Modélisation d'une file d'attente

Notions théoriques :

- variables aléatoires, espérance mathématique ;
- loi de Poisson, loi géométrique, loi exponentielle, loi d'Erlang ;
- processus de Poisson, processus de naissance-mort.

Table des matières

Position du problème

Les files d'attente sont aujourd'hui des phénomènes que l'on rencontre quotidiennement dans de très nombreux domaines et sous diverses formes. Citons quelques exemples parmi tant d'autres : queue à un guichet, saturation d'un trafic routier, d'un réseau de télécommunications, gestion d'un stock de production, maintenance d'un équipement informatique, mouvements de populations, prévisions météorologiques, etc. De nombreux modèles stochastiques existent, reposant sur certaines hypothèses adaptées au contexte en question. Le modèle le plus célèbre que nous allons étudier ci-après, le plus simple et le plus utilisé de manière générale est un modèle markovien (file $M/M/1$) qui repose sur l'absence de mémoire de certaines occurrences.

D'innombrables questions se posent naturellement, afin d'optimiser la rentabilité de certains services, de diminuer les attentes des différentes parties concernées ainsi que les coûts associés s'il y a des dépenses de fonds. Par exemple, quel est le temps passé par un client dans une file d'attente devant un guichet, quelle est la longueur de la queue à un instant donné, quelle est la durée de repos du serveur (c'est-à-dire lorsque la queue est vide avant l'arrivée d'un nouveau consommateur), à quelle vitesse minimale devrait travailler le serveur pour ne pas dépasser un seuil maximal de clients, combien de serveurs faudrait-il au minimum pour éviter la saturation de la salle d'attente... ? Autant de questions auxquelles la théorie des chaînes de Markov [1] et des processus stochastiques apporte des réponses plus ou moins explicites selon le modèle adopté.

1 Constitution d'une file d'attente

Un système d'attente comporte plusieurs caractéristiques. Typiquement, une file est composée de clients se succédant et demandant un service (voir Fig. 1). Les clients peuvent être des individus, des appels téléphoniques, des signaux électriques, des véhicules, des accidents, des turbulences atmosphériques... et le service peut être un serveur humain, un central téléphonique, un serveur informatique, un péage autoroutier, une compagnie d'assurance, la météorologie nationale...

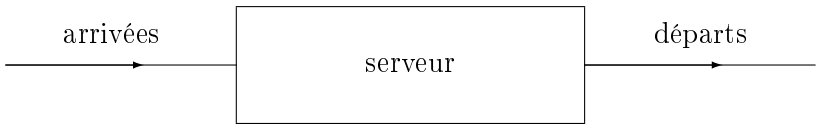


FIGURE 1 – Système d'attente

1.1 Flux d'arrivées

Les arrivées peuvent être régulières (déterministes) ou complètement aléatoires, individuelles ou groupées, provenir de populations différentes ou se répartir en plusieurs files. On devra modéliser les temps inter-arrivées. Dans certaines situations, on devra tenir compte de l'effectif de la population susceptible de se présenter dans le système. Si cette population n'est pas infinie, la quantité d'individus entrant dans le système diminue avec le temps. Ce facteur est pris en compte notamment dans l'étude de certains modèles d'émigration-immigration.

1.2 Organe de service

Le service peut être constitué d'un ou plusieurs serveurs, qui peuvent être disposés de diverses façons :

- *serveurs en parallèle*
  cette disposition concerne des files où le client a le choix du serveur : files de personnes en attente dans une administration offrant plusieurs services, files de consommateurs en attente aux caisses de paiement dans un hypermarché, files de voitures se présentant à un péage d'autoroute... Voir Fig. 2 (a) ;

1. Markov, Andreï Andreïevitch : mathématicien russe (Riazan 1856 – Pétrograd 1922).

– *serveurs en série*
cette disposition concerne des services à la chaîne : service de restauration, service des cartes grises dans une préfecture (nécessitant deux temps : enregistrement puis confection de cartes), visite médicale dans une infirmerie (nécessitant plusieurs contrôles successifs), chaînes de production avec contrôle de qualité... Voir Fig. 2 (b) ;
– *serveurs en réseau*
cette disposition est beaucoup plus complexe et réaliste : centraux de télécommunications, réseaux informatiques, internet... Voir Fig. 3.

On devra par ailleurs modéliser les temps de service.

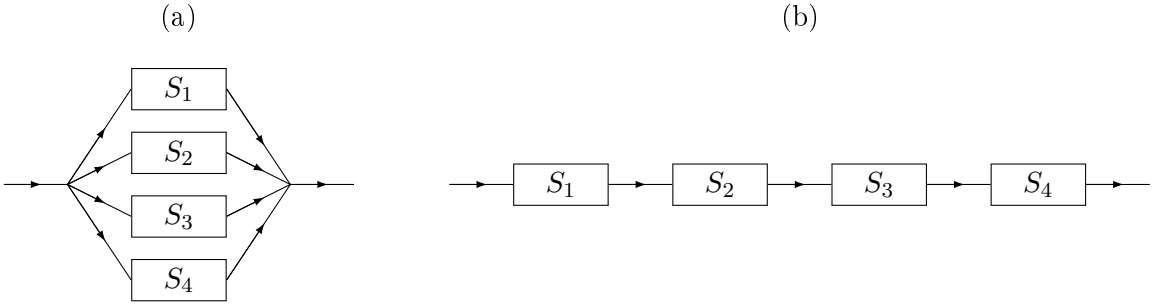


FIGURE 2 – (a) Serveurs en parallèle. (b) Serveurs en série

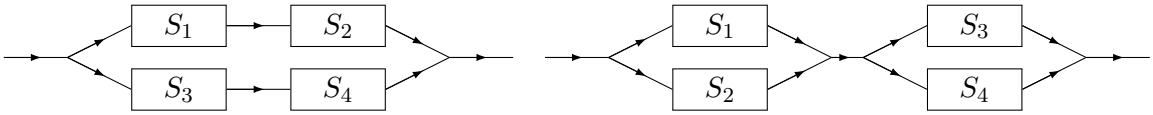


FIGURE 3 – Serveurs en réseaux

1.3 Discipline de service

La discipline de service indique dans quel ordre sont traités le clients. Un certain nombre de règles courantes sont adoptées.

– Une règle de courtoisie voudrait que l'on serve les personnes en respectant leur ordre d'arrivée, c'est la discipline FCFS (first come, first served). Une règle proche de cette dernière est la discipline FIFO (first in, first out) qui lui est équivalente dans le cas ou le service est effectué par un unique serveur.
– Une règle opposée à la règle de courtoisie consiste à servir en premier le dernier client arrivé. On rencontre cette situation lors de la gestion d'un stock : un magasinier déballe des cartons empilés sur une palette en démarrant du haut de la pile, le carton du haut étant le dernier arrivé. Le carton au bas de la pile sera déballé lorsque tous les autres au-dessus de lui (empilés après lui) auront été déballés. C'est la discipline LCFS (last come, first served). Une règle proche de la précédente est la discipline LIFO (last in, first out).
– Dans certains services d'urgence, des règles de priorités s'imposent : les clients se répartissent en plusieurs classes avec des ordres de priorité différents. Une personne prioritaire devra être servie avant une personne non prioritaire, même si celle-là est arrivée avant. Dans le cas où une personne prioritaire arriverait pendant le service d'une personne non prioritaire, on peut envisager plusieurs situations :

1. soit le service de la personne non prioritaire continue jusqu'à son terme avant le traitement de la personne prioritaire. C'est par exemple le cas des services d'urgence médicale : lorsqu'une personne est en train de subir une intervention chirurgicale, il faut terminer celle-là avant de traiter une autre urgence, même prioritaire ;
2. soit le service de la personne non prioritaire est interrompu par l'arrivée d'une personne prioritaire (jouissant d'un droit de « préemption »). Dans ce cas, à la fin du service de la personne prioritaire, le service de la personne non prioritaire reprend, s'il n'y a plus de personne prioritaire dans la file,
   (a) soit en l'état tel qu'il était à l'instant d'interruption. C'est le cas par exemple de traitement de tâches informatiques : lorsque certaines doivent être traitées d'ur-

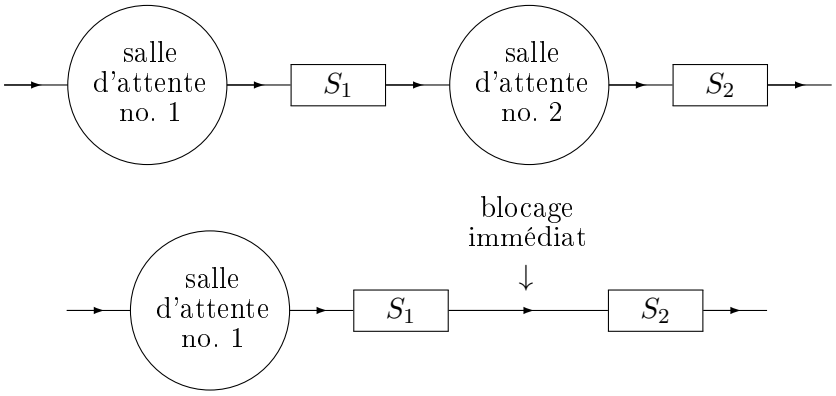


FIGURE 4 – Files en tandem avec salle d'attente intermédiaire ou non

gence, les moins urgentes restent en mémoire (s'il y a suffisamment de mémoire) et seront achevées ultérieurement ;

(b) soit depuis le début. C'est le cas par exemple de certaines réservations qui sont prioritaires pour des personnes handicapées ; lorsque ces réservations sont terminées, il faut reprendre celle qui était en cours depuis le début puisque les places disponibles ont diminués entre temps.

– Enfin, dans d'autres situations, on adopte le partage de service (processor sharing) : on rencontre ce cas notamment en informatique, cas dans lequel différentes tâches sont traitées simultanément.

1.4 Capacité du système

De nombreux systèmes d'attente comportent une salle d'attente à capacité limitée. Il y aura donc refoulement de personnes à l'entrée du système lorsque cette salle est pleine. Ce facteur a une importance notamment dans l'étude de files en tandem. Si par exemple une salle d'attente à capacité limitée est insérée entre deux services consécutifs, lorsque cette salle arrive à saturation, un blocage se produit au niveau du premier service.

On va maintenant étudier en détail une file simple provenant d'une population infinie avec un serveur, une discipline FCFS (ou FIFO) et une salle d'attente de capacité infinie (donc sans limitation au niveau des arrivées). On parle de file $M/M/1$.

2 Modélisation des arrivées

2.1 Processus des arrivées

Reprenons l'exemple des personnes se présentant à un guichet. On fera trois hypothèses :
– les arrivées sont aléatoires et les laps de temps inter-arrivées ont même loi de probabilité ;
– les arrivées sur des intervalles de temps disjoints sont indépendantes ;
– il n'y a pas d'arrivées simultanées, i.e. il n'arrive pas plus d'un client à la fois.

Divisons l'échelle du temps en sous-intervalles $[0, \Delta t[, [\Delta t, 2\Delta t[, \ldots, [(n-1)\Delta t, n\Delta t[, \ldots$ de longueur Δt très petite de telle sorte que pas plus d'une personne n'arrive dans chaque laps de temps $[(n-1)\Delta t, n\Delta t[$ (voir Fig. 5).

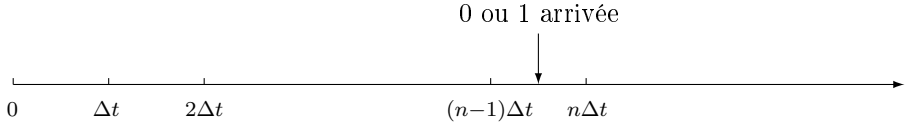


FIGURE 5 – Subdivision de l'échelle du temps

Notons X_n le nombre (aléatoire) d'arrivées durant l'intervalle de temps $[(n-1)\Delta t, n\Delta t[$. D'après ce qui précède, X_n est une v.a. prenant essentiellement les deux valeurs 0 et 1. Elle suit approximativement une loi de Bernoulli, elle est donc caractérisée par un paramètre $p_{\Delta t}$ qui n'est

autre que son espérance : $\mathbb{E}(X_n) \approx p_{\Delta t}$. Il est raisonnable de supposer que cette espérance est proportionnelle à la longueur de l'intervalle de temps $[(n-1)\Delta t, n\Delta t[$: $p_{\Delta t} = \lambda \Delta t$. On a plus précisément, lorsque $\Delta t \to 0^+$,

$$\begin{cases} \mathbb{P}(X_n = 1) = \lambda \Delta t + o(\Delta t), \\ \mathbb{P}(X_n = 0) = 1 - \lambda \Delta t + o(\Delta t), \\ \mathbb{P}(X_n \geqslant 2) = o(\Delta t). \end{cases}$$

On modélise le processus des arrivées par une fonction aléatoire (processus stochastique) croissante $t \in \mathbb{R}^+ \longmapsto A_t$. Ici, A_t représente le nombre (aléatoire) de consommateurs entrés dans le système pendant le laps de temps $[0, t]$. L'étude faite ci-dessus montre que $A_{n\Delta t} = X_1 + \cdots + X_n$ et alors la v.a. $A_{n\Delta t}$ suit approximativement la loi binomiale $\mathcal{B}(n, p_{\Delta t})$. Pour un instant t donné, t est dans un intervalle $[(n-1)\Delta t, n\Delta t[$ et

$$\mathbb{P}(A_t = k) \approx p_{k,\Delta t}(t) = C_n^k p_{\Delta t}^k (1 - p_{\Delta t})^{n-k} + o(\Delta t).$$

Faisons tendre Δt vers 0^+ ou encore n vers $+\infty$ avec $n\Delta t \sim t$ (fixé), donc $p_{\Delta t} \sim \lambda t/n$. En écrivant $p_{k,\Delta t}(t) = \dfrac{n!}{(n-k)!\, n^k} \dfrac{(\lambda t)^k}{k!} \left(1 - \dfrac{\lambda t}{n}\right)^{n-k} + o(\Delta t)$, on obtient la valeur de la limite $\lim\limits_{\Delta t \to 0^+} p_{k,\Delta t}(t)$:

$$\boxed{\mathbb{P}(A_t = k) = \frac{(\lambda t)^k}{k!}\, e^{-\lambda t}.}$$

C'est la loi de Poisson [2] $\mathcal{P}(\lambda t)$ (voir l'annexe 2). On dit que $(A_t)_{t \in \mathbb{R}^+}$ est un processus de Poisson d'intensité λ. On a $\mathbb{E}(A_t) = \lambda t$, ce qui fournit pour λ l'interprétation suivante :

$$\boxed{\lambda = \frac{\mathbb{E}(A_t)}{t}.}$$

Ainsi, le paramètre λ représente le nombre moyen d'arrivées par unité de temps (taux d'arrivée).

2.2 Temps inter-arrivées

Soit $T_1 = \inf\{t \geqslant 0 : A_t \geqslant 1\}$ l'instant d'arrivée (aléatoire) de la première personne. La condition $T_1 > t$ signifie que la première personne arrive après l'instant t et donc que $A_t = 0$. En conséquence,

$$\mathbb{P}(T_1 > t) = \mathbb{P}(A_t = 0) = e^{-\lambda t}$$

d'où la fonction de répartition de T_1 :

$$F_{T_1}(t) = \mathbb{P}(T_1 \leqslant t) = 1 - e^{-\lambda t}.$$

Ainsi, la v.a. T_1 suit la loi exponentielle $\mathcal{E}(\lambda)$ (voir l'annexe 2). Elle a pour densité

$$\boxed{f_{T_1}(t) = \lambda\, e^{-\lambda t}.}$$

Son espérance vaut $\mathbb{E}(T_1) = 1/\lambda$, ce qui donne pour λ une autre interprétation :

$$\boxed{\lambda = \frac{1}{\mathbb{E}(T_1)}.}$$

En fait, la loi de T_1 mesure aussi l'intervalle de temps séparant deux arrivées consécutives : si T_n est le laps de temps s'écoulant entre les $(n-1)^{\text{e}}$ et n^{e} personnes, on peut démontrer que les v.a. $T_1, T_2, \ldots, T_n, \ldots$ sont indépendantes de loi commune la loi $\mathcal{E}(\lambda)$.

2. Poisson, Siméon Denis : mathématicien français (Pithiviers 1781 – Paris 1840)

2.3 Temps d'arrivée de la n^{e} personne

Soit $s_n = \inf\{t \geqslant 0 : A_t = n\}$ l'instant d'arrivée de la n^{e} personne. La condition $s_n > t$ signifie que la n^{e} personne arrive après l'instant t ; il y a donc eu moins de n arrivées durant l'intervalle de temps $[0, t]$. En d'autres termes on a

$$\mathbb{P}(s_n > t) = \mathbb{P}(A_t < n) = \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!}\, e^{-\lambda t}$$

d'où l'on déduit la fonction de répartition de s_n :

$$F_{s_n}(t) = \mathbb{P}(s_n \leqslant t) = 1 - \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!}\, e^{-\lambda t}.$$

La densité de s_n s'obtient en dérivant l'expression ci-dessus :

$$f_{s_n}(t) = F'_{s_n}(t) = \lambda\, e^{-\lambda t} - \sum_{k=1}^{n-1} \left[\frac{\lambda^k\, t^{k-1}}{(k-1)!} - \frac{\lambda^{k+1}\, t^k}{k!} \right] e^{-\lambda t} = -\sum_{k=0}^{n-2} \frac{\lambda^{k+1}\, t^k}{k!}\, e^{-\lambda t} + \sum_{k=0}^{n-1} \frac{\lambda^{k+1}\, t^k}{k!}\, e^{-\lambda t}.$$

Il reste finalement

$$\boxed{f_{s_n}(t) = \frac{\lambda^n\, t^{n-1}}{(n-1)!}\, e^{-\lambda t}.}$$

C'est la célèbre loi d'Erlang[3] $E(n; \lambda)$ (voir l'annexe 2). Bien sûr, s_n est la somme des n premiers temps inter-arrivées : $s_n = T_1 + T_2 + \cdots + T_n$ où $T_k = s_k - s_{k-1}$ (voir Fig. 6), les v.a. $T_1, T_2, \ldots, T_n$ étant indépendantes de même loi $\mathcal{E}(\lambda)$ comme cela a été signalé dans le paragraphe précédent.

