Analysis of finite capacity queuing networks

Carolina Osorio and Michel Bierlaire

Transport and Mobility Laboratory, EPFL

Zinal, March 2007

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Outline

- Project objectives
- Finite capacity queuing network
  - analysis methods
  - decomposition methods
- Proposed decomposition method
  - description
  - validation

Objectives

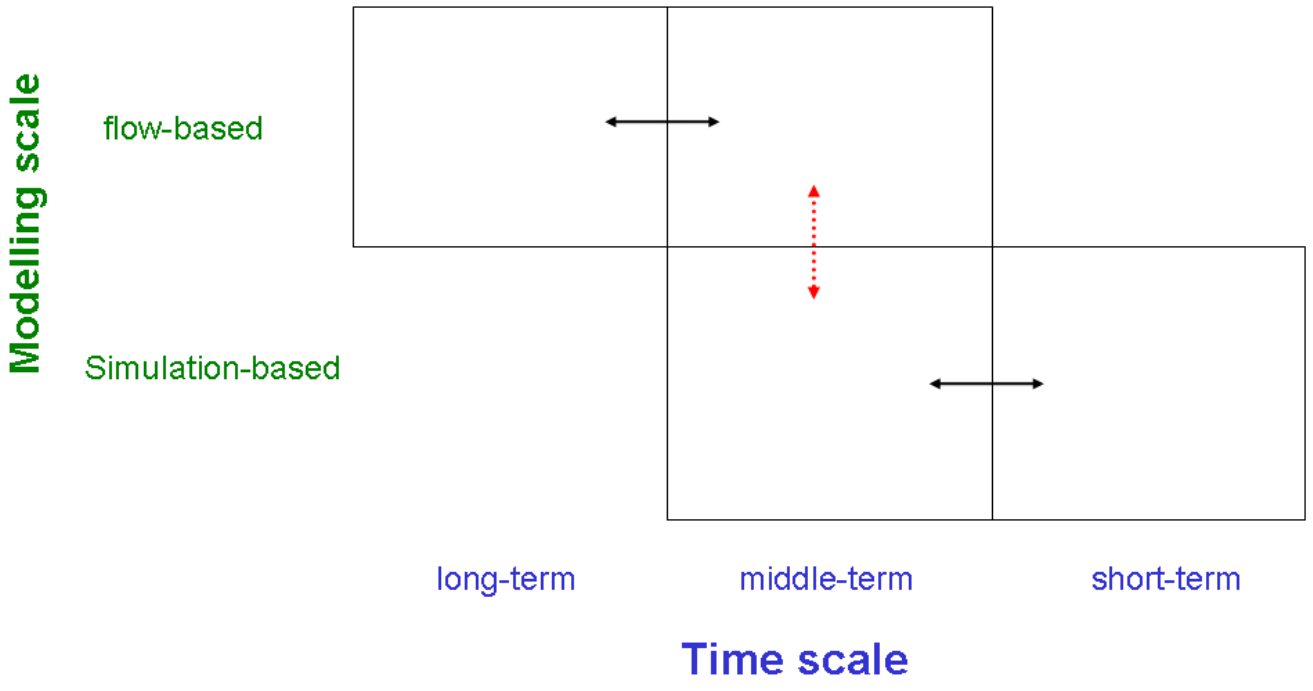
Context: hospital resource management.

2 main research tracks:

- macroscopic models:
  model aggregate flows: e.g. queuing theory.
  simpler to use and easier to integrate in an optimization process
- microscopic models:
  model specific details: simulation-based.
  more realistic model but cumbersome to optimize.

Long-term aim: define an optimization framework allowing the use of both approaches.

Objectives

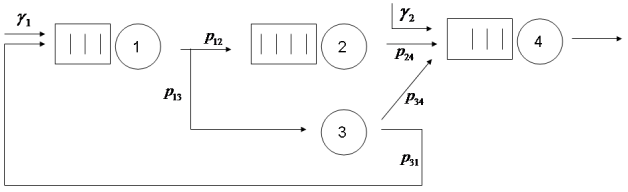


Current phase: define aggregate model using queuing theory

Finite capacity networks

- Jackson networks: infinite buffer size assumption violated in practice.
- Blocking may occur: **complex correlation structure** between the different queues in the network.

How can we model these networks?



Finite capacity queuing networks, FCQN

- General research:
  - Balsamo *et al*. 2001. Analysis of queuing networks with blocking.
  - Perros. 2001. Open queuing networks with blocking, a personal log.
  - Perros. 1984. Queuing networks with blocking: a bibliography.
- Field-specific research:
  - Balsamo *et al*. 2003. A review on queuing networks with finite capacity queues for software architectures performance prediction.
  - Artalejo *et al*. 1999. Accessible bibliography on retrial queues. Mathematical and computer modelling.
  - Papadopoulos *et al*. 1996. Queuing theory in manufacturing systems analysis and design : a classification of models for production and transfer lines.

