ZCloud

Consensus on Hardware for Distributed Systems

ZETAOPS

WORLD IS CONCURRENT!

“Simplicity is the ultimate sophistication.”
Leonardo Da Vinci

3rd Workshop on Dependability,
May 8, Monday 2017, İYTE, https://goo.gl/fSVnZy

Gökhan Boranalp, gokhan@zetaops.io

Road Map

- Problem Definition
- Our Solution
- ZCloud Components
  - ZCloud Hardware
  - ZCloud Cluster Management Tools
  - ZCloud Protocol
- Benefits
- Similar Work
- Discussion

Problem Definition

Increasing,

- computing power,
- data storage, analysis and
- sophisticated network communication requirements

in modern “data centers”, reveals the strong need for “distributed” operation for both networking devices and applications.

Problem Definition

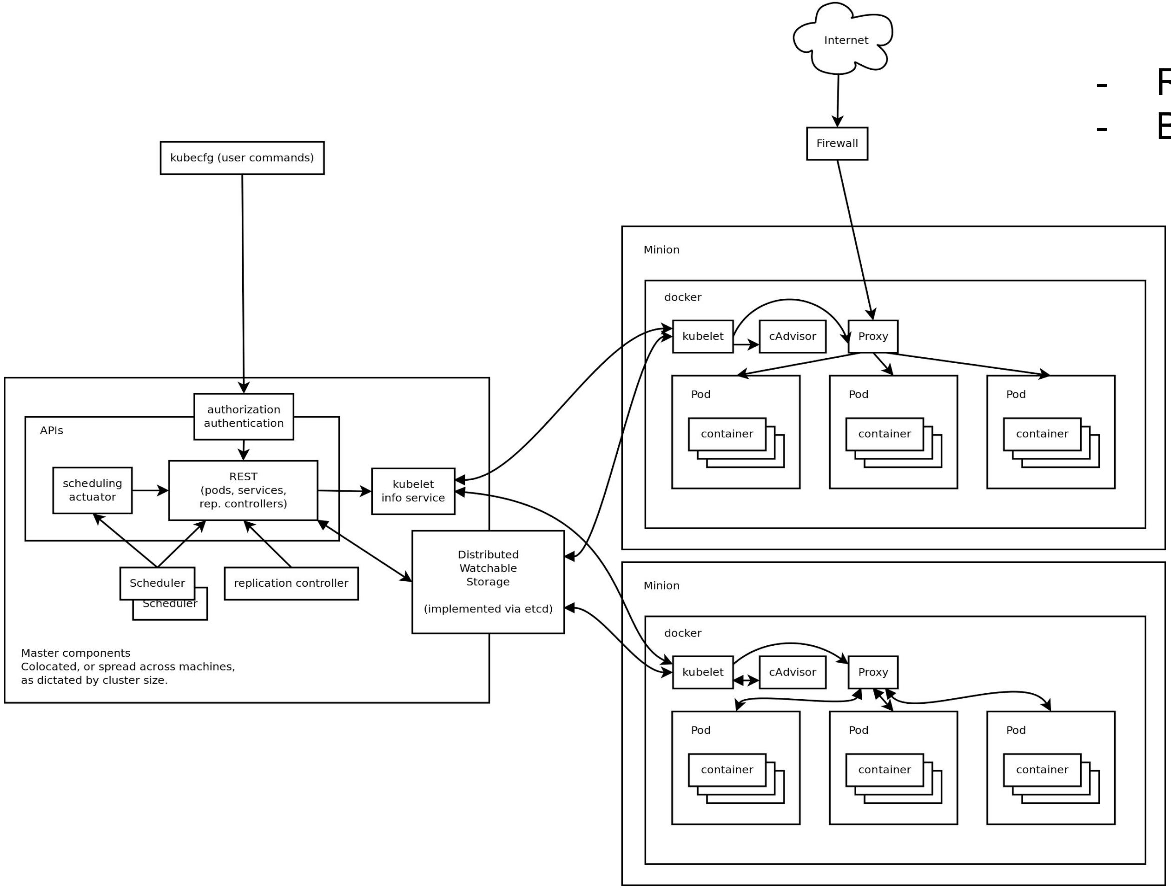
Apache Mesos and Google Kubernetes, which are using the container virtualization technique to develop, scale and manage applications in distributed systems, have emerged. These applications uses different "consensus" approaches in their internal processes.

