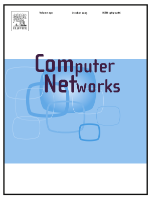
Computer Networks 275 (2026) 111807

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet



Computer network optimization using the power metric for multiple flows: Part II — Extension to continuous priority queueing disciplines

Meng-Jung Chloe Tsai *, Leonard Kleinrock

ARTICLE INFO

Keywords:
Optimal network performance
Optimized queueing system
System performance optimization
Power performance metrics
Throughput delay tradeoff
Priority queueing disciplines
Queueing
Delay dependent queueing disciplines
Beta priority queueing disciplines
Flow priority discrimination
Continuously adjustable priority queueing disciplines

ABSTRACT

Part I [1] introduced three performance metrics based on "power", namely *individual power* P_i, *sum of individual powers* P_{sum}, and *average power* P_{avg}, and analyzed their optimization in multi-flow queueing systems under two extreme scheduling disciplines, specifically the least discriminatory First-Come, First-Served (FCFS) and the most discriminatory Head-of-the-Line (HOL) preemptive resume priority discipline. Building on that foundation, this follow-up study provides the designer greater **flexibility** to range the flow priority discrimination continuously from FCFS to HOL by introducing families of queueing disciplines to accommodate the designers' mixed workloads with diverse SLO in their systems. We examine two such **continuous** families—the **delay-dependent system** and our newly created **beta-priority system**—each spanning the full spectrum from minimal to maximal flow priority discrimination. These systems provide a continuous control mechanism, enabling power metric optimization beyond the constraints of fixed-discipline analysis. These two are examples of the many possible families that exist, and each has its unique trajectory from one extreme discipline to the other.

We focus on selecting the flow utilization to optimize **individual power** P_i and **sum of individual powers** P_{sum}, respectively (we leave out *average power* in this study since its optimal value is invariant across the queueing disciplines, as shown in [1]). We begin with the two-flow case to compare both full spectrum systems. While both systems exhibit similar patterns in individual power optimization, they differ in their outcomes for sum of individual powers optimization. For an arbitrary number of flows n, we focus on the beta-priority system due to its analytical tractability. We derive closed-form solutions for the optimal utilization of lowest-priority flows, and use numerical methods to compute equilibrium outcomes. In particular, we observe an increase in both system utilization and the maximal total powers as the level of discrimination increases in optimizing sum of individual powers, highlighting the efficiency gains achievable through prioritization in multi-flow systems.

1. Introduction

Part I [1] discussed performance optimization in a multi-flow setting by introducing and analyzing power metrics under two fundamental queueing disciplines: First-Come, First-Served (FCFS) and Head-of-the-Line (HOL) preemptive resume, in M/M/1. That study proposed three distinct definitions of the power metric—(1) individual power, (2) sum of individual powers, and (3) average power—each motivated by different operational objectives and system contexts.[1] By optimizing each of these metrics through appropriate selection of flow utilization values, it provided foundational insights into the tradeoff between throughput and response time under fixed scheduling disciplines such as FCFS and HOL.

While analytically convenient, FCFS and HOL may impose, respectively, too little or too much discrimination for heterogeneous workloads in multi-tenant cloud environments and computer networks. FCFS is **too relaxed**; for example, without prioritization, urgent or short jobs can get stuck behind long ones, harming responsiveness and failing to meet latency Service Level Objectives (SLOs). Conversely, strict HOL is **too severe**; for example, a sustained stream of high-priority arrivals can starve lower-priority jobs, delaying them indefinitely. This is problematic for mixed workloads where some tasks are latency-critical, while

* Corresponding author.
E-mail addresses: chloe16808@ucla.edu (M.-J.C. Tsai), lk@cs.ucla.edu (L. Kleinrock).

[1] In Paper I and throughout this work, we primarily focus on a common exponential distribution for each flow's service time. This choice allows us to establish a key result for the first power metric, individual power: the optimized utilization is equal to half of the remaining available utilization when service times are exponentially distributed with a common mean across flows. To enable comparison of this result (Table 2) with the optimization outcome for the second power metric, sum of individual powers (Table 3), we likewise assume the common exponential distribution. For the third power metric, average power, we extend the analysis to generalized service time distributions.

https://doi.org/10.1016/j.comnet.2025.111807
Received 26 June 2025; Received in revised form 31 August 2025; Accepted 22 October 2025
Available online 8 November 2025

others must still meet a deadline despite being less time-sensitive. Because neither extreme provides sufficient flexibility, these realities highlight the need for more **flexible** scheduling disciplines that can reconcile the competing demands of high-priority tasks, which require immediate, latency-sensitive responses, and low-priority tasks, which, while less time-critical, still require a guaranteed, eventual completion.

Modern cloud data centers and networks concurrently serve **latency-critical interactive services** (e.g., RPCs and real-time analytics) alongside **throughput-oriented batch jobs**, backups, and large model or dataset transfers. On the compute side, Google's cluster management system, Borg, categorize their heterogeneous workloads into two main classes: long-running services for latency-sensitive user-facing requests (microseconds to hundreds of milliseconds), and batch jobs that run for seconds to days with far less sensitivity to performance fluctuations [2,3]. Kubernetes, an open-source orchestrator, formalizes a similar concept with Pod QoS classes—Guaranteed, Burstable, and BestEffort—with distinct resource guarantees and eviction semantics [4]. On the network side, a similar heterogeneity exists due to diverse workload types and data center structures. An empirical study spanning university, enterprise, and cloud environments found traffic to be bursty and dominated by many small, short-lived "mice" flows [5]. In contrast, a study of Facebook's network reported a very different pattern, with many flows being long-lived and exhibiting distinct properties in locality, stability, and predictability [6]. These differing observations suggest the need for a differentiated service model rather than a one-size-fits-all policy.

To support such mixed workloads in the data center, several solutions have been proposed. On the congestion-control side, approaches include delay-based control (e.g., TIMELY [7], Swift [8]) and protocols that use ECN signals to keep queues short (e.g., DCTCP [9], DCQCN [10], HPCC [11]). On the scheduling side, several protocols use shortest-job-first-style strategies to reduce flow completion time, including pFabric [12], pHost [13], Homa [14], and Karuna [15]. On the application-aware side, approaches incorporate workload-level information to guide resource allocation. Saba[16] uses application bandwidth sensitivity to minimize slowdown across applications. Coflow[17–19] introduces an abstraction capturing communication requirements in data-parallel jobs. Aequitas[20] maps different application-level priorities onto corresponding QoS-weighted queues with weighted fair queueing to enforce RPC-level latency SLOs even under overload.

These diverse approaches illustrate the broad proliferation of queueing disciplines and their many variants. Rather than analyze each queueing discipline in isolation, we model them along a single axis of *flow priority discrimination*: FCFS (minimal discrimination) at one end and strict HOL (maximal discrimination) at the other. Between them is a rich continuum that better accommodates mixed SLOs. We introduce two flexible families, each of which spans its own spectrum from FCFS to HOL: the **delay-dependent** system [21], where a flow's effective priority increases with experienced delay (capturing SLO or deadline driven promotion), and our newly created **beta-priority** system, which provides a tunable, weight-like control on prioritization intensity. Both families allow us to determine the utilization at which a system should operate along their respective spectra. In practice, operators may select a priority discrimination level aligned their workload or SLO portfolio and use our results as guidance when selecting a target utilization.[2,3]

Building on Part I [1] and assuming the operator has chosen a point in the spectrum of possible priority disciplines that match their SLO requirements, we use the two power metrics—**individual power** P_i and **sum of individual powers** P_{sum}—to select the utilizations that maximize the chosen metric under two policy families: the delay-dependent and beta-priority families. These families introduce tunable parameters that control the degree of flow priority discrimination and together span the range from FCFS (minimal discrimination) to HOL (maximal discrimination). We begin by analyzing the two-flow case ($n = 2$) to develop intuition and validate the modeling approach, and then extend to an arbitrary number of flows n using analytical and numerical methods. The results characterize the *operating point* that the average utilizations the system should maintain in order to maximize the chosen power metric for any fixed level of flow priority discrimination (i.e., for a chosen queueing discipline).

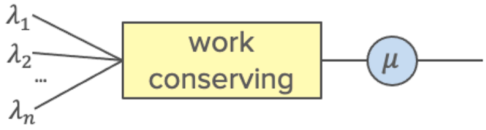


Fig. 1. Model for an M/M/1 system with multiple flows using work-conserving queueing disciplines.

2. Background

To begin our investigation, we adopt the same multi-flow single-server queueing model introduced in Part I [1]. This model, illustrated in Fig. 1, serves as the foundation for our analysis in determining the flow utilizations that optimize the power metrics under various scheduling disciplines. In the following, we briefly review the system setup and notation, the FCFS and HOL queueing disciplines along with their corresponding mean response times, and the three power metrics defined in Part I [1], along with their optimization results under both FCFS and HOL.

2.1. The single-server queueing system with multiple flows

We consider the same queueing model as in [1], illustrated in Fig. 1. The system consists of n priority groups indexed from $i = 1, 2, \dots, n$, where the i^{th} group represents a flow modeled as a Poisson arrival process with an average packet arrival rate of λ_i packets per second. Packet lengths of each flow i are independently and identically distributed, drawn from an exponential distribution with a common mean service time $\frac{1}{\mu_i} = \frac{1}{\mu}$ seconds.[4] The utilization factor of flow i is therefore given by $\rho_i = \frac{\lambda_i}{\mu}$. These n independent Poisson processes can be viewed as a single aggregated Poisson process with a total average arrival rate of $\lambda = \sum_{i=1}^{n} \lambda_i$, and a total system utilization of $\rho = \sum_{i=1}^{n} \rho_i = \sum_{i=1}^{n} \frac{\lambda_i}{\mu} = \frac{\lambda}{\mu}$. Following [1], we require $\rho = \sum_{i=1}^{n} \rho_i < 1$ to ensure system stability.[5]

The queueing discipline is assumed to be work-conserving [24], meaning it neither creates nor eliminates workload within the system. It merely determines the service order of packets without affecting the total amount of work to be processed. Different ways of ordering packets lead to different mean response times for each flow. There exist infinitely many ways to vary the service order, each resulting in a different distribution of response times across flows—and thus different degrees of flow priority discrimination. Among these, two extreme cases are First-Come, First-Served (FCFS), which imposes no flow discrimination, and Head-of-the-Line (HOL) preemptive resume priority, which represents the highest level of flow priority discrimination.

[2] This analysis is a first step. The standard power metric (throughput over delay)[22] weighs throughput and delay equally. If one or the other should be emphasized more heavily, then a generalized power $P^r = \rho^r/(\mu T)$ can be used, where $r > 0$ encodes the relative importance of throughput [22,23].

[3] The selected target utilization can be used for network capacity planning or serve as a reference for network-aware job placement.

[4] In general, the mean service time can depend on the priority class. For analytical and computational tractability, we assume a common mean service time across all classes in this paper.

[5] Do note there are some systems that allow some priority groups to exhibit stable behavior for $\rho \geq 1$; for example, HOL. However, in this paper, we will assume all HOL priority groups to be stable and have finite response time.

2.2. *Response time in FCFS and HOL*

We now describe the two extreme cases of flow priority discrimination in work-conserving queueing systems [24]—First-Come, First-Served (FCFS) and Head-of-the-Line (HOL)—and introduce their respective mean response time functions.

2.2.1. *First-Come, First-Served (FCFS)*

In an FCFS system, where there is no flow priority discrimination (i.e, minimum discrimination), each flow has the same mean response time:

$$T_i = T = \frac{1}{\mu(1-\rho)} \quad \text{for all} \quad i = 1, \dots, n \tag{1}$$

This shared mean response time is determined by the average service rate μ, and by the total system utilization ρ.

2.2.2. *Head-of-the-Line (HOL)*

In the HOL system [25], the flow priority discrimination is maximal because higher-priority flows are always served before all lower-priority ones present in the system. We assume that the lower the index of the priority group, the higher is that group's priority. Furthermore, we consider the preemptive priority case. This means that if a packet from group i arrives while a lower-priority packet (from group $j > i$) is being served, the service of the lower-priority packet is immediately interrupted, and the higher-priority packet from group i begins service. The mean response time for a flow of priority group i under this HOL preemptive priority is given by [25]:

$$T_i = \frac{1}{\mu(1-\sigma_{i-1})(1-\sigma_i)}, \quad \text{where } \sigma_i = \sum_{j=1}^{i} \rho_j \tag{2}$$

As the equation indicates, the response time for group i depends on its utilization and the utilizations of all higher-priority groups, but is independent of all lower-priority traffic.

2.3. *Three power definitions*

Below, we introduce the three power performance metrics defined in [1], along with their corresponding power optimization results under the FCFS and HOL queueing disciplines.

2.3.1. *Individual power*

The **individual power** of flow i quantifies the flow's performance by considering both throughput and response time;[6] it is defined as the ratio of its utilization to its normalized mean response time:

$$P_i = \frac{\rho_i}{\mu T_i(\rho_i)} \tag{3}$$

Using the response time expressions under FCFS and HOL, the individual power can be written as follows:

- FCFS:

$$P_i = \frac{\rho_i}{\mu T_i} = \rho_i(1-\rho) = \rho_i(1-\alpha_i-\rho_i) \quad \text{where } \alpha_i = \sum_{j=1, j\neq i}^{n} \rho_j \tag{4}$$

- HOL:

$$P_i = \frac{\rho_i}{\mu T_i} = \rho_i(1-\sigma_i)(1-\sigma_{i-1}) = \rho_i(1-\sigma_{i-1}-\rho_i)(1-\sigma_{i-1}) \quad \text{where } \sigma_i = \sum_{j=1}^{i} \rho_j \tag{5}$$

Table 1
Singly optimized individual power results for FCFS and HOL in [1].

	FCFS	HOL
ρ_i^*	$\frac{1-\alpha_i}{2}$	$\frac{1-\sigma_{i-1}}{2}$
P_i^*	$\frac{(1-\alpha_i)^2}{4}$	$\frac{(1-\sigma_{i-1})^3}{4}$

Table 2
Jointly optimized individual power results at convergence for FCFS and HOL in [1]. The table lists ρ_i^*, ρ^*, the optimized individual power P_i^*, their sum $P_{\text{sum-of-optimals}} = \sum_{i=1}^{n} P_i^*$, and the corresponding limiting values.

	FCFS	HOL
ρ_i^*	$\frac{1}{n+1}$	$(\frac{1}{2})^i$
$\rho^* = \sum_{i=1}^{n} \rho_i^*$	$\frac{n}{n+1}$	$1-(\frac{1}{2})^n$
$\lim_{n\to\infty} \rho^*$	1	1
P_i^*	$\frac{1}{(n+1)^2}$	$2(\frac{1}{8})^i$
$P_{\text{sum-of-optimals}} = \sum_{i=1}^{n} P_i^*$	$\frac{n}{(n+1)^2}$	$\frac{2}{7}(1-(\frac{1}{8})^n)$
$\lim_{n\to\infty} P_{\text{sum-of-optimals}}$	0	$\frac{2}{7}$

In [1], both *singly* and *jointly* optimized individual power metrics are analyzed. In the *singly* optimized case, the individual power of flow i is maximized by selecting ρ_i^*, assuming the utilizations of all other flows are known and fixed. From [1], the resulting optimal utilization ρ_i^* and the corresponding optimized individual power P_i^* under FCFS and HOL are summarized in Table 1.

For the *jointly* optimized case, each flow iteratively adjusts its own utilization to maximize its individual power, assuming the utilizations of all other flows remain fixed in each iteration. This process continues iteratively until convergence to an equilibrium point, where no flow can further increase its power unilaterally—thus constituting a Nash equilibrium [26]. Table 2 summarizes the results of this joint optimization under FCFS and HOL, as derived in [1]. The table presents the equilibrium individual utilizations ρ_i^*, the total optimized system utilization ρ^*, the corresponding optimized individual powers P_i^*, and their sum $\sum P_i^*$. It also includes the limiting behavior of ρ^* and $\sum P_i^*$ as the number of flows, n, increases.

Under FCFS, the joint optimization result shows that each individual power P_i^* and their sum $\sum P_i^*$ approach zero as the number of flows n tends to infinity. This behavior resembles the well-known *tragedy of the commons* [27], where optimizing for individual gain leads to reduced performance for both individual flows and the system as a whole. In contrast, the joint optimization under HOL yields a total power sum that approaches $\frac{2}{7}$, while the optimized system utilization ρ^* also approaches 1—similar to the asymptotic behavior observed under FCFS.

