Available online at www.sciencedirect.com
ScienceDirect
Performance Evaluation 64 (2007) 892–908
PERFORMANCE EVALUATION
An International Journal
www.elsevier.com/locate/peva

On processor sharing and its applications to cellular data network provisioning

Yujing Wu*, Carey Williamson, Jingxiang Luo

Department of Computer Science, University of Calgary, Calgary, AB, Canada T2N 1N4

Available online 27 June 2007

Abstract

To develop simple traffic engineering rules for the downlink of a cellular system using Proportional Fairness (PF) scheduling, we study the "strict" and "approximate" insensitivity of a Processor Sharing (PS) system, specifically for the Egalitarian (EPS) and Discriminatory (DPS) variants of PS. Assuming homogeneous channel conditions, all concurrent flows are allocated an equal share of downlink transmission slots regardless of flow types and locations. The cell system is modeled as an EPS queue. We prove the performance insensitivity of EPS in a relevant new case that has not been studied in the literature. Considering heterogenous channel conditions, the system is modeled as the DPS queue in which each traffic type is divided into subclasses with different assigned weights. Asymmetric weights among the subclasses model the unequal channel sharing that occurs with PF scheduling. Our results show that the first-order performance of the DPS is largely insensitive to the input traffic characteristics, as long as the weights among subclasses are not highly skewed. Our findings, confirmed by the simulation of a cellular system, imply reduced complexity for traffic provisioning procedures. However, our study also shows that the first-order performance is sensitive to the traffic details when there is discrimination among different traffic types. This observation implies that the introduction of differentiated services may pose a great challenge to network provisioning in future cellular systems.

Keywords: Cellular data network provisioning; Proportional fairness; Heterogeneous channels; Processor sharing; Approximate insensitivity

1. Introduction

Insensitivity is a desirable property in networks and systems. This property means that the system performance does not depend on the specific details of the input traffic offered to the system. A classic example is the Erlang B call blocking formula, for which only the traffic intensity is needed; the precise distribution of the call holding times is irrelevant. The presence of the insensitivity property facilitates robust approaches to traffic engineering. The engineering decisions made will be effective for a broad range of input traffic conditions (e.g., different flow size distributions, or session structures).

This paper studies the "strict" and "approximate" insensitivity properties of Processor Sharing (PS) systems. The primary motivation for this study arises from the need to develop simple traffic engineering rules for the downlink (from the base station to the mobile user) in 3G cellular systems. We are most interested in the impacts of traffic

* Corresponding address: QuIC Financial Technologies Inc., 3553-31 St. NW, T2L 2K7 Calgary, AB, Canada. Tel.: +1 403 5324254.
E-mail addresses: yujingwu@gmail.com (Y. Wu), carey@cpsc.ucalgary.ca (C. Williamson), jxluo@cpsc.ucalgary.ca (J. Luo).

0166-5316/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.peva.2007.06.011

characteristics on the first-order flow-level performance metrics, specifically, the mean number of active flows, blocking probability, response time, and throughput.

Egalitarian Processor Sharing (EPS) has been used to evaluate the flow-level performance of cellular data systems using Proportional Fairness (PF) scheduling [9]. It is a popular model in the study of bandwidth sharing on the Internet [8,14]. Some researchers also use it in the performance analysis of wireless LANs [18]. In an EPS queue, the server's capacity[1] is shared equally among all flows concurrently in service. Assuming Poisson flow arrivals, the EPS model has simple, explicit expressions for the distribution of the number of active flows in steady state, and the first-order flow-level performance metrics. These measures are insensitive to the flow size distribution. For an infinite-capacity EPS system, Bonald et al. [6,8,14] extend the insensitivity results to the case of Poisson session arrivals. In this paper, we further extend the analysis to a *finite-capacity* EPS system fed by Poisson session traffic.

Strictly speaking, EPS is applicable only if resources (e.g., bandwidth, time slots) are shared in a perfectly fair way. This may not be the case in real systems. For example, TCP bandwidth sharing on the Internet actually depends on the round-trip times of the flows. In a cellular system employing PF scheduling, flows with larger rate fluctuations around the long-term transmission rate tend to receive a smaller fraction of the transmission time [15]. These observations suggest that Discriminatory Processor Sharing (DPS) may be more appropriate than EPS for modeling real systems.

In a DPS system, all concurrent flows are served simultaneously at rates controlled by a weight vector $\vec{\mathbf{W}} := [w_1, \ldots, w_K]$, where K denotes the number of classes. When there are n_k active flows of class k, each flow in class j receives service at the rate $w_j / \sum_{k=1}^{K} w_k n_k$. If all weights w_k are equal, then DPS reduces to EPS. Throughout this paper, we use the term Processor Sharing (PS) to refer to the general scheduling principle that includes both EPS and DPS.

Rigorously speaking, the performance of DPS is sensitive to specific traffic characteristics [7]. However, several prior studies mention certain insensitivity properties of DPS. Bonald and Proutiere [7] derive bounds on the queue length distributions, and identify the limiting distributions obtained when nodes of PS networks operate at very different time scales. The bounds and the limiting distributions are shown to be insensitive to the service time (flow size) distributions. Work by van Kessel et al. [21] develops an approximation for the queue length distribution based on the time-scale decomposition described in [7]. Their numerical study reveals that the performance of a traffic class with relatively slow dynamics is mostly insensitive to the priority weights, and also insensitive to the service time distribution.

The existence of these insensitive bounds and limiting approximations naturally leads to the conjecture that the insensitivity properties of EPS systems may *approximately* carry over to DPS for a certain range of parameter choices. So far, the investigation of such "approximate insensitivity" is limited. In the context of statistical bandwidth sharing, several papers [6,14,19,21] conjecture about approximate insensitivity. However, no analytical study or extensive simulation investigation is given to support the statements there.

The purpose of this paper is to explore these (in)sensitivities in the context of 3G system provisioning. Our study makes two main contributions. First, the theoretical analysis of the EPS system and the extensive simulation investigation of the DPS system improve the understanding of processor sharing. Second, we apply the findings to a practical CDMA 1xEV-DO system model, and provide useful insights into network provisioning. Our results demonstrate the *practical insensitivity* of DPS for the purposes of traffic engineering in cellular data networks, while highlighting sensitivity issues for networks that support differentiated services.

The rest of the paper is organized as follows. The next section introduces flow-level modeling in cellular systems. Section 3 summarizes prior theoretical analysis of PS. Section 4 derives analytical results for a finite-capacity EPS. Section 5 presents an extensive DPS simulation study. The approximate insensitivity identified in this section is confirmed by the detailed simulation of an EV-DO cellular system in Section 6. Section 7 presents results when considering differentiated services. Section 8 concludes the paper.

2. Modeling cellular data networks

In high data rate 3G cellular systems, such as CDMA 1xEV-DO, a base station (BS) transmits elastic flows to mobile users[2] in a time-slotted fashion. The scheduler chooses one user for transmission at each time slot. The

[1] Server capacity may mean transmission slots, bandwidth, or CPU time, depending on the system being studied.

[2] For simplicity, we assume a one-to-one mapping between flows and users, and thus use these terms interchangeably throughout the paper.

scheduling algorithm greatly influences cellular system capacity and the flow-level performance perceived by users. A form of Proportional Fairness (PF) scheduling is typically implemented in commercial products. Our study focuses on the downlink in a cellular system that uses PF scheduling to transmit elastic data flows to mobile users.

The PF scheduler seeks to optimize the overall cell throughput, while at the same time maintaining a certain level of fairness among different users over longer time scales. In time slot t, flow j (destined to user j) is chosen for transmission if

$$j = \arg\max_i \left[\frac{r_i(t)}{T_i(t)} \right], \tag{1}$$

where $r_i(t)$ is the feasible rate of flow i at time t, and $T_i(t)$ is the average transmission rate realized by flow i over the past time window of length t_c. Specifically, $T_i(t)$ is calculated as

$$T_i(t+1) = \left(1 - \frac{1}{t_c}\right) T_i(t) + \frac{1}{t_c} r_i(t) \mathbf{I}_{\{\text{flow } i \text{ is scheduled at } t\}}. \tag{2}$$

The slot duration (1.667 ms in EV-DO) is very short compared to flow durations and inter-arrival times. In terms of flow-level performance, the downlink of the cellular system behaves like a PS queue, which operates in continuous time and serves all active flows concurrently.

