UNIVERSITY OF CALIFORNIA

Los Angeles

Computer Network Optimization Using the Power Metric

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Computer Science

by

Meng-Jung Tsai

2024

ABSTRACT OF THE DISSERTATION

Computer Network Optimization Using the Power Metric

by

Meng-Jung Tsai

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2024

Professor Leonard Kleinrock, Chair

Modern network research focuses on optimizing performance through congestion control, quality of service, and fairness. With the rapid expansion of networks and increasing traffic, balancing throughput and response time has become critical. This thesis explores this tradeoff and introduces the Power metric as a tool for optimizing network performance and expands its investigation to achieving optimized performance with optimum fairness.

The Power metric, defined as the ratio of normalized throughput to normalized mean response time, serves as our performance optimization goal. Previous research primarily focused on single-flow systems, but contemporary networks involve multiple flows with more complex scenarios. This work extends Power analysis in performance optimization to modern network environments, developing a model that also accommodates multiple flows. We further examine different queueing disciplines that implement various levels of flow discrimination. In addition, we examine fairness metrics coupled with performance optimization.

Our research focuses on three aspects: performance, flow priority discrimination, and fairness. We introduce performance metrics, including individual power, sum of power, and average power, and optimize these metrics using an M/M/1 system model with multiple flows under different queueing disciplines. We also explore fairness metrics such as throughput, delay, and power, and investigate scenarios where optimum performance and equal fairness

can be achieved simultaneously.

Additionally, we study generalized power, which allows specifying the relative preference for throughput versus delay, providing a flexible approach to optimizing network performance based on specific requirements.

In summary, this research represents a first step in incorporating performance, fairness, and priority flow discrimination into the Power metric analysis for modern multi-flow network environments. Our goal is to provide insights, guidance, and "rules of thumb" for system designers to create more efficient and equitable network systems.

The dissertation of Meng-Jung Tsai is approved.

George Varghese

Lixia Zhang

Omid Abari

Leonard Kleinrock, Committee Chair

University of California, Los Angeles

2024

Table of Contents

List of Figures

List of Tables

List of Theorems

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Leonard Kleinrock, for his invaluable mentorship and unwavering support throughout my PhD journey. Over the past six years, he has patiently guided me, offering insightful advice and thoughtful suggestions during our discussions. His dedication, hard work, and genuine passion for research have continuously encouraged and inspired me. His generosity with both his time and wisdom has been truly remarkable, and without his patient guidance, I would not have come this far. I am profoundly grateful for his support, and working with him has been an honor. I sincerely appreciate all the encouragement and help he has provided along the way.

I would like to express my heartfelt gratitude to the members of my doctoral committee—Professors George Varghese, Lixia Zhang, and Omid Abari—for taking the time to serve on my committee and for providing invaluable feedback. I am particularly grateful to Professors Varghese, Zhang, and Songwu Lu for serving on my committee for my candidate oral exam and for their inspiring courses, which greatly shaped my understanding and knowledge in the field of networks.

I would like to thank my labmates in the Connection Lab—Seunghyun, Kevin, Eli, Can, and Rohit—for the discussions and for making my time in the lab enjoyable. I am also grateful to all my friends who have supported me throughout this process, offering personal, emotional, and academic encouragement along the way.

Lastly, I would like to express my deepest gratitude to my parents and my family. Their unwavering support and constant encouragement have been the foundation of this journey. I am incredibly fortunate to have had their love, patience, and understanding through all the ups and downs.

Vita

Education

B.S. in Computer Science Sep 2014 – Jun 2018

National Chiao Tung University, Hsinchu, Taiwan

Ph.D. Student in Computer Science Sep 2018 – Sep 2024

University of California, Los Angeles, CA, USA

Internships

Cisco Meraki, San Francisco, CA Jun 2020 – Sep 2020

Software Engineer Intern at MX Routing Team

Google Cloud, Sunnyvale, CA Jun 2021 – Sep 2021

Software Engineer Intern at Borg Scheduling Team

Google Cloud, Sunnyvale, CA Jun 2022 – Sep 2022

Software Engineer Intern at Congestion Control Team

Google Cloud, Sunnyvale, CA Oct 2022 – Jun 2023

Student Researcher at Congestion Control Team

Chapter 1: Introduction

Modern network research has long focused on design decisions that optimize network performance, through consideration of congestion control, quality of service (QoS), differentiated services(DiffServ), and fairness. However, as networks expand and traffic explodes due to video streaming, cloud computing, and ubiquitous mobile devices [1, 2], factors such as effective congestion control, efficient quality of service, and balanced resource distribution have gained considerable importance. These factors are crucial for accommodating the ever-increasing demands without compromising network performance.

Given the multitude of factors influencing network performance, a central challenge emerges: **how to strike a balance between throughput and response time**. This balance is fundamental for delivering a seamless user experience while maintaining network efficiency. It's crucial not only within individual data flows (intra-flow) but also across multiple flows (inter-flow) competing for shared resources. The following sections study this throughput-response time tradeoff by introducing a metric called **Power** as a potential tool for optimizing network performance effectively. We then extend the study by addressing the issue of fairness among the competing network flows while simultaneously optimizing network performance. Subsequently, we discuss the limitations of previous research on Power, present our specific problem statement, and outline our focus on addressing these challenges.

1.1 Tradeoff between Throughput and Mean Response Time

Users interacting with networks often prioritize two key metrics: **throughput** and **mean response time**. 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**. Figure 1.1 visually represents this concept. The x-axis represents throughput (denoted by γ), which signifies the network's transmission rate, measured in packets (bytes) successfully delivered per second. The y-axis depicts mean response time (denoted by $T(\gamma)$), the average time taken for a packet to travel from source to destination.

As illustrated in Figure 1.1, response time typically exhibits a **positive correlation** with throughput. Generally, increasing traffic volume (higher throughput) leads to a corresponding rise in response time. Conversely, reducing traffic to achieve lower response times results in a decrease in throughput. Therefore, a critical aspect of network management lies in identifying an "optimal" operating point that effectively balances these two competing factors.

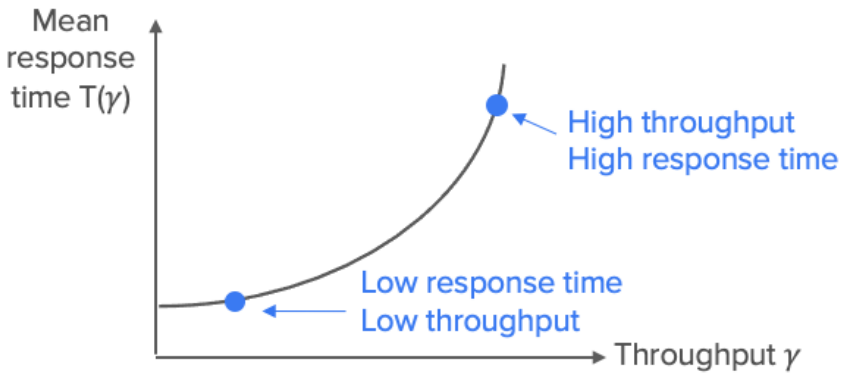


Figure 1.1: The tradeoff between throughput and mean response time

1.2 Power as the Optimization Metric

To quantitatively assess the intricate balance between throughput and mean response time, we utilize the **Power metric** (denoted as P). It was originally defined as the ratio of throughput to mean response time, $\frac{\gamma}{T(\gamma)}$. In this work, we use a slightly different definition for this metric, normalizing both the numerator and denominator, leading to:

$$P = \frac{\rho}{\mu T(\rho)} \tag{1.1}$$

Here, the numerator throughput γ is normalized to $\rho = \frac{\lambda}{\mu}$, where λ is the average input arrival rate, assuming $\lambda = \gamma$, and μ is the average system service rate[1]. The denominator, mean response time $T(\gamma)$, is normalized to $\mu T(\rho)$, based on the average service time per packet, $\frac{1}{\mu}$. We will explain the normalization in detail in Chapter 2. We adopt this form of the 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[2].

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. Note that the Power metric increases when throughput increases or when mean response time decreases, so our goal will be to maximize power.

---

[1] λ and μ are both measured in packets (or bytes) per second. λ represents the average number of packets (bytes) entering the system per second, and μ represents the average number of packets (bytes) the system can process per second. We assume a no-loss system, so the input rate equals the output rate, meaning $\lambda = \gamma$.

[2] In the remainder of this document, any mention of response time is explicitly intended to be interpreted as "mean response time."

Furthermore, the Power metric aligns with the intuitive principle of deterministic reasoning: *keep the pipe just full, but no fuller* [7], particularly in a stochastic system. 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.

1.3 Limitations of Previous Research on Power to Today's Networks

Previous research on power [3–7] has primarily focused on a *single* flow system, typically involving one hop or multiple hops. However, contemporary networks, as illustrated in Figure 1.2, presents a level of complexity that far exceeds these simplified scenarios. In this intricate network environment, several factors come into play.

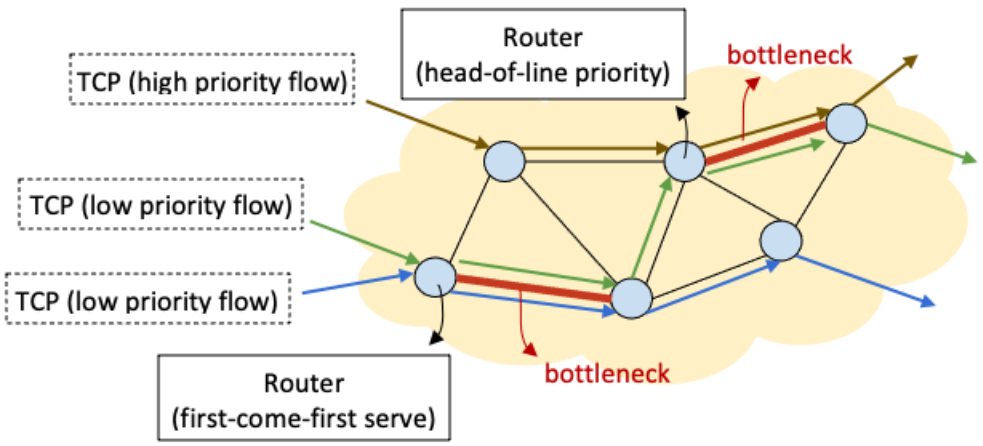


Figure 1.2: An example of current networks: multiple flows, multiple hops[3], different routes, and different queueing disciplines.

[3] We focus on one-hop analysis, paralleling the bottleneck, and do not consider the effect of multiple hops in this dissertation.

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 bandwidth, 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].

Furthermore, *ensuring* ***fairness*** *among flows while still optimizing performance is essential, especially when implementing differentiated services and priority queueing*, as it is essential to prevent starvation for lower priority flows. While prioritizing certain traffic, like voice calls, is essential, it shouldn't come at the expense of fairness for other flows. Without fairness, low-priority flows could experience starvation or excessive delays, leading to a degraded user experience. However, ensuring fairness becomes increasingly complex in practice as today's diverse network environments see a rise in applications with varying service requirements.

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], reacts to congestion encountered for a given flow along its path, aiming to achieve a balance between maximizing its own throughput and ensuring fairness, 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 and fair 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 and balance fairness among them.

Given these complexities, navigating the network landscape to achieve the optimal balance between throughput and response time using the power metric requires a thorough analysis. This analysis must delve into the intricacies introduced by multiple flows and various queueing disciplines, accounting for both end-to-end and router perspectives. A comprehensive understanding is essential for effectively addressing the diverse challenges posed by today's intricate networks.

1.4 Research Goal

As we see, modern networks present a complex landscape characterized by **multiple flows**, **differentiated services** (implemented through varied queueing disciplines), and the critical need for **fairness**. While the **power** metric effectively balances throughput and response time to identify optimal operating points, previous analyses using this metric are limited by models that focus on single-flow systems. These simplified models fall short of addressing the complexities of contemporary networks, which involve multiple flows and diverse requirements.

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 three core aspects:

1. **Performance**: Characterized by different transformed versions of power for multiple flows.

2. **Flow priority discrimination**: Represented by different functions of throughput to mean response time for different queueing disciplines.

3. **Fairness**: Essential for ensuring equitable resource allocation among competing flows and preventing starvation or excessive delays.

This research represents a first step towards a holistic understanding of the interrelationships between these aspects, how they affect each other, and how to optimize and balance them.

We aim to simplify the mathematical modeling of complex networks and explore their impact on the power metric. By evaluating various forms of power optimization criteria based on differing performance needs and applying these optimizations to different queueing disciplines, while addressing the issue of fairness, we generate valuable insights that can assist in designing effective and implementable congestion control algorithms. Our ultimate goal is to adapt power analysis to modern networks, **integrating considerations of performance, flow discrimination, and fairness**, thus offering practical guidance to system designers for creating more efficient and equitable network systems.

1.5 Problem Statements

The Power metric, while effective in single-flow scenarios, faces limitations when applied to modern, multi-flow network environments. These limitations, arising from the complexities of multi-flow systems, lead to the following key questions:

- **How can we define the Power metric in a way that explicitly accounts for throughput and delay in multi-flow systems?** The current definition of the Power metric is general and doesn't explicitly define how throughput and delay are calculated in a multi-flow system. This ambiguity makes it difficult to apply the Power metric effectively in multi-flow scenarios.

- **How can the model incorporate mechanisms for differentiated services to prioritize certain data flows?** This will ensure that critical data flows receive the necessary resources and achieve their performance targets. This is related to various priority queueing disciplines, leading to questions like: *What are the optimal operating points under different queueing disciplines? How do these points change with the queueing disciplines?*

- **How can we evaluate fairness in resource allocation across multiple flows?** What happens if we use Power, as well as throughput and delay, as a fairness metric?

- **Can we achieve both optimal performance and optimum fairness simultaneously in a multi-flow system?** If so, under what performance metrics and what fairness metrics can we achieve this, and at what operating point? This question delves into the fundamental trade-off between performance and fairness, and explores potential scenarios where both can be optimized.

- **How can the model adapt to diverse throughput and delay requirements for different flows?** The model should be able to adapt to these diverse requirements, providing tailored solutions that meet the specific needs of individual flows.

1.6 Summary of Results

To address the questions outlined in the problem statements, we define multiple forms of power as performance metrics for a multiple-flow system with n flows:

1. individual power of the i^{th} flow $\boldsymbol{P_i}$
2. sum of (individual) powers $\boldsymbol{P}_{\text{sum}}$
3. average power $\boldsymbol{P}_{\text{avg}}$

We then optimize these performance power metrics within an M/M/1 system with multiple flows under various queueing disciplines, spanning the full spectrum of flow discrimination from minimum to maximum. Furthermore, we discuss fairness metrics, including throughput, delay, and power. Additionally, we combine performance and fairness considerations along with the degree of flow discrimination to determine when optimum performance and optimum fairness[4] can be achieved simultaneously. Finally, we adopt the generalized power concept and extend it to our performance metrics.

We now present a brief outline of this dissertation and summarize key findings. In the first chapter, we discuss issues related to congestion control in computer networks, narrowing down to the throughput-delay tradeoff. We then choose to use the **Power** metric to address this competing relationship and discuss the limitations of previous work on the Power metric in contemporary networks, with a particular focus on multiple-flow systems. Following this, we present our overall goal, which is to develop a mathematical optimization model incorporating **performance**, **flow priority discrimination**, and **fairness** to provide high-level guidance for system designers.

[4] By "optimum fairness", we mean "equal fairness" throughout this dissertation. When we refer to "fairness," it implies "optimum fairness" throughout this dissertation.

In chapter 2, we describe the single-server queueing theory model used to model computer networks and the notations associated with that model. We then introduce the Power metric in detail, describe its normalization and present preliminary results about it in the context of maximizing the power metric in the M/M/1 and M/G/1 queueing systems[5].

In Chapter 3, we extend the model from a single-server, single-flow system to a single-server, *multiple-flow* system. We introduce priority queueing disciplines, with First-Come, First-Serve (FCFS) as the minimum flow discrimination and Head-of-the-Line (HOL) as the maximum flow discrimination. We then present two families of queueing disciplines that span the full spectrum of flow discrimination between FCFS and HOL.

In Chapter 4, we introduce our first multi-flow performance metric called **individual power**, which focuses on an end-to-end perspective and is denoted by P_i for the i^{th} flow. This metric utilizes the individual flow parameters ρ_i and μT_i for the numerator and denominator, respectively, expressed as $P_i = \frac{\rho_i}{\mu T_i}$. We then optimize this performance metric for multiple flows under both First-Come First-Served (FCFS) and Head-of-Line (HOL) queueing disciplines, identifying the equivalent optimal operating points when each flow simultaneously optimizes its individual power. These optimization results are summarized in Table 4.1. We further investigate the iterative process involved in achieving these equivalent operating points. Additionally, we explore alternative approaches for normalization and proceed with optimization under these new formulations. The results of this investigation are presented in Table 4.3.

In chapter 5, we introduce 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 identify the

[5] Note that the M/M/1 model is used throughout this dissertation, except in Chapter 6 and Section 2.2, where the M/G/1 model is applied.

operating points 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. Additionally, we investigate properties of this metric, such as symmetry. The optimization results for this metric are summarized in Table 5.1. Comparisons of using individual power and sum of power as optimization goals are presented in Figures 5.5 and 5.6.

In chapter 6, we propose our third performance metric, called **average power**, which is also based on a holistic system aspect and is denoted by $P_{\text{avg}} = \frac{\sum_{i=1}^{n} \rho_i}{\sum_{i=1}^{n} \left(\frac{\rho_i}{\rho} \mu T_i\right)}$. We apply the conservation law [33] to this metric and discover that optimizing it is equivalent to optimizing a single flow, as detailed in Theorem 6.2. This result is applicable not only to M/M/1 systems but also to the broader range of M/G/1 systems.

In Chapter 7, we extend the optimization analysis of performance metrics to a wider spectrum of queueing disciplines (i.e., those between FCFS and HOL). We utilize the two families of queueing disciplines with flow discrimination ranging from minimum to maximum, introduced in Chapter 3. We proceed with individual power optimization and sum of power optimization for these queueing disciplines. We present both the analytical and numerical results for $n = 2$ and for arbitrary n. Finally, we identify the queueing parameters β in the beta-priority system and k in the delay-dependent system[6] that maximize the sum of power for given values of ρ_1 and ρ_2. We found that the optimized sum of power for both systems increases monotonically with given values of ρ_1 and ρ_2.

In Chapter 8, we discuss common fairness metrics, namely, throughput and delay, and propose using "(individual) power" as an additional fairness metric. We then thoroughly

[6] Choice of β and k parameters allow us to range from FCFS to HOL.

examine the concept of equal power fairness (which we often refer to as "power fairness") in an M/M/1 system with two flows using the beta-priority system and the delay-dependent system. We define the feasible regions where power fairness is achievable and identify the queueing parameters β in the beta-priority system and k in the delay-dependent system that achieve power fairness for given values of ρ_1 and ρ_2 in a two-flow system.

In Chapter 9, we integrate the three performance metrics and three fairness metrics with various queueing disciplines, represented by flow discrimination, into a three-dimensional framework. We examine the 9 combinations of performance metrics and fairness metrics, derived from our 3 performance metrics (individual power, sum of power, average power) and our 3 fairness metrics (throughput, delay, individual power). We analyze these combinations across different queueing disciplines and under varying numbers of flows.

In Chapter 10, we study the concept of **generalized power**, introduced by Kleinrock, which allows for specifying the relative preference for throughput versus delay. We extend this definition to our individual power metric and optimize individual generalized power under the FCFS and HOL queueing disciplines to identify the optimal operating points.

In Chapter 11, we outline potential future work to extend this dissertation.

Chapter 2: Background

This chapter dives into the crucial concept of the Power metric within the context of queueing theory, specifically focusing on single-server queueing systems as models for computer networks. The Power metric is valuable tool for understanding and optimizing system performance by analyzing the balance between system resource utilization and mean queueing delays. We begin by establishing the network model, elucidating key parameters such as arrival rate, service rate, efficiency, loss rate, throughput, and mean response time.

Next, we introduce the definition of the power metric and review previous research related to it. Of particular note is Kleinrock's seminal work from 1979 [5], which identified the optimal power levels for both $M/M/1$ and $M/G/1$ systems. This research yielded valuable insights into congestion control, summarizing the optimal strategy as *"keep the pipe just full, but not fuller."*

2.1 The Single-Server Queueing System

A computer network system can be modeled as a global single server queueing system where packets arrive at a rate λ (packets/second), undergo processing within the system, and depart as in Figure 2.1. Though an idealized and simplified representation, this model remains a valuable tool for understanding and analyzing various aspects of network performance, particularly throughput and delay. We use the following notation to describe the key parameters of this system:

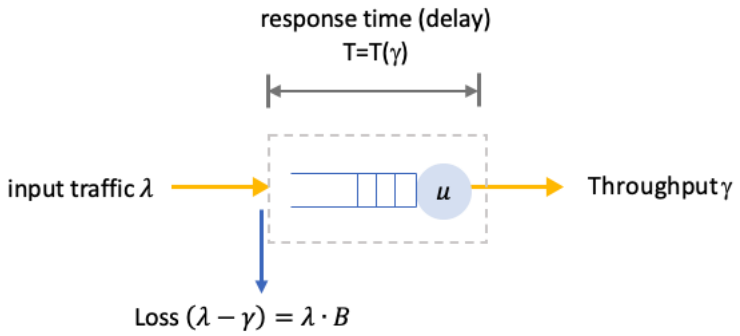


Figure 2.1: Model of a computer network system as a single-server queueing system.

- $\bar{t}$: 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 $\lambda = \frac{1}{\bar{t}}$.
- $\bar{x}$: Mean service time of a packet, measured in seconds.
- μ: Average service rate of the system, indicating the 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 $0 \leq \rho < 1$ for system stability.
- B: Loss rate or blocking probability, which denotes the likelihood that an incoming packet will be denied entry into the system or dropped.
- γ: Throughput, representing the volume of traffic that successfully traverses the system. It is calculated as $\gamma = \lambda(1 - B)$ and measured in packets/second.
- T=T(γ): 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. Specifically, $T = W + \overline{x}$, where W denotes the average waiting time.

This metric typically varies as a function of throughput, hence we use the notation $T(\gamma)$[1].

In the following analysis, we focus on a lossless system, where B=0, implying that input traffic λ is equivalent to throughput γ, and the mean response time $T=T(\gamma)=T(\lambda)$.

The distribution of service time and arrival time may be deterministic or stochastic. Different distributions are denoted as follows [34]: **D** represents a deterministic distribution, **M** represents an exponential distribution, **G** represents a general distribution. We use the notation $\boldsymbol{A/B/K}$ to represent the system characteristics, where $\boldsymbol{A}$ is instantiated by D, M, or G to represent the distribution of packet inter-arrival times, $\tilde{\lambda}$ [2]; $\boldsymbol{B}$ is also instantiated by D, M, or G to represent the distribution of packet service time, $\tilde{x}$; and $\boldsymbol{K}$ stands for the number of servers in the system.

This thesis focuses on single-server queueing systems as models for computer networks. To be more specific, we focus on two common systems: $\boldsymbol{M/M/1}$ and $\boldsymbol{M/G/1}$. In this context, the K parameter in the notation $A/B/K$ is set to 1, representing a single server. A is instantiated with M, since we assume a Poisson arrival process, meaning that inter-arrival times are independently and exponentially distributed. B is instantiated with either M or G, representing either the exponential distribution or a general distribution for service times.

---

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

[2] Note that the tilde symbol ˜ in $\tilde{t}$ indicates a random variable, while placing a bar over it as $\bar{t}$ denotes its average value (mean).

2.2 The Power Metric

2.2.1 Definition

Recall that Power is a metric that combines two competing performance measures, throughput and mean response time (delay), into a single metric. It was first introduced by Giessler in [3] as the ratio of throughput to mean response time, $\frac{\gamma}{T}$. 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.

Kleinrock, in [5], proposed an alternative definition of power that normalizes both throughput and mean response time. This normalized power is expressed as:

$$P = \frac{\rho}{\mu T} \tag{2.1}$$

Here, the throughput is transformed into the utilization factor (efficiency) using the equation $\rho = \frac{\gamma}{\mu} = \frac{\lambda(1-B)}{\mu}$. Since we consider a lossless system (B = 0), ρ can be simplified to $\frac{\lambda}{\mu}$. 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 results in a power metric that is dimensionless. We will use this normalized power definition throughout this dissertation.

It's worth noting that Little's Result [35], expressed as $\overline{N} = \lambda T$ also combines the utilization factor (efficiency) and the normalized mean response time, but through multiplication:

$$\overline{N} = \lambda T = \frac{\lambda}{\mu}(\mu T) = \rho(\mu T) \tag{2.2}$$

Here, $\overline{N}$ represents the average number of packets in the system, which is another commonly

used metric in queueing theory analysis.

2.2.2 The Maximal Power Operating Point

Recall the discussion in Section 1.1 regarding the tradeoff between throughput and mean response time. Our objective now is to identify the operating point for power that addresses this tradeoff. In essence, we aim to optimize power by increasing system utilization while keeping mean response time low. This involves finding the operating point that yields the maximum power value, often 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.

To find the maximal power operating point, we differentiate power with respect to efficiency and set the derivative equal to zero:

$$\frac{dP}{d\rho} = \frac{d\frac{\rho}{\mu T}}{d\rho} = \frac{\mu T - \rho\frac{d\mu T}{d\rho}}{(\mu T)^2} = 0$$

This implies that the numerator, $\mu T - \rho\frac{d\mu T}{d\rho}$, must be zero. Therefore, the condition for maximizing power is:

$$\frac{\mu T}{\rho} = \frac{d\mu T}{d\rho} \tag{2.3}$$

Equation 2.3 indicates that at the maximal power point, which we henceforth denote as ρ^*, the derivative of the curve $\mu T(\rho)$ is equal to the slope of the straight line connecting the origin to $(\rho^*, \mu T(\rho^*))$. This relationship is illustrated in Figure 2.2, where the derivative at

ρ^* is equal to $\frac{\mu T(\rho^*)}{\rho^*}$.

Another key observation from this figure is that, at any point $(a, \mu T(a))$ on the curve, the power value is inversely proportional to the slope of the line from the origin to that point as shown in blue. This relationship holds even when the response time function is non-differentiable, discontinuous, or non-convex. Consequently, *finding the maximal power point is equivalent to finding the line of minimum slope originating from the origin that intersects the curve.* When $T(\rho)$ is continuous and convex, then this line represents the tangent to the curve $\mu T(\rho)$. Hence, *in the continuous convex case, the maximal power point can be determined by identifying the tangent to the curve that passes through the origin.*

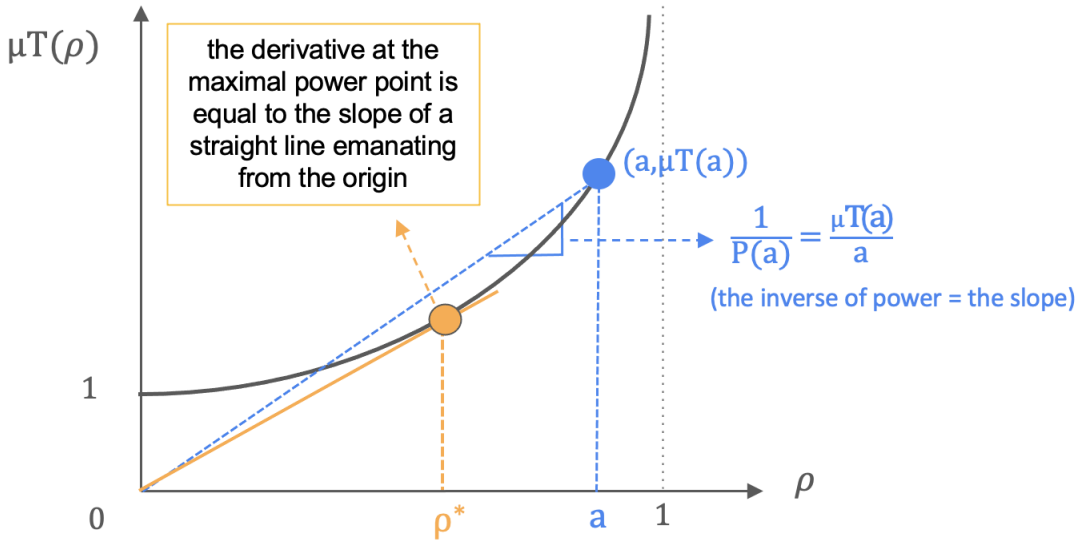


Figure 2.2: Relationship between slope and power. The slope of a straight line out of the origin to any point on the curve is the inverse of power at that point.

2.2.3 Maximal Power in M/M/1 queueing systems

According to [36], in the M/M/1 queueing system, we know that

$$\mu T(\rho) = \frac{1}{1-\rho} \tag{2.4}$$

and therefore, the power is

$$P = \frac{\rho}{\mu T} = \rho(1-\rho) \tag{2.5}$$

In [5], Kleinrock derived the optimal power point

$$\rho^* = 0.5 \tag{2.6}$$

Substituting Equation 2.6 into Equation 2.5, the corresponding maximal power value is

$$P^* = \rho^*(1-\rho^*) = 0.25 \tag{2.7}$$

Additionally, Kleinrock calculated the average number at optimality as

$$\overline{N^*} = \rho^* \frac{1}{1-\rho^*} = 1 \tag{2.8}$$

This implies the deterministic reasoning: *"keep the pipe just full, but no fuller"* [7].

2.2.4 Maximal Power in M/G/1 queueing systems

In [5], Kleinrock extended the same optimization to a more general system, the M/G/1 system, where

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

Here, C_b^2 represents the squared coefficient of variation for the service time, which is the ratio of variance ($\overline{x^2} - \overline{x}^2$) to the squared mean service time ($\overline{x}^2$):

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

He demonstrated that in the M/G/1 system, the optimal operating point is achieved when

$$\rho^* = \frac{1}{1 + \sqrt{\frac{1+C_b^2}{2}}} \tag{2.9}$$

and then $\overline{N^*}$, the average number in system at this optimal point is still:

$$\overline{N^*} = 1 \tag{2.10}$$

2.3 Limitations

This chapter introduced the single-server queueing model and explored preliminary research results concerning the power metric. While this model offers simplicity and leads to good intuition by accommodating only one flow, it has inherent limitations. Notably, it fails to fully capture the complexity of real-world networks, particularly the competition among multiple flows for network bandwidth.

The single-server queueing model treats multiple flows as a composite flow, leading to an optimization process that yields an optimal solution for the aggregate flow. However, this approach does not address how individual flows should share the available bandwidth. This limitation highlights the need for a more comprehensive queueing system with multiple flows to accurately determine the optimal operating point for each flow and identify the optimal

bandwidth allocation strategy.

Moreover, the single-flow system does not account for scenarios with different queueing disciplines. It assumes that packets from multiple flows are aggregated into a composite queue based solely on the first-come-first-serve policy. However, different queueing disciplines result in varying packet orderings from different flows when merging into a queue, consequently impacting the response time of each flow differently. Given the need to prioritize traffic flows based on their properties, various queueing disciplines are valuable in real-world networks. Therefore, it is crucial to incorporate the effects of diverse queueing disciplines into our model to achieve a more realistic representation of network behavior.

Chapter 3: Model

In the previous chapter, we discussed the limitations of the single-server queueing model in capturing the complexity of real-world network scenarios, particularly in handling multiple flows and various queueing disciplines. To address these limitations, this chapter introduces a queueing system with multiple flows, focusing on different queueing disciplines to handle various flows. This system serves as a model for the complex network outlined in Chapter 1 and depicted in Figure 1.2. Several simplifications are made below to render the model solvable while retaining sufficient complexity to effectively capture the impact of multiple flows and different queueing disciplines.

3.1 Multiple Flows System

The multiple flows queueing system we consider is an M/M/1 system illustrated in Figure 3.1. This will be used throughout this dissertation[1]. There are n flows entering the system, with each flow having a packet arrival rate of λ_i packets per second from a Poisson process. Each flow is assumed to have the same packet service rate, μ packets per second from an exponential distribution, with the average service time per packet being $\frac{1}{\mu}$ seconds. The utilization of each flow is thus

$$\rho_i = \frac{\lambda_i}{\mu} \tag{3.1}$$

[1] This dissertation 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 don't explicitly say it, we assume an M/M/1 system.

and the total system utilization is

$$\rho = \sum_{i=1}^{n} \rho_i \tag{3.2}$$

These n Poisson processes can be viewed as a combined Poisson process with total average arrival rate of $\lambda = \sum_{i=1}^{n} \lambda_i$. The total system utilization ρ can also be computed as

$$\rho = \sum_{i=1}^{n} \rho_i = \sum_{i=1}^{n} \frac{\lambda_i}{\mu} = \frac{\lambda}{\mu} \tag{3.3}$$

When combined into a single flow, the system can be seen as a single flow. However, we explicitly differentiate each flow here to observe the impact of multiple flows, particularly when different priorities are applied to each.

For the queueing discipline used to handle the order of packets from various flows being queued and entering service, this dissertation focuses on a family of **work-conserving queueing disciplines**, represented by the yellow box in Figure 3.1. 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 Section 5.2 of [37], this family of work-conserving queueing disciplines adheres to the following principles:

- No defections: Work does not leave the system before completion.
- No additional workload generation: No new work is created within the system.
- Preemption: Preemption is allowed only for exponentially distributed service times in a preemptive-resume setting.
- No server idleness: The server never idles when work is available.

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 boundaries of flow priority discrimination within this family of queueing

disciplines. Other disciplines of this family falling between these extremes will be introduced in the following section.

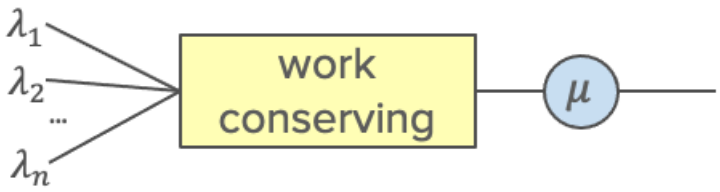


Figure 3.1: Model for a single hop M/M/1 system with multiple flows using work-conserving queueing disciplines.

3.2 Assumptions and Simplification

The model in Figure 3.1 resembles a single flow system but focuses on different queueing disciplines to handle tbe various flows. Compared to real network systems, several simplifications are made below to 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 Figure 1.2 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 typically occurs at a bottleneck, where most of the 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 accurately.

- **Assume M/M/1**

  As stated above, each flow is assumed to arrive from a Poisson process, and the required service time of each packet is exponentially distributed, and the average service time for

each flow is identical. We opt for the M/M/1 model here as it simplifies the computation regarding mean response time. This choice facilitates easier analysis, allowing us to uncover potentially hidden insights. The M/M/1 model is the default unless otherwise explicitly stated.

- **First 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 (HOL) [15–17]. This choice is primarily driven by their common use in queueing disciplines and their simplicity for calculation, particularly due to their uncomplicated form in terms of response time. Furthermore, these two disciplines represent the lower and upper bounds on discrimination priority among all work-conserving queueing disciplines that do not depend on the job size.

  In the FCFS system, each flow has the same mean response time:

$$T_i = T = \frac{1}{\mu(1-\rho)} \quad \text{for all} \quad i = 1, \ldots, n \tag{3.4}$$

  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 ρ.

  In the HOL system, a packet from group i has full preemptive priority over all packets from groups $i+1, i+2, ..., n$. This means that if a packet from group i arrives while a packet from any of the lower-priority groups is being served, the service of the lower-priority packet will be interrupted, and the group i packet will be served immediately. Essentially, group 1 has the highest priority, followed by group 2, and so on, with group

n having the lowest priority. The response time for each priority group is given by [16]:

$$T_i = \frac{1}{\mu(1-\sigma_{i-1})(1-\sigma_i)}, \quad \text{where } \sigma_i = \sum_{j=1}^{i} \rho_j \tag{3.5}$$

The 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 all work-conserving queueing policies [15]. Note that HOL is the most discriminatory and results in the maximum difference in each flow's response time, while FCFS is the least discriminatory and results in each flow having the same response time.

3.3 Spectrum of Queueing Disciplines From FCFS to HOL

Examining the two extreme cases allows us to effectively bound the impact of different queueing disciplines on power optimization. However, as we progress further in this research, we need to incorporate more comprehensive queueing policies. There are several approaches to transition from FCFS to HOL, and one approach is to consider the delay-dependent priority discipline [38]. Additionally, we will introduce the beta-priority system to illustrate other alternatices to study the spectrum of flow discrimination from HOL to FCFS. The following sections introduce these queueing disciplines that allow us to study the full spectrum from HOL to FCFS.

3.3.1 The Delay-Dependent System

The delay-dependent system was introduced by Kleinrock [38] and uses a set of variable parameters, b_i to provide the flexibility in adjusting the relative waiting time among different groups.

Each priority group is assigned a number b_i, where $0 \leq b_n \leq b_{n-1} \leq \ldots \leq b_2 \leq b_1$. $q_i(t)$, the priority of a packet from group i, is a function of time that linearly increases with the time it stays in the system, using the scalar b_i, namely $q_i(t) = (t - \delta)b_i$, where δ is the time when the packet enters the system to wait for service and $t \geq \delta$. Therefore, a larger value for b_i represents a higher growth rate for the i^{th} priority group. This mechanism also ensures that packets waiting for a long enough time can be served, preventing starvation (such as in the strict head-of-line priority system). (Note that the priority order used here is reversed compared to that used by Kleinrock in [38]. Kleinrock used a higher index to represent a higher priority group, while we use a lower index to indicate higher priority.)

In [38], Kleinrock derived the triangular set of equations for the waiting time of each group i. For the preemptive case, with the modification of the priority order so that a lower index corresponds to a higher priority, the mean waiting time is as follows:

$$W_i = \frac{\frac{W_0}{1-\rho} + \sum_{j=1}^{i-1} \frac{\rho_j}{\mu}[1 - \frac{b_i}{b_j}] - \sum_{j=i+1}^{n} \frac{\rho_j}{\mu}[1 - \frac{b_j}{b_i}] - \sum_{j=i+1}^{n} \rho_j W_j [1 - \frac{b_j}{b_i}]}{1 - \sum_{j=1}^{i-1} \rho_j [1 - \frac{b_i}{b_j}]}$$

where

$$W_0 = \sum_{i=1}^{n} \frac{\lambda_i \overline{x_i^2}}{2} = \sum_{i=1}^{n} \frac{\lambda_i \frac{2}{\mu^2}}{2} = \frac{\frac{\lambda}{\mu}}{\mu} = \frac{\rho}{\mu} \tag{3.6}$$

W_0 is the average residual service time and is equal to $\frac{\rho}{\mu}$ in our M/M/1 model, with the assumption that each flow has the same mean service time. For an exponential distribution of service time, the mean service time is $\frac{1}{\mu}$, the variance is $\frac{1}{\mu^2}$ and the second moment of service time is $\frac{2}{\mu^2}$ (the second moment $E[X^2] = Var(x) + (E[x])^2$). Since we assume each flow has the same mean service time, the second moment $\overline{x_i^2} = \frac{2}{\mu^2}$ for $i = 1, .., n$.

We note that the mean response time,

$$T_i = W_i + \frac{1}{\mu} \tag{3.7}$$

We take the case of two flows as an example to compute the response time:
For T_2:

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

For T_1:

$$T_1 = \frac{1}{\mu} + \frac{W_0}{1-\rho} - \frac{\rho_2}{\mu}(1 - \frac{b_2}{b_1}) - \rho_2 W_2(1 - \frac{b_2}{b_1}) = \frac{1 - \rho(1 - \frac{b_2}{b_1})}{\mu(1-\rho)[1 - \rho_1(1 - \frac{b_2}{b_1})]}$$

We define

$$k = (1 - \frac{b_2}{b_1}) \tag{3.8}$$

Then the mean response time can be written as:

$$T_1 = \frac{1 - k\rho}{\mu(1-\rho)(1 - k\rho_1)}, \quad T_2 = \frac{1}{\mu(1-\rho)(1 - k\rho_1)} \tag{3.9}$$

When $b_1 = b_2$, flows 1 and 2 gain priority at the same rate, and the one that arrives earlier receives service first, which is equivalent to the first-come, first-served (FCFS) system. In this case, the parameter k becomes 0, and the response times reduce to those in the FCFS case:

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

When $b_1 \gg b_2$, flow 1's priority is effectively infinite compared to flow 2 when flow 1 enters the system. This is equivalent to the head-of-line (HOL) case, as the priority of a newly arrived packet from flow 1 will always be higher than that of packets in the system from flow 2, making flow 1's priority strictly larger than flow 2's. With $\frac{b_2}{b_1} \to 0$, the parameter k becomes 1, making the response time the same as in the HOL case:

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

For $n > 2$, the scenario where $b_1 = b_2 = \cdots = b_n$ corresponds to the FCFS case, as each group has the same increasing rate. This means the priority is based on the time spent waiting in the system, resulting in a first-come, first-served order. Conversely, the scenario where $b_1 \gg b_2 \gg \cdots \gg b_n$ corresponds to the HOL case. In this scenario, flow 1 has the highest priority whenever it enters the system because its priority increase rate is very large compared to the others. This is followed by flow 2, and so on. This arrangement means that flow 1 always has the highest priority, flow 2 has the next highest priority, and this pattern continues for the subsequent flows.

3.3.2 The Beta-Priority System

In addition to the delay-dependent system, we consider a second approach, the beta-priority system, which can also describe the full range from FCFS to HOL. The idea behind this

system is to make the response time an average of the FCFS and HOL systems, weighted by the parameter β:

$$T = \beta * T_{HOL} + (1 - \beta) * T_{FCFS} \tag{3.10}$$

β of time the response time follows the head-of-line behavior and $1 - \beta$ of time follows the first-come-first-serve pattern. By substituting the mean response times of FCFS and HOL with two flows into Equation 3.10, we derive the mean response times for flow 1 and flow 2 as follows:

$$\begin{aligned} T_1 &= \frac{\beta}{\mu(1-\rho_1)} + \frac{1-\beta}{\mu(1-\rho)} \\ T_2 &= \frac{\beta}{\mu(1-\rho_1)(1-\rho)} + \frac{1-\beta}{\mu(1-\rho)} \end{aligned} \tag{3.11}$$

When $\beta = 0$, the system behaves like the FCFS system; when $\beta = 1$, it behaves like the HOL system. For $0 < \beta < 1$, the system behavior in terms of flow discrimination ranges between these two extremes. Unlike the delay-dependent system, which uses n variables, $b_1, b_2, \ldots, b_n$, to cover the full spectrum of flow discrimination from FCFS to HOL, the beta-priority system utilizes a single variable, β, to span this range. This simplification reduces complexity in the analysis by using just one parameter to model the transition between FCFS and HOL.

There may be several possible implementations of the beta-priority system, and Figure 3.2 illustrates one example. For each input traffic flow, a portion β is directed to the HOL queue, while the remaining $(1 - \beta)$ goes to the FCFS queue. In the HOL queue, traffic is sorted, with higher-priority flows at the front and lower-priority flows at the back. The red traffic represents externally introduced traffic, which establishes the proper response time in the beta-priority system. This red traffic is later routed away from our model.

For the FCFS queue, the externally introduced traffic is $\beta\lambda$. Combined with the $(1 - \beta)\lambda$ traffic from the input flow goes to that queue, the total flow to the FCFS queue is λ, resulting

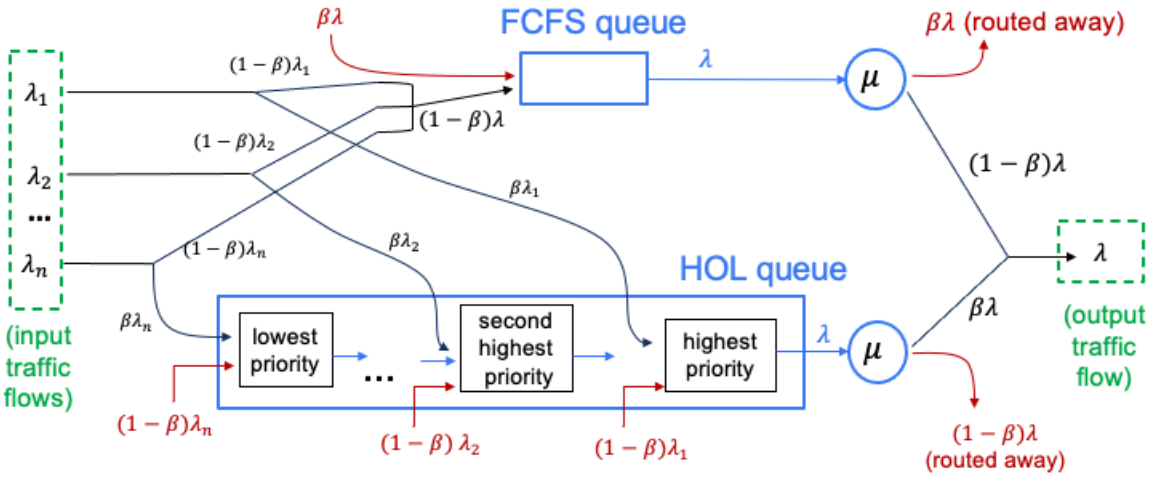


Figure 3.2: An example of implementing the beta-priority system. Each input flow i splits its traffic between the FCFS and HOL queues in portions $(1-\beta)$ and β. The red traffic represents externally introduced traffic and is used to achieve the average response time as Equation 3.10.

in a service time given by Equation 3.4, $T_{FCFS} = \frac{1}{\mu(1-\rho)}$. In the HOL queue, each priority group i receives external traffic in the amount of $(1-\beta)\lambda_i$ for $i = 1, \ldots, n$, resulting in the response time for each flow i following Equation 3.5, along with each $\beta\lambda_i$ input traffic.

