Chapitre 4.
Modèles de files d'Attente.

4.1.Introduction.

La *théorie des files* d'attente ou des Queues (Queuing Theory) est un outil de modélisation développé dans les années 1930 pour l'analyse statistique des systèmes téléphoniques (évaluation des performances, dimensionnement du système..). Cette technique se retrouve aujourd'hui pour l'analyse des systèmes informatiques et réseaux d'ordinateurs. Les cursus d'Informatique font souvent référence aux problèmes liés à la congestion ou à la saturation due au fait que des files d'attente se forment aux différents niveaux de traitement des informations et qui sont interconnectées de manière complexe.

L'une des missions principales de l'ingénieur est la conception ou l'exploitation de tels systèmes. Il est ainsi amener à quantifier les performances (ou métriques dans le jargon des développeurs de logiciels) du système : temps de réponse, débit, taux d'utilisation des différentes ressources, mais aussi estimer les ressources nécessaires au respect des critères de *Qualité de Service* (QoS) : on parle de dimensionnement de la capacité des ressources. La théorie des files d'attente permet de répondre partiellement à certaines de ces questions. Il faut noter que dans certains ouvrages, notamment d'expression slave, on parle de *Théorie du service de masse*. Ceci précise deux aspects de cet outil: (i) D'abord, la *file d'attente* n'est qu'un moyen d'ordonner les flux d'informations (les clients,demandes ou requêtes) qui se présentent en *masse* selon des lois mal connues, si ce n'est statistiquement. Ces lois seront donc définies au sens *probabiliste*. (ii) b. La file d'attente est une notion abstraite, pas toujours *physique* . Dans certains systèmes, elles sont parfois absentes (systèmes avec pertes ou refus). L'objectif de la théorie a pour finalité en fait, l'étude de l'influence des délais dans les files d'attente (ou dans le système) sur la qualité du service (QoS). Cette dernière sera évaluée à partir des différentes mesures de performance (débit, temps de réponse, taux de perte, coût…), et si elle n'est pas satisfaisante proposer des moyens de l'améliorer.

Pour ceux qui en ont le temps et la volonté, consulter le livre de Kleinrock L .[31]. Le second tome est particulièrement intéressant car l'auteur montre comment la théorie des files d'attente a été utilisée pour la conception de réseaux informatiques et de transmissions d'informations à l'exemple d'*ARPANET* (ancêtre d'*Internet* ou de l'actuel *World Wilde Web* : www) à laquelle l'auteur a participé [61-67] . Le livre de Tanenbaum [56] est plus orienté vers le côté technique de la conception de réseaux et logiciels réseaux. On y trouvera une introduction à la théorie des files d'attente (p 795) ainsi que des exemples pour évaluer les performances des réseaux (par exemple p. 185, 413).

4.2. Classification générique de Kendall . Un *système de files d'attente* est caractérisé par le mécanisme d'entrées des *clients* (abonnés, arrivées, demandes, requêtes) dans le système et le mécanisme de service. Par *client* , on entend un terme générique représentant l'entité « demande » (ou « requête ») qui peut être aussi bien celle d'un

article au niveau de sa fabrication, un appel à la mémoire d'un ordinateur, à une base de donnée (transactions) ou à un centre d'appels (call center), une sollicitation de l'organe central d'un réseau mobile…. La description d'un tel système peut être réalisée en définissant:

1.*la loi des arrivées* : Soient $t_0=0<t_1<t_2<...<t_n<...$ les instants d'arrivées des requêtes, où t_n =instant d'arrivée de la nième requête C_n.

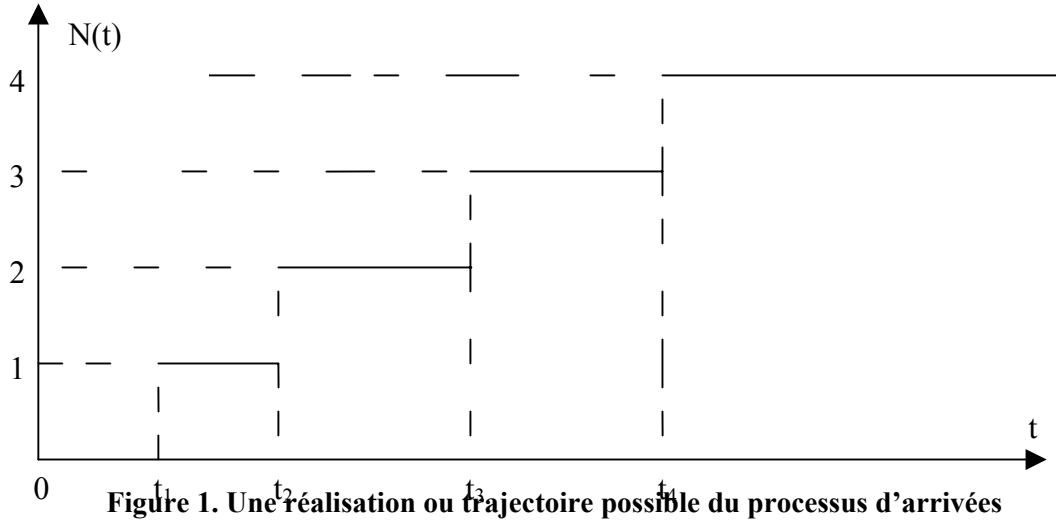


Figure 1. Une réalisation ou trajectoire possible du processus d'arrivées

En général, on observe les variables $\xi_n = t_n - t_{n-1}$, $n = 1,2,...$ supposées indépendantes et identiquement distribuées de loi $A(x) = P(\xi_n < x)$. On s'intéresse toutefois au nombre $N(t)$ d'évènements (arrivées) au cours d'une période donnée $(0,t)$, $N(t) = \max\{n: t_n < t\}$. Si $A(x) = 1 - e^{-\lambda x}$, nous avons vu que la suite $\{t_n\}$ formait un processus de Poisson de paramètre λ.

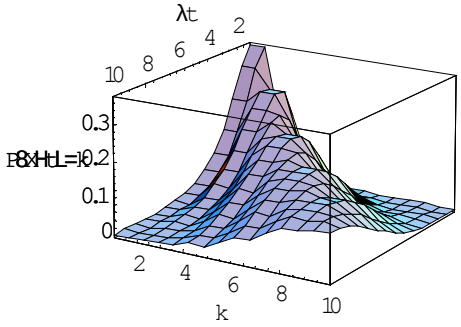
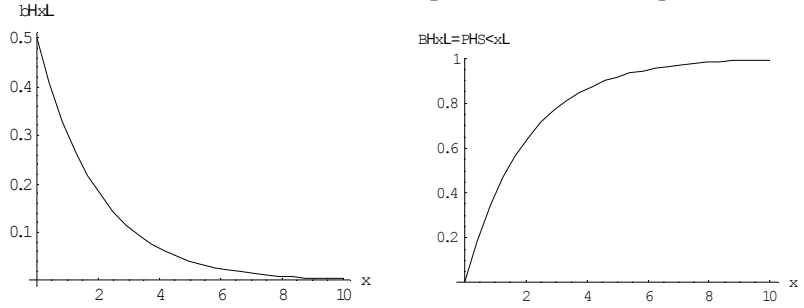


Figure 2 . Distribution de probabilité de la loi de Poisson.

Dans le cas le plus simple les clients sont supposés homogènes. On peut envisager le cas de 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 fonction de répartition. ; ces variables sont indépendantes et identiquement distribuées.



Figures 3. Densité $b(x)$ et fonction de répartition $B(x) = P(S < x)$ du temps de service (taux de service $\mu = 0.5$ et le temps moyen de service $E(S) = 2$ unités de temps).

3.*le nombre de serveurs* m (entier positif) ayant des caractéristiques 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* (FIFO, LIFO, RANDOM, ROUND ROBIN, PROCESSEUR PARTAGE,….).
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….

