Bench 2018@Seattle

Scalability Evaluation of Big Data Processing Services in Clouds

Wei Huang[1,2], **Congfeng Jiang**[1,2], Zujie Ren[1,2], Huayou Si[1,2], Jian Wan[3]

1 Key Laboratory of Complex Systems Modeling and Simulation, Ministry of Education, Hangzhou 310018,China
2 School of Computer Science and Technology
Hangzhou Dianzi University, Hangzhou 310018, China
3 Department of Software Engineering, Zhejiang University of Science and Technology, Hangzhou, China

Outline

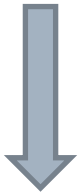
- Introduction
- Related Work
- Experiment and Analysis
- Implications

Introduction

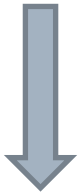
- Typical examples of cloud-based big data processing services include Amazon EMR, Microsoft Azure HDInsight, and AliCloud E-MapReduce.
- Among various cloud-based data processing services, how to scale the system is still challenging.
- How to evaluate the scalability of a big data processing system?
- Given a group of workload, should user scale-up or scale-out their deployed cluster? i.e., how to select the cluster configuration or rent a pre-configured big data processing platform for better performance?

Related Work

Big data benchmark: CloudSuite、BigDataBench、HiBench



Some research efforts have been done for evaluating big data system



Comparison of scalability of different service providers is still missing.

Our Work

- We proposed evaluation model for the scalability of big data processing system in clouds
- We evaluated the performance of Hadoop and Spark on AliCloud and BaiduCloud's big data processing platform in two dimensions of scale-out and scale-up configurations

Evaluation model

Speedup measurement:

- S_p represents the speed-up ratio:

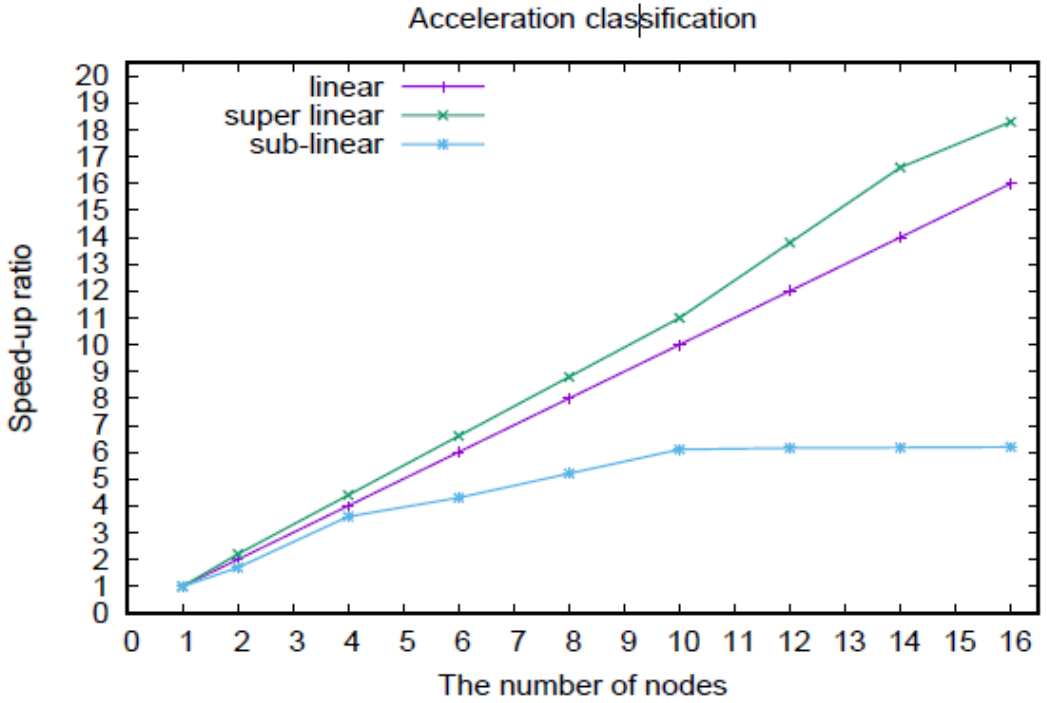
  $S_p = M_1/M_p$ (i.e., 1 node over multiple nodes)

- Scalability can be divided into three categories:

1. Linear acceleration
2. Sub-linear acceleration
3. Super linear acceleration

Evaluation model

- Acceleration classification



JIANG et al. Scalability Evaluation of Big Data Processing Services in Clouds. Bench 2018

