Analytic Queuing Network

Henry Neimeier
The MITRE Corporation
7525 Colshire Drive
McLean, Virginia, 22102, USA

A new simulation paradigm is proposed to overcome several of the limitations of discrete event simulation. It is based on the combination of analytic queuing networks and analytic uncertainty modeling. The analytic queue technique gives an approximate transient solution to the general inter arrival time and general service time single server queue. Analytic uncertainty analysis is based on the beta distribution. It provides the entire uncertainty probability distribution vice an uncertain estimate of the mean. The beta distribution can be fit based on the minimum, mean, maximum, and standard deviation statistics. In a complex results calculation, all that is required is to keep track of these statistics as the calculation proceeds. At any point in a calculation, the probability distribution of the result can be derived by fitting a beta distribution based on the four statistics. When analytic queuing is combined with analytic uncertainty, modeling dynamic uncertainty analysis becomes feasible. The time varying uncertainty distribution in resulting measures of effectiveness can be calculated at any specified time or over any user specified time interval. This new capability is not available in discrete event simulation.

Analytic Queuing Network

Henry Neimeier

Overview

A new simulation paradigm is proposed to overcome several of the limitations of discrete event simulation. It is based on the combination of analytic queuing networks and analytic uncertainty modeling. The analytic queue technique gives an approximate transient solution to the general inter arrival time and general service time single server queue. Analytic uncertainty analysis is based on the beta distribution. It provides the entire uncertainty probability distribution vice an uncertain estimate of the mean. The beta distribution can be fit based on the minimum, mean, maximum, and standard deviation statistics. In a complex results calculation all that is required is to keep track of these statistics as the calculation proceeds. At any point in a calculation, the probability distribution of the result, can be derived by fitting a beta distribution based on these four statistics. When analytic queuing is combined with analytic uncertainty, modeling dynamic uncertainty analysis becomes feasible. The time varying uncertainty distribution in resulting measures of effectiveness can be calculated at any specified time or over any user specified time interval. This new capability is not available in discrete event simulation.

Discrete event simulation can be viewed as a network of interrelated queues. Thus the new paradigm has wide applicability. Its deterministic solution greatly simplifies sensitivity and uncertainty analysis of complex many parameter models. The factor effects in a many factor model are very difficult to obtain in discrete event simulation since they are masked by the stochastic simulation uncertainty. In discrete event simulation the causal chain between input parameter change and resulting output measure effect is broken. Due to the stochastic random number generation process, many runs could be required to see the effect of an input parameter change. In training simulations this random learning effect can be a problem. Our proposed deterministic technique overcomes this problem.

Discrete event simulation requires a time period of many simulation events to determine sample statistics. Our technique provides instantaneous statistics. This feature, and the analytic uncertainty analysis technique, makes dynamic uncertainty anlysis feasible. A simple three node queuing network example is presented. The network has a time varying uncertain input workload. The time varying total processing delay distribution output is calculated using the proposed technique.

Discrete Event Simulation Times

Discrete simulation requires multiple long simulation runs to obtain a statistically significant point estimate. The different result values, from multiple runs with identical parameter values but different random number seeds, are averaged to obtain the point estimate of the mean result value. Conversely, the analytic solution gives the entire resulting probability distribution with minimal calculation. The analytic solution also considerably

simplifies sensitivity analysis. A single analytic run is done for each parameter setting vice multiple runs for a statistically significant result in discrete event simulation. The simulation time (T) required to be 95 percent confident in a relative error (ε) is approximated by the following equation for open GI/G/1 (general independent inter arrival times, general service times, 1 server) queuing networks:

$$T = 8\,\tau\,(C_a^2 + C_s^2)\,Z^2 / (\rho^2(1-\rho^2)\,\varepsilon^2)$$

Where:

- T = simulation time for a specified relative error
- τ = service time
- C_a^2 = square coefficient of variation in inter arrival time (variance in inter arrival time divided by the mean inter arrival time squared)
- C_s^2 = square coefficient of variation in service times
- Z = unit normal deviate (Z=2 for 95 percent confidence
- ρ = utilization (service time divided by inter arrival time)
- ε = tolerated relative error

Figure 1 is a semi-log plot of simulation time required for 95 percent confidence in a specified relative error as a function of utilization. It represents the exponential inter arrival and service time case ($C_a = C_s = 1$). Note that at high utilization and low relative errors extremely long simulation times are required. For example, to achieve 5 percent relative error in the mean on an 80 percent utilized queuing network requires one million service times.

