

Power and Performance Analysis of Key-Value Stores on ARM and x86 Based Servers

Styliani Mikrou

Thesis submitted in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science and Engineering

University of Crete
School of Sciences and Engineering
Computer Science Department
Voutes University Campus, 700 13 Heraklion, Crete, Greece


Thesis Advisor: Prof. *Angelos Bilas*

UNIVERSITY OF CRETE
COMPUTER SCIENCE DEPARTMENT

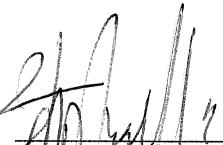
**Power and Performance Analysis of Key-Value Stores on ARM and
x86 Based Servers**

Thesis submitted by
Styliani Mikrou
in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science

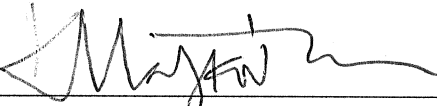
THESIS APPROVAL

Author: 

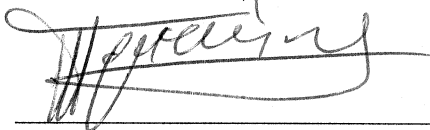
Styliani Mikrou

Committee approvals: 

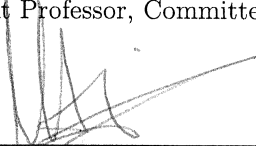
Angelos Bilas
Professor, Thesis Supervisor



Kostas Magoutis
Associate Professor, Committee Member



Polyvios Pratikakis
Assistant Professor, Committee Member

Departmental approval: 

Antonios Argyros
Professor, Director of Graduate Studies

Heraklion, November 2019

Power and Performance Analysis of Key-Value Stores on ARM and x86 Based Servers

Abstract

With the current rate of data growth, processing needs are becoming difficult to fulfill due to CPU power and energy limitations. Data serving systems and especially persistent key-value stores have become a substantial part of data processing stacks in the data center, providing access to massive amounts of data for applications and services. Key-value stores exhibit high CPU and I/O overheads because of their constant need to reorganize data on the devices.

In this master thesis, we examine the efficiency of two key-value stores on four servers of different generations and with different CPU architectures. We use RocksDB, a key-value that is deployed widely, e.g. in Facebook, and Kreon, a research key-value store that has been designed to reduce CPU overhead. We evaluate their behavior and overheads on an ARM-based microserver and three different generations of x86 servers. Our findings show that microservers have better power efficiency in the range of 0.68-3.6x with a comparable tail latency and they incur 1.1-2.7x lower energy cost. However, microservers to be more cost-effective they need to have a purchase price several times lower, and typically around or more than 3x, than higher end servers.

Ανάλυση Ισχύος και Απόδοσης Συστημάτων Αποθήκευσης Ζευγαριών Κλειδιού-Τιμής σε Διακομιστές Βασισμένους σε ARM και x86

Περίληψη

Ο συνεχώς αυξανόμενος ρυθμός παραγωγής δεδομένων, καθιστά την ολοκλήρωση της επεξεργασίας τους όλο και πιο δύσκολη λόγω των περιορισμών σε ισχύ και ενέργεια στα κέντρα δεδομένων. Τα συστήματα εξυπηρέτησης δεδομένων και ιδίως τα συστήματα μόνιμης αποθήκευσης ζευγαριών κλειδιού-τιμής αποτελούν σημαντικό μέρος της επεξεργασίας δεδομένων στα μοντέρνα κέντρα δεδομένων, παρέχοντας πρόσβαση σε δεδομένα για εφαρμογές και υπηρεσίες. Τα συστήματα αποθήκευσης ζευγαριών κλειδιού-τιμής αυξάνουν τον φόρτο του επεξεργαστή και των συσκευών αποθήκευσης λόγω της συνεχούς τους ανάγκης για αναδιοργάνωση των δεδομένων.

Σε αυτή τη μεταπτυχιακή εργασία, εξετάζουμε την αποδοτικότητα δύο συστημάτων αποθήκευσης ζευγαριών κλειδιού-τιμής σε τέσσερις διαφορετικές γενιές διακομιστών με διαφορετικές αρχιτεκτονικές επεξεργαστών. Χρησιμοποιούμε την RocksDB, ένα σύστημα αποθήκευσης ζευγαριών κλειδιού-τιμής που χρησιμοποιείται ευρέως, π.χ. στην Facebook, και το Kreon, ένα ερευνητικό σύστημα αποθήκευσης ζευγαριών κλειδιού-τιμής που έχει σχεδιαστεί για να μειώνει τον φόρτο του επεξεργαστή. Μέσα από τη μελέτη μας, αξιολογούμε τη συμπεριφορά των δύο συστημάτων και την επιβάρυνση που προκαλούν στον επεξεργαστή και στις συσκευές αποθήκευσης. Εξετάζουμε ένα μικρο-διακομιστή ARM και τρεις διαφορετικές γενιές διακομιστών x86. Τα αποτελέσματα της αξιολόγησής μας, δείχνουν ότι οι μικρο-διακομιστές έχουν 0.68-3.6x καλύτερη απόδοση ισχύος με συγκρίσιμο χρόνο καθυστέρησης και 1.1-2.7x χαμηλότερο κόστος σε ενέργεια. Ωστόσο, το συνολικό κόστος των μικρο-διακομιστών είναι καλύτερο, μόνο όταν η τιμή αγοράς τους είναι σημαντικά, π.χ 3 φορές, χαμηλότερη των τυπικών διακομιστών.

Acknowledgements

This work was carried out at the Computer Architecture and VLSI (CARV) laboratory of the Institute of Computer Science (ICS) of the Foundation of Research and Technology Hellas (FORTH), and was financially supported by a FORTH ICS scholarship. There are several people I would like to thank. First of all I would like to thank my thesis advisor and committee member, Prof. Angelos Bilas for his support and guidance throughout my graduate studies. I would also like to thank the other two members of the committee, Assistant Prof. Polyvios Pratikakis and Associate Prof. Kostas Magoutis for their time and effort they put to evaluate this work. A special thanks also goes to Anastasios Papagiannis and Giorgos Saloustros for their help and advice for this work. In addition, I would like to thank Manos Pavlidakis and Iacovos Kolokasis for reviewing this work and also for their comments both for the text and presentation. Also, I would like to thank Christi Symeonidou, Nikos Papakostantinou, Antonis Papaioanou, Klodjan Klodi Hidri, Stelios Mavridis, Nikos Batsaras, and Lena Kanellou for their comments for the presentation. Finally, I would like to thank my family, friends and colleagues for supporting me during my studies.

Contents

Table of Contents	i
List of Tables	iii
List of Figures	v
1 Introduction	1
2 Related Work	3
3 Experimental Methodology	5
3.1 Server characteristics	5
3.2 KV stores	6
3.3 Workloads	7
3.4 Power measurements	8
4 Experimental Analysis	9
4.1 Which server architecture is more power efficient?	9
4.2 Which server architecture achieves the highest absolute throughput?	11
4.3 Which micro-architectural features (do not) matter?	11
4.4 Does server performance translate to tail latency benefits?	16
5 Server cost analysis	21
6 Conclusions	25
Bibliography	27

List of Tables

3.1	Our evaluation servers and their hardware components.	5
3.2	Memory (DRAM) throughput for one thread and a number of threads equal to the number cores in the server, as measured with STREAM [17].	6
3.3	Workloads used with YCSB. All workloads use Zipf distribution except for D that uses latest distribution.	8
5.1	Values for the energy and equipment cost model.	22
5.2	Energy cost for each server type and workload.	22

List of Figures

4.1	Power efficiency (ops/joule) for Kreon (top) and RocksDB (bottom).	10
4.2	Multi-threaded absolute performance and IPC for Kreon (top) and RocksDB (bottom), under high CPU utilization.	12
4.3	Single-thread absolute performance and IPC for Kreon (top) and RocksDB (bottom).	13
4.4	Performance monitor counter measurements for branch and L3 cache miss ratio in multi-threaded experiment for Kreon (top) and RocksDB (bottom).	14
4.5	Single-thread L3 miss ratio for Kreon and RocksDB.	15
4.6	IPC for four hardware threads (running YCSB) with an increasing available memory throughput.	15
4.7	Average and tail latency (μs) for Load A and Run C with Kreon (top) and RocksDB (bottom).	17
4.8	Average and tail latency (μs) for Run C in RocksDB with different I/O block cache sizes.	18
5.1	Cost in \$/h for servers that run a KV store related to the purchase price of each server and depreciation time.	23

Chapter 1

Introduction

Projections for data growth show that data doubles roughly every two years [25] leading to high demand for more processing capacity to serve and process the data. Given current technology limitations for power and energy [13, 15], this increasing demand for CPU cycles can only be satisfied by increasing the processing density within existing power and energy budgets. One approach to achieve this, is to execute certain classes of applications in microservers rather than high-end servers. Microservers include CPUs with different, lower-power designs, such as ARM processors, compared to typical data center servers that use higher-end Intel or AMD processors.

Previous research [20, 4, 6, 23, 16] has examined the benefits of running applications on microservers and in certain cases microservers have been deployed in production setups [14], [26]. These previous works use mobile, desktop, web server, database, and other workloads to examine performance and energy trade-offs.

