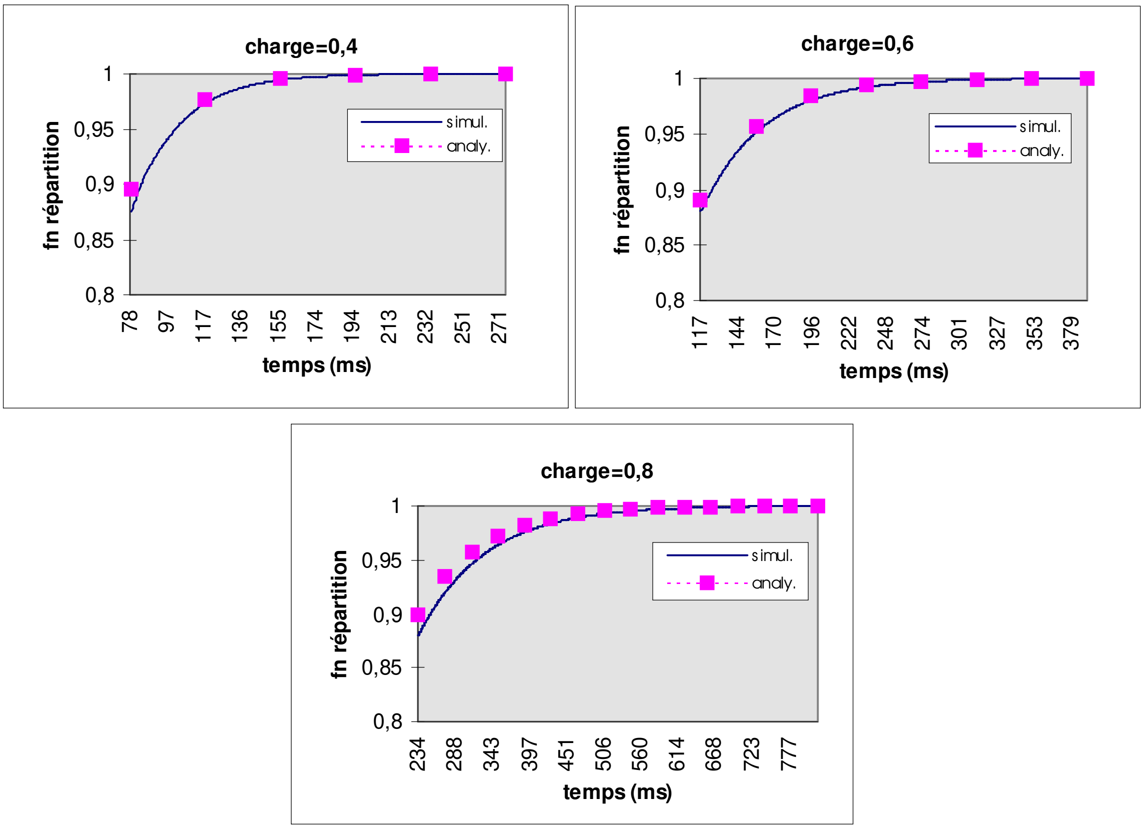

Figure 6. Fonction de répartition du temps de réponse

Nous avons dressé trois familles de courbes pour différentes charges (Fig.6). Ces courbes montrent que le modèle simplifié donne des résultats assez proches de ceux obtenus sur le modèle précis. Nous nous sommes limités aux valeurs élevées de la fonction de répartition du temps de réponse du disque puisque nous nous intéressons à un fonctionnement avec de faibles probabilités de retard de blocs (cf. équation 3).

5. Etude du réseau d'interconnexion

Le développement des architectures parallèles a beaucoup contribué au développement des réseaux d'interconnexion. Les réseaux d'interconnexion permettent de connecter les différents composants d'une machine parallèle : processeurs, mémoires, disques, etc. Nous nous intéressons dans cette étude à un réseau multiétage de type Oméga. Une étude bibliographique sur les réseaux multiétages est présentée dans [Siegle 87].

