Computer Networks 263 (2025) 111153

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 I

Meng-Jung Chloe Tsai *, Leonard Kleinrock

University of California Los Angeles, Los Angeles, CA 90095, United States

ARTICLE INFO

Keywords:
Optimal network performance
Optimized queueing system
System performance optimization
Power performance metrics
Throughput delay tradeoff
Priority queueing disciplines
Queueing

ABSTRACT

With the rapid expansion of networks and increasing traffic, optimizing network performance has become increasingly important, especially in balancing two competing objectives: increasing throughput and decreasing delay. This paper adopts the Power metric to address this tradeoff, extending the analysis to a general multi-flow model and examining the influence of different queueing disciplines. We introduce three forms of power metrics–**individual power**, **sum of powers**, and **average power**–to capture performance in a multi-flow context. Individual power optimizes each flow's end-to-end performance, while sum of powers and average power provide a system-wide perspective. These three power metrics are analyzed and optimized under an M/M/1 queueing systems setting, considering two extreme flow discrimination priority disciplines–First-Come, First-Served (FCFS) and Head-of-Line (HOL)–to capture their discriminatory effect on response time while maintaining power optimization. This work is a first step in examining the tradeoff of throughput and delay in queueing systems from various perspectives and across different priority group disciplines. The optimization results aim to provide theoretical insights and guidance for system designers in performance optimization.

1. Introduction

Throughput and **response time** (delay) have been two of the most important metrics when optimizing network performance. As networks rapidly expand and traffic continues to rise [1,2], optimizing these two performance metrics has become increasingly important. The desire for faster speeds (higher throughput) and quicker responses (lower response time) reflects our natural inclination to access information as quickly as possible. However, achieving both simultaneously presents a challenge: throughput and response time exhibit a **tradeoff**.

Fig. 1 visually represents this tradeoff. The x-axis represents throughput (denoted by λ), which signifies the network's transmission rate, measured in packets (bytes) successfully delivered per second. The y-axis depicts mean response time (denoted by $T(\lambda)$), the average time (sec) taken for a packet to travel from source to destination. To quantitatively optimize the intricate balance between throughput and mean response time, we utilize the **Power metric** (denoted as P). The Power metric was originally defined as the ratio of throughput to mean response time, $\frac{\lambda}{T(\lambda)}$. In this work, we use a slightly different definition for the power metric, normalizing both the numerator and denominator, leading to:

$$P = \frac{\rho}{\mu T(\rho)} \tag{1}$$

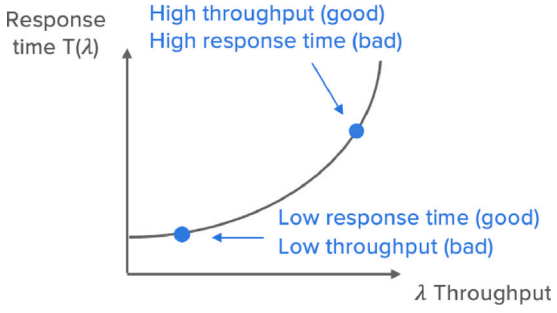

Fig. 1. The tradeoff between throughput and mean response time.

Here, the numerator is the well-known utilization factor, $\rho = \frac{\lambda}{\mu}$. It is the throughput λ normalized by the average service rate μ, where throughput, represented by λ, is equal to the average input arrival rate under the assumption of a no-loss system. For system stability, $\rho < 1$ is required. The denominator, $\mu T(\rho)$, is the mean response time $T(\lambda)$ normalized by the average service time per packet, $\frac{1}{\mu}$. We will explain the normalization in detail in Section 2. We adopt this form of the

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

https://doi.org/10.1016/j.comnet.2025.111153
Received 16 January 2025; Accepted 18 February 2025
Available online 27 February 2025

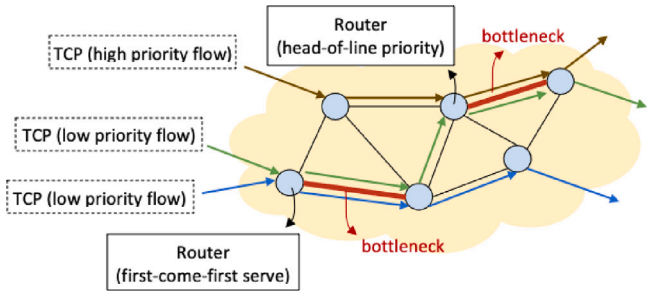

Fig. 2. An example of current networks: multiple flows, multiple hops,[1] different routes, and different queueing disciplines.

power metric to serve as our optimization goal. A higher Power metric signifies a network that efficiently utilizes resources, achieving both high throughput and low response time.

Introduced in [3] and further investigated in subsequent works [4–6], the Power metric has garnered attention for its potential in network congestion control [7]. Its unique strength lies in capturing both throughput and mean response time, providing a holistic view of network performance. Furthermore, the Power metric aligns with the intuitive principle of deterministic reasoning: *keep the pipe just full, but no fuller* [7], as it applies to stochastic systems. By maximizing the Power metric, we aim to strike a balance between high throughput and low response times, ultimately contributing to effective network congestion management.

Previous research on power [3–7] has primarily focused on a *single* flow, typically involving one hop or multiple hops. However, contemporary networks, as illustrated in Fig. 2, presents a level of complexity that far exceeds these simplified scenarios. In this intricate network environment, several factors come into play.

First, *the presence of* ***multiple flows*** *navigating diverse routes and encountering various bottlenecks introduces a heterogeneity*, with each flow serving distinct purposes and requiring a nuanced understanding. Moreover, network traffic is often divided into different classes, each following specific quality of service (QoS) standards [8,9] and assigned different scheduling priorities. For instance, multimedia applications like two-way video streaming and VoIP [10,11] require low latency and high throughput, necessitating prioritization over bulk data transfers that can tolerate higher delays but require high throughput. Frameworks like DiffServ (Differentiated Services) [12,13] and IntServ (Integrated Services) [14] address these needs by enabling differentiated service levels and employing mechanisms like priority queueing [15–17].

Adding to the complexity, networks employ congestion control at two critical points: **end-to-end control** and **router-based control**. *These approaches present distinct optimization challenges due to their different information access and objectives*. **End-to-end control**, implemented in TCP protocols with diverse algorithms like Tahoe [18], Reno [19,20], Vegas [21], Cubic [22], DCTCP [23], Timely [24], BBR [25], HPCC [26], and Swift [27], react to congestion encountered for a given flow along its path, aiming to achieve a balance between maximizing its own throughput, but without complete knowledge of other flows. In contrast, **router-based control** adopts a more holistic perspective, having knowledge of all flows traversing through that router. To achieve an overall efficient allocation of resources, router-based control must differentiate between various flows and strive for "good" performance for all users. Techniques utilized include congestion signaling (ECN [28] and XCP [29]), active queue management (RED [30] and CoDel [31]), and router buffer sizing [32] to address the overall performance of all flows.

Given these complexities, our objective in this paper is to use the power metric to navigate the network landscape to achieve an optimal balance between throughput and response time. Our analysis must look into the intricacies introduced by **multiple flows** and **various queueing disciplines**, accounting for both end-to-end and router perspectives. We study multiple-flow systems with n flows. The ith flow $(i = 1, 2., \dots, n)$ carries a flow of λ_i packets/sec at a utilization factor of $\rho_i = \frac{\lambda_i}{\mu}$.

This research aims to **extend power analysis to current network environments, deriving high-level insights for system designers**. We develop a comprehensive mathematical analysis that accommodates multiple flows and incorporates various aspects of today's network complexity. Our analysis will focus on two core aspects:

1. **Performance**: Characterized by different transformed versions of *power* for multiple flows.
2. **Flow priority discrimination**: Represented by different functions of mean response time as a function of throughput for priorities in different queueing disciplines.

This research represents an initial step towards a comprehensive understanding of the interrelationships between these aspects, how they affect each other, and how to optimize and balance them. In this paper, we focus on the performance optimization for different optimization metrics based on different congestion control algorithms using queueing disciplines with different levels of flow discrimination. To be more specific, we define three forms of power:

1. **Individual Power** of the ith flow $\boldsymbol{P_i} = \frac{\rho_i}{\mu T_i}$
2. **Sum of (Individual) Powers** $\boldsymbol{P_{\text{sum}}} = \sum_{i=1}^{n} P_i$
3. **Average Power** $\boldsymbol{P_{\text{avg}}}$ (see Eq. (58))

We then optimize each of these three performance power metrics for an M/M/1 queueing system with multiple flows, finding the set of optimal utilization factors for each flow $\rho_1^*, \rho_2^*, \dots, \rho_n^*$[2] and their corresponding optimal power values. We examine these optimizations under two queueing disciplines: FCFS (i.e., with minimal flow discrimination), and HOL [15–17] (i.e., with maximal flow discrimination). A more general discussion on flow discrimination in other queueing disciplines will be addressed in future papers.

2. Background

2.1. The single-server queueing system

In its most simplified form, we choose to model a computer network system as a single server queueing system where packets arrive at a rate λ (packets/second), undergo processing within the system, and depart as in Fig. 3.

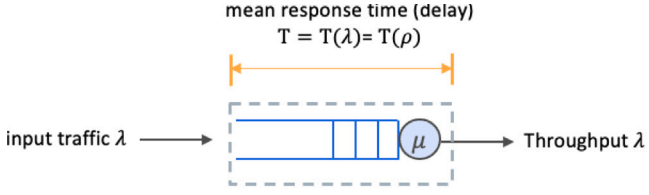

Fig. 3. Model of a computer network system as a loseless single-server queueing system.

[1] In this paper, we focus on one-hop analysis, paralleling the bottleneck, and do not consider the effect of multiple hops.

[2] Throughout this paper, the use of superscript * indicates an optimized value.

Though this is an idealized and simplified representation, this model remains a valuable tool for understanding and analyzing various aspects of network performance, particularly the tradeoff between throughput and delay. We use the following notation to describe the key parameters of this system:

- $\bar{t}$: Mean inter-arrival time of packets, measured in seconds.
- λ: Average arrival rate of packets into the system, measured in packets per second and calculated as $\lambda = \frac{1}{\bar{t}}$. Given that we consider a no-loss system, the number of packets going in equals the number of packets going out. Hence, λ is also considered as throughput in our model.
- $\bar{x}$: Mean service time of a packet, measured in seconds.
- μ: Average service rate of the system, indicating the average number of packets that can be processed per second, calculated as $\mu = \frac{1}{\bar{x}}$.
- ρ: Utilization factor (also known as efficiency), representing the proportion of time the server is actively engaged in serving packets. It is computed as $\rho = \bar{x}/\bar{t} = \lambda/\mu$, with the requirement that $0 \le \rho < 1$ for system stability.[3]
- T: Average (mean) response time (delay), indicating the average duration (measured in seconds) a packet spends within the system, inclusive of both waiting time and service times.[4] Specifically, $T = W + \bar{x}$, where W denotes the average waiting time in queue. This metric typically varies as a function of input traffic, hence we use the notation T(λ). Alternatively, we may denote it as a function of utilization factor, T(ρ) with the input traffic normalized by the service rate.

2.2. The power metric

Power is a metric that combines two competing performance measures, throughput and mean response time (delay), into a single metric[5]. Kleinrock, in [5,7], proposed an alternative definition of power that normalizes both throughput and mean response time. This normalized power is expressed as:

$$P = \frac{\rho}{\mu T} \tag{2}$$

Here, the throughput is transformed into the utilization factor (efficiency) using the equation $\rho = \frac{\lambda}{\mu}$. The mean response time is normalized by dividing it by the no-load response time, $T(0)$, which is equivalent to the average service time, $\frac{1}{\mu}$. This normalization makes the power metric dimensionless. We will use this normalized power definition throughout this document.

2.3. The maximal power operating point

Our objective is to optimize power in order to increase system utilization while keeping mean response time low. This involves finding the operating point that yields the maximum power value, [6] sometimes referred to as "*the knee point*" on the system's performance curve. Before reaching the knee point, increasing system utilization usually improves efficiency without significantly increasing mean response time. Therefore, we seek to augment utilization until this knee point is reached. Beyond this threshold, however, any further efficiency gains lead to a disproportionate rise in mean response time.

