CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2017; **00**:1–23
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cpe

KeyValueServe†: Design and Performance Analysis of a Multi-Tenant Data Grid as a Cloud Service‡

Anwesha Das[1]* Arun Iyengar[2] and Frank Mueller[1]

[1]*North Carolina State University*
[2]*IBM T. J. Watson Research Center*

SUMMARY

Distributed key-value stores have become indispensable for large scale cluster applications. Many cloud services have deployed in-memory data grids for their enterprise infrastructures and support multi-tenancy services. However, most services do not offer fine-grained multi-tenant resource sharing. To this front, we present *KeyValueServe*, a low overhead cloud service with features aiding resource management. Results based on Hazelcast, a popular open source data grid, indicate that KeyValueServe can efficiently provide services to tenants without degrading performance. Providing consistent performance to all tenants for fluctuating workloads is still difficult. Performance problems occur at scale with diverse tenant requirements. To address this, the paper provides insights to contention and performance bottlenecks. Through experimental analysis, we uncover scenarios of performance degradation and demonstrate optimized performance via coalescing multiple clients' requests. Our work indicates that a Hazelcast cluster can get congested with multiple concurrent connections when processing client requests, resulting in poor performance. KeyValueServe can reduce the number of parallel connections maintained for client requests, resulting in improved performance. Copyright © 2017 John Wiley & Sons, Ltd.

Received . . .

KEY WORDS: Multi-Tenancy; Performance; Service; NoSQL; In-memory; Data-Grid; Key-Value Store; Cloud Computing

1. INTRODUCTION

Innovative key-value stores for data intensive computations have been studied thoroughly in recent times such as [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. Both academically and commercially, high performance key-value stores have recently gained substantial attention. Table I enumerates the same. While some of these are in-memory, the others are disk-based, comprised of Erlang, C, C++ and Java based solutions. These applications use atleast one if not all of Java Native, ReST and RPC-style client protocols. *Platform* and *Persistence* in Table I indicate the main programming language used to develop the enlisted stores and what way persistence is enabled in them, respectively. It should be noted that many key-value stores can easily be used by applications written in a

*Correspondence to: E-mail: adas4@ncsu.edu

Contract/grant sponsor: National Science Foundation; contract/grant number: 1217748 and 0958311
†Key-Value Store as a Service
‡Part of this work was previously published at IEEE Cloud 2016 as a short paper. This paper presents KeyValueServe framework not present in our earlier paper. In addition to the cloud service model, the paper discusses various experimental results pertaining to different workload types, distribution types, and client-side thread count variation not described in the short paper. This paper presents the contention detection analysis and demonstrates optimization approach in considerably more detail than the previous paper.

Table I. Popular Key-Value Stores.

Name	Platform	Open Source	Persistence
BigTable	Java	No	Yes
PNUTS	C++	No	Yes, update logs
Dynamo	Java	No	Pluggable Backing store
MongoDB	C/C++	Yes	Yes
Voldemort	Java	Yes	Pluggable
Hadoop	Java	yes	Local OS file system
HBase	Java	Yes	On-disk
HyperTable	C++	Yes	On-disk
HyperDex	C/C++	Yes	On-disk
Comet	Lua	No	Based on Active Storage Objects
Silt	C++	No	Flash drive based
Cassandra	Java	Yes	Custom on-disk
Memcached	C	Yes	Berkeley DB as backend store
Redis	C	Yes	Snapshots on-disk
CouchDB	Erlang	Yes	On-Disk
Hazelcast	Java	Yes	In-Memory

wide variety of programming languages. NoSQL stores such as MongoDB and DynamoDB are already hosted on clouds such as Amazon Web Services. Google's Cloud Storage sits on its AppEngine, which uses various NoSQL solutions like Python dataStore, MongoDB, Cassandra and RabbitMQ. Similar services available include Joyent [14] offering Riak [15] and Cloudant [16] hosting CouchDB [17]. They have different pricing models based on their architecture and design. In spite of these available services, fine-grained multi-tenant resource sharing (such as a VM or a distributed data structure) is not well studied. Evaluating the trade-offs of designing an in-memory data grid as a cloud service can be an eye-opener to understand its benefits in the context of multi-tenant performance.

Memcached [11] in particular has been used by researchers [18, 19, 20, 21, 22, 23, 24] to solve problems related to key-value stores. Besides commercial usage [10, 19, 12, 25, 26, 16, 17, 14, 1, 2, 3, 15], novel key-value stores have been contributed from academia, e.g., Silt [13], Voldemort [8], Hyperdex [7] and Comet [9]. Sustained performance is a primary concern for tenants accessing such stores under fluctuating workloads especially in the context of cloud computing.

1.1. Challenges and Motivation

Key-value stores cater to a combination of read (get) and write (put) requests. Ensuring enhanced throughput for ever increasing numbers of tenants is challenging because:

- Data placement and eviction in such stores is oblivious to external tenant and data characteristics. For example, suppose tenant A's and B's keys share the same cluster instance. If every time B's data is accessed there is a high probability that A's data gets evicted, then A may experience poor performance due to B.
- Since each get/put task has relatively short task duration, the average task response time, including scheduling/queuing etc., cannot be very high. Executing such tasks while achieving consistent performance is hard.
- Every operation accesses some previously stored data in a typical key-value store unlike queries such as join or merge. Hence, *co-location of data and computation* is important in such clusters. Moreover, a single cluster instance may be over stressed serving multiple tenants, if that instance happens to store the requested keys of those tenants. In that case, it is difficult to ensure well balanced and distributed request handling, since other instances may be idle serving no requests.

DOI: 10.1002/cpe

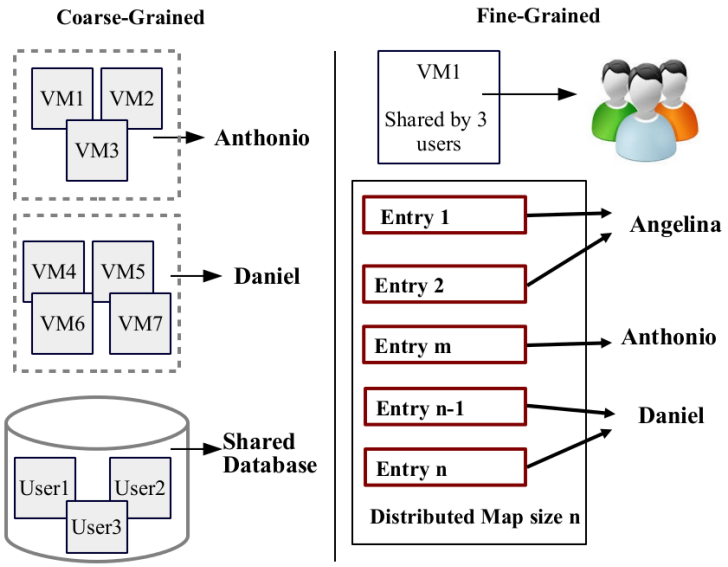


Figure 1. Multi-tenant Sharing.

- Prior work [22] has indicated an imbalance in data center environments such as workload skews and fluctuating request patterns. A system coping with such fluctuations that tries to deliver the desired throughput encounters resource contention.
- With an increase in the number of instances in a cluster and clients, the available network bandwidth becomes a bottleneck. Hence, performance suffers under high cluster load even with access to sufficient resources. This network inflated delay can be particularly *overwhelming* for low latency operations.
- Based on workload *size* and *type*, throughput tends to vary. Ensuring tenant performance is a challenge because of the inherent unpredictability of the workload.

These issues coupled with the maintenance problems constantly faced by existing cloud services have inspired the development of KeyValueServe and our corresponding performance study. Addressing the discussed challenges, our proposed framework handles issues related to co-location of data and computation, network-inflated delay, tenant starvation and isolation. Our objectives of improving multi-tenant performance and establishing a better service model for key-value stores conform to the major open problems mentioned by Agrawal et al. [27] while discussing the desiderata in the context of cloud computing. We have described them in detail in Section 2.

1.2. Terminology Description

We describe three commonly used terms here to indicate their meanings in the context of this work and avoid misunderstandings.

1.2.1. Multi-tenancy: Most cloud services [28, 29, 30] claim to provide multi-tenancy as a service. In reality tenants are sharing the overall infrastructure but not individual entities like virtual machines or containers. These coarse-grained solutions provide service simultaneously to several customers without sharing a virtual guest, data structure or container. Each customer is usually allocated a bunch of VMs for its individual use without any sharing. In our work, we define multi-tenancy as *sharing cloud resources at a finer granularity such as sharing a single instance of the data grid, a single VM part of the key-value store, or a single data structure by multiple tenants*. Since in-memory sharing and distributed computing have implications on data structure level sharing, this work discusses a service model pertaining to fine-grained multi-tenancy where unique tenants can seamlessly share a *map* data structure or a cluster *instance* without any inconsistencies. Figure 1 illustrates coarse-grained sharing where users Anthonio and Daniel have dedicated VMs

DOI: 10.1002/cpe

allocated for themselves and the database is shared with three users. Fine-grained sharing is highlighted in the same figure where a single VM and a map data structure are shared by the three users. `Map data structure` in this paper refers to Hazelcast's distributed map API used for storage and retrieval of keys.

1.2.2. Throughput Degradation: When we refer to performance degradation in this paper, we always refer to *reduced throughput or drop* in terms of operations per second. It should be noted that throughput has been considered from both the server and client's perspective, since both system throughput and per client throughput indicate quality of performance. We clarify early on that per client throughput refers only to the throughput perceived at each client and system throughput refers to the total amount of operations (considering all the clients) serviced from the server's perspective. In the context of multi-tenancy, per client throughput is important for evaluation. The overall system throughput helps us to understand the system saturation point and gives indications of overload situations.

We reiterate the terms *multi-tenancy* and *performance degradation*, hence understanding them in the right sense is important.

1.2.3. Cluster Instance: We use the terms cluster instance, server, node and client frequently. A cluster instance refers to a key-value store instance (Hazelcast in this case) in the cluster, which is a specific JVM or an object instance. This need not necessarily refer to a different physical node. These instances can be either servers or clients. Typically, we refer to servers since clients are external instances sending requests to the Hazelcast cluster in this paper.

