Key-value databázové systémy

Key-Value Database Management Systems

Bc. Jan Jedlička

Vedoucí: prof. Ing. Michal Krátký, Ph.D.

FEI, VŠB-TUO

2024

Úvod

- Studium problematiky key-value databázových systémů (KDBS [1])
- Návrh a implementace testovací prostředí pro porovnání KDBS s ostatními DBS (YCSB [2])
- Otestování vybraných KDBS a vyhodnocení výsledků experimentů (Redis [3], Aerospike [4], Memcached [5], Riak KV [6])

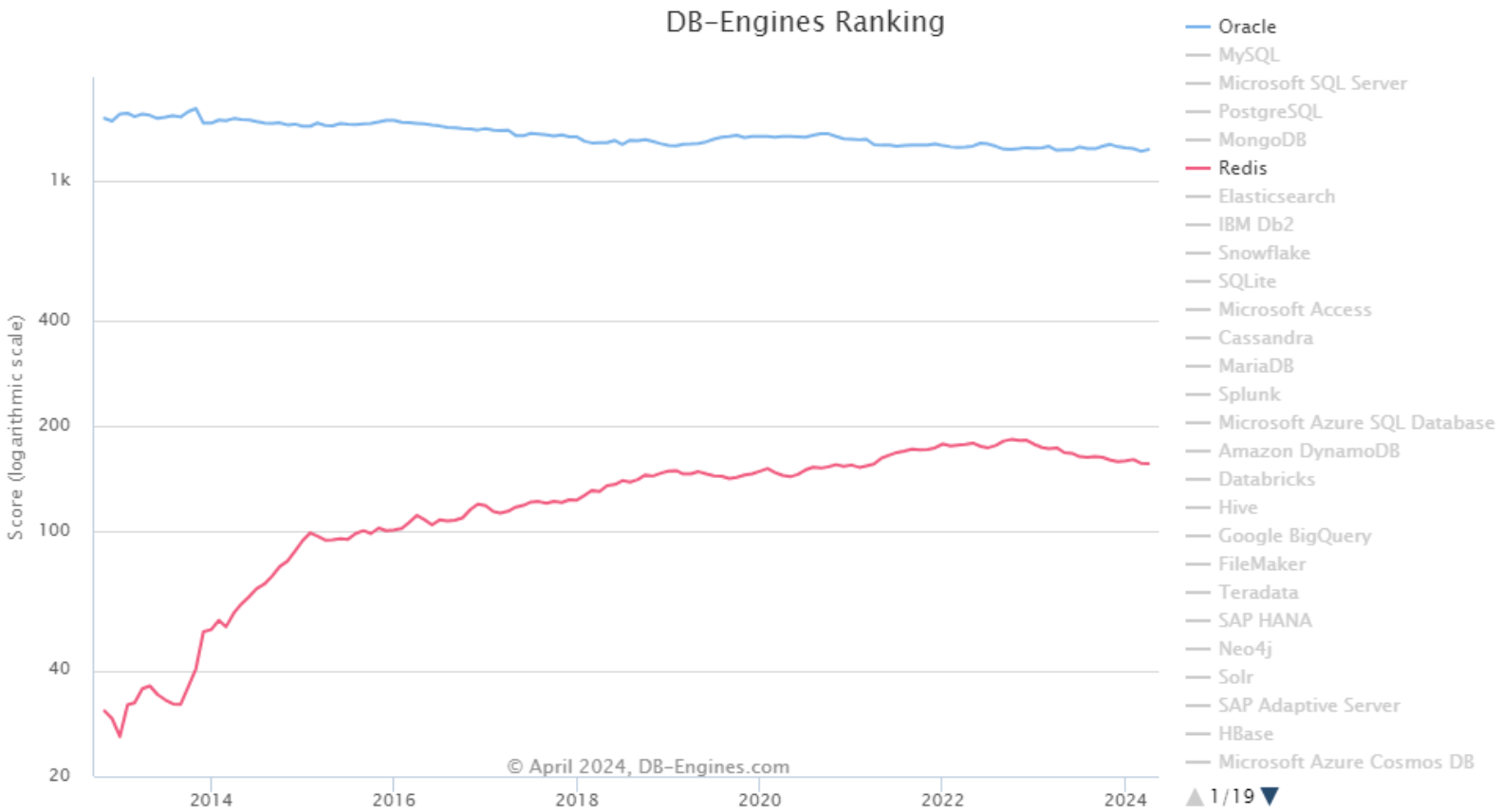
Not Only SQL (NoSQL)

