Gestion des files d'attentes

A. Bazeniar

Enseignant chercheur
Centre universitaire Abdelhafid Boussouf
Mila, Algérie

Se reporter à des manuels de base et à certaines livres de spécialités

Septembre 2024

Chapitre 2

Généralités sur les phénomènes d'attente

Introduction

La théorie des files d'attente utilise divers concepts mathématiques pour modéliser et analyser les processus d'arrivée et de départs des clients. Ces concepts permettent de décrire la façon dont les clients arrivent et sortent dans le système, leur fréquence et leur répartition dans le temps.

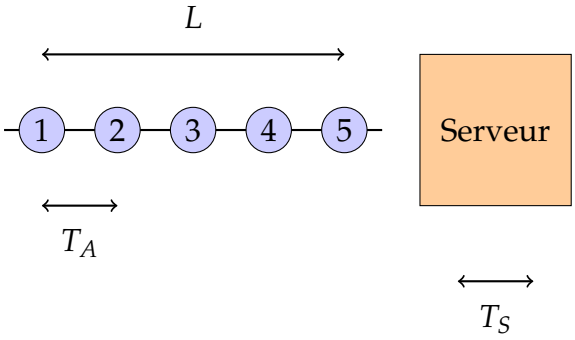

FIGURE 2.1 – Configuration de file d'attente.

2.0.1 Historique

Les étudiants qui sont intéressés par cette section peuvent consulter :

1. U. N. Bhat, *An Introduction to Queueing Theory : Modeling and Analysis in Applications*, Birkhäuser, 2015.
2. L. Kleinrock, *Queueing Systems, Volume I : Theory*, Wiley-Interscience, 1975.
3. D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, Wiley-Interscience, 1998.

La théorie des files d'attente trouve ses origines au début du XXe siècle, principalement grâce aux travaux pionniers du mathématicien danois **Agner Krarup Erlang** (1878–1929). Ce dernier est souvent considéré comme le père fondateur de la théorie des files d'attente.

Les travaux d'Erlang

Alors que Erlang travaillait pour la compagnie *Copenhagen Telephone Company*. Il a réussi à développer cette théorie dans le contexte des systèmes téléphoniques en 1909. Le problème qui la affronter Erlang est que les compagnies, de l'époque, de télécommunication faisaient face à des problèmes de congestion des lignes téléphoniques. Il était crucial de savoir combien d'opérateurs et de lignes téléphoniques étaient nécessaires pour répondre aux demandes d'appels entrants, tout en optimisant les coûts. Erlang a proposé de modéliser les flux d'appels entrants et la durée des appels à l'aide de probabilités.

L'une des premières contributions majeures d'Erlang fut la **loi d'Erlang**, qui modélise le nombre d'appels qui peuvent être servis par un certain nombre de lignes téléphoniques sans provoquer de saturation. C'est dans ce cadre qu'Erlang a formulé le modèle **M/M/1**, où :

— Les arrivées suivent un processus de Poisson.
— Les temps de service sont exponentiels.
— Il y a un seul serveur.

Erlang à continuer de définir des formules importantes pour la gestion des files d'attente, comme la **formule d'Erlang B** et la **formule d'Erlang C**; la première : permet de calculer la probabilité de rejet d'un appel dans un système de lignes téléphoniques occupées; la deuxième : utilisée pour calculer la probabilité qu'un appel soit mis en attente dans un système à plusieurs serveurs.

Développements au XXe siècle

Après les travaux pionniers d'Erlang, la théorie des files d'attente a rapidement attiré l'attention des chercheurs et des ingénieurs dans divers domaines. Durant les années 1930, des chercheurs comme **Felix Pollaczek** et **Andrey Kolmogorov** ont étendu les travaux d'Erlang en développant des modèles mathématiques plus généraux pour les files d'attente. Ces travaux ont permis d'inclure des temps de service aléatoires non nécessairement exponentiels et des processus d'arrivée plus complexes.

Puis dans années 1950 et grâce à l'essor des ordinateurs et des technologies de l'information, le mathématicien britannique **David George Kendall** a introduit la célèbre notation **Kendall** (A/B/c), qui permet de décrire les différents systèmes de files d'attente :

— A décrit le processus d'arrivée.
— B décrit le processus de service.
— c est le nombre de serveurs dans le système.

Applications modernes et avancées

Avec l'avènement des réseaux de communication modernes et de l'informatique dans la seconde moitié du XXe siècle, la théorie des files d'attente a trouvé de nombreuses applications pratiques dans des domaines tels que :

— **Les réseaux de télécommunication** : Optimisation du trafic de données, gestion des paquets d'information, équilibrage de charge dans les réseaux.
— **L'informatique et les systèmes distribués** : Gestion des tâches dans les serveurs, modélisation des temps d'attente dans les systèmes d'exploitation, gestion des bases de données et des serveurs web.
— **La production industrielle** : Optimisation des chaînes de production, gestion des stocks, et réduction des temps d'attente pour les clients.
— **Les services publics** : Gestion des files d'attente dans les hôpitaux, les banques, les transports publics et les centres d'appels.

