An Analytic Finite Capacity Queueing Network Model Capturing Congestion and Spillbacks

Carolina Osorio and Michel Bierlaire

Transport and Mobility Laboratory, EPFL

TRISTAN VI, June 2007

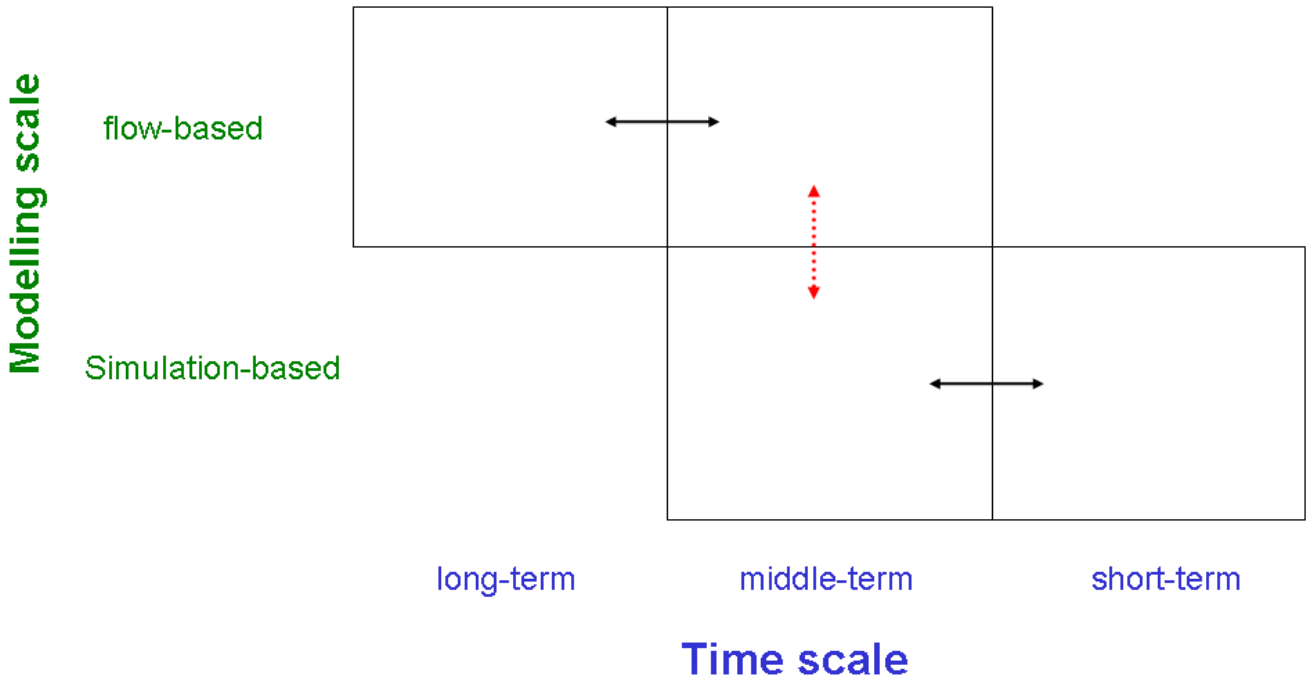
TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Outline

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

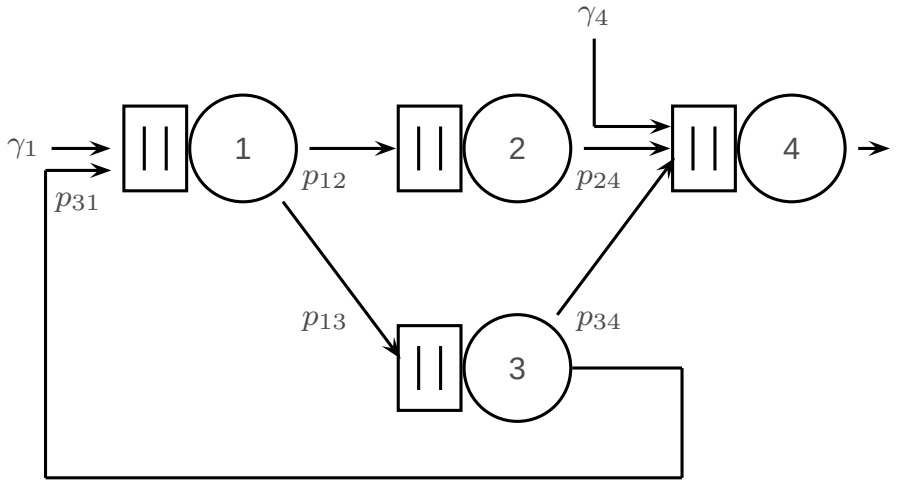
Overall objectives



Current phase: define aggregate analytic model

Finite capacity networks

Aim: evaluate 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:

finite capacity queueing networks

Finite capacity queueing networks FCQN

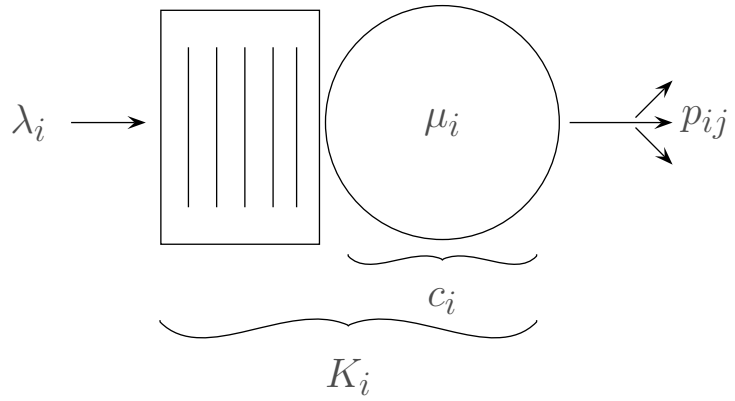
Main application fields:

- 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

Evaluate the main network performance measures using the joint stationary distribution.

State of the network: number of jobs per station.

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

1. Closed form expression
2. Exact numerical evaluation

} small networks (+ specific topologies)

A more flexible approach:

3. **Approximation methods: decomposition methods**

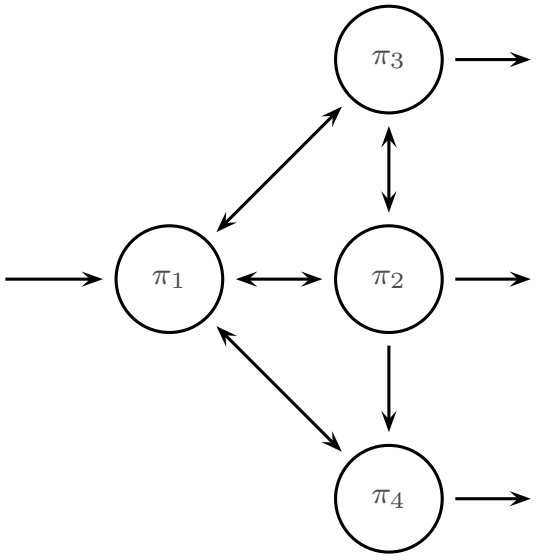
Decomposition methods

By decomposing we can aim at analysing:

- arbitrary topology and size

Method description

1. decompose into subnetworks
2. analyse each subnetwork independently
3. evaluate the main performance measures



Subnetwork analysis

- size: single stations
- method: global balance equations.
- output: estimates of the marginal dbns

Current objective

Existing methods adapted for multiple server + arbitrary topology:

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

Requires:

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

unsuitable for an optimization framework

Current objective

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

Global balance equations

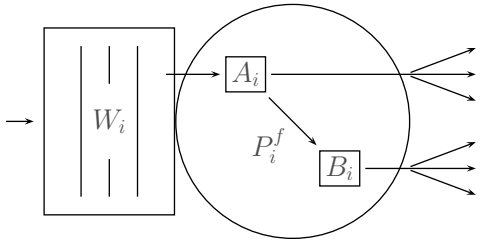
$$\begin{cases} \pi(i)Q(i) = 0 \\ \sum_{s \in \mathcal{S}(i)} \pi(i)_s = 1 \end{cases}$$

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

State space

Upon arrival to a station a job :

1. [queue]
2. is served (active phase)
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 evaluate:

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

Transition rates

$Q(i)$ is a function of:

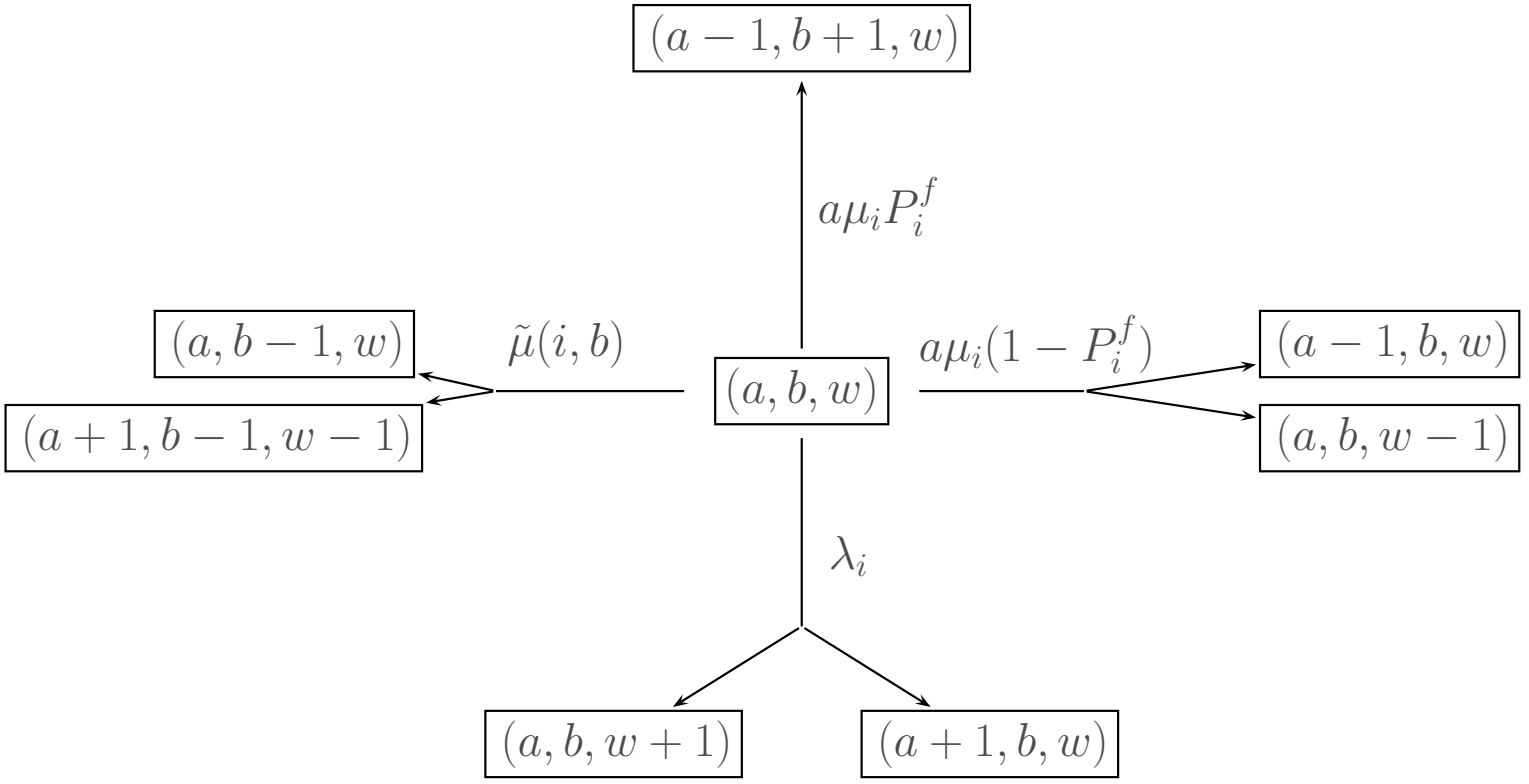
- λ_i, μ_i: average arrival and service rate
- P_i^f: average blocking probability
- $\tilde{\mu}(i, b)$: average unblocking rate given that there are b blocked jobs

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Transition rates

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



Transition rates

$$Q(i) = f(\lambda_i, \mu_i, P_i^f, \tilde{\mu}(i, b))$$

Main challenge and complexity

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

stationary dbn of each station $\leftrightarrow$ marginal dbn of the station

- approximations used to maintain a tractable model
- classical distributional assumptions

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

Summary

$$
\begin{cases}
\pi(i)Q(i) &= 0 \\
\sum\limits_{s \in \mathcal{S}(i)} \pi(i)_s &= 1 \\
\\
Q(i) &= f(\lambda_i, \mu_i, P_i^f, \tilde{\mu}(i,b)) \\
\lambda_i^{\text{eff}} &= \lambda_i(1 - P(N_i = K_i) \\
\lambda_i^{\text{eff}} &= \gamma_i(1 - P(N_i = K_i)) + \sum\limits_j p_{ji}\lambda_j^{\text{eff}} \\
P_i^f &= \sum\limits_j p_{ij} P(N_j = K_j)
\end{cases}
$$