For the well-known M/M/1 queueing system [33], the normalized response time is given by $\mu T = \frac{1}{1-\rho}$ and thus the power is $P = \frac{\rho}{\mu T} = \rho(1-\rho)$. In [5,7], Kleinrock derived that the optimal utilization operating point that maximizing power occurs at $\rho^* = \frac{1}{2}$, where the maximum power itself is $P^* = \rho^*(1-\rho^*) = \frac{1}{4}$. For the general M/G/1 system, where the normalized response time is $\mu T(\rho) = 1 + \frac{\rho(1+C_b^2)}{2(1-\rho)}$, the power is given by

$$P = \frac{\rho}{\mu T} = \frac{\rho}{1 + \frac{\rho(1+C_b^2)}{2(1-\rho)}} \tag{3}$$

Kleinrock derived the optimal operating point for the M/G/1 queueing system [5,7], which is achieved when

$$\rho^* = \frac{1}{1 + \sqrt{\frac{1+C_b^2}{2}}} \tag{4}$$

3. Model for multiple flows

3.1. Multiple flows system

Our focus in this paper is for *multiple flows*. The multiple flows queueing system we consider is an M/M/1 system illustrated in Fig. 4. This will be used throughout this document[7]. There are n independent Poisson flows entering the system, with the ith flow having a packet arrival rate of λ_i packets per second. The system service rate is μ packets per second. Packets length of each flow are independently and identically drawn from an exponential distribution where the average service time for the ith flow is $\frac{1}{\mu_i} = \frac{1}{\mu}$ seconds. The utilization factor of each flow is thus $\rho_i = \frac{\lambda_i}{\mu}$. These n independent Poisson processes can be viewed as a combined Poisson process with total average arrival rate of $\lambda = \sum_{i=1}^{n} \lambda_i$ and the total system utilization as $\rho = \sum_{i=1}^{n} \rho_i = \sum_{i=1}^{n} \frac{\lambda_i}{\mu} = \frac{\lambda}{\mu}$.

When the flows are combined, the system can be seen as a single flow. However, we explicitly differentiate each flow here to observe the impact of multiple flows when different approaches are applied to handle the order of packets from various flows being queued and entering service, particularly when different priorities are applied to each flow. There are many queueing disciplines and here we focus on a family of **work-conserving queueing disciplines**[8], represented by the yellow box in Fig. 4. Within this family, first-come, first-served (FCFS) represents the least discriminatory discipline, while head-of-line (HOL) represents the most discriminatory. FCFS and HOL define the upper and lower bounds of flow priority discrimination within this family of queueing disciplines.

3.2. Assumptions and simplification

The model in Fig. 4 resembles a single flow system but focuses on different queueing disciplines to handle the various flows. Compared to real network systems, several simplifications are made below to

[3] $\rho < 1$ here for all flows to be stable. There are situations of $\rho \ge 1$ for HOL where only some of the higher priority group flows are stable while the rest of the lower priority group flows are unstable. We do not consider those cases here.

[4] In the remainder of this paper, by any mention of response time (or delay) we explicitly intend it to be interpreted as "mean response time" (or "mean delay").

[5] It was first introduced by Giessler in [3] as the ratio of throughput to mean response time, $\frac{\lambda}{T}$. This definition parallels the concept of "power" in physics, where power is defined as energy divided by time. In this analogy, throughput corresponds to energy and mean response time (or delay) corresponds to time.

[6] Note that the power value must lie in the range $0 \le P \le 1$.

[7] This document primarily focuses on the M/M/1 queueing model. We will explicitly specify the use of other models, such as the M/G/1 model, when applicable. If we do not explicitly say it, we assume an M/M/1 system.

[8] A "work-conserving" discipline ensures that no work (or service requirement) is created or destroyed within the system, maintaining a constant system workload. As defined in [34], this family of work-conserving queueing disciplines adheres to the following principles: **no defections** (work does not leave the system before completion), **no extra work is created**, and **no server idleness** (the server never idles when work is available).

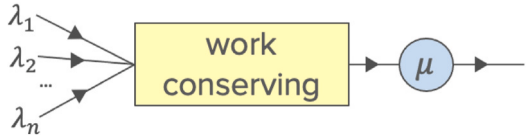

Fig. 4. Model for a single hop M/M/1 system with multiple flows using work-conserving queueing disciplines.

concentrate on understanding the impact of transitioning from a single flow to multiple flows, particularly how different flows with varying throughput and delay requirements compete for system resources. The simplifications we make include:

- **From Multiple Hops to Single Hop**
  The network graph depicted in Fig. 2 consists of multiple hops. However, for our analysis, we simplify the network to a single hop. This simplification is justified by the fact that congestion for a given flow typically occurs at that flow's bottleneck, where most of its waiting time arises. By focusing on the analysis only at the bottleneck, we can avoid the influence of multiple hops, allowing us to analyze the effect of multiple flows more clearly.
- **Assume M/M/1**
  As stated above, each flow is assumed to arrive from an independent Poisson process, and the required service time of each packet is independently and identically selected from an exponential distribution, and the average service time for each flow is identical. We opt for the M/M/1 model [33] here as it simplifies the computation regarding mean response time. This choice facilitates easier analysis, allowing us to uncover potentially hidden insights.[9] The M/M/1 model is the default unless otherwise explicitly stated.
- **Focus on Two Queueing Disciplines**
  The two queueing disciplines that we initially focus on are first-come, first-served (FCFS) and head-of-line preemptive-resume priority queueing (HOL) [15–17]. The choice of these two disciplines is motivated by three key factors: common use in practice, representation of the least (FCFS) and most (HOL) discriminatory behavior based on priority groups, and the relative simplicity of their response time formulas for theoretical analysis.

Having established these simplifications, we now detail the two chosen queueing disciplines and introduce their response times in the M/M/1 setting.

First-Come, First-Served (FCFS)

In the FCFS system, each flow is treated the same in that each flow joins the same single queue and has the same mean response time [33]:

$$T_i = T = \frac{1}{\mu(1-\rho)} \quad \text{for all} \quad i = 1, \ldots, n \tag{5}$$

where ρ represents the total utilization of the system. This mean response time is equal for all flows and is mainly determined by the total system utilization $\rho = \frac{\lambda}{\mu}$.

Head-of-Line Preemptive Resume Priority (HOL)

In the head-of-line (HOL) system, as depicted in Fig. 5, a packet from group i has priority and can cut in line ahead of all packets from groups $i+1, i+2, \ldots, n$. Specifically, under preemptive resume priority[10] (the most discriminative of work-conserving queueing disciplines), a higher-priority packet can preempt a lower-priority packet currently being served, with the lower-priority packet's service later resuming from the point of interruption. In our model, we assume that priority decreases with increasing group number; thus, group 1 has the highest priority, and group n has the lowest.

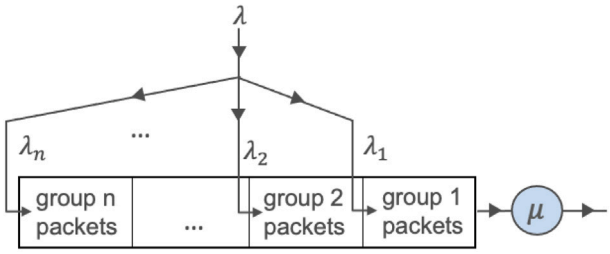

Fig. 5. Head-of-Line (HOL) priority queueing, where higher-priority packets, even arriving late, are placed ahead of lower-priority groups' packets.

The response time in HOL for each priority group i is [16]:

$$T_i = \frac{1}{\mu(1-\sigma_{i-1})(1-\sigma_i)} \tag{6}$$

where

$$\sigma_i = \sum_{j=1}^{i} \rho_j \tag{7}$$

This formula demonstrates the different response times for different priority groups, with higher-priority groups experiencing shorter response times compared to lower-priority groups.

These two disciplines represent the extremes in terms of the difference in each flow's response time among work-conserving queueing policies [15]: FCFS is the least discriminatory, resulting in identical response times for all flows, while HOL is the most discriminatory, leading to the largest difference in response times between priority groups.

In the following three sections, we introduce three different optimization power metrics, one in each section. For each, we will define the optimization metric, find the optimized ρ^* that maximizes each power metric, and compute the corresponding optimal power.

4. Performance optimization metric 1: Individual power, P_i

4.1. Description of the end-to-end viewpoint

The end-to-end perspective refers to congestion control mechanisms implemented at the endpoints of a communication system. For instance, this would be the TCP congestion control at the transport layer [18–27] or the adaptive bitrate algorithm for video streaming at the application layer [35–39]. This viewpoint emphasizes the experience of each end user, leading to the concept of "individual power". This term is defined in terms of the throughput and delay experienced by individual flows, providing a user-centric metric of network performance.

4.2. Definition

The definition of "individual power" for the ith flow is:

$$P_i = \frac{\rho_i}{\mu T_i(\rho_i)} \tag{8}$$

In this equation, ρ_i represents the utilization factor of the ith flow, and $\mu T_i(\rho_i)$ is its normalized mean response time. The term $T_i(\rho_i)$ denotes the mean response time for flow i, which depends on ρ_i. The subscript i in T indicates that the response time may vary for each flow, particularly when the queueing discipline is not FCFS. The denominator in Eq. (8) shows that the mean response time $T_i(\rho_i)$ is normalized by its

[9] Future studies involving other service time distributions, (e.g. deterministic) might reveal how general these insights beyond this distribution might apply.

[10] In the following, "HOL" refers to the preemptive resume form of head-of-line priority queueing.

no-load response time, $\frac{1}{\mu}$, which represents the average service time of a packet.

With the definition of individual power established, we proceed to the optimization, considering two scenarios: singly optimizing flow i's individual power and then jointly optimizing all flows' individual powers.

4.3. Singly optimizing individual power

We first focus on singly optimizing flow i's individual power ($1 \le i \le n$). Our objective is to determine the value of ρ_i^* that maximizes the metric $P_i = \frac{\rho_i}{\mu T_i}$, assuming fixed utilizations for all other flows. The optimal ρ_i^* is found by solving:

$$\frac{dP_i}{d\rho_i} = 0 \tag{9}$$

In the following, we derive the individual power P_i and determine the optimal utilizations ρ_i^* first for FCFS and then for the HOL queueing discipline.

4.3.1. FCFS

In an M/M/1 system with n flows under the FCFS queueing discipline, the response time T_i is the same for all flows and depends on the total system utilization ρ, as stated in Eq. (5). Thus, for the ith flow, the individual power given by Eq. (8) is

$$P_i = \frac{\rho_i}{\mu T_i} = \rho_i(1-\rho) \tag{10}$$

To emphasize the individual impact of utilization on response time, we separate the utilization of the ith flow, ρ_i, from the total system utilization, ρ. We use α_i to represent the sum of the utilizations of all other flows besides the ith, defined by the formula:

$$\alpha_i = \sum_{j=1, j\neq i}^{n} \rho_j \tag{11}$$

The individual power for flow i is then:

$$P_i = \rho_i(1-\rho) = \rho_i(1 - \sum_{j=1, j\neq i}^{n} \rho_j - \rho_i) = \rho_i(1-\alpha_i-\rho_i)$$

Setting the derivative of P_i with respect to ρ_i to zero (Eq. (9)) yields:

$$\frac{dP_i}{d\rho_i} = \frac{d\rho_i(1-\alpha_i-\rho_i)}{d\rho_i} = 1-\alpha_i-2\rho_i = 0$$

Solving for ρ_i gives the optimal utilization ρ_i^* as:

$$\rho_i^* = \frac{1-\alpha_i}{2} = \frac{1-\sum_{j=1, j\neq i}^{n}\rho_j}{2}$$

The corresponding optimal individual power P_i^* is:

$$P_i^* = \rho_i^*(1-\alpha_i-\rho_i^*) = \frac{1-\alpha_i}{2}(1-\alpha_i-\frac{1-\alpha_i}{2}) = (\frac{1-\alpha_i}{2})^2$$

Note that P_i^* is independent of ρ_i for FCFS (but does depend on all ρ_j for $j \neq i$). The above results are summarized in the following theorem:

Theorem 4.1. *In an M/M/1 system employing a FCFS queueing discipline, the optimal ρ_i^* for the flow i ($1 \le i \le n$) to maximize P_i is half of the remaining utilization (i.e., the utilization not used by other flows):*

$$\rho_i^* = \frac{1-\alpha_i}{2} \tag{12}$$

with $\alpha_i = \sum_{j=1, j\neq i}^{n}\rho_j$ indicating the portion of utilization occupied by other flows. The maximal individual power value for the ith flow is:

$$P_i^* = \left(\frac{1-\alpha_i}{2}\right)^2 \tag{13}$$

Notably, the well-known result that the optimal value $\rho^* = 0.5$ for a single flow [5–7] as stated in Section 2 aligns with this theorem. That is, it is the case when $\alpha_i = 0$, as there are no other flows in the system, allowing the entire channel to be available for that single flow, which leads to the optimal utilization value being $\rho^* = 0.5$.

