The Price of Anarchy in an Exponential Multi-Server*

Moshe Haviv† Tim Roughgarden‡

Abstract

We consider a single multi-server memoryless service station. Servers have heterogeneous service rates. Arrivals are routed to one of the servers, and the routing decisions are not based on the queue lengths. We consider two criteria for routing selection: the (Nash) equilibrium, under which each customer minimizes his own mean waiting time, given the behavior of the others; and social optimization, where the routing minimizes the average mean waiting time across all arrivals. The ratio between the social costs of these two routings is called the price of anarchy (PoA). We show that the PoA is upper bounded by the number of servers used in the socially optimal outcome. We also show that this bound is tight.

1 Introduction

We consider a single multi-server service station with a number of not necessarily identical servers. We assume that customers' costs are their expected waiting times, and that the time in service is part of the waiting time. Each customer selects a server, and no further information (such as the queue lengths upon arrival) is given. We assume a never-ending stream of arrivals which follows a Poisson process, and exponentially-distributed (server-dependent) service times. We assume that steady-state conditions have been reached, and in particular that the arrival rate is smaller than the total service rate.

Selfish customers choose a server to minimize their own mean waiting times, ignoring the social consequences of their actions. As a point of comparison, we also consider the outcome that directs customers to servers to minimize the average mean waiting time. Naturally, the outcome reached by selfish customers—a (Nash) equilibrium—need not coincide with

*February 2006; revised July 2006 and September 2006.

†Department of Statistics, The Hebrew University of Jerusalem, 91905 Jerusalem, Israel. The research leading to this paper was conducted in the School of Mathematics, Statistics and Computer Science at Victoria University of Wellington, New Zealand and at the School of Economics and Political Science at the University of Sydney, Sydney, NSW, Australia. Supported in part by The Israel Science Foundation Grant no. 237/02. E-mail: `haviv@mscc.huji.ac.il`.

‡Department of Computer Science, Stanford University, 462 Gates Building, 353 Serra Mall, Stanford, CA 94305. Supported in part by ONR grant N00014-04-1-0725, DARPA grant W911NF-05-1-0224, and an NSF CAREER Award. Email: `tim@cs.stanford.edu`.

the socially optimal one. We measure this inefficiency via the *price of anarchy (PoA)* [7, 8], defined as the ratio between the average mean waiting times of the (unique) equilibrium and of the socially optimal outcome. The PoA is by definition at least 1. A PoA close to 1 indicates that the equilibrium is approximately socially optimal, and thus the consequences of selfish behavior are relatively benign.

Our goal is to understand when the PoA in a multi-server exponential service station is reasonably small. On the negative side, previous work by Friedman [5] shows that if the number of servers can be arbitrarily large, then the PoA can also be arbitrarily large. On the positive side, the PoA is 1 when servers are homogeneous; this is implicit in [3]. Also, Roughgarden [9] showed that the PoA is upper bounded by a small constant in stations in which the equilibrium outcome leaves a constant fraction of the capacity of each server unused.

Our main result is that the PoA is small in exponential service stations with a bounded number of servers, even when service times are heterogeneous and an arbitrarily large fraction of the server capacities are used. Specifically, we show that the PoA in such stations is upper bounded by the number of servers used in the socially optimal outcome. We also give a family of examples achieving a matching lower bound.

We conclude with brief comments on additional related work. Our model can be viewed as a special case of the traffic routing model that is often referred to as "traffic equilibria" (see e.g. [11]) or "selfish routing" (see e.g. [10]). While the PoA in this general model has been extensively studied, the only results on the PoA in exponential multi-server stations are those discussed above. Also, although different terminology is used, Koutsoupias and Papadimitriou [7] study a multi-server system. However, they consider a finite number of players, each controlling a non-negligible fraction of the arriving traffic. They also define the social cost to be the maximum individual cost, whereas we consider the average player cost. For surveys of more recent work in this model, see [2, 4].

2 The Model and Preliminaries

Suppose there are n exponential servers who man a single service station. Let μ_i be the service rate of server i, $1 \leq i \leq n$. Assume without loss of generality that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n > 0$. For convenience assume that $\mu_{n+1} = 0$. Denote $\sum_{j=1}^{i} \mu_j$ by $\mu_{(i)}$. There is a single Poisson arrival process with rate λ. To guarantee stability, we assume that $\lambda < \mu_{(n)}$.