Problem Definition

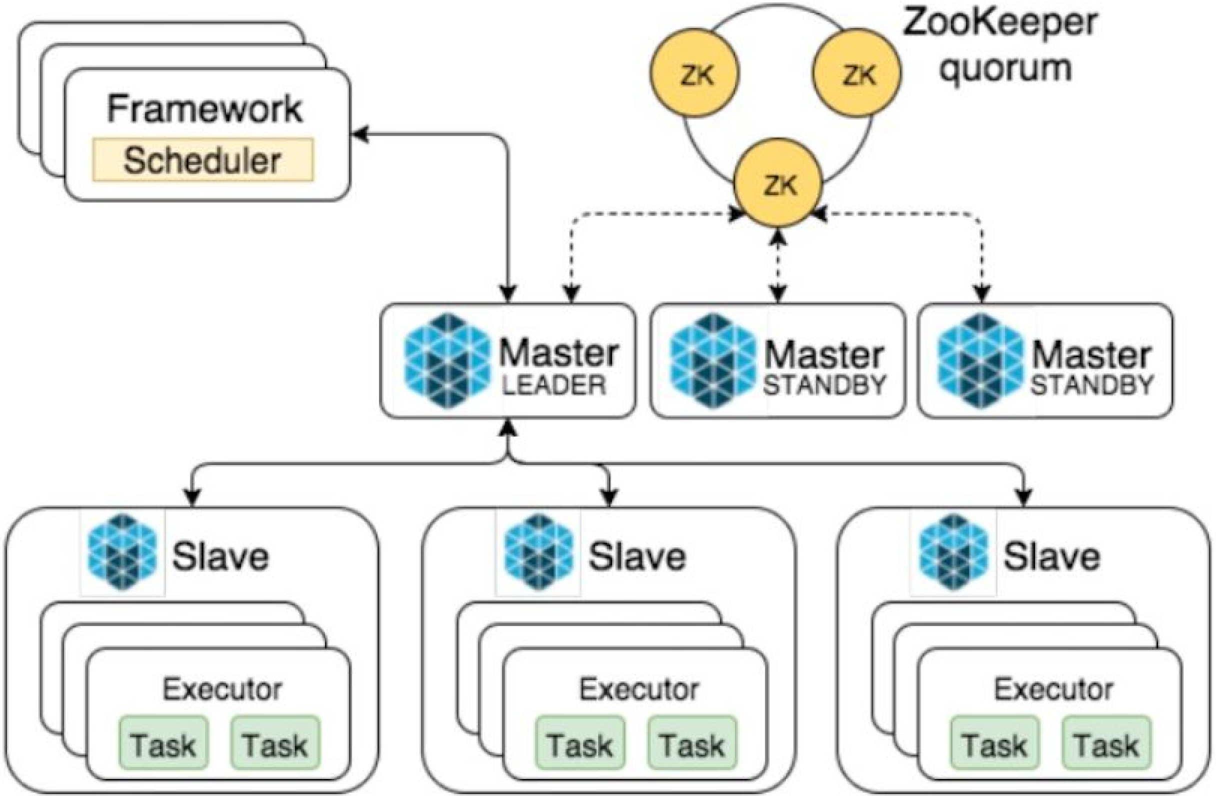
These approaches for distributed application development seem to be efficient enough for the time being, yet new approaches are needed in terms of **latency**, **number of transactions** and **throughput** in distributed systems, taking into account of physical boundaries and the increase in the size of future applications and the number of cluster members.

- With Mesos, practically 50,000 instance tests were performed on 24,000 core 500 physical servers while Kubernetes was tested on 500 physical servers.

