Étude et simulation du phénomène d'attente dans un système bancaire

Approximations des mésures de performances des files d'attente M/G/c

Donatien CHEDOM FOTSO* — Laure Pauline FOTSO**

* Département d'informatique
Université de Yaoundé I
BP 812Yaoundé
Cameroun
chedonat@yahoo.com

** Département d'informatique
Université de Yaoundé I
BP 812Yaoundé
Cameroun
l_fotso@yahoo.com

RÉSUMÉ. Des tests d'ajustements sur les intervalles de temps entre les arrivées et les durées de service d'une étude de cas ont abouti au modèle M/$\Gamma(\alpha, \beta)$/c. Il n'existe pas de formules analytiques pour calculer les mesures de performance d'un tel modèle. Nous généralisons la formule de Pollazeck-Khitchinne pour établir des formules approximatives de ces mesures et nous présentons un simulateur pour calculer les mesures de performance des modèles de file A/B/c où A, B appartiennent à l'ensemble des distributions {Déterministe, Exponentielle, Erlang, Gamma}. Nous vérifions la théorie selon laquelle le type d'organisation de la file (une file unique pour tous les serveurs ou multiple files avec une file devant chaque serveur) n'influence pas les mesures de performance du système d'attente bancaire.

ABSTRACT. Adjustment tests show that arrival times and service times in a banking queueing system of a case study led to the model M/$\Gamma(\alpha, \beta)$/c. There are no analytical formulae to calculate performance measures of such a model. In this article, we generalize Pollazeck-Khitchinne's formula to establish approximate formulae of these measures and we present a simulator that calculates performance measures of A/B/c queueing system where A, B can be one of the following distributions: determinist, exponential, Erlang or Gamma. We verify the theory according to which the type of organization of the queue (an unique queue for all the servers or multiple queues where each server has his/her own queue) does not influence the performance measures of the queueing system.

MOTS-CLÉS : Files d'attente, analyse statistique, simulation multi agent, distribution exponentielle, distribution Gama

KEYWORDS : Queues, statistical analysis, multi agent simulation, exponential distribution, Gamma distribution

1. Introduction

La Théorie des files d'attente est une technique de la Recherche opérationnelle qui permet de modéliser un système admettant un phénomène d'attente, de calculer ses performances et de déterminer ses caractéristiques pour aider les gestionnaires dans leurs prises de décisions. Des résultats et formulations théoriques sont bien établis pour les modèles de files d'attente avec arrivées poissonnières et durées de services exponentielles (M/M/c)[14][10]. Mais pas pour tous les systèmes tels que ceux avec arrivées poissonnières et durées de services non exponentielles M/G/c dont l'étude analytique est très complexe.

Nous proposons dans cet article une généralisation de la formule de Pollazeck-Kitchinne pour établir les mesures de performances du modèle M/G/c. Notre application numérique est basée sur une étude de cas d'un système bancaire d'une banque camerounaise Afriland First Bank dont les test statistiques sur les arrivées et les durées de service ont montré que leurs distributions étaient respectivement poissonnières et Gamma ; donc un modèle M/$\Gamma(\alpha, \beta)$/c.

Nous présentons un simulateur multi agent permettant d'implémenter tout système d'attente A/B/c où A, B appartiennent à l'ensemble des distributions {Déterministe, Exponentielle, Erlang, Gamma}. Ce simulateur est utilisé pour vérifier l'hypothèse selon laquelle dans un système bancaire, le type d'organisation de la file d'attente (file unique pour tout les serveurs ou files multiples, une par serveur) n'influence pas les mesures de performances.

Le reste de l'article est organisé comme suit. La section 2 présente brièvement les files d'attente bancaires ; une étude de cas est faite à la section 3 suivie par la généralisation de la formule de Pollaczeck-khitchinne aux mesures de performances des modèles M/G/c à la section 4 ; Le simulateur est présenté à la section 5. La section 6 calcul et interprète les résultats et la section 7 conclut l'article.

