BULGARIAN ACADEMY OF SCIENCES
INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Genka Ivanova Slavova-Torres

METHODS AND APPROACHES FOR OPTIMIZING WRITE OPERATIONS IN STORAGE AREA NETWORKS

ABSTRACT OF PhD THESIS

Doctoral Program **Computer Systems, Complexes and Networks**
Professional Field 5.3 **Communication and Computer Engineering**

Supervisor: Assoc. Prof. Vladimir Ivanov, PhD

Sofia, 2026

The dissertation has been reviewed and scheduled for defense at a meeting of the Scientific Council of the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences (IICT–BAS), held on ………….

The public defense of the dissertation will take place on …………. at …………. hours at IICT–BAS.

The dissertation is structured into an Introduction, 6 chapters, Conclusion, Scientific Contributions, List of Publications, Declaration of Originality of the Results, and Bibliography. The total length of the dissertation is 147 pages and it includes 24 figures and 16 tables.

The defense of the dissertation will take place on ………………….…... 2026
at ……………………….…… hours in hall ……………………. of block 2 of IICT–BAS
at an open session of a scientific jury composed of:

Scientific jury

1.……………………………………………………………………………………

2. ……………………………………………………………………………………

3. ……………………………………………………………………………………

4. …………………………………………………………………………………….

5. …………………………………………………………………………………….

The reviews and opinions of the members of the scientific jury and the abstract will be published on the website of IICT-BAS.

The materials for the defense are available to those interested in room 315 of IICT-BAS, "Akad. G. Bonchev", Str., bl. 2.

Author: **Genka Ivanova Slavova-Torres**
Title: **Methods and Approaches for Optimizing Write Operations in Sorage Area Networks**

Keywords:
Storage Area Networks (SAN), write path modeling, latency distribution analysis, tail latency behavior, queueing mechanisms, parallelism control, empirical cumulative distribution function (ECDF), system-level performance analysis, coordination effects

INTRODUCTION

Modern information systems are characterized by rapidly increasing data volumes and continuously growing requirements for performance, reliability, and scalability of storage infrastructures. Storage Area Networks (SAN) play a central role in enterprise data centers, virtualized environments, and cloud platforms, where multiple applications and systems share common storage resources and generate concurrent input/output workloads.

Within these environments, write operations are of critical importance, as they directly affect application response time, data consistency, and overall system stability. Despite significant technological advancements, including the widespread adoption of solid-state drives (SSD) and NVMe-based architectures, the behavior of write operations remains complex and difficult to predict under realistic workload conditions.

The primary reason for this complexity lies in the multi-layered nature of the write process in SAN environments. Write requests traverse a sequence of architectural layers, including the host system, the transport infrastructure, and the storage subsystem. Each of these layers introduces its own mechanisms for buffering, scheduling, and queue management. The interaction between these mechanisms forms a system of interdependent queues that collectively determine the observed latency of write operations.

Empirical observations in real-world systems show that as load increases, write latency does not grow linearly. While the majority of operations are completed within relatively short time intervals, a small fraction of requests experience significantly higher delays. These rare but extreme delays form the tail of the latency distribution and may have a disproportionate impact on application performance and service quality.

Traditional approaches to performance evaluation are typically based on aggregate metrics such as average latency or throughput. However, such metrics are often insufficient to capture the actual behavior of systems under dynamic and competitive workloads. In many cases, average latency remains within acceptable limits while high-percentile latency (e.g., p95, p99) increases significantly, indicating the presence of instability and queueing effects within the system.

Furthermore, modern SAN environments are characterized by high levels of concurrency and dynamic workload patterns. The widespread use of virtualization, containerization, and microservice-based architectures leads to multiple independent sources of I/O requests competing for shared resources. Under these conditions, the behavior of write operations cannot be attributed solely to the characteristics of individual components. Instead, it emerges from the interaction and coordination between different layers of the system.

These observations indicate that the problem of optimizing write operations in SAN environments has a strongly pronounced system-level nature. Improving the performance of individual components does not necessarily lead to stable and predictable system behavior. Therefore, there is a need for an approach that considers the write path as a system of interacting mechanisms for request management, parallelism, and queueing.

In this context, the present dissertation is focused on the investigation of write operation behavior in SAN environments through the development of a system-level model of the write path and an experimental analysis of the influence of workload characteristics, parallelism, and queue management on latency distribution. Particular attention is given to the behavior of high-percentile latency and to the factors that lead to the emergence of rare but significantly delayed operations.

The research aims not only to quantify system performance but also to provide a deeper understanding of the mechanisms that shape latency behavior under different operating conditions. This understanding creates a foundation for more effective approaches to analysis, configuration, and optimization of modern storage infrastructures.

1. METHODOLOGY

The research methodology adopted in this dissertation is based on a combined experimental and analytical approach aimed at investigating the behavior of write operations in Storage Area Network (SAN) environments from a system-level perspective. The primary objective of the methodology is to identify and analyze the relationships between workload characteristics, parallelism, queue management mechanisms, and the resulting latency behavior of write operations.

The study follows a structured research framework consisting of three main stages: conceptual modeling, experimental design and execution, and analytical interpretation of the obtained results.

Conceptual Modeling

The first stage of the methodology involves the development of a conceptual system model of the write path. In this model, write operations are represented as a sequence of interactions across three main architectural layers: the host system, the transport infrastructure, and the storage subsystem.

Each layer is characterized by its own mechanisms for request generation, buffering, scheduling, and queue management. The host layer defines the initial level of parallelism and request issuance, the transport layer governs the movement and coordination of requests across the SAN fabric, and the storage layer processes and persists the data using internal scheduling and caching mechanisms.

From a methodological perspective, the write path is interpreted as a system of interacting queues, where the latency of each operation is determined not only by service time but also by accumulated waiting time across multiple stages. This abstraction enables the analysis of latency as a system-level phenomenon emerging from the interaction between independent control mechanisms rather than as a property of a single component.

Experimental Design

The experimental design is based on a controlled approach aimed at isolating the impact of individual factors on latency behavior. Each experiment varies a single parameter while maintaining a fixed system configuration, enabling clear identification of causal relationships.

The experiments are conducted in a controlled SAN environment consisting of a host system generating I/O workloads, a Fibre Channel transport layer, and a flash-based storage system with active write caching. Workloads are generated using a dedicated I/O testing tool, allowing precise control over parameters such as request size, access pattern, and concurrency level.

A baseline scenario is defined to establish a reference point for system behavior under stable and uniform workload conditions. This scenario serves as a foundation for comparison across all subsequent experiments.

Building upon the baseline, four groups of experimental scenarios are defined, each targeting a distinct factor influencing latency:

Parallelism (Queue Depth):

The first group examines the impact of queue depth (QD) as a control parameter for concurrency. By systematically increasing QD, the experiments analyze how parallelism affects throughput and the distribution of latency.

Workload Structure:

The second group investigates the influence of workload structure by comparing uniform and dynamically varying request patterns. This enables evaluation of system behavior under non-stationary conditions.

Temporal Coordination:

The third group focuses on the effect of request coordination. By comparing coordinated and bursty request submission, the experiments analyze how synchronization influences queue formation and latency variability.

Resource Competition:

The fourth group examines the effect of competition between independent sources of I/O operations. Multiple hosts generate concurrent requests to shared storage resources, enabling analysis of interference effects and resource contention.

This structured design allows systematic evaluation of how different workload characteristics influence both typical latency behavior and high-percentile latency, providing a basis for identifying the mechanisms leading to variability and instability.

Measurement and Metrics

A key aspect of the methodology is the selection of appropriate metrics for evaluating system behavior. While traditional performance indicators such as throughput and average latency are included, the primary focus is placed on the analysis of latency distribution.

Latency is analyzed using percentile-based metrics, including median (p50), high percentiles (p95), and extreme percentiles (p99). These metrics provide insight into both typical system behavior and the presence of rare but significantly delayed operations.

In addition, empirical cumulative distribution functions (ECDF) are used as a central tool for visualizing and interpreting latency behavior. The ECDF enables direct observation of the proportion of operations completed within a given latency threshold and allows detailed examination of the tail of the distribution.

This approach provides a more comprehensive understanding of system performance, as it captures both central tendencies and extreme deviations, which are critical for evaluating stability under load.

Analytical Interpretation

The final stage of the methodology involves the analytical interpretation of the experimental results. The collected data is processed using statistical analysis techniques to identify patterns, trends, and dependencies between system parameters and observed latency behavior.