- Raft protocol,
- Based on etcd



Kubernetes Architecture



Mesos Architecture

Problem Definition

In distributed applications where the number of members in the cluster increases, the separation of the consensus related operations at the hardware level is essential for the following reasons:

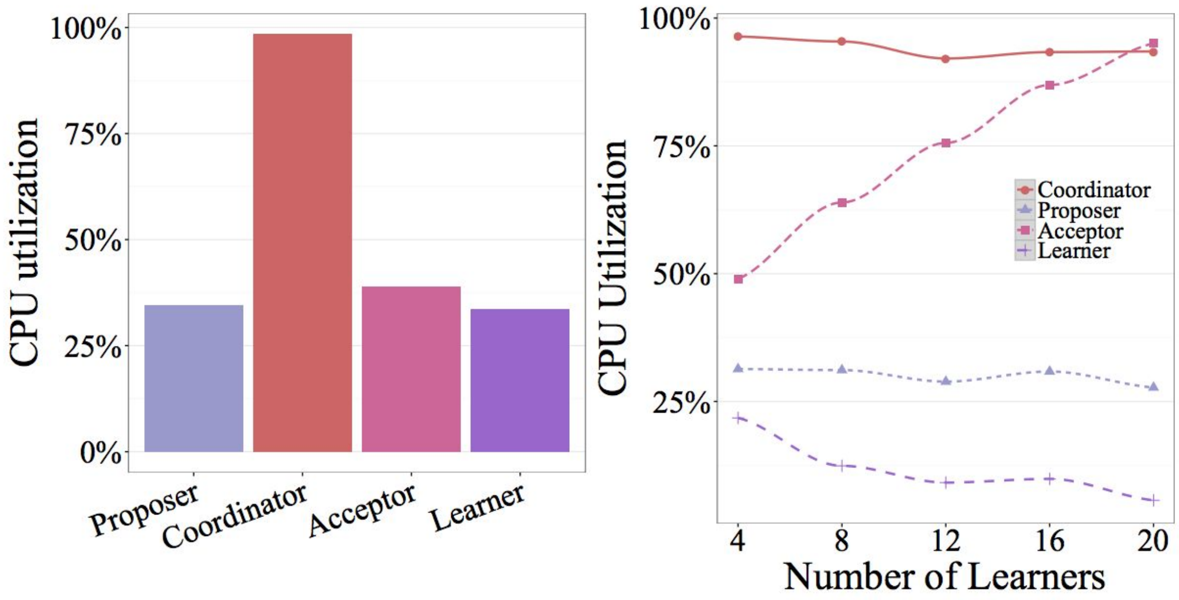
1. At the operating system level, messages broadcast on the protocol stack cause latency.

2. It is necessary to increase the number of completed transactions in the communication of distributed system components and on the network unit (throughput).

Problem Definition

3. For devices with limited storage and CPU computing facilities that use embedded operating systems such as IOT devices, it is also necessary to reduce the processing burden due to "consensus" operations.

4. A better common consensus communication model is needed for different applications that need to work together in (BFT) environment.



(a) Coordinator bottleneck

(b) Acceptor bottleneck

Figure 2: The coordinator and acceptor processes are the bottleneck in a software-based Paxos deployment. Acceptor utilization scales with the degree of replication.

