

Notes du cours de
Modélisation et simulation discrète

Cédric Dutoit
Laurent Burgbacher
EI5, EIVD

6 décembre 2001

Table des matières

1	Info	3
1.1	Informations sur le cours :	3
2	Buts	4
2.1	Etude et dimensionnement de systèmes informatiques à l'aide de réseaux de files d'attentes	4
2.1.1	Exemple : Serveur de base de données	4
3	Les files d'attente	7
3.1	Définition	7
3.2	Système de Kendall	7
3.2.1	Kendall/A, Kendall/S	8
3.2.2	Kendall/C (capacité)	8
3.2.3	Kendall/P(Taille population)	8
3.2.4	Kendall/Ds(discipline)	8
3.2.5	Exemple	9
3.3	Loi exponentielle de param $\lambda > 0$, rappel	10
3.3.1	Histogramme	10
3.3.2	Fonction de répartition	10
3.3.3	Espérance et variance d'une v.a. $X \sim \exp(\lambda)$	11
3.3.4	Absence de mémoire d'une v.a.	11
3.4	Formule de Little	12
3.5	Processus de comptage	13
3.6	Processus de Poisson	13
3.6.1	propriétés d'un processus de Poisson	14
3.7	Mélange de sources de clients	14
3.8	Séparation aléatoire des clients d'une source	14
3.9	Liens processus de Poisson / v.a. exponentielles	15
3.10	File M M 1	16
3.10.1	Rappel	16
3.10.2	Notation	16
3.10.3	Graphe	16
3.10.4	Equations	16
3.10.5	Résolution	17
3.10.6	Conclusion	18
3.10.7	Exemple	18
3.10.8	Mesures de performance d'une file M M 1	18
3.10.9	Exemple : Mesures de performance de notre serveur de base de données	19
3.11	Analyse de sensibilité	19
3.12	Mesures de performances M M 1	20
3.13	Exercices	21

3.13.1	Exercice 1	21
3.13.2	Exercice 2	21
3.13.3	Exercice 3	21
3.14	File M M K	23
3.14.1	Mesures de performance M M K	23
3.15	Exercices	25
3.15.1	Exercice 1	25
3.15.2	Exercice 2	25
3.15.3	Exercice 3	26
3.16	Autres files classiques	28
3.17	File M G 1	28
3.17.1	Mesures de performance	28
3.18	File G G 1	29
3.18.1	Mesures de performance	29

1 Info

1.1 Informations sur le cours :

- Professeur : Eric Thiémard; bureau H06a
- Cours :
 - 2-3 tests
 - exercices
 - pré-requis :
 - Probabilités et statistiques
 - Mathématiques
 - Algorithmie et structures de données

2 Buts

2.1 Etude et dimensionnement de systèmes informatiques à l'aide de réseaux de files d'attentes

2.1.1 Exemple : Serveur de base de données

Des requêtes sont soumises à un serveur de base de données suivant un taux moyen de $\lambda = 40$ requêtes par secondes. Lorsqu'une requête arrive :

```

Si le serveur est libre
    alors elle est traitée directement
Sinon
    elle attend son tour dans un buffer le temps que le serveur se libère
  
```

On mesure qu'en moyenne lorsque le serveur est libre, le temps moyen d'une requête est de $20ms$ (i.e. $\frac{1}{0.02} = 50 = \mu$ requêtes sont traitées par seconde en moyenne.)

1. Quelle est la taille moyenne du buffer ?
2. Quel est le temps moyen de réponse? (attente + service)
3. L'entreprise étant en expansion, une hausse de 50% des requêtes est à prévoir. Est-ce que le système pourra suivre? Sinon, il faut changer le serveur. Quel devra être son temps moyen de service μ pour que le temps moyen de réponse soit $\leq 100ms$? Réponse: théorie des files d'attente.

Mesure	Expérience 1	Estimateurs
Taux d'occupation du serveur	0.73	$\frac{t - \text{"t. système vide"}}{t} = \frac{t - \sum_{i=1}^n (t_i - t_{i-1})}{t}$
Taux d'arrivée du serveur	38.47	$\frac{\#\text{requêtes arrivées sur } [0,t]}{t}$
Durée moyenne de service	0.019	$\frac{\sum_{i=1}^k S_i}{k}$
Durée moyenne de réponse	0.055	$\frac{\sum_{i=1}^k (a_i + S_i)}{k}$
Nombre moy. requêtes présentes	2.12	$\frac{\sum_{i=1}^n (t_i - t_{i-1}) m_i}{t}$

TAB. 1 – Mesures de performances

Comment estime-t-on ces mesures de performances à partir d'une simulation sur un intervalle de temps $[0,t]$? On note t_1, t_2, \dots, t_n les temps auxquels des événements (arrivées, début/fin traitement) ont été observés et m le nombre de requêtes présentes entre t_{i-1} et t_i .

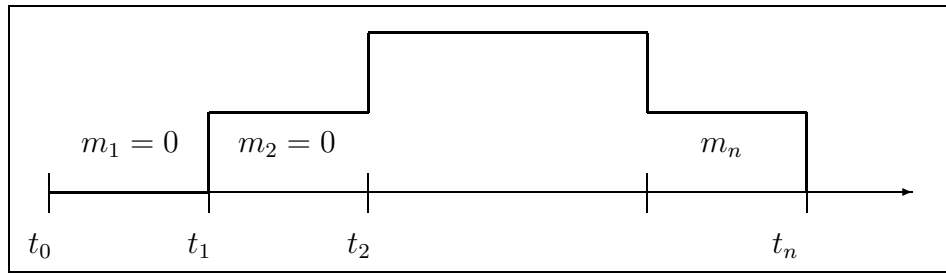


FIG. 1 – Requetes dans le système

Si k requêtes ont été traitées sur $[0, t]$, on note $a_1 \dots a_n$ leurs temps d'attente respectifs et $S_1 \dots S_n$ leurs temps de service respectifs.

