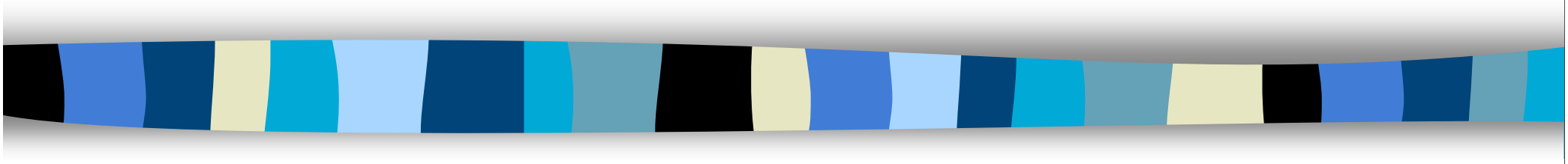
Modélisation et Evaluation de Performance



Introduction et analyse opérationnelle

Auteur: PHAM Cong-Duc

Modélisation et Evaluation de performance

Auteur: PHAM Cong-Duc

- L'évaluation de performance: qu'est-ce et pourquoi?
- Le processus de modélisation
- La simulation
  - Simulation à événements discrets
  - Les outils de simulation
  - Autres formes de simulation
- L' analyse opérationnelle
  - Formule opérationnelle de Little
- Les méthodes analytiques : files d'attente
  - Processus de naissance et de mort
  - File élémentaire M/M/1 et files dérivées (M/M/m M/M/1/K...)
  - Réseaux de files d'attente de type Jackson, Gordon-Newel, BCMP

Evaluation qualitative / quantitative

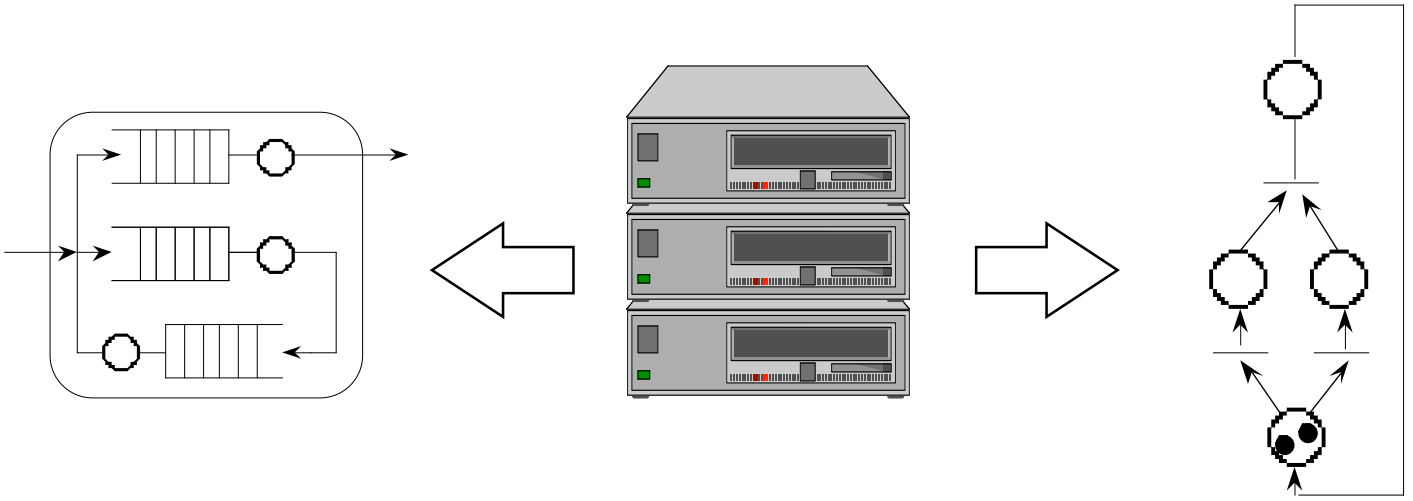
Auteur: PHAM Cong-Duc

- Il existe deux approches d'évaluation pour un système, l'approche qualitative et l'approche quantitative.
- L'évaluation qualitative s'intéresse à définir des propriétés structurelles et comportementales.
  - Absence de blocage
  - Existence d'une solution
  - Gestion de la concurrence
- L'évaluation quantitative consiste à calculer les critères de performances du système.
  - Débit
  - Temps de réponse
  - Etc...

Formalisme qualitatif / quantitatif

Auteur: PHAM Cong-Duc

- L'analyse quantitative est essentiellement réalisée à l'aide de files d'attente.



- L'analyse qualitative fait appel aux Réseaux de Pétri ou aux langages formels (Lotos, Esterelle...).

L'évaluation de performance

Auteur: PHAM Cong-Duc

- L'évaluation de performance s'intéresse au valeurs quantitatives d'un système.

- Guichet SNCF
  - Temps d'attente des usagers
  - Nombre de clients, débit d'un guichet

- Réseaux de communication
  - Débits en paquets, cellulles...
  - Taux de pertes, de retransmission...