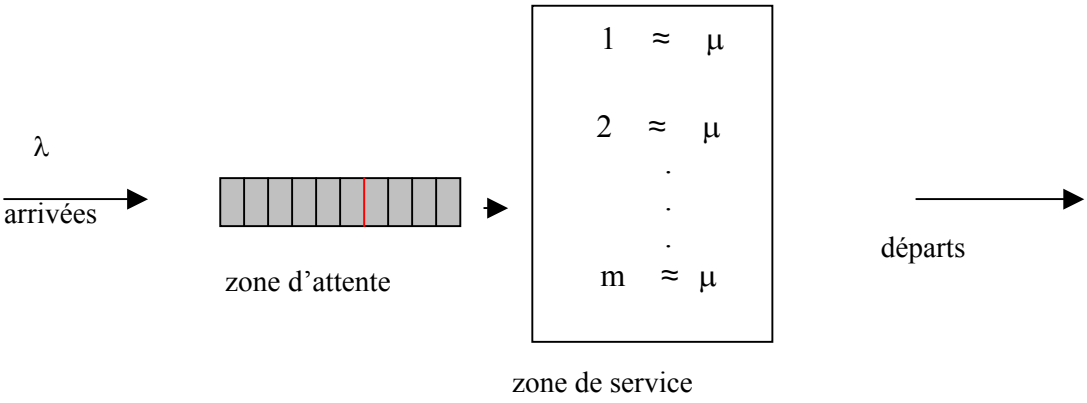


Figure 4 : Schéma simplifié d'un système de files d'attente

Le serveur est souvent assimilé à l'automate et le temps de service (ou temps d'émission) est un exemple de temporisation d'une transition. *Un réseau de files d'attente* est un ensemble de systèmes (appelés aussi *stations, nœuds* ou *routeurs*) 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). L'analyse réseau prend sa signification dès que l'on cherche à mettre en évidence les

dépendances entre les différentes stations (ou routeurs). En fait, la notion de système ou réseau est relative, et relève de l'étape de modélisation. Un réseau est bien entendu un système lui-même ; la distinction entre les deux réside uniquement dans la méthode d'analyse : il s'agit en général d'exploiter les résultats connus pour les systèmes et de les adapter à l'étude du réseau, par exemple en décomposant le réseau en sous-systèmes plus simples à étudier.

Type de système	Serveurs	Clients
Cabinet Médical (Poste,Banque)	Médecin (Guichets)	Malades (Clients)
Central téléphonique	Lignes téléphoniques	Abonnés (appels)
Aéroport (Port)	Pistes d'atterrissage (Quais)	Avions (Navires)
Atelier (Production,Maintenance)	Machines, convoyeurs (Réparateurs)	Pièces à traiter (Machines en panne)
Système Hydraulique(Réseau routier)	Barrage (Feux lumineux)	Volumes d'eau (Véhicules)
Systèmes informatiques (SGBD , routeur, serveur web…)	Ressources (processeurs, mémoires, périphériques) (Base de données)	Informations (messages, programmes, images, sons(transactions)
Système de Reconnaissance (Vision, Criminologie,…	Classifieur	Formes (caractères, images, sons,…
Serveur Web	Temps de service=temps de transfert d'une page Web	Arrivées des sessions (http)
Assurances	agence	sinistres

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). Par exemple, on note M/M/2 (arrivées poissoniennes, service exponentiel, $m = 2$ serveurs, capacité de la file infinie, source infinie) ; M/G/2 (loi de service générale ou arbitraire).

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

Remarques: (i) Du point de vue conceptuel, mathématique, le symbole *G* signifie qu'aucune hypothèse particulière n'est adoptée concernant la loi de probabilité. Du point de vue programmation, cela signifie que le programme comporte une bibliothèque de routines ou primitives pouvant faire appel à n'importe quelle loi selon l'application. Si cette bibliothèque est absente, le programme doit pouvoir permettre d'introduire n'importe quelle loi. (ii) Cette notation n'est pas très standardisée : ici K est la capacité du système, alors que dans certains ouvrages, ce symbole (noté autrement parfois) désigne la capacité de la file d'attente (service non compris). (iii) Cette notation a été adaptée aux problèmes d'ordonnancement discutés un peu plus loin.

4.3.Types de problèmes :

4.3.1. Etude de stabilité : 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. Si le système n'est pas stable, il y a congestion et il est inutile de procéder à une analyse de performance qui peut s'avérer coûteuse et inutile ; la conception du système doit être revue au moins du point de vue stabilité.

4.3.2. Analyse et Evaluation des performances du système: Le but de l'analyse est la description du système de manière à fournir de l'information sur son comportement. Cela se traduit par l'étude des principales caractéristiques du système qui correspondent souvent aux mesures (métriques) de *performances du système* . Ce sont :

- intensité (ou taux) du trafic.
- 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 dans la file;
- temps moyen de séjour $E(V)$ dans le système (attente +service) ou délai;(dans certains ouvrages, cette caractéristique est elle-même appelée temps d'attente).
- probabilité (parfois taux) qu'un client qui arrive dans le système ne soit pas accepté pour service pour cause de congestion (*probabilité* (parfois *taux) de refus , de perte* ou *d'encombrement* P_r) ;
- trafic offert et écoulé

- taux d'arrivées effectif $\bar{\lambda}$ ou λ_{eff} ;
- 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 $E(SO)$;
- coûts : d'attente, d'exploitation du système, d'inactivité (ou d'oisiveté)…

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.

4.3.3. Contrôle du système. Le contrôle est lié à des problèmes décisionnels et cette partie s'intéresse aux modèles normatifs qui permettent de calculer la configuration *optimale* du système: (i) Analyser des scénarios appropriés correspondants à des configurations données; (ii) Choisir la meilleure alternative par rapport à un critère de qualité (économique par exemple, ou technologique). La figure suivante illustre l'évolution des coûts d'exploitation du système. On note que le coût de service augmente avec le niveau de service et que le coût d'attente décroît. On remarque enfin l'existence d'une valeur optimale dans la courbe du coût total.

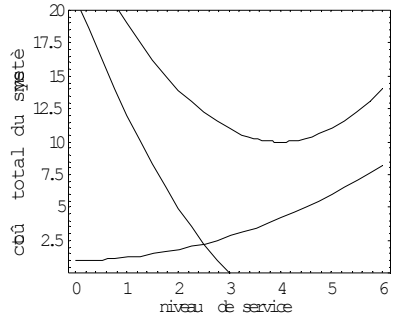


Figure 5. Evolution des coûts d'exploitation du système versus le niveau de service.

Voici d'autres critères possibles de performance (ou de qualité).

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	$E(C)$	Minimiser
Temps d'inactivité	$E(I)$	Minimiser
Disponibilité	$E(D)$	Maximiser
Probabilité de refus	π_k	Minimiser

Les variables de contrôle ou de décision 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). Ce peut être l'ordonnancement dans une même file, entre plusieurs files d'une même station (dans le cas de priorités) ou entre stations (dans le cas de réseaux).

4.4. Modèles markoviens. 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 (mesures de performance). Considérons quelques modèles markoviens décrits par des processus de naissance et de mort et qui conduisent à des solutions simples

4.4.1.Modèle M/M/1. 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. On vérifie sans peine (voir chapitre sur les chaînes à temps continu) que la probabilité d'arrivée d'un client durant $(t,t+h)$ est égale à $\lambda h+o(h)$; la probabilité qu'un client soit servi durant $(t,t+h)$ (sachant que le serveur est occupé) est égale à $\mu h+o(h)$. Représenter à titre d'exercice le graphe des états et en déduire que le processus $N(t)$ est un processus de naissance et de mort de paramètres λ et μ.