Mesure	Expérience 2	Théorique
Taux d'occupation du serveur	0.87	0.8
Taux d'arrivée du serveur	39.9	40
Durée moyenne de service	0.021	0.02
Durée moyenne de réponse	0.227	0.1
#moy. requêtes présentes	9.04	4

TAB. 2 – Performances de l'expérience 2 et résultats théoriques

- Morale:** une telle simulation peut mener à des résultats "peu fiable" (à forte variance)
- Remèdes:**
- Simuler plus longtemps
 - méthodes de réduction de variance
 - ...
- Mais:** dans notre cas particulier, il est possible de *calculer* exactement nos mesures de performance

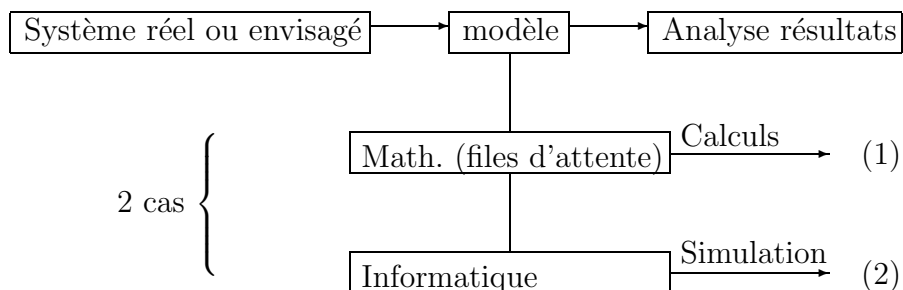


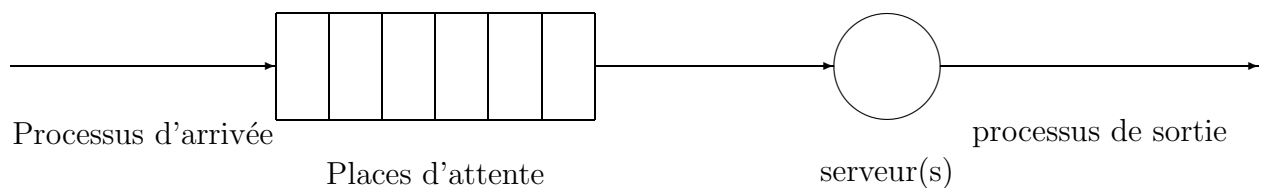
FIG. 2 – Démarche

1.
 - Analyse résultats
 - validation du modèle
 - Sensibilité
2.
 - Analyse résultats
 - validation du modèle
 - Sensibilité
 - Convergence

3 Les files d'attente

3.1 Définition

Etude des systèmes où des phénomènes d'attente apparaissent suite à des accès concurrentiels à des ressources limitées.



Une *file d'attente* est un modèle stochastique composé d'un certain nombre de *places d'attente*, d'un ou plusieurs *serveurs* et de *clients* (requêtes) qui *arrivent*, attendent, se font *servir* d'après certaines *règles de priorité* et quittent le système.

3.2 Système de Kendall

On utilisera le système de *Kendall* pour décrire une telle file :

$$A|S|K|C|P|D_s$$

A : Loi de distribution du temps écoulé entre 2 arrivées consécutives de client

S : Loi de distribution du temps de service d'un client

K : Nombre de serveurs

C : Capacité du système (attente + service)

P : Taille de la population

D_s : Discipline de service

On utilise également la notation abrégée

$$A|S|K$$

qui désigne

$$A|S|K|\infty|\infty|FIFO$$

3.2.1 Kendall/A, Kendall/S

Pour les processus d'arrivée A et de service S, on utilise généralement des réalisations de v.a.¹ i.i.d² parmi les lois suivantes :

M: loi exponentielle (processus de Poisson)

D: déterministe (constante)

Eq: loi d'Erlang d'ordre q (i.e. somme de q v.a. exponentielles i.i.d.)

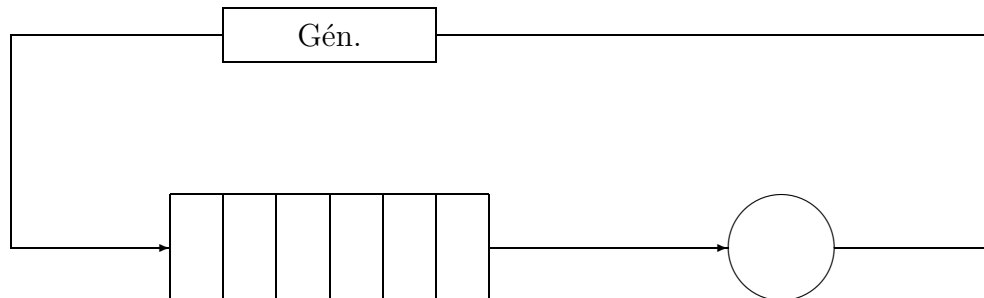
G: distribution quelconque

3.2.2 Kendall/C (capacité)

Si $C \neq \infty$, tout client arrivé alors que le système contient déjà C clients (en attente ou service) est *éliminé*

3.2.3 Kendall/P(Taille population)

Si $P \neq \infty$, les clients circulent généralement en circuit fermé.



3.2.4 Kendall/Ds(discipline)

Règle précisant dans quel ordre les clients dans la file sont servis :

FIFO : (First In First Out)

FILO : (First In Last Out)

SIRO : (Service In Random Order) : Choix d'un client au hasard parmi ceux dans la file.

PS : (Processor Sharing) : Si S clients sont dans la file, chacun reçoit une part $\frac{1}{n}$ de la capacité de service.

1. v.a. = variables aléatoires

2. i.i.d = indépendantes et identiquement distribuées

RR : (Round Robin) : Chaque client est servi à tour de rôle pendant un intervalle de temps de durée fixe.

PR : (PRioritaire) : les clients sont divisés en différentes classes auxquelles sont associées des priorités de traitement. Les clients de forte priorité sont servis avant les autres.