1.3. Contributions

With the rise of big data processing and cloud applications, it is understandable that such NoSQL based solutions enrich cloud computing. These solutions possess their own strengths and weaknesses. In spite of the existence of these solutions, certain facets of Quality of Service (QoS) and performance (e.g., fine-grained resource sharing) have not been emphasized enough. Features such as multi-tenancy, novel performance optimization, increased multiplexing with controlled scalability and flexible pricing models can all together provide a better cloud service. Keeping in mind *service* and *performance*, this paper makes the following contributions:

1. We design and implement *KeyValueServe*, a novel cloud service framework and discuss its architecture.
2. We describe the key features of such a multi-tenant service and demonstrate its negligible overhead.
3. We conduct experiments with the Hazelcast [31] key-value store to observe characteristics and client-performance.
4. We identify causes of performance degradation and develop an approach based on coalescing clients' requests to optimize performance of tenants. Our optimizations result in more than 10% improvement in throughput.

KeyValueServe is built on top of Hazelcast to provide a low-overhead service with several helpful features catering to tenant requirements. Our study substantiates the fact that performance degradation is indeed high as the number of clients increases and that an increase in the number of parallel connections with a rising number of clients is one cause for it. We further show how to alleviate performance degradation in Hazelcast through multiplexing multiple client connections maintaining fewer connection instances. Instead of processing every client request individually, we obtained better performance by processing multiple requests, which occur in close temporal proximity to each other in a single batch.

The results from our study can be used to improve performance in similar multi-tenant architectures. The salient features and key takeaways of this work are summarized below:

DOI: 10.1002/cpe

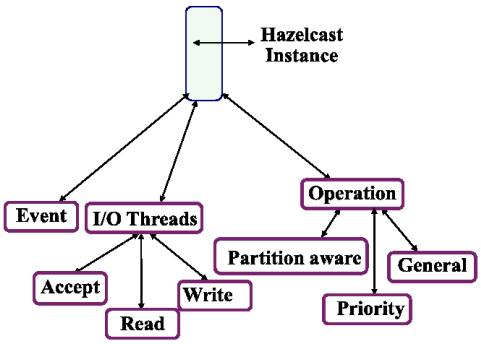


Figure 2. Threading Model.

1. Hazelcast can be used as a cloud service with fine-grained resource sharing such as multi-tenant access to distributed data structures (e.g., maps), cluster instances and virtual machines with negligible overhead. (Section 2)
2. JVM-VCPU pinning does not help in reducing contention arising in Hazelcast. (Section 3.3)
3. A well distributed workload across multiple clients improves the perceived response time. (Section 3.3)
4. With increasing number of clients, contention increases and per client throughput decreases. However, increasing the number of client side threads does not explicitly degrade performance. (Section 3.4)
5. If all the clients have a similar transaction pattern, concurrent data access invariably causes contention no matter where the data is stored. (Section 3.5)
6. Read (shared lock used unlike an exclusive lock-based write or update operation) throughput is considerably throttled due to contention. Improving operational latency is important for a multi-tenant service. (Section 3.6)
7. Every new client talking to an instance in the cluster gives rise to an independent connection instance followed by subsequent threads for handling the data read and write operations. Increasing the number of clients results in the creation of more simultaneous parallel connections. Coalescing multiple client requests for processing instead of handling them individually helps in improving overall per client performance. (Section 3.7)

The rest of this paper is organized as follows. Section 2 presents the KeyValueServe framework for supporting multi-tenancy in cloud services. Section 3 presents contention analysis and performance optimization. Section 4 surveys previous efforts on cloud based storage services, efficient performance isolation and resource sharing, and Section 5 concludes the paper.

2. KEYVALUESERVE FRAMEWORK

This section proposes KeyValueServe, a cloud based service that can offer a multi-tenant key-value store (Hazelcast) as a service in data centers for efficient computation. However, our ideas are applicable to other key-value stores as well.

2.1. Hazelcast

Hazelcast [31] is an open source in-memory data grid for distributed computing. Hazelcast's decentralized performance benefits and its easy deployment make it a good choice for our study. Moreover, it incorporates useful features of co-location of data and computation as well as inherent on-the-fly data redistribution on topology changes, conforming to a strict peer-to-peer model. This makes it suitable for our target stores and aids in addressing challenges of distributed request handling in a peer-to-peer service. Hazelcast sharding entails equal data and replica distribution on all the instances in the cluster intending to make all nodes fair (in terms of storage and redundancy).

DOI: 10.1002/cpe

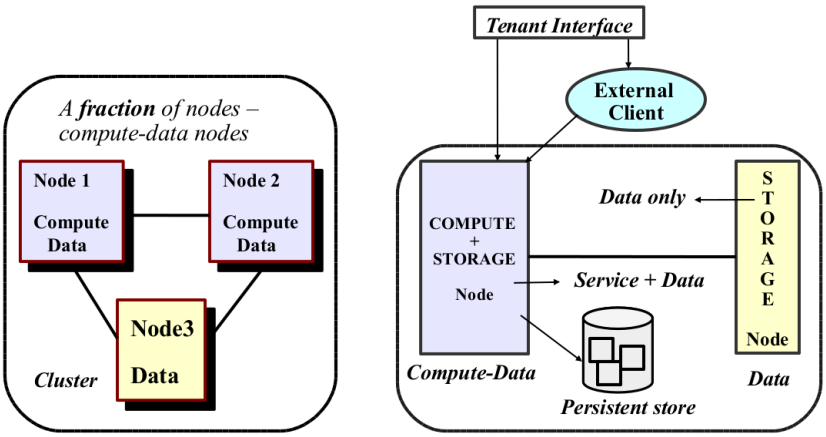


Figure 3. Components of KeyValueServe.

The partitions are equally distributed on all the cluster instances (see Section 1.2.3), and a hashing algorithm is used to map data, i.e., a key-value item, to a specific partition, i.e., an instance.

Figure 2 gives an overview of Hazelcast's threading model. Although several parameters like thread pool size and queue size are configurable, every cluster instance by default has 7 threads serving I/O operations and 5 threads handling events. Additionally, there are dedicated threads to perform partition aware, generic, or priority operations (see [31] for details). A new client operating in an instance is expensive since several multi-threaded operations are associated with a client. When a client connects to a Hazelcast server instance, the socket acceptor thread establishes communication to process client requests, which are typically various operations such as insert, read, update etc. Then subsequent threads are created such as SocketClientDataReader/Writer, read-handler/write-handler to process the requests. These threads are created every time a new client connects to the cluster with varying requests. Thus, with increasing number of clients per instance, there is an increase in the client threads per instance, and the more the clients talk to members, the higher will be the internal sharing of data structures across multiple threads. In other words, increasing the number of clients increases internal resource consumption through threads and work queues, which creates imbalance. Section 3.7 discusses the performance impact of increasing client connections, threads, and their impact on overall throughput.

2.2. *Framework*

This section describes the *KeyValueServe* prototype in more detail. The prototype is built as a decentralized peer-to-peer service model on Hazelcast. The only a priori is that the cluster has to be up and running with atleast one instance. Figure 3 shows the primary components of the design.

- *Data Nodes*: These are normal data nodes that store data and monitor resource usages sending them periodically to the compute-data nodes.
- *Compute-Data Nodes*: KeyValueServe chooses to delegate the functionalities of the service layer across multiple Hazelcast nodes without dedicating any single node for the same. This prevents single points of failure and facilitates efficient resource utilization across the nodes, which share data and execution functionalities. In other words, a subset of all the data nodes act as privileged nodes, which not only store distributed data but also perform some core services. These nodes contain the resource usage statistics and information about all the nodes of the cluster. Such privileged service nodes are called the compute-data nodes. These nodes perform services to enable KeyValueServe features we proposed in Section 2.3. These services are different from the inherent computation available in Hazelcast or any such stores where nodes perform both storage and computation. The idea is to have a peer-to-peer *cloud service model* as well apart from the native peer-to-peer computation and storage model of these stores (with compute and data nodes). This not only fits in the inherent peer-to-peer model

DOI: 10.1002/cpe

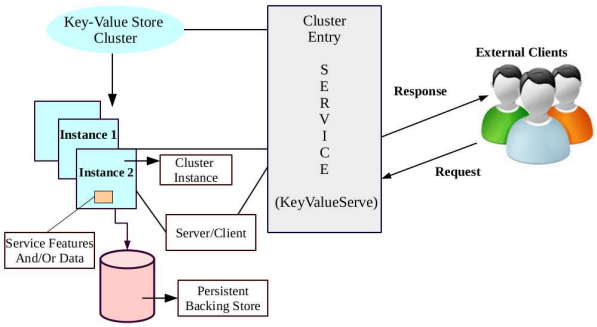


Figure 4. KeyValueServe Framework.

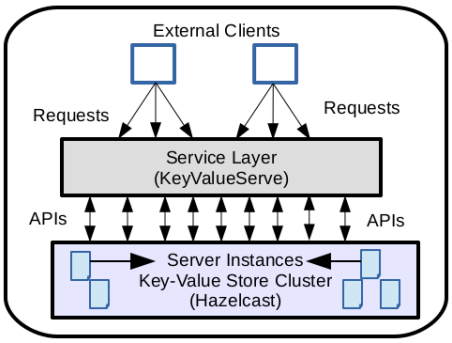


Figure 5. Service over Key-Value Store.

of these stores but also aids service resilience without a single point of failure since multiple compute-nodes are enabled to perform a specific service.

- *External Client*: Any node that connects to any Hazelcast instance in the cluster not being part of the cluster itself is a client. There may or may not be a dedicated client node. (Hazelcast offers different client interfaces where a client performs computations without itself being a part of the cluster).
- *Tenant Interface*: External tenants provide input specifications and other requests through the exposed tenant interface to one of the compute-data nodes or the external client, which in turn provides the necessary information to the cluster. Data nodes alone cannot suffice for this interface as services reside on the compute-data nodes.
- *Service Resilience*: Distribution of services across multiple compute-data nodes to avoid single points of failure and delegation of equal responsibilities to compute-data nodes instead of over-burdening a single node makes KeyValueServe resilient. The inherent peer-to-peer model as mentioned in Section 2.1 helps in restoring service to another available data node when a compute-data node goes down. For a stateful service, storing service objects in distributed stores for future restoration on node failure is a legitimate way to revert back to the same juncture of the service when failures occur. Although fair distribution is expected, making all the data nodes do some computation may not be a good idea. This is because cluster size will be large in any data center compared to the distinct services offered, and services starting or stopping on a node when nodes keep leaving or joining a dynamic cluster is a problem. Hence, only a subset of the cluster should be used as compute-data nodes, as shown in Figure 3. The fraction of compute-data nodes in a cluster is a function of the cluster size and the number of functionalities and features offered by the service. In other words, resilience is enabled by service restoration on another peer, when a specific peer goes down, both configured with the same functionality. As most NoSQL stores have a peer-to-peer model, this service resilience can be applicable to most stores, preventing complete service failures due to the node failures.

The next section discusses the major models of the service distinguishing KeyValueServe from other existing services.

