Gestion d'un centre d'appels téléphonique

Université de Grenoble - Préparation à l'agrégation

Christophe Leuridan

Dans ce texte, nous étudions un modèle de centre d'appels téléphonique, en nous intéressant aux questions suivantes : combien y-a-t-il d'appels dans le système (en cours de traitement ou en attente) ? Quelle est la durée de l'attente pour qu'un appel soit traité ? combien faut-il prévoir de serveurs ?

Ce texte est inspiré d'un livre de Ger Koole en cours de rédaction, dont une version préliminaire est consultable à l'adresse http ://www.cs.vu.nl/ koole/ccmath/book.pdf .

1 Le modèle Erlang C

Le modèle Erlang C est un modèle mathématique simple et très utilisé. Dans ce modèle, on suppose que

- les appels arrivent au centre suivant un procesus de Poisson de taux λ ;
- les durées de chaque appel sont indépendantes, indépendantes du processus des arrivées et sont de loi exponentielle de paramètre μ ;
- le nombre de serveurs s est supérieur au niveau de charge $a = \lambda/\mu$;
- les appels sont traités par ordre d'arrivée dès qu'il y a un serveur de libre ;
- lorsque tous les serveurs sont occupés, les appels sont mis en attente ;
- la capacité de la file d'attente est illimitée.

Soit X_t le nombre d'appels en train d'être traités ou dans la file d'attente à l'instant t. Le processus $(X_t)_{t\geq 0}$ est donc une file d'attente $M/M/s/\infty$.

Autrement dit, $(X_t)_{t\geq 0}$ est une chaîne de Markov à temps continu, d'espace d'états $\mathbf{N}$, dont les transitions sont $n \to n+1$ au taux λ et $n \to n-1$ au taux $\mu_n = \min(n, s)\mu$. En notant $p_n(t) = P[X_t = n]$, on obtient donc

$$\frac{d}{dt}p_n(t) = \lambda p_{n-1}(t) + \mu_{n+1}p_{n+1}(t) - (\lambda + \mu_n)p_n(t)$$

avec la convention $p_{-1}(t) = 0$. La chaîne de Markov est irréductible et récurrente, nous allons donc étudier le régime stationnaire, c'est-à-dire le cas où la loi de X_t ne dépend pas de t.

Suggestions de développement :

1. Pourquoi suppose-t-on que $s > a$? Que se passerait-il si $s < a$? Justifier de façon heuristique ces affirmations.
2. Donner les probabilités de transition de la chaîne induite. Vérifier l'irréductibilité.
3. Justifier de façon heuristique la formule donnant $\frac{d}{dt}p_n(t)$, en expliquant d'où proviennent les différents termes.
4. Pour quelle(s) raison(s) s'intéresse-t-on au régime stationnaire ?

2 La formule d'Erlang

Supposons que la chaîne $(X_t)_{t\geq 0}$ est stationnaire : pour tout $n \in \mathbf{N}$, la probabilité $P[X_t = n] = p_n$ ne dépend pas de t et

$$\lambda p_{n-1} + \mu_{n+1} p_{n+1} = (\lambda + \mu_n) p_n,$$

avec la convention $p_{-1} = 0$.

Lorsque $n < s$, la relation de récurrence s'écrit $(n+1)p_{n+1} - ap_n = np_n - ap_{n-1}$. On en déduit que pour tout $n \in [1 \dots s]$, $p_n = \frac{a}{n} p_{n-1}$, d'où pour tout $n \in [0 \dots s]$,

$$p_n = \frac{a^n}{n!} p_0.$$

Lorsque $n \geq s$, la relation de récurrence s'écrit $s(p_{n+1} - p_n) = a(p_n - p_{n-1})$. On en déduit que $p_n - p_{n-1} = \big(\frac{a}{s}\big)^{n-s}(p_s - p_{s-1}) = \big(\frac{a}{s}\big)^{n-s}\big(\frac{a}{s} - 1\big)p_{s-1}$ d'où

$$p_n = \Big(\frac{a}{s}\Big)^{n-s} p_s = \frac{a^n}{s! s^{n-s}} p_0.$$

Il reste à remarquer que la valeur de p_0 est donnée par

$$\Big(\sum_{n=0}^{s-1} \frac{a^n}{n!} + \frac{a^s}{(s-1)!(s-a)}\Big) p_0 = 1.$$

Le nombre moyen de serveurs occupés à un instant donné est

$$\mathbf{E}[\min(X_0, s)] = \sum_{n=0}^{+\infty} \min{(n, s)} p_n = 0 + \sum_{n=1}^{+\infty} a p_{n-1} = a.$$

La productivité (c'est-à-dire la proportion moyenne de serveurs occupés) est donc $\frac{a}{s}$.

