Université Abderrahmane-MIRA de Bejaia

Faculté des Sciences Exactes

Département d'Informatique

Module : Modélisation et Evaluation des Performances

Spécialité : Intelligence Artificielle

Chapitre IV : Files d'Attente

Présenté par : Prof. Mohand YAZID

PLAN DU CHAPITRE

PLAN DU CHAPITRE

PLAN DU CHAPITRE

PLAN DU CHAPITRE

PLAN DU CHAPITRE

PLAN DU CHAPITRE

PLAN DU CHAPITRE

PLAN DU CHAPITRE

PLAN DU CHAPITRE

PLAN DU CHAPITRE

PLAN DU CHAPITRE

PLAN DU CHAPITRE

PLAN DU CHAPITRE

1. Introduction

Le modèle général d'un système d'attente peut être résumé comme suit :

1. Des **clients** arrivent à un certain **endroit** et réclament un certain **service**.
2. Les **instants d'arrivées** et les **durées de service** sont généralement des **quantités aléatoires**.
3. Si un **poste** de service est **libre**, le client qui arrive se **dirige immédiatement** vers ce poste où il est servi.
4. Sinon, il prend **sa place** dans une **file d'attente** dans laquelle les clients **se rangent** suivant leur **ordre d'arrivée**.

1. Introduction

Un système d'attente comprend donc :

1. Un **espace de service** avec **une** ou **plusieurs** stations de service **montées en parallèle**.
2. Un **espace d'attente** dans lequel se forme une **éventuelle** file d'attente.

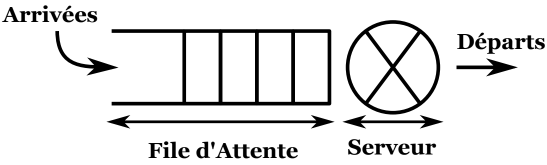

Figure: Représentation d'un système d'attente.

2. Classification des Systèmes d'Attente

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

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

2. Classification des Systèmes d'Attente

- Pour la classification des systèmes d'attente, on a recourt à la notation symbolique suivante **A/B/s/N**, où :
  - **A** : Distribution des temps entre deux arrivées successives,
  - **B** : Distribution des durées de service,
  - **s** : Nombre de postes de service en parallèle,
  - **N** : Capacité du système.
- Le dernier symbôle sera toutefois supprimé si $N = \infty$.
- Pour spécifier les distributions **A** et **B**, on introduit les symbôles suivants :
  - **M** : Distribution Exponentielle (qui vérifie la propriété de Markov).
  - E_k : Distribution d'Erlang d'ordre k.
  - **G** : Distribution générale.
  - **D** : Cas déterministe.

2. Classification des Systèmes d'Attente

- La notation **M/D/1/4** définit donc un système d'attente comprenant :
  - **Une seule** station de service,
  - La capacité de l'espace d'attente vaut $4 - 1 = 3$,
  - Le processus d'arrivée est **Poissonien**,
  - La durée de service est **constante**.
- En plus des notations introduite ci-dessus, on utilisera les grandeurs suivantes :
  - $1/\lambda$: Intervalle moyen séparant deux arrivées consécutives.
  - D'où, λ : Taux des arrivées.
  - $1/\mu$: Durée moyenne de service.
  - D'où, μ : Taux de service.

Un serveur web reçoit des requêtes HTTP en moyen toutes les **25 ms**. Les requêtes ont toutes la même priorité et sont donc traitées selon leurs ordres d'arrivée sur le serveur. La prise en charge d'une requête HTTP a une durée moyenne de **20 ms** (une seule requête peut être traitée à la fois). On suppose que les durées des inter-arrivées sont distribuées selon la loi exponentielle et les durées de traitement sont indépendantes et identiquement distribués selon la loi exponentielle. On suppose aussi que le serveur web possède une capacité suffisamment importante pour contenir toutes les requêtes des utilisateurs.

1. Proposer un système d'attente modélisant le serveur web.
2. Préciser le taux d'arrivée et le taux de service.

3. Caractéristiques d'un Système d'Attente

3.1. Analyse Mathématique

- L'étude mathématique d'un système d'attente se fait le plus souvent par l'introduction d'un processus stochastique.
- En premier lieu, on s'intéresse au nombre $X(t)$ de clients se trouvant dans le système à l'instant t $(t \geq 0)$, ensuite on cherche à calculer :
  - Les probabilités d'états $P_n(t) = P\left(X(t) = n\right)$ $(n = 0, 1, 2, \cdots)$ qui définissent le régime transitoire du processus stochastique $\{X(t), t \geq 0\}$.
  - Le régime stationnaire du processus stochastique, défini par :

$$P_n = lim_{t \to \infty} P_n(t) = P(X(\infty) = n). \quad (1)$$

- On écrira par la suite X au lieu de $X(\infty)$.
- A partir de la distribution stationnaire du processus $\{X(t), t \geq\}$, on pourra obtenir d'autres caractéristiques de performances du système d'attente :
  - Le nombre moyen de clients dans le système,
  - La durée moyenne d'attente d'un client,
  - La durée moyenne de séjour dans le système, etc.

