Files d'attente et modèles de Trafic

Alain Jean-Marie
LIRMM
161 Rue Ada
34392 Montpellier Cedex 5
ajm@lirmm.fr

8 décembre 1999
Deuxième École d'Hiver des Télécommunications
de Sophia-Antipolis

Plan

Introduction: Réseaux, contention, attente et pertes

Partie I: Processus stochastiques

- Stationarité, autocorrélation
- Chaînes de Markov

Partie II: La théorie des files d'attente

- Paramètres, Notation de Kendall
- Mesures de performances
- Modèles stochastiques de trafic

Partie III: Analyse exacte

- Analyse exacte dans le cas de tampons finis
- Analyse exacte dans le cas infini
- Réseaux de files d'attente

Partie IV: Analyse asymptotique

- Bornes et asymptotiques exponentielles
- Mémoire longue, autosimilarité, sous-exponentialité

Partie V: Modèles déterministes

- Enveloppes de trafic et bornes (σ, ρ)
- Régulateurs de trafic, courbes de service

Introduction

Dans un réseau de communication orienté datagramme, des files d'attente se créent tout le long du chemin de communication. Gestion du multiplexage statistique et de la contention.

Ces files créent attente et pertes .

Le problème est de savoir les quantifier.

On se place dans un cadre stochastique étant donnée la nature incertaine du trafic.

La théorie des files d'attente: un ensemble de concepts, d'outils et de résultats pour aborder ces problèmes.

Recherche de résultats permettant de définir puis garantir la fameuse *qualité de service*.

Partie I: Processus stochastiques

- Variables aléatoires
- Processus aléatoires
- Stationarité, ergodicité
- Covariance, autocorrélation
- Chaînes de Markov

Variables aléatoires

Espace probabilisé: Ω ensemble de *trajectoires* ou *réalisations*.

Variable aléatoire X: fonction de l'espace des trajectoires Ω dans un espace de valeurs.

Distribution:

$$\mathbb{P}\{X \leq x\} \;=\; \mathbb{P}\{\omega \mid X(\omega) \leq x\} \;.$$

Espérance, variance:

$$\begin{aligned} \mathbb{E}X &= \int x \mathrm{d}\mathbb{P}\{X \leq x\} \\ \mathrm{Var}(X) &= \int x^2 \mathrm{d}\mathbb{P}\{X \leq x\} \;-\; \mathbb{E}X^2 \end{aligned}$$

Si la variable aléatoire est discrète:

$$\begin{aligned} \mathbb{E}X &= \sum_n n\, \mathbb{P}\{X = n\} \\ \mathrm{Var}(X) &= \sum_n n^2\, \mathbb{P}\{X = x\} \;-\; \mathbb{E}X^2 \end{aligned}$$

Variance: mesure de la variabilité de X autour de sa moyenne.

Covariance de deux v.a.:

$$\mathrm{Cov}(X,Y) \quad = \quad \mathbb{E}(XY) - \mathbb{E}X \ \mathbb{E}Y \ .$$

Mesure de la dépendance de X et Y. Si X et Y sont *indépendantes*, $\mathrm{Cov}(X,Y) = 0$.

Transformée de Laplace (Laplace-Stiltjes) de X:

$$X^*(s) \ = \ \int_0^\infty e^{-st} \mathrm{d}\mathbb{P}\{X \leq t\} \ = \ \mathbb{E}(e^{-sX}) \ .$$

Fonction génératrice d'une variable aléatoire discrète:

$$X^*(z) \ = \ \sum_{n=0}^{\infty} z^n \ \mathbb{P}\{X = n\} \ = \ \mathbb{E}(z^X) \ .$$

Loi d'addition: si $X \perp\!\!\!\perp Y$ alors,

$$(X+Y)^*(s) \ = \ X^*(s) \ Y^*(s) \ .$$

Processus stochastiques

Un processus stochastique "vit" dans un espace d'état $\mathcal{E}$.

Deux catégories:

en temps discret	$\{X_n, n \in \mathbb{Z}\}$
en temps continu	$\{X(t), t \in \mathbb{R}\}$

Temps discret: suite de variables aléatoires.

Temps continu: famille de fonctions aléatoires $\omega \mapsto X(t;\omega)$.

Exemples classiques:

- suite de tirages Bernoulli (Pile ou Face) indépendants:

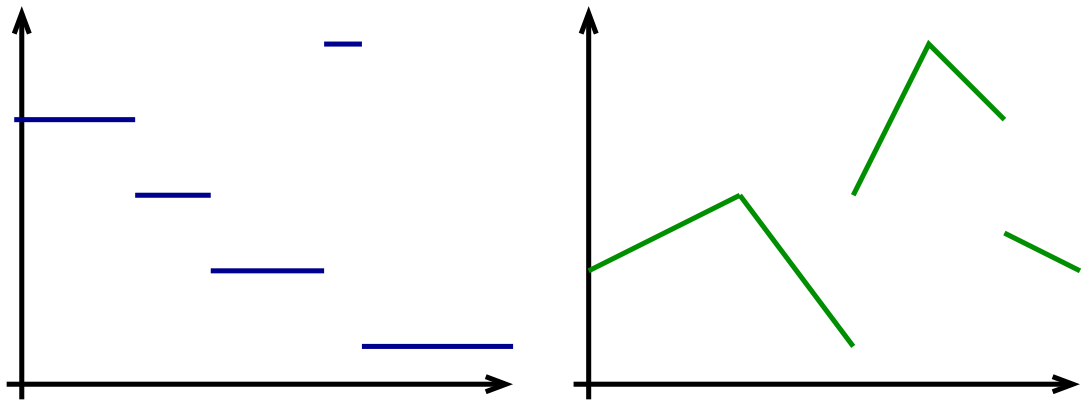
$$X_n = 0 \quad \text{avec proba } 1/2, \quad X_n = 1 \quad \text{avec proba } 1/2.$$

- mouvement Brownien: $\{X(t), t \in \mathbb{R}\}$ tel que

$$X(s+t) - X(t) \sim \mathcal{N}(0, \sigma t) \ .$$

Événements discrets

Dans le domaine des systèmes d'information (calculateurs, réseaux) on travaille avec des *systèmes à événements discrets* tels que $X(t)$ ou $\dot{X}(t)$ est constant par morceaux.



Processus à événements discrets

Modèles mathématiques de cette situation:

- Processus d'arrivée d'événements: Processus Ponctuels (Baccelli, Bremaud).
- Plus généralement: dynamique déterministe + sauts aléatoires dans le temps et dans l'espace
  $\Rightarrow$ PDP = Piecewise Deterministic Processes, processus (Markoviens) déterministes par morceaux (Davis).

Stationarité

Stationarité au sens strict: $X(\cdot) = X(\cdot + s)$ en distribution.

En particulier, $\mathbb{E}f(X(t_1)) = \mathbb{E}f(X(t_2))$

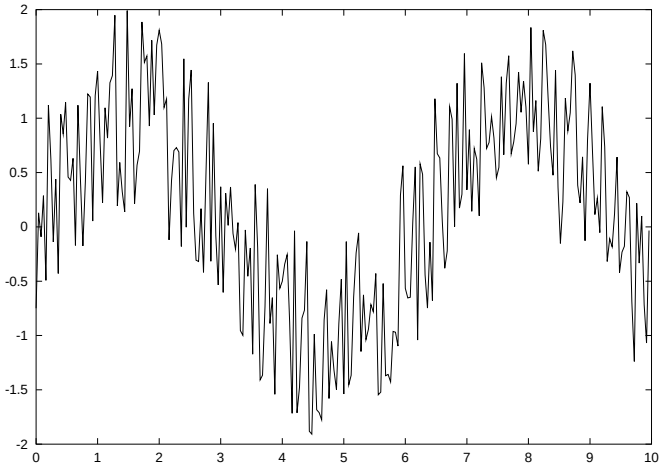
La stationarité exclut les périodicités. Exemple:

$$X(t) \;=\; \sin(t) \;+\; \xi_t$$

avec ξ_t aléatoire et petit.

Alors

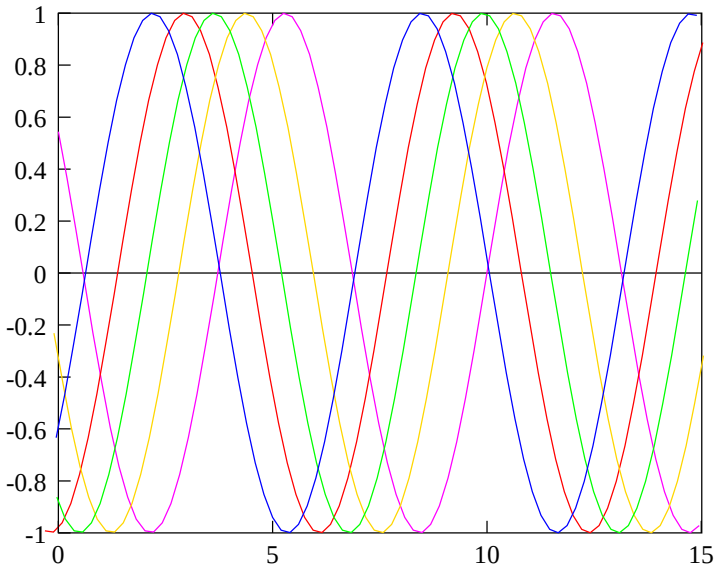
$$\mathbb{P}\{X(t+\pi) > 0\} \neq \mathbb{P}\{X(t) > 0\} \; .$$



Trajectoire de $sin(t) + U(-1, 1)$

Mais il existe des processus essentiellement périodiques stationnaires:

$$X(t) \;=\; sin(t+\xi), \quad \xi \sim U(0,\pi)\ .$$



Trajectoires de $sin(t+\xi(\omega))$

Convergence

Un processus n'est en général pas stationnaire mais il peut le devenir:

$$\begin{aligned} X(t) &\to X, \quad t \to \infty \\ X_n &\to X, \quad n \to \infty \end{aligned}$$

en distribution (ou autrement...).

Si pour tout s,

$$X[t, t+s] \to \hat{X}[0, s], \text{ en distribution} \qquad t \to \infty.$$

Le processus converge vers un *état stationnaire*.

Si la convergence est assez rapide, on peut utiliser la distribution de X comme approximation de celle de $X(t)$.

Ergodicité

Ergodicité: coïncidence des moyennes spatiales et temporelles:

$$\mathbb{E}f(X(s)) = \lim_{T\to\infty} \frac{1}{T} \int_0^T f(X(t))\mathrm{d}t \ ,$$

$$\mathbb{E}f(X(n)) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(X_n) \ .$$

Application: la loi des grands nombres pour les *estimateurs statistiques* de quantités.

