Première partie

Introduction à la Qualité de Service - Métrologie

1 Définitions

1.1 Débit :

On a des évènements discrets qui arrivent selon un certain processus d'arrivéequi suit une certaine loi à des dates successives [1]. Soient $a_0, a_1, a_2, \ldots, a_n$, les temps d'arrivées, on va noter la n^{ieme} interarrivée :

$$\tau_n = a_{n+1} - a_n$$

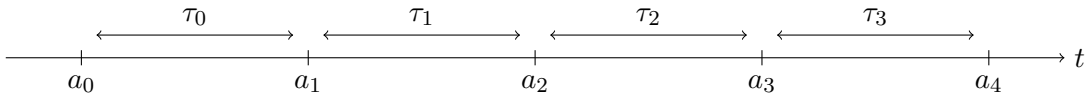

FIGURE 1 – Evolution du temps

On notera $N(t) =$ le nombre d'arrivées avant t :

$$N(t) = card\{n \ / \ a_n \leq t\}$$

Le débit d'arrivée est défini par :

$$\theta = \lim_{y \to \infty} \frac{N(t)}{t}$$

C'est le nombre moyen d'arrivées par unité de temps.

On estime le débit d'arrivées par :

$$\bar{\theta} = \frac{N(T)}{T} \text{pour T grand}$$

Le nombremoyen d'arrivées par unité de temps :

$$\bar{\tau} = \lim_{m \to \infty} \frac{1}{n} \sum_{i=0}^{n} \tau_i = \lim_{n \to \infty} \frac{a_n}{n}$$

On peut estimer :

$$\bar{\tau} = \frac{a_N}{N} \qquad \text{pour N grand}$$

On voit alors que :

$$\theta = \frac{1}{\tau} \quad \text{ou} \quad \tau = \frac{1}{\theta}$$

1. on est dans un principe de discètisation des évènements

1.2 Débit d'information :

Chaque paquet a une quantité d'information donnée. On définit alors le débit d'information comme le rapport de la quantité total d'information transmise par unité de temps. Soit σ_n la quantité d'information du paquet n. La quantité totale d'information est :

$$\sum_{i=1}^{n} \sigma_i$$

La durée total est :

$$\sum_{i=1}^{n} \tau_i$$

On définit alors $\bar{\delta}$, le débit moyen d'information, ainsi :

$$\bar{\delta} = \lim_{n \to \infty} \frac{\sum_{i=1}^{n} \sigma_i}{\sum_{i=1}^{n} \tau_i}$$

On peut définir la quantité moyenne d'information d'un paquet :

$$\bar{\sigma} = \lim_{n \to \infty} \frac{\sum_{i=1}^{n} \sigma_i}{n}$$

On peut alors écrire ce qui suit :

$$\bar{\delta} = \bar{\theta} \times \bar{\sigma}$$

Ce qui signifie que le débit d'information ($bit.s^{-1}$) est le produit du débit ($paquet.s^{-1}$) par la taille moyenne des paquets (bit).

1.3 Débit crête :

Le débit est irrégulier dans la plupart des systèmes. On considère des processus de type ON/OFF, soit ta' le débit crête, on a :

FIGURE 2 – Processus de type ON/OFF

On appelle : X_i = périodes d'activité, Y_i = périodes d'inactivité, $\bar{X}$ = période moyenne d'activité et $\bar{Y}$ période moyenne d'inactivité. On peut alors définir le débit moyen comme :

$$\bar{\theta} = \theta' \times \frac{\bar{X}}{\bar{X} + \bar{Y}}$$

1.4 Lois de conservation de débit

La loi est valable uniquement dans un système stable : à savoir aucun paquet ne se perd dans le réseau.