With the external traffic, even though the input is split, each portion experiences the response time as if the full traffic amount λ were processed by either the FCFS or HOL queues. The β portion directed to the HOL queue experiences the response time give by Equation 3.5, while the $(1-\beta)$ portion directed to the FCFS queue follows the response time described by Equation 3.4. As a result, the beta-priority system is implemented, with the average response time of this input flow given by Equation 3.10.

3.3.3 Summary: Response Times for Two Queueing Systems

Table 3.1 displays the response times of two flows across various preemptive queueing systems, from FCFS (First-Come, First-Served) to HOL (Head-of-Line). Meanwhile, Table 3.2 details the conditions under which each system transitions to FCFS or HOL.

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

Table 3.1: Response Times for Two Flows Across Various Queueing Disciplines

Discipline	**First-Come-First-Served**	**Head-of-line Priority**
Delay-Dependent System	$k = 0$ $(b_1 = b_2)$	$k = 1$ $(b_1 \gg b_2)$
Beta-Priority System	$\beta = 0$	$\beta = 1$

Table 3.2: Parameters That Make the System Equivalent to FCFS or HOL Scheduling

Chapter 4: Performance Optimization Metric: Individual Power, P_i

In these next three chapters (Chapter 4, 5, 6), we introduce three distinct performance metrics based on different forms of power. Here in Chapter 4, we focus on the first performance power metric: **Individual Power $\boldsymbol{P_i}$**. In Chapter 5, we introduce the second performance power metric: **Sum of Powers $\boldsymbol{P}_{\text{sum}}$**. In Chapter 6, we discuss the third performance power metric: **Average Power $\boldsymbol{P}_{\text{avg}}$**.

In this chapter, we utilize the multiple-flow single-hop model introduced in the previous chapter to investigate the impact of multiple flows and different queueing disciplines from an **end-to-end** perspective. We define **individual power** to account for the limited view that each flow has of the system. Subsequently, we optimize this power for each flow individually for both FCFS and HOL systems. Finally, we compare and discuss the optimization results to gain insights into the effectiveness of different queueing disciplines in optimizing individual power.

4.1 Individual Power

4.1.1 Description of 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 [39–43]. 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.

4.1.2 Definition

Here is a formal definition of "individual power" for flow i:

$$P_i = \frac{\rho_i}{\mu T_i(\rho_i)} \tag{4.1}$$

In this equation, ρ_i represents the utilization factor of the i^{th} 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 first-come-first-served (FCFS). The denominator in Equation 4.1 shows that the mean response time $T_i(\rho_i)$ is normalized by its no-load response time, $\frac{1}{\mu}$, which represents the average service time of a packet.

4.1.3 Example

Let's consider an M/M/1 system with n flows where FCFS is the queueing discipline. T_i is consistent across all i flows and depends on the total system utilization, given by, $\frac{1}{\mu(1-\rho)}$,

as stated in Equation 3.4. For the i^{th} flow, the individual power is given by Equation 4.1 as

$$P_i = \frac{\rho_i}{\mu T_i}$$

and so

$$P_i = \rho_i(1 - \rho) \tag{4.2}$$

To emphasize the individual impact of utilization on response time, we separate the utilization of the i^{th} flow, ρ_i, from the total system utilization. We use α to represent the sum of the utilizations of all other flows, defined by the formula:

$$\alpha = \sum_{j=1, j\neq i}^{n} \rho_j \tag{4.3}$$

The individual power for flow i is then:

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

4.1.4 Individual Power Optimization

With the definition of individual power established, we can proceed to optimize individual power. Our objective is to determine the value of ρ_i that maximizes the metric $P_i = \rho_i(1 - \alpha - \rho_i)$, assuming α is fixed. This involves calculating the derivative of P_i with respect to ρ_i and setting it to zero for each i:

$$\frac{dP_i}{d\rho_i} = 0 \quad \text{for } i = 1, \ldots, n$$

Using the example of FCFS as stated above, we have:

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

Solving this equation for ρ_i yields our optimized ρ_i^* as:

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

and the optimal individual power is:

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

Note that P_i^* is independent of ρ_i.

The results derived above can be summarized in the following theorem:

Theorem 4.1.

In an M/M/1 system employing a FCFS queueing discipline, the optimal ρ_i^* for maximizing P_i is half of the remaining utilization:

$$\rho_i^* = \frac{1-\alpha}{2} = \frac{1-\sum_{j=1,j\neq i}^{n}\rho_j}{2} \tag{4.4}$$

with α indicating the portion of utilization occupied by other flows $\alpha = \sum_{j=1,j\neq i}^{n}\rho_j$.

The corresponding maximal individual power value is:

$$P_i^* = \left(\frac{1-\alpha}{2}\right)^2 \tag{4.5}$$

Notably, the well-known result from Equation 2.6 (that the optimal value $\rho^* = 0.5$ for a single flow) aligns with this theorem. That is, it is the case when $\alpha = 0$, as there are no other flows in the system, allowing the entire available utilization for that single flow to be 1, which leads to the optimal value of ρ being ρ^* =0.5.

4.1.5 Limitation

The optimization of individual power yields the optimal value for ρ_i^* as half of the remaining utilization left over by the other flows, with the corresponding individual power value calculated as $\left(\frac{1-\alpha}{2}\right)^2$. This value of individual power varies with α. It is at its maximum when $\alpha = 0$, and decreases as α increases. This decrease arises because α is assumed to be fixed during the optimization process.

This assumption reflects the limitations inherent in the end-to-end perspective, where each flow lacks information about other flows at the bottleneck and only see the remaining utilization, underscoring a fundamental constraint in decentralized systems. Despite the use of FCFS, where each flow experiences the same average response time, the absence of information about other flows can still result in a different optimal operating point compared to a single flow system.

Let's illustrate this with the same example of a system using FCFS. If there is only one flow, that flow has complete information about the relationship between utilization and response time, as shown in Figure 4.1 below. Here, the response time depends solely on the utilization of that single flow. As discussed earlier in chapter 2, in such a scenario within the M/M/1 system, the power optimization yields a maximal power at $\rho^* = 0.5$, where the optimal power $P^* = \rho(1 - \rho) = 0.25$.

However, in a system with multiple flows, each individual flow observes a **truncated** response time, depicted by the blue curve and blue axes as in Figure 4.2. The presence of other flows occupying a certain portion of the system utilization introduces constraints on the achievable utilization for each individual flow, represented by α. Consequently, the achievable utilization for the i^{th} flow is bounded by the utilization that is left, which is $1 - \alpha$. Yet, the i^{th} flow is unaware of this constraint as it lacks information about other flows. It only knows the traffic it contributes to the system, denoted by the blue axis as ρ_i, and observes the mean response time (in blue) based on the total system utilization ρ, which includes the contributions of other flows.

Effectively, the i^{th} flow operates as if it were on a truncated response time curve, represented by the blue curve in Figure 4.2. When it attempts to optimize its power, it identifies the tangent that passes through the (0,0) point on this perceived blue curve, resulting in ρ_i^* being half of the remaining utilization, $\frac{1-\alpha}{2}$. Thus, the optimal operating point ρ_i^* varies depending on the utilization taken up by other flows, denoted as α.

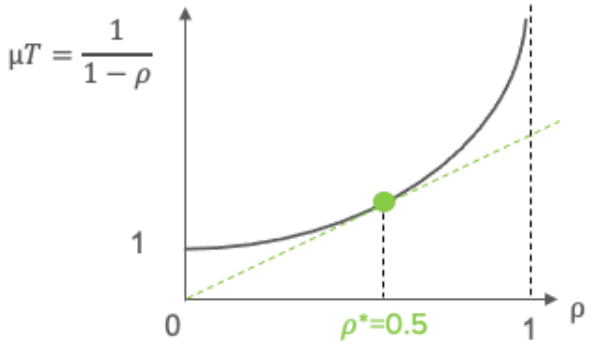


Figure 4.1: Single flow in an M/M/1 system with the whole view of the response time curve.

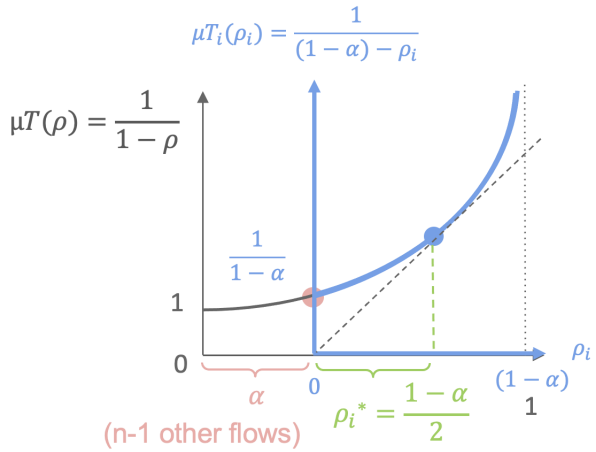


Figure 4.2: Blue curve is the view of the i^{th} flow in an M/M/1 system with n flows. The bound for ρ_i is $[0, 1-\alpha)$ given that α is taken by other flows.

4.2 Multiple flows optimize their individual power

Equation 4.4 presents the optimization result for the i^{th} flow under the assumption that α is fixed. However, if other flows are also optimizing their individual power, then α becomes dynamic (iterative), which in turn alters the optimal ρ_i. This leads to the question: What are the equilibrium optimal values of ρ_i when all flows in the system are optimizing simultaneously? Furthermore, how might different queueing disciplines impact these equilibrium optimal values of ρ_i^*?

To address these questions, we explore scenarios where each flow in the system optimizes its individual power. We will first discuss how to find equilibrium optimal solutions and then apply these concepts to practical systems. Specifically, we will examine two queueing disciplines: FCFS and HOL. Our analysis in the following section will begin with determining the equilibrium optimal operating point for each flow. We will then calculate optimum system utilization, optimal individual power, and the sum of optimized individual powers[1]. Lastly, we will also consider the limiting case where the number of flows approaches infinity. In Section 4.3, we will investigate the optimization process to understand if the sequence in which they perform their optimizations impacts the outcomes.

4.2.1 Finding Equilibrium Optimal Operating Points ρ_i^*

When n flows in a system are all optimizing their individual power, we now derive n equations. Each equation represents the partial derivative of each flow's power with respect to its respective utilization:

$$\frac{\partial P_i}{\partial \rho_i} = 0 \quad \text{for} \quad i = 1, \ldots, n \tag{4.6}$$

[1] In Chapter 5, we consider optimizing the sum of individual powers as opposed to simply showing the sum of the optimized individual powers in this chapter

If a solution exists for this system of n equations, then solving them will yield the equilibrium optimal set of ρ_i^*[2].

4.2.2 FCFS

If there are n flows in an M/M/1 system with FCFS, the individual power for each flow is given by:

$$P_i = \frac{\rho_i}{\mu T_i(\rho_i)} = \frac{\rho_i}{\mu T(\rho)} = \rho_i(1-\rho) \quad \text{for} \quad i = 1, \ldots, n \tag{4.7}$$

where $T_i(\rho_i)$ is equal to $T(\rho) = \frac{1}{\mu(1-\rho)}$ from Equation 3.4, indicating that each flow experiences the same mean response time.

To optimize individual power, we partially differentiate P_i with respect to ρ_i and set the derivative equal to zero for each i:

$$\frac{\partial P_i}{\partial \rho_i} = \frac{\partial(\rho_i(1-\rho))}{\partial \rho_i} = 1 - \rho - \rho_i = 0 \quad \text{for} \quad i = 1, \ldots, n$$

By rearranging the terms of this equation, we have:

$$\rho_i = 1 - \rho \quad \text{for} \quad i = 1, \ldots, n \tag{4.8}$$

This is equivalent to:

$$\rho_i = \frac{1}{2}(1 - \sum_{j=1, j\neq i}^{n} \rho_j) \quad \text{for} \quad i = 1, \ldots, n \tag{4.9}$$

If all n flows simultaneously optimize their individual power, we need to solve this system of n equations. Solving these n equations (Equation 4.9) is equivalent to solving the n equations

[2] Note that this set of ρ_i^* represents a Nash equilibrium [44].

in Equation 4.8. To solve these n equations (Equation 4.8), we sum all of them:

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

Rearranging this gives the *optimum total load*:

$$\rho^* = \frac{n}{n+1} \tag{4.10}$$

Consequently, we can calculate the optimized individual utilization factor for flow i:

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

Thus, we obtain:

$$\rho_i^* = \frac{1}{n+1} \tag{4.11}$$

Note that since $n < (n+1)$, $\rho^* = \frac{n}{n+1}$ is strictly less than 1, indicating that the system remains stable for any finite number of flows. However, in the limit as $n \to \infty$, the system becomes unstable as ρ^* approaches 1:

$$\lim_{n \to \infty} \rho^* = \lim_{n \to \infty} \frac{n}{n+1} = 1 \tag{4.12}$$

Taking Equation 4.11 back into Equation 4.7, we compute the optimal individual power:

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

Next, let us sum all optimized individual powers:

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

These results are summarized in the following theorem:

Theorem 4.2.

In an M/M/1 system with n flows using FCFS, when each flow optimizes its individual power $P_i = \rho_i(1-\rho)$*, the equilibrium optimal operating point* $(\rho_1, \rho_2, \ldots, \rho_n)$ *is*

$$\rho_i^* = \frac{1}{n+1} \quad \textit{for} \quad i = 1, \ldots, n. \tag{4.13}$$

The corresponding optimal individual power for each flow is

$$P_i^* = \frac{1}{(n+1)^2} \tag{4.14}$$

and the sum of optimal individual power is

$$P_{sum} = \sum_{i=1}^{n} P_i^* = \frac{n}{(n+1)^2} \tag{4.15}$$

Note that we don't use the superscript $*$ for P_{sum} as the sum of power may not be maximal[3] even though each individual power is optimal. This optimal individual power represents an equilibrium balance for each flow. The absolute optimal individual power for flow i alone is achieved when $\rho_i^* = 0.5$ and $\rho_j = 0$ for all $j = 1, \ldots, n$ and $j \neq i$, requiring the other flows to have zero utilization. However, we assume that the utilization factors of other flows cannot be altered during the optimization of flow i, resulting in outcomes different from $\rho_i^* = 0.5$. Nevertheless, any change in a flow's own utilization factor may prompt other flows to adjust theirs in response, until an equilibrium is reached.

[3] We will explore the optimal sum of power value in the next chapter.

As the number of flows n approaches infinity, the asymptotic behavior of the optimization results is summarized in the following corollary:

Corollary 4.2.1.

Consider an M/M/1 system with n flows under an FCFS queueing. 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.*

$$\lim_{n\to\infty} \rho^* = \lim_{n\to\infty} \frac{n}{n+1} = 1 \tag{4.16}$$

- *The optimized individual power P_i^* for each flow approaches 0.*

$$\lim_{n\to\infty} P_i^* = \lim_{n\to\infty} \frac{1}{n+1} = 0 \tag{4.17}$$

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

$$\lim_{n\to\infty} P_{sum} = \lim_{n\to\infty} \sum_{i=1}^{n} P_i^* = \lim_{n\to\infty} \frac{n}{(n+1)^2} = 0 \tag{4.18}$$

From the limiting behavior from Equations 4.16, 4.17, and 4.18, we observe that as the number of flows increases significantly and each flow optimizes its individual power simultaneously, the optimized individual power $P_i^* = \frac{1}{n+1}$ diminishes to zero, and the sum of the optimized powers $P_{\text{sum}} = \sum_{i=1}^{n} P_i^* = \frac{n}{(n+1)^2}$ also approaches zero. This indicates that each flow experiences a significant response time, causing its individual power to approach zero. Consequently, the summation of individual powers also trends toward zero. This scenario reflects reduced benefits for both individual flows and the system as a whole as the number of flows grows.

In addition, this finding from 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 [45–47]. For instance, Morris in [45] 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 theoretically analysis offers valuable insights into the potential challenges posed by a large number of TCP flows, aligning with empirical observations in practical scenarios.

4.2.3 HOL

Now, let's consider the scenario of maximal flow discrimination, employing the head-of-line (HOL) priority queueing discipline. Specifically, we examine the preemptive-resume case, which is work-conserving, where packets from higher-priority groups can preempt those of lower priority currently being served and the lower priority packet will resume from where it was preempted [15, 16]. We assume there are n flows, with the priority order among each flow as described in Chapter 3. Flow 1 has the highest priority, flow 2 has the second highest priority, and so on, with flow n having the lowest priority.

In this system, the response time for flow i is given by Equation 3.5 as:

$$T_i = \frac{1}{\mu(1-\sigma_i)(1-\sigma_{i-1})}, \quad \text{where } \sigma_i = \sum_{k=1}^{i} \rho_k$$

The individual power is expressed as:

$$P_i = \frac{\rho_i}{\mu T_i} = \rho_i(1-\sigma_i)(1-\sigma_{i-1}) = \rho_i(1-\sigma_{i-1}-\rho_i)(1-\sigma_{i-1}) \tag{4.19}$$

To find ρ_i^* that maximizes P_i, we take the partial derivative of P_i with respect to ρ_i and set it to zero:

$$\frac{\partial P_i}{\partial \rho_i} = \frac{\partial \rho_i(1-\sigma_{i_1}-\rho_i)(1-\sigma_{i-1})}{\partial \rho_i} = 0 \quad \text{for } i = 1, 2, \ldots, n$$

Solving this gives:

$$\rho_i^* = \frac{1-\sigma_{i-1}}{2} = \frac{1-\sum_{j=1}^{i-1}\rho_j}{2} \quad \text{for } i = 1, 2, \ldots, n \tag{4.20}$$

Equation 4.20 follows a logic similar to FCFS, where each flow optimizes its power by taking half of the available utilization. The difference lies in the effect of HOL queueing, where flows with higher priority are not influenced by flows with lower priority. Therefore, when calculating the remaining utilization for a flow, it is only necessary to consider the utilization taken by flows with higher priority. The remaining utilization is then determined as one minus the summation of utilization from these higher priority flows, which are the flows with lower index in our setting. For instance, for the k^{th} flow, the remaining utilization is $1 - \sum_{j=1}^{k-1} \rho_j$.

From Equation 4.20, we can derive the optimal values for each flow:

$$\rho_1^* = \frac{1-0}{2} = \frac{1}{2}$$

$$\rho_2^* = \frac{1-\rho_1^*}{2} = \frac{1-\frac{1}{2}}{2} = \frac{1}{4}$$

$$\rho_3^* = \frac{1-\rho_1^*-\rho_2^*}{2} = \frac{1-\frac{1}{2}-\frac{1}{4}}{2} = \frac{1}{8}$$

$$\rho_4^* = \frac{1-\rho_1^*-\rho_2^*-\rho_3^*}{2} = \frac{1-\frac{1}{2}-\frac{1}{4}-\frac{1}{8}}{2} = \frac{1}{16}$$

and so forth for additional flows. This sequence demonstrates how each subsequent flow's optimal utilization is calculated by halving the remaining available utilization after accounting for all prior flows. We observe that ρ_i^* follows a geometric sequence with each term being half of the preceding one, manifesting the pattern: $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and so forth. We now prove that each flow's optimal utilization ρ_i^* conforms to the formula:

$$\rho_i^* = \left(\frac{1}{2}\right)^i \quad \text{for } i = 1, 2, \ldots, n \tag{4.21}$$

Equation 4.20 tell us:

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

Rearranging, we have:

$$(1 - \sigma_{i-1}) = 2\rho_i^* \tag{4.22}$$

Equation 4.20 is also applicable for index $i + 1$:

$$\rho_{i+1}^* = \frac{1 - \sigma_i}{2} \tag{4.23}$$

Since $\sigma_i = \sigma_{i-1} + \rho_i^*$, we substitute this and Equation 4.22 into Equation 4.23:

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

From this, we derive the relationship between ρ_{i+1}^* and ρ_i^*:

$$\rho_{i+1}^* = \frac{\rho_i^*}{2} \tag{4.24}$$

Applying Equation 4.24 recursively, we obtain:

$$\rho^*_{i+2} = \frac{\rho^*_{i+1}}{2} = \frac{\rho^*_i}{2} \cdot \frac{1}{2} = \frac{\rho^*_i}{4}$$

Applying Equation 4.24 recursively, we obtain:

$$\rho^*_{i+k} = (\frac{1}{2})^k \rho^*_i \tag{4.25}$$

Since $\sigma_0 = 0$, we establish the base case:

$$\rho^*_1 = \frac{1 - \sigma_0}{2} = \frac{1}{2}$$

This enables us to use Equation 4.25 and calculate:

$$\rho^*_{1+k} \ = (\frac{1}{2})^k \rho^*_1 \ = (\frac{1}{2})^{1+k}$$

Simplifying further, we find:

$$\rho^*_i = \left(\frac{1}{2}\right)^i \quad \text{for } i = 1, 2, \ldots, n \tag{4.26}$$

This is also Equation 4.21, which we aim to prove.

Total Utilization

Now let's calculate the total utilization after each flow reaches equilibrium ρ^*_i by summing 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:

$$\rho^* = 1 - \left(\frac{1}{2}\right)^n \tag{4.27}$$

The limiting behavior as n approaches infinity is thus:

$$\lim_{n\to\infty} \rho^* = \lim_{n\to\infty} 1 - (\frac{1}{2})^n = 1 \tag{4.28}$$

Once again, it is important to note from Equation 4.27 that the total utilization ρ^* is strictly less than 1 for any finite number of flows. This ensures that the system remains stable under those conditions. The total utilization ρ^* approaches unity only as the number of flows n approaches infinity.

Optimized Individual Power

To determine the optimized individual power, we perform the following calculation:

$$\begin{aligned}
P_i^* &= \rho_i^* \cdot (1 - \sigma_i) \cdot (1 - \sigma_{i-1}) \\
&= \rho_i^* \cdot (1 - \sum_{j=1}^{i} \rho_j^*) \cdot (1 - \sum_{j=1}^{i-1} \rho_j^*) \\
&= (\frac{1}{2})^i \cdot (1 - \sum_{j=1}^{i} (\frac{1}{2})^j) \cdot (1 - \sum_{j=1}^{i-1} (\frac{1}{2})^j) \\
&= (\frac{1}{2})^i \cdot [1 - 1 - (\frac{1}{2})^i] \cdot [1 - (1 - (\frac{1}{2})^{i-1})] \\
&= (\frac{1}{2})^i \cdot (\frac{1}{2})^i \cdot (\frac{1}{2})^{i-1}
\end{aligned}$$

Simplifying this expression yields:

$$P_i^* = 2 \cdot (\frac{1}{8})^i \quad \text{for } i = 1, 2, \ldots, n \tag{4.29}$$

From Equation 4.21 and 4.29, 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 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:

Theorem 4.3.

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

The optimal load for each flow i is given by:

$$\rho_i^* = (\frac{1}{2})^i \quad for\ i = 1, 2, \ldots, n \tag{4.30}$$

and the optimized individual power for flow i is:

$$P_i^* = 2 \cdot (\frac{1}{8})^i \quad for\ i = 1, 2, \ldots, n \tag{4.31}$$

The total optimum utilization factor is:

$$\rho^* = 1 - \left(\frac{1}{2}\right)^n \tag{4.32}$$

and the limiting behavior this is:

$$\lim_{n\to\infty} \rho^* = \lim_{n\to\infty} 1 - (\frac{1}{2})^n = 1 \tag{4.33}$$

Sum of Optimized Individual Powers

The sum of the optimized individual powers in this case is computed as follows:

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

This expression simplifies to:

$$P_{\text{sum}} = \sum_{i=1}^{n} P_i^* = \frac{2}{7} \cdot (1 - (\frac{1}{8})^n) \tag{4.34}$$

When n goes to infinity:

$$\lim_{n\to\infty} P_{\text{sum}} = \lim_{n\to\infty} \sum_{i=1}^{n} P_i^* = \frac{2}{7} \tag{4.35}$$

This leads to the following theorem:

Theorem 4.4.

Using the preemptive HOL queueing policy in an M/M/1 system with n flows, where each flow optimizes its individual power, at equilibrium, the sum of optimal individual powers increases as n increases. The sum of optimized individual powers is given by:

$$P_{sum} = \sum_{i=1}^{n} P_i^* = \frac{2}{7} \cdot (1 - (\frac{1}{8})^n) \tag{4.36}$$

The limit of this value as n goes infinity is:

$$\lim_{n\to\infty} P_{sum} = \lim_{n\to\infty} \sum_{i=1}^{n} P_i^* = \frac{2}{7} \tag{4.37}$$

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 until the system utilization reaches a critical point where new flows experience very high waiting times, resulting in almost zero individual power for those 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 $\frac{2}{7}$.

4.2.4 Comparison of FCFS and HOL

Table 4.1 summarizes the equilibrium results for FCFS and HOL derived in the previous sections. It presents the equilibrium optimal ρ_i^* for each flow, total system utilization ρ^* (the sum of ρ_i^*), and the corresponding optimized individual powers along with their sum. In addition, the limiting behavior of ρ^* and the sum of P_i^* are also included.

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

Table 4.1: Individual Power Optimization at Equilibrium for FCFS and HOL.

Figure 4.3 shows the optimal system utilization $\rho^* = \sum_{i=1}^{n} \rho_i^*$ for different values of n for FCFS and HOL. These values are based on the second row in Table 4.1. As discussed in the previous section, we observe that as the number of flows tends towards infinity, the system utilization approaches unity for both the FCFS and HOL systems. For other work-conserving queueing disciplines, the curve of the optimized system utilization with the number of flows n at equilibrium is conjectured to lie between the curves of FCFS and HOL, with HOL as the upper bound and FCFS as the lower bound[4].

Additionally, we can observe from Figure 4.3 that the optimum system utilization approaches 1 in HOL faster than in FCFS. In HOL, it reaches about 0.995 at approximately $n = 8$, whereas the optimum system utilization in FCFS is about 0.888 at the same n. The optimum system utilization for FCFS, $\frac{n}{n+1}$, can be expressed as $1 - \frac{1}{n+1}$. This mathematical expression helps explain why HOL blows up faster than FCFS: $\left(\frac{1}{2}\right)^n$ approaches 0 exponentially, which is much faster than $\frac{1}{n+1}$, which approaches 0 hyperbolically. That is, the optimum system utilization in HOL, $1 - \left(\frac{1}{2}\right)^n$, approaches 1 faster than in FCFS, $\frac{n}{n+1}$.

To assess overall performance, we use the summation of optimized individual powers as the metric to compare the optimization results. The values for FCFS and HOL, indicated in the fifth row of Table 4.1, are used to plot Figure 4.4. This 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 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

[4] We conjecture that optimized rho and optimized power discussed later are also bounded by HOL and FCFS based on our numerical results in Chapter 7.

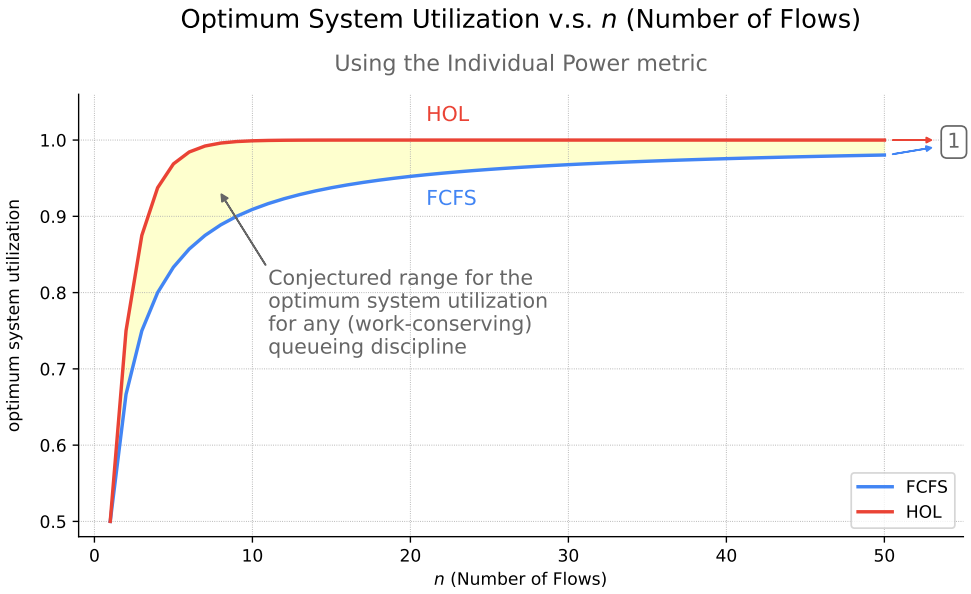


Figure 4.3: Trend of optimum system utilization ρ^* as the number of flows increases. The HOL and the FCFS are conjectured to be the upper and lower bound. The yellow region between FCFS and HOL is conjectured to be the range of possible optimum system utilization for any work-conserving priority discipline (see footnote 4).

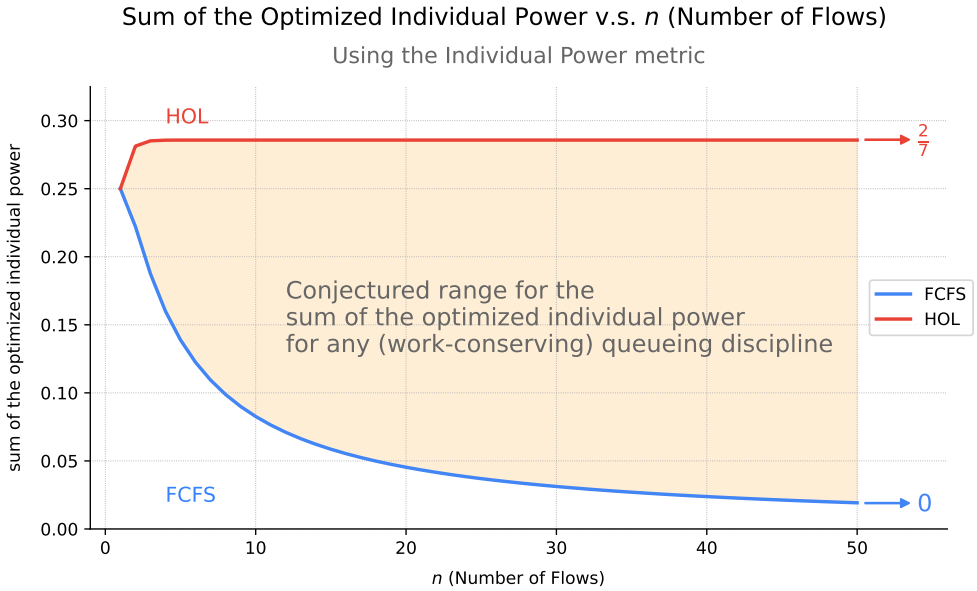


Figure 4.4: Trend of the optimized individual power summation versus the number of flows. HOL and FCFS are conjectured as the upper and lower bounds, respectively. The orange region between them is the conjectured range of possible sum of the optimized individual power for any work-conserving priority discipline.

divergent trends in overall performance. These two queueing disciplines are conjectured to serve as the bounds for the possible sum of optimized individual power values for any work-conserving priority discipline.

4.3 Iterative Optimization Process

In Section 4.2, we used mathematical equations to obtain the equilibrium results. Here, we explore the process of iterative optimization that leads to the same equilibrium results.

4.3.1 FCFS

The equilibrium can be achieved by allowing each flow to alternatively take turns optimizing its individual power. When one flow optimizes, the other flows are assumed to freeze. For example, for $n = 2$, flow 1 and flow 2 start with $\rho_i = 0$ for $i = 1, 2$, and flow 1 optimizes first. As illustrated in Figure 4.5, flow 1 sets its optimal utilization to $\rho_1^* = 0.5$. Next, flow 2 optimizes by taking half of the available (remaining) utilization and sets $\rho_2^* = \frac{1-\rho_1}{2} = \frac{1-0.5}{2} = 0.25$. Then, flow 1 takes its turn and re-optimizes its power by taking half of the now remaining available utilization: $\rho_1^* = \frac{1-\rho_2}{2} = \frac{1-0.25}{2} = \frac{3}{8} = 0.375$.

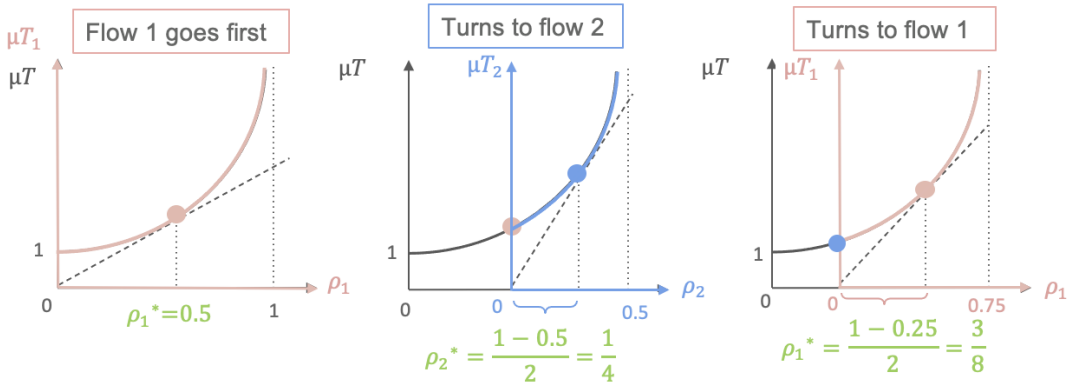


Figure 4.5: An example of two flows alternately optimizing their individual power, starting with $\rho_1 = \rho_2 = 0$ and flow 1 optimizing first. The figure illustrates the first three steps.

This process continues until they stabilize in the state where $(\rho_1, \rho_2) = \left(\frac{1}{3}, \frac{1}{3}\right)$, as shown in Figure 4.6. In the equilibrium state, both flows have the same utilization in the system. Consequently, each flow observes the same curve regarding their utilization and mean response time, leading to a steady-state. The iterative process of ρ_1 and ρ_2 is illustrated in Figure 4.7, showing the values of ρ_1 and ρ_2 at each iteration step. From the figure, we can see that they converge quickly, within about 10 steps. Similarly, iterative convergence occurs with n flows.

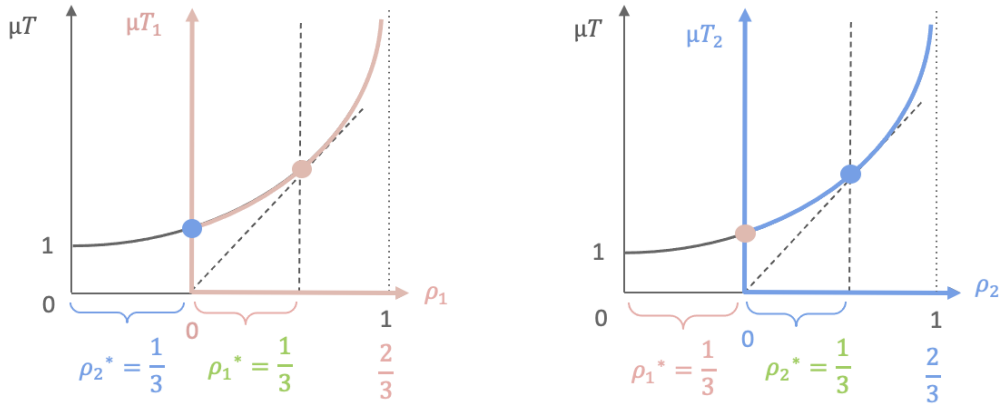


Figure 4.6: Equilibrium of two flows performing individual power optimization, with each flow perceiving the same mean response time curve.

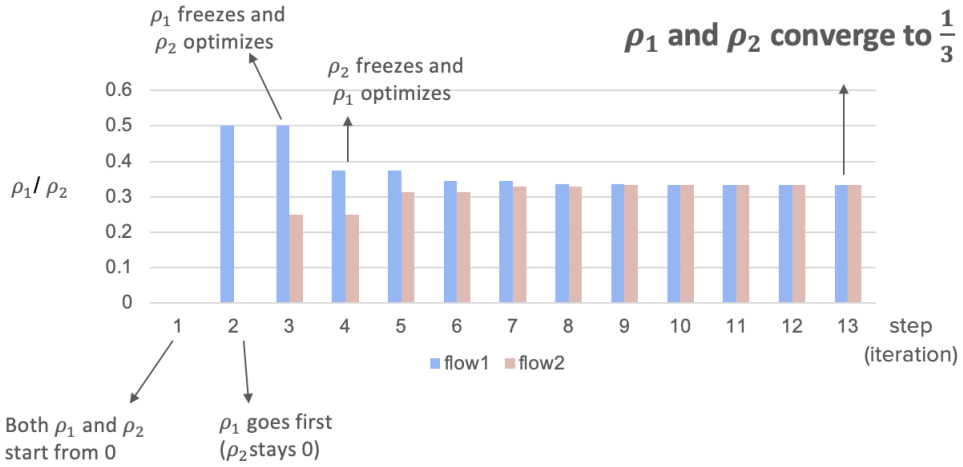


Figure 4.7: The evolution of ρ_1 and ρ_2 over individual power optimization iterations.

4.3.2 HOL

For the optimization iteration process in the HOL case, equilibrium can be reached after each flow optimizes its individual power based on the flow priority order once. Unlike the FCFS case, which requires multiple iterations to reach stability, the HOL case can achieve stability in just one iteration by following the priority order for optimization. Since higher priority flows are not affected by lower priority flows, they remain unchanged after optimization. Let's take an example of three flows in HOL case. The first flow takes 1/2, the second takes 1/4, and the third takes 1/8, after which all flows have reached their limiting optimization. This contrasts with FCFS, where each flow continuously adjusts its values during the optimization process until they collectively reach an equilibrium point.

4.4 Alternative Normalization Methods

In the previous sections, we used the average service time $\frac{1}{\mu}$ to normalize the response time in the definition of individual power. This was based on the assumption that each flow has the full capacity of the channel during its turn and thus the minimum time spent in the system, (i.e., without any waiting time) is the average service time $\frac{1}{\mu}$. However, if we consider other flows as uncontrollable, then the no-load delay used for normalization should be the average response time when the individual flow's load is zero but the other flows are non-zero, rather than the average response time when the total system load is zero. Additionally, the end user is unaware of the number of other flows in the system. To determine the minimum response time, the user sends infinitesimal traffic and measures the response time, which may account for delays caused by other flows. As a result, to better align with the limitations of the end user perspective, we now investigate the case using the no-load response time of the individual flow for normalization while allowing other flows in the system and analyze whether the result and the optimal operating point change.

4.4.1 Individual Power using Individual No Load Delay to Normalize

With normalization based on the delay when the individual load is zero, $T_i(0)$, (while other flows carry load), we use the power notation P_i^{NILD} to represent this case (NILD stands for No Individual Load Delay), defined as:

$$P_i^{\text{NILD}} = \frac{\rho_i}{\frac{T_i(\rho_i)}{T_i(0)}} \tag{4.38}$$

4.4.2 Optimization of Individual Power in FCFS

In an FCFS system, the individual flow's response time is given by

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

Therefore, the no-load delay (specifically, the delay with no utilization for flow i) is

$$T_i(0) = \frac{1}{\mu(1-\sum_{j=1,j\neq i}^{n}\rho_j)} = \frac{1}{\mu(1-\alpha)}$$

Using this for normalization, the individual power becomes:

$$P_i^{\text{NILD}} = \frac{\rho_i \cdot (1-\rho_i-\sum_{j=1,j\neq i}^{n}\rho_j)}{1-\sum_{j=1,j\neq i}^{n}\rho_j} = \frac{1-\rho_i-\alpha}{1-\alpha}$$

4.4.2.1 Optimizing Individual Power

To optimize this value, we take the derivative with respect to ρ_i and set it to zero:

$$\frac{\partial P_i^{NILD}}{\partial \rho_i} = \frac{\partial \frac{\rho_i \cdot (1-\rho_i-\sum_{j=1,j\neq i}^{n}\rho_j)}{(1-\sum_{j=1,j\neq i}^{n}\rho_j)}}{\partial \rho_i} = \frac{(1-2\rho_i-\sum_{j=1,j\neq i}^{n}\rho_j)}{(1-\sum_{j=1,j\neq i}^{n}\rho_j)} = 0$$

This leads to:

$$\rho_i^* = \frac{(1 - \sum_{j=1, j \neq i}^{n} \rho_j)}{2} = \frac{1 - \alpha}{2} \tag{4.39}$$

This result is half of the remaining utilization, which is the same outcome as when using the average service time for normalization as shown in Equation 4.4. This is because neither normalization term (average service time or the no individual load delay) contains ρ_i, and therefore each serves only as a constant in the optimization. As a result, the optimal ρ_i does not change when altering the value used for normalization. However, a change in the power value does occur as shown below.

4.4.2.2 Maximal Individual Power

The optimal individual power value in this case is:

$$P_i^{NILD*} = \frac{\rho_i(1 - \rho_i - \alpha)}{1 - \alpha} = \frac{\frac{1-\alpha}{2} \cdot \frac{1-\alpha}{2}}{1 - \alpha} = \frac{1 - \alpha}{4}$$

This value will be larger than the optimal individual power using average service time for normalization (AVS stands for the average service time):

$$P_i^{AVS*} = \rho_i(1 - \rho_i - \alpha) = \frac{(1 - \alpha)^2}{4}$$

This occurs because the response time is normalized by a larger value, resulting in a smaller normalized response time and thereby making the power value larger.

4.4.2.3 Multiple Flows Optimizing Individual Power

Since the ρ_i that maximizes individual power remains the same regardless of the normalization method used, we can infer that the optimal utilization factor for each flow i at equilibrium, when each flow optimizes its individual power, is identical to that which we derived earlier in Equation 4.11:

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

and the optimum system utilization is:

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

which is the same as in Equation 4.10.

Each individual power at this equilibrium point, where $\rho_i = \frac{1}{n+1} \quad$ for $i = 1, 2, \ldots, n$, is optimal and is given by:

$$P_i^{NILD*} = \frac{\rho_i(1-\rho)}{(1-\alpha)} = \frac{\frac{1}{n+1}(1-\frac{n}{n+1})}{1-\frac{n-1}{n+1}} = \frac{(\frac{1}{n+1})^2}{\frac{2}{n+1}} = \frac{1}{2(n+1)} \quad \text{for } i = 1, 2, \ldots, n$$

The sum of optimized individual power is then:

$$P_{\text{sum}}{}^{\text{NILD}} = \sum_{i=1}^{n} P_i^{NILD*} = \frac{n}{2(n+1)}$$

The limiting behavior as n approaches infinity is:

$$\lim_{n\to\infty} P_{\text{sum}}{}^{\text{NILD}} = \frac{1}{2}$$

4.4.3 Optimization of Individual Power in HOL

In HOL, the response time of each flow is $T_i = \frac{1}{\mu(1-\sigma_i)(1-\sigma_{i-1})}$ where $\sigma_i = \sum_{j=1}^{i-1} \rho_j$. The delay for flow i when its utilization is zero, while other flows carry load, is:

$$T_i(0) = \frac{1}{\mu(1-\sigma_{i-1})^2} \quad \text{for } i = 1, 2, \ldots, n$$

Hence, the individual power for flow i, normalized by the no-load delay of the individual flow while other flows carry load, is:

$$P_i^{NILD} = \frac{\rho_i}{\frac{T_i(\rho_i)}{T_i(0)}} = \frac{\rho_i(1-\sigma_i)}{(1-\sigma_{i-1})} \quad \text{for } i = 1, 2, \ldots, n$$

Optimizing this value by taking the derivative with respect to ρ_i, we have:

$$\frac{\partial P_i^{NILD}}{\partial \rho_i} = \frac{\partial \frac{\rho_i(1-\sigma_i)}{(1-\sigma_{i-1})}}{\partial \rho_i} = \frac{1-\sigma_i-\rho_i}{1-\sigma_{i-1}} = 0 \quad \text{for } i = 1, 2, \ldots, n$$

Solving this equation, we have:

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

This is the same as Equation 4.20, which is the result for individual power normalized using average service time.

Therefore, we know that this recursive equation can be solved as:

$$\rho_i^* = (\frac{1}{2})^i$$

and the optimum total utilization:

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

The optimized individual power is then:

$$P_i^{NILD*} = \frac{\rho_i(1-\sigma_i)}{(1-\sigma_{i-1})} = \frac{(1-\sigma_{i-1})}{2}\frac{(1-\sigma_i)}{(1-\sigma_{i-1})} = \frac{(1-\sigma_i)}{2} = \rho_{i+1} = (\frac{1}{2})^{i+1}$$

The sum of optimized individual power is:

$$P_{\text{sum}}{}^{\text{NILD}} = \sum_{i=1}^{n} P_i^{NILD*} = \sum_{i=1}^{n} (\frac{1}{2})^{i+1} = \frac{1}{2} \cdot [1 - (\frac{1}{2})^n]$$

The limiting behavior is:

$$\lim_{n\to\infty} P_{\text{sum}}{}^{\text{NILD}} = \lim_{n\to\infty} \frac{1}{2} \cdot [1 - (\frac{1}{2})^n] = \frac{1}{2}$$

4.4.4 Summary

4.4.4.1 FCFS

Table 4.2 summarizes the optimization results using these two different normalization methods for FCFS. As stated before, the optimal ρ_i^* and the sum of optimum utilization ρ^* are the same across different normalization approaches. The difference lies in the optimized individual power values and their trends as n increases. This is demonstrated in Figure 4.8. When using the average service time for normalization, the sum of optimized individual power values decreases as n increases. In contrast, when using the no individual load delay for normalization, the sum of optimized individual power values increases and approaches 0.5 as n increases.