Evaluation model

- Fit the speed-up ratio curve:

$$S = f(p)$$

- Measure the scalability of the system by:

$$Q = \int f(p)dp$$

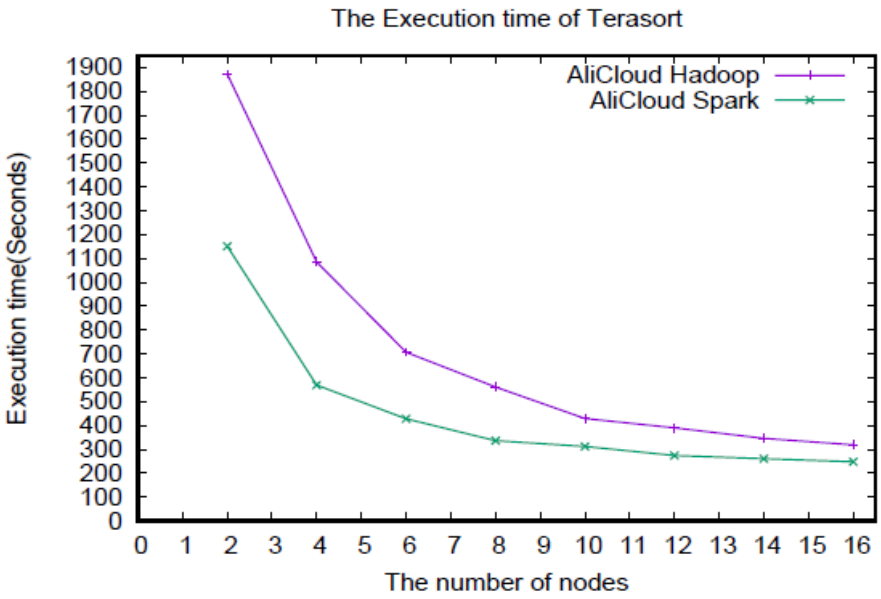
Experiment and Analysis

- Platforms: AliCloud E-MapReduce
  Baidu Cloud MRS
- Workloads: Terasort , WordCount
- System configuration for the host

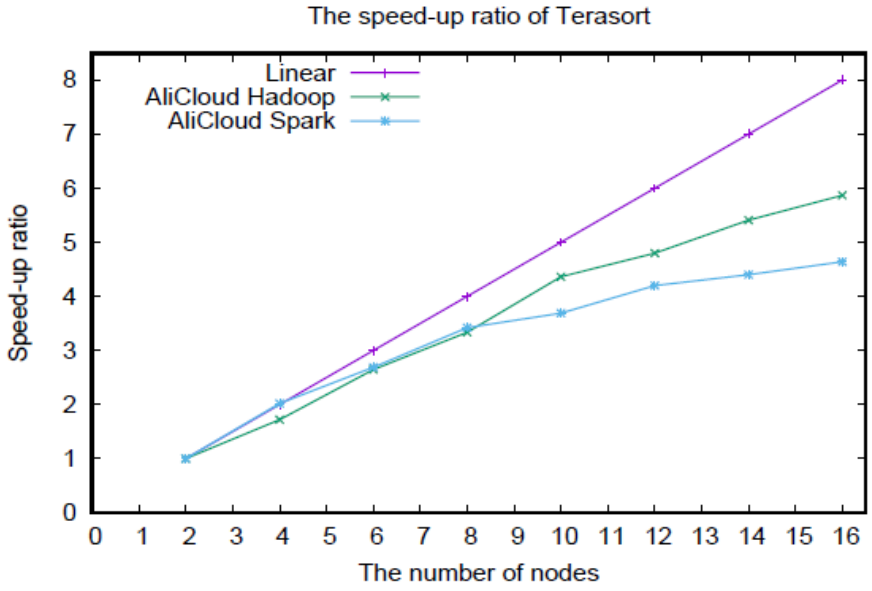
Configuration	NameNode	DataNode
CPU	4core	4core
Memory	16GB	16GB
Disk	SATA	SATA

Experiment and Analysis

- Scale-out on AliCloud(terasort)