In functional economic analysis we are interested in the relative future costs of alternative systems. There are uncertainties in process performance, resource requirements, cost estimates, investment required, workload, interest and inflation rates. There is also uncertainty in the future projection of these elements. Thus there is uncertainty in the discounted present value cost distribution for each alternative system. A plot of cumulative probability versus cost aids the decision process. The entire range in cost distribution is of interest. Figure 2 shows the expected number of simulation events required to obtain an event in the tail of the result distribution when using discrete event simulation. The equation plotted is:

$$E = 1 / P^C$$

Where:

- E = expected number of simulation events
- P = distribution tail probability
- C = uncertain model components

The lower the tail probability and the more components in the model, the greater the required number of simulation events to meet an accuracy criterion. For example, an average of one million simulation events are required in a six component model to simultaneously be in the 10 percent tail of all component distributions. To simultaneously be in the 1 percent tail requires an average of one trillion simulation events. In the limit it requires an infinite number of simulation events to capture the entire range of results. Thus discrete event simulation is not practical if one is interested in the entire result distribution in other than very small models with few components. If the minimum and maximum distribution values are not needed then discrete event simulation is practical. However, even in this case the model development, execution, and sensitivity analysis costs are higher.

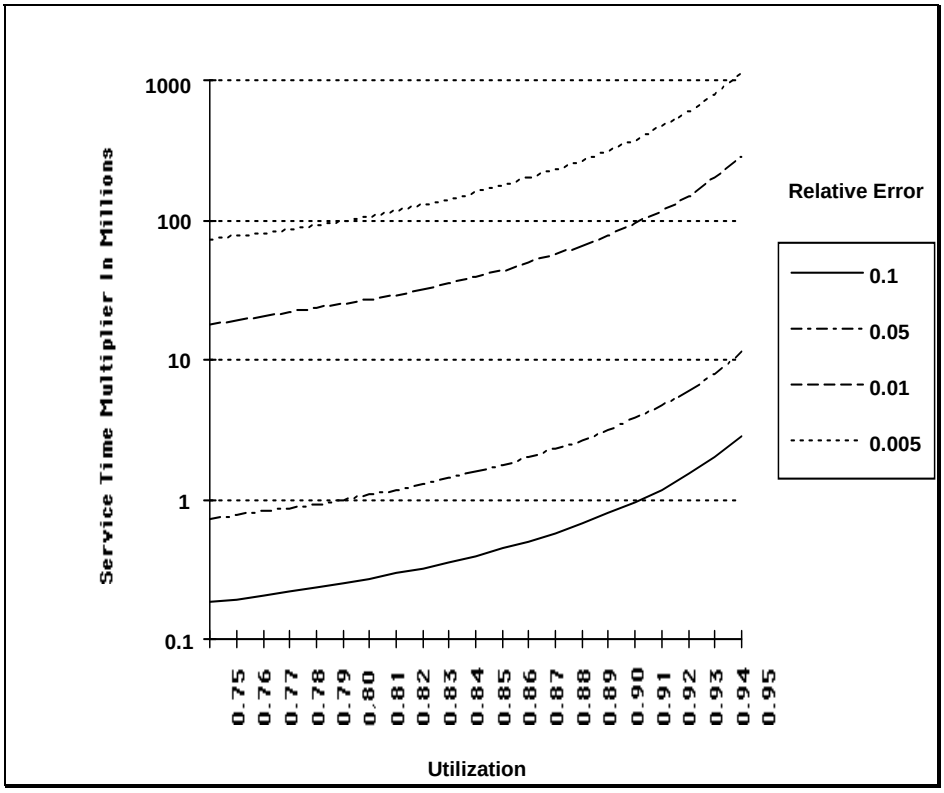

Figure 1. Simulation Time For Specified Relative Error

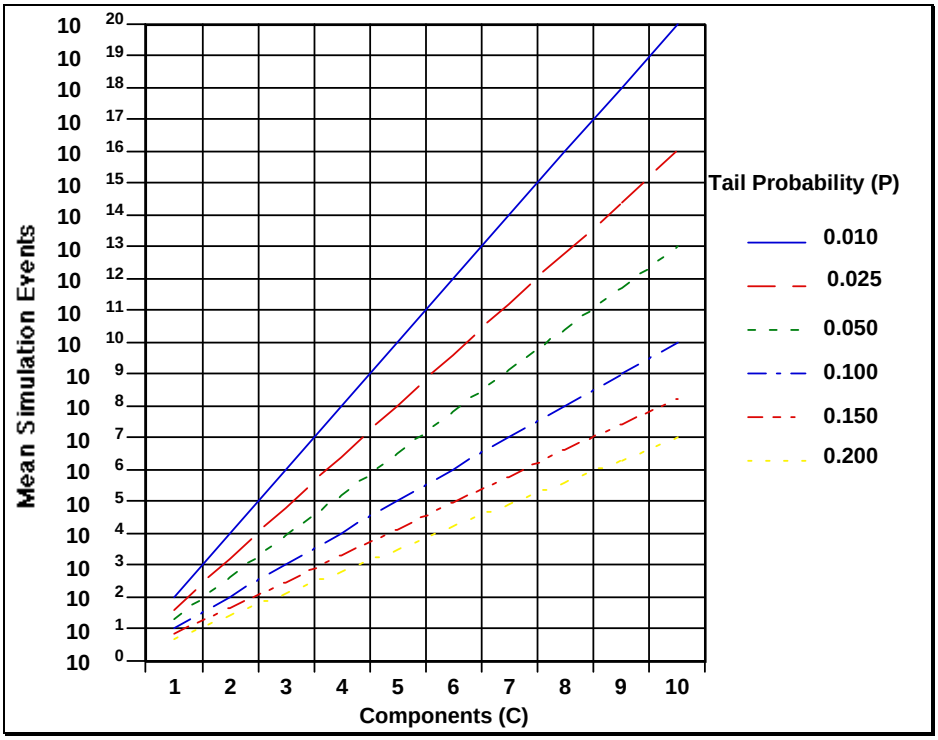

Figure 2. Simulation Events For A Specified Tail Probability Versus Model Components ($1/P^C$)

Analytic Queue Approximation

Figure 3 gives a summary of the analytic queuing algorithm for two tandem queues.

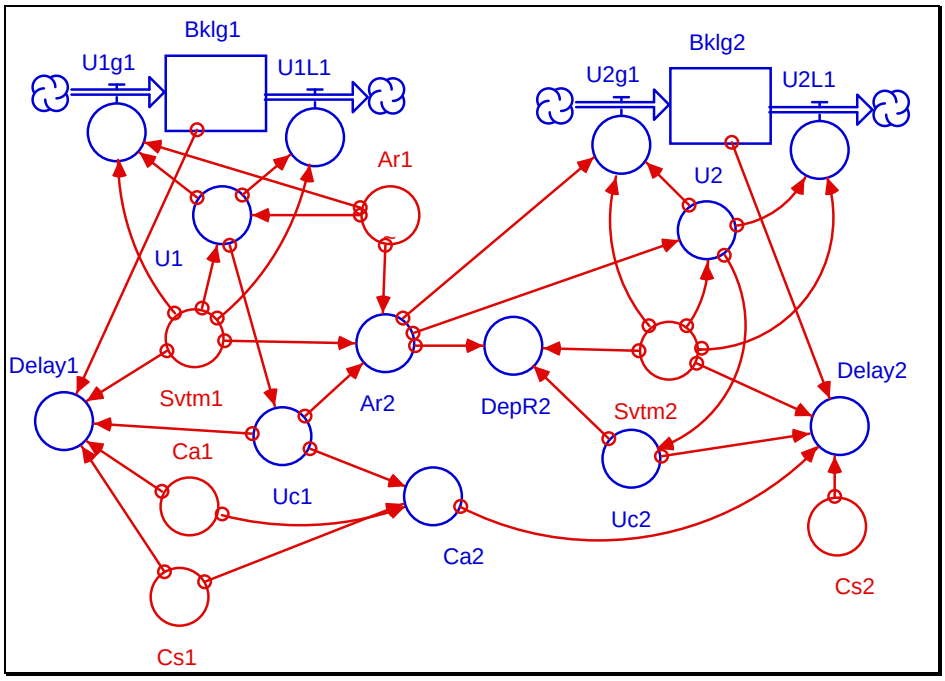

Figure 3. Tandem Queue Model

The mean steady state queue wait for a G/G/1 queue is given by:

$$\text{Wait} = \tau \rho(C_a^2+C_s^2) /(2(1-\rho))$$

Where:

τ = service time

C_a^2 = square coefficient of variation in inter arrival time

(variance in inter arrival time divided by the mean inter arrival time squared)

C_s^2 = square coefficient of variation in service times

ρ = utilization (service time divided by inter arrival time)

The relaxation time or time to reach steady state is two times the above wait divided by one minus the utilization. At high utilization the time to reach steady state is far longer than the time a dynamic workload remains close to a given value (the relaxation time equation has a $(1-\rho)^2$ term in denominator). In the limit it takes infinite time to reach steady state at unity utilization. At low utilization's (<.8) the steady state wait equation gives reasonable answers.

Between .8 and unity utilization we calculate a corrected utilization adding one half the utilization above .8 to .8. Above unity utilization, utilization is reset to .9. Thus the highest utilization used in the steady state delay equation is .9. More complex models have been developed accumulating time at different utilization's and correcting for the relaxation time, however the improvement in accuracy does not justify the added calculations.

When utilization exceeds unity, the excess over unity is added to work backlog. Conversely when utilization is less than unity, the work backlog is reduced. The mean node delay (Delay1,Delay2) is the queue wait, plus the service time, plus the service time times the number of jobs in the queue backlog. The coefficient of variation in service time (Cs = standard deviation in service time divided by mean service time) can calculated from service observations. The coefficient of variation in inter arrival times to the process can be calculated from the times between work requests. The coefficient of variation of inter arrival times to the second node is based on the utilization of the first node. At high utilization it is identical to the first node coefficient of variation of service times. At low first node utilization it is close to the first node inter arrival coefficient of variation (light traffic approximation principle for multi-class queuing networks with deterministic routing). In between we use a utilization weighted sum of the two coefficients of variation.

Priority Queuing Example

A three node two product queuing example was developed for both priority and common queue operation. S**4 supports arrays which considerably simplified implementation. One dimension was used for statistics (minimum, mean, maximum, standard deviation). The second dimension was used for product (products 1 & 2). Figure 4 shows the user interface cockpit for S**4. All parameters are given default values that can be reset at any time in the simulation. Hourly arrival rates are provided for each of twelve hours for product 1 and for product 2. Between hours the model employs linear interpolation. The twelve hour period is repeated in the simulation. In addition the coefficient of variation in inter arrival times and a load multiplier is specified for both products. The latter can be used to increase the workload over all hourly periods. An example of a time varying workload would be the phone loading during the workday.

In S**4 one can step any number of time steps, view available reports, graphs, and tables; change any of the displayed parameters; and then proceed with the simulation. In this simple example the minimum, maximum, and standard deviation in workload are calculated from multipliers of the time varying mean workload. The standard deviation is set to mean workload plus 7% minus mean workload. Node input data consists of mean service time and coefficient of variation in service time for each product.

