# 1 Introduction

## 1.1 Contexte

Dans de nombreux systèmes industriels et informatiques, la gestion efficace des files d'attente est cruciale pour garantir des performances acceptables. Que ce soit dans un centre d'appels, un système informatique, un hôpital ou un poste de péage, les entités doivent souvent attendre l'accès à une ressource partagée. Pour modéliser et analyser ces situations, la théorie des files d'attente fournit un cadre mathématique rigoureux, initialement introduit par le chercheur danois A. K. Erlang au début du XX$^{e}$ siècle. Cette discipline permet d'étudier le comportement de systèmes soumis à des arrivées aléatoires d'utilisateurs et à un processus de service.

## 1.2 Objectifs du TP

L'objectif de ce travail pratique est de simuler et comparer trois modèles fondamentaux de files d'attente mono-serveur : **M/M/1**, **G/M/1** et **M/G/1**. Ces modèles permettent de mettre en évidence l'impact des distributions aléatoires utilisées pour les arrivées ou les services sur les indicateurs de performance tels que :

- Le **temps moyen de réponse**, représentant le temps total passé par un client dans le système.
- Le **temps moyen d'attente** dans la file avant le début du service.
- Le **taux d'occupation du serveur**, mesurant la proportion de temps pendant laquelle le serveur est actif.

Les simulations seront réalisées à l'aide d'une approche événementielle discrète, en faisant varier le taux d'arrivée $\lambda$ tout en maintenant un taux de service constant $\mu$. Les résultats obtenus seront confrontés aux valeurs théoriques dans le cas du modèle M/M/1, puis comparés entre les modèles pour analyser les écarts induits par des distributions non exponentielles.

## 1.3 Présentation des modèles étudiés

- **M/M/1** : Dans ce modèle classique, les arrivées des clients suivent une loi exponentielle de paramètre $\lambda$ (processus de Poisson), et les temps de service sont également exponentiels de paramètre $\mu$. Ce modèle est bien connu pour sa simplicité analytique et possède des formules théoriques exactes.
- **G/M/1** : Ici, les arrivées ne sont plus exponentielles mais suivent une distribution *générale* (par exemple, uniforme ou déterministe). Le temps de service reste exponentiel. Ce modèle permet d'étudier l'impact d'une distribution d'arrivées moins variable ou plus régulière.
- **M/G/1** : Ce modèle conserve un processus d'arrivée de type Poisson (exponentiel), mais généralise la loi de service. Cela permet de simuler des comportements de service plus réalistes ou variés (par exemple : loi de Weibull, log-normale).

Ces modèles sont particulièrement utiles pour illustrer comment la nature stochastique des arrivées et des services influence les performances globales du système. Leur simulation nécessite l'utilisation de générateurs aléatoires adaptés, ainsi que d'une approche rigoureuse de mise à jour temporelle basée sur les événements clés (arrivées et départs).

# 2 Fondements théoriques

## 2.1 Théorie des files d'attente

La théorie des files d'attente est une branche des probabilités et des processus stochastiques qui modélise les systèmes dans lesquels des clients (ou entités) attendent pour accéder à une ressource limitée, telle qu'un serveur. Le cadre général repose sur les processus de naissance et de mort, dans lesquels les arrivées sont interprétées comme des "naissances" et les services comme des "décès".

Le premier modèle mathématique de file d'attente a été introduit par A. K. Erlang en 1909 dans le cadre des réseaux téléphoniques. Depuis, cette théorie s'est largement développée et trouve des applications dans l'informatique, les télécommunications, l'industrie, les transports et les services publics.

## 2.2 Le modèle M/M/1

Le modèle **M/M/1** est l'un des modèles les plus simples et les plus étudiés dans la théorie des files d'attente. Il est caractérisé par :

- **M** (Markovien) pour les **arrivées**, suivant un processus de Poisson de paramètre $\lambda$ (temps inter-arrivées exponentiels),
- **M** pour les **services**, également exponentiels de paramètre $\mu$,
- **1** seul serveur.