La chaîne est irréductible et apériodique. Si de plus, $\rho=\lambda/\mu<1$, alors elle est récurrente positive et il existe un régime stationnaire : les limites $\pi_i=\lim_{t\to\infty}P\big(N(t)=i\big)=$ probabilité pour qu'il y ait i clients dans le système en régime stationnaire sont strictement positives et leur somme est égale à 1. Ceci est une indication d'absence de congestion. Si par contre $\rho>1$, alors tous les $\pi_i=0$ (la chaîne est récurrente nulle ou transitoire), et dans ce cas la file va à l'infini (il y a saturation ou congestion). La conception du système doit être revue en contrôlant ou modifiant certains paramètres contrôlables (par exemple augmenter le nombre de serveurs ou l'intensité de service, modifier la règle d'ordonnancement). Ce système simple possède des propriétés remarquables :

(i) la loi stationnaire du nombre de clients dans le système est une loi géométrique de paramètre $\rho=\dfrac{\lambda}{\mu}$ i.e. $\pi_i=\rho^i(1-\rho),\ i=0,1,2,...$

En effet, les équations de Chapman-Kolmogorov s'écrivent en régime stationnaire (lorsque $t\to\infty$, et si un tel régime existe):

$$\lambda\pi_{i-1}+\mu\pi_{i+1}-(\lambda+\mu)\pi_i=0,\quad i=1,2,...$$
$$\lambda\pi_0-\mu\pi_1=0$$

Il n'est pas difficile de résoudre ce système par récurrence. Pour $i=1$, on a $\pi_1 = \frac{\lambda}{\mu}\pi_0 = \rho\pi_0$. Reportons l'expression de π_1 dans la seconde équation pour $i=2$, on a $\lambda\pi_0 + \mu\pi_2 (\lambda+\mu)\pi_1 = 0$, d'où $\pi_2 = \rho\pi_1 = \rho^2\pi_0$ …De proche en proche, on obtient finalement la solution à une constante π_0 près sous la forme $\pi_i = \rho^i\pi_0$. La constante inconnue s'obtient en utilisant la condition de normalisation $\sum_{i=0}^{\infty}\pi_i = 1$. On trouve que $1 = \sum_{i=0}^{\infty}\pi_i = \pi_0 + \sum_{i=1}^{\infty}\rho^i\pi_0 = \pi_0\left(\sum_{i=0}^{\infty}\rho^i\right)$. Par suite, $\pi_0 = \left(\sum_{i=0}^{\infty}\rho^i\right)^{-1}$. Nous sommes en mesure maintenant de préciser les conditions d'existence d'un régime stationnaire et son interprétation. Notons d'abord (on peut le constater à partir du graphe) que si $\lambda > 0, \mu > 0$, alors tous les états communiquent entre eux, et la chaîne est donc irréductible. De plus, elle est apériodique. Considérons maintenant la série $S = \sum_{i=0}^{\infty}\rho^i$ ci-dessus. C'est une progression géométrique de pas ρ. On distingue deux cas :

(a) Si $\rho > 1$, la série est divergente $S = \infty$, et $\pi_0 = 0$. Par suite toutes les probabilités $\pi_i = 0, \forall i$; Dans ce cas, les états sont tous transitoires ou récurrents nuls (proposition 4, chap 3) , et il n'y a pas de distribution stationnaire. Dans ce cas, le nombre de clients dans le système à l'instant t, $N(t) \to \infty$, et la file d'attente croît indéfiniment, ce qui correspond à la saturation du système (on parle encore d'engorgement ou de congestion).

(b) Si par contre $\rho < 1$, alors la série $S < \infty$ est convergente, et on peut calculer la somme de la série $S = \frac{1}{1-\rho}$. Par conséquent, $\pi_0 = 1-\rho > 0$ et tous les $\pi_i = \rho^i(1-\rho) > 0, \forall i$: la chaîne est récurrente positive donc ergodique . Il existe une seule distribution stationnaire qui est la loi géométrique de paramètre ρ.

(c) Pour ρ=1, les caractéristiques du système sont instables et il faut une étude particulière.

Par conséquent, si $\rho < 1$, le vecteur $\pi = (\pi_0, \pi_1, \pi_2, ...)$ forme la distribution stationnaire qui est unique.

(ii) la loi stationnaire du temps d'attente d'un client quelconque sachant que le serveur est occupé (avec une probabilité ρ) est une loi exponentielle de paramètre $\mu(1-\rho)$. Elle est égale à zéro (avec une probabilité 1-ρ i.e. si le serveur est libre, par conséquent la fonction de répartition du temps d'attente s'écrit $F_w(x) = 1 - \rho e^{-\mu(1-\rho)x}$, $x \geq 0$

(iii) La loi stationnaire du temps de séjour dans le système (attente+service) est

exponentielle de paramètre $\mu(1-\rho)$ i.e. $F_V(x) = 1 - e^{-\mu(1-\rho)x}$, $x \geq 0$

(iv) Les principales mesures de performance du système sont résumées dans le tableau récapitulatif (annexe C).

Montrons comment obtenir quelques unes des mesures de performance moyennes.

- **Nombre moyen de clients dans le système**

$$\overline{N} = \sum_{i=0}^{\infty} i.\pi_i = \sum_{i=0}^{\infty} i.\rho^i(1-\rho) = \frac{\rho}{1-\rho}$$

La figure 6 représente le nombre moyen de clients dans le système en fonction de l'intensité ρ .

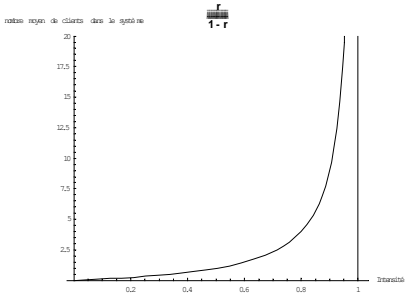


Figure 6. Nombre moyen de clients dans le système versus l'intensité ρ.

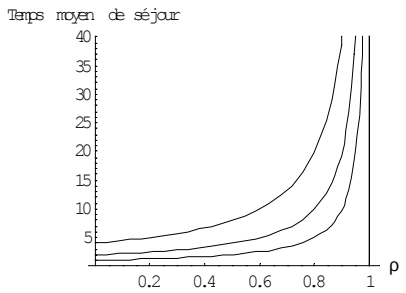


Figure 7. Courbe Temps moyen de séjour versus l'intensité ρ.

- **Temps moyen de séjour dans le système.** (*Loi de conservation ou Formule de Little*). Cette caractéristique peut être calculée directement à partit de la loi du temps de séjour $F_V(x)$ puisqu'elle est connue. Cette distribution peut être difficile à évaluer pour des modèles plus complexes. Aussi, on a coutume d'utiliser une méthode plus simple basée sur les *lois de conservation* qui reste une loi valable pour des modèles très généraux. La loi de conservation pour le système à un serveur (en particulier M/M/1) stipule que le nombre moyen de requêtes dans le système est égal au taux d'arrivées multiplié par le temps moyen de séjour $\overline{N} = \lambda.\overline{V}$, où $\overline{V}$ représente le temps moyen de séjour du client. On peut justifier intuitivement ce fait en remarquant que le client qui arrive dans le système doit trouver (en moyenne) autant de clients qu'il n'en laisse derrière lui à son départ du système , soit $\overline{N}$. Une requête fixée passe en moyenne $\overline{V}$ unités de temps dans le système et durant ce temps il arrive λ nouvelles requêtes par unités de temps, soit $\lambda\overline{V}$ requêtes en tout.

Soit $a(t)$ le nombre d'arrivées durant la période de temps (0,t) et $b(t)$ le nombre de départs (requêtes servies) au cours de cette période. Le nombre de requêtes N(t)

présentes dans le système (en attente ou en cours de service) à l'instant t s'écrit $N(t) = a(t) - b(t)$. Cette évolution est représentée sur la figure 8.

