Série de travaux dirigés de MODSIM N°

Thème: Modèles de files d'attente.

But du travail : Pratique de modélisation d'une situation concrète (files d'attente) à l'aide d'une chaîne de Markov à temps continu à espace des états très grand (infini). Pratique de l'identification statistique du modèle. de stationnarité et stabilité . Pratique de calcul de distribution stationnaire. Interprétation de cette dernière et ce que l'on peut en extraire comme questions pratiques (liées à l'évaluation des performances). Cet exemple montre en outre la notion pratique d'infini et son intérêt.

Exercice 1: 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 2 : (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.

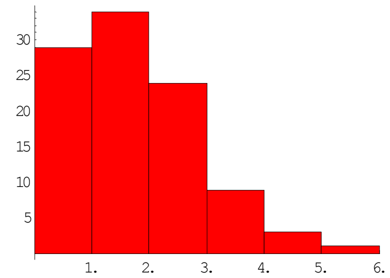

Figure 11. Distribution expérimentale des observations

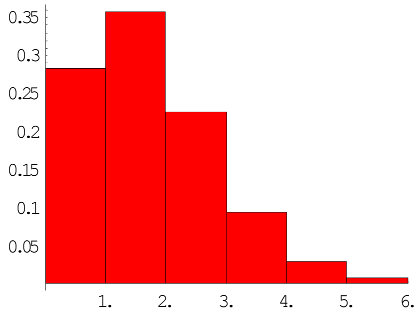

Figure 12. Distribution de la loi de Poisson de paramètre $\lambda=1.26$

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

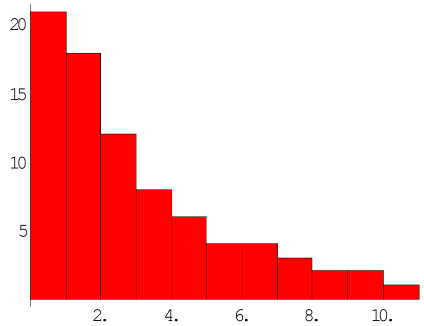

Figure 13. Distribution expérimentale du temps de service

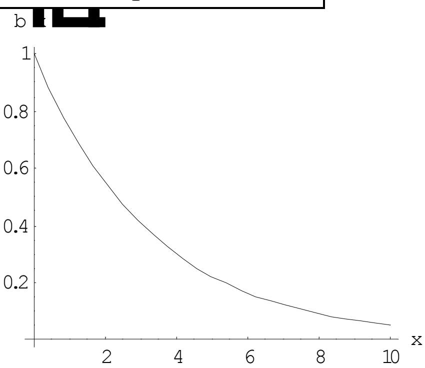

Figure 14. Distribution de la loi exponentielle de paramètre $\mu=0.30$

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 $c_2 = 25$.

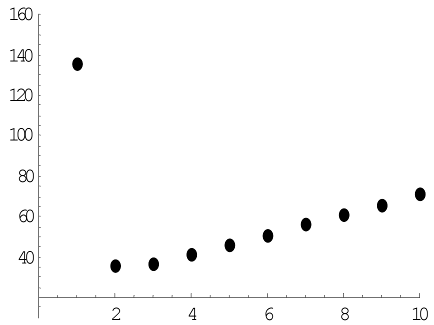

Figure 15. Evolution du coût en fonction du nombre de serveurs m.

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		35.4978
3	0.862899	0.157013	$\frac{c_1}{c_2}=\frac{5}{25}=0.2$	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. $m^*=2$: $\{0,2,3,4,5,6,7,8,9,10\}$