BigDataBench: An Open-source Big Data Benchmark Suite

Jianfeng Zhan
http://prof.ict.ac.cn/BigDataBench

Professor, ICT, Chinese Academy of Sciences and University of Chinese Academy of Sciences

WBDB 2015 Toronto, Canada
What is **BigDataBench**?

- An open source big data benchmarking project
  - [http://prof.ict.ac.cn/BigDataBench](http://prof.ict.ac.cn/BigDataBench)
  - Search Google using “**BigDataBench**”
## Why BigDataBench?

<table>
<thead>
<tr>
<th></th>
<th>Specification</th>
<th>Application domains</th>
<th>Workload Types</th>
<th>Workloads</th>
<th>Scalable data sets (from real data)</th>
<th>Multiple implementations</th>
<th>Multi tenancy</th>
<th>Sub sets</th>
<th>Simulat or version</th>
</tr>
</thead>
<tbody>
<tr>
<td>BigDataBench</td>
<td>Y</td>
<td>Five</td>
<td>Five</td>
<td>33</td>
<td>8</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y(3)</td>
</tr>
<tr>
<td>BigBench</td>
<td>Y</td>
<td>One</td>
<td>Three</td>
<td>10</td>
<td>3</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>CloudSuite</td>
<td>N</td>
<td>N/A</td>
<td>Two</td>
<td>8</td>
<td>3</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y(1)</td>
</tr>
<tr>
<td>HiBench</td>
<td>N</td>
<td>N/A</td>
<td>Two</td>
<td>10</td>
<td>3</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>CALDA</td>
<td>Y</td>
<td>N/A</td>
<td>One</td>
<td>5</td>
<td>N/A</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>YCSB</td>
<td>Y</td>
<td>N/A</td>
<td>One</td>
<td>6</td>
<td>N/A</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>LinkBench</td>
<td>Y</td>
<td>N/A</td>
<td>One</td>
<td>10</td>
<td>1</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>AMP Benchmarks</td>
<td>N</td>
<td>N/A</td>
<td>One</td>
<td>4</td>
<td>N/A</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>
BigDataBench evolution

- BigDataBench 1.0: Search engine, 6 workloads
- BigDataBench 2.0: Typical Internet service domains, An architectural perspective, 19 workloads & data generation tools
- BigDataBench 3.0: Multidisciplinary effort, 32 workloads: diverse implementations
- BigDataBench 3.1: 5 application domains: 14 data sets and 33 workloads, Same specifications: diverse implementations, Multi-tenancy version, BigDataBench subset and simulator version

2013.7

2013.12

2014.4

2014.12

2014.12

2014.4

2013.12

2013.7

Data sources: BigDataBench.
BigDataBench Users

- http://prof.ict.ac.cn/BigDataBench/users/
- Industry users
  - Accenture, BROADCOM, SAMSUNG, Huawei, IBM
- About 20 academia groups published papers using BigDataBench
Industry Standard: BigDataBench-DCA

- China’s first industry-standard big data benchmark suite
  - http://prof.ict.ac.cn/BigDataBench/industry-standard-benchmarks/
  - Telecom Research Institute of Ministry of Industry and Information Technology, ICT, CAS, Huawei, China Mobile, Sina, ZTE, Intel (China), Microsoft (China), IBM CDL, Baidu, INSPUR, ZTE, 21viane and UCloud
BigDataBench Publications


- Characterizing data analysis workloads in data centers. 2013 IEEE International Symposium on Workload Characterization (IISWC 2013) (Best paper award)

- BigOP: generating comprehensive big data workloads as a benchmarking framework. 19th International Conference on Database Systems for Advanced Applications (DASFAA 2014)


Outline

- BigDataBench 3.1 Overview

- Dwarf Workloads in Big Data Analytics

- Insights
What’s New in BigDataBench 3.1

Methodology
- New application domains
- Specification for each domain

Implementation
- New real world data sets
- New big data workloads

Specific-purpose Version
- Multi-tenancy version
- BigDataBench subset version
What is new?

