doi:10.3969/j.issn.1000-436x.2013.12.015

基于龙芯多核处理器的云计算节点机

阮利[1,2]，秦广军[1,2]，肖利民[1,2]，祝明发[1,2]
（1. 北京航空航天大学 软件开发环境国家重点实验室，北京 100191；2. 北京航空航天大学 计算机学院，北京 100191）

摘 要：提出了一种基于龙芯多核处理器的高效能云计算节点机的软硬件设计和实现方法，并研制成功相应原型系统。实验和测试表明，本系统单节点取得了每秒 0.256×10^{12} 次浮点运算能力（Tflops），单一机柜可容纳 42 个 1U 节点机箱，672 颗 CPU，2 688 个 CPU 核（672×4）的性能，总体具有基于龙芯多核处理器、高密度、高性能功耗比等优点，为基于龙芯多核处理器的云计算系统奠定了坚实基础。

关键词：龙芯处理器；计算节点； 多核处理器；云计算
中图分类号：TP301 **文献标识码：**B **文章编号：**1000-436X(2013)12-0131-11

Cloud computing node based on Loongson multi-core CPU

RUAN Li[1,2], QIN Guang-jun[1,2], XIAO Li-min[1,2], ZHU Ming-fa[1,2]
(1. State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China;
2. School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

Abstract: A cloud computing node based on Loongson multi-core CPUs was introduced. Both the design and implementation methods of its hardware and software were introduced. Moreover, the prototype was successfully implemented. Experimental results show that one computing node of this system can achieve the performance of 0.256×10^{12}TFlops per second, 42 1U ranks and 672 CPUs with 2688 CPU cores (672×4) in one cabinet. i.e. it is competing in metrics like Loongson multi-core CPU based, high performance with low power, high integration, etc. and lays a solid foundation for Loongson multi-core CPUs based cloud computing systems.

Key words: Loongson CPU; computing node; multi-core CPU; cloud computing

1 引言

高性能计算机是当前云计算数据中心[1]、大数据计算的核心装备。高性能计算机的研制水平、生产能力及应用程度是国家综合国力、信息化建设、数据和计算安全保障能力的重要体现，是世界各国特别是发达国家争夺的战略制高点[2]。美国国防部的“高效能计算系统”(HPCS, high productivity computing system)[3]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。自此，美国的 HPCS 以及我国的“863”计划“高效能计算机系统研制与关键技术研究”重大专项等项目均已相继开展了针对 HPCS 的研究。

目前，科学计算、云计算和大数据计算等新应用对计算性能无止境的需求推动了高性能计算机系统峰值性能的迅速提高。然而，在高效能计算机峰值性能和系统规模不断扩大的同时，如何有效解决国外处理器带来的可靠性和安全性，由于应用多样性、系统规模显著增大、复杂的体系结构等带来的性能功耗比、性能体积比和可扩展性等高效能瓶

收稿日期：2012-10-31；修回日期：2013-07-15
基金项目：国家高技术研究发展计划（“863”计划）基金资助项目（2011AA01A205）；国家自然科学基金资助项目（61003015, 61370059, 61232009）；软件开发环境国家重点实验室探索性自主研究课题基金资助项目（SKLSDE-2012ZX-23）；北京市自然科学基金资助项目（4122042）
Foundation Items: The Hi-tech Research and Development Program of China (863 Program) (2011AA01A205); The National Natural Science Foundation of China (61003015, 61370059, 61232009); The Fund of the State Key Laboratory of Software Development Environment (SKLSDE-2012ZX-23); Beijing Natural Science Foundation (4122042)

颈问题，是高效能计算机研究领域近期和未来亟待解决的热点和难点问题之一。首先，高性能计算机对高密度紧耦合计算资源的需求变得极为迫切。例如，在 2013 年 11 月公布的 Top 500 排名中，排名第一的 Tianhe-2 的 Linpack 能达到 33 862.7 Tflop/s，拥有 3 120 000 个处理核，1 024 000 GB 内存[4]。此外，因为高性能计算机在系统功耗方面也面临严峻的挑战。例如，据英特尔统计，数据中心 25%的成本是耗电。Top 500 排名第二的 Titan 功耗为 8 209 kW，而 Tianhe-2 更高达 17 808 kW。另一个是高性能计算机体积问题。从数据中心的占地面积来看，Google 在美国俄勒冈州哥伦比亚河畔建设的数据中心占地面积达到 10×10^4 平方公尺，总共拥有 8 180 个机柜[5,6]。

