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


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 \( \lambda_i \). (See e.g. [11]) 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 \( \lambda = (\lambda_i) \) is within the system rate region \( V \), 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 \( \lambda \rightarrow \lambda^* \), where \( \lambda^* \) is a point on the outer boundary of rate region \( V \), 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 \( V \) at point \( \lambda^* \). 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
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 $N_k$ of possible service-vector outcomes, i.e., with probability $p^k$, it is given by a non-negative vector $v^k$. The expected service vector for decision $k$ is denoted $\mu_k$. We assume that vectors $\mu_k$ are non-zero and different from each other; and that for each $i$ there exists $k$ such that $\mu_{ik} > 0$.

We denote by $S(t) = (S_1(t), \ldots, S_N(t))$ the (random) service vector at time $t$, and call $S(t)$ 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 $A(t)$. The values of $A(t)$ are i.i.d. across times $t$, and $A(\cdot)$ is called the arrival process. The mean arrival rates of this process are given by vector $\lambda$.

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

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

\[ k \in \arg \max_i \sum_{i} \gamma_i Q_i(t) \mu_i, \]

to that of the positive Harris recurrent Markov process $Y^*(\cdot)$, which, informally speaking, can be thought of as the queue-differential process of the critically loaded system (with mean arrival rates given by $\lambda^\star$). The structure of the process $Y^*(\cdot)$ 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 $Y^{(n)}(t)$ 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