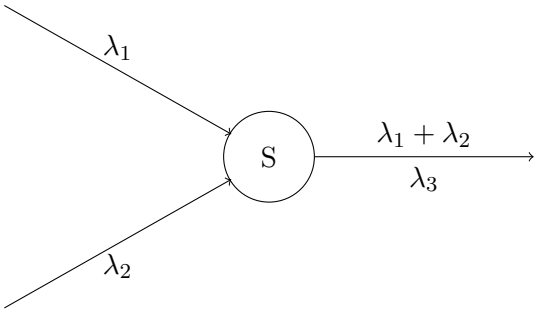

FIGURE 3 – Conservation du débit (1)

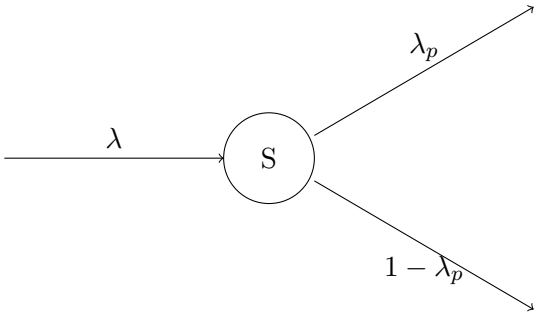

FIGURE 4 – Conservation du débit (2)

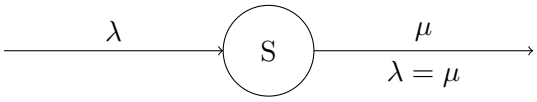

FIGURE 5 – Conservation du débit (3)

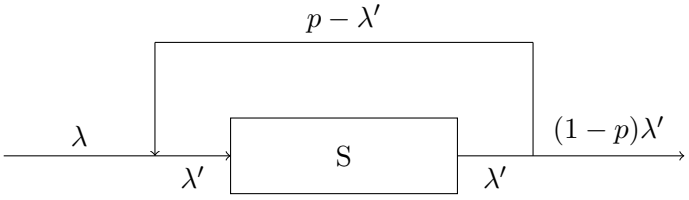

FIGURE 6 – Exercice : petit réseau

Un petit exercice

Soit le réseau présenté sur la figure 1.4.

Calculer λ'.

Par a loi de conservation des débits, on a :

$$\begin{aligned} & (1-p)\lambda' &=& \lambda \\ \Longrightarrow \quad & \lambda' &=& \frac{\lambda}{1-p} \end{aligned}$$

1.5 Temps de réponse

a_n est le début d'activité de la n^e activité. On note d_n la date de fin de la n^e activité. Le temps de réponse est alors défini par :

$$r_n = d_n - a_n$$

Le temps de réponse moyen est défini par :

$$\bar{r} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} r_i$$

qui peut aussi être défini ainsi :

$$\bar{r}_N = \frac{1}{N} \sum_{i=1}^{N} r_i \quad \text{pour N grand}$$

On peut alors calculer la variance de r :

$$var(r) = \frac{1}{N-1} \sum_{i=1}^{n} (r_m - \bar{r}_N)^2$$

L'objectif est alors de minimiser la variance.

1.6 Taux d'utilisation

Reconsidérons les processus ON/OFF. On définit le taux d'utilisation u par :

$$u = \frac{\bar{X}}{\bar{X} + \bar{Y}}$$

u est alors le taux d'utilisation du système. En pratique, il faut estimer le temps d'utilisation. On comptabilise la durée totale d'activité :

$$\bar{u}_T = \frac{1}{T} \times D \quad D \in [0, T], D \text{ étant la durée d'activité}$$

Où $\bar{u}_T$ est le taux moyen d'utilisation sur la période T.