2.3. Design

The goal of *KeyValueServe* as shown in Figure 4 is to offer additional valuable services in accordance with the normal requirements of a data center (e.g., availability, performance, scalability). Figure 5 illustrates that the service layer sits on top of the key-value store and uses the exposed store APIs to implement the features of the service and to access the store. It aims to enable tenants to efficiently utilize the key-value store with a fine-grained shared storage model, where the degree of multi-tenancy is high. This brings in the following design considerations as part of the service as shown in Figure 6:

DOI: 10.1002/cpe

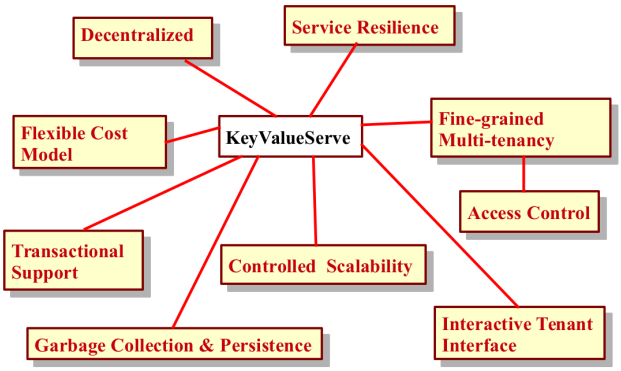


Figure 6. Salient Service Features.

- *Multi-Tenancy Model*: How are multiple tenants going to access the shared storage? Maintaining *tenant-ids* pertaining to specific clients that use the key-value store will enable *tenant authentication*. Access permissions are granted based on this unique tenant identifier. Authentication and controlled access privileges together enhance multi-tenant sharing of distributed storage enabling opportunities of key-level concurrency. In other words, key-value entries in the distributed data structures such as map can be shared by multiple tenants. However, ensuring explicit unlocking of keys for exclusive operations is important to avoid tenant starvation at such an increased level of sharing.
- *Transactional Model*: Transactional service is required for updates in a concurrent distributed system (such as additions done on maps, queues etc.). Once a tenant is authorized, it can rollback or commit a transaction on its data. This is a feature where users need not worry about the "under the hood" implementation specifics for transaction operations.
- *Garbage Collection*: In addition to default garbage collection, the service keeps track of data freshness based on recency of access and tenant activity. Storage flushing is performed based on infrequent data access during tenant inactivity. If users don't access their data for a considerable period of time (configurable parameter as a threshold) then that data gets stale because no one accessed it for the specified amount of time. It is better to replace that data with recent data accessed by active tenants. Active tenants are those who have not been inactive in terms of reading/writing/updating their data in the recent past. This feature considers both recency and frequency to replace or refill storage. This is *different* from built-in JVM garbage collection, which handles automatic memory deallocation for Java objects. During flushing, KeyValueServe either spills to the disk for *persistence* needed for future use or deletes the user's data forever. This makes room for fresh data in similar *in-memory* (non-disk-based) stores where insufficient RAM is often a performance bottleneck. This customized garbage collection scheme also eliminates the need to start additional nodes in the cluster for storing a tenant's data when the overall cluster memory size gets insufficient. KeyValueServe uses the `time-to-live` field of *map* configuration, which enables eviction of a map entry after the specified time in seconds has elapsed. This addresses the challenge of biased eviction of a tenant's keys without considering how active the tenant is. This feature ensures that an active tenant's keys are retained consciously and less active tenant's keys are evicted, freeing up space. This enables better performance for frequent users.
- *Persistence*: Interfaces such as *MapLoad* and *MapStore* are exposed for data persistence. Persistence avoids the risk of data loss or corruption, which arises when relying solely on in-memory storage, by storing data on the disk. All tenants are paying for their data. Now the question is *when does it make sense to store data persistently?* When the data structures reach their full capacity they can be stored in an external database. In addition to that there can be three circumstances that require for persistence:
  - When crucial/sensitive data needs to be retained, which cannot be lost at any cost. In this case we need the consent of the tenant.

DOI: 10.1002/cpe

  - When data is rarely accessed, in-memory storage is not a necessity. That space can be utilized for some other tenant's data. Based on access frequency, data is spilled to disk for effective memory utilization.
  - When tenants are willing to *pay more* for persistent storage exclusively for their data. This will be a rare situation since the idea behind such stores is good in-memory sharing.

  KeyValueServe follows the above persistence model to enable flexible tenant data storage and retention.
- *Cost Model*: In the cloud pay-as-you-go model, cost is based on the amount of resources consumed, the duration of usage and the granularity of access privileges provided to a tenant. The last factor is hard to define since more freedom implies more cost, and the definition of freedom may be different for different tenants. Our cost model charges the tenants based on their *amount of resource consumption* and *duration* of usage. During peak hours both *amount* and *duration* are considered as opposed to off-peak hours when only duration is considered. This is a weighted average over multiple resources. However, if a tenant expects to use a VM or map dedicated to itself, through reservation, he needs to pay more since it is free from sharing and consistency issues. Furthermore, VMs are charged higher than maps since the more coarse-grained resource or entity sharing is (i.e., more resource allocation), the higher is the cost. To retain reservations over a period of time, *nominal charges* are made. To summarize, the amount of resources used, the duration of usage, the sharing level (shared versus dedicated) and the granularity (VM versus physical node versus distributed map) govern this cost model. No upfront costs or long term commitments are required. Illustrative example: Say *memory* is charged \$0.065 per hour and \$0.08 per 500MB and a user uses 1GB of memory for 3 hours. As per this cost model, the *peak* hour cost is $(2 * 0.08) + (3 * 0.065) = \$0.355$ and the *off-peak* hour cost is $(3 * 0.065) = \$0.195$.
- *Controlled Scalability*: Increasing the number of cluster instances increases network congestion and overall resource consumption in the cluster. KeyValueServe aims to multiplex cluster resources as efficiently as possible. This generates the need for a threshold of resource consumption based on which a number of new instances are started or existing instances are terminated. The compute-data nodes collect resource usage statistics from the cluster instances and check for violation of this threshold. As an example, let the threshold be defined in terms of the CPU consumption to be maintained; suppose this threshold is between 50% and 90%. This implies that, if the overall CPU consumption of the cluster goes beyond 90%, a new server instance needs to be created for handling tenant requests. However, if the overall cluster load is below 50%, some instances can be shutdown. This is applicable for multiple types of resources and controls the size of the cluster based on the resources consumed depending on the tenant workload. Overload violation checks based on a specified threshold trigger termination or instantiation of cluster instances. Based on cluster load, the size of the cluster is dynamically varied. Consolidation of services by shutting down instances or starting new instances can help load balancing in the presence of fluctuating workload conditions. This works well in this data grid since data re-distribution is taken care of by Hazelast on-the-fly.

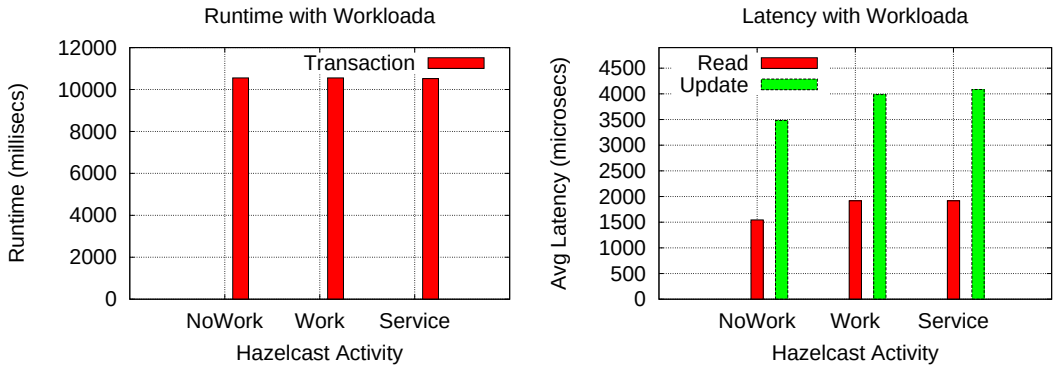
To put it into perspective, a multi-tenancy model in conjunction with garbage collection addresses the challenges of tenant starvation and isolation in fine-grained sharing. The challenge of network-induced congestion and performance optimization of multiple tenants is handled in Section 3.7 to enhance client throughput. Existing services may have considered a subset of the above mentioned features but a holistic service catering to similar functionalities has not been previously presented to the best of our knowledge.

2.4. Implementation

A prototype of *KeyValueServe* was built using Hazelcast version 3.3. Python along with bash scripts have been used for coding the service layer. Various APIs and interfaces were used to implement the service (e.g., utilities such as `forceUnlock` and `TransactionContext` have been used to prevent tenant starvation and enable transactional service support). We deliberately focus on our

DOI: 10.1002/cpe

Figure 7. Scalability Evaluation (Runtime and Latency).



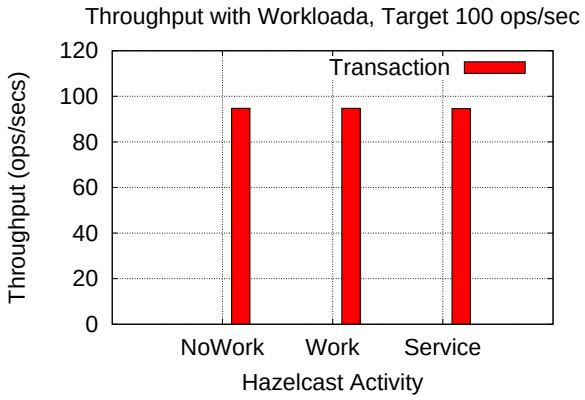


Figure 8. Scalability Evaluation (Throughput).

contribution in designing the key service models in the context of a modern cloud service instead of detailing how the built-in features of the key-value store are leveraged. The latter is less important and will vary based on the underlying store used. Our experiments indicated no additional service layer overhead.