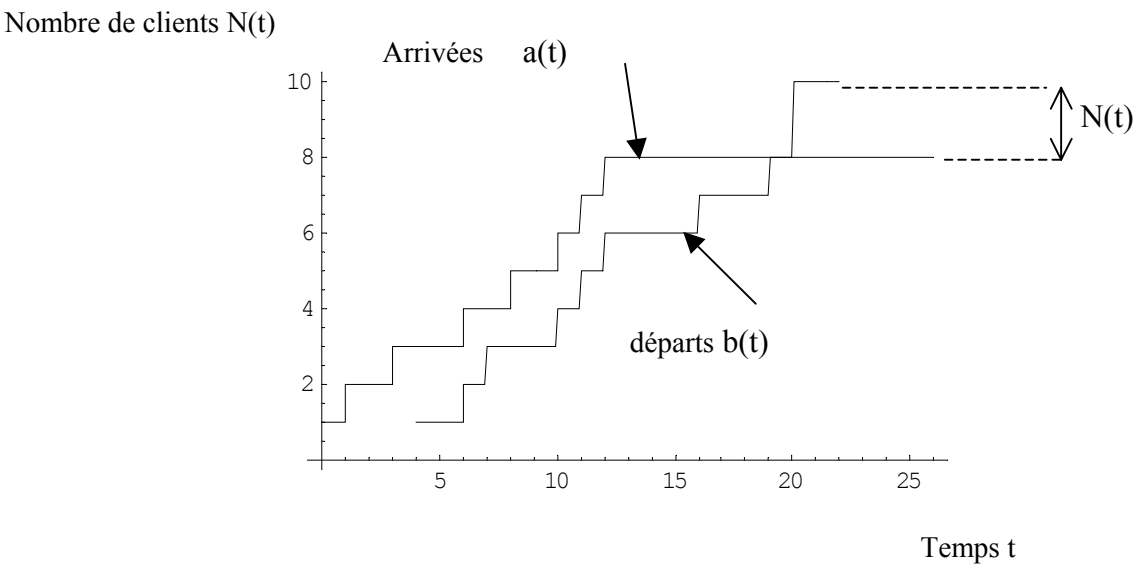


Figure 8. Evolution des arrivées et départs.

La surface comprise entre les deux courbes en escalier représente le nombre de clients présents à l'instant t , et qui a été obtenue en retranchant le nombre de départs au nombre d'arrivées. Elle représente aussi le temps total c(t) passé par l'ensemble des clients dans le système au cours de la période fixée (0,t). Comme le montre la figure ci-dessus , les deux surfaces sont identiques. Une manière alternative est de représenter le temps passé dans le système en fonction du nombre de clients, qui fournit de nouveau la même surface. L'intensité des arrivées dans le système s'écrit $\lambda(t) = \frac{a(t)}{t}$. D'autre part , définissons $\overline{V}(t) = \frac{c(t)}{a(t)}$ la moyenne (par rapport au nombre de clients présents au cours de la période t) du temps de séjour d'un client. Par conséquent, le nombre moyen de clients présents durant (0,t) s'écrit $\overline{N}(t) = \lambda(t).\overline{V}(t)$

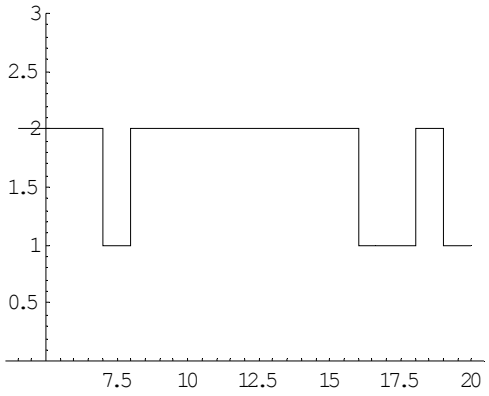


Figure 9. Evolution des arrivées séparément.

Si le système est stable et qu'il existe un régime stationnaire, alors les limites $\lambda = \lim_{t\to\infty} \lambda(t)$ et $\overline{V} = \lim_{t\to\infty} \overline{V}(t)$ existent et on obtient le résultat. La courbe suivante représente le temps moyen de séjour dans le système en fonction de la charge ρ.

- **Nombre moyen de serveurs occupés.** Le nombre de serveurs occupés D est une variable de Bernoulli et $\overline{D} = \rho$
- **Nombre moyen de clients dans la file d'attente.** On peut l'évaluer directement

$$\overline{Q} = \sum_{i=1}^{\infty} (i-1)\pi_i = \frac{\rho^2}{1-\rho}.$$

On peut le trouver d'une autre manière en notant que $\overline{Q} = \overline{N} - \overline{D}$.

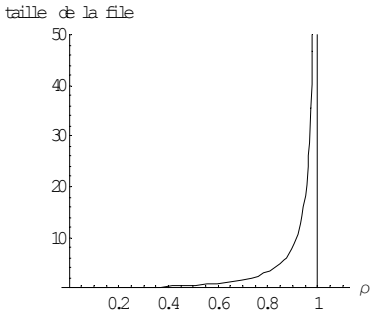


Figure 10. Courbe de la taille moyenne de la file en fonction de la charge ρ

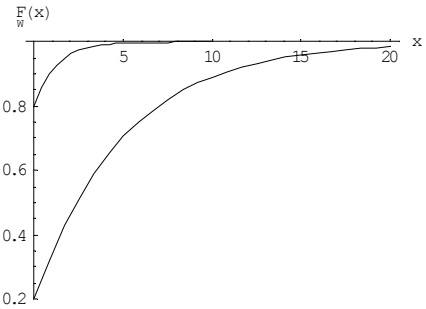


Figure 10a. Distribution du temps d'attente

- **Temps moyen d'attente.** Une manière directe est d'utiliser la loi du temps d'attente puisqu'elle est connue. On peut utiliser une méthode plus simple, valable pour des modèles plus complexes, qui utilise la loi de conservation pour le temps d'attente $\overline{Q} = \lambda \overline{W}$. Quand à la loi du temps d'attente, elle a été obtenue dans le point (ii), et représentée sur la figure 10a pour $\mu = 1; \rho = 0.2 \, et \rho = 0.8.$

- **Trafic offert.** On appelle *trafic offert* le pourcentage de temps pendant lequel une ressource (un serveur) est occupée. L'unité de mesure est l'Erlang. On dira qu'une ligne téléphonique occupée à 100% possède un trafic d'un Erlang. Les statistiques des années 2000 fournissent un ordre de grandeur et indiquent pour valeurs typiques d'une ligne résidentielle fixe un trafic de 70 mE . La valeur typique pour une ligne industrielle est de 150 mE et elle est de 25 mE pour un téléphone mobile. Le trafic offert d'une ligne téléphonique peut être ainsi défini par [26]

$$a = \frac{1}{T} \times \sum_{k=1}^{N(T)} S_k = \frac{N(T) \times E(S)}{T}$$

où T est la durée de la période d'observation ; S_k la durée du k-ème appel ; $E(S)$ la durée moyenne d'un appel ; $N(T)$ le nombre d'appels pendant la période d'observation, a le trafic en exprimé en Erlang.

- **Débit.** Dans le modèle M/M/1, il n'y a pas de contrainte, toute requête sera servie tôt ou tard**.** Par conséquent, le débit absolu $\overline{A} = \mu \overline{D} = \lambda$, et le débit relatif $A' = 1$.

4.5. Exercices.

Exercice 1 : (Application numérique du système M/M/1). Cet exercice montre également comment estimer les paramètres du modèle à partir des observations d'un échantillon.

