# Files d'attente

- Introduction
- Constitution d'une file d'attente
- Modélisation des arrivées
- Modélisation du temps de service
- Modélisation de la longueur de la queue
- Étude de la file en régime stationnaire
- Autres modèles de files d'attente
- Applications

# Introduction

- **La théorie des files d'attente** est principalement vue comme une branche **de la théorie des probabilités appliquées**.
- Cette théorie utilise des **outils probabilistes pour étudier et modéliser le comportement d'un système donné**. En quelques mots, cette théorie a pour objet **l'étude des systèmes** ou du comportement des "**entités**" appelées : **clients, services, gestionnaire**.
- Les clients cherchent à accéder à une ressource afin d'obtenir un service.
- Dans certains cas, un client a besoin de recevoir **plusieurs traitements avant de quitter le système**. Par exemple, dans les systèmes de production, les banques, les systèmes informatiques

# Introduction

## ❑ Définition

- La théorie de files d'attente est une technique **de la recherche opérationnelle** qui permet de **modéliser un système admettant un phénomène d'attente, de calculer ses performances et de déterminer ses caractéristiques** pour aider les praticiens dans leurs **prises de décisions**
- Ce domaine de recherches, né en **1917**, des travaux de l'ingénieur **Danois** Erlang sur **la gestion des réseaux téléphoniques** de **Copenhague** à partir de 1908, étudie notamment **les systèmes d'arrivée dans une queue**, les différentes priorités de chaque nouvel arrivant, ainsi que **la modélisation statistique des temps d'exécution**.

# Introduction

## ❑ 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: **queue à un guichet, saturation d'un trafic routier, 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** qui repose sur l'absence de mémoire de certaines occurrences.

# Introduction

## ❑ Position du problème

- 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**.

Q1: Quel est le temps passé par un client dans une file d'attente devant un guichet ?

Q2: Quelle est la longueur de la queue à un instant donné ?

Q3: Quelle est la durée de repos du serveur (c'est-à-dire lorsque la queue est vide avant l'arrivée d'un nouveau consommateur)?

Q4: À quelle vitesse minimale devrait travailler le serveur pour ne pas dépasser un seuil maximal de clients?

# Introduction

## ❑ Position du problème

Q5: Combien de serveurs faudrait-il au minimum pour éviter la saturation de la salle d'attente… ?

Autant de questions auxquelles **la théorie de files d'attente et des processus stochastiques apporte des réponses plus ou moins explicites** selon le modèle adopté.

# 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**.
- Les clients peuvent être **des individus, des appels téléphoniques, des signaux électriques, des véhicules, des accidents, des turbulences atmosphériques…**
- Le service peut être un **serveur humain, un central téléphonique, un serveur informatique, un péage autoroutier, une compagnie d'assurance, la météorologie nationale…**

# Constitution d'une file d'attente

## ❑ 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.

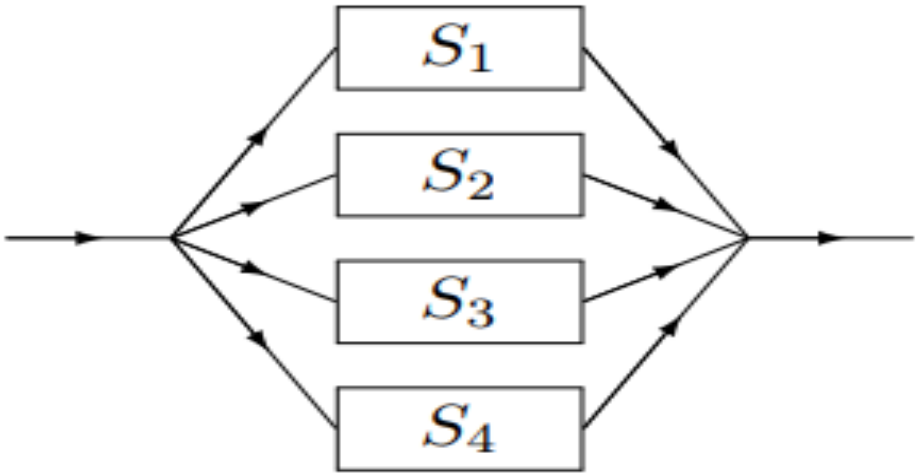
## ❑ Organe de service

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

# Constitution d'une file d'attente

- **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..**



# Constitution d'une file d'attente

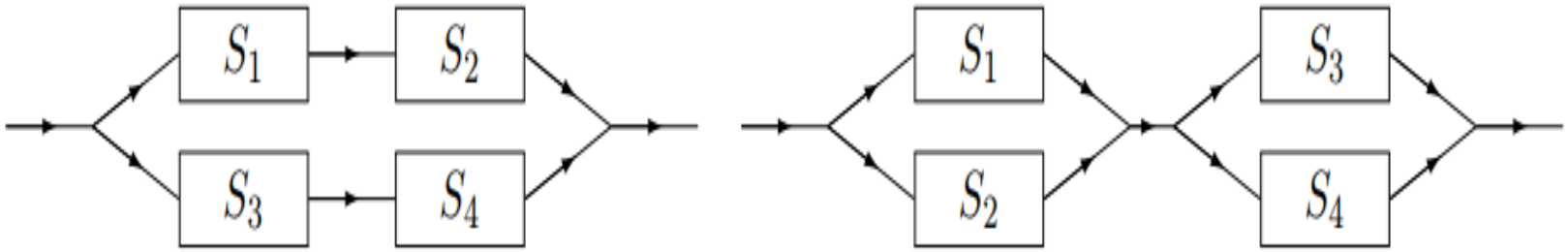
- **Serveurs en série**

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**...

# Constitution d'une file d'attente

- **Serveurs en réseau**

Cette disposition **est beaucoup plus complexe et réaliste** : centraux de télécommunications, réseaux informatiques, internet...




**On devra par ailleurs modéliser les temps de service.**

## ❑ Discipline de service

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

# Constitution d'une file d'attente

- Une règle de courtoisie voudrait que l'on serve les personnes en respectant leur ordre d'arrivée, **FCFS ( First come, First served),** c'est **FIFO** dans le cas où le service est exécuté par un unique serveur.
- Une règle opposée à la règle de courtoisie consiste à servir en premier le dernier client arrivé : gestion d'un stock. C'est la discipline **LCFS (last come, First served). LIFO** (last in first out ) dans le cas où le service est exécuté par un unique serveur.
- 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.

# Constitution d'une file d'attente

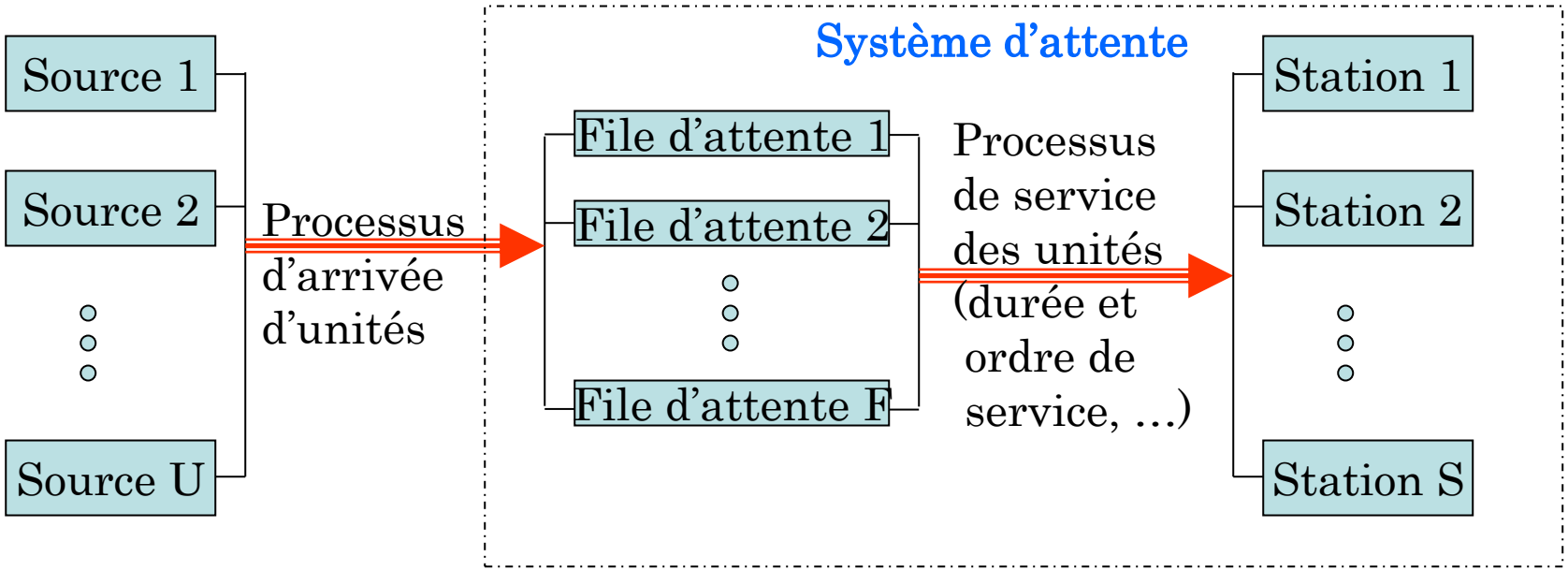
- Enfin, dans d'autres situations, on adopte **le partage de service** (processor sharing) : on rencontre ce cas notamment en informatique, cas dans lequel **différentes tâches sont traitées simultanément**.

## ❑ Capacité du système

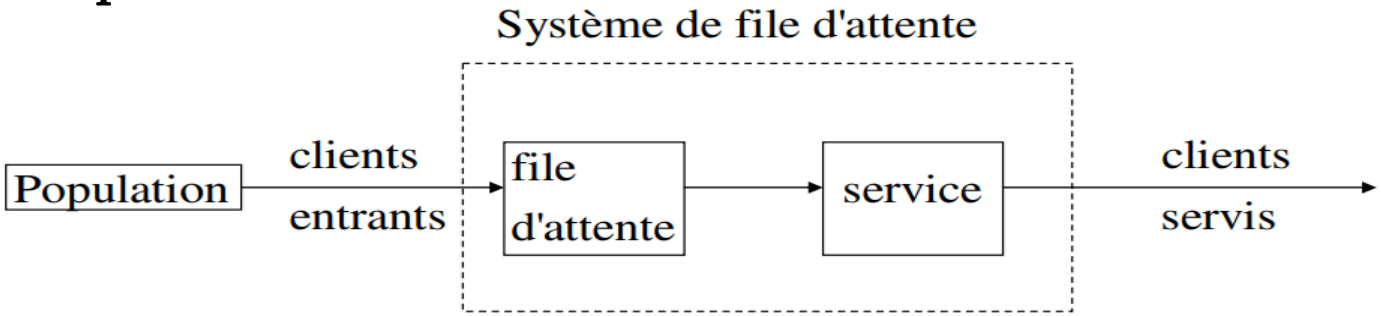
- De nombreux systèmes d'attente comportent **une salle d'attente à capacité limitée**. Il y aura donc refoulement de personnes à l'entrée du système lorsque cette salle est pleine. Ce facteur a une importance notamment dans **l'étude de files en tandem**.
- Si par exemple une **salle d'attente à capacité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

# Constitution d'une file d'attente

## ❑ Schéma de file d'attente




**Plus simplement**



# Constitution d'une file d'attente

## ❑ Classification des systèmes d'attente

Pour identifier un système d'attente, on a besoin des spécifications suivantes :

- **La nature stochastique du processus des arrivées**, qui est défini par la distribution des intervalles séparant deux arrivées consécutives ;
- **La distribution du temps aléatoire de service** ;
- **Le nombre *m* de serveurs** (stations de service) qui sont montées en parallèle. On admet généralement que **les temps de service correspondants suivent la même distribution et que les clients qui arrivent forment une seule file d'attente** (dans le cas homogène) ;
- **La capacité *N* du système**. Si $N < \infty$, la file d'attente ne peut dépasser une longueur de $N - m$ Unités. Dans ce cas, certains clients arrivant vers le système n'ont pas la possibilité d'y entrer ;
- **La source des clients potentiels**.

