Service Oriented Performance Analysis

Da Qi Ren and Masood Mortazavi
US R&D Center
Santa Clara, CA, USA

HUAWEI

Performance Model for Service in Data Center and Cloud

1. Service Oriented (end to end) big data performance analysis has become an extreme attention of ICT industry, the related techniques are being investigated by numerous hardware and software vendors.

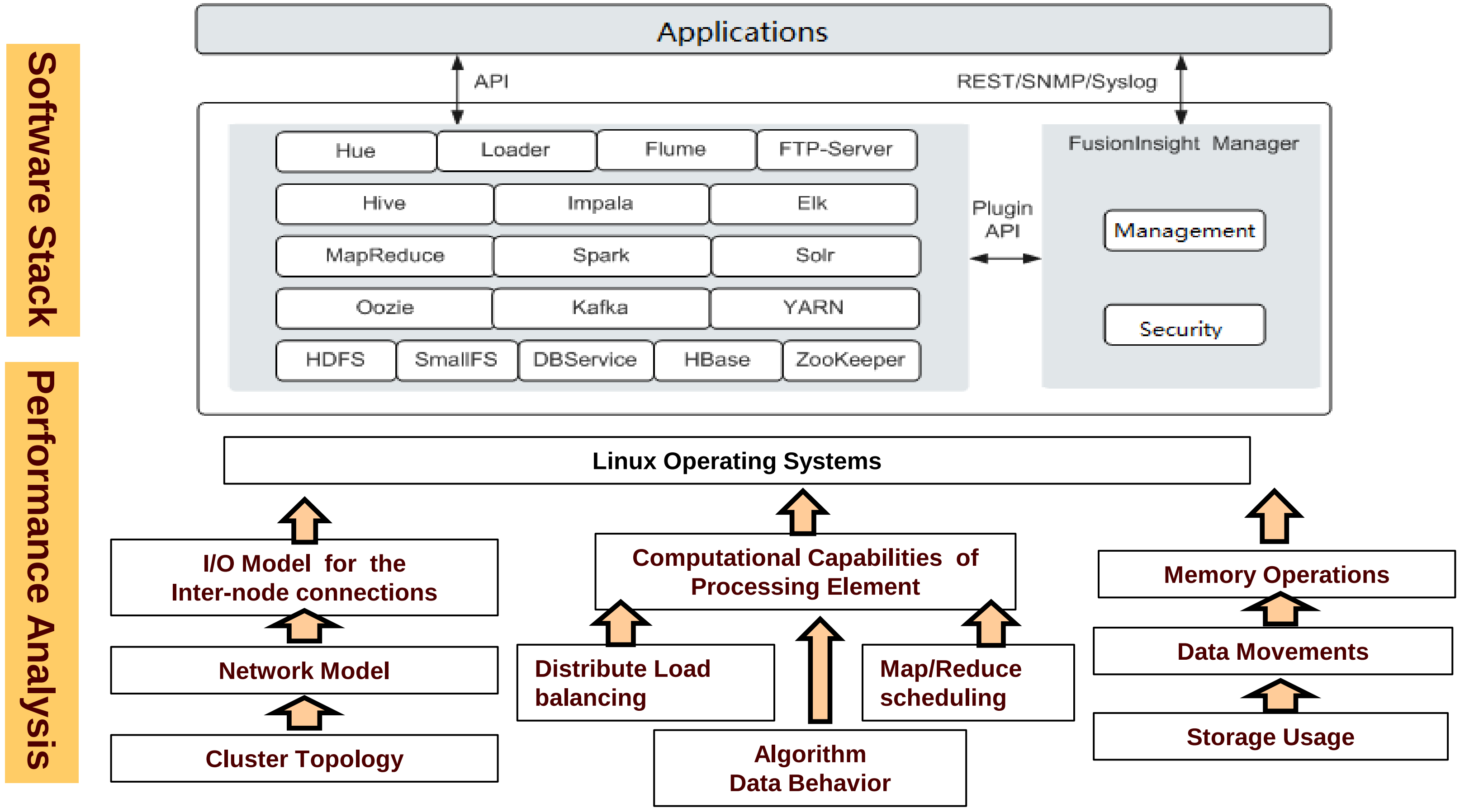
2. Compute and storage devices, as one of the core components of a data center system, need specially designed approaches to measure, evaluate and analyze their performance.

3. This talk introduces our methods to create the performance model based on workload characterization, algorithm level behavior tracing and capture, and software platform management.

4. The functionality and capability of our methodology have been validated through benchmarks and measurements performed on real data center system.

A FusionInsight System for Big Data

Performance Modeling and Analysis



Program Behavior Model

1. Correlates data from all the experiments based on the common sampling rate and common time line;
2. Analyzing performance parameters and the application’s (Data Sorting) behavior
3. Analyzing performance parameters and the application’s (Transactions) behavior
4. Analyzing performance parameters and the application’s (K-Means Clustering) behavior
5. Identified program characters and create leading markers;
6. Identified program segments and perform detailed analysis on each segment;
7. Developed a Model that can use data captured from the systems stimuli and explore bottlenecks and dependencies.

Performance Tuning

Application – platform design pattern, problem size, sub-domain partition, iterations.

