Stochastic Rendezvous Analysis on Multi Server Markovian Queueing Network

SivaselvanKasilingam and C. VijayalakshmiSeshadhri

***Abstract*—**Over the last few years have seen a incredible growth in the area of web-applications such as electronic commerce, web service, search engines etc., Multi-tiered system architectures have a client tier to provide an interface to the end users, a business logic tier to synchronize information salvage and processing. One of the most important quality of service parameters for these applications is expected response time, which is the total time it takes a users request to be processed. In multi-programmed computer system, scheduling plays a key role to optimize the CPU utilization. Thus, CPU scheduling is vital for the operating-system design. CPU scheduling determines which process runs when there are multiple processes. CPU scheduling is important because it can have a big effect on resource utilization and the overall performance of the system. The Multi-Layer Feedback Queue (MLFQ) can be considered as the apt scheduling algorithm for multi queues with different quanta. In MLFQ the optimized number of the queues is not define, and quantum of each queue, which is the starvation of this algorithm. In distributed multi-server network in which the customer transitions have exemplified by more than one closed Markov chain. Generating function has implemented to derive closed form of solutions and product form solution with the parameters such as stability, normalizations constant and marginal distributions.

***Keywords*—**Closed and Open Sub Chain, Marginal distribution, Markov chain, Multi-Layer Feedback Queue, Multiserver Queueing Network, Scheduling, Transition.

I. Introduction

RECENTLY communication infrastructure plays vital role in applications that share parts of the infrastructure. Web-based multi-tiered system architecture is the best example for where a client tier provides an interface to the end users and a logic tier to synchronize information salvage and processing, and a data tier with legacy system to store and access the customer data. In many applications servers, like Web servers file servers, database servers, a huge amount of transaction has to be handled properly in a specified time limit. Each transaction typically consists of several sub-transactions have to be processed in a fixed sequential order. By implementing a thread-pools, which perform a specific type of sub-transaction. In thread-pool size, the maximum number of threads execute the transactions simultaneously which explicit the optimization performance. An important application of this model is that at any moment in time the active threads share a common Central Processing Unit (CPU) hardware in a P.S fashion. If the multi-programmed jobs exceeds the number of threads in CPUs, a queue or waiting line is formed which can be managed by the CPU scheduling algorithm. In Multi-Server Queue scheduling algorithm, the queue is partitioned into several queues, each has its own scheduling, and its process assigned permanently to one queue. The processes can be moved to different queues in Multi-Server Queue scheduling. So that the process I/O bound moved to high priority queues and required a large amount of CPU time for low priority queue. The Multi-Server Queue scheduling is focused on maximum utilization of CPU and minimize the queue delay. For the distributed computer network model, an algorithm is designed in which the customer transitions have characterized by more than one closed Markov chain. A solution of product form algorithm is derived in the case of multiple closed sub chains and computational algorithm is presented for general class of queueing networks. The result is generalized to a queueing network in which the customer routing transitions are characterized by a Markov chain decomposable into multiple sub chains. Several aggregate states and their marginal distributions are discussed in conclusion.

K.Sivaselvan Assistant Professor, Department of Mathemaitcs Jeppiaar Engineering College,Chennai, India. Email: (sivajpr@gmail.com)

C.Vijayalkshmi Professor,Department of Mathematics, School of Advanced Sciences, VIT University, Chennai,India.Email:(vijusesha2002@yahoo.co.in)

II. Literature Survey

M. Andrews (2004).et.al., had developed the concept of the Scheduling in a Queueing system with asynchronously varying service rates. S.Balsamo (2001), et.al., had explained the analysis of queueuing networks with blocking. P.Cremonesi et.al.,(2002) had enlightened the approximate solution of closed multiclass queuing networks. M.Harchol-Baltere et.al.,(2005) has evidently envisaged the Multi-server queueing systems with multiple priority classes.W.K. Ehrlich.et.al.,(2001), had clearly explained the performance of web servers in a distributed computing environment. J.R.Ramos,et.al.,(2003)has approached an improved computational algorithm for Round Robin Service. The concept of node decomposition based approaches for multi-class closed queuing networks has envisioned by K.Satyamet.al.,(2005). M,Reiser et.al.,(1980) had analyzed multichain closed queueing networks. S.Stidham (2002) has clearly explained the methodology to design and control of queueing system. L.Tadj et.al.,(2005)had explained Optimal design and control of queues.

III. SCHEDULING

A scheduling algorithm is designed to optimize the performance measured which includes the following:
(i) Maximize CPU utilization
(ii) Minimize waiting time and response time
(iii)Minimize turnaround time

The following are well-known algorithms used in CPU scheduling.