Dans les années 1970 et 1980, des chercheurs comme **Leonard Kleinrock** ont développé des modèles plus sophistiqués adaptés aux réseaux de communication numériques, qui ont révolutionné la gestion des réseaux informatiques.

2.1 Éléments clés d'une file d'attente

Une file d'attente est décrite par plusieurs composants essentiels, permettant de comprendre et de modéliser son fonctionnement. Ces composants sont :

2.1.1 Processus d'arrivée des clients

Le processus de Poisson est le modèle le plus utilisé pour décrire les arrivées dans une file d'attente, surtout lorsque les événements d'arrivée sont indépendants. Voici quelques caractéristiques clés :

— **Indépendance** : Les arrivées sont indépendantes les unes des autres.
— **Taux constant (λ)** : Les clients arrivent à un taux constant, défini par le paramètre λ.
— **Distribution exponentielle des temps d'arrivée** : Le temps entre deux arrivées consécutives suit une distribution exponentielle de paramètre λ.

La fonction de densité de probabilité (f.d.p) est donnée par :

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0 \tag{2.1}$$

Arrivées déterministes

Dans un modèle d'arrivées déterministes, les clients ou entités arrivent dans la file d'attente à des intervalles de temps fixes. Cela signifie qu'il existe une période de temps constante, appelée période inter-arrivée ou intervalle de temps d'arrivée (souvent notée Δt), entre deux arrivées successives.
Dans un modèle d'arrivées déterministes, le temps d'arrivée de chaque client est donné par une formule simple :

$$t_n = n \cdot \Delta t. \tag{2.2}$$

Avec, t_n est le temps d'arrivée du n-ième client, Δt est l'intervalle de temps fixe entre deux arrivées consécutives, n est le numéro du client.

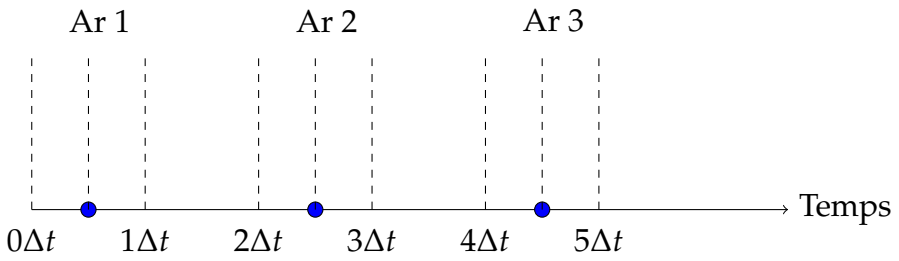

FIGURE 2.2 – Subdivision du Temps en Intervalles Δt.

Arrivées groupées (par lots)

Dans certains cas, les clients arrivent en groupes ou par lots. Ce processus peut être modélisé par :

— **Loi de Poisson par lots** : Les clients arrivent en groupes dont le nombre suit une loi de Poisson.

Dans un processus de Poisson avec un taux d'arrivée λ, le temps entre deux arrivées successives suit une loi exponentielle de paramètre λ. La fonction de densité de probabilité pour le temps d'attente T est :

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0.$$

La fonction de répartition est :

$$F_T(t) = P(T \leq t) = 1 - e^{-\lambda t}.$$

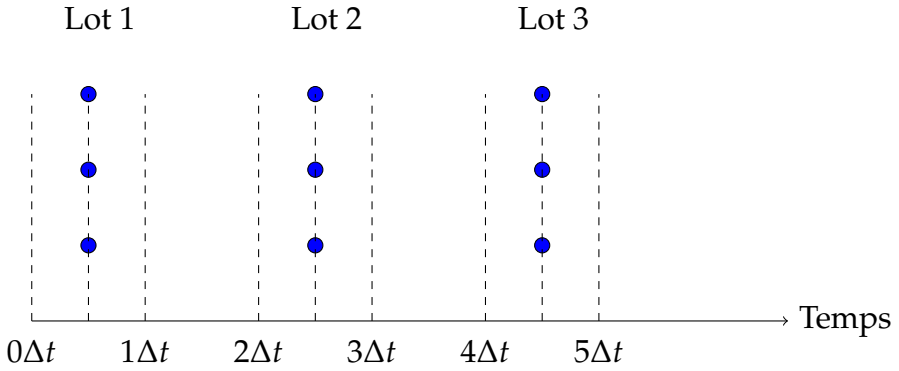

FIGURE 2.3 – Arrivées en Lots dans des Intervalles Δt.

