

# Fiche de File d'attente

V.R

April 18, 2015

## Contents

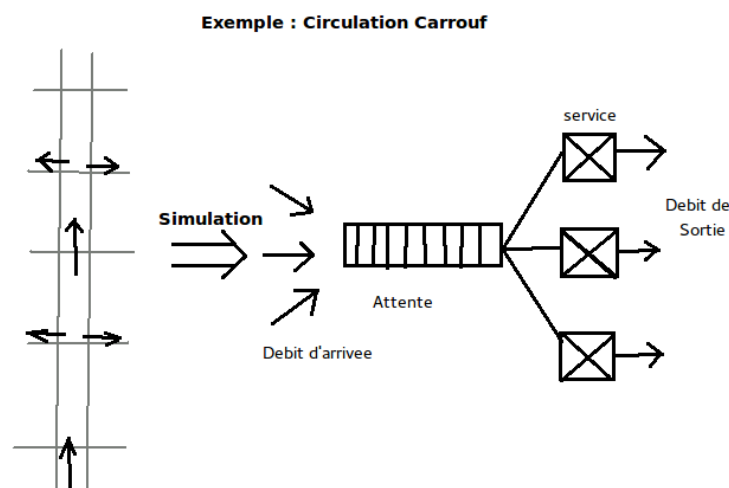
<b>I</b>	<b>Introduction</b>	<b>2</b>
1	Motivations	2
2	Critères de performance	2
<b>II</b>	<b>Analyse opérationnelle</b>	<b>4</b>
1	Formule de Little opérationnelle	4
2	Système à un serveur	5
3	Modèle de réseau fermé	5
4	Relations d'équilibre d'un système	6
5	Notations importantes	7
<b>III</b>	<b>Temps d'attente résiduel &amp; Lois de probabilité</b>	<b>8</b>
1	Temps d'attente résiduel	8
2	Lois de probabilité	8
<b>IV</b>	<b>Chaînes de Markov</b>	<b>10</b>
1	Chaînes de Markov à temps discret (CMTD)	10
2	Chaînes de Markov à temps continu (CMTC)	12
<b>V</b>	<b>Modèle à file simple</b>	<b>14</b>
1	Description d'une file d'attente	14
2	Généralités	15
<b>VI</b>	<b>Réseaux de files d'attente</b>	<b>17</b>
1	Théoreme de Jackson	17
2	Théorème de Gordon et Newell (Jackson fermé)	18

3	Théorème de Baskett, Chandy, Muntz et Palacios	19
4	Algorithme de Reiser pour les réseaux fermés possédant une solution en forme produit	19
VII Méthodes d'agrégation		20
1	Principe général	20
2	Méthode exacte pour les réseaux BCMP	20

## Part I Introduction

### 1 Motivations

**Évaluation des performances** On veut, à partir du modèle réel, créer un modèle virtuel afin d'en étudier les performances et ainsi améliorer les performances de la réalité. Pour ce faire on crée des modèles de file d'attente.



### 2 Critères de performance

**Le temps de réponse** C'est l'espérance mathématique du temps séparant l'arrivée d'une requête de la fin de son traitement.

$$T_{\text{reponse}} = T_{\text{attente}} + T_{\text{service}}$$

**Le débit** C'est l'espérance mathématique du nombre de requêtes traitées par unité de temps.

**Exemple**, débit internet :

-requête = 1Mo

-unité de temps = 1s

-soit débit = nbMo/s

**Le taux d'occupation** Probabilité pour qu'une ressource soit occupée, c'est à dire, La probabilité qu'une ressource soit en train de traiter une requête à un instant t donné.



**Le taux de perte des paquets** Probabilité pour qu'un paquet soit perdu. La perte d'un paquet peut se produire, par exemple, lorsque la file d'attente est pleine, le paquet est alors rejeté.

$$\mathbf{TauxDePerte} = \mathit{mesure} \left[ \frac{\mathbf{nbPaquetPerdu}}{\mathbf{nbPaquetEnvoyé}} \right]$$

$$\mathbf{ProbaPerte} = E[1/\mathit{perte}]$$

**Remarque importante** On s'intéressera, mathématiquement, surtout à des mesures générales, comme des probabilités ou des espérances qui permettent de définir des performances dans le cas général. (voir petit exemple page 11/94)

## Part II

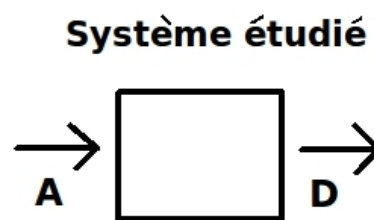
# Analyse opérationnelle

On s'intéresse ici à des mesures élémentaires que l'on effectue sur un système informatique. A partir de ces mesures on calcul un ensemble de critères de performance qui permettent de quantifier les performances du système.

