Delay and Effective Throughput of Wireless Scheduling in Heavy Traffic Regimes: Vacation Model for Complexity

Yung Yi*
School of EECS, KAIST
Daejon, South Korea
yiyung@ee.kaist.ac.kr

Junshan Zhang†
EE Dept. Arizona State Univ.
Tempe, AZ, USA
Junshan.Zhang@asu.edu

Mung Chiang‡
EE Dept. Princeton Univ.
Princeton, NJ, USA
chiangm@princeton.edu

ABSTRACT

Distributed scheduling algorithms for wireless ad hoc networks have received substantial attention over the last decade. The complexity levels of these algorithms span a wide spectrum, ranging from no message passing to constant/polynomial time complexity, or even exponential complexity. However, by and large it remains open to quantify the impact of message passing complexity on throughput and delay. In this paper, we study the effective throughput and delay performance in wireless scheduling by explicitly considering complexity through a vacation model, where signaling complexity is treated as "vacations" and data transmissions as "services," with a focus on delay analysis in heavy traffic regimes. We analyze delay performance in two regimes of vacation models, depending on the relative lengths of data transmission and vacation periods. State space collapse properties proved here enable a significant dimensionality reduction in the challenging problem of delay characterization. We then explore engineering implications and quantify intuitions based on the heavy traffic analysis.

Categories and Subject Descriptors

C.2 [**Computer-Communication Networks**]: Network Architecture and Design

General Terms

Algorithms, Performance, Theory

Keywords

Scheduling, Wireless Networks, Complexity, Heavy-Traffic

*Y. Yi was partially supported by IT R&D program [Ultra Small Cell Based Automatic Wireless Network] of MKE/IITA and Korea Research Council of Fundamental Science and Technology.

†J. Zhang is supported in part by National Science Foundation through the grants ANI-0238550 and CNS-0721820.

‡M. Chiang is supported in part by NSF grant CNS-0720570, ONR YIP, and PECASE awards.

MobiHoc'09, May 18–21, 2009, New Orleans, Louisiana, USA.

1. INTRODUCTION

1.1 Overview and Motivation

Since the seminal work [25] on throughput maximization for constrained queueing systems, there have been a surge of interest in wireless scheduling with provable, topology-independent performance-guarantees over wireless ad-hoc networks. As will be defined in the next section, *throughput* and the associated stability region, has been the main object to be maximized, often using the mathematical techniques of fluid-models and Lyapunov theory. To overcome exponentially high complexity and centralized operation of the optimal algorithm in [25], referred to as the Max-Weight (MW) scheduling, many distributed algorithms with lower complexity have recently been studied. These include maximal and greedy scheduling, e.g., [3, 14, 29], decentralized pick-and-compare, e.g., [8, 15, 18] as motivated by [24], and constant-time random access, e.g., [10, 13].

These distributed algorithms require the communication overhead of message passing, which can be quantified by the temporal measure of signaling complexity, or, in short, *time complexity*: the fraction of time spent on communicating and updating queue information. Such signaling overhead may have significant impact on performance of considered distributed scheduling schemes. For example, even distributed algorithms (e.g., [8, 15]), known to provably achieve optimality in throughput with polynomial-time computational complexity, need heavy message passing. The corresponding "effective" performance, in terms of throughput and delay after such complexity is taken into account, is not well understood.

In this paper, we aim at analyzing effective throughput, and then delay performance, of a large family of wireless scheduling algorithms by *explicitly considering complexity*. We model signaling time complexity by "vacation," where the system stops serving packets and takes vacation to compute a new schedule. Using the vacation model, we introduce *Generalized Max-Weight (GMW)* scheduling as a general family of scheduling algorithms that include many major classes of distributed algorithms in the literature.

The impact of signaling overhead on delay characterization presents a particularly demanding challenge. In spite of significant effort towards wireless scheduling research with throughput-guarantee, delay performance is generally an underexplored area, mainly due to technical intractability when applying the following three popular approaches:

1. Standard queueing analysis does not work well for scheduling algorithms in large-scale distributed systems, since it is intractable to explicitly characterize the stationary

queue-length distribution due to the complex coupling between the queue sizes and employed scheduling algorithms. There exists some work [9] that reduces the original system to a simpler system enabling computation of delay bounds.

2. Bounds (on the sum of stationary queue length over all links) have been studied using Lyapunov technique in e.g., [11, 16, 30]. Similar techniques have been used in [30] to study the 3-dimensional tradeoff among time complexity, throughput, and delay, which generalizes the 2-dimensional tradeoff between complexity and throughput in [17, 18]. However, Lyapunov bounds are often very loose.

3. Large deviation techniques provide another alternative to the study of delay, e.g., in [20, 23, 26, 31]. However, due to dimensionality the proof techniques are not scalable, and only small networks with restrictive topologies allow exact analysis.

In this paper, we advocate the use of a fourth approach: namely *heavy traffic approximation*, develop the corresponding theory for the problem of wireless scheduling with overhead accounting, and apply it to our vacation model. The benefits of this approach will be illustrated in the subsections on engineering implication. Simply put, heavy traffic analysis focuses on the network model with bottleneck links (i.e., arrival volumes are almost close to system capacities at these links). One can approximate the original system $Q(t) = \{Q_l(t)\}$[1] by considering a sequence of systems, appropriately scaled in time and space, and studying the limiting system. The limiting system is often an interesting one that allows mathematical tractability and lends insights to the engineering of scheduling beyond the existing analysis.

However, technical challenges of the heavy traffic approach are also substantial. In particular, as an important intermediate step for dimensionality reduction, one needs to prove the state space collapse property, which allows the original multi-dimensional queue length vectors to be expressed as a scaled version of a one-dimensional workload process. These proofs tend to be challenging.

Although the state-space collapse technique has also been used in multi-class queueing networks, e.g., [2, 28], considering vacations over various regimes to include signaling complexity entails additional technical difficulties. For example, heavy traffic analysis of scheduling in [19, 21, 22] does not consider signaling complexity, which is the main focus in this work. A more general stochastic network is considered for heavy traffic analysis in [7], again without considering signaling complexity. Similarly, vacation models for queueing networks in [1, 4, 12] has only studied much simpler systems, e.g., a single-server queue or polling systems.

1.2 Main Contributions and Organization

In Section 2, we first propose a new scheduling family, GMW (Generalized Max-Weight), that enables us to explicitly consider signaling complexity as well as cover many distributed algorithms in literature. In Section 3, we study the effective throughput of the GMW.

In Section 4.2, we consider delay in a regime where the average vacation duration and transmission duration are comparable and are both of $O(1)$. As expected, the limiting process of the scaled process in this regime turns out to be a reflected Brownian motion. We provide mean workload analysis for delay performance using the statistical properties of the reflected Brownian motion. Different from Lyapunov bounds on the delay, our results on the limiting workload process provide an exact characterization of the average delay.

In Section 4.3, in contrast to Regime I, we explore a regime where the average transmission duration is significantly larger than the average vacation duration. In this regime, vacation periods are relatively small, compared to transmission periods, where it is expected that possibly higher throughput can be achieved and yet infrequent schedule updates may lead to larger delay. Technically, we let the average transmission duration scale together with the diffusion scale n while fixing the average vacation duration to be $O(1)$. This regime generates much deeper technical difficulties, especially in terms of proving the state-space collapse property. Significantly different from Regime I, the limiting process can be either a reflected Brownian motion or its mixture with stable Levy motion, depending on the system parameters used in the analysis. Accordingly, our results indicate that due to infrequent updates, larger average transmission durations may lead to larger delay and even grow unbounded under some conditions.