With different assumptions about rate variations, the system can be abstracted to different PS models. Let $R_i := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r_i(t)$ be the mean transmission rate of flow i in the absence of any other flows. In the homogeneous rate variation scenario, the fluctuations in $r_i(t)$ around the mean R_i for all active flows are statistically identical. That is, $\forall i$, $\frac{r_i(t)}{R_i} = Y_i(t)$, where the $Y_i(t)$'s are independently and identically distributed for different flows and for different time slots. Based on Eqs. (1) and (2), all active flows are scheduled with equal probability at each time slot, and receive equal transmission allocations over long time scales. Such a cellular system works like an EPS system, for which known insensitivity results can be applied.

Homogeneous rate variation is an idealized situation. In practice, heterogeneous scenarios are more likely. The SINR (Signal to Interference-and-Noise Ratio) perceived by a user depends on complicated interactions among radio signals propagating to the user from the home BS and from interfering base stations.

The SINR values perceived by different users typically have different statistics. In addition, the relationship between the feasible rate and SINR is not linear, especially for high data rates [9]. Thus, the Y_i values for different users generally have different distributions, which causes unequal sharing of time slots over longer time scales.

As observed in [15], PF allocates more time to users with lower variability in the feasible rate. Such users are usually close to the home BS. When users are at the periphery of the home cell, the feasible rate shows higher variability, and the flows receive a smaller transmission share. Given that the PF scheduler is biased against flows with weak SINR values, DPS is probably a better abstraction of this system.

To evaluate (in)sensitivity to traffic characteristics, we consider two traffic models: Poisson flow arrivals and Poisson session arrivals. In the first model, flows of each type arrive as an independent Poisson process with generally distributed flow sizes. In the second (more general) model, flows are generated within sessions, for which the session starting times constitute a Poisson process. Each session contains multiple flow transmissions, each separated by a thinking time. The session structure, including the distribution of the number of flows per session, flow size distribution, thinking time distribution, and any correlation in successive flow and thinking time statistics, is general. The Poisson session model provides a more realistic traffic model.

For network provisioning, we are most interested in the effect on the first-order performance. In the homogeneous rate variation scenario (idealized situation), EPS applies. From the literature, the performance metrics are insensitive to traffic details in a few cases. We extend such insensitivity to a relevant new case. In the heterogeneous scenario, DPS applies. We ask "How pronounced are the performance sensitivities?" and "Do they manifest themselves in the relevant regime for practical parameter settings?".

3. Processor Sharing (PS) systems

We recap relevant theoretical analysis from the literature for PS systems, which provide the starting point for our study. The PS queue receives offered traffic from K distinct flow classes. In the finite-capacity case, it admits at most B flows at a time. Flows of class k are generated as a stationary point process with rate λ_k, and have generally

distributed flow sizes with mean β_k. Define $\lambda := \sum_{k=1}^{K} \lambda_k$ as the aggregate arrival rate, $\rho_k := \lambda_k \beta_k$ as the traffic intensity of class k, and $\rho := \sum_{k=1}^{K} \rho_k$ as the overall traffic intensity. It is always assumed that the system operates in the stable regime, i.e., $\rho < 1$. Let N denote the number of active flows at an arbitrary epoch in stochastic equilibrium, and let N_k denote the number of active class k flows at that time. Denote by P_b the aggregate blocking probability when admission control is used.

3.1. EPS results

EPS systems have elegant closed-form expressions for performance metrics in several cases.

Poisson flow arrivals. The joint queue length distribution and blocking probability are given by [11,16]:

$$P\{N_k = n_k, k = 1, \ldots, K\} = \frac{n!(1-\rho)}{1-\rho^{B+1}} \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{n_k!}, \tag{3}$$

where $n = \sum_{k=1}^{K} n_k \leq M$ and

$$P_b = P\{N = B\} = \frac{(1-\rho)\rho^B}{1-\rho^{B+1}}. \tag{4}$$

These measures do not depend on traffic characteristics other than the intensities.

Poisson session arrivals. For an infinite-capacity system, Bonald et al. [6,8,14] generalize the above results to the case with Poisson session traffic by using the classic stochastic network theory [11,16]. As long as session arrivals are Poisson, the joint distribution of flows in progress is the same as that in the case of Poisson flow arrivals, regardless of the session structure. This directly establishes the insensitivity of the first-order performance measures $E[N]$ and $E[N_k]$. For an EPS system with admission control, Borst [9] mentions that the previous analytical results should apply. However, to the best of our knowledge, no one has proved it.

3.2. DPS results

In moving from EPS systems to DPS systems, the insensitivity property and much of the modeling tractability disappear. So far, theoretical studies are limited to Poisson flow arrivals. We have not seen any study dealing with Poisson session arrivals. Even if exact analysis is possible, closed-form expressions exist in few cases.

Infinite capacity. Consider Poisson flow arrivals with exponentially distributed sizes. Let R_k denote the mean flow response time for class k. Based on *Lemma* 3 of [13], the average response times satisfy :

$$R_k \left[1 - \sum_{j=1}^{K} \frac{\lambda_j w_j}{w_j/\beta_j + w_k/\beta_k} \right] - \sum_{j=1}^{K} \frac{\lambda_j w_j R_j}{w_j/\beta_j + w_k/\beta_k} = \beta_k, \tag{5}$$

for $k = 1, 2, \ldots, K$. Closed-form expressions are available for the two-class case. The mean number of active flows in steady state can be readily obtained using Little's Law.

The work by van Kessel et al. [21] extends the foregoing results to flow sizes with phase-type distributions. The mean queue lengths can be obtained numerically, with details given by Theorem 1, 2, and 3 in [21]. For generally distributed flow sizes, performance depends on the distributions of flow sizes for all classes [13]. To date, there is little progress on exact analysis for this general case. We refer interested readers to [2] for recent progress in the analysis of infinite-capacity DPS systems.

Finite capacity. The literature on finite DPS analysis is sparse. One recent paper [10] provides exact analysis of the sojourn time distribution. Flows arrive as a Poisson process, and have exponentially distributed sizes (possibly extensible to phase-type distributions). When the number of classes K or the system capacity B is large, the numerical method becomes computationally onerous. Because of the difficulties associated with exact analysis, several studies [20,21] have applied asymptotic analysis to develop approximations and bounds for DPS.

4. Theoretical study of a finite-capacity EPS system

Assuming homogenous rate variation, a cellular system can be modeled as an EPS queue. As mentioned in Section 3.1, the performance insensitivity has been proved for the EPS queue fed by Poisson flow traffic, and for the infinite-capacity one fed by Poisson session traffic. In this section, we extend the insensitivity to the finite-capacity one with Poisson session traffic. We follow the approach in [5], where the insensitivity of the Erlang loss model is extended to the case with Poisson session traffic. In that paper, a queueing network with infinite-server nodes is used to prove the insensitivity. We consider a queueing network with generalized nodes, which leads to the analysis of the finite- capacity EPS system.

4.1. A queueing network

We start from an infinite-capacity network, and then address the finite-capacity network via restriction of the state space. Consider a queueing network of J nodes each with infinite capacity. Let $X(t)$ denote the J-dimensional vector whose ith element is the number of customers at node i. Customers arrive at node i as a Poisson process with rate ν_i. If the network is in state x, the time to the next departure from node i is exponentially distributed with rate $\phi_i(x)$. The rate is positive except for the case $\phi_i(x) = 0$ if $x_i = 0$. After being served at node i, customers move to node j with probability p_{ij}, and leave the network with probability $p_i = 1 - \sum_{j=1}^{J} p_{ij}$. Denote by λ_i the steady-state arrival rate at node i, including arrivals from other nodes. These arrival rates satisfy the traffic equations: $\lambda_i = \nu_i + \sum_{j=1}^{J} \lambda_j p_{ji}$, $i = 1, \ldots, J$.

Now consider a similar network but with a restricted state space. A set of admissible states is given as $\mathcal{A} \subset \mathbb{N}^J$. It is assumed that $\mathcal{A}$ is coordinate convex in the sense that if $x \in \mathcal{A}$ then $y \in \mathcal{A}$ for all y such that $y \leq x$ component-wise. Denote by e_i the J-dimensional vector whose ith element is 1 and all other elements are zero. In state x, customers arrive at node i as a Poisson process at the rate $\nu_i'(x)$, with $\nu_i'(x) = 0$ if $x + e_i \notin \mathcal{A}$. After being served at node i, customers move to node j with probability $p_{ij}'(x)$, where $p_{ij}'(x) = 0$ if $x - e_i + e_j \notin \mathcal{A}$. Customers leave the network with probability $p_i'(x) = 1 - \sum_{j=1}^{J} p_{ij}'(x)$. Let $\lambda_i'(x)$ be the arrival rate at node i in state x including arrivals from other nodes. The arrival rates satisfy the traffic equation (6) and traffic conservation equation (7):

$$\lambda_i'(x) = \nu_i'(x) + \sum_{j=1}^{J} \lambda_j'(x) p_{ji}'(x + e_j), \quad i = 1, \ldots, J \tag{6}$$

$$\sum_{i=1}^{J} \nu_i'(x) = \sum_{i=1}^{J} \lambda_i'(x) p_i'(x + e_i). \tag{7}$$

These equations are given by Bonald [5]. They are still applicable to our network with generalized nodes.

