Heavy Traffic Analysis of Approximate Max-Weight Matching Algorithms for Input-Queued Switches

Yu Huang
IIIS, Tsinghua University
Beijing, China
y-huang20@mails.tsinghua.edu.cn

Longbo Huang*
IIIS, Tsinghua University
Beijing, China
longbohuang@tsinghua.edu.cn

ABSTRACT

In this paper, we propose a class of approximation algorithms for the max-weight matching (MWM) policy for input-queued switches, called expected 1-APRX. We establish the state space collapse (SSC) result for expected 1-APRX, and characterize its queue length behavior in the heavy-traffic limit. Our results indicate that expected 1-APRX can approximately approach the optimal queue length scaling in the heavy-traffic regime. We further propose an expected 1-APRX based policy, called MWM with adaptive update (MWM-AU), for reducing communication cost due to queue information update. Our simulation results demonstrate that the proposed policy can significantly reduce queue update overhead, while maintaining the delay performance comparable to that of MWM.

Keywords

Max-Weight, Heavy Traffic, Delay

1. INTRODUCTION

The $n \times n$ switch model has been intensively studied, owing to its importance in modeling and studying network scenarios including high-speed routers, data centers networks and wireless networks The system is assumed to be a discrete-time network consisting of n^2 queues with n input ports and n output ports. At every time, a scheduler chooses a matching between the inputs and outputs to transmit one packet from each input port to its destination. The objective is to find optimal scheduling policies so as to achieve good throughput and delay performance.

A well-known scheduling algorithm for the input-queued switch system is the max-weight matching algorithm (MWM) Despite many attractive properties, e.g., throughput optimal[3], optimal heavy-traffic queue length[4], MWM suffers from a high computational complexity of $\mathcal{O}(n^3)$ for an $n \times n$ switch [5], due to the need to compute a maximum weighted matching in every time slot. To reduce the implementation complexity, [6] introduced a general class of approximation algorithms for MWM, called 1-APRX, which computes a schedule with weight difference to MWM upper bounded by a sublinear function, and showed that 1-APRX achieves throughput optimality.

In this paper, we propose the expected 1-APRX policy, an extended version of the 1-APRX algorithm [6], and focus on its heavy-traffic analysis for switch systems. We first show that expected 1-APRX exhibits the similar state space collapse (SSC) as MWM [4] in a weak sense. Then, we utilize the SSC result and drift technique to characterize the heavy-traffic behavior of the queue lengths, which indicates that expected 1-APRX approximates the optimal scaling achieved by MWM in the heavy-traffic regime. In addition, we apply the expected 1-APRX policy to design communication efficient scheduling algorithms. In particular, we propose a MWM policy with adaptive updates (MWM-AU) for systems where queue information update can be costly and affect algorithm performance, e.g., datacenters [1] and wireless networks [2]. We prove the throughput optimality of MWM-AU and demonstrate through numerical experiments that it significantly reduces queue update information while maintaining the same level of throughput and delay performance as MWM. We carry out our analysis under the general saturated setting. Our analysis also well handles the additional challenges raised by the weight approximation in scheduling, which cannot be directly addressed with existing analysis.

2. EXPECTED 1-APRX

The expected 1-APRX can be defined as follows.

Definition 1. Denote the weight of a schedule obtained by a scheduling algorithm π at time t by $W_\pi(t)$, and denote the weight of a schedule under MWM for the same switch state by $W^*(t)$. A policy π is defined to be a expected 1-APRX algorithm, if the following condition always holds:

$$\mathbb{E}\left\{W_\pi(t)|\mathbf{Q}(t)\right\} \geq W^*(t) - f\left(W^*(t)\right) \tag{1}$$

where $\mathbf{Q}(t)$ is the current queue state vector, the expectation $\mathbb{E}$ is taken with respect to the system randomness and policy π, and $f(\cdot)$ is a sub-linear function, i.e., $\lim_{x\to\infty} \frac{f(x)}{x} = 0$.

The original version of 1-APRX in [6] only considered a deterministic difference.

