Techniques élémentaires d'encadrement des temps de réponse dans une prévision de charge (capacity planning) .

1. Introduction

Les logiciels de modélisation mathématique utilisent des algorithmes de simulation relativement complexes, qui permettent de prévoir le comportement d'un système comprenant plusieurs classes de charges de travail (une classe de charge de travail est un ensemble homogène d'utilisateurs qui utilisent la même application). Ces algorithmes sont trop complexes pour pouvoir être appliqués " à la main ". Il est en revanche relativement simple d'obtenir un encadrement des temps de réponse et du débit (nombre de transactions par secondes) d'un système comportant une seule classe. Les termes temps de réponse et débit peuvent d'ailleurs prendre plusieurs sens, selon le domaine auquel on l'applique.

- Dans le cas d'un applicatif interactif, la notion de temps de réponse s'impose d'elle-même, le débit s'entend comme le nombre de transactions applicatives maximum que peut absorber l'ordinateur.
- Dans le cas de travaux batchs, on entend par temps de réponse le temps d'exécution, le débit n'a pas de sens réel.
- Dans le cas du système d'Entrées/ Sorties, le temps de réponse est le temps de traitement d'une E/S (y compris le temps d'attente), le débit est le nombre d'E/S maximum que peut supporter le système d'Entrées/ Sorties.

On peut appliquer cette notion de temps de réponse et de débit à pratiquement tous les centres de services d'un ordinateur (CPU, mémoire, disques, etc.). Cet article donne les formules de calcul pour les traitements batchs et interactifs et propose un exemple concret pour ce dernier cas de figure.
Intérêts de l'encadrement des temps de réponse et du débit dans un modèle :

- Simplicité.
- Mise en évidence du goulot d'étranglement du système et de son effet sur les temps de réponse.
- La simulation des différentes configurations possibles du système est simple à effectuer.

Variables de travail
Les variables de base sont les mêmes que celles utilisées dans l'article sur les corrélations entre les données issues des mesures de performances.

Pour mémoire :

On appellera "centre de service " un composant de l'ordinateur (CPU, disque, etc.)

K	**Nombre de centres de services**
U_k	**Utilisation (pourcentage de temps pendant lequel un centre de service k est utilisé)**
D_k	**Temps total passé dans le centre de service k par une transaction (si une transaction requiert 5 accès à un disque, que chacun de ces accès requiert 10 ms,D_k = 50ms). D_k varie en fonction de K (D est une constante).**
D	**Temps total passé par une transaction dans les différents centres. On considère que D est connu et est une constante.** $D = \sum_{k=1}^{n} D_k$
R_k	**Temps de résidence dans un centre de services R**
R	**Temps de résidence dans le système (c'est pour un ordinateur le temps deréponse, pour un ensemble de disques, ce sera le temps passé en E/S).** $R = \sum_{k=1}^{n} R_k$
N	**Population dans le système**
X	**Nombre de transactions par unités de temps sur le système entier**

2. Systèmes équilibrés

La première possibilité d'approximation consiste à supposer que les demandes sur les centres de services sont les mêmes. Les limites de ce type de modèle apparaissent immédiatement :

- On ne pourra pas y mêler CPU et systèmes d'Entrées/ Sorties, car les demandes sur ces deux centres ne peuvent pas être identiques.
- Il est difficile dans la pratique d'équilibrer parfaitement un système d'Entrées/ Sorties.

Ce type de modèle offre en revanche un certain nombre d'avantages :

- Pour un système borné par les E/S, ce modèle est une bonne approximation.
- Il plus précis que les déductions effectuées par bornes asymptotiques.

Le calcul ici proposé permet d'encadrer le débit du système et le temps de réponse en délimitant les demandes sur chaque service que demanderont les transactions. Connaissant le nombre d'Entrées/ Sorties nécessaires à l'exécution d'un batch, on peut

calculer D. D_k (le nombre d'Entrées/ Sorties sur chaque disque) variera en fonction du nombre de centres K entre les bornes D_{max} et D_{min}.