Persistent key-value (KV) stores are a main component of data analytics stacks and data access frameworks in general [9, 7, 2, 3, 14, 12, 10]. Typically, persistent KV stores are complex systems because they constantly re-organize data on storage devices to achieve high data rates for write, scan, and read operations. As such, each user-initiated operation in KV stores requires several thousands of CPU cycles in the common path [22]. Recently, new designs for KV stores have emerged that trade storage device efficiency for CPU efficiency, in an effort to increase data serving density [21, 14].

In this master thesis we explore the use of microservers to increase data serving density with persistent KV stores. We use two KV stores: RocksDB [10], a persistent KV store from Facebook that is widely deployed in production setups. Furthermore, we use Kreon [22], a research key-value store that reduces CPU overhead and therefore CPU cycles for each KV operation. In order to provide a thorough analysis, we use four different, server-grade systems that span a broad range of processor architectures, memory hierarchy characteristics, and fabrication process technologies. We run YCSB [8] using the default workloads that cover a

wide range of cloud use-cases.

We evaluate power efficiency, absolute performance, and architectural characteristics that affect performance and tail latency. We use one ARM microserver and three different generations of data center x86 servers. These servers cover a wide range of fabrication processes technology, micro-architectural features, and amount of processing and memory resources. For performance experiments we carefully select KV store configuration setups to make our evaluation realistic. For power measurements we use a power monitor connected right after the power supply unit (PSU). For both performance and power measurements we perform a large number of experiments and we present the most relevant data. Finally, we also provide an analysis for how different server types contribute to the total cost of ownership in data centers when used for data serving. Our results show that microservers:

1. Are 0.68-3.6x times more power efficient.
2. Result in a 1.27-5.3x lower absolute performance where a major factor to performance is memory (DRAM) throughput.
3. Do not have a big impact to tail latency.
4. Have on average 1.1-2.7x lower energy cost.
5. Are more cost-effective if they have a purchase price around 3x lower than high-end servers.

The rest of the master thesis is organized as follows. Section 2 compares our work with related research. Section 3 presents our experimental methodology and Section 4 presents our experimental analysis. Finally, Section 5 presents an analysis for energy and equipment costs and Section 6 concludes our master thesis.

Chapter 2

Related Work

Previous work has compared microservers and high-end servers in terms of performance and energy consumption. The authors in [11] show that current high-end Out-Of-Order processor micro-architectures are inefficient for running scale-out (cloud) workloads. They use performance counters to identify key micro-architectural needs for these workloads and sources of inefficiency. Also, authors in [6] revisit the RISC vs. CISC architecture using mobile, desktop, and server workloads. They find that RISC and CISC ISAs are irrelevant to power and performance characteristics of modern cores, whereas micro-architectural features have an important impact on them. In our work we use performance monitor counters with the difference that we want to identify micro-architectural features that affect the performance and power of KV stores. Moreover, we examine how micro-architectural features affect tail latency of KV stores.

Also, authors in [4] and [20] compare x86 and ARM architectures in terms of power efficiency for web server, database, and other workloads. They conclude that x86-based servers are more efficient for compute-intensive workloads and ARM-based servers are advantageous in computationally lightweight applications in terms of power efficiency. One major difference from our work is that we use KV stores and we evaluate them in an ARM server (not mobile) and x86 real deployed servers. Another difference is that we show how micro-architectural features affects performance.

In [1], the authors show that low-power embedded nodes with flash storage can deliver over an order of magnitude more queries per joule for random read-intensive workloads using a custom KV store on a custom cluster. Also in [14] they use a custom KV store that runs to a customized compact server design based on ARM processors and they show that is reliable, highly scalable and cost-effective. In contrast, in our work we use two persistent key-value stores, one research (Kreon [22]) and one widely deployed in production products (RocksDB [10]) and we run them on a range of commodity servers.

Authors in [16] show that x86 servers are more energy efficient in I/O-intensive workloads and ARM servers are more energy efficient for database query processing

with slightly lower performance. They present a total cost of ownership (TCO) analysis and they find out that an ARM-based cluster has lower TCO except for I/O intensive workloads, where it incurs 50% higher TCO compared to an x86-based cluster. Furthermore, authors in [20] use a monthly cost model to analyze cost-efficiency of ARM and x86-based data centers. They conclude that from the perspective of cost-efficiency ARM-based data centers are advantageous for computationally lightweight applications. In our work, we calculate energy cost (dollars per hour) to show the most cost efficient server type for KV stores. Furthermore, we include purchase price and depreciation time of server and we use a wide range of prices, 3, and 5 years depreciation time in order to see how these two affect total cost.

Chapter 3

Experimental Methodology

In this section we describe our experimental setups and how we perform our measurements.

Table 3.1: Our evaluation servers and their hardware components.

	CPU (all 64-bit)	# chips	Fabrication technology	ISA	# cores	# Threads	Clock GHz	L1 KB/core	L2 KB/2core	L3 MB/chip	DRAM x DIMM GB
S1	X-Gene 1 ARMv8	1	40 nm	ARM	8	8	2.40	32	256	8	16x1= 16 DDR3
S2	Xeon(R) E5520	2	44 nm	x86	8	16	2.27	32	256	8	2x6= 12 DDR3
S3	Xeon(R) E5620	2	32 nm	x86	8	16	2.40	32	256	12	2x12= 24 DDR3
S4	Xeon(R) E5-2630 v3	2	22 nm	x86	16	32	2.40	32	256	20	32x8= 256 DDR4

3.1 Server characteristics

In our work we use four different types of servers that span a broad range of processor architectures, memory hierarchy characteristics, and fabrication process technologies. Table 3.1 summarizes the characteristics of each server. Server S1 is an ARM-based server, whereas servers S2, S3, and S4 are x86 servers of different generations. All servers have similar clocks (2.27-2.4 GHz), similar size of L1 cache per core (32KB), and similar L2 cache size (256KB), shared in all cases by a pair of cores. L3 cache is shared by the whole chip and the normalized per-core size is similar. Servers S2, S3, and S4 have two NUMA nodes while S1 has a single NUMA node. The number of DIMMs and the total DRAM size differ in all cases, as shown in Table 3.1. Finally, servers S1 and S2 use similar fabrication process technology (40-44 nm), S3 uses a 32 nm process, and S4 uses a 22 nm process.

Table 3.2: Memory (DRAM) throughput for one thread and a number of threads equal to the number cores in the server, as measured with STREAM [17].

	#threads=1 GB/s	#threads=#cores GB/s	#threads=64 GB/s
S1	7.7	8.6	8.5
S2	9.0	21.5	21.2
S3	9.4	23.9	23.5
S4	14.2	68.9	60.6

We perform various point-to-point comparisons among these servers to identify characteristics that have an impact on system performance and power efficiency. To minimize the impact of the software stack, all servers run the same Linux kernel (version 4.4) with the same version of the GCC compiler (version 4.8) toolchain. Finally, all servers are equipped with the same type of NVMe storage device, a Samsung 950 PRO 256GB.

For calibration purposes we measure the memory throughput in each server using the STREAM [17] memory benchmark with an increasing thread count. Table 3.2 shows these results for the Triad scenario.

3.2 KV stores

RocksDB [10] is an LSM-based persistent key-value store that is widely used in production at Facebook. It is optimized for fast storage but it can be also used for hard disk drives. It contains multiple levels of increasing size where keys are sorted within each level. Each level is divided in multiple units of fixed size named SSTable (SST) and each of them is stored in a separate file.

In RocksDB, the first level is stored in memory (named memtable) and when it becomes full it is flushed in first levels SSTs. In order to provide write amortization, it keeps the ratio between 2 successive levels to be less than 10x. In the case where the ratio exceeds this value, a compaction is triggered. A compaction merges SSTs of level L_i with SSTs of the next level (L_{i+1}). RocksDB interacts with storage when it writes a memtable to the device, creating a new SST and during compactions. In both cases it uses read-write API to access files.

Kreon [22] is a persistent write-optimized key-value store designed for flash storage. The main design tradeoff is that it increases I/O randomness in order to reduce CPU overhead and I/O amplification. To achieve this, Kreon uses a write-optimized data structure and *memory-mapped I/O*.

Kreon uses a multi-level indexing data structure similar to the LSM-Tree [19], with levels of increasing size. This enables batched data transfers to lower levels to amortize insert costs. Kreon uses a per-level full index (B-tree) to enable partial data reorganization which reduces I/O amplification and CPU overhead. Furthermore, Kreon stores key-value pairs in an append-only log to avoid data movement

during spill (merge) operations.

Kreon uses *memory-mapped I/O* to further reduce CPU overheads related to the I/O cache in three ways: (a) It eliminates cache lookups for hits by using valid virtual memory mappings. Accesses to data that are not in memory result in page faults that are then handled by Linux kernel. (b) *Read/write* system calls require a data copy between user and kernel space for protection purposes. *Memory-mapped I/O* removes the need for data copies when performing I/O. (c) *Memory-mapped I/O* uses a single address space for both memory and storage, which eliminates the need for pointer translation between them. Therefore, this approach removes the need for serialization and deserialization when transferring data between memory and storage.

