MODELES DE FILES D'ATTENTE

I. INTRODUCTION

1. Définition

Un système de files d'attente est caractérisé par le mécanisme d'entrée des clients (**les arrivées**) dans le système et le mécanisme de **service**.

On associe à la file d'attente :

1) **La loi des arrivées :** Soit $\mathbf{t_n}$ l'instant d'arrivée du n-ième client C_n, et soit $t_0(=0) < t_1 < \ldots < t_n < \ldots$ les instants d'arrivée des clients. On observe en général les variables $\mathbf{E_n=(t_n-t_{n-1})}$, n=1,2 … qui sont supposées indépendantes et de même loi (i.i.d) :

$$\mathbf{A(x)= P(E_n<x)}.$$

Si $A(x)= P(E_n< x)= \mathbf{1-e^{-\lambda x}}$, la suite $\mathbf{\{t_n\}}$ forme un **processus de Poisson de paramètre λ.**

2) **La loi de service :** Soit S_n la durée de service du client C_n, de loi

$$\mathbf{B(x)= P(S_n<x)} ;$$

Ces variables sont i.i.d.

3) **Le nombre de serveurs** : m (entier positif) serveurs ayant des caractéristiques statistiques identiques

4) **La capacité** : maximale de la file d'attente K (finie ou infinie)

5) **La source de clients L** (finie ou infinie)

6) **La discipline de service** : FIFO, LIFO, RANDOM, …

Exemples :

Type de système	Clients	Ressource ou Serveur	Activité ou service
Informatique	processus	processeur	temps de traitement
	demande d'E/S	disque dur	lecture ou écriture
Réseau de	message	réseau	temps de transmission
communication	paquet	canal de transmission	temps de transmission

Un réseau de files d'attente est un ensemble de systèmes (dits stations ou nœuds) interconnectés de manière quelconque; Chaque client nécessite un service dans une ou plusieurs stations.

2. Notations de Kendall :

C'est une nomenclature de classification des modèles de file d'attente de la forme :

A/B/m/K/L

où :

A : code de la loi des arrivées

B : code de la loi de service

m : le nombre de serveurs

K : la capacité de la zone d'attente

L : la capacité de la source de clients (nb max de clients pouvant accéder dans le système)

Code	Type de loi de probabilité
M	Exponentielle : $F(x)= 1- e^{-\lambda x}$, $x>0$
D	Déterministe : $F(x)= 1$ si $x>c$, et $F(x)= 0$ sinon
G	Loi générale

Exemple :

M/M/2 : arrivées de Poisson, service exponentiel, m=2 serveurs, capacité de la file infinie, source infinie

M/G/2 : arrivées de Poisson, loi de service générale ou arbitraire

3. Types de problèmes :

On s'intéresse à :

1) Etude de stabilité et d'existence d'un régime stationnaire
2) Evaluation des performances du système :
    a. Taille moyenne de la file d'attente **E(F)** (nb moyen de clients dabs la file)
    b. Nombre moyen de clients dans le système **E(N)** (en attente ou en cours de service)
    c. Temps moyen d'attente d'un client **E(W)**
    d. Temps moyen de séjour dans le système **E(V)** (temps d'attente +temps de service)
    e. Prob de refus ou de perte ou d'encombrement **Pr**
    f. Taux d'arrivée effectif **λ'**
    g. Débit absolu **A** (nb moy de clients servis par unité de temps)
    h. Débit relatif **A'** (prob qu'un client qui arrive ds le syst soit servi)
    i. Nb moyen de serveurs actifs **E(SA)** ou oisifs **E(SO)**
    j. Traffic offert : % de temps pendant lequel un serveur est occupé. Il peut être estimé par :

$$\lambda = N(S1+S2+\ldots+Sn) / T$$

    Où : T : période d'observation
    N(t) : le nb d'appels (de clients) durant le temps t
    Si : durée de l'appel
3) Control du système:

Il s'agit de déterminer la configuration « optimale » du système ou d'élaborer des politiques de control optimales du système par rapport à un critère économique donné tel que :

- Le taux de service
- Le taux d'arrivée
- Le nombre de serveurs
- La capacité de la file
- La discipline de service

File M/M/1

Le système est décrit par le processus des arrivées, poissonnien de taux λ, la loi exponentielle de service (taux μ), la file d'attente de capacité infinie (les clients sont supposés servis dans l'ordre de leur arrivée).

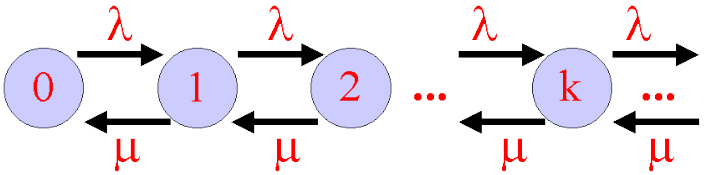
Soit n > 0 le nombre de clients dans le système (**un en service, n-1 en attente**) à un instant t.