Nombre de clients arrivant pendant une période de 5 minutes	Nombre observé m_i	Fréquence observée $f_i = \frac{m_i}{n}$	Fréquence théorique $p_i = P(X = i) = \frac{(1.26)^i}{i!} e^{-1.26}$	Nombre théorique $n_i = 100 \times p_i$
0	29	0.29	0.283654	28.36
1	34	0.34	0.357404	35.74
2	24	0.24	0.225165	22.51
3	9	0.09	0.094569	9.45
4	3	0.03	0.0207853	2.07
5	1	0.01	0.0075	0.
Total	100	=1	≈1.	≈100

Le propriétaire d'un magasin désire effectuer une analyse de performance de son activité. Il commence par réaliser une analyse statistique de la demande au niveau de ce service. Il a ainsi dénombré le nombre d'arrivées pendant 100 intervalles de 5 minutes et on a obtenu le tableau ci-dessus. Ici, f_i =nombre d'intervalles de 5 minutes où on observe i arrivées divisé par le nombre total d'intervalles, soit 100, $i = 0,1,2,3,4,5$. Ce tableau représente la loi empirique des arrivées que l'on peut représenter graphiquement sous forme d'histogramme.

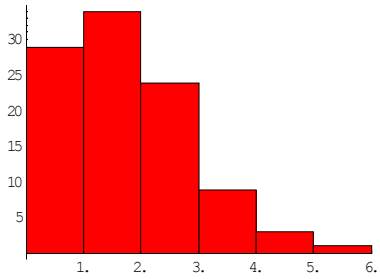


Figure 11. Distribution expérimentale des observations

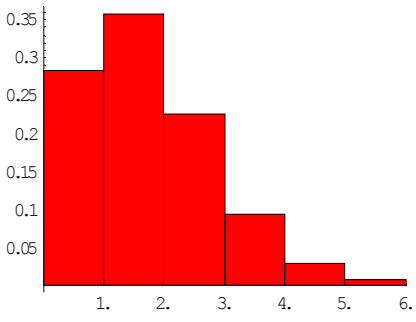


Figure 12. Distribution de la loi de Poisson de paramètre λ=1.26

On peut calculer les caractéristiques (paramètres de position) de cette distribution expérimentale : la moyenne empirique $m = 1.26$. La forme de l'histogramme (figure 11) suggère que la distribution pourrait être une loi de Poisson. En effet, la densité de la loi de Poisson de paramètre $\lambda = 1.26$ est de la forme représentée sur la figure 12. L'écart entre les distributions expérimentale et théorique (de Poisson) peut être mesuré à l'aide de la mesure du Khi-deux $\chi^2 = \sum_{i=1}^{5} \frac{(m_i - n_i)^2}{n_i} = 0.1903$. Si l'hypothèse d'une loi de Poisson est vraie, cet écart doit être plus petit qu'une valeur théorique correspondant à une loi du khi-deux à d=5−1−1=3 degrés de liberté pour un niveau de confiance de 95%. Cette valeur, lue dans des tables statistiques ou fournies par les logiciels spécialisés comportant des modules statistiques est égale à $0.352 > 0.1903$.On peut donc pour la suite supposer que les arrivées suivent une loi de Poisson et le taux d'arrivées est de $\lambda = 1.26/5 = 0.252$ arrivées par minute. Une étude similaire peut être effectuée pour la loi du temps de service. Les observations statistiques sont reportées dans le tableau suivant:

Temps de service en minutes	Nombre de services observés
<1 mn	23
de 1 à 2 mn	20
de 2 à 3 minutes	14
de 3 à 4 mn	12
de 4 à 5 mn	9
de 5 à 6 mn	5
de 6 à 7 mn	4
de 7 à 8 mn	5
de 8 à 9 mn	3
de 9 à 10 mn	2
de 10 à 11 mn	2
d e 11 à 12 mn	1

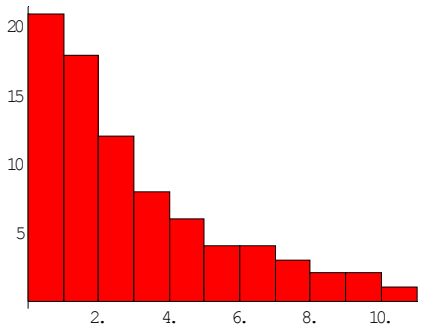


Figure 13. Distribution expérimentale du temps de service

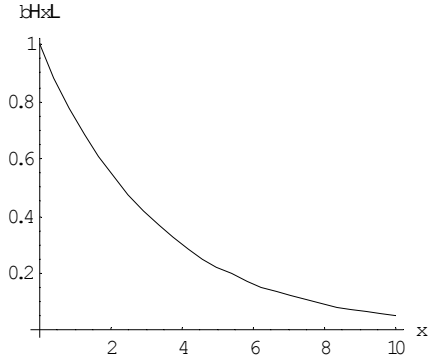


Figure 14. Distribution de la loi exponentielle de paramètre μ=0.30

La moyenne empirique est de 3.27 minutes. La loi exponentielle de paramètre $\mu = 1/3.27 = 0.30$ services par minute. Le graphe de la loi exponentielle de paramètre 0.30 est représenté sur la figure 14. La valeur calculée du khi-deux est 0.9609 et la valeur théorique de 2.167 pour un degré de liberté de $9-1-1=7$ et un niveau de confiance de 95%. La loi exponentielle ajuste donc bien les données de l'expérience. Par conséquent, si on suppose en outre que le service est *FIFO* et que la file d'attente n'est pas limitée, le modèle M/M/1 peut être utilisé pour l'analyse des performances du système (le magasin).

B. Mesures (ou métriques) de performance:

(i) Intensité du trafic $\rho = \lambda / \mu = 0.252 / 0.30 = 0.84 < 1$;

Il faut noter que cette quantité possède diverses interprétations

- Rappelons que $\rho = 1 - \pi_0$, c'est-à-dire la probabilité pour qu'il y ait un client ou plus dans le système. On peut donc l'interpréter comme la proportion de temps où le serveur (et le système) est occupé. On l'appelle parfois *coefficient d'utilisation* ou *charge* (load) du serveur. (notons toutefois que pour certains systèmes, même avec un serveur, les charges du serveur et du système ne coïncident pas forcément : systèmes avec rappels). Dans le cas de notre modèle, le trafic offert $a = \frac{\lambda}{\mu} = \rho$. Ceci permet de donner une autre interprétation de ρ.
- On peut interpréter également ici ρ comme la probabilité d'attente (i.e. qu'un client qui arrive ne puisse être servi immédiatement) . Ainsi, 84% des clients devront attendre le service. Nous avons vu également que ρ est le nombre moyen de clients en cours de service. 1-ρ=16% est la proportion de temps où le système ou le serveur est inactif (oisif) ; on l'interprète également comme la proportion de temps où le système est vide de clients.

(ii) La loi du nombre de clients dans le système suit une loi géométrique de paramètre $\rho = 0.84$. Ainsi la probabilité pour qu'il y ait k clients dans le système vaut $\pi_k = 0.16 \times (0.84)^k$. Les résultats des calculs sont reportés dans le tableau suivant.

i	0	1	2	3	4	5	6	7	8
π_k	0.16	0.1344	0.11289	0.094832	0.079659	0.066913	0.056207	0.047214	0.03966

Ici, π_i représente aussi le pourcentage de temps (sur une période d'observation du système suffisamment grande) pendant lequel le système contient i clients (i.e. se trouve à l'état i).

(iii) Nombre moyen de clients dans la file $E(Q) = 4.41$ clients

(iv) Nombre moyen de clients dans le système $E(N) = 5.25$. clients

(v) Temps moyen d'attente $E(W) = 17.5$ minutes

(vi) Temps moyen de séjour $E(V) = 20.8333$ minutes

(vii) Débit absolu= 0.252 (car tous les clients seront servis)

(viii) Débit relatif 1

C. Contrôle du système. Les clients se plaignent du temps d'attente trop long. On peut envisager diverses solutions pour améliorer la fonction production.

