Cloud-Sea Computing on ZB of Data

Zhiwei Xu
Institute of Computing Technology (ICT)
Chinese Academy of Sciences (CAS)
www.ict.ac.cn, zxu@ict.ac.cn

This research is supported in part by the National Basic Research Program of China (Grant 2011CB302502), the Strategic Priority Program of Chinese Academy of Sciences (Grant XDA06010400), and the Guangdong Talents Program
We Are Entering a ZB Computing Era

• Two historical observations:
  – Per-capita capacity: Mega $\rightarrow$ Giga $\rightarrow$ Tera;
  – Worldwide capacity: Peta $\rightarrow$ Exa $\rightarrow$ Zetta

• Two major challenges
  – Capacities increase 1000X, while power (and energy) 1X
  – Enable existing and new workloads (and values)

<table>
<thead>
<tr>
<th>Capacity</th>
<th>1986</th>
<th>2007</th>
<th>2030</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Per Capita</td>
<td>Worldwide</td>
<td>Per Capita</td>
</tr>
<tr>
<td>Storage</td>
<td>4.3 MB</td>
<td>21 PB</td>
<td>44.7 GB</td>
</tr>
<tr>
<td>Communication</td>
<td>12 MB</td>
<td>59 PB</td>
<td>9.86 GB</td>
</tr>
<tr>
<td>GP Computing</td>
<td>0.06 MIPS</td>
<td>0.3 PIPS</td>
<td>0.97 GIPS</td>
</tr>
<tr>
<td>SP Computing</td>
<td>0.09 MIPS</td>
<td>0.44 PIPS</td>
<td>28.6 GIPS</td>
</tr>
</tbody>
</table>

2030 projection: from a conservative estimation by ICT, CAS.
Workload Mega Trend: e-People

- e-People = Computing for the Masses
  - IT that directly benefits the masses (billions of individuals), not institutions
    - e-People, not e-Business, e-Science, e-Government
  - Computer science utilizing the human-cyber-physical ternary universe
    - Ternary computing, not just cyber computing (unary computing)
    - e-People is not fully realized if we have to use cyber devices

[Diagram showing billions of users, trillions of devices, millions of verticals, and human-facing devices are not enough. Currently videos are the #1 load, 2.88 TB = 8 HD movies per day!]

Institutional Computing
- e-Business
- e-Science
- e-Government

Cyberspace Computing
- IT services
- IT software
- IT hardware
The Chinese Academy of Sciences
NICT Project

• New generation ICT
  – 10-year research project (2012-2021)
  – 19 institutes, over 200 faculty members
  – Targeting potential mainstream markets of 2020-2030
  – Aiming at China’s needs in 2020-2050

• Human-cyber-physical ternary computing for ZB of data
  – Functional sensing
  – Customizable Internet
  – Cloud-sea computing
Functional Sensing: Acquisition of Home Appliances Data

• Application examples (2020-2030)
  – Web search → Grid search
    • “Top 100 green households in Beijing and London”
  – Appliances R&D
    • Utilizing field data for all appliances (better than software beta-test)

• Acquisition challenge
  – Can we timely acquire massive and accurate field data from billions of households, for each and every appliance (lamp, refrigerator, etc.) in every household, with 1(~3) sensors per home?
Traditional Sensing

- One sensor per device
  - ~50 devices per home, 220V@50Hz
  - Up to $128^{th}$ harmonics
    - 256 samples/cycle, 10 bytes/sample
  - 6.4 MB/s, or 200TB per year per home
  - For China, 200TB x 0.5 billion homes = 100 ZB per year

Current waveform of a heater in one cycle
Functional Sensing

• One sensor per home

• Function is formalized behavior
  – Type 0: human sensor
  – Type 1: current smart meters
  – Type 2: on-off behavior data for each device
  – Type 3: event behavior data
  – Type 4: finite behavior data
    (up to $k$th harmonics for a given finite $k$)
  – Type 5: infinite behavior data