2.1 Equilibrium

Suppose customers wish to minimize their mean waiting time and select a line to join accordingly. They do not observe the actual queue lengths at the time of deciding which line to join. Customers are allowed to randomize regarding their choice of servers. Thus, each line corresponds to a pure strategy and every distribution over the lines is a mixed (randomized) strategy.

We adopt the solution concept of a symmetric Nash equilibrium under steady-state conditions. Specifically, suppose all customers decide to join line i with probability p_i, $1 \leq i \leq n$. Under the resulting steady-state conditions, the arrival rate to a server i is Poisson with rate λp_i and the mean waiting time is $1/(\mu_i - \lambda p_i)$. The mixed strategy $\{p_i\}_{i=1}^n$ is a symmetric Nash equilibrium if when all customers follow this strategy, no customer can decrease its mean waiting time by unilaterally deviating to a different strategy.

In a symmetric Nash equilibrium, the mean waiting times in all utilized servers are identical and their common value is smaller than or equal to the mean service time of each idle server. Bell and Stidham [1] (see also [6, P.63]) show that at equilibrium, the only servers that are used are the first i_e servers, where

$$i_e = \min\left\{ i \geq 1 \,:\, \mu_{i+1} \leq \frac{\mu_{(i)} - \lambda}{i} \right\} \,. \tag{1}$$

Moreover, the Poisson arrival rate to server i, denoted here by λ_i^e, is $\lambda_i^e = \mu_i - (\mu_{(i_e)} - \lambda)i_e$. The common mean waiting time in queues 1 through i_e is

$$W^e = \frac{i_e}{\mu_{(i_e)} - \lambda} \tag{2}$$

and the total social cost per unit of time equals

$$L^e = \lambda \frac{i_e}{\mu_{(i_e)} - \lambda} \quad . \tag{3}$$

The social cost under equilibrium behavior is therefore a function only of the number of and the sum of the service rates of the used servers.

Equation (1) implies that i_e is non-decreasing in λ. We also require the answer to the inverse question: for a given integer $m \leq n$ what is the infimum value for λ such that the first m servers are used at equilibrium? Denote this value by $\lambda_e^{\min}(m)$. Equation (1) implies that

$$\lambda_e^{\min}(m) = \mu_{(m-1)} - \mu_m(m-1), \quad 1 \leq m \leq n. \tag{4}$$

To see this, note that equation (1) gives

$$\lambda_e^{\min}(m) = \min\{\lambda \geq 0 \,:\, \mu_{i+1} \geq \frac{\mu_{(i)} - \lambda}{i}, 1 \leq i \leq m-1\}.$$

Since μ_i is nondecreasing in i, this inequality holds for all $i \in \{1, 2, \ldots, m-1\}$ if and only if it holds for $i = m-1$. This implies (4). As expected, $\lambda_e^{\min}(m)$ is nondecreasing in m.

Remark 2.1 With one server with service rate $\mu_{(i_e)}$, the common mean waiting time would be $1/(\mu_{(i_e)} - \lambda)$, a $1/i_e$ fraction of (2). This increase in the mean waiting time is a function only of the number i_e of used servers, and not a function of how the $\mu_{(i_e)}$ total service rate is partitioned among these servers. Nonetheless, as we will see below, the PoA depends on the partition in the total capacity.

This phenomenon is also reminiscent of the following fact about queue splitting. Suppose instead of one server with an arrival rate of λ and a service rate of μ, there are n servers such that the arrival rate and service rate at each server $i \in \{1, 2, \ldots, n\}$ are, respectively, $p_i\lambda$ and $p_i\mu$ for some positive probabilities $\{p_i\}_{i=1}^n$ with $\sum_{i=1}^n p_i = 1$. Let $\rho = \lambda/\mu$. Then, for any choice of p_i's, all servers are utilized at the level of ρ and the mean queue length in *each* is $\rho/(1-\rho)$, as in the single queue. This implies that the total mean number in the system after the split is n times larger than the original mean and likewise, by Little's Law, the mean waiting time is n times larger than the corresponding mean before the split. Note also that the social cost after such a split coincides with the social cost under equilibrium (3) if all n servers are used in the equilibrium outcome.