La fin de service du client en cours se produit dans l'intervalle **(t, t+Δt)** avec une probabilité **μ Δt +o(Δt)**: propriété fondamentale d'une loi de service exponentielle.

Une arrivée se produira dans l'intervalle (t, t+Δt) avec une probabilité **λ Δt+o(Δt)**

L'état du système, représenté par le nombre total des clients dans le système (n) évolue selon un processus de Naissance et de Mort très simple (voir cours).

L'évolution de l'état de la file M/M/1 est décrit dans la figure suivante :



On en déduit les résultats suivants :

L'analyse du système, selon la procédure des équations de Chapman - Kolmogorov, aboutit aux équations suivantes:

$$\lambda \Pi_0 = \Pi_1 \mu$$
$$(\lambda + \mu) \Pi_k = \Pi_{k-1}\lambda + \Pi_{k+1} \mu$$

La sommation des équations de rang 0 à n conduit à la forme particulièrement simple:

$$\lambda \Pi_n = \mu \Pi_{n+1}$$

$$\Pi_{n+1} = \rho\Pi_n \qquad (\text{avec } \rho = \lambda / \mu)$$

Par récurrence, on obtient :

$$\mathbf{\Pi_{n+1} = \rho^{n+1} \Pi_0.}$$

La normalisation achève le traitement:

$$\sum \Pi_n = \Pi_0 \quad \sum_k \rho^k = \frac{\Pi_0}{1-\rho} = 1$$

Finalement:

$$\Pi_n = \rho^n (1-\rho) \qquad n \geq 0 \qquad \rho = \frac{\lambda}{\mu}$$

$$\Pi_W = \rho$$

$$E(n) = \rho/(1-\rho)$$

$$E(n_W) = \rho^2/(1-\rho)$$

Π_W : probabilité qu'un client ait à attendre
E(n) : nombre moyen de clients dans le système
$E(n_W)$: nombre moyen de clients en attente

Temps moyen d'attente : $[(\rho/\mu)/(1-\rho)]$

Temps moyen de séjour : $[(1/\mu)/(1-\rho)]$

Important : ρ doit être strictement inférieur à 1(condition de stabilité)
On montre que la condition $\rho < 1$ est une condition nécessaire et suffisante d'existence des probabilités stationnaires (condition de stabilité).

Modèles à capacité limitée(M/M/1/K)

Le modèle M/M/1/K correspond au cas d'une capacité de K clients (K représente le nombre total de clients dans le système, c'est à dire en attente ou en service). Le diagramme d'évolution est obtenu par la troncature du diagramme M/M/1, et les équations donnent :

$$\Pi(n) = \frac{\rho^n (1-\rho)}{1-\rho^{K+1}}$$

La **probabilité de rejet** est :

$$\Pi = \frac{\rho^K (1-\rho)}{1-\rho^{K+1}}$$

Evidemment, ici on peut lever la condition $\rho < 1$. On remarque que $\rho = 1$ conduit à une difficulté sur la formule. C'est qu'en réalité elle est ``arrangée'', et il faudrait lire par exemple

$$\Pi = \frac{\rho^K}{1 + \rho + \rho^2 + \rho^3 + \ldots + \rho^K}$$

File M/M/c

Mêmes hypothèses et notations que ci-dessus, avec un nombre c de serveurs identiques en pool.
Que deviennent les taux de service ?

Supposons que le système est dans l'état $k < c$: k clients sont présents, qui occupent k serveurs (les autres sont inactifs).

Dans l'intervalle Δt qui vient, chacun des services en cours peut se terminer, avec la probabilité $\mu \Delta t + o(\Delta t)$ (services exponentiels). La probabilité d'une et une seule fin de service est la somme des probabilités que chacun des services s'achève, diminuée de la probabilité de 2 ou plus fins de services - *événement de probabilité en* $o(\Delta t)^2$, donc négligeable dans l'opération de passage à la limite.

. Le taux de service vaut donc **k μ tant que k < c** et **c μ au-delà** (parce que c est la limite au nombre des serveurs actifs) .

. Le trafic offert est noté A et le **trafic par serveur est ρ = A/c**. La condition de stabilité reste **ρ < 1** , soit **A < c** .
On trouve:

$$\Pi(k) = \Pi(0) \frac{A^k}{k!} \qquad k \leq c$$

$$\Pi(k) = \Pi(0) \frac{A^c}{c!} (\rho)^{k-c} \qquad k > c$$

avec

$$\Pi(0) = \left(\sum_0^{c-1} \frac{A^k}{k!} + \frac{A^c}{c!} \frac{1}{1-\rho} \right)^{-1}$$

La probabilité d'attendre est (formule d'Erlang)

$$\Pi_W = \frac{A^c}{c!} \frac{1}{1-\rho} \Pi(0)$$

On en déduit le temps moyen d'attente :

$$E(W) = \frac{\Pi_W}{c(1-\rho)} E(s)$$