5.1. Architecture et fonctionnement

Un réseau multiétage est constitué par des éléments de commutation (commutateurs) interconnectés et organisés en étages. Chaque étage est connecté à l'étage suivant en respectant des règles de connexion qui caractérisent l'architecture considérée. Dans beaucoup de réseaux classiques les commutateurs sont symétriques avec le même nombre d'entrées et de sorties : a . Pour de tels réseaux à n étages, le nombre d'entrées est égal au nombre de sorties et égal à a^n . Le réseau Oméga est un réseau monochemin, entre une entrée et une sortie quelconques il existe un chemin unique. Les éléments de commutations considérés ont 2 entrées et 2 sorties ($a = 2$).

Nous supposons un fonctionnement avec un routage de messages (store-and-forward). Les blocs avancent dans le réseau vers leurs destinations en transitant dans les commutateurs intermédiaires. Lorsqu'un bloc arrive à un commutateur, il est stocké dans sa mémoire jusqu'à libération du lien à emprunter. A chaque étape le lien emprunté est aussitôt libéré. La taille des blocs implique que le délai de traitement induit dans la traversée d'un étage est principalement dû au temps de transmission sur le lien.

5.2. Modélisation

Les critères de performances généralement recherchés lors de l'étude de tels réseaux d'interconnexion sont : le temps de traversée et le débit maximum du système. Pour l'architecture retenue le débit maximal est égal au débit des liens multiplié par le nombre d'entrées si nous considérons qu'il n'y a pas de perte causée par la limitation des mémoires des commutateurs.

Le système peut être modélisé par un réseau de file d'attente. Chaque commutateur est modélisé par deux files à la sortie. Le temps de service de ces files correspond alors au temps de traversée sur un lien. Pour une taille de bloc donnée t_{bloc} ce service est constant égal à $T_{lien} = t_{bloc} / D_{lien}$, D_{lien} étant le débit du lien.

La résolution analytique d'un tel modèle est rendue difficile à cause de la dépendance entre les étages. Alors que nous trouvons beaucoup de travaux sur les réseaux à fonctionnement synchrones à cause de leurs utilisation dans les télécommunications (ATM), peu de travaux sont disponibles pour les réseaux d'interconnexion à temps continu qui nous intéressent. Les études disponibles sont généralement limitées à la détermination du temps moyen de traversée. [Mohapatra 96] présente un bref état de l'art sur le sujet.

Pour cette étude nous avons considéré des commutateurs avec des débits bidirectionnels de 20 Mo/s. On trouve actuellement sur le marché des commutateurs avec des débits bidirectionnels de 50 Mo/s (l'Elite de Pci) voire 200 Mo/s (Paragon Mesh Routing Chip d'Intel) [Intel 91].

Notons qu'avec les données techniques d'actualité choisies pour les disques et pour les commutateurs, la charge du RI reste inférieure à 0,2. Le débit global maximum d'un disque est égal à $D_{lec} = 5{,}2 Mo / s$ ou $D_{lec} / t_{bloc} \; (blocs / s)$ (cf. équation 8). Le débit maximum de l'ensemble des disques est alors :

$nb.\ disques \times D_{lec} / t_{bloc}\ (blocs / s)$. Le débit d'un lien de commutateur est $D_{lien} = 20\,Mo / s$. Le débit du réseau en entier est égal au nombre de ports d'entrée multiplié par ce débit ou encore $nb.\ noeuds\ de\ stockage \times D_{lien} / t_{bloc}\ (blocs / s)$. Si nous considérons qu'il y a un disque par noeud de stockage on aura : $nb.\ disques = nb.\ noeuds\ de\ stockage$ et le goulet d'étranglement est alors situé au niveau des disques. Pour une charge de disque inférieure à 0,8 la charge du RI est inférieure à 0,2.

Si on veut utiliser le RI avec une charge supérieure, on peut augmenter le nombre de disques par nœuds de stockage. Un tel fonctionnement induirait des conflits au niveau du noeud de stockage et une dégradation du temps de traversée du RI. L'utilisation du réseau avec une charge forte induit des temps de traversée élevés et dispersés à cause des problèmes de contention qui détérioreraient les performances du système [Lee Gyungho 89]. On voit sur la figure 7 la forte croissance du temps moyen de traversée et de sa variance pour une charge supérieure à 0,5. Ces résultats sont obtenus sur un réseau à 64 entrées/sorties, soit 6 étages de commutation, pour des blocs de 128 Ko.

