International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 4, April 2014 ISSN 2319 - 4847

DISTRIBUTED QUEUEING SYSTEM WITH FREE CONSERVATION: A MODEL FOR MOBILE COMMUNICATION NETWORK

K. M. Sharma* and Yogesh Kumar Bhardwaj

B.S.A. College, Mathura (U.P.) India

Abstract

Queueing system considered in this paper suits for investigation of the two following situations appearing in telephone networks. The first situation is typical for distributed telephone networks where the corresponding clients are not connected by the direct physical channel and the preliminary establishing of a connection is required. During this phase of serving the client call, reservation of some physical channels is performed and these channels become are unavailable for other calls. It may occur that at the end of this phase the client who is destination of this call or its area are busy. In this case the work related to preliminary establishing the connection is lost. The service provider does not get any profit from this call but he probably has a loss due to the possible rejection of other calls due to the busyness of channels reserved for this call. The second variant concerns mobile communication networks when provider promises that several first, seconds of conversation are free of charge for a client. Such a promise attracts potential users of the network. But at the same time it creates problem to the provider. Some clients make excessive use of possibility to have free conversation. Just before the free time expires, they stop conversation and make a new call. So, they do not pay for their calls and probably create problems with access to the network of other clients who are ready to pay for the calls but meet the server be busy.

Keywords: Mobile communication networks, Queueing system; Excessive use of possibility, Free conversation.

1. Introduction

The parallel system's total processing rate will fall below m when there are fewer than m jobs available. How close, then, can we get to matching the performance of the fast processor with the corresponding set of slow processors, and how should we schedule the parallel system to achieve its best performance? These issues are significant in the design and operation of complex service systems, such as flexible manufacturing systems and computer communication networks **Burke (1956), Finch (1959), Papangelou (1972)**. In this paper we address such problems in the idealized setting of a versatile service system model: a multiclass M/M/m queue with Bernoulli feedback. We shall thus consider the problem of allocating dynamically m identical servers to customers in distributed M/M/m queueing network to minimize a performance objective of expected linear holding costs, where $E_u[L_j]$ represents the steady-state expected number of class j customers in the system under policy u, and $C_j > 0$ their holding cost rate. Admissible scheduling policies make history-dependent decisions, allow customer preemptions, and are nonidling (no server can lie idle when there are customers waiting). Consider now the corresponding problem in which the m *slow* parallel servers are replaced by a *pooled resource* consisting of one *fast* m-fold speed single server. While the parallel-server optimal scheduling problem is likely to be computationally intractable, the solution for the pooled resource constitutes a classical result in the field of stochastic scheduling: Klimov (1974, 1978) showed that the optimal policy is characterized by class-dependent priority-indices, efficiently computed by an *adaptive greedy algorithm*, so it is optimal to give at each decision epoch higher service priority to a customer with larger index. Clearly, Klimov's rule extends naturally into a simple heuristic for the parallel server system: At each decision epoch, let servers select preemptively available customers with larger indices. The current paper investigates the performance of this heuristic. In related work Weiss (1990, 1992, 1995) has analyzed the performance of index-based heuristics in several models for the optimal scheduling of a *batch* of stochastic jobs on parallel machines. He has argued that the index rules considered, which may be thought of as policies whose aim is to drive down fastest the cost rate of waiting jobs, are suboptimal because of an *end effect* caused by the loss of processing efficiency when the number of machines exceeds that of jobs present. He was able to bound the magnitude of this effect by deriving and applying certain decomposition formulae for the system's total expected workload. He thus obtained suboptimality bounds, independent of the batch size, for the index rules considered. Asymptotic optimality as the batch size grows to infinity follows **(**Boxma, O. J.; 1989). Weiss further argued the importance of proceeding to analyze index rules in more complex models incorporating job arrivals, such as queueing networks. This is the task we undertake in the present paper.

In our analysis of the performance of Klimov's rule in the above multiclass M/M/m system we shall focus on the following issues.