$$
\begin{cases}
\tilde{\mu}(i,b) &= \tilde{\mu}_i^o \; \phi(i,b) \\
\frac{1}{\tilde{\mu}_i^o} &= \sum\limits_{j \in \mathcal{I}+} \frac{\lambda_j^{\text{eff}}}{\lambda_i^{\text{eff}} \hat{\mu}_j c_j} \\
\frac{1}{\hat{\mu}_i} &= \frac{1}{\mu_i} + P_i^f \frac{1}{\mu_i^{\tilde{avg}}} \\
\frac{1}{\tilde{\mu}_i^{avg}} &= \sum\limits_{b \geq 1} \frac{P(B_i = b)}{P(B_i > 0)} \sum\limits_{k=1}^{b} \frac{k}{b} \frac{1}{\tilde{\mu}(i,k)} \\
P(N_i = K_i) &= \sum\limits_{s \in \mathcal{F}(i)} \pi(i)_s \\
P(B_i = b) &= \sum\limits_{s=(.,b,.) \in \mathcal{S}(i)} \pi(i)_s \\
P(B_i > 0) &= 1 - \sum\limits_{s=(.,0,.) \in \mathcal{S}(i)} \pi(i)_s
\end{cases}
$$

- 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
  - triangular topology
  - tandem two-station
- simulation results on a set of small networks

Excellent results

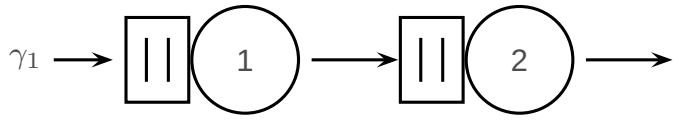
Validation

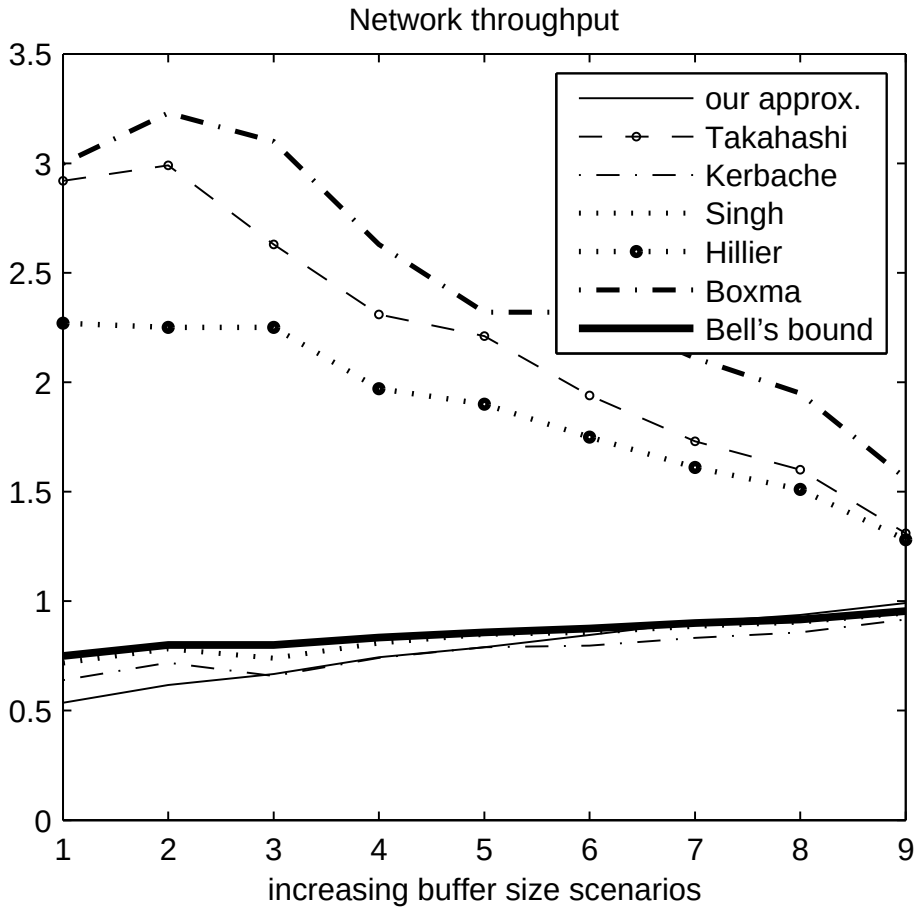
Theoretical bound on the throughput Bell (1982):

