Product Form Queueing Networks

S.Balsamo
Dept. of Math. And Computer Science
University of Udine, Italy

Abstract

Queueing network models have been extensively applied to represent and analyze resource sharing systems such as communication and computer systems and they have proved to be a powerful and versatile tool for system performance evaluation and prediction. Product form queueing networks have a simple closed form expression of the stationary state distribution that allow to define efficient algorithms to evaluate average performance measures. We introduce product form queueing networks and some interesting properties including the arrival theorem, exact aggregation and insensitivity. Various special models of product form queueing networks allow to represent particular system features such as state-dependent routing, negative customers, batch arrivals and departures and finite capacity queues.

1 Introduction and Short History

System performance evaluation is often based on the development and analysis of appropriate models. Queueing network models have been extensively applied to represent and analyze resource sharing systems, such as production, communication and computer systems. They have proved to be a powerful and versatile tool for system performance evaluation and prediction.
A queueing network model is a collection of service centers representing the system resources that provide service to a collection of customers that represent the users. The customers' competition for the resource service corresponds to queueing into the service centers. The analysis of the queueing network models consists of evaluating a set of performance measures, such as resource utilization and throughput and customer response time. The popularity of queueing network models for system performance evaluation is due to a good balance between a relative high accuracy in the performance results and the efficiency in model analysis and evaluation. In this framework the class of product form networks has played a fundamental role. Product form queueing networks have a simple closed form expression of the stationary state distribution that allow to define efficient algorithms to evaluate average performance measures. We introduce product form queueing networks and their properties.

Queueing network models extend the basic queueing systems that are stochastic models first introduced to represent the entire system by one service center. The basic queueing systems have been applied to analyze congestion in telephonic systems and then they have been applied to study congestion in computer and communication systems [Kleinrock 75, Lavenberg 83, Gelenbe-Mitrani 80, Trivedi 82, Lazowska et alt. 84, DeSouza-Muntz 89, Kant 92].
Queueing network models represent such systems as a network of interacting service centers whose analysis often provides quite accurate prediction of their performance. Despite of several assumptions of the class of queueing networks, they have been observed to be very robust models [Suri 83].
Queueing network models can be analyzed by analytical methods or by simulation. Simulation is a general technique of wide application, but its main drawback is the potential high development and computational cost to obtain accurate results. Analytical methods require that the model satisfies a set of assumptions and constraints and are based on a set of mathematical relationships that characterize the system behavior.
We consider analytical methods to analyze queueing network models and specifically product form queueing networks that have a simple closed form of the stationary state

probability distribution, which allow the definition of efficient algorithms to evaluate their performance.
Jackson [Jackson 63] introduced product form queueing network models for open exponential networks and Gordon and Newell [Gordon-Newell 67a] for closed exponential networks. They introduce several assumptions on the model characteristics and provide a simple closed form expression of the stationary state distribution and some average performance indices. This class of models was then extended to include various interesting and useful characteristics to represent more complex system. These features include different types of customers of the networks, various queueing discipline (i.e., the scheduling algorithms of the waiting queues), state-dependent service rate, state-dependent routing between the service centers and some constraints on the population of subnetworks.
The most famous result concerning product form queueing networks was presented by Baskett, Chandy, Muntz and Palacios in 1975 [Baskett et alt. 75] known as BCMP theorem. It defines the well-known class of BCMP queueing networks with product form solution for open, closed or mixed models with multiple classes of customers and various service disciplines and service time distributions. The stationary state distribution is expressed as the product of the distributions of the single queues with appropriate parameters and, for closed networks, with a normalization constant.
An important property of queueing networks with product form is the arrival theorem. It states that the distribution at arrival times at a service center is identical to the distribution at arbitrary times of the same network, for open networks, and of a network with one less customer for closed networks [Lavenberg-Reiser 80, Sevcik-Mitrani 81].
This led to the definition of a set of recurrence equations between average performance measure for closed networks from which it was derived a recursive computational algorithm, the Mean Value Analysis (MVA) [Reiser 81], that avoids the direct evaluation of the normalization constant.

We can analyze product form networks with various computational algorithms to evaluate the performance indices. These algorithms provide a powerful tool in the efficient analysis of large queueing network models. The most important ones are the Convolution Algorithm [Buzen 73] and the Mean Value Analysis [Reiser-Lavenberg 80, Reiser 81] for closed networks. They provide the evaluation of average performance indices with a polynomial space and time computational complexity in the network dimension, that is the number of service centers and the network population.
Product form networks with multiple classes of customers are more difficult to analyze. Various types of customers define the customers' classes in the network that are gathered in chains. Both Convolution and MVA algorithms have been extended to multiple classes networks [Reiser-Lavenberg 80, Reiser 81, Sauer 83, Lam 82], but their cost grows exponentially with the number of customer classes or chains. Other algorithms for multiclass queueing networks have been proposed. The tree Convolution and tree MVA algorithms for multichain networks are based on a tree data structure to optimize the algorithm computation [Lam-Lien 83, Tucci-Sauer 85, Hoyme et alt. 86]. Multichain networks with several types of customers can be analyzed by the algorithms named Recursion by Chain Algorithm (Recal) [Conway-Georganas 86, Conway-Georganas 89], Mean Value Analysis by Chain [Conway et alt. 89] and Distribution Analysis by Chain (DAC) [DeSouza-Lavenberg 89]. Their computational complexity is polynomial with the number of classes of customers, but exponential in the number of service centers.
The computational algorithms have been integrated in various software tools for performance modelling and analysis that include user friendly interfaces based on different languages to take into account the particular field of application, e.g. computer networks, computer systems. This allows not expert users to apply efficient performance modelling techniques. More recently the solution performance algorithms have been integrated with model specification techniques to provide tools for the combined functional and quantitative system analysis.

Product form networks yield various interesting properties. The insensitivity property states that the analytical results, i.e. the stationary state distribution and the average performance indices, depend on the service time requirements only through their average. Similarly, the performance indices depend on the customers routing only through the average visit ratio to each service center [Baskett et alt. 75, Chandy et alt. 77, Chandy-Martin 83, Shassenberger 78, Whittle 85].

Another important property of product form queueing network models is that aggregation methods yield exact results. Chandy, Herzog and Woo [Chandy et alt. 75] first introduced the aggregation theorem. It allows substituting a subnetwork with a single service center, so that the new aggregated network has the same behavior in terms of a set of performance indices. From the performance viewpoint exact aggregation allows us to apply the hierarchical system design process by relating the performance indices of the models at different levels in the hierarchy [Lazowska et alt. 84]. In a bottom-up analysis of systems represented by a succession of queueing network models exact aggregation defines the next model. Similarly, in a hierarchical top-down design of system with given performance requirements, the inverse process of disaggregation or development of the network can be applied to define a more detailed model with the same performance indices [Balsamo-Iazeolla 85].
Aggregation is an efficient technique when applied to the analysis of nearly complete decomposable systems. Informally, such a system can be decomposed into subsystems whose internal interactions are much higher than the interactions among the subsystems [Courtois 77]. Exact aggregation for product form queueing networks provides a basis for approximate solution methods of more general non-product form network models [Marie 79].

More recently further research has devoted to the extension of the class of product form network models and to its characterization. Some interesting new features have been defined such as networks with positive and negative customers proposed by Gelenbe [Gelenbe 91] that can be used to represent special dynamic of actual systems. Some other more complex models include various functions of state-dependent routing and several special cases of queueing networks with finite capacity queues, finite population constraints and blocking [Akyldiz 87, Balsamo-DeNitto 94, Balsamo-Clò 98, Boucherie-VanDijk 91, Gordon-Newell 67b, Lam 77, Towsley 80, VanDijk 93]. Nelson in [Nelson 93] has discussed the mathematics leading to the product form results and the properties of the stochastic process underlying the network model. Product form solution has been extended to queueing networks with batch arrivals and batch services [Henderson-Taylor 90a, Henderson-Taylor 90b] that are also related to discrete time queueing network models.

The goal of this paper is to provide an introduction to product form network models, their properties and applications to system performance evaluation. We will present the basic results, the key ideas and we discuss why this class of models is important in system performance evaluation, whereas we refer to the literature for the mathematical details of the properties.
In the next section we introduce queueing networks to represent and evaluate system performance. Section 3 deals with the key ideas of product form network models and their basic properties. The main algorithms and tools for product form network analysis are introduced in Section 4. Current and future directions of research and application of this class of models are discussed in Section 5.

