Optimal Round-Robin Routing to Parallel Servers in Heavy Traffic

[Extended Abstract]

Heng-Qing Ye [*]

ABSTRACT

We study a system with heterogeneous parallel servers. Upon arrival, a job is routed to the queue of one of the servers. We establish the diffusion limit for the round-robin (RR) policy, and show that with properly chosen parameters, it achieves the optimal performance asymptotically over all blind routing policies. Analysis of the diffusion limit yields a number of insights into the performance of the optimal RR policies.

1. INTRODUCTION

Consider a queueing system with *heterogeneous* parallel servers, each with an infinite waiting room. One stream of jobs arrives at the system following a renewal process, and each job upon its arrival is routed immediately to one of the servers. At each server, the jobs are served according to the first-in first-out discipline, and their service times are independent and identically distributed.

This study aims to answer the question: what would be the optimal routing policy when the routing controller cannot observe any state information of the system? However, it is very difficult to answer this question through exact analysis, and to overcome this difficulty we apply the heavy-traffic analysis in the spirit of the BIGSTEP method (cf. [5]).

To do so, we first formulate the diffusion limit model $\hat{Q}(t)$ that approximate the (original, discrete) routing control problem. Then, we identify the best possible limit that uses no state information in routing. It turns out that we interpret a blind routing policy, which is indeed a generalized round-robin (RR) policy, from the "best" limit. Next, we justify the interpretation by establishing the diffusion limit theorem (Theorem 1). That is, under the RR policy, the sequence of diffusion-scaled systems does converge to the limit we identify. Finally, the stationary performance of the limit, $\hat{Q}(\infty)$, is taken as an approximation of the stationary performance of the (original, discrete) system of interest, $\hat{Q}^n(\infty)$, which yields an approximation of the performance objective, i.e., the expected stationary queue length, immediately.

Furthermore, examining the performance of the diffusion limit under the RR policy reveals interesting insights. For example, the optimal RR policy can attain the performance of the JSQ policy (the globally optimal policy under heavy traffic) if and only if all service times are deterministic. But, on the other hand, it could perform arbitrarily worse than the JSQ policy, say, when there are many servers in the system. Most surprisingly, the proportional RR policy, a conventional option of the RR policy, can be arbitrarily worse than the optimal RR policy.

Many studies in the literature use the round-robin routing policy or its variations for the parallel server system when the controller cannot observe any state information. In the case of identical servers, the conventional RR routing policy is widely used, which in its simplest form assigns the incoming jobs to each server equally in a rotating fashion. It is shown that it minimizes the long-run average total queue length in the system over all blind routing policies ([4, 1]). Refer to, e.g., [6, 7] for more related studies. Here, we extend the research to the case of heterogeneous servers. (Technical details omitted here can be found in [8], in which optimal routing policies are also established when various kinds of state information such as the job arrival history are available for making routing decision.)

2. MODEL AND PRELIMINARY

We consider a queueing system with $K(\geq 2)$ servers, indexed by $k \in \mathcal{K} = \{1, \cdots, K\}$, described in the previous section. Let $E(t)$ be the (exogenous) renewal arrival process, which denotes the number of arrivals during the time interval $[0, t]$. Assume the interarrival times have mean $1/\lambda$ and coefficient of variation c_a. Let $S(t) = (S_k(t))_{k\in\mathcal{K}}$ be the renewal service process, where $S_k(t)$ denotes the number of class-k service completions (job departures) after server k is busy for a total of t time units. Assume the service times have mean $1/\mu_k$ and coefficient of variation $c_{b,k}$.

To describe the routing of jobs, we define the routing process as $\Phi(\ell) = (\Phi_k(\ell))_{k\in\mathcal{K}}$, $\ell = 0, 1, 2, \cdots$, where $\Phi_k(\ell)$ is the number of jobs among the the first ℓ arrivals that are dispatched to the server k. The total number of jobs routed to servers must be equal to the total arrivals: $\sum_{k\in\mathcal{K}} \Phi_k(\ell) = \ell$.