4.4.4.2 HOL

Table 4.3 presents the optimization results for HOL using these two different normalization methods. Similar to the FCFS case, the optimal ρ_i^* and the total utilization ρ^* of the optimization result are the same across different normalization methods. However, the two normalizations differ in the optimized individual power values and the sum of optimized power values, with larger values observed when using no individual load delay to normalize the response time. Figure 4.9 shows the sum of optimized individual power values using different normalization terms versus the number of flows n. The trend indicates that the sum of optimized power increases as more flows are added to the system, with each new flow contributing to the sum of optimized power values until it reaches the limit of $\frac{1}{2}$ or $\frac{2}{7}$.

FCFS	average service time $\frac{1}{\mu}$	no individual load delay $T_i(0)$
ρ_i^*	$\frac{1}{n+1}$	$\frac{1}{n+1}$
ρ^*	$\frac{n}{n+1}$	$\frac{n}{n+1}$
$\lim_{n\to\infty} \rho^*$	1	1
P_i^*	$\frac{1}{(n+1)^2}$	$\frac{1}{2(n+1)}$
$\sum_{i=1}^{n} P_i^*$	$\frac{n}{(n+1)^2}$	$\frac{n}{2(n+1)}$
$\lim_{n\to\infty} \sum_{i=1}^{n} P_i^*$	0	$\frac{1}{2}$

Table 4.2: Optimization results of individual power optimization for **FCFS** using different normalization approaches.

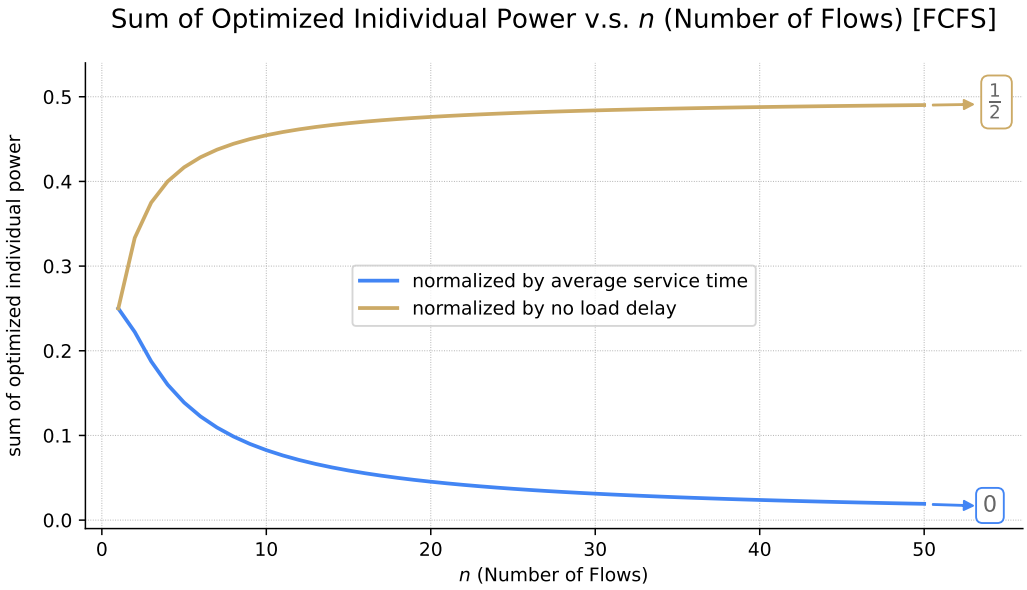


Figure 4.8: Comparison of sum of optimized individual power versus n using average response time and no individual load delay for normalization. Using the average service time approach shows a decreasing trend, while using the no individual load delay shows an increasing trend in the sum of optimized individual power values.

HOL	average service time $\frac{1}{\mu}$	no individual load delay $T_i(0)$
ρ_i^*	$(\frac{1}{2})^i$	$(\frac{1}{2})^i$
ρ^*	$1-(\frac{1}{2})^n$	$1-(\frac{1}{2})^n$
$\lim_{n\to\infty} \rho^*$	1	1
P_i^*	$2(\frac{1}{8})^i$	$(\frac{1}{2})^{i+1} = \frac{1}{2}(\frac{1}{2})^i$
$\sum_{i=1}^{n} P_i^*$	$\frac{2}{7}(1-(\frac{1}{8})^n)$	$\frac{1}{2}(1-(\frac{1}{2})^n)$
$\lim_{n\to\infty} \sum_{i=1}^{n} P_i^*$	$\frac{2}{7}$	$\frac{1}{2}$

Table 4.3: Optimization results of individual power optimization for **HOL** using different normalization approaches.

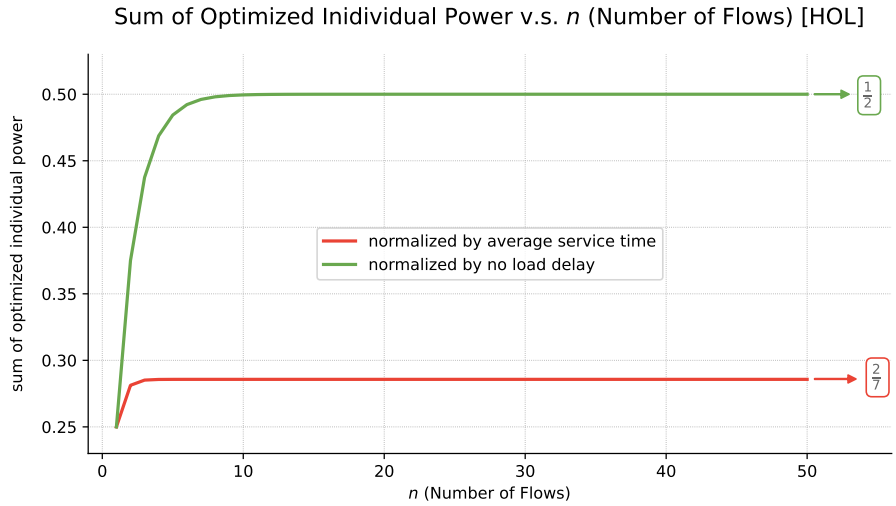


Figure 4.9: Comparison of sum of optimized individual power versus n using average response time and no individual load delay for normalization. Both approaches show an increasing trend as the number of flows increases but have different limit values. The average service time approach reaches $\frac{2}{7}$, while the no individual load delay approach reaches $\frac{1}{2}$.

Chapter 5: Performance Optimization Metric: Sum of Individual Powers, P_{sum}

In the previous chapter, we discussed optimizing individual power in FCFS and HOL and used the "sum of individual powers" (more properly "sum of optimized individual powers") to evaluate the optimization results for the two systems with different flow discrimination. However, the "sum of individual powers" computed there was not maximized because it was not the primary optimization objective, as we focused on the end-to-end viewpoint.

In this chapter, we shift our perspective from an end-to-end (individual flow) viewpoint to a system-wide viewpoint and change the optimization goal to **maximizing the "sum of individual powers"** to enhance the overall system performance. We will discuss how this system-wide viewpoint aligns with the metric of the "sum of individual powers". We will identify the optimal utilization factor ρ_i^* for each flow $i = 1, \ldots, n$, which collectively maximize the sum of powers for both the FCFS and HOL cases.

5.1 Description of the System-wide viewpoint

5.1.1 System operator

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 in Chapter 4.

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 metrics but also strives to utilize system resources efficiently and enhance overall performance.

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.

5.1.2 Active Queue Management in Routers

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 [48], ECN [28]) or by preemptively dropping packets before buffers reach capacity (e.g., RED [49]). 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. This is

increasingly vital with the emergence of the bufferbloat problem [50, 51]. The cause of bufferbloat is that buffers are getting larger and cheaper and help absorb bursts. However, these sizable buffers inadvertently exacerbate the issue by allowing TCP to flood the network with packets without experiencing loss, resulting in saturated buffers and significant delays.

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 in the system and the utilization ratio of high and low priority flows to prevent starvation becomes critical questions.

Therefore, it is crucial 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. Effective congestion control mechanisms in routers should be capable of determining when to trigger control actions and how to implement them under different queueing disciplines. This ensures the mitigation of issues like bufferbloat and starvation when priority services are provided for specific types of traffic.

5.2 The Metric: Sum of Individual Powers

5.2.1 Definition

The **sum of individual powers**[1] is defined as the total of the individual powers of all flows within a system. This metric provides a summary measure of each flow's system

[1] We may also use "*sum of powers*" to refer to "*sum of individual powers*".

experience, evaluated based on overall power. Mathematically, it is expressed as:

$$P_{\text{sum}} = \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} \frac{\rho_i}{\mu T_i} \tag{5.1}$$

Here, P_i represents the individual power of the i^{th} flow, ρ_i denotes the utilization of the i^{th} flow, μ indicates the average service rate, and T_i is the mean response time associated with the i^{th} flow.

The formula aggregates the power of each flow, ensuring every flow is considered. In a system with flow discrimination, focusing solely on high-priority flows can significantly impact lower-priority flows. By considering the sum of individual powers, we ensure that higher-priority flows do not operate independently of lower-priority flows, preventing potential negative impacts. This aggregation provides a more balanced resource allocation, especially in HOL systems, where higher-priority flows may be unaware of the needs of lower-priority flows.

With the definition of the sum of individual powers established, we now proceed to find its maximal value, which we denote as P^*_{sum}. Specifically, we aim to determine the operating points for the set of utilizations for each flow $\rho_1, \rho_2, \ldots, \rho_n$ that collectively maximize the sum of individual powers.

5.2.2 Optimizing Sum of Individual Powers

To find the operating point that maximizes the sum of individual powers, we need to identify the critical points of the function for sum of individual powers. This involves calculating the partial derivatives with respect to each variable and simultaneously setting them all to zero.

The process is expressed as follows:

$$\frac{\partial P_{\text{sum}}}{\partial \rho_i} = \frac{\partial}{\partial \rho_i} \sum_{i=1}^{n} \frac{\rho_i}{\mu T_i} = 0 \quad \text{for } i = 1, 2, \ldots, n.$$

By solving these equations simultaneously, we can find the critical points for each ρ_i, which as usual we denote as ρ_i^* (for $i = 1 \cdots, n$). These critical points maximize the sum of individual powers.

5.3 Optimizing Sum of Individual Powers in FCFS

Now, let's apply this optimization process to our M/M/1 system with n flows using FCFS queueing. 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 $i = 1, \ldots, n$. This uniform response time is determined by the overall system utilization, $\rho = \sum_{i=1}^{n} \rho_i$.

The sum of individual powers for all flows in the FCFS system is computed as follows:

$$\sum_{i=1}^{n} P_i = \sum_{i=1}^{n} \frac{\rho_i}{\mu T} = \sum_{i=1}^{n} \rho_i (1 - \rho) = \rho(1 - \rho)$$

Thus, the sum of the individual powers for the FCFS system is:

$$P_{\text{sum}} = \rho(1 - \rho) \tag{5.2}$$

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:

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

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 the above partial derivative equation:

$$\frac{\partial \rho(1-\rho)}{\partial \rho_i} = \frac{\partial \rho(1-\rho)}{\partial \rho} \cdot \frac{\partial \rho}{\partial \rho_i} = (1-2\rho) \cdot 1 = 1 - 2\rho = 0 \quad \text{for } i = 1, 2, \ldots, n$$

Solving this equation, we find:

$$\rho^* = \frac{1}{2} \tag{5.3}$$

This implies that when the total system utilization ρ^* is equal to $\frac{1}{2}$, the sum of the individual powers in the FCFS system reaches its maximum value. This occurs regardless of how the individual utilizations ρ_i are distributed among different flows within the system; they simply must sum to $\rho = \frac{1}{2}$. This is consistent with Section 2.2.3 since in FCFS, the sum of the n Poisson flows is equivalent to a single flow at a traffic level of ρ, and we know maximum power occurs at $\rho = \frac{1}{2}$ for M/M/1, as shown in Equation 2.6.

Substituting Equation 5.3 into Equation 5.2, we have the maximal sum of individual powers value:

$$P^*_{\text{sum}} = \rho(1-\rho) = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

This gives us the following unsurprising 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:

$$P^*_{sum} = \frac{1}{4}$$

when

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

where the distribution of ρ_i for $i = 1, \ldots, n$ is irrelevant as long as their sum $\rho^* = \frac{1}{2}$.

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 ρ. 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.

5.4 Optimizing Sum of Individual Powers in HOL

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

5.4.1 Two Flows

In the head-of-line priority system, the mean response time for flow 1 (the higher priority flow) is given by

$$T_1 = \frac{1}{\mu(1-\rho_1)}$$

and for flow 2, the mean response time is

$$T_2 = \frac{1}{\mu(1-\rho_1)(1-\rho_1-\rho_2)}$$

Therefore, the individual power of flow 1 is

$$P_1 = \rho_1(1-\rho_1)$$

and the individual power for flow 2 is

$$P_2 = \rho_2(1-\rho_1)(1-\rho_1-\rho_2).$$

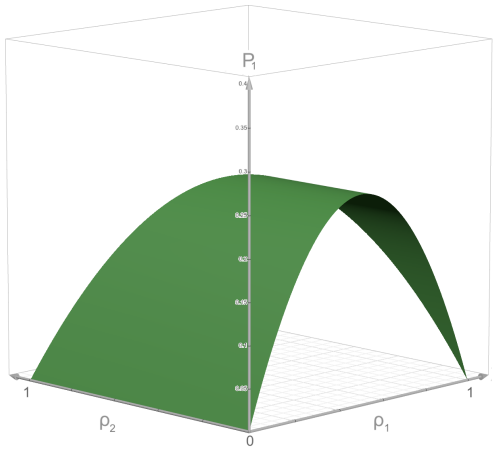
The sum of the individual powers in HOL is

$$P_{\text{sum}} = P_1 + P_2 = \rho_1(1-\rho_1) + \rho_2(1-\rho_1)(1-\rho_1-\rho_2) \tag{5.4}$$

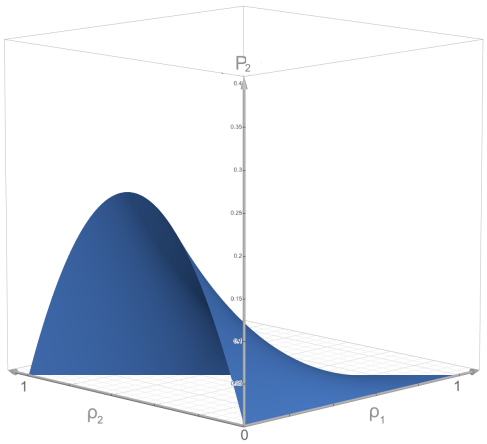
This can be simplified and expressed as:

$$P_{\text{sum}} = (\rho_1 + \rho_2)\ (1 - \rho_1)\ (1 - \rho_2) \tag{5.5}$$

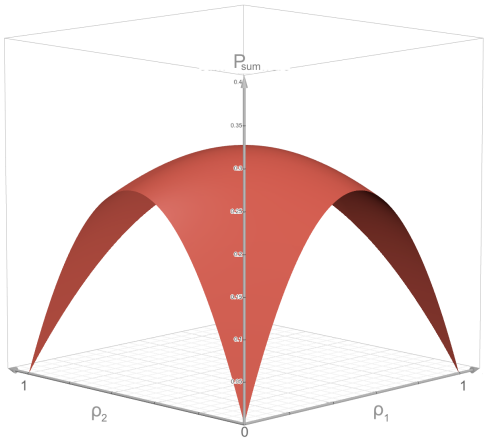
To better understand the practical implications of this equation, let us examine a visual representation. Figure 5.1 illustrates the interaction between the individual power of two separate flows and their cumulative effect on the system total powers.



(a) $P_1 = \rho_1(1 - \rho_1)$



(b) $P_2 = \rho_2(1 - \rho_1)(1 - \rho_1 - \rho_2)$



(c) $P_{\text{sum}} = (\rho_1 + \rho_2)(1 - \rho_1)(1 - \rho_2)$

Figure 5.1: These figures show the individual power of each flow and the sum of power in a two-flow system under HOL. All plots are 3D surfaces showing power as a function of the utilization factors of the two flows (ρ_1 and ρ_2).

Figure 5.1a illustrates the individual power of flow 1, $P_1 = \rho_1(1 - \rho_1)$, Figure 5.1b depicts the individual power of flow 2, $P_2 = \rho_2(1 - \rho_1)(1 - \rho_1 - \rho_2)$; and Figure 5.1c presents the sum of these individual powers, $P_{\text{sum}} = (\rho_1 + \rho_2)(1 - \rho_1)(1 - \rho_2)$. This visualization aids in understanding how changes in one flow's utilization impacts the other flow and the total sum of powers.

5.4.1.1 Properties

From Figure 5.1, we observe that the sum of powers is symmetric along the plane $\rho_1 = \rho_2$. This symmetry indicates that the values of ρ_1 and ρ_2 are interchangeable without affecting the total powers value. This property can also be confirmed by examining the equation for the sum of powers, given in Equation 5.5. Whether the values for the variables (ρ_1, ρ_2) are (x, y) or (y, x), the resulting sums of powers are the same with the value of $(x + y)(1 - x)(1 - y)$. This leads to the following theorem:

Theorem 5.2.

For an M/M/1 system with two flows using HOL as queueing discipline, the sum of individual powers can be expressed as a function of the two variables (ρ_1, ρ_2)

$$P_{sum}(\rho_1, \rho_2) = (\rho_1 + \rho_2)(1 - \rho_1)(1 - \rho_2) \tag{5.6}$$

This function is ***symmetric****, meaning that the value of the sum of individual powers remains unchanged when the variables* (ρ_1, ρ_2) *are switched. Mathematically, this symmetry is expressed as:*

$$P_{sum}(x, y) = P_{sum}(y, x) \tag{5.7}$$

where x *and* y *represent any values of the variables* ρ_1 *and* ρ_2*.*

This symmetry property implies that the order in which the utilizations ρ_1 and ρ_2 are considered does not affect the overall sum of powers, highlighting the interchangeable nature of the two variables in this context.

Another observation is that, when the total utilization ρ is fixed with c, the sum of power increases as the difference between ρ_1 and ρ_2 decreases. To visually clarify this statement, we refer to Figure 5.2, where the purple surface represents a plane defined by $\rho_1 + \rho_2 = 0.4$ (with c set to be 0.4, an arbitrary value chosen for illustration). The intersection of this plane with the sum of powers surface peaks at the midpoint where ρ_1 and ρ_2 are equal. This leads to the following theorem:

Theorem 5.3.

In an M/M/1 systems with two flows using HOL, given the constraint that the total system utilization ρ is a fixed number, say c (where $0 < c < 1$), then the sum of individual powers is maximal when each of the two flows shares the fixed amount c equally:

$$\rho_1 = \rho_2 = \frac{c}{2} \tag{5.8}$$

Proof:

Given that $\rho = \rho_1 + \rho_2 = c$ (where $0 < c < 1$), the sum of individual powers:

$$(\rho_1 + \rho_2)(1 - \rho_1)(1 - \rho_2)$$

can be expressed as

$$c(1 - \rho_1)[1 - (c - \rho_1)] = c(1 - \rho_1)(1 - c + \rho_1)$$

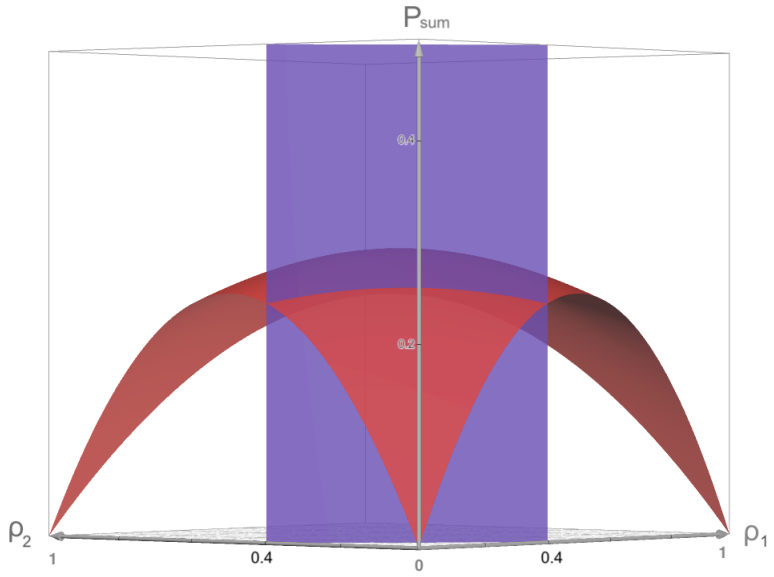


Figure 5.2: Visualization of distribution of sum of powers with a constraint of fixed total utilization ($\rho_1 + \rho_2 = c = 0.4$). The plot illustrates the sum of powers surface intersecting with a purple plane that represents the constraint of constant system utilization. Theorem 5.3 shows that the optimal sum of powers occurs when the individual utilizations are equal (in this example, $\rho_1 = \rho_2 = 0.2$).

To find the maximum of this function, we take the derivative of this expression with respect to ρ_1:

$$\frac{\partial c(1-\rho_1)(1-c+\rho_1)}{\partial \rho_1} = c \cdot (-(1-c+\rho_1) + (1-\rho_1)) = c(c-2\rho_1)$$

Setting the derivative equal to zero gives:

$$c(c-2\rho_1) = 0 \implies \rho_1 = \frac{c}{2}$$

since $c \neq 0$ by the constraint.

Substituting this back into $\rho_1 + \rho_2 = c$, we have:

$$\rho_2 = c - \rho_1 = c - \frac{c}{2} = \frac{c}{2}$$

Thus, it is shown that under the constraint of a fixed total system utilization $\rho = c$, where $0 < c < 1$, the sum of individual powers is maximal when $\rho_1 = \rho_2 = \frac{c}{2}$.

■

Based on the observation of symmetry in Figure 5.1c, we established Theorem 5.2 and Theorem 5.3. These theorems provide valuable insights into optimizing flow management for enhanced system performance. Such understanding is crucial for improving efficiency and ensuring that resources are allocated effectively to maximize the system's overall performance.

5.4.1.2 Optimization

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:

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

To simplify these equations, 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 both sides of the equations by 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 values of ρ_1^* and ρ_2^* that maximize the sum of power for HOL are:

$$\rho_1^* = \rho_2^* = \frac{1}{3} \tag{5.9}$$

This outcome reveals that the most effective 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, the ideal utilization for each flow should be exactly one-third of the total system utilization. This configuration ensures that the cumulative power of the system is optimized, reflecting a balanced approach where each flow contributes equally to achieving the highest possible efficiency. This balanced allocation not only enhances the system's performance in terms of power but also ensures a more equitable distribution of resources among the flows.

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

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

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

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

$$P^*_{\text{sum}} = P_1 + P_2 = \frac{2}{9} + \frac{2}{27} = \frac{6}{27} + \frac{2}{27} = \frac{8}{27} \approx 0.296$$

This value surpasses the sum of individual powers realized in an FCFS system, which is only 0.25. This trend is consistent with the findings discussed in the previous chapter, where the optimal value for HOL exceeds that of FCFS when optimizing individual power. Such consistency in performance metrics underscores the superior efficiency of HOL compared to FCFS, particularly in scenarios where maximizing power is paramount.

5.4.2 n Flows

From the two-flows case, we derived two theorems about the properties of the sum of individual powers metric and optimized this metric in an M/M/1 system using HOL as the queueing discipline. The key results are summarized as follows:

- The sum of individual powers is maximal with the value $\frac{8}{27}$ when the two flows have equal utilization of $\frac{1}{3}$.
- The sum of powers function is symmetric, indicating that ρ_1 and ρ_2 are interchangeable without changing the sum of power value.
- Given that the total utilization ρ is a fixed value, say c, the sum of powers is maximal when $\rho_1 = \rho_2 = \frac{c}{2}$.

In the following, we extend the analysis from two flows to n flows and focus on finding the maximal sum of powers. As the dimension n increases beyond 2, it becomes difficult to visualize the sum of powers metric as a function of each ρ_i (for $i = 1, 2, \ldots, n$), making it

challenging to directly determine whether the properties observed in the two-flow case—such as the symmetry of the sum of powers function and the maximization of the sum of powers when each flow i has the same utilization $\rho_i = \frac{c}{n}$ under a fixed total utilization c—still hold. Therefore, we leave these complexities for future work.

5.4.2.1 Optimizing Sum of Individual Powers

In the two-flow case, the system utilization for each flow is proven to be equal when optimizing the sum of individual powers in the preemptive head-of-line (HOL) queueing system. This principle can be extended to an arbitrary number of flows in a preemptive HOL priority system with n groups of flows and leads to the following theorem:

Theorem 5.4.

Given the HOL preemptive priority queueing discipline with n flows in an M/M/1 system, the sum of powers is

$$P_{sum} = \sum_{i=1}^{n} \rho_i (1 - \sigma_{i-1})(1 - \sigma_i) \tag{5.10}$$

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

The sum of powers is maximal when each flow has the same utilization:

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

and the optimum total system utilization is:

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

Proof:

Equation 4.19 provides the individual power of flow i. Thus, the sum of the individual powers is given by Equation 5.10:

$$P_{\text{sum}} = \sum_{i=1}^{n} \rho_i (1 - \sigma_{i-1})(1 - \sigma_i)$$

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:[2]

$$\begin{aligned} \frac{\partial}{\partial \rho_i} P_{\text{sum}} &= \frac{\partial}{\partial \rho_i} \sum_{j=1}^{n} \rho_j (1 - \sigma_{j-1})(1 - \sigma_j) \\ &= (1 - \sigma_n - \rho_i)(1 - \sigma_n + \rho_i) \quad \text{for } i = 1, \ldots, n \end{aligned} \tag{5.11}$$

Then in step 2, we will use this result to find the effective critical point:

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

by setting each partial differential equation to zero and solving them simultaneously, under the assumption that the total utilization $\rho < 1$:

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

Step 1: Proving the Partial Differentiation Equation

We now prove the partial differentiation equation (Equation 5.11) by induction:

- Base Case ($n = 2$): From Equation 5.4, the sum of individual powers when $n = 2$ is:

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

When optimizing the sum of power 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:

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

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 Equation 5.11 where $n = 2$ for $i = 1, 2$.

- Induction Hypothesis:

  Suppose the partial differentiation equation (Equation 5.11) works when the number of flows n is k:

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

- Induction Step:

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

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

Here's the computation:

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

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

Thus, by induction, we have shown that Equation 5.11 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 Equation 5.11 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:

- $(1 - \sigma_n + \rho_i) = 0$:

  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.

- $(1 - \sigma_n - \rho_i) = 0$:

  For $(1 - \sigma_n - \rho_i) = 0$, we have

$$\rho_i = 1 - \sigma_n \quad \text{for } i = 1, 2, \ldots, n \tag{5.12}$$

  Summing all equations, we get:

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

  Because $1 - \sigma_n$ is independent of i and $\sigma_n = \sum_{i=1}^{n} \rho_i$, the above equation can be expressed as:

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

Thus, we compute σ_n as:

$$\sigma_n = \frac{n}{n+1} \tag{5.13}$$

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

$$\rho^* = \frac{n}{n+1} \tag{5.14}$$

as was to be shown. With σ_n computed, we now determine ρ_i using Equation 5.12 and Equation 5.13:

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

Thus, we have also shown that:

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

■

Based on Theorem 5.4, we have $\rho^* = \frac{n}{n+1}$ when the sum of individual power is maximal. Since n is always less than $n+1$, $\rho^* = \frac{n}{n+1}$ is always less than 1. Additionally, ρ^* depends on n. The behavior of ρ^* as n approaches infinity is stated in the following corollary:

Corollary 5.4.1.

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

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

This corollary indicates that as the number of flows increases indefinitely, the optimized system utilization approaches its upper bound of 1.

5.4.2.2 Maximal Sum of Individual Power Value

Knowing that the sum of individual power is optimal when each flow's utilization is given by

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

we can now proceed to compute this maximal value. By substituting this utilization into the expression for the sum of individual power, we derive the following theorem:

Theorem 5.5.

Consider an M/M/1 system with n flows using the head-of-line preemptive priority queueing disciplines with, the maximal sum of individual powers is

$$P_{sum}^* = \frac{n(n+2)}{3(n+1)^2} \tag{5.15}$$

where each flow's individual power is:

$$P_i = \frac{(n+1-i)(n+2-i)}{(n+1)^3} \quad \textit{for } i = 1, 2, \ldots, n \tag{5.16}$$

Proof

First, we substitute $\rho_i^* = \frac{1}{n+1}$ for $i = 1, 2, \ldots, n$ into the formula for each individual power P_i. We know from Equation 4.19 that the individual power $P_i = \rho_i(1 - \sigma_{i-1})(1 - \sigma_i)$, where σ_i represents the cumulative utilization up to the i^{th} flow, calculated as $\sigma_i = \sum_{j=1}^{i} \rho_j$. After substituting the optimized $\rho_i^* = \frac{1}{n+1}$ into the formula in place of ρ_i and simplifying, we derive

the expression for P_i as follows:

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

Next, we compute the summation of individual power across all flows at optimization:

$$\begin{aligned}
\sum_{i=1}^{n} P_i &= \frac{1}{(n+1)^3}\sum_{i=1}^{n}(n+1-i)(n+2-i) \\
&= \frac{1}{(n+1)^3}\sum_{i=1}^{n}[(n+1)(n+2)-i(2n+3)+i^2] \\
&= \frac{1}{(n+1)^3}\left[(n+1)(n+2)n-(2n+3)\frac{n(n+1)}{2}+\frac{n(n+1)(2n+1)}{6}\right] \\
&= \frac{n(n+1)}{(n+1)^3}\left[(n+2)-\frac{(2n+3)}{2}+\frac{(2n+1)}{6}\right] \\
&= \frac{n(n+2)}{3(n+1)^2}
\end{aligned}$$

Thus, we have shown that the maximal sum of individual power P^*_{sum} for n flows in an M/M/1 system using HOL as queueing discipline is: $\frac{n(n+2)}{3(n+1)^2}$

■

To gain further insight into the behavior of this expression as the number of flows n becomes very large, we examine its limit as n approaches infinity. By evaluating the limit, we can determine the asymptotic value of the optimum sum of individual powers for HOL, as

shown in the following corollary:

Corollary 5.5.1.

Consider an M/M/1 system with n flows using the head-of-line preemptive queueing disciplines. As n approaches infinity, the limit of the optimum sum of individual powers is

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

This corollary indicates that as the number of flows increases indefinitely, the maximal sum of individual powers converges to $\frac{1}{3}$.

5.5 Comparison of Optimization Results

Before diving into the analysis of the properties of the optimal sum of individual powers with n flows, we first discuss the comparison of the various optimization results that have been derived. Specifically, we will compare maximal sum of individual powers across different queueing disciplines. Then, in Section 5.5.2, we will compare the results within the same queueing discipline but using different optimization metrics. This comparison provides a better understanding of each optimization metric across different queueing disciplines, offering insights into the effectiveness and efficiency of each approach.

5.5.1 Comparison with FCFS

We first compare the optimization results of FCFS and HOL derived in this chapter using the sum of individual power metric as the optimization goal in Table 5.1. In FCFS, the individual utilization factor ρ_i is not restricted as long as the sum of each flow is 0.5; therefore, we leave it blank in the table. The same applies to P_i for FCFS. Some values in the table depend on n.

	FCFS	**HOL**
ρ_i^*		$\frac{1}{n+1}$
ρ^*	$\frac{1}{2}$	$\frac{n}{n+1}$
$\lim_{n\to\infty} \rho^*$	$\frac{1}{2}$	1
P_i		$\frac{(n+1-i)(n+2-i)}{(n+1)^3}$
$P_{\text{sum}}^* = \sum_{i=1}^{n} P_i$	$\frac{1}{4}$	$\frac{n(n+2)}{3(n+1)^2}$
$\lim_{n\to\infty} P_{\text{sum}}^*$	$\frac{1}{4}$	$\frac{1}{3}$

Table 5.1: Optimization result of using "sum of individual powers" as the optimization goal. 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.

To better understand how P_{sum}^* and the optimized ρ^* that achieves this maximum behave across different values of n, we plot them against n in Figure 5.3 for ρ^* and in Figure 5.4 for P_{sum}^*, respectively. In Figure 5.3 and Figure 5.4, the curves for HOL and FCFS are conjectured to serve as the upper and lower bounds[3]. These two curves in each figure bound the region for the maximum sum of individual power and the optimized ρ^* from the summation of ρ_1^* and ρ_2^* that achieve the maximum. The curves for other disciplines are conjectured to lie within the regions bounded by these two curves.

The maximum sum of power for FCFS remains at 0.25 as n increases, with ρ^* kept at 0.5 regardless of the number of flows in the system. In contrast, HOL, which represents maximal flow discrimination, achieves a maximum for the optimized sum of individual power that increases as n increases, approaching an asymptotic value of $\frac{1}{3}$ as n becomes very large, with

[3] The conjecture that HOL is the upper bound and FCFS is the lower bound is based on the numerical results in Chapter 7.

ρ^* increasing to approach 1.

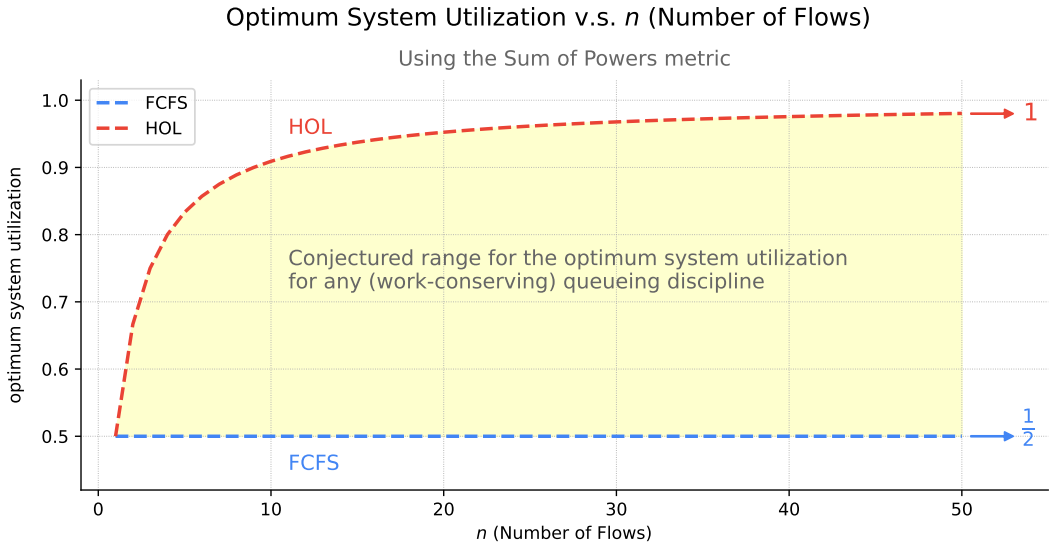


Figure 5.3: The ρ^* that achieves the maximal sum of individual powers is shown for both FCFS and HOL, corresponding to the variation of n.

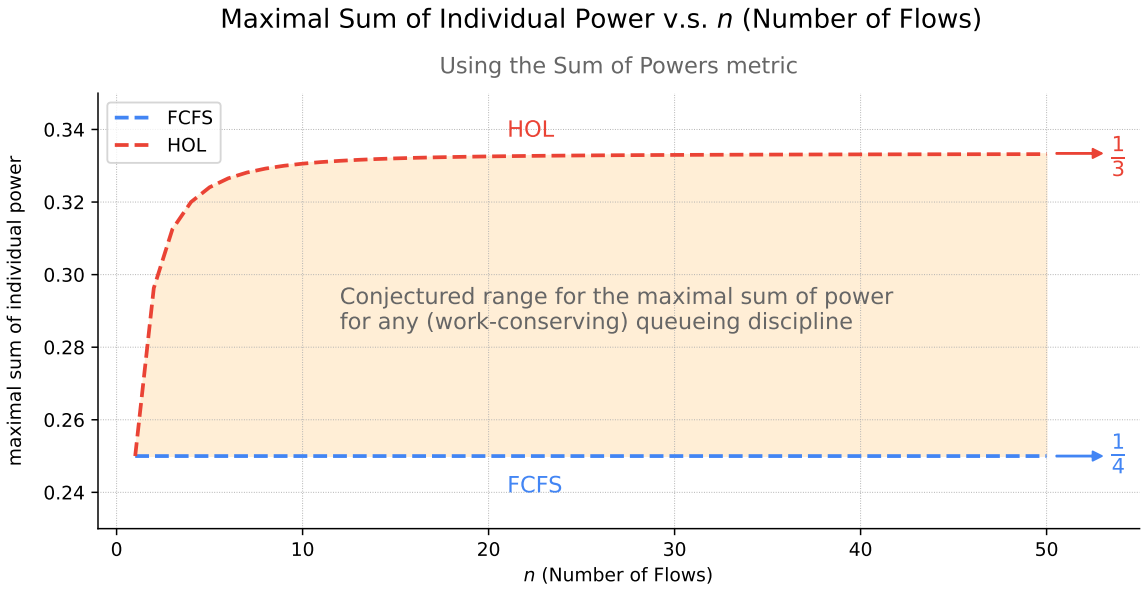


Figure 5.4: Optimized sum of individual power P^*_{sum} versus n for both FCFS and HOL.

5.5.2 Comparison of two Performance Metrics

Now, let's compare the optimization results with those derived in Chapter 4. Figures 5.5 and 5.6 display the variation of ρ and power with respect to n for different optimization metrics and queueing disciplines. In each figure, optimization results from optimizing individual power for all flows are represented by solid lines, while results from maximizing the sum of individual power are presented with dashed curves. Curves for FCFS results are shown in blue, while curves for HOL results are shown in red.

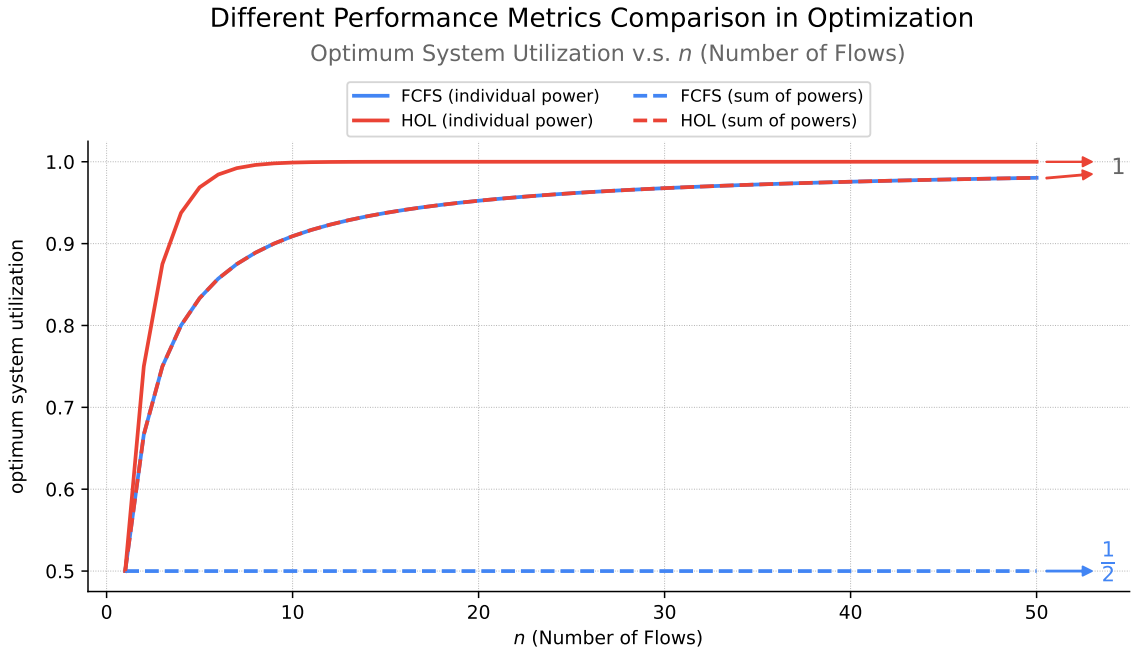


Figure 5.5: Optimum ρ vs. n in FCFS and HOL with different performance metrics used in optimization. The optimization results using "Individual Power" are represented by solid lines, while the results using "Sum of Powers" are presented by dashed lines.

In Figure 5.5, only the curve for maximizing the sum of power in FCFS remains at 0.5, while the other three curves approach 1 as n approaches infinity. For those three curves, maximizing individual power in HOL achieves 1 when n is around 10, which is much earlier than the other two. The other two curves—maximizing individual power in FCFS and

maximizing the sum of power in HOL—are the same.

In Figure 5.6, the red curves from HOL are both larger than the two blue curves from FCFS, indicating better system resource usage and overall benefits when evaluated based on the sum of individual power. Moreover, the solid lines enclose a larger region compared to the dashed lines, implying more divergent behavior in terms of the sum of power. Furthermore, both red curves of HOL exhibit a jump at around $n = 2$ and then increase slowly thereafter. This indicates that while the sum of power in HOL increases as n increases, a few flows in the system are already close to reaching its limit. Subsequently, adding more flows to the system does not significantly contribute to the sum of power.

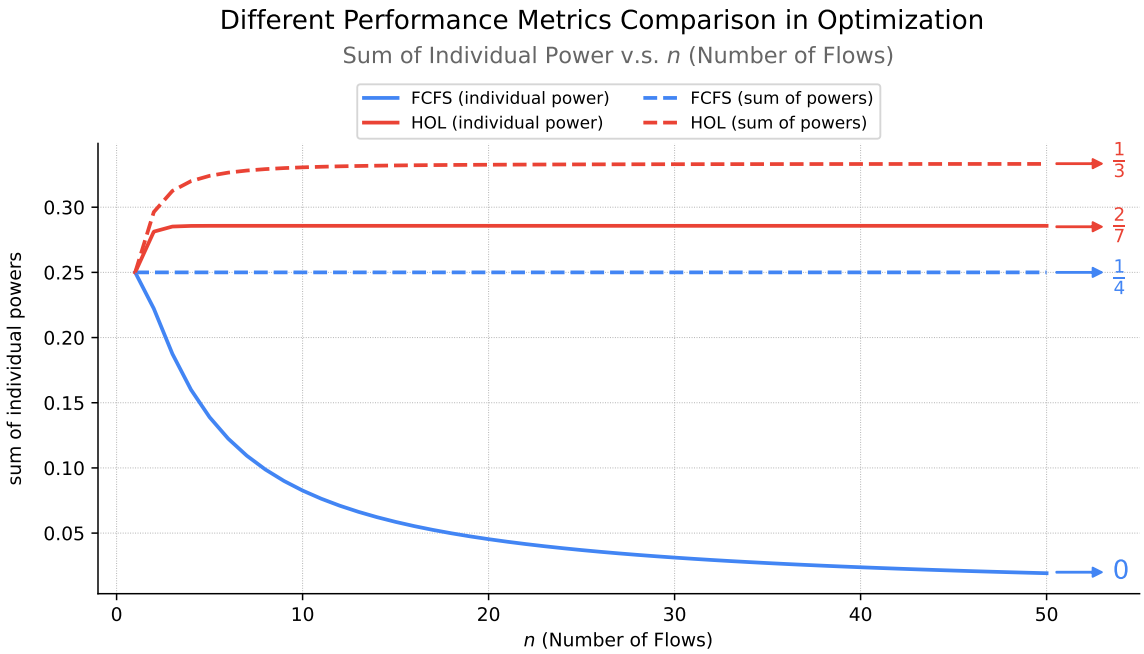


Figure 5.6: Optimized Sum of Powers vs. n in FCFS and HOL with different performance metrics used in optimization. When "Individual Power" is used as the metric, the sum of powers at the optimal operating point corresponds to the sum of **maximal individual power**, represented by solid lines. When "Sum of Powers" is used as the metric, the sum of powers at the optimal operating point reflects the **maximal sum of individual power**, represented by dashed lines.

Chapter 6: Performance Optimization Metric: Average Power, P_{avg}

In the previous chapter, we proposed the sum of individual powers from a system-wide perspective. In this chapter, we introduce an alternative metric—our third metric, average power—to represent the system's overall performance. Remarkably, by adopting this metric, we can utilize the conservation law [33] to demonstrate that system-level performance remains stable even when flow prioritization is implemented. This insight provides significant flexibility in system design without compromising performance.

6.1 Average Power

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:

$$P_{\text{avg}} = \frac{\sum_{i=1}^{n} \rho_i}{\sum_{i=1}^{n} \left(\frac{\rho_i}{\rho} \mu T_i \right)} \tag{6.1}$$

In this equation, the numerator represents the summation of the utilization factor for each traffic flow, where ρ_i corresponds to the utilization factor of the i^{th} flow, computed as $\rho_i = \frac{\lambda_i}{\mu}$.

