Brownian network models of ramp metering

Frank Kelly
University of Cambridge
www.statslab.cam.ac.uk/~frank

Conference on
Stochastic Processing Networks
in Honor of J. Michael Harrison

Outline

- Rate control in communication networks (relatively well understood)
- Philosophy: optimization vs fairness
- Ramp metering (very preliminary)

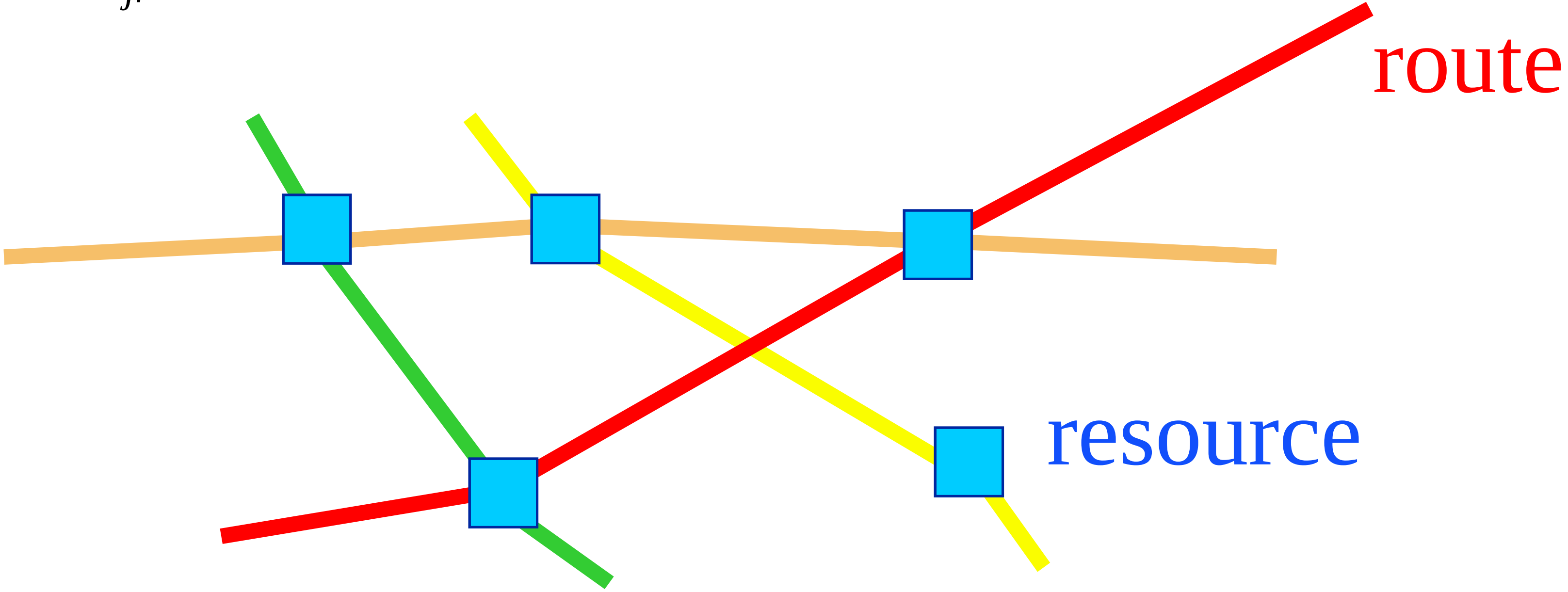
Network structure

J - set of resources

R - set of routes

$A_{jr} = 1$ - if resource j is on route r

$A_{jr} = 0$ - otherwise



Rate allocation

n_r - number of flows on route r

x_r - rate of each flow on route r

Given the vector how are the rates chosen ?

$$n = (n_r, r \in R)$$

$$x = (x_r, r \in R)$$

Optimization formulation

Suppose $x = x(n)$ is chosen to

maximize $\sum_r w_r n_r \frac{x_r^{1-\alpha}}{1-\alpha}$

subject to $\sum_r A_{jr}\, n_r\, x_r \le C_j \quad j \in J$

$$x_r \ge 0 \quad r \in R$$

(weighted α-fair allocations, Mo and Walrand 2000)

$0 < \alpha < \infty$ (replace $\frac{x_r^{1-\alpha}}{1-\alpha}$ by $\log(x_r)$ if $\alpha = 1$)

Solution

$$x_r = \left(\frac{w_r}{\sum_j A_{jr} p_j(n)} \right)^{1/\alpha} \qquad r \in R$$

where

$$\sum_r A_{jr} n_r x_r \le \quad C_j \quad j \in J; \quad x_r \ge 0 \quad r \in R$$

$$p_j(n) \ge 0 \quad j \in J$$

$$p_j(n) \left(C_j - \sum_r A_{jr} n_r x_r \right) \ge 0 \quad j \in J$$

KKT conditions

$p_j(n)$ - *shadow price* (Lagrange multiplier) for the resource j capacity constraint

Examples of α-fair allocations

maximize $\displaystyle \sum_r w_r n_r \frac{x_r^{1-\alpha}}{1-\alpha}$

subject to $\displaystyle \sum_r A_{jr} n_r x_r \le C_j \quad j \in J$

$$x_r \ge 0 \quad r \in R$$

$$x_r = \left(\frac{w_r}{\sum_j A_{jr} p_j(n)} \right)^{1/\alpha} \quad r \in R$$

$\alpha \to 0 \quad (w = 1)$ - maximum flow

$\alpha \to 1 \quad (w = 1)$ - proportionally fair

$\alpha = 2 \quad (w_r = 1/T_r^2)$ - TCP fair

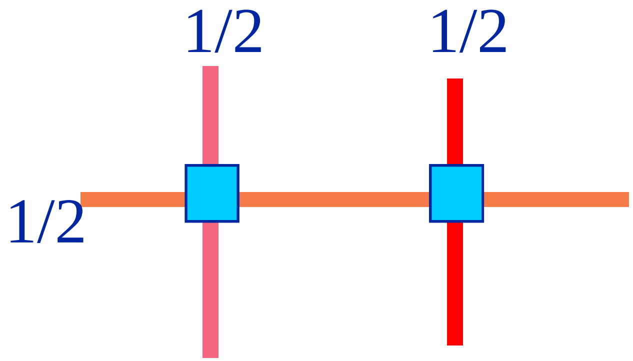
$\alpha \to \infty \quad (w = 1)$ - max-min fair

Example

$$n_r = 1,\ \ w_r = 1 \ \ r \in R,$$
$$C_j = 1 \ \ j \in J$$

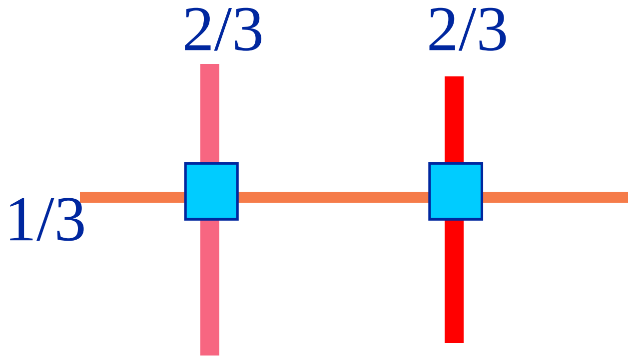
max-min fairness:

$$\alpha \to \infty$$



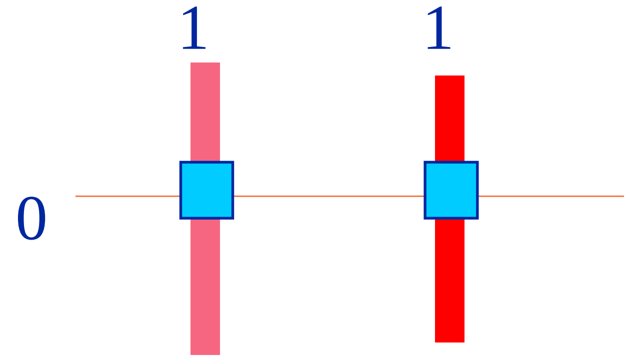
proportional fairness:

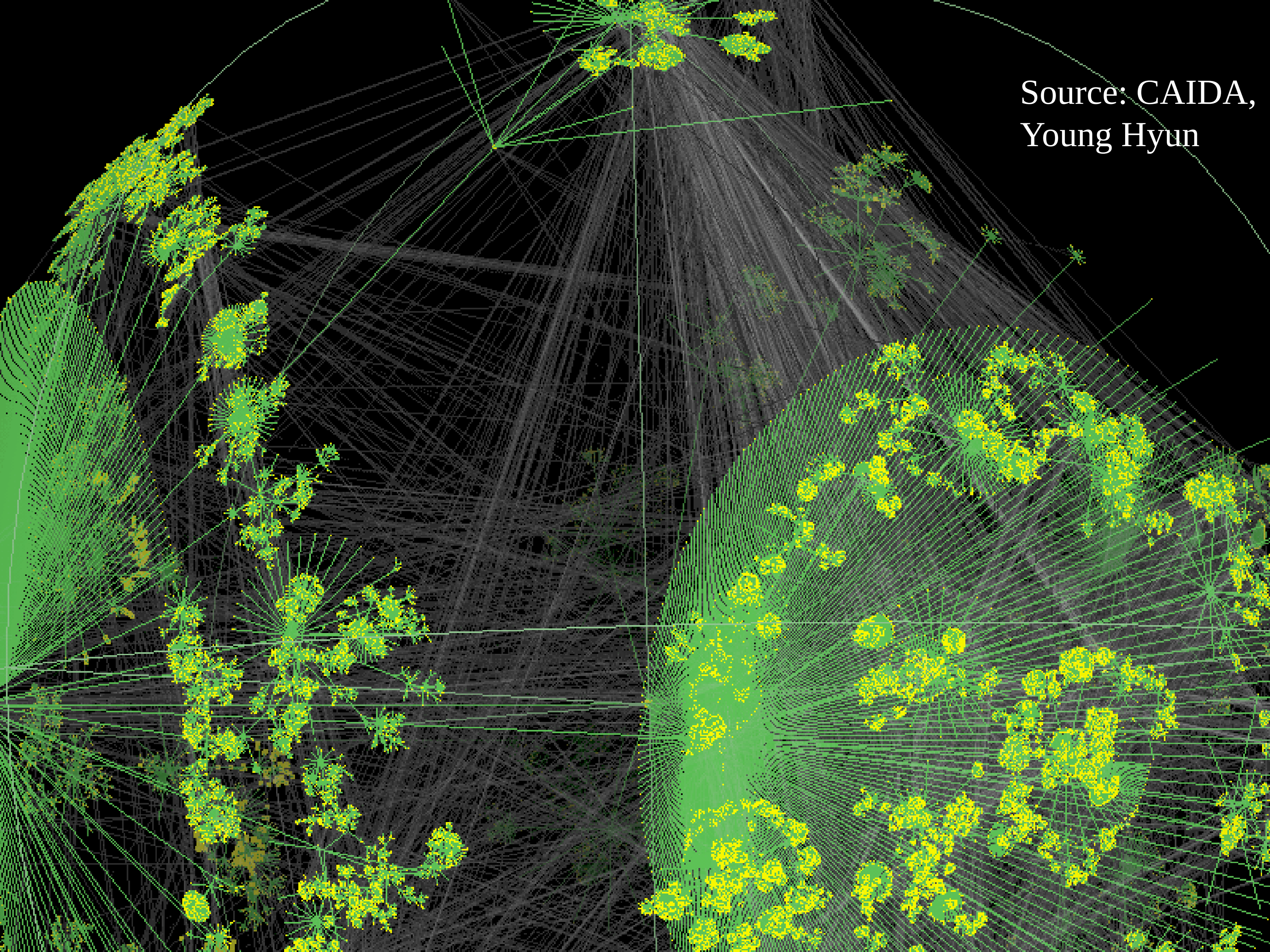
$$\alpha = 1$$



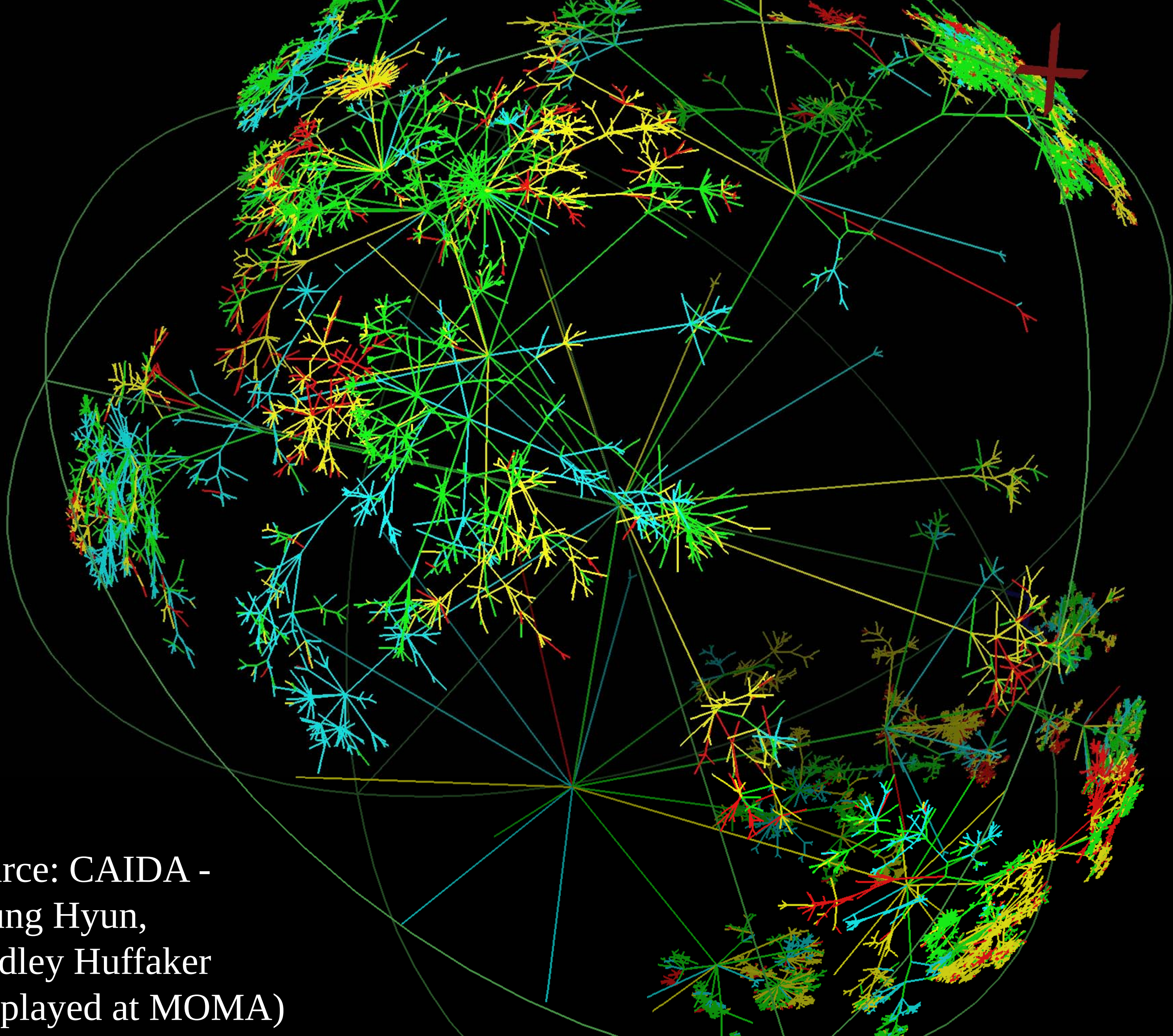
maximum flow:

$$\alpha \to 0$$





Source: CAIDA,
Young Hyun



Source: CAIDA -
Young Hyun,
Bradley Huffaker
(displayed at MOMA)

Flow level model

Define a Markov process $n(t) = (n_r(t), r \in R)$
with transition rates

$n_r \to n_r + 1$	at rate	ν_r	$r \in R$
$n_r \to n_r - 1$	at rate	$n_r \; x_r(n) \; \mu_r$	$r \in R$