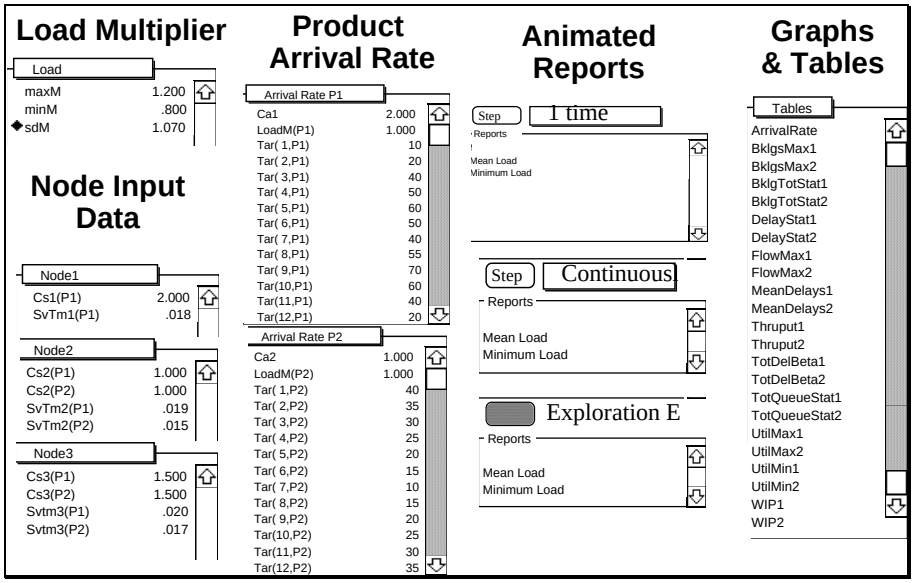

Figure 4. Priority Queuing Cockpit Interface

Figure 5 shows the animated report at minimum loading. As time proceeds each of the report numbers is updated. In priority queuing separate backlogs, delays, and utilization are kept for nodes serving both products. Product 1 is the sole user of node 1 but shares node 2 and 3 with product 2. Product 1 is given priority over product 2 which lowers its delay at the expense of greater product 2 delays. At time 20.25 hours product one workload requirements arrive at node one at a rate of 54 per hour. This utilizes .972 of node one capacity. However in the past workload at node one exceeded capacity and a backlog of .112 units developed. This backlog is being worked off with a departure rate of 55.56 units per hour. At node 2 this departing workload exceeds capacity (utilization=1.056) and will build the node two product 1 backlog in the next time step. The product 1 departure rate from node 2 is 48.19 units per hour that uses .964 of node three capacity. This leaves some capacity to work off the product 2 backlog. Final product 1 departure rate from the system is 44.58 units per hour. The squared coefficient of variation of node 1 product 1 inter arrival times was 2. The final node 3 product 1 inter departure squared coefficient of variation was 1.43.

Product 2 arrives at node 2 at a rate of 17 per hour. It forms a backlog there since all capacity is being used by product 1 which has priority. Thus there are no departures of product 2 from node 2 at his time. In past times, a product 2 backlog has developed at node 3. Since product 1 does not fully utilize node three this is worked off at a rate of 6.94 units per hour. In the upper-right are product summary statistics. Hour 20.25 total delay, backlog, queue, and work in process statistics are listed first. These are followed by total delay minimum, mean, maximum, and standard deviation statistics over the previous simulation time. Finally the fit beta distribution a and b parameters for the previous statistics are given. Note that product 2 delay statistics are only for that product that flows through the network. Most of the time simulation time priority product 1 takes all node capacity and there is no product 2 flow.

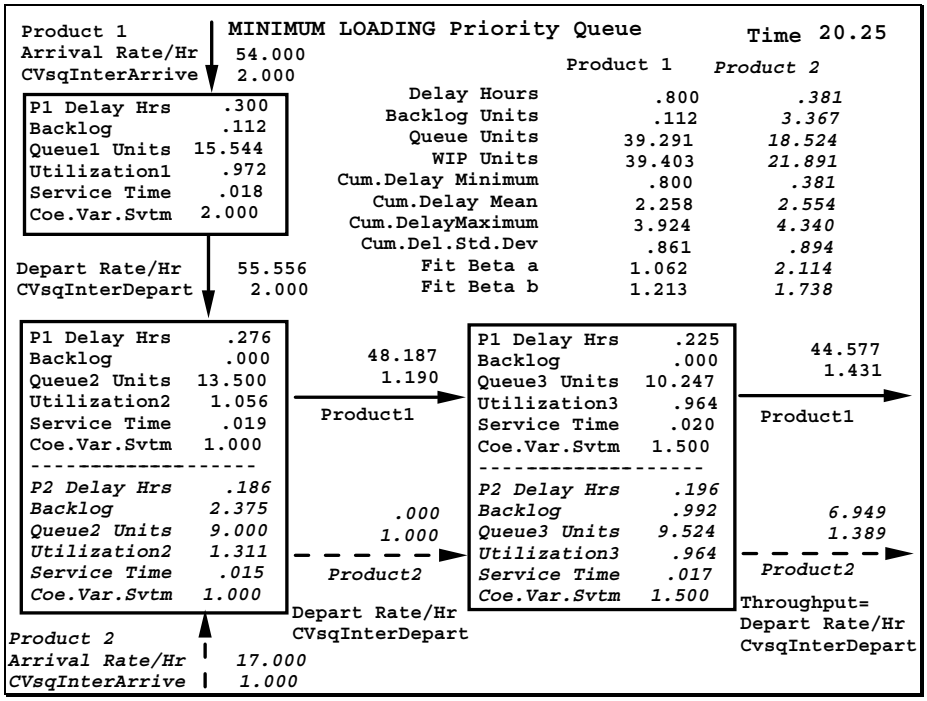

Figure 5. Two Product Three Node Priority Queuing Network

Figure 6 gives the product 1 instantaneous total delay statistics versus scenario time. The mean product 1 arrival rate is also shown. Note the repeating 12 hour workload pattern. At minimum workload product one backlogs are worked-off in the low loading times. However at the mean, mean plus standard deviation, and maximum workloads the backlogs are not worked-off, and over time will get very large. This is the time varying statistic data that is used to dynamically fit a beta distribution.

