QUEUEING NETWORK MODELING OF HUMAN PERFORMANCE OF CONCURRENT SPATIAL AND VERBAL TASKS

Yili Liu
Department of Industrial and Operations Engineering
The University of Michigan
Ann Arbor, MI 48109-2117

Technical Report 93-31

November 1993

Queueing Network Modeling of Human Performance of Concurrent Spatial and Verbal Tasks

Yili Liu, *Member*, IEEE

Yili Liu is on the faculty of Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109.

Correspondence should be addressed to Yili Liu at the above address, who can also be contacted at (313)763-0464 (phone), (313)764-3451 (fax), or yililiu@engin.umich.edu (Internet).

ABSTRACT

This article describes a 3-node queueing network model of human multitask performance to account for interferences between concurrent spatial and verbal tasks. The model integrates considerations of single channel, queuing theoretic models of selective attention and parallel processing, multiple resources models of divided attention, and provides a computational framework for modeling both the serial processing and the concurrent execution aspects of human multitask performance. The single channel and the multiple resources concepts and their applications in engineering models are reviewed. Experimental evidence in support of the queueing network model is summarized. The potential value of using queueing network methods to integrate currently isolated concepts of human multi-task performance and in modeling human machine interaction in general is discussed.

I. INTRODUCTION

One of the common characteristics of an operator's task in human-machine systems is the need to perform a number of concurrent activities at once. Examples of multi-task situations abound and include an automobile driver who has to ensure the smooth operation of a vehicle while time-sharing between the instrument panel and the forward view of the roadway, and a traffic controller who has to divide attention between various visual and auditory sources of information, while making time-critical decisions and performing intensive communications activities. The requirements for processing multiple sources of information often push the operators in multitask human machine systems to the upper bound of their attention capabilities.

Fortunately, the increasing power and sophistication of hardware and software technology are providing more and more options for human machine system design that could take into account the characteristics of the human operator. In this regard, as indicated by a recent report of the Committee on Human Factors of the National Research Council, comprehensive, engineering-based, predictive models of operator performance and workload in complex multitask systems become of increased importance [3]. These models can help assess the impact of the technological infusions and determine the most effective design before a system is configured, and will allow the human factors professionals to communicate their knowledge to the engineering community more effectively in a language that is compatible with the designers' existing terminology and conceptual base.

Many predictive models of multitask performance have been proposed to address the nature and the cause of task interference. Prominent among these models are the single channel, serial processing models and the multiple resources, parallel processing models. The two schools of models have fundamental differences in their views of the nature of multitask performance and in their research and modeling methodology. The single channel, serial processing models treat multitask performance as an issue of task selection and scheduling: human information processing system can only attend to one task at a time, and multitask

performance relies on the rapid switching of attention among the tasks competing for attention [9], [36], [45]. The multiple resources, parallel processing models, in contrast, treat multitask performance as an issue of parallel allocation and division of processing resources among simultaneous tasks: multiple tasks can be processed at the same time as long as the total demand does not exceed the limit of attentional capacity or processing resources [59].

Until recently, there has been a substantial gap between the two schools of models. As models of human behavior, both schools of models have received substantial support from a multitude of experimental studies. But at the same time, it has become increasingly evident that neither school of models alone is sufficient in providing fully satisfactory explanations to the empirical data. From the perspective of engineering modeling, the single channel assumptions have thus far enjoyed a greater success, as indicated by the existence of a set of well-established models such as the queueing theoretic models and the network models reviewed in the following section. These models provide formal mechanisms for representing and codifying the single channel assumptions of task selection in engineering terms. The multiple resources models, in contrast, have only recently started to see some of their concerns being gradually accommodated in several simulation models of human performance, and there is still a lack of a set of computational methods to represent the assumptions of simultaneous execution and resource allocation in engineering terms. Furthermore, there does not exist a set of integrated engineering-based methods to model the concerns of both schools of models and to bridge the gap between the two. As indicated in the recent National Research Council report, there is a lack of methods to model the two most important features of a macromodel: task selection and simultaneous execution [3].

Recently, Liu [29] [30] proposed that queueing network models and related methods employed widely in industrial engineering and system performance analysis may provide us an integrated computational framework for modeling the complex structural and temporal arrangements that multiple tasks might assume. The structural arrangements include both serial selection and parallel execution, and the temporal arrangements include both

immediate activities and delayed processing. The purpose of this article is to examine the gap between the single channel and the multiple resources models of multitask performance, to describe a 3-node queueing network model of interference between concurrent spatial and verbal tasks, and to illustrate the potential power of the queueing network approach for modeling multi-task performance. As described below, the queueing network model provides a computational framework to integrate the concerns of the single channel and the multiple resources models, which, in essence, can be treated as special cases of the queueing network model.

The structure of the article is as follows. The first two sections review the single channel and the multiple resources models in detail, in terms of their theoretical assumptions and their applications in engineering modeling. Both the successes and the limitations of the two modeling approaches will be discussed. Then, a specific pattern of task interference between concurrent spatial and verbal tasks will be discussed as an illustration of the gap between the two classes of models. A 3-node queueing network model is then described as a plausible and intuitive account of interference between spatial and verbal tasks and as an attempt to bridge the gap between the single channel and the multiple resources models. The value and the potential power of using queueing network methods to integrate some other currently isolated concepts of human performance and in modeling human-computer systems in general is discussed at the end of the article.

II. SINGLE CHANNEL, QUEUEING THEORETIC MODELS OF MULTITASK PERFORMANCE

As mentioned above, a prominent theory of multitask performance is the single channel theory of selective attention. Its root can be traced to the single channel theory of human information processing originally proposed by Craik [12] to explain the psychological refractory period in human information processing discovered by Telford [51].

Telford discovered that when two reaction time tasks are presented close together in time, the reaction time to the second task stimulus is consistently delayed from a single task control condition. Various forms of single channel theories have been subsequently proposed and elaborated [6] [14] [35] [56]. What is consistent about these theories is their common assumption that the human information processing system has bottlenecks that can only process one stimulus or piece of information at a time, and thus the system functions through a series of selections about which stimulus or piece of information to process. The focus of investigation is the identification of the bottlenecks, and the topics of debate among these theories are their different opinions regarding the locus of the bottlenecks and the factors that influence the selection processes.

The single channel psychological theory of selective attention has been the fundamental basis of numerous engineering models of human performance [9] [42] [45]. These engineering models postulate that the human is a single channel processor or a time-shared computer with a single central processing unit (CPU), which quickly switches and allocates its processing capacity among a variety of tasks in a sequential and all-or-none fashion. The models view human multitask performance as a single server queueing problem or multitask sequencing problem in which multiple tasks or diverse sources of information are queued for service of the single-server human information processing system.

Early models in this tradition have focused on modeling human visual sampling and monitoring behavior. Senders developed an instrument monitoring model, which integrated the single channel concept and the sampling theorem of Shannon's information theory in making its predictions about the observer's fractional dwell time on each monitored instrument [45]. Carbonell proposed a single server priority queueing model of multi-instrument visual sampling [9]. The priority of each instrument at any instant is modeled as the combined effect of both the probability and the cost of exceeding a prescribed limit. The model integrates concepts from queueing theory, information theory, and decision theory. Carbonell used simulation to solve the model, and showed a close fit between the model's

predictions and the subjects' actual performance in flying a simulator in terms of the fraction of attention devoted to each instrument. Senders and Posner further developed the queuing theoretic approach to instrument monitoring and provided analytical solutions to a model that they developed for display sampling [46]. Schmidt applied queueing theoretic method to the analysis of air traffic control task [43].

A systematic and extensive effort in applying the queuing theoretic methods to the modeling of human machine systems can be found in a series of studies conducted by Rouse and his colleagues. Rouse described human-computer interaction as a queueing system with the human and the computer as two servers [42]. He formulated a queueing theoretic model of dynamic allocation of responsibility between the human and the computer in multitask situations, and illustrated the potential utility of this model with simulation experiments. Chu and Rouse later investigated the predictive power of the model with a behavioral experiment that simulated a multitask flight management situation [10]. A similar task scenario was also used in an earlier study by Walden and Rouse that investigated the suitability of a single server queueing model of pilot decision making [55]. Greenstein and Rouse integrated a pattern recognition technique called discriminant analysis with queueing theory methods in their 2-stage model of human decision making in multi-process monitoring situations [19]. Discriminant analysis was used in the first stage to generate estimates of event occurrence probability, and queueing theory was then applied in the second stage to incorporate these probabilities into the solution of the attention allocation problem.