# Constitution d'une file d'attente

## ❑ Notation de Kendall A/B/s/N/K/D

- Un modèle de file d'attente est totalement décrit selon la notation de Kendall.
- Dans sa version étendue, un modèle est spécifie par une suite de six symboles : ***A/B/s/N/K/D***
  - *A* : Nature du processus des arrivées ;
  - *B* : Nature du processus de service ;
  - *s* : Nombre de serveurs en parallèle ;
  - *N* : Capacité du système (serveurs + file d'attente) ;
  - *K* : Taille de la population ;
  - *D* : Discipline de la file.

# Constitution d'une file d'attente

## ❑ Notation de Kendall A/B/s/N/K/D

- Dans la description des processus d'arrivée et de service, les symboles les plus courants sont :
  - M : Distribution exponentielle *(qui vérifie donc la propriété de Markov)* ;
  - E : Distribution d'Erlang ;
  - G : Distribution générale *(on ne sait rien sur ses caractéristiques)* ;
  - D : loi Déterministe *(temps d'inter-arrivés ou de service constant)* ;

- La forme abrégé : ***A/B/s*** signifie que $N$ et $K$ sont infinies

# Constitution d'une file d'attente

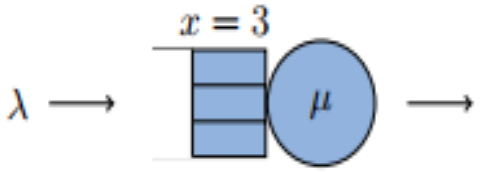
## ❑ Exemple 1: File M/M/1/ ∞ /PS

*M* : Processus d'arrivée : Poisson de paramètre $\lambda > 0$.

*M* : Distribution du temps de service : exponentielle de paramètre $\mu > 0$.

*1/∞* : 1 serveur, une infinité de places dans la file.

*PS* : Discipline de service processor-sharing.



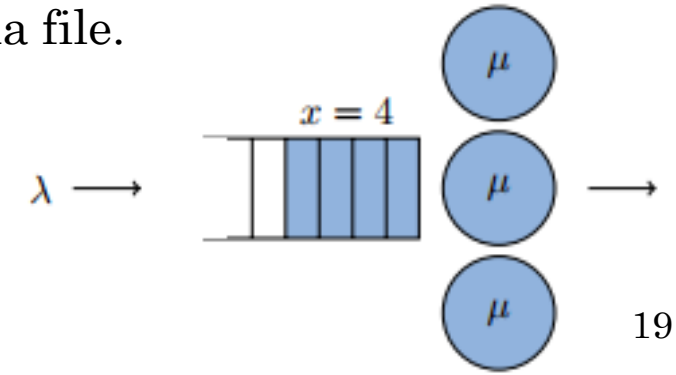

## ❑ Exemple 2 : File M/M/3/ ∞/ FIFO

*M* : Processus d'arrivée : Poisson de paramètre $\lambda > 0$.

*M* : Distribution du temps de service : exponentielle de paramètre $\mu > 0$.

*3/∞* : 3 serveurs, une infinité de places dans la file.

*FIFO* : Discipline de service first-in, first-out



# Constitution d'une file d'attente

## ❑ Mesures de performance d'une file d'attente :

L'étude d'une file d'attente a pour but **de calculer ou d'estimer les performances du système dans des conditions de fonctionnement données.** Ce calcul se fait le plus souvent pour **le régime stationnaire uniquement**, et les mesures les plus fréquemment utilisées sont :

- E(N ) : nombre moyen de **clients dans le système** ;
- E(Q) : nombre moyen de **clients dans la file d'attente** ;
- E(T) : temps moyen de **séjour d'un client dans le système** ;
- E(W) : temps moyen **d'attente d'un client dans la file** ;
- E(U) : taux **d'utilisation de chaque serveur** ;
- E(S) : le temps **moyen de service** ;
- E(A) : le temps **moyen entre deux arrivées**

# Constitution d'une file d'attente

## ❑ Mesures de performance d'une file d'attente :

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

On parle de le *M/M/*1.

# 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 $\Omega$ (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 :

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

# Annexes

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

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

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

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

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

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

# Annexes

## A.2 Quelques lois de probabilité

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

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

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

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

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

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

# Annexes

– *Loi d'Erlang* $E(n;\lambda)$ :

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

– *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 $\alpha_k = 1 \Big/ \prod_{\substack{1 \leqslant j \leqslant n \\ j \neq k}} \left(1 - \frac{\lambda_k}{\lambda_j}\right)$ ;

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

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

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

# Modélisation des arrivées

## ❑ Processus des arrivées

- Prenons l'exemple des personnes se présentant à un guichet. On fera trois hypothèses :

1. **les arrivées sur des intervalles de temps disjoints sont indépendantes ;**

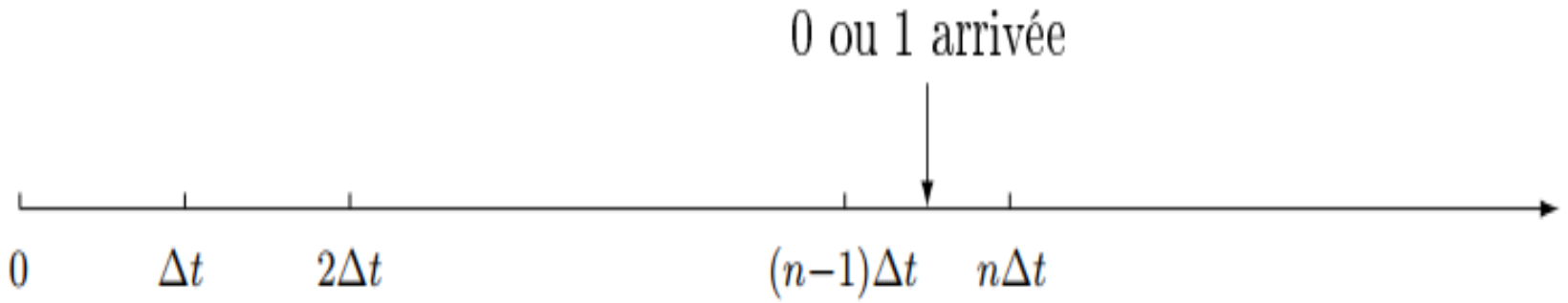
2. **les arrivées sont aléatoires et les laps de temps inter-arrivées ont même loi de probabilité ;**

3. **il n'y a pas d'arrivées simultanées, i.e. il n'arrive pas plus d'un client à la fois.**

# Modélisation des arrivées

## ❑ Processus des arrivées

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




- Notons $Xn$ le nombre (aléatoire) d'arrivées durant l'intervalle de temps $[(n-1)\Delta t, n\Delta t[$.
- $Xn$ est une v.a. prenant essentiellement les deux valeurs 0 et 1.

# Modélisation des arrivées

## ❑ Processus des arrivées

- $Xn$ suit approximativement une loi de Bernoulli, elle est donc caractérisée par un paramètre $P_{\Delta t}$ qui n'est autre que son espérance : $\mathrm{E}(Xn) \approx P_{\Delta t}$ .
- 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 $t \in \mathbf{R}+ \to At$

# Modélisation des arrivées

## ❑ Processus des arrivées

- $At$ représente le nombre (aléatoire) de consommateurs entrés dans le système pendant le laps de temps $[0, t]$.
- L'étude faite ci-dessus montre que $A_{n\Delta t} = X1 + \cdots + Xn$
- La v.a. $A_{n\Delta t}$ suit approximativement la loi binomiale $B(n, p\Delta t)$.
  Pour un instant $t$ est dans un intervalle $[(n-1)\Delta t, n\Delta t[$

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

- Faisons tendre $\Delta t$ vers 0+ ou encore $n$ vers $+\infty$ avec $n\Delta t \sim t$ (fixé), donc $P_{\Delta t} \sim \lambda t/n$.

donc
$$p_{k,\Delta t}(t) = \frac{n!}{(n-k)!\, n^k} \frac{(\lambda t)^k}{k!} \left(1 - \frac{\lambda t}{n}\right)^{n-k} + o(\Delta t),$$

# Modélisation des arrivées

## ❑ Processus des arrivées

$$\lim_{\Delta t \to 0^+} p_{k,\Delta t}(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$ donc $$\mathbb{P}(A_t = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

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

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

- **Le paramètre $\lambda$ représente le nombre moyen d'arrivées par unité de temps (taux d'arrivée)**

# Modélisation des arrivées

## ❑ Temps inter-arrivées

- Soit $T1 = \inf\{t \geq 0 : At \geq 1\}$ l'instant d'arrivée (aléatoire) de la première personne.
- La condition $T1 > t$ signifie que la première personne arrive après l'instant $t$ et donc que $At = 0$.
- En conséquence, $$\mathbb{P}(T_1 > t) = \mathbb{P}(A_t = 0) = e^{-\lambda t}$$
- La fonction de répartition de $T1$ est $$F_{T_1}(t) = \mathbb{P}(T_1 \leqslant t) = 1 - e^{-\lambda t}$$
- La v.a. $T_1$ suit la loi exponentielle $\boldsymbol{\varepsilon(\lambda)}$. Elle a pour densité $$f_{T_1}(t) = \lambda e^{-\lambda t}$$

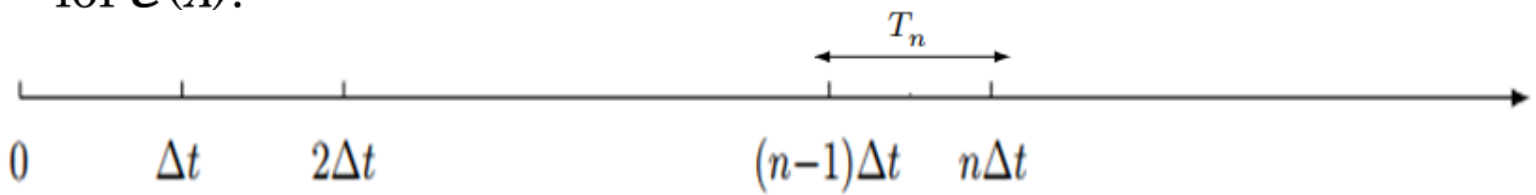
# Modélisation des arrivées

## ❑ Temps inter-arrivées

- Son espérance vaut $\mathbf{E(T1) = \frac{1}{\lambda}}$ , ce qui donne pour $\lambda$ une autre interprétation :

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

- En fait, la loi de T1 mesure aussi l'intervalle de temps séparant deux arrivées consécutives
- Si Tn est le laps de temps s'écoulant entre les *(n - 1) et n* personnes
- les v.a. T1, T2, . . . , Tn, . . . sont indépendantes de loi commune la loi $\mathcal{E}(\lambda)$.



