A Testbed for Datacenter Computing

詹剑锋 (@jfzhan)
中国科学院计算技术研究所
先进计算机系统试验室

Hadoop in China 2011
2011.12.2

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Outline

- **What is datacenter computing?**
- Motivation for a new testbed
- Current status of the testbed
- Benchmarks

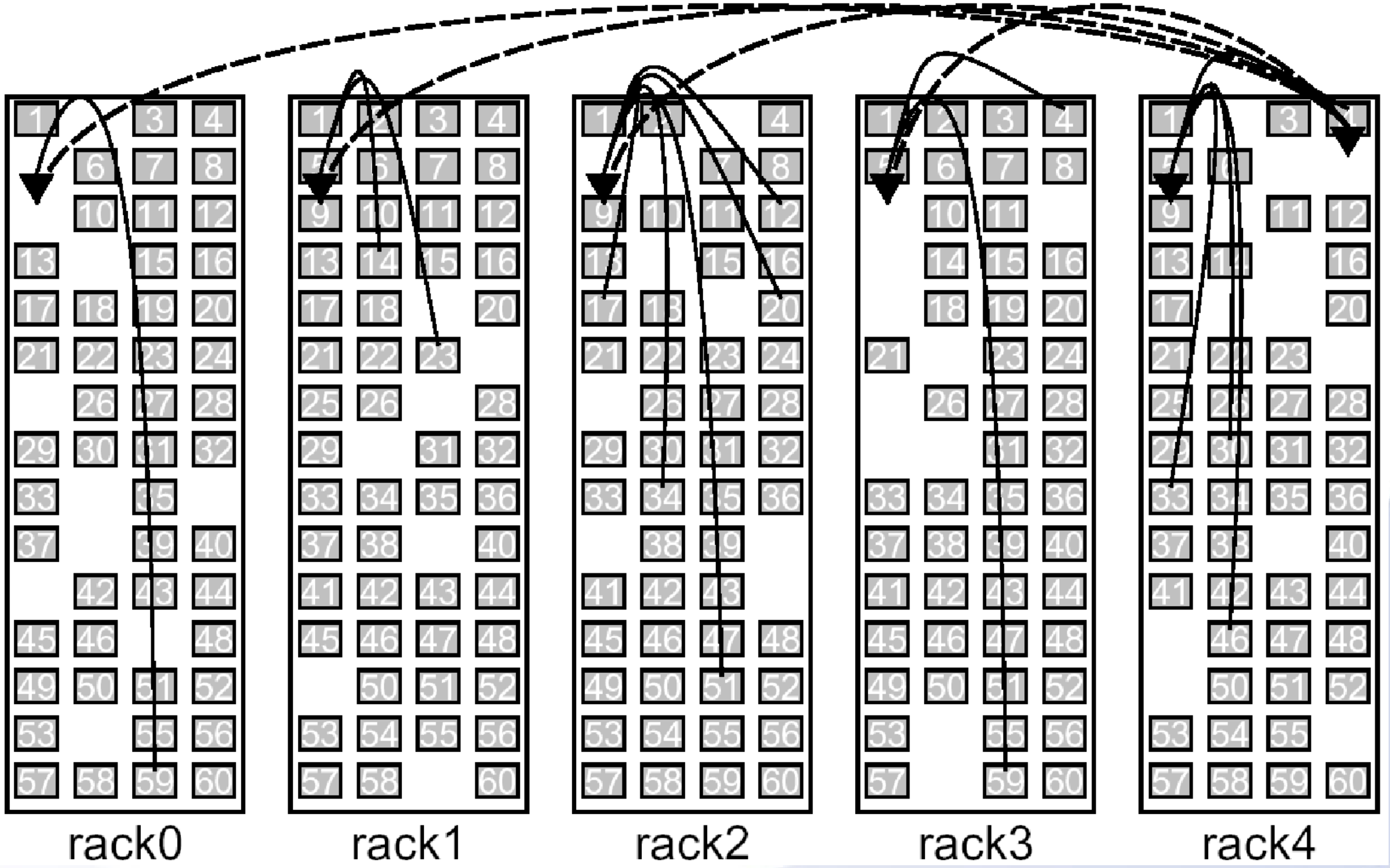
Datacenter hosting services

- Free services are ubiquitous and pervasive
  - Computing resources
    - Amazon EC2
  - Information
    - Google, Ebay, Baidu, Tencent, Taobao, and……
  - Knowledge
    - Cost-effective solutions

Datacenter racks
[copyright wikipedia.com]

Google Datacenter
[copyright google.com]

A typical Google datacenter.



Data streaming to/from major sources

- Twitter "fire hose"
  - 50M tweets/day * (140 + 54)B/tw = 10GB/day = 1Mb/sec
- Google search (estimate)
  - 2.36 Mb/sec input queries, 100 Mb/sec out
- LHC (particle accelerator)
  - 15 PB/year = 41 TB/day = 1.712 TB/h =4.8 Gb/sec
- Email (non spam)
  - Gmail 18 emails/day = 75 TB/day = 7Gb/sec
- SKA (radio telescope)
  - Raw data = 960PB/day, Final processed data = 10 Gb/sec
- Zynga (social network game）
  - 1PB/day = 92 Gb/sec

Courtesy. Dr. Dennis Gannon PDAC-11 Keynote

An informal definition of datacenter computing (*DC*)

- Data-intensive computing or services for the masses hosted on datacenters.
  - Data-intensive services
    - Massive concurrent requests, e.g., million
    - Data scales varying from TB to PB
  - Data-intensive computing (data analysis).
    - A large amount of **Jobs composed of independent tasks** (loosely coupled)
    - Data scales varying from TB to PB
      - Approaching EB in near future

DC vs. High end HPC

	Workload analysis	Parallelism	Reliability	Metrics
DC	● **Loosely coupled** ● Workload churn	Ample parallelism	No checkpoint need for single failures. Reliability requirements depend upon the nature of data.	High throughput
High end HPC	● **Tightly coupled**: a single job with huge resource demand. ● Depend on collective communication.	Difficult to exploit parallelism.	Checkpoint of a whole application for a single failure.	The turnaround time

J. Zhan, L. Wang and N. Sun, Performance Evaluation of a Datacenter Computer ,Communication of CCF. July, 2011.

Outline

- What is datacenter computing?
- **Motivation for a new testbed**
- Current status of the testbed
- Benchmarks

What is a testbed?

- A collection of connected machines?
  - – Yes, maybe geographically distributed

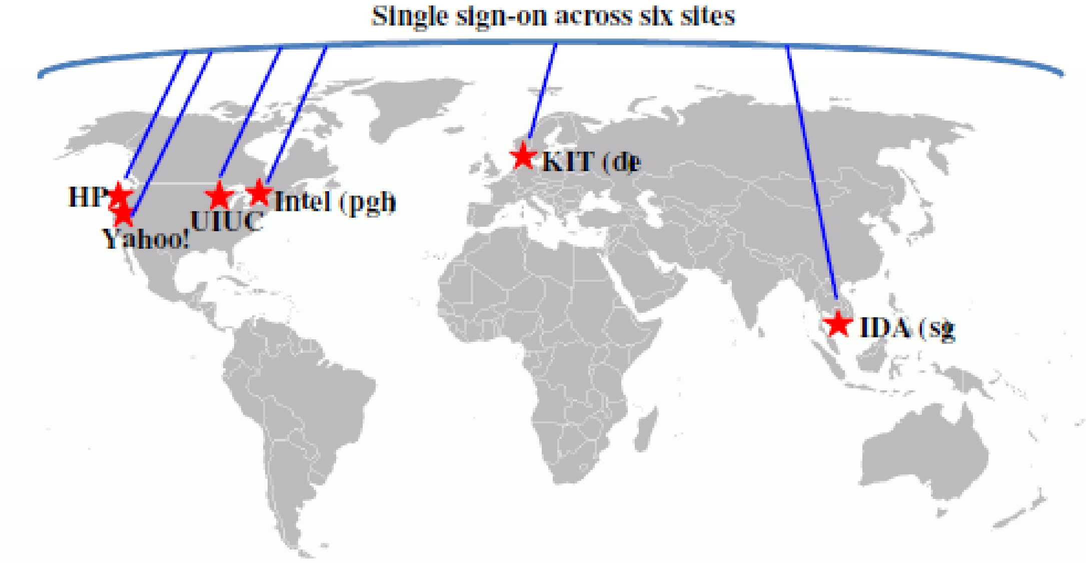
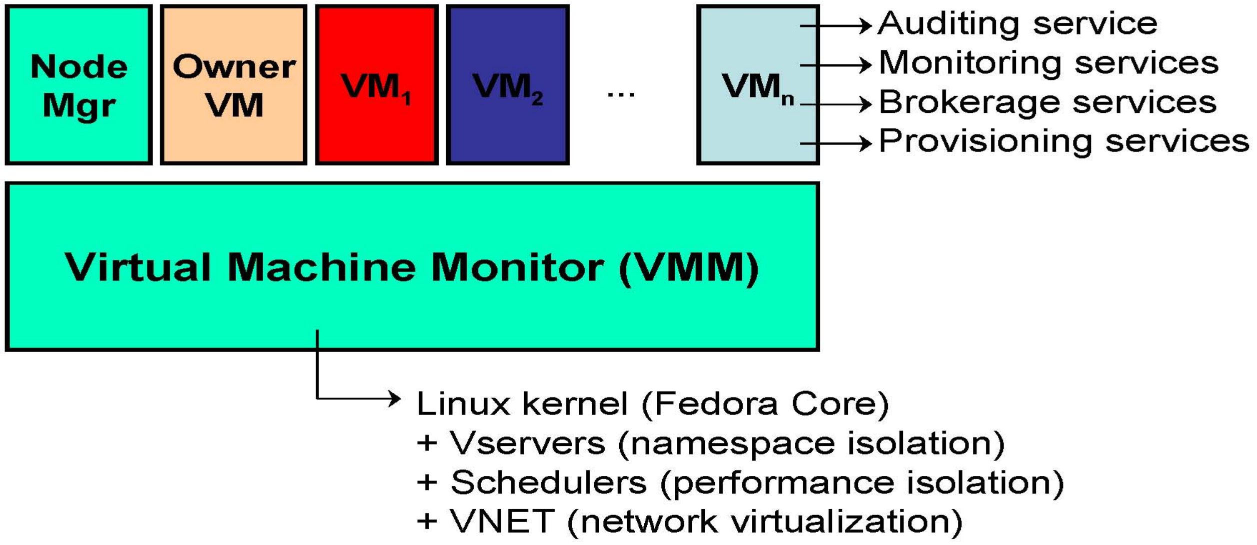