Network Hardware-Accelerated Consensus, page 4, USI-INF-TR-2016-03

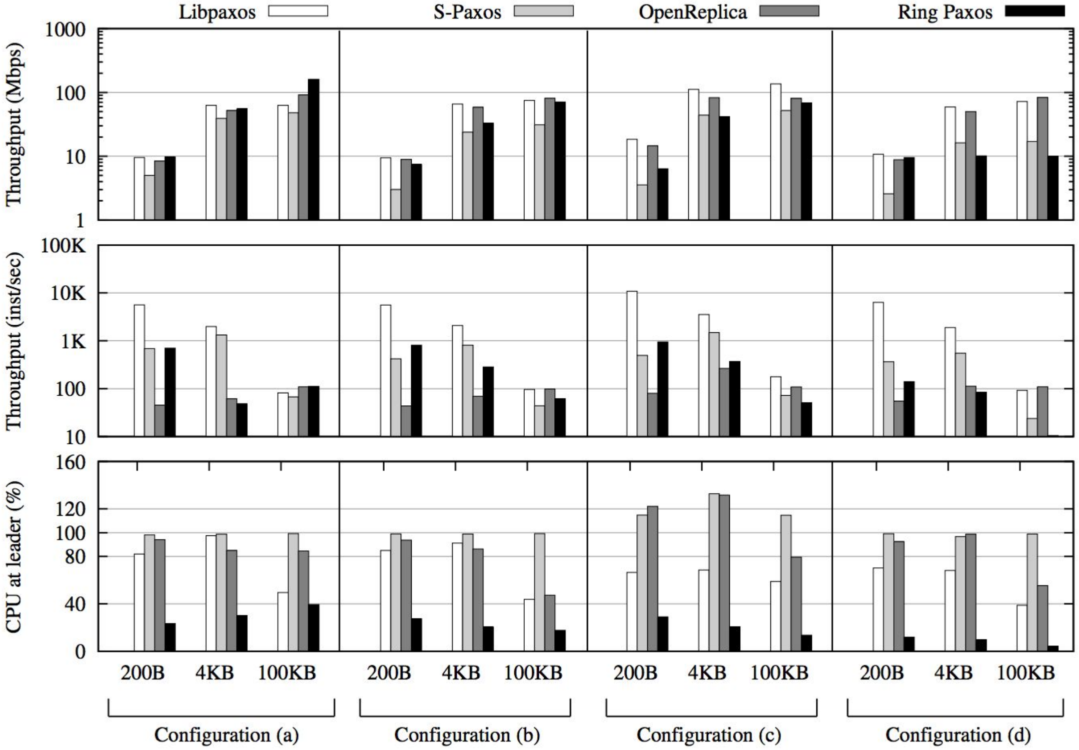
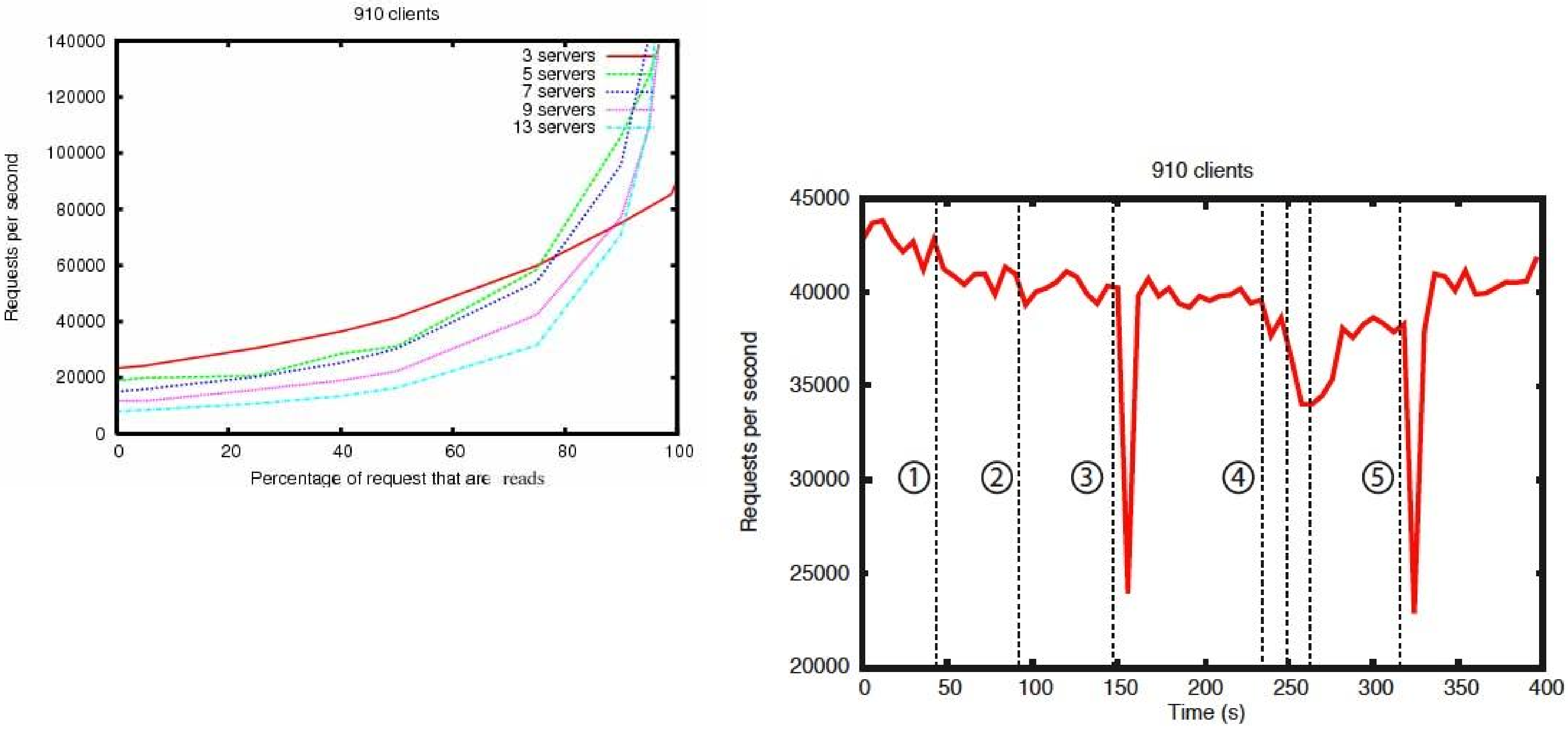