## 1 Formule de Little opérationnelle

**Définition du système** Le système, ici, est un mécanisme recevant des requêtes et les restituant à l'issue d'un certain temps de traitement. (le système peut contenir des requêtes non traitées) On connaît le système grâce à deux compteurs :

- le nombre total d'entrées
- le nombre total de sorties



### Mesures élémentaires considérées

1.  $T$  : durée de la mesure
2.  $A$  : nombre total d'arrivées de requêtes
3.  $D$  : nombre total de départs de requêtes
4.  $T(n)$  : durée cumulée pendant laquelle le système a contenu  $n$  requêtes lors de cette mesure
5.  $n_{max}$  : nombre maximum de requêtes dans le système

### Critères de performances

1.  $\bar{\Lambda}$  : débit du système (à la sortie) :  $\bar{\Lambda} = \frac{D}{T}$

2.  $\bar{L}$  : nombre moyen de requêtes dans le système :  $\bar{L} = \frac{\sum_1^{n_{max}} n.T(n)}{T}$

3.  $\bar{R}$  : temps de réponse moyen du système :  $\bar{R} = \frac{\sum_1^{n_{max}} n.T(n)}{D}$

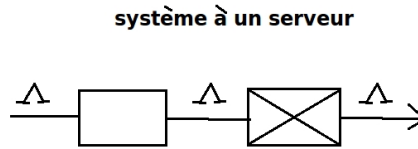
### Formule de Little

$$\boxed{\bar{L} = \bar{\Lambda} \cdot \bar{R}}$$

(Cf exemple du modèle du dentiste)

## 2 Système à un serveur

**Définition du système**  $U = P[\text{serveur occupé}] = \text{taux d'utilisation du serveur (si un seul serveur)}$



On applique la **Loi de Little** au service :

-S : Le temps de service

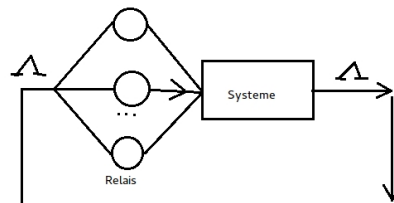
-nb de clients, 0 si serveur inoccupé  $(1 - U)$

-nb moyen de clients  $= 0 \times (1 - U) + 1 \times U = U$

soit :  $\bar{U} = \bar{\Lambda} \bar{S}$  et  $U = \bar{\Lambda} E[S]$

## 3 Modèle de réseau fermé

**Définition du système** on a N le nombre de client (constant)



Le processus passe alternativement par deux phases :

-réflexion : l'utilisateur crée sa requête

-traitement : la requête est traitée.

### Mesures élémentaires considérées

1.  $T$  : durée de la mesure
2.  $N$  : nombre de terminaux connectés
3.  $A$  : nombre total d'arrivées de requêtes
4.  $D$  : nombre total de départs de requêtes
5.  $r(k)$  : durée cumulée passée en traitement par le processus  $k$
6.  $z(k)$  : durée cumulée en réflexion par le processus  $k$

**Critères de performances**

1.  $\bar{\Lambda}$  : débit du système (à la sortie) :  $\bar{\Lambda} = \frac{D}{T}$

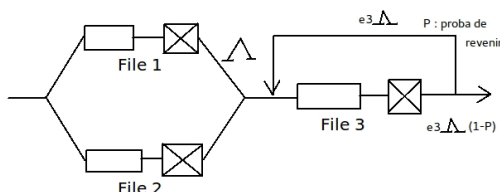
2.  $\bar{Z}$  : temps de réflexion moyen sur les terminaux :  $\bar{Z} = \frac{\sum_{k=1}^N z(k)}{A}$

3.  $\bar{R}$  : temps de réponse moyen du système :  $\bar{R} = \frac{\sum_{k=1}^N r(k)}{D}$

**Temps de réflexion dans un système interactif**  $\bar{R} = \frac{N}{\Lambda} - \bar{Z}$

**4 Relations d'équilibre d'un système**

**Définition du système** On a le système suivant :



**Mesures élémentaires considérées**

1.  $T$  : durée de la mesure
2.  $D$  : nombre total de requête globales sorties du système
3.  $D_i$  : nombre total de requêtes élémentaires traitées par la station  $i$
4.  $T_i(n)$  : durée cumulée pendant laquelle la station  $i$  a contenu  $n$  requêtes élémentaires

**Critères de performances**

1.  $\bar{\Lambda}_i$  : débit de la station  $i$  (au niveau des requêtes élémentaires) :  $\bar{\Lambda}_i = \frac{D_i}{T}$

2.  $\bar{U}_i$  : taux d'occupation de la station  $i$  (proportion de temps pendant lequel la station n'est pas vide) :  $\bar{U}_i = \frac{T - T_i(0)}{T}$

3.  $\bar{S}_i$  : durée moyenne de service à la station  $i$  :  $\bar{S}_i = \frac{T - T_i(0)}{D_i}$

4.  $\bar{e}_i$  : nombre moyen de visite à la station par travail :  $\bar{e}_i = \frac{D_i}{D}$

5.  $\bar{R}_i$  : temps de réponse de la station  $i$  :  $\bar{R}_i = \sum n \cdot \frac{T_i(n)}{D_i}$

6.  $\bar{L}_i$  : nombre moyen de requêtes élémentaires dans la station  $i$  :  $\bar{L}_i = \sum n \cdot \frac{T_i(n)}{T}$

7.  $\bar{\Lambda}$  : débit global du système (au niveau des travaux) :  $\bar{\Lambda} = \frac{D}{T}$



