

Une solution approchée pour les files $Ph/Ph/1$ et $Ph/Ph/1/N$

Thomas Begin ⁽¹⁾ et Alexandre Brandwajn ⁽²⁾

(1) Université Lyon 1 / LIP UMR CNRS - ENS Lyon - UCB Lyon 1 - INRIA 5668.

(2) University of California Santa Cruz / Jack Baskin School of Engineering.

Résumé

Nous proposons une approximation simple pour évaluer les probabilités stationnaires du nombre de clients et les probabilités d'état trouvées à l'arrivée dans les files $Ph/Ph/1$ et $Ph/Ph/1/N$. Pour cette dernière, ceci inclut la probabilité du dépassement de capacité. Les distributions de type phase considérées ici sont acycliques. Notre méthode s'appuie sur une itération entre les solutions d'une file $M/Ph/1$ à taux d'arrivée dépendant de l'état et une file $Ph/M/1$ à taux de service dépendant de l'état. Nous résolvons ces deux files à l'aide d'une récurrence simple et efficace. En itérant entre ces deux modèles, notre approximation divise l'espace d'états, et peut ainsi traiter des distributions avec un grand nombre de phases (plus de 100) nécessaires à la représentation de distributions à queue lourde qui risquent de poser problème aux méthodes numériques classiques. La méthode proposée converge généralement en quelques dizaines d'itérations. Notre approximation est asymptotiquement exacte, et sa précision est bonne : généralement à quelques pourcents de la valeur exacte, sauf quand à la fois les distributions des inter-arrivées et du temps de service présentent une faible variabilité. Dans ce cas, en particulier pour des niveaux de charge modérés, nous ne recommandons pas l'usage de notre méthode.

Keywords: File monoserveur, distribution de type phase, files $Ph/Ph/1$ et $Ph/Ph/1/N$, probabilités stationnaires, probabilités de dépassement de capacité, grand nombre de phases, solution approchée, stabilité numérique.

1 Introduction

Malgré la présence grandissante du parallélisme dans de nombreux domaines, pour beaucoup d'applications le service des requêtes (clients) demeure intrinsèquement lié à un serveur unique. C'est le cas pour les paquets sur une interface réseau ou pour les requêtes sur une base de données verrouillée. Dans de nombreux cas, le temps entre les arrivées et le temps de service présentent une forte variabilité, voire sont à queue lourde [FEL98] et plus généralement s'éloignent significativement d'une distribution exponentielle. Nous représentons le temps d'inter-arrivée et le temps de service par des distributions de type phase acycliques [BOB05, FEL98]. Il est connu que toute distribution peut être approchée aussi finement que voulu par une distribution de type phase [LAT99]. Par ailleurs, lorsqu'un système de type file d'attente est exposé à une charge de trafic importante, le modèle associé doit comporter une taille de tampon maximum pour permettre l'étude de la probabilité de dépassement de capacité. Ainsi, dans cet article nous considérons la file $Ph/Ph/1$ à tampon illimité et la file $Ph/Ph/1/N$ avec un tampon de taille N .

Bien qu'il existe une littérature considérable traitant de la file monoserveur [CHAU92, COH82, JAG88, ABA93], aucune solution simple n'existe pour le cas général (i.e., temps de service et d'inter-arrivée généraux), y compris lorsqu'il s'agit d'évaluer uniquement le nombre moyen de clients dans le cas d'une file $GI/G/1$ [BOL05, page 265]. Des méthodes numériques éprouvées pour résoudre les files de type $Ph/Ph/1$ existent (e.g. méthodes matrices-géométriques [LAT99]), mais, elles s'accommodent mal de la cardinalité de l'espace d'état lorsque le nombre de phases pour représenter les distributions s'accroît.

Les approximations existantes se limitent à évaluer le temps d'attente moyen mais elles souffrent déjà d'une description des distributions limitée à leurs 2 premiers moments [ALL90, BOL05, RAO99], et aucune ne semble adaptée pour le calcul de la probabilité de dépassement de capacité.

Une méthode simple stable numériquement a été proposée pour calculer les probabilités stationnaires du nombre de clients dans une file $M/Ph/1$ à taux d'arrivée dépendant de l'état [BRA08]. Plus récemment, une méthode analogue a été proposée pour une file $Ph/M/1$ à taux de service dépendant de l'état [BRA10]. Nous utilisons ces deux solutions par récurrence pour obtenir une approximation pour la distribution stationnaire du nombre de clients dans une file $Ph/Ph/1$ ou $Ph/Ph/1/N$. Cette approximation a l'avantage de prendre directement en compte la forme réelle des distributions du temps de service et d'inter-arrivée. Nous montrons comment en déduire les probabilités d'état vus par un client à l'arrivée dans la file, notamment celle d'un dépassement de capacité.