Arrivées selon des distributions aléatoires

D'autres distributions peuvent également être appliquées, comme :

— **Arrivées selon la loi de Gibbs** : Modéliser des phénomènes où les clients arrivent en fonction de l'état du système. La loi de Gibbs peut être exprimée par la formule suivante :

$$P(X = x) = \frac{1}{Z(\beta)} e^{-\beta E(x)}, \tag{2.3}$$

où :

— $Z(\beta)$ est la fonction de partition,

— $E(x)$ est l'énergie associée à l'état x,

— β est un paramètre lié à la température.

— **Arrivées selon la loi de Weibull** : La distribution de Weibull est utilisée pour modéliser les temps de défaillance et peut également s'appliquer aux temps d'arrivée (où k est un paramètre de forme et λ est un paramètre d'échelle.) :

$$f(t) = \frac{\lambda^k}{k!} \left(\frac{\lambda t}{\lambda} \right)^{k-1} e^{-\left(\frac{t}{\lambda}\right)^k} \quad (t \geq 0). \tag{2.4}$$

La modélisation du processus d'arrivée est essentielle pour effectuer des analyses quantitatives. Voici quelques concepts mathématiques clés :

— **Intensité d'arrivée** : L'intensité $\lambda(t)$ peut varier dans le temps.

— **Fonction de distribution des arrivées** : La fonction de distribution cumulative des temps d'arrivée peut être utilisée pour évaluer la probabilité que le temps d'attente soit inférieur à une certaine valeur.

2.1.2 Instant d'arrivée de la n-ième personne

— Soit $s_n = \inf\{t > 0 : N(t) = n\}$ l'instant d'arrivée de la n-ième personne. La condition $s_n > t$ signifie que la n-ième personne arrive après l'instant t. Par conséquent, cela indique qu'il y a eu moins de n arrivées durant l'intervalle de temps $[0, t]$. En d'autres termes, nous avons :

$$P(s_n > t) = P(N(t) < n).$$

Cela correspond à la probabilité d'observer moins de n arrivées au cours de l'intervalle de temps $[0, t]$. En utilisant la loi de Poisson, nous pouvons exprimer cette probabilité comme suit :

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

— La fonction de répartition de s_n, notée $F_{s_n}(t)$, est donnée par :

$$F_{s_n}(t) = P(s_n \leq t) = 1 - P(s_n > t).$$

En substituant l'expression de $P(s_n > t)$, nous obtenons :

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

— La densité de s_n, notée $f_{s_n}(t)$, s'obtient en dérivant l'expression de la fonction de répartition $F_{s_n}(t)$:

$$f_{s_n}(t) = F'_{s_n}(t).$$

En appliquant la règle de Leibniz et la dérivation, on a :

$$\begin{aligned} f_{s_n}(t) &= \lambda e^{-\lambda t} - \sum_{k=1}^{n-1} \left[\frac{\lambda^k t^{k-1}}{(k-1)!} - \frac{\lambda^{k+1} t^k}{k!} \right] e^{-\lambda t}, \\ &= -\sum_{k=0}^{n-2} \frac{\lambda^{k+1} t^k}{k!} e^{-\lambda t} + \sum_{k=0}^{n-1} \frac{\lambda^{k+1} t^k}{k!} e^{-\lambda t}. \end{aligned}$$

Il reste finalement :

$$f_{s_n}(t) = \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t}. \tag{2.5}$$

C'est la loi d'Erlang.

2.1.3 Temps inter-arrivées

Dans le cadre des processus de file d'attente, les temps inter-arrivées jouent un rôle essentiel dans la modélisation des comportements des clients qui arrivent dans un système. Soit T_k le temps d'inter-arrivée entre la $(k-1)$-ième et la k-ième arrivée. Nous pouvons définir les temps inter-arrivées comme suit :

$$T_k = s_k - s_{k-1}.$$

où s_k est le temps d'arrivée de la k-ième personne et s_{k-1} est le temps d'arrivée de la $(k-1)$-ième personne. Dans de nombreux systèmes, les temps inter-arrivées sont modélisés par une loi exponentielle de paramètre λ, c'est-à-dire que T_k suit une distribution exponentielle :

$$f_{T_k}(t) = \lambda e^{-\lambda t} \quad \text{pour } t \geq 0$$

Cette modélisation implique que les arrivées sont indépendantes les unes des autres et que le temps entre chaque arrivée suit un processus de Poisson. Ainsi, la probabilité que le temps d'inter-arrivée soit inférieur à t peut être décrite par la fonction de répartition :

$$F_{T_k}(t) = P(T_k \leq t) = 1 - e^{-\lambda t}.$$

Cette fonction indique que plus le paramètre λ est élevé, plus les arrivées sont fréquentes, ce qui réduit les temps d'attente.

