Capturing blocking and spillback in finite capacity queuing networks

Carolina Osorio

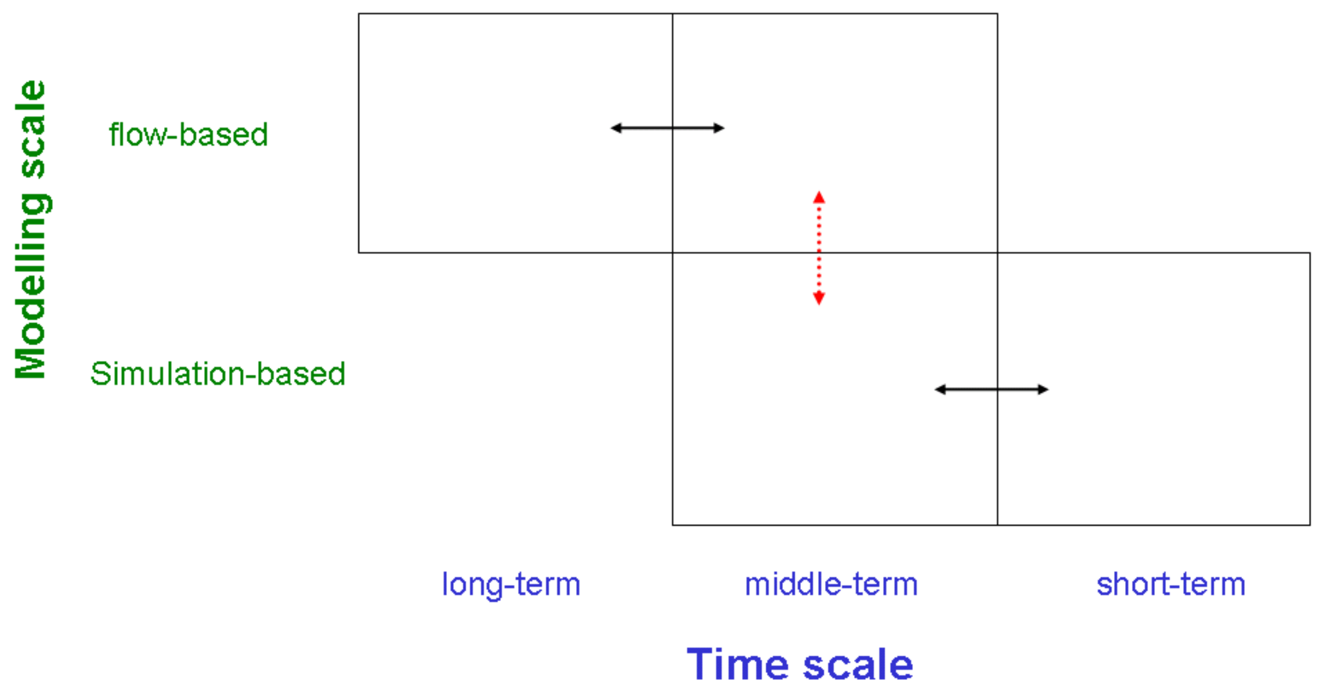
Transport and Mobility Laboratory, EPFL

May 2007

Outline

- finite capacity queuing network framework
- model description
- validation
- case study

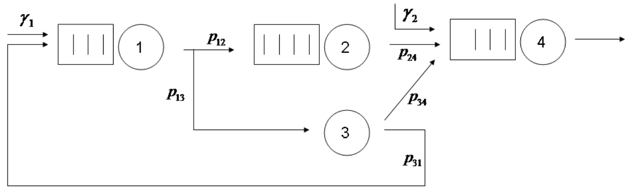
Overall objectives



Current phase: define aggregate model

Finite capacity networks

Aim: estimate network performance



How can we model these networks?

Approach: queueing theory.

Queueing networks

- Jackson networks
  - infinite buffer size assumption
  - violated in practice

Between-queue correlation structure

- complex to grasp
- helps explain: blocking, spillbacks, deadlocks, chained events

If these events want to be acknowledged:

$\boxed{\text{finite capacity queueing networks}}$

Finite capacity queueing networks FCQN

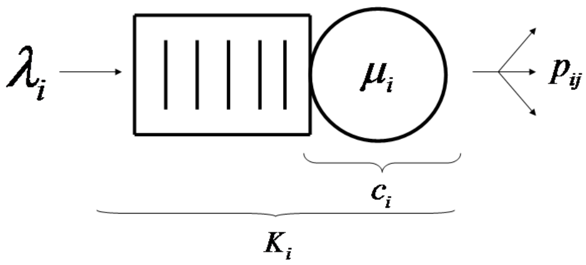
Main application fields:

- software architectures performance prediction
- telecommunications
- manufacturing systems

More uncommon applications:

- pedestrian flow through circulation systems
- prisoner flow through a network of prisons with varying security levels
- hospital patient flow

Queueing: framework



- c_i parallel servers
- K_i total capacity: nb serveurs + queueing slots
- λ_i: average arrival rate
- μ_i: average service rate
- p_{ij}: transition probabilities (routing)