1.3 Notation

We use capital letters, say X or $X(t)$, to refer to a random variable or random process. We use $a \cdot b$ to denote the inner product of two vectors a and b. The boundary of a set A is denoted by $\partial(A)$, and $[x]$, $x \in \mathcal{R}$ refers to the large integer no greater than x. As is standard, for a given random process $X(t)$, we use $\tilde{x}^n(t)$ and $\hat{x}^n(t)$ to refer to the *fluid-scaled* and the *diffusion-scaled* processes of $X(t)$ in the n-th system, respectively, i.e.,

$$\tilde{x}^n(t) = \frac{X(nt)}{n}, \quad \hat{x}^n(t) = \frac{X(n^2 t)}{n}.$$

Throughout this paper, we use the superscript n to index the sequence of random variables or processes. For a sequence of processes $\{X^n(t) : t \geq 0\}$, we use $X^n(t) \Rightarrow X(t)$ to denote "weak convergence," or equivalently "convergence in distribution." Let $L^\alpha(t)$ denote *α-stable Levy motion* where $0 < \alpha \leq 2$. In the special case $\alpha = 2$, $L^\alpha(t)$ boils down to a standard Brownian motion, which we denote by $B(t)$. We use $\Phi(f)$ to denote the one-dimensional reflection mapping of a process $f(t)$, $f(0) \geq 0$:

$$\Phi(f)(t) = f(t) - \inf_{0 \leq s \leq t} \Big(f(s) \wedge 0 \Big),$$

where $a \wedge b = \min(a, b)$. Roughly speaking, Φ generates the "reflection" of f around the horizontal axis.

2. GMW SCHEDULING AND VACATION MODEL FOR COMPLEXITY

2.1 System Model

Network Model. We model a wireless multi-hop network by a graph $G(L, N)$, where L and N denote the set of (bidirectional) links and the set of nodes, respectively. When it is clear from the context, we abuse the notation and use L and N to refer to the number of links and nodes. The wireless system has a single channel (e.g., frequency or code), and each node is time-synchronized and has a half-duplex radio. We assume that time is slotted, indexed by t. For

[1] $Q_l(t)$ is the queue length stochastic process of link l.

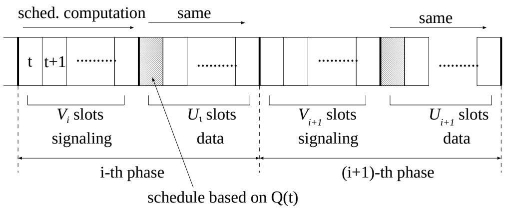


Figure 1: Generalized max-weight scheduling with vacation: V_i slots are used to compute a newly computed max-weight schedule, which is used for U_i slots starting at shaded time instances.

simplicity, we assume that one slot length is chosen to serve one fixed-size packet that is normalized to one unit.

Signaling Model. In a distributed scheduling, nodes typically exchange messages to inform their states (e.g., queue length information) to other nodes, then the states are used to determine a schedule during a slot. We refer to the time to exchange control message as *signaling time.* Therefore, a time slot can be one of the following two states: *signaling* or *data transmission* (simply *transmission*). In the state of signaling, the system stops serving packets and takes "vacations" to exchange signaling messages. We will discuss more details on a class of scheduling algorithms considered in this paper, as well as vacation modeling at Subsection 2.2. We denote by $V(t)$ and $U(t)$ the *cumulative* number of vacation and transmission slots until time t, respectively. Clearly $V(t) + U(t) = t$.

Traffic Model. In this study, we consider single-hop sessions, i.e., each link is fed by exogenous arrivals. We denote by $A_l(t)$ the number of arrivals over link l at slot t, assumed to be i.i.d, and also independent across links. Note that arrivals occur regardless of the state of a slot, i.e., signaling or transmission. We let $\text{E}[A_l(t)] = \lambda_l$, and $\text{VAR}[A_l(t)] = (\sigma_l^a)^2$. Denote by $F_l(t)$ and $G_l(t)$ the number of *cumulative* arrivals and departures over link l over the slot interval $[0, t]$, respectively.

Resource Model. The network resources are represented by a finite set $\mathcal{S} \subset \{0, 1\}^L$ of the link schedules. A *link schedule*, or simply a *schedule,* $S = (S_l \in \{0, 1\} : l = 1, \ldots, L)$, is a vector representing the set of scheduled links *without any interference*, where $S_l = 1$ if the link l is scheduled, and 0 otherwise. The set $\mathcal{S}$ depends on interference patterns among links. We model interference by a $L \times L$ symmetric matrix $I = [I_{ij}]$, where $I_{ij} = 1$ means that links i, j interfere with each other. This interference model generalizes popular one-hop (appropriate for Bluetooth and FH-CDMA) or two-hop (appropriate for IEEE 802.11) interference models. The set $\mathcal{S}$ of all link schedules are indexed by a set $\mathcal{J}$. We denote by $S^j = (S_l^j, l \in L)$, $j \in \mathcal{J}$ the j-th schedule in $\mathcal{S}$. Then, a scheduling algorithm chooses a sequence of schedules $(S(t), t = 0, 1, \ldots, S(t) \in \mathcal{S})$, where the schedule $S(t)$ is applied to the network whenever a slot t is not in vacation.

2.2 GMW with Vacation

2.2.1 Description

We now describe a class of scheduling algorithms, referred to as *generalized max-weight (GMW) scheduling.* In GMW, as conceptually depicted in Figure 1, entire time slots are divided into *phases,* each of which consists of *vacation phase* and *transmission phase.* One vacation phase (as well as a transmission phase) takes multiple consecutive slots. We

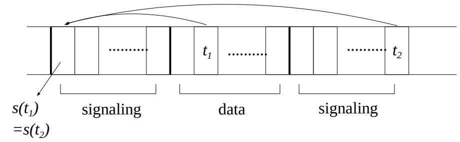


Figure 2: $s(t)$ for a given slot t.

denote by V_i and U_i the i.i.d. random lengths (i.e., number of slots) of i-th vacation and transmission phase (or simply i-th vacation and transmission duration), respectively[2]. We also denote by $T(t)$ the cumulative number of phases over the slot interval $[0, t]$.

To get a more concrete sense of how the schedules are updated, we recall the well-known Max-Weight (MW) rule: For a given queue length $Q = (Q_l)_{l \in L}$, a schedule by the MW rule is one that maximizes an "aggregate weight" over all schedules, i.e.,

$$S^\star = \max_{S \in \mathcal{S}} \Big(Q \cdot S \Big). \qquad (1)$$

The system spends V_i slots on computing a new max-weight schedule for the queue lengths at the beginning slot of the vacation phase V_i. The scheduled links by the new max-weight schedule performs data transmission during the subsequent transmission phase that spans U_i slots. GMW is essentially an algorithm that computes a max-weight schedule and update it after a vacation phase.

Denote by $Q_l(t)$ the queue length of link l at slot t. Let $C_S(t)$ be the cumulative number of slots that use the schedule $S \in \mathcal{S}$ until slot t. To facilitate the characterization of queueing dynamics, we keep incrementing $C_S(t)$ at a vacation phase, say i, when the schedule S was used at the previous transmission phase $i-1$. We have $\sum_{S \in \mathcal{S}} C_S(t) = t$ (i.e., (4)). However, obviously actual data transmissions do not occur at the i-th vacation phase.

The queueing dynamics is represented by:

$$Q_l(t) = Q_l(0) + F_l(t) - G_l(t), \qquad (2)$$

$$G_l(t) = \sum_{S \in \mathcal{S}} \sum_{\tau=1}^{t} \Big(S_l 1_{\{Q_l(t)>0\}} (U(\tau) - U(\tau - 1)) \times (C_S(\tau) - C_S(\tau - 1)) \Big) \qquad (3)$$

$$\sum_{S \in \mathcal{S}} C_S(t) = t, \quad C_S(\cdot) \text{ is non-decreasing.} \qquad (4)$$

Note that $U(\tau) - U(\tau - 1)$ is '1' when the slot τ is in a transmission phase, and '0' in a vacation phase. The fact that $C_S(\tau) - C_S(\tau - 1) = 1$ means that a schedule S is (potentially) used at slot τ, as long as τ is not in vacation phase, i.e., $U(\tau) - U(\tau - 1) = 1$.