1. **Optimisation du taux de service.** Supposons que l'on connaisse le coût d'exploitation de l'activité du magasin $c_1 = 1800$ DA/heure ou 30 DA/minute (différentes charges) et le coût unitaire de séjour d'un client $c_2 = 15$ DA/minute (cela peut être compris par exemple comme un manque à gagner, du fait que des clients potentiels sont découragés par un temps d'attente trop long). On cherche à savoir quelle est l'intensité optimale de service minimisant les coûts unitaires d'exploitation du système. Les coûts moyens unitaires d'exploitation s'écrivent

$$C(\mu) = c_1 \mu + c_2 E(N) = c_1 \mu + c_2 \frac{\lambda}{\lambda + \mu}$$

La valeur optimale de $\mu^* = \lambda + \sqrt{\frac{\lambda c_2}{c_1}} = 0.606965$ services par minute au lieu des 0.3 services par minute actuel. Le temps moyen d'attente avec cette valeur serait de 1.16964 minutes.

2. **Optimisation du nombre de serveurs.** La solution proposée précédemment n'est pas toujours techniquement envisageable : le serveur (qu'il soit humain ou technique) a-t-il la capacité « physique » de doubler son rythme de travail ? Une alternative, peut-être plus rentable serait d'affecter un (ou plusieurs autres) serveur(s) qui fonctionneraient en parallèle. Noter cependant que le fait de doubler le nombre de serveurs par exemple ne réduira pas forcément le temps d'attente (la taille de la file, ou le coût d'exploitation) de

moitié. Soit c_1 le coût unitaire lié à l'addition d'un serveur supplémentaire, et c_2 le coût unitaire d'attente par client comme précédemment. Les coûts d'exploitation s'écrivent maintenant $C(m) = mc_1 + c_2 N(m) \to \min_m$.Il est intuitivement clair qu'à partir d'une certaine valeur de m, l'addition de serveurs n'est plus rentable. Ceci est illustré sur le graphe suivant correspondant aux données de l'exercice précédent, lorsque les coûts valent $c_1 = 5$ et $c_2 = 25$.

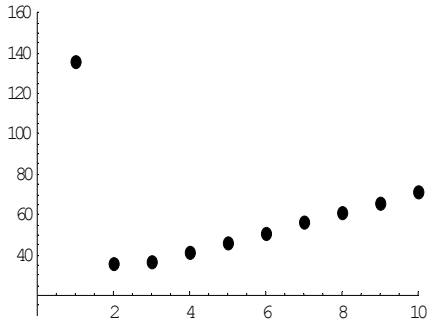


Figure 15. Evolution du coût en fonction du nombre de serveurs m.

Le graphe ci-dessus montre la zone qui contient l'optimum où l'on voit clairement que la valeur optimale est m*=2. La fonction C(m) est convexe en m, et par conséquent la technique heuristique bien connue consistant à ajouter un serveur jusqu'à obtention de l'optimum peut être utilisée. Cette technique consiste en l'algorithme suivant :

- incrémenter la valeur de m jusqu'à ce que la condition suivante soit réalisée

$$C(m-1) \geq C(m) \quad \text{et} \quad C(m+1) \geq C(m)$$

En fait, on peut améliorer la recherche en remarquant que la condition précédente s'écrit

$$N(m) - N(m+1) \leq \frac{c_1}{c_2} \leq N(m-1) - N(m)$$

Le code 9 (annexe C) confirme ce résultat. Représentons les résultats de l'exécution de ce programme qui simule 7 scénarios.

m	$N[m]$	$N[m]-N[m-1]$		$C[m]$
1	5.25			136.25
2	1.01991	4.23009	$\frac{c_1}{c_2} = \frac{5}{25} = 0.2$ ←	35.4978 ←
3	0.862899	0.157013		36.5725
4	0.843012	0.0198879		41.0753
5	0.840365	---		46.0091
6	0.84004	---		
7	0.840004	---		

La procédure a été automatisée au moyen d'un test If qui fournit la valeur 0 tant que la condition n'est pas réalisée. Elle fournit ainsi la liste où la première valeur non nulle est l'optimum. $m^* = 2$: $\{0,2,3,4,5,6,7,8,9,10\}$

Exercice 2. Reprendre l'exercice 1 avec d'autres problèmes de contrôle: sur quelles variables peut-on agir ? Ecrire la fonction coût ainsi que le programme d'optimisation correspondant.

Dimensionnement du système à capacité limitée. En guise d'illustration, considérons le problème de dimensionnement du système à capacité limitée M/M/1/K . A priori, on peut augmenter les performances d'un tel système en augmentant le taux de service (pour accélérer le service), sur la capacité du système (pour diminuer les pertes de clients insatisfaits et donc augmenter le débit), et éventuellement augmenter le nombre de serveurs. Définissons outre les coûts précédents, les coûts suivants:

c_3 = coût unitaire d'augmentation de la capacité d'attente

c_4 = dépenses liées à l'impossibilité d'accepter un nouveau client dans la file (ou bien manque à gagner par la perte d'un client insatisfait)

La fonction coût s'écrit alors

$$C(\mu, K) = c_1 \mu + c_2 E(N) + c_3 K + c_4 \lambda \pi_K \rightarrow \min_{\mu \in IR^+, K \in \{1,2,...\}}$$

Les premiers termes sont similaires au cas étudié précédemment. Le troisième donne le coût des K places dans le système ; et le dernier terme exprime le coût des pertes : en moyenne λ clients arrivent dans le système et trouvent le système bloqué avec une probabilité π_K . On aurait pu ajouter une variable m et considérer la fonction-coût $C(\mu, m, K)$ à trois variables, deux entières et l'autre continue. Il est « pratiquement » impossible de trouver une solution analytique explicite à ce problème. Il est possible dans certains cas de faire appel aux techniques de programmation mathématique, quoique l'existence de variables mixtes complique la situation. Les techniques de simulation évoquées au chapitre 3 et présentées au chapitre 5 permettent d'aborder cette simulation d'une autre façon expérimentale et visuelle.

Problème d'admission. Un problème qui paraît être développé ces dernières années dans les systèmes informatiques actuels est le contrôle de l'admission ou en d'autres termes des arrivées. L'optimisation du seul taux d'arrivées est en général peu intéressant quoique des modèles simples puissent être formulés de la même manière que pour les problèmes ci-dessus ;

Exercice 3: Pour chacun des modèles à discipline FIFO 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. Discuter.
5. Discuter les problèmes d'optimisation associés à ce modèle.

Les solutions sont résumées dans les tableaux 11 et 12 et peuvent être obtenus à l'aide des codes Mathematica en annexe.

(i) **Modèle M/M/1.** $\lambda_k = \lambda \ \forall k;\ \mu_k = \mu \ \forall k$ **;** $K = L = \infty$**.** Montrer que

(a) la loi stationnaire du nombre de clients dans le système est une loi géométrique de paramètre ρ i.e. $\pi_k = \rho^k (1-\rho),\ k = 0,1,2,...$

(b) la loi stationnaire du temps d'attente d'un client quelconque sachant que le serveur est occupé (avec une probabilité ρ) est une loi exponentielle de paramètre $\mu(1-\rho)$. Elle est égale à zéro (avec une probabilité 1-ρ i.e. si le serveur est libre, par conséquent la fonction de répartition du temps d'attente s'écrit $F_w(x) = 1 - \rho e^{-\mu(1-\rho)x},\ x \geq 0$

(c) La loi stationnaire du temps de séjour dans le système (attente+service) est exponentielle de paramètre $\mu(1-\rho)$ i.e. $F_V(x) = 1 - e^{-\mu(1-\rho)x},\ x \geq 0$

(d) Pour ce modèle, la minimisation du temps d'attente et la maximisation de l'utilisation du serveur sont-ils des objectifs compatibles ?

