基于龙芯 3A 处理器的高效能计算结点研制*

刘宇航[1,3] 祝明发[1,2] 肖利民[1] 高宇辉[1,3]
（1. 北京航空航天大学计算机学院 北京 100191；2. 联想集团 北京 100085；
3. 北京航空航天大学软件开发环境国家重点实验室 北京 100191）

摘 要 计算结点是超级计算机的核心部件，是构建超级计算机系统的重要基石。本文设计并研制了一款基于国产龙芯 3A 四核处理器的 16 路高效能计算结点机：首先研究了计算结点设计和研制中的一些关键问题；在此基础上，给出了计算结点系统总体结构以及处理器互联、时钟、上电与复位、内存、主板和结构等子系统的设计和实现方案，并对该结点机进行了功能和性能的评测。与当前市场主流的计算结点机相比，本文研制的结点机在性能功耗比、集成度等方面有较大优势。

关键词 计算结点 龙芯 3A 片上多处理器 高效能 超级计算机
中图法分类号 TP391

1 引 言

高性能计算机是科技创新的重要基础设施，对几乎所有的科学与工程领域（例如，美国高性能计算和通信（HPCC）计划指出的重大挑战性应用车辆及飞行器动力学、全球气候变化、人类基因、流体湍流等）的发展都具有巨大推动作用。这些领域对计算性能的需求呈现日益持续增长的趋势。同时，商业计算等事务处理领域也对计算性能提出了越来越高的需求[1]。

随着超大规模集成电路（VLSI）制造工艺的提升，芯片级、主板级、系统级的设计呈现出高集成度的特点，高性能计算机已步入千万亿次（Petascale）并向百亿亿次（Exascale）超级计算发展的时代，但当前面临着多方面的严峻挑战：

第一，扩展性（Scalability）问题，随着系统规模的扩张，延时成为影响并行应用性能的一个重要因素，应用的实际加速比和系统计算效率均成为问题；

第二，存储墙（Memory Wall）问题[2-3]，一直以来就是系统设计时必须考虑的一个重要方面。内存和 I/O 性能与处理器性能之间的差距呈日益扩大的趋势；

第三，编程（Programming）问题，大量处理器和并发线程背景下的并行编程是高性能计算机设计者和用户都必须面对的严峻挑战；

第四，功耗（Power Consumption）问题，功耗对系统可靠性、散热制冷、电能等众多方面都产生了重要影响[4]，高性能计算机的能效问题得到重视，以 MFLOPS/W 为指标的 Green500 超级计算机排行榜已成为 TOP500 排行榜的补充。

同时，超级计算机在系统的可靠性、容错、复杂性等方面也面临诸多挑战。

美国国防部于 2002 年制定的“高效能计算系统”（HPCS：High Productivity Computing Systems）研究计划首次提出了以高效能作为新一代高性能计算机研制的目标[5-6]。高效能代表了高性能计算机研究的新方向，对作为超级计算系统构建块的计算结点提出了可扩展性好、内存和 I/O 性能高、易编程、低功耗等多方面的要求。

同时，面向 P 级和 E 级计算的并行计算机，对这些方面提出更迫切的要求[4]。

计算结点是超级计算机的核心系统部件，是构建超级计算机系统的重要基石。设计与研制高性能功耗比、高集成度、低成本的计算结点机，对于应对上述挑战，实现大规模的高效能计算机系统具有重要意义。本文设计研制了一款基于国

收稿日期：2010-09-15
作者简介：刘宇航（1985-），男，北京航空航天大学博士研究生，主要研究方向为高性能计算机体系结构和并行计算。
*本文获得国家 863 计划重大项目课题（No. 2007AA01A127）、国家自然科学基金项目（No. 60973008）、软件开发环境国家重点实验室探索性自主研究课题（No. SKLSDE-2009ZX-01）资助。

产龙芯 3A 四核处理器 16 路高效能计算结点机。

2　系统设计的基本考虑

2.1　处理器及其互连结构

并行计算机是一组互相通信、相互协作以快速求解大型问题的处理单元。处理单元的选择和处理单元之间互连结构的设计是并行计算机设计的重要内容。

一方面，高效能计算系统计算能力的发挥需要均衡设计处理器单元、存储系统、I/O 和互连网以使之协同运行；另一方面，对于许多并行应用而言，提高单个处理器的能力仍然是提升整个并行计算系统能力的有效方法之一。

采用片上多处理器可不过度追求单核的高主频，多个较低主频的核构成的 CPU 功耗相对于高主频单处理器会有能效优势，同时由于片内通信延迟较小，容易增加片内的聚合计算能力，所以片上多处理器（CMP：Chip Multiprocessor）将可能是构建千万亿次级及更大规模并行计算系统的基础。通过先进互连网络使基于 CMP 的计算结点协同工作是超级计算较为可能的解决方案之一。