- Atelier de production
  - Taux d'utilisation d'une machine
  - Temps de fabrication

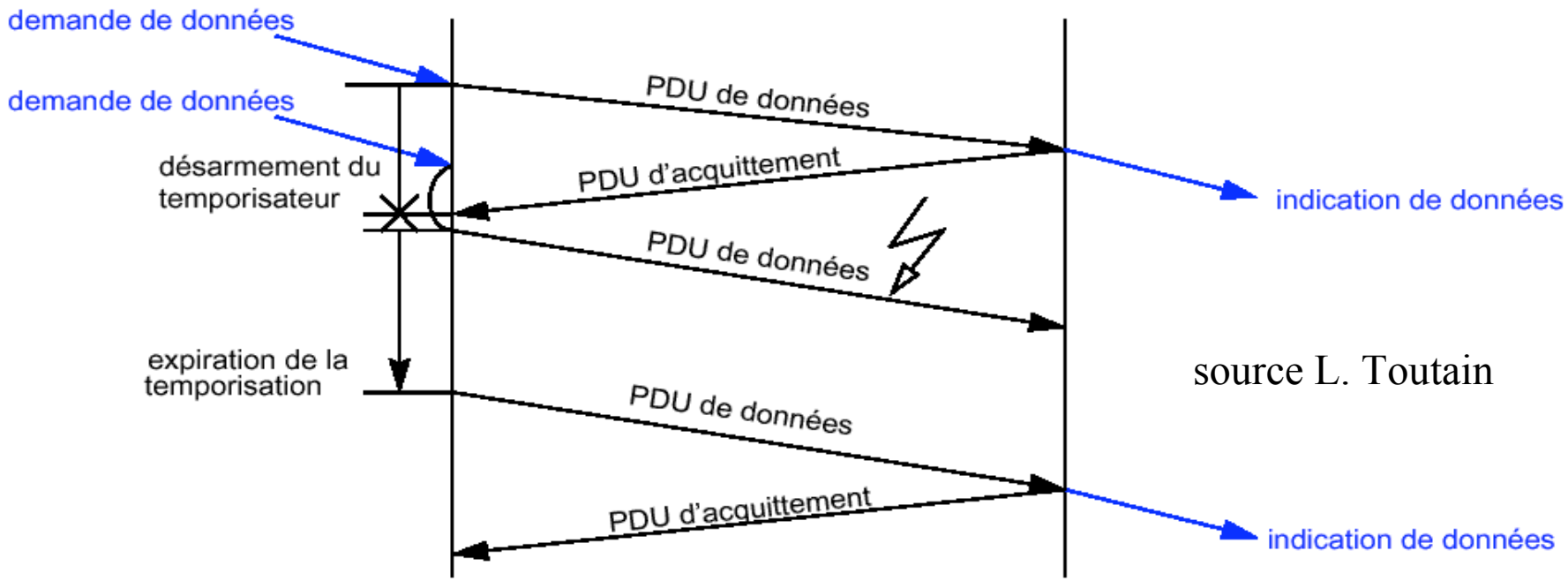
Pourquoi évaluer les performances ?

Auteur: PHAM Cong-Duc

- Phase de conception
  - Le système n'existe pas.
  - Dimensionner le système futur selon le cahier des charges
    - Sous-dimensionnement
      - Performances insuffisantes
      - Fiabilité aléatoire
      - Evolution onéreuse
    - Sur-dimensionnement
      - Sur-coût inutile
      - Réalisation parfois impossible
    - Exemple d'ATT avec les tables d'Erlangs
- Phase d'exploitation
  - Optimiser le système
  - Etudier le système sous des conditions critiques
  - Etudier l'évolution possible du système.

Exemple simple: étude de Idle RQ

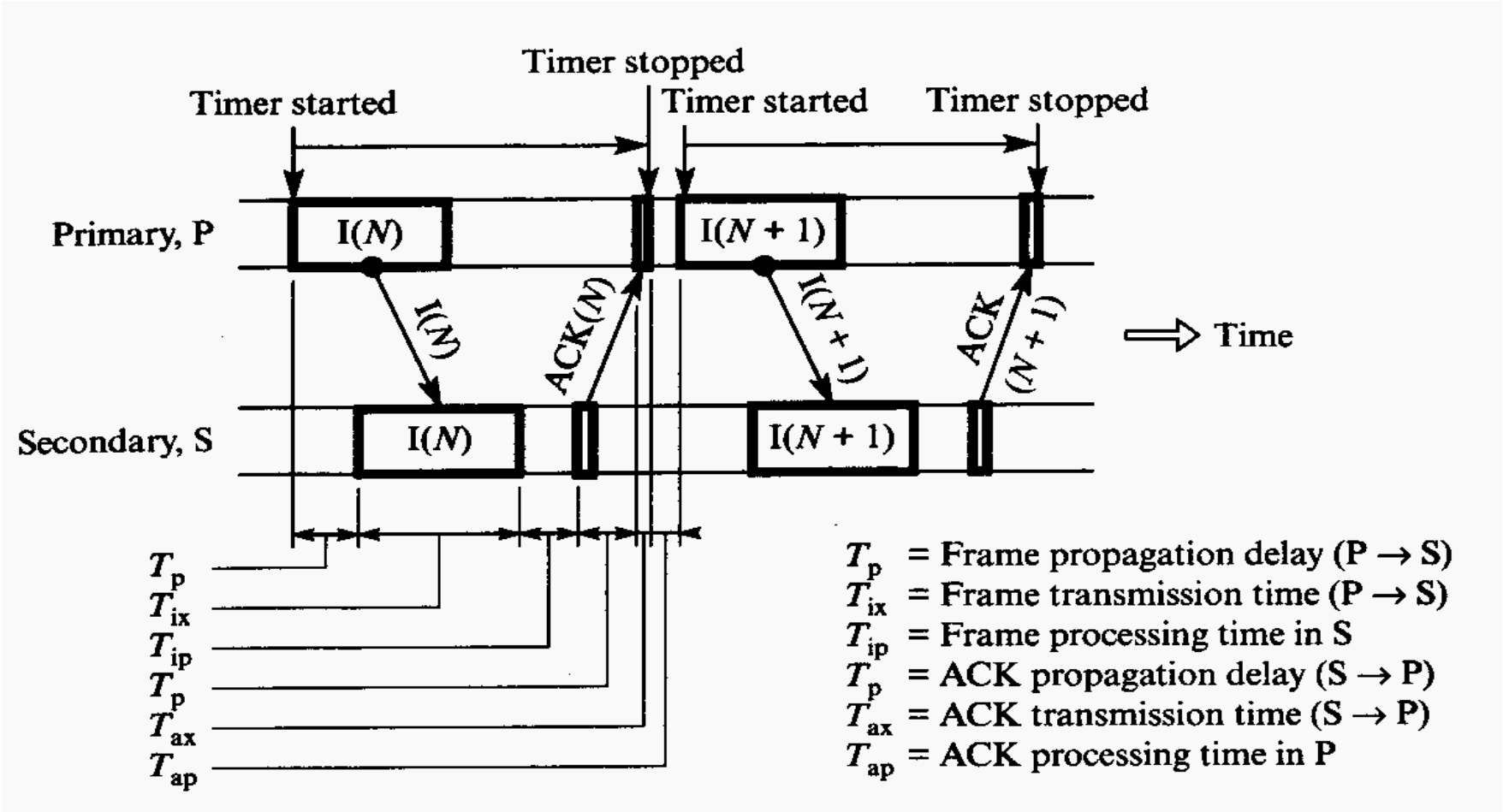
- De l'émetteur, une seule trame de données (I) non-acquitté à la fois (stop-and-go, send-and-wait)
- Le récepteur envoie un ACK pour chaque trame correcte