3.3 Workloads

We run a C++ version of YCSB [24, 8] using the proposed workloads in the recommended sequence: Load the database, run workloads in order A, B, C, F, and D, clear the database, load the database again and run workload E. In all cases the key is about 30 bytes and the value is 1000 bytes. We use datasets that fit in main memory for all servers, since we are mainly interested in CPU efficiency.

For our analysis, we run two different experiments for both key-value stores. For Kreon the first experiment uses a single YCSB thread and a single Kreon table with a dataset of 3M records that fits in memory and does not cause I/O traffic (no snapshots) for all servers. We run this experiment to identify the performance and energy characteristics of the core’s architecture. In the second experiment for Kreon we run YCSB with multiple threads to identify server performance and power consumption under high utilization. In this case, the duration of each run for 3M records is short, which does not allow us to observe server performance under steady state. For this reason, and since servers have different main memory sizes, we use for each server a dataset proportional to its memory. We use 350K records per 1GB of DRAM i.e., 2.4M for S1, 4.2M for S2, 8.4M for S3, and 89.6M for S4. Finally, we use 64 YCSB threads and 32 Kreon tables in all cases to keep all servers at high CPU utilization.

For RocksDB the first experiment uses a single YCSB thread and one database with dataset of 6M records. We enable direct I/O with 2GB block cache. In this experiment we identify the performance and energy characteristics of the core’s architecture with I/O traffic. In the second experiment for RocksDB we run YCSB with multiple threads again with direct I/O. To avoid thread contention and keep utilization high we choose to keep the number of YCSB threads equal to the number of hardware threads for all servers (S1 does not support hardware threads and therefore, this is the number of cores). We use a number of databases equal to half the number of YCSB threads.

Since our interest is CPU behavior, in our experiments we try to maintain similar I/O behavior across servers. For this purpose, we configure the datasets in a

Table 3.3: Workloads used with YCSB. All workloads use Zipf distribution except for D that uses latest distribution.

Workload	
A	50% reads, 50% updates
B	95% reads, 5% updates
C	100% reads
D	95% reads, 5% inserts
E	95% scans, 5% inserts
F	50% reads, 50% read-modify-write

manner where the LSM multi-level structure exhibits the same I/O behavior across servers: We use 4 databases and 5M records for S1, 8 databases and 10M records for S2/S3, and 16 databases and 20M records for S4. With these parameters, the number of flushes and compactions for each workload are the same for each server. Finally, we adjust the size of block cache, to fit the same amount of dataset for each server i.e, 2GB for S1, 4GB for S2/S3 and 8GB for S4.

3.4 Power measurements

We measure power for all platforms with a Microchip MCP39F511A Power Monitor Demonstration Board [18]. The Power Monitor is connected right after the power supply unit (PSU), which converts the AC wall socket supply to DC. It reports mean power readings every second. This measures the total power of the board, including CPU, memory, and storage devices.

We also use the Linux kernel perf tool to collect and analyze performance monitor counters (PMC) that are available in all servers. We calculate branch miss ratio, instructions per cycle (IPC), and Last-Level-Cache (LLC) (i.e. L3) miss ratio. We perform our calculation based on reading from the following PMCs: branches, branch misses, instructions, cycles, L3 cache references and L3 cache misses. The ARM-based server S1 has limited set of performance counters and it does not provide counters for the LLC (L3).

Chapter 4

Experimental Analysis

In this section we focus our analysis around four main questions:

- Which server architecture is more power efficient?
- Which server architecture achieves the highest absolute throughput?
- Which micro-architectural features (do not) matter?
- Does server performance translate to tail latency benefits?

Next, we discuss each of these in detail.

4.1 Which server architecture is more power efficient?

We measure power efficiency as ops/joule in all servers for both KV stores (Figure 4.1). Figures 4.1a and 4.1c show the efficiency of each server for Kreon and RocksDB under high utilization, above 80% in all runs shown. First, for both KV stores, we can categorize servers in two groups: S2, S3 and S1, S4. We see that S1 and S4 are more power efficient than S2 and S3 for both KV stores. Compared to servers S2 and S3, S1 executes 1.6-3.6x more ops/joule for Kreon and 1.8-3x more ops/joule for RocksDB. Similarly, compared to S2 and S3, S4 achieves 2.1-2.7x more ops/joule for Kreon and 1.4-2.5x more ops/joule for RocksDB. We note that servers S2, S3 have the same architecture as S4, but differ significantly in fabrication process technology (Table 3.1).

Between S1 and S4 there is greater variance. S4 has both a more aggressive architecture and more recent fabrication process (22nm for S4 vs. 40nm for S1). However, S1 achieves between 0.68-1.47x more ops/joule for Kreon and between 0.96-1.87x for RocksDB. We notice that S1 is better at writes (Load A, Load E) for both KV stores except Load E at Kreon that is slightly worse. Finally, S4 achieves always slightly more operations per joule in scans (Run E).

Figures 4.1b and 4.1d depict ops/joule for the single-threaded experiment (one YCSB thread and one database) for both KV stores. This experiment shows

the behavior of a single thread with abundant resources, including shared micro-architectural resources, memory throughput, caches, I/O. We notice that a single thread in server S1 is more power efficient compared to a single thread in server S4 by 1.46-1.86x for Kreon and by 1.03-1.74x for RocksDB.

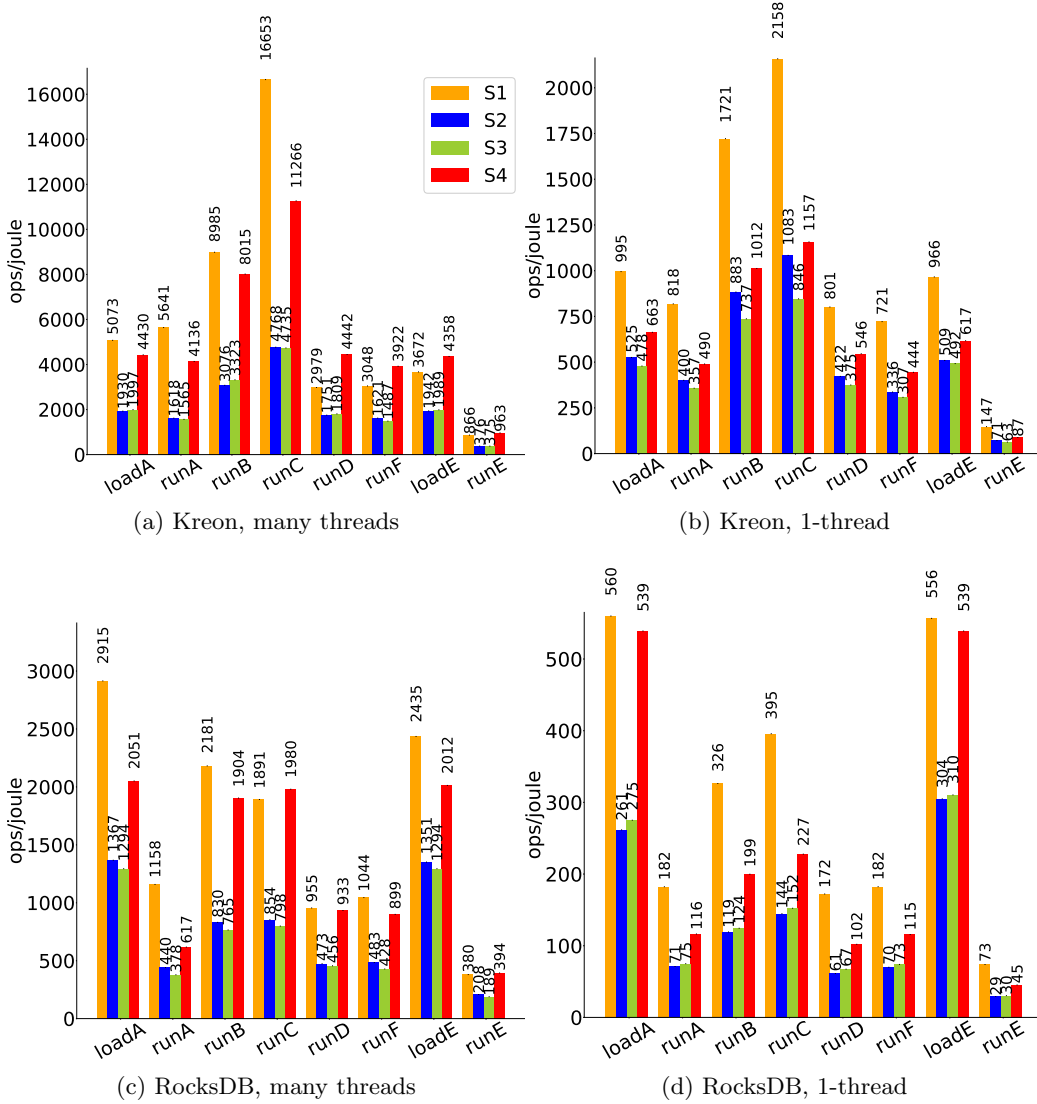


Figure 4.1: Power efficiency (ops/joule) for Kreon (top) and RocksDB (bottom).