The (generalized) RR policy is specified with the weight-parameter $p = (p_1, \cdots, p_K)$ satisfying $\sum_{k\in\mathcal{K}} p_k = 1$. By the policy, a fraction p_k of jobs is sent to the server k according to a pre-specified splitting sequence. That is, the sequence of arrivals should be split so that the number of jobs dispatched to each server k is "close" to its quota, a fraction p_k of the total arrival, at any time instance. More specifically, the RR policy should satisfy the following requirement: for some constant κ,

$$|\Phi_k(\ell) - p_k\ell| < \kappa, \quad \ell = 1, 2, \cdots, \quad k \in \mathcal{K}. \qquad (1)$$

[*] Supported in part by HK/RGC Grant 15501421. Dept of Logistics and Maritime Studies, Hong Kong Polytechnic University, Hong Kong; lgtyehq@polyu.edu.hk.

A detailed implementation can be found in [8].

The objective of the routing control is to minimize the performance objective, expected stationary (total) queue length, i.e., $\mathsf{E}Q_{\mathcal{K}}(\infty)$ $(= \mathsf{E}\sum_{k\in\mathcal{K}} Q_k(\infty))$, where $Q_k(\infty)$ represents the stationary queue length of server k, if exists.

Let $Q(t) = (Q_k(t))_{k\in\mathcal{K}}$ be the queue length process, where $Q_k(t)$ denotes the number of jobs in queue k at time t. The number of arrivals routed to server k during $[0, t]$, is given as $\Phi_k(E(t))$, and satisfies the requirement:

$$\sum_{k\in\mathcal{K}} \Phi_k(E(t)) = E(t). \tag{2}$$

Let $B(t) = (B_k(t))_{k\in\mathcal{K}}$, where $B_k(t)$ denotes the busy time, i.e., total amount of the time that server k has served jobs during $[0, t]$. The number of service completions at server k up to time t is given as $S_k(B_k(t))$. Then, the dynamics of the queueing system is characterized by

$$Q_k(t) = Q_k(0) + \Phi_k(E(t)) - S_k(B_k(t)) \geq 0, \tag{3}$$

$$B_k(t) = \int_0^t 1_{\{Q_k(s)>0\}} ds. \tag{4}$$

The first equation is a balanced equation, and the second specifies a work-conserving condition. Define the idling processes $Y(t) = (Y_k(t))_{k\in\mathcal{K}}$ as follows,

$$Y_k(t) = \mu_k(t - B_k(t)) = \mu_k \int_0^t 1_{\{Q_k(s)=0\}} ds. \tag{5}$$

It is immediately observed from the above expressions:

$$\int_0^\infty Q_k(s) dY_k(s) ds = 0. \tag{6}$$

$$Y_k(t) \text{ is non-decreasing in } t \geq 0, \text{ and } Y_k(0) = 0. \tag{7}$$

To carry out the heavy traffic analysis, we introduce a sequence of systems, indexed by n that represents a sequence of numbers increasing to $+\infty$. Each system is like the one introduced above, but may differ in their arrival rates. For example, for the n-th system, the arrival process, the arrival rate and the server-k service rate are denoted as $E^n(t)$, λ^n and μ_k^n, respectively. The processes satisfying the relationships in equations (2-7) are appended with the index n properly, too.

Assume the sequence of systems are linked via the limit:

$$\lambda^n \to \lambda := \mu_{\mathcal{K}} \quad \text{and} \quad c_a^n \to c_a, \quad \text{as } n \to \infty,$$

and furthermore the *heavy traffic condition* is satisfied:

$$n(\lambda^n - \mu_{\mathcal{K}}) \to \theta_{\mathcal{K}} < 0, \quad \text{as } n \to \infty. \tag{8}$$

From now on, the parameter λ denotes the limit of λ^n rather than the arrival rate of a particular system. Moreover, the above condition implies that the (nominal) traffic intensity approaches one, $\rho^n := \lambda^n/\mu_{\mathcal{K}} \to 1$, as $n \to \infty$.

