DTIC FILE COPY

(2)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY
N/A

EDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release: distributi
unlimited

MBER(S)

AD-A206 840

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 89-0424

Georgia Inst. of Tech.

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
AFOSR/NM

6c. ADDRESS (City, State, and ZIP Code)
School of Information and Computer Sci.
Georgia Institute of Technology
Atlanta, GA 30332

7b. ADDRESS (City, State, and ZIP Code)
AFOSR/NM
Bolling AFB DC 20332-6448

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
AFOSR

8b. OFFICE SYMBOL (If applicable)
NM

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR-88-0028

8c. ADDRESS (City, State, and ZIP Code)
Building 410, Bolling AFB,DC 20332

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61102F	2304	A2	

11. TITLE (Include Security Classification)
PARAMETRIC ANALYSIS OF QUEUEING NETWORKS WITH BLOCKING

12. PERSONAL AUTHOR(S)
I. F. Akyildiz

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM 1 Nov 87 TO 31Oct8

14. DATE OF REPORT (Year, Month, Day)

15. PAGE COUNT

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

As we already observed in the investigation of queueing networks with blocking that the throughput is a non-decreasing function of the number of jobs [3], i.e., the blocking events have the effect of violating the throughput results, which were mentioned on pages 7, 8, and 9 of the proposal [1].
Two questions arose from this observation:

1) How to distribute the total buffer capacity to the stations such that no deadlock will occur and a maximum (optimal) throughput will be achieved?

2) Given the buffer capacity of each station in the network. How to select the total number of jobs in the network such that the throughput will be maximum (optimum)?

To answer these questions first we assumed that all stations have infinite capacity and derived new formulas for optimal throughput and response times based on the well-known mean value analysis approach [4]. Then in [5[we found necessary and sufficient conditions for buffer allocation in cyclic networks with blocking such that an optimal throughput will be achieved.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED □ SAME AS RPT. □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
Dr Abraham Waksman

22b. TELEPHONE (Include Area Code)
202-767-5027

22c. OFFICE SYMBOL
NM

Previous editions are obsolete.

89 4 [illegible] 087

Final Report on the Project

"PARAMETRIC ANALYSIS OF QUEUEING NETWORKS WITH BLOCKING"

AFOSR-88-0028

November 1, 1987 - November 1, 1988

I. F. Akyildiz
School of Information and Computer Science
Georgia Institute of Technology
Atlanta, GA 30332

Tel.: 404-894-5141

Our investigation within this project was focused on two major areas: *transfer blocking and rejection blocking* protocols in queueing networks with finite capacities.

1. Networks with Transfer Blocking

In this type of networks blocking occurs when a job after completing its service at a station cannot proceed to its destination which is full. The job resides in the server of the source station and waits there until a place becomes available in the destination station.

1.1. Deadlocks

On page 5, equation (2), of the proposal [1] we stated that deadlocks may occur in this type of networks. We proved this statement formally in [2]. In [2] we also gave an algorithm for the distribution of total buffer capacities to stations such that no deadlock will occur. Additionally, we introduce an algorithm which automatically finds cycles in so-called cacti networks. *This paper was submitted to "QUESTA: Queueing Systems: Theory and Applications" in October 1987 for publication. The paper is accepted in June 1988 and will be published soon.*

1.2. Throughput Optimization

As we already observed in the investigation of queueing networks with blocking that the throughput is a non-decreasing function of the number of jobs [3], i.e., the blocking events have the effect of violating the throughput results which were mentioned on pages 7,8 and 9 of the proposal [1]. Two questions arose from this observation:

i) How to distribute the total buffer capacity to the stations such that no deadlock will occur and a maximum (optimal) throughput will be achieved?

ii) Given the buffer capacity of each station in the network. How to select the total number of jobs in the network such that the throughput will be maximum (optimal)?