- **Application Domains**
  - Previous version
    - Internet service: search engine, social network, e-commerce
  - Newly added: bioinformatics, multimedia

- **Specification**
  - Guidelines for benchmark implementation
Methodology evolution

**BigDataBench 3.0**

- **Data Complexity**
- **Typical Application Domains**
- **Scalable data sets and workloads**
- **Different implementations**
- **System characteristics**

**BigDataBench 3.1**

- Application Domain 1
- Application Domain 2
- Application Domain N

- Benchmark specification 1
- Benchmark specification 2
- Benchmark specification N

- Real-world data sets
- Data generation tools
- Workloads with diverse implementations

- Multi-tenancy version
- Mix with different percentages
- Reduce benchmarking cost

**BigDataBench | WBDB 2015**
Specification--Search Engine

General search and vertical search
Online server and Offline analytics
Search Engine: Parsing

- Parsing:
  - Extract the text content and links from the raw web pages
Search Engine: Indexing

- Indexing
  - create the mapping of terms to document id lists

Diagram:
- Web
- Parsing
- Indexing
- Content
- URL
- Out_link
- Key_word
- Classification
- Category
- Statistic
- PageRank
- Score
- Filter, semantic extract
- Semantic information
- Recommendation
- Query log
- Data access
- Sorting
- Index
- Web server
- Query
- BigDataBench WBDB 2015
Search Engine: PageRank

- PageRank
  - Compute the importance of the page according to the web link graph using PageRank
Search Engine: Search query

- Querying
  - online web search server serving users' requests

```
web
  ↓
parsing
  ↓
index
  ↓
search
  ↓
sorting
  ↓
data access
  ↓
recommendation
  ↓
query log

key_word classification category
  ↓
statistic
  ↓
content url out_link
  ↓
PageRank score
  ↓
filter, semantic extract
  ↓
metadata
```

BigDataBench | WBDB 2015
Search Engine: Sorting

- Sorting
  - Sort the results according to the page ranks and the relevance of between the query and the document.
Search Engine: Recommendation

- **Recommendation**
  - Recommend related queries to users by mining the search log

---

BigDataBench | WBDB 2015 | 中国科学院计算技术研究所

INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
Search Engine: Statistic counting

- Statistic counting
  - Counting the word frequency to extract the key words which represent the features of the page.
Search Engine: Classification

- Classification
  - Classify text content into different categories.
Search Engine: Filter & Semantic extraction

- **Filter**
  - Identify pages with specific topic which can be used for vertical search

- **Semantic extraction**
  - Extract semantic information

```
web
  ↓
parsing
  ↓
content → url → out_link → PageRank → score
  ↓
statistic → classification → category
  ↓
index
  ↓
sorting → data access
  ↓
query log
  ↓
web server
  ↓
query
```

BigDataBench | WBDB 2015 | IICT
Search Engine: Data access

- Data access operations
  - Read, write, and scan the semantic information.

BigDataBench WBDB 2015
Specification--Social network

- Data sets
  - User table
  - Relation table
  - Article table

- Workloads
  - Offline analytics
## Social network: Data schema

### User table

<table>
<thead>
<tr>
<th>attribute</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>user_id</td>
<td>the id of the user</td>
</tr>
<tr>
<td>sex</td>
<td>the sex of the user</td>
</tr>
<tr>
<td>age</td>
<td>the age of the user</td>
</tr>
<tr>
<td>education</td>
<td>the situation of education</td>
</tr>
<tr>
<td>tag</td>
<td>the terms showing characteristics of the user</td>
</tr>
</tbody>
</table>

### Relation table

<table>
<thead>
<tr>
<th>attribute</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>user_id</td>
<td>the id of the user</td>
</tr>
<tr>
<td>follow_user_id</td>
<td>the user id who is followed</td>
</tr>
</tbody>
</table>

### Tweet table

<table>
<thead>
<tr>
<th>attribute</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweet_id</td>
<td>the id of the tweet</td>
</tr>
<tr>
<td>content</td>
<td>the content of the tweet</td>
</tr>
<tr>
<td>user_id</td>
<td>the id of user who own the tweet</td>
</tr>
<tr>
<td>review_number</td>
<td>the number of review</td>
</tr>
<tr>
<td>transmit_number</td>
<td>the number of transmitting</td>
</tr>
<tr>
<td>time</td>
<td>the publish time of the tweet</td>
</tr>
</tbody>
</table>
Social network: Workloads

- **Hot review topic**
  - Select the top N tweets by the number of review

- **Hot transmit topic**
  - Select the tweets which are transmitted more than N times.

- **Active user**
  - Select the top N person who post the largest number of tweets.

- **Leader of opinion**
  - Select top ones whose number of review and transmit are both large than N.
Social network: Workloads

- Topic classify
  - Classify the tweets to certain category according to the topic.

- Sentiment classify
  - Classify the tweets to negative or positive according to the sentiment.

- Friend recommendation
  - Recommend friend to person according the relational graph.

