

# Evaluation de performances

## Analyse opérationnelle

### Plan du cours:

- *Introduction*
- *Formulation d'indicateurs de perf.*
- *Taux d'occupation d'un système*
- *Système interactif / Système complexe*
- *Analyse asymptotique*
- *Conclusions*

(adapté à partir de Georges Keryvel « Arte et Métiers »)

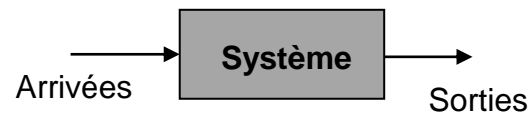
- ❑ Système physique donné quelconque
- ❑ Observation sur une période de temps finie
- ❑ Ensemble de **grandeurs mesurables**:
  - les sondes matérielles permettent de ne regarder que ce que l'on veut, si on peut identifier ce que l'on veut
  - les sondes logicielles permettent de mesurer ce qui n'est pas mesurable matériellement (e.g. nombre d'appels)
- ❑ Objectifs:
  - Contrôle des mesures (relations cohérentes, redondantes)
  - Explication des phénomènes observées
  - Définir des critères caractérisant le système, à partir des **mesures effectuées**
  - Donner d'autres critères, non directement mesurables

## Approche d'analyse opérationnelle

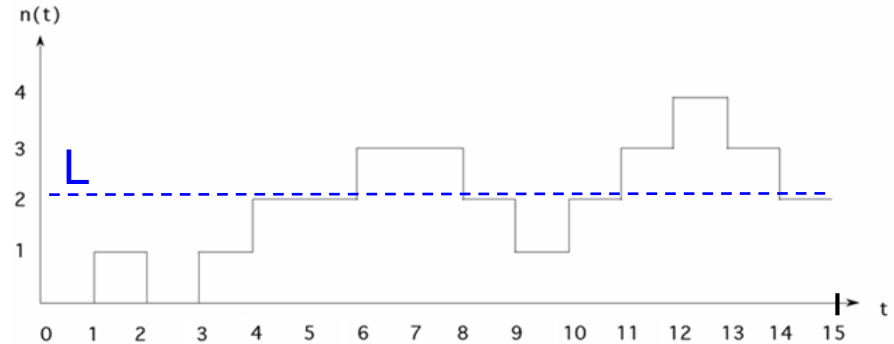
- ❑ L'analyse opérationnelle a été appliquée après la théorie des files d'attente
- ❑ En 1978, Denning et Buzen adoptent une approche opérationnelle qui consiste à dériver **un ensemble de relations à partir des observations** faites sur un système
- ❑ Ces relations fondamentalement sont vérifiées quel que soit le système et la période de mesure. Ces hypothèses se trouvent en théorie des files d'attente sous l'aspect probabilistes

### Principe de l'analyse opérationnelle

- ❑ Le système est vu comme une boîte noire recevant des requêtes et les restituant après un certain temps de traitement



- ❑ Deux compteurs permettent de connaître le nombre total de requêtes entrantes et sortantes du système
- ❑ Aucune hypothèse (ordre de traitement, parallélisme etc.)



## ❑ Mesures élémentaires :

- Durée de la mesure:  $T$
- Nombre total d'arrivées de requêtes:  $A$
- Nombre total de départs de requêtes:  $D$
- Durée cumulée pendant laquelle le système a contenu  $n$  requêtes:  $T(n)$
- Nombre maximum de requêtes dans le système :  $n_{\max}$

## ❑ On recherche les critères de performances suivants:

- Débit du système à l'entrée:  $\Lambda = \frac{A}{T}$
- Débit du système à la sortie:  $X = \frac{D}{T}$
- Nombre moyen de requêtes dans le système:  $L = \frac{\sum_{n=1}^{n_{\max}} n \cdot T(n)}{T}$
- Temps de réponse du système:  $R = \frac{\sum_{n=1}^{n_{\max}} n \cdot T(n)}{D}$