Compared to S2 and S3, S1 exhibits single-thread efficiency (ops/joule) between 1.89-2.55x for Kreon and 1.79-2.81x for RocksDB. Compared to S2 and S3, S4 achieves a single-thread efficiency (ops/joule) between 1.06-1.45x for Kreon and 1.49-2.06x for RocksDB. Therefore, despite its older fabrication process, S1 is more power efficient compared to S2 and S3, than S4.

Finally, between servers S2 and S3 we do not observe any significant differences in ops/joule in both the high-utilization and single-thread experiments. Their CPU architectures are similar, they both have 2 NUMA nodes and they have about the same fabrication process.

Overall, server S1 is better by 0.68-3.6x in terms of ops/joule in all our experiments for both KV stores, despite the fewer resources and older fabrication technology.

4.2 Which server architecture achieves the highest absolute throughput?

Figures 4.2a and 4.2c show absolute performance expressed in kops/s. We see that servers of different generations exhibit significant differences in performance for KV stores up to 5.3x. In Kreon, S1 exhibits up to 5.3x worse performance (kops/s) compared to S4 and between 1.34-2.0x worse performance compared to S2 and S3. Servers S2 and S3 have approximately the same absolute performance and 2.0-2.7x worse compared to S4. In RocksDB, S1 exhibits 1.75-3.23x fewer kops/s compared to S4 and 1.27-2.2x lower performance compared to S2 and S3. S4 achieves 1.24-2.2x higher performance compared to S2 and S3.

Next, we examine the achieved IPC per core (not per hardware thread). Figure 4.2b shows that IPC follows the same trend as absolute performance across all servers. S4 achieves the highest IPC among all servers in the range of 1.46-2.38, whereas S1 achieves the lowest IPC in the range of 0.64-1.03. If we multiply IPC with the number of cores in each server, we approximately get the same trend as absolute performance.

Figure 4.3a captures single thread performance for each server type. It measures absolute performance (ops/s) with a single YCSB thread assigned to one core. We see that a single thread in S1 performs 1.34-1.76x worse than S2, S3. Compared to S4, S1 has 1.9-2.1x lower throughput. Finally, S4 compared to S2, S3, achieves 1.3-1.4x higher throughput. This experiment, with all resources available to one thread, shows (Figure 4.3b) that S1 achieves again the lowest IPC, between 0.73-1.07. Server S4 achieves the highest IPC between 1.72-2.09. We get the same trend across servers, as single thread performance. In comparison in multi-threaded experiments, the differences in absolute performance and IPC between servers decreases, compared to single-thread experiment.

4.3 Which micro-architectural features (do not) matter?

Next, we examine CPU performance counters for several events to identify sources of performance differences. We study branch and L3 cache miss ratios, and the impact of hardware multi-threading, as shown in Figure 4.4. We perform these

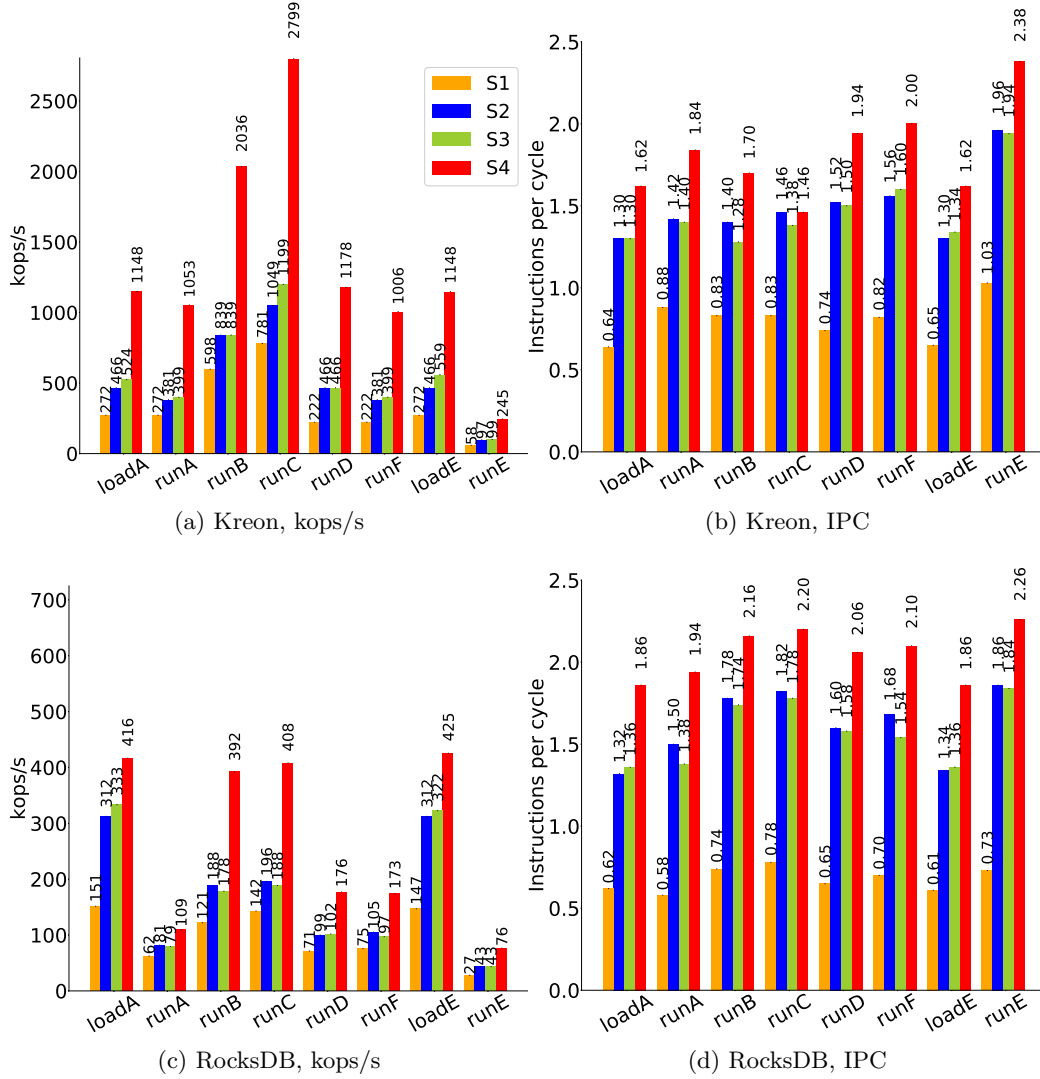


Figure 4.2: Multi-threaded absolute performance and IPC for Kreon (top) and RocksDB (bottom), under high CPU utilization.

measurements for both multi-threaded and single-threaded experiments for each server. Results are averages across all cores.

Branch misses Figures 4.4a and 4.4c show that the branch miss ratio does not exceed 3.18% for all servers and workloads for both KV stores. We observe that S4 has significantly lower branch miss ratio compared to the other servers and in most cases it incurs less than 50% of the misses. However, given the overall low branch miss ratio, this does not contribute significantly to the observed performance differences.