# Modélisation des arrivées

## ❑ Temps d'arrivée de la *n* personne

- Soit $\mathbf{s_n = \inf\{t \geq 0 : At = n\}}$ l'instant **d'arrivée de la *n* personne.**
- La condition $\mathbf{s_n > t}$ signifie que la *n* 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 $\mathbf{s_n}$

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

- La densité de $\mathbf{s_n}$ s'obtient en dérivant l'expression ci-dessus :

# Modélisation des arrivées

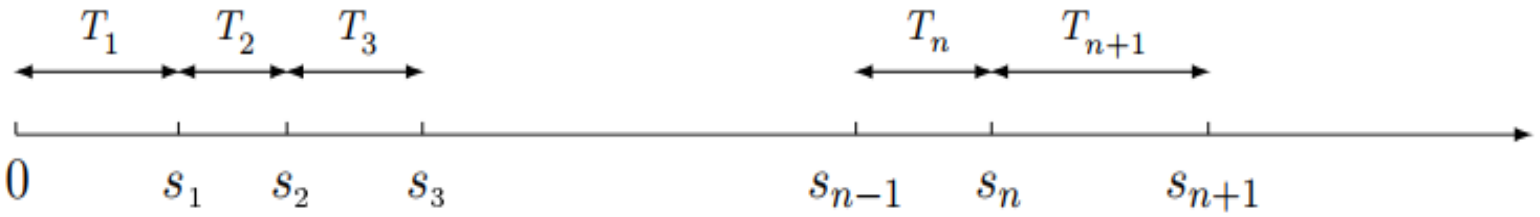
## ❑ Temps d'arrivée de la *n* personne

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

$$= -\sum_{k=0}^{n-2} \frac{\lambda^{k+1} t^k}{k!} e^{-\lambda t} + \sum_{k=0}^{n-1} \frac{\lambda^{k+1} t^k}{k!} e^{-\lambda t}$$

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

- C'est la célèbre loi d'Erlang $\mathbf{E(n;\lambda)}$. Bien sûr, $\mathbf{s_n}$ est la somme des *n* premiers temps inter-arrivées : $\mathbf{s_n} = \boldsymbol{T_1} + \boldsymbol{T_2} + \cdots \boldsymbol{T_n}$ où $\boldsymbol{T_k} = \mathbf{s_k} - \mathbf{s_{k-1}}$.



# Modélisation du temps de service

## ❑ Modélisation du temps de service

- Soit ***Sn*** la durée de service de la *n* 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 de Markov

- 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 $\varphi$ définie par $\varphi(t) = P(Sn > t)$ vérifié ainsi l'équation fonctionnelle $\varphi(s+t) = \varphi(s)\varphi(t)$.

# Modélisation du temps de service

## ❑ Modélisation du temps de service

- De plus, cette fonction est bornée $(0 \leq \varphi(t) \leq 1)$ et vérifié $\varphi(0) = 1$.
- En rajoutant l'hypothèse simplificatrice (non nécessaire a priori) que $\varphi$ 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 $\varphi(0) = 1$, admet pour solution

$$\varphi(t) = e^{at}$$

# Modélisation du temps de service

## ❑ Modélisation du temps de service

- Enfin, la fonction $\varphi$ recherchée doit être bornée, ceci entraîne donc

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

où la constante $\mu$ est l'inverse du temps moyen de service $\mu = \frac{1}{\mathbb{E}(S_n)}$

- La v.a. $Sn$ suit la loi exponentielle $\varepsilon(\mu)$.

- Le paramètre $\mu$ pourrait s'interpréter comme représentant le nombre moyen de services (taux de service) que le serveur peut effectuer par unité de temps.

# Modélisation de la longueur de la queue

## ❑ 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.
- Alors $Q_t$ la longueur de la file le à l'instant $t$, c'est-à-dire le nombre de personnes présentes dans le système (**en attente ou en service**) ; on a donc :

$$Q_t = A_t - D_t$$

(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 $\lambda$ et de taux de mort $\mu$

# Modélisation de la longueur de la queue

## ❑ Modélisation de la longueur de la queue

- Pour pouvoir représenter la courbe de départs $t \rightarrow (D_t)$, 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 $\boldsymbol{W_n}$ le temps d'attente du $n$ client présent dans la fille et $\boldsymbol{\sigma_n}$ son instant de sortie. Le temps $\boldsymbol{W_n}$ est donné par la relation de récurrence suivante (formule de Lindley, 1955)

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

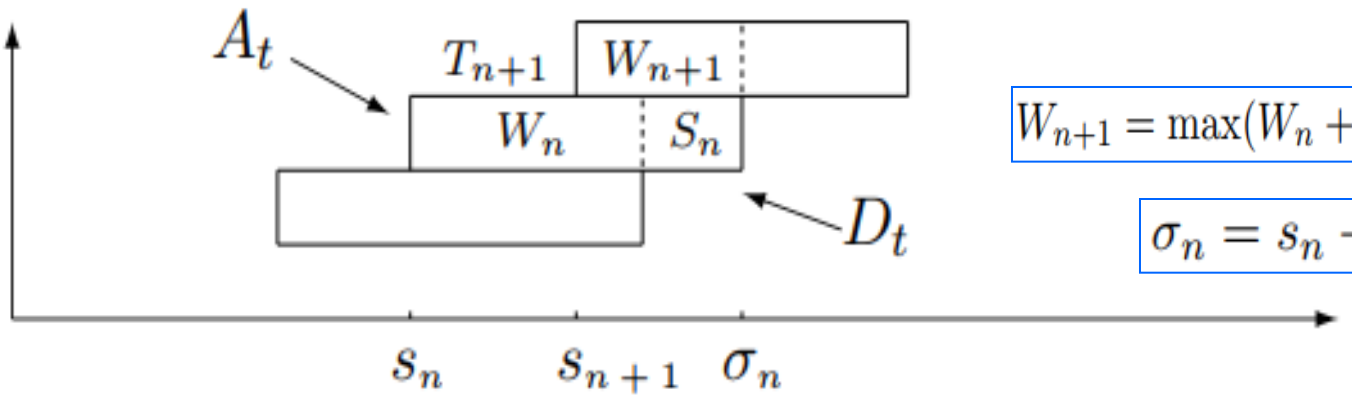
et alors

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

# Modélisation de la longueur de la queue

## ❑Modélisation de la 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 suivante, il est clair que si la $(n+1)$ personne arrive avant le départ de la $n$ , 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 où la $(n+1)$ personne arrive après le départ de la $n$, la $(n+1)$ n'a pas d'attente : $W_n+1 = 0$.



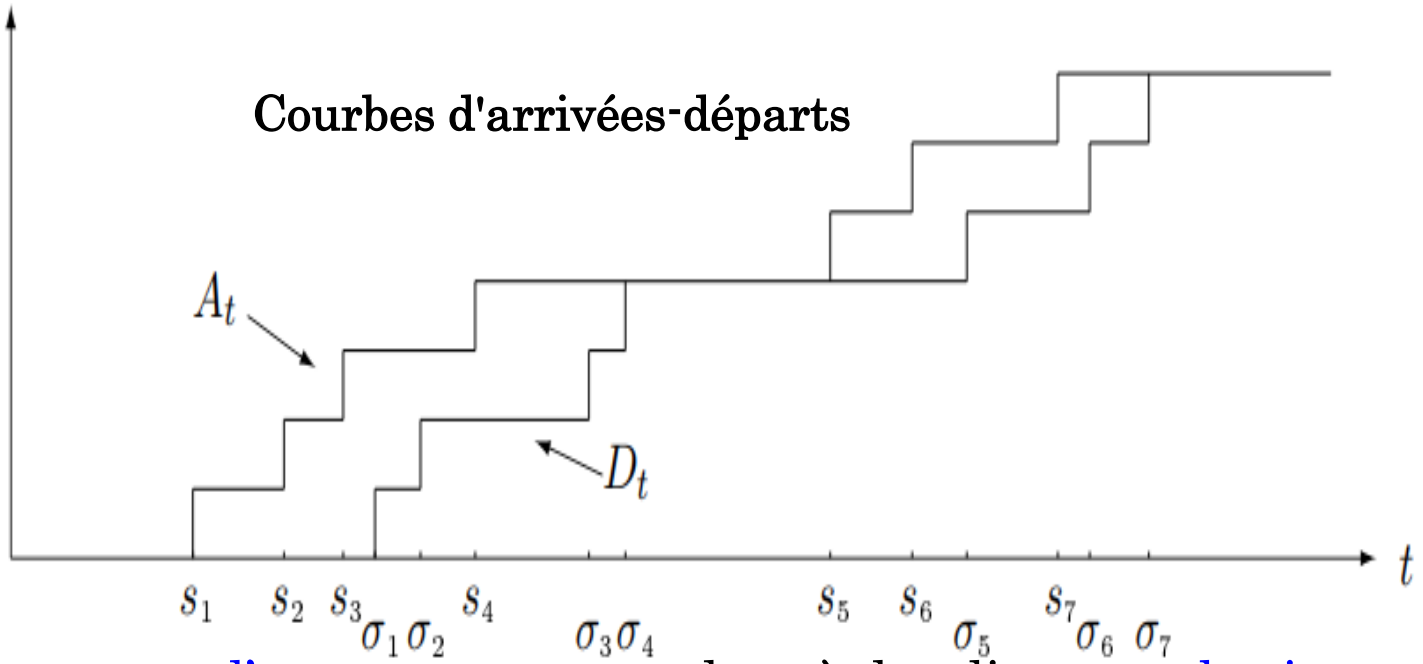

$$W_{n+1} = \max(W_n + S_n - T_{n+1}, 0)$$

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

# Modélisation de la longueur de la queue

## ❑Modélisation de la longueur de la queue

- Les graphes des fonctions **d'arrivées** $t \rightarrow A_t$, de **départs** $t \rightarrow D_t$ sont des courbes en **escaliers croissantes avec des sauts de 1**. Ils sont représentés ci-dessous.