2.3.2. *Sum of individual powers*

The **sum of individual powers**, denoted as P_{sum}, is defined as:

$$P_{\text{sum}} = \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} \frac{\rho_i}{\mu T_i} \tag{6}$$

This quantity represents the overall system performance in terms of power, calculated by summing the individual powers of all flows from 1 to n. Using the expressions for individual power in FCFS, namely Eq. (4), and HOL, namely Eq. (5), the corresponding sums of individual powers for FCFS and HOL are given below:

[6] In paper [22], it was noted that the individual power for a single flow has the pleasing property that the average number of users inside an individual power optimized system is exactly equal to 1. This satisfies the deterministic intuitive optimization that there should be exactly one customer in service and no customer in the queue in order to obtain the best balance between utilization and response time, which was discussed in [22].

Table 3
Results of optimizing the sum of individual powers, P^*_{sum}, in [1]. The table shows ρ^*_i and ρ^* that achieve the maximum sum of powers, along with P_i and P^*_{sum} and the limits of ρ^* and P^*_{sum} for both FCFS and HOL.

	FCFS	HOL
ρ^*_i	see Footnote 7	$\frac{1}{n+1}$
ρ^*	$\frac{1}{2}$	$\frac{n}{n+1}$
$\lim\limits_{n\to\infty} \rho^*$	$\frac{1}{2}$	1
P_i	see Footnote 7	$\frac{(n+1-i)(n+2-i)}{(n+1)^3}$
$P^*_{\text{sum}} = \sum_{i=1}^{n} P_i$	$\frac{1}{4}$	$\frac{n(n+2)}{3(n+1)^2}$
$\lim\limits_{n\to\infty} P^*_{\text{sum}}$	$\frac{1}{4}$	$\frac{1}{3}$

- FCFS:

$$P_{\text{sum}} = \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} \rho_i(1-\rho) = \rho(1-\rho) \tag{7}$$

- HOL:

$$P_{\text{sum}} = \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} \rho_i(1-\sigma_i)(1-\sigma_{i-1}) \quad \text{where } \sigma_i = \sum_{j=1}^{i} \rho_j \tag{8}$$

The optimization result for the sum of individual powers—obtained by finding the optimal set $\rho^*_1, \rho^*_2, \dots, \rho^*_n$—is summarized in Table 3. In the FCFS case, as shown in Theorem 5.1 in [1], optimizing the sum of individual powers is equivalent to optimizing for a single flow. As a result, the optimal total utilization is $\rho^* = 0.5$, and the corresponding maximum sum of powers is $P^*_{\text{sum}} = 0.25$. The individual utilization factors ρ^*_i are not uniquely determined as long as their sum is 0.5; therefore, the corresponding entries for ρ^*_i and P_i in the table are noted as a comment.[7] The optimal values of ρ^* and P^*_{sum} remain constant regardless of the number of flows n.

In contrast, in HOL, the optimal values vary with n. Optimizing the sum of individual powers in HOL results in an equal optimized utilization factor across all flows, with $\rho^*_i = \frac{1}{n+1}$. This leads to an optimized total utilization $\rho^* = \frac{n}{n+1}$, which approaches 1 as n approaches infinity. The corresponding maximum sum of individual powers also increases with n and approaches an asymptotic value of $\frac{1}{3}$, exceeding the optimal sum of powers in FCFS, which is $\frac{1}{4}$.

2.3.3. Average power

The **average power**, denoted P_{avg}, aggregates performance across flows by weighting each flow's normalized delay by its utilization share:

$$P_{\text{avg}} = \frac{\sum_{i=1}^{n} \rho_i}{\sum_{i=1}^{n} \left(\frac{\rho_i}{\rho} \mu T_i\right)} \tag{9}$$

For the average power analysis, we use an M/G/1 model with general service-time distributions. This is a significant generalization for this power metric. As shown in Theorem 6.2 from paper Part I [1], for an M/G/1 system with n flows using any work-conserving queueing discipline, the average power is equivalent to the power of a single aggregated flow served under FCFS. This equivalence implies that the average power remains invariant across all work-conserving disciplines, provided the assumptions of Theorem 6.2 hold.[8]

In other words, although the queueing discipline may change the order in which groups of packets are served, it does not affect the system's average power as long as it remains work-conserving. As a result, the optimal average power is the same across all such systems and is equal to 0.25—the optimal power of a single flow M/M/1 system. For this reason, we omit this metric from our subsequent analysis of extending power optimization beyond FCFS and HOL, since it remains constant under all work-conserving queueing disciplines.

3. Continuous spectrum of queueing disciplines from FCFS to HOL

To introduce additional control over response time and further extend power performance optimization, we move beyond FCFS and HOL and incorporate more flexible and comprehensive scheduling policies that span the continuous spectrum between these two extremes in its own way. Several approaches exist to transition between FCFS and HOL. One such known approach is the **delay-dependent priority** discipline [21], in which a flow's priority increases proportionally with its time in the system, with each group assigned a different rate of increase. Among the remaining other ways to span, we introduce and study the **beta-priority system** as an alternative framework for exploring their spectrum of flow priority discrimination.

The following sections provide a more detailed introduction to these two flexible queueing disciplines and their corresponding mean response time expressions, which together allow us to study the full range of flow prioritization—from minimal discrimination under FCFS to maximum discrimination under HOL.

3.1. The delay-dependent system

The delay-dependent system was introduced by Kleinrock [21] and uses a set of variable parameters, $\{b_i\}$ to provide the flexibility in adjusting the relative waiting time among different priority groups. The i^{th} priority group is assigned a number b_i, where $0 \le b_n \le b_{n-1} \le \dots \le b_2 \le b_1$. The priority of a packet from group i, denoted as $q_i(t)$, is a function of time that linearly increases with the time it stays in the system, using the scalar b_i, namely $q_i(t) = (t-\delta)b_i$, where δ is the time when the packet enters the system to wait for service and $t \ge \delta$. At any time t, the packet with the highest value of $q_i(t)$ is immediately taken into service preemptively. A larger value for b_i represents a higher growth rate for the i^{th} priority group, thus giving that group "higher" priority performance. This mechanism also ensures that packets waiting long enough will eventually be served, thereby preventing the starvation problem that occurs in strict head-of-the-line priority systems.[9]

In [21], Kleinrock derived a triangular set of equations to characterize the mean waiting time for each group i. For the preemptive case—and with a modified priority notation in which lower indices correspond to higher priority groups—the mean waiting time is as follows:

$$W_i = \frac{\frac{W_0}{1-\rho} + \sum_{j=1}^{i-1} \frac{\rho_j}{\mu_i}[1 - \frac{b_i}{b_j}] - \sum_{j=i+1}^{n} \frac{\rho_j}{\mu_j}[1 - \frac{b_j}{b_i}] - \sum_{j=i+1}^{n} \rho_j W_j [1 - \frac{b_j}{b_i}]}{1 - \sum_{j=1}^{i-1} \rho_j [1 - \frac{b_i}{b_j}]} \qquad i = 1, 2, \dots, n \tag{10}$$

where

$$W_0 = \sum_{i=1}^{n} \frac{\lambda_i \overline{x_i^2}}{2} \tag{11}$$

W_0 is the average residual service time experienced by an arriving packet. Here, the random variable $\tilde{x}_i$ denotes the service time for flow i, $\overline{x_i} = \frac{1}{\mu_i}$ is the first moment of $\tilde{x}_i$, and $\overline{x_i^2}$ is the second moment of

[7] Any set $(\rho^*_1, \rho^*_2, \dots, \rho^*_n)$ that sums to 0.5 is optimal. The corresponding individual power values $P_1, P_2, \dots, P_n$ for this set sum to the optimal total sum of powers P^*_{sum} of 0.25.

[8] Namely, that the first and second moments of the service time are identical across all flows.

[9] Note that the priority order used here is reversed compared to that used by Kleinrock in [21]. Kleinrock used a higher index to represent a higher priority group, while we use a lower index to indicate higher priority.

$\tilde{x}_i$. We note that only the ratios of $\frac{b_i}{b_j}$ appear in Eq. (10), and thus the delay-dependent system introduces only $n-1$ independent parameter ratios.[10]

In this paper, we focus on the M/M/1 model and assume that all flows have exponential service time distributions, with the same mean service time $\overline{x_i} = \frac{1}{\mu_i} = \frac{1}{\mu}$ for $i = 1, \dots, n$. Under this assumption, the second moment is also common across all flows, given by $\overline{x_i^2} = \frac{2}{\mu^2}$.[11] Substituting into the expression for W_0, we obtain:

$$W_0 = \sum_{i=1}^{n} \frac{\lambda_i \overline{x_i^2}}{2} = \sum_{i=1}^{n} \frac{\lambda_i \frac{2}{\mu^2}}{2} = \frac{\frac{\lambda}{\mu}}{\mu} = \frac{\rho}{\mu} \tag{12}$$

We now apply this result to the triangular set of equations for the delay-dependent priority system, Eq. (10), to derive explicit expressions for the mean response times in the case of two flows ($n = 2$). The mean response time T_i is the sum of the average waiting time W_i and the average service time $\frac{1}{\mu}$:

$$T_i = W_i + \frac{1}{\mu} \tag{13}$$

Using this relationship, we derive the following expressions:

- For T_2:

$$\begin{aligned} T_2 &= \frac{1}{\mu} + \frac{\frac{W_0}{1-\rho} + \frac{\rho_1}{\mu}(1 - \frac{b_2}{b_1})}{1 - \rho_1(1 - \frac{b_2}{b_1})} = \frac{\frac{W_0}{1-\rho} + \frac{1}{\mu}}{1 - \rho_1(1 - \frac{b_2}{b_1})} = \frac{\frac{\rho}{\mu(1-\rho)} + \frac{1}{\mu}}{1 - \rho_1(1 - \frac{b_2}{b_1})} \\ &= \frac{1}{\mu(1-\rho)[1 - \rho_1(1 - \frac{b_2}{b_1})]} \end{aligned}$$

- For T_1:

$$\begin{aligned} T_1 &= \frac{1}{\mu} + \frac{W_0}{1-\rho} - \frac{\rho_2}{\mu}(1 - \frac{b_2}{b_1}) - \rho_2 W_2 (1 - \frac{b_2}{b_1}) \\ &= \frac{1 - \rho(1 - \frac{b_2}{b_1})}{\mu(1-\rho)[1 - \rho_1(1 - \frac{b_2}{b_1})]} = T_2\,[1 - \rho\,(1 - \frac{b_2}{b_1})] \end{aligned}$$

To simplify the notation, we define:

$$k = 1 - \frac{b_2}{b_1} \qquad (0 \le k \le 1) \tag{14}$$

Then, the mean response times for the two flows can can be written as:

$$T_1 = \frac{1 - k\rho}{\mu(1-\rho)(1-k\rho_1)}, \quad T_2 = \frac{1}{\mu(1-\rho)(1-k\rho_1)} \tag{15}$$

When $b_1 = b_2$, packets from group 1 and group 2 gain priority at the same rate. As a result, the packet that arrives earlier is served first, regardless of which priority group it belongs to. This behavior is equivalent to the first-come, first-served (FCFS) system. In this case, the parameter k is 0, and the response times simplify to those in the FCFS case:

$$T_1 = T_2 = \frac{1}{\mu(1-\rho)}$$

When $b_1 \gg b_2$, the priority of packets from group 1 becomes effectively infinite compared to that of packets from group 2 present in the system at the time of arrival. As a result, packets from group 1 are always served before those from group 2. This behavior aligns with the head-of-the-line (HOL) preemptive discipline, where a newly arrived packet from group 1 always has higher priority than any packet from group 2 already in the system. In this case, the priority of group 1 is strictly greater than that of group 2. As $\frac{b_2}{b_1} \to 0$, the parameter k approaches 1, and the response times converge to those of the HOL case:

$$T_1 = \frac{1}{\mu(1-\rho_1)}, \quad T_2 = \frac{1}{\mu(1-\rho_1)(1-\rho)}$$

The general equation for T_i with arbitrary n is derived from Eqs. (10), (12), and (13) above, yielding:

$$T_i = \frac{\frac{1}{\mu(1-\rho)} - \sum_{j=i+1}^{n} \rho_j \left(1 - \frac{b_j}{b_i}\right) T_j}{1 - \sum_{j=1}^{i-1} \rho_j \left(1 - \frac{b_i}{b_j}\right)} \qquad i = 1, \dots, n. \tag{16}$$

For example, for $n = 3$, we get

$$T_3 = \frac{1}{\mu(1-\rho)\left[1 - \rho_1\left(1 - \frac{b_3}{b_1}\right) - \rho_2\left(1 - \frac{b_3}{b_2}\right)\right]}$$

$$\begin{aligned} T_2 &= \frac{\frac{1}{\mu(1-\rho)} - \rho_3\left(1 - \frac{b_3}{b_2}\right)T_3}{1 - \rho_1\left(1 - \frac{b_2}{b_1}\right)} \\ &= \frac{1}{\mu(1-\rho)} \frac{1 - \rho_1\left(1 - \frac{b_3}{b_1}\right) - \rho_2\left(1 - \frac{b_3}{b_2}\right) - \rho_3\left(1 - \frac{b_3}{b_2}\right)}{\left[1 - \rho_1\left(1 - \frac{b_2}{b_1}\right)\right]\left[1 - \rho_1\left(1 - \frac{b_3}{b_1}\right) - \rho_2\left(1 - \frac{b_3}{b_2}\right)\right]} \end{aligned}$$

$$\begin{aligned} T_1 &= \frac{1}{\mu(1-\rho)} - \rho_2\left(1 - \frac{b_2}{b_1}\right)T_2 - \rho_3\left(1 - \frac{b_3}{b_1}\right)T_3 \\ &= \frac{\begin{gathered}\left[1 - \rho_1(1 - \frac{b_2}{b_1})\right]\left[1 - \rho_1(1 - \frac{b_3}{b_1}) - \rho_2(1 - \frac{b_3}{b_2})\right] \\ -\rho_2\left(1 - \frac{b_2}{b_1}\right)\left[1 - \rho_1\left(1 - \frac{b_3}{b_1}\right) - (\rho_2 + \rho_3)\left(1 - \frac{b_3}{b_2}\right)\right] \\ -\rho_3\left(1 - \frac{b_3}{b_1}\right)\left[1 - \rho_1\left(1 - \frac{b_2}{b_1}\right)\right]\end{gathered}}{\mu(1-\rho)\left[1 - \rho_1\left(1 - \frac{b_2}{b_1}\right)\right]\left[1 - \rho_1\left(1 - \frac{b_3}{b_1}\right) - \rho_2\left(1 - \frac{b_3}{b_2}\right)\right]} \end{aligned}$$

We have not solved for T_i in Eq. (16) for $n > 3$ since it becomes unwieldy. However, for the case of arbitrary n, we see that the scenario where $b_1 = b_2 = \cdots = b_n$ corresponds to the classical FCFS case, as all groups gain priority at the same rate. This means the priority is solely based on the time spent in the system, resulting in a first-come, first-served order. Conversely, when $b_1 \gg b_2 \gg \cdots \gg b_n$, the system approaches the behavior of the classical HOL discipline. In this scenario, flow 1 has the highest priority whenever it enters the system because its priority increase rate is very large compared to the others. Flow 2 has the next highest priority, followed by flow 3, and so on. This significant difference in priority growth rates establishes a strict hierarchy among the flows. Consequently, flow 1 is always served before flow 2, flow 2 before flow 3, and so on for the remaining flows.

3.2. The beta-priority system

In addition to the delay-dependent system, we now introduce a second approach to span its spectrum from FCFS to HOL: the **beta-priority system**. The idea behind this system is to model the response time as a weighted average of the FCFS and HOL response times, controlled by a parameter $\beta \in [0, 1]$, namely,

$$T = \beta \cdot T_{HOL} + (1-\beta) \cdot T_{FCFS} \tag{17}$$

Substituting the expressions for the response times under FCFS and HOL, the mean response time of the flow i is:

$$T_i = \beta \cdot \frac{1}{\mu(1-\sigma_i)(1-\sigma_{i-1})} + (1-\beta) \cdot \frac{1}{\mu(1-\rho)} \quad \text{for } i = 1, \dots, n \tag{18}$$

In this formulation, the response time follows head-of-the-line (HOL) behavior for a fraction β of the time, and first-come, first-served (FCFS) behavior for a fraction $1-\beta$. This structure allows the beta-priority system to transition between FCFS ($\beta = 0$) and HOL ($\beta = 1$). By varying β

[10] There are n^2 ratios, $\frac{b_i}{b_j}$, for $i = 1, .., n$ and $j = 1, .., n$. Theses n^2 ratios can be simplified to $n-1$ independent ratios as $\frac{b_i+1}{b_i}$ for $i = 1, .., n-1$ since knowing these $n-1$ ratios allows us to calculate every $\frac{b_i}{b_j}$ ratio based on the chain rule.

