

# Files d'attente

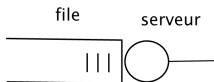
Enseignant : Dr Stephan Robert

Décembre 2019

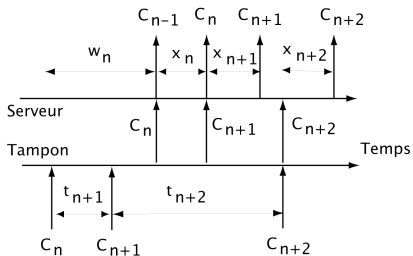


# Introduction

## Représentation d'une **file d'attente** simple

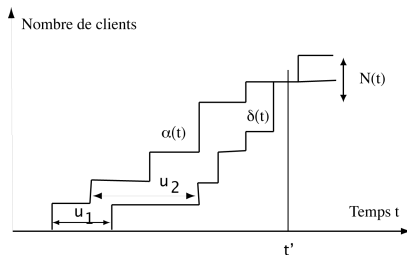


## Diagramme temporel



## Introduction (2)

**Question** : Quelle est la relation entre le **temps d'attente moyen** et le **nombre de clients** dans le système ?



$$\hat{T}_t = \frac{\sum_{i=0}^{\alpha(t)} u_i}{\alpha(t)}$$

**Nombre moyen de clients** dans la file d'attente :

$$\hat{N}_t = \frac{\int_0^t N(\tau) d\tau}{t} = \frac{\sum_{i=0}^{\alpha(t)} u_i}{t}$$

ce qui signifie

$$\hat{N}_t = \hat{\lambda}_t \hat{T}_t$$

Si le système est ergodique, nous obtenons la  
**LOI DE LITTLE :**

$$E[M] = \lambda E[T]$$

Fonctions de répartition

- ▶ Interarrivées :  $A(t)$
- ▶ Temps de service :  $B(x)$

Caractérisation d'une file d'attente (**Notation de Kendall**)

A/B/s/K/DS

Exemples : M/M/1, G/G/3/K

# Exercice 1

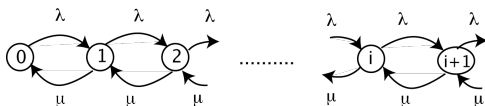


# Files d'attente élémentaires

## Généralités

- ▶ **Temps d'interarrivées** :  $E[T] = 1/\lambda$ 
  - ▶  $\lambda$  : Taux d'arrivées moyen.
- ▶ **Temps de service** :  $E[S] = 1/\mu$ 
  - ▶  $\mu$  : Taux de service moyen par serveur.
- ▶  $N(t)$  : Nombre de clients dans le système au temps  $t$ .

Représentation de la file d'attente M/M/1 (Processus de naissance et de mort) :



avec les paramètres suivants :

$$\lambda_k = \lambda \quad k = 0, 1, 2, 3, \dots$$

$$\mu_k = \mu \quad k = 1, 2, 3, \dots$$

## Remarques :

- ▶ Les interarrivées sont distribuées exponentiellement ( $E[T] = 1/\lambda$ )
- ▶ Les temps de service sont distribués exponentiellement ( $E[S] = 1/\mu$ )
- ▶ La file ne comporte qu'un serveur et est de longueur infinie
- ▶ Discipline de service : FIFO.
- ▶ **Intensité du trafic** :  $\rho = \lambda/\mu$
- ▶ La chaîne de Markov est ergodique si  $\rho < 1$

Probabilité d'état  $\pi_0$  (voir le cours) :

$$\pi_i = \pi_0 \left( \frac{\lambda}{\mu} \right)^i = \pi_0 \rho^i$$

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^i}$$

mais

$$\pi_0 = \frac{1}{1 + \frac{\lambda/\mu}{1-\lambda/\mu}}$$

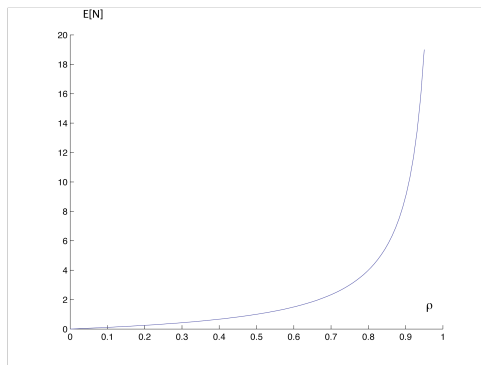
Ainsi

$$\pi_0 = 1 - \rho$$

## Caractéristiques de la file M/M/1

- ▶ Espérance mathématique du nombre de clients dans la file

$$E[M] = \sum_{i=0}^{\infty} ip_i = (1 - \rho) \sum_{i=0}^{\infty} i\rho^i = \dots = \frac{\rho}{1 - \rho}$$



