# Remerciements

Avant toute chose, nous rendons grâce à Dieu Tout-Puissant pour la patience et la force qu'Il nous a accordées tout au long de la réalisation de ce mémoire.

Nos plus sincères remerciements vont à nos chers parents et à notre famille, pour leur amour inconditionnel, leurs sacrifices quotidiens et leurs encouragements constants. Leur présence bienveillante a été notre fondation solide, leur confiance en nous notre plus précieuse motivation.

Nous tenons à exprimer notre profonde reconnaissance à notre encadreur, Mr Hamadouche Djamel, qui a été bien plus qu'un guide : un accompagnateur exceptionnel dans ce travail. Par ses conseils éclairés, sa disponibilité sans faille et son engagement constant, il a su nous orienter avec rigueur et bienveillance à chaque étape de ce parcours. Ce mémoire doit beaucoup à son accompagnement précieux et à son expertise.

Nous n'oublions pas nos amis et proches qui, par leur soutien moral, leurs relectures attentives et leurs paroles réconfortantes, nous ont aidés à surmonter les moments de doute.

Que toutes les personnes qui ont contribué, de près ou de loin, à l'aboutissement de ce travail trouvent ici l'expression de notre gratitude la plus sincère.

# Dédicaces

Je tiens à dédier ce travail à tous ceux qui comptent dans ma vie et qui m'ont accompagné jusqu'ici.

À mon père et ma chère mère, pour leur amour inconditionnel, leur soutien sans faille et leurs innombrables sacrifices. Votre confiance m'a portée dans les moments les plus difficiles. Que dieu vous garde toujours auprès de moi.

À mes sœurs et à mon frère. Merci d'avoir toujours été là, chacun à votre façon. Votre présence, vos mots et votre affection ont souvent fait toute la différence. Ce travail vous est dédié avec tout mon amour.

À toute ma famille, pour leur présence, leurs encouragements et leur patience.

À mes amis proches, qui ont su m'écouter, me motiver, et parfois simplement me faire rire quand j'en avais le plus besoin.

À tout mes enseignants, mes professeurs, que ce rapport vous fait preuve de toutes les connaissances que vous m'avez transmises.

À ma binôme, ma soeur Tassadit. Merci pour ta patience et ton engagement tout au long de cette aventure. Ton soutien, ton sérieux et ta bonne humeur ont rendu ce travail plus riche et plus agréable. Ce mémoire est le fruit de nos efforts partagés, et je suis fière de l'avoir réalisé avec toi.

À moi-même, pour avoir persévérée, même quand le chemin semblait long et incertain.

À la mémoire de mes grands-mères Messaouda et Hedjila, et ma petite cousine Ouiza, qu'elles reposent en paix.

***Belkadi Melyssa***

# Dédicaces

Ce mémoire est le fruit d'un travail mené à deux, avec engagement, rigueur et persévérance. Mais je me permets, humblement, de dédier ces pages à celle que j'ai été dans l'ombre.
À moi-même, À celle qui a traversé l'obscurité seule, blessée par le doute, tentée mille fois par l'abandon, mais qui a tenu bon, sans témoin, sans renfort. À chaque guerre intérieure, chaque excès, chaque silence. Ce mémoire est un aboutissement commun, mais c'est aussi, pour moi, une victoire intime et profonde. Je me le dédie, avec fierté.

**"L'échec n'est pas fatal, le succès n'est pas final : c'est le courage de continuer qui compte."**
**-Winston Churchill-**

***Ammar Tassadit***

# Table des matières

**Introduction générale** **6**

**1 Systèmes de files d'attente markoviens** **8**
Partie I : Processus de Markov . . . . . 8
1.1 Introduction . . . . . 8
1.2 Processus stochastique . . . . . 8
1.3 Processus markoviens . . . . . 9
1.3.1 Générateur infinitésimal d'un processus de Markov 9
1.3.2 Loi stationnaire d'un processus de Markov . . . . 12
1.3.3 Ergodicité d'un processus de Markov . . . . . 12
1.4 Exemples de processus de Markov . . . . . 13
1.4.1 Processus de Poisson . . . . . 13
1.4.2 Processus de naissance et de mort . . . . . 14
1.5 Conclusion . . . . . 15
Partie II : Files d'attente . . . . . 16
1.1 Introduction . . . . . 17
1.2 Caractéristiques d'un système de file d'attente . . . . . 18
1.3 Notation de Kendall . . . . . 18
1.4 Étude du modèle (M/M/1) . . . . . 19
1.4.1 *Étude du processus* . . . . . 20
1.4.2 *Loi stationnaire du système* . . . . . 21
1.4.3 Mesures de performances du système . . . . . 22
1.5 Étude du modèle (M/M/1/K) . . . . . 26
1.5.1 *Étude du processus* . . . . . 26
1.5.2 *Loi stationnaire du système* . . . . . 27
1.5.3 Performances du système . . . . . 28
1.6 Étude du modèle (M/M/s) . . . . . 30
1.6.1 *Étude du processus* . . . . . 30
1.6.2 *Stabilité du système* . . . . . 31
1.6.3 *Performances de la file (M/M/s), liées aux clients* 33
1.7 Étude du modèle (M/M/s/K) . . . . . 35
1.7.1 *Étude du processus* . . . . . 35
1.7.2 *Loi stationnaire du système* . . . . . 36

1.7.3 Caractéristiques (performances) du système . . . 37
1.8 Conclusion . . . 38

**2 Modélisation et optimisation des services bancaires 40**
Partie I : Modélisation du système bancaire . . . 39
2.1 Introduction . . . 40
2.2 Description de CPA Bank . . . 41
2.2.1 Présentation de l'agence . . . 41
2.2.2 Organisation du système . . . 42
2.3 Étude et analyse des cas . . . 43
2.4 Position du problème . . . 46
2.5 Modélisation par type de service . . . 49
2.5.1 Modélisation du service d'Animation commerciale 49
2.5.2 Modélisastion du service financier . . . 51
2.5.3 Modélisation du service crédit . . . 54
2.6 Conclusion . . . 55
Partie II : Optimisation des services du système . . . 55
2.1 Introduction . . . 56
2.2 Optimisation des services bancaires . . . 56
2.2.1 Service d'ouverture de compte . . . 57
2.2.2 Service des opérations courantes . . . 59
2.3 Conclusion . . . 66

**3 Simulations numériques et autres approches 67**
Partie I : Simulations numériques . . . 66
3.1 Introduction . . . 67
3.2 Simulations par mesures de performances . . . 67
3.2.1 Service d'ouverture de compte . . . 68
3.2.2 Service des opérations courantes (M/M/2) . . . 77
3.2.3 Service des opérations courantes (M/M/2/5) . . 84
3.3 Conclusion . . . 91
Partie II : Améliorations possibles et approches intelligentes . 90
3.1 Introduction . . . 92
3.2 Approches classiques et innovantes pour l'optimisation des files d'attente . . . 92
3.2.1 Approches classqiues . . . 92
3.2.2 Approches innovantes . . . 94
3.3 Conclusion . . . 97

**Conclusion générale 96**

**Bibliographie 97**

# Introduction générale

La théorie des files d'attente est une branche des mathématiques utilisée dans la modélisation et l'analyse des systèmes de gestion de masse. La théorie appliquée des files d'attente permet de construire un modèle suffisamment simple, permettant une analyse mathématique contenant suffisamment de détails pour que ses mesures de performance reflètent le comportement réel du système. Les modèles de files d'attente sont largement utilisés dans divers domaines tels que l'informatique, l'ingénierie industrielle, les services d'urgence, les télécommunications, la finance et la logistique militaire.

Ce domaine de recherche trouve son origine dans les travaux de l'ingénieur danois Erlang, menés dès 1908 sur la gestion des réseaux téléphoniques de Copenhague, et publiés en 1917. Il 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. Les apports des mathématiciens Khintchine, Palm, Kendall, Pollaczek et Kolmogorov ont ensuite permis à cette théorie de se développer considérablement.

Dans un contexte de concurrence accrue et d'exigences croissantes en matière de qualité de service, les établissements bancaires sont confrontés à la nécessité d'optimiser leurs performances tout en améliorant l'expérience client. L'attente excessive constitue l'un des points sensibles dans les agences bancaires, notamment au niveau des guichets les plus sollicités. Dans ce travail on s'intéresse à l'agence bancaire CPA Bank de Tizi-Ouzou, où certains services souffrent d'une congestion importante, générant insatisfaction, perte de temps, et baisse d'efficacité opérationnelle.

L'objectif de ce mémoire est d'analyser, modéliser puis optimiser les systèmes de files d'attente dans cette agence bancaire afin de réduire les délais d'attente et d'améliorer le rendement des services. Cette démarche s'inscrit dans une approche scientifique combinant outils mathématiques et simulations numériques, avec pour finalité la proposition d'approches

intelligentes d'amélioration adaptées à la réalité du terrain.

Le premier chapitre de ce travail est composé de deux parties. La première partie traite de fondements théoriques des processus stochastiques, et en particulier des processus markoviens, qui offrent un cadre rigoureux pour modéliser les phénomènes aléatoires observés dans les files d'attente. Cette base probabiliste permet de mieux comprendre les comportements dynamiques des systèmes bancaires.

Dans la deuxième partie, nous abordons les modèles classiques de files d'attente markoviennes ((M/M/1), (M/M/s),(M/M/1/K),...),en détaillant leurs caractéristiques et les mesures de performance associées (temps d'attente, longueur de la file, risque de saturation, etc.). Cette modélisation vise à représenter de façon fidèle les services bancaires étudiés.

Le deuxième chapitre est consacré à la modélisation des différents services de l'agence bancaire CPA Bank de Tizi-Ouzou et à l'optimisation de ces modèles. L'objectif ici est d'identifier les configurations optimales (nombre de serveurs, durée de service) permettant d'améliorer significativement les performances sans investissements démesurés. Les contraintes opérationnelles telles que la capacité maximale du système ou la stabilité du trafic sont rigoureusement prises en compte.

Dans le troisième chapitre, nous avons simulé les différentes performances des services congestionnés, telles que le risque de saturation, la durée de séjour d'un client dans le système et la longueur de la file d'attente, à l'aide du logiciel MATLAB.Les simulations numériques obtenues viennent étayer et confirmer les résultats préalablement établis au chapitre 2. Par la suite, des approches plus intelligentes, qu'il s'agisse d'ajustements organisationnels ou structurels, ou encore de solutions basées sur l'intelligence artificielle sont proposées afin d'affiner les recommandations et atteindre une gestion plus agile des flux de clients.

Ainsi, ce travail articule des fondements théoriques solides et des applications concrètes, dans l'objectif ultime d'optimiser la performance des services bancaires à la CPA Bank de Tizi-Ouzou, tout en replaçant la satisfaction client au cœur des priorités.

# Chapitre 1

# Systèmes de files d'attente markoviens

## Partie I : Processus de Markov

### 1.1 Introduction

La théorie des processus stochastiques (aléatoires) est essentiellement basée sur la notion de stationnarité. En particulier, cette notion a rendu facile la modélisation des plusieurs phénomènes réels, notamment dans le cadre des systèmes de files d'attente. En effet, bon nombre des phénomènes dans la vie réelle montrent une évolution qui ne s'écarte pas trop loin d'un état d'équilibre statistique. Ces phénomènes peuvent être assimilés à des processus aléatoires stationnaires dans la mesure où leurs propriétés statistiques et probabilistes restent les mêmes au cours du temps.

### 1.2 Processus stochastique

**Définition :**

Soit $T \subseteq \mathbb{R}_+$ un espace temps et $E \subseteq R$ un espace des états.
Un processus stochastique $(N_t)_{t\in T}$ est une fonction du temps dont la valeur à chaque instant dépend de l'issue d'une expérience aléatoire.
À chaque instant $t \in T$, N(t) est donc une variable aléatoire. Il peut donc être considéré comme une famille de variables aléatoires (généralement non indépendantes).

Généralement, N(t) représente l'état du processus stochastique au temps t.
– Si T est dans [0 ;∞[ alors le processus stochastique est dit un processus à temps continu.

– Si T est dénombrable, i.e. $T \subseteq N$, $(N_t)_{t\in T}$ est dit un processus à temps discret.

## 1.3 Processus markoviens

En Mathématiques, un processus de Markov est un processus stochastique possèdant la propriété de Markov. Dans un tel processus, la prédiction du futur à partir du présent n'est pas rendue plus précise par des éléments d'information concernant le passé.

**Définition**

On dit que $(N_t)_{t\in T}$ est un processus de Markov si,
pour tout $u \leq s \leq t$ et pour tout $x, i, j \in \mathbf{E}$ :

$$P(N_t = j \mid N_s = i, N_u = x) = P(N_t = j \mid N_s = i) = P_{ij}(t, s),$$

le processus est dit sans mémoire.

Si $P_{ij}(t, s) = P_{ij}(t - s)$, alors le processus $(N_t)_{t\in \mathbf{T}}$ est dit processus markovien homogène.

Dans ce qui suit, on suppose que $(N_t)_{t\in T}$ est markovien et homogène, donc

$$P(N_{t+s} = j \mid N_s = i) = P_{ij}(t + s - s) = P_{ij}(t).$$

Elle est appelée probabilité de transition de l'état i à l'état j pendant le laps de temps t.

Soit $P(t) = (P_{ij}(t))$ avec $(i, j) \in \mathbf{E} \times \mathbf{E}$, elle est appelée matrice de transition du système à l'instant $t$ (matrice de fonctions), et elle caractérise le processus $(N_t)_{t\geq 0}$.

**Proposition 1.1**

Soit $\pi(t) := (P(N_t = j), j \in \mathbf{E})$ la loi t-instantanée du processus $(N_t)_{t\geq 0}$ ,

on a la proposition : $\pi(t) = \pi(0) \times P(t)$.

### 1.3.1 Générateur infinitésimal d'un processus de Markov

On suppose que $\forall (i, j) \in \mathbf{E} * \mathbf{E}$, la fonction $Pij(t)$ est continue en 0, c'est à dire :

$$\lim_{t\to 0^+} P'ij(t) = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} = Pij(0).$$

Pour i$\neq j, q \neq ij = \lim_{t\to 0+} \frac{Pij(t)}{t} = P'ij(0)$, quand elle existe.

Pour i=j, q$ii = \lim_{t\to 0+} \frac{Pij(t)-1}{t} = Pii$, quand elle existe.

On appelle générateur infinitésimal du processus de markov $(N_t)_{t\in T}$, la matrice :

$Q = (qij),\ avec\ (i,j) \in \mathbf{E} * \mathbf{E}.$

**Propriétés :**

1) $\sum_{j\in E} qij = 0, \forall i \in \mathbf{E}.$

En effet, on a $\sum_{j\in E} Pij(t) = 1, \forall i \in \mathbf{E}.$

$$\Rightarrow (\sum_{j\in E} Pij(t))'\big|_{t=0} = 1' = 0 \iff \sum_{j\in E} Pij'(t)\big|_{t=0} = 0.$$

( Les séries sont convergentes)

$$\iff \sum_{j\in E} Pij'(0) = 0 \iff \sum_{j\in E} qij = 0, \forall i \in \mathbf{E}.$$

2) $\sum_{j\in E} qij = 0.$

Ceci implique, q$ii + \sum_{j\in E} qij = 0 \Rightarrow -qii = \sum_{j\in E} qij,\ avec\,(j \neq i).$

Et si on note q$i = -qii$, on aura :

q$i = \sum_{j\in E} qij,\ avec\,(j \neq i).$

**Remarque 1.1**

$q_{ij}$ est dit le taux de transition de l'état $i$ vers l'état $j$.

$q_i$ est dit le taux de transition à partir de $i$.

**Equations de Chapman-Kolmogorov**

Elles décrivent l'évolution des probabilités de transition dans un processus de Markov. Elles s'expriment sous la forme suivante :

$$P_{ik}(s+t) = \sum_{j \in E} P_{ij}(s) P_{jk}(t)$$

pour tout $i, k \in E$ et $s, t \in T$.

**Proposition 1.2**

L'équation différentielle matricielle gouvernant l'évolution d'un processus de Markov en temps continu s'écrit :

$$\frac{d}{dt}P(t) = Q.P(t), \quad \text{avec la condition initiale} \quad P(0) = I_E,$$

où $I_E$ est la matrice identité de l'espace d'état $E$.
La solution de cette équation peut être exprimée sous la forme de la série de Taylor :

$$P(t) = \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!}.$$

On peut également la réécrire sous forme compacte en utilisant l'exponentielle de matrice :

$$P(t) = \exp(Qt).$$

Cette notation met en évidence la structure exponentielle de la solution, ce qui est fondamental dans l'étude des processus de Markov en temps continu.

**Proposition 1.3**

Si l'espace d'états $E$ est fini et si la matrice des taux de transition $Q$ est diagonalisable, c'est à dire il existe une matrice inversible $B$ et une matrice diagonale $D$ telles que :

$$Q = BDB^{-1}$$

où $D$ est une matrice diagonale de la forme :

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

où $\lambda_i$ sont les valeurs propres de $Q$, et $B$ est la matrice des vecteurs propres associés à ces valeurs propres.

Alors, la solution de l'équation différentielle est donnée par :

$$P(t) = B\Delta(t)B^{-1},$$

où $\Delta(t)$ est l'exponentielle matricielle de $D$ :

$$\Delta(t) = \exp(Dt) = \begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n t} \end{bmatrix}.$$

Cette décomposition simplifie le calcul de $P(t)$, car elle réduit l'exponentielle d'une matrice à celle d'une matrice diagonale.

### 1.3.2 Loi stationnaire d'un processus de Markov

Une loi $\pi = (\pi_0, \pi_1, \pi_2, ..., \pi_n)$ est stationnaire si elle est solution du système :

$$(S) \quad \begin{cases} \pi Q = 0, \\ \sum_{j \geq 0} \pi_j = 1. \end{cases}$$

### 1.3.3 Ergodicité d'un processus de Markov

Un processus de Markov est ergodique si, indépendamment de l'état initial, il converge vers une distribution stationnaire unique $\pi$ c'est à dire :

$$\lim_{t \to +\infty} P_{ij}(t) = \pi_j, \quad \forall i, j.$$

Cela signifie que la probabilité d'être dans un état $j$ après un temps suffisamment long ne dépend plus de l'état initial $i$.

## 1.4 Exemples de processus de Markov

### 1.4.1 Processus de Poisson

Soit $(N_t)_{t \in T}$ un processus stochastique, si $(N_t)_{t \in T}$ vérifie les hypothèses suivants :

1) $(N_t)_{t \in T}$ est un processus à accroissements stationnaires, c'est-à-dire que $N_t - N_s$ a la même loi que $N_{t-s}$, ce qui signifie que le contrôle de l'aléa ne dépend pas de la localisation du temps, mais plutôt de $t - s$ (durée d'observation).

2) $(N_t)_{t \in T}$ est un processus à accroissements indépendants (PAI), c'est-à-dire :
$N_{t_n} - N_{t_{n-1}}, N_{t_{n-1}} - N_{t_{n-2}}, \ldots, N_{t_2} - N_{t_1}, N_{t_1}$ sont indépendants, $\forall t_1 < t_2 < \cdots < t_n \in \mathbb{R}^+$, $\forall n \geq 1$.

3) Dans un laps de temps dt très petit, au plus une occurence (relation d'événement) s'est produit .

Alors le processus est dit processus de Poisson de taux d'entrée $\lambda$.

**Proposition 1.4** (Loi de $N_t$ pour un processus de Poisson - **HAMADOUCHE, D**. *Cours de Master RO : Processus stochastiques et files d'attente*).

Soit $(N_t)_{t \geq 0}$ un processus de poisson de taux $\lambda$, avec les hypothèses précédentes,

la loi de $(N_t)_{t \geq 0}$ est donnée par :

$\forall n \in \mathbb{N}$, $P_n(t) = P[N_t = n] = e^{-\lambda t} \frac{(\lambda t)^n}{n!} ....(\mathcal{P})$

c'est à dire $N_t \overset{\text{v.a.}}{\sim} \mathcal{P}(\lambda t)$.

**Remarque 1.2**

$\lambda$ est appelé le taux du processus de Poisson $(N_t)_{t\geq 0}$, et on note $(N_t)_{t\geq 0} \sim P(\lambda)$.

**Processus de Poisson et loi exponentielle**

Considérons un processus de Poisson de paramètre $\lambda$. Soit $(T_n)_{n\geq 0}$ la suite des instants d'occurrence des événements, avec $T_0 = 0$. On définit les temps inter-arrivées par :

$$N_n = T_n - T_{n-1}, \quad \text{pour } n \geq 1.$$

Dans un processus de Poisson, les variables $(N_n)_{n\geq 1}$ sont indépendantes et suivent une loi exponentielle de paramètre $\lambda$ :

$$N_n \sim \text{Exp}(\lambda).$$

La fonction de répartition de $N_n$ est donnée par :

$$F(t) = P(N_n \leq t) = 1 - e^{-\lambda t}, \quad t \geq 0.$$

Ce qui entraîne que la densité de probabilité de $N_n$ est donnée par la dérivée de la fonction de répartition :

$$f_{N_n}(t) = F'(t) = \lambda e^{-\lambda t}, \quad t \geq 0.$$

Cette propriété permet d'interpréter $N_n$ comme le temps d'attente entre deux événements successifs dans un processus de Poisson. De plus, la loi exponentielle est caractérisée par la propriété d'absence de mémoire :

$$P(N_n > s + t \mid N_n > s) = P(N_n > t), \quad \forall s, t \geq 0.$$

Cela signifie que la distribution du temps restant avant l'occurrence d'un événement ne dépend pas du temps déjà écoulé.

Ainsi, dans un processus de Poisson de paramètre $\lambda$, les instants d'occurrence des événements forment un processus ponctuel, et les durées séparant ces instants suivent une loi exponentielle indépendante et identiquement distribuée.

**Proposition 1.5** (Propriété de Markov du processus de Poisson - **HAMADOUCHE, D**. *Cours de Master RO : Processus stochastiques et files d'attente*).

Le processus de Poisson $(N_t)_{t\in T}$ est un procesuus de Markov.

### 1.4.2 Processus de naissance et de mort

On dit que $(N_t)_{t\in T}$ est un processus de naissance et de mort (PNM) de taux $\lambda_n$ et $\mu_n$, si les hypothèses suivantes sont vérifiées :

- H1 - Transitions limitées : à partir d'un état $n \in E$, à l'instant $t$, le processus ne peut évoluer à l'instant $t + dt$ que vers :

  $n + 1$ (augmentation d'un état, appelée naissance),
  $n - 1$ (diminution d'un état, appelée mort),
  ou rester en $n$.