[11] For an exponential service time distribution with mean $\frac{1}{\mu}$, the variance is $\frac{1}{\mu^2}$, resulting in a second moment of $\frac{2}{\mu^2}$.

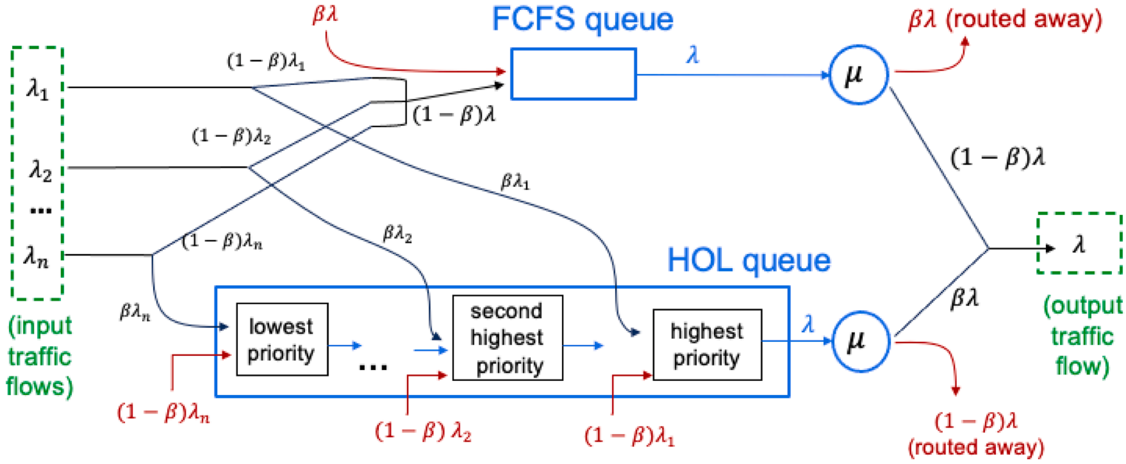


Fig. 2. An example of implementing the beta-priority system. Each input flow i splits its traffic between the FCFS and HOL queues in portions $(1-\beta)$ and β. The red traffic represents externally introduced traffic and is used to achieve the average response time as Eq. (17).

from 0 to 1, we capture a full spectrum of flow priority discrimination—from minimal discrimination in FCFS to maximal discrimination in HOL. Unlike the delay-dependent system, which introduces $n-1$ ratios, $\frac{b_{i+1}}{b_i}$ (for $i = 1, .., n-1$), to span the full spectrum of flow discrimination from FCFS to HOL, the beta-priority system can achieve this using a single variable, β. This simplification reduces analytical complexity by providing a one-dimensional control over the degree of flow priority discrimination.

To implement the beta-priority system, we adopt the structure illustrated in Fig. 2. Each input flow splits into two portions: a fraction β is sent to the HOL queue, and the remaining fraction $(1-\beta)$ is sent to the FCFS queue. Within the HOL queue, traffic is sorted in priority order, with higher-priority flows at the front and lower-priority flows at the back. To ensure each flow segment experiences the response time precisely defined by Eq. (18), we introduce external traffic, depicted in red in the Fig. 2. While this traffic is subsequently routed away from our model, its presence is essential for accurately creating the response time according to Eq. (18).

In the FCFS queue, we insert red external traffic, $\beta\lambda$ and combine it with the $(1-\beta)\lambda$ traffic from the input traffic flow to result a total flow of λ entering the FCFS queue; this results in an average response time given by Eq. (1), $T_{FCFS} = \frac{1}{\mu(1-\rho)}$. In the HOL queue, each priority group i receives external traffic in the amount of $(1-\beta)\lambda_i$ for $i = 1, \ldots, n$, resulting in the response time for each flow i following Eq. (2), along with each $\beta\lambda_i$ input traffic.

With the red external traffic, even though the input traffic flow is split, each portion experiences the response time as if the full traffic amount λ were processed by either the FCFS or HOL queues. The β portion directed to the HOL queue experiences the response time give by Eq. (2), while the $(1-\beta)$ portion directed to the FCFS queue follows the response time described by Eq. (1). As a result, the beta-priority system is implemented with the average response time of this input flow given by Eq. (18) for each flow.

To explore system behavior under varying degrees of priority discrimination, we now present an example. Fig. 3 illustrates the average normalized response time, $\frac{1}{\mu T_i(\rho)}$, versus ρ for $n = 5$ flows, all having equal utilization $\rho_i = \rho/n$. The mean response times are computed using Eq. (18) and normalized by the no-load response time, $1/\mu$. Curves are shown for $\beta = 0, 0.25, 0.5, 0.75, 1$, capturing the progression from FCFS ($\beta = 0$) to HOL ($\beta = 1$) and illustrating the effect of increasing priority discrimination.

As β increases, the degree of flow discrimination becomes more pronounced, as reflected in the widening separation among the response time curves. In the FCFS case ($\beta = 0$), all flows experience identical response time—represented by the overlapping curves and the black dotted line. However, for $\beta > 0$, flows begin to experience differentiated service based on their priority level, and the level of differentiation grows progressively with increasing β. This effect is clearly visible in Fig. 3, where each flow has equal utilization but exhibits increasingly unequal response times. The divergence among the normalized response time curves illustrates the growing influence of priority in the system.

In addition, Fig. 3 reveals that the response times of lower-priority flows begin to grow rapidly at lower values of ρ, well before the system becomes heavily loaded. This effect becomes most evident when $\beta = 1$, where higher-priority flows (e.g., flows 1 and 2) continue to maintain low response times even as ρ increases, while lower-priority flows (e.g., flows 4 and 5) experience growing delays beginning at relatively low values of ρ. For clarity in visualization, the y-axis is truncated at 6, though actual response times for lower-priority flows under high load can grow significantly beyond this threshold. It is interesting to note that as long as $\beta < 1$, each flow retains some first-come-first-served component in its response time, and therefore none of them can demonstrate finite response time when $\rho \geq 1$. However, for $\beta = 1$, there is no first-come-first-served component, so the higher-priority flows (i.e., flows 1 and 2) can demonstrate finite response time when $\rho \geq 1$.[12]

Similar to the delay-dependent system, we examine the simple case of two flows $n = 2$. The mean response times for flow 1 and flow 2 under the beta-priority system are given by:

$$\begin{aligned} T_1 &= \frac{\beta}{\mu(1-\rho_1)} + \frac{1-\beta}{\mu(1-\rho)} \\ T_2 &= \frac{\beta}{\mu(1-\rho_1)(1-\rho)} + \frac{1-\beta}{\mu(1-\rho)} \end{aligned} \tag{19}$$

When $\beta = 0$, the system behaves like the FCFS system; when $\beta = 1$, it behaves like the HOL system. For $0 < \beta < 1$, the system exhibits

[12] An additional observation from Fig. 3 follows from the conservation law [28], which states that $\sum_{i=1}^{n} \rho_i W_i = \frac{\rho W_0}{1-\rho}$. Under our M/M/1 model and the assumption of equal service time distribution across flows, we have $W_0 = \frac{\rho}{\mu}$, yielding $\sum_{i=1}^{n} \rho_i W_i = \frac{\rho^2}{\mu(1-\rho)}$, a quantity that depends only on μ and ρ, and is independent of the scheduling discipline. Similarly, $\sum_{i=1}^{n} \rho_i T_i = \sum_{i=1}^{n} \rho_i \left(W_i + \frac{1}{\mu}\right) = \sum_{i=1}^{n} \rho_i W_i + \frac{\rho}{\mu} = \frac{\rho^2}{\mu(1-\rho)} + \frac{\rho}{\mu} = \frac{\rho}{\mu(1-\rho)}$. After normalizing by $\frac{1}{\mu}$, the utilization-weighted average response time, given by $\sum_{i=1}^{n} \rho_i \cdot \mu T_i$, becomes a function of ρ alone. Therefore, in each plot, a vertical line at a fixed ρ intersects the five flow response time curves at points whose utilization-weighted average corresponds to the FCFS response time. In Fig. 3, for the same ρ, the response times of individual flows vary with different values of β, but their utilization-weighted averages remain identical.

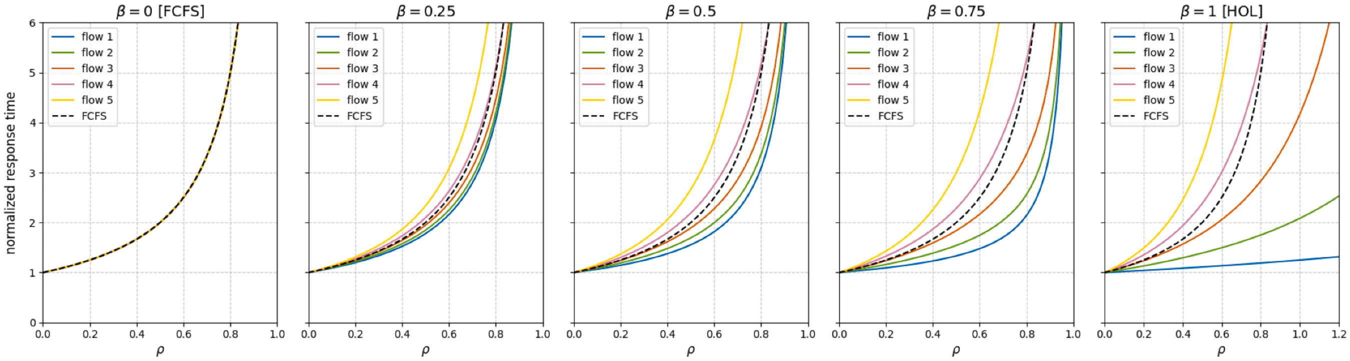


Fig. 3. Normalized response times $\frac{1}{\mu T_i(\rho)}$ versus ρ for $n = 5$ flows, each with equal utilization $\rho_i = \rho/n$. Curves are shown for $\beta = [0, 0.25, 0.5, 0.75, 1]$, illustrating the transition from FCFS ($\beta = 0$) to HOL ($\beta = 1$). The black dotted curve represents the FCFS baseline where all flows experience identical delays. As β increases, response times diverge across flows, with lower-priority flows (e.g., flows 4 and 5) exhibiting rapidly increasing delays—especially as ρ approaches 1. The y-axis is truncated at 6 to aid visualization.

Table 4
Response Times for Two Flows Across Various Queueing Disciplines.

Discipline	Flow 1 Response Time	Flow 2 Response Time
Delay-Dependent System $k = (1 - \frac{b_2}{b_1})$	$T_1 = \frac{1-k\rho}{\mu(1-\rho)(1-k\rho_1)}$	$T_2 = \frac{1}{\mu(1-\rho)(1-k\rho_1)}$
Beta-Priority System	$T_1 = \frac{\beta}{\mu(1-\rho_1)} + \frac{1-\beta}{\mu(1-\rho)}$	$T_2 = \frac{\beta}{\mu(1-\rho_1)(1-\rho)} + \frac{1-\beta}{\mu(1-\rho)}$

Table 5
Parameters for k and β that make the system equivalent to FCFS or HOL scheduling, which are the lower and upper bounds.

Discipline	First-Come-First-Served ($T_1 = T_2 = \mu\frac{1}{1-\rho}$)	Head-of-the-Line Priority ($T_1 = \frac{1}{\mu(1-\rho_1)}$; $T_2 = \frac{1}{\mu(1-\rho_1)(1-\rho)}$)
Delay-Dependent System	$k = 0$ ($b_1 = b_2$)	$k = 1$ ($b_1 \gg b_2$)
Beta-Priority System	$\beta = 0$	$\beta = 1$

intermediate levels of flow discrimination, transitioning between the two extremes.

3.3. Summary for $n = 2$: response times for two spectrum-spanning queueing systems

Table 4 presents the response times of two spectrum-spanning flows ($n = 2$) under the two priority systems introduced above. Each system spans a spectrum from FCFS to HOL by adjusting a single parameter—k in the delay-dependent system[13] and β in the beta-priority system. By varying k and β from 0 to 1, we capture the entire range of flow priority discrimination, from minimal discrimination in FCFS to maximal discrimination in HOL.

In Table 5, we summarize the specific conditions under which each system reduces to either FCFS or HOL. These two disciplines represent the lower and upper bounds of the parameters k and β, respectively. The transition between the two extremes is governed by varying these parameters across the interval $[0, 1]$.

3.4. Difference between two spectrum-spanning queueing systems

As noted above, there are multiple trajectories that each span the priority discrimination range from FCFS to HOL. Here, we focus on two trajectories: the *delay-dependent* system and the *beta-priority* system. Our objective in this paper is to identify target utilizations that optimize the power metrics; accordingly, we do not detail every aspect of these trajectories. Instead, we highlight one salient distinction between them, namely the ordering of service within each priority class (**intra-class priority ordering**).

For the delay-dependent system, arrivals within the same priority class are served in *first-come-first-serve* order. For the beta-priority system, intra-class first-come-first-serve ordering is not guaranteed. In the implementation of Fig. 2 (with an FCFS queue and an HOL queue and probabilistic server selection), an earlier packet may join the FCFS queue while a later packet joins the HOL queue. If both packets belong to the same highest-priority class, the FCFS packet must wait behind earlier arrivals (including lower-priority ones), whereas the packet that joins the HOL queue can move to the highest-priority queue because ordering there is by priority. Since the server selects the next packet probabilistically across the two queues, it may choose the HOL queue, in which case the later arrival (which joins the HOL queue) is served before the earlier one (which joins the FCFS queue).

In contrast, in the delay-dependent system, a packet's service priority increases with its time in the system, scaled by a *class-specific coefficient*. Packets in different classes therefore accumulate priority at different rates, whereas packets within the same class accrue at the *same* rate. Consequently, intra-class ordering reduces to *first-come-first-serve*: among packets in the same priority class, the one that has waited longer (i.e., arrived earlier) attains the higher priority and is served first in its class.

This example illustrates the key difference of these two trajectories: the delay-dependent system preserves first-come-first-serve within a priority class, while the beta-priority system can break intra-class first-come-first-serve ordering.

[13] In the delay-dependent system, flow discrimination can be represented by a single parameter only in the two-flow case. When the number of flows $n > 2$, additional parameters are required to capture the full range of flow priority discrimination. In contrast, the beta-priority system requires only a single parameter, β, to achieve this for any number of flows n.

4. Extending the analysis of the power metrics: The continuum from FCFS to HOL for n = 2

In Section 3, we introduced two families of queueing disciplines that span the full range of priority discrimination. Now, we proceed to perform power optimization using these two families for the simplified case of two flows ($n = 2$). In the following subsections, we will choose the utilizations to optimize the two power metrics for each of the two queueing discipline families. Specifically, we will only apply to the individual power and to the sum of powers metric. The third average power metric is not included in the analysis since Theorem 6.3 in [1] showed that there is no difference in the optimization result for the third power metric for all conservative queueing disciplines.

4.1. Optimizing power metric 1: Individual power

First, we consider our first power metric—individual power. Specifically, we focus on the jointly optimized case, where each flow aims to select its utilization to optimize its own power in equilibrium. In the following, we investigate the joint individual power optimization for two flows $n = 2$ in the delay-dependent system and the beta-priority system, respectively. We aim to find the optimal average utilization of each flow, $\rho_1^*, \rho_2^*, .., \rho_n^*$ that jointly optimize each individual power $P_1^*, P_2^*, .., P_n^*$ in the equilibrium.

