自主高端云计算服务器模型机*

崔吉顺, 刘宇航, 张楠赓, 祝明发, 肖利民

(北京航空航天大学 计算机学院,北京 100191)

Independent Developing of High-end Cloud Computing Server Model Machine

CUI Ji-Shun, LIU Yu-Hang, ZHANG Nan-Geng, ZHU Ming-Fa, XIAO Li-Min

(School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

Abstract: With the rapid development of cloud computing technology and the consolidation of massive computing resources, the servers carrying cloud computing application are demanded higher requirements on high-density, reliability, stability and low power consumption. Independent development and design of the high-end cloud computing server, which is the most important information infrastructure for many application domains, is of great security significance to change the situation that high-end server market is mainly occupied by foreign manufacturers. System architecture, multi-core processor, motherboard layout, thermal analyses are studied, and a model machine prior to product but involves the same key engineering technologies is developed. Furthermore, the critical choices of reliability design are discussed.

Key words: cloud computing; high-end server; Godson-3B; multi-objective design; model machine

摘 要: 云计算技术的快速发展和大量计算资源的聚集,对承载云计算应用的服务器在高密度、可靠性、稳定性、低功耗等方面提出了新的要求.为改变国外厂商占据高端云计算服务器的局面,高端云计算服务器作为应用领域中的重要信息化设备,研究和设计自主产权的国产高端云计算服务器具有重大安全战略意义.本文从服务器体系结构、多核处理器、可靠性设计（主板布局、系统散热）方面对服务器进行研究和产品模型的研制,着重探讨了可靠性设计时的关键抉择,掌握了关键工程化技术,达到了可产品化的目标.

关键词: 云计算;高端服务器;龙芯 3B;多目标设计;模型机

云计算是将海量可扩展的IT资源和能力通过服务形式提供给用户的一种网络计算模式,也是一种新兴的IT资源配置和交付模式.云计算主要代表如 Google 和 Amazon 的应用均以数据冗余为基础,配合高可扩展的并发技术,向用户提供高响应度的服务.这种应用模式以大规模的计算资源聚集为基础设施,以计算设备和流程的多路性保证对外服务的可用性和响应速度.除互联网应用之外,将有更多传统的企业应用以云服务形式对外提供,由此也将导致分散的 IT 资源集中到规模不等的数据中心内[1,2].John L.Hennessy 和 David A. Patterson 将这样的云计算服务器称之为仓库级计算机(Warehouse-Scale Computer),同时开发请求级并行（Request-level parallelism）和数据级并行（Data-level parallelism）[3].

云计算技术的快速发展和数量众多的计算资源聚集,对承载云计算应用的服务器系统提出了新的要求[2],对高密度紧耦合计算资源的需求以解决规模庞大的群集系统在可管理性方面存在的巨大缺陷；对服务器低功耗的架构设计以解决庞大规模云计算数据中心在系统功耗方面都面临严峻的挑战.面向云计算的高端服务器必须尽可能提升单位体积计算能力或处理器个数以减少数据中心占地面积.高端的云计算服务器需要在高密度、低功耗方面突破系统关键技术[4].高端服务器是关键应用领域中的重要信息化设备,主要用于银行、证券、电信等国家关键行业的业务运营,服务器的自主安全性越来越重要.结合我国高端服务器市场的情况,目前国外厂商占据了主导地位,大量的关系到国计民生的关键行业核心系统都是采用国外厂商的产品,从国家和社会安全的角度考虑,我国必须拥有自主知识产权的高端服务器系统.高端云计算服务器系统,其一方面应具有较强的事务处理能力,另外一方面应具有较高的可靠性,可长期提供高速、稳定的信息处理服务,因此需要在可靠性设计上进行研究和实现.在云计算催生大量开放性技术和我国核心处理器实现突破

*Supported by the National Core electronic devices high-end general purpose chips and fundamental software project under Grant No.2010ZX01036-001-001(核高基重大专项); the National Natural Science Foundation of China under Grant No. 60973008 (国家自然科学基金); the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.20101102110018(国家教育部博士点基金); the Hi-Tech Research and Development Program (863) of China under Grant No.2011AA01A205(国家高技术研究发展计划(863)).