FCQN methods

Aim: evaluate the main network performance measures using the joint stationary distribution, π.

1. Closed form expression
   exists only for a small set of networks:
   - product-form dbn: (Jackson, BCMP)
   - two-station single server with either tandem or closed topology

For more general topology networks:

2. Exact numerical evaluation: solve $\pi Q = 0$
   Pb: construction of Q for the whole network: limited to small networks.
3. Approximation methods: decomposition methods

Decomposition methods

Aim: reduce dimensionality of the system under study by simplifying the correlation structure between the stations.

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

Analysis of a subnetwork:

i. use a network with a similar behaviour (e.g. Expansion method).

ii. analyse each subnetwork exactly and model their correlation via structural parameters.

Current objective

”Most existing blocking research takes either a tandem configuration with a single or multiple servers or an arbitrarily linked network model with feed-forward flows with a single server.” Koizumi (2005)

”No algorithms have been reported on networks of bufferless multiple server queues with the blocking-after-service rule.” Korporaal et al (2000)

Existing methods mainly concern:

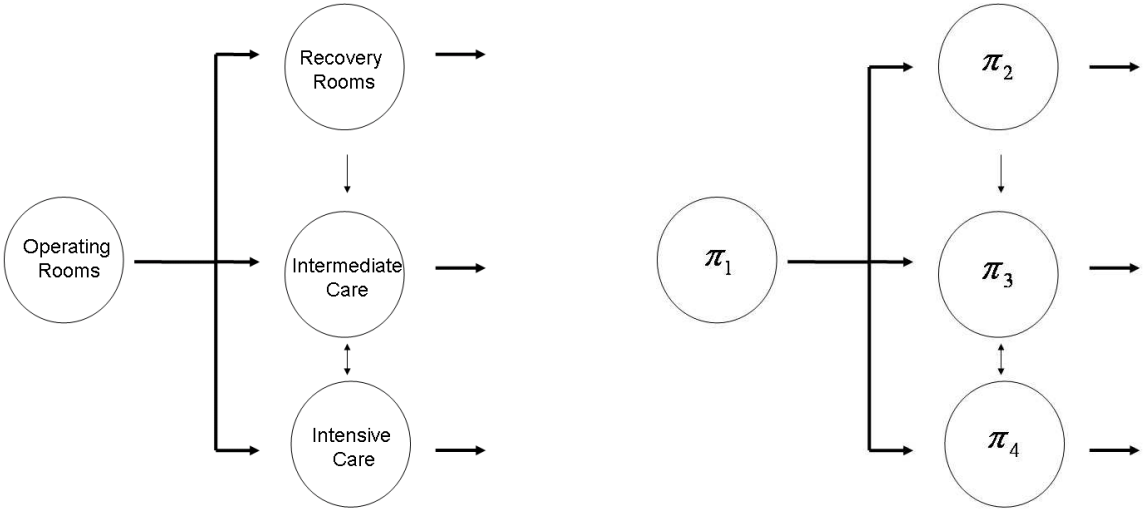
- single server queues in a feed-forward network
- multiple server queues in tandem

Current aim: **generalize to multiple server queuing networks with an arbitrary topology (allowing for feedback)**.

Decomposition method

Subnetwork size: single queues.

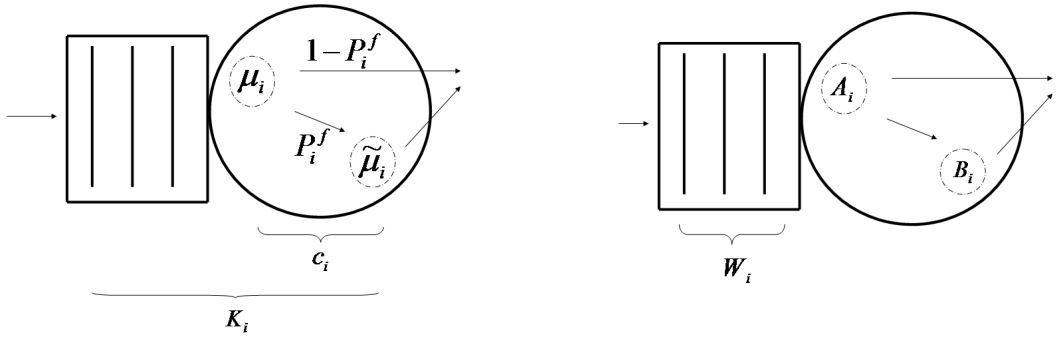
Aim: for each station i estimate the marginal distribution π_i.



