Module : Evaluation de Performances Master 2 Informatique Haute Performance (MIHP)

Soraya Zertal

Li-PaRAD-Université de Versailles
soraya.zertal@uvsq.fr

Cours 3 : Modélisation mathématique

Introduction

Définition

La modélisation mathématique est un outil relativement simple, qui consiste en l'**abstraction** du système réel en un ensemble de **fonctions mathématiques**, représentant ses principales **fonctionnalités/services** afin d'analyser le comportement dudit système.

Formalismes de description

Le système peut être décrit par différents formalismes : files d'attente, chaines de Markov, réseaux de Petri...etc

Introduction aux files d'attente

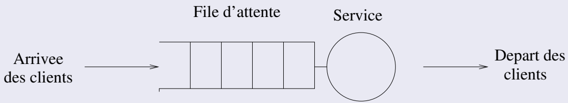
Files d'attente

Une théorie mathématique, basée sur **les probabilités** et utilisée pour l'évaluation des performances des systèmes où des **clients** peuvent arriver et demandent l'acquisition de **ressources** pour une certaine durée.
Certains clients doivent donc **patienter** avant d'**être servi** et quitter le système.
La file d'attente est alimentée par le flux d'arrivée des clients et vidée par les serveurs (ressources).

Introduction

Files d'attente

Théorie valable dans toutes les situations où le taux de service est inférieur à la demande dudit service (taux d'arrivée).
Situation dans laquelle le système n'est pas saturé.



Modélisation analytique : intérêt et limitation

Intérêt

La modélisation mathématique permet l'évaluation d'un nombre considérable de configurations possibles du système à analyser ainsi que des **charges** qu'il est **susceptible de supporter**.

Limitation

Les hypothèses et les approximations sur lesquelles se base la modélisation peuvent restreindre le champ d'application de ses résultats et impacter leur précision.

Plan

Modèles de files d'attente

Vocabulaire

Ressource (CPU,Mémoire,Disque...)$\equiv$ serveur
Job, Processus, IO req $\equiv$ client

Définition

Les clients attendent dans une file que le serveur soit libre pour les servir.
Le serveur sélectionne selon une stratégie définie, un client en attente dans la file à servir.

Modèles de files d'attente

Définition

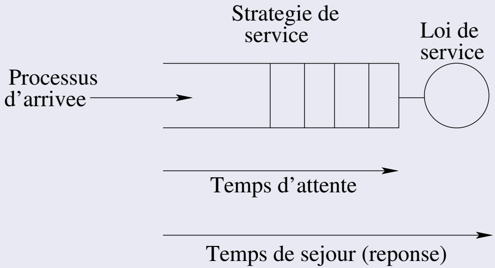


Figure : File d'attente élémentaire

Modèles de files d'attente

Stratégies de sélection

- **FIFO** premier arrrivé dans la file, le premier servi
- **LIFO** dernier arrivé dans la file, le premier servi
- **Random** aléatoire (au hasard)
- **Round Robin** cyclique
- **Time sharing (TS)** temps partagé, valable surtout pour la ressource processeur

Réseaux de files d'attente

Connecter les files d'attente ...

Les serveurs et les files peuvent être interconnectés pour former un réseau de files d'attente

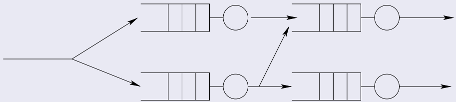
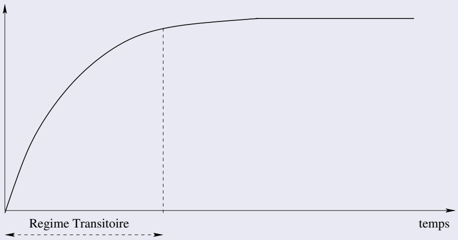


Figure : Réseau de files d'attente

Régime permanent et transitoire

Régime transitoire

Régime d'évolution d'un système qui n'a pas encore atteint son régime stable ou permanent.
Il peut apparaitre lors de l'amorce du système ou de sa modification.