Compte tenu du matériel disponible actuellement (performances/coût), l'architecture choisie est telle que le goulet d'étranglement de ces systèmes se trouve au niveau des disques. Pour optimiser l'utilisation de l'ensemble du serveur, on est amené à utiliser le RI avec une charge faible. Les simplifications qui suivent sont alors généralement applicables pour tous les serveurs ayant cette architecture.

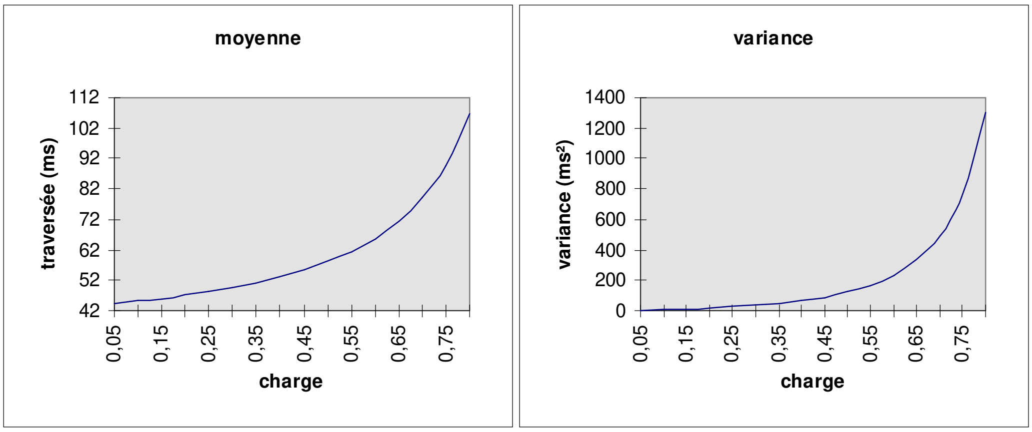

Figure 7. Temps de traversée et sa variance

L'étude du RI dans les conditions de fonctionnement qui nous intéressent (faible charge et confrontation avec les disques) montre que nous pouvons réduire le réseau en entier à un simple délai égal à la moyenne de son temps de traversée. En effet pour les charges considérées nous avons un coefficient de variation qui est inférieure à 10^{-1}. La confrontation de la variance du temps de traversée du réseau d'interconnexion à celle du temps de réponse des disques montre que la part de variation induite par le réseau dans la durée totale de traversée (cycle de transmission) est négligeable par rapport à celle causée par les disques. Le rapport de ces variances varie entre 10^{-3} et 10^{-2} pour les valeurs de charge qui nous intéressent. Cette variation dans le temps de traversée du RI peut alors être négligée. Ces observations nous ont amenés à modéliser le RI par un délai constant.

Dans la suite de notre étude le réseau d'interconnexion sera remplacé par un simple délai égal à la moyenne de son temps de traversée. La validation de la modélisation du réseau sera présentée dans la suite lors de la validation du système en entier. Nous présentons dans ce qui suit le calcul de cette moyenne et la validation de ce calcul.

5.3. Calcul du délai de traversée et validation

L'étude détaillée du réseau d'interconnexion montre que le temps moyen de traversée d'un étage se stabilise pour devenir constant à partir d'un certain rang. Ce résultat est exposé dans [Mohapatra 96], il s'explique par l'établissement d'une certaine régularité dans le trafic sortant des étages à partir d'un certain rang. Pour les faibles charges qui nous intéressent nous constatons que ce temps de traversée est pratiquement le même pour tous les étages sauf pour le premier où nous notons une légère différence.

L'ensemble du réseau d'interconnexion est modélisé par un délai de valeur constante égale au nombre d'étages multiplié par le temps de réponse moyen d'une file M/D/1 modélisant un étage. Ce dernier est un résultat classique [Jain 92]. Cette approximation est validée par comparaison avec la simulation. Les résultats présentés sur la figure 8 sont relatifs a un réseau à 64 entrées/sorties, soit 6 étages de commutateurs, pour des blocs de 128 Ko.