Propriétés des temps inter-arrivées

1. Indépendance : Les temps inter-arrivées $T_1, T_2, \ldots$ sont indépendants, ce qui signifie que le temps d'inter-arrivée d'une personne n'affecte pas le temps d'inter-arrivée d'une autre.

2. Moyenne et Variance : La moyenne des temps inter-arrivées est donnée par :

$$\mathbb{E}[T_k] = \frac{1}{\lambda}.$$

et la variance par :

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

3. Somme des Temps Inter-Arrivées : Si l'on considère n temps inter-arrivées, la somme des temps inter-arrivées S_n suit une distribution gamma :

$$S_n = T_1 + T_2 + \ldots + T_n \sim \text{Gamma}(n, \lambda).$$

Cela implique que la somme des n premiers temps inter-arrivées a une distribution qui capture la variabilité et les caractéristiques des arrivées dans le système.

2.1.4 Discipline et Capacité de la file, Temps moyen d'attente

Dans le cadre des systèmes de **files d'attente**, il est essentiel de comprendre plusieurs concepts clés pour analyser et modéliser l'efficacité du système. Voici les définitions et explications concernant les **discipline**, **capacité**, et **temps moyen d'attente** dans une file d'attente.

2.1.4.1 Capacité de la file d'attente

La capacité de la file d'attente fait référence au nombre maximal de clients que la file peut contenir en même temps. En fonction de la capacité, les files d'attente peuvent être classées en deux catégories :

— **Capacité infinie** : La file peut contenir un nombre illimité de clients en attente. Ce modèle est souvent utilisé dans des scénarios théoriques pour simplifier les calculs.

— **Capacité limitée (ou finie)** : Il y a une limite au nombre de clients qui peuvent attendre. Lorsque cette limite est atteinte, les clients qui arrivent sont rejetés ou perdus (modèle de rejet), ou ils peuvent aller vers une autre file d'attente si un tel mécanisme est en place.

2.1.4.2 Probabilité de Dépasser une Attente

Le temps d'attente entre deux arrivées successives suit une distribution exponentielle de paramètre λ. La probabilité que le temps d'attente entre deux arrivées dépasse un certain temps t_0 est :

$$P(T > t_0) = e^{-\lambda t_0}.$$

Exemple

Supposons que des clients arrivent à un guichet selon un processus de Poisson avec un taux d'arrivée de $\lambda = 2$ clients par minute. Nous voulons calculer la probabilité que le temps d'attente entre deux arrivées dépasse 3 minutes.

La probabilité est donnée par :

$$P(T > 3) = e^{-2\times 3} = e^{-6} \approx 0{,}00248.$$

Ainsi, il est très peu probable (environ 0,25%) que le temps d'attente entre deux arrivées dépasse 3 minutes.

2.1.4.3 Modélisation de la longueur de la file (queue)

Formulant le problème comme suit : les processus d'arrivées A_t, de départs D_t et la longueur de la file Q_t.

1. Processus des arrivées A_t : Le nombre (aléatoire) de clients arrivées dans le système. Graphiquement, chaque saut de 1 représente une arrivée d'un client.
2. Processus des départs D_t : Le nombre (aléatoire) de clients sortis du système. Graphiquement, chaque saut de 1 représente une sortie après avoir terminé son service.
3. Longueur de la file Q_t : La différence entre le nombre de clients entrés et le nombre de clients sortis. C'est une courbe en escalier qui augmente ou diminue de 1 à chaque fois qu'un client entre ou sort, voir les deux Figures (2.5) et (2.4).

$$Q_t = A_t - D_t,$$

Ensuite, la formule de Lindley en dessous, montre que le temps d'attente du $n+1$-ème client dépend du temps d'attente du n-ième client W_n, du temps de service S_n et du temps entre

l'arrivée des deux clients T_{n+1}. Notons aussi que σ_n est l'instant de sortie du n-ième client.

$$W_{n+1} = \max(W_n + S_n - T_{n+1}, 0). \tag{2.6}$$

$$\sigma_n = s_n + W_n + S_n.$$

La courbe de la longueur de la file Q_t de la Figure (2.5) est indépendante en nombres par