1. Approximate optimality. How far from the optimal cost can the expected cost under Klimov's rule be? How large can the gap be between the expected cost achieved by Klimov's rule in the parallel and in the pooled systems? Can one obtain simple bounds for the corresponding gaps?
2. Heavy traffic optimality. Does the relative suboptimality gap for Klimov's rule vanish in heavy traffic, as arrival rates approach system capacity? Our findings support the claim that Klimov's rule is a good heuristic for the parallel-server system:
We show that both its sub optimality gap and the gap between its expected cost in the parallel and in the pooled systems are uniformly bounded with respect to (i) external arrival rates, as long as they stay within system capacity; and (ii) the number of servers. The first such uniform boundness result implies its *heavy-traffic optimality*, in the following sense: The *relative* sub optimality gap of Klimov's rule vanishes as external arrival rates approach system capacity. We note that this notion of heavy-traffic optimality is not the standard one in the literature on queueing systems control (cf., Harrison 1998), where one typically considers the asymptotic behaviour of a sequence of systems appropriately scaled in time and space. The form of heavy-traffic optimality established in this paper is technically simpler, yet we believe it has the advantage of being intuitive.
In fact, we establish a stronger result, namely that the relative gap between the expected performance of Klimov's rule in the parallel and in the pooled systems vanishes in heavy traffic, in the sense stated above. The fact that intelligent dynamic scheduling of a queueing network may lead to an effective pooling of processing resources in heavy traffic has been studied in a variety of models (see, e.g., the review paper by Kelly and Laws 1993). However, as pointed out by Harrison (1998), "studies of resource pooling have been largely heuristic to date." Harrison proves a resource pooling result, and establishes a strong form of heavy-traffic optimality for a specific policy in the context of a model different from the one discussed here.
The approach in this paper to a rigorous development of a resource pooling/heavy-traffic optimality result is radically different and is based on an analysis of the system's region of achievable mean queue lengths. We believe that this approach has the potential to be extended to more complex systems. Our mode of analysis is the so-called achievable region approach to stochastic optimization. This approach was introduced in a seminal paper by Coffman and Mitrani (1980) and has since been extended to ever more encompassing frameworks in Gelenbe and Mitrani (1980), Federgruen and Groenevelt (1988), Ross and Yao (1989), Tsoucas (1991), Shanthikumar and Yao (1992), Bertsimas and Nirio-Mora (1996), and Glazebrook, K.D. and Garbe, R. (1999). In all these analyses, the agenda outlined in Equations (1)–(3) is carried through in full. Bertsimas and Nirio-Mora (1996) use the achievable region approach to unify classical priority index optimality results in a variety of problem domains, including deterministic machine scheduling (see Smith 1956), multi-armed bandits (see Gittins and Jones 1974), and multiclass queueing networks (see Klimov 1974, 1978).
The paper proceeds as follows. The parallel-server system that is our prime object of study is described in §2. To assist the reader we also give a brief account of an achievable region analysis of this system in the single-server (or pooled) case, when Klimov's rule is optimal. In §§3–5 we analyse the parallel server system according to the following three-step plan:

2. Mathematical Formulation of the Problem and Solution

The problem of enhancement of the system operation in these situations is very complicated. One of the possible ways for its solving is separation of two phases of the service of a customer (channels reservation or free phase of the call is the first phase and conversation, which is paid by a client, is the second phase) physically or virtually into separate stages with the controlled access to the first phase depending on the situation at the second phase. To simplify capacity planning, performance evaluation and parameters tuning for such a way of the system operation organization, the following queueing model looks to be useful.
Tandem queueing system consists of two sequential multi-server systems. We suggest that the number of servers at the first phase is infinite. Alternative assumption that it is finite, say V, simplifies investigation of the model. The number of servers at the second phase is assumed to be equal to $N, N \geq 1.$ The servers at each phase are suggested to be identical. Service time at a server at the k^{th} phase is assumed to be exponentially distributed with the rate $\mu_k, k = 1,2.$
The customers arrive to the system according to a MAP ($Markovian\,Arrival\,\Pr ocess$). The notion of the MAP and its detailed description is given by Lucantoni (1991) were the currently standard in literature denotations for the MAP are introduced. We denote the directing process of the MAP by $\nu_t, t \geq 0.$ The process $\nu_t, t \geq 0,$ is an irreducible continuous time Markov chain with state space $\{0,1,...,W\}$. The sojourn time of this chain in state ν is exponentially distributed with the positive finite parameter λ_ν. When the sojourn time in the state ν expires, with probability $p_{\nu,\nu'}^{(k)}$ the process ν_t jumps into the state ν' and k customers arrive, $k = 0,1, \nu, \nu' = \overline{0,W}$. The behavior of the MAP is completely characterized by the matrices $D_k, k = 0,1,$ defined by their entries