2.1.1　处理器

从自主创新和安全可控的角度，使用国产处理器作为超级计算机的核心部件可扭转长期以来依赖从国外进口处理器的被动局面，对提高我国高性能计算机的自主研发水平、维护国家安全具有重要意义。

龙芯 3A 处理器是中国科学院计算技术研究所自主研制的一款面向高端计算的片上多处理器。片内集成四个 64 位的四发射超标量主频 1GHz 的 GS464 高性能处理器核。每个处理器核流水线采用四发射动态超标量，10 级的超流水线结构，支持寄存器重命名、动态调度、分支预测和其他乱序执行技术[7-9]。

从缩短升级周期和减少升级成本角度，如果系统处理器等部件因升级换代被更换时，其他部件基本不动，系统能以更高的性能工作，对大规模超级计算机具有重要意义。微处理器芯片升级换代较快，一年半左右性能就翻一番，如果每次处理器芯片升级都要重新设计超级计算机系统的结点机，超级计算机系统的升级换代就较为困难。

当前，国际上可扩展并行系统的主流是从芯片级集成过渡到板级集成、系统级集成。由于龙芯 3B 八核处理器与龙芯 3A 四核处理器具有相同的封装，高效能结点机具有系统级集成的优点，可保证计算机系统与处理器同步升级，升级延迟较小。这也是龙芯 3 系列处理器的优势之一。

2.1.2　互连结构

无论是芯片级、主板级，还是系统级的设计，互连网拓扑结构的选择都是设计决策中最基本、最重要的问题之一。首先，芯片级、主板级及系统级处理器之间的读写访问通信需要经过互连网实现，互连网的效率直接关系到系统性能；其次，芯片级、主板级及系统级可集成的处理器数量，将随着集成电路工艺水平的提升而进一步增加，对互连这些处理器网络的设计选择将直接决定整体系统的规模可扩展性。互连结构应既能满足当前规模下系统需要，又能为未来的规模扩展提供潜在的设计空间；第三，对于较为流行的分布式共享存储系统 DSM 尤其是 CC-NUMA 而言，高速缓存一致性的维护需要结合互连结构实现，互连网的特征直接影响缓存管理机制的实现[8]。

评价度量一个互连网结构，一种直观的认识是，度和直径是互连网结构较重要的两个参数。度是与系统成本相关的参数之一。度越大，处理机的接口越复杂，造价越高，而且链路数将越多，系统的成本将随结点的度增大而增大。度和直径相互影响，同时这两个参数还影响系统的多重路由能力、可扩展性。结点的度越大，网络的直径越短，多重路由能力越好。若互连网度的大小不随结点数的增加而增加，将有利于网络的扩展[10]。

本文提出一个衡量性能和成本（P_C：Performance and Cost）的指标如下：

$$P_C(t,N,Topo)=\frac{1}{L(t,N,Topo)*M(t,N,Topo)}$$

$$L(t,N,Topo)=\frac{1}{N}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}p_{ij}(t)d_{ij}(t)$$

$$M(t,N,Topo)=M(N,Topo)=\frac{1}{2N}\sum_{k=0}^{N-1}\lambda_k$$

$p_{ij}(t)$ 为在整个系统范围内结点对 (i,j) 之间通信的概率。$d_{ij}(t)$ 为结点对 (i,j) 之间的最短路径长度。$d_{ij}(t)$ 之所以可能随 t 变化，是因为在实际应用运行期间，存在多源情形（Multi-Source Situations）的通信，不同通信对之间的最短路径可能路由冲突。在路由冲突时实时最短路径长度可能大于静态最短路径长度，而且实时最短路径

长度可能是随时间变化的。

以在整个系统范围内任意结点对 (i, j) 之间通信的概率相同（均匀分布）时为例，分析队列（Array）、环（Ring）、二叉树（Binary Tree）、二维网格（2D-Mesh）、二维环绕（2D-Torus）等网络拓扑的指标随 N 变化。基于提出的度量，不同拓扑结构随结点数目 N 的可扩展性如图 1 所示。

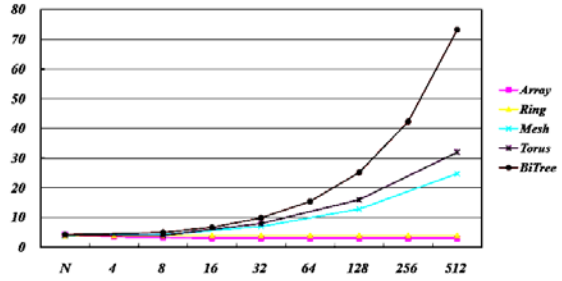

图 1　$P_C(N,Topo)$ 随 N 和拓扑变化曲线图

BinaryTree（BiTree）具有更好的性能和成本可扩展优势，但对称性相对较差，易引起负载的不均衡，且寻径算法相对 Mesh、Torus 的 X-Y 路由较复杂，所以不用 BiTree 作为处理器互连拓扑。