**Identité** On en déduit :

$$\bar{\Lambda} = \frac{\bar{\Lambda}_i}{\bar{e}_i} = \frac{\bar{U}_i}{\bar{S}_i \cdot \bar{e}_i} = \frac{\bar{L}_i}{\bar{R}_i \cdot \bar{e}_i}$$

et :  $e_3 = \frac{1}{1-p}$

## 5 Notations importantes

**Notations :**

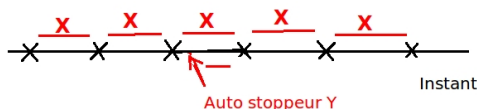
1.  $S$  : Temps de service (temps passé par un client dans le serveur)
2.  $U$  : Taux d'occupation du serveur (probabilité pour que ce dernier soit occupé)
3.  $\Lambda$  : Débit du système à la sortie
4.  $L$  : Nombre de clients dans le système (serveur et file d'attente)
5.  $R$  : Temps de séjour dans le système (somme du temps d'attente et du temps de service)

Part III

# Temps d'attente résiduel & Lois de probabilité

## 1 Temps d'attente résiduel

**Définition du système** On a une série de voiture et un auto stoppeur. On veut connaître à un instant  $t$  la durée avant la prochaine arrivée.  $Y$  la durée résiduelle d'attente pour le piéton, durée séparant son arrivée de l'arrivée du premier bus/voiture.



on a les cas :

$-E[Y] = E[X]$

$-E[Y] < E[X]$  (pas toujours vrai)

**Formule de Pollaczek Khinchine**  $E[Y] = E[X] \left( \frac{1+C^2[X]}{2} \right)$  avec  $C^2[X] = \frac{Var[X]}{E[X]^2}$

On peut avoir  $E[Y] > E[X]$  car les intervalles  $X$  ne sont pas forcément réguliers.

**Voiture à intervalle regulier**  $X = cste$  d'où,  $E[X] = X$  et  $E[Y] = X/2$

**repartition exponentielle**  $C^2[X] = 1$  et  $\sigma^2[X] = (E[X])^2$

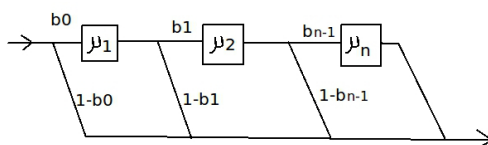
$X$  est sans mémoire. On arrive à un instant quelconque et le temps au bout duquel un événement se produit a la même distribution que le temps séparant 2 événements successifs. Pas besoin de savoir ce qui s'est passé avant.

**Arrivées uniformément réparties**  $C^2[X] > 1$  et  $E[Y] > E[X]$

Si les arrivées sont uniformément réparties, on a 2 intervalles (un grand et un petit). On a une probabilité plus forte d'arriver dans le grand intervalle plutôt que dans le petit.

## 2 Lois de probabilité

**Loi de cox** Si on considère dans une file d'attente que le service suit une loi de Cox, on peut décomposer le service en réseau de services de la façon suivante :



Mais il n'y qu'un seul client à la fois dans l'ensemble du réseau. On a le temps d'attente  $r$  avec une densité de probabilité  $f(t)$  dont la transformée de Laplace :

$$f(z) = \int_0^\infty e^{-z.t} f(t)dt \text{ vérifie } f(z) = (1 - b_0) + \sum_{i=1}^n (1 - b_i) \cdot a_i \prod_{j=1}^i \left( \frac{\mu_j}{z + \mu_j} \right)$$



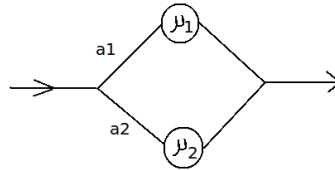


avec :

$$a_1 = b_0.b_1.b_2...b_{i-1} \quad E[\tau] = \sum_{i=1}^n \frac{a_i}{\mu_i} \quad Var[\tau] = \sum_{i=1}^n \frac{a_i}{\mu_i^2}$$

**Loi exponentielle** On a les variables comme pour une loi de probabilité exponentielle (Cf pougne de Proba)

**Loi hyper-exponentielle** Cette fois on a :

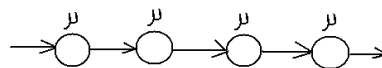


On prend les deux lois exponentielles en parallèle :  $x_1 = \frac{1}{\mu_1}$  et  $x_2 = \frac{1}{\mu_2}$

la densité est :  $f(x) = a_1.\mu_1.e^{-\mu_1 x} + a_2.\mu_2.e^{-\mu_2 x}$

Carré du coefficient de variation :  $C^2 = 1 + 2.a_1.\frac{a_2.(x_1-x_2)^2}{(a_1.x_1+a_2.x_2)^2} \leq 1$

**Loi d'Erlang** Cette loi correspondà la situation où on a  $r$  lois exponentielles en série



-La moyenne :  $E[X] = \frac{r}{\mu}$

-Carré du coefficient de Variation :  $C^2 = \frac{1}{r}$

**Loi de Poisson** (cf pougne de proba)

## Part IV

# Chaînes de Markov

**Un processus stochastique** est une famille de Variables Aléatoires  $\{\mathcal{X}_t, t \in T\}$  définie sur un même espace de probabilité et indexée par  $t \in T$  ( $t$  est le paramètre "temps").

## 1 Chaînes de Markov à temps discret (CMTD)

### Définition

-  $E$  : espace d'état. Les états sont numérotés  $\{1, 2, \dots, k, \dots\}$

-  $\{\mathcal{X}_n, n \in N\}$  une chaîne à valeur dans  $E$ .

