Visualizing Performance in the Frequency Plane

D. G. Keehn
Computer Engineering Department
San Jose State University
San Jose, California

Abstract

A method of showing the performance limiting effects of a product form queueing network as lines, planes, etc in a J dimensional space is given. The location of a certain critical point (Little's Law Point) in this space allows the asymptotic calculation of the normalizing constant G(**K**) of the network. This Little's Law point (LLP) is found by applying Little's Law to the augmented system generating function of the BCMP[1] network. The computational complexity of this algorithm is the Order(number of chains cubed * number of service centers in the system). Comparisons of numerical accuracy with other methods (Convolution, and another asymptotic method) are given.

Introduction

In the early stages of a project, either a new proposed system or an evolving system, a designer may have some idea on the resource usage (disks, processors, buses, and critical application code) that an application system requires. What is not always easy to understand is how the combined effect of these resource usage along with queueing effects of processors, disks etc. determine the overall system performance. Typically, one can build a detailed simulation model to estimate the relative importance of the critical parts of the design. As an alternative, an analytic model can be constructed in a shorter time, with fewer parameters. This allows the designer to widely vary these system parameters at little cost compared to simulation. However, not all effects can be accounted for by analytic models and a detailed simulation model or a approximate analytic model is usually needed as more detailed design parameters are understood and quantified.

This paper describes an analytic modeling technique which growths in accuracy and relative efficiency as the queueing network grows in size since it relies on a form of the central limit theorem of probability. If the reader is familiar with transfer functions from linear circuits, he/she will find the methods quite familiar as the product form queueing network is analyzed as a product of "transfer functions" to produce a system generating function in a complex frequency domain. To use the central limit theorem it is required that we find a special point in this frequency domain. This point is found by applying Little's Law { Average jobs in the system = arrival frequency * mean system time} to the overall system generating function. The major part of the algorithm is in finding this Little's Law point. We will refer to this algorithm as LLP. The location of the LLP within the feasible region of the frequency domain gives this approach a simple geometric interpretation which we will demonstrate shortly.

One can compare analytic modeling techniques by examining the computational complexity and performance measures that each method provides. The computational complexity of the LLP algorithm to compute the thruput for a chain K network, utilizations etc. is the $O(K^3 J)$. Here K is the number of chains and J the number of service centers in the system. The dominant term in the algorithm is the calculation of a loop which has $O(K^2 J)$ for each component of the K chains. This estimates, which has been experimentally verified, compares favorably with RECAL[4] and DAC[3]. The RECAL algorithm requires approximately J(J+K-1 choose J) operations to compute the normalization constant. The dominant term in the computational complexity of DAC is (3J+1)(J+K choose J). The storage requirement of LLP grows as the square of the number of chains. DAC and RECAL are theoretically exact which leaves open the question of the accurate LLP even though it takes fewer operations. The accuracy question is addressed in the *Numerical Experiments* section. An asymptotic expansion for LLP which quantifies the error in the asymptotic terms can be obtained. This expansion follows the ideas of the accuracy of the gaussian distribution in probability theory. Rather than show this expression, we have chosen to use numerical experiments to illustrate the accuracy of LLP.

LLP computes the normalization constant in a **non-recursive manner**. The

underflow/overflow problems which has presented difficulties for Convolution[7] and later RECAL[2],[4] has no impact on LLP. By contrast, LLP does not require a sequence of accurate calculations to obtain the next accurate value.

Turning to the question of which performance measures are available, we note that LLP computes the normalization constant using an asymptotic technique. Once the normalization constant is obtained most performance measures are at hand[7]. DAC[3][2] has an advantage in calculating the joint distribution of queue lengths at some or all service centers.