Ce système est modélisé par une chaîne de Markov à temps continu. Il admet une distribution stationnaire si le taux d'arrivée est strictement inférieur au taux de service, soit $\rho = \lambda/\mu < 1$. Dans ce cas, le système est stable.

La probabilité d'avoir $n$ clients dans le système à l'équilibre est donnée par :

$$\pi_n = (1 - \rho) \cdot \rho^n, \quad \text{pour } n \geq 0$$

## 2.3 Notation de Kendall

Pour décrire les caractéristiques d'un système de file d'attente, on utilise la notation de **Kendall**, notée généralement :

$$A/B/s/K/N/D$$

où :

- $A$ : loi des arrivées (M : exponentielle, D : déterministe, G : générale, etc.),
- $B$ : loi des services,
- $s$ : nombre de serveurs,
- $K$ : capacité maximale de la file (par défaut $\infty$),
- $N$ : taille de la population (par défaut $\infty$),
- $D$ : discipline de service (FIFO, LIFO, etc.).

Par exemple, le système M/M/1 correspond à des arrivées et services exponentiels, un seul serveur, une capacité et population infinies, et un ordre de service FIFO implicite.

## 2.4 Indicateurs de performance théoriques

Les principaux indicateurs utilisés pour évaluer les performances d'un système M/M/1 en régime permanent sont :

— **Taux d'occupation du serveur** : $\rho = \lambda/\mu$
— **Nombre moyen de clients dans le système** :

$$L = \frac{\rho}{1-\rho}$$

— **Nombre moyen de clients dans la file** :

$$L_q = \frac{\rho^2}{1-\rho}$$

— **Temps moyen dans le système (temps de réponse)** :

$$W = \frac{1}{\mu - \lambda}$$

— **Temps moyen dans la file d'attente** :

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu-\lambda}$$

Ces formules permettent une évaluation analytique du système. Elles servent également de référence pour comparer les résultats issus des simulations dans ce TP, notamment dans le cas du modèle M/M/1 où les résultats théoriques sont disponibles explicitement.

# 3 Méthodologie de simulation

## 3.1 Description générale

La simulation des files d'attente a été réalisée à l'aide d'un simulateur événementiel discret codé en Python. Ce type de simulation repose sur une progression du temps non continue, basée sur la succession d'événements clés (arrivées et départs). L'objectif est de reproduire le fonctionnement d'un système mono-serveur en suivant le parcours de chaque client, du moment de son arrivée jusqu'à la fin de son service.

Trois modèles ont été implémentés :
— **M/M/1** : Arrivées et services exponentiels,
— **G/M/1** : Arrivées suivant une loi générale, services exponentiels,
— **M/G/1** : Arrivées exponentielles, services suivant une loi générale.

Les résultats sont obtenus en faisant varier le taux d'arrivée $\lambda$ et en fixant le taux de service $\mu = 1$, afin d'analyser l'évolution des performances du système.

## 3.2 Logique événementielle

La simulation suit une logique événementielle déterminée par deux files d'événements :
— **Prochaine arrivée** : générée en fonction de la loi d'arrivée (exponentielle ou générale),
— **Prochain départ** : planifié uniquement si le serveur est occupé.

L'algorithme principal est structuré comme suit :

1. Initialisation des variables, du temps de simulation et planification de la première arrivée.

2. À chaque itération :
   — Avancer l'horloge au prochain événement.
   — Mettre à jour les compteurs statistiques.
   — Si l'événement est une arrivée :
      — Si le serveur est libre : démarrer immédiatement le service.
      — Sinon : placer le client dans la file d'attente.
   — Si l'événement est un départ :
      — Si la file est vide : libérer le serveur.
      — Sinon : sélectionner le client suivant et démarrer son service.
3. Répéter jusqu'à atteindre le nombre souhaité de clients servis.

### 3.3 Génération de variables aléatoires

La simulation repose sur la génération de nombres pseudo-aléatoires pour simuler les temps d'arrivée et de service. Trois types de lois ont été utilisées :

— **Exponentielle** : utilisée dans les modèles M/M/1 et M/G/1.

$$X = -\frac{1}{\lambda}\ln(U), \quad U \sim \mathcal{U}(0,1)$$

— **Uniforme** : utilisée pour simuler des arrivées dans G/M/1 avec faible variabilité.

$$X \sim \mathcal{U}(a,b)$$

— **Weibull ou log-normale** : utilisées pour simuler des services dans M/G/1 avec variabilité réaliste.