当 N=16 时，Mesh 以较少的链接数却获得与 Torus 在性能和价格方面几乎相同的优势。从性能和成本角度，相对于 Ring、Array，总体上看，Mesh 和 Torus 具有很好的可扩展性，随着结点数目 N 的增长，性能和成本评估指标呈上升趋势，尤其是 Torus 在结点数较多时此趋势更为明显。因此，在众核 CMP 设计时，可重点考虑将 Torus 作为片上互连拓扑结构。对于高效能结点机，单个 1U 机箱内有 16 个 CMP，且考虑到相对 Torus，Mesh 结构的路由较便捷、布线较简易，本文选择以 Mesh 作为结点内处理器互连网拓扑结构，在后续系统扩展时可考虑采用 Torus 结构。

2.2　访存模式

从访存角度，高效能结点机每四片龙芯 3A 以目录协议硬件实现一个 CC-NUMA 结构。计算结点内部在主板级物理上有分布的处理器本地内存，并形成一个全局地址空间的共享存储器。作为一种分布式共享存储（DSM）结构，其优势是：

第一，易编程方面，它提供了单一系统映像，可几乎不加修改地运行单处理器的二进制代码；

第二，硬件资源共享方面，多处理器、内存、I/O 紧耦合在单一系统内，提供了强大的计算能力，从而使得这些资源能有效共享；

第三，访存时延方面，由于程序的时间局部性和空间局部性，CC-NUMA 结构的大部分存取操作在本地完成，因此相对于 SMP 结构，CC-NUMA 单元具有较小的整体访存时延；

第四，作为系统可扩展性的一个重要方面，随着处理器数量的增加，访存或通信延迟如何增加，称为时延可扩展性（Latency Scaling）[11]。基于与第三点相同的原因，当进一步扩展 CC-NUMA 单元的规模时，整体时延增加相对于规模增加较小，系统具有较好的时延可扩展性。

2.3　互连方式

超级传输总线（HT：HyperTransport）是一种为主板上集成电路互连而设计的端到端总线[12]，可用于处理器的互连和处理器的 I/O，龙芯 3A 处理器集成了 HT 接口，在总线宽度为 32bit 时 HT 总线带宽为 6.4 GB/s。因此，HT 可作为主板级 CPU 之间及 CPU 与芯片组的互连总线。

结点机之间的互连网络选择也是影响系统通信性能的一个重要方面。从 2010 年 6 月发布的 TOP500 数据看，Gigabit Ethernet 和 InfiniBand 是主流的互连网络，比例之和已超过 80%[13]。本文选择 Gigabit Ethernet 和 QDR InfiniBand 作为高效能结点机的互连网。为每个 CC-NUMA 单元配置一个 QDR InfiniBand 接口和一个 Gigabit Ethernet 接口。从而在 1U 空间内，16 路处理器以 Mesh 结构通过 HT 总线互连，QDR InfiniBand 和 Gigabit Ethernet 提供结点之间的互连。同时，QDR InfiniBand 和 Gigabit Ethernet 也可作为结点机内部 CC-NUMA 单元之间的互连手段。一个 QDR 4X 的带宽为 32Gb/s[14-15]。这就为结点机设计了一个具有多重路由选择的高带宽、低延迟互连方案。

2.4　机箱结构

构建大规模超级计算机系统，应采用模块化结构来设计其计算结点。就特定性能目标的单个结点而言，在物理上究竟采用高密度设计还是松散设计，是一个基本的设计问题。一方面，对于系统级总线如 HT，其链路由于能耗和信噪比的原因具有最大长度的限制，而且链路的传输时延随长度的增加而增加[12]。采用高密度设计方案有利于提高此类高速总线的信号完整性和传输性能。同时，对于特定的性能目标，采用高密度设计，结点在物理空间上可紧密排放，有利于大规模结点的封装和装配，例如刀片服务器，它在节约空间、便于管理、可扩展性方面具有显著优势。

另一方面，松散设计允许更多地采用市售器件，有较大的设计空间和较多的可选方案，有利于降低工程上的难度。例如塔式服务器，占用空间较大，可选设计空间较大，可以较低的成本进行灵活的设计。