2. Les files d'attente

Une file d'attente est constituée des clients qui demandent un service à un ou plusieurs serveurs et d'une salle d'attente. Le taux des clients qui arrivent et le taux de service par unité de temps sont respectivement notés λ et μ. Un modèle de file d'attente est complètement décrit selon la notation de Kendall-lee[10][14] par *A/B/C/Disc/N/P* où : **A** et **B** représentent les lois des processus des arrivées et des durées de services. Par convention, M signifie exponentielle, D déterministe, $\Gamma(\alpha, \beta)$ Gamma de paramètres α et β, et G une loi générale (quelconque). **C** est le nombre de serveurs, **Disc** la discipline de service, **N** la capacité du système et **P** la taille de la population source. Lorsque les trois derniers paramètres facultatifs sont omis ils sont considérés $FCFS/\infty/\infty$ où FCFS veut dire premier arrivé premier servi (First Come, First served).

Les mesures de performances d'un modèle de file d'attente sont : Le temps moyen de séjour d'un client dans le système (W), le temps d'attente moyen d'un client (W_q), le temps de service moyen (W_s), le nombre moyen de clients dans le système (L), le nombre moyen de clients dans la file d'attente (L_q), le nombre moyen de clients en service(L_s), le taux d'utilisation des serveurs ou intensité du trafic $\rho = \dfrac{\lambda}{c.\ \mu}$ (où c est le nombre de serveurs), Les probabilités d'équilibre noté P_n (probabilité d'avoir n client dans le

système) et la probabilité d'attente Π_w (probabilité qu'un client qui arrive attende avant d'être servi).

Dans les banques, deux grandes écoles se partagent le type d'organisation des files d'attente[7],[11] : (ii) type 1, une unique file d'attente pour servir tous les guichets ; ce qui permet d'empêcher les clients de former eux même les files d'attente. (ii) type 2 chaque guichet a sa propre file d'attente laissant le client joindre la plus courte file.

3. Etude de cas

Pendant une période de deux semaines, nous avons effectué une collecte des données à Afriland First Bank (une banque privée du Cameroun) sur les arrivées des clients et les durées des services. Le hall d'entrée de l'immeuble (abritant le siège social) où se trouvaient trois guichets (serveurs) étaient notre cadre d'étude. Les tests d'ajustement statistiques ont montré que par seconde, les arrivées étaient poissonnières de paramètre $\lambda = 0,015$ et les durées de service Gamma de paramètres $\alpha = 2,42$ et $\beta = 77,17$.

Contrairement aux modèles multiserveurs avec des arrivées poissonnières et des durées de services exponentielles M/M/c, ceux avec des distributions quelconques sont plus complexes à analyser[8]. Aucun résultat théorique exact sur le calcul des mesures de performance des modèles M/G/c n'a été formellement établi et prouvé[1][3]. Ivo Adan[1] n'a proposé une approximation que de la durée moyenne d'attente pour ces modèles M/G/c. Les seules bonnes formulations théoriques rencontrées traitent des modèles n'admettant pas d'attente, à savoir le modèle $M/G/c/G_D/c/\infty$ avec perte de clients et le modèle $M/G/\infty$ avec un nombre infini de serveurs[5][8].

4. Approximation des mesures de performance des modèles M/G/c

D'après la formule de la valeur moyenne de Pollaczek-khintchinne[2], un nouveau client qui arrive doit attendre qu'un client en service complète son service et que tous les clients présents dans la file se servent avant d'être servi à son tour. Le temps d'attente moyen est alors définie par :

$$W_q = \rho E(R) + L_q \, E(B) \qquad [1]$$

où R est le temps de service résiduel d'un client en service, B le temps de service et $\rho = \frac{\lambda}{\mu}$ la probabilité de trouver un client en service.

En accord avec Ivo Adan [1], nous supposons que le temps nécessaire pour vider la file d'attente avec c serveurs est c fois plus petit qu'avec un serveur, on obtient alors :

$$W_q = \frac{1}{c} \, .(\Pi_w E(R) + L_q \, E(B)) \qquad [2]$$

Où cette fois la probabilité qu'un client attende avant d'être servi est notée Π_w (La valeur de Π_w du modèle M/M/c peut être utilisée comme une approximation de Π_w du modèle M/G/c). R est le temps de service résiduel et B le temps moyen de service.