4.3.2. HOL

Having analyzed the minimal flow discrimination case under FCFS, we now consider the maximal flow discrimination case under the head-of-line (HOL) preemptive resume priority queueing discipline. The response time for flow i is given by Eq. (6) in Section 3. The individual power for flow i therefore:

$$P_i = \frac{\rho_i}{\mu T_i} = \rho_i(1-\sigma_i)(1-\sigma_{i-1}) = \rho_i(1-\sigma_{i-1}-\rho_i)(1-\sigma_{i-1}) \tag{14}$$

To find ρ_i^* that maximizes P_i, we set its derivative with respect to ρ_i to zero (Eq. (9)):

$$\frac{dP_i}{d\rho_i} = \frac{d\rho_i(1-\sigma_{i-1}-\rho_i)(1-\sigma_{i-1})}{d\rho_i} = 0$$

Solving this gives:

$$\rho_i^* = \frac{1-\sigma_{i-1}}{2}$$

The optimal individual power P_i^* is then:

$$\begin{aligned} P_i^* &= \rho_i^*(1-\sigma_{i-1}-\rho_i^*)(1-\sigma_{i-1}) \\ &= \frac{1-\sigma_{i-1}}{2}\left(1-\sigma_{i-1}-\frac{1-\sigma_{i-1}}{2}\right)(1-\sigma_{i-1}) = \frac{(1-\sigma_{i-1})^3}{4} \end{aligned}$$

For HOL, P_i^* is also independent of ρ_i, as was the case for FCFS. The preceding results are summarized in the following theorem:

Theorem 4.2. *In an M/M/1 system with HOL, the optimal ρ_i^* for the flow i ($1 \le i \le n$) to maximize P_i is half of the remaining utilization after accounting for higher-priority flows:*

$$\rho_i^* = \frac{1-\sigma_{i-1}}{2} \tag{15}$$

with $\sigma_{i-1} = \sum_{j=1}^{i-1}\rho_j$ denoting the aggregate utilizations of higher-priority flows. The resulting maximal individual power is then:

$$P_i^* = \frac{(1-\sigma_{i-1})^3}{4} \tag{16}$$

4.4. Jointly optimizing individual power

Having singly optimized individual power for the ith flow, we now consider the joint optimization of individual power across all flows. Unlike the singly optimized case, where we focused on the flow i and assumed other flows' utilizations were fixed, the joint optimization accounts for the dynamic (iterative) nature of these utilizations. This dynamic behavior arises because each flow's optimization alters the remaining utilization available to other flows, affecting their optimal ρ_i^*. To determine the equilibrium optimal values of ρ_i^*, we solve the following system of n equations, which represent the optimization condition for each flow:

$$\frac{\partial P_i}{\partial \rho_i} = 0 \quad \text{for} \quad i = 1, \dots, n \tag{17}$$

If a solution exists for this system of n equations (Eq. (17)), simultaneously solving them yields the equilibrium optimal set of ρ_i^*.[11]

Below we perform the joint optimization under both the FCFS and HOL queueing disciplines. The mathematical operation for finding the optimal utilization–setting the ordinary derivative (in the singly optimized case, Eq. (9)) or the partial derivative (in the jointly optimized case, Eq. (17)) of a flow's power with respect to its utilization to zero–is fundamentally the same. Therefore, we use the expressions for ρ_i^* derived in the singly optimized cases (Eq. (12) for FCFS and Eq. (15) for HOL) as components of the systems of n equations to solve for each discipline. We then calculate the optimized total system utilization ρ^*, the optimal individual power for each flow P_i^*, and the sum of these optimized individual powers $P_{\text{sum-of-optimals}}$.[12] Finally, we consider the limiting case where the number of flows approaches infinity.

[11] Note that this set of ρ_i^* represents a Nash equilibrium [40].

4.4.1. FCFS

To determine the equilibrium optimal utilizations ρ_i^* for $i = 1, \ldots, n$ under FCFS, we consider the system of n equations of the form given by Eq. (12). This equation can be rewritten as:

$$\rho_i = 1 - \rho \quad \text{for} \quad i = 1, \ldots, n \tag{18}$$

where $\rho = \sum_{j=1}^{n} \rho_j$. Summing Eq. (18) over all i from 1 to n yields:

$$\sum_{i=1}^{n} \rho_i = \sum_{i=1}^{n} (1 - \rho) \Longrightarrow \rho = n(1 - \rho)$$

Rearranging this gives the *optimal total system utilization*:

$$\rho^* = \frac{n}{n+1} \tag{19}$$

Substituting ρ^* into Eq. (18), we have:

$$\rho_i^* = 1 - \rho^* = 1 - \frac{n}{n+1}$$

After simplification, we obtain the convergent optimal utilization ρ_i^* for each flow:

$$\rho_i^* = \frac{1}{n+1} \quad \text{for} \quad i = 1, \ldots, n \tag{20}$$

Taking Eq. (20) back into Eq. (10), we compute the optimized individual power at convergence:

$$P_i^* = \frac{1}{n+1}\left(1 - \frac{n}{n+1}\right) = \frac{1}{(n+1)^2} \tag{21}$$

Let us now define the **sum of individual powers**:

$$P_{\text{sum}} = \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} \frac{\rho_i}{\mu T_i} \tag{22}$$

For the sum of *optimized* individual powers, we use the notation $P_{\text{sum-of-optimals}}$ to emphasize that each individual power value is optimal:

$$P_{\text{sum-of-optimals}} = \sum_{i=1}^{n} P_i^* \tag{23}$$

Using Eq. (21), we have:

$$P_{\text{sum-of-optimals}} = \sum_{i=1}^{n} P_i^* = \frac{n}{(n+1)^2} \tag{24}$$

These results are summarized in the following theorem:

Theorem 4.3. *In an M/M/1 system with n flows using FCFS, when each flow jointly optimizes its individual power* $P_i = \rho_i(1 - \rho)$

- *The optimal operating point* $(\rho_1^*, \rho_2^*, \ldots, \rho_n^*)$ *at convergence is:*

$$\rho_i^* = \frac{1}{n+1} \quad \textit{for} \quad i = 1, \ldots, n \tag{25}$$

- *The optimized total system utilization is:*

$$\rho^* = \frac{n}{n+1} \tag{26}$$

- *The corresponding optimized individual power for each flow at convergence is:*

$$P_i^* = \frac{1}{(n+1)^2} \quad \textit{for} \quad i = 1, \ldots, n \tag{27}$$

- *The sum of optimized individual powers is:*

$$P_{\textit{sum-of-optimals}} = \sum_{i=1}^{n} P_i^* = \frac{n}{(n+1)^2} \tag{28}$$

[12] Note that in this section we show the *sum of the optimized individual powers* (as defined in Eq. (23)). However, the sum of individual powers (as defined in Eq. (22)) might not be maximal. We explore the *optimal sum of individual powers* in Section 5.

We do not use the superscript * for $P_{\text{sum-of-optimals}}$ because optimizing each individual power does not guarantee a maximum sum of powers P_{sum}^*. The optimal individual power given by Eq. (27) represents a convergent balance for each flow when optimizing its individual power.

Note that $\rho^* = \frac{n}{n+1}$ (Eq. (26)) is strictly less than 1 for finite n, indicating that the system remains stable. However, in the limit as $n \to \infty$, the system becomes unstable as ρ^* approaches 1. The asymptotic behavior of the optimization results as n approaches infinity is summarized in the following corollary:

Corollary 4.1. *Consider an M/M/1 system with n flows under the FCFS queueing discipline and where each flow i jointly optimizes its individual power* $P_i = \rho_i(1 - \rho)$. *As n approaches infinity, the limiting behavior is as follows:*

- *The optimized total system load* ρ^* *approaches 1:*

$$\lim_{n\to\infty} \rho^* = \lim_{n\to\infty} \frac{n}{n+1} = 1 \tag{29}$$

- *The optimized individual power* P_i^* *at convergence for each flow approaches 0:*

$$\lim_{n\to\infty} P_i^* = \lim_{n\to\infty} \frac{1}{(n+1)^2} = 0 \tag{30}$$

- *The sum of optimized individual powers* $P_{\textit{sum-of-optimals}} = \sum_{i=1}^{n} P_i^*$ *also approaches 0:*

$$\lim_{n\to\infty} P_{\textit{sum-of-optimals}} = \lim_{n\to\infty} \sum_{i=1}^{n} P_i^* = \lim_{n\to\infty} \frac{n}{(n+1)^2} = 0 \tag{31}$$

Based on the limiting behavior from Eqs. (29), (30), and (31), we observe that as the number of flows increases significantly and each flow optimizes its individual power simultaneously, both the optimized individual power $P_i^* = \frac{1}{(n+1)^2}$ and the sum of the optimized individual powers $P_{\text{sum-of-optimals}} = \sum_{i=1}^{n} P_i^* = \frac{n}{(n+1)^2}$ diminish to zero. This indicates that each flow experiences a significant response time, causing its individual power to approach zero. Morever, the summation of individual powers also trends towards zero. This scenario reflects reduced benefits for both individual flows and the system as a whole as the number of flows increases, resembling the well-known concept of ***the tragedy of the commons*** [41].

In addition, this finding from our theoretical analysis is consistent with practical simulation results that highlight how numerous TCP flows can lead to high loss rates and delays, as demonstrated in studies [42–44]. For instance, Morris in [42] found that the loss rate can reach as high as 17% with 1500 TCP flows. While our analysis assumes a lossless system, with a theoretical very large buffer, in reality, the significant delays identified in our results would likely translate to high loss rates when taking into account the limited size of actual buffers. Thus, this theoretical analysis offers valuable insights into the potential challenges posed by a large number of TCP flows, aligning with empirical observations in practical scenarios.