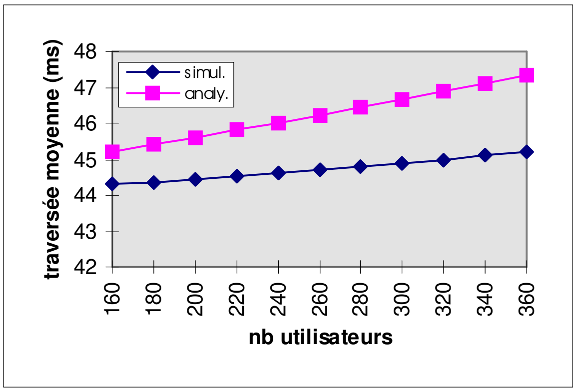

Figure 8. Temps moyen de traversée global du réseau d'interconnexion

On constate (Fig. 8) que l'erreur commise dans l'estimation du temps de traversée du RI en utilisant l'approximation M/D/1 pour la détermination du temps de traversée de chaque étage est faible dans les conditions de fonctionnement considéré (erreur relative maximale = 4%).

6. Modèle de simulation

Pour valider cette approche analytique nous avons développé un modèle précis du système en entier sous forme de réseau de files d'attente. Le modèle tient compte des principales opérations qui interviennent dans le fonctionnement du système. L'approche de modélisation a été validée dans une étude antérieure [kaddeche 96] sur un prototype développé à cet effet. Les simulations ont étés réalisées avec le simulateur QNAP2 (*Queueing Networks Analysis Package*) [POT 84].

La résolution est effectuée en réseau fermé : nous avons un nombre fixe d'utilisateurs simultanés constamment actifs. Dès que le système a terminé avec la transmission d'une séquence, fin de la séquence ou demande d'interruption par l'utilisateur, une nouvelle requête est générée. Cette requête peut provenir de ce même utilisateur ou d'un utilisateur nouvellement connecté. Le choix de la séquence à visualiser est effectué suivant une loi uniforme modélisant la demande des utilisateurs. Pour avoir une bibliothèque de séquences assez variée, nous avons considéré 300 séquences dont les durées sont choisies au départ suivant une loi exponentielle de moyenne 10 minutes. Pour tenir compte du comportement de l'utilisateur nous avons considéré qu'il interrompt la visualisation de la séquence, si celle-ci n'est pas terminée, au bout d'un temps qui suit une loi exponentielle de moyenne 6 minutes. Les simulations sont effectuées sur un serveur de 64 disques à raison d'un disque par noeud de stockage.

La probabilité de retard est obtenue en divisant le nombre de blocs arrivés en retard par le nombre total de blocs arrivés.

7. Validation et utilisation du modèle analytique

La figure 9 présente les probabilités de retard de blocs obtenues par simulation (courbes en pointillé) et par la méthode analytique approchée (courbes en continu). Le serveur considéré a 64 disques à raison d'un disque par noeud de stockage. La taille des blocs est de 128 Ko. Le nombre d'anticipation varie entre 1 et 3. Les nombres d'utilisateurs considérés varient entre 160 et 340 ce qui pour la taille de blocs considérée représente une charge du disque comprise entre 0.4 et 0.85.