Plusieurs variantes d'interruption :

- on n'interrompt pas le service entamé d'un client de priorité inférieure.
- on interrompt son service et on le remet dans la file. (il recommencera ou reprendra son service plus tard).

3.2.5 Exemple

Exemple : Notre problème de serveur de base de données était une file

$$\boxed{M|M|1}$$

Explication :

M : Arrivée; $\text{Exp}(\lambda = 40)$

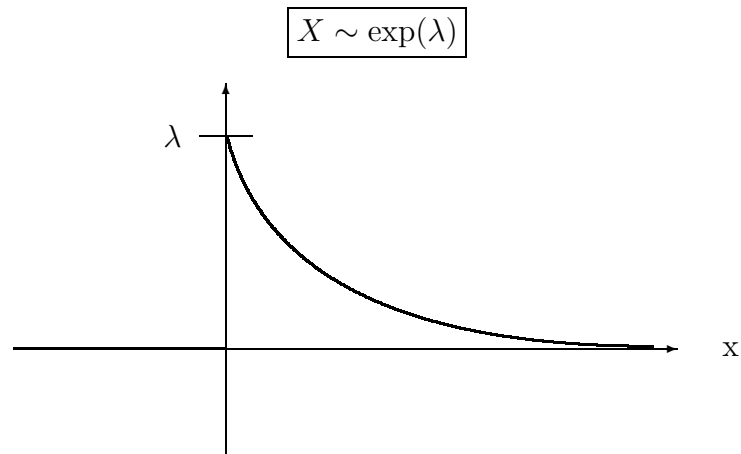
M : Service; $\text{Exp}(\mu = 50)$

1 : 1 serveur

3.3 Loi exponentielle de param $\lambda > 0$, rappel

Utilisation : Temps d'attente ; durée de vie.

3.3.1 Histogramme



$$\text{densité: } f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

$$E(x) = \frac{1}{\lambda}$$

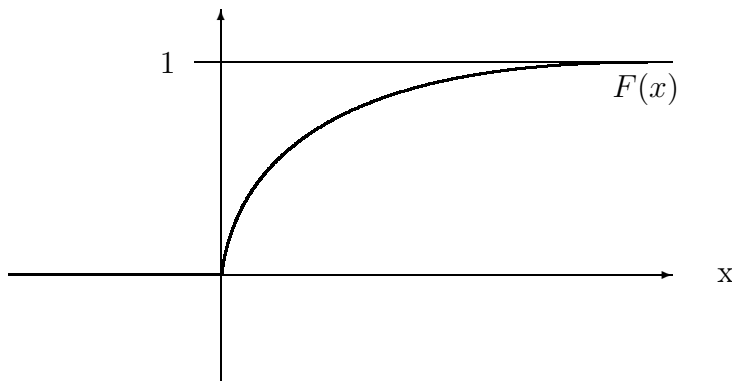
$$\text{Var}(x) = \frac{1}{\lambda^2}$$

3.3.2 Fonction de répartition

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

Pour une loi expo :

$$\text{pour } x \geq 0 : F(x) = \int_{-\infty}^x \lambda e^{-\lambda t} dt = -e^{-\lambda t} \Big|_0^x = 1 - e^{-\lambda x}$$



3.3.3 Espérance et variance d'une v.a. $X \sim \exp(\lambda)$

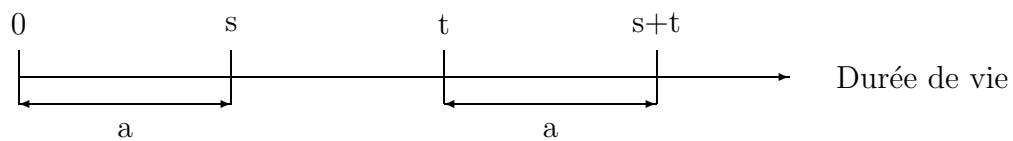
$$E(x) = \int_{-\infty}^{\infty} x f(x) dx = \dots = \frac{1}{\lambda}$$

$$\text{var}(X) = \int_{-\infty}^{\infty} (x - E(x))^2 f(x) dx = \dots = \frac{1}{\lambda^2}$$

3.3.4 Absence de mémoire d'une v.a.

Soit $X \sim \exp(\lambda)$, $\lambda > 0$ pour tout $t, s \geq 0$, on a

$$P\{X > s + t | x > t\} = \frac{P\{x > s+t \text{ et } x > t\}}{P\{x > t\}} = \frac{P\{x > s+t\}}{P\{x > t\}} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = P\{x > s\}$$



3.4 Formule de Little

Il s'agit d'une formule valide pour *n'importe quel* système *stable* (i.e. pour lequel un équilibre stochastique³ s'est établi) où l'on peut *définir* :

λ : le taux d'arrivée des clients dans le système

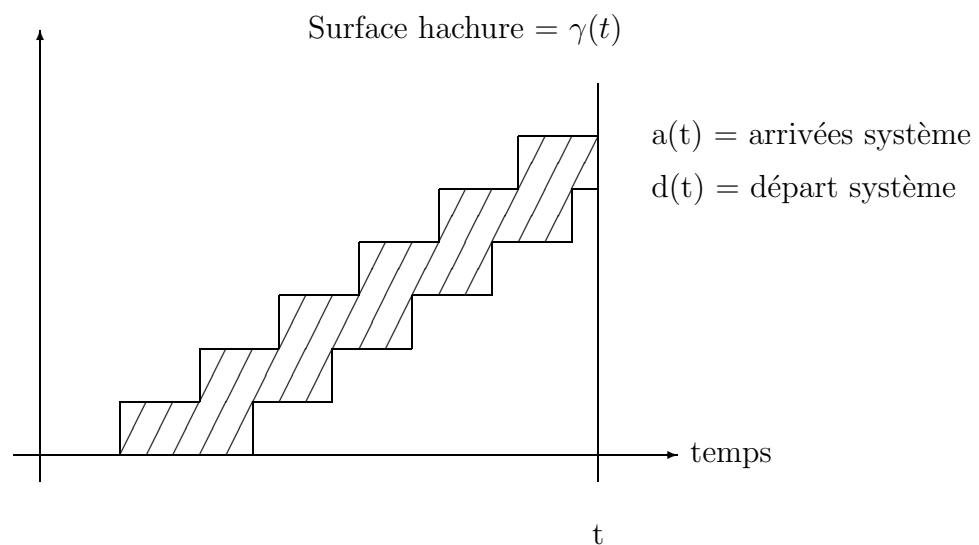
N : le taux moyen de clients dans le système