This is done by solving the global balance equations. $\begin{cases} \pi_i Q_i = 0 \\ \sum_j \pi_{ij} = 1 \end{cases}$

Process description

Jobs go through an **active** phase, and may eventually go through a **blocked** phase:



Description of the state of a station: $\{(A_i, B_i, W_i) \in \mathbb{N}^3, A_i + B_i \leq c_i, W_i \leq K_i - c_i\}$

- c_i parallel servers
- K_i total capacity: number of servers + buffer size
- λ_i, μ_i: average arrival and service rate
- $\tilde{\mu}_i$ average unblocking rate
- P_i^f average blocking probability

Transition rate estimations

Acknowledge correlation between stations: revise structural parameters.

Main challenge:

effective arrival rates
effective service rates

} stationary dbn of the subnetwork

$\updownarrow$

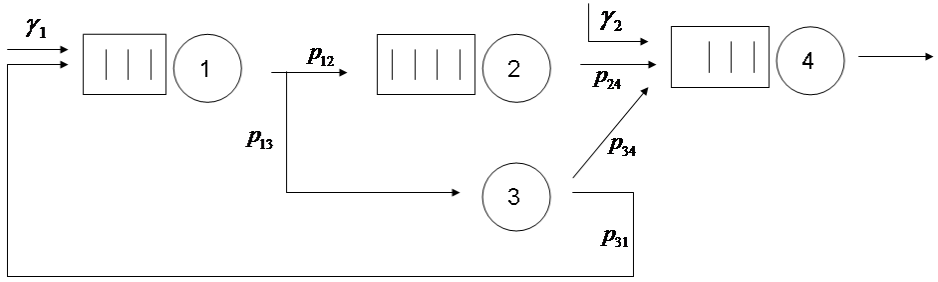
marginal stationary dbn of the network

$Q_i = f(\lambda_i, \mu_i, \tilde{\mu}_i)$

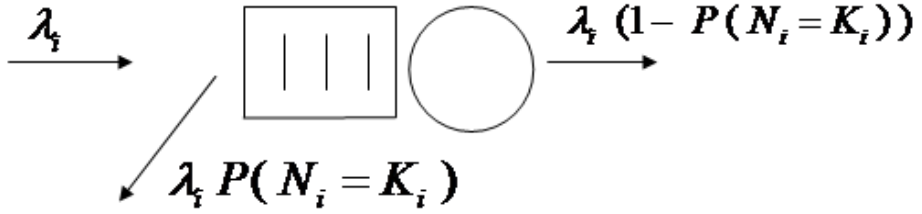
How can we estimate λ_i and $\tilde{\mu}_i$?.

Arrival rates

Flow conservation laws: $\lambda_i^* = \gamma_i + \sum_{j \in i^-} p_{ji} \lambda_j^*$



Each station is modelled as a (two-dimensional) M/M/c/K queue, which are known as *loss models*:



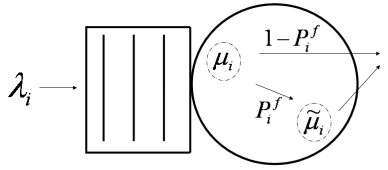
Arrival rates

The effective arrival rates are:

$$\lambda_i = \gamma_i + \sum_{j \in i^-} p_{ji} \lambda_j^* (1 - P(N_j = K_j))$$

Inter-arrival times: $T_i^A \sim \varepsilon(\lambda_i)$, iid
(i.e. Poisson arrival rates)

Service parameters



- active time: $T_i^A \sim \varepsilon(\mu_i)$, iid
- blocked time: $T_i^B \sim \varepsilon(\tilde{\mu}_i)$, iid

The average **effective** service time: $\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + P_i^f \frac{1}{\tilde{\mu}_i}$

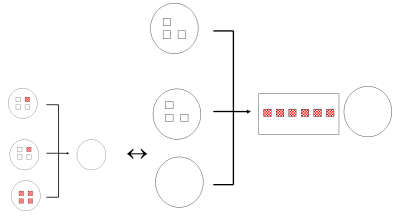
$P_i^f = \sum_{j \in i^+} p_{ij} P(N_j = K_j)$

How can we estimate the average blocked time $\frac{1}{\tilde{\mu}_i}$?

Service parameters

Blocked jobs can be seen as forming a virtual single server queue with a FIFO unblocking mechanism.

Aim: estimate the average waiting time in the virtual queue.



$$\frac{1}{\tilde{\mu}_i} = \sum_{j \in i^+} p_{ij} E[T_{ij}^B]$$

$$E[T_{ij}^B] \approx \frac{1}{r_{ij}\hat{\mu}_j c_j}(E[B_i]+1)\frac{p_{ij}P(N_j=K_j)}{P_i^f}$$

Model properties [1]

In order to acknowledge the finite capacity property of real networks pre-existing models have either:

- revised queue capacities which makes these parameters endogenous
- varied the network topologies

In both cases approximations need to be used to ensure the integrality of the capacity parameters and their positivity is only checked a posteriori. This makes these methods unsuitable for use within an optimization framework.