source L. Toutain

Idle RQ - éléments temporels

Auteur: PHAM Cong-Duc



Idle RQ - taux d'utilisation du lien parfait

- Quel est le taux d'utilisation d'un mécanisme comme Idle RQ?
  - le taux d'utilisation U est défini comme le rapport entre Tx (temps de transmission d'une trame) et Tt (temps d'attente pour un acquittement)
  - on utilisera Tp (temps de propagation sur le lien) et on négligera les temps de traitement (d'une trame et d'un ACK à la réception).

$$U = \frac{T_x}{T_x + 2T_p} = \frac{1}{1 + 2T_p / T_x}$$

Tp/Tx est souvent écrit a

Auteur: PHAM Cong-Duc

Idle RQ - taux d'utilisation du lien avec erreur

Auteur: PHAM Cong-Duc

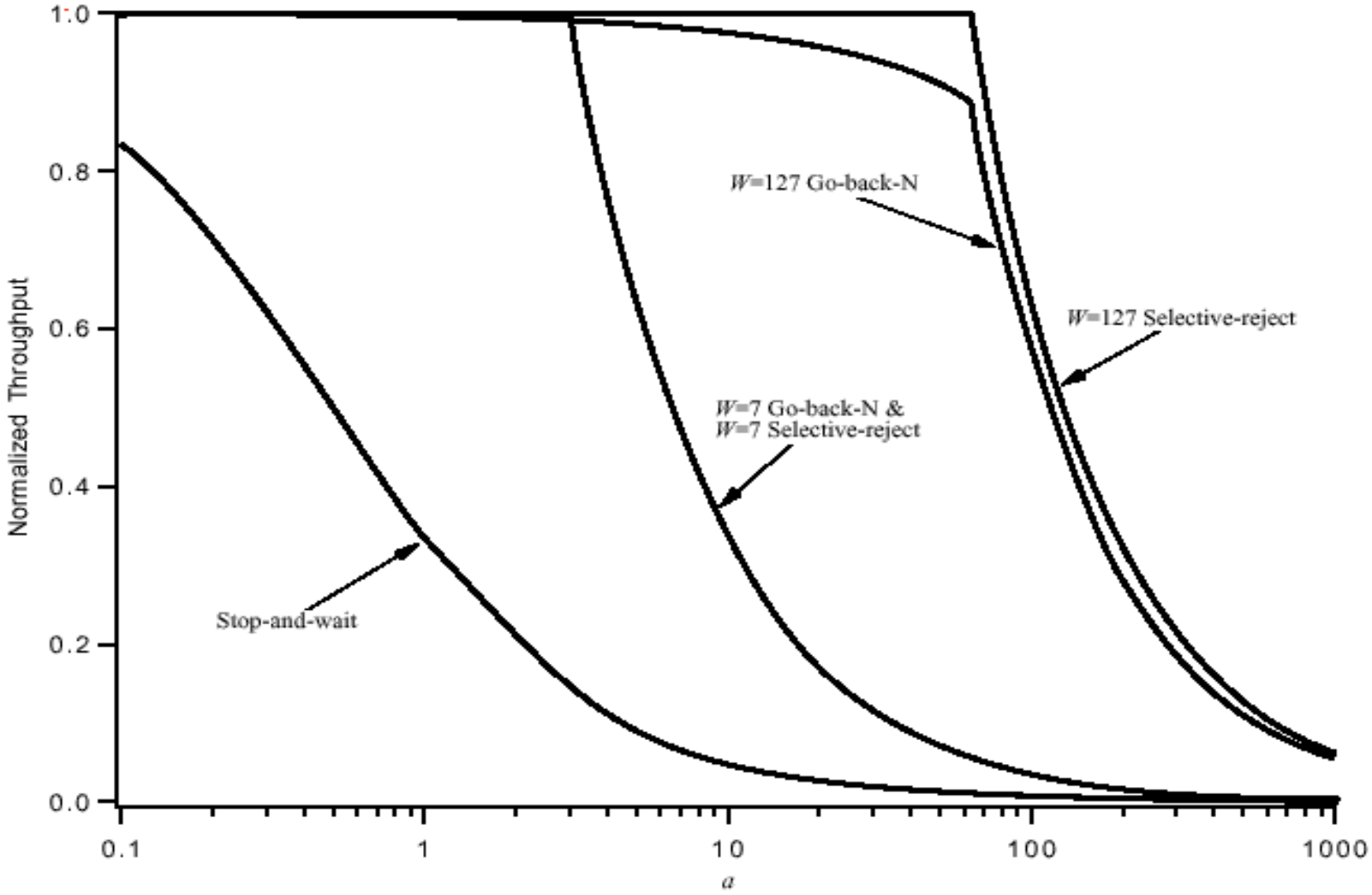
- On introduit P, le taux d'erreur bit (BER)
  - on introduira Nr le nombre moyen de retransmissions pour une trame et Ni la taille d'une trame en bit. De même, on négligera les temps de traitement (d'une trame et d'un ACK à la réception).