作者简介: **崔吉顺**(1972－),男,山东昌乐县人,博士研究生,主要研究领域为计算机体系结构;**刘宇航**(1985－),男,博士研究生,主要研究领域为计算机体系结构,并行计算;**张楠赓**(1983－),男,北京人,博士研究生,主要研究领域为计算机体系结构;**祝明发**(1945－),男,博士,教授, 博士生导师,主要研究领域为计算机体系结构,超级计算机;**肖利民**(1970－),男,博士,教授,博士生导师,主要研究领域为计算机体系结构,计算机系统软件,高性能计算机系统.

的背景下,独立研制我国自主知识产权的高端服务器也具备了相应的技术基础.本文研究适于云计算的高端服务器体系结构模型,并从系统高密度、自主、可靠性方面进行精细分析和设计.

1 系统体系结构

适于云计算的高端服务器系统设计,是根据云计算的特点和结合应用需求确定的.如图 1,系统的设计从处理器及芯片组结构、服务器结构两大方面考虑,其中服务器结构包括机箱结构、主板、存储、电源、网络互连、整机散热等. 在本文研发团队的前期技术积累之上[5],本文采用的是基于 CMP 的高密度计算机多目标设计方法[6]。

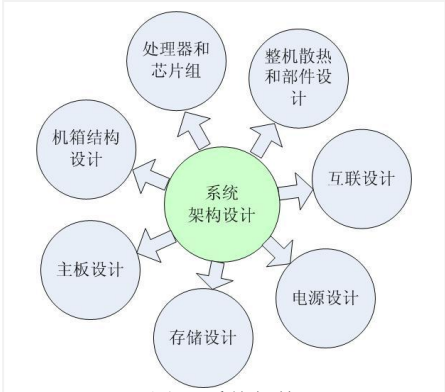

图 1 系统架构

1.1 基于的国产龙芯 3B 处理器

在云计算高端服务器处理器的选择使用上,采用国产处理器.一方面国产处理器的技术已经取得良好的进步.通过“十五”和“十一五”期间的发展,在国家的大力支持下,我国在微处理器设计方面已经有了长足的基础,在微处理器研制方面取得群体性发展.在单核处理器的研发方面达到世界先进水平,并快速切入多核处理器的研发,研制出多款高性能通用处理器[7,8],如“龙芯”、“申威”、“飞腾”、 “众志”.另一方面,云计算高端服务器的安全性非常重要,尤其在一些关键领域中的应用.自主研发的国产处理器在安全性上能够满足这些关键领域的要求.国产处理器的选择,除了技术水平,安全性满足要求外,软件平台的适配在服务器生态系统的建设和云计算应用推广至关重要.龙芯处理器在国产 BIOS、OS、数据库和中间件生态系统匹配上越来越完善[9].处理器选择上采用国产高性能低功耗的龙芯 3B 处理器,能够满足一些云计算应用的性能要求.龙芯 3B 处理器的研制目标是足我国信息化建设基本需求,尤其是满足国家安全需求的面向服务器和高性能机的高性能、低成本、低功耗的多核 CPU 芯片产品和与之配套的基础软件,并作为主 CPU 用于国产高性能机及国产服务器[10,11].如图 2,龙芯 3B 处理器采用 65 纳米设计,主频为 1GHz; 集成 8 个精度浮点峰值可达 128GFLOPS； 片上集成共享的 4M 二级 Cache； 两个 64 位 400M 强型 HT 控制器(可支持多个芯片间的 Cache 一致性互连); 一个 32 位标准 PCI/PCIX GPIO 等.总之,龙芯多核处理器是因为它相对其他主流多核 CPU 具有结构上的代表性软件生态环境上相对完善,适合云计算服务的应用,因此用于研制高密度、低功耗、自