T : le temps moyen de séjour d'un client dans le système

$$\text{Loi de Little: } N = \lambda T$$

Remarque : Ce résultat n'exige que l'existence d'un régime stationnaire. Il ne dépend pas de la (ou des) files sous-jacentes : nombre de serveurs, règles de priorités, lois probabilistes.

Preuve :



$\mathbf{a}(t)$: nb clients arrivés sur $[0, t]$

$\mathbf{d}(t)$: nb clients partis sur $[0, t]$

$\mathbf{N}(t)$: nb clients présents au temps $t = a(t) - d(t)$

$\gamma(t)$: temps total passé par tous les clients dans le système = $\int_0^t N(s) ds$

– Taux moyen d'arrivée au temps t :

$$\lambda_t = \frac{a(t)}{t}$$

3. stochastique : qui est de nature aléatoire

- Nb moyen de clients dans le système sur $[0,t]$:

$$N_t = \frac{\gamma(t)}{t}$$

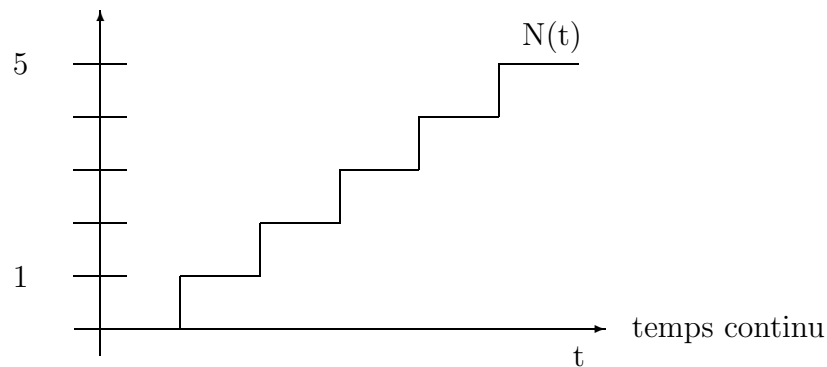
- Temps moyen de séjour d'un client sur $[0,t]$:

$$T_t = \frac{\gamma(t)}{a(t)} \Rightarrow N(t) = \frac{\gamma(t)}{t} = T_t \frac{a(t)}{t} = T_t \lambda_t$$

pour $t \rightarrow \infty$, en invoquant l'existence d'un régime stationnaire, on obtient $N = \lambda T$

3.5 Processus de comptage

(pour processus d'arrivée G)



Un processus stochastique $N(t), t \geq 0$ est un processus de comptage si $N(t)$ représente le nombre total d'événements aléatoires qui se sont produits sur $[0,t]$

Propriétés

- $M(t) \geq 0, \forall t \geq 0$
- $M(t)$ est un entier $\forall t \geq 0$
- $s < t \Rightarrow N(s) \leq N(t), \forall s, t \geq 0$
- pour $s < t, N(t) - N(s)$ = "Nombre d'événements ayant eu lieu sur $[s,t]$ "

3.6 Processus de Poisson

Un processus de comptage $N(t), t \geq 0$ est un processus de *Poisson* de paramètre λ si :

- $N(0) = 0$
- Les incréments $N(t)$ et $[N(s+t) - N(t)]$ sont des v.a. indépendantes $\forall s, t \geq 0$

- Le nombre d'événements dans l'intervalle $[s, s+t]$ suit une *loi de Poisson* de paramètre λ .

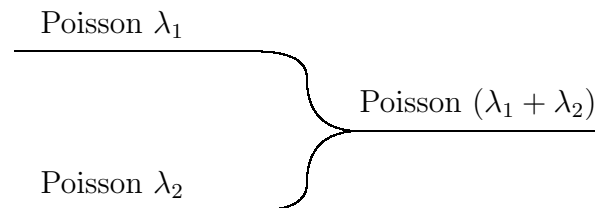
$$P\{N(t+s) - N(t) = n\} = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \forall n \in (\text{ensemble } \mathbb{N})$$

3.6.1 propriétés d'un processus de Poisson

- les incréments sont *stationnaires*: $[N(t_2 + s) - N(t_1 + s)]$ et $[N(t_2) - N(t_1)]$ ont la même distribution $\forall t_1 < t_2$ et $s > 0$
- $P\{N(t) = n\} = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \forall n \in (\text{ensemble } \mathbb{N})$
- Espérance $E(N(t)) = \lambda t$
- Variance $\text{var}(N(t)) = \lambda t$

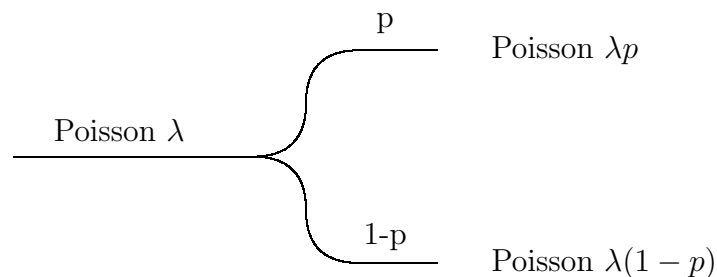
3.7 Mélange de sources de clients

La superposition de deux processus de Poisson *indépendants* de paramètres λ_1 et λ_2 engendre un processus de Poisson de paramètre $\lambda_1 + \lambda_2$