2 Approximation

La Figure 1 présente la file $Ph/Ph/1$ considérée ici. Nous notons a le nombre de phases dans la distribution du temps inter-arrivées, et b le nombre de phases dans la distribution du temps de service. Le nombre courant de requêtes dans le système est noté n . L'état stationnaire de cette file peut être décrit par la phase courante du processus d'arrivée j , la phase courante du processus de service (si la file n'est pas vide) i , et par le nombre total de requêtes dans le système, c'est-à-dire (j,i,n) . La Table 1 présente les notations utilisées dans cet article.

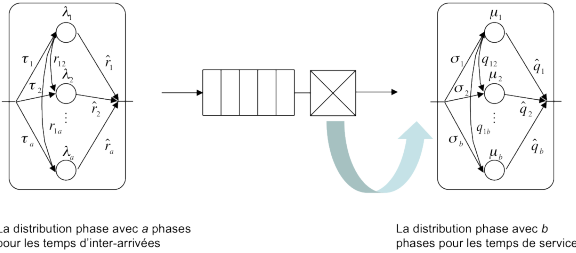


Fig. 1 – La file $Ph/Ph/1$.

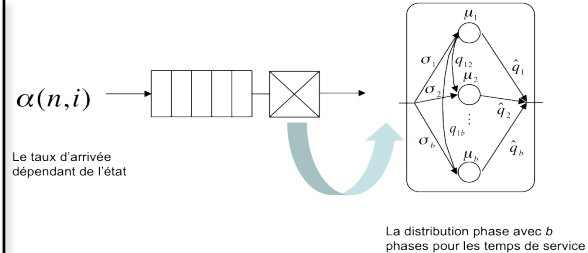


Fig. 2 – La file $M/Ph/1$.

τ_j	Probabilité que le processus d'arrivée démarre à la phase j , $j = 1, \dots, a$
λ_j	Taux de la phase j du processus d'arrivée
r_{jl}	Probabilité que le processus d'arrivée continue en phase l après la phase j , $j, l = 1, \dots, a$, $l > j$
\hat{r}_j	Probabilité que le processus d'arrivée se termine (arrivée d'une nouvelle requête) après la phase j , $j = 1, \dots, a$
σ_i	Probabilité que le service d'une requête démarre à la phase i , $i = 1, \dots, b$
μ_i	Taux de la phase i pour le processus de service
q_{ih}	Probabilité que le processus de service continue en phase h après la phase i , $i, h = 1, \dots, b$, $h > i$
\hat{q}_i	Probabilité que le processus de service se termine (la requête quitte la file) après la phase i , $i = 1, \dots, b$
n	Nombre total courant de requêtes dans le système; $n = 0, \dots, N$ pour un tampon de taille fini
$p(i n, j)$	Probabilité conditionnelle que la phase de service soit i sachant que le nombre dans le système est n et la phase d'arrivée est j
$p(j n, i)$	Probabilité conditionnelle que la phase d'arrivée soit j sachant que le nombre dans le système est n et la phase de service est i
$p(n)$	Probabilité stationnaire que le nombre de requêtes dans le système soit n
$P_A(n)$	Probabilité qu'une requête arrivant trouve n requêtes déjà présente dans le système

Table 1 – Notations principalement utilisées

En considérant la description d'état marginale (i, n) , notre file peut être représentée comme sur la Figure 2 où le taux d'arrivée des clients dépendant de l'état $\alpha(n, i)$ est donnée par

$$\alpha(n, i) = \sum_{j=1}^a \lambda_j \hat{r}_j p(j|n, i) \quad (1)$$

De façon analogue, en considérant la description d'état marginale (j, n) , notre file peut être représentée comme sur la Figure 3 où le taux de service dépendant de l'état $u(n, j)$ est donnée par

$$u(n, j) = \sum_{i=1}^b \mu_i \hat{q}_i p(i|n, j) \quad (2)$$

Nous dérivons notre approximation en considérant que $p(j|n, i) \approx p(j|n)$ et $p(i|n, j) \approx p(i|n)$. Autrement dit, nous supposons que la probabilité conditionnelle que la phase d'arrivée soit j sachant le nombre dans le système et la phase courante du processus de service dépend surtout du nombre dans le système; et, inversement, que la probabilité conditionnelle que la phase du service soit i sachant le nombre dans le système et la phase du processus d'arrivée dépend surtout du nombre dans le système. Ainsi on a $\alpha(n, i) \approx \alpha(n)$ et $u(n, j) \approx u(n)$.