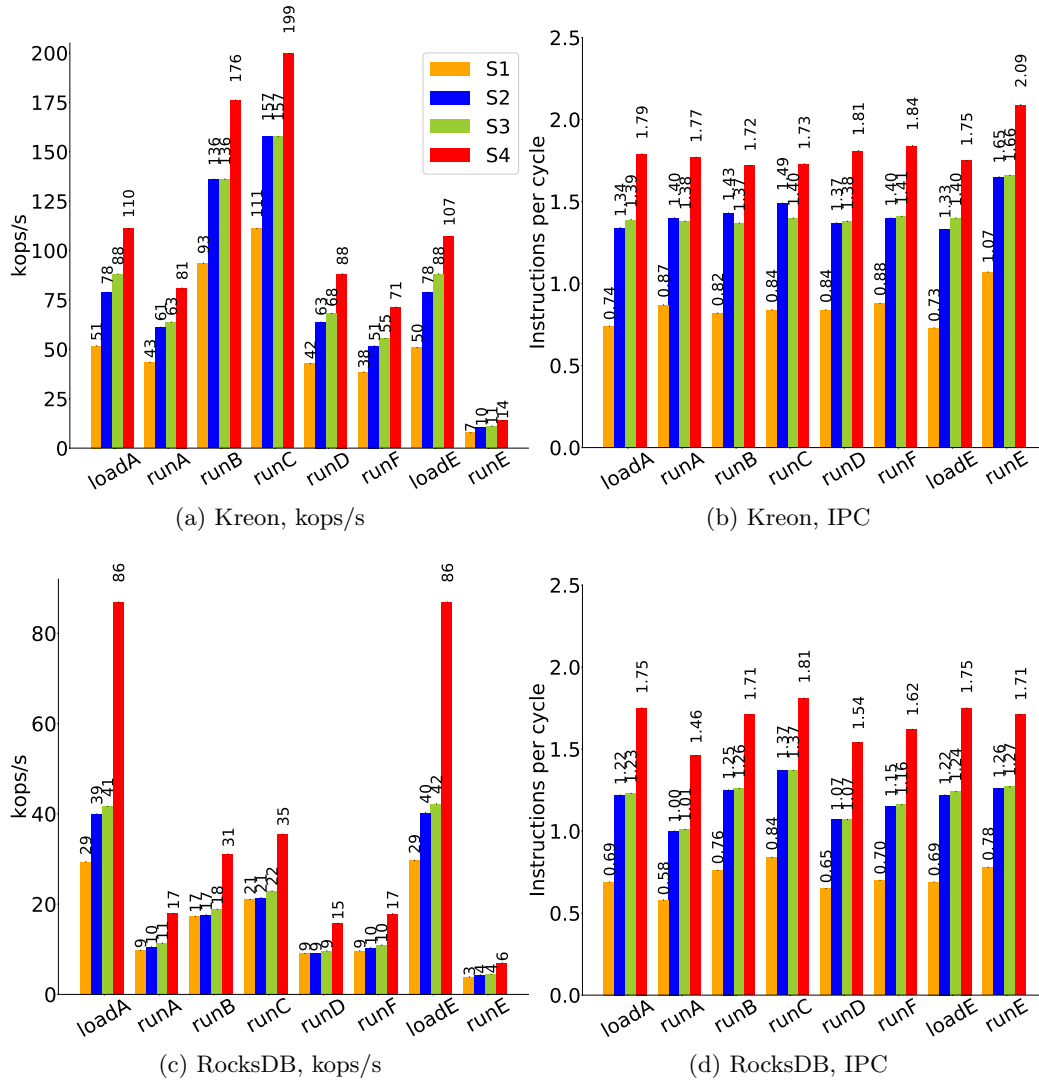


Figure 4.3: Single-thread absolute performance and IPC for Kreon (top) and RocksDB (bottom).

L3 misses As a note, L3 miss ratio is not available for S1 because of counter limitations on the specific platform. Figures 4.4b and 4.4d show that the L3 miss ratio differs between 1-6% of L3 references across servers for both Kreon and RocksDB. Although total L3 cache sizes differ across servers, the amount of L3 cache per core is about the same: S1 and S2 have 1 MB/core, whereas S3 and S4 have 1.5 and 1.25 MB/core, respectively, resulting in similar L3 miss ratios.

To examine how larger L3 caches and other components contribute to performance, we run the single-threaded experiment, with a single YCSB thread assigned to one core. Figure 4.5a shows that the L3 cache miss ratio differs up to 6% of

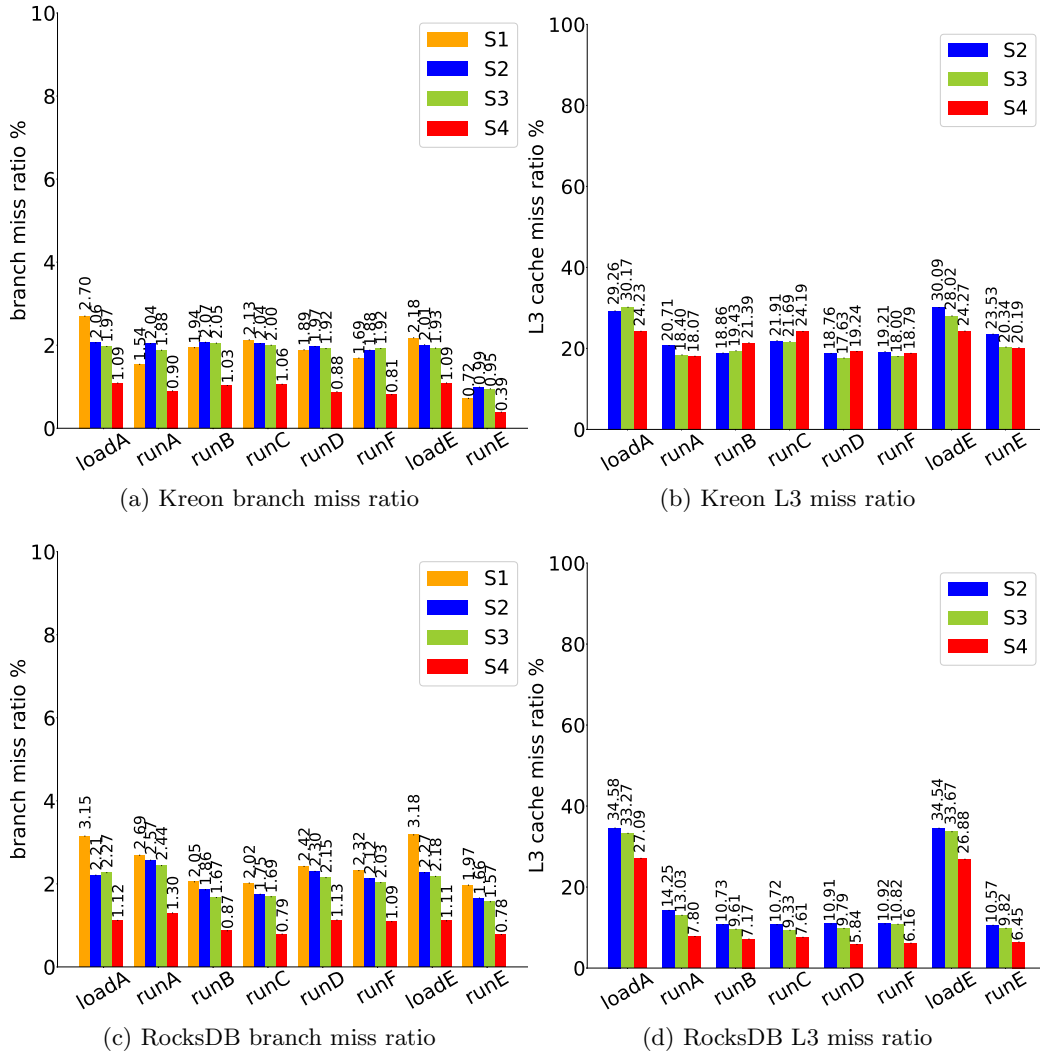


Figure 4.4: Performance monitor counter measurements for branch and L3 cache miss ratio in multi-threaded experiment for Kreon (top) and RocksDB (bottom).

all L3 cache references among all servers for both KV stores. This difference is similar to multi-threaded runs (Figures 4.4b and 4.4d), although in this case the per core L3 size differs significantly across servers. Therefore, L3 cache size does not contribute significantly to performance.

Hardware multi-threading S1 supports a single hardware thread per-core while S2, S3, and S4 have hyper-threading and thus they provide two hardware threads per-core. We perform the experiment of Figure 4.2a with hyper-threading disabled. We find that servers S2 and S3 perform 1.22-1.31x fewer kops/s, compared to the same experiment with hyper-threading enabled, whereas S4 performs

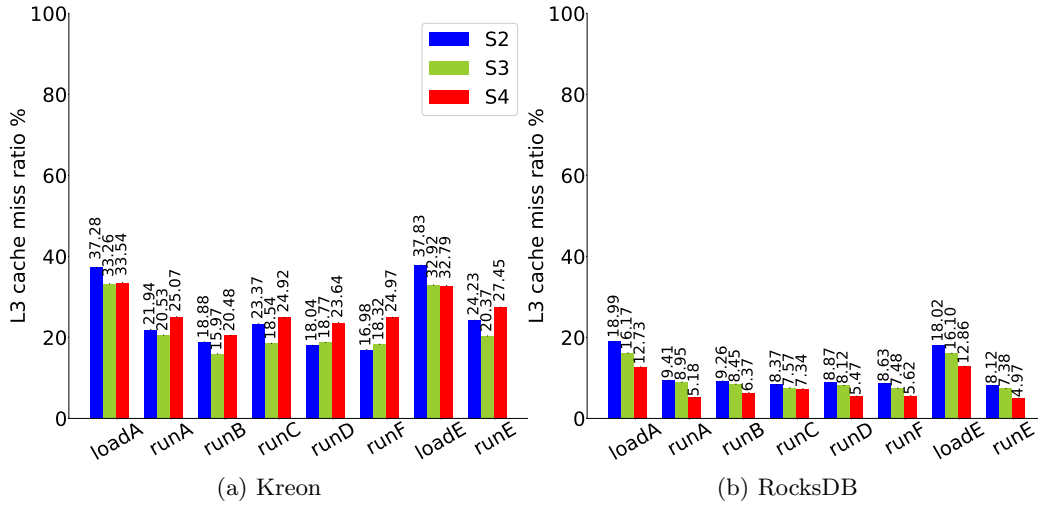


Figure 4.5: Single-thread L3 miss ratio for Kreon and RocksDB.

1.15-1.22x fewer kops/s. Similarly, IPC with hyper-threading disabled, has a drop of 1.22-1.34x and 1.16-1.29x for servers S2/S3 and S4 respectively. This shows that using twice the number of hardware threads (hyper-threading) only increases performance between 1.15-1.34x across all cases.

