# 龙芯 3A 处理器的高效能计算节点电源管理设计

高宇辉　祝明发　刘宇航　肖利民
（北京航空航天大学 软件开发环境国家重点实验室 北京 100191）
（北京航空航天大学 计算机学院 北京 100191）
（michaeldoer@hotmail.com）

# Power management design of high-productivity computing node based on Godson-3A CPU

Gao Yuhui, Zhu Mingfa, Liu Yuhang, Xiao Limin
(State Key Laboratory of Software Development Environment, Beihang University，Beijing 100191)
(School of Computer Science and Engineering, Beihang University, Beijing 100191)

**Abstract** The power system provides the power to run the whole computing system, so the power management becomes the key to make the system run safely and reliably. Analyses the power management design problems of the high-productivity computing node based on Godson-3A CPU, especially power monitoring and power sequencing which is the most important problem of power management, and gives the solution of power management design for the computing node. From power monitoring, the solution provides voltage distribution design, unit control switch design and voltage monitor design. Through the voltage distribution design, the power system can meet the needs of all the components, and also reduce the loss of voltage conversion. The unit control switch can avoid voltage pressure drop at the boot moment, and must meet the ACPI specifications. The voltage monitor design can accurately monitor the voltage, and guarantees the stabilization of power supply quality. From power sequencing, the solution provides power sequencing design and reset sequencing design. According to the result of application, this design can make the computing node run safely, incessantly and normally, and meet the high-productivity requirements of the computing node.
**Key words** Godson-3A CPU; high-productivity; south bridge; power; power sequencing

**摘要**　电源系统作为计算机运行的“动力系统”，其电源管理设计是确保系统安全可靠运行的关键。分析了国产龙芯 3A 处理器的高效能计算节点在电源管理方面的问题，尤其是电源管理设计中最为重要的电源监控和时序控制问题，提出了计算节点电源管理设计的解决方案。电源监控方面，分为电压分配方案、单元控制开关设计和电压监控设计。电压分配方案能够确保电源满足所有器件的供电要求，同时最小化电压转换所造成的损耗。单元控制开关可以避免计算节点开机瞬间产生的较大压降，同时又能够满足 ACPI 高级配置和电源管理接口规范。电压监控设计能够对电压实施精确的监控，保证各器件的供电质量。时序控制方面，分为上电时序设计和复位时序设计。上电与复位时序设计既要满足各器件的时序要求，也应满足 ACPI 高级配置和电源管理接口规范。经过实际应用证明，该设计能够确保计算节点安全、持续和正常的运行，并满足计算节点高效能的设计需求。
**关键词** 龙芯 3A 处理器；高效能；南桥；电源；上电时序
**中图法分类号** TP368.5

---

本课题得到国家自然科学基金：分布式虚拟机监控器时钟系统性能优化方法研究（609730008）、面向沉浸式科研环境的统一资源框架与云计算模型研究（SKLSDE-2009ZX-01）以及 E 级高性能计算机文件系统中 I/O 服务器间负载均衡方法预先研究（YWF-10-02-058）的资助。

## 1. 引言

高效能计算系统（High Productivity Computing Systems，HPCS）研究计划[1]将高效能作为新一代高性能计算机研制的目标，计算节点作为高性能计算机的核心组成部分，其设计必须考虑高效能的需求。基于国产龙芯 3A 多核处理器研制的计算节点，利用了龙芯 3A 处理器高性能、低功耗、低成本的优势，并采用高密度、高集成度的板级设计，充分体现了高效能的设计特点，但这同时也给计算节点的电源系统和电源管理设计提出了更高的要求。

调查数据表明，80%造成服务器出现瘫痪的故障及用户终端 45%左右的数据丢失和出错均与电源供电质量密切相关。因此，电源管理设计是确保系统安全可靠运行的关键。计算节点为满足高效能需求而采用的设计给电源管理带来了一些设计难题，主要包括电源监控和时序控制两方面内容。

