第 35 卷 第 6 期 计 算 机 工 程 2009 年 3 月
Vol.35 No.6 Computer Engineering March 2009
·博士论文· 文章编号：1000—3428(2009)06—0017—03 文献标识码：A 中图分类号：TP303

曙光 5000A 高效能计算节点的设计与实现

曾 宇[1,2]，王 洁[1,2]，孙凝晖[1]
(1. 中国科学院计算技术研究所计算机系统结构重点实验室，北京 100080；2. 中国科学院研究生院，北京 100039)

摘 要：由于求解问题和系统规模的不断扩大，基于 cluster 架构的高性能计算机面临扩展性、可靠性、功耗、占地面积、均衡性等诸多挑战。该文针对计算模块、交换管理模块、自适应功率管理、专用 FPGA 硬件加速部件、高速 PCI-E 全交换扩展等方面，设计并实现高效能计算节点。基于该节点构建的曙光 5000A 百万亿次计算机能有效解决计算密度、I/O 扩展及带宽瓶颈和能耗等方面的瓶颈。
关键词：高效能；节能；硬件加速

Design and Realization of Dawning 5000A High-productivity Computing Node

ZENG Yu[1,2], WANG Jie[1,2], SUN Ning-hui[1]
(1. Key Laboratory of Computer System and Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080;
2. Graduate University of Chinese Academy of Sciences, Beijing 100039)

【Abstract】 As for the scale of the problem and the system continues to expand, the cluster-based high-performance computer is facing scalability, reliability, power consumption, footprint, balance, and many other challenges. This paper introduces the design and realization of high-productivity computing node such as computing module, switch module, management module, adaptive power management, FPGA-based hardware accelerator board, high-speed PCI-E switch extend module and other aspects. It resolves the computing density, I/O expansion and bandwidth bottleneck as well as energy consumption and other bottlenecks in Dawning 5000A 100 Teraflops supercomputer based on the high-productivity computing node.
【Key words】 high-productivity; energy saving; hardware accelerate

1 概述

随着社会的发展，机群逐渐成为市场主流，但其简单的体系结构在耗电、空间、散热、效率、可靠性和可管理性方面的问题使其性能无法延续到千万亿次(Petaflops)[1]。美国国防部于 2002 年制定的高效能计算系统(High Productivity Computing Systems, HPCS)研究计划，首先提出以高效能作为新一代高性能计算机研制的目标，IBM PERCS, Cray Cascade, SUN Hero 成为首批入选计划。高效能包含了高性能、可编程性、可移植性、稳定性等多个方面的要求[2]。其他千万亿次研发计划，如 IBM Roadrunner, Cray Baker, SUN Constellation、日本京速计算机计划等，也将高效能列为其关键实现目标。高效能代表了高性能计算机研究的新方向[3-4]。

计算节点(主要由处理器、芯片组和交换芯片构成)是高性能计算机的核心硬件部件，不同的高性能计算机体系结构有不同的实现方式。MPP 系统多采用 CPU+Router 方案，Router 把芯片组和交换芯片集成在一个芯片中，一个端口连接 CPU，其他端口连接相邻的 Router，构成 3D-Torus 网络，如 Cray XT3 的 Seastar 芯片。IBM BlueGen/L 没有独立的 Router 芯片，而是集成在 CPU 中[4]。cc-NUMA 系统采用 CPU+芯片组+交换机的方案，如 SGI Altix4000 的芯片组 SHUB 通过共享总线连接 2 个 Itannium2 CPU，通过网络接口连接 NUMA Link 交换芯片。Cluster 系统采用 CPU+芯片组+NIC+交换机的结构，芯片组提供标准的 I/O 接口(如 PCI-E)，高速网络接口采用独立芯片实现，一端通过标准 I/O 接口与芯片组连接，一端连接互联网络交换芯片。节点的结构设计主要考虑密度、散热、扩展和管理要求。Cray XD1 节点采用 CPU 卡+底板结构，Cray XT3 节点采用 CPU 板+通信背板结构，IBM BlueGene/L 节点采用计算卡+中板结构。

