QUEUEING NETWORK MODELING OF ELEMENTARY MENTAL PROCESSES

Yili Liu, Ph.D.
Dept of Industrial and Operations Engineering
The University of Michigan
1205 Beal Avenue
Ann Arbor, Michigan 48109-2117

Phone: 313-763-0460
FAX: 313-764-3451
Email: yililiu@engin.umich.edu

Technical Report 94-21

September 1994

Queueing Network Modeling of Elementary Mental Processes

Abstract

This article presents a queueing network model of elementary mental processes. As a continuous-transmission network in its general form, the model unifies the existing discrete and continuous serial models and discrete network models in a larger modeling framework, and covers a broader range of temporal and structural arrangements that mental processes might assume and can be subjected to empirical tests. Five elementary but important types of queueing networks are described in detail and they are used to reexamine existing models for reaction time. These networks include a tandem network that takes existing discrete and continuous serial models as special cases, a fork-join network that takes the PERT network model as a special case, a feedback network that mimics a serial network with identical stages accurately at the level of mean RT, a Simon-Foley network that allows noise components to overtake signal components and predicts overadditive factor effects, and a cyclic network that processes a fixed number of stimulus components at a time and predicts underadditive factor effects.

Why is there a delay between stimulus presentation and response initiation? This has been one of the most enduring and fundamental questions that psychologists have been fascinated with. The current belief of cognitive psychologists is that this delay is a reflection of the dynamic activities of an underlying mental structure that transforms stimulus into response. And most importantly, since the cognitive system is not amenable to open inspection, the characteristics of this delay--also called reaction time (RT)--may offer important clues to the possible configurations of the mental structure.

Theoretical models that use reaction time as the primary performance measure to infer the general structure of mental systems are often called models for RT. Of great interest to the present article are two issues that are central to RT modeling and theory in cognitive psychology. The two issues also define two dimensions along which RT models can be classified. One of the two is a temporal dimension distinguishing discrete-transmission models from continuous-transmission models, and the other an architectural dimension distinguishing serial-stage models from network models. All of the models assume that the psychological activity that transforms stimulus into response is composed of a system of mental processes. Discrete-transmission models assume that a mental process transmits its processing output in an indivisible unit and will not make its output available to other processes until it is completed. Therefore, a process can not begin until all of its preceding processes are completed. Continuous-transmission models, in contrast, assume that each process transmits its partial outputs to other processes continuously as soon as they are available rather than waiting for the full completion of processing, and thus a process can begin even though its preceding processes are still active. Serial-stage models assume a serial arrangement of mental processes, whereas network models assume a network configuration. The two dimensions jointly define four classes of models as shown in Figure 1.

While the distinction between the terms "serial" and "network" is usually quite standard in the literature, there exist some differences in the use of the terms "discrete"

and "continuous" by different authors. As discussed by Miller (1988, 1990), the terms "discrete" and "continuous" have been used in at least four different senses in cognitive models--discrete versus continuous information representations, discrete versus continuous information transformation, discrete versus continuous information transmission, and discrete versus continuous variation in a priori state of an information processing stage (see, Miller, 1988, 1990, for excellent discussions of this topic). The aim of the present paper is to address the issue of discrete versus continuous information transmission in conjunction with the issue of serial versus network arrangements. Several important RT models are therefore not included in Figure 1, primarily because their concerns were on discrete versus continuous information representation or transformation. Prominent among these models include the model developed by Meyer and his colleagues (Meyer, Irwin, Osman, and Kounios, 1988) and the stochastic diffusion model (Ratcliff, 1988).

It should also be noted here that, although the terms "continuous" and "discrete" have been used extensively in the literature to refer to models that do or do not allow partial output and temporal overlap of process durations, there is no intrinsic relationship between continuity of transmission and temporal overlapping of process activities. A process may continuously transmit its partial output to its successors, but processes could still be in strict temporal sequence if each process has to wait until it has accumulated all of the continuously-arrived inputs before it starts. Similarly, although partial outputs transmitted in the form of a continuous flow will support overlapping process activities, those transmitted as "discrete packets" will do so as well, as long as the number of packets that can be separately transmitted is greater than one. However, due to the lack of a better term and to be consistent with the common practice in RT modeling, I will continue to use "continuous" and "discrete" transmission to distinguish whether or not a series of processes could be active concurrently.

Miller (1982, 1988) has suggested that discrete and continuous transmissions be viewed as the extremes of a continuum defined by the extent to which the output of a stage can be divided and separately transmitted to other stages (the grain size of transmission). At one extreme, discrete transmission models have the largest possible grain size because the output must be transmitted as a whole unit. At the other extreme, outputs in continuous-*flow* models can be divided into an arbitrarily large number of small units. Intermediate models assume grain sizes between the two extremes, which are also called "nondiscrete models" (Miller, 1993). In this article, I will use the term "continuous" to refer to both truly continuous flows and continuous transmission of discrete packets of partial outputs.

Historically, the modeling work covered in Figure 1 started with the serial discrete-stage models shown in the top-left quadrant. These models assume non-overlapping durations of serially arranged processes or stages. The underlying models for the subtraction method developed by Donders (1868/1969) and the additive factor method developed by Sternberg (1969) belong to this class of models. Donders assumed that processes can be added or deleted from a chain of processes while leaving intact the rest of the chain (called the assumption of pure insertion). Based on this assumption, Donders proposed that the mean duration of an inserted or deleted process can be inferred by examining the difference between the mean duration of a task that does not include the process in question and one that does--a method known as the subtraction method for mean RT analysis. Since pure insertion appears to be a strong assumption, Sternberg tried to relax this assumption by addressing the issue of how experimental manipulations might change the durations of processes rather than insert or delete them. Sternberg assumed that the mean duration of a process depends on experimental manipulations that influence it, but not directly on the mean durations of other processes, and a change in the mean duration of a process will not produce indirect effects on the mean durations of other processes in the processing chain (called the assumption of selective influence).

Based on this assumption, Sternberg proposed an additive factor method for mean RT analysis, according to which experimental factors that influence a common process will interact with each other in an analysis of variance of the RT data, whereas those influencing separate processes will be additive. The serial discrete-stage model and the additive factor method have been the fundamental basis of a large body of experimental literature.

While the models underlying Donders's and Sternberg's methods are models for mean RT, numerous authors have examined properties of RT at the distributional level, since much more can be obtained from examining RT distributions than from examining mean RTs alone. It has been shown that examining RT distributions could be critical in discriminating models that would demonstrate similar behavior at the mean level. When the durations of serial processes are independent of each other, RT distribution is the convolution of the process durations. McGill and Gibbon (1965) noted that reaction time in a serial discrete-stage model can be described by the general-gamma distribution, if the independent stage durations are exponentially distributed with different duration means. Several authors argued that the convolution of normal and exponential distributions provides a close approximation to observed RT distributions (Hockley, 1984; Hohle, 1967; Ratcliff and Murdock, 1976). Ashby and Townsend (1980, Ashby, 1982; Townsend and Ashby, 1983) extended the assumptions of pure insertion and selective influence to the distributional level, and proved a set of theorems that can be employed to test these assumptions.

Models in the other quadrants of Figure 1 try to relax the assumption of serial and non-overlapping process activities adopted by serial discrete-stage models with an aim to generalize the class of RT models and broaden the range of possible mental structures for elementary psychological processes. In the bottom-left quadrant of Figure 1, we find the models that permit temporal overlap of sequentially arranged processes. Prominent

among this class of RT models include McClelland's (1979) Cascade model, and more recently, Miller's (1993) queue-series model.

The cascade model assumes that the human information processing system functions like a series of parallel linear integrators. These linear integrators take a weighted sum of a subset of the outputs of the integrators at the preceding level and produces continuous output that is always available for processing at the next level. The heart of the cascade model is a cascade equation, which expresses the activation of a linear integrator at a processing level as a function of the asymptotic activation of the linear integrator that would result if the stimulus were left on indefinitely, and the rate constants of the different processes in the system. McClelland examined the effects of manipulating rate constants and asymptotic levels on RT and derived a set of predictions for RT behavior. Similar to the additive factor method, the cascade model shows that experimental factors affecting the rate of the same process will interact, whereas those affecting the rates of different processes are additive. However, the predictions become more complicated and start to diverge from those of the additive factor method when at least one of the experimental factors affects the asymptotic level of activation.

The queue-series model recently developed by Miller (1993) assumes that the cognitive system is composed of a series of stages and the stimulus is regarded as consisting of a number, M, of distinct components. The important concept of grain size of transmission is mathematically represented by the parameter M. Discrete-stages and cascade-flows are treated as special cases of the queue series, corresponding to the cases of $M=1$ and $M=\infty$, respectively. Other positive values of M represent intermediate degrees of transmission continuity. In the queue series model, components are processed separately through the series of stages. After finishing processing a component, each stage immediately passes along that component to the next stage. The first stage processes components either serially or in parallel, but each of the subsequent stages selects and processes one component at a time. Thus, components may have to wait in

front of slow stages, and a series of queues could be observed in the system. The response is made when all of the M components have passed through the series. Based on the results of numerical simulations of the time for M customers to traverse through the system, Miller concluded that, within the class of queue series models he considered, experimental factors affecting different processing stages always have additive effects on reaction time with discrete sequential stages but rarely do so with overlapping stages, and thus, observations of factor additivity support discrete-stage models.

Insert Figure 1 about Here

Along another line of investigation, several authors have challenged the notion of serial arrangement of mental processes and examined the possibility of parallel or network configurations of process activities. Although the debate about whether the human mind is capable of engaging in more than one mental activities at a time can be traced back to the work of ancient philosophers and the founders of modern experimental psychology, Christie and Luce (1956) appear to have been the first to address the issue of parallel versus serial processing from a mathematical perspective. The most systematic work in this area is a series of studies of Townsend (1972, 1974, 1976) and that of Ashby and Townsend (1980, Townsend and Ashby, 1983). Townsend extended the notion of processing stages from serial to parallel systems and analyzed the identifiability and the equivalence of serial and parallel systems. Ashby and Townsend extended the assumptions of pure insertion and selective influence from the level of mean RTs to the level of distributions in serial and parallel systems.

A central concept in Townsend's analysis is the notion of intercompletion times (ICTs), which refers to the time interval from the completion of one task element to the completion of the next. In serial systems like those of Donders and Sternberg, task elements are processed one at a time in strict serial stages, and each ICT is the duration of

a corresponding stage. In parallel systems, however, individual elements are processed simultaneously, but they may be completed at different times. The fundamental basis for Townsend's definition of a stage is serial completion rather than serial initiation or concatenation of separate processes. Successive processing stages are defined as the state of the system between successive completions of task elements. For example, stage i is the state of the system from the completion of the (i-1)th element to the completion of the ith. The concept of ICT played a critical role in Townsend's analysis of the identifiability and equivalence of serial and parallel models (Townsend, 1974) and in Ashby and Townsend's extension of the assumptions of pure insertion and selective influence from the level of mean RTs to the level of distributions in serial and parallel systems (Ashby and Townsend, 1980).

Townsend (1974) mentioned the possibility of hybrid mental systems that process task elements in neither a parallel nor a serial manner. But a mathematical analysis of this class of system in cognitive modeling did not appear until Schweikert (1978) introduced PERT (program evaluation and review technique) methods to this research area. Schweikert (1978) developed a class of PERT network models of mental processes, which assume that the processes can be arranged as a network with serial and parallel structures as special cases. The models adopted two fundamental assumptions: First, in a PERT network, processes that are not on the same path are allowed to be active at the same time, but those on the same path are assumed to operate in strict sequence. In other words, a process can not start until all the preceding processes on the same path are completed. Second, the PERT method for RT analysis follows the postulate of selective influence of Sternberg (1969) and assumes that each experimental manipulation prolongs the duration of one process, but does not change the duration of any other process. An important result of PERT analysis is that when all the processes are arranged in sequence, the predictions of the PERT method is consistent with that of the additive factor method, in that experimental factors affecting the durations of separate processes are additive.

However, if all the processes are not arranged in a series, the effects on RT of experimental manipulations prolonging separate processes can be interactive.

Schweikert (1978) started with deterministic PERT networks in which process durations are assumed to be constants. This assumption was later relaxed in the subsequent developments of stochastic PERT network models, in which process durations are random variables (Fisher and Goldstein, 1983; Schweikert, 1982; Schweikert and Townsend, 1989; Townsend and Schweikert, 1989). Townsend and Schweikert (1985, 1989; Schweikert and Townsend, 1989) later generalized the results of Townsend (1974), Schweikert (1978) and Ashby and Townsend (1980) on serial, parallel, and hybrid network processes, and identified a set of conditions under which additivity, overadditivity and underadditivity of experimental factors are expected. All this work, however, assumes strict serial operations for processes on the same path, and thus, in essence, a PERT network is a network of discrete process chains. This class of models have naturally been referred to as discrete mental networks (Townsend and Schweikert, 1989) and is shown in the top-right quadrant of Figure 1.