The single channel assumptions can also be found in several other engineering models. For example, Sheridan assumed that there is a mental cost to switching attention which will determine how often different information sources in the environment are sampled, and thereby influence the sampling behavior [47]. Kleinman and Curry developed a model for human operator display monitoring, which assumed that the human is a single channel time-shared processing channel [26]. Tulga modeled the multitask attention allocation problem as a dynamic, single machine sequencing problem [53]. The concepts and

assumptions of single channel processing and sequencing were also employed in the semi-Markov dynamic decision model of human task selection performance proposed by Pattipati, Kleinman and Ephrath [40].

The single channel concepts have been the fundamental basis of numerous simulation models of human performance. Notable examples include the task network models [28], PROCRU [4] and HOS [57]. Started with the SAINT (Systems Analysis of Integrated Networks of Tasks) modeling methodology developed by Siegal and Wolf [48], the task network approach models the human interaction with the environment as a sequence of tasks (also called paths), and acknowledges the existence of alternative paths to accomplish a goal or different goals in certain circumstances. These alternative paths form a task network. Parallel paths in a task network represent alternatives rather than concurrence of processing. Furthermore, a task in a network can not be started until the preceding task on the same path of the network has been completed. Thus, at any instant only one task on a path can be executed.

PROCRU (Procedure-Oriented Crew Model) is a control theory-based simulation model that has received widespread recognition [4]. The model is a closed-loop system model incorporating submodels for the aircraft, aircraft crew members, and the air traffic controller. The crew member submodel is a detailed human performance model and has a comprehensive coverage of human activities in monitoring and control of a large system. The model assumes that the crew members have a set of procedures or tasks to perform, and one task is chosen at a given instant in time, which is the one perceived to have the highest expected gain for execution at that time. The model contains a procedure selector, which is responsible for task selection and sequencing.

The Human Operator Simulator (HOS) uses a library of human performance micromodels to simulate the operator's perceptual, cognitive and motor responses [21] [50] [57]. The original versions of the HOS approach assumes that humans are single channel processors, and that human behavior is goal-oriented, and can be defined as a sequence of

discrete micro-tasks, which can be aggregated to predict task performance. Other simulation models of complex task performance that are based on single channel assumptions include SIMWAM [17] [33], STALL [11], the model developed by Tulga and Sheridan [54] and its subsequent modified versions (e.g., [37]).

In spite of their differences, one of the common features of these models is their reliance on the fundamental assumption that humans can only process one piece of information at a time. Human multitask performance is modeled as a process of selecting tasks for sequential action according to some service discipline or cost function, which is usually based on the assumption that there is a mental cost to switching attention and/or there is a cost of being unable to attend to a critical instrument in a timely fashion [39] [42] [47]. Another common characteristic of these models is their focus on time as the underlying dimension and the metric of processing. Time is what is competed for by multiple tasks in a serial fashion and completion time defines the difficulty or demand of each task or task component. The models are relatively silent as to the intensity aspects of task demand.

These single channel based models have demonstrated tremendous success in modeling two aspects of human performance: visual sampling in process monitoring and strategic task scheduling in high workload situations. The primary concern of visual sampling is to find the optimal tactics for a single channel sampler to sample sources of information sequentially when the information sources can not be attended to at once and thus compete for the operator's focal attention [36]. For example, in monitoring instrument panels, owing to the need for foveal vision when accurate reading is required, the eyes must be pointed in an appropriate direction to scan and sample a source of information. Since the distances between instruments are most often greater than the radius of foveal vision (2 to 3 degrees of visual angle), accurate reading of spatially separated instruments can only be done sequentially. Similarly, in high workload situations when time pressure is the major source of workload and the operators are free to choose the order in which the tasks should be done to avoid overload [1] [37], these single channel models have been quite successful in capturing

the strategic scheduling aspect of task performance.

However, intuition and experimental evidence both support the view that humans do have the ability to perform multiple tasks in a truly concurrent fashion under many real-world circumstances. "We must recognize that people in fact can do more than one thing at a time and normally do ([1]; p.5)". In these task situations, single channel assumptions and the analogy of a single-CPU time-shared computer or a single server queueing system may only capture part of the nature of human performance and may not be adequate to portray the complex cognitive mechanisms for concurrent processing. Furthermore, each of the concurrent tasks may have attentional demands varying in intensity as well as in time. Some task pairs may be more similar with each other in their task structure than with others, and thus produce different patterns of task interference when they are performed concurrently with each other than when they are with other tasks. Thus, it may be necessary to address the parallel processing aspect of performance, to consider the intensity as well as the time characteristics of task demand, and to analyze the structural aspects of concurrent tasks. These issues have been the focus of investigation of models of divided attention.

III. MULTIPLE RESOURCES, PARALLEL PROCESSING MODELS OF MULTITASK PERFORMANCE

In contrast to the single channel assumption adopted by selective attention theorists that attention capacity can only be switched in a sequential and all-or-none fashion among competing tasks, divided attention theorists have generally suggested that the limited human information processing capacity can be simultaneously allocated to multiple tasks in a graded fashion, and that information for simultaneous tasks can be processed in parallel as long as the total processing load does not exceed a person's processing capacity. Since the early 1970s, the concept of capacity has become more commonly referred to as a "resource", a concept first formalized by Kahneman as a mental energy necessary to support task performance [24].

Early resource theory holds that there is a single undifferentiated pool of processing resource, and performance of concurrent tasks should deteriorate as they compete for more of the same scarce resource [24]. However, it has become increasingly evident from the results of a multitude of experimental studies that human information processing may include multiple pools of resources [59]. Foremost among these results was the phenomenon of difficulty insensitivity: the failure of performance on one task to reflect changes in the difficulty of a concurrent task. A related phenomenon is that two tasks may be found to have perfect time-sharing with each other, but each can be shown to interfere with other tasks.

Based on a review of experimental literature, Wickens proposed a multiple resources model that assumes the existence of multiple pools of processing resources in the human information processing system [59]. The model represents resource composition as three dichotomous dimensions, which include codes of processing (that distinguishes verbal-linguistic vs. analog-spatial coding and representation of information material), stages of processing (that distinguishes perceptual/cognitive vs. response selection and execution), and modalities of processing (that distinguishes auditory vs. visual perceptual channel and voice

vs. manual response). According to the multiple resources model, the three dimensions define separate resources, and interference between concurrent tasks should only be manifest to the extent that the tasks compete for common resources.

The multiple resources model has had significant impact in the human factors community, particularly as a conceptual framework for understanding multitask performance and as a heuristic for guiding multitask system design. Recently, several simulation models of human performance, e.g. the MICROSAINT model [28] and the WINDEX model [38], have started to accommodate some of the assumptions of the multiple resources models. As briefly reviewed in the following paragraphs, these models have started to address the structural aspects of task demands and that they have taken a step toward acknowledging aspects of task demands that are not defined only in terms of time.

Operator workload is modeled in MICROSAINT--an elaboration and development of the original SAINT model [48] -- with a variant of a technique that had been used by McCracken and Aldrich [34]. This technique characterizes operator activities as imposing some amount of workload demand along each of four attention channels: the auditory channel, the visual channel, the cognitive processing channel, and the psychomotor output channel. A set of benchmark scales developed by McCracken and Aldrich is used for determining the workload intensity or demand for each channel.

The WINDEX (Workload INDEX) model [38] perhaps represents the highest level of sophistication among the simulation models thus far in accommodating the multiple resources assumptions. The model combines mission, task and timeline analysis with the multiple resources considerations to predict attentional demands in a crewstation. A critical component of the WINDEX model is a conflict matrix, which determines the amount of interference between concurrent activities according to their similarity in the multiple resources space defined by the Wickens's model.

Of the three dimensions in the multiple resources model, the one that has received the most consensus in the literature is the one based on a distinction between spatial and verbal

processing codes. Briefly stated, since separate stages are used to perform different functions, this dimension is less relevant for predicting alternative methods of carrying out equivalent tasks. The role of perceptual modalities has been found, in a variety of experiments, to be more complicated than it was first proposed. Other factors such as visual scanning, auditory masking and cross-modality switching may influence the pattern of task interference (for a detailed discussion, see Wickens and Liu, 1988, [61]). Therefore, the present article will focus on the dimension of processing codes as defining separate resources.