4.1.1. The delay-dependent system

The response time for two flows in the delay-dependent system is given by Eq. (15), allowing us to calculate the individual power, where recall $k = 1 - \frac{b_2}{b_1}$ as follows:

$$P_1 = \frac{\rho_1}{\mu T_1} = \frac{\rho_1(1-\rho)(1-k\rho_1)}{1-k\rho} \tag{20}$$

$$P_2 = \frac{\rho_2}{\mu T_2} = \rho_2(1-\rho)(1-k\rho_1) \tag{21}$$

To find the optimal utilizations, we set the partial derivative of Eq. (20) with respect to ρ_1, and the partial derivative of Eq. (21) with respect to ρ_2 equal to 0 to obtain:[14]

$$\begin{aligned}\frac{\partial P_1}{\partial \rho_1} &= \frac{[(1-\rho-\rho_1)(1-k\rho_1) - k\rho_1(1-\rho)]\cdot(1-k\rho) + k\rho_1(1-\rho)(1-k\rho_1)}{(1-k\rho)^2} \\ &= 0\end{aligned} \tag{22}$$

$$\frac{\partial P_2}{\partial \rho_2} = (1-k\rho_1)(1-\rho-\rho_2) = 0 \tag{23}$$

By Eq. (23), either $(1-k\rho_1) = 0$ or $(1-\rho-\rho_2) = 0$. Stability requires $k\rho_1 < 1$ (and in particular $\rho_1 < 1$ and $0 \le k \le 1$), so $1 - k\rho_1 > 0$; thus the admissible root is $1 - \rho - \rho_2 = 0$. Using $\rho = \rho_1 + \rho_2$,

$$1 - (\rho_1 + \rho_2) - \rho_2 = 0 \implies 1 - \rho_1 - 2\rho_2 = 0 \implies \rho_2 = \frac{1-\rho_1}{2}$$

This indicates that flow2's power is maximized when it takes half of the remaining utilization after flow 1 has taken its share—a result also shown in Part I paper [1]. This holds not only for the two boundary cases, FCFS and HOL, but also for all intermediate queueing disciplines within the delay-dependent class. This leads to the following theorem:

Theorem 4.1. *In an M/M/1 system with two flows under a* ***delay-dependent*** *queueing discipline, the individual power of the lower-priority flow (flow 2) is maximized when ρ_2 equals half of the remaining utilization after the higher-priority flow (flow 1) has taken its share. That is,*

$$\rho_2^* = \frac{1-\rho_1}{2} \tag{24}$$

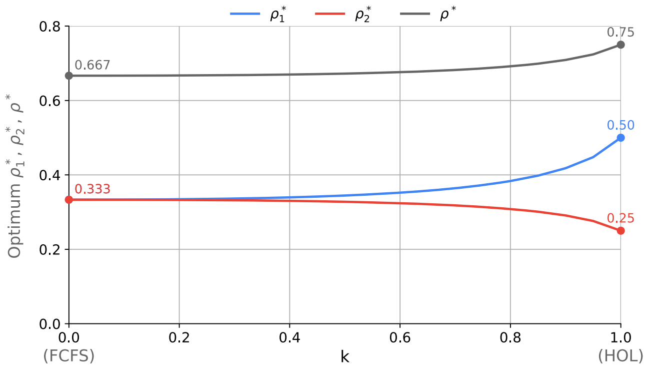


Fig. 4. Optimal values ρ_1^*, ρ_2^*, and ρ^* as functions of k, under joint optimization of "individual power" for both flow 1 and flow 2 in an M/M/1 system with two flows ($n = 2$) using a delay-dependent queueing discipline. As k increases, ρ_1^*, ρ^*, and the difference between ρ_1^* and ρ_2^* increase, while ρ_2^* decreases.

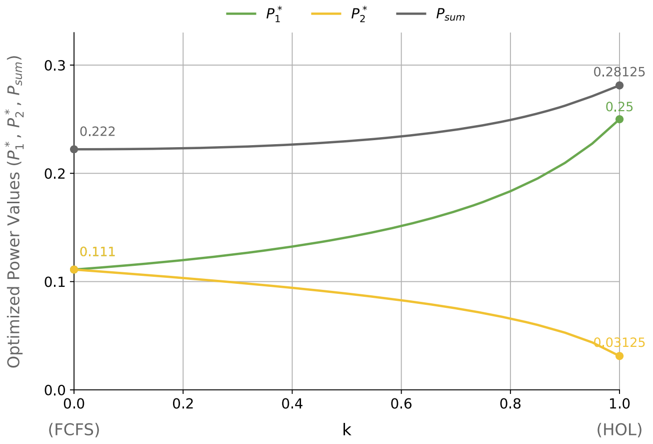


Fig. 5. Optimized power values of P_1^*, P_2^*, P_{sum} versus k, under joint optimization of "individual power" for both flow 1 and flow 2 in an M/M/1 system with two flows ($n = 2$) using a delay-dependent queueing discipline. Similar to the behavior of utilizations, as k increases, P_1^*, P_{sum}, and the difference $(P_1^* - P_2^*)$ increase, while P_2 decreases.

The theorem follows by setting $\partial P_2/\partial\rho_2 = 0$ in Eq. (23) and solving for ρ_2, as shown above.

Having determined that the optimal utilization for flow 2 is half of the remaining capacity after flow 1, we now proceed to find the value of ρ_1 that maximizes flow1's individual power. To do so, we solve the equation obtained by taking the partial derivative of P_1 with respect to ρ_1, as shown in Eq. (22). Setting the numerator of that expression to zero gives the simplified form:

$$(1-k\rho)(1-2\rho_1-\rho_2)(1-k\rho_1) + k^2\rho_1\rho_2(1-\rho) = 0 \tag{25}$$

Substituting the result from Eq. (24), namely $\rho_2^* = \frac{1-\rho_1}{2}$, into Eq. (25) yields a cubic equation in ρ_1 of the form:

$$-2k^2{\rho_1}^3 + (9k - 4k^2){\rho_1}^2 + (2k^2 - 6)\rho_1 + (2 - k) = 0 \tag{26}$$

By numerically solving for the root of the cubic equation in Eq. (26) within the interval $[0, 1]$, we obtain the value of ρ_1^* that maximizes flow1's individual power for each value of k, as shown in Fig. 4. The corresponding values of ρ_2^*, derived from Eq. (24) as $\rho_2^* = \frac{1-\rho_1}{2}$ based on the solved ρ_1^*, along with the total utilization $\rho^* = \rho_1^* + \rho_2^*$, are also presented.

In Fig. 4, the two extreme cases—FCFS and HOL—are marked on the curves of ρ_1^*, ρ_2^*, and total utilization ρ^* as functions of k. Specifically, at the left bound ($k = 0$, corresponding to FCFS), we have $\rho_1^* = \rho_2^* = \frac{1}{3}$ and $\rho^* = \frac{2}{3}$. At the right bound ($k = 1$, corresponding to HOL), we have $\rho_1^* =$

[14] Similar to Part I [1], we use an asterisk (*) to denote optimized values.

0.5, $\rho_2^* = 0.25$ and $\rho^* = 0.75$. The area between these bounds represents the range of optimal operating points for ρ_1^*, ρ_2^*, and ρ^* as k varies from 0 to 1, reflecting the transition from FCFS to HOL.

From the figure, we observe that when k increases, ρ_1^* increases while ρ_2^* decreases. This trend arises because increasing k reflects greater flow discrimination in favor of the higher-priority flow. As flow 1 receives more prioritized treatment, its delay decreases, enabling it to maintain the same response time at a higher utilization—or achieve a lower response time under the same utilization. As a result, the optimized value of ρ_1^* increases with k when maximizing its power. Consequently, less capacity remains for flow 2, reducing its optimal ρ_2^* as k increases, which is given by $\rho_2^* = \frac{1-\rho_1^*}{2}$. In addition, both the sum $\rho_1^* + \rho_2^*$ and the difference $\rho_1^* - \rho_2^*$ increase as k increases. This reflects a growing overall system utilization. However, it also results in a larger disparity in utilization between the two flows, indicating reduced fairness in resource allocation between flow 1 and flow 2.

Fig. 5 presents the maximized individual power for flow 1 and flow 2, computed using the optimized utilization values ρ_1^* and ρ_2^* from Fig. 4. In other words, by substituting the optimized individual ρ_1^* and ρ_2^* from Fig. 4 into the individual power equation expressions, namely, Eq. (20) for flow 1 and Eq. (21) for flow 2, we obtain the results shown in Fig. 5. At $k = 0$ (the least discriminatory case, corresponding to FCFS), both flows achieve equal optimized individual power: $P_1^* = P_2^* = \frac{1}{9} \approx 0.111$, resulting in the sum of individual powers $P_{\text{sum}} = \frac{2}{9} \approx 0.222$. At $k = 1$ (the most discriminatory case, corresponding to HOL), the optimized individual powers are $P_1^* = \frac{1}{4} = 0.25$ and $P_2^* = \frac{1}{32} = 0.03125$, yielding a total power $P_{\text{sum}} = \frac{9}{32} = 0.28125$. These two results were previously reported in [1].

The delay-dependent system allows us to explore the full spectrum of operating points between these two extremes and quantify the tradeoff between discrimination and overall performance. Similar to the behavior of ρ_1^*, ρ_2^* and ρ^*, Fig. 5 shows that P_1^* increases with k, while P_2^* decreases. In addition, both the sum and the difference of P_1^* and P_2^* become larger as flow discrimination, represented by k, increases. This is evident from the rising curve of P_{sum}, as well as the widening gap between P_1^* and P_2^*. Combined with the growing disparity between ρ_1^* and ρ_2^*, larger values of k lead to increased discrimination in utilization, but also yield greater overall performance in terms of the sum of individual powers.

4.1.2. The beta-priority system

In the beta-priority system, the response time for flow 1 and flow 2 are given in Eq. (15). Using the individual power definition from Eq. (3), the corresponding individual power values for flows 1 and 2 are computed as follows:

$$P_1 = \frac{\rho_1}{\mu T_1} = \frac{\rho_1}{\frac{\beta}{(1-\rho_1)} + \frac{1-\beta}{(1-\rho)}} = \frac{\rho_1(1-\rho_1)(1-\rho)}{\beta(1-\rho) + (1-\beta)(1-\rho_1)} = \frac{\rho_1(1-\rho_1)(1-\rho)}{1-\rho_1-\beta\rho_2} \tag{27}$$

$$P_2 = \frac{\rho_2}{\mu T_2} = \frac{\rho_2}{\frac{\beta}{(1-\rho_1)(1-\rho)} + \frac{1-\beta}{(1-\rho)}} = \frac{\rho_2(1-\rho_1)(1-\rho)}{\beta + (1-\beta)(1-\rho_1)} = \frac{\rho_2(1-\rho_1)(1-\rho)}{1-(1-\beta)\rho_1} \tag{28}$$

Taking the partial derivative of P_2 (Eq. (28)) with respect to ρ_2 and setting it equal to zero, we factor out the term that is independent of ρ_2 and proceed with the computation:

$$\frac{\partial P_2}{\partial \rho_2} = \frac{1-\rho_1}{1-(1-\beta)\rho_1} \cdot \frac{\partial \rho_2(1-\rho)}{\partial \rho_2} = \frac{1-\rho_1}{1-(1-\beta)\rho_1} \cdot (1-\rho-\rho_2) = 0 \tag{29}$$

Since $\rho_1 < 1$ and $\beta \in [0,1]$, the prefactor $\frac{1-\rho_1}{1-(1-\beta)\rho_1} > 0$ is strictly positive and cannot be zero; therefore $(1-\rho) - \rho_2 = 0$. With $\rho = \rho_1 + \rho_2$, this gives $1 - \rho_1 - 2\rho_2 = 0$, hence

$$\rho_2^* = \frac{1-\rho_1}{2} \tag{30}$$

Interestringly, this result matches the one obtained in Eq. (24) for the delay-dependent system. It also leads to the following theorem:

Theorem 4.2. *In an M/M/1 system with two flows using a **beta-priority** system as the queueing discipline, the individual power of the lower-priority flow is maximized when ρ_2 is equal to half of the remaining utilization after accounting for the higher-priority flow, that is,*

$$\rho_2^* = \frac{1-\rho_1}{2} \tag{31}$$

Now taking the partial derivative of P_1 (Eq. (27)) with respect to ρ_1 and setting it to zero, we get:

$$\frac{\partial P_1}{\partial \rho_1} = \frac{[(1-2\rho_1)(1-\rho) - \rho_1(1-\rho_1)](1-\rho_1-\beta\rho_2) + \rho_1(1-\rho_1)(1-\rho)}{(1-\rho_1-\beta\rho_2)^2} = 0 \tag{32}$$

Plugging Eq. (31) into the numerator of Eq. (32) and solving the equation, we have the equilibrium point for the optimized ρ_1^* that optimizes the individual power of flow 1:

$$\rho_1^* = \frac{2-\beta}{2(3-2\beta)} \tag{33}$$

Substituting Eq. (33) back to Eq. (31), we have the equilibrium point for the optimized ρ_2^* as a function of β:

$$\rho_2^* = \frac{4-3\beta}{4(3-2\beta)} \tag{34}$$

and the total optimized utilization:

$$\rho^* = \rho_1^* + \rho_2^* = \frac{8-5\beta}{4(3-2\beta)} \tag{35}$$

The plot of ρ_1^*, ρ_2^*, and ρ^* versus β ranging from 0 to 1 is presented in Fig. 6. The left bound where $\beta = 0$ represents the FCFS case with minimal flow discrimination, results in equal utilization for flow 1 and flow 2, $\rho_1^* = \rho_2^* = \frac{1}{3}$, and thus $\rho^* = \frac{2}{3}$. The right bound where $\beta = 1$ represents the HOL case with maximal flow discrimination, resulting in the highest ρ^* among all queueing disciplines within the beta-priority system, with $\rho_1^* = 0.5$, $\rho_2^* = 0.25$, and $\rho^* = 0.75$. The trend of curves in the figure resembles that in Fig. 4, where ρ_1^* increases and ρ_2^* decreases as the level of flow discrimination, represented by β, increases. Moreover, both the sum of ρ_1^* and ρ_2^* as well as the difference between ρ_1^* and ρ_2^* grow with increase of β, this point is the same as in the delay-dependent system. As the value of β increases, it indicates a higher probability that flow 1 can cut in line ahead of flow2's packets, thereby reducing the waiting time for flow 1 and leading to a higher level of flow discrimination.

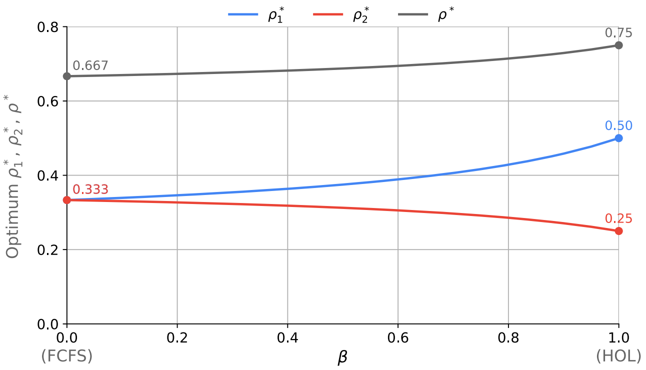


Fig. 6. Optimal values of ρ_1^*, ρ_2^*, and ρ^* versus β, derived from individual power optimization in an M/M/1 system with two flows using a beta-priority queueing discipline.

Greater flow discrimination results in a higher ρ_1^* for optimizing individual power and consequently a lower ρ_2^*, but a higher sum and difference of ρ_1^* and ρ_2^*.

Given that ρ_1^* and ρ_2^* are functions of β as shown in Eq. (33) and Eq. (34), we compute the corresponding optimized power values for each flow in terms of β as follows:

$$P_1^* = \frac{(4-3\beta)}{4(3-2\beta)^2} \tag{36}$$

$$P_2^* = \frac{(4-3\beta)^3}{16(3-2\beta)^2(-\beta^2-\beta+4)} \tag{37}$$

and the sum of optimized individual power is:

$$P_{\text{sum}} = P_1^* + P_2^* = \frac{(4-3\beta)(4-\beta)(8-5\beta)}{16(3-2\beta)^2(-\beta^2-\beta+4)} \tag{38}$$

Substituting $\beta = 0$ for the FCFS case into Eq. (36) and Eq. (37), we have $(P_1^*, P_2^*) = \left(\frac{1}{9}, \frac{1}{9}\right)$, yielding a total sum of powers $P_{\text{sum}} = \frac{2}{9}$. Substituting $\beta = 1$ for the HOL case gives $(P_1^*, P_2^*) = \left(\frac{1}{4}, \frac{1}{32}\right)$, resulting in $P_{\text{sum}} = \frac{9}{32}$. These values are consistent with the results previously reported in [1].