电源监控方面的问题包括：计算节点内部器件众多，包含 16 片龙芯 3A 处理器，以及多种桥片、网络芯片和电源模块等。电源管理既要满足各类器件多种电压的供电要求，也要尽量减少电压转换造成的功耗损失。保持电流稳定，确保各器件的供电质量，避免开机瞬间压降过大等问题。能够精确地监视各电压是否在工作范围内，在电压异常时采取相应措施保护芯片。

时序控制方面的问题包括：满足各芯片的上电和复位时序控制要求，确保各芯片的正确启动和运行。在满足上电和复位时序的基础上，进一步满足统一的电源管理接口规范，实现可节能的电源管理模式。此外，时序设计还需与电源管理的其它设计相适应。

## 2. 研究背景

### 2.1.龙芯 3A 处理器

龙芯 3A 处理器作为龙芯系列处理器的最新产品，采用65nm工艺制造，在单芯片内集成了四个主频为1GHz的64b超标量通用处理器核[2,3,4]。其内部集成了两级AXI（Advanced eXtensible Interface）交叉开关：一级开关用于连接四个处理器核心、四个二级Cache模块和两个HT（HyperTransport）端口，HyperTransport是一种为主板上的集成电路互连而设计的端到端总线技术[5]，用于处理器的互联和处理器的IO通信；二级开关连接四个二级Cache模块和两个64位内存控制器，GPIO、LPC、PCI/PCIX和UART等IO控制器共享同一个AXI端口也连接在第二级AXI开关上。四片龙芯3A可通过HT总线连接成共享二级缓存的CC-NUMA结构，也可以通过片外的扩展方法可以支持更多龙芯3A处理器全局地址共享的互联方式。

### 2.2. 高效能计算节点

计算节点采用 1U 机箱，内部结构图如 1 所示，共包括一个 AC/DC 电源和和两块完全相同的主板，其中一块主板平转 180 度与另一块主板通过接插件对接，每个主板均包含两个功能相同、结构对称的单元，每个单元内部采用四片龙芯 3A 处理器的环状互联标准系统，南桥芯片通过 HT 总线与其中一片处理器互联实现内部 IO 通信，与以太网芯片和 InfiniBand[6]芯片互联实现网络通信。这种高密度设计能够使单个计算节点实现 0.256Tflops 的理论峰值计算能力。

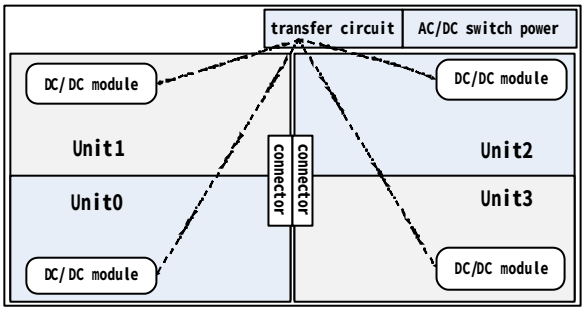


Fig.1 The structural drawing of computing node

图1 计算节点结构图

## 3. 电源监控设计

常见的服务器AC/DC电源标准有ATX和SSI（Server System Infrastructure）两种，ATX标准使用较为普遍，主要用于台式机、工作站和低端服务器，而SSI标准适用于各种档次的服务器，随着其标准日益规范化，更能适应服务器的发展。计算节点采用符合SSI规范的服务器电源，SSI电源可提供12V和3.3V_SB（Standby，待机）电压。外置电源再通过电源转接电路将12V和3.3V_SB分别引入计算节点内的四个单元。

单元内电压分配利用多种DC/DC电源模块将外部12V电压转换为指定电压提供给处理器、芯片

组、内存等芯片，应尽量较少电压转换次数，降低功耗损失；设计单元控制开关，避免开机瞬间压降过大，确保供电质量；电压监控电路可以实现对各电压的精确监控，当电压超出正常范围时由该电路向控制芯片发出告警信号。