- Data nezávislá na schématu (schema-less), schéma je možné využít
- Zpracování nesouvisejících nebo rychle se měnících dat [7]
- Vyžadujeme výkon a dostupnost oproti silné konzistenci
- Distribuované (sdílené) úložiště
- Horizontální škálovatelnost (paralelní přidávání nezávislých výpočetních kapacit, rozkládání zátěže mezi tyto kapacity)

Not Only SQL (NoSQL)

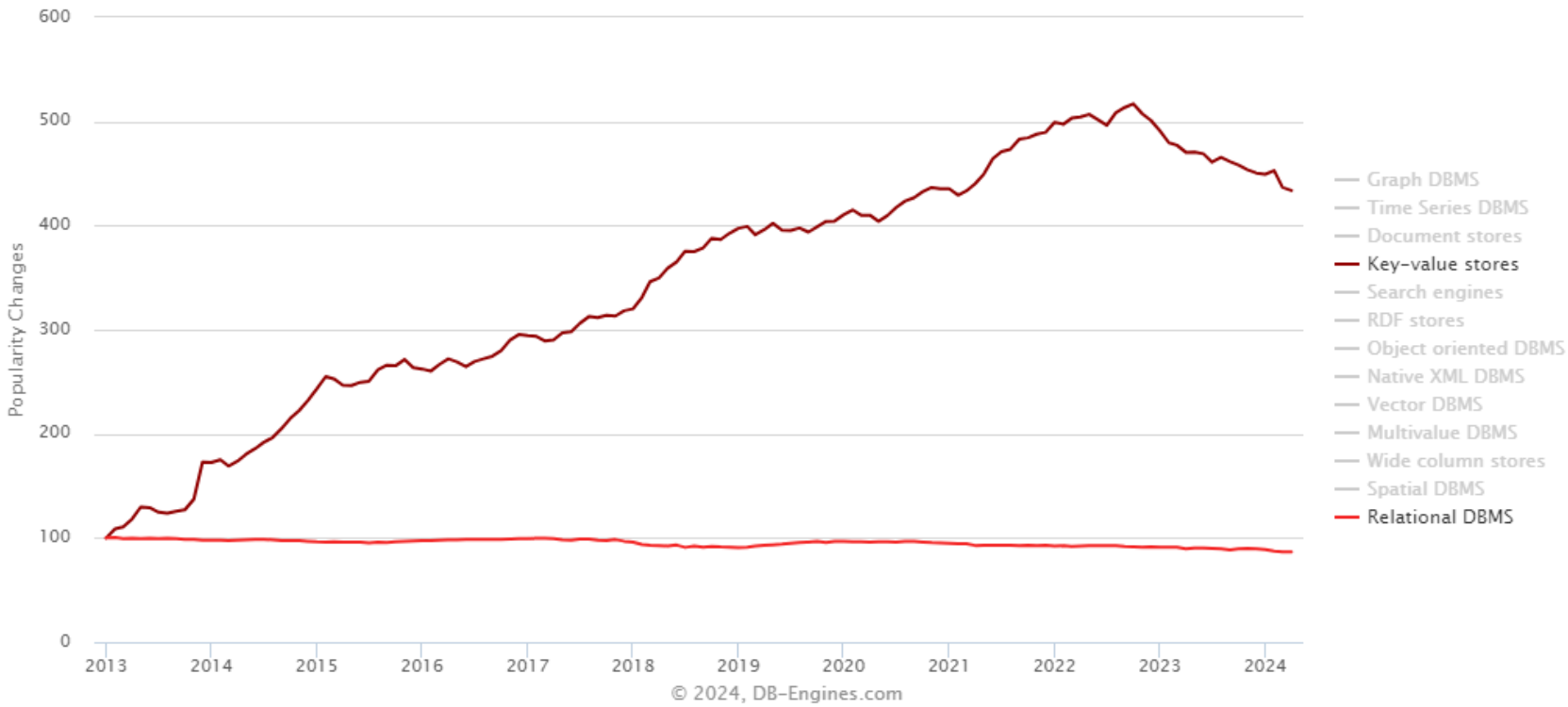
- Klíč-hodnota DBS (Redis [3], Aerospike [4])
- Dokumentové DBS (MongoDB [8], CouchDB [9])
- Sloupcové DBS (Apache Cassandra [10], Apache HBase [11])
- Grafové DBS (Neo4J [12], Azure Cosmos DB [13])

