Chapitre 5
Modèles de files d'Attente

1. Introduction.

La théorie des files d'attente est un outil de modélisation développé initialement pour l'analyse statistique des réseaux téléphoniques (évaluation des performances, dimensionnement du système..). Cette technique a été étendue pour l'analyse des réseaux d'ordinateurs ; voir par exemple [1,2,3] .

Définition. Un système de files d'attente est caractérisé par le mécanisme d'entrées des clients (arrivées) dans le système et le mécanisme de service. Ces derniers peuvent être définis en donnant:

1.**la loi des arrivées** : Soient $t_0=0< t_1<t_2<\ldots t_n<\ldots$ les instants d'arrivées des clients, où t_n=instant d'arrivée du nième client C_n. En général, on observe les variables $\xi_n=t_n-t_{n-1}$, n=1,2,… qui sont supposées indépendantes et identiquement distribuées (i.i.d.) de loi $A(x)=P(\xi_n<x)$. Si $A(x)=1-e^{-\lambda x}$, nous avons vu que dans ce cas, la suite $\{t_n\}$ formait un processus de Poisson de paramètre λ. Dans le cas le plus simple les clients sont supposés statistiquement homogènes. Cependant on peut envisager le cas où il existe plusieurs classes de clients, chaque classe étant définie par ses propres paramètres.

2.la **loi de service** : Soit S_n la durée de service du client C_n ; $B(x)=P(S_n<x)$ sa f.r. ; ces variables sont i.i.d.

3.Le **nombre de serveurs** m (entier positif) ayant des caractéristiques statistiques identiques.

4.La **capacité** maximale de la file d'attente K (finie ou infinie).

5.La **source** de clients L (finie ou infinie).

6.La **discipline de service** ou ordonnancement (FIFO, LIFO, RANDOM, ….).

7.D'autres paramètres éventuels décrivant diverses contraintes de fonctionnement (temps d'attente borné, règles de priorité entre classes de clients, serveurs sujets à des pannes, répétition d'appels, arrivées ou service par paquets, etc….

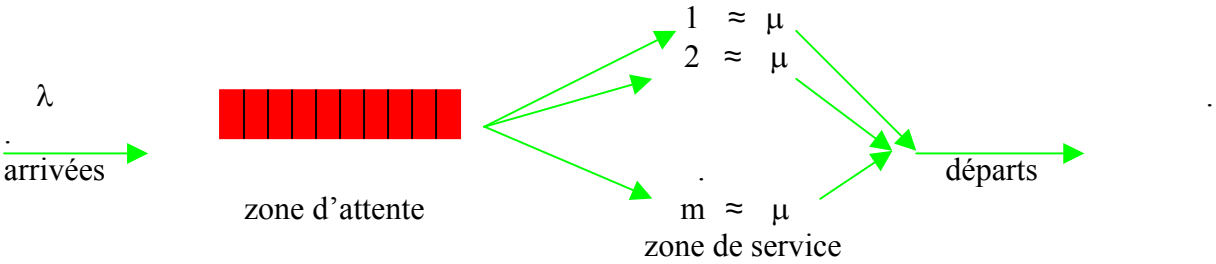


Figure 1 : Schéma simplifié d'un système de files d'attente

Exemples:

Type de système	Serveurs	Clients
Cabinet Médical	Médecin	Malades
Facultés	Filières	Etudiants
Central téléphonique	Lignes téléphoniques	Abonnés (appels)
Aéroport	Pistes d'atterrissage	Avions
Port	Navires	Quais
Atelier de Fabrication	Machines, convoyeurs	Pièces à traiter
Atelier de maintenance	Réparateurs	Machines en panne
Systèmes informatiques	Ressources (processeurs, mémoires, périphériques)	Informations (messages, programmes

Réseau routier	Feux lumineux	Véhicules
Serveur Web	Temps de service=temps de transfert d'une page Web	Arrivées des sessions (http)

Pour les applications aux systèmes informatiques, voir par exemple [4,5].

Un réseau de files d'attente est un ensemble de systèmes (appelés aussi « stations » ou « nœuds ») interconnectées de manière quelconque ; chaque client nécessite un service dans une ou plusieurs stations. Ce cas englobe le cas précédent où les serveurs ne sont pas identiques (statistiquement).