A l'instant 0, la probabilité pour que les serveurs soient tous occupés est

$$C(s,a) = P[X_0 \geq s] = \frac{a^s}{(s-1)!(s-a)} \Big(\sum_{n=0}^{s-1} \frac{a^n}{n!} + \frac{a^s}{(s-1)!(s-a)}\Big)^{-1}.$$

Sachant que les serveurs sont tous occupés à l'instant 0, un nouvel appel qui arrive doit attendre que $X_0 - s + 1$ appels soient traités pour qu'un serveur se libère. Le temps d'attente W_0 est donc une somme de variables aléatoires exponentielles de paramètre $s\mu$ dont le nombre suit une loi géométrique sur $\mathbf{N}^*$ de raison $\frac{a}{s}$. Sachant que $X_0 \geq s$, W_0 suit donc une loi exponentielle de paramètre $s\mu(1 - \frac{a}{s}) = s\mu - \lambda$.

On en déduit la formule d'Erlang donnant la loi de W_0 :

$$P[W_0 > t] = C(s,a) e^{-(s\mu - \lambda)t} \quad \text{pour tout } t \geq 0,$$

Le temps d'attente est moyen est donc $\mathbf{E}[W_0] = \frac{C(s,a)}{s\mu - \lambda} = \frac{C(s,a)}{s-a} \frac{1}{\mu}$, autrement dit

$$\text{temps moyen d'attente} \;=\; \frac{\text{probabilité de saturation}}{\text{surcapacité}} \times \;\text{temps de service moyen} .$$

Suggestions de développement :

1. Justifier les affirmations et le calcul donnant la loi de W_0
2. Expliquer par un argument heuristique pourquoi la productivité est $\frac{a}{s}$.
3. Réaliser un programme donnant la valeur de $C(s,a)$, $\mathbf{E}[W_0]$ (en secondes) et $P[W_0 > t]$ en fonction des paramètres fournis par l'utilisateur : λ (en minutes^{-1}), μ (en minutes^{-1}), s et t (en secondes).

3 Applications de la formule d'Erlang et résultats asymptotiques

Le centre d'appel doit traiter les appels qui arrivent au bout d'un délai raisonnable. Un critère fréquemment utilisé est la règle des « 80/20 » : 80% des appels doivent être traités avec un délai inférieur à 20 secondes.

Pour un taux d'arrivée λ et une durée moyenne de service $1/\mu$ donnés, la formule d'Erlang fournit la probabilité que le temps d'attente dépasse 20 secondes en fonction du nombre s de serveurs. On prend s égal au plus petit entier tel que $P[W_0 \leq 20\text{s}] \geq 0,8$.

Application numérique : on prend $\lambda = 1\,\text{min}^{-1}$, $\frac{1}{\mu} = 5\,\text{min}$, d'où $a = 5$.
Si $s = 7$, alors $C(s,a) = 0,3241$, $P[W_0 \leq 20\text{s}] = 0,7163$ et $\mathbf{E}[W_0] = 48,62$ s.
Si $s = 8$, alors $C(s,a) = 0,1673$, $P[W_0 \leq 20\text{s}] = 0,8631$ et $\mathbf{E}[W_0] = 16,73$ s.

Prenons maintenant $\lambda = 2\,\text{min}^{-1}$, $\frac{1}{\mu} = 5\,\text{min}$, d'où $a = 10$.
Si $s = 13$, alors $C(s,a) = 0,2853$, $P[W_0 \leq 20\text{s}] = 0,7419$ et $\mathbf{E}[W_0] = 57,05$ s.
Si $s = 14$, alors $C(s,a) = 0,1741$, $P[W_0 \leq 20\text{s}] = 0,8476$ et $\mathbf{E}[W_0] = 26,12$ s.

Ces exemple suggère qu'on peut réaliser des économies d'échelle en groupant des centres d'appel : bien que dans le deuxième exemple, le centre d'appel ait deux fois plus d'appels à traiter, le nombre de serveurs nécessaires passe seulement de 8 à 14. En revanche, ces économies rendent le centre d'appels plus sensible au risque de saturation en cas de sous-estimation du taux d'arrivée : des erreurs de l'ordre de 10% sont fréquentes.