The trend of research reviewed above is clearly in the direction of expanding the scope of modeling to cover a broader range of temporal and architectural arrangements that mental processes might assume. Later models try to relax certain assumptions of the earliest model assuming serial and discrete stages, while at the same time, try to mimic the behavior of serial discrete-stage models, and identify the conditions under which these models converge or diverge in their predictions and explanations of RT data in factorial experiments. The cascade model matches closely the predictions of the additive factor method if experimental manipulations only affect the rate of activations but not the asymptotic level of processes. The queue-series model supports the additive factor method only when it functions like a discrete series. A PERT network is fully consistent with the additive factor method if all the PERT processes are arranged in a sequence.

In this article, I present a queuing network model for reaction time and elementary mental processes. The model, in its most general form, is a continuous-transmission-network model. As will be shown below, the model takes the existing models in the other three quadrants of Figure 1 as special cases, and thus attempts to unify them in a larger modeling framework. I will also reexamine the logic and conclusions of these models. It turns out that many of the conclusions based on the previous models are open to alternative explanations. Furthermore, I will show that the queueing network model allows us to cover a broader range of possible mental structures that can be subjected to empirical testing.

The idea of a queueing network arises naturally when one thinks of a network of service stations (also called service centers, or simply, nodes), each of which provides a service of some kind to the demanders for service (called customers), either immediately or after a delay. Each node has a waiting space for customers to wait if they cannot immediately receive their requested service, and thus multiple queues may exist simultaneously in the system. The nodes are connected by arcs over which customers flow from node to node in the network. Telephone communications systems, computer networks and road traffic networks are examples of queueing networks. The study of networks of queues started with the work of Erlang (1917) in the area of telephone communications. However, much of queueing network theory stems from two papers by J.R. Jackson (1957, 1963). Since then, queueing network theory has been extensively applied to the modeling and analysis of a large variety of real world systems, and has become one of the most commonly used tools for system performance analysis.

It is not difficult to see, at least at the conceptual level, the close resemblance between a queueing network and the current views of a human cognitive system. However, in order to link queueing network theory with observable RT behavior, we need to specify the model in concrete terms and make a set of specific assumptions. In the following I will first present the queueing network model for reaction time in its general

form, and discuss the basic assumptions of the model. This general discussion is followed by 4 sections, each of which discusses an important and elementary type of queueing network. I will show that the discrete and continuous-flow serial models are special cases of a special type of queueing network called tandem queues, and PERT networks can be regarded as a special case of another type of queueing network called fork-join networks. A queueing network that allows noise components to overtake or bypass signal components is described as a possible cause of positive dependencies between process durations. Negative process duration dependencies are then shown to be an important characteristic of closed queueing networks. Neither of the two classes of networks follows the postulate of selective influence and both can be subjected to empirical testing. Overadditive and underadditive factor effects observed in some factorial experiments are given an interesting alternative explanation in terms of these queueing networks.

General Descriptions of the Queueing Network Model

General Assumptions and Notations

The queueing network model for reaction time assumes that a reaction time task is carried out by a network of processing nodes, each of which provides a distinct type of information processing service to the customers. I will also use the term "stimulus components" adopted in Miller (1993) to refer to these customers. There are at least two types of nodes in the network--those that receive customers from outside the network (called input nodes) and those that only receive customers from other nodes of the network (called internal nodes). A special class of queueing network called fork-join network requires two additional types of nodes, which will be described in the section about fork-join networks. For reasons that will be described below, an input node may or may not immediately transmit its outputs to other nodes. But each internal node begins processing as soon as it receives some customer, and immediately transmits any available output (a serviced customer) to other nodes or to the outside of the network without

waiting for the full completion of processing all of its customers. A node processes one customer at a time (a single-channel server) and processes its customers according to their order of joining the queue in front of the node (First-come-first-serve, FCFS). The assumption of single channel processing is commonly made in psychological theories, a recent example of which can be found in Miller's queue-series model. The implications of assuming FCFS single-channel nodes for RT modeling will be discussed later. Consistent with another common assumption of RT models, including the cascade model and the queue-series model, I assume that there is a separate response unit at the end of the processing network, which is responsible for the actual response.

Queueing network theory assumes that the sequence of customer arrival times and the sequence of customer service times are random processes. The arrival pattern or input to a node is described by the probability distribution of successive arrivals, which include both external arrivals from outside the network and internal arrivals from other nodes in the network. The service pattern or output of a node is described by the probability distribution for service time. In order to describe a queueing network, a set of rules called a switching process is also needed, which serves to route customers through the nodes of the network from entrance to exit. This collection of nodes, arcs, external and internal arrival processes and switching processes constitute the basic elements of all queueing networks, including the queueing network model for RT.

In order to represent the queueing network model for RT mathematically, the following notations are needed, which are now rather standard in the queueing network literature. Two sets of notations are needed, the first for describing a stochastic queueing process at a node, and the second for stochastic processes in a queueing network. A queueing process at a service node in a network is described by a series of symbols and slashes such as A/B/C/D/E, where A indicates the arrival pattern of customers as described by the probability distribution for interarrival-time or arrival rate, B the probability distribution for service time, C the number of parallel service channels at the

node, D the restriction on waiting room capacity in front of the node, and E the queue discipline (the manner by which the customers are selected from the queue for service). The following symbols are used in this article to represent a queueing network:

1) K: the total number of nodes,

2) i: the identity of a node,

3) γ_i: the mean arrival rate to node i from outside the network (the external arrival rate),

4) λ_i: the total mean arrival rate into node i (from outside the network and from other nodes),

5) p_{ij}: the probability that a customer visits node j immediately after departing from node i (the routing probability or switching probability), i=1,..., K, j=0,.., K, with pi0 representing the probability that a customer leaves the network immediately after visiting node i,

6) μ_i: the mean service rate for each channel of node i.

The state of a queueing network at any time instant, t, is commonly described by a stochastic process called the queue length process, which characterizes the number of customers (including those in service and those waiting in the queue) at each service center of the network at that instant. More formally, the queue length process describes the joint probability distribution for the number of customers at each service center, and is expressed mathematically as a vector-valued stochastic process with the vector element, Ni(t), i=1, ..., K, specifying the queue length at node i at time t.

Stimulus Components as Customers

The model assumes that a stimulus is composed of a number, C, of distinct classes of components, with Ni components of class i, i=1,..., C. In the simplest case, there is only one class of stimulus component that is responsible to RT (this is the case considered in the queue-series model of Miller). We may call them "signal" components. In a more general case, there may be two classes of stimulus components-- "signal"

components and "noise" components. For most of the networks considered in this article, the model assumes that the two types of components take the same path and have the same service time requirements and thus they are indistinguishable. It will be shown in a later section that this distinction becomes critical when noise and signal components are assumed to take separate paths of a network, and this distinction will be shown to have significant implications for queueing network analysis of certain types of RT behavior. It is easy to image situations in which a finer distinction between the classes of stimulus components may be necessary, but this paper will not extend the discussion further to include those cases. As nicely summarized in Miller (1993), "the stimulus components may be regarded as elementary stimulus features, complex semantic codes, objects, or the associated neural activations" (p.703), and as in Miller, this article will not attempt to develop the empirical means of identifying stimulus components.

Component Arrivals, Services and Routing Characteristics

As pointed out by Pachella (1974), the definition of stimulus onset is not always psychologically obvious. It seems improbable that all stimulus components will arrive at the perceptual receptor precisely at the same time. It is consistent with our intuition that auditory stimuli are spread out in time, but at a finer level of analysis, even visual stimuli are spread out in time. The main difference between the two is in the rate at which the stimulus components arrive at the perceptual system. In probability theory and in the queueing literature, the most commonly adopted assumption is that the interarrival times of successive customers are independent and exponentially distributed with a mean arrival rate of λ. In other words, the component arrival process is a Poisson process with rate parameter λ. The queueing network model for RT will adopt this assumption, except for a few special classes of networks discussed below for which more general results are available that do not rely on this assumption.

Another commonly made assumption in the queueing network literature is that at node i, customers have an exponentially distributed service time requirement with a mean

duration of $\frac{1}{\mu_i}$. μ_i is often called the service rate of node i. This assumption is similar to the assumption of exponential process durations, which is commonly made in RT modeling. As discussed by numerous authors, this assumption is not as strong as it appears to be, and has received support from experimental studies (Ashby, 1982; Fisher and Goldstein, 1983; Hockley, 1984; Hohle, 1967; Ratcliff and Murdock, 1976; Rumelhart, 1970; Townsend and Schweikert, 1989). The queueing network model for RT will adopt the same assumption, unless discussed otherwise.

Using the notations introduced in the previous section, this queueing process can be denoted as M/M/1/∞/FCFS (or M/M/1 for short), representing a queueing process with Poisson arrivals (also called exponential interarrival times), exponential service times, a single-channel server at each node, no restriction on the maximum number of customers allowed in the queue, and first-come, first-served queue discipline. Detailed discussions about the importance and justifications of employing this type of queueing process in performance modeling can be found in all standard textbooks on queueing theory (Kleinrock, 1976).

Another most commonly made assumption in the queueing network literature is that the routing probabilities are independent of the state of the system (i.e., the p_{ij}'s are independent of $N_i(t)$, i=1,..., K). For the most part of this article, the queueing network model for RT will adopt this assumption. The only class of networks discussed in this article that do not follow this assumption is a special type of queueing networks called fork-join queueing networks, which will be shown to take PERT networks as a special case.

Product-form Networks (Jackson Networks)

It turns out that networks that have these three properties (Poisson arrivals, exponential services, and state-independent routing probabilities) are the class of queueing networks that have received the most research attention and enjoyed a most

fruitful history of producing usable analytical results for applications. This class of networks are called separable networks or product-form networks. They are also called Jackson networks, named after the author who showed that this class of networks have the following amazing property (Jackson, 1957): The joint probability distribution for the number of customers at each node can be written as a product of the marginal probability distributions at each of the nodes. In other words, in terms of queueing length distribution, the network *acts as if* each node can be viewed as an independent M/M/1 queue, with parameters λi and μi, although the actual internal flow in this kind of networks is not always Poisson. More formally, we have,

$$P(N_1=n_1, ..., N_K=n_K) = \prod_{i=1}^{K} (1 - \frac{\lambda i}{\mu i}) (\frac{\lambda i}{\mu i})^{n_i} \qquad (1)$$

where

n_i is the number of customers at node i,

λi can be determined by the following equation, which is commonly referred to as the "traffic equation":

$$\lambda i = \gamma i + \sum_{j=1}^{K} (p_{ji} \lambda_j) \qquad (2)$$

This amazing property makes it possible to derive many important results for product-form networks that are often not possible to obtain or analytically intractable for other types of networks. Jackson networks have subsequently enjoyed a great success in model development and have been successfully applied to diverse areas of applications. Furthermore, numerous studies have demonstrated that many of the results for Jackson networks provide close approximations to non-Jacksonian networks. It has been pointed out that, in practical applications, inaccuracies resulting from violations of Jackson's assumption typically are not worse than those arising from other error sources such as inadequate measurement data (Boxma and Daduna, 1990; Denning and Buzen, 1978).

Reaction Time as Network Sojourn Time

From the perspective of RT analysis, the most relevant performance measure of a queueing network is a random variable called customer sojourn time--the total time for a customer to traverse portions of or the entire network. More formally, if an arbitrary customer, m, traverses nodes 1, 2, ..., J, between entering the network at node 1 and exiting at node J, then its network sojourn time, Tm, is,

$$T_m = T_{m1} + T_{m2} + ... + T_{mJ} \quad (3)$$

where $T_{mi} = W_{mi} + S_{mi}$, i=1, ..., J. T_{mi} is the sojourn time of customer m at node i, which includes the time it spends in waiting (W_{mi}) and the time it spends receiving service (S_{mi}).

In order to link customer sojourn time with reaction time, the queueing network model for RT assumes that a response is made when the response unit has accumulated M of the N signal components (M and N are usually defined arbitrarily and can be made arbitrarily close to each other). This assumption is similar to that in the accumulator model (see, e.g., Pachella, 1974) and in the queue-series model of Miller (1993). According to this assumption, total RT is the time interval between the instant of stimulus presentation and the instant at which the Mth signal component arrives at the response unit.