The dichotomy of spatial and verbal processing codes distinguishes operations of perception, working memory and response that have a linguistic and symbolic base from those that have a spatial analog base. Researchers of multitask performance have started to realize the role of processing codes in accounting for variances in task interference since the pioneering study of Brooks conducted in the late 1960s [7]. In a series of experiments, Brooks compared the effects of spatial and verbal response methods on the subjects' performance of an imagery task, which involved mental operations in either spatial working memory or verbal working memory. The spatial memory task required the subjects to mentally "walk" around the perimeter of an imagined capital letter and indicate in turn whether each corner was in a designated orientation. The verbal memory task required the subjects to mentally "walk" through an imagined familiar sentence and indicate in turn whether each word belongs to a particular grammatical category. The results showed that the spatial working memory task was consistently more disrupted by the spatial response method (pointing to a column of Y's and N's) than by the verbal response method (vocal articulation). But the reverse pattern of interference was observed for the verbal memory task.

This pattern of task interference has been observed in a variety of experimental tasks, ranging from skill-based tasks (e.g., [63]); to rule-based tasks (e.g., [52]) and to knowledge-based tasks (e.g., [64]). Converging evidence strongly supports the definition of processing codes as separate resources in accounting for task interference, and suggests that analysis of human performance must address this structural aspect of task demands and distinguish

between within-code and between-codes tasks.

While the multiple resources concepts represent a significant advancement in multitask performance theory, there are at least two issues that need substantial further research. First, although the single channel assumptions have enjoyed a great success in engineering modeling as reviewed in the previous section, there is still a lack of a set of engineering methods to transform and codify the multiple resources assumptions of simultaneous execution and resource allocation in a computational form. Second, since the multiple resources models were originally proposed to address the characteristics of parallel allocation of scarce yet divisible processing resources to concurrent activities, the models do not provide a formal mechanism to model the role of serial processing bottlenecks and the selective and scheduling aspects of task performance. Thus, a typical research strategy of investigators in this research paradigm has been to treat these bottlenecks of serial processing as extraneous factors or to keep the influence of these factors as small or constant as possible (see, Liu and Wickens, 1992, [32], for an analysis). However, as reviewed in the previous section, processing bottlenecks do exist and play an important role in many task situations. "Given extremely high and time-intense processing demands, it may be impossible to parallel process the information, and the subject may be forced into a serial processing mode" ([61]; p. 612). The next section will argue that the existence of processing bottlenecks might be a major cause of interference between tasks that do not use the same processing code.

IV. GAPS BETWEEN QUEUEING THEORETIC AND MULTIPLE RESOURCES MODELS

Apparently, substantial gaps exist between the single channel, queueing theoretic models and the parallel processing, multiple resources models. Both schools of models have received substantial support from a multitude of experimental studies and demonstrated great value in analyzing certain types of task situations. But at the same time, it has become increasingly clear that each school only emphasizes one aspect of multitask performance, and neither school of models alone appears to be sufficient in providing fully satisfactory explanations to both the serial processing and the parallel allocation aspects of task performance. Furthermore, from the perspective of engineering modeling, there does not exist a set of integrated engineering-based methods to model the concerns of both schools of models in a quantitative manner [3].

To further illustrate the need for integrating the concerns of the two schools of models, let us examine, as an example, a pattern of task interference that has been demonstrated in experimental studies and is consistent with intuition. Imagine a situation in which a spatial task is time-shared with a concurrent spatial task or a verbal task. The single channel models would predict that interference between the concurrent tasks should be determined by the level of difficulty of the concurrent tasks, regardless of whether they use the same or separate processing codes. The most logical prediction of the multiple resources models, however, would be that the performance/difficulty tradeoff should be observed only if concurrent tasks use the same processing code. It has become increasingly evident that neither of the two predictions is fully consistent with empirical data. What has been demonstrated in a variety of experimental studies is that task interference is expected to be greater when concurrent tasks use the same processing code than when they require separate codes. However, the data do not indicate that two tasks demanding separate codes will always be perfectly time-shared.

Experimental evidence and intuition suggest a pattern of task interference as follows. First, task interference will not be observed when the total demand of the concurrent tasks is low, regardless of the processing codes involved. Second, when the total task demand is sufficiently high, both within-code and between-codes interferences could be observed, but within-code interference is more likely to occur; third, increases in task difficulty tend to produce faster increases in within-code than between-codes interferences; fourth, when the processing demand of each of the concurrent tasks is high, the human appears to behave more like a single channel processor and can only process one demanding task at a time.

It appears that this pattern of task interference is consistent with everyday experiences. For example, in a perfect driving environment an easy secondary task can be performed concurrently with the primary driving task without causing any performance decrement, no matter whether the easy task is spatial (e.g., imagining a simple road map or tuning a radio) or verbal (e.g., recollecting a previous conversation or talking to a passenger). But under the same driving condition, a difficult secondary spatial task (such as mentally "walking" around a complex spatial layout or using a complicated navigational device) would be more likely to disrupt driving than would an equally difficult verbal task (such as reciting a difficult poem or engaging in a challenging conversation). As the driving environment becomes more hostile, a driver would very likely to experience great difficulty in performing even a simple secondary spatial task but could still perform some easy secondary verbal tasks such as a light conversation. But as the difficulty of driving further increases--such as driving on a winding and narrow road in a stormy weather or in a heavy-traffic area--the driver would have to concentrate on driving, which could be easily disturbed by a slight disruption of any kind.

The following is a brief review of some experimental evidence that demonstrate this pattern of task interference. Wickens, Mountford, and Schreiner investigated dual-task performance decrements in an experiment in which four individual component tasks were selected to be performed in all pairwise combinations [62]. The four tasks included a one-

dimensional compensatory manual tracking task, a line judgment task, an auditory short-term memory task and a digit classification task. The tasks were selected specifically to place demands upon qualitatively different capacities of processing resources. The tracking and the line judgment task both placed heavy demands upon spatial processing, whereas the short-term memory and the digit classification task both placed heavy demands on verbal processing. The experimental results showed performance decrements in all dual task conditions, but the decrements were greater when two concurrent tasks required the same processing code than when they required separate codes. For example, the tracking task suffered the largest decrement when it was time-shared with another tracking task, a smaller decrement with the line judgment task, and the smallest, yet still significant, decrement with the verbal memory or classification task.

In a series of four experiments Klapp and Netick compared the amount of interference of a verbal reference task called probe digit (PD) and a spatial reference task called missing digit (MD), with several concurrently performed verbal or spatial distracting tasks. Both tasks required the subjects to remember a list of eight digits [25]. The PD task required the subjects to respond, when they were presented with one of the eight digits as a "probe", with the digit that had followed the probe in the original sequence. The MD task required the subjects to identify which of the nine possible digits (from the population of 1-9) was absent from the original list. The authors argued that PD used auditory-verbal memory storage but MD used visual-spatial memory storage. In the four experiments performance decrements were observed in most of the dual task conditions, no matter whether the concurrent tasks required the same or separate processing codes. However, the decrements were found to be greater when they required the same code. For example, vocalization as a concurrent verbal task interfered with PD more than with MD (in Experiment 1), but tracking as a concurrent spatial task interfered with MD more than with PD (in Experiment 2). A concurrent verbal memory load interfered with PD more than with MD (in Experiment 3), but a spatial memory load was shown to interfere only with MD and not with PD (in Experiment 4).

There is also clear evidence that between-codes interference could be absent in some other task situations. For example, Baddeley conducted a series of studies which tested subjects' memory for complex chess positions while performing a concurrent spatial or verbal task. All subjects-- ranging from the modest club player to the international grand master-- showed the same basic pattern of task interference: performance of this spatial memory task was not disrupted by the concurrent verbal task but showed clear impairment from the concurrent spatial task [2].