Theorem 4.1. *The stationary distribution π of the Markov process $X(t)$ satisfies*

$$\pi(x)\phi_i(x) = \pi(x - e_i)\lambda_i'(x - e_i), \quad \forall i \in \{1, \ldots, J\}, \ \forall x \in \mathbb{N}^J. \tag{8}$$

Proof. We show that Eq. (8) leads to the global balance equation. Substituting (6) into (8) yields

$$\begin{aligned}\pi(x)\phi_i(x) &= \pi(x - e_i)\nu_i'(x - e_i) + \sum_{j=1}^{J} \pi(x - e_i)\lambda_j'(x - e_i) p_{ji}'(x - e_i + e_j) \\ &= \pi(x - e_i)\nu_i'(x - e_i) + \sum_{j=1}^{J} \pi(x + e_j - e_i)\phi_j(x + e_j - e_i) p_{ji}'(x + e_j - e_i) \quad \text{[use (8)]}.\end{aligned}$$

Summing the above equation for $i = 1, 2, \ldots, J$, we obtain:

$$\pi(x) \sum_i \phi_i(x) = \sum_i \pi(x - e_i)\nu_i'(x - e_i) + \sum_i \sum_j \pi(x + e_j - e_i)\phi_j(x + e_j - e_i) p_{ji}'(x + e_j - e_i). \tag{9}$$

On the other hand, multiplying (7) by $\pi(x)$ and applying (8) yield:

$$\pi(x)\sum_i \nu_i'(x) = \sum_i \pi(x+e_i)\phi_i(x+e_i)p_i'(x+e_i). \tag{10}$$

The global balance equation can be obtained by adding (9) and (10). Note that the sum of the l.h.s. of (9) and (10) gives the transition rate out of the state x, while the sum of the r.h.s. gives the transition rate from all other states into state x. □

Define that node i is *free* if $x + e_i \in \mathcal{A}, \forall x \in \mathcal{A}$ (this definition is given in [5]).

Corollary 4.2. *Assume that, for all nodes i and all states $x \in \mathbb{N}^J$, the arrival rates satisfy*

$$\lambda_i'(x) = \begin{cases} \lambda_i, & \text{if } x + e_i \in \mathcal{A}, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Then customers leaving any free node see the network in steady state immediately after their departure (but before reaching the next node).

Proof. Denote by $\pi_i(x)$ the probability that customers leaving free node i see the network in state x immediately after their departure.

$$\pi_i(x) = \frac{\pi(x+e_i)\phi_i(x+e_i)}{\sum_{y\in\mathcal{A}} \pi(y+e_i)\phi_i(y+e_i)}. \tag{12}$$

Since $x + e_i \in \mathcal{A}$, $\pi(x+e_i)\phi_i(x+e_i) = \pi(x)\lambda_i$ by Theorem 4.1 and (11). Similarly, $\pi(y+e_i)\phi_i(y+e_i) = \pi(y)\lambda_i$. Substituting these two equations into (12), we obtain $\pi_i(x) = \pi(x)$. □

Consider a specific network with two types of nodes. Type 1 nodes are in set $\theta_1 = \{1, \ldots, \tilde{J}\}$ while type 2 nodes are in set $\theta_2 = \{\tilde{J}+1, \ldots, J\}$, where $\tilde{J} < J$. The service rates are:

$$\phi_i(x) = \begin{cases} \mu_i x_i / T(x), & i \in \theta_1, \\ \mu_i x_i, & i \in \theta_2, \end{cases} \tag{13}$$

where $T(x) = \sum_{i\in\theta_1} x_i$. The admissible states are $\mathcal{A} = \left\{x \in \mathbb{N}^J : T(x) \le B\right\}$. Define $\tilde{\rho}_i := \lambda_i/\mu_i$, $i = 1, \ldots, J$, and $\rho := \sum_{i\in\theta_1} \tilde{\rho}_i$. We partition θ_1 into K disjoint sets: $\theta_1 = \bigcup_{k=1}^K \alpha_k$. Denote by N the number of customers in set θ_1, and by N_k those in set α_k. Then, $N = \sum_{i\in\theta_1} x_i$ and $N_k = \sum_{i\in\alpha_k} x_i$, $k = 1, \ldots, K$.

Corollary 4.3. *Under the condition of Corollary 4.2, the stationary distribution of the Markovian process $X(t)$ is given as*

$$\pi(x) = \pi(0)T(x)! \prod_{i\in\theta_1\bigcup\theta_2} \frac{\tilde{\rho}_i^{x_i}}{x_i!}, \quad \text{if } x \in \mathcal{A}, \tag{14}$$

where

$$\pi(0) = \exp\left(-\sum_{i\in\theta_2}\tilde{\rho}_i\right)\frac{1-\rho}{1-\rho^{B+1}}. \tag{15}$$

Furthermore

$$\Pr\{N_1 = n_1, \ldots, N_K = n_K\} = \frac{n!(1-\rho)}{1-\rho^{B+1}} \prod_{k=1}^K \frac{1}{n_k!}\left(\sum_{i\in\alpha_k}\tilde{\rho}_i\right)^{n_k}, \tag{16}$$

and

$$\Pr\{N = n\} = \frac{\rho^n(1-\rho)}{1-\rho^{B+1}}, \tag{17}$$

where $n = \sum_{k=1}^K n_k \le B$.

Proof. Substituting (13) and (11) into (8) yields

$$\pi(x) = \begin{cases} \pi(x - e_i)\dfrac{\lambda_i T(x)}{\mu_i x_i}, & i \in \theta_1, \\ \pi(x - e_i)\dfrac{\lambda_i}{\mu_i x_i}, & i \in \theta_2. \end{cases}$$

We derive the relationship between $\pi(x)$ and $\pi(0)$ by iteratively applying the above formula.

$$\begin{aligned} \pi(x_1, \ldots, x_{\tilde{J}}, x_{\tilde{J}+1}, \ldots x_J) &= \pi(0, x_2 \ldots, x_{\tilde{J}}, x_{\tilde{J}+1}, \ldots x_J)\frac{\tilde{\rho}_1^{x_1}}{x_1!}T(x)[T(x)-1]\cdots[T(x)-x_1+1] \\ &\vdots \\ &= \pi(0, \ldots, 0, x_{\tilde{J}+1}, \ldots x_J)T(x)! \prod_{i\in\theta_1}\frac{\tilde{\rho}_i^{x_i}}{x_i!} \\ &\vdots \\ &= \pi(0, \ldots, 0)T(x)! \prod_{i\in\theta_1\bigcup\theta_2}\frac{\tilde{\rho}_i^{x_i}}{x_i!}. \end{aligned}$$

Now we are ready to derive the stationary probabilities for the number of customers in the sets.

$$\begin{aligned} \Pr\{N_1 = n_1, \ldots, N_K = n_K\} &= \sum_{\substack{N_k = n_k \\ k=1,\ldots,k}} \sum_{x_{\tilde{J}+1}=0}^{\infty} \cdots \sum_{x_J=0}^{\infty} \pi(0)T(x)! \prod_{i\in\theta_1}\frac{\tilde{\rho}_i^{x_i}}{x_i!}\prod_{i\in\theta_2}\frac{\tilde{\rho}_i^{x_i}}{x_i!} \\ &= \pi(0)n!\left[\sum_{x_{\tilde{J}+1}=0}^{\infty}\cdots\sum_{x_J=0}^{\infty}\prod_{i\in\theta_2}\frac{\tilde{\rho}_i^{x_i}}{x_i!}\right] \times \left[\sum_{N_K=n_K}\cdots\sum_{N_1=n_1}\prod_{k=1}^{K}\prod_{i\in\alpha_k}\frac{\tilde{\rho}_i^{x_i}}{x_i!}\right] \\ &= \pi(0)n!\exp\left(\sum_{i\in\theta_2}\tilde{\rho}_i\right) \times \prod_{k=1}^{K}\left(\sum_{N_k=n_k}\prod_{i\in\alpha_k}\frac{\tilde{\rho}_i^{x_i}}{x_i!}\right) \\ &= \pi(0)\exp\left(\sum_{i\in\theta_2}\tilde{\rho}_i\right) \times \frac{n!}{\prod_{k=1}^{K} n_k!}\left[\prod_{k=1}^{K}\left(\sum_{i\in\alpha_k}\tilde{\rho}_i\right)^{n_k}\right]. \end{aligned} \tag{18}$$