- ❑ Relation:  $L = \lambda . R$ 
  - Si  $A=D$  alors  $L = \lambda . R$

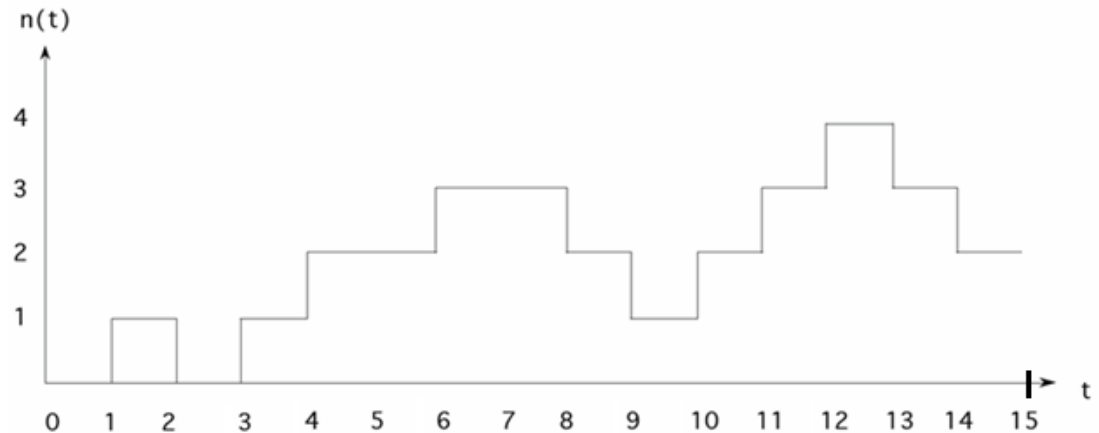
- ❑ Formule de Little:

*Le nombre moyen de requêtes dans un système est égal au produit du débit de ce système par le temps moyen d'une requête passé dans ce système*

## Formule de Little: exemple

☞ Déterminer:

- $X$
- $\Lambda$
- $L$
- $R$



### 1. T=15

$$\begin{aligned} A &= 7 \\ D &= 5 \\ X &= 1/3 \\ \Lambda &= 7/15 \\ L &= 29/15 \\ R &= 29/5 \end{aligned}$$

$$\begin{aligned} T(0) &= 2 \\ T(1) &= 3 \\ T(2) &= 5 \\ T(3) &= 4 \\ T(4) &= 1 \end{aligned}$$



### 2. T=12

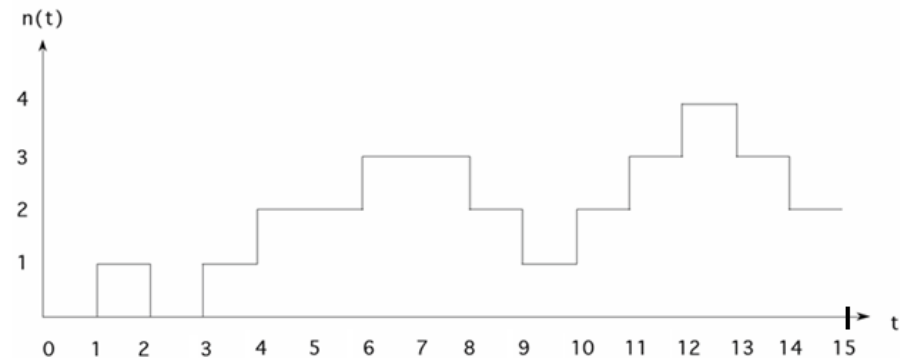
$$\begin{aligned} A &= 6 \\ D &= 3 \\ X &= 1/4 \\ \Lambda &= 1/2 \\ L &= 20/12 \\ R &= 20/3 \end{aligned}$$



## Définitions

- ❑ Soit  $B$ , la durée d'occupation d'un système pendant une période d'observation  $T$ :  $B = T - T(0)$

- ❑ Taux d'occupation:  $U = B/T$ 
  - $U$  est toujours inférieur ou égal à 1



- ❑ Durée apparente du service  $S$ :
  - Temps de service moyen, demandé par requête
  - $S = B/D$

Relation:  $U = X.S$  car  $U = B/T = D/T.B/D$

ou  $X = U/S$

## Exemple

- ❑ Soit un système ayant un processeur et un disque
- ❑ Caractéristique du disque:
  - *Temps de service:  $S_d=25ms$*
  - *Taux d'utilisation:  $U_d=20\%$*
  - *Chaque transaction du système génère 8 requêtes sur disque*
- ❑ *Calculer le débit du système en nombre de transactions par seconde ?*

