Markov-modulated models for traffic modeling

Alain Jean-Marie
INRIA et LIRMM, University of Montpellier
161 Rue Ada, 34392 Montpellier Cedex 5, France
ajm@lirmm.fr

Universidad Tecnica Federico Santa Maria
Valparaiso
October 2004

Plan of the talk

Introduction

The mathematical modeling of communication network necessitates an accurate representation of the arrival process of information.

Depending on the level of the model, this may be:

- the quantity of packets arrived in some network element before some time t,
- a quantity of frames (video), requests (transactions), or any other Application Data Unit,
- a quantity of bytes or bits,
- a quantity of time elapsed in some continuous media: vidéo, audio streaming...

Mathematical models of arrivals

The appropriate mathematical object is a counting process:

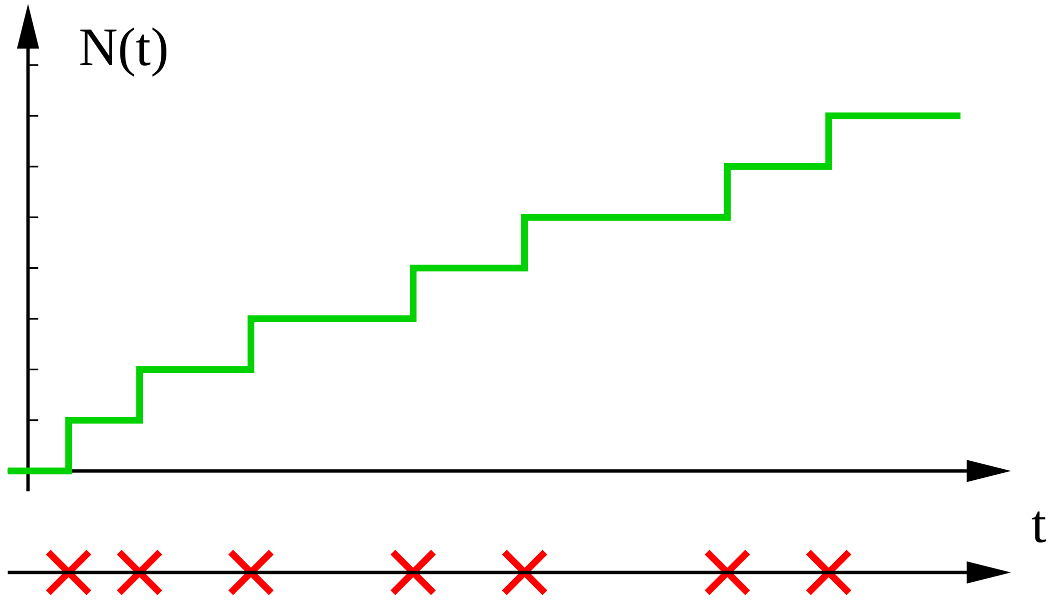
$$N(t) \ = \ \text{quantity arrived in the interval } [0, t) \ .$$

Several cases:

- discrete time: $t \in \mathbb{N}$
- continuous time: $t \in \mathbb{R}$
- discrete space: $N(t) \in \mathbb{N}$
- continuous space: $N(t) \in \mathbb{R}$

Counting process: illustration

Process of arrivals of events (arrivals, departures, changes, starts, stops, etc).



Modeling constraints

The variety of situations makes the following features necessary:

- relatively complex processes (bursts, temporal correlations, ...)
- possibly large number of sources
- ease of use, for simulation and stochastic calculus: distributions, queueing networks, asymptotics...

... with a mastered algorithmic complexity.

$\longrightarrow$ **Markov-modulated processes have these features**

Markov chains

A discrete-time Markov chain is a process $\{X(n), n \in \mathbb{N}\}$ such that:

- if $X(n) = i$, then $X(n+1) = j$ with probability p_{ij},
- jumps are independent.

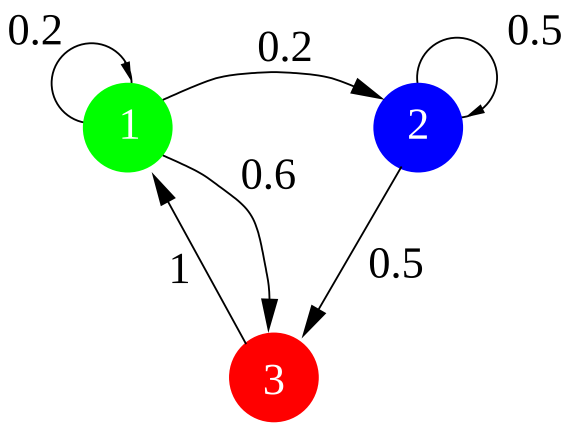
A Markov chain is fully described by its

transition probabilities: $p_{i,j}, (i,j) \in \mathcal{E} \times \mathcal{E}$, or its

transition matrix $\mathbf{P}$.