Régime permanent et transitoire

Régime permanent

Régime de stabilité que le système atteint aprés un certain temps de fonctionnement. Il peut être continu ou périodique. Dans les deux cas, les valeurs atteintes par le système ne dépendent plus du temps.

Réseaux de files d'attente : paramètres

Paramètres de performance

Dans un réseau de files d'attente, ils se définissent aussi bien dans le régime **transitoire** que **permanent** par :

- Débit ou fréquence d'arrivée et de sortie
- Nombre moyen de clients dans le système
- Temps moyen de séjour
- Taux d'utilisation du serveur

Réseaux de files d'attente : stabilité

Condition de stabilité

La stabilité d'un système n'a de sens que pour le régime **permanent**

La stabilité est atteinte **si et seulement si**

$$\text{Le débit moyen en entrée} \equiv \text{Le débit moyen en sortie}$$

Réseaux de files d'attente

Topologies et classes

Selon leurs topologies, les réseaux de files d'attente peuvent être:

- Ouverts : la sortie du serveur n'est pas connectée à l'entrée de la file.
- Fermés : Il y a une boucle entre la file d'attente et le serveur.

Plan

Réseaux de files d'attente ouverts

Définition

Un réseau de files d'attente ouvert est un réseau où :

1. Les clients arrivent d'une source externe,
2. Sont servis par les différents serveurs,
3. Peuvent se terminer complètement ou bien génèrent de nouveaux clients entrants.

Réseaux de files d'attente ouverts

Remarques

- Tous les clients dans de tels réseaux se terminent et quittent le système un jour.
- Un nombre illimité de clients compose donc un flux entrant et sortant du système.

Réseaux de files d'attente ouverts

Exemple

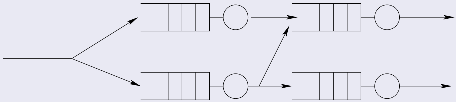


Figure : Réseau de files d'attente ouvert

Réseaux de files d'attente ouverts

Réseaux de jackson

- Mono-classe de clients
- Les arrivées Poissonniènes
- Un serveur par station
- File de capacité illimitée, Discipline FIFO
- Routage probabiliste

Réseaux de files d'attente ouverts

Exemple :

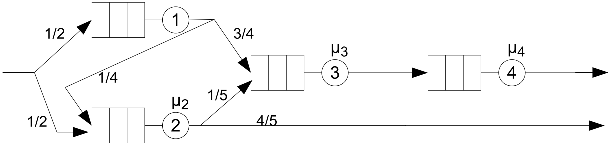


Figure : Réseau de Jackson

Calcul des taux d'arrivées :

$\lambda_1 = \frac{\lambda}{2}$, $\lambda_2 = \frac{5\lambda}{8}$, $\lambda_3 = \frac{\lambda}{2}$, $\lambda_4 = \frac{\lambda}{2}$

Plan

Réseaux de files d'attente fermés

Définition

Un réseau de files d'attente fermé a un **nombre fixe de clients** qui circulent à travers les différents serveurs et files d'attente.
La sortie est directement connectée à l'entrée du réseau. donc, il n y a ni arrivée et ni départ dans ce réseau.

Réseaux de files d'attente fermés

Remarques

- Ces modèles de réseaux fermés représentent avec plus de précision la réalité et les capacités limitées des buffers dédiés au stockage des clients.
- La résolution de tels modèles reste plus complexe.