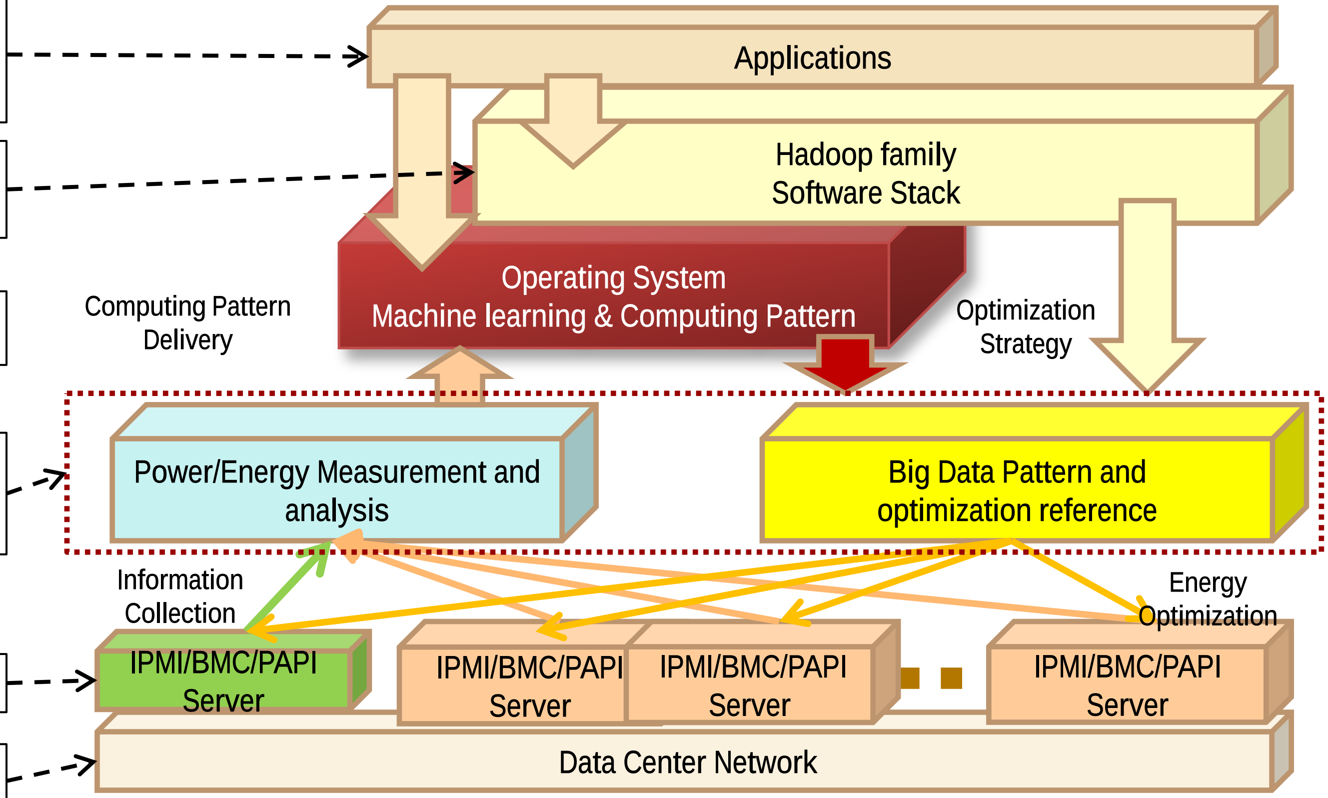
Hadoop configurations and parameters, scalability, parallel efficiency…

OS strategies, DFS solutions

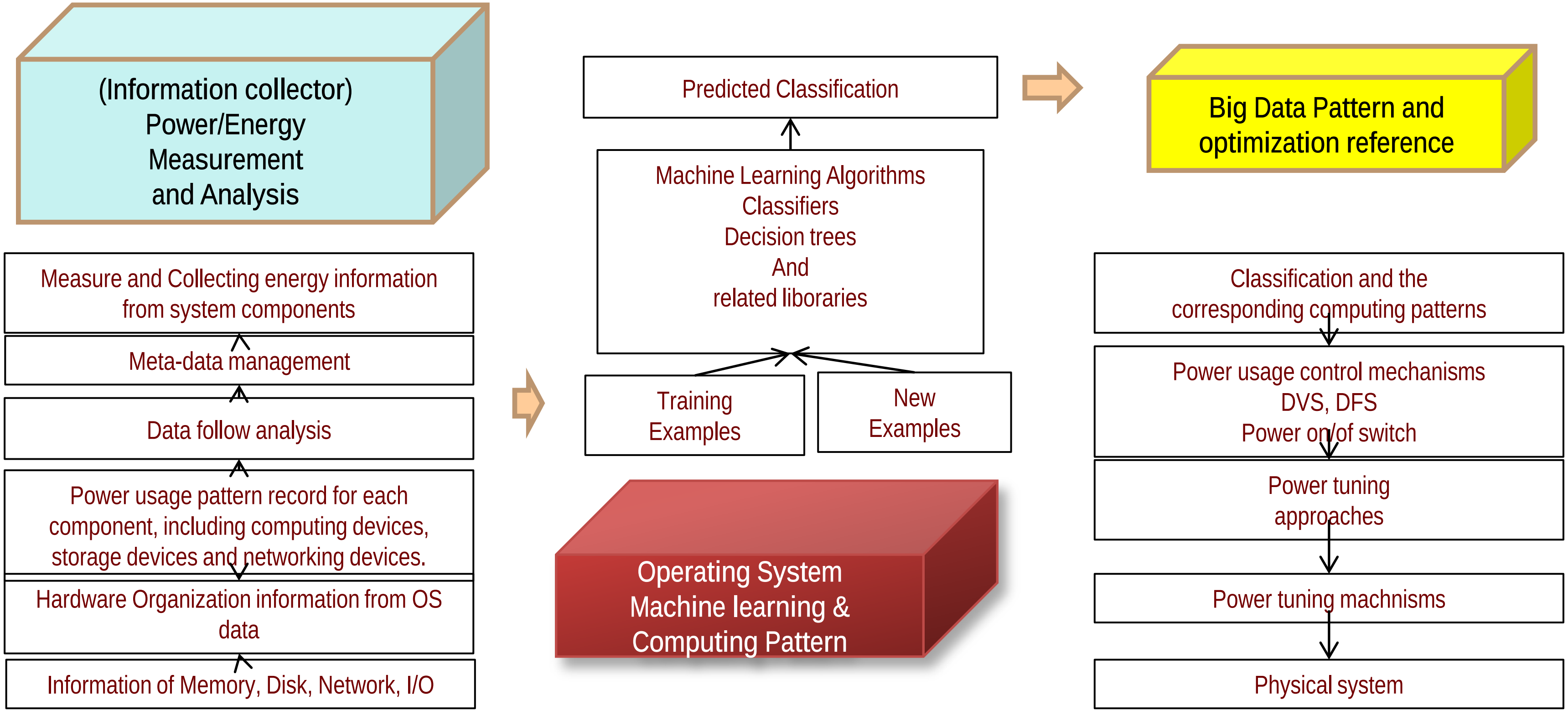
The Performance and Power optimization clock