C'est une chaîne de markov si la probabilité que l'état soit  $j$  en  $t = n + 1$  est connue lorsque l'état  $i_n$  en  $t = n$  est connu.

$$P(X_{n+1} = j / X_1 = i_1 \cap X_2 = i_2 \cap \dots \cap X_n = i_n) = P(X_{n+1} = j / X_n = i_n)$$

### Notations

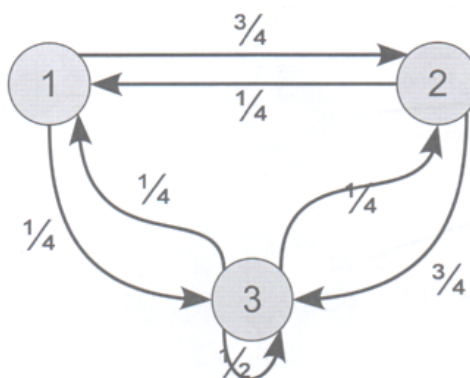
$P_{ij} = P(X_{n+1} = j / X_n = i)$  : Probabilité de transition

$P_{ij}^{(n)} = P(X_{r+n} = j / X_r = i)$  : probabilité d'aller de  $i$  à  $j$  en  $n$  étapes

$$\text{on a : } P_{ij}^{(n)} = \sum_{k \in E} P_{ik}^{(n-1)} \cdot P_{kj}$$

### Représentation des probabilités de transition

1. les noeuds correspondent aux états
2. les arcs correspondent aux transitions entre états



On a la matrice stochastique :  $P = [P_{ij}] \cdot P^{(m)} = [P_{ij}^{(m)}]$ .

$$P = \begin{bmatrix} 0 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

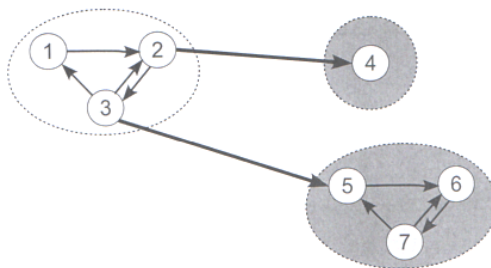
**Distributions** Notons  $\Pi_j^{(n)} = P(X_n = j)$  et  $\Pi^{(n)} = (\Pi_1^{(n)}, \dots, \Pi_j^{(n)}, \dots)$  la distribution à l'instant  $n$ . Il s'agit de la loi de  $X_n$ .

On a :  $\Pi^{(n+1)} = \Pi^{(n)} \cdot P$

En notant  $\Pi^{(0)}$  la distribution initiale, on a :  $\Pi^{(n)} = \Pi^{(0)} \cdot P^n$

**Chaînes de Markov irréductible** C'est une chaîne dont tous ses états peuvent être atteints, depuis tout autre état, au bout d'un nombre fini de pas, avec une probabilité non nulle. Le graphe doit être fortement connexe.

Ex : Ici cette chaîne est réductible. Depuis l'état 4 on ne peut rejoindre aucun autre état. Et un nombre limité pour les états 5, 6 et 7.



**Classification des états** Soit :

$-f_j^{(n)}$  la probabilité pour que le premier retour en  $j$ , après un départ de  $j$ , se produise au bout de  $n$  pas.

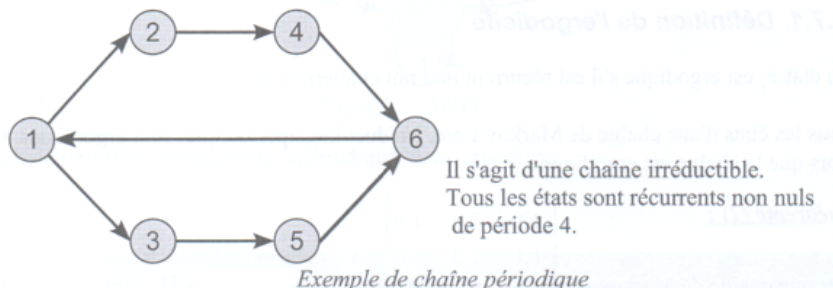
$f_j = \sum_{n=1}^{\infty} f_j^{(n)}$  la probabilité de repasser en  $j$ , étant parti de  $j$ .

**Définition :**

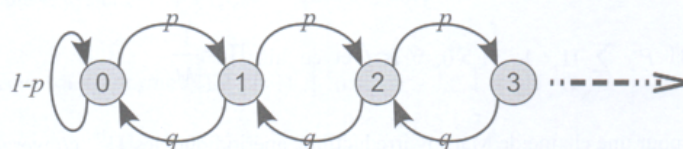
- Si  $f_j = 1$  on dit que  $j$  est récurrent : on reviendra toujours en  $j$
- Si  $f_j < 1$  on dit que  $j$  est transitoire

De plus, si  $j$  est récurrent, on définit le temps moyen de retour en  $j$  :  $M_j = \sum_{n=1}^{\infty} n \cdot f_j^{(n)}$ .