RDBS Oracle vs KDBS Redis



Obrázek: Graf hodnot popularity RDBS Oracle a KDBS Redis [14]

RDBS vs KDBS



Obrázek: Graf změn hodnot popularity RDBS a KDBS [15]

KDBS

- Databáze párů klíč-hodnota libovolného formátu
- Klíč je unikátní identifikátor umožňující rychlý přístup k hodnotám
- Základní operace (get(k), insert(k,h), delete(k), update(k,h)) [16]

KDBS - hodnota

```
{
   "jmeno": "Jan Novak",
    "vek": 35,
    "adresa": null,
    "kontakt": [
        {"email": "jan.novak.01@gmail.com"},
        {"telefon": "406-792-448"}
    ]
}
```

Obrázek: Hodnota (JSON dokument) uložená pro klíč "userid_135"

Testovací prostředí

- Porovnání vlastností systémů (propustnost, bezpečnost, škálovatelnost)
- Nástroj pro nastavitelné a opakovatelné testy
- Řada nástrojů pro různorodé využití
- ApexSQL Diff, Apache JMeter, QuerySurge, Redgate SQL Test [17]

TPC

- Transaction Processing Performance Council (TPC) [18]
- Relační DBS
- Komplexní testování, náročné operace (agregace, seřazení, průměr, spojení) nad velkými daty
- Počet transakcí za minutu (tpm)
- TPC-C, TPC-H, TPCx-AI, TPCx-IoT, TPCx-BB

TPC-H - SQL Test Q1

```sql
--Q1
select
  l_returnflag,
  l_linestatus,
  sum(l_quantity) as sum_qty,
  sum(l_extprice) as sum_base_price,
  sum(l_extprice * (1 - l_discount)) as sum_disc_price,
  sum(l_extprice * (1 - l_discount) * (1 + l_tax)) as sum_charge,
  avg(l_quantity) as avg_qty,
  avg(l_extprice) as avg_price,
  avg(l_discount) as avg_disc,
  count(*) as count_order
from lineitem
where  l_shipdate <= date '1998-12-01'
group by  l_returnflag,  l_linestatus
order by  l_returnflag,  l_linestatus;
```

Obrázek: TPC-H, SQL Test Q1 [19, 20]

YCSB

- Yahoo! Cloud Serving Benchmark (YCSB) [2]
- Porovnávání výkonu NoSQL DBS (KDBS)
- Scénáře využití (Workload A-F)
- Klíč - řetězec 'user_id'
- Hodnota - JSON dokument s poli ('field0'-'fieldN')

YCSB - operace vložení

```
1  {
2    "operation": "insert",
3    "table": "usertable",
4    "key": "user1474",
5    "values": {
6      "field0": "value0",
7      "field1": "value1",
8      "field2": "value2"
9    }
10 }
```

Obrázek: YCSB - příklad definice operace vložení záznamu

Zprovoznění testů

- Java JDK 8 (1.8) [21]
- Nastavení %JAVA_HOME% systémové proměnné [22] (OS Windows)
- Docker [23]
- YCSB 0.17 [24]
- Apache Maven 3 [25]