### 3.1. 电压分配

将 12V 电压转换为各芯片所需电压即 DC/DC 转换。低压差（low drop out , LDO）线性稳压器具有结构简单、低噪声等优点[7,8]，但效率较低、发热量大、负载不能过大，无法满足计算节点高效能和散热的需求。DC/DC开关电源具有高效率、低功耗的特性[9]，因此在计算节点中所有DC/DC电源均采用开关电源。电压分配方案见图2，其中1.1V是处理器的核心电压，1.2V为HT、南桥和InfiniBand芯片的核心电压，其余电压为各芯片的IO电压或参考电压。

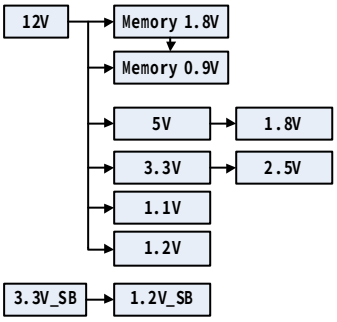


Fig.2 Voltage distribution

图2 电压分配

### 3.2. 单元控制开关

AC/DC电源提供了PS_ON信号，可由其它电路控制该电源的12V输出，3V_SB输出始终保持有效，这种设计满足ACPI规范[10]对电源管理模式的要求，可实现电源的多种电源节能管理。但由于整个计算节点仅使用了一个AC/DC电源，如果允许任一单元对该电源的PS_ON信号实施控制，将会干扰其它单元的供电。

此外，当外置电源开启时，四个单元将同时获得12V输入，如直接进行后续的DC/DC转换，瞬间过大的启动电流会造成线路压降增大，AC/DC电源将启用自我保护功能，自动关闭电源。

综合考虑以上两种情况，计算节点采用了单元控制开关的方式：首先将PS_ON信号固定为有效状态，四个单元在AC/DC电源开启时将持续获得12V输入，之后在单元内部采用PMOS电路作为12V进入单元的控制开关，由南桥的S3信号完成开关控制。根据电流的强弱可相应增加PMOS的数目，具体参见图3。

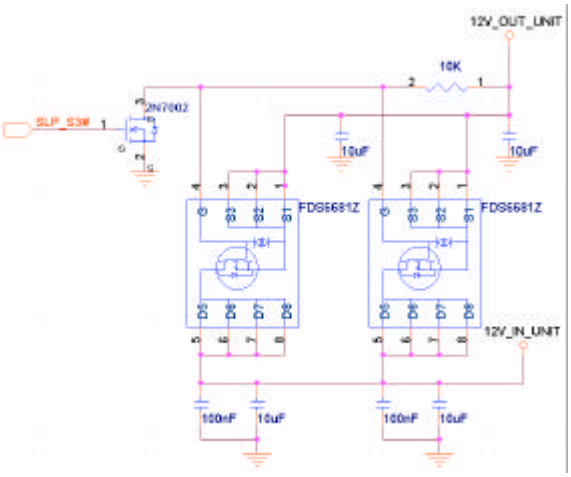


Fig.3 Unit control switch

图3 单元12V控制开关

AC/DC 电源启动时，其 3V_SB 输出将直接进入各单元内部，而 12V 输出作为 12_OUT_UNIT 受图 3 所示电路的控制。在单元内 3V_SB 有效时，南桥内部分电路工作，可检测到单元开机电路状态，如为开机状态，经过正确的上电时序后，南桥的 SLP_S3#信号为高电平，PMOS 电路因此置为打开状态，12V 进入单元内部，此后该单元继续完成上电时序。