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.

6.1.1 Property

With this definition, the average power can be expressed in a form that solely depends on ρ, leveraging the conservation law and under the assumption that each flow has the same mean service rate, represented by $\mu_i = \mu$ for $i = 1, \ldots, n$. This characteristic is not confined to M/M/1 queue systems but extends to all M/G/1 work-conserving systems as well. To formalize this result, we establish the following theorem:

Theorem 6.1.

*For an **M/G/1** system with n flows, all sharing the same mean service rate and operating under any **work-conserving** queueing discipline, the average power* $P_{avg} = \frac{\sum_{i=1}^{n} \rho_i}{\sum_{i=1}^{n} \left(\frac{\rho_i}{\rho} \mu T_i\right)}$ *is given by:*

$$P_{avg} = \frac{\rho(1-\rho)}{\mu W_0 + (1-\rho)} \tag{6.2}$$

where

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

is the average remaining service time for the customer found in service by a new arrival from a Poisson arrival process. The term $\overline{x_i^2}$ *denotes the second moment of the service time for the* i^{th} *flow.*

Proof

For each flow, the average response time, T_i, can be broken down into two components: the average waiting time, W_i, and the average service time, $\frac{1}{\mu}$. This leads to the equation:

$$T_i = W_i + \frac{1}{\mu} \tag{6.3}$$

The subscript i specifies the i^{th} flow, highlighting that the waiting times may vary between flows. It is assumed that the average service time, $\frac{1}{\mu}$, is the same for all flows, as $\mu_i = \mu$ for $i = 1, \ldots, n$.

We substitute Equation 6.3 into Equation 6.1:

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

After substitution, we find that $\sum_{i=1}^{n} \rho_i W_i$ remains constant under an M/G/1 system utilizing any work-conserving queueing discipline. "Work-conserving" means that no work (or service requirement) is created or destroyed within the system, and the system's workload neither increases nor decreases. The detailed definition of the work-conserving queueing class, as outlined in Section 5.2 of [37], includes the following: no work leaves before completion

(no defections); no workload is created within the system; preemption is allowed only for exponentially distributed service times in a preemptive-resume setting; and no server remains idle when work is present.

This work-conserving principle applies to various queueing models, including non-preemptive priority and preemptive-resume priority disciplines. Of all work-conserving queueing types, the head-of-line preemptive-resume queueing results in the most significant discrimination in response times among flows, whereas the first-come-first-served discipline exhibits the least discrimination, effectively showing no discrimination.

The constancy of $\sum_{i=1}^{n} \rho_i W_i$ is demonstrated by the conservation law [33] and is given by:

$$\sum_{i=1}^{n} \rho_i W_i = \frac{\rho W_0}{1-\rho} \quad \text{for} \quad \rho < 1 \tag{6.4}$$

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

By substituting the constant value of $\sum_{i=1}^{n} \rho_i W_i$, Equation 6.4, into the equation for P_{avg}, Equation 6.1, we have:

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

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

■

From Equation 6.1, we see that the term W_0 is influenced by the second moment of each flow's service time. If we further assume that each flow's service time shares the same second moment, that is $\overline{x_i^2} = \overline{x^2}$ for $i = 1, \ldots, n$, we can establish the following theorem:

Theorem 6.2.

For an M/G/1 system with multiple 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 be expressed as:

$$P_{avg} = \frac{\rho}{1 + \frac{\rho(1+C_b^2)}{2(1-\rho)}} \tag{6.5}$$

Proof:

If the service times for all flows have identical second moments $\overline{x^2}$ [1], then the average residual 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

$$W_0 = \frac{\lambda \overline{x^2}}{2} \tag{6.6}$$

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

$$\overline{x^2} = (1 + C_b^2)\frac{1}{\mu^2} \tag{6.7}$$

The coefficient of variation squared C_b^2 measures the variability of service times relative to their mean, calculated by:

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

where $\overline{x}$ represents the first moment of service time, which is $\frac{1}{\mu}$.

---

[1] 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 $\frac{2}{\mu^2}$.

The relationship between the second moment $\overline{x^2}$ and the coefficient of variation squared C_b^2 is derived as follows:

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

Substituting this back into the equation for W_0, Equation 6.6, we get:

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

This allows W_0 to be simplified as:

$$W_0 = \frac{\rho(1 + C_b^2)}{2\mu} \tag{6.8}$$

Thus, by expressing W_0 in terms of C_b^2, μ, and ρ, Equation 6.2 for P_{avg} can be reformulated as:

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

Thus, we have proven Equation 6.5. 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 detailed in Chapter 2. Hence, we have proven Theorem 6.2.

■

Theorem 6.2 illustrates that the performance of a single-server system with multiple flows, as depicted in Figure 6.1, remains unaffected by the specific queueing discipline used, provided it is work-conserving and the first and second moments of the service rate are the same for all flows. This invariance means that even if the order in which flows are processed changes, as long as the queueing discipline remains work-conserving, the overall performance of the system in terms of average power will not be impacted. Consequently, the whole system can

be viewed as a single-flow system.

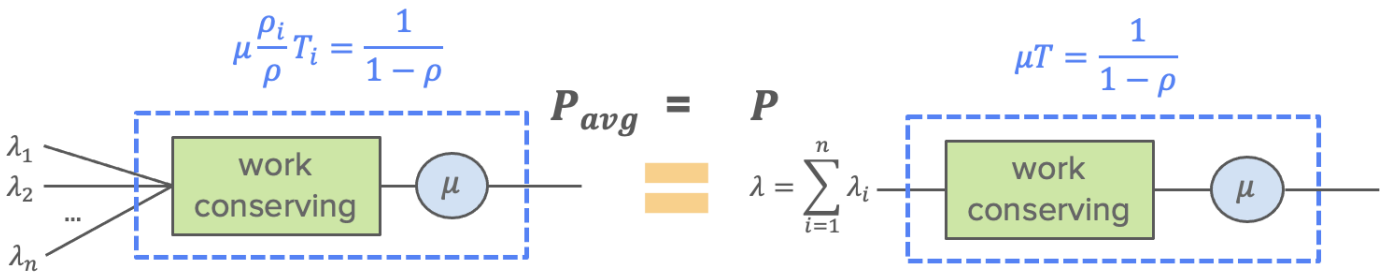


Figure 6.1: An M/G/1 system with multiple flows single hop using any work-conserving queueing discipline. The average power of a system with multiple flows is equivalent to the power of a single-flow system.

6.2 Average Power Optimization

Now we turn our attention to the optimization of the average power. 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], as outlined in Chapter 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 Equation 2.9:

$$\rho^* = \frac{1}{1 + \sqrt{\frac{1+C_b^2}{2}}} \tag{6.9}$$

It 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 Equation 6.9, we have established the following theorem:

Theorem 6.3.

Any M/G/1 system with any work-conserving queueing discipline, where each flow has the same first and second moments of service time, has an average power $\boldsymbol{P_{avg}}$ *that reaches its optimal level when:*

$$\rho^* = \sum_{i=1}^{n} \rho_i = \frac{1}{1 + \sqrt{\frac{1+C_b^2}{2}}} \tag{6.10}$$

Additionally, we have the following corollary for a specific system type:

Corollary 6.3.1.

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

$$\rho^* = \sum_{i=1}^{n} \rho_i = \frac{1}{2} \tag{6.11}$$

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 critically on the effective management of total system utilization.

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

More importantly, the insight that optimal performance is linked to system-wide utilization provides significant **flexibility** for network administrators to pursue additional operational goals, such as **fairness**, which is also a key objective in this thesis. Ensuring **fairness**, which facilitates reasonable resource distribution across multiple flows and users, has also been a focus of much research for a long time, highlighting its importance in network management strategies [52–54].

The capability to simultaneously achieve **optimal performance and fairness** in network operations is a significant benefit of applying this average power definition. As we show in later chapters addressing fairness, this dual achievement could facilitate the design of more effective network congestion control strategies, where both performance efficiency and equitable fairness are jointly considered. Therefore, it is highly advisable for system operators to consider integrating this average power metric into their performance evaluation frameworks. This strategic integration can lead to more informed decision-making and enhanced overall network management, aligning operational practices with the *theoretical insights* provided by the study of queueing theory.

Chapter 7: Extending the Analysis of the Power Metric: the Continuum From FCFS to HOL

In the previous three chapters, we defined three distinct performance metrics, each related to power definitions that incorporate both throughput (represented by the utilization factor) and response time, but combined in different ways to account for multiple flows. We applied power optimization using these definitions to both FCFS and HOL queueing disciplines and analyzed the optimization results.

In this chapter, we introduce another degree of freedom to adjust the power optimization of these metrics beyond FCFS and HOL by studying other queueing disciplines with flow discrimination ranging between these two extremes. We utilize the delay-dependent priority queueing and beta-priority queueing disciplines introduced in Chapter 3 to extend our power optimization analysis across the full spectrum of flow discrimination, ranging from FCFS to HOL. We begin with the base case where $n = 2$ and then proceed to scenarios with an arbitrary number of flows.

7.1 Extension to Full Range in Two Flows n=2

In Chapter 3, two families of queueing disciplines were introduced, and the response times for two flows in an M/M/1 system for each family were presented in Table 3.1. In this chapter,

we will use this table to compute the power values and perform optimization analysis with different power definitions of performance metrics within each family.

The performance metrics we consider here are **individual power** and **sum of individual powers**. We do not include average power because the results derived in Chapter 6 completely answer the optimization solution for all the disciplines we are going to discuss. These results show that average power is maximized when $\rho^* = \frac{1}{2}$ for all work-conserving queueing systems with the same mean service time[1] for n flows in an M/M/1 system. Therefore, when extending the power optimization analysis to the full range of queueing disciplines, we consider only individual power and sum of individual power as optimization metrics.

7.1.1 The Delay-Dependent System

7.1.1.1 Individual Power Optimization

The response time for two flows in the delay-dependent system is given by Equation 3.9, allowing us to calculate the individual power, with $k = 1 - \frac{b_2}{b_1}$ as follows:

$$\begin{aligned} P_1 &= \frac{\rho_1}{\mu T_1} = \frac{\rho_1(1-\rho)(1-k\rho_1)}{1-k\rho} \\ P_2 &= \frac{\rho_2}{\mu T_2} = \rho_2(1-\rho)(1-k\rho_1) \end{aligned} \tag{7.1}$$

Taking the partial derivatives with respect to ρ_1 and ρ_2 respectively, we have:

$$\frac{\partial P_1}{\partial \rho_1} = \frac{[(1-\rho-\rho_1)(1-k\rho_1) - k\rho_1(1-\rho)] \cdot (1-k\rho) + k\rho_1(1-\rho)(1-k\rho_1)}{(1-k\rho)^2} = 0$$

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

[1] Note that for an exponential distribution, the same mean implies the same second moment.

By solving the partial derivative of P_2 with respect to ρ_2, we obtain:

$$\rho_2 = \frac{1 - \rho_1}{2}$$

This indicates that flow 2's power is maximized when it takes half of the remaining utilization after flow 1 has taken its share, a result also discussed in Chapter 4. It holds not only for the two extreme cases, FCFS and HOL, but also for all the other queueing disciplines that lie between them when using a delay-dependent system as the queueing discipline. Consequently, this leads to the following theorem:

Theorem 7.1.

In an M/M/1 system with two flows using a ***delay-dependent*** *queueing discipline, the individual power of the lower priority flow, flow 2, is maximized when ρ_2 is half of the remaining utilization after the higher priority flow, flow 1, has taken its share. The mathematical expression for ρ_2 is:*

$$\rho_2 = \frac{1 - \rho_1}{2} \tag{7.2}$$

This theorem is proved by solving the partial differential of P_2 with respect to ρ_2 above. To continue find the ρ_1 optimizes flow 1's individual power, we solve the equation obtained from the partial derivative of P_1 with respect to ρ_1. We set the numerator to zero, leading to the simplified form:

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

By substituting Equation 7.2, we obtain the cubic function of ρ_1 as:

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

By numerically finding the root within the interval $[0, 1]$ of this cubic function, we obtain ρ_1 that maximizes the individual power of flow 1 corresponding to k, as shown in Figure 7.1. The values of ρ_2, derived from Equation 7.2, and the total utilization ρ are also presented.

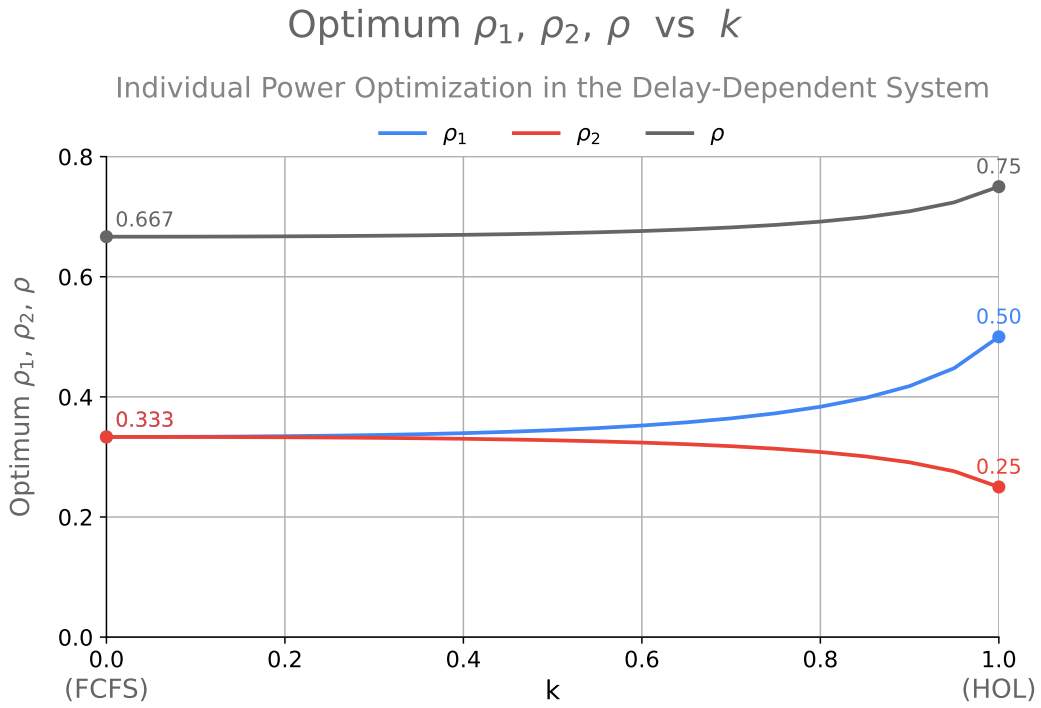


Figure 7.1: Optimum ρ_1, ρ_2, and ρ versus k when optimizing "individual power" for both flow 1 and flow 2 in an M/M/1 system with two flows using the delay-dependent queueing discipline. As k increases, ρ_1, ρ, and the difference between ρ_1 and ρ_2 increase, while ρ_2 decreases.

7.1.1.1.1 Optimization Results for Rho - Figure 7.1

In Figure 7.1, the two extreme cases, FCFS and HOL, are marked on the curve of ρ vs k. The left bound at $k = 0$ (FCFS) has $\rho_1 = \rho_2 = \frac{1}{3}$, and $\rho = \frac{2}{3}$. The right bound at $k = 1$ (HOL) has $\rho_1 = 0.5$, $\rho_2 = 0.25$, and $\rho = 0.75$. The area between these bounds show the range of optimal operating points for ρ_1, ρ_2, and ρ as k changes from 0 to 1, corresponding to the shift from FCFS to HOL.

From the figure, we can observe that when k increases, the ρ_1 value increases while the ρ_2 value decreases. This is because as flow discrimination increases, represented by the increase of k, the impact of lower priority flow is reduced, leading to less waiting time and thus a higher value of ρ_1 for flow 1 to optimize its power. Consequently, there is less remaining utilization in the system left for flow 2. The more flow 1 takes, the less flow 2 can take, resulting in a lower value of ρ_2 as it takes half of the flow 1's remaining utilization.

In addition, the sum of ρ_1 and ρ_2, as well as the difference between ρ_1 and ρ_2, become larger as k increases. This is because both the sum and the difference are functions of ρ_1 that increase as ρ_1 increases. Specifically, the sum of ρ_1 and ρ_2 is given by $\rho = \rho_1 + \rho_2 = \rho_1 + \frac{1-\rho_1}{2} = \frac{1+\rho_1}{2}$. As ρ_1 increases, $\frac{1+\rho_1}{2}$ also increases, indicating that the total utilization of the system grows with an increasing ρ_1. Similarly, the difference between ρ_1 and ρ_2 is given by $\rho_1 - \rho_2 = \rho_1 - \frac{1-\rho_1}{2} = \frac{3\rho_1-1}{2}$. As ρ_1 increases, $\frac{3\rho_1-1}{2}$ also increases, demonstrating that the disparity in utilization between the two flows grows as ρ_1 increases. Therefore, given that ρ_1 increases with k, both the sum and the difference of ρ_1 and ρ_2, which increase with ρ_1, also increase with k.

7.1.1.1.2 Optimization Results for Power - Figure 7.2

Figure 7.2 presents the individual power for flow 1 and flow 2 using the values of ρ_1 and ρ_2 from Figure 7.1 after individual power optimization. The left bound (FCFS with $k = 0$) has the individual power for flow 1 and flow 2 as $P_1 = P_2 = \frac{1}{9} \approx 0.111$, resulting in the sum of individual power $P_{\text{sum}} = \frac{2}{9} \approx 0.222$. The right bound (HOL with $k = 1$) has $P_1 = \frac{1}{4} = 0.25$ and $P_2 = \frac{1}{32} = 0.03125$, leading to $P_{\text{sum}} = \frac{9}{32} = 0.28125$. Similar to the behavior of ρ_1, ρ_2 and ρ, P_1 increases with k while P_2 decreases with k. In addition, the sum and difference of P_1 and P_2 become larger as flow discrimination increases. This can be observed from the curve of P_{sum} increasing as k increases and the gap between P_1 and P_2 widening as well.

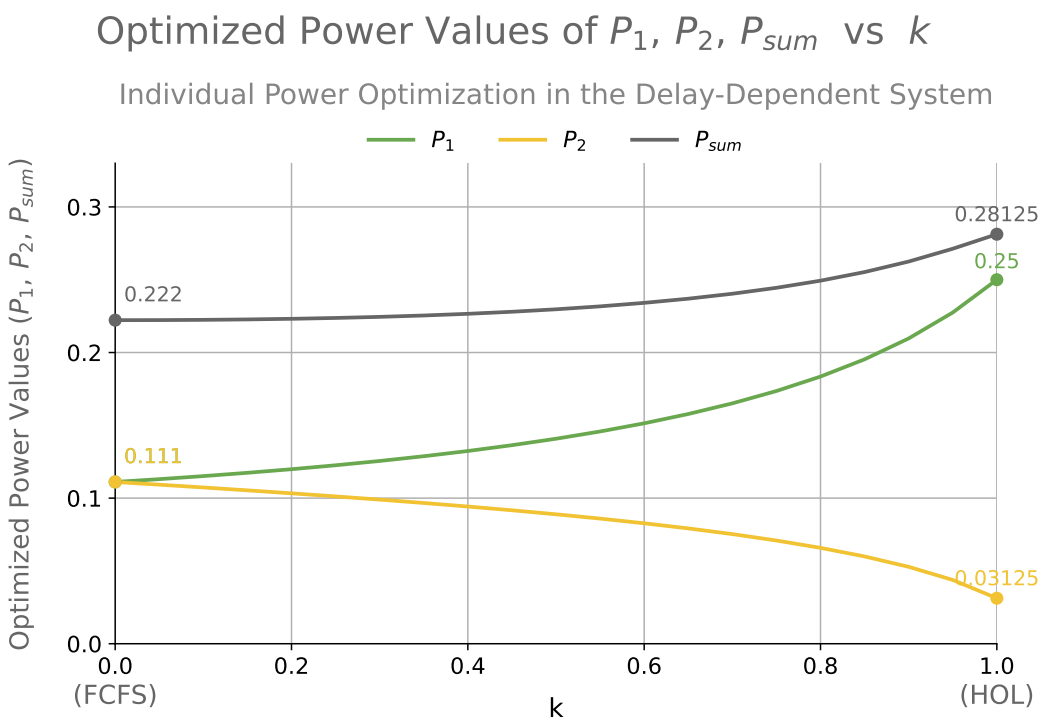


Figure 7.2: Optimized power Values of P_1, P_2, P_{sum} versus k, when optimizing "individual power" for both flow 1 and flow 2 in an M/M/1 system with two flows using the delay-dependent queueing discipline. As k increases, P_1, P_{sum}, and the difference $(P_1 - P_2)$ increase while P_2 decreases.

Note that the power values of P_1 and P_2 shown in Figure 7.2 are not the global maximal power for flow 1 or flow 2. The global maximal individual power value for flow 1 or flow 2 is 0.25, which occurs when their utilization factor for one flow is 0.5 while the other flow has a utilization factor of zero. This global maximal individual power value is independent of k because, with no utilization from other flows, the system can be viewed as a single-flow system, and thus no flow discrimination is needed.

In contrast, the individual power optimization here assumes that the portion of the other flow cannot be modified. Each flow tries to adjust its amount of utilization factor in the system in order to maximize its individual power based on the current scenario, where other flows may also be present with a certain amount of utilization. Such adjustment may trigger other flows to modify their utilization factors as well. This process continues until an equilibrium point is reached. The values derived in Figure 7.1 and Figure 7.2 represent the equilibrium optimal operating points where both flow 1 and flow 2 are optimizing their own power.

7.1.1.2 Sum of Individual Powers Optimization

Now, we change the optimization goal to the sum of individual powers metric. The sum of individual powers in a delay-dependent system is given by:

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

This can be simplified to:

$$P_{\text{sum}} = \frac{\rho(1-k\rho_2)(1-\rho)(1-k\rho_1)}{1-k\rho} \tag{7.3}$$

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

$$\begin{cases} \frac{\partial}{\partial \rho_1} \frac{\rho(1-k\rho_2)(1-\rho)(1-k\rho_1)}{1-k\rho} = 0 \\ \\ \frac{\partial}{\partial \rho_2} \frac{\rho(1-k\rho_2)(1-\rho)(1-k\rho_1)}{1-k\rho} = 0 \end{cases} \tag{7.4}$$

and establish the following theorem:

Theorem 7.2.

In an M/M/1 system using ***delay-dependent*** *queueing disciplines (except for FCFS case where $k = 0$), the* ***sum of individual powers*** *is maximized when*

$$\rho_1 = \rho_2 \tag{7.5}$$

However, $\rho_1 = \rho_2$ does not sufficiently imply that the sum of power is maximal. It is a ***necessary*** *but not sufficient condition for optimizing the sum of individual power.*

Proof

From the partial differentials of Equation 7.4, we have:

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

Since neither $1-k\rho_2$ nor $1-k\rho_1$ can be zero (as ρ_1 and ρ_2 are assumed to be less than one to prevent system overloading)[2], the other terms in each numerator must be zero:

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

$$\begin{cases} [(1-2\rho)(1-k\rho_1)-k\rho(1-\rho)](1-k\rho) \ + \ k\rho(1-\rho)(1-k\rho_1)=0 \\ \\ [(1-2\rho)(1-k\rho_2)-k\rho(1-\rho)](1-k\rho) \ + \ k\rho(1-\rho)(1-k\rho_2)=0 \end{cases} \tag{7.6}$$

Rewriting these equations yields:

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

If $k=0$, which is the FCFS case, substituting $k=0$ into the above equations give $1-2\rho=0$, without indicating the exact value of ρ_1 and ρ_2. This result was discussed in Chapter 5. For other cases where $k \neq 0$, we continue solving the equations and simplifying the right side of the two equations to obtain:

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

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

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

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

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

Leading to

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

and thus

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

which simplifies to:

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

Since $1 - k(\rho_1 + \rho_2)$ cannot be zero under the assumption that the system's total utilization $\rho = \rho_1 + \rho_2$ is less than one, this implies that $k(\rho_1 + \rho_2)$ is less than one, and thus $1 - k(\rho_1 + \rho_2) > 0$. Therefore, for the above equation to be zero, it must be that $\rho_1 - \rho_2 = 0$.

Thus, we obtain:

$$\rho_1 = \rho_2 \quad \text{for } k \neq 0 \tag{7.7}$$

■

Given that $\rho_1 = \rho_2$, we further solve the partial differentials to find the values of ρ_1 and ρ_2 that maximize the sum of individual powers. By substituting ρ_2 with ρ_1 in the first equation from Equation 7.6:

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

we obtain:

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

which simplifies to:

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

This can be written as:

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

The root of this equation that falls within the interval $[0, 1]$ is the value of ρ_1 we are looking for. This root, ρ_1, and ρ_2 (where $\rho_2 = \rho_1$), optimizes the sum of individual power for different values of k. Figure 7.3 illustrates the values of k with their corresponding roots of ρ_1 as found numerically.

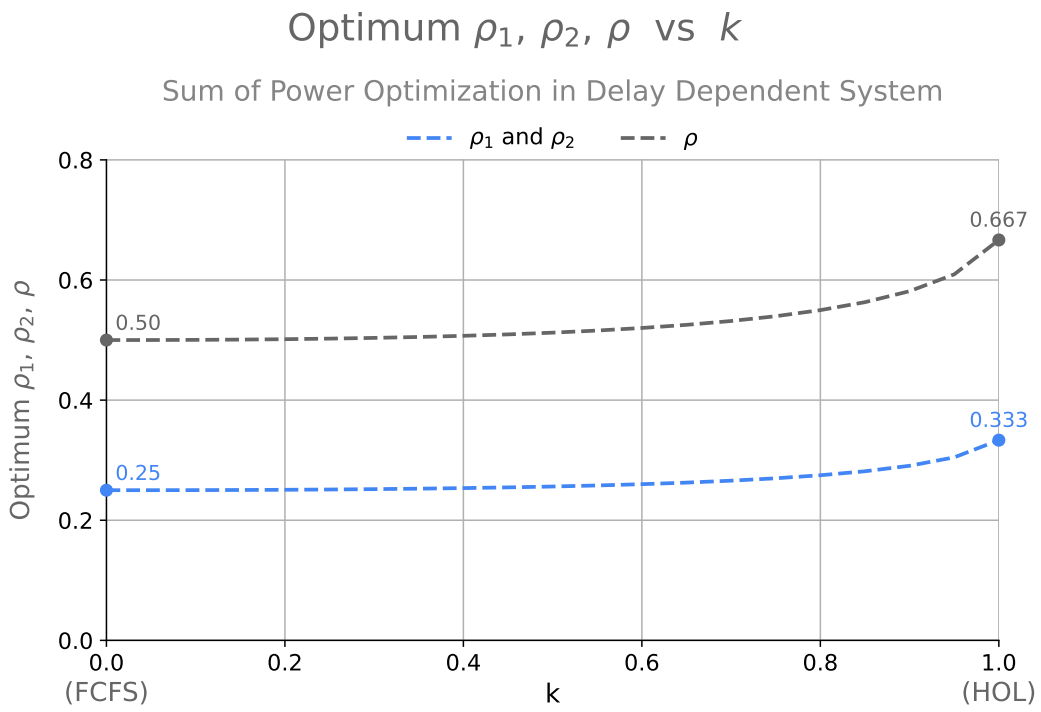


Figure 7.3: Optimum ρ_1, ρ_2, and ρ (where $\rho_2 = \rho_1$) that maximizes the sum of individual powers versus k in an M/M/1 system with two flows using the delay-dependent queueing discipline. At $k = 0$ (FCFS), where $\rho_1 = \rho_2$ are not required for optimal P^*_{sum} as long as $\rho = 0.5$, we explicitly set $\rho_1 = \rho_2 = 0.25$ to align with the requirement $\rho_1 = \rho_2$ for other values of k.

7.1.1.2.1 Optimization Results for Rho - Figure 7.3

In Figure 7.3, the values of k range from $[0, 1]$, corresponding to the two extreme flow discrimination queueing disciplines: FCFS and HOL. For the HOL case (when $k = 1$), as stated in Chapter 5, the sum of power is maximal when $\rho_1 = \rho_2 = \frac{1}{3}$. For the FCFS case (when $k = 0$), the sum of power is maximal as long as $\rho = 0.5$, without the requirement that ρ_1 equals ρ_2. However, to align with other k values that have this requirement, we set $\rho_1 = \rho_2 = 0.25$.

From this figure, we can observe that the optimal ρ_1 value increases slowly and remains almost flat in the first half of the curve. The optimal ρ_1 at $k = 0.5$ is 0.256, which is only about a 2% increase from the optimal ρ_1 of 0.25 at $k = 0$. Not until k reaches 0.8 does ρ_1 start to show a relatively large increase, reaching a value of approximately 0.275, which reflects about a 10% increase.

7.1.1.2.2 Optimization Results for Power - Figure 7.4

Figure 7.4 presents the corresponding power values of P_1, P_2, and P_{sum}, where the sum of individual power is maximal at the operating points shown in Figure 7.3. At $k = 0$ (FCFS), $P_1 = P_2 = 0.125$, and $P_{\text{sum}} = 0.25$. At $k = 1$ (HOL), $P_1 = \frac{2}{9} \approx 0.222, P_2 = \frac{2}{27} \approx 0.074$, and $P_{\text{sum}} = \frac{8}{27} \approx 0.296$.

The trend of these power values is consistent with that observed when optimizing individual power for both flows. As the flow discrimination k increases, P_1 increases and P_2 decreases, leading to a widening difference between them. In addition, since the increase in P_1 is greater than the decrease in P_2, the sum of the power values for P_1 and P_2 also becomes larger. These two observations are evident in the Figure 7.4, where the black curve (P_{sum}) increases with k, and the gap between the green curve (P_1) and the yellow curve

(P_2) expands. Furthermore, given that $\rho_1 = \rho_2$ in this optimization and they both increase with k, the figure indicates that with the same increase in utilization factor, the individual power values for each flow can change in opposite directions and at different rates as k increases. Specifically, while ρ_1 and ρ_2 both increase with k, P_1 increases and P_2 decreases, with the magnitude of the increase in P_1 being larger than the magnitude of the decrease in P_2.

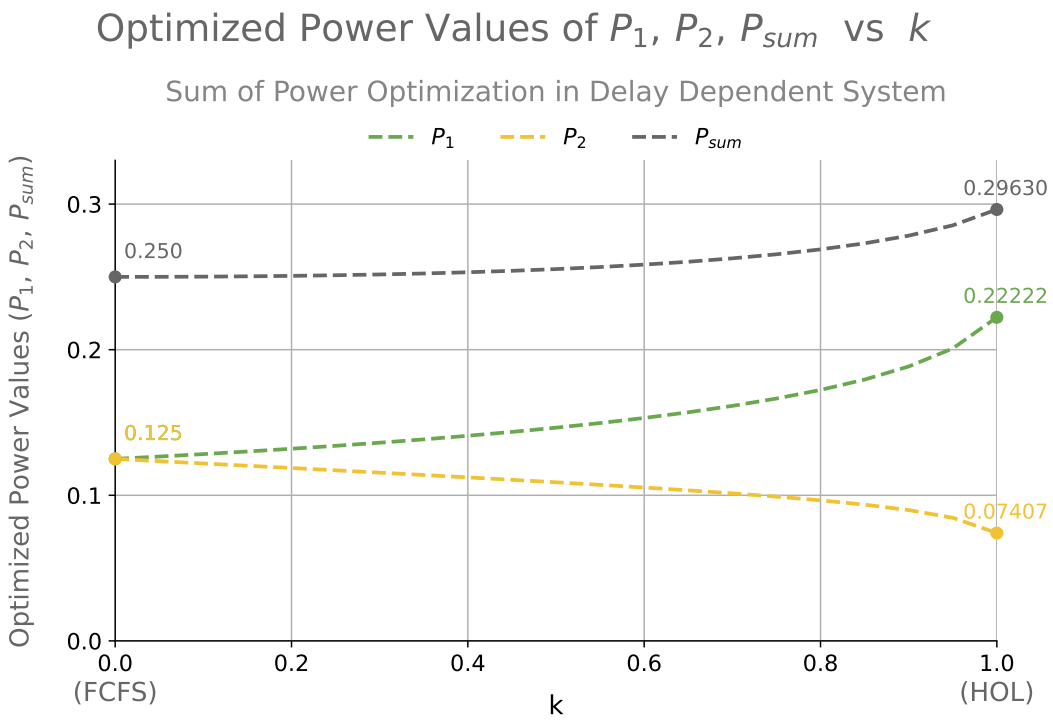


Figure 7.4: The maximal sum of power P^*_{sum} along with the individual powers P_1 and P_2 versus k in an M/M/1 system with two flows using the delay-dependent queueing discipline.

Figure 7.5 compares the different performance metrics by presenting the sum and difference of the individual power values derived from each optimization. Solid lines represent the results from optimizing "individual power", while dashed lines represent the results from optimizing the "sum of individual powers". For the sum of individual powers, represented by black curves, the dashed line is higher than the solid line, indicating higher sum of power values. This is expected, as the dashed line results from optimizing the sum of powers,

making it the maximal sum of powers. The difference in individual powers, denoted as P_{diff}, is computed as $P_1 - P_2$, represented by brown curves. The dashed brown line is below the solid brown line, indicating less variation in individual power values and implying better fairness in terms of individual power. We will discuss the topic of fairness more in the next chapter.

To sum up, using the sum of individual power as the optimization criterion not only achieves the maximal sum of power but also results in better fairness in terms of individual power compared to using individual power as the optimization criterion.

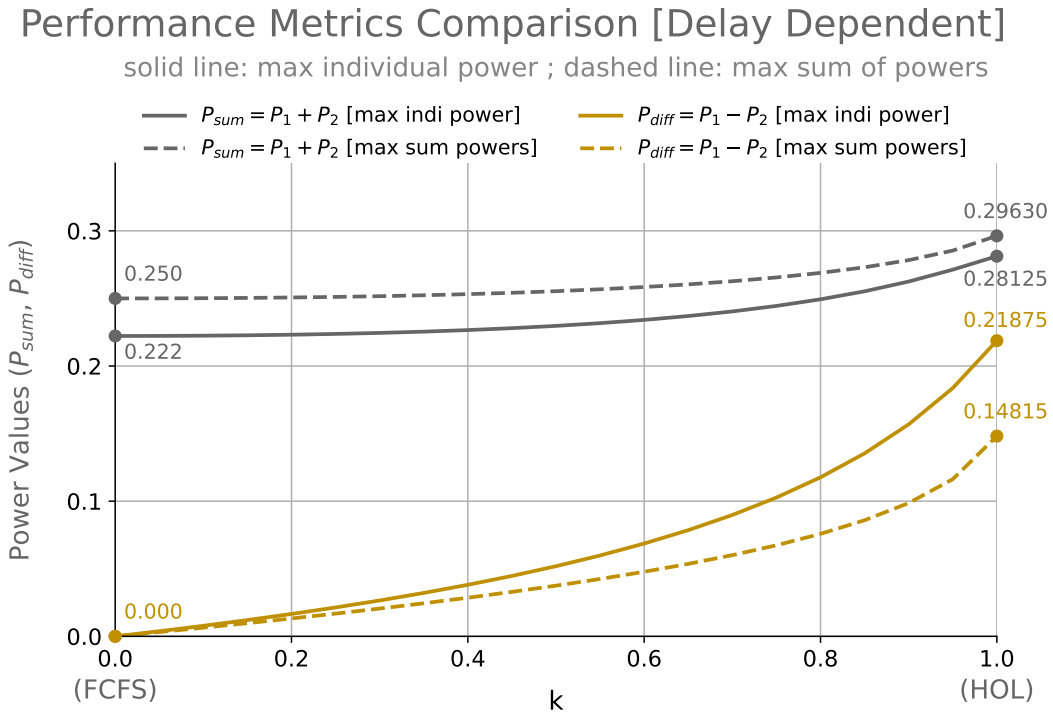


Figure 7.5: The sum and difference of P_1 and P_2 versus k using different performance metrics as the optimization goal. Black curves represent the sum of powers P_{sum}, while brown curves represent the power difference P_{diff}. Solid lines are the results of optimizing "individual power", whereas dashed lines are the results of optimizing "sum of individual powers".

7.1.2 The Beta-Priority System

7.1.2.1 Individual Power Optimization

In the beta-priority system, the power for flow 1 and flow 2 are given:

$$\begin{aligned} P_1 &= \frac{\rho_1}{\mu T_1} = \frac{\rho_1}{\frac{\beta}{(1-\rho_1)} + \frac{1-\beta}{(1-\rho)}} = \frac{\rho_1(1-\rho_1)(1-\rho)}{\beta(1-\rho) + (1-\beta)(1-\rho_1)} = \frac{\rho_1(1-\rho_1)(1-\rho)}{1-\rho_1-\beta\rho_2} \\ P_2 &= \frac{\rho_2}{\mu T_2} = \frac{\rho_2}{\frac{\beta}{(1-\rho_1)(1-\rho)} + \frac{1-\beta}{(1-\rho)}} = \frac{\rho_2(1-\rho_1)(1-\rho)}{\beta + (1-\beta)(1-\rho_1)} = \frac{\rho_2(1-\rho_1)(1-\rho)}{1-(1-\beta)\rho_1} \end{aligned} \tag{7.8}$$

Taking the partial derivative of P_2 with respect to ρ_2 and setting it to zero, we proceed by factoring out the term that is independent of ρ_2 and continuing the computation:

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

Leading to

$$\rho_2 = \frac{1-\rho_1}{2} \tag{7.9}$$

This result is the same as in Equation 7.2 in the delay-dependent system. This equation gives us the following theorem:

Theorem 7.3.

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

$$\rho_2 = \frac{1-\rho_1}{2} \tag{7.10}$$

Now taking the partial derivative of P_1 with respect to ρ_1 and setting it to zero, we get:

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

Plugging Equation 7.10 into the numerator and solving the equation, we have the equilibrium point for ρ_1 that optimizing individual power:

$$\rho_1 = \frac{2-\beta}{2(3-2\beta)} \tag{7.11}$$

Substitute this back to Equation 7.10, we have the equilibrium point for ρ_2:

$$\rho_2 = \frac{4-3\beta}{4(3-2\beta)} \tag{7.12}$$

and the total utilization:

$$\rho = \rho_1 + \rho_2 = \frac{8-5\beta}{4(3-2\beta)} \tag{7.13}$$

7.1.2.1.1 Discussion of Optimization Results for Rho - Figure 7.6

The plot of β ranging from 0 to 1 versus the corresponding ρ_1, ρ_2, and ρ is presented in Figure 7.6. The left bound where $\beta = 0$ represents the FCFS case with minimal flow discrimination, results in equal utilization for flow 1 and flow 2, $\rho_1 = \rho_2 = \frac{1}{3}$, and thus $\rho = \frac{2}{3}$. The right bound where $\beta = 1$ represents the HOL case with maximal flow discrimination, resulting in the maximum ρ among all queueing disciplines within the beta-priority system, with $\rho_1 = 0.5$, $\rho_2 = 0.25$, and $\rho = 0.75$. The trend of curves in the figure resembles that in Figure 7.1, where ρ_1 increases and ρ_2 decreases as the level of flow discrimination increases, represented by the parameter β. Moreover, both the sum of ρ_1 and ρ_2 as well as the difference between ρ_1 and ρ_2 grow with increase of β, this point is the same as in the delay-dependent

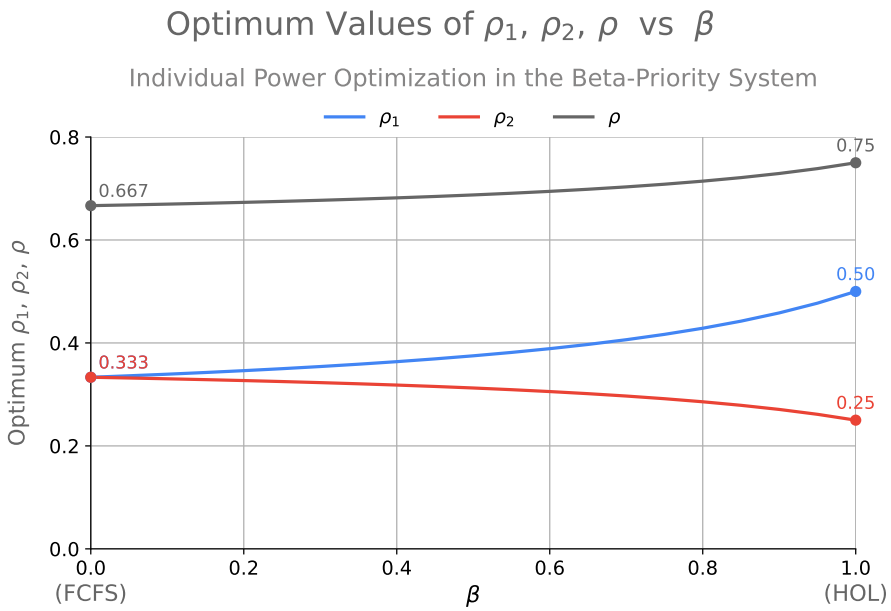


Figure 7.6: Optimal values of ρ_1, ρ_2, and ρ versus β, derived from individual power optimization in an M/M/1 system with two flows using the beta-priority queueing discipline.

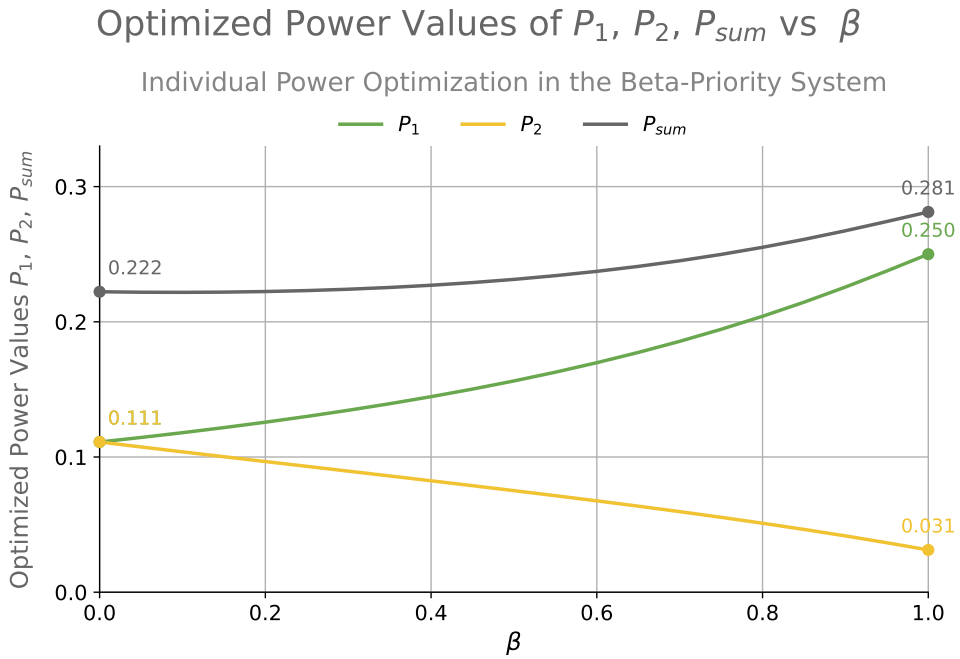


Figure 7.7: Optimized power values of P_1, P_2, and P_{sum} versus β, derived from individual power optimization in an M/M/1 system with two flows using the beta-priority queueing discipline.

system. As the value of β increases, it indicates a higher probability that flow 1 can cut in line ahead of flow 2's packets, thereby reducing the waiting time for flow 1 and leading to a higher level of flow discrimination. Greater flow discrimination results in a higher ρ_1 for optimizing individual power and consequently a lower ρ_2, but a higher sum and difference of ρ_1 and ρ_2.

7.1.2.1.2 Discussion of Optimization Results for Power - Figure 7.7

Given that ρ_1 and ρ_2 are functions of β, we compute the corresponding optimized power values for each flow as follows:

$$\begin{aligned} P_1 &= \frac{(4-3\beta)}{4(3-2\beta)^2} \\ P_2 &= \frac{(4-3\beta)^3}{16(3-2\beta)^2(-\beta^2-\beta+4)} \end{aligned} \tag{7.14}$$

and the sum of optimized individual power is:

$$P_{\text{sum}} = P_1 + P_2 = \frac{(4-3\beta)(4-\beta)(8-5\beta)}{16(3-2\beta)^2(-\beta^2-\beta+4)} \tag{7.15}$$

Substituting $\beta = 0$ into the equations (FCFS case), we have $(P_1, P_2) = \left(\frac{1}{9}, \frac{1}{9}\right)$ and $P_{\text{sum}} = \frac{2}{9}$. Substituting $\beta = 1$ into the equations (HOL case), we have $(P_1, P_2) = \left(\frac{1}{4}, \frac{1}{32}\right)$ and $P_{\text{sum}} = \frac{9}{32}$. The values derived are expected to be the same as those derived in Chapter 4.