Zprovoznění testů - Docker

```
PS E:\ycsb-0.17.0\bin> docker pull redis:latest
latest: Pulling from library/redis
Digest: sha256:f14f42fabc123example456c7e824b9
Status: Image is up to date for redis:latest
docker.io/library/redis:latest

PS E:\ycsb-0.17.0\bin> docker run --name my-redis -p 6379:6379 -d redis:latest
aba5fabc123example4568aa97

PS E:\ycsb-0.17.0\bin> docker stop aba5fabc123example4568aa97
aba5fabc123example4568aa97

PS E:\ycsb-0.17.0\bin> docker rm aba5fabc123example4568aa97
aba5fabc123example4568aa97
```

Obrázek: Docker - příkazy pro stažení, spuštění, zastavení a odstranění DBS Redis

Zprovoznění testů - YCSB Load

```
PS E:\ycsb-0.17.0\bin> .\ycsb load redis -P ..\workloads\workloada
        -p redis.host=127.0.0.1 -p redis.port=6379 -p recordcount=10000
        -p threadcount=4
```

```
[OVERALL], RunTime(ms), 4211
[OVERALL], Throughput(ops/sec), 2374.7328425552128
```

```
[INSERT], Operations, 10000
[INSERT], AverageLatency(us), 1647.61
[INSERT], MinLatency(us), 904
[INSERT], MaxLatency(us), 13895
[INSERT], 95thPercentileLatency(us), 2481
[INSERT], 99thPercentileLatency(us), 3389
[INSERT], Return=OK, 10000
```

Obrázek: YCSB Redis, příkaz pro vložení dat (load)

Zprovoznění testů - YCSB Run

```
PS E:\ycsb-0.17.0\bin> .\ycsb run redis -P ..\workloads\workloada
        -p redis.host=127.0.0.1 -p redis.port=6379 -p operationcount=10000
        -p threadcount=4
```

```
[OVERALL], RunTime(ms), 2560
[OVERALL], Throughput(ops/sec), 3906.25
```

```
[READ], Operations, 5157
[READ], AverageLatency(us), 998.8885010665116
[READ], MinLatency(us), 389
[READ], MaxLatency(us), 17695
[READ], 95thPercentileLatency(us), 1692
[READ], 99thPercentileLatency(us), 3479
[READ], Return=OK, 5157
```

```
[UPDATE], Operations, 4843
[UPDATE], AverageLatency(us), 985.7646087136072
[UPDATE], MinLatency(us), 364
[UPDATE], MaxLatency(us), 17567
[UPDATE], 95thPercentileLatency(us), 1744
[UPDATE], 99thPercentileLatency(us), 3675
[UPDATE], Return=OK, 4843
```

Obrázek: YCSB Redis, příkaz pro spuštění testů (run)

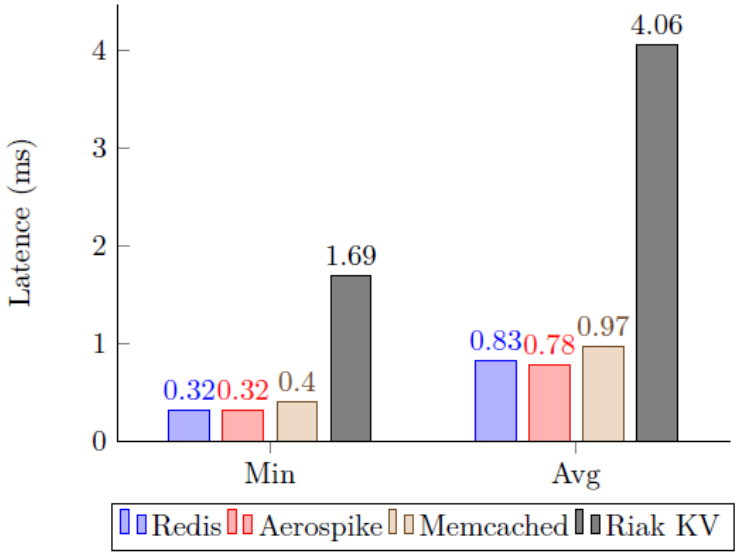
Zprovoznění testů - Spouštěcí skript

- Windows PowerShell ISE (Integrované skriptovací prostředí)
- Automatizace, nahrazení ruční manipulace se skripty