An important factor that must be considered in comparing within-code and between-codes interferences is the level of difficulty of the spatial or verbal tasks involved. Although the data is limited in which the processing codes of concurrent tasks have been manipulated along with a simultaneous manipulation of task demand, the reported data supports the view that within-code task interference is greater than between-codes interference. For example, Wickens found that performance on a manual tracking task was more disrupted by a manual response task than by an auditory signal detection task, even though the latter was judged by the subjects to be more difficult [58]. Another example is the study by Wickens, Sandry and Vidulich [63], in which subjects were required to time-share a tracking task with a verbal reaction time (RT) task. When the RT task presented the stimuli auditorily and required voice response, an increase in tracking difficulty produced no increase in task interference. However, task interference increased significantly when visual input and manual response were employed with the RT task.

In order to systematically examine the joint effects of task difficulty and processing codes on task interference, Liu and Wickens conducted a series of experiments, in which the experimental conditions were carefully controlled to achieve a simultaneous manipulation of processing code and task difficulty, while keeping all other aspects of the spatial and verbal task conditions equivalent to ensure that the contrast between verbal and spatial code was not confounded with other variables [29] [32] [61]. Performance measures and the NASA-TLX subjective workload ratings [22] were used to establish the difficulty levels of the tasks. The

results of the experiments are consistent with the pattern of task interference described earlier in this section.

In the first study, Wickens and Liu required the subjects to perform a simulated flight task, which was time-shared with a verbal or spatial decision task whose degree of demand was varied between two levels [61]. The simulated flight task was a one-dimensional compensatory tracking task with second-order control dynamics (acceleration control). The spatial task required the subjects to predict the future position of a displayed vector and the verbal task required the prediction of the future value of a displayed numerical variable. The spatial and the verbal tasks were designed in a way that they imposed analogous demands on the respective spatial or verbal working memory systems, by imposing a continuous running memory task with overlapping encoding, storage and retrieval processing. The difficulty level of the two types of tasks were equated by comparing single task performance measures and subjective ratings. The results showed clear evidence that both the spatial and the verbal decision tasks produced significant interference with the tracking task, but the amount of interference produced by the spatial task was found to be greater than that by the verbal task. A particularly interesting finding was that a further increase in the difficulty of the spatial task produced a significant increase in tracking error, whereas the same amount of increase in the verbal task difficulty produced a small and nonsignificant increase in tracking error.

In the second study [32], half of the experimental trials were similar to those of the first study--the subjects were required to time-share a primary second-order tracking task with a secondary spatial or verbal decision task, and the relevant information that was needed to perform the decision tasks was displayed at a fixed location on the visual display. On each of the other half of the trials, however, the decision information was displayed at either of two possible locations with equal probability. The subjects were thus required to search the visual display to find the needed information before making a decision. Since visual search has been identified in the literature as a spatial exploratory process demanding spatial resources, the presence or the absence of the visual search requirement created two levels of

demand on the spatial resource (for a detailed description of the rationale and literature review, see [32]). Not surprisingly, for the dual task conditions when target search was not involved, the experimental result was similar to that of the first study. A particularly interesting finding was that on the other half of the trials when target search was required to perform the tasks, the increased demand of the visual search task for spatial resources produced a significantly greater increase in the interference between the spatial decision and the tracking task than that between the verbal decision and the tracking task.

The third study extended the first two by expanding the range of task difficulty levels. The study consisted of two experiments [29], both of which required the subjects to perform a primary first-order pursuit tracking task and a secondary spatial or verbal task, which was an easy information acquisition task in the first experiment and a difficult information integration task in the second experiment. The task demand for spatial resources was further manipulated by creating four levels of demand for target searching--the relevant information that was needed for the decision tasks was displayed with 4 levels of uncertainty regarding its spatial location on the visual display. The four levels corresponded to the task conditions in which the relevant information was displayed at one of 1, 2, 4, or 8 locations on each trial with equal probability.

In the first experiment in which an easy first-order tracking task was time-shared with an easy information acquisition task, there was no evidence of task interference in almost all conditions. The only evidence of a significant task interference was found when the demand for target search reached the highest level (N=8). However, in the second experiment in which a difficult information integration task was employed, significant task interference was observed in all dual task conditions. Furthermore, increases in the demand for spatial search produced a faster increase in the interference between the tracking and the spatial task than between the tracking and the verbal task.

These observations and experimental results pose a challenge to both the queueing theoretic models and the multiple resources models of multitask performance. As mentioned

above, the queueing theoretic models predict that task interference should be determined by the level of difficulty of the concurrent tasks, regardless of whether they use the same or separate processing codes. The most logical prediction of the multiple resources models is that the performance/difficulty tradeoff should be observed only if concurrent tasks use the same processing code.

Since the queueing theoretic models have not attempted to address the parallel and structural aspects of task demand, it is not clear how the models would account for the differential effects of concurrent spatial and verbal tasks. A natural approach to this problem from the single channel perspective might be to develop a set of complicated scheduling algorithms that allow the single server to vary task priorities and service rates according to the processing codes of the to-be-processed tasks. But it is not clear whether this is an intuitive or feasible approach, and the problem remains until the algorithms are developed.

A concept that might be invoked by the resource theorists in explaining the finding of between-codes interference is the concept of concurrence cost, which says that the act of time-sharing itself requires resources and thus produces worse performance than single task conditions. But this concept by itself appears limited in this context, since it fails to explain why between-codes interference could be absent in some situations, but present in other situations, and perhaps more importantly, why between-codes interference could show a pattern of gradual increase as seen in a performance/difficulty tradeoff. Furthermore, the concept itself does not address the more fundamental question of why the concurrence costs exist.

While it remains to be seen how the single channel and the multiple resources models would address these issues, the next section proposes a three-node queueing network model as a plausible and intuitive account of the research findings regarding interference between spatial and verbal concurrent tasks. As described below, this model provides a computational framework to integrate the concerns of the single channel and the multiple resources models, which, in essence, can be treated as special cases of the queueing network model. Thus the

model does not in any way obviate the explanatory value of the fundamental concepts of the two classes of models.

V. A 3-NODE QUEUEING NETWORK MODEL OF MULTITASK PERFORMANCE

Recently, Liu proposed that human multi-task information processing system and human-computer systems are, in many respects, analogous to queueing networks (examples of which include telephone communications systems, computer networks and road traffic networks, and industrial production systems) [29] [30] [31]. Queueing network methods of performance analysis and systems modeling employed widely in industrial engineering and systems analysis may provide an integrated computational framework for modeling multitask performance and human-computer systems. In general, human information processing system can be modeled as a network of information processing nodes (called servers), with each node representing a service facility of some kind. These nodes do not have to be, and many are not, of the same kind. The nodes are connected by arcs over which information processing tasks (called customers) flow without delay. Each node has a waiting space for customers to wait if their service demand cannot be immediately satisfied, and thus each node can have a queue of customers formed in front of it, and multiple queues may exist simultaneously in the system. Queueing networks support the modeling of a wide range of complex structural and temporal arrangements that multiple tasks might assume. The structural arrangements include both serial selection and parallel execution, and the temporal arrangements include both immediate activities and delayed processing.

In order to give mathematical substance to the model presented below, we introduce the following notations, 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 service station, and the second for stochastic processes in a queueing network.

A queueing process at a service station 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 station, D the restriction on waiting room capacity in front of the station, and E the queue discipline (the manner by which the customers are selected from the queue for service). For the most part, this article will focus on the class of queueing process that has received most research attention and enjoyed a most fruitful history of producing usable analytical results. This queueing process is denoted as M/M/c/∞/FCFS (or M/M/c for short), representing a queueing process with exponential interarrival times (also called Poisson arrival), exponential service times, c identical servers at a station, no restriction on the maximum number of customers allowed in the queue, and first-come, first-served queue discipline. The importance and justifications of employing this type of queueing process in performance modeling are discussed in all standard textbooks on queueing theory [20] [27].

We use the following symbols to represent a queueing network:

1) K: number of nodes,

2) i: identity of nodes,

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

4) pij: the probability that a customer visits node j immediately after departing from node i (also called 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,

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

6) μ_i: mean service rate of node i.

We assume that we will mainly be concerned with queueing networks with the following characteristics:

1. Arrivals from the "outside" to node i follow a Poisson process with mean rate γi,
2. Service rates for each node are independent of the arrival rates and are mutually independent of each other, and the service rate for node i is exponentially distributed with parameter μi,
3. The routing probabilities (pij's) are independent of the state of the system, which represents the number of customers at each node.