通过以上设计，每个单元使用独立的控制开关，因此可以避免上述的两个问题，同时也为大规模机群系统中设计统一的开机管理提供了方便。

### 3.3. 电压监控设计

电压监控电路用于确定指定电压是否在工作范围内，这可以通过由精密电阻分压器、比较器和基准电压源组成的电路来实现，具体参见图4。监控结果可输出给相应控制电路。

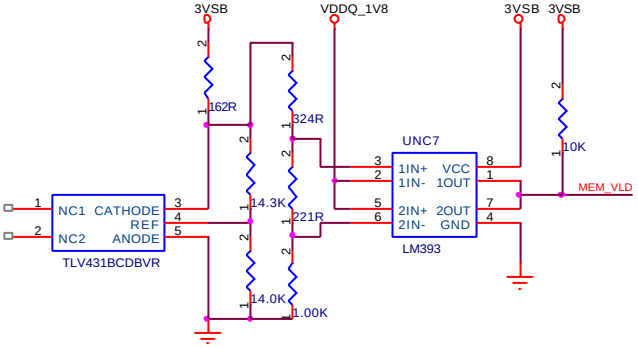

Fig.4 Voltage monitoring circuit

图4 电压监控电路

## 4. 时序控制设计

电压分配为各种芯片提供核电压和 IO 接口电压，每种芯片的正常工作不仅取决于这些电压的工作范围，而且取决于各种电压间的时序控制，包括先后次序和延时等。时序控制包括上/断电时序控制和复位时序控制。由于上电时序和断电时序通常是完全可逆的，在设计完成上电时序后，只要验证该控制电路能够满足断电时序的要求即可。复位时序包括复位掉电和复位上电两个阶段，通常是掉电时序和上电时序的子集，因此也只需在设计完成上电时序后进行验证。

### 4.1.系统上电及复位相关信号

计算节点采用的 MCP68 南桥芯片[11,12]提供了控制输出和状态输入管脚，这些管脚与处理器、其它芯片及时序控制电路相连完成系统的上电及复位时序控制，具体参见图 5。时序控制电路可分为两部分：电压确认及延时电路和 DC/DC 模块控制电路。

电压确认及延时电路将外部电压平面的状态准确的反应给南桥芯片，并保证南桥芯片对状态输入信号的有效反应时间，这可以通过电压监控电路和 RC 电路来实现。DC/DC 控制电路完成将南桥输出控制信号进行适当转换后使能各种 DC/DC 电源模块，随后 DC/DC 电源的输出将导致相应的电压平面状态发生变化，再通过电压确认及延时电路将状态信号反馈给南桥芯片，由此，逐一完成各种电压的生成和确认。

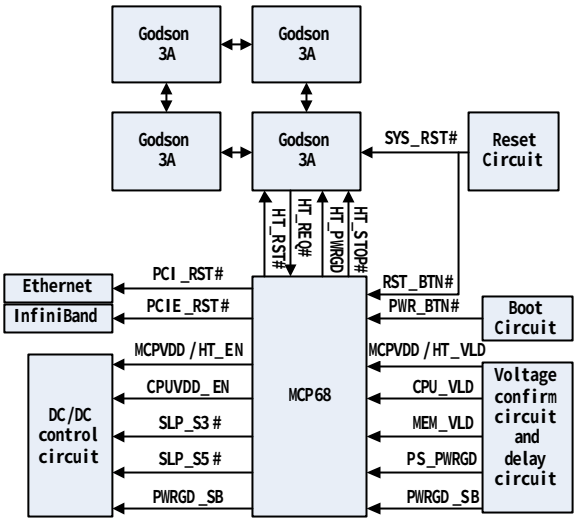


Fig.5 System power and reset signals

图5 系统上电及复位相关信号

### 4.2. MCP68 上电时序设计

MCP68 上电时序支持最新的 ACPI 3.0 规范，可支持 ACPI 的 6 种工作状态模式，分别是 S0~S5。S0 表示正常工作状态；S1 表示 POS（Power on Suspend)，这时除了通过 CPU 时钟控制器将 CPU 关闭之外，其他部件仍然正常工作；S2 表示此时 CPU 处于停止运作状态，总线时钟也被关闭，但其余的设备仍然运转；S3 表示 STR（Suspend to RAM），此时系统主电源关闭，但利用待机电源为内存供电，确保内存数据不丢失；S4 表示 STD（Suspend to Disk），此时内存数据保存到硬盘，被唤醒后可立即恢复之前状态；S5 表示连电源在内的所有设备全部关闭。其中，S1、S2 和 S4 状态可以在硬件实现 S3 和 S5 状态的基础上通过操作系统实现。