Réseaux de files d'attente fermés

Exemple

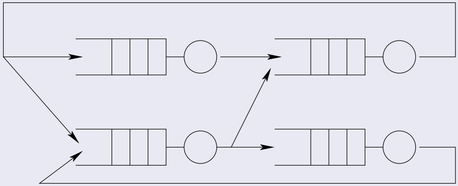


Figure : Réseau de files d'attente fermé

Plan

Hypothèses

Les hypothèses

A considérer avant la modélisation sont :

- Equilibre du flux des clients
  Le nombre de clients sortants doit être équivalent à celui des clients entrants en considérant une observation à long terme.
- Comportement à une étape
  A chaque moment, seul un client peut entrer ou sortir du système. L'état du système change donc de manière incrémentale

Hypothèses

Remarques

Concernant le comportement à chaque étape :

- Les arrivées ainsi que les départs multiples ne sont pas autorisés.
- Les mouvements simultanés des clients entre les files d'attente au sein du même système sont non autorisés.

Hypothèses

Les hypothèses

- Homogénéité
  Le taux moyen d'arrivée et le taux moyen de service sont indépendants de l'état du système.
- Exclusivité
  Un client peut se trouver au niveau d'un **seul serveur** (en service ou en attente).
  Un client ne peut pas demander deux services simultanément et quand il obtient un, il est le seul à être servi par ce serveur.

Hypothèses

Les hypothèses

- Service non-bloquant
  Le serveur est seul maître du service fourni.
  Celui-ci ne peut être contrôlé par aucun composant du système.

Plan

Notation

Dans le reste du cours, on utilisera la notation suivante:

Symbole	Commentaire
s	temps de service moyen pour un client
μ	taux de service $(1/s)$
λ	taux d'arrivée moyen
ρ	l'intensité du trafic (λ/μ)
r	temps de réponse moyen
w	temps d'attente moyen
q	nombre moyen de clients dans la file d'attente
n	nombre de clients dans le système (en attente + en service)
U	fraction d'utilisation du système
a	nombre d'arrivées pendant un temps d'observation T
d	nombre de départs pendant un temps d'observation T

Plan

L'analyse opérationnelle

Vue globale

L'analyse débute en considérant **le système à analyser comme une boite noire** à l'entrée de laquelle arrivent des jobs (clients) à traiter (servir) et en sortie de laquelle figure des jobs traités.

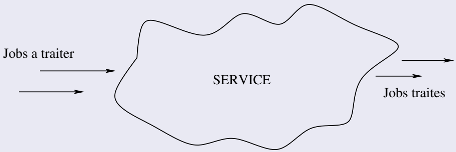


Figure : Vue globale de l'analyse opérationnelle

L'analyse opérationnelle

Caractérisation

- Les caractéristiques du système sont **évalués** par mesure ou par hypothèse (utilisateur,fabriquant/constructeur).
- Elles serviront à fixer des valeurs pour **les paramètres des lois mathématiques**.
- L'évaluation du comportement du système pourra alors s'effectuer dans ce contexte précis.

Lois d'utilisation

Taux d'arrivée

On observe le système pendant une période T, durant laquelle a clients sont arrivés,

on en déduis le taux d'arrivée λ :

$$\lambda = \frac{a}{T}$$

Lois d'utilisation

Utilisation du serveur

Si le serveur a été occupé pendant seulement b unités de temps durant toute la période d'observation T,
On en déduis l'utilisation du serveur U :

$$U = \frac{b}{T} = \frac{b}{d} \times \frac{d}{T}$$

d étant le nombre de clients servis,
$\frac{b}{d}$ est le temps moyen pour servir un client, et
$\frac{d}{T}$ le taux de départ des clients.

Lois d'utilisation

Utilisation du serveur

Selon l'hypothèse d'équilibre, le flux en entrée doit être équivalent à celui en sortie, d'où :

$$U = \frac{b}{d} \times \frac{d}{T} = s \times \frac{d}{T}$$

$$U = \lambda \times s$$

L'analyse opérationnelle

L'intensité du trafic

$$\rho = \frac{\lambda}{\mu}$$

- $\rho > 1$: signifiant que le taux d'arrivée est supérieur à celui du départ.
  Le nombre de clients en attente dans la file va augmenter et engendrer des temps d'attente infinis.
  $\Rightarrow$ Le système n'est pas stable

L'analyse opérationnelle

L'intensité du trafic

$$\rho = \frac{\lambda}{\mu}$$

- $\rho < 1$: siginifiant que le système est dans un état stable.
  $\Rightarrow$ L'utilisation U n'excède jamais les 100%.