From the equation above, we sum over $n = n_1 + n_2 + \cdots + n_K$ to obtain

$$\begin{aligned} \Pr\{N = n\} &= \pi(0)\exp\left(\sum_{i\in\theta_2}\tilde{\rho}_i\right)\sum_{n_1+n_2+\cdots+n_K=n}\frac{n!}{n_1!\cdots n_K!}\left[\prod_{k=1}^{K}\left(\sum_{i\in\alpha_k}\tilde{\rho}_i\right)^{n_k}\right] \\ &= \pi(0)\exp\left(\sum_{i\in\theta_2}\tilde{\rho}_i\right)\rho^n. \end{aligned} \tag{19}$$

Since $\sum_{n=0}^{B}\Pr\{N = n\} = 1$, we get (15). Eqs. (16) and (17) follow by substituting (15) into Eqs. (18) and (19). □

4.2. A finite-capacity EPS system fed by poisson session traffic

Based on the queueing network results from the previous section, we now show that the joint queue length distribution, mean number of active flows, and blocking probabilities are insensitive to the session structure in the finite-capacity EPS queue fed by Poisson session arrivals.

Following [8,14], we model the EPS queue as a queueing network. For example, the network in Fig. 1 illustrates an EPS queue with two types of Poisson session traffic. The succession of flow transfer and thinking times of a session is viewed as a customer visiting two stations. Each node in a station represents an exponential phase of the phase-type

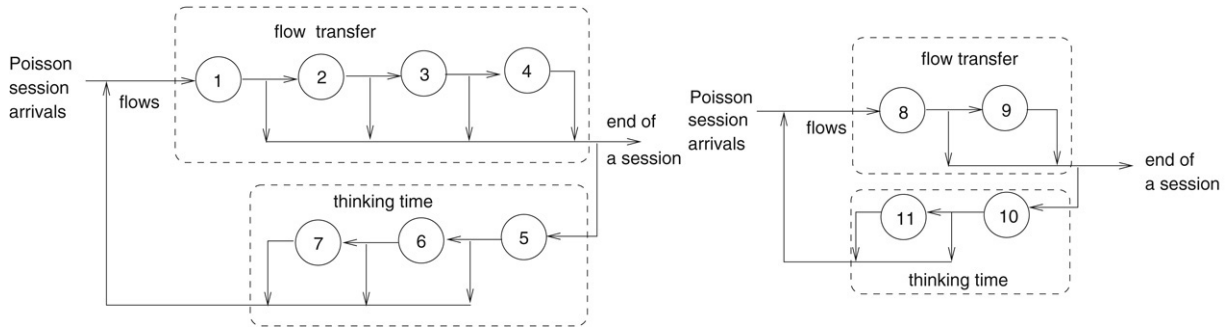


Fig. 1. A queueing network models an EPS system fed by two types of Poisson session arrival traffic.

distribution for that station. Since phase-type distributions form a dense subset within the set of all distributions with real, non-negative support, flow size and thinking time distributions can be very general. All nodes in transfer stations form set θ_1 while all nodes in thinking time stations form set θ_2. If node i is in a transfer station, it simultaneously serves x_i flows at the rate $x_i\mu_i/T(x)$, where $T(x)$ is the total number of flows in all transfer stations, and μ_i is the service rate of node i if there is only one flow at this node and in all transfer stations. Provided the number of nodes J is sufficiently large, we can represent a general session structure using such a network.

The EPS queue has a maximum capacity of B flows. We consider two retrial behaviors proposed by Bonald [5]. *Jump-over blocking* allows a session to continue in case of flow blocking. The following thinking time starts immediately. If the blocked flow is the last one, the session ends. For *random retrials*, each blocked flow is reattempted with a fixed probability p after an idle period of random duration. The session goes on as in the jump-over blocking model with probability $1-p$. The duration of the idle period between two attempts has the same distribution as the idle period that precedes the blocked flow.

It is proved that, with either retrial behavior, the network in [5] satisfies the conditions of Corollary 4.2. The derivation is independent of $\phi_i(x)$, but is dependent on the network topology. Similar to the derivation in [5], we can prove that networks like the one in Fig. 1 satisfy the conditions of Corollary 4.2. Details are omitted here to save space.

Corollaries 4.2 and 4.3 hold in the considered EPS system. Nodes in the flow transfer station of class k sessions form set α_k. The traffic intensity of class k traffic is ρ_k. Clearly, $\rho_k = \sum_{i\in\alpha_k}\tilde{\rho}_i$. Then Eq. (16) reduces to Eq. (3). By PASTA, the first flow of an arbitrary session sees the system in steady state at arrival. The successive flows also have such a property since nodes in thinking time stations are free and Corollary 4.2 applies. Therefore, the blocking probability for an arbitrary flow is (4) according to Eq. (17). For Poisson session traffic, the joint distributions of the number of active flows and the blocking probability are the same as that for Poisson flow traffic. In other words, they are independent of traffic details other than the intensity.

4.3. Summary

Our study extends the theoretical analysis of EPS systems in the literature. Assuming homogenous rate variation in an EV-DO system, we can model the complicated Poisson session traffic using simple Poisson flow traffic with exponentially distributed flow sizes for provisioning purposes.

5. DPS simulations

5.1. Model description

In the real world, the channel conditions experienced by different mobile users are different. Thus, heterogeneous variations in feasible rates are more likely to occur. We present the following DPS queue as an abstract model for a cellular data network using the PF scheduler.

A DPS queue handles multiple flow types, based on applications such as Web browsing and FTP downloading. Within each flow type, flows follow the same stochastic arrival and service time processes. These flows are further divided into multiple *subclasses* with different weights, to reflect the fact that flows share time slots unequally in the

heterogeneous rate variation scenario. Flows of type m have l_m subclasses, each of which is assigned weight $w_{m,i}$, $1 \leq i \leq l_m$ for its $\psi_{m,i}$ fraction of type m flow arrivals. Note that $\sum_{i=1}^{l_m} \psi_{m,i} = 1$. Vector $W_m := [w_{m,1}, w_{m,2}, \ldots, w_{m,l_m}]$ is referred to as the weight vector of type m. Assume that there are M types in total. For the entire DPS system, the number of classes is $K = \sum_{m=1}^{M} l_m$, and the weight vector $\vec{\mathbf{W}}$ is $[W_1, \ldots, W_M]$. The queue may have an infinite or finite capacity. In the finite case, jump over blocking is used for retrials.

Assume that all flows are geographically placed uniformly at random in the cell site, independently of their types. When there is no differential service, all flow types have the same number of subclasses and the same weight vector, namely $l_m = l$ and $W_m = W$, $\forall m \in \{1, \ldots, M\}$, where $W = [w_1, w_2, \ldots, w_l]$. This is assumed in this section. When QoS is introduced into the 3G system, flows from different applications may be treated unequally during transmission. Distinct types have different weight vectors. For example, $W_m = a_m \times W$, $m = 1, \ldots, M$. Adjusting the a_m coefficients controls the share allocated to each traffic type. Section 7 presents results for this case.

Note that the foregoing DPS queue has been tailored to the physical system under consideration. As discussed in Section 3.2, there are limited theoretical tools to study DPS systems. We resort to an extensive simulation study. Results are presented for the mean number of active flows ($E[N]$) and the blocking probabilities (when applicable) for the aggregate traffic and individual traffic types. Note that $E[N_i]$ and P_{bi} are for traffic type i instead of class i in this section. The other metrics such as mean response time and throughput can be obtained using Little's Law. Since these metrics show the same qualitative trends as $E[N]$, we do not include these results here.

By default, all traffic types have the same flow arrival rates and mean flow sizes. Within each flow type, all subclasses have equal arrival rates.

We consider several flow size distributions, including Deterministic, Exponential, (second-order) HyperExponential (H2), LogNormal (LN), and Pareto distributions. To characterize the variability of a distribution, we use the coefficient of variation (CV) if the variance is finite, and the shape parameter α if the variance is infinite. The α parameter measures the heaviness of the tail for the Pareto distribution; the smaller the α, the heavier the tail.

Session structure is determined by four characteristics: the distribution for the number of flows per session, the thinking time distribution, the flow size distribution, and the correlation in successive flow and thinking time statistics. For ease of presentation, we mainly show the results for a default session structure, in which the number of flows per session is geometrically distributed with mean 10, thinking times are exponentially distributed with mean 0.05 s, flow sizes have one of the five foregoing distributions with mean 1 s, and there is no correlation in flow and thinking-time statistics.