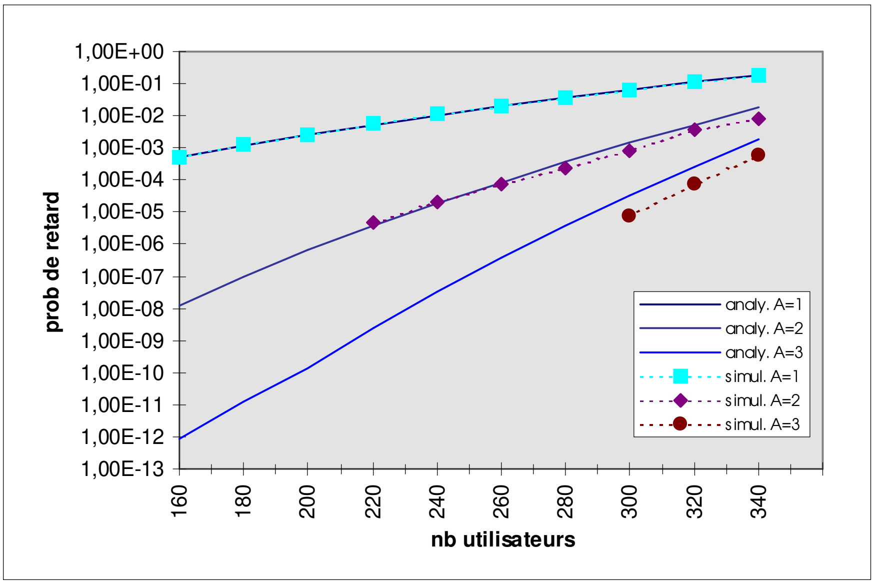

Figure 9. Comparaison méthode analytique - Simulations

La comparaison entre la simulation et le modèle analytique approché peut être effectuée lorsque les simulations sont possibles, c'est à dire lorsque les retards ne sont pas trop rares. Cette comparaison montre que le modèle analytique approché donne des résultats valides.

L'analyse de ces courbes montre que les probabilités de retard augmentent avec la charge et diminuent lorsque le nombre d'anticipations augmente. Lorsque la charge augmente les temps de réponse des disques et du réseau d'interconnexion augmentent. La durée du cycle de transmission (extraction sur disque + traversée du réseau d'interconnexion) du bloc augmente, ce qui provoque l'augmentation du nombre de blocs retardés. L'anticipation permet d'augmenter la durée allouée à l'opération de transmission d'un bloc qui est directement proportionnelle à ce nombre (cf. équation 1). Ainsi l'augmentation du nombre d'anticipation favorise le bon déroulement de l'opération de transmission.

Pour une taille de bloc de 128 Ko on peut fixer la probabilité de retard maximum à 10^{-4} , ce qui se traduit au niveau de l'utilisateur par au maximum une interruption de 250 ms toutes les 41 minutes, interruption non sensible pour ce type d'application. On voit sur la figure 9 que de telles probabilités de retards sont obtenues pour une charge inférieure à 260 et 300 utilisateurs respectivement pour 2 et 3 anticipations. Avec une seule anticipation les probabilités de retards sont toujours supérieures à ce seuil pour le domaine d'utilisation étudié. Pour une taille de bloc donnée on choisit ainsi le nombre d'anticipation qui nous permet d'atteindre une charge (en nombre d'utilisateurs) avec une probabilité de retard seuil fixée en fonction de la qualité de transmission requise. Notons qu'on peut calculer une borne supérieure théorique de la charge par le rapport du débit total de tous les disques par le débit requis pour une séquence. Pour les données techniques considérées cette borne vaut 400 utilisateurs.