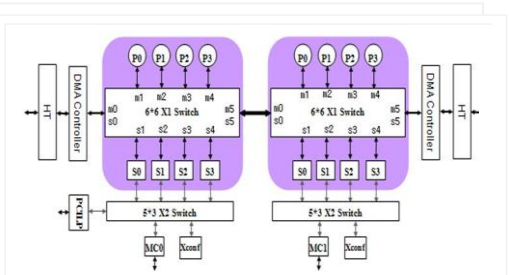

图 2 龙芯 3B 八核处理器

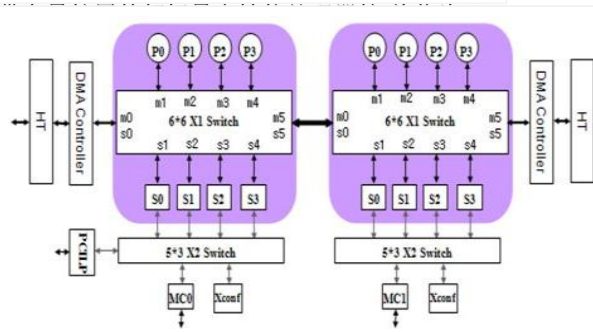

图 2 龙芯 3B 八核处理器

1.2 多胞胎节点服务器架构

为应对云计算中对服务器的高密度低功耗问题,服务器架构设计上采用多胞胎节高的计算密度[12,13,14].在形态上,1U 结构的 Rack 机箱空间内安装两块处理器主板,到 8 颗处理器,64 个处理器核心的高密度.在数据中心每机柜功耗限制下,相比普通服务统总体 TCO.在低功耗设计上,采用转换效率的电源,主板所有的器件采用低功耗器件,并设计有主板管理控制器,根据系统的负载调整系统风扇等实现降低能耗的管理,有效降低空闲状态下的能耗损失.为达到云计算服务器的可靠性,结构设计上采用冗余电源,在系统设计上,重点解决主板布局和系统散热对系统可靠性的影响.

针对云计算中一些商用计算应用软件,要求 I/O 系统具有较高的带宽性能和较高的扩展能力.在系统架构和主板设计中也充分考虑了这一点.每个服务器集成了高速 Infiniband,总带宽 4×40Gbps,千兆以太网,总带宽 4×1Gbps.千兆以太网接口可以连接到以太网交换机上组成管理网络,同时也可以作为计算网络的后备使用.服务器具有对外高带宽的网络,可以通过外部的交换机连接成一个更大的系统,具有很好的扩展性.多胞胎节点机架构同时满足云计算对存储的要求[15].存储子系统包含本地存储和远程存储,其中远程存储是通过 I/O 对外连接实现的,本地存储是与主板直接连接的硬盘,用于安装操作系统和保存本地的数据.同时,主板上的每颗处理器支持 2 个内存条,作为板级的存储,用于计算数据的保存.这三部分组成了服务器的存储系统.在由多个服务器组成的系统中,I/O 系统和存储合为一体,直接连接到高速 Inifiniband 网络上,提高 I/O 的带宽和访问存储的速度.每块主板上有两个结点机,其中每个结点机由两个 3B 8 核处理器、AMD 北桥和南桥芯片组组成.处理器之间采用 HyperTransport（超级传输总线）互连,处理器和北桥芯片之间 HyperTransport 互连,北桥芯片扩展 PCIE 接口,千兆网口和 VGA 视频接口.南桥芯片扩展 USB,硬盘 SATA,串口和监控管理芯片接口. 每颗处理器有两个内存通道,每个内存通道扩展一条内存条,支持 DDR2 内存.处理器之间采用 16bit 宽的 HT 连接,使用 HT0 互连.处理器 0 和北桥 R780E 之间采用 HT1 接口互连; 北桥和南桥之间采用 PCIEX4 通道互连;

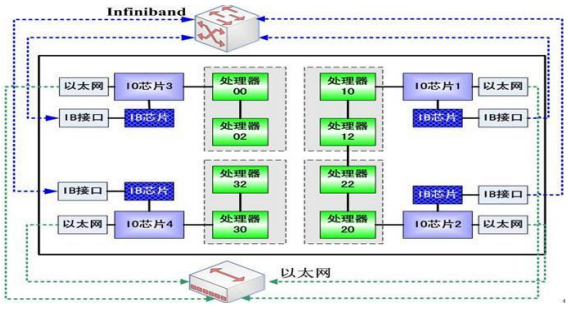

图 3 多胞胎节点服务器架构

北桥芯片 R780E 扩展一个 PCIEX16 接口,VGA 视频接口,两个 Intel 的 82574 千兆以太网芯片,通过 PCIEX1 与北桥连接,另外扩展 VGA 缓存 DDR2 颗粒.南桥芯片 S710 扩展 6 个 USB 接口,6 个 SATA 硬盘接口,Super I/O 等接口.

2 系统硬件可靠性

高端云计算服务器长期运行关键应用,服务器必须具有很高的可靠性.主板的布局设计和结构散热设计是整机可靠性设计中的关键因素,与系统的稳定性有直接的关系.后两个方面又是相互影响,器件的布局影响主板信号的布线质量和完整性,还影响系统散热.设计中在系统结构上,结构和要求结合机箱结构全面考虑各个部件的尺寸是否干涉,布局是否有利于散热.通过模拟软件进行结构和风流系统的分析,进而改进布局设计,这个过程也是反复的.主板是最复杂的系统部件,在主板的设计上,选定可靠的部件和型号,根据散热模拟结果调整布局.