  Dans un laps de temps infinitésimal $dt$, une seule transition peut se produire.

- H2 - Processus à Accroissements Indépendants et Stationnaires (PAIS) : Le processus possède la propriété des accroissements indépendants, c'est-à-dire que les transitions futures ne dépendent que de l'état actuel. Il est également stationnaire, ce qui signifie que les probabilités de transition ne dépendent pas du temps absolu $t$, mais seulement de la durée $dt$.

- H3 - Unicité des événements : à chaque instant $t$, au plus un événement (naissance ou mort) peut se produire.

**Proposition 1.6** (Propriété de Markov des processus de naissance et de mort - **HAMADOUCHE, D.** ***Cours de Master RO : Processus stochastiques et files d'attente***).

Le processus de naissance et de mort $(N_t)_{t \in T}$ est un processus de Markov.

## 1.5 Conclusion

Dans cette première partie du premier chapitre, nous avons exploré les fondements des processus stochastiques, en mettant particulièrement l'accent sur les processus markoviens, qui sont essentiels dans l'analyse et l'optimisation des files d'attente. Nous avons défini les caractéristiques clés des processus stochastiques, soulignant l'importance de l'indépendance et de la stationnarité des accroissements.

Nous avons examiné plusieurs types de processus, tels que les processus de Markov, notamment ; le processus de Poisson et le processus de naissance et de mort. Chacun de ces modèles offre des outils puissants pour modéliser les comportements aléatoires des arrivées et des services dans les systèmes de files d'attente.

L'intégration de ces concepts dans l'étude des files d'attente permet d'optimiser les performances en termes de temps d'attente, de taux de service et d'efficacité opérationnelle. La compréhension approfondie des

processus stochastiques et de leurs applications dans le domaine des files d'attente ouvre la voie à des solutions innovantes et efficaces pour gérer les défis liés à l'afflux de clients dans divers secteurs.

# Partie II : Files d'attente

## 1.1 Introduction

Les systèmes de files d'attente constituent un outil fondamental pour modéliser les processus d'attente dans divers domaines tels que les télécommunications, la logistique ou les services publics et privés. Cette partie présente les concepts de base des files d'attente markoviennes, où le processus des arrivées et le service ont des propriétés markoviennes particulières.
Généralement, l'étude de ces files d'attente porte sur la qualité et le rendement du service fourni. Elle sera caractérisée par la description des éléments suivants :

- La file d'attente : longueur de la file, temps de séjour (d'attente), etc.
- Les serveurs : nombre de clients servis par période d'activité, durée de répit, etc.

Cette étude a pour objectif :

- L'amélioration du fonctionnement dans le cadre des structures actuelles du système ;
- L'analyse des investissements à consentir (augmentation du nombre de serveurs ou de la capacité du système, etc).

Ces analyses visent à améliorer la qualité et le rendement du service fourni.

Le schéma suivant illustre la structure générale du système de files d'attente.

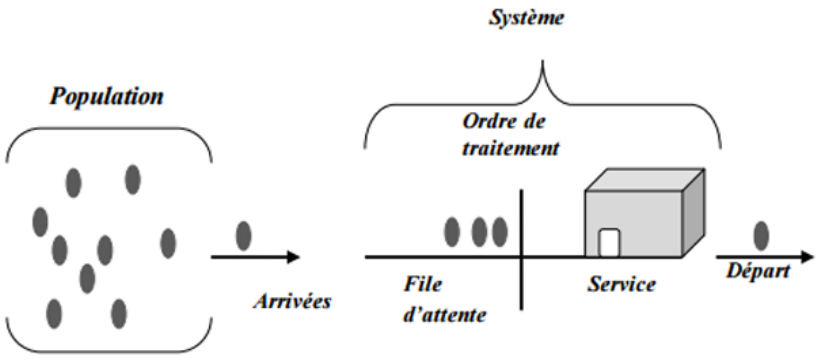


Figure 1.1 : Structure générale d'un système de file d'attente

## 1.2 Caractéristiques d'un système de file d'attente

Les caractéristiques d'une file d'attente sont :

1. Loi des entrées (arrivées) des clients dans le système, par exemple Poisson), notée $L_1$ ;
2. Loi des durées de service (indépendante des arrivées), par exemple exponentielle, notée $L_2$ ;
3. Nombre de serveurs, noté $s$ ($s \geq 1$) ;
4. Capacité du système (longueur maximale permise par le système de la file), notée $K$ ($K \leq \infty$) ;
5. Discipline de service : la discipline de service définit la manière dont les clients sont pris en charge lorsqu'ils attendent dans une file d'attente. Elle influence fortement les performances du système et l'expérience des clients.

   Les principales disciplines sont :

   - FIFO (*First In, First Out*) ou PAPS (Premier Arrivé, Premier Servi).
     - Principe : les clients sont servis dans l'ordre d'arrivée ;
     - Exemple : file d'attente à la caisse d'un supermarché.
   - LIFO (*Last In, First Out*) ou DAPS (Dernier Arrivé, Premier Servi)
     - Principe : le dernier client arrivé est servi en premier ;
     - Exemple : pile d'assiettes propres dans un restaurant (on prend toujours celle du dessus).
6. Systèmes ouverts [O] (accepte tout les clients) ou fermés [F] (accepte des clients particuliers).

## 1.3 Notation de Kendall

Mathématiquement, une file d'attente est la donnée de ces six caractéristiques notées par :

$$(L_1, L_2, s, K, \text{FIFO ou LIFO ou PS ou RS}, [O] \text{ ou } [F]).$$

- $L_1$ : loi des arrivées des clients,
- $L_2$ : loi des durées de service,
- $s$ : nombre de serveurs ($s \geq 1$),

- $K$ : capacité maximale du système ($K \leq \infty$),
- FIFO/LIFO/PS/RS : discipline de service,
- $[O]$ ou $[F]$ : type de système (ouvert/fermé).

**Remarque 1.3 :**

Par défaut, on note ($L_1/L_2/$ s) la file d'attente
$(L_1, L_2, s, \infty, \text{FIFO}, [O])$.

## 1.4 Étude du modèle (M/M/1)

Une file d'attente (M/M/1) peut être définie par le processus stochastique $(N_t)_{t\geq 0}$, qui compte le nombre de personnes dans la file, représentant ainsi la taille de la file d'attente.

On rappelle que, dans ce cas, les instants d'arrivée des clients sont distribués selon un processus de Poisson d'intensité $\lambda$, et que les temps de service sont indépendants (et indépendants du processus d'arrivée) et suivent la loi exponentielle de paramètre $\mu$.

L'indépendance des arrivées implique que $(N_t)_{t\geq 0}$ est un processus de Markov homogène (c'est-à-dire que les probabilités de transition entre les états sont constantes dans le temps), en particulier, un processus de naissance et de mort.

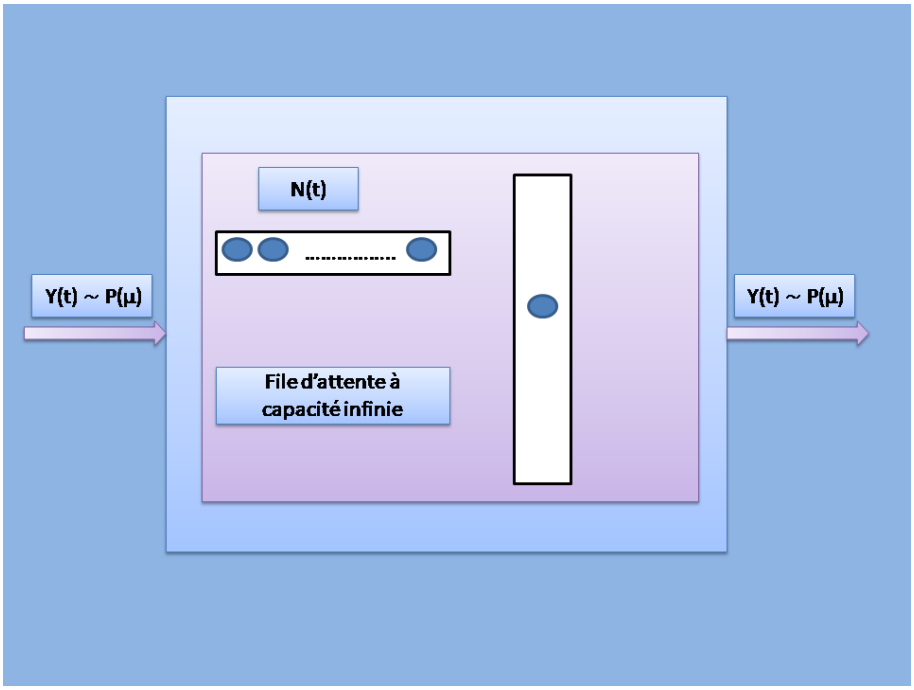


Figure 1.2 : Système d'attente $(M/M/1)$

### 1.4.1 *Étude du processus*

Soit $N_t$ le nombre de clients dans le système.
Soit $M_t$ le nombre de clients dans la file.

On remarque que $(N_t)_{t\geq 0}$ est un processus de naissance et de mort (PNM) avec des taux de naissance $\lambda_n$ et de mort $\mu_n$ à déterminer.

a) *Taux de naissance*

On a :

$$\lambda_n\, dt + o(dt) = \mathbb{P}[N_{t+dt} = n+1 \,|\, N_t = n]$$

$$= \mathbb{P}[X_{t+dt} = X_t + 1] = \mathbb{P}[X_{t+dt} - X_t = 1] = \mathbb{P}[X_{dt} = 1] = \lambda\, dt + o(dt)$$

Donc :

$$\lambda_n = \lambda, \quad \forall n \geq 0.$$

b) *Taux de mort*

On a :

$$\mu_n\, dt + o(dt) = \mathbb{P}[N_{t+dt} = n-1 \,|\, N_t = n]$$

$$= \mathbb{P}[Y_{t+dt} = Y_t + 1] = \mathbb{P}[Y_{t+dt} - Y_t = 1] = \mathbb{P}[Y_{dt} = 1] = \mu\, dt + o(dt)$$

Donc :

$$\mu_n = \begin{cases} \mu, & \text{si } n \geq 1, \\ 0, & \text{si } n = 0. \end{cases}$$

On en déduit que $(N_t)_{t\geq 0} \sim \text{PNM}(\lambda, \mu)$.

### 1.4.2 *Loi stationnaire du système*

Quand le système se stabilise, la loi est donnée par :

$$P_n = \begin{cases} \dfrac{1}{1 + \sum\limits_{n\geq 1} a_n}, & \text{si } n = 0, \\ \dfrac{a_n}{1 + \sum\limits_{n\geq 1} a_n}, & \text{si } n \geq 1, \end{cases}$$

avec

$$a_n = \prod_{i=1}^{n} \left( \frac{\lambda_{i-1}}{\mu_i} \right),$$

à condition que $\sum\limits_{n\geq 1} a_n$ converge, c'est-à-dire

$$\sum_{n\geq 1} \left( \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right) < +\infty.$$

$$\Leftrightarrow \sum_{n\geq 1} \left( \frac{\lambda}{\mu} \right)^n < +\infty.$$

En posant $\rho = \dfrac{\lambda}{\mu}$, on aura :

$$\sum_{n\geq 1} \rho^n < +\infty \Leftrightarrow \rho < 1 \quad \text{si } (\lambda < \mu).$$

Donc, si $\lambda < \mu$ ;

$$P_n = a_n \cdot \frac{1}{1 + \sum\limits_{n\geq 1} \rho^n} = \rho^n \cdot \frac{1}{\sum\limits_{n\geq 1} \rho^n}.$$

$$\Rightarrow P_n = \rho^n \cdot (1 - \rho), \forall\ n \geq 0.$$

### 1.4.3 Mesures de performances du système

***Performances associées aux clients***

*1) Nombre moyen de clients dans le système*

$$\eta = E(N_t) = \sum_{n\geq 0} nP_n = \sum_{n\geq 0} n\rho^n(1-\rho)$$

$$= (1-\rho)\cdot\rho\sum_{n\geq 1} n\rho^{n-1} = (1-\rho)\cdot\rho\sum_{n\geq 1}(\rho^n)'$$

$$= (1-\rho)\cdot\rho\left(\frac{1}{1-\rho}\right)',$$

(car les séries $\sum \rho^n$ et $\sum n\rho^{n-1}$ sont convergentes).

$$\eta = (1-\rho)\cdot\rho\cdot\frac{1}{(1-\rho)^2}$$

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

$$\Rightarrow \eta = \frac{\lambda}{\mu-\lambda}.$$

*2) Nombre moyen de clients dans la file*

Soit $M_t$ : nombre de clients dans la file.

- *Première méthode*

$$\eta_q = E(M_t)$$

$$= \sum_{n\geq 0} n\cdot\mathbb{P}[M_t = n]$$

$$= \sum_{n\geq 0} n\cdot\mathbb{P}[N_t = n+1]$$

$$= \sum_{n\geq 0} n\cdot P_{n+1} \quad (\text{posons } n' = n+1 \Rightarrow n = n'-1)$$

$$= \sum_{n'\geq 1}(n'-1)\cdot P_{n'}$$

$$= \sum_{n' \geq 1} n' \cdot P_{n'} - \sum_{n' \geq 1} P_{n'}$$
$$= \eta - (1 - P_0)$$
$$\eta_q = \eta - \rho = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu} = \frac{\lambda\mu - \lambda(\mu - \lambda)}{\mu(\mu - \lambda)} = \frac{\lambda^2}{\mu(\mu - \lambda)},$$
$$\Rightarrow \eta_q = \frac{\lambda}{\mu} \cdot \frac{\lambda}{\mu - \lambda}.$$

- *Deuxième méthode*

On a $N_t = M_t + Z$ où :

$$Z = \begin{cases} 1, & \text{si } N_t \geq 1 \\ 0, & \text{si } N_t = 0, \end{cases}$$

c'est à dire $Z \overset{\text{v.a.}}{\sim} \text{Ber}(p)$ avec

$$p = \mathbb{P}[Z = 1] = \mathbb{P}[N_t \geq 1] = 1 - \mathbb{P}[N_t = 0] = 1 - P_0 = \rho$$

$$\Leftrightarrow E(N_t) = E(M_t + Z)$$
$$\eta = E(M_t) + E(Z) \Rightarrow \eta = \eta_q + \rho,$$
$$\eta_q = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu}.$$

En utilisant le résultat de la première méthode :

$$\eta_q = \frac{\lambda}{\mu} \cdot \frac{\lambda}{\mu - \lambda}.$$

*3) Temps moyen d'attente (séjour) d'un client dans le système*

On a la formule de Little

$$E(W) = W = \frac{\text{Nombre moyen de clients dans le système}}{\text{Taux d'entrée global}}$$
$$\Leftrightarrow W = \frac{\eta}{\Lambda} = \frac{\frac{\lambda}{\mu - \lambda}}{\lambda},$$
$$\Rightarrow W = \frac{1}{\mu - \lambda}.$$

De même : on a $\Lambda = \sum_{i \geq 0} \lambda_i P_i = \sum \lambda P_i = \lambda \sum_{i \geq 0} P_i = \lambda$,

d'où $W = \frac{\eta}{\lambda}$,

*4) Temps moyen d'attente d'un client dans la file*

$$E(W_q) = W_q = \frac{\text{Nombre moyen de clients dans la file}}{\text{Taux d'entrée global}}$$
$$= \frac{\eta_q}{\Lambda} = \frac{\eta \cdot \rho}{\lambda},$$
$$\Rightarrow W_q = \frac{\lambda}{\mu(\mu - \lambda)}.$$

**Remarque 1.4 :**

$(\eta, \eta_q, W, W_q)$ sont appelées les caractéristiques (performances) du système associées aux clients.

***Performances liées au serveur***

*1) Période de répit et durée de période de répit R*

Soit $R$ une variable aléatoire et $\overline{R}$ sa moyenne.

$$F_R(t) = \mathbb{P}[R \leq t] = 1 - \mathbb{P}[R > t].$$
$$\mathbb{P}[R > t] = \mathbb{P}[X_{s+t} = X_s] = \mathbb{P}[X_{s+t} - X_s = 0].$$

Une période de répit commence à la fin du service du dernier client et se termine à l'entrée d'un nouveau client dans le système, c'est-à-dire :

$$\mathbb{P}[R > t] = \mathbb{P}(X_{s+t} - X_s = 0) \text{ avec}$$
$$(X_t)_{t \geq 0} \sim \mathcal{P}(\lambda) \text{ qui est à accroissement stationnaire}$$

$$\text{donc } \mathbb{P}(X_t = 0) = e^{-\lambda t} \cdot \frac{(\lambda t)^0}{0!} = e^{-\lambda t}$$
$$\Rightarrow \mathbb{P}[R > t] = e^{-\lambda t} \quad \text{(fonction de survie)}$$
$$\Rightarrow F_R(t) = \begin{cases} 1 - e^{-\lambda t}, & \text{si } t \geq 0 \\ 0, & \text{si } t < 0, \end{cases}$$

$\Rightarrow R \sim \mathcal{E}(\lambda)$, c'est-à-dire que $R$ suit une loi exponentielle de paramètre $\lambda$.

d'où : $\overline{R} = \frac{1}{\lambda}$.

*2) Période de répit (oisivité)*

Soit A le nombre de périodes de répit et $\overline{A}$ le nombre moyen. On a, sur la période d'activité globale T, un temps de répit qui est un certain pourcentage P du temps T.

$$\mathbb{P}(N_t = 0) = P_0 = 1 - \rho \quad \text{(en régime permanent)}$$

c'est-à-dire le temps de répit global est :

$$P \cdot T = (1 - \rho) \cdot T$$

$$\Rightarrow \overline{A} \cdot \overline{R} = (1 - \rho) \cdot T$$
$$\Rightarrow \overline{A} = \lambda(1 - \rho) \cdot T.$$

**Remarque 1.5 :**

Le nombre de périodes d'activité est, à une unité près, égal au nombre de périodes de répit à cause de l'alternance (activité-répit).

C'est-à-dire, en moyenne :

$$\overline{B} = \overline{A} = (1 - \rho)\lambda T,$$

ce qui représente le nombre moyen de périodes d'activité $\overline{B}$ (en régime permanent).

*3) Durée d'une période d'activité*

Soit $C$ la durée d'une variable d'activité, $C$ est une variable aléatoire.

Si le serveur est oisif pendant $(1 - \rho)$ unité de temps, donc il est en activité pendant :

$$T - (1 - \rho) \cdot T = \rho \cdot T \text{ unité de temps.}$$

$$\text{On a } \overline{C} \cdot \overline{B} = \rho \cdot T$$
$$\Rightarrow \overline{C} = \frac{\rho \cdot T}{\overline{B}} = \frac{\rho \cdot T}{(1 - \rho) \cdot \lambda \cdot T},$$
$$\text{D'où} \quad \overline{C} = \frac{1}{\mu - \lambda}.$$

*4) Le nombre $N$ de clients servis par période d'activité*

On a

$$N \cdot D = \overline{C}$$
$$\Rightarrow N = \frac{\overline{C}}{D} = \frac{\frac{1}{\mu-\lambda}}{\frac{1}{\mu}},$$
$$\text{D'où } N = \frac{\mu}{\mu-\lambda} = \frac{1}{1-\rho}.$$

**Remarque 1.6 :**

$(R, A, B, C, N)$ sont appelées caractéristiques (performances) du système liées aux serveurs.

## 1.5 Étude du modèle (M/M/1/K)

C'est un système qui consiste à servir des clients selon leur ordre d'arrivée où les arrivées sont poissonniennes ($\lambda$) et La durée de service (indépendante des arrivées) suit une loi exponentielle de paramètre $\mu$, notée $\mathcal{E}(\mu)$ , avec une capacité du système limitée à $K$.

### 1.5.1 *Étude du processus*

Soit $N_t$ : nombre de clients dans le système. Le processus $(N_t)_{t\geq 0}$ est un processus de naissance et de mort (PNM) de paramètres $(\lambda_n, \mu_n)$ à déterminer.

*a) Taux de naissance*

$$\lambda_n\, dt + o(dt) = \mathbb{P}[N_{t+dt} = n+1 \mid N_t = n] = \mathbb{P}[\text{avoir une naissance}]$$
$$= \begin{cases} \mathbb{P}[X_{t+dt} = X_t + 1], & \text{si } 0 \leq n \leq K-1 \\ 0, & \text{si } n \geq K \end{cases}$$
$$= \begin{cases} \mathbb{P}[X_{t+dt} - X_t = 1], & \text{si } 0 \leq n \leq K-1 \\ 0, & \text{si } n \geq K \end{cases}$$
$$= \begin{cases} \mathbb{P}[X_{dt} = 1], & \text{si } 0 \leq n \leq K-1 \\ 0, & \text{si } n \geq K \end{cases}$$
$$= \begin{cases} \lambda\, dt + o(dt), & \text{si } 0 \leq n \leq K-1 \\ 0, & \text{si } n \geq K \end{cases}$$

$$\Rightarrow \lambda_n = \begin{cases} \lambda, & \text{si } 0 \leq n \leq K-1 \\ 0, & \text{si } n \geq K \end{cases}$$

*b) Taux de mort*

$$\mu_n\, dt + o(dt) = \mathbb{P}[N_{t+dt} = n-1 \mid N_t = n] = \mathbb{P}[\text{avoir une mort}]$$

$$= \begin{cases} \mathbb{P}[Y_{t+dt} = Y_t + 1], & \text{si } 1 \leq n \leq K \\ 0, & \text{si } n > K \text{ ou } n = 0 \end{cases}$$

$$= \begin{cases} \mathbb{P}[Y_{t+dt} - Y_t = 1], & \text{si } 1 \leq n \leq K \\ 0, & \text{si } n = 0 \end{cases}$$

$$= \begin{cases} \mathbb{P}[Y_{dt} = 1], & \text{si } 1 \leq n \leq K \\ 0, & \text{si } n = 0 \end{cases}$$

$$= \begin{cases} \mu\, dt + o(dt), & \text{si } 1 \leq n \leq K \\ 0, & \text{si } n = 0 \end{cases}$$

$$\Rightarrow \mu_n = \begin{cases} \mu, & \text{si } 1 \leq n \leq K \\ 0, & \text{si } n = 0 \end{cases}$$

### 1.5.2 *Loi stationnaire du système*

La loi du système en régime permanent (quand elle existe) est donnée par :

$$P_n(t) = P_n = \begin{cases} P_0 = \dfrac{1}{1 + \sum\limits_{n \geq 1} a_n}, & \text{si } n = 0 \\ P_0 \cdot a_n, & \text{si } n \geq 1 \end{cases}$$

À condition que $\sum\limits_{n \geq 1} a_n < +\infty$ avec $a_n = \prod\limits_{i=1}^{n} \left( \dfrac{\lambda_{i-1}}{\mu_i} \right)$.