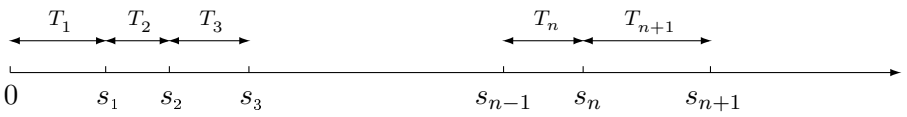


FIGURE 6 – Arrivées

3 Modélisation du temps de service

Soit S_n la durée de service de la n^{e} personne. Un modèle très répandu, reposant sur l'absence de mémoire, consiste à stipuler que $\mathbb{P}(S_n > s + t \mid S_n > s) = \mathbb{P}(S_n > t)$ (modèle markovien ; voir l'annexe 1 pour la définition d'une probabilité conditionnelle) : sachant qu'à un instant donné le service a démarré depuis une durée s, le temps d'achèvement de service est le même que s'il débutait à cet instant. Alors

$$\mathbb{P}(S_n > s + t) = \mathbb{P}(S_n > s)\mathbb{P}(S_n > s + t \mid S_n > s) = \mathbb{P}(S_n > s)\mathbb{P}(S_n > t).$$

La fonction φ définie par $\varphi(t) = \mathbb{P}(S_n > t)$ vérifie ainsi l'équation fonctionnelle $\varphi(s + t) = \varphi(s)\varphi(t)$. De plus, cette fonction est bornée ($0 \leqslant \varphi(t) \leqslant 1$) et vérifie $\varphi(0) = 1$. En rajoutant l'hypothèse simplificatrice (non nécessaire *a priori*) que φ est dérivable, on obtient, en dérivant par rapport à s :

$$\varphi'(s + t) = \varphi'(s)\varphi(t)$$

puis pour $s = 0$, en posant $\varphi'(0) = a$:

$$\varphi'(t) = a\varphi(t).$$

Cette équation différentielle, avec la condition initiale $\varphi(0) = 0$, admet pour solution $\varphi(t) = e^{at}$. Enfin, la fonction φ recherchée doit être bornée, ceci entraîne donc

$$\boxed{\mathbb{P}(S_n > t) = e^{-\mu t}}$$

où la constante $\mu = \dfrac{1}{\mathbb{E}(S_n)}$ est l'inverse du temps moyen de service. Ainsi, la v.a. S_n suit la loi exponentielle $\mathcal{E}(\mu)$. Le paramètre μ pourrait s'interpréter comme représentant le nombre moyen de services (taux de service) que le serveur peut effectuer par unité de temps.

3. Erlang, Agner Krarup : mathématicien danois (Lønborg 1878 – Copenhagen 1929)

4 Modélisation de la longueur de la queue

Introduisons D_t le nombre (aléatoire) de clients sortis du système pendant le laps de temps $[0,t]$; $(D_t)_{t\in\mathbb{R}^+}$ est le processus des départs. Notons alors Q_t la longueur de la file à l'instant t, c'est-à-dire le nombre de personnes présentes dans le système (en attente ou en service) ; on a bien sûr :

$$\boxed{Q_t = A_t - D_t}$$

(longueur = nombre de personnes entrées − nombre de personnes sorties). On dit que le processus $(Q_t)_{t\in\mathbb{R}^+}$ est un processus de naissance-mort de taux de naissance λ et de taux de mort μ.

Pour pouvoir représenter la courbe de départs $t \longmapsto D_t$, il est nécessaire d'évaluer l'instant de sortie de chaque client, lequel dépend à la fois de l'instant d'entrée de ce client, son temps d'attente et son temps de service. Notons W_n le temps d'attente du n^{e} client présent dans la file et σ_n son instant de sortie. Le temps W_n est donné par la relation de récurrence suivante (formule de Lindley, 1955) :

$$\boxed{W_{n+1} = \max(W_n + S_n - T_{n+1}, 0)}$$

et alors

$$\boxed{\sigma_n = s_n + W_n + S_n.}$$

Les graphes des fonctions d'arrivées $t \longmapsto A_t$, de départs $t \longmapsto D_t$ sont des courbes en escaliers croissantes avec des sauts de 1 et le graphe de la longueur de la file $t \longmapsto Q_t$ est une courbe en escaliers avec des sauts de ± 1. Ils sont représentés ci-dessous (Fig. 7 et 8).

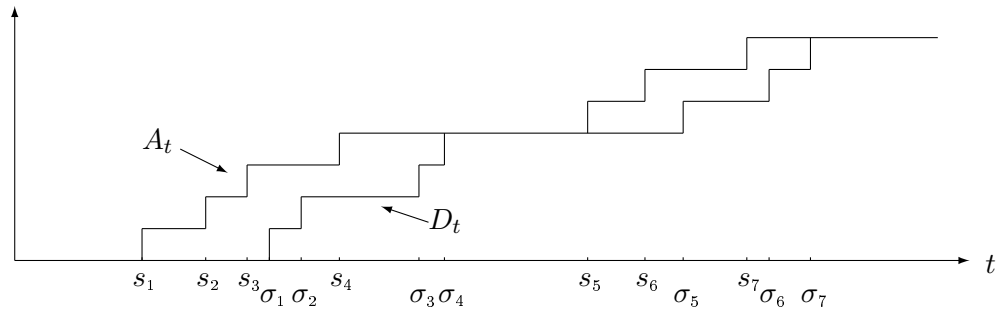


FIGURE 7 – Courbes d'arrivées-départs

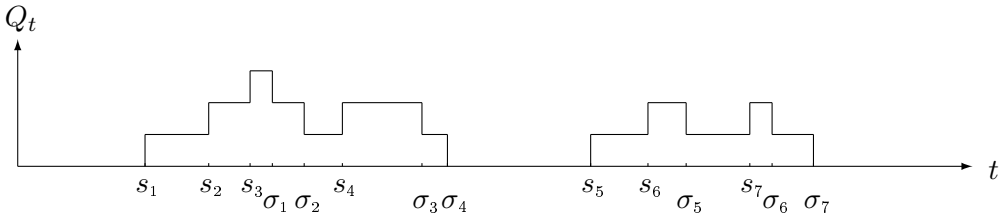


FIGURE 8 – Longueur de la queue

La formule de Lindley se démontre facilement en interprétant les laps de temps W_n, W_{n+1}, S_n, T_{n+1} comme les aires de rectangles de base les temps précédents et de hauteur 1. D'après la figure 9, il est clair que si la $(n+1)^{\text{e}}$ personne arrive avant le départ de la n^{e}, i.e. $s_{n+1} < \sigma_n$ ou encore $T_{n+1} < W_n + S_n$, alors $W_n + S_n = T_{n+1} + W_{n+1}$. Dans le cas ou la $(n+1)^{\text{e}}$ personne arrive après le départ de la n^{e}, la $(n+1)^{\text{e}}$ n'a pas d'attente : $W_{n+1} = 0$.

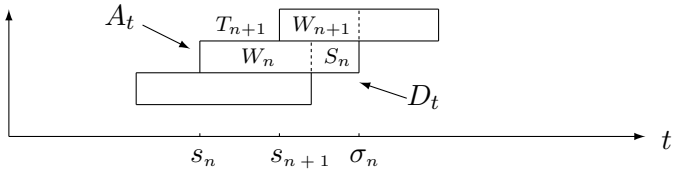


FIGURE 9 – Formule de Lindley

5 Étude de la file en régime stationnaire

On étudie maintenant le système en temps grand. Les consommateurs arrivent selon un processus de Poisson d'intensité λ et le service s'effectue selon la loi $\mathcal{E}(\mu)$ avec la discipline FCFS. Lorsque $\lambda < \mu$, le système tend vers un état stationnaire c'est-à-dire indépendant du temps. On suppose cette condition satisfaite dans toute la suite.

5.1 Loi de la longueur de la queue

Soit Q_∞ la longueur « limite » de la queue, et $\pi_n = \mathbb{P}(Q_\infty = n)$. On comptabilise les arrivées en l'état $Q_\infty = n$ ainsi que les départs depuis cet état. Les états voisins de l'état $Q_\infty = n$ sont les états $Q_\infty = n-1$ et $Q_\infty = n+1$ (voir Fig. 10).

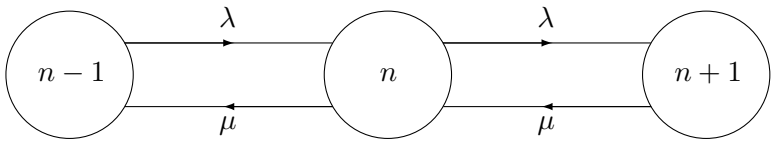


FIGURE 10 – Changements d'états

(i) Flux entrant en l'état n :
- soit la file contient $n-1$ personnes (avec une probabilité π_{n-1}) et il en arrive une de plus au taux λ ;
- soit la file contient $n+1$ personnes (avec une probabilité π_{n+1}) et il en part une au taux μ.

Ceci conduit au schéma suivant : $n-1 \xrightarrow{\lambda} n \xleftarrow{\mu} n+1$. Le taux entrant en l'état n est donc $\lambda\pi_{n-1} + \mu\pi_{n+1}$.

(ii) Flux sortant de l'état n : la file contient n personnes avec une probabilité π_n ;
- soit il en arrive une de plus au taux λ et la longueur de la file devient $n+1$;
- soit il en part une au taux μ et la longueur de la file devient $n-1$.

Le schéma est cette fois le suivant : $n-1 \xleftarrow{\mu} n \xrightarrow{\lambda} n+1$. Le taux sortant de l'état n est donc $(\lambda+\mu)\pi_n$.

Ainsi en égalant les flux entrant et sortant, on obtient les équations d'équilibre suivantes (équations de balance globale) :

$$\begin{cases} \lambda\,\pi_0 = \mu\,\pi_1 \\ \lambda\,\pi_{n-1} + \mu\,\pi_{n+1} = (\lambda+\mu)\,\pi_n \ \text{ si } n \geqslant 1. \end{cases}$$

Pour résoudre ce système, on pose $\rho = \dfrac{\lambda}{\mu}$; on a par hypothèse $\rho \in]0,1[$. Le paramètre ρ est appelé intensité du trafic (ou encore charge du système). On a affaire à une suite définie par une relation de récurrence linéaire à trois indices. La recherche de suites géométriques particulières conduit à l'équation caractéristique

$$\mu\,r^2 - (\lambda+\mu)\,r + \lambda = 0$$

dont les solutions sont $\dfrac{\lambda}{\mu} = \rho$ et 1. La forme générale des suites vérifiant la relation de récurrence ci-dessus est alors

$$\pi_n = \alpha\rho^n + \beta.$$

La condition initiale $\lambda\,\pi_0 = \mu\,\pi_1$ donne $\lambda\alpha + \lambda\beta = \lambda\alpha + \mu\beta$, soit encore $\beta = 0$ et donc $\pi_n = \alpha\,\rho^n$ et $\alpha = \pi_0$. De plus, la suite $(\pi_n)_{n\in\mathbb{N}}$ doit être une probabilité : $\displaystyle\sum_{n=0}^{+\infty} \pi_n = 1$, ce qui fournit au passage, grâce à la relation $\displaystyle\sum_{n=0}^{\infty} \rho^n = \frac{1}{1-\rho}$, la probabilité de trouver la file vide :

$$\boxed{\pi_0 = \mathbb{P}(Q_\infty = 0) = 1 - \rho.}$$

Cette probabilité est non nulle. La file connaît des oscillations qui se reproduisent de manière similaire au cours du temps, on parle de file récurrente. La solution de l'équation de balance est donc

$$\boxed{\pi_n = \mathbb{P}(Q_\infty = n) = (1-\rho)\,\rho^n, \quad n \in \mathbb{N}.}$$

La v.a. $Q_\infty + 1$ suit la loi géométrique $\mathcal{G}(1-\rho)$ (voir l'annexe 2) et la longueur moyenne de la queue est

$$\boxed{\mathbb{E}(Q_\infty) = \frac{\lambda}{\mu - \lambda}.}$$

Remarque.

- Lorsque $\lambda > \mu$, on montre que Q_t devient infini en temps infini ; le flux des arrivées est plus important que celui des sorties et le système est saturé. On parle de file transitoire (Fig. 11).

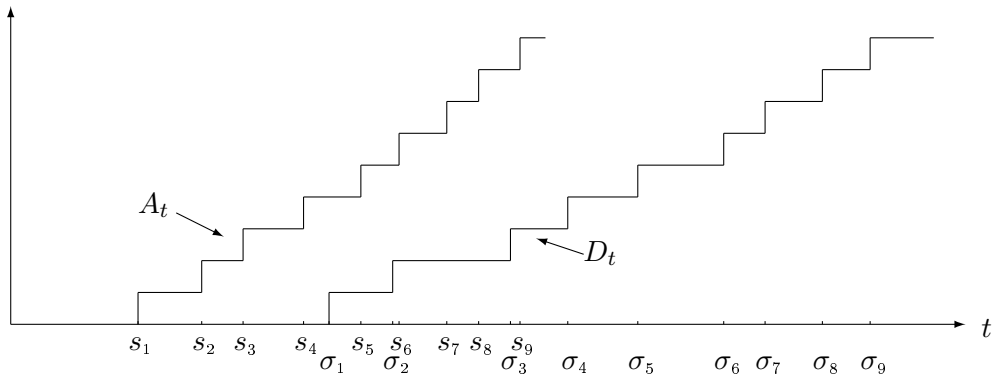


Figure 11 – Courbes d'arrivées-départs : cas transitoire

- Lorsque $\lambda = \mu$ (cas critique), Q_t n'a pas de limite ; on observe dans ce cas des oscillations d'amplitude non bornées.

5.2 Lois des temps d'attente et temps de séjour d'un client

Il est possible de déterminer la loi de probabilité du temps d'attente W_∞ (waiting time) d'un client générique en régime stationnaire grâce à la formule des probabilités totales (voir l'annexe 1) :

$$\mathbb{P}(W_\infty \leqslant t) = \sum_{n=0}^{\infty} \mathbb{P}(W_\infty \leqslant t \mid Q_\infty = n)\,\mathbb{P}(Q_\infty = n).$$

Or la v.a. conditionnelle $(W_\infty \leqslant t \mid Q_\infty = n)$ représente l'attente pendant laquelle n personnes consécutives doivent être servies selon un service exponentiel $\mathcal{E}(\mu)$, c'est donc la somme de n v.a. indépendantes de loi $\mathcal{E}(\mu)$. Si $n = 0$, le temps d'attente est nul. Si $n \geqslant 1$, cette v.a. conditionnelle suit une loi d'Erlang $E(n;\mu)$. D'où

$$\begin{aligned}
\mathbb{P}(W_\infty \leqslant t) &= \pi_0 + \sum_{n=1}^{\infty} \pi_n \int_0^t \frac{\mu^n s^{n-1}}{(n-1)!}\, e^{-\mu s}\, ds \\
&= (1-\rho)\Big[1 + \rho\mu \int_0^t \sum_{n=0}^{\infty} \frac{(\rho\mu s)^n}{n!}\, e^{-\mu s}\, ds\Big] \\
&= (1-\rho)\Big[1 + \lambda \int_0^t e^{-\mu(1-\rho)s}\, ds\Big]
\end{aligned}$$

soit

$$\boxed{\mathbb{P}(W_\infty \leqslant t) = 1 - \frac{\lambda}{\mu}\, e^{-(\mu-\lambda)t}.}$$

C'est une loi exponentielle $\mathcal{E}(\mu - \lambda)$ avec un poids en 0 : $\mathbb{P}(W_\infty = 0) = 1 - \rho$. Notons l'égalité $\mathbb{P}(W_\infty = 0) = \mathbb{P}(Q_\infty = 0)$ qui traduit le fait que la probabilité de ne pas attendre coïncide

avec celle de trouver une file vide. L'attente moyenne d'un client peut se calculer selon le même procédé :

$$\mathbb{E}(W_\infty) = \sum_{n=0}^{\infty} \mathbb{E}(W_\infty \mid Q_\infty = n)\, \mathbb{P}(Q_\infty = n).$$

Le raisonnement précédent montre que $\mathbb{E}(W_\infty \mid Q_\infty = n) = \dfrac{n}{\mu}$ et donc

$$\mathbb{E}(W_\infty) = \sum_{n=0}^{\infty} \frac{n}{\mu}\, \mathbb{P}(Q_\infty = n) = \frac{1}{\mu}\, \mathbb{E}(Q_\infty),$$

soit

$$\boxed{\mathbb{E}(W_\infty) = \frac{\lambda}{\mu(\mu - \lambda)}.}$$

Enfin, l'expression $W_\infty + S_\infty$ représente le temps de séjour total (attente+service) du client générique. Sa loi s'obtient comme précédemment, c'est à présent une véritable loi exponentielle $\mathcal{E}(\mu - \lambda)$:

$$\boxed{\mathbb{P}(W_\infty + S_\infty \leqslant t) = 1 - e^{-(\mu-\lambda)t}.}$$

Le temps de séjour moyen d'un client dans le système vaut donc

$$\boxed{\mathbb{E}(W_\infty + S_\infty) = \frac{1}{\mu - \lambda}.}$$

Ces différentes quantités sont reliées par la célèbre formule de Little (1961) :

$$\boxed{\mathbb{E}(Q_\infty) = \mu\, \mathbb{E}(W_\infty) = \lambda\, \mathbb{E}(W_\infty + S_\infty).}$$

Remarquons que la longueur de la file Q_∞ dépend du rapport $\rho = \dfrac{\lambda}{\mu}$ (qui est l'écart relatif entre λ et μ), alors que le temps de séjour W_∞ dépend de l'écart absolu $\mu - \lambda$.