• Data storage needs can be reduced 10,000 times
  – **20GB**/year per home for aggregated data
  – 1TB/year per home for disaggregated data for each device
The REST 2.0 Architecture for Cloud-Sea Computing

**Sea-side functions**
- sensing, interaction, local processing

**Cloud-side functions**
- aggregation, request-response, big data

- **EB-scale Billion-thread Servers**
- **PB-scale Servers**
- **CDN/CGN**

**Sea HTTP**
- Sea Zone
- Billions, GB-TB
- Trillions, KB-GB

**Sea Zone**
- Seaport
- Billions units TB-PB/unit

**Network:** HTTP 2.0+
- 100s units
- 10Ks units
- Millions
New Gadgets for Homes

- GB sensor nodes @0.2W
- TB “smart phones” @2W
- PB wuTV (home datacenter) @20W
- PB Personal Watson (iPC) @200W

[Diagram showing SeaHTTP, wuTV, iPC connected to Home with HTTP 2.0+]
Three examples of Data Computing

• Off-line (back end):
  RCFile for Apache Hive
  – Production use: Facebook, Taobao, Netflix, Twitter, Yahoo!, Linkedin, AOL, Salesforce.com, etc.

• On-line (front end):
  CCIndex on Hbase
  – Production use in Taobao, Tencent

• High-speed communication: DataMPI

Alexa Top Sites (2013.06.14)
1. Facebook
2. Google
3. YouTube
4. Yahoo!
5. Baidu
7. Windows Live
8. Twitter
9. QQ (Tencent)
10. Taobao
22. eBay
DataMPI open sourced at datampi.org

**Sort**

Hadoop

EXEC Time 99 sec

**PageRank**

Hadoop

EXEC Time 364 sec

DataMPI

EXEC Time 18 sec

**PageRank**

Hadoop

EXEC Time 103 sec
Billion-Thread Server

Traditional Architecture of Datacenters

- Core
- Aggregation
- Access

Applications
- Application Management
- Runtime Environment

OS
- Hypervisor

CPU
- Memory

Chipset

Disk
- NIC

REST 1.0 Requests

Reduce Datacenter Layers

Architecture of Cloud-Sea Server

Applications
- Micro OS
- Nano Kernel

Workload Processing Unit (WPU)
- Memory
- Storage

REST 2.0 Requests

Simplify SW/HW Stacks
Cloud-Sea Storage

- Emphasize power-on efficiency (70% HW peak), while matching latency, scalability, resilience needs
- Innovations
  - stable sets
  - metadata clustering
  - network RIAD

40 benchmark apps: reduces latency 123 times, backend load 50 times
Elastic Processor

• A new architecture style (FISC)
  – Featuring function instructions executed by programmable ASIC accelerators
  – Targeting 1000 GOPS/W applications

• Results: 932 GOPS/W for machine learning

<table>
<thead>
<tr>
<th>Chip types:</th>
<th>10s</th>
<th>1K</th>
<th>10K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power:</td>
<td>10~100W</td>
<td>1~10W</td>
<td>0.1~1W</td>
</tr>
<tr>
<td>Apps/chip:</td>
<td>10M</td>
<td>100K</td>
<td>10K</td>
</tr>
</tbody>
</table>

Chip types: Intel X86, ARM

Power: 10~100W, 1~10W, 0.1~1W

Apps/chip: 10M, 100K, 10K
References

- Xiaoyi Lu, Bing Wang, Li Zha, Zhiwei Xu: Can MPI Benefit Hadoop and MapReduce Applications? ICPP Workshops 2011: 371-379
- Yongqiang Zou, Jia Liu, Shicai Wang, Li Zha, Zhiwei Xu: CCIndex: A Complemental Clustering Index on Distributed Ordered Tables for Multi-dimensional Range Queries. NPC 2010: 247-261
谢谢！
Thank you!

zxu@ict.ac.cn