Evaluation de performances

Théorie des files d'attente

Phuc Do

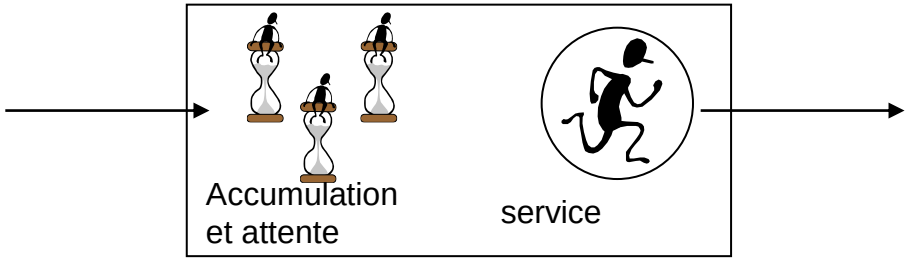
TELECOM Nancy – Université de Lorraine

Files d’attente

- Généralités sur les files d’attente
- Réseaux de files d’attente
- Paramètres de performances
- Files simples markoviennes

Notion de partage de ressources

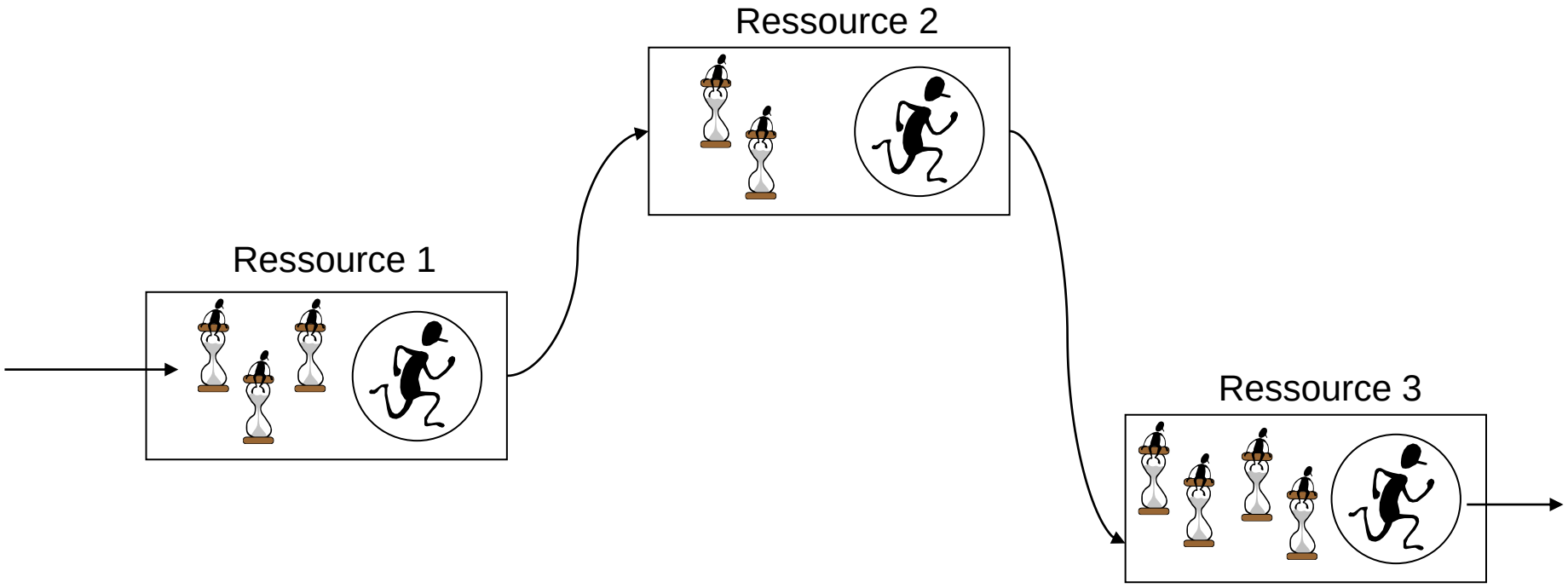
- Le formalisme des files d'attente permet de modéliser des phénomènes de partage de ressources
  - **Ressource** : composante physique, logique ou humaine que les clients d'un système doivent obtenir afin de réaliser une activité
  - Un client faisant une demande d'accès à une ressource devra, en général, attendre que celle-ci soit disponible
  - Dès l'instant où la ressource est disponible, le client entre en service et conserve la ressource pendant toute la durée du service
  - À la fin de son service, le client libère la ressource qui devient alors disponible pour d'autres clients en attente
  - Dans un système simple, l'activité d'un client peut se résumer à l'attente d'une **seule ressource** et à la réalisation du service associé. Ce type de système sera modélisé par **une file simple**



Illustration

Une succession de ressources

– La plupart du temps, l'activité d'un client nécessite tout au long du cycle de vie l'accès à une succession de ressources. Ce type de système sera modélise par **un réseau de files d'attente**