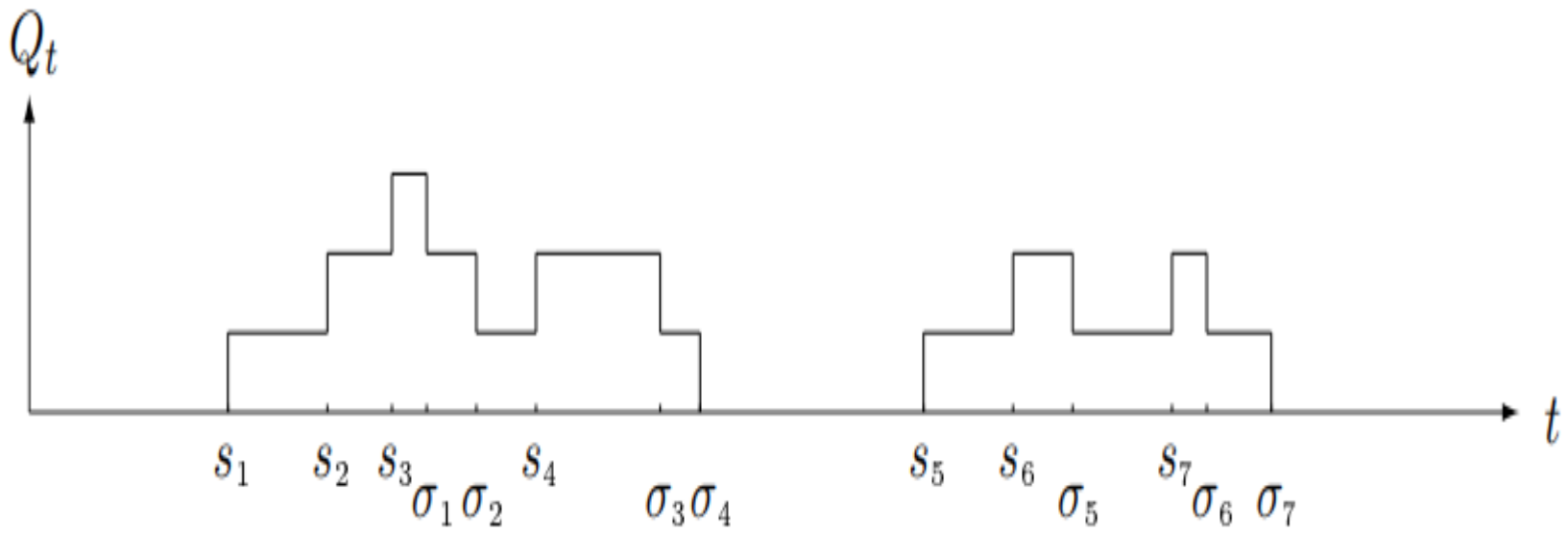

Les temps d'attente correspondent à des distances horizontales;

La longueur de la file correspond à des distances verticales

# Modélisation de la longueur de la queue

## ❑ Modélisation de la longueur de la queue

- Le graphe de la longueur de la file $t \rightarrow Q_t$ est une courbe en escaliers avec des sauts de ±1. Il est représenté ci-dessous.

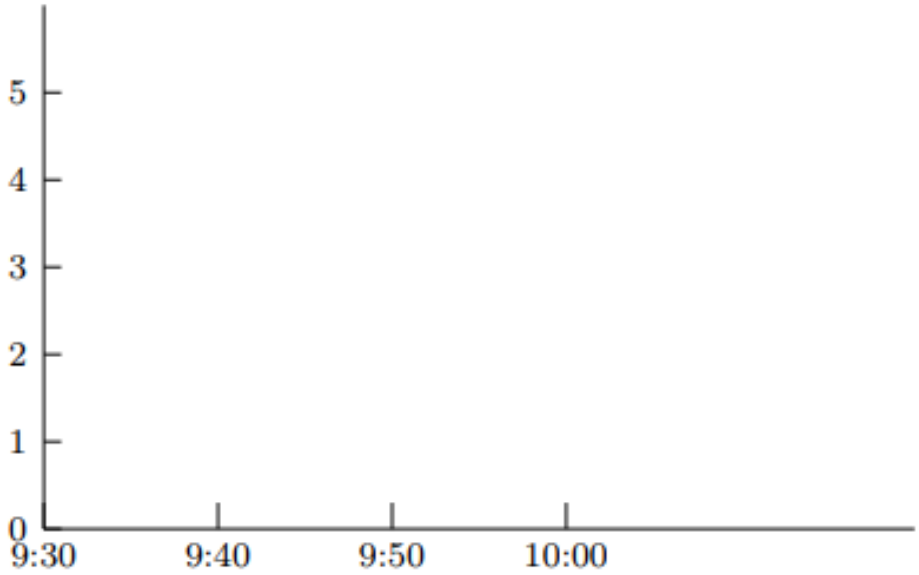



Longueur de la queue

# Modélisation de la longueur de la queue

- Exemple

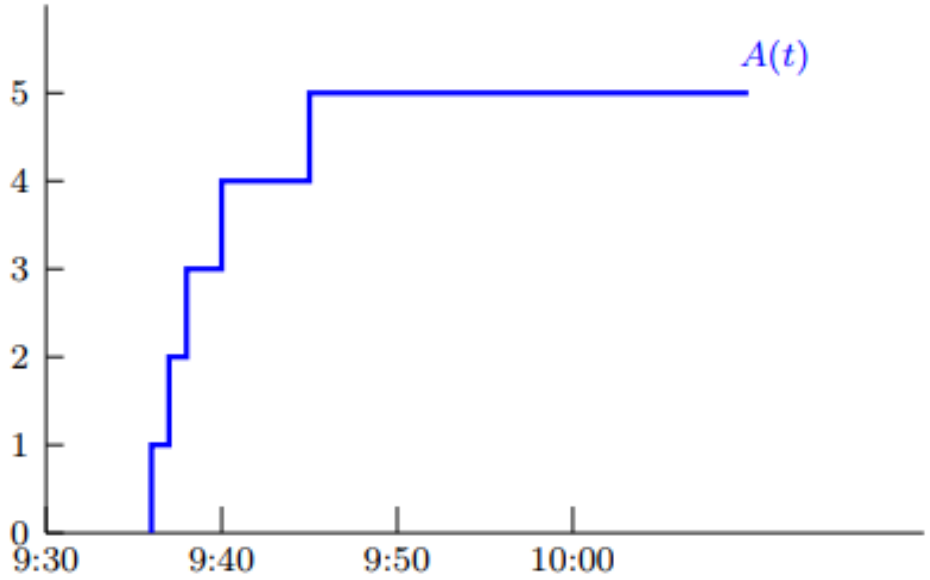
| Client | Arrivée | Service | Départ du système |
|---|---|---|---|
| 1 | 9:36 | 9:36 | 9:40 |
| 2 | 9:37 | 9:40 | 9:44 |
| 3 | 9:38 | 9:44 | 9:48 |
| 4 | 9:40 | 9:48 | 9:52 |
| 5 | 9:45 | 9:52 | 9:56 |



# Modélisation de la longueur de la queue

❑ Exemple

| Client | Arrivée | Service | Départ du système |
|---|---|---|---|
| 1 | 9:36 | 9:36 | 9:40 |
| 2 | 9:37 | 9:40 | 9:44 |
| 3 | 9:38 | 9:44 | 9:48 |
| 4 | 9:40 | 9:48 | 9:52 |
| 5 | 9:45 | 9:52 | 9:56 |

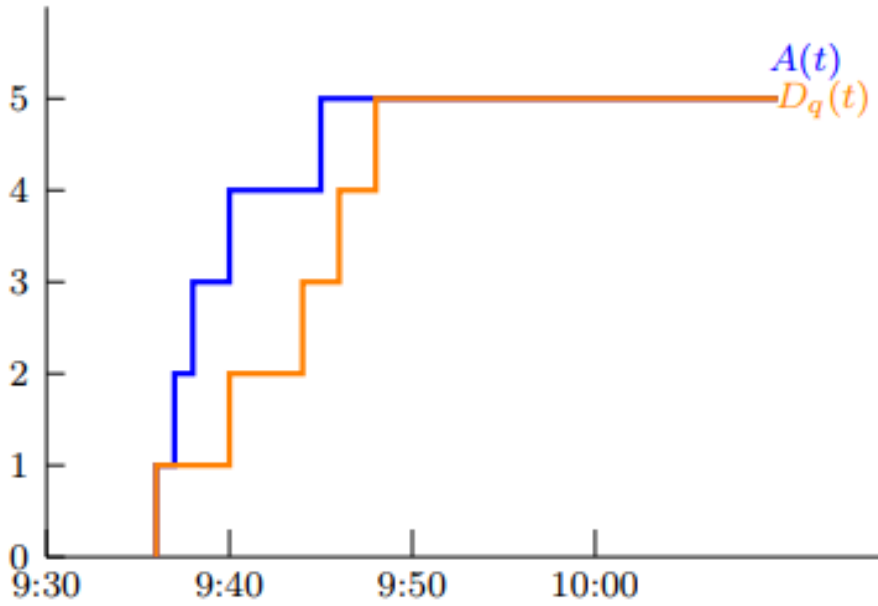



$A(t)$ est le nombre de clients arrivés entre 0 et $t$

# Modélisation de la longueur de la queue

## ❑ Exemple

| Client | Arrivée | Service | Départ du système |
|---|---|---|---|
| 1 | 9:36 | 9:36 | 9:40 |
| 2 | 9:37 | 9:40 | 9:44 |
| 3 | 9:38 | 9:44 | 9:48 |
| 4 | 9:40 | 9:48 | 9:52 |
| 5 | 9:45 | 9:52 | 9:56 |

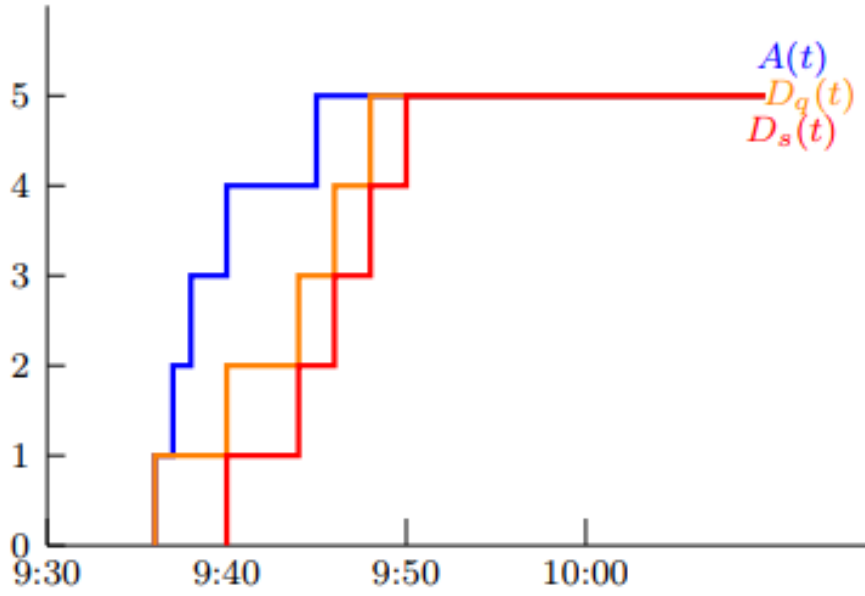



$D_q(t)$ est le nombre de clients qui ont quitté la file entre 0 et $t$

# Modélisation de la longueur de la queue

## ❑ Exemple

| Client | Arrivée | Service | Départ du système |
|---|---|---|---|
| 1 | 9:36 | 9:36 | 9:40 |
| 2 | 9:37 | 9:40 | 9:44 |
| 3 | 9:38 | 9:44 | 9:48 |
| 4 | 9:40 | 9:48 | 9:52 |
| 5 | 9:45 | 9:52 | 9:56 |