This method does not relax the finite capacity assumption. It therefore preserves the network topology and its configuration (number of queues and their capacities) as exogenous parameters.

Model properties [2]

- The transition probability matrix remains exogenous: arbitrary topologies are allowed (e.g. allowing for feedback)

- The state of a station explicitly models the blocked jobs, B_i.
  There is a recently recognized need for modelling the bed blocking phenomenon in the hospital context (Cochran 2006).

Summary

$$\mathcal{S}_i = \left\{ \begin{array}{rcl} \pi_i Q_i & = & 0 \\ \sum_j \pi_{ij} & = & 1 \\ Q_i & = & f(\lambda_i, \tilde{\mu}_i, P_i^f, \mu_i) \\ \\ \lambda_i & = & \gamma_i + \sum_{j \in i^-} p_{ji} \lambda_j^* (1 - P(N_j = K_j)) \\ \\ \frac{1}{\tilde{\mu}_i} & = & \sum_{j \in i^+} (E[B_i] + 1) \frac{p_{ij} P(N_j = K_j)}{P_i^f} \frac{\lambda_j}{\lambda_i \hat{\mu}_j c_j} \\ \frac{1}{\tilde{\mu}_i} & = & \frac{1}{\mu_i} + P_i^f \frac{1}{\tilde{\mu}_i} \\ \\ P_i^f & = & \sum_{j \in i^+} p_{ij} P(N_j = K_j) \\ P(N_i = K_i) & = & \sum_{j \in \mathcal{F}(i)} \pi_{ij} \\ \\ E[B_i] & = & \sum_{j \in \mathcal{B}(i)} b_j \pi_{ij} \end{array} \right.$$

- Exogenous : $\{\mu_i, \gamma_i, p_{ij}, c_j, \lambda_i^*\}$
  All other parameters are endogenous.
- Of main interest are the estimates of the:
  - marginal dbns: $\{\pi_i\}$
  - behavioral parameters: $\{\lambda_i, \tilde{\mu}_i, \hat{\mu}_i\}$
- MATLAB **fsolve** routine for systems of nonlinear equations.

Method validation

Validation versus:

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

Validation

Validation versus pre-existing methods

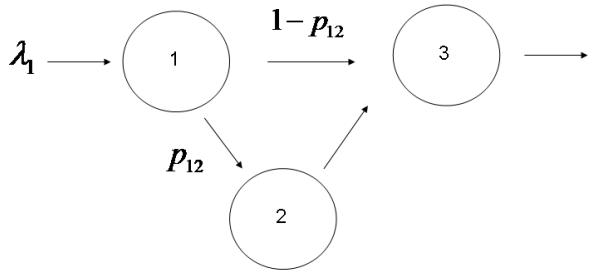
- Kerbache and MacGreggor Smith. 1988. Asymptotic behaviour of the Expansion method for open finite queuing networks. *Computers and Operations Research*
- Altiok and Perros. 1987. Approximate analysis of arbitrary configurations of open queuing networks with blocking. *Annals of Operations Research*
- 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 ($c_j = 1$)

$\lambda_1 = 1, p_{12} = \frac{1}{2}$

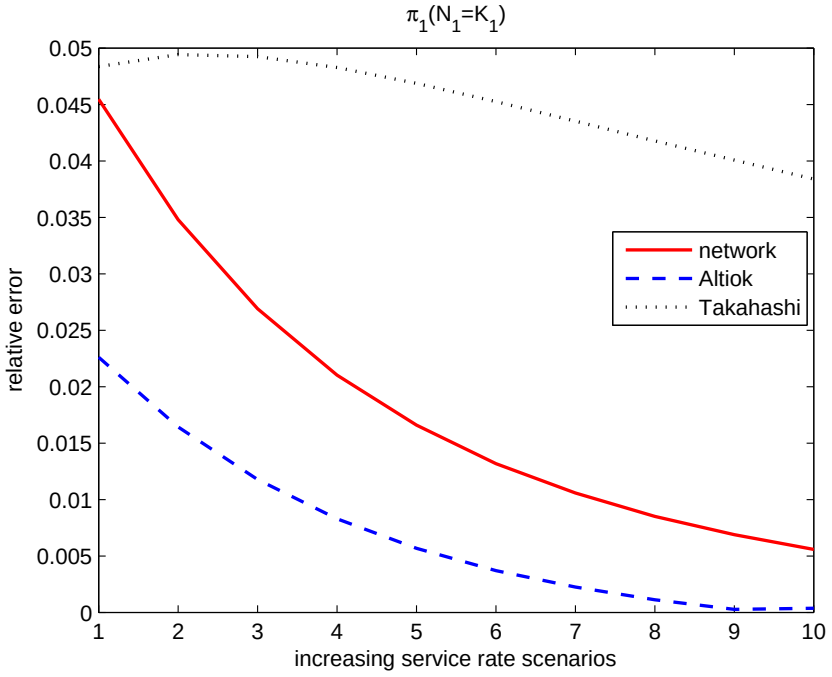
μ_1	μ_2	μ_3
1	1.1	1.2
1	1.2	1.4
1	1.3	1.6
1	1.4	1.8
1	1.5	2
1	1.6	2.2
1	1.7	2.4
1	1.8	2.6
1	1.9	2.8
1	2	3

