MaxWeight Scheduling: Asymptotic Behavior of Unscaled Queue-Differentials in Heavy Traffic

Rahul Singh
ECE Department
Texas A&M University
College Station, TX 77843
rsingh1@tamu.edu

Alexander Stolyar
ISE Department
Lehigh University
Bethlehem, PA 18015
stolyar@lehigh.edu

ABSTRACT

The model is a “generalized switch", serving multiple traffic flows in discrete time. The switch uses MaxWeight algorithm to make a service decision (scheduling choice) at each time step, which determines the probability distribution of the amount of service that will be provided. We are primarily motivated by the following question: in the heavy traffic regime, when the switch load approaches critical level, will the service processes provided to each flow remain “smooth" (i.e., without large gaps in service)? Addressing this question reduces to the analysis of the asymptotic behavior of the unscaled queue-differential process in heavy traffic. We prove that the stationary regime of this process converges to that of a positive recurrent Markov chain, whose structure we explicitly describe. This in turn implies asymptotic “smoothness" of the service processes.

Categories and Subject Descriptors

[**Network Services**]: Cloud Computing, Real-Time Systems; [**Probability and Statistics**]: Markov Processes, Queueing Theory, Stochastic Processes; [**Design and Analysis of Algorithms**]: Scheduling Algorithms, Real-Time Scheduling

General Terms

Algorithms, Performance, Theory

Keywords

Dynamic scheduling; MaxWeight algorithm; Heavy traffic asymptotic regime; Markov chain; Queue length differentials; Smooth service process

1. INTRODUCTION

Suppose we have a system in which several data traffic flows share a common transmission medium (or channel). Sharing means that in each time slot a scheduler chooses a transmission mode – the subset the flows to serve and corresponding transmission rates; the outcome of each transmission is random. Scheduler has two key objectives: (a) the time-average (successful) transmission rate of each flow i has to be at least some $\lambda_i > 0$; (b) the successful transmissions for each flow need to be spread out "smoothly" in time – without large time-gaps between successful transmissions. Such models arise, for example, when the goal is *timely delivery* of information over a shared wireless channel [5].

A very natural way to approach this problem is to treat the model as a queueing system, where services (transmissions) are controlled by a so called MaxWeight scheduler (see [3, 9, 10]), which serves a set of *virtual queues* (one for each traffic flow), each receiving new work at the rate λ_i. (See e.g. [1].) This automatically achieves objective (a), if this is feasible at all; MaxWeight is known to be *throughput optimal* – stabilize the queues if this is feasible at all. The MaxWeight stability results, however, do not tell whether or not the objective (b) is achieved. Specifically, when the system is heavily loaded, i.e. the vector $\boldsymbol{\lambda} = (\lambda_i)$ is within the system *rate region* $\boldsymbol{V}$, but close to its boundary, the steady-state queue lengths under MaxWeight are necessarily large, and it is conceivable that this may result in large time-gaps in service for individual flows. (Note that, if (a) and (b) are the objectives and the queues are virtual, the large queue lengths in themselves are not an issue. As long as (a) and (b) are achieved, minimizing the queue lengths is not important.) Our main results show that this is *not* the case. Namely, in the heavy traffic regime, when $\boldsymbol{\lambda} \to \boldsymbol{\lambda}^\star$, where $\boldsymbol{\lambda}^\star$ is a point on the outer boundary of rate region $\boldsymbol{V}$, the service process remains "smooth", in the sense that its stationary regime converges to that of a positive recurrent Markov chain, whose structure is given explicitly.

We emphasize two features of our model. First, our heavy traffic results do *not* require the *complete resource pooling* (cf. [3, 9]), namely the condition that there is a unique outer normal vector to region $\boldsymbol{V}$ at point $\boldsymbol{\lambda}^\star$. Second, we assume that the (random) amounts of new work arriving into the queues have *absolutely continuous* distribution; this assumption is non-restrictive if we create virtual queues, artificially, for the purpose of applying MaxWeight algorithm – in this case the structure of the virtual arrival process is within our control.

The details of our results and proofs are in [11].

2. MODEL

We consider a system of N flows served by a “switch", which evolves in discrete time $t = 0, 1, \ldots$. At the beginning of each

SIGMETRICS'15, June 15i£¡19, 2015, Portland, OR, USA.
ACM 978-1-4503-3486-0/15/06.
http://dx.doi.org/10.1145/2745844.2745878.

time-slot, the scheduler has to choose from a finite number K of "service-decisions". If the service decision k is chosen, then independently of the past history the flows get an amount of service, given by a random non-negative vector. Furthermore, we assume that (if decision k is chosen), there is a finite number $\mathcal{O}_k$ of possible service-vector outcomes, i.e. with probability $p^{k,j}$, it is given by a non-negative vector $\boldsymbol{v}^{k,j}$. The expected service vector for decision k is denoted $\boldsymbol{\mu}^k$. We assume that vectors $\boldsymbol{\mu}^k$ are non-zero and different from each other; and that for each i there exists k such that $\mu_i^k > 0$.

We denote by $\boldsymbol{S}(t) = (S_1(t), \ldots, S_N(t))$ the (random) service vector at time t, and call $\boldsymbol{S}(\cdot)$ the service process.

After the service at time t is completed, a random amount of work arrives into the queues, and it is given by a non-negative vector $\boldsymbol{A}(t)$. The values of $\boldsymbol{A}(t)$ are i.i.d. across times t, and $\boldsymbol{A}(\cdot)$ is called the arrival process. The mean arrival rates of this process are given by vector $\boldsymbol{\lambda}$.