2 计算节点的总体结构设计

曙光 5000A 全系统采用 90 个计算节点，每个计算节点包含 10 个计算模块、1 个 20G Infiniband 交换模块、2 个 Gigabit Ethernet 交换模块和冗余管理模块。每个计算模块采用 4 个 AMD 64 位 2.0 GHz Barcelona 4Core CPU，系统实现了 115.2 Teraflops 的理论峰值计算能力，计算节点总体结构如图 1 所示。

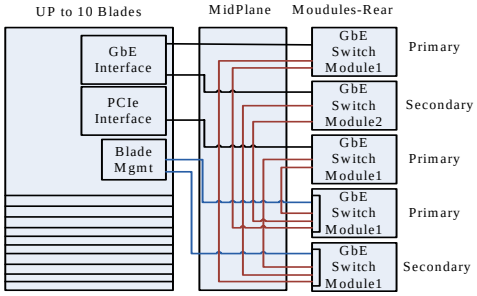


图 1 计算节点总体结构

基金项目：国家“863”计划基金资助重大项目(2006AA01A102)
作者简介：曾 宇(1973－)，男，高级工程师、博士研究生，主研方向：高性能计算机体系结构；王 洁，博士研究生；孙凝晖，研究员
收稿日期：2008-08-19 **E-mail：**zengyu@ncic.ac.cn

2.1 计算模块

计算模块采用 4 路 SMP 架构，基于 AMD 4 core 64 位 Barcelona 处理器，16 条 DDR2 服务器内存，PCI-E 8X 至中板，可提供生物、加密等硬件加速功能扩展、网络扩展及存储区域网络功能扩展，如 Fiber Channel 等。计算模块监控子卡 BMC 基于 485 总线和管理模块通信，有两大功能：监控管理本地各项环境参数，并传递参数到管理模块；配合管理模块，提供总线接入的管理功能。

2.2 交换模块

曙光 5000A 全系统采用 CLOS 网络构建。每个计算节点上的 2 层 Infiniband 交换模块上传级联端口全等分设计(即每组 2 条上行链路)。InfiniBand 交换模块其外接端口包括：1 个与中板相连的内部端口(PCIE 8X)；10 个与计算节点外部 InfiniBand 交换机相连的外部接口(InfiniBand 4X)；与管理模块相连的 IIC 接口以及其他控制信号和电源接口。交换模块采用第 3 代 Infiniband Switch 芯片 InfiniScale III MT47396D 芯片；集成线性数据包缓存。提供 10 个 downlink 4X ports 与 10 个 blade 相连，10 个 uplink 4X ports 外联，每个 port 的传送速率 20 Gb/s；共提供 400 Gb/s 传送速率，800 Gb/s 的交换带宽。80 条集成的 5 Gb/s DDR version SerDes，SerDes 传送速率能够静态分配或自适应。支持 VL(Virtual Lane)划分；同时在交换机内部采用了一个嵌入式 CPU 进行本地初始化、管理模块进行通信和上层管理使用，此外这个处理器还要提供 IB 网络管理功能，作为 IB 子网的 Subnet Manager。Infiniband 交换机可通过 SNMP, Web/GUI, CLI 等方式远程管理，支持冗余管理模块。Infiniband 交换模块其设计框图如图 2 所示。