4.4.2. HOL

We now proceed to the HOL case. To solve the system of n equations given by Eq. (15) for $i = 1, \ldots, n$, we first examine the equations for $i = 1$ and $i = 2$, finding that $\rho_1^* = \frac{1}{2}$ and $\rho_2^* = \frac{1}{4}$. We observe that ρ_i^* follows a geometric sequence where each term is half of the preceding one. This pattern for $i = 1, 2, 3, 4, \ldots$ is $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$, and so forth. Based on this pattern, we hypothesize that the formula for each flow's optimal utilization ρ_i^* at convergence is:

$$\rho_i^* = \left(\frac{1}{2}\right)^i \quad \text{for } i = 1, 2, \ldots, n \tag{32}$$

Substituting Eq. (32) into Eq. (15) confirms the equation, indicating the correctness of the formula. Next, we calculate the optimum total utilization at convergence by summing the individual optimal utilizations

ρ_i^*:

$$\rho^* = \sum_{i=1}^{n} \rho_i^* = \sum_{i=1}^{n} (\frac{1}{2})^i = \frac{\frac{1}{2}(1-(\frac{1}{2})^n)}{1-\frac{1}{2}}$$

Simplifying this gives:

$$\rho^* = 1 - \left(\frac{1}{2}\right)^n \tag{33}$$

Substituting Eq. (32) into Eq. (14), we compute the optimized individual power at convergence:

$$P_i^* = \rho_i^*(1-\sigma_i)(1-\sigma_{i-1}) = \rho_i^*(1-\sum_{j=1}^{i}\rho_j^*)(1-\sum_{j=1}^{i-1}\rho_j^*) = (\frac{1}{2})^i\ (\frac{1}{2})^i\ (\frac{1}{2})^{i-1}$$

Simplifying this expression yields:

$$P_i^* = 2\cdot(\frac{1}{8})^i \quad \text{for } i = 1, 2, \ldots, n \tag{34}$$

Summing the optimized individual powers, we have:

$$P_{\text{sum-of-optimals}} = \sum_{i=1}^{n} P_i^* = \sum_{i=1}^{n} 2\cdot(\frac{1}{8})^i = 2\cdot\frac{\frac{1}{8}(1-(\frac{1}{8})^n)}{1-\frac{1}{8}}$$

This expression simplifies to:

$$P_{\text{sum-of-optimals}} = \sum_{i=1}^{n} P_i^* = \frac{2}{7}\cdot\left(1-(\frac{1}{8})^n\right) \tag{35}$$

From Eq. (32) and Eq. (34), we can see that the optimal value for ρ_i^* and P_i^* are independent of the number of flows, n. This is because the response time of higher priority groups is not affected by lower priority groups. Hence, as the number of flows increases, the value of ρ_i^* and P_i^* for higher priority flows already in the system remains unchanged. They are not affected by the joining of subsequent lower priority flows. For example, the first flow always has its $\rho_1^* = 0.5$ with $P_1^* = 0.25$ regardless of how many subsequent flows join the system. This leads us to the following theorem, which formalizes the independence of the optimal equilibrium values from the number of flows and summarizes the optimization result:

Theorem 4.4. *For an M/M/1 system with n flows using the preemptive resume HOL queueing discipline, when each flow jointly optimizes its individual power $P_i = \rho_i(1-\sigma_i)(1-\sigma_{i-1})$, at convergence, the optimum utilization factor ρ_i^* for each flow i (where $i \le n$) and the corresponding value of optimized individual power P_i^* are both independent of the number of flows in the system.*

In other words, for $i \le n$, the values of ρ_i^* and P_i^* are invariant to increases in n, meaning ρ_i^* and P_i^* remain the same as the number of flows increases from n to $n+k$, where k is any arbitrary positive integer.

At convergence, the optimal values are as follows:

- *The optimal utilization factor for each flow i at convergence is:*

$$\rho_i^* = (\frac{1}{2})^i \quad \text{for } i = 1, 2, \ldots, n \tag{36}$$

- *The optimum total utilization factor is:*

$$\rho^* = 1 - \left(\frac{1}{2}\right)^n \tag{37}$$

- *The optimized individual power for flow i at convergence is:*

$$P_i^* = 2\cdot(\frac{1}{8})^i \quad \text{for } i = 1, 2, \ldots, n \tag{38}$$

- *The sum of optimized individual powers is:*

$$P_{\text{sum-of-optimals}} = \sum_{i=1}^{n} P_i^* = \frac{2}{7}\cdot(1-(\frac{1}{8})^n) \tag{39}$$

Given that the optimal individual power for each flow remains constant regardless of the number of flows in the system, the sum of the optimized individual powers for the entire system increases as the number of flows increases. Each newly added flow, with lower priority, does not affect the power sum of higher priority flows and contributes its own power value to the system. This accumulation continues with new flows experiencing increasing waiting times, resulting in almost zero individual power for very low priority flows. Consequently, as more flows are added, the sum of optimized individual powers gradually converges to a limiting value. Specifically, as the number of flows approaches infinity, the total sum of the individual powers converges to $\frac{2}{7}$.

Corollary 4.2. *Consider an M/M/1 system with n flows under an HOL queueing system where each flow i optimizes its individual power $P_i = \rho_i(1-\rho)$. As n approaches infinity, the limiting behavior is as follows:*

- *The optimized total system load ρ^* approaches 1*

$$\lim_{n\to\infty} \rho^* = \lim_{n\to\infty} 1-(\frac{1}{2})^n = 1 \tag{40}$$

- *The sum of optimized individual powers $P_{\text{sum-of-optimals}} = \sum_{i=1}^{n} P_i^*$ approaches $\frac{2}{7}$*

$$\lim_{n\to\infty} P_{\text{sum-of-optimals}} = \lim_{n\to\infty} \sum_{i=1}^{n} P_i^* = \lim_{n\to\infty} \frac{2}{7}\cdot(1-(\frac{1}{8})^n) = \frac{2}{7} \tag{41}$$

4.5. Comparison of joint individual power optimization results for FCFS and HOL

Table 1 summarizes the convergent results when each flow jointly optimizes its individual power under FCFS and HOL, as derived in Section 4.4. It presents the equilibrium optimal individual utilizations ρ_i^* for each flow, the optimum total system utilization ρ^*, and the corresponding optimized individual powers along with their sum. The table also includes the limiting behavior of ρ^* and the sum of P_i^*.

Table 1
Jointly optimized individual power optimization at convergence for FCFS and HOL.

	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	$\frac{2}{7}$

Fig. 6 shows the optimal system utilization $\rho^* = \sum_{i=1}^{n} \rho_i^*$ for different values of n under FCFS and HOL (values from the second row in Table 1). As discussed previously, as the number of flows n tends to infinity, the optimum system utilization ρ^* approaches unity for both FCFS and HOL. For other work-conserving queueing disciplines, we conjecture that the curves of the optimized system utilization ρ^* with the number of flows n at equilibrium lie between the curves of FCFS and HOL, with HOL as the upper bound and FCFS as the lower bound. Furthermore, Fig. 6 also shows that the optimum system utilization ρ^* approaches 1 in HOL faster than in FCFS.

To assess overall performance, we use the summation of optimized individual powers to compare the optimization results. The values for FCFS and HOL, indicated in the fifth row of Table 1, are used to plot Fig. 7, which shows that the sum of powers is greater in HOL than in FCFS for each n. Furthermore, the figure shows how the sum of optimized individual powers changes differently with the number of flows in the system. In the HOL system, the sum of optimized individual powers

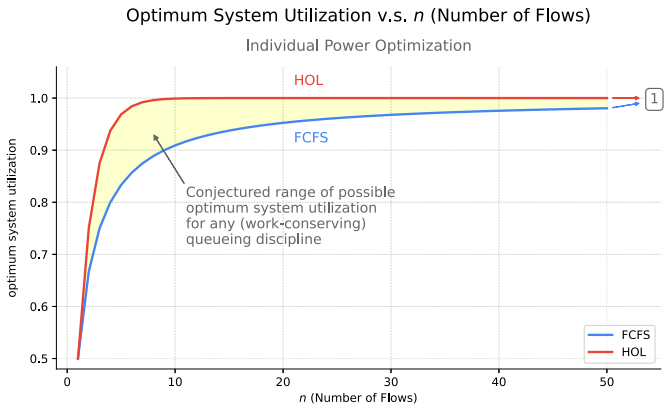

Fig. 6. Trend of the optimum system utilization ρ^* as the number of flows n increases. The HOL and the FCFS are conjectured to be the upper and lower bounds. The yellow region between FCFS and HOL is conjectured to be the range of possible optimum system utilization for any other work-conserving priority discipline.

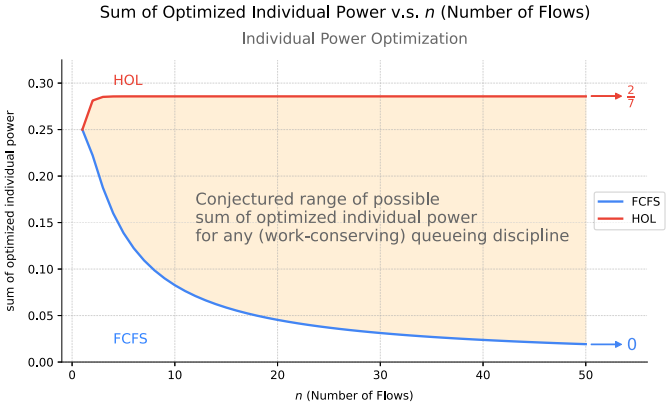

Fig. 7. Trend of the optimized individual power summation versus the number of flows n. HOL and FCFS are conjectured as the upper and lower bounds, respectively. The tan region between them is the conjectured range of possible sum of the optimized individual power for any other work-conserving priority discipline.

increases with the number of flows and rapidly converges towards $\frac{2}{7}$. Conversely, in the FCFS system, this metric decreases, approaching 0 as the number of flows grows to infinity. This demonstrates that different queueing disciplines, which introduce varying levels of discrimination among flows, lead to divergent trends in overall performance. The two queueing disciplines, FCFS and HOL, are conjectured to serve as the bounds for the possible sum of optimized individual power values for any other work-conserving priority discipline.

To summarize, Section 4 introduced individual power and explored both singly optimizing and jointly optimizing individual power under FCFS and HOL. We specifically used the "sum of optimized individual powers" to evaluate the convergent results of the joint optimization for these two disciplines with different flow discrimination priorities. However, the "sum of individual powers" computed here was not maximized because it was not the primary optimization objective, as we focused on the end-to-end (individual flow) perspective.

In the next section, we shift our perspective from this end-to-end viewpoint to a system-wide viewpoint and change the optimization goal from maximizing individual power to maximizing the *sum of individual powers*, P^*_{sum}, to enhance overall system performance; this is our metric 2. We will discuss how this system-wide viewpoint aligns with the sum of individual powers as a metric. We will then identify the optimal utilization factor ρ_i^* for each flow $i = 1, \ldots, n$, which collectively maximizes the sum of individual powers for both FCFS and HOL.

5. Performance optimization metric 2: Sum of individual powers, P_{sum}

5.1. Description of the system-wide viewpoint

When each flow optimizes its individual power, the total system resources may not be efficiently utilized. For example, in the FCFS case, the system can become overwhelmed when each flow optimizes its own power, causing long mean response times for all flows and leading to an almost zero sum of individual power, as discussed previously.

From a system operator point of view, the goal should perhaps be optimizing the overall benefit for the system. A straightforward approach to achieving this is to take the sum of individual powers as the optimization target. Compared to considering just the throughput (which corresponds to system utilization) or just the mean response time of each flow, using the sum of individual powers not only retains the benefits of balancing two competing performance metrics (throughput and response time) but also strives to utilize system resources efficiently and enhance overall performance.

One corresponding congestion control mechanism that requires a system-wide perspective is the active queue management mechanism in routers. Even though routers lack direct control over incoming traffic from sources managed by end systems, they can indirectly influence traffic by transmitting congestion signals to the end systems (e.g., DECbit [45], ECN [28]) or by preemptively dropping packets before buffers reach capacity (e.g., RED [46]). These actions prompt TCP to adjust its congestion window, thereby reducing input rates. Consequently, effective congestion control mechanisms in routers necessitate an understanding of when to trigger these control mechanisms and how to execute them. Additionally, routers may need to prioritize certain types of traffic, such as delay-sensitive traffic, to ensure lower response times for these flows. However, this prioritization may negatively impact other traffic, resulting in higher response times or even causing starvation. In that case, how routers manage each flow's volume and the utilization ratio of high and low priority flows to prevent starvation becomes critical, making it essential for routers under different queueing disciplines to ascertain the optimal traffic volume each input flow should maintain within the system, i.e., each flow's utilization factor ρ_i.

Moreover, a system operator could use individual power to charge each user, as a higher value of power typically indicates more throughput usage or higher priority in being served to achieve lower mean response times. The goal of maximizing the sum of power could be used to maximize revenue for the system operator, as it aligns the operator's financial incentives with the efficient utilization of system resources and improved overall performance.

5.2. Definition

Recall that P_{sum} is defined in Eq. (22) as:

$$P_{\text{sum}} = \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} \frac{\rho_i}{\mu T_i} \tag{42}$$

This formula aggregates the power of each flow, ensuring every flow is considered. A system with flow discrimination focusing solely on high-priority flows can negatively impact lower-priority flows. By focusing on improving (maximizing) the sum of individual powers here, we ensure that higher-priority flows do not operate independently of lower-priority flows, preventing potential negative impacts to lower priorities. This aggregation provides a more balanced resource allocation, especially in HOL systems, where higher-priority flows may be unaware of lower-priority flows, as they can bypass them in the queue.

