Should I Run My Cloud Benchmark on Black Friday?

Sören Henning,[1] Adriano Vogel,[1] Esteban Perez-Wohlfeil,[1] Otmar Ertl,[1] Rick Rabiser[2]
[1]Dynatrace Research, Linz, Austria
[2]LIT CPS Lab, Johannes Kepler University Linz, Austria
{firstname.lastname}@{dynatrace.com, jku.at}

Abstract

Benchmarks and performance experiments are frequently conducted in cloud environments. However, their results are often treated with caution, as the presumed high variability of performance in the cloud raises concerns about reproducibility and credibility. In a recent study, we empirically quantified the impact of this variability on benchmarking results by repeatedly executing a stream processing application benchmark at different times of the day over several months. Our analysis confirms that performance variability is indeed observable at the application level, although it is less pronounced than often assumed. The larger scale of our study compared to related work allowed us to identify subtle daily and weekly performance patterns. We now extend this investigation by examining whether a major global event, such as Black Friday, affects the outcomes of performance benchmarks.

1 Introduction

With the ongoing transition to cloud-based deployments in many organizations, conducting benchmarks and performance experiments in the cloud has also become common practice in both research and engineering. Cloud environments are widely available in many organizations and provide a realistic testbed that closely reflects production deployments. However, caution is urged when interpreting performance measurements in such settings [3]. Cloud workloads share underlying infrastructure with other tenants and the high abstraction from hardware introduces variability [1, 2]. As a result, performance measurements may fluctuate, impairing reproducibility and hindering meaningful comparisons across studies.

In a recent study [8], we empirically assessed the significance of performance variability in the cloud through a longitudinal investigation for the case of distributed stream processing applications. By repeatedly executing the same benchmark over several months, we collected a large dataset that enables quantitative analysis of performance fluctuations and, beyond related work,[1] provides an updated, application-level characterization over longer time spans. In this paper, we summarize our key findings and complement them by an investigation of cloud performance variability in the context of the Black Friday event. Black Friday is a global shopping event that is known to cause massive increases in web traffic and, hence, cloud resource demand [5]. We investigate whether such an event measurably affects the observed performance of benchmark executions.

[1]See our conference paper for a more detailed discussion of the literature [8].

2 Experiment Design

Test subject. Our study design focuses on stream processing applications as a representative type of data-intensive, performance-critical distributed systems [6]. These applications process continuous streams of data with low (often sub-second) latency, involve heavy CPU and network usage, while also having to maintain properties such as fault-tolerance, scalability, resource efficiency. As test subject for our experiments we use the open-source, cloud-native stream processing benchmark ShuffleBench [7] with its Kafka Streams implementation.

Performance measurements. As performance metric for which we quantify variability, we focus on throughput captured according to ShuffleBench's adhoc measurement method [7]. The obtained throughput values provide a good estimate of the load a similar real-world application could sustain under typical operating conditions. Achieving high throughput while minimizing required computing resources is a core optimization target. Within one benchmark execution, we continuously measure the application's throughput, discard the first 3 minutes as warm-up, and take the average over the remaining duration as result of the benchmark. Each benchmark execution thus produces one average throughput value for the subsequent analysis in Section 3 and Section 4.

Execution environment. We run our benchmarks in managed Kubernetes environments, which is a common choice for operating large-scale, distributed, and data-intensive software systems in the cloud. We use the largest cloud provider Amazon Web Services (AWS) with its Elastic Kubernetes Service (EKS) offering. The Kubernetes cluster consists of 10 nodes provisioned with *m6i* instances of different size [8] in the *us-east-1* region.

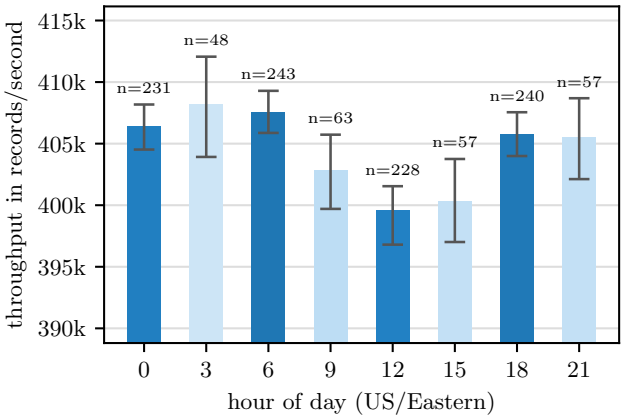

Figure 1: Measured throughput summarized by the hour of the day; color intensity reflects sample size [8].