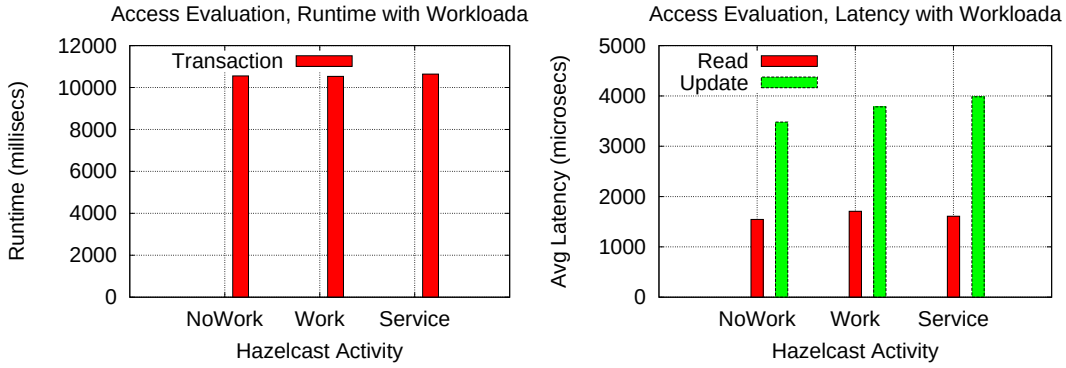
2.5. *Evaluation*

Results in this section show that the KeyValueServe service layer has a negligible impact on performance. The YCSB [32] benchmarking tool was used to generate load on the system. Experiments were conducted on Fedora and CentOS-based VMs hosted on the IBM RC2 (Research Compute Cloud) cluster [33], a cloud computing platform used by IBM Research. As mentioned in Section 1.2.2, both per client throughput and system throughput have been assessed in the results. Parameters such as the number of clients and the workload type used in the experiments, if varied, have been identified. The environments do not change in any other way. We have observed that, even if the experiments are repeated, the throughput numbers do not vary much when the system configuration parameters are kept constant. This indicates a stable, well warmed up system.

The overall *runtime*, *throughput* and *average latency* are shown in Figures 7, 8, 9, 10, 11 and 12. *Workloada* consists of 50% get and 50% put requests for objects of size 100000. 10 threads with 100 ops/sec as the target throughput are used for the experimental runs. A target of 10 ops/sec means each YCSB client aims to perform 10 operations per second. YCSB allows us to set target throughput using the `target` parameter. This option determines the number of operations per second based on the supplied value of `target` throughput. We use this `target` parameter to enforce target throughput in our evaluation; accordingly, ops/sec gets throttled. Each YCSB client can function as a single worker. Specifying 10 threads enables 9 additional workers for each client.

DOI: 10.1002/cpe

Figure 9. Map Access Evaluation (Runtime and Latency).



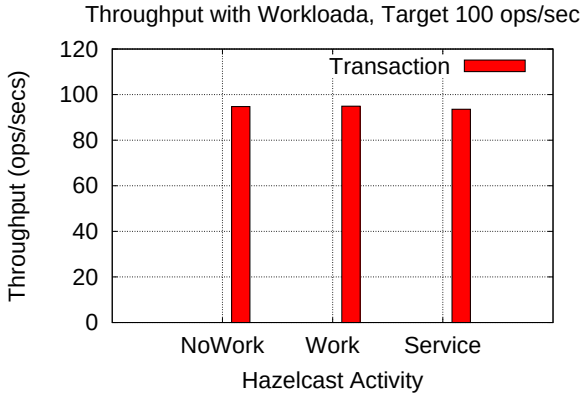


Figure 10. Map Access Evaluation (Throughput).

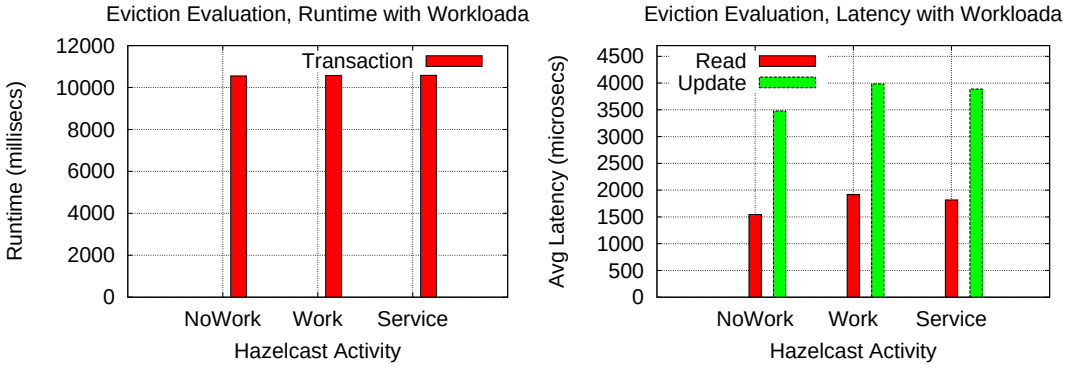
This increases the load offered. A combination of target throughput and number of client threads is used to fine-tune the workload. The load phase of the workload inserts data into the store. In Figures 7, 8, 9, 10, 11 and 12, the *transaction* phase throughput is shown, which indicates the operations executed on that inserted data such as read, write etc. These figures refer to the client side *read* and *update latencies* (perceived at the YCSB client) using `Workloada` not to be confused with the additional activity (Section 2.5.1) performed by the store/service such as spawning new instances, map updates and evictions. The Y axis in these figures pertaining to Hazelcast activity refers to the following:

- NoWork - This refers to the naive case where a Hazelcast instance is up and running doing nothing as part of the activity. It serves requests of the clients internally but there are no additional activities going on as part of the cluster or service. (no explicit action)
- Work - This case has some additional activity (see Section 2.5.1) going on in addition to serving the requests of YCSB clients such as map read/write/delete etc. However, there is no service layer; this implies the absence of KeyValueServe. The activity is performed by the Hazelcast cluster directly. (Java APIs)
- Service - This refers to the presence of KeyValueServe performing the activities. The service performs the required activity using Hazelcast APIs in addition to serving the client requests. (Python, Java API)

  To summarize, the task of serving client requests is common in all the cases. The first "NoWork" case has no additional activity going on, the second "Work" case does some additional work (see Section 2.5.1) and carries it out *directly* but without the service layer, and the third "Service" case does the same work (see Section 2.5.1 for fair comparison) *indirectly*.

Prepared using cpeauth.cls
DOI: 10.1002/cpe

Figure 11. Map Eviction Evaluation (Runtime and Latency).



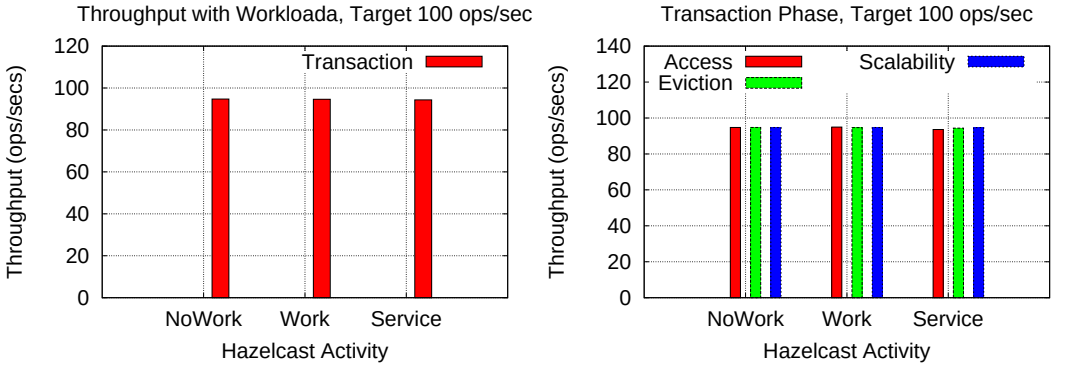


Figure 12. Map Eviction (Throughput).

Figure 13. Transaction Phase Throughput.

It uses the Service layer to perform the activity via the Hazelcast APIs. This additional activity is described below in Section 2.5.1.

2.5.1. Fair Comparison: Three cases of activity performed by the cluster or service have been considered in the experiments:

- *Scalability*: By scalability we refer to the feature of controlled scalability as mentioned in Section 2.3. The service can either spawn new cluster instances or shut them down dynamically for consolidation based on the cluster load. In this case, experiments were conducted to see the overhead of starting new instances. The objective was to check if those additional instances created during the experimental run caused any significant change in performance. The operation being performed here is instantiating new instances. The client requests are served by default.
- *Map access*: In addition to serving the client requests, the distributed map is being accessed for updates. Read operations are performed during experiments.
- *Map Eviction*: In addition to serving the client requests, map entry eviction is performed by the service. Deletion operations are performed during experiments.

The goal was to evaluate perceived response at the client end with additional activity or service related functionality happening in parallel on the cluster. We wanted to check the impact on performance both with and without the service layer and with and without any additional cluster or service activity. The service layer is tightly coupled with the key-value store and leverages its exposed APIs. From Figures 8, 10 and 12 we observe that near target throughput is achieved for all the three activities. For a target of 100 ops/sec, we procure around 95 ops/sec. There is a slight reduction in throughput from the target, because of locking and unlocking operations on

DOI: 10.1002/cpe

data structures. Moreover, for each activity, across the three cases of NoWork, Work and Service, the disparity is not much indicating that the service layer is not impeding the performance. From Figures 7, 9 and 11 we observe that the overall latency increases slightly with a negligible drop in throughput for the *Work* and *Service* cases, which is due to the distributed map operations being done as opposed to the naive case *NoWork* when no operation is performed. Figure 13 clearly shows the transaction phase throughput, where the perceived throughput is seen close to 100 ops/sec for all the three cases. There is no overhead imposed by the service layer in particular since we do not see any performance degradation from *Work* and *Service* cases. Thus, the KeyValueServe service layer is not contributing to any performance degradation.

2.6. Discussion

The design considerations of multi-tenancy, persistence, controlled scalability and flexible cost model are the core strengths of *KeyValueServe*. Such a service model is apt for high volume data intensive computations with performance guarantees. Both efficient resource multiplexing as well as quality of service can be achieved through such a service design.

2.6.1. Implications of Disk-based Access: Our work focuses on in-memory data grids since memory is frequently a bottleneck in low-latency fast-access stores. The premise of this paper is cached data in RAM. Hence, we propose ideas of garbage collection to store recent, frequently accessed data. Key-value stores offer APIs to support disk-based access. Disk-based access is one possibility but it will increase response time and definitely lower the overall throughput. Instead, we utilize a multi-tenant performance *optimization* that depends on channel reuse (see Section 3.7), and is orthogonal to ram and disk-based access conflicts. Our customized garbage collection feature will function correctly even if, most of the keys are not cached and tenant's requests are mostly fetched from the disk. Certain other features such as cost model and persistence will remain unaffected. It will be difficult to ensure fine-grained disk access, since concurrent accesses are more expensive for disks. It should be noted that, unlike prior works such as Argus [34] and Libra [35], KeyValueServe does not focus on disk-I/O cost models or deal with disk or cache reservations and scheduling.