Fig. 2. Peak performance of Libpaxos, S-Paxos, Ring Paxos, and OpenReplica in four configurations (see Table I); y-axis in the two top-most graphs is in log scale; note that S-Paxos, Ring Paxos, and OpenReplica are multithreaded and therefore in some scenarios (configuration (c)) the CPU usage at the leader is higher than 100% for some of the libraries.

The Performance of Paxos in the Cloud, p. 46, DOI: 10.1109/SRDS.2014.15



ZooKeeper Throughput as the Read-Write Ratio Varies, https://goo.gl/8aR35E

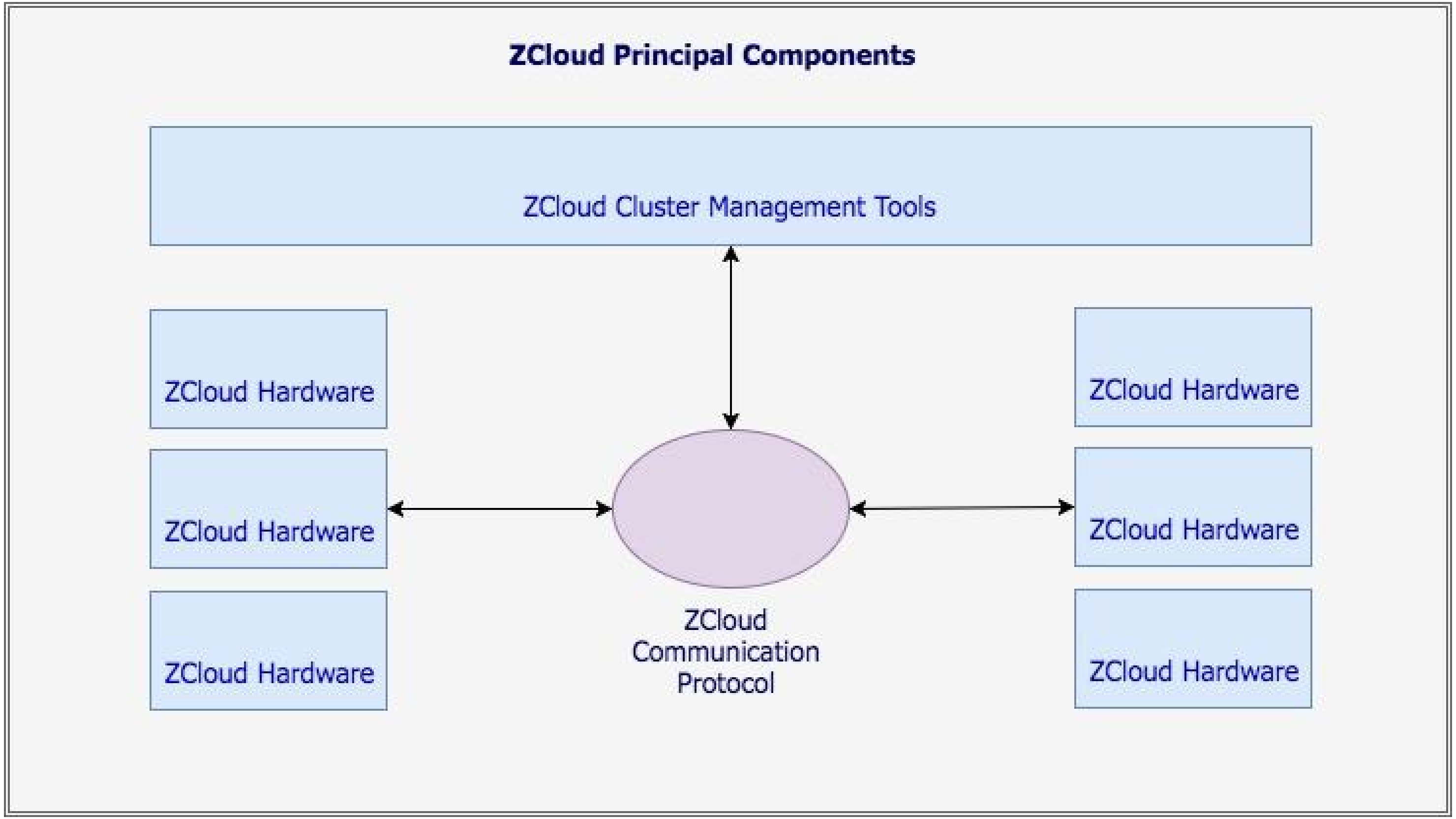
Our Solution

ZCloud is an hardware and software solution for distributed systems and conceived to provide Byzantine fault tolerance (BFT) consensus primitives.

Our Solution

ZCloud came out as a solution for today's demands such as distributed hardware and software solutions and network components in data centers, distributed software architectures with very high number of components, IOT components for distributed architecture, distributed data processing and distributed data storage.

- **ZCloud is ongoing research effort and is in early stage.**


ZCloud Principal Components
ZCloud Cluster Management Tools
ZCloud Hardware
ZCloud Hardware
ZCloud Hardware
ZCloud Hardware
ZCloud Hardware
ZCloud Hardware
ZCloud Communication Protocol

ZCloud Components - Hardware

* CPU, changeable, low energy consumption, high processor power CPU

* SoC and related hardware to manage distributed operations

* RAM, variable 4, 8, 16, 32, 64 GB, type will be determined

* 2 or 4 x Ethernet connections, changeable 40 GbE, 100 GbE ethernet

* SD card slot, Mini USB input, USB-C, EEPROM, Flash memory

* Wireless 802.11b, g, n, a Ethernet

Notes:

- The hardware components that can be used for IOT devices will be diversified during the research.

- Storage is separate and distributed in cloud, see CEPH.