## Caractéristiques de la file M/M/1 (suite)

- ▶ Avec la formule de Little :

$$E[T] = \frac{E[N]}{\lambda} = \frac{1}{\mu(1 - \rho)}$$

- ▶ Variance du nombre de clients en fonction de l'intensité du trafic :

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

Une entreprise effectue des mesures sur son serveur de base de donnée :

- ▶ taux moyen d'arrivées : 30 requêtes/seconde
- ▶ temps moyen de service : 20 ms

**Question** : A partir de quelle charge faut-il augmenter la vitesse du processeur pour maintenir un **service de même qualité** ? Si la firme augmente la charge du serveur de 40% par exemple, de combien doit augmenter la vitesse du processeur ?



## Serveur de bases de données (2)

**Réponse** : Modèle : M/M/1.

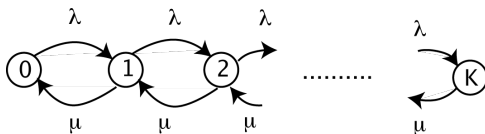
- ▶ Taux de traitement de requêtes :  $\mu = \frac{1}{20 \cdot 10^{-3}} = 50$  requêtes/seconde
- ▶ Taux d'utilisation du serveur :  $\rho = \lambda/\mu = 30/50 = 0.6 = 60\%$
- ▶ Temps de traitement (à garder constant !) :  $E[T] = \frac{1/\mu}{1-\rho} = \frac{1/50}{1-0.6} = 50$  ms

Le nombre de requêtes s'accroît de 40% :

- ▶  $\lambda' = 30 + 30 \cdot 0.4 = 42$
- ▶  $\rho' = \lambda'/\mu = 42/50 = 0.84 = 84\%$

Si nous voulons que  $E[T] = 50$  ms =  $\frac{1/\mu'}{1-\rho'}$ ,  $\mu' = 62$  requêtes/seconde. Donc le processeur devra **augmenter sa vitesse** de  $(62 - 50)/50 = 0.24 = 24\%$

Différence avec la file M/M/1 : Limitation du nombre de places dans le système (serveur + tampon).



Nous connaissons les probabilités d'état en fonction de  $\pi_0$  (processus de naissance et de mort) :

$$\pi_k = \pi_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu} = \pi_0 \left( \frac{\lambda}{\mu} \right)^k \quad \text{pour } k \leq K$$

D'autre part,

$$\pi_0 + \pi_1 + \dots + \pi_K = 1$$

ce qui donne

$$\pi_0 = \left[ 1 + \sum_{k=1}^K \left( \frac{\lambda}{\mu} \right)^k \right]^{-1} = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}}$$

Espérance mathématique du nombre de clients dans le système

$$E[M] = L = \left( \frac{\lambda}{\mu} \right) \frac{1 - (K+1)(\lambda/\mu)^K + K(\lambda/\mu)^{K+1}}{(1 - \lambda/\mu)(1 - (\lambda/\mu)^{K+1})}$$

Taux moyen d'arrivées :  $(1 - \pi_K)\lambda$ .

Temps moyen d'attente :

$$E[T] = \frac{E[M]}{(1 - \pi_K)\lambda} = \frac{L}{(1 - \pi_K)\lambda}$$

# Exemple : Dimensionnement d'un site Web

Quelle place mémoire faut-il allouer dans un serveur Web lorsque le nombre de requêtes/seconde =  $\lambda$  et que nous ne voulons pas avoir des pertes supérieures à 1% ?