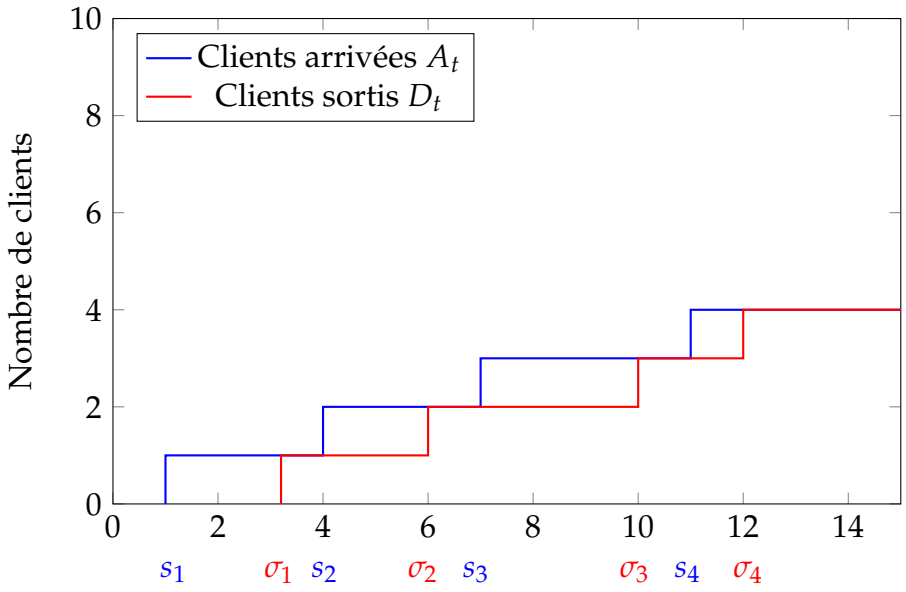

FIGURE 2.4 – Courbes d'arrivées A_t et des sortis D_t.

rapport aux deux courbes d'arrivées A_t et de sorties D_t.

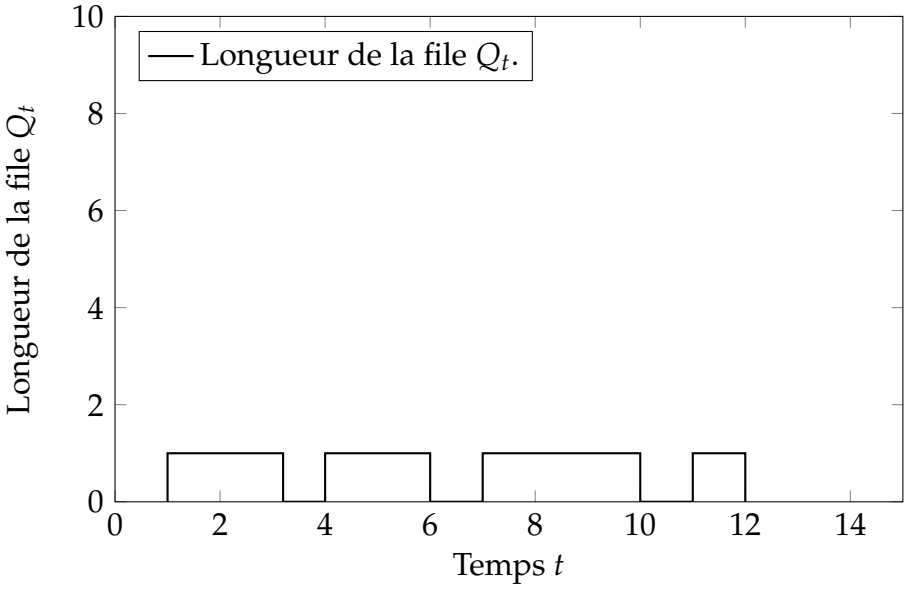

FIGURE 2.5 – Courbe de la longueur de la file Q_t.

2.1.4.4 Synthèse des Principales Probabilités dans une File d'Attente

Les files d'attente sont modélisées à l'aide de différentes distributions de probabilité qui décrivent divers aspects du système, comme les arrivées des clients, les temps de service, le nombre de clients dans la file, et les temps d'attente. Voici un aperçu des principales probabilités utilisées dans la théorie des files d'attente.

Loi	**Utilisation**	**Formule**
Loi de Poisson	Modélisation des arrivées	$P(X = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$
Loi Exponentielle	Modélisation des temps entre arrivées ou des temps de service	$f_T(t) = \lambda e^{-\lambda t}$, $F_T(t) = 1 - e^{-\lambda t}$
Loi Géométrique	Longueur de la file dans $M/M/1$	$P(L = n) = (1 - \rho)\rho^n$
Loi d'Erlang	Probabilité d'attente ou de blocage dans les systèmes multi-serveurs	$P_{\text{blocage}} = \frac{\frac{(\lambda/\mu)^c}{c!}}{\sum_{k=0}^{c} \frac{(\lambda/\mu)^k}{k!}}$
Loi Gamma	Modélisation des temps de service en plusieurs étapes	$f_T(t) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{(k-1)!}$
Loi Uniforme	Arrivées ou services non Poissonniens	$f_T(t) = \frac{1}{b-a}, \quad a \leq t \leq b$
Loi Normale	Modélisation des variations de service ou d'attente	$f_T(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$

TABLE 2.1 – Lois Mathématiques dans les Files d'Attente

2.1.5 Discipline, Nombre et Temps de service

Le temps de service est le temps pris pour traiter un client. Il suit généralement une loi exponentielle de paramètre μ, bien que d'autres distributions soient possibles. Le service peut être constitué d'un ou plusieurs serveurs, qui peuvent être disposés de diverses façons : Serveurs en séries, serveurs en parallèles et serveurs en réseaux, voir les figures (**??**).