Nous utiliserons l'exemple suivant afin d'illustrer les résultats obtenus :
On considère que chaque transaction nécessite 18 Entrées/ Sorties de 12ms (quelque soit le disque) chacune. On en déduit D=18*0,012=0,216. En fonction du nombre de disques, et sachant que le système est équilibré, on aboutit aux résultats suivants :

Nombre de disques	Nombre d'E/S/disque	D_k
4	4,5	0,048
3	6	0,072
2	9	0,096

$$D = \sum_{k=1}^{n} D_k = 0,216$$

$$D_{max} = 0,096$$

$$D_{min} = 0,072$$

2.1 Cas des traitements batchs

Le calcul de base s'appuie sur la fonction qui donne l'utilisation d'un centre en fonction de la population N présente dans le système :

$$Uk = \frac{N}{N+K-1}$$

(la démonstration de ce théorème est donnée en paragraphe 2.3).

On peut d'après la loi de Little exprimer le débit X(N) du système en fonction de la population et de la demande de service à chaque centre par

$$X(N) = \frac{U_k}{D_k} = \frac{N}{N+K-1}\frac{1}{D_k}$$

D'où :

$$\frac{N}{N+K-1} x \frac{1}{D_{max}} \leq X(N) \leq \frac{N}{N+K-1} x \frac{1}{D_{min}}$$

et le temps de réponse R(N) est encadré par

$$(N+K-1)D_{min} \le R(N) \le (N+K-1)D_{max}$$

On peut trouver une seconde borne optimiste obtenue par la constation intuitive suivante: en supposant que le système ne soit pas équilibré, et que la charge de travail demandée au système soit forte, l'utilisation de chaque centre augmente, sans pour autant dépasser 1. Le centre le plus sollicité, celui pour lequel la demande est D_{max}, est tel que :

$$U_k(N) = D_k X(N) \le 1 \Rightarrow X(N) \le \frac{1}{D_{max}}$$

et

$$D + (N-1)D_{min} \le R(N)$$

2.2 Cas des traitements interactifs

Les formules deviennent

$$\frac{N}{D+Z+\frac{(N-1)D_{max}}{1+\frac{Z}{ND}}} \le X(N) \le \frac{N}{D+Z+\frac{(N-1)D_{min}}{1+\frac{Z}{ND}}}$$