- A partir de la distribution stationnaire du processus $\{X(t), t \geq\}$, on peut calculer les caractéristiques suivantes d'un système d'attente :
  - $L = E(X)$: Nombre moyen de clients dans le système.
  - L_q : Nombre moyen de clients dans la file.
  - W : Temps moyen de séjour d'un client dans le système.
  - W_q : Temps d'attente moyen d'un client.
- Ces caractéristiques permettent de juger le comportement opérationnel d'un système d'attente. Elles sont liées par les relations suivantes :
  - $L = \lambda \times W$.
  - $L_q = \lambda \times W_q$.
  - $W = W_q + 1/\mu$.
  - $L = L_q + \lambda/\mu$.
- Ces quatres relations sont valables dans des conditions assez générales, i.e., pour des systèmes d'attente du type $G/G/s/N$.
- Les deux premières relations sont connues sous le nom de formules de **Little**.

3. Caractéristiques d'un Système d'Attente

3.3. Caractéristiques du système M/M/1

$$L = E(X) = \sum_{n=0}^{\infty} n \times P_n. \tag{2}$$

$$P_n = (1 - \frac{\lambda}{\mu}) \times \left(\frac{\lambda}{\mu}\right)^n. \tag{3}$$

On note $\frac{\lambda}{\mu} = \rho$, ρ doit être inférieur strictement à 1 pour que la condition de stabilité du système M/M/1 soit vérifiée. $\rho < 1$, nous avons donc :

$$\begin{aligned} L &= \sum_{n=0}^{\infty} n \times (1-\rho) \times \rho^n, \\ &= (1-\rho) \times \sum_{n=0}^{\infty} n \times \rho^n, \\ &= \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1-\rho}. \end{aligned} \tag{4}$$

Soit X_q le nombre de clients se trouvant dans la file d'attente :

$$X_q = \begin{cases} 0 & X = 0, \\ X - 1 & X \geq 1. \end{cases} \tag{5}$$

$$\begin{aligned} L_q &= E(X_q), \\ &= \sum_{n=1}^{\infty} (n-1) \times P_n, \\ &= \frac{\lambda^2}{\mu \times (\mu - \lambda)} = \frac{\rho^2}{1-\rho}. \end{aligned} \tag{6}$$

On peut maintenant calculer W et W_q soit d'après les formules de Little, soit directement à partir de la distribution stationnaire du système :

$$W = \frac{1}{\mu - \lambda} = \frac{1}{\mu \times (1 - \rho)}. \tag{7}$$

$$W_q = \frac{\lambda}{\mu \times (\mu - \lambda)} = \frac{\rho}{\mu \times (1 - \rho)}. \tag{8}$$

3.4. Exercice d'Application 2

On considère un réseau de communication très simple permettant de transmettre des paquets de données d'un noeud A vers un noeud B avec une vitesse de 300 Mbits/minute. Les paquets de données sont générés par une source en A à un taux moyen de 240 Mbits/minute. On suppose que les paquets de données ont une taille fixe de 100 000 bits. On suppose aussi que les durées des inter-arrivées sont distribuées selon la loi exponentielle et les durées de transmission sont indépendantes et identiquement distribués selon la loi exponentielle.

1. Proposer un modèle de file d'attente correspondant au réseau de communication.
2. Préciser les paramètres du système d'attente.
3. Déduire la durée moyenne des inter-arrivées et la durée moyenne de transmission.
4. Vérifier la condition de stabilité du système.
5. Calculer les mesures de performances du réseau de communication :
   - Le nombre moyen de paquets dans le réseau de communication.
   - Le nombre moyen de paquets en attente de transmission.
   - La durée moyenne d'attente d'un paquet dans la file de transmission.
   - Le temps moyen de réponse d'un paquet.

4. Systèmes d'Attente Autres M/M/1

4.1. Système à Plusieurs Stations M/M/s

- Envisageons d'abord le cas d'un système d'attente comprenant *s* stations en parallèle.
- Nous admettons que les durées de service correspondantes suivent la **même distribution exponentielle** de paramètre μ.
- Nous admettons que les clients qui arrivent **forment une seule file d'attente**.

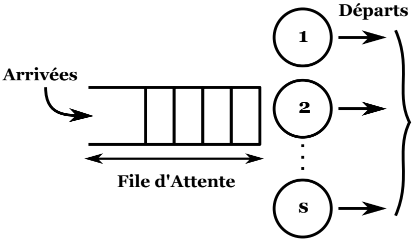

Figure: Représentation d'un système d'attente M/M/s.

