Evaluation of Big Data Processing Capabilities based on Chinese Hardware and Software Platform

Li Yi, Cui Yun-Fei, Li Kang, Liu Dong

***Abstract*—Based on domestic and foreign hardware and software environment, a heterogeneous big data processing platform is built. This paper experimental tests Chinese processor utilization on dealing with different amount of data through the experimental method, measures the ability to handle big data of domestic processors, and compares the capability with the advanced foreign processor. The experiments show that the single-core power of domestic processors is to be improved, while the ability of overall multi-core processor is par with the foreign advanced processor. The domestic platform could support processing big data well.**

***Index Terms*—domestic platform; multi-core; processor; big data**

I. INTRODUCTION

National long-term science and technology development plan (2006-2020) issued by Chinese State Council in 2006, brought forward 16 major science and technology projects, one of which is the "core electronic devices, high-end universal CMOS chips and basic software product" [1]. And then, the research and development of domestic hardware and software platforms reached the national strategic level.

With the advent of the era of big data, large-scale data processing infrastructure construction is full swing across the country, and big data applications are also around rapid development. We urgently need to develop domestic hardware and software platform to support large-scale data processing, in order to get rid of the control of foreign IT giants, achieve independent controllable, to ensure the security of our information systems. At present, the development of domestic hardware and software platforms is at the key period, while tests are orderly formed. In the current era of big data, we should be concerned about the ability domestic hardware and software platform to support large-scale data processing. The develop way of domestic platform is a multi - core processor. Thus, how to take full use of multi-core processor becomes an important issue.

Manuscript received February 06, 2014; revised February 14, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant 60904082.

Li Yi is with the Academy of Equipment, Beijing 101416, China (e-mail: ylili@139.com).

Cui Yun-Fei is with Beijing Aerospace Control Center, Beijing 100094, China (e-mail: sducyf123@yahoo.com.cn).

Li Kang is with the Academy of Equipment, Beijing 101416, China (phone: 86-13520876718; fax: 86-10-66364116; e-mail: xtlikang1987@126.com).

Liu Dong is with the Academy of Equipment, Beijing, 101416 China (e-mail: ld4m@139.com).

II. DOMESTIC HARDWARE AND SOFTWARE TESTING PLATFORM

Domestic hardware and software platform covers a collection of hardware, operating system and other basic software. Representational domestic hardware is YinHe FeiTeng and Loongson, while the representational domestic software is NeoKylin. In this paper, the test environment is mainly used NeoKylin FeiTeng CPU, NeoKylin CPU and NeoKylin operating system.

A Domestic CPU

***a** YinHe FeiTeng CPU*

The FeiTeng processor is independently developed by the School of Computer Science, National University of Defense Technology. Its technical performance is superior to common international mainstream high-end digital signal processor. It not only breaks the monopoly of the Chinese high-end general-purpose digital signal processor market by foreign products, and marks this design technology has reached the world advanced level. Its successful development has great significance to improve Chinese national security, can help improve economic information systems and defense sector security and confidentiality issues arising in the application of high-end DSP chip.

In March 2011, our domestic server manufacturer Inspur Group officially launched the first server products based FeiTeng processor (FT-1000). FT-1000 clocked at 1GHz, 65nm process, the chip integrates 350 million transistors. The performance of FT-1000 is corresponding to Intel/AMD mainstream multi-core processor at 2006, and its computational efficiency is better than Intel's latest six-core processor.

b Loongson CPU

September 28, 2002, domestic general-purpose CPU Godson-1, developed by the Institute of Computing Technology, Chinese Academy of Sciences, is officially released. Since then, the Godson-2 and Godson-3 have been successfully developed. The Godson-1 processor is suited for low power embedded applications, is 32-bit processor core which considered both general-purpose and embedded CPU features. It adopts a MIPS III instruction set, has a 32-bit integer unit and 64-bit floating-point unit.