Networks that have these properties 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: the network *acts as if* each node can be viewed as an independent M/M/c queue, with parameters λi and μi. The joint probability distribution for the number of customers at each node can be written as a product of marginal M/M/c's [16] [23]. This amazing property makes it possible to derive many important results for the Jackson network that are often not available or analytically intractable for other types of networks. Jackson networks have subsequently received the most research attention and enjoyed a great success in model development. The models have also been successfully applied in diverse areas, because separable networks can be evaluated quite efficiently. Furthermore, many authors have demonstrated that many of the results for Jackson networks provide close approximations to non-Jacksonian networks [5] [16]. In computer system analysis, the pragmatic, "operational" framework for queueing network analysis, pioneered by Buzen and Denning, relies heavily on the assumption of separable queueing 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 (e.g., inadequate measurement data) [8] [15].

It is not the purpose of the present article to fully explicate the formalisms and the modeling capabilities of the queueing network methods. Rather, the focus of this article is on the gap between the single channel, queueing theoretic models and the parallel allocation,

multiple resources models discussed in the previous sections. For this purpose, the present article proposes a simple, three-node queuing network as shown in Figure 1. This queuing network model has a parallel processing component and a serial processing component. The parallel processing component refers to the parallel operation of a "spatial server" (S) and a "verbal server" (V), analogous to the spatial and verbal processing mechanisms or resources advocated in the multiple resources theory. The serial processing component refers to the serial operation from either S or V to the third server, which can be referred to as the central server (C).

Insert Figure 1 about here

The model assumes that a task is composed of a number of task components (called customers), and the mean arrival rate of the task components, λ (the number of arriving customers per unit of time), is a measure of the difficulty or the service demand of a task. The model assumes that spatial and verbal task components take separate routes of the network: all spatial task components must be serviced by the server S, and all verbal task components must be serviced by the server V, before they receive the required service of the server C prior to their completion. The capacity of a server in meeting the service needs of the arriving tasks can be represented as the mean service rate of the server (μ), which is the number of customers that can be serviced per unit of time. For the purpose of the present discussion, the model assumes that all servers are single channel servers with constant service rates. This is consistent with one of the most widely adopted assumptions in engineering models and psychological theories of human performance.

An important property of separable queueing networks is that the total arrival rate into any node (from outside and from other nodes) satisfies the "traffic equation": $\lambda i = \gamma i + \sum_{j=1}^{K}(p_{ji}\lambda_j)$. Therefore, for the 3-node network, the arrival rate at the central server is

the sum of the arrival rates at the spatial and the verbal servers (i.e., $\lambda 3 = \lambda 1 + \lambda 2$, or equivalently, $\lambda c = \lambda s + \lambda v$). Using the symbols introduced earlier, the essential constituents of the 3-node queueing network can be formally represented as follows:

1) K=3,

2) i=1, 2, 3, representing the spatial (S), verbal (V), and the central (C) server, respectively,

3) $\gamma i = \lambda i$, for i = 1, 2

$= 0$, for i = 3

4) $\lambda i = \lambda i$, for i = 1,2

$= \lambda 1 + \lambda 2$, for i=3

5) $p12 = p21 = p31 = p32 = 0$,

$pii = 0$, for $\forall i$,

$p13 = p23 = 1$

6) $p10 = p20 = 0$, $p30=1$

The following discussion will use i=s, v, c, rather than i=1, 2, 3, to refer to the spatial, verbal, and central servers. This is simply for the purpose of making the description and understanding easier.

A number of performance measures can be computed using queueing network methods. Of most interest to the present analysis is the waiting time of the customers in front of each server. Customer waiting time is a stochastic variable that describes the amount of time that a customer has to wait in front of a server before receiving its service. As in the existing queueing theoretic models of human performance, the queueing network model assumes that there is a performance cost associated with delaying service to a task (called the cost of waiting). Performance decrement due to delayed service at a busy server is assumed to be determined jointly by the waiting time and the cost of waiting in front of the server as described in the following expression:

$$PDi = Wi \times Ci \quad (1)$$

where

PDi is task performance decrement due to delayed service at server i,

Wi is the customer waiting time in front of server i,

Ci is the cost of waiting at server i.

As seen in Figure 1, customer waiting could occur either in front of S or V or both servers (similar to the predictions of the multiple resources models), or only in front of C (similar to the predictions of the single channel or queueing theoretic models), or a combination of both cases. More specifically, two spatial tasks will interfere with each other if they total service demand on the spatial server is high. Similarly, two verbal tasks will interfere with each other if they total service demand on the verbal server is high. This interference will be progressively greater as their competition for the service of the same spatial (or verbal) server increases. This is what is predicted by the multiple resources model. However, it is easy to see from the 3-node model that a spatial task and a verbal task could also interfere with each other when their total service demand for the central server is high, although they do not compete with each other for the service of the spatial or the verbal server. This interference would also show a progressive increase as their total demand at the central server increases, regardless of whether the tasks use the same or separate processing codes. This is consistent with the predictions of the queueing theoretic, single channel models. It can be seen that, in essence, the multiple resources model focuses on the role of spatial and verbal servers, whereas the single channel models focus on the role of the central server.

For separable queueing networks, customer waiting time in front of a server satisfies the following equation [5] [20]:

$$W_i = 1/(\mu_i - \lambda_i) - 1/\mu_i \qquad (2)$$

where

W_i is the mean customer waiting time in front of server i,

μ_i is the mean service rate of server i,

λ_i is the mean customer arrival rate at server i.

This relationship between waiting time and arrival rate for servers with constant service rates is illustrated in Figures 2. It can be seen that waiting time increases monotonically as the arrival rate increases. Waiting time approaches infinity as the mean arrival rate approaches the mean service rate, indicating a situation in which the tasks are too difficult to be performed simultaneously (beyond the processing limit). Of most interest to the model is the change in performance when λ is smaller than μ at each server.

By making necessary assumptions about the capacities of each server (i.e., the μi's) and the cost of waiting in front of each server (i.e., the Ci's), equations (1) and (2) allow us to model a variety of task interference patterns, considering both the difficulty and the processing codes of concurrent tasks. The following discussion will use the pattern of task interference discussed in the previous section as an example to illustrate this point. A way to model that pattern of interference is to assume that the serial processing bottleneck (server C) has a capacity that is smaller than the sum of the capacities of the spatial and the verbal servers, but greater than the capacity of each server. That is,

$\mu_C < \mu_S + \mu_V$,

$\mu_C > \mu_S$, and

$\mu_C > \mu_V$.

As a numerical demonstration, we could assume that μ_C=14, μ_S=10, μ_V=10. In this case the relationships between waiting time and arrival rate at the three servers are shown in Figure 2 and Figure 3. If we further make the simplest assumption that the costs of waiting at the three servers are identical constants (i.e., Ci=C, for i=1, 2, 3, and C is a constant), then the relationship between performance decrements due to delayed service at a server (PDi) and the total task demands at that server (λi) should show the same pattern as that between waiting time and arrival rate, shown in Figure 2 and Figure 3.

Insert Figure 2 and Figure 3 about here

Based on Figure 2 and Figure 3, we can distinguish four important cases of customer waiting (task interference):

Case 1: When the customer arrival rate at each of the centers is significantly smaller than the service rate (i.e., $\lambda i << \mu i$, for $\forall i$), minimal waiting is expected, and thus performance decrements would be minimal in performing the simultaneous tasks. This is the case when easy tasks are time-shared with each other, regardless of the processing code involved. For example, suppose two spatial tasks are performed concurrently with each other. One of the two spatial tasks (e.g., driving on an easy road) has a spatial task demand corresponding to a mean arrival rate of spatial task components of 3 ($\lambda_{S1}=3$). The concurrent spatial task has a task demand corresponding to a mean arrival rate of 2 to the spatial server ($\lambda_{S2}=2$). The total demand of the two tasks is 5 ($\lambda_S=\lambda_{S1}+\lambda_{S2}=5$) both at the spatial and at the central server, which is significantly smaller than the capacities of each server ($\lambda_S << \mu s = 10$, and $\lambda_S << \mu c = 14$). Minimal waiting would be observed in front of the two servers, and thus no interference would be observed between the two easy spatial tasks.