- Community detection
  - Detecting clusters or communities in large social networks.

- Breadth first search
  - Sort persons according to the distance between two people.
Specification: E-commerce

Order table

<table>
<thead>
<tr>
<th>attribute</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>order_id</td>
<td>the id of the order</td>
</tr>
<tr>
<td>buyer_id</td>
<td>the id of person who own the order</td>
</tr>
<tr>
<td>time</td>
<td>the time of the order occurred</td>
</tr>
</tbody>
</table>

Item table

<table>
<thead>
<tr>
<th>attribute</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>item_id</td>
<td>the id of the item</td>
</tr>
<tr>
<td>order_id</td>
<td>the id of order which the item belongs to</td>
</tr>
<tr>
<td>goods_id</td>
<td>the id of goods</td>
</tr>
<tr>
<td>goods_number</td>
<td>the number of goods</td>
</tr>
<tr>
<td>price</td>
<td>the price of goods</td>
</tr>
<tr>
<td>amount</td>
<td>the total assumption of the item</td>
</tr>
<tr>
<td>score</td>
<td>the score the buyer gave</td>
</tr>
<tr>
<td>review</td>
<td>the text commence the buyer gave</td>
</tr>
</tbody>
</table>
E-commerce

- Data sets
  - Order table
  - Item table

- Workloads
  - Offline analytics

Data sets:
- Order table
- Item table

Workloads:
- Offline analytics

Diagram:
- Buyers
  - Order table
  - Item table
  - Data analytics
E-commerce: Workloads

- Select query
  - Find the items of which the sales amount is over 100 in a single order.
- Aggregation query
  - Count the sales number of each goods.
- Join query
  - Count the number of each goods that each buyer purchased between certain period of time.
- Recommendation
  - Predict the preferences of the buyer and recommend goods to them.
- Sensitive classification
  - Identify positive or negative review.
- Basic data operation
  - Unit of operation of the data

The workloads of select, aggregation, and join are similar to the queries in A. Pavlo’s sigmod09 paper, but are specified in the e-commerce environment.
Specification--Multimedia

Front end cameras

Video Data → MPEG Encoder → MPEG Decoder → Intelligent Video Analysis

Monitoring data analysis

Voice Data Extraction → Speech Recognition
Frame Data Extraction → Feature Extraction

Image Segmentation → Face Detection

Three-Dimensional Reconstruction → Tracing

BigDataBench | WBDB 2015
Multimedia: Workloads

- **MPEG Decoder.**
  - Decode video streams using MPEG-2 standard.

- **Feature extraction**
  - For a given video frame, extract features which are invariant to scale, noise, and illumination.

- **Speech Recognition.**
  - For a given audio file, recognize the content of the file and find whether there are sensitive words.
Multimedia: Workloads

- **Ray Tracing.**
  - Render a 2-Dimensional video frame to a 3-Dimensional scene.

- **Image Segmentation.**
  - Segment the input video frame according to color, intensity, and texture, and extract concerned regions.

- **Face Detection.**
  - Detect whether face exists in the input data, if exists, then extract the face.

- **Deep Learning.**
  - The input images are classified into different categories, and then detect human face.
Specification--Bioinformatics

- **Sequence assembly.**
  - Assemble scattered and repetitive DNA fragments to original long sequence.

- **Sequence alignment.**
  - Align assembled DNA sequence to known sequences in the database, and detect disease.
What’s New in BigDataBench 3.1

**Methodology**
- New application domains
- Specification for each domain

**Implementation**
- New real world data sets
- New big data workloads

**Specific-purpose Version**
- Multi-tenancy version
- BigDataBench subset version
BigDataBench 3.1 Overview

**BDGS (Big Data Generator Suite) for scalable data**

| Wikipedia Entries | Amazon Movie Reviews | Google Web Graph |
| Facebook Social Network | E-commerce Transaction | ProfSearch Resumes |
| ImageNet | English broadcasting audio | DVD Input Streams |
| Image scene | Genome sequence data | Assembly of the human genome |
| SoGou Data | MNIST |

**14 Real-world Data Sets**

- Wikipedia Entries
- Amazon Movie Reviews
- Google Web Graph
- Facebook Social Network
- E-commerce Transaction
- ProfSearch Resumes
- ImageNet
- English broadcasting audio
- DVD Input Streams
- MNIST
- Genome sequence data
- Assembly of the human genome
- SoGou Data