A. First Come First Serve Algorithm:

The processes depend on the arrival time in the queue. Operating system runs the processes in queue depending on their arrival time, without regarding to the priority or the burst time of the processes.

B. Shortest Job First Algorithm:

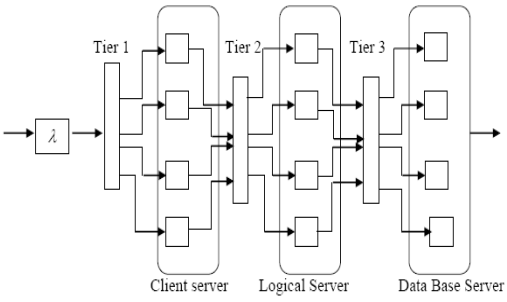
The processes depend on the shortest burst time in queue. Operating system runs the processes in queue depending on their burst time, without regarding to the priority or the arrival time of the processes.

C. Priority Algorithm:

The processes depend on the priority number in the queue assigned to each process. Operating system runs the processes in queue depending on their priority, without regarding to the arrival time or burst time of the processes.

D.Round Robin:

In the round robin scheduling, processes are dispatched in a FIFO manner but are given a limited amount of CPU time called a time-slice or a quantum. Round Robin fashion runs the algorithm based on the length of the quantum that will drastically decrease the waiting time compared with the other scheduling algorithm.



FCFS and SJF are examples for non-primitive algorithms. Scheduling primitive algorithms like MLFQ and RR, which provides response time and fair dispatching of CPU time. Round Robin function technique will provide a proper response time and no starvation with low overhead. Feedback scheduling algorithm has not possibility for starvation since this algorithm gives better results of I/O bound processes and there is no importance for throughput and response time.

Scheduling Algorithm	CPU Utilization	Through put	Turn around time	Response time	Deadline Handling	Starvation free
FCFS	Short	Short	Soaring	Short	No	Yes
SJF	Medium	Soaring	Medium	Medium	No	No
Priority	Medium	Short	Soaring	Soaring	Yes	No
Round Robin	Soaring	Medium	Medium	Short	No	Yes

IV. STOCHASTIC ASSIGNATION NETWORK

A set of interconnected queues that describes a multi-class queueing network, where a job moves from a queue to another queue with some probability after getting a service. A multiple class of customer could be open or closed where each class has its own set of queueing parameters. A closed queuing network model is suitable for large number of job arrivals. These parameters are obtained by analyzing each station in isolation under the assumption that the arrival process of each class is a state-dependent Markovian process. In inter-task closed queueing networks jobs must synchronized with other jobs at some point which is called as Stochastic Assignation Network (SAN). When a customer enters a fork or join queueing node, it will generate a fixed number of independent jobs. These jobs are served at different sub queues with different service time distributions. The entire sub queues have been served completely, then only the customer leave the system. Queueing network with finite capacity queues to represent the system with finite capacity resources and population constraints. When a queue reaches its maximum capacity then the flow of customers is blocked into the service centre and external cause in networks. The various blocking mechanisms Blocking Before Service (BBS), Repetitive Service Blocking (RBS),Blocking After Service(BAS) are analyzed to represent the behavior of queuing systems.

V.NETWORK DECOMPOSITION

The decomposition principle which performs the following steps based on the aggregation theorem: (i) network decomposition into a set of sub networks (ii) analysis of each sub network in isolation to define an aggregate component. The throughput of the closed network with Blocking Before Service(BBS) is approximated by a network decomposition method. The network is partitioned into M one –node sub networks. In next step each sub network is analyzed in isolation as M/M/1/ Ni network with arrival rate $\wp_i^*$ and load dependent service rate $\Psi_i^*(n), 0 \le n \le N_i$ to derive the marginal queue length distribution $\Re_i^*(n), 0 \le n \le N_i, 1 \le i \le E$. Consider two cases depending on whether all the nodes have finite capacity or there is one infinite capacity node, denoted by 1.

$$\Psi_i^*(n)=\left\{\left(\frac{1}{\Psi_i}\right)+\sum_{i=1}^{E} b_{ij}(n)\left[\sum_{k=i+1}^{j}\left(\frac{1}{\Psi_k}\right)\right]\right\}^{-1} \quad 1\le i\le E\text{-}1, 1\le N_i\le E\text{-}1 \text{ .....(1)}$$

$$\Psi^* E^{(n)}=\Psi_M \text{ , } 1\le N_i\le E\text{-}1\text{.....(2)}$$

$$\wp_i^*=\frac{X}{(1-\Re_i^*(N_i))} \text{ .....(3)}$$

where X is the network throughput and $b_{ij}(n)$ denotes the probability that nodes i+2……j are full ,given n customers in node i, $1\le i, j\le E$.

