Improved Queue-Size Scaling for Input-Queued Switches via Graph Factorization

Jiaming Xu

Duke University, Fuqua School of Business
jiaming.xu868@duke.edu

Yuan Zhong

University of Chicago, Booth School of Business
yuan.zhong@chicagobooth.edu

ABSTRACT

This paper studies the scaling of the expected total queue size in an $n \times n$ input-queued switch, as a function of both the load ρ and the system scale n . We provide a new class of scheduling policies under which the expected total queue size scales as $O\left(n(1-\rho)^{-4/3} \log\left(\max\left\{\frac{1}{1-\rho}, n\right\}\right)\right)$, over all n and $\rho < 1$, when the arrival rates are uniform. This improves over the previously best-known scalings in two regimes: $O\left(n^{1.5}(1-\rho)^{-1} \log \frac{1}{1-\rho}\right)$ when $\Omega(n^{-1.5}) \leq 1-\rho \leq O(n^{-1})$ and $O\left(\frac{n \log n}{(1-\rho)^2}\right)$ when $1-\rho \geq \Omega(n^{-1})$. A key ingredient in our method is a tight characterization of the largest k -factor of a random bipartite multigraph, which may be of independent interest.¹

KEYWORDS

input-queued switch; queue-size scaling; graph factorization

ACM Reference Format:

Jiaming Xu and Yuan Zhong. 2019. Improved Queue-Size Scaling for Input-Queued Switches via Graph Factorization. In *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '19 Abstracts)*, June 24–28, 2019, Phoenix, AZ, USA. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3309697.3331497>

1 INTRODUCTION

An $n \times n$ input-queued switch is a discrete-time queueing network that consists of n^2 queues, which are arranged in the form of an $n \times n$ matrix. Packets arrive to these n^2 queues exogenously according to independent Bernoulli processes. In each time slot, packets are processed according to *schedules*, subject to the following constraints: at most one queue can be served from each row and each column, and when a queue is chosen to be served, at most one packet can be processed. After the packets are processed, they immediately depart the system. The main goal is to design a scheduling policy that decides which queues to serve in each time slot, in order to optimize certain performance measures.

The input-queued switch plays a pivotal role in the study of data packets scheduling in an internet router [9] and more recently in a data center network [2, 4]. It also serves as a prominent example of “stochastic processing networks” (SPN) [5, 6], which is a canonical

model of dynamic resource allocation problems that arise in a wide range of applications. The study of the switch has paved the way towards a better understanding of more general SPNs; indeed, insights generated from policy design and performance analysis in a switch often extend to broader classes of SPNs.

Even though basic performance measures such as throughput and stability, are relatively well understood for switches (and for general SPNs), there is a wide gap in our understanding of more refined performance measures, such as moments of queue sizes. In this paper, we reduce this gap by designing new scheduling policies that achieve tighter bounds on the expected total queue size in switches. Our study is motivated by [13], which stated the conjecture that the expected total queue size scales as $\Theta\left(\frac{n}{1-\rho}\right)$ in an $n \times n$ switch, where $\rho \in (0, 1)$ is the load factor and equals the largest total arrival rate to each row/column of the switch.

It can be easily seen that $\Omega\left(\frac{n}{1-\rho}\right)$ is a lower bound on the expected total queue size, over all n and ρ (see e.g., [13]). However, matching upper bounds have been only established for restricted parameter regimes. [10] proposed a batching policy under which the expected total queue size is upper bounded by $O\left(\frac{n \log n}{(1-\rho)^2}\right)$, so that when $\rho \in (0, 1)$ is treated as a constant, this upper bound matches the lower bound up to a logarithmic factor in n . [15] proposed a policy under which the switch emulates a product-form queueing network, and established an upper bound of $O\left(\frac{n}{1-\rho} + n^3\right)$, which matches the lower bound up to a constant factor when $1-\rho = O(n^{-2})$. More recently, a significant work [8] shows that the celebrated Max-Weight policy achieves an upper bound of $O\left(\frac{n}{1-\rho} + \frac{n^{5-1/r}}{(1-\rho)^{1/r}}\right)$ for any integer-valued $r \geq 2$, for systems with uniform arrival rates, which matches the lower bound up to a constant factor when $1-\rho = O(n^{-a})$ for any $a > 4$. Let us also note that using a simple Lyapunov function argument, it can be shown that the Max-Weight policy achieves an upper bound of $O\left(\frac{n^2}{1-\rho}\right)$ (see e.g., [9, 12, 13, 17]).