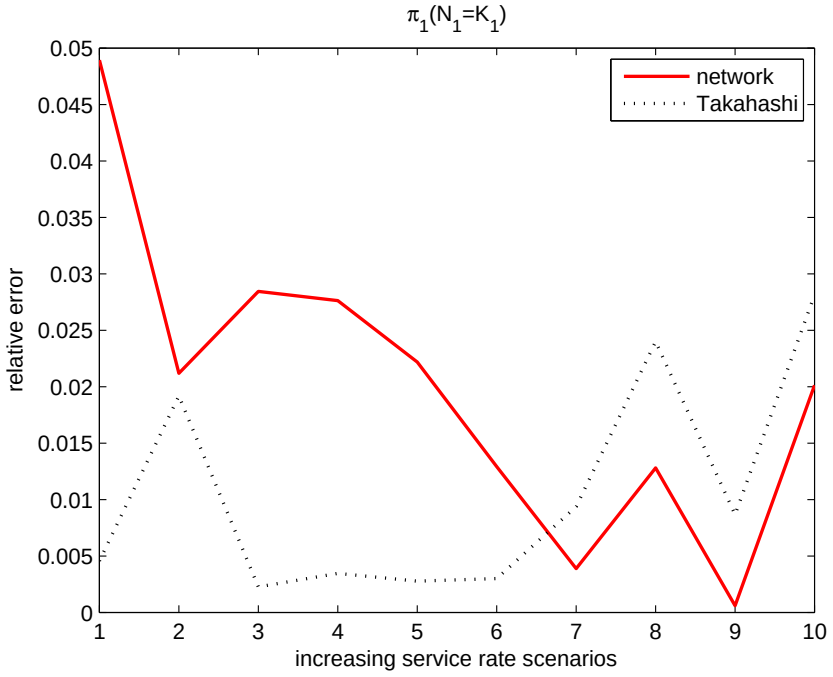


Validation [1]

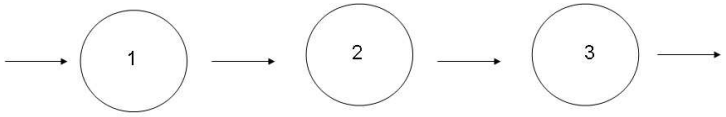
2 sets of scenarios:

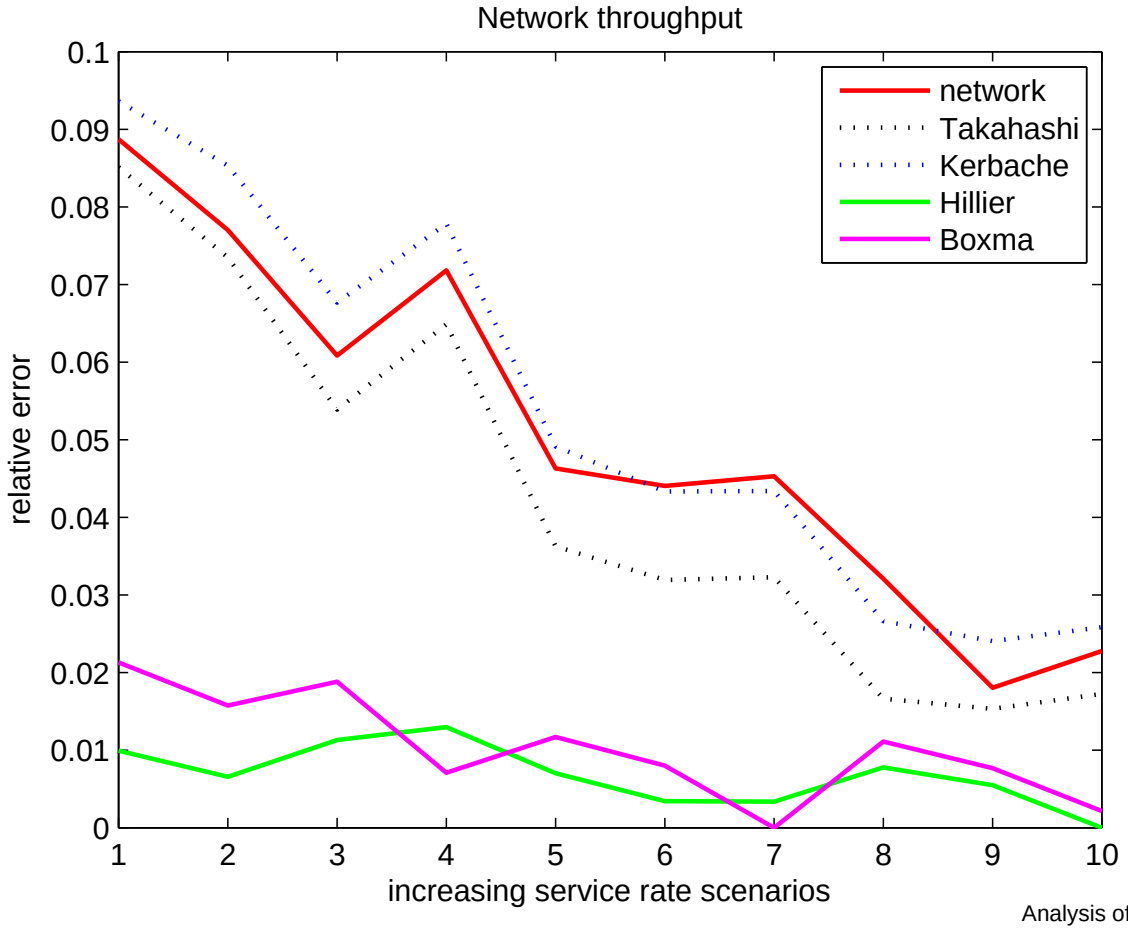
bufferless: $K_j = c_j = 1$, and non-bufferless: $K_j = 3$.





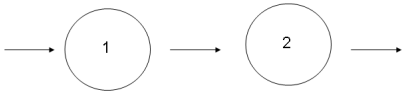
Validation [2]

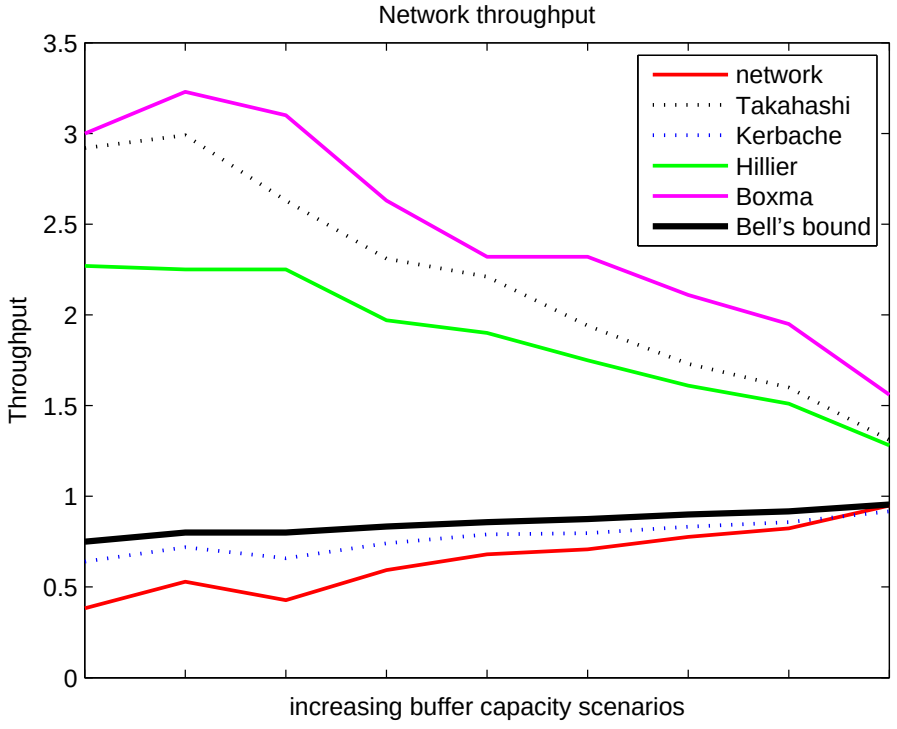




Validation [3]

Theoretical bound on the throughput Bell (1982):

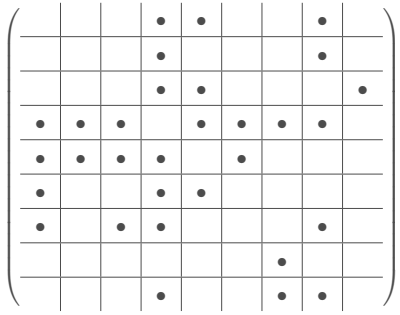




Validation vs. simulation results

- **Network of interest**: network of operative and post-operative rooms in the HUG, Geneva University Hospital.
- **Dataset**: 3475 records of arrivals to units of interest (initially 25246 records). Each record contains the time of arrival and departure, current unit, following unit.