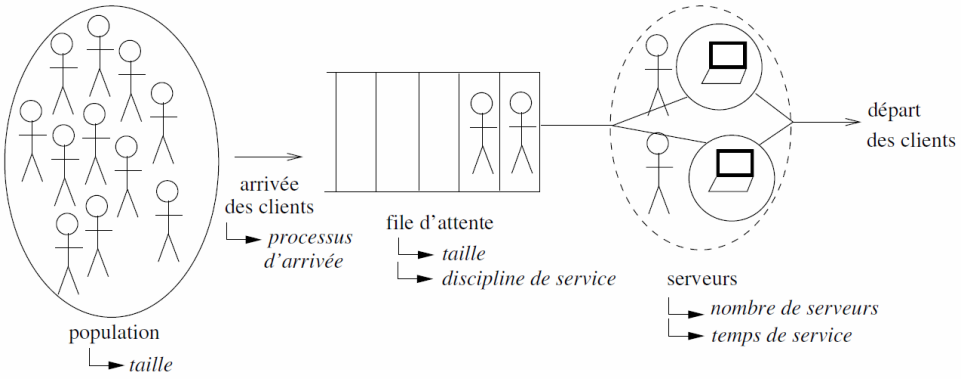


Quelques exemples d'application :

	Clients	Ressource ou serveur	Activité ou service
Systèmes informatiques	processus	processeur	temps de traitement
	demande d'E/S	disque dur	lecture ou écriture
Réseaux de communication	message	réseau	temps de transmission
Système de production	pièce	machine	temps d'usinage
	palette	station de chargement	temps de chargement d'une pièce brute
Guichet SNCF	usager	employé	réservation du billet

File simple

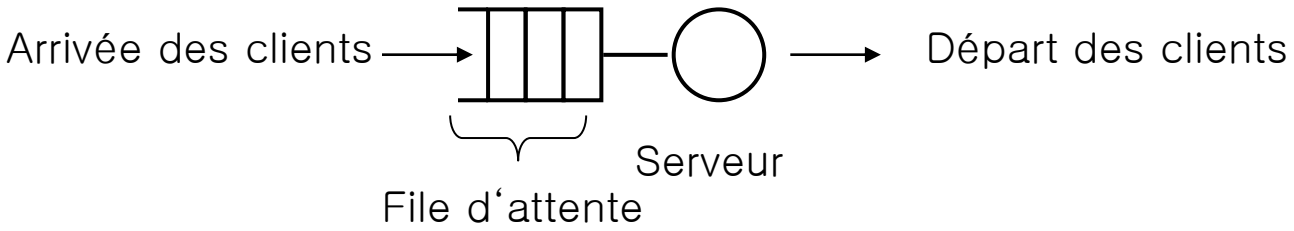
- Une file simple (ou station) est une entité constituée d'une file d'attente (ou buffer) et d'un ou plusieurs serveurs. Les clients arrivent de l'extérieur, patientent éventuellement dans la file d'attente, reçoivent un service, puis quittent la file.



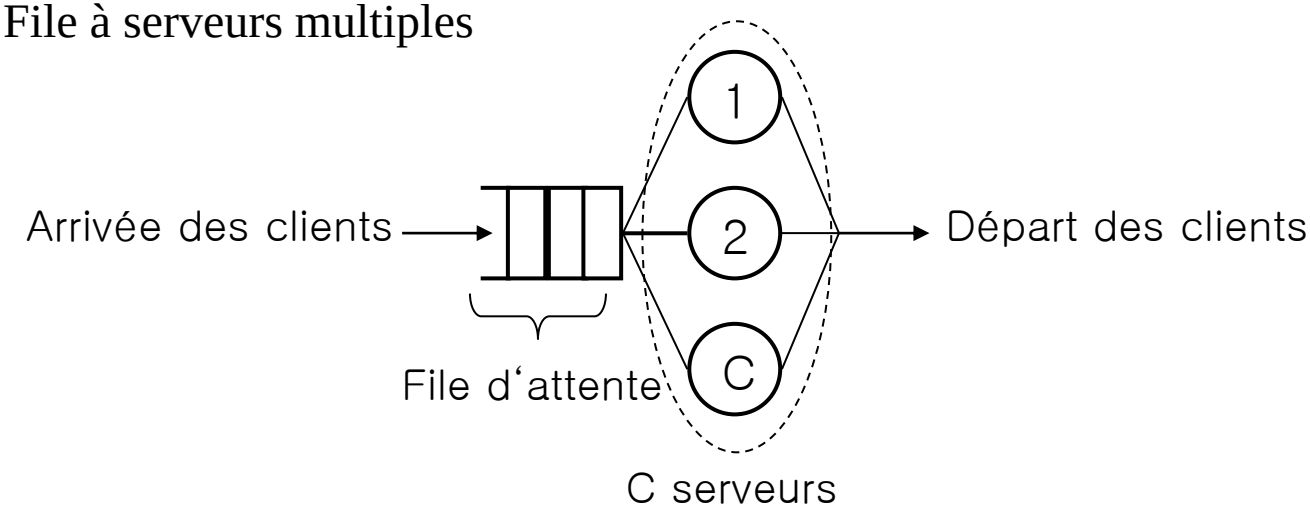
- Une file simple est définie par:
  - Le nombre de serveurs
  - La capacité de la file
  - Le disciple de service qui donne l'ordre dans lequel seront servi les clients:
  - La suite des instants d'arrivées des clients
  - La suite des temps de service des clients