ZCloud Cluster Management Tool

ZCloud Cluster Manager enables ZCloud hardware to be managed under Openstack for cloud adaptation and use with existing systems.

ZCloud SDK will be developed for use with the Nova, Neutron, Heath, Telemetry, Ironic, Manila, Magnum components found in the Openstack system.

ZCloud Cluster Management Tool

Related information and operations in Openstack Horizon panel will be available below;

- Add new hardware to the cluster,
- Remove hardware manually from cluster,
- Automatically mark and remove faulty hardware,
- Ensemble requested topologies,
- Collect logs (CPU, RAM, network load, parametric values)

ZCloud Communication Protocol

- Newly designed Paxos based protocol.
  - Masquerade agents
  - Separation of request types
  - More to come, still in development.

Some Benefits

- Drop in usage with an existing cloud operated apps.
- Ability to develop fast applications for network communicating devices.
- Ability to develop very *very* large applications.
- Reduction of electricity consumption (power consumption, cooling, etc.) in data centers.
- The possibility of accelerating the analysis results by efficiently distributing the analysis of the data derived from web applications in a large cluster.
- Prevention of energy and resource losses due to inefficient management of resource usage on the cloud.

Similar Work

- Network Hardware-Accelerated Consensus
  - CAANS provides a complete Paxos protocol, is a dropin replacement for software-based implementations of Paxos, makes no restrictions on network topologies, and is implemented in a higher-level, data-plane programming language, allowing for portability across a range of target devices.
- Consensus in a Box: Inexpensive Coordination in Hardware
  - Zookeeper's atomic broadcast at the network level using an FPGA.