Hardware Organization and Architecture

Memory, Disk, Network, I/O



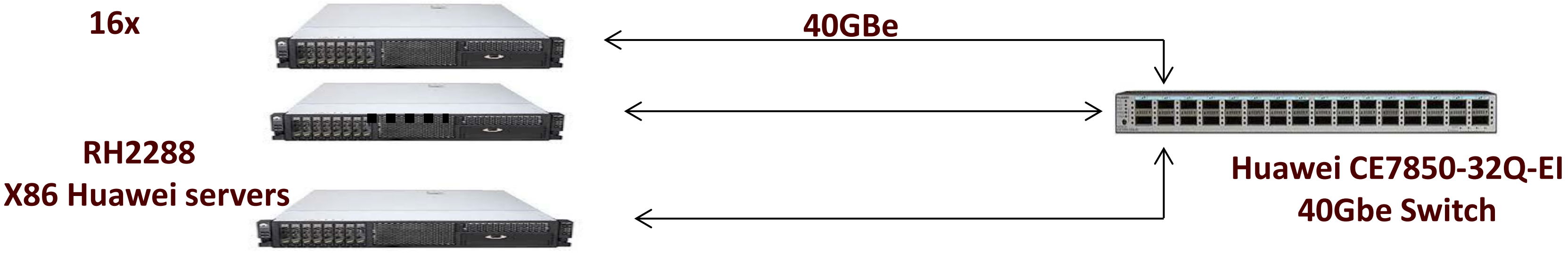
Power/Energy Tunings



Major Scenarios and the WL Characterization

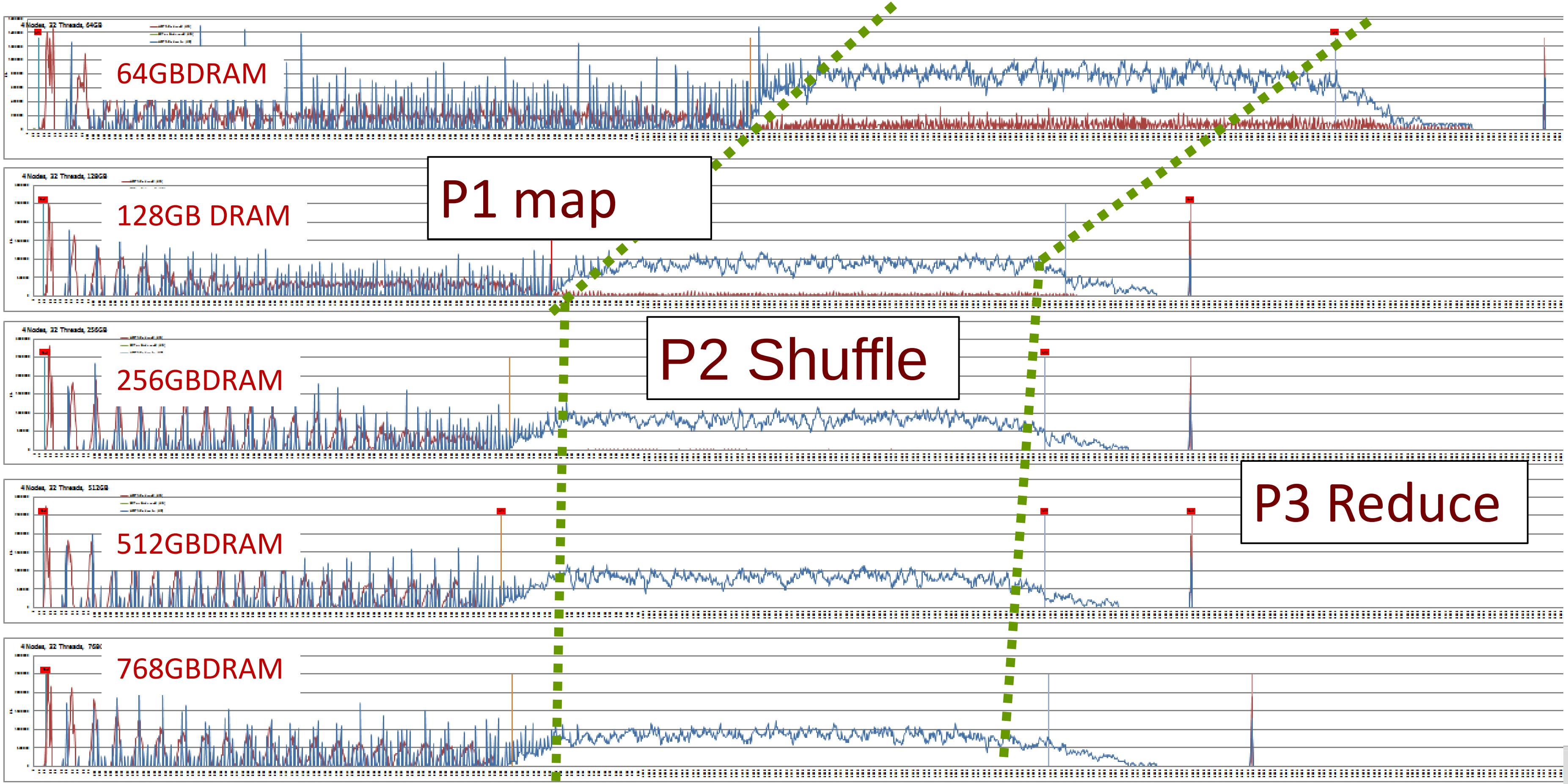
Application Scenarios	Workloads	Algorithms	Characterization	Computing methods	Benchmarks
Business Intelligent	TeraSort	Sorting Algorithms	I/O intensive Parallel Computing	Massive data Parallel sorting Matrix Computing	Hadoop Sort
	PageRank K-means clustering	Search Engines / Ranking pages	Multiple resource Utilization	Massive Matrix Computing	Hadoop/Spark +
	Naive Bayes/ Aprioi Algorithms	Decision Supports Queries Association Rules	Computing intensive	Parallel computing	Hadoop/Spark
VMall	Joint Query	List , Tables, Hash tables, Search Tech.	Computing Intensive	List , Tables, Hash tables, Search Tech.	Hive
Cloud+	Read/Write /Scan	Multiple Data Operations	I/O Cloud Different DBs	Parallel Database Access / Virtual machines / Queries	HBase CBTOOL, (SPEC Cloud)