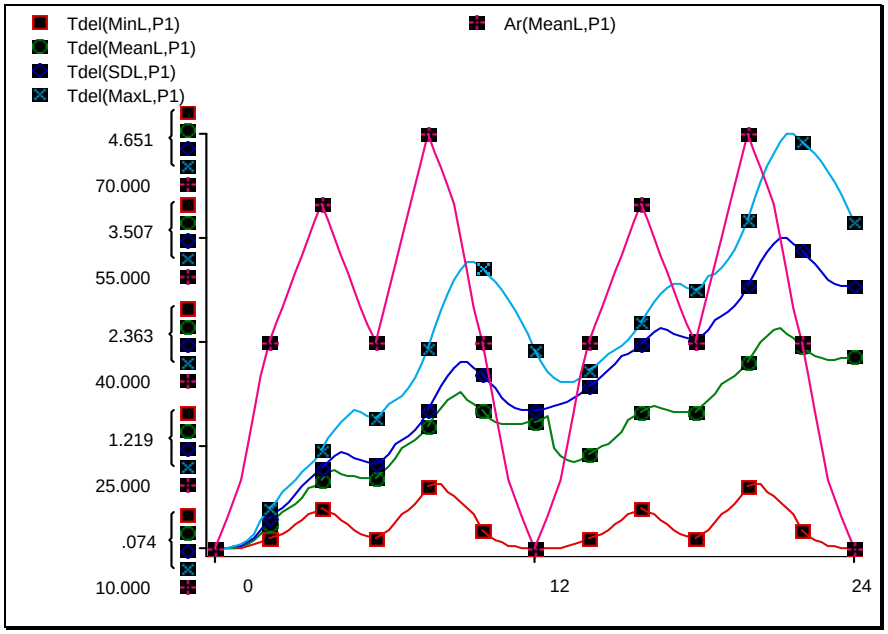

Figure 6. Priority Product 1 Total Delay Statistics

Common Queuing Example

Figure 7 shows the same situation without priority queuing. Product 1 delays are significantly increased from .8 hours to 2.47 hours. Product 1 throughput is reduced from 44.57 to 39.44 units per hour. Product 2 delay is also increased from .38 to 2.17 hours. In priority queuing there was little product 2 flow. With no priority capacity is shared and product 2 flow increases. At time 20.25 priority product 2 flow was 6.95 units per hour; without priority the flow is 12.41 units per hour.

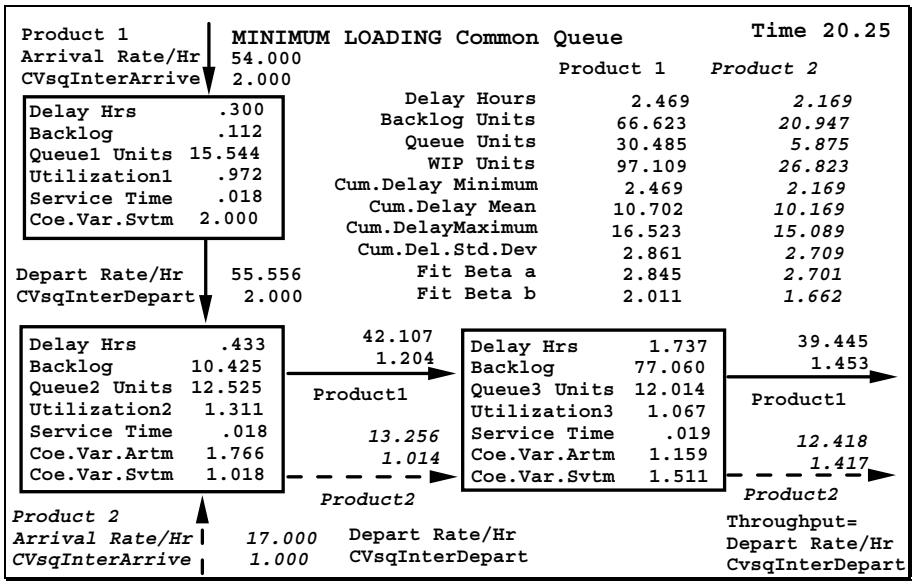

Figure 7. Common Queue Network

Figure 8 shows the increased product one delay (Tdel) statistics and the total product 1 and 2 arrival rate (Art). Note even at minimum workload delay is now increasing with time.