The model assumes that an input node in a discrete network has the function of accumulating all the independently arrived M components and then transmitting them as an "assembled package" to internal nodes. Components that arrive later than the Mth component are not allowed to enter the network while the current "package" is being processed by the network. Thus, in a discrete network, all nodes operate in strict sequence without temporal overlap of node activities. In contrast, an input node in a continuous-transmission network, like all internal nodes, transmits each customer immediately after it has received and processed it. Therefore, all nodes could operate concurrently.

To use an observable natural phenomenon as an analogy, we can imagine a discrete network as a special type of highway transportation system in which shipping materials arrive at the highway entrance independently, being assembled into one big package there and then shipped through an otherwise empty highway (empty except for the only package). Similarly, a continuous network can be imagined as a "normal" highway, where shipping materials arrive at and pass through the entrance independently and travel through the network individually like in a traffic-flow situation. As will be discussed below, in some special classes of networks (discrete PERT networks or continuous fork-join networks), the shipping materials may be disassembled into parcels after having entered the network. Each parcel may take a separate path of the network and then they are reassembled at the destination.

For most of the networks considered in this article, there is no need to distinguish noise and signal components. For discrete networks of this type, reaction time (denoted as RT_d) is apparently the sum of the time required by the input node to accumulate M components (T_1) and the time for the assembled "package" to traverse through the network (T_d), i.e.,

$$RT_d = T_1 + T_d \tag{4}$$

For Poisson arrivals, the time interval between the first and the Mth arrival to the input node (T_1) follows the ordinary gamma distribution with parameters M and λ (Ross, 1983; Townsend and Ashby, 1983), and is independent of T_d.

If the structure of a continuous-transmission network does not permit components to overtake each other, then the Mth component to depart from the network is also the Mth to arrive at the input node from the outside. Apparently, reaction time in this case (denoted as RT_c) is the sum of the time interval between the first and the Mth arrival (T_2) and the Mth customer's network sojourn time (T_c), i.e.,

$$RT_c = T_2 + T_c \tag{5}$$

It is easy to see that $T_1=T_2$, because both can be described as the same ordinary gamma distribution with parameters (M, λ). For values of M that are not too small, this distribution approximates a normal distribution. Several authors have shown that the convolution of a normal and an exponential distribution provides a close approximation to experimental data (Ashby, 1982; Hockley, 1984; Hohle, 1967; Ratcliff and Murdock, 1976). This finding may be borrowed as a tentative support of the role of T_1 and T_2.

Because $T_1=T_2$ and they are independent of T_d and T_c, respectively, in order to compare the RT behavior of a discrete and a corresponding continuous network (RT_d and RT_c), it suffices to compare T_d and T_c. To continue to use the highway system analogy, equations (4) and (5) tell us that in order to compare the time needed for shipping M pieces of materials in a discrete network (RT_d) and that in its continuous counterpart (RT_c), what is needed is to compare the sojourn time of a large package containing the M components in an empty network (T_d) with the sojourn time of one of the components (the Mth one to arrive at the network) in a crowded network (T_c).

Sojourn Time in Queueing Networks

Several decades of queueing network research has shown that determining sojourn time of a customer in queueing networks at the distributional level is a very complicated problem and among the hardest in queueing network theory. For non product-form (non-Jackson) networks, almost no explicit results exist. For product-form networks, although we know that in terms of queueing length distribution, the network *acts as if* each node can be viewed as an independent Poisson-arrival, exponential-service queue, it does not imply that the sojourn times of a customer at successive nodes are independent of each other. On the contrary, a well-established result is that the sojourn times of a customer at successive nodes are, in general, not independent of each other (Simon and Foley, 1979; Walrand and Varaiya, 1980).

Until now, exact expressions of network sojourn time distributions are usually not available, because little is known about the complicated dependencies among sojourn

times at successive nodes. An important exception to this statement is provided by the sojourn time distribution of a customer along an "overtake-free" path in a product-form network. A path is overtake-free if customers can not overtake or bypass one another on that path. As a rule of thumb, it suffices to say that a path is overtake-free if all the nodes on that path are single-channel FCFS nodes and every two nodes are connected by at most one directed path. Historically, Reich (1957, 1963) was the first to have proved that the successive sojourn times of a customer at the nodes of an overtake-free series system are independent and exponentially distributed. This result remained the state-of-the-art until the late 1970s when it was generalized by Walrand and Varaiya (1980), who proved that in all product-form networks, the sojourn times of a customer at successive nodes along an overtake-free path are independent and exponentially distributed with parameter (μi-λi) for node i (Walrand and Varaiya, 1980; Boxma and Daduna, 1990).

From elementary probability theory, we know that if X and Y are independent random variables having respective distribution functions F and G, then the distribution of X + Y is the convolution of F and G, denoted by F * G. Since the network sojourn time of a customer is the sum of its sojourn times at the successive nodes from the input node to the exit node, in an overtake-free product-form network, we have,

$$F(t) = \Pr\{T<t\} = (1- e^{-(\mu_1-\lambda_1)t}) * \ldots * (1- e^{-(\mu_J-\lambda_J)t}), \qquad t>=0 \qquad (6)$$

$$E[T] = \frac{1}{\mu_1-\lambda_1} + \ldots + \frac{1}{\mu_J-\lambda_J} \qquad (7)$$

where

F(t) is the cumulative probability distribution of network sojourn time of a customer, visiting nodes 1 to J on an overtake-free path,

E[T] is the mean of the network sojourn time,

$(1- e^{-(\mu_i-\lambda_i)t})$ is the exponential probability distribution of a customer's sojourn time at node i,

$\frac{1}{\mu_i-\lambda_i}$ is the mean sojourn time at node i,

μ_i and λ_i follow the standard interpretation introduced earlier, and λ_i satisfies the traffic equation.

If a path linking node 1 to node J in a product-form network is not overtake-free, equation (6) for sojourn time distributions will, in general, not valid. However, for the whole class of product-form networks, equation (7) for mean sojourn times still holds for any sequence of nodes. Actually, Lemoine (1987) has proved that equation (7) is a special case of the following recursive equation, which characterizes the mean sojourn time of a customer in all product-form networks:

$$E[T_i] = \frac{1}{\mu_i - \lambda_i} + \sum_{j=1}^{K} p_{ij} E[T_j] \qquad (8)$$

where

$E[T_i]$ is the expected value of the remaining network sojourn time of a customer at the instant when it arrives at node i of a network with K nodes.

Apparently, if a customer visits node 1 to node J successively, without skipping any node or visiting any node more than once, then we have $p_{ij}=1$ for i=1 to J-1 and j=2 to J, and $p_{ij}=0$ for all other values of i and j. In this case, equation (8) specializes to equation (7). Lemoine (1987) also derived the recursive relations for computing the second moment of network sojourn times, which involves more unknown variables than the number of equations. In general, exact computations of the second or higher moments of sojourn times in a product-form network are not possible without additional information about some characteristics of the network.

With these general descriptions and assumptions at hand, I am ready to examine RT behavior in several interesting classes of queueing networks. I will first compare T_c and T_d in the simplest network called tandem queues, in which nodes are arranged in sequence. Then I will examine fork-join queues to show that they include PERT networks as special cases. Tc and Td in a simple feedback queueing system will also be compared.

In the last two sections, I will discuss the characteristics of Tc in two classes of continuous-transmission queueing networks, the first of which allows noise components to overtake signal components, and the second only allows a fixed number of customers to exist in the system.

Tandem Queues as a Model for Reaction Time

Network Sojourn Time in Tandem Queues

The simplest type of queueing networks is tandem queues, also called series queues, in which the service stations form a series system with flows always in a single direction from the first node to the last node. As shown in Figure 2, customers may enter from the outside only at node 1 and depart only from the last node. More formally, assuming that external Poisson arrival to node 1 has a mean arrival rate of λ, then we have an open K-node network where

$$\gamma i \begin{cases} = \lambda \ (i=1) \\ \\ = 0 \ (\text{elsewhere}) \end{cases}$$

and

$$p_{ij} \begin{cases} = 1 & (j=i+1;\ 1 \le i \le K-1) \\ = 1 & (i=K,\ j=0) \\ = 0 & (\text{elsewhere}) \end{cases}$$

Using the traffic equation for Jackson networks introduced earlier as equation (2), it is easy to see that in a Jackson tandem queueing system, the mean arrival rate for each internal node is the same as that for the input node, i.e.,

$\lambda i = \lambda$ (for $\forall i$)

When the tandem queueing system is formed by FCFS single-channel nodes, it is impossible for the customers to overtake each other while traversing the system. Therefore, equation (6) for computing sojourn time distributions holds. Substituting λi in

equation (6) with λ for all i, we obtain the following expression for the network sojourn time of a customer, Tc, in this continuous tandem queueing system (see, e.g., Boxma and Daduna, 1990):

$$\Pr\{Tc<t\} = (1- e^{-(\mu_1^c-\lambda)t}) * \ldots * (1- e^{-(\mu_K^c-\lambda)t}), \qquad t>=0 \tag{9}$$

If we assume that the mean service rates at the K nodes are all different from each other (i.e., $\mu_i^c \neq \mu_j^c$, for $i \neq j$), then the convolution in equation (9) can be described by the general gamma distribution, and can be written more simply as:

$$\Gamma_K^c(t) = 1 - \sum_{i=1}^{K} C_{ik}\, e^{-(\mu_i^c-\lambda)t} \tag{10}$$

where

$\Gamma_K^c(t)$ is the cumulative distribution function (CDF) of T_c, and

$$C_{ik} = \prod_{j=1\ j\neq i}^{K} \frac{(\mu_j^c-\lambda)}{(\mu_j^c-\lambda) - (\mu_i^c-\lambda)} \tag{11}$$

As previous authors have noted, there is no practical need for considering the case in which a subset of the K nodes have identical mean durations, since equal values can be approximated with any degree of accuracy by using values that are almost, but not exactly , equal (see, e.g., McClelland, 1979).

Insert Figure 2 about Here

Because Tc characterizes the RT behavior of a continuous-transmission queueing network model, what we have learned here is that reaction time in a tandem queueing system can be described by general-gamma distribution. In the following I will discuss the implications of this result on RT modeling in the context of existing discrete and continuous serial models for RT.

Serial Discrete-Stage Model of McGill and Gibbon

In a serial discrete network, stimulus components are transmitted as an indivisible unit from the first node to the Kth node, and there is no temporal overlap of stage activities. The passage time for the unit to go through the K nodes is the sum of the passage time at each of the K nodes. McGill and Gibbon (1965) have shown that general gamma is the RT distribution if the passage time at each stage is exponentially distributed with parameter μ_i^d, i=1 to K. More specifically, they showed that the passage time for a serial discrete-stage model with K exponential stages with non-identical mean durations has the following form:

$$\Gamma_K^d(t) = 1 - \sum_{i=1}^{K} C_{ik}\, e^{-(\mu_i^d)t} \qquad (12)$$

where

$\frac{1}{\mu_i^d}$ is the mean duration of the exponentially distributed passage time through stage i, and $\mu_i^d \neq \mu_j^d$, for $i \neq j$, and

$$C_{ik} = \prod_{j=1\ j\neq i}^{K} \frac{\mu_j^d}{\mu_j^d - \mu_i^d} \qquad (13)$$

Apparently, $\Gamma_K^d(t)$ is the cumulative distribution function for T_d in equation (4).

The close resemblance of (10) (11) and (12) (13) is obvious, although the two sets of equations were independently formulated by different authors for continuous and discrete systems respectively. It is clear that the continuous-transmission tandem model and the serial discrete-stage model demonstrate the same RT behavior, which is characterized by the general-gamma distribution. Actually, the serial discrete stage model of McGill and Gibbon can be treated as a special case of the tandem queuing model by replacing $(\mu_i^c - \lambda)$ in equation (10) with μ_i^d. In the queueing network literature, $(\mu_i^c - \lambda)$ is often called the "effective service rate" of node i.

The major conceptual difference between the two general gamma functions is that the serial discrete stage model has the largest possible grain size of transmission, and only one "large" customer exists in the tandem network. Since no other customers are allowed to enter the network until the current one has completed processing, we have a situation in which $\lambda=0$ and thus $\mu_i^c = \mu_i^d$. In the more general case of the tandem queueing model, individual stimulus components enter and traverse the network like a traffic flow with $\lambda > 0$. In order to have $\mu_i^c - \lambda = \mu_i^d$, we must have $\mu_i^c > \mu_i^d$. This is consistent with intuition--a node serves a small customer faster than serving a large one, but the small customers in a continuous system may have to wait a long time in queue due to the presence of other customers.

To continue to use the transportation system analogy, it appears that the total time required for packaging all the independently-arrived shipping materials into a package and then shipping the package through an empty series system is identical to the corresponding situation in which all the shipping materials are allowed to enter and traverse through a "crowded" series system separately. It appears that, for the class of models we have examined here, the detection of a general gamma distribution in a set of RT data would not distinguish whether the underlying mental structure is discrete or continuous. All the conclusions and inferences about discrete models based on general gamma distributions are also applicable to the continuous tandem queueing model. The additive stages models discussed next is an example.