2.1 布局

主板上电子元器件较多,超过 5000 个,主要的部件包括处理器、桥片、网络芯片、电源芯片和内存等,在布局上主要是把这些部件的布局和电子线路设计好.考虑处理器之间的互连和各自连接内存条,在布局上采用规则的方阵式布局,处理器放在中间,内存条放置在两侧.这样的布局能够使得处理器之间的高速信号连接线路最短,同时处理器与内存的连接线路也最短,电路信号上保证主板运行的可靠性.电路板上解决高速、低速数字信号和电源等模拟信号之间的干扰,从 PCB 印制板上再次保证信号的完整性,从而进一步确保系统的稳定.

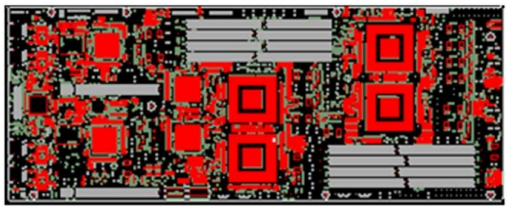

图 4 主板器件布局方式 A

在布局上,还考虑了内存条和处理器的散热,处理器放在中间,较高的内存条放在两侧也是有效的考虑制冷通道,所有的内存排列是前后走向,这样有利于冷风从主板的前端向后端流动制冷.其余的芯片如 IO 放置在不影响通道的位置,同时也考虑芯片信号的完整性.针对处理器和内存的具体布局,考虑了图 4 和图 5 两种方式,并通过实际布线和参数模拟,设计出更有利于系统稳定的布局. 经过实际布线、软件模拟和分析,采用主板器件布局 B 的方式系统设计（如表 1）.

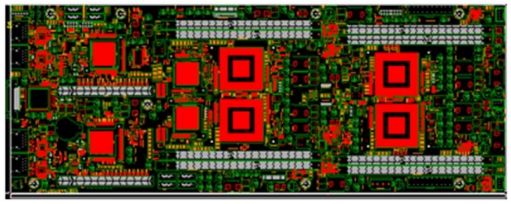

图 5 主板器件布局方式 B

表 1 主板布局对比

	主板器件布局 A	**主板器件布局 B**
布局	板上每 2 颗处理器节点划分不清晰; 电源布局难度大	板上每 2 颗处理器节点划分清晰,左右对称; 电源区域容易分割和布局
走线	内存布线有交叉,布线难度大,内存到处理器的信号长度不均等	处理器和内存之间的连线顺畅; 内存到处理器的信号长度均等
散热	散热通道整体均衡,中间的两个处理器前后风道重叠,后侧处理器散热略差	散热风道集中在中间,后端两个处理器的散热压力较大
PCB 布线层数	20 层	16 层
稳定性和信号质量	处理器到内存的运行频率较低,700MHz 以上出现不稳定现象	内存频率可以运行到 800Mhz 以上,系统稳定

2.2 散热

系统的散热设计包括风扇的选用,散热片和风道的设计,主板布局的散热结构三个方面.

（1）风扇风速测量和处理器热参数

系统风扇由主板上的 LM93 芯片和 SuperI/O 实现风速的检测和调速控制.对整个系统散热的设计和模拟,需要确定和选用哪种风扇更合适,选用 DELTA GFB0412SHS 风扇,转速 15000rpm,对该风扇进行实际测速.系统的模拟需要处理器温度要求.为保证制冷的风流通过机箱内部,保证主板上的处理器、内存条的发热及时带走达到良好的散热效果,整机系统布局上,系统风扇安装在机箱前端,采用前吸风后送风的方式进行系统风流的驱动,使得通过机箱内部的风流速度加快,同时会在处理器、内存上辅以导风罩,减少风流的短路,更有效地导出器件的热量.

（2）散热片设计

在主板上,处理器是主要热量产生源之一,由于两颗处理器器相邻,采用一个大处理器的散热片.散热片的材料采用铜基底,散热效果好.芯片组的热量较小,采用铝合金散热片.处理器的散热片采用两种方案,最后通过散热模拟选择一种最合适的散热方案.

