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 « Arts et Métiers»)

Phuc Do
phuc.do@univ-lorraine.fr

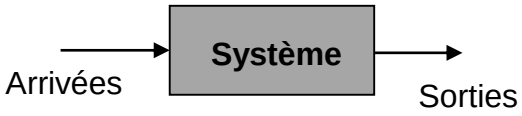
- Système physique donné quelconque sur une période de temps finie
- Ensemble de grandeurs mesurables:
  - les sondes matérielles permettent de ne regarder que ce que l'on veut, si on peut identifier ce que l'on veut
  - les sondes logicielles permettent de mesurer ce qui n'est pas mesurable matériellement (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 quelques soient le système et la période de mesure. Ces hypothèses se trouvent en théorie des files d'attente sous l'aspect probabilistes

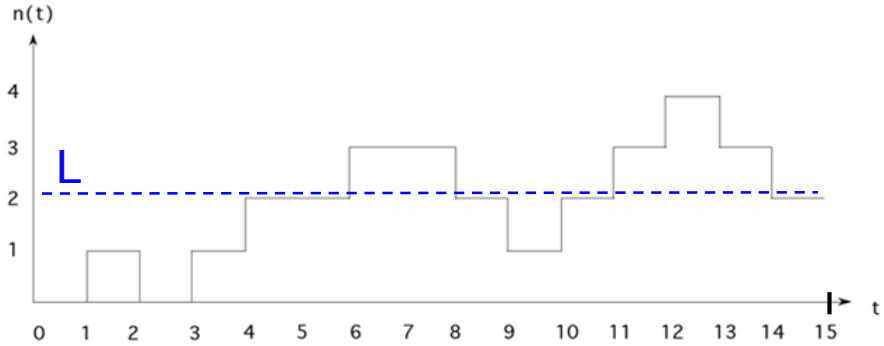
Principe de l'analyse opérationnelle

- 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.)

Formule de Little



- 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 : nmax

- 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:
  - Temps de réponse du système:

Formule de Little

- Relation:
  - Si ***A***=***D*** alors:

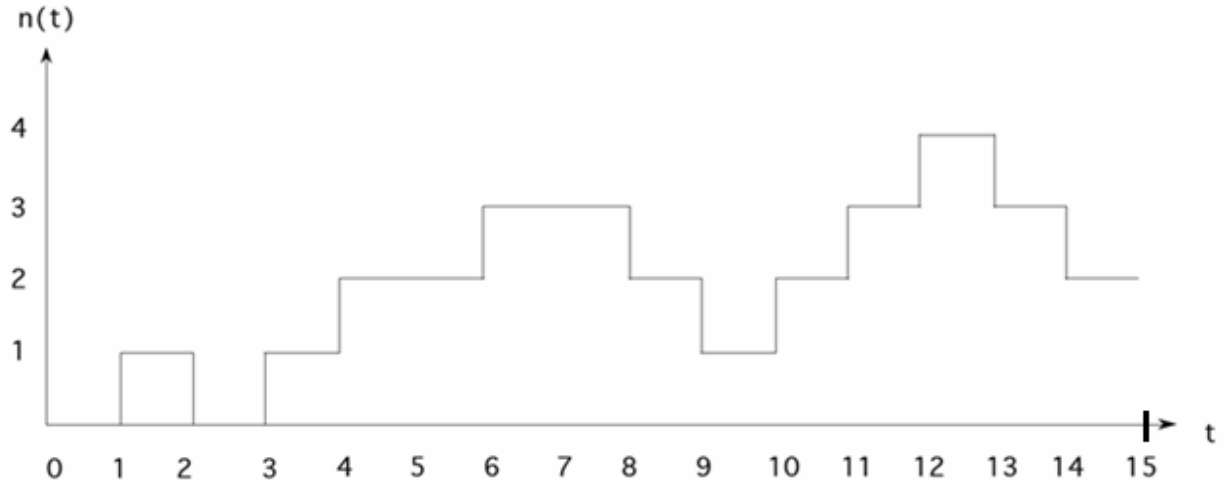
- Formule de Little:

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
- Λ
- L
- R

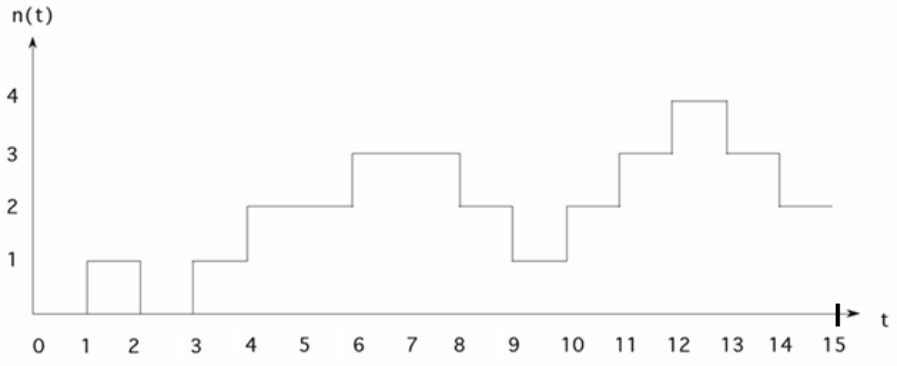


1. T=15

2. T=12

Définitions

- Soit *B*, la durée d'occupation d'un système pendant une période d'observation *T:*



- Taux d'occupation: *U=B/T*
  -

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

Relation:

Exemple

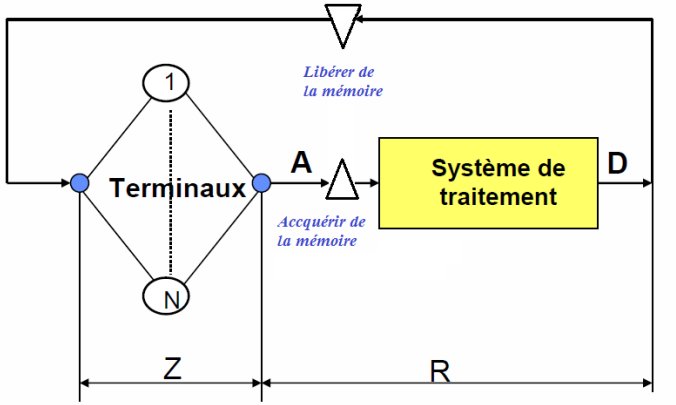
- Soit un système ayant un processeur et un disque
- Caractéristique du disque:
  - *Temps de service: S_d=25ms*
  - *Taux d'utilisation: U_d=20%*
  - *Chaque transaction du système génère 8 requêtes sur disque*
- *Calculer le débit du système en nombre de transactions par seconde ?*

Solution:

- Débit du disque:
  -
- Débit du système:
  -

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

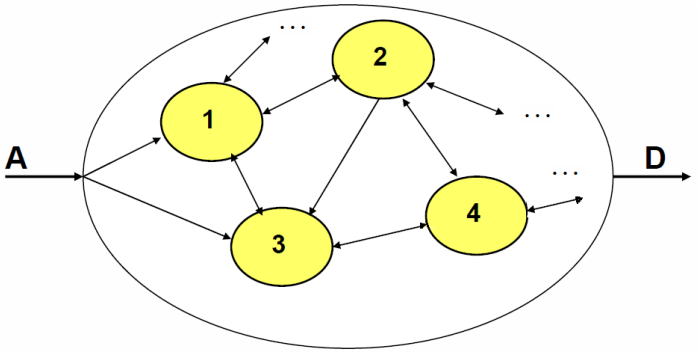
- Mesures élémentaires :
  - Durée de la mesure: T
  - Nombre de terminaux connectés: N
  - Nombre de requêtes envoyées depuis les terminaux: A
  - Nombre requêtes traitées par le système: D
  - Durée cumulée passée en réflexion par le processus k: z(k)
  - Durée cumulée passée en traitement par le processus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

- On a:
  - Temps de réponse moyen du système:
  - Temps de réflexion moyen du système:
  - Débit en sortie du système: $X = \frac{D}{T}$

❖ **En régime stationnaire**:

Système complexe

- On considère un système constitue de **plusieurs stations mono-serveur** de traitement.
  - Chaque travail (transaction/requête) envoyé au système 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*:
  - Nombre moyen de visite à la station i par travail:
  - 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: $X = \frac{D}{T}$

$S_i.e_i$ = temps total de service demandé à la station *i*

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

$$L = \sum_i L_i$$

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

$$L = \boldsymbol{X} \cdot R$$

$$L_i = \boldsymbol{X}_i \cdot R_i$$

- 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

$$\boldsymbol{X}_s =$$

$$U_s =$$

où j désigne une population, s la station

- Objectif: Etudier le débit (X) et le temps de réponse (R) du système en fonction du nombre de transactions N
  - Considérons un système fermé. Comment varie la performance en fonction du nombre de transactions existant dans le système ?
  - Hypothèse: temps moyen de service S_i et le taux de visite e_i sont constants indépendants de N (c'est à dire que les services globaux sur chaque station sont indépendants de la charge)

Analyse de saturation

-
-

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

Débit du système

- 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:

- L'analyse opérationnelle a permit 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'arrivé 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