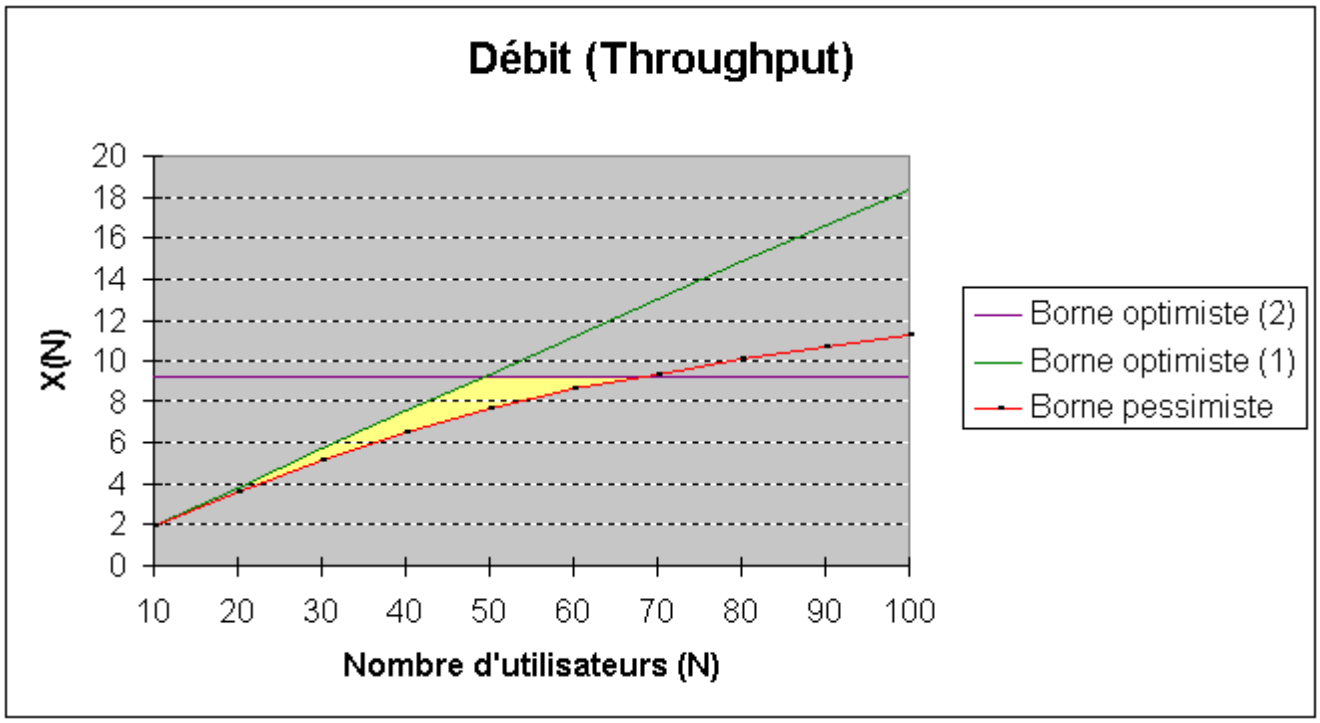
et

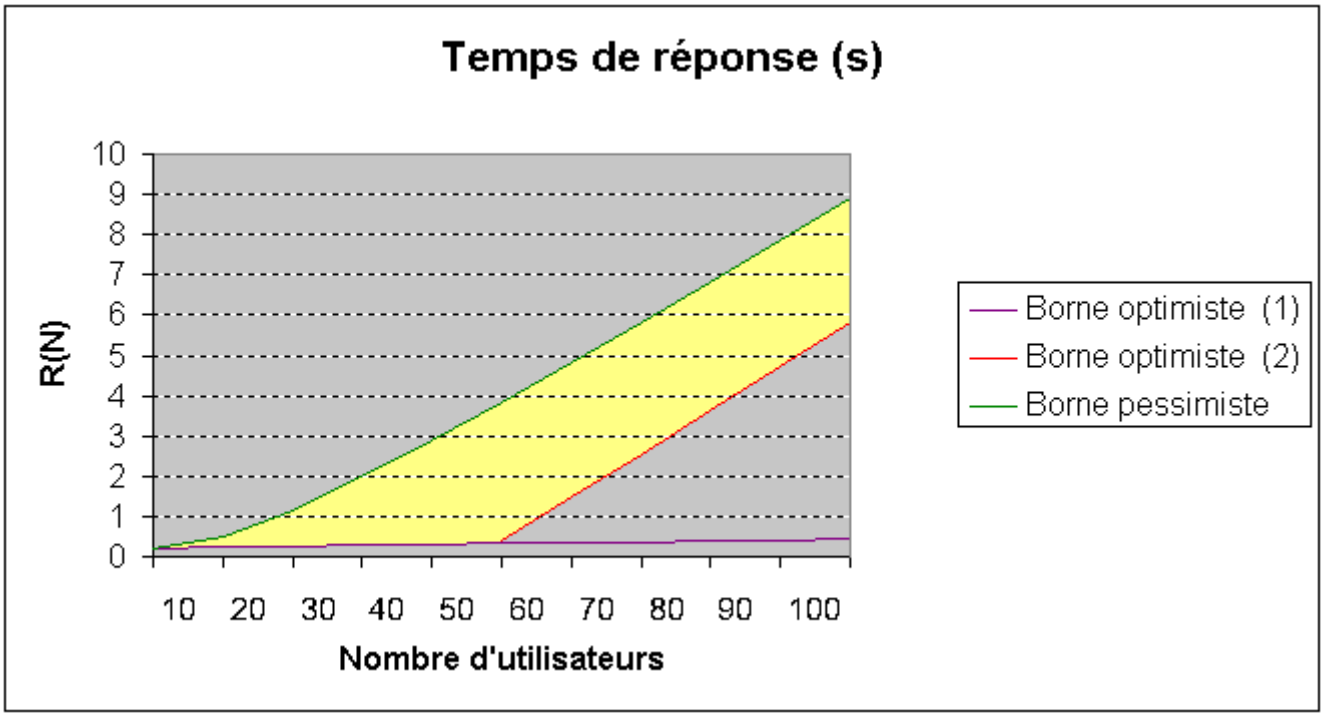
$$\frac{N}{D+Z+\frac{(N-1)D_{max}}{1+\frac{Z}{ND}}} \le R(N) \le \frac{N}{D+Z+\frac{(N-1)D_{min}}{1+\frac{Z}{ND}}}$$