Hadoop Cluster HW Configuration

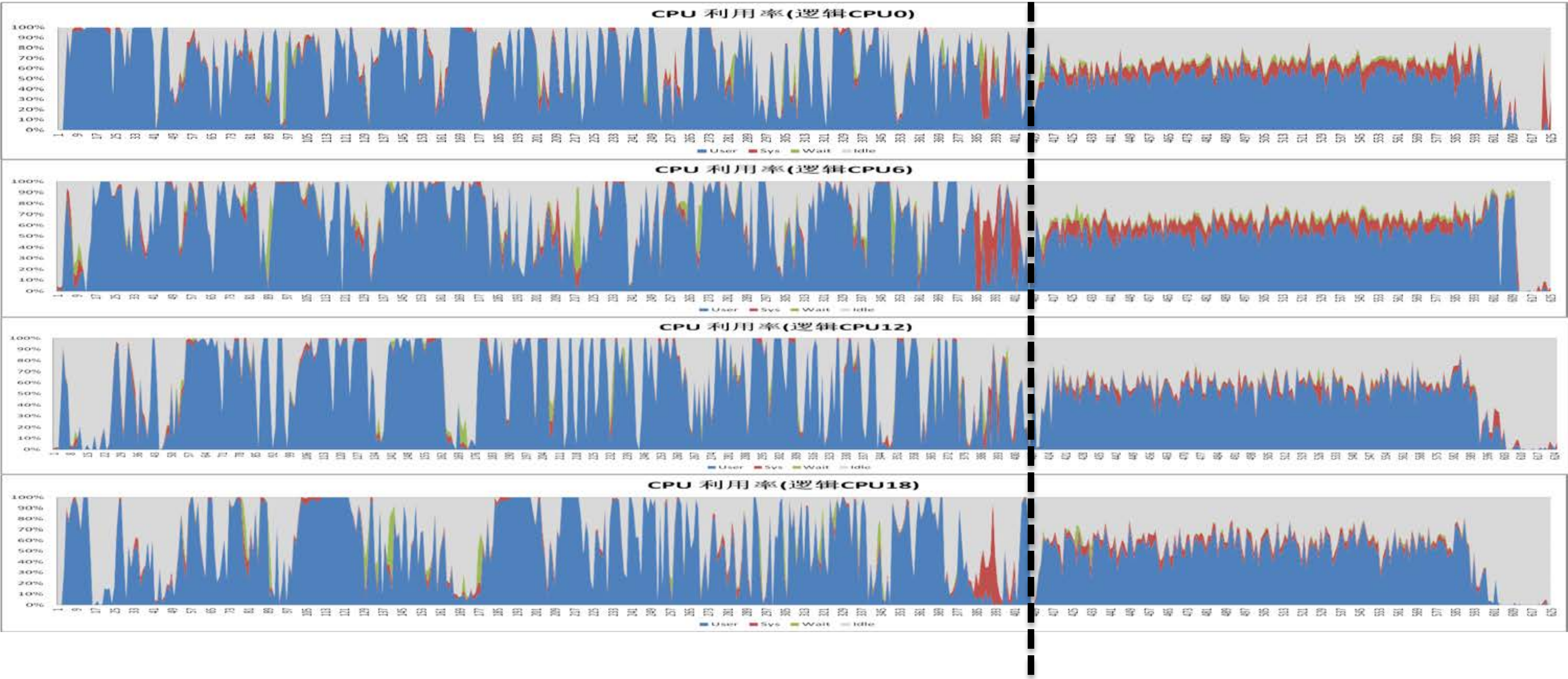


Servers:	16 x Huawei Tecal RH2288 v2 Server
Total Processors/Cores/Threads	32/256/512
Processor	2 x Intel® Xeon® Processor E5-2680 v2, 2.70 GHz, 20M L3
Memory	256GB
Storage Controller	1 x Symbios Logic MegaRAID SAS 2208
Storage Device	1 x 600GB 10K SAS HDD, 1 x 2.4TB Huawei ES3000 PCIe SSD Card
Network	1 x Mellanox ConnectX-3 Pro EN 40GbE SFP+CNA
Connectivity:	1 x Huawei CE7850-32Q-EI 40GigE

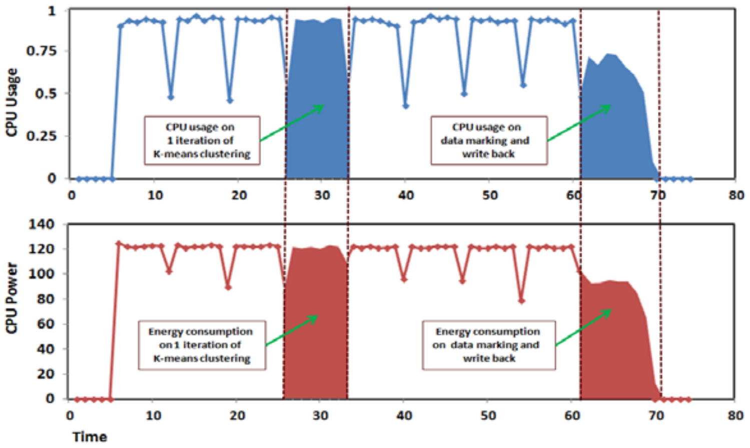
Memory Capacity Scalability

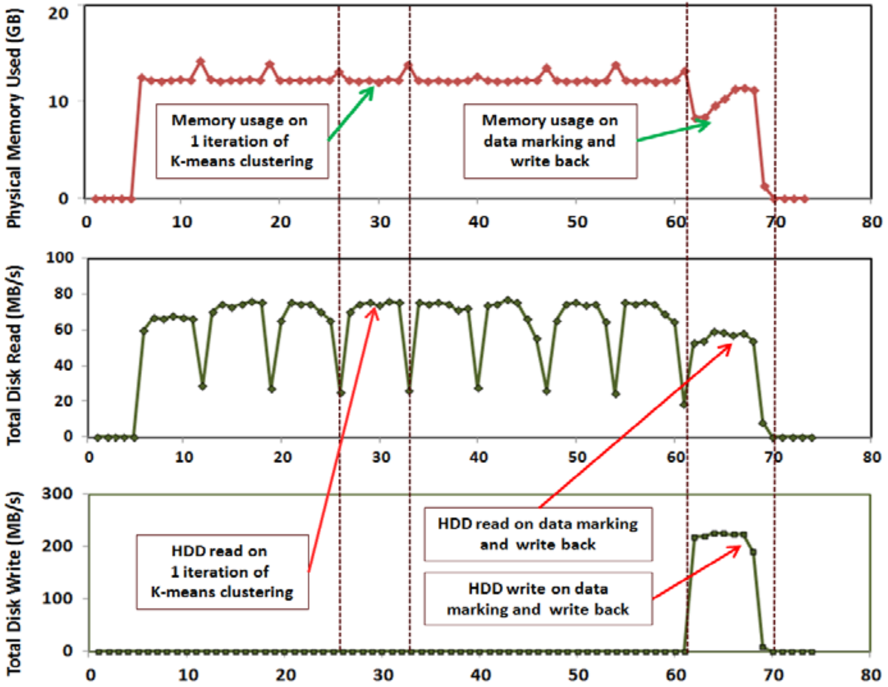


TeraSort CPU Utilization (24CPUs)