Particular attention is given to the identification of nonlinear effects, queue accumulation phenomena, and coordination effects between different layers of the system. The analysis aims to distinguish between different operational regimes, including stable operation, transition zones, and states characterized by increased variability and tail amplification.

By combining experimental observations with system-level interpretation, the methodology enables the formulation of generalized insights into the behavior of write operations in SAN environments.

Methodological Summary

The proposed methodology provides a reproducible and systematic framework for analyzing write operations under controlled conditions. By integrating conceptual modeling with targeted experimentation and distribution-based analysis, it allows the identification of key factors influencing latency stability and supports the development of more effective approaches for system analysis and optimization.

2. AIM AND TASKS OF THE RESEARCH

The The main objective of this dissertation is to investigate the behavior of write operations in Storage Area Network (SAN) environments through the development of a system-level model of the write path and through experimental analysis of the influence of workload characteristics, parallelism, and I/O queue management on write latency.

The research aims to identify the factors that determine the stability and predictability of write latency in SAN environments and to develop an analytical approach for the analysis and optimization of write operations based on the interactions between the host system, the transport layer, and the storage subsystem. Particular emphasis is placed on latency distributions, high-percentile latency (p95 and p99), and queueing effects that emerge under concurrent workloads.

To achieve this objective, the following research tasks are defined:

- to analyze the architectural characteristics of modern SAN environments and the factors influencing the behavior of write operations;

- to develop a conceptual system-level model of the write path, describing the interaction between the host system, the transport layer, and the storage subsystem;
- to define appropriate metrics for latency analysis, with emphasis on latency distributions and high-percentile latency (p95 and p99);
- to design an experimental methodology for the controlled investigation of write operations under different workload conditions;
- to conduct a series of controlled experiments examining the influence of I/O queue depth, workload characteristics, parallelism, and host competition on write latency;
- to perform an analytical synthesis of the experimental results in order to identify the principal factors governing latency stability and predictability;
- to develop an analytical framework for the optimization of write operations in SAN environments based on latency distributions, queueing effects, and workload coordination.

3. OBJECT AND SUBJECT OF THE RESEARCH

The object of this research is **Storage Area Network (SAN)** environments used in modern information infrastructures to provide block-level access to shared storage resources.

The subject of the research is the behavior of write operations in SAN environments, analyzed through write latency and its distribution under different workload conditions. Particular attention is given to the influence of parallelism, **I/O queue management**, and the interaction between the architectural layers along the write path, namely the host system, the transport infrastructure, and the storage subsystem.

The research focuses on the mechanisms governing the distribution and variability of write latency, with particular emphasis on high-percentile latency (e.g., **p95** and **p99**) and **tail latency**, which characterize rare but significant delays beyond the average system behavior. In this context, write operations are considered as a system-level process resulting from the interaction of multiple mechanisms for request scheduling, buffering, queue management, and resource sharing across the different layers of the SAN infrastructure.

Special attention is devoted to the role of **I/O queues** and parallelism as key factors influencing the stability and predictability of write latency. The analysis is aimed at identifying the system-level relationships between the architectural components of SAN infrastructures and the observed latency behavior under different workload conditions.

4. RESEARCH THESIS

The main thesis of this dissertation is that the behavior **and optimization** of write operations in **Storage Area Network (SAN)** environments are determined by the interaction between workload characteristics, parallelism control mechanisms, and **I/O queue management** along the write path.

Accordingly, the performance of SAN systems cannot be adequately evaluated using average latency metrics alone. A comprehensive assessment requires the analysis of latency distributions and high-percentile latency values (e.g., **p95** and **p99**), which reveal queueing effects and rare but significantly delayed write operations that remain hidden when only average metrics are considered.

Within this framework, the stability and predictability of write latency are analyzed through a system-level model of the write path, in which the SAN infrastructure is interpreted as a set of interacting mechanisms responsible for request management, parallelism, buffering, and queueing across the host system, transport layer, and storage subsystem.

The dissertation further argues that sustainable optimization of write operations is achieved not by maximizing the performance of individual components, but by maintaining effective coordination between the architectural layers and controlling the interactions between workload intensity and I/O queues. Consequently, stable latency behavior is treated as the primary criterion for optimization rather than peak throughput alone.

5. LIMITATIONS OF THE STUDY

The present study focuses on the analysis of write operations in **Storage Area Network (SAN)** environments under controlled and well-defined experimental conditions. The scope of the research is deliberately constrained in order to isolate the principal factors affecting write latency and to ensure the reproducibility of the experimental results.

First, the research is limited to **write operations**, while read operations are not investigated in detail. This decision is motivated by the fact that write operations involve more complex coordination mechanisms and generally have a greater impact on system stability under high workload conditions.

Second, the experimental study is conducted in a controlled environment with fixed infrastructure parameters, where only workload characteristics - such as **I/O queue depth**, workload profile, and concurrency - are systematically varied. Consequently, the reported results should be interpreted within the context of the adopted experimental methodology rather than as universally applicable configurations for all SAN implementations.

Third, the research focuses on the **system-level behavior** of write operations and does not investigate the internal implementation details of specific commercial storage systems. Internal controller algorithms, scheduling policies, caching mechanisms, and vendor-specific optimizations are intentionally treated as a **black box**, allowing the analysis to concentrate on the end-to-end latency behavior observed along the write path.

Similarly, the study does not examine in detail the internal behavior of the SAN transport infrastructure, including switch architectures, traffic management mechanisms, or protocol-specific optimizations. Although these factors may influence the quantitative characteristics of latency, they are outside the scope of the present investigation.

These limitations do not diminish the validity of the obtained results. Instead, they clearly define the architectural and methodological boundaries of the research and allow the identified relationships between workload, parallelism, queueing, and latency behavior to be interpreted within a consistent system-level framework.

6. PRACTICAL SIGNIFICANCE OF THE RESEARCH

The results of this dissertation have practical significance for the design, configuration, optimization, and operation of **Storage Area Network (SAN)** environments.

The proposed system-level model of the write path provides a structured framework for understanding how write latency is formed through the interaction of the host system, transport layer, storage subsystem, and **I/O queueing mechanisms**. This enables more effective identification of bottlenecks, request accumulation points, and coordination effects across the storage infrastructure.

The experimental methodology developed in this research provides a reproducible approach for evaluating write behavior under controlled workload conditions. Combined with the analysis of latency distributions, empirical cumulative distribution functions (ECDF), and high-percentile latency values (p95 and p99), it enables the identification of performance degradation that remains hidden when only average latency or throughput metrics are considered.

The analytical synthesis of the experimental results forms the basis of a practical framework for the optimization of write operations in SAN environments. The proposed optimization framework supports the identification of stable operating regions, the early detection of coordination stress, and the selection of appropriate optimization actions through the management of parallelism, **I/O queue depth**, and workload characteristics.

The findings demonstrate that maximizing throughput alone does not necessarily result in stable system behavior. Instead, sustainable performance is achieved by maintaining stable latency distributions and controlling queueing effects under concurrent workloads.

The proposed methodology is directly applicable to latency-sensitive environments, including **OLTP databases, financial transaction systems, multi-tenant storage platforms, and other enterprise applications** where predictable write latency is more important than peak throughput.

The results may assist storage architects, system engineers, and performance analysts in designing and operating SAN infrastructures that provide improved stability, predictability, and efficient resource utilization under conditions of high concurrency and shared resource access.

7. STRUCTURE OF THE DISSERTATION

The dissertation consists of six chapters, followed by the conclusion, the scientific contributions, references, and appendices.

Chapter 1 introduces the research problem and justifies its relevance. It discusses the role of write operations in modern **Storage Area Network (SAN)** environments, outlines the limitations of conventional performance evaluation based primarily on average latency, and formulates the research objective, tasks, object and subject of the study, research thesis, methodology, and scope of the investigation.

Chapter 2 develops a conceptual system-level model of the write path in SAN environments. The write operation is interpreted as a sequence of interacting architectural layers, including the host system, transport infrastructure, and storage subsystem. Particular attention is given to the role of **I/O queues**, parallelism, and queueing mechanisms in shaping latency behavior.

Chapter 3 presents the experimental methodology adopted in the research. It describes the experimental environment, workload generation, measurement procedures, and evaluation metrics. A controlled experimental design is introduced, allowing the systematic investigation of the influence of workload characteristics, queue depth, and concurrency while maintaining a fixed system configuration.