（3）系统散热模拟

基于不同的处理器散热片设计方案进行散热模拟,从而选择最有利的散热方案.从散热片的翅片间距、铜底厚度、散热片方案 A/B 三个方面组合进行散热模拟. 如图 6,在同一款散热片上选择不同的铜基底和翅片,散热效果不同.按照处理器散热片方案 A,散热片铜底选用 3mm、4mm 和翅片间距 1.5mm、2mm 和 2.2mm 进行典型的 5 种组合模拟.在 5 种方案中,通过表 2 的模拟结果可以看出：第一,在系统中 2 个

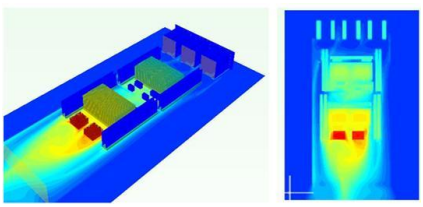

图 6 系统散热方案 A 模拟效果

散热器前后交错放置,后端散热器比较难解,前后 CPU 温度在 10 度左右.第二,从模拟数据分析在 Solution-1 中前端(CPU-1 和 CPU-2)性能较好,但后端(CPU-3 和 CPU-4)性能没有 Solution-2 好,Solution-2 与 Solution-4 性能相当,Solution-4 成本相对低,首先推 Solution-4,散热片铜底使用 3mm,翅片间距 2mm.对处理器散热片方案 B 进行了模拟,方案 B 比方案 A 加长了翅片宽度.如图 7 和表 3,方案 B 相对方案 A 具有以下特点：第一,总模拟分析中,增加散热翅片 FIN 对散热的整体影响不大,新方案对前端 CPU(1、2)的散热有提高作用,对后端的 CPU(3、4)散热没有帮助.第二,增加 FIN 的宽度后,影响了经过后端的 CPU（3、4）风流,从而影响对后端 CPU 的散热.

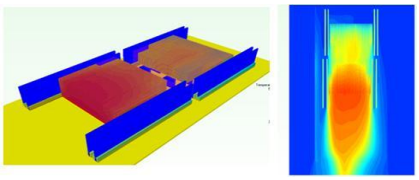

图 7 系统散热方案 B 模拟效果

3 整机测试

作为云计算的高端服务器体系架构及核心技术,系统硬件是整个系统的基础,硬件研发完成了包括自主设计服务器的系统架构、机箱结构、系统散热、多处理器主板、高带宽 I/O 子系统、互连子系统、存储子系统、电源子系统等,研制出了适合云计算特点的高端服务器.在各部分完成详细设计和样品之后,需要进行样机的研制,包括机箱、主板、电源等各个部分,每个部分先进行样板的调试和测试,最后进行整机样机的联合调试和测试.对作为系统的关键部件主板完成了详细功能、稳定性测试.整机样机联调和测试主要进行硬件功能性验证和调试、跌落、高低温、EMI、一致性、可靠性和兼容性等方面的测试,然后根据测试和调试结果再次修改主板、机箱等部件的设计,直到产品达到稳定,整机样机如图 8.为了尽快完善国产龙芯服务器的生态系统,研发的云服务器与国产 BIOS、OS、数据库和中间件进行了匹配评测和验证.实现了一些典型应用在龙芯服务器上的移植和测试,主要涉及常用的国产及通用数据库、浏览器、办公软件、邮件服务器、负载均衡等.

表 2 散热方案 A 的 5 种具体设计散热效果对比

Lenovo Server Simulation							
Solution	Description	T_{cpu-1}	T_{cpu-2}	T_{cpu-3}	T_{cpu-4}	$T_{Adapter-1}$	$T_{Adapter-2}$
Solution-1	AL FIN 62pcs,P=1.5mm,T=0.3mm Copper BASE T=4.0mm	42.98	43.2	54.2	54.28	88.23	86.8
Solution-2	AL FIN 47pcs,P=2.0mm,T=0.3mm Copper BASE T=4.0mm	43.77	44.11	53.03	53.37	88.11	86.67
Solution-3	AL FIN 42pcs,P=2.2mm,T=0.3mm Copper BASE T=4.0mm	44.9	45.12	53.53	53.63	88.21	86.71
Solution-4	AL FIN 62pcs,P=2.0mm,T=0.3mm Copper BASE T=3.0mm	43.85	44.21	53.13	53.49	88.15	86.69
Solution-5	AL FIN 62pcs,P=2.0mm,T=0.3mm AL BASE T=4.0mm,2pcs Heatpipe	45.55	45.64	55.24	55.38	88.25	86.77