Exemple.

Considérons deux files indépendantes de paramètres respectifs

$\lambda_1 = 4$ personnes/h, $\quad 1/\mu_1 = 10$ mn/personne $\quad$ (ou encore $\mu_1 = 6$ personnes/h),

$\lambda_2 = 8$ personnes/h, $\quad 1/\mu_2 = 5$ mn/personne $\quad$ (ou encore $\mu_2 = 12$ personnes/h).

Ces paramètres sont dans un rapport de 2. Les intensités des deux files coïncident :

$$\rho_1 = \rho_2 = \frac{2}{3};$$

il en est de même pour les longueurs moyennes :

$$\mathbb{E}(Q_{1,\infty}) = \mathbb{E}(Q_{2,\infty}) = 2 \text{ personnes}$$

alors que les temps d'attente et de séjour sont dans un rapport de 2 inversé :

$$\mathbb{E}(W_{1,\infty}) = 20 \text{ mn}, \quad \mathbb{E}(W_{2,\infty}) = 10 \text{ mn},$$

$$\mathbb{E}(W_{1,\infty} + S_{1,\infty}) = 30 \text{ mn}, \quad \mathbb{E}(W_{2,\infty} + S_{2,\infty}) = 15 \text{ mn}.$$

5.3 Loi du temps d'activité du serveur

Pour le serveur, les périodes suivantes sont importantes :

- une période d'activité est une période (aléatoire) durant laquelle il sert les clients de manière continuelle. Elle démarre à partir de l'arrivée d'un client entrant dans une file vide et cesse dès la fin du service du prochain client laissant derrière lui la file vide, puis une nouvelle période d'activité redémarre avec l'arrivée d'un autre client ;
- une période de vacance est une période (aléatoire) durant laquelle il n'a aucune personne à servir ;

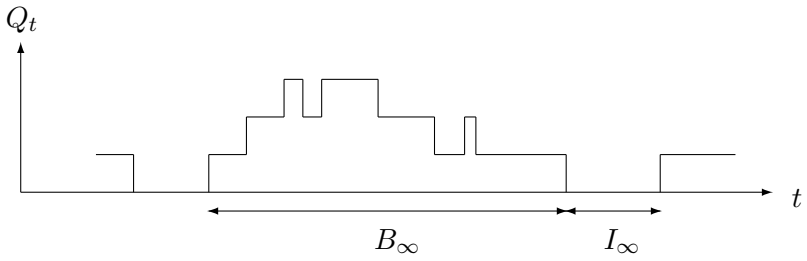


FIGURE 12 – Cycle d'activité

– un cycle d'activité est le laps de temps séparant deux personnes consécutives arrivant dans un système vide. Un tel cycle est donc la somme d'une période d'activité et d'une période de vacance du serveur (voir Fig. 12).

On peut démontrer que la loi du temps d'activité B_∞ (busy time) en régime stationnaire a pour densité

$$\boxed{f_{B_\infty}(t) = \sqrt{\frac{\mu}{\lambda}}\,\frac{1}{t}\,e^{-(\lambda+\mu)t}\,I_1(2t\sqrt{\lambda\mu}\,)}$$

où $I_1(z) = \displaystyle\sum_{n=0}^{+\infty} \frac{1}{(n+1)!\,n!}\left(\frac{z}{2}\right)^{2n+1}$ est une fonction de Bessel. Le temps d'activité moyen du serveur se calcule selon $\displaystyle\int_0^\infty t\,f_{B_\infty}(t)\,dt$, il est donné par

$$\boxed{\mathbb{E}(B_\infty) = \frac{1}{\mu-\lambda}.}$$

Ce résultat peut s'obtenir de manière empirique en multipliant le temps de service moyen d'un client par le nombre de clients présents dans la file lorsque cette dernière est non vide :

$$\mathbb{E}(B_\infty) = \mathbb{E}(Q_\infty \mid Q_\infty \geqslant 1) \times \mathbb{E}(S_\infty).$$

On a

$$\mathbb{E}(Q_\infty \mid Q_\infty \geqslant 1) = \frac{\mathbb{E}(Q_\infty)}{\mathbb{P}(Q_\infty \geqslant 1)} = \frac{\mu}{\mu-\lambda} \qquad \text{et} \qquad \mathbb{E}(S_\infty) = \frac{1}{\mu}$$

et l'on retrouve immédiatement le résultat énoncé.

Remarquons que le temps d'activité moyen du serveur est identique au temps de séjour moyen d'un client dans le système.

D'autre part, la propriété d'absence de mémoire de la loi exponentielle montre que le temps de vacance I_∞ du serveur (idle time), qui est la durée d'attente pour le serveur d'une nouvelle arrivée lorsque la file est vide, suit la loi $\mathcal{E}(\lambda)$:

$$\boxed{f_{I_\infty}(t) = \lambda\,e^{\lambda t}.}$$

5.4 Processus des départs

En régime stationnaire, il est possible de décrire le processus des départs des clients après avoir été servis. Introduisons le laps de temps τ_n séparant les départs des $(n-1)^{\text{e}}$ et n^{e} personnes : $\tau_n = \sigma_n - \sigma_{n-1}$. En remarquant que $Q_{\sigma_{n-1}}$ représente le nombre de clients que laisse le $(n-1)^{\text{e}}$ derrière lui en quittant le système,

– si $Q_{\sigma_{n-1}} \geqslant 1$, la n^{e} personne était en attente pendant le service de la précédente et se fait servir à partir de l'instant σ_{n-1} jusqu'à son instant de sortie σ_n pendant une durée S_n. Le temps τ_n n'est donc autre que S_n ;

– si $Q_{\sigma_{n-1}} = 0$, la $(n-1)^{\text{e}}$ personne laisse un système vide après son départ, le serveur entre dans une période de vacance jusqu'à l'arrivée de la n^{e} personne, d'une durée qu'on notera I_{n-1}. Cette nouvelle personne entrant dans un système vide sera immédiatement servie et ce pendant un temps S_n. On a dans ce cas $\tau_n = I_{n-1} + S_n$.

Ainsi (voir Fig. 13),

$$\tau_n = \begin{cases} S_n & \text{si } Q_{\sigma_{n-1}} \geqslant 1, \\ S_n + I_{n-1} & \text{si } Q_{\sigma_{n-1}} = 0. \end{cases}$$

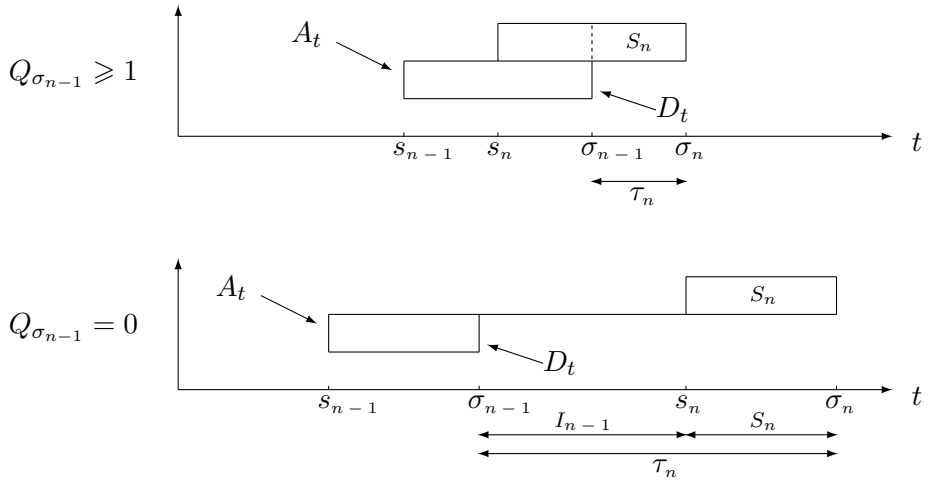


FIGURE 13 – Laps de temps inter-départs

D'après la formule des probabilités totales,

$$\mathbb{P}(\tau_n \leqslant t) = \mathbb{P}(Q_{\sigma_{n-1}} \geqslant 1)\,\mathbb{P}(S_n \leqslant t \mid Q_{\sigma_{n-1}} \geqslant 1) + \mathbb{P}(Q_{\sigma_{n-1}} = 0)\,\mathbb{P}(I_n + S_n \leqslant t \mid Q_{\sigma_{n-1}} = 0).$$

Or la v.a. S_n est indépendante de la taille de la queue, donc $\mathbb{P}(S_n \leqslant t \mid Q_{\sigma_{n-1}} \geqslant 1) = \mathbb{P}(S_n \leqslant t)$ et la v.a. conditionnelle $(I_n \mid Q_{\sigma_{n-1}} = 0)$ représente le laps de temps entre la dernière personne sortie d'une période d'activité du serveur et la personne suivante. Par absence de mémoire de la loi exponentielle, c'est la durée d'attente d'une nouvelle personne et la v.a. $(I_n \mid Q_{\sigma_{n-1}} = 0)$ suit donc la loi $\mathcal{E}(\lambda)$. En conséquence, après dérivation, on obtient la densité à l'équilibre de la v.a. τ_∞ :

$$f_{\tau_\infty}(t) = \mathbb{P}(Q_\infty \geqslant 1)\, f_{S_\infty}(t) + \mathbb{P}(Q_\infty = 0)\, f_{I_\infty + S_\infty}(t) = \rho\mu\, e^{-\mu t} + (1-\rho)(f_{I_\infty} \star f_{S_\infty})(t)$$

où (voir l'annexe 1)

$$(f_{I_\infty} \star f_{S_\infty})(t) = \int_0^t f_{I_\infty}(s) f_{S_\infty}(t-s)\, ds = \lambda\mu \int_0^t e^{-\lambda s}\, e^{-\mu(t-s)}\, ds = \frac{\lambda\mu}{\mu - \lambda}\,(e^{-\lambda t} - e^{-\mu t}).$$

Finalement, on trouve

$$\boxed{f_{\tau_\infty}(t) = \lambda\, e^{-\lambda t}.}$$

La v.a. τ_∞ suit donc la loi $\mathcal{E}(\lambda)$. En fait, on peut démontrer que tous les laps de temps inter-départs sont indépendants en régime stationnaire, ce qui signifie que le processus des départs des clients du système est un processus de Poisson d'intensité λ (tout comme celui des arrivées).

Ce résultat est important pour pouvoir étudier des files d'attente couplées en tandem (deux ou plusieurs services disposés en série). Considérons par exemple une file de clients devant passer par deux guichets successifs. Ce système est en fait la concaténation de deux files simples $M/M/1$. En effet, les clients se présentent au premier guichet selon un processus de Poisson d'intensité λ, ressortent de ce premier guichet une fois leur service accompli selon le même processus de Poisson, puis se présentent au deuxième guichet toujours selon le même de processus de Poisson d'intensité λ (voir Fig. 14).

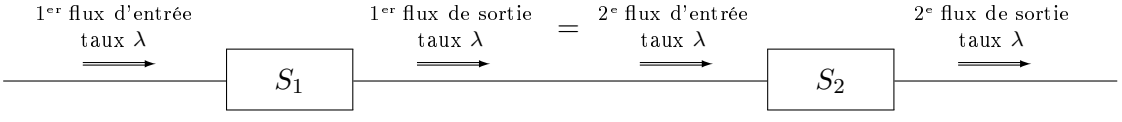


FIGURE 14 – Files en tandem

6 Autres modèles de files d'attente

La file d'attente que l'on vient d'étudier est une file avec des arrivées poissonniennes (temps inter-arrivées exponentiels) et un serveur fournissant un service exponentiel, elle porte la nomenclature de file $M/M/1$ (notation de Kendall) ; la lettre M fait référence au nom de Markov, la loi exponentielle utilisée en arrivée et en service possédant la propriété d'absence de mémoire (propriété « markovienne »). La nomenclature générale d'une file d'attente est de la forme $A/B/n/m/\text{dis}$ où A définit un type d'arrivées, B un type de service, n est le nombre de serveurs, m est la capacité du système (infinie si elle n'est pas précisée), dis est la discipline de service adoptée (FCFS si elle n'est pas précisée). Les types possibles pour A et B sont M (markovien), D (déterministe), G (général)...

Mentionnons quelques autres modèles importants et couramment utilisés :
- file $M/M/1/N$: système avec une salle d'attente de capacité limitée à N places ;
- file $M/M/n_0$: service assuré par n_0 serveurs ;
- file $M/M/\infty$: système assuré par une infinité de serveurs, donc sans attente pour les clients ;
- file $M/M/n_0/n_0$: système comportant n_0 serveurs avec une capacité de n_0 personnes, donc avec refoulement dès que tous les serveurs sont occupés. Par exemple, un parc automobile disposant de n_0 places rentre dans le cadre d'une telle file : les places matérielles jouent le rôle de serveur et les automobilistes sont refoulés dès que le parc est complet ;
- files $G/G/1$: files avec des arrivées et services plus généraux (aléatoires ou non). Par exemple :
  (a) file $M/D/1$: arrivées poissonniennes et service déterministe (constant) ;
  (b) file $D/M/1$: arrivées déterministes (régulières) et service exponentiel ;
  (c) service erlangien : succession de services exponentiels. Un client doit passer successivement par n serveurs proposant des services de durée les v.a. indépendantes $S_1, S_2, \dots, S_n$ suivant respectivement les lois $\mathcal{E}(\mu_1), \mathcal{E}(\mu_2), \dots, \mathcal{E}(\mu_n)$ (voir Fig. 15). À l'issue du dernier service effectué, le client suivant pourra démarrer son premier service.

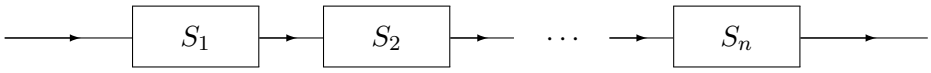


FIGURE 15 – Service erlangien

La durée totale des services dispensés pour le client sera $S = S_1 + S_2 + \cdots + S_n$. Cette situation se rencontre dans des files avec arrivées groupées : le serveur traite des groupes de n personnes, ces personnes demandant des services non nécessairement identiques. La v.a. S représente pour le serveur la durée totale de service d'un groupe. Elle suit une loi d'Erlang généralisée $E(\mu_1, \dots, \mu_n)$ (voir l'annexe 2), elle a pour densité, dans le cas où les paramètres $\mu_1, \dots, \mu_n$ sont tous distincts,

$$\boxed{f_S(t) = \sum_{k=1}^{n} \alpha_k \mu_k \, e^{-\mu_k t} \text{ avec } \alpha_k = 1 \Big/ \prod_{\substack{1 \leqslant j \leqslant n \\ j \neq k}} \Big(1 - \frac{\mu_k}{\mu_j}\Big).}$$

Le service total aura pour durée moyenne

$$\boxed{\mathbb{E}(S) = \sum_{k=1}^{n} \frac{1}{\mu_k}.}$$

Si les services sont similaires, $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$, S suit la loi d'Erlang ordinaire $E(n; \mu)$;

(d) service hyperexponentiel : choix au hasard d'un service exponentiel. Un client a le choix entre n serveurs proposant des services de durée les v.a. indépendantes $S_1, S_2, \dots, S_n$ suivant respectivement les lois $\mathcal{E}(\mu_1), \mathcal{E}(\mu_2), \dots, \mathcal{E}(\mu_n)$. Il choisit le k^{e} serveur avec probabilité p_k indépendamment du service (voir Fig 16). Ce choix est modélisé par une v.a. N indiquant le numéro du serveur choisi ; sa loi de probabilité est donnée par

$$\mathbb{P}(N = k) = p_k, \ 1 \leqslant k \leqslant n.$$

Cette situation se rencontre par exemple dans le cas d'une file simple à un seul serveur, ce dernier étant capable de proposer n services différents, p_k étant alors la probabilité que le client demande le k^{e} service.