If $\boldsymbol{Q}(t)$ is the vector of queue lengths at time t, then for each i, $Q_i(t+1) = \max\{Q_i(t) - S_i(t),\ 0\} + A_i(t)$.

Let parameters $\gamma_i > 0$ be fixed for all i. MaxWeight scheduling algorithm chooses, at each time t, a service decision

$$k \in \arg\max_l \sum_i \gamma_i Q_i(t) \mu_i^l,$$

with ties broken according to any well defined rule.

We will consider a sequence of systems, indexed by $n \to \infty$, operating under MaxWeight scheduling. (Variables pertaining to n-th system will be supplied superscript (n).) The switch parameters will remain unchanged, but the distribution of $\boldsymbol{A}^{(n)}(t)$ changes with n, for each n it has density $f^{(n)}$, and $f^{(n)}$ uniformly converges to some density f^*. The arrival process $\boldsymbol{A}^\star(\cdot)$, such that the distribution of $\boldsymbol{A}^\star(t)$ has density $f^\star$, has the arrival rate vector $\boldsymbol{\lambda}^\star$. Correspondingly, $\boldsymbol{\lambda}^{(n)} \to \boldsymbol{\lambda}^\star$.

We assume that $\boldsymbol{\lambda}^\star$ has positive components and is a maximal element of rate region $\boldsymbol{V}$. We further assume that for each n, $\boldsymbol{\lambda}^{(n)}$ lies in the *interior* of $\boldsymbol{V}$; therefore, the system is stable for each n (under the MaxWeight algorithm).

3. MAIN RESULTS

We define the notion of *asymptotic smoothness* of the steady-state service process. Informally, it means the property that as the system load approaches critical, the steady state service processes are such that for each flow the probability of a T-long gap (without any service at all) uniformly vanishes, as $T \to \infty$.

For each n, consider the cumulative service process $\boldsymbol{G}^{(n)}(\cdot)$ in steady state. Namely,

$$\boldsymbol{G}^{(n)}(t) \triangleq \sum_{\tau=1}^{t} \boldsymbol{S}^{(n)}(\tau), \quad t = 1, 2, \ldots$$

DEFINITION 1. *We call the service process asymptotically smooth, if*

$$\max_i \lim_{T\to\infty} \left(\limsup_{n\to\infty} P\left(G_i^{(n)}(T) = 0\right)\right) = 0.$$

Our key result, Theorem 21 in [11], is concerned with the following "queue-differential" process $\boldsymbol{Y}^{(n)}(t)$, which is the orthogonal projection of Markov process $(\gamma_1 Q_1^{(n)}(t), \ldots, \gamma_N Q_N^{(n)}(t))$ on the subspace orthogonal to the outer normal cone to $\boldsymbol{V}$ at point $\boldsymbol{\lambda}^\star$. We prove that the stationary version of the *non-Markov* process $\boldsymbol{Y}^{(n)}(\cdot)$ converges to that of the positive Harris recurrent *Markov* process $\boldsymbol{Y}^\star(\cdot)$, which, informally speaking, can be thought of as the queue-differential process of the critically loaded system (with mean arrival rates given by $\boldsymbol{\lambda}^\star$). The structure of the process $\boldsymbol{Y}^\star(\cdot)$ is defined explicitly. Moreover, as $n \to \infty$, the steady-state probability that the choice of a scheduling decision is uniquely determined by the current queue-differential $\boldsymbol{Y}^{(n)}(t)$ converges to 1. These results, in particular, imply the following

THEOREM 2. *Consider the sequence of systems described in Section 2, in the heavy traffic regime. Under MaxWeight scheduling, the service process is asymptotically smooth.*

4. REFERENCES

[1] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, P. Whiting. Providing Quality of Service over a Shared Wireless Link. *IEEE Communications Magazine*, 2001, Vol.39, No.2, pp.150-154.

[2] M. Andrews, K. Jung, A. L. Stolyar. Stability of the Max-Weight Routing and Scheduling Protocol in Dynamic Networks and at Critical Loads. STOC'07, San Diego, CA, June 11-13, 2007.

[3] A. Eryilmaz and R. Srikant. Asymptotically Tight Steady-State Queue Length Bounds Implied by Drift Conditions. *Queueing Systems*, 2012, Vol.72, No.3-4, pp.311-359.

[4] B. Hajek. Hitting-Time and Occupation-Time Bounds Implied by Drift Analysis with Applications. *Advances in Applied Probability*, 1982, Vol.14, No.3, pp. 502-525.

[5] I-H. Hou, V. Borkar and P.R. Kumar. A Theory of QoS for Wireless. *INFOCOM 2009*.

[6] R. Li, A. Eryilmaz, B. Li. Throughput-Optimal Wireless Scheduling with Regulated Inter-Service Times. *INFOCOM 2013*.

[7] B. Li, R. Li, A. Eryilmaz. Heavy-Traffic-Optimal Scheduling with Regular Service Guarantees in Wireless Networks. *MobiHoc 2013*.

[8] S. P. Meyn and R.L. Tweedie. Stability of Markovian Processes I: Criteria for Discrete-Time chains. *Advances in Applied Probability*, 1992, Vol.24, No. 3, pp. 542-574.

[9] A.L. Stolyar. Max Weight Scheduling in a Generalized Switch: State Space Collapse and Workload Minimization in Heavy Traffic. *Annals of Applied Probability*, 2004, Vol.14, No. 1, pp. 1-53.

[10] L. Tassiulas, E. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Automat. Control*, 1992, Vol. 37, No. 1, pp. 1936 - 1948.

[11] R. Singh and A. Stolyar. MaxWeight Scheduling: Asymptotic Behavior of Unscaled Queue-Differentials in Heavy Traffic. `http://arxiv.org/abs/1502.03793`