Comme $\sum\limits_{n \geq 1} a_n = \sum\limits_{n=1}^{K} a_n$ (somme finie) donc elle existe toujours.

$\Rightarrow$ La loi stationnaire existe toujours $\forall \lambda, \forall \mu$.

D'autre part, on a :

$$a_n = \begin{cases} \prod\limits_{i=1}^{n} \left( \dfrac{\lambda_{i-1}}{\mu_i} \right) = \left( \dfrac{\lambda}{\mu} \right)^n, & \text{si } 1 \leq n \leq K \\ 0, & \text{si } n > K \end{cases}$$

et

$$P_n = \begin{cases} P_0 = \dfrac{1}{1 + \sum_{n=1}^{K} a_n}, & \text{si } n = 0 \\ P_0 \cdot \rho^n, & \text{si } 1 \leq n \leq K \text{ avec } \rho = \dfrac{\lambda}{\mu} \end{cases}$$

$\Rightarrow$

$$P_n = \begin{cases} \dfrac{1}{\sum_{n=0}^{K} \rho^n}, & \text{si } n = 0 \\ \dfrac{\rho^n}{\sum_{n=0}^{K} \rho^n}, & \text{si } 1 \leq n \leq K \end{cases}$$

D'où $P_n = \dfrac{\rho^n}{\sum_{n=0}^{K} \rho^n}$ pour $0 \leq n \leq K$.

*Premier cas : Si* $\lambda = \mu$

$\rho = 1 \Rightarrow P_n = \dfrac{1}{K+1}$ pour $0 \leq n \leq K$.

*Deuxième cas : Si* $\lambda \neq \mu$

$\rho \neq 1 \Rightarrow P_n = \dfrac{\rho^n}{\sum_{i=0}^{K} \rho^i}$ pour $0 \leq n \leq K$.

*Finalement*

$$P_n = \begin{cases} \dfrac{\rho^n(1-\rho)}{1-\rho^{K+1}}, & \text{si } 0 \leq n \leq K,\ \rho \neq 1 \\ \dfrac{1}{K+1}, & \text{si } 0 \leq n \leq K,\ \rho = 1. \end{cases}$$

### 1.5.3 Performances du système

***Performances associées aux clients***

1) *Nombre moyen de clients dans le système*

On a $\eta = \sum_{n=0}^{K} nP_n$

Si $\rho = 1$ : $$\eta = \frac{1}{K+1}\sum_{n=0}^{K} n = \frac{K}{2}$$

Si $\rho \neq 1$ : $$\eta = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}$$

Détail des calculs pour $\rho \neq 1$ :

$$\begin{aligned}\eta &= \frac{1-\rho}{1-\rho^{K+1}}\rho\sum_{n=0}^{K} n\rho^{n-1} \\ &= \frac{1-\rho}{1-\rho^{K+1}}\rho\left(\frac{1-\rho^{K+1}}{1-\rho}\right)' \\ &= \frac{\rho[1-(K+1)\rho^K + K\rho^{K+1}]}{(1-\rho)(1-\rho^{K+1})}\end{aligned}$$

donc

$$\eta = \begin{cases} \frac{K}{2}, \rho = 1 \\ \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}, \rho \neq 1. \end{cases}$$

*2) Temps d'attente d'un client dans le système*

D'après la formule de Little, on a

$$\begin{aligned} W &= \frac{\eta}{\Lambda} \text{ où } \Lambda = \sum_{n=0}^{K} \lambda_n p_n \\ &= \lambda p_0 + \lambda p_1 + \cdots + \lambda p_{K-1} + 0 \cdot p_K \\ &= \lambda \sum_{n=0}^{K-1} p_n = \lambda(1-p_K). \end{aligned}$$

$$\text{De même, } \Lambda = \sum_{n=0}^{K} \mu_n p_n = \mu \sum_{n=0}^{K} p_n = \mu(1-p_0)$$

$$\Rightarrow W = \frac{\eta}{\mu(1-p_0)} = \begin{cases} \dfrac{K}{2\mu(1-p_0)}, \rho = 1 \\ \left(\dfrac{\rho}{1-\rho} - \dfrac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}\right)\dfrac{1}{\mu(1-p_0)}, \rho \neq 1. \end{cases}$$

## 1.6 Étude du modèle (M/M/s)

Il s'agit d'un système à une seule file de clients arrivant avec un flux poissonnien $\lambda$ et $s$ serveurs (guichets) où les durées de service $D_i$ $(1 \leq i \leq s)$ sont i.i.d. $\sim \mathcal{E}(\mu)$ avec une capacité illimitée, discipline FIFO et ouvert à tous les clients. La figure suivante illustre le système d'attente $(M/M/s)$.

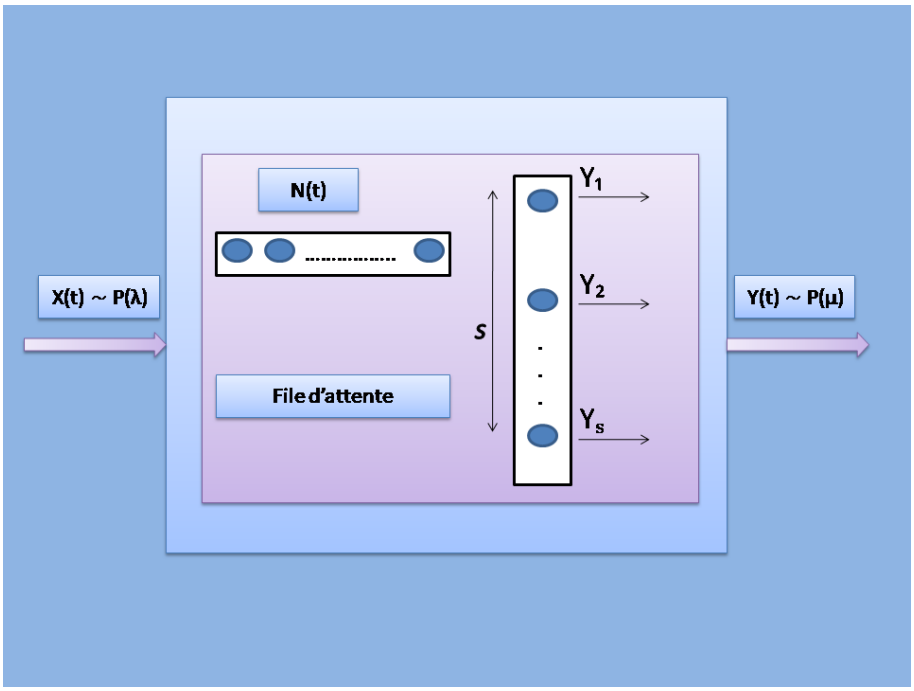


Figure 1.3 : Système d'attente $(M/M/s)$

### 1.6.1 *Étude du processus*

Soit $N_t$ : le nombre de clients dans le système (file + guichets).
On a $(N_t)_{t\geq 0}$ est un processus de naissance et de mort $(\lambda_n, \mu_n)$ où $\lambda_n$ et $\mu_n$ sont des paramètres à déterminer.

*a) Taux de naissance*

$$\begin{aligned}\lambda_n \cdot dt + o(dt) = \ & \mathbb{P}[\text{avoir une naissance pendant } dt \\ & \text{quand le système est à l'état } n \text{ à l'instant } t] \\ & = \mathbb{P}[N_{t+dt} = n+1 \mid N_t = n] \\ & = \mathbb{P}[X_{t+dt} = X_t + 1] \\ & = \mathbb{P}[X_{t+dt} - X_t = 1] = \mathbb{P}[X_{dt} = 1] \\ & = \lambda \cdot dt + o(dt),\end{aligned}$$

$$\Rightarrow \lambda_n = \lambda, \ \forall n \geq 0.$$

*b) Taux de mort*

$$\begin{aligned}\mu_n \cdot dt + o(dt) = \ & \mathbb{P}[\text{avoir une mort pendant } dt \\ & \text{quand le système est à l'état } n \text{ à la date } t] \\ & = \mathbb{P}[N_{t+dt} = n-1 \mid N_t = n]\end{aligned}$$

Cas 1 : $1 \leq n \leq s$

$$= C_n^1 \cdot \mathbb{P}[Y^i_{t+dt} = Y_t^{i_0} + 1]$$

$$= n \cdot (\mu \cdot dt + o(dt)) = n \cdot \mu \cdot dt + o(dt)$$

$$\mu_n = \begin{cases} 0, & \text{si } n = 0 \\ n \cdot \mu, & \text{si } 1 \leq n \leq s. \end{cases}$$

Cas 2 : $n \geq s$

$$\mu_n \cdot dt + o(dt) = C_s^1 \cdot \mathbb{P}[Y^{i_0}_{t+dt} = Y_t^{i_0} + 1]$$

$$= s \cdot (\mu \cdot dt + o(dt))$$

$$= s \cdot \mu \cdot dt + o(dt)$$

$$\Rightarrow \mu_n = s \cdot \mu, \ n \geq s.$$

Donc,

$$\mu_n = \begin{cases} 0, & \text{si } n = 0 \\ (n \wedge s) \cdot \mu, & \text{si } n \geq 1 \end{cases}$$

$$\text{avec} \quad (n \wedge s) = \inf(n, s).$$

### 1.6.2 *Stabilité du système*

La loi stationnaire existe ou bien le régime permanent (stationnaire, stable,...) s'établit si et seulement si :

$$\sum_{n \geq 1} a_n < +\infty \quad \text{avec} \quad a_n = \prod_{i=1}^{n} \left( \frac{\lambda_{i-1}}{\mu_i} \right).$$

On a

$$a_n = \prod_{i=1}^{n} \left( \frac{\lambda_{i-1}}{\mu_i} \right) = \prod_{i=1}^{n} \left( \frac{\lambda}{i \cdot \mu} \right)$$

$$= \frac{\lambda}{\mu} \cdot \frac{\lambda}{2\mu} \cdot \frac{\lambda}{3\mu} \cdots \frac{\lambda}{n\mu} = \frac{\rho^n}{n!},$$

$$\Rightarrow a_n = \frac{\rho^n}{n!}, \quad 1 \leq n \leq s.$$

D'autre part,

$$a_n = \prod_{i=1}^{n} \left( \frac{\lambda_{i-1}}{\mu_i} \right)$$

$$= \prod_{i=1}^{s} \left( \frac{\lambda}{\mu} \right) \cdot \prod_{i=s+1}^{n} \left( \frac{\lambda}{s \cdot \mu} \right),$$

$$\Rightarrow a_n = \frac{\rho^s}{s!} \cdot \left( \frac{\rho}{s} \right)^{n-s}, \quad n \geq s.$$

Ainsi :

$$\sum_{n \geq 1} a_n = \sum_{n=1}^{s} a_n + \sum_{n \geq s+1} a_n = \sum_{n=1}^{s} \frac{\rho^n}{n!} + \frac{\rho^s}{s!} \sum_{n \geq s+1} \left( \frac{\rho}{s} \right)^{n-s}.$$

La loi stationnaire existe si

$$\sum_{n \geq s+1} (\tilde{\rho})^{n-s} < +\infty \quad \text{avec} \quad \tilde{\rho} = \frac{\rho}{s} = \frac{\lambda}{s \cdot \mu}.$$

On pose $m = n - s$,

$$\sum_{m \geq 1} (\tilde{\rho})^m < +\infty$$

$$\iff \tilde{\rho} < 1 \iff \lambda < s \cdot \mu.$$

Finalement, on aura le régime stationnaire qui va s'établir si

$$\lambda < s \cdot \mu \quad \text{(condition de stabilité)}.$$

De là, on a

$$1 + \sum_{n \geq 1} a_n = 1 + \sum_{n=1}^{s} \frac{\rho^n}{n!} + \frac{\rho^s}{s!} \cdot \sum_{m \geq 1} (\tilde{\rho})^m$$

$$= \sum_{n=0}^{s} \left( \frac{\rho^n}{n!} \right) + \frac{\rho^s}{s!} \cdot \sum_{m \geq 1} (\tilde{\rho})^m$$

$$\text{avec } \sum_{m \geq 1} (\tilde{\rho})^m = \tilde{\rho} \cdot \frac{1}{1 - \tilde{\rho}} = \frac{\rho/s}{1 - \rho/s} = \frac{\rho}{s - \rho}$$

$$P_0 = \frac{1}{\sum_{n=0}^{s} \frac{\rho^n}{n!} + \frac{\rho^{s+1}}{s!(s-\rho)}}.$$

$$\text{Si } n \geq 1, \quad P_n = a_n \cdot P_0 = \begin{cases} \dfrac{\rho^n}{n!} \cdot P_0, & \text{si } 0 \leq n \leq s, \\ \dfrac{\rho^s}{s!} \cdot \left(\dfrac{\rho}{s}\right)^{n-s} \cdot P_0, & \text{si } n \geq s. \end{cases}$$

Donc,

$$P_n = \begin{cases} \dfrac{\rho^n}{n!} \cdot P_0, & \text{si } 0 \leq n \leq s, \\ \dfrac{\rho^n}{s! \cdot s^{n-s}} \cdot P_0, & \text{si } n \geq s. \end{cases}$$

**Remarque 1.7 :**

Si $n \geq s \Rightarrow n = s + j,\ j \geq 0$, et $P_n = P_{s+j},\ j \geq 0$, et

$$P_{s+j} = \frac{\rho^s}{s!} \left(\frac{\rho}{s}\right)^j P_0 = (\tilde{\rho})^j \cdot \frac{\rho^s}{s!} \cdot P_0,$$

$$\Rightarrow P_n = P_{s+j} = (\tilde{\rho})^j \cdot \frac{\rho^s}{s!} \cdot P_0, \quad j \geq 0,\ n \geq s.$$

### 1.6.3 *Performances de la file (M/M/s), liées aux clients*

*a) Risque de saturation*

$$\pi^* = \mathbb{P}[N_t \geq s] = \sum_{n \geq s} \mathbb{P}[N_t = n] = \sum_{i \geq 0} \mathbb{P}[N_t = s + i]$$

$$\pi^* = \sum_{i \geq 0} P_s \cdot \tilde{\rho}^i = \frac{\rho^s}{s!} \cdot P_0 \cdot \sum_{i \geq 0} \tilde{\rho}^i$$

$$\pi^* = \frac{\rho^s}{s!} \cdot P_0 \cdot \frac{1}{1 - \frac{\rho}{s}}$$

$$\pi^* = \frac{\rho^s}{s!} \cdot \frac{s}{s - \rho} \cdot P_0.$$

*b) Nombre moyen de clients dans la file*

$$\eta_q = \mathbb{E}(M_t) = \sum_{n \geq 0} n \cdot \mathbb{P}[M_t = n]$$

$$= \sum_{n\geq 0} n \cdot \mathbb{P}[N_t = n+s]$$

$$= \sum_{i\geq 0} i \cdot \mathbb{P}[N_t = s+i]$$

$$= \sum_{i\geq 0} i \cdot P_{s+i} = \sum_{i\geq 0} i \cdot \tilde{\rho}^i \cdot \frac{\rho^s}{s!} \cdot P_0$$

$$= \frac{\rho^s}{s!} \cdot P_0 \cdot \left( \sum_{i\geq 0} i \cdot \tilde{\rho}^i \right) = \frac{\rho^s \cdot \tilde{\rho}}{s!} \cdot P_0 \cdot \left( \sum_{i\geq 1} i \cdot \tilde{\rho}^{i-1} \right)$$

$$= \frac{\rho^{s+1}}{s! \cdot s} \cdot P_0 \cdot \sum_{i\geq 0} (\tilde{\rho}^i)'.$$

(La série $\sum_{i\geq 0} \tilde{\rho}^i$ étant convergente ainsi que $\sum_{i\geq 1} i \cdot \tilde{\rho}^{i-1}$, car $\tilde{\rho} < 1$).

Donc,

$$\eta_q = \frac{\rho^{s+1}}{s! \cdot s} \cdot P_0 \cdot \left( \sum_{i\geq 0} \tilde{\rho}^i \right)'$$

$$= \frac{\rho^{s+1}}{s! \cdot s} \cdot P_0 \cdot \left( \frac{1}{1-\tilde{\rho}} \right)'$$

$$= \frac{\rho^{s+1}}{s! \cdot s} \cdot P_0 \cdot \frac{1}{(1-\tilde{\rho})^2}$$

$$= \frac{\rho^{s+1}}{s! \cdot s} \cdot P_0 \cdot \frac{s^2}{(s-\rho)^2}$$

$$= \frac{\rho^{s+1} \cdot s}{s! \cdot (s-\rho)^2} \cdot P_0$$

$$\Rightarrow \eta_q = \frac{\rho^{s+1}}{(s-\rho)^2 \cdot (s-1)!} \cdot P_0.$$

*c) Durée moyenne d'attente dans la file*

On a par la formule de Little :

$$\overline{W_q} = \frac{\eta_q}{\Lambda} \quad \text{avec : } \Lambda = \sum_{i\geq 0} \lambda_i \cdot P_i = \lambda \cdot \sum P_i = \lambda$$

$$\Rightarrow \overline{W_q} = \frac{\eta_q}{\lambda} = \frac{\rho^{s+1}}{\lambda \cdot (s-\rho)^2 \cdot (s-1)!} \cdot P_0$$
$$= \pi^* \cdot \frac{\rho}{\lambda \cdot (s-\rho)} = \pi^* \cdot \frac{1}{\mu \cdot (s-\rho)}$$
$$\Rightarrow \overline{W_q} = \frac{1}{\mu s - \lambda} \quad (\text{car } \pi^* = 1).$$

## 1.7 Étude du modèle (M/M/s/K)

C'est un système d'attente ouvert, où le nombre de clients ayant accès au service est limité à K clients, qui représente sa capacité et il comporte s serveurs. Les arrivées sont supposées poissoniennes de taux $\lambda$ et les durées de service sont supposées indépendantes, de même loi exponentielle de paramètre $\mu$.

### 1.7.1 *Étude du processus*

Soit $N_t$ : le nombre de clients dans le système. $(N_t)_{t\geq 0}$ est un processus de naissance et de mort de paramètres $(\lambda_n, \mu_n)$ à déterminer.

*a) Taux de naissance*

$$\text{On a} \quad \lambda_n dt + o(dt) = P[N_{t+dt} = n+1 / N_t = n]$$
$$= \begin{cases} P[X_{t+dt} = X_t + 1] & \text{si } 0 \leq n \leq K-1 \\ 0 & \text{si } n \geq K \end{cases}$$
$$= \begin{cases} P[X_{dt} = 1] = \lambda dt + o(dt) & \text{si } 0 \leq n \leq K-1 \\ 0 & \text{si } n \geq K \end{cases}$$
$$\Rightarrow \lambda_n = \begin{cases} \lambda & \text{si } 0 \leq n \leq K-1, \\ 0 & \text{sinon.} \end{cases}$$

b) *Taux de mort*

$$\text{On a} \quad \mu_n dt + od(dt) = P[N_{t+dt} = n-1 / N_t = n]$$

Si $1 \leq n \leq s$

$$P[N_{t+dt} = n-1 / N_t = n] = C_n^1 P[Y_{i_0,t+dt} = Y_{i_0,t} + 1] = C_n^1 P[Y_{i_0,dt} = 1]$$
$$= \frac{n!}{1!(n-1)!}(\mu dt + o(dt)) = n\mu dt + o(dt).$$

Si $s \leq n \leq K$

$$P[N_{t+dt} = n - 1/N_t = n] = C_s^1 P[Y_{i_0,t+dt} = Y_{i_0,t} + 1] = C_s^1 P[Y_{i_0,dt} = 1]$$
$$= \frac{s!}{1!(s-1)!}(\mu dt + o(dt)) = s\mu dt + o(dt).$$
$$\Rightarrow \mu_n = \begin{cases} n\mu & 1 \leq n \leq s, \\ s\mu & s \leq n \leq K, \\ 0 & sinon. \end{cases}$$
$$= \begin{cases} n \wedge s, & 1 \leq n \leq K, \\ 0 & sinon. \end{cases}$$

### 1.7.2 *Loi stationnaire du système*

$$\text{On note} \quad P_n = P[N_t = n],\ 0 \leq n \leq K.$$

On a $P_n = \begin{cases} a_n P_0 & 1 \leq n \leq K, \\ P_0 & n = 0, \end{cases}$

avec $a_n = \prod_{i=1}^{n} \left(\frac{\lambda_{i-1}}{\mu_i}\right)$ et $P_0 = \frac{1}{1+\sum_{n=1}^{K} a_n}$.

Déterminons $(a_n)$ :

Si $1 \leq n \leq s$

$$a_n = \prod_{i=1}^{n} \left(\frac{\lambda_{i-1}}{\mu_i}\right) = \frac{\lambda_0\lambda_1\lambda_2 \ldots \lambda_{n-1}}{\mu_1\mu_2\mu_3 \ldots \mu_n} = \frac{\lambda\lambda\lambda \ldots \lambda}{1\mu 2\mu 3\mu \ldots n\mu}$$
$$= \frac{\lambda^n}{n!\mu^n} = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n$$
$$= \frac{\rho^n}{n!}, \text{ avec } \rho = \frac{\lambda}{\mu}.$$

Si $s < n \leq K$

$$a_n = \prod_{i=1}^{n} \left(\frac{\lambda_{i-1}}{\mu_i}\right) = \frac{\lambda_0\lambda_1\lambda_2 \ldots \lambda_{n-1}}{\mu_1\mu_2\mu_3 \ldots \mu_s\mu_{s+1} \ldots \mu_n} = \frac{\lambda\lambda \ldots \lambda}{\mu_1\mu_2\mu_3 \ldots s\mu s\mu \ldots s\mu}$$
$$= \frac{\lambda^n}{s!\mu^s s^{n-s}\mu^{n-s}} = \frac{\lambda^n}{s!s^{n-s}\mu^n} = \frac{\rho^n}{s!s^{n-s}}.$$

Déterminons $P_0$ :

$$P_0 = \frac{1}{1+\sum_{n=1}^{K} a_n} = \frac{1}{1+\sum_{n=1}^{s} \frac{\rho^n}{n!} + \sum_{n=s+1}^{K} \frac{\rho^n}{s!s^{n-s}}}$$

$$\Rightarrow P_0 = \frac{1}{\sum_{n=0}^{s} \frac{\rho^n}{n!} + \sum_{n=s+1}^{K} \frac{\rho^n}{s!s^{n-s}}}.$$

Finalement, on obtient

$$P_n = \begin{cases} \frac{\rho^n}{n!} P_0 & 1 \leq n \leq s, \\ \frac{\rho^n}{s!s^{n-s}} P_0 & s < n \leq K, \\ \frac{1}{\sum_{n=0}^{s} \frac{\rho^n}{n!} + \sum_{n=s+1}^{K} \frac{\rho^n}{s!s^{n-s}}} & n = 0. \end{cases}$$

### 1.7.3 Caractéristiques (performances) du système

*1) Nombre moyen de clients dans le système*

$$\eta = E(N_t) = \sum_{n=1}^{K} nP_n = \sum_{n=s+1}^{K} nP_n$$

$$= \sum_{n=s+1}^{K} n\frac{\rho^n}{s!s^{n-s}}P_0 = \sum_{n=s+1}^{K} \frac{n}{(s-1)!s^{n-s}}P_0 = \sum_{n=s+1}^{K} \frac{s^{s-1}}{(s-1)!}\frac{\rho^n}{s^n}P_0$$

$$\Rightarrow \eta = P_0.\frac{s^{s-1}}{(s-1)!} \sum_{n=s+1}^{K} n(\bar{\rho})^n.$$

avec $\bar{\rho} = \frac{\rho}{s}$.