Fig. 7 shows the optimized individual power values P_1^*, P_2^*, and P_{sum} plotted against β, exhibiting the same trend as observed in Fig. 5. As β increases, flow1's individual power P_1^*, along with the sum and the difference of P_1^* and P_2^*, increases, while the individual power for flow 2, P_2^*, decreases. This trend is consistent with the behavior of ρ_1^*, ρ_2^*, and ρ^* in Fig. 6. It is also consistent with the behavior of P_1^*, P_2^*, and P_{sum} in the delay-dependent system when optimizing individual power as shown in Fig. 5. Although the parameters k and β both range from 0 to 1, they represent different approaches to spanning their spectrum from FCFS to HOL, and thus capture different notions of flow discrimination. While the overall behavior and trends in joint individual power optimization are similar in both the delay-dependent and beta-priority systems, the rate of change in power values differs due to the distinct characteristics of the k and β parameterizations—except at the endpoints corresponding to FCFS ($k = \beta = 0$) and HOL ($k = \beta = 1$).

4.2. Optimizing power metric 2: Sum of individual powers

We now shift the optimization objective to the **sum of individual powers**. Following the same analysis structure used for power metric 1, we first examine the delay-dependent system, followed by the beta-priority system.

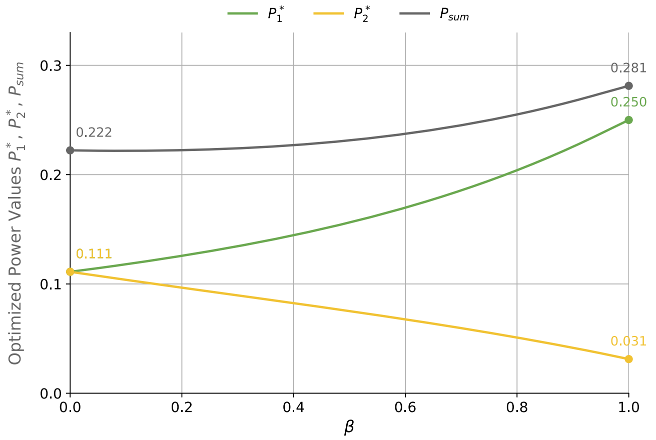


Fig. 7. Optimized power values of P_1^*, P_2^*, and P_{sum} versus β, derived from individual power optimization in an M/M/1 system with two flows using a beta-priority queueing discipline.

4.2.1. The delay-dependent system

The sum of individual powers in a delay-dependent system is the summation of Eq. (20) and (21), resulting in:

$$P_{\text{sum}} = P_1 + P_2 = \frac{\rho_1(1-\rho)(1-k\rho_1)}{1-k\rho} + \rho_2(1-\rho)(1-k\rho_1) \tag{39}$$

This can be simplified to:

$$P_{\text{sum}} = \frac{\rho(1-k\rho_2)(1-\rho)(1-k\rho_1)}{1-k\rho} \tag{40}$$

To find the maximal sum of individual powers, we solve the following equations:

$$\begin{cases} \frac{\partial}{\partial\rho_1} P_{\text{sum}} = \frac{\partial}{\partial\rho_1} \frac{\rho(1-k\rho_2)(1-\rho)(1-k\rho_1)}{1-k\rho} = 0 \\ \frac{\partial}{\partial\rho_2} P_{\text{sum}} = \frac{\partial}{\partial\rho_2} \frac{\rho(1-k\rho_2)(1-\rho)(1-k\rho_1)}{1-k\rho} = 0 \end{cases} \tag{41}$$

and establish the following theorem:

Theorem 4.3. *In an M/M/1 system with two flows ($n = 2$) using a* ***delay-dependent*** *queueing discipline (excluding the FCFS case where $k = 0$), the* ***sum of individual powers*** *is maximized when*

$$\rho_1^* = \rho_2^* \tag{42}$$

This establishes $\rho_1^* = \rho_2^*$ as a ***necessary condition*** *for maximizing the sum of individual powers.*

Proof. From the partial differentials of Eq. (41), we have:

$$\begin{cases} (1-k\rho_2)\frac{[(1-2\rho)(1-k\rho_1)-k\rho(1-\rho)](1-k\rho)\ +\ k\rho(1-\rho)(1-k\rho_1)}{(1-k\rho)^2} = 0 \\ (1-k\rho_1)\frac{[(1-2\rho)(1-k\rho_2)-k\rho(1-\rho)](1-k\rho)\ +\ k\rho(1-\rho)(1-k\rho_2)}{(1-k\rho)^2} = 0 \end{cases}$$

Since neither $1 - k\rho_2$ nor $1 - k\rho_1$ can be zero (as ρ_1 and ρ_2 are assumed to be less than one to prevent system overloading[15]), the other terms in each numerator must be zero, leading to:

$$\begin{cases} [(1-2\rho)(1-k\rho_1) - k\rho(1-\rho)](1-k\rho) + k\rho(1-\rho)(1-k\rho_1) = 0 \\ [(1-2\rho)(1-k\rho_2) - k\rho(1-\rho)](1-k\rho) + k\rho(1-\rho)(1-k\rho_2) = 0 \end{cases} \tag{43}$$

Rewriting these equations yields:

$$\begin{cases} (1-2\rho)(1-k\rho_1)(1-k\rho) = k\rho(1-\rho)(1-k\rho-1+k\rho_1) \\ (1-2\rho)(1-k\rho_2)(1-k\rho) = k\rho(1-\rho)(1-k\rho-1+k\rho_2) \end{cases}$$

If $k = 0$, corresponding to the FCFS case, substituting $k = 0$ into the above equations yields $1 - 2\rho = 0$, resulting in the optimal total utilization $\rho^* = 0.5$, without specifying the individual values of ρ_1 and ρ_2. This result was previously discussed in [1]. For other cases where $k \neq 0$, we proceed by solving the equations and simplifying the right-hand sides to obtain:

$$\begin{cases} (1-2\rho)(1-k\rho_1)(1-k\rho) = -k\rho(1-\rho)k\rho_2 \\ (1-2\rho)(1-k\rho_2)(1-k\rho) = -k\rho(1-\rho)k\rho_1 \end{cases}$$

We rearrange the terms to express the relationship between ρ_1 and ρ_2 in a simpler form:

$$\frac{1-k\rho_1}{\rho_2} = \frac{1-k\rho_2}{\rho_1} = \frac{-k^2\rho(1-\rho)}{(1-2\rho)(1-k\rho)}$$

From this relationship, we multiply both sides of $\frac{1-k\rho_1}{\rho_2} = \frac{1-k\rho_2}{\rho_1}$ by $\rho_1\rho_2$ to obtain:

$$(1-k\rho_1)\rho_1 = (1-k\rho_2)\rho_2$$

Leading to

$$(1-k\rho_1)\rho_1 - (1-k\rho_2)\rho_2 = \rho_1 - k(\rho_1)^2 - \rho_2 + k(\rho_2)^2 = 0$$

[15] This is because if either $1 - k\rho_2$ or $1 - k\rho_1$ were to equal zero, it would require either $k = 1$ and $\rho_1 = 1$ or $k = 1$ and $\rho_2 = 1$. However, both ρ_1 and ρ_2 are assumed to be less than 1 to prevent system overloading, making it impossible for either $1 - k\rho_2$ or $1 - k\rho_1$ to equal zero.

and thus

$$\rho_1 - k\rho_1^2 - \rho_2 + k\rho_2^2 = (\rho_1 - \rho_2) - k(\rho_1^2 - \rho_2^2)$$
$$= (\rho_1 - \rho_2) - k(\rho_1 - \rho_2)(\rho_1 + \rho_2) = 0$$

Factoring out the common term $(\rho_1 - \rho_2)$ yields:

$$(\rho_1 - \rho_2)\ [1 - k(\rho_1 + \rho_2)] = 0$$

Since $1 - k(\rho_1 + \rho_2)$ cannot be zero under the assumption that the system's total utilization $\rho = \rho_1 + \rho_2 < 1$, it follows that $k(\rho_1 + \rho_2) < 1$, and thus $1 - k(\rho_1 + \rho_2) > 0$. Therefore, for the equation to equal zero, it must be that $(\rho_1 - \rho_2) = 0$.

Thus, we obtain:

$$\rho_1^* = \rho_2^* \quad \text{for } k \neq 0 \tag{44}$$

Although $\rho_1^* = \rho_2^*$ is a necessary condition for maximizing the sum of individual powers, it is not **sufficient** on its own to guarantee optimality. To continue the search for the optimal sum, we further solve the partial derivatives to identify the values of ρ_1^* and ρ_2^* that maximize the sum of individual powers, using the necessary condition $\rho_1^* = \rho_2^*$. By substituting $\rho_1^* = \rho_2^*$ into the first equation from Eq. (43), we get:

$$[(1 - 2\rho)(1 - k\rho_1) - k\rho(1 - \rho)](1 - k\rho) + k\rho(1 - \rho)(1 - k\rho_1) = 0$$

we then obtain:

$$[(1 - 4\rho_1)(1 - k\rho_1) - 2k\rho_1(1 - 2\rho_1)](1 - 2k\rho_1) + 2k\rho_1(1 - 2\rho_1)(1 - k\rho_1) = 0$$

which simplifies to:

$$(1 - 4\rho_1 - k\rho_1 + 4k{\rho_1}^2 - 2k\rho_1 + 4k{\rho_1}^2)(1 - k\rho_1) - 2k\rho_1(1 - 2\rho_1) = 0$$

This can be written as:

$$(-12k^2){\rho_1}^3 + (4k^2 + 12k){\rho_1}^2 - (3k + 4)\rho_1 + 1 = 0$$

The root of this cubic equation within the interval $[0, 1]$ gives the optimized value ρ_1^*. Since $\rho_2^* = \rho_1^*$ under the necessary condition for maximizing the sum of individual powers, this pair (ρ_1^*, ρ_2^*) yields the optimal sum of powers for each value of k. Fig. 8 presents the numerically computed roots of ρ_1^* as a function of k. For the HOL case ($k = 1$), as discussed in [1], the sum of powers is maximized when $\rho_1^* = \rho_2^* = \frac{1}{3}$. In contrast, for the FCFS case ($k = 0$), the maximum sum is achieved when $\rho^* = \rho_1^* + \rho_2^* = 0.5$, without requiring that $\rho_1^* = \rho_2^*$. However, to maintain consistency with other k values where equality holds, we set $\rho_1^* = \rho_2^* = 0.25$ for the FCFS case in the plot. For intermediate values of k, representing queueing disciplines between FCFS and HOL, the optimal values are shown in the figure. We observe that ρ_1^* (as a function of k) increases gradually as k increases and remains nearly flat across the lower half of the curve (for $k < 0.5$). In contrast to the behavior of optimized ρ_1^*, ρ_2^*, and ρ^* under joint individual power optimization shown in Fig. 4, the optimized utilizations in Fig. 8 remain $\rho_1^* = \rho_2^*$ across different values of k. This is different from Fig. 4, where ρ_1^* and ρ_2^* diverge as k increases. □

Fig. 9 presents the corresponding optimal power values P_1, P_2, and P_{sum}^*, computed using the optimized pair (ρ_1^*, ρ_2^*) shown in Fig. 8. At $k = 0$ (FCFS), we have $P_1 = P_2 = 0.125$, resulting in $P_{\text{sum}}^* = 0.25$. At $k = 1$ (HOL), the values are $P_1 = \frac{2}{9} \approx 0.222$, $P_2 = \frac{2}{27} \approx 0.074$, and $P_{\text{sum}}^* = \frac{8}{27} \approx 0.296$.

The trend in these optimal power values under sum of individual powers optimization aligns with the behavior observed under joint individual power optimization. As the flow discrimination parameter k increases, P_1 increases while P_2 decreases, leading to a growing disparity between the two. Moreover, since the increase in P_1 outweighs the decrease in P_2, the total sum P_{sum}^* also increases with k. These patterns are illustrated in Fig. 9, where the black curve representing P_{sum}^* rises with k, and the gap between the green curve for P_1 and the yellow curve for P_2 widens. Furthermore, although $\rho_1^* = \rho_2^*$ in this optimization both increase with k, the corresponding power values (P_1 and P_2) move in opposite directions and at different rates as k increases. Specifically, as k grows, both utilization factors increase by the same amount, yet P_1 increases while P_2 decreases—demonstrating that identical increases in utilization can lead to diverging power outcomes depending on the priority structure imposed by k.

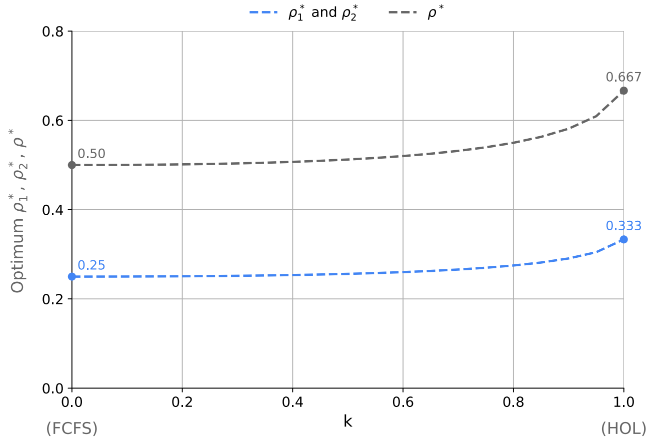


Fig. 8. Optimum ρ_1^*, ρ_2^*, and ρ^* (where $\rho_2^* = \rho_1^*$) that maximizes the sum of individual powers versus k in an M/M/1 system with two flows ($n = 2$) under the delay-dependent queueing discipline. At $k = 0$ (FCFS), where $\rho_1^* = \rho_2^*$ are not required for optimal P_{sum}^* as long as $\rho^* = 0.5$, we explicitly set $\rho_1^* = \rho_2^* = 0.25$ to align with the requirement $\rho_1^* = \rho_2^*$ for other values of k.

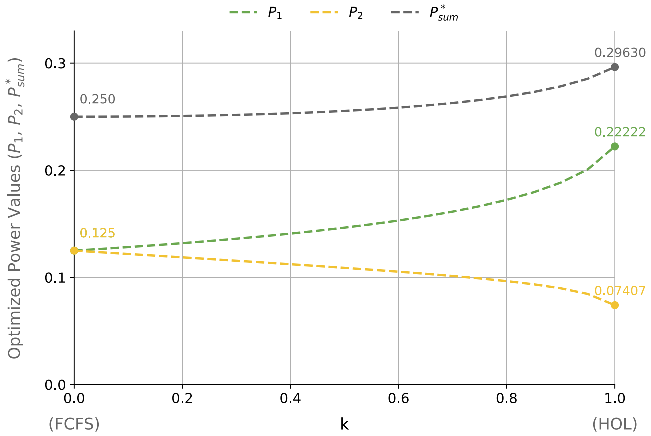


Fig. 9. The maximal sum of power P_{sum}^* along with the individual powers P_1 and P_2 versus k in an M/M/1 system with two flows under the delay-dependent queueing discipline.

4.2.2. *The beta-priority system*

The sum of individual powers in the beta-priority system of two flows is given by:

$$P_{\text{sum}} = P_1 + P_2 = \frac{\rho_1(1 - \rho_1)(1 - \rho)}{1 - \rho_1 - \beta\rho_2} + \frac{\rho_2(1 - \rho_1)(1 - \rho)}{1 - (1 - \beta)\rho_1}$$

This can be simplified to:

$$P_{\text{sum}} = \frac{\rho\,(1 - \rho)\,(1 - \rho_1)\,[1 - (1 - \beta)\rho_1 - \beta\rho_2]}{(1 - \rho_1 - \beta\rho_2)\,[1 - (1 - \beta)\rho_1]} \tag{45}$$

Since the coefficients for ρ_1 and ρ_2 in the simplified form of P_{sum} are different, the equilibrium optimal values of ρ_1^* and ρ_2^* in optimizing the sum of individual powers may not be the same. This is different from the observation in the delay-dependent system where $\rho_1^* = \rho_2^*$, as specified in Theorem 4.3.