- Poisson arrivals, exponentially distributed file sizes

Roberts and Massoulié 1998

Stability

Let $$\rho_r = \frac{\nu_r}{\mu_r} \quad r \in R$$

If $$\sum_r A_{jr}\,\rho_r \; < \; C_j \quad j \in J$$

then the Markov chain is positive recurrent $$n(t) = (n_r(t), r \in R)$$

De Veciana, Lee & Konstantopoulos 1999;
Bonald & Massoulié 2001

Heavy traffic: balanced fluid model

$\boxed{\alpha = 1}$

The following are equivalent:

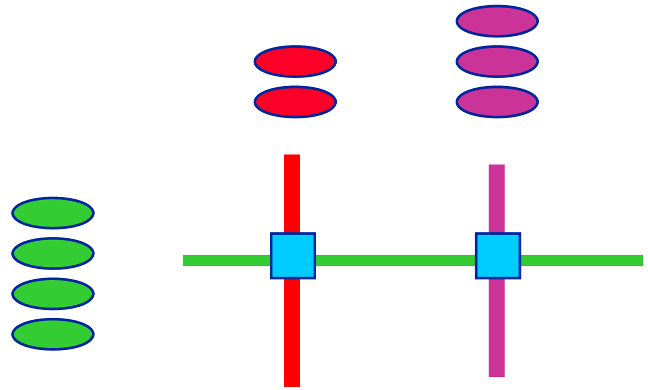
- n is an invariant state
- there exists a non-negative vector p with

$$n_r = \frac{\nu_r}{\mu_r} \sum_j A_{jr} p_j \quad r \in R$$

Thus the set of invariant states forms a J dimensional subspace, parameterized by p.

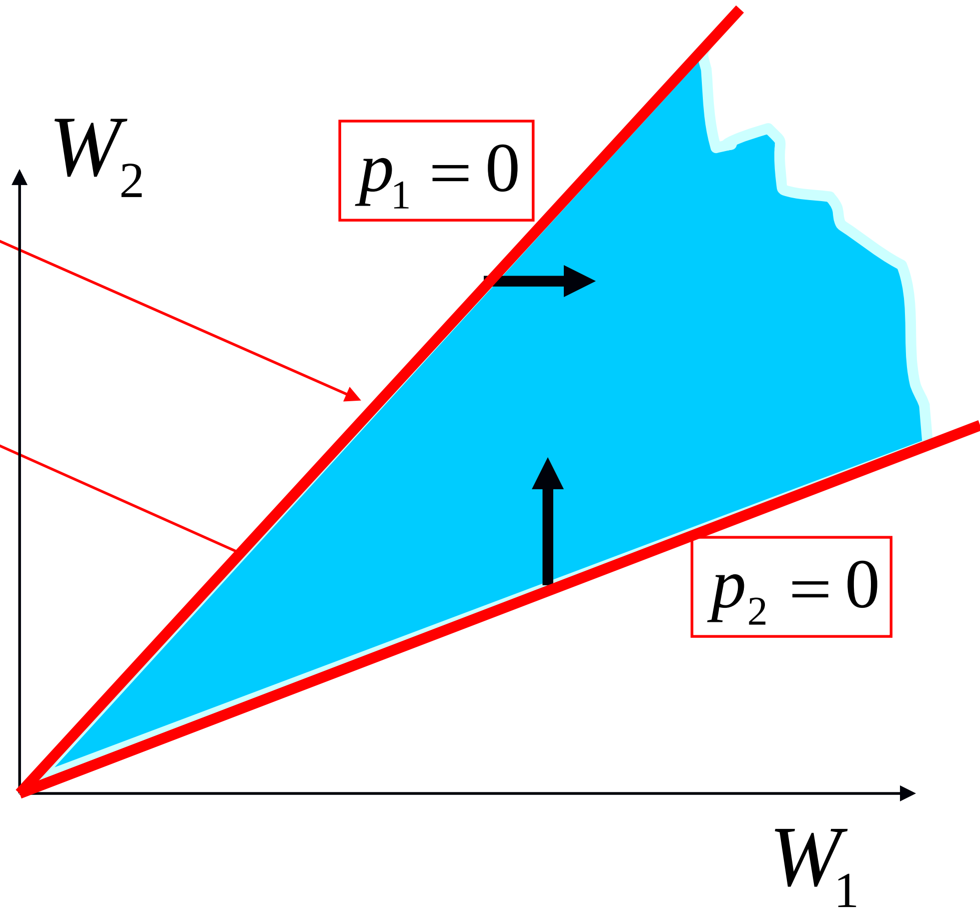
Example

$$\mu_r = 1, \quad r \in R$$



slope $\dfrac{\rho_2 + \rho_0}{\rho_0}$

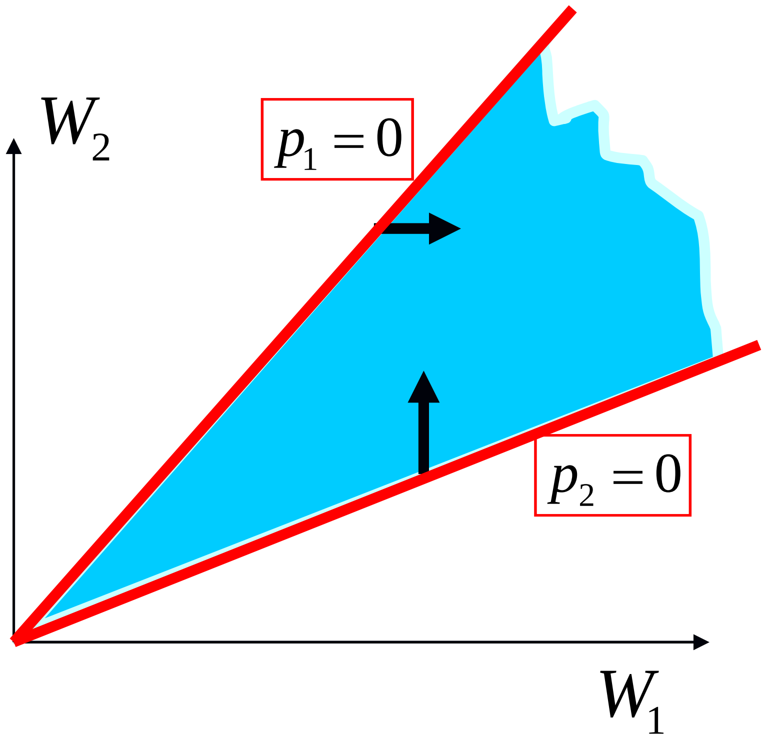
slope $\dfrac{\rho_0}{\rho_1 + \rho_0}$

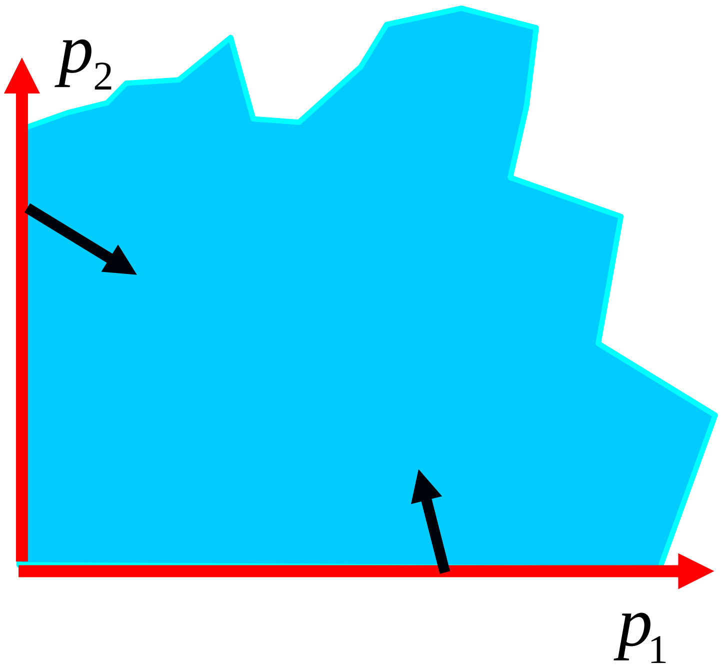


Each bounding face corresponds to a resource not working at full capacity

Entrainment: congestion at some resources may prevent other resources from working at their full capacity.