$(D_k)_{\nu,\nu'} = \lambda_\nu p^{(1)}_{\nu,\nu'}, k = 1$ and $k = 0, \nu \neq \nu', (D_0)_{\nu,\nu} = -\lambda_\nu, \nu = \overline{0,W}.$

The matrix $D(1) = D_0 + D_1$ represents the generator of the process $\nu_t, t \geq 0$. The average arrival rate λ is defined by $\lambda = \theta D_1 \mathbf{e}$ where θ is the invariant vector of the stationary distribution of the Markov chain $\nu_t, t \geq 0$. The vector θ is the unique solution to the system $\theta D(1) = \mathbf{0}, \theta \mathbf{e} = 1.$ Here and in the sequel $\mathbf{e}$ is the column-vector of appropriate size consisting of 1's and $\mathbf{0}$ is the row-vector of appropriate size consisting of zeroes. The set of MAP s is dense in the set of all point processes, hence, any ordinary arrival process can be approximated by a MAP. It is worth mentioning again that the MAP in general is a correlated process. Hence, it is suitable to model the flows in modern communication networks. The survey of research in the queues with the MAP can be found in Chakravarthy (2001).

Admission of customers to the system is controlled by means of a parameter $M, M = \overline{1,N}.$ If, at a customer arrival to the first system of tandem, the number of customers at the second system does not exceed the threshold M then the customer is admitted into the system and starts the service at the first phase. Otherwise, it leaves the system without service (is lost). After the service at the first phase, the customer leaves the system with probability $q, 0 \leq q < 1.$ With the supplementary probability $1 - q,$ the customers moves to the second system of tandem. If all servers of this system are busy, the customer leaves the system (is lost). Otherwise, it is processed by a server at the second system and then leaves the tandem.

We assume that the quality of the system operation is evaluated by means of the following cost criterion:

$$J(M) = a\lambda_{out} - c_1\lambda P^{(1)}_{loss} - c_2\lambda P^{(2)}_{loss} - d(N - M + 1), \quad \ldots.(1)$$

where λ_{out} is the rate of the flow of the customers which get successful service at the both systems of tandem, a is an average profit obtained by the system from the service of one customer, $P^{(r)}_{loss}$ is probability of a loss of a customer who wishes to get the service at the r th phase of the system, c_r is the charge of the system when a customer is lost at the r th phase of the system, $r = 1,2,$ d is a charge paid for reservation of one server at the second phase per time unit.

Our goal is to calculate the main performance measures of this tandem system and to show the effect of maximization of cost criterion (1) via an appropriate choice of the threshold M.

To this end, we consider the three-dimensional continuous time Markov chain $\xi_t = \{i_t, k_t, \nu_t\}, t \geq 0,$ where i_t is the number of customers at the first phase, k_t is the number of customers at the second phase at the moment t. Here $i_t \geq 0, k_t = \overline{0,N}, \nu_t = \overline{0,W}.$ Denote $\overline{W} = W + 1,\ K = (N+1)(W+1).$

Theorem:

Generator A of the Markov chain $\xi_t, t \geq 0,$ has blocking structure $A = (A_{i,j})_{i,j \geq 0}.$ Non-zero blocks $A_{i,j}$ of dimension $K \times K$ have the following form:

$$A_{i,i-1} = (qi\mu_1 E_N + i\mu_1(I_{N+1} - E_N) + (1-q)i\mu_1 \hat{E}_{N+1}) \otimes I_{\overline{W}}, i \geq 1,$$

$$A_{i,i+1} = E_M \otimes D_1, i \geq 0,$$

$$A_{i,i} = E_M \otimes D_0 + (I_{N+1} - E_M) \otimes D(1) - i\mu_1 I_K + \mu_2 \tilde{E}_{N+1} \otimes I_{\overline{W}}, i \geq 0.$$

where

$$E_l = \begin{pmatrix} I_l & 0_{l \times (N+1-l)} \\ 0_{(N+1-l) \times l} & 0_{(N+1-l) \times (N+1-l)} \end{pmatrix}, l = \overline{0,N},$$