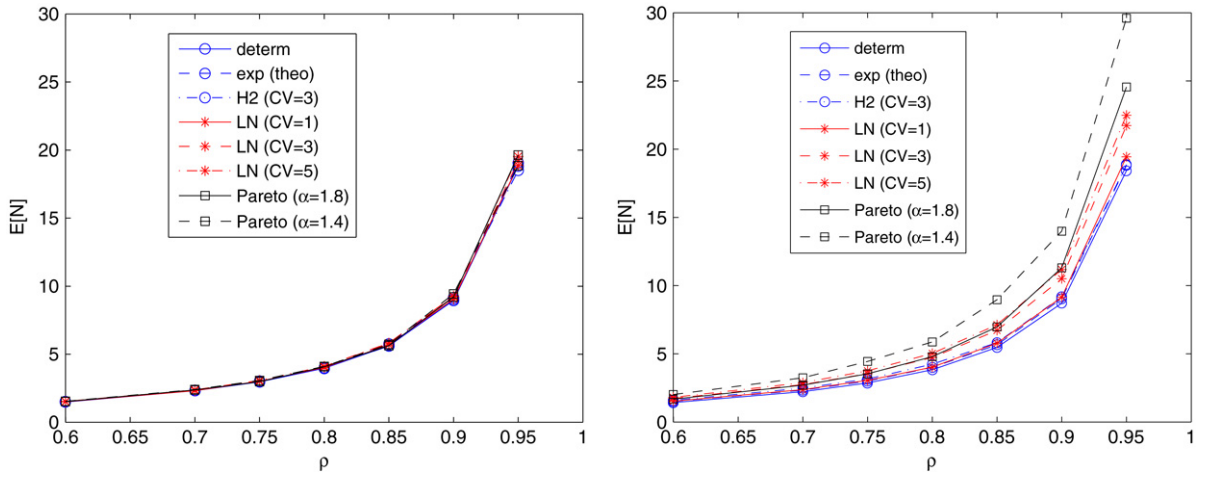
In a few selected cases, we can calculate the performance measures analytically using the results given in Section 3. When present, these results are marked "theo" in the graphs. In all other cases, simulation is used. The results for each simulation scenario are the average from 40 runs, each with 1 million flow arrivals. For results involving heavy-tailed distributions, each simulation run uses 10 million flow arrivals.

5.2. Infinite-capacity DPS

Experiment I considers the simplest scenario: a single type of Poisson flow arrivals with two subclasses. The weights for the two subclasses are [1, 2] in Fig. 2(a), and [1, 10] in Fig. 2(b).

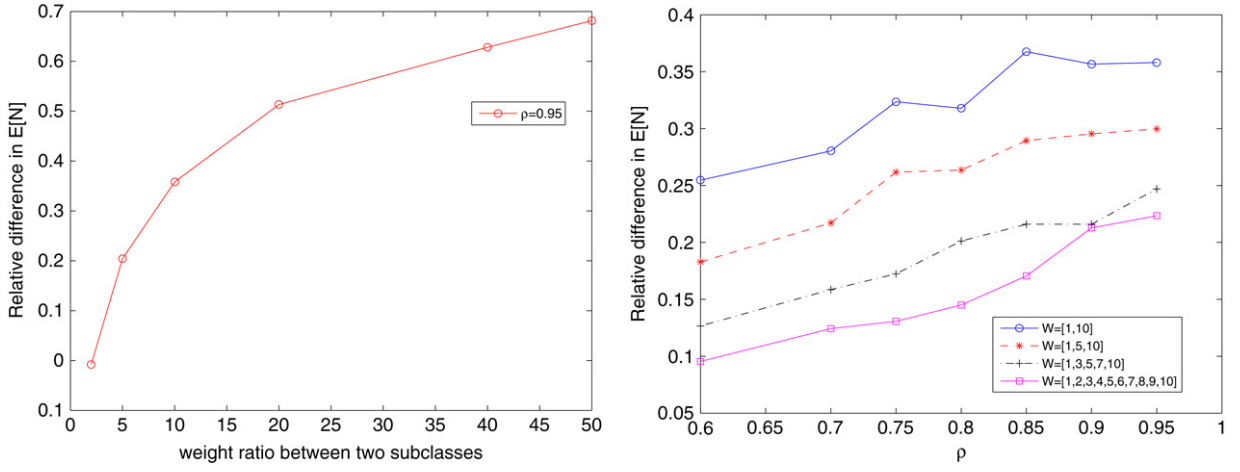
Fig. 2 illustrates three observations. First, there is little sensitivity to the flow size distribution when the ratio between subclass weights is low (specifically, 2 in Fig. 2(a)). Second, sensitivity emerges when the subclass weights differ greatly (e.g., by a factor of 10 in Fig. 2(b)). Furthermore, the sensitivities are most pronounced at high utilization (e.g., $\rho > 0.8$). Third, when sensitivities are present, they have the intuitive behavior. For example, within a given distribution type (e.g., LN), a larger CV leads to worse performance (i.e., greater value of $E[N]$). Similarly, for the heavy-tailed Pareto distribution, the heavier the tail, the worse the performance. Across different distributions, no precise characterization of the results is possible, since the performance does not depend solely on the CV and the heaviness of the tail.

The degree of sensitivity depends on the asymmetry of the weights within subclasses. In the previous experiment with weight ratio 10, the $E[N]$ value for the Pareto distribution ($\alpha = 1.4$) differs the most from that for the exponential distribution. We can characterize this sensitivity by the relative difference $(E[N_P] - E[N_E])/E[N_P]$, where $E[N_P]$ and $E[N_E]$ are for the Pareto and Exponential distributions, respectively. The smaller this difference is, the lower the sensitivity is and the smaller the error is when replacing the Pareto flow size distribution with the exponential one. Experiment II focuses on these relative differences.



(a) $W = [1, 2]$. (b) $W = [1, 10]$.

Fig. 2. Experiment I: Effect of weight vector for a single type of Poisson flow traffic with two subclasses.

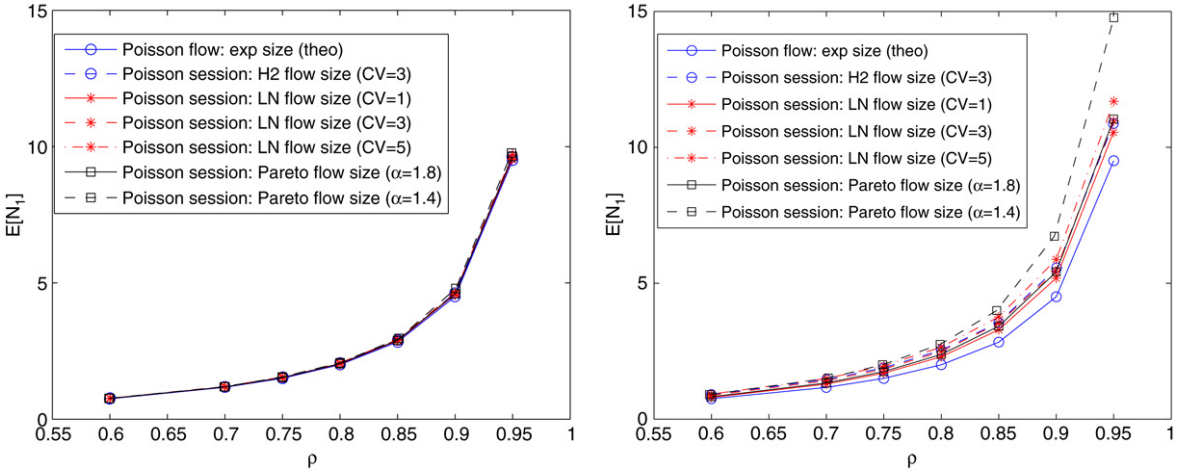


(a) Effect of weight ratio between two subclasses. (b) Effect of subclasses number.

Fig. 3. Experiment II: Effect of weight asymmetry among subclasses for a single type of Poisson flows with Pareto distributed sizes ($\alpha = 1.4$).

Fig. 3 illustrates the impact of the weights on the sensitivity. Fig. 3(a) plots the relative difference for utilization 0.95. As the ratio of weights increases from 2 to 5, the difference dramatically increases. In Fig. 3(b), we increase the number of subclasses from 2 to 10. Each subclass contributes equal load. While the relative difference always increases with load, the asymmetry of weights decreases as the number of subclasses increases. Thus the sensitivity to the flow size distribution diminishes.