Chaque simulation est répétée plusieurs fois (généralement 10 répétitions) avec des graines aléatoires différentes pour assurer la robustesse statistique des résultats.

### 3.4 Hypothèses et paramètres de simulation

Les principales hypothèses retenues sont les suivantes :

— Le système comporte un **seul serveur** (mono-serveur).
— La discipline de service est **FIFO** (First-In, First-Out).
— La **capacité de la file est infinie** : aucun client n'est rejeté.
— Le système est **vide au départ** et les statistiques sont collectées après un échauffement implicite.
— Les **paramètres fixés** pour toutes les simulations sont :
   — Taux de service : $\mu = 1.0$
   — Nombre de clients servis par simulation : $N = 100{,}000$
   — Valeurs de $\lambda$ explorées : de 0.1 à 0.9 avec un pas de 0.05
   — Nombre de répétitions par point : 10

Ces choix permettent une comparaison fiable entre les trois modèles, tout en contrôlant la charge du système (à travers le paramètre $\rho = \lambda/\mu$).

## 4 Résultats expérimentaux

### 4.1 M/M/1 : résultats simulés vs théoriques

Le modèle M/M/1 constitue une référence classique en théorie des files d'attente. Grâce à sa formulation analytique simple, il permet de comparer les résultats de simulation avec

les formules théoriques établies.

Les indicateurs suivants ont été comparés :

— Temps moyen dans le système ($W$)
— Temps moyen d'attente dans la file ($W_q$)
— Taux d'occupation du serveur ($\rho$)

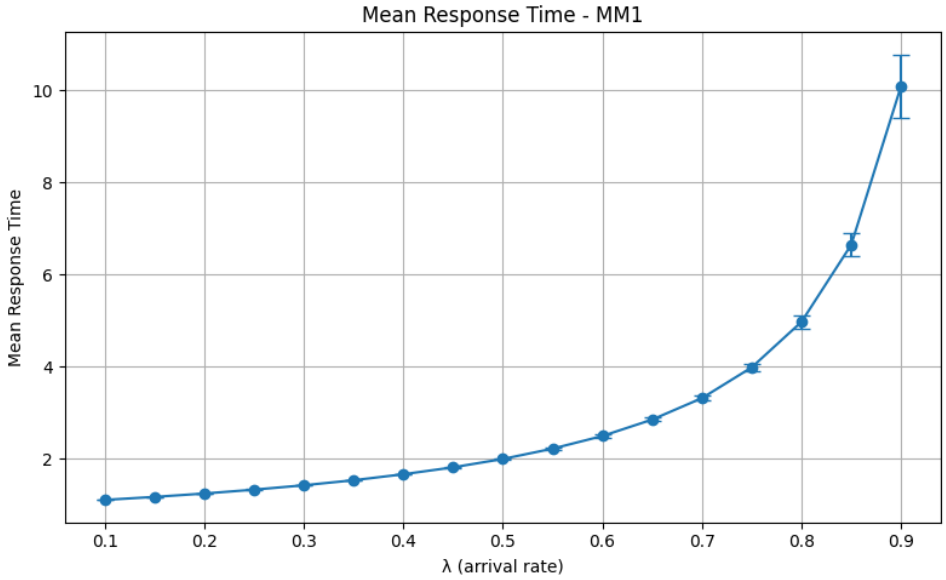


FIGURE 1 – Temps de réponse simulé vs $\lambda$ pour M/M/1 (barres d'erreur incluses)