For a given slot t, denote by $s(t)$ the beginning slot of the vacation phase prior to the phase (either vacation or transmission) that a slot t belongs to, as shown in Figure 2. Then, the GMW scheduling adds the its own unique dynamics (based on the MW rule) to (2), (3), and (4):

$$C_S(t + 1) = C_S(t) \text{ if } Q(s(t)) \cdot S < Q(s(t)) \cdot S', \qquad (5)$$

for some schedule S'. This is due to the fact that a schedule at some slot t is a max-weight schedule based on the queue

[2]The i.i.d assumption of $A_l(t)$, V_i, and U_i can be readily extended to more general correlated models under reasonable conditions.

length at slot $s(t)$. We assume a random tie-breaking rule when there exists multiple MW schedules.

2.2.2 Generality of GMW with Vacation

Delay analysis of the GMW scheduling family, which includes many existing scheduling mechanisms, is useful for a number of reasons.

First, the Max-Weight rule [25] is known to be provably throughput-optimal, if the complexity of passing information to the centralized controller is ignored. The Max-Weight rule requires very high complexity as it is reduced to an NP-hard WMIS (Weighted Maximum Independent Set) problem. Its various distributed implementations take significant signaling time too, as discussed in Section 1. The GMW scheduling framework explicitly considers signaling time by introducing vacations, and the vacation duration would be dependent on network size.

Second, the idea of infrequent updates with max-weight schedules can be exploited to provide a good way to develop a low-complexity scheduling and understand its fundamental principle. To elaborate, in [24], a randomized technique has been introduced, called *pick-and-compare* to overcome high complexity in the Max-Weight rule: The basic idea was that (i) first randomly choose any schedule out of all possible schedules (with guarantee of a lower bound on probability to find a max-weight schedule), (ii) compare the aggregate weight of the randomly-chosen schedule to that of the previous schedule, and finally (iii) choose the schedule with larger aggregate weight. Extensions of pick-and-compare is also provably throughput-optimal with polynomial complexity, again ignoring the time complexity that forms the focus of this paper, and has been widely applied to many types of sophisticated distributed scheduling algorithms, e.g., [8,15,18]. It turns out that pick-and-compare is a "version" of the Max-Weight scheduling with infrequent schedule update, such that once a schedule is determined, then the same schedule (or a better schedule at least due to weight comparison) is repeated for some duration [30].

Therefore, the GMW framework generalizes an array of scheduling algorithms based on the idea of max-weight, yet is more practically accurate by considering signaling complexity. To reflect the random nature of signaling complexity as well as transmission durations (e.g., pick-and-compare algorithms), we model vacation and transmission durations by random variables in the GMW.

2.3 Performance Metrics

Throughput. A central performance metric of scheduling is the throughput, for which we first define a notion of throughput region as follows:

DEFINITION 1 (THROUGHPUT REGION). *The throughput region Λ is the set of all mean arrival rate vectors $\lambda = (\lambda_l : l \in L)$ that is rate-stabilized by some scheduling scheme, where the system is rate-stable for a given λ, if* $\lambda_l = \lim_{t\to\infty} G_l(t)/t$, *a.s., for all $l \in L$.*

The throughput region is interpreted as the maximum achievable throughput. As mentioned earlier, the resource occupied by signaling would incur throughput loss, indicating that even the Max-Weight rule may become sub-optimal after taking into account signaling complexity. To quantify such sub-optimality due to signaling complexity, in what follows we introduce γ-effective throughput optimality.

DEFINITION 2 (γ-EFFECTIVE-THROUGHPUT-OPTIMALITY). *A scheduling algorithm Π is γ-effective-throughput-optimal, if it can stabilize the system for any $\lambda \in \gamma\Lambda$, after taking into account the throughput loss due to signaling overheads. Then, γ is said to be effective-throughput-ratio*[3] *of the scheduling algorithm Π.*

Delay. Characterizing the exact delay performance metrics, such as the stationary distribution of the queue lengths or even the exact aggregate total queue length over links, has been known to be very challenging and been open for a long while, due to complex queueing dynamics. As an efficient means to extract implications on delay, we focus on scenarios when the system has (unique) bottleneck links, i.e., these links are heavily loaded, i.e., arrival intensities are very close to the boundary of the achieved throughput region.

3. EFFECTIVE THROUGHPUT OF GMW AND TWO REGIMES

3.1 Effective Throughput of GMW

We first study the effective throughput of GMW. Recall that $\{V_i\}$ and $\{U_i\}$ denote the i.i.d. vacation durations and transmission durations. Let α and β be the corresponding expectations, i.e.,

$$\mathrm{E}[V_i] = \alpha, \quad \mathrm{E}[U_i] = \beta.$$

For convenience, we refer to GMW with such parameters as GMW(α, β). Theorem 1 characterizes the effective throughput of GMW(α, β).

THEOREM 1 (EFFECTIVE THROUGHPUT OF GMW). *The GMW(α, β) is $(\frac{\beta}{\alpha+\beta})$-throughput optimal.*

The effective throughout can be obtained by using the fluid-limit approach [6], while taking into account the additional complication due to interleaved vacations and data transmissions. We outline the key steps as follows: 1) on average, $\frac{\beta}{\alpha+\beta}$ per unit time is used for data transmission, which leads to $(\frac{\alpha}{\alpha+\beta})$ throughput-reduction, and 2) a repeated use of the same scheduling (e.g., a max-weight schedule at a transmission phase for the queue length at the beginning slot of each vacation phase) during a *finite* number of slots would not impact the asymptotic stability region. More specifically, it has the same stability region that is achieved by the *modified* system, where, instead of using a same schedule repeatedly during a transmission phase, *max-weight schedules are computed and used per-slot during the transmission phase* (note that vacation phases still exist as in the original system).

3.2 Two Regimes

We shall develop heavy traffic approximations for the delay performance of the GMW family. Similar to congestion control in spirit, heavy traffic approximation focuses on network models with bottleneck links (i.e., arrival volumes are almost close to system capacities at these bottleneck links). For such network models, one can approximate the original system $Q(t) = \{Q_l(t)\}$ by considering a sequence of systems, appropriately scaled in time and space, and studying the limiting system. Roughly speaking, the trajectory of the

[3] We henceforth omit the word 'effective' for simplicity, unless necessary.

scaled system over a closed interval, where the time axis is stretched by n^2 and the space is scaled down by n, offers good approximation for the original system for large n.

Consider the vacation model for GMW scheduling. Intuitively speaking, the higher the complexity is, the longer the vacation phases are. The larger the transmission phases are, then less frequent the scheduling updates are. Recall that in general the vacation and transmission durations are random. We shall develop diffusion approximations of the delay performance, for the following continuum of regimes that are differentiated by transmission and vacation durations: the average vacation duration is $O(1)$; and the average transmission duration depends on the scaling parameter n, i.e., $O(n^{2-r})$, for some constant $1 < r \leq 2$. Roughly speaking, for $r > 0$, the transmission duration also increases as the scaling parameter n increases, and this is applicable to systems when the network is operated at relatively lower signaling complexity and schedules are updated less and less frequently. In an extreme case with $r = 2$, both average transmission and average vacation durations are $O(1)$, which models a system with constant (average) signaling complexity and transmission duration. More formally, we define two regimes[4] to differentiate the case with $1 < r < 2$ from the degenerate case with $r = 2$.

Regime I:	Both average transmission duration and average vacation duration are $O(1)$.
Regime II:	The average vacation duration is $O(1)$, but the average transmission duration increases with the scaling parameter n, i.e., $O(n^{(2-r)})$, for some constant $1 < r < 2$.