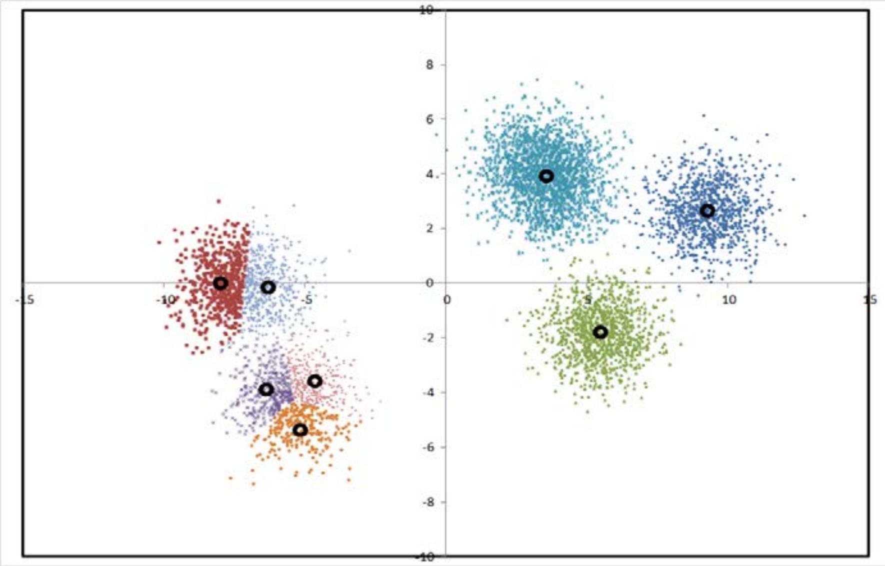


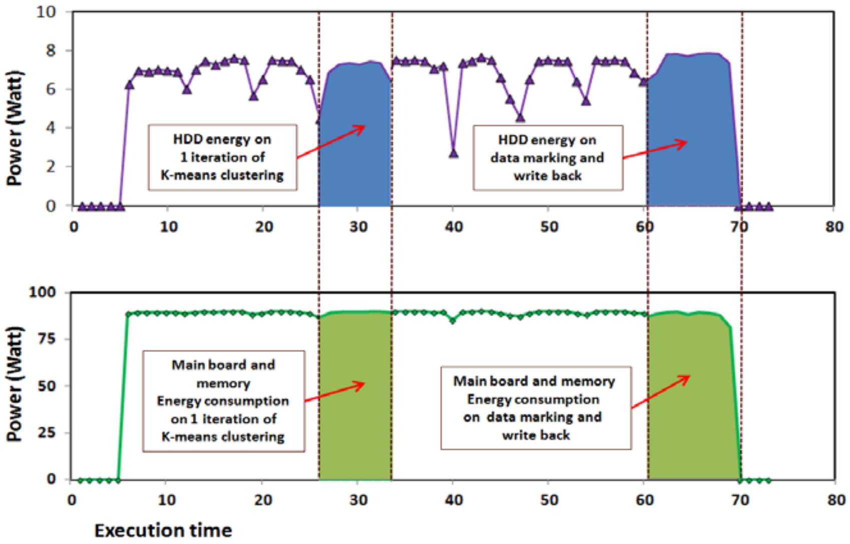
Algorithm Behaviors in K-Means Clustering