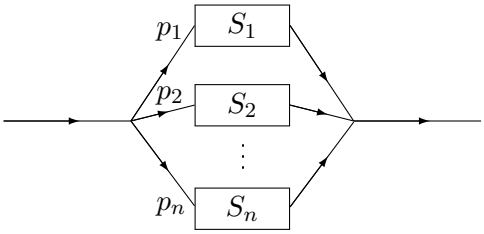


FIGURE 16 – Service hyper-exponentiel

La v.a. N est indépendante des v.a. $S_1, S_2, \ldots, S_n$. Le serveur ainsi choisi délivrera donc un service d'une durée S_N. C'est une v.a. composée. Sa fonction de répartition se calcule facilement en recourant à la formule des probabilités totales :

$$\begin{aligned}\mathbb{P}(S_N \leqslant t) &= \sum_{k=1}^{n} \mathbb{P}(N=k)\,\mathbb{P}(S_N \leqslant t \mid N=k) \\ &= \sum_{k=1}^{n} \mathbb{P}(N=k)\,\mathbb{P}(S_k \leqslant t) \\ &= \sum_{k=1}^{n} p_k\,(1-e^{-\mu_k t_k}).\end{aligned}$$

Sa densité en découle par dérivation :

$$\boxed{f_{S_N}(t) = \sum_{k=1}^{n} p_k \mu_k\, e^{-\mu_k t}.}$$

La v.a. S_N suit la loi hyper-exponentielle $\mathcal{H}(p_1, \ldots, p_n; \mu_1, \ldots, \mu_n)$ (voir l'annexe 2). Le service choisi au hasard aura pour durée moyenne

$$\boxed{\mathbb{E}(S_N) = \sum_{k=1}^{n} \frac{p_k}{\mu_k}.}$$

Lorsque les serveurs proposent le même service, c'est-à-dire $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$, on récupère, puisque $\displaystyle\sum_{k=1}^{n} p_k = 1$,

$$f_{S_N}(t) = \sum_{k=1}^{n} p_k \mu\, e^{-\mu t} = \mu\, e^{-\mu t}$$

et S_N suit la loi exponentielle $\mathcal{E}(\mu)$.

A Annexes

A.1 Probabilité conditionnelle, espérance et variance

– *Probabilité conditionnelle :* si $\mathbb{P}(B) \neq 0$, on définit la probabilité conditionnelle de A sachant B selon

$$\boxed{\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.}$$

Si $B_1, \ldots, B_n$ est une partition de l'univers Ω (i.e. $\Omega = B_1 \cup \ldots \cup B_n$ et les B_i sont non vides et deux à deux disjoints) telle que $\mathbb{P}(B_i) \neq 0$ pour tout i, on a la formule des probabilités totales :

$$\boxed{\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i)\,\mathbb{P}(A \mid B_i).}$$

– *Espérance :* l'espérance mathématique d'une v.a. discrète à valeurs dans $\mathbb{N}$ est définie par

$$\boxed{\mathbb{E}(X) = \sum_{n=0}^{+\infty} n\,\mathbb{P}(X=n)}$$

et celle d'une v.a. continue à valeurs dans $\mathbb{R}^+$ de densité f_X par

$$\boxed{\mathbb{E}(X) = \int_0^{+\infty} t f_X(t)\,dt.}$$

– *Variance :* la variance d'une v.a. est donnée par

$$\boxed{\mathrm{var}\,(X) = \mathbb{E}[(X-\mathbb{E}(X))^2] = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.}$$

– *Somme de deux v.a. :* la loi de la somme de deux v.a. indépendantes discrètes à valeurs dans $\mathbb{N}$ et donnée par

$$\boxed{\mathbb{P}(X+Y=n) = \sum_{k=0}^{+\infty} \mathbb{P}(X=k)\,\mathbb{P}(Y=n-k)}$$

et la somme de deux v.a. indépendantes continues à valeurs dans $\mathbb{R}^+$ de densités f_X et f_Y a pour densité

$$\boxed{f_{X+Y}(t) = (f_X \star f_Y)(t) = \int_0^t f_X(s)\,f_Y(t-s)\,ds.}$$

A.2 Quelques lois de probabilité

– *Loi de Poisson $\mathcal{P}(\lambda)$:*

$$\boxed{\mathbb{P}(X=n) = \frac{\lambda^n}{n!}\,e^{-\lambda} \text{ pour } n \in \mathbb{N}\,;\, \mathbb{E}(X) = \lambda,\ \mathrm{var}\,(X) = \lambda.}$$

– *Loi exponentielle $\mathcal{E}(\mu)$:* densité : f_X, fonction de répartition : F_X ;

$$\boxed{f_X(t) = \mu\,e^{-\mu t} \text{ et } F_X(t) = 1 - e^{-\mu t} \text{ pour } t \in \mathbb{R}^+\,;\, \mathbb{E}(X) = \frac{1}{\mu},\ \mathrm{var}\,(X) = \frac{1}{\mu^2}.}$$

– *Loi géométrique $\mathcal{G}(\rho)$: $\rho \in]0,1[$;*

$$\boxed{\mathbb{P}(X=n) = \rho\,(1-\rho)^{n-1} \text{ pour } n \in \mathbb{N}^*\,;\, \mathbb{E}(X) = \frac{1}{\rho},\ \mathrm{var}\,(X) = \frac{1-\rho}{\rho^2}.}$$

– *Loi d'Erlang $E(n;\lambda)$:*

$$\boxed{f_X(t) = \frac{\lambda^n t^{n-1}}{(n-1)!}\,e^{-\lambda t} \text{ pour } t \in \mathbb{R}^+\,;\, \mathbb{E}(X) = \frac{n}{\lambda},\ \mathrm{var}\,(X) = \frac{n}{\lambda^2}.}$$

– *Loi d'Erlang généralisée $E(\lambda_1,\ldots,\lambda_n)$:* dans le cas où les $\lambda_1,\ldots,\lambda_n$ sont tous distincts, on pose $\alpha_k = 1 \Big/ \prod_{\substack{1\leqslant j\leqslant n\\ j\neq k}} \left(1 - \frac{\lambda_k}{\lambda_j}\right)$;

$$\boxed{f_X(t) = \sum_{k=1}^{n} \alpha_k \lambda_k\,e^{-\lambda_k t} \text{ pour } t \in \mathbb{R}^+\,;\, \mathbb{E}(X) = \sum_{k=1}^{n} \frac{\alpha_k}{\lambda_k},\ \mathrm{var}\,(X) = 2\sum_{k=1}^{n} \frac{\alpha_k}{\lambda_k^2} - \left(\sum_{k=1}^{n} \frac{\alpha_k}{\lambda_k}\right)^2}$$

– *Loi hyper-exponentielle $\mathcal{H}(p_1,\ldots,p_n;\lambda_1,\ldots,\lambda_n)$:* $\displaystyle\sum_{k=1}^{n} p_k = 1$;

$$\boxed{f_X(t) = \sum_{k=1}^{n} p_k \lambda_k\,e^{-\lambda_k t} \text{ pour } t \in \mathbb{R}^+\,;\, \mathbb{E}(X) = \sum_{k=1}^{n} \frac{p_k}{\lambda_k},\ \mathrm{var}\,(X) = 2\sum_{k=1}^{n} \frac{p_k}{\lambda_k^2} - \left(\sum_{k=1}^{n} \frac{p_k}{\lambda_k}\right)^2.}$$

A.3 File $M/M/n_0$

Pour la file d'attente à n_0 serveurs, les lois du régime stationnaire qui existe lorsque la charge est inférieure aux nombres de serveurs, i.e. $\rho = \dfrac{\lambda}{\mu} < n_0$, sont données par

$$\mathbb{P}(Q_\infty = n) = \begin{cases} \pi_0 \dfrac{\rho^n}{n!} & \text{si } 0 \leqslant n < n_0 \\ \pi_0 \dfrac{n_0^{n_0}}{n_0!} \Big(\dfrac{\rho}{n_0}\Big)^{n_0} & \text{si } n \geqslant n_0 \end{cases}$$

$$\text{avec } \pi_0 = \mathbb{P}(Q_\infty = 0) = \left[\sum_{j=0}^{n_0-1} \frac{\rho^j}{j!} + \frac{\rho^{n_0}}{n_0!\,(1-\frac{\rho}{n_0})} \right]^{-1},$$

$$\mathbb{E}(Q_\infty) = \rho + \pi_0 \frac{\rho^{n_0+1}}{(n_0-1)!\,(n_0-\rho)^2},$$

$$\mathbb{P}(W_\infty \leqslant t) = 1 - \frac{\rho^{n_0}}{n_0!\,(1-\frac{\rho}{n_0})}\, e^{-(n_0\mu-\lambda)t} \qquad \text{loi } \mathcal{E}(n_0\mu-\lambda) \text{ pondérée en } 0,$$

$$\mathbb{E}(W_\infty) = \frac{\pi_0\, \rho^{n_0}}{n_0!\,(1-\frac{\rho}{n_0})} \frac{1}{n_0\mu - \lambda}.$$

La formule de Little s'écrit ici

$$\mathbb{E}(Q_\infty) = \lambda\, \mathbb{E}(W_\infty + S_\infty).$$

A.4 Un autre exemple de file d'attente

On observe à une station de taxis les échanges entre taxis et clients. Les arrivées des taxis sont modélisées par un processus de Poisson d'intensité λ_1, $(A_1(t))_{t\in\mathbb{R}^+}$, celles des clients par un processus de Poisson d'intensité λ_2, $(A_2(t))_{t\in\mathbb{R}^+}$. On suppose les deux processus indépendants et l'on pose $Q(t) = A_1(t) - A_2(t)$. Cette quantité représente en valeur absolue le nombre de taxis ou de clients en attente à l'instant t, et son signe indique si ce sont des taxis ou des clients qui sont en attente.

Déterminons la loi de probabilité de la v.a. $Q(t)$. On a pour $m \geqslant 0$:

$$\mathbb{P}(Q(t) = m) = \sum_{n=0}^{+\infty} \mathbb{P}(A_1(t) = m+n)\, \mathbb{P}(A_2(t) = n).$$

En effet, dans la somme ci-dessus l'indice courant n représente le nombre de clients en attente à l'instant t. Comme $Q(t) = m$, il y a $m+n$ taxis disponibles à ce même instant. On a ensuite

$$\mathbb{P}(Q(t) = m) = e^{-(\lambda_1+\lambda_2)t} \sum_{n=0}^{+\infty} \frac{(\lambda_1 t)^{m+n}}{(m+n)!} \frac{(\lambda_2 t)^n}{n!} = \left(\frac{\lambda_1}{\lambda_2}\right)^{m/2} e^{-(\lambda_1+\lambda_2)t} \sum_{n=0}^{+\infty} \frac{(t\sqrt{\lambda_1\lambda_2}\,)^{m+2n}}{(m+n)!\,n!}$$

soit encore

$$\boxed{\mathbb{P}(Q(t) = m) = \left(\frac{\lambda_1}{\lambda_2}\right)^{m/2} e^{-(\lambda_1+\lambda_2)t} I_m(2t\sqrt{\lambda_1\lambda_2}\,)}$$

où $I_m(z) = \displaystyle\sum_{n=0}^{+\infty} \frac{1}{(m+n)!\,n!} \Big(\frac{z}{2}\Big)^{m+2n}$ est une fonction de Bessel. Cette formule est également valable pour $m \leqslant 0$. Par exemple, la quantité $\mathbb{P}(Q(t) = 0) = e^{-(\lambda_1+\lambda_2)t} I_0(2t\sqrt{\lambda_1\lambda_2}\,)$ est la probabilité que la station soit vide à l'instant t.

A.5 Le paradoxe de l'autobus

Dans une gare routière, des autobus arrivent selon un processus de Poisson d'intensité λ. On sait que dans un tel modèle, le laps de temps moyen séparant les passages de deux autobus consécutifs est $1/\lambda$. Un usager se présente à un instant t fixé et se demande combien de temps (en moyenne) il va attendre le prochain autobus.

Notons T_t^- et T_t^+ les laps de temps séparant l'instant t des passages consécutifs des deux autobus respectivement antérieur et postérieur à l'instant t, puis $\mathcal{T}_t = T_t^- + T_t^+$. L'instant s_{N_t} est l'instant de passage du dernier autobus avant l'arrivée de l'usager et s_{N_t+1} est celui du prochain autobus que l'usager pourra emprunter. Le temps $\mathcal{T}_t$ est l'intervalle de temps encadrant l'instant fixé t écoulé entre deux passages successifs d'autobus. On a précisément (voir Fig. 17)

$$T_t^- = t - s_{N_t} \quad \text{et} \quad T_t^+ = s_{N_t+1} - t.$$

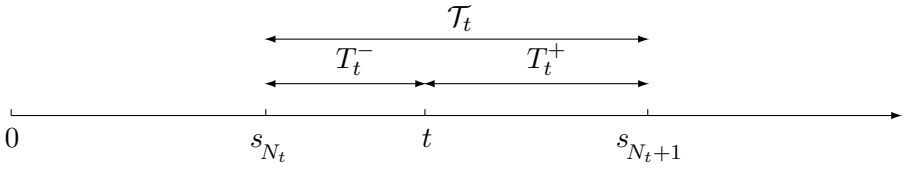


FIGURE 17 – Passages successifs encadrant l'instant t

Lorsque $N_t = n$, on a évidemment $s_{N_t} = n$ et $s_{N_t+1} = s_{n+1}$. Par ailleurs, la relation $N_t = n$ signifie qu'exactement n autobus sont passés à la gare durant l'intervalle de temps $[0, t]$, soit encore que le n^{e} est passé avant l'instant t et le $(n+1)^{\text{e}}$ est passé après t. On a ainsi l'équivalence $N_t = n \iff s_n \leqslant t < s_{n+1}$.

Ces considérations permettent de calculer la loi conjointe des v.a. T_t^- et T_t^+ à l'aide de la formule des probabilités totales. On a pour $0 \leqslant u < t$ et $v \geqslant 0$:

$$\begin{aligned}
\mathbb{P}(T_t^- > u, T_t^+ > v) &= \mathbb{P}(s_{N_t} < t - u, s_{N_t+1} > t + v) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(s_n < t - u, s_{n+1} > t + v, N_t = n) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(s_n < t - u, s_{n+1} > t + v, s_n \leqslant t < s_{n+1}) \\
&= \sum_{n=0}^{\infty} \mathbb{P}(s_n < t - u, s_{n+1} > t + v).
\end{aligned}$$

Rappelant que $s_{n+1} = s_n + T_{n+1}$ et que T_{n+1} a même loi que T_1 et est indépendante de s_n, on a pour $n \geqslant 1$:

$$\begin{aligned}
\mathbb{P}(s_n < t - u, s_{n+1} > t + v) &= \mathbb{P}(s_n < t - u, T_{n+1} > t + v - s_n) \\
&= \int_0^{t-u} \mathbb{P}(T_1 > t + v - s) f_{s_n}(s)\, ds \\
&= \int_0^{t-u} e^{-\lambda(t+v-s)} \frac{\lambda^n s^{n-1}}{(n-1)!} e^{-\lambda s}\, ds.
\end{aligned}$$

Poursuivons les calculs :

$$\begin{aligned}
\mathbb{P}(T_t^- > u, T_t^+ > v) &= \mathbb{P}(s_1 > t + v) + \int_0^{t-u} e^{-\lambda(t+v-s)} \left[\sum_{n=1}^{\infty} \frac{\lambda^n s^{n-1}}{(n-1)!} e^{-\lambda s} \right] ds \\
&= e^{-\lambda(t+v)} + \lambda \int_0^{t-u} e^{-\lambda(t+v-s)}\, ds \quad = \quad e^{-\lambda(u+v)}.
\end{aligned}$$

De cette relation, on tire en faisant $u = 0$ ou $v = 0$:

$$\mathbb{P}(T_t^- > u) = e^{-\lambda u} \quad \text{et} \quad \mathbb{P}(T_t^+ > v) = e^{-\lambda v}$$

d'où les fonctions de répartition :

$$\mathbb{P}(T_t^- \leqslant u) = \begin{cases} 1 - e^{-\lambda u} & \text{si } 0 \leqslant u < t \\ 1 & \text{si } u \geqslant t \end{cases} \qquad \text{et} \quad P(T_t^+ \leqslant v) = e^{-\lambda v}.$$

Ceci montre que la v.a. T_t^+ suit la loi exponentielle $\mathcal{E}(\lambda)$ et que la v.a. T_t^- suit une loi exponentielle $\mathcal{E}(\lambda)$ pondérée en t :

$$\mathbb{P}(T_t^- = t) = \mathbb{P}(s_{N_t} = 0) = \mathbb{P}(N_t = 0) = e^{-\lambda t}.$$

Par ailleurs, on voit que

$$\mathbb{P}(T_t^- > u, T_t^+ > v) = \mathbb{P}(T_t^- > u)\mathbb{P}(T_t^+ > v),$$

égalité signifiant que les v.a. T_t^- et T_t^+ sont indépendantes. L'espérance de T_t^+ est bien sûr $1/\lambda$ et celle de T_t^- se calcule comme suit :

$$\mathbb{E}(T_t^-) = \int_0^t \lambda s\, e^{-\lambda s}\, ds + t\, e^{-\lambda t} = \frac{1}{\lambda}\,(1 - e^{-\lambda t}).$$

On a donc

$$\mathbb{E}(\mathcal{T}_t) = \mathbb{E}(T_t^-) + \mathbb{E}(T_t^+) = \mathbb{E}(s_{N_t+1} - s_{N_t}) = \frac{1}{\lambda}\,(2 - e^{-\lambda t}).$$

Asymptotiquement, en temps grand, la v.a. T_t^- suit approximativement une véritable loi exponentielle d'espérance $1/\lambda$, et alors la v.a. $\mathcal{T}_t$ suit approximativement la loi d'Erlang $E(2, \lambda)$. En particulier :

$$\mathbb{E}(\mathcal{T}_t) \underset{t\to+\infty}{\longrightarrow} \frac{2}{\lambda}.$$

On peut répondre à la question de l'usager se présentant à la gare à l'instant t : il attendra un temps moyen de $1/\lambda$ avant l'arrivée d'un prochain autobus. Ce résultat est un paradoxe dans la mesure où le laps de temps moyen entre deux arrivées consécutives étant déjà $1/\lambda$ et l'usager arrivant au hasard entre deux autobus, on aurait pu s'attendre à un temps moyen d'attente moitié moindre : $1/(2\lambda)$. Une explication à ce phénomène réside en le fait que si les temps inter-arrivées suivent des lois exponentielles, il s'agit en fait de durées de la forme $s_{n+1} - s_n$ relatives aux n^{e} et $(n+1)^{\text{e}}$ autobus, alors que le laps de temps inter-arrivées réellement associé au problème doit recouvrir l'instant fixé t, il est relatif à un numéro d'autobus **aléatoire** : $s_{N_t+1} - s_{N_t}$, durée qui ne suit plus une loi exponentielle. Les calculs précédents ont montré que la durée d'attente $s_{N_t+1} - t$ de l'usager suivait la même loi exponentielle que les $s_{n+1} - s_n$ et cela peut s'expliquer par la propriété d'absence de mémoire de la loi exponentielle.