- Si  $M_j = \infty$ ,  $j$  est récurrent nul
- Si  $M_j < \infty$ ,  $j$  est récurrent non nul



Exemple de la chaîne de Markov suivante :



*Chaîne infinie*

- Si on a un nombre fini d'états, ils sont tous récurrents non nuls.
- Si on a un nombre infini d'états :
  1.  $p > q$  les états sont transitoires.
  2.  $p = q$  les états sont récurrents nuls.
  3.  $p < q$  les états sont récurrents non nuls.

**Théorème 1** Les états d'un chaîne de Markov irréductibles sont :

- soit tous transitoires, soit tous récurrents nuls, soit tous récurrents non nuls,
- soit tous apériodiques, soit tous périodiques de même période.

Les états d'une chaîne de Markov finie irréductible sont tous récurrents non nuls.

**Définition de l'ergodicité** Un état  $e_j$  est ergodique s'il est récurrent non nul et apériodique.

**Théorème 2** Pour une chaîne de Markov ergodique, il existe une distribution limite  $\Pi_i = \lim_{n \rightarrow \infty} \Pi_i^{(n)}$ . C'est une distribution stationnaire indépendante de la distribution initiale.

$$\text{Elle vérifie : } \Pi = \Pi.P, \sum \Pi_i = 1 \quad \Pi_i > 0, \forall i \in E \text{ et, en fait, } \Pi_i = \frac{1}{M_i}$$

Dans tous les cas, pour une chaîne de Markov irréductible, apériodique, les  $\Pi_j^{(n)}$  convergent. Si les états sont transitoires ou récurrents nuls,  $\Pi_j = 0$ . Dans tous les cas,  $\Pi_j$  est la limite de la proportion du temps passé dans l'état  $j$  pour une durée tendant vers l'infini.

**Convergence** Si la chaîne de Markov est **irréductible, récurrente positive et apériodique**, alors  $P^k$  converge vers une matrice dont chaque ligne est l'unique distribution stationnaire  $\Pi$ . On note, si elle existe :

$$\Pi^\infty = \lim_{n \rightarrow \infty} \Pi^{(n)}$$

(Cf exemple de Doudou le Hamster)

**Théorème de convergence**

1. Chaîne périodique  $\Rightarrow$  pas de convergence
2. Chaîne réductible  $\Rightarrow$  pas de convergence
3. chaîne apériodique et irréductible :
  - convergente :  $\Pi = \Pi P$  a une solution  $\Pi$  non nulle.
  - non convergente :  $\Pi = \Pi P$  n'a que la solution nulle

Conséquence : Si la chaîne est finie, apériodique et irréductible  $\Rightarrow$  convergente.

**Technique de coupes IMPORTANT** On peut faire les calculs de probabilité d'état en faisant des coupes dans l'ensemble des états. On considère alors qu'à l'état stationnaire le flux des transitions traversant la coupe est le même dans les deux sens. Ainsi :

$$\sum_{i \in \text{Région1}, j \in \text{Région2}} p_{ij} \Pi_i = \sum_{i \in \text{Région2}, j \in \text{Région1}} p_{ij} \Pi_i$$

## 2 Chaînes de Markov à temps continu (CMTC)

**Définition**  $(X_t, t \in \mathbb{R})$  est une chaîne de Markov à temps continu si:  
 $\forall (t_1, t_2, \dots, t_n) \in \mathbb{R}^n$  avec  $t_1 < t_2 < \dots < t_n$ , et  $\forall (i_1, i_2, \dots, i_n) \in \mathbb{N}^n$  on a:

$$P(X_{t_n} = j / X_t = i, \dots, X_{t_{n-1}} = i_{n-1}) = P(X_{t_n} = j / X_{t_{n-1}} = i_{n-1})$$

**Distribution** Soir les probabilités :  $p_{ij}(s, t) = P(X_t = j / X_s = i)$  et  $H_{s,t} = [p_{ij}(s, t)]$  la matrice de transition.

De plus, notons  $Q(t) = \lim_{\Delta t \rightarrow 0} \frac{H(t, t + \Delta t) - I}{\Delta t}$  le générateur infinitésimal de la matrice de transition.

En posant :  $\Pi_i(t) = P(X_t = i)$  et  $\Pi(t) = (\Pi_1(t) \dots \Pi_k(t) \dots)$

On obtient au final pour la distribution :

$$\Pi(t) = \Pi(0) \cdot H(0, t) \text{ soit } \Pi(t) = \Pi(0) e^{\int_0^t Q(u) du}$$

**Théorème 3**

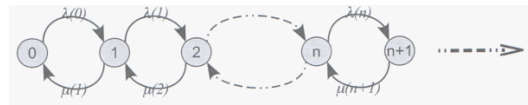
Pour une chaîne de Markov à temps continu, homogène, irréductible, il y a bien convergence de  $\Pi(t)$  quand  $t$  tend vers l'infini.

- états récurrents nuls ou transitoires  $\Pi_i(t) \rightarrow 0, \forall i$
- états récurrents non nuls, il existe une distribution stationnaire  $\Pi$  telle que

$$\lim_{t \rightarrow \infty} \Pi(t) = \Pi \text{ vérifiant } \Pi \cdot Q = 0 \text{ et } \sum \Pi_i = 1$$

- Dans tous les cas  $\Pi_i$  est la limite de la proportion du temps passé en  $i$ , pour une durée totale tendant vers l'infini.