**33 Workloads**

- Search Engine
- Social Network
- E-commerce
- Multimedia
- Bioinformatics

**Software Stacks**

- Impala
- Shark
- Hadoop RDMA
- NoSQL
- MVAPICH
- MPI
- DataMPI

**Tools**

- Hadoop
- Spark
- Impala
- Shark

**BigDataBench | WBDB 2015 | ICT**
What’s New in BigDataBench 3.1

Methodology
- New application domains
- Specification for each domain

Implementation
- New real world data sets
- New big data workloads

Specific-purpose Version
- Multi-tenancy version
- BigDataBench subset version
BigDataBench Multi-tenancy Version

- Scenarios of multiple tenants running heterogeneous applications in cloud datacenters
  - Latency-critical online services
  - Latency-insensitive offline batch applications

- Mining real-world Workload traces (Google and Facebook)
- Profiling Real-world Workload traces
- Workload matching using Machine learning techniques
- Parametric workload generation tool

Benchmarking scenarios
- Mixed workloads in public clouds
- Data analytical workloads in private clouds

BigDataBench WBDB 2015
Motivation

- It is expensive to run all the benchmarks for system and architecture researches
  - multiplied by different implementations
  - BigDataBench 3.0 provides about 77 workloads
Methodology of Subsetting

- Identify a comprehensive set of workload characteristics from a specific perspective
- Eliminate the correlation data in those metrics
  - Map the high dimension metrics to a low dimension
- Use the clustering method to classify
  - Choose representative workloads from each category
Outline

- BigDataBench 3.1 Overview
- Dwarf Workloads in Big Data Analytics
- Insights
Why Dwarf Workloads in Big Data Analytics

How to define a representative big data benchmark?

• One attempt
  – Using popular workloads
  – Subjective

• Another attempt
  – Specific domains of systems

How can we construct a benchmark suite using a minimum set of units of computation to represent diversity of big data analytics? Dwarf workloads!
Significance

A minimum set of necessary functionality

A direction for evaluation and performance optimization

A highly abstraction of patterns

Strong expression power

Dwarf Workloads

http://cacs.usc.edu/education/cs596/DavidPatterson.pdf
**Inspiration**

**Successful Compute Abstractions**

- Relational algebra
  - 5 primitive operations
  - Select, Project, Product, Union, Difference

- Parallel computing
  - Computational & communication patterns
  - 13 dwarfs

**Successful Benchmarks**

- TPC-C
  - OLTP domain
  - Functions of abstraction

- HPCC
  - High performance computing
  - Seven basically tests
Fundamental Issues

**What** is the dwarf workloads in big data analytics?

**How** to find?
Challenges #1--Massive Application Domains

Massive application domains make us wonder where to start or how to achieve a wide range of coverage.
Challenges #2—Multiple Research Fields and Techniques

Many techniques for processing big data exist, which bring greater complexity for identifying dwarfs workloads.
Challenges #3—Large Numbers of Algorithms and Variants

A machine learning library – scikit learn, implements so many algorithms, which is still much less than the total number of algorithms

Challenge #4—Unstructured Data with Complicated Operations

- 80% data growth are unstructured data
- Operations on big data are *complicated*
  - Pipeline? Parallel?

https://www.capgemini.com/blog/capping-it-off/2014/07/are-you-effectively-using-big-data
Methodology of Dwarfs

Big Data Analytics

Processing Techniques
- Machine learning
- Data mining
- Deep learning
- Computer vision
- Natural language processing...

Libraries
- Mlib, Mahout

Frameworks
- Spark, Hadoop, GraphLab

Benchmarks
- BigBench, LinkBench...

Representative Algorithms

- Frequently appearing operations
- Different combinations

Application
- Domain 1
- Domain 2
- Domain N

Dwarfs
- 1
- 2
- M

BigDataBench | WBDB 2015
Internet Services

Taking up 80% of internet services according to page views and daily visitors

Top 20 websites

http://www.alexa.com/topsites/global;0
# The Explosive Growth of Multimedia Data

<table>
<thead>
<tr>
<th>600+ new VIDEOS on YouTube every minute</th>
<th>13000+ hours MUSIC streaming on PANDORA every minute</th>
<th>6600+ new PHOTOS on FLICKR every minute</th>
</tr>
</thead>
<tbody>
<tr>
<td>100’s VIDEO feeds from surveillance cameras</td>
<td>370000+ minutes VOICE calls on Skype every minute</td>
<td>80% data growth are IMAGES, VIDEOS, documents, ...</td>
</tr>
</tbody>
</table>