- station (queue)
- job

FCQN methods

We can evaluate the main network performance measures using the joint stationary distribution, π.

$$\pi = (P(N_1 = n_1, ..., N_S = n_S), \quad (n_1, ..., n_S) \in (\mathcal{S}_1, ..., \mathcal{S}_S))$$

1. **Closed form expression**
   - product-form dbn: (Jackson, BCMP)
   - small networks: two-station single server with either tandem or closed topology

For more general topology networks:

2. **Exact numerical evaluation**
3. **Approximation methods: decomposition methods**

Exact numerical methods

$$\begin{cases} \pi Q = 0 \\ \sum_{s \in \mathcal{S}} \pi_s = 1 \end{cases}$$

π: stationary dbn of the network

Q: network transition rate matrix

$\mathcal{S}$: state space

For **each network state** we define:

- all possible transitions to other states
- their corresponding rates

Disadvantages:

- **untractable**: limited to small networks
- **not flexible**: changes in the configuration or topology: redefine Q

A more flexible approach: decomposition methods.

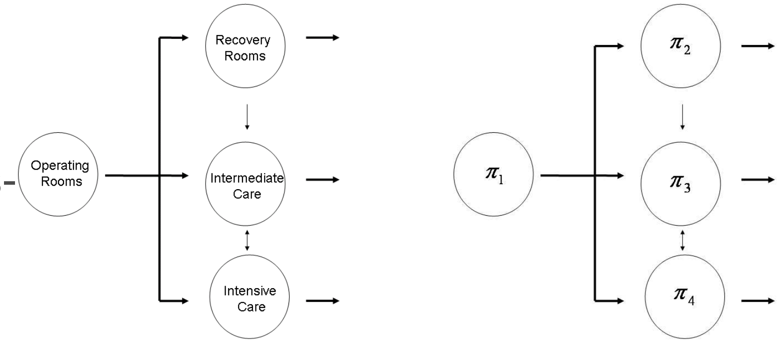
Decomposition methods

By decomposing we can aim at analysing:

- arbitrary topology and size

Method description

1. decompose the network into subnetworks
2. analyse each subnetwork independently: estimates of the marginal dbns
3. estimate the main performance measures



Subnetwork

- size: single queues
- analysis using global balance equations.
- obtain estimates of the marginal dbns

Current objective

Existing methods mainly concern

- single server + feed-forward network
- multiple server + tandem

For multiple server + arbitrary topology:

- revise queue capacities (endogenous)
- vary network topologies (analogy with closed form dbn networks)

Requires:

- approximations to ensure integrality of endogenous capacities
- aposteriori validation (e.g. check positivity)

$\boxed{\text{unsuitable for an optimization framework}}$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Current objective

- multiple server + arbitrary topology + BAS
- preserving initial network configuration (topology + capacities)
- **explicitly** model blocking events

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Global balance equations

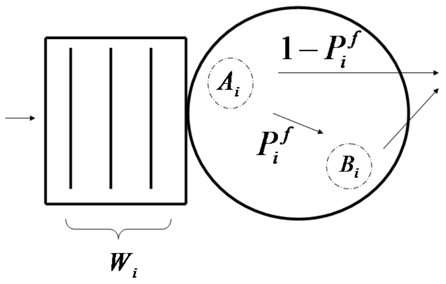
$$\begin{cases} \pi(i)Q(i) = 0 \\ \sum_{s \in \mathcal{S}(i)} \pi(i)_s = 1 \end{cases}$$

$\pi(i)$: stationary dbn of station i
$Q(i)$: transition rate matrix
$\mathcal{S}(i)$: state space

State space

Upon arrival to a queue a job :

1. [queue]
2. is served
3. [blocked]
4. departs



State space of station i :

$$\mathcal{S}_i = \{(A_i, B_i, W_i) \in \mathbb{N}^3, A_i + B_i \leq c_i, W_i \leq K_i - c_i\}$$

We want to estimate:

$$\pi(i) = (P((A_i, B_i, W_i) = (a, b, w))\ \forall (a, b, w) \in \mathcal{S}(i))$$

Transition rates

$Q(i)$: effective arrival rates, effective service rates $\Big\}$ stationary dbn of the subnetwork

$\updownarrow$

marginal stationary dbn of the network

For a given station how can we estimate the

- effective arrival rates ?
- effective service rates ?

Main challenge and complexity lies in appropriatly acknowledging the correlation between the stations i.e. in approriatly revising these structural parameters.

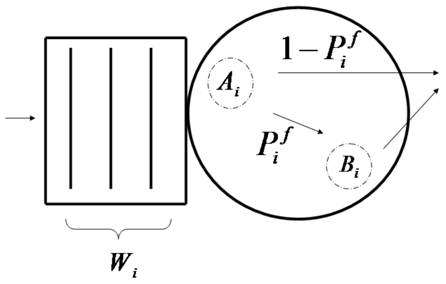
TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Transition rates

Upon arrival to a queue a job :

1. [queue]
2. is served
3. [blocked]
4. departs



Grasping the between station correlation implies appropriately estimating the transition rates between these states.