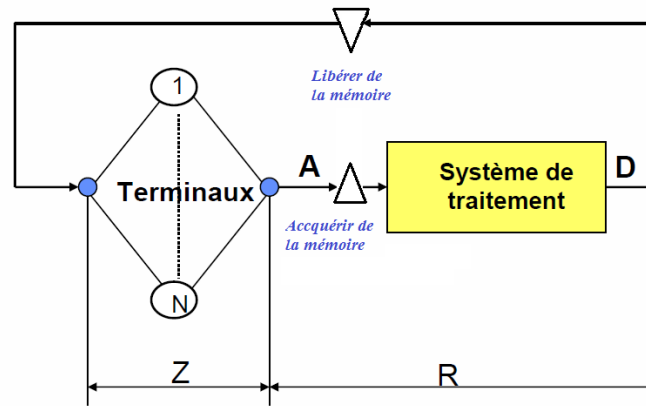
## Solution:

- ❑ Débit du disque:
  - $X_d = U_d / S_d = 0,2 / (25 * 10^{-3}) = 8 \text{ requêtes/seconde}$
- ❑ Débit du système:
  - $X = X_d / 8 = 1 \text{ transaction/seconde}$



## Système interactif

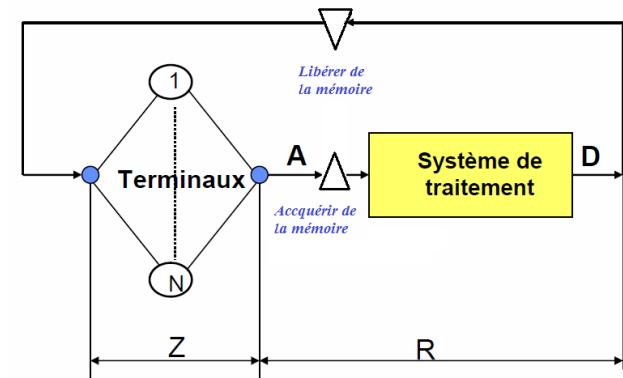
- ❑ Considérons un serveur accédé à partir d'un ensemble de terminaux.
- ❑ A chaque terminal, on associe un processus alternant entre 2 phases:
  - **Réflexion** : l'utilisateur réfléchit ou frappe au clavier (avant le ENTER)
  - **Traitement** : la requête est traitée par le serveur, attente de réponse.



- ❑ On s'intéresse à évaluer les critères de performance suivantes:
  - *Temps de réponse moyen du système*
  - *Temps de réflexion moyen du système*
  - *Débit en sortie du système*

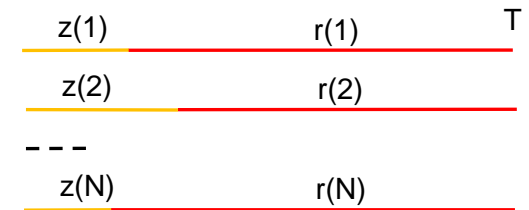
## ❑ Mesures élémentaires :

- Durée de la mesure:  $T$
- Nombre de terminaux connectés:  $N$
- Nombre de requêtes envoyées depuis les terminaux:  $A$
- Nombre requêtes traitées par le système:  $D$
- Durée cumulée passée en réflexion par le processus  $k$ :  $z(k)$
- Durée cumulée passée en traitement par le processus  $k$ :  $r(k)$



## ❑ On a:


- Temps de réponse moyen du système:  $R = \sum_{k=1}^N r(k) / D$
- Temps de réflexion moyen du système:  $Z = \sum_{k=1}^N z(k) / A$
- Débit en sortie du système:  $X = \frac{D}{T}$