La démonstration est similaire à celle donnée pour les traitements batchs, en tenant compte du fait que l'évolution du temps de réponse en fonction du débit est

$$R = \frac{N}{X} - Z$$

On obtient les courbes de débit et de temps de réponse (on utilise pour cet exemple un temps de réflexion pour **Z**=5s).





2.3 Démonstration pour les traitements batchs :

Elle s'appuie sur le principe de base suivant : le temps de séjour d'une requête dans un service est égal au temps passé dans ce service (pendant laquelle la requête est traitée) auquel est ajouté le temps passé dans la file d'attente de ce service. Le temps passé dans la file d'attente est égal au temps de traitement des requêtes qui précèdent la requête, donc au produit de la longueur de la file d'attente par le temps passé par une

requête dans le service. Il faut néanmoins tenir compte du fait que la queue d'attente est celle qui existe avant que n'arrive la transaction courante : il y a donc N-1 clients dans le référentiel, le Nième étant la transaction courante. Si on nomme $Q_k(N-1)$ la longueur de la file d'attente pour N-1 clients, on obtient donc

$$R_k(N) = D_k + D_k Q_k(N-1)$$

La relation entre les longueurs des files d'attente $Q_k(N)$ pour N clients et celles $Q_k(N-1)$ qui existaient pour N-1 clients peut être approximée par :

$$Q_k(N) - Q_k(N-1) = \frac{Q_k(N)}{N}$$

On parle ici d'une approximation, car un client de plus dans un référentiel équilibré ayant des files d'attente non nulles sur les services a pour effet de le déséquilibrer, sauf s'il ne compte qu'un seul service.

D'où :

$$R_k = D_k(1 + \frac{N-1}{N} Q_k(N))$$

or d'après la loi de Little

$Q_k = XR_k$, et $U_k = XD_k$, donc $Q_k = U_k R_k / D_k$

On en déduit que

$$R_k = D_k(1 + \frac{N-1}{N} \frac{U_k}{D_k} R_k)$$

Donc (équation 1) :

$$U_k = \frac{N-1}{N}(1 - \frac{D_k}{R_k})$$

Par ailleurs, le système étant équilibré, les temps de résidence dans les différents centres de service sont égaux et leur valeur est $R_k = R/K$.

Or $R = N/X$ (loi de Little) et $X = U_k / D_k$, on en déduit que $R_k = (ND_k)/(KU_k)$.
En réintroduisant R_k dans l'équation (1), on en déduit U_k :
d'où :

$$Uk = \frac{N}{N+K-1}$$

3. Bornes asymptotiques

Que faire si on ne peut pas considérer que les demandes sur les services sont équilibrées ? On peut alors raisonner en calculant les asymptotes sur les encadrements réalisés ci-dessus.

1. **Première borne pessimiste sur le débit:**

Nous nous trouvons donc ici dans un cas de charge importante. Le centre qui occasionnera le goulot d'étranglement principal sera celui pour lequel le débit sera minimal, donc celui pour lequel l'utilisation sera la plus importante. Mais l'utilisation est par définition inférieure à 1, donc $X(N)$ trouvera une première limite dans le centre de service qui saturera le plus vite :

$$U_k(N) = D_k X(N) \leq 1 \Rightarrow X(N) \leq \frac{1}{D_{max}}$$

Attention : D_{max} est ici défini comme le service maximal observé sur un centre, alors que D est constant. Lors de l'étude des systèmes équilibrés, D_{max} était une valeur commune à tous les centres. L'indice "max" traduisait une valeur maximale de D, non de D_k.