**Processus de naissance et de mort** Processus de Markov où on ne peut passer de l'état  $n$  que dans l'état  $n + 1$  (naissance), ou  $n - 1$  (mort).



**Limite stationnaire ? :**

$$P(n) = \frac{\lambda(n-1)\lambda(n-2)\dots\lambda(0)}{\mu(n)\mu(n-1)\dots\mu(1)} P(0) \quad (1)$$

Soit série est convergente :

$$\sum_{n=0}^{\infty} P(n) = 1 \quad (2)$$

**Théorème 4** Pour un processus de naissance et de mort, s'il y a convergence, la probabilité d'état est solution du système (1),(2).



## Part V

# Modèle à file simple

## 1 Description d'une file d'attente

**Définition d'une file d'attente simple** Un modèle de file d'attente simple est caractérisé par des serveurs (en attente de travail ou en cours de traitement d'un client) et des clients (ne peut être traité que par un serveur à la fois ou ne traite qu'un client à la fois).

**Processus d'arrivées** La loi qui régit le temps entre les arrivées des clients. Ex :  $P_r[X \leq t] = 1 - e^{-\lambda t}$

**Loi de service** La loi de la durée du service

**Nombre de serveurs**

**Loi de priorité**

- FIFO (First In First Out) = shotgun
- LIFO (devinez) = dernier arrivé premier servi
- classes de priorité

**Capacité de la file**

**Réseau ouvert ou fermé** fermé = nb de clients constant / ouvert = nb de clients variable.

**Paramètres**

- Nombre de serveurs : m
- Capacité de la file : k clients (y compris ceux qui reçoivent du service)
- Distribution des services : Durée moyenne  $E[S]$ , carré du coef de variation  $C^2 = \frac{\sigma^2}{E^2[S]}$  ( $\sigma^2$  la variance)
- Distribution des inter-arrivées : durée moyenne  $[Z]$
- Nombre total de clients N

**Notation de Kendall** Notation qui caractérise une file d'attente : A/B/m/F/K/N

- A : Loi des arrivées
- B : loi des services
- m : nombre de serveurs
- F : discipline de la file (par défaut FCFS = FIFO)
- K : nombre de places dans la file (par défaut  $\infty$ )
- N : nombre total de clients éventuels (par défaut  $\infty$ )

Les lettres A et B peuvent être :

- M : Loi exponentielle (Markovien)
- G : loi quelconque (Générale)
- D : durée constante (déterministe)



- E : Loi d'Erlang
- H : Loi hyper-exponentielle

La discipline de service peut être :

- FCFS (= FIFO abus de langage, vrai ssi un seul serveur) First Come First Served
- LCFS (= LIFO abus de langage) Last Come First Served
- LCFS préemptif : dernier arrivé interrompt le service en cours, il sera repris ensuite quand le client aura la priorité
- QUANTUM le service dur au max  $q$  sinon interrompu et on passe au suivant en remettant le client interrompu à la fin
- PS (Processor Sharing) quand  $g$  tend vers 0. Pendant chaque unité de temps, chaque client reçoit un service de durée  $\frac{1}{k}$
- RANDOM clients choisis au hasard
- PRIORITÉ (avec ou sans préemption) système de classes

## 2 Généralités

**Supplément d'amphi**  $N(t)$  = nb de clients dans la file à l'instant  $t$ . Il suffit de connaître l'état de la file pour connaître la loi d'évolution dans le futur.  $N(t) = i$  events possibles entre  $t$  et  $t + dt$

$$P_r[S(t) < \theta] = 1 - e^{-\mu\theta}$$

$$P_r[\text{service résiduel} < \theta] = 1 - e^{-\mu\theta}$$

et

$$E[S] = \frac{1}{\mu}$$

$i \geq 3$  un des trois clients dont le service est en cours va sortir (car service terminé) on cherche  $\inf(3 \text{ services résiduels})$  sachant que : temps service résiduel = temps de service restant avant de sortir du serveur.

**Lemme** inf de 3 variables exponentielles de même taux  $\mu$  est une variable exponentielle de taux  $3\mu$  (sachant que : vitesse serveur = nb de clients servis / unité de temps pendant une période d'occupation)

$$\begin{aligned} P[\sum_1(t) \leq \theta] &= 1 - e^{-\mu\theta} \\ P[\sum_2(t) \leq \theta] &= 1 - e^{-\mu\theta} \\ P[\sum_3(t) \leq \theta] &= 1 - e^{-\mu\theta} \end{aligned} \Rightarrow P[\inf(\sum_1(t), \sum_2(t), \sum_3(t)) < \theta] = 1 - e^{-3\mu\theta}$$

$$P[\sum_i(t) > \theta] = e^{-\mu\theta} \rightarrow P[\inf(\sum_i(t)) > \theta] = e^{-3\mu\theta}$$

### PASTA 1

**Théorème d'échantillonnage** La distribution limite obtenue par échantillonnage avec la probabilité  $p$  parmi des arrivées quelconques de taux  $\lambda$ , lorsque :

- $\lambda \rightarrow \infty$
- $p \rightarrow 0$
- $\lambda p \rightarrow \text{Constante}$

est une loi de Poisson



**Théorème de superposition** La distribution limite obtenue par superposition de  $N$  arrivées indépendantes de même loi, de distribution quelconque, lorsque :