Define the diffusion scaling (along with centering) for the primitive processes:

$$(\hat{E}^n(t), \hat{S}_k^n(t)) = \frac{1}{n}\left(E^n(n^2t) - \lambda^n n^2 t, S_k^n(n^2t) - \mu_k n^2 t\right).$$

By the functional central limit theorem for the renewal process (e.g., [2]), we have the following weak convergence,

$$(\hat{E}^n(t), \hat{S}^n(t)) \quad \Rightarrow \quad (\hat{E}(t), \hat{S}(t)), \quad \text{as } n \to \infty,$$

where $\hat{E}(t)$ is a Brownian motion with zero mean and variance λc_a^2; and $\hat{S}(t) = (\hat{S}(t))_{k\in\mathcal{K}}$ is a K-dimensional Brownian motion with independent coordinates, whose kth coordinate, $\hat{S}_k(t)$, is a Brownian motion with zero mean and variance $\mu_k c_{b,k}^2$. $\hat{E}(t)$ and $\hat{S}(t)$ are independent.

For the routing process, denote formally:

$$\hat{\Phi}_k^n(t) := \frac{1}{n}\left(\Phi_k^n(\lfloor n^2 t\rfloor) - p_k^n \lfloor n^2 t\rfloor\right).$$

Now, a routing policy for the sequence of systems actually refers to a sequence of policies, with the n-th policy associated with the n-th system.

For the other derived processes, we write:

$$\left(\hat{Q}_k^n(t), \hat{Y}_k^n(t)\right) := \frac{1}{n}\left(Q_k^n(n^2t), Y_k^n(n^2t)\right).$$

Rewrite equation (3) for the n-th system as:

$$\begin{aligned} Q_k^n(t) &= Q_k^n(0) + X_k^n(t) + Y_k^n(t), & (9)\\ X_k^n(t) &= [\Phi_k^n(E^n(t)) - p_k^n E^n(t)] + p_k^n[E^n(t) - \lambda^n t] \\ &\quad -[S_k^n(B_k^n(t)) - \mu_k B_k^n(t)] + (p_k^n\lambda^n - \mu_k)t(10) \end{aligned}$$

Then, the dynamics given in equations (2-7) can be written as a Skorohod problem (e.g., [2]): for all $k \in \mathcal{K}$ and $t \geq 0$,

$$\hat{Q}_k^n(t) = \hat{Q}_k^n(0) + \hat{X}_k^n(t) + \hat{Y}_k^n(t) \geq 0, \tag{11}$$

$$\int_0^\infty \hat{Q}_k^n(s) d\hat{Y}_k^n(s) ds = 0, \tag{12}$$

$$\hat{Y}_k^n(t) \text{ is non-decreasing in } t \geq 0, \text{ and } \hat{Y}_k^n(0) = 0(13)$$

$$\hat{X}_k^n(t) = \hat{\Phi}_k^n(\tilde{E}^n(t)) + p_k^n \hat{E}^n(t) - \hat{S}_k^n(\tilde{B}_k^n(t)) + \theta_k^n t,$$

where $\theta_k^n := n(p_k^n\lambda^n - \mu_k)$ and $\sum_{k\in\mathcal{K}} \hat{\Phi}_k^n(t) = 0$.

For the parameters $\theta^n = (\theta_k^n)_{k\in\mathcal{K}}$ just introduced, assume that for some constants $\theta = (\theta_k)_{k\in\mathcal{K}} < 0$, the following holds

$$\theta_k^n \to \theta_k, \quad \text{as } n \to \infty. \tag{14}$$

The weight p_k^n can also be interpreted as the (approximate) routing rate to server k in the n-th system. Then, the (approximate) arrival rate to and traffic intensity of server k are then denoted as $p_k^n\lambda^n$ and $\rho_k^n := p_k^n\lambda^n/\mu_k$, respectively. Hence, the above condition requires that the arrival rates to the queues $p_k^n\lambda^n$ are within the service capacities (rates) μ_k and approach the capacities proportionally. This condition also implies

$$p_k^n \to p_k := \frac{\mu_k}{\mu_{\mathcal{K}}}, \quad \sum_{k\in\mathcal{K}} \theta_k = \theta_{\mathcal{K}}. \tag{15}$$

Unlike the parameter $\theta_{\mathcal{K}}$, which is given in (8) and is fixed, we have some room to adjust the parameters θ_k's when we try to find the optimal routing policy below.