Figure 7.7 shows the plot of power values P_1, P_2, and P_{sum} versus β, exhibiting the same trend as observed in Figure 7.2. As β increases, flow 1's individual power P_1, along with the sum and the difference of P_1 and P_2, increases, while the individual power for flow 2, P_2, decreases. This trend is consistent with the behavior of ρ_1, ρ_2, and ρ in Figure 7.6. It is also consistent with the behavior of P_1, P_2, and P_{sum} in the delay-dependent system when optimizing individual power as shown in Figure 7.2, although the rates differ since the

parameters k and β are both variables ranging from 0 to 1, and except at the endpoints 0 and 1, they represent different levels of flow discrimination.

7.1.2.2 Sum of Individual Power Optimization

The sum of individual power in the beta-priority system is given by:

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

This can be simplified to:

$$P_{\text{sum}} = \frac{\rho\,(1-\rho)\,(1-\rho_1)\,[1-(1-\beta)\rho_1-\beta\rho_2]}{(1-\rho_1-\beta\rho_2)\,[1-(1-\beta)\rho_1]} \tag{7.16}$$

Since the coefficients for ρ_1 and ρ_2 in the simplified form of P_{sum} are different, the equilibrium optimal values of ρ_1 and ρ_2 in optimizing individual power may not be the same. This is different from the observation in the delay-dependent system where $\rho_1 = \rho_2$, as specified in Theorem 7.2.

To find the maximum sum of individual power, we establish the following partial differentials:

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

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

7.1.2.2.1 Optimization Results for Rho - Figure 7.8

The β values corresponding to each set of (ρ_1, ρ_2) optimizing the sum of individual power are shown in Figure 7.8. There is no data point for (ρ_1, ρ_2) when $\beta = 0$, since the only constraint in this optimization process is $\rho = \rho_1 + \rho_2 = 0.5$ without specifying the exact values of (ρ_1, ρ_2). We plot the initial point with data at $\beta = 0.01$, leading to $\rho_1 \approx 0.333$, $\rho_2 \approx 0.167$, and $\rho \approx 0.5$.

In Figure 7.8, ρ_1 and ρ_2 are equal only when $\beta = 1$ in the HOL case. For other values of β, the optimization results show $\rho_1 \neq \rho_2$, which differs from the behavior in the delay-dependent system when optimizing the same target performance metric, sum of individual power.

In addition, the trend for ρ_1 and ρ_2 is different from the trend observed when optimizing individual power, where ρ_1 increases and ρ_2 decreases as the level of flow discrimination increases, as shown in Figure 7.6. Here, in contrast, ρ_1 shows only slight changes as β increases from 0 to 1, with the minimum value being about 0.307 and the maximum value being 0.333. Meanwhile, ρ_2 increases with β since flow 2's power is also included in the optimization target.

Moreover, the difference between ρ_1 and ρ_2 decreases and becomes zero when β reaches its maximum value of 1. This can be observed from the curves of ρ_1 and ρ_2, as the gap between them narrows and they converge at $\beta = 1$. This behavior contrasts with what is observed when optimizing the individual power of both flows, where the difference between ρ_1 and ρ_2 increases, as shown in Figure 7.6. However, a consistent observation when changing the optimization metric from individual power to the sum of power is that the total utilization ρ (the sum of ρ_1 and ρ_2) increases as the level of flow discrimination rises.

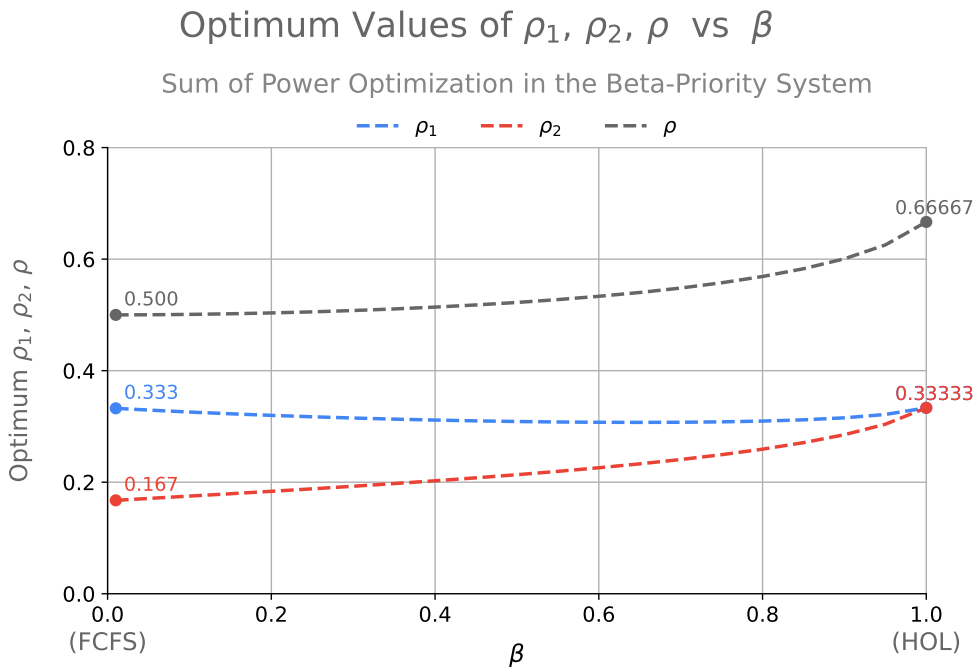


Figure 7.8: Optimum values of ρ_1, ρ_2, and ρ vs β, derived from the sum of individual power optimization in an M/M/1 system with 2 flows using the beta-priority queueing discipline.

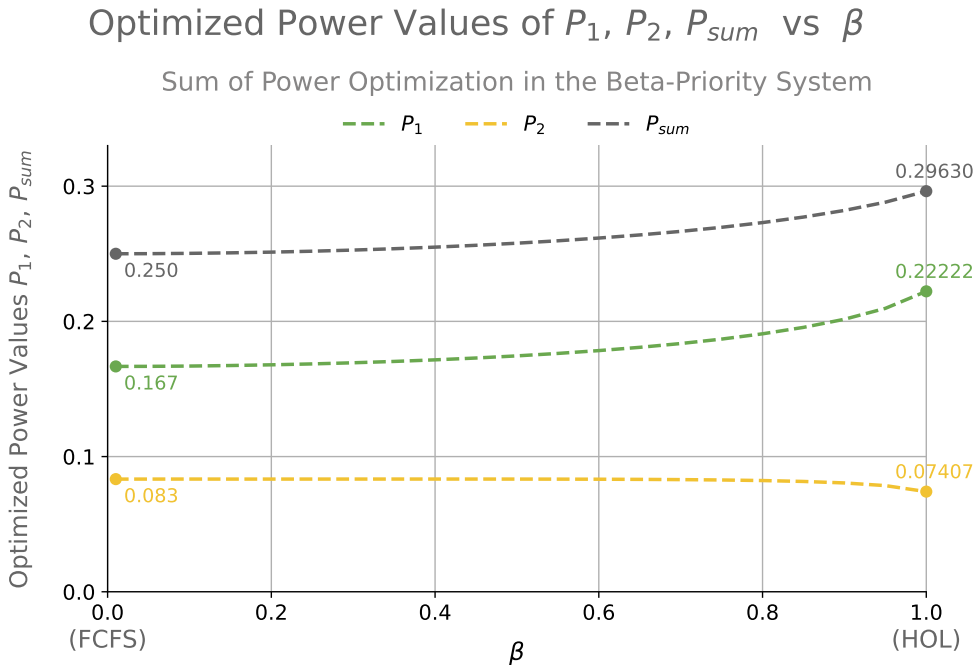


Figure 7.9: Optimized power values of P_1, P_2, and P_{sum} versus β, derived from the sum of individual power optimization in an M/M/1 system with 2 flows using the beta-priority queueing discipline.

7.1.2.2.2 Optimization Results for Power - Figure 7.9

In Figure 7.9, the corresponding maximal sum of individual power, P_{sum}, along with the individual power values for P_1 and P_2 versus β, are presented. For the left bound of the FCFS case where $\beta = 0$, only the sum of individual power with a value of 0.25 is shown in the plot. The exact values for P_1 and P_2 are not marked since there are several combinations for P_1 and P_2 that satisfy the condition $\rho = \rho_1 + \rho_2 = 0.5$ with the sum of power being 0.25. The starting point we use in the curves is when $\beta = 0.01$, resulting in $P_1 \approx 0.166$, $P_2 \approx 0.083$, and $P_{\text{sum}} \approx 0.25$.

In Figure 7.9, the maximum of sum of individual power P_{sum} increases as β increases. The increase is driven by the rise in flow 1's power, while P_2 shows a slight decrease. As depicted by the yellow curve in the figure, P_2 decreases from 0.0833 to 0.074. Conversely, P_1 increases significantly, from 0.166 to 0.222. Consequently, the sum of the powers increases from 0.25 to 0.296.

Even though ρ_1 changes subtly and ρ_2 increases, as shown in Figure 7.8, the individual power for flow 1, P_1, still increases with the increase in β. This is because as flows become more discriminative, the higher priority flow is less affected by the lower priority flow, resulting in reduced waiting and response times. Therefore, this leads to a higher individual power value for flow 1 with the same value of ρ_1. Subsequently, the increase in P_1 also leads to an increase in the sum of power, implying the positive effect of flow discrimination.

7.2 Extension to Full Range from FCFS to HOL - Arbitrary Number of Flows in the Beta-Priority System

Now we extend the investigation from $n = 2$ (two flows) to an arbitrary number of flows. Given the complexity of the response time equations in the delay-dependent system, we focus on the **beta-priority** system, as its equation form is relatively straightforward compared to the recursive form in the delay-dependent system.

7.2.1 Maximizing Individual Power in the Beta-Priority System for Arbitrary n Flows

In this section, we discuss using individual power as an optimization metric. We first establish the analytical results that apply to an arbitrary number n of flows in the beta-priority system. However, given the complexity of the computations, analytical results are limited, and we cannot derive the equilibrium results for an arbitrary number n. Therefore, we adopt numerical methods to find the equilibrium results when each flow optimizes individual power for n greater than 2.

7.2.1.1 Analytical Results

From Theorem 7.2 and Theorem 7.10, we have the result:

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

for maximizing flow 2's individual power with only two flows in both the beta-priority system and the delay-dependent system[3]. This result indicates that the lowest priority flow takes

[3] We will only consider the beta-priority system. Here, we just use the observation from the delay-dependent system

half of the remaining utilization left by the higher priority flow.

Given this interesting finding, it raises the question of whether a similar principle can be applied to systems with more than two flows. Specifically, we are curious if the lowest priority n^{th} flow in a system with an arbitrary number of flows would also take half of the remaining utilization left by the higher priority flows.

7.2.1.1.1 The Lowest Priority n^{th} Flow

To address this question, we extend our analysis to systems with an arbitrary number of flows. By doing so, we aim to determine if the observed pattern holds true in more complex scenarios. This leads us to the following theorem:

Theorem 7.4.

In an M/M/1 system with arbitrary n flows using beta-priority as queueing system, the lowest priority flow has its maximal individual power value when it takes half of the remaining utilization left by the other higher priority flows.

The mathematical expression is:

$$\rho_n^* = \frac{1 - \sum_{i=1}^{n-1} \rho_i}{2} = \frac{1 - \sigma_{n-1}}{2} \tag{7.17}$$

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

Proof

In the beta-priority system, we have the mean response time for each flow i being average of the mean response time in HOL and FCFS with the weight β. The mean response time is

represented by:

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

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

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

The corresponding individual power is

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

Taking the partial derivative of this function with respect to ρ_n and taking the factor that is not related to ρ_n out of the differential equation as constant:

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

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

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

Leading to

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

Since $\sigma_{n-1} = \sum_{j=1}^{n-1} \rho_j$, this theorem shows that ρ_n^*, which optimizes the individual power P_n, is achieved when ρ_n^* takes half of the remaining utilization after accounting for the amount occupied by all higher priority flows. ■

Since HOL and FCFS are both members (and in fact, the extremes) of the beta-priority system, clearly HOL and FCFS yield the same result for the n^{th} flow when maximizing individual power. Specifically for HOL, substituting i with n into Equation 4.20:

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

yields the same result for the n^{th} flow as in Equation 7.17.
Similarly, for FCFS, substituting i with n into Equation 4.4:

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

also results in the same outcome as Equation 7.17. Given this consistency between HOL and FCFS, it is not surprising that the beta-priority system, where the mean response time is the average of response times from HOL and FCFS weighted by β, also produce the same optimization result for the n^{th} flow. To repeat, we have proven from the above theorem that the optimal ρ_n^* for the lowest priority flow to maximize its individual power P_n in any beta-priority system is half of the remaining utilization left by other flows.

Since the n^{th} flow can utilize the individual power optimization results from HOL and FCFS, we might wonder if these results can be combined to derive the optimal outcomes for all other priority groups such as $n-1, \ldots, 2, 1$ in the beta-priority system, represented by the value of β. However, *this is not the case*. To illustrate this point, we now turn our

attention to the second lowest priority flow $n-1$.

7.2.1.1.2 The Second Lowest Priority $(n-1)^{\text{th}}$ Flow

To understand the behavior of the second lowest priority flow, we derive the scenario in which its individual power maximal is optimal. The following theorem presents the optimal ρ^*_{n-1} for the $(n-1)^{\text{th}}$ flow in an M/M/1 system with n flows using a beta-priority queueing system.

Theorem 7.5.

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

$$\rho^*_{n-1} = \frac{1-\sigma_{n-2}}{2} \cdot \frac{\beta + 2(1-\beta)(1-\sigma_{n-2})}{\beta + 3(1-\beta)(1-\sigma_{n-2})} \tag{7.18}$$

Proof

The mean response time for the $(n-1)^{\text{th}}$ flow is given by:

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

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

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

Taking the partial derivative with respect to ρ_{n-1} and factoring out the factor not related to ρ_{n-1}:

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

Setting the partial differential equation to zero and solving it:

$$\begin{aligned}[\beta\cdot(1-\sigma_n)+(1-\beta)\cdot(1-\sigma_{n-1})(1-\sigma_{n-2})]\cdot[(1-\sigma_{n-1}-\rho_{n-1})(1-\sigma_n)-\rho_{n-1}(1-\sigma_{n-1})]\\ -\rho_{n-1}(1-\sigma_{n-1})(1-\sigma_n)\ [-\beta-(1-\beta)(1-\sigma_{n-2})]=0\end{aligned} \tag{7.19}$$

With Theorem 7.4, we have[4]:

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

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

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

Therefore,

$$1-\sigma_n = \frac{1-\sigma_{n-1}}{2} \tag{7.20}$$

[4] Note that the scenario we consider is one in which each flow optimizes its own individual power. Therefore, we apply the individual power optimization result for flow n from Theorem 7.4

Substituting Equation 7.20 into Equation 7.19:

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

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

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

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

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

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

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

This can be rewritten as:

$$[\beta + 2(1-\beta)(1-\sigma_{n-2})] \cdot [(1-\sigma_{n-2}) - 4\rho_{n-1}] = 2\rho_{n-1}\ [-\beta - (1-\beta)(1-\sigma_{n-2})]$$

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

$$[\beta + 2(1-\beta)(1-\sigma_{n-2})] \cdot [(1-\sigma_{n-2})] = 2\rho_{n-1}\ [-\beta - (1-\beta)(1-\sigma_{n-2})] \\ +4\rho_{n-1}\ [\beta + 2(1-\beta)(1-\sigma_{n-2})]$$

This can be expressed as

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

Thus, we have:

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

This completes the proof of the theorem.

■

Equation 7.18 agrees with the optimal results for maximizing individual power derived in Chapter 4. Substituting $\beta = 1$ in Equation 7.18, we obtain:

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

This result matches the HOL outcome when substituting $i = n-1$ into Equation 4.20. Additionally, substituting $\beta = 0$ in Equation 7.18, we derive:

$$\rho^*_{n-1} = \frac{1-\sigma_{n-2}}{3}$$

This is consistent with the optimal individual power result of FCFS. By using Equation 4.11,

we have:

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

Thus,

$$\rho^*_{n-1} = \frac{1 - \sigma_{n-2}}{3} = \frac{\frac{3}{n+1}}{3} = \frac{1}{n+1}$$

Based on the above derivations and results, we can state the following corollary:

Corollary 7.5.1.

The optimal ρ^*_{n-1} will be bounded by the optimal result of FCFS, namely, $\frac{1-\sigma_{n-2}}{3}$ and the optimal result of HOL, namely, $\frac{1-\sigma_{n-2}}{2}$. Mathematically, this is expressed as:

$$\frac{1 - \sigma_{n-2}}{3} \leq \rho^*_{n-1} \leq \frac{1 - \sigma_{n-2}}{2} \tag{7.21}$$

Proof

From Theorem 7.5, we have the optimal ρ^*_{n-1} expressed as in Equation 7.18:

$$\rho^*_{n-1} = \frac{1 - \sigma_{n-2}}{2} \cdot \frac{\beta + 2(1-\beta)(1-\sigma_{n-2})}{\beta + 3(1-\beta)(1-\sigma_{n-2})}$$

Substituting $\beta = 1$ to get the HOL result:

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

Substituting $\beta = 0$ to get the FCFS result:

$$\rho^*_{n-1} = \frac{1 - \sigma_{n-2}}{3}$$

To show the optimal ρ^*_{n-1} is bounded by the FCFS and HOL results, that is:

$$\frac{1-\sigma_{n-2}}{3} \leq \rho^*_{n-1} = \frac{1-\sigma_{n-2}}{2} \cdot \frac{\beta + 2(1-\beta)(1-\sigma_{n-2})}{\beta + 3(1-\beta)(1-\sigma_{n-2})} \leq \frac{1-\sigma_{n-2}}{2}$$

we only need to show the following, given that $1-\sigma_{n-2} > 0$:

$$\frac{2}{3} \leq \frac{\beta + 2(1-\beta)(1-\sigma_{n-2})}{\beta + 3(1-\beta)(1-\sigma_{n-2})} \leq 1$$

For the right part, it is proved since clearly

$$\beta + 2(1-\beta)(1-\sigma_{n-2}) \ \leq \beta + 3(1-\beta)(1-\sigma_{n-2})$$

For the left part, it is also true since clearly:

$$2 \cdot (\beta + 3(1-\beta)(1-\sigma_{n-2})) \leq 3 \cdot (\beta + 2(1-\beta)(1-\sigma_{n-2}))$$

as $2\beta \leq 3\beta$ for $0 \leq \beta \leq 1$.

Therefore, we have shown that:

$$\frac{2}{3} \leq \frac{\beta + 2(1-\beta)(1-\sigma_{n-2})}{\beta + 3(1-\beta)(1-\sigma_{n-2})} \leq 1$$

Thus,

$$\frac{1-\sigma_{n-2}}{3} \leq \rho^*_{n-1} \leq \frac{1-\sigma_{n-2}}{2}$$

This completes the proof. ■

In conclusion, Equation 7.18 effectively encapsulates the optimal ρ^*_{n-1} results, confirming that they are bounded by the values derived for FCFS and HOL, thereby validating the consistency of our approach.

For subsequent flows, such as the $(n-2)^{th}$ flow, given the complexity of the form of Equation 7.18, solving for the $(n-2)^{th}$ optimal condition is not trivial. Therefore, we resort to numerical methods to find the equilibrium point where each flow optimizes its individual power.

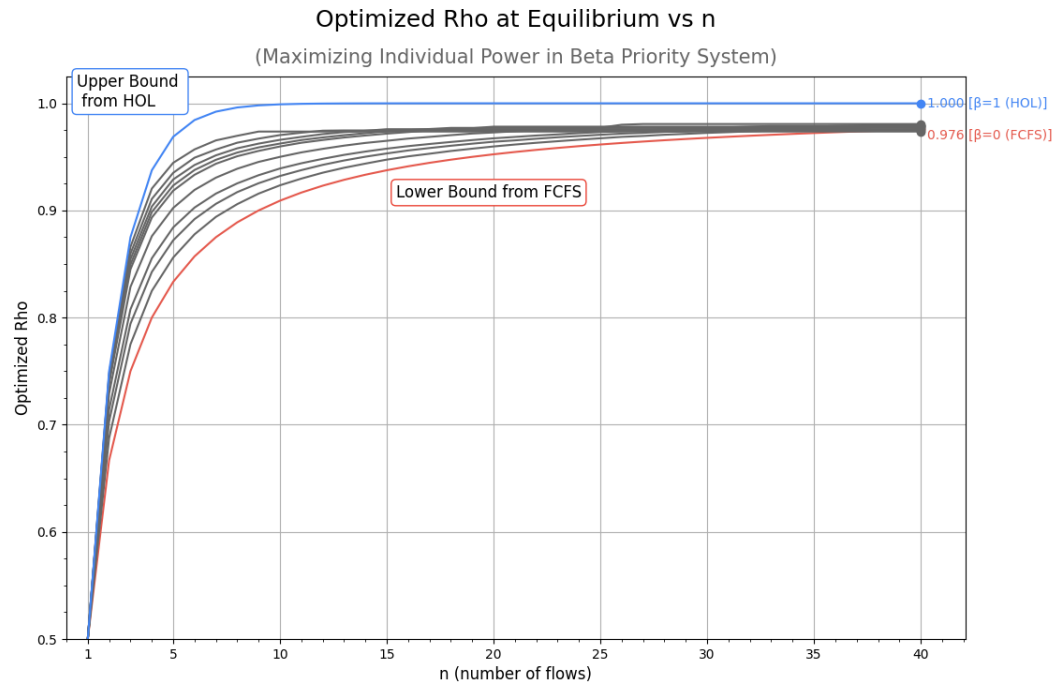
7.2.1.2 Numerical Results

The numerical results are based on the following algorithm. For each iteration, we update each flow's utilization based on their priority order. Specifically, we proceed from $i = 1, 2, \ldots, n$ and optimize the individual power of the i^{th} flow in this sequence within a single iteration. When optimizing the individual power of the i^{th} flow, we assume the utilizations of the other flows are fixed and update the value of ρ_i to maximize the individual power of the i^{th} flow within the feasible region where $0 \leq \rho_i < 1$ and $0 \leq \rho = \sum_{j=1}^{n} \rho_j < 1$.

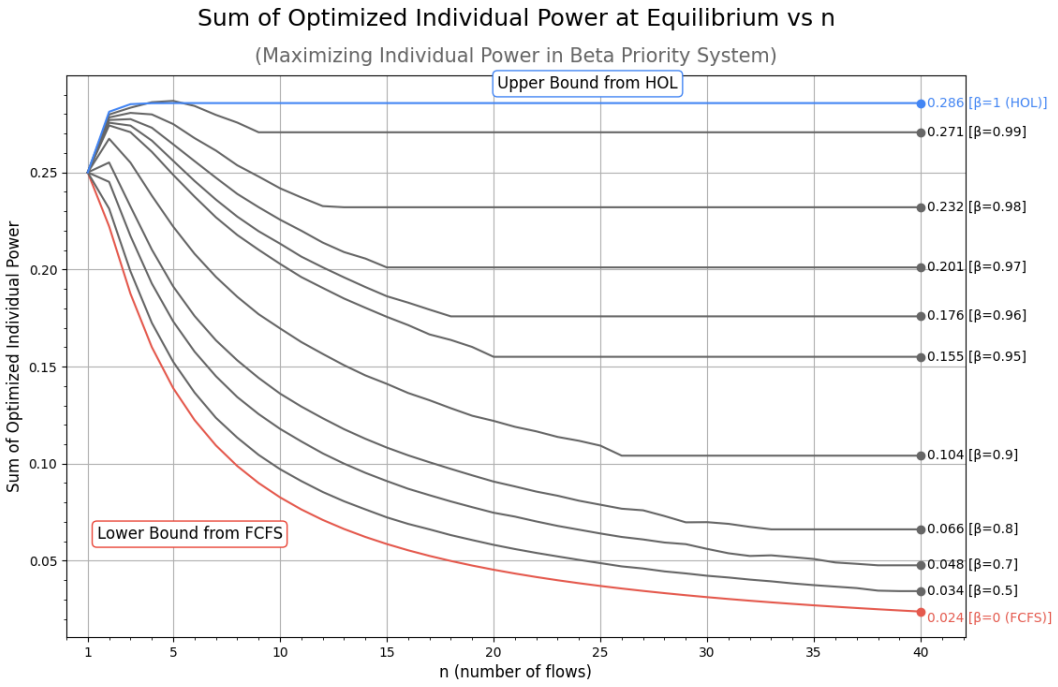
We repeat the iteration up to a maximum of 300 times. After each iteration, we check if each value in the set $(\rho_1, \rho_2, \ldots, \rho_n)$ has converged. Convergence is considered achieved if the difference between each ρ_i value in the current set $(\rho_1, \rho_2, \ldots, \rho_n)$ and the corresponding ρ_i value in the set from the previous iteration is smaller than epsilon, which is set to 10^{-8}. If this condition is met, we stop the iteration. We test for $n = 1, 2, \ldots, 40$ with various β values from 0 to 1 with a step size of 0.05. Additionally, we use a finer resolution for β in the range of 0.95 to 1, with a step size of 0.01.

Figure 7.10 presents the numerical results[5]. FCFS and HOL serve as the upper and lower bounds, respectively, for the optimized ρ^* and the sum of optimized individual power. For other queueing disciplines with flow discrimination falling between these two, their results lie between those of HOL and FCFS. (In Figure 7.10b, the results for $\beta = 0.99$ at $n = 3, 4, 5$ are insignificantly larger than HOL at n=3-5, which may be due to numerical error.)

[5] The results for $0 < \beta < 0.5$ are omitted from the figure, as they fall within a small region, making them difficult to annotate.



(a) Optimized ρ^* at Equilibrium vs n



(b) Sum of Optimized Individual Power at Equilibrium vs n

Figure 7.10: Equilibrium Point in Maximizing Individual Power for Various β in the Beta-Priority System.

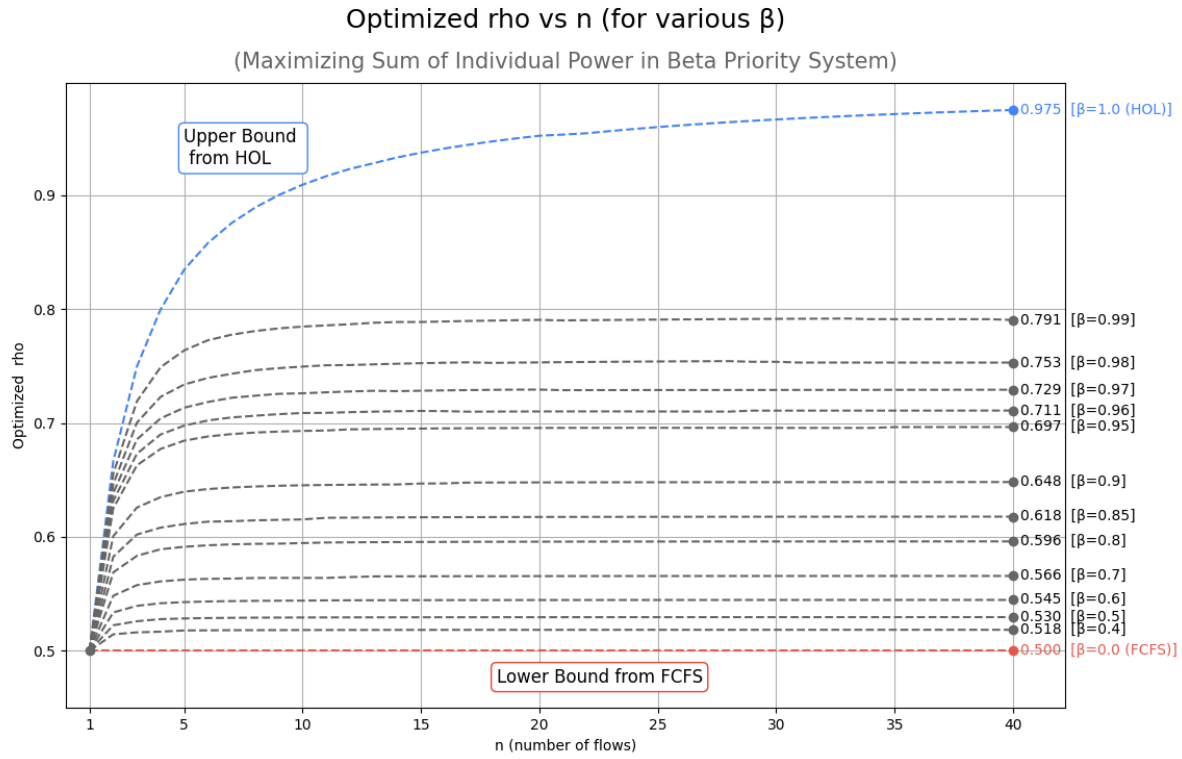
7.2.2 Maximizing Sum of Individual Power in the Beta-Priority System for Arbitrary n Flows

Given the complexity of computing the **sum of individual powers** in the beta system, even for two flows, solving the n differential equations becomes infeasible when n is arbitrary or greater than 2. Therefore, we resort to numerical methods to explore the sum of power optimization for n greater than 2.

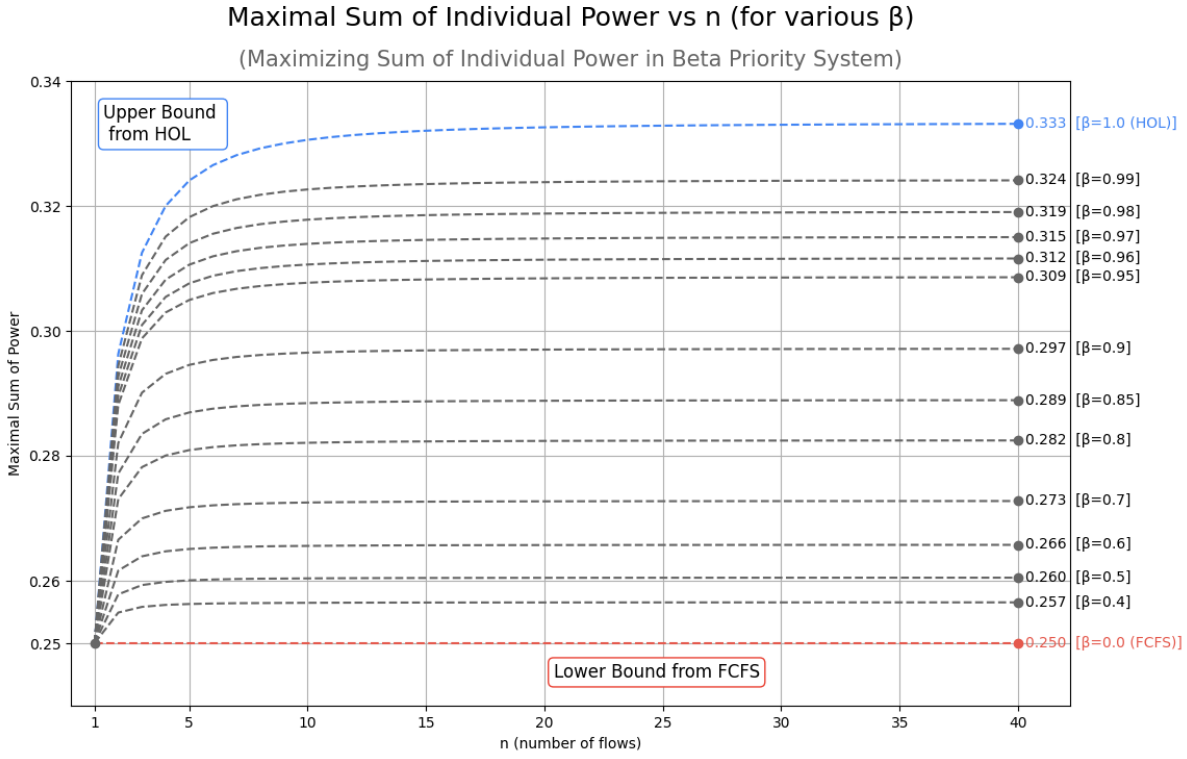
We first compute the response time and power value from Equation 3.10 with the given n and β. Then, we form the objective function to numerically find the set of ρ_i that maximizes the sum of power. Our goal is to determine how the maximal sum of individual power changes with n and to identify its limiting value. To achieve this, we test different values of n from 1 to 40 to see if it converges. To better understand the beta system as a whole, we also examine various values of β from 0 to 1 with a step size of 0.05. Additionally, we use a finer resolution for β in the range of 0.95 to 1, with a step size of 0.01.

The optimization results for maximizing the sum of individual power are presented in Figure 7.11, with each curve representing a queueing discipline characterized by β. Within each curve, n ranges from 1 to 40, and the corresponding optimization results for each value of n are shown. Figure 7.11a shows the optimization result of system utilization, which is the sum of the utilizations of each flow that achieves the maximal sum of individual power. Figure 7.11b shows the maximal sum of individual power for different values of n.

In Figures 7.11a and 7.11b, the curves are bounded by the upper bound (HOL) and the lower bound (FCFS). As stated in Chapter 5, the optimization results for other queueing disciplines, as long as they are work-conservative, fall within the region bounded by HOL and



(a) ρ^* vs n at the maximal sum of individual power



(b) P^*_{sum} vs n

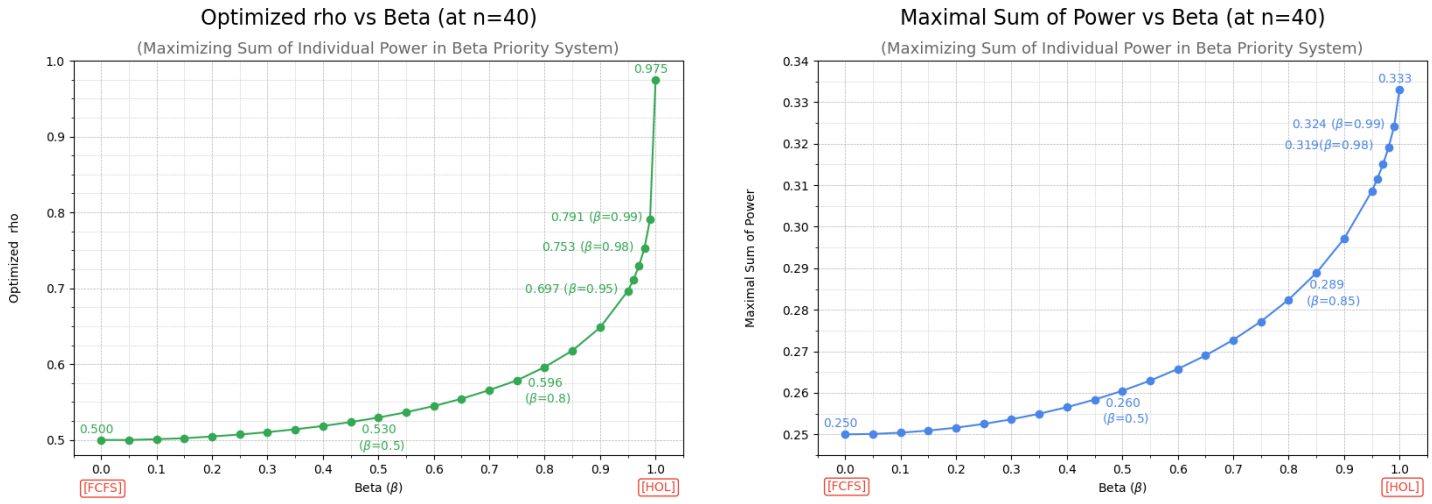
Figure 7.11: Maximizing sum of individual power for various β in the beta-priority system.

FCFS. Here, we use beta-priority as an example to verify this point, further demonstrating that various queueing disciplines characterized by the β value have results that fall within this region, as shown in Figure 7.11. The curves are ordered by the value of β. As we move from the lower bound curve (corresponding to $\beta = 0$, FCFS) to the upper bound curve (corresponding to $\beta = 1$, HOL), we observe that the β value corresponding to each curve increases. This is no surprise. As β increases, the level of flow discrimination increases, leading higher priority flows to be less affected by lower priority flows. Thus, the individual power contributions add up, resulting in an overall growing trend.

In Figure 7.11a, there is a noticeable gap between the curve for $\beta = 0.95$ and the upper bound curve of $\beta = 1$. When $\beta = 1$, the optimal system utilization approaches 1, while for the slightly smaller β value of 0.95, the optimal system utilization remains around 0.7. To better understand the behavior in this range, we added a finer resolution from 0.95 to 1 with a step size of 0.01. However, even with this finer resolution, the gap between $\beta = 0.99$ and $\beta = 1$ becomes smaller but remains noticeable, with the optimal ρ^* at $\beta = 0.99$ being 0.79. Despite this, the maximal sum of individual power at $\beta = 0.99$ is not far from the HOL's optimal point. $\beta = 0.99$ represents a system where 1% of the time behaves like HOL and 99% behaves like FCFS. This suggests that a slightly head-of-line system can prevent system overloading while maintaining near-optimal performance. In other words, a slight head-of-line system, where $\beta \approx 1$, allows us to sacrifice a little performance in exchange for preserving system utilization for other uses.

Convergence

Given that each curve, except for the HOL curve ($\beta = 1$), nearly converges when n reaches 40, we take the value at $n = 40$ as the convergent value for each queueing discipline represented by $\beta$and create a plot against β, as presented in Figure 7.12.



(a) β vs ρ at maximal sum of power

(b) β vs maximal sum of power

Figure 7.12: β vs maximal sum of power and ρ in maximizing sum of power for n=40. The data are marked at β values ranging from 0 to 1 with a step size of 0.05, with a finer resolution of 0.01 close to 1.

Figure 7.12a shows the convergent ρ^* value at the maximal sum of individual power, and Figure 7.12b shows the corresponding optimal power, P^*_{sum}, for each β. These figures illustrate how the optimal sum of power transitions from FCFS to HOL in the limit. Both curves exhibit exponential growth with β, starting with a small increase and then accelerating rapidly as β surpasses a certain point.

In Figure 7.12a, ρ^* begins to accelerate rapidly at around $\beta = 0.95$. At the midpoint of β, the corresponding ρ^* is about 0.53, indicating only about a 6% increase whereas the maximum growth to HOL is 95% $\left(\left(\frac{0.975}{0.5} - 1\right) \times 100\right)$. For the midpoint of ρ^* between FCFS and HOL, where ρ^* is $(1 + 0.5)/2 = 0.75$, the corresponding β is approximately 0.98, which is close to 1 (HOL).

In Figure 7.12b, the curve also exhibits exponential growth but at a slower rate compared to the curve in Figure 7.12a. At the midpoint of β, the corresponding P^*_{sum} is 0.26, represent-

ing about a 4% increase $\left(\left(\frac{0.53}{0.5} - 1\right) \times 100\right)$, whereas the maximum growth to HOL is 32% $\left(\left(\frac{0.33}{0.25} - 1\right) \times 100\right)$. The β that reaches the midpoint of the maximal sum of individual power between FCFS and HOL, which is $P^*_{\text{sum}} = 0.29$, occurs at $\beta = 0.85$.

7.3 Constraint on (ρ_1, ρ_2) in a two-flow system

In the previous sections, we focused on determining the optimal utilization factors ρ_i for each flow $i = 1, \ldots, n$ that maximize various power performance metrics, including "individual power" and the "sum of individual power" of all flows. However, if the utilization factor values (ρ_i for $i = 1, \ldots, n$) are fixed, a different question arises: what queueing discipline leads to the highest power performance?

In the following discussion, we use a M/M/1 system with two flows as an example and adopt the "sum of individual power" P_{sum} as our performance metric. Our goal is to find the optimal queueing parameter k (where $k = 1 - \frac{b2}{b1}$) in the delay-dependent system and the optimal priority parameter β in the beta-priority system that maximizes P_{sum}.

7.3.1 The Delay-Dependent System

The sum of individual power in the delay-dependent system is given by Equation 7.3:

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

where $k = 1 - \frac{b2}{b1}$

For this system, we have the following:

Theorem 7.6.

In an M/M/1 system with two flows employing the ***delay-dependent*** *queueing discipline, given specific utilization factors* (ρ_1, ρ_2) *for the two flows where* $0 < \rho_1 < 1$ *and* $0 < \rho_2 < 1$*, the sum of individual powers* P_{sum} ***monotonically increases*** *as the priority parameter* k *increases. Moreover, this sum of powers is maximized when* $k = 1$ *(i.e., the HOL case).*

Proof

To demonstrate that P_{sum} increases as k increases for $k \in [0, 1]$ and is maximal at $k = 1$ with fixed (ρ_1, ρ_2) in the delay-dependent system, we show that P_{sum} is monotonically increasing within this interval. We establish this by demonstrating the following two points:

1. **Positive Derivative for $k \in (0, 1]$**: The derivative of P_{sum} with respect to k is positive for all $k \in (0, 1]$. This indicates that the function is increasing within this open interval.
2. **Comparison with Endpoints at $k = 0$**: $P_{\text{sum}}(k)$ is larger than $P_{\text{sum}}(0)$ for any $k \in (0, 1]$. In other words, P_{sum} for any $k \in (0, 1]$ is larger than P_{sum} for $k = 0$.

Together, these two points prove that P_{sum} is monotonically increasing for $k \in [0, 1]$ and therefore attains its maximum at $k = 1$.

We now show the first point: **Positive Derivative for $k \in (0, 1]$**.

The derivative of P_{sum} with respect to k is:

$$\begin{aligned}
\frac{dP_{\text{sum}}}{dk} &= \rho\,(1-\rho)\,\frac{d\,\frac{(1-k\rho_1)(1-k\rho_2)}{1-k\rho}}{dk} \\
&= \rho\,(1-\rho)\,\frac{(1-k\rho)\,[-\rho_1(1-k\rho_2)-\rho_2(1-k\rho_1)]+\rho\,(1-k\rho_1)\,(1-k\rho_2)}{(1-k\rho)^2} \\
&= \rho\,(1-\rho)\,\frac{(1-k\rho)\,(2k\rho_1\rho_2-\rho)+\rho\,(1-k\rho+k^2\rho_1\rho_2)}{(1-k\rho)^2} \\
&= \rho\,(1-\rho)\,\frac{(1-k\rho)\,(2k\rho_1\rho_2)+\rho\,(k^2\rho_1\rho_2)}{(1-k\rho)^2} \\
&= \rho\,(1-\rho)\,\frac{k\rho_1\rho_2\,(2-2k\rho+k\rho)}{(1-k\rho)^2}
\end{aligned}$$

Simplifying further:

$$\frac{dP_{\text{sum}}}{dk} = \rho\,(1-\rho)\,\frac{k\,\rho_1\,\rho_2\,(2-k\rho)}{(1-k\rho)^2} \tag{7.22}$$

Given that $0 < \rho_1 + \rho_2 = \rho < 1$, we have $(2-k\rho) > 1$ and $0 < (1-\rho) < 1$ for $k \in (0,1]$. In addition, ρ_1, ρ_2, ρ are all within the interval (0, 1). The term k and the square term $(1-k\rho)^2$ are positive. Since all terms in the numerator and denominator of the derivative expression are positive for $k \in (0,1]$, we can conclude that:

$$\frac{dP_{\text{sum}}}{dk} > 0 \quad \text{for } k \in (0,1]$$

This shows that P_{sum} is monotonically increasing for $k \in (0,1]$.

We now show the second point: **Comparison with Endpoints at $k = 0$**.

P_{sum} at $k = 0$ is:

$$P_{\text{sum}}\Big|_{k=0} = \frac{\rho\;(1-k\rho_2)\;(1-\rho)\;(1-k\rho_1)}{1-k\rho}\Big|_{k=0} = \rho(1-\rho)$$

To show P_{sum} for $k \in (0, 1]$ is larger than P_{sum} at $k = 0$, consider:

$$\frac{\rho\ (1 - k\rho_2)\ (1 - \rho)\ (1 - k\rho_1)}{1 - k\rho}\bigg|_{k\in(0,1]} > \frac{\rho\ (1 - k\rho_2)\ (1 - \rho)\ (1 - k\rho_1)}{1 - k\rho}\bigg|_{k=0} = \rho(1 - \rho)$$

We cancel the common positive term $\rho(1 - \rho)$, yielding:

$$\frac{(1 - k\rho_2) \cdot (1 - k\rho_1)}{1 - k\rho}\bigg|_{k\in(0,1]} > 1$$

This simplifies to:

$$(1 - k\rho_2) \cdot (1 - k\rho_1) > 1 - k\rho$$

Expanding the left-hand side:

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

This simplifies to:

$$k^2\rho_1\rho_2 > 0$$

Since k, ρ_1, ρ_2 are all positive, the term $k^2\rho_1\rho_2$ is positive, confirming that this inequality holds. Therefore, we have proven that P_{sum} for $k \in (0, 1]$ is indeed larger than P_{sum} at $k = 0$.

Combining these two points, we conclude that P_{sum} is monotonically increasing in $k \in [0, 1]$. Hence, the maximum is at the upper boundary of the interval, which is $k = 1$. Therefore, we have proved that P_{sum} increases as k increases and is maximal when $k = 1$ for a given (ρ_1, ρ_2).

■

7.3.2 The Beta-Priority System

The sum of individual power in the beta-priority system is given by Equation 7.16:

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

where β represent the probability that a packet from flow 1 can cut in line before flow 2 packets. For this system, a result similar to Theorem 7.6 is presented as follows:

Theorem 7.7.