Il existe des processus stationnaires mais non ergodiques.

Autocorrélations

Autocorrélation:

$$R(s, s+t) \;=\; \mathbb{E}[X(t)\; X(t+s)]\;.$$

Autocovariance: dépendance de l'état de X à la date $t+s$ par rapport à l'état à t.

$$h(t, t+s) \;=\; \mathrm{cov}(X(t)\; X(t+s)) \;=\; \mathbb{E}[X(t)\; X(t+s)] \;-\; \mathbb{E}X(t)\; \mathbb{E}X(t+s)\;.$$

Si $X(t) \perp\!\!\!\perp X(t+s)$, alors $h(t, t+s) = 0$.

Définition: X stationnaire au sens large (ou à l'ordre deux): pour tout t:

$$h(t, t+s) \;=\; h(s) \;=\; \mathbb{E}X(0)\; \mathbb{E}X(s) \;-\; (\mathbb{E}X)^2\;.$$

Note: $h(0) \;=\; \mathrm{Var}(X)$.

Mémoire

Autocorrélation totale:

$$\text{temps continu} \quad \int_0^\infty |h(s)| \, \mathrm{d}s \qquad \text{temps discret} \quad \sum_{n=0}^{\infty} |h(n)| \; .$$

Un processus est à mémoire courte si

$$\int_0^\infty |h(s)| \, \mathrm{d}s \; < \; \infty.$$

Sinon il est à *mémoire longue*.

Mémoire longue $\Rightarrow$ décroissance lente de la dépendance de $X(t+s)$ et $X(t)$.

Les chaînes de Markov

$\{X(n), n \in \mathbb{N}\}$ est une **chaîne de Markov en temps discret** homogène si:

i/ (propriété de Markov) $\forall t \in \mathbb{N}$, et $\forall \ (j_0, j_1, \ldots, j_t, j_{t+1}) \in \mathcal{E}^{t+2}$:

$$\mathbb{P}\{X(t+1) = j_{t+1} | X(t) = j_t, \ldots, X(0) = j_0\} \ = \ \mathbb{P}\{X(t+1) = j_{t+1} | X(t) = j_t\} \ ;$$

ii/ (homogénéité) $\forall t \in \mathbb{N}$, et $(i, j) \in \mathcal{E} \times \mathcal{E}$,

$$\mathbb{P}\{X(t+1) = j | X(t) = i\} \ = \ P_{i,j} \ .$$

$P_{i,j}, (i, j) \in \mathcal{E} \times \mathcal{E}$: probabilités de transition

$\mathbf{P}$ *matrice de transition.*

Dynamique des probabilités

On cherche les *probabilités de transition à n pas*:

$$p(i,j;n) \;=\; \mathbb{P}\{X(n)=j \mid X(0)=i\}\;,$$

Soit $P(n)$ la matrice des $p(i,j;n)$. Alors:

$$\boxed{P(n) \;=\; \mathsf{P}^n.}$$

Soit maintenant, pour $n \in \mathbb{N}$ et $j \in \mathcal{E}$,

$$\pi_n(j) \;=\; \mathbb{P}\{X(n)=j\}.$$

On a alors:

$$\pi_n(j) \;=\; \sum_{i\in\mathcal{E}} \pi_0(i)\; p(i,j;n)\;.$$

Sous forme matricielle: Pour tout $n \in \mathbb{N}$:

$$\boxed{\boldsymbol{\pi}_n \;=\; \boldsymbol{\pi}_0\, \mathsf{P}^n.}$$

Exemple de chaîne de Markov

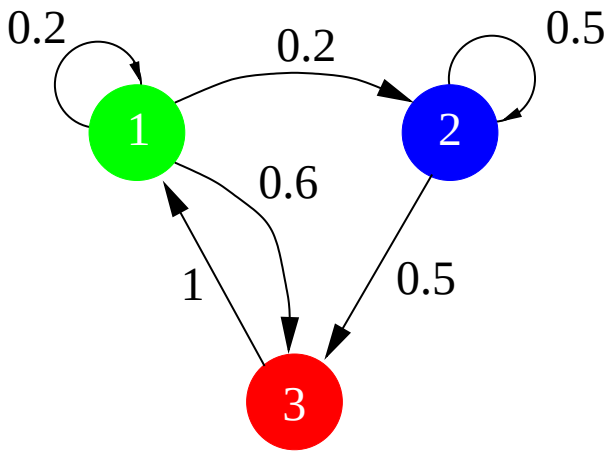


Diagramme de transition

Matrice de transition:

$$\mathsf{P} = \begin{pmatrix} 0.2 & 0.2 & 0.6 \\ 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \end{pmatrix} .$$

Vecteurs de probabilité:

$$\begin{aligned} \mathbf{p}_0 &= (1, 0, 0) \\ \mathbf{p}_1 &= (0.2 \quad 0.2 \quad 0.6) \\ \mathbf{p}_2 &= (0.64 \quad 0.14 \quad 0.22) \\ \mathbf{p}_3 &= (0.348 \quad 0.198 \quad 0.454) \\ \mathbf{p}_4 &= (0.5236 \quad 0.1686 \quad 0.3078) \\ \vdots \quad &\quad \vdots \quad \vdots \\ \mathbf{p}_\infty &= (5/11, 2/11, 4/11) \end{aligned}$$

Équations d'équilibre

Si $\lim_n \boldsymbol{\pi}_n = \boldsymbol{\pi}$ existe, alors:

$$\boldsymbol{\pi} \quad = \quad \boldsymbol{\pi}\, \mathsf{P}\ .$$

Ces équations d'équilibre s'écrivent: $\forall i \in \mathcal{E}$,

$$\pi(i) \ = \ \sum_{j \in \mathcal{E}} \pi(j)\ P_{j,i}\ .$$

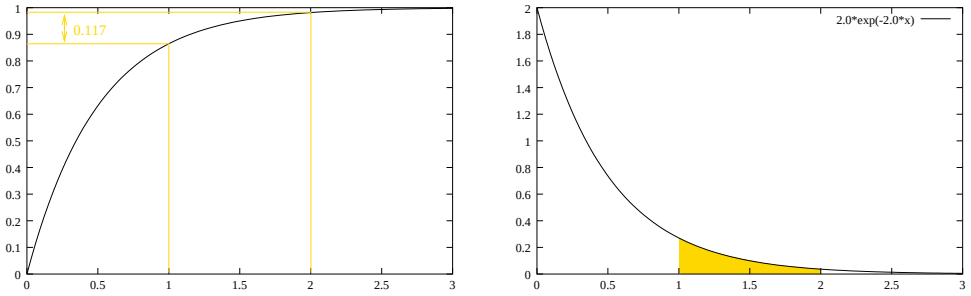
Le calcul des probabilités stationnaires se réduit à la résolution d'un système linéaire!

Problème: il est souvent très grand voire infini.

Le temps continu

Une variable aléatoire X a une distribution *exponentielle* de paramètre $\lambda > 0$ ($X \sim \text{Exp}(\lambda)$) si:

$$F_X(x) \;=\; \mathbb{P}\{X \leq x\} \;=\; 1 \;-\; e^{-\lambda x}\;.$$



Fonction de répartition et densité de la variable exponentielle

La distribution exponentielle est *sans mémoire*: $\forall s, t > 0$,

$$\mathbb{P}\{X > s + t \mid X > s\} \;=\; \mathbb{P}\{X > t\}\;.$$

La famille des distributions exponentielles est stable pour la minimisation: si $X_1 \sim \text{Exp}(\lambda_1)$, $X_2 \sim \text{Exp}(\lambda_2)$ et X_1 et X_2 sont indépendantes alors $\min\{X_1, X_2\} \sim \text{Exp}(\lambda_1 + \lambda_2)$.

Le processus de Poisson

Soit une suite aléatoire $T_0 \leq T_1 \leq \ldots \leq T_n \leq T_{n+1} \leq \ldots$.
Le processus de comptage:

$$N(a,b) \;=\; \#\{n \mid a \leq T_n < b\} \;=\; \sum_{n=0}^{\infty} \mathbf{1}_{\{a \leq T_n < b\}}$$

est un **Processus de Poisson** de paramètre λ si $\{T_{n+1} - T_n\}$ est une suite i.i.d. de variables Exp(λ).

Pour tout u:

$$\mathbb{P}\{N(x, x+u) \;=\; k\} \;=\; \frac{(\lambda u)^k}{k!} e^{-\lambda u} \;.$$

En particulier, $\mathbb{E}N(x, x+u) = \lambda u$: λ est le *taux d'arrivée* du processus.

Théorème limite: si on superpose un grand nombre de processus rares, le processus résultant est asymptotiquement Poisson.

Chaînes de Markov en temps continu

Soit $\{X(t), t \in \mathbb{R}^+\}$, ayant les propriétés suivantes. Quand X entre dans l'état i:

- X reste dans l'état i un temps aléatoire, exponentiellement distribué avec paramètre τ_i, indépendemment du passé; puis
- X saute instantanément dans l'état j avec probabilité p_{ij}. On a $p_{ij} \in [0, 1]$, $p_{ii} = 0$ et

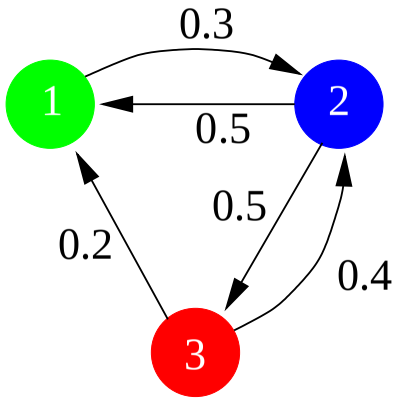
$$\sum_j p_{ij} = 1.$$

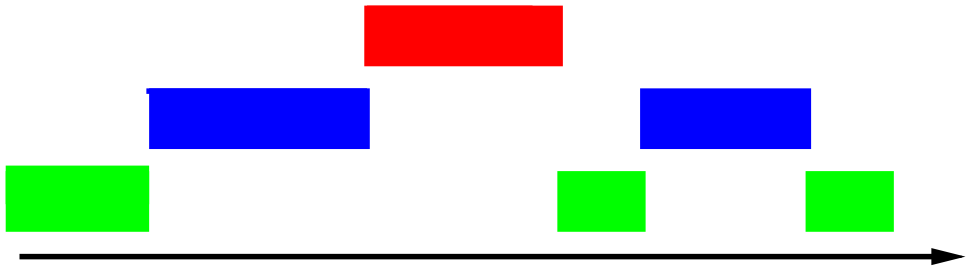
La loi exponentielle étant *sans mémoire*, on obtient un processus qui a la propriété que:

$$\begin{aligned} &\mathbb{P}\{X(t_{n+1}) = j_{n+1} | X(t_n) = j_n, \ldots, X(t_0) = j_0\} \\ &= \quad \mathbb{P}\{X(t_{n+1}) = j_{n+1} | X(t_n) = j_n\} \ . \end{aligned}$$