La file de la Figure 2 devient alors une file $M/Ph/1$ avec un taux d'arrivée dépendant de l'état $\alpha(n)$, et la file de la Figure 3 devient une file $Ph/M/1$ avec un taux de service dépendant de l'état $u(n)$ (cf. Figure 4). Une simple récurrence numériquement stable permet d'obtenir efficacement la solution de cette file $M/Ph/1$ queue [BRA08] et donc de son taux de service $u(n)$. Une simple récurrence permet également d'obtenir les taux d'arrivées dépendant de l'état $\alpha(n)$. Ainsi, l'idée consiste à itérer entre les solutions de ces deux files jusqu'à ce qu'un point fixe sur les taux d'arrivées et de service soit trouvé. Une fois obtenus les taux $\alpha(n)$ et $u(n)$, la probabilité stationnaire du nombre de clients dans le système $p(n)$ peut être calculée comme

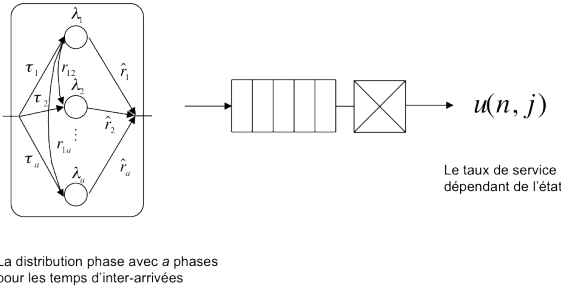


Fig. 3 - La file Ph/M/1.

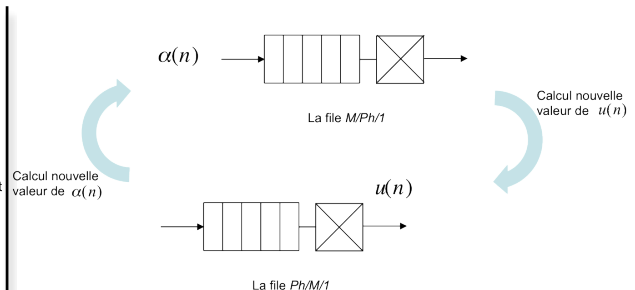


Fig. 4 - Itérations entre les files M/Ph/1 et Ph/M/1.

$$p(n) \approx \frac{1}{G} \prod_{m=1}^n \alpha(m-1)/u(m), \quad n = 0, 1, \dots \quad (3)$$

G est une constante de normalisation telle que $\sum_n p(n) = 1$. Le nombre moyen de clients dans le système est $\bar{n} = \sum_n np(n)$. La probabilité qu'un client arrivant trouve n clients déjà présent dans le système, $P_A(n)$ est

$$P_A(n) \approx \frac{\alpha(n)p(n)}{\sum_{i=0}^n \alpha(i)p(i)}, \quad n = 0, 1, \dots \quad (4)$$

Note itération part d'un taux d'arrivées égal à l'inverse du temps moyen entre arrivées pour la solution de la file M/Ph/1 (cf. [BRA11]).

3 Précision et vitesse de convergence

Nous avons intensément testé notre approximation en comparant ses résultats pour plusieurs quantités comme la probabilité qu'un client doit attendre avant d'être servi et la forme générale de la distribution de probabilité stationnaire $p(n)$ à celles données par une méthode numérique exacte. Notons que notre approximation produit toujours l'utilisation correcte du serveur dans le cas d'un tampon infini. Généralement la précision de l'approximation est bonne, à quelques pourcents de la valeur exacte. Dans la quasi-totalité des cas, un petit nombre d'itérations (quelques dizaines) suffit pour atteindre la convergence (avec une précision de 10^{-7}) vers le point fixe de notre solution approchée.

Nous présentons un exemple de comportement typique de notre méthode. Nous considérons une file Ph/Ph/1/N avec un tampon de taille $N = 15$. Le temps entre arrivées est représenté par une distribution d'Erlang à 100 phases et de moyenne égale à 1, et le temps de service est représenté par une distribution mixte d'Erlang à 4 phases de moyenne égale à 1 et de coefficient de variation égal à 3. La Table 2 montre les résultats obtenus et la Figure 5 compare les valeurs exactes et approchées pour la distribution stationnaire du nombre de clients $p(n)$.