从上述两个方面折中考虑实际方案，本文最终采用了机架式（Rack）服务器方案。机架式服务器是按照统一标准设计的、配合机柜使用的服务器。相比塔式服务器，其占用空间较小，单位空间可放置更多的服务器，同时可降低托管成本（托管费用按机器占用空间大小收取），对千万亿次计算系统这样的计算结点数量来说尤其如此。同时，相比刀片服务器，其拥有较大的设计空间。

3　结点机系统设计

3.1　系统总体结构

综合第 2 节所述的各种考虑，结点机总体设计方案确定为：处理器全部采用国产自主研发的龙芯 3A；结点机采用 1U Rack 箱体；结点机内部具有以 Mesh 结构互连的 16 路处理器（如图 2 所示），每四路处理器构成一个 CC-NUMA 单元（图 2 中的虚线边框），单元之间以 HT 总线相连；结点间采用商业化的 QDR InfiniBand 和 Gigabit Ethernet 互连，结点内的 CC-NUMA 单元之间也可通过这两种方式互连。以上设计使单结点具有每秒 0.256 万亿次浮点运算能力（TFLOPS），单一机柜可容纳 42 个 1U 结点机箱、672 颗 CPU、2688 个 CPU 核，完成千万亿次计算需用 100 个左右的机柜。

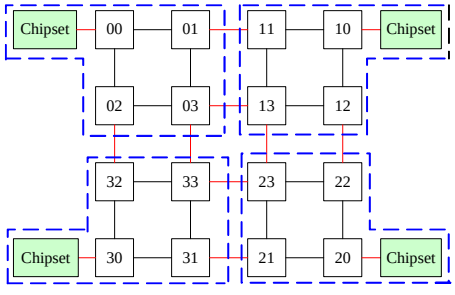

图 2　结点机系统总体结构

3.2　CC-NUMA 单元

如图 3 所示，每个结点机包括四个 CC-NUMA 单元，每个单元在主板上集成了 4 片龙芯 3A 四核处理器，8 个 DDR2 内存插槽，1 个 RTL8110SC 千兆以太网芯片和 QDR，1 个 BIOS Flash，1 个串口芯片，时钟系统电路和电源变换电路等。

对每一个 CC-NUMA 单元，下面依次介绍其处理器互联逻辑、时钟、上电与复位电路、内存、主板和箱体结构的设计。

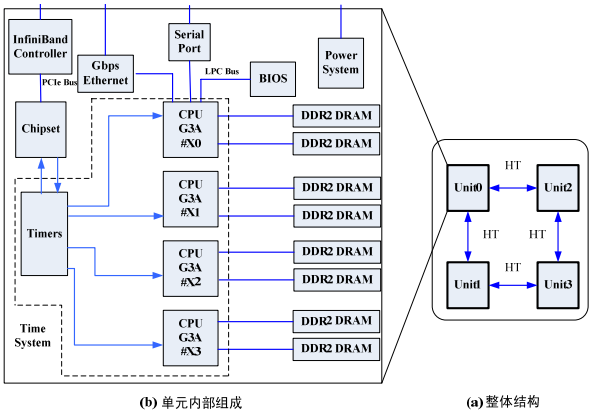

图 3　结点机 CC-NUMA 单元

3.2.1　处理器互连逻辑

龙芯 3A 处理器内部集成了两级 AXI（Advanced eXtensible Interface）交叉开关（图 4 中所示为一级交叉开关），其中一级交叉开关用于连接四个处理器核心、四个二级 Cache 模块、HT0 和 HT1 两个 HT 通道。每个 HT 通道都是 16bit，但均可拆分为两个 8bit 的通道，这样每个处理器就可有 4 个 8bit 的 HT 通道。4 片龙芯 3A 处理器进行互连时，HT0 分成两个 8bit 的链路，分别与上行的处理器和下行的处理器进行互连，而 HT1 可用来与南桥芯片连接。