	BO BOU	BO OPERA	BO ORL	IC chir	IC med	IM med	IM neuro	Recovery OPERA	Recovery 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

$(p_{ij}) =$ 

- bufferless stations ($K_i = c_i$)
- Number of unknowns/equations: 635

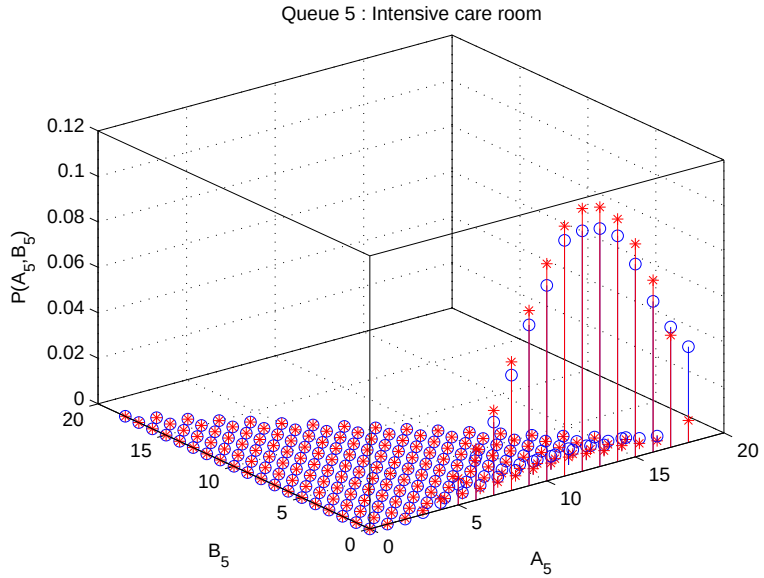
Validation vs. simulation results

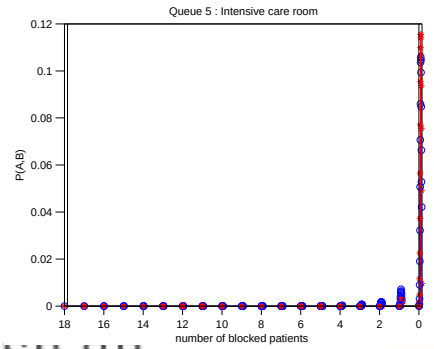
- Commercial simulator: ProModel.

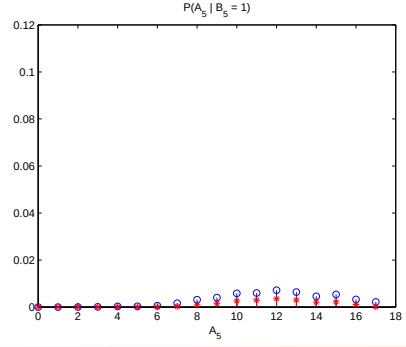
- FCQN model initialization:
  i) simulator estimates
  ii) random initialization
  iii) iterative method