Quelques exercices corrigés...

Exercice 1 (Calculs d'espérances et de variances)
Calculer les espérances et variances des lois de Poisson $\mathcal{P}(\lambda)$, géométrique $\mathcal{G}(\rho)$, exponentielle $\mathcal{E}(\mu)$, d'Erlang $E(n;\lambda)$ et hyper-exponentielle $\mathcal{H}(p_1,\ldots,p_n;\lambda_1,\ldots,\lambda_n)$.

Solution.

1. *Loi de Poisson $\mathcal{P}(\lambda)$*

   (a) L'espérance est donnée par

   $$\mathbb{E}(X) = \sum_{n=0}^{+\infty} n\,\mathbb{P}(X=n) = \sum_{n=0}^{+\infty} n\,\frac{\lambda^n}{n!}\,e^{-\lambda} = \lambda\,e^{-\lambda}\sum_{n=1}^{+\infty}\frac{\lambda^{n-1}}{(n-1)!}.$$

   Or

   $$\sum_{n=1}^{+\infty}\frac{\lambda^{n-1}}{(n-1)!} = \sum_{n=0}^{+\infty}\frac{\lambda^n}{n!} = e^{\lambda},$$

   donc

   $$\boxed{\mathbb{E}(X) = \lambda.}$$

   (b) La variance se calcule selon

   $$\mathrm{var}\,(X) = E(X^2) - [\mathbb{E}(X)]^2.$$

   On a

   $$\mathbb{E}(X^2) = \sum_{n=0}^{+\infty} n^2\,\mathbb{P}(X=n) = \sum_{n=0}^{+\infty} n^2\,\frac{\lambda^n}{n!}\,e^{-\lambda}.$$

   En écrivant $n^2 = n(n-1)+n$, on obtient $\dfrac{n^2}{n!} = \dfrac{1}{(n-1)!} + \dfrac{1}{(n-2)!}$ et alors

   $$\begin{aligned}\mathbb{E}(X^2) &= \lambda\,e^{-\lambda}\sum_{n=1}^{+\infty}\frac{\lambda^{n-1}}{(n-1)!} + \lambda^2\,e^{-\lambda}\sum_{n=2}^{+\infty}\frac{\lambda^{n-2}}{(n-2)!}\\ &= \lambda\,e^{-\lambda}\sum_{n=0}^{+\infty}\frac{\lambda^n}{n!} + \lambda^2\,e^{-\lambda}\sum_{n=0}^{+\infty}\frac{\lambda^n}{n!} = \lambda+\lambda^2.\end{aligned}$$

   D'où

   $$\boxed{\mathrm{var}\,(X) = \lambda.}$$

2. *Loi géométrique $\mathcal{G}(\rho)$*

   (a) L'espérance vaut

   $$\mathbb{E}(X) = \sum_{n=1}^{+\infty} n\,\mathbb{P}(X=n) = \sum_{n=1}^{+\infty} n\rho\,(1-\rho)^{n-1}.$$

   Or

   $$\sum_{n=1}^{+\infty} n\,x^{n-1} = \frac{d}{dx}\Big(\sum_{n=0}^{+\infty} x^n\Big) = \frac{d}{dx}\Big(\frac{1}{1-x}\Big) = \frac{1}{(1-x)^2},$$

   donc

   $$\boxed{\mathbb{E}(X) = \frac{1}{\rho}.}$$

(b) De même,

$$\mathbb{E}(X^2) = \sum_{n=1}^{+\infty} n^2\, \mathbb{P}(X=n) = \sum_{n=1}^{+\infty} n^2\, \rho\,(1-\rho)^{n-1}.$$

En écrivant $n^2 = n(n-1)+n$, on obtient

$$\begin{aligned}\sum_{n=1}^{+\infty} n^2\, x^{n-1} &= x \sum_{n=1}^{+\infty} n(n-1)\, x^{n-2} + \sum_{n=1}^{+\infty} n\, x^{n-1} \\ &= x\, \frac{d^2}{dx^2}\Big(\sum_{n=0}^{+\infty} x^n\Big) + \frac{d}{dx}\Big(\sum_{n=0}^{+\infty} x^n\Big) \\ &= \frac{2x}{(1-x)^3} + \frac{1}{(1-x)^2} = \frac{1+x}{(1-x)^3},\end{aligned}$$

et alors

$$\mathbb{E}(X^2) = \frac{2-\rho}{\rho^2}.$$

Ainsi

$$\boxed{\mathrm{var}\,(X) = \frac{1-\rho}{\rho^2}.}$$

3. *Loi exponentielle $\mathcal{E}(\mu)$*

(a) L'espérance vaut

$$\mathbb{E}(X) = \int_0^{+\infty} x\, f_{_X}(x)\, dx = \int_0^{+\infty} \lambda x\, e^{-\lambda x}\, dx.$$

On rappelle l'expression de la fonction eulérienne

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}\, e^{-x}\, dx.$$

On a pour tout $n \in \mathbb{N}^*$, $\Gamma(n) = (n-1)!$, et alors

$$\int_0^{+\infty} x^n\, e^{-\lambda x}\, dx = \frac{n!}{\lambda^{n+1}}.$$

Ici, cela donne

$$\boxed{\mathbb{E}(X) = \frac{1}{\lambda}.}$$

(b) On a de la même manière

$$\mathbb{E}(X^2) = \int_0^{+\infty} x^2\, f_{_X}(x)\, dx = \int_0^{+\infty} \lambda x^2\, e^{-\lambda x}\, dx = \frac{2\lambda}{\lambda^3} = \frac{2}{\lambda^2},$$

et donc

$$\boxed{\mathrm{var}\,(X) = \frac{1}{\lambda^2}.}$$

4. *Loi d'Erlang $E(n;\lambda)$*

Une v.a. X suivant la loi d'Erlang $E(n;\lambda)$ est la somme de n v.a. exponentielle $\mathcal{E}(\lambda)$ indépendantes : $X = X_1 + \cdots + X_n$. Il vient immédiatement

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n), \\ \mathrm{var}\,(X) &= \mathrm{var}\,(X_1) + \cdots + \mathrm{var}\,(X_n),\end{aligned}$$

ce qui donne

$$\boxed{\mathbb{E}(X) = \frac{n}{\lambda} \text{ et } \mathrm{var}\,(X) = \frac{n}{\lambda^2}.}$$

5. *Loi hyper-exponentielle $\mathcal{H}(p_1,\ldots,p_n;\lambda_1,\ldots,\lambda_n)$*

(a) L'espérance vaut

$$\mathbb{E}(X) = \int_0^\infty x \sum_{k=1}^n p_k \lambda_k \, e^{-\lambda_k x} \, dx = \sum_{k=1}^n p_k \lambda_k \int_0^\infty x \, e^{-\lambda_k x} \, dx$$

soit

$$\boxed{\mathbb{E}(X) = \sum_{k=1}^n \frac{p_k}{\lambda_k}.}$$

(b) On a de même

$$\mathbb{E}(X^2) = \int_0^\infty x^2 \sum_{k=1}^n p_k \lambda_k \, e^{-\lambda_k x} \, dx = \sum_{k=1}^n p_k \lambda_k \int_0^\infty x^2 \, e^{-\lambda_k x} \, dx = 2 \sum_{k=1}^n \frac{p_k}{\lambda_k^2}$$

et donc

$$\boxed{\mathrm{var}\,(X) = 2 \sum_{k=1}^n \frac{p_k}{\lambda_k^2} - \left(\sum_{k=1}^n \frac{p_k}{\lambda_k} \right)^2.}$$

Une autre approche consiste à considérer une v.a. X suivant la loi hyper-exponentielle $\mathcal{H}(p_1, \ldots, p_n; \lambda_1, \ldots, \lambda_n)$ comme étant une variable exponentielle choisie au hasard parmi n : $X = T_N$ où N est une v.a. à valeurs dans $\{1, 2, \ldots, n\}$ suivant la loi discrète $\{p_1, p_2, \ldots, p_n\}$ et les $T_1, T_2, \ldots, T_n$ sont des v.a. indépendantes et indépendantes de N à valeurs dans $\mathbb{R}^+$ suivant les lois respectives $\mathcal{E}(\lambda_1), \mathcal{E}(\lambda_2), \ldots, \mathcal{E}(\lambda_n)$. En faisant appel à une formule similaire à celle des probabilités totales, on trouve

$$\mathbb{E}(X) = \mathbb{E}(T_N) = \sum_{k=1}^n \mathbb{P}(N = k) \, \mathbb{E}(T_N \mid N = k) = \sum_{k=1}^n p_k \, \mathbb{E}(T_k) = \sum_{k=1}^n \frac{p_k}{\lambda_k}$$

et

$$\mathbb{E}(X^2) = \mathbb{E}(T_N^2) = \sum_{k=1}^n \mathbb{P}(N = k) \, \mathbb{E}(T_N^2 \mid N = k) = \sum_{k=1}^n p_k \, \mathbb{E}(T_k^2) = 2 \sum_{k=1}^n \frac{p_k}{\lambda_k^2}.$$

Exercice 2 (Monoserveur $M/M/1$)
On utilise une ligne à 512 kb/s (en moyenne) pour faire un transfert de fichier. Le fichier est transféré par blocs de 100 000 caractères de 8 bits. La ligne joue le rôle de serveur.

1. Quel est le temps moyen $1/\mu$ nécessaire pour transférer un bloc ?
2. La ligne délivre un trafic Poissonnien avec une charge limitée à 60% (de 512 kb/s).
   (a) Quel est le taux d'arrivée en blocs/s ?
   (b) Calculer le temps moyen d'attente d'un bloc dans la ligne, le temps de réponse de la ligne ainsi que le nombre moyen de blocs transitant dans la ligne en régime stationnaire.

Solution.

1. Un bloc contient $100\,000 \times 8 = 800$ kb. Le temps de transfert vaut $1/\mu = \frac{800}{512} = 1,56$ s/bloc.
2. (a) La charge est de $\rho = 0,6$ donc le taux d'arrivée vaut $\lambda = \rho\mu = 0,38$ blocs/s.
   (b) On a $\rho < 1$, on peut donc utiliser les résultats concernant le régime stationnaire : le temps d'attente d'un bloc par

$$\mathbb{E}(W_\infty) = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu(1 - \rho)} = \frac{0,6}{0,4 \times 0,64} = 2,34 \text{ s},$$

le temps de réponse de la ligne est donné par

$$\mathbb{E}(W_\infty + S_\infty) = \frac{1}{\mu - \lambda} = \frac{1}{0,256} = 3,9 \text{ s},$$

et le nombre moyen de blocs transitant dans la ligne par

$$\mathbb{E}(Q_\infty) = \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1-\rho} = \frac{0,6}{0,4} = 1,5 \text{ blocs.}$$

On peut vérifier sur cet exemple les formules de Little :

$$\mathbb{E}(Q_\infty) = \mu\, \mathbb{E}(W_\infty) = \lambda\, \mathbb{E}(W_\infty + S_\infty).$$

Exercice 3 (File $M/M/n_0$)
Préliminaire. — En utilisant l'annexe 3, écrire le nombre moyen de clients ainsi que le temps moyen passé par un client dans un système décrit par le modèle $M/M/n_0$ dans les cas $n_0 = 1, 2, 3$ serveurs.
On considère une cabine téléphonique avec une loi d'arrivée de Poisson avec un taux d'arrivée de 3 personnes/heure. Le temps passé par un individu dans la cabine suit la loi exponentielle de moyenne 10 mn.

1. Quel est le temps moyen d'attente de chaque individu ?
2. Quel est le nombre total moyen de personnes dans le système ?
3. Quelle est la probabilité que le temps total passé dans le système soit plus grand ou égal à 10 mn ? 15 mn ? 20 mn ? Quelle conclusion peut-on en tirer ?
4. On suppose maintenant que le taux d'arrivée s'élève à 10 personnes/heure. On souhaiterait limiter le nombre moyen de personnes à moins de trois (resp. le temps d'attente moyen à moins de 10 mn). De combien de cabines faudrait-il disposer au minimum ?

Solution.

Préliminaire. — On a d'abord

$$\pi_0 = \mathbb{P}(Q_\infty = 0) = \begin{cases} \left[1 + \dfrac{\rho}{1-\rho}\right]^{-1} = 1-\rho & \text{si } n_0 = 1, \\ \left[1 + \rho + \dfrac{\rho^2}{2(1-\frac{\rho}{2})}\right]^{-1} = \dfrac{2-\rho}{2+\rho} & \text{si } n_0 = 2, \\ \left[1 + \rho + \dfrac{\rho^2}{2} + \dfrac{\rho^3}{6\,(1-\frac{\rho}{3})}\right]^{-1} = \dfrac{2(3-\rho)}{6+4\rho+\rho^2} & \text{si } n_0 = 3, \end{cases}$$

puis

$$\mathbb{E}(Q_\infty) = \begin{cases} \rho + \pi_0 \dfrac{\rho^2}{(1-\rho)^2} = \dfrac{\rho}{1-\rho} & \text{si } n_0 = 1, \\ \rho + \pi_0 \dfrac{\rho^3}{(2-\rho)^2} = \dfrac{4\rho}{4-\rho^2} & \text{si } n_0 = 2, \\ \rho + \pi_0 \dfrac{\rho^4}{2(3-\rho)^2} = \dfrac{\rho(18+6\rho-\rho^2)}{(3-\rho)(6+4\rho+\rho^2)} & \text{si } n_0 = 3 \end{cases}$$

et

$$\mathbb{E}(W_\infty) = \begin{cases} \dfrac{\pi_0\rho}{1-\rho}\,\dfrac{1}{\mu-\lambda} = \dfrac{\rho}{\mu(1-\rho)} & \text{si } n_0 = 1, \\ \dfrac{\pi_0\rho^2}{2(1-\frac{\rho}{2})}\,\dfrac{1}{2\mu-\lambda} = \dfrac{\rho^2}{\mu(4-\rho^2)} & \text{si } n_0 = 2, \\ \dfrac{\pi_0\rho^3}{6(1-\frac{\rho}{3})}\,\dfrac{1}{3\mu-\lambda} = \dfrac{\rho^3}{\mu(3-\rho)(6+4\rho+\rho^2)} & \text{si } n_0 = 3. \end{cases}$$

Le taux d'arrivée vaut $\lambda = 3$ personnes/h=0,05 personnes/mn et le temps d'appel moyen vaut $1/\mu = 10$ mn/personne, d'où l'intensité du trafic $\rho = \lambda/\mu = 0,5 < 1$. Dans ces conditions :

1. le temps moyen d'attente d'un usager est donné par

$$\mathbb{E}(W_\infty) = \frac{\rho}{\mu(1-\rho)} = 10 \text{ mn} ;$$

2. le nombre moyen de personnes dans le système est donné par

$$\mathbb{E}(Q_\infty) = \frac{\rho}{1-\rho} = 1.$$

3. Le temps de séjour (temps d'attente + temps de communication), $W_\infty + S_\infty$ suit la loi exponentielle $\mathcal{E}(\mu - \lambda = 0,05)$ et donc

$$\begin{array}{rcccl} \mathbb{P}(W_\infty + S_\infty \geqslant 10 \text{ mn}) & = & e^{-0,05\times 10} & = & 0,606, \\ \mathbb{P}(W_\infty + S_\infty \geqslant 15 \text{ mn}) & = & e^{-0,05\times 15} & = & 0,470, \\ \mathbb{P}(W_\infty + S_\infty \geqslant 20 \text{ mn}) & = & e^{-0,05\times 20} & = & 0,367. \end{array}$$

Le temps de séjour moyen vaut

$$\mathbb{E}(W_\infty + S_\infty) = 20 \text{ mn}.$$

La probabilité de rester longtemps dans le système est de plus en plus faible. La probabilité de rester plus du temps de séjour moyen est de 36,7 %.

4. Maintenant $\lambda = 10$ personnes/h, $\mu = 6$ personnes/h$= 0,1$ personnes/mn et alors $\rho = \frac{5}{3} \in]1,2[$. Si l'on dispose de n_0 cabines téléphoniques,
   (a) pour $n_0 = 1$, $\rho > 1$, il n'y a pas de régime stationnaire, la longueur de la file devient de plus en plus grande et $\mathbb{E}(Q_\infty) = \mathbb{E}(W_\infty) = +\infty$;
   (b) pour $n_0 = 2$, $\rho < 2$, il y a un régime stationnaire et $\mathbb{E}(Q_\infty) = \frac{60}{11} = 5,45$ personnes, $\mathbb{E}(W_\infty) = \frac{25}{10\times 11} = 22$ mn 42 s ;
   (c) pour $n_0 = 3$, $\rho < 3$, il y a un régime stationnaire et $\mathbb{E}(Q_\infty) = \frac{1135}{556} = 2,04$ personnes, $\mathbb{E}(W_\infty) = \frac{125}{10\times 556} = 2$ mn 12 s.