2.2 Social Optimization

Every symmetric mixed strategy leads to a set of arrival rates $\{\lambda_i\}_{i=1}^n$ to the servers. The search for a socially optimal symmetric strategy is therefore equivalent to solving the following mathematical program:

$$\min_{\lambda_1,\ldots,\lambda_n} \sum_{j=1}^{n} \frac{\lambda_j}{\mu_j - \lambda_j}$$

subject to the constraints $\sum_{j=1}^n \lambda_j = \lambda$ and $0 \le \lambda_j < \mu_j$ for all $j \in \{1, 2, \ldots, n\}$. Bell and Stidham [1] showed that, in an optimal solution, the first i_s servers are used, where

$$i_s = \min\left\{ i \ge 1 \,:\, \mu_{i+1} \le \frac{(\mu_{(i)} - \lambda)^2}{(\sum_{j=1}^{i} \sqrt{\mu_j})^2} \right\} . \tag{5}$$

Moreover, the socially optimal arrival rate λ_i^s to each server $i \le i_s$ is

$$\lambda_i^s = \mu_i - \frac{\sqrt{\mu_i}}{\beta}, \tag{6}$$

where

$$\beta = \frac{\sum_{j=1}^{i_s} \sqrt{\mu_j}}{\mu_{(i_s)} - \lambda}.$$

The corresponding socially optimal policy is for each customer to join server i, $1 \le i \le i_s$, with probability λ_i^s/λ, and to join the other servers with probability zero. Finally, the mean waiting time in the utilized server i equals

$$W_i^s = \beta/\sqrt{\mu_i}, \quad 1 \le i \le i_s. \tag{7}$$

In particular, the mean waiting time is smaller in the fast servers who nevertheless receive more traffic than the slow servers. This should be compared with the equilibrium criterion in which mean waiting times in all utilized servers are identical. Note also that the utilization levels $\lambda_i^s/\mu_i = 1 - 1/(\beta\sqrt{\mu_i})$, for $1 \le i \le i_s$, are higher at the faster servers.

The social cost (mean number in the system) under the socially optimal behavior, denoted by L^s, is $L^s = \sum_{i=1}^{i_s} \lambda_i^s W_i^s$ where λ_i^s and W_i^s are defined as in (6) and (7), respectively. This, after minimal algebra, equals

$$L^s = \frac{(\sum_{i=1}^{i_s} \sqrt{\mu_i})^2}{\mu_{(i_s)} - \lambda} - i_s \ . \tag{8}$$

Note that the socially optimal cost is a function only of the number of used servers, and of the sum and the sum of the square roots of the service rates of these servers.

Equation (5) implies that i_s is non-decreasing in λ. Let $\lambda_s^{\min}(m)$ denote the infimum value of λ such that the first m servers are utilized in the socially optimal outcome. We then obtain

$$\lambda_s^{\min}(m) = \mu_{(m-1)} - \sqrt{\mu_m} \sum_{j=1}^{m-1} \sqrt{\mu_j}, \quad 1 \leq m \leq n \tag{9}$$

via the same argument used to establish (4). Comparing (4) and (9), we see that $\lambda_e^{\min}(m) \geq \lambda_s^{\min}(m)$ for all $m \in \{1, 2, \ldots, n\}$. It follows that $i_e \leq i_s$.

Remark 2.2 Starting from socially optimal behavior as a point of departure, customers will migrate from queues with high mean waiting times to queues with smaller mean waiting times—from servers with high indices to those with lower indices. In particular, this may cause some servers that are used in the socially optimal outcome to become empty in the equilibrium. There are thus two related reasons that the PoA can be large: (1) over-congestion at faster servers; and (2) servers which are slower but socially useful may stay idle with selfish behavior.

3 Main Results

Our main results are that the price of anarchy in every multi-server exponential service station is upper bounded by the number of servers, and that no better upper bound is possible. The following lemma, which considers the special case of service stations in which the socially optimal and equilibrium outcomes use the same set of servers, is crucial for our upper bound proof.