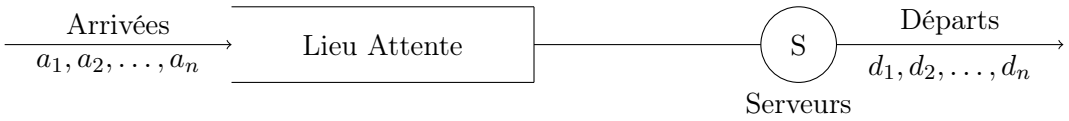

FIGURE 7 – Une file d'attente

1.7 Taux de perte

Le taux de perte définit la proportion de requêtes qui ont échoué, on peut aussi le voir comme la probabilité qu'une requête quelconque échoue. Considérons un système de traitement qui peut perdre des paquets, il y a alors 2 flux de données :

- Celui des paquets perdus
- Celui des paquets servis ou acceptés [2]

La somme des deux est appelé débit offert. Le taux de perte Π_p est donné par :

$$\Pi_p = \frac{\lambda - \lambda_{eff}}{\lambda}$$

$$\Longrightarrow 1 - \Pi_p = \frac{\lambda_{eff}}{\lambda}$$

2 Description des files d'attente

2.1 Caractéristiques communes

On peut définir un modèle de file d'attente par ce qui sit :

1. des entités défilent dans le système
2. utilisation de ressources
3. des comportements pas complètement prévisibles

Une file d'attente est constituée :

- d'un ou plusieurs serveurs
- de clients demandant un service
- d'un lieu où les clients attendent d'être servis lorsqu'aucun serveur n'est disponible
- d'arrivées étant des processus stochastiques [3]
- de processus de renouvellement des populations [4]
- durée de services [5]

2.2 Notation de Kendall

$A/S/P/K/D$

Un peu d'éclaircissement :

- A : décrit la loi des arrivées
- S : décrit la loi des services
- P : nombre de serveurs (par défaut $P = 1$)
- K : capacité de traitement du système (i.e. le nombre maximum de clients $K = \infty$)

2. appelé débit efficace
3. Il s'agit de modèles probabilistes dont on connaît les distributions : moyenne, variances, écart-type, ...
4. Une variable aléatoire modèlisant les arrivées
5. Aussi définie par une variable aléatoire

– D : politique de services (FIFO, LIFO, Priorités [6])

2.2.1 Loi d'inter-arrivée et processus de Poisson

La loi des inter-arrivées peut être une loi déterministe (D), on observe en général ce cas lorsque le processus est périodique.

Elle peut être une loi d'Erlang (E_m), elle est alors définie ainsi :

$$Q(N) = \frac{\frac{x^N}{\mu^N N!}}{1 + \frac{x}{\mu} + \cdots + \frac{x^N}{\mu^N N!}}$$

$\frac{1}{\mu}$ est alors la durée moyenne d'un appel [7]. $\frac{1}{x}$ représente le temps moyen entre 2 appels, et $Q(N)$ est la probabilité qu'un appel trouve toutes les lignes occupées.

Elle peut être de distribution générale (G), elle englobe tous les cas.

La loi qui nous intéresse est notée M pour markov, il s'agit d'une loi de distribution exponentielle, dans ce cas le processus d'arrivée est un processus de Poisson.

Considérons une suite de variables aléatoires :

$$0 \leq t_1 < t_2 < \cdots < t_n < t_{n+1} < \ldots$$

La variable aléatoire enregistre la date d'occurence du n^e évènement dans une expérience [8]. Ce processus est un processus stochastique de Poisson avec un taux (paramètre) de $\lambda > 0$ si :

1. $\mathbb{P}$(un seul évènement dans un intervalle h) $= \lambda \times h + o(h)$
2. $\mathbb{P}$(plus d'un évènement durant h) $= O(h)$
3. Le nombre d'évènements qui arrivent dans des intervalles qui ne chevauchent pas sont indépendants les uns des autres

La probabilité d'avoir n arrivées dans un intervalle de durée h est alors donnée par :

$$\begin{aligned} \mathbb{P}_n(h) &= \mathbb{P}(N(h) = n) \\ &= \frac{(\lambda h)^n}{n!} e^{-\lambda h} \end{aligned}$$

La probabilité nous sert juste à calculer le nombre moyen d'arrivées dans un intervalle de temps h qui nous est donnée par :

$$\mathbb{E}(N(h)) = \lambda \times h$$

Intuitivement on comprends alors que le nombre moyen d'arrivées par unité de temps est λ. La probabilité que la durée d'attente séprant deux arrivées régies par un processus de Poisson soit inférieure à x est donnée par :

$$\mathbb{P}(\tau \leq x) = 1 - e^{-\lambda x}$$

Lors d'arrivées régies par une loi de Poisson, la loi d'inter-arrivée suit une loi exponentielle.