- $N \rightarrow \infty$
- $\lambda_i \rightarrow 0$
- $\sum_i^N \lambda_i \rightarrow \text{Constante}$

**Exemple de files simples** Voir les exemples du cours.

### Modèle M/M/1 IMPORTANT

Caractéristiques :

- M : Arrivées poissonniennes, taux  $\lambda$
- M : Service exponentiel, taux  $\mu$
- 1 : serveur

Résultats importants

- système stable si  $\lambda < \mu$ . En posant  $\rho = \frac{\lambda}{\mu} = \lambda \cdot E[S]$ . Et,  $P(n) = \rho^n \cdot (1 - \rho)$ ,  $P(0) = 1 - \rho$
- Probabilité qu'il y ait au moins  $N$  clients dans la file :  $P(n \geq N) = \sum_{n=N}^{\infty} P(n) = (1 - \rho) \cdot \sum_{n=N}^{\infty} \rho^n = \rho^N$
- Probabilité que le serveur soit occupé :  $U = 1 - P(0) = \rho$
- Nombre moyen de clients dans le système :  $E[L] = \sum_{n=1}^{\infty} n \cdot P(n) = \frac{\rho}{(1-\rho)} = \frac{\lambda}{\mu - \lambda}$
- Temps de réponse moyen :  $E[R] = \frac{1}{\mu - \lambda}$
- Variance du temps de réponse :  $\sigma_R^2 = (E[R])^2$
- Temps d'attente moyen :  $E[W] = E[R] - E[S] = \frac{\rho}{1-\rho} \cdot \frac{1}{\mu}$
- Variance du temps d'attente :  $\sigma_W^2 = \sigma_R^2 - \sigma_S^2$
- Temps attente moyen, attente non nulle :  $E[W] = E[R]$
- $\Lambda = \lambda = U \cdot \mu$



## Part VI

## Réseaux de files d'attente

## 1 Théoreme de Jackson

## Théoreme

Soit un reseau de F.A à capacité  $\infty$ , un serveur à temps de service exponentiel (M/M/1), FIFO, à routage probabiliste entre les files dont les arrivées de l'exterieur sont de poisson.

On note

- $\lambda$  : débit des arrivées de Poisson
- $\mu_1$  : le taux de service de la file
- $q_i$  : la probabilité qu'un client arrivant de l'esterieur aille sur la  $F.A_i$
- $p_{ij}$  : la probabilité qu'un client quittant la  $F.A_i$  aille dans la  $F.A_j$
- $e_i$  le nombre de passage d'un client pour la  $F.A_i$

Sous conditions de stabilité :

$$P(k_1, k_2, \dots, k_n) = \prod_{i=1}^n p_i^{k_i} (1 - p_i) \text{ où } p_i = \frac{\lambda_i}{\mu_i} = \frac{\lambda e_i}{\mu_i}$$

**Corollaire 1** Pour chaque file  $i$ ,  $P(k_i) = p_i^{k_i} (1 - p_i)$  :

- $L_i = \frac{p_i}{1-p_i}$
- $R_i = \frac{L_i}{\lambda_i} = \frac{L_i}{\lambda e_i} = \frac{p_i}{(1-p_i)\lambda e_i}$

## Corollaire 2

Le nombre moyen de clients dans tout le reseau est :  $L = \sum_i L_i = \sum_i \frac{p_i}{1-p_i}$

Le temps de reponse moyen du reseau entier :  $\sum_i e_i R_i = \sum_i e_i \frac{L_i}{\lambda_i} = \sum_i e_i \frac{L_i}{e_i \lambda} = \frac{\sum_i L_i}{\lambda}$

**Calcul des  $e_i$**  On raisonne sur les flux :  
sur le noeud  $i$  ( $FA_i$ ), le débit d'entrée.

$$\lambda q_i + \sum_{j=1}^n p_{ij} \lambda_j = \lambda q_i + \sum_{j=1}^n p_{ij} e_j \lambda$$

Or, le débit qui traverse la  $FA_i$  est  $\lambda e_i$

$$\lambda e_i = \lambda q_i + \sum_{j=1}^n p_{ij} e_j \lambda$$

$$e_i = q_i + \sum_{j=1}^n p_{ji} e_j$$

$$e = q + eP \text{ où } e(e_1, e_2, \dots, e_n), P = \{P_{ij}\}$$

Le théorème est en fait plus général dans le cas où  $\lambda$  varie en fonction du nombre de client dans le réseau et les taux de service varient en fonction des  $k_i$ :

$$P(k_1, \dots, k_n) = C \prod_{n=0}^{\sum_i k_i - 1} \lambda(n) \cdot \prod_{i=1}^n \prod_{j=1}^{k_i} \frac{e_i}{\mu_i(j)}$$

avec C tel que :

$$\sum_{k_1, \dots, k_n} P(k_1, \dots, k_n) = 1$$

## 2 Théorème de Gordon et Newell (Jackson fermé)

Soit un réseau de F.A à capacité infinie, à un seul serveur exponentiel, FIFO, à routage probabiliste entre les F.A et comportant k clients à l'intérieur. On reprend les mêmes notations pour  $\mu_i, e_i, p_{ij}$  sous conditions de stabilité :

$$P(k_1, \dots, k_n) = C \prod_{i=1}^n p_i k_i \text{ (Jackson : } \prod_{i=1}^n p_i k_i (1 - p_i))$$

$$P(k_1, \dots, k_n) = \prod_{i=1}^n \left(\frac{\lambda_i}{\mu_i}\right)^{k_i} = C \prod_{i=1}^n \left(\frac{\lambda e_i}{\mu_i}\right)^{k_i} = C \lambda^{\sum k_i} \prod_{i=1}^n \left(\frac{e_i}{\mu_i}\right)^{k_i} = \lambda^K \prod_{i=1}^n \left(\frac{e_i}{\mu_i}\right)^{k_i}$$