Dans les systèmes de **files d'attente**, les variables aléatoires jouent un rôle crucial dans la modélisation et l'analyse des événements comme :

— Les **temps entre les arrivées** (souvent modélisés par une loi exponentielle),
— Le **nombre de clients dans la file** (souvent modélisé par une loi de Poisson),
— Le **temps d'attente** ou le **temps de service** (souvent modélisés par des distributions exponentielles ou normales).

2.1.5.1 Discipline de service

La discipline de la file d'attente détermine la manière dont les clients sont servis une fois qu'ils sont en attente dans le système. Il existe plusieurs types de disciplines, parmi les plus courantes :

— **FIFO (First In, First Out)** : Les clients sont servis dans l'ordre dans lequel ils sont arrivés. C'est la discipline la plus commune dans de nombreux systèmes.

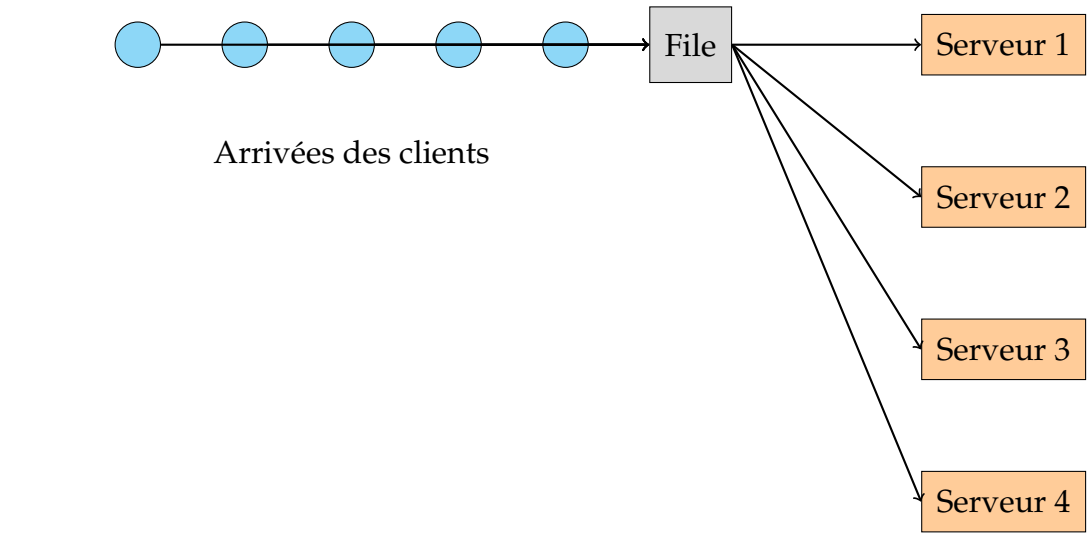

FIGURE 2.6 – Arrivées dans une File avec des serveurs en parallèle.

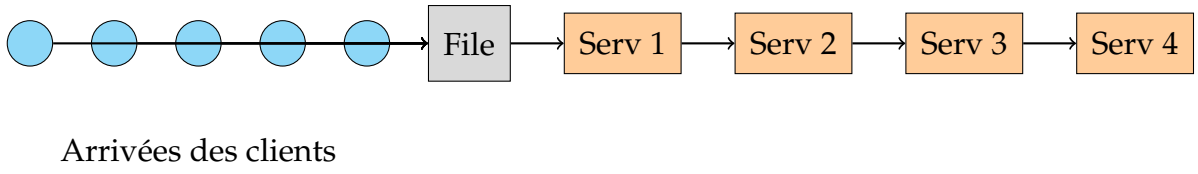

FIGURE 2.7 – Arrivées dans une File avec des serveurs en series.