*2) Nombre moyen de clients dans la file*

$$\text{Soit} \quad \eta_q = \sum_{h=0}^{K-s} hP[N_t = s+h] = \sum_{h=1}^{K-s} hP_{s+h}$$

$$= \sum_{h=1}^{K-s} h\frac{\rho^{s+h}}{s!s^h}P_0 = \frac{P_0\rho^s}{s!} \sum_{h=1}^{K-s} h\left(\frac{\rho}{s}\right)^h$$

$$\Rightarrow \eta_q = P_0.\frac{\rho^s}{s!} \sum_{h=1}^{K-s} h(\bar{\rho})^h, \text{ avec } \bar{\rho} = \frac{\rho}{s}.$$

*3) Durée moyenne d'attente d'un client dans le système*

On utilise la formule de Little : $\overline{W} = \frac{\eta}{\Lambda}$ avec $\Lambda = \sum_{i=0}^{K} \lambda_i P_i$

$$= \sum_{i=0}^{K-1} \lambda P_i = \lambda(1 - P_K), \text{ tel que } P_K = \frac{\rho^K}{s! s^{K-s}} P_0, \text{ donc on aura}$$

$$\overline{W} = \frac{\eta}{\Lambda} = \frac{\eta}{\lambda(1 - P_K)} = \frac{\eta}{\lambda\left(1 - \frac{\rho^K}{s! s^{K-s}} P_0\right)}.$$

*4) Durée moyenne d'attente d'un client dans la file*

On utilise la formule de Little : $\overline{W}_q = \frac{\eta_q}{\Lambda}$ avec $\Lambda = \sum_{i=0}^{K} \lambda_i P_i$

$$= \sum_{i=0}^{K-1} \lambda P_i = \lambda(1 - P_K) \text{ et } P_K = \frac{\rho^K}{s! s^{K-s}} P_0.$$

$$\text{Ainsi} \quad \overline{W}_q = \frac{\eta_q}{\Lambda} = \frac{\eta_q}{\lambda(1 - P_K)} = \frac{\eta_q}{\lambda\left(1 - \frac{\rho^K}{s! s^{K-s}} P_0\right)}.$$

*5) Risque de saturation*

$$\text{Soit} \quad \pi(\rho, s, K) = P(N_t \geq s) = \sum_{n=s}^{K} P(N_t = n) = \sum_{n=s}^{K} P_n$$

$$= \sum_{n=s}^{K} \frac{\rho^n}{s! s^{n-s}} P_0 = \frac{\rho^s}{s!} P_0 \sum_{n=s}^{K} \frac{\rho^{n-s}}{s^{n-s}} = \frac{\rho^s}{s!} P_0 \sum_{n=s}^{K} \left(\frac{\rho}{s}\right)^{n-s}.$$

On pose $h = n - s$, $\bar{\rho} = \frac{\rho}{s}$ et on obtient

$$\pi(\rho, s, K) = \frac{\rho^s}{s!} P_0 \sum_{h=0}^{K-s} (\bar{\rho})^h.$$

## 1.8 Conclusion

Dans ce chapitre, nous avons examiné les systèmes de files d'attente, en mettant l'accent sur les différents modèles tels que les modèles (M/M/1), (M/M/1/K), (M/M/1/K/[F]), (M/M/s), $(M/M/\infty)$ .... Chaque modèle présente des caractéristiques (performances) spécifiques qui permettent d'analyser et de comprendre le comportement des systèmes de service

en fonction de divers paramètres, tels que le nombre moyen de clients, le nombre de serveurs, le taux d'arrivée des clients, le temps de service, ...Ces modèles nous permettent non seulement de quantifier les performances des systèmes de files d'attente, mais aussi de développer des stratégies d'optimisation adaptées aux besoins spécifiques des entreprises. En comprenant les dynamiques de ces systèmes, nous pouvons mieux anticiper les problèmes potentiels et mettre en œuvre des solutions efficaces pour améliorer le service du client.

# Chapitre 2

# Modélisation et optimisation des services bancaires

## Partie I : Modélisation du système bancaire

### 2.1 Introduction

La modélisation est une méthode scientifique visant à représenter un système réel à l'aide d'outils mathématiques. Son objectif principal est d'analyser le fonctionnement du système, de prédire son comportement et de proposer des solutions pour en optimiser l'efficacité. Cette démarche repose sur la formalisation du problème concret sous forme d'équations, en identifiant les paramètres clés qui gouvernent sa dynamique.

Dans ce travail, nous appliquons cette approche à l'étude des files d'attente dans une agence de la *CPA Bank* à Tizi-Ouzou. Plus précisément, notre modélisation permettra :

- d'évaluer les temps d'attente des clients,
- d'analyser l'allocation des ressources (guichets, personnel),
- de proposer des scénarios d'optimisation pour améliorer la qualité de service.

Cette analyse s'appuiera sur des outils classiques de la théorie des files d'attente, tels que le modèle $(M/M/s)$, afin de fournir des recommandations opérationnelles mathématiquement fondées.

## 2.2 Description de CPA Bank

### 2.2.1 Présentation de l'agence

Le Crédit Populaire d'Algérie (CPA), banque publique fondée en 1966, compte parmi les principaux établissements bancaires du pays. Avec son réseau de 165 agences, le CPA offre une gamme complète de services bancaires aux particuliers, professionnels et entreprises, tout en développant des solutions digitales innovantes (*Mobile CPA*, *e-banking*).

L'agence CPA de Tizi-Ouzou, située stratégiquement au centre-ville, joue un rôle clé dans l'économie locale. Elle dessert une clientèle diversifiée comprenant des particuliers, des professionnels et des entreprises, qu'elles soient physiques ou morales. Elle propose une large gamme de services, notamment des opérations courantes (retraits, dépôts, virements) et des services spécialisés tels que l'octroi de crédits, les produits d'assurance et la gestion de comptes professionnels.

Cette agence accompagne également des entreprises et des institutions spécifiques, notamment dans les secteurs de l'artisanat, de la promotion immobilière (CACOBATPH, AADL), ainsi que dans les domaines de la sécurité sociale (CNAS) et de la retraite. Les heures d'ouverture standard sont de 8h30 à 15h30, avec des pics d'affluence notables entre 10h et 12h, ainsi qu'en début d'après-midi. Malgré le développement de solutions digitales externes comme Wimpay pour les paiements par carte, l'agence repose principalement sur un fonctionnement interne traditionnel, sans système de ticket électronique ni outils avancés de gestion des files d'attente. Les opérations sont assurées manuellement par le personnel, ce qui engendre des temps d'attente prolongés, particulièrement pour les services complexes comme les demandes de crédit ou les dossiers AADL. Ces contraintes opérationnelles, couplées à une forte affluence aux heures de pointe, justifient pleinement la nécessité d'une étude approfondie visant à optimiser la gestion des flux clients et l'allocation des ressources.

***Objectifs de l'étude :***

- Modéliser mathématiquement les flux clients (processus d'arrivée et temps de service).
- Analyser l'allocation actuelle des ressources.
- Proposer des scénarios d'optimisation pour :
  - Réduire les temps d'attente moyens,
  - Améliorer la répartition du personnel,
  - Maintenir la qualité de service.

### 2.2.2 Organisation du système

L'agence CPA Bank s'organise autour d'une structure fluide pour offrir un service optimal. Les opérations courantes sont gérées en salle d'accueil par une équipe dédiée, avec deux guichets polyvalents (Guichet 1 + Caisse 1 / Guichet 2 + Caisse 2). Le pôle clientèle repose sur une équipe commerciale dynamique et cinq conseillers financiers, appuyés par des experts juridico-commerciaux. En soutien, trois services clés opèrent en interne : administratif (logistique), télécompensation (échanges interbancaires) et contrôle interne (conformité et sécurité). Le schéma ci-dessous détaille la structure complète de l'agence, incluant tous les services.

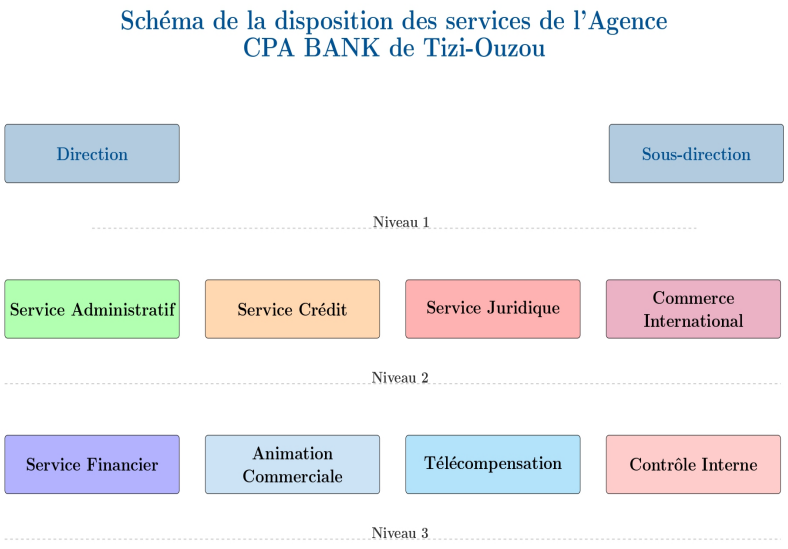


Figure 2.1 : Organisation du système

## 2.3 Étude et analyse des cas

### 2.3.1 Service de la direction

Le service de la directrice joue un rôle central au sein de l'agence. En plus de traiter les réclamations complexes et les demandes spécifiques des clients, la direction assume plusieurs responsabilités clés :

- Supervision du bon fonctionnement de l'ensemble des services.
- Coordination entre les différentes entités internes.
- Arbitrage des décisions importantes.
- Communication avec les autres agences et les clients professionnels.

Dotée d'une excellente aisance relationnelle et de solides capacités de persuasion, la directrice constitue un pilier essentiel de l'agence.

- Nombre de serveur : 1
- Temps de service : Variable selon le type des clients et la complexité des opérations

### 2.3.2 Service de la sous-direction

Responsable de la validation et de la signature des documents, notamment pour les transactions importantes ou les modifications de comptes. Elle gère également les demandes de clôture de compte et les situations nécessitant une validation hiérarchique.

- Nombre de serveurs : 1
- Temps de service : Variable selon le type des clients et la complexité des opérations
- Temps d'attente : Généralement les clients passent directement sans faire la file, ils sont directement pris en charge par la sous-direction.

### 2.3.3 Service administratif

Ce service assure la gestion logistique interne, l'archivage des documents et le support technique aux autres départements. Son rôle est essentiel pour maintenir la fluidité des opérations quotidiennes.

- Nombre de serveurs : 1
- Tâches principales :
  - Archivage des dossiers clients.

  - Gestion des fournitures de bureau.
  - Support technique basique.
- Temps de traitement moyen :
  - Documents courants : 10 à 20 minutes
  - Demandes complexes : 1 à 2 heures
- Taux de demande : 15 à 20 requêtes/jour

### 2.3.4 Service des crédits bancaires

Ce service est chargé du traitement des demandes de prêts.

- Nombre de serveurs : 5 conseillers financiers
- Règle de service : FIFO
- Temps de service :
  - Crédit ordinaire : 20 min
  - Crédit d'investissement : 30 à 40 minutes
- Taux d'arrivée : 5 clients/heure

### 2.3.5 Service commerce international et service juridique

- Nombre de serveurs : 2 (1 commercial + 1 juriste)
- Opérations traitées :
  - Commercial : Commerce international, import/export, visas
  - Juridique : Procurations, oppositions, successions, vérifications légales
- Temps de service : 30 min

### 2.3.6 Service financier (opérations courantes)

Ce service concerne les opérations courantes telles que les renseignements, les retraits, les dépôts, les virements et les mises à jour.

- Nombre de serveurs : 2 (deux caisses fonctionnant en parallèle)
- Temps d'attente : Variable selon la complexité des opérations

Bien que ce secteur ne dispose que de deux guichets, le chef de service fait également partie du personnel et intervient pour les opérations complexes, les explications ou la gestion des tensions.

*Exemple d'analyse de performance (données fictives) :*

- Temps moyen d'attente : 15 min (jusqu'à 1 heure en heure de pointe)
- Temps moyen de traitement :
  - Retrait en DA : 5 min (30% des opérations),
  - Retrait en devise : 10 min,
  - Demande de solde/mises à jour : 3 min,
  - Versement en devise : 30 min,
  - Versement en dinars : 10 min,
  - Virement : 10 min,
  - Renseignements : 5 min.
- Affluence : 100 à 200 clients/jour (pics entre 10 h et 13 h)

### 2.3.7 Service d'ouverture de compte

Les nouveaux clients doivent passer par le service d'animation commerciale, où un seul serveur est disponible pour les accueillir.

- Nombre de serveurs : 1
- Règle de service : Premier arrivé, premier servi (FIFO)
- Temps de service moyen : Variable selon le type d'opération

Ce service repose sur la prise en charge individuelle des clients. Ces derniers arrivent aléatoirement et attendent leur tour s'il y a déjà un client en cours de traitement. Le serveur traite plusieurs tâches en parallèle telles que l'ouverture de compte ( 45 minutes jusqu'à 1 heure) et le renouvellement ou la récupération de carte bancaire (de 15 à 20 minutes).

### 2.3.8 Service de télécompensation

Ce service gère les transactions interbancaires et la compensation des opérations financières.

- Nombre de serveurs : 1 opérateur
- Tâches principales :
  - Traitement des chèques et virements
  - Compensation interbancaire
  - Gestion des transactions électroniques
- Durée de service :

- Chèque local : 5 à 10 minutes
- Virement international : 15 à 30 minutes
- Réconciliation comptable : 20 minutes

### 2.3.9 Contrôle interne

Ce service vérifie la conformité des opérations et assure la sécurité financière.

- Nombre de serveurs : 1 contrôleur
- Tâches principales :
  - Vérification aléatoire des transactions.
  - Audit des procédures.
  - Détection des anomalies.
- Durée de service :
  - Contrôle standard : 30 minutes
  - Audit complet : 1 à 2 heures
  - Investigation complexe : 3 heures

**Remarque 2.1**

Notre étude se concentre principalement sur les cas où la congestion est plus importante, comme pour les services nécessitant une attention particulière ou une période de forte affluence, notamment pour le service d'ouverture de compte et le service Financier (opérations courantes).

## 2.4 Position du problème

Nous considérons un système de files d'attente dans une agence bancaire CPA Bank où trois types de services sont offerts :

- Service d'ouverture de comptes
- Service financier
- Service crédit

Chaque service peut être caractérisé par les paramètres (A,B,C) suivants :

## *A : Flux des arrivées et flux de sorties des clients*

Dans le cadre de la modélisation des arrivées et des sorties des clients de l'agence CPA Bank de Tizi-Ouzou, il est essentiel de représenter mathématiquement ces flux à l'aide d'un processus stochastique de comptage. Ce processus permet de quantifier, en fonction du temps, le nombre de clients se présentant à l'agence ainsi que celui des clients la quittant.

Soit $t_0 = 0 < t_1 < t_2 < \cdots < t_n$, les instants des arrivées successifs des clients dans l'agence. On note $T_n = t_n - t_{n-1}$, le temps des inter-arrivées entre deux clients consécutifs. On suppose que ces temps des inter-arrivées $\{T_n\}_{n\geq 1}$ sont des variables aléatoires indépendantes et identiquement distribuées, suivant une loi exponentielle de paramètre $\lambda$, de densité de probabilité : $f(t) = \lambda e^{-\lambda t}, \quad t \geq 0$.

En effet, dans notre cas d'étude, les observations montrent qu'entre 2 à 4 clients arrivent chaque minute à l'agence. C'est-à-dire en moyenne 3 clients par minute, ce qui correspond à un temps moyen des inter-arrivées de $\frac{1}{\lambda}$ minute, soit environ 20 secondes entre deux clients.

Ainsi, le nombre de clients arrivés jusqu'au temps $t$ peut être alors modélisé par une variable aléatoire $X(t)$, définie par :
$X(t) = \sup\{n \in \mathbb{N} \mid t_n \leq t\}$.

Ainsi défini, le processus $\{X(t), t \geq 0\}$ peut être considéré comme un processus de Poisson homogène de paramètre $\lambda = 3$. Il possède donc la propriété d'absence de mémoire, ce qui en fait un processus de Markov d'après le chapitre 2.

En ce qui concerne les flux de sorties de cette agence, les observations montrent qu'entre 1 et 3 clients quittent l'agence toutes les 3 minutes. C'est-à-dire en moyenne 2 clients par intervalle de 3 minutes, ce qui correspond à un taux moyen de sortie de $\mu = \frac{2}{3}$ client par minute, soit un temps moyen entre deux départs de $\frac{1}{\mu} = 1{,}5$ minute (soit environ 90 secondes entre deux clients).

Soit $s_0 = 0 < s_1 < s_2 < \cdots < s_n$, les instants des départs successifs des clients de l'agence. On note $S_n = s_n - s_{n-1}$, le temps des inter-sorties entre deux clients consécutifs. On suppose que ces temps des inter-sorties $\{S_n\}_{n\geq 1}$ sont des variables aléatoires indépendantes et identiquement distribuées, suivant une loi exponentielle de paramètre $\mu$, de densité de probabilité $f(t) = \mu e^{-\mu t}$ pour $t \geq 0$.

Le nombre de clients ayant quitté l'agence jusqu'au temps $t$ peut

être alors modélisé par une variable aléatoire $Y(t)$, définie par : $Y(t) = \sup\{n \in \mathbb{N} \mid s_n \leq t\}$.
Ainsi défini, le processus $\{Y(t),\ t \geq 0\}$ peut être considéré comme un processus de Poisson homogène de paramètre $\mu = \frac{2}{3}$. Il possède donc la propriété d'absence de mémoire, ce qui en fait un processus de Markov, d'après le chapitre 2.

### *B : Temps de service*

Variables aléatoires i.i.d. exponentielles de paramètre $\mu$ :

- Ouverture de comptes :
  - Service exclusif (1 client à la fois)
  - Durée $\sim \text{Exp}(\mu_1)$
- Opérations courantes :
  - Opérations courantes : $\text{Exp}(\mu_2)$
  - CNAS : $\text{Exp}(\mu_3)$
  - AADL : $\text{Exp}(\mu_4)$
- Crédit :
  - Accès libre (sans rendez-vous)
  - Durée $\sim \text{Exp}(\mu_5)$

### *C : Configuration des serveurs*

- Ouverture de comptes :
  - Modèle $(M/M/1/10)$ (1 serveur, capacité maximale de 10 clients par jour).
  - Pas de file d'attente physique.
- Opérations courantes :
  - Modèle $(M/M/2/K)$ (2 serveurs).
  - Capacité $K$ variable :
    - Normalement illimitée
    - $K \approx 5$ pendant les pics AADL
- Crédit :
  - Modèle $(M/M/5)$ (5 serveurs)
  - File d'attente illimitée
  - Accès libre sans rendez-vous.

## 2.5 Modélisation par type de service

### 2.5.1 Modélisation du service d'Animation commerciale

***Analyse du modèle (M/M/1/10)***

Le système est représenté par une file d'attente (M/M/1/10) caractérisée par :

- Arrivées markoviennes (M) : Processus de Poisson homogène de taux $\lambda > 0$.
- Service markovien (M) : Temps de service exponentiels de taux $\mu > 0$.
- 1 serveur disponible ($s = 1$).
- Capacité maximale de 10 clients ($K = 10$).

La capacité maximale est imposée par les contraintes opérationnelles et vise à éviter la saturation du service. Le serveur traite tous les types de demandes selon leur ordre d'arrivée. Les horaires de fonctionnement sont de 9h à 12h, puis de 13h30 à 15h30, soit un total de 5 heures de service par jour. La durée de traitement est environ 45 minutes pour une ouverture de compte.
Nous calculons d'abord le taux moyen d'arrivée $\lambda$ à l'aide d'une moyenne pondérée, en supposant que le nombre total d'heures d'observation est de 5 heures. Les clients sont répartis comme suit :

- Heure 1 : 2 clients
- Heure 2 : 3 clients
- Heure 3 : 2 clients
- Heure 4 : 0 client
- Heure 5 : 1 client

La fréquence pour chaque heure est $f_i = \dfrac{1}{5} = 0{,}2$. On applique la formule de la moyenne pondérée :

$$\begin{aligned} \bar{x} &= \sum x_i \cdot f_i \\ &= 2 \cdot 0{,}2 + 3 \cdot 0{,}2 + 2 \cdot 0{,}2 + 0 \cdot 0{,}2 + 1 \cdot 0{,}2 \\ &= 0{,}4 + 0{,}6 + 0{,}4 + 0 + 0{,}2 = 1{,}6 \text{ clients/heure.} \end{aligned}$$

Ainsi, $\lambda \approx 2$ clients par heure.

Nous supposons maintenant que chaque client passe exactement 45 minutes dans le service, soit 0,75 heure. Le taux de service $\mu$ est alors donné par l'inverse du temps de service :

$$\mu = \frac{1}{\text{temps de service}} = \frac{1}{0{,}75} = 1{,}333 \text{ clients/heure.}$$

Ainsi, $\mu \approx 1$ client par heure.