A. Single Chain Closed Sub Networks:

The multi-chain closed queueing network is decomposed to single-chain sub networks. Under the arrival of Poisson process , each server is analyzed in isolation with the same chain-dependent service times as in the original problem. The solution from each single server isolated queue provides the parameters for the service in each single chain sub network.

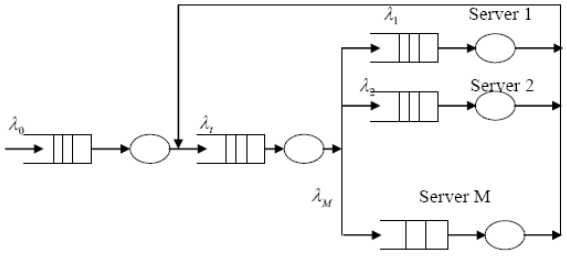


Fig. 1 Single chain closed subnetworke

The closed queueing network is decomposed into N single chain closed queueing sub networks. Each sub network consists of the M server queues and one client as shown in figure 1. There is only one customer in each sub network. The arrival rate of class j at server queue k, j = 1,2,….Kn condition on the queue having zero customers of chain n is given by $\Delta_{(n,j,k)}=\dfrac{r_{(n,j,k)}}{\sum_j\sum_k r_{(n,j,k)}M_{(n,j,k)}+\sum_j Z_{(n,j)}}$

where M(n ,j ,k) signifies the mean service time of a class j customer at server k and Z(n,j) symbolize the mean service time of a class j customer at the nth client.The message from nth client are synchronous , the arrival rate of class (n,j) customers at queue k, conditioned on the queue having i customers of chain n, is 0 for $i\ge 1$. All the customers are served in a First in First Out (FIFO) manner.

VI. Design and Analysis

Consider a queueing network system consisting of K service centers as $1\le i\le K$. There exist R different classes of customers and their transition are presiding over from one state to another by a first-order Markov chain M. $P_{ir(i'r')}$ represent the probability that a customer of class r which completes service at centre i will go to service centre i′ and changes to class r′ in M. Then the Markov chain M is decomposable into L sub chains M1,M2,…,ML which are all irreducible. Assume that without loss of generality $R\ge L\ge 1$. The sub chains M_l's are either open are closed. The sub chains are driven by L independent Poisson arrival streams with rate $\lambda_l(n_l)$, where n_l is the total number of customers in sub chain M_l at a given system. Representation of general service time distribution in the method of stages is, only one stage can accommodate a job at a given time .In FCFS discipline, a customer waiting at the head of the line is not allowed to enter the first stage until the job currently in service completes its last stage and departs from this centre. That is the entrance stage is blocked as longs as a job exists in the some stage. The steady state distribution provides a solution in a product form when it is not blocking. In the situation of blocking, the solution is complicated for the queueing system. Hence the service centre is assumed to be a queue dependent exponential server in FCFS discipline.

In Processor sharing (PS) queue discipline , the problem of blocking in the exponential server could not exist. In a multi server queue, there are many servers available than jobs and no waiting line is formed, thus blocking is nonexistent. In an infinite server queue, where the service rate is lowered according to the number of jobs in the centre at a given time.In FCFS, when a new job enters, the first stage of the server is provided to it. If a new job is entered prior this has been served by its own server. If a job is stayed at any stage restart the service among those remaining in the system. When a new job entered the service centre without blocking, this leads to provide a product form solution. In Processor Sharing(PS) and Last Come First Serve(LCFS),the service stages are specified in the system.

VII. Class –Aggregation Technique:

When the customer class is large, the steady-state probabilities of the Markov chain is difficult to calculate for large class of customers associated with synchronization station. Aggregation technique is intended to reduce the complexity of analysis in multi class queuing network. With respect to class j customers, the customers of the other classes can be consider as resources that are required in order for class j customers to proceed through the synchronization station. The synchronization mechanism forces every class j customer to synchronize with a customer of every other class , condition with at least one customer of each of the other classes be present class. The arrival process of class j in the jth aggregate synchronization is same as that of original synchronization station. $\lambda_j^j(n_j)=\lambda_j(n_j)$ for j = 1,2,3,….,R and nj = 0 ,1,2…..,Nj-1. The arrival process is modeled by a Markovian process whose rate $\lambda_a^j(n_a)$ depends on the total number of customers present in the queue. The service rate is equal to the throughput X(n) of the sub network analyzed in isolation with n customers, for each n. The aggregated network is obtained by substituting the sub network with the aggregated node. The aggregated network and the parent network have the same marginal queue length distribution and

average performance indices. Exact aggregation in queueing networks holds for any sub network. For multiple entry and exit points of sub networks there exists a new routing matrix for the aggregated network and for multi chain network.