Exemples

- File avec un seul serveur

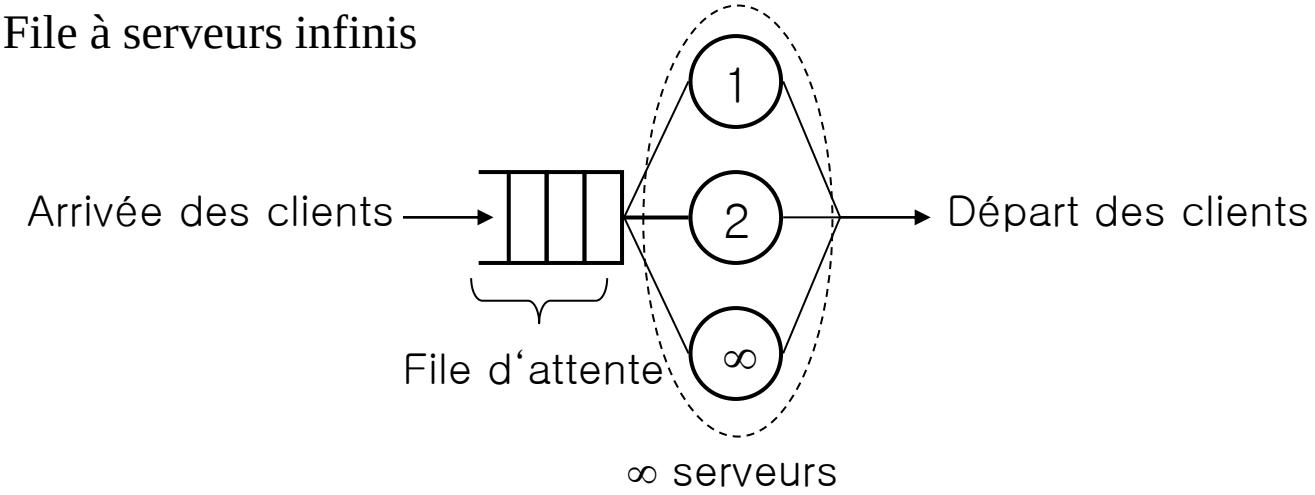


- File à serveurs multiples



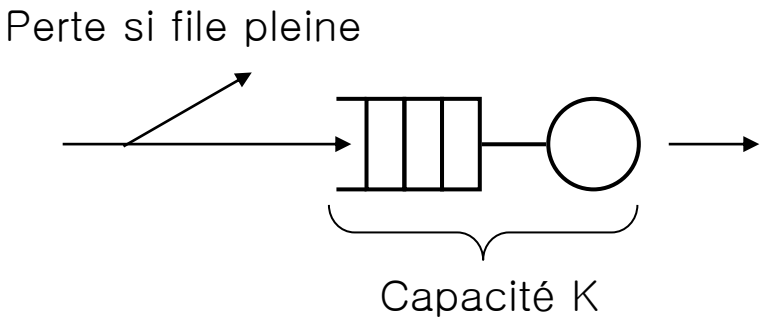
Exemples

- File à serveurs infinis



– **Capacité de la file**

- La capacité de la file peut être finie ou infinie. Lorsque la capacité de la file est limitée et qu'un client arrive alors que la file est pleine, le client est perdu.

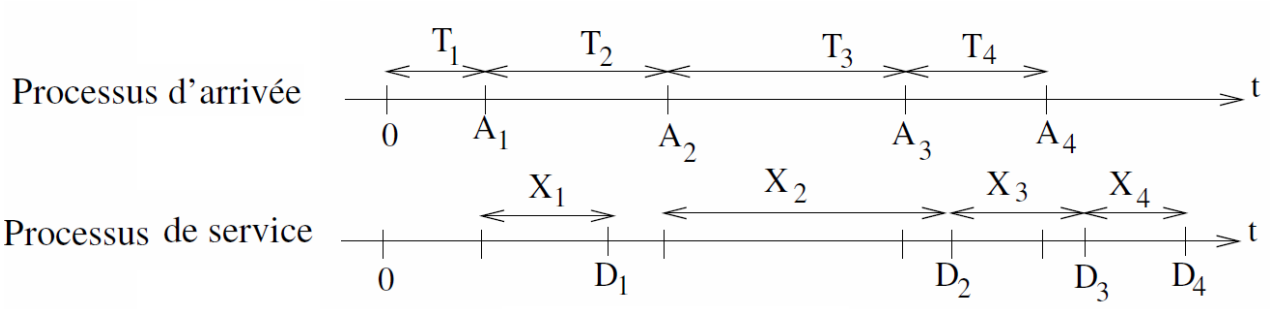


Files simples:

Illustration

- **Processus d'arrivée**
  - La arrivée des clients à la station est généralement décrite par un processus stochastique de comptage.
    - A_n variable aléatoire mesurant l'instant d'arrivée du $n^{ième}$ client
    - T_n variable aléatoire mesurant le temps séparant l'arrivée du $(n-1)^{ième}$ client et celle du $n^{ième}$ client

- **Temps de service**
  - Le temps de service/traitement d'une station est généralement décrite par un processus stochastique.
    - D_n variable aléatoire mesurant l'instant de départ du $n^{ième}$ client
    - X_n variable aléatoire mesurant le temps de service du $n^{ième}$ client (temps entre début et fin de service)



File simple: Notation de Kendall

- Une file d'attente se note:

A/S/C (DS/K/L)
A/S/C/K/L/ (DS)
A/S/C/K/L/DS

- Avec:

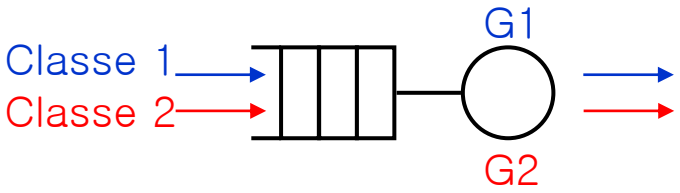
A : processus d'arrivée
S : processus de sortie
C : nombre de serveurs
K : capacité maximale de la file
L : population de clients
DS : discipline de service