Chapter 4 presents and analyzes the experimental results obtained under different workload conditions. The influence of **I/O queue depth**, workload dynamics, bursty traffic, and competing hosts on write latency is investigated. The analysis includes empirical cumulative distribution functions (ECDF), latency distributions, and high-percentile latency metrics (p95 and p99), demonstrating the emergence of queueing effects and latency instability under increasing contention.

Chapter 5 provides an analytical synthesis and interpretation of the experimental findings. The observed behavior is interpreted using principles of queueing theory, leading to the identification of distinct operating regimes - stable operation, transition, and coordination stress. On this basis, an analytical framework for the optimization of write operations in SAN environments is developed, together with practical optimization guidelines and a closed-loop optimization framework based on latency monitoring, workload analysis, and adaptive decision-making.

Chapter 6 summarizes the principal findings of the dissertation, discusses the limitations of the research, outlines directions for future work, and presents the overall conclusions regarding the analysis and optimization of write operations in SAN environments.

8. CONTENT OF THE DISSERTATION

Chapter 1: Problem Statement, Research Thesis, and Motivation

Chapter 1 introduces the research problem and establishes the motivation for the study. It examines the role of write operations in modern **Storage Area Network (SAN)** environments and their impact on system performance, latency, and operational stability.

The chapter shows that, despite the significant evolution of storage technologies - from mechanical disks to flash-based and NVMe architectures - the behavior of write operations remains difficult to predict under concurrent and dynamic workloads. This is because the write path represents a complex, multi-stage process involving interactions between the host system, the transport infrastructure, and the storage subsystem, each introducing its own mechanisms for request scheduling, buffering, and queue management.

Particular attention is devoted to **tail latency**, where a small fraction of write operations experience significantly higher response times than the majority of requests. It is demonstrated that conventional performance metrics, such as average latency and throughput, are insufficient for characterizing this behavior, as they fail to capture latency variability and the emergence of rare but critical delays.

On this basis, the chapter argues that write latency should be interpreted as a **system-level property** rather than as a characteristic of the storage subsystem alone. The observed latency behavior is presented as the result of interactions between workload characteristics, parallelism, and **I/O queue management** across the entire write path.

The chapter formulates the main research thesis, stating that the behavior **and optimization** of write operations in SAN environments are determined primarily by the coordination between workload characteristics, parallelism control mechanisms, and I/O queues, rather than by the nominal performance of individual hardware components.

Finally, the chapter defines the research objective, research tasks, object and subject of the investigation, research methodology, limitations of the study, and its practical significance, thereby establishing the conceptual foundation for the system-level model and the experimental investigation presented in the subsequent chapters.

Chapter 2: System-Level Model of Write Operations in SAN Environments

Chapter 2 is devoted to the development of a conceptual system-level model of write operations in **Storage Area Network (SAN)** environments. The primary objective of this chapter is to establish an analytical framework that enables the consistent interpretation of write latency as the result of interactions between multiple architectural components and **I/O queueing mechanisms** distributed along the write path.

The chapter begins with an analysis of the evolution of performance bottlenecks in storage systems. It is shown that, with the transition from mechanical storage devices to flash-based and NVMe technologies, the dominant limiting factors have shifted from physical constraints to architectural coordination and queueing effects. While device-level latency has been substantially reduced, the overall behavior of modern storage systems is increasingly determined by the interaction between workload characteristics, parallelism, and queue management.

Based on this observation, the concept of the **write path** is introduced as a system-level abstraction. A write operation is modeled as a process traversing a sequence of interconnected architectural layers—the host system, the transport infrastructure, and the storage subsystem. Each layer contributes to the overall latency through its own mechanisms for request generation, scheduling, buffering, and queue management. The conceptual model of the write path is illustrated in **Figure 1**.

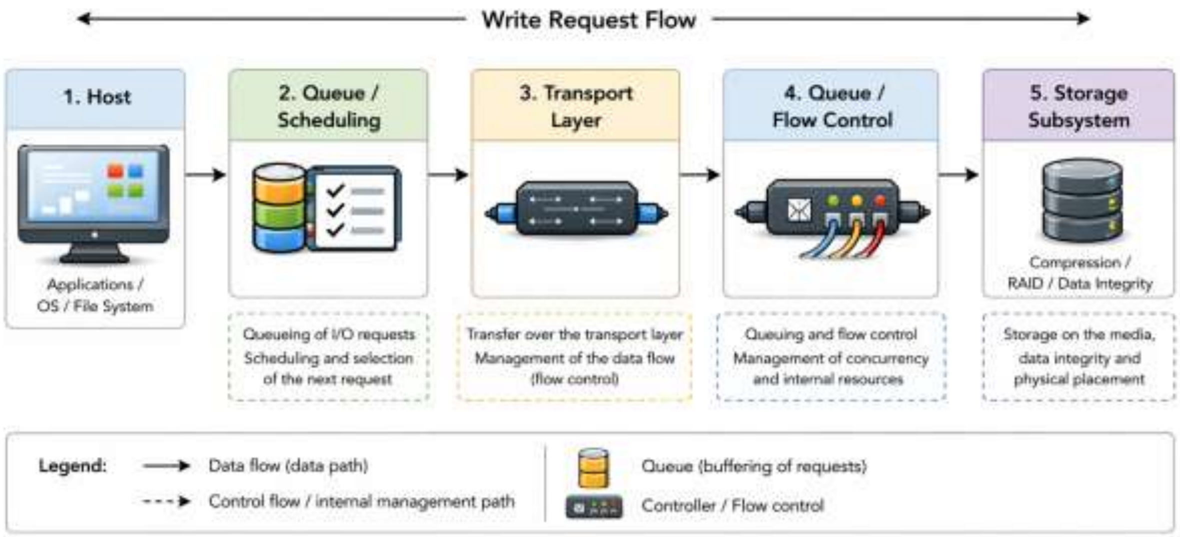

Figure 1. Conceptual system-level model of the write path illustrating interactions between host, transport, and storage layers

The proposed model represents the write path as a sequence of interacting queues distributed across these layers. The host system generates I/O requests and determines the initial level of parallelism through operating system queues and I/O schedulers. The transport layer forwards requests across the SAN fabric while introducing its own flow-control mechanisms. Finally, the storage subsystem processes incoming requests through internal scheduling, caching, and persistence mechanisms.

To provide an analytical interpretation of this behavior, the chapter introduces principles derived from queueing theory, where the total write latency is interpreted as the cumulative effect of service time and waiting time across the different stages of the write path. Consequently, the overall latency can be expressed conceptually as

$$T_{write} = \sum_{i=1}^{n} (S_i + W_i),$$

where $S_i$denotes the service time at the i-th stage and $W_i$represents the corresponding queue waiting time.

This formulation emphasizes that write latency is a system-level property resulting from the cumulative interaction of multiple service and queueing mechanisms rather than from the performance of a single component. As workload intensity increases, temporary queue buildup may occur at different stages of the write path, leading to nonlinear latency growth and increased latency variability.

A key contribution of the proposed model is the explicit representation of the relationship between parallelism and queueing. Increasing the degree of parallelism may improve throughput up to a certain operating point, but it simultaneously increases the probability of temporary congestion within one or more architectural layers. These effects propagate across the write path and give rise to emergent system behavior that cannot be explained by analyzing individual components in isolation.

The model also provides the conceptual basis for interpreting latency distributions, high-percentile latency (p95 and p99), and tail latency. Rather than attributing latency spikes to isolated bottlenecks, the proposed interpretation explains them as the consequence of transient imbalances between workload intensity and processing capacity across interacting queues.

Furthermore, the chapter emphasizes that stable write behavior depends on effective coordination between the architectural layers. Even when individual components operate within their nominal performance limits, insufficient coordination between queue management mechanisms may result in increased latency variability and reduced predictability.

The conceptual model developed in this chapter serves as the analytical foundation for the experimental investigation presented in the following chapters and for the analytical synthesis and optimization framework developed later in the dissertation. It establishes the system-level perspective used to interpret latency behavior and provides the basis for identifying stable operating regimes and optimization opportunities in SAN environments.

Chapter 3: Experimental Methodology

Chapter 3 presents the experimental methodology used to investigate the behavior of write operations in Storage Area Network environments. The primary objective of this chapter is to establish a reproducible experimental framework for the systematic investigation of the influence of workload parameters on write latency. The methodology follows a sequential research framework consisting of conceptual analysis, development of a system-level model, controlled experimentation, and analytical synthesis of the obtained results.