Stationary distribution?





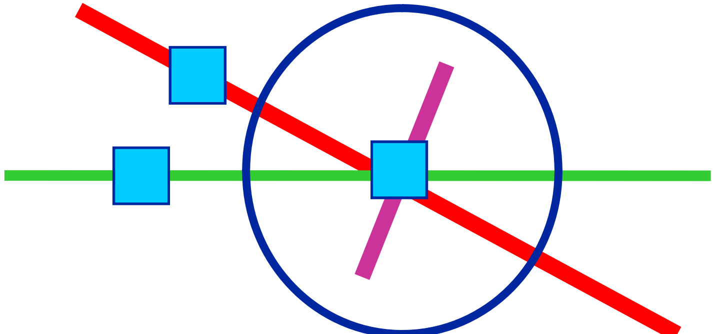
Williams (1987) determined sufficient conditions, in terms of the reflection angles and covariance matrix, for a SRBM in a polyhedral domain to have a product form invariant distribution – a skew symmetry condition

Local traffic condition

Assume the matrix A contains the columns of the unit matrix amongst its columns:

$$A = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 1 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 1 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 1 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

i.e. each resource has some local traffic -



Product form under proportional fairness

Kang, K, Lee and Williams 2009

$$\alpha = 1,\ w_r = 1,\ \ r \in R$$

Under the stationary distribution for the reflected Brownian motion, the (scaled) components of p are independent and exponentially distributed. The corresponding approximation for n is

$$n_r \approx \rho_r \sum_j A_{jr} p_j \qquad r \in R$$

where

$$p_j \sim \mathrm{Exp}(C_j - \sum_r A_{jr} \rho_r) \qquad j \in J$$

Dual random variables are independent and exponential

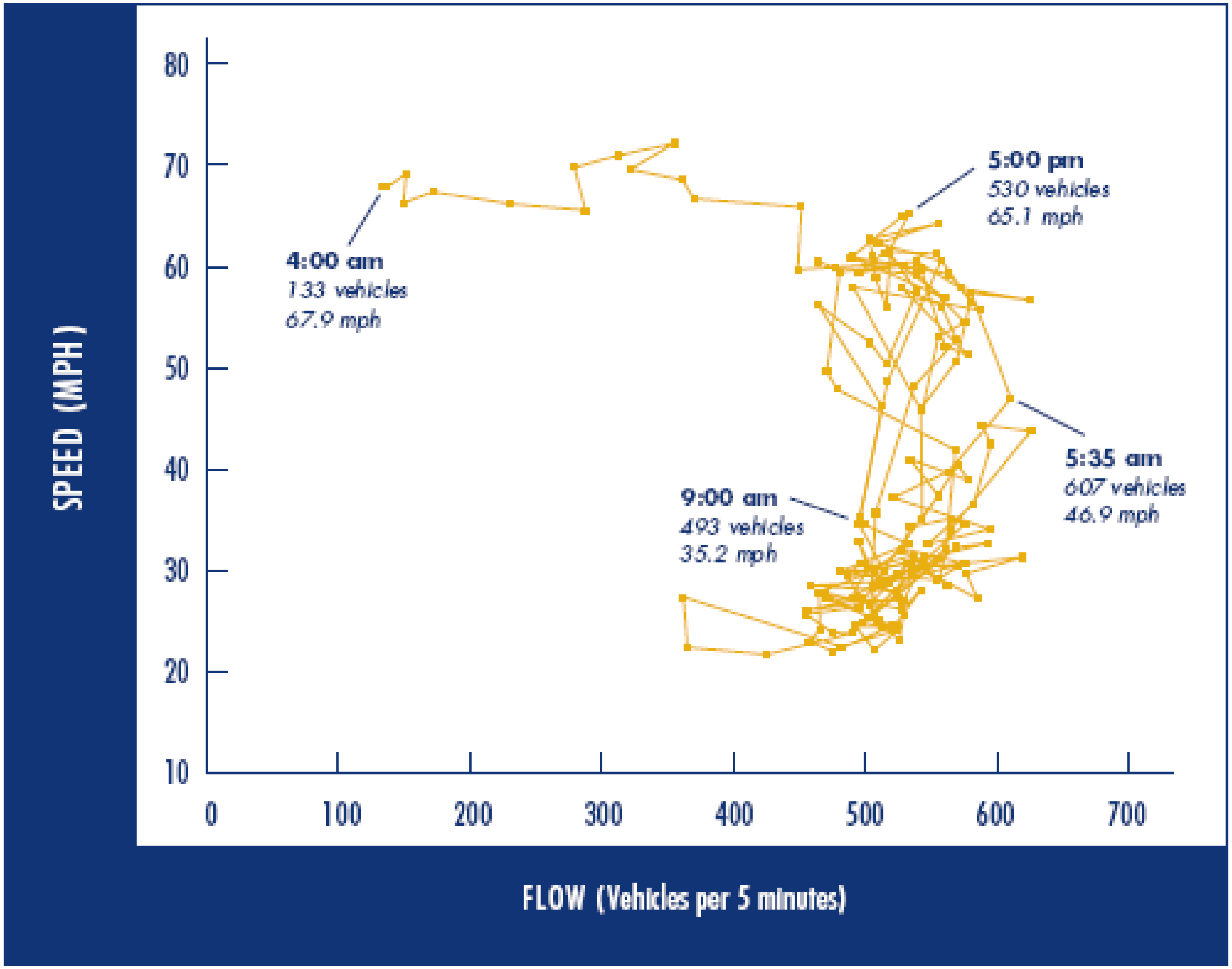


FIGURE 1

Speed vs. flow on I-10 westbound in 5 minute intervals from 4:00 am to 6:00 pm

[What we've learned about highway congestion]