参见图 6，上电时序具体如下：

1) 3.3V_BAT（电池电压）有效；
2) 由 AC/DC 直接提供的 3.3V_SB 有效；
3) 南桥输入晶振开始工作，同时发出 SUS_CLK 通知外部设备；
4) 外围电路确认 3.3V_SB 电压有效后，向南桥输入 PWRGD_SB 高电平信号；
5) 南桥置输出 SLP_S5#信号为高电平，通知外围设备由 S5 状态向其它状态转换；
6) 外围电路建立内存相关电压平面；
7) 外围电路确认内存电压有效后向南桥输入 MEM_VLD 高电平信号；
8) 南桥置输出 SLP_S3#信号为高电平，通知外

围设备由 S3 状态向其它状态转换；

9) 外围电路建立 IO 接口电压平面；
10) 外围电路确认接口电压有效后向南桥输入 PS_PWRGD 高电平信号；
11) 南桥置输出 CPUVDD_EN 信号为高电平，通知 CPU 核电压的 DC/DC 模块工作；
12) 外围电路确认 CPU 核电压有效后向南桥输入 CPU_VLD 高电平信号；
13) 南桥置输出 MCPVDD/HT_EN 信号为高电平，通知 MCP 和 HT 核电压的 DC/DC 模块工作；
14) 外围电路确认 MCP 和 HT 核电压有效后向南桥输入 MCPVDD/HT_VLD 高电平信号；
15) 南桥对外输出各种时钟信号，使得相关芯片正常工作。

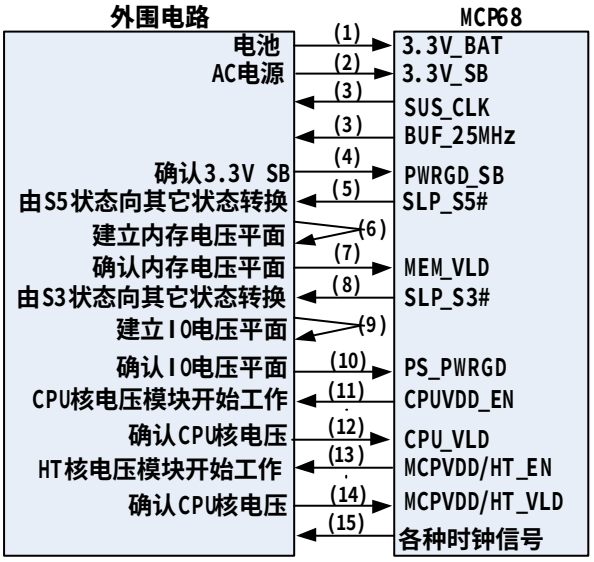


Fig.6 power sequencing diagram

图6 上电时序图

分析 MCP68 断电时序与复位时序后，发现利用上述的上电时序控制电路可同时实现断电时序与复位时序。

### 4.3. 单元上电时序设计

按照以上 MCP68 上电时序设计：上电时在 SLP_S5#由低电平转为高电平之后，SLP_S3#还保持低电平之前，MCP68 通知外围设备由 S5 状态转为 S3 状态；掉电时在 SLP_S3#由高电平转为低电平之后，SLP_S5#还保持高电平之前，MCP68 通知外围设备由其它状态转为 S3 状态，在这两种情况下电源都只需向内存供电，因此可以使用 SLP_S3#信号控制 AC/DC 电源的 PS_ON 信号，在 S3 状态下由 3.3V_SB 提供内存供电。相应的电压分配如图 7，通过切换开关实现 3.3V_DUAL 的双电压平面，再由 3.3V_DUAL 统一生成内存相关电压。