$\hat{E}_{N+1}$ is the square matrix of dimension $N+1$ which has entries

$$(\hat{E}_{N+1})_{n,n'}, n, n' = \overline{0,N},$$

such that $(\hat{E}_{N+1})_{n,n+1} = 1, n = \overline{0, N-1},\ (\hat{E}_{N+1})_{n,n'} = 0, n' \neq n+1, n, n' = \overline{0,N},$

$\tilde{E}_{N+1}$ is the square matrix of dimension $N+1$ which has entries $(\tilde{E}_{N+1})_{n,n'}, n, n' = \overline{0,N},$ such that

$$(\tilde{E}_{N+1})_{n,n} = -n, n = \overline{0,N},\ (\tilde{E}_{N+1})_{n,n-1} = n, n = \overline{1,N},$$

$$(\tilde{E}_{N+1})_{n,n'} = 0, n' \neq n, n' \neq n-1, n, n' = \overline{0,N},$$

I_R is identity matrix of dimension R, $\otimes$ is symbol of Kronecker product of matrices.

Let us denote $p(i,k,\nu) = \lim_{t\to\infty} P\{i_t = i, k_t = k, \nu_t = \nu\}, i \geq 0, k = \overline{0,N}, \nu = \overline{o,W}$ stationary probabilities of the Markov chain $\xi_t, t \geq 0.$ It can be shown that, due to restricted access of customers into the system, the stationary probabilities $p(i,k,\nu)$ exist for any choice of the system parameters. Let us enumerate probabilities $p(i,k,\nu)$ in the lexicographic order and form row vectors $\mathbf{p}_i$ of these probabilities corresponding to the value i of the first component of the Markov chain.

It is well-known that the probability vectors $\mathbf{p}_i$ satisfy the following system of linear algebraic equations: $(\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \ldots)A = \mathbf{0}$, $(\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \ldots)\mathbf{e} = 1.$

This system is infinite, so the problem of its solving is non-trivial. Effective and numerically stable algorithm for computing the vectors $\mathbf{p}_i, i \geq 0,$ of stationary probabilities for the Markov chain of the type, to which belongs the Markov chain under consideration $\xi_t, t \geq 0,$ is described in Klimenok et al. (2005). So, the problem of computing these vectors can be considered solved. Having these vectors been computed, we can calculate the main performance measures of the tandem system including the measures involved in the cost criterion $J(M)$:

- Probability $P_{loss}^{(1)}$ that an arbitrary customer will be lost at the first phase of the system is computed by
$$P_{loss}^{(1)} = \frac{1}{\lambda}\sum_{i=0}^{\infty}\mathbf{p}_i(I_{N+1} - E_M) \otimes D_1\mathbf{e},$$
- Average number $N^{(1)}$ of busy servers at the first phase of the system is computed by $N^{(1)} = \sum_{i=0}^{\infty} i\mathbf{p}_i\mathbf{e}.$
- Intensity $\lambda_{out}^{(1)}$ of flow of customers, which get the service at the first phase of the system, is computed by $\lambda_{out}^{(1)} = N^{(1)}\mu_1.$
- Average number $N^{(2)}$ of busy servers at the second phase of the system is computed by $N^{(2)} = \sum_{i=0}^{\infty}\mathbf{p}_i(\mathbf{0}, 1\mathbf{e}, 2\mathbf{e}, \ldots, N\mathbf{e})^T.$
- Intensity λ_{out} of flow of customers, which get the service at the second phase of the system, is computed by $\lambda_{out}^{(2)} = N^{(2)}\mu_1.$
- Probability $P_{loss}^{(2)}$ that an arbitrary customer will be lost at the second phase of the system (it leaves the system after the service at the first phase due to the busyness of all servers at the second phase) is computed by $P_{loss}^{(2)} = 1 - \frac{\lambda_{out}}{\lambda_{out}^{(1)}(1-q)}.$