4. HEAVY TRAFFIC DELAY ANALYSIS

4.1 CRP and Workload Process

The *workload process* is a critical metric in heavy traffic analysis, where the workload is defined as the amount of *time* to drain all the queues. In what follows, we formally define the workload process through a linear optimization problem, by which we introduce a standard condition, namely CRP (Complete Resource Pooling). Simply put, the CRP condition holds the key to establish the state space collapse property, which makes it possible to express the L-dimensional queue length process as a scaled version of a one-dimensional workload process, thus overcoming the key bottleneck of high dimensionality in delay characterization.

Consider a γ-throughput-optimal scheduling algorithm in the GMW family. Since we explicitly consider signaling time using vacations, the algorithm may have sub-optimal throughput performance (i.e., $\gamma < 1$), whereas in the "traditional" analysis of MW scheduling, γ is always one. We refer the readers to [2, 7, 19, 21, 22] for the similar CRP condition in other types of networks.

Consider the following *primal* problem:

$$
\begin{array}{lll}
\textbf{(Primal)} & \max & c, \\
 & \text{s. t.} & \displaystyle\sum_{j\in\mathcal{J}} \zeta_i S_l^j \geq c\lambda_l, \quad \forall l \in L, \\
 & & \displaystyle\sum_{i\in\mathcal{J}} \zeta_j = \gamma, \\
 & & c \in \mathcal{R}, \zeta_j \geq 0, \forall j \in \mathcal{J}, \\
 & \text{variables} & c,\ \zeta = (\zeta_j : j \in \mathcal{J}) \\
 & \text{constants} & \lambda = (\lambda_l), S = (S^j), \gamma
\end{array} \quad (6)
$$

[4]We exclude the case with $0 < r \leq 1$ since it is not meaningful.

First, note that for any stabilizable arrival vector λ (i.e., $\lambda \in \gamma\Lambda$), we can find the constants $(\zeta_j : j \in \mathcal{J})$, such that $\sum_{j\in\mathcal{J}} \zeta_j S^j \geq \lambda$, and $\sum_{j\in\mathcal{J}} \zeta_j \leq \gamma$. The ζ_j can be interpreted as the long-term fraction of time that the schedule S^j is used. Therefore, the above problem is find the maximum value by which λ can be scaled up-to the boundary of the throughput region, $\gamma\Lambda$. Intuitively, c can be regarded as "distance" between λ and the boundary of the achieved throughput region.

We can regard $\rho = 1/c^\star$ as the *system load*, where $c^\star$ is the optimal solution of the primal problem. Obviously, if $c^\star = 1$, then λ turns out to be exactly at the boundary of $\gamma\Lambda$.

Next, consider the corresponding Lagrange dual problem:

$$
\begin{array}{lll}
\textbf{(Dual)} & \min & \gamma\ \beta, \\
 & \text{s. t.} & \displaystyle\sum_{l\in L} \eta_l \lambda_l = 1, \\
 & & \beta \geq \displaystyle\sum_{l\in L} \eta_l S_l^i, \forall i \in \mathcal{I} \\
 & & \beta \in \mathcal{R}, \eta_l \geq 0, \forall l \in L, \\
 & \text{variables} & \beta, \eta = (\eta_l : l \in L) \\
 & \text{constants} & \lambda = (\lambda_l), S = (S^j), \gamma.
\end{array} \quad (7)
$$

Each pair of (η, β) that are in the constraints of **Dual** is said to be a *resource pool.* The η_l is interpreted as the work dedicated to a unit of the buffer of the link l by the resource pool. A *bottleneck pool* is defined to be an optimal solution $(\eta^\star, \beta^\star)$ (that may not be unique) to the **Dual** problem. If the bottleneck pool is unique, we say that the given arrival vector λ *satisfies the complete resource pooling (CRP) condition.*

To get a more concrete sense of the physical meaning of the CRP condition, the arrival vector $\lambda = (\lambda_l : l \in L)$ satisfying CRP condition is such that it lies in the relative interior of *one of the faces* of the corresponding throughput region $\gamma\Lambda$. As a simple example, consider two interfering links with unit capacity, and fed by dynamic arrivals with means λ_1 and λ_2, respectively. Then, the maximum throughput region (that may be achieved by MW scheduling assuming that signaling complexity is zero) is the convex polyhedron connecting $(0, 1)$ and $(1, 0)$ having one face. Then, $\lambda = (0, 1)$ or $(1, 0)$ does not satisfy CRP condition, whereas $(0.5, 0.5)$ does.

The CRP condition enables us to uniquely define the workload process, as well as provide useful properties in heavy traffic analysis (e.g., see Step 1 of the proof of Theorems 2 and 3 at the appendix). Formally, for a unique bottleneck pool $(\eta^\star, \beta^\star)$, we define the notion of *workload process*, given by:

Definition 3 (Workload Process). *The workload process $W(t)$ is the average total work of the bottleneck pool, i.e.,*

$$
\begin{aligned}
W(t) &\triangleq \eta^\star \cdot Q(t) \\
&= \eta^\star \cdot Q(0) + \eta^\star \cdot F(t) - \eta^\star \cdot G(t)
\end{aligned} \quad (8)
$$

We will consider a system with arrival vectors satisfying CRP condition. For notational simplicity, we henceforth use just η instead of $\eta^\star$ to be the solution of the **Dual** problem. In the next two sections, we study the workload processes in different regimes under the CRP condition.

4.2 Delay Performance in Regime I

We start with the regime where the transmission and vacation durations are of $O(1)$, both with finite variances, i.e.,

$$\textbf{Regime I:} \qquad \mathrm{E}[U_i] = u, \quad \mathrm{E}[V_i] = v,$$

where u and v are positive constants. Let $\mathrm{VAR}[U_i] = (\sigma^u)^2$, and $\mathrm{VAR}[V_i] = (\sigma^v)^2$. Recall that GMW in Regime I is $(\frac{u}{u+v})$-throughput optimal.

4.2.1 Heavy Traffic Diffusion Approximation

A heavy traffic regime refers to a model with bottleneck links, and if the arrival rates at these links satisfy the CRP condition, we will show that the L-dimensional queue lengths can be well approximated by a scaled version of a one-dimensional process workload process.

Consider a sequence of systems, indexed by n. In particular, the arrival vectors are denoted by λ_n. The heavy traffic condition in Regime I is given by:

A1. As $n \to \infty$, for some $b > 0$,

$$n \times \eta \cdot (\lambda^n - \lambda) \to -b, \tag{9}$$

where $\{\lambda^n\}$ and $\lambda \in \partial((\frac{u}{u+v})\Lambda)$ satisfy CRP condition.

In Regime I, the throughput ratio is $u/(u+v)$. Thus, for an arrival rate λ at the boundary of the throughput region, the fact that λ_n approaches λ with the rate of $\eta(\lambda^n - \lambda) \sim 1/n$, implies that for a large n, $\lambda^n \approx \lambda$, i.e., the network is heavily loaded.

Recall that $\hat{w}^n(t)$ and $\hat{q}^n(t)$ are the diffusion-scaled workload and the queue-length process in the n-th system, respectively. For convenience, define $\xi_i = \eta_i / \sum_{l=1}^{L} \eta_l^2$. We have the following result for Regime I.

THEOREM 2. *In Regime I, under Assumption* **A1***, as* $n \to \infty$,

$$(\hat{w}^n, \hat{q}^n) \Rightarrow (w^\star, q^\star), \tag{10}$$

where $w^\star = \Phi(z^\star)$, $q^\star = \xi \times w^\star$, *and*

$$z^\star(t) = w^\star(0) - bt + \sum_{l \in L} \eta_l \sigma_l^a B_l(t) + \mu \left(\frac{u\sigma^v}{u+v} B_v\Big(\frac{t}{u+v}\Big) - \frac{v\sigma^u}{u+v} B_u\Big(\frac{t}{u+v}\Big) \right), \tag{11}$$

where $\mu \triangleq \max_{S \in \mathcal{S}} \eta \cdot S$, *and* $B_l(t), B_v(t)$, *and* $B_u(t)$ *are independent standard Brownian motions.*

The proof follows the same line as that of Theorem 3 presented in the Appendix.