3. HEAVY TRAFFIC RESULT

In this paper, we consider the switch where $n_1 \leq n$ input ports (rows) and $n_2 \leq n$ output ports (columns) are

[*] Longbo Huang is the corresponding author

saturated. Without loss of generality, we assume that input ports (rows) $1, 2, \ldots, n_1$ and output ports (columns) $1, 2, \ldots, n_2$ are saturated.

3.1 State-Space Collapse

Consider the following subspace:

$$\mathcal{S}_{n_1 n_2} \triangleq \Big\{ \mathbf{x} \in \mathbb{R}^{n^2} : \mathbf{x} = \sum_{i=1}^{n_1} w_i \mathbf{e}^{(i)} + \sum_{j=1}^{n_2} \widetilde{w}_j \widetilde{\mathbf{e}}^{(j)} \text{ where}$$

$$w_i \in \mathbb{R} \text{ for } 1 \le i \le n_1, \widetilde{w}_j \in \mathbb{R} \text{ for } 1 \le j \le n_2 \}$$

We first show that the system will collapse on the $\mathcal{S}_{n_1 n_2}$ in steady state by theorem 1, which means that as the parameter ϵ approaches zero, the mean arrival rate approaches the boundary of the capacity region and $\mathbf{Q}_{\perp\mathcal{S}}^{(\epsilon)}$ can be neglected in comparison to the dominant term $\mathbf{Q}_{\|\mathcal{S}}^{(\epsilon)}$.

Theorem 1. Consider a set of switch systems indexed by ϵ under an expected 1-APRX scheduling policy π. For any fixed $\beta > 0$, and each system with $0 < \epsilon \le \nu'_{\min}/4(1 + 2\beta)\|\boldsymbol{\eta}\|$, the steady state queue lengths vector satisfies:

$$\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp\mathcal{S}}^{(\epsilon)}\right\| - \beta\left\|\overline{\mathbf{Q}}_{\|\mathcal{S}}^{(\epsilon)}\right\|\right] \le M_\beta$$

where M_β is independent of ϵ.

The above result indicates that for any $\beta > 0$, as $\epsilon \to 0$

$$\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp\mathcal{S}}^{(\epsilon)}\right\|\right] / \mathbb{E}\left[\left\|\overline{\mathbf{Q}}^{(\epsilon)}\right\|\right] < \beta$$

Therefore, $\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp\mathcal{S}}^{(\epsilon)}\right\|\right]$ can be controlled by $\left\|\overline{\mathbf{Q}}^{(\epsilon)}\right\|$.

3.2 Upper Bound

Then, we can obtain an asymptotically tight upper bound for heavy-traffic queue length .

Theorem 2. Consider a set of switch systems under an expected 1-APRX scheduling policy , the steady state queue lengths vector satisfies

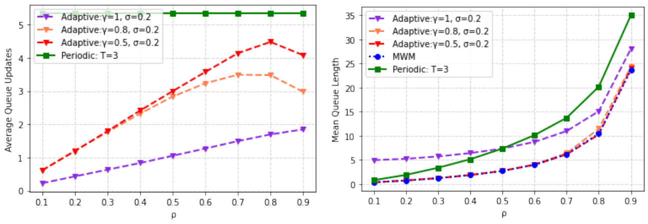
$$\epsilon\left(\mathbb{E}[\langle\overline{\mathbf{Q}}^{(\epsilon)}, \boldsymbol{\alpha}\rangle] - (\|\boldsymbol{\alpha}\| + 2n^2 \min\{n_1 + n_2, n\})\mathbb{E}\left[\left\|\overline{\mathbf{Q}}_{\perp\mathcal{S}}^{(\epsilon)}\right\|\right]\right)$$
$$\le \frac{1}{2}\left\langle(\boldsymbol{\sigma}^{(\epsilon)})^2, \boldsymbol{\zeta}\right\rangle + B(\epsilon)$$