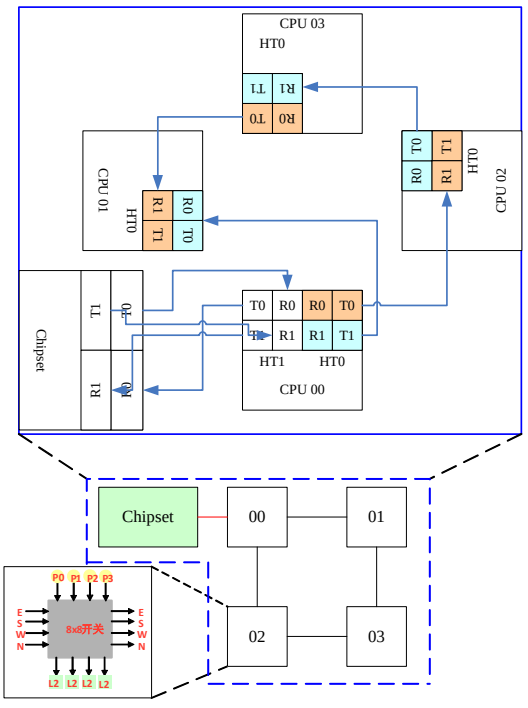

图 4　CC-NUMA 内部龙芯 3A 处理器互连逻辑

为便于描述，将处理器编号为 CC-NUMA 单元号+CC-NUMA 内位置号，由 0 号处理器提供该 CC-NUMA 单元与 BIOS 和南桥芯片的连接。

每个处理器包含 HT0 接收（R0）、HT0 发送（T0）、HT1 接收（R1）和 HT1 发送（T1）共 4 个接口。图 2 中除 03 号、13 号、23 号和 33 号 4 个处理器的 4 个接口全部用完外，其他处理器均只用了两个 HT0 接口和一个 HT1 接口。

图 2 中黑色线表示龙芯 3A 处理器的 CC-NUMA 内部互连，用 HT0 接口实现，首尾相联形成闭环；红色线表示 4 个 CC-NUMA 之间的互连和 CC-NUMA 与南桥芯片的互连，用各 CC-NUMA 边界处龙芯 3A 处理器的 HT1 接口实现，03 号、13 号、23 号和 33 号处理器形成内部闭环，其余处理器形成外围闭环。

3.2.2　时钟子系统

时钟子系统电路为系统提供实时时钟和保存存储器配置信息。

CC-NUMA 中的每个 CPU 都需要提供相同的内存时钟输入（33MHz）、系统时钟输入（33MHz）、HT 参考时钟（100MHz）和 HT 时钟（200MHz）。如图 5 所示，采取的设计方案是：

1）将 33MHz 晶振通过缓存芯片分为 4 路驱动 CPU 的内存时钟输入；

2）将 33MHz 晶振通过缓存芯片分为 4 路驱动 CPU 的系统时钟输入；

3）将 100MHz 晶振通过缓存芯片分为 4 路驱动 CPU 的 HT 参考时钟输入；

4）由于每个 CC-NUMA 含 1 个 MCP68 Chipset 芯片，可将南桥芯片的 200MHz 时钟输出通过缓存芯片分为 4 路驱动 CPU 的 200MHz 时钟输入。

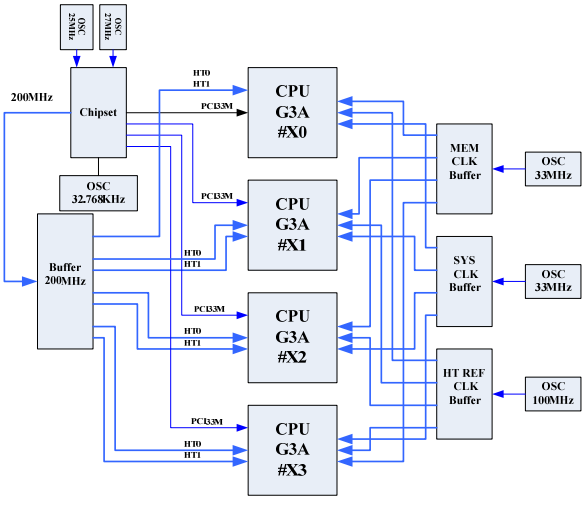

图 5　CC-NUMA 的时钟子系统设计

3.2.3　上电与复位电路

结点机采用的 MCP68 芯片组提供了控制输出和状态输入管脚，这些管脚与处理器等芯片及时序控制电路相连完成 CC-NUMA 单元的上电及复位时序控制。MCP68 上电及复位的主要相关信号包括检测开机或复位输入（PWR_BTN#、RST_BTN#）、PCI 复位信号（PCI_RST#）、PCIE 复位信号（PCIE_RST#）、HT 总线控制信号、状态控制信号（SLP_S5#、SLP_S3#）和电平面状态及控制信号。当 CC-NUMA 单元复位时，四片龙芯 3A 处理器将收到 SYS_RST#信号，同时 MCP68 芯片组接收 RST_BTN#信号。龙芯 3A 处理器构成 CC-NUMA 单元时，各处理器需按固定的次序和延迟完成复位，因此需要对各 SYS_RST#信号分别进行延时处理。

时序控制电路分为两部分：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地反应给 MCP68 芯片组，并保证 MCP68 芯片组对状态输入信号的有效反应时间，这可通过电压监控电路和 RC 电路来实现。DC/DC 控制电路将 MCP68 芯片组输出的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 MCP68 芯片组，由此，逐一完成各种电压的生成和确认（如图 6 所示）。