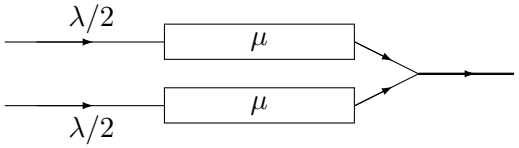
   Il faut donc disposer d'au moins trois cabines pour avoir un nombre moyen de personnes inférieur à trois ou un temps d'attente moyen inférieur à 10 mn.

Exercice 4 (Comparaison entre les files $2 \times M(\lambda/2)/M(\mu)/1$ et $M(\lambda)/M(\mu)/2$)

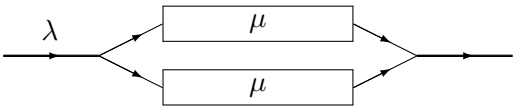
1. En utilisant l'annexe 3, calculer pour chacune des deux files $2 \times M(\lambda/2)/M(\mu)/1$ (deux files indépendantes avec pour chacune un taux d'arrivée moyen $\lambda/2$ et un serveur offrant un service d'une durée moyenne $1/\mu$) et $M(\lambda)/M(\mu)/2$ (une seule file de taux global d'arrivée moyen λ répartie entre deux serveurs offrant chacun un service d'une durée moyenne $1/\mu$) en régime stationnaire ($\lambda < 2\mu$) :
   – la longueur moyenne de la file $\mathbb{E}(Q_\infty)$;
   – le temps d'attente moyen avant service $\mathbb{E}(W_\infty)$;
   – le temps de séjour moyen dans le système $\mathbb{E}(W_\infty + S_\infty)$.
2. Dresser un tableau comparatif des résultats obtenus puis en tirer une conclusion.

Solution.

Voir le tableau ci-après.



Deux files $M(\lambda/2)/M(\mu)/1$



Une file $M(\lambda)/M(\mu)/2$

	2 files $M(\lambda/2)/M(\mu)/1$	comparaison	$M(\lambda)/M(\mu)/2$
$\mathbb{E}(Q_\infty)$	pour chaque file : $\dfrac{\lambda}{2\mu-\lambda}$ pour l'ensemble des deux files : $\dfrac{2\lambda}{2\mu-\lambda}$	$<$ $>$	$\dfrac{4\lambda\mu}{4\mu^2-\lambda^2}$
$\mathbb{E}(W_\infty)$	$\dfrac{\lambda}{\mu(2\mu-\lambda)}$	$>$	$\dfrac{\lambda^2}{\mu(4\mu^2-\lambda^2)}$
$\mathbb{E}(W_\infty+S_\infty)$	$\dfrac{2}{2\mu-\lambda}$	$>$	$\dfrac{4\mu}{4\mu^2-\lambda^2}$

Vérification des comparaisons :

$$\frac{4\lambda\mu}{4\mu^2-\lambda^2}-\frac{\lambda}{2\mu-\lambda}=\frac{\lambda}{2\mu+\lambda}>0\ ;\qquad \frac{4\lambda\mu}{4\mu^2-\lambda^2}-\frac{2\lambda}{2\mu-\lambda}=-\frac{2\lambda^2}{4\mu^2-\lambda^2}<0\ ;$$

$$\frac{\lambda^2}{\mu(4\mu^2-\lambda^2)}-\frac{\lambda}{\mu(2\mu-\lambda)}=-\frac{2\lambda}{4\mu^2-\lambda^2}<0\ ;\qquad \frac{4\mu}{4\mu^2-\lambda^2}-\frac{2}{2\mu-\lambda}=-\frac{2\lambda}{4\mu^2-\lambda^2}<0.$$

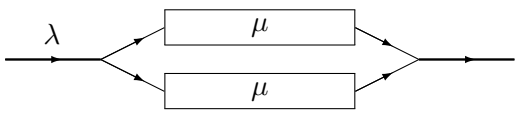
Commentaire : dans les deux cas, le nombre de serveurs est le même. On compare donc ici deux répartitions des arrivées. Les temps d'attente et de séjour pour la file $M(\lambda)/M(\mu)/2$ sont plus courts que ceux de la file dédoublée $M(\lambda/2)/M(\mu)/1$. De même, la longueur de la file $M(\lambda)/M(\mu)/2$ est plus courte que la longueur totale des deux files $M(\lambda/2)/M(\mu)/1$. En conséquence, du point de vue du consommateur, la file unique est plus avantageuse que la file fragmentée. Cela étant, la longueur individuelle de chaque file $M(\lambda/2)/M(\mu)/1$ est plus courte que celle de la file $M(\lambda)/M(\mu)/2$. Dans certaines infrastructures telles que les hypermarchés, les péages d'autoroute, pour des raisons de dimensionnement, on préférera plutôt multiplier les files (fragmentation) pour réduire la longueur de chaque file parallèle.

Exercice 5 (Comparaison entre les files $M(\lambda)/M(\mu)/2$ et $M(\lambda)/M(2\mu)/1$)

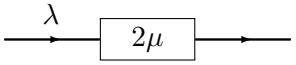
1. En utilisant l'annexe 3, calculer pour chacune des deux files $M(\lambda)/M(\mu)/2$ (deux serveurs offrant chacun un service d'une durée moyenne $1/\mu$) et $M(\lambda)/M(2\mu)/1$ (un serveur offrant un service d'une durée moyenne $1/(2\mu)$) en régime stationnaire ($\lambda<2\mu$) :
   - la longueur moyenne de la file $\mathbb{E}(Q_\infty)$;
   - le temps d'attente moyen avant service $\mathbb{E}(W_\infty)$, le temps de séjour moyen dans le système $\mathbb{E}(W_\infty+S_\infty)$;
   - la probabilité de trouver le système vide $\mathbb{P}(Q_\infty=0)=\mathbb{P}(W_\infty=0)$.
2. Dresser un tableau comparatif des résultats obtenus puis en tirer une conclusion.

Solution.

Voir le tableau ci-après.



File $M(\lambda)/M(\mu)/2$



File $M(\lambda)/M(\mu)/2$

	$M(\lambda)/M(2\mu)/1$	comparaison	$M(\lambda)/M(\mu)/2$
$\mathbb{E}(Q_\infty)$	$\dfrac{\lambda}{2\mu-\lambda}$	$<$	$\dfrac{4\lambda\mu}{4\mu^2-\lambda^2}$
$\mathbb{E}(W_\infty)$	$\dfrac{\lambda}{2\mu(2\mu-\lambda)}$	$>$	$\dfrac{\lambda^2}{\mu(4\mu^2-\lambda^2)}$
$\mathbb{E}(W_\infty+S_\infty)$	$\dfrac{1}{2\mu-\lambda}$	$<$	$\dfrac{4\mu}{4\mu^2-\lambda^2}$
$\mathbb{P}(W_\infty=0)$	$1-\dfrac{\lambda}{2\mu}$	$<$	$\dfrac{2\mu^2+\lambda\mu-\lambda^2}{\mu(2\mu+\lambda)}$

Vérification des comparaisons :

$$\frac{4\lambda\mu}{4\mu^2-\lambda^2}-\frac{\lambda}{2\mu-\lambda}=\frac{\lambda}{2\mu+\lambda}>0\ ;\qquad \frac{\lambda^2}{\mu(4\mu^2-\lambda^2)}-\frac{\lambda}{2\mu(2\mu-\lambda)}=-\frac{\lambda}{2\mu(2\mu+\lambda)}<0\ ;$$

$$\frac{4\mu}{4\mu^2-\lambda^2}-\frac{1}{2\mu-\lambda}=\frac{1}{2\mu+\lambda}>0\ ;\qquad \frac{2\mu^2+\lambda\mu-\lambda^2}{\mu(2\mu+\lambda)}-\Big(1-\frac{\lambda}{2\mu}\Big)=\frac{\lambda}{2\mu}\,\frac{2\mu-\lambda}{2\mu+\lambda}.$$

Commentaire : le temps d'attente pour la file $M(\lambda)/M(2\mu)/1$ est plus long que celui de la file $M(\lambda)/M(\mu)/2$; en revanche le temps de séjour total est plus court. De même, la longueur de la file $M(\lambda)/M(2\mu)/1$ est moins longue que celle de la file $M(\lambda)/M(\mu)/2$. En conséquence, il est préférable de remplacer deux serveurs par un seul serveur deux fois plus efficace.

Exercice 6 (Blocage pour des files en tandem)
On considère un système composé de deux services disposés en série ne comportant pas de salle d'attente entre les deux services. Cela signifie qu'un client reste bloqué à l'issue du premier service, bloquant ainsi tout le système, tant que le client précédent n'a pas terminé son service auprès du deuxième serveur. Les clients se présentent au premier guichet selon un processus de Poisson d'intensité λ et les services suivent des lois exponentielles indépendantes de paramètres μ_1 et μ_2.

1. Calculer la probabilité de blocage du système en régime stationnaire.
2. Quel service est-il préférable de mettre en première position afin d'éviter le blocage : le plus rapide ou le plus lent ?

Solution.

1. Examinons deux clients consécutifs (les $(n-1)^{\text{e}}$ et n^{e}) dans le système. Le n^{e} client démarre son premier service d'une durée $S_n^{(1)}$ après le départ du $(n-1)^{\text{e}}$ client de ce service. Plus précisément, il démarre son premier service
   - immédiatement après le départ du $(n-1)^{\text{e}}$ client vers le deuxième service s'il est présent dans le système à cet instant ;
   - ou durant le deuxième service du $(n-1)^{\text{e}}$ client, ou plus tard, s'il n'est pas encore entré dans le système à l'instant de sortie du $(n-1)^{\text{e}}$ client hors du premier service.

   Le n^{e} client reste bloqué au premier service si à l'issue de celui-ci, le $(n-1)^{\text{e}}$ client est encore au deuxième service, et tout le système est bloqué.

   Notons $S_{n-1}^{(2)*}$ le temps de service restant à fournir par le deuxième serveur auprès du $(n-1)^{\text{e}}$ client à compter du moment où le n^{e} client entre dans le premier service. Le système est bloqué si et seulement si le n^{e} client achève son premier service (de durée $S_n^{(1)}$) avant la fin du deuxième service (de durée $S_{n-1}^{(2)*}$) du $(n-1)^{\text{e}}$ client, comptée depuis l'entrée dans le premier service du n^{e} client ; voir Fig. 18. La probabilité d'un tel blocage est alors donnée par

$$\mathbb{P}(\text{le } n^{\text{e}} \text{ client est bloqué}) = \mathbb{P}(S_{n-1}^{(2)*} > S_n^{(1)}).$$

Faisant appel à la propriété d'absence de mémoire de la loi exponentielle, on peut voir que la v.a. $S_{n-1}^{(2)*}$ suit la même loi exponentielle que $S_{n-1}^{(2)}$, et ainsi

$$\mathbb{P}(S_{n-1}^{(2)*} > S_n^{(1)}) = \int_0^\infty \mathbb{P}(S_{n-1}^{(2)*} > s) f_{S_n^{(1)}}(s)\, ds = \int_0^\infty \mu_1\, e^{-(\mu_1+\mu_2)s}\, ds = \frac{\mu_1}{\mu_1+\mu_2}.$$

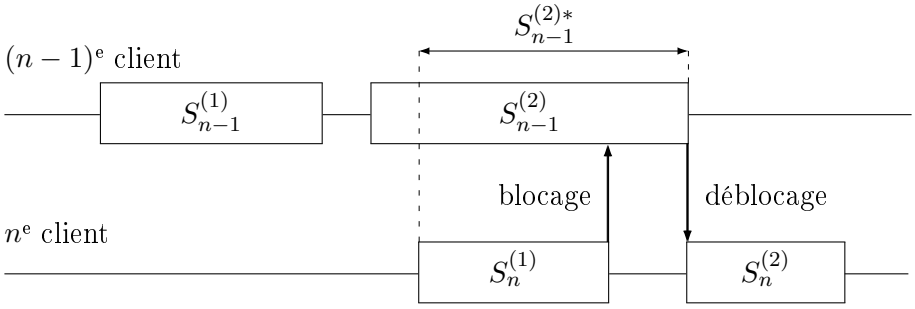


FIGURE 18 – Files en tandem : blocage

2. On voit donc que la probabilité de blocage est proportionnelle au paramètre de la loi exponentielle du premier service, ou encore, inversement proportionnelle au temps moyen du premier service. En conséquence, cette probabilité est minimale lorsque le service le plus lent ($\mu_1 \leqslant \mu_2$) est placé en première position.

Exercice 7 (Monoserveur-multiserveur)
On doit relier deux ordinateurs qui utilisent huit applications parallèles. Chaque application génère un trafic Poissonnien de 2 paquets/s en moyenne. La longueur moyenne des paquets est de 16 kb. On propose deux solutions :

solution 1 : dédier une bande de base 64 kb/s à chaque application (8 monoserveurs $M/M/1$) ;

solution 2 : utiliser une ligne à 512 kb/s pour toutes les applications (1 monoserveur $M/M/1$).

1. Quelle est la solution qui donne le meilleur temps de réponse (*i.e.* temps de séjour dans le système) ?
2. On envisage d'utiliser la solution 1 mais avec une répartition équilibrée entre les serveurs, c'est-à-dire que l'on ouvre 8 serveurs (1 multiserveur $M/M/8$, voir l'annexe 3).
   (a) Quel est le temps de réponse ?
   (b) Quel est le nombre moyen de paquets transitant dans le système ?
   (c) Quel semble être le meilleur dispositif ?

Solution.

1. **Solution 1 :** On a huit files $M/M/1$ indépendantes, voir Fig. 19. Pour chaque file, les paramètres sont $\lambda = 2$ paquets/s et $1/\mu = \frac{16}{64} = 0,25$ s/paquet ou encore $\mu = 4$ paquets/s ; la charge est donc $\rho = \lambda/\mu = 0,5$. Comme $\rho < 1$, il y a un régime stationnaire. D'où le temps d'attente puis le temps de réponse :

$$\mathbb{E}(W_\infty) = \frac{\rho}{\mu(1-\rho)} = 250 \text{ ms},$$
$$\mathbb{E}(W_\infty + S_\infty) = \frac{1}{\mu-\lambda} = 500 \text{ ms}.$$

   En utilisant la formule de Little, on obtient le nombre moyen de paquets transitant dans le système :

$$\mathbb{E}(Q_\infty) = \lambda\, \mathbb{E}(W_\infty + S_\infty) = 2 \times 0,5 = 1 \text{ paquet}.$$

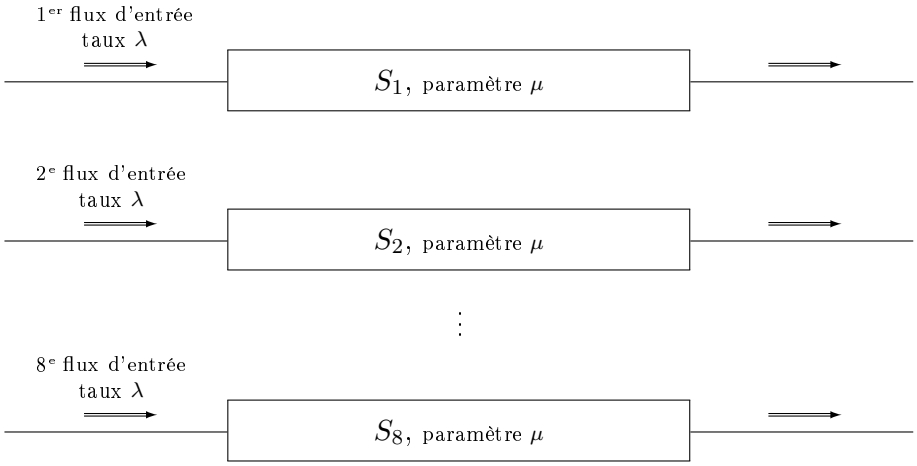


FIGURE 19 – Huit monoserveurs $M(\lambda)/M(\mu)/1$

Solution 2 : On a une file $M/M/1$ de paramètres $\lambda' = 16$ paquets/s $(= 8\lambda)$ et $1/\mu' = \frac{16}{512} = \frac{1}{32}$ s/paquet ou encore $\mu' = 32$ paquets/s $(= 8\mu)$; voir Fig. 20. La charge est toujours $\rho' = \lambda'/\mu' = 0,5 = \rho$. D'où le temps d'attente puis le temps de réponse :

$$\mathbb{E}(W'_\infty) = \frac{\rho'}{\mu'(1-\rho')} = 31,25 \text{ ms},$$
$$\mathbb{E}(W'_\infty + S'_\infty) = \frac{1}{\mu' - \lambda'} = 62,5 \text{ ms}.$$

Avec la formule de Little, le nombre moyen de paquets transitant dans le système est de

$$\mathbb{E}(Q_\infty) = \lambda\,\mathbb{E}(W_\infty + S_\infty) = 16 \times 62,5 = 1 \text{ paquet}.$$

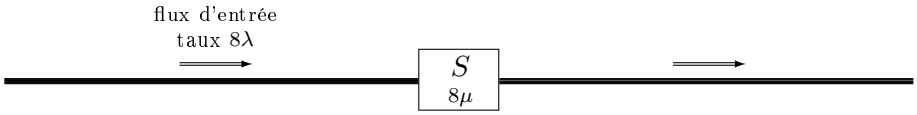


FIGURE 20 – Un monoserveur $M(8\lambda)/M(8\mu)/1$

Le gain de temps est considérable : lorsqu'on partage une voie en n_0 voies parallèles $M(\lambda)/M(\mu)/1$, le temps de réponse est $\dfrac{1}{\mu(1-\rho)} \Big/ \dfrac{1}{\mu'(1-\rho')} = n_0$ fois plus grand que celui pour une unique voie à débit n_0 fois plus important $M(n_0\lambda)/M(n_0\mu)/1$. Ceci peut s'expliquer par le fait que le temps de traitement perdu par chacune des voies parallèles n'est pas récupérable par les autres.