Dans cet exemple, l'erreur relative de l'approximation reste inférieure à quelques pourcents de la valeur exacte, et la méthode converge en quelques dizaines d'itérations. Ceci semble être le comportement typique de cette méthode. Toutefois, la méthode proposée perd en précision lorsqu'à la fois le temps entre arrivées et le temps de service présentent une faible variabilité (disons, des coefficients de variation inférieurs à 0.3), en particulier pour des charges modérées. Par conséquent, notre approximation est déconseillée dans ce cas. Notons que si uniquement la distribution du temps entre arrivée ou celles du temps de service présente une faible variabilité, la précision de la méthode reste bonne.

Utilisation du serveur		Probabilité d'aucune attente		Probabilité de dépassement de capacité		Nombre d'itérations
Exact	Appr.	Exact	Appr.	Exact	Appr.	
0.8258	0.8355	0.298	0.276	0.1742	0.1645	58

Table 2 – Précision et vitesse de convergence

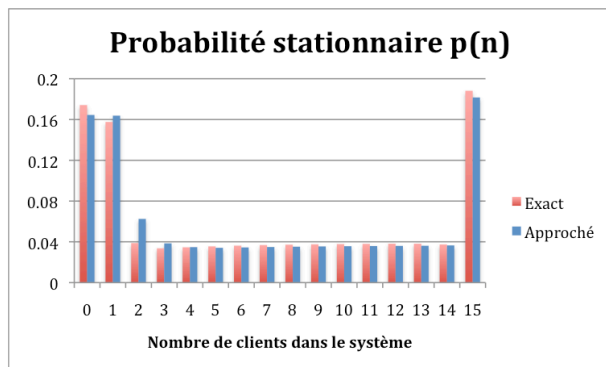


Fig. 5 – Probabilités exactes et approchées

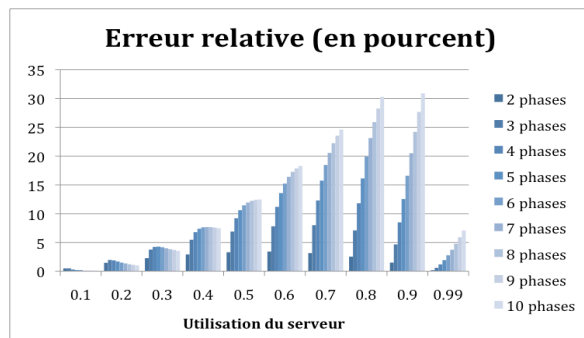


Fig. 6 – Précision de l'approximation en fonction du nombre de phases dans les distrib. Erlang du service et arrivées

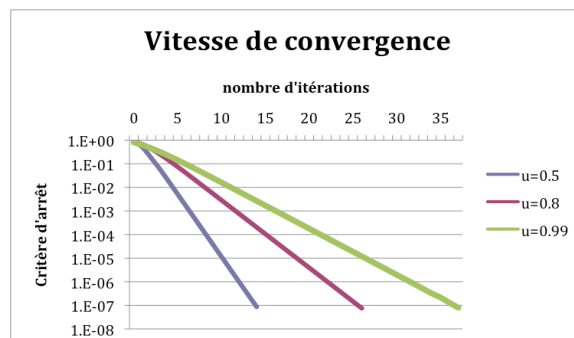


Fig. 7 – Vitesse de convergence de l'approximation pour plusieurs niveaux d'utilisation du serveur

Pour illustrer cette baisse de précision lorsqu'à la fois le temps entre arrivées et le temps de service présentent une faible variabilité, nous considérons une file $Ek/Ek/1$ avec le même nombre de phases pour les distributions des arrivées et du service. La Figure 6 montre l'erreur relative du nombre moyen de clients dans ce système en fonction de l'utilisation du serveur pour un nombre de phases variant de 2 à 10 (l'approximation est naturellement exacte pour le cas d'une file $M/M/1$.) Nous remarquons que les erreurs relatives les plus fortes tendent à apparaître pour des utilisations modérées à assez élevées (disons, de 0.6 à 0.9). Dans cette plage, la précision de l'approximation tend à se dégrader avec l'ordre de la distribution d'Erlang, et dépasse 20% avec 7 phases ou plus. On note que lorsque le serveur approche de la saturation, la précision de l'approximation s'améliore. En fait, on peut montrer que notre approximation est asymptotiquement exacte quand $n \rightarrow \infty$ [BRA11]. Cette baisse de précision est liée au fait qu'en présence de distributions hypo-exponentielles, quand le nombre de clients dans le système est faible, la connaissance de la phase courante du processus de service fournit une information non-négligeable sur la phase possible du processus d'arrivée (et vice versa), connaissance perdue dans notre approximation.