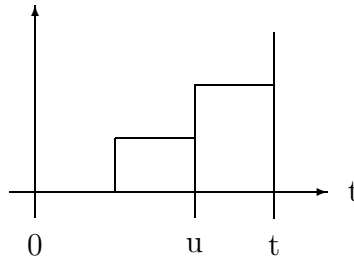
3.8 Séparation aléatoire des clients d'une source

Soit une probabilité $p \in [0,1]$ et un processus de Poisson de paramètre λ . On génère deux processus de comptage en leur affectant chacune des arrivées avec probabilités respectives p et $(1-p)$. On peut montrer que les deux processus obtenus sont des processus de Poisson *indépendants*, de paramètre λp et $\lambda(1-p)$



3.9 Liens processus de Poisson / v.a. exponentielles

Soit $N(t), t \geq 0$ un processus de Poisson de paramètre $\lambda > 0$



Soit x la durée entre la $i^{\text{ème}}$ arrivée (qui a lieu en $t = u$) et la $(i + 1)^{\text{ème}}$. Pour $x \geq 0$, on a :

$$\begin{aligned}
 P\{X \leq x\} &= 1 - P\{N(u+x) - N(u) = 0\} \\
 &= 1 - P\{N(x) = 0\} \\
 &= 1 - e^{-\lambda x} \frac{(\lambda x)^0}{0!} \\
 &= 1 - e^{-\lambda x} \\
 &\Rightarrow X \text{ est une v.a. exponentielle de paramètre } \lambda
 \end{aligned}$$

Processus de Poisson $\lambda \Leftrightarrow$ Temps inter-arrivées i.i.d. $\text{Exp}(\lambda)$.

3.10 File M|M|1

3.10.1 Rappel

M : Arrivées selon processus de Poisson de paramètre $\lambda > 0$

M : Temps de service i.i.d; $\text{Exp}(\mu)$, $\mu > 0$

1 : Un serveur

... : Capacité et population infinis

... : Discipline de service FIFO (ou autre parmi celles proposées au cours)... n'intervient pas ici

3.10.2 Notation

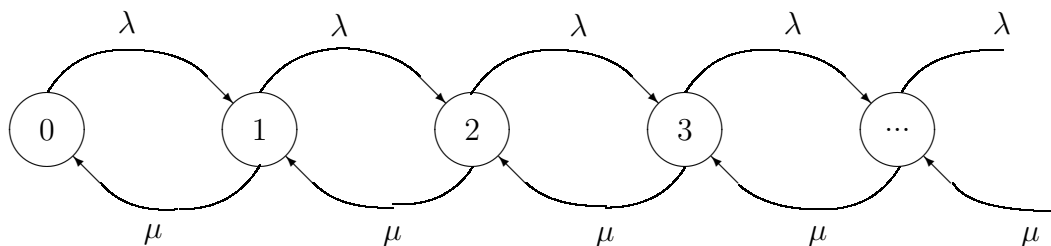
On note (si elle existe) $\Pi = (\Pi_0, \Pi_1, \Pi_2, \dots)$ la distribution stationnaire du nombre de clients dans le système :

$\Pi_i = \lim_t \rightarrow P\{\text{on a } i \text{ clients dans le système}\}$

\cong proportion du temps où i clients sont dans le système.

3.10.3 Graphe

Un événement est soit un départ, soit une arrivée et l'on peut donc représenter les transitions d'état du système à l'aide du graphe suivant :



Les sommets de ce graphe symbolisent le nombre de clients dans le système et les arcs les intensités de transitions possibles. La conservation des clients au cours du processus implique le flot entrant dans le sommet i (le taux auquel le nombre de clients dans le système devient i) est égal au flot sortant de i .

3.10.4 Equations

On a donc les équations du bilan :

$$* \{ \lambda \Pi_0 = \mu \Pi_1 \}$$

$$(\lambda + \mu)\Pi_i = \lambda\Pi_{i-1} + \mu\Pi_{i+1} \quad \forall i \in (\text{ensemble } \mathbb{N}^*)$$

auxquelles on ajoute la contrainte de normalisation

$$** \left\{ \sum_{i=0}^{\infty} \Pi_i = 1 \right\}$$

3.10.5 Résolution

En introduisant le paramètre $\rho = \frac{\lambda}{\mu}$,
en divisant par μ , le système (*) devient :

$$\begin{cases} \rho\Pi_0 = \Pi_1 \\ (\rho + 1)\Pi_i = \rho\Pi_{i-1} + \Pi_{i+1} \end{cases} \quad \forall i \in (\text{ensemble } \mathbb{N}^*)$$

pour i=1 :

$$\begin{aligned} (\rho + 1)\Pi_1 &= \rho\Pi_0 + \Pi_2 \\ \rho\Pi_1 &= \Pi_2 (= \rho^2\Pi_0) \end{aligned}$$

pour i=2 :

$$\begin{aligned} (\rho + 1)\Pi_2 &= \rho\Pi_1 + \Pi_3 \\ \rho\Pi_2 &= \Pi_3 (= \rho^3\Pi_0) \end{aligned}$$

...

en procédant par récurrence sur i , on obtient

$$\begin{aligned} \Pi_{i+1} &= \rho\Pi_i & \forall i \in (\text{ensemble } \mathbb{N}) \\ \Rightarrow \Pi_i &= \rho^i\Pi_0 & \forall i \in (\text{ensemble } \mathbb{N}) \end{aligned}$$

la contrainte (**) devient :

$$\begin{aligned} \sum_{i=0}^{\infty} \Pi_0 \rho^i &= 1 \\ \Rightarrow \Pi_0 \underbrace{\left(\sum_{i=0}^{\infty} \rho^i \right)}_a &= 1 \end{aligned}$$

$$\begin{aligned} a \text{ converge vers } \frac{1}{1-\rho}, \text{ pour } \rho \in [0,1[\\ \Rightarrow \boxed{\Pi_0 = 1 - \rho} \end{aligned}$$