Figure 1. Open Cirrus testbed circa Q1 2009.

Source: Open CirrusTM Cloud Computing Testbed: Federated Data Centers for Open Source Systems and Services Research

What is a testbed?

- A collection of software stack for supporting experiments?
  - e.g., PlanetLab supports distributed virtualization.


Node Mgr
Owner VM
VM1
VM2
...
VMn
Auditing service
Monitoring services
Brokerage services
Provisioning services
Virtual Machine Monitor (VMM)
Linux kernel (Fedora Core)
+ Vservers (namespace isolation)
+ Schedulers (performance isolation)
+ VNET (network virtualization)

Source: Larry Peterson, PlanetLab: Evolution vs Intelligent Design in Global Network Infrastructure

The summary of testbed projects

Characteristics	Testbeds				
	Open Cirrus	TerraGrid	PlanetLab	EmuLab	Open Cloud Consortium
Type of research	Systems & services	Scientific applications	Systems and services	Systems	interoperability across clouds using open APIs
Approach	Federation of heterogeneous data centers	Multi-site hetero clusters super comp.	A collection of nodes hosted by research instit.	A single-site cluster with flexible control	Multi-site
Participants	HP, Intel, IDA, KIT, UIUC, Yahoo!	Many univ. & organizations	Many univ & organizations	University of Utah 4	4 centers

Most important requirements for a DC testbed

- Data: from TB or PB
  - Real data, not synthetic data
- Applications
  - State-of-the-art algorithms
- User access traces
  - A search engine
    - Query rate variance
    - Query locality
    - Query frequencies
      - Some search terms are hot.

Most important requirements for a DC testbed

Unfortunately, no testbed provides big data, application, and real user access traces (**live workloads**).

Data lock-in issue

- Internet service companies indeed own big data, and real applications.

- Commercial confidentiality
  - They would not like to share data, applications with research communities.

- Current open data projects
  - Limited purposes: algorithms researches.

Our targets

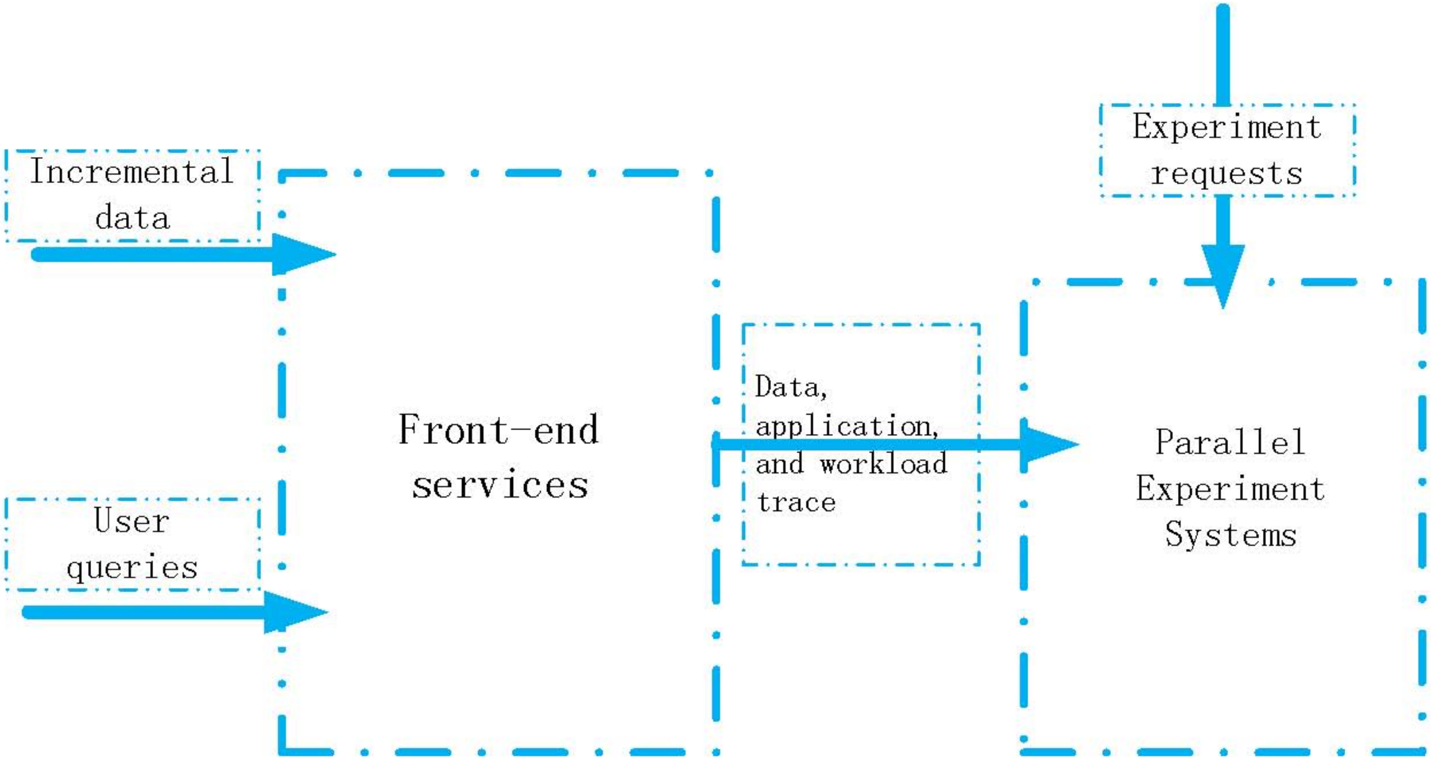
- Build a testbed, providing real big data, applications, and user access traces for research communities.
  - Architecture
  - OS/VM
  - Hadoop-like systems
  - Data management
  - Reliability
  - Power management

- Promote innovations
  - Support Web-based experiments for innovative technologies

Outline

- What is Datacenter Computing?
- Motivation for a New Testbed
- **Current Status of the Testbed**
- Benchmarks

