Assessing the Impact of Diverse Scheduling Strategies on Digital Library System Performance

Haoqin Shi [1, *]

[1] Beijing University of Posts and Telecommunication, Beijing, 102200, China
* jp2021212630@qmul.ac.uk

Abstract. With the evolution and maturity of cloud computing and big data technologies, the shift from traditional libraries to digital libraries has emerged as a prevailing trend. To comprehend the current scheduling strategies employed within digital libraries, and to address challenges such as server congestion and the absence of robust scheduling strategies, this study aims to evaluate and compare various scheduling techniques found in digital libraries. This article offers a comprehensive overview of four primary categories of scheduling strategies, along with the essential components of digital library systems. These insights, gleaned from an exhaustive review of existing literature, set the stage for a more in-depth discussion on the application of these scheduling strategies within digital library systems. Subsequently, the study delves into a comparative analysis of the application, advantages, and drawbacks of these four key categories of scheduling strategies in the context of managing virtual resources. The discussion is divided into four segments, each devoted to a particular class of scheduling strategies.

Keywords: Scheduling Strategies, Digital Library Systems, Performance

1 Introduction

With the rapid advancement of digital technology, numerous traditional libraries and literature resource institutions have begun their journey of digital transformation. The process entails digitizing physical documents and resources, as well as constructing digital library systems to facilitate online access and services [1]. The open access concept has been instrumental in propelling the growth of digital libraries [2]. A rising number of academic publications, research data, and literary resources are being made available via open access, granting users worldwide the liberty to freely access and utilize these resources. Digital libraries go beyond offering e-books and academic literature, extending their services to the storage and management of multimedia resources such as audio, video, and images. This diversification enhances the learning and research experiences of users [3].

Within a digital library system, various scenarios necessitate the implementation of scheduling strategies. For instance, digital library systems frequently encounter circumstances where vast numbers of users seek access to resources simultaneously

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_49

[4]. Consequently, efficient scheduling strategies are required to manage and allocate virtual resources effectively, thereby ensuring equitable access and delivering an optimal user experience. The scheduling of virtual resources in a digital library system should aim to maximize resource utilization to enhance system efficiency and performance. Resources should be distributed judiciously to optimize their use, thereby minimizing waste and idleness. As users of digital libraries have high expectations regarding response times, scheduling strategies ought to focus on reducing user wait times, ensuring swift resource access and response to bolster user satisfaction [5].

Given that resource demands and user access patterns in a digital library system may fluctuate over time, scheduling strategies should exhibit dynamic adaptability. This involves making scheduling decisions based on real-time system load and resource conditions, thereby accommodating evolving needs [6].

2 Theoretical Background

2.1 Understanding Scheduling Strategies: An Overview

Scheduling policies can be divided into four categories. They are non-preemptive, non-size based policies, preemptive, size-based policies, non-preemptive, size-based policies, preemptive, size-based policies.

Non-preemptive, non-size based policies refer to scheduling strategies that do not allow tasks to be interrupted or preempted once they start executing and do not consider the size or resource requirements of tasks when making scheduling decisions. These policies prioritize other factors such as task priority, arrival order, or specific criteria. Common Non-preemptive, non-size based policies are First-Come-First-Serve (FCFS), Last-Come-First-Serve (LCFS), Random strategy [7].

Preemptive, size-based policies are scheduling strategies that allow tasks to be interrupted or preempted by higher-priority tasks and take into account the size or resource requirements of tasks when making scheduling decisions. These policies prioritize tasks based on their sizes or resource needs and allow for dynamic allocation and reallocation of resources.Common Preemptive, size-based policies are Preemptive Shortest Job First (PSJF), Shortest Remaining Processing Time (SRPT).

Non-preemptive, size-based policies are scheduling strategies that consider the size or resource requirements of tasks but do not allow for preemption or interruption of tasks once they start executing. These policies prioritize tasks based on their sizes or resource needs and schedule them in a non-preemptive manner. Common Non-preemptive, size-based policies are Shortest Job First (SJF) strategy [8].