Experiment III involves two types of Poisson traffic. We are interested in the performance sensitivity of the individual traffic types to the traffic details. Both types have two subclasses and the same weight vectors (namely, $W_1 = W_2$). Traffic type 1 has Poisson flow arrivals with exponentially distributed sizes, while traffic type 2 has Poisson session arrivals with our default session structure. We fix the characteristics of traffic type 1, and vary the flow size distribution for traffic type 2. The results are compared to those for the case when traffic type 2 has Poisson flow arrivals with exponentially distributed sizes. Fig. 4(a) shows that approximate insensitivity of $E[N_1]$ (for type 1) to the session structure holds when the subclass weights are $[1, 2]$. When the weights are $[1, 10]$ in Fig. 4(b), the effects of the session details are manifest. We have the same observations for $E[N_2]$ and $E[N]$. Increased variability in the flow size distribution adversely impacts the system performance, when the other session characteristics are



(a) $W_1 = W_2 = [1, 2]$. (b) $W_1 = W_2 = [1, 10]$.

Fig. 4. Experiment III: Effect of session structure for traffic type 2 (traffic type 1 is Poisson flow arrivals with exponentially distributed flow sizes).

fixed. However, it is difficult to make general observations about the impact of different session structures, as the four session characteristics influence the performance in a complicated fashion.

5.3. Finite-capacity DPS

Cellular systems usually implement admission control to prevent overload and maintain the quality of service for accepted flows. We now check the applicability of the previous observations to a DPS system with admission control. As discussed in Section 3, moving from an infinite-capacity system to a finite one is not trivial. It is not obvious that the same properties should hold in both systems.

In Experiment IV, we simulate a finite-capacity system that allows at most $B = 15$ concurrent flows. The retrial behavior in the case of flow blocking is jump over blocking. The queue is fed by two types of Poisson session traffic. Traffic type 1 has 5 flows per session, with lognormal flow sizes (mean 2, CV 3) and hyperexponential thinking times (mean 1, CV 3). Traffic type 2 has our default session structure.

We vary the session details of type 2 and compare the results to those for the case when both types of traffic are Poisson flow arrivals with exponentially distributed sizes. Fig. 5 shows the overall blocking probability P_b. The other metrics such as $E[N]$, P_{b1}, and P_{b2} have the same qualitative trends.

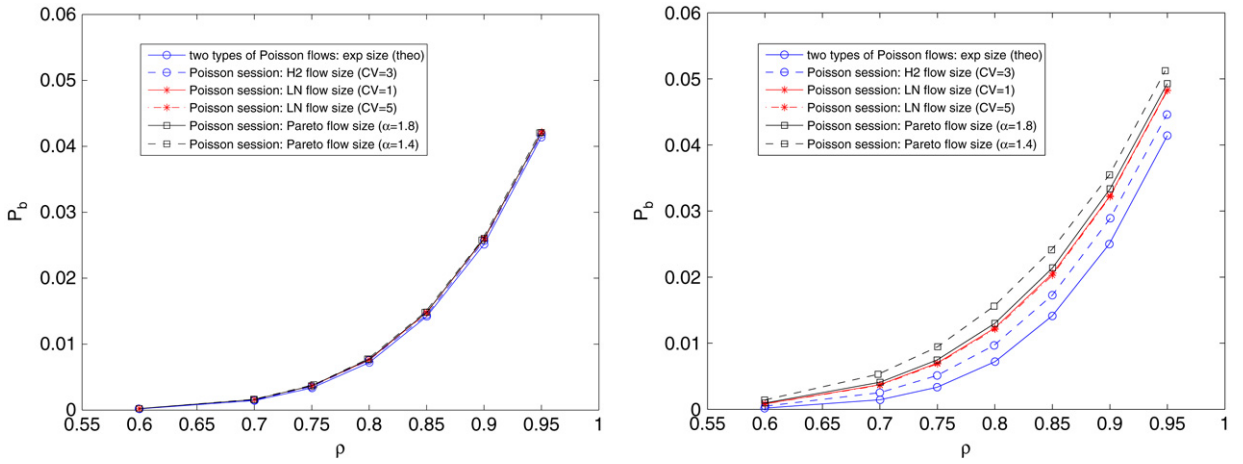
As in the infinite-capacity system, approximate insensitivity holds for individual traffic types and aggregate traffic provided that the subclass weights are not highly skewed. In this case, we can ignore the complicated session structure and replace the Poisson session arrivals with Poisson flow arrivals of the same intensity. When sensitivity emerges for highly skewed weights, increased variability in the flow size distribution adversely affects the performance (other session characteristics remaining fixed).

When we make the subclass weights equal in the above experiment, the simulation results indicate that Eqs. (3) and (4) do apply. This verifies our theoretical analysis in Section 4.

5.4. Summary

The four experiments presented here are selected illustrative examples of the extensive simulations that we have conducted. We varied the simulation scenarios by increasing the number of traffic types and the number of subclasses per type, and using different session structures. The previous observations always hold. When there is no differential treatment of traffic types, the first-order performance measures show little sensitivity to the flow details (flow size distribution, variability, and the session structure), particularly when the subclass weights are not highly skewed.

The latter scenario is likely to be representative of practical cellular systems. That is, the unequal sharing caused by the interaction between PF scheduling and heterogeneous rate variations is unlikely to exceed a factor of two. Thus, we conjecture that the traffic details in practical cellular systems will not affect the metrics relevant for network provisioning. In the next section, we confirm this conjecture by simulating an EV-DO cell.



(a) $W_1 = W_2 = [1, 2]$. (b) $W_1 = W_2 = [1, 10]$.

Fig. 5. Experiment IV: Effect of session structure of traffic type 2 on blocking probability ($B = 15$).

Table 1
Slot shares allocated to each node in the static user scenario

Node	1	2	3	4	5	6
Slots (%)	18.44	18.44	17.91	16.91	14.57	13.73

6. EV-DO System simulations

6.1. EV-DO Model

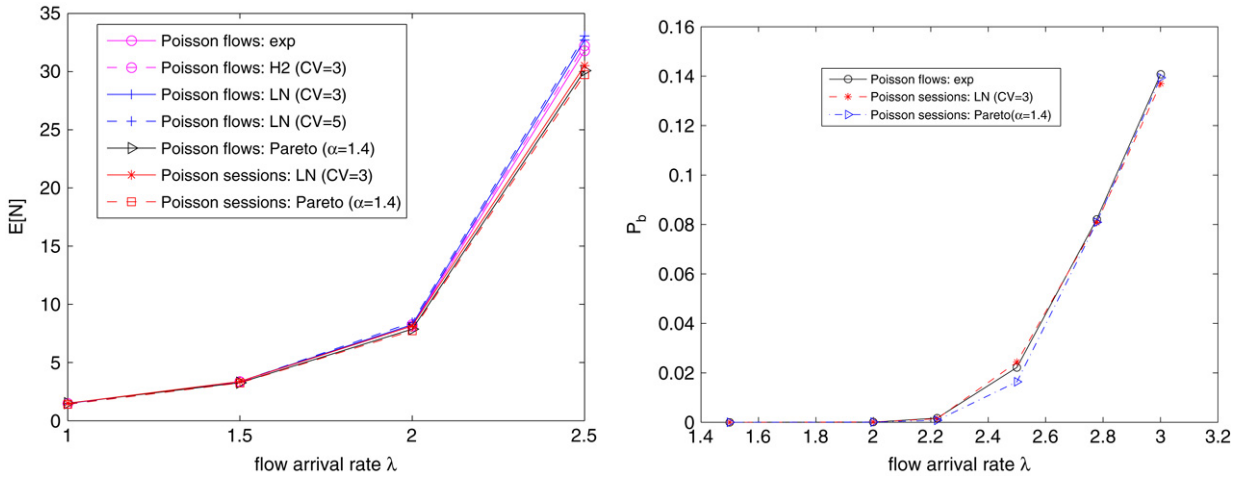
In our EV-DO model, we simulate a shared downlink data channel of a central cell site surrounded by 18 interfering cells. The cell size is 1 km, and all BSs transmit at full power P on the downlink. The channel model includes propagation loss (represented by $L(d)$, where d is the distance of the user from the BS), slow fading (χ), and fast fading (ξ). Propagation loss uses the modified Hata urban model, and slow fading is lognormal with standard deviation 8.9 dB. These models are based on the 3GPP2 specification [1]. The fast fading is Rayleigh fading. Flows are placed uniformly at random in the center cell, and users do not move during flow transmission. Each active flow has a time-varying SINR, which is updated at every slot using:

$$\mathrm{SINR}_j(t) = \frac{PL(d_{0,j})\chi_{0,j}(t)\xi_{0,j}(t)}{N_0 + \sum_{i=1}^{18} PL(d_{i,j})\chi_{i,j}(t)\xi_{i,j}(t)}, \tag{20}$$

where N_0 is the thermal noise power, and subscript $\{i, j\}$ denotes the path from the ith BS ($i = 0$ is the home BS) to the jth mobile user in the home cell. The corresponding feasible rate $r_i(t)$ is obtained by consulting the "SINR versus rate" table used in practical systems [9]. The time constant of the PF scheduler t_c is set to the typical value 100.