- **Symbole pour les arrivées et les services**
  - M : loi exponentielle (Markovienne)
  - D : loi constante
  - E_k : loi Erlang-k
  - H_k : loi hyper-exponentielle ordre k
  - GI : loi générale indépendante
  - G : loi générale
- **Symbole pour les discipline de service**
  - FCFS : First Come First Serve (Preempt)
  - LCFS : Last Come First Serve (Preempt)
  - QUANTUM : Round Robin
  - PS : Processor Sharing
  - RANDOM
  - PRIORITY

Notion de classe de clients

- Une file d'attente peut être parcourue par différentes classes de clients, qui se distinguent par :
  - Des processus d'arrivée différents
  - Des temps de service différents

- Un ordonnancement dans la file d'attente en fonction de leur classe



- Il est ainsi possible de définir des disciplines de service avec priorités entre les types de clients

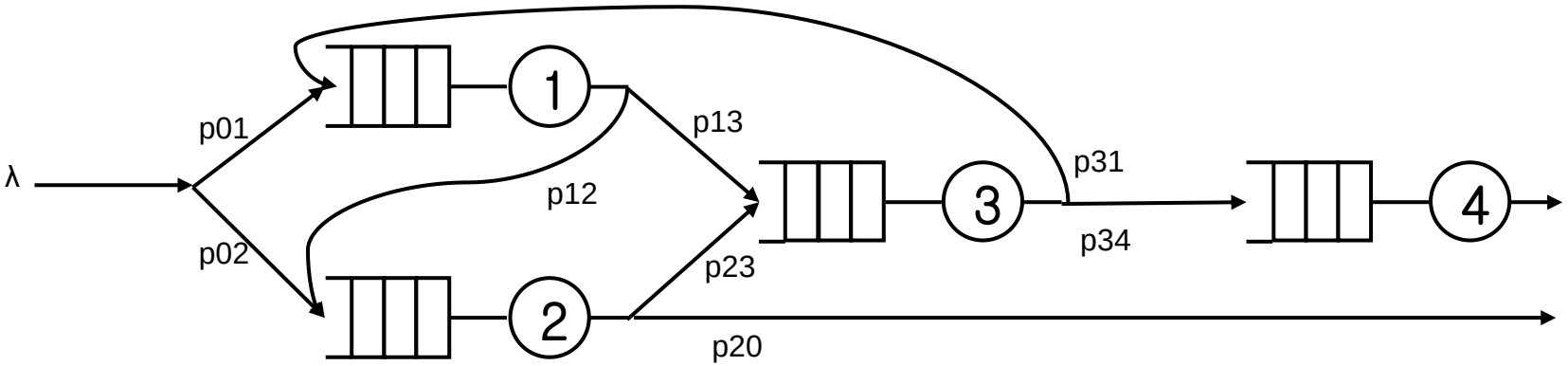
Réseaux de files d'attente

Les réseaux de file d'attente: un ensemble de files simples interconnectées

1. Les réseaux ouverts
2. Les réseaux fermés
3. Les réseaux multiclasses
4. Les réseaux mixtes
5. Les réseaux à capacité limitée

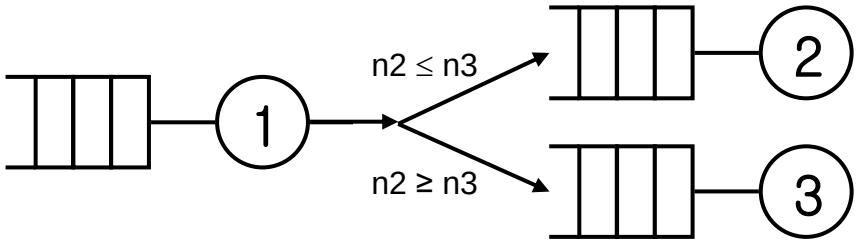
Les réseaux ouverts

- Réseaux dans lesquels les clients arrivent de l'extérieur, circulent dans le réseau à travers les stations, puis quittent le réseau.
- Spécification du réseau :
  - Caractéristiques de chaque station
  - Processus d'arrivée des clients
  - Routage (cheminement) des clients dans le réseau

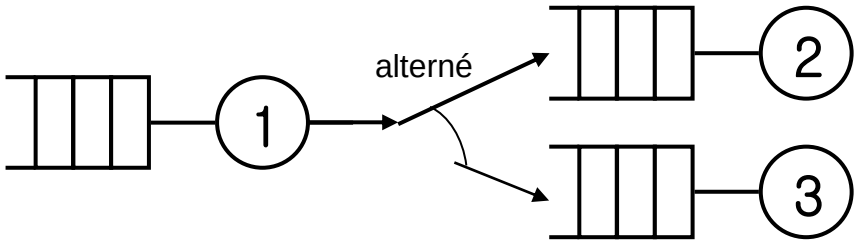


Les réseaux ouverts (suite)

- Processus d'arrivée : comme dans une file simple, caractérisé par un processus de renouvellement (G, M …)
- Routage de clients caractérisé de manière probabiliste : p_{ij}
- D'autres types de routage :
  - Routage vers la file la plus courte (routage dynamique) : un client quittant une station choisira la station qui comporte le moins de clients