4. Numerical Computation

Consider numerical example. Let the number N of servers at the second phase be equal to 50. Intensities of service are given by $\mu_1 = 2, \mu_2 = 0.1.$ Probability q is equal to 0.3. To illustrate the effect of maximization of the cost criterion $J(M)$ by means of appropriate choice of the threshold M and also to demonstrate importance of investigation of the system with the MAP arrival process, we consider three different arrival processes. Arrival rate λ of all these three processes is equal to 10. The first arrival process is assumed to be the stationary Poisson arrival process. It has correlation between the successive inter-arrival times equal to 0 and variation equal to 1. The second process has correlation between the successive inter-arrival times equal to 0.3868 and variation equal to 6.4788. The third process has correlation between the successive inter-arrival times equal to 0.4627 and variation equal to 15.26. The curves 1-3 on the Fig.1 present the dependence of cost criterion on the threshold M for these three arrival processes. Although arrival rate is the same, the dependencies are quite different. One can see, that the optimal values of the threshold M for these curves are equal to 47, 45, and 38 correspondingly. Optimal values of the cost criterion are equal to 1045.02, 817.47, 245.24 while the values of this criterion when no channel reservation is made at the second phase (i.e. $M = N = 50$) are equal to 714.45, 293.35, - 924.31 correspondingly. So, the optimal channel reservation gives essential profit.

5. Discussion of Results and Conclusions

We have analyzed a simple heuristic index policy that extends Klimov's classical solution for the single-server case to the general parallel-server model, presenting closed form sub optimality bounds that imply its asymptotic optimality in a heavy-traffic limit. Ideas that emerge from our analysis include the following:

(1) Understanding of a simple single-server system has yielded useful insights into the performance of its more complex parallel-server counterpart;
(2) Understanding the fundamental laws of a complex parallel-server model (flow balance and work decomposition) has yielded a key to its analysis;
(3) Investigating strong linear programming relaxations of a complex stochastic optimization problem has yielded an approximate and asymptotic analysis of a heuristic, which had resisted traditional approaches.

We believe these ideas, which guided our approach, should prove fruitful for addressing other complex stochastic optimization problems. We refer the reader back to the discussion in the paragraph following Lemma 5 for indications of further work to be done on the current model. In a companion paper (see Glazebrook and Nino-Mora 1999) we carry out a corresponding analysis of priority index rules for scheduling Markovian multiclass queueing networks with multiple service stations. Although such rules are known to perform poorly in general for the latter type of networks, in that paper we present suboptimality bounds under appropriate light-traffic conditions. It should be remarked that while Klimov's optimal solution for the single-server model applied to multiclass
M/G/1 networks, i.e., it was valid under general service time distributions, our analysis requires the latter to be exponential. Extending our approach to a model with general service time distributions and nonpreemptive policies would require the development and application of work decomposition laws for such a model. Carrying out this extension remains a challenging problem.

