Aalto University
School of Science
Degree Programme in Computer Science and Engineering

Yangjun Wang

Stream Processing Systems Benchmark: StreamBench

Master's Thesis
Espoo, Nov 20, 2015

DRAFT! — December 30, 2015 — DRAFT!

Supervisors: Assoc. Prof. Aristides Gionis
Advisor: D.Sc. Gianmarco De Francisci Morales

Aalto University
School of Science
Degree Programme in Computer Science and Engineering

ABSTRACT OF
MASTER'S THESIS

Author: Yangjun Wang

Title:
Stream Processing Systems Benchmark: StreamBench

Date: Nov 20, 2015 **Pages:** 7

Major: Data Communication Software **Code:** T-110

Supervisors: Assoc. Prof. Aristides Gionis

Advisor: D.Sc. Gianmarco De Francisci Morales

Batch processing technologies(Such as MapReduce, Hive, Pig) have matured and been widely used in the industry. These systems solved the issue processing big volumes of data successfully. However, first big data need to be collected and stored in a database or file system. Then it takes time to finish batch processing analysis job before get any results. While there are many cases that need analysed results from streaming data immediately. The demand for processing real time stream data is increasing a lot these days. A big data architecture contains several parts. Often, masses of structured and semi-structured historical data are stored in Hadoop (Volume + Variety). On the other side, stream processing is used for fast data requirements (Velocity + Variety)[**?**]. Several streaming processing systems are implemented and widely adopted, such as Apache Storm, JStorm, Apache Spark, IBM InfoSphere Streams and Apache Flink. They all support real-time stream processing, high scalability, and awesome monitoring. How to evaluate a real time stream processing system before choosing it to use in production development is a open question. Before these real time stream processing systems are implemented, Michael demonstrated the 8 requirements[**?**] of real-time stream processing, which gives us a standard to evaluate whether a real time stream processing system satisfies these requirements. A very common and traditional approach to verify whether the performance of a system meets the requirements is benchmarking. Published benchmarking results from industry standard benchmark systems could help users compare products and understand features of a system easily.

Keywords: Big Data, Stream, Benchmark, Storm, Flink, Spark

Language: English

Acknowledgements

I want to thank ProfessorAristides Gionis and my advisor Gianmarco De Francisci Morales for their good guidance.

Espoo, Nov 20, 2015

Yangjun Wang

Abbreviations and Acronyms

Symbols

$\mathbf{B}$	magnetic flux density
c	speed of light in vacuum $\approx 3 \times 10^8$ [m/s]
ω_{D}	Debye frequency
ω_{latt}	average phonon frequency of lattice
$\uparrow$	electron spin direction up
$\downarrow$	electron spin direction down

Operators

$\nabla \times \mathbf{A}$	curl of vectorin $\mathbf{A}$
$\frac{\mathrm{d}}{\mathrm{d}t}$	derivative with respect to variable t
$\frac{\partial}{\partial t}$	partial derivative with respect to variable t
$\sum_i$	sum over index i
$\mathbf{A} \cdot \mathbf{B}$	dot product of vectors $\mathbf{A}$ and $\mathbf{B}$

Abbreviations

AC	alternating current
APLAC	an object-oriented analog circuit simulator and design tool (originally Analysis Program for Linear Active Circuits)
BCS	Bardeen-Cooper-Schrieffer
DC	direct current
TEM	transverse eletromagnetic

Contents

List of Figures

Chapter 1

Introduction

Introduce big data and the four ***V***s of Big Data

1.1 Stream Processing Systems and Evaluation

Describe common stream processing systems and current evaluation and comparison these platforms.

1.2 Structure of the Thesis

Structure description of the thesis

1.3 Background

Background knowledge of StreamBench which includes Big Data, Cloud Computing and widely accepted benchmark systems.

1.3.1 Cloud Computing

Concept of Cloud Computing and how Cloud Computing solves Big Data issues

1.3.1.1 Parallel Computing

1.3.1.2 Computing Cluster

1.3.1.3 Batch Processing and Stream Processing

1.3.2 Apache Hadoop

Introduce Apache Hadoop and several important modules

1.3.2.1 MapReduce

1.3.2.2 Hadoop Distribution File Systems

1.3.2.3 YARN

1.3.2.4 Zookeeper

1.3.3 Benchmark

Describe benchmark systems of traditional database and cloud service systems. Demonstrate design and components of benchmark system

1.3.3.1 Traditional Database Benchmark

1.3.3.2 Cloud Service Benchmark

1.4 Stream Processing Platforms

Introduce three widely used stream processing platforms, point out core concepts and key features

1.4.1 Apache Storm

1.4.1.1 Storm Architecture

1.4.1.2 Computing Model

1.4.2 Apache Flink

1.4.2.1 Flink Architecture

1.4.2.2 Memory Management

1.4.2.3 Flink Streaming

1.4.3 Apache Spark

1.4.3.1 Resilient Distributed Datasets(RDDs)

1.4.3.2 Spark Streaming

1.5 Benchmark Design

1.5.1 Architecture

1.5.2 Experiment Environment Setup

1.5.3 Data Source

1.5.3.1 Test Data Generation

1.5.3.2 Kafka

1.5.4 Experiment Log and Statistic

1.5.5 Extensibility

1.6 Experiment

1.6.1 Experiment Environment

1.6.2 Classic Workload

1.6.2.1 WordCount

1.6.2.2 Data Source

1.6.2.3 Algorithm Description

1.6.2.4 Results and Discussion

1.6.3 Multi-Streams Join Workload

1.6.3.1 Advertisements Click

1.6.3.2 Data Source

1.6.3.3 Algorithm Description

1.6.3.4 Results and Discussion

1.6.4 Iterate Workload

1.6.4.1 WordCount

1.6.4.2 Data Source

1.6.4.3 Algorithm Description

1.6.4.4 Results and Discussion

1.7 Conclusions

Summary of experiment results

1.7.1 Selection in Practice

Summarize several factors which affect selection of stream processing systems in practic

1.7.1.1 Performance Summary

1.7.1.2 Issues

1.7.2 Future Work

Future works

1.7.2.1 Scale-out and Elasticity Evaluation

1.7.2.2 Evaluation of Other Platforms

.1 Source Code

.1.1 WordCount

.1.2 Advertisements Click