Example of Markov chain

Transition diagram



Transition matrix

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

Continuous-time

X has an exponential distribution of parameter $\lambda > 0$ ($X \sim \text{Exp}(\lambda)$) if:

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

Counting process of the sequence $T_0 \leq T_1 \leq \ldots \leq T_n \leq T_{n+1} \leq \ldots$:

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

This is a **Poisson process** of parameter λ if $\{T_{n+1} - T_n\}$ is a i.i.d. sequence of variables $\text{Exp}(\lambda)$.

Continuous time Markov chains

Let $\{X(t), t \in \mathbb{R}^+\}$, having the following properties. When X enters state i:

- X stays in state i a random time, exponentially distributed with parameter τ_i, independent of the past; then
- X jumps instantly in state j with probability p_{ij}. We have $p_{ij} \in [0, 1]$, $p_{ii} = 0$ and

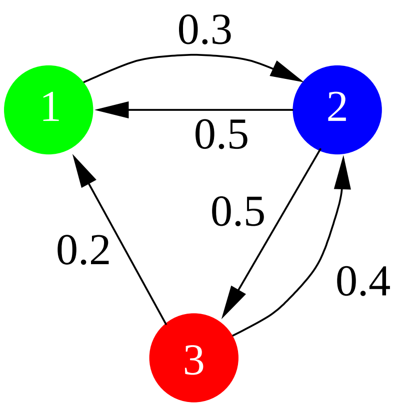
$$\sum_j p_{ij} = 1.$$

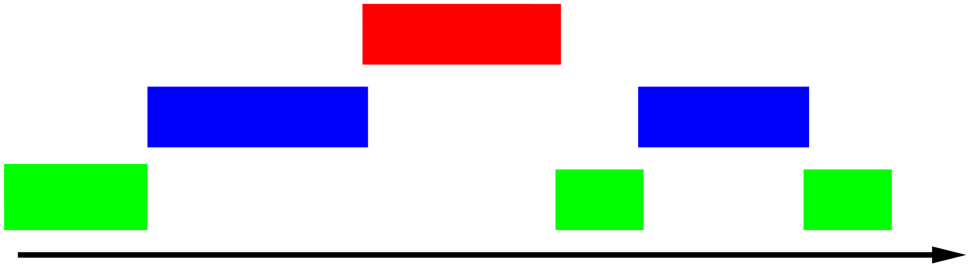
This process is a continuous-time Markov chain with transition rates

$$q_{ij} \ = \ \tau_i p_{ij} \ .$$

Example

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





Properties and Analysis

The most useful properties of Markov processes are:

- they are described by matrices,
- computing distributions involves the solution of linear problems
- their superposition leads to simple matrix computations.

Superposition of sources

If on superposes several Markov-modulated sources, the resulting process is still Markov-modulated.

The matrices (generators and rates) are obtained using Kronecker sums.

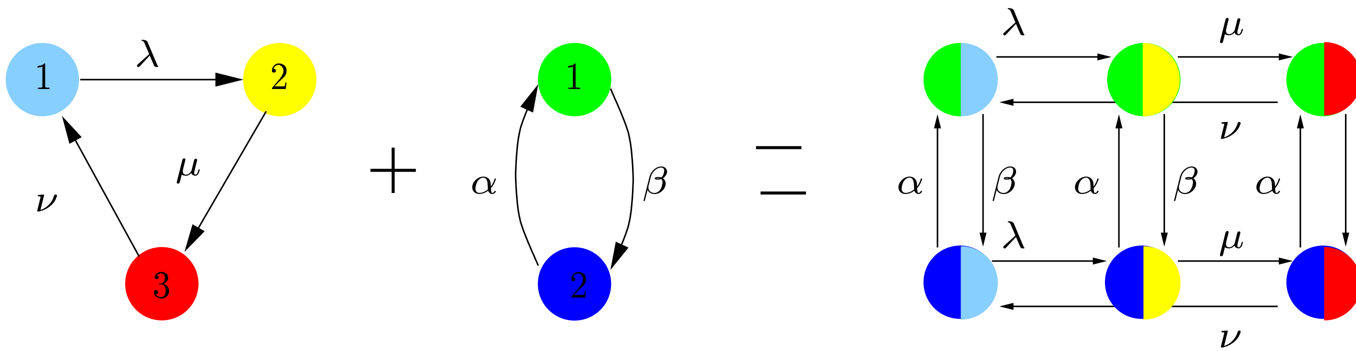
Kronecker product: consider two matrices A $(n \times n)$ and B $(m \times m)$. Their Kronecker product is a matrix $nm \times nm$ with

$$A \otimes B = \begin{pmatrix} A_{11}B & \dots & A_{1n}B \\ \vdots & & \vdots \\ A_{n1}B & \dots & A_{nn}B \end{pmatrix}.$$