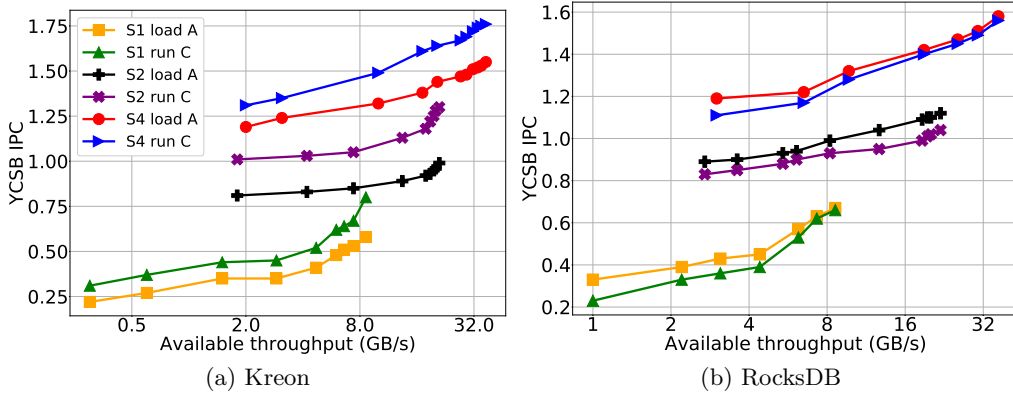


Figure 4.6: IPC for four hardware threads (running YCSB) with an increasing available memory throughput.

Memory throughput DRAM throughput affects IPC. Table 3.2 shows that S1 has 2.5x lower maximum memory throughput than S2, S3 and 7x lower maximum memory throughput than S4. The differences are smaller for the maximum memory throughput observed by a single thread.

To examine how memory throughput affects KV store performance, we create one microbenchmark which can consume a specific amount of memory throughput.

We run, for both KV stores, four YCSB threads concurrently with four threads of our microbenchmark, each of them pinned to one physical core. We choose four threads because this is the minimum number of threads that can consume the total memory throughput. In every run we decrease the throughput consumed by our microbenchmark to increase the available throughput for Kreon and RocksDB. Figure 4.6 shows the resulting IPC for four YCSB threads running Load A and Run C. We observe that in all cases we achieve better IPC when available memory throughput increases. Overall, all systems are underprovisioned and they can benefit from more memory throughput than the currently provisioned 2.1, 5.4, 9.1 GB/s/core for S1, S2, and S4 respectively.

4.4 Does server performance translate to tail latency benefits?

In this section we examine the impact of server type on tail latency. To capture how tail latency deteriorates as load increases, we increase the number of application threads per hardware threads on each server from 1-to-1 up to 8-to-1. We also examine lower and higher loads, but we find this range to be representative. Figure 4.7 shows the average and tail latency for S1 and S4 and two workloads, Load A and Run C, for both Kreon and RocksDB. We use only S1 and S4 because these two server types exhibit the largest difference in performance. For Kreon we focus on in-memory KV store performance, where the server and CPU type has the highest impact. For this reason, we use 3M keys in Load A for both S1 and S4, where they fit in memory in both servers. Then, for Run C we run a larger number of get operations to extend the execution time to 5 minutes on each server and obtain reliable 99.9% tail latency measurements.

In Figures 4.7a and 4.7b, we see that average response time differs as follows: S4 has slightly lower average latency compared to S1 for Load A. More specifically, S4 average response time is $17 \mu\text{s}$, compared to $23 \mu\text{s}$. In Run C however, S1 and S4 have almost the same average latency $8 \mu\text{s}$ for S1 vs. $9 \mu\text{s}$ for S4. We notice that as load increases, tail latency deteriorates significantly for both workloads and both servers, to hundreds of times compared to average latency at high load. In Load A, tail latency becomes up to 107x worse (S4, 8-to-1 99.9%), whereas in Run C, tail latency becomes up to 626x worse (S1, 4-to-1, 99.9%) compared to average latency in the same run. We also observe that generally, tail latency deteriorates in a similar manner on both servers, without one of the two servers exhibiting worse behavior compared to the other.

For RocksDB we run two different experiments. In the first experiment, we examine the performance of RocksDB with I/O traffic. For both Load A and Run C, we use the same dataset size as in the multi-threaded experiment (5M keys for S1 and 20M keys for S4). As we use a 4x larger dataset in S4 we also use a 4x larger user-space block cache (2GB for S1 and 8GB for S4) in combination with direct I/O to bypass the Linux kernel buffer cache. For Run C, we use a number

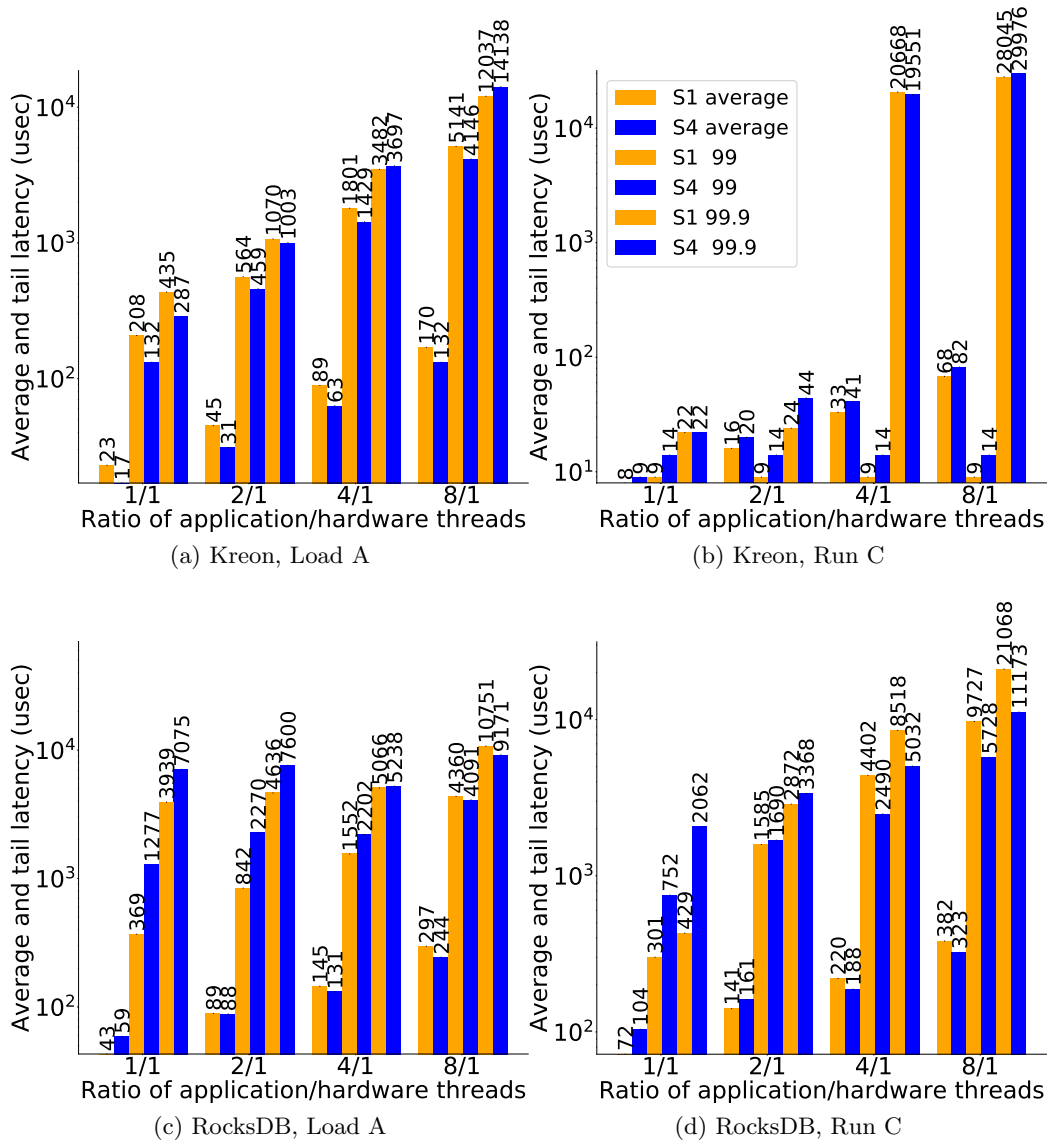


Figure 4.7: Average and tail latency (μs) for Load A and Run C with Kreon (top) and RocksDB (bottom).

of operations (*gets*) required for 5 minute run for all cases. Figure 4.7c shows for Load A that in all cases (except 1-to-1) S4 has slightly better average latency compared to S1. In the case of 1-to-1, S4 has an average latency of $59 \mu\text{s}$ compared to $43 \mu\text{s}$ for S1. Figure 4.7d shows for Run C that when we have low load (1-to-1 and 2-to-1) S1 has better latency time compared to S4. In case of high load (4-to-1 and 8-to-1) S4 becomes better compared to S1. Finally, we notice that as the load is increase further, the tail latency also increases for both S1 and S4 for both

workloads (Load A and Run C). More specifically, in Load A, tail latency becomes up to 120x worse (S4, 1-to-1, 99.9%), whereas in Run C, tail latency becomes up to 55x (S1, 8-to-1, 99.9%) compared to average latency in the same run.

In the second experiment we examine how the size of the block cache affects the average and tail latency. We provide results only for Run C as in loads RocksDB bypasses the block cache. We use the same setup as in the first experiment and we only keep the case with high load (8-to-1). We keep the ratio of the dataset to the block cache size to be the same for both S1 and S4 for each case. We start with the block cache disabled and we increase the size as follows: 512MB/2GB, 1GB/4GB and 2GB/8GB for S1/S4 respectively.