Notations de Kendall : C'est une nomenclature de classification des modèles de files d'attente de la forme A/B/m/K/L où

A est le code de la loi de service ; B celui de la loi de service ; m le nombre de serveur ; K la capacité de la zone d'attente et L celle de la source de clients (nombre maximal de clients pouvant accéder dans le système).

Code	Type de loi de probabilité
M	Exponentielle : $F(x)=1-e^{-\alpha x}$, $x>0$
D	Déterministe : $F(x)=1$ si $x>c$ $F(x)=0$ si $x<c$
E_k	Erlang d'ordre k : loi de la somme de k v.a. Exp(α)
H_k	Mélange d'exponentielle
G	Loi générale
GI	Loi générale indépendante

Exemple : On note M/M/2 (arrivées poissoniennes, service exponentiel, m=2 serveurs, capacité de la file infinie, source infinie) ; M/G/2 (ici la loi de service est générale ou arbitraire).

Types de problèmes :

1.Etude de stabilité et d'existence d'un régime stationnaire: Au cours de cette étape on tente d'étudier les conditions d'existence d'un régime stationnaire (ou permanent) qui correspond au non-engorgement du système.

2.Evaluation des performances du système: C'est l'étude des principales caractéristiques du système appelées aussi « performances du système ». Ce sont :

-taille moyenne de la file d'attente E(Q) (nombre moyen de clients dans la file) ;
-nombre moyen de clients dans le système E(N) (en attente ou en cours de service)
-temps moyen d'attente E(W) d'un client
-temps moyen de séjour E(V) dans le système (temps d'attente +temps de service)
-probabilité qu'un client qui arrive dans le système ne soit pas accepté pour service pour cause de congestion (probabilité de refus ou de perte ou d'encombrement P_r).

-taux d'arrivées effectif $\bar{\lambda}$.
-débit absolu A (nombre moyen de clients servis par unité de temps)
-débit relatif A'(probabilité qu'un client qui arrive dans le système soit servi).
-nombre moyen de serveurs actifs E(SA) ou oisifs E(SO)

-traffic offert : pourcentage de temps pendant lequel un serveur (ressource) est occupé : mesuré en Erlang. Il peut être estimé de la manière suivante :

$\lambda = \frac{N(S_1 + S_2, ..., S_N)}{T}$, où T est la période d'observation, N(T) le nombre d'appels (de clients) durant le temsp T , S_i durée de l'appel.

La littérature de télécommunication mentionne qu'une ligne téléphonique occupée à 100% a un trafic de 1 Erlang ; une ligne résidentielle fixe : 70 Me. Une ligne industrielle 150mE ; et une ligne mobile 25 mE .
etc….

La plupart de ces mesures de performance sont calculées en général en régime stationnaire, mais on peut également tenter de les obtenir en régime transitoire. Nous verrons que seul le régime stationnaire présente un intérêt. En régime transitoire, on s'intéresse beaucoup plus à la conception du système lui-même.

3.Contrôle du système :

Cette partie s'intéresse aux modèles normatifs qui permettent de calculer la configuration « optimale » du système ou d'élaborer des politiques de contrôle optimal du système par rapport à un critère économique donné. Par exemple :

Critère	Notation	Nature de l'optimisation
Temps d'attente	E(W)	Minimiser
Temps de séjour	E(V)	Minimiser
Taille de la file	E(Q)	Minimiser
Débit du système	A	Maximiser
Coûts d'exploitation	E©	Minimiser
Temps d'inactivité	E(I)	Minimiser
Disponibilité	E(D)	Maximiser
Probabilité de refus	π_R	Minimiser

Les variables de contrôle peuvent être :

(i) le taux de service ;
(ii) le taux d'arrivées ;
(iii) le nombre de serveurs
(iv) la capacité de la file ;
(v) la discipline de service (on parle alors d'ordonnancement).

Lorsque les lois d'arrivées et de service sont exponentielles, on dit que le système est **markovien**, car dans ce cas le processus N(t)=nombre de clients dans le système est une chaîne de Markov à temps continu. Ce processus comme nous le verrons suffit pour obtenir les principales caractéristiques du système (appelées aussi **mesures de performance**).