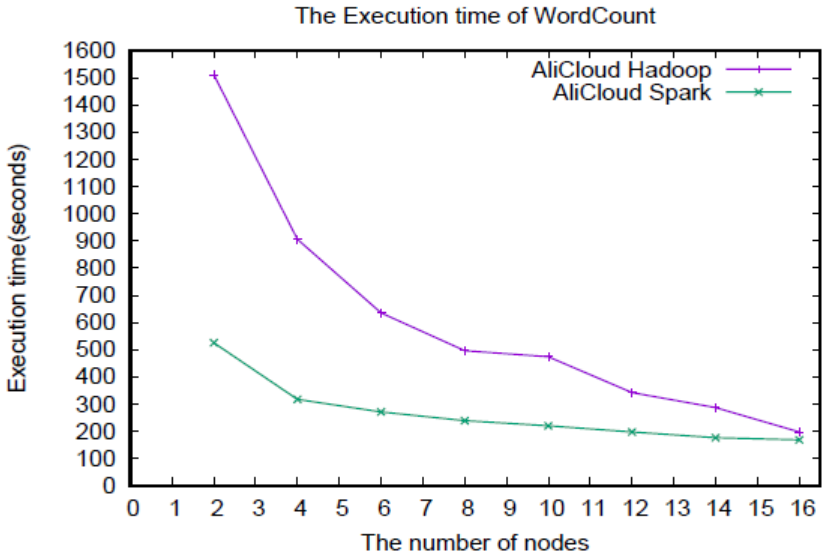
AliCloud Terasort execution time



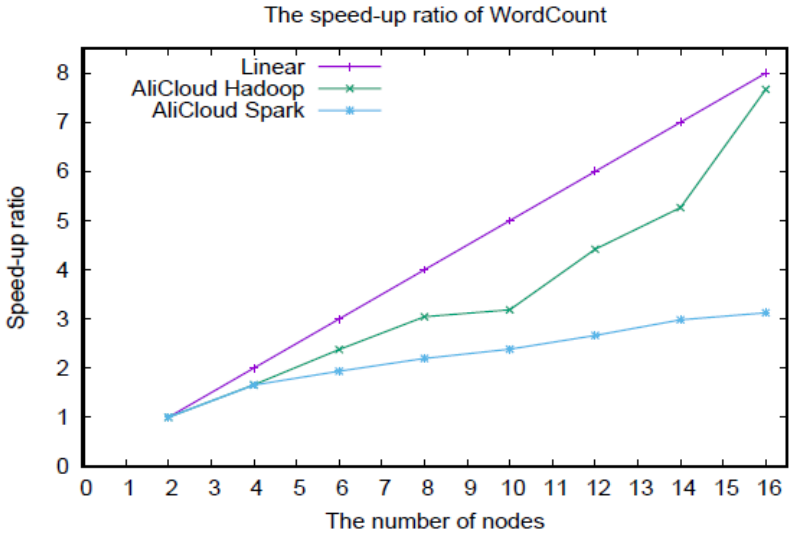
AliCloud Terasort speed-up ratio

Experiment and Analysis

- Scale-out on AliCloud (wordcount)



WordCount execution time

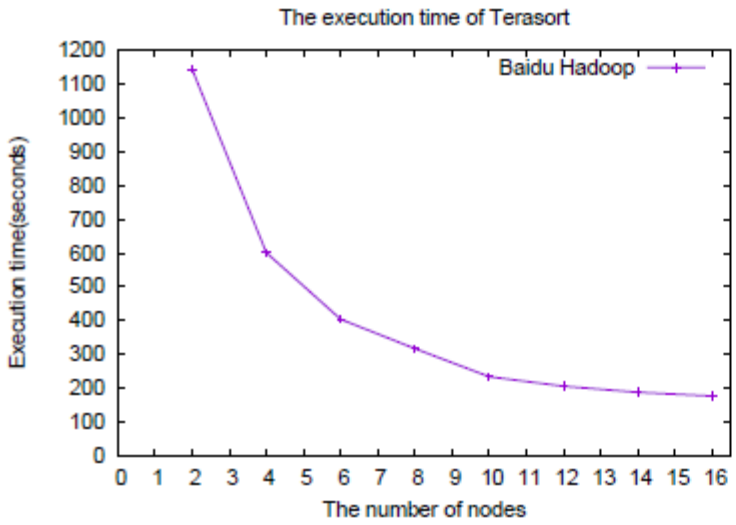


WordCount speed-up ratio

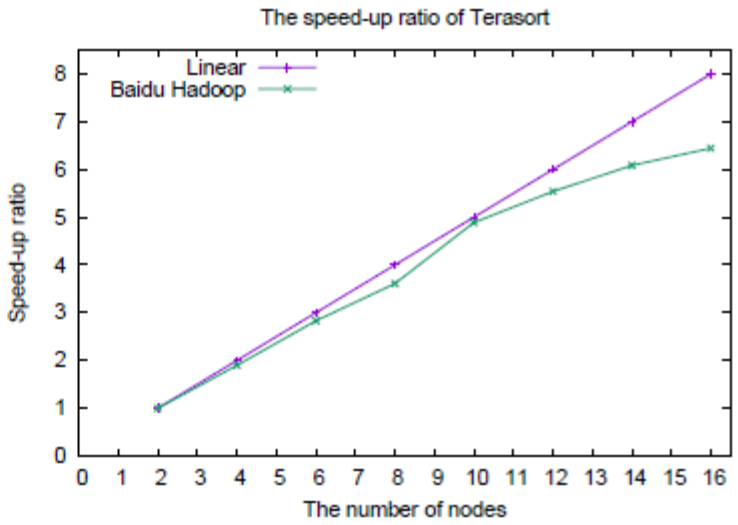
JIANG et al. Scalability Evaluation of Big Data Processing Services in Clouds. Bench 2018