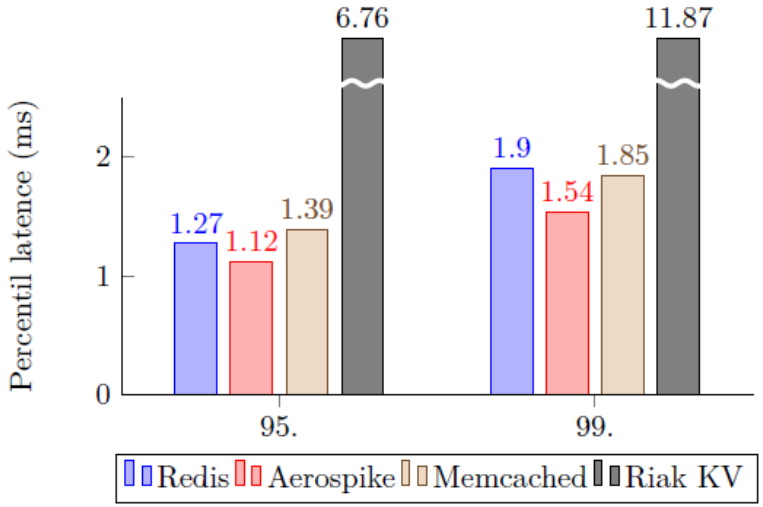
Parametry testů

- Počet záznamů v databázi (recordcount)
- Počet testovaných operací (operationcount)
- Název DBS
- Workload (A-C)
- Případná konzistence (Riak KV měl chybně nastavenou silnou konzistenci)

Vyhodnocení testů - Workload A (50% čtení, 50% aktualizace)