D'après le théorème de Little $L_q = \lambda \, W_q$. Comme $E(B) = \frac{1}{\mu}$ nous avons :

$$W_q \approx \frac{\Pi_w \, E(R)}{c(1-\rho)} \text{ avec } \rho = \frac{\lambda}{c \, \mu} \qquad [3]$$

La moyenne distribution du temps résiduel[10],[14]est : $E(R) = \frac{1}{2}\ (c_B^2 + 1)\ E(B)$

où c_B^2 est le coefficient de variation des durées de service (quotient de l'écart-type sur la moyenne).

Nous avons alors :

$$W_q = \frac{1}{c\ \mu} . \frac{\Pi_w}{2(1-\rho)}\ (c_B^2 + 1) \quad [4]$$

D'où :

$$W_s = E(B) = \frac{1}{\mu} \quad [5]$$

$$W = W_s + W_q = \frac{1}{\mu} + \frac{1}{c\ \mu} . \frac{\Pi_w}{2(1-\rho)}\ (c_B^2 + 1) \quad [6]$$

$$L_q = \lambda\ W_q = \lambda .\ \frac{1}{c\ \mu} . \frac{\Pi_w}{2(1-\rho)}\ (c_B^2 + 1) \quad [7]$$

$$L_s = \lambda\ W_s = \rho = \frac{\lambda}{\mu} \quad [8]$$

$$L = L_s + L_q = \frac{\lambda}{\mu} + \lambda .\ \frac{1}{c\ \mu} . \frac{\Pi_w}{2(1-\rho)}\ (c_B^2 + 1) \quad [9]$$

Ne pouvant établir des formules pour trouver des valeurs approximatives des probabilités d'équilibre du nombre de clients présents dans le système (P_n). Nous les avons simulées.

5. Simulation

La simulation multi agent, propose de créer un monde artificiel dans lequel interagissent des agents évoluant dans un environnement[4].

5.1. Fonctionnalités

Le simulateur programmé en Java, est multi agent. Il implémente les deux types d'organisation des files d'attente bancaires pour les modèles A/B/c où A, B appartiennent à l'ensemble des distributions {Déterministe,Exponentielle,Erlang,Gamma}.

Les paramètres d'entrée du simulateur sont : le type d'organisation de la file du modèle à simuler, la durée de la simulation, les distributions des durées de services et des arrivées et le nombre de guichets. Le simulateur fournit en sortie les mesures de performance du modèle. La figure 1 donne une vue de ce simulateur.

Figure 1. *Une vue du simulateur*

5.2. Description

Un agent est représenté par un thread indépendant, son savoir-faire par une méthode, son état par l'ensemble des valeurs de ses propriétés et son comportement par la méthode Run du thread. Les agents sont dotés de deux facultés : Percevoir et Agir. Un agent doit percevoir d'abord son environnement avant d'y agir. Le système a trois types d'agents : les clients, les serveurs et le super agent qui contrôle et cordonne le déroulement de la simulation.

Le choix des unités de temps et la détermination des valeurs qui en dépendent sont importants. L'unité de temps est régie par les taux de service et taux des arrivées des clients. Ces taux sont exprimés par rapport à la même unité de temps. Si par exemple le taux des arrivées s'exprime en nombre de clients qui arrivent par heure, alors le taux de service devra être exprimé en nombre de clients servis par heure. La durée de la simulation, doit également être exprimée dans la même unité de temps.

Dans le type 2, quand un client choisi une file, il y reste jusqu'à obtenir son service. Le simulateur utilise la méthode de la transformation inverse[9] pour générer les variables aléatoires.

Pour valider le simulateur nous avons comparé ses sorties aux résultats théoriques connus.

5.3. Calcul des mésures de performances

Calcul de L, L_q, et L_s

Nous utilisons deux tableaux d'entiers Lq[ ] et Ls[ ]. Un pas de simulation correspond à 100 ms (1/10 secondes). Au i^{eme} pas on affecte à Lq[i] (repectivement Ls[i]) le nombre de clients dans la file d'attente Qsize(respectivement nombre de clients en service Ssize). A la fin de la simulation nous avons :

$$L_q = \frac{\sum_{i=1}^{T_p} Lq[i]}{T_p}; \qquad L_s = \frac{\sum_{i=1}^{T_p} Ls[i]}{T_p}; \qquad L = L_q + L_s \qquad [10]$$

où T_p est le nombre total de pas de simulation. Si T est la durée de la simulation, $T_p = 10 * T$.