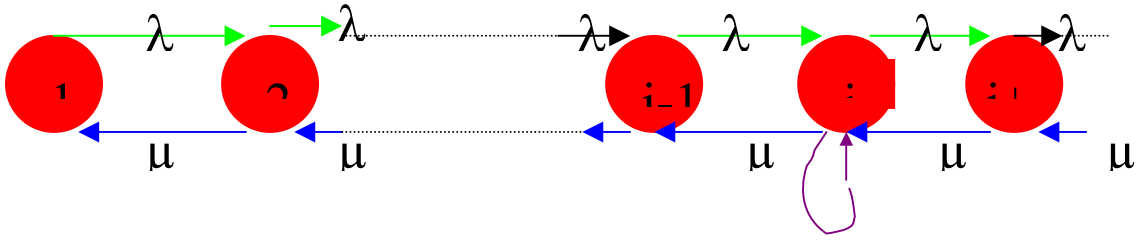
Considérons quelques modèles markoviens décrits par des processus de naissance et de mort et qui conduisent à des solutions simples

Modèle M/M1. Soit un système d'attente tel que $A(x)=1-e^{-\lambda x}$, $B(x)=1-e^{-\mu x}$; On note N(t)=nombre de clients présents dans le système à l'instant t.

Exercice 1:

1. Montrer que la probabilité d'arrivée d'un client durant (t,t+h) est égal à λh+o(h) ; la probabilité qu'un client soit servi durant (t,t+h) (sachant que le serveur est occupé) est égale à μh+o(h). Représenter le graphe des états et en déduire que le processus N(t) est un processus de naissance et de mort dont on précisera les paramètres. Ecrire les équations de Chapman-Kolmogorov.

Solution : Découle immédiatement des propriétés du processus de Poisson et de la loi exponentielle. N(t) est donc un processus de naissance et de mort de paramètres : $\lambda_i \equiv \lambda$, $\mu_i \equiv \mu$ "i≥0.



Automate correspondant

Equations de Chapman-Kolmogorov : Soit $\pi_i(t)=P(N(t)=i), i\geq 0$. Ces probabilités sont solutions du système (voir graphe) :

$$\frac{d\pi_i(t)}{dt}=\lambda\pi_{i-1}(t)+\mu\pi_{i+1}(t)-(\lambda+\mu)\pi_i(t)\text{ , i}\geq 1.$$

$$\frac{d\pi_0(t)}{dt}=\mu\pi_1(t)-\lambda\pi_0(t)$$

2. Etudier la classification des états de la chaîne et en déduire sa nature .Vérifier que la condition d'ergodicité ρ=λ/μ<1 est une condition de stabilité du système au sens où il y a absence de congestion.

Notons $\pi_i = \lim_{t\to\infty}\pi_i(t) = \lim_{t\to\infty} P(N(t)=i)$ la distribution stationnaire des états, si un tel régime stationnaire existe. Dans ce cas le système se ramène à un système d'équations linéaires algébriques (π_i ne dépend pas de t, et la dérivée s'annule) :

$$\lambda\pi_{i-1}+\mu\pi_{i+1}-(\lambda+\mu)\pi_i=0\text{ , i}\geq 1.$$

$$\mu\pi_1-\lambda\pi_0=0$$

Ce système s'écrit encore sous forme matricielle $\pi\Lambda=0$, où Λ est la matrice des taux de transition instantané (ou infinitésimaux) (on l'appelle parfois générateur infinitésimal) :

$$\Lambda=\begin{pmatrix} -\lambda & \lambda & 0 & ..... & 0 & 0 & 0 & ..... \\ \mu & -(\lambda+\mu) & \lambda & ..... & 0 & 0 & 0 & ..... \\ 0 & \mu & -(\lambda+\mu) & ..... & & & & ..... \\ 0 & 0 & \mu & ..... & & & & ..... \\ 0 & 0 & 0 & ..... & & & & ..... \\ ..... & ..... & ..... & ..... & ..... & ..... & ..... & ..... \\ & & & ..... & & & & ..... \\ & & & ..... & & & & ..... \end{pmatrix}$$

On note que c'est une matrice tri-diagonale (tous les éléments se trouvant en dehors des trois diagonales principales sont nuls). Les seuls éléments négatifs se trouvent sur la diagonale principal. La somme des éléments d'une ligne est nulle.

On peut résoudre ce système de manière récursive :

Pour i=1, on a $\pi_1=\frac{\lambda}{\mu}\pi_0=\rho\pi_0$, en notant $\rho=\frac{\lambda}{\mu}$.