(e) Dans la perspective d'améliorer le temps de service, on étudie les deux scénarios suivants :
Scénario 1 : Acquérir un serveur deux fois plus rapide (un taux de service égal à 2μ.
Scénario 2 : Acquérir un second serveur de même taux μ fonctionnant en parallèle avec l'ancien.
Quelle est la meilleure variante ?

(ii) **Modèle M/M/1/K** La capacité du système est finie : $\lambda_k = \lambda$ si $k \leq K$; $\lambda_k = 0$ si $k > K$; $\mu_k = \mu,\quad k = 1,2,...,K.$

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 $k \leq L$; $\lambda_k = 0$ λ_k= si $k > L$; $\mu_k = \mu,\quad k = 1,2,...$

Exemples :

- Système constitué de L machines (les clients) et d'un réparateur (le serveur) ; la capacité d'attente est illimitée (*Repairman problem*).
- Une machine de production et une surface à capacité de stockage limitée
- Réseau d'ordinateur : Un serveur central connecté à L terminaux.

Exercice 4 : On considère deux ordinateurs qui sont reliés par une ligne de $64\ kbits/\text{sec}$ 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 $8\ kbits/\sec$ 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 $64\ kbits/\sec$. Cela revient à un seul système de taux de service μ=4x8=32 paquets/seconde et un taux d'arrivées de $\lambda = 2 \times 8 = 16$ paquets/s. Quelle est la meilleure variante ?

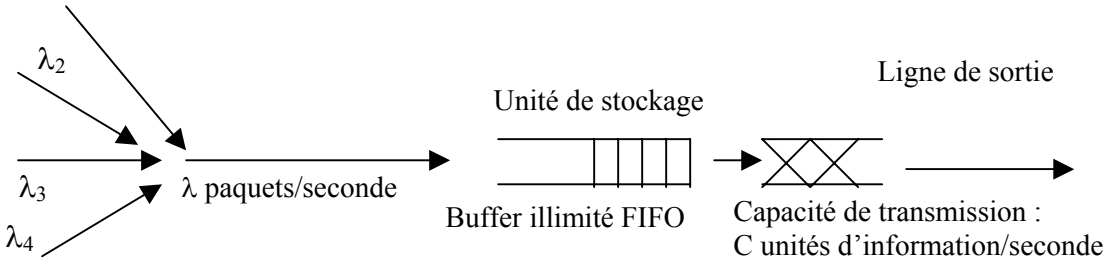
Exercice 5 : (retour à l'exemple (ii).3). Les demandes provenant des terminaux sont traitées en FIFO et une seule à la fois. L'usager a un comportement cyclique avec une phase de réflexion (usager oisif), pendant laquelle le programmeur réfléchit et tape sa demande, suivie d'une phase d'attente de la 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 : $M = 4, \quad 1/\lambda = 25 \quad \sec ondes \; 1/\mu = 100 \quad \sec ondes$.

Exercice 6 : On considère un système *M/M/m* de paramètres λ=1 et μ=2 Déterminer le nombre optimal de serveurs qui minimise le temps d'attente dans le système.

Exercice 7: Même question pour un système M/M/1/K (λ,μ), lorsqu'on donne:c_1=coût unitaire d'exploitation du serveur, c_2=coût unitaire de séjour d'un client, c_3=coût unitaire d'une place dans la file d'attente, c_4=coût unitaire de refus d'un client qui demande service. (1) Trouver la valeur optimale de la capacité K* qui minimise les coûts unitaires d'exploitation du système. (2) Trouver les valeurs optimales de la capacité K* et du taux de service minimisant ces coûts.

Exercice 8 : 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 $\theta = 1\ 000\ bits$ et la capacité de la ligne de sortie est de $C = 9\ 600\ bits/\sec onde$, de telle sorte que le temps de service (transmission du paquet) suit une loi exponentielle de paramètre $\theta C = 9.6$ paquets/seconde. Chaque

ligne d'entrée a une capacité de 4800 bits/seconde et délivre un trafic Poissonien de taux $\lambda_i = 2$ 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. Quelle est la capacité optimale ?



Exercice 9 : Un atelier comprend 3 machines automatiques entretenues par un réparateur La durée de bon fonctionnement de chaque machine est une variable aléatoire de loi exponentielle de moyenne $1/\lambda = 9$ heures. Le réparateur ne peut réparer qu'une machine à la fois et le temps de réparation est une variable aléatoire de loi exponentielle de moyenne $1/\mu = 2$ heures. Soit $X(t)$ le nombre de machines en panne à l'instant t.
(i) Calculer les probabilités d'état stationnaires et le nombre moyen de machines en panne. Quel est le temps moyen d'oisiveté du réparateur ? Quel est le temps moyen d'immobilisation d'une machine en panne ? Interpréter les débits absolus et relatifs en termes de productivité du réparateur. Quelle est la productivité de l'atelier (compte tenu des pannes) si chaque machine produit $n = 5$ articles par unités de temps
(ii) On suggère en guise d'approximation grossière que le parc de machines est infini (grand) de telle sorte que le processus de pannes soit poissonien avec un taux de pannes de 3 pour chaque période de 9 heures. Comparer les résultats avec ceux du point (i).
(iii) On se place dans le cas (ii) (parc infini) avec les même données. Trouver le nombre optimal de réparateurs à affecter à la maintenance des machines si le salaire (unitaire) d'un réparateur est $C_1 = 10$ *D.A./heure* et le coût unitaire d'immobilisation d'une machine est $C_2 = 20$ *D.A./heure*.

Exercice 10: Les mesures effectuées sur le portail d'un réseau montrent que les paquets de données arrivent selon un processus de Poisson de moyenne 125 paquets par seconde. Le portail met environ 2 millisecondes pour les transmettre. Afin d'analyser les performances du portail, on opte pour un modèle M/M/1/K. Le paramètre K=8 représente le buffer (zone de stockage des paquets en attente de transmission). (i) Peut-on considérer ce système comme stable ? Eventuellement sous quelle condition ? (ii) Quelle est la charge du système ? (iii) Déterminer la probabilité de perte d'un paquet ? (iv) Quel est le

taux d'arrivée réel des paquets (i.e. les paquets admis pour transmission) ? (v) En déduire le taux de perte des paquets (overflow) ; (vi) Quel est le débit relatif ? absolu ? (qui mesurent en quelque sorte les proportion et nombre espéré de paquets réellement transmis). (vii) Quel est le temps moyen de séjour d'un paquet dans le portail ? (viii) Comment choisir la capacité du buffer pour limiter la probabilité de perte à une valeur inférieure à 10^{-4} ? (ix) Utiliser le modèle M/M/1 pour répondre aux questions ci-dessus ? Que peut-on suggérer ? Avez-vous une explication ?

Exercice 11. Une clinique dispose d'un service d'urgence tenu par un seul médecin. Les malades se présentent selon un processus de Poisson de taux 96 malades par jour (24 heures) et les durées des soins sont indépendantes et suivent une loi exponentielle de moyenne 12 minutes pour chaque malade. Les malades sont soignés dans le cabinet du médecin suivant leur ordre d'arrivée et il n'y a pas de limitation de places dans le service d'urgence.

A. Donner la notation de Kendall de cette file d'attente ; préciser ses paramètres λ et μ. (1) Montrer que la condition d'ergodicité est vérifiée. (2) Donner la loi du nombre de malades dans le système en régime stationnaire et en déduire la probabilité qu'il y ait 4 malades en attente. (3) Déterminer les mesures de performance de cette file d'attente.