***Étude du processus***

Soit $N_t$ : nombre de clients dans le système. Le processus $(N_t)_{t\geq 0}$ est un processus de naissance et de mort (PNM) de paramètres $(\lambda_n, \mu_n)$

$$\text{avec } \lambda_n = \begin{cases} \lambda, & \text{si } 0 \leq n \leq K-1 \\ 0, & \text{si } n \geq K \end{cases}, \quad \mu_n = \begin{cases} \mu, & \text{si } 1 \leq n \leq K \\ 0, & \text{si } n = 0 \end{cases}$$

$$\text{et } \mathrm{P}_n = \begin{cases} \dfrac{\rho^n(1-\rho)}{1-\rho^{K+1}}, & \text{si } 0 \leq n \leq K,\ \rho \neq 1 \\ \dfrac{1}{K+1}, & \text{si } 0 \leq n \leq K,\ \rho = 1. \end{cases}$$

$(P_n)_{n\geq 0}$ est la loi du système en régime permanent.

***Performances du système associées aux clients***

Ces perfermanes sont données par les formules suivantes établies au chapitre 2.

*1) Nombre moyen de clients dans le système :*

$$\eta = \begin{cases} \frac{K}{2}, & \rho = 1 \\ \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}, & \rho \neq 1. \end{cases}$$

*2) Temps d'attente d'un client dans le système :*

$$\Rightarrow W = \frac{\eta}{\mu(1-p_0)} = \begin{cases} \dfrac{K}{2\mu(1-p_0)}, & \rho = 1 \\ \left(\dfrac{\rho}{1-\rho} - \dfrac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}\right).\dfrac{1}{\mu(1-p_0)}, & \rho \neq 1. \end{cases}$$

### 2.5.2 Modélisastion du service financier

Nous considérons un service composé de deux guichets traitant deux types de clients : les clients habituels et les clients particuliers. Les clients habituels arrivent individuellement selon un processus de comtage qu'on peut considérer comme un processus de Poisson de paramètre $\lambda_1$ et sont servis selon une loi exponentielle de paramètre $\mu$, avec un nombre illimité de clients autorisés dans le système. Les clients particuliers, quant à eux, arrivent selon un processus de Poisson de paramètre $\lambda_2$ et sont également servis selon une loi exponentielle de paramètre $\mu$, mais leur nombre total dans le système (en file d'attente et en service) est limité à 5, correspondant donc à un modèle $(M/M/2/5)$. Les deux types de clients sont pris en charge indifféremment par les guichets, sans priorité, selon une sélection équitable ou aléatoire.
Pour simplifier la modélisation, nous considérons deux files d'attente classiques en parallèle, avec un partage de ressources (les 2 guichets communs) :

- File 1 : $(M/M/2)$ (clients habituels)
- File 2 : $(M/M/2/5)$ (clients particuliers)

Chaque file est modélisée séparément afin d'appliquer les formules analytiques classiques, tout en notant que le taux de service effectif dépend du partage des serveurs.

#### *Analyse du modèle ($M/M/2$)*

Le système est représenté par une file d'attente $(M/M/2)$ caractérisée par :

- Arrivées markoviennes (M) : Processus de Poisson homogène de taux $\lambda > 0$.
- Service markovien (M) : Temps de service exponentiels de taux $\mu > 0$.
- 2 serveurs disponibles ($s = 2$).

Le modèle $(M/M/2)$ est adapté à la modélisation du service des opérations courantes, où deux serveurs traitent simultanément des opérations variées :

- Retrait en DA : 5 min (30% des opérations),

– Retrait en devise : 10 min (20 % ),
– Demande de solde/mises à jour : 3 min (15%),
– Versement en devise : 30 min (5%),
– Versement en dinars : 10 min (25%),
– Virement : 10 min (4%),
– Renseignements : 5 min (1%).

Ainsi, ceci correspond à une La durée moyenne pondérée de service est calculée comme suit :

$$\begin{aligned}\bar{x} &= \sum x_i.\ f_i \\ &= 5 \times 0.30 + 10 \times 0.20 + 3 \times 0.15 \\ &\quad + 30 \times 0.05 + 10 \times 0.25 \\ &\quad + 10 \times 0.04 + 5 \times 0.01 \\ &= 1.5 + 2.0 + 0.45 + 1.5 + 2.5 + 0.4 + 0.05 \\ &= 8.4 \text{ minutes/opération.}\end{aligned}$$

C'est-à-dire à une durée de service moyenne $d \approx 8.4$ minutes $\approx \frac{1}{7}$ heure par opération, ceci correspond à un taux $\mu = \frac{1}{d} = 7$ clients servis par unité de temps (heure).
Pour les arrivées, en tenant compte de la durée journalière de 6,5 heures par jour et une affluence de 100 à 200 clients par jour, le taux moyen d'arrivée est calculé en prenons la moyenne du nombre des clients servis par jour $\frac{100+200}{2} = 150$, divisée par la duréé journalière de travail, donc $\lambda = \frac{150}{6{,}5} = 23$ clients par heure.

***Étude du processus***

Soit $N_t$ : le nombre de clients dans le système (file+guichets)
Ona $(N_t)_{t\geq 0}$ est un processus de naissance et de mort $(\lambda_n, \mu_n)$ où :

$\lambda_n = \lambda, \forall n \geq 0,$

et $\quad \mu_n = \begin{cases} 0, & \text{si } n = 0 \\ n \cdot \mu, & \text{si } 1 \leq n \leq s. \end{cases}$

La loi stationnaire est donnée par :

$$P_n = \begin{cases} \frac{\rho^n}{n!}.P_0, & \text{si } 0 \leq n \leq s \\ \frac{\rho^n}{s!.s^{n-s}}, & \text{si } n \geq s. \end{cases}$$

***Performances du système associées aux clients***

Ces perfermances sont données par les formules suivantes établies dans le chapitre 2.

*1) Nombre moyen de clients dans la file :*

$$\eta_q = \frac{\rho^{s+1}}{(s-\rho)^2.(s-1)!}.P_0.$$

*2) Temps moyen d'attente d'un client dans la file :*

$$\overline{W_q} = \frac{1}{\mu s - \lambda}.$$

*3) Temps moyen d'attente d'un client dans le système*

$$\overline{W} = \frac{\overline{\eta_q}}{\lambda} + \frac{1}{\mu} = \overline{W_q} + \frac{1}{\mu}.$$

***Analyse du modèle (*$M/M/2/5$*)***

C'est un système d'attente ouvert, où le nombre de clients ayant accès au service est limité à $K$ clients particuliers ($K = 5$), qui représente sa capacité et il comporte 2 serveurs ($s = 2$). Pour les arrivées, comme précédemment, il est logique de supposer qu'elles sont poissoniennes de paramètre $\lambda = 5$ (valeur $\lambda = 5$ clients/heure représente le taux d'arrivée effectif observé). Les durées de service, supposées indépendantes et exponentiellement distribuées. En tenant compte de la durée journallière de 6,5 heures et une durée moyenne de service de 15 minutes par client, ceci correspond à un taux de service $\mu = \frac{1}{0.25} = 4$ clients par heure par serveur.

***Étude du processus***

$(N_t)_{t\geq 0}$ est un processus de naissance et de mort (PNM) défini sur l'ensemble fini {0, 1, 2,..., K} de taux $\lambda_n$ et $\mu_n$ tel que ;

$$\lambda_n = \begin{cases} \lambda, & \text{si} \quad 0 \leq n \leq K-1, \\ 0, & \text{sinon.} \end{cases} \quad et \quad \mu_n = \begin{cases} n\mu, & \text{si} \quad 1 \leq n \leq s, \\ s\mu, & \text{si} \quad s \leq n \leq K, \\ 0, & \text{sinon.} \end{cases}$$

La loi stationnaire est donnée par :

$$P_n = \begin{cases} \frac{\rho^n}{n!} P_0 & 1 \leq n \leq s, \\ \frac{\rho^n}{s! s^{n-s}} P_0 & s < n \leq K, \\ \frac{1}{\sum_{n=0}^{s} \frac{\rho^n}{n!} + \sum_{n=s+1}^{K} \frac{\rho^n}{s! s^{n-s}}} & n = 0. \end{cases}$$

***Performances du système associées aux clients***

Ces perfermances sont données par les formules suivantes établies au chapitre 2.

*1) Nombre moyen de clients dans le système*

$$\eta = P_0 . \frac{s^{s-1}}{(s-1)!} \sum_{n=s+1}^{K} n(\bar{\rho})^n.$$

avec $\bar{\rho} = \frac{\rho}{s}$.

*2) Nombre moyen de clients dans la file :*

$$\eta_q = P_0 . \frac{\rho^s}{s!} \sum_{h=1}^{K-s} h(\bar{\rho})^h, \text{ avec } \bar{\rho} = \frac{\rho}{s}.$$

*3) Temps d'attente moyen de clients dans le système :*

$$\overline{W} = \frac{\eta}{\lambda \cdot (1 - P_K)} = \frac{\eta}{\lambda \left(1 - \frac{\rho^K}{s! . s^{K-s}} \cdot P_0\right)}.$$

*4) Temps moyen d'attente de clients dans la file :*

$$\overline{W_q} = \frac{\eta_q}{\Lambda} = \frac{\eta_q}{\lambda(1 - P_K)} = \frac{\eta_q}{\lambda(1 - \frac{\rho^K}{s! s^{K-s} P_0})}.$$

### 2.5.3 Modélisation du service crédit

Nous considérons le service des crédits bancaires, qui assure le traitement des demandes de prêts par cinq conseillers financiers travaillant en parallèle. Les clients se présentent individuellement selon un processus de comptage qu'on peut considérer comme un processus de Poisson de paramètre $\lambda$ et sont servis selon une loi exponentielle de paramètre $\mu$.

En effet, le modèle $(M/M/5)$ s'applique ici, avec une file d'attente en FIFO et des temps de service variables : 20 minutes pour un crédit ordinaire et 30 à 40 minutes pour un crédit d'investissement. Avec un taux d'arrivée moyen de 5 clients par heure ($\lambda = 5$) et un taux de service estimé à $\mu = 2$ clients par heure par conseiller (basé sur un temps moyen de 30 minutes par demande), le système dispose d'une capacité globale suffisante pour absorber la charge tout en limitant les temps d'attente.

***Analyse du modèle*** $(M/M/5)$

C'est un système d'attente ouvert, où le nombre de clients autorisés dans le système est illimité et qui comporte $s = 5$ serveurs. Pour les arrivées, comme précédemment, il est logique de supposer qu'elles suivent un processus de Poisson de paramètre $\lambda$ et que les durées de service sont indépendantes, suivant une même loi exponentielle de paramètre $\mu$.

**Remarque 2.2**

L'analyse du processus et des performances du système dans le modèle $(M/M/5)$ est identique à celle obtenue dans l'étude du service financier, dans le cas des clients habituels, c'est-à-dire le modèle $(M/M/2)$.

D'autre part, on a observé que le service crédit fonctionne efficacement avec ses 5 serveurs actuels ($s = 5$). En effet, les files d'attente sont toujours courtes par rapport au reste des services. De plus, il n'y a jamais de saturation et les clients sont servis rapidement. Par conséquent, il apparaît que le système actuel est déjà optimal pour ce service. Ainsi, une analyse d'optimisation supplémentaire ne semble pas nécessaire dans ce cas de service.

## 2.6 Conclusion

Cette partie a été consacrée à la modélisation des systèmes bancaires, notamment à travers les modèles de files d'attente $(M/M/1/K)$, $(M/M/s)$ et $(M/M/s/K)$. L'étude de ces modèles à l'état stationnaire nous a permis d'identifier et d'analyser plusieurs indicateurs de performance, tels que le risque de saturation, la longueure moyenne de la file d'attente et le temps de séjour d'un client dans le système.

# Partie II : Optimisation des services du système

## 2.1 Introduction

L'optimisation est une branche des mathématiques appliquées qui s'intéresse à l'analyse et à la modélisation des contraintes d'un problème donné, dans le but de déterminer la solution optimale ou le meilleur compromis au sein d'un ensemble de solutions possibles. Ainsi, dans cette partie, on s'attellera à optimiser quelques mesures de performance pour chacun des services modélisés précédemment (hormis le service crédit, comme on l'a relaté dans la remarque 2.2). Pour ces mesures de performance, on optera essentiellement pour :

- le risque de saturation des services ;
- la durée de séjour d'un client dans le service ;
- la longueur de la file d'attente dans le système.

Comme on ne peut pas agir sur le comportement des clients et leurs habitudes, le taux d'arrivée $\lambda$ est considéré comme un paramètre fixe. Comme on l'a expliqué dans la modélisation du système, la capacité maximale K du système est fixée pour chacun des services.

Les optimisations seront réalisées par rapport aux variables :

- $\mu$ : durée moyenne de service $(\frac{1}{\mu})$ ;
- $s$ : nombre de serveurs.

Avec les contraintes communes : $\lambda < s\mu \quad \text{et} \quad s \in \mathbb{N}^*$.

## 2.2 Optimisation des services bancaires

On opte pour la façon d'optimiser suivante :
Au lieu de faire le calcul explicite et donc faire de la différentiation et la dérivation des fonctions objectives par rapport aux différents paramètres entrants, on raisonne avec la théorie des voisinages sur ces différents paramètres.

### 2.2.1 Service d'ouverture de compte

On rappelle que ce service a été modélisé précédemment par le modèle $(M/M/1/10)$, tel que :

- L'unité de temps est l'heure ;
- $\lambda = 2$ clients par heure ;
- $\mu = 1$ clients par heure.

Ce modèle $(M/M/1/10)$, bien qu'utile pour décrire l'état actuel, ne permet pas d'étudier l'impact de l'ajout de serveurs ou de l'amélioration des processus, car il s'agit d'une observation et pas d'une règle explicite du système, donc le passage à $(M/M/s)$ offre cette flexibilité dans le but d'explorer la vrai capacité du système sans la capacité théorique de 10. On s'intéressera donc à ces 3 mesures de performance qu'on va optimiser, en prenant le modèle (M/M/s).

#### a. Le risque de saturation

Pour le Modèle $(M/M/s)$, le risque de saturation est défini par :

$$\pi(\rho, s) = \frac{\rho^s}{s!} \cdot \frac{s}{s-\rho} \cdot P_0.$$

Avec

$$\rho = \frac{\lambda}{s\mu}, \quad P_0 = \frac{1}{\sum_{n=0}^{s} \frac{\rho^n}{n!} + \frac{\rho^{s+1}}{s!(s-\rho)}}.$$

Afin d'optimiser $\pi$, il convient d'agir sur les paramètres clés qui sont le taux de service $\mu$ et le nombre de serveurs $(s)$. Le problème s'énonce formellement :

$$\begin{cases} \min\limits_{s,\mu} & \pi(\rho, s), \\ & \lambda < s\mu, \\ & s \in \mathbb{N}^*. \end{cases}$$

D'après les contraintes de ce problème d'optimisation, $\rho$ doit être inférieur à 1.

– Si $\rho \in v(1)$
Quand le système approche de sa limite de capacité, le terme $\frac{s}{s-\rho}$ diverge car le dénominateur tend vers zéro. Bien que la probabilité $P_0$ d'un système vide diminue, compensant partiellement cette croissance, le risque de saturation $\pi(\rho, s)$ suit approximativement une loi en $\frac{C}{1-\rho}$, où $C$ est une constante positive. Cette relation montre que le risque augmente de manière critique à mesure que $\rho$ se rapproche de 1.

– Si $\rho << 1$
Lorsque le taux d'occupation est très faible, les termes $\rho^s$ et $\rho$ deviennent négligeables. Le risque de saturation se réduit alors à une expression simplifiée $\frac{\rho^s}{s!}$, extrêmement petite en pratique.

**b. La durée de séjour d'un client dans le système**

Le temps d'attente d'un client dans le service est donné par :

$$W(\rho, s) = \frac{\overline{\eta_q}}{\lambda} + \frac{1}{\mu} = \overline{W_q} + \frac{1}{\mu}.$$

On pose le problème d'optimisation comme suit :

$$\begin{cases} \min\limits_{s,\mu} & W(\rho, s), \\ & \lambda < s\mu, \\ & s \in \mathbb{N}^*. \end{cases}$$

l'expression de $W$ montre que la durée de séjour d'un client dans le système dépend à la fois de la file d'attente ($\overline{\eta_q}$) et de la vitesse du service ($\mu$).

– Si $\rho \in v(1)$
La file d'attente s'allonge considérablement, ce qui augmente $\overline{\eta_q}$ et donc $\overline{W_q}$. Le temps d'attente devient critique.

– Si $\rho \ll 1$
La file d'attente reste courte et stable, maintenant $\overline{W_q}$ à des valeurs faibles. Le temps de séjour est principalement déterminé par $1/\mu$.

On cherche donc à déterminer une configuration $(s, \mu)$ qui minimise le temps total $W$, tout en respectant la contrainte de stabilité $\rho = \frac{\lambda}{s\mu} < 1$.

**c. La longueur de la file d'attente**

Le nombre moyen de clients dans la file est donné par :

$$\eta_q(\rho, s) = \frac{\rho^{s+1}}{(s-\rho)^2.(s-1)!}.P_0$$

On pose le problème d'optimisation suivant :

$$\begin{cases} \min\limits_{s,\mu} & \eta_q(\rho, s), \\ & \lambda < s\mu, \\ & s \in \mathbb{N}^*. \end{cases}$$

- Si $\rho \in v(1)$
  Le système fonctionne presque à pleine capacité et la file s'allonge rapidement. Le dénominateur $(s-\rho)^2$ tend vers 0, faisant diverger $\eta_q$. Les arrivées dépassent quasiment les départs.
- Si $\rho \ll 1$
  Les clients sont traités plus rapidement que leur rythme d'arrivée, ce qui maintient une file courte. Le terme $\rho^{s+1}$ au numérateur devient très petit, rendant $\eta_q$ négligeable.

Pour optimiser cette performance, on cherche à réduire $\eta_q$ en ajustant les paramètres $s$ et $\mu$, tout en gardant $\rho < 1$. Cette mesure reflète l'encombrement de la file : plus elle est élevée, plus le système est saturé. Elle dépend directement du taux d'occupation $\rho = \frac{\lambda}{s\mu}$.

### 2.2.2 Service des opérations courantes

Ce service étant modélisé précédemment par le modèle $(M/M/2)$ pour le cas des clients habituels et $(M/M/2/5)$ pour le cas des clients AADL, avec :

- L'unité de temps est l'heure ;
- $\lambda = 23$ clients par heure (habituels) ;

– $\lambda = 5$ clients par heure (particuliers) ;

– $\mu \approx 7$ clients par heure par serveur (habituels) ;

– $\mu \approx 7$ clients par heure par serveur (particuliers) ;

– $s = 2$.

On s'interessera donc à l'optimisation de ces trois mesures de performances pour chaque cas.

**Cas 1 : Modèle (M/M/2)**

**a. Le risque de saturation**

Pour le Modèle $(M/M/2)$, le risque de saturation est défini par :

$$\pi(\rho, s) = \frac{\rho^s}{s!} \cdot \frac{s}{s-\rho} \cdot P_0.$$

Avec

$$\rho = \frac{\lambda}{s\mu}, \quad P_0 = \frac{1}{\sum_{n=0}^{s} \frac{\rho^n}{n!} + \frac{\rho^{s+1}}{s!(s-\rho)}}.$$

Afin d'optimiser $\pi$, il convient d'agir sur les paramètres clés qui sont le taux de service $\mu$ et le nombre de serveurs $(s)$. Le problème s'énonce formellement :

$$\begin{cases} \min\limits_{s,\mu} & \pi(\rho, s), \\ & \lambda < s\mu, \\ & s \in \mathbb{N}^*. \end{cases}$$

D'après les contraintes de ce problème d'optimisation, $\rho$ doit être inférieur à 1.

– Si $\rho \in v(1)$
Quand le système approche de sa limite de capacité, le terme $\frac{s}{s-\rho}$ diverge car le dénominateur tend vers zéro. Bien que la probabilité $P_0$ d'un système vide diminue, compensant partiellement cette croissance, le risque de saturation $\pi(\rho, s)$ suit approximativement une loi en $\frac{C}{1-\rho}$, où $C$ est une constante positive. Cette

relation montre que le risque augmente de manière critique à mesure que $\rho$ se rapproche de 1.

– Si $\rho << 1$
Lorsque le taux d'occupation est très faible, les termes $\rho^s$ et $\rho$ deviennent négligeables. Le risque de saturation se réduit alors à une expression simplifiée $\frac{\rho^s}{s!}$, extrêmement petite en pratique.

**Remarque 2.3**

La situation réelle du système (service des opérations courantes) donne un $\rho$ très grand que 1, vu les données recueillies et citées précédemment. Dans ce cas, le système est instable et sans l'abondon de certains clients, on aurait eu une congestion très importante et une saturation certaine. D'où la nécessité d'ajouter au moins un autre serveur pour décongestionner le système et ainsi réduire la durée de séjour d'un client et la longueur de la file d'attente.

### b. La durée de séjour d'un client dans le système

Le temps d'attente d'un client dans le service est donné par :

$$W(\rho, s) = \frac{\overline{\eta_q}}{\lambda} + \frac{1}{\mu} = \overline{W_q} + \frac{1}{\mu}.$$

On pose le problème d'optimisation comme suit :

$$\begin{cases} \min\limits_{s,\mu} & W(\rho, s), \\ & \lambda < s\mu, \\ & s \in \mathbb{N}^*. \end{cases}$$

l'expression de $W$ montre que la durée de séjour d'un client dans le système dépend à la fois de la file d'attente ($\overline{\eta_q}$) et de la vitesse du service ($\mu$).

– Si $\rho \in v(1)$
La file d'attente s'allonge considérablement, ce qui augmente $\overline{\eta_q}$ et donc $\overline{W_q}$. Le temps d'attente devient critique.

– Si $\rho \ll 1$
La file d'attente reste courte et stable, maintenant $\overline{W_q}$ à des valeurs faibles. Le temps de séjour est principalement déterminé par $1/\mu$.

On cherche donc à déterminer une configuration $(s, \mu)$ qui minimise le temps total $W$, tout en respectant la contrainte de stabilité $\rho = \frac{\lambda}{s\mu} < 1$.

### c. La longueur de la file d'attente

Le nombre moyen de clients dans la file est donné par :

$$\eta_q(\rho, s) = \frac{\rho^{s+1}}{(s-\rho)^2.(s-1)!}.P_0$$

On pose le problème d'optimisation suivant :

$$\begin{cases} \min\limits_{s,\mu} & \eta_q(\rho, s), \\ & \lambda < s\mu, \\ & s \in \mathbb{N}^*. \end{cases}$$

– Si $\rho \in v(1)$
Le système fonctionne presque à pleine capacité et la file s'allonge rapidement. Le dénominateur $(s-\rho)^2$ tend vers 0, faisant diverger $\eta_q$. Les arrivées dépassent quasiment les départs.