To answer these questions first we assumed that all stations have infinite capacity and derived new formulas for optimal throughput and response times based on the well-known mean value analysis approach [4]. Then in [5] we found necessary and sufficient conditions for buffer allocation in cyclic networks with blocking such that an optimal throughput will be achieved.

1.3. Norton's Theorem Application

Our major breakthrough for this proposal [1] was that we found a very efficient solution for the proposed problems stated on pages 5 through 11. The entire solution is described in detail in [6]. The execution of the application of Norton's theorem on blocking queueing networks as stated on pages 6 and 7 of [1], needs 4 steps for the analysis, (see page 7 of [6]):

i) Construction of the subnetwork Γ
ii) Throughput Analysis of the Subnetwork
iii) Construction of the Phases for the Selected Station
iv) Analysis of the Two-Station Network

As mentioned in the previous report for (*AFOSR*−87−0160), the throughput algorithm suggested in the proposal [1] (on pages 7,8,9) and used in step ii) above, has extensively been studied for its accuracy. The algorithm is described in detail in [3].

Instead of determing the capacity of the composite station as proposed on page 9 in [1], we solved that problem in a different way. We replace the subnetwork by a composite station with infinite capacity. However, the blocking events at selected station are neglected due to some full stations in the subnetwork. Therefore, we modify the service mechanism of the selected station such that all delays a job might undergo due to blocking events could be represented. By this modification we obtain a multiphase server representing all blocking delays in the selected station. We intorduce an iterative technique to determine the parameters such as the branching probabilites and delay service times for the multiphase server of the selected station. Finally, the entire network was reduced to two-station network as it was planned in the proposal [1]. The reduced two-station network containing the selected finite capacity station with multiphase-server representing the blocking delays and the composite (flow-equivalent) infinite capacity station representing all other stations in the originally given network is analyzed by a numerical technique as discussed in [6].

The results of our technique have been validated by executing several examples and comparing them with simulation counterparts which are obtained by IBM-RESQ package. The validation study shows very good accuracy of our results.

2. Extended Parametric Analysis of Queueing Networks with Blocking (PART II)

In the first step we, in collaboration with colleagues from the University of Erlangen-Nuernberg, applied the extended parametric analysis concept on queueing networks with infinite capacities and implemented a parallel processing on a multiprocessor system with pyramid type architecture [7]. The experiments showed that parallel cumputing provides some advantages regarding storage space. The synchronization between processors take some time which reflects in run time of the programs. An open research question is how to optimize the run time as well as how to optimally decompose the network for parallel computation.

As proposed in section 4 in [1], we also applied the extended parametric analysis concept on queueing networks with blocking where the blocking network can arbitrarily be partitioned into disjoint subnetworks. A modified and extended version of the algorithm presented in [6] can be applied solving the problem of parallel analysis of the subnetworks. We briefly outline the major steps of the method using the example from section 4 in [1]. The network to be analyzed is given in Figure 1.

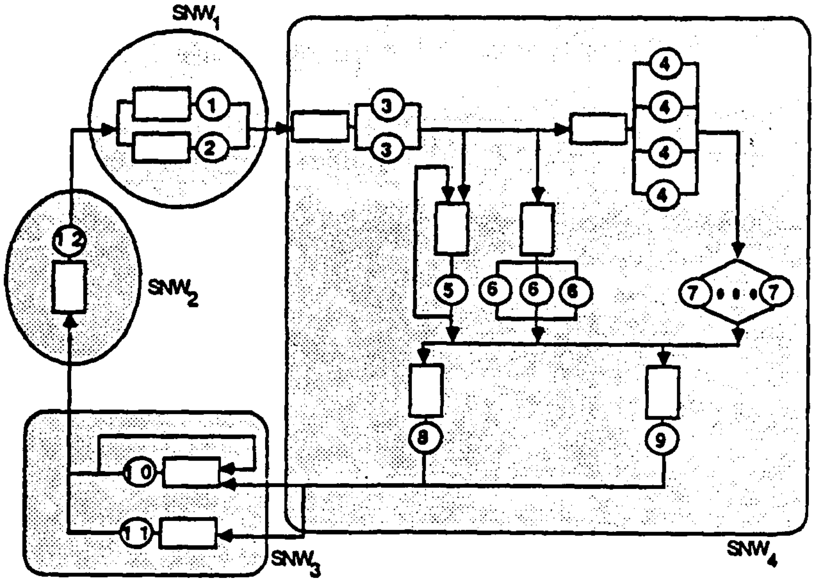

Figure 1.