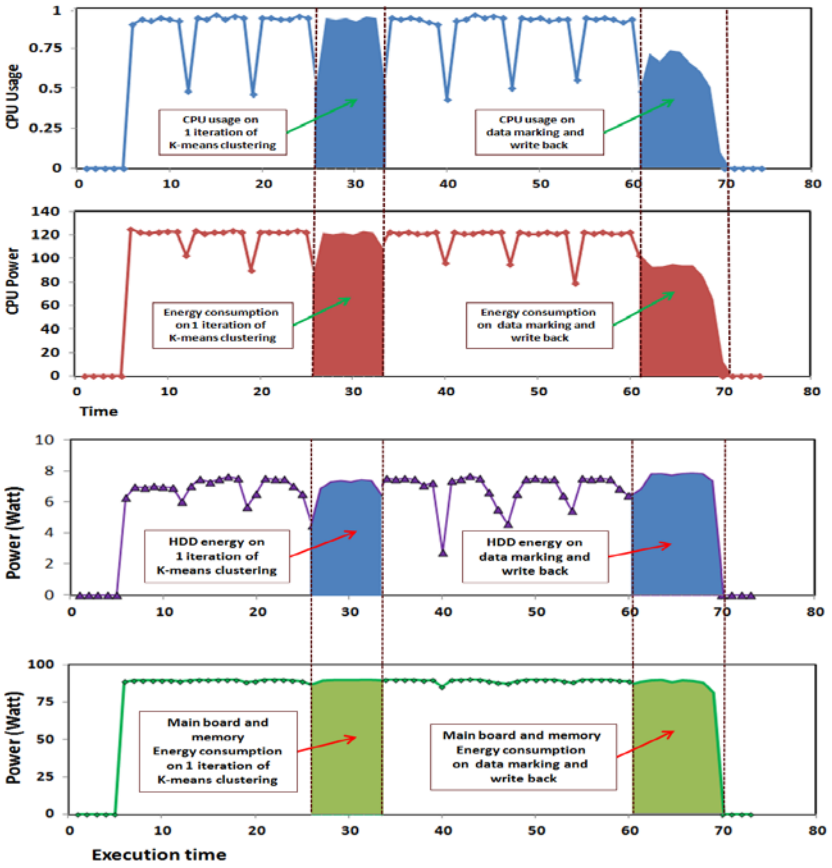


Algorithm Behaviors in K-Means Clustering





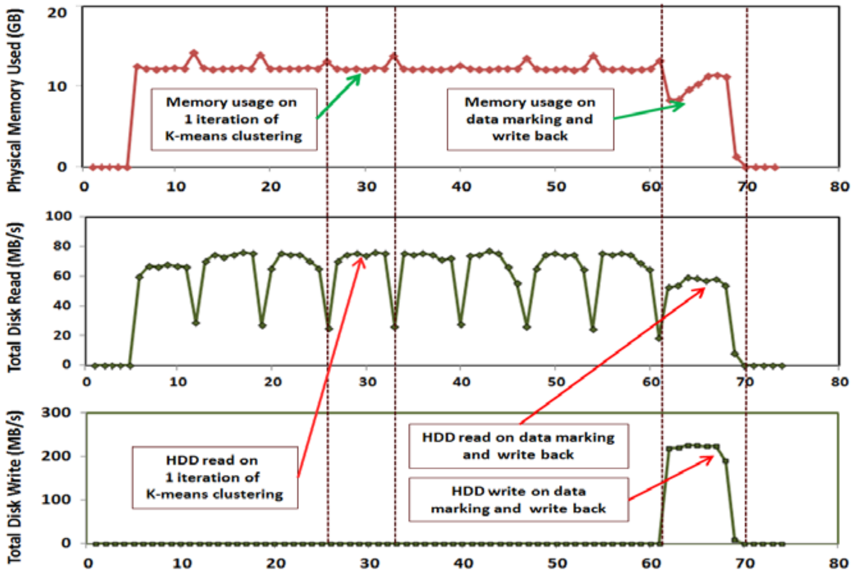
Characters of K-means Clustering



CPU usage and the corresponding power measurement results, when executing K-means clustering operations following the system configuration.

HDD and Main board (include memory) power measurement results, when executing K-means clustering operations following the system configuration.

Characters of K-means Clustering



K-Means Job Parameter	Results (Clustering Phase)	Results (Data Writing Phase)
CPU Power (average)	122.21W	98.67W
Main board Power (incl. memory)	72.14W	72.09W
HDD Energy	7.03W	7.17W
Results		
Data Size / Watt	126.04M/Watt	158.13M/Watt

The Physical memory usage, total disk read and write measurement results, when executing K-means clustering operations following the system configuration.

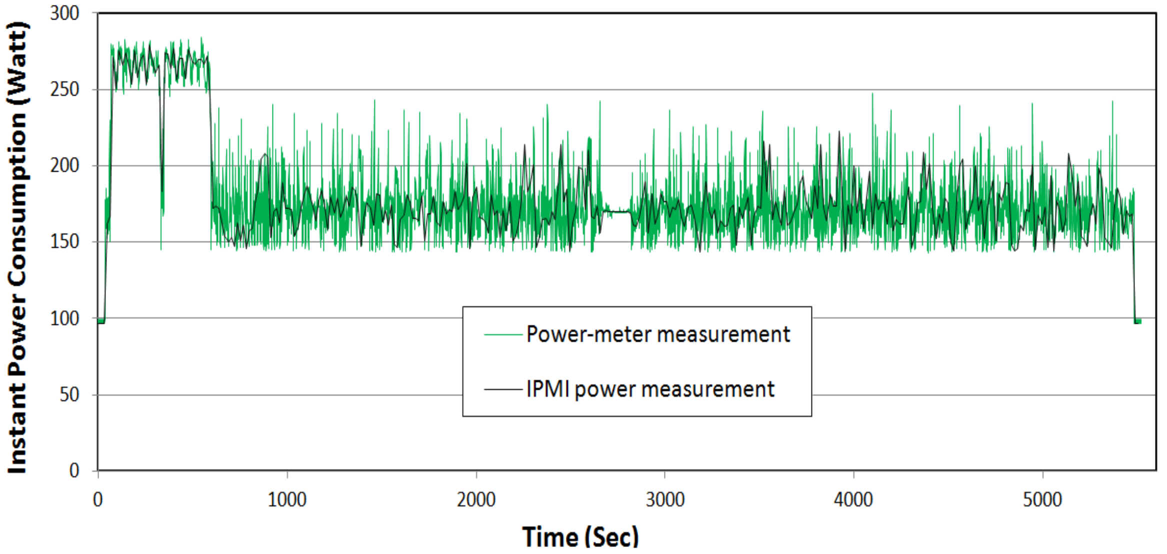
The average of power measurement results for k-means clustering at clustering operation phase and data writing phase. The throughput and power performance of the target platform is concluded.