– Si $\rho \ll 1$
Les clients sont traités plus rapidement que leur rythme d'arrivée, ce qui maintient une file courte. Le terme $\rho^{s+1}$ au numérateur devient très petit, rendant $\eta_q$ négligeable.

Pour optimiser cette performance, on cherche à réduire $\eta_q$ en ajustant les paramètres $s$ et $\mu$, tout en gardant $\rho < 1$. Cette mesure reflète l'encombrement de la file : plus elle est élevée, plus le système est saturé. Elle dépend directement du taux d'occupation $\rho = \frac{\lambda}{s\mu}$.

## Cas 2 : Modèle (M/M/2/5)

### a. Le risque de saturation

Pour le Modèle $(M/M/2/5)$, le risque de saturation est défini par :

$$\pi(\rho, s, K) = \frac{\rho^s}{s!} P_0 \sum_{h=0}^{K-s} (\bar{\rho})^h, \text{ avec } \quad \overline{\rho} = \frac{\rho}{s}$$

et

$$P_0 = \frac{1}{1 + \sum_{n=1}^{s} \frac{\rho^n}{n!} + \sum_{n=s+1}^{K} \frac{\rho^n}{s! s^{n-s}}}$$

Cette fonction objective sera optimée par rapport à $s$ et $\mu$, donc

$$\begin{cases} \min\limits_{s,\mu}, & \pi(\rho, s, K), \\ & \lambda \neq s\mu, \\ & s \in \mathbb{N}^*. \end{cases}$$

On a $\overline{\rho} = \frac{\lambda}{s\mu}$ qui est le taux d'occupation global, et $P_0$ est la probabilité que le système soit vide. L'objectif est donc de contrôler ce risque de saturation en ajustant les paramètres $s$ et $\mu$ pour maintenir $\overline{\rho} < 1$ et garantir un fonctionnement stable.

- Lorsque $\overline{\rho} \in v(1)$
  La file a tendance à se remplir complètement, ce qui augmente fortement $\pi$. Le risque se rapproche de 1, indiquant un fort taux de rejet potentiel.
- Lorsque $\overline{\rho} \ll 1$
  Le risque $\pi$ reste proche de 0, indiquant une faible probabilité de blocage. Les termes $\rho^s$ et $(\overline{\rho})^h$ deviennent négligeables.
- Lorsque $\overline{\rho} > 1$
  Le risque de saturation devient certain ($\pi \approx 1$) car les arrivées dépassent durablement la capacité de traitement. Le système ne peut plus maintenir un état stable.

L'objectif est donc de contrôler ce risque de saturation en ajustant les paramètres $s$ et $\mu$ pour maintenir $\rho < 1$ et garantir un fonctionnement stable.

### b. La durée de séjour d'un client dans le système

La durée de séjour d'un client est donnée par :

$$W(\rho, s, K) = \frac{\eta}{\lambda \cdot (1 - P_K)} = \frac{\eta}{\lambda \left(1 - \frac{\rho^K}{s! \cdot s^{K-s}} \cdot P_0\right)}.$$

On pose le problème d'optimisation comme suit :

$$\begin{cases} \min\limits_{s,\mu} & W(\rho, s, K), \\ & \lambda \neq s\mu, \\ & s \in \mathbb{N}^*. \end{cases}$$

- Lorsque $\overline{\rho} \ll 1$ La probabilité $P_K$ est très faible, ce qui maintient le dénominateur proche de 1. La durée de séjour $W$ reste modérée, principalement déterminée par $1/\mu$.
- Lorsque $\overline{\rho} \in v(1)$
  $P_K$ augmente significativement, réduisant le dénominateur $(1 - P_K)$. Cela provoque une divergence de $W$, traduisant un engorgement du système.
- Lorsque $\overline{\rho} > 1$
  La durée de séjour $W$ devient infinie en régime stationnaire, car le système ne peut plus traiter toutes les arrivées. La file d'attente croît indéfiniment.

L'enjeu est donc d'équilibrer les paramètres $\mu$ et $s$ du système pour maintenir une durée de séjour acceptable, tout en respectant la contrainte de stabilité $\overline{\rho} < 1$.

### c. La longueur de la file d'attente

La formule est donnée par :

$$\eta_q(\rho, s, K) = P_0 . \frac{\rho^s}{s!} \sum_{h=1}^{K-s} h(\bar{\rho})^h.$$

Afin d'optimiser $\eta_q$ du système (clients AADL), il convient d'agir sur les paramètres clés qui sont le taux de service $\mu$ et le nombre de serveurs $(s)$. Le problème s'énonce formellement :

$$\begin{cases} \min\limits_{s,\mu} & \eta_q(\rho, s, K), \\ & \lambda \neq s\mu, \\ & s \in \mathbb{N}^*. \end{cases}$$

- $\overline{\rho} \ll 1$
  Le terme $(\overline{\rho})^h$ devient négligeable rapidement. La file d'attente $\eta_q$ reste très courte car les clients sont servis presque immédiatement.
- $\overline{\rho} \in v(1)$
  Les puissances de $\overline{\rho}$ deviennent significatives. La longueur de file $\eta_q$ augmente rapidement, traduisant un engorgement progressif du système
- $\overline{\rho} > 1$
  La somme diverge théoriquement. En pratique, la file d'attente croît indéfiniment jusqu'à saturation complète du système ($\eta_q \to \infty$).
  En effet, lorsque $\overline{\rho}$ augmente, chaque terme de la somme croît, ce qui allonge la file moyenne.
  En particulier, lorsque $\overline{\rho} \to 1$, les puissances $(\overline{\rho})^h$ deviennent plus importantes, et $\eta_q$ augmente rapidement.

Cela signifie que plus le système approche la saturation, plus le nombre moyen de clients en attente augmente.

Inversement, si $\overline{\rho}$ est faible, les puissances $(\overline{\rho})^h$ deviennent négligeables, et la file reste courte.

Il est donc crucial de maintenir $\overline{\rho} < 1$ pour contrôler la taille de la file d'attente, et éviter l'engorgement du système.

Pour cela, on agit sur $s$ (en ajoutant des serveurs) ou sur $\mu$ (en augmentant la capacité de traitement), de façon à diluer la charge globale et raccourcir la file.

## 2.3 Conclusion

Dans cette partie, les résultats obtenus ont été exploités dans une démarche d'optimisation visant à améliorer l'efficacité des services, en ajustant certains paramètres clés comme le taux de service $\mu$ ou le nombre de serveurs s. Les systèmes modélisés et les formulations d'optimisation proposées serviront de support à des simulations numériques, qui seront réalisées à l'aide du logiciel MATLAB dans le chapitre suivant.

# Chapitre 3

# Simulations numériques et autres approches

## Partie I : Simulations numériques

### 3.1 Introduction

La simulation est une méthode qui consiste à reproduire le comportement d'un système réel ou théorique à l'aide d'un modèle mathématique ou informatique. Dans ce travail, les simulations ont été réalisées à l'aide du logiciel **MATLAB**, un environnement de calcul numérique et de programmation largement utilisé dans les domaines scientifiques et techniques. MATLAB offre de puissantes fonctionnalités de calcul matriciel, de modélisation et de visualisation, ce qui en fait un outil particulièrement adapté pour la simulation et l'analyse des systèmes de files d'attente.

Ces simulations permettront d'une part de visualiser le comportement dynamique des services (financier et service d'ouverture de compte), et d'autre part d'analyser systématiquement l'impact des paramètres opérationnels critiques : le taux de service $\mu$ (nombre moyen de clients traités par heure et par serveur) et le nombre de serveurs $s$ (capacité parallèle de traitement). Cette analyse quantitative éclairera les compromis entre ressources allouées et performances (temps d'attente, taux de rejet), identifiant ainsi les facteurs d'amélioration prioritaires pour chaque service.

### 3.2 Simulations par mesures de performances

Dans cette section, nous allons réaliser des simulations basées sur des mesures de performances afin d'analyser et d'évaluer le comportement des services étudiés en fonction de $\mu$ et $s$ .

### 3.2.1 Service d'ouverture de compte

Nous allons simuler différentes mesures de performance du modèle $(M/M/s)$ associé à ce service, afin d'évaluer précisément son comportement.

***Simulation du risque de saturation***

Nous fixons $\lambda = 2$ (voir le chapitre 2, section 2.5.1).

**i) Risque de saturation en fonction de $\mu$ (quand s=1)**

Nous obtenons les résultats suivants :

| $\mu$ | Risque de saturation $\pi$ |
|---|---|
| 3 | 0.6667 |
| 4 | 0.5000 |
| 5 | 0.4000 |
| 6 | 0.3333 |
| 7 | 0.2857 |
| 8 | 0.2500 |
| 9 | 0.2222 |

Table 3.1 : Risque de saturation

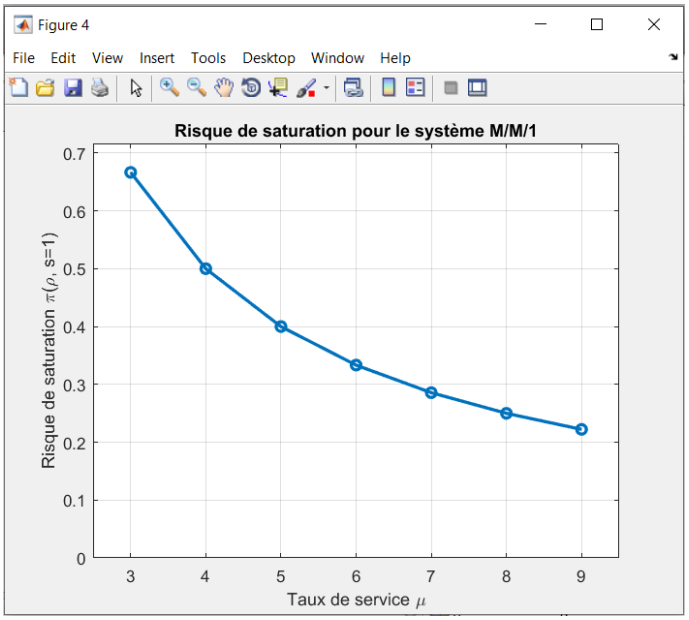


Figure 3.1 : Risque de saturation

### Interprétation des résultats obtenus

L'analyse des résultats met en évidence deux comportements distincts :

- De 3 à 9 : on observe une décroissance significative du risque de saturation $\pi$.
- Lorsque $\mu = 3$, le risque de saturation est très élevé ($\pi = 0{,}6667$), ce qui pourrait provoquer des files d'attente fréquentes.
- À l'inverse, avec $\mu = 9$, le système est beaucoup plus fluide avec un risque de saturation réduit à seulement 22,22 %..

Nous proposons donc de maintenir le taux de service $\mu$ à un niveau élevé (idéalement $\mu \geq 9$) pour garantir la fluidité du système et éviter les files d'attente fréquentes associées aux faibles valeurs de $\mu$.

### ii) Risque de saturation en fonction de $\mu$ (quand s=2)

L'évolution du risque de saturation est donnée comme suit :

| $\mu$ | Risque de saturation $\pi$ |
|---|---|
| 3 | 0.0476 |
| 4 | 0.0278 |
| 5 | 0.0182 |
| 6 | 0.0128 |
| 7 | 0.0095 |
| 8 | 0.0074 |
| 9 | 0.0058 |

Table 3.2 : Risque de saturation

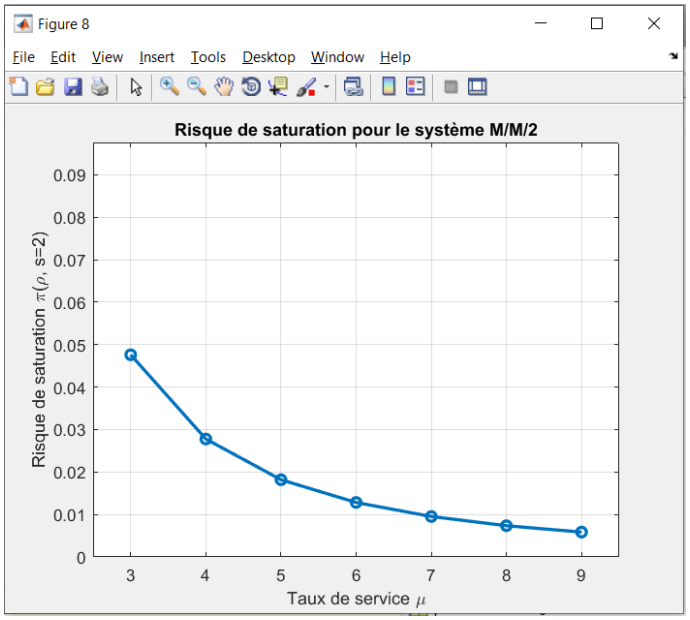


Figure 3.2 : Risque de saturation

**Interprétation des résultats obtenus :**

- La probabilité de saturation devient quasi nulle dès $\mu = 3$ avec ($\pi$= 0.0476).

– Cette observation permet de conclure que l'augmentation du nombre de serveurs à 2 suffit à éliminer le risque de saturation, même pour des valeurs relativement petites de $\mu$.

***Simulation de la duré de séjour dans le système***

Dans cette section, nous allons décrire la variation du temps moyen d'attente d'un client dans le service en fonction de $\mu$ et $s$.
Pour cela, on fixe $\lambda = 2$.

### i) Durée de séjour en fonction de $\mu$ (quand s=1)

Nous obtenons les résultats suivants :

| Taux de service $\mu$ | Durée de séjour W |
|---|---|
| 3 | 1.3333 |
| 4 | 0.7500 |
| 5 | 0.5333 |
| 6 | 0.4167 |
| 7 | 0.3429 |
| 8 | 0.2917 |
| 9 | 0.2540 |

Table 3.3 : Durée de séjour dans le système

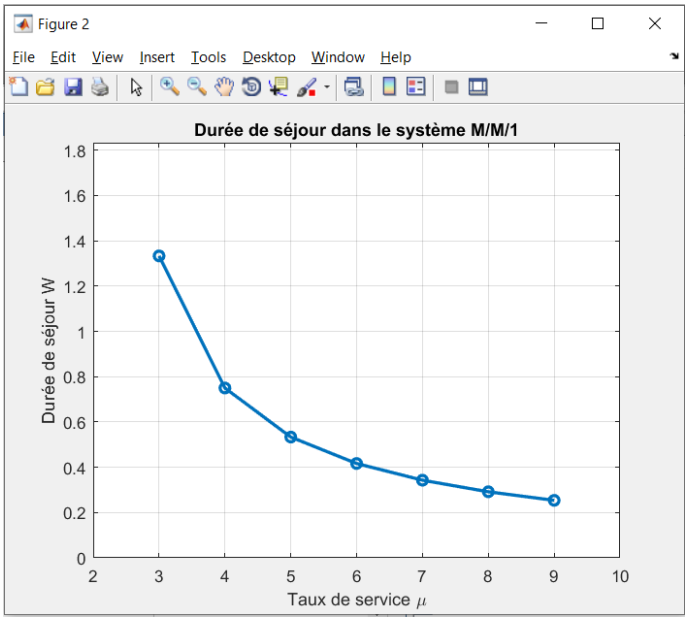


Figure 3.3 : Durée de séjour dans le système

**Interprétation des résultats obtenus :**

- Lorsque $\mu = 3$, la durée moyenne de séjour est longue ($W = 1{,}3333$), ce qui indique un service lent et un système congestionné.
- A partir de $\mu = 4$ à $\mu = 9$, la durée de séjour diminue à traduisant une amélioration notable de la performance du système.
- L'augmentation de $\mu$ entraîne une diminution progressive du temps d'attente dans le système.

### ii) Durée de séjour en fonction de $\mu$ (quand s=2)

Nous obtenons les résultats suivants :

| Taux de service $\mu$ | Durée de séjour W |
|---|---|
| 3 | 0.5833 |
| 4 | 0.4167 |
| 5 | 0.3250 |
| 6 | 0.2667 |
| 7 | 0.2262 |
| 8 | 0.1964 |
| 9 | 0.1736 |

Table 3.4 : Durée de séjour dans le système

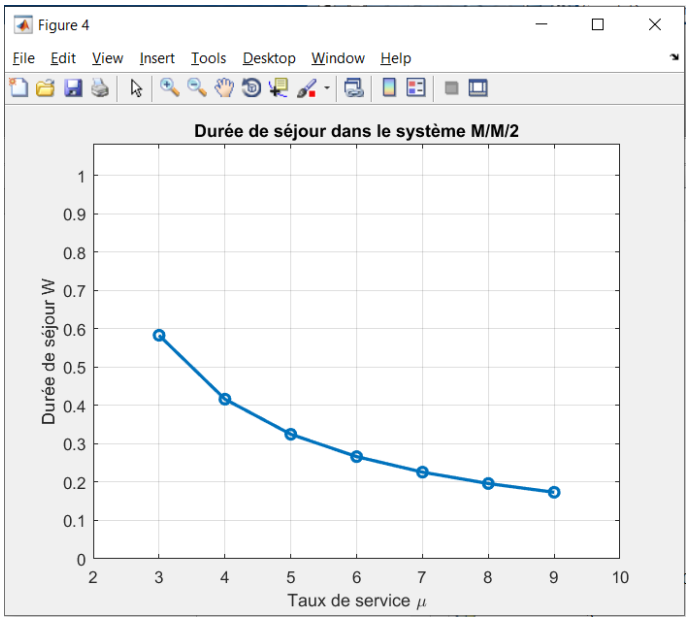


Figure 3.4 : Durée de séjour dans le système

**Interprétation des résultats obtenus :**

- Les résultats affichés pour la durée de séjour dans le cas s=2 sont un peu plus petits par rapport à ceux trouvés dans le cas où s=1.
- Augmenter le taux de service $\mu$ suffit donc pour améliorer le temps d'attente dans le système.

***Simulation de la longueur de la file***

Dans cette partie, nous allons expliquer comment le nombre moyen de client dans la file varie en fonction de $\mu$ et $s$.
On fixe $\lambda$= 2.

### i) Simulation en fonction de $\mu$ (quand s=1)

Nous obtenons les résultats suivants :

| Taux de service $\mu$ | Longueur moyenne $\eta_q$ |
|---|---|
| 3 | 1.3333 |
| 4 | 0.5000 |
| 5 | 0.2667 |
| 6 | 0.1667 |
| 7 | 0.1143 |
| 8 | 0.0833 |
| 9 | 0.0635 |

Table 3.5 : Longueur de la file d'attente

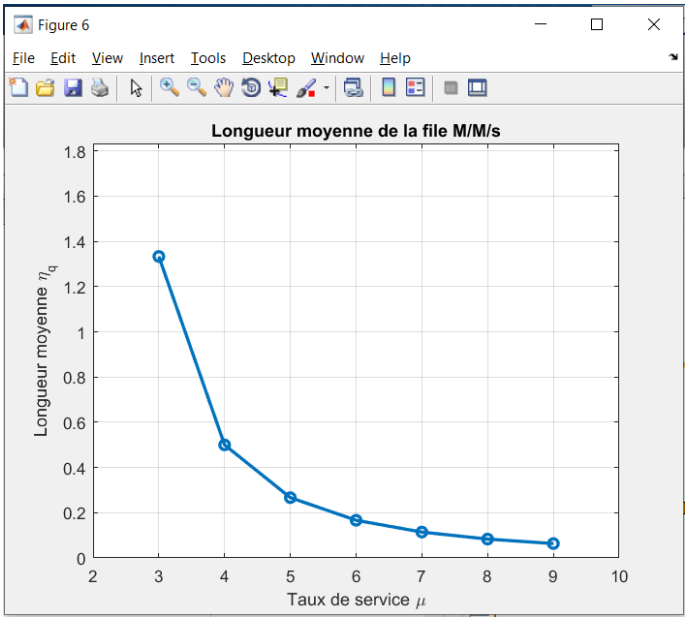


Figure 3.5 : Longueur de la file d'attente

**Interprétation des résultats obtenus :**

L'analyse des résultats nous permet de déduire les comportements suivants :

- L'augmentation de $\mu$ entraîne une diminution progressive de la longueur de la file d'attente dans le système.
- La décroissance la plus importante est observée de 3 à 4.
- De 4 à 9 la fonction diminue légérement, delà on peut conclure que le nombre de clients en attente devient négligeable.

### i) Simulation en fonction de $\mu$ (quand s=2)

Nous obtenons les résultats suivants :

| Taux de service $\mu$ | Longueur moyenne $\eta_q$ |
|---|---|
| 3 | 0.0095 |
| 4 | 0.0040 |
| 5 | 0.0020 |
| 6 | 0.0012 |
| 7 | 0.0007 |
| 8 | 0.0005 |
| 9 | 0.0003 |

Table 3.6 : Longueur de la file d'attente

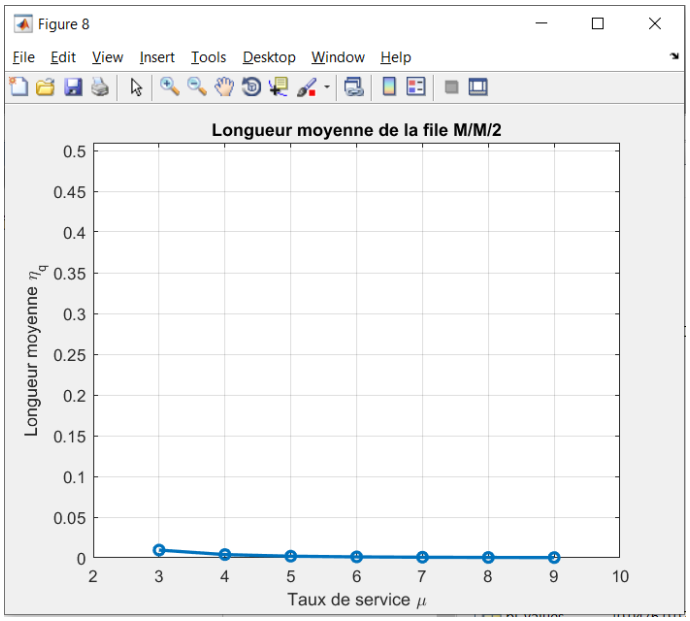


Figure 3.6 : Longueur de la file d'attente

**Interprétation des résultats obtenus :**

- La longueur de la file dans le système diminue d'une manière significative en fonction de $\mu$.
- Le fait que $\eta_q$ atteigne des valeurs aussi faibles dès $\mu = 3$ indique

que le système M/M/2 est performant même à des taux de service modérés, d'où le service est pratiquement sans attente quans s=2.