Kronecker sum: a matrix $nm \times nm$ defined as

$$
\begin{aligned}
A \oplus B &= A \otimes I(m) + I(n) \otimes B \\
&= \begin{pmatrix} A_{11}B & & \\ & \ddots & \\ & & A_{nn} \end{pmatrix} + \begin{pmatrix} B_{11}I & \dots & B_{1m}I \\ \vdots & & \vdots \\ B_{n1}I & \dots & B_{nn}I \end{pmatrix}.
\end{aligned}
$$

Example: for two Markov chains $\{X_1(t)\}$ and $\{X_2(t)\}$, we have:



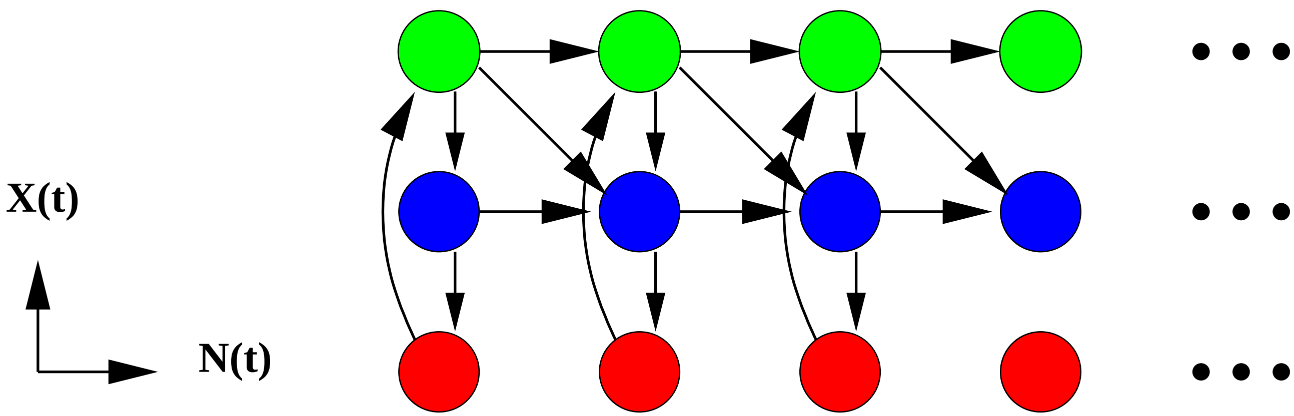
$$\begin{pmatrix} -\lambda & \lambda & 0 \\ 0 & -\mu & \mu \\ \nu & 0 & -\nu \end{pmatrix} \oplus \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} = \left(\begin{array}{ccc|ccc} - & \lambda & 0 & \alpha & 0 & 0 \\ 0 & - & \mu & 0 & \alpha & 0 \\ \nu & 0 & - & 0 & 0 & \alpha \\ \hline \beta & 0 & 0 & - & \lambda & 0 \\ 0 & \beta & 0 & 0 & - & \mu \\ 0 & 0 & \beta & \nu & 0 & - \end{array}\right)$$

Plan of the talk

Markov modulated arrivals

General idea:

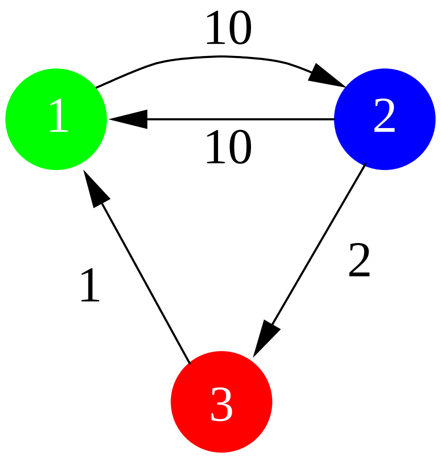
- A Markov chain $\{X(t); t \in \mathbb{R} \text{ or } \mathbb{N}\} \in \mathcal{E}$, the phase
- A counting process $N(t)$ such that $\{(X(t), N(t))\} \in \mathcal{E} \times \mathbb{N}$ is a Markov chain.