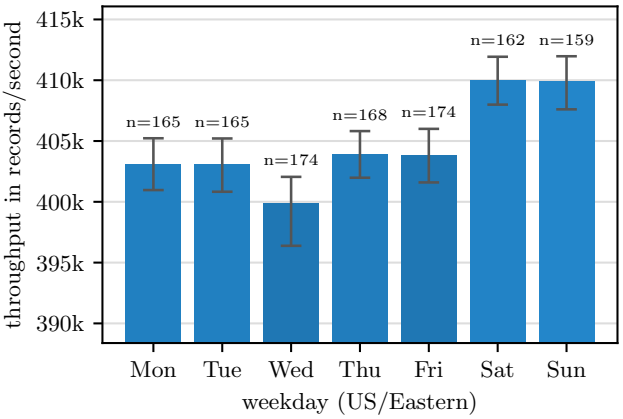

Figure 2: Measured throughput summarized by the day of the week; color intensity reflects sample size [8].

Automated benchmark execution. A scheduled task in AWS Elastic Container Service (ECS) automates periodic benchmark execution. It provisions a new EKS cluster and installs the benchmarking infrastructure, including Apache Kafka, monitoring tooling, and the Theodolite benchmarking framework [4]. Once the setup is complete, the ECS task initiates the execution of the ShuffleBench benchmark through Theodolite. Theodolite launches the stream processing application and ShuffleBench's load generator, keeps them running for 15 minutes, and collects the monitored throughput data. The benchmark execution is repeated three times according to Theodolite's configuration. Finally, benchmark results are stored for later analysis, the benchmarking infrastructure is uninstalled, and the cluster is decommissioned.

Time spans. Between May and July 2024 as well as for one week in September 2024, the periodic benchmarking task was configured to run every 6 hours to cover a full daily cycle. For a period of 3 weeks, we additionally reduced the time between experiments to 3 hours to capture a more fine-grained daily pattern. Each task execution runs the benchmark three times to account for performance variability within the same infrastructure. In Section 4, we report on an additional week of experiments around Black Friday 2024.

3 Cloud Performance Patterns

From our more than one thousand benchmark executions, we obtain a clear picture of performance variability in the cloud. The distribution of throughput measures shows a clear central tendency and almost symmetry in the interquartile range, but is not normal distributed due to slight skewness toward lower throughput results. We observe a coefficient of variation (CV), a common measure for quantifying performance variability, of 3.69%. This is on the lower end of the wide range of variability reported for micro and system-level benchmarks in the literature. Moreover, 50% of all measurements are within -2.4% and $+2.3\%$ of the median (i.e., the interquartile range). This leads us to conclude that cloud performance variability clearly exists, but contrary to what is sometimes assumed, it is not inherently detrimental when benchmarking on the application level.

Daily pattern. To break down our results, we investigate whether the performance variability exposes a daily pattern. For this purpose, we summarize all results by the hour of the day when the corresponding experiment was executed. Figure 1 shows the mean observed throughput per hour of day with its corresponding confidence intervals (obtained via bootstrapping). We observe a subtle yet statistically significant daily pattern in performance. Benchmarks executed around noon tend to exhibit slightly lower performance, whereas those conducted during late-night and early-morning hours achieve the highest results with a difference of the mean of 2.15%.

Weekly pattern. A similar analysis breaks down the observed throughput by the day of the week as shown in Fig. 2. Again we observe a modest yet statistically significant pattern in performance. Benchmarks executed over the weekend show slightly higher performance compared to weekdays, with Wednesday standing out as the day with the lowest performance. The maximum variability is similar to the daily pattern with a difference of 2.52% in mean throughput from Saturdays to Wednesday.

Long-term pattern. Breaking down the results by the week of execution reveals small performance fluctuations over time. However, from our result we cannot see a long-term pattern or trend. As our experiments span only part of the year, we cannot rule out the possibility of differences during other periods.

Further observations. Besides these findings for the *us-east-1* cloud region with *m6i* instances, we made similar observations in experiments conducted in the *eu-central-1* region or with *m6g* instances.

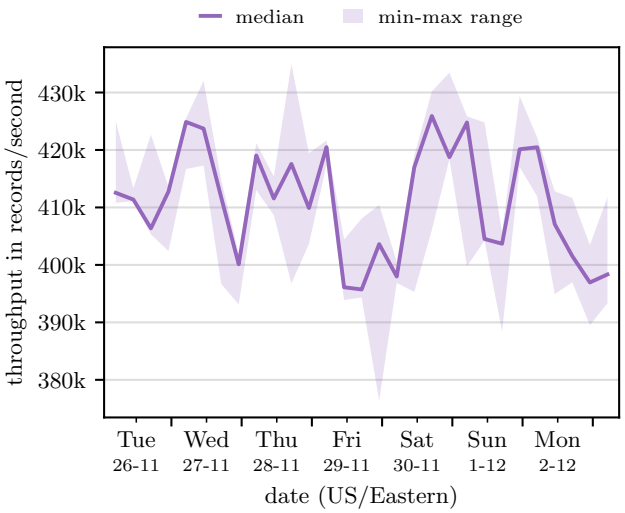

Figure 3: Measured throughput in benchmark executions around Black Friday 2024.