An alternative asymptotic method of was given by McKenna et al [5]. This method overcomes the numerical instability of Convolution and storage requirement limitations of MVA[5]. In addition, it has shown good accuracy even for small networks by using a few terms in its asymptotic expansion. Currently, this method applies to mixed networks with at least one IS center visited by each closed chain, single server fixed rate centers, and an assumption of "normal usage". This means a CPU utilization which is not too close to 100%. The requirement of at least one IS center in each chain is frequently met in practice. Different asymptotic expansions are required for "heavy usage" and are anticipated for general networks. LLP by contrast has the following characteristics.

a)IS centers (a set of users at terminals) need not exists in the network. In addition, no special expansion is needed for heavy usage of large networks. The same LLP approximation is used over the entire range of CPU utilizations. See[5] pp 346.

b)The accuracy of one term using the LLP compares favorably with the Convolution and Asymptotic methods even for moderate networks.

d)The storage requirement of LLP method does not depend on the number jobs/processes in each chain. The storage requirement does grow as the square of number of chains in the network.

e)Very heavy usage (>95%) is easily calculated by LLP.

f) LLP does allow calculation of *sensitivities* by applying the central limit technique to the partial derivative of the augmented system generating function. See Appendix B for more details. If new service centers are shown to yield to the"product form" the LLP method will continue to apply provided that the generating function for that service center can be summed in a simple form.

Restrictions of the LLP method include:

a) All product form networks with the exception of general queue length dependent (QLD) service rate centers can be calculated. LLP requires that one can sum the generating function for a single QLD service center.

b)If a single service center has an extremely large queue size as compared to other centers in the model, the central limit theorem breaks down. In this case however, all performance is determined by this single service center.

Using the Frequency Plane and the System Generating Function

To illustrate LLP consider a closed model of a signal processing system [9] shown in Figure 1. This model was derived from a data flow diagram of a radar signal processing system which removes clutter from radar signals. There are two chains (job types) in the system each with its own visit ratio and mean service times. A system designer often needs to understand the performance limits of a system. We illustrate this use of LLP for the system of Figure 1 as follows.

A.*Create the System Generating Function*

For each service center in the model construct the system generating function based on the type of service center, queueing discipline, and number of servers. These generating functions are selected from the basic results of Baskett, Chandy, Muntz and Palacios[1]. See Basic Result below. Form the product of these functions to get the overall system generating function (SGF). In addition, multiply the SGF by a factor which accounts for the population of chain 1 and chain 2 customers in the network. We will follow the notation described in E. de Souza de Silva and R.R. Muntz[2] using the following notation

J = number of service centers.
K = Total number of chains.
T_{jk} = mean service time of chain k customers at center j.
a_{jk} = relative utilization of chain k customers at center j. This term combines the visit ratio and the average service time.
n_{jk} = number of chain k customers at center j.
FCFS = First come first serve service center
SSFR = Single Server Fixed Rate per chain service center
IS = Infinite number of parallel Server
QLD = Queue Length Dependent service centers, the rate of service changes with the total number of jobs at the service center.

In the case at hand, with SSRF service centers we have :

$$W(z_1, z_2) = z_1^{-(N_1+1)} z_2^{-(N_2+1)} \prod_{j=1}^{6} \frac{1}{(1-a_{j,1}z_1-a_{j,2}z_2)} \ldots\ldots Eq\ 1$$

z_1 and z_2 are complex variables

the factors which account for the population are $z_1^{-(N_1+1)} z_2^{-(N_2+1)}$

B. *Determine the Feasibility Region in the Frequency Domain*
Find the enclosed region bounded by the $\{x_1, x_2\}$ axis and the line(s) nearest the origin as defined by the denominator of Eq.1. Figure 2 shows this region X for the model in Figure 1 and the specific values for the a_{ij} . Note in this case the system bottle neck is determined by the S_5 server. We see immediately a version of Amdahl's Law which bounds the thruput rate for each of the chains to lie inside the region X.