B. On souhaite que le nombre moyen de malades en attente dans la salle soit $\leq 1/2$ (condition *C*). (a) Une première solution est d'agir sur le temps de soin. Quelle doit être la durée moyenne de soin minimale pour que la condition *C* soit vérifiée ? Quelle est alors la probabilité qu'un malade attende plus de *2* heures ? (b) Une solution alternative serait d'affecter d'autres serveurs identiques parallèles, chacun assurant le même temps moyen de service de *12* minutes. (i) Déterminer le nombre minimal *m* de médecins nécessaires pour assurer un régime stationnaire du système. (ii) *Expliquer* (sans calculs) comment déterminer le nombre de serveurs vérifiant la condition *C*.

Exercice 12. Un *Cyber-Espace* comporte *m* postes. Les clients arrivent selon un processus de Poisson de paramètre $\lambda=2$ clients/heure. L'analyse statistique des données a montré que la durée T de la connexion de chaque client suit une loi exponentielle de moyenne $E(T)=30$ minutes. Le client qui trouve tous les postes occupés va à la recherche d'un autre *Cyber*. (i) Y a t-il une condition d'existence d'un régime stationnaire ? Si oui, laquelle ? (ii) Dans l'affirmative, évaluer la distribution de probabilités stationnaire. (iii) Calculer le nombre moyen de postes occupés et le nombre moyen de clients dans le *Cyber*. (iv) Quelle est la proportion de clients satisfaits (servis)? Non satisfaits ? (v) Combien de postes devrait-on avoir pour satisfaire 95% de la clientèle ? (vi) Combien en faudrait-il pour satisfaire 80% de la clientèle lorsque la durée moyenne de connexion est heure.

Exercice 13: *Modèle avec impatience***.** Une manière de modéliser l'impatience des clients est la suivante. On suppose qu'un client qui arrive dans la file a le choix de quitter le système immédiatement sans se faire servir. Ce choix dépend du nombre de clients présents devant lui dans la file d'attente. On notera b_i la probabilité d'impatience (quitter le système) lorsque i clients sont présents dans le système. (i) On suppose que la

probabilité d'impatience vaut $b_i = \frac{i}{i+1}$. Vérifier que l'on a un processus de, naissance et de mort de taux $\lambda_i = \lambda\left(1-\frac{i}{i+1}\right) = \frac{\lambda}{i+1}$, $i \geq 0$ et μ. En déduire que la distribution stationnaire est une loi de Poisson de paramètre ρ.

Exercice 14: Dans une fabrique de semelle de chaussures on dispose de deux ateliers contenant plusieurs machines chacun (un atelier pour les chaussures d'hommes et l'autre pour les chaussures de femmes). La qualité du travail dépend du bon réglage des machines. Dès qu'une machine se dérègle (ou tombe en panne), on fait appel à l'équipe de maintenance (constituée de 2 personnes) pour réglage (réparation). Les études statistiques sur les dérèglements et les temps de remise en marche ont montré que les pannes se produisait à raison de 4 pannes l'heure en moyenne selon un processus de Poisson et que le temps de remise en marche était exponentiellement distribué de paramètres 6 réparations/heure.

(I) **Etude de la situation actuelle.** (1) Donner la notation de Kendall du modèle dont on précisera les paramètres ; (2) Quelle est la charge ou intensité du trafic ? que peut-on en déduire ? (3) En moyenne combien de machines restent inutilisées en permanence ? (4) Quel est le nombre moyen de machines en attente de réglage ? Combien chaque machine attend-elle en moyenne sa prise en charge ? (5) Si l'heure d'immobilisation d'une machine revient à 1000 DA, quelle est la perte (ou plutôt le manque à gagner) sur un mois ? (on suppose que la production est continue sur 30 jours à raison de 8 heures par jour). (6) Quel est le coût dû seulement à l'attente d'être réparé ? (7) Quelle est la proportion de temps où l'équipe de maintenance est inactive ?

(II) Variante d'amélioration. Pour améliorer le service on examine les différents scénarios suivants :

Scénario 1 : On renforce l'équipe existante qui passe à trois personnes de telle sorte à augmenter le taux de réparation à 7 réparations par heure. Mêmes questions qu'en (I). Y a-t-il une économie de réalisée, si oui de combien est-elle?

Scénario 2 : On met en place une seconde équipe de maintenance avec un coût fixe de 50 000 DA le mois. Quel est ce nouveau modèle ? Préciser ses paramètres. Même questions que précédemment. Y a-t-il une économie de réalisée, si oui de combien ?

Scénario 3 : On garde comme précédemment deux équipes (avec le même coût fixe de mise en place), mais chacune s'occupant d'un seul atelier. L'atelier 1 (homme) comprend 25% du parc machine contre 75% pour le second. Cette solution permet de réduire les déplacements des équipes de maintenance et augmenter ainsi le taux de réparation qui passe à 8 réparations de l'heure. Quel est ce nouveau modèle ? Calculer les mesures de performance associées et l'économie réalisée dans ce cas, s'il y en a une.

Enfin, conclure quel est le meilleur choix parmi les trois scénarios. Discuter.

Exercice 15 : Soit un système d'attente M/M /1 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. (1) Ecrire les équations de

Chapman-Kolmogorov en régime stationnaire. (2) Montrer à partir du système d'équation obtenu que la fonction génératrice de la distribution stationnaire du nombre de clients dans le système est de la forme

$$\Pi(z) = \sum_{i=0}^{\infty} \pi_i \, z^i = \frac{1-\rho}{1-\rho z},$$

(3) En déduire une autre manière d'obtenir les principales mesures de performance (nombre moyen de clients dans le système, temps moyens d'attente et de séjour).Indication : dériver la fonction génératrice au point z=1 (pourquoi ?), puis utiliser la formule de conservation de Little). (4) Retrouver le résultat obtenu en cours selon lequel la loi stationnaire du nombre de clients dans le système est une loi géométrique de paramètre ρ. Indication : Décomposer la fonction génératrice en série de Taylor).

(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).

(5) Evaluer les principales mesures de performance du système. Conclusions ?

Exercice 16: Utiliser la méthode des fonctions génératrices pour prouver que le nombre de clients dans un système M/M/1 (en régime stationnaire) suit une loi géométrique. Vérifier que la connaissance seule de la fonction génératrice permet d'en déduire toutes les mesures de performance d'un tel système.

Exercice 17 : Soit un système constitué d'un serveur exponentiel de taux μ et tel que le flux d'arrivées dépend du nombre de clients dans le système. On suppose que si i clients sont présents dans le système à l'instant t, alors la probabilité d'arrivées d'une nouvelle requête dans (t,t+h) vaut $q_i = \lambda \alpha^i h + o(h)$. Trouver la distribution stationnaire des états dans le système.

Exercice 18 : La poste du quartier comprend deux (02) guichets, le premier G_1 spécialisé pour les retraits et envoi de mandats, chèques etc…et le second G_2 pour les affranchissements et vente de timbres divers. Les processus d'arrivées de clients à ces guichets sont poissoniens de taux λ_1=4 clients/heure et λ_2=2 clients/heure respectivement. On suppose que le temps de service au niveau de chaque guichet obéit à une loi exponentielle de moyenne μ=12 minutes (la même pour chaque guichet). En vue d'étudier les possibilités d'amélioration des prestations on se propose de comparer les deux variantes :

V1 : conserver cette spécialisation des guichets (chacun avec sa propre file d'attente).

V2 : Regrouper les deux guichets en un seul avec une file commune, les temps de service de chacun des employés ayant la même loi exponentielle de moyenne 12 minutes.

(1).Comparer les deux variantes du point de vue du temps d'attente et de séjour de chaque client. (2) Quel est le nombre optimal de guichets qui minimisent les coûts unitaires d'exploitation si on connaît (a) le coût unitaire d'exploitation d'un serveur=10 unités monétaires/heure, (b) le coût unitaire de séjour (attente) d'un client dans le système (file)=1 unité/heure.