Case 2: When the customer arrival rate approaches the service rate of either the spatial (S) or the verbal server (V) but not that of the central server (C), significant waiting is expected in front of S or V, respectively, but not in front of C. For example, when one of the two spatial tasks mentioned in Case 1 has increased its difficulty to $\lambda s1=6$, then we have $\lambda s=\lambda s1+\lambda s2=8$, which is close to μs and causes significant within-code interference at the spatial server. But since the arrival rate at the central server is still much smaller than its service rate ($\lambda c=8 << \mu c=14$), minimal waiting would be expected in front of the central server. Performance bottleneck is at the spatial rather than the central server. This is the case in which within-code interference is the only source of task interference. This within-code interference will increase quickly as the arrival rate at the congested server (the spatial server in the present example) continues to increase.

Case 3: When the customer arrival rate approaches the service rate of the central server but not that of the spatial or the verbal server, customers are expected to wait in front of C, but

not in front of S or V (e.g., when $\lambda s=6$, $\lambda v=6$, $\lambda c=\lambda s+\lambda v=12$). Since λc is close to μc, between-codes interference at the central server would be the primary source of task interference. Progressive increase in λs or λv would produce progressive increase in λc, and produce a corresponding progressive increase in the between-codes interference. However, as long as λs and λv are both still much smaller than μs and μv, within-code interference would be minimal. The central server is the bottleneck.

Case 4: This is the most general case, in which performance decrements are caused by a combination of within-code and between-codes interferences--a combination of Case 2 and Case 3. Due to the heavy demands of the tasks, significant waiting could occur in front of two or all three servers. For example, when $\lambda s=8$ and $\lambda v=5$, within-code and between-codes interferences would both occur due to congestion in front of both the spatial and the central server.

To use an observable system as an analogy, we can imagine a roadway transportation system with a configuration like the 3-node network of Figure 1. One or two types of vehicles pass through the network: trucks and cars. All "trucks" have to pass through a "truck" section, which has, for example, 4 parallel lanes. All "cars" have to pass through a "car" section of 4 parallel lanes, and all vehicles must pass through a "final" section of, say, 6 parallel lanes. Congestion and traffic delay could be observed at any one, or any two, or all three sections, depending on the arrival rates of trucks and cars. The different types of traffic delay are similar to the several cases discussed above.

The 3-node queueing network model provides a conceptual and computational framework for modeling a much broader range of task interferences than what has been described thus far. The following is a discussion of some interesting cases.

First, in the analysis presented above, we have assumed that the capacity of the spatial server is the same as that of the verbal server (i.e., $\mu_S=\mu_V$). This assumption implies that a pair of concurrent spatial tasks will interfere with each other in the same way as a pair of verbal tasks, as long as the service demands of the spatial tasks on the spatial server are the same as

the service demands of the verbal tasks on the verbal server. This tentative assumption is based on the current state of knowledge in that there does not exist an empirical database for comparing the capacities of the two servers. Although ample evidence has demonstrated that within-code interference is likely to be greater than between-codes interference, no data is available that has compared within-code interference between spatial tasks with that between verbal tasks while keeping the difficulty levels of the two types of tasks equivalent. Thus it is not clear which of the two types of tasks is more likely to produce within-code interference.

This tentative assumption about the equivalence of the capacities of the spatial and the verbal servers suggests an interesting research question rather than offers a final conclusion. Future experimental results along this direction would undoubtedly provide deeper insights into this assumption, and can be easily accommodated by the model. For example, suppose that future research concludes that human cognitive system is more capable of processing multiple concurrent spatial tasks than verbal tasks of the same difficulty (or vice versa), then an intuitive and easy way to modify the model would be to assume that the spatial server has a greater capacity than the verbal server (or vice versa).

Second, we have thus far adopted the simplest assumption that the cost of waiting in front of the servers are identical and are all constants. This assumption implies that the relationship between performance decrements due to delayed service at a server and the total task demands at that server is the same as the relationship between waiting time and arrival rate as shown in Figure 2 and Figure 3. Under this assumption, performance decrement increases slowly at the lower end of difficulty levels and rises sharply as the difficulty levels approach the processing limits. Although this pattern of performance decrements appear to be consist with many task situations, it is conceivable that other relationship may emerge in some other task situations. For example, some task situations may show a linear or a negatively accelerated data pattern, rather than the positively accelerated data pattern shown in Figure 2 and Figure 3. One way to model these relationship patterns is to make corresponding changes in the model's assumption about the cost of waiting.

Third, the 3-node model allows us to model the effects of task selection and performance strategy on multitask performance. Our discussions thus far have assumed that the central server (C) gives equal priority to spatial and verbal tasks. However, a number of experimental investigations have demonstrated that subjects can prioritize concurrent tasks and allocate their attention flexibly between tasks in any proportion desired (e.g., [18] [44] [63]). This is also clearly consistent with our intuition and everyday experience. For example, we could give a higher priority to a primary spatial task than to a concurrent verbal task (e.g., driving is more important than talking to a passenger). As task demand increases, we could keep the performance of the spatial task at a relatively constant level at the expense of the performance on the verbal task.

An approximative approach to model this aspect of multitask performance is to allow the central server to have a faster service rate for the high-priority task and a slower service rate for the low-priority task than the equal-priority situation, while keeping the mean service rate (μc) as a constant defined for the equal-priority situation. The task with a higher priority would thus have a shorter waiting time and a smaller performance decrement than if it were not given a higher priority. Since μc is assumed to be a constant, an increase in the service rate of a task can only be realized by a corresponding decrease in the service rate for the concurrent task that competes for the same server. This results in a deteriorated performance on the low-priority task. The relative priority level between two tasks competing the same server would determine the difference between the service rates they receive.

Fourth, although our discussion thus far has grouped all task components of the same processing code as one class of customers, the model can be extended to consider more than one class of spatial (or verbal) customers when needed. For example, in order to analyze the performance of two concurrent spatial tasks with different priority levels, the model could assume that the two tasks represent two classes of spatial customers. The spatial server provides service to the two classes of spatial customers in the same way as what has been discussed thus far about how the central server serves spatial and verbal customers. When the

total service demand at the spatial server approaches its capacity, the spatial server may behave as if it could give a higher service rate to the higher-priority spatial task so as to maintain its performance level, while slowing down its service to the low-priority spatial task so as to keep its mean service rate as a constant.

VI. DIRECTIONS FOR FUTURE RESEARCH

The 3-node queueing network model presented in this article integrates the concerns of the queueing theoretic and the multiple resources models in a larger computational modeling framework and presents the qualitative theories of multiple resources in engineering terms. The model provides a formal mechanism to consider both the serial bottleneck and the parallel allocation aspects of multitask performance. The model is simple, but it is integrative in that it unifies the concerns of isolated models into a larger modeling framework, it is computational in that it supports quantitative predictions and tradeoff analysis, it is intuitive in that it is quite consistent with everyday experiences, and it is testable in that its assumptions can be subjected to empirical investigations.

In addition to the concept of attention switching employed by serial processing theories and the concept of resource sharing employed by parallel processing theories, a number of other important concepts have been proposed in the multitask literature. These concepts attempt to account for aspects of multitask performance that can not be readily explained by the concepts of serial processing and resource sharing. One example of these important concepts is the degree of similarity between two concurrent tasks. Ample experimental evidence has demonstrated that task similarity can sometimes improve dual-task performance through task cooperation and integration, and sometimes degrade it through task confusion and cross-talk, which has also been labeled as outcome conflict, referring to the situation in which stimuli for one task activate responses for a different task (see, for example, [60]). Along another line of investigation, Wickens and Liu have analyzed the role of visual scanning and auditory preemption in multitask performance, together with an

analysis of the concepts of resources, switching, and task integration [61]. Visual dominance, perceptual and cognitive tunneling, multitask management and scheduling, and the concept of automaticity in skill acquisition and training are also examples of important concepts in the multitask literature [13].

While these concepts are all of critical importance to multitask modeling, there are at least two issues that need substantial further research. First, these concepts are relatively isolated from each other, each focusing on one aspect of the whole picture, and second, they are not presented in a computational form. What is needed is an integrated computational framework that unifies these concepts and represents them in computable forms. The queueing network approach of modeling multitask performance offers one possible means for achieving this goal. The 3-node queueing network model represents one step toward this goal.