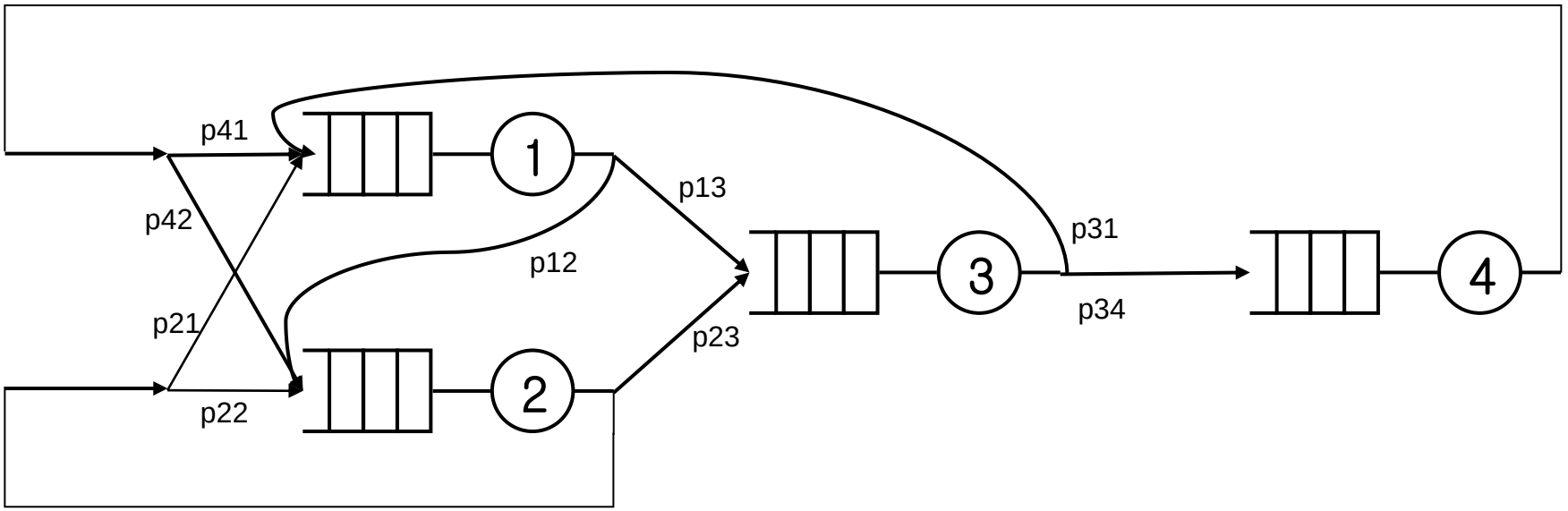


  - Routage cyclique (routage déterministe) : les clients quittant une station choisiront à tour de rôle chacune des station possibles



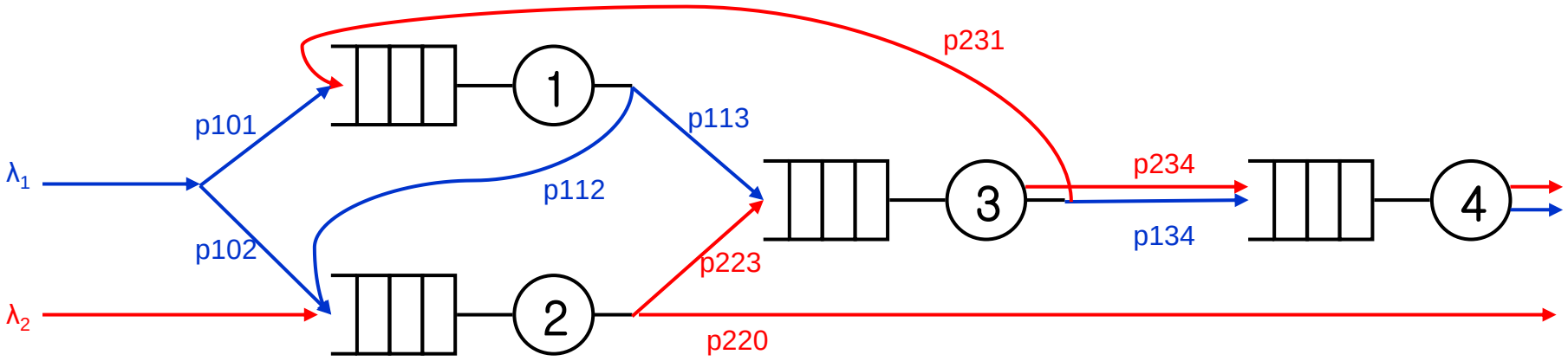
Les réseaux fermés

- Dans un réseau de files d'attente fermé, le nombre des clients est constant
- Pour spécifier un réseau fermé il ne faut spécifier que les stations et le routage