Big Bench Performance Chart

Query	Input Datatype	Processing Model	Query	Input Datatype	Processing Model
#1	Structured	Java MR	#16	Structured	Java MR (OpenNLP)
#2	Semi-Structured	Java MR	#17	Structured	HiveQL
#3	Semi-Structured	Python Streaming MR	#18	Unstructured	Java MR (OpenNLP)
#4	Semi-Structured	Python Streaming MR	#19	Structured	Java MR (OpenNLP)
#5	Semi-Structured	HiveQL	#20	Structured	Java MR (Mahout)
#6	Structured	HiveQL	#21	Structured	HiveQL
#7	Structured	HiveQL	#22	Structured	HiveQL
#8	Semi-Structured	HiveQL	#23	Structured	HiveQL
#9	Structured	HiveQL	#24	Structured	HiveQL
#10	Unstructured	Java MR (OpenNLP)	#25	Structured	Java MR (Mahout)
#11	Unstructured	HiveQL	#26	Structured	Java MR (Mahout)
#12	Semi-Structured	HiveQL	#27	Unstructured	Java MR (OpenNLP)
#13	Structured	HiveQL	#28	Unstructured	Java MR (Mahout)
#14	Structured	HiveQL	#29	Structured	Python Streaming MR
#15	Structured	Java MR (Mahout)	#30	Semi-Structured	Python Streaming MR

Most OEMs do provide the capability to implement power measurement.

