Exploiting Analytical Models to Realize the Full Performance and Parallelization Potential of Modern Storage Architectures for Big Data Applications

Anshul Gandhi and Erez Zadok, Computer Science

Storage is often the performance bottleneck for today's applications, including data analytics frameworks such as HDFS-backed big data processing algorithms and online services such as database-backed, customer-facing web services (e.g., Amazon and Facebook). While state-of-the-art storage hardware promises much parallelism and improved performance, the workload characteristics of today's applications are complex and heavy-tailed (in terms of latency), making it difficult to fully exploit the potential of modern storage systems. As a result, applications continue to rely on the sub-optimal storage stack software that was designed for previous device generations. We propose to develop novel I/O schedulers that are tailored for the heavy-tailed workload characteristics of big data applications, enabling the full utilization of the parallelism offered by modern storage hardware. Our key idea is to first investigate the distribution of request-level characteristics (e.g., inter-arrival times) in big data applications, and then develop stochastic models that can accurately predict the latency for the storage stack. These models will then be leveraged by the I/O scheduler to redistribute requests in the storage stack to increase throughput and parallelism while avoiding I/O contention. We will use an integrated theory-systems approach to design a rigorous solution. We will employ techniques from queueing theory, control theory, and machine learning to analyze big data applications and model their performance. We will then design and implement a Linux I/O scheduler that leverages our model to significantly improve the storage performance, closing the gap between theoretical and achieved performance for modern storage devices. We will systematically evaluate our approach using several big data applications, including those used by our data science faculty. We have already obtained some preliminary and promising results to validate our proposed approach. We experimented with storage workloads, such as file servers and database servers, and analyzed their workload characteristics. We found that the heavy-tailed characteristics of storage workloads can be modeled fairly well via the seldom used Hyper-exponential distribution. Based on this result, we developed a simple stochastic model that can estimate the performance of storage workloads. We applied this model to a Flash-based device and showed that we can accurately predict the mean response time for storage workloads, unlike other models that are typically used in the literature.

We have already obtained some preliminary and promising results to validate our proposed approach. We experimented with storage workloads, such as file servers and database servers, and analyzed their workload characteristics. We found that the heavy-tailed characteristics of storage workloads can be modeled fairly well via the seldom used Hyper-exponential distribution. Based on this result, we developed a simple stochastic model that can estimate the performance of storage workloads. We applied this model to a Flash-based device and showed that we can accurately predict the mean response time for storage workloads, unlike other models that are typically used in the literature.

PI Qualifications. PIs Gandhi and Zadok are from the CS department at SBU, and have collaborated in the past on research proposals, one funded project, and publications. The PIs have complementary expertise and just enough overlap in research interests to ensure a synergistic collaboration and a successful project. Gandhi's relevant expertise includes various aspects of performance modeling for computer systems including queueing theory, control theory, and machine learning. Zadok is an expert on OS software and storage-stack development, with a long history of performance analysis, benchmarking, and optimizations.

Timeline. We will pursue the above research plan aggressively over the next 12 months and target a full proposal to the NSF CNS Medium deadline next October. We believe that a prototype I/O scheduler that delivers performance improvements for big data applications will significantly bolster our NSF proposal. The goal of this SEED proposal is to do exactly that.