5.3. Optimizing sum of individual powers, P^*_{sum}

We now seek to find the **maximal** value of the sum of individual powers, which we denote as P^*_{sum}. Specifically, we aim to determine the operating points for the set of utilizations for each flow $\rho_1^*, \rho_2^*, \dots, \rho_n^*$ that collectively maximize this sum of individual powers. To this end, we identify the critical points of the sum of individual powers by calculating the partial derivatives with respect to each ρ_i and setting them to zero:

$$\frac{\partial}{\partial \rho_i} P_{\text{sum}} = \frac{\partial}{\partial \rho_i} \sum_{j=1}^{n} P_j = \frac{\partial}{\partial \rho_i} \sum_{j=1}^{n} \frac{\rho_j}{\mu T_j} = 0 \quad \text{for } i = 1, 2, \dots, n$$

By solving these equations simultaneously, we find the critical points for each ρ_i, which, as usual, we denote as ρ_i^* (for $i = 1, \dots, n$). These critical points maximize the sum of individual powers. In the following, we determine the optimal utilization for each flow that maximizes the sum of powers in FCFS and HOL, respectively.

5.3.1. FCFS

In FCFS, where jobs are processed in the order they arrive without prioritization, each flow has the same response time, denoted as $T_i = T = \frac{1}{\mu(1-\rho)}$ for $i = 1, \dots, n$. With the individual power given by Eq. (10), the sum of individual powers for the FCFS system is:

$$P_{\text{sum}} = \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} \frac{\rho_i}{\mu T} = \sum_{i=1}^{n} \rho_i (1 - \rho) \tag{43}$$

resulting in:

$$P_{\text{sum}} = \rho(1 - \rho) \tag{44}$$

To maximize the sum of powers, we take the partial derivative of this sum with respect to each flow's utilization and set it equal to zero:

$$\frac{\partial P_{\text{sum}}}{\partial \rho_i} = \frac{\partial \rho(1-\rho)}{\partial \rho_i} = 0 \quad \text{for } i = 1, 2, \dots, n \tag{45}$$

Since $\frac{\partial \rho}{\partial \rho_i} = \frac{\partial \sum_{i=1}^{n} \rho_i}{\partial \rho_i} = 1$, we apply the chain rule to change the variable in Eq. (45), resulting in:

$$\frac{\partial P_{\text{sum}}}{\partial \rho_i} = \frac{\partial \rho(1-\rho)}{\partial \rho_i} = \frac{\partial \rho(1-\rho)}{\partial \rho} \frac{\partial \rho}{\partial \rho_i} = 1 - 2\rho = 0 \qquad \text{for } i = 1, 2, \dots, n$$

Solving this equation, we find:

$$\rho^* = \frac{1}{2} \tag{46}$$

This implies that the sum of individual powers in the FCFS system is maximized when the total system utilization ρ^* is equal to $\frac{1}{2}$. This maximum is achieved regardless of the distribution of individual utilizations ρ_i^* among the flows, as long as they are non-negative and sum to $\rho^* = \frac{1}{2}$. This result is consistent with the result mentioned in Section 2 that maximum power occurs at $\rho^* = \frac{1}{2}$ for a single flow in an M/M/1 system. This consistency is not surprising because in FCFS, the superposition of n Poisson flows is equivalent to a single Poisson flow at a traffic level of ρ.

Substituting Eq. (46) into Eq. (44), we have the maximal sum of individual powers value: $P^*_{\text{sum}} = \rho^*(1 - \rho^*) = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4}$. This gives us the following interesting theorem:

Theorem 5.1. *In an M/M/1 system with n flows using the FCFS queueing discipline, the sum of individual powers reaches it maximal value when the sum of utilizations $\rho^* = \sum_{i=1}^{n} \rho_i^* = \frac{1}{2}$. This is **independent** of the distribution of ρ_i^* for $i = 1, \dots, n$ as long as their sum*

$$\rho^* = \frac{1}{2} \tag{47}$$

In this case, the maximal sum of power value is:

$$P^*_{sum} = \sum_{i=1}^{n} P_i = \frac{1}{4}$$

where the individual power P_i depends on each ρ_i^* but sums to $\frac{1}{4}$.

This result indicates that the system performs most efficiently when the total utilization is at $\frac{1}{2}$, irrespective of the individual allocations of utilization among different flows. This is because each flow's utilization contributes equally to the sum of powers, allowing the function to be expressed in terms of only ρ; this can be seen from Eq. (44). Therefore, this multiple-flow system can be considered as a single-flow system where the total power $\rho(1 - \rho)$ is a quadratic function of ρ and reaches its maximum value at the midpoint of ρ.

The fact that the sum of powers depends solely on system utilization is crucial for optimizing performance and resource management. This characteristic allows for *flexible allocation of individual utilizations*, facilitating the achievement of various optimization objectives. By understanding and leveraging this property, system administrators can dynamically adjust individual flow utilizations to maintain the total system utilization around the optimal point, thereby ensuring maximum efficiency and resource utilization. This **flexibility** is particularly beneficial in complex systems where workload and flow characteristics can vary over time.

5.3.2. HOL

FCFS offers flexibility in optimizing system performance in terms of sum of individual powers. We now proceed to investigate the other extreme of flow priority discrimination: Head-Of-Line (HOL) priority, the most discriminatory priority queueing discipline. We begin with the two-flow case and then extend the analysis to an arbitrary number of flows.

Two Flows

From Eq. (14), we have the individual power of flow 1 is $P_1 = \rho_1(1-\rho_1)$ and the individual power for flow 2 is $P_2 = \rho_2(1-\rho_1)(1-\rho_1-\rho_2)$. The sum of the individual powers in HOL is thus

$$P_{\text{sum}} = P_1 + P_2 = \rho_1(1 - \rho_1) + \rho_2(1 - \rho_1)(1 - \rho_1 - \rho_2) \tag{48}$$

This can be simplified and expressed as:

$$P_{\text{sum}} = (\rho_1 + \rho_2)\,(1 - \rho_1)\,(1 - \rho_2) \tag{49}$$

To find the maximal sum of powers, we need to find the critical point where the partial derivatives of P_{sum} with respect to ρ_1 and ρ_2 are both equal to zero. This can be represented by the following system of equations:

$$\begin{cases} \frac{\partial}{\partial \rho_1} P_{\text{sum}} = 0 \\ \frac{\partial}{\partial \rho_2} P_{\text{sum}} = 0 \end{cases}$$

Substituting $P_{\text{sum}} = (\rho_1 + \rho_2)(1 - \rho_1)(1 - \rho_2)$ into the system of equations above, we get

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

We observe that the terms $(1 - \rho_2)$ and $(1 - \rho_1)$ are non-zero, given the assumption of system stability where $\rho_1 < 1$ and $\rho_2 < 1$. Therefore, we can divide each of these equations by one of these non-zero terms to yield the two equations:

$$\begin{cases} (1 - \rho_1 - \rho_1 - \rho_2) = 0 \\ (1 - \rho_2 - \rho_2 - \rho_1) = 0 \end{cases}$$

Solving these equations simultaneously, we determine that the unique values of ρ_1^* and ρ_2^* that maximize the sum of power for HOL are:

$$\rho_1^* = \rho_2^* = \frac{1}{3} \tag{50}$$

This outcome reveals that the optimal utilization strategy for each flow, aimed at maximizing the summation of individual powers, occurs when both flows are allocated identical amounts of the system's resources. Specifically, each flow should utilize one-third of the system capacity.

This configuration maximizes the system's total power and reflects a balanced approach where each flow contributes equally to achieving optimal efficiency, ensuring both enhanced system performance and a more equitable distribution of resources.

The corresponding power values for flow 1 and flow 2 are:

$$P_1 = \rho_1(1-\rho_1) = \frac{1}{3}\left(1-\frac{1}{3}\right) = \frac{2}{9}$$

$$P_2 = \rho_2(1-\rho_1)(1-\rho_1-\rho_2) = \frac{1}{3}\left(1-\frac{1}{3}\right)\left(1-\frac{1}{3}-\frac{1}{3}\right) = \frac{2}{27}$$

The maximal sum of powers for the head-of-line priority system with two flows is thus:

$$P^*_{\text{sum}} = P_1 + P_2 = \frac{2}{9} + \frac{2}{27} = \frac{8}{27} \approx 0.296$$

This value exceeds the maximum sum of individual powers achievable in an FCFS system, which is 0.25. This trend is consistent with the findings in Section 4, where the sum of powers in HOL was shown to be greater than that in FCFS when jointly optimizing individual power. This consistency across optimizing different performance metrics highlights the greater efficiency of HOL compared to FCFS.

n Flows

In the two-flow case, optimizing the sum of individual powers in HOL results in equal optimal system utilization for each flow. This principle, where each flow's optimum utilization is equal, can be generalized to an arbitrary number of flows with n flows. The optimal utilization for each flow is then given by (see the Appendix for the proof):

$$\rho_i^* = \frac{1}{n+1} \quad \text{for } i = 1,\dots,n \tag{51}$$

Substituting Eq. (51) into the individual power expression for HOL (Eq. (14), where $\sigma_i = \sum_{j=1}^{i}\rho_j$) for each i:

$$P_i = \rho_i(1-\sigma_{i-1})(1-\sigma_i) = \frac{1}{n+1}\left(1-\frac{i-1}{n+1}\right)\left(1-\frac{i}{n+1}\right)$$
$$= \frac{1}{n+1}\cdot\frac{n+2-i}{n+1}\cdot\frac{n+i-i}{n+1} = \frac{(n+1-i)(n+2-i)}{(n+1)^3} \quad \text{for } i = 1,2,\dots,n$$

Next, summing the individual powers across all flows yields the maximal sum of powers:

$$\begin{aligned}
P^*_{\text{sum}} &= \sum_{i=1}^{n} P_i = \frac{1}{(n+1)^3}\sum_{i=1}^{n}(n+1-i)(n+2-i) \\
&= \frac{1}{(n+1)^3}\sum_{i=1}^{n}\left[(n+1)(n+2) - i(2n+3) + i^2\right] \\
&= \frac{1}{(n+1)^3}\left[(n+1)(n+2)n - (2n+3)\frac{n(n+1)}{2} + \frac{n(n+1)(2n+1)}{6}\right] \\
&= \frac{n(n+1)}{(n+1)^3}\left[(n+2) - \frac{2n+3}{2} + \frac{2n+1}{6}\right] \\
&= \frac{n}{6(n+1)^2}\left[6n+12-6n-9+2n+1\right] \\
&= \frac{n(n+2)}{3(n+1)^2}
\end{aligned}$$

The results derived above are summarized in the following theorem:

Theorem 5.2. *Given the HOL preemptive resume priority queueing discipline with n flows in an M/M/1 system, the sum of individual powers $P_{sum} = \sum_{i=1}^{n}\rho_i(1-\sigma_{i-1})(1-\sigma_i)$, where $\sigma_i = \sum_{j=1}^{i}\rho_j$ is optimal when* ***each flow has the same utilization.***

- *Each flow's optimized utilization is*

$$\rho_i^* = \frac{1}{n+1} \quad \text{for } i = 1,2,\dots,n \tag{52}$$

- *The optimum total system utilization is*

$$\rho^* = \frac{n}{n+1} \tag{53}$$

- *The maximal sum of powers is*

$$P^*_{sum} = \frac{n(n+2)}{3(n+1)^2} \tag{54}$$

- *The individual power of each flow i is*

$$P_i = \frac{(n+1-i)\,(n+2-i)}{(n+1)^3} \quad \text{for } i = 1,2,\dots,n \tag{55}$$

Based on Theorem 5.2, the sum of individual powers is maximized when $\rho^* = \frac{n}{n+1}$. Clearly, $\rho^* = \frac{n}{n+1}$ is always less than 1 for positive n. In addition, as n approaches infinity, the asymptotic behavior of the optimization results (optimized utilization and optimal powers) are stated in the following corollary:

Corollary 5.1. *Consider an M/M/1 system with n flows using the HOL preemptive resume queueing discipline. As n approaches infinity, the limit of the optimized system utilization ρ^*, when the sum of individual powers is maximal, is*

- *The optimized total system load ρ^* approaches 1*

$$\lim_{n\to\infty}\rho^* = \lim_{n\to\infty}\frac{n}{n+1} = 1 \tag{56}$$

- *The optimal sum of individual powers $P_{sum} = \sum_{i=1}^{n}P_i^*$ approaches $\frac{1}{3}$*

$$\lim_{n\to\infty}P^*_{sum} = \lim_{n\to\infty}\frac{n(n+2)}{3(n+1)^2} = \frac{1}{3} \tag{57}$$

5.4. Comparison of sum of powers optimization results for FCFS and HOL

We compare the optimization results of FCFS and HOL derived in this section using the sum of individual power metric as the optimization goal in Table 2. In FCFS, the individual utilization factor ρ_i^* is not uniquely determined as long as their sum is 0.5; therefore, we leave the corresponding entry as a comment,[13] in the table. The same applies to P_i for FCFS.