2. Seconde borne pessimiste sur le débit

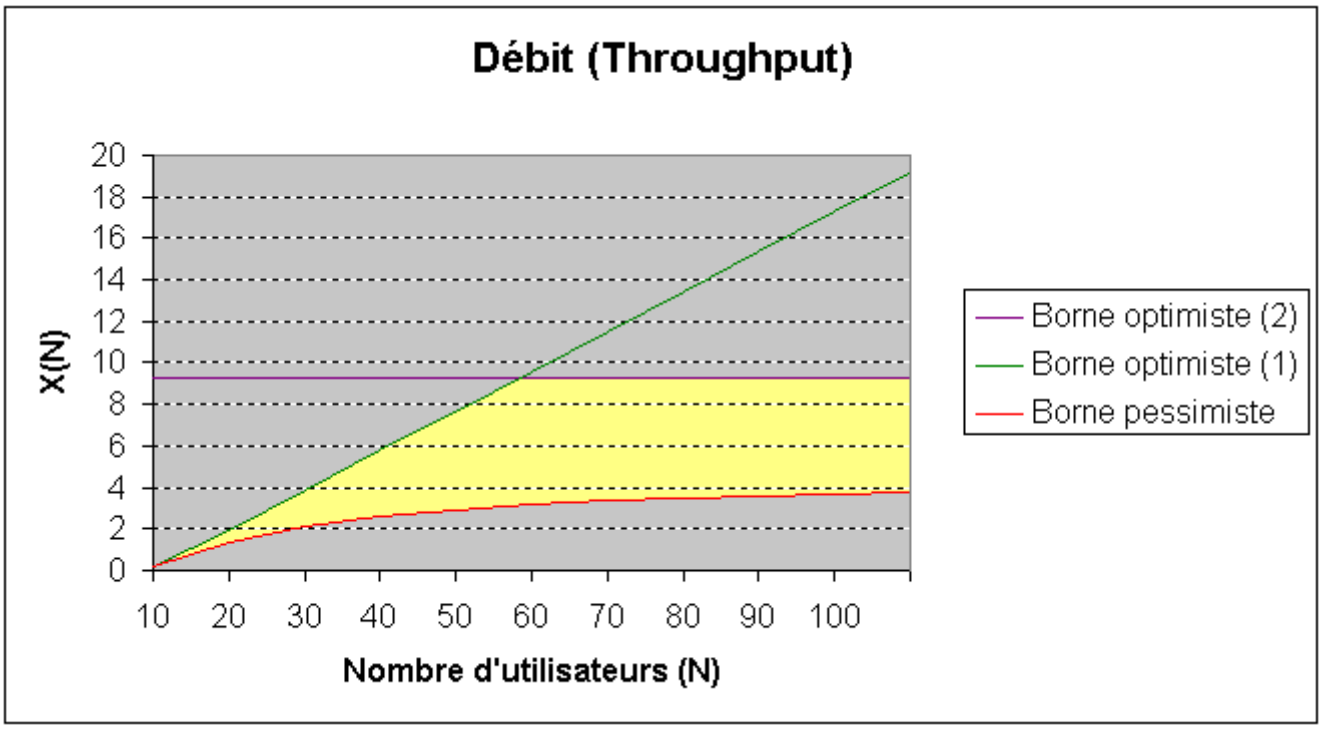
Le débit le plus faible que puisse atteindre un système est celui pour lequel une transaction doit attendre que toutes les autres transactions soient servies. Pour un environnement batch, si N est la population de ce système, ce temps d'attente est pour une transaction de (N-1)D car chaque transaction a besoin de passer une fraction de temps D dans le système. Pour un environnement transactionnel, on devra ajouter à l'attente sur les autres transactions le temps de réflexion. Pour chaque client, le débit maximum sera donc de 1/(ND+Z), pour N clients il sera donc borné par :