3.10.6 Conclusion

Sous la condition

$$\rho = \frac{\lambda}{\mu} < 1 \Leftrightarrow \lambda < \mu$$

on obtient la solution

$$\begin{cases} \Pi_0 = 1 - \rho \\ \Pi_i = (1 - \rho)\rho^i \end{cases} \quad \forall i \in (\text{ensemble } \mathbb{N}^*)$$

3.10.7 Exemple

Problème d'avoir 7 clients dans le système :

$$\Pi_7 = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^7$$

une file M|M|1 ne présente un régime stationnaire que pour

$$\rho = \frac{\lambda}{\mu} < 1$$

Dans le cas contraire, la longueur de la file a tendance à croître indéfiniment.

3.10.8 Mesures de performance d'une file M|M|1

1. Taux moyen d'occupation du serveur :

$$1 - \Pi_0 = 1 - (1 - \rho) = \rho = \frac{\lambda}{\mu}$$

2. Nombre moyen de clients dans le système (attente + service)

$$\begin{aligned} N &= \sum_{i=0}^{\infty} i\Pi_i = \sum_{i=0}^{\infty} i(1 - \rho)\rho^i \\ &= (1 - \rho) \sum_{i=0}^{\infty} i\rho^i = \frac{\rho}{(1 - \rho)^2} \quad \text{pour } \rho \in [0,1[\\ &= \frac{\rho}{1 - \rho} = \frac{\frac{\lambda}{\mu}}{1 - \frac{\lambda}{\mu}} = \frac{\lambda}{\mu - \lambda} \end{aligned}$$

3. Temps moyen de réponse (attente + service)

$$T = \frac{N}{\lambda} = \frac{1}{\mu - \lambda} \quad (\text{par la formule de Little})$$

3.10.9 Exemple : Mesures de performance de notre serveur de base de données

Modélisée par une file M|M|1 donnée par $\lambda = 40$ requêtes par seconde et $\mu = 50$ requêtes traitées par secondes. Comme $\rho = \frac{\lambda}{\mu} = \frac{4}{5} < 1$, le système est *stable*.

Mesures de performance :

Taux moyen d'occupation du serveur : $\rho = \frac{4}{5} = 80\%$

Taux moyen d'arrivées : $\lambda = 40$

Durée moyenne de service : $\frac{1}{\mu} = 0.02$ sec.

Temps moyen de réponse : $T = \frac{1}{\mu - \lambda} = 0.1$ sec.

Nb moyen de requetes dans le système : $N = \frac{\lambda}{\mu - \lambda} = \frac{40}{10} = 4$

Si une hausse de 50% des requêtes est à prévoir :

$$\lambda' = \frac{3}{2} \quad \lambda = 60$$

$$\Rightarrow \rho' = \frac{\lambda'}{\mu} = \frac{60}{50} > 1$$

\Rightarrow Le système ne pourra pas suivre

Donc, nous devons changer de serveur. Quel doit être son temps moyen de service pour que le temps moyen de réponse reste à 100ms?

$$T' = \frac{1}{\mu' - \lambda'} = \frac{1}{\mu' - 60} = 0.1s$$

$$\Rightarrow \mu' = 70 \text{ requetes par seconde}$$

Pour maintenir le temps moyen de réponse à 100ms, une augmentation de 50% de λ nécessite une augmentation de 40% de μ

\Rightarrow Pas linéaire!

3.11 Analyse de sensibilité

Soit une file M|M|1 stable donnée par son taux d'arrivée $\lambda = 1$. Considérons le temps moyen de réponse $T = \frac{1}{\mu - \lambda} = \frac{1}{\mu - 1}$ comme une fonction du taux moyen de service $\mu > 1$:

1. Réduction du temps moyen de réponse de 50% pour une augmentation de 9.09% du temps moyen de service. (μ de 1.1 à 1.2)

2. Réduction du temps moyen de réponse de 50% pour une augmentation de 66.6% du temps moyen de service. (μ de 3 à 5)

Morale : le dimensionnement d'un système n'est pas linéaire, donc pas intuitif.

3.12 Mesures de performances $M|M|1$

Nombre moyen de clients dans le système :

$$N = \frac{\lambda}{\mu - \lambda}$$

Temps moyen de réponse :

$$T = \frac{1}{\mu - \lambda}$$

Nombre moyen de clients en attente :

$$Na = \sum_{i=0}^{\infty} i\pi_{i+1} = \sum_{i=0}^{\infty} i(1-\rho)\rho^{i+1} = \rho(1-\rho) \sum_{i=0}^{\infty} i\rho^i = \rho(1-\rho) \frac{\rho}{(1-\rho)^2} = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

Temps moyen d'attente (par la formule de Little) :

$$Ta = \frac{Na}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)}$$

Temps moyen de service :

$$Ts = \frac{1}{\mu}$$

Nombre moyen de clients en service = taux d'occupation du serveur :

$$Ns = \underbrace{\lambda Ts}_{\text{par Little}} = \frac{\lambda}{\mu} = \rho = 1 - \pi_0$$

On peut voir que :

$$N = Na + Ns$$

$$T = Ta + Ts$$

3.13 Exercices

3.13.1 Exercice 1

On a une file M|M|1 avec $\lambda = 5$ et $\mu = 6$.

1. $\frac{\lambda}{\mu} = \rho = \frac{5}{6} > \frac{3}{4}$, donc il faut une nouvelle cabine.
2. $Na = \frac{\rho^2}{1-\rho} = \frac{25}{6}$
3. $Ta = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{5}{6}$ d'heure
4. $T = \frac{1}{\mu-\lambda} = 1$ heure

3.13.2 Exercice 2

$\lambda = 100$ et $N = 5 = \frac{\lambda}{\mu-\lambda}$, donc $\mu = 120$.