The testbed architecture

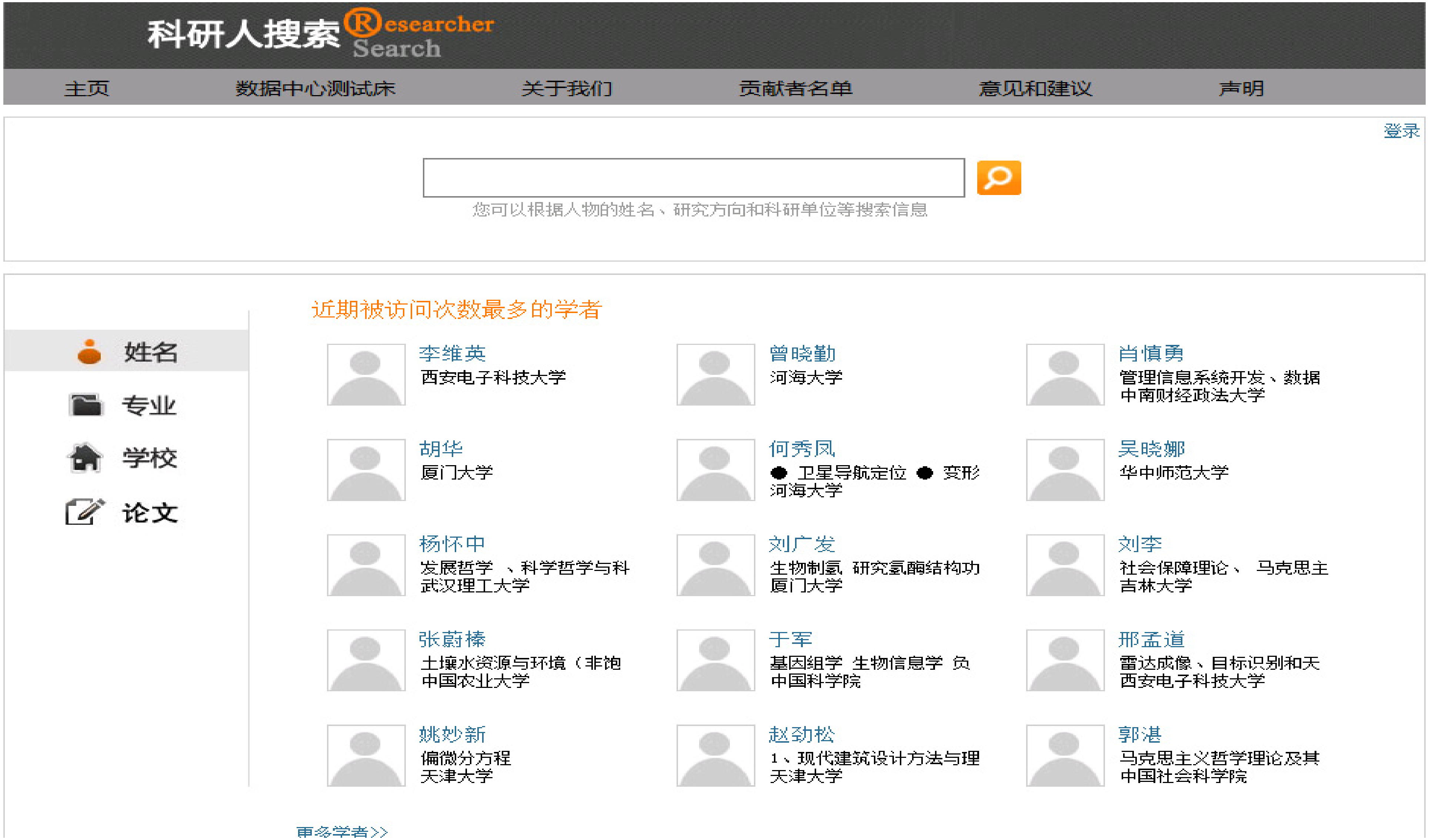


An ideal application

- Data-intensive
  - TB or PB
- A challenging application
  - **E.g. machine reading of the World Wide Web**
- Valuable services attracts more searches

ProfSearch (http://prof.ncic.ac.cn), online since Sep. 2011.

Front-end service



您可以根据人物的姓名、研究方向和科研单位等搜索信息

近期被访问次数最多的学者

李维英
西安电子科技大学

曾晓勤
河海大学

肖慎勇
管理信息系统开发、数据
中南财经政法大学

胡华
厦门大学

何秀凤
● 卫星导航定位 ● 变形
河海大学

吴晓娜
华中师范大学

杨怀中
发展哲学 、科学哲学与科
武汉理工大学

刘广发
生物制氢 研究氢酶结构功
厦门大学

刘李
社会保障理论、 马克思主
吉林大学

张蔚榛
土壤水资源与环境（非饱
中国农业大学

于军
基因组学 生物信息学 负
中国科学院

邢孟道
雷达成像、目标识别和天
西安电子科技大学

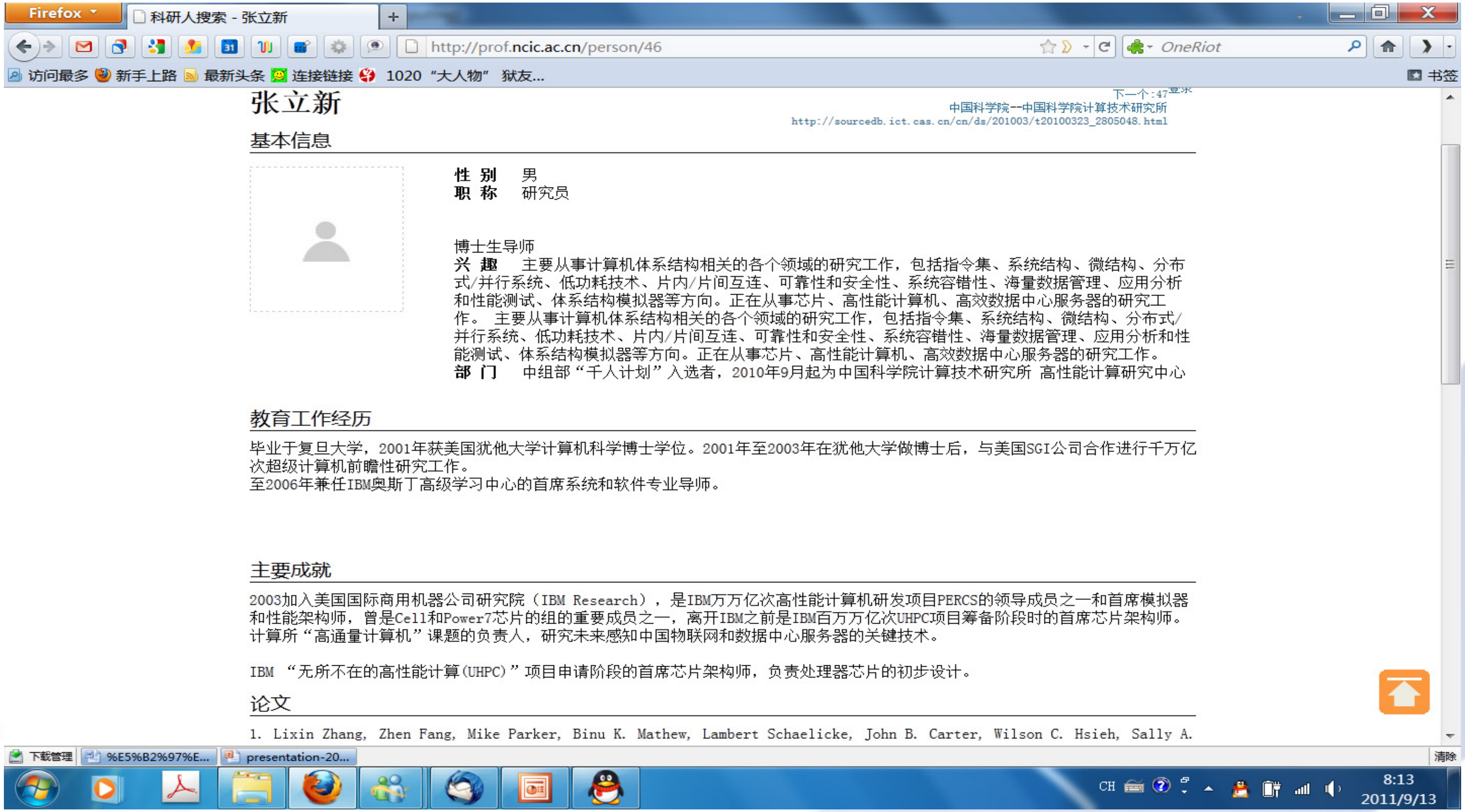
姚妙新
偏微分方程
天津大学

赵劲松
1、现代建筑设计方法与理
天津大学

郭湛
马克思主义哲学理论及其
中国社会科学院

更多学者>>

Information extraction



下一个:47登录

中国科学院--中国科学院计算技术研究所
http://sourcedb.ict.cas.cn/cn/ds/201003/t20100323_2805048.html