2 Queueing network models for system performance evaluation

Queueing network models have been extensively applied as performance evaluation models of congestion systems, such as production, communication and computer systems. They

provide a simple model at a high level of abstraction, intuitively understandable and that can clearly represent resource contention. System performance evaluation with queueing network models consists in the definition and parameterization of the model to evaluate of a set of figures of merit that are performance indices, such as resource utilization, system throughput and customers' response time.
First simple queueing systems have been proposed to model a system as a unique service center. These stochastic models were originally proposed for the congestion analysis in telephonic systems and then they have been applied to study congestion in various systems including computer and communication systems [Kleinrock 75, Gelenbe-Mitrani 80, Lavenberg 83, Trivedi 82].

2.1 Queueing systems with a single service center

A single resource model is described by an arrival process of incoming customers, a service process, a buffer space for holding the waiting customers, a scheduling algorithm of the queue, a set of servers that provide the service to customers. Figure 1 illustrates a single service center. The Kendall's notation A/B/c denotes a queueing system with arrival process A, service process B and c service centers, by assuming infinite buffer and First Come First Server scheduling. For example M/M/1 denotes the system with Poisson (Markov) arrival process, exponential (Markov) service process and a single server and M/G/1 the same system except for the service time that has a general or arbitrary distribution. The single resource queueing systems are analyzed by defining an associated discrete-space continuous-time stochastic process, whose state include the system population, denoted by n. Under independent and exponential assumptions the associated Markov process has a simple stationary solution in terms of state probability [Kleinrock 75]. Some queueing systems, such as the M/M/1 and M/M/m systems have an associated Markov process with special structure, that is a birth-death Markov process, which yield a simple closed-form solution of the stationary state probabilities. Hence such queueing systems can be easily analyzed and the average performance indices show simple analytical expressions.

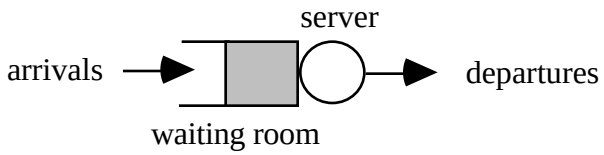

Figure 1 - A single service center queueing system.

For example an M/M/1 system with exponential arrival rate λ and service rate μ has an associated birth and death Markov process whose state is given by k=0,1,… with constant birth rate λ from any state k to state k+1 and constant death rate μ from any state k+1 to state k. If the system is stable, i.e. if the arrival rate is less than the service rate ($\lambda<\mu$), then the queue length distribution is geometric with parameter $\rho = \lambda/\mu$, that is the stationary probability $\pi(k)$ of k customers in the system is

$$\pi(k) = \rho^k (1-\rho) \qquad k\geq 0 \qquad (1)$$

The average queue length is $E[n]=\rho/(1-\rho)$ and the average response time is $R=1/(\mu-\lambda)$. Similarly the M/M/m system with exponential arrival rate λ, m servers with service rate μ has an associated birth and death Markov process whose state is given by k=0,1,… with constant birth rate λ from any state k to state k+1 and state-dependent death rate min{k+1,

m} μ from any state k+1 to state k. The system is stable when $\lambda < m\mu$ and then the stationary queue length probability $\pi(k)$ of k customers in the system is given by

$$\pi(k) = \pi(0)\,(m\,\rho)^k / k! \qquad 1 \le k \le m \qquad (2.1)$$

$$\pi(k) = \pi(0)\,(m^m \rho^k) / m! \qquad k > m \qquad (2.2)$$

where $\rho = \lambda/m\mu$ and $\pi(0)$ is determined by the normalizing condition $\sum_{k=0}^{\infty} \pi(k) = 1$. The average queue length is $E[k] = [m\rho + \pi(m)\rho/(1-\rho)^2]$ and the average response time is $R = [1/\mu + \pi(m)/(m\mu(1-\rho)^2)]$.
Hence we can immediately apply these simple performance model to evaluate several performance indices of a system that can be represented by M/M/1 or M/M/m models. For a detailed discussion of these and other single service center models see [Kleinrock 75].

2.2 Queueing Networks

With more details we can represent a system as a network of resources. A queueing network model is a collection of interconnected single service center queueing systems that provide service to a set of customers. Informally, a queueing network is defined by the service centers, the customers and network topology. Service center characteristics include the service time, the buffer space with its queueing scheduling and the number of servers. Customers are described by their number for closed models and by the arrival process to each service center for open models, the service demand to each service center and the types of customer. Network topology models how the service centers are interconnected and how the customers move between them. Figure 2 illustrates some open and closed networks with various topologies.

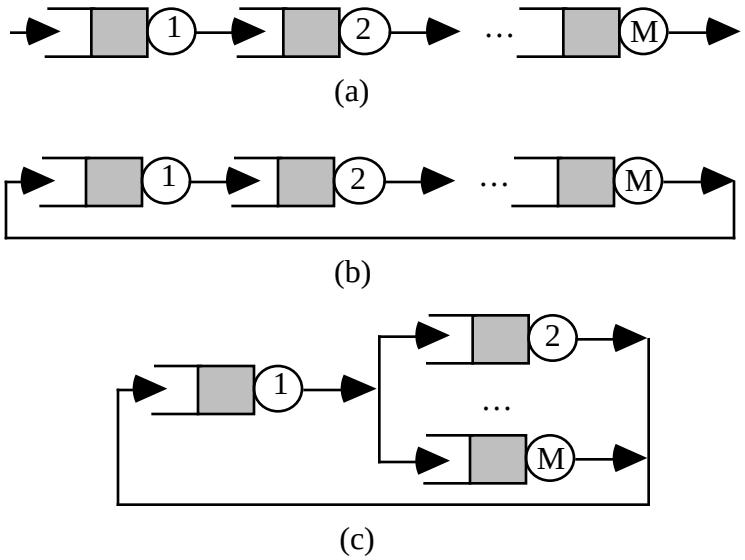

Figure 2 - Example of queueing network topologies: (a) tandem, (b) cyclic, (c) central server.

Different types of customer in the queueing network model can model different behaviors of the customers. This allows representing various types of external arrival process, different service demands and different types of network routing. A chain gathers the customers of

the same type. A chain consists of a set of classes that represent different phases of processing in the system for a given type of customer. Classes are partitioned on the service centers and each customer in a chain moves between the classes. A chain can be used to represent a customer routing behavior dependent on the past history. For example two classes of the same chain in a service centers can represent the customer requirement of two successive services (e.g. a customer representing a job in a computer system that requires two services: program loading and execution). Each chain can be open or closed depending on whether external arrivals and departures are allowed. Multiclass or multichain networks may be open or closed if all the chains are open or closed, respectively. A mixed network has both open and closed chains. A simple example of a multiclass network with two chains and four classes is illustrated in Figure 3. The service time requirement for each class can be different. Chain 1 is open and describes the type 1 customer routing behavior of two successive visits to the same service center first in class a and then in class b. Chain 2 is closed and there is a constant number of type 2 customers circulating between the service centers in class c and d. Multiclass models can be used for a more precise representation of system behavior and to obtain more detailed performance indices.

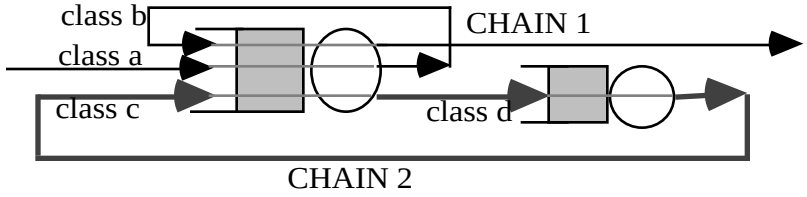

Figure 3 - Example of a mixed network with two service centers, an open chain with two classes (a and b) and a closed chain with two classes (c and d).

The analysis of a queueing network model provides information on the performance of each system component modeled by a service center, and on the overall system performance. Performance indices are obtained by the model analysis.