1. Dans les trois cas, $\rho = \frac{\lambda}{\mu} = \frac{100}{120} = \frac{5}{6}$
2. Dans les trois cas, $T = \frac{1}{\mu-\lambda} = \frac{N}{\lambda} = 50\text{ms}$
3. Dans les trois cas, $Na = \frac{\rho^2}{1-\rho} = \frac{25}{6}$
4. Dans les trois cas, $Ta = \frac{Na}{\lambda} = \frac{25}{600} = \frac{1}{24}$
 $\text{Var}(\text{Ta, FIFO}) \leq \text{Var}(\text{Ta, SIRO}) \leq \text{Var}(\text{Ta, LIFO})$
5. $\mu' = \frac{120}{2} = 60$ requêtes traitées par seconde en moyenne.
 $\rho' = \frac{\lambda}{\mu'} = \frac{100}{60} > 1$
 Le système n'est pas stable.

3.13.3 Exercice 3

Temps moyen de réponse :

$$T = \frac{1}{\mu - \lambda}$$

Temps moyen de réponse désiré :

$$T = \frac{1}{\mu - \lambda} * \frac{100 - r}{100}$$

Nouveau temps moyen de service proposé :

$$\mu' = \mu(1 + (1 - \rho)\frac{r}{100 - r})$$

Calculons le nouveau temps moyen de réponse :

$$T = \frac{1}{\mu' - \lambda} = \frac{1}{\mu(\frac{100-r+r-\rho r}{100-r}) - \lambda} = \frac{1}{\frac{100\mu - \mu\rho r - 100\lambda + r\lambda}{100-r}} = \frac{100 - r}{100(\mu - \lambda)}$$

CQFD

3.14 File M|M|K

- Une seule file d'attente.
- Arrivée: processus de Poisson de paramètre λ
- K serveurs indépendants.
- Temps de service i.i.d $\exp(\mu)$

Quel que soit l'état du système (nombre i de clients présents), le taux d'arrivée est $\lambda_i = \lambda$. Par contre, globalement, le taux de service dépend du nombre de clients i dans le système :

$$\mu_i = \begin{cases} i\mu & \text{pour } 0 \leq i \leq K \\ K\mu & \text{pour } i \geq K \end{cases}$$

Après avoir écrit et résolu les équations du bilan, on obtient la distribution stationnaire suivante pour le nombre de clients dans le système :

$$\Pi_i = \begin{cases} \Pi_0 \frac{(K\rho)^i}{i!} & \text{pour } 0 \leq i \leq K \\ \Pi_0 \frac{\rho^i K^K}{K!} & \text{pour } i \geq K \end{cases}$$

où

$$\Pi_0 = \left[\frac{(K\rho)^K}{K!(1-\rho)} + \sum_{i=0}^{K-1} \frac{(K\rho)^i}{i!} \right]^{-1}$$

et

$$\rho = \frac{\lambda}{K\mu}$$

Le régime stationnaire n'existe que si $\rho < 1$

3.14.1 Mesures de performance M|M|K

ω = probabilité qu'un client qui arrive doive attendre = probabilité que les K serveurs soient déjà occupés.

$$\omega = \sum_{i=K}^{\infty} \Pi_i = \Pi_0 \frac{(K\rho)^K}{K!(1-\rho)}$$

N Nombre moyen de clients présents

$$N = \sum_{i=0}^{\infty} i\Pi_i = K\rho + \frac{\rho\omega}{1-\rho}$$

N_a Nombre moyen de clients en attente

$$N_a = \sum_{i=K+1}^{\infty} (i-K)\Pi_i = \frac{\rho\omega}{1-\rho}$$

N_s Nombre moyen de clients en train de se faire servir

$$N_s = N - N_a = K\rho = \frac{\lambda}{\mu}$$

La formule de Little fournit les temps de réponse T , d'attente T_a et de service T_s correspondants :

T Temps de réponse moyen

$$T = \frac{N}{\lambda} = \frac{1}{\mu} + \frac{\omega}{K\mu(1-\rho)}$$

T_a Temps d'attente moyen

$$T_a = \frac{N_a}{\lambda} = \frac{\omega}{K\mu(1-\rho)}$$

T_s Temps de service moyen

$$T_s = \frac{N_s}{\lambda} = \frac{1}{\mu}$$

3.15 Exercices

3.15.1 Exercice 1

On compare les temps moyen de réponse des deux systèmes.

M|M|1: taux d'arrivée λ et de service 2μ , $\rho = \frac{\lambda}{2\mu} < 1$

$$T_1 = \frac{1}{2\mu(1-\rho)} = \frac{1}{2\mu - \lambda}$$

M|M|K: taux d'arrivée λ et de service μ , $\rho = \frac{\lambda}{2\mu} < 1$

$$\begin{aligned} \Pi_0 &= \left[\frac{4\rho^2}{2(1-\rho)} + 1 + 2\rho \right]^{-1} = \frac{1-\rho}{1+\rho} \\ \omega &= \Pi_0 \frac{(2\rho)^2}{2(1-\rho)} = \frac{2\rho^2}{1+\rho} \\ T_2 &= \frac{1}{\mu} + \frac{2\rho^2}{(1+\rho)2\mu(1-\rho)} = \frac{1}{\mu(1-\rho^2)} \end{aligned}$$

Comparaison :

$$\frac{T_1}{T_2} = \frac{\mu(1-\rho^2)}{2\mu(1-\rho)} = \frac{1+\rho}{2} \leq 1$$

Donc, il vaut mieux un processeur deux fois plus rapide.

Remarque : par contre,

$$T_{a,1} = \frac{\rho^2}{\lambda(1-\rho)} T_{a,2} = \frac{\rho^2}{\mu(1-\rho^2)} \frac{T_{a,1}}{T_{a,2}} = \frac{1+\rho}{2\rho} > 1$$

Donc l'attente est plus courte avec deux processeurs lents.