Discussion

Resources

Dual-leader Master Election for Distributed Systems (Obiden), http://www.cse.scu.edu/~mwang2/projects/Distributed_dualLeaders_15s.pdf

The Performance of Paxos in the Cloud, http://sci-hub.cc/10.1109/SRDS.2014.15

The Performance of Paxos and Fast Paxos, http://www.ic.unicamp.br/~reltech/2008/08-35.pdf

Consensus in the Cloud: Paxos Systems Demystified, https://www.cse.buffalo.edu/tech-reports/2016-02.pdf

Seamless Paxos coordinators, http://sci-hub.cc/10.1007/s10586-013-0264-9

Implementing Fault-Tolerant Services Using the State Machine Approach: A Tutorial

http://www-users.cselabs.umn.edu/classes/Spring-2014/csci8980-sds/Papers/ProcessReplication/p299-schneider.pdf

Holistic Configuration Management at Facebook, http://sigops.org/sosp/sosp15/current/2015-Monterey/printable/008-tang.pdf

Optimistic Replication, http://sci-hub.cc/10.1145/1057977.1057980

Dotted Version Vectors: Logical Clocks for Optimistic Replication, https://arxiv.org/pdf/1011.5808.pdf

In Search of an Understandable Consensus Algorithm, https://web.stanford.edu/~ouster/cgi-bin/papers/raft-atc14

Fast Quantum Byzantine Agreement, https://pdfs.semanticscholar.org/73ab/ef762dd61fdd388173f24f811e8693a79d7c.pdf

Asynchronous Consensus and Broadcast Protocols, http://zoo.cs.yale.edu/classes/cs426/2013/bib/bracha85asynchronous.pdf

Customizable and Extensible Deployment for Mobile/Cloud Applications, https://sapphire.cs.washington.edu/papers/sapphire-osdi14.pdf

Ovid: A Software-Defined Distributed Systems Framework,
https://www.usenix.org/system/files/conference/hotcloud16/hotcloud16_altinbuken.pdf

Resources

Network Hardware-Accelerated Consensus, https://arxiv.org/pdf/1605.05619.pdf

Consensus in a Box: Inexpensive Coordination in Hardware, https://www.usenix.org/system/files/conference/nsdi16/nsdi16-paper-istvan.pdf

HT-Paxos- High Throughput State-Machine Replication Protocol for Large Clustered Data Centers, https://arxiv.org/abs/1407.1237

Ring Paxos: A High-Throughput Atomic Broadcast Protocol, http://www.inf.usi.ch/phd/jalili/RingPaxos-DSN2010.pdf

http://sci-hub.cc/10.1109/SRDS.2014.15

https://infoscience.epfl.ch/record/49946/files/HUS+02b.pdf

http://libpaxos.sourceforge.net/files/Primim-SPLab08.pdf

http://www.ic.unicamp.br/~reltech/2008/08-35.pdf

http://www.inf.usi.ch/faculty/soule/2015-06-22-disn.pdf

Megastore: Providing Scalable, Highly Available Storage for Interactive Servicesx

https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/36971.pdf

A Beginner’s Guide to Understanding the Leaf-Spine Network Topology

http://blog.westmonroepartners.com/a-beginners-guide-to-understanding-the-leaf-spine-network-topology/

http://pbs.cs.berkeley.edu/

There Is More Consensus in Egalitarian Parliaments, https://www.cs.cmu.edu/~dga/papers/epaxos-sosp2013.pdf

Resources

Sinfonia: a new paradigm for building scalable distributed systems, http://www.sosp2007.org/papers/sosp064-aguilera.pdf

https://web.archive.org/web/20131017235612/http://www.temple.edu/cis/icdcs2013/data/5000a011.pdf

A Scalable Conflict-free Replicated Set Data Type, https://web.archive.org/web/20131017233249/http://www.temple.edu/cis/icdcs2013/data/5000a186.pdf

FChain: Toward Black-box Online Fault Localization for Cloud Systems, https://web.archive.org/web/20131017235525/http://www.temple.edu/cis/icdcs2013/data/5000a021.pdf

Diagnosing Data Center Behavior Flow by Flow, https://web.archive.org/web/20131017235612/http://www.temple.edu/cis/icdcs2013/data/5000a011.pdf

Experimental Demonstration of a Quantum Protocol for Byzantine Agreement and Liar Detection, https://arxiv.org/pdf/0710.0290v2.pdf

https://tendermint.com/intro

Performance Comparison Between the Paxos and Chandra-Toueg Consensus Algorithms, https://infoscience.epfl.ch/record/49946/files/HUS+02b.pdf

A Distributed Lock Manager Using Paxos Design and Implementation of Warlock, a Consensus Based Lock Manager, http://uu.diva-portal.org/smash/get/diva2:615805/FULLTEXT01.pdf