UK Systems Conference 2024

Dixon, Iain
i.g.dixon1@@ncl.ac.uk

Forshaw, Matthew
matthew.forshaw@ncl.ac.uk

Matthews, Joe
joe.matthews@ncl.ac.uk

February 2024

Abstract

Systems research depends on reproducible artefacts to verify experimental findings and enable follow-on research [1, 2]. While artefact availability is not mandatory, it makes verification through artefact evaluation committees possible and papers are assigned artefact badges to signal the work's commitment to replicability and reproducibility [1].

Selecting benchmark parameters that explore the full space of system inputs and behaviours is crucial as limited parameters can obfuscate undesirable results and be prohibitively expensive to evaluation teams. For example, a paper may select an arrival rate which maximizes observed throughput while avoiding complications seen at higher rates. A thorough benchmark test scenarios which a system would experience as well as extremes [3]. The hardware a system is tested under can influence results, allowing for inefficient practices that would be obvious on a scaled down system to be mitigated by increased compute power. Authors can avoid demonstrating negative results by using large amounts of compute power, limiting the potential of reproduction to groups with similar resources.

Even a well-defined artefact can appear to perform differently than how it would in practice. An unoptimised system may scale better than the same system optimised [4], or a system could fail to correctly measure benchmark accuracy metrics [5]. Authors rely on metrics collected from a benchmark to ensure validity of a test, for instance by looking at the difference between desired and observed measurement interval or load [6]. A benchmark can coordinate with the system being tested and avoid capturing metrics which would alert authors to an invalid test [7]. This "coordinated omission" would mean metrics collected from a test would appear valid and be reported on regardless of the fact that the system failed [8]. A clear example of coordinated omission is YSCB failing to capture latency spikes due to the data structure which captures latency blocking the load generator [5].

To facilitate reproducibility, papers commonly adopt closed-loop where the generator and system are attached and operate on state changes in each component [9]. This allows for easier evaluation as the entire arte-

fact and benchmark are available together, and are not impacted by network behaviours. Experimentally, we demonstrate that stream generators and stream processing pipelines (including popular Nexmark and YCSB benchmarks) are susceptible to a coordinated omission problem induced by backpressure mechanisms [10]. Backpressure occurs when a stream processing system receives data faster than it can be processed, causing operators to halt, and processing delays propagate through the pipeline to upstream operators. In the real world the entire pipeline would halt up to the ingest operator, causing new tuples entering the system to be dropped. Meanwhile, under a closed system backpressure causes a benchmark generator to also halt, causing a coordinated halt between the benchmark and system and the collected metrics to show all tuples were ingested.

This presentation will: (**a**) highlight best practices for systems benchmarking spanning ACM guidelines, JSys, and industry standard benchmarks by SPEC [6]; (**b**) demonstrate how undetected failures and coordinated omission can obfuscate benchmark results; (**c**) demonstrate experimentally that backpressure can induce a coordinated omission problem in stream benchmarking; (**d**) offer recommendations for better design and running of experiments. This presentation represents our ongoing work to solve challenges to reproducibility in systems. We hope to provide guidance to practitioners in the design of artifact evaluation checklists and performance benchmarks.

[1] Noa Zilberman and Andrew W. Moore. "Thoughts about artifact badging." In: *ACM SIGCOMM Computer Communication Review 50*. 2020.

[2] Stefan Winter et al. "A retrospective study of one decade of artifact evaluations." In: *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 2022.

[3] Stefan Bouckaert et al. "BONFIRE: benchmarking computers and computer networks". In: *EU FIRE Workshop*. 2011.

[4] Frank McSherry, Michael Isard, and Derek G. Murray. "Scalability! But at what COST?" In: *15th Workshop on Hot Topics in Operating Systems (HotOS XV)*. Kartause Ittingen, Switzerland: USENIX Association, 2015. URL: `https://www.usenix.org/conference/hotos15/workshop-program/presentation/mcsherry`.

[5] Nitasan Wakart. 2015. URL: `https://psy-lob-saw.blogspot.com/2015/03/fixing-ycsb-coordinated-omission.html`.

[6] *SPECpower*$_{s}$*sj*2008*RunandReportingRules*. `https://www.spec.org/power/docs/SPECpower_ssj2008-Run_Reporting_Rules.html#2.1`.

[7] Gil. Tene. "How NOT to Measure Latency." In: *Strange Loop Conference*. 2015.

[8] Ivan Prisyazhynyy. 2021. URL: https://www.scylladb.com/2021/04/22/on-coordinated-omission/.

[9] Bianca Schroeder, Adam Wierman, and Mor Harchol-Balter. "Open Versus Closed: A Cautionary Tale". In: *USENIX 3rd Symposium on Networked Systems Design Implementation*. 2006.

[10] Ufuk Celebi. 2015. URL: https://www.ververica.com/blog/how-flink-handles-backpressure.