La précision de l'approximation varie en fonction de la forme des distributions d'inter-arrivée et du service. Elle tend à être particulièrement bonne lorsque ces distributions sont asymétriques (e.g. hyper-exponentielles déséquilibrées). Il est important de souligner que ces cas sont justement ceux qui posent le plus de problèmes pour les méthodes numériques classiques [CHAU92] et pour la simulation à événements discrets [ASM00].

Pour conclure, la Figure 7 illustre la vitesse de convergence de notre approximation vers son point fixe pour le premier exemple. Elle montre l'évolution de la différence relative pour le nombre moyen de clients entre les modèles $M/Ph/1$ et $Ph/M/1$ en fonction du nombre d'itérations pour plusieurs utilisations du serveur, 0.5, 0.8 and 0.99. Dans ces 3 cas, la diminution de la différence relative suit une forme géométrique. Enfin, notons que la vitesse et la stabilité numérique de la méthode deviennent particulièrement visibles par rapport à une méthode exacte numérique classique lorsque les distributions comportent un grand nombre de phases.

Références

- [ABA93] Abate, J., Choudhury, G. L. and Whitt, W. 1993. Calculation of the $GI/G/1$ waiting time distribution and its cumulants from Pollaczek's formulas. Arch. Elektr. Uebertragung 47 (Pollaczek memorial volume, 1993), 311-321.
- [ALL90] Allen, A.O. 1990. Probability, Statistics and Queueing Theory. Second Edition, Academic Press, London.
- [ASM00] Asmussen, K., Binswanger, K. and Hojgaard B. 2000. Rare events simulation for heavy-tailed distributions. Bernoulli. 6, 2, 303-322.
- [BOB05] Bobbio, A., Horváth, A. and Telek, M. 2005. Matching Three Moments with Minimal Acyclic Phase Type Distributions. Stochastic Models. 21, 2, 303-326.
- [BOL05] Bolch, G., Greiner, S., Meer, H. d. and Trivedi, K. S. 2005. Queueing Networks and Markov Chains. Second Edition, Wiley-Interscience.
- [BRA08] Brandwajn, A. and Wang, H. 2008. A conditional probability approach to $M/G/1$ -like queues. Performance Evaluation. 65, 5 (May. 2008), 366-381.
- [BRA10] Brandwajn, A. and Begin, T. 2010. A Recurrent Solution of $Ph/M/c/N$ -like and $Ph/M/c$ -like Queues. INRIA Research Report 7321.
- [BRA11] Brandwajn, A. and Begin, T. 2011. An Approximate Solution for $Ph/Ph/1$ and $Ph/Ph/1/N$ Queues. Under submission.
- [CHAU92] Chaudhry, M. L., Agarwal, M. and Templeton, J. G. 1992. Exact and approximate numerical solutions of steady-state distributions arising in the queue $GI/G/1$. Queueing Systems Theory Applications. 10, 1-2 (Jan. 1992), 105-152.
- [COH82] Cohen, J.W. 1982. The single server queue. North- Holland (second edition).
- [FEL98] Feldmann, A. and Whitt, W. 1998. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. Performance Evaluation. 31, 3-4, 245-279.
- [JAG88] Jagerman, D. 1988. Approximations for waiting time in $GI/G/1$ systems. Queueing Systems Theory Applications. 2, 4 (Feb. 1988), 351-361.
- [LAT99] Latouche, G., Ramaswami, V., 1999. Introduction to Matrix Analytic Methods in Stochastic Modeling, ASA, 1999.
- [RAO99] Rao, B. V. and Feldman, R. M. 1999. Numerical approximations for the steady-state waiting times in a $GI/G/1$ queue. Queueing Systems Theory Applications. 31, 1/2 (Jan. 1999), 25-42.
- [WHI89] Whitt, W. 1989. An Interpolation Approximation for the Mean Workload in a $GI/G/1$ Queue. Operations Research. 37, 6 (Nov. - Dec. 1989), 936-952.