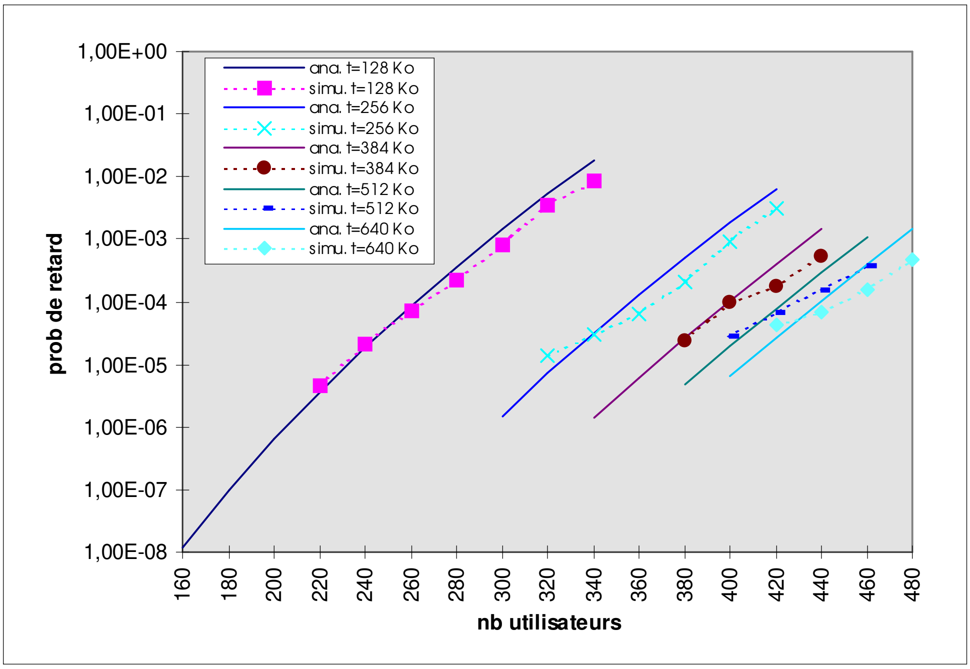

Figure 10. Effet de la taille des blocs sur la probabilité de retard

Pour étudier l'effet de la taille des blocs, nous considérons un nombre d'anticipation égal à 2. La figure 10 présente la probabilité de retard en fonction du nombre d'utilisateurs pour des tailles de blocs de 256, 354, 512 et 640 Ko. Les courbes en pointillé correspondent aux résultats des simulations, les courbes en trait plein correspondent aux résultats obtenus par le modèle analytique approché.

Sur les diverses courbes, on peut estimer la précision de la méthode analytique. Cette dernière nous permet dans tous les cas d'obtenir rapidement un ordre de grandeur de la probabilité de retard.

Lorsque la taille des blocs augmente, le débit global des disques augmente et ainsi la probabilité de retard diminue. On remarque que le gain en charge est plus important pour les premières variations de la taille (premières courbes plus espacées). Ceci s'explique en partie par le fait que le débit des disques croit d'une façon hyperbolique en fonction de la taille des blocs (cf. équation 8). Il n'est donc pas intéressant d'accroître indéfiniment la taille des blocs d'autant que l'on est limité par le coût des mémoires dans les disques et dans les autres composants du système. Une trop grande taille de blocs pourrait aussi remettre en cause les propriétés de l'architecture à savoir le bon équilibrage de la charge sur les disques et sur le réseau d'interconnexion. Pour les valeurs des paramètres que nous avons considérées, une taille de blocs de 128Ko semble raisonnable.

8. Conclusion

Nous avons présenté dans cette étude un modèle analytique simple pour l'analyse des performances d'une architecture générale de serveurs multimédias. Il s'agit d'une architecture multidisques en grappe qui consiste en un ensemble de nœuds de stockage reliés à un ensemble de nœuds de transmission par un réseau d'interconnexion.

Nous nous sommes intéressés à la qualité de transmission exprimée en terme de probabilité de retards des blocs de données qui induiraient des petites interruptions au niveau de l'utilisateur. Vue la complexité du système étudié, nous avons essayé de trouver un compromis entre la complexité du modèle et sa précision. Pour ce faire nous avons étudié en détail chacun des composants en essayant de simplifier au maximum sa modélisation. A chaque étape les approximations effectuées sont justifiées et validées. Le modèle final du système complet est à son tour validé. Les validations sont effectuées par de longues simulations réalisées sur des modèles précis sous forme de réseau de files d'attente. Notons que la validation a été effectuée avec les données techniques de matériel performant actuellement disponible sur le marché. Dans le modèle final les disques sont représentés par des files M/D/1 et le réseau d'interconnexion par une file infinie à service déterministe soit un délai constant.

Le modèle analytique est alors utilisé pour dimensionner les paramètres internes de fonctionnement du système. Nous avons montré que l'architecture présente un bon comportement multi-utilisateur et que la technique d'anticipation est intéressante. L'augmentation de la taille des blocs améliore les performances du système jusqu'à un seuil donné au dessus duquel la faible amélioration ne justifie plus le coût des mémoires nécessaires aux composants du système.

La méthode analytique présentée est d'un intérêt assez général et s'adapte pour beaucoup de variantes de l'architecture proposée. Nous travaillons actuellement sur des approximations moins fortes pour le temps de service disque. Des études pourraient être faites pour améliorer le modèle du réseau d'interconnexion.