处理器作为高效能云计算节点的核心处理单元，是达成安全性、高性能、低功耗等高效能指标的关键。龙芯 3A 处理器[7,8]是中国科学院计算技术研究所自主研制的一款面向高性能计算的片上多处理器，采用 65 nm 生产工艺，在 1 GHz 主频下可实现 16 Gflop 的运算能力，功耗约 10 W，性能功耗比为 1.6。研究表明，相对于主流的 Intel 和 AMD 处理器，新兴的以龙芯 3A 和 3B 为代表的龙芯 3 系列处理器在性能功耗比等高效能指标和自主创新等上有较明显的综合优势。此外，从自主创新和安全可控的角度来看，使用国产处理器作为高效能云计算节点的核心部件可扭转高效能计算机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、维护国家安全等都具有重要意义。另一方面，计算节点是高性能计算机体系结构中除加速节点、主机单元、通信网络等外的核心部件之一，计算机节点的高效能已成为整机“高效能”目标能否达成的关键。目前在研和投入使用的一批大规模高性能计算机均十分注重计算节点高效能问题的突破，代表性的系统主要包括：Gary Titan[4]，日本富士通 K 计算机[9]、Tianhe-1A、Jaguar[10]、Cray 的 Cascade 和 Baker，IBM 的 Roadrunner 等[11~18]。可见，研究一种具有高性能功耗比、高密度和较强可扩展性的、基于国产龙芯多核处理器的高效能云计算节点具有非常重要而深远的研究意义和工业实践价值。

本文关注于云计算节点的高效能软硬件设计和实现问题，提出了一种基于龙芯 3A 多核处理器的云计算节点机的软硬件设计和实现方法。

2 逻辑结构

基于龙芯 3A 多核处理器的高效能云计算节点的基本设计思路为：作为高性能计算机体系结构中除加速节点、主机单元、通信网络等外的核心部件之一，高效能云计算节点主要承载大规模或超大规模的计算密集型应用任务，是超级计算能力的主要实现者，主要应对高效能目标中的高性能、低功耗、高密度、低成本和高安全等挑战。在拓扑结构上，计算节点总体由 4 个 SMP 构成。从处理器角度来看，节点在 1U 高度的 Rack 机箱中集成 16 颗龙芯 3 四核处理器，构成一个 cc-NUMA 结构的系统，系统实现高性能功耗比、高性能体积比和高集成度等目标。

处理器是高效能云计算节点的核心运算单元，处理器的选择和互连方法是本文计算节点设计中的关键。在处理器的选择上，本文选取的龙芯 3A 处理器总体具有如下特征：每个龙芯 3A 片内集成 4 个 64 bit 的四发射超标量主频 1 GHz 的 GS464 高性能处理器核。处理器预留的互连端口决定了高效能云计算节点互连技术的设计。对于龙芯 3A 多核处理器来说，每个处理器有 2 个 16 bit HT 端口，而每个端口可以拆分成 2 个 8 bit 端口使用(HT0 和 HT1)，因此共有 4 个可用的 8 bit HT 端口(HT00 和 H01 分别表示 HT0 的低 8 bit 和高 8 bit，HT10 和 HT11 分别表示 HT1 的低 8 bit 和高 8 bit)。基于龙芯 3A 的 HT 互连端口的配置，计算节点通过每 4 颗处理器互连即可自动构成一个 cc-NUMA 结构的 SMP。

为便于对本文计算节点设计和实现技术做进一步详细介绍，首先给出如下定义。

定义 1 基于龙芯 3A 多核处理器的高效能云计算节点 $\Gamma(n,r,m)$ 可定义为一个二元组 $\Gamma(n,r,m)=<S_{n,r},C_m>$。其中，$S_{n,r}$ 是 SMP 集合，n 表示计算节点中所含的 SMP 的个数，即 $S_{n,r}$ 集合中的元素个数；r 表示每个基本 SMP 单元包含的处理器个数，即 $S_{n,r}$ 集合中每个元素包含的处理器个数。C_m 是处理器间直连端口集合；m 表示每个处理器可用的直连端口数目。

具体地，$\Gamma(n,r,m)$ 表示该计算节点是由 n 个包含 r 个处理器，且每个处理器的可用直连端口数为 m 的 SMP 组成的计算节点。例如，$\Gamma(4,4,4)=<S_{4,4},C_4>$，就表示计算节点 $\Gamma(4,4,4)$ 在体系结构上由 4 个 SMP

组成，每个 SMP 包含 4 个处理器，每个处理器有 4 个互连端口。也即表示了一个 16 路 4SMP 的计算节点。对 $S_{n,r}$ 和 C_m 更详细的定义如定义 2~定义 4 所示。

定义 2　计算节点的 SMP 集合定义为 $S_{n,r}=\{s_i\}$, $n-1\geq i\geq 0$。其中，s_i 表示第 i 个 SMP 单元。

定义 3　给定 $S_{n,r}$，第 i 个 SMP 单元定义为 $s_i=\{p_{i,j}\}$, $r-1\geq j\geq 0$。其中，$p_{i,j}$ 表示 s_i 中的第 j 个处理器。r 表示每个基本 SMP 单元 s_i 包含 r 个处理器。