Dans le cas où les  $\mu_i$  sont fonctions du nombre de clients dans la  $FA_i$  :

$$P(k_1, \dots, k_n) = \prod_{i=1}^n \prod_{j=1}^{k_i} \frac{e_i}{\mu_i(j)} \text{ sans contrainte } \sum k_i = K$$

- Dans les 2 cas il n'existe pas d'expression générale de C.
- Attention, un même client passe un  $\infty$  de fois par la même file  $i$ . On définit alors les  $e_i$  relativement à un passage par un point de référence.

Le débit qui arrive sur la file  $i$  est alors :

$$\sum_{j=1}^n p_{ji} \lambda_j = \sum_{j=1}^n p_{ji} \lambda e_j \text{ avec } \lambda \text{ débit moyen au point de référence}$$

$$\sum_{j=1}^n p_{ji} \lambda_j = \lambda e_i \Leftrightarrow e_i = \sum_{j=1}^n p_{ji} e_j \Leftrightarrow e = eP$$

### 3 Théorème de Baskett, Chandy, Muntz et Palacios

**Introduction** Ce théorème généralise les deux théorèmes précédents et permet d'agrandir le champ d'application. Il s'applique à des réseaux :

- dont les clients appartiennent à un nombre  $\xi$  de classes différentes et tels que la probabilité de transition, d'un client de classe  $c$  sortant de la file  $i$  vers un client de classe  $c'$  et allant vers  $j$  est :  $p_{i,c';j,c}$
- Et dont chacune des stations appartient à l'une des 4 catégories suivantes :
  1. FCFS : service exponentiel de même coeff pour toutes les classes mais dont le coeff peut dépendre du nombre de clients.
  2. Processor Sharing : Service de Cox, peut être différent pour chaque classe, 1 serveur, priorité en Processor Sharing
  3. Pas d'attente : Service de Cox, nb de serveurs  $>$  nb max de clients.
  4. LCFS préemptif : 1 serveur, service de Cox, service préemptif, discipline LCFS.

(Catégories Cf p77 poly)

#### Résultats

$$e_{jc} = \sum_{c'=1}^{\xi} \left[ \sum_{i=1}^n (e_{i,c'} p_{i,c';j,c}) + q_{c';j,c} \right]$$

Où  $q_{c';j,c}$  désigne la probabilité pour qu'un client de classe  $c'$  arrivant dans le système devienne de classe  $c$  et aille vers la file  $j$ . (Cf poly .p77 pour les différentes catégories).

### 4 Algorithme de Reiser pour les réseaux fermés possédant une solution en forme produit

Aussi appelé Mean Value Analysis.

**Cas du théorème de Gordon et Newell avec taux de service constant** Lorsqu'un client arrive dans une file. Il voit la file des client qui suivent la loi du nombre de clients de la file sans le compter. Si la priorité est FCFS on a :

$$E[R_i(N)]E[S_i].[1 + E[L_i(N - 1)]]$$

$\lambda_{i_0}$  débit global,  $\lambda_i = \lambda_{i_0} \cdot e_i$ .  $e_i$  représente le nb moyen de passage d'un client dans la file  $i$  entre 2 passages par la file  $i_0$ . Avec la formule de Little appliquée au réseau :

$$E[L_i(N)] = \lambda_{i_0}(N) \cdot e_i \cdot E[R_i(N)]$$

En sommant par rapport à  $i$  les relations :

$$N = \lambda_{i_0}(N) \sum_{i=1}^N e_i \cdot E[R_i(N)]$$

L'algorithme de Reiser est itératif par rapport à  $N$ . Puis faire de  $N = 1$  à  $K$ :

- calculer  $E[R_i(N)]E[S_i].[1 + E[L_i(N - 1)]]$
- calculer débit global :  $\lambda_{i_0} = (N) = \frac{N}{\sum_{i=1}^M e_i \cdot E[R_i(N)]}$
- calculer  $E[L_i(N)] = \lambda_{i_0}(N) \cdot e_i \cdot E[R_i(N)]$

**Part VII****Méthodes d'agrégation****1 Principe général**

Pour étudier un système complexe on peut le séparer en sous-systèmes plus faciles à étudier.

**Étape 1 :** étudier les sous-systèmes indépendamment les uns des autres.

**Étape 2 :** étudier le système global en remplaçant les sous-systèmes par leur solution.

**2 Méthode exacte pour les réseaux BCMP**

Dans un réseau BCMP, pour un sous-système convenable, il est possible de remplacer le sous-système par une seule file d'attente de taux de service  $\mu(n)$ , si  $n$  est le nombre de clients dans la file,  $\mu(n)$  est égal au débit global du réseau fermé obtenu en fermant le sous-système sur lui-même avec  $n$  clients dans le réseau.

(Cf poly pour les schémas)