Transition rates

$Q(i)$ is a function of:

- λ_i, μ_i: average arrival and service rate
- P_i^f: average blocking probability
- $\tilde{\mu}(i,b)$: average unblocking rate given that there are b blocked jobs

Consider station i which is in state $(A_i, B_i, W_i) = (a, b, w)$.
Then the possible transitions and their rates are:

new state l	rate q^i_{kl}	condition
$(a, b, w+1)$	λ_i	$a + b == c_i$ & $w + 1 \leq K_i - c_i$
$(a+1, b, w)$	λ_i	$a + b + 1 \leq c_i$
$(a-1, b, w)$	$a\mu_i(1 - P_i^f)$	$w == 0$
$(a, b, w-1)$	$a\mu_i(1 - P_i^f)$	$w \geq 1$
$(a-1, b+1, w)$	$a\mu_i P_i^f$	always possible
$(a, b-1, w)$	$\tilde{\mu}(i,b)$	$w == 0$
$(a+1, b-1, w-1)$	$\tilde{\mu}(i,b)$	$w \geq 1$

Lets estimate these parameters ...

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Average blocking probability

$$\boxed{P_i^f = \sum_j p_{ij} P(N_j = K_j)}$$

where $P(N_j = K_j)$ is the probability that station j is full.

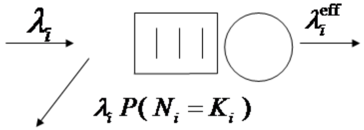
Arrival rates

- λ_i : total arrival rate (includes potentially lost arrivals)
- λ_i^{eff} : the effective arrival rate (excludes lost arrivals)
- γ_i: external arrival rate

1) Loss model:

$$\boxed{\lambda_i^{\text{eff}} = \lambda_i(1 - P(N_i = K_i))}$$

where N_i denotes the total number of jobs at station i ($N_i = A_i + B_i + W_i$).



2) Flow conservation laws hold for the effective arrrival rates:

$$\boxed{\lambda_i^{\text{eff}} = \gamma_i(1 - P(N_i = K_i)) + \sum_j p_{ji}\lambda_j^{\text{eff}}}$$

Inter-arrival times $\sim \varepsilon(\lambda_i)$, i.i.d

Service and unblocking rates

When station i is in state (a, b, w):

1) service rate:

a parallel servers $\Rightarrow$ service rate: $a\mu_i$.

2) unblocking rate:

if there are b blocked jobs at station i:

how many parallel blocked queues are there ?

aim: $a\mu_i \quad \longleftrightarrow \quad \tilde{\mu}(i, b) = \phi(i, b)\, \tilde{\mu}_i^o$

Service and unblocking rates

aim: $a\mu_i \quad \longleftrightarrow \quad \tilde{\mu}(i,b) = \phi(i,b)\ \tilde{\mu}_i^o$

- one station blocking : $\tilde{\mu}_i^o$
- d distinct destination stations : $d\tilde{\mu}_i^o$
  d virtual **parallel** queues

$\phi(i,b)$ **represents: the average number of blocking stations given that there are b blocked jobs at station** i

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Service and unblocking rates

- $\tilde{\mu}_i^o$ approach: average "inter-unblocking times" across destination stations

$$\boxed{\frac{1}{\tilde{\mu}_i^o} = \sum_{j \in \mathcal{I}+} \frac{\lambda_j^{\text{eff}}}{\lambda_i^{\text{eff}} \hat{\mu}_j c_j}}$$

- $\phi(i,b)$ approach: condition on the number of distinct stations that are blocking the b jobs.

$$\frac{1}{\tilde{\mu}(i,b)} = \sum_{d=1}^{min(b,card(\mathcal{I}^+))} P(D(i,b)=d) \frac{1}{d\,\tilde{\mu}_i^o} = \frac{1}{\tilde{\mu}_i^o} \sum_{d=1}^{min(b,card(\mathcal{I}^+))} \frac{1}{d} \sum_{l_i \in L} \frac{b!}{\prod_{j \in \mathcal{I}+} l_{ij}!} \prod_{j \in \mathcal{I}+} \tilde{p}_{ij}^{l_{ij}}$$

adding an assumption ...

$$\boxed{\tilde{\mu}(i,b) = \tilde{\mu}_i^o\, \phi(i,b)}$$

where $\phi(i,b)$ is now exogenous

- Service time $\sim \varepsilon(\mu_i)$, i.i.d
- Time between unblockings $\sim \varepsilon(\tilde{\mu}_i^o)$, i.i.d

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Summary

Aims were:

- decompose the network into single stations
- solve the global balance equations associated to each station:

$$\begin{cases} \pi(i)Q(i) = 0 \\ \sum_{s \in \mathcal{S}(i)} \pi(i)_s = 1 \end{cases}$$