$(1 - \rho)$: la probabilité que le serveur soit inactif.

L'analyse opérationnelle

Loi de Little

- Une loi générale qui peut s'appliquer à tout système.
- Aucune hypothèse sur la "boite noire"
- Aucune hypothèse sur les variables aléatoires qui caractérisent le système.
- La seule condition est la stabilité $\Rightarrow$ la satisfaction de l'hypothèse de **l'équilibre des flux** et donc ne concerne que le régime permanent.

L'analyse opérationnelle

Loi de Little

Le nombre moyen de clients (N), le temps moyen de séjour ou (t) et le débit moyen λ d'un système en régime permanent se relient selon la loi :

$$N = t \times \lambda$$

L'analyse opérationnelle

Loi de Little : importance

Permet de calculer l'un des trois paramètres (N,t,λ) en fonction des 2 autres.
Peut s'appliquer à n'importe quel système :

- Un buffer uniquement
- Un buffer + serveur
- Le serveur de la file

L'analyse opérationnelle

Loi de Little : importance

- **Un buffer uniquement**
  N représente le nombre moyen de clients dans la file d'attente, t représente le temps moyen d'attente avant le service et λ le flux d'arrivée (ou de départ) des clients.
- **Un buffer + serveur**
  N représente le nombre moyen de clients dans le système (dans la file ou en service), t représente le temps moyen de séjour (attente et service) et λ le flux d'arrivée.
- **Le serveur de la file**
  N représente le nombre moyen de clients en service, t représente le temps moyen de service et λ le flux d'arrivée.

Plan

L'analyse stochastique

Limite des lois d'utilisation et de Little

Elles ne considèrent aucune hypothèse concernant les lois de probabilité des inter-arrivées et du service.
Elles sont complètement indépendantes des distributions

Intérêt de l'analyse stochastique

La connaissance des distributions permet d'effectuer une analyse stochastique et d'affiner ainsi l'analyse du système en fournissant des résultats plus détaillés que ceux de l'analyse opérationnelle.

L'analyse stochastique

Définition

Avec l'analyse stochastique, le système n'est plus une boite noire. Ce qui permet d'avoir des résultats bien détaillés sur le système analysé, tel que :

- temps d'attente moyen pour un client,
- le nombre moyen de clients dans la file,
- la probabilité d'avoir un certain nombre de clients en attente,
- le temps de réponse moyen...

L'analyse stochastique

Notation de Kendall

Le système est composé d'une file d'attente et d'un ou plusieurs serveurs. Il sera représenté par le mot $A/S/C/B/N/D$ avec :

- A : La distribution du processus d'arrivée, symbolisée par M pour Exponentielle (cas typique), D pour constante, E_k pour Erlang, G pour général.
- S : La distribution du service, utilisant des lois similaires à celles pour le processus d'entrée.

L'analyse stochastique

Notation de Kendall

- C : le nombre de serveurs identiques dans le système. Si un système contient des serveurs différents, il faut le décomposer en sous systèmes ayant chacun des serveurs identiques.
- B : Le nombre total de clients dans le système (sa capacité), supposé infini pour simplifier l'analyse.
- N : le nombre de clients pouvant entrer au système, supposé infini pour simplifier l'analyse.
- D : L'ordre ou la stratégie de service: FIFO, LIFO...etc

L'analyse stochastique

Cas M/M/1

C'est le cas le plus simple :

- L'arrivée est Poissonienne de taux λ,
- Le service exponentiel de taux μ,
- Il y a un seul serveur,
- Capacité infinie de la file (par défaut),
- Nombre de clients pouvant entrer dans le système est infini (par défaut) et
- Stratégie de service FIFO (par défaut).