Exemple

$$\tau = \begin{pmatrix} 0.3 \\ 1 \\ 0.6 \end{pmatrix} \qquad \mathsf{P} = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} & 0 \end{pmatrix} \qquad \mathsf{Q} = \begin{pmatrix} -0.3 & 0.3 & 0 \\ 0.5 & -1.0 & 0.5 \\ 0.2 & 0.4 & -0.6 \end{pmatrix}.$$





La définition

Un processus $\{X(t), t \in \mathbb{R}^+\}$ est une chaîne de Markov en temps continu (ou un *processus de Markov*) homogène si et seulement si:

i/ (propriété de Markov) pour tout $n \in \mathbb{N}$, tout $n+2$– uplet de réels $t_0 < t_1, < \ldots < t_n < t_{n+1}$ et tout $n+2$– uplet $(j_0, j_1, \ldots, j_n, j_{n+1})$ d'éléments de $\mathcal{E}$:

$$\begin{aligned} &\mathbb{P}\{X(t_{n+1}) = j_{n+1} | X(t_n) = j_n, \ldots, X(t_0) = j_0\} \\ = \quad &\mathbb{P}\{X(t_{n+1}) = j_{n+1} | X(t_n) = j_n\} \ ; \end{aligned}$$

ii/ (homogénéité) pour tout réels s, t et u, et toute paire (i, j) de $\mathcal{E}$, indépendamment de t on a:

$$\mathbb{P}\{X(t+u) = j | X(s+u) = i\} \ = \ \mathbb{P}\{X(t) = j | X(s) = i\} \ = \ P_{t-s}(i, j) \ .$$

Dynamique des probabilités

Équations de Chapman-Kolmogoroff

$$P_{t+s}(i,j) \;=\; \sum_{k\in\mathcal{E}} P_t(i,k)\; P_s(k,j)\;,$$

ou encore, sous forme matricielle,

$$\mathsf{P}_{t+s} \;=\; \mathsf{P}_t\,\mathsf{P}_s\;,$$

Si le processus $\{X(t)\}$ est suffisamment "régulier", alors il existe une matrice $\mathsf{Q} = \mathsf{P}'(t)$ telle que:

$$\frac{d\mathsf{P}_t}{dt} \;=\; \mathsf{Q}\mathsf{P}_t \;=\; \mathsf{P}_t\mathsf{Q}\;.$$

C'est le *générateur infinitésimal*.

Alors:

$$\boxed{\mathsf{P}_t \;=\; e^{t\mathsf{Q}}\;, \qquad \mathsf{p}_t \;=\; \mathsf{p}_0\mathsf{P}_t \;=\; \mathsf{p}_0 e^{t\mathsf{Q}}\;.}$$

Construction des générateurs

Sous les hypothèses d'évolution ci-dessus: le processus $\{X(t), t \in \mathbb{R}^+\}$ est une CMTC de générateur infinitésimal:

$$\begin{cases} q(i,j) &= \tau_i p_{ij} \quad \text{si } i \neq j \\ q(i,i) &= -\tau_i \ . \end{cases}$$

Équations d'équilibre

Si $\lim_t \boldsymbol{\pi}_n = \boldsymbol{\pi}$ existe, alors:

$$\boxed{\mathbf{0} \quad = \quad \boldsymbol{\pi}\, \mathbf{Q}\ .}$$

Ces équations d'équilibre s'écrivent: $\forall i \in \mathcal{E}$,

$$(\sum_{j \neq i} q_{i,j})\pi(i) \ = \ \sum_{j \neq i} \pi(j) q_{j,i}\ .$$

Interprétation: flux entrant = flux sortant.

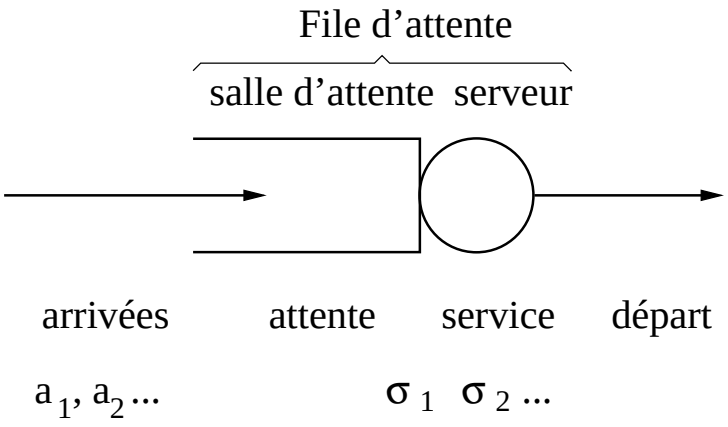
Généralisation: équations d'équilibre global. Pour $S \subset \mathcal{E}$:

$$\sum_{i \in S, j \in \overline{S}} \pi(i)\ q_{i,j} \ = \ \sum_{i \in \overline{S}, j \in S} \pi(i) q_{i,j}\ .$$

Partie II: La théorie des files d'attente

- Files discrètes, files fluides
- Processus d'arrivée, processus des services; notation de Kendall
- Mesures de performances: nombre de clients, temps d'attente/réponse, probabilité de perte, gigue
- Dynamique de la file; courbes de charge; Équations de la dynamique
- Capacité: finie ou infinie?
- Files simples ou réseaux?
- Modèles stochastiques de trafic
  - Processus i.i.d.
  - Processus de Poisson
  - Processus modulés par Markov

Les files d'attente



Représentation habituelle d'une file d'attente

Les éléments constitutifs d'une file d'attente sont:

- Un ou plusieurs serveurs
- Une salle d'attente
- (éventuellement) des classes de clients
- Un processus d'arrivée par classe
- Un processus des durées de service des clients
- Une discipline de service

Notation de Kendall

Cette notation permet de s'y retrouver dans la grande variété de possibilités.

Un modèle de file d'attente se note par:

$$A/S/P/K/D$$

A la loi des inter-arrivées
S loi des services
P nombre de serveurs
K taille de la salle d'attente (par défaut: ∞)
D discipline de service (par défaut: FIFO)

Exemples: la file M/M/1, M/GI/1/K, etc.

Mesures de performance

Condition(s) de stabilité Sous quelles conditions la file d'attente admet un état stationnaire?
$X(t)$ quantité dynamique:

$$\lim_{t\to\infty} \mathbb{P}\{X(t) \leq x\} \;=\; ?$$

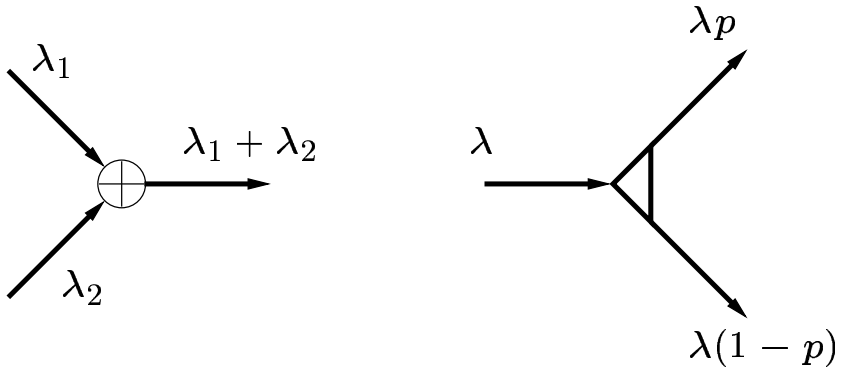
Débit Si $N(a,b)$ mesure le nombre d'arrivées dans $[a,b[$, le débit des arrivées est:

$$\lambda \;=\; \text{limsup}_{t\to\infty}\, \frac{N(0,t)}{t} \;=\; \mathbb{E}N(0,1) \;=\; \text{limsup}_{n\to\infty}\, \frac{n}{a_n}\,.$$

Si les dates de départ des clients sont $d_1,\ldots,d_n,\ldots$, le débit de sortie est:

$$\theta \;=\; \text{limsup}_{n\to\infty}\, \frac{n}{d_n}\,.$$

Les débits se conservent:



Lois de conservation des débits

Si stabilité:

$$\lambda \;=\; \theta \;.$$

Taux d'utilisation Fraction du temps d'utilisation de la ressource:

$$\rho \;=\; \text{limsup}_{T\to\infty}\, \frac{U(0,T)}{T} \;.$$

Temps de réponse $R_n = d_n - a_n$.

Également: temps d'attente W_n et temps de service σ_n:

$$R_n \;=\; W_n \;+\; \sigma_n \;.$$

Taux/probabilité de perte Fraction de clients "perdus". Rapport du débit "efficace" au débit "offert".

Temps de cycle Pour les systèmes en boucle.

Gigue Mesure de variabilité de la réponse du réseau:

$$J_n \;=\; |(d_{n+1} - d_n) - (a_{n+1} - a_n)| \;=\; |R_n - R_{n+1}| \;.$$

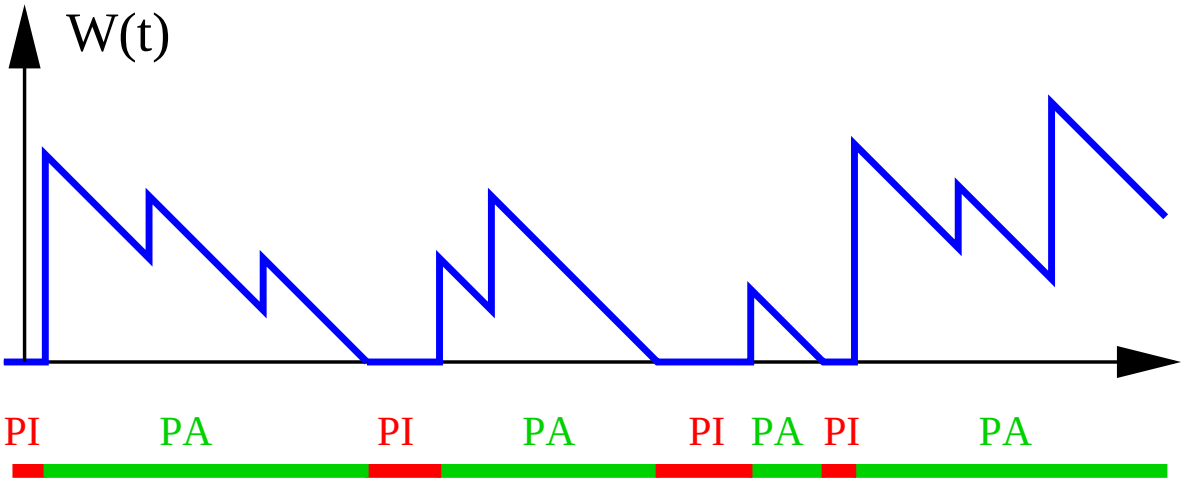
Dynamique d'une file d'attente

Quantités fondamentales:

$N(t)$ nombre de clients présents dans la file à la date t;

$W(t)$ quantité de travail présente dans la file à la date t (workload, backlog).