Additive Stages Model of Ashby and Townsend

Donders's assumption of pure insertion and Sternberg's assumption of selective influence are both assumptions at the level of the mean stage durations. Ashby and Townsend (1980) extended the two assumptions to the distributional level and supplied a set of important theorems for testing the applicability of these assumptions to RT data. Ashby and Townsend (1980) assumed that the RT process can be decomposed into a number of additive stages. Their definition of a stage is based on serial completion rather

than serial initiation or concatenation of separate processes--processes could start and be active either strictly in sequence or simultaneously in parallel, but there exists a serial order in their completion. Observable RT can be decomposed into a number of intercompletion times (ICTs), with the jth ICT defined as the time between the completion of the (j-1)th and the jth additive stages.

According to Ashby and Townsend, at the level of distribution, the assumption of pure insertion becomes

$$RT_k = RT_{k-1} + ICT_k \tag{14}$$

where RT_k is the observable RT when there are k ICTs in a task, ICT_k is the unobservable duration of the kth ICT, and RT_{k-1} is independent of ICT_k. Assuming that the kth ICT is exponentially distributed with parameter V_k, Equation (14) can be written equivalently as

$$g_k(t) = g_{k-1}(t) * V_k e^{-V_k t} \tag{15}$$

where $g_k(t)$ is the density function of RT_k.

If there are two levels of two experimental factors, A and B, and each factor influences a different stage of processing at the level of distribution, then the probability density function when factor A is at level i and factor B at level j can be written as

$$g_{A_iB_j}(t) = b(t) * f_{A_i}(t) * f_{B_j}(t) \tag{16}$$

where b(t) is the density function of all stages not influenced by either factor A or factor B. Ashby and Townsend showed that, at the distributional level, the assumption of selective influence becomes

$$g_{A_1B_1}(t) * g_{A_2B_2}(t) = g_{A_1B_2}(t) * g_{A_2B_1}(t) \tag{17}$$

It is a simple exercise to show that general gamma distributions satisfy equations (14)-(17), considering the fact that general gamma is the convolution of a number of exponential distributions. It is not surprising, of course, that the discrete serial stage model of McGill and Gibbon satisfies the assumptions of pure insertion and selective influence at the distributional level. The importance of this result is that the continuous

tandem queueing model satisfies the two assumptions as well. Ashby and Townsend have pointed out that their extension of the model "does not rule out the possibility that stages make use of partial information from preceding stages" (Ashby and Townsend, 1980, p.96). The tandem queueing model presented here is a concrete evidence in support of this statement, and indicates that if experimental factors demonstrate additive effects on reaction time, they do not imply that they must affect discrete processes.

Serial Continuous Models

McClelland's Cascade model. The cascade model of McClelland (1978) is a continuous-flow serial-processing model. The model assumes that the human information processing system functions like a series of parallel linear integrators. These linear integrators take a weighted sum of a subset of the outputs of the integrators at the preceding level and produces continuous output available for processing at the next level. A central assumption of the cascade model is that the rate of change of activation of a linear integrator unit is determined by its rate constant times the difference between the asymptotic level it is being driven to and the activation level the unit has already reached. Interestingly, as shown in McClelland's derivation and in standard electronics texts, the type of units that satisfy this assumption behave like exponential servers.

The heart of the cascade model is a cascade equation, which gives an expression for the activation of linear integrator l at processing level K to a stimulus S presented at time t = 0. The equation has the following form:

$$a_{Kl/S}(t) = a_{Kl/S}\left(1 - \sum_{i=1}^{K} C_i e^{-(\mu_i)t}\right)$$

$$= a_{Kl/S}\,\Gamma_K[t] \qquad (18)$$

where

μ_i is the rate constant of a linear integrator at level i, and $\mu_i \neq \mu_j$, for $i \neq j$;

$$C_i = \prod_{j=1}^{K} \prod_{j \neq i} \frac{\mu_j}{\mu_j - \mu_i} \qquad (19)$$

There are two independent terms on the right hand side of equation (18). The first term, $a_{Kl/S}$, does not vary over time. It represents the asymptotic activation value of linear integrator l at level K that would result as time goes to infinity. The second term, $\Gamma_K[t]$, is a dynamic term that characterizes how the activations of units at level K vary with time. It can be easily seen that $\Gamma_K[t]$ changes from 0 to 1 as time t goes from 0 to infinity. Correspondingly, $a_{Kl/S}(t)$ approaches the asymptotic level of $a_{Kl/S}$ as time goes to infinity. It is assumed that subjects adopt a response criterion level, and a response occurs when the activation level at time t exceeds the criterion. According to the cascade model, all of the units at the same processing level have the same activation function.

The appearance of a general-gamma function in equation (18) as the dynamic term is not surprising, considering the fact that the units at each level behave like exponential servers. The function of the cascade system is, in essence, similar to a series of independent exponential servers with overlapping service durations. In fact, the tandem queueing model could mimic the cascade model precisely by making an assumption about how the response unit located at the end of the tandem queues works. Instead of assuming that the response unit functions like a binary unit that initiates a response unconditionally if and only if it has accumulated M signal components, a modified tandem queueing model could assume that the response unit has a time-varying response-activation strength, which at time t, is the product of the probability that it has received the Mth signal component and the activation value of the Mth signal component when it has arrived at the response unit. Analogous to the cascade model, the modified model assumes that a response is made when the response activation strength of the response unit exceeds the activation criterion set by a subject.

Since the probability that the Mth signal component has passed through K nodes and arrived at the response unit at time t is exactly what is expressed by $\Gamma_k^c(t)$ of equation (10), we have,

$$a(t) = a\,\Gamma_k^c(t) \qquad (20)$$

where

a(t) is the response activation strength of the response unit at time t,

a is the response activation value of the Mth signal component when it has arrived at the response unit.

Comparing $\Gamma_K[t]$ in equation (18) with $\Gamma_k^c(t)$ of equation (10), we can see that the only difference between the two is in the rate constants. If we use the "effective service rates " of the tandem queues as the rate constants of the cascade model, the tandem queueing model could mimic the behavior of the cascade model accurately, and all the conclusions and inferences of the cascade model are applicable to the modified tandem queueing model.

McClelland (1978) examined the effects of manipulating rate constants and asymptotic activation levels on RT and derived a set of predictions for RT behavior. Similar to the serial discrete-stage models, the cascade model shows that experimental factors affecting the rate of the same process will interact, whereas those affecting the rates of different processes are additive. However, the predictions become more complicated when at least one of the experimental factors affects the asymptotic level of activation. For example, two factors would interact if one affects the rate of the slowest process and the other affects the asymptotic activation level.

Another interesting result of the cascade model is that the model is able to fit the shape of the well-known time-accuracy curve closely. The tandem queueing model could mimic this result as follows. Analogous to the cascade model, we assume that in yes/no experiments, the response activation value of the Mth signal component when it arrives at

the response unit is $a_{y/y}$ if it is from a stimulus that is appropriate for a yes response, but is $a_{y/n}$ if it is from a stimulus that is appropriate for a no response. To be consistent with the cascade model, we assume that actual response execution is a discrete event that adds the duration of a single discrete stage (e.g., 0.1 sec.) to the time between the stimulus presentation and the registration of the overt response. Then, following the same steps of derivations in McClelland (1978, p.327), it can be shown that for the tandem queueing model, the observed value of d' at time t is given by

$$d'(t) = (a_{y/y} - a_{y/n}) \frac{\Gamma_n[t - .1]}{\{1+\sigma^2(\Gamma_n[t - .1])^2\}^{1/2}} \qquad (21)$$

Equation (21) is identical to equation (13) of McClelland (1979, p. 298), which has been shown to fit the time-accuracy curve closely.

It should be emphasized here that although the modified tandem queueing model could mimic the behavior of the cascade model, the two models have a fundamental difference in their interpretations of the general-gamma function. In the cascade model, the general gamma function, $\Gamma_K(t)$, is an activation function that represents the relative activation of a unit at level K at time t. In the tandem queueing model, the same function represents the probability that the Mth signal component has passed through the Kth node of the network at time t.

Of course, while attempting to mimic the cascade model, the modified tandem model would also expose itself to the same "infinite-RT" problem of the cascade model, pointed out by Ashby (1982). Ashby has pointed out that the cascade model always predicts a nonzero probability that a response never occurs, because the activations may never exceed their criterion level on some of the experimental trials. The lower the asymptotic activation level relative to the response criterion, the more likely that a response will never happen. This prediction of the cascade model is clearly inconsistent with reality in simple RT experiments in which no incorrect responses are made. Clearly, the modified tandem model suffers from the same problem, which can be dealt with by following the same corrective steps discussed in Ashby (1982) for the cascade model.

It should be noted here, however, that discrete-stage models as characterized by general gamma distributions also have an "infinite-RT" problem, because $\Gamma^{d}_{K}(t)$ approaches but is always smaller than 1 as t goes to infinity. $\Gamma^{d}_{K}(t)$ is the probability that the Kth discrete exponential process in the process chain has completed by time t. Or equivalently, $\Gamma^{d}_{K}(t)$ is the probability that the total duration of the first K discrete exponential processes is smaller than or equal to t. Because $\Gamma^{d}_{K}(t)$ is smaller than 1 for t smaller than infinity, the general-gamma model also predicts a nonzero probability for infinite RT. In essence, the discrete-stage model and the unmodified tandem queueing model--both are characterized by a general-gamma distribution-- could be regarded as having an implicit assumption that their response units have an infinite asymptotic activation level, which is assumed to be finite for the response units of the cascade model and the modified tandem model. Therefore, the "infinite-RT" problem becomes worse for the latter two models because the finite difference between the asymptotic activation level and response criterion provides an additional cause for infinite RT.

Miller's queue series model. The tandem queueing model that I am considering in this article is one of many possible types of tandem queues, and the reason that I have selected this particular one for discussion is because it is the only one for which analytical results are available for customer sojourn times and the assumptions are consistent with those most commonly made in the literature. But undoubtedly, other types of tandem queues exist and could make different predictions for RT behavior. One excellent example is provided by Miller in his queue-series model for RT presented recently (Miller, 1993).

In Miller's queue-series model, the stimulus is regarded as consisting of a number, M, of distinct components, and they are serviced by a series of processing nodes, each of which functions as a queue. The queue-series model assumes that customers arrive at the first node of the queue series at the same time, and they are processed either serially or in

parallel there. The second node, like all subsequent ones, is a single-channel node, and it processes components one at a time, not necessarily in the same order as they join the queue (not FCFS). The queue-series model considers discrete-stage models, continuous-flow models and intermediate models as special cases, corresponding to the cases in which $M=1$, $M=\infty$, and $1<M<\infty$, respectively.

Miller evaluated the behavior of the queue-series model through a novel application of the PERT method. The queue-series model was represented as a PERT network consisting of $M \times N$ separate processes--M is the total number of components that constitute a stimulus and N is the total number of nodes in a queue-series. The passage of the jth component through the ith node is represented as a PERT process with a duration of t_{ij} (i=1 to N; j=1 to M). The ith node cannot begin processing its jth component until it has finished processing its (j-1)th components and the (i-1)th node has finished processing at least j components. These temporal contingencies are naturally represented as unidirectional paths from node to node in a PERT network, and the path with the largest sum of t_{ij}s is called the critical path. The length of the critical path is the RT for the task.

Miller performed extensive numerical simulations of the PERT network representation of the queue-series model to examine how experimental manipulations might affect the time for M customers to traverse through the queue-series and concluded that, within the class of queue-series models he considered, experimental factors affecting different processing stages always have additive effects on reaction time with discrete stages but rarely do so with overlapping stages, and thus, observations of factor additivity support discrete-stage models. Nondiscrete queue-series models were shown to be more likely to produce underadditive factor effects on RT.

Apparently, the conclusions from Miller's queue-series model is different from those of the tandem queueing model, which has been shown to be able to mimic McGill and Gibbon's discrete serial model and McClelland's cascade model closely, both of

which could produce additive effects. The tandem queueing model also satisfies Ashby and Townsend's (1980) assumptions of pure insertion and selective influence at the distributional level. Since no analytical result is available in the queueing network literature about sojourn times in the type of queue-series considered in Miller's model, it is difficult to identify exactly why the two models behave differently. But it appears that at least part of the explanation is offered by Miller in his discussion of the relationship between the queue-series model and the cascade model.