Godson-2 processor adopts a four superscalar super pipeline structure. Its chip level and data cache is 64KB, chip secondary cache up to 8MB, the highest frequency is 1GHz,

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

consumes is 5 watts to 7 watts, much lower than similar foreign chip. Measured by SPEC CPU2000 procedures, the performance of Godson-2 CPU has reached the level of low-end Intel Pentium4 series processor, is suitable for desktop and thin client applications. The family of Godson-3 processors is mainly suited for high-performance computing, low-power data centers as well as high-end embedded applications. It is a multi-core processor using 65nm technology, the clock frequency of 900MHz ~1GHz，the number of transistors is 425 million. Its design has a distributed architecture, scalability characteristics. Loongson said that the future of the three series will develop in parallel, in order to meet the needs of a variety of different applications [2].

Auroral company has delivered 1000 sets automatic longteng server based on Loongson processor in 2012. There are 400 sets in the cloud computing industry area in Chongqing Jiangjin. There are 100 sets in an institute of green intelligent, the Chinese academy of sciences. Solutions based on a long-time server have successfully applied in aviation, electronics, electronic government and cloud computing, which involve the countries of the importance of safety field.

B Kylin Operating System

The operating system is a kernel to control other programs to run and manage system resources and to provide users with a set of operating system software interface. For a long time, the operating system market is almost monopolized by Microsoft and other companies.

The Kylin operating system is a set of operating system with independent intellectual property developed by the PLA National University of Defense Technology. As a “863” major technological research project, the goal of Kylin is to break the foreign monopoly of the operating system. The Kylin operating system is the first one via product quality supervision by the Chinese Ministry of Public Security, and is B+ level security certification by the People’s Liberation Army. It is currently the highest level of security of the operating system. The Kylin operating system is divided into three main structures, the bottom is free to replace the loaded base layer to ensure system security and real-time, in the middle is FreeBSD kernel, and the upper is Linux compatibility library, in order to achieve binary compatibility of the Linux platform.

III. Capabilities Demand of Big Data Processing

The amount of data in many areas, such as scientific research, Internet applications and computer simulation shows a trend of rapid growth. Moore's Law tells us that the CPU processing speed doubling every 18 months, and Turing Award winner Jim Gray proposed an empirical law that the amount of data generated every 18 months under the network environment is equal to the sum of the history. And a new term is created for this phenomenon: large data.

IDC defines big data new generation architectures and technologies that in order to economically get value from frequency and large capacity, and different structures and types of data. There are four characteristics of big data: volume, variety, velocity and value. Big data needs to be stored and be computed out to be able to reflect the advantage of the vast amounts of data. As long as there are enough original data, good algorithms coupled with powerful data processing capability, the laws and trends hidden in the data will be able to be revealed [3].

Big Data + Large-scale Data Processing=Value

At present, the many domestic and foreign IT companies, scientific research units and government agencies are in intensifying research or ongoing big data processing, the representative are Google [4], Microsoft [5], Facebook [6], Amazon [7], Ali Baba [8], etc., which mainly USES is the map-reduce programming model. This paper will use map-reduce on domestic hardware and software platform for performance test.

IV. Domestic Hardware and Software Platform Performance Test Based on the Map-Reduce

A Experimental Environment

The hardware of experimental environment consists of four Great Wall FeiTeng PC, three Great Wall LongMeng PC and three Lenovo PC. FeiTeng PC uses a 32-core FT-1000 CPU, 4G RAM, 420G disk space. LongMeng PC uses a 4-core Loongson CPU, 2G RAM, 20G disk space. Lenovo PC uses a dual-core Pentium E5500CPU, 2G RAM, 500G disk space. The experimental environment installs the NeoKylin operating system and centos operating system, and nodes are connected through 1000Mbit/s Ethernet.

The prototype system is developed based on Hadoop0.34.0. The Master module is deployed on a Great Wall FeiTeng PC, while slave modules are deployed on other nodes, and 1000Mbit/s Ethernet as a data transmission network. Input files are managed by Hadoop Distributed File System, while file blocks are stored in the compute node's local disk.

B Experimental Set

In order to compare the capability of domestic hardware and software platform to support processing different scale of data, the experiment is divided into three groups according to the number of tasks. Each task is answer for dealing with the amount of data is 32M [9,10]. The specific settings are shown in Table 1.