In an M/M/1 system with two flows using the ***beta-priority*** *as queueing discipline, given fixed utilization factors (ρ_1, ρ_2) for the two flows where $0 < \rho_1 < 1$ and $0 < \rho_2 < 1$, the sum of individual powers P_{sum} is* ***monotonically increasing*** *for the queueing parameter $\beta \in [0, 1]$. This sum of powers P_{sum} is maximized at $\beta = 1$, i.e. the HOL case.*

Proof

The proof follows a similar approach to that used for the delay-dependent system. We demonstrate that P_{sum} increases as β increases and is maximal when $\beta = 1$ for fixed (ρ_1, ρ_2) in the beta-priority system. First, we show that the derivative with respect to β is positive for $\beta \in (0, 1]$. Then, we show that $P_{\text{sum}} \in (0, 1]$ is larger than P_{sum} at $\beta = 0$.

The derivative of P_{sum} with respect to β is:

$$\begin{aligned}
\frac{dP_{\text{sum}}}{d\beta} &= \rho\,(1-\rho)\,(1-\rho_1)\,\frac{d\left[\frac{1-(1-\beta)\rho_1-\beta\rho_2}{(1-\rho_1-\beta\rho_2)[1-(1-\beta)\rho_1]}\right]}{d\beta} \\
&= \frac{\rho\,(1-\rho)(1-\rho_1)}{(1-\rho_1-\beta\rho_2)^2\,[1-(1-\beta)\rho_1]^2} \cdot \Big\{ (\rho_1-\rho_2)\,(1-\rho_1-\beta\rho_2)\,[1-(1-\beta)\rho_1] \\
&\qquad - [1-(1-\beta)\rho_1-\beta\rho_2]\,[-\rho_2\,(1-(1-\beta)\rho_1)+\rho_1\,(1-\rho_1-\beta\rho_2)] \Big\}
\end{aligned}$$

The derivative can be rewritten as:

$$\frac{dP_{\text{sum}}}{d\beta} = \frac{\rho\,(1-\rho)(1-\rho_1)}{(1-\rho_1-\beta\rho_2)^2\,[1-(1-\beta)\rho_1]^2} \cdot \left[\,\beta^2(\rho_1-\rho_2)\,\rho_1\,\rho_2 + 2\rho_1\rho_2\beta(1-\rho_1)\,\right]$$

Simplifying further:

$$\frac{dP_{\text{sum}}}{d\beta} = \frac{\rho\,(1-\rho)(1-\rho_1)\,\beta\,\rho_1\,\rho_2\,\left[\,\beta(\rho_1-\rho_2)\,+2(1-\rho_1)\right]}{(1-\rho_1-\beta\rho_2)^2\,[1-(1-\beta)\rho_1]^2} \tag{7.23}$$

The term $\beta(\rho_1-\rho_2)+2(1-\rho_1)$ in the numerator can be rewritten as:

$$\beta\,\rho_1 + 2(1-\rho_1-\frac{\beta\,\rho_2}{2})$$

This term is positive since $(\beta\rho_1) > 0$ and $(1-\rho_1-\frac{\beta\rho_2}{2}) > 0$. Given that $1-\rho_1-\rho_2 > 0$ and $0 \le \frac{\beta}{2} < 1$, it follows that $(1-\rho_1-\frac{\beta\rho_2}{2}) \ge (1-\rho_1-\rho_2) > 0$.

The remaining terms in the numerator — ρ, $(1-\rho)$, $(1-\rho_1)$, β, ρ_1, ρ_2 — as well as the squared terms in the denominators, are also positive.

Thus, we conclude:

$$\frac{dP_{\text{sum}}}{d\beta} > 0 \quad \text{for } \beta \in (0,1]$$

Next, we show that P_{sum} for $\beta \in (0,1]$ is larger than P_{sum} at $\beta = 0$.

P_{sum} at $\beta = 0$ is $\rho(1-\rho)$. Thus, we need to show the following inequality holds for $\beta \in (0, 1]$

$$\left.\frac{\rho\,(1-\rho)\,(1-\rho_1)\,[1-(1-\beta)\rho_1-\beta\rho_2]}{(1-\rho_1-\beta\rho_2)\,[1-(1-\beta)\rho_1]}\right|_{\beta\in(0,1]} > \rho(1-\rho)$$

Cancelling the positive common term $\rho(1-\rho)$, we have:

$$\frac{(1-\rho_1)\,[1-(1-\beta)\rho_1-\beta\rho_2]}{(1-\rho_1-\beta\rho_2)\,[1-(1-\beta)\rho_1]} > 1$$

Since the denominator is positive as $1-\rho_1-\beta\rho_2 > 1-\rho_1 > 0$ and $[1-(1-\beta)\rho_1] > 1-\rho_1 > 0$, multiplying both sides by the denominator yields:

$$(1-\rho_1)\,[1-(1-\beta)\rho_1-\beta\rho_2] > (1-\rho_1-\beta\rho_2)\,[1-(1-\beta)\rho_1]$$

Expanding both the left and right sides:

$$(1-\rho_1)\cdot[1-(1-\beta)\rho_1]-(1-\rho_1)\beta\rho_2 > (1-\rho_1)\cdot[1-(1-\beta)\rho_1]-[1-(1-\beta)\rho_1]\beta\rho_2$$

Removing the common term $(1-\rho_1)\cdot[1-(1-\beta)\rho_1]$ from both sides yields:

$$-(1-\rho_1)\beta\rho_2 > -[1-(1-\beta)\rho_1]\beta\rho_2$$

Since β and ρ_2 are positive, we cancel the negative term $-\beta\rho_2$ and reverse the inequality, giving:

$$(1-\rho_1) < [1-(1-\beta)\rho_1]$$

This can be rewritten as:

$$\rho_1 > (1-\beta)\rho_1$$

Dividing both sides by ρ_1 (which is positive) gives:

$$1 > (1 - \beta)$$

This inequality is true for all $\beta \in (0, 1]$. Therefore, P_{sum} is larger for all $\beta \in (0, 1]$ compared to P_{sum} at $\beta = 0$.

Since P_{sum} is monotonically increasing for $\beta \in (0, 1]$ and P_{sum} values for $\beta \in (0, 1]$ are all larger than its value at $\beta = 0$, we can conclude that P_{sum} is monotonically increasing for $\beta \in [0, 1]$. Consequently, it reaches its maximum at the upper boundary of the interval, which is $\beta = 1$. This confirms that P_{sum} increases as β increases and is maximized when $\beta = 1$ for a given (ρ_1, ρ_2). Therefore, the theorem is proved.

■

Chapter 8: Fairness

So far, we have examined various performance measures in the context of different power forms, namely, individual power, sum of powers, and average power. We now introduce **fairness** as another critical dimension of analysis. This chapter explores several metrics commonly used to evaluate fairness in network systems. Beyond traditional metrics like throughput and delay[1], we propose **individual power** as an additional fairness measure.

We then conduct an analysis of equal power fairness in an M/M/1 system with two flows, considering both the beta-priority and delay-dependent queueing disciplines. Our analysis includes:

- Defining the feasible regions where equal power fairness is achievable.
- Identifying the optimal values of β (for the beta-priority system) and k (for the delay-dependent system) that achieve equal power fairness for given values of ρ_1 and ρ_2.

In the next chapter, we will explore how to integrate performance optimization with these fairness considerations, combining both performance and fairness measures into a unified framework.

[1] When mentioning throughput and delay here, we are considering them as fairness metrics, not performance measures.

8.1 Fairness Metrics

8.1.1 Throughput

Throughput is a commonly used fairness measure in network analysis [52–54]. Here are some key concepts in throughput fairness :

- **Equal Throughput Fairness**: This aims for an equal distribution of throughput across all flows. Often, Jain's index [55] is used to evaluate how close the actual throughput distribution is to an equal distribution.
- **Max-Min Fairness** [56]: This aims to maximize the minimum throughput for each flow. It ensures that resources are allocated to the flow with the least throughput requirement first, similar to filling buckets with water where the smallest bucket is filled first. This prevents any flow from being starved and promotes equitable resource distribution.
- **Proportional Fairness** [57, 58]: It aims to allocate resources, such as network bandwidth, in proportion to the demand of each user or flow. This method ensures that each user receives a resource share corresponding to their demand. It accomplishes this by maximizing the product of users' utilities, such that any small adjustment in allocation that benefits one user at the expense of another would proportionally decrease the overall system utility.

Note that, in our model, equal throughput is equivalent to equal utilization factors:

$$\rho_1 = \rho_2 = \ldots = \rho_i = \ldots = \rho_n$$

This is because we assume a no-loss system, where the throughput equals the arrival rate. Therefore, achieving equal throughput for multiple flows implies equal arrival rates. Additionally, we assume that each flow has the same average service rate (as discussed in Chapter 3). Therefore, equal throughput is equivalent to equal utilization factors, given that each λ_i is divided by the same μ, as in $\rho_i = \frac{\lambda_i}{\mu}$.

8.1.2 Delay (Response Time)

Delay (response time) is another crucial metric for network performance evaluation. One approach to defining fairness in the context of delay is to ensure equal delay (response time) for all flows, where each flow experiences the same average response time:

$$T_1 = T_2 = \ldots = T_i = \ldots = T_n$$

If the system is an M/M/1 system, achieving equal delay (response time) for all flows, where $T_1 = T_2 = \ldots = T_i = \ldots = T_n$, under the constraint of a fixed ρ, is only possible with work-conserving priority queueing disciplines that do not discriminate between flows, such as FCFS.

For other priority queueing disciplines that involve flow discrimination, some flows must be prioritized for lower response time, while others will experience longer response times. In these cases, equal delay is not achievable.

Therefore, only non-discriminatory queueing disciplines will result in delay fairness across all flows. Additionally, this equal delay time for each flow can be derived through the conservation law.

This can be formally stated as follows:

Theorem 8.1.

In an M/M/1 system with n flows, achieving equal delay (response time) for all flows under a fixed total utilization ρ is only possible with work-conserving queueing disciplines that have no flow discrimination, such as FCFS. Furthermore, in this case, each flow's delay is given by:

$$T_1 = T_2 = \ldots = T_i = \ldots = T_n = \frac{1}{\mu\,(1-\rho)} \tag{8.1}$$

Proof

In an M/M/1 system, the squared coefficient of variation of service time $C_b^2 = 1$. Substituting $C_b^2 = 1$ into the equation for the average residual service time, Equation 6.8, we have:

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

Substituting this into the conservation law, Equation 6.4, we obtain:

$$\sum_{i=1}^{n} \rho_i W_i = \frac{\rho W_0}{1-\rho} = \frac{\rho^2}{\mu(1-\rho)}$$

Using the relationship between response time T_i, waiting time W_i, and service time $\frac{1}{\mu}$ from Equation 6.3, we derive:

$$\begin{aligned}\sum_{i=1}^{n} \rho_i T_i &= \sum_{i=1}^{n} \rho_i (W_i + \frac{1}{\mu}) = (\sum_{i=1}^{n} \rho_i W_i) + (\sum_{i=1}^{n} \frac{\rho_i}{\mu}) \\ &= \frac{\rho^2}{\mu(1-\rho)} + \frac{\rho}{\mu} = \frac{\rho}{\mu} \cdot (\frac{\rho}{1-\rho} + 1)\end{aligned}$$

This leads to:

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

Now, assuming that each flow has the same response time:

$$T_1 = T_2 = \ldots = T_i = \ldots = T_n$$

we have,

$$\sum_{i=1}^{n} \rho_i T_i = \sum_{i=1}^{n} \rho_i T_1 = \rho \cdot T_1 = \frac{\rho}{\mu(1-\rho)}$$

This implies that:

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

Since all flows share the same response time, we have:

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

This result, derived from the conservation law, precisely matches the response time formula for the FCFS case in an M/M/1 system.

■

8.1.3 Individual Power

Here, we define another metric to evaluate fairness: equal individual power[2]. Mathematically, this is expressed as:

$$P_1 = P_2 = \ldots = P_i = \ldots = P_n$$

[2] We may also simply use "equal power". Whenever we refer to "equal power", we mean equal individual power.

Using power as a fairness measure offers several benefits, particularly in scenarios where resource allocation and user experience are crucial. Power, often expressed as a ratio of throughput to response time, directly reflects a user's perceived quality of service. A higher power value indicates that a user is receiving more data per unit of time, resulting in a smoother and more enjoyable experience.

Power uniquely captures the inherent trade-off between throughput and delay. As we said in the beginning, while high throughput might seem desirable, it can often lead to increased response times. This interrelationship between throughput and delay is not reflected when using only throughput or only delay as fairness metrics. However, power provides a more holistic view, balancing the desire for high throughput with the need for low response times.

In the following sections, we provide a detailed analysis of the concept of power fairness in an M/M/1 system with two flows. One section will focus on the beta-priority system, while another will examine the delay-dependent system.

8.2 Power Fairness in the Beta-Priority System: Two-Flows Analysis

8.2.1 Analyzing the Equal Power Condition

The beta-priority system prioritizes one flow over another, with the degree of prioritization determined by a parameter β. Equal power fairness in this system occurs when both flows achieve the same individual power. The individual power for flow 1 and flow 2 in the

beta-priority system is given by Equation 7.16. Setting them equal leads to:

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

This can be rewritten as:

$$\frac{\rho_1(1-\rho_1)(1-\rho_1-\rho_2)}{\beta \cdot (1-\rho_1-\rho_2) + (1-\beta) \cdot (1-\rho_1)} = \frac{\rho_2(1-\rho_1)(1-\rho_1-\rho_2)}{\beta + (1-\beta) \cdot (1-\rho_1)}$$

By canceling the common terms and multiplying by the denominators of both sides, we get:

$$\rho_1 \cdot [\beta + (1-\beta) \cdot (1-\rho_1)] = \rho_2 \cdot [\beta \cdot (1-\rho_1-\rho_2) + (1-\beta) \cdot (1-\rho_1)]$$

Simplifying this equation yields the condition on the parameters ρ_1, ρ_2 and β to achieve equal power:

$$\rho_1 \cdot [1 - (1-\beta) \cdot \rho_1] = \rho_2 \cdot (1 - \rho_1 - \beta \cdot \rho_2) \tag{8.2}$$

Equation 8.2 defines the condition for ρ_1 and ρ_2 to achieve equal individual power fairness in the beta-priority system. Figure 8.1 visualizes this condition by plotting Equation 8.2 for different β values ranging from 0 to 1. Each green curve in the figure represents the condition for equal individual power fairness for a specific value of β in the beta-priority system.

Importantly, these curves exist only for $\rho_1 \leq \rho_2$. This constraint arises because the beta-priority system prioritizes flow 1 when $\beta > 0$, leading to a shorter response time for flow 1 and a longer response time for flow 2. To achieve equal power, flow 2 needs a higher utilization factor ρ_2 to compensate for its longer response time, while flow 1 needs a lower utilization factor ρ_1. This explains why the curves for equal power fairness in the beta-priority system always lie to the left of the line $\rho_1 = \rho_2$ (the FCFS case), where $\rho_1 \leq \rho_2$.

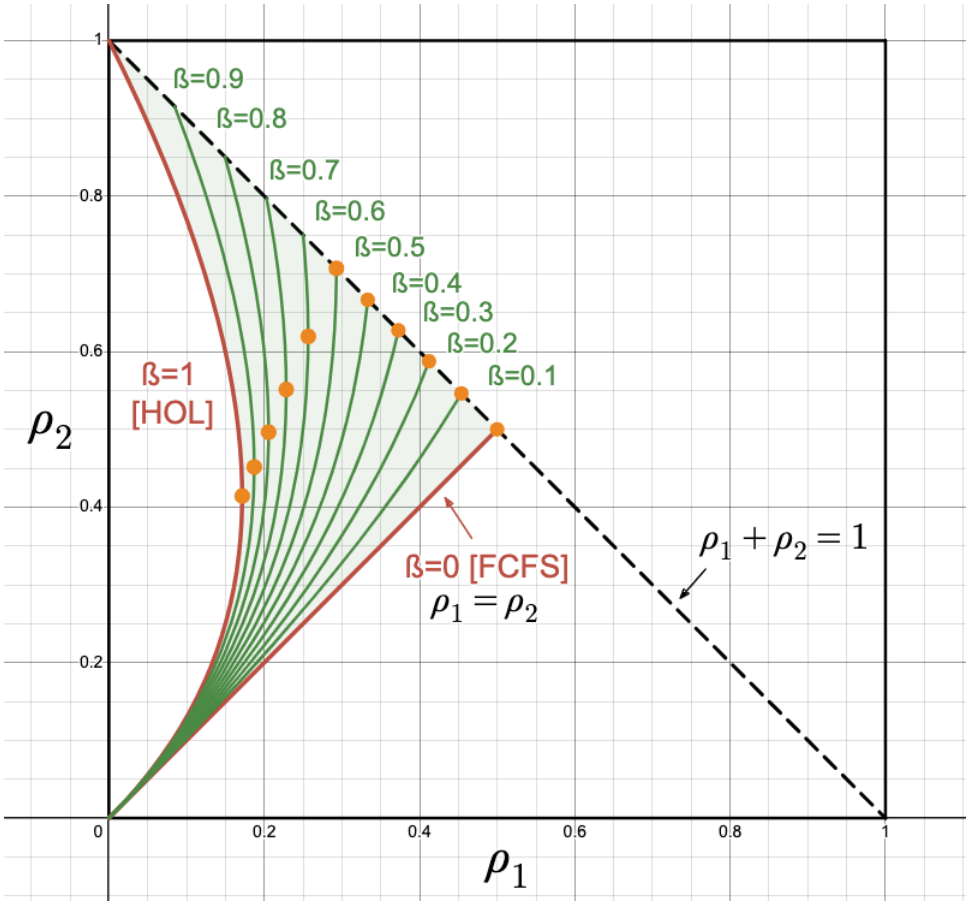


Figure 8.1: Equal power fairness for ρ_2 vs ρ_1 in an M/M/1 system with 2 flows using the beta-priority system. The green shaded region is the feasible region for equal power.

8.2.1.1 Feasible Region for Equal Power Fairness

Figure 8.1 depicts the feasible region for equal power fairness, which is shaded green. This region is bounded by three curves. The first curve, the black dashed curve $\rho_1 + \rho_2 = 1$, represents the system stability constraint $\rho = \rho_1 + \rho_2 < 1$, which restricts the feasible region to lie below this line.

The other two curves, shown in red, represent the extreme cases of the beta-priority queueing system. The left curve corresponds to HOL scheduling ($\beta = 1$), while the right curve corresponds to FCFS scheduling ($\beta = 0$). Substituting these values into Equation 8.2 results in the following equations for the red curves:

- HOL Case ($\beta = 1$): $\rho_2 = \frac{1-\rho_1 \pm \sqrt{{\rho_1}^2 - 6\rho_1 + 1}}{2}$, which is equivalent to $\rho_1 = \frac{\rho_2(1-\rho_2)}{1+\rho_2}$
- FCFS Case ($\beta = 0$): $\rho_2 = \rho_1$ (This also corresponds to equal throughput.)

In Figure 8.1, any pair of (ρ_1, ρ_2) within the feasible region can achieve equal individual power. This means that for a given pair of (ρ_1, ρ_2) within the boundary, there exists a β value that can be used to adjust the response time of the two flows to make their individual power values equal. Conversely, points outside this boundary represent (ρ_1, ρ_2) combinations where equal power is not attainable.

8.2.1.2 Determining the Optimal β

This observation about the boundary of achievable equal power leads to the following theorem:

Theorem 8.2.

Given a beta-priority system with two flows, each with a utilization factor (ρ_1, ρ_2), and under the constraints $0 < \rho_1 < 1$ and $0 < \rho_2 < 1$, there exists a β value that achieves equal individual power fairness for a specific pair of (ρ_1, ρ_2) if the following conditions hold:

- $\rho_2 < 1 - \rho_1$
- $\rho_2 \geq \rho_1$
- $\rho_2 \geq \frac{1-\rho_1+\sqrt{{\rho_1}^2-6\rho_1+1}}{2}$ *and* $\rho_2 \leq \frac{1-\rho_1-\sqrt{{\rho_1}^2-6\rho_1+1}}{2}$.
  (The union of these two conditions is equivalent to $\rho_1 \geq \frac{(1-\rho_2)\rho_2}{1+\rho_2}$.)

The corresponding value of β that achieves equal individual power fairness is given by:

$$\beta = \frac{(1-\rho_1)(\rho_1-\rho_2)}{({\rho_1}^2+{\rho_2}^2)} \tag{8.3}$$

Proof

These conditions correspond to the feasible equal power solution region in Figure 8.1.

- $\boldsymbol{\rho_2 < 1 - \rho_1}$:

  Rewriting this gives $\rho_1 + \rho_2 < 1$, which ensures that the system is not overloaded. This corresponds to the left side of the curve $\rho_1 + \rho_2 = 1$.

- $\boldsymbol{\rho_2 \geq \rho_1}$:

  it ensures that the corresponding β value is non-negative. This arises from the FCFS case where $\rho_2 = \rho_1$. Since the FCFS case represents the lower bound with $\beta = 0$, all cases with $\beta \geq 0$ fall on the left side of the curve $\rho_2 = \rho_1$, leading to $\rho_2 \geq \rho_1$.

- $\boldsymbol{\rho_2 \geq \frac{1-\rho_1+\sqrt{\rho_1^2-6\rho_1+1}}{2}}$ and $\boldsymbol{\rho_2 \leq \frac{1-\rho_1-\sqrt{\rho_1^2-6\rho_1+1}}{2}}$:

  These two conditions correspond to $\boldsymbol{\rho_1 \geq \frac{(1-\rho_2)\rho_2}{1+\rho_2}}$, indicating the region to the right of the parabolic curve $\rho_1 = \frac{(1-\rho_2)\rho_2}{1+\rho_2}$, which is derived from the HOL case. Since the HOL case represents the upper bound with $\beta = 1$, all cases with $\beta \leq 1$ fall to the right of this parabolic curve, leading to $\rho_1 \geq \frac{(1-\rho_2)\rho_2}{1+\rho_2}$ and therefore corresponding to $\rho_2 \geq \frac{1-\rho_1+\sqrt{\rho_1{}^2-6\rho_1+1}}{2}$ and $\rho_2 \leq \frac{1-\rho_1-\sqrt{\rho_1{}^2-6\rho_1+1}}{2}$.

For the corresponding β values, we start with the condition for equal individual power fairness, Equation 8.2. This equation is given by:

$$\rho_1 \cdot [1 - (1 - \beta) \cdot \rho_1] = \rho_2 \cdot (1 - \rho_1 - \beta \cdot \rho_2)$$

Rewriting it, we get:

$$\rho_1 \cdot (1 - \rho_1 + \beta\rho_1) = \rho_2 \cdot (1 - \rho_1 - \beta \cdot \rho_2)$$

Rearranging the equation by grouping terms with β and terms without β, we get:

$$\beta {\rho_1}^2 + \beta {\rho_2}^2 = (1 - \rho_1) \cdot (\rho_2 - \rho_1)$$

Isolating β and factoring out β on the left-hand side, we have:

$$\beta({\rho_1}^2 + {\rho_2}^2) = (1 - \rho_1) \cdot (\rho_2 - \rho_1)$$

Solving for β by dividing both sides by $({\rho_1}^2 + {\rho_2}^2)$ yields:

$$\beta = \frac{(1 - \rho_1)\,(\rho_1 - \rho_2)}{({\rho_1}^2 + {\rho_2}^2)}$$

Therefore, given a specific pair of utilization factor parameters (ρ_1, ρ_2) that fall within the feasible region, Equation 8.3 provides the corresponding β value that achieves equal individual power fairness in the beta-priority system. This completes the proof.

■

8.2.2 Analyzing the Maximum Achievable ρ_1

Let us back to Figure 8.1, which depicts the relationship between ρ_1, ρ_2, and β within the beta-priority system. For power fairness, dots on each curve mark the upper bound for achievable ρ_1 values at each value of β while maintaining equal power allocation. These dots represent the maximum value for ρ_1 and the corresponding ρ_2 values for each of these equal power curves.

As β increases, the maximum achievable ρ_1 for equal power decreases. This indicates a tighter constraint on ρ_1 when higher flow discrimination is applied to maintain individual

flow power fairness. *Such power fairness prevents the higher priority flow from transmitting excessive traffic, thereby protecting the lower priority flow by imposing reasonable limits on the value of* ρ_1.

For $0.5 < \beta \leq 1$, the dots align with the vertical tangents to the respective green curves. These tangents are determined by differentiating Equation 8.2 with respect to ρ_2, resulting in

$$\rho_1 = 1 - 2 \cdot \beta \cdot \rho_2$$

Substituting this expression into the equation of the green curve yields the maximum achievable ρ_1 values for each β and the corresponding ρ_2:

$$\rho_1 = \frac{(2 - 2\sqrt{2})\beta + 1}{-4\beta^2 + 4\beta + 1}$$

$$\rho_2 = \frac{-2\beta + (1 + \sqrt{2})}{-4\beta^2 + 4\beta + 1}$$

These points are feasible for $0.5 < \beta \leq 1$. However, for $0 \leq \beta \leq 0.5$, they become infeasible because they coincide with the boundary curve where $\rho_1 + \rho_2 = 1$, a constraint imposed to prevent system overload. Since the sum of the calculated ρ_1 and ρ_2 exceeds 1 when $0 \leq \beta \leq 0.5$, the upper bound for achievable ρ_1 in this range is dictated by the $\rho_1 + \rho_2 = 1$ boundary curve.

To find the intersection points of the green curves and the boundary curve, we substitute

$\rho_1 + \rho_2 = 1$ into Equation 8.2, resulting in:

$$\rho_1 = \frac{3 - 2\beta - \sqrt{4\beta - 4\beta^2 + 1}}{4(1 - \beta)}$$

$$\rho_2 = \frac{1 - 2\beta + \sqrt{4\beta - 4\beta^2 + 1}}{4(1 - \beta)}$$

In addition, when $0.5 < \beta \leq 1$, a vertical line intersecting the green curve can have two equal power intersection points within a specific region of ρ_1. This region is limited by the line $\rho_1 + \rho_2 = 1$ and the vertical tangent to the curve, defined by:

$$\frac{3 - 2\beta - \sqrt{4\beta - 4\beta^2 + 1}}{4(1 - \beta)} \leq \rho_1 \leq \frac{(2 - 2\sqrt{2})\beta + 1}{-4\beta^2 + 4\beta + 1} \quad \text{for} \quad 0.5 < \beta \leq 1$$

These two intersection points correspond to different ρ_2 values, each of which achieves equal power, but offering distinct performance characteristics. A higher ρ_2 prioritizes throughput, while a lower ρ_2 favors lower response time.

8.2.3 Summary and Applications

The behavior of the system varies based on the value of β. For $0.5 < \beta \leq 1$, the system offers multiple performance profiles, with the maximum achievable ρ_1 for a given β determined by the tangent to the green curve at that value of β. For $0 \leq \beta \leq 0.5$, the system is constrained by the boundary curve $\rho_1 + \rho_2 = 1$, and there is only a single corresponding ρ_2 value for each ρ_1. This analysis reveals that as the priority disparity between flows increases (larger β), the system imposes stricter limitations on the lower priority flow to maintain power fairness, offering fewer options.

Figure 8.1 delineates the parameter space for achieving power fairness in terms of ρ_1 and ρ_2. It is a versatile tool for various optimization tasks. For instance, given ρ_1 and ρ_2, the figure can determine the appropriate β with Theorem 8.2.

Conversely, when the total system utilization is constrained to a fixed value, say c (i.e., $\rho_1 + \rho_2 = c$), power fairness for ρ_1, ρ_2, and β can be determined. For example, when targeting maximal average power with $\rho = 0.5$, we draw the line $\rho_1 + \rho_2 = 0.5$ and find the intersection with feasible power fairness regions. At this intersection, the set of ρ_1, ρ_2, and β that achieves both optimal average power and power fairness can be identified.

Moreover, for points outside the feasible fairness region, the figure quantifies the proximity to the optimal fairness curve for a specific β, revealing the performance gap to ideal power fairness.

8.3 Power Fairness in the Delay-Dependent System: Two-Flows Analysis

To determine the condition when the power of flow 1 equals the power of flow 2 for the delay-dependent system, we set P_1 equal to P_2 and proceed with the analysis. The power for flow 1 and flow 2 are given by Equation 7.1:

$$P_1 = \frac{\rho_1(1-\rho)(1-k\rho_1)}{1-k\rho}$$

$$P_2 = \rho_2(1-\rho)(1-k\rho_1)$$

Setting $P_1 = P_2$ gives:

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

Cancelling the common factor $(1-\rho)(1-k\rho_1)$ on both sides:

$$\frac{\rho_1}{1-k\rho} = \rho_2$$

Rewriting this and substituting $\rho = \rho_1 + \rho_2$, we get:

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

Solving for ρ_1, we have the relationship equation for ρ_1 and ρ_2 under which the powers of flow 1 and flow 2 are equal:

$$\rho_1 = \frac{\rho_2(1-k\rho_2)}{1+k\rho_2} \tag{8.4}$$

Equation 8.4 defines the condition for the delay-dependent system with $n = 2$ to achieve equal power fairness. This condition is illustrated in Figure 8.2 by plotting Equation 8.4 for different values of k.

8.3.1 Analyzing the Equal Power Condition

In Figure 8.2, each blue curve in the figure represents the condition for equal power fairness for a specific value of k in the delay-dependent system. The two red curves in the figure represent the same extreme cases: HOL with $k = 1$ and FCFS with $k = 0$, which align with the beta-priority system.

8.3.1.1 Feasible Region for Equal Power Fairness

The feasible region for achieving equal power fairness in the delay-dependent system is the same as that in the beta-priority system. This region is constrained by the following inequalities:

1. **System Stability**: $\rho_2 < 1 - \rho_1$ (given that $\rho_1 + \rho_2 < 1$)

2. **FCFS Boundary**: $\rho_2 \geq \rho_2$

3. **HOL Boundary**: $\rho_2 \geq \frac{1-\rho_1+\sqrt{\rho_1{}^2-6\rho_1+1}}{2}$ and $\rho_2 \leq \frac{1-\rho_1-\sqrt{\rho_1{}^2-6\rho_1+1}}{2}$ (This is equivalent to $\rho_1 \geq \frac{(1-\rho_2)\rho_2}{1+\rho_2}$)

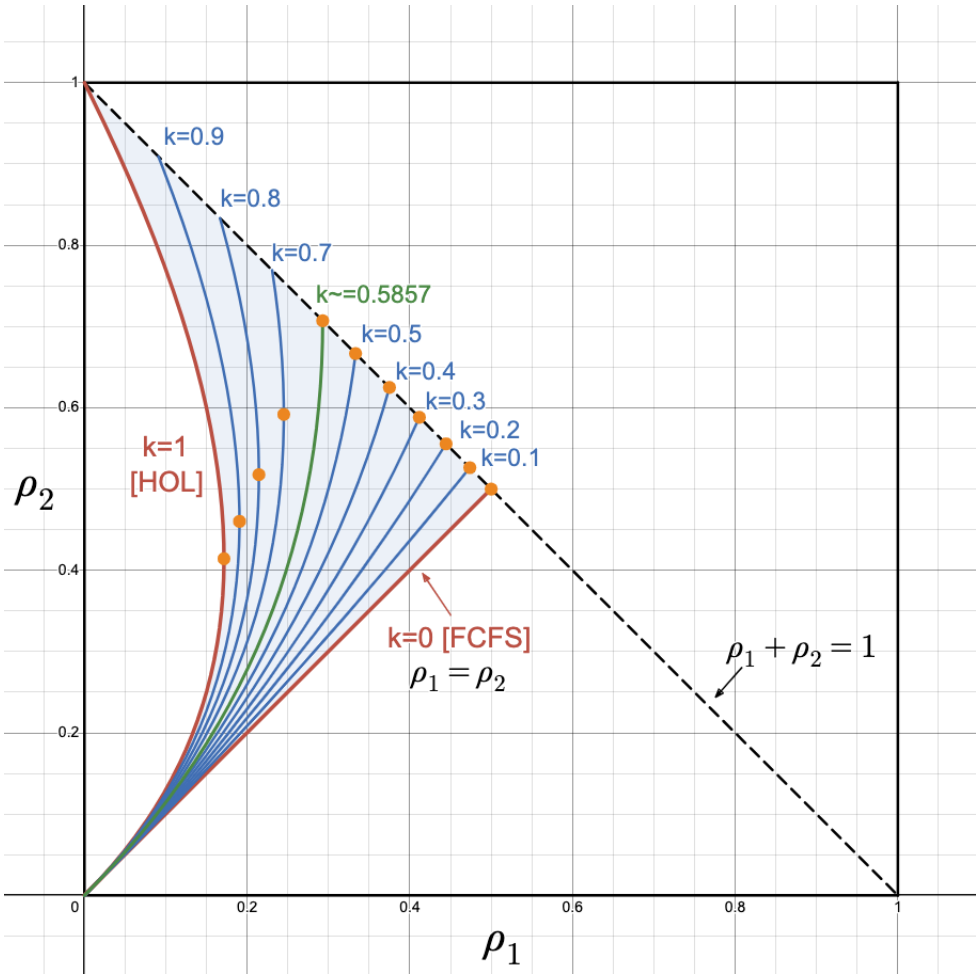


Figure 8.2: Equal power fairness for ρ_2 vs ρ_1 in an M/M/1 system with 2 flows using the delay-dependent system. The blue shaded region is the feasible region for equal power.

8.3.1.2 Determining the Optimal k

Within this feasible region, we derive a theorem similar to the one for the beta-priority system:

Theorem 8.3.

Given any pair of (ρ_1, ρ_2) *inside the feasible region bounded by the three curves above, there exists a* k *value that can arrange the queueing discipline in the delay-dependent system to result in equal power for flow 1 and flow 2. The* k *value is given by:*

$$k = \frac{\rho_2 - \rho_1}{\rho_2 \left(\rho_2 + \rho_1\right)} \tag{8.5}$$

Proof

The following derivation of the k value in Equation 8.5 follows from the power fairness condition (Equation 8.4):

$$\rho_1 = \frac{\rho_2(1 - k\rho_2)}{1 + k\rho_2}$$

Rearranging the terms to group terms with k and without k:

$$k\,\rho_2\left(\rho_1 + \rho_2\right) = \rho_2 - \rho_1$$

Solving for k leads to Equation 8.5:

$$k = \frac{\rho_2 - \rho_1}{\rho_2 \left(\rho_2 + \rho_1\right)}$$

■

8.3.2 Analyzing the Maximum Achievable ρ_1

Similar to the beta-priority case, the dots in Figure 8.2 represent the maximum achievable ρ_1 value for each k while maintaining equal power fairness. These dots correspond to the vertical tangents to the blue curves, which represent the equal power condition for different k values. Importantly, this applies specifically for $k > 2 - \sqrt{2}$, with the green curve in the figure representing the case where $k = 2 - \sqrt{2} \approx 0.5857$. When $k \leq 2 - \sqrt{2}$, these dots fall on the boundary curve $\rho_1 + \rho_2 = 1$, indicating that the maximum achievable ρ_1 is constrained by the system stability condition.

To find the equation of the points where these tangents occur, we differentiate Equation 8.4 with respect to ρ_2, which gives us:

$$\rho_1 = \frac{1}{k} - 2\rho_2$$

Substituting this expression into the equation of each blue curve yields:

$$\rho_1 = \frac{3 - 2\sqrt{2}}{k}$$

$$\rho_2 = \frac{\sqrt{2} - 1}{k}$$

These points correspond to the vertical tangent curves for each equal power fairness curve at different values of k. The sum of these $\rho_1 + \rho_2$ is equal to 1 when $k = 2 - \sqrt{2}$, showing that the point of the vertical tangent line that hits the boundary curve of $\rho_1 + \rho_2 = 1$. Therefore, when $k <= 2 - \sqrt{2}$, the maximum achievable ρ_1 is bounded by $\rho_1 + \rho_2 = 1$.

The equal power condition curve at this $k = 2 - \sqrt{2}$ is represented by the green curve in the figure. It marks the dividing line:

- For $k \leq 2 - \sqrt{2}$, the maximum achievable ρ_1 is bounded by the $\rho_1 + \rho_2 = 1$ curve.
- For $2 - \sqrt{2} < k$, the maximum achievable ρ_1 is determined by the tangent point of the vertical line to the blue curve.

The intersection point of the boundary line $\rho_1 + \rho_2 = 1$ with each equal power condition blue curve is at:

$$\rho_1 = \frac{1-k}{2-k}$$

$$\rho_2 = \frac{1}{2-k}$$

For $2 - \sqrt{2} < k \leq 1$, the points of vertical tangency remain within the feasible equal power region, indicating that the equal power curves at these k values can produce a vertical line with two intersection points. This means that for each ρ_1, it is possible to have two corresponding ρ_2 values that achieve equal power with flow 1. The range of ρ_1 that can have two ρ_2 intersection points is:

$$\frac{1-k}{2-k} \leq \rho_1 \leq \frac{2-2\sqrt{2}}{k} \quad \text{for} \quad 2 - \sqrt{2} < k \leq 1$$

For each k, a vertical line within the corresponding ρ_1 region results in two intersection points of ρ_2 with the equal power curve at this k. These two ρ_2 values offer different performance characteristics: one provides higher throughput, while the other results in lower throughput and thus shorter response time.

Chapter 9: Performance, Fairness, and Priority Flow Discrimination

In previous chapters (Chapters 4 through 7), we defined three performance optimization metrics: individual power of the i^{th} flow for $i = 1, .., n$ (P_i), sum of powers (P_{sum}), and average power (P_{avg}). We then applied these metrics to different queueing disciplines to perform optimization. In Chapter 8, we explored common fairness metrics, namely, throughput and delay, and proposed an additional fairness metric, namely, individual power.

This chapter investigate the relationship between **performance** and **fairness**. We look into this in various scheduling disciplines and explore the full spectrum of **priority flow discrimination** from the least to the most discriminatory. We investigate the fundamental question: *Can we optimize performance without compromising fairness for a variety of priority disciplines that create different levels of flow discrimination?*

To systematically analyze this relationship, we first introduce a three-dimensional framework that encompasses various performance metrics, various fairness metrics, and various priority disciplines (flow discrimination). We begin by investigating the interplay between performance and fairness in a simple case with only two flows. Subsequently, we extend the analysis to scenarios with an arbitrary number of flows.

9.1 A Three-Dimensional Framework

The conservation law principle, as articulated by Kleinrock [33], states that the sum of the products of each flow's utilization factor and average waiting time remains constant over all conservative queueing disciplines:

$$\sum_{i=1}^{n} \rho_i \cdot W_i = \text{constant value} \tag{9.1}$$

This mathematical relationship implies that prioritizing certain flows – leading to reduced response times for some – inevitably results in increased response times for others. This inherent trade-off raises a critical question: *Can we optimize performance without compromising fairness, especially when different flows receive differential priority treatment?*

The answer to this question depends heavily on the specific metrics used to evaluate performance and fairness. In this dissertation, we adopt the performance metrics defined and examined in Chapters 4 to 7 and the fairness metrics introduced in Chapter 8. For the performance metrics, we aim for the optimal value of each separately: **optimal individual power P_i^*** [1], **optimal sum of (individual) powers P_{sum}^***, and **optimal average power P_{avg}^***. For the fairness metric, we target equal values across flows[2], namely, **equal throughput fairness**, **equal delay fairness**, and **equal power fairness**.

Our goal is to shed light on the complex interplay between performance and fairness in scheduling systems with flow discrimination. The following introduces a three-dimensional framework that encompasses performance, fairness, and the degree of flow discrimination:

[1] Note that optimal individual power refers to the situation where each flow optimizes its individual power and reaches an equilibrium point.

[2] From this point on, whenever we talk about fairness, we refer to equal value of the metric for each flow.

- **Performance Metrics**: We consider three types of power as previously discussed: individual power P_i, sum of individual powers P_{sum}, and average power P_{avg}.

- **Fairness Metrics**: We select throughput ρ_i[3], delay T_i, and individual power P_i as our fairness metrics.

- **Degree of Priority Discipline Flow Discrimination**: We evaluate a range of work-conserving priority queueing disciplines that vary in their level of flow discrimination, from first-come-first-serve (FCFS), which is the least discriminatory, to preemptive head-of-line (HOL) priority queueing, which is the most discriminatory. To span the range of disciplines from FCFS to HOL, we consider the delay-dependent system as well as the beta-priority system as introduced in Chapter 3. The parameter k in the delay-dependent system and β in the beta-priority system represent the parameter that determines the degree of flow discrimination.

This framework allows us to compare the performance and fairness characteristics of different priority scheduling disciplines in a comprehensive way. For visualization, we represent the framework as a three-dimensional space:

- ***x*-axis**: discrete space for performance metrics — individual power P_i^* $(x = 1)$, sum of powers P_{sum}^* $(x = 2)$, average power P_{avg}^* $(x = 3)$

- ***y*-axis**: discrete space for fairness metrics — delay T_i $(y = 1)$, throughput ρ_i $(y = 2)$, individual power P_i $(y = 3)$

- ***z*-axis**: continuous space for degree of flow discrimination ranging from FCFS (no flow discrimination at $z = 0$) to HOL (maximum flow discrimination at $z = 1$)[4].

[3] We use the utilization factor to represent throughput. As explained in Chapter 8, assuming a no-loss system and that the average service rate for each flow is the same, equal throughput for each flow is equivalent to an equal utilization factor for each flow.

[4] Note that minimum or maximum flow discrimination can be achieved through methods other than FCFS and HOL. For example, last-come-first-serve (LCFS) can also result in the same response time as FCFS and also represents minimum flow discrimination.

An example is shown in Figure 9.1 (the details of this figure will be discussed in the subsequent section). At each layer of queueing discipline, there is a surface, and each surface has 9 intersection points, composed of 3 performance metrics multiplied by 3 fairness metrics. At each intersection point in a given surface layer, we have a performance metric and a fairness metric corresponding to that queueing discipline.

For all figures in this chapter, values are only entered when the performance metric is maximal and the fairness metrics have equal values for each flow under that queueing discipline. Entries are made only at the intersection point where optimal performance and optimal fairness coincide.

The value is represented by a circle at each z-layer, with the radius corresponding to the optimal value of the performance metric for the associated queueing discipline. If the performance metric is average power, the circle represents the average power value. For the other two performance metrics, the circles both represent the sum of individual powers values. Given that individual power may differ even when equal throughput or equal delay is achieved, we choose to use the sum of individual powers (which is the sum of maximal individual powers) to represent the optimal operating point status.

9.2 Analysis in a Two-Flow System

Figure 9.1 exemplifies the three-dimensional framework we introduced for analyzing both optimal performance and optimal fairness in systems with two flows. This figure demonstrates how different queueing disciplines are assessed based on maximum performance and maximum fairness metrics, emphasizing the intersection points where both criteria are met. In other words, only regions of simultaneous optimum performance with fairness are shown in this figure.

The lowest layer (i.e., $z = 0$) in the figure represents zero discriminatory case (i.e., First-Come-First-Serve, FCFS), while the upper layer (i.e., $z = 1$) represents the most discriminatory case, (Head-of-Line Priority, HOL) case. We will illustrate this framework by first examining various performance metrics and then evaluating whether the corresponding fairness metrics can be satisfied[5].

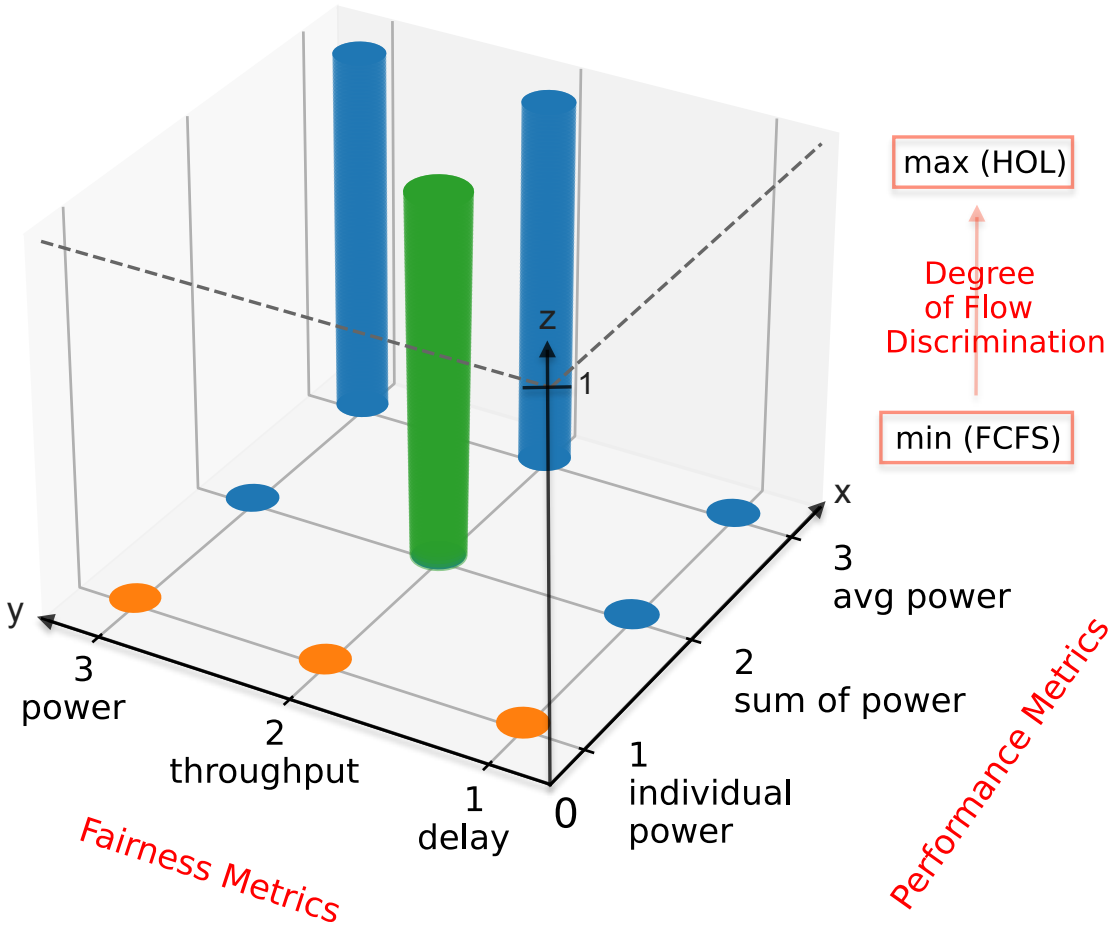


Figure 9.1: A three-dimensional framework for simultaneous performance optimization and fairness in an M/M/1 system with two flows using work-conserving queueing disciplines.