The methodology is based on a controlled experimental approach, in which the system configuration remains fixed while individual workload parameters are varied. This principle allows the isolation of specific factors and supports the identification of causal relationships between system inputs and observed behavior.

The experimental environment consists of three main components corresponding to the layers of the write path: a host system generating input/output workloads, a transport infrastructure based on Fibre Channel connectivity, and a storage subsystem built on a dual-controller architecture with flash-based media and active write caching. The host operates in a Linux-based environment and generates workloads using a dedicated I/O benchmarking tool, enabling precise control over request characteristics. The fixed parameters of the experimental design, including hardware configuration, workload characteristics, measurement duration, warm-up period, and repeatability, are summarized in Table 1.

Category	Parameter	Value
Host (Initiator)	Number of Hosts	1
	Access Type	Block-level SAN access (no file system)
	Operating System	Linux-based
	Workload Generator	fio
	Multipathing	Enabled (2 independent I/O paths)
SAN Transport	Technology	Fibre Channel
	Connectivity	2 × 16 Gbit/s
	Topology	Redundant, no single-path dependency
Storage System	Architecture	Dual-controller SAN storage system
	Write Cache	Enabled
	Back-end Storage	Flash-based storage
Working Set	Logical Capacity	1 TB
	Working Set Size	256 GB
Baseline Workload	Operation Type	100% Write
	Block Size	4 KB Random Write
Measurement	Test Duration	300 s
	Warm-up Period	30 s
	Repeatability	3 runs

Table 1. Fixed Parameters of the Experimental Design

The selected configuration reflects a representative enterprise SAN environment while ensuring experimental repeatability and consistent comparison across all test scenarios.

A baseline scenario is defined to establish a reference for system behavior under stable conditions. In this scenario, the workload consists of random write operations with fixed block size and uniform intensity. The system is allowed to reach a steady state before measurements are collected, ensuring that transient effects do not influence the results.

Building on this baseline, several experimental scenarios are designed to examine the impact of key parameters on latency behavior. The first experimental scenario investigates the influence of I/O queue depth, which represents the primary mechanism controlling the degree of parallelism within the storage system. By increasing the number of concurrent requests, the study analyzes how higher load levels influence both throughput and latency distribution.

The second set of experiments focuses on workload structure. Different patterns of request generation are compared, including uniform and dynamically varying workloads. This allows evaluation of the system's response to non-stationary conditions and provides insight into how temporal variations in load affect queueing behavior.

The third experimental scenario examines the effect of competition between independent sources of I/O operations. In this case, multiple hosts generate concurrent write requests to shared storage resources. This setup enables analysis of interference effects and illustrates how shared resource contention contributes to variability in latency.

The measurement process is designed to capture both aggregate performance indicators and detailed distributional characteristics. Latency is collected for each operation and analyzed using statistical methods. In addition to average latency and throughput, percentile-based metrics such as median (p50), p95, and p99 are used to capture both typical and extreme behavior. The workflow of latency measurement and interpretation used in the experimental analysis is illustrated in Figure 2.

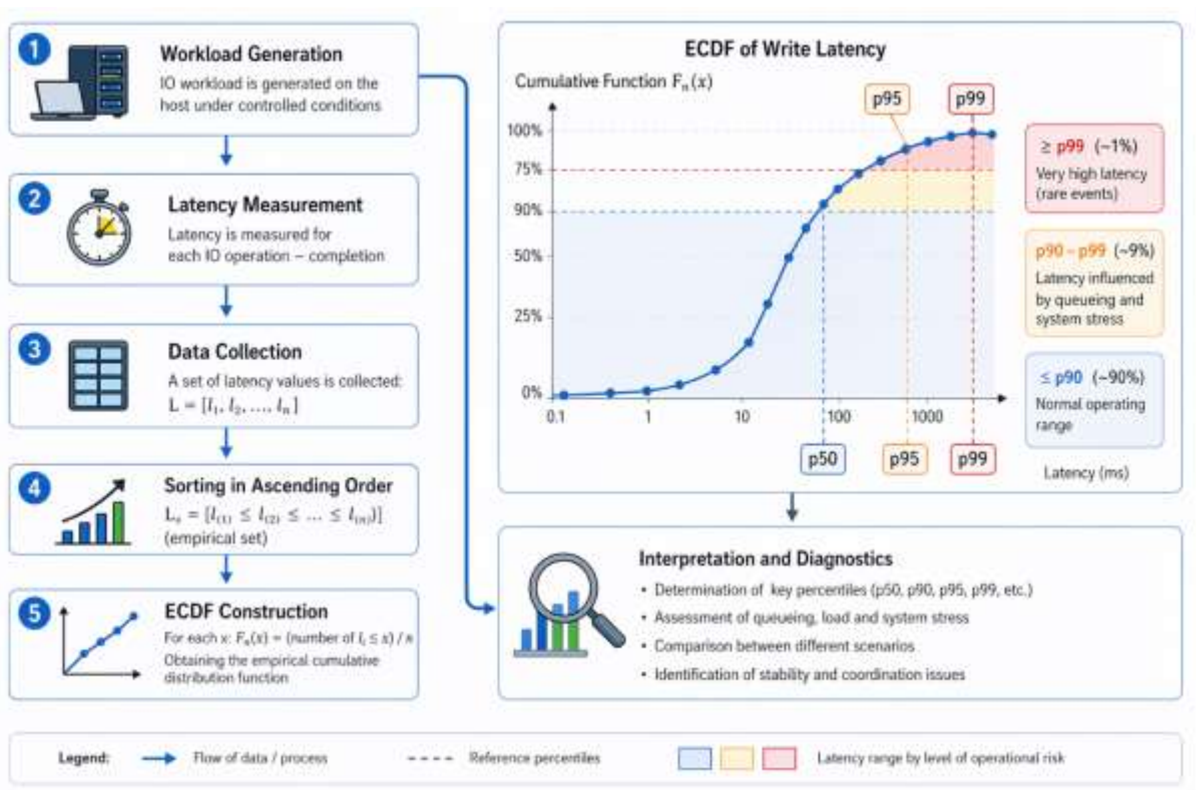

Figure 2. Workflow for measurement and interpretation of write latency using ECDF analysis

A central element of the methodology is the use of empirical cumulative distribution functions (ECDF) for analyzing latency. This approach provides a direct representation of the proportion of operations completed within a given latency threshold and enables detailed observation of the tail of the distribution. Combined with percentile-based metrics (p95 and p99), ECDF analysis enables the identification of queueing effects and early manifestations of coordination stress that remain hidden when only average performance metrics are considered.

The combination of controlled experimentation, latency distribution analysis, and ECDF visualization provides a reproducible framework for investigating the behavior of write operations under different workload conditions. This methodological approach enables the identification of nonlinear effects, queue accumulation phenomena, and interactions between the architectural layers of the SAN infrastructure. The obtained measurements constitute the experimental basis for the analytical synthesis presented in the following chapter, where the observed effects are interpreted using queueing theory and integrated into the proposed optimization framework for SAN write operations.

Chapter 4: Experimental Results

Chapter 4 presents the results of the experimental investigation of write operations in **Storage Area Network (SAN)** environments. The objective of this chapter is to evaluate the influence of workload characteristics, **I/O queue depth**, and concurrent access on write latency and to identify the conditions under which latency instability and coordination effects emerge.

The analysis begins with a **baseline experimental scenario**, which serves as a reference for system behavior under stable operating conditions. The obtained results show that, under moderate workload, the majority of write operations are completed within a narrow latency range, indicating efficient coordination between the host system, transport layer, and storage subsystem.

The influence of **I/O queue depth**, representing the degree of parallelism, is then investigated through a series of controlled experiments. The results demonstrate that increasing parallelism initially improves throughput but also leads to a gradual widening of the latency distribution. While the average latency and median remain relatively stable, the high-percentile latency values (p95 and p99) exhibit nonlinear growth, indicating the emergence of queue accumulation effects.