### 3.2.2 Service des opérations courantes (M/M/2)

Nous allons simuler différentes mesures de performance du modèle $(M/M/2)$ associé à ce service, afin d'évaluer précisément son comportement.

***Simulation du risque de saturation***

Dans cette partie, nous allons expliquer comment le risque de saturation varie en fonction de $\mu$ et $s$, en fixant $\lambda < s\mu$.

**Données**

On fixe $\lambda$= 23.

**i) Simulation en fonction de $\mu$ (quand s=2)**

Nous obtenons les résultats suivants :

| Taux de service $\mu$ | Risque de saturation $\pi$ |
|---|---|
| 12 | 0,937943 |
| 14 | 0,740896 |
| 16 | 0,601136 |
| 18 | 0,498117 |
| 20 | 0,419841 |
| 22 | 0,358887 |
| 24 | 0,310446 |

Table 3.7 : Risque de saturation

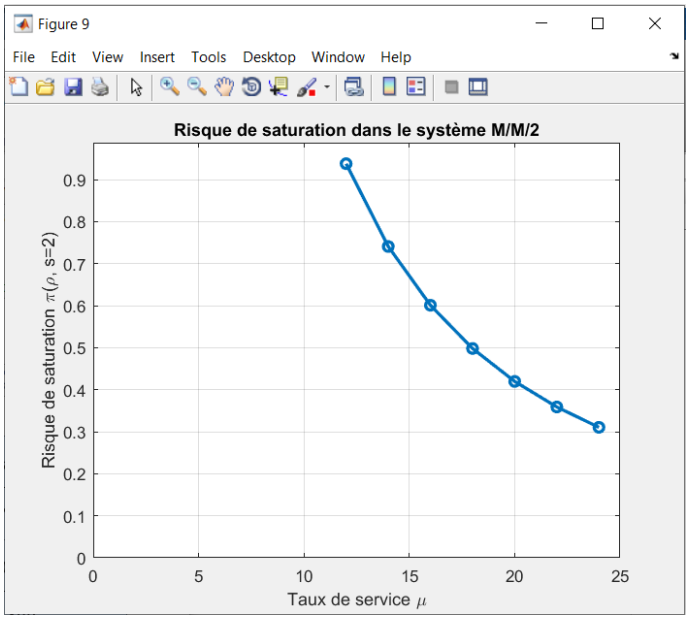


Figure 3.7 : Risque de saturation

**Interprétation des résultats obtenus :**

– Entre un taux de service de 12 à 24, le risque de saturation décroît progressivement, mais reste non négligeable (avec $\pi = 0.310446$), sans atteindre une valeur quasi nulle.

### ii) Simulation en fonction de $\mu$ (quand s=3)

Les résultats sont donnés comme suit :

| Taux de service $\mu$ | Risque de saturation $\pi$ |
|---|---|
| 12 | 0,406054 |
| 14 | 0,290368 |
| 16 | 0,214995 |
| 18 | 0,163707 |
| 20 | 0,127563 |
| 22 | 0,101346 |
| 24 | 0,081860 |

Table 3.8 : Risque de saturation

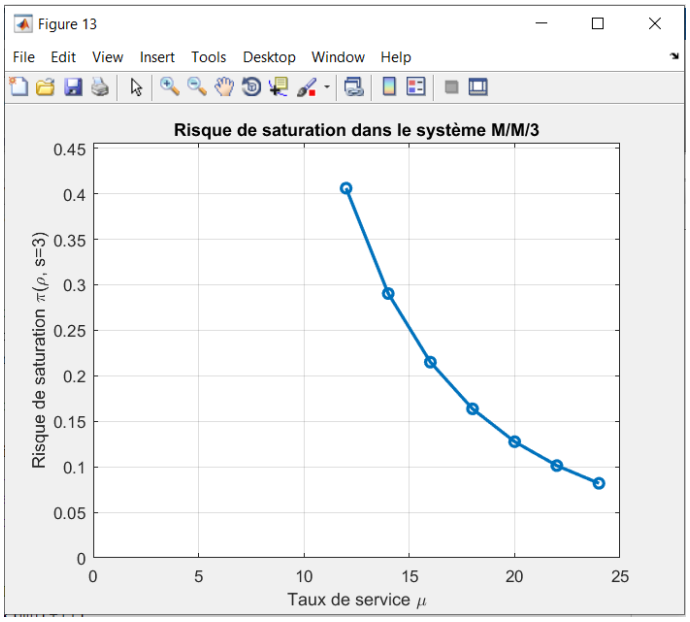


Figure 3.8 : Risque de saturation

**Interprétation des résultats obtenus :**

- Le risque de saturation diminue progressivement de 12 à 24.
- On observe que lorsque s=3, la probabilité de saturation est plus petite pour $s = 2$.
- Cette analyse montre qu'augmenter le nombre de serveurs à 3 réduit effectivement $\pi$, mais ne suffit pas à éliminer complètement le risque de saturation.

***Simulation de la durée de séjour dans le système***

Dans cette section, nous allons décrire la variation du temps moyen d'attente d'un client dans le service en fonction de $\mu$ et $s$.
Pour cela, on fixe $\lambda = 23$.

**i) Simulation en fonction de $\mu$ (quand s=2)**

Les résultats sont les suivants :

| Taux de service $\mu$ | Durée moyenne de séjour W |
|---|---|
| 12 | 0,095751 |
| 14 | 0,078676 |
| 16 | 0,067134 |
| 18 | 0,058712 |
| 20 | 0,052253 |
| 22 | 0,047121 |
| 24 | 0,042935 |

Table 3.9 : Durée de séjour dans le système

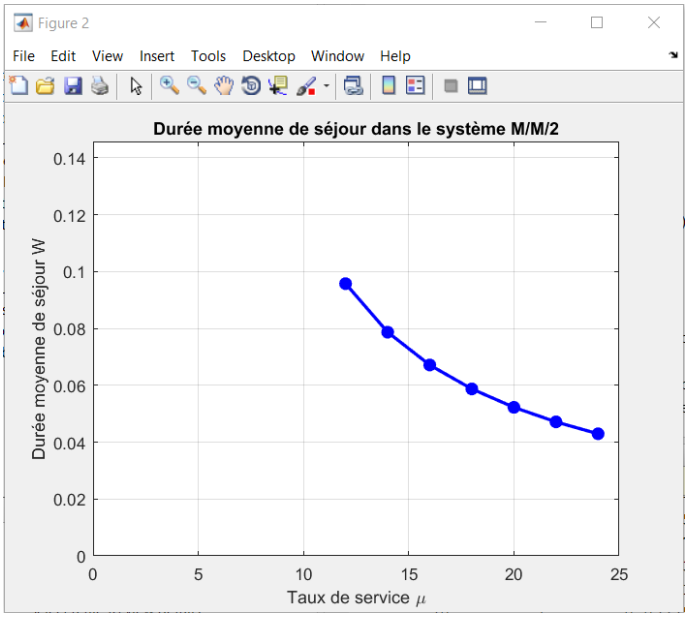


Figure 3.9 : Durée de séjour dans le système

**Interprétation des résultats obtenus :**

- De 12 à 24 on observe une décroissance de la durée de séjour dans le système.
- Auguementer le taux de service $\mu$ diminue la durée de séjour des clients dans le système.

**ii) Durée de séjour en fonction de $\mu$ (quand s=3)**

Nous obtenons les résultats suivants

| Taux de service $\mu$ | Durée moyenne de séjour W |
|---|---|
| 12 | 0,083676 |
| 14 | 0,071616 |
| 16 | 0,062612 |
| 18 | 0,055626 |
| 20 | 0,050047 |
| 22 | 0,045487 |
| 24 | 0,041690 |

Table 3.10 : Durée de séjour dans le système

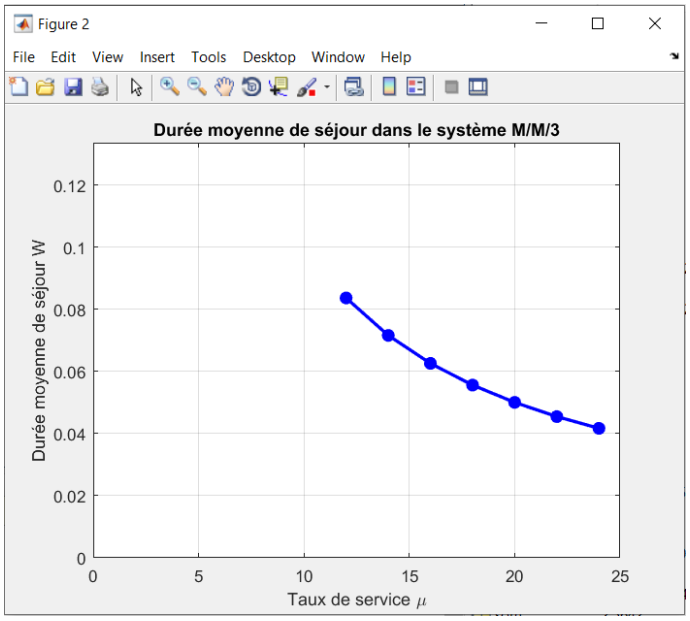


Figure 3.10 : Durée de séjour dans le système

**Interprétation des résultats obtenus :**

- De 12 à 24 la décroissance de $W$ est progressive.
- L'ajout d'un serveur à s=3 ne suffit pas pour réduire le temps de séjour dans le système.

### *Simulation de la longueur de la file*

Dans cette section, nous allons simuler la longueur de la file d'attente dans le service en fonction de $\mu$ et $s$.

On fixe $\lambda$=23.

### i) La longueur de la file en fonction de $\mu$ (quand s= 2)

L'évolution de la longueur de la file est donnée comme suit :

| Taux de service $\mu$ | Longueur de la file $\eta_q$ |
|---|---|
| 12 | 0,285610 |
| 14 | 0,166680 |
| 16 | 0,106593 |
| 18 | 0,072604 |
| 20 | 0,051810 |
| 22 | 0,038326 |
| 24 | 0,029179 |

Table 3.11 : Longueur de la file d'attente

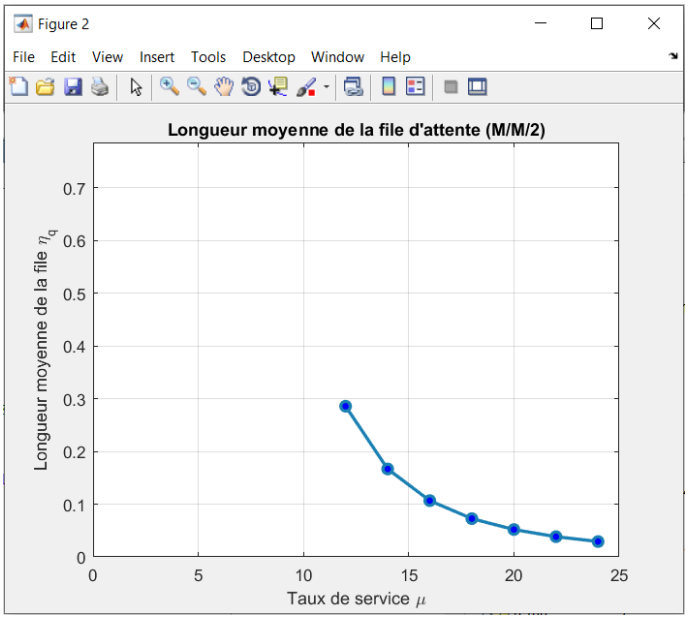


Figure 3.11 : Longueur de la file d'attente

**Interprétation des résultats obtenus :**

- De 12 à 24, la décroissance est progressive.
- La longueur de la file décroît avec $\mu$ et devient très faible dès $\mu = 18$.

### ii) La longueur de la file en fonction de $\mu$ (quand s= 3)

Nous obtenons les résultats suivants

| Taux de service $\mu$ | Longueur de la file $\eta_q$ |
|---|---|
| 12 | 0,007872 |
| 14 | 0,004319 |
| 16 | 0,002567 |
| 18 | 0,001621 |
| 20 | 0,001074 |
| 22 | 0,000740 |
| 24 | 0,000526 |

Table 3.12 : Longueur de la file d'attente

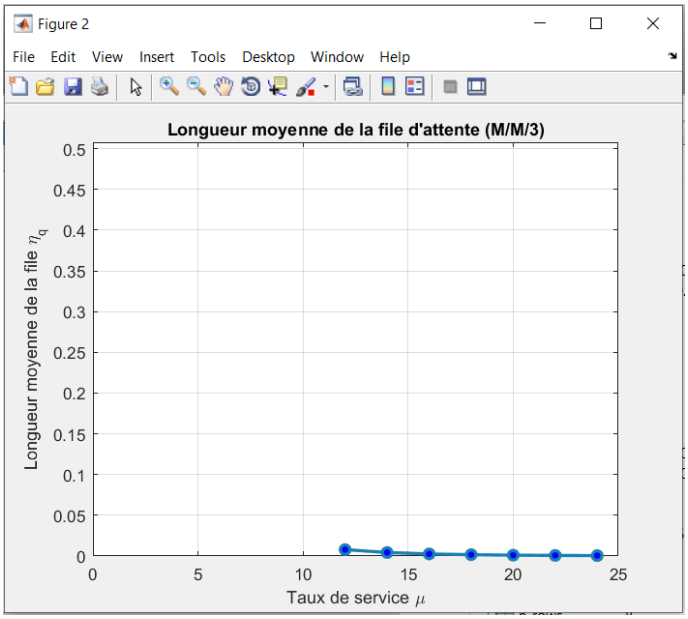


Figure 3.12 : Longueur de la file d'attente

**Interprétation des résultats obtenus :**

- Plus $\mu$ augmente, plus la longueur de la file diminue, atteignant des valeurs négligeables pour $\mu$ compris entre 12 et 24 (avec des valeurs inférieures à 0,02).
- L'ajout d'un troisième serveur réduit significativement la longueur

de la file d'attente dans le système.

### 3.2.3 Service des opérations courantes (M/M/2/5)

Dans cette partie, nous allons expliquer comment le risque de saturation varie en fonction de $\mu$ et $s$.

#### Données

On fixe $\lambda = 5$, $K = 5$

#### i) Simulation en fonction de $\mu$ (quand $s = 2$)

Les résultats sont les suivants :

| Taux de service $\mu$ | Risque de saturation |
|---|---|
| 1 | 0.9814 |
| 3 | 0.6180 |
| 5 | 0.3191 |
| 8 | 0.1476 |
| 11 | 0.0840 |
| 15 | 0.0476 |
| 20 | 0.0278 |

Table 3.13 : Risque de saturation

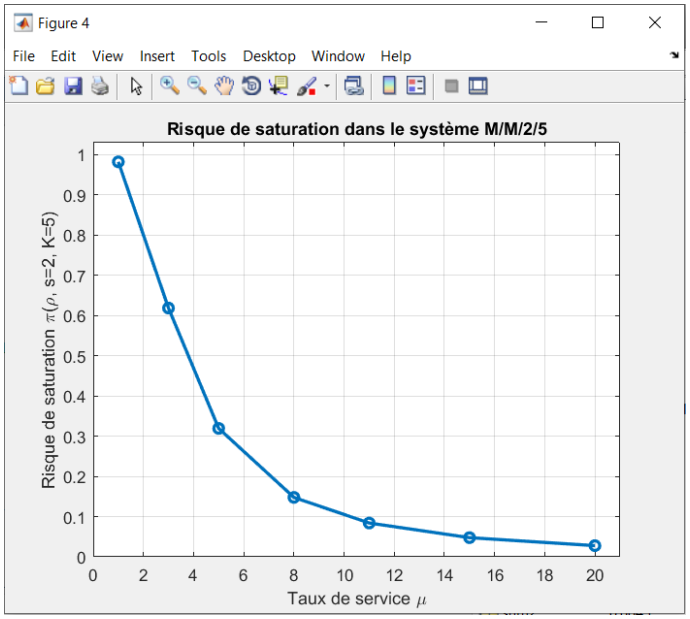


Figure 3.13 : Risque de saturation

**Interprétation des résultats obtenus :**

- De 1 à 8 on observe une décroissance importante du risque de sturation.
- De 8 à 20 la probabilité de saturation devient quasi nulle.
- On conclut que l'augmentation de $\mu$ diminue le risque de saturation dans le système.

**ii) le risque de saturation en fonction de $\mu$ (quand s=3)**

Le risque de saturation évolue comme suit :

| Taux de service $\mu$ | Risque de saturation |
|---|---|
| 1 | 0.8598 |
| 3 | 0.2618 |
| 5 | 0.0878 |
| 8 | 0.0272 |
| 11 | 0.0117 |
| 15 | 0.0050 |
| 20 | 0.0022 |

Table 3.14 : Risque de saturation

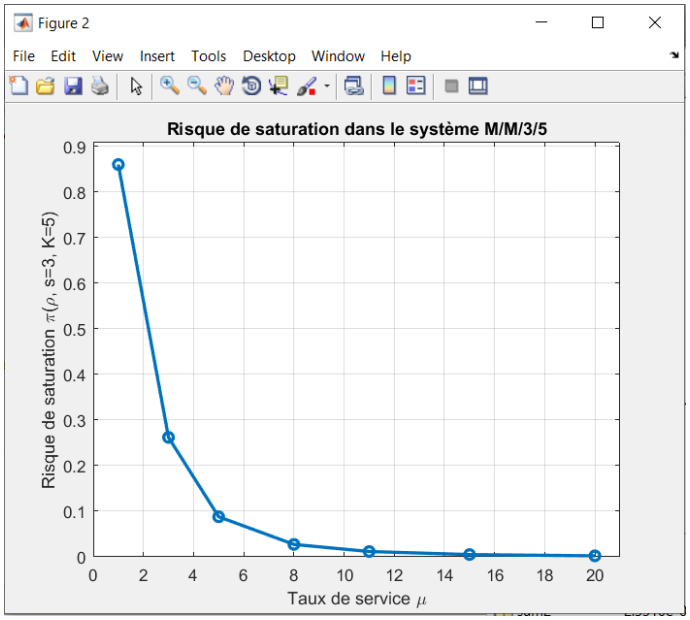


Figure 3.14 : Risque de saturation

**Interprétation des résultats obtenus :**

- La diminution la plus importante est observée de 1 à 3.
- De 3 à 20 le risque de saturation est très proche de zéro.
- Augmenter le nombre de serveurs à s=3 suffit donc de minimiser la saturation dans le système.

*Simulation de la durée de séjour dans le système*

**Données**

$\lambda = 5$, $K = 5$.

**i) La durée de séjour en fonction de $\mu$ (quand s=2)**

Nous obtenons les résultats suivants

| Taux de service $\mu$ | Durée moyenne W |
|---|---|
| 1 | 2.2097 |
| 3 | 0.5113 |
| 5 | 0.2478 |
| 8 | 0.1373 |
| 11 | 0.0957 |
| 15 | 0.0685 |
| 20 | 0.0508 |

Table 3.15 : Durée de séjour dans le système

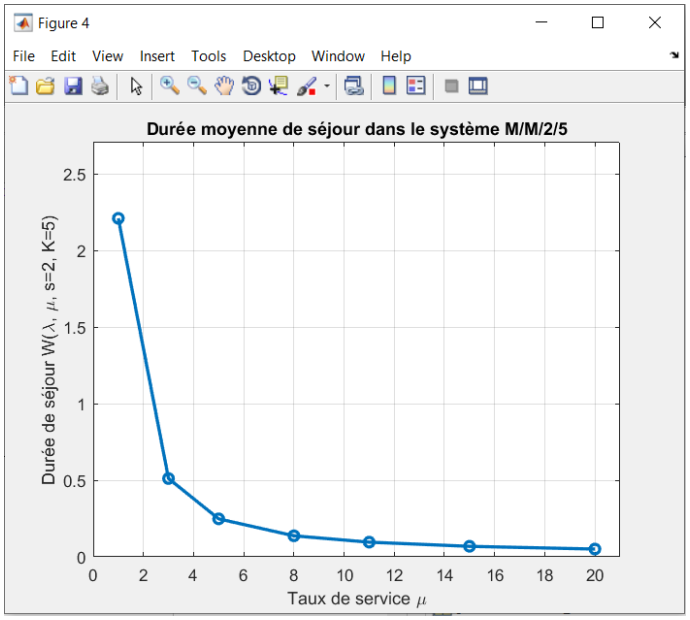


Figure 3.15 : Durée de séjour dans le système

**Interprétation des résultats obtenus :**

- La diminution la plus importante est observée de 1 à 3.
- De 3 à 20 la durée de séjour diminue progressivement.

### ii) Durée de séjour en fonction de $\mu$ (quand s=3)

Nous obtenons les résultats suivants

| Taux de service $\mu$ | Durée moyenne W |
|---|---|
| 1 | 1.4064 |
| 3 | 0.3678 |
| 5 | 0.2068 |
| 8 | 0.1263 |
| 11 | 0.0913 |
| 15 | 0.0668 |
| 20 | 0.0500 |

Table 3.16 : Durée de séjour dans le système

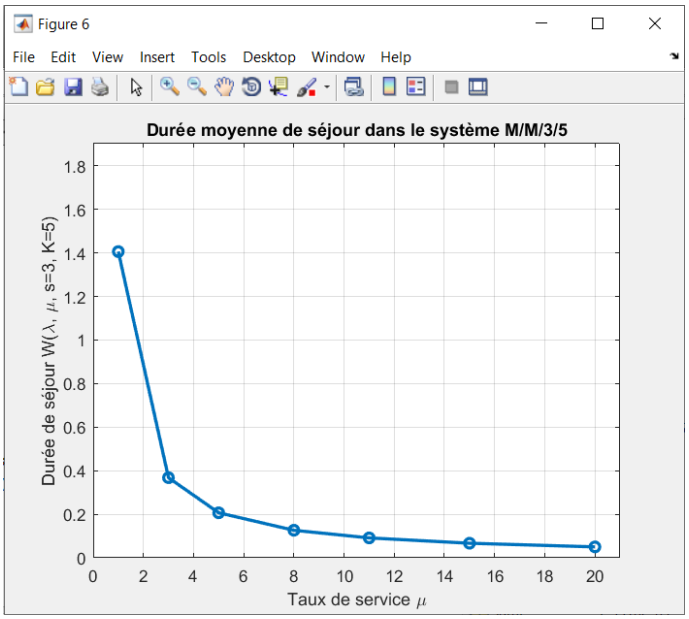


Figure 3.16 : Durée de séjour dans le système

**Interprétation des résultats obtenus :**

- La diminution la plus importante est observée de 1 à 3.
- De 3 à 20 la durée de séjour est très proche de zéro.
- Augmenter le nombre de serveurs à s=3 suffit donc d'améliorer la durée de séjour dans le système.

