An analytic finite capacity queueing network capturing congestion and spillbacks

Carolina Osorio and Michel Bierlaire

Analytic queueing networks constitute a flexible tool for the study of network flow. They are simple to manipulate and their analytic aspect renders them suitable for use within an optimization framework. This paper presents an analytic queueing network model which acknowledges the finite capacity of the different queues. By explicitly modelling the blocking phase the model yields a description of the congestion effects and their scope upon the rest of the network. A decomposition method allowing the evaluation of the model is also described.

The framework that we shall present can be applied to study pedestrian flow through circulation systems (e.g. corridors) as in Cheah and Smith (1994). Traffic queues and airport operations are also potential applications. Existing applications in other fields include the study of patient flow throughout a hospital room network (Koizumi et al. (2005) and Cochran and Bharti (2006)), and the study of criminals through a network of prison cells (Korporaal et al. (2000)). The paper is structured as follows. The finite capacity queueing network (FCQN) framework is described, then an overview of the possible analysis methods is given. The proposed model and decomposition method are presented, followed by a description of their validation.

1 General Framework

A queueing network is composed of a set of linked queues, also called stations. Of interest is the study of the flow of "jobs" throughout the network. A job is the generic name for the units of interest, e.g. a vehicle or a pedestrian. A vehicle traffic network can be studied by using this approach: the jobs denote the vehicles and the stations are the locations where between-vehicle interactions are of interest. A station can represent for example a signalized intersection, where vehicles are served as they traverse the intersection. We first describe the general process that a job goes through upon arrival at a station. Station i has c_i parallel servers. The total number of jobs allowed in the station is called the capacity of the station, K_i, the buffer size is $K_i - c_i$. Jobs arriving to station i are either served immediately or queue until a server becomes available. Once a job is served it is routed to its next station, p_{ij} denoting the probability that a job at station i is routed to station j. If this downstream station is full the job will be **blocked** at its current station until the downstream station becomes available. This blocking mechanism, which is at the heart of spillbacks, is known as blocking-after-service (BAS). The jobs are unblocked with a First In First Out (FIFO) mechanism.
The most researched analytic queueing network model is the Jackson network model (Jackson (1957), Jackson (1963)) which assumes infinite capacity for all stations. For real systems this infinite capacity assumption does not hold, but is often maintained due to the difficulty of grasping the between-station correlation structure present in finite capacity networks. This correlation structure helps explaining bottleneck effects and spillbacks, the latter being of special interest in networks containing loops because they are a source of potential gridlocks (Daganzo (1996)). In order to capture this correlation

and to estimate these congestion effects we resort to models with finite capacities. We now describe the existing methods allowing the analysis of FCQN.

2 FCQN Methods

A first survey of FCQN models was made by Perros (1984), who later on also wrote a great historical overview of the research motivations and advances in networks with blocking (Perros (2003)). A detailed introductory book was written by Balsamo et al. (2001). Surveys focusing on specific application fields are given for the production and manufacturing sector (Papadopoulos and Heavey (1996)), for software architecture performance (Balsamo et al. (2003)), and on retrial queues for the telecommunications sector (Artalejo (1999)).
The joint stationary distribution of the network allows us to derive all network performance measures of interest. Exact analysis of FCQN, that is exact evaluation of this joint distribution, is limited to very small networks: single server 2 or 3 station tandem topologies (i.e. stations that are one behind the other) (Grassman and Derkic (2000), Stewart (1999), Akyildiz and von Brand (1994), Latouche and Neuts (1980)), If the network has a more general topology or an arbitrary size then approximation methods are required.