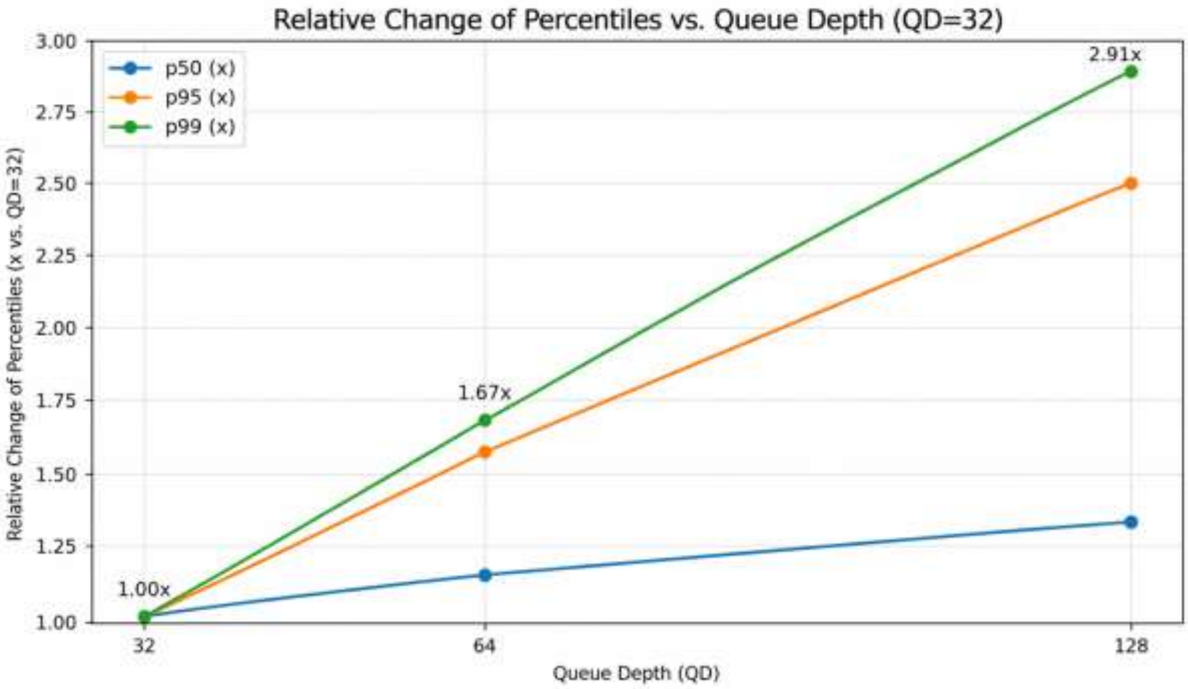

Figure 3. Relative change of latency percentiles (p50, p95, p99) as a function of queue depth

The observed behavior shows that increasing parallelism affects only a subset of write operations. The growing separation between the median and the upper latency percentiles reflects the formation of **tail latency** and suggests that the system gradually approaches a regime where queueing effects begin to dominate the latency behavior.

Additional experiments investigate the influence of workload structure by comparing uniform and dynamically varying request patterns. The results indicate that bursty workloads introduce temporary request accumulation, increasing latency variability even when the average workload intensity remains unchanged.

The effect of concurrent access is further examined through experiments involving multiple hosts accessing shared storage resources. Under these conditions, the results demonstrate a significant increase in latency variability caused by competition for shared resources, illustrating the **noisy neighbor** effect commonly observed in enterprise SAN environments.

To obtain a more comprehensive view of latency behavior, the collected measurements are analyzed using **empirical cumulative distribution functions (ECDF)**.

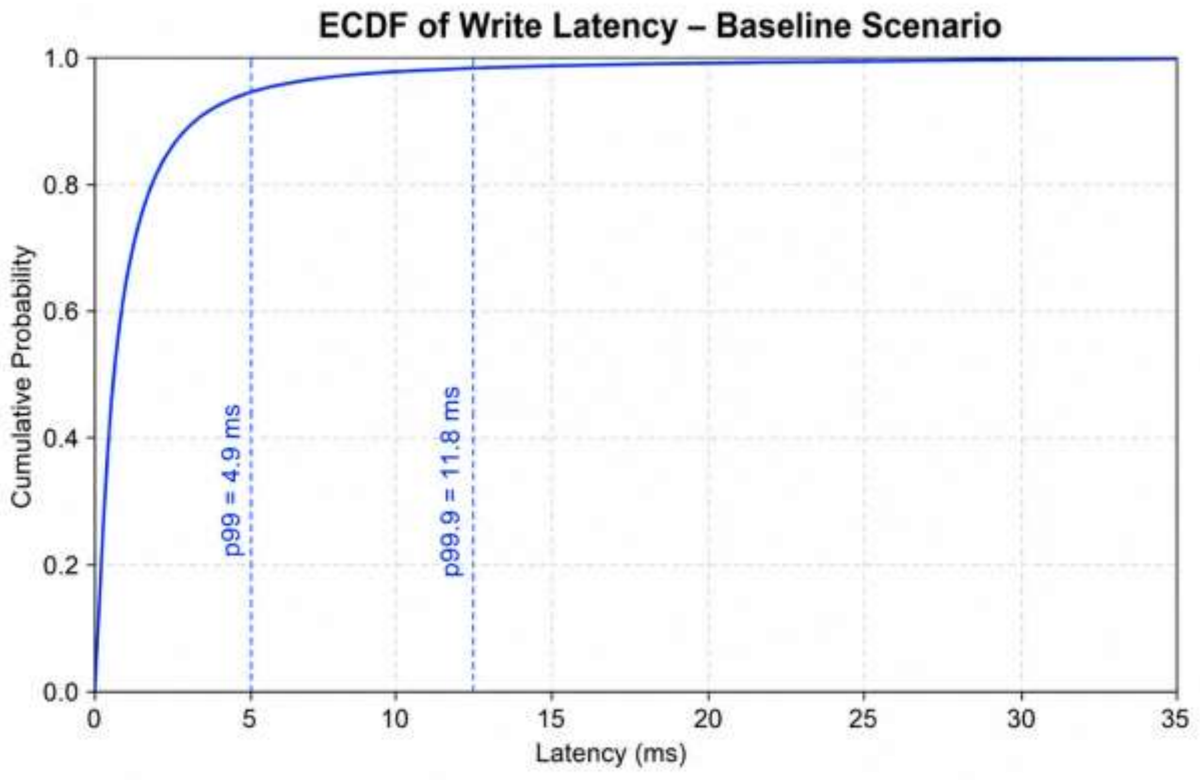

Figure 4. Figure 4. Empirical cumulative distribution function (ECDF) of write latency for the baseline experimental scenario

The ECDF analysis shows that the majority of write operations remain concentrated within a relatively narrow latency interval, whereas a small proportion of requests forms an extended latency tail. This tail determines the behavior of the high-percentile latency values and represents the first experimental indication of the emergence of coordination stress under increasing workload.

The experimental results further demonstrate that the combination of ECDF analysis and high-percentile latency metrics (p95 and p99) provides substantially more information about system behavior than average latency alone. These indicators reveal queueing effects and transient latency amplification that remain hidden when conventional performance metrics are used.

Overall, the experimental investigation confirms that write latency in SAN environments is governed not only by the performance of individual components but primarily by the interaction between workload characteristics, **I/O queue management**, and parallelism across the entire write path.

The observed relationships provide the experimental foundation for the analytical synthesis presented in the following chapter. Based on these findings, distinct operating regimes are identified, and the experimental observations are transformed into an analytical interpretation and an optimization framework for write operations in SAN environments.

Chapter 5: Analytical Synthesis and Interpretation

Chapter 5 presents the analytical synthesis of the experimental results and identifies the fundamental mechanisms governing the behavior of write operations in **Storage Area Network (SAN)** environments. Unlike the previous chapter, which reports the experimental observations individually, this chapter integrates the results into a unified analytical interpretation of the relationships between workload intensity, **I/O queue management**, parallelism, and latency behavior.

The analysis demonstrates that SAN performance cannot be adequately characterized using conventional aggregate performance metrics alone. Instead, the experimental results reveal the existence of distinct operating regimes that differ in both resource utilization and latency behavior. As workload intensity increases, the system progressively transitions from **stable operation**, through a **transition regime**, to a state of **coordination stress**, where latency variability increases rapidly.

A central aspect of the analytical synthesis is the investigation of latency distributions. Rather than focusing solely on average latency, the analysis considers the evolution of the latency tail using the **p99/p95 ratio**, which provides a normalized indicator of tail expansion.