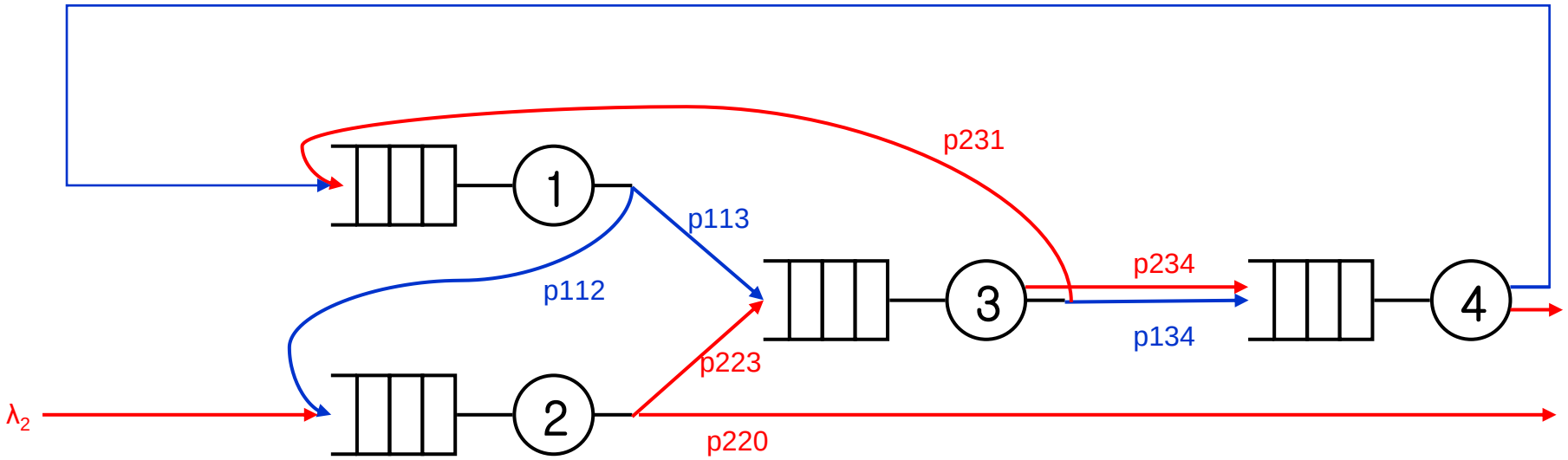
Les réseaux multiclasses

- Comme pour les files simples, les réseaux de file d'attente peuvent être parcourus par différentes classes de clients
- Pour chaque classe de client :
  - Processus d'arrivée différents (si le réseau est ouvert)
  - Comportements différents à chaque station
  - Routage différents dans le réseau



Les réseaux mixtes

- À partir de la notion de réseau multiclasses, nous pouvons définir le notion de réseau mixte
  - Ouvert vis à vis de certaines classes
  - Fermé vis à vis d'autres classes



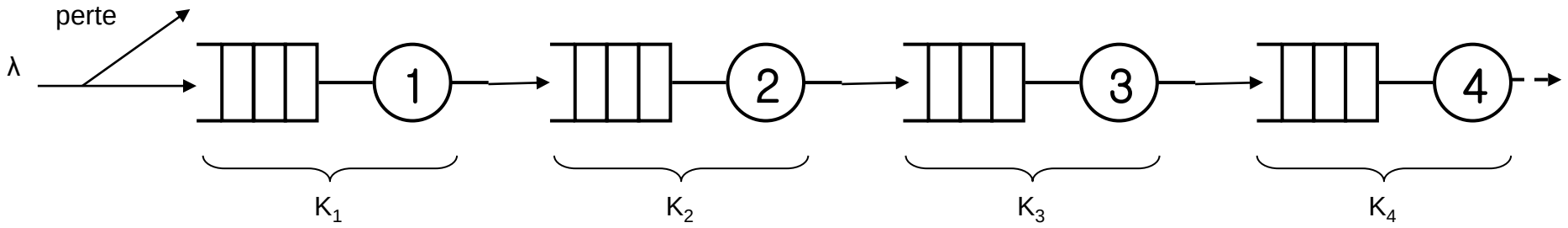
Classe 1 : fermée
Classe 2 : ouverte

Les réseaux de file d'attente à capacité limitée

- Les différentes stations du réseau peuvent avoir des capacités limitées. Lorsqu'une file est pleine, plus aucun client ne peut entrer

➔ blocages dans les stations amont

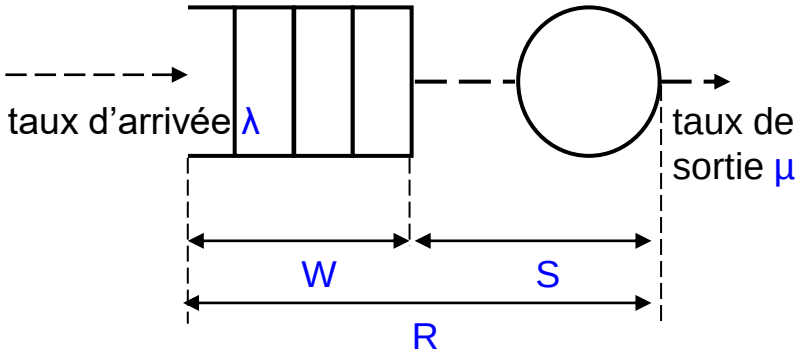
➔ perte de clients à l'entrée du système (si le système est ouvert)



Paramètres de performances

Quelques critères de performance intéressantes à déterminer:

- Le *temps moyen de séjour* d'un client dans le système (temps moyen de réponse de la station): R
- Le *temps moyen d'attente*: W
- Le *nombre moyen de clients en attente de service*: Qw
- Le *nombre moyen de clients dans le système*: Q
- Le *taux d'utilisation* du serveur: ρ



- On note le taux d'utilisation du serveur : $\rho = \frac{\lambda}{\mu}$
- Il constitue une mesure pour le degré de saturation du système. Il correspond au nombre moyen d'arrivées par durée moyenne de service
- Condition de stabilité : $\lambda \leq \mu$
- Propriétés :
  - On a les relations suivantes:

$$R = W + S$$

$$W = Q.S \Rightarrow R = Q.S + S = (Q+1).S$$

$$S = \frac{1}{\mu} \Rightarrow R = \frac{Q+1}{\mu}$$

W : temps d'attente
S : temps du service
R : temps de séjour
Q : nombre de clients
μ : taux de service d'un serveur
λ : taux d'arrivée

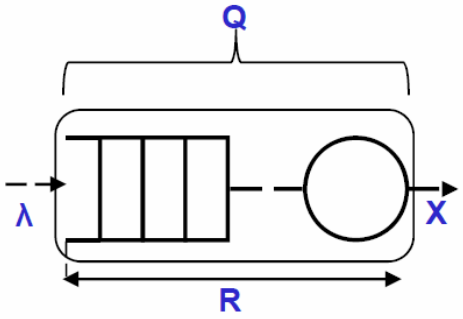
Théorème de Little

- Soit une file d'attente dont les lois d'arrivée et de service sont *indépendantes*
- On suppose que le système évolue en *régime stationnaire*