Based on Theorem 2, we have a few remarks in order:

1) The process $z^\star(t)$ is the *virtual* workload process, assuming that there always exist packets to serve in every queue. Thus, $z^\star(t)$ can be sometimes negative, since random arrivals allow queues to be un-backlogged. The reflection map Φ is responsible for "correcting" $z^\star(t)$ by reflecting it and leads to the real workload process $w^\star(t)$ that is always non-negative.
2) The vector ξ corresponds to the "inverse" of the vector η. The fact that $q^\star = \xi \times w^\star$ points to the state space collapse property: L-dimensional queue length process can be represented by a *constant multiple* of the one-dimensional workload process.
3) The parameter μ corresponds to the maximum possible amount of workload that could *potentially* be served at a time-slot when the system is in a transmission phase.

4.2.2 Engineering Implications

(1) Approximation for average workload

We elaborate further on the engineering implications of Theorem 2. Let $\mathrm{RBM}_X(\theta, \sigma^2)$ denote the reflected process of a Brownian motion X with drift θ and variance σ^2. Then, the process $\mathrm{RBM}_X(\theta, \sigma^2)$ has a stationary distribution if $\theta < 0$, in which case the stationary distribution is exponential with mean $-\sigma^2/2\theta$ [27].

Theorem 2 indicates $w^\star(t) = RBM_{z^\star}(-b, (\sigma_w^*)^2)$, where

$$(\sigma_w^*)^2 = \sum_{l \in L} (\eta_l \sigma_l^a)^2 + \frac{\mu^2}{(u+v)^{3/2}} \left((u\sigma^v)^2 + (v\sigma^u)^2 \right).$$

Then, it follows that

$$\text{average total workload} \quad = \quad (\sigma_w^*)^2 / 2b. \tag{12}$$

The above result on the average total workload can then be used to approximate the delay performance of GMW scheduling while taking into consideration of the throughput loss due to signaling complexity.

(2) Impact of network size on delay

A system designer may desire to design a scalable system that provides *fixed* throughput regardless of the network size (i.e., the number of links). The main design parameters are u, v and σ^u, and σ^v, determined by a scheduling algorithm chosen by the designer. We consider the case when the designer can control only u and v, but σ_l^a, b, σ^v, and σ^u stays the same. Note that due to the NP-hard property, $v \sim 2^L$. Thus, to achieve a fixed throughput $\gamma \triangleq u/(u+v)$, u should also scale as 2^L.

Consider the *average workload per link* defined as the average total workload divided by L. Remarking that $\sum_{l \in L} (\eta_l \sigma_l^a)^2 \sim L$, $\mu^2 \sim L^2$ (because $\mu = \eta \cdot \lambda$), and $u, v \sim 2^L$, we have:

$$\text{average workload} \sim \frac{1}{L}(L + L^2 4^L) = O(4^L). \tag{13}$$

That is to say, the delay increases exponentially (to support a fixed throughput performance) with the network size L. This result, based on Theorem 2, provides an exact characterization of the average delay, different from the delay bound computed from the Lyapunov approach, e.g., in [11, 30]. It is known that the bound based on the Lyapunov approach is tight in terms of order only for restricted network topologies with special structures. In contrast, the complementary approach in this paper produces results that apply to any network topology, but under the heavy traffic regime only.

4.3 Delay Performance in Regime II(r)

In this section, we consider a regime, parameterized by r, where the transmission duration is much larger than that of a vacation phase. Technically, we impose that $\mathrm{E}[U_i]$ scales together with the diffusion parameter n, and $\mathrm{E}[V_i]$ remains constant, i.e., for some random variable U_i',

$$\textbf{Regime II}(r)\textbf{:} \quad U_i^n = n^{(2-r)} U_i', \ \mathrm{E}[U_i'] = u, \ \mathrm{E}[V_i] = v, \tag{14}$$

for some $1 < r < 2$, and u and v are positive constants. Let $\mathrm{VAR}[U_i'] = (\sigma^u)^2$, and $\mathrm{VAR}[V_i] = (\sigma^v)^2$.

By the Renewal Theorem, for large n, the total number of phases (thus also vacation phases as well as transmission phases) at the diffusion-scale, is given by[5]: $T(n^2 t) \sim n^r t/u$.

[5]We abuse the notation and use $T(\cdot)$ to denote the random process $T(n^2 t)$ and T to denote time T.

Then, it is not hard to see (14) implies the following on the vacation process: for the same r in (14), there exists $0 < \bar{v} < \infty$, such that

$$\frac{V(n^2t)}{n^r} \Rightarrow \bar{v}t, \text{ where } \quad \bar{v} = v/u. \tag{15}$$

4.3.1 Heavy Traffic Diffusion Approximation

Next, we study the limiting workload process on the diffusion scale. Compared to **A1** for Regime I, the impact of vacation is directly taken into account in the heavy traffic condition for Regime II(r).

A2. As $n \to \infty$, for some $b > 0$,

$$n \times \eta \cdot \left(\lambda^n - \lambda(1 - \bar{v}n^{r-2})\right) \to -b, \tag{16}$$

where the term $\bar{v}n^{r-2}$ corresponds to the vacation per unit time in the n-th system, and $\{\lambda^n\}$ and $\lambda(1 - \bar{v}n^{r-2})$ satisfy the CRP condition.

Note that in **A2**, the quantity $\lambda(1 - \bar{v}n^{r-2})$ corresponds to the throughput region of the n-th system for large n. We emphasize that Regime II(r) is an approximation model for the cases where the network operates at lower complexity and the scheduling is updated less and less frequently, so as to achieve high throughput. However, the throughput is still governed by Theorem 1. We will elaborate this in the proof of Theorem 3.

We present the main result in Regime II(r), whose proof is relegated to the Appendix.

THEOREM 3. *In Regime II(r), under Assumption* **A2***, as* $n \to \infty$,

$$(\hat{w}^n, \hat{q}^n) \Rightarrow (w^\star, q^\star), \tag{17}$$

with

$$z^\star(t) = w^\star(0) - bt + \sum_{l \in L} \eta_l \sigma_l B_l(t) + \mu(C(t) - \bar{v}D(t)),$$

where $\{B_l(t) : l \in L\}$ *are independent standard Brownian motions, and* $C(t)$ *and* $D(t)$ *are some random processes for which, as* $n \to \infty$,

$$\frac{V(n^2t) - \bar{v}tn^r}{n} \Rightarrow C(t) \tag{18}$$

$$\frac{U(n^2t) - tn^2}{n^{(3-r)}} \Rightarrow D(t). \tag{19}$$

In order to quantify the delay performance in Regime II(r) (Subsection 4.3.3), it is essential to characterize $C(t)$ and $D(t)$ rigorously (Subsection 4.3.2). We first make two remarks.

1) The state-space collapse property is a cornerstone step for establishing the heavy traffic limiting processes. As is clear in the proof (in the appendix), the local-fluid limit is of critical importance to state-space collapse. Furthermore, to construct the local-fluid limit, each piece of fluid-scaled process should have enough realizations of transmission and vacation phases for FSLLN (Functional Law of Large Numbers) to be applicable, which is only possible when $n^{2-r} < n$, making it necessary to have $1 < r < 2$.
2) Note that (18) and (19) have similar structures, since $U(n^2t)$ includes n^r random realizations of transmission phases. Recall that $U_i = n^{(2-r)}U'_i$, and define $U'(t) \triangleq \frac{U(t)}{n^{(2-r)}}$. Then, we can rewrite (19) as:

$$\frac{U'(n^2t) - tn^r}{n} \Rightarrow D(t), \tag{20}$$

which takes a similar form to (18).