C. *Find the Little's Law Point*
For a specific load $\{N_1, N_2\}$, locate the Little's Law Point in the real plane $\{x_1, x_2\}$ at which the augmented SGF (Eq 1) is a minimum. To see that the minimum coincides with Little's Law Point we take the natural log of Eq 1 and set both partial derivatives to zero.

$$\ln W(z_1, z_2) = -(N_1+1)\ \ln(z_1) - (N_2+1)\ \ln(z_2)$$
$$-\sum_{1}^{6} \ln(1-a_{j1}z_1-a_{j2}z_2) \ldots\ldots Eq\ 2$$

Taking partial derivatives and rearranging we have Eq 3.

$$N_1+1 = \sum_{1}^{6} \frac{a_{j1}z_1}{(1-a_{j1}z_1-a_{j2}z_2)} \ldots\ldots\ldots\ldots Eq\ 3$$
$$N_2+1 = \sum_{1}^{6} \frac{a_{j2}z_2}{(1-a_{j1}z_1-a_{j2}z_2)}$$

To notice Little's Law we see that each term on the right side of Eq. 3 has the form of arrival rate z_k for a given chain, times the average response time with Poisson arrivals for the queue type. The left shows the total number of customers in a chain (plus 1) which are distributed throughout the network in an

average sense as shown by the right hand side.

Hence the point of zero derivatives is determined by the constraint of distributing the N_k customers of each chain over the 6 service centers in the network of Figure 1. The large fraction of the number in a chain piling up where the Poisson wait times are large. Notice the interference between chains is expressed in the denominator of each of the above terms.

Good numerical techniques for finding the minimum of $W(z_1,z_2)$ in Eq 1 are available[8]. An algorithm for locating the LLP is given in [11] where the Davidon-Fletcher-Powell algorithm is used.

D. *Obtain Numerical Results*

Numerical values are obtained by plugging the LLP into Eq.8 in Section *Calculating Performance Using the Central Limit Theorem* . In the more lengthy report [11] it is show how one can calculate the normalizing constant G(**K**). It turns out that the LLP will provide a vector (of dimension the number of chains) at which we can simply evaluate a function (Eq 8) to get G(**K**). Once G(**K**) is available many performance measures are readily calculated using this normalization constant. These include queue length distributions for each chain, utilizations, throughput, and response times.

E. *Visualize the Effect of System Changes*

Changes in the system performance can be estimated by viewing the motion of the Little's Law Point (LLP) and the boundary of the feasible region as shown in Figure 2a.

Changes in the $\{a_{ij}\}$ show up geometrically as shifts in the lines which bound the region X shown in Figure 2a,b. If the line determined by S_5 shifts to the right, the LLP reacts in the same direction. If this line shifts beyond an adjacent line then that line will provide a new boundary of the possible thruput values.

For instance, if we change the service rate of the server S5 from 0.5 to 1.0 then the new feasible will increase as shown in Figure 2b. Note that the line S5 has shifted to the right making the servers S2 and S7 the limiting system factors. In this case increasing the system load (N_1,N_2) could achieve higher throughput.

Changes in loads $\{N_1,N_2\}$ cause the LLP to shift along the axis corresponding to that chain. Figure 2b illustrates the effect of increasing N_1 while holding N_2 constant.

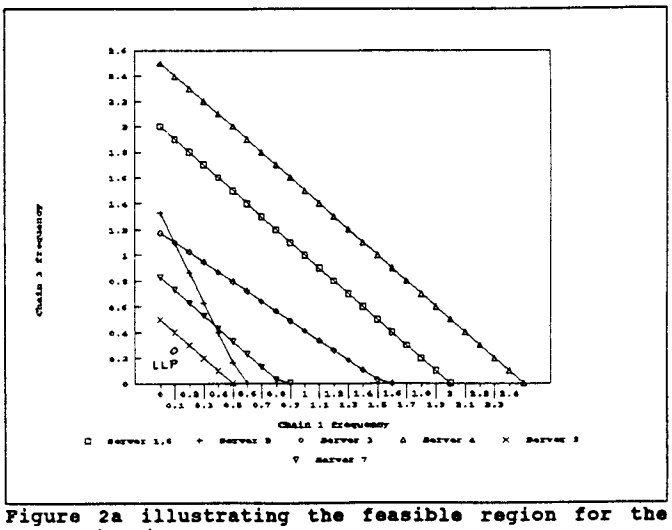


Figure 2a illustrating the feasible region for the system in Figure 1.