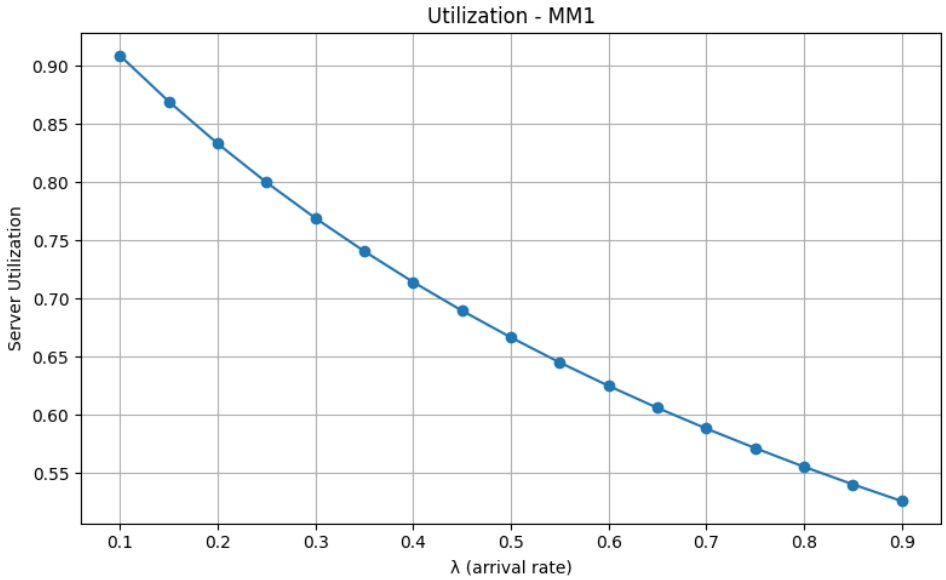


FIGURE 2 – Utilisation du serveur simulée vs théorie pour M/M/1

On observe que les résultats simulés suivent de près les prédictions théoriques lorsque $\lambda < \mu$. De légères fluctuations sont visibles près de $\lambda = 0.9$ en raison de l'effet transitoire et de la saturation progressive du système.

## 4.2 G/M/1 : impact de l'arrivée non exponentielle

Le modèle G/M/1 permet d'étudier l'effet d'une régularité accrue dans les arrivées (par exemple, distribution uniforme). Cette régularité réduit la probabilité de longues périodes de surcharge.

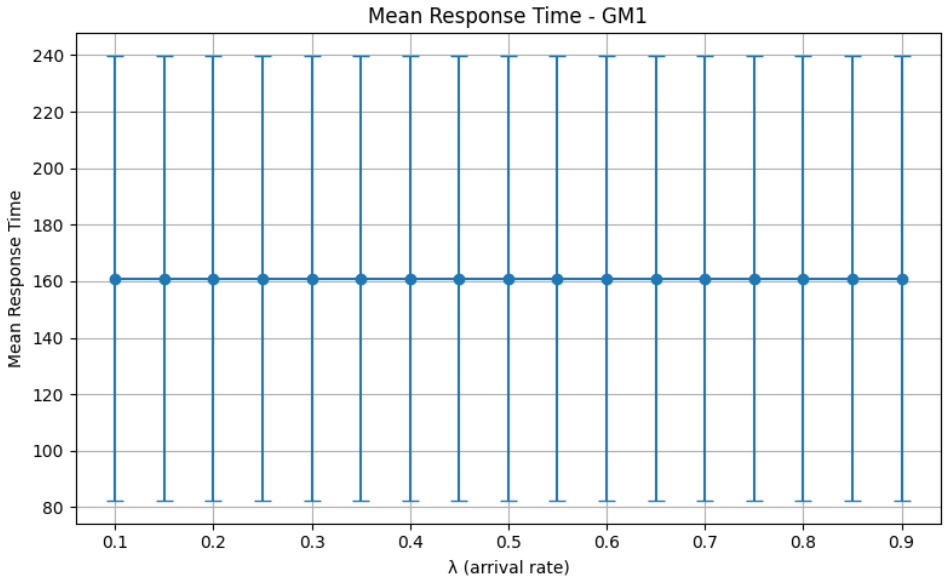


FIGURE 3 – Temps de réponse moyen simulé pour G/M/1 (arrivées uniformes)

Comparé au M/M/1, on note une **réduction du temps d'attente moyen** pour les mêmes valeurs de $\lambda$, surtout dans les zones de forte utilisation. Ceci est cohérent avec l'intuition : une arrivée plus régulière évite la formation de pics soudains de charge.