Évolution de $W(t)$: la courbe de charge.



Une courbe de charge

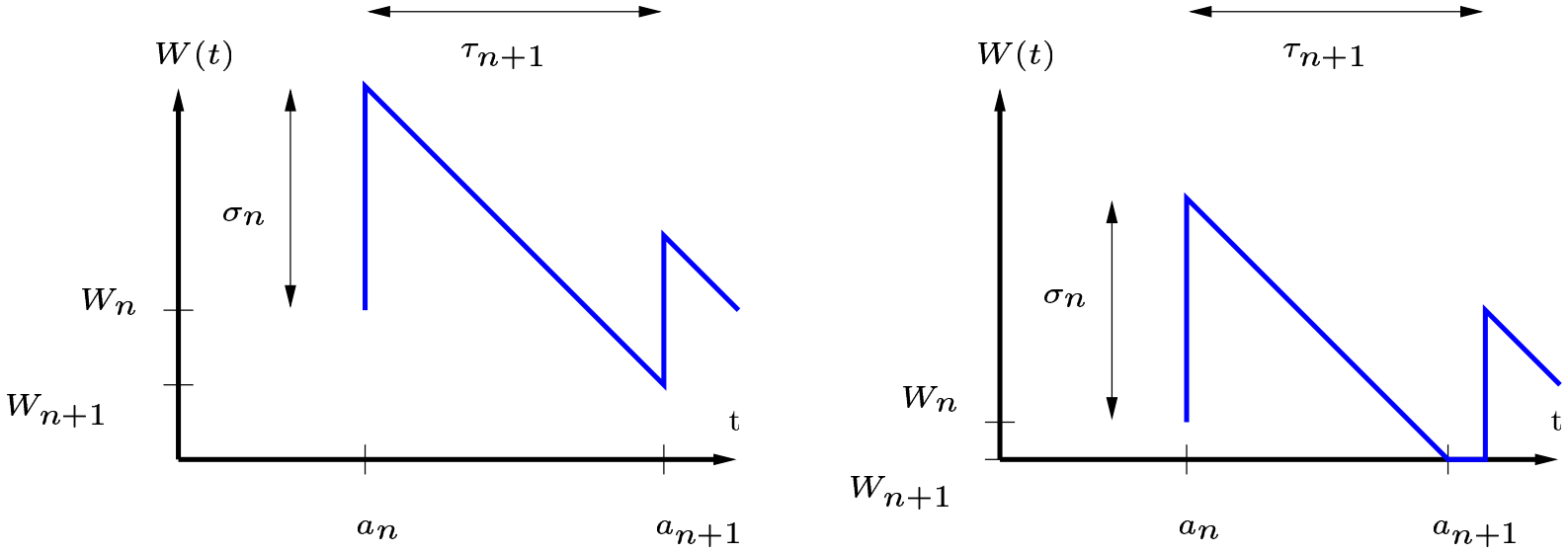
Périodes d'activité (PA) et inactivité (PI).

Temps d'attente – Le cas FIFO

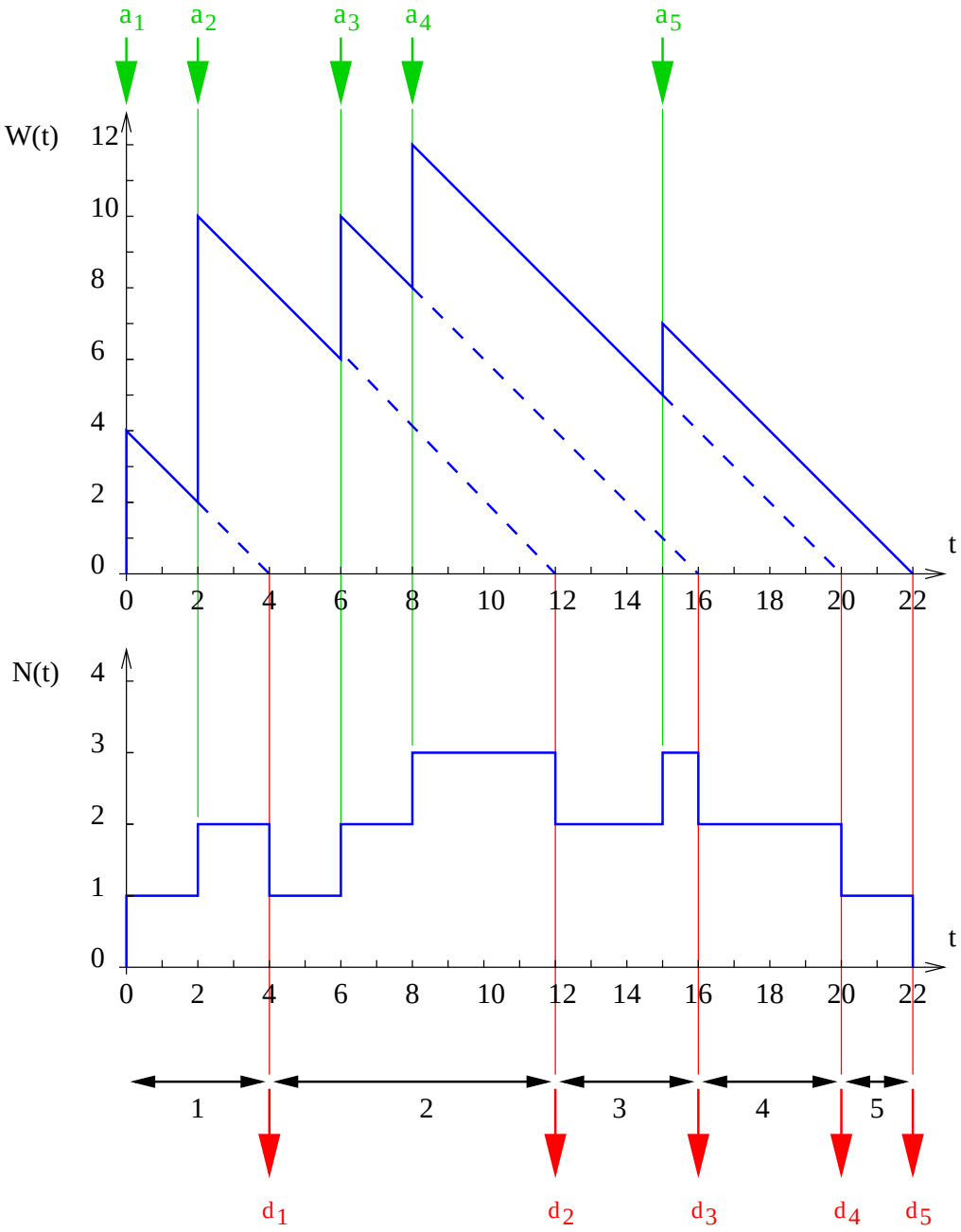
W_n: temps d'attente du client n avant service. σ_n: durée du service du client n.

Équation de Lindley:

$$W_{n+1} = [W_n + \sigma_n - \tau_{n+1}]^+.$$



Relation entre nombre de clients et charge



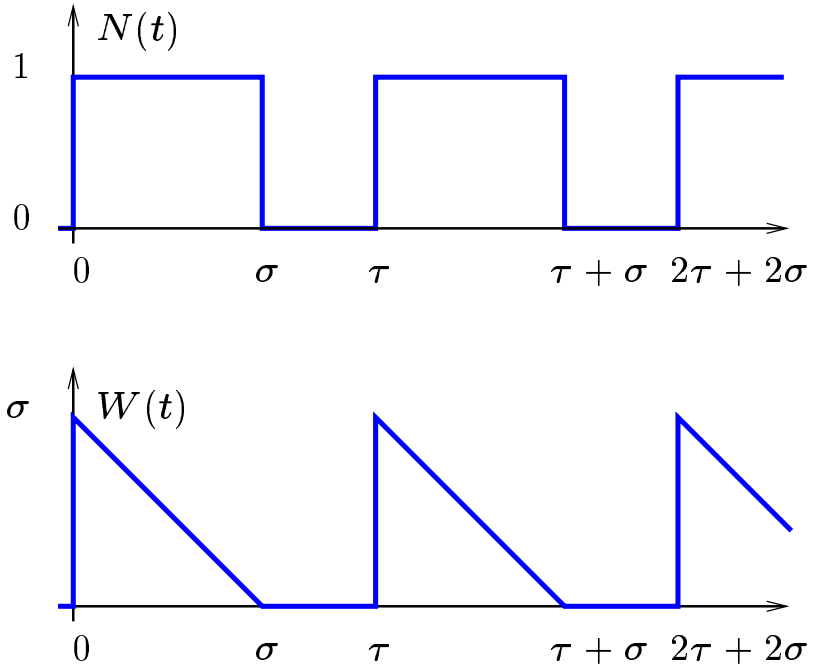
Attente virtuelle, attente réelle

Si a_n est la date d'arrivée du client n et si FIFO:

$$W_n \;=\; W(a_n^-)\;.$$

$\Rightarrow W(t)$ est aussi nommé temps d'attente virtuel.

Attention! W_n et $W(t)$ n'ont pas nécessairement la même distribution! Exemple: la file D/D/1.