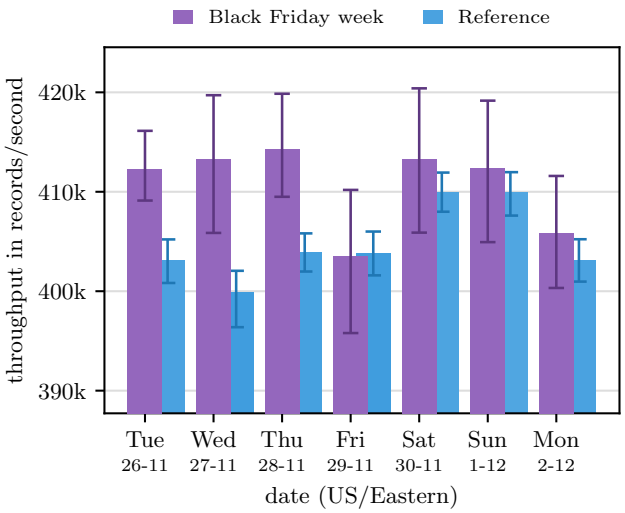

Figure 4: Mean daily throughput around Black Friday compared to the daily pattern described in Section 3.

4 Cloud Performance on Black Friday

We repeated the periodic benchmark execution around Black Friday 2024. Figure 3 shows the average throughput of each run, revealing a drop on Friday morning followed by recovery on Saturday morning. To better highlight this effect, Fig. 4 summarizes the average throughput per day and contrasts it with the baseline daily pattern identified in Section 3. The results show a clear dip in performance on the Friday compared to the three preceding and two following days, although the differences are not statistically significant due to overlapping confidence intervals.

Interestingly, the three days before Black Friday exhibit a statistically significantly throughput increase of 2.3% to 3.3% compared to the corresponding weekdays in the reference pattern. In contrast, Black Friday itself shows no measurable deviation from the typical Friday performance. For the days following Black Friday, we observe slightly higher performance, the effect is not statistically significant when compared to the reference pattern.

A possible explanation for these observations is the generally higher performance we measured in November 2024 compared to the summer months, followed by a temporary dip on Black Friday itself. Another explanation could be that the cloud provider proactively provisions additional computing resources in anticipation of Black Friday, which may elevate performance in the preceding days. In either case, our results suggest that Black Friday introduces a small but noticeable source of variability in cloud performance.

5 Conclusions

Our study confirms that application-level benchmark performance in the cloud exhibits noticeable variability. In contrast to related work, we identify clear effects of the time of day, the weekday, and global events such as Black Friday. Nevertheless, these effects are relatively small and only become relevant when targeting very small performance differences below 5%.

Acknowledgments We would like to thank JKU and Dynatrace for co-funding this research.

References

[1] P. Leitner and J. Cito. "Patterns in the Chaos—A Study of Performance Variation and Predictability in Public IaaS Clouds". *ACM Trans. Internet Technol.* 16.3 (2016). DOI: 10.1145/2885497. URL: https://doi.org/10.1145/2885497.

[2] A. Abedi and T. Brecht. "Conducting Repeatable Experiments in Highly Variable Cloud Computing Environments". *ACM/SPEC International Conference on Performance Engineering*. 2017. DOI: 10.1145/3030207.3030229.

[3] A. V. Papadopoulos et al. "Methodological Principles for Reproducible Performance Evaluation in Cloud Computing". *IEEE Transactions on Software Engineering* 47.8 (2021). DOI: 10.1109/TSE.2019.2927908.

[4] S. Henning and W. Hasselbring. "A Configurable Method for Benchmarking Scalability of Cloud-Native Applications". *Empir. Softw. Eng.* 27.6 (2022). DOI: 10.1007/s10664-022-10162-1.

[5] L. Horwitz. *Black Friday traffic exposes gaps in observability strategies*. 2022. URL: https://www.dynatrace.com/news/blog/solving-black-friday-traffic-issues/.

[6] A. Vogel et al. "A systematic mapping of performance in distributed stream processing systems". *Euromicro Conference on Software Engineering and Advanced Applications*. 2023. DOI: 10.1109/SEAA60479.2023.00052.

[7] S. Henning et al. "ShuffleBench: A Benchmark for Large-Scale Data Shuffling Operations with Distributed Stream Processing Frameworks". *ACM/SPEC International Conference on Performance Engineering*. 2024. DOI: 10.1145/3629526.3645036.

[8] S. Henning et al. "When Should I Run My Application Benchmark?: Studying Cloud Performance Variability for the Case of Stream Processing Applications". *ACM International Conference on the Foundations of Software Engineering*. 2025. DOI: 10.1145/3696630.3728563.