REFERENCES

[1] Chakravarthy, S. (2001): "The batch Markovian arrival process: a review and future work", Advances in Probability Theory and Stochastic Processes, Notable Publications, New Jersey, pp. 21–49.
[2] Klimenok, V., Kim, C.S., Orlovsky, D. and Dudin, A. (2005): "Lack of invariant property of Erlang loss model in case of the MAP input", Queueing Systems, Vol. 49, pp. 187–213.
[3] Lucantoni, D. M. (1991): "New results on the single server queue with a batch Markovian arrival process", Communications in Statistics-Stochastic Models, Vol. 7, pp. 1–46.
[4] Bertsimas, D. and Nino-Mora, J. (1996): "Conservation laws, extended polymatroids and multi-armed bandit problems: a polyhedral approach to indexable systems", *Math. Oper. Res., vol.* 21, pp. 257–306.
[5] Bertsimas, D. and Nino-Mora, J. (1999a): "Optimization of multiclass queueing networks with changeover times via the achievable region approach: Part I, the single-station case", *Math. Oper. Res., vol.* 24, pp. 306–330.
[6] Bertsimas, D. and Nino-Mora, J. (1999b): "Optimization of multiclass queueing networks with changeover times via the achievable region approach: Part II, the multi-station case", *Math. Oper. Res., vol.* 24, pp. 331–361.
[7] Boxma, O. J. (1989): "Workloads and waiting times in single-server systems with multiple customer classes", *Queueing Syst., vol.* 5, pp. 185–214.
[8] Burke, P. J. (1956): "The output of a queueing system", *Oper. Res., vol.* 4, pp. 699–704.
[9] Coffman, E. and Mitrani, I. (1980): "A characterization of waiting time performance realizable by single server queues", *Oper. Res., vol.* 28, pp. 810–821.
[10] Federgruen, A. and Groenevelt. H. (1988): "Characterization and optimization of achievable performance in general queueing systems", *Oper. Res., vol.* 36, pp. 733–741.
[11] Finch, P. D. (1959): "On the distribution of queue size in queueing problems", *Acta Math. Hungar., vol.* 10, pp. 327–336.
[12] Gelenbe, E. and Mitrani, I. (1980): *"Analysis and Synthesis of Computer Systems"*, Academic Press, London.
[13] Gittins, J.C. and Jones, D.M. (1974): "A dynamic allocation index for the sequential design of experiments", Gani, J., Sarkadi, K., and Vince, I., eds., *Progress in Statistics: European Meeting of Statisticians, Budapest, 1972*, North-Holland, Amsterdam, pp. 241–266.
[14] Glazebrook, K.D. and Garbe, R. (1999): "Almost optimal policies for stochastic systems which almost satisfy conservation laws", *Ann. Oper. Res., vol.* 92, pp. 19–43.
[15] Harrison, J.M. (1998): "Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies", *Ann. Appl. Probab., vol.* 8, pp. 822–848.
[16] Kelly, F.P. and Laws, C.N. (1993): "Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling", *Queueing Syst. Theory Appl., vol.* 13, pp. 47–86.
[17] Klimov, G.P. (1974): "Time sharing service systems I", *Theory Probab. Appl., vol.* 19, pp. 532–551.
[18] Klimov, G.P. (1978): "Time sharing service systems II". *Theory Probab. Appl., vol.* 23, pp. 314–321.

[19] Kumar, S. and Kumar, P.R. (1994): "Performance bounds for queueing networks and scheduling policies", *IEEE Trans. Autom. Control, vol*. 39, pp. 1600–1611.

[20] Papangelou, F. (1972): "Integrability of expected increments of point processes and a related random change of time scale", *Trans. Amer. Math. Soc., vol.* 165, pp. 483–506.

[21] Ross, K.W. and Yao, D.D. (1989): "Optimal dynamic scheduling in Jackson networks", *IEEE Trans. Aut. Control, vol.* 34, pp. 47–53.

[22] Shanthikumar, J.G. and Yao. D.D. (1992): "Multi-class queueing systems: polyhedral structure and optimal scheduling control", *Oper. Res., vol*. 40, pp. 293–299.

[23] Smith, W.E. (1956): "Various optimizers for single stage production", *Naval Res. Logist. Quart., vol*. 3, pp. 59–66.

[24] Tsoucas, P. (1991): "The region of achievable performance in a model of Klimov", Technical Report RC16543, IBM T.J. Watson Research Center, Yorktown Heights, NY.

[25] Weiss, G. (1990): "Approximation results in parallel machines stochastic scheduling Special Volume on Production Planning and Scheduling M. Queyranne (ed.), *Ann. Oper. Res., vol.* 26, pp. 195–242.

[26] Weiss, G. (1992): "Turnpike optimality of Smith's rule in parallel machines stochastic scheduling", *Math. Oper. Res., vol.* 17, pp. 255–270.

[27] Weiss, G. (1995): "On almost optimal priority rules for preemptive scheduling of stochastic jobs on parallel machines", *Adv. Appl. Probab., vol.* 27, pp. 821–839.

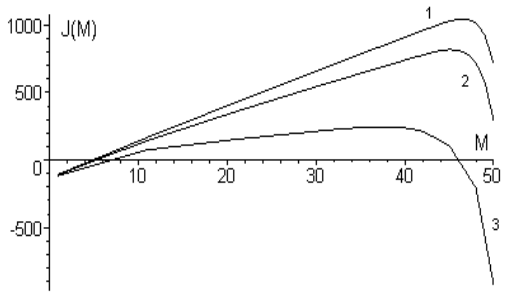

Figure 1: Dependence of cost criterion on the threshold M