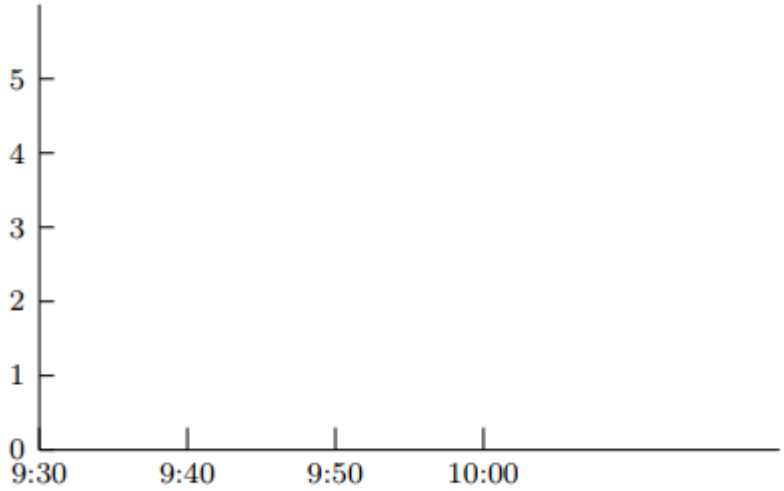



$D_s(t)$ est le nombre de clients qui ont quitté le système entre 0 et $t$

# Modélisation de la longueur de la queue

- Exemple

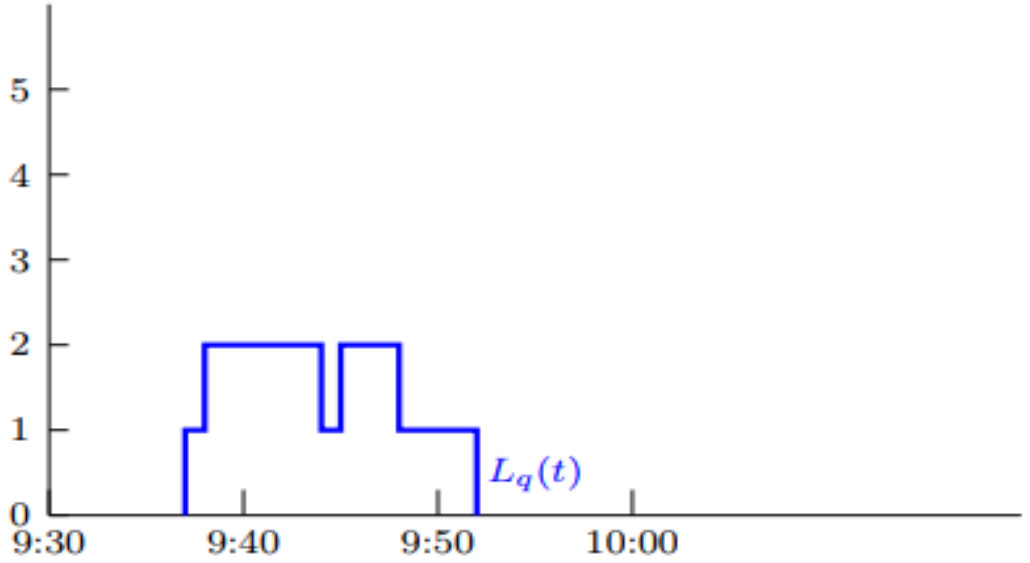
| Client | Arrivée | Service | Départ du système |
|---|---|---|---|
| 1 | 9:36 | 9:36 | 9:40 |
| 2 | 9:37 | 9:40 | 9:44 |
| 3 | 9:38 | 9:44 | 9:48 |
| 4 | 9:40 | 9:48 | 9:52 |
| 5 | 9:45 | 9:52 | 9:56 |



# Modélisation de la longueur de la queue

- Exemple

| Client | Arrivée | Service | Départ du système |
|---|---|---|---|
| 1 | 9:36 | 9:36 | 9:40 |
| 2 | 9:37 | 9:40 | 9:44 |
| 3 | 9:38 | 9:44 | 9:48 |
| 4 | 9:40 | 9:48 | 9:52 |
| 5 | 9:45 | 9:52 | 9:56 |

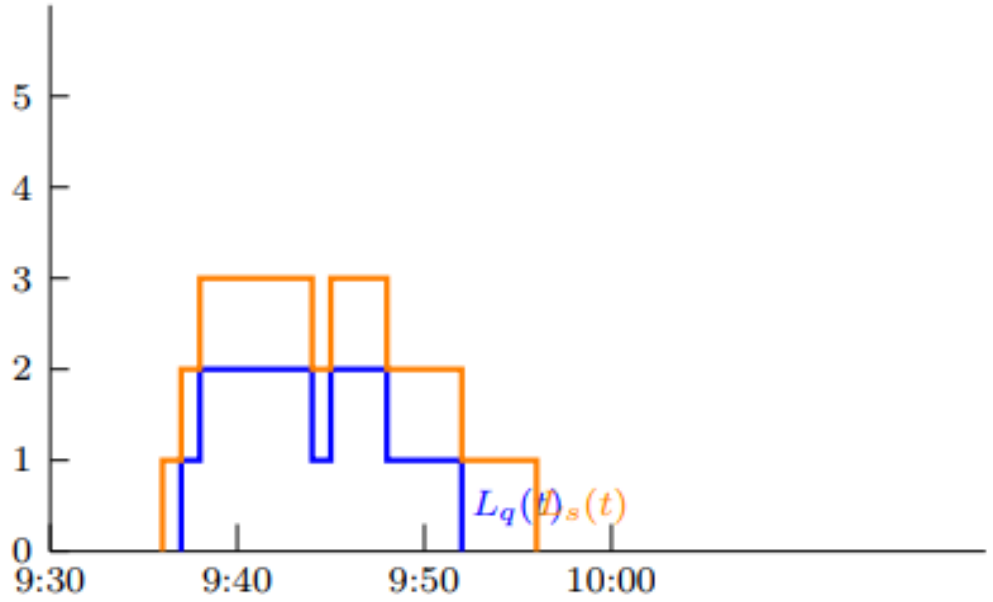



$Lq(t)$ est le nombre de clients dans la file au temps $t$

# Modélisation de la longueur de la queue

❑ Exemple

| Client | Arrivée | Service | Départ du système |
|---|---|---|---|
| 1 | 9:36 | 9:36 | 9:40 |
| 2 | 9:37 | 9:40 | 9:44 |
| 3 | 9:38 | 9:44 | 9:48 |
| 4 | 9:40 | 9:48 | 9:52 |
| 5 | 9:45 | 9:52 | 9:56 |




$Ls(t)$ est le nombre de clients dans le système au temps $t$;

# Récapitulatif

| | |
|---|---|
| le moyen d'arrivées par unité de temps (taux d'arrivée) | $\lambda = \frac{\mathbb{E}(A_t)}{t}$ |
| L'intervalle de temps entre deux arrivées consécutives | $\lambda = \frac{1}{\mathbb{E}(T_1)}$ |
| le nombre moyen de services (taux de service) | $\mu = \frac{1}{\mathbb{E}(S_n)}$ |
| la longueur de la queue | $Q_t = A_t - D_t$ |
| le temps d'attente du *n+1* client présent dans la fille | $W_{n+1} = \max(W_n + S_n - T_{n+1}, 0)$ |
| L'instant de sortie du *n* client | $\sigma_n = s_n + W_n + S_n$ |

# Étude de la file en régime stationnaire

## ❑ Exercice 1 : Temps d'attente d'un train

On considère une voie ferrée sur laquelle les passages des trains sont séparés par des durées (durée entre deux trains successifs) de deux types possibles :

- 90% de ces durées sont constantes et égales à 6 mn.
- 10% de ces durées sont constantes et égales à 54 mn.

1- Calculer la durée moyenne séparant deux trains successifs

2- Un individu arrive à un instant quelconque. Au bout de combien de temps en moyenne pourrait-il prendre un train ? On fera le calcul de probabilité pour que l'individu arrive pendant un intervalle court entre deux trains. On en déduira le temps d'attente résiduelle.

# Étude de la file en régime stationnaire

## ❑ Solution: Temps d'attente d'un train

1- $E[X] = 0.9 * 6 + 0.1 * 54 = 10.8$ soit 10 minutes 48 secondes.

2- Sur 100 intervalles de temps, il y a en moyenne 90 intervalles courts de 6 minutes qui durent 540 minutes et 10 intervalles longs de 54 minutes qui durent 540 minutes. Donc sur 1080 minutes, il y en a 540 qui correspondent à des intervalles courts et 540 qui correspondent à des intervalles longs.

Donc pour un voyageur qui arrive à un instant quelconque, il a une probabilité 0.5 d'arriver pendant un intervalle court (et pas du tout 90%) et 0.5 d'arriver pendant un intervalle long.

Si on arrive pendant un intervalle court, il faudra attendre 3 minutes en moyenne. Si on arrive pendant un intervalle long, on attendra 27 minutes en moyenne.

Donc $E[Y] = 0.5 * 3 + 0.5 * 27 = 15$ minutes.

# Étude de la file en régime stationnaire

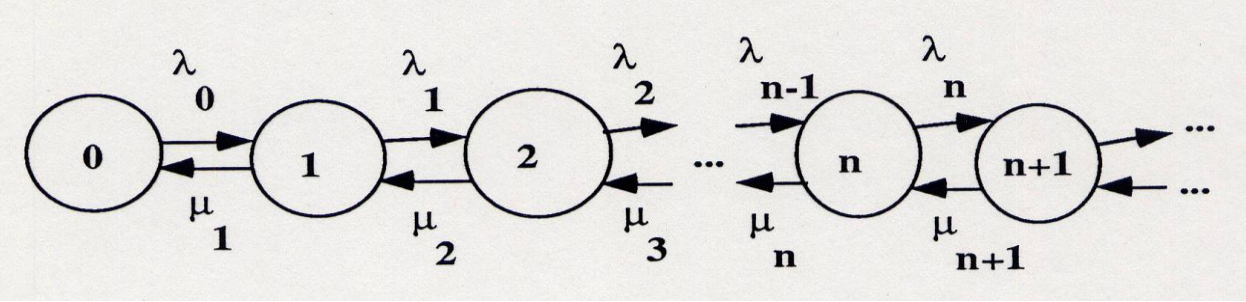
On étudie maintenant le système en temps grand. Les consommateurs arrivent selon un processus de Poisson d'intensité $\lambda$ et le service s'effectue selon la loi $\varepsilon(\mu)$ avec la discipline FCFS.