Reportons l'expression de π_1 dans la seconde équation pour i=2,

$$\lambda\pi_0+\mu\pi_2-(\lambda+\mu)\pi_1=0 \quad \Leftrightarrow \quad \lambda\pi_0+\mu\pi_2-(\lambda+\mu)\rho\pi_0=0$$

On obtient après simplification : $\pi_2=\rho\pi_1=\rho^2\pi_0$. En général , $\pi_i=\rho^i\pi_0$,i≥0.

Il reste à déterminer la constante π_0. On peut utiliser dans ce cas la condition de normalisation $\sum_{i=1}^{\infty}\pi_i=1$.

En substituant l'expression de π_i, on obtient $1=\sum_{i=0}^{\infty}\pi_i=\pi_0+\sum_{i=1}^{\infty}\rho^i\pi_0=\pi_0\left(\sum_{i=0}^{\infty}\rho^i\right)$.

Par suite, $\pi_0=\frac{1}{\sum_{i=0}^{\infty}\rho^i}$.

Nous sommes en mesure maintenant de déduire les conditions d'existence d'un régime stationnaire et son interprétation. Notons d'abord (on peut le constater à partir du graphe) que si $\lambda>0$ et $\mu>0$, alors tous les états communiquent entre eux, et la chaîne est donc irréductible. De plus elle est apériodique. Considérons maintenant la série $S=\sum_{i=0}^{\infty}\rho^i$. C'est une progression géométrique de pas r. On distingue deux cas :

(a) Si $r>1$, la série est divergente $S=\infty$, et $\pi_0=0$. Il s'ensuit que $\pi_i=0$ "i≥1 . Dans ce cas (voir chapitre 3) les états sont tous transitoires ou récurrents nuls et il n'y a pas de distribution stationnaire. Dans ce cas, $N(t)\to\infty$, la file d'attente croît indéfiniment ce qui correspond à la saturation du système (on dit encore engorgement, ou congestion).

(b) Si par contre r<1, alors la série est convergente S<∞, π_0>0, et tous les π_i>0 . De plus, on peut calculer la somme de la série $S=\frac{1}{1-\rho}$. Par conséquent, $\pi_0=1-\rho$; $\pi_i=\rho^i(1-\rho), i\geq 1$. La loi stationnaire du nombre de clients dans un système M/M/1 est donc de la forme

$$\pi_i=\begin{cases} 1-\rho, \ si\ i=0 \\ \rho^i(1-\rho), \ si\ i\geq 1 \end{cases}$$

qui n'est rien d'autre que la loi géométrique de paramètre r. Cette loi (voir cours de probabilités) à l'instar de la loi exponentielle dans le cas continu, est la seule loi discrète possédant la propriété d'absence de mémoire.

Par conséquent, si r<1, le vecteur $\pi=(\pi_0,\pi_1,\pi_2,...)$ forme ainsi la distribution stationnaire qui est unique (car dans ce cas la chaîne est récurrente positive).

Remarques :

(a) On a vu que π_0=1-r. Ainsi, r=1-π_0 est la probabilité pour que le serveur soit actif occupé par le service d'un client ; c'est aussi la probabilité que le système soit non vide. C'est pourquoi on l'appelle charge du serveur ou du système. En général, par exemple dans le cas de modèles à plusieurs serveurs, les charges du serveur et du système sont différentes.

(b) Notons que r<1 est équivalent à λ<μ . Si les lois d'arrivées et de service sont déterministes (c'est à dire non aléatoires, connues), alors il n'y a jamais de files d'attente, ce qui est intuitivement clair. Par contre, lorsque les lois sont aléatoires (markoviennes dans notre cas), il existe une certaine probabilité non nulle (positive) d'avoir i clients à tout instant du régime stationnaire ; cette probabilité est égale à $\pi_i=\rho^i(1-\rho)$. Si r>1 ou λ>μ le système est saturée est la file d'attente croît indéfiniment. Il faut revoir la conception du système, par exemple en augmentant le nombre de serveurs ou en diminuant le temps de service par des moyens techniques ou de gestion. Ce type de problèmes sera discuté dans les exercices plus loin.

(c) Rappelons que $\rho=\frac{\lambda}{\mu}=\frac{taux\,de\,service}{intensité\,maximale\,de\,service}$ (le taux de service est nul avec une probabilité π_0 et égal à μ avec une probabilité r). C'est pourquoi on l'appelle également taux d'utilisation du serveur ou intensité du trafic.