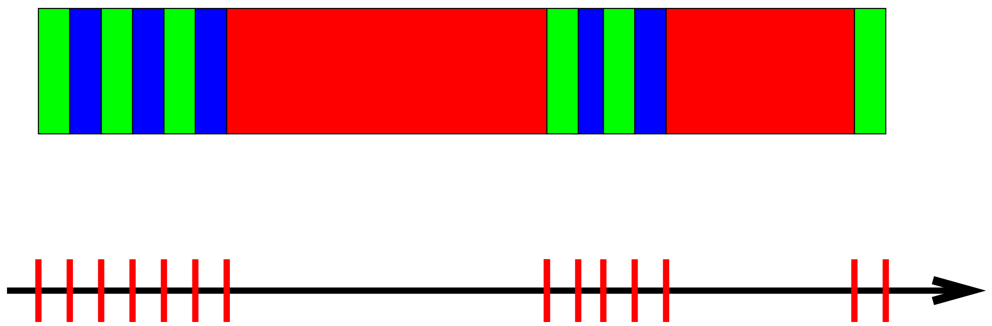


MAP: Markov Arrival Process

Let $\{X(t); t \in \mathbb{R}\}$ be a continuous-time Markov chain.

$\{N(t); t \in \mathbb{R}\}$ counts the number of jumps of X in $[0, t)$.



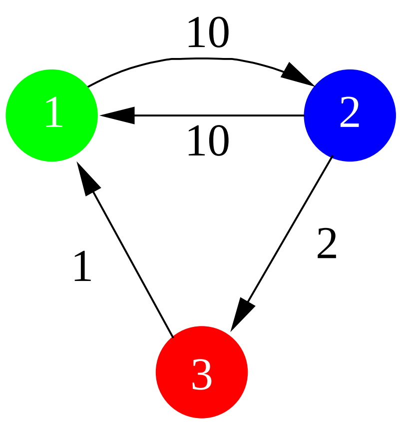


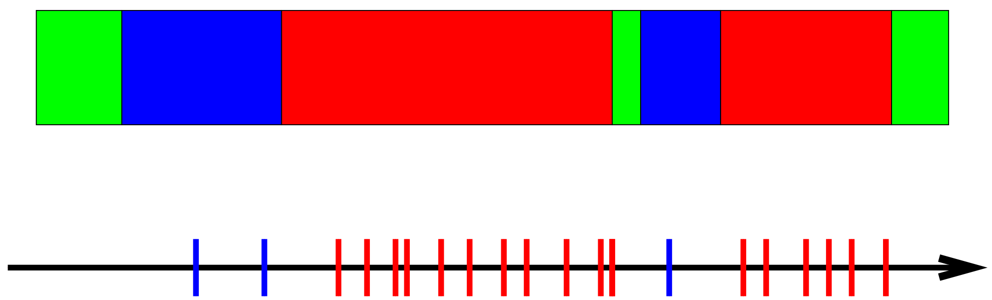
MMPP: Markov Modulated Poisson Process

Let $\{X(t); t \in \mathbb{R}\}$ be a continuous-time Markov chain in $\mathcal{E}$.

Let $\lambda_i \geq 0$ be an arrival rate, for each $i \in \mathcal{E}$.

Arrivals occur according to a Poisson process of time-varying rate $\lambda_{X(t)}$: that is, λ_i as long as $X(t) = i$.





BMAP: Batch Markov Arrival Process

Also known as "N-process" (N = Neuts), or the "versatile" process.

$\{(X(t), N(t)); t \in \mathbb{R}\}$ is a continuous-time Markov chain with a generator structured as:

$$Q = \begin{pmatrix} D_0 & D_1 & D_2 & \dots & \\ & D_0 & D_1 & D_2 & \\ & & D_0 & D_1 & \ddots \\ & & & \ddots & \ddots \end{pmatrix}$$

A process in the family of Markov additive process.

MMRP: Markov Modulated Rate Process

Let $\{X(t); t \in \mathbb{R}\}$ be a continuous-time Markov chain over a finite state space $\mathcal{E}$.

Let r_i be arrival rates (or accumulation rates), for each $i \in \mathcal{E}$.

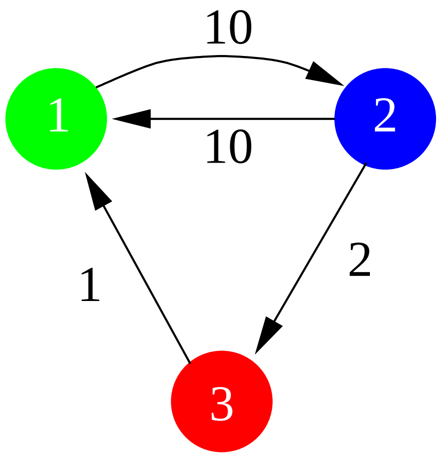
Arrivals occur according to a fluid process with rate $r_{X(t)}$, that is: with rate r_i as long as $X(t) = i$.