Pour que le centre d'appel soit en mesure de traiter les appels dans un délai raisonnable, il faut bien sûr que le nombre de serveurs s soit supérieur à la charge a. La question est de savoir de quel ordre de grandeur doit être la surcapacité $s - a$.

Une réponse partielle est donnée par le résultat asymptotique suivant : si s et a tendent vers $+\infty$ et $\frac{s-a}{\sqrt{s}} \to c > 0$, alors

$$C(s,a) \to \frac{g(c)}{cG(c) + g(c)},$$

en notant g et G la densité et la fonction de répartition de la loi $\mathcal{N}(0,1)$:

$$g(c) = \frac{1}{\sqrt{2\pi}} e^{-c^2/2} \quad \text{et} \quad G(c) = \int_{-\infty}^{c} g(x)dx.$$

En particulier, si $\frac{s-a}{\sqrt{s}} \to 1$, alors $C(s,a) \to \frac{g(1)}{G(1)+g(1)} = 0,22$, ce qui signifie qu'environ 78% des appels sont traités immédiatement.

Suggestions de développement :

1. Expliquer par des arguments heuristiques pourquoi on réalise des économies d'échelle en regroupant des centres d'appel.
2. Que se passe-t-il si l'on multiplie λ par une constante et si on divise μ par la même constante ? Expliquer le résultat obtenu.
3. Démontrer la convergence annoncée. Indication : chercher des équivalents de $\sum_{n=0}^{s-1} \frac{a^n}{n!}$ et de $\frac{a^s}{(s-1)!(s-a)}$. On utilisera la convergence en loi de $\frac{N_a - a}{\sqrt{a}}$ vers la loi $\mathcal{N}(0,1)$ lorsque N_a suit la loi de Poisson de paramètre a et la formule de Stirling.
4. Simuler des trajectoires du processus $(X_t)_{t\geq 0}$ pour différentes valeurs des paramètres pour illustrer certains des résultats.

4 Discussion du modèle et estimation des paramètres.

Le modèle proposé ci-dessus est critiquable pour deux raisons : en réalité la capacité de la file d'attente est finie et certains clients abandonnent lorsque l'attente est trop longue. Par ailleurs, l'estimation des paramètres est essentielle pour prévoir le nombre de serveurs nécessaires.

Suggestions de développement :

1. Comment doit-on modifier la chaîne $(X_t)_{t\geq 0}$ lorsque la capacité (nombre maximum d'appels traités ou en attente) est $S \geq s$) ? Que devient la mesure invariante ? Quelle est la probabilité pour qu'un appel soit rejeté ?
2. On suppose que chaque client en attente a une durée de patience de loi exponentielle de paramètre ν, indépendante des autres clients, des arrivées des appels et des durées de service : autrement dit, si un appel en attente n'est pas traité avant la durée de patience du client, le client abandonne définitivement. Que deviennent les probabilités de transition dans ce cas ? Que se passe-t-il si $s < a$?
3. Comment estimer les paramètres λ et μ en observant la chaîne ?
4. On considère n clients dont les durées de patience $Z_1, \ldots, Z_n$ sont indépendantes et suivent une loi exponentielle de paramètre ν inconnu. On soumet le client numéro k à une attente W_k de loi exponentielle de paramètre α connu. On observe les variables $Y_k = \min(Z_k, W_k)$ et $I_k = \mathbf{I}_{[Z_k < W_k]}$: au bout de la durée Y_k (attente effective du client numéro k), le k-ième client abandonne si $Z_k < W_k$ et est servi dans le cas contraire. On suppose que les variables aléatoires $Z_1, \ldots, Z_n, W_1, \ldots, W_n$ sont indépendantes. Calculer $P[I_k = i; Y_k \in dy]$ pour $i \in \{0, 1\}$ et $y \geq 0$. En déduire la densité de $(I_1, \ldots, I_n, Y_1, \ldots, Y_n)$ par rapport à la mesure produit de la mesure de comptage sur $\{0,1\}^n$ par la mesure de Lebesgue sur $\mathbf{R}_+^n$. En déduire que l'estimateur du maximum de vraisemblance de ν est donné par

$$\frac{1}{\hat{\nu}} = \frac{Y_1 + \ldots + Y_n}{I_1 + \ldots + I_n}.$$

Que représentent le numérateur et le dénominateur ?