L'analyse stochastique

Cas M/M/1

Les paramètres de performance de la file M/M/1 en régime permanent (stable) sont :

- Débit La file est stable, donc il y a un équilibre entre les flux, d'ou :

$$Debit = \lambda$$

- Taux d'utilisation du serveur C'est la probabilité que le serveur soit occupé :

$$U = \frac{\lambda}{\mu} = \rho$$

L'analyse stochastique

Cas M/M/1

- Probabilité que le système soit vide

$$P_0 = 1 - \rho$$

- Probabilité qu'il y ait attente dans le système

$$P_a = \rho$$

L'analyse stochastique

Cas M/M/1

- Nombre moyen de clients dans le système :

$$n = \frac{\rho}{(1-\rho)}$$

- Temps moyen de séjour se déduit de la loi de Little :

$$r = \frac{n}{\lambda} = \frac{1}{\mu(1-\rho)}$$

et peut se décomposer en :

$$\frac{1}{\mu} + \frac{\rho}{\mu(1-\rho)}$$

L'analyse stochastique

Cas M/M/1

- Temps passé dans la file d'attente en découle :

$$w = \frac{\rho}{\mu(1-\rho)}$$

- Nombre moyen de clients dans la file

$$q = w \times \lambda = \frac{\rho^2}{(1-\rho)}$$

L'analyse stochastique

Cas M/M/1

En régime stationnaire :

- La probabilité d'avoir 0 clients dans le système est :
$$P_0 = (1 - \rho)$$
- La probabilité d'avoir n clients dans le système est :
$$P_n = \rho^n P_0$$

L'analyse stochastique: cas M/M/C

Cas M/M/C

Le cas précédent avec C serveurs de même type et indépendants (processeurs ou disques).
Les jobs/requêtes IO attendent d'être servis dans la même file

Si un des C serveurs est disponible, il sert un des jobs/requêtes IO de la file.

L'analyse stochastique: cas M/M/C

Cas M/M/C

Condition de stabilité

Pour atteindre la stabilité, il faut que le nombre moyen de clients qui arrivent à la file par unité de temps, soit inférieur au nombre de clients que les C serveurs sont capables de traiter par unité de temps.

D'où :

$$\lambda < C\mu$$

L'analyse stochastique: cas M/M/C

Cas M/M/C

Les paramètres de performance de la file M/M/C en régime permanent (stable) sont :

- Débit La file est stable, donc débit moyen d'entrée est égal au débit moyen de sortie. d'ou :

$$Debit = \lambda$$

- Intensité du trafic

$$\rho = \frac{\lambda}{C\mu}$$

L'analyse stochastique

Cas M/M/C

En régime stationnaire :

- La probabilité que le système soit vide :

$$P_0 = \frac{1}{\Sigma_{k=0}^{C-1} \frac{\rho^k}{k!} + \frac{\rho^C}{C!} \frac{1}{1-\rho/C}}$$

- La probabilité qu'il y ait attente dans le système :

$$P_a = P_0 \frac{\rho^C}{(C-1)!(C-\rho)}$$

L'analyse stochastique

Cas M/M/C

- Nombre moyen de clients :

$$n = \rho(1 + \frac{P_a}{(C-\rho)})$$

- Temps moyen de séjour en découle par application de la loi de Little :

$$r = \frac{n}{\lambda} = \frac{1}{\mu}(1 + \frac{P_a}{(C-\rho)})$$

  et peut se décomposer en :

$$r = \frac{1}{\mu} + \frac{P_a}{\mu(C-\rho)}$$

L'analyse stochastique: cas M/M/C

Cas M/M/C

- Temps moyen d'attente en est déduit :

$$w = \frac{P_a}{\mu(C-\rho)}$$

$$w = \frac{\rho^C}{\mu(C-1)!(C-\rho)^2} P_0$$