[5] Note that we only discuss scenarios where full fairness is achieved, based on equal values of the metric for each flow. Fairness can be assessed with unequal values and quantified using approaches like Jain's index [55], but we do not cover these cases.

9.2.1 Fairness for Individual Power P_i^* Optimization

Along the plane of individual power (i.e., $x = 1$) in Figure 9.1 represents the individual power metric, with three orange circles located at the FCFS layer. These circles have an identical radius size of $\frac{2}{9}$. They represent the same operating point in the FCFS case, where each flow has the same throughput (utilization factor ρ_i) (at $y = 2$), the same delay T_i (at $y = 1$), and thus the same individual power P_i (at $y = 3$). Since all three fairness metrics are satisfied at this optimal performance metric, i.e., individual power, there are three circles corresponding to these fairness metrics. Only at the layer of FCFS where $z = 0$, are all three fairness metrics satisfied. It could have been that not all three fairness metrics are satisfied like in other layers where $0 < z \leq 1$. In the following, we provide a detailed explanation of why only FCFS achieves this result and how this value is derived.

Using individual power as a performance optimization metric for FCFS and HOL was discussed in Chapter 4, and the optimization results are summarized in Table 4.1. We utilize the delay-dependent system and the beta-priority system to explore the full range of flow discrimination spanning from FCFS to HOL. The individual power optimization for both systems is analyzed in Chapter 7. The delay-dependent results can be found in Figure 7.1 and Figure 7.2. The beta-priority system results can be found in Figure 7.6 and Figure 7.7.

The following examines the fairness using different fairness metrics after maximizing individual power:

9.2.1.1 Throughput Fairness

In Figure 9.1, when examining the intersection of maximum individual power performance ($x = 1$) and throughput fairness ($y = 2$) along the z-axis (from $z = 0$ FCFS layer to $z = 1$ HOL layer), only the bottom plane $z = 0$ (FCFS) contains a solution represented by a circle. This

indicates that *no other queueing disciplines in the full spectrum of disciplines except FCFS can achieve both optimal individual power performance and throughput fairness simultaneously.*

To better understand this, let's look at the specifics of each case:

- **FCFS Case**: From Table 4.1, the optimal ρ_i^* is $\frac{n}{n+1}$ for all flows in FCFS, demonstrating equal throughput. The individual power in the FCFS case with two flows is $P_1 = P_2 = \frac{1}{9}$, and thus the sum of maximal individual powers is $\frac{2}{9}$, which is the radius of the orange circle in Figure 9.1.

- **HOL Case**: From Table 4.1, HOL has the result $\rho_i^* = (\frac{1}{2})^i$ for optimal individual power, showing that ρ_1 and ρ_2 are not identical. This confirms that throughput fairness is not achievable with HOL.

- **Other Queueing Disciplines**: For other queueing disciplines between FCFS and HOL, we refer to the results in Chapter 7. Figures 7.1 and 7.6 demonstrate the ρ_1 and ρ_2 that optimize individual power simultaneously for the delay-dependent system and the beta-priority system, respectively. These figures show that throughput fairness is achieved only when $k = 0$ or when $\beta = 0$, both of which correspond to the FCFS case. For other queueing disciplines with any flow discrimination, the optimal ρ_1 and ρ_2 are different and therefore cannot achieve throughput fairness.

Note: Although the delay-dependent system and the beta-priority system both represent a full range from FCFS to HOL, there may be other families of queueing disciplines that also span the full spectrum. We believe that no other queueing disciplines except FCFS can simultaneously result in optimal individual power and throughput fairness. However, we cannot definitively prove this. We can only show that the solution doesn't exist for HOL. For other disciplines, we haven't provided proof. We have only studied these two families of queueing disciplines and conjecture that that the solution does not exist for any family.

9.2.1.2 Power Fairness

In the intersection of performance being individual power ($x = 1$) and fairness being equal power ($y = 3$) in Figure 9.1, only one circle appears at the FCFS layer, paralleling the situation in the equal throughput case. This indicates that, similar to throughput, *no other queueing disciplines except FCFS can achieve both optimal individual power and equal individual power simultaneously.*

To better understand this, let's look at the specifics:

- **FCFS Case**: From Table 4.1, FCFS achieves equal individual power with $P_i^* = \left(\frac{1}{n+1}\right)^2$ for $i = 1, \ldots, n$. In our two-flow case, this is $P_1^* = P_2^* = \frac{1}{9}$, and the sum of maximal individual powers is $\frac{2}{9}$. Thus, the radius of the circle at this intersection point on the FCFS layer is $\frac{2}{9}$. Thus, the radius of the circle at this intersection point on the FCFS layer is $\frac{2}{9}$. Note that this optimal operating point is the same as in the equal throughput case, and it also leads to the same delay.

- **HOL Case**: Table 4.1 shows that for HOL, $P_i^* = 2\left(\frac{1}{8}\right)^i$, indicating $P_1^* \neq P_2^*$. This confirms that power fairness is not achievable with HOL when maximizing individual power.

- **Other Queueing Disciplines**: Figures 7.2 and 7.7 illustrate the optimal individual powers P_1^* and P_2^* for the delay-dependent system and the beta-priority system, respectively. These figures show that power fairness, $P_1^* = P_2^*$, is achieved only when $k = 0$ or $\beta = 0$, both of which correspond to the FCFS case. For other queueing disciplines with any flow discrimination, the optimal P_1^* and P_2^* are different.

In summary, both throughput fairness and power fairness demonstrate that only the FCFS discipline can achieve both optimal individual power and fairness in terms of throughput or

power simultaneously. When each flow is optimizing individual power under FCFS, all flows experience the same response time, leading them to select the same ρ_i and consequently the same P_i^*. However, in queueing disciplines where flows are discriminated against, some flows are prioritized and experience shorter response times, resulting in higher ρ_i and higher P_i^*. Conversely, flows that are not prioritized encounter longer response times, leading to lower ρ_i and lower P_i^*. Therefore, only when all flows observe the same response time, as in the FCFS case, do they select the same ρ_i and achieve equal P_i^*.

It's important to note that while the sum of the product of utilization factor and response time, $\sum_{i=1}^{n} \rho_i \cdot T_i$, is a constant term based on conservation law [33], the sum of optimal individual powers, which is calculated as the division of utilization factor over response time, $\sum_{i=1}^{n} P_i^*$, is not a constant term. This can be observed from Figures 7.2 and 7.7, where the sum of optimal individual powers increases as the degree of flow discrimination increases.

Note: Similar to the equal throughput case, we only formally prove that the HOL case cannot achieve equal power. We did not formally prove this for other queueing disciplines. However, we infer that no other queueing disciplines except FCFS can result in optimal individual power and equal power, based on the optimization results observed in both the delay-dependent system and the beta-priority system.

9.2.1.3 Delay Fairness

From Chapter 8, we know that delay fairness occurs when each flow has the same response time: $T_1 = T_2 = \frac{1}{\mu(1-\rho)}$. This implies that delay fairness occurs only in the absence of flow discrimination. With flow discrimination, even if the degree of discrimination is subtle, we cannot find a pair (ρ_1, ρ_2) where $0 < \rho_1 < 1$ and $0 < \rho_2 < 1$ as well as $(0 < \rho_1 + \rho_2 < 1)$ that leads to equal average response time. The response time with flow discrimination is not in

the form $T_1 = T_2 = \frac{1}{\mu(1-\rho)}$. Therefore, when maximizing individual power for a system using a queueing discipline with flow discrimination (excluding FCFS), equal delay does not exist. This is because even without optimization, we cannot find a feasible pair (ρ_1, ρ_2) within the region where $0 < \rho_1 < 1$ and $0 < \rho_2 < 1$ as well as $(0 < \rho_1 + \rho_2 < 1)$. Not to mention, individual power optimization falls within this region, the region where equal power doesn't exist.

As a result, only FCFS has a data point in the three-dimensional framework for the intersection of individual power metric and equal delay. This is represented in Figure 9.1 by a single circle at the bottom layer (FCFS). The radius of this circle represents the sum of optimal individual powers, which is the same as in the equal throughput and equal power cases, with the circle radius being $\frac{2}{9}$.

9.2.2 Fairness for Sum of Power P^*_{sum} Optimization

The sum of power plane at $x = 2$ in Figure 9.1 shows the results for using the sum of power as the performance optimization goal. For power fairness ($y = 3$) and delay fairness ($y = 1$), only the FCFS can achieve both optimal P_{sum} and fairness simultaneously when evaluated by these two metrics as shown by the two circles at the FCFS plane. However, for throughput fairness ($y = 2$), there is a green column[6] extending from the bottom to the top, indicating that optimized sum of power and throughput fairness is achievable across the full range of queueing disciplines. The radius of the green column increases with the degree of flow discrimination, showing that the corresponding maximal sum of power increases as one moves up from FCFS to HOL.

Next, we examine each fairness criterion in detail.

[6] The column is green because the sum of power increases as the degree of flow discrimination increases. This is to differentiate it from a column where the radius size does not change.

9.2.2.1 Throughput Fairness

When maximizing the sum of individual power, P_{sum}, it was shown in Chapter 7, specifically in Theorem 7.2, that P_{sum} is maximal when $\rho_1 = \rho_2$[7] [8]. These results demonstrate that *both* ***optimal sum of power*** $\boldsymbol{P^*_{sum}}$ *and* ***throughput fairness*** *can be achieved across a full range of queueing discrimination, from FCFS to HOL, in an M/M/1 system with two flows.* While the beta-priority system does not exhibit this result, the fact that the delay-dependent system achieves it indicates the existence of this intersection point of optimal sum of powers and throughput fairness across the entire range of discrimination.

In the three-dimensional framework, the intersection of maximum sum of powers ($x = 2$) and throughput fairness ($y = 2$) is represented by a green column extending from the $z = 0$ layer (FCFS) to the $z = 1$ layer (HOL) in Figure 9.1. The radius size of the column, which corresponds to the optimal sum of power, increases as the degree of flow discrimination increases. The value at the bottom is 0.25, and the value at the top is $\frac{8}{27} \approx 0.296$. The values in between can be found in Figure 7.4.

9.2.2.2 Power Fairness

While maximizing the sum of individual power, P_{sum}, we observe that power fairness cannot be maintained across all queueing disciplines, unlike throughput fairness. As stated in Theorem 7.2, equal throughput is necessary to maximize the sum of power in the delay-dependent system. However, when flow discrimination is applied, each flow experiences a different response time, leading to unequal individual power when P_{sum} is maximized since

[7] Note that equal throughput (utilization factor) is a necessary but not sufficient condition for maximal sum of power. While maximal sum of power results in equal utilization factors, the converse is not always true. Equal utilization factors do not guarantee maximal sum of power.

[8] Note that for FCFS, there is no restriction on having equal throughput to maximize the sum of power in the theorem, but the throughput (utilization) of each flow can be set identically while still maintaining the maximal sum of power.

equal throughput is required.

Only in the FCFS case, where each flow has the same response time, can power fairness be achieved. Other queueing disciplines with flow discrimination cannot satisfy both optimal sum of power, P^*_{sum}, and power fairness simultaneously. This confirms that only FCFS can achieve both optimal performance in the context of maximizing the sum of power and power fairness, shown as an blue dot at $(x, y, z) = (2, 3, 0)$.

Figure 7.4 again demonstrates the full range of individual power values at the maximal sum of power, showing that only when $k = 0$, with no flow discrimination, will P_1 be identical to P_2. Therefore, in Figure 9.1, the intersection of the sum of power and power fairness $(x, y) = (2, 3)$ is represented by a single circle with a radius of 0.25 at the bottom layer $(z = 0)$. This radius corresponds to the optimal sum of power for FCFS, which is 0.25.

9.2.2.3 Delay Fairness

Similar to using individual power as the performance optimization goal, only flow without discrimination, i.e., FCFS, can achieve delay fairness. Other queueing disciplines with flow discrimination can only have delay fairness when $\rho_1 = \rho_2 = 0$, which falls outside the region of power optimization as that leads to minimal power, which is zero. Therefore, queueing disciplines with any flow discrimination cannot reach the maximal sum of individual power $(x = 2)$ and delay fairness $(y = 1)$ simultaneously.

For the FCFS case, the sum of power is maximal when $\rho = 0.5$. Given that each flow sees the same delay, ensuring the system's total utilization is 0.5 can result in delay fairness at maximal sum of power, P^*_{sum}. The maximal power value for FCFS in this case is 0.25, which is the radius of the circle at $(x, y, z) = (2, 1, 0)$, the intersection of P^*_{sum} and delay fairness

at the bottom layer, where the degree of flow discrimination is zero. We see that no other queueing disciplines besides FCFS can achieve this.

9.2.3 Fairness for Average Power $P^*_{\mathbf{avg}}$ Optimization

The average power, P_{avg}, was examined in Chapter 6. It was shown in Corollary 6.3.1 that when the total utilization ρ is 0.5, the average power is optimized in an M/M/1 system for all work-conserving queueing disciplines. Given this condition, we further explore if it is possible to allocate the utilization factor for each flow in a way that ensures fairness in terms of throughput, delay, and power.

The results are presented in the plane of $x = 3$ in Figure 9.1. We show that both power fairness ($y = 3$) and throughput fairness ($y = 2$) can be satisfied while maximizing average power. Importantly, the maximal average power value remains unchanged across all degrees of flow discrimination. However, for delay fairness (along $y = 1$), only the FCFS queueing discipline ($z = 0$) can achieve this while maintaining maximal average power ($x = 3$); this result is similar to the cases for maximal individual power ($x = 1$) and maximal sum of power ($x = 2$).

The following sections discuss the details of each fairness measure under the condition of maximal average power:

9.2.3.1 Throughput Fairness

It is straightforward to achieve throughput fairness as long as each flow equally shares the system utilization, resulting in each flow having a utilization of

$$\rho_1 = \rho_2 = 0.25$$

This holds true across all work-conserving queueing disciplines from HOL to FCFS. Regardless of the queueing discipline, the optimal average power and throughput fairness both yield an average power value of 0.25 (average power $= \rho \cdot (1-\rho) = 0.5 \cdot 0.5 = 0.25$). Consequently, there is a column shown in blue with a uniform radius of 0.25 at $(x, y) = (3, 2)$, the intersection of P^*_{avg} and throughput fairness.

Note that the individual power under this condition where $\rho_1 = \rho_2 = 0.25$ results in different individual power values with different degrees of flow discrimination. In the delay-dependent system, where the degree of flow discrimination is represented by k, substituting $(\rho_1, \rho_2) = (0.25, 0.25)$ into Equation 7.1 yields the corresponding individual powers for flow 1 and flow 2 as a function of k, as follows:

$$P_1 = \frac{(4-k)}{8\,(1-\frac{1}{2}k)}$$

$$P_2 = \frac{1}{8}\,(4-k)$$

In summary, ***throughput fairness*** *can be maintained across all work-conserving queueing disciplines while achieving* ***optimal average power*** $\boldsymbol{P^*_{avg}}$ *in an M/M/1 system with two flows.* However, the individual power values vary depending on the degree of flow discrimination.

9.2.3.2 Power Fairness

Figure 9.1 shows a column with a constant radius size of 0.25 at $(x, y) = (3, 3)$, the intersection of maximal average power and power fairness across the full range of flow discrimination. This visually suggests that ***optimal average power*** $\boldsymbol{P^*_{avg}}$ *and* ***power fairness*** *can be achieved for all work-conserving queueing disciplines in an M/M/1 system with two flows.*

To demonstrate that this is indeed the case, we utilize the delay-dependent system to

derive the utilization factor allocations for ρ_1 and ρ_2 that maximize average power while also ensuring power fairness for flow 1 and flow 2. We start with the power fairness condition given by Equation 8.4, where $\rho_1 = \frac{\rho_2(1-k\rho_2)}{1+k\rho_2}$. Substituting this into the maximal average power condition, $\rho_1 + \rho_2 = 0.5$, we have:

$$\rho_1 + \rho_2 = \frac{\rho_2(1-k\rho_2)}{1+k\rho_2} + \rho_2 = 0.5$$

Solving this gives:

$$\begin{aligned} \rho_1 &= \frac{2-k}{2\,(4-k)} \\ \rho_2 &= \frac{1}{4-k} \end{aligned} \tag{9.2}$$

The corresponding individual powers for flow 1 and flow 2 are:

$$P_1 = P_2 = \frac{k^2 - 4k + 8}{4(4-k)^2} \tag{9.3}$$

For the maximal flow discrimination where $k = 1$, we have $\rho_1 = \frac{1}{6}$ and $\rho_2 = \frac{1}{3}$. The corresponding $P_1 = P_2 = \frac{5}{36}$. For the minimal flow discrimination where $k = 0$, it is the same as in the equal throughput case where $\rho_1 = \rho_2 = 0.25$ and $P_1 = P_2 = 0.125$.

Therefore, Equation 9.2 provides the utilization factor allocation for ρ_1 and ρ_2 that leads to maximal average power and power fairness, where the power of flow 1 and flow 2 is equal to Equation 9.3 in the delay-dependent system. This serves as proof of the existence of maximal average power and power fairness at the same time.

Note that the average power value remains constant at 0.25 for all queueing disciplines for this case of power fairness, which is the same as in the case of throughput fairness. Given

that the maximal average power depends only on total utilization, throughput fairness and power fairness differ in their utilization factor allocations for each flow, but their sums are identical.

9.2.3.3 Delay Fairness

As stated in the case of individual power, no pair (ρ_1, ρ_2) that is larger than 0 can achieve delay fairness unless it is FCFS. Therefore, at the intersection of average power and delay fairness where $(x, y) = (3, 1)$, only the bottom layer has a circle with a radius equal to the maximal average power of 0.25 at $(x, y, z) = (3, 1, 0)$.

9.2.4 Summary

Figure 9.1 illustrates all we need to know about the 9 possible combinations of optimal performance and optimum fairness criteria for all work-conserving queueing disciplines in an M/M/1 system with two flows. The figure represents the intersection points of different performance objectives (optimal individual power P_i^* at $x = 1$, optimal sum of power P_{sum}^* at $x = 2$, and optimal average power P_{avg}^* at $x = 3$) with various fairness metrics (throughput fairness at $y = 2$, delay fairness at $y = 1$, and power fairness at $y = 3$).

Among these 9 intersection points, three cases allow simultaneous achievement of optimal performance and optimum fairness for all work-conserving priority queueing schemes in an M/M/1 system with two flows. These three cases are marked by a column in the figure, which are:

- P_{sum}^* with throughput fairness at $(x, y) = (2, 2)$
- P_{avg}^* with throughput fairness at $(x, y) = (3, 2)$
- P_{avg}^* with power fairness at $(x, y) = (3, 3)$

For the FCFS queueing discipline ($z = 0$), all 9 intersection points are marked with a circle, indicating that under any of the optimal performance metrics (optimal individual power P_i^*, optimal sum of power P_{sum}^*, or optimal average power P_{avg}^*), throughput fairness, delay fairness, and power fairness can all be maintained simultaneously.

9.3 Extending the Analysis to an Arbitrary Number of Flows

Extending the analysis to an arbitrary number of flows adds a new dimension, namely, the number of flows (n), to the interplay between queueing disciplines with various flow discrimination and both optimal performance and fairness. As the number of flows increases, the optimal performance value itself may change, making it dependent on the number of flows (n) and this introduces an additional factor to consider alongside fairness. Additionally, we must consider whether the fairness achievable in a two-flow system (for example, optimal sum of power and throughput fairness in HOL) can still be achieved in an n-flow system.

To facilitate our analysis, we divide our study into three distinct subsections based on the level of flow discrimination exhibited by different queueing disciplines:

- No Flow Discrimination (FCFS)
- Max Flow Discrimination (HOL)
- Intermediate Flow Discrimination (All Other Disciplines)

This approach simplifies our investigation by reducing the number of dimensions considered within each subsection. This categorization is essential because different queueing disciplines introduce varying complexities and analytical scenarios. By isolating and focusing on the key characteristics and trade-offs specific to each case, we can effectively analyze each

before integrating the findings for a broader understanding of the interplay between optimal performance and fairness.

9.3.1 FCFS

9.3.1.1 Fairness for Individual Power P_i^* Optimization

The individual power optimization results for an arbitrary number of flows, summarized in Table 4.1, show that the optimal resource allocation for each flow, $\rho_i{}^* = \frac{1}{n+1}$, results in an optimal individual power of $P_i^* = (\frac{1}{n+1})^2$, leading to throughput fairness and power fairness simultaneously. In addition, delay fairness is also achievable in this scenario because each flow experiences the same response time under FCFS, regardless of power optimization. This means that even after optimizing power, delay fairness is still maintained. These findings lead to the following theorem:

Theorem 9.1.

In an M/M/1 system with n flows using FCFS, maximizing each flow's individual power achieves throughput fairness, power fairness, and delay fairness simultaneously. The resource utilization allocation that achieves these fairness metrics and maximizes individual power is

$$\rho_i{}^* = \frac{1}{n+1} \quad \textit{for } i = 1, 2, \ldots, n \tag{9.4}$$

This behavior is visualized in Figure 9.2 as three cones, each located at the intersection of the "individual power" at $x = 1$ with throughput ($y = 2$), delay ($y = 1$), and power fairness ($y = 3$). Each cone's radius, which represents the achieved performance level (sum of optimized individual power in this case), varies with the number of flows n along the z-axis, becoming smaller as n increases. This consistency is because those states are from the same operating points with resource allocation, $\rho_i{}^* = \frac{1}{n+1}$. The variation in size with the value of n occurs because the sum of maximal individual power, $\sum_{i=1}^{n} P_i^* = \frac{n}{(n+1)^2}$, depends on

the number of flows, n. Figure 9.2 closely resembles Figure 9.1, but with the z-axis now representing the additional dimension of n. Here, we range n from 2 to 100, resulting in a radius of $\frac{2}{9}$ at $z = 0$ ($n = 2$) and a radius of approximately 0.0098 at $z = 1$ ($n = 100$). Therefore, as depicted in the figure, the radius of the circle in each column decreases as the sum of maximal individual power decreases with increasing n.

This analysis highlights the unique characteristics of FCFS in achieving multiple fairness metrics (delay at $y = 1$, throughput at $y = 2$, power at $y = 3$) under individual power optimization, i.e., $x = 1$. However, as we move to more complex queueing disciplines, the relationship between performance, fairness, and the number of flows becomes more intricate, and some fairness properties may disappear.

9.3.1.2 Fairness for Sum of Power P^*_{sum} Optimization

When considering the sum of individual power as the performance optimization metric ($x = 2$), Table 5.1 reveals that the optimal resource allocation is $\rho^* = \frac{1}{2}$, resulting in an optimal sum of power of $P^*_{\text{sum}} = \frac{1}{4}$. Notably, this result remains constant regardless of the number of flows (n) and does not impose any specific restrictions on the individual utilization factors, ρ_i.

To achieve **throughput fairness** ($y = 2$) in this context, we can simply distribute the total utilization, $\rho = \frac{1}{2}$, evenly across the n flows. This leads to a uniform allocation of $\rho_i = \frac{1}{2n}$ for each flow $i = 1, .., n$. Therefore, this demonstrates that throughput fairness is achievable under FCFS when maximizing the sum of power, using this allocation. The resulting optimal sum of power value is $P^*_{\text{sum}} = \frac{1}{4}$. This behavior is visualized in Figure 9.2 at $(x, y) = (2, 2)$, the intersection point of maximal sum of power and throughput fairness, represented by a cylinder with a uniform radius of $\frac{1}{4}$, regardless of the value of n.

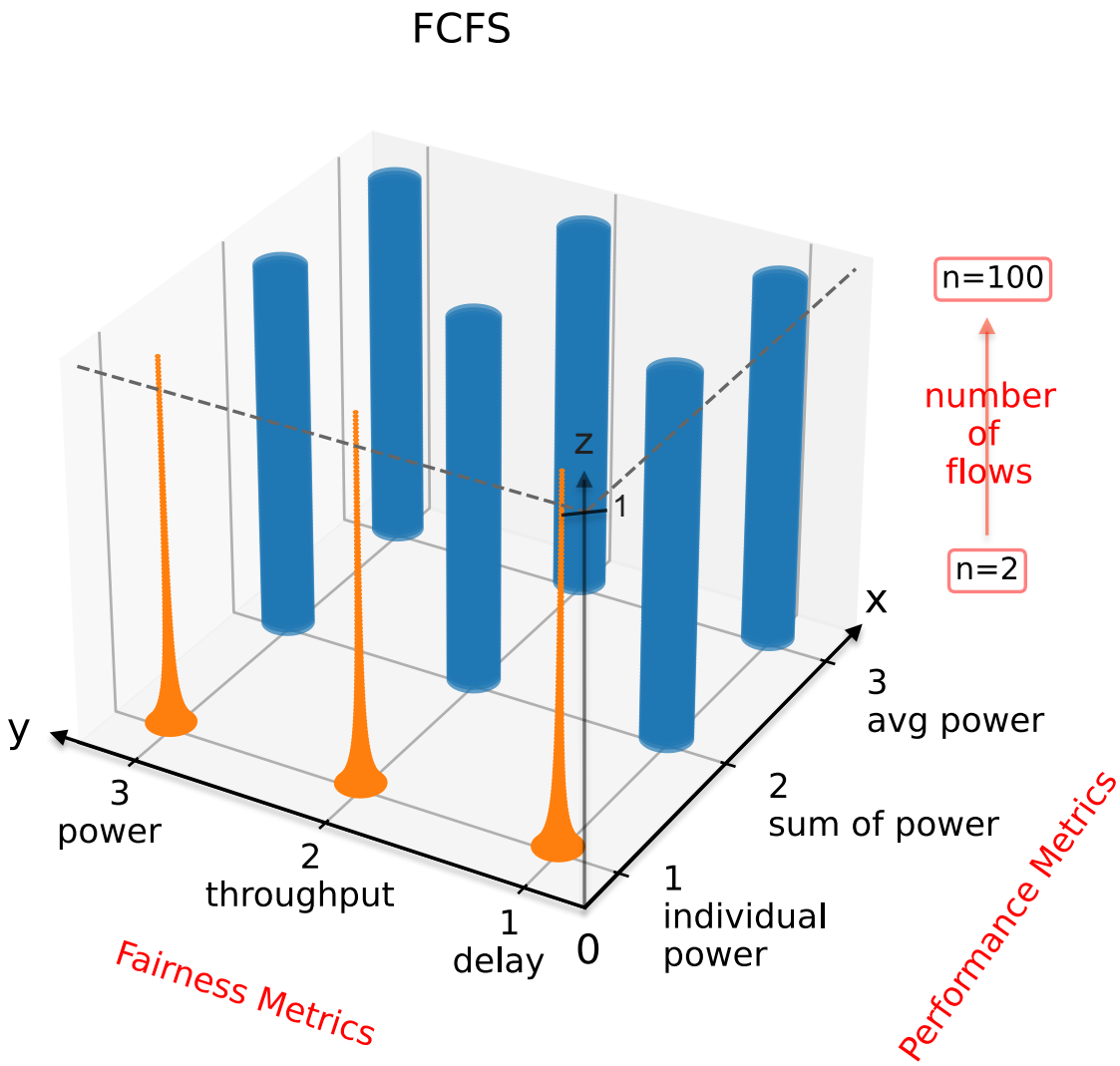


Figure 9.2: A three-dimensional graph of performance (x-axis), fairness (y-axis), and the number of flows n (z-axis), in an M/M/1 systems using **FCFS** (no flow discrimination). The blue columns are all cylinders with a radius of 0.25. The orange columns have a radius of $\frac{2}{9} \approx 0.2222$ at $z = 0$, which gradually decreases to approximately 0.0098 at $z = 1$.

Given that **delay** is equal across all flows under FCFS, even without power optimization, the allocation $\rho_i = \frac{1}{2n}$ for each flow $i = 1, ..n$ also maintains delay fairness ($y = 1$). Furthermore, under this condition, **power fairness** ($y = 3$) is achieved because both throughput and

delay are the same for each flow, and individual power is calculated as the ratio of each flow's utilization factor to its mean normalized delay. The optimal sum of power value in this case, $P^*_{\text{sum}} = \frac{1}{4}$, is identical to the value achieved with throughput fairness, as both scenarios originate from the same utilization factor allocation, $\rho_i = \frac{1}{2n}$ for each flow $i = 1, .., n$. This behavior is visualized in Figure 9.2, where three cylinders along the "sum of power" plane at $x = 2$ are located at the intersection points with each fairness metric (i.e., $(x, y) = (2, 1)$, $(x, y) = (2, 2)$, and $(x, y) = (2, 3)$), each with a uniform radius of $\frac{1}{4}$, regardless of the value of n. These lead to the following theorem:

Theorem 9.2.

In an M/M/1 system with n flows using FCFS, maximizing the sum of power P^*_{sum} simultaneously achieves throughput, delay, and power fairness. This occurs at:

$$\rho^*_i = \frac{1}{2n} \quad \textit{for } i = 1, 2, \ldots, n \tag{9.5}$$

and thus

$$\rho^* = \frac{1}{2} \tag{9.6}$$

The individual power values are:

$$P_i = \frac{1}{4n} \quad \textit{for } i = 1, 2, \ldots, n \tag{9.7}$$

and optimized sum of powers:

$$P^*_{sum} = \frac{1}{4} \tag{9.8}$$

9.3.1.3 Fairness for Average Power P^*_{avg} Optimization

The results of average power optimization for an M/M/1 system with n flows are summarized in Corollary 6.3.1 in Chapter 6. The maximal average power $P^*_{\text{avg}} = \frac{1}{4}$ is achieved

with a total utilization of $\rho = \frac{1}{2}$. Interestingly, this condition is identical to the condition for optimal sum of power, P^*_{sum}.

Therefore, to achieve throughput fairness, we can apply the same strategy as in the sum of power case: distribute the system utilization $\rho = \frac{1}{2}$ uniformly across the n flows, resulting in $\rho_i = \frac{1}{2n}$ for each flow $i = 1, .., n$. This allocation leads to simultaneous throughput, delay, and power fairness, as demonstrated in the earlier discussion of sum of power and fairness. This outcome is visualized in Figure 9.2, where three cylinders are located at the intersection points on the "average power" plane at $x = 3$ with each fairness metric (i.e., $(x, y) = (3, 1)$, $(x, y) = (3, 2)$, and $(x, y) = (3, 3)$), each with a uniform radius of $\frac{1}{4}$, regardless of the value of n. This behavior exhibits the same pattern as seen with the sum of power on the plane at $x = 2$. Thus, we have the following theorem:

Theorem 9.3.

In an M/M/1 system with $\boldsymbol{n}$ *flows using* ***FCFS****, the optimized utilization factor allocation*

$$\rho_i^* = \frac{1}{2n} \quad \textit{for } i = 1, \ldots, n \tag{9.9}$$

results in both ***optimal sum of power and optimal average power****, while simultaneously achieving* ***throughput, delay, and power fairness****. The optimal values for both sum of power and average power are*

$$P^*_{sum} = P^*_{avg} = \frac{1}{4} \tag{9.10}$$

9.3.2 HOL

9.3.2.1 Fairness for Individual Power P_i^* Optimization

Given that maximizing individual power in HOL fails to achieve fairness in either throughput, delay, or power in the two-flow case, we can infer that this remains unachievable when extended to n flows. Table 4.1 serves as proof of this: the optimal utilization factor is ${\rho_i}^* = (\frac{1}{2})^i$ and the optimal individual power is ${P_i}^* = 2(\frac{1}{8})^i$. Consequently, the normalized mean delay, $\mu T_i = \frac{\rho_i}{P_i}$, becomes $2^{(2i-1)}$.

Since these values for ρ_i, P_i, and μT_i all differ for each flow i, it's clear that achieving throughput fairness, delay fairness, and power fairness is impossible when maximizing individual power in HOL with n flows.

Therefore, along the "individual power" plane of $x = 1$ in Figure 9.3, there is nothing at the intersection points with the fairness metrics — throughput fairness at $(x, y) = (1, 2)$, delay fairness at $(x, y) = (1, 1)$, and power fairness at $(x, y) = (1, 3)$ — indicating that none of these fairness metrics can be achieved when maximizing individual power.

9.3.2.2 Fairness for Sum of Power $P_{\mathbf{sum}}^*$ Optimization

Theorem 5.4 in Chapter 5 states that the optimal sum of power, P_{sum}^*, is achieved when each flow's utilization factor is $\rho_i^* = \frac{1}{n+1}$ for $i = 1, .., n$. This theorem demonstrates that ***maximizing the sum of individual power can be achieved simultaneously with throughput fairness in an M/M/1 system using HOL with n flows***. The maximal sum of power is $P_{\text{sum}}^* = \frac{n(n+2)}{3(n+1)^2}$. This outcome is represented by a tapered cylindrical shape at the intersection of "sum of power" at $x = 2$ and throughput fairness at $y = 2$, i.e., $(x, y) = (2, 2)$ in Figure 9.3, with the radius gradually increasing along the z-axis. At

the bottom ($z = 0$), where $n = 2$, the radius is approximately 0.296, and it increases to approximately 0.3333 at the top ($z = 1$), where $n = 100$.

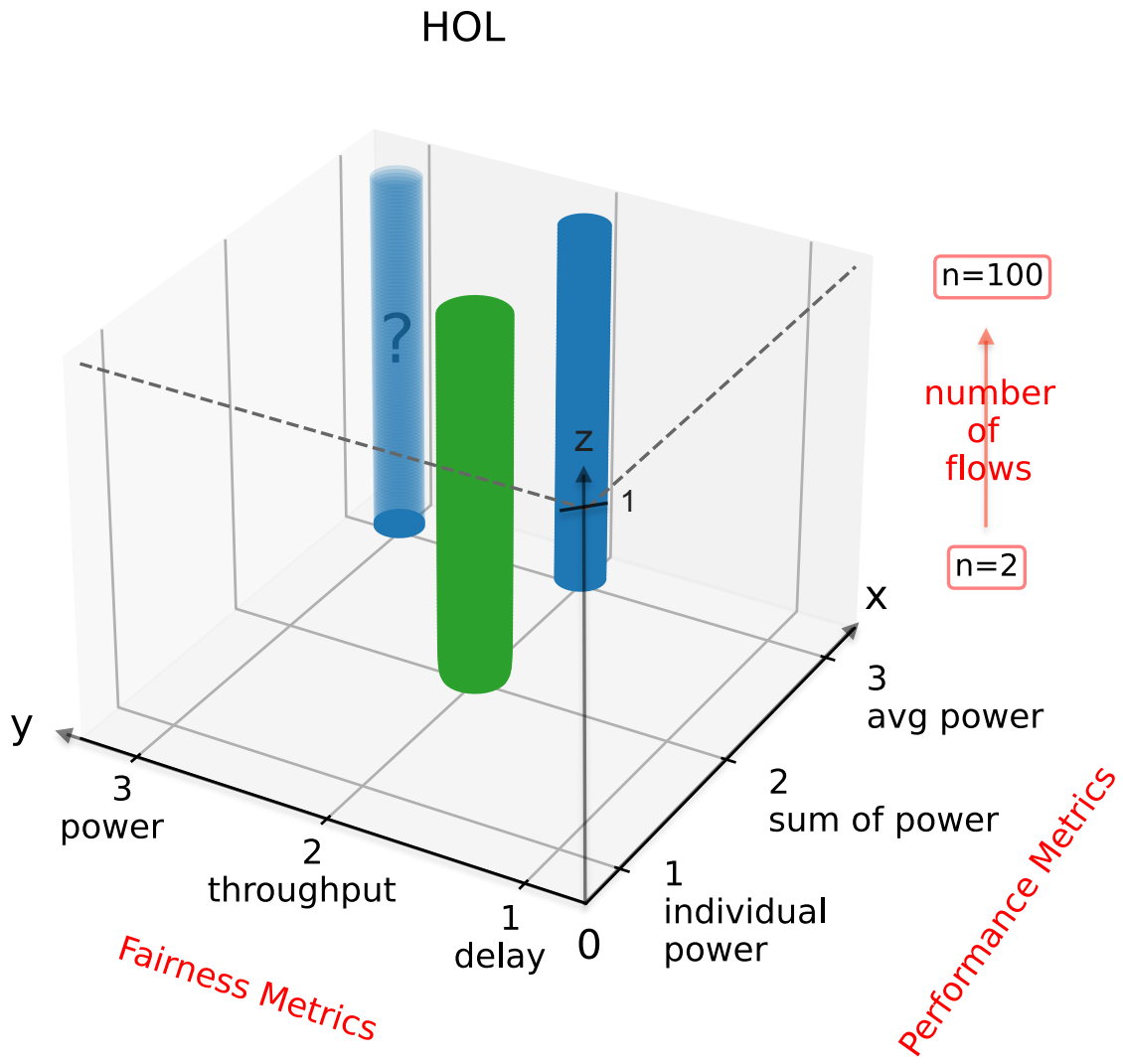


Figure 9.3: A three-dimensional graph of performance, fairness, and the number of flows, n, in an M/M/1 system using **HOL** (max flow discrimination). The green column at $(x, y) = (2, 2)$ has a radius of approximately 0.296 at $z = 0$, which gradually increases to approximately 0.3333 at $z = 1$. Two blue columns along $x = 3$ (average power) are cylinders with a radius of 0.25. One column is opaque at $(x, y) = (3, 2)$, indicating certainty of its existence, while the other is semi-transparent at $(x, y) = (3, 3)$, representing uncertainty about its existence.

However, Theorem 5.5 reveals that when the sum of power is maximized, individual power, $P_i = \frac{(n+1-i)(n+2-i)}{(n+1)^3}$ for $i = 1, 2, ..., n$, is not uniform across flows, varying with the flow index, i. This indicates that power fairness is not achievable when optimizing the sum of power, P^*_{sum}. Furthermore, as previously discussed, delay fairness is only achievable under FCFS. Therefore, at those two intersection points in Figure 9.3—one at $(x, y) = (2, 3)$, the intersection of optimal sum of power with power fairness, and the other at $(x, y) = (2, 1)$, the intersection of optimal sum of power with delay fairness—there is no solution. This indicates that *achieving both optimal performance (maximal sum of power P^*_{sum}) and fairness for either delay or power is not possible in HOL.*

9.3.2.3 Fairness for Average Power P^*_{avg} Optimization

Corollary 6.3.1 shows that for an M/M/1 system with any work-conserving queueing discipline, maximizing average power occurs when the total utilization ρ is 0.5. Under this condition, **throughput fairness** is achievable by uniformly distributing the system utilization, $\rho = 0.5$, across the n flows, resulting in $\rho_i = \frac{1}{2n}$ for each flow $(i = 1, .., n)$. This outcome exhibits the same pattern as observed for average power and equal throughput in the FCFS case.

In Figure 9.3, the cylinder at $(x, y) = (3, 2)$, the intersection of optimal average power and throughput fairness has a uniform radius corresponding to the maximal average power value of $P^*_{\text{avg}} = \frac{1}{4}$. This demonstrates that *maximal average power and throughput fairness can be achieved at the same time for an arbitrary number of flow using HOL in an M/M/1 system.* This cylinder is also identical to the one at the same intersection point, i.e., $(x, y) = (3, 2)$, in Figure 9.2 for the FCFS case.

Since HOL is not FCFS, **delay fairness** is not achievable under optimal average power. Therefore, there is no solution at $(x, y) = (3, 1)$, the intersection of the optimal average power

and delay fairness in Figure 9.3.

The achievability of **power fairness** under the constraint of $\rho = 0.5$ for maximizing average power remains unclear. While it has been shown that equal power can be achieved in the two-flow case, the situation becomes more complex with additional flows. As more flows compete for the system utilization factor, each flow's impact on the overall system balance becomes significant if power fairness is enforced. For $n = 3$, we can find a set of solutions for ρ_1, ρ_2, and ρ_3 that lead to power fairness while maintaining maximal average power. The solution is $(\rho_1, \rho_2, \rho_3) \approx (0.106, 0.141, 0.252)$. However, these values do not follow a clear pattern that can be easily extended to an arbitrary number of flows. Therefore, it remains uncertain whether a set of ρ_i for $i = 1, \ldots, n$ exists for equal power and $\rho = 0.5$ when $n > 3$.

To represent this uncertainty, we use a semi-transparent blue cylinder with a question mark inside at $(x, y) = (3, 3)$, the intersection of optimal average power and power fairness in Figure 9.3. If such a solution exists, the maximal average power value would be 0.25. Since the solution exists for $n = 2$, at the bottom of the intersection, i.e., $(x, y, z) = (3, 3, 0)$, the blue color is opaque, not semi-transparent.

9.3.3 Intermediate Queueing Disciplines

All other queueing disciplines exhibit flow discrimination, which falls somewhere between the extremes of minimum (FCFS) and maximum (HOL) discrimination. These disciplines, which we will refer to as "intermediate queueing disciplines," do not include the minimum and maximum points of this range.

The performance and fairness results for a system using intermediate queueing disciplines are presented in Figure 9.4. Based on our current understanding, the existence of some results

remains uncertain. In this figure, we observe that among the 9 intersection points, 6 do not exist in their corresponding performance and fairness metrics, and nothing is depicted at those points. We assume non-existence at these locations because they do not exist in the HOL case in Figure 9.3 for the same locations. One intersection point features a opaque blue cylinder at $(x, y) = (3, 2)$, the intersection of optimal average power and throughput fairness, representing the only result we are certain of.

The other two cylinders with semi-transparent colors, located at $(x, y) = (2, 2)$, the intersection of optimal sum of power with throughput fairness, and at $(x, y) = (3, 3)$, the intersection of average power with power fairness, include a question mark to indicate uncertainty regarding their existence[9]. Only at $n = 2$ (where $z = 0$) are we certain of the existence of a solution, and for those two intersection points at the bottom layer, the circles are in an opaque color.

9.3.3.1 Fairness for Individual Power P_i^* Optimization

When maximizing individual power for intermediate queueing disciplines, fairness in terms of throughput, power, and delay are all unattainable. Even with just two flows, fairness cannot be achieved at the equilibrium point of optimizing individual power. This difficulty is exacerbated in systems with more flows, as each additional flow adds another constraint by optimizing its own power, making it even harder to achieve fairness. Therefore, along the individual power plane of $x = 1$ in Figure 9.4, there are no solutions.

[9] Note that for the semi-transparent green column at $(x, y) = (2, 2)$, we set the radius to P^*_{sum} in HOL as an upper bound. Yet, the actual value of P^*_{sum} depends on the specific queueing discipline used and will be smaller than that in HOL.