2.6.2. Extension to other Key-Value Stores: We stress that KeyValueServe is *key-value store agnostic*, i.e., it is not Hazelcast specific. However, KeyValueServe is not independent of the key-value store used. It definitely needs some key-value store. In other words, the service is compatible with most contemporary key-value stores such as Memcached, Redis etc. Ideas of controlled scalability, fined grained multi-tenancy, persistence model etc. can certainly be adapted on other key-value stores, provided they are compatible in terms of having a decentralized, peer-to-peer model. Today, most key-value stores, if not all, fulfill this requirement. Needless to mention, necessary customizations and implementation specific tailoring is required based on which key-value store is used.

3. CONTENTION DETECTION AND PERFORMANCE OPTIMIZATION

In this section, Hazelcast is evaluated to determine and measure the presence of contention leading to performance degradation with increasing numbers of clients. This section illustrates a drop in per client throughput and presents a novel way to optimize performance in the presence of increasing numbers of clients.

3.1. Evaluation Methodology

Experiments were conducted on a local cluster, where each cluster node is equipped with a quad-core Xeon 2.53GHz CPU and 8 GB memory connected to a Gigabit network switch. Each host ran Ubuntu 12.04 64-bit and RedHat 64 bit with KVM 0.9.8 and KVM 0.10.0, respectively. The guest VMs run Ubuntu 12.04 32-bit and are configured with two virtual CPUs and 4 GB memory. An 8 instance Hazelcast cluster is set up across 2 hosts and 4 VMs as shown in Figure 14. *JVM*

DOI: 10.1002/cpe

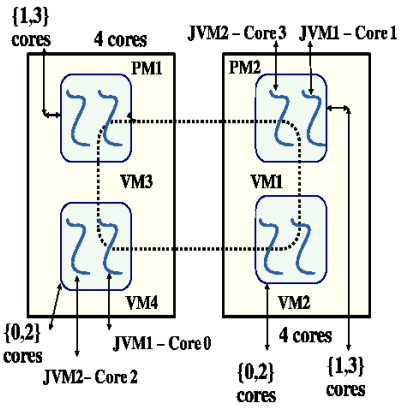


Figure 14. Experimental Set-up.

Figure 15. Performance Drop.

Figure 16. JVM-VCPU Pinned vs. Unpinned.

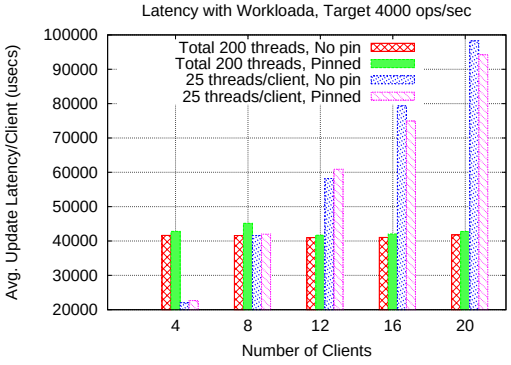


Figure 17. Update Latency Pinned vs. Unpinned.

and *Hazelcast server instance* are used interchangeably to indicate a Hazelcast cluster instance henceforth. A separate host outside the cluster ran multiple instances of the YCSB [32] client to create multi-tenant workloads. The client host did not run any other application and is not a bottleneck in any of the experiments since it did not interfere with the cluster resources. The observed performance is described in terms of overall throughput in *ops/sec*.

3.2. Performance Degradation

To understand client performance, experiments are conducted on the 8 instance cluster set-up. Figures 15 and 16 clearly indicate that throughput decreases and latencies increase with more clients. This is expected as the overall cluster load increases and the system saturates after a point. Investigations regarding what causes this performance drop were carried out across two major directions: a) JVM-VCPU pinning (Section 3.3) to check whether context switches or migrations of threads across cores affect performance and b) Multiplexing (Section 3.7) client channels to reduce the number of runnable instances per client. These are described below.

3.3. JVM-VCPU Pinning

This section observes the effects of pinning a JVM to a specific physical core. The idea is to prevent thread migration across cores, to increase cache locality, and to reduce overall context switch overhead, which might arise due to *contention*. Figure 14 depicts a two-level pinning, where every JVM has access to two physical cores. A symmetric set-up was deliberately designed to avoid any unnecessary bias in resource allocation for a JVM. 8 JVMs were running on 4 VMs, two on each. Each VM was assigned 2 vcpus. Each vcpu was pinned to 2 physical cores. Each JVM was pinned to 1 vcpu. *taskset*, a Linux utility that internally uses *sched_setaffinity*, was used for pinning, along with changes in the VM configuration file. The number of clients was varied from 4 to 20.

DOI: 10.1002/cpe

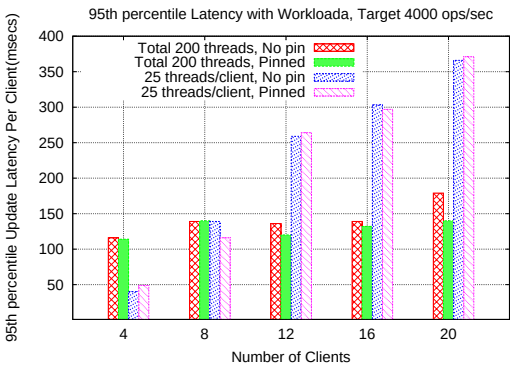


Figure 18. 95th percentile latency plot.

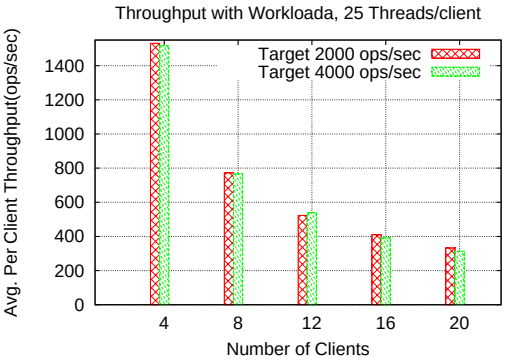


Figure 19. Client variation with different target throughput.

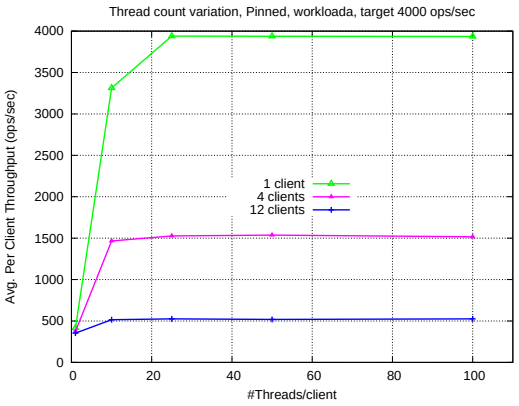


Figure 20. Client side thread variation.

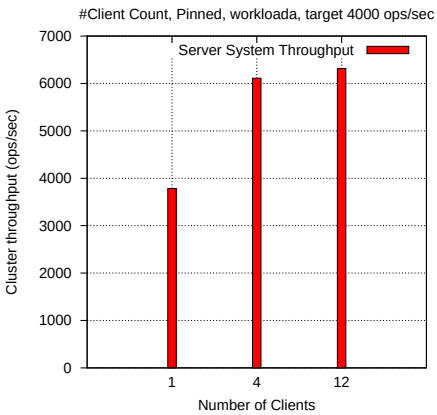


Figure 21. Cluster Throughput.

Two sets of experiments were conducted: a) every YCSB client was started with 25 threads, b) the overall thread count in the system from the client's perspective remained fixed. Clients were distributed equally among the 8 server instances to avoid unevenness. In the former case, with an increase in clients there is an increase in the total number of threads used to create the workload (4 clients, 25 threads each, for a total of 100 threads, 8 clients, 25 threads each, for a total of 200 threads). In the latter case, irrespective of the number of clients, the total number of threads is fixed at 200. In other words, for the latter case, the number of threads per client was varied to generate the workload based on the number of clients. This thread count refers to the *number of client threads*. The overall number of system-level threads created by the Hazelcast cluster remained fixed at all times from the server's perspective. The goal was to check if the client thread count had any significant impact on performance. In the load phase, `Workloada` was used consisting of 50% gets and 50% puts. In the transaction phase, a *Zipfian* distribution was used setting the *target* (expected per client throughput) to 4000 ops/sec for all experimental runs.

As seen in Figure 16, there is no change in performance with pinning. Even a thread count variation does not affect performance. Hence, thread migration context switches do not contribute *significantly* to overhead. Figure 17 shows an increase in update latency with an increase in number of clients with a *larger* thread count. However, the average update latency perceived by each client does not deviate much if the overall number of threads generated by the workload remains fixed. The same trend is observed in Figure 18 for the 95th percentile latency illustrating the fact that 95% of the operations completed within the indicated latency on the Y-Axis. This indicates that *when the workload is well distributed across the clients by keeping the overall client thread count constant, there is less variation of average response time*. Every client connects to one instance and delegates its requests to that member. Increasing the number of threads per client further increases parallelization, which increases overall latency. Figures 16, 17, and 18 clearly indicate that *pinning*

DOI: 10.1002/cpe

Figure 22. Hashed & Zipfian vs. Ordered & Uniform distribution.

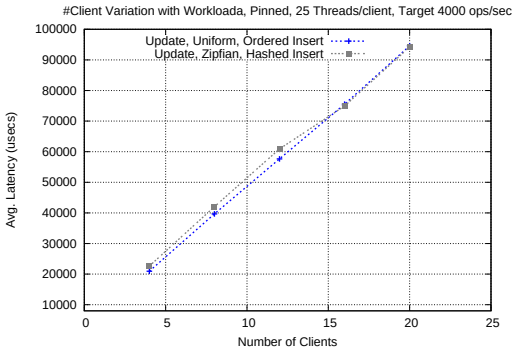


Figure 23. Hashed & Zipfian vs. Ordered & Uniform distribution.

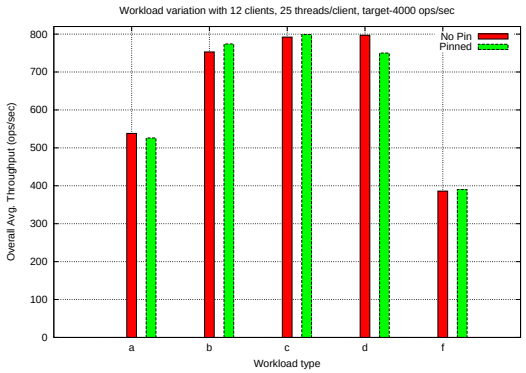


Figure 24. Workload Type Variation.