$$U = \frac{T_x}{N_r T_x + 2 N_r T_p} = \frac{1}{N_r \left(1 + 2T_p / T_x\right)}$$

  - proba qu'une trame soit juste, $(1-P)^{Ni}$
  - Nr = $1/(1-P)^{Ni}$

$$U = \frac{(1-P)^{Ni}}{(1+2a)}$$

Débit en fonction de *a=Tp/Tx*



ARQ Throughput as a Function of *a* (*P* = 10 $^{-3}$)

Les méthodes d'évaluation de performance

- Les principales méthodes quantitatives sont :
  - **La mesure**
    - Les sondes matérielles
    - Les sondes logicielles
  - **La simulation**
    - A événements discrets
    - Autres formes de simulation
  - **L'analyse opérationnelle**
  - **Les méthodes analytiques**
    - Les files d'attente

Auteur: PHAM Cong-Duc

La mesure

Auteur: PHAM Cong-Duc

- Elle demande l'existence d'un système ce qui réduit la classe des cas possibles.
- Par contre, les mesures de performances sont ceux du système réel et non ceux d'un modèle.
  - 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 (nbr d'appels système par exemple) mais introduisent des pertubations dans les mesures.

⇨ Problème de la collecte des information

⇨ Instrumentation lourde

⇨ Interprétation délicate des résultats

Le processus de modélisation

Auteur: PHAM Cong-Duc

- L'étude d'un système réel dans un environnement opérationnel est rarement réalisable (coût, difficulté).
- Le système peut ne pas encore exister !

  ⇨ On va représenter le fonctionnement d'un système de manière plus ou moins précise.

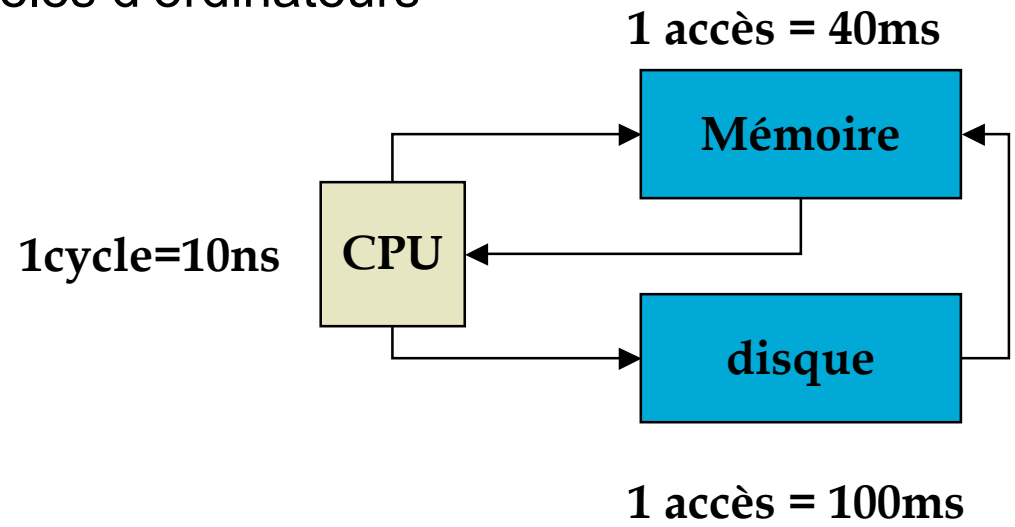
  ⇨ Pour cela on va s'appuyer sur des outils permettant d'approcher le comportement du système.

- Cette phase de substitution d'un système réel par un modèle se nomme la *modélisation.*
- *Modèle mathématique ou modèle logique*
- Cette étape, longtemps ignorée, s'impose de plus en plus.

Exemple de modèles

Auteur: PHAM Cong-Duc

- Modèles physiques de perturbations atmosphériques
- Modèles de propagation d'ondes
- Modèles de communications téléphonique. Il peut y avoir plusieurs variantes: fixe, mobile, satellites, ToIP…
- Modèles de protocoles
- Modèles d'ordinateurs