One of our current research activities is aimed at developing a queueing network model of multitask performance at a more general level. The model would integrate more of the other currently isolated concepts into a unified computational framework. In this general queueing network model of multitask performance, information processing tasks may enter the information processing system at some node, traverse from node to node in the system, and depart from some node, not all tasks necessarily entering and leaving at the same nodes (e.g., different sensory and motor modalities), or taking the same path once having entered the system (separate processing routines or demands). Tasks may return to nodes previously visited (e.g., performance feedback), skip some nodes entirely (e.g., skill acquisition and automaticity), and even remain in the system for a long time (e.g., memory rehearsal). Some customers may failed to enter a busy system (e.g., perceptual tunneling), leave a busy system before they have been fully serviced (a cause for imperfect performance), jockey for position by switching from one queue to another, or preempt earlier customers if the queue discipline allows this to happen (task scheduling). Multiple queues may improve their joint performance by adopting some coordinated service schemes (e.g., task integration), or lose effective communication in the face of overwhelming information (e.g., confusion, cross-talk, and

outcome conflict). We are currently developing this conceptual description into a testable computational model that is able to predict and explain empirical data.

The prospect for future research and application of queueing network methods in human performance modeling is exciting. It should be noted here that the value of queueing network methods is not limited to the level of multitask modeling and analysis as presented here. In fact, a very important characteristic of queueing network methods is that they support the modeling of human performance and human-computer systems at various analysis levels. For example, at the most microscopic level of human performance analysis, a queueing network model for reaction time has been proposed to address one of the most fundamental and enduring questions in psychological theory: why is there a delay between stimulus presentation and response initiation? As a continuous-flow network model of elementary mental processes, the queueing network model for reaction time includes existing discrete serial-stage and continuous-flow overlapping stage models as well as discrete critical path network models of reaction time as special cases [31].

At the macro-level of engineering applications, a number of queueing network models of human-computer interaction have been developed for both standalone and network environments [30]. The models consider a network of *distinct* human operators and *distinct* computers organized with complicated functional allocation schemes and processing delays, which cannot be modeled by the queueing theoretic or discrete network approaches. The models can be used to answer a number of important system design questions by computing related performance measures of the network. The model can also be extended to model competitive as well as cooperative relations among components in a human-computer network and to predict task processing times in complicated human-computer organizations. Since queueing network methods have been extensively applied to performance analysis of computer systems, a desirable feature of queueing network models is that they allow us to model the human and the computer components in a human-computer system with the same modeling language in a cohesive framework.

REFERENCES

[1] M. Adams, Y. Tenney and R. Pew, *Strategic workload and the cognitive management of advanced multi-task systems*, CSERIAC SOAR Report 91-6, 1991.

[2] A. D. Baddeley, *Human memory: Theory and practice*, Boston: Allyn and Bacon, 1990.

[3] S. Baron, D. Kruser and B. Huey (Eds.), *Quantitative modeling of human performance in complex, dynamic systems*, Washington, DC.: National Academy Press, 1990.

[4] S. Baron, G. Zacharias, R. Muralidharan and R. Lancraft, "PROCRU: a model for analyzing flight crew procedures in approach to landing," NASA-CR-1523397, 1980.

[5] O. Boxma and H. Daduna, "Sojourn times in queueing networks," In H. Takagi (Ed.), *Stochastic analysis of computer and communications systems*, pp. 401-150. North Holland, 1990.

[6] D. E. Broadbent, *Perception and communication*, London: Pergmon, 1958.

[7] L. R. Brooks, "Spatial and verbal components in the act of recall," *Canadian Journal of Psychology*, vol. 22, pp. 349-368, 1968

[8] J. Buzen, "Fundamental operational laws of computer system performance," *Acta Informatica*, vol-7, pp.167-182, 1976.

[9] J. Carbonell, "A queueing model of many-instrument visual sampling," *IEEE Trans. HFE-4*, pp. 157-164, 1966.

[10] Y. Chu, and W. Rouse, "Adaptive allocation of decision making responsibility between human and computer in multi-task situations," *IEEE Trans. SMC-9*, pp. 769-778, 1979.

[11] G. P. Chubb, N. Stodolsy, W. D. Fleming and J. A. Hasson, "STALL: A simple model for workload analysis in early system development," In *Proceedings of the 31st Annual Meeting of the Human Factors Society*, 1987.

[12] K. W. J. Craik, "Theory of the human operator in control systems I: The operator as an engineering system," *British Journal of Psychology*, vol. 38, pp. 56-61, 1947

[13] D. Damos (Ed.), *Multiple task performance*, London: Taylor and Francis, 1992.

[14] J. A. Deutsch and D. Deutsch, "Attention: Some theoretical considerations," Psychological Review, vol. 70, pp. 80-90, 1963.

[15] P. Denning and J. Buzen, "The operational analysis of queueing network models," *Computing Surveys*, vol-10, pp. 225-261, 1978

[16] R. Disney, and D. Konig, "Queueing networks: A survey of their random processes," *SIAM Review*, vol-27, 335-403, 1985.

[17] R. Edwards, R. Curnow and R. Ostrand, *Workload assessment model (WAM) user's manual (Report D180-20247-3)*, Seattle, WA: Boeing Aerospace Co., 1977.

[18] D. Gopher, M. Brickner and D. Navon, "Different difficulty manipulations interact differently with task emphasis: Evidence for multiple resources," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 8, pp. 146-158, 1982.

[19] J. Greenstein, and W. Rouse, "A model of human decision making in multiple process monitoring situations," IEEE Trans. SMC-12, pp. 182-193, 1982.

[20] D. Gross and C. Harris, *Fundamentals of queueing theory*, Wiley, 1985.

[21] S. Harris, J. Owen and R. A. North, "A system for the assessment of human performance in concurrent verbal and manual control tasks," *Behavioral Research Methods and Instrumentation*, vol. 10, pp. 329-333, 1987.

[22] S. G. Hart and I. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," In P. A. Hancock and N. Meshkati (Eds.), *Human mental workload*, Amsterdam: North Holland, 1989.

[23] J. Jackson, "Networks of waiting lines," *Operations Research, vol-5*, pp. 518-521, 1957.

[24] D. Kahneman, *Attention and effort*, Englewood Cliffs, NJ: Prentice Hall, 1973.

[25] S. T. Klapp and A. Netick, "Multiple resources for processing and storage in short-term working memory," *Human Factors*, vol. 30, pp. 617-632, 1988.

[26] D. L. Kleinman and R. E. Curry, "Some new control theoretic models for human

operator display monitoring," *IEEE Trans. SMC-7*, pp. 778-784, 1977.

[27] L. Kleinrock, *Queueing systems*, Vol. 1 and Vol. 2. Wiley, 1976.

[28] K. Laughery, "Micro SAINT: A tool for modeling human performance in systems," In G. McMillan et al., (Eds.), *Applications of Human Performance Models to System Design*, pp. 219-230, New York: Plenum, 1989.

[29] Y. Liu, "Visual scanning, memory scanning, and computational human performance modeling," *Proc. of the Human Factors Society 37th Annual Meeting*, pp. ???, 1993.

[30] Y. Liu, "Queueing networks as models of human performance and human-computer systems," In *Proc. of the 1994 Symposium on Human Interaction with Complex Systems,* pp. 256-270, Greensboro, NC., 1994.

[31] Y. Liu, "Queuing network modeling of elementary mental processes," Manuscript submitted to *Psychological Review*, 1994.

[32] Y. Liu and C. Wickens, "Visual scanning with or without spatial uncertainty and selective and divided attention," *Acta Psychologica*, vol. 79, pp. 131-153, 1992.

[33] T. B. Malone, M. Kirkpatrick and W. H. Kopp, "Human factors impact of system workload and manning levels," In *Proc. of the 30th Human Factors Society Annual Meeting*, 1986

[34] J. H. McCracken and T. B. Aldrich, "Analysis of selected LHX mission functions: Implications for operator workload and system automation goals," Technical Note ASI 479-024-84(B), Anacapa Sciences Inc., 1984.