Bibliographie

[Berson 94] S. Berson, S. Ghandeharizadeh, R. Muntz, and X. Ju. Staggered striping in multimedia information systems. Proceedings of the fifth intl. Conf. On management of data, may 1994.

[Damm 94] G. Damm, G. Babonneau, A.L. Beylot and M. Becker, Performance evaluation of a multimedia server for ATM networks, Proceedings of the 27th Annual Simulation Symposium, La Jolla, April 1994, pp. 41-50.

[Ganger 94] Gregory R. Ganger, Bruce L. Worthington, Robert Y. Hou, and Yale N. Patt. Disk Arrays, High-Performance, High-Reliability Storage Subsytems. IEEE Computer, March 1994, pp. 30-36.

[Gemmell 95] D. James Gemmell, Harrick M. Vin, Dilip D. Kandlur, P. Venkat Rangan. Multimedia Storage Servers : A Tutorial and Survey. IEEE Computer, May 1995.

[Haskin 95] Roger Haskin and Frank L. Stein. A system for delivery of interactive television programming. COMPCON, Spring 1995.

[Hennesy 90] J.L. Hennesy and Patterson. Computer architecture a quantitative approach. Morgan Kauffman, 1990.

[IBM] http: //www.ibm.com, http: //www.storage.ibm.com/oem/tinfo.html

[Intel 91] Intel Corporation. PARAGON XP/S, product overview, 1991.

[ISO 93] ISO/IEC JTC1/SC29/WG11, N0601. Coding of Moving Pictures and Associeted Audio. Nov., 1993.

[Jain 92] Raj Jain. The art of computer systems performance analysis. Wiley, wiley professional computing. 1992.

[Kaddeche 96] H. Kaddeche, G. Damm, G. Babeaunneau and M. Becker. Design and performance analysis of a multimedia server for high speed networks. In proceedings of TDP96, pp. 359-374, La Londe les Maures, France, June 1996.

[Lee Edward 93] Edward K. Lee, Randy H. Katz. An analytic performance model of disk arrays. In Proceedings of the ACM Sigmetrics, May 1993, pp. 98-109.

[Lee Gyungho 89] Gyungho Lee. A performance bound of multistage combining networks. IEEE Transactions on computers, vol. 38, n° 10, October 1989.

[Livny 87] M. Livny, S. Khoshafians and H. Boral, Multi-disk management algorithms, Proceedings of the 1987 ACM SIGMETRICS International Conference on Measurement and Modelling of Computer Systems, May 1987, pp. 69-77.

[Mohapatra 96] Parsant Mohapatra and Chita R. Das. Performance analysis of finite-buffered asynchronous multistage interconnection networks. IEEE Transactions on parallel and distributed systems, vol. 7, n° 1, January 1996.

[Potier 84] D. Potier and M. Veran, QNAP2: a Portable Environment for Queuing Systems Modelling, International Conference on Modelling and Performance Analysis Tools, May 1984.

[Ruemmler 94] Chris Ruemmler and John Wilkes. An introduction to disk drive modeling. IEEE Computer, March 1994, pp. 17-28.

[Scranton 83] R.A. Scranton, D.A. Thompson, and D. W. Hunter. The access time myth. IBM Res. Rep. RC10197, Sept. 1983.

[Siegle 87] H. J. Siegle, T. Schwederski, D. G. Meyer, Hsu Tsun-Yunk : Large scale parallel processing systems. Microprocessors and microsystems. Vol. 11, n° 1, Jan./Feb. 1987, pp. 3-20.

[Tewari 96] Renu Tewari, Rajat Mukherjee, Daniel M. Dias, Harrik M. Vin. Design and Performance tradeoffs in clustered video servers. In the proceedings of IEEE-ICMCS-Tokyo, June 1996.

[Vetter 95] Ronald Vetter. ATM concept, architecture and protocols. CACM, february 1995.

[Vin 94] Harrick M. Vin, Multimedia systems architecture, In Defining the Global Information Infrastructure : Infrastructure, Systems, and Services, Ed. Stephen F. Lundstrom, Vol. CR56, SPIE Press, Pages 287-296, November 1994.