Calcul de W, W_q et W_s

Nous utilisons également deux tableaux d'entiers Wq[ ] et Ws[ ] correspondant respectivement aux temps d'attente et de service des clients dans le système.

Chaque agents Client possède deux variables Date (sa date de création Dc et sa date de début de service Dd) et une variable durée de service Ds. A la création du client, sa date de création est mise à jour à la date courante et l'ordinateur génère le temps avant la création du prochain client suivant la distribution des intervalles de temps entre les arrivées. Le client créé va être servi si un serveur est libre, sinon il se voit attribuer une position $Pos = Qsize + 1$ dans la file d'attente (la plus courte dans le cas à plusieurs files) qui est décrémentée au fur et à mesure qu'un serveur se libère (celui de sa file, dans le cas à plusieurs files).

Quand le i^{eme} client occupe la position 1 et qu'un serveur se libère (son serveur dans le cas à plusieurs file), il rejoint le service, sa date de début de service Dd est mise à jour à la date courante et l'ordinateur génère la durée de son service Ds suivant la distribution des durées de service. Puis on met à jour :

$$Wq[i] = Dd - Dc, \text{ et } Ws[i] = Ds \qquad [11]$$

A la fin de la simulation nous avons :

$$W_q = \frac{\sum_{i=1}^{N_q} Wq[i]}{N_q}; \qquad W_s = \frac{\sum_{i=1}^{N_s} Ws[i]}{N_s}; \qquad W = W_q + W_s; \qquad [12]$$

où N_s est le nombre total de clients servis pendant la simulation et N_q le nombre total de clients ayant attendus (la somme du nombre de clients déjà servis et du nombre de clients en service).

Calcul des probabilités d'équilibres

Nous utilisons un tableau d'entiers P[ ] dont les 200 premières composantes sont initialisées à 0 ($P[i] = 0, \text{ pour } 1 < i <= 200$). A chaque pas de simulation, on incrémente P[j] de 1 si $j = Qsize + Ssize$. A la fin de la simulation nous avons :

$$P_n = \frac{P[n]}{T_p} \qquad [13]$$

T_p représentant le nombre de pas de simulation.

6. Résultats et interprétation

Pour le type 1, nous avons utilisé les approximations établies à la section 3 et simulé les probabilités d'équilibre. Pour le type 2, il n'existe pas de modèles théoriques prenant en charge tous les aspects de cette approche, les résultats ont été simulés sous les mêmes conditions initiales.

Données initiales

L'analyse statistique de l'ensemble des données collectées de notre étude de cas a abouti à un modèle M/$\Gamma(\alpha, \beta)$/3 et nous a permis de conclure que les arrivées des clients sont poissonnières de taux $\lambda = 0,015$ par seconde et que les durées de service sont Gamma de moyenne $1/\mu = 187,38$ secondes et de variance $\sigma^2 = 14461,96$. D'où $\alpha = 2,42$ et $\beta = 77,17$.

Résultats

Le tableau 1 donne les valeurs des indicateurs de performance obtenues dans les deux cas.

Mesures de performance	file unique	files multiples
ρ	93,7%	93,5%
Π_0	0,015	0,015
Π_w	0,87	0,81
W	796,107	778
W_q	608,727	591,42
W_s	187,38	186,61
L	11,941	11,76
L_q	9,13	8,86
L_s	2,81	2,8

Tableau 1. *Mesures de performance comparées des deux approches étudiées*

La figure 2 présente les probabilités d'équilibre en fonction de n pour les deux types d'organisation de la file.