- define $\mathcal{S}(i)$
- estimate $Q(i) = f(\lambda_i, \mu_i, P_i^f, \tilde{\mu}(i,b))$
- estimate the transition rates

Summary

$$
\mathcal{E}(i) = \left\{
\begin{array}{rcl}
\pi(i)Q(i) & = & 0 \\
\sum_{s \in \mathcal{S}(i)} \pi(i)_s & = & 1 \\
 & & \\
Q(i) & = & f(\lambda_i, \mu_i, P_i^f, \tilde{\mu}(i,b)) \\
\lambda_i^{\text{eff}} & = & \lambda_i (1 - P(N_i = K_i) \\
\lambda_i^{\text{eff}} & = & \gamma_i (1 - P(N_i = K_i)) + \sum_j p_{ji} \lambda_j^{\text{eff}} \\
P_i^f & = & \sum_j p_{ij} P(N_j = K_j) \\
\tilde{\mu}(i,b) & = & \tilde{\mu}_i^o \; \phi(i,b) \\
\frac{1}{\tilde{\mu}_i^o} & = & \sum_{j \in \mathcal{I}+} \frac{\lambda_j^{\text{eff}}}{\lambda_i^{\text{eff}} \hat{\mu}_j c_j} \\
\frac{1}{\hat{\mu}_i} & = & \frac{1}{\mu_i} + P_i^f \frac{1}{\mu_i^{\tilde{avg}}} \\
\frac{1}{\tilde{\mu}_i^{avg}} & = & \sum_{b \geq 1} \frac{P(B_i = b)}{P(B_i > 0)} \sum_{k=1}^{b} \frac{k}{b} \frac{1}{\tilde{\mu}(i,k)} \\
P(N_i = K_i) & = & \sum_{s \in \mathcal{F}(i)} \pi(i)_s \\
P(B_i = b) & = & \sum_{s = (.,b,.) \in \mathcal{S}(i)} \pi(i)_s \\
P(B_i > 0) & = & 1 - \sum_{s = (.,0,.) \in \mathcal{S}(i)} \pi(i)_s
\end{array}
\right.
$$

- Exogenous : $\{\mu_i, \gamma_i, p_{ij}, c_i, K_i, \phi(i,b)\}$
- All other parameters are endogenous
- MATLAB **fsolve** : route for systems of nonlinear equations.

Method validation

Validation versus:

- pre-existing decomposition methods
- simulation results on a set of small networks
- simulation results on a network of hospital rooms

Validation

Validation versus pre-existing methods

- Kerbache and MacGreggor Smith. 1988. Asymptotic behaviour of the Expansion method for open finite queuing networks. *Computers and Operations Research*
- Altiok and Perros. 1987. Approximate analysis of arbitrary configurations of open queuing networks with blocking. *Annals of Operations Research*
- Boxma and Konheim. 1981. Approximate Analysis of Exponential Queueing Systems with Blocking. *Acta Informatica*
- Takahashi *et al*. 1980. An approximation method for open restricted queuing networks. *Operations research*
- Hillier and Boling. 1967. Finite queues in series with exponential or erlang service times. A numerical approach. *Operations research*

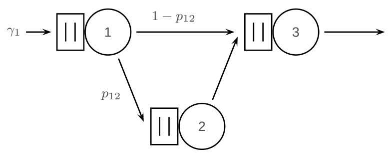
Validation [1]

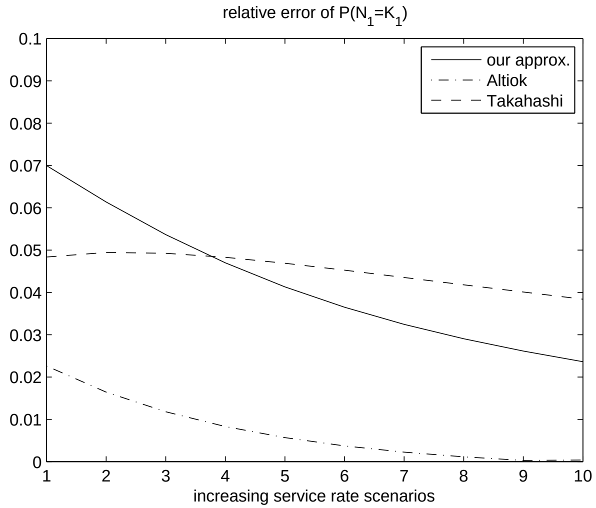
Setting: triangular topology with single-server stations ($c_j = 1$)