In the latest generation IPMI can measure Power for the Entire system power,
CPU subdomain power
Memory Subdomain power.



Performance Analysis on Cloud

SPEC Cloud 2016

CloudBecch (CBTOOL)

- CloudBench kit， CloudOS, Virtual machine perofmrnace and reports.
- 4 date bases performance： Cassandra； Hbase； PNUTS； MySQL
- Uniform distribution; zipfian distribution ; popularity distribution； Latest distribution; multinomial distribution.

Also Benchmarked

- SPECvirt
- Refer: Cloudsuite, CBTooL

From Application to Performance

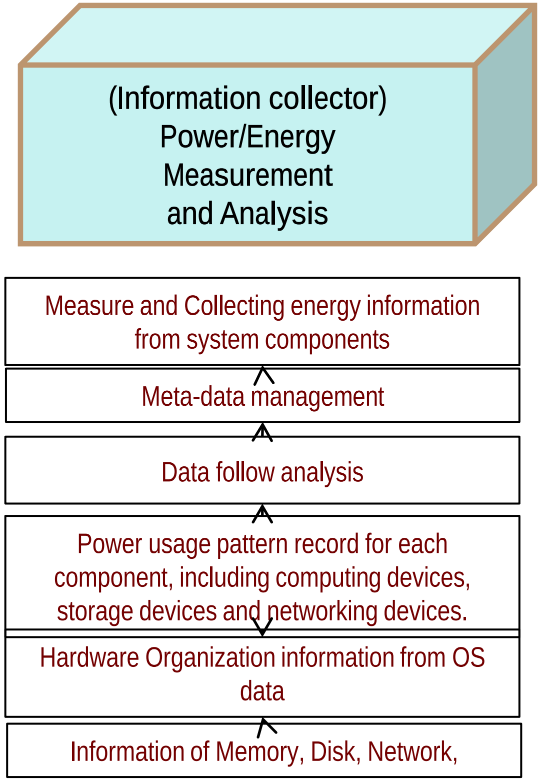


TABLE I
POWER FEATURES OF EACH PE

PE Components	MFlops /W	CPU Frequency Scaling (GHz)	GPU NV/Memory Frequency (MHz)
Intel QX9650 + GF 8800 GTS/512	83.5	2	650 / 972
	73.6	3	650/ 972
Intel QX9650 + 2 GF 8800 GTS/512	76.9	2	650 / 972
	75.1	3	650 / 972
QX9650 (Share Work)+ 2 GF 8800 GTS/512	70.2	2	650 / 972
	71.3	3	650 / 972
Intel i7 + GF 8800 GTS /640M	81.3	2.67	513 / 792
	70.8	1.60	513 / 792
Intel QX9650	2.4	2	-
	1.7	3	-
Intel i7	2.3	2.67	-
	4.4	1.60	-

$$T_{CPU^{3GHz}_{QX9650} \times GPU_{8800}} = 2N^3 / GFLOPS_{CPU^{3GHz}_{QX9650} \times GPU_{8800}}$$

$$P_{CPU^{3GHz}_{QX9650} \times GPU_{8800}} = \frac{GFLOPS_{CPU^{3GHz}_{QX9650} \times GPU_{8800}}}{GFLOPSperWatt_{CPU^{3GHz}_{QX3650} \times GPU_{8800}}}$$

$$E(w) = T_{CPU^{3GHz}_{QX9650} \times GPU_{8800}} \cdot P_{CPU^{3GHz}_{QX9650} \times GPU_{8800}}$$

$$P(w) = \sum_{i=1}^{N} P^{i}_{Compute}(w^{i}) + \sum_{j=1}^{M} P^{j}_{Storage}(w^{j}) + P_{network}(w)$$

Energy Tuning

Controlled by OS Through IPMI or other mechanisms

Controlled by Fan in the server, OR the cooling system in Datacenter

Controlled by higher level software

Components	Frequencies	Voltages	Temperature	Usage	Total
CPUs (processors, cores)	f (Frequency steps)	V (VIDs)	T	$P = ACV^2f$	
Memories (RAM, NVM)	f (Frequency steps)	V (VIDs)	T	$P = ACV^2f$	
GPUs (cores)	f (Frequency steps)	V (VIDs)	T	$P = ACV^2f$	
Board (buses, slots)	f (Frequency steps)	V (VIDs)	T		
Buses (PCI, Optical)	f (Frequency steps)	V (VIDs)	T		
Disk (HDD, SSD)	f (Frequency steps)	V (VIDs)	T		
I/O (Network, DISK I/O)	f (Frequency steps)	V (VIDs)	T		

$$Power = DynamicPower + ShortCircuitPower + LeakagePower$$

where A is the active fraction of the gates in the CMOS element that are switching each clock cycle, C is the total capacitance driven by all the gate outputs in the CMOS element (thereby making AC the capacitance being switched per clock cycles), V is the operating voltage of the CMOS element, and f is the frequency of the clock.

Thank you!