The main motivation of approximation methods is to reduce the dimensionality of the system under study. The main idea is to decompose the network into subnetworks and analyze each subnetwork in isolation. The structural parameters of each subnetwork (e.g. average arrival and service rates) depend on the status of other subnetworks and thus acknowledge the correlation with other subnetworks. The main difficulty lies in obtaining good approximations for these parameters so that the stationary distribution of the subnetwork is a good estimate of the marginal distribution of the network.
Given a subnetwork the distributional estimates can be obtained by either establishing a behavioural analogy with a network whose distribution has a closed (and often product) form, or by exact numerical evaluation of the global balance equations (these equations will be defined in section 3) which now have a smaller dimension but are often non-linear. We have chosen the latter approach because it allows us to preserve the network topology and its configuration.
Existing approximation methods have analysed simple subnetworks consisting of single stations, pairs of stations and triplets. The most commonly used approach is a single station decomposition, which dates back to the work of Hillier and Boling (1967) who considered tandem single server networks. One of the most used approaches concerns single server feed-forward networks where each station is modelled as an M/M/1 station (Takahashi et al. (1980)). An extension of this method to multiple servers (i.e. M/M/c stations) is given by Koizumi et al. (2005). Here the buffers are considered infinite for each isolated station and their average queue length updates the capacity of the predecessor stations. This approximation holds if the capacity of adjacent predecessor stations can accommodate this average queue length; this is checked only a posteriori. A method applicable to networks with an arbitrary topology is given by Korporaal et al. (2000). The individual stations are modelled as M/M/c/K stations for which closed form performance measures are used. As for the method of Koizumi et al. (2005) the capacity of the stations are revised and the validity of these capacity adjustments are verified a posteriori.
The Expansion method, proposed by Kerbache and Smith (Kerbache and Smith (1987), Kerbache and Smith (1988)), was developed for networks of M/M/1/K stations. Here a network reconfiguration expands all finite capacity stations to artificial infinite capacity holding stations, which register the blocked jobs. This method was later extended to multiple servers and applied to pedestrian traffic flows by Cheah and Smith (1994). A similar transformation where all GE/GE/c/K stations are

transformed into GE/GE/c stations, and thus the joint distribution is approximated by a product form joint distribution, is proposed in Tahilramani et al. (1999). Single server networks with phase-type service distributions have been proposed for tandem (Altiok (1982)) and feed-forward topologies (Altiok and Perros (1987)). Jun and Perros (1988) have extended this work to an arbitrary topology and have also considered general service times for an open tandem network in Jun and Perros (1990). The use of a phase-type service distribution accounts for all possible blockings but as stated in Altiok and Perros (1987) it requires the construction of very detailed phase-type service mechanisms, which is rather time consuming, and may become computationally expensive for more complex networks (e.g. multiple server stations). In these methods queue capacity is also augmented in order to allow for storage of all predecessor station capacities. Few authors have considered subnetworks larger than single stations. Two-station decomposition methods have been proposed by van Vuuren et al. (2005), Alfa and Liu (2004), Lee et al. (1998), Perros (1994), Brandwajn and Jow (1988), Brandwajn and Jow (1985). As an extension of the work by Brandwajn and Jow (1988), Schmidt and Jackman (2000) proposed a three-station decomposition method. These methods can provide more accurate results than single station decomposition, but are computationally more intensive as confirmed by Perros (1994).

In order to acknowledge the finite capacity property of these networks the existing methods either revise station capacities or vary the network topologies (e.g. including a holding station such as the Expansion method). The revision of the station capacities renders them dynamic parameters. Thus approximations need to be used to ensure their integrality and their positivity is only checked a posteriori. Nevertheless we believe that a flexible and optimization-friendly model is one that preserves the network topology and its configuration (number of stations and their capacities) as static parameters. We are also interested in explicitly modelling the blocking phase within our analytical approach, yielding performance measures such as the probability distribution of the number of blocked jobs in a station, and that of the blocked time. Since we have not found methods with these characteristics we have developed the method that we shall now describe.

3 Method

We now describe the decomposition method that allows the analysis of multiple server FCQN with an arbitrary topology and blocking-after-service. The method is based on a decomposition of the network into single stations whose structural parameters are approximated so that they can account for the between-station correlation. The general process that a job goes through upon arrival to a station has been described in section 1.

We decompose the network into single stations. Let π_i denote the stationary distribution of the isolated station i. π_i can be obtained via the global balance equations along with the use of a normalizing constraint:

$$\begin{cases} \pi_i Q_i = 0 \\ \sum_{s \in \mathcal{S}_i} \pi_i(s) = 1 \end{cases} \tag{1}$$

where Q_i is the transition rate matrix and $\mathcal{S}_i$ is the state space of station i.

The transition rate matrix Q_i is a function of the following structural parameters:

- the average arrival rate, λ_i.
- the average service rate, μ_i.

- the average unblocking rate, $\tilde{\mu}_i$.
- the average probability of being blocked, $\mathcal{P}_i^f$.

These transition rates are illustrated in figure 1.

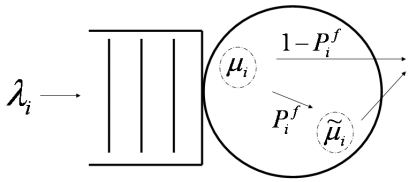

Figure 1: **Transition rates of station** i

The aim and main challenge of decomposition methods is to appropriately approximate these structural parameters so that π_i is a good estimate of the marginal stationary distribution of station i. We now describe our approximation procedure. Hereafter all rates are average rates.