[35] N. Moray, "Where is attention limited? A survey and a model," *Acta Psychologica*, vol. 27, pp. 84-92, 1967

[36] N. Moray, "Monitoring behavior and supervisory control," In K. R. Boff, L. Kaufman, and J. P. Thomas (Eds.), *Handbook of perception and human performance, vol-II: Cognitive processes and performance*, New York: Wiley, 1986.

[37] N. Moray, M. Dessouky, B. Kijowski, and R. Adapathya, "Strategic behavior, workload, and performance in task scheduling," *Human Factors*, vol. 33, pp. 607-629,

1991.

[38] R. North and V. Riley, "W/INDEX: A predictive model of operator workload," In G. McMillan et al., (Eds.), *Applications of Human Performance Models to System Design,* pp. 81-89, New York: Plenum, 1989.

[39] K. R. Pattipati and D. Kleinman, "A review of the engineering models of information-processing and decision-making in multi-task supervisory control," In Damos, D. (Ed.), *Multiple task performance*, pp. 35-68, London: Taylor and Francis, 1992.

[40] K. R. Pattipati, D. L. Kleinman and A. R. Ephrath, "A dynamic decision model of human task selection performance," *IEEE Trans. SMC-13*, pp. 145-166, 1983.

[41] W. B. Rouse, "Human-computer interaction in multitask situations," *IEEE Trans. on SMC-7*, 384-391, 1977.

[42] W. B. Rouse, *Systems engineering models of human-machine interaction*, New York: Elsevier North-Holland, 1980.

[43] D. Schmidt, "A queueing analysis of the air traffic controller's workload," *IEEE Trans. SMC-8*, pp. 492-493, 1978.

[44] W. Schneider and A. D. Fisk, "Concurrent automatic and controlled visual search: Can processing occur without cost?" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 8, pp. 261-278, 1982.

[45] J. W. Senders, "The human operator as a monitor and controller of multidegree freedom systems," *IEEE Trans. on HFE, HFE-5*, pp. 2-5, 1964.

[46] J. W. Senders and M. Posner, "A queueing model of monitoring and supervisory behavior," In T. Sheridan and G. Johannsen (Eds.), *Monitoring Behavior and Supervisory Control,* New York: Plenum, 1976.

[47] T. Sheridan, "On how often the supervisor should sample. IEEE Trans., SSS-6, pp. 140-145, 1972.

[48] A. I. Siegal and J. J. Wolf, *Man-machine simulation models*, New York: Wiley, 1969.

[49] G. Sperling and B. A. Dosher, "Strategy and optimization in human information

processing", In K. Boff, L. Kaufman and J. Thomas (eds.), *Handbook of perception and performance*, vol. 1, New York: Wiley.

[50] M. Strieb, N. Lane, F. Glenn and R. J. Wherry, "The human operator simulator: An overview," In J. Moraal and K. Kraiss (eds.), *Manned system design: Methods, equipment, and applications*, New York: Plenum Press, 1981.

[51] C. W. Telford, "Refractory phase of voluntary and associate response," *Journal of Experimental Psychology*, vol. 14, pp. 1-35, 1931.

[52] P. S. Tsang and R. A. Rothschild, "To speak or not to speak: A multiple resource perspective," In *Proc. of the 29th Annual Meeting of the Human Factors Society* , 1985.

[53] M. K. Tulga, "Dynamic decision and workload in multitask supervisory control," *IEEE Trans. SMC-10*, pp. 217-232, 1980.

[54] M. K. Tulga and T. Sheridan, "Dynamic decisions and workload in multitask supervisory control," IEEE Trans., SMC-10, 217-232, 1980.

[55] R. Walden and W. Rouse, "A queueing model of pilot decisionmaking in a multitask flight management situation," *IEEE Trans. SMC-8*, pp. 867-875, 1978.

[56] A. T. Welford, "Single channel operation in the brain," *Acta Psychologica*, vol. 27, pp. 5-21, 1967.

[57] R. J. Wherry, The human operator simulator-HOS," In T. B. Sheridan and G. Johannsen (eds.), *Monitoring behavior and supervisory control*, New York: Plenum Press, 1976.

[58] C. D. Wickens, "The effects of divided attention on information processing in tracking," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 1, pp. 1-13, 1976.

[59] C. D. Wickens, "Processing resources in attention," In Parasuraman, R. and D. Davis (Eds.), *Varieties of attention*, New York: Academic Press, 1984.

[60] C. D. Wickens, "Processing resources and attention," In D. Damos (Ed.), *Multiple task performance*, London: Taylor and Francis, pp. 3-34, 1992.

[61] C. D. Wickens and Y. Liu, "Codes and modalities: A success and a qualification,"

Human Factors, vol. 30, pp. 599-616, 1988.

[62] C. D. Wickens, S. J. Mountford and W. Schreiner, "Time-sharing efficiency: Evidence for multiple resource, task hemispheric integrity and against a general ability," *Human Factors*, vol. 23, pp. 211-229, 1981.

[63] C. D. Wickens, D. Sandry, and M. Vidulich, "Compatibility and resource competition between modalities of input, central processing, and output: Testing a model of complex task performance," *Human Factors*, vol. 25, pp. 227-248, 1983.

[64] C. D. Wickens and A. Weingartner, "Process control monitoring: The effects of spatial and verbal ability and current task demand," In R. Eberts and C. Eberts (eds.), *Trends in ergonomics and human factors*, North Holland, 1985.

List of Figures

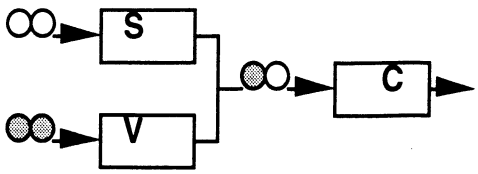

Figure 1: A 3-Node Queueing Network Model of Human Performance of Concurrent Spatial and Verbal Tasks

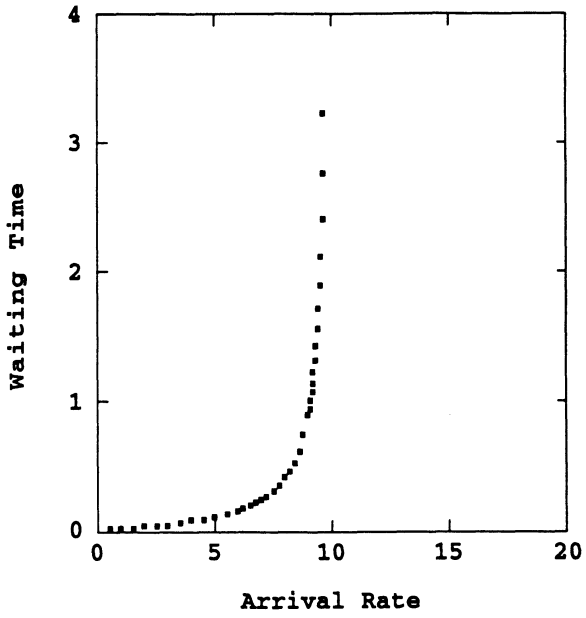

Figure 2: Relationship between waiting time and arrival rate at the spatial or the verbal server

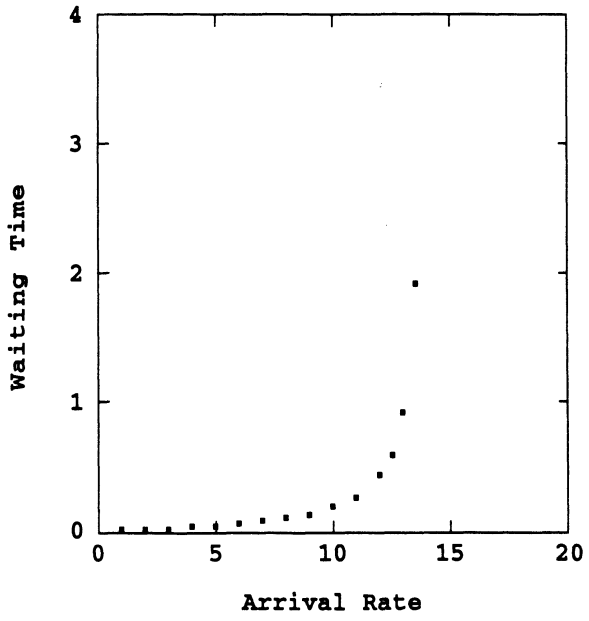

Figure 3: Relationship between waiting time and arrival rate at the central server