张立新

基本信息

性 别　男
职 称　研究员

博士生导师
兴 趣　主要从事计算机体系结构相关的各个领域的研究工作，包括指令集、系统结构、微结构、分布式/并行系统、低功耗技术、片内/片间互连、可靠性和安全性、系统容错性、海量数据管理、应用分析和性能测试、体系结构模拟器等方向。正在从事芯片、高性能计算机、高效数据中心服务器的研究工作。 主要从事计算机体系结构相关的各个领域的研究工作，包括指令集、系统结构、微结构、分布式/并行系统、低功耗技术、片内/片间互连、可靠性和安全性、系统容错性、海量数据管理、应用分析和性能测试、体系结构模拟器等方向。正在从事芯片、高性能计算机、高效数据中心服务器的研究工作。
部 门　中组部“千人计划”入选者，2010年9月起为中国科学院计算技术研究所 高性能计算研究中心

教育工作经历

毕业于复旦大学，2001年获美国犹他大学计算机科学博士学位。2001年至2003年在犹他大学做博士后，与美国SGI公司合作进行千万亿次超级计算机前瞻性研究工作。
至2006年兼任IBM奥斯丁高级学习中心的首席系统和软件专业导师。

主要成就

2003加入美国国际商用机器公司研究院（IBM Research），是IBM万万亿次高性能计算机研发项目PERCS的领导成员之一和首席模拟器和性能架构师，曾是Cell和Power7芯片的组的重要成员之一，离开IBM之前是IBM百万万亿次UHPC项目筹备阶段时的首席芯片架构师。计算所“高通量计算机”课题的负责人，研究未来感知中国物联网和数据中心服务器的关键技术。

IBM “无所不在的高性能计算(UHPC)”项目申请阶段的首席芯片架构师，负责处理器芯片的初步设计。

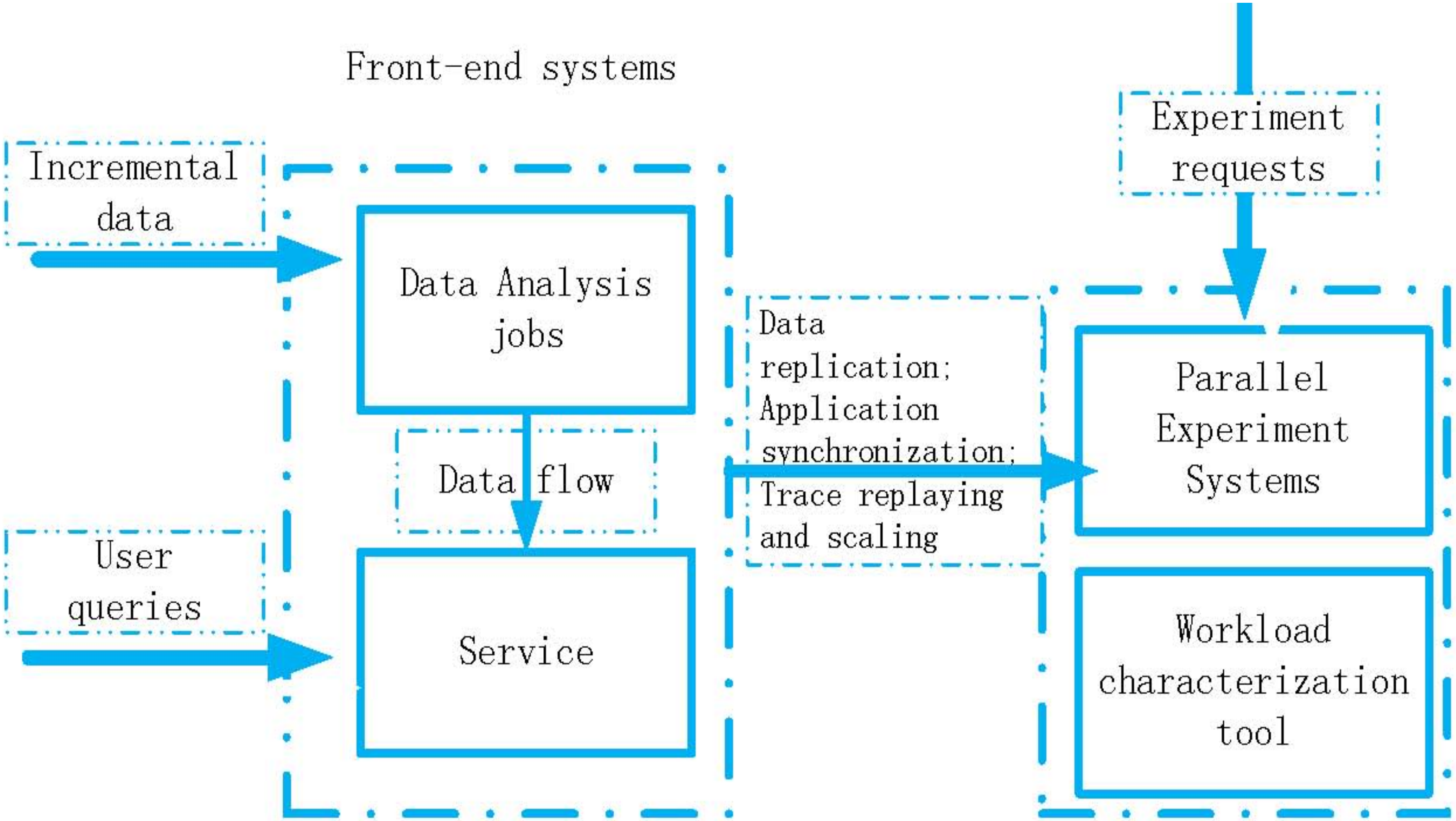
论文

1. Lixin Zhang, Zhen Fang, Mike Parker, Binu K. Mathew, Lambert Schaelicke, John B. Carter, Wilson C. Hsieh, Sally A.

Milestones

- ProfSearch 1.0, 2011.9
  - Baseline services
- ProfSearch 1.5, 2011.12
  - Incremental data processing
  - Autonomic management
  - Upgraded algorithms
- ProfSearch 2.0, 2012.2
  - Worldwide scholars from all disciplines
  - Papers
- ProfSearch 3.0, 2012.12
  - Full-fledged services

The detail of the testbed


Front-end systems
Incremental data
Data Analysis jobs
Data flow
User queries
Service
Data replication; Application synchronization; Trace replaying and scaling
Experiment requests
Parallel Experiment Systems
Workload characterization tool