Miller has pointed out that although the queue-series model is able to approximate the shape of the activation functions of the cascade model by increasing the value of M, the two models produce different effects on RT. The explanation was that the cascade model allows experimental factors to have downstream effects, whereas the queueing series model does not consider such propagation. More specifically, the queue-series model has two explicit restrictions about the effects of experimental manipulations on critical path membership for the t_{ij}s. The first is that at most one experimental factor influences whether each t_{ij} is on the critical path, and the second is that when a t_{ij} is influenced by an experimental factor, only that factors may determine whether the t_{ij} is on the critical path. As Miller pointed out, these restrictions essentially require that a factor affecting an earlier stage must not have non-local, "downstream" influence on critical path membership for the tijs of a stage affected by another factor. It is possible that this could at least partially explain why the predictions of the cascade model and the tandem queueing models converge, while both diverge from the queue-series model. Experimental manipulations only change the mean process durations (i.e., t_{ij}s) in the queue-series model, but both the tandem queueing model and the cascade model allow experimental manipulations to change process durations at the distributional level.

Future research will undoubtedly offer deeper insight about why the two models make significantly different predictions for RT behavior. At a more general level, the present discussion of the two classes of series queueing systems clearly is a

demonstration of the diversity of possible queueing models, both in their structures and assumptions and in their predictions of RT behavior. Queueing network theories and methods can be regarded as a unified conceptual framework as well as a set of mathematical tools for developing and evaluating these models.

Mean RT in Fork-Join and Feedback Queueing Networks

A tandem queueing system is the simplest type of queueing network, in which all customers visit the same sequence of service nodes and each node is visited by every customer exactly once. Non-tandem network configurations arise when customers have more complicated service requirements. For example, when the service requirements of a customer do not have to be processed in strict sequence, a customer may split into several new customers, each of which takes a separate path so that the customer's service requirements can be processed by separate parts of the network in parallel. When all customers do not have identical service requirements, they do not necessarily take the same path of a network, and they may return to nodes previously visited or skip some nodes entirely.

In this section I will discuss two classes of network configurations. I will first consider what happens if a customer splits itself in a network. The reason that I discuss this class of networks first is because they are closely related to the existing network models of psychological processes. Then I will examine RT behavior in a simple feedback queueing system in which a customer may visit a node several times before it departs from a system. It turns out that for both classes of networks, it appears that continuous-transmission models are not distinguishable from their discrete counterparts at the level of mean RTs. At the distributional level, no conclusions can be drawn for fork-join networks. For the feedback system, however, discrete and continuous models make different predictions about RT distributions.

Fork-Join Queueing Networks

As described earlier in this article, psychologists have challenged the notion of serial arrangement of mental processes and examined the possibility of network configurations of process activities. The current state of knowledge in this area is represented by the class of PERT network models, originally proposed by Schweikert (1978) and further developed by subsequent studies (e.g., Fisher and Goldstein, 1983; Schweikert and Townsend, 1989; Townsend and Schweikert, 1989). In this part of the article, I will show that PERT networks can be treated as a special case of a class of queueing networks called fork-join networks in the same way that the serial discrete stage models can be treated as a special case of the tandem queueing model.

Fork-join networks arise naturally when the service requirements of a customer do not have to be processed in strict sequence (Baccelli and Makowski, 1990). In addition to service nodes, which exist in all queueing networks, there are two special types of nodes in a fork-join queueing network: fork nodes and join nodes. A customer arriving at a "fork" node splits into several new customers, which are sent to separate service nodes, and the corresponding join occurs at a "join" node when the services of all these new customers are completed. These new customers themselves may also be forked and joined while traversing the network. The multiple arcs emanating from a fork node or entering a join node represent simultaneous creation or synchronized merging of multiple customers. The function of a fork node is to divide the customer's service requirements into subsets of demands that can be serviced by separate parts of the network in parallel, while the function of a join node is to ensure that all the service requirements of a customer are met before it departs from the system.

As a continuous-transmission system in its general form, a fork-join network allows customers to enter, traverse, and leave the network separately like a flow. Apparently, if a fork-join network behaves in a discrete way and processes only one customer or its offsprings at a time, it becomes a PERT network. As models for RT, the

relationship between a PERT model and a corresponding fork-join model is similar to that between a serial discrete-stage model and a corresponding tandem queueing model. A PERT model assumes that all the components of a stimulus are tied together in a single unit before entering the network, although the unit may be forked and joined while traversing the network. A response is made when the unit departs from the network. A corresponding continuous-transmission fork-join network would allow individual stimulus components to enter and leave the system separately and they would be forked and joined in the same way as in the corresponding PERT network. A response is initiated when M components have departed the system. Similar to the acyclic characteristic of PERT networks, fork-join networks do not allow a customer to visit the same node more than once, and they are often called acyclic fork-join queueing networks (AFJQNs) in the literature.

The simplest instance of a non-trivial fork-join network is a parallel network consisting of a number, K, of parallel queueing systems, as shown in Figure 3. Customers arrive at the fork node as a Poisson flow with mean arrival rate λ, and upon arrival, a customer forks into K offsprings. The ith offspring is assigned to the ith queueing system which consists of a single-channel FCFS service node and an infinite capacity queue. The service times of the nodes are independent and exponentially distributed with mean $1/\mu_i$ for node i. A customer leaves the system as soon as all its K offsprings have completed their service and are merged at the join node. The network sojourn time, Tc, of any arbitrarily selected customer is the maximum of the sojourn times of its K offsprings, i.e.,

$$T_C = \max(T_1, T_2, ..., T_K) \tag{22}$$

where

$T_j = S_j + W_j$, is the sojourn time of the jth offspring of the customer at queue j (j=1,...K), including both service time (S_j) and waiting time (W_j).

In the extreme case in which only one customer is allowed to enter the system through the fork node, we have a corresponding parallel PERT network. All the

offsprings of the admitted customer are processed immediately at the servers, and thus the T_j's only include service times (i.e., $T_j=S_j$, for all j), which are usually referred to as process durations in PERT terms and are commonly assumed to be independent random variables. The problem of determining customer sojourn time in this discrete network, T_d, is that of finding the maximum of K independent random variables (referred to as determining the length of the critical path). Unfortunately, the problem becomes more difficult for fork-join networks, because the continuous nature of customer arrival makes it necessary to consider the queueing effects at the K service nodes. T_j's are no longer service times but sojourn times--the sum of service times and waiting times. Determining the network sojourn time, T_c, becomes that of finding the maximum of K random variables that are not necessarily independent of each other.

Insert Figure 3 about Here

Recently, Nelson and Tantawi (1988) proved that the T_j's in this simple parallel fork-join system, j=1, ..., K, are associated random variables. Random variables T_1, T_2, ..., T_K are said to be associated if $\text{cov}[f(T_1, T_2, ..., T_K), g(T_1, T_2, ..., T_K)] \geq 0$ for all pairs of increasing functions of f and g. The properties of associated random variables that are relevant to the present discussion are: All independent random variables are associated, but associated variables are not necessarily independent. If T_1, T_2, ..., T_K are associated, then

$$P[\max_{1\leq i\leq K} T_i > t] \leq 1 - \prod_{i=1}^{K} P[T_i \leq t] \tag{23}$$

and the expected value has an upper bound expressed as,

$$E[\max_{1\leq i\leq K} T_i < t] \leq \int_0^{\infty} (1 - \prod_{i=1}^{K} P[T_i \leq t])dt \tag{24}$$

In (23) and (24), equality holds if and only if T_1, T_2, ..., T_K are independent.

For a parallel system consisting of exponential servers, (24) specializes to,

$$E[Tc] = E[\max_{1\le i\le k} T_i < t] \le \int_0^{\infty} (1 - \prod_{i=1}^{K}(1-e^{-(\mu_i-\lambda)t})dt \qquad (25)$$

In the case of K identical servers, (25) becomes

$$E[Tc] = E[\max_{1\le i\le k} T_i < t] \le \int_0^{\infty} (1-(1-e^{-(\mu-\lambda)t})^k)dt = \frac{1}{\mu-\lambda} H_K \qquad (26)$$

where,

Hk is the harmonic series: $H_K = \sum_{i=1}^{K} \frac{1}{i}$

As Nelson and Tantawi (1988) pointed out, the lower bound for E[Tc] is obtained by ignoring queueing effects. Let $\lambda=0$, we have:

$$\frac{1}{\mu} H_K \le E[T] \le \frac{1}{\mu-\lambda} H_K \qquad (27)$$

Baccelli and Makowski (1990) later generalized this result to include customer arrivals that are not necessarily Poisson, and showed that as long as the parallel servers are identical and exponential, the following expression holds,

$$\frac{1}{a} H_K \le E[T] \le \frac{1}{b} H_K \qquad (28)$$

where a and b are uniquely determined by the rate of exponential service (μ) and the probability distribution of customer arrival. For Poisson arrivals, $a=\mu$.

As pointed out by these authors, since both bounds grow at the same rate H_K, $E[T_c]$ itself must grow at the same rate. An interesting property of the harmonic series is that H_K approximates logK for large K, which implies that mean customer sojourn time grows logrithmically in the number of parallel servers (denoted as O(logK)).

Since the extreme case of $\lambda=0$ corresponds to a PERT network of K parallel exponential processes with identical mean durations, it is not surprising that the lower bound obtained in (27) has the same form as that derived by Hartley and Wortham (1966) for stochastic PERT networks and that by Townsend (1972) for discrete parallel processes. This result suggests that the continuous-transmission fork-join network and the

corresponding discrete PERT network demonstrate the same O(logK) behavior, and thus it is impossible to distinguish whether the underlying mental network for a logarithmic relationship between mean RT and an experimental factor is discrete or continuous.

An classic example in the psychological literature that shows a logarithmic relationship between mean RT and an independent variable is the famous Hick-Hyman Law of choice reaction time. Hick (1952) and Hyman (1953) independently discovered that choice reaction time increased logrithmically with the number of stimulus-response alternatives and can be expressed by the equation $RT=a+b\log_2 K$, where a and b are constants, and K is the number of stimulus-response alternatives. Apparently, this RT relation could be explained equally well by a discrete parallel PERT model with K identical processes or a continuous fork-join model with K identical servers.

When the service rates of the exponential servers are non-identical, the expected sojourn time can be computed similarly with equation (25). Use the simplest case of two parallel servers as an example, we have,

$$\begin{aligned} E[Tc] &= E[\max_{1\le i\le k} T_i < t] \le \int_0^\infty (1-(1-e^{-(\mu 1-\lambda)t})(1-e^{-(\mu 2-\lambda)t}))dt \\ &= \int_0^\infty (e^{-(\mu 1-\lambda)t} + e^{-(\mu 2-\lambda)t}) - e^{-[(\mu 1-\lambda)+(\mu 2-\lambda)]t})dt \\ &= 1/(\mu 1-\lambda) + 1/(\mu 2-\lambda) - 1/[(\mu 1-\lambda) + (\mu 2-\lambda)] \end{aligned} \quad (29)$$

Let $\lambda=0$, we have,

$$E[T_C] \ge \frac{1}{\mu 1} + \frac{1}{\mu 2} - \frac{1}{\mu 1+\mu 2} \quad (30)$$

The right-hand side of expression (30) is identical to what was obtained in Townsend and Ashby (1983). Through some simple algebraic manipulations, (30) can be written equivalently as,

$$E[T_C] \ge \frac{1}{\mu 1+\mu 2} + \frac{\mu 1}{\mu 2(\mu 1+\mu 2)} + \frac{\mu 2}{\mu 1(\mu 1+\mu 2)} \quad (31)$$

which is the same as the result obtained by Townsend (1974) and that by Fisher and Goldstein (1983) using Order-of-Processing Diagrams.

Although the exact expression for $E[T_c]$ is still unknown, the similarity between the expressions for the upper- and the lower- bounds suggests that changes in $E[T_c]$ as a function of μ_1 and μ_2 must follow the same trend defined by the similarly-shaped upper and lower bounds. Therefore, if all we know is that RT data collected in a behavioral experiment can be described by the right-hand side of (30), then the data does not seem to be sufficient for distinguishing whether the underlying mental structure is discrete or continuous.

An interesting question is whether factorial experiments would allow us to distinguish the two classes of models. Townsend and Ashby (1983) have shown that if the durations of two parallel PERT processes (say, T_1 and T_2) are independent and if experimental factors A and B separately and monotonically affects T_1 and T_2 respectively, then manipulations of A and B will always produce underadditive interactions. A natural question is whether a corresponding parallel fork-join network would show the same pattern of underadditive interactions.