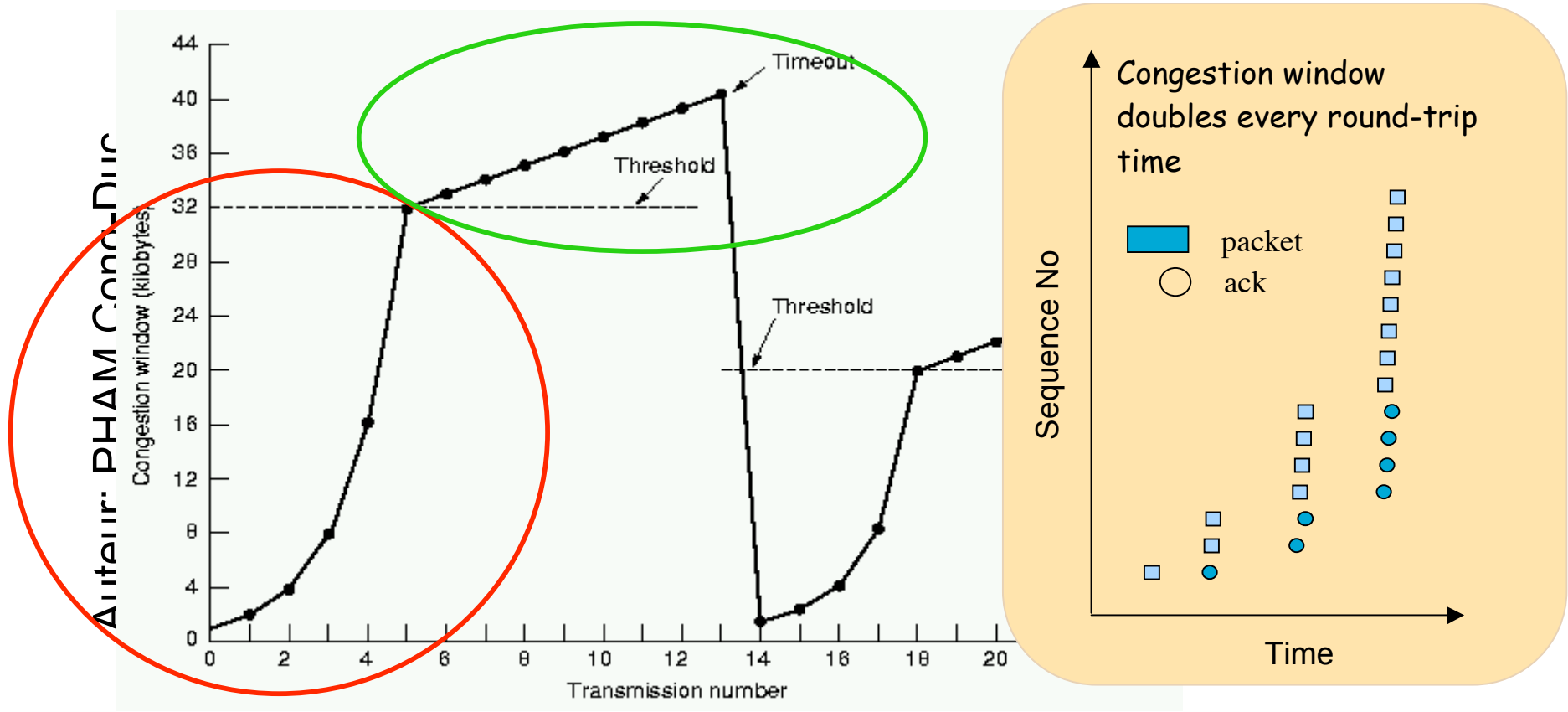
IL NE FAUT RETENIR QUE CE QUI EST IMPORTANT POUR LES RESULTATS QUE L'ON RECHERCHE

Comment feriez vous un modèle d'Ethernet?

- Quels sont les éléments importants/indispensable du protocole?
- Comment prendre en compte les collisions?

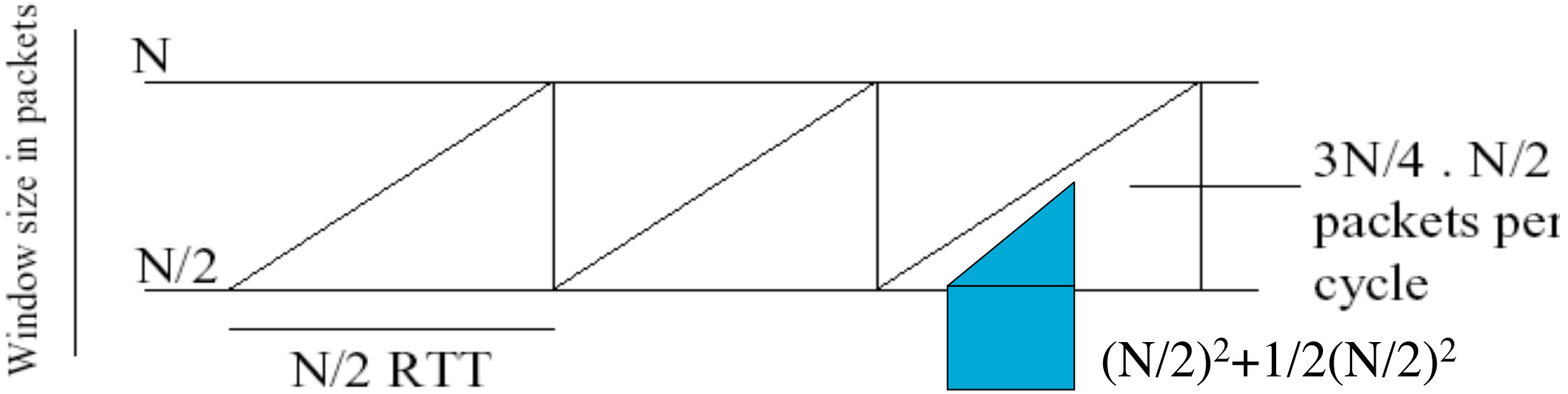
Auteur: PHAM Cong-Duc

Exemple de contrôle de congestion de TCP



- cwnd grows exponentially (slow start), then linearly (congestion avoidance) with 1 more segment per RTT
- If loss, divides threshold by 2 (multiplicative decrease) and restart with cwnd=1 packet

TCP throughput in steady state



Average window size (in packets) = W = 3N/4 , from (N+N/2)/2

Number of packets per cycle = 3N/4 . N/2 = $3N^2/8 = 1/p$

- Where p is the packet loss ratio (which should remain small enough)
- So $N = \sqrt{8/3p}$

Average throughput (in packets/sec) = B = W / RTT = 3N / 4 RTT

Average throughput (in bps) $= \sqrt{3/2}\,\frac{MTU}{RTT\sqrt{p}}$ $= \sqrt{3/2}\,\frac{1}{RTT\sqrt{p}}$

- MTU in bits

From Guy Leduc, RHDM 2002

La simulation

- Pas de contraintes contrairement à la mesure et aux méthodes analytiques. La simulation permet un niveau de détail arbitraire, mais on veut souvent faire trop précis.