***Simulation de la longueur de la file d'attente***

**Données** : $\lambda = 5$, $K = 5$

### i) La longueur de la file en fonction de $\mu$ (quand s=2)

Nous obtenons les résultats suivants

| Taux de service $\mu$ | Longueur de la file $\eta_q$ |
|---|---|
| 1 | 2.3932 |
| 3 | 0.7875 |
| 5 | 0.2340 |
| 8 | 0.0614 |
| 11 | 0.0238 |
| 15 | 0.0094 |
| 20 | 0.0039 |

Table 3.17 : Longueur de la file d'attente

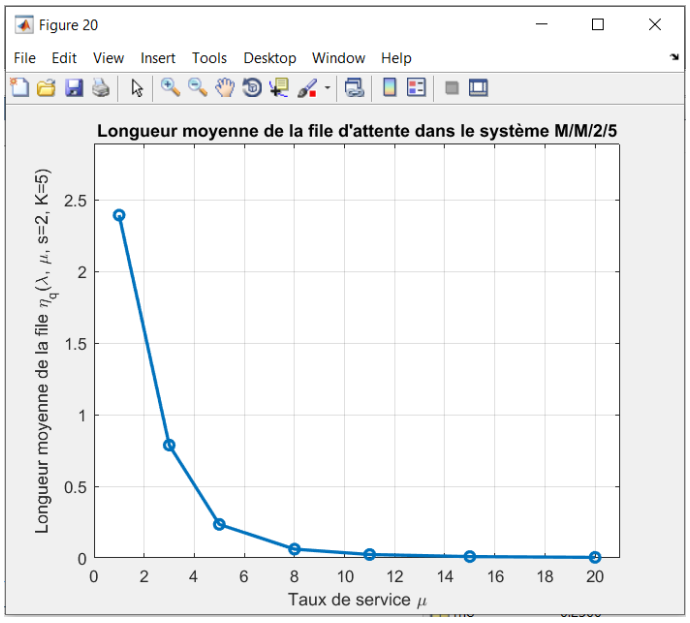


Figure 3.17 : Longueur de la file d'attente

**Interprétation des résultats obtenus :**

- La diminution la plus importante est observée de 1 à 3.
- De 3 à 8 la longueur de la file est très proche de zéro.
- De 8 à 20 les valeurs sont quasi nulles.

### ii) La longueur de la file en fonction de $\mu$ (quand s= 3)

Nous obtenons les résultats suivants

| Taux de service $\mu$ | Longueur de la file $\eta_q$ |
|---|---|
| 1 | 1.1405 |
| 3 | 0.1647 |
| 5 | 0.0338 |
| 8 | 0.0064 |
| 11 | 0.0020 |
| 15 | 0.0006 |
| 20 | 0.0002 |

Table 3.18 : Longueur de la file d'attente

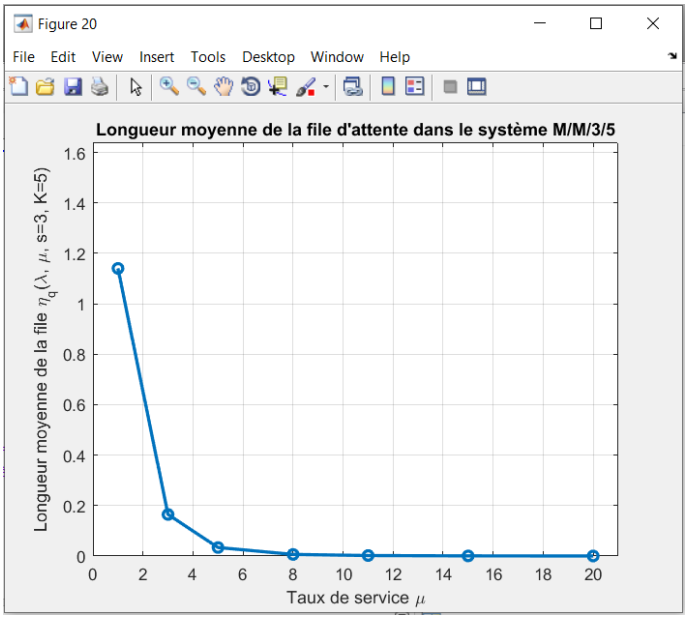


Figure 3.18 : Longueur de la file d'attente

**Interprétation des résultats obtenus :**

- De $\mu = 1$ à $\mu = 5$, la baisse de $\eta_q$ est spectaculaire (de 1,1405 à $0,0338$).
- De $\mu = 8$ à $\mu = 20$, la baisse continue mais de manière plus marginale (de $0,0064$ à $0,0002$).
- L'ajout d'un troisième serveur ($s = 3$) améliore considérablement les performances du système, en réduisant fortement la longueur moyenne de la file $\eta_q$.

## 3.3 Conclusion

Dans cette partie, nous avons simulé les différentes performances des services congestionnés, telles que le risque de saturation, la durée de séjour d'un client dans le système et la longueur de la file d'attente, à l'aide du logiciel MATLAB.
Les simulations numériques obtenues viennent étayer et confirmer les résultats préalablement établis au chapitre 2. L'analyse comparative révèle des besoins distincts selon les services :

- Pour le service financier, l'ajout d'un troisième serveur s'avère crucial pour réduire significativement le risque de saturation
- Pour le service d'Animation commerciale, qui fonctionne déjà avec un seul serveur, nous avons proposé l'ajout d'un deuxième serveur combiné à l'accélération du temps de traitement des demandes

Nous proposons donc des stratégies différenciées adaptées à chaque contexte de service. Ces approches d'optimisation, que nous explorerons en détail dans la partie suivante, permettront une amélioration ciblée des performances de chaque système.

# Partie II : Améliorations possibles et approches intelligentes

## 3.1 Introduction

Dans cette section, nous utilisons les résultats des simulations et les stratégies d'optimisation présentées précédemment pour proposer des solutions classiques et concrètes permettant d'optimiser les performances des systèmes. Cependant, bien que ces approches soient efficaces dans certains contextes, elles atteignent souvent leurs limites face à la complexité croissante des environnements réels. C'est pourquoi, dans un second temps, nous introduirons des approches plus avancées basées sur l'intelligence artificielle (IA), qui permettent une gestion plus fine, adaptative et prédictive des flux de clients, là où les méthodes traditionnelles montrent leurs limites.

## 3.2 Approches classiques et innovantes pour l'optimisation des files d'attente

### 3.2.1 Approches classqiues

Il s'agit de mettre en place des solutions simples et peu coûteuses, notamment à travers des aménagements digitaux adaptés aux spécificités de chaque service. Déjà éprouvées dans d'autres établissements bancaires ou de services, ces mesures ont pour objectif d'alléger la charge des guichets et d'améliorer la fluidité des files d'attente, sans recourir, dans un premier temps, à des technologies complexes. Les différentes solutions seront développées dans les sous-sections qui suivent.

**Solutions digitales**

**A. Espace "Renseignements"**

Pour améliorer l'efficacité de l'accueil et favoriser une focalisation accrue des agents sur les tâches à forte valeur ajoutée, on suggère pour cette agence de CPA de Tizi Ouzou, la création d'un espace "Renseignements" . Cette zone serait dotée d'un écran interactif conçu pour permettre aux clients de consulter de manière autonome les informations actuelles. Cette solution vise à désengorger les guichets en orientant les demandes d'informations simples vers un dispositif automatisé, libérant ainsi les ressources humaines pour les interventions complexes. Cela entraîne un gain de temps notable pour toutes les parties concernées et

une amélioration concrète de l'expérience client.

### B. Formulaire Digital pour Ouvertures de Compte :

En mettant en place un système de pré-remplissage en ligne, les clients pourraient saisir leurs données à l'avance depuis chez eux. Ces informations seraient ensuite automatiquement intégrées dans le système de l'agent. Cela permettrait à l'agent de se concentrer uniquement sur l'analyse de la demande et la prise de décision, au lieu de passer du temps à recueillir les informations de base.
Cette approche de pré-remplissage en ligne est déjà testée et a fait ses preuves dans d'autres établissements, notamment à l'ANEM (Agence Nationale de l'Emploi).

### C. Système de tickets classique et Affichage :

Les tickets orientent automatiquement les clients vers un guichet selon l'ordre des arrivées, créant une files d'attente virtuelle et une transparence dans les différent services.

## Solutions physiques (sans investissement)

### A. Réallocation Dynamique des Guichets :

Affecter un guichet aux opérations rapides : faire une mise à jour, demande de solde ou bien un retrait en DA ($<$5min) et le deuxième guichet aux opérations longues (plus complexes) comme le versement et le virement de devise ($>$25min).

### B. Guichets supplémentaires :

La mise en place d'un troisième guichet supplémentaire, spécifiquement destiné aux opérations des clients particuliers, est proposée. Ce guichet offrira une flexibilité opérationnelle accrue : il permettra de traiter les tâches courantes en cas de files d'attente prolongées et assurera une continuité de service en cas de panne de l'un des autres guichets. Cette configuration vise à optimiser l'efficacité globale de l'accueil et à améliorer significativement l'expérience client.

### 3.2.2 Approches innovantes

Dans cette section, nous introduisons quelques concepts clés relatifs aux solutions innovantes en gestion des files d'attente par IA. Cette présentation vise à poser les bases des approches qui seront développées ultérieurement.

## A. Prédiction des temps d'attente (Machine learning)

Pour optimiser la gestion des flux en temps réel, le machine learning exploite des données historiques afin de prédire les temps d'attente avec précision. Des modèles comme les Support Vector Machines (SVM), notamment en régression (SVR), permettent d'anticiper ces délais et d'améliorer la prise de décision, par exemple dans le secteur bancaire via des tableaux de bord interactifs.
Pour optimiser la gestion des files d'attente en temps réel, une banque peut implémenter un modèle SVR selon les étapes suivantes :

- **Collecte des données historiques**
  - Fréquentation horaire
  - Durée moyenne des opérations
  - Nombre de guichets ouverts
- **Apprentissage du modèle**
  - Le SVR analyse les données pour identifier les motifs récurrents
  - Construction de l'hyperplan optimal pour les prédictions
- **Prédiction en temps réel**
  - Le système ajuste continuellement ses estimations
  - Adaptation à la situation actuelle (affluence, personnel disponible)
- **Utilisation des résultats**
  - Alimentation de tableaux de bord interactifs
  - Information des clients via applications mobiles
  - Aide à la décision pour le personnel

Le schéma suivant illustre les différentes étapes de traitement mises en œuvre pour prédire les temps d'attente dans un système bancaire, depuis la collecte des données jusqu'à l'exploitation des résultats.

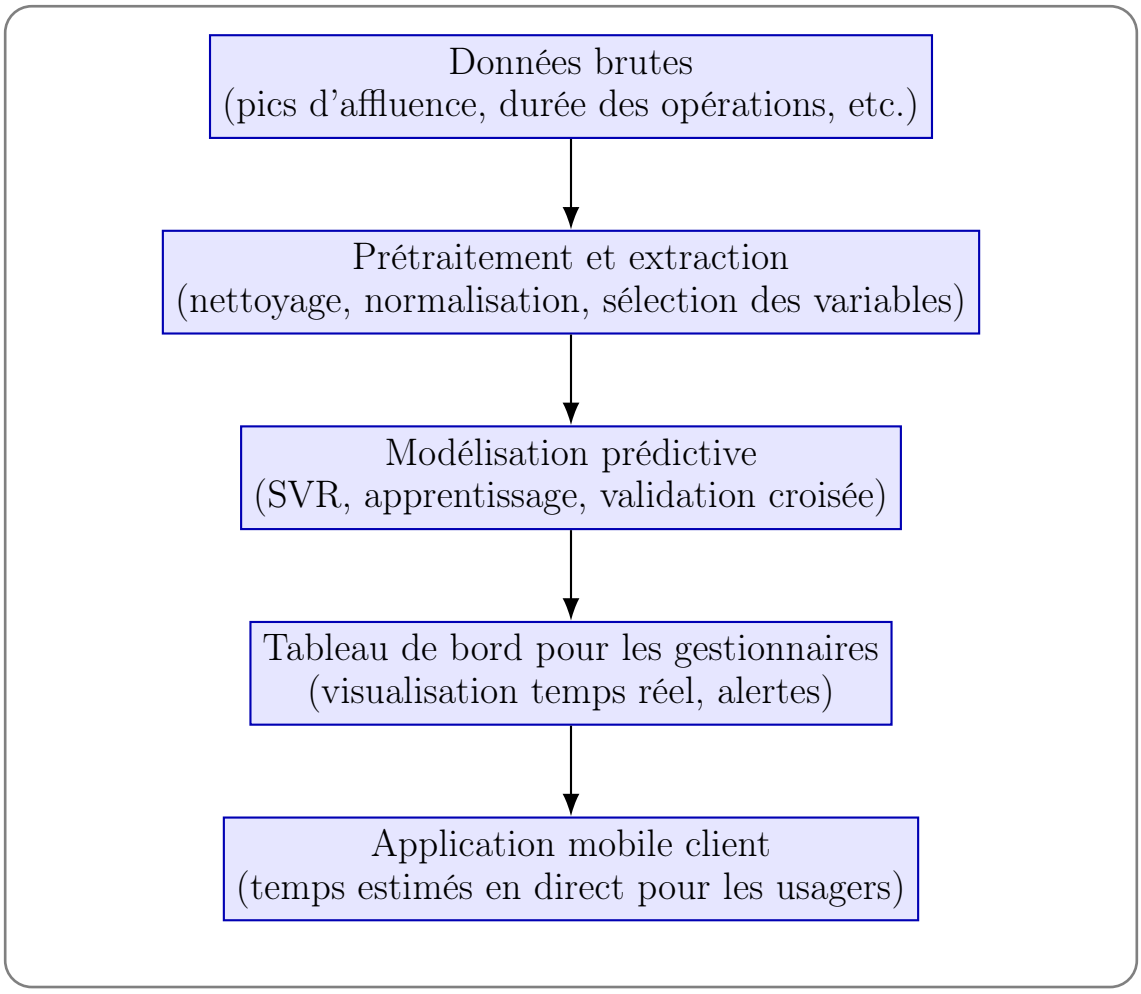


FIGURE 3.1 – Étapes de traitement pour la prédiction des temps d'attente dans un système bancaire

### B. Priorisation intelligente (Reinforcement learning)

Des algorithmes classent les clients selon des critères (durée estimée du service, statut VIP, urgence). Parmi ces algorithmes, citons le Priority Queue couplé à du Reinforcement Learning (RL), où un agent apprend à attribuer des priorités via des récompenses (ex : minimiser l'attente moyenne). Transposé aux banques, il optimiserait l'allocation des guichets.
Dans les banques modernes, la gestion des files d'attente repose souvent sur un système de tickets électroniques couplé à des automates de

libre-service. Ces derniers incluent :

- Les distributeurs automatiques de billets (DAB) pour les retraits et consultations de solde,
- Les bornes de dépôt pour espèces/chèques,
- Les terminaux de virements/recharges, permettant de délester les guichets des opérations simples.

Le Reinforcement Learning (RL) optimise ce système de deux manières complémentaires :

Classification intelligente des tickets :

- Un agent RL analyse en temps réel chaque demande client (durée estimée, statut, urgence),
- Il attribue dynamiquement :

  Un ticket prioritaire (ex : VIP ou demande courte → guichet dédié),

  Un ticket standard (ex : dépôt → automate si possible).

Apprentissage continu :

- L'agent reçoit des récompenses (ex : +1 si le temps d'attente moyen diminue) ou des pénalités (ex : -2 si un client abandonne),
- Il affine sa politique de priorisation via des algorithmes comme le Q-learning, s'adaptant aux variations d'affluence.

Le système de priorisation intelligente repose sur un cycle d'apprentissage automatique continu. Le schéma suivant illustre ce processus.

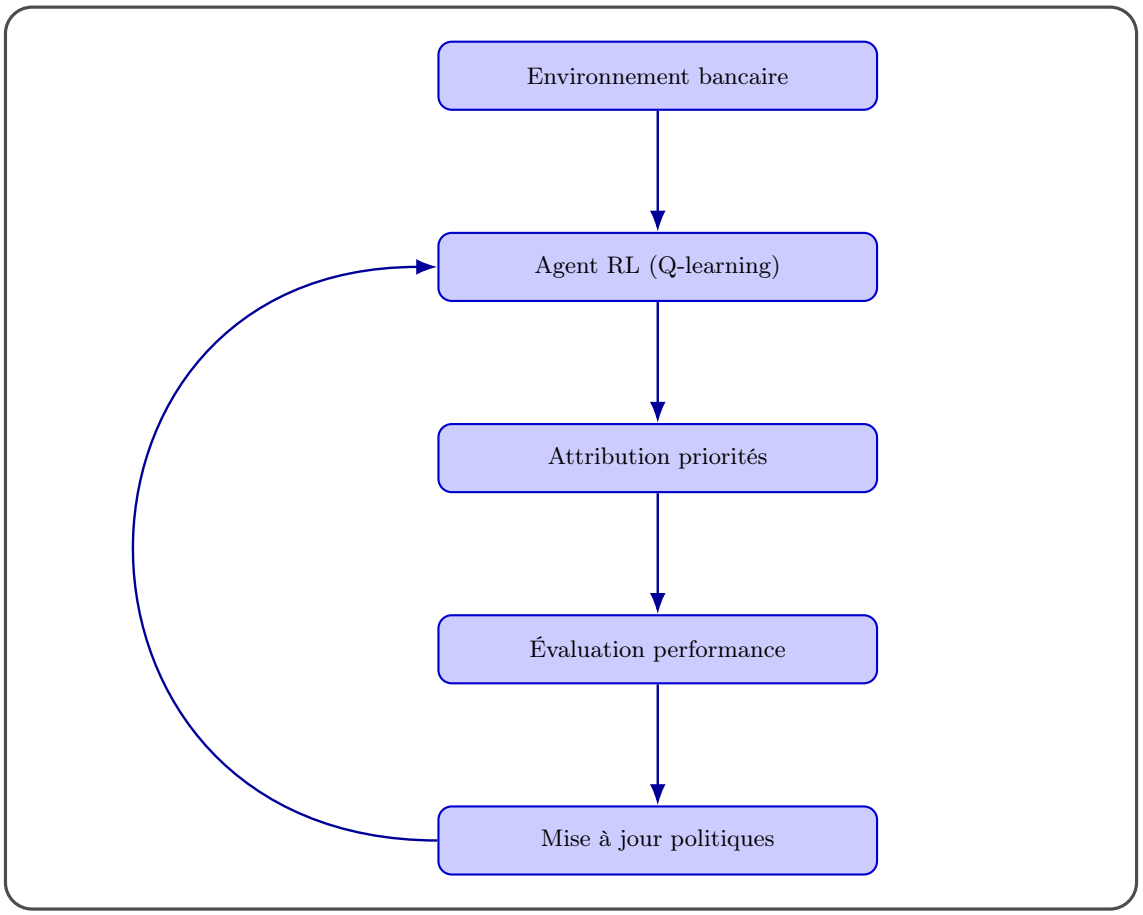


FIGURE 3.2 – Architecture du système de priorisation par Reinforcement Learning

## 3.3 Conclusion

Dans cette partie, nous avons proposé des solutions concrètes pour améliorer la qualité de service et fluidifier les systèmes d'attente, en les catégorisant selon deux approches : des solutions sans investissement majeur (physiques) et des solutions digitales.
Nous avons également exploré le potentiel de modèles d'intelligence artificielle, tels que le modèle **Machine Learning** et le modèle **Reinforcement Learning**, pour accélérer l'expérience client et optimiser l'efficacité des services. Ces avancées transforment ainsi la gestion des files d'attente en un système dynamique intelligent et centré sur le client.

# Conclusion générale

Ce mémoire propose des approches théoriques rigoureuses et des applications concrètes pour améliorer la qualité du service dans un environnement bancaire. En effet l'objectif principal de ce travail est de réduire les files d'attente et d'améliorer les performances des services les plus sollicités de l'agence CPA Bank de Tizi-Ouzou. Pour cela, nous avons mené une étude structurée en plusieurs étapes, combinant des outils mathématiques, des modèles de files d'attente, des analyses de performance et des simulations.

Dans un premier temps, nous avons présenté les bases théoriques nécessaires à notre étude, notamment les processus stochastiques, avec un accent particulier sur les processus de Markov. Ces outils sont essentiels pour comprendre les phénomènes aléatoires liés aux arrivées et aux services dans les systèmes de file d'attente.

Par la suite, nous avons étudié différents modèles de files d'attente comme (M/M/1), (M/M/s), (M/M/1/K) et (M/M/s/K). Ces modèles nous ont permis d'analyser les performances des systèmes de service selon plusieurs paramètres : le nombre de serveurs, le taux d'arrivée des clients, la durée de service, etc.

Ensuite, nous avons appliqué ces modèles au contexte réel de l'agence bancaire étudiée. À l'aide du logiciel MATLAB, nous avons réalisé plusieurs simulations qui ont permis d'évaluer les performances des services dans différentes situations. Des ajustements ont alors été proposés, notamment l'augmentation du nombre de serveurs ou l'optimisation de la vitesse de traitement.

Nous avons également introduit des approches plus intelligentes, qu'il s'agisse d'ajustements organisationnels ou structurels, ou encore de solutions basées sur l'intelligence artificielle dans le but de mieux anticiper l'affluence, d'adapter les ressources en temps réel et d'offrir un service plus rapide et plus efficace.

# Bibliographie

[1] BABES, Malika. *Statistiques, files d'attente et simulation.* Alger : OPU, 1992.
[2] BAYNAT, Bruno. *Théorie des files d'attente.* Paris : Hermès - Lavoisier, 1970.
[3] BHATTACHARYA, Chandrima et SINHA, Manish. "Role of Artificial Intelligence in Banking for Leveraging". *Accounting and Finance Research*, vol. 16, n° 5, 2022.
[4] HAMADOUCHE, D. *Cours de Master RO : Processus stochastiques et files d'attente.*
[5] HASSANI, Karima et OUHACHI, Dyhia. *Modélisation et Optimisation du télétrafic dans un centre d'appels téléphoniques.* Mémoire de master, Université Mouloud Mammeri de Tizi Ouzou, 2018-2019.
[6] IWAN, Adhicandra, NURHIDAYATI, Safitri et FAUZAN, Tribowo Rachmat. "Optimization of Hospital Queue Management Using Priority Queue Algorithm and Reinforcement Learning for Emergency Service Prioritization". *International Journal of Science, Engineering and Computer Technology*, vol. 4, n° 2, 2024.