Lemma 3.1 *In a multi-server exponential service station in which both the socially optimal and the equilibrium outcomes use the same number m of servers, the price of anarchy is at most m.*

Proof: By definition, the PoA is L^e/L^s, where L^e and L^s are defined as in (3) and (8), respectively. Assume that $i_e = i_s = m$—i.e., that the socially optimal and equilibrium outcomes both use precisely the first m servers. We can assume that $m > 1$, as otherwise

the PoA is clearly 1. The PoA is then

$$\begin{aligned}\frac{L^e}{L^s} &= \lambda i_e \frac{\mu_{(i_s)} - \lambda}{\mu_{(i_e)} - \lambda} \left[\left(\sum_{i=1}^{i_s} \sqrt{\mu_i} \right)^2 - i_s(\mu_{(i_s)} - \lambda) \right]^{-1} \\ &= \frac{\lambda m}{(\sum_{i=1}^{m} \sqrt{\mu_i})^2 - m(\mu_{(m)} - \lambda)}. \end{aligned} \tag{10}$$

First, note that the socially optimal and equilibrium outcomes, along with the ratio of their social costs (3) and (8), remain unchanged if all service and arrival rates are scaled by a common positive value. We can therefore assume, without loss of generality, that $\mu_{(m)} = 1$. Second, the expression (10) is strictly decreasing in λ, and $\lambda \geq \lambda_e^{\min}(m)$ by definition. Combining our expression (4) for $\lambda_e^{\min}(m)$ with (10) then gives

$$\frac{L^e}{L^s} \leq \frac{m(1 - m\mu_m)}{(\sum_{i=1}^{m} \sqrt{\mu_i})^2 - m^2\mu_m}. \tag{11}$$

Our next goal is to maximize the right-hand side of (11) where the decision variables μ_i, $1 \leq i \leq m$, are constrained so that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m \geq 0$ and $\sum_{i=1}^{m} \mu_i = 1$.

We claim that the optimal solution is of the form $\mu_i = \mu_m$ for $2 \leq i \leq m$ and $\mu_1 = 1 - (m-1)\mu_m$. To see this, fix a value for μ_m; since $\mu_1 \geq \cdots \geq \mu_m \geq 0$ and $\sum_{i=1}^{m} \mu_i = 1$, we have $\mu_m \leq 1/m$. Now maximize the right-hand side of (11) with respect to the other variables $\mu_1, \mu_2, \ldots, \mu_{m-1}$. This optimization problem is equivalent to minimizing $\sum_{i=1}^{m-1} \sqrt{\mu_i}$ subject to $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$ and $\sum_{i=1}^{m-1} \mu_i = 1 - \mu_m$. Since this is minimizing a strictly concave function over a convex set, the optimal solution is on the boundary, which in this case implies that $\mu_2 = \mu_3 = \cdots = \mu_{m-1} = \mu_m$ and $\mu_1 = 1 - (m-1)\mu_m$.

Maximizing the right-hand side of (11) thus reduces to maximizing the following single variable objective:

$$\frac{m(1 - m\mu_m)}{1 - (3m-2)\mu_m + 2(m-1)\sqrt{\mu_m(1 - (m-1)\mu_m)}}.$$

Finally, trivial algebra shows that

$$1 - mx \leq 1 - (3m-2)x + 2(m-1)\sqrt{x(1 - (m-1)x)}$$

for every $x \in [0, 1/m]$, which proves that the PoA (10) is at most m. ■

We next show that the bound m derived in Lemma 3.1 is applicable also when more servers are used in the socially optimal outcome than in the equilibrium one. Thus, in a multi-server service station with n servers, the price of anarchy is at most n.

Theorem 3.1 *In every multi-server exponential service station, the price of anarchy is at most the number of servers used in the socially optimal outcome.*

Proof: Recall from Subsection 2.2 that the number of servers i_s used in the socially optimal outcome can only be greater than the number i_e used by the equilibrium outcome. Suppose that $i_e < i_s = m$. Define augmented service rates μ'_j as follows: $\mu'_j = \mu_j$ for $j \notin \{i_e + 1, \ldots, m\}$, and

$$\mu'_j = \frac{\mu_{(i_e)} - \lambda}{i_e}, \quad i_e + 1 \leq j \leq m.$$

The definition of these augmented service rates ensures that the equilibrium outcome is unaffected: the rates of the unused servers are increased to the largest value at which customers still do not switch to a server with an index $i_e + 1$ or higher. Thus, L_e (the social cost under equilibrium) is not changed but L_s (the social cost under optimal behavior) has been reduced due to the improvement in the service capacities. In particular, these capacity augmentations increase the PoA. We can therefore prove the theorem by showing that the PoA is at most m in the augmented service station.