Unfortunately, an answer to this question relies on a clear understanding of several unsolved issues. All we know now is that sojourn times in a parallel fork-join network are associated. It is still not clear whether or not and under what conditions the sojourn times are independent. Suppose future research will show that they are independent under certain conditions, then a parallel fork-join model satisfying these conditions should be able to mimic the corresponding PERT model accurately. However, the problem remains even if the sojourn times are proved to be dependent, because there is another set of questions that need to be answered. For example, will a fork-join network be able to approximate the PERT counterpart to any desired level of accuracy? Furthermore, what Townsend and Ashby (1983) have shown is that independent parallel PERT processes must predict RT underadditivity, but this does not necessarily imply that RT underadditivity can only be observed when the processes are independent.

In fact, Townsend and Schweikert (1989, Schweikert and Townsend, 1989) later generalized this result on RT underadditivity to include a broader range of PERT networks, where independence is not required as long as the postulate of selective influence is satisfied. This postulate is stated in the form of a set of conditions that must be satisfied by the density functions of process durations and slacks. This set of conditions effectively preclude considerations of networks in which an experimental factor directly affecting a process is able to exert an indirect influence on another process through stochastic dependencies among the processes--referred to as an "indirect nonselective influence"(Townsend, 1984). As will be discussed in the next two sections, this "indirect nonselective influence" is an important characteristic of many queueing networks without fork and join nodes. However, it is not clear whether the synchronization constraints induced by the forks and joins are able to enforce fork-join networks to follow the postulate of selective influence.

Compared to non-fork-join queueing networks and stochastic PERT networks, which have been topics of much research since the late 1950s (Jackson, 1957) and the early 1960s (Fulkerson, 1962), fork-join networks represent one of the new and most difficult areas in queueing network research (Baccelli and Makowski, 1990; Nelson and Tantawi, 1988). Fork-join networks belong to the class of non-product-form networks, for which very little result is available in the literature and every problem becomes computationally hard. The present discussion illustrates that, at least at the conceptual level, PERT networks can be treated as a special case of fork-join networks. Furthermore, in the special case in which the networks are formed by identical parallel exponential nodes, the two classes of networks are able to mimic each other accurately at the level of mean RTs. For network with more complicated structures, this article has raised more questions than it has answered. Future research may find it worthwhile to further examine the equivalence and the identifiability of the two classes of networks.

A Single-Server Feedback Queueing System

The network in Figure 4 is a single server queueing system with instantaneous Bernoulli feedback. Customers arrive at the system in accordance with a Poisson process with mean arrival rate of γ. The server is a single-channel FCFS exponential server with rate parameter μ and an infinite queue capacity. After receiving service each customer may immediately return to the end of the queue in front of the server with probability p or depart the system with probability q=1-p. The feedback probability is independent of the state of the system.

Since this feedback system satisfies all the three Jackson assumptions, it belongs to the class of product-form networks--the joint probability distribution of the number of customers being in their first, second, ..., and Kth loop has a product form. However, this feedback system is not overtake-free, and the order of customer arrival is not preserved in the order of their departure from the system. Because customers may overtake each other while traversing the system, the sojourn time of an arbitrary customer is not only influenced by the number of customers (and their remaining service requests) found upon its arrival, but also by later arrivals. Thus, the sojourn times of a customer's successive visits at the server are not independent of each other.

Insert Figure 4 about Here

Takacs (1963) was the first to have examined this feedback system and derived an exact expression for the mean network sojourn time of a customer, $E[T_C]$, as follows:

$$E[T_C] = \frac{1}{q\mu-\gamma} \qquad (32)$$

This expression can also be derived from equation (8) directly as follows (Lemoine, 1987). Because each customer is expected to visit the server 1/q times, the total arrival rate at the server, λ, is γ/q (including both external and feedback arrivals). Therefore, according to equation (8), the expected network sojourn time can be computed as:

$$E[T] = \frac{1}{q}\,\frac{1}{\mu - \frac{\gamma}{q}} = \frac{1}{q\mu - \gamma}$$

As described earlier, RT is characterized by the time for the response unit to accumulate M components. Of great interest to RT modeling, therefore, is the sojourn time of the customer who is the Mth to depart from the system. However, queueing network research investigates the issue of customer sojourn times from the perspective of arrivals--How long does it take for the Mth arrival rather than the Mth departure to traverse through the system. This difference in perspective does not pose a problem when a network is overtake-free, as in the case of tandem queues and fork-join networks discussed thus far, because the Mth departure is also the Mth arrival. But this simple relation between arrival and departure does not hold for the feedback queueing system, because of customer overtaking. An important result in this regard is that of Whitt (1984), who has proved that in this feedback system the expected number of customers that overtake a particular customer is the same as the expected number of customers that are overtaken by this customer. This result implies that although on a particular trial of an RT experiment, the Mth stimulus component to arrive at the system is not necessarily the Mth to depart, over a large number of repeated trials, the Mth arrival is still expected to be the Mth to depart. Therefore, in a typical RT experiment involving a large number of trials, the network sojourn time of the arbitrarily selected Mth arrival as expressed in equation (32) can still be used to infer the RT behavior of this feedback queueing system.

What is particularly interesting about equation (32) to RT modeling is that it tells us that, at the level of the mean RT, this feedback system is able to mimic a serial system with N identical exponential servers accurately. In equation (32), if we let $q=1/N$ and $\gamma=\lambda/N$, with N take integer values, we have

$$E[T] = \frac{1}{q\mu - \gamma} = \frac{N}{\mu - \lambda} = \sum_{i=1}^{N} \frac{1}{\mu - \lambda} = kN \qquad (33)$$

Clearly, $E[T] = \sum_{i=1}^{N} \frac{1}{\mu-\lambda}$ in (33) is the expression for the expected sojourn time of a customer in a system of a series of N identical exponential servers with parameters μ and λ for each server. Furthermore, E[T] = kN in (33) tells us that the detection of a linear relationship between mean RT and N in a set of RT data is not sufficient to distinguish whether the underlying mental system is consisted of a sequence of N identical servers or of a single server with feedback probability of $\frac{1}{N}$.

In psychological experiments a linear relationship between mean RT and a discrete independent variable has traditionally been interpreted as an evidence in support of a serial stage model. A classic example is Sternberg's memory scanning task, in which the subjects are asked to remember a list of items (called "positive set") and then to make a yes/no type of binary response about whether a displayed item is or is not a member of the positive set. A large number of studies using this experimental paradigm have shown a robust linear relationship between mean RT and the size of the positive set. The slope of this linear relation is interpreted as the duration of a new stage inserted in the processing chain when the size of the positive set increases by one. Townsend (1974) has shown that a system of identical and independent parallel processes could predict the same linear relationship with arbitrary slope k by assuming that the processing rate of the parallel processes decreases as the number of items increases. More specifically, Townsend showed that since $E[T] = \frac{1}{\mu} \sum_{i=1}^{N} \frac{1}{i}$ in parallel systems, E[T] = kN could be obtained if $\mu = \frac{\sum_{i=1}^{N} \frac{1}{i}}{kN}$. Using the results discussed earlier in this section about fork-join networks, it is easy to see that a corresponding continuous-transmission parallel fork-join network could predict the same linear relation equally well. As Townsend pointed out, this

interpretation based on parallel systems is not necessarily intuitive or natural, particularly considering the complicated relation between μ and N.

The single-node feedback system offers us another plausible explanation of the linear RT relation. The effect of adding a new item to the positive set may very well be that of decreasing the departure probability q in this feedback system rather than that of inserting an additional stage. Since q is related to set size N in a simple reciprocal relation ($q=\frac{1}{N}$), this interpretation does not appear to be unnatural. A single-node system with a feedback loop appears to be perhaps more parsimonious as a model for RT than a chain of N nodes, particularly when N is large.

If the feedback system is a discrete system, then it seems impossible to distinguish, even at the distributional level, whether a process visits the same node N times or visits N identical nodes in series, because both can be characterized by the same ordinary gamma distribution with parameter (N, μ). However, for continuous-transmission systems, the two classes of systems dissociate in their predictions of sojourn time at levels higher than the means. Takacs (1963) has derived an exact expression for the Laplace-Stieltjes transform of the network sojourn time distribution, and its form is far more complex than that of the ordinary gamma distributions. Furthermore, the existence of customer overtaking in the feedback system tend to produce a greater variance in network sojourn time than the series system (Takacs, 1963; Lemoine, 1987). Therefore, detection of large RT variances in conjunction with a linear mean RT relationship appears to be an evidence in favor of the continuous-transmission feedback model over a series model, although not necessarily a definitive evidence.

Simon-Foley Network and Overadditive Factor Interactions

If a Jackson network does not allow customers to overtake each other, then sojourn times of a customer at successive nodes are mutually independent and exponentially distributed random variables and network sojourn time can be described as

general gamma distributions, which have been shown to play a central role in the McGill and Gibbon's model, the cascade model, and the tandem queueing model. In this section I will discuss a classic example of a non-overtake-free network shown in Figure 5. This network is often referred to as a Simon-Foley network, which is a three-node Jackson network with a single server at each node (Simon and Foley, 1979). Customers only enter the system at node 1 and exit the system at node 3. After visiting node 1 a customer goes directly to node 3 with probability (1-p), or goes to node 2 and node 3 in sequence with probability p. We may also think of this system as having two types of customers. Type 1 customers take the indirect route, whereas type 2 customers take the direct route. The two types of customers have identical service requirements and priority level at the nodes they visit, and the value of p decides the proportion of type 1 customers in the total customer population.

This network has an interesting property: the sojourn time in the first and the third queue (T_1 and T_3) are not independent for those customers who go through the second queue, but they are independent for those customers who go directly from node 1 to node 3. T_1 and T_2 are independent. T_2 and T_3 are also independent. Recent research has shown that T_3 is stochastically increasing in T_1 for a customer that goes through node 2, i.e., $P\{T_3 > t|T_1\}$ is increasing in T_1 (Foley and Kiessler, 1989). A result that is particularly useful for mean sojourn time analysis is derived by Walrand and Varaiya (1980), who showed that the expected value of T_3 increases as T_1 increases. That is,

$$E[T_3|T_1=t'] > E[T_3|T_1=t], \quad t'>t>0 \tag{34}$$

where $E[T]$ represents the mean of T.

This relationship has a quite intuitive interpretation. Let A be a customer who goes to queue 3 via queue 2 after leaving queue 1. Some customers who arrived and departed from queue 1 later than A may arrive at queue 3 before A arrives there because they have taken a direct route. The longer A had spent at queue 1, the more likely that A had left a long queue waiting behind it there, and the more likely that many of these late

arrivals would have arrived at queue 3 earlier than A because they took the direct route. Thus, the longer A had stayed at queue 1, the longer A would have to wait at queue 3. Clearly, this dependence between T_3 and T_1 can be found if and only if $0<p<1$ (i.e., both the direct and the indirect routes exist) and only for those customers who take the indirect route.

Insert Figure 5 about Here

These results can be used to offer a new and interesting explanation for overadditive interactions observed in some factorial experiments. Let us examine what happens when one factor affects T_1 and another factor determines whether p equals 1 or takes an intermediate value between 0 and 1. When p equals 1, there is only one type of stimulus components, which can be regarded as signal components and they all take the indirect route from node 1 to node 3. In this case, the network is overtake-free and the sojourn times of a customer at the three nodes are all independent of each other. When $0<p<1$, some stimulus components take the direct route and others take the indirect route. We may assume that those taking the direct route are noise components and those taking the indirect route are signal components. In this case, noise components could overtake signal components but signal components never overtake each other. In both cases, the Mth signal component to arrive at the system is also the Mth to depart, so its network sojourn time characterizes the RT behavior of this task network. Comparing the two cases, it is clear that the effect of increasing T_1 on RT would be greater when p takes intermediate values than when p equals 1, since increasing T_1 produces a corresponding increase in T_3 only when it is between 0 and 1. Therefore, when two experimental factors affect T_1 and p respectively, their joint effects on RT would be an interaction of the overadditive type.

As an example, let us examine the results of the lexical decision study of Meyer, Schvaneldt and Ruddy (1975). This study has also been used by McClelland (1979) throughout his article to illustrate the differences between the predictions of the cascade model and the discrete-stage model in their explanations of RT behavior. Meyer and his colleagues studied reaction time to decide whether a string of letters was a word or a nonword and manipulated the visual quality of the display of the target word (with or without "noise dots" superimposed on the letters). The factor of visual degrading was manipulated in factorial combination with the relatedness of a preceding context word to the target. The result was an overadditive interaction of the quality and relatedness factors: Responding to associated words was 38 msec faster than to unassociated words in the intact condition, and this RT difference was increased to 71 msec in the degraded condition. Subjects responded to intact associated words 129 msec faster than to degraded associated words. These authors suggested that, according to the discrete stage model, visual degrading by dots and context priming have their effects on a common stage. According to the predictions of the cascade model, McClelland suggested an alternative explanation in that the interaction could be due to the joint effects of two factors each affecting asymptotic activation or to the joint effects of a factor affecting the asymptotic activation and another affecting the rate of the rate-limiting process.