TABLE I
OPERATING PARAMETERS SET

the group number	the number of tasks	the amount of data
1	7	224
2	40	1280
3	100	3200

C Experimental Results and Analysis

a Domestic CPU Utilization Comparison

Domestic CPU’s develop direction is multi-core, and how to effectively use its multi-core advantage is a matter be worthy of concern. This paper tests domestic CPU utilization for different amount of data. Fig 1 shows FT-1000’s (32-core) utilization before running task. Fig 2 shows FT-1000’s (32-core) utilization running 7 tasks. Fig 3 shows FT-1000’s (32-core) utilization running 32 tasks. Fig 4 shows Loongson’s (4-core) utilization running 4 tasks.

As you can see from the figures, the utilization of domestic CPU can reach more than 90% when running big data

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

processing procedures, indicating that the domestic platform has powerful parallel processing capabilities.

Fig1 FT's utilization before running task

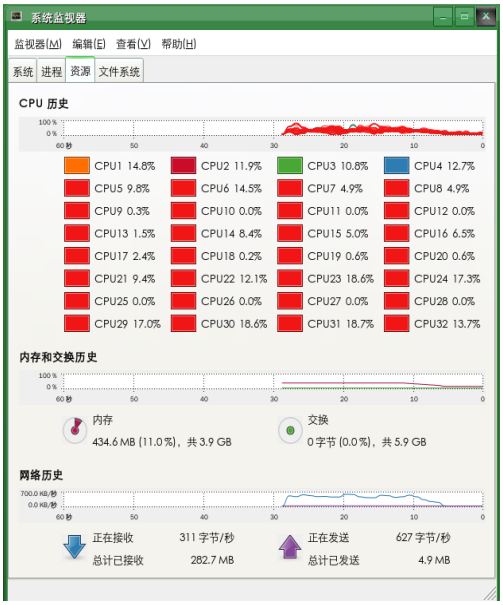

Fig 2 FT's utilization running 7 tasks

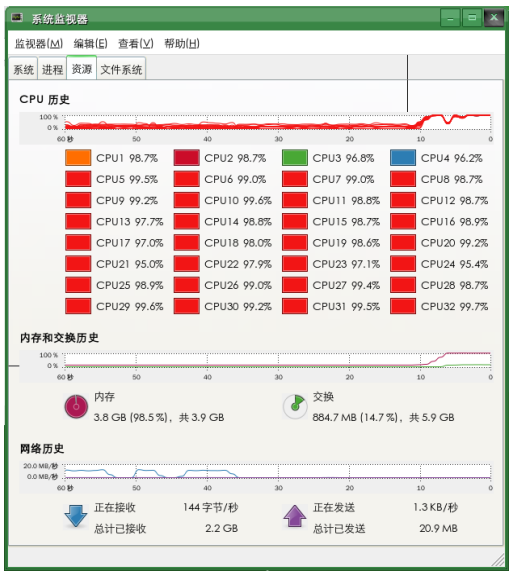

Fig 3 FT's utilization running 32 tasks

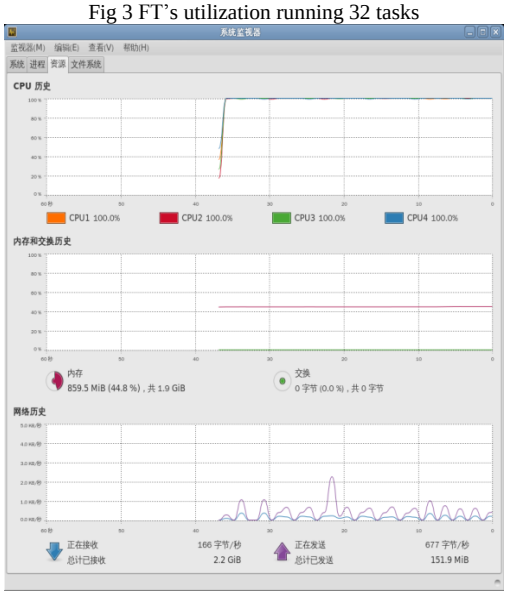

Fig 4 LS's utilization running 4 tasks

b Performance Compare between Domestic Platform and Foreign Platform

Experiments are carried through based Table 1. Fig 5 shows the time needed to process the same amount of data by each core of FeiTeng, Loongson and Pentium. Fig 6 shows the time needed to process the same amount of data by FeiTeng PC, Loongson PC and Pentium PC.