4.3.2 Characterizing $C(t)$ and $D(t)$

We now focus on quantifying $C(t)$. Similar results can be carried over to $D(t)$.

Observe that when $r = 2$, $C(t)$ boils down to a Brownian motion. Due to large transmission phases in Regime II(r), the process $V(n^2t)$ needs to be centered with $\bar{v}tn^r$ (see (15)) to characterize its variability. We will present that $C(t)$ differs, depending on the variance of V_i (U'_i for $D(t)$).

(1) The finite variance case.

In this case, we show that $C(t) = 0$, a.s. Specifically, note that $T(n^2t) = O(n^rt)$ and that $T(n^2t)/n^r \to \frac{t}{u}$ u.o.c.. Define

$$\begin{aligned} \text{A} &\triangleq \frac{\sum_{i=1}^{n^r \frac{T(n^2t)}{n^r}} V_i}{n^r} - v\frac{T(n^2t)}{n^r} \\ \text{B} &\triangleq n^{r-1}\left(v\frac{T(n^2t)}{n^r} - \bar{v}tn^r\right). \end{aligned} \tag{21}$$

Then, it is easy to see that

$$\frac{V(n^2t) - \bar{v}tn^r}{n} = n^{(r/2-1)}n^{r/2}A + n^{r-1}B.$$

Appealing to Random-Time Change Theorem [5], we conclude that $n^{r/2}A$ converges to a Brownian motion. Since $n^{(r/2-1)}$ is decreasing for $1 < r < 2$, $n^{(r/2-1)}n^{r/2}A$ vanishes. The term $n^{r-1}B$ also vanishes since $T(n^2t)/n^r \to \frac{t}{u}$.

(2) The infinite variance case.

By definition, infinite variance indicates that V_i is heavy-tailed. We have the following assumption[6].

A3. The complementary CDF of V_i satisfies the following:

$$\text{P}\{|V_i| > x\} \approx Kx^{-\alpha}, \quad \text{as} \quad t \to \infty,$$

for some $0 < \alpha < 2$, and some constant K.

Next, we show that the limiting process $C(t)$ hinges on the relationship between α and r. We first state the following intuitions, using FCLT (Functional Central Limit Theorem) for partial sums of heavy-tailed i.i.d. random variables [27]:

$$\frac{V(n^2t) - \bar{v}tn^r}{n^{r/\alpha}} \Rightarrow L^\alpha(t/u). \tag{22}$$

Recall that for the finite variance case, it is enough to space-scale the system by $n^{r/2}$ for the convergence to a Brownian motion. A key observation of (22) is that for the partial sums with heavy-tailed random variables, we need to space-scale the process with *higher order* (depending on the constant α) for the convergence to a limiting process.

To treat formally, we first re-arrange (21) by:

$$\frac{V(n^2t) - \bar{v}tn^r}{n} = n^{(\frac{r}{\alpha}-1)}n^{r(1-\frac{1}{\alpha})}A + n^{r-1}B.$$

[6]To be more precise, we need more rigorous mathematical description that the complementary CDF of V_i belongs to *the normal domain of attraction*. However, we omit its details for brevity, and describe its key behavior in the assumption. We refer the readers to [27] for details.

Note that $n^{r-1}B$ vanished similar to the finite variance case. Next, again from Random-Time-Change Theorem [5], the term of "$n^{r(1-\frac{1}{\alpha})}A$" converges to a α-stable Levy motion, as n tends to ∞, as summarized in (22). In particular, since $T(n^2t)/n^r \to \frac{t}{u}$ from FSLLN, we get

$$n^{r(1-\frac{1}{\alpha})}A \Rightarrow L^{\alpha}(t/u), \tag{23}$$

where $L^{\alpha}(t)$ is α-stable Levy motion.

We consider the following two sub-cases:

(i) $1 < r < \alpha$. In this case, we observe that $n^{(\frac{r}{\alpha}-1)}$ tends to 0 as $n \to \infty$, indicating that $C(t) = 0$.

(ii) $r = \alpha$. In this case, $n^{(\frac{r}{\alpha}-1)} = 0$, and it follows that $C(t)$ is a α-stable Levy motion.

(iii) $r > \alpha$. In this case, $n^{(\frac{r}{\alpha}-1)}$ is increasing in n, i.e., there does not exist a converging process $C(t)$. In other words, the model here is not appropriate to study this case, and different methods are needed.

We conclude our characterization in this subsection with the following remark on the limiting processes:

1) When the sample paths of converging stochastic processes (e.g., Brownian motion in Regime I, or $C(t) = 0$ in some cases of Regime II(r)) have continuous sample paths, it is sufficient to consider a standard Skorohod J_1 topology to discuss all notions of convergence.
2) In the case when $r = \alpha$ of Regime II(r), stable Levy motion does not have continuous sample path, i.e., there are jumps in the sample path. We need to use a weaker topology, called Skorohod M_1 topology. Another technical difficulty we overcame in Regime II(r) is such an enlargement of topological spaces to establish the convergence.

4.3.3 Engineering Implications

We now discuss engineering implications based on Theorem 3. Different from Regime I, except for the special case of $C(t) = D(t) = 0$, the limiting workload process now does not admit a closed form. Fortunately, it is possible to study the asymptotic tail behavior based on the stationary distribution of α-stable Levy motion, based on which we next study study the *impact of transmission durations on the tail behavior of workload process.*

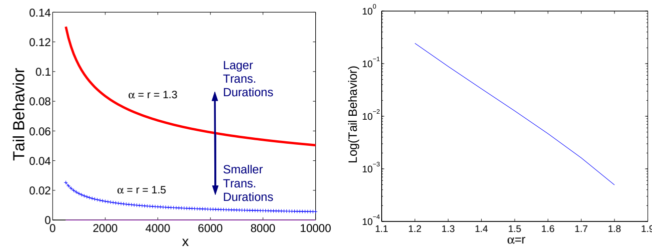
(1) $C(t) = D(t) = 0$.

In this case, it is clear that $w^\star$ is $\mathrm{RBM}_{w^*}(-b, \sum_{l\in L}(\eta_l\sigma_l^a)^2)$, where its average total workload is simply $\sum_{l\in L}(\eta_l\sigma_l^a)^2)/2b$. That is, when the transmission duration is large and the vacation duration has a "light" tail, the delay performance is affected only by the heavy traffic condition and the variability of the arrival process. Thus, this regime is not appropriate for investigating different delay properties depending on average transmission and vacations durations as well as their variability.

(2) $C(t) = L^{\alpha}(t/u)$ and/or $D(t) = L^{\alpha}(t/u)$.

In contrast to the earlier case when $C(t)$ and $D(t)$ vanish, this case reveals much more interesting implications that larger transmission durations leads to larger delays, which is intuitive since schedules are less frequently updated for larger transmission durations.

In this case, Theorem 3 reveals that $w^\star(t)$ is the reflection mapped process of the sum of a Brownian motion and a stable Levy motion, where a closed form solution for the stationary distribution is often unknown. Fortunately, the tail behavior of its stationary distribution offers an attractive



(a) (b)

Figure 3: (a) Impact of the order of transmission durations on tail behavior. (b) Power-law of tail behavior.

structure. Specifically, we study the *tail probability* of the workload process $w^\star(t)$, i.e., for a given $x \geq 0$,

$$\lim_{t\to\infty} \mathrm{P}\{w^\star(t) \geq x\}. \tag{24}$$

The stationary distribution corresponding to a Brownian motion has a light tail, i.e., its tail probability decays exponentially. In sharp contrast, the stationary distribution that corresponds to a stable Levy motion has a "heavy" tail, its tail probability decays in a power law manner. Therefore, for large x, the Levy motion part dominates in the large delay regime.