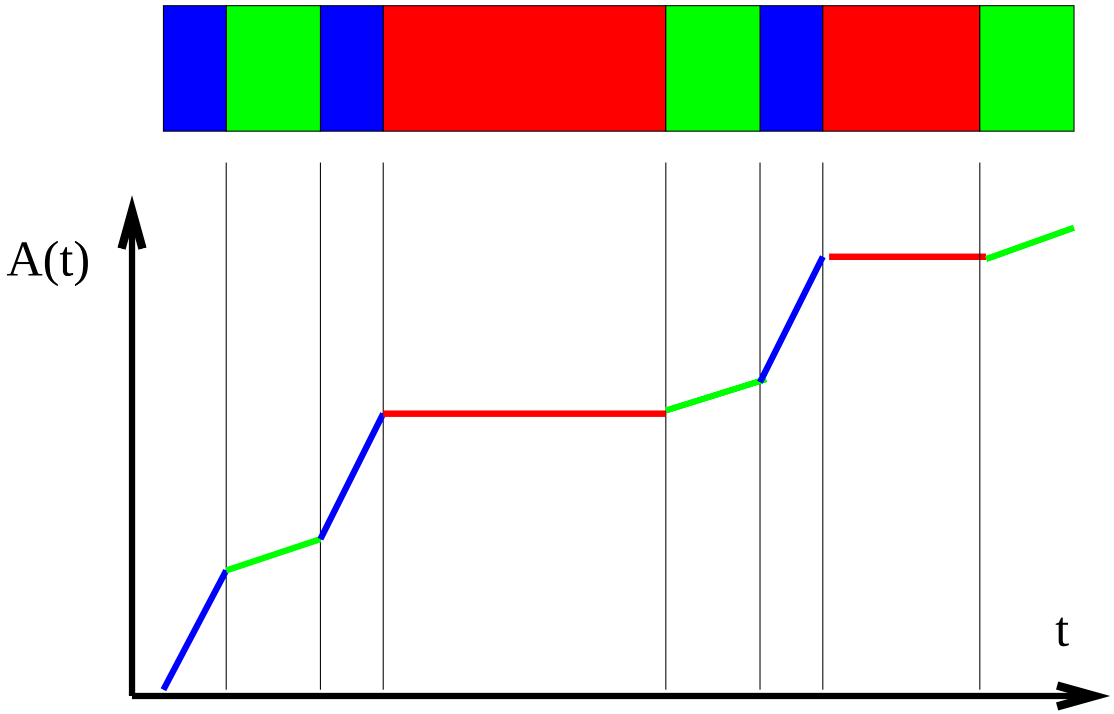
Let $N(t)$ the quantity arrived at time t:

$$\frac{dN}{dt}(t) \;=\; r_{X(t)} \;.$$

Note: also known as "Markov drift process".

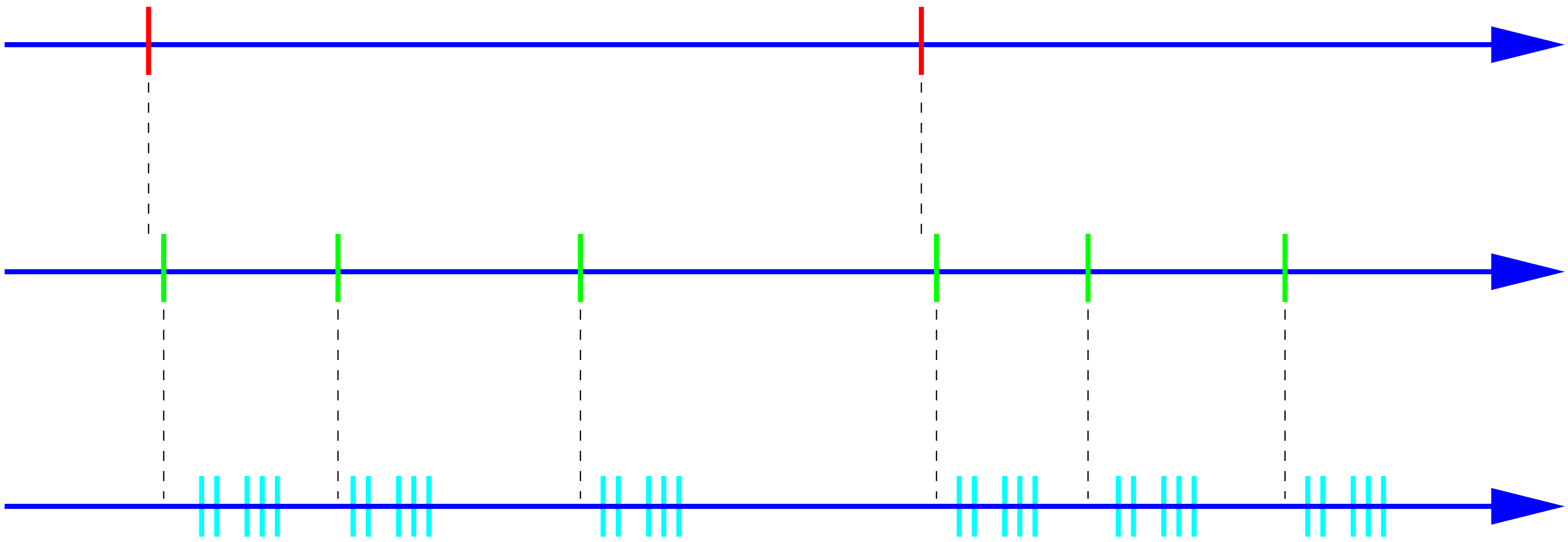
Example. $\mathcal{E}$ with three states, $r_3 = 0$, $0 < r_1 < r_2$:





Elaborate multiscale processes

Process with arrivals of sessions, requests, packets:



Synthesis

Markov modulated sources of arrivals are described by matrices

- For a MAP:

  the generator $\mathbf{Q}$

- For a MMPP/MMRP:

  the generator $\mathbf{Q}$, and the rate matrix $\mathbf{\Lambda}$

- For a BMAP:

  the collection of transition rate matrices $D_0, D_1, \ldots$

  Most distributions and performance measures are computed using these matrices.

Examples of computations

Average arrival rate

For a MMPP/MMRP, with π the stationary probability of X,

$$\overline{\lambda} \;=\; \boldsymbol{\pi}\,\boldsymbol{\Lambda}\,\mathbf{1} \;=\; \sum_{i\in\mathcal{E}} \pi_i\,\lambda_i\;.$$

Distribution of arrivals

For a MMPP, if $A_{ij}(k,T) = \mathbb{P}\{k \text{ arrivals and} X(T) = j \mid X(0) = i\}$, then

$$\sum_k z^k A_{ij}(k,T) \;=\; \left(e^{(\mathbf{Q}-(1-z)\boldsymbol{\Lambda})T}\right)_{ij}\;.$$

Plan of the talk

Markov modulated speeds

Consider a Markov chain Z which evolves in some state space with a generator $\mathbf{M} = (m_{ab})$.

There is an "environment" X which is a CTMC with generator $\mathbf{G} = (g_{ij})$.

When X is in state i, the speed of $Z(t)$ (transition rates) is multiplied by v_i:

$$\text{rate } a \to b \;=\; m_{ab} \times v_i \;.$$

The generator of the process $(Z(t), X(t))$ has transition rates:

$$\begin{array}{lcll} (i, a) & \to & (i, b) & \text{with rate } m_{ab} v_i \\ (i, a) & \to & (j, a) & \text{with rate } g_{ij} \end{array}$$

In block-matrix form:

$$\mathbf{Q} \;=\; \begin{pmatrix} v_1\mathbf{M} + g_{11}\mathbf{I} & g_{12}\mathbf{I} & \dots & g_{1K}\mathbf{I} \\ g_{21}\mathbf{I} & v_2\mathbf{M} + g_{22}\mathbf{I} & & g_{2K}\mathbf{I} \\ \vdots & & \ddots & \\ g_{K1}\mathbf{I} & g_{K2}\mathbf{I} & \dots & v_K\mathbf{M} + g_{KK}\mathbf{I} \end{pmatrix}$$

Or, with the Kronecker notation:

$$\mathbf{Q} \;=\; \mathbf{G} \otimes \mathbf{I} + \mathbf{V} \otimes \mathbf{M} \,.$$

where

$$\mathbf{V} \;=\; \mathtt{diag}(v_1, \dots, v_K) \,.$$

Problem: compute the transition probabilities, whith are the elements of the matrix $e^{\mathbf{Q}t}$. A standard method is to diagonalize $\mathbf{Q}$: find its eigenvalues and eigenvectors.

