

Degré de service et files d'attente

Qui ne s'est pas déjà trouvé bloqué dans une file d'attente ? On pense immédiatement à celles rencontrées derrière les guichets de la poste ou d'une banque. Il y évidemment une grande quantité d'autres exemples, comme les voitures derrière un stop, des camions en attente de chargement, des documents sur un bureau ou des pièces stockées entre les étapes d'un processus de fabrication. Même s'il n'y a pas formation de file d'attente, le phénomène concerne également les clients d'un restaurant, l'accès à un serveur internet, des machines à réparer, un article au rayon d'un magasin ou les clients d'un service de prêt.

Description d'un système de service avec files d'attente

Un système de service est défini par :

- des *canaux* (« points de service » ou « guichets »)
- un flux de demandes ou *taux d'arrivées* d'unités ou de clients (défini comme nombre moyen d'arrivées par unité de temps)
- un temps de traitement ou *taux de service* (défini comme nombre moyen d'unités traitées par unité de temps).

Par simplification, il ne sera pas fait état ici du phénomène d'abandon de la file d'attente.

Des modèles mathématiques permettent alors de déterminer, par exemple :

- un temps d'attente moyen jusqu'au début du service
- la longueur moyenne de la file d'attente.

Le degré de service, quant à lui, est supposé correspondre à la tolérance au temps d'attente. Il est donc directement déterminé par :

- le nombre de guichets disponibles
- l'organisation au guichet
- la rapidité et la qualité du traitement

A noter que le système peut très bien remplir des fonctions multiples, soit, par exemple, des opérations de caisse, de change et de conseils financiers. Il faut cependant prendre la précaution de contrôler que la répartition observée correspond bien au modèle retenu.

Le modèle

Sans grand risque de se tromper, on peut admettre que la plupart des systèmes de service présentent les caractéristiques suivantes :

- le nombre d'unités qui entrent dans le système durant un intervalle de temps donné (taux d'arrivées) est aléatoire et distribué selon une fonction de Poisson
- le temps de traitement d'une demande (taux de service) est aléatoire et distribué selon une loi exponentielle.

Pour avoir une bonne perception du fonctionnement du système, il n'est alors pas nécessaire d'effectuer une simulation ; il suffit de calculer les caractéristiques de la courbe qui résulte de la composition des flux de clients et des temps de traitement. On obtient alors, par exemple :

- la probabilité qu'à un instant donné, un ou plusieurs client(s) attend(ent) d'être servi
- la probabilité qu'aucun client n'attende
- la probabilité qu'un client doive attendre moins de 2 minutes
- le nombre moyen de clients qui attendent
- le temps effectif d'utilisation du système de service

Un exemple

Une société compte trois employés pour servir ses clients. On constate que 20 clients en moyenne par jour (8 heures) sollicitent ces employés selon une loi de Poisson et que le temps moyen qu'un employé passe avec un client est de 40 min ; le temps de traitement est réparti selon une loi exponentielle. On peut alors calculer le temps qu'un employé passe avec les clients (ici 22,2 heures par semaine) et le temps moyen qu'un client doit consacrer jusqu'à qu'il ait pu être servi (en l'occurrence, 49 minutes).

Implications pratiques

La modélisation d'un système de service permet de traiter toute sorte de problèmes de manière objective et de préparer des mesures adaptées. En particulier, on peut redéfinir :

- l'effectif ou le nombre de guichets
- les heures d'ouverture ou d'activité
- les équipements.

Dans une perspective d'investissement ou dans le cadre d'une réflexion stratégique, il est souvent intéressant de revoir le temps consacré au service du client en fonction d'hypothèses de rationalisation ou de projections sur la fréquentation du système. Lors de l'interprétation des résultats, il faut alors avoir une bonne idée du taux de service que l'on veut atteindre.

Le suivi des temps de traitement et des flux devrait permettre de réagir assez tôt pour éviter des engorgements ou, à l'inverse, une mobilisation inadéquate des ressources. Il faut alors procéder par sondages et relever le nombre d'unités dans le système à des moments déterminés de manière aléatoire ; le nombre moyen de personnes dans le système ainsi que le nombre total d'unités traitées durant la période d'observation fournissent alors toute l'information nécessaire au calcul du taux d'arrivées et du taux de service. De là, il est facile de remonter à la source du problème que peut poser la gestion de files d'attente et d'évaluer les moyens de y remédier. Il est notamment possible de comparer la performance des canaux et d'examiner les effets d'une spécialisation sur le système de files d'attente.

Claude Meylan
10.6.1999