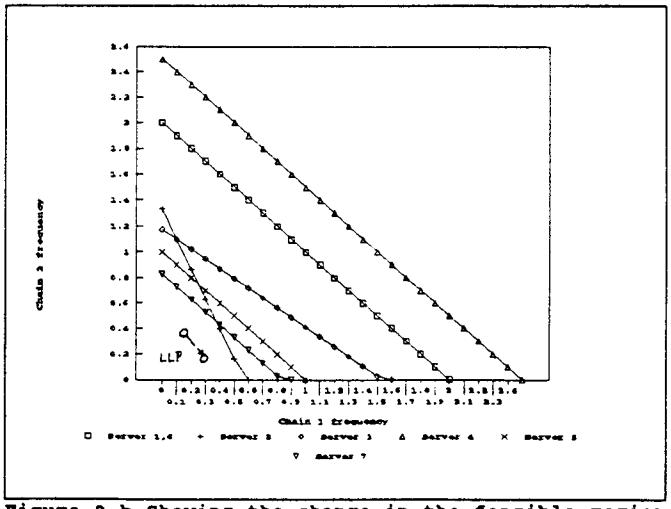


Figure 2 b Showing the change in the feasible region as $\{a_{ij}\}$ changes.

Calculating Performance Using the Central Limit Theorem

Little's Law Point Gives an Asymptotic Expansion

The basic result of the LLP method is stated here. A detailed description of this result is available in report [11].

Basic Result Define J functions $h_j(z_1,z_2,\ldots,z_K)$ of R complex variables, one variable for each chain. Here j ranges over all service centers {j= 1, ...,J}. For each service center of a given type select $h_j()$ from one of four distinct types as shown below.
Form the product $W(\mathbf{z}) = h_1(\mathbf{z})\ h_2(\mathbf{z})\ \ldots\ h_J(\mathbf{z})$,

$$\textit{for type 1 serive centers } h_j(\boldsymbol{z}) = \frac{1}{(1-\sum_{k\epsilon K} e_{jk}T_j\ z_k)}$$

$$\textit{for type 2 or 4 } h_j(\boldsymbol{z}) = \frac{1}{(1-\sum_{k\epsilon K} a_{jk}\ z_k)} \ldots\ldots\ldots$$

$$\textit{for type 3 } h_j(\boldsymbol{z}) = \exp\left(\sum_{k\epsilon K} a_{jk} z_k\right) \ldots\ldots\ldots$$

then $G(K_1, K_2, \ldots, K_K)$ is the coefficient of the term with powers $\{K_1,K_2,\ldots K_K\}$ in the power series expansion of the analytic function $W(\mathbf{z})$. For a large number of network serves ln $W(\mathbf{z})$ is basically a quadratic function of $\{z_j\}$ in the neighborhood of the LLP. Consequently, the technique relies on an integral which is the K dimensional version of the gaussian integral arising in the central limit theorem.

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{C}\boldsymbol{y}\right) dy_1 \ldots dy_K = \frac{(2\pi)^{\frac{K}{2}}}{(\det \boldsymbol{C})^{\frac{1}{2}}} \ldots\ldots Eq\ 4$$

In our context the matrix **C** is a positive definite matrix of partial derivative evaluated at the LLP. We will use the results of Eq 4 after applying Cauchy's Integral Theorem to Eq 5 below.

The flexibility afforded by Cauchy's Integral theorem in K-complex variables allows us to chooses a path through the K dimensional space which passes through the critical point at which all first order partial derivatives of the integrand vanish. This point coincides with the Little's Law point. The question of uniqueness of the minimum for $G(\mathbf{K})$ can be addressed by the use of convexity arguments.