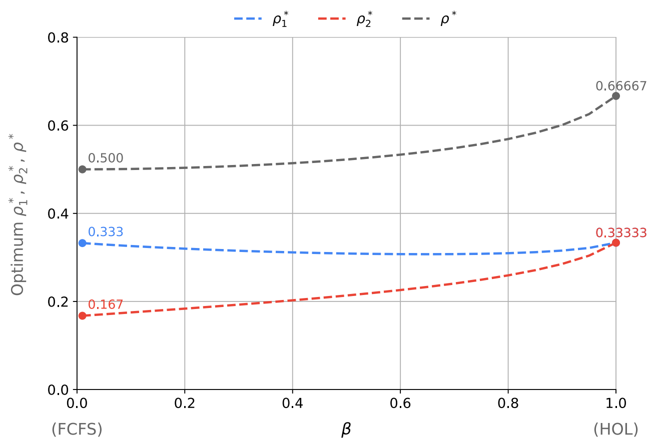


Fig. 10. Optimum values of ρ_1^*, ρ_2^*, and ρ^* versus β, derived from the sum of individual powers optimization in an M/M/1 system with two flows ($n = 2$) under the beta-priority queueing discipline.

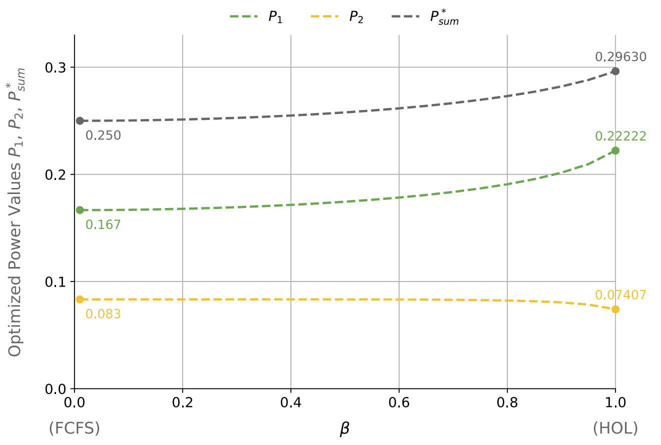


Fig. 11. Optimized power values of P_1, P_2, and P_{sum}^* versus β, derived from the sum of individual powers optimization in an M/M/1 system with two flows ($n = 2$) under the beta-priority queueing discipline.

To find the optimal utilizations that achieve the maximum sum of individual powers, we establish the following partial differentials:

$$\begin{cases} \frac{\partial}{\partial \rho_1} P_{\text{sum}} = \frac{\partial}{\partial \rho_1} \frac{\rho(1-\rho)(1-\rho_1)[1-(1-\beta)\rho_1-\beta\rho_2]}{(1-\rho_1-\beta\rho_2)[1-(1-\beta)\rho_1]} = 0 \\ \frac{\partial}{\partial \rho_2} P_{\text{sum}} = \frac{\partial}{\partial \rho_2} \frac{\rho(1-\rho)(1-\rho_1)[1-(1-\beta)\rho_1-\beta\rho_2]}{(1-\rho_1-\beta\rho_2)[1-(1-\beta)\rho_1]} = 0 \end{cases}$$

Given that those equations are complex and not easy to solve explicitly, we choose to find the values of ρ_1^* and ρ_2^* that optimize the sum of individual powers numerically. The following section discusses the optimization results.

The values of each set of (ρ_1^*, ρ_2^*) optimizing the sum of individual powers versus β are shown in Fig. 10. There is no data point for (ρ_1^*, ρ_2^*) when $\beta = 0$, since the only constraint in this optimization process is $\rho = \rho_1 + \rho_2 = 0.5$ without specifying the exact values of (ρ_1^*, ρ_2^*). We plot the initial point with data at $\beta = 0.01$, leading to $\rho_1^* \approx 0.333$, $\rho_2^* \approx 0.167$, and $\rho^* \approx 0.5$. In Fig. 10, ρ_1^* and ρ_2^* are equal only when $\beta = 1$ in the HOL case. For other values of β, the optimization results show $\rho_1^* \neq \rho_2^*$, which differs from the behavior in the delay-dependent system when optimizing the same target performance metric, sum of individual powers.

In addition, the trend for ρ_1^* and ρ_2^* is different from the trend observed when optimizing individual power, where ρ_1^* increases and ρ_2^* decreases as the level of flow discrimination increases, as shown in Fig. 6. Here, in contrast, ρ_1^* shows only slight changes as β increases from 0 to 1, with the minimum value being about 0.307 and the maximum value being 0.333. Meanwhile, ρ_2^* increases with β since flow2's power is also included in the optimization target. Moreover, the difference between ρ_1^* and ρ_2^* decreases and becomes zero when β reaches its maximum value of 1. This can be observed from the curves of ρ_1^* and ρ_2^*, as the gap between them narrows and they converge at $\beta = 1$. This behavior contrasts with what is observed when optimizing the individual power of both flows, where the difference between ρ_1^* and ρ_2^* increases, as shown in Fig. 6. However, a consistent observation when changing the optimization metric from individual power to the sum of power is that the total utilization ρ^* (the sum of ρ_1^* and ρ_2^*) increases as the level of flow discrimination rises (i.e., as β increases).

In Fig. 11, the corresponding maximal sum of individual powers, P_{sum}^*, along with the individual power values for P_1 and P_2 versus β, are presented. For the left bound of the FCFS case where $\beta = 0$, only the sum of individual powers with a value of 0.25 is shown in the plot. The exact values for P_1 and P_2 are not marked since there are several combinations for P_1 and P_2 that satisfy the condition $\rho^* = \rho_1^* + \rho_2^* = 0.5$ with the sum of power being 0.25. The starting point we use in the curves is when $\beta = 0.01$, resulting in $P_1 \approx 0.167$, $P_2 \approx 0.083$, and $P_{\text{sum}}^* \approx 0.25$. From Fig. 11, the maximum of sum of individual powers P_{sum}^* increases as β increases. The increase is driven by the rise in flow1's power, while P_2 shows a slight decrease. As depicted by the yellow curve in the figure, P_2 decreases from 0.0833 to 0.074. Conversely, P_1 increases from 0.167 to 0.222. Consequently, the sum of the powers increases from 0.25 to 0.296.

Even though ρ_1^* changes subtly and ρ_2^* increases, as shown in Fig. 10, the individual power for flow 1, P_1, still increases with the increase in β. This is because as flows become more discriminative, the higher-priority flow is less affected by the lower-priority flow, resulting in reduced waiting and response times. Therefore, this leads to a higher individual power value for flow 1 with the same value of ρ_1^*. Subsequently, the increase in P_1 also leads to an increase in the sum of powers P_{sum}^*, implying the positive effect of flow discrimination.

5. Extension to full range from FCFS to HOL: Arbitrary number of flows in the beta-priority system

We now extend the investigation from $n = 2$ (two flows) to an arbitrary number of flows. Given the complexity of the response time equations in the delay-dependent system, we focus on the **beta-priority** system, as its equation form is explicit compared to the recursive form in the delay-dependent system. As in Section 4, we examine both power metric 1 (individual power) and power metric 2 (the sum of individual powers), optimizing each within the beta-priority system. This allows us to explore a spectrum from FCFS to HOL for an arbitrary n number of flows.

5.1. *Optimizing power metric 1: individual power*

We first consider the power metric 1—individual power—as the optimization objective. Based on this, we begin with deriving analytical results that apply to an arbitrary number n of flows in the beta-priority system. However, due to the complexity of the computations, analytical solutions become increasingly difficult to obtain as n grows. To address this, we then adopt numerical methods to compute equilibrium outcomes when each flow jointly optimizes its own individual power for $n > 2$, providing a detailed view of how the equilibrium solutions evolve across a spectrum—from FCFS (the least flow-discriminatory) to HOL (the most flow-discriminatory).

5.1.1. *Analytical results*

We begin by establishing the analytical expressions for individual power in the beta-priority system with an arbitrary number n of flows. Theorem 4.2 shows that the optimized utilization for the flow 2 is $\rho_2^* = \frac{1-\rho_1}{2}$, which is half the remaining utilization after flow 1 take's its share.

This result generalizes to an arbitrary number of flows, as stated in the following theorem:

Theorem 5.1. *In an M/M/1 system with arbitrary number n of flows using the beta-priority queueing discipline, the lowest-priority flow, the n^{th} flow, achieves its maximum individual power when it takes half of the remaining utilization left by the higher-priority flows (i.e., flows 1 through $n-1$). The expression is given by:*

$$\rho_n^* = \frac{1-\sum_{i=1}^{n-1}\rho_i}{2} = \frac{1-\sigma_{n-1}}{2} \tag{46}$$

where $\sigma_i = \sum_{j=1}^{i}\rho_j$

Proof. In the beta-priority system, the mean response time for each flow i is a weighted average of its mean response times under HOL and FCFS, with weight β, as given in Eq. (18). The corresponding normalized mean response time—using the no-load delay $\frac{1}{\mu}$ as the normalization factor—is given by:

$$\mu T_i = \beta \cdot \frac{1}{(1-\sigma_{i-1})(1-\sigma_i)} + (1-\beta)\cdot\frac{1}{1-\rho} \tag{47}$$

For the lowest-priority flow n, the normalized mean response time can be simplified to the following as $\rho = \sigma_n$:

$$\mu T_n = \beta \cdot \frac{1}{(1-\sigma_{n-1})(1-\sigma_n)} + (1-\beta)\cdot\frac{1}{1-\rho} = \frac{[\beta + (1-\beta)\cdot(1-\sigma_{n-1})]}{(1-\sigma_{n-1})(1-\sigma_n)}$$

The corresponding individual power for the flow n is:

$$P_n = \frac{\rho_n}{\mu T_n} = \frac{\rho_n(1-\sigma_{n-1})(1-\sigma_n)}{[\beta + (1-\beta)\cdot(1-\sigma_{n-1})]} \tag{48}$$

Taking the partial derivative of Eq. (48) with respect to ρ_n, and taking the factor that is not related to ρ_n out of the differential equation as constant:

$$\begin{aligned}\frac{\partial P_n}{\partial \rho_n} &= \frac{\partial}{\partial \rho_n}\left(\frac{\rho_n(1-\sigma_{n-1})(1-\sigma_n)}{[\beta + (1-\beta)\cdot(1-\sigma_{n-1})]}\right)\\ &= \left(\frac{(1-\sigma_{n-1})}{[\beta + (1-\beta)\cdot(1-\sigma_{n-1})]}\right)\cdot\frac{\partial \rho_n(1-\sigma_n)}{\partial \rho_n}\end{aligned}$$

Proceeding the differential and setting it to zero, we have:

$$\frac{\partial \rho_n(1-\sigma_n)}{\rho_n} = 1-\sigma_n-\rho_n = 1-\sigma_{n-1}-2\rho_n = 0$$

Leading to

$$\rho_n^* = \frac{1-\sigma_{n-1}}{2}$$

Since $\sigma_{n-1} = \sum_{j=1}^{n-1}\rho_j$, this theorem shows that the optimal utilization ρ_n^*, which maximizes the individual power P_n^*, is achieved when ρ_n^* equals half of the remaining utilizations after all higher-priority flows have taken their shares. □

Similarly, we can derive the optimal utilization for the second-lowest-priority flow ρ_{n-1}^*, which is affected only by the higher-priority flows (i.e., flow 1 through $n-2$) and depends on their cumulative utilizations $\sigma_{n-2} = \sum_{j=1}^{n-2}\rho_j$. The result is stated in the following theorem:

Theorem 5.2. *In an M/M/1 system with n flows using the beta-priority queueing discipline, when each flow optimizes its individual power, the second-lowest-priority flow, i.e., the $(n-1)^{th}$ flow, achieves its optimal individual power P_{n-1}^* when*

$$\rho_{n-1}^* = \frac{1-\sigma_{n-2}}{2}\cdot\frac{\beta+2(1-\beta)(1-\sigma_{n-2})}{\beta+3(1-\beta)(1-\sigma_{n-2})} \tag{49}$$

Proof. From Eq. (47), the normalized mean response time for the $(n-1)^{\text{th}}$ flow is given by:

$$\begin{aligned}\mu T_{n-1} &= \frac{\beta}{(1-\sigma_{n-1})(1-\sigma_{n-2})} + \frac{1-\beta}{(1-\sigma_n)}\\ &= \frac{\beta\cdot(1-\sigma_n) + (1-\beta)\cdot(1-\sigma_{n-1})(1-\sigma_{n-2})}{(1-\sigma_{n-1})(1-\sigma_{n-2})(1-\sigma_n)}\end{aligned}$$

Thus, the individual power for the $(n-1)^{\text{th}}$ flow is:

$$P_{n-1} = \frac{\rho_{n-1}}{\mu T_{n-1}} = \frac{\rho_{n-1}(1-\sigma_{n-1})(1-\sigma_{n-2})(1-\sigma_n)}{\beta\cdot(1-\sigma_n) + (1-\beta)\cdot(1-\sigma_{n-1})(1-\sigma_{n-2})} \tag{50}$$

Taking the partial derivative of Eq. (50) with respect to ρ_{n-1}, and factoring out the terms that are not related to ρ_{n-1}:

$$\frac{\partial P_{n-1}}{\partial \rho_{n-1}} = (1-\sigma_{n-2})\,\frac{\partial}{\partial \rho_{n-1}}\left(\frac{\rho_{n-1}(1-\sigma_{n-1})(1-\sigma_n)}{[\ \beta\cdot(1-\sigma_n) + (1-\beta)\cdot(1-\sigma_{n-1})(1-\sigma_{n-2})\ ]}\right)$$

Setting the partial differential equation to zero, we have:

$$\begin{aligned}&[\beta\cdot(1-\sigma_n) + (1-\beta)\cdot(1-\sigma_{n-1})(1-\sigma_{n-2})]\\ &\quad\cdot[(1-\sigma_{n-1}-\rho_{n-1})(1-\sigma_n) - \rho_{n-1}(1-\sigma_{n-1})]\\ &\quad-\rho_{n-1}(1-\sigma_{n-1})(1-\sigma_n)\ [-\beta-(1-\beta)(1-\sigma_{n-2})] = 0\end{aligned} \tag{51}$$

With Theorem 5.1, we have:

$$\rho_n^* = \frac{1-\sigma_{n-1}}{2}$$

Using this result, we can compute $1-\sigma_n$ as follows:

$$1-\sigma_n = 1-\sigma_{n-1}-\rho_n = 1-\sigma_{n-1}-\frac{1-\sigma_{n-1}}{2} = (1-\sigma_{n-1})(1-\frac{1}{2})$$

Therefore,

$$1-\sigma_n = \frac{1-\sigma_{n-1}}{2} \tag{52}$$

Substituting Eq. (52) into Eq. (51):

$$\begin{aligned}&[\beta\cdot\frac{1-\sigma_{n-1}}{2} + (1-\beta)\cdot(1-\sigma_{n-1})(1-\sigma_{n-2})]\\ &\quad\cdot[(1-\sigma_{n-1}-\rho_{n-1})(\frac{1-\sigma_{n-1}}{2}) - \rho_{n-1}(1-\sigma_{n-1})]\\ &\quad-\rho_{n-1}(1-\sigma_{n-1})(\frac{1-\sigma_{n-1}}{2})\ [-\beta-(1-\beta)(1-\sigma_{n-2})] = 0\end{aligned}$$

Factoring out the $1-\sigma_{n-1}$ and moving the second line part to the right side of the equation, we get:

$$\begin{aligned}&(1-\sigma_{n-1})\cdot\left[\frac{\beta}{2}+(1-\beta)(1-\sigma_{n-2})\right]\cdot(1-\sigma_{n-1})\cdot\left[\frac{1-\sigma_{n-1}-\rho_{n-1}}{2}-\rho_{n-1}\right]\\ &\quad= (1-\sigma_{n-1})^2\cdot\left(\frac{\rho_{n-1}}{2}\right)[-\beta-(1-\beta)(1-\sigma_{n-2})]\end{aligned}$$

Canceling the $(1-\sigma_{n-1})^2$ term, we have :

$$\begin{aligned}&\left[\frac{\beta}{2}+(1-\beta)(1-\sigma_{n-2})\right]\cdot\left[\frac{1-\sigma_{n-1}-\rho_{n-1}}{2}-\rho_{n-1}\right]\\ &\quad= (\frac{\rho_{n-1}}{2})\ [-\beta-(1-\beta)(1-\sigma_{n-2})]\end{aligned}$$

Isolating ρ_{n-1} from σ_{n-1} with $\sigma_{n-1} = \sigma_{n-2}+\rho_{n-1}$, we have:

$$\begin{aligned}&\left[\frac{\beta}{2}+(1-\beta)(1-\sigma_{n-2})\right]\cdot\left[\frac{1-\sigma_{n-2}-\rho_{n-1}-\rho_{n-1}}{2}-\rho_{n-1}\right]\\ &\quad= \left(\frac{\rho_{n-1}}{2}\right)[-\beta-(1-\beta)(1-\sigma_{n-2})]\end{aligned}$$

This can be rewritten as:

$$\begin{aligned}&[\beta+2(1-\beta)(1-\sigma_{n-2})]\cdot[(1-\sigma_{n-2})-4\rho_{n-1}]\\ &\quad= 2\rho_{n-1}\ [-\beta-(1-\beta)(1-\sigma_{n-2})]\end{aligned}$$

Moving the ρ_{n-1} term to the right side of the equation:

$$\begin{aligned}&[\beta+2(1-\beta)(1-\sigma_{n-2})]\cdot[(1-\sigma_{n-2})]\\ &\quad= 2\rho_{n-1}\ [-\beta-(1-\beta)(1-\sigma_{n-2})] + 4\rho_{n-1}\ [\beta+2(1-\beta)(1-\sigma_{n-2})]\end{aligned}$$

This can be expressed as

$$[\beta+2(1-\beta)(1-\sigma_{n-2})]\cdot(1-\sigma_{n-2}) = 2\rho_{n-1}\ [\beta+3(1-\beta)(1-\sigma_{n-2})]$$

Thus, we have:

$$\rho_{n-1} = \frac{1-\sigma_{n-2}}{2}\cdot\frac{\beta+2(1-\beta)(1-\sigma_{n-2})}{\beta+3(1-\beta)(1-\sigma_{n-2})}$$

which is Eq. (49) □

This completes the proof of the theorem, showing that the optimized utilization ρ^*_{n-1} is expressed as a function of $(1-\sigma_{n-2})$, the remaining utilization after all higher-priority flows (i.e., flows 1 through $n-2$) have taken their shares.