Arrival pattern
In most existing decomposition methods the arrival rate is obtained via the flow conservation equations:

$$\lambda_i^* = \gamma_i + \sum_{j \in i^-} p_{ji} \lambda_j^* \tag{2}$$

where λ_i^* and γ_i denote respectively the total arrival rate and the external arrival rate to station i, and i^- denotes the set of predecessor stations of i.
We model each station as a two-dimensional M/M/c/K station (the distributional assumptions will be detailed further on). For these models, known as loss models, the arrivals that arise while the station is full are considered to be lost. Given an arrival rate λ_i, the portion of flow that is effectively processed corresponds to $\lambda_i(1 - P(N_i = K_i))$; where N_i denotes the total number of jobs at station i, and $P(N_i = K_i)$ is known as the blocking probability. We use this loss model information to approximate the effective arrival rates to station i by:

$$\lambda_i = \gamma_i + \sum_{j \in i^-} p_{ji} \lambda_j^* (1 - P(N_j = K_j)) \tag{3}$$

where λ_i is the effective arrival rate to station i, γ_i the external arrival rate and λ_i^* satisfy the flow conservation equations (2). Inter-arrival times to station i are assumed to be independent and identically distributed exponential variables with parameter λ_i.

Service pattern
The state of a station is described by the number of jobs that are active, those that are blocked and those that are queueing (waiting), i.e. $N_i = (A_i, B_i, W_i)$, such that $A_i + B_i \leq c_i$. Service time and blocked time are each assumed to follow an exponential distribution with parameters μ_i and $\tilde{\mu}_i$ respectively. For a given station all service times are independent and identically distributed, as are all blocked times. By explicitly modeling both of these exponential phases, the number of jobs in front of the servers becomes a two dimensional system (A_i, B_i) composed of the active and the blocked jobs. We are thus in the presence of an M/M/c/K model with a two-dimensional state space.

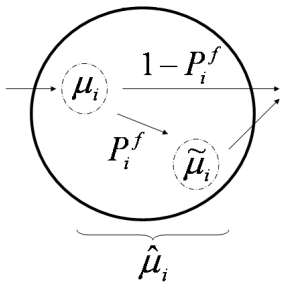

Figure 2: **Service pattern.** For a given station i, all servers serve on average at rate μ_i, jobs are blocked on average with probability $\mathcal{P}_i^f$, and are unblocked with an average rate of $\tilde{\mu}_i$. Accounting for both the service and the possible blocking we obtain the average effective service rate $\hat{\mu}_i$

The total time spent by a job in front of a server, called the effective service time, is composed of the service time and with probability P_i^f of the blocked time. The average effective service rate, noted $\hat{\mu}_i$, is approximated by:

$$\frac{1}{\hat{\mu}_i} = \frac{1}{\mu_i} + P_i^f \frac{1}{\tilde{\mu}_i} \tag{4}$$

where P_i^f, the average probability of being blocked at station i, is given by:

$$P_i^f = \sum_{j \in i^+} p_{ij} P(N_j = K_j) \tag{5}$$

Figure 2 summarizes this service pattern. μ_i is an exogenous parameter, whereas $\tilde{\mu}_i$ needs to be estimated. We now describe how this is done.

The average unblocking rate $\tilde{\mu}_i$
Blocked jobs can be seen as forming a virtual single server queue with a FIFO unblocking mechanism as pictured in figure 3.

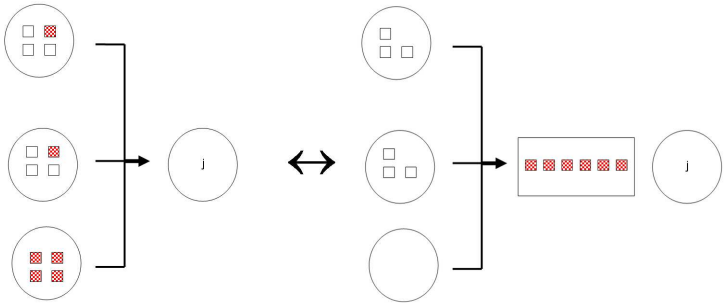

Figure 3: **Unblocking process**. Here we illustrate how the blocked jobs behave as a virtual queue for their next station. This is the main idea underlying the estimation of the unblocking rate. The downstream station j is considered to be full. The squares within each station represent the parallel servers, the filled squares denote blocked servers (i.e. blocked jobs).