- 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.

**Principe permettant d'écrire une équation d'équilibre pour tout état n :**

- pour tout état n = 0, 1, 2, …, le taux d'entrée moyen de clients doit être égal au taux de départ moyen.

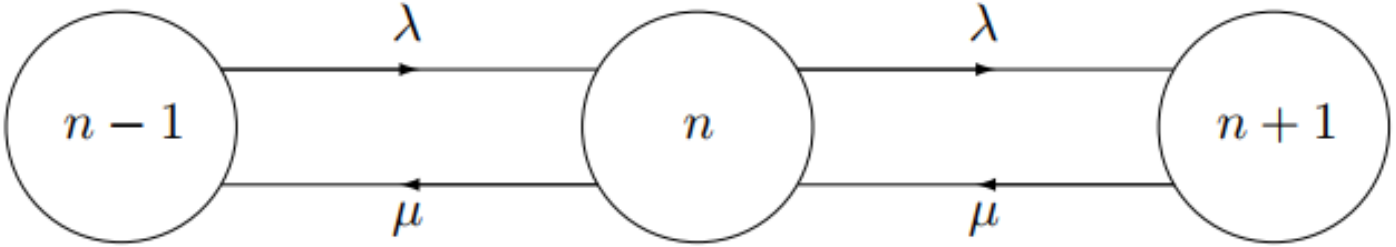
diagramme d'états



# Étude de la file en régime stationnaire

## ❑ Loi de la longueur de la queue

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

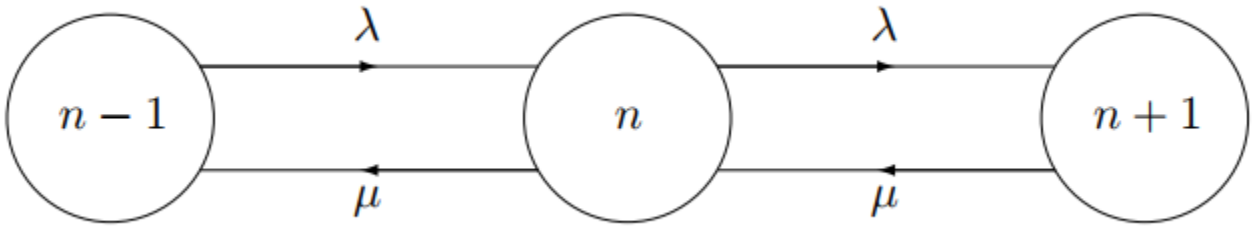



Changements d'états

# Étude de la file en régime stationnaire

## Flux entrant en l'état $n$ :

- Soit la file contient $n$ - 1 personnes (avec une probabilité $\pi_{n-1}$ ) et il en arrive une de plus au taux $\lambda$ ;
- Soit la file contient $n + 1$ personnes (avec une probabilité $\pi_{n+1}$ ) et il en part une au taux $\mu$.
- Ceci conduit au schéma suivant : $n-1 \xrightarrow{\lambda} n \xleftarrow{\mu} n+1$
- Le taux entrant en l'état $n$ est donc $\lambda\pi_{n-1} + \mu\pi_{n+1}$




**Changements d'états**

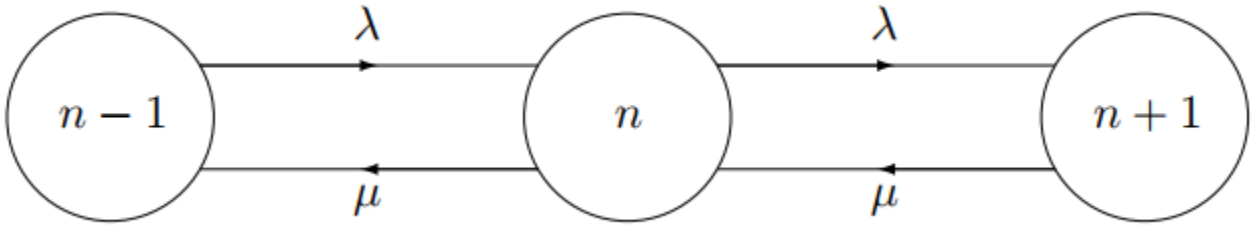
# Étude de la file en régime stationnaire

## Flux sortant en l'état $n$ :

- la file contient $n$ personnes avec une probabilité $\pi_n$ ;
- soit il en arrive une de plus au taux $\lambda$ et la longueur de la file devient $n + 1$ ;
- soit il en part une au taux $\mu$ 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 $(\lambda + \mu)\pi_n$




**Changements d'états**

# Étude de la file en régime stationnaire

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

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

- Pour résoudre ce système, on pose $\rho = \lambda/\mu$ ; on a par hypothèse $\rho \in ]0, 1[$.
- Le paramètre $\rho$ 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

# Étude de la file en régime stationnaire

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

- Les solutions sont $\lambda/\mu=\rho$ 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 $\lambda\alpha + \lambda\beta = \lambda\alpha + \mu\beta$, donc encore $\beta = 0$ et donc $\pi_n = \alpha \rho^n$ et $\alpha = \pi_0$
- Déterminons $\pi_0$
- On a : $\sum_{n=0}^{+\infty} \pi_n = 1$ et $\sum_{n=0}^{\infty} \rho^n = \frac{1}{1-\rho}$

# Étude de la file en régime stationnaire

- La probabilité de trouver la file vide

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

- 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

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

- La v.a. $\boldsymbol{Q_{\infty+1}}$ suit la loi géométrique $G(1-\rho)$ et la longueur moyenne de la queue est

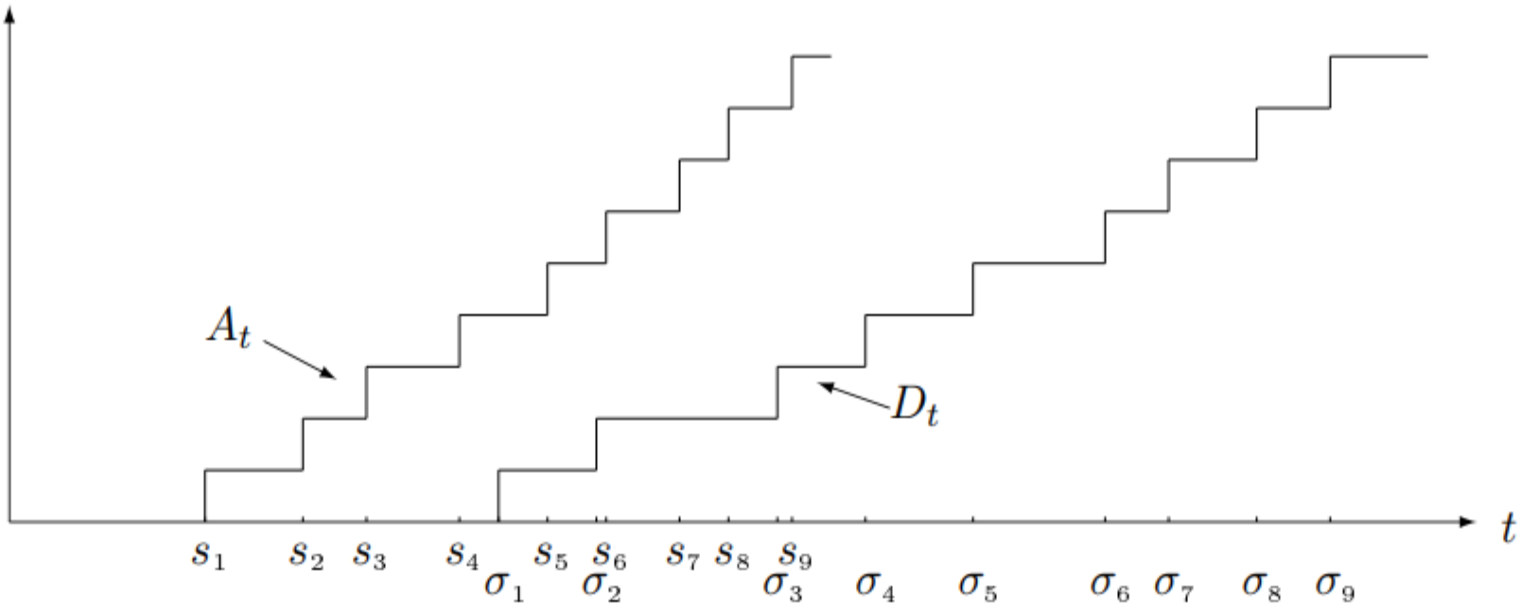
$$\mathbb{E}(Q_\infty) = \frac{\lambda}{\mu - \lambda}.$$

# Étude de la file en régime stationnaire

- **Remarque**

  Lorsque $\lambda > \mu$, on montre que 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

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




Courbes d'arrivées-départs : cas transitoire

# Étude de la file en régime stationnaire

## ❑ 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 $W_\infty$ (waiting time) d'un client générique en régime stationnaire grâce à **la formule des probabilités totales**

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

- 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 $\varepsilon(\mu)$, c'est donc la somme de $n$ v.a. indépendantes de loi $\varepsilon(\mu)$.
- Si $n = 0$, le temps d'attente est nul. Si $n > 1$, cette v.a. conditionnelle suit une loi d'Erlang $E(n; \mu)$.

# Étude de la file en régime stationnaire

## ❑ Lois des temps d'attente et temps de séjour d'un client

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

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

- C'est une loi exponentielle $\mathcal{E}(\mu - \lambda)$ avec un poids en 0 :

$$\mathbb{P}(W_\infty = 0) = 1 - \rho$$

# Étude de la file en régime stationnaire

## ❑ Lois des temps d'attente et temps de séjour d'un client

- 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 $\mathbb{E}(W_\infty \mid Q_\infty = n) = \dfrac{n}{\mu}$

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

Et par suite $$\mathbb{E}(W_\infty) = \frac{\lambda}{\mu(\mu - \lambda)}$$

# Étude de la file en régime stationnaire

## ❑ Lois des temps d'attente et temps de séjour d'un client

- 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 $\varepsilon(\mu - \lambda)$ :

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

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

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

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

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

# Étude de la file en régime stationnaire

## ❑ Récapitulatif

| | |
|---|---|
| La solution de l'équation de balance | $\pi_n = \mathbb{P}(Q_\infty = n) = (1-\rho)\,\rho^n, \quad n \in \mathbb{N}$ |
| La probabilité de trouver la file vide | $\pi_0 = \mathbb{P}(Q_\infty = 0) = 1 - \rho$ |
| La longueur moyenne de la queue | $\mathbb{E}(Q_\infty) = \dfrac{\lambda}{\mu - \lambda}$ |
| L'attente moyenne d'un client | $\mathbb{E}(W_\infty) = \dfrac{\lambda}{\mu(\mu - \lambda)}$ |
| Le temps de séjour moyen d'un client | $\mathbb{E}(W_\infty + S_\infty) = \dfrac{1}{\mu - \lambda}$ |

- Remarquons que la longueur de la file $Q\infty$ dépend du rapport $\rho = \lambda/\mu$ (qui est l'écart relatif entre $\lambda$ et $\mu$), alors que le temps de séjour $W\infty$ dépend de l'écart absolu $\mu - \lambda$.