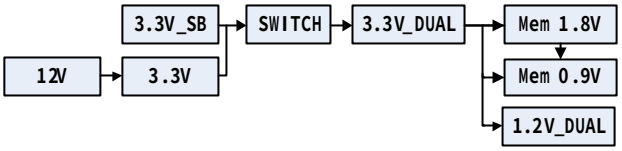


Fig.7 Voltage distribution according to MCP68

图 7 MCP68 时序规定的电压分配图

计算节点中为了避免开机瞬间电流过大已经将 AC/DC 电源的 PS_ON 信号固定为有效状态，同时为了提高电源的转换效率直接由 12V 生成内存相关电压，因此单元上电时序必须对上述时序做调整。首先将电压分配图细化为图 8，内存相关电压在 S3 状态和正常工作状态下均由单元外的 12V 产生，内存电压 DC/DC 模块均工作，但负载不同。由 SLP_S5#信号控制内存电压 DC/DC 模块的使能，由 SLP_S3#信号控制图 3 所示的单元控制开关。经分析和测试，这种实现方式仍符合 MCP68 的上电时序要求。

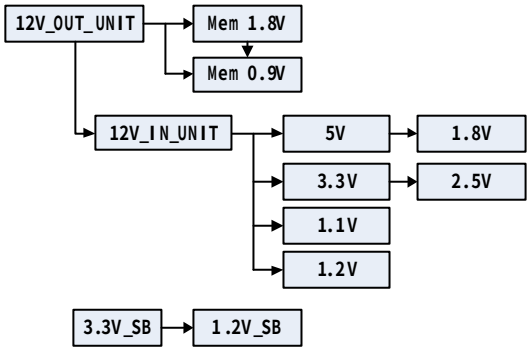


Fig.8 Adjusted voltage distribution

图8 细化后的电压分配图

## 5. 实验测试

对高效能计算节点进行了实际测试，龙芯 3A 处理器主频采用 725MHz，内存频率为 180MHz，启用以太网和 InfiniBand 网络，测试工具采用 Linpack。

### 5.1. 电源监控测试

图 9 为开机瞬间 AC/DC 电源输出 12V 的变化情况。由测试结果可见，单元控制开关能够在开机瞬间有效保持电流和电压的稳定。

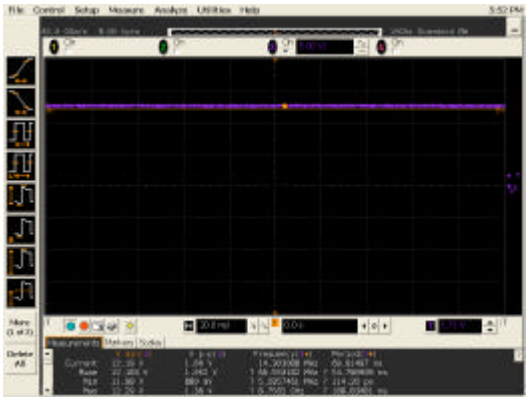

Fig.9 12V signal at the boot moment

**图 9 开机瞬间 12V 变化情况**

**启动计算节点内全部四个节点进行 Linpack 测试时，整机功耗经测试小于 300W，满足高效能的设计要求。**

### 5.2. 时序控制测试

**经测试，计算节点能够完成正确的启动、关机和复位功能，其中开机过程中，S3 和 S5 信号变化见图 10，其中 S5 与 S3 信号有效时刻相差 15 毫秒，符合设计要求和 ACPI 规范要求。配合 Debian5 操作系统，计算节点能够正确实现了 S0~S5 状态，因此满足计算节点高效能的设计需求。**

Fig.10 S3 and S5 signals

**图 10 开机时 S3 与 S5 信号**

## 6. 结束语

**本文基于龙芯3A处理器的高效能计算节点，提出了一种电源管理设计方案，详细介绍了电源监控和时序控制两个方面的内容。该设计还存在着有待改进的地方，比如：利用RC电路无法实现精确的延时控制；电压监控分散在多个电路中，可采用芯片实现集中式监控等。**

## 参 考 文 献


[1] DARPA. High productivity computing systems (HPCS) program [EB/OL]. 2002[2010-05-30]. http:// www.darpa.mil/ipto/programs/hpcs/index.htm.