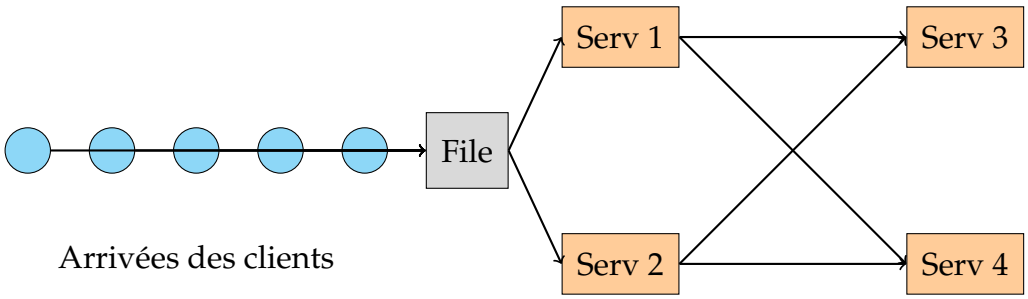

FIGURE 2.8 – Arrivées dans une File avec des serveurs en réseau.

— **LIFO (Last In, First Out)** : Le dernier client arrivé est le premier à être servi. Cette discipline est moins fréquente et peut être utilisée dans des systèmes spécifiques (par exemple, gestion des stocks de type pile).
— **SIRO (Serve In Random Order)** : Les clients sont servis dans un ordre aléatoire, sans considération de leur ordre d'arrivée.
— **PS (Processor Sharing)** : Tous les clients reçoivent une part égale du temps de service simultanément. Ce modèle est souvent utilisé pour la modélisation de partage de ressources, comme en informatique.
— **Priorités** : Certains systèmes attribuent des priorités aux clients, et les clients avec des priorités plus élevées sont servis avant ceux avec des priorités inférieures.

2.1.5.2 La durée de service du n-ième client

Supposons que S_n soit une variable aléatoire représentant la durée de service du n-ième client. La propriété d'absence de mémoire qui se traduit par l'égalité :

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

D'autre part, on peut écrire la probabilité non conditionnelle $P(S_n > s + t)$ comme suit,

$$P(S_n > s + t) = P(S_n > t) \cdot P(S_n > s),$$

Cela conduit à l'introduction d'une fonction φ définit par $\varphi(t) = P(S_n > t)$ qui vérifié,

$$\varphi(t + s) = \varphi(t) \cdot \varphi(s).$$

Et comme la probabilité que la durée de service dépasse 0 est de 1 ($\varphi(0) = 1$). Alors, par dérivation par rapport à s on aura

$$\varphi^{'}(t + s) = \varphi(t) \cdot \varphi'(s).$$

En prenant $s = 0$, et en posant $\varphi'(0) = a$, on obtient l'équation différentielle suivante,

$$\varphi'(t) = c \cdot \varphi(t).$$

dont la solution générale est donnée par :

$$\varphi(t) =^{ct} .$$

Mais, $\varphi(t)$ est entre 0 et 1, on peut remplacer $c = -\mu$ comme

$$\varphi(t) = e^{-\mu t}.$$

Donc, la variable aléatoire S_n, représentant la durée de service et suit une loi exponentielle où μ est le nombre moyen de services par unité de temps.

2.2 La file en régime stationnaire

Dans le cadre des files d'attente, le régime stationnaire désigne un état d'équilibre où les caractéristiques statistiques du système (comme la longueur de la file, le temps d'attente, et les probabilités de transition entre états) ne changent plus en moyenne. Ce régime n'est atteint que lorsque Condition de Stabilité vérifiée $\lambda < \mu$.

2.2.1 Probabilités Stationnaires et équilibre

Les probabilités stationnaires sont les probabilités qu'il y ait n clients dans le système en régime stationnaire, noté $\pi_n = P(Q_\infty = n)$, où Q_∞ est le nombre de clients dans le système à long terme. Ces probabilités vérifient les équations d'équilibre.

2.2.1.1 Équations de Balance

La figure ci-dessus montre les transitions entre les états voisins dans une file d'attente où l'on examine les états $n-1$, n et $n+1$.

1. **Flux entrant en l'état n :**
   Entrant : La probabilité d'être dans l'état n avec un client qui arrive avec le taux λ, donc $\lambda\pi_n$.
   Sortant : La probabilité d'être dans l'état $n+1$ avec un client qui part avec le taux μ, donc $\mu\pi_{n+1}$.
   L'équation de balance pour l'état $n+1$ est :

$$\lambda\pi_n = \mu\pi_{n+1}. \tag{2.7}$$

2. **Flux sortant de l'état n :**
   Entrant : La probabilité d'être dans l'état $n-1$ et d'avoir un client qui arrive avec le taux λ, ou la probabilité d'être dans l'état $n+1$ avec un client qui part avec le taux μ, soit $\lambda\pi_{n-1} + \mu\pi_{n+1}$.
   Sortant : La probabilité d'être dans l'état n avec un client qui arrive avec le taux λ, ou d'avoir un client qui part avec le taux μ, donc $(\lambda+\mu)\pi_n$.
   L'équation de balance pour l'état n est donc :

$$\lambda\pi_{n-1} + \mu\pi_{n+1} = (\lambda+\mu)\pi_n. \tag{2.8}$$

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

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

Pour résoudre ce système d'équations de balance, nous cherchons les probabilités stationnaires π_n en fonction de π_0. Nous allons utiliser les équations du système au dessus.