(d) Pour r=1, les caractéristiques du système sont instables.

(ii) la loi stationnaire du temps d'attente d'un client quelconque sachant que le serveur est occupé est une loi exponentielle de paramètre μ(1-ρ).

3. Evaluer les principales mesures de performance du système. Conclusions ?

4. Représenter graphiquement

(i) le nombre moyen de clients dans le système en fonction de la charge ρ.

(ii) le temps moyen de séjour dans le système en fonction de la charge ρ. Conclusions ?

Exercice 2: Pour chacun des modèles ci-dessous,

1.Représenter le graphe de Markov et en déduire les équations de Chapman-Kolmogorov.

2.En déduire la solution (par récurrence ou en utilisant la méthode de la fonction génératrice).

3.Retrouver les principales mesures de performance données dans les tableaux 1 et 2.

4.Représenter graphiquement quelque unes de ces mesures de performances.

Les solutions sont résumées dans les tableaux ci-dessous 1 et 2.

(i) Modèle M/M/1/K La capacité du système est finie : $\lambda_k=\lambda$ si $k\leq K$; $\lambda_k=$ si $k>K$; $\mu_k=\mu$ $k=1,2,\ldots,K$.

Exemples: 1. Système constitué d'un processeur et d'une mémoire de capacité finie
2. une machine de production et une surface capacité de stockage limitée.
3. Un service de maintenance : les clients sont les machines en panne ; le serveur est le réparateur.

(ii) Modèle M/M/1/∞/L (noté habituellement M/M/1//L). La source des clients est finie : $\lambda_k=\lambda(L-k)$ si $k\leq L$; $\lambda_k=$ si $k>L$; $\mu_k=\mu$ $k=1,2,\ldots$.

Exemples : 1. Système constitué de L machines (les clients) et d'un réparateur (le serveur) ; la capacité d'attente est illimitée (Repairman problem).

2. une machine de production et une surface à capacité de stockage limitée

3.Réseau d'ordinateur : Un serveur central connecté à L terminaux .

Exercice 3 : On considère deux ordinateurs qui sont reliés par une ligne de 64 kbit/seconde et 8 applications parallèles se partagent cette ligne. Chaque application génère un trafic poissonien de 2 paquets/seconde en moyenne.

Le concepteur doit choisir entre deux solutions : (i) La première est de dédier une bande de base de 8 kbits/seconde à chaque application. Dans ce cas, chaque ligne de 8 kbit/s agit comme une file d'attente indépendante (illimitée, FIFO) de taux de service $\mu=4$ paquets/seconde. (ii)La seconde solution est d'utiliser un accès multiple à la même ligne de transmission de 64 kbits/s. Cela revient à un seul système de taux de service $\mu=4x8=32$ paquets/seconde et un taux d'arrivées de $\lambda=2x8=16$ paquets/s. Quelle est la meilleure variante ?

Exercice 4 : (retour à l'exemple (ii).3). Les demandes provenant des terminaux sont traitées en FIFO et une seule à la fois . L'usager a un comportement cyclique avec une phase de réflexion (usager oisif), pendant laquelle le programmeur réfléchit et tape sa demande, suivie d'une phase d'attente de la réponse (usager actif). Le temps de réflexion suit une loi exponentielle de paramètre λ. Le temps de traitement de la demande suit une loi exponentielle

de paramètre μ (ce temps de traitement comprend le temps de chargement du programme correspondant et le temps d'exécution). Soit X(t) le nombre d'usagers actifs à l'instant t. Si i usagers sont actifs à l'instant t, alors L-i usagers sont susceptibles de passer à l'état actif pendant la période (t,t+h).

1.Quelle est la condition d'existence d'un régime stationnaire ?

2.Quelle est la fraction de temps pendant laquelle le système est oisif i.e. sans usagers actifs.

3.Quel est le nombre moyen d'usagers actifs en régime stationnaire

4.Quel est le temps de réponse moyen (durée moyenne de la période d'activité d'un usager
Application numérique : M=4, $1/\lambda$=25 secondes , $1/\mu$=100 secondes.

Exercice 5 : On considère un système M/M/m de paramètres λ= et μ= . Déterminer le nombre optimal de serveurs qui minimise le temps d'attente dans le système.