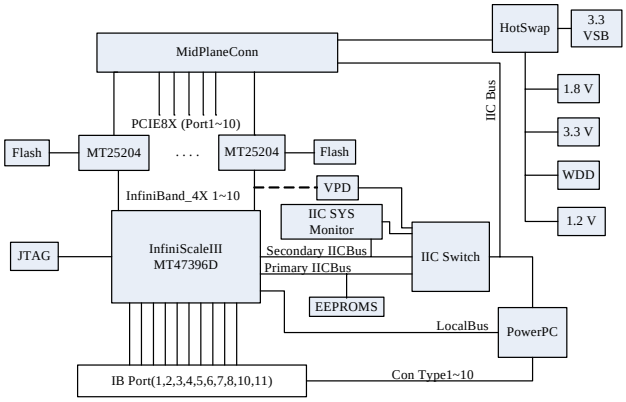


图 2　InfiniBand 交换模块设计框图

计算节点内置 2 个千兆模块以支持百万亿次扩展分区全线速无阻塞骨干网和非线速骨干网连接。分区全线速无阻塞骨干网连接所有计算节点，用于千兆计算网络通信；非线速骨干网连接所有被管理的设备，用于管理等功能。

2.3 管理模块

管理模块架构如图 3 所示，左侧为管理模块对外接口，分别为 KVM 部分 PS2 键盘鼠标接口和 VGA 接口，用于连接外置的键盘鼠标和显示器；Share Media 部分的 USB 接口，用于连接外置移动存储介质。另外，还存在一个系统网络接口，用于刀片系统网络管理接口，以及 KVM 的远程管理。

系统存在 5 种管理和监控网络，分别实现不同的功能如下：(1)KVM 网络，进行刀片 KVM 系统管理；(2)USB 网络，进行刀片 Share Media 和 Virtual Media 管理；(3)RS485 监控网络，刀片监控和上下电控制；(4)I2C 网络，进行资源信息管理和监控；(5)100 Mb/s 管理网络(交换机以及管理模块心跳)。

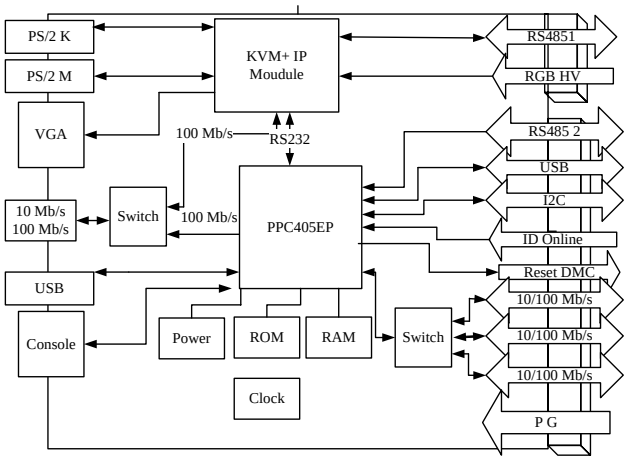


图 3　管理模块架构

2.4 中板及 I/O 扩展模块

在架构设计中，将 PCI-E 总线引入中板，这样可以根据系统需要灵活地设计各种高速交换设备，如 InfiniBand 交换、Myrinet 交换、FC 交换等，同时外接的 PCI-E 插槽可以很好地支持专用的 ASIC 硬件加速卡，使很多商业应用可以通过硬件加速获得几倍到几十倍的性能提升，最大限度地满足用户的需求。高效能计算节点中板带宽和延迟是其性能瓶颈。在高效能计算节点设计中，中板总带宽为 42.5 Gb/s。

计算节点同时设计了高效 PCI-E 全交换 I/O 扩展模块，每个计算模块采用 1 颗透明桥芯片 IDT PES8NT2，中板采用 2 颗 64 通道，16 口交换芯片 IDT PES64H16，系统提供高达 18 Tb/s 的单向 I/O 聚合带宽。I/O 扩展模块和计算模块之间采用 PCI-E 4X 总线实现互联，PCI-E 全交换 I/O 扩展模块系统架构如图 4 所示。