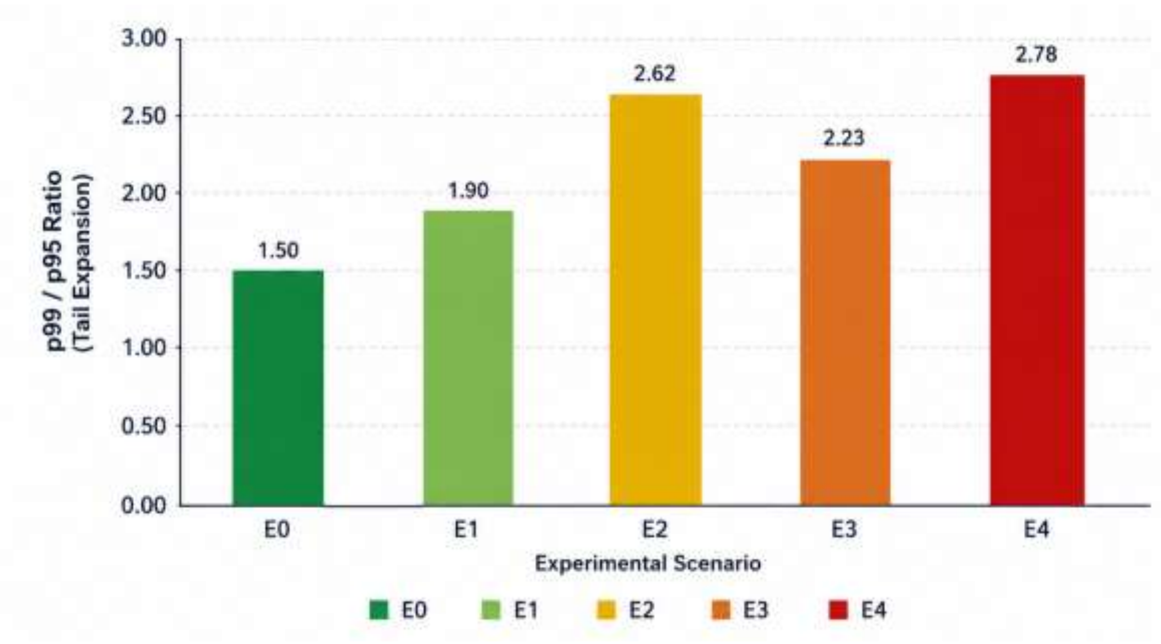

Figure 5. Tail expansion (p99/p95) under different experimental scenarios

The results demonstrate that, under moderate workload, the p99/p95 ratio remains relatively stable, indicating compact latency distributions. As workload intensity and concurrency increase, the ratio grows significantly, reflecting the emergence of rare but highly delayed write operations and the progressive expansion of the latency tail.

To further investigate this behavior, the high-percentile latency values are compared across the different experimental scenarios.

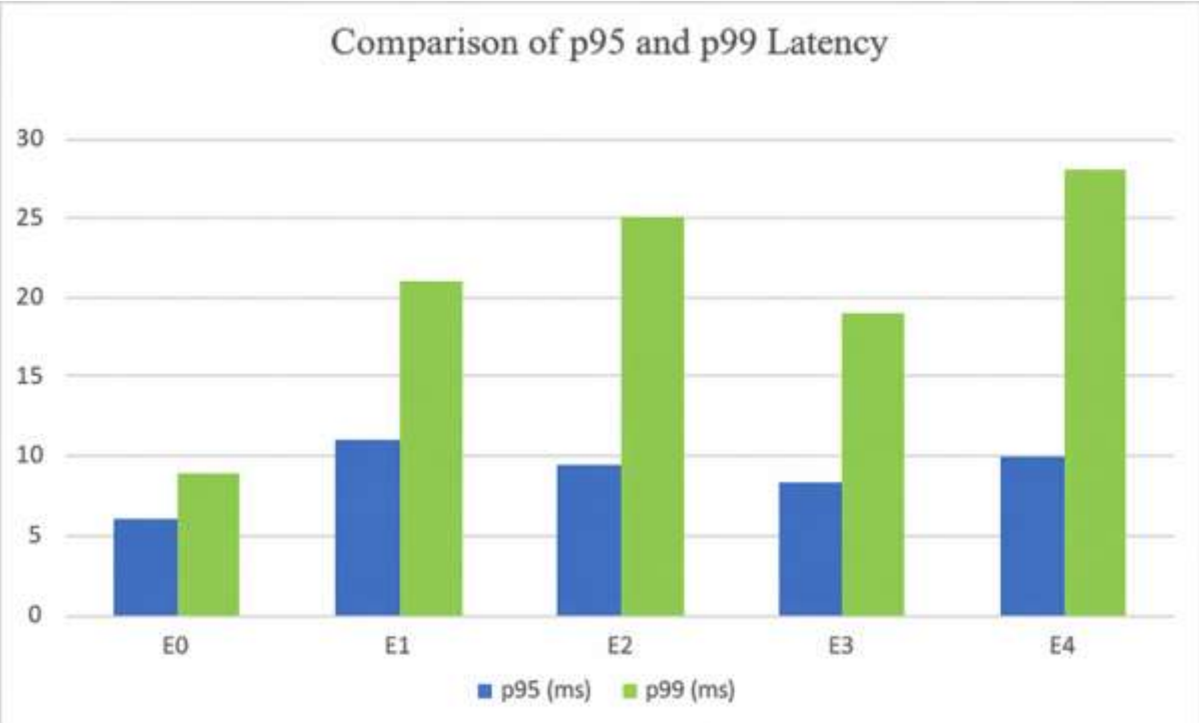

Figure 6. Comparison of p95 and p99 latency across different experimental scenarios

The comparison shows that p95 remains relatively stable, whereas p99 increases substantially faster as workload grows. This observation confirms that latency degradation first appears in the extreme tail of the distribution and therefore cannot be detected using average latency alone. Consequently, high-percentile latency represents a significantly more sensitive indicator of queue accumulation and coordination effects.

The observed behavior is interpreted using the principles of queueing theory. As the arrival rate approaches the effective processing capacity of the storage system, waiting time within the distributed I/O queues increasingly dominates the total write latency. Conceptually, this relationship may be expressed as

$$T_{latency} = T_{service} + T_{queue},$$

where T_{service}represents the effective service time and T_{queue}denotes the cumulative waiting time introduced by the interacting I/O queues along the write path.

This interpretation explains why latency variability grows even when individual hardware components remain below their nominal utilization limits. The results indicate that latency instability is primarily a consequence of coordination between distributed queueing mechanisms rather than simple resource saturation.

The analytical synthesis further demonstrates that increasing parallelism initially improves throughput but eventually amplifies queueing interactions between the architectural layers. Consequently, the system progresses through three characteristic operating regimes.

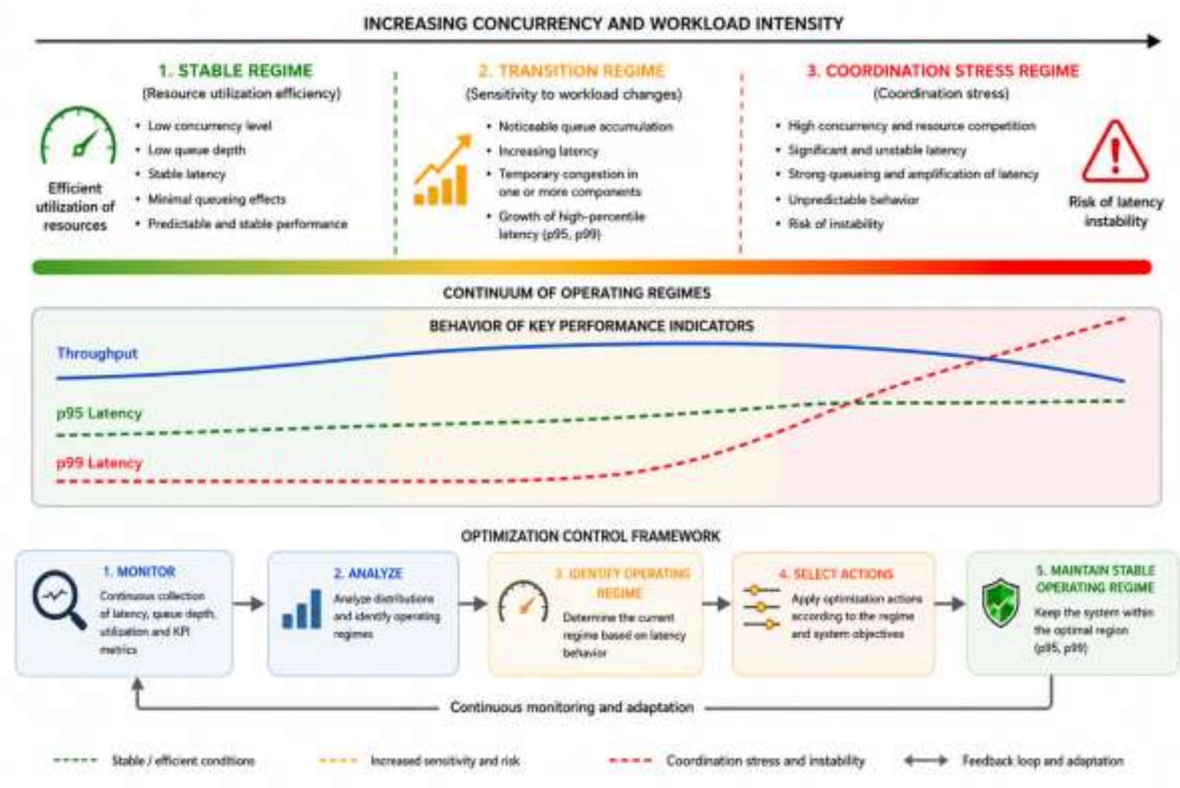

Figure 7. Operating regimes under increasing workload: Stable Operation – Transition – Coordination Stress