It is known that when the drift (i.e., $-b$) is negative, there exists a stationary distribution, and the tail asymptotics for the stationary distribution of some process $Z(t) = x_0 + \theta t + L^{\alpha}(t)$ follows (where x_0 is the starting point and $\theta < 0$ is the drift in standard notation [27]): for large x,

$$\lim_{t\to\infty} \mathrm{P}\{Z(t) \geq x\} = H(x),$$

where

$$H(x) = \begin{cases} \frac{1}{\Gamma(2-\alpha)x^{\alpha-1}} - \frac{1}{\Gamma(3-2\alpha)x^{2(\alpha-1)}}, & 1 < \alpha < 2, \alpha \neq \frac{3}{2}, \\ \frac{1}{\sqrt{\pi}x^{1/2}} - \frac{1}{2\sqrt{\pi}x^{3/2}}, & \alpha = \frac{3}{2}. \end{cases}$$

In the special case when $\alpha = 3/2$, $H(x) = 2e^x\Psi^c(\sqrt{2x})$, where Ψ^c is the standard normal complementary CDF.

As a numerical illustration, consider the case when $u = 1$, and an arrival vector and a network, such that $\mu = 1$ (recall that $\mu = \eta \cdot \lambda$ and η is determined by the considered network and the arrival vector λ). Figure 3(a) shows the tail behavior for $1000 < x < 10000$ and two values of $\alpha = 1.3, 1.5$. It can be seen that the tail probability follows a power-law and decays slowly. Note that larger r leads to a smaller transmission duration which in turn results in a "lighter" tail behavior, since the system is updated with new schedules more frequently.

Then, a natural question would be how the tail behavior generally decays with r. To get a more concrete sense, in Figure 3(b), we fix $x = 2000$, and vary α (equal to r), ranging from 1.2 to 1.8. This example examines the impact of r (the order of transmission duration) on the delay performance, measured by the tail probability. For a large fixed n, note that the increase of order r leads to the exponential decrease of transmission durations, since $\mathrm{E}[U_i] = n^{(2-r)}u$. Observe that the increase of r results in an exponential decrease in the tail probability (y-axis is log-scaled), indicating that the tail probability scales *linearly* with transmission durations in this regime. This was also observed in Regime I, where we extended the earlier results based on Lyapunov bounds, under a regime with large transmission durations, for arbitrary network topologies.

In a nutshell, in the case with $C(t) = L^{\alpha}(t/u)$ and/or $D(t) = L^{\alpha}(t/u)$, the Levy motion part in the limiting workload process dominates in the large delay regime, and the tail probability of the stationary distribution of the delay decays in a power law manner. Accordingly, the average delay can be significant and even grow unbounded.

5. CONCLUDING REMARKS

The main thrust of this paper is devoted to quantifying the impact of signaling complexity on delay and throughput performance of wireless scheduling. Complementary to previous studies, we advocate the approach of heavy traffic analysis to shed new light on traditionally challenging issues in this research area. We model the signaling complexity as "vacations" and data transmissions as "services," and characterize the effective throughput and workload/queueing processes in heavy traffic regimes. Such heavy traffic analysis focuses on the cases with bottleneck links and is used to prove the desirable state space collapse property, which is substantially simpler than the original model while providing excellent approximation. The heavy traffic model with vacations is applicable to a general family of max-weight based scheduling rule, namely GMW, while explicitly considering signaling complexity.

In particular, we analyze the delay performance in various regimes of vacation models which hinges on the durations of vacation and data transmission. For Regime I where both vacation and transmission durations are of $O(1)$, we prove that the workload process at the diffusion scale converges to a reflected Brownian motion. We also study the impact of growing network sizes on average workload. For Regime II(r) where vacation duration is of $O(1)$ and transmission duration is of order n^{2-r} with $1 < r < 2$, we discover that the diffusion-scaled workload process converges to a weighted mixture of a reflected Brownian motion and a reflected α-stable Levy motion. In this regime, the analysis reveals that the average delay can be significantly large and even grow unbounded.

This paper aims at making initial steps towards heavy traffic delay analysis of wireless scheduling. We have focused on the case where the signaling complexity is finite, corresponding to a fixed network size. A next step is to study the model where that network size $L = L(n)$, vacation and transmission durations scale jointly with the diffusion scale n. This regime is challenging due to the fact that the dimensionality of queue vector process $Q(t) = (Q_l(t))$ also grows. A key step here will again be the characterization of the conditions under which state space collapse follows.

Appendix

Proofs of Theorem 2 and Theorem 3

We outline the main steps for proving Theorem 3. The proof of Theorem 2 follows the same line and is simpler indeed, with differences described at the end.

Step 1: We first prove that $\mu = \eta \cdot \lambda$. Let $S' \in \arg\max_{S\in\mathcal{S}} \eta \cdot S$. Then, we have $S' \in \arg\max_{a\in\partial\Lambda} \eta \cdot a$, since any $a \in \partial\Lambda$ is a convex combinations of $S \in \mathcal{S}$. Now, from the CRP condition of λ, $\lambda = \arg\max_{a\in\Lambda} \eta \cdot a$, implying that $\mu = \eta \cdot \lambda$.

Step 2: We prove that $\hat{z}^n(t) \Rightarrow z^\star(t)$.

Denote by $P^n(t)$ be the amount of workload that could be served by slot t at the n-th system. When $n \to \infty$, $P^n(t)/U^n(t) \to \mu$.) Let $Y^n(t) \triangleq P^n(t) - \eta \cdot G^n(t)$, which is the amount of workload service "wasted" by time t during the transmission phase. Let $Z^n(t) \triangleq W^n(t) - Y^n(t)$.

Then, using (8) and $V^n(t)+U^n(t) = t$, and adding/subtracting $\eta \cdot \lambda^n t$, we get:

$$\begin{aligned} Z^n(t) &= W^n(0) + \eta \cdot F^n(t) - P^n(t) \\ &= W^n(0) + \eta \cdot (F^n(t) - \lambda^n t) + \eta \cdot \lambda^n t - \mu U^n(t) \\ &= W^n(0) + \eta \cdot (F^n(t) - \lambda^n t) + \eta \cdot \lambda^n t \\ &\quad -\mu t + \mu V^n(t) \\ &= W^n(0) + \eta \cdot (F^n(t) - \lambda^n t) + \eta \cdot \lambda^n t - \eta \cdot \lambda t \\ &\quad +\mu V^n(t), \end{aligned}$$

where in the last inequality, we use $\mu = \eta \cdot \lambda$.

Taking the diffusion scaling yields that

$$\hat{z}^n(t) = \hat{w}^n(0) + \eta \cdot (\hat{f}^n(t) - \lambda^n nt) + n\eta(\lambda^n - \lambda)t + \mu\hat{v}^n(t). \quad (25)$$

A key observation is that $T(t)$ is a random process, where $T(t) = \max\{k \geq 0 \mid \sum_{i=1}^{k}(V_i + U_i) \leq t\}$. Observe that

$$\hat{v}^n(t) = V(n^2t)/n = \frac{1}{n}\sum_{i=1}^{T(n^2t)}(V_i - v) + \frac{1}{n}T(n^2t)v,$$

which can be expressed as the sum of the following four terms:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{T(n^2t)}(V_i - v)}_{(a)} - \underbrace{\frac{v}{n^{(2-r)}u+v}\frac{1}{n}\sum_{i=1}^{T(n^2t)}(U_i - n^{(2-r)}u)}_{(b)} - \underbrace{\frac{v}{n^{(2-r)}u+v}\frac{1}{n}\sum_{i=1}^{T(n^2t)}(V_i - v)}_{(c)} + \underbrace{\frac{v}{n^{(2-r)}u+v}\frac{1}{n}\sum_{i=1}^{T(n^2t)}(V_i + U_i)}_{(d)}. \quad (26)$$

Accordingly, in the diffusion scale limit, we have that

- From (18), $(a) \Rightarrow C(t)$.
- $(b) \Rightarrow \bar{v}D(t)$, from (19) and (20).
- (c) vanishes, since $(c) \sim \frac{1}{n^{(2-r)}}C(t)$, where $1 < r < 2$.
- $(d) \Rightarrow \bar{v}n^{(r-1)}t$, since $\sum_{i=1}^{T(n^2t)}(V_i + U_i) = n^2t$.