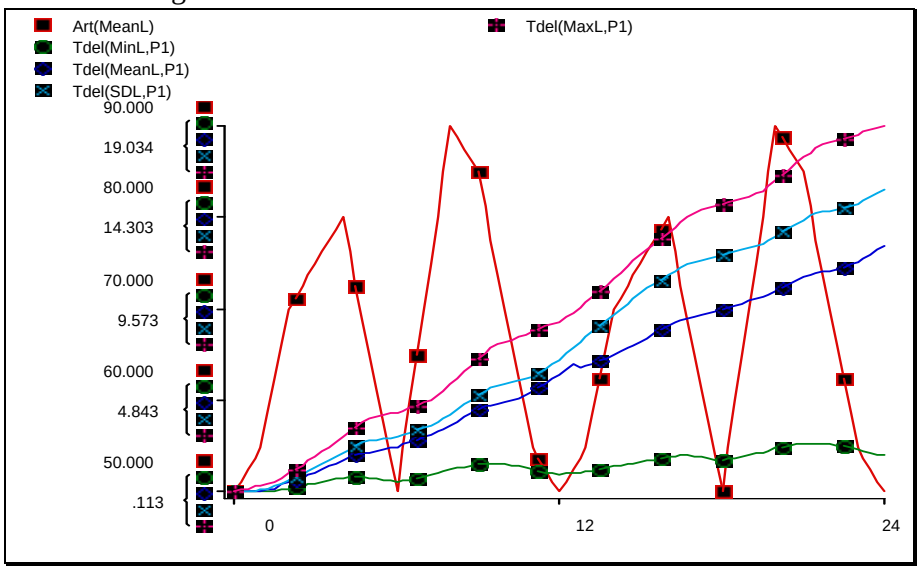

Figure 8. Common Queue Product 1 Total Delay Statistics

Figure 9 shows the common queue product 1 total delay fit beta distribution a and b parameters (ao(P1),bo(P1)). The a and b parameters change through simulation time giving an example of dynamic uncertainty analysis. Also shown is the mean product one arrival rate (Ar(MeanL,P1)) and the mode (mo(P1)) of the fit beta distribution. Note in this overloaded situation the mode of the fit beta distribution is increasing throughout the simulation. The minimum and maximum product 1 delays were shown on the previous figure 8.

In the first 6 simulation hours the fit beta a parameter is less than the b parameter, giving a positive skewed distribution. Later a exceeds b yielding a negative skewed distribution. At the low loading 12 hour point there is a large difference between a and b. Other program features provide probability distributions over any user selected time period.