Queueing network analysis is based on the definition and analysis of an underlying stochastic process that is usually a discrete space continuous time homogeneous Markov process. The process state definition typically includes the number of customers in each queue. The behavior of the queueing network models is represented by the evolution of the associated process. Consider a network with M service centers. Let n_i denote the customer population at node i, $\mathbf{n} = (n_1,\ldots,n_M)$ the network joint queue length and $\pi(\mathbf{n}) = \pi(n_1,\ldots,n_M)$ the stationary joint queue length distribution. Let π denote the stationary state probability vector of the Markov process and Q its transition rate matrix. If the queueing network is stable, then the stationary state probability π is defined by the normalized solution of the following linear system:

$$\pi Q = 0 \qquad (3)$$

also called system of *global balance equations*. Performance indices of the queueing network are derived by the stationary state distribution of the process.
Unfortunately the generality of this approach is limited by its computational complexity. One can easily observe that the process state space cardinality, that is the number of global balance equations, often makes the solution of the system of global balance equation intractable. More precisely, for an open network the process state space is infinite and we can obtain an exact solution only in some special cases, when the matrix Q shows a particular regular structure. For a closed network the process state space grows exponentially with the network parameters that are the number of service centers, customers and customers types. For example, for a single class exponential queueing networks with M service centers and K customers the state space cardinality is $\binom{M+K-1}{K}$.

So why queueing networks became so popular as performance models? The answer is that in some cases, such as for product form networks, we can obtain a simple and efficient solution of the network model analysis.
There is a trade-off between the accuracy and the efficiency of the model analysis. Some interesting approaches provide simple solutions of the model that are useful for system performance evaluation in many practical cases.
Operational analysis of queueing network models was proposed to derive simple results in terms of performance bound and of asymptotic analysis under very general assumptions. The method is appropriate for a first application of performance modelling in the early phases of system design when the system can be not completely specified and we want to compare the potentialities of design alternatives [Lazowska et alt. 84]. However, this approach provides only bounds on asymptotic performance measures.

Product form queueing networks provide more precise and detailed results than operational analysis, in terms of performance indices such as queue length distribution, average response time, resource utilization and throughput. These performance indices are evaluated for each component and for the overall network. Product form network analysis is based on a set of assumptions on the system parameters that lead to a closed form expression of the stationary state distribution. The stationary joint queue length probability π defined by the solution of the associated Markov process, given by the linear system (3) has a *product form solution,* as follows:

$$\pi(\mathbf{n}) = \frac{1}{G} V(n) \prod_{i=1}^{M} g_i(n_i) \qquad (4)$$

where G is a normalizing constant, n is the total network population, function V is defined in terms of network parameters and g_i is a function of state n_i and depends on the type of service center i, $1 \le i \le M$. For open networks G=1, whereas for closed networks V(n)=1. For open network function g_i is the stationary queue length distribution of node i in isolation with appropriate parameters.
Similarly, for networks with multiple types of customers let R denote the number of chains and S the network state that include the customer population at each service center. For a multiclass product form network we can express the stationary state probability π as follows:

$$\pi(S) = \frac{1}{G} \prod_{r=1}^{R} V_r(K_r) \prod_{i=1}^{M} g_i(n_i) \qquad (5)$$

where G is a normalizing constant, K_r is the total network population in chain r, function V_r, $1 \le r \le R$ is defined in terms of network parameters and function g_i depends on the state and the type of service center i, $1 \le i \le M$.

Product form networks can be analyzed by efficient algorithms with a polynomial time computational complexity in the number of network components. This class of models allows a good balance between a relative high accuracy in the performance results and the efficiency in model analysis and evaluation. Moreover product form networks yield several interesting properties such as insensitivity and exact aggregation that greatly influenced the application of this class of models as a powerful tool for performance evaluation.
Various degrees of details can be used to define the performance model at the appropriate level of abstraction, depending on the goal of the system performance analysis. System components are represented by the model components that are the service centers, the classes of customers and the customer routing according to the objective of the performance

evaluation study and depend on the performance measures of interest. This concerns the model definition.

Product form networks are simple and intuitive models that can be solved by efficient algorithms and tools that we shall introduce in Section4. But first, in order to apply this class of models a question is: how can we characterize product form queueing networks?

3. Product form queueing network models: basic ideas

Product form solution of queueing networks holds under special assumptions. The precise characterization of the class of product form network is not easy. The product form solution is related to some properties of the queueing network model that are defined on the Markov process underlying the queueing model. Some sufficient conditions for product form solution based on these properties has been derived. Important properties are *quasireversibility* and partial balance. Informally, quasireversibility of a service center states that the current state, the past departures and the future arrivals are mutually independent. This property refers to the relation between the arrival and departure process. The so called

M M property was first proved by Burke [Burke 56] and states that in an M/M/1 system a Poisson arrival process produces a Poisson departure process, independent of the queue state. Examples of *quasireversible* queues are:

I Multiclass service center with First Come First Served (FCFS) queueing discipline and exponential service time distribution, identical for each customer class.
II Multiclass service center with Processor Sharing (PS) scheduling and arbitrary phase type service time distribution, i.e. formed by a network of exponential stages [Kleinrock 75].
III Multiclass service center with infinite number of servers, that is with IS scheduling and arbitrary phase type service time distribution.
IV Multiclass service center with Last Come First Served with preemption (LCFS-Pr) scheduling and arbitrary phase type service time distribution.

Other examples of quasireversible queues are defined in terms of the particular class of symmetric queueing discipline. A service discipline is called *symmetric* [Kelly 79] if the probability that an arrival enters the queue in the i-th position when there are n customer is equal to the fraction of the service capacity destined to the customer in the i-th queue position when there are n+1 customer in the queue. Examples of symmetric disciplines are LCFS-Pr and PS. However, this is only a sufficient condition for product form solution. A similar condition was defined as station balancing [Chandy-Martin 83] to characterize the queueing disciplines that yield product form queues. In this case, by assuming a special form of the product form expression one can define a necessary and sufficient condition for product form solution that requires station balance discipline for non-exponential queues.

Given a single queue with product form solution a problem is how to combine a set of queues into a network in order to maintain the product form solution.

3.1 Preliminary results

First we can simply connect the queues so that the M M property holds. Tandem exponential networks with Poisson arrivals where the service times are mutually independent satisfy this condition. Similarly, acyclic exponential networks with Poisson external arrivals and routing with Bernoulli splitting can be analyzed as a set of independent M/M/1 queueing system with appropriate arrival rates. This immediately derives from the decomposition and superposition of Markov independent processes. However, when we consider networks with feedback, for example a tandem network with feedback, even if the

external arrival is a Poisson process the total arrival process is not Poisson. Nevertheless under exponential and independence assumptions one can still derive a product form solution for the associated Markov process.

The first important result concerning product form queueing networks was proved by Jackson [Jackson 63] for open exponential networks with FCFS queues and arbitrary Markovian routing. For this network let λ denote the overall arrival rate to the network, p_{0i} the probability that an arrival enters queue i and μ_i the exponential service rate of center i, $1 \le i \le M$. Customers' behavior between service centers of the network is described by routing matrix $P=[p_{ij}]$ where p_{ij} denotes the probability that a customer leaving center i immediately goes to center j, whereas p_{i0} is the probability that it leaves the network, $1 \le i,j \le M$. Hence $\sum_{1 \le j \le M} p_{ij} + p_{i0} = 1$. The routing matrix defines the set of *traffic equations* that determine the visit ratio x_i for each service center i as follows:

$$x_i = \lambda\, p_{0i} + \sum_{1 \le j \le M} x_j\, p_{ji} \,. \qquad (6)$$

The visit ratio x_i is the stationary average arrival rate of customers at center i from outside and inside the network. For stationary and stable open networks x_i is equal to node i throughput. So the traffic equations immediately provide this performance measure for each network component. Jackson proved that for a stable network the stationary joint queue length distribution $\pi(\mathbf{n}) = \pi(n_1, \ldots, n_M)$ is given by formula (4) where G=1, $V(n)=\lambda^n$, $n = \sum_{1 \le i \le M} n_i$ and function $g_i(n_i)$ is the stationary state distribution of center i analyzed as an isolated $M/M/m_i$ queue where m_i is the number of center i servers, with arrival rate x_i and service rate μ_i. In particular let $\rho_i = x_i/m_i\mu_i$. The stationary queue length probability of service center i is given by formulas (2.1) and (2.2). When service center i has a single server ($m_i=1$) the queue length distribution reduces to the solution of the M/M/1 system given by formula (1).

The stability condition requires that each queue is stable, i.e. $\rho_i<1$ for each center i.

Note that a surprising property of Jackson networks is that the service centers *behaves as* independent M/M/m type queueing systems, although in general they *are not* independent.