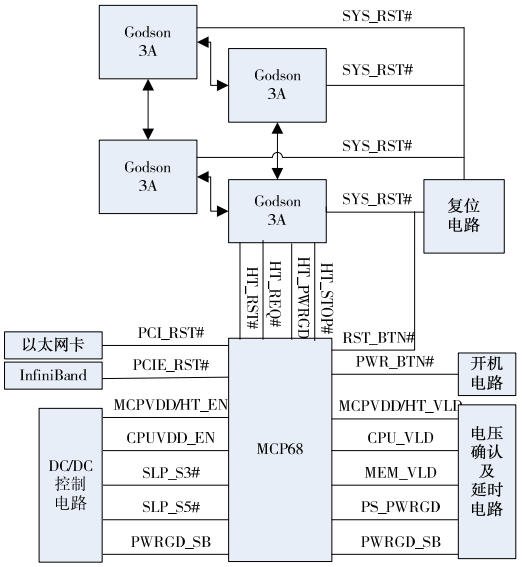

图 6　CC-NUMA 单元上电及复位子系统设计

3.2.4　内存子系统

如图 7 所示，每片龙芯 3A CPU 的 2 个内存控制器各连接一个 DDR2 240PIN DIMM 插槽。

DDR2 内存总线带宽可达到 800MB/s。

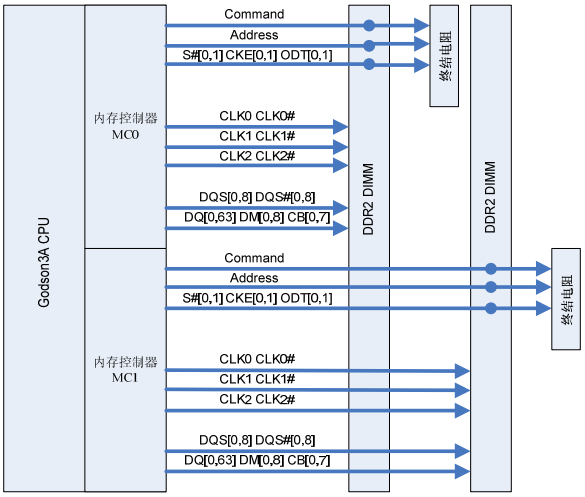

图 7　龙芯 3A 内存控制器与 DDR2 DIMM 连接框图

3.2.5　主板和箱体结构

结点机采用 1U 机箱，内部结构如图 8 所示，包括一个 AC/DC 电源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公头-母头式高速板间连接器对接。

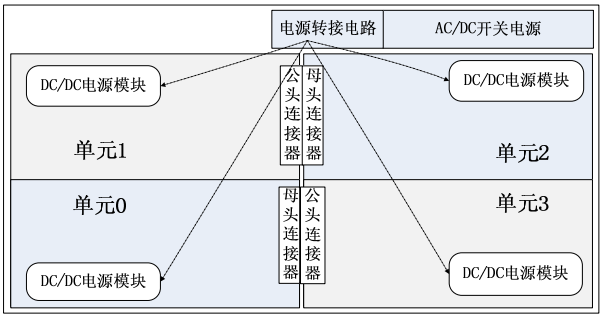

图 8　结点机机箱内部结构

4　结点机系统评测

基于上述结点机设计方案，本文实现了基于龙芯 3A 处理器的高效能计算结点机，图 9 为结点机实物图。

图 9　结点机实物图

作为龙芯系列首款多核处理器，龙芯 3A 四核处理器在性能等方面还有很大的提升空间，即将推出的龙芯 3B 八核处理器性能有数倍的提升，且与龙芯 3A 完全兼容，因此，本文基于龙芯 3A 处理器研制的结点机将来可直接支持龙芯 3B 八核处理器。

4.1　功能测试

基于自主研发的 PMON 和移植的 Linux 内核，调试后的结点机加电自检在 POST（Power On Self Test）阶段检测出了 CPU 时钟、内存 DDR2 倍频后的时钟和 CPU 的内存控制器地址空间等硬件配置信息，完成了 HT 总线的握手。

在加载内核后，系统检测出了 QDR InfiniBand 和 Gigabit Ethernet 等互连设备。例如 QDR Infiniband 控制器启动信息如下：

```
mlx4_core: Mellanox ConnectX core driver v0.01 (May 1, 2007)
mlx4_core: Initializing 0000:02:00.0
PCI: Enabling device 0000:02:00.0 (0000 -> 0002)
mlx4_ib: Mellanox ConnectX InfiniBand driver v1.0 (April 4, 2008)
```