Several papers [15,22] have studied the unfairness of the PF algorithm. Here we use a simple experiment to show the existence of PF unfairness in our simulation setting. Consider the static user scenario. Six mobile nodes are placed along a line drawn eastward from the BS to the cell boundary. Nodes are numbered in order of their distance from the BS. Node 1 is 100 m away, and Node 6 is 1 km away. The remaining nodes are at 180 m intervals between these two. An infinite amount of data is destined to each user.

Table 1 shows the time fraction allocated to each node for downlink transmission. The scheduler indeed discriminates against users who are far from the home BS, with higher variability in the feasible rates. Equal time sharing does not occur in cellular systems with a PF scheduler.



(a) Infinite. (b) Finite ($B = 40$).

Fig. 6. EV-DO simulation results demonstrating approximate insensitivity to the session structure.

6.2. Simulation results

Returning to the dynamic user scenario, we now investigate the impact of such unequal sharing on the performance insensitivity. We simulate Web browsing, where a flow corresponds to the downloading of one Web page with mean size 50 KB. Flows arrive as a Poisson arrival process, or as a component within a Poisson session arrival process. For Poisson sessions, the number of flows per session is geometrically distributed with mean 30, and thinking times are exponentially distributed with a mean of 30 s. We change the flow size distribution and compare the results to that of Poisson flow arrivals with exponentially distributed sizes. Each simulation run corresponds to 200,000 flows, or 800,000 flows if the Pareto distribution is used.

Fig. 6 shows the results for infinite- and finite- ($B = 40$) capacity systems. The first-order performance is largely insensitive to the precise traffic characteristics. For the finite-capacity system, $E[N]$ (not shown here) is also insensitive.

The previous DPS study shows that the impacts of traffic details, if any, are most pronounced at high utilization. However, in the cellular system, high load may not necessarily compromise insensitivity if flows are uniformly distributed. The reason is that increasing the number of concurrent flows (i.e., higher load) actually ameliorates the weight asymmetry among subclasses. To show this effect, we change the number of nodes in the previous static user scenario. All nodes are equally spaced along the line from the home BS to the edge. For each number of nodes considered, we measure the unfairness by the coefficient of variation of the slot shares. As shown in Fig. 7, the degree of unfairness tends to diminish with more users, which helps to preserve insensitivity even at high load.

The next experiment studies the case in which flows are non-uniformly distributed within the cell. Specifically, flows are placed at random into one of the two doughnut-shaped rings (one near the cell center, and one at the cell periphery, each of width 10 m) according to a certain probability. There is no admission control. Fig. 8 shows that the approximate insensitivity of the mean number of active flows is still maintained. In this hypothetical case, the asymmetry among subclasses, though exaggerated compared to that in the real world, is still less than 2.

6.3. Summary

The observed phenomenon in the EV-DO simulation comes from the fundamental properties of DPS: approximate insensitivity holds as long as the subclass weights are not highly asymmetric. Approximate insensitivity is a valuable property. It eliminates the necessity of modeling the characteristics of applications that may change over time. Replacing complicated traffic by Poisson flow arrivals with exponentially distributed sizes generates results that are sufficiently accurate for traffic engineering purposes. This approach can greatly reduce the complexity of the provisioning procedure.

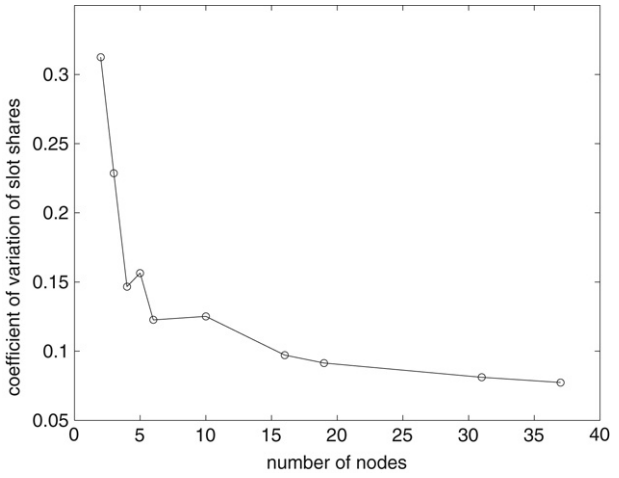


Fig. 7. Dispersion of slot shares versus the number of static nodes.

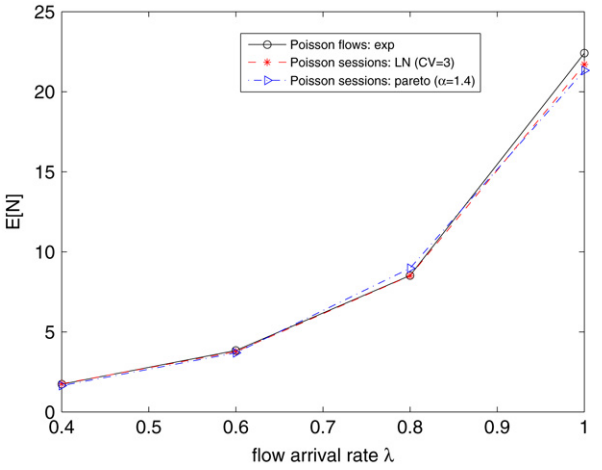


Fig. 8. Simulation results when flows are put into two rings with one at the center and one at the edge.

7. Further discussion and caveats

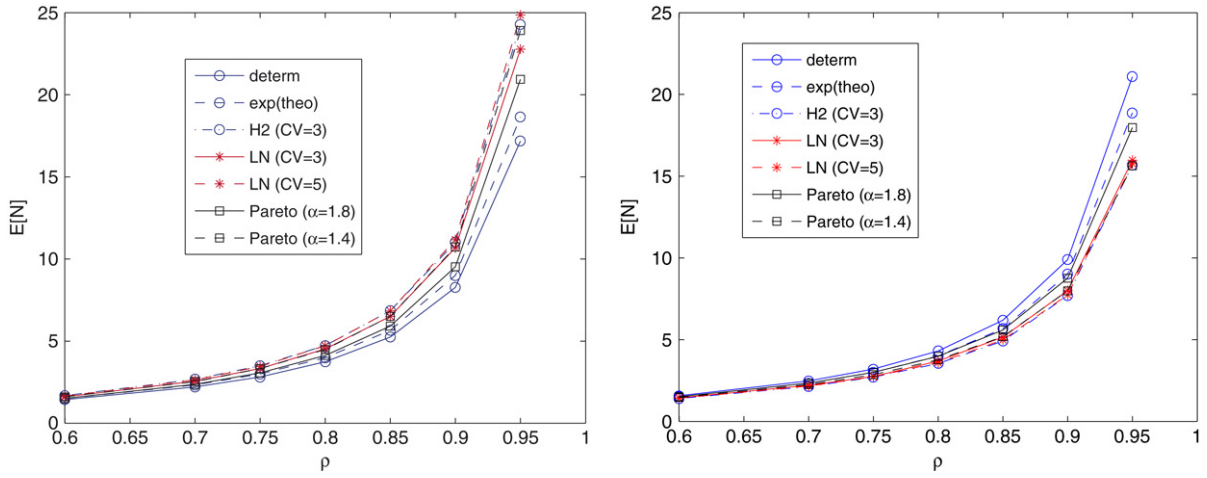
7.1. Service differentiation

So far we have considered systems with a low-layer bias among subclasses. Such bias is caused unintentionally by the interaction between PF scheduling and heterogeneous rate variation.

When QoS is introduced into cellular networks, differentiated service mechanisms may deliberately treat traffic unequally (e.g., at the type level). It is interesting to know whether such features change the guidelines for traffic engineering.

We return to DPS as an abstract model of the cellular system. For the sake of clarity, Experiment V only considers two types of Poisson flow arrivals, each with a single subclass. The single-subclass assumption is reasonable when the bias among subclasses is minimal. One example of such a scheduling scheme is score-based scheduling [3,4].