Closed exponential queueing networks with FCFS discipline and arbitrary Markovian routing has been studied by Gordon and Newell [Gordon-Newell 67a] that proved that product form solution (4) holds where G is a normalizing constant and V(n)=1. Similarly to the Jackson open networks, the routing matrix P defines the customers' behavior. Since in closed networks customers cannot enter or leave the network, P is a stochastic matrix, i.e. $\sum_{1 \le j \le M} p_{ij} = 1$ and $p_{0i}=p_{i0}=0$ for each i. Hence the system of traffic equations (6) has M-1 linear dependent equations and has infinite solutions. In other words the visit ratio x_i is defined up to an arbitrary constant and it represents the *relative* throughput of node i.

Function $g_i(n_i)$ in formula (4) is proportional to the stationary distribution of center i analyzed as an isolated $M/M/m_i$ with arrival rate x_i, service rate μ_i and m_i servers, i.e. it is defined by formulas (2.1)-(2.2) without factor $\pi(0)$ for multiple servers ($m_i>1$) and by formula (1) without factor $(1-\rho)$ for single server ($m_i=1$) where $\rho_i = x_i/m_i\mu_i$. Like Jackson networks in such a closed network the service centers *behaves as* independent M/M/m type queueing systems, although this is not the case.

Note that in closed networks the queue length distribution of any center i does not correspond to function g_i as for open networks, but is derived from the joint queue length distribution $\pi(\mathbf{n})$ given by formula (4). This requires the computation of functions g_j for each node j and of the normalization constant G that guarantees that $\sum_{\mathbf{n}} \pi(\mathbf{n}) = 1$. Hence the

computation of the performance indices in closed queueing networks is a non trivial problem. In Section 4 we shall deal with this problem.

3.2 Main result

The quasireversibility of the queues in network models was discussed and studied by Kelly [Kelly 79]. A sufficient condition for product form solution network is that it consists of quasireversible queues interconnected by a Markovian routing. Such a routing is defined when the customer routing decision only depends on the state of the current customer's class and service center and is independent of the state of the rest of the network. The characterization of product form solution related to quasireversibility of queueing network is discussed in [Kelly 79, Warland 88, Nelson 93].
If this sufficient condition for product form holds then the stationary state distribution π has the closed form expression given by formulas (4) and (5) for single class and multiclass queueing networks, respectively.

The well-known BCMP theorem proved by Baskett, Chandy, Muntz and Palacios in [Baskett et alt. 75] defines the so-called BCMP queueing networks with product form solution for open, closed or mixed models with multiple classes of customers, various service disciplines and service time distributions. In particular they defined the four types I-II-III-IV of service centers introduced above that are quasireversible queues. The stationary state distribution is expressed as the product of the distributions of the single queues with appropriate parameters and, for closed networks, with the normalization constant.
External arrivals in BCMP networks are Poisson process and the average arrival rate may depend on the total network population or on the population of a chain. Let $\lambda(n)$ and $\lambda_r(K_r)$ denote the overall arrival rate to the network dependent on the total number of customers in the network (n) and in chain r (K_r), respectively. The routing for each chain r is described by a routing matrix $P^{(r)}=[p^{(r)}_{ic;jd}]$ where $p_{ic;jd}$ denotes the probability that a customer leaving center i from class c immediately goes to center j in class d, whereas $p_{ic;0}$ is the probability that it leaves the network. Then the traffic equations (6) are defined for each chain r and provide the visit ratio $x^{(r)}_{ic}$ for each class c in service center i. Let C_{ir} the set of the classes in service center i that belong to chain r. Hence $x_{ir}=\sum_{c\ C_{ir}} x^{(r)}_{ic}$ is the visit ratio for node i and chain r. Let μ_{ir} denote the service rate of service center i and chain r and let $\rho_{ir}=x_{ir}/\mu_{ir}$. Then the BCMP product form solution is given by formulas (4) and (5) with $V(n)=\prod_{k=0}^{n-1}\lambda(k)$, $V_r(K_r)=\prod_{r=1}^{R}\prod_{k=0}^{K^{(r)}-1}\lambda_r(k)$, $g_i(n_i)=\prod_{r=1}^{R}\frac{\rho_{ir}^{n_{ir}}}{n_{ir}!}$ for type I-II-IV queues and $g_i(n_i)=n_i!\prod_{r=1}^{R}\frac{\rho_{ir}^{n_{ir}}}{n_{ir}!}$ for type III queue, G=1 for open networks and the normalizing constant for mixed and closed networks.
The service rate of each node i can be dependent on the service center load, i.e. the number of customers in node i and chain r. For type I node (with exponential service and FCFS scheduling) it can depend only on node i population.
Note that as observed for FCFS-exponential networks, in a BCMP network the service centers *behaves as* a set of independent queueing systems (M/M/m for type I queue and M/G/m with PS discipline for type II queue, IS for type III queue and LCFS-Pr for type IV queue) although this is *not true.*

Quasireversible queues and Markovian routing provide a sufficient condition for product form solution of queuing networks. As discussed above, queues with symmetric queueing discipline are quasireversible [Kelly 69, Warland 88]. Another similar sufficient condition

for product form solution is station balance, a property of queueing discipline similar to symmetric queues, that characterizes the scheduling for non-exponential queues that yield product form solution by assuming a special closed form solution [Chandy-Martin 83].
Partial balance is a necessary condition for quasireversibility. It is defined on the Markov process associated to the queueing network and states that the probability flux, i.e. the time average transition rate, out of a state **S** due to arrivals of type r customers is equal to the probability flux in state **S** due to departures of type r customers. An extensive discussion of partial balance, quasireversibility, product form and other properties was presented in [Nelson 93].
The partial balance condition as a characterization of product form networks is given on the underlying process and it allows identifying more general cases of product form networks. However, it cannot be always easily translated in terms of a simple characterization of queueing network components, i.e. types of service centers, queueing scheduling, number of servers and routing.

3.3 Extensions

Various extensions of the class of BCMP product form networks have been derived.
They include state dependent routing [Boucherie-VanDijk 91, Lam 77, Towsley 80], i.e. the definition of routing probabilities are special functions that may depend on the state of the entire network or of subnetworks and/or single service centers. This allows representing systems with more complex features such as dynamic load balancing algorithms or adaptive routing strategies.
Such models usually assume some additional constraints on the network parameters and a special structure of the routing state dependent functions. For example Towsley [Towsley 80] considered closed queueing networks where the routing for some service centers may be a rational function of the queue length of the service centers belonging to a downstream subnetwork with a particular topology, called parallel subnetwork.

Example 3.1. A simple example is the central server network illustrated in Fig. 2c with BCMP type service centers and where the routing probability from the central node 1 to the other nodes may depend on the state of the downstream node and the state of subnetwork {2,…,M}, i.e. the routing probability from service center 1 to i, $2 \le i \le M$, can be defined as the following state dependent function: $p_{1i}(\mathbf{n})=h_i(n_i)h(n_2+\ldots+n_M)$ where h_i and h are arbitrary nonnegative functions. The network has a product form solution (4) where $V(n)=\prod_{k=0}^{n-1} h(k)$ and g_i is defined as for BCMP networks time a factor $\prod_{k=0}^{n_i-1} h_i(k)$ for each node i, $2 \le i \le M$.

Boucherie and VanDijk have proposed an extension to more complex state dependent routing by considering a more detailed definition of routing functions dependent on the state of subnetworks called clusters and the state of service centers [Boucherie-VanDijk 91]. The model assumes that the service centers are partitioned into a set of subnetworks that are linked by a state dependent routing. Then the routing function between two service centers i and j that respectively belong to two disjoint subnetworks I and J has the following expression: $p^{(I)}_{i0}\ p'_{IJ}\ p^{(J)}_{0j}$, where $p^{(I)}_{i0}$ and $p^{(J)}_{0j}$ are routing functions internal to subnetworks I and J, respectively, and p'_{IJ} denotes the routing between subnetworks. This model can be useful to represent hierarchical and decomposable systems.

Queueing networks with finite capacity queues, subnetwork population constraints and blocking have product form solution in some special cases [Akyldiz 87, Balsamo-DeNitto 94, Balsamo-Clò 98, Gordon-Newell 67b]. Various blocking types that describe different behaviors of customer arrivals at full capacity service centers and the servers' activity in the network have been defined. For several special combinations of network topology, types of

service centers and blocking mechanisms one can derive a product form solution for the stationary state distribution. Moreover, one can derive various equivalence properties between product form networks with and without blocking and between networks with different blocking type, as discussed in [Balsamo-De Nitto 94].