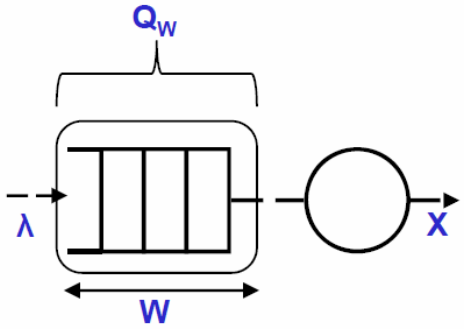
☞ **Loi de Little**

$Q = R\,X$ (Q nombre moyen de clients, R temps moyen de réponse, X débit moyen)

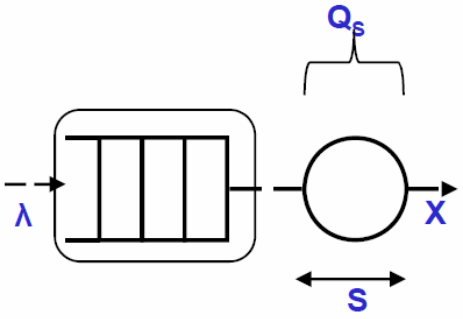
Pour une file simple avec un unique serveur



$Q = R\,X = R\,\lambda$



$Q_W = W\,X = W\,\lambda$



$Q_S = S\,X = S\,\lambda$

Files simples markoviennes

Les files simples markoviennes, notées M/M…, sont telles que:

- Les inter-arrivées ont une distribution exponentielle de paramètre λ
- Le temps de service d'un client au sein d'un serveur est une variable aléatoire ayant une distribution exponentielle de taux μ

Quelques files simples markoviennes

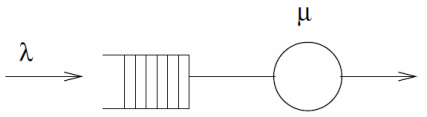
- La file M/M/1
- La file M/M/1/K
- La file M/M/C
- La file M/M/C/C

La file M/M/1:

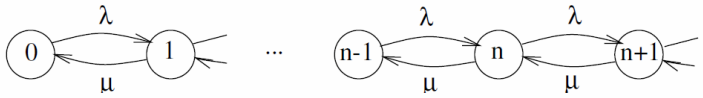
File d'attente à un seul serveur, et une capacité de file infinie

Dans le cas où $\lambda < \mu$, la file admet un régime permanent et les probabilités d'état sont :

$$\pi_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \quad \forall n \geq 0$$



File M/M/1



Chaîne de Markov associée à la file M/M/1

Paramètres de performances

- Débit moyen : $X = \lambda$
- Nombre moyen de clients : $Q = \dfrac{\lambda}{\mu - \lambda}$
- Temps moyen de séjour : $R = \dfrac{1}{\mu - \lambda}$
- Taux d'utilisation du serveur : $\rho = \dfrac{\lambda}{\mu}$
- Nombre moyen de clients dans la file d'attente :

$$Q_W = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

- Temps moyen de séjour dans la file d'attente :

$$W = \frac{\lambda}{\mu(\mu - \lambda)}$$

- Nombre moyen de clients dans le serveur :

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

- Temps moyen de service : $S = \dfrac{1}{\mu}$

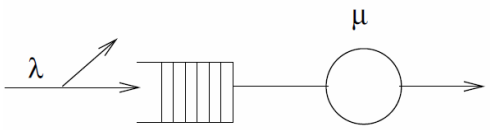
La file M/M/1: (suite)

- Exemple :
  - 1 client arrive en moyenne toutes les 12 ms, durée de service moyen : 8 ms.
  - Quelle est la probabilité p que deux clients au moins attendent d'être servis ?
  - Calculer les mesures de performances (Q, R, Qw, W, Qs) ?

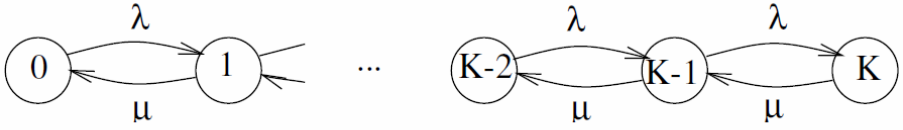
Solution ?

La file M/M/1/K:

File d'attente à un seul serveur, et la capacité est limitée à K clients, e.i. K-1 clients dans la file d'attente



File M/M/1/K



Chaîne de Markov associée à la file M/M/1/K

Régime permanent : La file est toujours stable, puisque sa capacité est limitée.