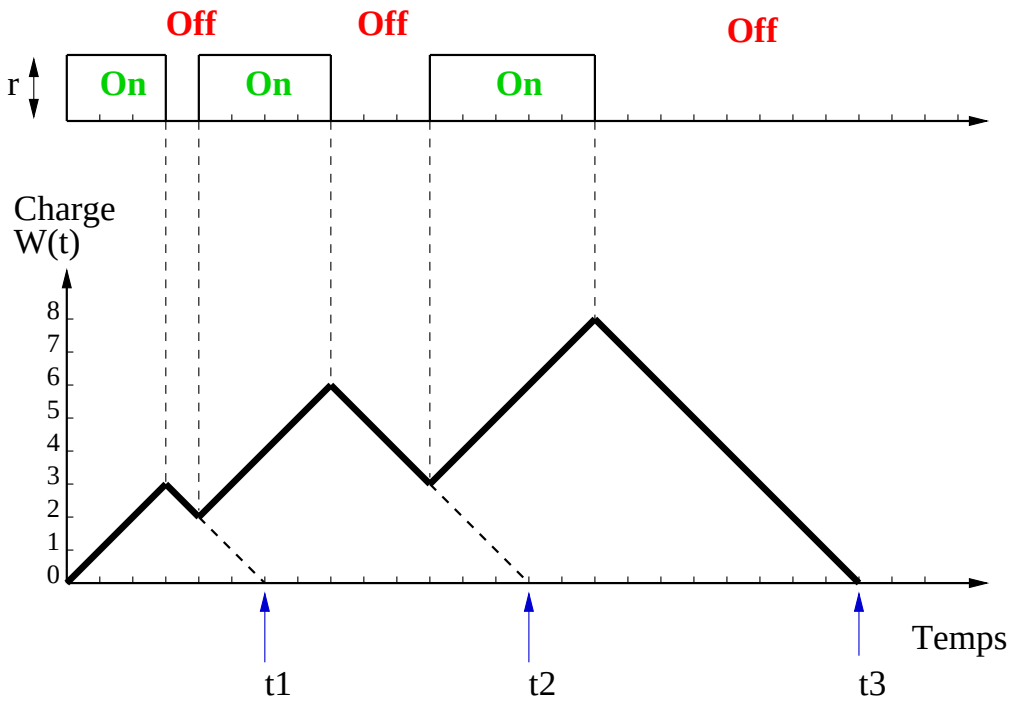


Propriété PASTA (Poisson Arrivals See Time Averages)
les arriveés sont Poisson, les distributions stationnaires de W_n et $W(t)$ coïncicent.

Files fluides

Plus de clients, mais du "fluide" arrivant avec un certain débit $r(t)$ (variable) et servi à une certaine vitesse C (possiblement variable aussi).

Exemple: arrivées selon un processus "on/off" (typique de la voix numérisée):



ti = fin du traitement de la charge apportée dans la i-ème période «On»

Résultats généraux

Stabilité

Stabilité: W_n admet un régime stationnaire.

Résultat:

La file G/G/1 est stable si et quasiment seulement si

$$\mathbb{E}\sigma \quad < \quad \mathbb{E}\tau$$

Formule de Little

Le temps moyen de réponse R et le nombre moyen de clients N son liés par la formule:

$$\lambda\, T \quad = \quad N$$

Modèles de trafic

Le trafic est décrit par:

- le processus de arrivées $\{a_n\}_{n\in\mathbb{N}}$ ou des inter-arrivées $\{\tau_n\}_{n\in\mathbb{N}}$.
- le processus des services $\{\sigma_n\}_{n\in\mathbb{N}}$.

Modèles "iid". Distribution de proba. du temps entre arrivées fixe + indépendance. Idem pour les services.

Cas classiques: déterministe, loi exponentielle

$$\mathbb{P}\{\tau > x\} \;=\; e^{-\lambda x}$$

Lois Gamma/Erlang (sommes d'exponentielles).

Nouvelles tendances: lois à "queue lourde": Pareto

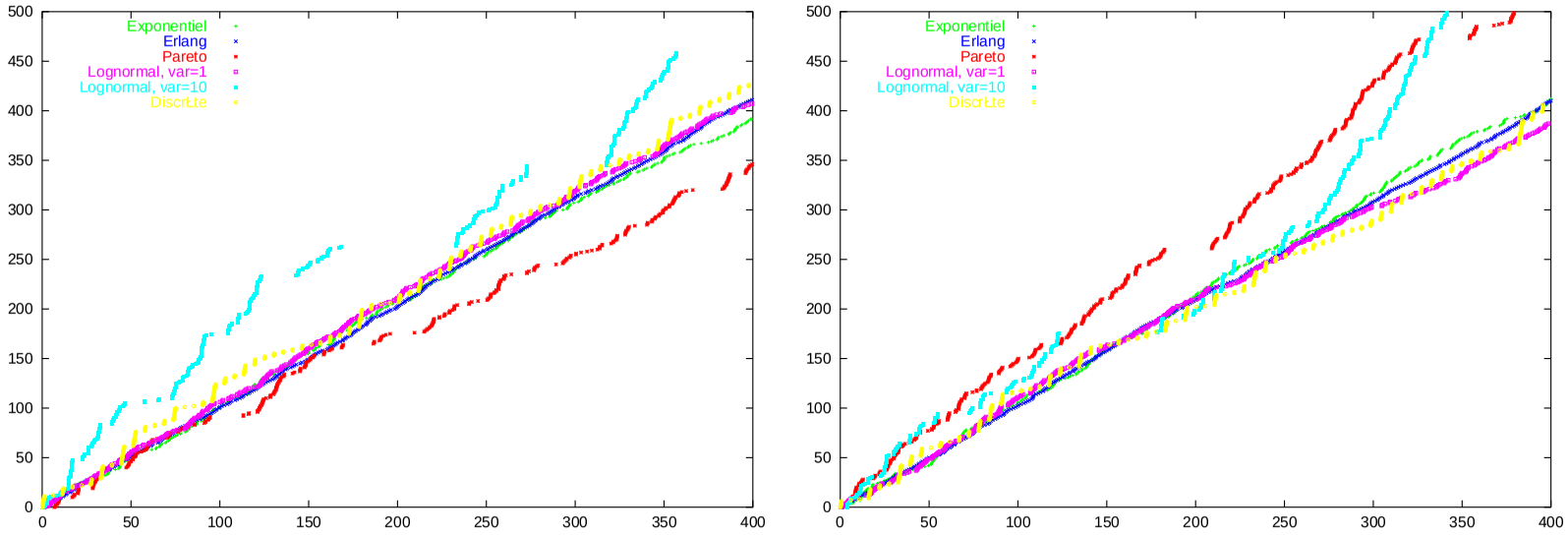
$$\mathbb{P}\{\tau > x\} \;=\; \left(\frac{a}{a+x}\right)^{\alpha} .$$

Weibull, LogNormale.

$$\mathbb{P}\{\tau > x\} \;=\; \mathbb{P}\{X > \log(x)\}, \quad X \;\sim\; \mathcal{N}(m,\sigma) \; .$$

Exemple

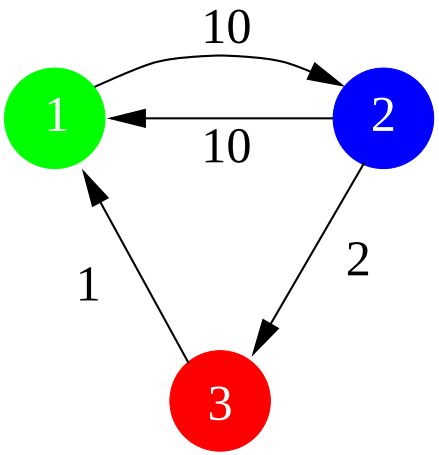
Comparaison des irrégularités dans les dates d'arrivées pour diverses lois de τ.

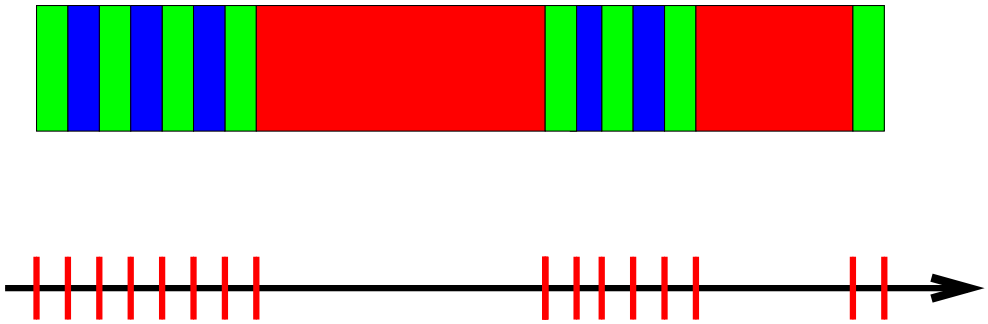


Abscisse: a_n, Ordonnée: n.

Arrivées Modulées par Markov

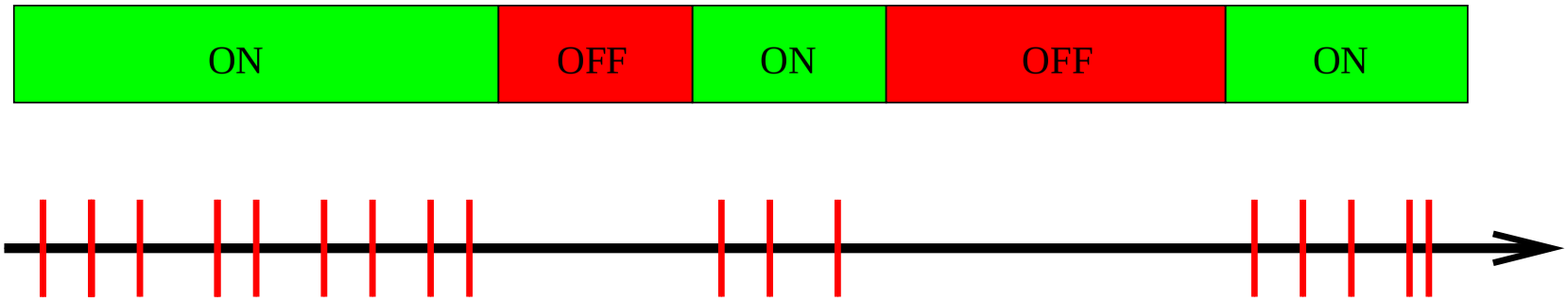
MAP: Markov Arrival Process: Les arrivées se font au moment des changement d'état d'une chaîne de Markov.