P. Varaiya, Access 27, Fall 2005, 2-9.

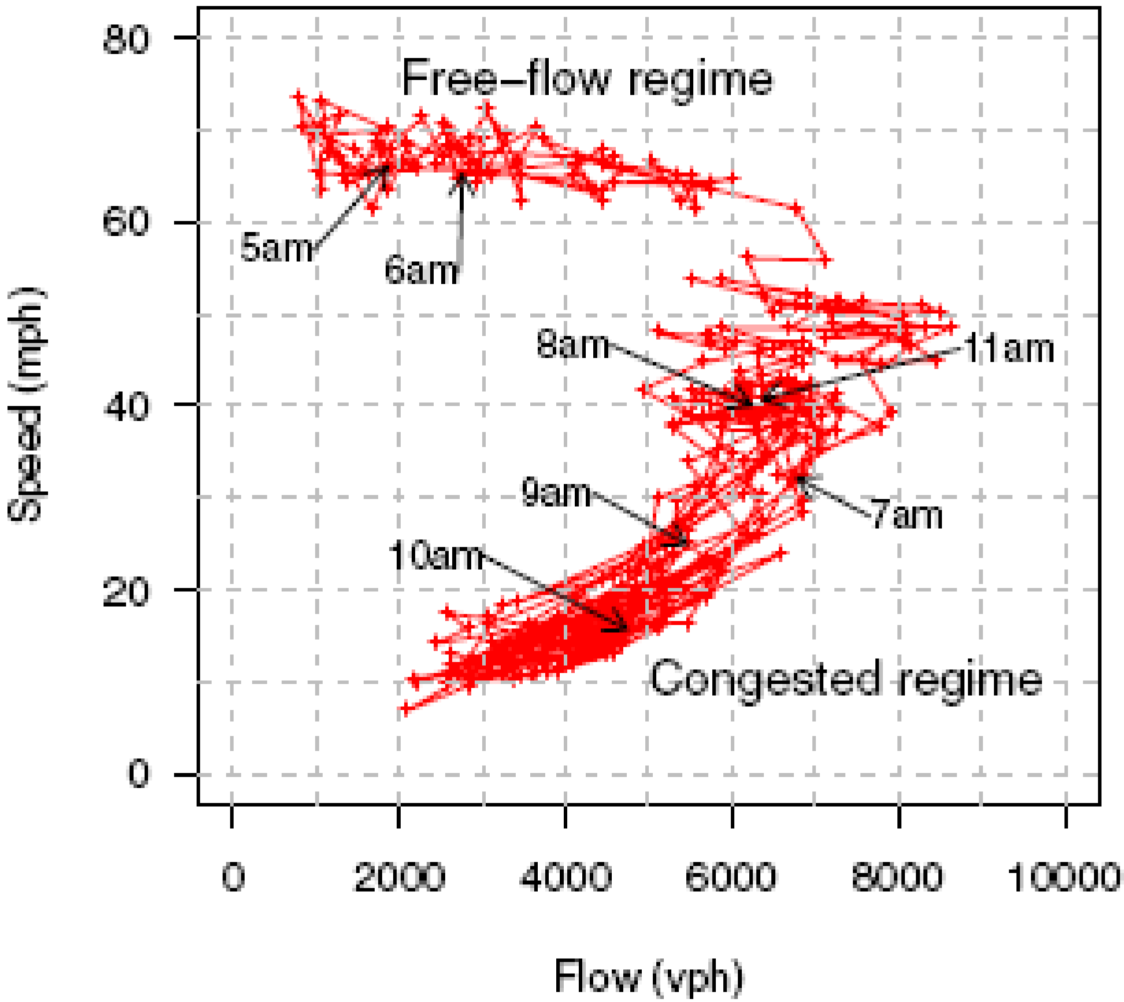
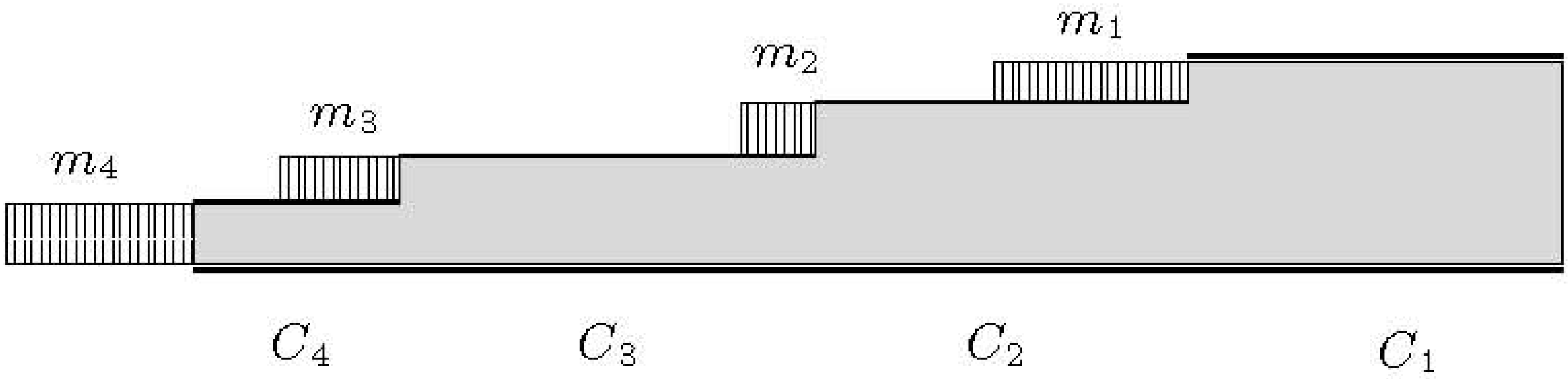


Figure 2. The relationship between the speed and flow of vehicles observed on the morning of Wednesday, 14 July 2004 using the M25 midway between junctions 11 and 12 in the clockwise direction. In the free-flow regime, flow rapidly increases with only a modest decline in speeds. Above a critical occupancy of vehicles there is a marked drop in speed with little, if any, improvement in flow which is then followed by a severe collapse into a congested regime where both flow and speed are highly variable and attain very low levels. Finally, the situation recovers with a return to higher flows and an improvement in speeds

Data, modelling and inference in road traffic networks
R.J. Gibbens and Y. Saatci
Phil. Trans. R. Soc. A366 (2008), 1907-1919.

A linear network



$$m_i(t) = m_i(0) + e_i(t) - \int_0^t \Lambda_i(m(s))\mathrm{d}s, \quad t \geq 0$$

queue size

cumulative inflow

metering rate

Metering policy

Suppose the metering rates can be chosen to be any vector $\Lambda = \Lambda(m)$ satisfying

$$\sum_i A_{ji}\, \Lambda_i \le \quad C_j, \quad j \in J$$

$$\Lambda_i \ge 0, \quad i \in I$$

$$\Lambda_i = 0, \quad m_i = 0$$

and such that

$$m_i(t) = m_i(0) + e_i(t) - \int_0^t \Lambda_i(m(s))\mathrm{d}s \ge 0, \quad t \ge 0$$

Optimal policy?

For each of $i = I, I\text{-}1, \ldots\ldots 1$ in turn choose

$$\int_0^t \Lambda_i(m(s))\mathrm{d}s \geq 0$$

to be maximal, subject to the constraints.

This policy minimizes

$$\sum_i m_i(t)$$

for all times t

Proportionally fair metering

Suppose $\Lambda(m) = (\Lambda_i(m), i \in I)$ is chosen to

maximize $$\sum_i m_i \log \Lambda_i$$

subject to $$\sum_i A_{ji} \Lambda_i \leq C_j, \quad j \in J$$

$$\Lambda_i \geq 0, \quad i \in I$$

$$\Lambda_i = 0, \quad m_i = 0$$

Proportionally fair metering

$$\Lambda_i(m) = \frac{m_i}{\sum_j p_j A_{ji}}, \quad i \in I$$

where

$$\boxed{\begin{aligned} &\Lambda_i \geq 0, \quad i \in I \\ &\sum_i A_{ji}\,\Lambda_i \leq C_j, \quad j \in J \\ &p_j \geq 0, \quad j \in J \\ &p_j\left(C_j - \sum_i A_{ji}\,\Lambda_i\right) \geq 0, \quad j \in J \end{aligned}}$$

KKT conditions

p_j - *shadow price* (Lagrange multiplier) for the resource j capacity constraint

Brownian network model

Suppose that $(e_i(t), t \geq 0)$ is a Brownian motion, starting from the origin, with drift ρ_i and variance $\rho_i \sigma^2$. Let

$$X_j(t) = \sum_i A_{ji} e_i(t) - C_j t$$

Then $X(t) = (X_j(t), j \in J)$ is a J-dimensional Brownian motion starting from the origin

with drift $-\theta = A\rho - C$

and variance $\Gamma = \sigma^2 A[\rho]A'$

Brownian network model

Let $\quad \mathbf{W} = A[\rho]A'\mathbf{R}_+^J$

and $\quad \mathbf{W}^j = \{ A[\rho]A' : \quad q \in \mathbf{R}_+^J, \quad q_j = 0 \}.$

Define $W(t)$ by the following relationships:

$(i) \quad W(t) = X(t) + U(t)$ for all $t \geq 0$

$(ii) \quad W$ has continuous paths, $W(t) \in \mathbf{W}$

$(iii) \quad$ for each $j \in J, U_j$ is a one-dimensional process such that

$(a) \ U_j$ is continuous and non-decreasing, with $U_j(0) = 0$,

$(b) \ U_j(t) = \int_0^t I\{W(s) \in \mathbf{W}^j\} \mathrm{d}U_j(s)$ for all $t \geq 0$.