**BigDataBench** | **WBDB 2015** | **ICT** INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
The Explosive Growth of Bioinformatics Data

DDBJ/EMBL/GenBank database Growth

http://www.ddbj.nig.ac.jp/breakdown_stats/dbgrowth-e.html#dbgrowth-graph
The Explosion Growth of Astronomy Data

- The last decade
  - hundreds of terabytes of astronomical data for hundreds of millions of sources

- The next decade
  - Large Synoptic Survey Telescope (LSST)

One 6-Gigabyte image every 20 seconds
30 Terabytes every night for 10 years
100-Petabyte final image data archive anticipated

Chosen Application Domain

Internet Service
*Search engine, Social network, E-commerce*

Multimedia
Top 20 websites

**Bioinformatics**

DDBJ/EMBL/GenBank database Growth

- **Nucleotides**

- **Entries (million)**

- **Entries (billion)**

**Astronomy**

- **Gigabytes**

- **Images, Videos, Documents, ...**

Large Synoptic Survey Telescope (LSST)

- **Images, Videos, Documents, ...**

- **Anticipated**

BigDataBench | WBDB 2015
Methodology of Dwarfs

Big Data Analytics
- Processing Techniques
  - Machine learning
  - Data mining
  - Deep learning
  - Computer vision
  - Natural language processing ...
- Libraries
  - MLlib, Mahout
- Frameworks
  - Spark, Hadoop, GraphLab
- Benchmarks
  - BigBench, LinkBench ...

Representative Algorithms
- Frequently appearing operations
- Different combinations

Application Domain 1
- Dwarfs 1

Application Domain 2
- Dwarfs 2

Application Domain N
- Dwarfs M
Big Data Analytics

- Structured
- Semi-Structured
- Unstructured

- Text
- Graph
- Table
- Multimedia

Data Model

Semantics

Processing Techniques

Open Source Projects

- Data Mining
- Machine Learning ...

- Libraries
- Frameworks
- Benchmarks

BigDataBench WBDB 2015
Libraries & Frameworks & Benchmarks

- **Libraries**
  - Opencv, Mlib, Weka, AstroML ...

- **Frameworks**
  - Spark, Hadoop, Graphlab ...

- **Benchmarks**
  - BigDataBench, Linkbench ...
Methodology of Dwarfs

Big Data Analytics

- Processing Techniques
  - Machine learning
  - Data mining
  - Deep learning
  - Computer vision
  - Natural language processing ...

- Libraries
  - Mlib, Mahout

- Frameworks
  - Spark, Hadoop, GraphLab

- Benchmarks
  - BigBench, LinkBench ...

Representative Algorithms

- Frequently appearing operations
- Different combinations

Application Domain 1
- Dwarfs 1

Application Domain 2
- Dwarfs 2

... Application Domain N
- Dwarfs M
Algorithms Chosen to Investigate

**Computer Vision**
- MPEG-2, Scale-Invariant feature transform, Image segmentation, Ray Tracing

**Bioinformatics**
- Needleman-Wunsch, Smith-Waterman, BLAST

**Deep Learning**
- CNN, DBN

**Data Mining & Machine Learning**
- C4.5/CART/ID3, Logistic regression, SVM, KNN, HMM, Maximum-entropy Markov model, Conditional random field, PageRank, HITS, Aporiori, FP-growth, Principal component analysis, Back Propagation, Adaboost, MCMC, Connected component, Random forest

**Natural Language Processing**
- Latent semantic indexing, pLSI, Latent Dirichlet allocation, Index, Porter Stemming, Sphinx speech recognition

**Recommendation**
- Demographic/Content based recommendation, Collaborative Filtering

---

BigDataBench | WBDB 2015 | INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
Frequently-appearing Operations

- **Similarity Analysis**
  - KNN, K-means, Recommendation, Feature matching, Image segmentation

- **Neural Network**
  - Back propagation, CNN, DBN, Neural network

- **Linear Algebra**

- **Multimedia Representation**
  - Sphinx speech recognition, MPEG-2, SIFT, Image segmentation, Ray tracing

- **Matrix/Vector Calculation**
  - SVM, HMM, MEMM, CRF, PageRank, HITS, Logistic regression
Frequently-appearing Operations (cont’)

- Jaccard similarity
- Locality sensitive hashing...

Set Operations

Association Rules Mining
- Apriori
- FP-Growth
- ...

Similarity Analysis

Database
- Set union
- Set difference
- ...