# Étude de la file en régime stationnaire

- **Exemple**

Considérons deux files indépendantes de paramètres respectifs

| | |
|---|---|
| $\lambda 1 = 4$ personnes/h | $1/\mu 1 = 10$ mn/personne (ou $\mu 1 = 6$ personnes/h), |
| $\lambda 2 = 8$ personnes/h | $1/\mu 2 = 5$ mn/personne (ou $\mu 2 = 12$ personnes/h) |

- Ces paramètres sont dans un rapport de 2.
- Les intensités des deux files coïncident : $\rho_1 = \rho_2 = \frac{2}{3}$
- Les longueurs moyennes $\mathbb{E}(Q_{1,\infty}) = \mathbb{E}(Q_{2,\infty}) = 2$ personnes
- Les temps d'attente et de séjour sont dans un rapport de 2 inversé :

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

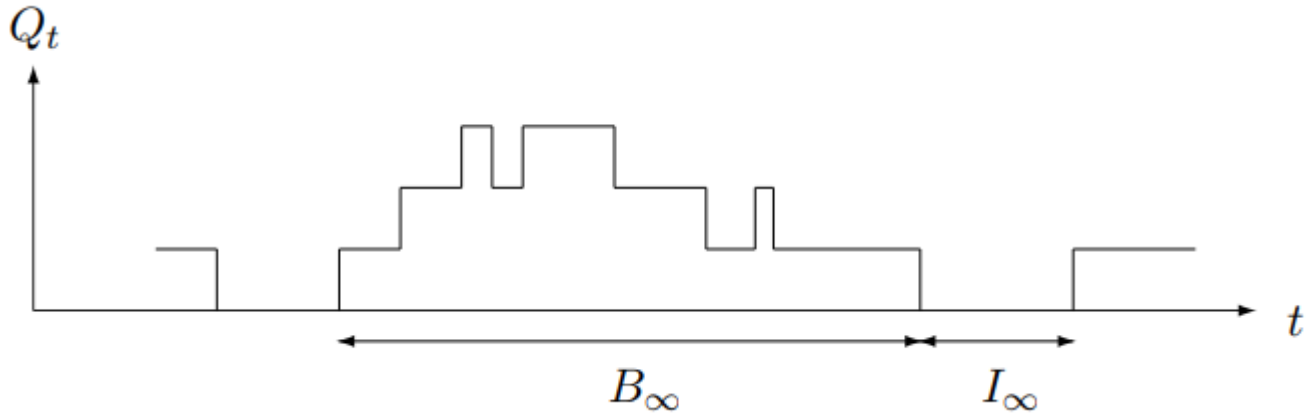
$$\mathbb{E}(W_{1,\infty} + S_{1,\infty}) = 30 \text{ mn}, \quad \mathbb{E}(W_{2,\infty} + S_{2,\infty}) = 15 \text{ mn}.$$

# Étude de la file en régime stationnaire

## ❑ 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 ;**



# Étude de la file en régime stationnaire

## ❑ Loi du temps d'activité du serveur

- **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**
- **Le temps d'activité moyen du serveur est donné par :** $\mathbb{E}(B_\infty) = \frac{1}{\mu - \lambda}$

- Ce résultat s'obtient 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** :

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

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

$$P(Q_\infty \geq 1) = 1 - \mathrm{P}(Q_\infty < 1) = 1 - \mathrm{P}(Q_\infty = 0) = 1 - (1 - \rho) = \frac{\lambda}{\mu}$$

# Étude de la file en régime stationnaire

- **Période de vacance**

- D'autre part, la propriété d'absence de mémoire de la loi exponentielle montre que le temps de vacance $I\infty$ 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 $\mathcal{E}(\lambda)$ : $f_{I_\infty}(t) = \lambda e^{\lambda t}$

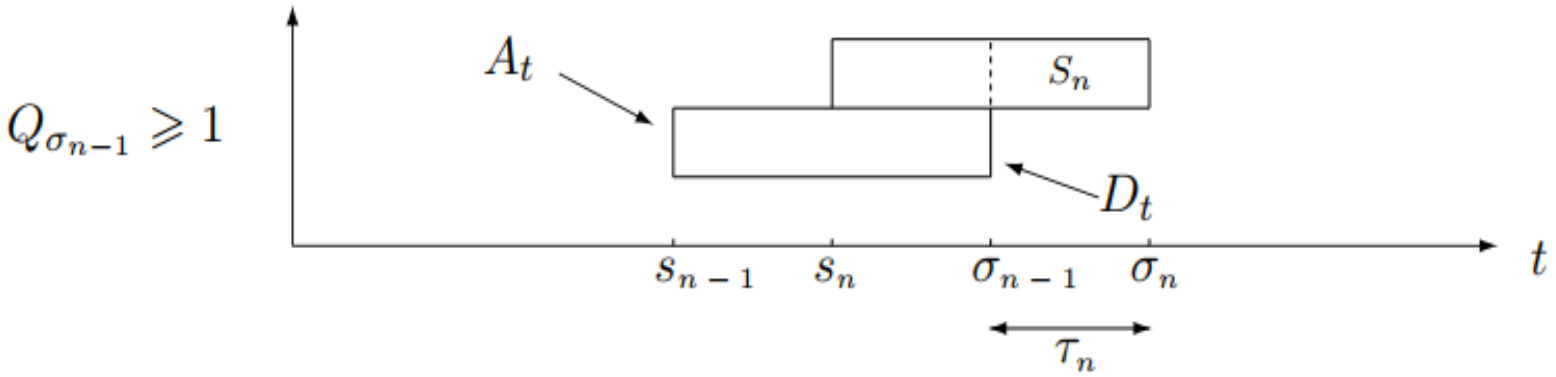
## ❑ Processus des départs

- En régime stationnaire, il est possible de décrire le processus des départs des clients après avoir été servis.
- Introduisons le laps de temps $\tau_n$ séparant les départs des ($n$-1) et $n$ 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) derrière lui en quittant le système,

# Étude de la file en régime stationnaire

## ❑ Processus des départs

- Si $Q_{\sigma_{n-1}} \geq 1$, la $n$ personne était en attente pendant le service de la précédente et se fait servir à partir de l'instant $\sigma_{n-1}$ jusqu'à son instant de sortie $\sigma_n$ pendant une durée $S_n$. Le temps $\tau_n$ n'est donc autre que $\boldsymbol{S_n}$ ;

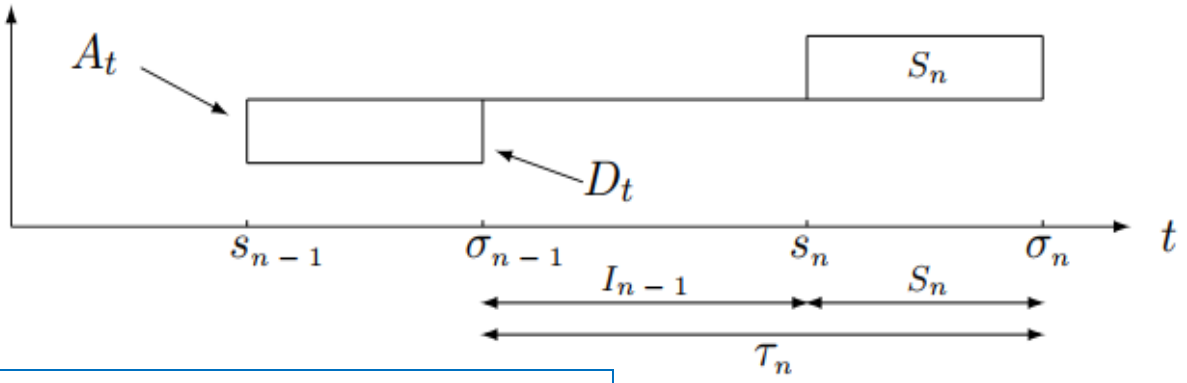


# Étude de la file en régime stationnaire

## ❑ Processus des départs

- si $Q_{\sigma_{n-1}} = 0$, la $(n-1)$ personne laisse un système vide après son départ, le serveur entre dans une période de vacance jusqu'à l'arrivée de la $n$ 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 $\boldsymbol{\tau_n = I_{n-1} + S_n}$.

$Q_{\sigma_{n-1}} = 0$




Ainsi

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

# Étude de la file en régime stationnaire

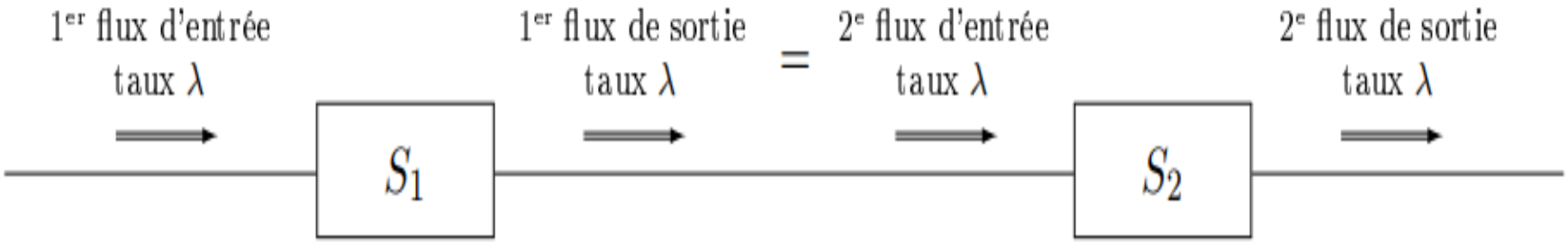
## ❑ Processus des départs

- La v.a. $\boldsymbol{\tau}_{\infty}$ suit donc la loi $\varepsilon(\lambda)$. 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 $\lambda$ (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 $M/M/1$.