$$r(k) + z(k) = T \Rightarrow \forall k \Rightarrow R.D + Z.A = N.T \quad d'où$$

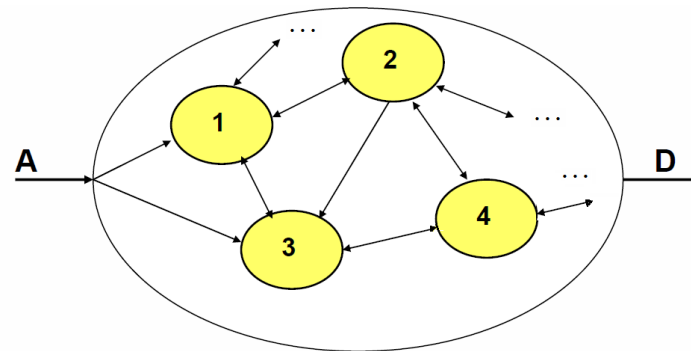
$$R = \frac{N}{X} - \frac{A}{D} Z$$

❖ En régime stationnaire:  $\frac{A}{D} \approx 1$

  $R = \frac{N}{X} - Z$

## Système complexe

- ❑ On considère un système constitué de **plusieurs stations mono-serveur** de traitement.
  - Les travaux envoyés au système engendrent des requêtes élémentaires.
  - Un travail peut engendre plusieurs requêtes et celles-ci peuvent être traitées simultanément sur différentes stations.
  - Aucune autre hypothèse n'est faite sur le fonctionnement interne du système
  - On considère chaque station comme un **sous-système**



## ❑ Mesures élémentaires :

- $T$  = durée de la mesure
- $D$  = nombre total de requêtes globales traitées par le système
- $D_i$  = nombre total de requêtes élémentaire traitées par la station  $i$
- $T_i(n)$  = durée cumulée pendant laquelle la station  $i$  a contenu  $n$  requêtes élémentaires

## ❑ Evaluation de performances: on cherche les critères suivantes:

- Débit de la station  $i$ :  $X_i = D_i / T$
- Taux d'occupation de la station  $i$ :  $U_i = (T - T_i(0)) / T$
- Durée moyenne de service de la station  $i$ :  $S_i = (T - T_i(0)) / D_i$
- Nombre moyen de visite à la station  $i$  par travail:  $e_i = D_i / D$
- Temps de réponse de la station  $i$ :  $R_i = \frac{\sum n.T_i(n)}{D_i}$
- Nombre moyen de requêtes élémentaires dans la station  $i$ :  $L_i = \frac{\sum n.T_i(n)}{T}$
- Débit global du système:  $X = \frac{D}{T}$

$$\Rightarrow \boxed{X = \frac{X_i}{e_i} = \frac{U_i}{S_i \cdot e_i} = \frac{L_i}{R_i \cdot e_i} \quad \text{Th de Chang-Lavenberg (version opérationnelle)}}$$

$S_i \cdot e_i$  = temps total de service demandé à la station  $i$

$e_i$  = nombre moyen de requêtes envoyées au sous-système  $i$  par transaction

## Cas particulier

- Hypothèses supplémentaires : si une requête globale (travail) ne peut générer plusieurs requêtes élémentaires alors:

$$L = \sum_i L_i$$

$$R = \sum_i R_i \cdot e_i$$

$$L = X \cdot R$$

$$L_i = X_i \cdot R_i$$

- Toutes les relations précédentes peuvent être appliquées à des populations (classes) distinctes de travaux
  - Il suffit de restreindre les mesures aux requêtes issues de chaque population.
  - Au niveau d'une station, additionner des débits et des taux d'occupation

$$X_s = \sum_j X_s^j$$

$$U_s = \sum_j U_s^j$$

où j désigne une population, s la station

## Analyse de saturation

- ❑ Un système est dit saturé si au moins un de ses sous-système l'est.
- ❑ Le taux d'occupation du sous-système saturé est 1
  
- ❑ Considérons un sous-système saturé, noté sous-système b:
  - Son taux d'occupation:  $U_b=1$
  - Son débit maximum:  $X_b = \frac{1}{S_b}$
  
- ❑ Débit maximum du système:

$$X_{\max} = \frac{X_b}{e_b} = \frac{1}{S_b e_b}$$

- ❑ L'analyse opérationnelle a permis d'introduire de manière **simple** quelques critères de performance en se basant uniquement sur des **observations**.

## ❖ Limites de l'analyse opérationnelle

- Problème de collecte des informations
- Instrumentation lourde
- Interprétation délicate des résultats
- *Réalisation impossible dans la phase de conception*
- Si on désire connaître, par exemple, le temps de réponse d'une station, connaissant le débit d'arrivée et le temps moyen de service, on en est incapable



***Il est nécessaire d'étudier plus finement les interactions entre arrivées et service.***

- Introduction des hypothèses de nature statistique sur le comportement des requêtes.
- Processus stochastiques et files d'attente fournissent des résultats utilisables dans un grand nombre de situations