The results indicate that the optimal operating region is not the point of maximum throughput, but rather the workload range in which throughput increases while latency distributions remain stable and the expansion of the latency tail is limited.

Based on these observations, the experimental findings are transformed into a practical optimization methodology. The analytical interpretation is summarized through optimization guidelines relating observed latency effects to appropriate system-level actions.

Observed Effect	Underlying Cause	Indicator	Optimization Action	Expected Outcome
Increase in p99	Request accumulation	p99, ECDF	Reduce I/O queue depth	Lower tail latency
ECDF expansion	Coordination stress	p95–p99	Limit the degree of parallelism	Stabilized latency distribution
Bursty workload	Uneven workload	ECDF	Smooth the workload (traffic shaping)	Reduced latency peaks
Competition for shared resources	Resource contention	p99, latency variability	Apply Quality of Service (QoS) mechanisms / workload isolation	Reduced resource contention
High throughput with unstable latency	Overloaded queues	p99/p50	Operate within the stable operating region	Balance between performance and stability

Table 2. Optimization guidelines derived from the experimental analysis of write operations in SAN environments

The proposed optimization guidelines summarize the practical implications of the experimental analysis and provide decision support for maintaining stable latency behavior under different workload conditions.

Building upon these guidelines, the dissertation proposes a closed-loop optimization framework consisting of latency monitoring, operating regime identification, optimization decision-making, and continuous feedback.

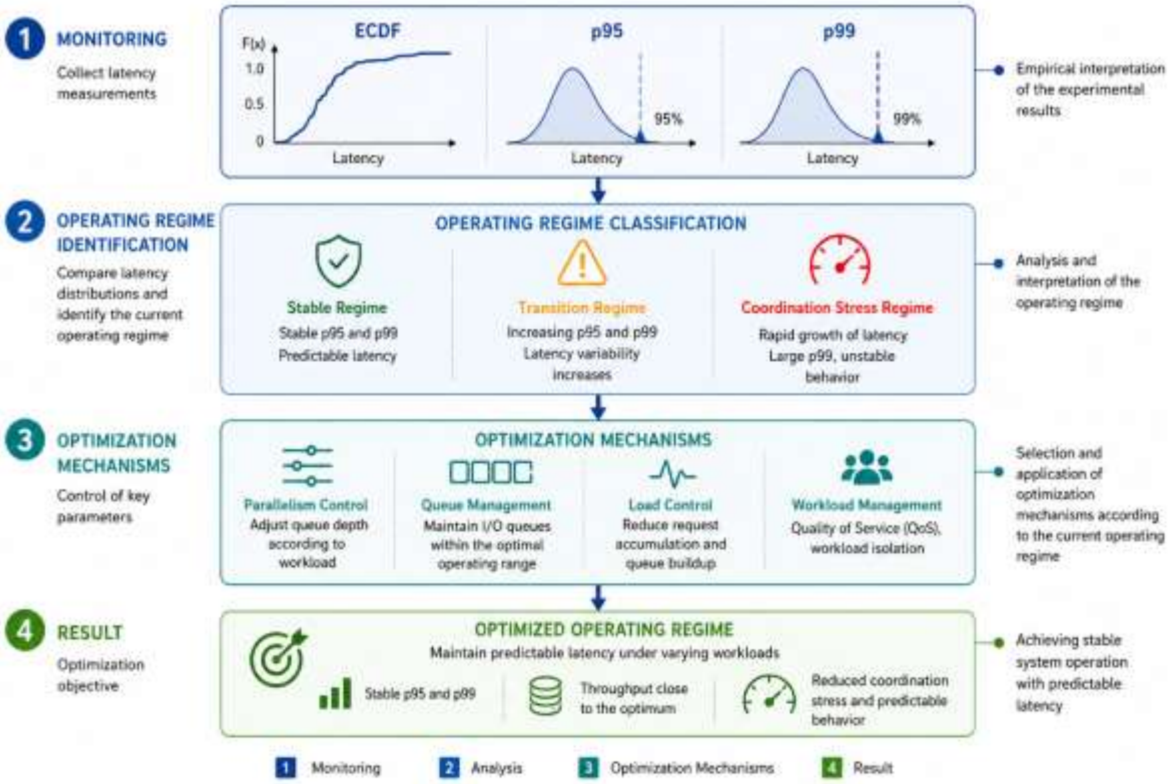

Figure 8. Optimization framework for write operations in SAN environments

The proposed framework translates the analytical findings into a practical methodology for identifying stable operating regimes, detecting early signs of coordination stress, and selecting optimization actions based on latency distributions, **I/O queue behavior**, and workload characteristics.

The relationship between the observed latency effects and the corresponding optimization actions is summarized in **Table 3**, providing a direct link between the analytical interpretation and practical system management.

Observed Effect	**Primary Indicator**	**Underlying Cause**	**Optimization Action**	**Objective**
Gradual increase in p95/p99	p95, p99	Increasing parallelism	Limit the degree of parallelism	Maintain a stable operating regime
Rapid increase in p99	p99/p95	Request accumulation	Reduce I/O queue depth	Minimize queue waiting time
ECDF expansion	ECDF	Coordination stress	Adjust workload characteristics	Stabilize the latency distribution
Latency spikes	p99	Bursty workload	Smooth the workload	Reduce latency variability

Observed Effect	Primary Indicator	Underlying Cause	Optimization Action	Objective
Resource contention	p99	Shared resources	Apply Quality of Service (QoS) mechanisms; isolate and prioritize competing workloads	Workload isolation
High throughput	Throughput and p99	Operation near system capacity	Operate within the stable operating region	Balance performance and stability

Table 3. Relationship Between Observed Latency Effects and Recommended Optimization Actions

The mapping between observed latency effects and optimization actions establishes a direct connection between the analytical interpretation and practical management of SAN environments.

The analytical synthesis confirms the central research thesis of the dissertation. It demonstrates that stable write latency is determined primarily by the interaction between workload characteristics, parallelism, and distributed queueing mechanisms rather than by the peak performance of individual storage components. Consequently, effective optimization is achieved by maintaining the system within a stable operating region in which latency distributions remain controlled, coordination effects are minimized, and high throughput is sustained without compromising latency predictability.

Chapter 6: Conclusions and Future Work

Chapter 6 summarizes the principal findings of the dissertation and outlines directions for future research.

The conducted research demonstrates that the latency of write operations in **Storage Area Network (SAN)** environments cannot be interpreted solely as a characteristic of the storage subsystem. Instead, write latency is shown to be a system-level property emerging from the interaction between workload characteristics, parallelism, and distributed I/O queueing mechanisms across the host system, transport infrastructure, and storage subsystem.

The experimental investigation confirms that increasing parallelism initially improves throughput but also leads to progressive expansion of the latency distribution. While average latency remains relatively stable over a wide operating range, the high-percentile values (p95 and p99) increase significantly under higher workload levels, indicating the emergence of queue accumulation and coordination effects. These observations make it possible to distinguish three characteristic operating regimes: **stable operation, transition, and coordination stress**.

The analytical synthesis demonstrates that latency instability is primarily caused by the interaction between distributed queueing mechanisms rather than by simple resource saturation. Interpreting the experimental results through the principles of queueing theory provides a consistent explanation of the observed latency behavior and supports the proposed system-level interpretation of SAN write operations.