Let T_{ij}^b denote the blocked time at station i due to station j. $\tilde{\mu}_i$ is given by:

$$\frac{1}{\tilde{\mu}_i} = \sum_{j \in i^+} p_{ij} E[T_{ij}^b] \tag{6}$$

$$E[T_{ij}^b] = \frac{1}{r_{ij}\hat{\mu}_j c_j}(E[B_i]+1)\frac{p_{ij}P(N_j = K_j)}{\mathcal{P}_i^f} \tag{7}$$

Let us detail the main underlying assumptions in this approximation. If station j is blocking jobs at predecessor stations, it is therefore full and is serving at rate $\hat{\mu}_j c_j$. r_{ij} denotes the proportion of arrivals to station j that arise from station i. The unblocking rate of jobs at i due to j is thus given by $r_{ij}\hat{\mu}_j c_j$. The estimation of the expected blocked time is a function of the expected number of blocked jobs at i due to j, denoted $E[B_{ij}]$. At station i we assume that the proportion of blocking that is due to a downstream station j is given by $\frac{p_{ij}P(N_j=K_j)}{\mathcal{P}_i^f}$. Thus $E[B_{ij}]$ is approximated by $E[B_i]\frac{p_{ij}P(N_j=K_j)}{\mathcal{P}_i^f}$.

In this method p_{ij}, μ_i, γ_i are exogenous parameters. The set of equations 1-7 are solved simultaneously for all stations in order to obtain the distributions $\{\pi_i\}_i$, which allow us to derive the performance measures of interest.

4 Validation

This method has been validated by comparison to both pre-existing methods and to exact models. Validation versus the methods described in Kerbache and Smith (1988), Altiok and Perros (1987), Boxma and Konheim (1981), Takahashi et al. (1980) and Hillier and Boling (1967), has been carried out on tandem 2 station, tandem 3 station and triangular topology networks with varying buffer sizes and service rates, namely under high intensity traffic. Validation has also been carried out by comparing to a theoretical upper bound of the throughput of a tandem 2 station network given by Bell (1982). Comparison with results obtained from exact models has been carried out on both a tandem 3 station and a triangular topology. This has allowed us to validate distributional information concerning blocked jobs, which will be used in the description of congestion effects. In both types of validation the results are very encouraging. These results are available and will be presented. This is an ongoing project and we are currently working on the application of this method on a large scale case study which will also be presented.

5 Conclusion

We have presented a method allowing the analysis of network flows via the use of analytic queueing networks that acknowledge the finite capacity property of the real system. The network is decomposed into single stations which are analysed individually. The between-station correlation is captured via structural parameters. Unlike pre-existing methods the network topology and its configuration are preserved throughout the analysis, this renders the method suitable for use within an optimization framework. The originality of this method also lies in its capacity to explicitly model the blocking phase that jobs may go through under congested traffic conditions.

References

Akyildiz, I. F. and von Brand, H. (1994). Exact solutions to networks of queues with blocking-after-service, *Theoret. Comput. Sci.* **125**(1): 111–130.

Alfa, A. S. and Liu, B. (2004). Performance analysis of a mobile communication network: the tandem case, *Comp. Comm.* **27**(3): 208–221.

Altiok, T. (1982). Approximate analysis of exponential tandem queues with blocking, *Europ. J. Operational Res.* **11**(4): 390–398.

Altiok, T. and Perros, H. G. (1987). Approximate analysis of arbitrary configurations of open queuing networks with blocking, *Ann. Oper. Res.* **9**(1): 481–509.

Artalejo, J. R. (1999). Accesible bibliography on retrial queues, *Math. Comput. Modelling* **30**: 1–6.

Balsamo, S., De Nitto Persone, V. and Inverardi, P. (2003). A review on queueing network models with finite capacity queues for software architectures performance prediction, *Perf. Evaluation* **51**: 269–288.

Balsamo, S., De Nitto Persone, V. and Onvural, R. (2001). *Analysis of Queueuing Networks with Blocking*, Vol. 31 of *International Series in Operations Res. and Management Sci.*, Kluwer Academic Publishers.

Bell, P. C. (1982). Use of decomposition techniques for the analysis of open restricted queuing networks, *Operations Res. Letters* **1**(6): 230–235.

Boxma, O. J. and Konheim, A. J. (1981). Approximate analysis of exponential queueing systems with blocking, *Acta Inform.* **15**(1): 19–66.