We can form an expression for the normalizing constant $G(\mathbf{K})$ in terms of a transform integral of the function $W(\mathbf{z})$ as follows (See [11] for detailed steps).

$$G(\boldsymbol{K}) = \oint\oint \ldots \oint_{\Gamma} \exp\left\{-\sum_{k=1}^{K} (K_k+1)\ln z_k + \sum_{j=1}^{J} \ln\ h_j(z_1,\ldots,z_K)\right\} \prod_{k=1}^{K} \frac{dz_k}{2\pi i} \ldots Eq\ 5$$

Here the $h_j()$ are selected from those functions detailed in Basic Result. This

choice depends on how the designer models the system at hand.
We now expand the analytic function w(**z**) in a power series about a point $\mathbf{z_0}$. w(**z**) is the argument inside the exp{..} term of Eq 5.

$$w(\mathbf{z}) = w(\mathbf{z_0}) + \sum_{i=1}^{K} D_i w(\mathbf{z_0})\,(\mathbf{z}-\mathbf{z_0}) + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K} D_{ij} w(\mathbf{z_0})\,(z_i - z_{0i})\,(z_j - z_{0j})$$
$$+ \frac{1}{3!} d^3 w(\xi; \mathbf{z}-\mathbf{z_0})$$
here ξ *lies on the line ending in* $(\mathbf{z}, \mathbf{z_0})$ *Eq* 6

and choose a path which passes through the point $\mathbf{z_0} = \mathbf{x_0} + i\,\mathbf{y_0}$ with $y_0=0$ so that the first partial derivatives in the x_{0j} vanish. Note the value of this partial derivative is independent of the path through x_0. Since w(**z**) is an analytic function, this point is a minimum of w(**z**) in the **x** direction but a maximum in the **y** direction. We neglect the third and higher order terms in Eq.6 . This leaves the term $w(\mathbf{z_0})$ and the second order term involving the second partial derivatives evaluated at the LLP.
If the number of servers is moderate to large there are a large number of positive second partial terms contributed for each service center. Consequently, the magnitude of the second partial term will be large, resulting in sharp behavior of the integrand in the vicinity of the LLP. Note this argument is the same one used in the central limit theorem[12], in which the log of the characteristic function is shown to essentially a quadratic function for a large number of random variables. In Eq. 7, the role of the number of random variables is played by the number of service centers in the network. The behavior of the integrand in the **y** dimension is a maximum since the second order term is quadratic in the variables iy_k . Invoking Eqs 5,6 we have the final equation which can be used to compute G(**K**) .

$$G(\mathbf{K}) \sim \exp\{w(\mathbf{z_0})\} \oint .. \oint_{x_0+iy_r} \exp\{-\frac{1}{2}\sum_{r=1}^{r=R}\sum_{s=1}^{s=R} D_{ij} w(\mathbf{z_0})\, y_r y_s\}\; d\mathbf{y} \qquad Eq\ 7$$

Applying Eq 4 to integrate of this equation we have a result which allows us to calculate G(**K**) in terms of the load for each chain, the service center transforms and the LLP. Using the log from we have:

$$\ln G(\mathbf{K}) = -\sum_{k=1}^{K} (K_k+1)\ln(x_{ok}) + \sum_{j=1}^{J} \ln\, h_j(\mathbf{x_0})$$
$$-(K/2)\ln(2\pi) - \frac{1}{2}\ln\{\ \det\ [D_{ij} w(\mathbf{x_0})] \quad ...Eq\ 8$$

Numerical Experiments

We first consider experiments which directly compare the Asymptotic method of McKenna et al with the LLP method. Note that choosing this network is a test of the asymptotic method; a larger number of servers would increase the accuracy. The same table numbers are used as in [5] to make references easier. Reference [11] contains more extensive experiments for high CPU utilization. These results are shown to demonstrate the stability of the LLP method for high CPU usage and a large number of chains. In the last part, limitations of the LLP Method are discussed.

Comparing to Prior Results

The first experiment in Figure 3 compares Table II in [5] with the results of LLP. The parameters of this experiment are noted in Figure 3. The network model consist of two chains, a set of terminals with various average think times accessing a single CPU with the specified processing times. In examining Figure 3 we note a general agreement in the CPU utilization, deviations from are usually no more than 5 %. For the case of 90/60 jobs/processes there is a substantial disagreement which the author believes is a typing error. The last five rows show the high CPU utilization extension of this experiment. LLP easily converged to a relative change of one part in a million.
NOTE: Only CPU utilizations are shown, as the response times in [5] can be directly calculated from the utilizations for these networks.