$\forall i \; c_i = 1, \; p_{12} = \frac{1}{2}$

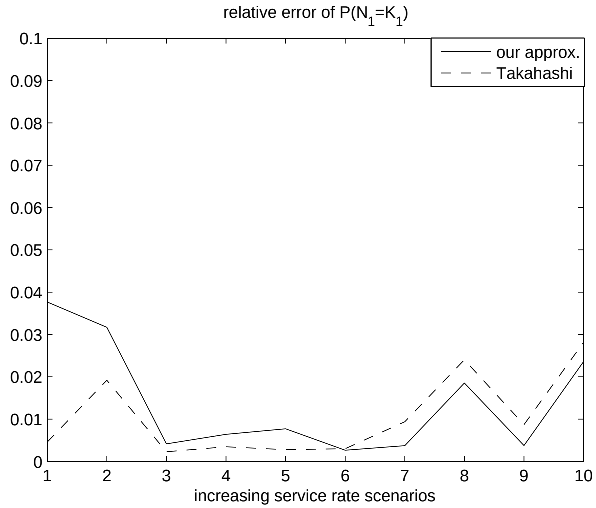
$\gamma_1 = 1, \gamma_2 = \gamma_3 = 0$

scenario	μ_1	μ_2	μ_3
1	1	1.1	1.2
2	1	1.2	1.4
3	1	1.3	1.6
4	1	1.4	1.8
5	1	1.5	2
6	1	1.6	2.2
7	1	1.7	2.4
8	1	1.8	2.6
9	1	1.9	2.8
10	1	2	3





(a) $\forall i \; K_i = 1$



(b) $\forall i \; K_i = 3$

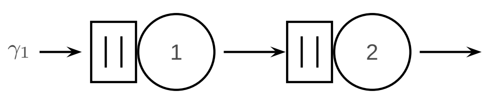
Validation [2]

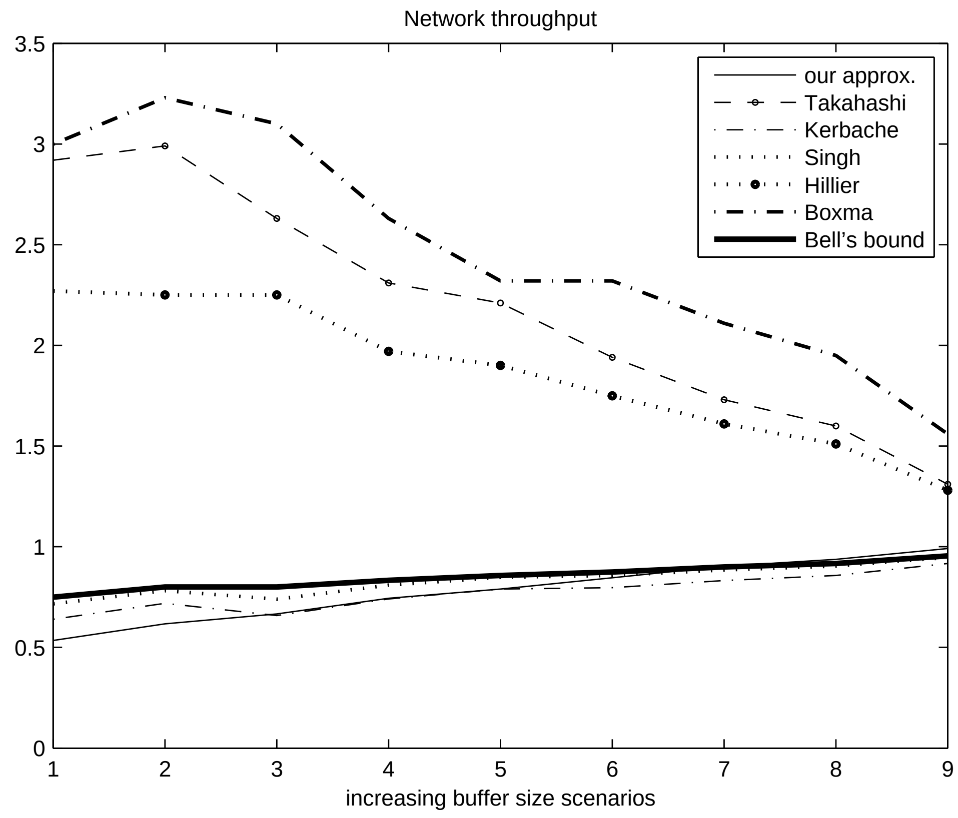
Theoretical bound on the throughput Bell (1982):

$\mu_1 = 3, \mu_2 = 1,\ c_1 = c_2 = 1$

$\gamma_1 = 1, \gamma_2 = 0$

scenario	$K_1 - c_1$	$K_2 - c_2$
1	1	1
2	1	2
3	2	1
4	2	2
5	2	3
6	3	3
7	4	4
8	5	5
9	10	10





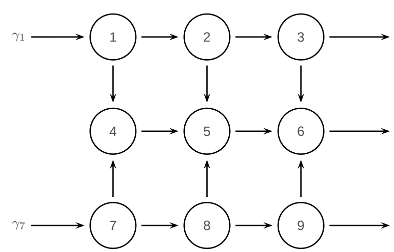
Validation vs. simulation results

station index i:	1	2	3	4	5	6	7	8	9
γ_i	-	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0
μ_i	0.3	0.3	0.3	0.1	0.01	0.014	0.1	0.4	0.5