for any fixed weight vector $\boldsymbol{\alpha} \in \mathbb{R}^{n^2}$ such that $\left\langle\boldsymbol{\alpha}, \mathbf{e}^{(i)}\right\rangle = n\left\langle\boldsymbol{\eta}, \mathbf{e}^{(i)}\right\rangle$ for $i \le n_1$ and $\left\langle\boldsymbol{\alpha}, \widetilde{\mathbf{e}}^{(j)}\right\rangle = n\left\langle\boldsymbol{\eta}, \widetilde{\mathbf{e}}^{(j)}\right\rangle$ for $j \le n_2$, where $\lim_{\epsilon\to 0} B(\epsilon) = 0$, and the vector $\boldsymbol{\zeta}$ only depends on n_1, n_2.

As stated in Section 3.1, $\mathbb{E}\left[\|\overline{\mathbf{Q}}_{\perp\mathcal{S}}\|\right] / \mathbb{E}\left[\|\overline{\mathbf{Q}}\|\right] \to 0$ when ϵ approaches 0. Thus, the upper bound for weighted queue length $\epsilon\mathbb{E}[\langle\overline{\mathbf{Q}}, \boldsymbol{\alpha}\rangle]$ is close to $\frac{1}{2}\left\langle\boldsymbol{\sigma}^2, \boldsymbol{\zeta}\right\rangle$ in the heavy-traffic limit.

4. MWM WITH ADAPTIVE UPDATE

Most existing works on switch operating under MWM consider instantaneous updates, where each queue updates its length to the scheduler at every time. However, in scenarios where the decision maker needs to collect queue information in a distributed system, e.g., in a datacenter network [1] or a wireless network [2], this can involve high communication overhead and impact system performance.[1] We propose an Adaptive Update scheme, which updates for each queue whenever the absolute value of the difference between the current length and the last update exceeds some threshold. The threshold has an adaptive form, $g\left(Q_{ij}(t)\right)$, where $g : \mathbb{R}^+ \cup \{0\} \mapsto \mathbb{R}^+ \cup \{0\}$ is a sub-linear, increasing and concave function, e.g. $\gamma x^{1-\sigma}$ with $0 < \sigma < 1$.



(a) Average Queue Updates

(b) Mean queue length

Figure 1: Communication frequency and delay performance versus traffic load under uniform traffic. $n = 4$

It can be regarded as the expected 1-APRX policy in the following sense:

$$W_a(t) \ge W^*(t) - 2ng\left(W^*(t)/n\right)$$

Simulations in Fig.1 illustrate that MWM with Adaptive Update can significantly reduce communication frequency without sacrificing delay performance.

5. REFERENCES

[1] M. Alizadeh, S. Yang, M. Sharif, S. Katti, N. McKeown, B. Prabhakar, and S. Shenker. pfabric: Minimal near-optimal datacenter transport. ACM Sigcomm, 47(8):1260–1267, 2013.

[2] A. Eryilmaz, R. Srikant, and J. R. Perkins. Stable scheduling policies for fading wireless channels. IEEE/ACM Transactions on Networking, 13(2):411–424, 2005.

[3] L. Georgiadis, M. J. Neely, and L. Tassiulas. Resource Allocation and Cross-Layer Control in Wireless Networks, volume 1. IEEE, 2006.

[4] S. T. Maguluri, S. K. Burle, and R. Srikant. Optimal heavy-traffic queue length scaling in an incompletely saturated switch. Queueing Systems, 88(3-4):279–309, 2018.

[5] M. J. Neely, E. Modiano, and C. E. Rohrs. Tradeoffs in delay guarantees and computation complexity for n× n packet switches. In Proc. of Conf. on Information Sciences and Systems (CISS. Citeseer, 2002.

[6] D. Shah and M. Kopikare. Delay bounds for approximate maximum weight matching algorithms for input queued switches. In Proceedings IEEE INFOCOM 2002, The 21st Annual Joint Conference of the IEEE Computer and Communications Societies, volume 2, pages 1024–1025, New York, NY, USA, 2002. IEEE.

[1] The switch setting can model general single-hop networks.