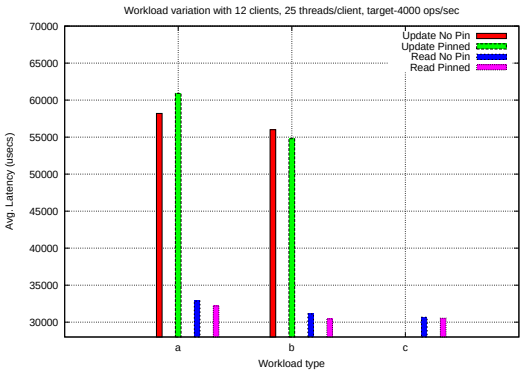


Figure 25. Latency Variation with Diverse Workloads.

does not help in mitigating contention effects. However, a well distributed workload across multiple clients improves the perceived response time.

Figure 19 illustrates that when demand (*target throughput per client*) is doubled, the performance does not degrade further, which means the system is already serving at its full capacity. Demanding more will not improve per client throughput. Once the system reaches saturation, increasing demands do not affect throughput; no further degradation is seen.

3.4. System Throughput

This section discusses performance behavior of the overall server system. Figure 20 illustrates that increasing the number of clients degrades per client throughput. The overall system throughput considering all clients from the cluster's perspective never exceeds 6300 ops/sec. When using 25 threads/client, which is our system threshold, the aggregate throughput of all clients reaches as high as 6300 ops/sec as shown in Figure 21.

Considering the limited scale of our experiments, beyond 25 threads/client there is no performance improvement. With increasing clients, contention increases and per client throughput decreases. However, increasing the number of client side threads *does not* degrade performance.

3.5. Insert Order and Distribution Type

The YCSB load phase uses hashed inserts by default where keys get hashed to specific slots in the database. In case of ordered inserts, the keys get inserted based on the order of keys. Experiments were conducted with ordered inserts and uniform distribution to see if the way keys are inserted and retrieved impacts performance. Intuitively, it depends on the Hazelcast instance location where keys get stored based on the hashing algorithm (see Section 2.1.)

DOI: 10.1002/cpe

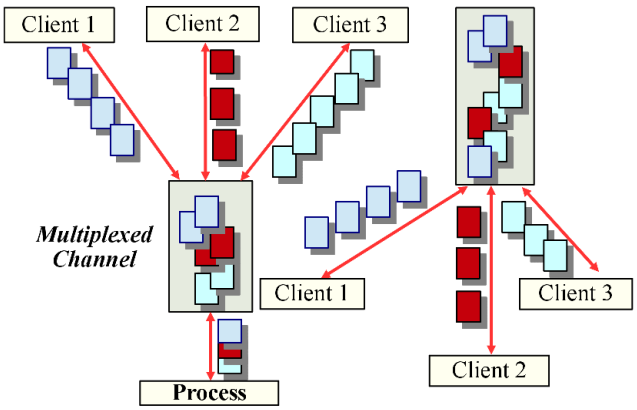


Figure 26. Multiplexing Channels in Hazelcast.

Figure 22 shows that hashed inserts are as good as ordered inserts. We do not see any tangible performance variation whether the transaction phase follows a zipfian or uniform distribution in Figure 23. This implies that *if all the clients have a similar transaction pattern, concurrent data accesses invariably cause contention no matter where data is stored.*

3.6. Workload-Type Variation

Most experiments were conducted with `Workloada` consisting of 50% gets and 50% puts. Two experiments were conducted with other workload types to observe characteristic behavior. While `workloadc` is read only, `workloadb` has 95% reads and 5% updates. `Workloadd` with 95% reads and 5% inserts and `Workloadf` with 50% reads and 50% *read-modify-writes* were also used in one experiment. Figure 24 confirms that pinning has no impact on throughput for different workload types. The larger the fraction of updates, the lower is the throughput. However, *read-modify-writes* in `Workloadf` are costlier than updates. Figure 25 shows that updates take more than twice the time than reads.

Prior work (Atikoglu et al. [22]) mentions that reads are more popular than writes in such stores. Even read throughput (shared lock over exclusive lock used in writes/updates) is considerably throttled due to contention. Improving operational latency is indeed important for better tenant service.

3.7. Multiplexing Client Channels

We successfully detect contention and throughput degradation with increasing numbers of clients in Section 3.2. We also found that thread migrations and context switches are not causing the performance drop in Section 3.3. For multi-tenancy, concurrent connections use the network more aggressively. This motivates our next question: If we reduce the number of parallel simultaneous connections, and feed the requests with fewer channels, will the per client throughput decrease the same way or will it improve? Is it I/O contention? How should we handle I/O to reduce contention? This inspired the idea of ensemble data stream processing that we shall discuss in this section.

We describe our observation with modifications to Hazelcast. We studied and then modified the *Java* source code in an effort to pipeline multiple client requests through a single proxy entry point. Our observations indicate a performance improvement with increasing number of clients. Prior experiments confirmed that keeping the number of system-level threads created by the Hazelcast cluster fixed while increasing clients invariably degrades throughput. We wanted to see if performance is affected by reducing the number of concurrent active connections through multiplexing multiple client connections. This reduces the number of runnable I/O threads needed to handle the client threads. Our study indicates a performance improvement over non-multiplexed connections.

DOI: 10.1002/cpe

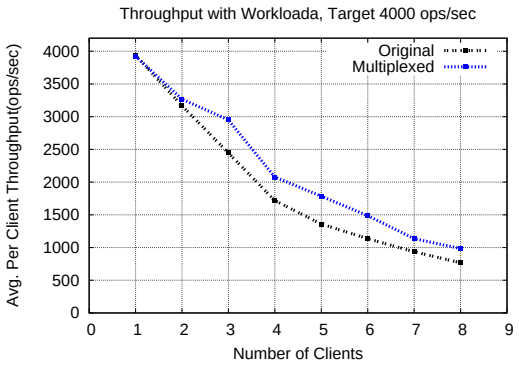


Figure 27. Multiplexed connections.

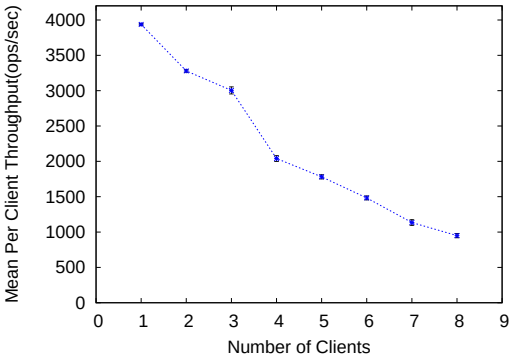


Figure 28. Throughput Mean and Std. Deviation.

As mentioned in Section 2.1, every time a client connects to the Hazelcast server instance, the socket acceptor thread establishes a connection. Data is read from the client channels and processed for further read or write operations through different handlers. When requests of multiple clients are coalesced into a single channel before processing the read or write requests, there is a small initial delay to start with. The idea is to form an ensemble of "n" data streams forming a long-lived connection rather than multiple independent short-lived data streams. Multiple concurrent connections exacerbate network congestion compared with a single multiplexed channel, resulting in contention. Every YCSB client generates requests in close temporal proximity, which further increases the number of connections at the Hazelcast cluster, thereby increasing overhead. Using our optimized approach, as the number of clients increases, the number of runnable instances created decreases substantially, with the coalesced channel not only making up for the initial time lag but improving the overall throughput as seen in Figure 27. A single long-lived connection is less aggressive in the limited network bandwidth compared to multiple parallel connections, weakening the adverse effects of network congestion.

Using fixed size thread pools to limit additional thread creation for task execution is a well known technique. Hazelcast can be configured with a fixed size thread pool for the purpose of distributed task execution. However, it should be noted that such fixed sized thread pools aim to reduce excessive thread creation overhead during voluminous task execution. They do not aid in reducing the overhead of maintaining multiple active client channels right after the clients establish connections with the cluster. In all our experiments, the default thread pool size is configured and used. Our approach tries to reduce the overall number of simultaneous active connections created when the number of clients increases, restricting the amount of congestion in the system. This should not be confused with the internal executor's threads or queues, which can be limited with the thread pool size configuration. In addition to this, in certain situations leveraging the benefits of required thread pool size for better performance may not be always possible. Results imply that throughput is well correlated to the number of active client connection instances (threads); a lower number of connections indicates better performance. This observation conforms to the claims made in the TCP analysis work by Balakrishnan et al. [36].

The `netty` [37] library was used to multiplex multiple client connections with changes in Hazelcast's `nio`-based connection implementation. The contents of multiple registered clients (inbound channels) are multiplexed to the contents of a single outbound channel containing the requests of multiple different clients. This single outbound channel interacting with the `nio` selector reduces the overall I/O because unlike before, the selector does not need to select between multiple channels anymore. This improves the overall balance per cluster instance. We noticed that Hazelcast created fewer threads (I/O, service, execution) than the naive implementation owing to the reduced outbound channel. Unlike before, the selector does not need to poll between multiple channels reducing the overall imbalance per instance. Figure 26 shows the modifications made by passing the contents of multiple socket channels into a single channel before processing it. Each experiment was repeated 3 times, and both mean and standard deviation are reported.

DOI: 10.1002/cpe

Figure 27 shows the improved performance with multiplexed connections with as many as 8 clients. As the number of clients increases, pipelining overhead increases and there arises a problem with buffer allocation and writing on the outbound channel. However, a marginal improvement in *per client throughput* justifies our claim that, indeed, multiple connection instances started for every client right at the outset cause contention, even though the internal data structures used are *asynchronous and non-blocking*. These parallel connections, spawning multiple additional threads along the way, hurt overall performance. Figure 28 shows the mean and standard deviation of the experiments. The standard deviation did not exceed 50, and the percentage increase in throughput is more than 10% for certain cases (see Figure 27).

Some experiments were conducted by over-stressing the system over limited scale. This is a limitation considering the fact that demanding much higher than the system's maximum threshold (which is a possibility in a real-life scenario) may not be appropriate for performance optimization evaluations. Figure 27 confirms that extrapolating the idea of limiting entry points to a cluster is a viable method to deal with contention problems compared to other solutions of changing partitioning or scheduling schemes in data stores based on tenant demand.

There are no side-effects of this approach. However, there always exist buffer size limitations. Hence, there will always be a bound on the number of outbound channels required to multiplex a specific number of inbound channels. So, appropriate selection of n (no. of client connections) and m (no. of channels) is required, through experimentations in a specific configuration. We successfully demonstrate the idea that per client throughput improves if the number of parallel concurrent connections is decreased. This holds true for higher numbers of clients as well. However, what m (multiplexed channels) fits what n (incoming client requests) depends on system buffer size, request size, the maximum acceptable system load, network bandwidth, etc. Making this technique scalable is part of our future work. Nevertheless, our results indicate that the abstraction of ensemble processing can be useful to derive a statistical multiplexing scheme by which n clients can be processed through m channels conforming to buffer size limitations.