Table 2
Optimization results of using "sum of individual powers" as the optimization objective. 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 13	$\frac{1}{n+1}$
ρ^*	$\frac{1}{2}$	$\frac{n}{n+1}$
$\lim_{n\to\infty}\rho^*$	$\frac{1}{2}$	1
P_i	See Footnote 13	$\frac{(n+1-i)(n+2-i)}{(n+1)^3}$
$P^*_{\text{sum}} = \sum_{i=1}^{n}P_i$	$\frac{1}{4}$	$\frac{n(n+2)}{3(n+1)^2}$
$\lim_{n\to\infty}P^*_{\text{sum}}$	$\frac{1}{4}$	$\frac{1}{3}$

In HOL, some optimized values still depend on n. To better understand how P^*_{sum} and the optimized ρ^* that achieves this maximum vary with n, we plot ρ^* against n in Fig. 8 and P^*_{sum} against n in Fig. 9. In these figures, the HOL and FCFS curves are conjectured to serve as the upper and lower bounds, respectively, for the maximum sum of individual powers P^*_{sum} and the optimized ρ^*. We further conjecture that the curves for other work-conserving queueing disciplines fall within the regions bounded by these two curves. For FCFS, the maximum sum of powers remains constant at 0.25 as n increases in Fig. 9, with ρ^* fixed at 0.5 regardless of the number of flows in Fig. 8. In contrast, HOL,

[13] From Theorem 5.1 any set $(\rho_1^*, \rho_2^*, \dots, \rho_n^*)$ that sums to $\frac{1}{2}$ is optimal. The corresponding individual power values $P_1, P_2, \dots, P_n$ for this set sum to the optimal total power of $\frac{1}{4}$.

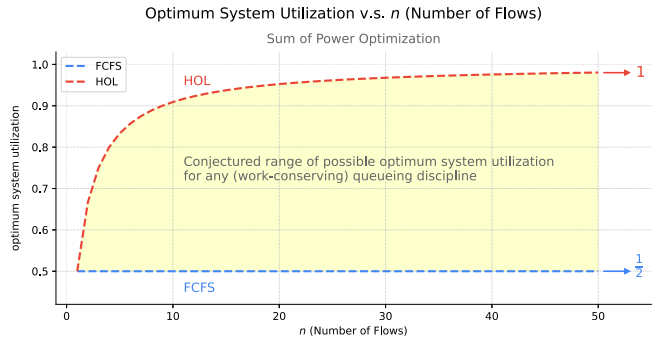

Fig. 8. The ρ^* that maximizes the sum of individual powers is shown for both FCFS and HOL as a function of n.

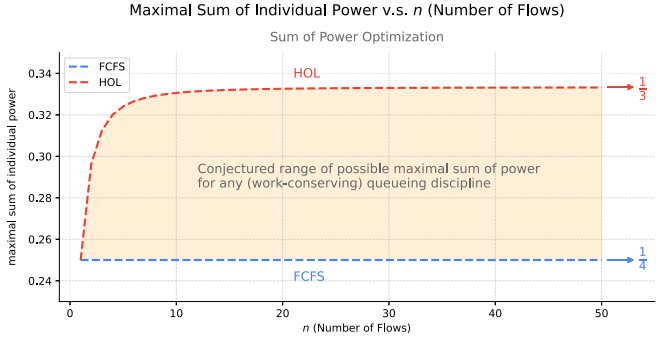

Fig. 9. The maximum sum of individual powers P^*_{sum} as a function of n for both FCFS and HOL.

which represents maximal flow priority discrimination, has a maximum sum of individual powers that increases with n and approaches an asymptotic value of $\frac{1}{3}$ as n becomes large, as shown in Fig. 9, while ρ^* increases towards 1, as shown in Fig. 8.

6. Performance optimization metric 3: Average power, P_{avg}

In this section, we introduce an alternative metric–our third metric, average power–as another approach to evaluate the system's overall performance.

6.1. Definition

Another approach to assessing system performance involves treating the system as a black box, focusing on measuring total system utilization and sampling packets to obtain an average view of response times. From this perspective, we define the average power, denoted by P_{avg}, with the following mathematical expression:

$$P_{\text{avg}} = \frac{\sum_{i=1}^{n} \rho_i}{\sum_{i=1}^{n} \left(\frac{\rho_i}{\rho} \mu T_i \right)} \tag{58}$$

P_{avg} is a form of utilization divided by response time, but specifically a load-weighted average response time. In the definition given by Eq. (58), the numerator represents the summation of the utilization factor for each traffic flow, where ρ_i corresponds to the utilization factor of the ith flow, computed as $\rho_i = \frac{\lambda_i}{\mu}$. Clearly, the numerator is simply ρ. The denominator represents the *weighted* average response time of each flow, with weights determined by their respective utilization factor fractions within the system, namely $\frac{\rho_i}{\rho}$. As usual. the mean response time for each flow, denoted as T_i, is a function of ρ_i and may vary depending on the queueing discipline employed.

With this definition, the average power can be expressed in a form that solely depends on ρ, leveraging the conservation law [47] and using the assumption that each flow has the same mean service time, represented by $\frac{1}{\mu_i} = \frac{1}{\mu}$ for $i = 1, \ldots, n$. The results in this section are not confined to M/M/1 queue systems but extend to all M/G/1 work-conserving systems as well. We establish the following theorem:

Theorem 6.1. *For an **M/G/1** system with n flows, all having the same mean service time $\frac{1}{\mu}$ and operating under any **work-conserving** queueing discipline, the average power, defined as:*

$$P_{avg} = \frac{\sum_{i=1}^{n} \rho_i}{\sum_{i=1}^{n} \left(\frac{\rho_i}{\rho} \mu T_i \right)} \tag{59}$$

can be reformulated as a function of ρ and W_0:

$$P_{avg} = \frac{\rho(1-\rho)}{\mu W_0 + (1-\rho)} \tag{60}$$

where W_0 is the average remaining service time for the customer found in service by a new arrival from a Poisson arrival process, namely,

$$W_0 = \sum_{i=1}^{n} \frac{\lambda_i \overline{x_i^2}}{2} \tag{61}$$

The term $\overline{x_i^2}$ denotes the second moment of the service time for the ith flow.

Proof. For each flow i, the average response time, T_i, can be decomposed into two components: the average waiting time, W_i, and the average service time, $\frac{1}{\mu}$, namely,

$$T_i = W_i + \frac{1}{\mu} \tag{62}$$

The subscript i specifies the ith flow, emphasizing that the waiting times may vary between flows. Regarding the average service time, we assume in this paper that it is the same for all flows and is given by $\frac{1}{\mu}$. We substitute Eq. (62) into Eq. (58):

$$\begin{aligned} P_{\text{avg}} &= \frac{\sum_{i=1}^{n} \rho_i}{\sum_{i=1}^{n} \left(\frac{\rho_i}{\rho} \mu T_i \right)} = \frac{\rho}{\frac{\mu}{\rho} \sum_{i=1}^{n} \rho_i T_i} = \frac{\rho}{\frac{\mu}{\rho} \sum_{i=1}^{n} \rho_i \left(W_i + \frac{1}{\mu} \right)} \\ &= \frac{\rho}{\frac{\mu}{\rho} \left(\sum_{i=1}^{n} \rho_i W_i \right) + \frac{\mu}{\rho} \left(\sum_{i=1}^{n} \rho_i \frac{1}{\mu} \right)} = \frac{\rho}{\frac{\mu}{\rho} \left(\sum_{i=1}^{n} \rho_i W_i \right) + 1} \end{aligned} \tag{63}$$

From the conservation law [47], we know that $\sum_{i=1}^{n} \rho_i W_i$ remains constant under an M/G/1 system utilizing any work-conserving queueing discipline, and it is expressed as:

$$\sum_{i=1}^{n} \rho_i W_i = \frac{\rho W_0}{1-\rho} \quad \text{for} \quad \rho < 1 \tag{64}$$

By substituting the constant value of $\sum_{i=1}^{n} \rho_i W_i$ from Eq. (64) into Eq. (63), we obtain:

$$P_{\text{avg}} = \frac{\rho}{\frac{\mu}{\rho} \left(\sum_{i=1}^{n} \rho_i W_i \right) + 1} = \frac{\rho}{\frac{\mu}{\rho} \left(\frac{\rho W_0}{1-\rho} \right) + 1} = \frac{\rho(1-\rho)}{\mu W_0 + (1-\rho)} \quad \blacksquare$$

This shows that the average power, P_{avg}, can be expressed simply as a function of ρ and W_0.

From Theorem 6.1, we know that the average power P_{avg} can be represented as a function of ρ and W_0. Eq. (61) shows that the term W_0 is influenced by the second moment of each flow's service time. If we further assume that all flows have the same second moment of service time, i.e., $\overline{x_i^2} = \overline{x^2}$ for $i = 1, \ldots, n$, we can establish the following theorem:

Theorem 6.2. *For an M/G/1 system with n flows using any work-conserving queueing discipline, if each flow has the same first and second moments of the service time, then the average power is equivalent to the power of a single flow system. Specifically, the average power can then be*

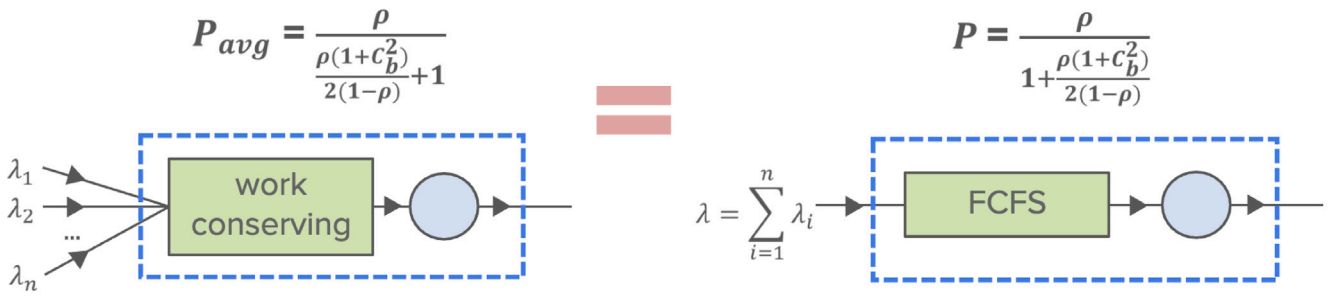

Fig. 10. The average power of an M/G/1 system with n flows using any work-conserving queueing discipline as shown on the left is equivalent to the power of a single-flow M/G/1 system where all n flows are aggregated into one and served FCFS as shown on the right.

expressed as follows:

$$P_{avg} = \frac{\rho}{1 + \frac{\rho(1+C_b^2)}{2(1-\rho)}} \tag{65}$$

Proof. If the service times for all flows have identical second moments, i.e., $\overline{x_i^2} = \overline{x^2}$ for $i = 1, \ldots, n$,[14] then the average remaining service time, W_0, can be expressed as follows:

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

Leading to

$$W_0 = \frac{\lambda \overline{x^2}}{2} \tag{66}$$

Next, we show that W_0 can be expressed as:

$$W_0 = \frac{\rho(1 + C_b^2)}{2\mu} \tag{67}$$

This expression is derived from the relationship between the second moment and the coefficient of variation squared, C_b^2, which quantifies the variability of service times relative to their mean ($\overline{x} = \frac{1}{\mu}$). It is calculated as:

$$C_b^2 = \frac{\overline{x^2} - \overline{x}^2}{\overline{x}^2}$$

By multiplying both sides by the denominator and rearranging the equation to isolate the second moment, we get:

$$\overline{x^2} = \overline{x}^2 + C_b^2 \cdot \overline{x}^2 = (1 + C_b^2)\overline{x}^2$$

Thus, the second moments $\overline{x^2}$ can be related to the service time coefficient of variation squared C_b^2 as:

$$\overline{x^2} = (1 + C_b^2)\frac{1}{\mu^2} \tag{68}$$

Substituting Eq. (68) into the equation for W_0 (Eq. (66)), we get:

$$W_0 = \frac{\lambda \overline{x^2}}{2} = \frac{\lambda}{2}(1 + C_b^2)\frac{1}{\mu^2} = \frac{\rho(1 + C_b^2)}{2\mu}$$

This is Eq. (67), which expresses W_0 in terms of C_b^2, μ, and ρ. We can then rewrite Eq. (60) for P_{avg} as:

$$P_{\text{avg}} = \frac{\rho(1-\rho)}{\mu W_0 + (1-\rho)} = \frac{\rho(1-\rho)}{\mu \cdot \frac{\rho(1+C_b^2)}{2\mu} + (1-\rho)} = \frac{\rho}{\frac{\rho(1+C_b^2)}{2(1-\rho)} + 1}$$ ■

[14] In an M/M/1 system with exponentially distributed service times, if all flows have the same mean service time (equal first moment), their second moments will also be equal. This is a direct consequence of the fact that the exponential distribution is fully characterized by its mean—once the mean is known, all other moments are determined. For an exponential distribution, the mean is $\frac{1}{\mu}$ and the second moment is $\frac{2}{\mu^2}$.

Thus, we have proven Eq. (65). In addition, this expression for P_{avg} is the same as the expression for the power of a single flow in an M/G/1 system, as given in Eq. (3) in Section 2.

Theorem 6.2 can be interpreted as shown in Fig. 10, which demonstrates that the average power for an M/G/1 system with multiple flows using any work-conserving queueing discipline, is equivalent to the power of a single-flow system where all n flows are combined into one flow based on FCFS. The power of a single flow follows the definition provided in Eq. (2) [5,7]. This equivalence means that the performance of a single-server system with multiple flows remains unaffected by the specific queueing discipline used, provided it is work-conserving and the first and second moments of the service time are the same for all flows. As long as the queueing discipline is work-conserving, even if the order in which flows are processed changes, the system's overall performance in terms of average power will remain unchanged. Consequently, the average power in a work-conserving system with multiple flows is equal to the power in a single-flow system.

6.2. Average power optimization

Now we turn our attention to the optimization of the average power for a work-conserving system of n flows, with each flow having the same first and second moments of service time. According to Theorem 6.2, the average power of a multiple-flow system can be equated to the power of a single-flow system. Therefore, optimizing the average power is equivalent to optimizing the power of a single-flow system. Hence, the only factor in affecting the optimization of the average power is the total amount of traffic entering the system.

To determine this optimal level of system utilization to maximize the power of a single-flow system, we refer to the findings in [5,7], as outlined in Section 2. According to this study, the optimal traffic load, ρ^*, for a single flow M/G/1 system that achieves the best performance in terms of power is given by Eq. (4), namely,

$$\rho^* = \frac{1}{1 + \sqrt{\frac{1+C_b^2}{2}}} \tag{69}$$

This identifies the ideal utilization factor for a single-flow system that balances system load with response time and is tailored to the specific variability of the input flows' service times. Combining Theorem 6.2 and the optimal result of a single flow represented by Eq. (69), we have established the following theorem:

Theorem 6.3. *An M/G/1 system with any work-conserving queueing discipline, where all flows have identical first and second moments of service time, achieves its optimal average power, P^*_{avg}, when:*

$$\rho^* = \sum_{i=1}^{n} \rho_i = \frac{1}{1 + \sqrt{\frac{1+C_b^2}{2}}} \tag{70}$$

The system's optimum average power is:

$$P^*_{\mathbf{avg}} = \frac{2}{(\sqrt{2}+\sqrt{1+C_b^2})^2} \tag{71}$$

Additionally, we have the following corollary for the specific M/M/1 system where $C_b^2 = 1$:

Corollary 6.1. *When the M/G/1 system of Theorem 6.3 is an **M/M/1** system[15] (where $C_b^2 = 1$), then optimal average power is achieved when:*

$$\rho^* = \sum_{i=1}^{n} \rho_i = \frac{1}{2} \tag{72}$$

The optimal average power is:

$$P^*_{\mathbf{avg}} = \frac{1}{4} \tag{73}$$

Eq. (72) and Eq. (73) are the same as the well-known single flow optimal power result for M/M/1 as in [5,7]. These results are valuable for system controllers and can guide the design of congestion control strategies, as they demonstrate that optimizing average power as the system performance hinges solely on the effective management of total system utilization.

7. Summary

This paper introduced and optimized three power-based performance metrics–individual power, sum of powers, and average power–for multi-flow systems, explicitly solving for the optimum values of the utilization factors and powers to address the competing goals of increasing throughput and reducing delay from different viewpoints. We analyzed and optimized these metrics within an M/M/1 system, considering two common queueing disciplines: First-Come, First-Served (FCFS) and Head-of-Line (HOL). Since queueing disciplines in multi-flow systems influence each flow's response time as well as performance in terms of power, we examined both the upper (HOL) and lower (FCFS) bounds of flow priority discrimination across all work-conserving queueing disciplines in the context of power optimization.

In Section 4, we introduced our first multi-flow performance metric, **individual power**, which focuses on the end-to-end perspective and is denoted by P_i for the ith flow. This metric, calculated as $P_i = \frac{\rho_i}{\mu T_i}$, was optimized singly and jointly for each flow. The convergent optimal operating points resulting from joint optimization under both FCFS and HOL are summarized in Table 1.

In Section 5, we introduced our second performance metric, called **sum of power**, which takes an overall system perspective and is denoted by $P_{\text{sum}} = \sum_{i=1}^{n} P_i$. We identified the operating points of flow utilizations that maximize this performance metric under both FCFS and HOL queueing disciplines. Surprisingly, the optimum sum of power for HOL is achieved under equal utilization factors for each flow. The optimization results for this metric are summarized in Table 2.

In Section 6, we proposed our third performance metric, called **average power**, which is also based on an overall system perspective and is denoted by $P_{\text{avg}} = \frac{\sum_{i=1}^{n} \rho_i}{\sum_{i=1}^{n} \left(\frac{\rho_i}{\rho} \mu T_i\right)}$. Applying the conservation law [47] to this metric, we discovered that optimizing it is equivalent to optimizing a single flow, as detailed in Theorem Theorem 6.2. This result is applicable not only to M/M/1 systems but also to the broader class of M/G/1 systems.

[15] Note that in an M/M/1 system, the service time is exponentially distributed. Therefore, if the first moment is the same across flows, we conclude that their second moments will also be the same.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Sunday Group, Inc.

Appendix A. Proof of Eq. (51)

To prove Eq. (51), we first demonstrate that each flow's utilization is equal when the sum of powers is maximized. Then, we find the optimal utilization for each flow and show Eq. (52) and Eq. (53).

Eq. (14) provides the individual power of flow i. Thus, the sum of the individual powers is: $P_{\text{sum}} = \sum_{i=1}^{n} \rho_i(1-\sigma_{i-1})(1-\sigma_i)$, where $\sigma_i = \sum_{j=1}^{i} \rho_j$. When aiming to optimize the sum of individual power for all flows in a system, we encounter a system of n equations. Each equation emerges from the partial differentiation of the sum of power with respect to each flow's utilization, ρ_i. We first show in step 1 below that each partial differentiation equation can be expressed as[16]:

$$\begin{aligned}\frac{\partial}{\partial \rho_i} P_{\text{sum}} &= \frac{\partial}{\partial \rho_i} \sum_{j=1}^{n} \rho_j(1-\sigma_{j-1})(1-\sigma_j) \\ &= (1-\sigma_n-\rho_i)(1-\sigma_n+\rho_i) \quad \text{for } i = 1,\dots,n\end{aligned} \tag{A.1}$$

Then in step 2, we use Eq. (A.1), set it to zero for each $i = 1,\dots,n$, and solve the n equations collectively (assuming the total utilization $\rho < 1$) to find the effective critical point, which is the optimized utilization operating point:

$$\rho_i^* = \frac{1}{n+1} \quad \text{for } i = 1,\dots,n$$

Step 1: Prove the Partial Differentiation Equation, Eq. (A.1)
We now prove the partial differentiation equation (Eq. (A.1)) by induction:

Base Case ($n = 2$):

From Eq. (48), the sum of individual powers when $n = 2$ is:

$$\begin{aligned}P_{\text{sum}} &= \sum_{i=1}^{2} \rho_i(1-\sigma_{i-1})(1-\sigma_i) \\ &= \rho_1(1-\rho_1) + \rho_2(1-\rho_1)(1-\rho_1-\rho_2)\end{aligned}$$

When optimizing the sum of powers with respect to ρ_1 and ρ_2 simultaneously, we have a system of two equations:

$$\begin{cases} \frac{\partial}{\partial \rho_1} P_{\text{sum}} = \frac{\partial}{\partial \rho_1}(\rho_1(1-\rho_1) + \rho_2(1-\rho_1)(1-\rho_1-\rho_2)) \\ \frac{\partial}{\partial \rho_2} P_{\text{sum}} = \frac{\partial}{\partial \rho_2}(\rho_1(1-\rho_1) + \rho_2(1-\rho_1)(1-\rho_1-\rho_2)) \end{cases}$$

The first equation:

$$\begin{aligned}\frac{\partial}{\partial \rho_1} P_{\text{sum}} &= \frac{\partial}{\partial \rho_1}(\rho_1(1-\rho_1) + \rho_2(1-\rho_1)(1-\rho_1-\rho_2)) \\ &= 1 - 2\rho_1 + \rho_2(-(1-\rho_1) - (1-\rho_1-\rho_2)) \\ &= 1 - 2\rho_1 - \rho_2(1-2\rho_1) - \rho_2(1-\rho_2) \\ &= (1-2\rho_1-\rho_2)(1-\rho_2) \\ &= (1-\rho_1-\rho_2-\rho_1)(1-\rho_1-\rho_2+\rho_1) \\ &= (1-\sigma_2-\rho_1)(1-\sigma_2+\rho_1)\end{aligned}$$

[16] We change the summation index in P_{sum} from i to j to avoid confusion with i in ρ_i.

The second equation:

$$\begin{aligned}\frac{\partial}{\partial\rho_2}P_{\text{sum}} &= \frac{\partial}{\partial\rho_2}(\rho_1(1-\rho_1)+\rho_2(1-\rho_1)(1-\rho_1-\rho_2))\\ &= (1-\rho_1)(1-\rho_1-2\rho_2)\\ &= (1-\rho_1-\rho_2+\rho_2)(1-\rho_1-\rho_2-\rho_2)\\ &= (1-\sigma_2-\rho_2)(1-\sigma_2+\rho_2)\end{aligned}$$

The two equations match Eq. (A.1) where $n=2$: $(1-\sigma_2-\rho_i)(1-\sigma_2+\rho_i)$ for $i=1,2$.