2. En considérant maintenant une file unique mais avec huit serveurs (file $M/M/8$), les caractéristiques sont les suivantes : $\lambda'' = \lambda' = 16$ paquets/s, $1/\mu'' = 1/\mu = 0,25$ s/paquet (voir Fig. 21) ; la charge est cette fois $\rho'' = \lambda''/\mu'' = 4$ $(= 8\rho)$. Comme $\rho < 8$, il y a un régime stationnaire.

   (a) Le temps d'attente et le temps de réponse ont pour valeurs dans ce cas

$$\mathbb{E}(W_\infty) = \frac{\pi_0(\rho'')^8}{8!\,(1-\frac{\rho''}{8})}\,\frac{1}{8\mu'' - \lambda''} = 0,203\,\pi_0$$

avec

$$\pi_0 = \mathbb{P}(Q_\infty = 0) = \left[\sum_{j=0}^{7}\frac{(\rho'')^j}{j!} + \frac{(\rho'')^8}{8!\,(1-\frac{\rho''}{8})}\right]^{-1} = 0,0181,$$

soit

$$\mathbb{E}(W_\infty) = 3,68 \text{ ms}$$

et

$$\mathbb{E}(W_\infty + S_\infty) = 253,68 \text{ ms}.$$

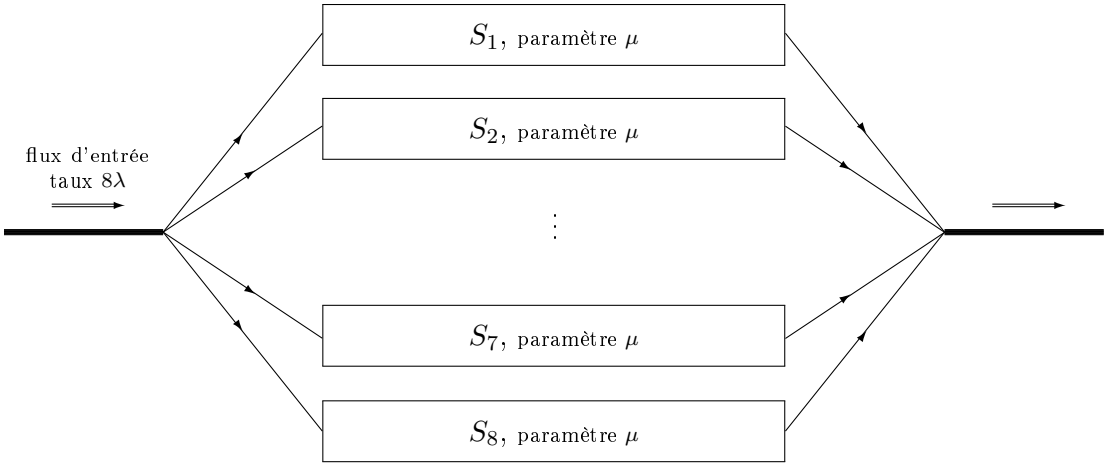


FIGURE 21 – Multiserveur $M(8\lambda)/M(\mu)/8$

(b) La formule de Little fournit le nombre moyen de paquets transitant dans le système :

$$\mathbb{E}(Q_\infty) = \lambda\,\mathbb{E}(W_\infty + S_\infty) = 16 \times 253,68 = 4,05 \text{ paquets.}$$

(c) On gagne énormément au niveau du temps d'attente par rapport aux deux dispositifs de type monoserveurs précédents. Par contre le temps de traitement est minimal dans le cas du monoserveur de la solution 2, ce dernier offrant un service huit fois plus rapide que les autres serveurs. Globalement, le temps de réponse est minimal pour le monoserveur de la solution 2.

Un TP avec Excel

Le but de ce TP est de simuler une file d'attente de type $M/M/1$ (arrivées poissonniennes, service exponentiel, 1 serveur) à l'aide du tableur Excel. On note, pour $n \geqslant 1$,

- T_n le laps de temps séparant les arrivées des $(n-1)^{\text{e}}$ et n^{e} personnes ;
- $s_n = T_1 + \cdots + T_n$ l'instant d'arrivée de la n^{e} personne ;
- S_n le temps de service de la n^{e} personne ;
- W_n le temps d'attente de la n^{e} personne ;
- σ_n l'instant de départ de la n^{e} personne ;
- A_t le nombre d'arrivées durant l'intervalle de temps $[0,t]$;
- D_t le nombre de départs durant l'intervalle de temps $[0,t]$;
- $Q_t = A_t - D_t$ la longueur de la file à l'instant t.

Les v.a. T_n sont i.i.d. de loi $\mathcal{E}(\lambda)$ et les v.a. S_n sont i.i.d. de loi $\mathcal{E}(\mu)$. On illustrera la manipulation sur trois exemples avec un effectif de 100 personnes : $\begin{cases} \lambda = 3 < \mu = 5 \,; \\ \lambda = 5 > \mu = 3 \,; \\ \lambda = \mu = 3. \end{cases}$

1. **Génération des temps inter-arrivées et des temps de service**
   À l'aide du générateur de nombres aléatoires ALEA(), simuler les v.a. exponentielles T_n et S_n. On rappelle que si U_n et V_n sont des nombres aléatoires uniformes sur $[0,1]$ indépendants, on peut facilement simuler des v.a. $\mathcal{E}(\lambda)$ et $\mathcal{E}(\mu)$ en posant $T_n = -\dfrac{1}{\lambda}\ln U_n$ et $S_n = -\dfrac{1}{\mu}\ln V_n$. Consigner dans deux colonnes particulières les valeurs numériques de T_n et S_n à l'aide d'un collage spécial sans les formules.
2. **Calcul des instants d'arrivée et de départ, du temps d'attente**
   Faire calculer successivement les v.a. s_n, W_n, σ_n en utilisant les relations
   (a) $s_1 = T_1, \quad s_{n+1} = s_n + T_{n+1},\ n \geqslant 1$;
   (b) $W_1 = 0, \quad W_{n+1} = \max(W_n + S_n - T_{n+1}, 0),\ n \geqslant 1$;
   (c) $\sigma_n = s_n + W_n + S_n$.
3. **Calcul des temps d'attente et de séjour moyens**
   Faire calculer les moyennes des temps d'attente W_n et des temps de séjour $W_n + S_n$, $n \geqslant 1$, puis dans le cas où $\lambda < \mu$, les comparer à leur valeurs théoriques respectives $\mathbb{E}(W_\infty) = \dfrac{\lambda}{\mu(\mu-\lambda)}$ et $\mathbb{E}(W_\infty + S_\infty) = \dfrac{1}{\mu-\lambda}$.
4. **Tracé des courbes arrivées-départs**
   On souhaite tracer sur une même figure les graphes des fonctions d'arrivées et de départs $t \longmapsto A_t$ et $t \longmapsto D_t$. Ce sont des fonctions en escaliers croissantes avec des sauts de $+1$.
   Pour cela, on doit trier et ordonner les instants d'arrivée et de départ pour les placer correctement sur l'axe des abscisses.
   (a) Introduire à côté de la colonne des instants d'arrivée s_n (resp. des instants de départ σ_n) une colonne de $+1$ (resp. -1). Les $+1$ et -1 servent d'indicateurs pour repérer ultérieurement les instants d'arrivée et de départ une fois ces instants mélangés.
   (b) Superposer les deux doubles-colonnes $(s_n, +1)$ et $(\sigma_n, -1)$ en ne conservant que les valeurs numériques (sans les formules) à l'aide d'un collage spécial, puis effectuer un tri croissant sur l'ensemble de la double-colonne ainsi obtenue, relativement à la première colonne. On trouve une double-colonne de la forme (t_n, ε_n) avec $\varepsilon_n = \pm 1$ et $t_n = \begin{cases} s_n & \text{si } \varepsilon_n = +1 \\ \sigma_n & \text{si } \varepsilon_n = -1 \end{cases}$.
   (c) Rajouter ensuite une colonne contenant les valeurs de A_t. La fonction cumulative $t \longmapsto A_t$ saute de $+1$ à chaque instant s_n et reste constante en chaque σ_n (saut nul). Ainsi, à chaque instant s_n ou σ_n, le saut vaut $+1$ ou 0 et peut s'écrire à l'aide des indicateurs $\varepsilon_n = \pm 1$ selon $(1+\varepsilon_n)/2$; en effet, $(1+\varepsilon_n)/2 = \begin{cases} +1 & \text{si } \varepsilon_n = +1 \\ 0 & \text{si } \varepsilon_n = -1 \end{cases}$.
   (d) Rajouter selon le même procédé une colonne contenant les valeurs de D_t. Les sauts sont ici de $(1-\varepsilon_n)/2$.

(e) Pour construire les « marches d'escalier » point par point, il faut effectuer une translation (voir Fig. 22). Superposer à cet effet les deux triples-colonnes $(t_n, A_{t_{n-1}}, D_{t_{n-1}})$, $n \geqslant 0$ et $(t_n, A_{t_n}, D_{t_n}), n \geqslant 1$ avec $t_{-1} = t_0 = 0$ en ne conservant que les valeurs numériques (sans les formules) à l'aide d'un collage spécial. Effectuer un tri croissant sur l'ensemble de la triple-colonne ainsi obtenue, relativement aux première, deuxième, troisième colonnes dans ces ordres de priorité.

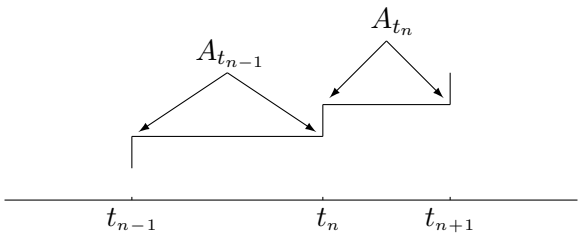


FIGURE 22 – Marches d'escalier : $(t_{n-1}, A_{t_{n-1}})$, (t_{n-1}, A_{t_n}), (t_n, A_{t_n}), (t_{n+1}, A_{t_n}), $(t_{n+1}, A_{t_{n+1}})$...

(f) Relier enfin les points ainsi obtenus à l'aide d'un nuage de points à lignes droites. On obtient sur un même graphique les courbes $t \longmapsto A_t$ et $t \longmapsto D_t$. Les t_n servent de graduations en abscisses communes aux deux fonctions précédentes.

5. Tracé de la courbe longueur de la file

On souhaite enfin tracer le graphe de la longueur de la file $t \longmapsto Q_t$. C'est une fonction en escaliers avec des sauts de $+1$ et -1.

(a) Rajouter en face de la dernière triple colonne construite en 4.(e) — formant les marches d'escalier des deux fonctions $t \longmapsto A_t$ et $t \longmapsto D_t$ — une colonne « longueur de la file » en utilisant la formule $Q_t = A_t - D_t$.

(b) Copier les valeurs numériques des première et quatrième colonnes de la quadruple-colonne précédente. En reliant les points de la double-colonne ainsi obtenue à l'aide d'un nuage de points, on récupère la courbe $t \longmapsto Q_t$.

(c) Faire calculer la longueur moyenne de la file

$$\frac{1}{\sigma_{100}} \int_0^{\sigma_{100}} Q_t \, dt = \frac{1}{t_{200}} \sum_{n=0}^{n=199} (t_{n+1} - t_n) \, Q_{t_n},$$

puis la comparer, dans le cas où $\lambda < \mu$, à sa valeur théorique $\mathbb{E}(Q_\infty) = \dfrac{\lambda}{\mu - \lambda}$.

(d) Construire l'histogramme de la longueur de la file. (On recherchera d'abord son maximum.)

6. **Calcul du temps d'activité du serveur**

On souhaite évaluer le temps moyen d'activité du serveur. Il est plus simple de calculer d'abord son temps moyen d'inactivité.

(a) Introduire une colonne fournissant les temps d'inactivité (écart entre un départ laissant la file vide et l'arrivée suivante) à l'aide de la fonction SI. Ces laps de temps se rencontrent lors de l'apparition de zéros dans la colonne donnant la longueur de la file : si $Q_{t_n} = 0$, le serveur rencontre une période de repos de durée $t_{n+1} - t_n$; voir Fig. 23. (Note : la colonne fournissant les marches d'escalier de la fonction longueur contient en fait des valeurs dédoublées. On voit donc apparaître des successions de deux zéros consécutifs et la période de repos est comprise entre deux tels zéros. La durée de cette période est alors la différence des deux instants correspondants.)

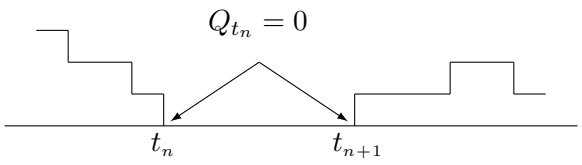


FIGURE 23 – Période d'inactivité du serveur

(b) Faire calculer la moyenne des temps d'inactivité, puis celle des temps d'activité et les comparer, dans le cas où $\lambda < \mu$, à leurs valeurs théoriques respectives $\mathbb{E}(I_\infty) = \dfrac{1}{\lambda}$ et $\mathbb{E}(B_\infty) = \dfrac{1}{\mu - \lambda}$.

Agner Krarup Erlang

Born: 1 January 1878 in Lønborg, Denmark
Died: 3 February 1929 in Copenhagen, Denmark

Agner Erlang was descended on his mother's side from Thomas Fincke. His father was a schoolmaster and Erlang was educated at his father's school when he was young. He took his examinations in Copenhagen at the age of 14 and passed with special distinction after having to obtain special permission to take the examinations because he was below the minimum age.

He returned to Lønberg and taught at his father's school for two years. In 1896 he passed the entrance examination to the University of Copenhagen with distinction and, since his parents were poor, he was given free board and lodgings in a College of the University of Copenhagen.

His studies at Copenhagen were in mathematics and natural science. He attended the mathematics lectures of Zeuthen and Juel and these gave him an interest in geometrical problems which were to remain with him all his life.

After graduating in 1901 with mathematics as his major subject and physics, astronomy and chemistry as secondary subjects, he taught in schools for several years. During this time he kept up his interest in mathematics, and he received an award for an essay on Huygens solution of infinitesimal problems which he submitted to the University of Copenhagen.

His interests turned towards the theory of probability and he kept up his mathematical interests by joining the Mathematical Association. At meetings of the Mathematical Association he met Jensen who was then the chief engineer at the Copenhagen Telephone Company. He persuaded Erlang to apply his skills to the solution of problems which arose from a study of waiting times for telephone calls.

In 1908 Erlang joined the Copenhagen Telephone Company and began applying probability to various problems arising in the context of telephone calls. He published his first paper on these problems *The theory of probability and telephone conversations* in 1909. In 1917 he gave a formula for loss and waiting time which was soon used by telephone companies in many countries including the British Post Office.

In addition to his work on probability Erlang was also interested in mathematical tables. This interest is described as :—

> *A subject that interested Erlang very much was the calculation and arrangement of numerical tables of mathematical functions, and he had an uncommonly thorough knowledge of the history of mathematical tables from ancient times right up to the present. Erlang set forth a new principle for the calculation of certain forms of mathematical tables, especially tables of logarithms...*

Article by: *J.J. O'Connor* and *E.F. Robertson*

School of Mathematics and Statistics
University of St Andrews, Scotland

Andrei Andreyevich Markov

Born: 14 June 1856 in Ryazan, Russia
Died: 20 July 1922 in Petrograd (now St Petersburg), Russia

Andrei Andreyevich Markov's mother was Nadezhda Petrovna, who was the daughter of a state worker, and his father was Andrei Grigorievich Markov, the son of a country deacon. Andrei Grigorievich Markov studied at a church seminary, then got a job as a clerk. The family moved to St Petersburg where Andrei Grigorievich served in the Forestry Department and then became a manager of various households and estates. Andrei Grigorievich married twice ; with his first wife Nadezhda he had two sons and several daughters. Andrei Andreyevich was the oldest of the two boys while the younger was Vladimir. Although Vladimir died from tuberculosis at the age of 25, he had already gained an international reputation as a mathematician.

In his early years Markov was in poor health and up to the age of ten he could only walk with the assistance of crutches. His secondary schooling was at St Petersburg Gymnasium No 5 where he showed outstanding talents for mathematics but performed rather poorly in other subjects. He wrote his first mathematics paper while at the Gymnasium but his results on integration of linear differential equations which were presented in the paper were not new. However, writing the paper did result in him meeting Korkin and Zolotarev, two of the leading professors at the university. It was clear that mathematics was the right subject for Markov to study at university and, in 1874, he entered the Physics and Mathematics Faculty of St Petersburg University. There he enrolled in the seminar run by Korkin and Zolotarev but also attended lectures by Chebyshev, the head of the mathematics department. These were particularly stimulating to Markov, since Chebyshev often encouraged an atmosphere of research by posing new questions and problems for his students to investigate.