The next set of experiments is defined in Figure 4A. The network configuration is the same as Table IV in [5]. There are 17 chains with a variety of CPU service rates as well as IS service rates. Note chains 4 and 5 should have the same utilization as they are symmetric in IS service rates as well as CPU service rates.

a) The left two columns Figure 4B shows a comparison with [5]. Once more the results are generally within 5% as measured by CPU usage. The total CPU utilization compares 44% in [5] and 43% for LLP . Chains 4 and 5 show identical CPU% for LLP, where as [5] shows 0.046 for chain 4 versus 0.040 for chain 5.

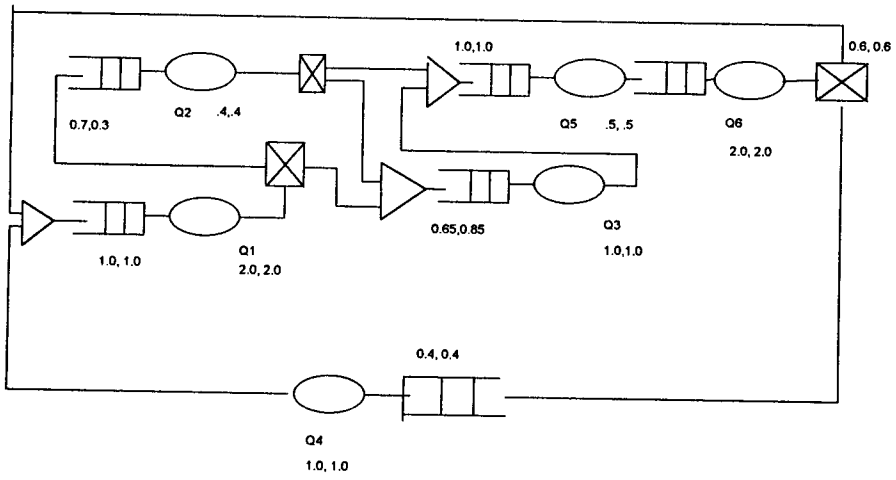
b)The right most two columns in Figure 4B shows the effect of increasing the CPU usage by each chain by a factor of 2.5 and 5.0 respectively. Alternately, we could say that the network's CPU has been replaced by slower models in the same ratio. These results show good convergence of the LLP and maintenance of good symmetry between chains 4 and 5. These results show the added range which LLP can provide. More extensive experiments which extend the number of chains and cpu utilization are given in [11].

Conclusions

The LLP Method for large product form networks proves effective with computational complexity of $O(K^3J)$ for a large number of chains. Service center utilizations can be near 100%, and underflow/overflow problems are not encountered. This method complements Convolution, RECAl, DAC methods by solving networks. For small to moderate size networks Convolution and MVA are still very effective.

Figure One - Modified network taken from "Queueing Network Model for radar Signal Processing System " of McCabe & Associates . A closed loop has been added and two chains are introduced. The numbers nearest the queues indicate the normalized visit ratios for each chain. The numbers nearest the Q's are service rates for each service center.
The system generating function for the network is shown here :

$$\frac{1}{(1-0.5z_1-0.5z_2)^2(1-1.75z_1-.75z_2)(1-0.65z_1-0.85z_2)(1-0.4z_1-0.4z_2)(1-2.0z_1-2.0z_2)}$$



Degree of Multiprogramming	Total %CPU CADS	Total % CPU LLP-Method	Total % CPU Euler(McKenna et al)
10/10	0.118	0.121	0.119
20/20	0.239	0.214	0.23
30/30	0.358	0.362	0.35
40/40	0.476	0.481	0.48
50/50	0.593	0.599	0.60
60/60	BD	0.71	0.70
80/60	BD	0.75	0.72
90/60	BD	0.776	0.97*
100/50	BD	0.705	0.69
110/50	BD	0.726	0.71
140/50	BD	0.788	0.79
200/10	BD	0.54	0.54
170/40	BD	0.76	0.75
200/25	BD	0.683	NA
200/50	BD	0.894	NA
200/100	BD	0.982	NA
200/200	BD	0.994	NA
400/400	BD	0.997	NA

NA-not available BD breakdown of method
Input for this experiment:
No. of classes = 2, Think time class 1 = 450 sec.,
Think time class 2 = 150 seconds,
Processing Time are 1.0 seconds and 1.5 seconds respectively.