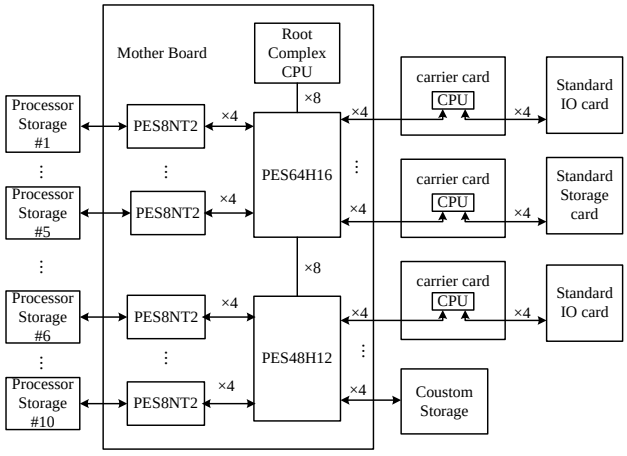


图 4　PCI-E 全交换 I/O 扩展模块系统架构

3　节能设计

高效能计算节点具备自动功率管理功能，主要体现在 3 个层面：

(1)根据实时功耗确定工作电源个数，使电源工作在最佳效率曲线上。系统电源功率设计分成静态功率设计及动态功率设计。控制方式如下：上电前电源功率预分配，计算模块把 Flash 中存储的最大满配功率需求发送给管理模块进行审批，等计算模块完成了 BIOS 自检之后还要提交电源实际分配功率给管理模块，由管理模块进行最佳电源效率判断，并

确定是否需要关闭或打开电源。控制算法如下：

输入 电源集合 $\{D_1, D_2, \cdots, D_n\}$，最大满配功率 *GSUM*，初始满功率需求 *G*

输出 电源集合 $\{D_1, D_2, \cdots, D_k\}$ 子集 SET，最佳电源功率 $GSUM\_SET = \sum_{i \in SET} D_i$，*k* 为实际工作的电源个数

```
PROCEDURE GEN_GSUM_SET
SET ← ∅, GSUM_SET ← 0, k = 0
/*上电前电源功率预分配*/
FOR i = 1 TO n DO
  { IF GSUM_SET <= G THEN{
        IF  i ∉ SET  THEN{
            GSUM_SET ← GSUM_SET + D_i
            SET ← SET ∪ {D_i}
             k = i
    } } }
/*实时调整电源功率*/
REPEAT{
Computing G_BEST
 IF GSUM_SET is not best THEN
 { IF GSUM_SET<G_BEST THEN{
        REPEAT{
            GSUM_SET ← GSUM_SET + D_{k+1}
            SET ← SET ∪ {D_{k+1}}
            k = k+1
          }UNTIL GSUM_SET=G_BEST
       }ELSE{REPEAT{
            GSUM_SET ← GSUM_SET – D_k
            SET ← SET – {D_k}
            k = k-1
       }UNTIL GSUM_SET=G_BEST
    }  }
}UNTIL Computing is finished
```

(2)优化计算刀片操作系统内核。通过优化程序执行队列或根据负载情况动态调整 CPU 频率。主要原理是对 CPU 的运行状态进行计算，分析任务队列，对不同时刻进行功耗计算，同时建立 CPU 散热器的散热模型，各个内部事件所需要的能耗为

$$E=a\times c1+b\times c2+\cdots$$

其中，*E* 代表能耗；*c* 代表所记录的事件发生的次数；*a* 和 *b* 则是这些事件对应的参数，这些参数可以通过实验来确定。任务在当前时间片中的能耗指数平均值在下式中给出：

$$\overline{x_i} = p \times x_i + (1-p) \times \overline{x_{i-1}}$$

前面一项是任务在当前时间段中的能耗，后面一项是任务在之前时间段中的能耗。参数 *P* 是一个变量，与当前时间片的长短有关，如果当前时间片大于设定值，那么将选择相对大的 *P*，反之亦然。

在工作过程中，尽量把功耗高的任务与功耗低的任务交叉进行，这样可以保持 CPU 在稳定的负载下运行，减少热能的散发并提高运行效率。同时，当发现 CPU 任务队列对功耗需求较低，则通过 BIOS 接口进行 CPU 功率的动态调整。