表 3 散热方案 A 和 B 整体散热效果对比

Solution	T_{cpu-1}	T_{cpu-2}	T_{cpu-3}	T_{cpu-4}	CFM(cpu-1\cpu-2)	CFM(cpu-3\cpu-4)
散热片A	43.22	43.92	50.09	50.5	15.31	11.11
散热片B	42.82	43.38	50.34	50.78	15.99	9.97

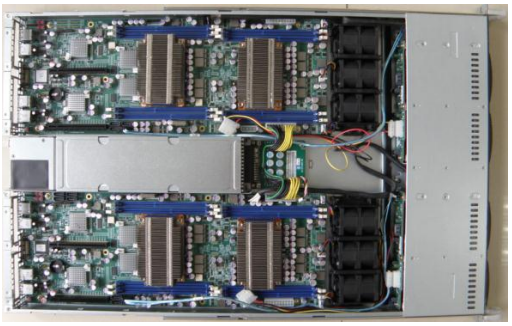

图 8 高端云服务器样机

4 结束语

云计算及其应用的快速发展对云计算服务器提出了多方面的要求.从发展国产信息产业的角度出发,并且结合当前服务器市场的实际情况,研究和设计基于国产多核处理器的适于云计算的高端服务器产品的需求十分迫切.从系统架构、主板、系统散热等方面完成了高端服务器的产品级精细设计,且自主国产处理器适配的云计算应用软件的测试和优化,使其在硬件和软件方面获得较好的稳定性和易用性.

参考文献：

[1] The art of service.Cloud Computing-A Complete Guide to Cloud Computing: Brisbane, Australia. http://theartofservice.com

[2] Sequel.中国云计算调查报告:SACC 2010 系统架构师大会,2010.

[3] JL Hennessy, DA Patterson.Computer architecture: a quantitative approach. 5th edtion. 2012.1.

[4] Kogge P, Bergman K, Borkar S, et al, Exascale computing study: Technology challenges in achieving exascale systems[R].USA. DARPA IPTO, 2008.

[5] 刘宇航,祝明发,肖利民,等. 基于龙芯3A处理器的高效能计算结点研制[J]. 高性能计算技术.2010,6: 46-53.(Liu Yuhang, Zhu Mingfa, Xiao Limin, et al. Design of high productivity computing node based on Godson 3A CPU[J]. High Performance Computing Technology,2010,6:46-53.)

[6] 刘宇航,祝明发,崔吉顺,肖利民.基于CMP的高密度计算机多目标设计方法[J]. 系统工程与电子技术.2012,4: 806-812.

[7] 程旭,陆俊林,易江芳,等. 面向UMPC的北大众志-SK系统芯片设计[J]. 计算机学报.2008, 11.

[8] 房振满,张为华,藏斌宇. 国产通用处理器芯片进展报告.中国计算机学会.2011, 11.

[9] 谢向辉,胡苏太,李宏亮. 多核处理器及其对系统结构设计的影响. 计算机科学与探索,2008,2(6):641-650.(Xie Xianghui, Hu Sutai, Li Hongliang. Multi-core/many-core processor and its influences on computer architecture design[J].Journal of Frontiers of Computer Science and Technology,2008,2(6):641-650.)

[10] Gao Xiang, Chen Yunji, Wang Huandong, et al. System architecture of Godson-3 multi-core processors [J]. Journal of Computer Science and Technology, 2010, 25(2): 181-191.

[11] Wang Huandong, Gao Xiang, Chen Yunji, et al. Interconnection of Godson-3 multi-core processor[J]. Journal of Computer Research and Development, 2008,45(12):2001-2010.

[12] D.Culler, J.Singh and A.Gupta. Parallel computer architecture: a hardware/software approach[M]. Morgan Kaufmann Pub, 1999.

[13] 沈绪榜, 张发存, 冯国臣, 等. 计算机体系结构的分类模型[J]. 计算机学报, 2005, 28(11): 1759-1766

[14] 沈绪榜, 刘泽响, 王茹, 等. 计算机体系结构的统一模型[J]. 计算机学报, 2007, 30(5): 729-736

[15] Zhu Mingfa, Xiao Limin, Ruan Li, et al. DeepComp: towards a balanced system design for high performance computer systems[J]. Frontiers of Computer Science in China, 2010, 4(4): 475-479.