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

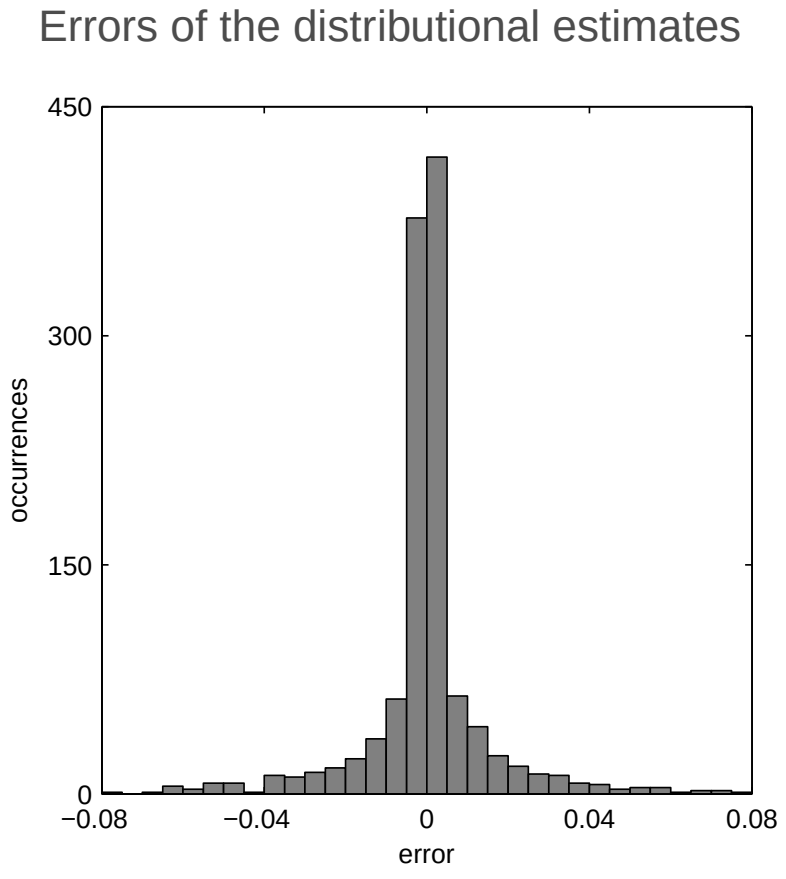
$\gamma_1 = 1, \gamma_2 = 0$

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



Runs: 3 network topologies with 9 stations each under 5 scenarios.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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
    Oct 2nd 2004 - Oct 2nd 2005
  - redunduncies in the dataset eliminated
  - used to estimate γ, μ, p_{ij} (MLE estimators)

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 ($K_i = c_i$)
- **Validation** of the results vs. DES.

HUG application

Transition probabilities conditional on a patient being blocked

unit id	1	2	3	4	5	6	7	8	9
unit	BO U	BO OPERA	BO ORL	IF CHIR	IF MED	IM MED	IM NEURO	REV OPERA	REV ORL
BO U	-	-	-	0.76	0.04	-	-	0.19	-
BO OPERA	-	-	-	0.59	-	-	-	0.41	-
BO ORL	-	-	-	0.87	0.13	-	-	-	0.01
IF CHIR	0.12	-	-	-	0.02	0.04	0.82	-	-
IF MED	0.11	-	-	0.05	-	***0.83***	-	-	-
IM MED	0.13	-	-	0.16	***0.71***	-	-	-	-
IM NEURO	0.34	-	0.01	0.65	-	-	-	0.01	-
REV OPERA	-	-	-	-	-	-	1.00	-	-
REV ORL	-	-	-	0.18	-	-	0.82	-	-

Sources of blocking:

- IF MED $\leftrightarrow$ IM MED
  IF CHIR $\leftrightarrow$ IM NEURO
- operating suites: BO U, BO OPERA, BO ORL $\longrightarrow$ IF CHIR
- REV OPERA, REV ORL $\longrightarrow$ IM NEURO

HUG application

Other performance measures

unit id	1	2	3	4	5	6	7	8	9
unit	BO U	BO OPERA	BO ORL	IF CHIR	IF MED	IM MED	IM NEURO	REV OPERA	REV ORL
K_i	4	8	5	18	18	4	4	10	6
P_i^f	0.02	0.01	0.00	0.06	0.02	0.01	0.01	0.00	**0.03**
$E[B_i]$	0.04	0.01	0.01	0.22	0.04	0.01	0.01	0.00	**0.06**
$E[N_i]$	1.37	2.00	0.77	14.03	12.56	2.46	3.19	4.04	**0.53**
$\frac{1}{\mu_i}$	3.15	3.92	2.99	76.92	66.67	71.43	66.67	4.55	1.93

Blocking may be rare but have a strong impact upon the units:

REV ORL:

- $P_i^f = .03$
- $\frac{E[B_i]}{E[N_i]} = .11$

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.