If one chooses x and y such that:

$$\begin{aligned} x\,\mathsf{M} &= \lambda\, x \\ y &= (a_1 x, \ldots, a_N x) = a \otimes x\,. \end{aligned}$$

Then

$$\begin{aligned} y\,\mathsf{Q} &= (a \otimes x)\;(\mathsf{G} \otimes \mathsf{I} + \mathsf{V} \otimes \mathsf{M}) \\ &= a\mathsf{G} \otimes x\mathsf{I} \;+\; a\mathsf{V} \otimes x\mathsf{M} \\ &= a\,(\mathsf{G} + \lambda \mathsf{V}) \otimes x\,. \end{aligned}$$

It is enough to choose a such that $a(\mathsf{G} + \lambda\mathsf{V}) = \mu a$ for $y\mathsf{Q} = \mu y$ to hold.

Diagonalization Algorithm

Data: an infinitesimal generator $\mathbf{Q}$ obtained by modulating a matrix $\mathbf{G}$ with speeds $v_1, \ldots, v_K$:

$$\mathbf{Q} = \mathbf{G} \otimes \mathbf{I} + \mathbf{V} \otimes \mathbf{M} .$$

Result: the eigenvalues, left and right eigenvectors of the matrix $\mathbf{Q}$ ($\Longrightarrow$ diagonalization of $\mathbf{Q}$).

Algorithm:

- Find the spectral elements of $\mathbf{G}$:

$$\rightarrow \quad (\lambda_i; x_i, y_i) \qquad i = 1..K \ .$$

- For each i, find the spectral elements of $\mathbf{G} + \lambda_i \mathbf{V}$:

$$\rightarrow \quad (\mu_{ij}; a_{ij}, b_{ij}) \qquad i = 1..K, \ j = 1..N \ .$$

- Obtain the spectral elements of $\mathbf{Q}$:

$$\rightarrow \quad (\mu_{ij}; a_{ij} \otimes x_i, b_{ij} \otimes y_i) \qquad i = 1..K, \ j = 1..N \ .$$

Complexity:

- soit N be the sise of the state space, K the number of speeds
- $\mathbf{Q}$ is of size $NK \times NK$
- diagonalizing directly is $O(N^3K^3)$
- this algorithm is $O(K^3 + KN^3)$.

It is not even necessary to store the "big" matrix.

Markov modulated queues

Discrete queues: Markov-modulated arrivals

- exponential/Erlang/Cox service distribution $\longrightarrow$ method of phases,
- general IID services: method of the embedded Markov chain.

Fluid queues:

- partial differential equations (Chapman-Kolmogoroff).

In both cases, the results are:

- Computation through matrix formulas, generating functions, Laplace transforms.
- Spectral expansions of stationary and transient probabilities:

$$\mathbb{P}\{W \leq x; X = i\} \;=\; \sum_p a_{i,p}\, e^{-z_i x} \;.$$

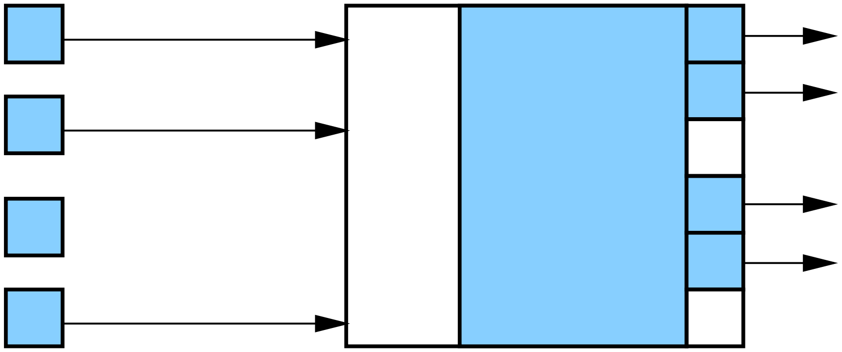
$\longrightarrow$ asymptotics, or bounds.

$$\mathbb{P}\{W \leq x; X = i\} \;\sim\; a_{i,1}\, e^{-z_i x} \;, \qquad x \to \infty \;.$$

The model of Mitra

A fluid model of producer/consumer coupled by a buffer with finite capacity.

Kosten (1982), Anick-Mitra-Sohdhy (1982), Stern-Elwalid-Mitra (198x).



Caracteristics:

- m sources, activity on/off exponential, peak rate r,
- n consumers, activity on/off exponential capacity c,
- buffer capacity X.