4.2　内存访问带宽测试

作为非一致存储访问结构，处理器访问本地内存和远程内存的带宽不同，表 1 的 stream 测试结果体现了这一系统结构特征。

表 1　CC-NUMA 内存访问带宽（单位：MB/s）

测试项目	CPU0 访问本地内存	CPU0 访问 CPU1 的本地内存	CPU0 访问 CPU2 的本地内存	CPU0 访问 CPU3 的本地内存
copy	290	111	128	83
scale	286	107	123	83
add	316	102	123	89
triad	317	102	122	89

4.3　点到点通信性能测试

使用 NetPIPE 测试软件，测试不同计算单元间的点到点通信延迟与带宽，如表 2 所示。

表 2　点到点通信延迟和带宽

测试项目	TCP 最小延迟（us）	TCP 最大带宽（Mbps）
同一虚拟子网内两个计算单元	37.49	502.00
同一 IB 子网内两个计算单元	71.40	781.31
相邻虚拟子网内两个计算单元	95.23	228.03

4.4　MPI 全局性能测试

使用 IMB 2.2 测试软件[16]，对结点机不同规模的计算结点通过 16384 字节的 SendRecv、Allreduce、Reduce、Allgather、Allgatherv、Alltoall、Bcast 测试，得到相应规模的带宽或延迟；通过

Barrier 测试，得到同步延迟，MPI[17]全局性能测试测试结果如表 3 所示。

表 3　MPI 全局通信带宽和延迟

处理器芯片数	2	4	8	16
SendRecv（MB/s）	35.54	16.39	13.18	8.56
Reduce（us）	1494.70	3506.47	5117.93	7045.03
Allreduce（us）	1695.87	4145.02	6758.90	10651.46
Allgather（us）	889.95	3852.05	22095.85	49207.19
Allgatherv（us）	926.36	3929.65	29082.62	48134.99
Alltoall（us）	915.45	3602.50	15085.09	45889.98
Bcast（us）	680.53	1548.56	4680.32	7172.58
Barrier（us）	243.83	630.96	983.71	1523.70

Barrier 是 MPI 程序中常用的用来进行线程同步的函数。如图 10 所示，从 8 到 64 核，耗费的时间增加约 8 倍。处理器核心数目扩展趋势线与线程同步耗时增长趋势线大体平行，整体上看线程同步具有较好的可扩展性。

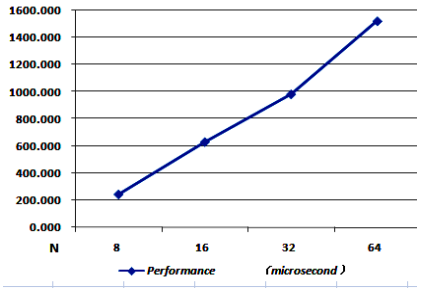

图 10　系统线程同步耗时的可扩展性

4.5　系统 Linpack 性能测试

16 个 CPU（64 核）规模的系统，矩阵阶数从 10000 增加到 48000 时，Linpack 浮点运算性能如图 11 所示。

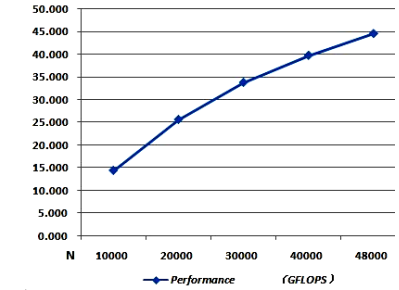

图 11　系统浮点性能随矩阵阶数的扩展

系统 Linpack 峰值情况下的可扩展性如图 12 所示。由图 11 和图 12，尤其图 12 可以看出，性能增长趋势线与处理器规模增长呈较好的正相关关系，说明系统具有较好的性能可扩展性。

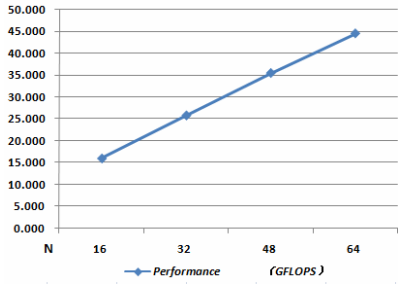

图 12　系统浮点性能随处理器核数的扩展

计算结点的上述性能指标在软件上还有一定的优化空间。

5　相关工作

当前市场上主流的计算结点以两路居多，四路及四路以上的计算结点较少。基于龙芯 3A 处理器的结点机，是一款 16 路 1U Rack 计算结点，占地 0.46 平方米，峰值性能 256 GFlops，峰值功耗不超过 300W/U，计算/功耗比约 0.853 GFlops/W。如表 4 所示，与 IBM 等一些主流高性能计算结点相比，其具有高密度、低占地的显著特点。在处理器 3A 升级为 3B 和软件进一步优化后，龙芯结点机的高性能、低功耗特点在实际运行中将体现得更为明显。