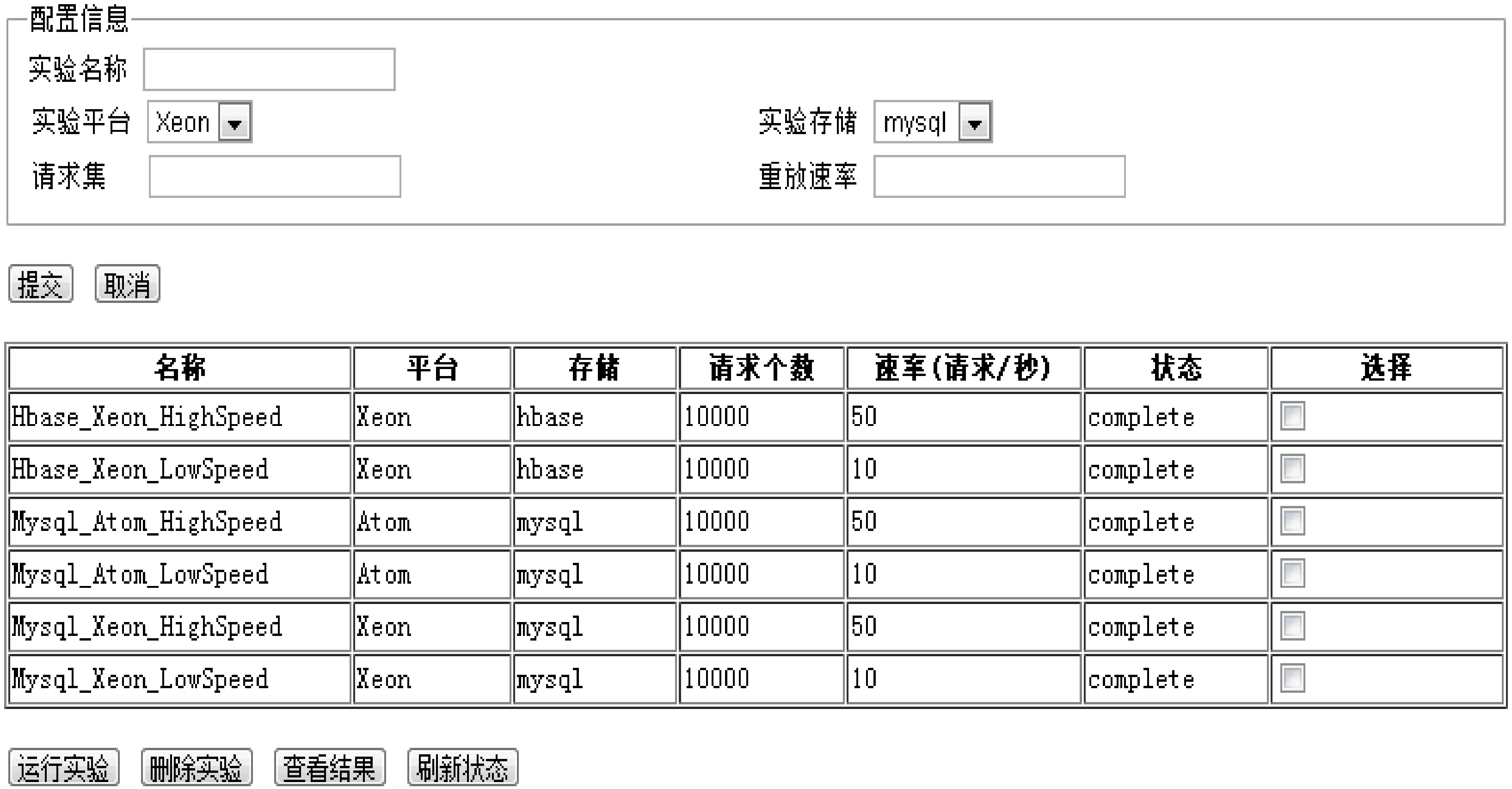
Main workloads

- Incremental data analysis jobs
- Search
  - File system-based
  - Database-based
- Web server
- Database
- NoSQL
  - Memcached
  - BigTable

Parallel experiments

- Users specify
  - Workloads
  - Data sets
  - Optional hardware configurations
  - Workload traces
    - Scaling factors
- Upload the tested systems
  - E.g., system or VM images
- Perform several experiments simultaneously

The snapshot of the current system



实验名称			
实验平台	Xeon	实验存储	mysql
请求集		重放速率	

名称	平台	存储	请求个数	速率(请求/秒)	状态	选择
Hbase_Xeon_HighSpeed	Xeon	hbase	10000	50	complete	□
Hbase_Xeon_LowSpeed	Xeon	hbase	10000	10	complete	□
Mysql_Atom_HighSpeed	Atom	mysql	10000	50	complete	□
Mysql_Atom_LowSpeed	Atom	mysql	10000	10	complete	□
Mysql_Xeon_HighSpeed	Xeon	mysql	10000	50	complete	□
Mysql_Xeon_LowSpeed	Xeon	mysql	10000	10	complete	□

Parallel evaluation experiments

实验配置

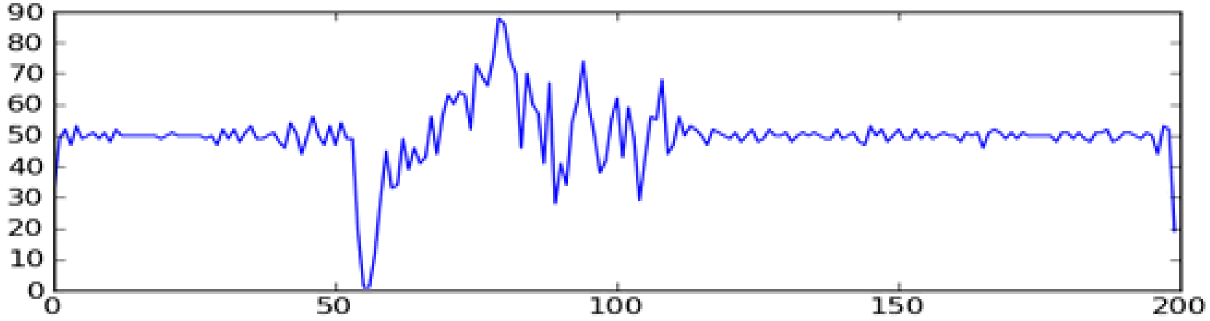
实验名称：Hbase_Xeon_HighSpeed
实验平台：Xeon
实验存储：hbase
请求集： 10000
重放速率：50

实验平台配置

CPU类型：Intel(R) Xeon(R) E5310
CPU个数：4
CPU频率：1600.136 MHZ
内存容量：3.86716 GB
操作系统：Linux
内核版本：2.6.34.7
Gcc版本：4.1.2