As can be seen from Fig 5, the Loongson's single-core processing power slightly stronger than FeiTeng, but significantly worse than single-core processing power of the Pentium processor. As can be seen from Fig 6, although the single-core poor, but domestic CPU, especially FeiTeng processor fully integrate the advantages of multi-core, the processor overall processing power is stronger than Pentium.

Additionally, the number of slots in hadoop can be set. In the experiments we find that if the number of slots is large, the execution rate of processor with fewer cores will be small, because of the task blocking and buffer overflow. If the set number of slots is less than the number of CPU cores, the number of tasks at the same time will be less than the number of processor cores, cannot give full play to the advantage of the processor multi-core. Thus the experiments finally set the number of slots equal to the number of cores of each processor.

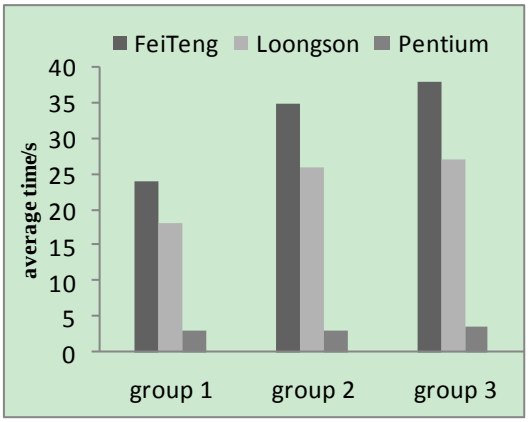

Fig 5 processing time of single-core

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

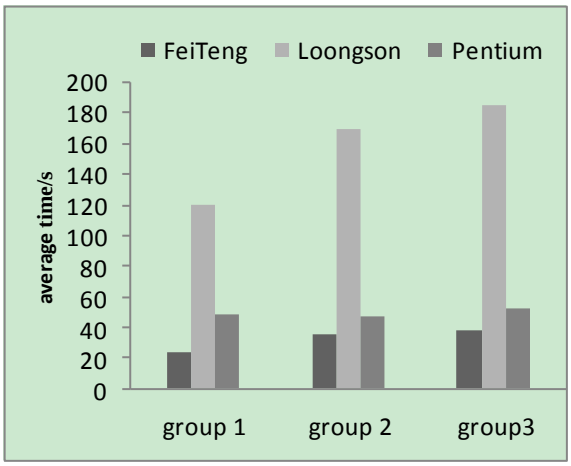

Fig6 processing time of whole PC

V. Conclusion

A Chinese platform has been already underway, but there is still a certain gap compared with the foreign advanced platform. The capability of domestic CPU is about 10% of foreign CPU's. However, our superiority is multi-core. The capability of whole PC, such as FT-100, is 1.6 times of foreign PC's. On the other hand, the use of a big data processing mechanism can play domestic processors multi-core advantage.

References

[1] Chinese State Council l. National long-term Science and Technology Development Plan (2006-2020). 2006
[2] Cheng jian, Wu wei. The Research of Apply Domestic Computing Platform in Command& Control System. Automation &Information Engineering, 2011 (3): 41-44.
[3] Wang Peng. Research on Programming Models for Massive Data Processing. Beijing, 2011.
[4] DEAN J, GHEMAWAT S. Map-Reduce: simplified data processing on large clusters. Common ACM，2008, 51 (1): 107-13.
[5] ISARD M, BUDIU M, YU Y, et al. Dryad: distributed data-parallel programs from sequential building blocks. Proceedings of the EuroSys, 07. Lisbon, Portugal, 2007:59-72
[6] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain,et. Hive - A petabyte scale data warehouse using hadoop. In Proceedings of the IEEE International Conference on Data Engineering, 2010.
[7] Amazon Elastic Compute Cloud (EC2). http://www.amazon.com/ec, 2012
[8] Tang Hong. "Flying" large-scale distributed computing system. Beijing: The 4th China Cloud Computing Conference.2012
[9] ZAHARIA M, BORTHAKUR D, SEN SARMA J, et al. Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. Proceedings of the EuroSys'10. Paris, France, F, 2010.265-278
[10] JIN Jia-hui, LUO Jun-zhou, SONG Ai-bo. Adaptive delay scheduling algorithm based on data center load analysis. Journal of communications, 2011,32 (7): 47-56.

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)