Example 3.2. For example consider an exponential cyclic network illustrated in Fig. 2b where each queue i has a finite capacity B_i. When a queue becomes full the upstream service center is blocked until there is a free buffer position in the destination node. This is called the Blocking Before Service (BBS). If the number of customers in the network K is such that any node can never be empty, i.e. $K > \sum_{1\le i\le M} B_i - \min_{1\le i\le M} B_i$ [Gordon-Newell 67b] then product form solution given by formula (4) holds with V(n)=1, G the normalizing constant, $g_1(n_1)=(\mu_M)^{n1}$, $g_i(n_i)=(\mu_{i-1})^{ni}$, $2\le i\le M$.

Example 3.3. Consider a network with BCMP type service centers and finite capacity queues. When a job attempts to enter a destination node with full capacity, it goes back to the sending node where it receives a new service according to the service discipline. This is called Repetitive Service Blocking (RS). If the network has reversible routing, i.e. if matrix P is such that $x_ip_{ij}=x_jp_{ji}$ and $\lambda p_{0i}=x_ip_{i0}$ for each i and j, then the network has the same product form solution (4) as the BCMP network with infinite capacity queues, but normalized on the restricted state space. The central server network shown in Fig. 2c is an example of reversible routing network.

Another extension of queueing networks with product form is the class of networks proposed by Gelenbe [Gelenbe 91] with positive and negative customers that can be used to represent special system behaviors. For example negative customers may represent commands to delete some transactions in databases or in a distributed computer system due to inconsistency or data locking. A negative customer arriving to a service center reduces the total queue length by one if the queue length is positive and it has no effect otherwise. Negative customers do not receive service. A customer moving between service centers can become either negative or remain positive. Such a queueing network has product form solution under exponential and independence assumptions and with a Markovian routing and the solution is based on a set of non linear traffic equations of the customers.

Extension of BCMP networks to different service discipline has been derived. Le Boudec proved product form solution for queueing networks with multiserver nodes with concurrent class of customers that allow to represent special systems [Le Boudec 86].

Product form solution has been extended to queueing networks with batch arrivals and batch services [Henderson-Taylor 90a, Henderson-Taylor 90b] that are also related to discrete time queueing network models. The model evolution is described by a discrete time Markov chain and assumes special expressions for the probability of batch arrivals and departures and correlated batch routing. The product form solution is based on a generalized expression of the traffic equations and the quasireversibility property of the network. The product form solution holds for continuous time and discrete time queueing networks.

3.4 Properties

Product form networks yield various properties.
Insensitivity is an interesting property that states that some performance indices are insensitive to certain network parameters [Baskett et alt. 75, Chandy et alt. 77, Chandy-Martin 83, Shassenberger 78, Whittle 85]. The stationary queue length distribution and the average performance indices (throughput, resource utilization, average waiting time and response time) depend on the service time distributions only through the average. Hence in

BCMP networks different service time distributions with the same mean value for a given node of type II, III or IV do not affect the queue length distribution and average performance indices. Insensitivity is related to the station balance property as discussed in [Chandy et alt. 77].
A practical consequences of insensitivity is that when a system is represented by a product form network one has to estimate only the first moment of the service time distribution for each resource to define the model parameter.
Insensitivity of product form networks holds also for the customers routing. Indeed, the product form solution definition depends on the customers routing only through the average visit ratio to each service center. They are obtained by the linear system of traffic equations (6). Hence product form networks with different routing matrix P but with the same visit ratios x_i's, for each service center i, provide the same queue length distribution and average performance indices. Moreover in multiclass networks if we want to evaluate performance indices for each chain and not for each class it is sufficient to estimate the visit ratios at each service center for each customer chain. In the example of Figure 3, one can simply estimate the visit ratios for chain 1 and 2 to each service center to derive the queue length distribution for each queue and chain.
As a consequence in order to define the network parameters for a system represented by a product form network it is not necessary to describe the routing matrix but it is sufficient to estimate the visit ratios at each service center for each customer chain.

Another property of product form queueing network models is *exact aggregation*. The aggregation theorem or Norton's theorem for queueing networks proved by Chandy, Herzog and Woo [Chandy et alt. 75] allows substituting a subnetwork with a single service center, so that the new aggregated network has the same behavior in terms of a set of performance indices. The aggregated or flow-equivalent service center is usually defined as a FCFS service center with exponential service time and load dependent service rate. This service rate when there are n customers represents the throughput of the subnetwork analyzed in isolation with n customers circulating. A simple example of aggregation is shown in Figure 4 where subnetwork {2,…,M} is aggregated into the flow equivalent node C. The service rate $\mu_C(n)$ is set equal to the throughput X(n) of the subnetwork analyzed in isolation when there are n customers, for each n. The aggregated network in Figure 4c is obtained by substituting the subnetwork with the aggregated node C. The aggregated network and the original one have the same marginal queue length distribution and average performance indices.

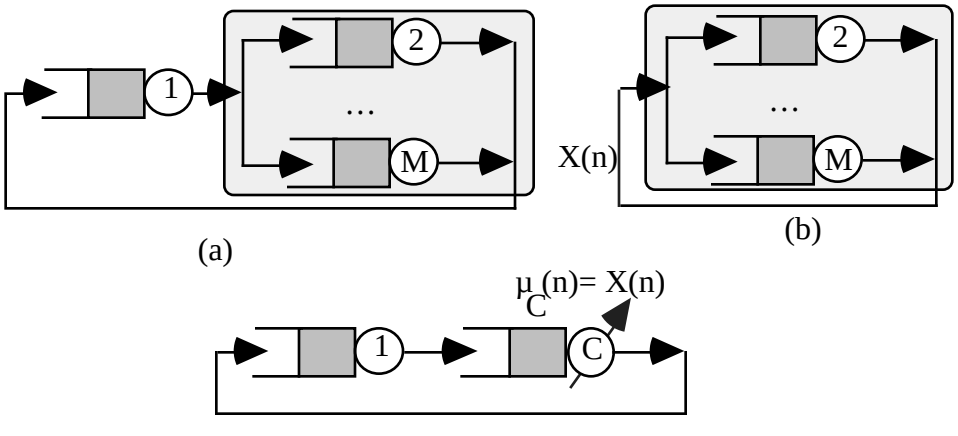

Figure 4 - Aggregation theorem: (a) original network, (b) isolated subnetwork, (c) aggregated network.

Exact aggregation in queueing networks holds for any subnetwork, i.e. for subnetworks with multiple entry and exit points and for which we have also to define a new routing matrix for the aggregated network [Balsamo-Iazeolla 82], and for multichain networks [Kritziger et alt. 82].
Exact aggregation can be used in hierarchical system analysis. In a bottom-up system design process we can relate the performance indices of the network models at different levels in the hierarchy [Lazowska et alt. 84]. Exact aggregation provides a tool to define an equivalent aggregated model at the higher level. Similarly, in a hierarchical top-down system design with predefined performance requirements, we can apply the inverse process called disaggregation or synthesis of the network to define a more detailed model with the same performance indices [Balsamo-Iazeolla 85]. The disaggregation process answer the question of what the system topology and parameter should be in order to achieve the given performance goal.
An important application of exact aggregation for product form queueing networks is the definition of various *approximate methods for non product form networks* [Marie 79]. These algorithms are usually based on an iterative scheme and they basically apply the aggregation theorem to non product form networks, although in this case it provides only approximate results. At each iteration step several subnetworks are analyzed and aggregated in the flow-equivalent service centers. This principle has been applied for the approximate analysis of various types of non product-form networks, such as for example networks with simultaneous resource possession and finite capacity queues.

Another interesting property of product form queueing networks is the *arrival theorem*. It states that the stationary queue length distribution at arrival times at a service center is identical to the stationary distribution at arbitrary times of the same networks, for open networks, and of a network with one less customer for closed networks [Lavenberg-Reiser 80, Sevcik-Mitrani 81]. A practical important consequence of this theorem is the definition of a set of recurrence equations between the performance indices of closed networks with K customers and those of the same network with K-1 customers. This led to the definition of Mean Value Analysis (MVA) [Reiser-Lavenberg 80], a recursive algorithm to evaluate performance indices of closed product form networks, without the direct evaluation of the normalization constant. This was a significant contribution in the research area of algorithms and tools for the efficient evaluation of product form queueing networks.

We shall now discuss the computational algorithms to analyze product form queueing networks.

4. Product form queueing networks: computational algorithms and tools

The main advantage of product form queueing networks is that several efficient algorithms have been developed for their performance analysis. As a consequence efficient and powerful performance evaluation tools based on product form network models have been developed and applied to obtain performance indices for large networks with many service centers and customers.
We shall now introduce the most used algorithm for product form BCMP networks. Two well-known algorithms for closed networks are Convolution Algorithm [Buzen 73] and the Mean Value Analysis [Reiser-Lavenberg 80, Reiser 81]. They provide the evaluation of a set of performance indices with a polynomial space and time computational complexity in the number of service centers and the network population.