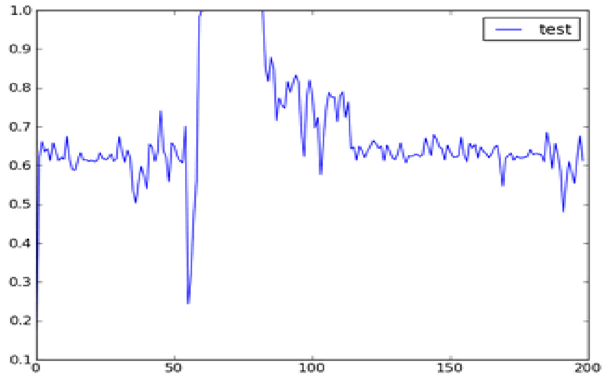
实验结果

请求强度： 49.9555reqs/s
持续时间： 200.038s
吞吐率：49.5856reqs/s
平均响应时间：0.737728s

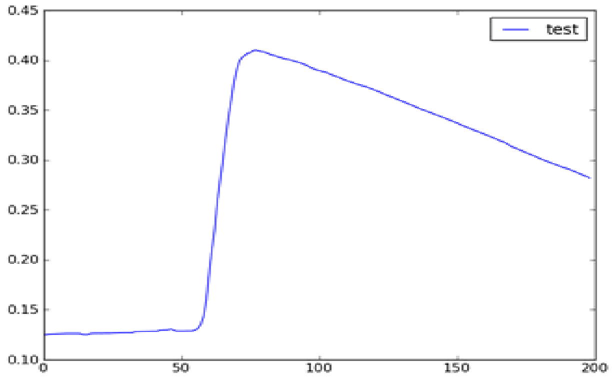


用户请求速率

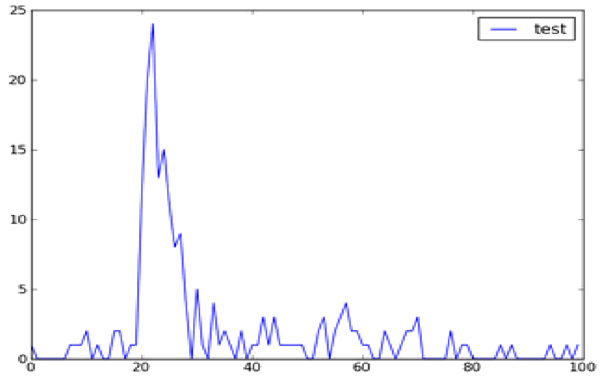
Parallel evaluation experiments



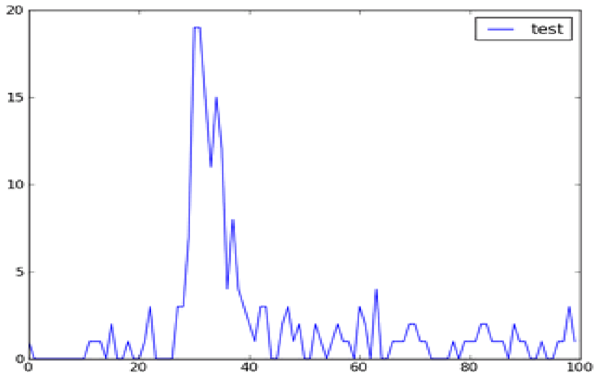
cpu利用率



内存利用率



每秒的收包个数



每秒的发包个数

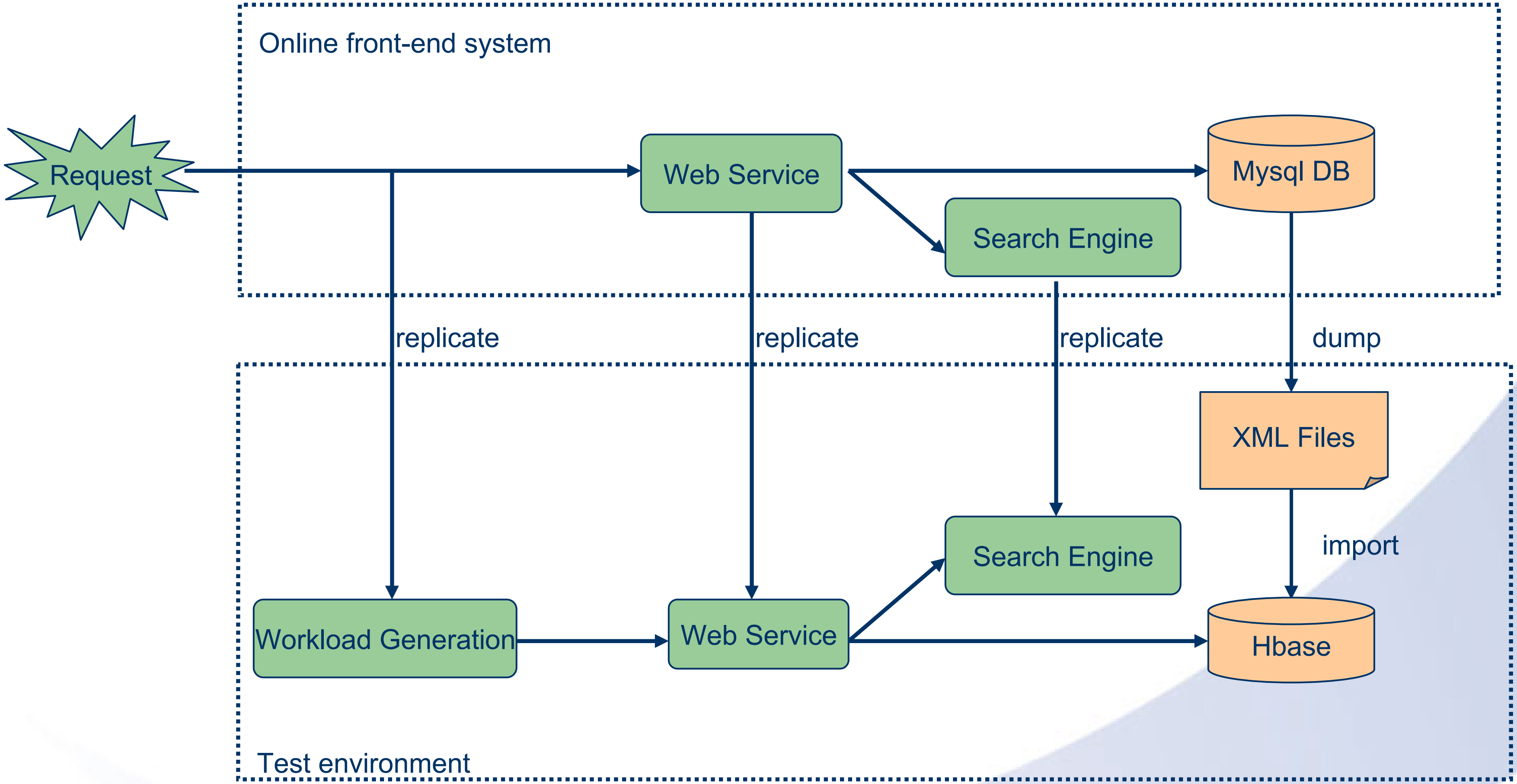
Four case studies

- different hardware
- Different data stores
- Domain-specific algorithms
- Performance behavior analysis

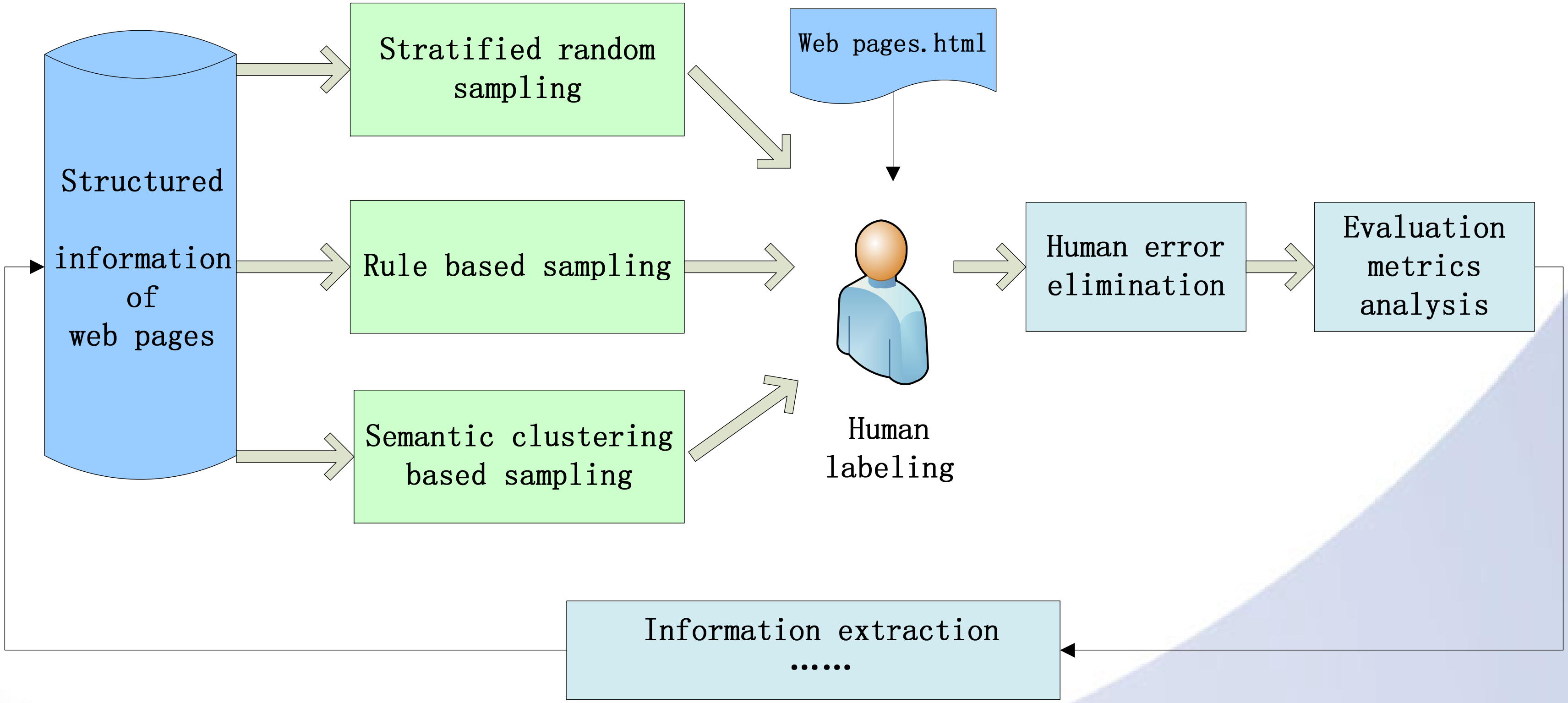
Different hardware: Xeon vs. Atom

- MySQL based search
  - Xeon: CPU Type: Intel(R) Xeon(R) E5310; CPU Numbers: 4; CPU Frequency: 1600.136 MHZ; **Memory Size: 3.86716 GB**
  - Atom: CPU Type: Intel(R) Atom(TM) D510; CPU Numbers: 4; CPU Frequency: 1666.428 MHZ; **Memory Size: 1.95093 GB**
- Atom
  - Intensity： 9.92923reqs/s
  - Duration： 1007.13s
  - Throughput: 9.04156reqs/s
  - **Average Response Time: 3.59891s**
- Xeon
  - Intensity： 9.9713reqs/s
  - Duration： 1002.78s
  - Throughput: 9.22338reqs/s
  - **Average Response Time: 1.334s**