Brownian network model

If $\theta_j > 0,\ j \in J,$ then there is a unique stationary distribution W under which the components of

$$Q = (A[\rho]A')^{-1}W$$

are independent, and Q_j is exponentially distributed with parameter

$$\frac{\sigma^2}{2}\theta_j, \quad j \in J$$

and queue sizes are given by

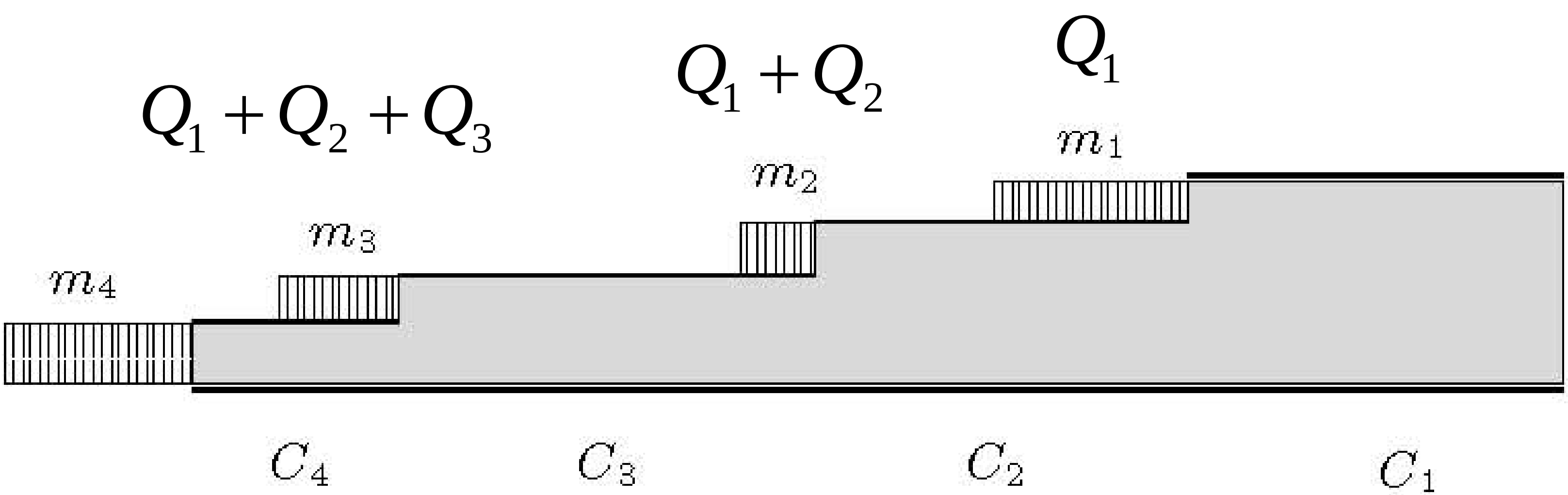
$$M = [\rho]A'Q$$

Delays

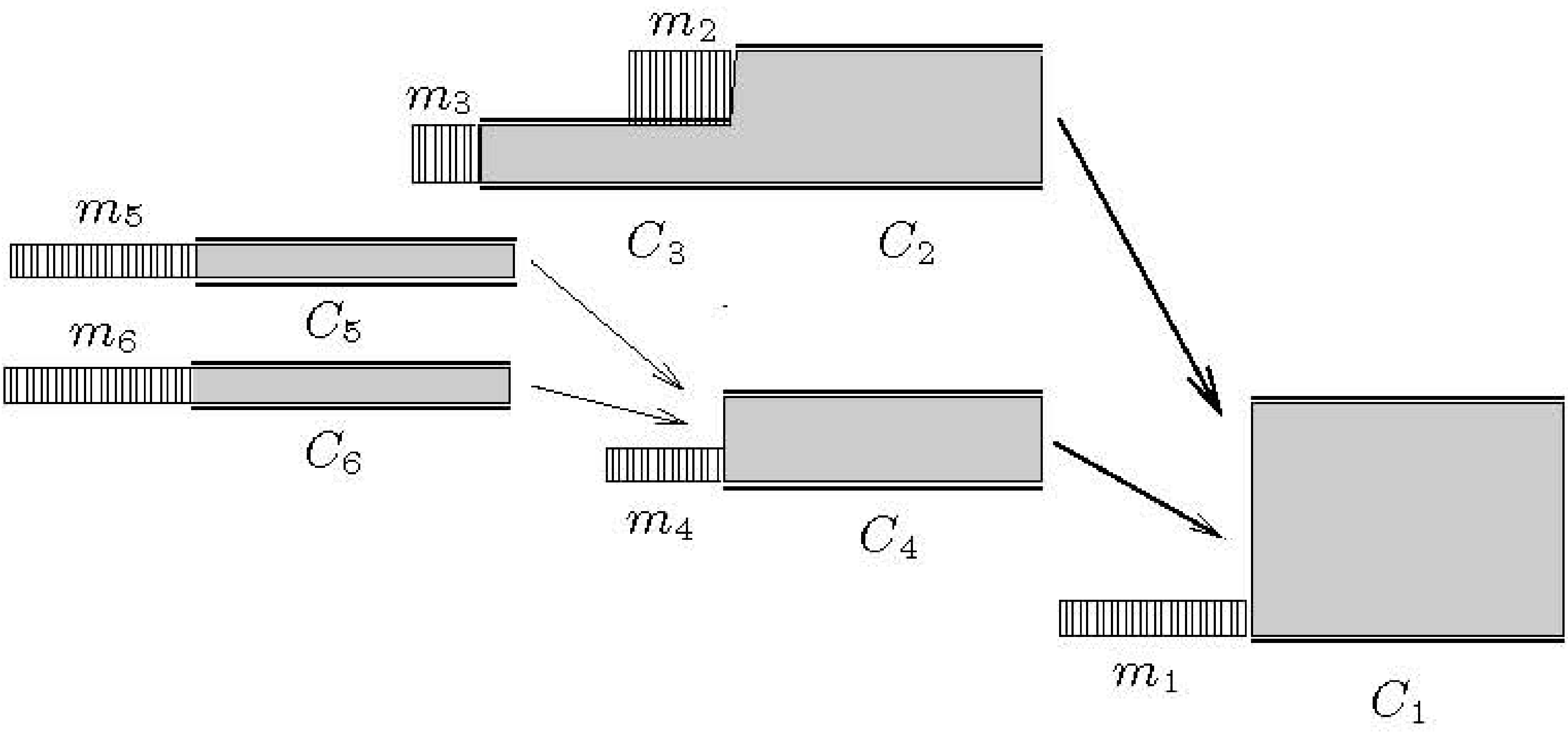
Let $$D_i(m) = \frac{m_i}{\Lambda_i(m)}$$

- the time it would take to process the work in queue i at the current metered rate. Then