表 4　相关结点机的比较

计算结点	密度、性能、功耗特点
基于龙芯 3A 的结点机	支持 16 颗龙芯 3A/1.0GHz 处理器，占地 1U，功耗约 300W。
IBM Blade® JS20 刀片服务器	支持 2 颗 IBM PowerPC970/2.2GHz 处理器。峰值性能约 17.6 GFlops，占地约 1/2 U，功耗约 395W。
IBM Blade Center® JS22 Express 刀片服务器	支持 4 颗 64 位 4.00GHz POWER6 处理器，占地约 1/2 U，功耗约 350W。
TYAN GT62B8230-LE 机架服务器	支持两颗 AMD 2.1Ghz 12-Core Opteron 6100 处理器，占地 1U，功耗约 350W，峰值性能 201.6 GFlops。

6　结 束 语

本文基于国产龙芯 3A 四核处理器设计和研制的 16-way 高效能结点机具有高性能、低功耗、高密度、可扩展的特点，初步验证了基于国产处理器构建超级计算机系统的可行性，为采用国产处理器研制千万亿次及更大规模的超级计算机系统奠定了坚实基础。

参 考 文 献

[1] Transaction Processing Performance Council[EB/OL]. TPC-C. http：//www. tpc. org/tpcc.

[2] Wm. A. Wulf, Sally A. McKee. Hitting the Memory Wall：Implications of the Obvious[J]. ACM SIGARCH Computer Architecture News, 1995-09, 23(1)：20-24.

[3] Sally A. McKee. Reflections on the memory wall[C]. Proceedings of the 1st Conference on Computing frontiers Ischia, Italy, 2004-04：14-16.

[4] P. Kogge, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, etc. Exascale Computing Study：Technology Challenges in Achieving Exascale Systems[R]. Technical Report, DARPA IPTO, 2008.

[5] HPCS. HPC Challenge [EB/OL]. [2010-07-12]. http：//icl.cs.utk.edu/hpcc/index.html.

[6] DARPA. High Productivity Computing Systems(HPCS) Program[EB/OL]. [2010-07-02]. http：//www. highproductivity.org.

[7] Gao X, Chen YJ, Wang HD et al. System Architecture of Godson-3 multi-core Processors[C]. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY, 2010-03, 25(2)：181-191.

[8] Wang H, Gao X, Chen Y, Hu W. Interconnection of Godson-3 Multi-core Processor[C]. Journal of Computer Research and Development, 2008-12, 45(12)：2001-2010.

[9] Institute of Computing Technology Chinese Academy of Science[S]. Godson3A Processor User Manual, 2010.

[10] Wang Ding xing, Chen Guo liang. Interconnection Network Analysis. Science Press, 1990：36-41.

[11] Culler D, Singh J, Gupta A. Parallel Computer Architecture[M]. San Francisco：Morgan Kaufmann, 1996.

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

[13] TOP500 Supercomputing Sites[DB/OL]. http：//www. top500. org/list/2010/06/100.

[14] G. Pfister. An Introduction to the InfiniBand Arechitecture[C]. IEEE Press, 2001.

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

[16] Performance Evaluation of Supercomputers using HPCC and IMB Benchmarks[C]. Journal of Computer and System Sciences, 2007.

[17] Intel MPI Benchmarks：Users Guide and Methodology Description[S]. Intel GmbH, Germany, 2007.

Design of High Productivity Computing Node based on Godson 3A CPU

LIU Yu-hang[1,3]　ZHU Ming-fa[1,2]　XIAO Li-min[1]　GAO Yu-hui[1,3]

(1. School of Computer Science and Engineering, Beihang University, Beijing, 100191；
2. Lenovo Group, Beijing, 100085；
3. State Key Laboratory of Software Development Environment, Beihang University, Beijing, 100191)

Abstract　As a core part of a supercomputer, computing node is of cornerstone significance for building supercomputing systems. This paper has designed and developed a 16-way high-productivity computing node using Godson3A which is a quad-core domestic CPU. First, a few key issues in design and development of the computing node are researched. Second, on this basis, whole system architecture of the computing node has been presented. Third, the design of the computing node involves processor interconnecting logic, subsystems that clock, power and reset, memory, motherboards and chassis structure, etc. The design and implementation scheme of all these are presented in detail respectively. At last, for the computing node, functional and performance evaluations are conduced. Compared with mainstream computing nodes in current market, the computing node developed in this paper has more advantages on performance, power efficiency (the ratio of performance per watt drawn), integration density etc.

Key words　computing node, Godson3A, Chip Multiprocessor, high productivity, supercomputer

Class number　TP391　　(责任编辑：刘　洋)