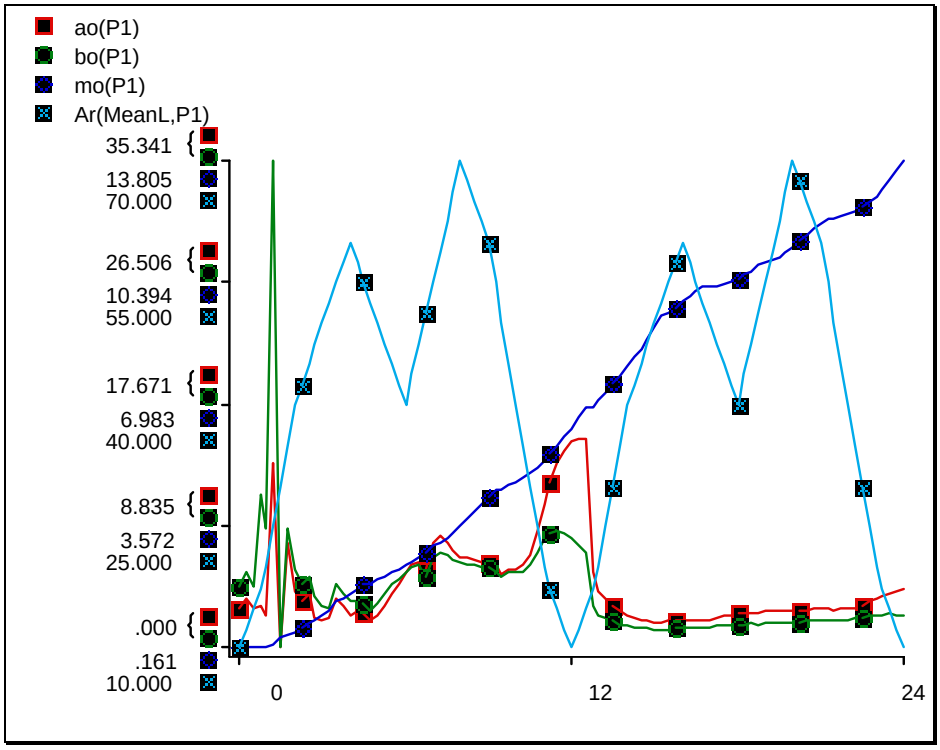

Figure 9. Common Queue Total Delay
Dynamic Uncertainty Analysis Fit Beta a and b Parameters

Figure 10 shows an example of a dynamic uncertainty distribution. The time varying beta distribution of product two delay is shown versus simulation time. Note how uncertainty increases with simulation time. The minmum, maximum, mean, and standard deviation statistics vary as the simulation proceeds. This results in time varying fit beta a and b parameters. The new technique provides an instantaneous probability distribution, which makes dynamic uncertainty plots possible. This new technique is especially helpful in forecasting uncertainty for non-linear feedback models.

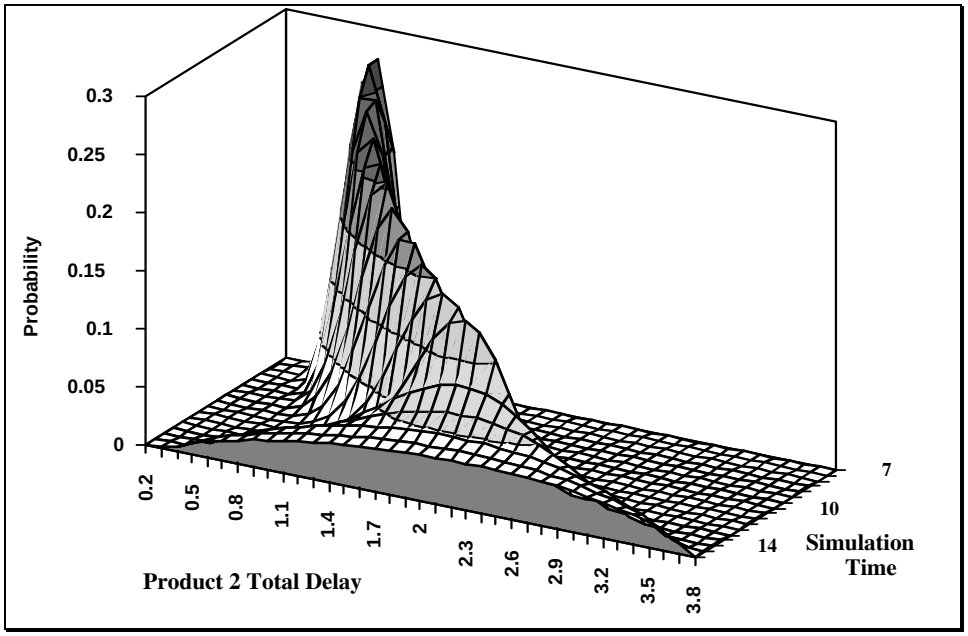

Figure 10. Dynamic Uncertainty Distribution

Conclusions

The technique demonstrated has wide applicability, especially in many factor sensitivity analysis situations. It overcomes the relaxation time problems of discrete event simulation. Thus it can be applied in time varying workload situations. It also provides a unique dynamic uncertainty analysis capability unavailable by other means.

References

Whitt,W.*Planning Queuing Simulations,* Management Science, Vol. 35, No., November 1989.

Whitt,W.*The Queuing Network Analyzer,* The Bell System Technical Journal, Vol62,No.9, November 1983.

Kleinrock,L.1976. *Queuing Systems Volume2: Computer Applications*. John Wiley & Sons, New York

Diehl,E.1992.*S**4 The Strategy Support Simulation System.* Microworlds Inc. Boston, Massachusetts