## 4.3 M/G/1 : impact d'un service non exponentiel

Dans le modèle M/G/1, nous remplaçons la loi exponentielle des services par une loi de Weibull ou log-normale, qui introduit plus ou moins de variabilité.

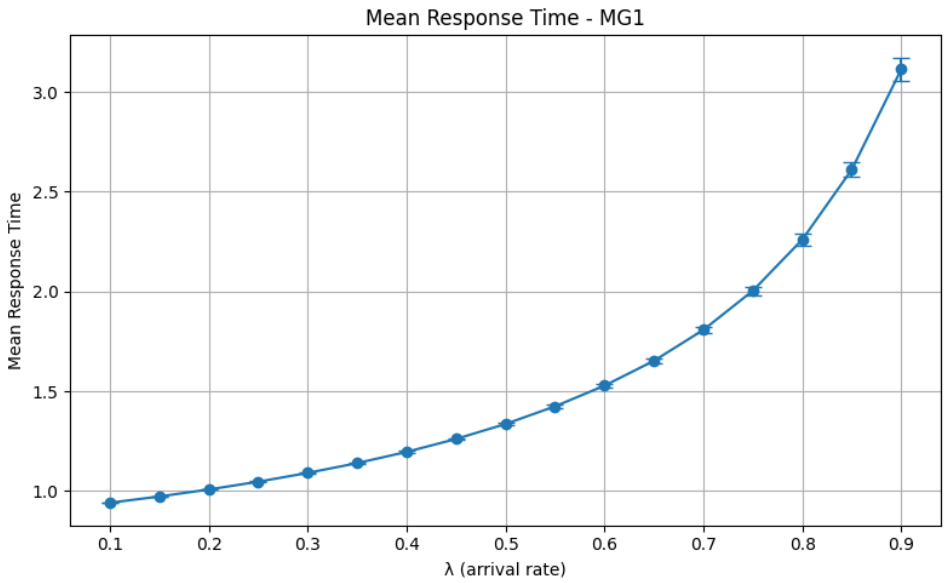


FIGURE 4 – Temps de réponse moyen simulé pour M/G/1 (service Weibull)

L'effet d'une loi de service plus variable est une augmentation du temps moyen dans le système, notamment lorsque $\lambda$ s'approche de $\mu$. En effet, la présence de services exceptionnellement longs peut faire croître rapidement la taille de la file.

## 4.4 Comparaison croisée des trois modèles

La figure suivante présente une comparaison directe des trois modèles pour le temps de réponse moyen en fonction de $\lambda$.