In terms of the Simon-Foley network, I can offer another alternative explanation which is interesting and intuitive. It seems reasonable to assume that there are two types of stimulus components--letters are formed by signal components and the dots are noise components. Both types of components must go through node 1 and node 3. After leaving node 1, signal components must go through node 2 for some type of analysis that is not required for noise components. Noise components are "filtered" out by node 1 and they all go to node 3 directly from node 2. I further assume that associated words are served faster at node 1 than unassociated words, but their sojourn times at node 2 are identical.

In the intact condition, noise components do not exist, and all customers go from node 1 to node 2 and then to node 3. In this case, time requirements at the three nodes (T_1, T_2, and T_3) are all independent. The 38 msec difference between unassociated and associated words reflects the difference in T_1 for the two priming conditions. In the degraded condition, the presence of noise components will cause an increase in T_1 and T_3 because signal components must compete for service with noise components at the two nodes. This is reflected by the 129 msec RT increase in RT that is shared by both associated and unassociated words. More importantly, the presence of a direct route taken by noise components produces a positive dependence of T_3 on T_1, and the increase in T_3 is expected to be larger for unassociated words (with a longer T_1) than for associated words, as seen in the 33 msec increase in the RT difference between the two types of words in the degraded condition (71-38=33).

Another example that can be used to further illustrate this point is the result reported by Miller (1976). Miller found that degrading by dots produced an overadditive interaction with the probability of stimulus occurrence, but degrading by contrast reduction had additive effects with the probability manipulation. In terms of the queueing network model, this result could be because degrading by dots creates "noise" customers that overtake signal components, whereas degrading by contrast reduction do not. Therefore, only degrading by dots will destroy the independence of T_1 and T_3 and produce an overadditive interaction between degrading-by-dots and occurrence probability.

The above discussion has suggested a plausible explanation to overadditive interactions discovered in psychological experiments such as the lexical decision tasks, in addition to that offered by Sternberg based on serial discrete stages and that by McClelland based on cascade processes. It should be noted here that the presence of an overadditive interaction does not confirm the presence of a network shown in Figure 5, just as it does not exclusively confirm the hypotheses of Sternberg or McClelland.

For the purpose of detecting the presence of a network arrangement of Figure 5, there does exist a test that is stronger than detecting the presence of interactions. This test is based on equation (34), and is provided by directly measuring the durations of T_3 and T_1 if related measurement methods are assumed to be available. This assumption is not stronger than those for testing the validity of the Schweikert's PERT methodology for RT analysis, which assume that we are able to prolong the duration of a process of interest, and "we may be able to record time at several points in the network. We may know the times at which various stimuli are presented and responses made, and we may also know the times at which various physiological events occur" (Schweikert, 1978; p. 123). According to Equation 34, if in a task situation in which prolonging a process produces a corresponding increase in the duration of another process, but not vice versa, then there is a great possibility that the task situation involves a continuous network of mental processes shown in Figure 5, particularly if such a network also " 'makes sense' in terms of other knowledge" (Sternberg, 1969, p. 283).

It should be noted here that this relationship between T1 and T3 is different from the type of possible correlation of stage durations induced by factors such as motivation or preparation. Several authors (e.g., Sternberg, 1969; Ashby and Townsend, 1980; Townsend and Ashby, 1983) have pointed out that a subject-controlled factor (such as preparation or motivation), that either vary from trial to trial or is controlled by experimental manipulations such as reward magnitudes, would induce a correlation of stage durations. For example, stage durations could both be short (or long) when the motivation is high (or low). This type of correlation would not destroy the additivity of factors that influence the two stages separately, because for any given level of the subject-controlled factor, stage durations would still be independent. For the Simon-Foley network, experimental manipulations that increase T_1 would be expected to produce a corresponding increase in T_3, but not vice versa. The dependence of T_3 on T_1 is not under the subject's control.

Closed Queueing Networks and Underadditive Factor Interactions

This section considers a special class of queueing networks that predict underadditive factor effects when they are used as models for RT. This class of networks are called closed queueing networks. In a closed network, the same customers circulate eternally through the network. A closed network can also be viewed as an open network with the total number of customers held fixed. All the networks discussed above in this article are open queueing networks. If for some psychological tasks, the cognitive system has an upper limit in terms of the number of stimulus components or task components that it could process at once, then a closed network would appear to be an appropriate candidate for modeling the cognitive system when it functions at its full capacity. An open network, on the other hand, would be useful for modeling a cognitive system that has not reached its full capacity.

Closed networks have a special characteristic that has significant implications for RT modeling. In a closed network, there always exists a negative correlation between the number of customers at the various nodes. This is because in order to keep the total number of customers fixed, an increase in the number of customers at one of the queues will inevitably produce a corresponding decrease in the number of customers at other queues. This negative correlation between the queue lengths at separate nodes will also induce a corresponding negative correlation between the sojourn times of a customer at these nodes.

Insert Figure 6 about Here

The simplest type of a closed queueing network is called cyclic queues, which is essentially a tandem queue with a fixed number, N, of customers allowed in the system. The N customers can also be viewed as N "containers", each of which is able to carry one customer. When a customer departs from the last node, the container that carried it

becomes empty and is immediately cycled back to the front of the system to admit a new customer. To simplify the discussion, let us consider the simplest cyclic queue with two nodes or stages as shown in Figure 6 (K=2). The following discussion uses N to represent the maximum number of customers allowed in the system (the total capacity of the system), N_i the number of customers at queue i, S_i the mean service time requirement of a customer at node i, and T_i the mean sojourn time of a customer at queue i (service time plus waiting time), with i=1, 2.

Clearly, a customer's network sojourn time is the sum of its sojourn times at queue 1 and queue 2. Suppose that the system is functioning at its full capacity, then at the time instant at which a new customer enters the system, there are exactly a total of N-1 customers already in the system. N_1-1 of them are at queue 1, each having a service time requirement of S_1, and the remaining N-N_1 customers are at queue 2, each having a service time requirement of S_2. Therefore, the total network sojourn time, T, of the new customer is

$$\begin{aligned} T &= T_1 + T_2 \\ &= N_1 \times S_1 + (N-N_1) \times S_2 \end{aligned} \qquad (35)$$

Let us examine what happens in a factorial experiment when one factor, A, affects S_1 and another factor, B, affects S_2. Without losing generality, suppose B is held fixed and a change of A from level 1 to level 2 results in a decrease in the processing rate of node 1 and thus an increase in S_1. Because queues tend to form in front of a slow server, this increase in S_1 would produce an increase in N_1, which inevitably produces a corresponding decrease in N_2 (N_2=N-N_1). Thus, a customer would experience a shorter delay at queue 2, although the service rate of node 2 has not been changed. In short, factor B is expected to have a smaller effect on total customer sojourn time when factor A is at a higher level than when A is at a lower level, resulting in an underadditive interaction between the two factors.

As discussed earlier, RT is defined as the total time for M customers to pass through a network. Clearly, when the number of customers are not too small, RT should demonstrate the same characteristic as the sojourn time of an arbitrary customer. To illustrate this point, let us suppose that M is a multiple of N ($M=F \times N$), and the multiplication factor F is not too close to 1. Then, the total time for M customers to pass through the network is the sum of the network sojourn time of the Nth, the 2Nth,..., and the FNth customer. Because the system does not reach its full capacity until the Nth customer has arrived at the system, the network sojourn time of the Nth customer may be somewhat different from that of the 2Nth, ..., FNth customer. But when the value of F is not too small, this difference becomes negligible, and we have

$$RT = F \times T$$

This result indicates that the effects of the experimental factors, A and B, on RT should be the same as their effects on T. That is, A and B should produce underadditive effects in an RT experiment.

Underadditive effects have been observed in a number of studies, an example of which is the study of Miller and Pachella (1976), in which the authors examined the joint effects of the meaningfulness of the stimuli and stimulus contrast on reaction time in a Sternberg-type memory scanning task. The high-meaningfulness condition used the digits 1-8 and the low-meaningfulness condition used pseudoletters as stimuli. The data showed a 140 msec RT difference between the low and the high meaningfulness conditions when stimulus contrast was high, and a 80 msec difference when stimulus contrast was low--demonstrating an underadditive interaction between the two experimental factors.

In another study, Pachella and Miller (1976) found an underadditive interaction between match-type and stimulus contrast in a letter-matching task, which required the subjects to decide whether two simultaneously displayed letters had the same name or not. The two match-type conditions corresponded to the situations in which the matched letters are the same or different in case. The data showed that contrast reduction had a

greater effect on the faster physical match trials (letters do not differ in case) than on the slower name match trials. Evidence of underadditive factor effects were also found in a number of other studies, including that of Stanovich and Pachella (1977) who showed that reducing stimulus contrast increased RT more when the experimental task employed a compatible rather than an incompatible stimulus-response mapping, and that of Miller (1983) who found that whether or not the subjects were given a response cue had a larger effect when a related perceptual process was easy than when it was hard.

These authors argued that these underadditivity results cannot be easily interpreted by discrete serial-stage models, and have interpreted these results as an evidence in support of models that assume temporal overlap of process activities (Stanovich and Pachella, 1977; Miller, 1983). This interpretation was further supported recently by Miller's queue-series model. As discussed earlier, underadditivity was predicted by the queue-series model when the queue-series functions in a nondiscrete manner (Miller, 1993). From the perspective of the queueing network model discussed in this article, however, an underadditive interaction between two experimental factors *may* indicate the presence of a cyclic queueing network as discussed here. We have seen earlier in this article that concurrent processing in a continuous-transmission tandem-queueing system do not necessarily predict underadditivity. Within the classes of models considered in this article, underadditivity in a continuous-transmission serial system are not caused by concurrent processing per se, but by its full-capacity operational status.

The type of negative correlation between sojourn times at successive nodes in a cyclic queue is different from the type of possible trial-by-trial correlation pointed out by Sternberg (1969). Sternberg considered a hypothetical task situation in which the stage duration is supposedly to be shorter if its input is of higher quality, and the input is supposedly to be of higher quality if the preceding stage has worked on it for a longer time. Then *on trials* on which the first stage *happened* to take longer, the second would be shorter. However, factor manipulations might remain additive if this negative

correlation is observed only on a trial-by-trial basis, because trial-by-trial negative correlation does not necessarily imply that independence at the level of means is destroyed. As Sternberg pointed out, this type of relations are beyond the current scope of empirical investigations.

For the cyclic queue, the presence and the magnitude of negative correlations are under the control of experimental investigations, rather than caused by some random extraneous factors, and they are observed at the level of mean durations, rather than only on the trial by trial basis. The cause for the negative correlation is not because an earlier stage has changed its output to be used by a later stage, but because the system is functioning at its full capacity for a certain task situation, and the negative correlation could be observed by making any of the nodes a processing bottleneck.

Another important interpretation of underadditive effects is that based on parallel independent processes. As discussed earlier in the discussion about parallel fork-join networks, Townsend and Ashby (1983) have shown that two experimental factors influencing parallel independent processes separately would produce an underadditive interaction. This result was later generalized by Schweikert and Townsend (1989) to concurrent processes in a larger class of discrete networks. Just as what we have seen in the last section that the discrete serial stage model, the cascade model and the Simon-Foley queueing network model offer three alternative explanations for overadditive interactions, the queue-series model, the parallel processes model and the cyclic queueing model offer three alternative interpretations for underadditive interactions. The purpose of this article is not to advocate one interpretation over another, but to broaden the scope of thinking about the possible causes for certain RT phenomenon. Before more diagnostic tests are available, it appears that the choice of a model among the alternatives should rely on "intuitive and sometimes logical grounds for rejecting one model or interpretation in favor of another" (Townsend, 1976; p.237).

The results in the last two sections on Simon-Foley network and cyclic queueing network have special theoretical significance in that they may suggest some new classes of mental architectures that do not follow the postulate of selective influence but can be subjected to empirical tests. Previous studies have discussed the possible existence of indirect influence of experimental factors. However, as Townsend and Schweikert (1989) have pointed out: "Although dependencies are extremely important in processing networks, very little is known about how to identify them or show they may affect the results of psychological experiments"(Townsend and Schweikert, 1989; p. 321). Simon-Foley network and the cyclic queueing network are examples of networks in which process dependencies exist and they are clearly defined and can be tested through experimental investigations. Furthermore, since both types of networks are continuous-transmission systems, further research along this direction can also offer deeper insights into the debate about the discreteness and continuity in information processing.