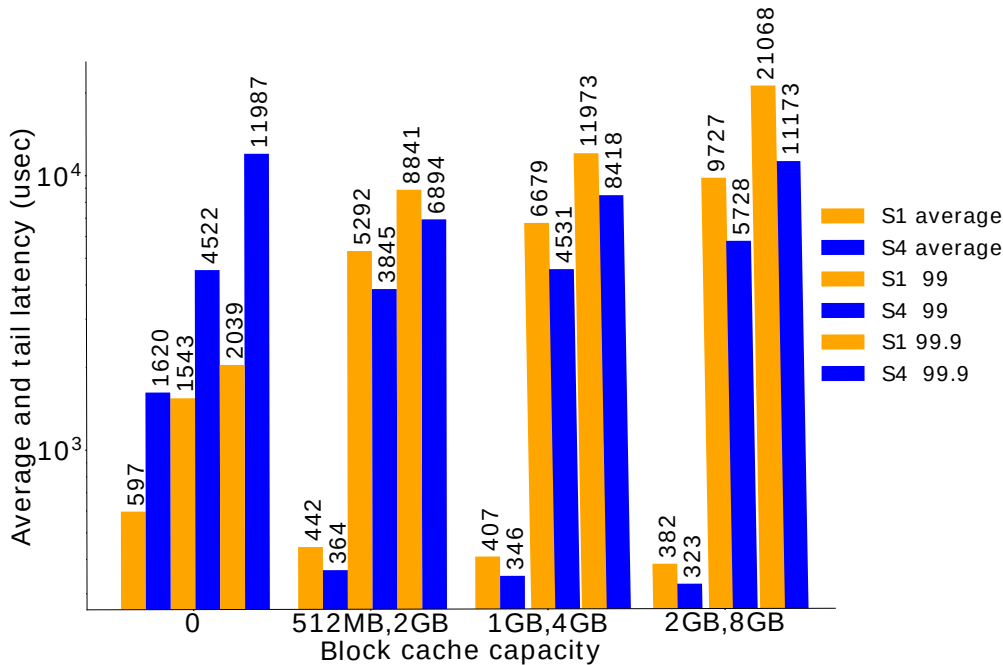


Figure 4.8: Average and tail latency (μs) for Run C in RocksDB with different I/O block cache sizes.

Figure 4.8 shows the results of this experiment. When there is no block cache, S1 is much better and up to 5.9x compared to S4 for both average and tail latencies. In this case we observe that the average disk queue depth is 254 for S4 and 10 for S1. As S4 has faster CPU compared to S1, it also processes requests and sends them to the device at a higher rate. This results in more pressure to the device and higher average and tail latencies.

Finally, as we increase the size of the block cache, both S1 and S4 achieve better average and worse tail latencies. Increasing the size of the block cache, also means that a larger part of the dataset fits in the cache and this results in lower average latency (higher hit ratio). On the other hand a larger cache also results

in slower hits, misses, and evictions, which increases the tail latency. The average latencies of S1 are slightly worse compared to S4 and the biggest difference of 1.21x is when the block cache ratio is 512MB/2GB (S1/S4). In this case S1 has average latency of 442 μ s compared to 364 μ s of S4. By increasing the block cache size, S1 always has higher tail latency compared to S4 up to 1.88x for a block cache ratio of 2GB/8GB for 99.9%.

Chapter 5

Server cost analysis

In this section we examine the tradeoffs across servers to purchase and energy costs. For this purpose, we calculate the cost of a cluster which provides a cumulative throughput of 100M ops/s. We use a cost model that calculates the total cost (C_t) based on energy and equipment (server) cost [16] as:

$$C_t = C_s + C_p \quad (5.1)$$

where C_s is the purchase price of the server and C_p is the energy cost for the entire server lifetime.

C_p is further defined as:

$$C_p = T_s \cdot C_{ph} \cdot (U \cdot P_a + (1 - U) \cdot P_i) \quad (5.2)$$

T_s is the server lifetime in years. We use two representative values for the typical server lifetime, 3 years as the low end and 5 years as the high end [5, 27]. C_{ph} is the electricity cost in \$/kWh. We choose a higher value of 0.25 \$/kWh (price in Australia) and a lower value of 0.07 \$/kWh (price in Russia) as in [16]. U and P_a represent the average utilization and power for each workload from our measurements. Finally, P_i is the idle power of each server. Table 5.1 summarizes the values we use in the cost model, based on our experimental setup and measurements.

We also present a variant of this model, C_e , where we remove the purchase price and essentially equate total cost with energy consumption. The purchase cost of each server can vary significantly for reasons such as market volume and is not easy to identify a representative value. For this reason C_e refers only to energy cost per hour to achieve the required throughput, given our performance and power measurements:

$$C_e = C_{ph} \cdot (U \cdot P_a + (1 - U) \cdot P_i) \quad (5.3)$$

Table 5.2 shows the results for C_e . The first block of columns is the required number of servers to achieve the target throughput for each configuration (server

Table 5.1: Values for the energy and equipment cost model.

Variable	Value	Description
C_s	\$	the price of a server
T_s	3-5 years	server lifetime
U	%	server utilization
C_p	\$	electricity total costs
C_{ph}	\$/kWh	electricity cost per hour
P_a	W	average server power
$P_{i,S1}$	41 W	S1 server idle power
$P_{i,S2}$	133 W	S2 server idle power
$P_{i,S3}$	163 W	S3 server idle power
$P_{i,S4}$	104 W	S4 server idle power

type, workload type). Low and high cost is the energy cost per hour (\$/h) that each cluster requires at the low and high prices we consider for electricity. Normalized costs are all the costs normalized to S1. We notice that normalized costs for low and high costs are the same because C_{ph} is the same for all servers and consequently it is cancelled in ratios. We calculate energy cost for two workloads (Load A and Run C) for each KV store (RocksDB and Kreon) and we include one summary line for the average across all workloads and KV stores, assuming all are running concurrently in each cluster. We use this average number to calculate the required number of servers in each case. Furthermore, for U and P_a we use the average values that each server achieves for both KV stores on Load A and Run C in our measurements.

Table 5.2: Energy cost for each server type and workload.

Workload	# Servers				Low cost (\$/h)				High cost (\$/h)				Normalized cost to S1			
	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Kreon Load A	368	215	191	87	1.20	3.20	3.44	1.44	4.30	11.42	12.29	5.14	1	2.66	2.86	1.20
Kreon Run C	128	95	83	36	0.45	1.51	1.48	0.60	1.60	5.39	5.27	2.15	1	3.37	3.30	1.34
RocksDB Load A	662	321	311	240	2.12	4.50	4.84	2.55	7.59	16.05	17.29	9.10	1	2.11	2.28	1.20
RocksDB Run C	704	510	532	245	3.23	6.82	7.57	3.33	11.55	24.36	27.04	11.89	1	2.11	2.34	1.03
Average	297	198	179	84	1.08	2.88	2.92	1.19	3.86	10.27	10.41	4.26	1	2.66	2.70	1.10

Table 5.2 shows that S1 always requires the largest number of servers (for both KV stores and both workloads). More interestingly, for both the low and high electricity price, S1 incurs the lowest cost per hour compared to all other server types to achieve the required throughput of 100M ops/s. The normalized results show that S2 and S3 that have similar fabrication technology incur between 2.11-3.37x higher cost, whereas S4 with more recent fabrication technology incurs between 1.03-1.34x higher energy cost.

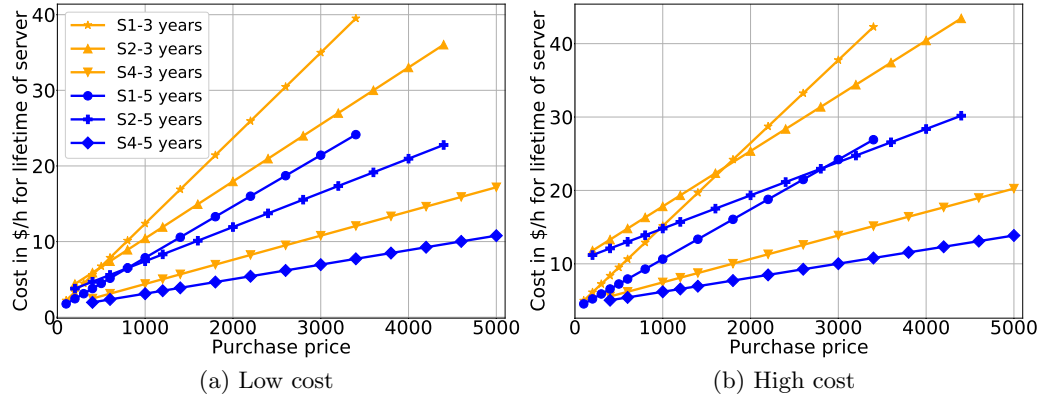


Figure 5.1: Cost in $\$/h$ for servers that run a KV store related to the purchase price of each server and depreciation time.