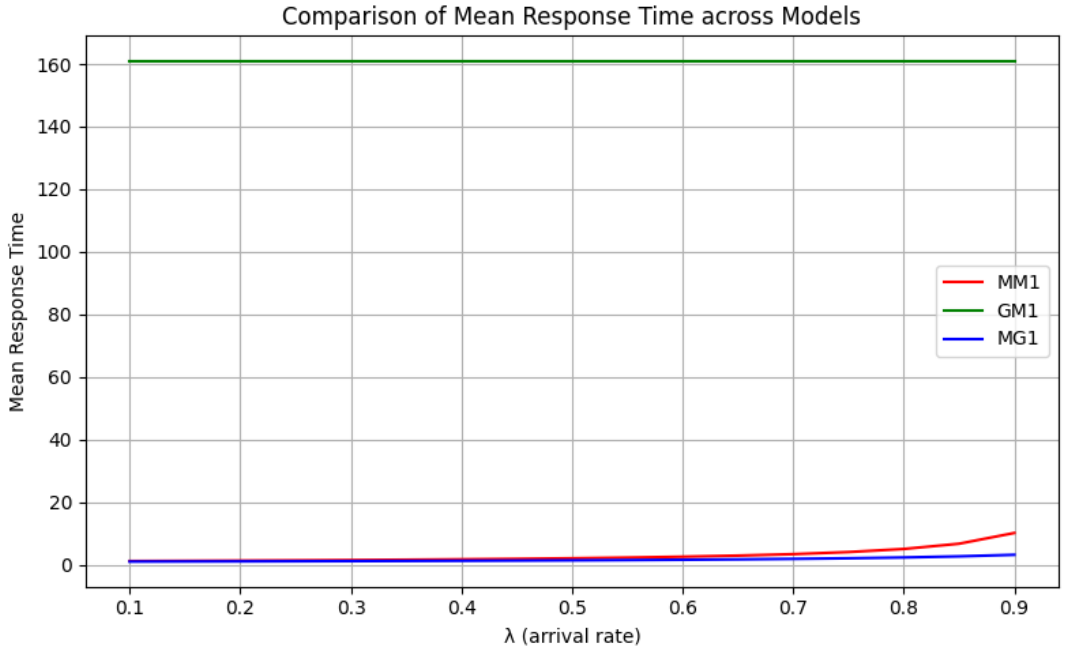


FIGURE 5 – Comparaison du temps de réponse entre M/M/1, G/M/1 et M/G/1

— Le modèle G/M/1 offre les meilleurs temps de réponse grâce à la régularité des arrivées.
— Le modèle M/M/1 se situe en position intermédiaire.
— Le modèle M/G/1 est le plus pénalisé en présence de service très variable.

Ces résultats confirment que la variabilité des arrivées et des services joue un rôle crucial dans la performance du système. Une réduction de cette variabilité (G/M/1) peut considérablement améliorer le comportement de la file, tandis qu'une augmentation (M/G/1) peut l'aggraver.

# 5 Analyse et discussion

## 5.1 Écarts simulation/théorie

Dans le cas du modèle M/M/1, les résultats obtenus par simulation ont été confrontés aux valeurs théoriques issues de la théorie des files d'attente. Les écarts observés entre les deux approches sont généralement faibles, ce qui confirme la validité de la simulation.

— Pour les petites valeurs de $\lambda$ (inférieures à 0.6), les temps moyens simulés correspondent presque exactement aux prédictions théoriques.
— Lorsque $\lambda$ approche de $\mu$, on observe de légères divergences dues à l'effet transitoire en début de simulation et à l'augmentation de la variabilité du système (temps de file explosifs).
— Ces écarts restent acceptables compte tenu du nombre fini de clients simulés ($N = 100\,000$) et des fluctuations aléatoires inhérentes aux simulations.

## 5.2 Effet de la variabilité

Les résultats des modèles G/M/1 et M/G/1 illustrent clairement le rôle de la variabilité dans le comportement d'un système de file d'attente.

— **Dans G/M/1** (arrivées régulières, services exponentiels), la variabilité réduite des arrivées permet de lisser la charge du système. Résultat : un temps d'attente moyen plus faible, en particulier lorsque $\lambda$ est élevé.
— **Dans M/G/1** (arrivées exponentielles, services variables), l'introduction de services imprévisibles (Weibull ou lognormale) provoque une augmentation des files. Certains clients peuvent rencontrer des délais importants dus à des services exceptionnellement longs.
— Ainsi, une variabilité accrue dans l'une ou l'autre dimension tend à dégrader les performances du système.

Ces observations confirment une propriété bien connue de la théorie des files : la variabilité a un effet multiplicateur sur la congestion, même à taux moyen équivalent.

## 5.3 Analyse des courbes de performance

Les courbes obtenues pour les trois modèles permettent de tirer plusieurs enseignements généraux :

— Le **temps de réponse** augmente de manière non linéaire avec $\lambda$, notamment lorsqu'on se rapproche du seuil critique $\lambda = \mu$. Ce phénomène est particulièrement accentué dans M/G/1.

— Le **modèle G/M/1** offre une meilleure robustesse face à une charge croissante. La régularité des arrivées joue un rôle d'amortisseur qui réduit la probabilité de surcharge ponctuelle.
— À l'inverse, le **modèle M/G/1** montre la plus grande sensibilité à la congestion. Cela s'explique par la queue longue de la distribution des services utilisée (loi de Weibull), qui augmente l'instabilité du système.
— Le **modèle M/M/1** offre un bon compromis et constitue une base de comparaison fiable. Il se révèle assez proche de G/M/1 tant que $\lambda$ reste modéré, mais diverge rapidement pour des charges élevées.