Markov graduated in 1878 having won the gold medal for submitting the best essay for the prize topic set by the faculty in that year–*On the integration of differential equations by means of continued fractions*. He now wished to train to become a university professor and worked for his Master's degree over the next two years (this was at a level equivalent to a doctorate). He was awarded the degree in 1880 for his thesis *On the binary quadratic forms with positive determinant*. This thesis was outstanding :—

> *This work, very highly esteemed by Chebyshev, represents one of the finest achievements of the St Petersburg school of number theory, and perhaps even of all Russian mathematics. It is enough to recall the sorts of questions in the field of rational approximation which at that time preoccupied the most prominent number theorists of France and Germany, to appreciate how much deeper into the field Markov had penetrated. It is therefore perhaps not surprising that, although the dissertation was published immediately (in French in Mathematische Annalen), it did not become generally absorbed by west European mathematicians, until from* 1910 *to the* 1920s *the Berlin mathematicians Frobenius and Remak attempted to master the set of ideas contained in Markov's work.*

After submitting his master's thesis, Markov began to teach at St Petersburg University as a privatdozent while working for his doctorate (equivalent to the habilitation). He was awarded his doctorate in 1884 for his dissertation *On certain applications of continued fractions*.

Markov had known Maria Ivanova Valvatyeva since they were children for she was the daughter of the owner of the estate which his father was managing. Markov had tutored Maria Ivanova

in mathematics and later he proposed marriage to her. However Maria Ivanova's mother would not allow her daughter to marry the son of her estate manager until Markov had gained sufficient social status. In 1883 Maria Ivanova's mother agreed to the marriage which took place in that year.

Markov became an extraordinary professor at St Petersburg University in 1886 and an ordinary professor in 1893. Chebyshev proposed Markov as an adjunct of the Russian Academy of Sciences in 1886. He was elected as an extraordinary member in 1890 and an ordinary academician in 1896. He formally retired in 1905 but continued to teach for most of his life.

Markov's early work was mainly in number theory and analysis, algebraic continued fractions, limits of integrals, approximation theory and the convergence of series. After 1900 Markov applied the method of continued fractions, pioneered by his teacher Pafnuty Chebyshev, to probability theory :—

> *Markov was the most elegant spokesman for Chebyshev's ideas and directions of research in probability theory. Especially remarkable is his research relating to the theorem of Jacob Bernoulli known as the Law of Large Numbers, to two fundamental theorems of probability theory due to Chebyshev, and to the method of least squares.*

He also studied sequences of mutually dependent variables, hoping to establish the limiting laws of probability in their most general form. He proved the central limit theorem under fairly general assumptions. Markov is particularly remembered for his study of Markov chains, sequences of random variables in which the future variable is determined by the present variable but is independent of the way in which the present state arose from its predecessors. This work founded a completely new branch of probability theory and launched the theory of stochastic processes. In 1923 Norbert Wiener became the first to treat rigorously a continuous Markov process. The foundation of a general theory was provided during the 1930s by Andrei Kolmogorov.

Sergei Bernstein, who continued to develop the theory of Markov chains, wrote :—

> *A Markov's classic course on the computation of probabilities, and his original memoirs, models of accuracy and clarity of exposition, contributed to a very large extent to the transformation of the theory of probability into one of the most perfected areas of mathematics, and to the wide dissemination of Chebyshev's methods and directions of research. His profound analysis in the spirit of Chebyshev of the dependencies among observed random phenomena allowed Markov to extend probability theory in an essential way through the introduction and investigation of dependent random quantities.*

Markov was also interested in poetry and he made studies of poetic style–perhaps surprisingly Kolmogorov had similar interests. It is worth pointing out, however, that although Markov developed his theory of Markov chains as a purely mathematical work without considering physical applications, he did apply the ideas to chains of two states, namely vowels and consonants, in literary texts. His interest in poetry was not, therefore, an entirely separate interest from his mathematical work.

As a lecturer, Markov demanded much of his students :—

> *His lectures were distinguished by an irreproachable strictness of argument, and he developed in his students that mathematical cast of mind that takes nothing for granted. He included in his courses many recent results of investigations, while often omitting traditional questions. The lectures were difficult, and only serious students could understand them... During his lectures he did not bother about the order of equations on the blackboard, nor about his personal appearance.*

Markov lived through a period of great political activity in Russia and, having firm opinions, he became heavily involved. Maksim Gorky, the Russian short-story writer, novelist and left wing activist, was elected a member of the Russian Academy of Sciences in 1902, but his election was soon withdrawn for political reasons on the Tsar's orders. Markov protested strongly and refused to accept honours awarded him on the following year. In June 1907 Tsar Nicholas dissolved the Second Duma which had been elected with majority on the left. Markov repudiated his membership and might have expected to suffer severe consequences but the authorities chose not

to make an example of an elderly and distinguished academician. In 1913 the Romanov dynasty, which had been in power in Russia since 1613, celebrated their 300 years of power. This was not likely to improve their already weak position. Markov showed his disapproval of the celebration but holding celebrations of his own–he celebrated 200 years of the Law of Large Numbers! The Russian Revolution began early in 1917 as food supplies ran low. In September of that year Markov requested the Academy to send him to a disadvantaged town in the Russian interior. He was sent to Zaraisk, a small country town, where he taught mathematics in the secondary school without receiving any remuneration. He returned to St Petersburg but his health was now deteriorating and he had an eye operation. Although by 1921 he was in such a bad way that he was hardly able to stand, yet he continued to lecture on probability at the university. His death in July 1922 came after months of the most severe suffering.

Markov had a son (of the same name) who was born on 9 September 1903 and followed his father in also becoming a renowned mathematician.

Article by: *J.J. O'Connor* and *E.F. Robertson*

School of Mathematics and Statistics
University of St Andrews, Scotland

Siméon Denis Poisson

Born: 21 June 1781 in Pithiviers, France
Died: 25 April 1840 in Sceaux, France

Siméon-Denis Poisson's parents were not from the nobility and, although it was becoming increasingly difficult to distinguish between the nobility and the bourgeoisie in France in the years prior to the Revolution, nevertheless the French class system still had a major influence on his early years. The main reason for this was that the army was one of the few occupations where the nobility enjoyed significant institutional privileges and Poisson's father had been a soldier. Certainly Poisson's father was discriminated against by the nobility in the upper ranks of the army and this made a large impression on him. After retiring from active service he was appointed to a lowly administrative post which he held at the time that his son Siméon-Denis was born. There is no doubt that Siméon-Denis's family put a great deal of energy into helping him have a good start in life.

Now Siméon-Denis was not the first of his parents children but several of his older brothers and sisters had failed to survive. Indeed his health was also very fragile as a child and he was fortunate to pull through. This may have been because his mother, fearing that her young child would die, entrusted him to the care of a nurse to bring him through the critical period. His father had a large influence on his young son, devoting time to teach him to read and write.

Siméon-Denis was eight years old when the Parisian insurrection of 14 July 1789 heralded the start of the French Revolution. As might be expected of someone who had suffered discrimination at the hands of the nobility, Poisson senior was enthusiastic about the political turn of events. One immediate consequence for his support of the Revolution was the fact that he became president of the district of Pithiviers which is in central France, about 80 km south of Paris. From this position he was able to influence the future career of his son.

Poisson's father decided that the medical profession would provide a secure future for his son. An uncle of Poisson's was a surgeon in Fontainebleau and Poisson was sent there to become an apprentice surgeon. However, Poisson found that he was ill suited to be a surgeon. Firstly he lacked coordination to quite a large degree which meant that he completely failed to master the delicate movements required. Secondly it was quickly evident that, although he was a talented child, he had no interest in the medical profession. Poisson returned home from Fontainebleau having essentially failed to make the grade in his apprenticeship and his father had to think again to find a career for him.

Times were changing quite quickly in France which was by this time a republic. No longer were certain professions controlled by the nobility as they had been and there had been moves towards making education available to everyone. In 1796 Poisson was sent back to Fontainebleau by his father, this time to enrol in the École Centrale there. On the one hand he had shown a great lack of manual dexterity, but he now showed that he had great talents for learning, especially mathematics. His teachers at the École Centrale were extremely impressed and encouraged him to sit the entrance examinations for the École Polytechnique in Paris. He sat these examinations and proved his teachers right, for although he had far less formal education than most of the young men taking the examinations he achieved the top place.

Few people can have achieved academic success as quickly as Poisson did. When he began to study mathematics in 1798 at the École Polytechnique he was therefore in a strong position to cope with the rigours of a hard course, yet overcome the deficiencies of his early education. There were certainly problems for him to overcome for he had little experience of the social or academic

environment into which he was suddenly thrust. It was therefore to his credit that he was able to undertake his academic studies with great enthusiasm and diligence, yet find time to enjoy the theatre and other social activities in Paris. His only weakness was the lack of coordination which had made a career as a surgeon impossible. This was still a drawback to him in some respects for drawing mathematical diagrams was quite beyond him.

His teachers Laplace and Lagrange quickly saw his mathematical talents. They were to become friends for life with their extremely able young student and they gave him strong support in a variety of ways. A memoir on finite differences, written when Poisson was 18, attracted the attention of Legendre. However, Poisson found that descriptive geometry, an important topic at the École Polytechnique because of Monge, was impossible for him to succeed with because of his inability to draw diagrams. This would have been an insurmountable problem had he been going into public service, but those aiming at a career in pure science could be excused the drawing requirements, and Poisson was not held back. In his final year of study he wrote a paper on the theory of equations and Bezout's theorem, and this was of such quality that he was allowed to graduate in 1800 without taking the final examination. He proceeded immediately to the position of répétiteur in the École Polytechnique, mainly on the strong recommendation of Laplace. It was quite unusual for anyone to gain their first appointment in Paris, most of the top mathematicians having to serve in the provinces before returning to Paris.

Poisson was named deputy professor at the École Polytechnique in 1802, a position he held until 1806 when he was appointed to the professorship at the École Polytechnique which Fourier had vacated when he had been sent by Napoleon to Grenoble. In fact Poisson had little time for politics for rather his whole energies were directed to support mathematics, science, education and the École Polytechnique. When the students at the École had been about to publish an attack on Napoleon's ideas for the Grand Empire in 1804, Poisson had managed to stop them, not because he supported Napoleon's views but rather because he saw that the students would damage the École Polytechnique by their actions. Poisson's motives were not understood by Napoleon's administration, however, and they saw Poisson as a supporter which did his career no harm at all.

During this period Poisson studied problems relating to ordinary differential equations and partial differential equations. In particular he studied applications to a number of physical problems such as the pendulum in a resisting medium and the theory of sound. His studies were purely theoretical, however, for as we mentioned above, he was extremely clumsy with his hands:—

> *Poisson ... was content to remain totally unfamiliar with the vicissitudes of experimental research. It is quite unlikely that he ever attempted an experimental measurement, nor did he try his hand at drafting experimental designs.*

His first attempt to be elected to the Institute was in 1806 when he was backed by Laplace, Lagrange, Lacroix, Legendre and Biot for a place in the Mathematics Section. Bossut was 76 years old at the time and, had he died, Poisson would have gained a place. However Bossut lived for another seven years so there was no route into the mathematics section for Poisson. He did, however, gain further prestigious posts. In addition to his professorship at the École Polytechnique, in 1808 Poisson became an astronomer at Bureau des Longitudes. In 1809 he added another appointment, namely that of the chair of mechanics in the newly opened Faculté des Sciences.

In 1808 and 1809 Poisson published three important papers with the Academy of Sciences. In the first *Sur les inégalités des moyens mouvement des planètes* he looked at the mathematical problems which Laplace and Lagrange had raised about perturbations of the planets. His approach to these problems was to use series expansions to derive approximate solutions. This was typical of the type of problem which he found interesting. Libri wrote:—

> *...he especially liked unresolved questions that had been treated by others or areas in which there was still work to be done.*

In 1809 he published two papers, the first *Sur le mouvement de rotation de la terre* and the second, *Sur la variation des constantes arbitraires dans les questions de mécanique* was a direct consequence of developments in Lagrange's method of variation of arbitrary constants which had been inspired by Poisson's 1808 paper. In addition he published a new edition of Clairaut's

Théorie de la figure de la terre in 1808. The work had been first published by Clairaut in 1743 and it confirmed the Newton-Huygens belief that the Earth was flattened at the poles. In 1811 Poisson published his two volume treatise *Traité de mécanique* which was an exceptionally clear treatment based on his course notes at the École Polytechnique.

Malus was known to have a terminal illness by 1811 and his death would leave a vacancy in the physics section of the Institute. The mathematicians, aiming to have Poisson fill that vacancy when it occurred, set the topic for the Grand Prix on electricity so as to maximise Poisson's chances. The topic for the prize was as follows :—

> *To determine by calculation and to confirm by experiment the manner in which electricity is distributed at the surface of electrical bodies considered either in isolation or in the presence of each other–for example at the surface of two electrified spheres in the presence of each other. In order to simplify the problem, the Class asks only for an examination of cases where the electricity spread on each surface remains always of the same kind.*

Poisson had made considerable progress with the problem before Malus died on 24 February 1812. Poisson submitted the first part of his solution to the Academy on 9 March entitled *Sur la distribution de l'électricité à la surface des corps conducteurs.* As the mathematicians had intended, this was the deciding factor in Poisson being elected to the physics section of the Institute to replace Malus. It also marked a move away from experimental research towards theoretical research in what was considered to constitute physics, and in this the Institute was following the lead given by Laplace.

Poisson continued to add various responsibilities to his already busy life. In 1815 he became examiner for the École Militaire and in the following year he became an examiner for the final examinations at the École Polytechnique.

It is remarkable how much work Poisson put in ; to his research, to his teaching and to playing an ever increasingly important role in the organisation of mathematics in France. When he married Nancy de Bardi in 1817 he found that family life put yet another pressure on him yet somehow he survived the pressures continuing to take on further duties. His research contributions covered a wide range of applied mathematics topics. Although he devised no innovative new theories, he made major contributions to further developing the theories of others often being the first to exhibit their real significance. We mention now just a few of the topics he studied after his election to the Academy.

In 1813 Poisson studied the potential in the interior of attracting masses, producing results which would find application in electrostatics. He produced major work on electricity and magnetism, followed by work on elastic surfaces. Papers followed on the velocity of sound in gasses, on the propagation of heat, and on elastic vibrations. In 1815 he published a work on heat which annoyed Fourier who wrote :—

> *Poisson has too much talent to apply it to the work of others. to use it to discover what is already know is to waste it...*

Fourier went on to make valid objections to Poisson's arguments which he corrected in later memoirs of 1820 and 1821.

In 1823 Poisson published on heat, producing results which influenced Sadi Carnot. Much of Poisson's work was motivated by results of Laplace, in particular his work on the relative velocity of sound and his work on attractive forces. This latter work was not only influenced by Laplace's work but also by the earlier contributions of Ivory. Poisson's work on attractive forces was itself a major influence on Green's major paper of 1828 although Poisson never seems to have discovered that Green was inspired by his formulations.

In *Recherches sur la probabilité des jugements en matière criminelle et matière civile,* an important work on probability published in 1837, the Poisson distribution first appears. The Poisson distribution describes the probability that a random event will occur in a time or space interval under the conditions that the probability of the event occurring is very small, but the number of trials is very large so that the event actually occurs a few times. He also introduced the expression "law of large numbers". Although we now rate this work as of great importance,

it found little favour at the time, the exception being in Russia where Chebyshev developed his ideas.

It is interesting that Poisson did not exhibit the chauvinistic attitude of many scientists of his day. Lagrange and Laplace recognised Fermat as the inventor of the differential and integral calculus ; he was French after all while neither Leibniz nor Newton were ! Poisson, however, wrote in 1831 :—

> *This* [*differential and integral*] *calculus consists in a collection of rules ... rather than in the use of infinitely small quantities ... and in this regard its creation does not predate* Leibniz, *the author of the algorithm and of the notation that has generally prevailed.*

He published between 300 and 400 mathematical works in all. Despite this exceptionally large output, he worked on one topic at a time. Libri writes :—

> *Poisson never wished to occupy himself with two things at the same time ; when, in the course of his labours, a research project crossed his mind that did not form any immediate connection with what he was doing at the time, he contented himself with writing a few words in his little wallet. The persons to whom he used to communicate his scientific ideas know that as soon as he had finished one memoir, he passed without interruption to another subject, and that he customarily selected from his wallet the questions with which he should occupy himself. To foresee beforehand in this manner the problems that offer some chance of success, and to be able to wait before applying oneself to them, is to show proof of a mind both penetrating and methodical.*

Poisson's name is attached to a wide variety of ideas, for example :— Poisson's integral, Poisson's equation in potential theory, Poisson brackets in differential equations, Poisson's ratio in elasticity, and Poisson's constant in electricity. However, he was not highly regarded by other French mathematicians either during his lifetime or after his death. His reputation was guaranteed by the esteem that he was held in by foreign mathematicians who seemed more able than his own colleagues to recognise the importance of his ideas. Poisson himself was completely dedicated to mathematics. Arago reported that Poisson frequently said :—

> *Life is good for only two things, discovering mathematics and teaching mathematics.*

Article by: *J.J. O'Connor* and *E.F. Robertson*

School of Mathematics and Statistics
University of St Andrews, Scotland

The URL of this page is
http://www-history.mcs.st-andrews.ac.uk/Biographies/Poisson.html