A major outcome of the dissertation is the development of an **optimization framework** that transforms the experimental findings into a practical methodology for monitoring latency behavior, identifying operating regimes, and selecting optimization actions based on latency distributions, I/O queue behavior, and workload characteristics. This framework emphasizes that effective optimization is achieved by maintaining the system within a stable operating region rather than by maximizing throughput alone.

The dissertation contributes to a deeper understanding of the mechanisms governing latency behavior in modern SAN environments. The proposed system-level model, the experimental methodology, the analytical interpretation, and the optimization framework together provide an integrated approach for analyzing and optimizing write operations in enterprise storage infrastructures.

Future research may extend the proposed methodology by investigating read operations, mixed read/write workloads, heterogeneous storage environments, and transport-layer dynamics. Additional work may also explore adaptive optimization techniques, machine learning–based prediction of operating regimes, and intelligent mechanisms for coordinated management of parallelism and distributed I/O queues under dynamically changing workloads.

9. MAIN RESULTS OF THE DISSERTATION

As a result of the conducted research, the following main results have been obtained:

1. A conceptual system-level model of the write path has been developed, representing write operations in Storage Area Network (SAN) environments as a system of interacting request management, parallelism control, and distributed I/O queueing mechanisms across the host, transport, and storage layers.
2. A reproducible experimental methodology has been developed for investigating write operations in SAN environments under controlled conditions through systematic variation of workload characteristics while maintaining a fixed system configuration.

3. An extensive experimental investigation has been performed, demonstrating the influence of workload characteristics, I/O queue depth, parallelism, and competing workloads on write latency and latency distribution.
4. It has been experimentally demonstrated that increasing I/O queue depth results in nonlinear growth of high-percentile latency (p95 and p99) and progressive expansion of the latency distribution, even when average latency and throughput remain relatively stable.
5. The experimental results have identified three characteristic operating regimes—stable operation, transition, and coordination stress—revealing the relationship between workload intensity, queue accumulation, and latency instability.
6. An analytical synthesis of the experimental results has been developed, interpreting the observed latency behavior through the principles of queueing theory and demonstrating that latency instability is primarily caused by interactions between distributed queueing mechanisms rather than by simple resource saturation.
7. A practical optimization framework has been proposed, integrating latency monitoring, operating regime identification, analytical interpretation, and optimization actions to support stable and predictable operation of SAN write workloads.

10. CONTRIBUTIONS

The following scientific contributions have been achieved in this dissertation:

1. A system-level representation of write operations in Storage Area Network (SAN) environments has been developed, in which write latency is interpreted as the result of interactions between multiple architectural layers and distributed queueing mechanisms along the write path.
2. An analytical model has been developed for the interpretation and optimization of SAN system behavior under increasing workload, enabling the identification of three characteristic operating regimes: stable operation, transition, and coordination stress.
3. A reproducible experimental methodology has been developed for investigating SAN systems under controlled conditions through systematic variation of workload parameters while maintaining a fixed system configuration.

4. A methodology for analyzing and evaluating write operations has been developed, based on empirical cumulative distribution functions (ECDF) and high-percentile latency metrics (p95 and p99), enabling the identification of queueing effects and their use as optimization criteria for latency behavior.
5. It has been established that increasing the degree of parallelism does not necessarily lead to sustainable performance improvement, but under certain operating regimes amplifies latency variability and increases high-percentile latency values.
6. It has been demonstrated that queueing effects constitute a key indicator of coordination stress, which cannot be adequately characterized using conventional average performance metrics.
7. An analytical framework for the optimization of write operations in SAN environments has been developed, based on latency distribution analysis, queueing indicators, and the interaction between workload characteristics and I/O queue behavior, enabling the identification of stable operating regimes and the formulation of practical optimization guidelines.
8. Principles for optimizing write operations in SAN environments have been formulated through the coordinated management of parallelism, I/O queues, and latency percentiles, aimed at limiting coordination stress and maintaining stable latency behavior under concurrent workloads.

12. PUBLICATIONS RELATED TO THE DISSERTATION

[1] Torres G.. *Entropy-Aware AI-Powered Ransomware Detection Using IBM FlashCore Module 4 in SAN Environments*. International Conference Automatics and Informatics (ICAI), 2026, ISBN:979-8-3315-8611-9, DOI:10.1109/ICAI67591.2025.11325084, 100-106; 2025

[2] Torres G.. *Optimization of Write Operations in SAN Environments Through Intelligent Compression, Deduplication and Thin Provisioning*. TELECOM 2025, 33, 33rd National Conference with International Participation (TELECOM), 2026, ISBN:978-1-6654-6564-9, ISSN:2837-5246, DOI:10.1109/TELECOM66943.2025.11304087, 1-4; 2025

[3] Torres G.. Storage Virtualization: *Optimizing Data Center Efficiency and Performance*. 2/2024, 2024, ISSN:2535-0668, 65-66; 2024

[4] Torres G., Ivanov Vl.. *A Survey on Storage Area Network*. Robotics, Automation and Mechatronics '23, Prof. Marin Drinov Publishing House of Bulgarian Academy of Science, 2023, ISSN:1314-4634, 22-26; 2023

[5] Торрес Г., Иванов Вл.. *Приложения на Fiber Channel, ISCSI, Infiniband и FCOE в Storage Area Network*. Robotics, Automation and Mechatronics '23, RAM 2023, Prof. Marin Drinov Publishing House of Bulgarian Academy of Sciences, 2023, ISSN:1314-4634, 30-32; 2022

[6] Торрес Г.. *Анализ на методи за оптимизация на записи в SAN среда*. Prof. Marin Drinov Publishing House of Bulgarian Academy of Sciences, 2022, ISSN:1314-4634, 69-72; 2022

13. REFERENCES

L. Kleinrock, *Queueing Systems, Volume I: Theory*. New York, NY, USA: Wiley-Interscience, 1975.

E. D. Lazowska, J. Zahorjan, G. S. Graham, and K. C. Sevcik, *Quantitative System Performance: Computer System Analysis Using Queueing Network Models*. Englewood Cliffs, NJ, USA: Prentice Hall, 1984.

R. Jain, *The Art of Computer Systems Performance Analysis*. New York, NY, USA: Wiley, 1991.

M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge, U.K.: Cambridge University Press, 2013.

A. B. Downey, *Think Stats: Probability and Statistics for Programmers*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2014.

S. Asmussen, *Applied Probability and Queues*, 2nd ed. New York, NY, USA: Springer, 2003.

A. S. Tanenbaum and H. Bos, *Modern Operating Systems*, 4th ed. Boston, MA, USA: Pearson, 2015.

J. H. Saltzer and M. F. Kaashoek, *Principles of Computer System Design*. Burlington, MA, USA: Morgan Kaufmann, 2009.

C. Min et al., "Understanding the characteristics of flash storage," in *Proc. IEEE MSST*, 2014.

J. Dean and L. A. Barroso, "The tail at scale," *Communications of the ACM*, vol. 56, no. 2, pp. 74–80, 2013.

A. Fox and E. Brewer, "Harvest, yield, and scalable tolerant systems," in *Proc. HotOS*, 1999.

S. Kanev et al., "Profiling a warehouse-scale computer," in *Proc. ISCA*, 2015.

A. Klimovic et al., "Understanding ephemeral storage performance in the cloud," in *Proc. USENIX ATC*, 2018.

J. Leverich and C. Kozyrakis, "Reconciling high server utilization and sub-millisecond QoS," in *Proc. EuroSys*, 2014.

R. Lu et al., "NVMe SSD failures in the field: The fail-stop and the fail-slow," in *Proc. USENIX ATC*, 2022.

I. Jo, "Toward ultra-low latency SSDs: Analyzing the impact on data-intensive workloads," *Electronics*, vol. 12, no. 1, 2023.

Y. Zhang et al., “Characterizing tail latency in cloud storage systems,” in *Proc. IEEE IC2E*, 2022.

Z. Huang et al., “FairNVMe: State-driven fairness control for efficient I/O queue scheduling,” *Future Generation Computer Systems*, vol. 155, 2025.

A. Yang et al., “Set-associative mapping to improve data throughput and reduce long tail latency in SSDs,” *Journal of Systems Architecture*, vol. 155, 2025.

S. Li et al., “Understanding performance variability in modern NVMe SSDs,” in *Proc. USENIX FAST*, 2023.

M. Bjørling et al., “The Linux block layer and NVMe: Performance implications for modern storage,” *ACM Transactions on Storage*, vol. 19, no. 2, 2023.