(3)多计算模块任务调整调度。首先从管理模块中获得各计算模块实际负载情况，一旦发现某计算模块实际利用率较低，查询对其进行控制的负载均衡和作业调度系统，并重新进行工作状态的调整。工作状态的调整分为多级：最轻量级的调整只是降低 CPU 的主频和内核电压至工作状态最低允许值。计算节点提供基于 ACPI 的工作状态调整模式，分别是 *S*0~*S*5。*S*0 表示正常工作状态；*S*1 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；*S*2 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；*S*3 表示 STR(Suspend to RAM)，这时的功耗不超过 10 W；*S*4 表示 STD(Suspend to Disk)，此时系统主电源关闭，但硬盘仍然带电并可以被唤醒；*S*5 表示连电源在内的所有设备全部关闭，功耗为 0。在 *S*1 状态下系统功耗将比空载状态再降低 30%，唤醒时间小于 3 s，在 *S*4 状态下功耗相比空载状态降低 70%以上，唤醒时间小于 1 min。

4 专用硬件加速器设计

由于 FPGA 可以根据不同的应用实现可重构计算，适应 HPC 面临的不同的计算模型，同时 FPGA 在内存带宽、并行处理和低功耗方面有突出的优势，因此与主处理器配合，可实现提高特定应用性能和降低系统功耗的双重目标，应用前景广阔，是实现高效能计算的有效途径之一。

高效能计算节点专用硬件加速器体系结构如图 5 所示，支持标准 PCI-X、PCI-E 总线协议、1 GB 片上 DDR2 内存、Vertex5 系列 FPGA 芯片，其内部工作频率为 200 MHz。

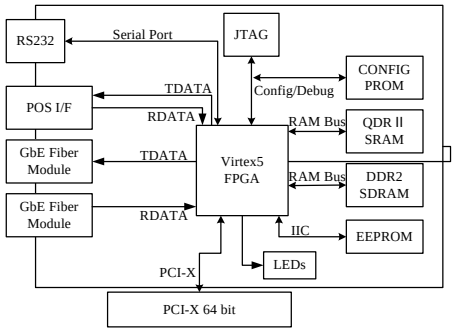
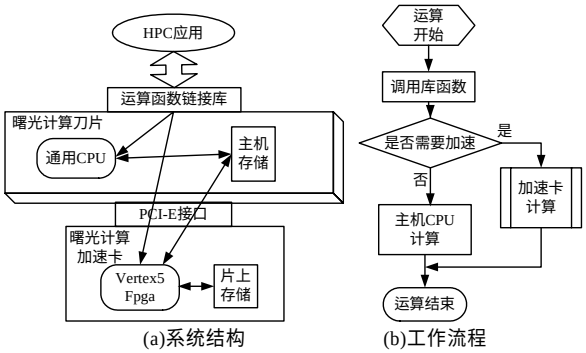


图 5 高效能计算节点专用硬件加速器

应用程序对数学运算库中函数的调用，会根据计算任务的特征，自动分配到主机 CPU 或加速器上运行。如果一个计算任务分配给加速器，加速器会访问主机的内存，获取原始数据，进行计算，并在板载内存中缓存计算的中间结果，当计算完成后，加速器会把最终结果存入主机内存，并通知主机 CPU。整个系统的结构和运算加速过程如图 6 所示。



(a)系统结构　　(b)工作流程

图 6 专用硬件加速器结构与工作流程

FPGA 内部完全用硬件逻辑实现计算加速，FPGA 内部的逻辑主要包括 PCI-E 控制、内存控制、指令解析、并行计算模块等组成。测试结果表明，与 1 个 Intel Xeon 2.8 GHz CPU 相比，单个专用加速器运行全局序列联配算法时，最高可以

(下转第 22 页)