4. Systèmes d'Attente Autres M/M/1

4.1. Système à Plusieurs Stations M/M/s

- On appelle :
  - $s \times \mu$: le taux de service global du système,
  - $\rho = \lambda / s\mu$: l'intensité global du trafic.
- La distribution stationnaire du système M/M/s :

$$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} \times P_0 & n \leq s, \\ \frac{(\lambda/\mu)^n}{s! \times s^{n-s}} \times P_0 & n \geq s. \end{cases} \tag{9}$$

- Où :

$$P_0 = \left[\sum_{n=0}^{s} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^{s+1}}{s! \times (s - \lambda/\mu)} \right]^{-1}. \tag{10}$$

- Ces relations sont valables pour $\rho = \lambda / s\mu < 1$.

- La probabilité q'un client qui entre dans le système doive attendre est alors donnée par :

$$P(attente) = P(X \geq s) = \sum_{n=s}^{\infty} P_n = \frac{P_s}{1-\rho}. \quad (11)$$

- A partir de la distribution stationnaire, on peut calculer les mesures usuelles suivantes :

$$\begin{aligned} L_q &= E(X_q) = \sum_{n=s+1}^{\infty} (n-s) \times P_n = \frac{\rho \times P_s}{(1-\rho)^2}, \\ &= \frac{\lambda^s \times \rho \times P_0}{\mu^s \times s! \times (1-\rho)^2}. \end{aligned} \quad (12)$$

$$L = L_q + \lambda/\mu = s \times \rho + \frac{\rho \times P_s}{(1-\rho)^2}. \quad (13)$$

- A l'aide des formules de little, on trouve ensuite les expressions de W et W_q.

4. Systèmes d'Attente Autres M/M/1

4.2. Exercice d'Application 3

On souhaite modéliser un ordinateur, équipé de deux CPUs qui traitent des programmes en parallèle, par un système d'attente simple. Les programmes arrivent suivant un processus de Poisson. La durée moyenne entre deux arrivées successives est de 1/20 secondes. La durée moyenne d'exécution d'un programme sur une CPU suit une loi exponentielle avec moyenne de 80 ms. Chaque CPU exécute un programme différent jusqu'à sa terminaison.

1. Proposer un système d'attente représentant l'ordinateur.
2. Préciser le taux d'arrivée et le taux de service du système.
3. Vérifier la condition de stabilité du système d'attente.
4. Calculer le nombre moyen de programmes qui sont en attente d'être exécutés.
5. Calculer le nombre moyen de programmes dans l'ordinateur.
6. Déduire la durée moyenne d'attente et la durée moyenne de séjour d'un programme.

4.3. Système à Pertes M/M/1/k

Nous avons considéré jusqu'ici uniquement des systèmes d'attente à capacité illimitée; on peut interpréter ces modèles comme des cas limites de systèmes réels qui comprennent un nombre fini de places d'attente.

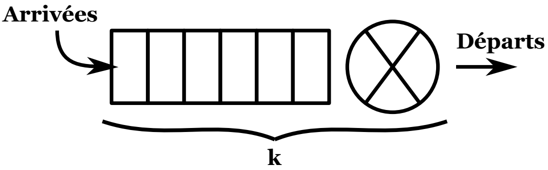

Figure: Représentation d'un système d'attente M/M/1/k.

$$\rho = \frac{\lambda}{\mu}. \tag{14}$$

$$P_n = \rho^n \times P_0. \tag{15}$$

4. Systèmes d'Attente Autres M/M/1

4.3. Système à Pertes M/M/1/k

$$P_0 = \frac{1-\rho}{1-\rho^{k+1}}. \tag{16}$$

$$L = \frac{\rho}{1-\rho} - (k+1) \times \frac{\rho^{k+1}}{1-\rho^{k+1}}. \tag{17}$$

$$W = \frac{L}{\lambda \times (1-P_k)}. \tag{18}$$

$$L_q = L - (1-P_0). \tag{19}$$

$$W_q = \frac{L_q}{\lambda(1-P_k)}. \tag{20}$$

On considère une mémoire pouvant contenir 4 fichiers. Les entrées des fichiers dans la mémoire se font selon un processus de Poisson de paramètre $\lambda = 2$ fichiers/seconde. Si la mémoire est pleine, les fichiers qui arriveront seront perdues. Par ailleurs, la prise de fichiers a lieu aussi selon un processus de Poisson de paramètre $\mu = 3$ fichiers/seconde.

1. Proposer un modèle de file d'attente correspondant à ce système.
2. Calculer la probabilité de perte d'un fichier.
3. Calculer le nombre moyen de fichiers dans la mémoire.
4. Calculer la durée moyenne d'attente d'un fichier dans la mémoire
5. Calculer la durée de séjour d'un fichier dans la mémoire.