Figure 3.Comparison with the Table II (Euler Approximation) from [5]

Class of Customer	Service Rate of Infinite Server	Service rate for CPU
1	0.0033	20
2	0.033	2
3	0.0033	4
4	0.033	4
5	0.033	4
6	0.033	6
7	0.033	20
8	0.00033	0.6
9	0.00055	0.6
10	0.0033	0.6
11	0.00033	0.2
12	0.00055	0.2
13	0.0003	0.2
14	0.033	1
15	0.00033	1
16	0.00055	1
17	0.0003	1

gure 4A. Problem Specification - from Table IV in [5]
e degree of multiprogramming for each class is 5.

Class of Customer	CPU% x1 McKenna et al	CPU % x1 LLP Point	CPU % x2.5 LLP Point	CPU % x 5.0 LLP Point
1	0.008	0.0008*	0.002	0.004
2	0.080	0.081	0.164	0.148
3	0.004	0.0041	0.010	0.019
4	0.046	0.041	0.09157	0.109296
5	0.040	0.041	0.09157	0.109296
6	0.027	0.027	0.063	0.097
7	0.008	0.008	0.020	0.035
8	0.003	0.0027	0.007	0.013
9	0.005	0.0046	0.011	0.021
10	0.027	0.0273	0.063	0.097
11	0.008	0.0083	0.020	0.035
12	0.013	0.0137	0.033	0.053
13	0.007	0.0075	0.018	0.032
14	0.156	0.157	0.27	0.181
15	0.002	0.0017	0.004	0.008
16	0.003	0.0027	0.007	0.013
17	0.001	0.0015	0.004	0.007
Total CPU	0.44	0.43	0.882	0.967

Figure 4B.Results of a Comparison of Euler and LLP aproximations. Compare to Table IV(b) in [5].

References

[1] F. Baskett et al, "Open Closed, and Mixed Networks of Queues with Different Classes of Customers", Journal of the Association for Computing Machinery, Vol. 22, No. 2 , April 1975 , pp 248-260.

[2] Edmundo de Souza e Silva and Richard R Muntz, "Queueing Networks:Solutions and Applications", Stochastic Analysis of Computer and Communications Systems, Hideaki Takagi (Editor) Elsevier Science Publishers, 1990

[3] Edmundo de Souza e Silva and S. S. Lavenberg, "Calculating the Joint Queue-Length Distribution in Product-Form Networks, Journal ACM, vol 36, No 1 january 1989, pp 194-207.

[4] Conway, A. E. and Georganas, N.D. " A new Efficient Algorithm for the exact analysis of multiple chain closed queueing networks" J. ACM 33, 4 (Oct 1986) 768-791.

[5] J McKenna, et al , "A Class of Closed Markovian Queuing Networks:Integral Representations, Asymptotic Expansions, Generalizations", The Bell System Technical Journal, Vol 60,
No. 5, May-June 1981.

[6] K. G. Ramakrishnan and D. Mitra, "An overview of PANACEA, a Software Package for Analyzing Markovian Queueing Networks", Bell Systems Technical Journal 61, pp 2849-2877.

[7] S.S. Lavenberg, "Computer Performance Modeling Handbook", Academic Press 1983.

[8] W. H. Press et al, "Numerical Recipes in C", (the art of scientific computing) Cambridge University Press, 1988,pp 324-328.

[9] T.J. McCabe et al "Structured Real Time Analysis and Design" COMPSAC-85 IEEE, OCt 1985, pp 40-51.

[10] A. C. Williams and R. A. Bhandiwad "A generating function approach to queueing network analysis of multiprogrmamed computers." Networks 6:1, 22, 1976.

[11] D.G. Keehn "Visualizing Performance in the Complex Plane" SJSU Technical Report.

[12] Harald Cramer "Mathematical Methods of Statistics", Princeton University Press, 1961, pp 217.