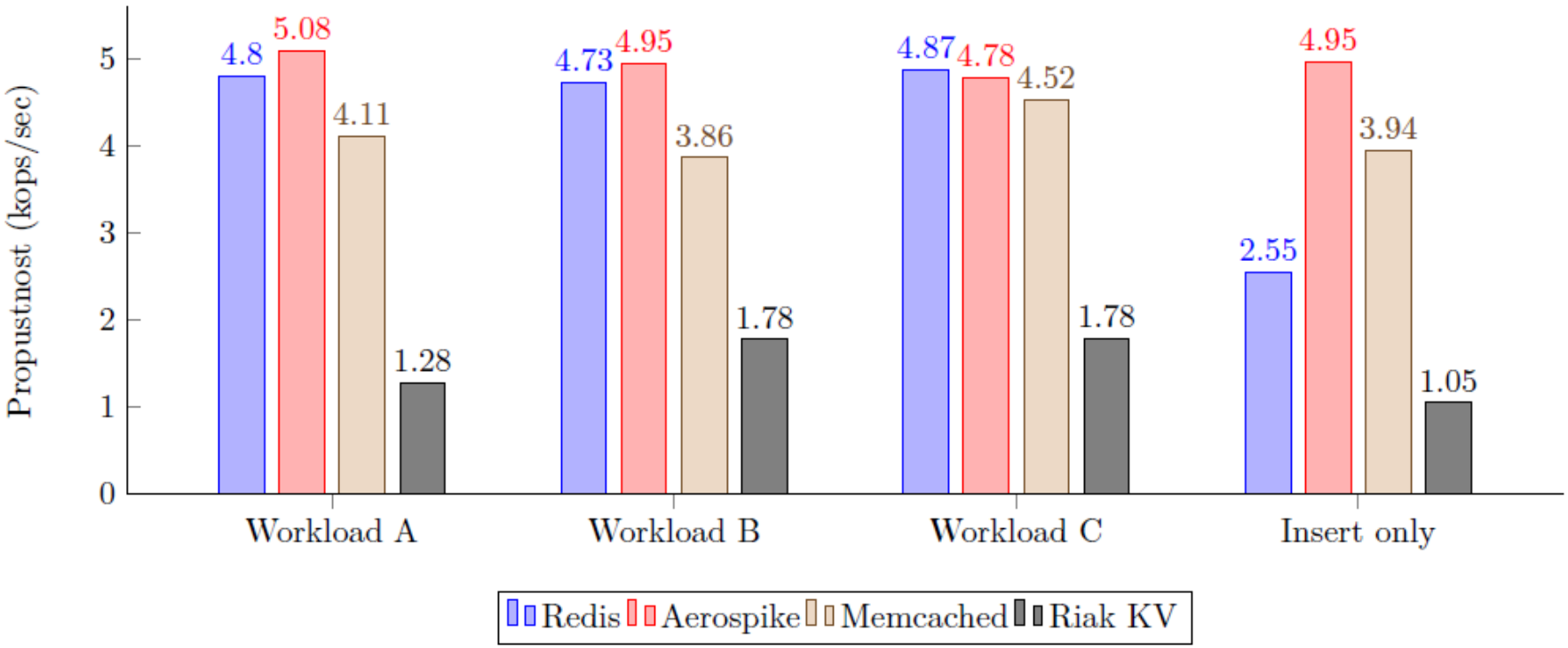


Obrázek: Min a Avg latence (ms)



Obrázek: Percentil latence (ms)

Vyhodnocení testů - Propustnost (kops/sec)



Obrázek: Workload A, B, C + Insert only - Propustnost (kops/sec)