We decompose the network into four subnetworks SNW_j:

SNW_1 = {1 , 2} ; SNW_2 = {12} ; SNW_3 = {3 , 4 , 5 , 6 , 7 , 8 , 9 } ; SNW_1 = {10 , 11 }

Each subnetwork is analyzed by shortening all stations in other subnetworks, i.e., the service times of the stations not belonging to a subnetwork are set equal to zero. The analysis of the subnetworks is done as described in [6] and load-dependent throughput values $\lambda_j(k)$ are obtained for each subnetwork SNW_j (for j = 1,2,3,4). As pointed out in the proposal the analysis of each subnetwork is independent from other subnetworks. Therefore, this part of the algorithm can be carried out in parallel and speeding up the total computation.

In the next step we construct a composite station for each subnetwork SNW_j with load-dependent service rates $\mu_{cj}(k)$ which are obtained by setting $\mu_{cj}(k)$ equal to the throughput values $\lambda_j(k)$. The composite stations are composed to a closed queueing network according to the configuration of the subnetworks in the originally given network. For the example given in Figure 1 we obtain a serially switched network as given in Figure 2.

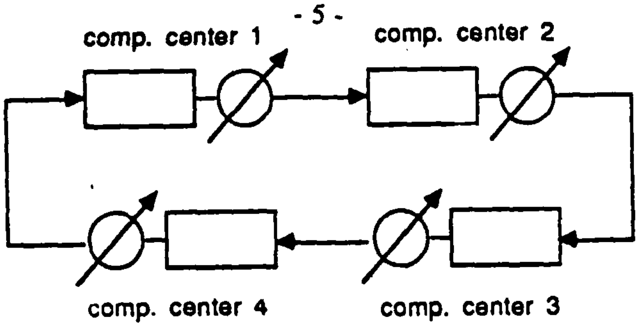

Figure 2.

The blocking events which occur during the transition of jobs traversing from one subnetwork to another are considered by adding delay phases to the load-dependent service stations in Figure 2. The construction of the delay phases is done as described in [6]. The application of our method to this example produces a network given in Figure 3. Note that the phase construction in Figure 3 is simplified in order to make a graphical representation possible.

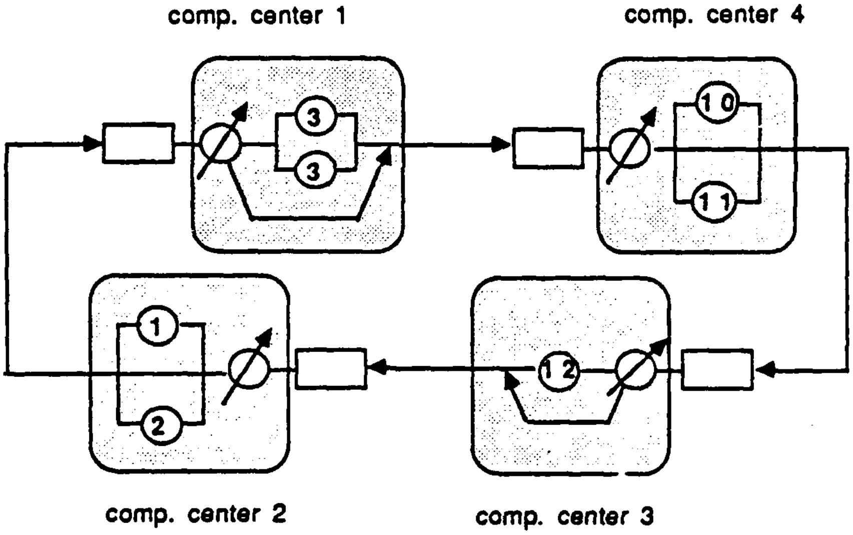

Figure 3.