We also observe that the analytical expression for the optimal utilization ρ^*_i becomes increasingly complex as we proceed from the lowest-priority flow (n) to the second-lowest-priority flow ($n-1$). While the expression for ρ^*_n is linear and straightforward, the result for ρ^*_{n-1} is already nonlinear—though it can still be written in terms of the remaining utilization after higher-priority flows, $(1-\sigma_{n-2})$. This trend suggests increasing complexity in deriving analytical expressions for even higher-priority flows such as $n-2$, $n-3$, …, 2, and 1. Therefore, we turn to numerical methods to continue our analysis for jointly optimizing individual powers (power metric 1) for arbitrary n number of flows across the full spectrum from FCFS to HOL.

5.1.2. Numerical results

We update per-flow utilizations one at a time in priority order, namely, round-robin over classes $1,2,\dots,n$, then back to 1 repeatedly. For class i, we hold all the other ρ_j fixed and choose ρ_i that maximizes its individual power, P_i, subject to $0 \le \rho_i < 1$ and $0 \le \rho = \sum_{j=1}^{n} \rho_j < 1$. One full sweep constitutes an iteration. We repeat until convergence, namely, until

$$\max_{1\le i\le n}\left|\rho_i^{(t+1)} - \rho_i^{(t)}\right| < \varepsilon, \quad \varepsilon = 10^{-8}$$

i.e., no ρ_i changes by more than ε from one sweep to the next. We test cases with $n = 1,2,\dots,40$ across various β values from 0 to 1 with a step size of 0.05, i.e., [0, 0.05, 1, .., 0.9, 0.95, 1].[16]

Fig. 12 presents the numerical results[17] for the optimized ρ^* and the sum of optimized individual powers $\sum_{i=1}^{n} P_i^*$. [18] FCFS and HOL form the lower and upper bounds, respectively, and the results for intermediate values of β, corresponding to queueing disciplines with varying degrees of flow discrimination, lie between these two extremes.

As shown in Fig. 12(a), the optimized total utilizaiton ρ^* increases monotonically with β for each n. As n grows, the curves for intermediate β values all concentrate near the FCFS bound. This occurs because any beta-priority queueing discipline with $\beta < 1$ retains a nonzero FCFS component, and therefore none of the flows can maintain finite response time at $\rho = 1$. Since maximizing individual power is the objective, and an infinite response time at full load ($\rho = 1$) results in zero individual power for every flow, no flow will choose to push the total utilization to 1. As a result, for all $\beta < 1$, the presence of the FCFS component prevents the system from operating near the HOL limit when n becomes large, causing the intermediate-β curves to cluster near the FCFS bound instead. Fig. 12(b) shows that the maximal sum of individual powers $\sum_{i=1}^{n} P_i^*$ decreases with increasing n but increases consistently with β. All intermediate β values lie between the FCFS and HOL bounds.[19]

5.2. Optimizing power metric 2: sum of individual powers

We now turn to the second power metric, sum of individual powers, as the optimization objective for arbitrary n number of flows under the beta-priority queueing system. Given the complexity of solving for the optimized utilizations ρ^*_i (for $i = 1,..,n$) that maximize the **sum of individual powers** in the beta-priority queueing system—even in the two-flow case ($n = 2$), as shown in Section 4—the problem becomes significantly more challenging when n is arbitrary or greater than 2. Therefore, we again resort to numerical methods to find the optimal set of utilizations ρ^*_i for each flow i, as well as the corresponding maximal sum of individual powers within the beta-priority queueing system across different β values for $n > 2$. This provides one way[20] to study the optimization of the sum of individual powers across the full spectrum of flow discrimination from FCFS to HOL for $n > 2$.

5.2.1. Numerical results

Eq. (47) gives the normalized mean response time and Eq. (48) specifies the individual power for the i^{th} flow. The sum of individual powers for arbitrary n number of flows is:

$$P_{\text{sum}} = \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} \frac{\rho_i}{\mu T_i} = \sum_{i=1}^{n} \frac{\rho_i(1-\sigma_{i-1})(1-\sigma_i)(1-\rho)}{\beta\cdot(1-\rho) + (1-\beta)\cdot(1-\sigma_{i-1})(1-\sigma_i)} \tag{53}$$

We use Eq. (53) as the optimization goal in the numerical method, evaluating different values of n from 1 to 40 and sweeping β from 0 to 1 with a step size of 0.05.[21]

The optimization results for maximizing the sum of individual powers are presented in Fig. 13, with each curve representing a queueing discipline under the beta-priority system characterized by β. Within each curve, n ranges from 1 to 40, and the corresponding optimization results for each value of n are shown. Fig. 13(a) shows the result of optimized system utilization, which is the sum of the optimized utilizations of each flow that achieves the maximal sum of individual powers. Fig. 13(b) shows the maximal sum of individual powers for different values of n.

In Figs. 13(a) and 13(b), the curves are bounded by the upper bound (HOL) and the lower bound (FCFS). As stated in paper I [1], the optimization results for other queueing disciplines, as long as they are work-conservative, fall within the region bounded by HOL and FCFS. Here, we use the beta-priority system as a representative example to illustrate this point. The curves in Fig. 13 are ordered by the value of β. As we move from the lower bound curve (corresponding to $\beta = 0$, FCFS) to the upper bound curve (corresponding to $\beta = 1$, HOL), the β values increase, and both the optimized system utilization and the maximal sum of individual powers increase as well. This trend indicates that as the level of flow discrimination increases (i.e., as β increases), higher-priority flows become less affected by lower-priority ones. As a result, the individual power of higher-priority flows can increase, contributing to a rising sum of powers. This suggests that a more discriminative queueing approach can improve overall system performance in terms of the sum of individual powers.

In Fig. 13(a), there is a noticeable gap between the curve for $\beta = 0.99$ and the upper bound curve of $\beta = 1$. When $\beta = 1$, the optimal system utilization approaches 1, while for the slightly smaller β value of 0.99, the optimal system utilization remains around 0.79. The reason that the gap between 0.99 and 1 does not vanish is structural: in the beta-priority system, any nonzero FCFS component (i.e., any $\beta < 1$) injects an FCFS-like term into the response time, and FCFS response time becomes infinite as $\rho \to 1$. Even a tiny weight on an infinite term remains large, so the optimizer keeps ρ^* bounded away from 1 for all $\beta < 1$. Only at $\beta = 1$ (pure HOL system) does this infinite FCFS response time contribution disappear, allowing higher-priority flows in the system to remain stable as ρ approaches 1.

In addition, comparing Figs. 12(a) and 13(a), we see that at $n = 40$, the optimal total utilization ρ^* under **individual-power** optimization across the β values remains high, namely 0.976 to 1 (Fig. 12(a)), whereas **sum-of-individual-powers** optimizations yields total utilizations spanning 0.5 to 0.975 (Fig. 13(a)). To understand this difference, consider the FCFS lower-bound case. Under sum-of-individual-powers optimization, both ρ^* and the sum of individual powers remain unchanged as

[16] We additionally use a finer resolution for β in the range of 0.95 to 1, with a step size of 0.01 as we observe a gap in the results between $\beta = 0.95$ and $\beta = 1$.

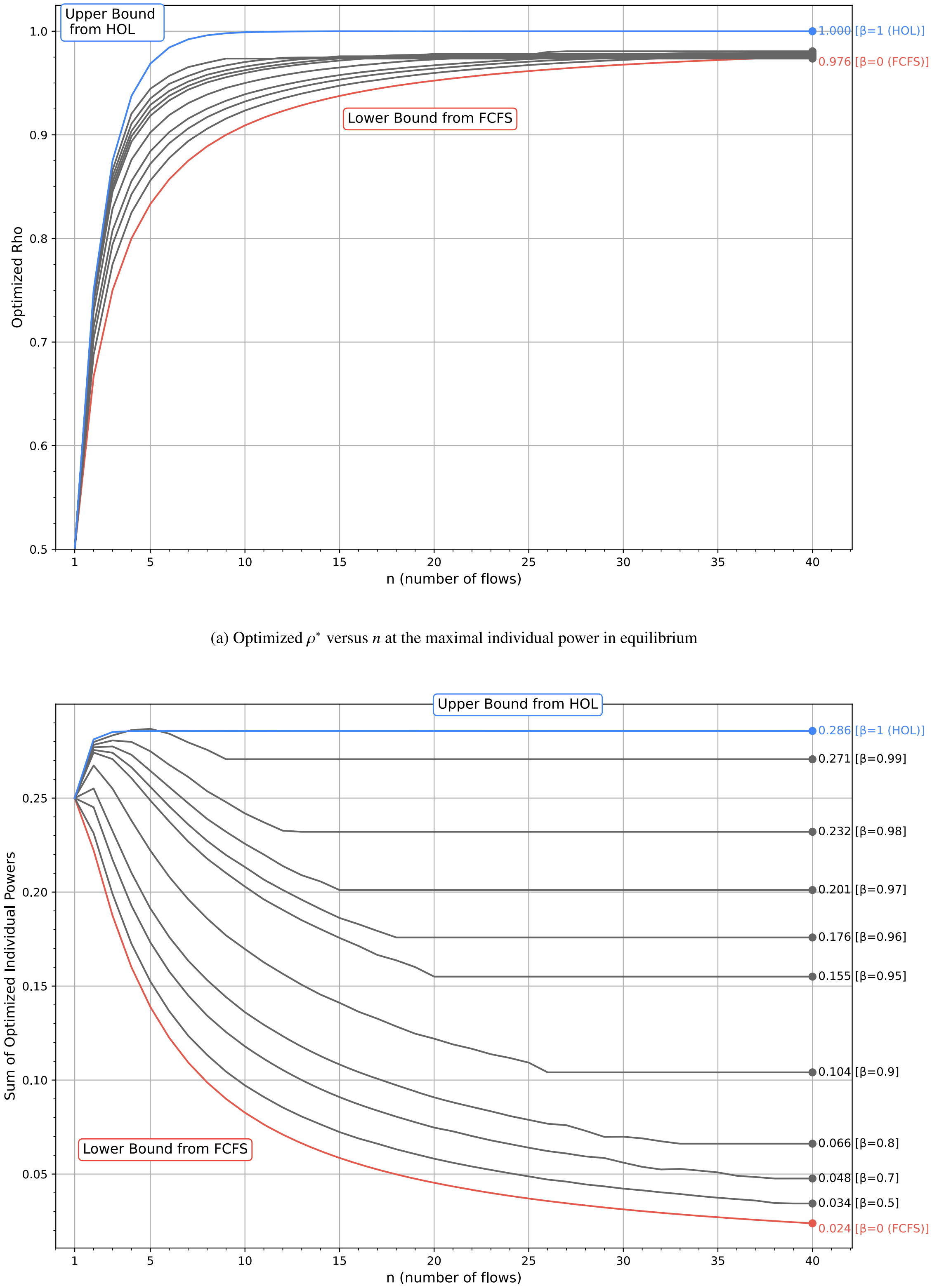
[17] For clarity, we omit the results for $0 < \beta < 0.5$ from the figure, as the curves lie in a narrow region.

[18] We use the sum of optimized individual powers to summarize the results, since listing each P_i^* individually becomes impractical when n is large.

[19] The results for $\beta = 0.99$ at $n = 3,4,5$ slightly exceed the HOL values. This is likely attributable to numerical error.

[20] As stated earlier, there are multiple ways to span the full spectrum of flow discrimination from FCFS to HOL, and the beta-priority system represents one such path; different paths may lead to different optimization outcomes.

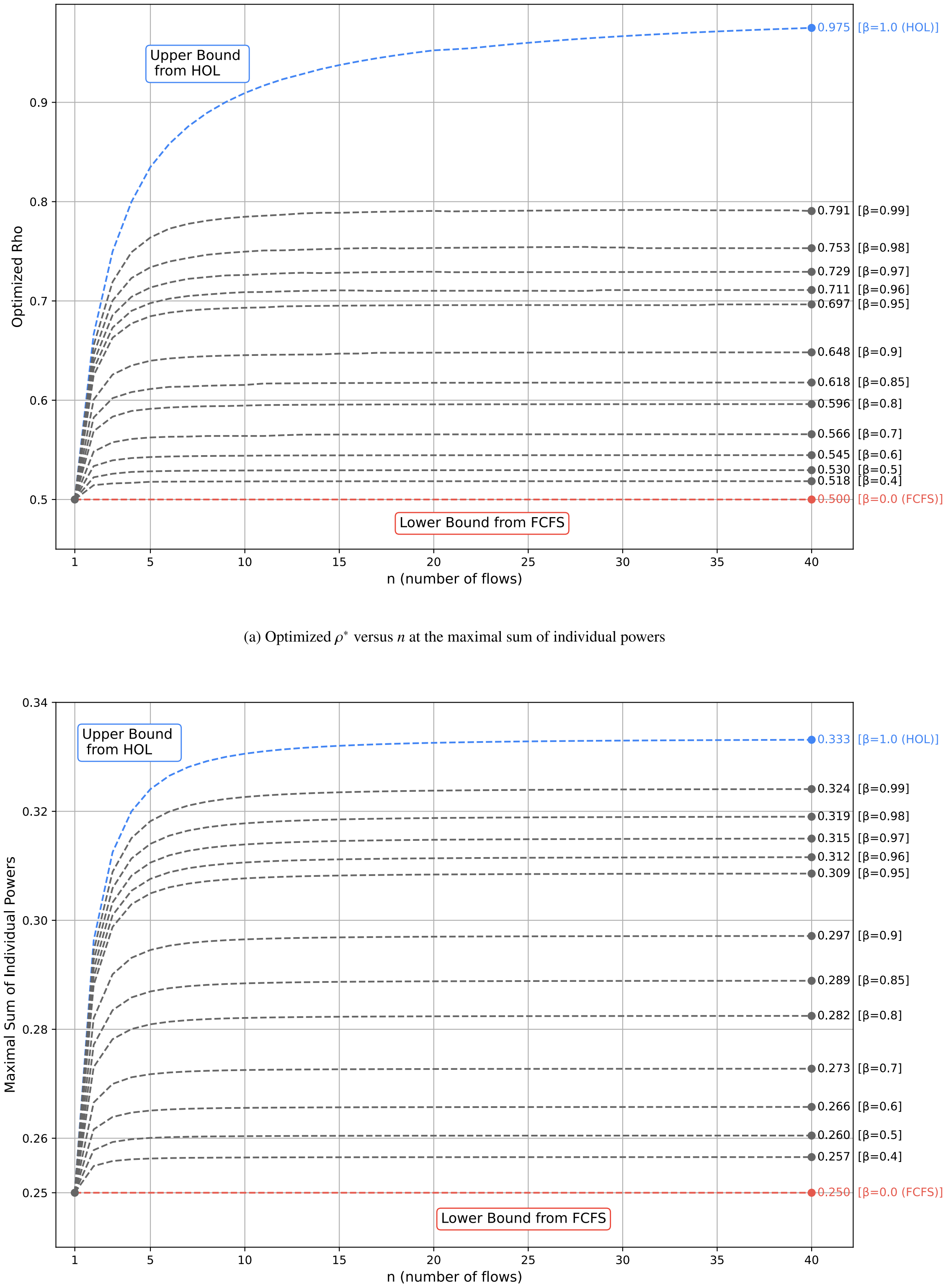
[21] As in the individual power optimization, we also apply a finer resolution for β in the range of 0.95 to 1, using a step size of 0.01.