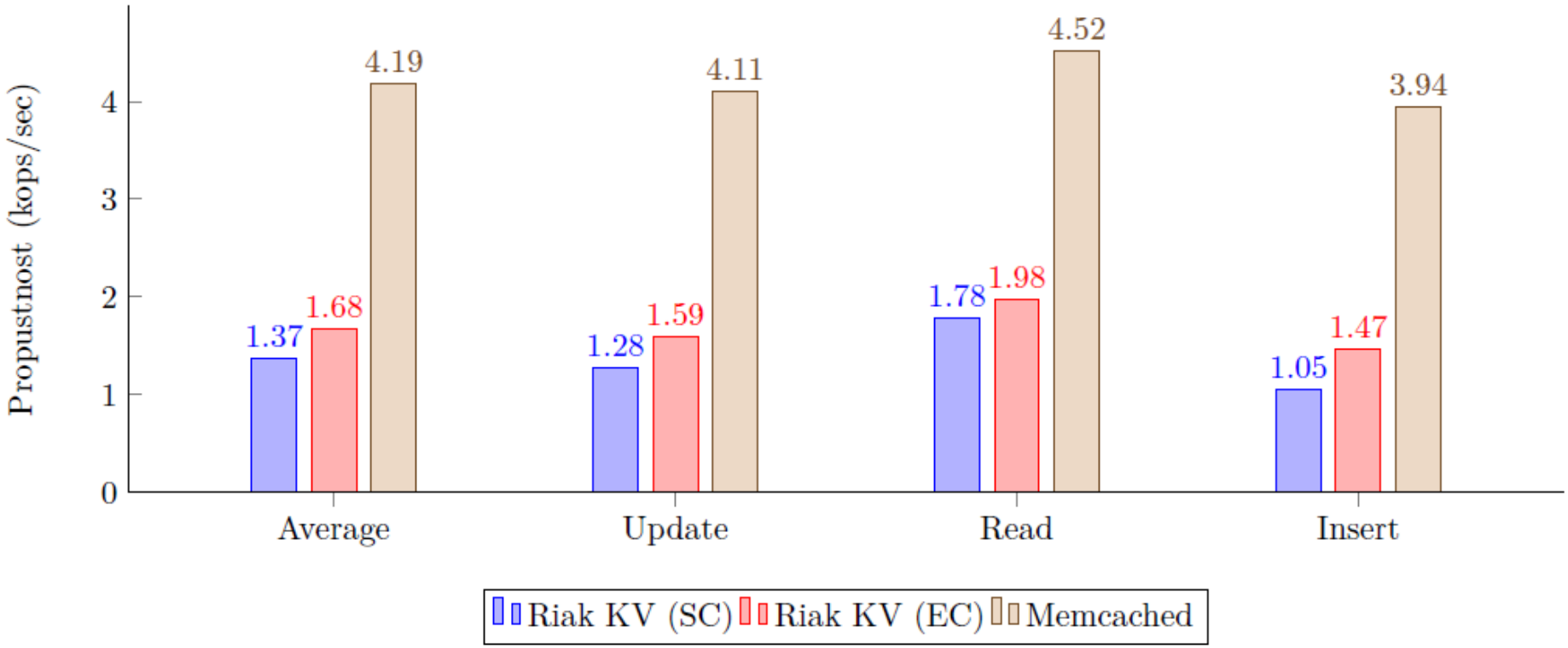
Vyhodnocení testů - Souhrn

Propustnost	1.	2.	3.	4.
Průměr všech operací	Aerospike 4,9 kops/s	Redis 4,2 kops/s	Memcached 4,1 kops/s	Riak KV 1,5 kops/s
Vkládání	Aerospike	Memcached	Redis	Riak KV
Aktualizace	Aerospike	Redis	Memcached	Riak KV
Čtení	Redis	Aerospike	Memcached	Riak KV

Na třetím řádku je uvedena průměrná propustnost všech testů (tisíce operací za sekundu) pro výše vypsaný KDBS

Tabulka: Porovnání propustnosti výsledků testů

Riak KV, Silná konzistence (SC) vs Případná konzistence (EC)



Obrázek: Riak KV SC vs EC - Propustnost (kops/sec), zlepšení o 22,6 %

Riak KV, Nízká propustnost

- Přenastavení z SC na EC zvýšilo průměrnou propustnost o ∼310 operací za sekundu (22,6 %), KDBS Memcached je stále rychlejší o 2,51 kops/sec (149,41 %)
- Úplné vyřešení problému s chybou u aktualizačních operací (∼5 %) přechodem na EC, nedochází ke konfliktům a operace nejsou zrušeny
- MapReduce je určen pro dávkové zpracování, nikoli pro dotazy v reálném čase

Poděkování

Děkuji za pozornost

Připomínky

- Využití 4 vláken?
  - Simulace paralelního provádění operací 4 klientů současně
  - Více vláken nebylo použito skrz HW omezení stroje
- Amazon DynamoDB úrovně izolace
  - Podpora Read Uncommitted, Read Committed a Repeatable Reads
  - Nepodporuje úroveň izolace Serializable
- Redis datová struktura hash
  - Redis hash - Hashovací tabulka
  - Ideální pro objekty s řadou vlastností
- Aerospike Hybrid-memory
  - Index je v paměti, data na disku (nejedná se o In-memory)
  - Při změně dat dojde k aktualizaci indexu v paměti a následně k zápisu změn na disk
  - Index pro klíč uchovává metadata o lokaci záznamu na disku pro získání chybějících hodnot

Připomínky

- Server-side clustering
  - Dynamické nahrazení nedostupných serverů v rámci clusteru
  - Horizontální škálovatelnost přidáním serverů do clusteru
- Customer deployment
  - Instalace, konfigurace a spuštění systému pro koncové uživatele
  - Plánování, testování, implementace a podpora
- Fail-over
  - Dostupnost služeb při výpadku primárního serveru
  - Přesun na záložní server či repliku při nedostupnosti primárního serveru
- Hot-standby
  - Poskytuje rychlý přechod a minimální dobu výpadku
  - Hot-standby server je v reálném čase synchronizován s primárním serverem (okamžité replikování dat, dostupnost náhradní aktualizované kopie)