$$X(N) \leq \frac{N}{ND + Z}$$

3. Borne optimiste sur le débit

En prenant le même raisonnement que ci-dessus, le meilleur débit du système est obtenu lorsque les centres sont toujours libres : aucune file d'attente n'existe. Comme chaque transaction nécessite une fraction de temps D pour le batch et de D+Z pour l'interactif, le débit par client sera alors dans ce cas de 1/(D+Z), pour N clients il sera de N/(D+Z), soit:

$$X(N) \geq \frac{N}{D + Z}$$

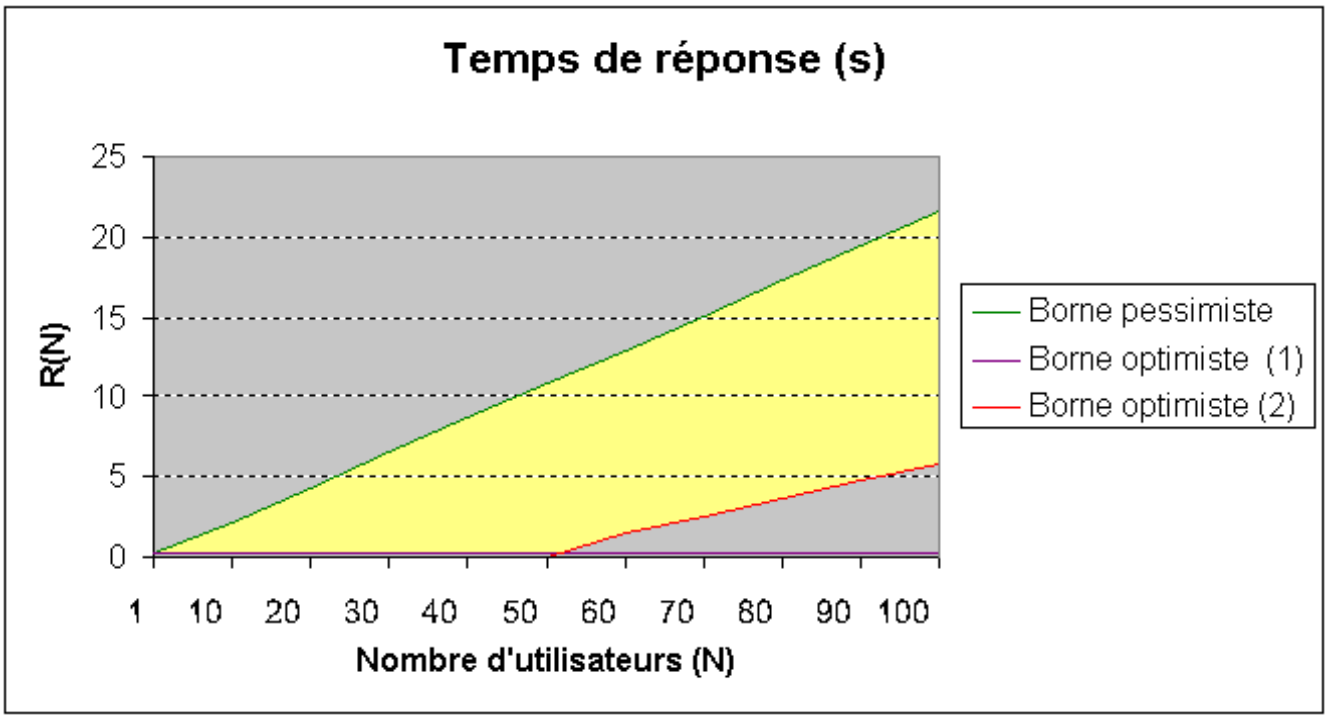


Temps de réponse :
On utilise la loi de Little pour convertir le débit en temps de réponse, soit les bornes :

$$ND_{max} - Z \leq R(N) \leq ND$$

Bien entendu, on aura également :

$$R(N) \geq D$$



Références:

Quantitative system performance Lazowska, Zahorjan, Graham, Sebcik chez Prentice All