Unlike the previous experiments, the two traffic types are assigned different weights. Our interest is to what extent the weight asymmetry between traffic types changes the insensitivity property. Traffic type 1 is assigned a weight of unity, while traffic 2 is assigned a weight of a ($a > 1$). The weight vector for the aggregate traffic is $\vec{\mathbf{W}} = [1, a]$. Traffic type 2 is high priority since it has a larger weight. Each type of traffic contributes equally to the total system load. We fix the flow size distribution of one traffic type (Exponential), and vary the flow size distribution for the other type.



(a) High priority.

(b) Low priority.

Fig. 9. Experiment V: Impact of flow size distributions when the weight ratio between two traffic types is 2.

Fig. 9 shows the impacts of the high and low priority traffic. We use $a = 2$ to facilitate comparison with previous simulation results.

There are two important observations in these figures. First, compared to the bias among subclasses, the bias among traffic types manifests sensitivity in a much more dramatic way. When $a = 2$, the impact of the flow size distribution is evident even for the hyperexponential distribution. If the ratio is reduced (e.g., $a = 1.2$), the impact of the flow size distribution becomes negligible (not shown here). Second, depending on the traffic priority, variability in the flow size distribution can have different impacts. Specifically, for a given flow size distribution, higher variability in the flow size distribution for high priority flows (type 2) makes performance worse (i.e., increases $E[N]$). Similarly, for the Pareto distribution, the heavier the tail, the worse the performance. This observation is consistent with intuition: the more variability in flow sizes, the worse the usual queueing performance measures are. In Fig. 9(b), however, the impact of variability in the low priority traffic has the opposite effect. For example, deterministic flow sizes result in the worst performance.

This counter-intuitive result has also been observed in work by others. For example, the results in Table I through X in [21] reflect the exact same phenomenon, though the authors do not highlight the trend. In an EPS system with time-varying service rate, a similar phenomenon is observed, and theoretical support for the result is given [12,17].

7.2. Summary

The results from this experiment suggest that the *deployment of differential services may fundamentally change the traffic engineering approach used in cellular systems*. The performance is sensitive to traffic details if differentiation among traffic types is present. Actual network traffic is highly variable, with complicated structure. Assuming Poisson flow arrivals with exponentially distributed sizes may lead to under-estimation or over-estimation of system performance.

8. Conclusions

Motivated by the desire for simple and robust traffic engineering rules for the downlink of cellular systems using the PF scheduler, we study the "strict" and "approximate" insensitivity of a Processor Sharing (PS) system. The physical system is modeled by an EPS or DPS queue depending on the assumption of channel conditions. In the EPS queue, all flows are allocated an equal share of transmission slots regardless of flow types and locations. In the DPS queue, each traffic type is further divided into subclasses. Different weights are assigned to the subclasses to reflect the unequal sharing that occurs in heterogeneous channel conditions.

The performance insensitivity of the EPS queue is a known result in a few cases. We extend the analysis of this property to a new case in which the input is Poisson session traffic and admission control exists.

Our DPS study shows that the first-order performance is insensitive to flow size distributions and session structure, provided that the asymmetry of weights amongst subclasses is moderate (the expected case in practice). With insensitivity, the only traffic characteristics needed for performance modeling are the traffic intensity. Our findings (confirmed using an EV-DO simulation model) are encouraging, because they may greatly reduce the complexity of provisioning for cellular systems using PF scheduling.

However, a different conclusion is drawn for systems with differentiated services. If discrimination across traffic types exists, then the performance metrics are much more sensitive to the traffic details. The imminent introduction of differentiated services in cellular systems may pose a great challenge for future network provisioning.

Our ongoing work is on studying the sensitivities in a cellular system with channel-aware and QoS-aware scheduling schemes.

References

[1] 3GPP2, CDMA2000 Evaluation Methodology, C.R1002-0 V1.0, 2005.
[2] E. Altman, K. Avrachenkov, U. Ayesta, A survey on discriminatory processor sharing, Queueing Systems: Theory and Applications 53 (2006) 53–63.
[3] T. Bonald, A score-based opportunistic scheduler for fading radio channels, in: Proceedings of European Wireless, Barcelona, 2004.
[4] T. Bonald, Flow-level performance analysis of some opportunistic scheduling algorithms, European Transactions on Telecommunications 16 (2005) 65–75.
[5] T. Bonald, The Erlang model with non-poisson call arrivals, in: Proceedings of ACM SIGMETRICS, Saint Malo, France, June 2006.
[6] T. Bonald, L. Massoulie, Impact of fairness on internet performance, in: Proceedings of ACM SIGMETRICS, June 2001, pp. 82–91.
[7] T. Bonald, A. Proutiere, On stochastic bounds for monotonic processor sharing networks, Queueing Systems: Theory and Applications 47 (2004) 81–106.
[8] T. Bonald, A. Proutiere, G. Regnie, J. Roberts, Insensitivity results in statistical bandwidth sharing, in: ITC-17, Salvador, Brazil, December 2001.
[9] S. Borst, User-level performance of channel-aware scheduling algorithms in wireless data networks, in: Proceedings of IEEE INFOCOM, San Francisco, April 2003.
[10] O. Boxma, N. Hegde, R. Nunez-Queija, Exact and approximate analysis of sojourn times in finite discriminatory processor sharing queues, International Journal of Electronics and Communications 60 (2006) 109–115.
[11] J. Cohen, The multiple phase service network with generalized processor sharing, Acta Informatica 12 (1979) 245–284.
[12] F. Delcoigne, A. Proutiere, G. Regnie, Modeling integration of streaming and data traffic, Performance Evaluation 55 (2004) 185–209.
[13] G. Fayolle, I. Mitrani, R. Iasnogorodski, Sharing a processor among many classes, Journal of the ACM 27 (1980) 519–532.
[14] S. Fredj, T. Bonald, A. Proutiere, G. Regnie, J. Roberts, Statistical bandwidth sharing: A study of congestion at flow level, in: Proceedings of ACM SIGCOMM, San Diego, USA, August 2001.
[15] J. Holtzman, Asymptotic analysis of proportional fair algorithm, Proceedings of IEEE PIMRC 2 (2001) 33–37.
[16] F. Kelly, Reversibility and Stochastic Networks, Wiley, Chichester, 1979.
[17] R. Litjens, R. Boucherie, Elastic calls in an integrated services network: The greater the call size variability the better the qos, Performance Evaluation 52 (2003) 193–220.
[18] R. Litjens, F. Roijers, J. van den Berg, R. Boucherie, M. Fleuren, Performance analysis of wireless LANs: An integrated packet/flow-level approach, in: Proceedings of ITC-18, Berlin, 2003, pp. 931–940.
[19] J. Roberts, W. Queslati-Boulahia, Quality of service by flow-aware networking, Philosophical Transactions of the Royal Society of London A (2002) 2197–2207.
[20] G. van Kessel, R. Nunez-Queija, S. Borst, Asymptotic regimes and approximations for discriminatory processor sharing, SIGMETRICS Performance Evaluation Review 32 (2004) 44–46.
[21] G. van Kessel, R. Nunez-Queija, S. Borst, Differentiated bandwidth sharing with disparate flow size, in: Proceedings of IEEE INFOCOM, Miami, Florida, 2005, pp. 2425–2435.
[22] C. Westphal, Monitoring proportional fairness in CDMA2000 high data rate networks, in: Proceedings of IEEE GLOBECOM, Dallas, TX, 2004.

Yujing Wu received the Ph.D. degree in Electrical and Computer Engineering from the University of Massachusetts at Amherst. She worked on performance evaluation of 3G cellular networks as a Research Associate in the Department of Computer Science at the University of Calgary during February 2004–May 2007. Currently, she is with QuIC Financial Technologies Inc.

Carey Williamson holds an iCORE Chair in the Department of Computer Science at the University of Calgary, specializing in *Wireless Internet Traffic Modeling*. He has a B.Sc. (Honours) in Computer Science from the University of Saskatchewan, and a Ph.D. in Computer Science from Stanford University. His research interests include Internet protocols, wireless networks, network traffic measurement, network simulation, and Web server performance.

Jingxiang Luo currently works as a Research Associate in the Department of Computer Science, University of Calgary, Canada. He received his Ph.D. (2004) and M.Sc. (2000) in Computer Science, both from the University of Saskatchewan. His interests are in the areas of performance evaluation, queueing and network theories, modeling and simulation methodology, and algorithm analysis. Currently, he is investigating evaluation of computer and communication systems, especially the traffic analysis and performance modeling in mobile and wireless networks.