In summary, this article presented a queueing network model for reaction time that considers the temporal dimension of discrete versus continuous information transmission in conjunction with the architectural dimension of serial versus network configurations of elementary mental processes. A number of elementary but important types of queueing networks are examined with regard to their predictions for RT behavior if they are used as models for psychological processes. These networks include tandem networks, fork-join networks, feedback networks, Simon-Foley networks, and cyclic networks. Interestingly, each of the networks makes some testable predictions for reaction time and offers some unique insights into the theoretical debates about possible mental structures. These models also provide a new perspective for reexamining existing models of psychological processes, and it turns out many of the conclusions of existing models are subject to alternative explanations.

It should be emphasized here that queueing network theory, in its general form, has the capacity to model a greater variety of processing systems than what have been

considered in this article. And this capacity is rapidly expanding, because queueing network is one of the most active areas of research in operations research and applied mathematics. Undoubtedly, progresses in this area will make more methods, tools and concepts available. This article has illustrated the power of queueing network methods in establishing new models of human cognition and in serving as a larger modeling framework for unifying some existing models.

REFERENCES

Ashby, F. G. (1982). Testing the assumptions of exponential, additive reaction time models. Memory and Cognition, 10, 125-134.

Ashby, F. G. (1982). Deriving exact predictions from the cascade model. Psychological Review, 89, 599-607.

Ashby, F. G., & Townsend, J. (1980). Decomposing the reaction time distribution: Pure insertion and selective influence revisited. Journal of Mathematical Psychology, 21, 93-123.

Baccelli, F. & Makowski, E. (1990). Synchronization in queueing systems. In H. Takagi (Ed). Stochastic analysis of computer and communication systems (pp. 57-129). Amsterdam: North-Holland.

Boxma, O. & Daduna, H. (1990). Sojourn times in queueing networks. In H. Takagi (Ed.), Stochastic Analysis of Computer and Communications Systems, p.401-450, North Holland.

Buzen, J. (1976). Fundamental operational laws of computer system performance. Acta Informatica, vol-7, pp.167-182.

Christie, L. S. & Luce, R. D. (1956). Decision structure and time relations in simple choice behavior. Bulletin of Mathematical Biophysics, 18, 89-112.

Denning, P. & Buzen, J. (1978). The operational analysis of queueing network models. Computing Surveys, vol-10, pp. 225-261.

Disney, R. & Konig, D. (1985). Queueing networks: A survey of their random processes. SIAM Review, 27, 335-403.

Donders, F. C. (1969). Over de snelheid van psychische processen [On the speed of mental processes]. In W. G. Koster (Ed.), Attention and performance (Vol. 2, pp. 412-431). Amsterdam: North-Holland. (Original work published in 1868).

Erlang, A. K. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges, Electroteknikeren (Danish) 13, 5-13. (English translation in the Post Office Electrical Engineer's Journal, 10, 189-197.

Fisher, D. & Goldstein, W. (1983). Stochastic PERT networks as models of cognition: Derivation of the mean, variance, and distribution of reaction time using Order-of-Processing (OP) diagrams. Journal of Mathematical Psychology, 27, 121-151.

Foley, R., & P. Kiessler (1989). Positive correlations in a three-node Jackson queueing network. Advances in Applied Probability, 21, 241-242.

Fulkerson, D. R. (1962). Expected critical path length in PERT networks. Operations Research, 10, 808-817.

Hartley, H. O., & Wortham, A. W. (1966). A statistical theory for PERT path analysis. Management Sciences, 12, pp. B-469-B.481.

Hockley, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 598-615.

Hohle, R. (1965). Inferred components of reaction times as functions of foreperiod duration. Journal of Experimental Psychology, 69, 382-386.

Jackson, J. R. (1957). Networks of waiting lines. Operations Research, 5, pp. 518-521.

Jackson, J. R. (1963). Job-shop like queueing systems. Management Science, 10, 131-142.

Kleinrock, L. (1975). Queueing Systems. New York: Wiley.

Lemoine, A. J. (1987). On sojourn time in Jackson networks of queues. Journal of Applied Probability, 24, 495-510.

McGill, W. & Gibbon, J. (1965). The general-gamma distribution and reaction times. Journal of Mathematical Psychology, 2, 1-18.

McClelland, J. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. Psychological Review, 86, 287-330.

Meyer, D. E., Irwin, D. E., Osman, A. M., & Kounios, J. (1988). The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. Psychological Review, 95, 183-237.

Miller, J. O. (1982). Discrete versus continuous stage models of human information processing: In search of partial output. Journal of Experimental Psychology: Human Perception and Performance, 8, 273-396.

Miller, J. O. (1983). Can response preparation begin before stimulus recognition finishes? Journal of Experimental Psychology: Human Perception and Performance, 9, 161-182.

Miller, J. O. (1988). Discrete and continuous models of human information processing: Theoretical distinctions and empirical results. Acta Psychologica, 67, 1-67.

Miller, J. O. (1990). Discreteness and continuity in models of human information processing, Acta Psychologica, 74, 297-318.

Miller, J. O. (1993). A queue-series model for reaction time, with discrete-stage and continuous-flow models as special cases. Psychological Review, 100, 702-715.

Miller, J. O., & Pachella, R. (1976). Encoding processes in memory scanning tasks. Memory and Cognition, 4, 501-506.

Pachella, R. (1974). The interpretation of reaction time in information processing research. In B. Kantowitz (Ed.), Human information processing: Tutorials in Performance and Cognition, p.41-82, Hillsdale, N.J.: Erlbaum.

Pachella, R., & Miller, J. O. (1976). Stimulus probability and same-different classification. Perception and Psychophysics, 19, 29-34.

Ratcliff, R. (1988). Continuous versus discrete information processing: Modeling accumulation of partial information. Psychological Review, 95, 238-255.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. Psychological Review, 86, 190-214.

Ross, S. M. (1983). Stochastic processes. New York: Wiley.

Rumelhart, D. E. (1970). A multi-component theory of the perception of briefly exposed visual displays. Journal of Mathematical Psychology, 7, 191-218.

Schweikert, R. (1978). A critical path generalization of the additive factor methods: Analysis of a Stroop task. Journal of Mathematical Psychology, 18, pp. 105-139.

Schweikert, R. (1982). The bias of an estimate of coupled slack in stochastic PERT networks. Journal of Mathematical Psychology, 26, 1-12.

Schweikert, R., & Townsend, J. T. (1989). A trichotomy: Interactions of factors prolonging sequential and concurrent mental processes in stochastic discrete mental (PERT) networks. Journal of Mathematical Psychology, 33, 328-347.

Simon, B. & Foley, R. Some results on sojourn times in acyclic Jackson networks. Management Science, vol-25, 1027-1034.

Stanovich, K., & Pachella, R. (1977). Encoding, stimulus-response compatibility, and stages of processing. Journal of Experimental Psychology: Human Perception and Performance, 3, 411-421.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders's method. Acta Psychologica, 30, p.276-235.

Takacs, L. (1963). A single-server queue with feedback. Bell System Technical Journal, 42, pp. 505-519.

Townsend, J. (1972). Some results concerning the identifiability of parallel and serial processes. British Journal of Mathematical and Statistical Psychology, 25, 168-199.

Townsend, J. (1974). Issues and models concerning the processing of a finite number of inputs. In B. Kantowitz (Ed.), Human information processing: Tutorials in Performance and Cognition (pp. 41-82), Hillsdale, N.J.: Erlbaum.

Townsend, J. T. (1976). Serial and within-stage independent parallel model equivalence on the minimum completion time. Journal of Mathematical Psychology, 14, 219-238.

Townsend, J. T. (1984). Uncovering mental processes with factorial experiments. Journal of Mathematical Psychology, 28, 363-400.

Townsend, J. T., & Ashby, F. (1983). The Stochastic Modeling of Elementary Psychological Processes. Cambridge: Cambridge University Press.

Townsend, J. T., & Schweikert, R. (1985). Interactive effects of factors affecting processes in stochastic PERT networks. In G. d'Ydewalle (Ed.), Cognition, information processing, and motivation. Amsterdam: North-Holland.

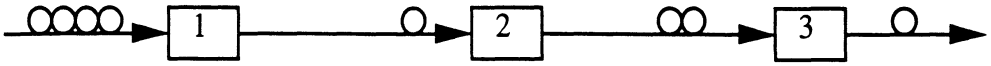
Townsend, J. T., & Schweikert, R. (1989). Toward the trichotomy method of reaction times: Laying the foundation of stochastic mental networks. Journal of Mathematical Psychology, 33, 309-327.

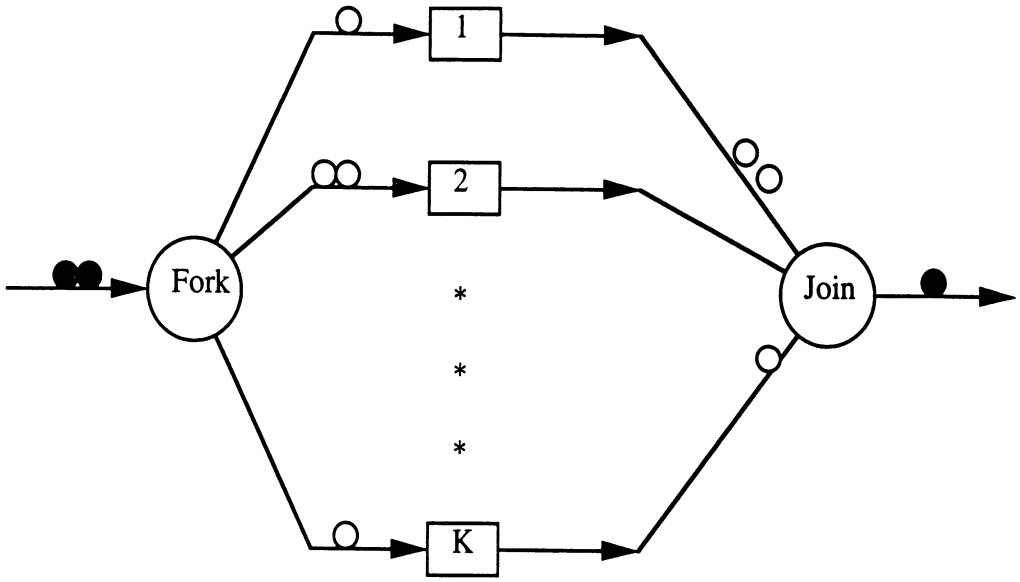
Walrand, J. & Varaiya, P. (1980). Sojourn times and the overtaking condition in Jacksonian networks. Advances in Applied Probability, 12, 1000-1018.

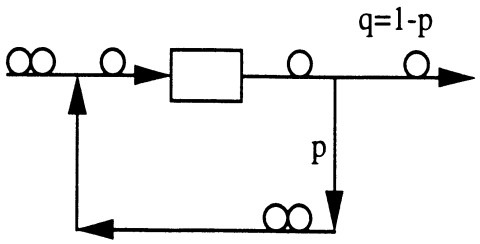
Whitt, W. (1984). The amount of overtaking in a network of queues. Networks, 14, 411-426.

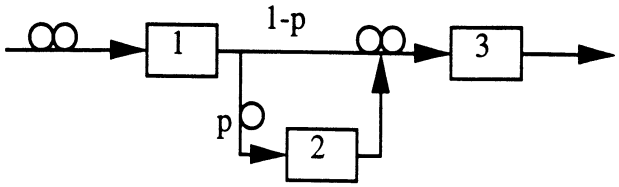
List of Figures

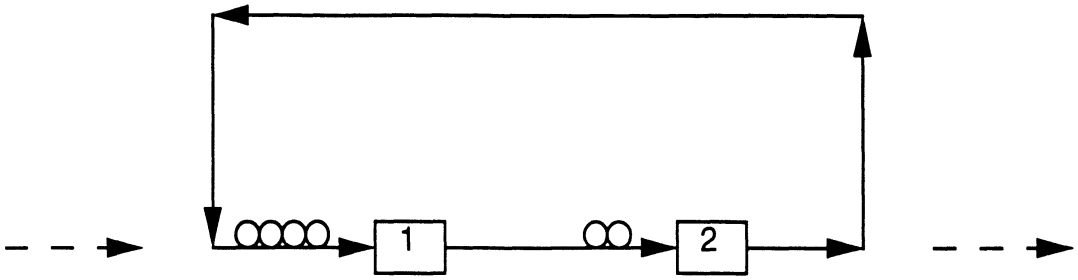
	Series	Network
Discrete	1. Subtractive 2. Additive Factors 3. General-Gamma	1. PERT (Critical-Path) Network
Continuous	1. Cascade 2. Queue-Series	1. Queueing Network


1
2
3


1
2
Fork
*
*
*
Join
K


q=1-p
p


1-p
1
3
p
2


1
2