VIII. Conclusion

This paper shows that the interactive behaviors in multiserver queueing network which has been modeling by a stochastic technique. The stochastic isolatiodecomposition technique is to maximizing the scalability and minimizing the complexity in large networks. In distributed multi-server network in which the customer transitions have exemplified by more than one closed Markov chain. Generating function has implemented to derive closed form of solutions and product form solution with the parameters such as stability, normalizations constant and marginal distributions.

Acknowledgment

I wish to express my gratitude and thanks to my parents, family members and my guide Dr.C. Vijayalakshmi for their valuable support and cooperation extended to design this model in a successful way.

References

[1] Andrews,M., Kumaran.K., Ramanan.K., Stolyar.A.,Vijayakumar.R and Whiting.P: Scheduling in a Queueing system with asynchronously varying service rates, Probability in the Engineering and Informational Sciences, Vol.18,no.02,(2004),pp.191-217.

[2] Balsamo.S, De Nitto Personè.V, Onvural.R, Analysis of Queueuing Networks with Blocking, Kluwer Academic Publishers, Dordrecht, (2001).

[3] Boxma.O.J. and Daduna.H, Sojourn times in queueing networks, In : Stochastic Analysis of Computer and Communication Systems, Elsevier Science Publishers, 1990.

[4] Cremonesi.P, Schweitzer.P.J, and Serazzi.G A unifying framework for the approximate solution of closed
multiclass queuing networks. IEEE Trans.Comp., 51:1423 1434,(2002).

[5] Harchol-Balter. M., Osogami. T., Scheller-Wolf. A and Wierman. A.,Multi-server Queueing systems with
Multiple priority classes, Queueing Systems: Theory andApplications 51 (2005) ,331-360.

[6] Ehrlich.W.K, Hariharan.R, Reeser.P.K and Van der Mei.R.D., Performance of Web servers in a distributed computing environment, In : Teletraffic Engineering in the Internet Era, Proceedings ITC-17 (Salvador-de Balia,Brazil), 137-148, 2001.

[7] Konheim.A, Meilijson.I and Melkman.A, Processor-sharing of two parallel lines, Journal of Applied Probability,18, 952-956, 1981.

[8] Ramos,J.R., Rego,V., and Sang. J : An improved computational algorithm for Round Robin Service. Proceeding of the 2003 Winter Simulation Conference,Dec.7-10,IEEE -Xplore Press, USA ,pp : 721-728.

[9] Reiser. M. and Lavenberg. S., "Mean value analysis of closed Multichain queueing networks",J. ACM 27 (2), 13–322(1980).

[10] Satyam.K, A. Krishnamurthy.A, and Kamath.M, "Node decomposition based approaches for multi-class closed queuing networks,"Technical Report Decision Sciences and Engineering Systems, (2005).

[11] Shakkotai.S and Stolyar.A. : Scheduling for multiple flows sharing a time-varying channel. The exponential rule.Transactions of the AMS, series 2, A volume in memory of Karpelevich.F., (2002) 207: 185-202.

[12] Stidham. S., "Analysis, design and control of queueing systems",Operations Research 50 ,197–216 (2002).

[13] Siva Selvan .K and Vijayalakshmi C.: Design and Analysis of Multi server queueing model networks for webbased system, Proceedings of International conference – ICOREM, Anna University, Thiruchirapalli (2009), PP.1296-1313.

[14] Siva Selvan .K and Vijayalakshmi C.: Algorithmic Approach For the Design Of Markovian Queueing Network with Multiple Closed Chains InternationalConference on TRENDZ information Sciences and Computing. Proceedings IEEE xplore, Sathyabama University ,TISC-2010.

[15] Sleptchenko, A. Harten and M. Heijden, An exact solution for the state probabilities of the multi-class, multi-server queue with preemptive priorities Queuing systems, Queueing Systems: Theory and Applications 50, (2005)81-107.

[16] Tadj, L. and Choudhury, G. (2005). Optimal design and control of queues. Top, 13, 359–412