BigDataBench | WBDB 2015
Frequently-appearing Operations (cont’)

- **Sampling**
  - MCMC, LDA, Random sampling, Downsampling...

- **Transform Operation**
  - FFT, Convolution computation, DCT, Speech recognition, MPEG-2...

- **Graph Operation**
  - BFS, DFS, Decision tree, Connected Component...

- **Logic Operation**
  - Encryption, index, Fingerprint, SimHash, MinHash...

BigDataBench | WBDB 2015 | ICT

INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
Frequently-appearing Operations (cont’)

- Two more primitive operations

**Statistic Operation**
- Probability calculation
- LSI, pLSI, Latent Dirichlet Allocation...

**Sort**
- Partial sort, quick sort, Top k sort...
- K-means, Decision tree...
Methodology of Dwarfs

Big Data Analytics

Processing Techniques
- Machine learning
- Data mining
- Deep learning
- Computer vision
- Natural language processing ...

Libraries
- Mllib, Mahout

Frameworks
- Spark, Hadoop, GraphLab

Benchmarks
- BigBench, LinkBench ...

Representative Algorithms
- Frequently appearing operations
- Different combinations

Application Domain 1
- Dwarfs 1

Application Domain 2
- Dwarfs 2

Application Domain N
- Dwarfs M
Finalized 8 Dwarfs workloads

- Linear Algebra
- Sampling
- Transform operation
- Graph operation
- Logic operation
- Set operation
- Statistic operation
- Sort

BigDataBench | WBDB 2015

INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES
Properties

• Composability
  – Algorithm can be composed of one or more dwarfs

• Irreversibility
  – Combination is sensitive to the order of dwarfs

• Uniqueness
  – Represent different patterns
Outline

- BigDataBench 3.1 Overview
- Dwarf Workloads in Big Data Analytics
- Insights
System Behaviors

- Diversified system level behaviors:
  - High CPU utilization & less I/O time
  - Low CPU utilization relatively and lots of I/O time
  - Medium CPU utilization and I/O

BigDataBench | WBDB 2015
Workloads Classification

- Finding from system behaviors
  - System behaviors vary across different workloads
  - Workloads can be divided into 3 categories:

<table>
<thead>
<tr>
<th>Type</th>
<th>Workloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>I/O Intensive</td>
<td>H-Read, H-Difference, I-SelectQuery, S-WordCount, S-Project, S-OrderBy, M-Grep and S-Grep</td>
</tr>
<tr>
<td>Hybrid</td>
<td>H-TPC-DS-query3, I-OrderBy, S-TPC-DS-query10, S-TPC-DS-query8, S-Sort, M-WordCount and M-Sort</td>
</tr>
</tbody>
</table>
The average IPC of the big data workloads larger than CloudSuite, SPECFP and SPECINT, same like PARSEC and slightly smaller than HPCC.

The average IPC of BigDataBench is **1.3 times** of that of CloudSuite.

Some workloads have high IPC (M_Kmeans, S-TPC-DS-Query8)
**Instructions Mix of BigDataBench vs. other benchmarks**

- Big data workloads are data movement dominated computing with more branch operations
  - 92% percentage in terms of instruction mix (Load + Store + Branch + data movements of INT)
Cache Behaviors of BigDataBench vs. other benchmarks

- **L1I MPKI**
  - *Larger than traditional benchmarks, but lower than that of CloudSuite (12 Vs. 31)*
  - *Different among big data workloads*
    - CPU-intensive(8), I/O intensive(22), and hybrid workloads(9)
  - *One order of magnitude differences among diverse implementations*
    - M_WordCount is 2, while H_WordCount is 17
Cache Behaviors

- L2 Cache:
  - The IO-intensive workloads undergo more L2 MPKI

- L3 Cache:
  - The average L3 MPKI of the big data workloads is smaller than all of the other workloads

- The underlying software stacks impact data locality
  - MPI workloads have better data locality and less cache misses
Locality

*Instructions Cache miss ratio versus Cache size*

*Data Cache miss ratio versus Cache size.*

Hadoop workloads have larger instructions footprint
Our observation from BigDataBench

- **Unique characteristic**
  - data movement dominated computing with more branch operations
    - 92% percentage in terms of instruction mix

- **Locality**
  - Hadoop workloads have larger instructions footprint

- **Different behaviors among Big Data workloads**
  - Disparity of ILP and memory access behaviors
    - CloudSuite is a subclass of Big Data

- **Software stacks impacts**
  - The L1I cache miss rates have one order of magnitude differences among diverse implementations with different software stacks.