4.1 Convolution Algorithm

For closed networks the computation of the stationary state distribution π requires the evaluation of the normalizing constant G in formula (4). Since V(n)=1, constant G is

defined as $G = \sum_{\mathbf{n}} \prod_{i=1}^{M} g_i(n_i)$. Direct computation of G as a summation over all the feasible states **n** of the network would take an exponential time in the number of service centers and customers of the network, i.e. proportional to the number of states of the network. The Convolution Algorithm avoids this direct computation and evaluates G recursively. For a network with M service centers and K customers let $G_j(k)$ denote the normalizing constant of the network with k customers and the first j service centers, $1 \le j \le M$ and $0 \le k \le K$. Then $G=G_M(K)$ and we can write the recursive relation that is the convolution

$$G_j(k) = \sum_{n=0}^{k} g_j(n) G_{j-1}(k-n) \qquad (7)$$

where $G_j(0)=1$. This basic scheme can be further simplified in some cases. If the first j service centers have infinite server discipline (type III BCMP node) then we can immediately write $G_j(k) = \left(\sum_{i=0}^{j} \rho_i\right)^k / k!$, with $\rho_i=x_i/\mu_i$. If service center j has a single server then we simply have $g_j(k)= \rho_j^k$. Hence convolution (7) reduces to

$$G_j(k) = G_{j-1}(k) + \rho_j^k G_{j-1}(k-1)$$

for $0 \le k \le K$. Therefore the time computational complexity of evaluating $G=G_M(K)$ is O(MK) operations. It is worthwhile noticing that several performance measures can be directly evaluated by function G_M. For a single server node j with load independent service rate we can write:

queue length distribution	$\pi_j(k) = \rho_j^k[G_M(K-k)-\rho_j G_M(K-k-1)]/G_M(K)$, $0 \le k \le K$
average queue length	$N_j = \sum_{k=1}^{K} \rho_j^k G_M(K-k)/G_M(K)$
throughput	$X_j = x_j G_M(K-1)]/G_M(K)$
utilization	$U_j = X_j / \mu_j$

However, for a service center j with load dependent service rate the queue length distribution can be written as follows $\pi_j(k) = g_j(k)\, G_{M-\{j\}}(K-k)/G_M(K)$, where $G_{M-\{j\}}$ is the normalizing constant of the entire network except for node j. This requires the solution of another network. Hence, the Convolution Algorithm efficiency is reduced when the network has several load dependent service centers.
A limitation of this algorithm is its potential numerical instability, i.e. possible overflow or underflow in the computation of constant G. Some scaling techniques to overcome this problem have been proposed [Lam 82].

4.2 MVA Algorithm

MVA Algorithm avoids the direct evaluation of the normalization constant. Consider a closed networks with M load independent service centers and K customers. Let $R_j(K)$, $X_j(K)$ and $N_j(K)$ denote respectively the average response time, the throughput and the average queue length of service center j. The algorithm is based on the following recursive scheme:

$$R_j(K) = [1/\mu_j] (1+N_j(K-1)) \qquad (8.1)$$

$$X_j(K) = K / [\sum_{i=1}^{M} x_i R_i(K)/x_j] \qquad (8.2)$$

$$N_j(K) = X_j(K) R_j(K) \qquad (8.3)$$

for $1 \le j \le M$, and $N_j(0)=0$. Formula (8.1) derives from the arrival theorem for product form closed networks while formulas (8.2) and (8.3) are Little's law applied to the entire network and node j, respectively. For infinite server queueing discipline the first relation simplifies in $R_j(K) = 1/\mu_j$. For load dependent service centers it is necessary to compute the queue length distribution of node j when there are K customers in the network, denoted by $\pi_j(k|K)$, $0 \le k \le K$. Then the first recursive relation of the algorithm becomes:

$$R_j(K) = \sum_{k=1}^{K} k \, \pi_j(k-1|K-1)/\mu_j(k)$$

and probability π_j is recursively evaluated as follows:

$$\pi_j(k|K) = \pi_j(k-1|K-1) \, X_j(K)/\mu_j(k), \; 1 \le k \le K, \quad \pi_j(0|K) = 1 - \sum_{k=1}^{K} \pi_j(k|K) .$$

Such a computation of the queue length distribution can lead to numerical instability. A modified MVA algorithm was proposed to overcome this drawback at the expenses of increased computational complexity. The empty node probability is recursively computed as follows: $\pi_j(0|K) = \pi_j(0|K-1)(X_i(K)/X_i^{M-\{j\}}(K))$ where $X_i^{M-\{j\}}(K)$ is the throughput of a service center i in the network obtained by the original one without node j. Hence for a network with J load dependent service center this modified version of the MVA algorithm requires the solution of 2^{J-1} additional networks to evaluate the throughput [Lavenberg 83].

Another interesting result of recursive relations for product form networks is the set of recursive expressions for the derivatives of higher moments of the queue length derived by McKenna and Mitrani [McKenna-Mitrani 84] that with the asymptotic expansion method allows obtained bounds for the higher moments of the queue length.

4.3 Multichain models

Convolution and MVA algorithms apply also to multiclass and multichain networks [Reiser-Lavenberg 80, Reiser 81, Lavenberg 83, Sauer 83, Lam 82]. However, their computational complexity for a network with M load independent service center, R closed chains and K_r customers in chain r=1,…,R is of $O(MR \prod_{1 \le r \le R} K_r)$ operations, i.e. it is exponential with the number of closed chains. This limitation led to the definition of special exact and approximate algorithms for multichain networks.

Exact methods are the tree Convolution and tree MVA algorithms that are efficient when customers of any given chain visit only a small number of service centers. This feature has been observed in models of computer and communication systems and in communication networks [Lam-Lien 83]. Tree Convolution and tree MVA algorithms extend respectively Convolution and MVA and are based on a tree data structure to optimize the algorithm computation [Lam-Lien 83, Tucci-Sauer 85, Hoyme et alt. 86].
Recursion by Chain Algorithm (Recal) [Conway-Georganas 86, Conway-Georganas 89] has a computational cost polynomial in the number of closed chains but exponential in the number of service centers. Recal recursively computes the normalization constant G of the product form solution and then one can observe overflow and underflow instability. The analogue extension of MVA to multichain is the Mean Value Analysis by Chain algorithm [Conway et alt. 89] that avoids the computation of the normalizing constant. If we want to

evaluate the joint queue length distribution we can use the Distribution Analysis by Chain (DAC) [DeSouza-Lavenberg 89] that also provides the average performance indices. The three algorithms Recal, MVA by Chain and DAC are efficient for a small number of service centers and many closed chains. Details on the algorithms can be found in literature [Chandy-Sauer 80, Lavenberg 83, Conway-Georganas 89, Kant 94]
Exact solution of multichain product form networks with a large number of customers, classes, chains and service centers is possible only if they have few closed chains by using tree structured algorithms or few service centers with Recal, MVA by Chain and DAC algorithms.

Approximate algorithms for product form networks can be non-iterative or iterative.

Simple non-iterative methods provide bounds on the performance measures. Various bounding methods such as Balanced Job Bounds, Proportional Bound and Performance Bound Hierarchies (PBH) techniques [Zahorjan et alt. 81, Eager-Sevcik 86, Hsieh-Lam 87, Hsieh-Lam 89] provide bounds on the average performance measures. Some techniques are based on the MVA equations, such that the PBH method that provides increasingly tighter bounds at the expense of increasing computation cost.
A different approach is the asymptotic expansion method [Mitra-McKenna 86] where the normalization constant G can be rewritten as a linear combination of terms that can be interpreted as normalizing constants of simple networks, with few chains and a small population. Hence G is approximated by these simpler computations. Moreover the method provides error bounds on the average performance measures.

Most iterative approximate methods are based on the MVA algorithm. The Bard-Schweitzer Proportional Estimation [Schweitzer 79] is a popular and widely applied approximate algorithm [Pattipati et alt. 90]. The average queue length $N_j(K-1)$ of a service center j for a network with K-1 customers is approximated as follows:

$$N_j(K-1)=N_j(K)\ K-1/K$$

Then by substituting this equation in formula (8.1) MVA becomes an approximate iterative algorithm. An improved algorithm called Linearizer [Chandy-Neuse 82] defines the difference between the fractional queue length of each service center at population K and K-1 as $D_j(K)=[N_j(K-1)/(K-1)]-[N_j(K)/K]$. While Schweitzer's approximation assumes that $D_j(K)=0$ for each service center j, Linearizer assumes that $D_j(K)$ is independent of K and approximates its value by iterations starting with $D_j(K)=0$. Linearizer is a quite accurate algorithm and further improvements have been developed [Pattipati 90]. Special extensions of these algorithms have been defined for networks with load dependent service centers.