Brandwajn, A. and Jow, Y. (1985). Tandem exponential queues with finite buffers, *Comp. Networking and Perf. Evaluation* pp. 245–258.

Brandwajn, A. and Jow, Y. (1988). An approximation method for tandem queues with blocking, *Operations Res. Letters* **36**(1): 73–83.

Cheah, J. Y. and Smith, M. G. J. (1994). Generalized m/g/c/c state dependent queueing models and pedestrian traffic flows, *Queueing Syst.* **15**: 365–386.

Cochran, J. and Bharti, A. (2006). Stochastic bed balancing of an obstetrics hospital, *Health Care Management Sci.* **9**(1): 31–45.

Daganzo, C. F. (1996). The nature of freeway gridlock and how to prevent it., *Proceedings of the 13th International Symposium on Transportation and Traffic Theory* pp. 629–646.

Grassman, W. and Derkic, S. (2000). An analytical solution for a tandem queue with blocking, *Queueing Syst.* **36**: 221–235.

Hillier, F. S. and Boling, R. W. (1967). Finite queues in series with exponential or erlang service times-a numerical approach, *Operations Res.* **15**(2): 286–303.

Jackson, J. R. (1957). Networks of waiting lines, *Operations Res.* **5**: 518–521.

Jackson, J. R. (1963). Jobshop-like queuing systems, *Management Sci.* **10**: 131–142.

Jun, K. P. and Perros, H. G. (1988). Approximate analysis of arbitrary configurations of queueing networks with blocking and deadlock, *Queueing Networks with Blocking: Proceedings of the First international workshop* .

Jun, K. P. and Perros, H. G. (1990). An approximate analysis of open tandem queueing networks with blocking and general service times, *Europ. J. Operational Res.* **46**(1): 123–135.

Kerbache, L. and Smith, M. G. J. (1987). The generalized expansion method for open finite queueing networks, *Europ. J. Operational Res.* **32**(3): 448–461.

Kerbache, L. and Smith, M. G. J. (1988). Asymptotic behaviour of the expansion method for open finite queueing networks, *Comp. and Operations Res.* **15**(2): 157–169.

Koizumi, N., Kuno, E. and Smith, T. E. (2005). Modeling patient flows using a queueing network with blocking, *Health Care Management Sci.* **8**(1): 49–60.

Korporaal, R., Ridder, A., Kloprogge, P. and Dekker, R. (2000). An analytic model for capacity planning of prisons in the netherlands, *J. Oper. Res. Soc.* **51**(11): 1228–1237.

Latouche, G. and Neuts, M. F. (1980). Efficient algorithmic solutions to exponential tandem queues with blocking, *SIAM. J. Alg. Disc. Meth* **1**: 93–106.

Lee, H. S., Bouhchouch, A., Dallery, Y. and Frein, Y. (1998). Performance evalutaion of open queueing networks with arbitrary configuration and finite buffers, *Ann. Oper. Res.* **79**: 181–206.

Papadopoulos, H. T. and Heavey, C. (1996). Queueing theory in manufacturing systems analysis

and design: A classification of models for production and transfer lines, *Europ. J. Operational Res.* **92**(1): 1–27.

Perros, H. (1984). Queueing networks with blocking: A bibliography, *Perf. Evaluation Review, ACM SIGMETRICS* **12**: 8–12.

Perros, H. (1994). *Queueing networks with blocking: Exact and Approximate Solutions*, Oxford Press.

Perros, H. (2003). *Open queueing networks with blocking a personal log. Performance Evaluation: Stories and Perspectives*, Austrian Computer Society.

Schmidt, L. C. and Jackman, J. (2000). Modeling recirculating conveyors with blocking, *Europ. J. Operational Res.* **124**(2): 422–436.

Stewart, W. J. (1999). *Numerical methods for computing stationary distribution of finite irreducible Markov chains*, Advances in Computational Prob., Kluwer Academic Publishers.

Tahilramani, H., Manjunath, D. and Bose, S. K. (1999). Approximate analysis of open network of GE/GE/m/N queues with transfer blocking, *Seventh IEEE International Symposium on Modeling; Analysis, and Simulation of Computer and Telecommunications Systems (MASCOTS'99)* .

Takahashi, Y., Miyahara, H. and Hasegawa, T. (1980). An approximation method for open restricted queuing networks, *Operations Res.* **28**(3): 594–602.

van Vuuren, M., Adan, I. J. B. F. and Resing-Sassen, S. A. E. (2005). Performance analysis of multi-server tandem queues with finite buffers and blocking, *OR Spectrum* **27**: 315–338.