3.15.2 Exercice 2

Arrivées avions en panne : $\lambda = 0.2 \frac{\text{pannes}}{\text{semaines}}$

Traitement : $\mu = 1 \frac{\text{réparation}}{\text{semaine}}$

Une seule équipe : M|M|1

Nombre moyen d'avions dans le système : $\frac{\lambda}{\mu-\lambda} = \frac{1}{4}$

Coût moyen hebdomadaire : $100'000\frac{1}{4} + 4'000 = 29'000.-$

Deux équipes : M|M|2

$$\lambda = 0.2, \mu = 1, \rho = \frac{\lambda}{2\mu} = 0.1$$

$$\omega = \frac{2\rho^2}{1+\rho}$$

Nombre moyen d'avions dans le système :

$$N = 2\rho + \frac{\rho\omega}{1-\rho} = \frac{2\rho}{1-\rho^2} = \frac{20}{99}$$

Coût moyen hebdomadaire : $100'000\frac{20}{99} + 2 * 4000 = 28'202.-$

Donc, il vaut mieux deux équipes.

3.15.3 Exercice 3

Système monoprocesseur coûte 10.- / seconde de calcul

On a une M|M|1 avec $\lambda = 0.3, \mu = 0.4 \frac{\text{job}}{\text{heure}}$

Nombre moyen de jobs dans le système : $N = \frac{\lambda}{\mu-\lambda} = 3 \text{ jobs.}$

Taux d'utilisation du processeur : $1 - \Pi_0 = \rho = \frac{\lambda}{\mu} = 0.75$

Coût moyen par seconde : $15 * 3 + 10 * 0.75 = 52.50$

Système biprocesseur coûte :

- 20.- / seconde de calcul pour un seul processeur
- 30.- / seconde de calcul pour les deux

On a une M|M|2, λ et μ sont les mêmes. $\rho = \frac{\lambda}{2\mu} = \frac{3}{8}$

Nombre moyen de jobs dans le système : $N = \frac{2\rho}{1-\rho^2} = \frac{48}{55} \text{ jobs.}$

$$\Pi_0 = \frac{1-\rho}{1+\rho} = \frac{5}{11}$$

$$\Pi_1 = \Pi_0 2\rho = \frac{15}{44}$$

La fraction du temps où on utilise un processeur :

$$\Pi_1 = \Pi_0 2\rho = \frac{15}{44}$$

La fraction du temps où on utilise deux processeurs :

$$1 - (\Pi_0 + \Pi_1) = \frac{9}{44}$$

Coût moyen par seconde :

$$15\frac{48}{55} + 20\frac{15}{44} + 30\frac{9}{44} = 26.045$$

Donc, le système biprocesseur est préférable.

3.16 Autres files classiques

- M|M|1|1|C
- M|M|1|1| ∞ |P
- M|M|1|1| ∞ | ∞ |Pr

Références :

- Kleinrock 75
- Ross 97
- Nelson 95
- Fdida & Pajolle 89
- Ajmone et al 86

3.17 File M|G|1

Modèle plus riche, plus réaliste. Arrivée selon un processus de Poisson(λ), traitement selon une distribution non négative quelconque.

Exemples pour S :

Déterministe : 50ms tout le temps

Uniforme : entre 20 et 100ms

Erlang d'ordre 3 et de paramètre $\mu = \frac{1}{50}$: Somme de trois traitements indépendants distribués selon une $Exp(\mu)$

Discret :

$$\begin{cases} 10ms \text{ avec prob } \frac{1}{4} \\ 20ms \text{ avec prob } \frac{1}{4} \\ 40ms \text{ avec prob } \frac{1}{2} \end{cases}$$

Exponentielle : redonne une M|M|1

3.17.1 Mesures de performance

Taux moyen d'occupation du serveur : $\rho = \lambda E(s)$

Condition de stabilité : $\rho < 1$

Distribution Π_0, Π_1, \dots est inconnue sous forme explicite dans le cas général.

Formules de Pollaczek-Khintchine pour les nombres moyen de clients dans le système ou en attente.

$$\begin{aligned}
 N &= \rho + \frac{\lambda^2}{2(1-\rho)} [(E(S))^2 + Var(S)] \\
 Na &= N - \rho = \frac{\lambda^2}{2(1-\rho)} [(E(S))^2 + Var(S)]
 \end{aligned}$$

Par Little, on obtient les temps moyen correspondants :

$$\begin{aligned}
 T &= \frac{N}{\lambda} = E(S) + \frac{\lambda}{2(1-\rho)} [(E(S))^2 + Var(S)] \\
 Ta &= \frac{Na}{\lambda} = \frac{\lambda}{2(1-\rho)} [(E(S))^2 + Var(S)] = T - E(S)
 \end{aligned}$$

3.18 File G|G|1

Distributions quelconques pour les temps inter-arrivées A et le temps de service S. C'est un cas très difficile à analyser. On ne dispose que d'approximations.

Condition de stabilité: $\rho = \frac{E(S)}{E(A)} < 1$

3.18.1 Mesures de performance

Approximations :

$$\begin{aligned}
 N &\cong \rho + \frac{Ca+Cs}{2} \frac{\rho^2}{1-\rho} \\
 Na &\cong \frac{Ca+Cs}{2} \frac{\rho^2}{1-\rho} = N - \rho
 \end{aligned}$$

Par Little, on obtient les temps moyen correspondants :

$$\begin{aligned}
 T &\cong E(S) + \frac{Ca+Cs}{2} \frac{\rho E(S)}{1-\rho} \\
 Ta &\cong \frac{Ca+Cs}{2} \frac{\rho E(S)}{1-\rho} \\
 &\text{où} \\
 Ca &= \frac{Var(A)}{(E(A))^2} \\
 Cs &= \frac{Var(S)}{(E(S))^2}
 \end{aligned}$$

Les approximations N, Na, T, Ta sont des bornes supérieures.