Different data stores



Different domain specific algorithms



Structured information of web pages

Stratified random sampling

Rule based sampling

Semantic clustering based sampling

Web pages.html

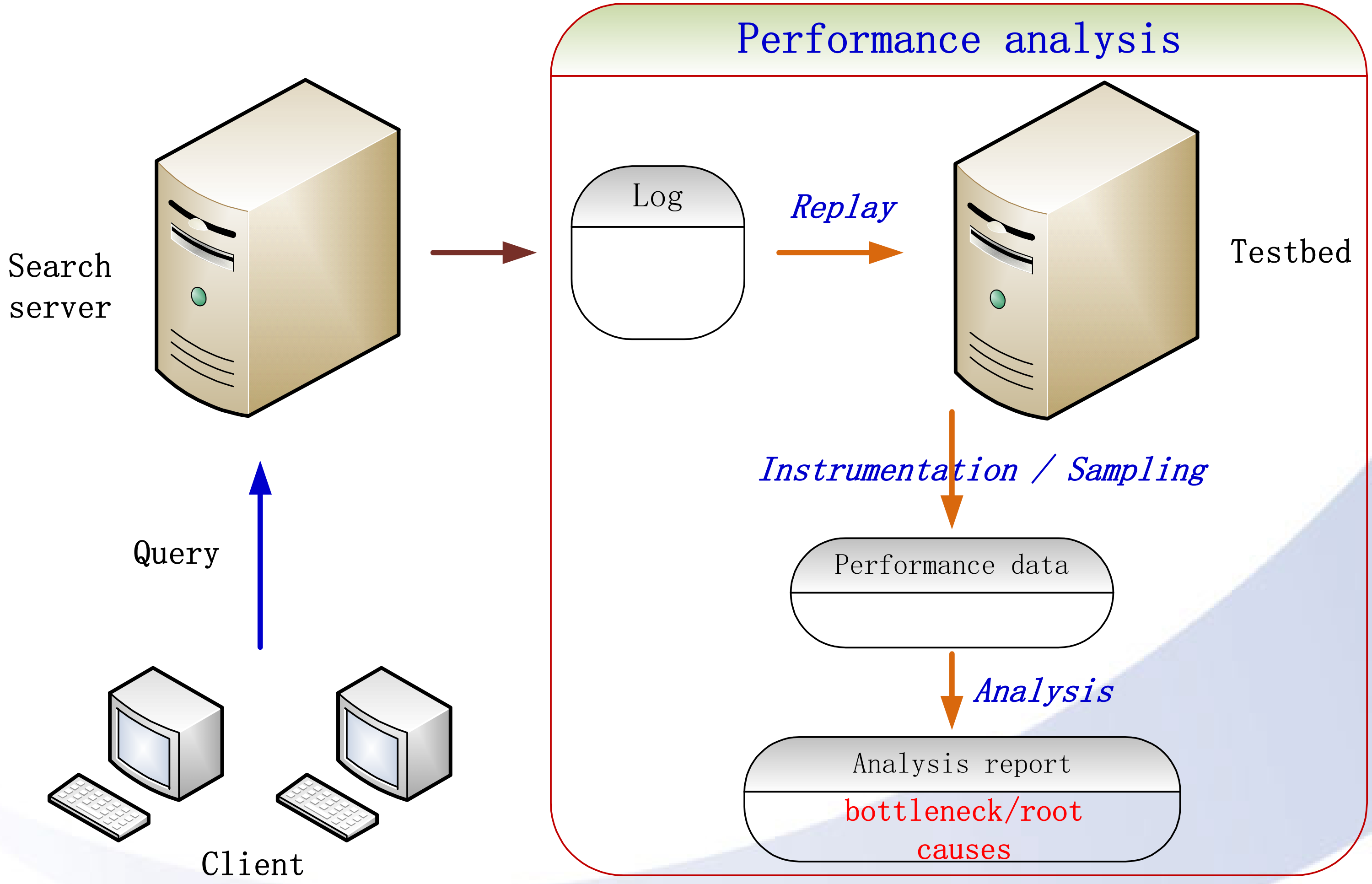
Human labeling

Human error elimination

Evaluation metrics analysis

Information extraction ……

Performance behavior analysis



中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

The current configuration of the deployed front-end service

cores	memory	storages	nodes	Workload types
48	128G	1T	x1	Machine learning
16	12G	1T	x2	Natural language processing
4	4G	4T	x5	Crawler
4	4G	4T	x2	Data cleaning and information extraction
4	4G	120G	x8	Web server, database, search engine

Milestones

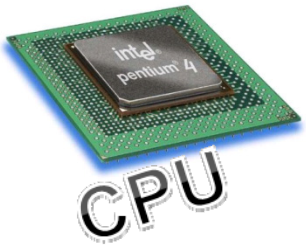
- Version 1.0, 2011.9
  - Demo
- Version 1.5, 2011.12
  - Internal use
- version 2.0, 2012.2
  - A part of features open to external users
- ProfSearch 2.5, 2012.6
  - Full-fledged functions: 140 nodes
- ProfSearch 3.0, 2012.12
  - 1000+ nodes
  - Federated testbeds
  - More applications deployed
    - diverse data and applications

Outline

- What is Datacenter Computing?
- Motivation for a New Testbed
- Current Status
- **Benchmarks**

Current Benchmarks

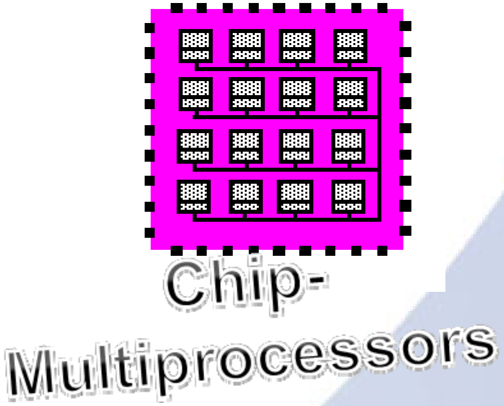
SPEC CPU



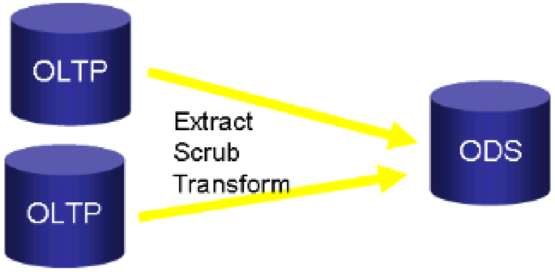
SPEC Web

HPCC

PARSEC



TPCC

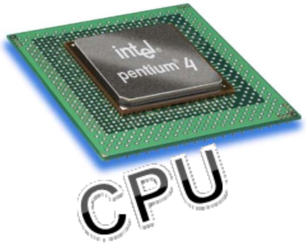


Gridmix

YCSB

Current Benchmarks

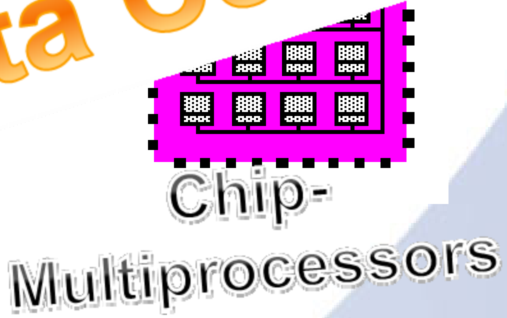
SPEC CPU



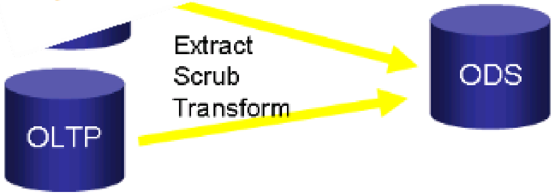
SPEC Web

HPCC

PARSEC



No benchmark for Data Center



Gridmix

YCSB

Our contributions

- For search engines, we find:
  - Real-world query traces do not follow well-defined probability models
  - Synthetic traces do not accurately reflect the real traces
- We develop and open source :
  - **Search**: a benchmark for datacenter computing
  - **DCAngel** : a comprehensive workload characterization tool
  - Available at http://prof.ncic.ac.cn/DCBenchmarks

Evaluation Methodology

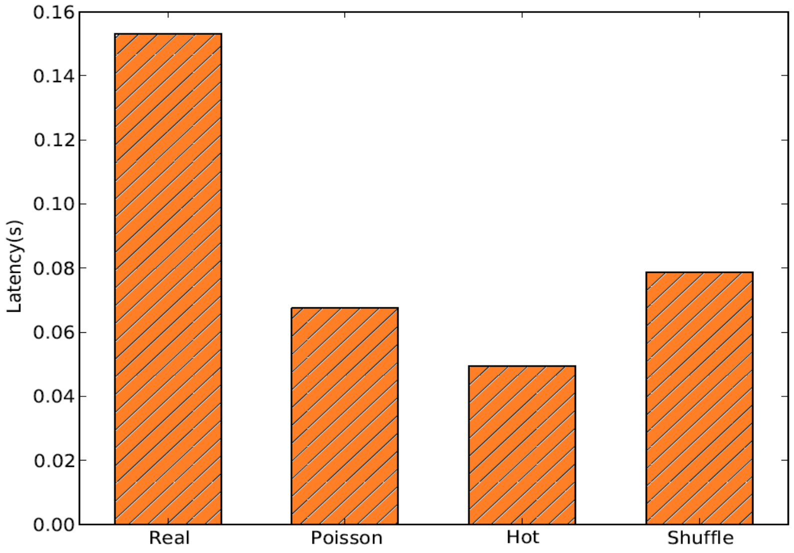
- Workload traces:

Name	Time sequence	Query sequence	Remark
Real	Original	Original	SoGou workload trace
Poisson	Poisson	Original	
Hot	Poisson	Frequency order	Only top 1000 distinct queries
Shuffle	Poisson	Random	Poor temporal locality