[2] Hu Weiwu, Zhang Fuxin, Li Zusong. Microarchitecture of the Goodson-2 processor [J]. Journal of Computer Science and Technology, 2005, 20(2): 243-249

[3] Hu Weiwu, Wang Jian, Gao Xiang. Micro-architecture of Godson-3 multicore processor [C/OL] //Proc of the 20th Hot Chips. 2008 [2010-05-30] . http://www. hotchips.org/hc20/mainpage.htm

[4] Hu Weiwu, Wang Jian, Gao Xiang. Godson-3: A Scalable Multicore RISC Processor with X86 Emulation [J]. IEEE Micro, 2009, 29 (2): 7229.

[5] HyperTransport Technology Consortium. Hyper Transport TM I/O Link Specification Revision 1.03 [M/OL].2001[2010-05-30]. http://www.hypertransport.org/default.cfm?page=HyperTransportSpecifications

[6] InfiniBand Architecture Specification [EB/OL]. 2000 [2010-05-30]. .http://www.InfiniBandta.org/

[7] Kwok K C, Mok P K T. Pole-zero tracking frequency compensation for low dropout regulator . IEEE J International Symposium on Circuit s and Systems, 2002, 4 :735

[8] Lee H, Mok P K T. Design of low-power analog drivers based on slew-rate enhancement circuits for CMOS low-drop-out regulators. IEEE Trans Circuit s Systems II: Express Briefs, 2005, 52 (9): 563

[9] Erickson R W, Maksimovic D. Fundamentals of power electronics [M]. 2nd Edition. New York : Kluwer Academic Publishers, 2001.

[10] Advanced Configuration and Power Interface Specification 3.0[S]. 2006.

[11] nVidia Corporation. MCP68 Media and Communications Processor Datasheet. Revision 04[Z]. 2007.

[12] nVidia Corporation. MCP68 Media and Communications Processor Design Guide. Revision 04[Z]. 2007.



**Gao Yuhui,** born in 1978. Postgraduate, Engineer. His main research area is computer architecture.

**高宇辉，1978 年生，男，硕士研究生，工程师，主要研究方向为计算机体系结构。**



**Zhu Mingfa,** born in 1945. Ph.D, Professor, Senior membership of China Computer Federation. His main research areas are computer architecture, computer system software, high performance computing, virtualization and cloud computing.

**祝明发，1945 年生，男，博士，教授，博士生导师，CCF**

**高级会员。主要研究领域：计算机体系结构、计算机系统软件、高性能计算、虚拟化与云计算。**

**Liu Yuhang,** born in 1985. Ph.D candidate. His main research areas are high performance computer architecture, and parallel processing.

**刘宇航，**1985 **年生，男，博士生。主要研究领域：高性能计算机体系结构、并行计算。**

**Xiao Limin,** born in 1970. Ph.D, Professor, Senior membership of China Computer Federation. His main research areas are computer architecture, computer system software, high performance computing, virtualization and cloud computing.

**肖利民，**1970 **年生，男，博士，教授，博士生导师，**CCF **高级会员。主要研究领域：计算机体系结构、计算机系统软件、高性能计算、虚拟化与云计算**

**Research Background**

In the high-productivity computing node based on Godson-3A CPU, the power management is the key to make the system run safely and reliably. The most import problem is how to solve the power monitoring and power sequencing of computing node. From power monitoring, the solution provides voltage distribution design, unit control switch design and voltage monitor design. Through the voltage distribution design, the power system can meet the needs of all the components, and also reduce the loss of voltage conversion. The unit control switch can avoid voltage pressure drop at the boot moment, and must meet the ACPI specifications. The voltage monitor design can accurately monitor the voltage, and guarantees the stabilization of power supply quality. From power sequencing, the solution provides power sequencing design and reset sequencing design. This study is sponsored by the fund of the State Key Laboratory of Software Development Environment under Grant No. SKLSDE-2009ZX-01, the Fundamental Research Funds for the Central Universities under Grant No. YWF-10-02-058 and the National Natural Science Foundation of China under Grant No. 60973008.