$\forall i \;\; c_i = K_i = 3, \;\; card(\mathcal{S}_i) = 10$

$$(p_{ij}) = \begin{pmatrix} & & & \bullet & \bullet & & & \bullet & \\ & & & \bullet & & & & \bullet & \\ & & & \bullet & \bullet & & & & \bullet \\ \bullet & \bullet & \bullet & & \bullet & \bullet & \bullet & \bullet & \\ \bullet & \bullet & \bullet & \bullet & & \bullet & & & \\ \bullet & & & \bullet & \bullet & & & & \\ \bullet & & \bullet & \bullet & & & & \bullet & \\ & & & & & & \bullet & & \\ & & & \bullet & & & \bullet & \bullet & \end{pmatrix}$$

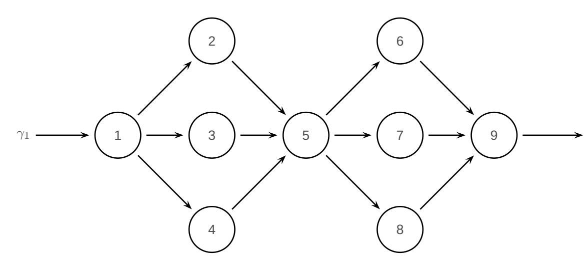
scenario	γ_1
1	0.1
2	0.2
3	0.3
4	0.4



station index i:	1	2	3	4	5	6	7	8	9
γ_i	-	0	0	0	0	0	-	0	0
μ_i	0.3	0.3	0.3	0.6	0.6	0.6	0.3	0.3	0.3

$\forall i \;\; c_i = K_i = 3, \;\; card(\mathcal{S}_i) = 10$

scenario	γ_1	γ_7
1	0.1	0.1
2	0.3	0.3
3	0.5	0.5
4	0.7	0.7
5	0.9	0.9



station index i:	1	2	3	4	5	6	7	8	9
γ_i	-	0	0	0	0	0	0	0	0
μ_i	0.3	0.1	0.1	0.1	0.3	0.1	0.1	0.1	0.3

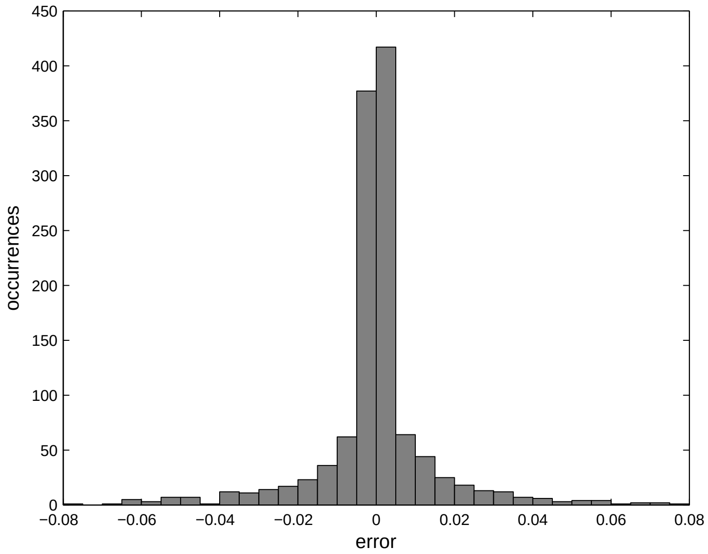
$\forall i \;\; c_i = K_i = 3, \;\; card(\mathcal{S}_i) = 10$

scenario	γ_1
1	0.1
2	0.3
3	0.5
4	0.7
5	0.9

TRANSP-OR

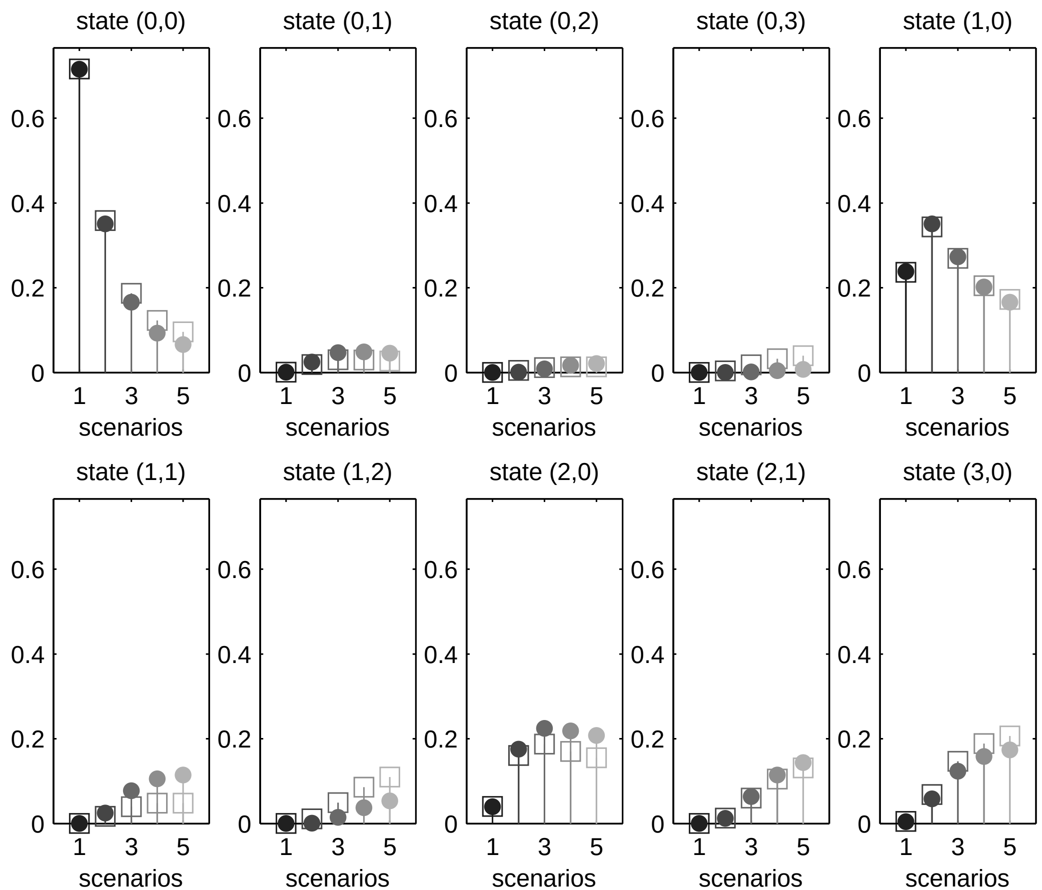
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Validation [3]