Then, plugging (26) into (25), and based on **A2** in (16), we conclude that $z^n(t) \Rightarrow z^\star(t)$. Recall that $z^\star(t)$ can be either a Brownian motion or stable Levy motion, depending on the tail behavior of V_i. In particular, we need to introduce Skorohod M_1 topology to establish the convergence to stable Levy motions, in order to handle the jumps in the limiting processes.

It remains to prove the state-space collapse property, which we turn our attention to next.

Step 3: The state-space collapse (SSC) property is formally defined as follows: for each $T \geq 0$,

$$||\hat{q}^n(t) - \xi \times \hat{w}^n(t)||_T \to 0, \quad \text{in probability}, \quad (27)$$

where $||\cdot||_T$ is the uniform norm over the time interval $[0, T]$. To prove SSC, we takes the following sub-steps:

(s1) *SSC for fluid-scaled processes.* We first prove the state-space collapse property for the fluid-scaled process $\tilde{q}(t)$, i.e., there exists a slot $t_0 < \infty$, such that

$$|\tilde{q}^n(t) - \xi \times \tilde{w}^n(t)| = 0, \quad \text{for all } t \geq t_0. \quad (28)$$

(s2) For a fixed $T > 0$, we divide n^2T slots into $[nT]+1$ intervals of length n, and view a diffusion-scaled process $\{\hat{q}^n(t) = \frac{Q^n(n^2t)}{n}, t \in [0,T]\}$ as a concatenation of $[nT]+1$ fluid scaled processes, where j-th piece of fluid scaled process starts at slot nj and ends at slot $n(j+1)$.

(s3) *Local fluid-limit.* As $n \to \infty$, each of these fluid-scaled processes over a compact interval with length n converge to a deterministic fluid solution which enjoys the fluid-level SSC in (28).

(s4) Then, at every point $t \in [0,T]$, we approximate $\hat{q}^n(t)$ by one piece of fluid-model solution that satisfies SSC, leads to SSC at the diffusion-scaled system.

Step 4: Finally, let $\hat{w}^n = \Phi(\hat{z}^n)$. Then, from the continuous mapping theorem, we have: as $n \to \infty$ and the result at Step 3 that $\hat{z}^n \Rightarrow z^\star$, we have $\hat{w}^n \Rightarrow w^\star$. Using all the above results, the theorem follows.

Modifications for proving Theorem 2.

The proof of Theorem 2 involves only a few minor modifications: (i) set $r = \alpha = 2$, and (ii) $\bar{\mu} = u/(u+v)\mu = \eta \cdot \lambda$, where $\lambda \in u/(u+v)\Lambda$. □

6. REFERENCES

[1] E. Altman and H. J. Kushner. Control of polling in presence of vacations in heavy traffic with applications to satellite and mobile radio systems. *SIAM Journal on Control and Optimization*, 41:217–252, 2002.

[2] M. Bramson. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Syst.*, 30(1-2):89–148, 1998.

[3] P. Chaporkar, K. Kar, and S. Sarkar. Throughput guarantees through maximal scheduling in wireless networks. In *Proceedings of Allerton*, 2005.

[4] H. Chen and W. Whitt. Diffusion approximations for open queueing networks with service interruptions. *Queueing Systems*, 13:335–359, 1993.

[5] H. Chen and D. Yao. *Fundamentals of Queueing Networks*. Springer-Verlag, 2001.

[6] J. G. Dai. On positive harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability*, 5:49–77, 1995.

[7] J. G. Dai and W. Lin. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Annals of Applied Probability*, 18(6):2239–2299, 2008.

[8] A. Eryilmaz, A. Ozdaglar, and E. Modiano. Polynomial complexity algorithms for full utilization of multi-hop wireless networks. In *Proceedings of IEEE Infocom*, 2007.

[9] G. R. Gupta and N. Shroff. Delay analysis of multi-hop wireless networks. In *Proceedings of IEEE Infocom*, 2009.

[10] C. Joo and N. B. Shroff. Performance of random access scheduling schemes in multi-hop wireless networks. In *Proceedings of IEEE Infocom*, 2007.

[11] K. Jung and D. Shah. Low delay scheduling in wireless network. In *Proceeding of ISIT*, 2007.

[12] O. Kella and W. Whitt. Diffusion approximations for queues with server vacations. *Advances in Applied Probability*, 22:706–729, 1990.

[13] X. Lin and S. Rasool. Constant-time distributed scheduling policies for ad hoc wireless networks. In *Proceedings of IEEE CDC*, 2006.

[14] X. Lin and N. B. Shroff. The impact of imperfect scheduling on cross-layer rate control in wireless networks. In *Proceedings of IEEE Infocom*, 2005.

[15] E. Modiano, D. Shah, and G. Zussman. Maximizing throughput in wireless networks via gossiping. In *Proceedings of ACM Sigmetrics*, 2006.

[16] M. J. Neely, E. Modiano, and C. E. Rohrs. Tradeoffs in delay guarantees and computation complexity for nxn packet switches. In *Proceedings of CISS*, 2002.

[17] S. Ray and S. Sarkar. Arbitrary throughput versus complexity tradeoffs in wireless networks using graph partitioning. In *Proceedings of Information Theory and Applications Second Workshop*, 2007.

[18] S. Sanghavi, L. Bui, and R. Srikant. Distributed link scheduling with constant overhead. In *Proceedings of ACM Sigmetrics*, 2007.

[19] D. Shah and D. J. Wischik. Optimal scheduling algorithms for input-queued switches. In *Proceedings of IEEE Infocom*, 2006.

[20] S. Shakkottai. Effective capacity and qos for wireless scheduling. *IEEE Transactions on Automatic Control*, 53(3):749–761, 2008.

[21] S. Shakkottai, R. Srikant, and A. Stolyar. Pathwise optimality of the exponential scheduling rule for wireless channels. *Advances in Applied Probability*, 36(4):1021–1045, 2004.

[22] A. L. Stolyar. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Annals in Applied Probability*, 14(1):1–53, 2004.

[23] A. L. Stolyar. Large deviations of queues under qos scheduling algorithms. In *Proceedings of Allerton*, 2006.

[24] L. Tassiulas. Linear complexity algorithms for maximum throughput in radionetworks and input queued switches. In *Proceedings of IEEE Infocom*, 1998.

[25] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37(12):1936–1949, December 1992.

[26] V. J. Venkataramanan and X. Lin. Structural properties of ldp for queue-length based wireless scheduling algorithms. In *Proceedings of Allerton*, 2007.

[27] W. Whitt. *Stochastic-Process Limits*. Springer-Verlag, 2001.

[28] R. J. Williams. An invariance principle for semimartingale reflecting brownian motions in an orthant. *Queueing Syst. Theory Appl.*, 30(1-2):5–25, 1998.

[29] X. Wu and R. Srikant. Bounds on the capacity region of multi-hop wireless networks under distributed greedy scheduling. In *Proceedings of IEEE Infocom*, 2006.

[30] Y. Yi, A. Proutiere, and M. Chiang. Complexity in wireless scheduling: Impact and tradeoffs. In *Proceedings of ACM Mobihoc*, 2008.

[31] L. Ying, R. Srikant, A. Eryilmaz, and G. E. Dullerud. A large deviations analysis of scheduling in wireless networks. *IEEE Transactions on Information Theory*, 52(11):5088–5098, 2006.