Remark. The computational algorithms surveyed in this section solve BCMP product form networks. Note that, as discussed in the previous section, various extensions of this class of product form networks have been defined. However, the solution algorithms do not always immediately apply to non-BCMP product form networks. We have pointed out how load dependent service centers often lead to special recursive formulas, like in Convolution and MVA algorithms. Similarly, solving product form networks with special features, such as state dependent routing, finite capacity queues and blocking, special queueing disciplines, negative customers and batch arrivals and services is in general a non trivial problem. Some algorithms have been defined for some classes of product form networks. For example special algorithms have been defined for some product form networks with a particular queueing discipline that models multiserver centers with concurrent classes of customers [Le Boudec 88] or for networks with finite capacity queues and blocking [Balsamo-Clò 98].

4.4 Queueing networks tools

Beside performance measurement tool, performance modelling tools based on queueing networks have been developed. The workload characterization tools provide the quantitative characterization of the system resource demands of workload that is used to define the performance models. Software tools for performance modelling and analysis integrate the computational algorithms to solve queueing network models with a model specification language. Such tools usually have user friendly interfaces based on different languages to take into account the particular field of application, e.g. computer networks, communication networks, distributed computer systems. This allows not expert users to apply efficient performance modelling techniques.
Some tools include hierarchical modelling techniques and allow the definition of various system performance models at various levels. Two models at different level are related by the aggregation and disaggregation technique.
Most performance evaluation packages include exact BCMP product form solution methods, e.g. at least Convolution and/or MVA algorithms and possibly other algorithms. Some tools provide approximate solution methods usually based on an approximate product form solution. Many packages give the user the choice between analytical methods and simulation. Examples of performance evaluation packages are Best-1, RESQ/IBM, QNAP2, HIT [Lavenberg 83, Lazowska et alt. 84, Beilner et alt. 95, Potier 86] just to mention a few.
More recently the solution performance algorithms have been integrated with model specification techniques to provide tools for the combined functional and quantitative system analysis [Smith 90].

5.Status and future directions

Product form queueing networks has proved to be a very useful class of models for system performance evaluation. This is due to a good balance between the relative high accuracy and robustness of performance results and the efficiency in model solution. The precise characterization of product form networks is not trivial in terms of model characteristics, as discussed in Section 3, since the properties are basically related to the associated Markov process.
Several special extensions of the class of BCMP product form networks have been obtained to include various interesting system features, such as state dependent routing, blocking, negative customers and batch customer movements. However, most of these models have product form solution under several constraints on the system structure and parameters. An open problem is the definition of efficient algorithms for the computation of performance indices of non-BCMP queueing networks, such as for example the models with batch arrivals and departures. Efficient analysis of discrete-time queueing networks is a related open problem.

Today the product form class seems to be well defined and it is difficult to expect that further wide extensions will be discovered or defined.
Product form networks provide the basis for many approximate algorithms to solve more general non product form models. Hierarchical modelling and decomposition-aggregation techniques are the main tools in this area. Hence exploiting the robustness of the properties of product form networks can still be useful to solve more general networks. Interesting and useful properties are insensitivity, exact aggregation and the arrival theorem. A research issue is how to efficiently combine subnetwork solution in a decomposition aggregation framework to obtain an approximate possibly error bounded solution. This potentially leads to develop simple and efficient performance modelling tools.
Another research issue is the integration and/or the relation between queueing networks and other classes of models with different characteristics, such as for example stochastic Petri nets or stochastic process algebra. A challenge could be to develop efficient and integrated tools that combine qualitative and quantitative system analysis, such as software architecture specification and system performance.

References

[Akyidiz 87] I.F. Akyildiz "Exact product form solution for queueing networks with blocking" IEEE Trans. on Computer, Vol. C-36-1, pp. 122-125, 1987.

[Balsamo-DeNitto 94] S. Balsamo, V.De Nitto "A survey of Product-form Queueing Networks with Blocking and their Equivalences" Annals of Operations Research, Vol. 48, pp. 31-61, 1994.

[Balsamo-Clò 98] S. Balsamo, M.C. Clò "A Convolution Algorithm for Product-form Queueing Networks with Blocking" Annals of Operations Research, Vol. 79, pp. 97-117, 1998.

[Balsamo-Iazeolla 82] S. Balsamo, G. Iazeolla "An Extension of Norton Theorem for Queueing Networks" IEEE Trans. on Software Engineering, Vol. SE-8, 1982.

[Balsamo-Iazeolla 85] S. Balsamo, G. Iazeolla "Product-form Synthesis of Queueing Networks" IEEE Trans. on Software Engineering, Vol. SE-11, No. 2, pp. 194-199, 1985.

[Basket et alt. 75] F. Baskett , K.M. Chandy, R.R.Muntz, G. Palacios "Open, closed, and mixed networks of queues with different classes of customers" Journal of the ACM, Vol. 22, No.2, pp. 248-260, 1985.

[Beilner et alt. 95] H. Beilner, J. Mäter, C. Wysocki "The Hierarchical Evaluation Tool HIT" 581/1995, University of Dortmund, Dortmund, Germany, 6-9, 1995.

[Boucherie-VanDijk 91] R. Boucherie, N.M. van Dijk "Product-form queueing networks with state dependent multiple job transitions" Adv. in Applied Prob. , Vol. 23, pp. 152-187, 1991.

[Boucherie-VanDijk 97] R. Boucherie, N.M. van Dijk "On the arrival theorem for product-form queueing networks with blocking" Performance Evaluation, Vol. 29, pp. 155-176, 1997.

[Burke 56] P.J. Burke "The output of a queueing system" Oper. Res., Vol. 4, pp. 699-704, 1956.

[Buzen73] J. P. Buzen "Computational algorithms for closed queueing networks with exponential servers" Comm. of the ACM, Vol. 16, No. 9, pp. 527-531, 1973.

[Chandy et alt. 75] K.M. Chandy, U. Herzog and L. Woo "Parametric analysis of queueing networks" IBM Journal of Res. and Dev., Vol. 1, No. 1, pp. 36-42, 1975.

[Chandy et alt. 77] K.M. Chandy, J.H. Howard and D. Towsley "Product form and local balance in queueing networks" Journal of the ACM, Vol. 24, No.2, pp. 250-263, 1977.

[Chandy-Martin 83] K.M. Chandy, A.J. Martin "A characterization of Product-Form Queueing Networks" Journal of the ACM, Vol. 30, No. 2, pp. 286-299, 1983.

[Chandy-Neuse 82] K.M. Chandy, D. Neuse "Linearizer: a heuristic algorithm for queueing network models of computer systems" Comm. of the ACM, Vol.25, pp.126-134, 1982.

[Chandy-Sauer 80] K.M. Chandy, C.H. Sauer "Computational algorithms for product form queueing networks" Comm. of the ACM, Vol. 23, No. 10, pp. 573-583, 1980.

[Conway et alt. 89] A.E. Conway, E. de Souza e Silva, S.S. Lavenberg "Mean Value Analysis by Chain of Product-Form Queueing Networks" IEEE Trans. on Computers, Vol. C-38, No. 10, pp. 573-583, 1989.

[Conway-Georganas 86] A.E. Conway, N.D. Georganas "RECAL - a new efficient algorithm for the exact analysis of multiple-chain closed queueing networks" Journal of the ACM, Vol. 33, pp. 768-791, 1986.

[Conway-Georganas 89] A.E. Conway, N.D. Georganas "Queueing Networks - Exact Computational Algorithms" MIT Press, Cambridge, Massachusetts, 1989.

[Courtois 77] P.J. Courtois "Decomposability" Academic Press, New York, 1977.

[DeSouza-Lavenber 89] E. De Souza e Silva, S.S. Lavenberg "Calculating the joint queue length distribution in product-form queueing networks" Journal of the ACM, Vol. 36, pp. 194-207, 1989.