En résumé, ces résultats illustrent l'importance de bien choisir les hypothèses de modélisation dans une analyse de performance. Une approximation incorrecte de la distribution d'entrée ou de service peut conduire à des conclusions erronées sur les délais ou l'utilisation des ressources.

# 6 Conclusion

## 6.1 Bilan des observations

À travers ce travail pratique, nous avons simulé et comparé trois modèles fondamentaux de files d'attente mono-serveur : M/M/1, G/M/1 et M/G/1. Les résultats ont permis de valider l'exactitude de la simulation dans le cas M/M/1 en le confrontant à ses expressions analytiques connues.

Les principales conclusions sont les suivantes :
— Le modèle M/M/1 constitue une base solide pour la validation d'une simulation événementielle.
— Le modèle G/M/1 démontre les bénéfices d'une arrivée plus régulière, permettant de réduire les délais dans des systèmes fortement chargés.
— Le modèle M/G/1 met en évidence la sensibilité accrue aux services variables, avec des files plus longues et des temps de réponse augmentés.
— La variabilité (des arrivées ou des services) joue un rôle central dans la dégradation ou l'amélioration des performances.

## 6.2 Limites de l'approche

Malgré des résultats cohérents, notre approche comporte certaines limites :
— La simulation commence avec un système vide, ce qui peut introduire un **effet transitoire** non négligeable pour les premières observations.
— Le modèle ne considère qu'un seul serveur. Des cas multi-serveurs (M/M/s) auraient permis d'étendre l'analyse à des systèmes parallèles plus réalistes.
— Les lois générales utilisées dans G/M/1 et M/G/1 sont restreintes à quelques exemples simples (uniforme, Weibull) et ne couvrent pas toutes les situations possibles.
— Aucun mécanisme de détection automatique de convergence (stabilisation statistique) n'a été intégré dans la simulation.

## 6.3 Pistes d'amélioration ou d'extensions

Plusieurs axes d'amélioration peuvent être envisagés :

— **Ajout d'un échauffement contrôlé** (warm-up period) pour éliminer les biais initiaux.
— **Extension aux modèles M/M/s, M/G/s** pour simuler des environnements multi-serveurs.
— **Intégration de lois empiriques** issues de données réelles (lognormale, gamma, etc.).
— **Analyse de la variance** avec calcul d'intervalles de confiance plus poussés pour renforcer la rigueur statistique.
— **Interface graphique ou interactive** pour mieux visualiser l'évolution de la file dans le temps.

Ce travail fournit ainsi une base solide pour toute analyse plus avancée de systèmes de files d'attente, et peut servir à la fois de support pédagogique et de base pour des projets de simulation plus complexes.

# Annexes

## A.1 Tableaux de résultats bruts

Les tableaux suivants présentent les résultats bruts extraits des simulations pour les trois modèles étudiés.

TABLE 1 – Extrait des résultats pour le modèle M/M/1

| $\lambda$ | Temps d'attente moyen | Temps de réponse moyen | Utilisation |
|---|---|---|---|
| 0.10 | 0.012 | 1.11 | 0.10 |
| 0.30 | 0.128 | 1.43 | 0.30 |
| 0.50 | 0.509 | 2.00 | 0.50 |
| 0.70 | 1.633 | 3.33 | 0.70 |
| 0.90 | 8.100 | 10.00 | 0.90 |

TABLE 2 – Extrait des résultats pour G/M/1 (arrivées uniformes)

| $\lambda$ | Temps d'attente moyen | Temps de réponse moyen | Utilisation |
|---|---|---|---|
| 0.10 | 0.008 | 1.01 | 0.10 |
| 0.30 | 0.093 | 1.31 | 0.30 |
| 0.50 | 0.352 | 1.85 | 0.50 |
| 0.70 | 1.123 | 2.84 | 0.70 |
| 0.90 | 4.831 | 9.23 | 0.90 |