4. RELATED WORK

There has been considerable recent research on storage services, performance isolation, and resource management. In this section, we discuss how our work distinguishes itself from the prior state-of-the-art.

Existing services such as S3 and DynamoDb [30, 28] from Amazon, and Cloudant [29] from IBM do not exhibit multi-tenancy in the truest sense at par with our definition (see Section 1.2.1). Clients use separate containers or virtual machines for their work without really sharing resources at a finer granularity, which KeyValueServe provides. Our work discusses how tenant authentication with controlled access can aid multi-tenant sharing across distributed data structures and cluster instances, distinguishing KeyValueServe from the existing storage services.

Pisces [18] enforces system-wide fair sharing and high resource utilization by performing partition placement, weight allocation, replica selection and fair queuing. Membase key-value store is evaluated. They partition based on demands, assign local weights based on global sharing, choose replica in a weight sensitive manner and prioritize dominant resource sharing to guarantee fairness. It is based on a centralized controller and requires huge modification of the key-value store, which is unwanted for a suitable cloud service. KeyValueServe supports multi-tenancy at the middleware level, which is less costly than the infrastructure level. Google's Borg [38] is a centralized scheduler, which uses priority-based round-robin scheduling through feasibility check and scoring. It has a master-slave structure where Borgmaster polls borglets offering good scale and performance through replicas, cached copies and stateless link shards. Pisces and Borg [18, 38] are intrusive centralized schedulers focusing on resource allocation and management complementing our work focusing on contention analysis.

MROrchestrator [39] is a resource allocator (typically for Hadoop map reduce jobs) agnostic of interference. Delay scheduling [40] has been used on HDFS systems. But how HFS (Hadoop's Fair

DOI: 10.1002/cpe

scheduler) alone will work for data grids remains an open question. These centralized solutions do not aim at fine-grained multi-tenant data store services.

Cake [41] proposes two level schedulers using HBase and HDFS to provide differentiated scheduling. They chunk large requests, provide different queues for batch and front end requests and enforce allocations based on (service level objective) slo-compliance and queue occupancy. Mercury [42] proposes a hybrid resource management framework that supports the full spectrum of scheduling from centralized to distributed. They regulate the knobs of execution guarantees and scheduling overhead by offloading work from centralized scheduler to auxillary set of fast schedulers making distributed decisions extending the Hadoop Yarn framework. Sparrow [43] proposes a stateless decentralized scheduler, which achieves improved performance through batch sampling and late binding strategies. Omega [44] designs a distributed multi-level scheduler focusing on scalability without any emphasis on multi-tenant performance. Mesos [45] and Yarn [46] propose two-level schedulers with offer and request-based resource managers. Apollo [47] aims at highly utilized, load balanced clusters. Hawk [48] proposes a hybrid scheduler for long and short jobs to leverage the advantages of both centralized and distributed schedulers. Hawk performs better than Sparrow [43] for short jobs under high load by leveraging the idea of work stealing. These [41, 42, 43, 44, 45, 47, 48] either propose a comprehensive scheduler infrastructure, complex and unfit for our target systems, or are not fine-grained enough focusing on Hadoop-style applications. Such decentralized solutions do not focus on ensuring consistent tenant performance such as ours.

SmartSLA [49] evaluates popular machine learning techniques such as linear regression and boosting to optimize resource allocation in an intelligent fashion. KeyValueServe does not focus on infrastructure cost or statistical analysis but on coalescing of I/O requests for better multi-tenant performance.

Gdwheel [24] and Camp [23] propose new key replacement strategies in caches to have high cache hits based on recency or frequency and re-computation costs, completely orthogonal to our objectives. MBal [20], CloudScale [50] and Scads [51] perform cluster load balancing mitigating hot-spots through key replication or data migration differing from our work. Argon [52], Walraven et al. [53], EyeQ [54], SqlVm [55], Pulsar[21], Das et al. [56], IOFlow[57], Zeng et al. [58] and PriDyn [59] focus on fairness, either at a coarser granularity dealing with virtual machines, network isolation, higher level abstraction or centralized quantum-based scheduling, in contrast to our work. Anderson et al. [60] quantify key-value store consistency through evaluation. Atikoglu et al. [22] unveil workload insights of real-life traces, which can have implications on the cache configuration. These give useful hints about patterns of real-life usage to analyze the impact of their performance under load and scale.

Wenqi et al. [61] assess key-value store performance analysis too, but they use Memcached and Redis unlike Hazelcast. They focus only on resource overhead unlike our proposal of service modeling and performance optimization schemes. Das et al. [62] proposes the channel reuse based performance optimization technique to improve client performance. KeyValueServe extends the earlier work by providing a multi-tenant cloud service model with detailed experimental evaluation pertaining to service and performance. Met [63] proposes a cloud-enabled framework to reconfigure a cluster for dynamic workload changes. Met takes homogeneity and heterogeneity of nodes into account and proposes a decision maker, which distributes data partitions based on classification and node grouping. While KeyValueServe's controlled scalability adjusts instances based on cluster load, the partitioning logic is unaltered. Different stores have unique data placement and redistribution policies, we do not want to tamper with data partitioning logic to make our model generic. KeyValueServe, unlike Met, proposes a multiplexed channel reuse-based scheme to improve throughput along with several design features, without considering node classification and group data assignments. Argus [34] proposes a workload-aware resource reservation NoSql store. Argus targets system-wide fair share like Pisces [18] and uses stochastic hill climbing to find optimal resource reservations with changing workload resource demands. Unlike Argus and Pisces, KeyValueServe does not perform elastic resource reservation but combines multiple client requests to alleviate network contention while improving throughput. Libra [35] proposes an I/O

DOI: 10.1002/cpe

scheduling framework for handling I/O interference by charging an I/O operation proportional to its actual resource usage. Unlike Libra, KeyValueServe schedules I/O requests before getting to the persistence engine. Libra is situated below the persistence engine to schedule I/O in a deficit round robin fashion. KeyValueServe leverages the idea of adjusting I/O requests based on ensemble I/O through fewer outgoing channels instead of weighted fair-share approaches. A-Cache [64] resolves cache interference for multiple workloads by tracking cache re-use ratios of different workloads. A-Cache uses cache throttling to improve throughput while KeyValueServe improves throughput by leveraging the strength of combined client requests over multiple parallel requests. The fundamental ideas are different; moreover, KeyValueServe discusses a service model with features aiding multi-tenant performance.

Existing database and key-value store services perform sharing at a much coarser granularity, leaving scopes of resources under-utilization and resulting in lower degrees of concurrent accesses. Agrawal et al. [27] discuss several forms of multi-tenancy in databases as explained in Section 1.2.1. Directing research efforts in enriching the functionality supported by key-value stores and designing scalable, elastic and autonomic multi-tenant systems is emphasized. We have taken a step in that direction. This paper discusses KeyValueServe (Section 2.3), a service which enables multi-tenancy by governing access privileges at the level of data structures, cluster instances and VMs. Prior work has investigated performance of distributed storage systems, but focusing on sources of contention and ensuring consistent client response still remains challenging. The existence of performance degradation was demonstrated (Figure 16) and a solution (Section 3.7) was shown to especially aid cloud computing infrastructures.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose KeyValueServe, a low-overhead service with its major models of multi-tenancy, persistence and cost. The Hazelcast in-memory data grid was evaluated in the context of multi-tenancy to assess its performance. Our findings unveil the following interesting insights: JVM-VCPU pinning does not help; maintaining parallel independent client connections with their own thread data structures and work queues degrades throughput. Even though internal data structures used are *asynchronous and non-blocking* intended for sharing, an increase in end-to-end separate connection instances with increasing clients causes resource contention. Our results indicate that an ensemble of multiple client connections together forming a single coalesced data stream can alleviate contention.

Future work aims to look at *novel performance optimization* techniques and overlay service delegation protocols to make KeyValueServe more resilient and performance efficient. Consideration of real cloud deployment and a thorough QoS evaluation of KeyValueServe on a larger scale with complex workloads are planned in the future. This performance study also provides a new perspective of looking at the contention problem. The best way to process data through a reduced number of data streams should be analyzed further. Another area for further investigation is comparing and integrating techniques using thread pools and optimized thread pool sizes with our approach of using a combined channel for requests from multiple clients.

ACKNOWLEDGEMENTS

The authors thank Prof. Xiaohui (Helen) Gu for her support and insightful suggestions in the initial stages of this work. Part of this work was done while Anwesha Das was a summer intern at IBM Research. This work was supported in part by NSF grants 1217748 and 0958311.

REFERENCES

1. Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE. Bigtable: A distributed storage system for structured data. *ACM Trans. Comput. Syst.* Jun 2008; **26**(2):4:1–4:26, doi: 10.1145/1365815.1365816. URL http://doi.acm.org/10.1145/1365815.1365816.