Les probabilités d'état sont : $\pi_n = \frac{1-\rho}{1-\rho^{K+1}}\rho^n$ pour $0 \leq n \leq K$

Paramètres de performances

- Débit moyen : $X = \frac{1-\rho^K}{1-\rho^{K+1}}\lambda$
- Nombre moyen de clients :

$$Q = \frac{\rho}{1-\rho}\frac{1-(K+1)\rho^K+K\rho^{K+1}}{1-\rho^{K+1}}$$

- Temps moyen de séjour : $R = Q/X$
- Taux d'utilisation du serveur :

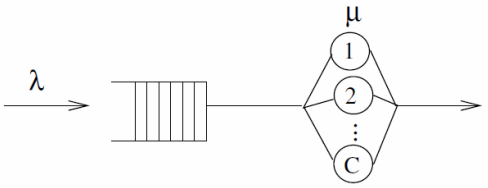
$$U = \rho\frac{1-\rho^K}{1-\rho^{K+1}}$$

- Probabilité de pertes (probabilité qu'un client qui arrive ne puisse entrer) :

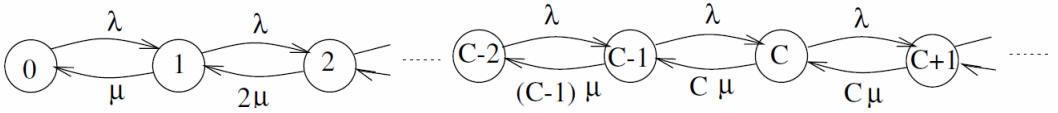
$$P(pertes) = \pi_K$$

La file M/M/C:

File d'attente à **une capacité infinie** et **C serveurs identiques** avec un temps de service exponentiel de paramètre μ pour chacun d'eux



File M/M/C



Chaîne de Markov associée à la file M/M/C

Régime permanent: la file est stable si $\lambda < C.\mu$

Les probabilités d'état en régime permanent sont : $\pi_0 = \left[\sum_{n=0}^{C-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^C}{(C-1)!\,(C-\lambda/\mu)}\right]^{-1}$

$$\pi_n = \frac{(\lambda/\mu)^n}{n!}\pi_0 \quad \text{pour } n \le C \qquad \pi_n = \frac{(\lambda/\mu)^n}{C!\,C^{n-C}}\pi_0 \quad \text{pour } n \ge C$$

Paramètres de performances

- Débit moyen : $X = \lambda$
- Temps moyen de service : $S = \frac{1}{\mu}$
- Nombre moyen de clients en attente :

$$Q_W = \frac{(\lambda/\mu)^{C+1}}{(C-1)!\,(C-\lambda/\mu)^2}\pi_0$$

- Temps moyen dans la file d'attente :

$$W = Q_W/\lambda$$

- Temps moyen de séjour :

$$R = \frac{(\lambda/\mu)^C}{\mu\,(C-1)!\,(C-\lambda/\mu)^2}\pi_0 + \frac{1}{\mu}$$

- Nombre moyen de clients :

$$Q = R\lambda = \frac{(\lambda/\mu)^{C+1}}{(C-1)!\,(C-\lambda/\mu)^2}\pi_0 + \lambda/\mu$$

La file M/M/C (suite):

– Exemple:

- On considère une arrivée aléatoire (markovienne) de clients à un taux de 9 clients/heure
- On dispose de commutateurs de capacité 6 paquets par heure (temps de service exponentiel)
- On voudrait savoir le nombre minimal de guichet nécessaires pour que le système soit en régime stationnaire

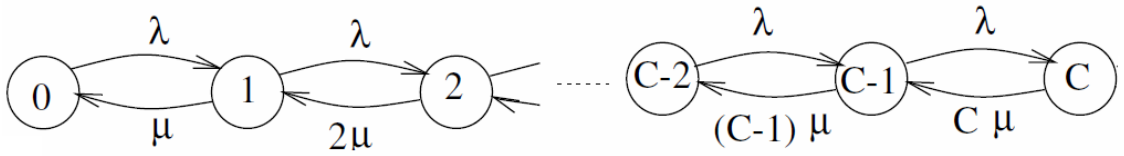
  $\rho < C \Rightarrow 1{,}5 = \lambda/\mu < C$ donc $C = 2$

- Calculer le temps d'attente (dans la file) moyen ?

Solution ?

La file M/M/C/C:

- File d'attente à **C serveurs identiques** avec un temps de service exponentiel de paramètre μ pour chacun d'eux, pas de file d'attente.
- Lorsqu'un client arrive, s'il y a au mois un serveur disponible, le client entre directement en service. Sinon, le client est rejeté

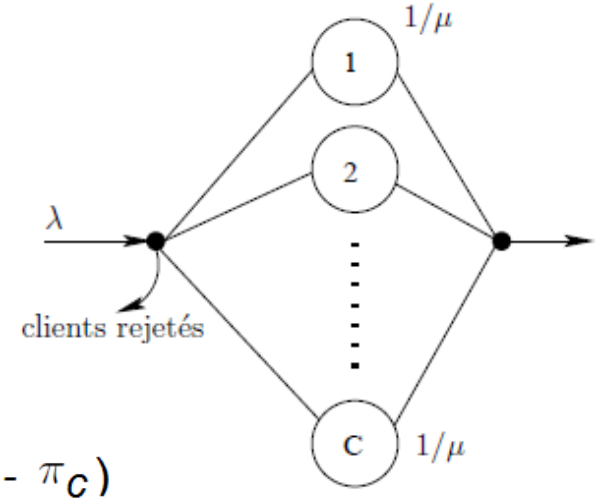


Chaîne de Markov associée à la file M/M/C/C

$$\pi_i = \pi_0 \frac{\rho^i}{i!} \quad i = 0, 1, \ldots, C$$

où,

$$\pi_0 = \frac{1}{\sum_{i=0}^{C} \rho^i / i!}.$$



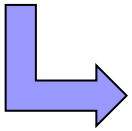
– Temps moyen de séjour : $R = \frac{1}{\mu}$

– Nombre moyen de clients :

$$Q = R\lambda_{entre} = R\lambda \sum_{i=0}^{c-1} \pi_i = R\lambda (1 - \pi_c)$$

$$= \rho (1 - \pi_0 \frac{\rho^c}{c!})$$

Evaluation de performances



Sûreté de fonctionnement

à suivre…