2.3 Evolution de la file d'attente

On appelle $N(t)$: le nombre de clients par unités de temps. Considérons le tableau des processus de la figure 8.

On prend la courbe de charge et on regarde le nombre de clients par unités de temps.

Num. Client	Arrivée	Tps Service	Prio
1	0	4	
2	2	8	
3	6	4	
4	8	4	
5	15	2	

FIGURE 8 – Tableau des clients

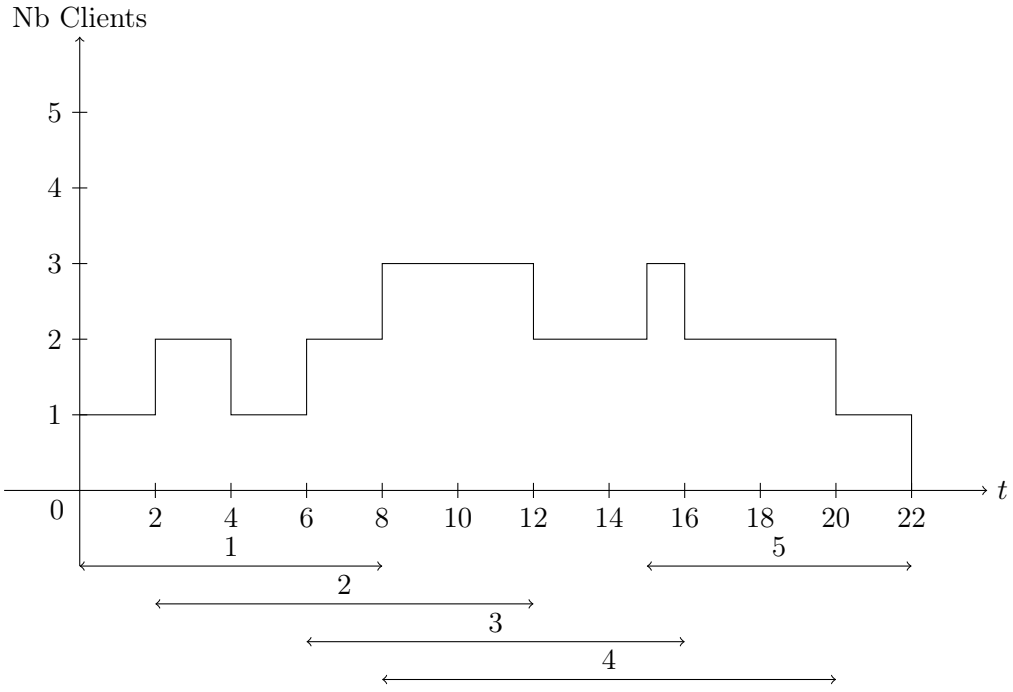

FIGURE 9 – Nombre de clients en fonction du temps en mode FIFO

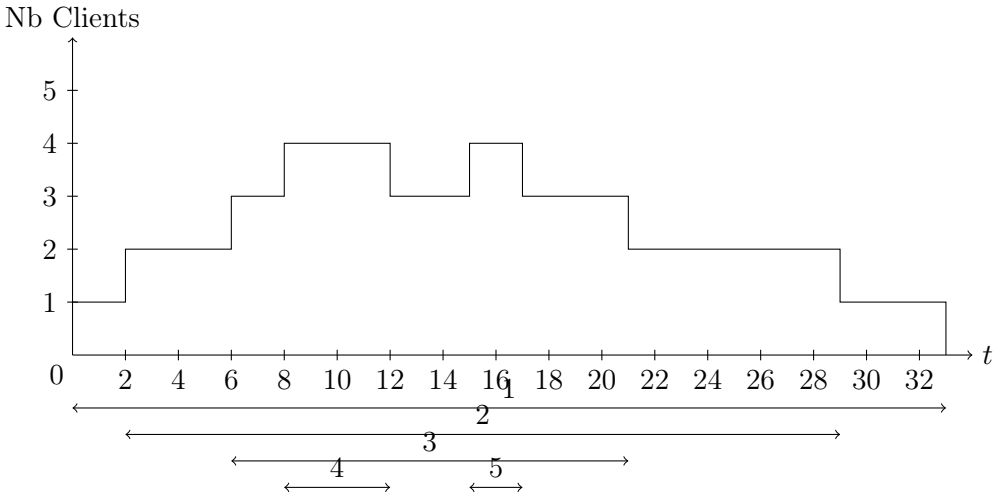

FIGURE 10 – Nombre de clients en fonction du temps en mode LIFO sans mémoire de travail

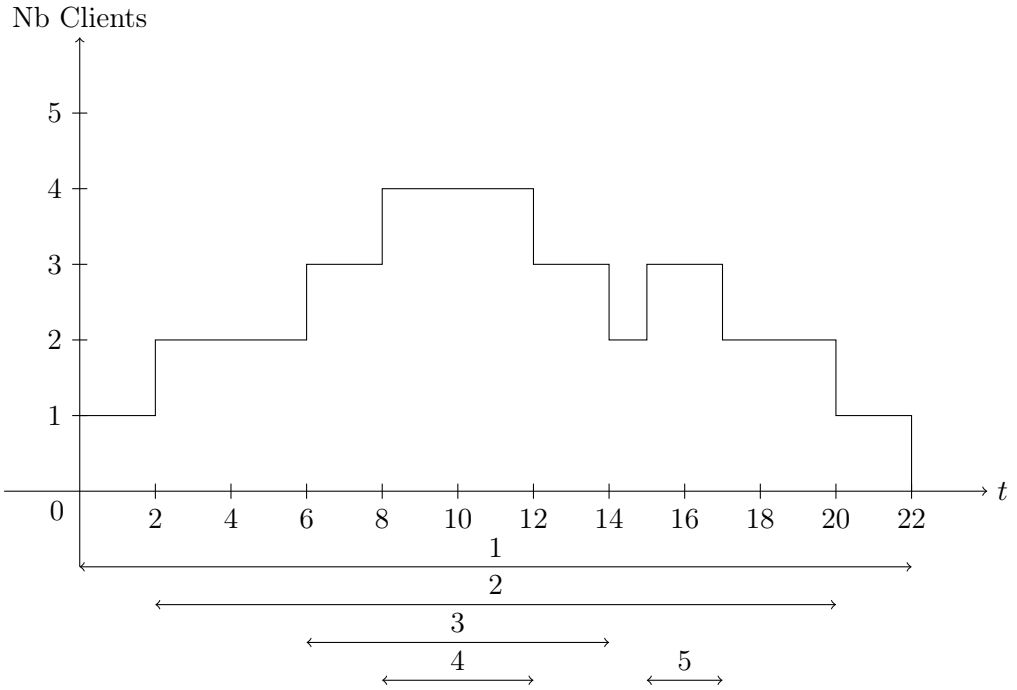

FIGURE 11 – Nombre de clients en fonction du temps en mode LIFO avec mémoire de travail