Induction Hypothesis:

Suppose the partial differentiation equation (Eq. (A.1)) works when the number of flows n is k:

$$\frac{\partial}{\partial\rho_i}P_{\text{sum}} = \frac{\partial}{\partial\rho_i}\sum_{j=1}^{k}\rho_j(1-\sigma_{j-1})(1-\sigma_j) = (1-\sigma_k-\rho_i)(1-\sigma_k+\rho_i) \quad \text{for } i=1,2,..,k$$

Induction Step:

We want to show that the equation also works for the number of flows $k+1$:

$$\frac{\partial}{\partial\rho_i}P_{\text{sum}} = \frac{\partial}{\partial\rho_i}\sum_{j=1}^{k+1}\rho_j(1-\sigma_{j-1})(1-\sigma_j) = (1-\sigma_{k+1}-\rho_i)(1-\sigma_{k+1}+\rho_i) \quad \text{for } i=1,2,..,k,k+1$$

Here's the computation:

$$\begin{aligned}&\frac{\partial}{\partial\rho_i}P_{\text{sum}} = \frac{\partial}{\partial\rho_i}\sum_{j=1}^{k+1}\rho_j(1-\sigma_{j-1})(1-\sigma_j)\\ &= \frac{\partial}{\partial\rho_i}\left[\sum_{j=1}^{k}\rho_j(1-\sigma_{j-1})(1-\sigma_j)+\rho_{k+1}(1-\sigma_k)(1-\sigma_{k+1})\right]\\ &= (1-\sigma_k-\rho_i)(1-\sigma_k+\rho_i)+\rho_{k+1}\frac{\partial}{\partial\rho_i}[(1-\sigma_k)(1-\sigma_{k+1})]\\ &= (1-\sigma_k-\rho_i)(1-\sigma_k+\rho_i)+\rho_{k+1}[-(1-\sigma_{k+1})-(1-\sigma_k)]\\ &= (1-\sigma_k-\rho_i)(1-\sigma_k+\rho_i)-\rho_{k+1}[(1-\sigma_{k+1})+(1-\sigma_k)-\rho_i+\rho_i]\\ &= (1-\sigma_k-\rho_i)(1-\sigma_k+\rho_i)-\rho_{k+1}(1-\sigma_k-\rho_i)-\rho_{k+1}(1-\sigma_{k+1}+\rho_i)\\ &= (1-\sigma_k-\rho_i)(1-\sigma_{k+1}+\rho_i)-\rho_{k+1}[1-\sigma_{k+1}+\rho_i]\\ &= (1-\sigma_k-\rho_i-\rho_{k+1})(1-\sigma_{k+1}+\rho_i)\\ &= (1-\sigma_{k+1}-\rho_i)(1-\sigma_{k+1}+\rho_i)\end{aligned}$$

This shows that the equation also works for the number of flows $k+1$.

Thus, by induction, we have shown that Eq. (A.1) holds for an arbitrary number of flows.

Step 2: Finding the Critical Point

In this step, we demonstrate that the critical point, where the n partial differential equations equals zero, is when each flow's optimum utilization $\rho_i^* = \frac{1}{n+1}$ and the total optimized utilization $\rho^* = \frac{n}{n+1}$.

From step 1, we know that Eq. (A.1) holds. Now, we set each partial differential equation to zero:

$$\frac{\partial}{\partial\rho_i}P_{\text{sum}} = (1-\sigma_n-\rho_i)(1-\sigma_n+\rho_i) = 0 \quad \text{for } i=1,2,\ldots,n$$

This implies that either:

$$(1-\sigma_n-\rho_i)=0 \quad \text{or} \quad (1-\sigma_n+\rho_i)=0 \quad \text{for } i=1,2,\ldots,n$$

We now discuss each case:

- $(1-\sigma_n+\rho_i)\stackrel{?}{=}0$:
  If $(1-\sigma_n+\rho_i)=0$, then $\sigma_n = 1+\rho_i$. This contradicts to the constraint that $\sigma_n = \sum_{j=1}^{n}\rho_j < 1$ because there must be at least one flow with $\rho_i > 0$. If all ρ_i values were 0, then σ_n would be zero, not 1 as indicated by the equation. Therefore, this scenario is not valid.
- $(1-\sigma_n-\rho_i)\stackrel{?}{=}0$:
  For $(1-\sigma_n-\rho_i)=0$, we have

$$\rho_i = 1-\sigma_n \quad \text{for } i=1,2,\ldots,n \tag{A.2}$$

Summing all equations, we get:

$$\sum_{i=1}^{n}\rho_i = \sum_{i=1}^{n}(1-\sigma_n),$$

From Eq. (A.2) and since $\sigma_n = \sum_{i=1}^{n}\rho_i$, the above equation can be expressed as:

$$\sigma_n = n(1-\sigma_n)$$

Thus, we compute σ_n as:

$$\sigma_n = \frac{n}{n+1} \tag{A.3}$$

Since $\sigma_n = \rho$, this shows that the optimum system utilization ρ^* when sum of individual power is maximized is:

$$\rho^* = \frac{n}{n+1} \tag{A.4}$$

as was to be shown. As $n < n+1$, this implies $\rho < 1$ for finite n, thereby confirming that this scenario is valid.

With σ_n computed, we now determine ρ_i using Eq. (A.2) and Eq. (A.3):

$$\rho_i = 1-\sigma_n = 1-\frac{n}{n+1} = \frac{1}{n+1} \quad \text{for } i=1,2,\ldots,n$$

Thus, we have also shown that:

$$\rho_i^* = \frac{1}{n+1} \quad \text{for } i=1,2,\ldots,n \quad \blacksquare$$

Data availability

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

References

[1] U. Cisco, Cisco annual internet report (2018–2023) white paper, Cisco: San Jose, CA, USA 10 (1) (2020) 1–35.
[2] I. Sandvine, Global internet phenomena report, North Am. Lat. Am. (2024).
[3] A. Giessler, J. Haenle, A. König, E. Pade, Free buffer allocation—An investigation by simulation, Comput. Networks (1976) 2 (3) (1978) 191–208.
[4] L. Kleinrock, On flow control in computer networks, in: Proceedings of the International Conference on Communications, Vol. 2, 1978, pp. 27–2.
[5] L. Kleinrock, Power and deterministic rules of thumb for probabilistic problems in computer communications, in: ICC'79; International Conference on Communications, Vol. 3, 1979, pp. 43–1.
[6] R. Gail, L. Kleinrock, An invariant property of computer network power, in: Proceedings of the International Conference on Communications, 1981, pp. 63.1.1–63.1.5.
[7] L. Kleinrock, Internet congestion control using the power metric: Keep the pipe just full, but no fuller, Ad Hoc Networks 80 (2018) 142–157.
[8] X. Xiao, L.M. Ni, Internet QoS: A big picture, IEEE Netw. 13 (2) (1999) 8–18.
[9] Z. Wang, Internet QoS: architectures and mechanisms for quality of service, Morgan Kaufmann, 2001.
[10] U. Black, Voice over IP, Prentice-Hall, Inc., 1999.
[11] B. Goode, Voice over internet protocol (VoIP), Proc. IEEE 90 (9) (2002) 1495–1517.
[12] K. Nichols, S. Blake, F. Baker, D.L. Black, Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers, Request for Comments RFC 2474, Internet Engineering Task Force, 1998, Available at https://www.rfc-editor.org/rfc/rfc2474.
[13] S. Blake, D. Black, M.A. Carlson, E. Davies, Z. Wang, W. Weiss, An Architecture for Differentiated Services, Request for Comments RFC 2475, Internet Engineering Task Force, 1998, Available at https://www.rfc-editor.org/rfc/rfc2475.
[14] R. Braden, D. Clark, S. Shenker, Integrated Services in the Internet Architecture: an Overview, Request for Comments RFC 1633, Internet Engineering Task Force, 1994, Available at https://www.rfc-editor.org/rfc/rfc1633.
[15] L. Kleinrock, Queueing Systems, Volume II: Computer Applications, Wiley New York, 1976.
[16] A. Cobham, Priority assignment in waiting line problems, J. Oper. Res. Soc. Am. 2 (1) (1954) 70–76.
[17] N. Jaiswal, Preemptive resume priority queue, Oper. Res. 9 (5) (1961) 732–742.
[18] V. Jacobson, Congestion avoidance and control, ACM SIGCOMM Comput. Commun. Rev. 18 (4) (1988) 314–329.

[19] V. Jacobson, Modified TCP congestion avoidance algorithm, 1990.
[20] M. Allman, V. Paxson, W. Stevens, RFC2581: TCP congestion control, 1999.
[21] L.S. Brakmo, S.W. O'Malley, L.L. Peterson, TCP vegas: New techniques for congestion detection and avoidance, in: Proceedings of the Conference on Communications Architectures, Protocols and Applications, 1994, pp. 24–35.
[22] S. Ha, I. Rhee, L. Xu, CUBIC: a new TCP-friendly high-speed TCP variant, ACM SIGOPS Oper. Syst. Rev. 42 (5) (2008) 64–74.
[23] 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.
[24] 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.
[25] N. Cardwell, Y. Cheng, C.S. Gunn, S.H. Yeganeh, V. Jacobson, BBR: congestion-based congestion control, ACM Queue 14 (5) (2016) 20–53.
[26] 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 19-23, 2019, ACM, 2019, pp. 44–58.
[27] G. Kumar, N. Dukkipati, K. Jang, H.M. 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.
[28] K. Ramakrishnan, S. Floyd, A Proposal to Add Explicit Congestion Notification (ECN) to IP, Request for Comments RFC 2481, Internet Engineering Task Force, 1999, Available at https://datatracker.ietf.org/doc/rfc2481/.
[29] D. Katabi, M. Handley, C. Rohrs, Congestion control for high bandwidth-delay product networks, in: Proceedings of the 2002 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, 2002, pp. 89–102.
[30] S. Floyd, V. Jacobson, Random early detection gateways for congestion avoidance, IEEE/ ACM Trans. Netw. 1 (4) (1993) 397–413.
[31] K.M. Nichols, V. Jacobson, A. McGregor, J.R. Iyengar, Controlled delay active queue management, RFC 8289 (2018) 1–25.
[32] G. Appenzeller, I. Keslassy, N. McKeown, Sizing router buffers, ACM SIGCOMM Comput. Commun. Rev. 34 (4) (2004) 281–292.
[33] L. Kleinrock, Queueing Systems, Volume I: Theory, Wiley New York, 1975.
[34] L. Kleinrock, Message delay in communication nets with storage (Ph.D. thesis), Massachusetts Institute of Technology, 1963.
[35] T. Stockhammer, Dynamic adaptive streaming over HTTP– standards and design principles, in: Proceedings of the Second Annual ACM Conference on Multimedia Systems, 2011, pp. 133–144.
[36] A. Zambelli, IIS smooth streaming technical overview, Microsoft Corp. 3 (40) (2009).
[37] Adobe, Adobe HTTP dynamic streaming (HDS), 2016.
[38] R. Pantos, W. May, Apple inc.,"http live streaming,", 2013.
[39] A. Bentaleb, B. Taani, A.C. Begen, C. Timmerer, R. Zimmermann, A survey on bitrate adaptation schemes for streaming media over HTTP, IEEE Commun. Surv. & Tutorials 21 (1) (2018) 562–585.
[40] J. Nash, Non-cooperative games, Ann. Math. (1951) 286–295.
[41] 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.
[42] R. Morris, TCP behavior with many flows, in: Proceedings 1997 International Conference on Network Protocols, IEEE, 1997, pp. 205–211.
[43] L. Qiu, Y. Zhang, S. Keshav, On individual and aggregate TCP performance, in: Proceedings. Seventh International Conference on Network Protocols, IEEE, 1999, pp. 203–212.
[44] L. Qiu, Y. Zhang, S. Keshav, Understanding the performance of many TCP flows, Comput. Netw. 37 (3–4) (2001) 277–306.
[45] K.K. Ramakrishnan, R. Jain, A binary feedback scheme for congestion avoidance in computer networks, ACM Trans. Comput. Syst. (TOCS) 8 (2) (1990) 158–181.
[46] D. Lin, R. Morris, Dynamics of random early detection, in: Proceedings of the ACM SIGCOMM'97 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, 1997, pp. 127–137.
[47] L. Kleinrock, A conservation law for a wide class of queueing disciplines, Nav. Res. Logist. Q. 12 (2) (1965) 181–192.