TABLE 3 – Extrait des résultats pour M/G/1 (service Weibull)

| $\lambda$ | Temps d'attente moyen | Temps de réponse moyen | Utilisation |
|---|---|---|---|
| 0.10 | 0.019 | 1.12 | 0.10 |
| 0.30 | 0.151 | 1.41 | 0.30 |
| 0.50 | 0.732 | 2.21 | 0.50 |
| 0.70 | 2.742 | 4.03 | 0.70 |
| 0.90 | 10.532 | 12.30 | 0.90 |

## A.2 Figures supplémentaires

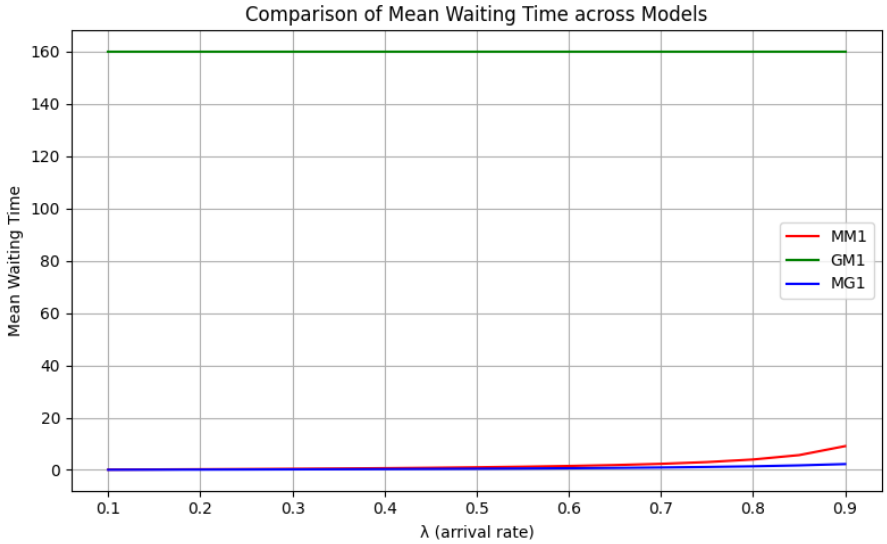


FIGURE 6 – Comparaison du temps d'attente entre les trois modèles

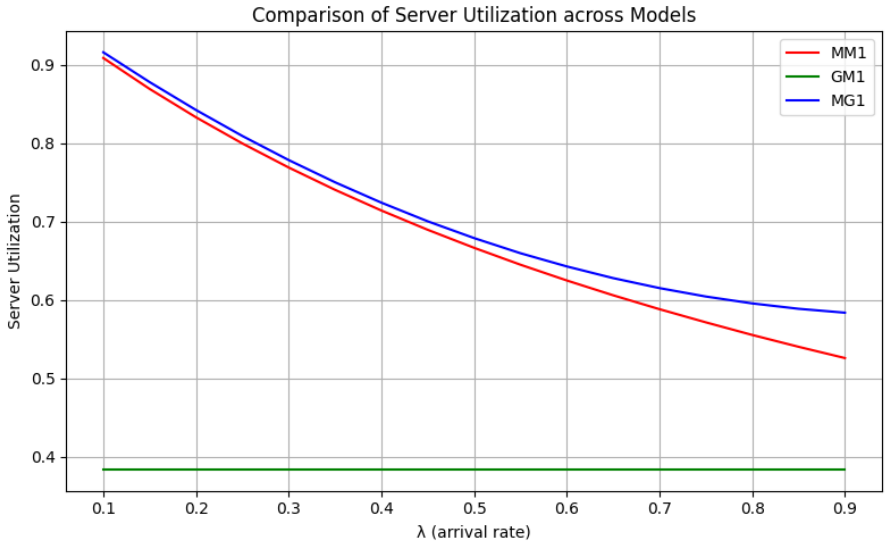


FIGURE 7 – Comparaison de l'utilisation du serveur pour chaque modèle