Citace I

- *What Is a Key-Value Database?* [online]. 2022. [cit. 2022-11-18]. Dostupné z: https://aws.amazon.com/nosql/key-value/.
- *Yahoo! Cloud Serving Benchmark (YCSB)* [online]. 2019. [cit. 2024-04-22]. Dostupné z: https://github.com/brianfrankcooper/YCSB.
- *Redis* [online]. 2022. [cit. 2022-11-18]. Dostupné z: https://redis.io/.
- *Aerospike* [online]. 2022. [cit. 2022-11-20]. Dostupné z: https://aerospike.com/.
- *What is Memcached?* [online]. 2024. [cit. 2024-04-17]. Dostupné z: https://memcached.org/.
- *Riak KV* [online]. 2022. [cit. 2022-11-21]. Dostupné z: https://riak.com/products/riak-kv/index.html.

Citace II

- *Databáze NoSQL* [online]. 2024. [cit. 2024-08-03]. Dostupné z: https://azure.microsoft.com/cs-cz/resources/cloud-computing-dictionary/what-is-nosql-database.
- *MongoDB* [online]. 2023. [cit. 2023-01-28]. Dostupné z: https://www.mongodb.com/.
- *CouchDB* [online]. 2024. [cit. 2024-06-10]. Dostupné z: https://couchdb.apache.org/.
- *Cassandra* [online]. 2023. [cit. 2023-01-28]. Dostupné z: https://cassandra.apache.org/_/index.html.
- *Welcome to Apache HBase* [online]. 2024. [cit. 2024-06-10]. Dostupné z: https://hbase.apache.org/.
- *Neo4j* [online]. 2024. [cit. 2024-06-10]. Dostupné z: https://neo4j.com/.

Citace III

- *Azure Cosmos DB* [online]. 2024. [cit. 2024-04-22]. Dostupné z: https://azure.microsoft.com/cs-cz/products/cosmos-db.
- *DB-Engines Ranking - Trend Popularity* [online]. 2024. [cit. 2024-04-23]. Dostupné z: https://db-engines.com/en/ranking_trend.
- *DBMS popularity broken down by database model* [online]. 2024. [cit. 2024-04-23]. Dostupné z: https://db-engines.com/en/ranking_categories.
- BAČA, Radim. *Nerelační distribuované databáze* [online]. 2024. [cit. 2024-08-03]. Dostupné z: https://db.cs.vsb.cz/Download.ashx?id=6.
- *Top 10 Database Testing Tools With Features, Cons and Pros* [online]. 2024. [cit. 2024-04-27]. Dostupné z: https://testsigma.com/blog/database-testing-tools/.

Citace IV

- *TPC* [online]. 2023. [cit. 2023-02-11]. Dostupné z: https://www.tpc.org/.
- *TPC-H Benchmark - Test SQL* [online]. 2024. [cit. 2024-04-27]. Dostupné z: https://docs.starrocks.io/docs/benchmarking/TPC-H_Benchmarking/.
- *TPC-H Vesion 2 and Version 3* [online]. 2024. [cit. 2024-04-27]. Dostupné z: https://www.tpc.org/tpch/.
- *Oracle - Java Downloads* [online]. 2024. [cit. 2024-04-25]. Dostupné z: https://www.oracle.com/java/technologies/downloads/.
- *Windows Environment Variables* [online]. 2024. [cit. 2024-04-26]. Dostupné z: https://ss64.com/nt/syntax-variables.html.
- *Docker Builds: Now Lightning Fast* [online]. 2024. [cit. 2024-03-19]. Dostupné z: https://www.docker.com/.

Citace V

Download the latest release of YCSB [online]. 2024. [cit. 2024-04-25]. Dostupné z: https://github.com/brianfrankcooper/YCSB/releases/download/0.17.0/ycsb-0.17.0.tar.gz.

Apache Maven Project [online]. 2024. [cit. 2024-04-25]. Dostupné z: https://maven.apache.org/download.cgi.