$$D_i(M) = \sum_j Q_j A_{ji}$$

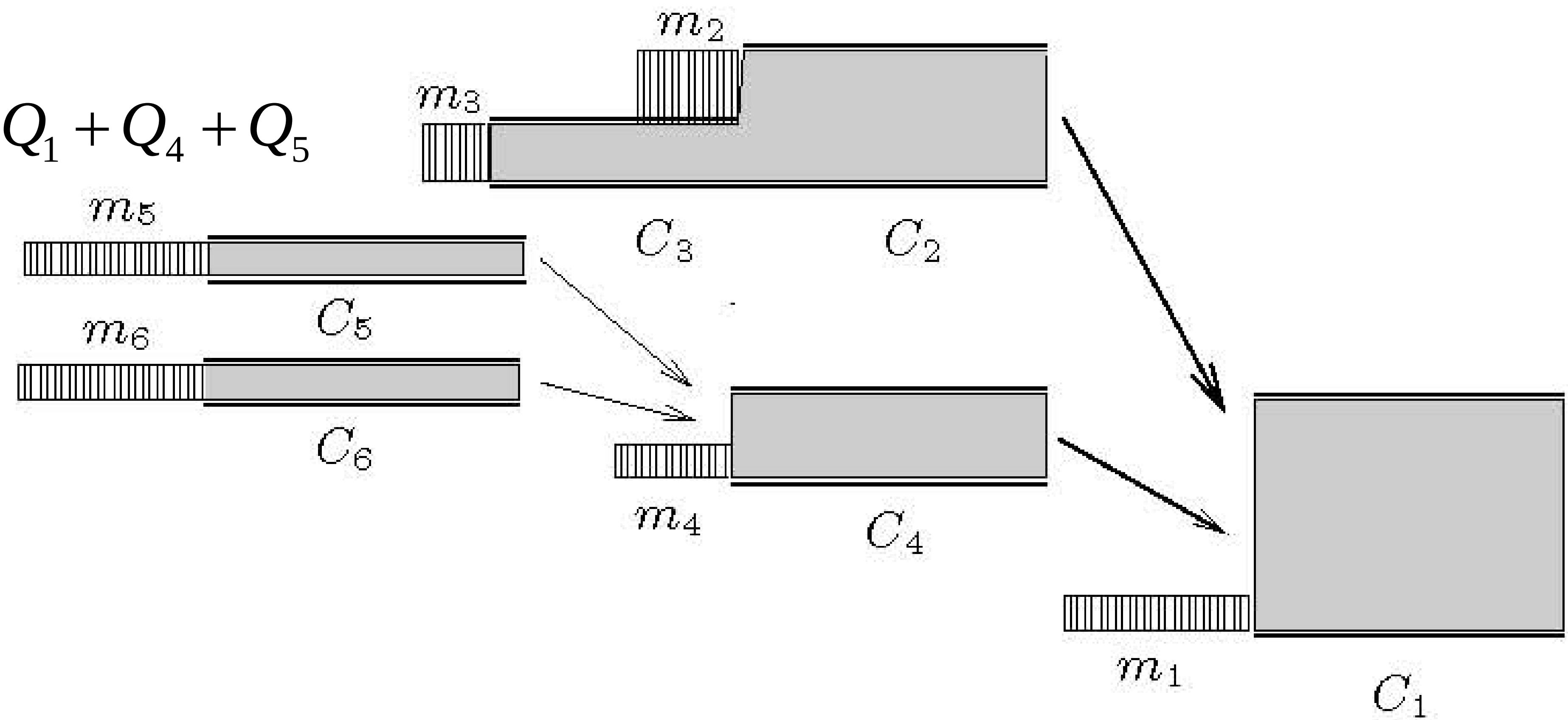


A tree network

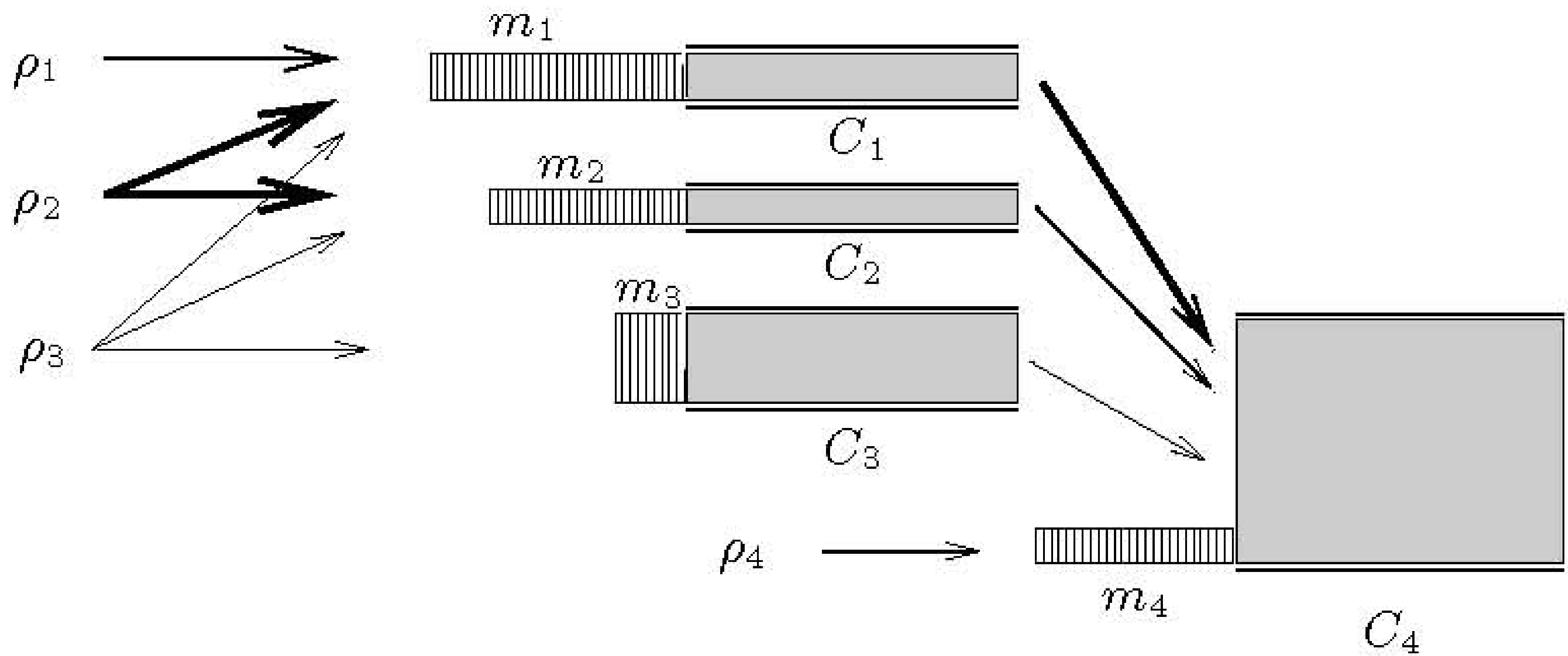


A tree network

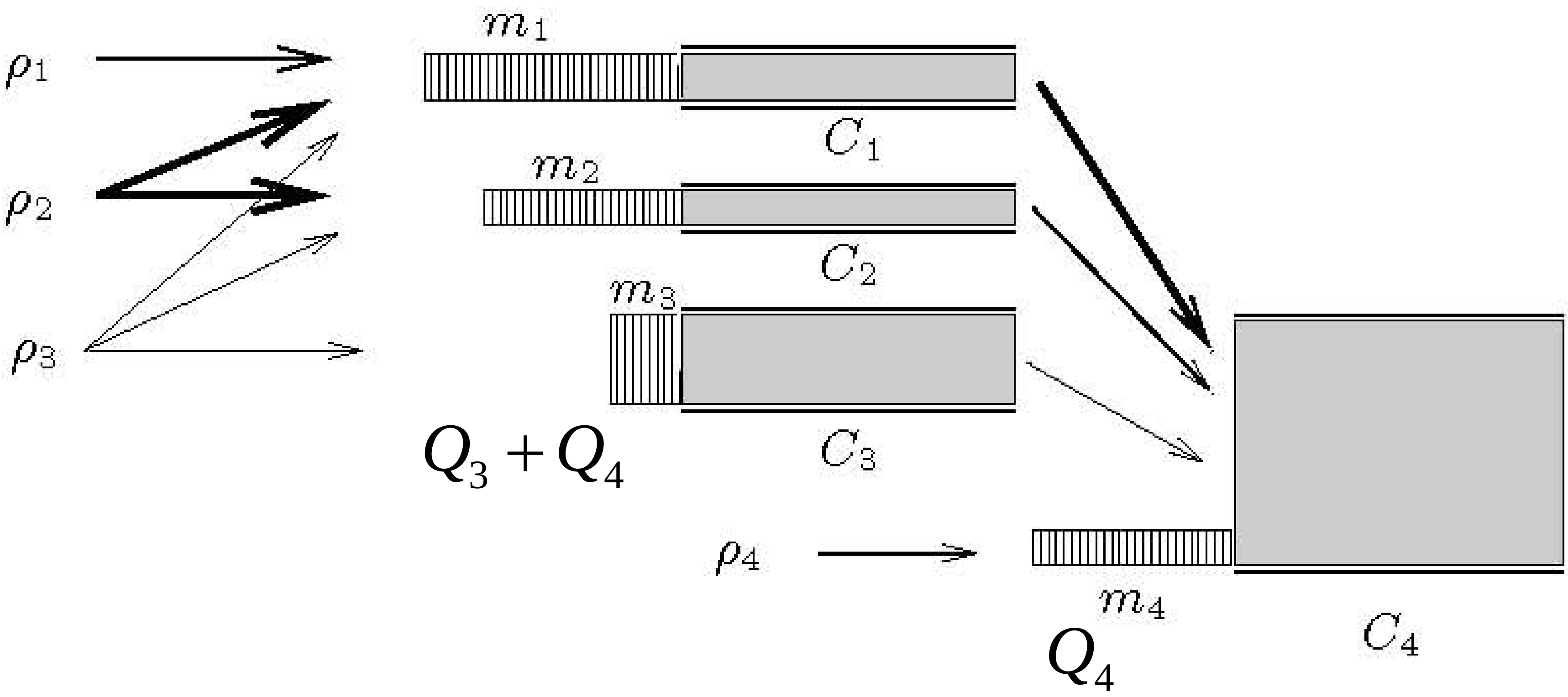
$Q_1 + Q_4 + Q_5$



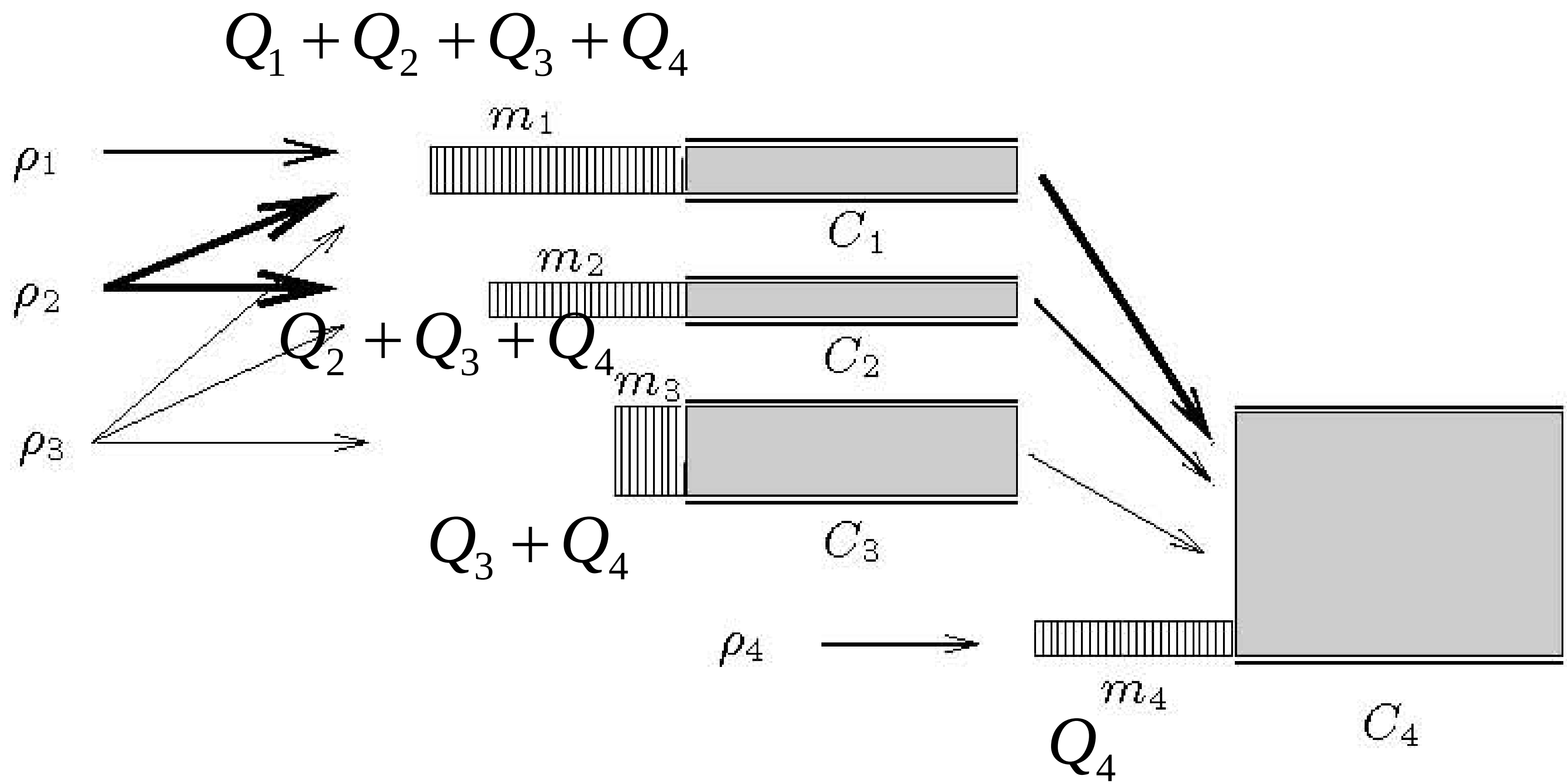
Route choices



Route choices

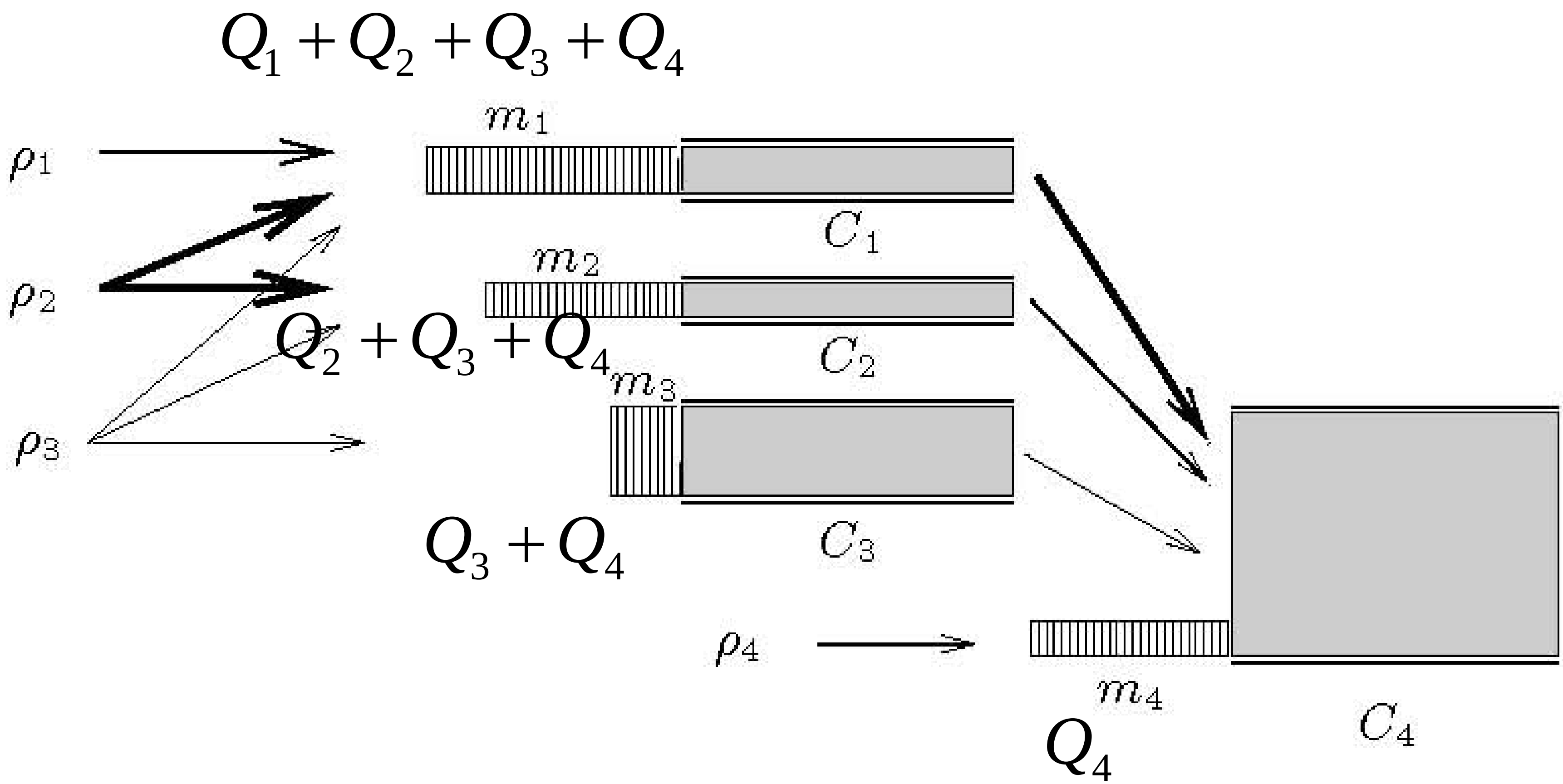


Route choices



Route choices

$$Q_2 \sim \frac{\sigma^2}{2}\mathrm{Exp}(C_1 + C_2 - \rho_1 - \rho_2)$$



Brownian network model

- As Mike has cogently argued, for many applications it may be ***easier*** to describe the workload arrival process in terms of the mean and variance of a Brownian motion.
- If relevant time periods long enough, negative increments less likely.
- The Brownian model exposes structure in a way that more detailed models (e.g. MDP models) do not.