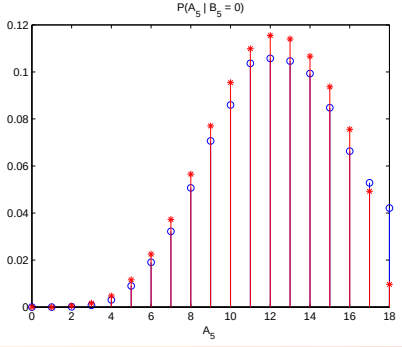
Validation vs. simulation results

Initialization: simulation estimates







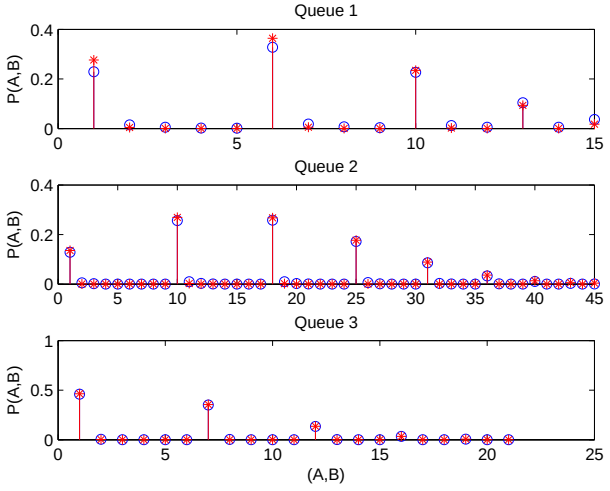


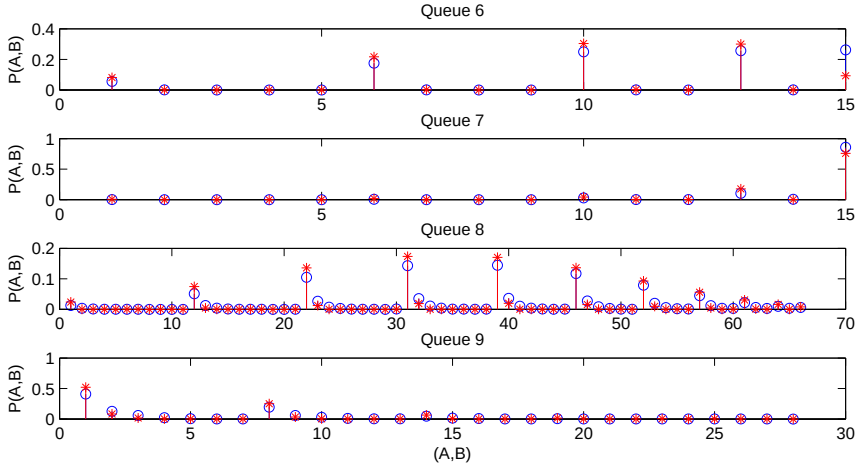
TRANSP-OR

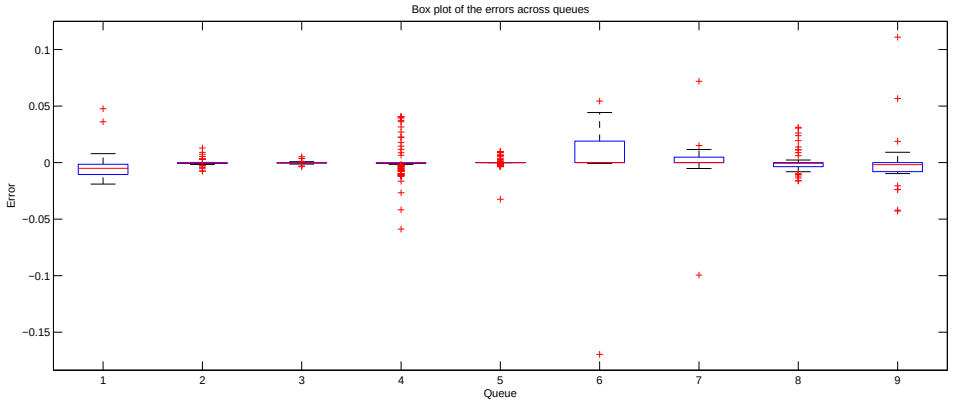
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Validation vs. simulation results

Initialization: simulation estimates





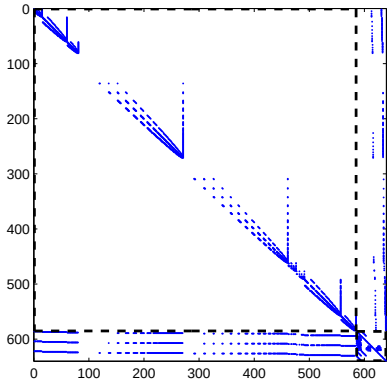


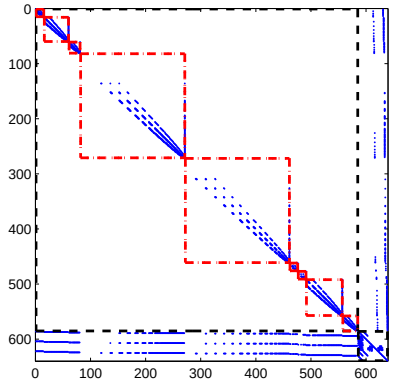
Validation vs. simulation results

ii. **Random initialization**:
We have obtained another numerical solution: analysis is to be carried out.

iii. **Initialization using an iterative method**:
Problem decomposed into 2 subproblems:

1. global balance equations given the behavioral parameters: a set of linear systems that are independent across queues
2. behavioral parameter equations given the distributions: system of nonlinear equations.





TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Conclusions and current aims

Conclusions:

- a decomposition method allowing the analysis of finite capacity queuing networks has been proposed.
- this method explicitly models the blocking phase
- unlike pre-existing methods it preserves the original network topology and configuration (number of stations and their capacity)
- its validation versus both pre-existing methods and simulation estimates shows encouraging results

Aims:

- Carry out further validation versus simulation on various networks.
- general framework: integrate with DES simulator.