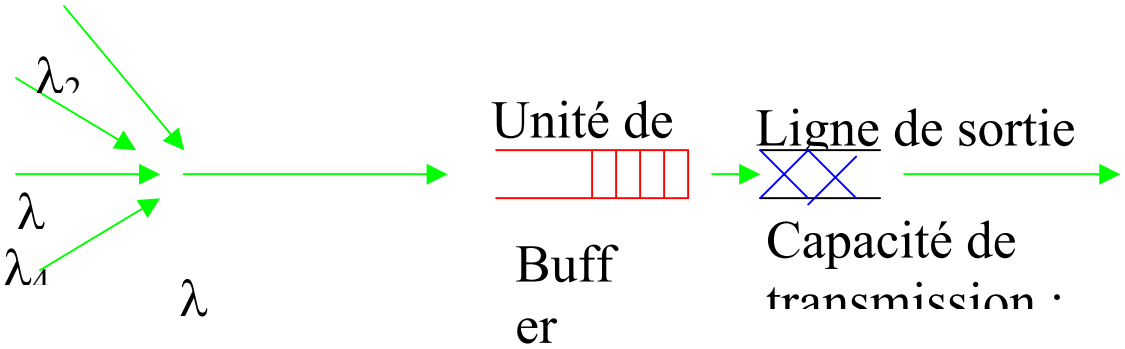
Exercice 6 : Chaque nœud d'un réseau à commutation de paquets peut-être modélisé comme un système de files d'attente. On considère ici le processus de mémorisation et de transmission des informations dans un concentrateur muni de n lignes d'entrées (en provenance des autres terminaux et autres nœuds du réseau) et une ligne de sortie(figure 1). La longueur moyenne d'un paquet est de θ=1000 bits et la capacité de la ligne de sortie est de C=9600 bits/seconde, de telle sorte que le temps de service (transmission du paquet) suit une loi exponentielle de paramètre θC=9,6 paquets/seconde. Chaque ligne d'entrée a une capacité de 4800 bits/seconde et délivre un trafic Poissonien de taux λ_i=2 paquets/seconde. (i=1,2,…,n). On prendra pour l'application numérique n=4.

1.Montrer que le flux des arrivées (de paquets) au concentrateur est poissonien de taux λ=8 paquets par seconde (Justifier votre réponse sans démonstration).

2.Quel est le nombre moyen de paquets dans le buffer et dans le concentrateur ?

3.Quel est le temps moyen d'attente (dans le buffer) et de séjour d'un paquet dans le concentrateur (attente dans le buffer+temps de transmission) ?

4.On cherche à dimensionner la capacité de la ligne de manière à minimiser le temps de séjour d'un paquet dans le concentrateur. Quel est la capacité optimale ?



Récapitulatif : Systèmes modélisables par des processus de

Table 1 :Modèles markoviens à m=1 serveur M/M/1/./. (FIFO)

Modèle →	K=∞, L =∞	K<∞, L=∞	K=∞ L<∞	K<∞,L<∞ (L≥K).
Paramètres ↓				
λ_k	λ	λ si k<K ; 0 si k≥K	λ(L-k) si 0≤k≤L ; 0 si k>L	
μ_k	μ ∀k	μ ∀k	μ ∀k	
Performance ↓				
Condition de stabilité	$\rho=\frac{\lambda}{\mu}<1$	Toujours stable (le régime stationnaire existe toujours)	Toujours stable	Toujours stable
$\bar{\lambda}=\lambda_{effectif}$	I	I(1-π_K)<$I_{M/M/1}$	λ(L-E(N))	λ(L-E(N))
π_k loi du nombre de clients dans le système	$\rho^k(1-\rho)$	$=\rho^k\frac{1-\rho}{1-\rho^{K+1}}$ si k ≤ =0 si k>K	$\frac{L!}{(L-k)!}\rho^k\pi_0$	$\rho^k\frac{L!}{(L-k)!}\pi_0$, k=0,K
Débit absolu A	$\lambda=\mu(1-\pi_0)$	$\lambda(1-\pi_K)=\mu(1-\pi_0)$	$\mu(1-\pi_0)$	μE(SA)
Débit relatif A’	1	1-π_K	$(1-\pi_0)/\rho$	
P_{refus}	0	π_K	0	π_K
E(SA)	ρ	$\rho(1-\pi_K)$	ρ	
E(Q) Taille de la file	$\frac{\rho^2}{1-\rho}$	E(N)-(1-π_0)	L-	

			$\frac{\lambda+\mu}{\lambda}(1-\pi_0)$	
E(N)	$\frac{\rho}{1-\rho}$	$\frac{\rho}{1-\rho}-\frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}$	L-(1-π_0)/ρ	
E(W) Temps d'at				
E(V) Temps de séjour	$\frac{1}{\mu-\lambda}$	E(N)/ $\bar{\lambda}$	E(N)/ $\bar{\lambda}$	

Table 2 : Modèles markoviens à m serveurs parallèles M/M/m/./.