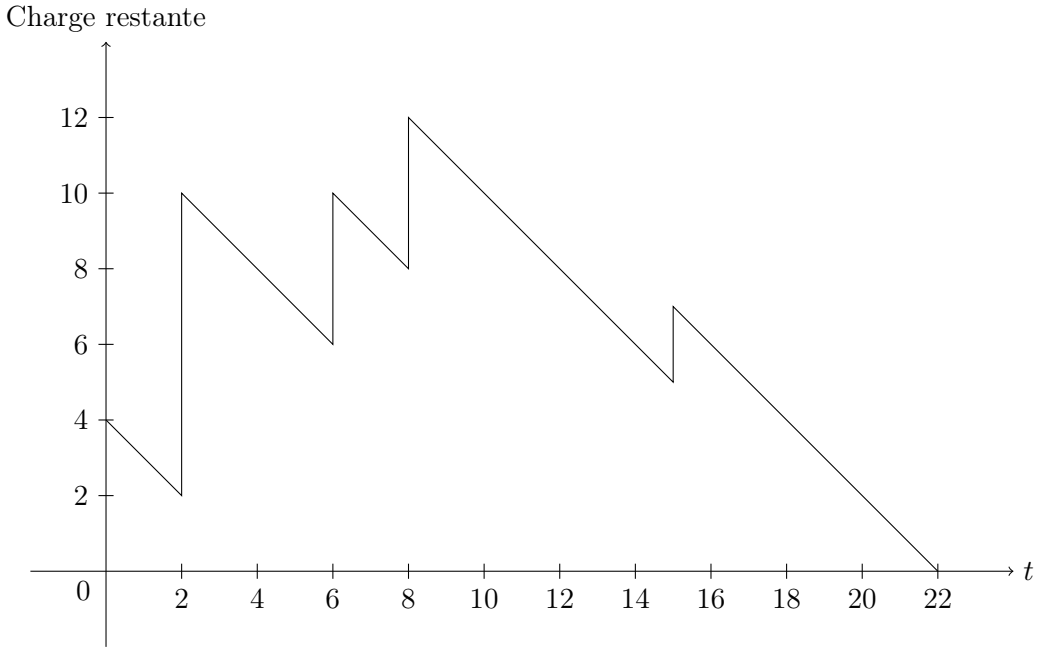

FIGURE 12 – Courbe de charge en mode FIFO

On définit la charge du système à l'instant t comme la quantité de travail qui reste à effectuer par le système à cet instant.

2.4 Etude de la stabilité d'un système

Le nombre moyen de client est donné par :

$$E(N) = \lim_{T\to\infty} \frac{1}{T} \int_0^T N(t)dt$$

On a aussi vu que le temps de réponse moyen est donné par :

$$\bar{r} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} r_i$$

On peut alors écrire le théorème de Little :
Pour tout système stable on a :

$$\bar{N} = \lambda\bar{r} \quad \text{où } \lambda \text{ est le débit et } \bar{r} \text{ est le temps de réponse moyen.}$$

Si le serveur est occupé avec une robabilité de u l'utilisation moyenne (dans le cas d'un système ON/OFF), la formule de Little devient :

$$\bar{u} = \lambda\bar{r}$$

Comme $\bar{u} \in [0, 1]$, le débit ne peut être supérieur à $\frac{1}{\bar{r}}$, on appelle alors cette valeur, la capacité de traitement et on la note μ.
Si $\lambda < \mu$ on a :
- système stable
- le débit de sortie est exactement λ
- le taux d'utilisation est alors : $u = \frac{\lambda}{\mu}$

Si $\lambda > \mu$ on a :
- le système est instable
- $N(t) \underset{t\to\infty}{\longrightarrow} \infty$
- le débit de sortie est μ
- $\bar{u} = 1$

2.5 Etudes des files M/M/1

Ce genre de file permet de modéliser des serveurs ftp et web. Commençons par en énoncer les caractéristiques :
- arrivées poissonniennes
- le temps entre deux arrivées suit une loi exponentielle $(1 - e^{-\lambda t})$
- services poissonniens
- le temps de service suit une loi exponentielle $(1 - e^{-\mu t})$
- un seul serveur

6. LIFO et Priorités sont préemptifs (i.e. si un client est servi et qu'un autre arrive avec une priorité supérieure, le premier client sort de la file d'attente)

7. Cette loi ressemble à un développement limité d'exponentielle, les lois exponentielles sont en général utilisées pour les évènements sans mémoire : l'évènement reçu à un temps t est indépendant du précédent et n'influera pas sur le suivant

8. On peut par exemple considérer la date d'arrivée du n^e paquet dans un réseau

– traitement FIFO

Le débit moyen est λ, le temps de service moyen est $\frac{1}{\mu}$ et la capacité de traitement est μ. D'après la formule de Little, le système est stable si $\lambda < \mu$. On cherche maintenant à calculer le nombre moyen de clients.

