Optimal Queue-Size Scaling in Switched Networks

Devavrat Shah[1], Neil Walton[2], and Yuan Zhong[1]

[1]Massachusetts Institute of Technology
[2]University of Amsterdam

Abstract

We consider a switched (queueing) network in which there are constraints on which queues may be served simultaneously; such networks have been used to effectively model input-queued switches and wireless networks. The scheduling policy for such a network specifies which queues to serve at any point in time, based on the current state or past history of the system. In the main result of this paper, we provide a new class of online scheduling policies that achieve optimal average queue-size scaling for a class of switched networks including input-queued switches. In particular, it establishes the validity of a conjecture (documented in [3]) about optimal queue-size scaling for input-queued switches.

1 Introduction.

Switched networks are fairly general constrained queueing networks which have been used to faithfully model a variety of networked processing systems. They have been used to effectively model input-queued switches that reside in Internet routers, the packet-level behavior of medium access in wireless networks, as well as operations in a data center. The key operational challenge in switched networks is deciding dynamically which queues to schedule simultaneously, subject to system constraints.

In this paper, we consider *online* scheduling policies, that is, policies that only utilize historical information (i.e., past arrivals and scheduling decisions). The performance objective of interest is the long-run average total queue size in the network. The questions that we wish to answer are: (a) what is the minimal value of the performance objective among the class of online scheduling policies, and (b) how does it depend on the network structure, $\mathcal{S}$, as well as the effective load.

As the main result of this paper, we propose a new online scheduling policy for any single-hop switched network. This policy effectively emulates an insensitive bandwidth sharing network with a product-form stationary distribution with each component of this product-form behaving like an M/M/1 queue. This crisp description of stationary distribution allows us to obtain precise bounds on the average queue sizes under this policy. This leads to establishing, as a corollary of our result, the validity of the conjecture stated in [3] for input-queued switches. In general, it provides explicit bounds on the average total queue size for any single-hop switched network. Furthermore, due to the explicit bound on the stationary distribution of queue sizes under our policy, we are able to establish a form of large-deviations optimality of the policy for *any* single-hop switched network.

We note that the validity of the conjecture in [3] for input-queued switches, stating that optimal average total queue size scales as $\sqrt{N}/(1-\rho)$, is a significant improvement over the best known bounds of $O(N/(1-\rho))$ (due to the moment bounds of [1] for the maximum weight policy) or $O\left(\frac{\sqrt{N}\log N}{(1-\rho)^2}\right)$ (obtained by using a batching policy [2]).

The related paper can be found at [4].

References

[1] S. Meyn and R. Tweedie. *Markov chains and stochastic stability*. Springer New York, 1993.

[2] M. Neely, E. Modiano, and Y. Cheng. Logarithmic delay for n× n packet switches under the crossbar constraint. *IEEE/ACM Transactions on Networking (TON)*, 15(3):657–668, 2007.

[3] D. Shah, J. N. Tsitsiklis, and Y. Zhong. Optimal scaling of average queue sizes in an input-queued switch: an open problem. *Queueing Systems*, 68(3-4):375–384, 2011.

[4] D. Shah, N. Walton, and Y. Zhong. Optimal queue-size scaling in switched networks. *http://arxiv.org/pdf/1110.4697v1.pdf*.