2. Cooper BF, Ramakrishnan R, Srivastava U, Silberstein A, Bohannon P, Jacobsen HA, Puz N, Weaver D, Yerneni R. Pnuts: Yahoo!'s hosted data serving platform. *Proc. VLDB Endow.* Aug 2008; **1**(2):1277–1288, doi: 10.14778/1454159.1454167. URL http://dx.doi.org/10.14778/1454159.1454167.
3. DeCandia G, Hastorun D, Jampani M, Kakulapati G, Lakshman A, Pilchin A, Sivasubramanian S, Vosshall P, Vogels W. Dynamo: Amazon's highly available key-value store. *SIGOPS Oper. Syst. Rev.* Oct 2007; **41**(6):205–220, doi:10.1145/1323293.1294281. URL http://doi.acm.org/10.1145/1323293.1294281.
4. Shvachko K, Kuang H, Radia S, Chansler R. The hadoop distributed file system. *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, 2010; 1–10, doi:10.1109/MSST.2010.5496972.
5. Hbase. https://hbase.apache.org/ 2008.
6. Hypertable. http://hypertable.org/ 2008.
7. Escriva R, Wong B, Sirer EG. Hyperdex: A distributed, searchable key-value store. *SIGCOMM Comput. Commun. Rev.* Aug 2012; **42**(4):25–36, doi:10.1145/2377677.2377681. URL http://doi.acm.org/10.1145/2377677.2377681.
8. Sumbaly R, Kreps J, Gao L, Feinberg A, Soman C, Shah S. Serving large-scale batch computed data with project voldemort. *Proceedings of the 10th USENIX Conference on File and Storage Technologies*, FAST'12, USENIX Association: Berkeley, CA, USA, 2012; 18–18. URL http://dl.acm.org/citation.cfm?id=2208461.2208479.
9. Geambasu R, Levy AA, Kohno T, Krishnamurthy A, Levy HM. Comet: An active distributed key-value store. *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation*, OSDI'10, USENIX Association: Berkeley, CA, USA, 2010; 1–13. URL http://dl.acm.org/citation.cfm?id=1924943.1924966.
10. Menon P, Rabl T, Sadoghi M, Jacobsen HA. Cassandra: An ssd boosted key-value store. *International Conference on Data Engineering*, 2014; 1162–1167, doi:10.1109/ICDE.2014.6816732.
11. Fitzpatrick B. Distributed caching with memcached. *Linux J.* Aug 2004; **2004**(124):5–.
12. Carlson JL. *Redis in Action*. Manning Publications Co.: Greenwich, CT, USA, 2013.
13. Lim H, Fan B, Andersen DG, Kaminsky M. Silt: A memory-efficient, high-performance key-value store. *Symposium on Operating Systems Principles*, 2011; 1–13.
14. Joyent cloud services. https://www.joyent.com/ 2004.
15. Riak nosql solution. http://docs.basho.com/riak/latest/ 2009.
16. Cloudant cloud service. https://cloudant.com/ 2010.
17. Couchdb nosql database. http://couchdb.apache.org/ 2005.
18. Shue D, Freedman MJ, Shaikh A. Performance isolation and fairness for multi-tenant cloud storage. *OSDI*, 2012; 349–362.
19. Nishtala R, Fugal H, Grimm S, Kwiatkowski M, Lee H, Li HC, McElroy R, Paleczny M, Peek D, Saab P, *et al.*. Scaling memcache at facebook. *NSDI*, vol. 13, 2013; 385–398.
20. Cheng Y, Gupta A, Butt AR. An in-memory object caching framework with adaptive load balancing. *Proceedings of the Tenth European Conference on Computer Systems*, ACM, 2015; 4.
21. Angel S, Ballani H, Karagiannis T, OShea G, Thereska E. End-to-end performance isolation through virtual datacenters. *Proceedings of the 11th USENIX conference on Operating Systems Design and Implementation*, USENIX Association, 2014; 233–248.
22. Atikoglu B, Xu Y, Frachtenberg E, Jiang S, Paleczny M. Workload analysis of a large-scale key-value store. *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, 2012; 53–64.
23. Ghandeharizadeh S, Irani S, Lam J, Yap J. Camp: a cost adaptive multi-queue eviction policy for key-value stores. *Proceedings of the 15th International Middleware Conference*, ACM, 2014; 289–300.
24. Li C, Cox AL. Gd-wheel: a cost-aware replacement policy for key-value stores. *Proceedings of the Tenth European Conference on Computer Systems*, ACM, 2015; 5.
25. Mongodb. https://www.mongodb.com/ 2009.
26. Amazon dynamodb. https://aws.amazon.com/dynamodb/ 2012.
27. Agrawal D, Das S, El Abbadi A. Big data and cloud computing: current state and future opportunities. *Proceedings of the 14th International Conference on Extending Database Technology*, ACM, 2011; 530–533.
28. Sivasubramanian S. Amazon dynamodb: a seamlessly scalable non-relational database service. *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, ACM, 2012; 729–730.
29. Bienko CD, Greenstein M, Holt SE, Phillips RT, *et al.*. *IBM Cloudant: Database as a Service Advanced Topics*. IBM Redbooks, 2015.
30. Palankar MR, Iamnitchi A, Ripeanu M, Garfinkel S. Amazon s3 for science grids: a viable solution? *Proceedings of the 2008 international workshop on Data-aware distributed computing*, ACM, 2008; 55–64.
31. Johns M. *Getting Started with Hazelcast*. Packt Publishing Ltd, 2015.
32. Ycsb. https://github.com/brianfrankcooper/YCSB/wiki/Getting-Started 2010.
33. Ryu KD, Zhang X, Ammons G, Bala V, Berger S, Da Silva DM, Doran J, Franco F, Karve A, Lee H, *et al.*. Rc2–a living lab for cloud computing. *Proceedings of LISA10: 24th Large Installation System Administration Conference*, 2010; 201.
34. Zeng J, Plale B. Argus: A multi-tenancy nosql store with workload-aware resource reservation. *Parallel Computing* 2016; **58**:76–89.
35. Shue D, Freedman MJ. From application requests to virtual iops: Provisioned key-value storage with libra. *Proceedings of the Ninth European Conference on Computer Systems*, ACM, 2014; 17.
36. Balakrishnan H, Padmanabhan VN, Seshan S, Stemm M, Katz RH. Tcp behavior of a busy internet server: Analysis and improvements. *INFOCOM'98. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1, IEEE, 1998; 252–262.
37. Netty. http://netty.io/ 2003.
38. Verma A, Pedrosa L, Korupolu M, Oppenheimer D, Tune E, Wilkes J. Large-scale cluster management at google with borg. *European Conference on Computer Systems*, 2015; 18.

39. Sharma B, Prabhakar R, Lim SH, Kandemir MT, Das CR. Mrorchestrator: A fine-grained resource orchestration framework for mapreduce clusters. *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*, IEEE, 2012; 1–8.
40. Zaharia M, Borthakur D, Sen Sarma J, Elmeleegy K, Shenker S, Stoica I. Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. *Proceedings of the 5th European conference on Computer systems*, ACM, 2010; 265–278.
41. Wang A, Venkataraman S, Alspaugh S, Katz R, Stoica I. Cake: enabling high-level slos on shared storage systems. *ACM Symposium on Cloud Computing*, 2012; 14.
42. Karanasos K, Rao S, Curino C, Douglas C, Chaliparambil K, Fumarola G, Heddaya S, Ramakrishnan R, Sakalanaga S. Mercury: Hybrid centralized and distributed scheduling in large shared clusters. *USENIX Annual Technical Conference*, 2015; 485–497.
43. Ousterhout K, Wendell P, Zaharia M, Stoica I. Sparrow: distributed, low latency scheduling. *Symposium on Operating Systems Principles*, 2013; 69–84.
44. Schwarzkopf M, Konwinski A, Abd-El-Malek M, Wilkes J. Omega: flexible, scalable schedulers for large compute clusters. *European Conference on Computer Systems*, 2013; 351–364.
45. Hindman B, Konwinski A, Zaharia M, Ghodsi A, Joseph AD, Katz RH, Shenker S, Stoica I. Mesos: A platform for fine-grained resource sharing in the data center. *NSDI*, 2011; 22–22.
46. Vavilapalli VK, Murthy AC, Douglas C, Agarwal S, Konar M, Evans R, Graves T, Lowe J, Shah H, Seth S, *et al.*. Apache hadoop yarn: Yet another resource negotiator. *Symposium on Cloud Computing*, 2013; 5.
47. Boutin E, Ekanayake J, Lin W, Shi B, Zhou J, Qian Z, Wu M, Zhou L. Apollo: scalable and coordinated scheduling for cloud-scale computing. *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, 2014; 285–300.
48. Delgado P, Dinu F, Kermarrec AM, Zwaenepoel W. Hawk: Hybrid datacenter scheduling. *USENIX Annual Technical Conference*, 2015; 499–510.
49. Xiong P, Chi Y, Zhu S, Moon HJ, Pu C, Hacigümüş H. Intelligent management of virtualized resources for database systems in cloud environment. *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*, IEEE, 2011; 87–98.
50. Shen Z, Subbiah S, Gu X, Wilkes J. Cloudscale: elastic resource scaling for multi-tenant cloud systems. *Proceedings of the 2nd ACM Symposium on Cloud Computing*, ACM, 2011; 5.
51. Trushkowsky B, Bodík P, Fox A, Franklin MJ, Jordan MI, Patterson DA. The scads director: Scaling a distributed storage system under stringent performance requirements. *FAST*, 2011; 163–176.
52. Wachs M, Abd-El-Malek M, Thereska E, Ganger GR. Argon: Performance insulation for shared storage servers. *FAST*, vol. 7, 2007; 5–5.
53. Walraven S, Monheim T, Truyen E, Joosen W. Towards performance isolation in multi-tenant saas applications. *Proceedings of the 7th Workshop on Middleware for Next Generation Internet Computing*, ACM, 2012; 6.
54. Jeyakumar V, Alizadeh M, Mazieres D, Prabhakar B, Kim C, Greenberg A. Eyeq: practical network performance isolation at the edge. *REM* 2013; **1005**(A1):A2.
55. Narasayya VR, Das S, Syamala M, Chandramouli B, Chaudhuri S. Sqlvm: Performance isolation in multi-tenant relational database-as-a-service. *CIDR*, 2013.
56. Das S, Narasayya VR, Li F, Syamala M. Cpu sharing techniques for performance isolation in multi-tenant relational database-as-a-service. *Proceedings of the VLDB Endowment* 2013; **7**(1).
57. Thereska E, Ballani H, O'Shea G, Karagiannis T, Rowstron A, Talpey T, Black R, Zhu T. Ioflow: A software-defined storage architecture. *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, ACM, 2013; 182–196.
58. Zeng J, Plale B. Multi-tenant fair share in nosql data stores. *Cluster Computing (CLUSTER), 2014 IEEE International Conference on*, IEEE, 2014; 176–184.
59. Jain N, Lakshmi J. Pridyn: Framework for performance specific qos in cloud storage. *Cloud Computing (CLOUD), 2014 IEEE 7th International Conference on*, IEEE, 2014; 32–39.
60. Anderson E, Li X, Shah M, Tucek J, Wylie JJ. What consistency does your key-value store actually provide. *Proceedings of the Sixth international conference on Hot topics in system dependability*, USENIX Association, 2010; 1–16.
61. Cao W, Sahin S, Liu L, Bao X. Evaluation and analysis of in-memory key-value systems. *Proceedings of the IEEE International Conference on Cloud Computing*, IEEE, 2016.
62. Das A, Mueller F, Gu X, Iyengar A. Performance analysis of a multi-tenant in-memory data grid. *Proceedings of the IEEE International Conference on Cloud Computing*, IEEE, 2016.
63. Cruz F, Maia F, Matos M, Oliveira R, Paulo J, Pereira J, Vilaça R. Met: workload aware elasticity for nosql. *Proceedings of the 8th ACM European Conference on Computer Systems*, ACM, 2013; 183–196.
64. Ravi B, Amur H, Schwan K. A-cache: Resolving cache interference for distributed storage with mixed workloads. *Cluster Computing (CLUSTER), 2013 IEEE International Conference on*, IEEE, 2013; 1–8.