Preemptive, non-size based policies are scheduling strategies that allow tasks to be interrupted or preempted by higher-priority tasks and do not consider the size or resource requirements of tasks when making scheduling decisions. These policies prioritize tasks based on factors other than their sizes, such as task priority, arrival time, or specific criteria. Common preemptive, non-size-based policies are Preemptive LCFS, processor sharing(PS)

2.2 The Mechanics of Digital Library Systems

The author considers digital libraries as a new type of library. In his book Redesigning Library Services, Professor M. Buckland, President of the American Society for Information Science, divides the history of libraries into three phases according to the technology used in libraries: paper-based traditional libraries, mechanized modern libraries and digital libraries of the future [9]. In digital libraries, the management and scheduling of cloud resources require a high level of intelligence, mainly including the virtualization management of cloud computing resources, cloud storage resources and cloud network resources, as well as the virtualization of cloud reader application services, dynamic scheduling of cloud resources, automated deployment of cloud systems and resources, and intelligent and optimal configuration management of scheduling policies.

This paper will discuss the application and comparison of four major types of scheduling strategies in these sessions [10].

3 Evaluation and Comparative Analysis of Diverse Scheduling Strategies

3.1 The Application of Non-Preemptive, Non-Size-Based Policies

When performing background data synchronization and backup tasks in digital library systems, a non-preemptive, non-size dependent scheduling policy such as First Come First Serve (FCFS) is appropriate.

The FCFS policy schedules tasks in the order of their arrival, ensuring that tasks are executed sequentially in the order in which they are submitted. For background data synchronization and backup tasks, this policy provides simple and efficient scheduling to ensure that tasks can be executed in an orderly fashion. Specifically, when background data synchronization and backup tasks are submitted to the system, the system processes the tasks in the order in which they are submitted, and each task is executed in turn until it completes.

Using the FCFS policy can avoid competition and preemption among tasks, ensure that background tasks are executed in the order they are submitted, and reduce resource conflicts and unnecessary context switching overhead. In addition, FCFS policies are easy to implement and manage, making them suitable for background task scenarios with relatively small task volumes and relatively predictable task times. However, please note that the FCFS policy does not consider the priority of tasks or the size of tasks, so if there are different priorities or large differences in resource requirements between background tasks, other scheduling policies may need to be considered or appropriate optimizations made.

3.2 The Role of Preemptive, Size-Based Policies

In digital library system, Shortest Remaining Time First (SRTF) policy can be used to handle the book reservation queue scheduling tasks, and the tasks are scheduled according to the principle of shortest remaining execution time. The specific processing process is as follows:

When a user submits a book reservation request, the system adds the request to the book reservation queue and records the reservation time and the remaining available quantity of books.

When a book is available for reservation, the system selects the task with the shortest remaining execution time from the book reservation queue. The remaining execution time of a task can be obtained by estimation or dynamic update, usually based on the current number of available books and the reservation time.

The system removes the selected task from the book reservation queue and assigns the book to the corresponding user for reservation.

During the reservation period, if another user submits a shorter reservation request, the system dynamically adjusts the currently executed task according to the change in the remaining execution time. If there is a new reservation request with a shorter remaining execution time than the current task, the system will seize the current task and allocate resources to the new task.

When the reservation time of a task ends or the task is completed, the system checks the next task in the book reservation queue and continues to process the task according to the principle of the shortest remaining execution time.

By using the SRTF policy for book reservation queue scheduling, the system can prioritize the tasks with the shortest remaining execution time and satisfy the users' reservation needs as soon as possible. This can improve the efficiency and user experience of the library system, reduce user waiting time, and ensure fair distribution of books.

3.3 Advantages and Disadvantages of Non-Preemptive, Size-Based Policies

In a parallel book download scenario in a digital library, the Shortest Job First strategy can optimize the user's download speed very well.

First, all the book jobs to be downloaded are sorted by file size, from smallest to largest. This allows the smaller book jobs to be ranked first and the larger ones to be ranked second. In the sorted order, download tasks are assigned and processed sequentially. The system processes the book task with the smallest file size first, and then progressively processes the larger tasks. When there are multiple available download threads or resources, the smaller tasks are assigned resources and downloaded first. During parallel book downloads, dynamic scheduling can be performed based on completed tasks and remaining tasks. If smaller tasks arrive or previous tasks have completed, these smaller tasks can be immediately prioritized to ensure the shortest download time.

3.4 A Closer Look at Preemptive, Non-Size-Based Policies

In digital library systems, multi-user search and filtering functions involve processing search requests from multiple users simultaneously. the Processor Sharing (PS) strategy can be used to equitably allocate processing resources to each user to improve the efficiency of search and filtering.