- La simulation est très gourmande en temps de calcul et un compromis doit être trouver entre le niveau de détail et la pertinence des résultats (modèle du type "usine à gaz" à éviter).

- Les résultats ont une nature statistique avec laquelle il faut tenir compte (validation, intervalle de confiance)

Auteur: PHAM Cong-Duc

La simulation à événements discrets

Auteur: PHAM Cong-Duc

- On s'intéresse à des systèmes discrets où les changements se produisent à des instants particuliers
  - Les variables d'états qui définissent le système prennent des valeurs discrètes.
  - Arrivée de paquets, déclenchement d'alarme...
- Deux types d'approches existent pour la gestion du temps
  - Approche synchrone: évolution par incréments fixes.
  - Approche asynchrone: évolution par incréments variables.
- Deux visions existent pour l'écriture d'une simulation
  - Vision orientée événements.
  - Vision orientée processus.

Les outils de simulation

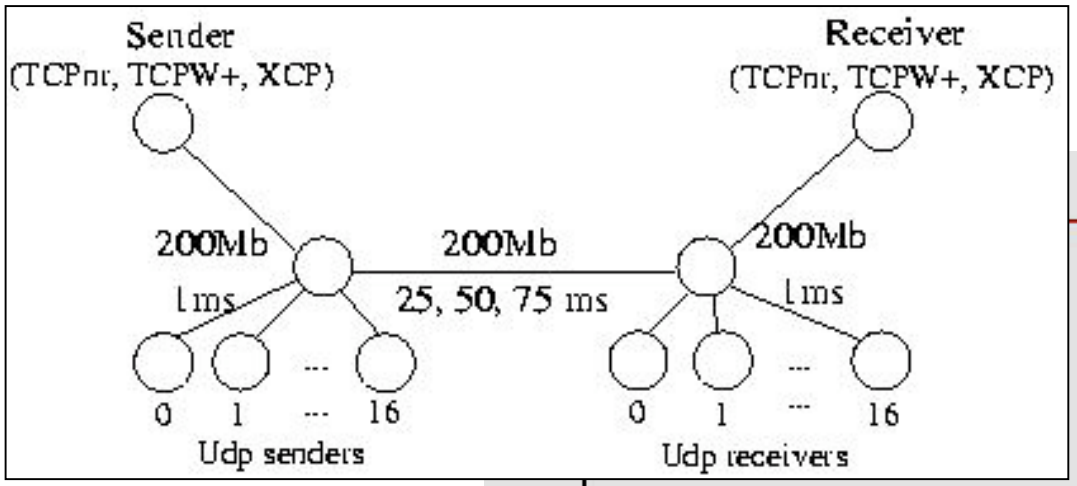
Auteur: PHAM Cong-Duc

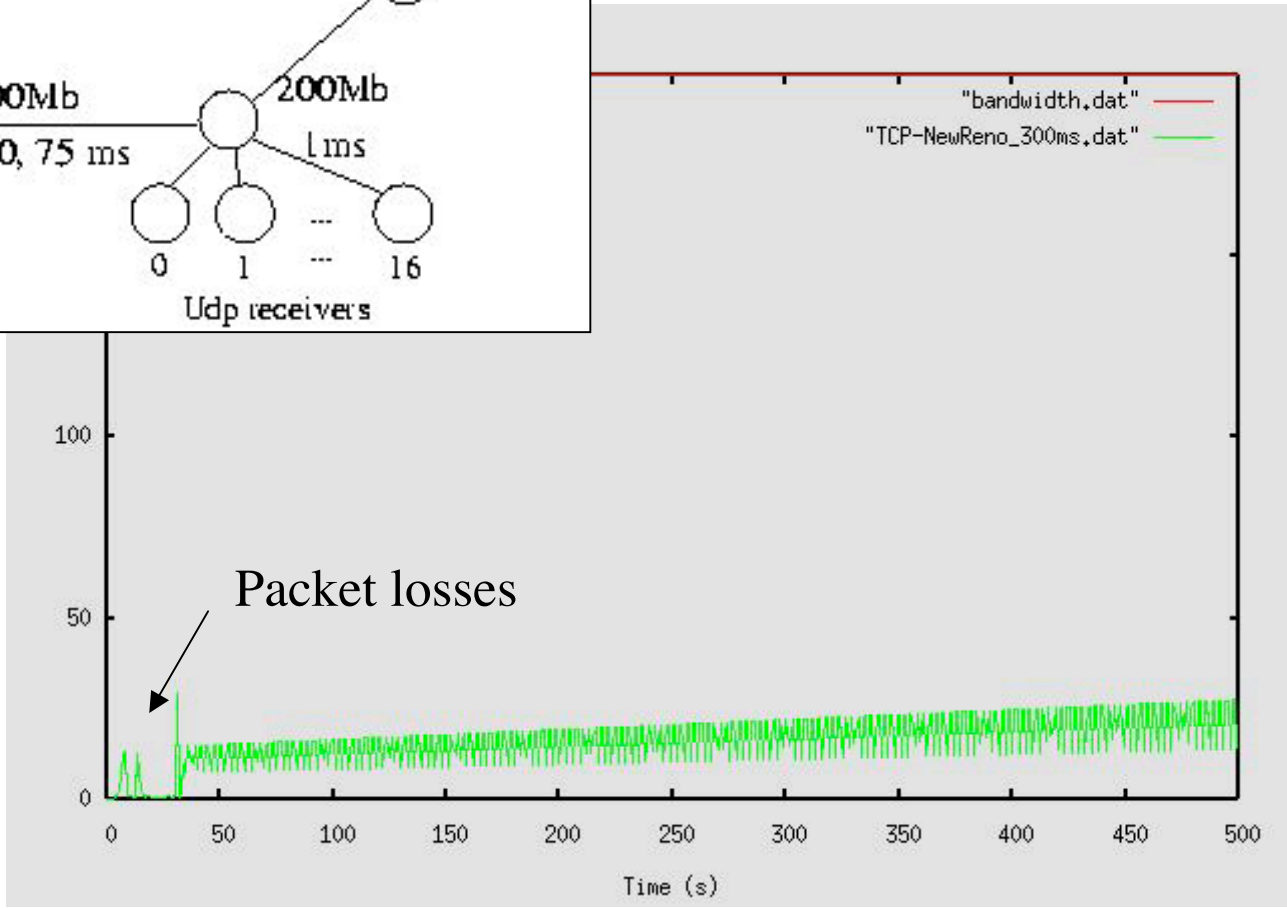
- Les langages spécialisés pour la simulation possèdent quelques avantages sur les langages généraux.
  - Simula
  - SIMSCRIPT II.5
  - GPSS
- Les logiciels de simulation sont plus conviviaux et nécessitent moins de programmation. Ils sont généralement dédiés à une classe d'applications.
  - QNAP2
  - OPNET
  - BONeS