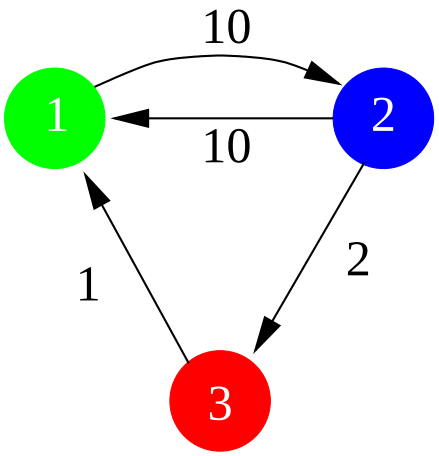


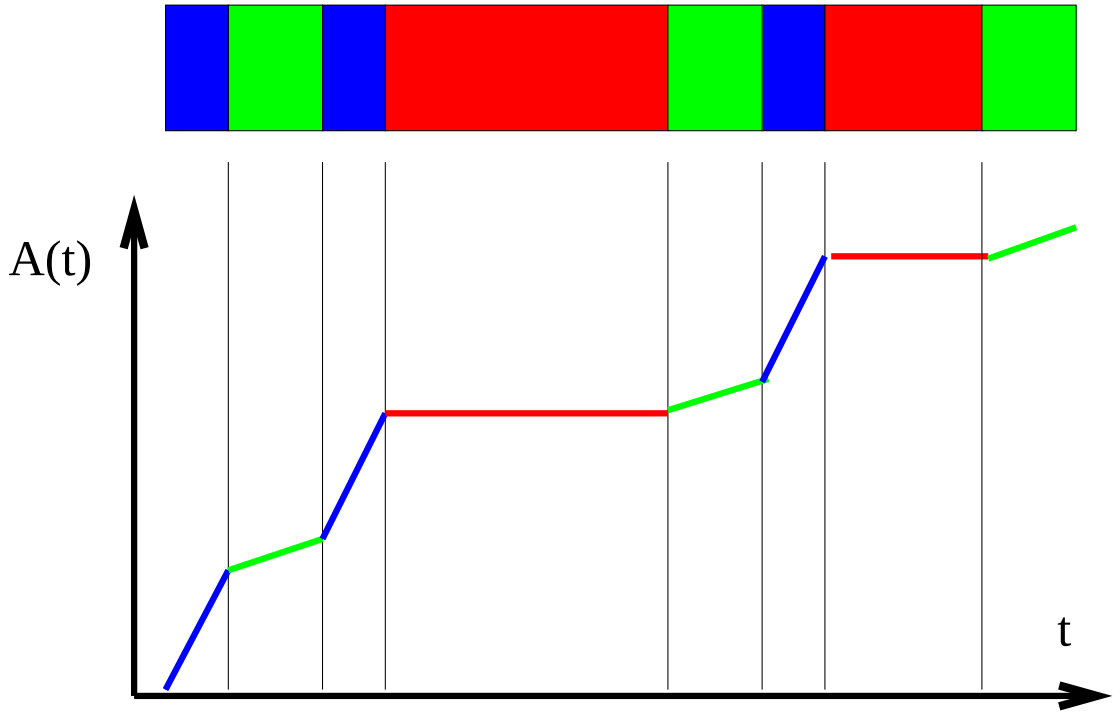
MMPP: Markov Modulated Poisson Process: Arrivées selon des processus de Poisson d'intensité dépendante de l'état d'une chaîne de Markov (ou d'un processus semi-Markovien).

En particulier, les processus IPP: Interrupted Poisson Processes.



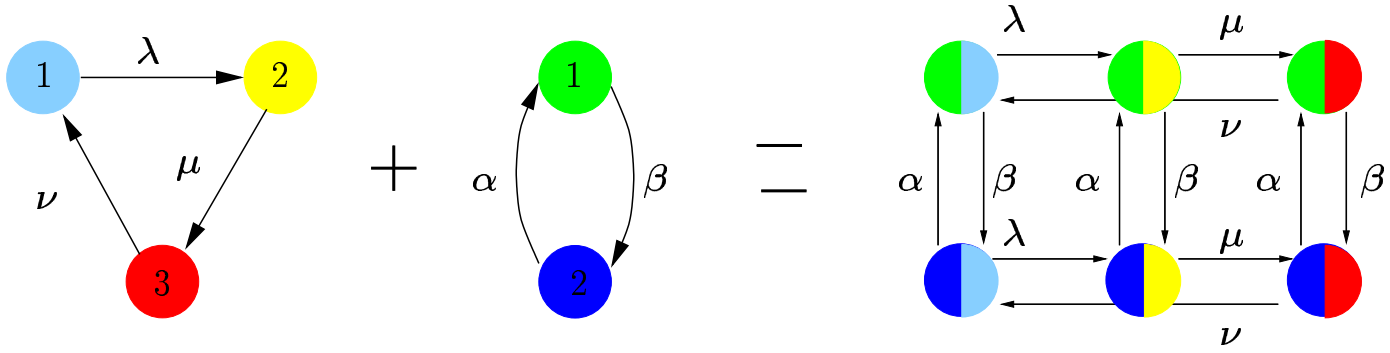
MMRP: Markov Modulated Rate Process: selon un processus fluide de taux dépendant de l'état d'une chaîne de Markov.





Superposition de sources

Si plusieurs sources de trafic sont superposées, le processus résultant est toujours modulé par Markov.



$$\begin{pmatrix} -\lambda & \lambda & 0 \\ 0 & -\mu & \mu \\ \nu & 0 & -\nu \end{pmatrix} \oplus \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} = \begin{pmatrix} - & \lambda & 0 & \alpha & 0 & 0 \\ 0 & - & \mu & 0 & \alpha & 0 \\ \nu & 0 & - & 0 & 0 & \alpha \\ \beta & 0 & 0 & - & \lambda & 0 \\ 0 & \beta & 0 & 0 & - & \mu \\ 0 & 0 & \beta & \nu & 0 & - \end{pmatrix}$$

Hypothèses de modélisation

- Capacité finie ou infinie?
  Les files d'attente à capacité infinie sont plus faciles à résoudre: on peut les utiliser comme approximations.

$$\mathbb{P}\{\text{perte}\} \leftrightarrow \mathbb{P}\{N = K\} \leftrightarrow \mathbb{P}\{W > K\}$$

- Modéliser les réseaux?
  Peu de résultats existent sur les réseaux de files d'attente. On compte sur l'hypothèse (souvent validée) du goulôt d'étranglement solitaire (single bottleneck)
- Quels modèles de trafic? Compromis entre ce qu'on sait calculer et ce qui est pratiquement raisonnable.

Partie III: Analyse exacte

- Analyse exacte dans le cas de tampons infinis
  - la file M/M/1, la file M/GI/1, la file GI/M/1.
  - la file MMPP/GI/1.
- Analyse exacte dans le cas de tampons finis
  - la file M/M/1/K.
  - les QBD
- Réseaux de files d'attente

La file M/M/1

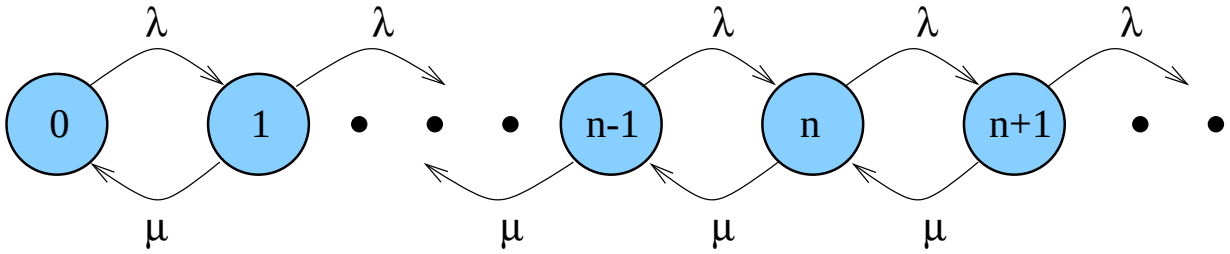
Caractéristiques: Salle d'attente infinie, 1 serveur.

inter-arrivées: loi exponentielle de paramètre λ:
services: loi exponentielle de paramètre μ:

$$\mathbb{P}\{\tau \leq x\} = 1 - e^{-\lambda x}, \qquad \mathbb{P}\{\sigma \leq x\} = 1 - e^{-\mu x}.$$

Stabilité: $\lambda < \mu$.

$\{N(t)\}$ est une Chaîne de Markov: un *processus de naissance et de mort*



Performances:

$$\mathbb{P}\{W > x\} = \frac{\lambda}{\mu}\, e^{-(\mu-\lambda)x}$$

$$\mathbb{P}\{N \geq n\} = \left(\frac{\lambda}{\mu}\right)^n$$

Autres résultats classiques

La file M/GI/1

Arrivées: exponentielles, taux λ,

Services: loi quelconque, transformée de Laplace $S^*(s)$.

La transformée de Laplace du temps d'attente et du nombre de clients (formule de *Pollaczek-Khinchine*):

$$\begin{aligned} W^*(s) &= \frac{1-\rho}{s-\lambda(1-S^*(s))} \\ N^*(z) &= S^*(\lambda(1-z))\,\frac{(1-\lambda/\mu)(1-z)}{S^*(\lambda(1-z))-z}\,. \end{aligned}$$

En particulier, les moyennes valent:

$$\mathbb{E}W \;=\; \frac{\lambda\mathbb{E}\sigma^2}{2(1-\rho)} \qquad \mathbb{E}N \;=\; \rho \;+\; \rho^2\,\frac{\mu^2\,\mathbb{E}\sigma^2}{2(1-\rho)}\,.$$

$\rho = \lambda/\mu$: est le taux d'utilisation.

La file GI/M/1

Arrivées: loi quelconque, transformée de Laplace $A^*(s)$,
Services: exponentielles, moyenne $1/\mu$.

Distribution du temps d'attente:

$$\boxed{\mathbb{P}\{W > x\} \quad = \quad \theta \; e^{-\mu(1-\theta)x} \; ,}$$

avec:

$$\theta \; = \; A^*(\mu(1-\theta)) \; .$$

$\Rightarrow$ distribution exponentielle!

Taux de service équivalent

$$\widehat{\lambda} \; = \; \theta\mu \; .$$

$\Rightarrow$ Bande Passante équivalente pour les réseaux.

La file MMPP/GI/1

Arrivées: MMPP avec N états, un générateur $\mathbf{Q}$ et matrice de taux $\mathbf{\Lambda}$;

Services: indépendants avec une distribution générale $H(x)$, de transformée de Laplace $H^*(s)$.

Distribution de la quantité de travail W:

$$\mathbf{W}^*(s) \;=\; s(1-\rho)\;\mathbf{g}\;[s\mathbf{I}+\mathbf{Q}-(1-H^*(s))\mathbf{\Lambda}]^{-1}\;\mathbf{1}\;,$$

$\mathbf{g}$ vecteur à déterminer.

Si $\sigma \sim \text{Exp}(\mu)$ (la file MMPP/M/1), alors:

$$\mathbb{P}\{W > x\} \;=\; \sum_{k=0}^{N} a_k\, e^{-\theta_k x} \;\sim\; a_1\, e^{-\theta_1 x}\;,$$

avec θ_k tel que:

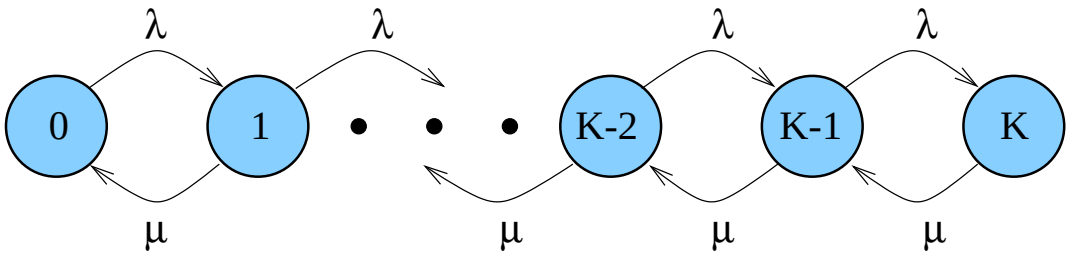
$$det\{-\theta_k s\mathbf{I}+\mathbf{Q}-(1-H^*(-\theta_k))\mathbf{\Lambda}\} \;=\; 0.$$

$\Rightarrow$ à nouveau: queue de distribution asymptotiquement exponentielle.