Now consider slightly increasing the arrival rate λ to the augmented service station. At equilibrium, the added flow of jobs is evenly spread among servers $1, \ldots, i_s$, making the number of used servers equal to the number m of such servers in the optimal outcome. Of course, this infinitesimal change in λ has only infinitesimal effects on both L_e and L_s—formally, L_e and L_s are both continuous functions of λ. We can therefore conclude that the PoA of the augmented multi-server service station is bounded above by the PoA of a service station in which the socially optimal and equilibrium outcomes both use the same m servers. Lemma 3.1 then implies that the PoA of the augmented service station is at most m, which completes the proof of the theorem. ■

We conclude with an example that demonstrates that our upper bound of n in Theorem 3.1 is the best possible. In other words, we show that for every $n \geq 1$, there are multi-server exponential service stations with n servers and PoA arbitrarily close to n.

Example 3.1 Fix $n \geq 1$, an arrival rate $\lambda > n$, and a constant $\epsilon > 0$, let $\mu_1 = \lambda + 1 + \epsilon$ and $\mu_i = 1 + \epsilon$ for $i \in \{2, \ldots, n\}$. In the equilibrium outcome, all customers choose the first server, and

$$L^e = \frac{\lambda}{1 + \epsilon}\ . \tag{12}$$

Now consider assigning customers to servers so that $\lambda_1 = \lambda - n + 1$ and $\lambda_i = 1$ for every $i \in \{2, \ldots, n\}$. The mean number in the n queues in this outcome upper bounds L^s from above and equals

$$\frac{\lambda - n + 1}{n + \epsilon} + (n - 1)\frac{1}{\epsilon}\ . \tag{13}$$

Now divide (12) by (13) and take λ to infinity. For each fixed $n \geq 1$ and $\epsilon > 0$, the limit of this ratio is $(n + \epsilon)/(1 + \epsilon)$. This expression can be made arbitrarily close to n by choosing the constant $\epsilon > 0$ to be sufficiently small. Hence, there are n-server service stations with PoA arbitrarily close to n, and the upper bound in Theorem 3.1 is optimal.

Acknowledgment: Thanks are due to Yoav Kerner for some helpful comments.

References

[1] C. H. Bell and S. Stidham. Individual versus social optimization in the allocation of customers to alternative servers. *Management Science*, 29:831–839, 1983.

[2] A. Czumaj. Selfish routing on the Internet. In J. Leung, editor, *Handbook of Scheduling: Algorithms, Models, and Performance Analysis*, chapter 42. CRC Press, 2004.

[3] S. C. Dafermos and F. T. Sparrow. The traffic assignment problem for a general network. *Journal of Research of the National Bureau of Standards, Series B*, 73(2):91–118, 1969.

[4] R. Feldmann, M. Gairing, T. Lücking, B. Monien, and M. Rode. Selfish routing in non-cooperative networks: A survey. In *Proceedings of the 28th International Symposium on Mathematical Foundations of Computer Science (MFCS)*, volume 2747 of *Lecture Notes in Computer Science*, pages 21–45, 2003.

[5] E. J. Friedman. Genericity and congestion control in selfish routing. In *Proceedings of the 43rd Annual IEEE Conference on Decision and Control (CDC)*, pages 4667–4672, 2004.

[6] R. Hassin and M. Haviv. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer's International Series, Boston, MA, 2003.

[7] E. Koutsoupias and C. H. Papadimitriou. Worst-case equilibria. In *Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science (STACS)*, volume 1563 of *Lecture Notes in Computer Science*, pages 404–413, 1999.

[8] C. H. Papadimitriou. Algorithms, games, and the Internet. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing (STOC)*, pages 749–753, 2001.

[9] T. Roughgarden. The price of anarchy is independent of the network topology. *Journal of Computer and System Sciences*, 67(2):341–364, 2003.

[10] T. Roughgarden. *Selfish Routing and the Price of Anarchy*. MIT Press, 2005.

[11] Y. Sheffi. *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Prentice-Hall, 1985.

[12] J. G. Wardrop. Some theoretical aspects of road traffic research. In *Proceedings of the Institute of Civil Engineers, Pt. II*, volume 1, pages 325–378, 1952.