Exemple de simulation de TCP

- Simulation de TCP NewReno avec ns (Network Simulator)

Auteur: PHAM Cong-Duc





Autres formes de simulation

Auteur: PHAM Cong-Duc

- Simulation de type Monte-Carlo
  - La simulation de type Monte-Carlo sont des simulations stocastiques utilisant des nombres aléatoires (U(0,1)).
- Simulation par trace (trace-driven)
  - La simulation *trace driven* consiste à réinjecter dans un modèle des valeurs de mesures.
- Simulation à temps continue
  - En simulation continue, le temps évolue de manière continue. On modélise souvent le système par un ensemble d'équation différentielles.
- Simulation hybride
  - Combinaison entre simulation et modèles analytiques.
- Simulation orientée objet
- Simulation parallèle

Conclusions sur la simulation

Auteur: PHAM Cong-Duc

- Avantages
  - C'est un outils indispensable pour évaluer les performances des systèmes complexes.
  - La simulation permet de répondre à des question de types "qu'est-ce qui se passe si..."
  - Le contrôle des expérimentations est plus grand sur un modèle que sur un système réel.
  - On peut étudier le système de manière très précise en changeant l'échelle du temps.

- Inconvénients
  - Une simulation ne fournit que des estimations de ce que l'on cherche.
  - Le modèle est généralement très lourd et requiert beaucoup de temps de développement.
  - Il faut définir des scénario bien précis qui précise les données à appliquer en entrée (fréquence d'arrivée, traffic perturbateur, corrélation temporelle…)

Historique sur l'analyse opérationelle

- L'analyse opérationelle a été appliquée bien 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 obsevations faites sur un système.
- Ces relations fondamentales sont vérifiées quelque soit le système et la période de mesure. Ces hypothèses se retrouvent en théorie des files d'attente sous l'aspect probabilistes.
- 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.

Auteur: PHAM Cong-Duc

Formule de Little

Auteur: PHAM Cong-Duc

- **T : Durée de la mesure**
- **A : Nbr total d'arrivée de requêtes**
- **D : Nbr total de départ de requêtes**
- **T(n) : Durée cumulée pendant laquelle le système a contenu n requêtes.**
- **nmax : nombre maximum de requêtes dans le système.**

On définit à partir de ces mesures les critères de performances suivants:

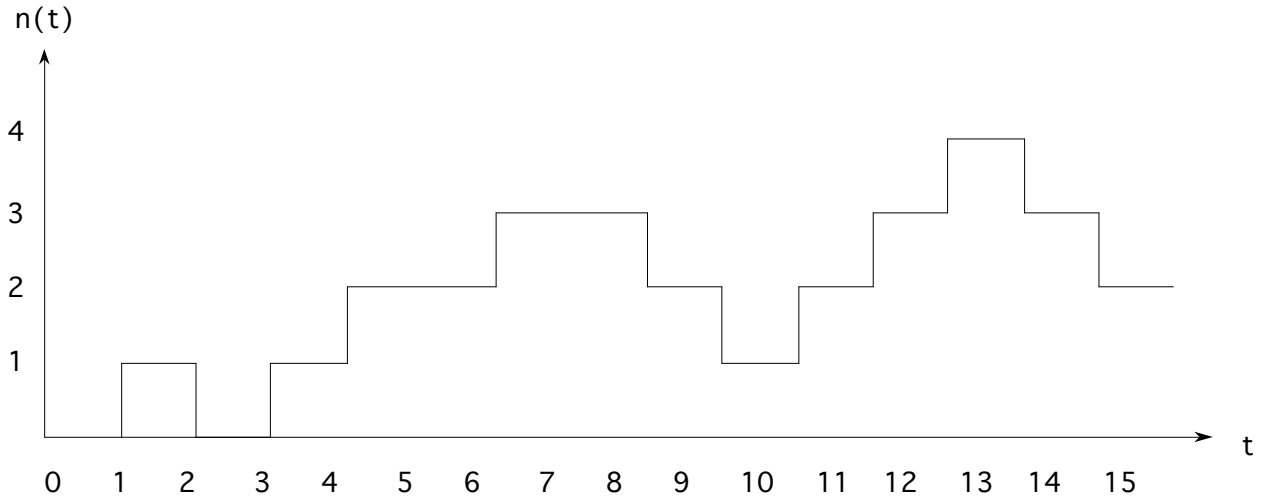
- Λ : débit du système à la sortie $\Lambda = \frac{D}{T}$
- L : nbr moyen de requêtes dans le système $L = \frac{\sum_{1}^{n\max} n.T(n)}{T}$
- R : temps de réponse du système $R = \frac{\sum_{1}^{n\max} n.T(n)}{D}$