La file M/M/1/K

Comme la M/M/1 mais avec capacité finie K.

Chaîne de Markov: elle est finie



Performances: soit $\rho = \lambda/\mu$.

$$\mathbb{P}\{N = K\} \;=\; \rho^K \, \frac{1-\rho}{1-\rho^{K+1}} \; .$$

Probabilité de perte d'un client: c'est justement $\mathbb{P}\{N = K\}$.

Note: avec l'approximation tampon infini, on aurait obtenu:

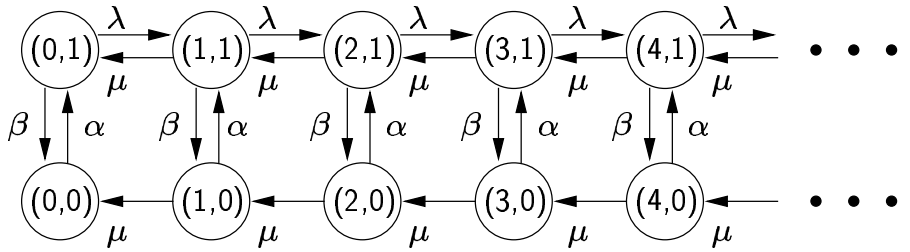
$$\mathbb{P}\{N = K\} \;\simeq\; \rho^K (1-\rho) \; .$$

Les QBD processes

QBD: "quasi birth-death".

C'est une structure de chaîne de Markov qu'on obtient quand le processus d'arrivée ou le processus de service est "à phases".

Par exemple:



On les résoud par la méthode de Neuts (par exemple).

La méthode de Neuts pour les QBD

L'espace d'états $\mathcal{E}$ est partitionné en blocs finis de même taille

$$\mathcal{E}_k = \{(k,1),(k,2),\ldots,(k,N)\}$$

tels que la matrice de transition de la chaîne de Markov a la structure: *tridiagonale par blocs* $N \times N$:

$$P = \begin{pmatrix} S_0 & L & & & & \\ M & S & L & & & \\ & M & S & L & & \\ & & \ddots & \ddots & \ddots & \\ & & & M & S & L \\ & & & & M & S_K \end{pmatrix}$$

Les probabilités stationnaires sont elles aussi regroupées en blocs:

$$\pi_k \;=\; (\pi_{k,1}, \ldots, \pi_{k,N})$$

L'équation d'équilibre $\pi P \;=\; \pi$ devient:

$$\left\{ \begin{array}{rcrcrcll} & & \pi_0 S_0 & + & \pi_1 M & = & \pi_0 & \\ \pi_{k-1}\, L & + & \pi_k S & + & \pi_{k+1} M & = & \pi_k & 0 < k < K \\ \pi_{K-1}\, L & + & \pi_K S_K & & & = & \pi_K & \end{array} \right.$$

$\Rightarrow$ résolution numérique de la récurrence par des méthodes itératives.

On a la même analyse pour les chaîne de Markov en temps continu.

Les réseaux à forme produit: Réseaux de Jackson

N files d'attente (stations) dont les services sont $\sim$ Exp:

- le vecteur $\boldsymbol{\lambda}_0 = (\lambda_{0,1}, \ldots, \lambda_{0,N})$ des taux d'arrivées extérieures dans chaque file,
- le vecteur $(\mu_1, \ldots, \mu_N)$ des taux de service,
- la matrice carrée $N \times N$ de routage interne R: $r_{i,j} = \mathbb{P}\{\text{un client sortant de } i \text{ va vers } j\}$.

Flux entrant dans les stations: vecteur $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ solution de:

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 + \boldsymbol{\lambda}\,\mathsf{R}\,.$$

La condition de stabilité du système est:

$$\forall 1 \leq i \leq N, \qquad \lambda_i < \mu_i.$$

Si stabilité, la distribution de probabilité stationnaire est:

$$p(n_1, \ldots, n_N) \;=\; \prod_{i=1}^{N} \left(1 - \frac{\lambda_i}{\mu_i}\right) \left(\frac{\lambda_i}{\mu_i}\right)^{n_i} ,$$

$\Rightarrow$ comme si les files étaient des $M/M/1$ en isolation, indépendantes

$\Rightarrow$ justification de la formule de temps de réponse de bout en bout:

$$T \;=\; \sum_{i=1}^{N} \frac{1}{C_i - L_i}$$

C_i: capacité du lien/routeur, L_i: trafic entrant.

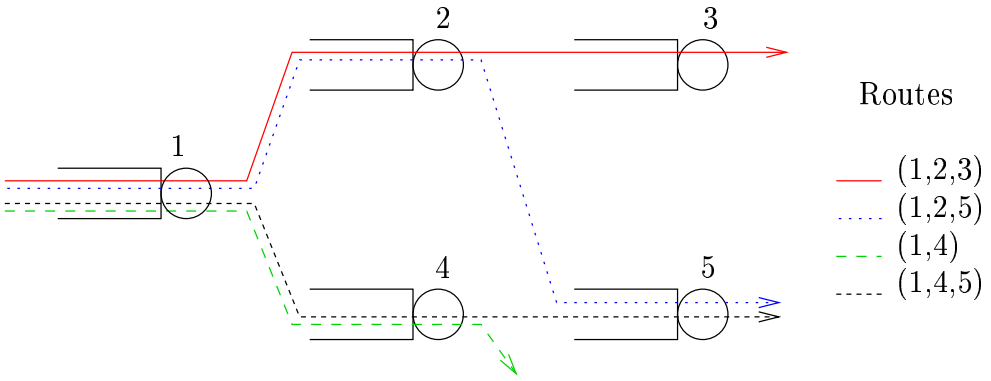
Les réseaux de Kelly

Les clients appartiennent à des classes.

À chaque classe k est affecté une route dans le réseau:

$$r_k = (r_k^1, \ldots, r_k^{n_k}) \ .$$

Les clients arrivent suivant des processus de Poisson, et les serveurs délivrent des temps de service exponentiels.



Un réseau de Kelly

Le flux $\hat{\lambda}_{ik}$ de clients de classe k entrant dans la file d'attente i:

$$\hat{\lambda}_{ik} \ = \ \lambda_k \ \times \ (\text{nombre de } i \text{ dans } r_k).$$

Flux total, toutes classes confondues:

$$\hat{\lambda}_i \ = \ \sum_k \ \hat{\lambda}_{ik} \ .$$

Probabilités stationnaires

Soit $M = ((m_{ik}))$ une matrice de populations par file et par classe. La probabilité stationnaire que le réseau se trouve dans l'état M est:

$$\mathbb{P}\{M\} = \prod_{i=1}^{N} \left(1 - \frac{\hat{\lambda}_i}{\mu_i}\right) \left(\sum_{k=1}^{K} m_{ik}\right)! \prod_{k=1}^{K} \frac{1}{m_{ik}!} \left(\frac{\hat{\lambda}_{ik}}{\mu_i}\right)^{m_{ik}} .$$

Statistiques

Nombre moyen de clients dans la file i,

$$\overline{N}_i = \frac{\hat{\lambda}_i}{\mu_i - \hat{\lambda}_i}$$

Temps de réponse de bout en bout sur la route k:

$$\overline{T}_k = \frac{1}{\lambda_k} \sum_{j=1}^{n_k} \frac{\hat{\lambda}_{r_k^j,k}}{\mu_{r_k^j} - \hat{\lambda}_{r_k^j}} ,$$

Temps moyen toutes classes confondues:

$$\overline{T} = \frac{1}{\sum_{k=1}^{K} \lambda_k} \sum_{i=1}^{N} \frac{\hat{\lambda}_i}{\mu_i - \hat{\lambda}_i} .$$

Questions pour les réseaux

Il existe des extensions aux formes produits des Réseaux de Jackson et Kelly: réseaux multiclasses, mixtes ouverts/fermés, et avec des politiques de service variés: Théorème BCMP.

De nombreuses questions restent ouvertes. Par exemple:

- hypothèses moins contraignantes sur le modèle de trafic: processus d'arrivée non Poissonnien, services non exponentiels
- capacités finies, pertes, rétroaction
- disciplines de service et stabilité
- distributions de temps de réponse de bout en bout

Partie IV: Analyse asymptotique

- Principe
- Bornes et asymptotiques exponentielles
  - Bornes de Chernoff et borne de Kingman
  - Processus markoviens additifs
  - Bande passante équivalente
- Mémoire longue, autosimilarité, sous-exponentialité
  - Processus autosimilaires dans la nature
  - Sous-exponentialité et dominance asymptotique
  - Mémoire longue et capacité finie

Principe

Analyse asymptotique brute: trouver un équivalent à:

$$\mathbb{P}\{W > x\}, \qquad x \to \infty$$

dont on espère tirer une approximation.

Typiquement: un *équivalent asymptotique exponentiel*:

$$\mathbb{P}\{W > x\} \sim C\, e^{-\theta x}, \quad x \to \infty.$$

Bornes: on essaie de trouver des bornes de cette nature:

$$B(\theta)\, e^{-\theta x} \leq \mathbb{P}\{W > x\} \leq C(\theta)\, e^{-\theta x},$$

soit pour x "grand", soit pour tout x.

Outils: borne de Chernoff, grandes déviations.

Borne de Chernoff

Soit t un nombre réel fixé, et X une variable aléatoire.