- ▶ Taux moyen d'arrivées :  $\lambda = 30$  requêtes/seconde
- ▶ Taux de traitement des requêtes :  $\mu = 50$  requêtes/seconde

Nous avons

$$\pi_K = \pi_0 \left( \frac{\lambda}{\mu} \right)^K$$

avec

$$\pi_0 = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{K+1}}$$

donc

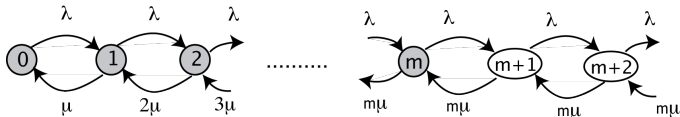
$$\pi_K = \frac{1 - 0.6}{1 - (0.6)^{K+1}} (0.6)^K < 0.01$$

# Exercices 4 et 5

# Applications aux dimensionnement de réseaux téléphoniques et cellulaires

# File M/M/m (Erlang C)

Système de file d'attente ayant un nombre illimité de places, avec  $m$  serveurs.



Probabilité d'état du  $k^{\text{ième}}$  état :

$$\pi_k = \pi_0 \frac{\lambda_0 \lambda_1 \dots \lambda_{k-1}}{\mu_1 \mu_2 \dots \mu_k}$$

En remplaçant ( $k \leq m$ ) :

$$\pi_k = \pi_0 \frac{\lambda \dots \lambda}{\mu(2\mu)(3\mu) \dots (k\mu)} = \frac{\pi_0}{k!} \left( \frac{\lambda}{\mu} \right)^k$$

Et pour  $k > m$

$$\begin{aligned}\pi_k &= \pi_m \prod_{i=m}^{k-1} \frac{\lambda}{m\mu} \\ &= \pi_m \frac{\lambda^{k-m}}{(m\mu)^{k-m}} \\ &= \frac{\pi_0 \lambda^m}{m! \mu^m} \frac{\lambda^{k-m}}{(m\mu)^{k-m}} \\ &= \frac{\pi_0}{m! m^{k-m}} \left( \frac{\lambda}{\mu} \right)^k\end{aligned}$$



## File M/M/m (Erlang C) (3)

Il reste à calculer  $\pi_0$  :

$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \frac{\lambda_0 \lambda_1 \dots \lambda_{i-1}}{\mu_1 \mu_2 \dots \mu_i}}$$

et après quelques calculs (cours) :

$$\pi_0 = \left( \sum_{k=0}^{m-1} \frac{(\lambda/\mu)^k}{k!} + \frac{(\lambda/\mu)^m}{m!(1 - \lambda/(m\mu))} \right)^{-1}$$

Question intéressante : Quelle est la probabilité d'attendre avant d'être servi ?

$$\sum_{k=m}^{\infty} \pi_k = \frac{\pi_0}{m!} \left( \frac{\lambda/\mu}{1 - \lambda/(m\mu)} \right)$$

qui est la **Formule d'Erlang C**. (En téléphonie : probabilité d'être mis en attente alors que toutes les lignes sont occupées).

## Réseaux 4G et 5G :

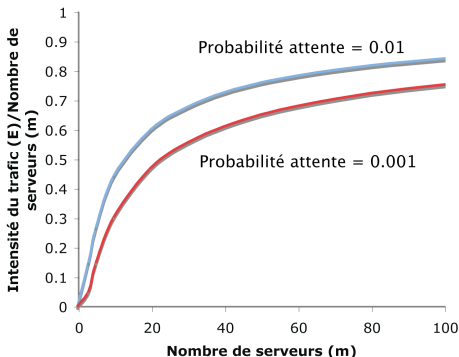
- ▶ Arrivées Poissonniennes (hypothèse valide dans le cas de la téléphonie)
- ▶ Temps de service distribués exponentiellement.
- ▶ Intensité du trafic :  $\rho = \lambda/\mu$ . 1 **Erlang** = occupation d'une ligne en permanence.

## Question :

Comment dimensionner le système de telle sorte à ce que la probabilité d'être mis en attente ne dépasse pas un certain seuil (par exemple 0.01) pour un trafic estimé ?