➡ Formule de Little opérationnelle : $L=\Lambda.R$

Formule de Little : exemple

Auteur: PHAM Cong-Duc

Exemple d'évolution d'un système



T=15
A=7
D=5

T(0)=2
T(1)=3
T(2)=5
T(3)=4
T(4)=1

$$\Rightarrow \begin{cases} \Lambda = \dfrac{5}{15} \\ L = \dfrac{29}{15} \\ R = \dfrac{29}{5} \end{cases}$$

Temps de réponse d'un système interactif

Auteur: PHAM Cong-Duc

- On considère un système informatique accédé à partir d'un ensemble de terminaux.

- A chaque terminal est associé un processus unique passant alternativement par une phase de réflexion et une phase de traitement.

On définit les mesures suivantes :

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

Temps de réponse d'un système interactif

Auteur: PHAM Cong-Duc

On a

- R : temps de réponse moyen du système
- Z : temps de réflexion moyen du système
- Λ : débit en sortie du système

$$R = \sum_{k=1}^{N} r(k) \,/\, D$$

$$Z = \sum_{k=1}^{N} z(k) \,/\, A$$

$$\Lambda = \frac{D}{T}$$

$$r(k) + z(k) = T \quad \forall k \Rightarrow R.D + Z.A = N.T$$

$$d'\,où \quad R = \frac{N}{\Lambda} - \frac{A}{D} Z$$

$$en \;\; régime \;\; stationnaire \;\; \frac{A}{D} \approx 1$$

$$\Rightarrow \quad R = \frac{N}{\Lambda} - Z$$

Relation d'équilibre d'un système

Auteur: PHAM Cong-Duc

- On considère un système constitué de plusieurs stations de traitement. Les travaux envoyés dans le système engendrent un certain nombre de requêtes élémentaires sur chacunes des stations qui ne peuvent en traiter qu'une à la fois.
- Aucune autre hypothèse n'est faite sur le fonctionnement interne du système.

- On considère chaque station comme un sous-système indépendant et les mesures suivantes sont effectuées :

- **T** **Durée de la mesure**
- **D** **Nbr total de requêtes globales sorties du système**
- **D_i** **Nbr total de requêtes élémentaires traitées par la station i**
- **$T_i(n)$** **Durée cumulée pendant laquelle la station i a contenu n requêtes élémentaires.**

Relation d'équilibre d'un système

Auteur: PHAM Cong-Duc

- Λ_i : **débit de la station i**
- U_i : **taux d'occupation de la station i**
- S_i : **durée moyenne de service de la station i**
- e_i : **nbr moyen de visite à la station i par travail**
- R_i : **temps de réponse de la station i**
- L_i : **nbr moyen de requêtes élémentaires dans la station i**
- Λ : **débit global du système**

$$\Lambda_i = D_i / T$$

$$U_i = (T - T_i(0)) / T$$

$$S_i = (T - T_i(0)) / D_i$$

$$e_i = D_i / D$$

$$R_i = \frac{\sum n.T_i(n)}{D_i}$$

$$L_i = \frac{\sum n.T_i(n)}{T}$$

$$\Lambda = \frac{D}{T}$$

$$\Lambda = \frac{\Lambda_i}{e_i} = \frac{U_i}{S_i.e_i} = \frac{L_i}{R_i.e_i}$$

Th de Chang-Lavenberg
(version *opérationnelle*)

Relation d'équilibre d'un système

- L'expression locale de la formule de Little s'obtient alors :

$$L_i = R_i . \Lambda_i$$

- Si on suppose qu'un travail ne peut génèrer simultanément plusieurs requêtes, on obtient bien la formule de Little globale.

> Toutes les relations précédentes peuvent être appliquées à des populations distinctes de travaux (batch, interactif...). Il suffit de restreindre les mesures aux requêtes issues de chaque population. Au niveau d'une station, il y a additivité des débits et des taux d'occupation.

Auteur: PHAM Cong-Duc

Etude de la saturation d'un système

Auteur: PHAM Cong-Duc

- Un système est dit saturé si au moins un de ses sous-système l'est.
- Si on suppose que le produit $S_i.e_i$ est invariant, c'est à dire que les services globaux sur chaque station sont indépendants de la charge, on peut calculer le débit maximum du système.
- Le taux d'occupation du sous-système saturé est 1, le taux des autres stations est donné par :

$$\Lambda_{\max} = \frac{1}{S_i^*.e_i^*},\ S_i^*.e_i^* = \max(S_i.e_i)$$

$$U_i = \frac{S_i.e_i}{S_i^*.e_i^*} < 1$$

Limite de l'analyse opérationnelle

Auteur: PHAM Cong-Duc

- L'analyse opérationnelle a permit d'introduire de manière simple quelques critères de performance en se basant uniquement sur des observations.
- Maintenant si on désire connaître, par exemple, le temps de réponse d'une station, connaissant le débit d'arrivé Λ_A et le temps moyen de service S, on en est incapable.
- Il est nécessaire d'étudier plus finement les interactions entre arrivées et service.
  - ⇨ On va introduire des hypothèses de nature statistique sur le comportement des requêtes.
  - ⇨ La théorie des files d'attente fournit des résultats utilisables dans un grand nombre de situations.