The Performance Analysis of Cache Architecture based on Alluxio over Virtualized Infrastructure

Xu Chang, Li Zha
Contents

- Background
- Related Works
- Motivation
- Experiments
- Results
- Conclusion
- Future Work
Background

• **Cloud Computing**
  – Computing as a service
  – Application of resources on demand and payment on demand

• **Virtualization**
  – Integrates and encapsulates the resources
  – Provide the resource in piece
  – Transparent to users
Decoupling vs Traditional Architecture

Advantage:
- More flexible
- Overall cost is reduced

Shortcoming:
- Performance decline
Related Works

For making up the loss of performance

• Traditional optimization method
  – Speed up the shuffle part of jobs with SSDs
  – [kambatla2014truth] [ruan2017improving]

• Reduce the frequency of accessing the object storage
  – Construct the cache layer between applications and object storage
  – [shankar2017performance] [qureshi2014cache]
Related Works

Alluxio (Tachyon)

- The world’s first memory speed virtual distributed storage system
- Resides between computation frameworks and storage systems

Source: https://www.alluxio.org/
Motivation

• Only concern about performance, do not care about cost
• Cost reduction is critical
• Question:
  – How to design the caching architecture to make the cost performance highest?
Experiments

System architecture

Cloud Storage

MapReduce

Alluxio

MapReduce

Alluxio

MapReduce

Alluxio

Source: https://www.alluxio.org/
Experiments

Experimental environment

Experiment 1:  
Platform: AWS  
Servers: m3.2xlarge * 4  
Object storage: S3

Experiment 2:  
Platform: G-Cloud  
Servers: 8 cores & 30G memory * 4  
Object storage: Ceph
Experiments

Experimental scheme

• Experiment 1:
  – Workload: Terasort * 6

• Experiment 2:
  – Workload: Hive-Join * 3

• Data Size: 120G

• Cost ratio of memory to SSD

<table>
<thead>
<tr>
<th>Memory : SSD</th>
<th>8:0</th>
<th>7:1</th>
<th>5:3</th>
<th>3:5</th>
<th>1:7</th>
<th>0:8</th>
</tr>
</thead>
</table>

* Indicates a multiplication factor.
Results

Experimental 1:

Performance

Cost Performance
Results

Experimental 2:

![Performance Graph]

Throughput (MB/s)

- No Cache
- 100%MEM
- 87.5%MEM 12.5%SSD
- 62.5%MEM 37.5%SSD
- 37.5%MEM 62.5%SSD
- 12.5%MEM 87.5%SSD
- 100%SSD

![Cost Performance Graph]

Cost Performance

- 100%MEM
- 87.5%MEM
- 62.5%MEM
- 37.5%MEM
- 12.5%MEM
- 100%SSD
Conclusion

• Hybrid cache architecture is recommended.

• For the workload with large size of output and small size of hot data, the cost ratio of memory to SSD in cache should be around 1:7

• For the workload with small size of output and large size of hot data, the cost ratio of memory to SSD in cache should be around 5:3
Future Work

- Study several aspects that affect the cost performance, and try to give a configuration scheme with the best cost performance.

- Increase workload types and application scenarios, so that the conclusion is closer to the real scene and has generality.
Q & A

Thanks!