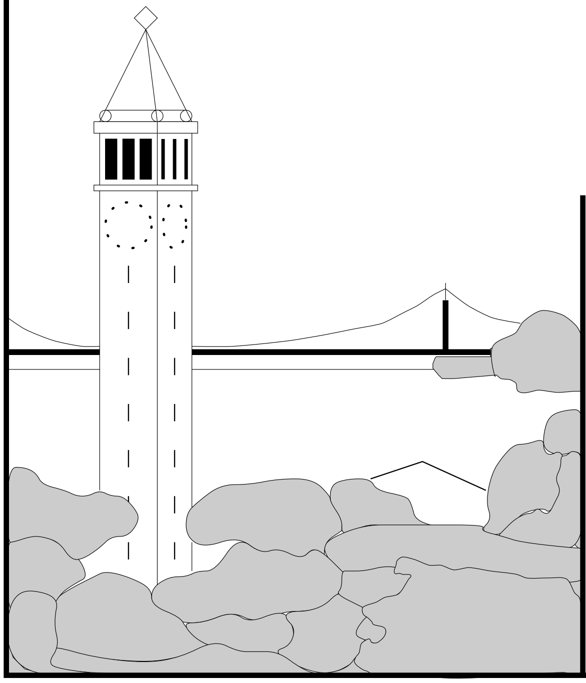
Constant Beats Memoryless for Service Times in a Markovian Queueing Network

Mor Harchol-Balter and David Wolfe



Report No. UCB/CSD-94-820

July 1994

Computer Science Division (EECS)
University of California
Berkeley, California 94720

Constant Beats Memoryless for Service Times in a Markovian Queueing Network

Mor Harchol-Balter* David Wolfe †

Abstract

We prove that a Markovian Open Queueing Network in which the service times are constant has lower average packet delay than the same network in which the service times are exponential (with the same mean). The proof is elementary, generalizing a similar result of Stamoulis and Tsitsiklis by removing the requirement that the network be layered.

1 Motivation

Many real-world packet-routing network algorithms in which packets are routed to random destinations can be modeled by Markovian[1] Jackson Queueing Networks (M.J.Q.N.), except for the fact that the real-world networks require the service times at the servers to all be equal and constant [HBB94], [ST91]. In this paper we show that the average delay for a M.J.Q.N. with constant service times is upper-bounded by the average delay for the corresponding traditional M.J.Q.N. (with exponential service times). [ST91] proved this result for layered networks. Our proof *exactly* parallels the [ST91] proof, except that whereas their proof used induction on the layers of the network, we induct on time, thereby obviating the need for a layered network.

2 Proof

Theorem 1 *The average delay for a M.J.Q.N. where all service times are exactly one is smaller than the average delay for the corresponding M.J.Q.N. where all service times are independent, exponentially distributed random variables with mean one.*

*Computer Science Division, UC Berkeley, CA 94720. Supported by National Physical Science Consortium (NPSC) Fellowship. Also supported by NSF grant number CCR-9201092.
Email: harchol@cs.berkeley.edu

†Computer Science Division, UC Berkeley, CA 94720. Email: wolfe@cs.berkeley.edu

[1] In a Markovian network, after each packet is served, the packet's next server (or whether it leaves the system altogether) depends only on its previous server and not on past history.

Proof: Recall that the average delay for a M.J.Q.N. whose servers have exponentially distributed service times is the same as that of the corresponding M.J.Q.N. whose servers are deterministic processor sharing, since both models are product form networks [Wal89, p. 94]. Therefore it is sufficient to prove that the average delay for a M.J.Q.N. with constant service times (hereafter denoted by FIFO network) is smaller than the average delay for the corresponding M.J.Q.N. with deterministic processor sharing servers (hereafter denoted by PS network).

Claim 1 *First, note that if the sequence of arrivals to a (single server) FIFO queue is no later than the arrivals to a PS queue, then the i^{th} departure from the FIFO queue occurs no later than the i^{th} departure from the PS queue.*

Proof: In both queues, each packet must wait for all packets with earlier arrivals to depart, but only in the PS queue must a packet also wait while later arrivals get service. ∎

To generalize the statement from the single server to the network, we'll use a coupling argument. Consider the behavior of the two networks when coupled to run on the same *sample point* consisting of:

1. the sequence of *outside* inter-arrival times at each server, and
2. the choices for where the j^{th} packet served at each server proceeds next.

Note the above quantities are all independent for a Markovian network with poisson arrivals. Also, the j^{th} packet to complete at a particular server in the two networks may not be the same packet.

Claim 2 *For a given sample point,*

1. *The i^{th} packet to arrive at any server of the FIFO network arrives no later than the i^{th} arrival at the corresponding server of the PS network.*
2. *the j^{th} service completion at any server of the FIFO network occurs no later than the j^{th} service completion at the corresponding server of the PS network.*

Proof: Assume the claim is true at time t. We'll now show it's true at time $t' > t$, where t' is the time of the next outside arrival or service completion. (Note, arrivals from inside the network occur at service completions). *If t' is an outside arrival*, that arrival gets queued at the same server in both networks, at the same time, so the above claim is still true. *Suppose t' is a service completion at server q.* During time $[0, t]$, the arrivals into FIFO server q were no later than the corresponding arrivals at PS server q. Since no arrivals occur during $[t, t']$, the previous remark holds for the interval $[0, t']$. Therefore, by Claim 1, during time $[0, t']$, every service completion at FIFO server q happened no later than the corresponding service completion at PS server q. This includes this current service completion. ■

By Claim 2, it follows that the i^{th} departure from the FIFO *network* occurs no later than the i^{th} departure from the PS network. Therefore, the expected delay is greater for the PS network. ■

3 Future Research

The natural open question is whether Theorem 1 is also true for all (non-Markovian) J.Q.N. networks. [HBW94] give strong evidence to the contrary, but there still may be other large classes of networks for which the result is true.

References

[HBB94] Mor Harchol-Balter and Paul E. Black. Queueing analysis of oblivious packet-routing algorithms. In *Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 583–592, January 1994.

[HBW94] Mor Harchol-Balter and David Wolfe. Greater variance does not necessarily imply greater delay. Technical Report UCB//CSD-94-821, University of California, Berkeley, 1994. Also submitted to *Journal of Applied Probability*.

[ST91] George D. Stamoulis and John N. Tsitsiklis. The efficiency of greedy routing in hypercubes and butterflies. *Journal of the ACM*, 1991.

[Wal89] Jean Walrand. *Introduction to Queueing Networks*. Prentice Hall, New Jersey, 1989.