[DeSouza-Muntz 89] E. De Souza e Silva, R.R. Muntz "Queueing Networks: Solutions and Applications" in "Stochastic Analysis of Computer and Communication Systems" (H. Takagy Ed.), pp.319-399, Elsevier, North Holland, 1990.
[Eager-Sevcik 86] D.L. Eager, K.C. Sevick "Bound Hierarchies for multiple-class queueing networks" Journal of the ACM, Vol. 33, pp. 179-206, 1986.
[Gelenbe-Mitrani 80] E. Gelenbe, I. Mitrani "Analysis and Synthesis of Computer Systems", Academic Press, New York, 1980.
[Gelenbe 91] E. Gelenbe "Product form networks with negative and positive customers" Journal of Applied Prob., Vol. 28, No. 3, pp. 656-663, 1991.
[Gordon-Newell 67a] W.J. Gordon, G.F. Newell "Cyclic Queueing Networks with exponential servers" Operations Research, Vol. 15, No. 2, pp. 254-265, 1967.
[Gordon-Newell 67b] W.J. Gordon, G.F. Newell "Cyclic Queueing Networks with restricted length queues" Operations Research, Vol. 15, No. 2, pp. 266-277, 1967.
[Henderson-Taylor 90a] W. Henderson, P. Taylor "Product form in networks of queues with batch arrivals and batch services" Queueing Systems, Vol. 6, pp. 71-88, 1990.
[Henderson-Taylor 90b] W. Henderson, P. Taylor "Some new results on queueing networks with batch movements" J. of Applied Prob., Vol. 28, pp. 409-421, 1990.
[Hoyme et alt. 86] K.P. Hoyme, S.C. Buell, P.V.Afshari, R.Y Jain "A tree structured Mean Value Analysis Algorithm" ACM Trans. on Computer Systems, Vol.4, pp. 178-185, 1986.
[Hsieh-Lam 87] C.T. Hsieh, S.S. Lam "Two classes of performance bounds for closed queueing networks" Performance Evaluation, Vol. 7, pp. 3-30, 1987.
[Hsieh-Lam 89] C.T. Hsieh, S.S. Lam "Pam - a noniterative approximate solution method for closed multichain queueing networks" Performance Evaluation, Vol. 9, pp. 119-133, 1989.
[Jackson 63] J.R. Jackson "Jobshop-Like Queueing Systems" Management Science, Vol. 10, pp. 131-142, 1963.
[Jain 90] R. Jain "The Art of Computer System Perforance Analysis" John Wiley, New York, 1990.
[Kant 92] K. Kant "Introduction to Computer System Performance Evaluation" MacGraw-Hill, 1992.
[Kelly 79] F.P. Kelly "Reversibility and Stochastic Networks" Wiley, New York, 1979.
[King 90] P.J.B. King "Computer and Communication System Performance Modelling" Prentice-Hall, Englewood Cliffs, 1990.
[Kleinrock 75] L. Kleinrock "Queueing Systems, Vol.1: Theory" John Wiley, New York, 1975.
[Kritziger et alt. 82] P. Kritzinger, S. van Wyk, A. Krezesinski "A generalization of Norton's theorem for multiclass queueing networks" Performance Evaluation, Vol. 2, pp. 98-107, 1982.
[Lam 77] S.S. Lam "Queueing networks with capacity constraints" IBM Journal of Res. and Dev., Vol. 21, No. 4, pp. 370-378, 1977.
[Lam 82] S.S. Lam "Dynamic scaling and grow behavior of queueing networks normalization constant" Journal of the ACM, Vol. 29, No. 2, pp. 492-513, 1982.
[Lam-Lien 83] S.S. Lam, Y.L. Lien "A tree convolution algorithm for the solution of queueing networks" Comm. of the ACM, Vol. 26, No. 3, pp. 203-215, 1983.
[Lavenberg 83] S.S. Lavenberg "Computer Performance Modeling Handbook" Academic Press, New York, 1983.
[Lavenberg-Reiser 80] S.S. Lavenberg and M. Reiser "Stationary State Probabilities at Arrival Instants for Closed Queueing Networks with multiple Types of Customers" Journal of Applied Prob., Vol. 17, pp. 1048-1061, 1980.
[Lazowska et alt. 84] E.D. Lazowska, J.L.Zahorjan, G.S. Graham, K.C. Sevcick "Quantitative System Performance: Computer System Analysis Using Queueing Network Models" Prentice Hall, Englewood Cliffs, NJ, 1984.
[Le Boudec 86] J.Y. Le Boudec "A BCMP extension to multiserver stations with concurrent classes of customers" Proc. Performance '86 and 1986 ACM Sigmetrics Conf., pp. 79-81, 1986.

[Le Boudec 88] J.Y. Le Boudec "The Multibus algorithm" Performance Evaluation" Vol.8, pp.1-18, 1988.
[McKenna-Mitrani 84] J. McKenna, I. Mitrani "Asymptotic Expansions and integral representations of moments of queue lengths in closed Markovian networks" Journal of the ACM, Vol. 31, pp. 346-360, 1984.
[Marie 79] R. Marie "An Approximate Analytical Method for General Queueing Networks" IEEE Trans. on Software Eng., Vol. 5, No. 5, pp. 530-538, 1979.
[Mitra-McKenna 86] D. Mitra, J. McKenna "Asymptotic Expansions for closed Markovian networks with state dependent service rates" Journal of the ACM, Vol. 33, pp. 568-592, 1986.
[Nelson 93] R. Nelson "The Mathematics of Product-Form Queueing Networks" ACM Computing Survey, Vol. 25, No. 3, pp. 339-369, 1993.
[Pattitapati et alt. 90] K.R. Pattitapati , M.M. Kostreva, J.L. Teele "Approximate Mean Value Analysis Algorithms for Queueing Networks: Existence, Uniqueness and Convergence Results" Journal of the ACM Vol. 37, pp.643-673, 1990.
[Potier 86] D. Potier "The modelling package QNAP2 and applications to computer networks simulation" in Computer Networks and Simulation, North Holland, 1986.
[Reiser-Lavenberg 80] M. Raiser, S.S. Lavenberg "Mean Value Analysis of Closed Multichain Queueing Networks" Journal of the ACM, Vol. 27, No. 2, pp. 313-320, 1980.
[Reiser81] M. Raiser "Mean Value Analysis and Convolution Method for Queue-Dependent Servers in Closed Queueing Networks "Performance Evaluation, Vol. 1, No. 1, pp. 7-18, 1981.
[Sauer 83] C.H. Sauer "Computational Algorithms for State-Dependent Queueing Networks" ACM Trans. on Computer Systems, Vol. 1, No. 1, pp. 67-92, 1983.
[Schweitzer 79] P.J. Schweitzer "Approximate analysis of multiclass closed networks of queues" Proc. of Int. Conference on Stochastic Control and Optimization, pp.25-29, Amsterdam, The Netherlands, 1979.
[Sevcik-Mitrani 81] K.S. Sevcik and I. Mitrani "The Distribution of Queueing Network States at Input and Output Instants" Journal of the ACM, Vol. 28, No. 2, pp. 358-371, 1981.
[Shassenberger 78] R. Shassenberger "The insensitivity of stationary probabilities in networks of queues" Journal of Applied Prob., Vol.10, pp. 85-93, 1978.
[Smith 90] C. Smith "Performance Engineering of Software Systems" Addison-Wesley, Reading, MA, US, 1990.
[Suri83] R. Suri "Robustness of queueing network formulas" Journal of the ACM, Vol. 30, No. 3, pp. 564-594, 1983.
[Towsley 80] D. Towsley "Queueing Network with State Dependent Routing "Journal of the ACM, Vol. 27, No. 2, pp. 323-337, 1980.
[Trivedi 82] K.S. Trivedi "Probability and Statistics with Reliability, Queueing and Computer Science Applications" Prentice Hall, Englewood Cliffs, 1982.
[Tucci-Sauer 85] S. Tucci, C.H. Sauer "The tree MVA algorithm" Performance Evaluation, Vol. 5, pp. 187-196, 1985.
[Van Dijk 93] N. van Dijk "Queueing networks and product forms" John Wiley, 1993.
[Warland 88] J. Warland "An Introduction to Queueing Networks", Prentice-Hall 1988.
[Whittle 85] P. Whittle "Partial balance and insensitivity" J. of Applied Prob., Vol. 22, pp. 168-175, 1985.
[Zahorjan et alt. 81] J.L.Zahorjan, K.C. Sevcick, D.L. Eaer, B. Galler "Balanced Job Bound analysis of Queueing Networks" Comm. of the ACM, Vol. 25, pp. 134-141, 1981.