The final step of the algorithm consists of an iteration between the network of Figure 3 and the analysis of each station of the originally given network in isolation (as a simple $M/M/1$ station with finite capacity). Trying to apply a similar method as described in [6] to this type of network we had to solve the following problems:

The multi-phase servers of the stations in Figure 3 contain load-dependent phases. We did not encounter this situation in [6] since the phase construction was only done for the selected station and not for a composite station representing a subnetwork. This situation is solved as follows. As in [6] for the selected station we make the multi-phase servers computationally tractable by reducing the service and delay phases of each station to a Coxian server with two phases. The load-dependence is considered by a computation of the Coxian server under all possible loads. For the given example the following calculations are necessary:

$$\hat{\mu}_{j1}(k) = 2 \cdot \hat{\mu}_j(k)$$

$$\hat{\mu}_{j2}(k) = \hat{c}_j^2(k) \cdot \hat{\mu}_j(k)$$

$$\hat{a}_{j1}(k) = \frac{1}{2 \cdot \hat{c}_j^2(k)}$$

for $j = 1, 2, 3, 4$; $k = 1,2,...,K$

where K is the total number of jobs in the network. The values for $\hat{\mu}_j(k)$ and $\hat{c}_j(k)$ are obtained by calculating mean value and coefficient of variation of the multi-phase servers. This is straightforward since the phases are exponentially distributed and independent from each other. As a result we obtain a load-dependent server with a Coxian-2 service time distribution for each composite station. From [8] we know that this representation of a general service time distribution provides good results as long as the squared coefficient of variation for each $\hat{c}_j^2(k)$ are greater than 0.5. In [6] only one subnetwork was constructed for the analysis. That one subnetwork was then analyzed by the numerical method. In this case, however, the numerical method is not applicable because of the dimension of the state space. By applying the load-dependent method of Marie [8] we obtained very accurate results for performance measures.

The applicability of the extended parametric analysis has numerous advantages. We mentioned the acceleration of computing time by carrying out subnetwork computations in parallel. Furthermore, some existing algorithms for very large finite capacity queueing networks with have overflows during the computation. Using the extended extended parametric analysis large networks can be decomposed into subnetworks and therefore they can become tractable for the analysis.

An application of the extended parametric analysis is given in [9]. There, we investigate the performance of internetwork traffic in a communication network where Local Area Networks (LANs) are connected using a Long Haul Network. We abstract from the individual structure of a LAN consisting of a possibly large number of nodes by defining each LAN as a subnetwork which is - applying the described method - reducable to one composite station. Therefore, we are able to parametrize each LAN, thus allowing us to focus our study on the internetwork traffic without neglecting intranetwork traffic in the LANs.

3. Rejection Blocking

Within this project we also attacked queueing networks with rejection blocking. In this case blocking occurs when a job after completing its service finds its destination full. The job is rejected and goes immediately back to the server of the source station and receives another round of service. This is repeated until a space becomes available in the destination station. This type of blocking protocol appears in communication networks. For example, in a token ring network suppose a source node has the token and wants to send a message to a destination node. If the destination node does not receive the message, the value of the token remains unchanged and the token comes back to the source node which then assumes that its message has not been received. The source node tries so many times until the message will be received. This situation can be modeled by a queueing network with rejection blocking.

In [10] we give an algorithm for computation of performance measures in reversible networks. We also analyze non-reversible queueing networks in [10] where the analysis is based on the duality concept. Based on this concept we give exact solutions for performance measures. *This paper [10] has been presented at the International Workshop* "Queueing Networks with Blocking" in May 1988 and it is appeared in the Procceedings of that conference.