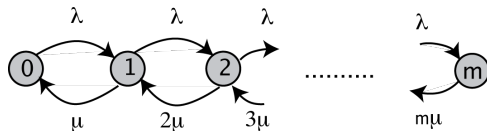
# Dimensionnement de réseaux téléphoniques et cellulaires (2)

Effet du nombre de serveurs (lignes) sur l'intensité du trafic avec une probabilité d'attente donnée



# File M/M/m/m (Erlang B)

Système de file d'attente avec  $m$  serveur, sans tampon ! Si tous les serveurs sont occupés, les clients sont rejetés !



Probabilité d'état du  $k^{\text{ième}}$  état :

$$\pi_k = \pi_0 \frac{\lambda_0 \lambda_1 \dots \lambda_{k-1}}{\mu_1 \mu_2 \dots \mu_k}$$

En remplaçant ( $k \leq m$ ) :

$$\pi_k = \pi_0 \frac{\lambda \dots \lambda}{\mu(2\mu)(3\mu) \dots (k\mu)} = \frac{\pi_0}{k!} \left( \frac{\lambda}{\mu} \right)^k$$

# File M/M/m/m (Erlang B)

Calcul de  $\pi_0$  avec  $\pi_0 + \pi_1 + \dots + \pi_m = 1$

$$\pi_0 = \left( \sum_{k=0}^m \left( \frac{\lambda}{\mu} \right)^k \frac{1}{k!} \right)^{-1}$$

La formule exprimant  $\pi_m$  est appelée “formule d'**Erlang B**”.

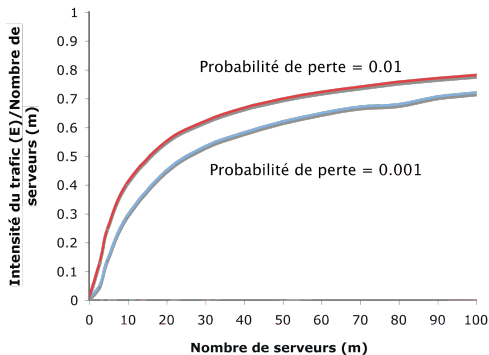
$$\pi_m = \frac{\pi_0}{m!} \left( \frac{\lambda}{\mu} \right)^m$$

Cette formule exprime la probabilité de voir tous les serveurs occupés alors qu'un nouveau client arrive dans le système. Un client perdu n'insiste pas et abandonne.

**Calculs** : <http://www.erlang.com/calculator>

# File M/M/m/m (Erlang B) (2)

Effet du nombre de serveurs (lignes) sur l'intensité du trafic avec une probabilité de perte donnée



# Exemple de dimensionnement de réseaux cellulaires

Dans le contexte de la téléphonie cellulaire :

Probabilité de blocage = **Grade of Service (GOS)**. Donnée par la formule d'Erlang B.

## Exemple dans une ville de 150'000 habitants (Lausanne)

- ▶ Opérateur A : 185 cellules avec 11 canaux chacune
- ▶ Opérateur B : 48 cellules avec 27 canaux chacune
- ▶ Opérateur C : 24 cellules avec 50 canaux chacune
- ▶ GOS=2%
- ▶ Chaque utilisateur fait en moyenne 2 appels/heure, de 3 minutes chacun.

**Question** : Quelle est la pénétration de marché de chacun ?

Trafic par utilisateur : 2 appels/heure \* (3/60) heures = 0.1 Erlang

# Exemple de dimensionnement de réseaux cellulaires

- ▶ L'opérateur A peut écouler 5.8 E sur 11 canaux et un GOS de 0.02, donc  $5.8/0.1=58$  utilisateurs/cellule. Puisqu'il en a 185, il peut servir  $185*58=10'730$  utilisateurs.
- ▶ L'opérateur B peut écouler 19.25 E sur 27 canaux, au total 192 utilisateurs/cellule. Donc il peut servir 9240 utilisateurs.
- ▶ Quant à l'opérateur C il peut écouler 40.25 E sur 50 canaux, donc 402 utilisateurs/cellule. En tout il peut servir 9660 utilisateurs. En tout 29'630 utilisateurs peuvent être servis.