Experiment and Analysis

- Scale-out on Baidu MRS



(a) Baidu Terasort execution Time



(b) Baidu Terasort speed-up ratio

Terasort execution time

Terasort speed-up ratio

JIANG et al. Scalability Evaluation of Big Data Processing Services in Clouds. Bench 2018

Experiment and Analysis

Summary of Scale-out comparison:

1. In the comparison of the speed-up ratio on AliCloud, (less than 8 nodes), scalability of Spark is better than Hadoop, then Spark's scalability is worse than Hadoop(larger than 8 nodes).
2. When Hadoop and Spark scale out to 16 nodes, the scale-out performance is good, and Hadoop overall performance(execution time) is better than the Spark in AliCloud.

Experiment and Analysis

- Scale-up experiment(only on AliCloud)

Experimental group	CPU	Memory
1	4core	16GB
2	8core	32GB
3	16core	64GB
4	32core	128GB

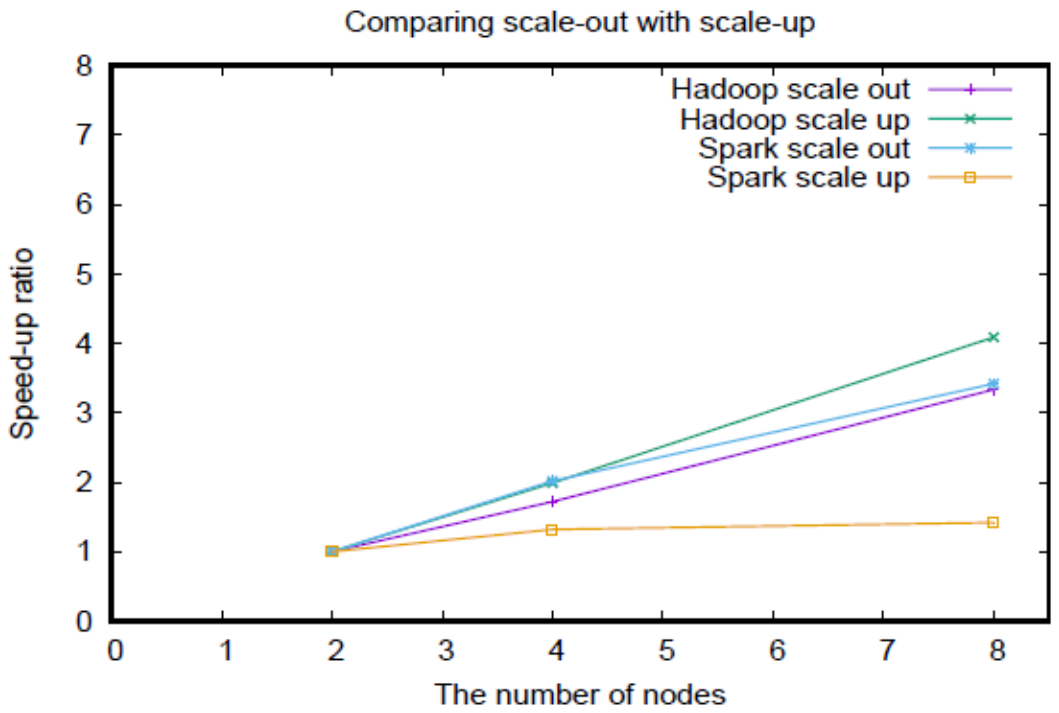
Experiment and Analysis

- Execution time for scale-up config

Experimental group/Task execution time	Hadoop	Spark
4core,16GB	1872 seconds	1151 seconds
8core,32GB	940 seconds	870 seconds
16core,64GB	457 seconds	810 seconds

Experiment and Analysis

- Comparison between scale-out and scale-up



JIANG et al. Scalability Evaluation of Big Data Processing Services in Clouds. Bench 2018

Implication #1

- The scalability of Hadoop and Spark are good enough on AliCloud and Baidu Cloud
- Hadoop's scalability is slightly better than Spark on AliCloud.
- Spark's speed is faster than Hadoop on AliCloud under WordCount workload
- The scalability of Hadoop on Baidu Cloud, is better than that on AliCloud.

Implication #2

- For Hadoop, scale-up is better than scale-out under the metric of processing performance(execution time).However, it's not true for Spark. This means that scale-up the Spark cluster may not achieve expected performance improvement.
- Here a dirty little secret is that scale-out is not more expensive than scale-up.

- The results presented here can be suggestions for Cloud services provider to design more scalable big data processing services avoid loss of customers.

Conclusions

- Different big data processing systems have different scalability
- Users should choose scale-out or scale-up wisely
- Cloud services provider can do more to provide more scalable big data processing services

杭州電子科技大学
HANGZHOU DIANZI UNIVERSITY

Thanks!