In [11] we studied queueing networks with mixed exponential and non-exponential parallel queues with interdependent service capacities and finite common pools and/or accessibility constraints. We showed that this type of networks have exact solution.

Note that our basic concept of parametric analysis as well as extended parametric analysis can easily be applied with some minor modifications also on networks of queues with rejection blocking.

References

1. I. F. Akyildiz, "Parametric Analysis of Queueing Networks with Blocking", *Proposal to Airforce Office of the Scientific Research, AFOSR, 1987.*
2. I. F. Akyildiz and S. Kundu, "Deadlock Free Buffer Allocation in Closed Queueing Networks", *to appear in "Queueing Systems: Theory and Applications" Journal.*
3. I. F. Akyildiz and G. Bolch, "Throughput and Response Time Optimization in Queueing Network Models of Computer Systems", *Proc. of Int. Conference on Distributed and Parallel Systems,* North Holland, December 1988, pp. 241-259.
4. I. F. Akyildiz and W. Liu, "Selecting Buffer Capacities in Cyclic Queueing Networks with Blocking", *Short Note; Submitted for Publication.*
5. I. F. Akyildiz, "Product Form Approximations for Queueing Networks with Multiple Servers and Blocking", *"IEEE Transactions on Computers", January 1989, pp. 99-114.*
6. I. F. Akyildiz and J. Liebeherr, "Application of Norton's Theorem on Queueing Networks with Finite Capacitics", *Technical Report, Georgia Tech, ICS-GIT-88-024, July 1988.* Accepted to INFOCOM 89, (Int. Computer Networking Conference) in Ottawa, Canada, April 1989.
7. I. F. Akyildiz, G. Bolch and M. Paterok, "Parallel Computation of Performance Measures in Computer Systems" *Submitted to "Parallel Computing Journal" in July 1988;* Revision is requested in January 1989; The revised version will be submitted in March 1989.
8. I. F. Akyildiz and A. Sieber, "Approximate Analysis of Load Dependent General Queueing Networks", *IEEE Transactions on Software Engineering,* November 1988, pp. 1537-1546.
9. J. Liebeherr and I. F. Akyildiz, "Performance Analysis of Gateways in Computer Networks", *Technical Report, Georgia Tech, (in preparation).*
10. I. F. Akyildiz, "Analysis of Queueing Networks with Rejection Blocking", Proc. of the Int. Workshop "Queueing Networks with Blocking", ed. H. G. Perros and T. Altiok, North Holland Publ. Co., pp. 24-36, May 1988.
11. N. van Dijk and I. F. Akyildiz, "Networks with Mixed Processor Sharing Parallel Queues and Common Pools", *Technical Report, Georgia Tech, ICS-GIT-88-22, June 1988.* Submitted for Publication.

CONFERENCES VISITED (Supported by AFOSR)

1. **Performance 87 Conference in Brussels/ Belgium in December 7-9, 1987.**

Presented the paper *"General Closed Queueing Networks with Blocking"*,

Acceptance rate was 20%.

2. **INFOCOM 88 Conference in New Orleans, LA, in March 28-31, 1988.**

Presented the paper *"Performance Analysis of Computer Communication Networks* with Local and Global Window Flow Control",

Acceptance Rate was 60%.

3. **International Workshop on "Queueing Networks with Finite Capacities" in Raleigh, NC, in May 19-21, 1988.**

Presented the paper *"Analysis of Queueing Networks with Rejection Blocking"*,

Acceptance Rate was 50%.

4. **International Conference on "Analysis and Optimization of Large Scale Stochastic Systems", in Chapel Hill, NC, in May 23-25, 1988.**

Presented

i) "On Optimal Performance Measures of Computer Systems"

ii) "Open, Closed and Mixed Queueing Networks with Rejection Blocking"

Both papers were invited.