Modèle →	K=∞, L=∞	K=m, L=∞	M=K=L =∞	K=m+M<∞,L=∞	K=∞,L<∞
Paramètres ↓					
λ_k	λ	λ	λ	λ si k<m+M, 0 sinon	λ(L-k), si k<L ; 0 sinon
μ_k	μmin(k,m)	μk	μk	μmin(k,m)	μmin(k,m)
Performance ↓					
Condition de stabilité	$\varphi=\frac{\lambda}{\mu m}=\frac{\rho}{m}<1$	Toujours stable	ρ<∞	Toujours stable	Toujours stable
π_k	$\frac{\rho^k}{k!}\pi_0,\ k\le m$ $\frac{\rho^k}{m!m^{k-m}}\pi_0,\ k\ge m$	$\frac{\rho^k}{k!}\pi_0$	$\frac{\rho^k}{k!}e^{-\rho}$ loi de Poisson	$\frac{\rho^k}{k!}\pi_0,\ k\le m$ $\frac{\rho^k}{m!m^{k-m}}\pi_0,\ m\le k\le m+$	$\frac{L!}{k!(N-k)!}\rho k\pi 0,\ k\le m$ $\frac{L!}{m!mk-m(L-k)!}\rho k\pi 0, m\le k\le L$
π_0					
Débit absolu A	λ	λ(1-π_m)	λ	λ(1-π_{m+M})	
Débit relatif A'	1	1-π_m	1	1-π_{m+M}	
P_{refus}	0	π_m		$\pi_{m+M}=\frac{\rho^{m+M}}{m^M m!}$	
E(S)	ρ	ρ(1-π_m)	ρ	ρ(1-π_{m+M})	
E(Q) Taille de la file	$\frac{\varphi\pi_m}{(1-\varphi)^2}$		0	$\frac{\rho^{m+1}}{m!m}\frac{1-\varphi^M(M+1)+M\varphi}{(1-\varphi)^2}$	
E(N)	E(Q)+ρ		ρ		
E(W) Temps d'attente					
E(V) Temps de séjour					

Bibilographie.

1.Andrew Tanenbaum , Réseaux : Architectures, protocoles, applications, Inter-éditions (iia : informatique, intelligence artificielle), Paris, 1990.(par exemple $3.1.1 page 184).

2.Claude Servin, Télécoms 1, De la transmission à l'architecture des réseaux, Editions Dunod Informatique, Paris 2000, 2ème édition. (chap.4 $5 pages 131-151).

3.G.Pujolle et al, Réseaux et Télématiques, tome 2, Editions Eyrolles, Paris 1990.(chapitre 22 page 258).

4.Crocus, Systèmes d'exploitation des ordinateurs, Dunod Informatique, Paris , 1974.(chapitre gestion des ressources ; chapitre 6, mesures et modèles de systèmes)

5. Cornafion, Systèmes informatiques répartis, Dunod Informatique, Paris, 1981. (chapitre 4, partage des voies).

Quelques sites utiles traitant des files d'attente et leurs applications :

1. Karim Fadel, Petite histoire d'Internet, Revue du palais de la découverte, http://www.palais-decouverte.fr/feteint/html/histoire.html
2. http://www.Cybersciences.com
3. Douillet, ensait, http://193.43.37.48/~douillet/cours/oprea/node3.html
4. Nicolas Navet, -INRIA/TRIO, Evaluation des performances par simulation : introduction générale et présentation du logiciel QNAP2
5. Les files d'attente dans la vie courante, http://membres.lycos.fr/dthiery
6. http://rfv.insa-lyon.fr/~jolion/STAT/node45.html
7. http://www.Microsoft .Com .html
8. http://ite.informs.org/Vol2No2/IngolfssonGrossman/