Observe that when $1-\rho = \Theta(n^{-1})$, the conjectured scaling is $\Theta(n^2)$, whereas all of the aforementioned upper bounds are at least a multiplicative factor $O(n)$ away. Motivated by this observation, [14] considered the regime where $1-\rho \leq O(n^{-1})$, and proposed a scheduling policy that gives an upper bound of $O\left(n^{1.5}(1-\rho)^{-1} \log \frac{1}{1-\rho}\right)$, for systems with uniform arrival rates. When $1-\rho = \Theta(n^{-1})$, this upper bound becomes $O(n^{2.5} \log n)$, a multiplicative factor of $O(n^{0.5} \log n)$ away from the conjectured scaling. While this was a significant improvement, whether this gap can be further reduced remains elusive.

¹Extended abstract. The full paper can be found at [18].

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGMETRICS '19 Abstracts, June 24–28, 2019, Phoenix, AZ, USA

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6678-6/19/06.

<https://doi.org/10.1145/3309697.3331497>

The main contribution of this paper is a new class of policies that achieves an upper bound of $O\left(n(1-\rho)^{-4/3} \log(\max\{n, 1/(1-\rho)\})\right)$ uniformly over all n and $\rho < 1$, when the arrival rates are equal. When $\Omega(n^{-1.5}) \leq 1-\rho \leq O(n^{-1})$, this upper bound improves over the previously best-known scaling $O\left(n^{1.5}(1-\rho)^{-1} \log \frac{1}{1-\rho}\right)$ [14]. In particular, when $1-\rho = \Theta(n^{-1})$, our upper bound reduces to $O(n^{7/3} \log n)$, an improvement over the previously best-known scaling $O(n^{2.5} \log n)$. When $1-\rho \geq \Omega(n^{-1})$, our upper bound improves over the previously best-known scaling $O\left(\frac{n \log n}{(1-\rho)^2}\right)$ [10]. Finally, we would also like to point out that besides the upper bound $O\left(\frac{n \log n}{(1-\rho)^2}\right)$ in [10], ours is the only other upper bound that scales almost linearly in n , the system scale, for any fixed value of $\rho \in (0, 1)$.

At the heart of our new policy is an efficient scheduling mechanism which depletes the packets in the n^2 queues as much as possible without any waste of service opportunities. Mathematically, this can be formulated as an integer linear optimization problem: call a matrix $\mathbf{q} = (q_{ij})_{i,j=1}^n$ with non-negative integer entries a *queue matrix*; given $\mathbf{q}$, find the maximum integer k and a queue matrix $\mathbf{g} = (g_{ij})_{i,j=1}^n$ such that $\sum_i g_{ij} = \sum_j g_{ij} = k$ and $0 \leq g_{ij} \leq q_{ij}$. Such $\mathbf{g}$ can then be written as the sum of k maximal schedules, which can be used to deplete the n^2 queue sizes of $\mathbf{q}$. By representing the rows as left vertices of a bipartite (multi)graph, and the columns as right vertices, the queue size q_{ij} can be viewed as the number of edges between left vertex i and right vertex j . In this vein, this optimization problem is equivalent to finding the largest k -factor (spanning k -regular subgraph) in a bipartite (multi)graph G . Under the Bernoulli arrival assumption, the queue sizes q_{ij} are independently and binomially distributed for some common parameters m and p . We show that if $pmn \geq 152 \log n$, then G has a k -factor with $k = \lfloor pmn - \sqrt{304pmn \log n} \rfloor$ with probability $1 - n^{-16}$. This result is new, and essentially tight: for example, when G is an Erdős-Rényi bipartite simple graph, i.e., when $m = 1$, then the condition $np = \omega(\log n)$ implies that G has a k -factor with $k = np - \Theta(\sqrt{np \log n})$, which is asymptotically the best possible, because the minimum degree of G is asymptotically equal to $np - \Theta(\sqrt{np \log n})$ with high probability.

We point out that there is a vast literature on graph factors and factorization (see the survey [11] and the book [1] and the references therein). Most previous work focuses on deriving sufficient conditions for the existence of a k -factor for worst-case possible graphs G . For instance, Csaba [3] showed that for any bipartite simple graph G with n left and n right vertices, if the minimum degree $\delta n \geq n/2$, then G has a k -factor with $k = \lfloor n \frac{\delta + \sqrt{2\delta - 1}}{2} \rfloor$. Applying this result to an Erdős-Rényi bipartite simple graph, we can only conclude with high probability the existence of k -factor with $k \sim n \frac{p + \sqrt{2p - 1}}{2}$, which is suboptimal, when compared to our result.