The specific steps for applying the Processor Sharing policy are as follows:

User request reception: When a user submits a search request or filtering criteria, the system receives and records the content and relevant parameters of the request.

Resource allocation: The system allocates processing resources (e.g. processor time slice) to each user proportionally according to the Processor Sharing policy. The ratio can be adjusted according to the system's configuration and needs to achieve fair sharing.

Concurrent execution: Each user's search and filtering operations will be performed simultaneously, with each user receiving a certain amount of execution time based on its allocated processing resources. The system can switch the execution of different users' operations through time slice rotation or other scheduling algorithms.

Response time control: To improve user experience, the system can set a maximum response time limit. If a user's search or filtering operation exceeds the set time limit, the system can interrupt the operation and switch to the next user's operation to ensure that other users can also be processed.

Result return: When each user's search or filtering operation is completed, the system will return the corresponding result to the user.

By using the Processor Sharing policy, the digital library system can process multiple users' search and filtering requests at the same time and allocate the processing resources to each user in a fair manner. This can improve the efficiency of searching and filtering, reduce the waiting time of users, and ensure that each user can get a certain amount of processing time to achieve justice and fairness.

4 Discussion

Through the study of the specific applications of four major scheduling strategies in digital library systems, the author has identified the limitations of existing scheduling strategies in addressing specific system problems. In most cases, employing a simple strategy like First-Come-First-Serve (FCFS) does not pose significant issues as background data synchronization and backup tasks are relatively straightforward. However, such a scheduling strategy is not adept at handling cases with varying priorities or significant disparities in virtual resource requirements. For instance, if the backup process fails to accumulate a large amount of low-priority data, utilizing this scheduling strategy for backup would result in the neglect of critical real-time backup requirements. Consequently, important data would go unprocessed, which is unfavorable for short jobs with high priority. Therefore, future research endeavors should delve into exploring alternative scheduling strategies to be employed after backup failures.

5 Conclusion

By examining the specific applications of four major scheduling strategies in digital library systems, the author has reached the conclusion that scheduling strategies play a vital role in optimizing resource utilization, enhancing system performance, and delivering a satisfactory user experience. The selection of an appropriate scheduling strategy necessitates a comprehensive evaluation of the system's characteristics, task attributes, and user requirements. During the investigation of scheduling strategies in digital libraries, the author also identified limitations and shortcomings in the existing approaches for handling associated tasks, which can be addressed in future research endeavors.

References

1. Ma Xiaoting, Chen Chen. Research on cloud computing resource scheduling and virtualized resource management mechanism for digital libraries. Journal. (2012)
2. Author, Michael Buckland.: Redesigning Library Services. A Manifesto. Chicago: American Library Association, USA (1992).
3. Zheng, Q., & Xiang, M. (2016). Scheduling strategies for mixed-load AGVs in automated container terminal. Automation in Construction, 72, 51-61.
4. Hsiao, H. F., & Chu, P. C. (2012). A novel scheduling strategy for cloud-based distributed data-intensive applications. Journal of Computer and System Sciences, 78(5), 1342-1356.
5. Qing, C., Zhang, L., & Liu, H. (2015). Task scheduling in heterogeneous distributed systems with digital library applications. Journal of Supercomputing, 71(6), 2121-2142.
6. Yu, D. C., & Lin, C. N. (2013). A digital library system architecture for displaying cultural and historical exhibits. Journal of Network and Computer Applications, 36(2), 811-820.
7. Chen, Y., Wu, R., & Wang, G. (2012). Research on the construction of local digital library system in university libraries: Taking Huazhong University of Science and Technology Library as an example. Journal of Academic Libraries, 30(6), 25-31.
8. Zhang, L., Li, H., & Zhao, J. (2018). A hybrid scheduling strategy for parallel job scheduling in cloud computing systems. Journal of Systems and Software, 137, 456-469.
9. Huang, H. C., Chou, S. Y., & Lio, C. H. (2014). A novel metadata-based information retrieval system for digital libraries. Journal of Intelligent Information Systems, 43(2), 261-290.
10. Li, Z., Zhang, H., & Chen, Y. (2014). A hybrid strategy for scheduling tasks in heterogeneous cluster systems. Journal of Systems Architecture, 60(9), 726-737.