Transformée de Laplace-Stiltjes de X:

$$X^*(s) \;=\; \mathbb{E}(e^{-sX}) \;.$$

On a:

$$\begin{array}{rcll}
\mathbf{1}_{\{x\leq t\}} & \leq & e^{\theta(x-t)} & \forall x,\theta \\
\mathbb{E}\mathbf{1}_{\{x\leq t\}} & \leq & \mathbb{E}e^{\theta(X-t)} & \forall \theta \\
\mathbb{P}\{X\leq t\} & \leq & X^*(-\theta)\; e^{-\theta t} & \forall \theta \\
\mathbb{P}\{X\leq t\} & \leq & \inf_\theta\{X^*(-\theta)\; e^{-\theta t}\} &
\end{array}$$

Borne de Kingman

On considère la file GI/GI/1.

$$\mathbb{P}\{W > x\} \leq e^{-\theta x}$$

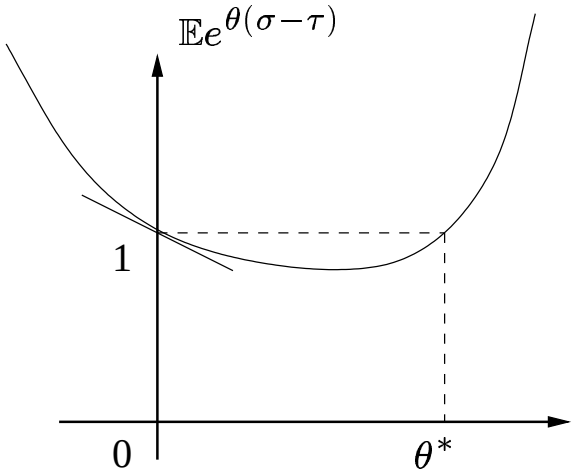
pour tout nombre $\theta \geq 0$, tel que:

$$\mathbb{E}(e^{\theta(\sigma-\tau)}) \leq 1 .$$

Donc, en prenant le plus grand θ possible:

$$\theta^* = \sup\{\theta \geq 0 \mid \mathbb{E}(e^{\theta(\sigma-\tau)}) \leq 1\}$$

Conclusion: décroissance exponentielle dans le cas où $\theta^* > 0$.



Grandes déviations

On généralise ce résultat à des processus d'arrivée/service moins simples:

Si le processus $\{U_n\} = \{\sigma_n - \tau_n\}$, stationnaire et ergodique, satisfait:

$$\Phi(\theta) \;=\; \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}[e^{\theta(U_0+\ldots+U_{n-1})}] \;,$$

alors:

$$\boxed{\lim_{x\to\infty} \frac{1}{x} \mathbb{P}\{W > x\} \quad = \quad -\theta^* \;.}$$

avec

$$\theta^* \;=\; \sup\{\theta \geq 0 \mid \Phi(\theta) = 0\} \;.$$

Mémoire longue et autosimilarité

Des mesures ont montré que le processus d'arrivée d'information montre une certaine autosimilarié et une corrélation à long terme.

Or les modèles "classiques" n'ont pas cette propriété.

D'ou provient ce phénomène?

Quelle est l'influence de cette mémoire longue sur les probabilités de perte? Faut il jeter les modèles connus?

Quels nouveaux modèles peut-on analyser? Modèles avec arrivées/services "à queue lourde".

Notion de sous-exponentialité des distributions de probabilité.

Autosimilarité

Soit $\mathbf{X} = \{X(n)\}_n$ un processus stationnaire au sens large.

$\mathbf{X}$ est autosimilaire si:

$$\mathbf{X} =_d \frac{1}{m^H} (X_{t(m+1)+1} + \ldots + X_{tm})$$

pour tout m.

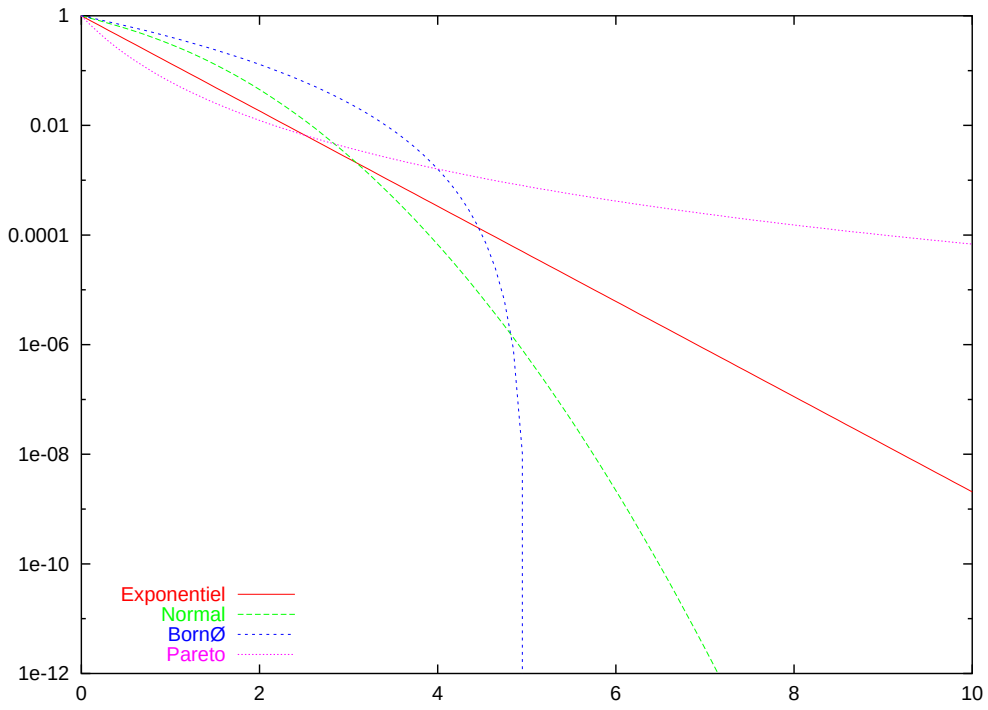
H: paramètre de Hurst.

Exemple: le mouvement Brownien fractionnaire est autosimilaire.

C'est également un processus à mémoire longue.

Sous-Exponentialité

Le problème, graphiquement.



Exemple: deux sources On/Off indépendantes.

- A_1 durées de "On" *à queue lourde*
- A_2 durées de "On" arbitraires, de débit moyen ρ
- C: capacité du serveur.

W^{1+2}: charge stationnaire quand A_1 et A_2 sont superposées.

W^1: charge avec A_1 seul mais capacité $C - \rho$.

Alors:

$$\boxed{\mathbb{P}\{W^{1+2} > x\} \sim \mathbb{P}\{W^1 > x\}}$$

Conclusion: A_1 "dicte" le comportement asymptotique de W.

Partie V: Modèles déterministes

- Enveloppes de trafic et bornes (σ, ρ)
- Régulateurs de trafic, courbes de service

Principe

Fonction *d'arrivée de travail*:

$$\begin{aligned} S(a,b) &= \sum_{a \le a_n < b} \sigma_n \qquad \text{(discret)} \\ &= \int_a^b r(t)\mathrm{d}t \qquad \text{(fluide)} \end{aligned}$$

Enveloppe de l'arrivée de travail: le pire des situations pour les intervalles d'une certaine longueur t:

$$\hat{S}(t) = \sup_s S(s, s+t) \ .$$

Exemple: bornes "(σ, ρ)":

$$S(s, s+t) \le \sigma + \rho\, t, \quad \forall s.$$

Résultats

Si un processus admettant une enveloppe bornée par une fonction affine "(σ, ρ), alimente une file d'attente, alors:

$$W(t) \quad \leq \quad \sigma \ ,$$

De plus, le temps de réponse est borné:

$$R_n \quad \leq \quad \frac{\sigma}{1-\rho} \ ,$$

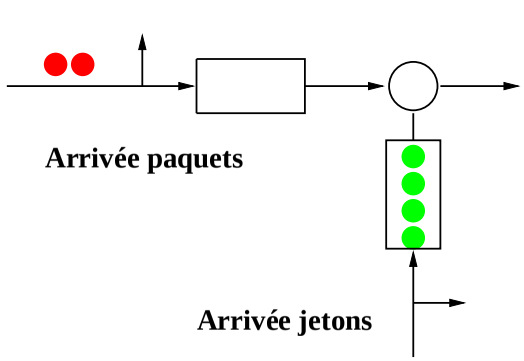
quelle que soit la politique de service.

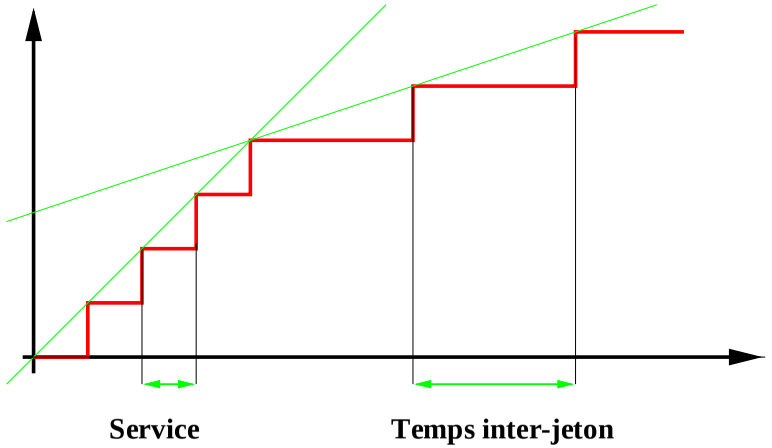
$\Rightarrow$ dimensionnement des tampons.

Régulateurs de trafic

Éléments de réseau chargés de réduire l'impact des rafales: *lissage* du trafic.

Exemple: le Leaky Bucket (aussi: Token Bucket).





Applications possibles: définition de *contrats de trafic* (débit crête/débit moyen).

Bibliographie succinte

Processus stochastiques:

- Ross, *Stochastic Processes*
- Cinlar, *Introduction to Stochastic Processes*, Prentice Hall, 1975.
- Baccelli, Brémaud, *Elements of Queueing Theory, Applications of Mathematics*, vol. 26, Springer-Verlag, 1994.
- Davis, *Markov Models and Optimisation*, Prentice Hall, 1993.

Théorie des files d'attente et son application aux réseaux:

- Kleinrock, *Queueing Networks* (2 volumes), Wiley, 1975.
- J. Walrand, *Introduction to Queuing Networks*, Prentice-Hall, 1989.
- Pujolle et Fdida, *Modèles de systèmes et de réseaux*, deux tomes, Eyrolles, Paris, 1989.

Aspects numériques, QBD, MMPP:

- Neuts, *Matrix geometric solutions*,
- Tijms, *Stochastic Models, an algorithmic approach*, Wiley, 1994.
- Mitra *et. al*, nombreux papiers.

– A. Jean-Marie, Z. Liu, Ph. Nain, D. Towsley, "Computational aspects of the Workload Distribution in the MMPP/GI/1 queue", *JSAC 99*.

Mémoire longue (très succint)

– Ph. Nain: "Impact of Bursty Traffic on Queues", à paraître dans *Statistical Inference in Stochastic Processes*.
`http://www-sop.inria.fr/mistral/personnel/Philippe.Nain/PAPERS/LRD/impact_bursty.pdf`.
– Bolot et Grossglauser: "On the relevance of long-range dependence in network traffic", Rapport de recherche INRIA RR-2830, 1996.