3. OPTIMAL ROUND-ROBIN ROUTING

By observing the Skorohod representation of the systems in (11-13), we *expect* the weak convergence, $\hat{Q}^n(t) \Rightarrow \hat{Q}(t)$, where the limit (i.e., the diffusion limit) is the unique solution of the following Skorohod problem:

$$\hat{Q}_k(t) = \hat{Q}_k(0) + \hat{X}_k(t) + \hat{Y}_k(t) \geq 0, \tag{16}$$

$$\hat{Y}_k(t) \text{ is non-decreasing in } t \text{ with } \hat{Y}_k(0) = 0, \tag{17}$$

$$\int_0^\infty \hat{Q}_k(t) d\hat{Y}_k(t) = 0, \tag{18}$$

with $\hat{X}(t) = (\hat{X}_k(t))_{k\in\mathcal{K}}$, $\hat{X}_k(t) = \hat{\Phi}_k(\lambda t) + p_k\hat{E}(t) - \hat{S}_k(t) + \theta_k t$. Given that the routing process is independent of the arrival and service processes, the expected stationary queue length in the limit can be evaluated as (cf. [2]:

$$\begin{aligned}\mathsf{E}\hat{Q}_k(\infty) &= \frac{\mathsf{Var}(\hat{\Phi}_k(\lambda) + p_k\hat{E}(1) - \hat{S}_k(1))}{-2\theta_k}\\ &= \frac{\mathsf{Var}(\hat{\Phi}_k(\lambda)) + p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}{-2\theta_k}.\end{aligned}$$

By letting $\mathsf{Var}(\hat{\Phi}_k(\lambda) = 0$ and optimizing over $\{\theta_k\}$, we can derive a lower bound of expected stationary total queue length and the associated parameter,

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty) \leq \frac{\left(\sum_k \sqrt{p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}\right)^2}{-2\theta_{\mathcal{K}}}, \text{ with} \quad (19)$$

$$\theta^* = (\theta_k^*)_{k\in\mathcal{K}},\ \theta_k^* = \frac{\sqrt{p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}}{\sum_j \sqrt{p_j^2\lambda c_a^2 + \mu_j c_{b,j}^2}}\theta_{\mathcal{K}}. \quad (20)$$

Indeed, the following theorem shows that the round-robin routing policy, denoted as $RR(\theta)$, can achieve this lower bound with $\theta = \theta^*$ (cf. [8] for a proof).

THEOREM 1. *(a) Under the RR policy $RR(\theta)$ (and along with some regular conditions), the weak convergence described in (16-18) holds with $\hat{\Phi}_k(t) = 0$. The "free process" $\hat{X}_k(t)$ is a Brownian motion with drift θ_k and variance $(p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2)$. The expected stationary queue lengths is:*

$$\mathsf{E}\hat{Q}_k(\infty; RR(\theta)) = \frac{p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}{-2\theta_k}, \quad k \in \mathcal{K}. \quad (21)$$

(b) Let θ be set to $\theta^* = (\theta_k^*)_{k\in\mathcal{K}}$ given in (19), or alternatively, choose "routing rates" $\{p_k^n\}$ such that

$$\frac{\mu_k - p_k^n\lambda^n}{\mu_{\mathcal{K}} - \lambda^n} = \frac{\sqrt{(p_k^n)^2\lambda^n(c_a^n)^2 + \mu_k c_{b,k}^2}}{\sum_j \sqrt{(p_j^n)^2\lambda^n(c_a^n)^2 + \mu_j c_{b,j}^2}}. \quad (22)$$

Then, the RR policy $RR^* = RR(\theta^*)$ is asymptotically optimal: $\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR^*) \leq \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; H)$ for any blind policy H. Moreover, the expected stationary queue length is given as,

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR^*) = \frac{\left(\sum_k \sqrt{p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}\right)^2}{-2\theta_{\mathcal{K}}}. \quad (23)$$

Note that in part (b), $(\mu_k - \lambda^n p_k^n)$ and $(\mu_{\mathcal{K}} - \lambda^n)$ are the surplus capacities of the server k and the whole system, respectively; and $((p_k^n)^2\lambda^n(c_a^n)^2 + \mu_k c_{b,k}^2)$ is the combined variability owing to the arrival and service processes of the class k. Hence, under the optimal RR policy, the overall surplus capacity is distributed to each server in proportional to the square root of its combined variability. This is reminiscent of the *square-root rule* in various queueing models.