The generator of one source is:

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}$$

The arrival process is a MMRP with matrices: generator

$$Q_m = \begin{pmatrix} -m\lambda & m\lambda & & & \\ \mu & -(m-1)\lambda-\mu & (m-1)\lambda & & \\ & 2\mu & -(m-2)\lambda-\mu & (m-2)\lambda & \\ & & \ddots & \ddots & \\ & & & m\mu & -m\mu \end{pmatrix}$$

and rate matrix.

$$\mathbf{\Lambda}_m = \begin{pmatrix} 0 & & & & \\ & 1 & & & \\ & & 2 & & \\ & & & \ddots & \\ & & & & mr \end{pmatrix}$$

The consumption follows a similar process with parameters (ν, τ) and $c \to Q_c$, $\mathbf{\Lambda}_c$.

Observe: $M(t)$ is the superposition of m on/off sources, but the generator has been *agregated* and its size is $m+1$ and not 2^m.

The superposed process $(M(t), C(t))$ has a generator:

$$Q = Q_1 \oplus Q_2 = Q_1 \otimes I(n+1) + I(m+1) \otimes Q_2 .$$

Each state (i, j) corresponds to a *drift* of the buffer contents $ir - jc$. Hence a drift rate matrix:

$$\mathbf{\Lambda} = \mathbf{\Lambda}_1 \oplus \mathbf{\Lambda}_2 = \mathrm{diag}(ir - jc) .$$

In the stationary regime, the probability $P(x)$ that the buffer has a level less than x, is solution of:

$$\boxed{\frac{\partial}{\partial x}\mathsf{P}(x)\mathbf{\Lambda} \;=\; \mathsf{P}(x)Q\;.}$$

It is proved that:

$$\mathsf{P}(x) \;=\; \sum_i \alpha_i \boldsymbol{\phi}_i \, e^{z_i x}\;.$$

The vectors $\boldsymbol{\phi}_i$ have a decomposition: $\boldsymbol{\phi}_i = \boldsymbol{\phi}_{i,1} \otimes \boldsymbol{\phi}_{i,2}$, where each $\boldsymbol{\phi}_{i,j}$ is solution of a smaller linear problem: find (z, ϕ) such that:

$$z\,\boldsymbol{\phi}\,\mathbf{\Lambda} \;=\; \boldsymbol{\phi}\,Q\;.$$

Bibliography

Fluid models

D. Anick, D. Mitra, and M.M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Sys. Tech. J.*, 61:1871–1894, October 1982.

D. Mitra. Stochastic theory of a fluid models of producers and consumers coupled by a buffer. *Adv. Appl. Prob.*, 20:646–676, 1988.

T.E. Stern and A.I. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Adv. Appl. Prob.*, 23:105–139, 1991.

A.I. Elwalid, D. Mitra, and T.E. Stern. Statistical multiplexing of Markov modulated sources:theory and computational algorihms. In A. Jensen and V.B. Iversen, editors, *Proc. 13th International Teletraffic Congress*, pages 495–500, Copenhagen, 1991. Elsevier Science.

A.I. Elwalid, D. Mitra, and T.E. Stern. A theory of statistical multiplexing of Markov modulated sources: Spectral expansions and algorihms. In W.J. Stewart, editor, *Numerical solution of Markov Chains*, 1991.

A.I. Elwalid and D. Mitra. Statistical multiplexing with loss priorities in rate-based congestion control of high speed networks. *IEEE Trans. Comm.*, 42(11):2989–3002, November 1994.

A.I. Elwalid and D. Mitra. Markovian arrival and service communication systems: Spectral expansions, separability and Kronecker-product forms. In W.J. Stewart, editor, *Computations in the Markov Chains*, pages 507–546. Kluwer, 1995.

MMPP, MAP, BMAP...

M.F. Neuts. The fundamental period of a queue with Markov-modulated arrivals. In *Probability, Statistics and Mathematics: papers in honour of Samuel Karlin*. Academic Press, NY, 1989.

W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.

D.M. Lucantoni, G.L. Choudhury, and W. Whitt. The transient $BMAP/G/1$ queue. *Commun.*

Statist.-Stochastic Models, 10(1):145–182, 1994.

A. Jean-Marie, Z. Liu, P. Nain and D. Towsley, "Computational Aspects of the Workload Distribution in the MMPP/GI/1 Queue". *JSAC*, 1999.

Asymptotics, bounds and equivalent bandwidth

W. Whitt. Tail probabilities with statistical multiplexing and effective bandwidth. *Telecommun. Syst.*, 3:71–107.

D. Artiges and P. Nain. Upper and lower bounds for the multiplexing of multiclass Markovian on/off sources. *Performance Evaluation*, **27&28**, pp. 673–698, 1996.

V.G. Kulkarni. Effective bandwidth for Markov regenerative sources. *Queueing Systems*, **24**, pp. 137–153, 1996.

Z. Liu, P. Nain, and D. Towsley. Exponential bounds with applications to call admission. *JACM*, 44 (2):366–394, 1997.