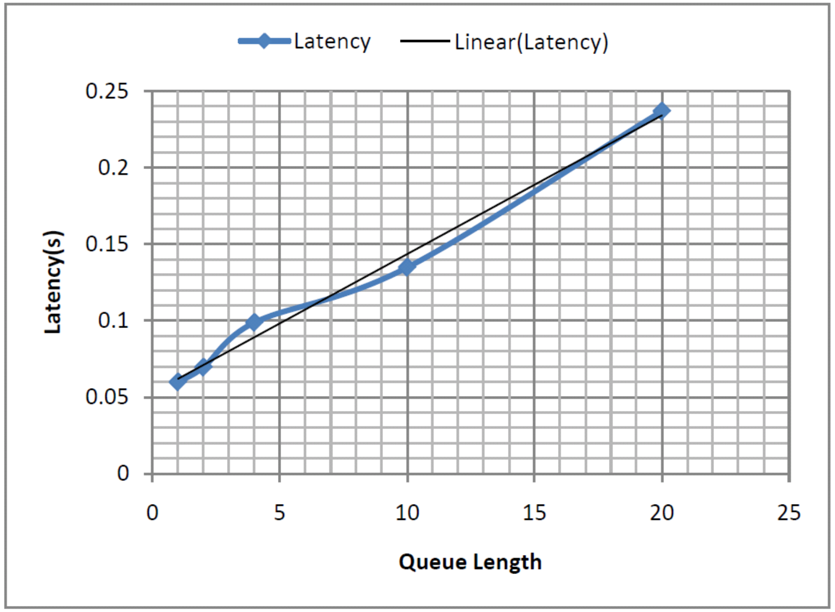
H. Xi, J. Zhan, et al. Characterization of Real Workloads of Web Search Engines. 2011 IEEE International Symposium on Workload Characterization （IISWC-2011）. 2011.

Response time

Four workload traces' response time



Response time and queue length

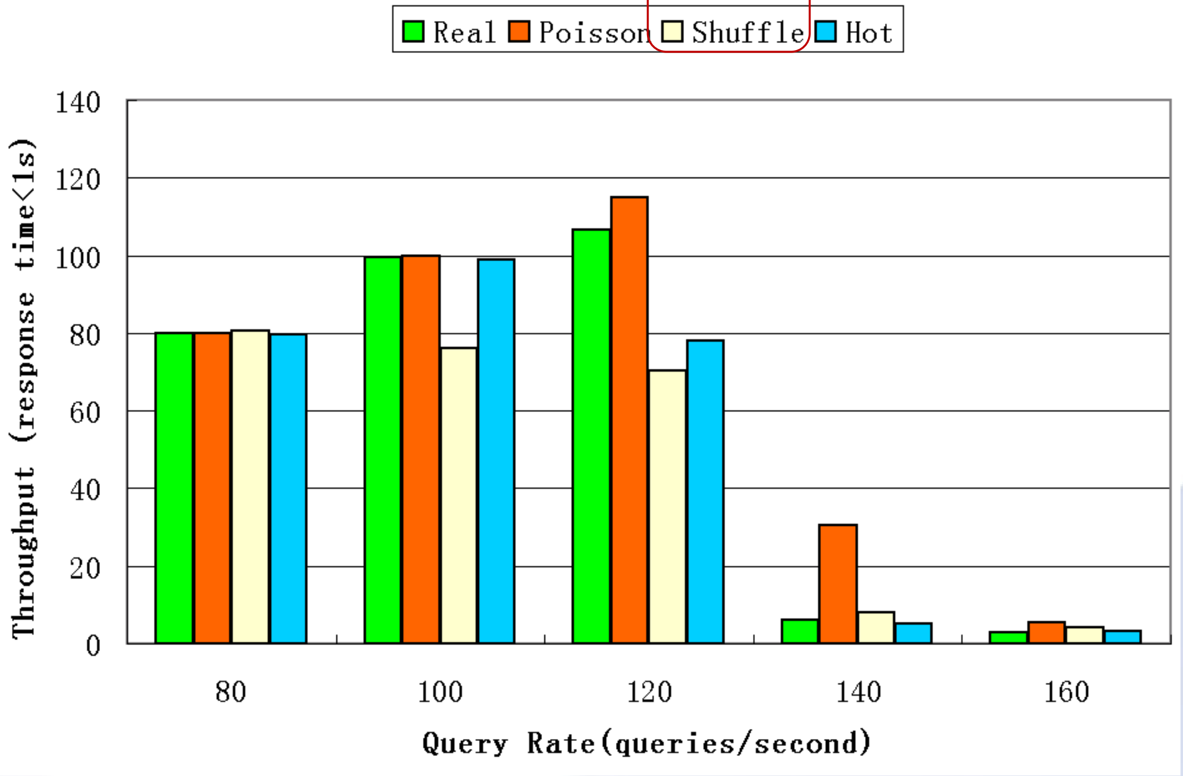


- *T_response* = *T_queue* + *T_service*, response time and queue length have a linear relationship
- Rate variation can make the queue become longer

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

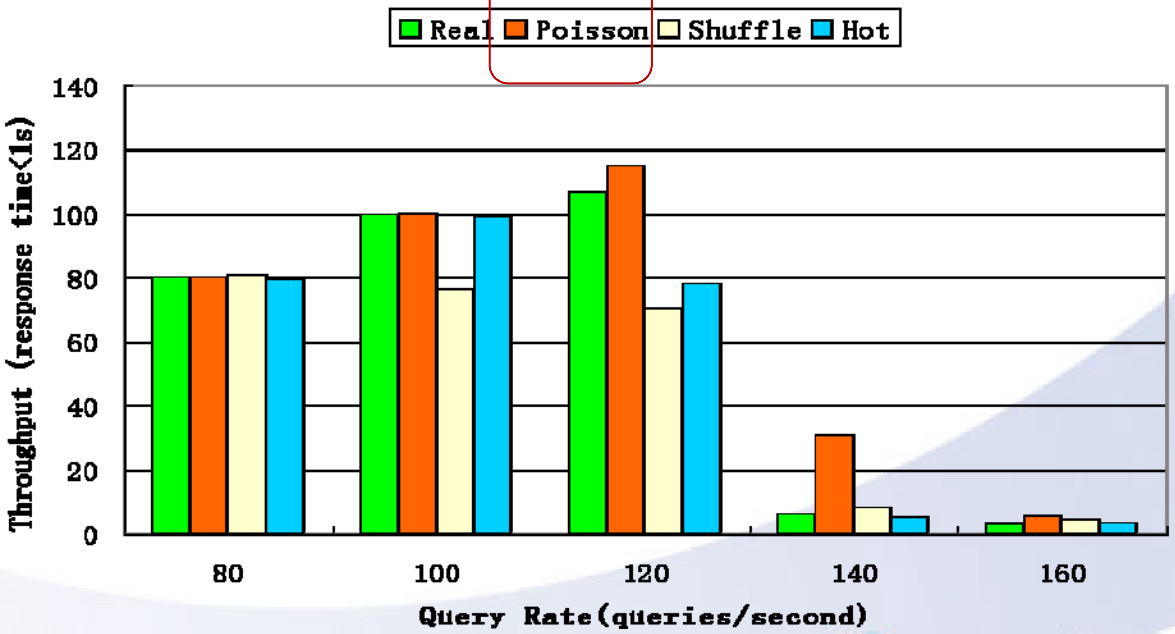
Throughput

- Shuffle has the worst throughput for its worst temporal locality



Throughput

- Shuffle has the worst throughput for its worst temporal locality.
- Poisson has the best throughput for its rate variation is not as severe as the real trace.



中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Current status

- We are publishing more benchmarks for datacenter computing
  - NoSQL based system
  - Data mining and machine learning algorithm
  - A benchmark for shared datacenters

- Hope that you can join!

Summary

- We have built a testbed for datacenter computing
  - Now 5 TB data, 36 nodes.
  - Expected 100TB+ data, 1000+ nodes in Dec. 2012
  - More applications deployed on federated testbeds
- The testbed provides real big data and live workloads.
  - Resolving data lock-in issue.
- Parallel experiment systems
  - Varying from architecture, OS, and domain-specific algorithms.
- Benchmarks
  - Hope you can join!

Contact information

- Homepage
  - http://prof.ncic.ac.cn/jfzhan
- Mail
  - zhanjianfeng@ict.ac.cn
- Weibo
  - http://weibo.com/jfzhan

Thank you!

Q&A

http://prof.ncic.ac.cn/jfzhan

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY ,CHINESE ACADEMY OF SCIENCES