Calcul de π_2 en fonction de π_0. À partir de l'équation de récurrence avec $n = 1$, on obtient :

$$\lambda\pi_0 + \mu\pi_2 = (\lambda+\mu)\pi_1$$

En substituant $\pi_1 = \frac{\lambda}{\mu}\pi_0$,

$$\lambda\pi_0 + \mu\pi_2 = (\lambda + \mu)\left(\frac{\lambda}{\mu}\pi_0\right)$$

En isolant π_2 :

$$\pi_2 = \frac{\lambda^2}{\mu^2}\pi_0$$

Calcul de π_n en général. Par récurrence, on peut démontrer que :

$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0$$

Condition de Normalisation : Pour que les probabilités stationnaires soient valides, la somme des probabilités doit être égale à 1,

$$\sum_{n=0}^{\infty} \pi_n = 1$$

Substituons $\pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0$:

$$\pi_0 \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = 1$$

Cela est une série géométrique, et pour qu'elle converge, il faut que $\frac{\lambda}{\mu} < 1$ (c'est-à-dire que le système soit stable, avec $\lambda < \mu$). La somme de cette série est alors :

$$\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = \frac{1}{1 - \frac{\lambda}{\mu}}.$$

Donc :

$$\pi_0 \cdot \frac{1}{1 - \frac{\lambda}{\mu}} = 1$$

En isolant π_0 et on pose $\rho = \frac{\lambda}{\mu}$, on obtient :

$$\pi_0 = P(Q_\infty = 0) = 1 - \frac{\lambda}{\mu} = 1 - \rho. \tag{2.9}$$

Les probabilités stationnaires sont donc données par :

$$\pi_n = P(Q_\infty = n) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n = (1 - \rho)\rho^n. \tag{2.10}$$

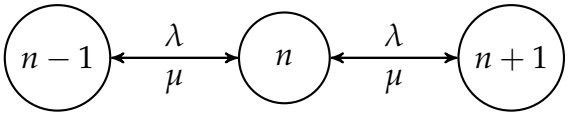

FIGURE 2.9 – Transitions entre les états.

Donc on peut dire que cet variable suit la loi géométrique et la longueur moyenne de la file (queue) est,

$$L = E(Q_\infty) = \frac{\lambda}{\mu - \lambda}. \tag{2.11}$$

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

Le temps d'attente W_∞ d'un client en régime stationnaire suit une loi de probabilité que nous pouvons déterminer comme suit. Lorsque le nombre de clients dans le système est n, le temps d'attente suit une distribution d'Erlang. Le temps d'attente conditionnel est donné par :

$$P(W_\infty \leq t \mid Q_\infty = n) = 1 - \sum_{k=0}^{n-1} \frac{(\mu t)^k}{k!} e^{-\mu t}.$$

En utilisant la probabilité $P(Q_\infty = n) = (1-\rho)\rho^n$, nous obtenons la probabilité totale que le temps d'attente soit inférieur à t :

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

Le temps d'attente conditionnel $(W_\infty \leq t \mid Q_\infty = n)$ représente l'attente durant laquelle n clients consécutifs doivent être servis avec un service exponentiel $E(\mu)$. Cela correspond donc à la somme de n variables aléatoires indépendantes de loi $E(\mu)$. Ainsi :

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

En simplifiant, nous obtenons :

$$P(W_\infty \leq t) = 1 - \frac{\lambda}{\mu} e^{-(\mu-\lambda)t}. \tag{2.12}$$

L'attente moyenne d'un client, calculée en utilisant le même procédé, est donnée par :

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

Sachant que $\mathbb{E}(W_\infty \mid Q_\infty = n) = \frac{n}{\mu}$, nous avons :

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

soit,

$$\mathbb{E}(W_\infty) = \frac{\lambda}{\mu(\mu - \lambda)}. \tag{2.13}$$

L'expression $W_\infty + S_\infty$ représente le temps de séjour total (attente + service) du client générique, qui suit une loi exponentielle $E(\mu - \lambda)$:

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

Le temps de séjour moyen d'un client dans le système est alors :

$$\mathbb{E}(W_\infty + S_\infty) = \frac{1}{\mu - \lambda}. \tag{2.14}$$

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

$$\mathbb{E}(Q_\infty) = \mu\mathbb{E}(W_\infty) = \lambda\mathbb{E}(W_\infty + S_\infty). \tag{2.15}$$