To take into account how the server equipment affects the total cost, we include the server purchase price. This requires using a depreciation period and calculating energy cost for the full server lifetime (depreciation period). Given that purchase price can vary significantly, Figure 5.1 plots in the y axis the cost per hour at different purchase prices in the x axis. For each server type we calculate two curves, for a 3-year and a 5-year lifetime, respectively. Furthermore, we plot costs for both low (Figures 5.1a) and high (Figures 5.1b) electricity price. As expected, in all cases, increasing the depreciation period from 3 to 5 years results in lower per-hour total cost.

More interestingly, Figure 5.1a shows that S1 has a higher sensitivity to purchase price and total cost increases faster (larger slope) because of the increased number of servers required to achieve 100M ops/s. Despite lower energy costs, for S1 to be more cost-effective compared to S4, S1 needs to have several times lower purchase price, e.g. \$3K for S4 and less than \$1K for S1 for more than a 3x difference. The same observation holds for a 5-year depreciation period. Furthermore, we notice that for 3 years depreciation time the distance (horizontal) between any two server type is bigger than the distance for 5-year lifetime. This happens because as depreciation time increases, the total energy cost increases, whereas the purchase price remains the same. Consequently, energy costs affect more the total cost per hour. Effectively, the shorter the lifetime of servers the larger the difference in purchase price of which microservers become more cost-effective.

Figure 5.1b shows that at higher electricity cost S1 and S4 are generally significantly more attractive than S2. Similar to the low electricity cost, for S1 to be more attractive compared to S4, S1 needs to maintain several times, and typically more than 3x, lower purchase cost compared to S4.

In summary, if we consider only energy costs (Table 5.2), S1 is always more cost-effective, when trying to achieve a performance mark of 100 Mops/s. However, if we take into account purchase price and depreciation period, then S1 needs to

have several times, and typically at least 3x, lower purchase cost compared to S4 to be more cost-effective.

Chapter 6

Conclusions

Persistent KV stores are an important component for modern software stacks in the data center. In this work, we examine how the processor micro-architecture and memory hierarchy affect data serving systems. We use four server types and two different KV stores (Kreon [22] and RocksDB) to measure power efficiency and absolute performance.

A microserver (S1) results in 1.6-3.6x better power efficiency compared to an x86 server with the same fabrication technology (S2). S1 is up to 1.87x more power efficient compared to S4, a more powerful server of newer process technology (22nm vs. 40nm). Although all processors have similar CPU clocks, servers with more cores result in higher performance. S4, with 2x more physical cores from S1 and hyper-threading enabled, achieves up to 5.3x more operations per second than S1. All of these come with small impact in tail latency. Our analysis shows that architectural features such as aggressive branch predictors, large caches, and hyper-threading do not provide significant benefits in performance. The most significant performance benefit comes from better memory throughput.

We perform a cost analysis based on energy cost, which shows that S1 has 1.1-2.7x lower energy cost. If in addition we include server equipment cost and a depreciation period of 3 to 5 years (server lifetime), then total cost efficiency depends on server purchase price. For microservers, such as S1 to be more cost-effective they need to have a purchase price several times lower, and typically around or more than 3x, than higher end servers, such as S4.

In summary, the most appropriate solution for KV stores is microservers with large numbers of cores, relatively simple branch prediction, small caches, no hyper-threading, and large memory throughput, if however, they also have a significantly lower purchase price compared to high-end servers.

Bibliography

- [1] David G. Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, and Vijay Vasudevan. Fawn: A fast array of wimpy nodes. In *Proceedings of the ACM SIGOPS 22Nd Symposium on Operating Systems Principles, SOSP '09*, pages 1–14, New York, NY, USA, 2009. ACM.
- [2] Apache. Cassandra. <http://cassandra.apache.org/>. Accessed: November 18, 2019.
- [3] Apache. Hbase. <https://hbase.apache.org/>. Accessed: November 18, 2019.
- [4] Rafael Vidal Aroca and Luiz Marcos Garcia Gonçalves. Towards green data centers: A comparison of x86 and arm architectures power efficiency. *Journal of Parallel and Distributed Computing*, 72(12):1770–1780, 2012.
- [5] Luiz André Barroso, Jimmy Clidaras, and Urs Hölzle. The datacenter as a computer: An introduction to the design of warehouse-scale machines. *Synthesis lectures on computer architecture*, 8(3):1–154, 2013.
- [6] Emily Blem, Jaikrishnan Menon, and Karthikeyan Sankaralingam. Power struggles: Revisiting the risc vs. cisc debate on contemporary arm and x86 architectures. In *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on*, pages 1–12. IEEE, 2013.
- [7] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C Hsieh, Deborah A Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):4, 2008.
- [8] B. F. Cooper. Core workloads. <https://github.com/brianfrankcooper/YCSB/wiki/Core-Workloads>. 2018.
- [9] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilch, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: amazon’s highly available key-value store. In *ACM SIGOPS operating systems review*, volume 41, pages 205–220. ACM, 2007.

- [10] Facebook. Rocksdb. <http://rocksdb.org/>, 2018.
- [11] Michael Ferdman, Almutaz Adileh, Onur Kocberber, Stavros Volos, Mohammad Alisafae, Djordje Jevdjic, Cansu Kaynak, Adrian Daniel Popescu, Anastasia Ailamaki, and Babak Falsafi. Clearing the clouds: A study of emerging scale-out workloads on modern hardware. In *Proceedings of the Seventeenth International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS XVII*, pages 37–48, New York, NY, USA, 2012. ACM.
- [12] Google. Leveldb. <http://leveldb.org/>, 2018.
- [13] Vasileios Kontorinis, Liuyi Eric Zhang, Baris Aksanli, Jack Sampson, Houshan Homayoun, Eddie Pettis, Dean M Tullsen, and Tajana Simunic Rosing. Managing distributed ups energy for effective power capping in data centers. In *2012 39th Annual International Symposium on Computer Architecture (ISCA)*, pages 488–499. IEEE, 2012.
- [14] Chunbo Lai, Song Jiang, Liqiong Yang, Shiding Lin, Guangyu Sun, Zhenyu Hou, Can Cui, and Jason Cong. Atlas: Baidu’s key-value storage system for cloud data. In *Mass Storage Systems and Technologies (MSST), 2015 31st Symposium on*, pages 1–14. IEEE, 2015.
- [15] Harold Lim, Aman Kansal, and Jie Liu. Power budgeting for virtualized data centers. In *2011 USENIX Annual Technical Conference (USENIX ATC’11)*, volume 59, 2011.
- [16] Dumitrel Loghin, Bogdan Marius Tudor, Hao Zhang, Beng Chin Ooi, and Yong Meng Teo. A performance study of big data on small nodes. *Proc. VLDB Endow.*, 8(7):762–773, February 2015.
- [17] John D. McCalpin. Memory bandwidth and machine balance in current high performance computers. *IEEE Computer Society Technical Committee on Computer Architecture (TCCA) Newsletter*, pages 19–25, December 1995.
- [18] microchip. Mcp39f511a power monitor demonstration board. https://www.microchip.com/DevelopmentTools/ProductDetails/adm00667#utm_medium=Press-Release&utm_term=MCP39F511_PR_4-21-15&utm_content=AIPD&utm_campaign=Board.
- [19] Patrick O’ Neil, Edward Cheng, Dieter Gawlick, and Elizabeth O’ Neil. The log-structured merge-tree (lsm-tree). *Acta Informatica*, 33(4):351–385, 1996.
- [20] Zhonghong Ou, Bo Pang, Yang Deng, Jukka K Nurminen, Antti Yla-Jaaski, and Pan Hui. Energy-and cost-efficiency analysis of arm-based clusters. In *Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012)*, pages 115–123. IEEE Computer Society, 2012.

- [21] Anastasios Papagiannis, Giorgos Saloustros, Pilar González-Férez, and Angelos Bilas. Tucana: Design and implementation of a fast and efficient scale-up key-value store. In *USENIX Annual Technical Conference*, pages 537–550, 2016.
- [22] Anastasios Papagiannis, Giorgos Saloustros, Pilar González-Férez, and Angelos Bilas. An efficient memory-mapped key-value store for flash storage. In *Proceedings of the ACM Symposium on Cloud Computing*, SoCC '18, pages 490–502, New York, NY, USA, 2018. ACM.
- [23] Nikola Rajovic, Lluís Vilanova, Carlos Villavieja, Nikola Puzovic, and Alex Ramirez. The low power architecture approach towards exascale computing. *Journal of Computational Science*, 4(6):439–443, 2013.
- [24] Jinglei Ren. Ycsb-c. <https://github.com/basicthinker/YCSB-C>, 2016.
- [25] Seagate. Data age 2025. <https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>. Accessed: November 18, 2019.
- [26] Yevgenly Sverdlik. Paypal deploys arm servers in data centers. <https://www.datacenterknowledge.com/archives/2015/04/29/paypal-deploys-arm-servers-in-data-centers>. April 29, 2015.
- [27] Kashi Venkatesh Vishwanath and Nachiappan Nagappan. Characterizing cloud computing hardware reliability. In *Proceedings of the 1st ACM Symposium on Cloud Computing*, SoCC '10, pages 193–204, New York, NY, USA, 2010. ACM.