Pour ce faire, notons P_i la probabilité d'avoir i clients dans le système. Soit la chronologie suivante :

1. J'ai 0 client
2. 1 client arrive avec un débit de λ
3. 1 client part avec un débit μ

D'un point de vue stationnaire on a :
– le nombre de clients qui arrivent est λP_0
– le nombre de clients qui partent est μP_1
Généralisons à n clients :
L'état stationnaire du noeud k nous donne (par conservation du débit), l'égalité suivante :

$$\lambda P_k + \mu P_k = \lambda P_{k-1} + \mu P_{k+1}$$

Sur l'ensemble du système on peut écrire :

$$\begin{cases} \lambda P_0 &= \mu P_1 \\ \vdots &= \vdots \\ \lambda P_{k-1} + \mu P_{k-1} &= \lambda P_{k-2} + \mu Pk \\ \lambda P_k + \mu P_k &= \lambda P_{k-1} + \mu P_{k+1} \\ \vdots &= \vdots \\ \mu P_n &= \lambda P_{n-1} \end{cases}$$

Par l'énoncé on a :

$$P_n = \left(\frac{\lambda}{\mu}\right)^n . P_0$$

Preuve : Procédons par récurrence. De par la première ligne, on a :

$$P_1 = \frac{\lambda}{\mu} P_0$$

Et de par la seconde ligne on a :

$$\begin{aligned} \lambda P_1 + \mu P_1 &= \lambda P_0 + \mu P_2 \\ \Rightarrow \quad \tfrac{\lambda^2}{\mu} P_1 + \lambda P_0 - \lambda P_0 &= \mu P_2 \\ \Rightarrow \quad P_2 &= \left(\frac{\lambda}{\mu}\right)^2 \end{aligned}$$

Supposons la propriété vraie au rang $k-1$ et au rang k et intéressons nous au rang $k+1$, $\forall k \in [1; n-1]$. :

$$\begin{aligned} \lambda P_k + \mu P_k &= \lambda P_{k-1} + \mu P_{k+1} \\ \Rightarrow \quad P_0 \left(\frac{\lambda}{\mu}\right)^k \left[\lambda + \mu - \lambda \left(\frac{\mu}{\lambda}\right)\right] &= \mu P_{k+1} \\ \Rightarrow \quad P_0 \lambda \left(\frac{\lambda}{\mu}\right)^k &= \mu P_{k+1} \\ \Rightarrow \quad P_{k+1} &= \left(\frac{\lambda}{\mu}\right)^{k+1} \end{aligned}$$

On définit de plus la valeur :

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

Par définition des probabilités, on a :

$$\begin{aligned} \sum_{k=1}^{n} P_k &= 1 \\ \Longrightarrow \quad P_0 \sum_{k=1}^{n} \rho^k &= 1 \\ \Longrightarrow \quad P_0 . \frac{1}{1-\rho} &= 1 \end{aligned}$$

De la même manière on peut écrire P_n sous la forme suivante :

$$P_n = (1-\rho).\rho$$

On peut alors calculer le nombre moyen de clients :

$$\begin{aligned} E(N) &= \sum_{i=0}^{n-1} i P_i \\ &= \sum_{i=0}^{n-1} i(1-\rho)\rho^i \\ &= \frac{\rho}{1-\rho} \end{aligned}$$

Par la formule de Little, on a :

$$\bar{r} = \frac{\bar{N}}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda}$$