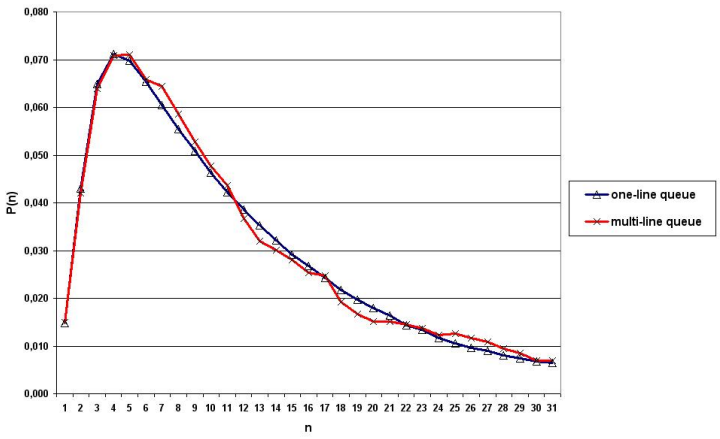

Figure 2. *Probabilités d'équilibre pour les deux types d'organisation*

On constate que l'intensité du service est élevée car les serveurs passent plus de 93% de leur temps occupés ce qui assure la banque d'une pleine utilisation de ses ressources. On a en permanence pratiquement 3 clients en service (exactement 2,81). Il y'a en moyenne 12 clients dans le système dont un peu plus de 9 clients en moyenne dans la file d'attente. Un client séjourne en moyenne 13 minutes 16 secondes dans la banque dont 10 minutes 8 secondes dans la file d'attente. Un nouvel client qui arrive a 80% de chance d'attendre avant d'être servi.

Nous constatons que les deux approches fournissent pratiquement les mêmes résultats. Nous pouvons donc conclure qu'il n'existe pas de différences significatives entre les deux types de disposition de la file si ce n'est du point de vue organisationnel (Ordre, confort

et espaces). En effet, le modèle à file d'attente unique possède certains avantages : (i) pour la banque, la présence de zéro ou une seule file d'attente assure l'ordre et le confort des clients, et une exploitation rationnelle de l'espace ; (ii) pour les clients, le système de numérotation n'exige pas la présence physique dans la file, donc ils peuvent s'asseoir et attendre ou s'occuper pendant l'attente.

7. Conclusion

Dans cet article, nous avons utilisé la formule de pollaczek-kitchinne pour proposer des formules permettant d'obtenir des approximations des mesures de performances des modèles de file d'attente M/G/c. Nous avons implémenté un simulateur multi agent qui fournit des mesures de performances des modèles A/B/c où A, B sont des distributions soit Déterministe, Exponentielle, Erlang, ou Gamma. Les application numériques avec des données issues d'une étude de cas dans une banque privé Camerounaise nous ont permis de vérifier la théorie qui dit que dans un système bancaire, le type d'organisation de la file d'attente (file unique pour tout les serveurs ou files multiples, une par serveur) n'influence pas les mesures de performances du système.

8. Bibliographie

[1] IVO ADAN, « Stochastic models for Design and Planning », University of Amsterdam 2000.

[2] IVO ADAN„ JACQUES RESING, « Queueing Theory », 2001. 188pp.

[3] IVO. ADAN,„ W.A VAN DE WAARSENBURG„ J. WESSELS, « Analyzing Ek/Er/c Queues », *ACM Portal of technologie*, 2004.

[4] FRANÇOIS BOUSQUET„ CHRISTOPHE LE PAGE„ JEAN-PIERRE MÜLLER, « Modélisation et simulation multi-agent », CIRAD.

[5] ROBERT B. COOPER , « Introduction to Queueing Theory », *Elsevier North Holland, Inc*, 1981, 347pp.

[6] J. FERBER « Les systèmes multi-agents. Vers une intelligence collective », Paris . 1999.

[7] HUYNH B. „ PHAN D. AND NGO Q.„ PHAM D, « Queues in Banks », GEM 2503 Project.

[8] RICHAR C. LARSON„ AMADEO R.ODONI, « Urban Operations Research », Prentice-Hall N.J 1997-1999.

[9] LAW„ KELTON, « Simulation Modeling and Analysis »,

[10] PHILIPPE, NAIN, « Basic Element of Queueing Theory : Application to the Modeling of Computer Systems », 1999.

[11] NICO M. VAN DIJK, « On hybrid combination of queueing and simulation », *University of Amsterdam*,

[12] JOHN S. SADOWSKY„ WOJCIECH SZPANKOWSKIY, « MAXIMUM QUEUE LENGTH AND WAITING TIME REVISITED », August 24, 1996.

[13] STALLINGS, WILLIAM, « Queuing Analysis (A Practical Guide for Computer Scientists) », 2000.

[14] ANDREA WILLIG, « A Short Introduction to Queueing Theory », *Technical University Berlin, Telecommunication Networks Group*, 1999.