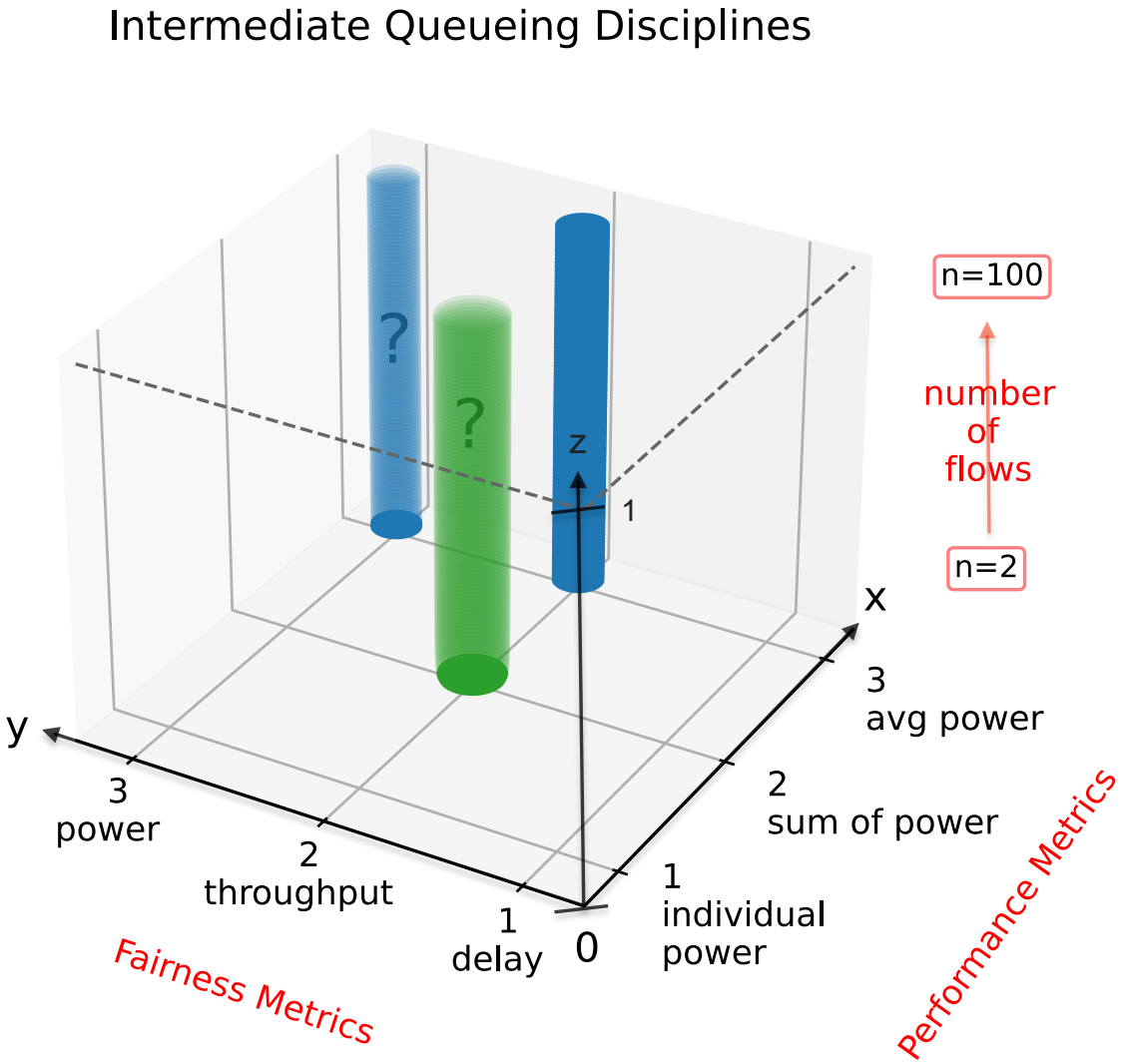


Figure 9.4: A three-dimensional graph of performance, fairness, and the number of flows, n, in an M/M/1 system using **intermediate queueing disciplines** with flow discrimination ranging between FCFS and HOL, but not including the two extremes. One opaque blue column at $(x, y) = (3, 2)$ is a cylinder with radius = 0.25. Two semi-transparent columns indicate uncertainty about their existence. One is a semi-transparent blue cylinder with a radius of 0.25 at $(x, y) = (3, 3)$, and the other is a semi-transparent green cylinder at $(x, y) = (2, 2)$ with a radius set to the upper bound of P^*_{sum} in HOL.

9.3.3.2 Fairness for Sum of Power P^*_{sum} Optimization

Along the sum of power line where $x = 2$, the intersections with delay fairness ($y = 1$) and power fairness ($y = 3$) both have no solution. The non-existence of delay fairness is evident

because the queueing disciplines here involve flow discrimination, leading to different response times for each flow. The non-existence of power fairness is inferred from the non-existence of power fairness in the HOL case, suggesting that power fairness cannot be achieved under the optimal sum of power for intermediate queueing disciplines.

For throughput fairness of $y = 2$, we are uncertain, so we used a lightly colored cylinder with a question mark inside at the intersection of sum of power and throughput fairness, i.e., $(x, y) = (2, 2)$ in Figure 9.4. We reserve the possibility of existence because it is shown that throughput fairness can be maintained while maximizing the sum of power in HOL and FCFS. Additionally, using a delay-dependent queueing discipline in a system with two flows, throughput fairness can be satisfied. Therefore, the green color is darker at the bottom layer of $(x, y) = (2, 2)$, corresponding to $(x, y, z) = (2, 2, 0)$.

Attempt to Find a solution at $(x, y) = (2, 2)$ in Three-Flow Systems Using the delay-dependent

Given solutions exist for $(x, y) = (2, 2)$ (maximal sum of power and throughput fainess) when $n = 2$ in the delay-dependent system for all queueing disciplines between FCFS to HOL (where any feasible combination of (b_1, b_2) has a solution), we attempt to extend the analysis to $n = 3$ in the same system. We test various combinations of (b_1, b_2, b_3) and apply a numerical approach to find the maximal sum of power for each set. Among the combinations tested, the resulting optimized utilization factors for ρ_1, ρ_2, and ρ_3 are all different.

For example, in a delay-dependent system with three flows using the parameters $(b_1, b_2, b_3) = (1, 2, 4)$, the maximal sum of power occurs at the operating point of $(\rho_1, \rho_2, \rho_3) \approx (0.27, 0, 0.27)$. Another example where the optimal (ρ_1, ρ_2, ρ_3) are different is under $(b_1, b_2, b_3) = (0, 1, 1)$. This results in the maximal sum of power at the operating point of $(\rho_1, \rho_2, \rho_3) = \left(\frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right)$.

These examples show that throughput fairness cannot be achieved for delay-dependent systems with arbitrary feasible combinations of b_1, b_2, and b_3 where $0 \leq b_1 \leq b_2 \leq b_3$ in a system with three flows, let alone for an arbitrary number of flows. Extending this to an arbitrary number of flows further complicates the analysis. Throughput fairness exists when $b_1 = b_2 = \ldots = b_n$, which is equivalent to the FCFS queueing discipline, and when $b_1 \ll b_2 \ll \ldots \ll b_n$, which corresponds to the HOL case. Aside from these two, we have not identified any other combination of $b_1, b_2, .., b_n$ that achieves throughput fairness.

Although we were unable to find a solution in the set tested for the delay-dependent systems with three flows, we also could not demonstrate non-existence. Given that throughput fairness is achievable for the maximum discrimination queueing discipline (HOL), there is potential for a family of queueing disciplines representing flow discrimination between FCFS and HOL to exist. The possibility of achieving throughput fairness while maximizing the sum of power for other such families of queueing disciplines remains an open question, requiring further investigation in future work.

Observations from Counter-Examples in Three-Flow Systems

Note that the counter-examples used earlier to demonstrate the existence of throughput fairness in a three-flow delay-dependent system do not hold when maximizing the sum of power. Yet, these examples lead to interesting observations.

In the first example where $(b_1, b_2, b_3) = (1, 2, 4)$, the optimal allocation results in $\rho_1 \approx \rho_3$, while $\rho_2 \approx 0$. This suggests that flow 2 may be experiencing starvation, although the reasons for this behavior are not immediately clear.

Furthermore, in the second example where $(b_1, b_2, b_3) = (0, 1, 1)$, we find that $\rho_2 = \rho_3$ and $\rho_2 + \rho_3 = \rho_1$. This configuration exhibits a hierarchical equal throughput scenario, where flows 2 and 3 are treated as a single group, sharing the same utilization factor. Importantly, this combined utilization for flows 2 and 3 is equal to the utilization factor for flow 1. This hierarchical structure implies that while individual flows within a group might not experience throughput fairness, the combined throughput of each group is equal.

9.3.3.3 Fairness for Average Power P^*_{avg} Optimization

The existence of throughput fairness and non-existence of delay fairness under maximal average power P^*_{avg} using intermediate queueing disciplines are straightforward. Achieving throughput fairness is simple by uniformly distributing the system utilization $\rho = \frac{1}{2}$ across n flows. Thus, there is a cylinder with a radius of 0.25 at the intersection of average power and throughput fairness at $(x, y) = (3, 2)$ in Figure 9.4. For delay fairness, it has been previously stated that each flow will have different response times due to flow discrimination. Thus, no solution is provided at $(x, y) = (3, 1)$, the intersection of optimal average power and delay fairness in Figure 9.4.

The existence of power fairness and optimal average power for an arbitrary number of flows using intermediate queueing disciplines in a system remains unclear. This uncertainty is inherited from the HOL case, where the form of the response time for each flow is relatively simple but has the maximum degree of flow discrimination. Given that power fairness exists for a two-flow system using delay-dependent queueing disciplines, we believe there is potential for the existence of this power fairness and optimal average power for a system with n flows, but we are unsure. To represent this uncertainty, we use a lightly colored cylinder with a radius of 0.25, indicating the maximal average power value if such existence is confirmed, and a question mark inside at the intersection of average power and power fairness at $(x, y) = (3, 3)$

in Figure 9.4.

9.3.4 Summary

Figures 9.2, 9.3, and 9.4 show **all possible and impossible results** for **optimal performance and optimum fairness** in a system with an arbitrary number of flows under different queueing disciplines. They highlight the values that are known and those that remain uncertain. Specifically, while we have determined the values for some columns, the realizability of certain columns marked with a "?" is still unknown.

In the FCFS case, where there is no flow discrimination, fairness can be maintained in the context of throughput, delay, and power while optimizing the performance metrics in terms of individual power, sum of power, and average power. Thus, at the nine intersection points in 9.2, each intersection point has a cylinder or a cone. Among these performance metrics, the individual power case indicates diminishing performance value as n increases, leading to a decrease in radius at the corresponding intersection points, thereby forming cone-like shapes. On the other hand, the performance values of the other two metrics remain invariant to n, which is represented by a cylinder.

In the HOL case, which is the case of maximal flow discrimination, the number of points where optimal performance and optimum fairness can both be achieved reduces to three, with one point uncertain regarding the existence of a solution, as shown in Figure 9.3. This reduction is in contrast to the FCFS case, where are all nine intersection points are achievable. The other six points, out of a total of nine intersection points, are confirmed to have no solution. The remaining two points, which are certain to have a solution, are located $(x, y) = (2, 2)$ and $(x, y) = (3, 2)$, corresponding to the sum of powers with throughput fairness and average power with throughput fairness, respectively. A notable property is that the optimal sum of

power values increases with n, while throughput fairness is still maintained, as represented by a cylinder with a radius increasing along z-axis at $(x, y) = (2, 2)$ in Figure 9.3. This behavior is not observed in the FCFS case, where the performance metric values, represented by the radius of the circle in each z-axis layer, either decrease or remain unchanged along the z-axis. This distinct characteristic of the HOL case highlights its unique advantage in maintaining throughput fairness while improving the sum of power as the number of flows increases.

For other intermediate disciplines with flow discrimination ranging between the maximum and minimum (but not including the two extremes), Figure 9.4 reveals a similar pattern to the HOL case in Figure 9.3, but with a second uncertainty. An interesting characteristic in HOL is that the optimal sum of power P^*_{sum} and throughput fairness can be achieved simultaneously. However, this property does not necessarily extend to intermediate queueing disciplines. The complexity of the power function formulas (stemming from the complicated response time formula) for these disciplines currently prevents us from proving the existence or non-existence of such a solution. Even though there are two uncertainties and six points confirmed for non-existence in other intermediate disciplines, one column remains certain: the intersection of maximal average power and throughput fairness, marked by an opaque blue column at $(x, y) = (3, 2)$ in our analysis. This intersection point demonstrates the simultaneous achievement of optimal average power P^*_{avg} and throughput fairness.

Chapter 10: Generalized Power Analysis

In the previous chapters, we defined three power metrics for a system with multiple flows. These metrics follow the pattern of normalized throughput over mean normalized delay, but in different variations. The power metric can be further extended to a generalized form to account for the preference of throughput versus delay. In this chapter, we introduce the concept of generalized power, summarize previous work on it, and extend this analysis to a system with multiple flows.

10.1 Generalized Power

Different applications may have varying tradeoffs between throughput and delay, which may not align with the power definition in Equation 2.1 that treats them equally. For example, interactive data may tolerate some loss of throughput, but the delay should be kept minimal. Conversely, for data transfer, throughput might be the primary concern, while delay is less critical. In [59], it is shown that voice, video, and data have different characteristics in terms of data rate and delay. To account for these diverse requirements, it is useful to adopt a generalized power definition. This extension allows the analyst to prioritize either throughput or delay, depending on the application's needs.

10.1.1 Definition

In [5], Kleinrock introduced a generalized definition of Power, P^G, by placing non-negative real variable r into the formula:

$$P^G = \frac{\rho^r}{\mu T} \tag{10.1}$$

The value of r determines the preference between throughput and delay:

- For $r > 1$, it represents a preference for throughput over delay.
- For $r < 1$, it indicates a favoring of delay over throughput.
- For $r = 1$, throughput and delay have equal importance, corresponding to the definition used in our earlier chapters.

Note that we use the superscript G in P^G to indicate the form of generalized power, as the subscript is reserved for P_i (individual power for flow i), P_{sum} (sum of power), and P_{avg} (average power).

Using this definition, the optimal operating point that maximizes generalized power can be discovered through the following condition [7]:

$$\left.\frac{d\mu T(\rho)}{d\rho}\right|_{\rho=\rho^*} = r \cdot \left.\frac{\mu T(\rho)}{\rho}\right|_{\rho=\rho^*} \tag{10.2}$$

where at the optimal operating point ρ^*, the slope of the normalized mean response time function at that point is r times the slope of a line from the origin to the point $[\rho^*, \mu T(\rho^*)]$.

In an M/M/1 system, the generalized power is given by:

$$P_G = \rho^r \cdot (1 - \rho) \tag{10.3}$$

Kleinrock derived the optimal operating point that maximizes this generalized power [5, 7]:

$$\rho^* = \frac{r}{r+1} \tag{10.4}$$

and under this operating point, the optimized average number in the system is:

$$\overline{N^*} = r \tag{10.5}$$

For $r = 1$, this corresponds to the definition in Equation 2.1, leading to $\rho^* = \frac{1}{2}$ and $\overline{N} = 1$, as presented in Chapter 2.

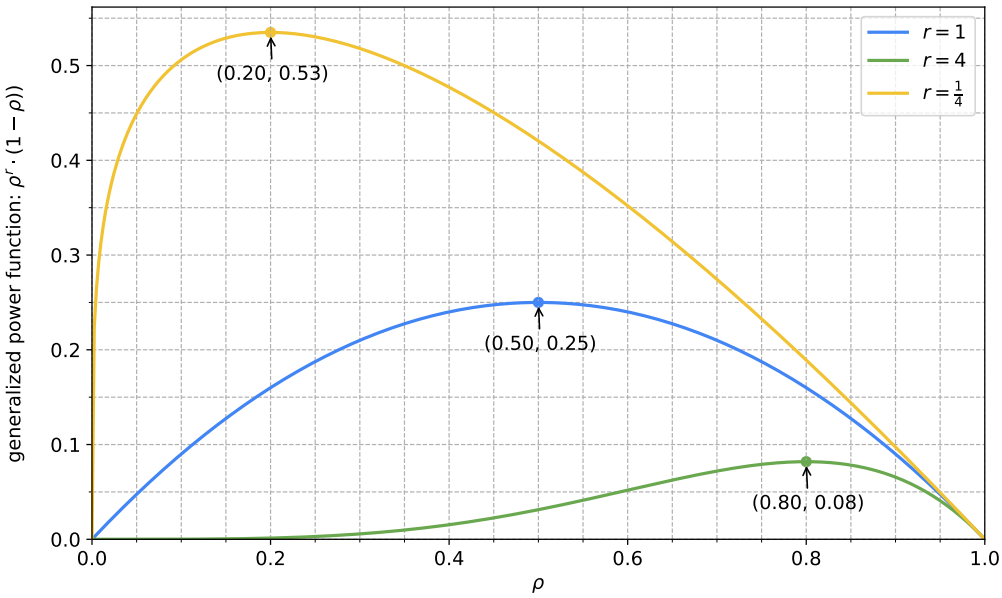


Figure 10.1: The generalized power function, $P^G = \rho^r(1 - \rho)$ for different values of r ($r = 1$, $r = 4$, $r = \frac{1}{4}$). The maximal value of power occurs at $\rho^* = \frac{r}{r+1}$ for each corresponding value of r.

Figure 10.1 illustrates the generalized power function, given by Equation 10.3, for different values of r: $r = 1$, $r = 4$, and $r = 1/4$. The figure shows that the maximal power for each power function occurs at $\rho^* = \frac{r}{r+1}$. With larger r, the optimal ρ^* will be larger, indicating a preference for throughput, while smaller r results in a smaller ρ^* and thus aslo a smaller response time.

By incorporating this generalized power definition, analysts can better tailor system performance to meet specific application requirements, balancing throughput and delay according to the needs of different data types.

10.2 Generalized Power in Systems with Multiple Flows

We now extend the generalized power into our analysis of multiple flows. We first look at individual power optimization.

10.2.1 Individual Power Optimization

Suppose each flow i has its own preference variable r_i. Then individual generalized power for flow i is:

$$P_i^G = \frac{{\rho_i}^{r_i}}{\mu T_i} \tag{10.6}$$

10.2.1.1 FCFS

In FCFS, each flow's normalized mean response time is $\mu T_i = \frac{1}{1-\rho}$. Thus, generalized individual power in FCFS is:

$$P_i^G = {\rho_i}^{r_i} \cdot (1-\rho) = {\rho_i}^{r_i} \cdot (1 - \rho_i - \sum_{j=1, j\neq i}^{n} \rho_j) \tag{10.7}$$

Taking the partial derivative of this function with respect to ρ_i and setting it to zero:

$$\frac{\partial P_i^G}{\partial \rho_i} = r_i \cdot {\rho_i}^{r_i - 1} \cdot (1 - \rho_i - \sum_{j=1, j \neq i}^{n} \rho_j) - {\rho_i}^{r_i} = 0$$

Leading to:

$$r_i \cdot (1 - \rho_i - \sum_{j=1, j \neq i}^{n} \rho_j) - \rho_i \; = \; 0 \tag{10.8}$$

This can be expressed as:

$$\rho_i = r_i \cdot (1 - \rho_i - \sum_{j=1, j \neq i}^{n} \rho_j) \tag{10.9}$$

This gives us:

$$\rho_i = \frac{r_i}{r_i + 1} \cdot (1 - \sum_{j=1, j \neq i}^{n} \rho_j) \tag{10.10}$$

This applies to all flows for $i = 1, .., n$. Equation 10.10 implies that each flow takes the $\frac{r_i}{r_i+1}$ fraction of the available utilization left over by other flows in optimizing its generalized individual power. This is similar to the idea discussed in Chapter 4, but with a different fraction to take for this generalized power definition. When each $r_i = 1$ for $i = 1, \ldots, n$, the fraction is $\frac{1}{2}$, as mentioned in Chapter 4, where the optimal individual power is achieved by taking half of the remaining utilization.

To find the equilibrium point when each flow optimizes its own generalized power, we solve the n partial derivatives and form the sum $\rho = \sum_{i=1}^{n} \rho_i$.
From Equation 10.9, we can rewrite ρ_i as:

$$\rho_i = r_i \cdot (1 - \rho) \tag{10.11}$$

Summing Equation 10.11 over all flows $i = 1, .., n$, we obtain:

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

Solving for ρ, we get optimized total utilizations ρ^*:[1]

$$\rho^* = \frac{\sum_{j=1}^{n} r_j}{1 + \sum_{j=1}^{n} r_j} \tag{10.12}$$

Each individual optimized utilization factor is:

$$\rho_i^* = \frac{r_i}{1 + \sum_{j=1}^{n} r_j} \quad \text{for} \quad i = 1, \ldots, n \tag{10.13}$$

This shows that the optimal results in terms of each flow's utilization factor are proportional to their r_i. The ratios of ρ_i^* depend on the ratio of r_i.

Optimized individual generalized power of each flow i is derived using Equation 10.12 and 10.13:

$$P_i^{G*} = {\rho_i}^{r_i}(1 - \rho) = \frac{r_i}{1 + \sum_{j=1}^{n} r_j} \cdot (1 - \frac{\sum_{j=1}^{n} r_j}{1 + \sum_{j=1}^{n} r_j}) = \frac{r_i}{1 + \sum_{j=1}^{n} r_j} \cdot \frac{1}{1 + \sum_{j=1}^{n} r_j}$$

This leads to:

$$P_i^{G*} = \frac{r_i}{(1 + \sum_{j=1}^{n} r_j)^2} \quad \text{for} \quad i = 1, \ldots, n$$

This also shows that optimized individual generalized power of each flow is proportional to r_i as the denominator for each flow i is the same.

[1] To avoid confusion with the index of the specific flow, i, we use the index j for both the numerator and denominator in the summation.

The sum of optimized individual generalized power is thus:

$$P_{\text{sum}}{}^{G*} = \sum_{i=1}^{n} P_i^{G*} = \frac{\sum_{i=1}^{n} r_i}{(1 + \sum_{j=1}^{n} r_j)^2}$$

These results can be summarized into the following theorem:

Theorem 10.1.

*In an M/M/1 system with n flows using **FCFS**, where each flow i has a preference variable r_i reflecting its relative preference for throughput versus delay, the equilibrium point that results when each flow individually optimizes its generalized power is characterized by:*

$$\rho_i^* = \frac{r_i}{1 + \sum_{j=1}^{n} r_j} \quad \textit{for} \quad i = 1, \ldots, n \tag{10.14}$$

This leads to a optimized total system utilization of:

$$\rho^* = \sum_{i=1}^{n} \rho_i = \frac{\sum_{j=1}^{n} r_j}{1 + \sum_{j=1}^{n} r_j} \tag{10.15}$$

The optimized individual generalized power for each flow is:

$$P_i^{G*} = \frac{r_i}{(1 + \sum_{j=1}^{n} r_j)^2} \quad \textit{for} \quad i = 1, \ldots, n \tag{10.16}$$

The sum of these optimized individual generalized powers is:

$$P_{sum}{}^{G} = \sum_{i=1}^{n} P_i^{G*} = \frac{\sum_{i=1}^{n} r_i}{(1 + \sum_{j=1}^{n} r_j)^2} \tag{10.17}$$

When $r_i = 1$ for $i = 1, \ldots, n$, this case was addressed in earlier analysis, and the optimization results for ρ_i^*, ρ^*, ${P_i}^{G^*}$, and ${P_{\text{sum}}}^G = \sum_{i=1}^{n} {P_i}^{G^*}$ correspond to those derived in Chapter 4 and are summarized in the FCFS column of Table 4.1.

10.2.1.2 HOL

For HOL, each flow's normalized mean response time is given by $\mu T_i = \frac{1}{(1-\sigma_i)(1-\sigma_{i-1})}$ from Equation 3.5, where $\sigma_i = \sum_{j=1}^{i} \rho_j$. This leads to the following expression for the individual generalized power in HOL:

$$P_i^G = {\rho_i}^{r_i} \cdot (1-\sigma_i)(1-\sigma_{i-1}) = {\rho_i}^{r_i} \cdot (1 - \rho_i - \sum_{j=1}^{i-1} \rho_j)(1 - \sum_{j=1}^{i-1} \rho_j) \tag{10.18}$$

To find the optimal ρ_i, we take the partial derivative of this function with respect to ρ_i and set it to zero:

$$\frac{\partial P_i^G}{\partial \rho_i} = (1 - \sum_{j=1}^{i-1} \rho_j) \cdot \frac{\partial}{\partial \rho_i} {\rho_i}^{r_i} \cdot (1 - \rho_i - \sum_{j=1}^{i-1} \rho_j) = 0$$

Leading to:

$$r_i \cdot {\rho_i}^{r_i - 1} \cdot (1 - \rho_i - \sum_{j=1}^{i-1} \rho_j) - {\rho_i}^{r_i} = {\rho_i}^{r_i - 1} \cdot [r_i \cdot (1 - \rho_i - \sum_{j=1}^{i-1} \rho_j) - \rho_i] = 0$$

Solving for the optimized ρ_i, we obtain:

$$\rho_i^* = \frac{r_i}{r_i + 1} \cdot (1 - \sum_{j=1}^{i-1} \rho_j)$$

This can be expressed more compactly as:

$$\rho_i^* = \frac{r_i}{r_i + 1} \cdot (1 - \sigma_{i-1}) \tag{10.19}$$

This result implies that the optimal individual utilization factor, ρ_i^*, takes a fraction of $\frac{r_i}{r_i+1}$ from the remaining utilization after higher-priority flows have taken their shares. While this is similar to the results derived in Chapter 4, the fraction is now variable and depends on the set of r_i values for $i = 1, .., n$, as opposed to the fixed value of $\frac{1}{2}$ derived in Chapter 4, which corresponds to the case of $r_i = 1$ for $i = 1, .., n$.

Substituting $i = 1$ into Equation 10.19, we obtain:

$$\rho_1^* = \frac{r_1}{r_1 + 1} \tag{10.20}$$

and for $i = 2$, we have:

$$\rho_2^* = \frac{r_2}{(r_2 + 1)} \cdot (1 - \sigma_1) = \frac{r_2}{(r_2 + 1) \cdot (r_1 + 1)}$$

To find a general expression for ρ_i in terms of the r_i values, we use induction to first prove the following:

$$1 - \sigma_i = \frac{1}{\Pi_{j=1}^{i}(r_j + 1)} \tag{10.21}$$

Proof:

For the base case of $i = 1$, we have:

$$1 - \sigma_1 = 1 - \rho_1 = 1 - \frac{r_1}{r_1 + 1} = \frac{1}{r_1 + 1}$$

This result is consistent with Equation 10.21 with $i = 1$.

Assuming the equation holds for $i = k$, we have:

$$1 - \sigma_k = \frac{1}{\Pi_{j=1}^{k}(r_j + 1)}$$

Now, for $i = k + 1$, we need to show:

$$1 - \sigma_{k+1} = \frac{1}{\Pi_{j=1}^{k+1}(r_j + 1)}$$

We prove this as follows:

$$\begin{aligned}
1 - \sigma_{k+1} &= 1 - \sigma_k - \rho_{k+1} \\
&= 1 - \sigma_k - \frac{r_{k+1}}{r_{k+1} + 1} \cdot (1 - \sigma_k) \\
&= (1 - \sigma_k) \cdot (1 - \frac{r_{k+1}}{r_{k+1} + 1}) \\
&= (1 - \sigma_k) \cdot (\frac{1}{r_{k+1} + 1}) \\
&= \frac{1}{\Pi_{j=1}^{k}(r_j + 1)} \cdot (\frac{1}{r_{k+1} + 1}) \\
&= \frac{1}{\Pi_{j=1}^{k+1}(r_j + 1)}
\end{aligned}$$

Therefore, the formula holds for $i = k + 1$. By induction, Equation 10.21 is true for all $i = 1, .., n$.

■

Substituting Equation 10.21 into Equation 10.19, we have:

$$\rho_i = \frac{r_i}{r_i + 1} \cdot \frac{1}{\Pi_{j=1}^{i-1}(r_j + 1)}$$

This leads to:

$$\rho_i^* = \frac{r_i}{\Pi_{j=1}^{i}(r_j + 1)} \quad \text{for } i = 1, 2, \ldots, n \tag{10.22}$$

This result shows that the optimal individual ρ_i^*, is solely determined by the r_i values of equal or higher-priority flows.

The optimum system utilization ρ^* is then:

$$\rho^* = \sum_{i=1}^{n} \rho_i^* = \sum_{i=1}^{n} \frac{r_i}{\Pi_{j=1}^{i}(r_j + 1)} \tag{10.23}$$

For the optimized individual generalized power in HOL, we compute with Equation 10.21 and 10.22:

$$P_i^{G*} = {\rho_i}^{r_i} \cdot (1 - \sigma_i)(1 - \sigma_{i-1}) = \frac{r_i}{\Pi_{j=1}^{i}(r_j + 1)} \cdot \frac{1}{\Pi_{j=1}^{i}(r_j + 1)} \cdot \frac{1}{\Pi_{j=1}^{i-1}(r_j + 1)}$$

Leading to:

$$P_i^{G*} = \frac{r_i}{\left(\Pi_{j=1}^{i}(r_j + 1)\right)^2 \cdot \Pi_{j=1}^{i-1}(r_j + 1)}$$

and the sum of optimized individual generalized power:

$$P_{\text{sum}}{}^{G} = \sum_{i=1}^{n} P_i^{G*} = \sum_{i=1}^{n} \frac{r_i}{\left(\Pi_{j=1}^{i}(r_j+1)\right)^2 \cdot \Pi_{j=1}^{i-1}(r_j+1)} \tag{10.24}$$

Summarizing these results, we have the following theorem:

Theorem 10.2.

Consider an M/M/1 system with n flows using ***HOL****, where each flow i has a preference variable r_i representing its relative preference for throughput over delay. At the equilibrium point where each flow individually maximizes its generalized power, the following holds:*

Optimized Individual Utilization Factor:

$$\rho_i^* = \frac{r_i}{\Pi_{j=1}^{i}(r_j+1)} \quad \textit{for } i = 1, 2, \ldots, n \tag{10.25}$$

Optimized Total System Utilization:

$$\rho^* = \sum_{i=1}^{n} \rho_i^* = \sum_{i=1}^{n} \frac{r_i}{\Pi_{j=1}^{i}(r_j+1)} \tag{10.26}$$

Optimized Individual Generalized Power:

$$P_i^{G*} = \frac{r_i}{\left(\Pi_{j=1}^{i}(r_j+1)\right)^2 \cdot \Pi_{j=1}^{i-1}(r_j+1)} \quad \textit{for } i = 1, 2, \ldots, n \tag{10.27}$$

Sum of Optimized Individual Generalized Power:

$$P_{sum}{}^{G} = \sum_{i=1}^{n} P_i^{G*} = \frac{r_i}{\left(\Pi_{j=1}^{i}(r_j+1)\right)^2 \cdot \Pi_{j=1}^{i-1}(r_j+1)} \tag{10.28}$$

As previously noted, the case where $r_i = 1$ for $i = 1, \ldots, n$ was analyzed earlier, and the optimization results for ρ_i^*, ρ^*, ${P_i}^{G*}$, and ${P_{\text{sum}}}^G = \sum_{i=1}^{n} {P_i}^{G*}$ align with those derived in Chapter 4 and are summarized in the HOL column of Table 4.1.

10.2.1.3 Limitations

Using the variable r_i for each flow provides the benefit of allowing each flow to represent its own importance of throughput relative to delay. However, without proper queueing disciplines, the individual power optimization result with each flow specifying a different r_i may not be effectively realized. For example, in the FCFS discipline, if flow 1 chooses $r_1 = 2$ and flow 2 chooses $r_2 = \frac{1}{2}$, both flows experience the same delay. Flow 2 does not experience a shorter response time simply because flow 1 values throughput more and sends more traffic based on its individual power optimization result. A better approach to handling different flow requirements should adjust the flow discrimination of queueing disciplines rather than relying on a fixed queueing discipline and simply adjusting each flow's utilization factor.

The situation in which some flow's requirements may not be met, especially when emphasizing the importance of delay over throughput with r_i smaller than 1, is also possible to occur in the HOL case. With the same $(r_1, r_2) = (2, \frac{1}{2})$, if flow 1 has a higher priority and flow 2 has a lower priority, flow 2 still experiences longer response times as it is affected by other flows that prefer throughput. It is unfair to flow 2 unless flow 1 pays more for its strictly higher priority. Therefore, these examples suggest a mechanism to adjust the degree of flow discrimination based on the values of each r_i. This necessitates further investigation as future work.

Chapter 11: Future Research Directions

11.1 Multiple Hops

In this dissertation, we have focused on single-hop systems, assuming that each flow has the same round-trip time (RTT). However, in real-world networks, flows often traverse different numbers of hops, leading to varying RTTs. This variation in RTT is a crucial factor that affects network performance. There is extensive research on the issue of RTT unfairness [60–63]. Flows with longer RTTs have a larger bandwidth-delay product, allowing them to send more traffic compared to flows with shorter RTTs. This disparity not only results in unfair bandwidth allocation but also leads to unexpected standing queues at bottleneck links.

In addition, multiple hops may cause bottlenecks to occur at different locations in the network, adding another dimension of complexity. Therefore, extending our power optimization analysis to multi-hop systems to investigate the impact of multiple hops on power optimization is necessary and will be an important direction for future work.

11.2 M/G/1

Most of the results in this dissertation are based on the M/M/1 model, which assumes exponential service times. Only a few analyses, such as the average power discussion in Chapter 6, have been extended to the more general M/G/1 model, where the service time distribution can be arbitrary. While the M/M/1 model provides valuable insights, it is limited by the assumption of exponentially distributed service times, which may not always

reflect real-world scenarios. Extending the analysis to the M/G/1 model across other metrics would allow for a more comprehensive examination of systems with general service time distributions, enhancing the applicability and robustness of the findings. This represents a significant opportunity for future work, as it would enable the development of more generalized and versatile optimization strategies.

11.3 Quantitative Fairness Measures

In Chapter 8 and Chapter 9, we define fairness as the equal value of the corresponding fairness metric. However, quantitative approaches, such as Jain's index [55], can provide deeper insights into the relationship between fairness and performance. Additionally, the trade-off between performance and fairness can be explored more thoroughly when fairness is measured quantitatively. In our analysis, we have only considered fairness as a binary measure—1 indicating fairness and 0 indicating the absence of fairness. If the fairness metric were expanded to an integer scale, a more comprehensive optimization framework could be developed. Our current approach is limited by the binary fairness measure, which restricts us to investigating only cases where a specific performance metric and fairness metric can simultaneously achieve optimal performance and fairness. If the fairness measure could be quantified more precisely and integrated with the performance power metric, this would allow for a more nuanced understanding of the interplay between performance and fairness, enabling the development of more sophisticated optimization strategies.

11.4 Dynamic Behavior of Networks

While our results are based on steady-state analysis, real networks are inherently dynamic. A critical challenge lies in designing network control mechanisms that can effectively guide the system toward the optimal steady state. Furthermore, the optimal operating point itself

may constantly shift due to fluctuations in real-world traffic patterns. Therefore, developing iterative algorithms that can adapt rapidly to these dynamic changes is crucial. One promising approach might involve leveraging the property that the maximal power point corresponds to the tangent line on the performance curve. However, further exploration is needed to determine how best to implement this property in practical algorithms.

11.5 Applications

Our work, primarily theoretical in nature, requires further development to address the dynamic complexities of real-world systems before it can be applied. Potential applications include areas such as TCP congestion control, router buffer sizing, and queue management within networks. Additionally, the scheduling components within data center networks, such as Borg in Google [64], represent promising targets for applying these concepts. Data center networks require scheduling systems that can efficiently and fairly balance resources such as CPU, memory, and GPU for a multitude of jobs and users. Designing the scheduling logic and determining the optimal resource allocation for each user presents a challenge analogous to our three-dimensional framework that encompasses performance, fairness, and priority flow discrimination.

References

[1] U Cisco. Cisco annual internet report (2018–2023) white paper. *Cisco: San Jose, CA, USA*, 10(1):1–35, 2020.

[2] I Sandvine. Global internet phenomena report. *North America and Latin America*, 2024.

[3] Alfred Giessler, J Haenle, Andreas König, and E Pade. Free buffer allocation—An investigation by simulation. *Computer Networks (1976)*, 2(3):191–208, 1978.

[4] Leonard Kleinrock. On Flow Control in Computer Networks. In *Proceedings of the International Conference on Communications*, volume 2, pages 27–2, 1978.

[5] Leonard Kleinrock. Power and Deterministic Rules of Thumb for Probabilistic Problems in Computer Communications. In *ICC'79; International Conference on Communications, Volume 3*, volume 3, pages 43–1, 1979.

[6] R. Gail and L. Kleinrock. An Invariant Property of Computer Network Power. In *Proceedings of the International Conference on Communications*, pages 63.1.1–63.1.5, June 14-18, 1981.

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

[8] Xipeng Xiao and Lionel M Ni. Internet qos: A big picture. *IEEE network*, 13(2):8–18, 1999.

[9] Zheng Wang. *Internet QoS: architectures and mechanisms for quality of service*. Morgan Kaufmann, 2001.

[10] Uyless Black. *Voice over IP*. Prentice-Hall, Inc., 1999.

[11] Bur Goode. Voice over internet protocol (voip). *Proceedings of the IEEE*, 90(9):1495–1517, 2002.

[12] K. Nichols, S. Blake, F. Baker, and 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, December 1998. Available at https://www.rfc-editor.org/rfc/rfc2474.

[13] Stephen Blake, David Black, Mark A. Carlson, Elwyn Davies, Zheng Wang, and Walter Weiss. An architecture for differentiated services. Request for Comments RFC 2475, Internet Engineering Task Force, December 1998. Available at https://www.rfc-editor.org/rfc/rfc2475.

[14] Robert Braden, David Clark, and Scott Shenker. Integrated services in the internet architecture: an overview. Request for Comments RFC 1633, Internet Engineering Task Force, June 1994. Available at https://www.rfc-editor.org/rfc/rfc1633.

[15] Leonard Kleinrock. *Queueing Systems, Volume II: Computer Applications*, volume 66. Wiley New York, 1976.

[16] Alan Cobham. Priority Assignment in Waiting Line Problems. *Journal of the Operations Research Society of America*, 2(1):70–76, 1954.

[17] NK Jaiswal. Preemptive Resume Priority Queue. *Operations Research*, 9(5):732–742, 1961.

[18] Van Jacobson. Congestion avoidance and control. *ACM SIGCOMM computer communication review*, 18(4):314–329, 1988.

[19] Van Jacobson. Modified TCP congestion avoidance algorithm, 1990.

[20] Mark Allman, Vern Paxson, and William Stevens. RFC2581: TCP congestion control, 1999.

[21] Lawrence S Brakmo, Sean W O'Malley, and Larry L Peterson. TCP Vegas: New techniques for congestion detection and avoidance. In *Proceedings of the conference on Communications architectures, protocols and applications*, pages 24–35, 1994.

[22] Sangtae Ha, Injong Rhee, and Lisong Xu. CUBIC: a new TCP-friendly high-speed TCP variant. *ACM SIGOPS operating systems review*, 42(5):64–74, 2008.

[23] Mohammad Alizadeh, Albert Greenberg, David A Maltz, Jitendra Padhye, Parveen Patel, Balaji Prabhakar, Sudipta Sengupta, and Murari Sridharan. Data center tcp (dctcp). In *Proceedings of the ACM SIGCOMM 2010 Conference*, pages 63–74, 2010.

[24] Radhika Mittal, Vinh The Lam, Nandita Dukkipati, Emily Blem, Hassan Wassel, Monia Ghobadi, Amin Vahdat, Yaogong Wang, David Wetherall, and David Zats. Timely: Rtt-based congestion control for the datacenter. *ACM SIGCOMM Computer Communication Review*, 45(4):537–550, 2015.

[25] Neal Cardwell, Yuchung Cheng, C. Stephen Gunn, Soheil Hassas Yeganeh, and Van Jacobson. BBR: congestion-based congestion control. *ACM Queue*, 14(5):20–53, 2016.

[26] Yuliang Li, Rui Miao, Hongqiang Harry Liu, Yan Zhuang, Fei Feng, Lingbo Tang, Zheng Cao, Ming Zhang, Frank Kelly, Mohammad Alizadeh, and Minlan Yu. HPCC: high precision congestion control. In Jianping Wu and Wendy Hall, editors, *Proceedings of the ACM Special Interest Group on Data Communication, SIGCOMM 2019, Beijing, China, August 19-23, 2019*, pages 44–58. ACM, 2019.

[27] Gautam Kumar, Nandita Dukkipati, Keon Jang, Hassan MG Wassel, Xian Wu, Behnam Montazeri, Yaogong Wang, Kevin Springborn, Christopher Alfeld, Michael 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, pages 514–528, 2020.

[28] Kadangode Ramakrishnan and Sally 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] Dina Katabi, Mark Handley, and Charlie Rohrs. Congestion control for high bandwidth-delay product networks. In *Proceedings of the 2002 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 89–102, 2002.

[30] Sally Floyd and Van Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Trans. Netw.*, 1(4):397–413, 1993.

[31] Kathleen M. Nichols, Van Jacobson, Andrew McGregor, and Janardhan R. Iyengar. Controlled delay active queue management. *RFC*, 8289:1–25, 2018.

[32] Guido Appenzeller, Isaac Keslassy, and Nick McKeown. Sizing router buffers. *ACM SIGCOMM Computer Communication Review*, 34(4):281–292, 2004.

[33] Leonard Kleinrock. A Conservation Law for a Wide Class of Queueing Disciplines. *Naval Research Logistics Quarterly*, 12(2):181–192, 1965.

[34] David G Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, pages 338–354, 1953.

[35] John D.C. Little. A Proof for the Queuing Formula: L= λW. *Operations research*, 9(3):383–387, 1961.

[36] Leonard Kleinrock. *Queueing Systems. Volume I: Theory.* wiley New York, 1975.

[37] Leonard Kleinrock. *Message delay in communication nets with storage.* PhD thesis, Massachusetts Institute of Technology, 1963.

[38] Leonard Kleinrock. A Delay Dependent Queue Discipline. *Naval Research Logistics Quarterly*, 11(4), December 1964.

[39] Thomas Stockhammer. Dynamic adaptive streaming over http– standards and design principles. In *Proceedings of the second annual ACM conference on Multimedia systems*, pages 133–144, 2011.

[40] Alex Zambelli. Iis smooth streaming technical overview. *Microsoft Corporation*, 3(40), 2009.

[41] Adobe. Adobe http dynamic streaming (hds), 2016.

[42] R Pantos and W May. Apple inc.,"http live streaming,", 2013.

[43] Abdelhak Bentaleb, Bayan Taani, Ali C Begen, Christian Timmerer, and Roger Zimmermann. A survey on bitrate adaptation schemes for streaming media over http. *IEEE Communications Surveys & Tutorials*, 21(1):562–585, 2018.

[44] John Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

[45] Robert Morris. Tcp behavior with many flows. In *Proceedings 1997 International Conference on Network Protocols*, pages 205–211. IEEE, 1997.

[46] Lili Qiu, Yin Zhang, and Srinivasan Keshav. On individual and aggregate tcp performance. In *Proceedings. Seventh International Conference on Network Protocols*, pages 203–212. IEEE, 1999.

[47] Lili Qiu, Yin Zhang, and Srinivasan Keshav. Understanding the performance of many tcp flows. *Computer Networks*, 37(3-4):277–306, 2001.

[48] Kadangode K Ramakrishnan and Raj Jain. A binary feedback scheme for congestion avoidance in computer networks. *ACM Transactions on Computer Systems (TOCS)*, 8(2):158–181, 1990.

[49] Dong Lin and Robert Morris. Dynamics of random early detection. In *Proceedings of the ACM SIGCOMM'97 conference on Applications, technologies, architectures, and protocols for computer communication*, pages 127–137, 1997.

[50] Jim Gettys. Bufferbloat: Dark buffers in the internet. *IEEE Internet Computing*, 15(3):96–96, 2011.

[51] Jim Gettys and Kathleen Nichols. Bufferbloat: dark buffers in the internet. *Communications of the ACM*, 55(1):57–65, 2012.

[52] Thomas Bonald and Laurent Massoulié. Impact of fairness on internet performance. In *Proceedings of the 2001 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 82–91, 2001.

[53] Tian Lan, David Kao, Mung Chiang, and Ashutosh Sabharwal. *An axiomatic theory of fairness in network resource allocation*. IEEE, 2010.

[54] Dah Ming Chiu and Adrian SW Tam. Network fairness for heterogeneous applications. In *Proceedings of ACM SIGCOMM ASIA Workshop*, 2005.

[55] Rajendra K Jain, Dah-Ming W Chiu, William R Hawe, et al. A quantitative measure of fairness and discrimination. *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, 21:1, 1984.

[56] E.L. Hahne. Round-robin scheduling for max-min fairness in data networks. *IEEE Journal on Selected Areas in Communications*, 9(7):1024–1039, 1991.

[57] Frank Kelly. Charging and rate control for elastic traffic. *European transactions on Telecommunications*, 8(1):33–37, 1997.

[58] Frank P Kelly, Aman K Maulloo, and David Kim Hong Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3):237–252, 1998.

[59] Fouad Tobagi, Waël Noureddine, Benjamin Chen, Athina Markopoulou, Chuck Fraleigh, Mansour Karam, Jose-Miguel Pulido, and Jun-ichi Kimura. Service differentiation in the internet to support multimedia traffic. In *Evolutionary Trends of the Internet: 2001 Tyrrhenian International Workshop on Digital Communications, IWDC 2001 Taormina, Italy, September 17–20, 2001 Proceedings*, pages 381–400. Springer, 2001.

[60] Lisong Xu, Khaled Harfoush, and Injong Rhee. Binary increase congestion control (bic) for fast long-distance networks. In *IEEE INFOCOM 2004*, volume 4, pages 2514–2524. IEEE, 2004.

[61] Mario Hock, Roland Bless, and Martina Zitterbart. Experimental evaluation of bbr congestion control. In *2017 IEEE 25th international conference on network protocols (ICNP)*, pages 1–10. IEEE, 2017.

[62] Shiyao Ma, Jingjie Jiang, Wei Wang, and Bo Li. Fairness of congestion-based congestion control: Experimental evaluation and analysis. *arXiv preprint arXiv:1706.09115*, 2017.

[63] Geon-Hwan Kim, Yeong-Jun Song, Imtiaz Mahmud, and You-Ze Cho. Enhanced bbr congestion control algorithm for improving rtt fairness. In *2019 eleventh international conference on ubiquitous and future networks (ICUFN)*, pages 358–360. IEEE, 2019.

[64] Muhammad Tirmazi, Adam Barker, Nan Deng, Md E Haque, Zhijing Gene Qin, Steven Hand, Mor Harchol-Balter, and John Wilkes. Borg: the next generation. In *Proceedings of the fifteenth European conference on computer systems*, pages 1–14, 2020.