La pénétration des systèmes mobiles est de 19% :

- ▶ 7.1% pour l'opérateur A
- ▶ 6.1% pour l'opérateur B
- ▶ 6.4% pour l'opérateur C



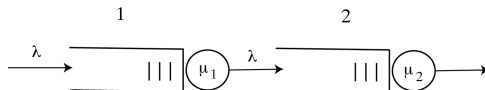
**Confirmer les résultats obtenus avec  
<http://www.erlang.com/calculator>**

# Réseaux de files d'attente

**Cas général** : Très compliqué.

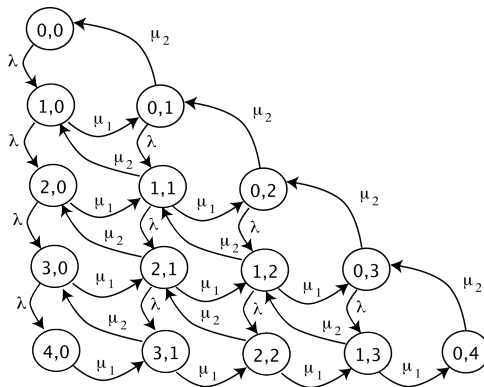
**Hypothèses** :

- ▶ Les flux entrant dans le réseau sont Poissoniens
- ▶ Les temps de service de tous les serveurs sont exponentiellement distribués
- ▶ Discipline de service : FIFO
- ▶ Routage entre les différentes files : Probabiliste



# Réseaux de files d'attente (2)

Chaîne de Markov pour deux files d'attente en série avec les états  $(n_1, n_2)$ ,  $n_1 \leq 4$ ,  $n_2 \leq 4$



Equations de balance :

$$\lambda\pi_{0,0} = \mu_2\pi_{0,1}$$

$$\mu_2\pi_{0,n_2} + \lambda\pi_{0,n_2} = \mu_1\pi_{1,n_2-1} + \mu_2\pi_{0,n_2+1}$$

$$\mu_1\pi_{n_1,0} + \lambda\pi_{n_1,0} = \lambda\pi_{n_1-1,0} + \mu_2\pi_{n_1,1}$$

$$\lambda\pi_{n_1,n_2} + \mu_1\pi_{n_1,n_2} + \mu_2\pi_{n_1,n_2} =$$

$$\lambda\pi_{n_1-1,n_2} + \mu_1\pi_{n_1+1,n_2-1} + \mu_2\pi_{n_1,n_2+1}$$

$$n_1 = n_2 = 0$$

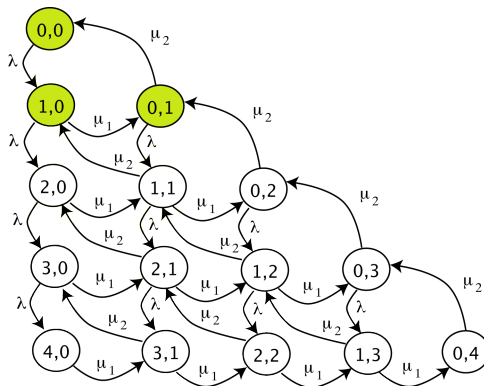
$$n_1 = 0, n_2 > 0$$

$$n_1 > 0, n_2 = 0$$

$$n_1, n_2 > 0$$

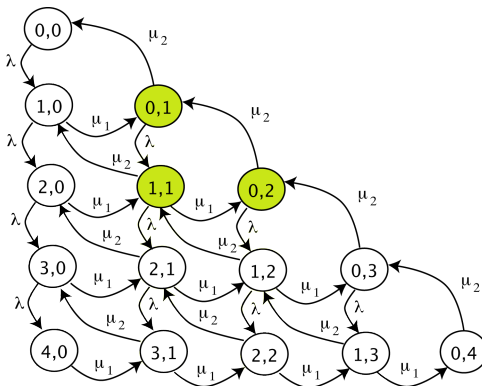
# Réseaux de files d'attente (4)

Chaîne de Markov pour deux files d'attente en série avec les états  $(n_1, n_2)$ ,  $n_1 \leq 4$ ,  $n_2 \leq 4$



# Réseaux de files d'attente (5)

Chaîne de Markov pour deux files d'attente en série avec les états  $(n_1, n_2)$ ,  $n_1 \leq 4$ ,  $n_2 \leq 4$ .