(a) Optimized ρ^* versus n at the maximal individual power in equilibrium

(b) Sum of optimized individual powers in equilibrium, $\sum_{i=1}^{n} P_i^*$, versus n

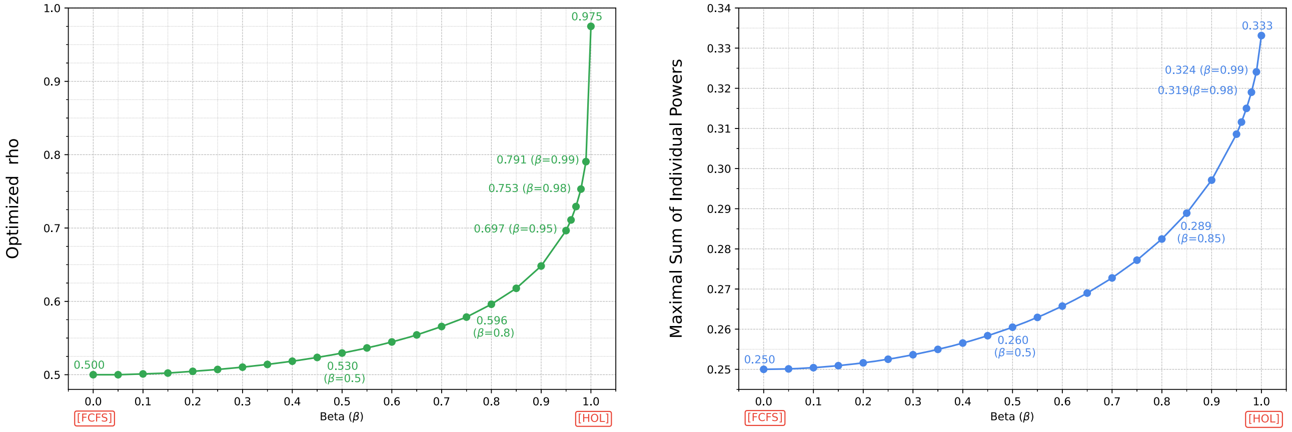
Fig. 12. Maximizing individual power in equilibrium for various β in the beta-priority system for $n = 1, \ldots, 40$.



(a) Optimized ρ^* versus n at the maximal sum of individual powers

(b) Maximal sum of individual powers, P^*_{sum}, versus n

Fig. 13. Maximizing the sum of individual powers for various β in the beta-priority system for $n = 1, \ldots, 40$.



(a) β versus ρ^* at the maximal sum of individual powers for $n = 40$

(b) β versus maximal sum of individual powers, P^*_{sum}, for $n = 40$

Fig. 14. Convergent sum-of-individual-powers optimization results across the spectrum of queueing disciplines in the beta-priority system. The plot shows β versus the maximal sum of individual powers P^*_{sum} and the optimized utilization ρ^* that attains this maximum for each β at $n = 40$ (taken as the convergence point). Data points are sampled for $\beta \in [0, 1]$ in increments of 0.05, with a finer resolution of 0.01 near 1.

n grows, as shown in Fig. 13. However, individual-power optimization yields a Nash equilibrium[26]: each flow greedily increases its own utilization to improve its own individual power, ignoring the response time externality it imposes on others. Any attempt by one flow to back off its utilization (to lower delay) may be promptly absorbed by others who claim the freed capacity. The equilibrium therefore settles with each flow targeting $\rho^*_i = \frac{1}{n+1}$, giving total utilization $\rho^* = \frac{n}{n+1} \to 1$ as n grows. This near-saturation of utilization inflates response times for all flows; thus, although each flow is individually “content”, the sum of individual powers actually deteriorates, as shown in Fig. 12.

5.2.2. Convergence

Given that each curve, except for the HOL curve ($\beta = 1$), nearly converges when n reaches 40, we take the value at $n = 40$ as the convergent value for each queueing discipline represented by β and plot these values against β, as presented in Fig. 14. Fig. 14(a) shows the convergent ρ^* value at the maximal sum of individual powers, and Fig. 14(b) shows the corresponding optimal sum of powers, P^*_{sum}, for each β. These figures illustrate how the optimal sum of powers transitions from FCFS to HOL in the limit. Both curves exhibit exponential growth with β, starting with a small increase and then accelerating rapidly as β surpasses a certain point.

6. Summary

This paper extended the work of Part I [1] by moving beyond the fixed queueing disciplines of FCFS and HOL to introduce a flexible framework for flow priority discrimination to accommodate diverse workloads and SLOs in modern cloud queueing and networked systems. Part I [1] proposed three power-based performance metrics—**individual power** P_i, **sum of individual powers** P_{sum}, and **average power** P_{avg}—and optimized them under the two extreme scheduling policies of FCFS (no discrimination) and HOL (maximal discrimination).

While analytically convenient, these two extremes are often inadequate for heterogeneous environments: FCFS is *too relaxed*, offering no prioritization for latency-critical tasks, whereas strict HOL is *too severe*, risking starvation of lower-priority flows. To address these limitations, this paper introduced continuous, tunable control over the degree of flow priority discrimination across spectra from FCFS to HOL, enabling response times to vary in a controlled manner across flows. We provided this flexibility by studying two parametric families of queueing disciplines that traverse a **continuous spectrum** between FCFS and HOL. When embedded into the power-metric optimization framework, these families enrich the design space: operators may first select an appropriate level of priority discrimination for their workload or SLO portfolio and then determine the utilizations that maximize the chosen power metric. This generalizes the analysis in Part I [1] and reveals how power metrics behave across a continuum of scheduling behaviors rather than only at the two endpoints.

There are multiple paths to span this full range, each allowing scheduling behavior to vary continuously between the two classical extremes.[22] In this study, we focused on two such systems: the **delay-dependent system** [21] and our newly created **beta-priority system**. The delay-dependent system adjusts priority discrimination among flows using a set of $n-1$ ratios $\frac{b_{i+1}}{b_i}$ for $i = 1, \ldots, n-1$. In contrast, the beta-priority system offers a simpler, one-dimensional control via a single parameter β, which interpolates smoothly between FCFS ($\beta = 0$) and HOL ($\beta = 1$). Section 3 introduced these systems and their corresponding response time formulations. We used these two systems as representative approaches to span the full range of flow discrimination and apply them to power metric optimization. Among the three power metrics proposed in Part I [1]—individual power, sum of individual powers, and average power—we focused on the first two since as shown in Theorem 6.3 in Part I [1], the maximum value of average power is unaffected by flow priority discrimination. Therefore, we omitted that metric in this follow-up study.

We began with the two-flow case ($n = 2$) in Section 4 as a starting point to examine the behavior of both systems. When optimizing **individual power** P_i, both systems exhibit similar trends: as flow discrimination increases, the sum of the optimized utilizations ($\rho^*_1 + \rho^*_2$) and the sum of the optimal individual powers ($P^*_1 + P^*_2$) both increase. The differences between the optimized utilizations ($\rho^*_1 - \rho^*_2$) and between the individual powers ($P^*_1 - P^*_2$) also grow. Most importantly, both systems exhibit the same key property: the optimal utilization for the lower-priority flow is always exactly half of the remaining utilization after the higher-priority flow has taken its share, as formalized in Theorem 4.1 for the delay-dependent system and Theorem 4.2 for the beta-priority system.

When optimizing the **sum of individual powers** P_{sum} for $n = 2$, the two systems yield different qualitative results. In the delay dependent system, the optimal solution produces **equal utilizations** for both flows

[22] Different ways of traversing the spectrum from FCFS to HOL give rise to different trajectories, but the FCFS and HOL endpoints remain unchanged.

across its spectrum of queueing disciplines from FCFS to HOL. That is, regardless of the degree of flow priority discrimination, the sum of individual powers is maximized when both flows receive the same utilization. In contrast, this property does not hold in the beta-priority system: equal utilizations arise only at the end points of the spectrum (FCFS and HOL), while at intermediate β values, the optimal utilizations for flow 1 and flow 2 differ. This example illustrates that although both systems span the full range of flow priority discrimination, the corresponding power optimization outcomes along these paths can be different.

For the case of an arbitrary number of flows n in Section 5, we focused exclusively on the beta-priority system due to the complexity of the recursive response time equations in the delay-dependent system. When optimizing **individual power** P_i, we derived an analytical result in Theorem 5.1 showing that the lowest-priority (i.e., n^{th}) flow achieves maximum individual power, P_n^*, when its utilization equals half of the remaining system utilization after higher-priority flows have taken their shares. While similar expressions can be derived for the second-lowest-priority flow, the analytical form becomes increasingly complex as we move toward higher-priority flows. Therefore, we used numerical methods to compute the optimized utilization ρ_i^* and corresponding individual powers P_i^* for all flows. When optimizing **sum of individual powers** $P_{\textbf{sum}}$, we also employed numerical optimization due to the intractability of closed-form solutions for large n. In both optimization settings, the results for intermediate values of β—representing varying degrees of flow discrimination—are bounded between the two extremes of FCFS ($\beta = 0$) and HOL ($\beta = 1$). Notably, under sum-of-individual-powers optimization, both the optimal system utilization and the maximal sum of individual powers increase as β increases.

CRediT authorship contribution statement

Meng-Jung Chloe Tsai: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization; **Leonard Kleinrock:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Meng-Jung Tsai reports financial support was provided by Sunday Group Incorporated. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.-J. Tsai, L. Kleinrock, Computer network optimization using the power metric for multiple flows: Part I, Computer Networks (2025) 111153.

[2] A. Verma, L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, J. Wilkes, Large-scale cluster management at Google with Borg, in: Proceedings of the tenth european conference on computer systems, 2015, pp. 1–17.

[3] M. Tirmazi, A. Barker, N. Deng, M.E. Haque, Z.G. Qin, S. Hand, M. Harchol-Balter, J. Wilkes, Borg: the next generation, in: Proceedings of the fifteenth European Conference on Computer Systems, 2020, pp. 1–14.

[4] B. Burns, J. Beda, K. Hightower, L. Evenson, Kubernetes: up and running: dive into the future of infrastructure, O'Reilly Media, Inc., 2022.

[5] T. Benson, A. Akella, D.A. Maltz, Network traffic characteristics of data centers in the wild, in: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, 2010, pp. 267–280.

[6] A. Roy, H. Zeng, J. Bagga, G. Porter, A.C. Snoeren, Inside the social network's (datacenter) network, in: Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication, 2015, pp. 123–137.

[7] R. Mittal, V.T. Lam, N. Dukkipati, E. Blem, H. Wassel, M. Ghobadi, A. Vahdat, Y. Wang, D. Wetherall, D. Zats, TIMELY: RTT-based congestion control for the datacenter, ACM SIGCOMM Comput. Commun. Rev. 45 (4) (2015) 537–550.

[8] G. Kumar, N. Dukkipati, K. Jang, H.M.G. Wassel, X. Wu, B. Montazeri, Y. Wang, K. Springborn, C. Alfeld, M. Ryan, et al., Swift: Delay is simple and effective for congestion control in the datacenter, in: Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication, 2020, pp. 514–528.

[9] M. Alizadeh, A. Greenberg, D.A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, M. Sridharan, Data center tcp (dctcp), in: Proceedings of the ACM SIGCOMM 2010 Conference, 2010, pp. 63–74.

[10] Y. Zhu, H. Eran, D. Firestone, C. Guo, M. Lipshteyn, Y. Liron, J. Padhye, S. Raindel, M.H. Yahia, M. Zhang, Congestion control for large-scale RDMA deployments, ACM SIGCOMM Comput. Commun. Rev. 45 (4) (2015) 523–536.

[11] Y. Li, R. Miao, H.H. Liu, Y. Zhuang, F. Feng, L. Tang, Z. Cao, M. Zhang, F. Kelly, M. Alizadeh, M. Yu, HPCC: high precision congestion control, in: J. Wu, W. Hall (Eds.), Proceedings of the ACM Special Interest Group on Data Communication, SIGCOMM 2019, Beijing, China, August 19deldel-delins–23, 2019, ACM, 2019, pp. 44–58.

[12] M. Alizadeh, S. Yang, M. Sharif, S. Katti, N. McKeown, B. Prabhakar, S. Shenker, pFabric: Minimal near-optimal datacenter transport, ACM SIGCOMM Comput. Commun. Rev. 43 (4) (2013) 435–446.

[13] P.X. Gao, A. Narayan, G. Kumar, R. Agarwal, S. Ratnasamy, S. Shenker, pHost: Distributed near-optimal datacenter transport over commodity network fabric, in: Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies, 2015, pp. 1–12.

[14] B. Montazeri, Y. Li, M. Alizadeh, J. Ousterhout, Homa: A receiver-driven low-latency transport protocol using network priorities, in: Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication, 2018, pp. 221–235.

[15] L. Chen, K. Chen, W. Bai, M. Alizadeh, Scheduling mix-flows in commodity datacenters with karuna, in: Proceedings of the 2016 ACM SIGCOMM Conference, 2016, pp. 174–187.

[16] M.R.S. Katebzadeh, P. Costa, B. Grot, Saba: Rethinking Datacenter Network Allocation from Application's Perspective, in: Proceedings of the Eighteenth European Conference on Computer Systems, 2023, pp. 623–638.

[17] M. Chowdhury, I. Stoica, Coflow: A networking abstraction for cluster applications, in: Proceedings of the 11th ACM Workshop on Hot Topics in Networks, 2012, pp. 31–36.

[18] M. Chowdhury, I. Stoica, Efficient coflow scheduling without prior knowledge, ACM SIGCOMM Comput. Commun. Rev. 45 (4) (2015) 393–406.

[19] S. Agarwal, S. Rajakrishnan, A. Narayan, R. Agarwal, D. Shmoys, A. Vahdat, Sincronia: Near-optimal network design for coflows, in: Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication, 2018, pp. 16–29.

[20] Y. Zhang, G. Kumar, N. Dukkipati, X. Wu, P. Jha, M. Chowdhury, A. Vahdat, Aequitas: admission control for performance-critical RPCs in datacenters, in: Proceedings of the ACM SIGCOMM 2022 Conference, 2022, pp. 1–18.

[21] L. Kleinrock, A Delay Dependent Queue Discipline, Naval Res. Logist. Q. 11 (4) (1964).

[22] L. Kleinrock, Power and Deterministic Rules of Thumb for Probabilistic Problems in Computer Communications, in: ICC'79; International Conference on Communications, Volume 3, 3, 1979, pp. 43.

[23] L. Kleinrock, Internet congestion control using the power metric: Keep the pipe just full, but no fuller, Ad hoc networks 80 (2018) 142–157.

[24] L. Kleinrock, Message delay in communication nets with storage, Ph.D. thesis, Massachusetts Institute of Technology, 1963.

[25] A. Cobham, Priority Assignment in Waiting Line Problems, J. Oper. Res. Soc. Am. 2 (1) (1954) 70–76.

[26] J. Nash, Non-cooperative games, Annals of mathematics (1951) 286–295.

[27] G. Hardin, The tragedy of the commons: the population problem has no technical solution; it requires a fundamental extension in morality, Science 162 (3859) (1968) 1243–1248.

[28] L. Kleinrock, A Conservation Law for a Wide Class of Queueing Disciplines, Naval Res. Logist. Q. 12 (2) (1965) 181–192.