Random graph factorization was studied earlier in Shamir and Upfal [16]. It was shown that for an Erdős-Rényi random graph G , if the average degree np satisfies $np - \log n - (k-1) \log \log n \rightarrow +\infty$ for a fixed $k \geq 1$, then G has a k -factor with probability converging to 1. However, this result only holds for a fixed k and does not provide the largest possible value of k in the dense graph regime

where $np = \omega(\log n)$. A related random capacitated transportation problem was studied in [7], where we observe a $n \times n$ capacity matrix $C = (c_{ij})$ with random c_{ij} and aim to find a matrix $X = (x_{ij})$ with the row sums and column sums as large as possible under the constraint that $0 \leq x_{ij} \leq c_{ij}$. Their general result allows for c_{ij} to be not independent and not identically distributed. In the special case where c_{ij} is Bernoulli with the success probability p , their result implies that with high probability an Erdős-Rényi random bipartite graph G has k -factor with $k \sim n^{1/3}$, which is much smaller than np for large n and hence highly suboptimal.

Finally, let us make two remarks regarding the proposed policy and its performance scaling. First, our policy and current analysis rely crucially on the assumption of uniform arrival rates, while some existing policies and results such as the standard batching policy [10] and the Max-Weight policy [9] apply to heterogeneous arrival rates. Second, our policy is required to know the arrival rates *a priori*. In contrast, some existing policies, such as Max-Weight policy, are based solely on the observed queue sizes.

ACKNOWLEDGMENTS

J. Xu was supported in part by the NSF Grants CCF-1850743, IIS-1838124, and CCF-1856424.

REFERENCES

- [1] AKIYAMA, J. AND KANO, M. (2011). *Factors and factorizations of graphs: Proof techniques in factor theory* vol. 2031. Springer.
- [2] M. Alizadeh, S. Yang, M. Sharif, S. Karri, N. McKeown, B. Prabhakar, S. Shenker. pFabric: Minimal near-optimal datacenter transport. *Proceedings of the ACM SIGCOMM*, pp. 435–446 (2013)
- [3] CSABA, B. (2007). Regular spanning subgraphs of bipartite graphs of high minimum degree. *The Electronic Journal of Combinatorics* 14, 21.
- [4] M. Chowdhury, Y. Zhong and I. Stoica. Efficient coflow scheduling with Varys. *Proceedings of the ACM SIGCOMM*, pp. 443–454 (2014)
- [5] J. M. Harrison. Brownian models of open processing networks: canonical representation of workload. *The Annals of Applied Probability* 10, 75–103 (2000). Also see [6]
- [6] J. M. Harrison. Correction to [5]. *The Annals of Applied Probability* 13, 390–393 (2003)
- [7] HASSIN, R. AND ZEMEL, E. (1988). Probabilistic analysis of the capacitated transportation problem. *Mathematics of Operations Research* 13, 80–89.
- [8] S. T. Maguluri and R. Srikant. Heavy traffic queue length behavior in a switch under the Maxweight algorithm. *Stochastic System*, 6(1), 211–250 (2016)
- [9] N. McKeown, V. Anantharam and J. Walrand. Achieving 100% throughput in an input-queued switch. *Proceedings of IEEE Infocom*, pp. 296–302 (1996)
- [10] M. Neely, E. Modiano and Y. S. Cheng. Logarithmic delay for $n \times n$ packet switches under the cross-bar constraint. *IEEE/ACM Transactions on Networking* 15(3) (2007)
- [11] PLUMMER, M. D. (2007). Graph factors and factorization: 1985–2003: a survey. *Discrete Mathematics* 307, 791–821.
- [12] D. Shah and M. Kopikare. Delay bounds for the approximate Maximum Weight matching algorithm for input queued switches. *Proceedings of IEEE Infocom* (2002)
- [13] 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)
- [14] D. Shah, J. N. Tsitsiklis and Y. Zhong. On queue-size scaling for input-queued switches. *Stochastic Systems* 6(1), 1–25 (2016)
- [15] D. Shah, N. Walton and Y. Zhong. Optimal queue-size scaling in switched networks. *The Annals of Applied Probability* 24(6), 2207–2245 (2014)
- [16] SHAMIR, E. AND UPFAL, E. (1981). On factors in random graphs. *Israel Journal of Mathematics* 39, 296–302.
- [17] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control* 37, 1936–1948 (1992)
- [18] J. Xu and Y. Zhong. Improved queue-size scaling for input-queued switches via graph factorization. *Preprint*, 2019. URL: <https://arxiv.org/pdf/1903.00398.pdf>