Below, we summarize key observations about the performance of the optimal RR policy RR^* (cf. [8] for details).

Comparison between the optimal RR policy and the JSQ policy. First, from the optimality of the JSQ policy (e.g., [3]), we know that the expected stationary queue length under the optimal RR policy RR^* cannot be smaller than the JSQ policy. On the other hand, by comparing their performances (cf. [3] for the performance under JSQ), both policies attain the same expected queue length when

$$c_{b,k}^2 = 0 \text{ for all } k \in \mathcal{K},$$

and the parameters θ_k^* and p_k^n specified in Theorem 1 can be simplified as:

$$\theta_k^* = p_k\theta_{\mathcal{K}}, \quad p_k^n = \frac{\mu_k}{\mu_{\mathcal{K}}}.$$

In other words, the optimal RR policy can attain the performance of the JSQ policy (and thus achieve the optimal performance over all non-anticipating policies) if and only if all service times are deterministic, and in this case, jobs are routed to each server in proportion to the service rate.

Next, we consider an example, in which all servers are identical, services times are exponentials, and arrivals follow the Poisson process. The expected queue lengths are reduced to

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR^*) = \frac{(1+K)\lambda}{-2\theta_{\mathcal{K}}}, \quad \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; JSQ) = \frac{\lambda}{-\theta_{\mathcal{K}}}.$$

Hence, when the number of servers K grows, the performance under the optimal RR policy can be arbitrarily worse than the JSQ policy.

Comparison between the optimal RR policy and the proportional RR policy. For the RR policy $RR(\theta)$, a conventional option is to distribute jobs to servers in proportion to the service rates, i.e., to set the parameters as $p_k^n = \mu_k/\mu_{\mathcal{K}}$, and thus from the condition in (14), $\theta_k = \theta_k' := (\mu_k/\mu_{\mathcal{K}})\theta_{\mathcal{K}}$. We call it the proportional RR policy, and its expected stationary queue length can be derived from Theorem 1.

First, the proportional RR policy is generally suboptimal within the class of RR policies. Second, we examine an example in which we assume Poisson arrival and exponential service and pick $p_1 = 1 - 1/K + 1/K^2$ and $p_k = 1/K^2$ for $k = 2, \cdots, K$. We show that the performance ratio $(\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR(\theta'))/\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR(\theta^*)))$ can be arbitrarily large as K increases. That is, the proportional RR policy can perform arbitrarily worse than the optimal RR policy.

4. REFERENCES

[1] Altman E., B. Gaujal and A. Hordijk. (2003). *Discrete-Event Control of Stochastic Networks: Multimodularity and Regularity*, Springer, Heidelberg.

[2] Chen H. and D.D. Yao. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization*, Springer, New York.

[3] Chen, H. and H.Q. Ye. (2012). Asymptotic optimality of balanced routing. *Oper. Res.*, **60**, 163-179.

[4] Hajek B. (1985). External splittings of point processes. *Math of Oper. Res.*, **10**, 543-556.

[5] Harrison, J.M. (1996). The BIGSTEP approach to flow management in stochastic processing networks. In *Stochastic Networks* (F.P. Kelly, S. Zachary and I. Ziedins, eds.), 57-90. Oxford Univ. Press.

[6] Liu, Z. and R. Righter. (1998). Optimal load balancing on distributed homogeneous unreliable processors. *Oper. Res.*, **46**, No.4, 563-573.

[7] Tsoukatos, K.P. and A.M. Makowski. (2006). Asymptotic optimality of the Round-Robin policy in multipath routing with resequencing. *Queueing Systems*, **52**, 199-214.

[8] YE, H.Q., Optimal Routing to Parallel Servers in Heavy Traffic. Available ssrn.com/abstract=3996615