Nous n'avons qu'une équation pour les décrire

$$\mu_2 \pi_{0,n_2} + \lambda \pi_{0,n_2} = \mu_1 \pi_{1,n_2-1} + \mu_2 \pi_{0,n_2+1}$$

mais on peut la décomposer en deux équations de balance locales :

$$\mu_2 \pi_{0,n_2} = \mu_1 \pi_{1,n_2-1}$$

$$\lambda \pi_{0,n_2} = \mu_2 \pi_{0,n_2+1}$$

Nous pouvons procéder de manière identique pour l'équation suivante :

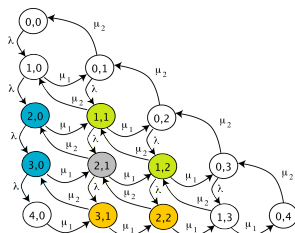
$$\mu_1 \pi_{n_1,0} = \lambda \pi_{n_1-1,0}$$

$$\lambda \pi_{n_1,0} = \mu_2 \pi_{n_1,1}$$



# Réseaux de files d'attente (7)

Chaîne de Markov pour deux files d'attente en série avec les états  $(n_1, n_2)$



Il nous reste à écrire les équations de balance locales pour les états  $\pi_{n_1, n_2}$ . En observant les cycles sur la figure il vient

$$\lambda \pi_{n_1, n_2} = \mu_2 \pi_{n_1, n_2 + 1}$$

$$\mu_1 \pi_{n_1, n_2} = \lambda \pi_{n_1 - 1, n_2}$$

$$\mu_2 \pi_{n_1, n_2} = \mu_1 \pi_{n_1 + 1, n_2 - 1}$$

Solution :

$$\pi_{n_1, n_2} = \left(1 - \frac{\lambda}{\mu_1}\right) \left(\frac{\lambda}{\mu_1}\right)^{n_1} \left(1 - \frac{\lambda}{\mu_2}\right) \left(\frac{\lambda}{\mu_2}\right)^{n_2}$$

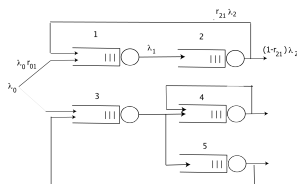
que nous pouvons écrire

$$\pi_{n_1, n_2} = (1 - \rho_1) (\rho_1)^{n_1} (1 - \rho_2) (\rho_2)^{n_2}$$

avec  $\rho_1 = \lambda/\mu_1$  et  $\rho_2 = \lambda/\mu_2$ .

## Hypothèses :

- ▶ Arrivées poissonniennes
- ▶  $r_{ij}$  : probabilité de routage, qu'un client servi par la file  $i$  aille vers la file  $j$ .
- ▶  $0 \leq r_{ij} \leq 1$  pour  $1 \leq i \leq J$  avec  $\sum_{j=1}^J r_{ij} = 1$

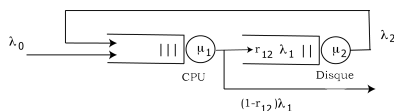


Alors

$$\pi_{n_1, n_2, \dots, n_J} = \prod_{i=1}^J \rho_i^{n_i} (1 - \rho_i)$$

# Réseau de files d'attente avec feedback

Les taux d'arrivées dans chacune des files d'attentes sont de  $\lambda_1 = \lambda_0 + \lambda_2$  et  $\lambda_2 = r_{12}\lambda_1$ .



Nous trouvons facilement que  $\lambda_1 = \frac{\lambda_0}{1-r_{12}}$  et  $\lambda_2 = \lambda_0 \frac{r_{12}}{1-r_{12}}$   
Chaque file d'attente se comporte comme une file d'attente M/M/1 donc nous pouvons facilement la probabilité d'état du système

$$\pi_{n_1, n_2} = (1 - \rho_1) \rho_1^{n_1} (1 - \rho_2) \rho_2^{n_2}$$

avec  $\rho_1 = \lambda_1 / \mu_1$  et  $\rho_2 = \lambda_2 / \mu_2$ .