# Étude de la file en régime stationnaire

## ❑ Processus des départs

- En effet, les clients se présentent au premier guichet selon un processus de Poisson d'intensité $\lambda$, 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 $\lambda$




**Files en tandem**

# Autres modèles de files d'attente

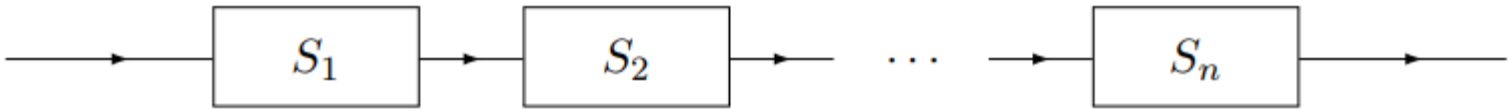
- La file d'attente qu'on a étudié est **une file avec des arrivées poissonniennes** (temps inter-arrivées exponentiels) et **un serveur fournissant un service exponentiel**, elle porte la nomenclature de file ***M/M/1*** (notation de Kendall) ; la lettre *M* fait référence au nom de Markov, la loi exponentielle 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/*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)...

# Autres modèles de files d'attente

- Mentionnons quelques autres modèles importants et couramment utilisés :
  - le $M/M/1/N$ : système avec une salle d'attente de capacité limitée à $N$ places ;
  - le $M/M/n_0$ : service assuré par $n_0$ serveurs ;
  - le $M/M/\infty$ : système assuré par une infinité de serveurs, donc sans attente pour les clients ;
  - le $M/M/n_0/n_0$ : 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 ;

# Autres modèles de files d'attente

- les *G/G/1* : files avec des arrivées et services plus généraux (aléatoires ou non). Par exemple :
  - le *M/D/1* : arrivées poissonniennes et service déterministe (constant) ;
  - le *D/M/1* : arrivées déterministes (régulières) et service exponentiel ;
  - service erlangien : succession de services exponentiels. Un client doit passer successivement par $n$ serveurs proposant des services de durée les v.a. indépendantes $S1, S2, \ldots, Sn$ suivant respectivement les lois $\varepsilon(\mu 1), \varepsilon(\mu 2), \ldots, \varepsilon(\mu n)$. À l'issue du dernier service effectué, le client suivant pourra démarrer son premier service




Service Erlangien

# Autres modèles de files d'attente

- La durée totale des services dispensés pour le client sera $S = S1 + S2 + \cdots + Sn$. 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 $E(\mu 1, \ldots, \mu n)$, elle a pour densité, dans le cas où les paramètres $\mu 1, \ldots, \mu n$ sont tous distincts,

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

- Le service total aura pour durée moyenne $\mathbb{E}(S) = \sum_{k=1}^{n} \frac{1}{\mu_k}$

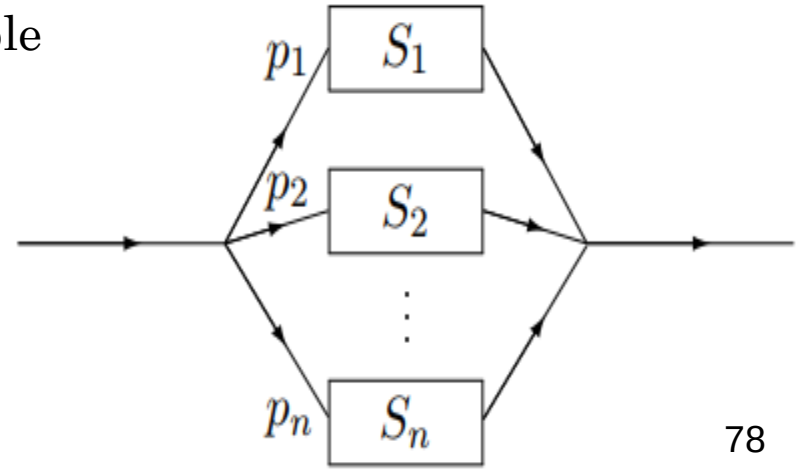
- Si les services sont similaires, $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$, $S$ suit la loi d'Erlang ordinaire $E(n; \mu)$ ;

# Autres modèles de files d'attente

- 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 $S1, S2, \ldots, Sn$ suivant respectivement les lois $\varepsilon(\mu 1), \varepsilon(\mu 2), \ldots, \varepsilon(\mu n)$. Il choisit le $k$ serveur avec probabilité $p_k$ indépendamment du service. 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 $k$ service



# Autres modèles de files d'attente

- La v.a. $N$ est indépendante des v.a. $S1, S2, \ldots, Sn$. 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 :

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

# Autres modèles de files d'attente

- La v.a. $S_N$ suit la loi hyper-exponentielle $H(p1, \ldots, pn; \mu1, \ldots, \mu n)$. Le service choisi au hasard aura pour durée moyenne

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

- Lorsque les serveurs proposent le même service, c'est-à-dire $\mu1 = \mu2 = \ldots = \mu n = \mu$, on récupère, puisque $\sum_{k=1}^{n} p_k = 1$

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

et $S_N$ suit la loi exponentielle $\varepsilon(\mu)$

# 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. $\rho = \frac{\lambda}{\mu} < n_0$, sont données par :

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

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

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

# File M/M/$n_0$

$$\mathbb{P}(W_\infty \leqslant t) = 1 - \frac{\rho^{n_0}}{n_0!\,(1-\frac{\rho}{n_0})}\, e^{-(n_0\mu-\lambda)t}$$

loi $\mathcal{E}(n_0\mu - \lambda)$ 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 :

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

# File M/M/$n_0$

si $n_0 = 2$,

$$\pi_0 = \mathbb{P}(Q_\infty = 0) = \left[1 + \rho + \frac{\rho^2}{2(1-\frac{\rho}{2})}\right]^{-1} = \frac{2-\rho}{2+\rho}$$

$$\mathbb{E}(Q_\infty) = \rho + \pi_0 \frac{\rho^3}{(2-\rho)^2} = \frac{4\rho}{4-\rho^2}$$

$$\mathbb{E}(W_\infty) = \frac{\pi_0 \rho^2}{2(1-\frac{\rho}{2})} \frac{1}{2\mu - \lambda} = \frac{\rho^2}{\mu(4-\rho^2)}$$

si $n_0 = 3$

$$\pi_0 = \mathbb{P}(Q_\infty = 0) = \left[1 + \rho + \frac{\rho^2}{2} + \frac{\rho^3}{6\,(1-\frac{\rho}{3})}\right]^{-1} = \frac{2(3-\rho)}{6+4\rho+\rho^2}$$

$$\mathbb{E}(Q_\infty) = \rho + \pi_0 \frac{\rho^4}{2(3-\rho)^2} = \frac{\rho(18+6\rho-\rho^2)}{(3-\rho)(6+4\rho+\rho^2)}$$

$$\mathbb{E}(W_\infty) = \frac{\pi_0 \rho^3}{6(1-\frac{\rho}{3})} \frac{1}{3\mu - \lambda} = \frac{\rho^3}{\mu(3-\rho)(6+4\rho+\rho^2)}$$

# Fin

# Exercices

## ❑ Exercice 2 : Modèle du dentiste

On considère une file d'attente à un serveur. On suppose que :

- le débit moyen est $\Lambda$,
- le temps moyen de réponse est E[R],
- le temps moyen d'attente est E[W],
- le temps moyen de service est E[S],
- l'espérance de longueur de la file d'attente est E[L],
- le nombre moyen de clients en train d'attendre est E[LW],
- le nombre moyen de clients en train d'être servis est E[LS]
- la probabilité pour que le serveur soit occupé est U.

1- Ecrire une relation entre E[R], E[S] et E[W] (relation 1).

2- Ecrire une relation entre E[L], E[LW] et E[LS] (relation 2).

3- Exprimer E[LS] en fonction de U.

# Exercices

4- Montrer que l'on passe de la relation 1 à la relation 2 en faisant une opération simple et montrer qu'on trouve ainsi une relation connue entre U, $\Lambda$ et E[S].

5- On considère un dentiste. Le nombre moyen de patients présents chez lui est 2.8, le nombre moyen de patients dans la salle d'attente est 2, le nombre moyen de clients arrivant en une heure est 4. Déduire les autres critères de performances et caractéristiques du traitement.

# Exercices

## ❑ Solution : Modèle du dentiste

1- Pour chaque client, on a R = W + S. Ce résultat est donc vrai en terme d'espérance.

2- Pour chaque client, on a L = LW + LS. Ce résultat reste donc vrai en terme d'espérance.

3- Le nombre de clients en train d'être servis est de 1 avec la probabilité U et de 0 avec la probabilité 1-U. On a donc : $E[LS] = 0 * (1 - U) + 1 * U$ d'o`u $E[LS] = U$.

4- En multipliant l'équation (1) par $\Lambda$, on a : $E[R] * \Lambda = E[S] * \Lambda + E[W] * \Lambda$. En appliquant la loi de Little aux systèmes global, file seule puis serveur seul, on a : $E[L] = E[R] * \Lambda$, $E[LW] = E[W] * \Lambda$ et $E[LS] = E[S] * \Lambda$ L'équation (2) vient immédiatement.

# Exercices

## ❑ Solution : Modèle du dentiste

La relation connue entre U, $\Lambda$ et E[S] s'obtient en utilisant le résultat obtenu en appliquant la loi de Little au serveur et celui de la question 3. On a alors : $U = E[S] * \Lambda$

5- L'énoncé donne E[L] = 2.8 client, E[LW ] = 2 client et $\Lambda$ = 4 clients par heure. On en déduit :

- En moyenne, le nombre E[Ls] de clients soignés est 0.8, ce qui revient à dire que le dentiste est occupé à 80% du temps.
- De plus, aller chez le dentiste occupe pendant $E[R] = E[L]/\Lambda = 2.8/4 =$ 0.7 heure, c'est à dire 42 minutes.
- Plus précisément, on passe $E[W] = E[LW]/\Lambda = 2/4 = 0.5$ heure dans la salle d'attente, c'est à dire 30 minutes.
- On a $E[S] = E[LS]/\Lambda = 0{,}8/4 = 0.2h=12min$, on en déduit que le dentiste nous soigne en 12 minutes...

# Exercices

## Exercice 3

Un organisme public est ouvert, chaque jour ouvrable, de 9h à 17h sans interruption. Il accueille, en moyenne, 64 usagers par jour; un guichet unique sert à traiter le dossier de chaque usager, ceci en un temps moyen de 2,5 minutes. Les usagers si nécessaire, font la queue dans l'ordre de leur arrivée ; même si la queue est importante, on ne refuse aucun usager. Une étude statistique a permis de conclure que la durée aléatoire des services suit une loi exponentielle et que le régime les arrivées des usagers forment un processus de Poisson.

1. Donner la notation de Kendall de cette file.
2. Donner l'expression de la probabilité invariante $\pi k$, donner la justification de son existence.

3. Quel sont les temps moyens passés : à attendre ? dans l'organisme par chaque usager ?

4. Quelles sont les probabilités qu'il n'arrive aucun client entre 15H et 16H ? Que 6 clients arrivent entre 16H et 17H ?

5. Quelle est, en moyenne et par heure, la durée pendant laquelle l'employé du guichet ne s'occupe pas des usagers ?

6. Quelle est la probabilité d'observer une file d'attente de 4 usagers, derrière celui en cours de service ?