Validation [3]

Network C: $\pi(5)$



Case study

Hospital bed blocking: recent demand for modeling and acknowledging this phenomenon:

- patient care and budgetary improvements (Cochran (2006), Koizumi (2005))
- flexibility responsiveness of the emergency and surgical admissions procedure (Mackay (2001)).

The existing analytic hospital network models are limited to:

- feed-forward topologies
- at most 3 units
- Koizumi (2005), Weiss (1987),Hershey (1981).

HUG application

- **Network of interest**: network of operative and post-operative rooms in the HUG, Geneva University Hospital.
- **Dataset**
  - records of arrivals and transfers between hospital units
  - 25336 patient records
  - redundancies in the dataset eliminated
  - used to estimate γ, μ, p_{ij}

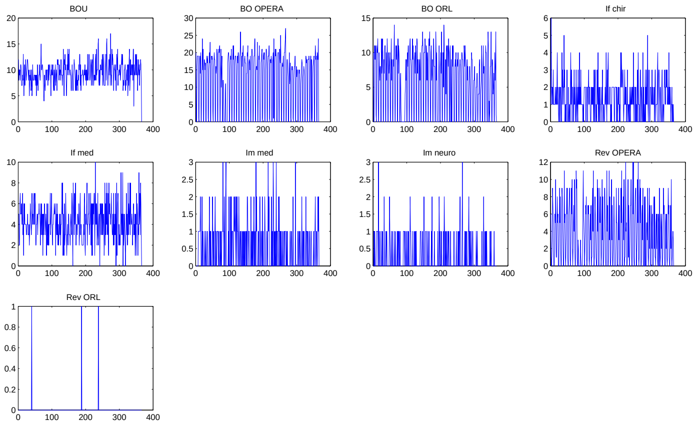
Network model:

Unit	BO U	BO OPERA	BO ORL	IF CHIR	IF MED	IM MED	IM NEURO	REV OPERA	REV ORL
c_i	4	8	5	18	18	4	4	10	6

- beds $\leftrightarrow$ servers
- no waiting space $\leftrightarrow$ bufferless ($K_i = c_i$)

HUG application

γ: avg external arrival rates



- observations:
  Oct 2nd 2004 - Oct 2nd 2005
- estimator: MLE
  (avg nb of occurences)

μ: avg service rate

- estimator: MLE ($\frac{1}{\bar{LOS}}$)
- Assumption: departure time includes no blocking

p_{ij}: transition probabilities:

- frecuency of each transition

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

HUG application

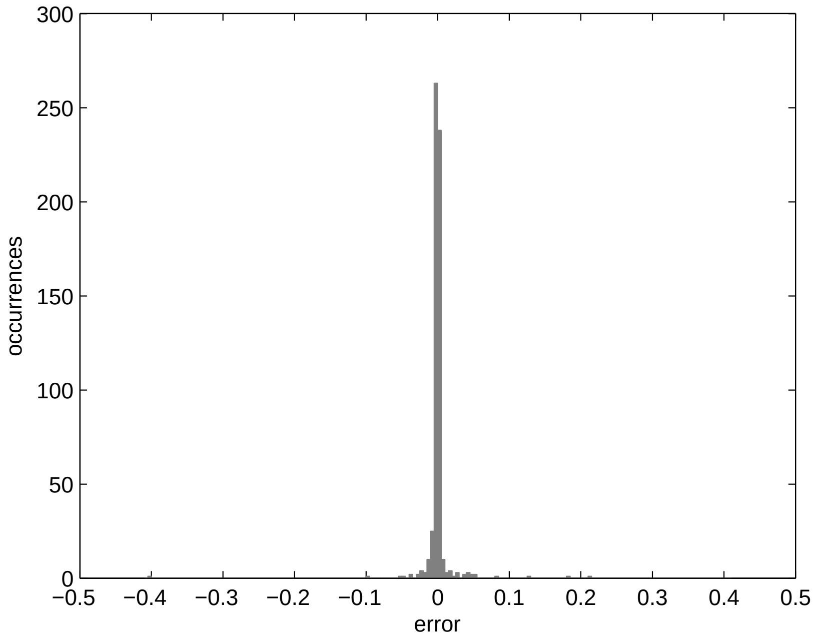
	BOU	BO OPERA	BO ORL	IF CHIR	IF MED	IM MED	IM NEURO	REV OPERA	REV ORL
c_i	4	8	5	18	18	4	4	10	6
γ_i	0.392	0.502	0.246	0.059	0.176	0.025	0.013	0.155	0
μ_i	0.317	0.255	0.335	0.013	0.015	0.014	0.015	0.22	0.518

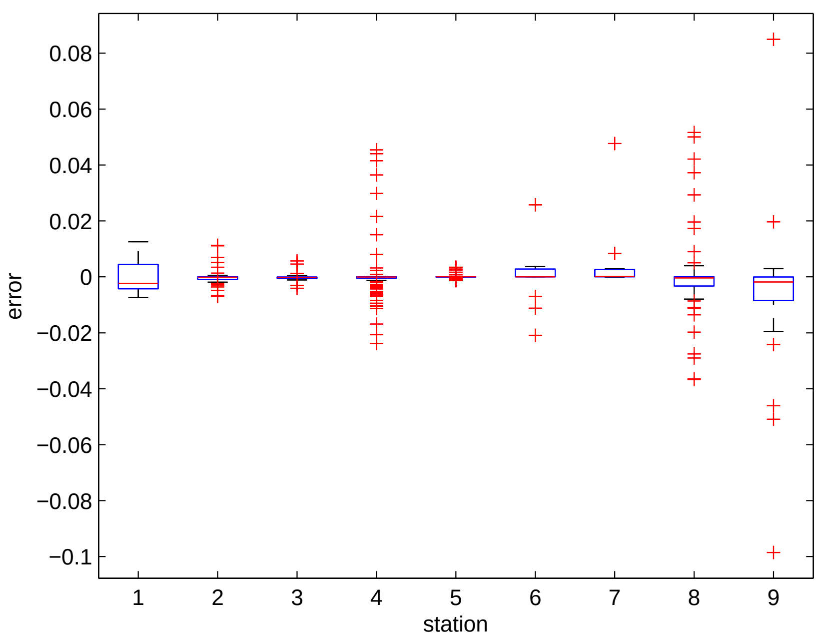
$$(p_{ij}) = \begin{pmatrix} 0 & 0 & 0 & 0.16 & 0.02 & 0 & 0 & 0.71 & 0 \\ 0 & 0 & 0 & 0.07 & 0 & 0 & 0 & 0.84 & 0 \\ 0 & 0 & 0 & 0.03 & 0.01 & 0 & 0 & 0 & 0.95 \\ 0.18 & 0.01 & 0.03 & 0 & 0.03 & 0.01 & 0.11 & 0.03 & 0 \\ 0.05 & 0.01 & 0.01 & 0.01 & 0 & 0.07 & 0 & 0 & 0 \\ 0.02 & 0 & 0 & 0.01 & 0.1 & 0 & 0 & 0 & 0 \\ 0.05 & 0 & 0.05 & 0.04 & 0 & 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0 & 0.05 & 0 & 0 & 0.05 & 0.02 & 0 \end{pmatrix}$$

- Number of unknowns/equations: 635

HUG application

validation of the results





HUG application

Estimation results

	BO U	BO OPERA	BO ORL	IF CHIR	IF MED	IM MED	IM NEURO	REV OPERA	REV ORL
c_i	4	8	5	18	18	4	4	10	6
$P(N_i = K_i)$	0.042	0.001	0.001	0.102	0.046	0.226	0.471	0.006	0.000
$P(N_i = 0)$	0.244	0.136	0.464	0.000	0.000	0.053	0.009	0.017	0.591
P_i^f	0.021	0.012	0.004	0.063	0.020	0.006	0.006	0.005	0.029
$\frac{1}{\hat{\mu}_i}$ LOS	3.2510	3.9499	3.0067	78.0939	66.8836	71.7699	66.8884	4.5497	2.1668
$P(B_i > 0)$	0.0399	0.0142	0.0055	0.1918	0.0400	0.0117	0.0105	0.0038	0.0559

Conclusions and current aims

Conclusions:

- a decomposition method allowing the analysis of FCQN
- explicitly models the blocking phase
- preserves network topology and configuration
- validation versus both pre-existing methods and simulation estimates shows encouraging results
- application on a real case study

Aims:

- come back to general framework:
  integrate with DES.