定义 4　给定 $S_{n,r}$ 的处理器间直连端口集定义为 $C_m=\{c_{i,j,k}\}$，$m-1\geq k\geq 0$，其中，$c_{i,j,k}$ 表示 $S_{n,r}$ 中的 SMP 单元 s_i 的第 j 个处理器 $p_{i,j}$ 的第 k 个直连端口。n 和 r 如定义 2 和定义 3 所示，m 表示每个处理器可用的直连端口数目，端口从正南方向开始顺时针编号。在定义 4 中，处理器互连的具体拓扑结构与 C_m 的值密切相关，m 值不同，互连规则不同，得到的拓扑结构也将不同。

基于定义 1~定义 4，本文研制的基于龙芯 3A 多核处理器的云计算节点表示为 $\Gamma(4,4,4)$，其逻辑结构布局如图 1 所示。具体地，该计算节点由 4 个 SMP 组成，每个 SMP 包含 4 个处理器，每个处理器有 4 个互连端口，也可看作一个由 16 个处理器组成的二维四元的 mesh 结构。图 1 中，$p_{i,j}$ 表示龙芯 3A 处理器。“chipset”表示芯片组，用于连接 USB、IDE、Ethernet 等外部接口；“InfiniBand”表示 InfiniBand 互连设备，用于连接 InfiniBand 网络。其中，SMP 内部通过 HT0 互连，SMP 间通过 HT1 互连。$\Gamma(4,4,4)$ 总体相当于一个基于 HT 总线的 4SMP 板上机群。

3　互连规则

本文所研制的高效能云计算节点单节点上具有 16 个处理器，如何将这 16 个处理器进行互连并形成一个统一的计算平台，这是系统设计首要解决的关键技术问题。2.1 节重点介绍了计算节点的核心组成部件和逻辑结构。本节将重点介绍本文所研制的 $\Gamma(4,4,4)$ (16 路 4 SMP)的互连方法，并讨论其性质。

HyperTransport(HT)是一种为主板上的集成电路互连而设计的端到端总线，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 32 bit 时 HT 总线带宽为 6.4 Gbit/s。因此，HT 可作为主板级 CPU 之间及 CPU 与芯片组的互连总线。

由龙芯 3A 组成的 2D mesh 结构的计算节点 $\Gamma(n,4,4)$，其处理器间互连总体包括 SMP 组内与 SMP 组间 2 种。下面将详细介绍互连规则。

定义 5　对计算节点 $\Gamma(n,4,4)$，其中，$n=xy$, $x\geq 1, y\geq 1$，如 x 表示横向(或 x 轴方向)的 SMP 个数，y 表示纵向（或 y 轴方向）的 SMP 个数。其 2D mesh 构图规则如下所示。

规则 1　SMP 内部处理器互连。在 s_i 内，$p_{i,j}$ 与 $p_{i,k}$ 若满足 $(j+1)\bmod 4=k$，则 $p_{i,j}$ 与 $p_{i,k}$ 直接互连，且互连端口分别为 c_{i,j,h_1} 和 c_{i,k,h_2}, $|h_1-h_2|\equiv 2$。

规则 2　SMP 间处理器互连。在 s_k 与 s_j 间，$p_{k,i}$ 与 $p_{j,i}$，满足 $j+1=k$ (在 x 轴方向)，或 $j+x=k$ (在 y 轴方向)，则 $p_{j,i}$ 与 $p_{k,i}$ 直接互连，且互连端口分别为 c_{j,i,h_1} 和 c_{k,i,h_2}, $|h_1-h_2|\equiv 2$。

规则 1 定义了 SMP 内部的处理器间互连方式。

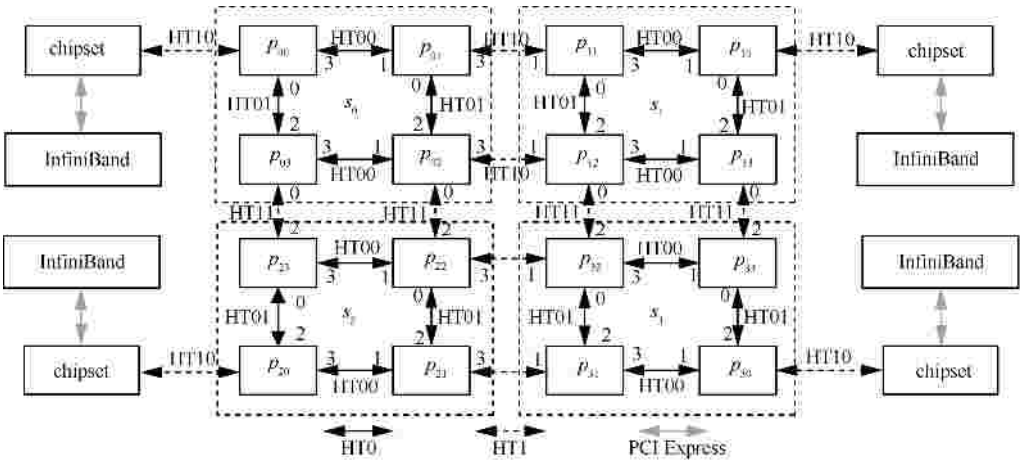


图 1　基于龙芯 3A 多核处理器的高效能计算节点 $\Gamma(4,4,4)$ 2D mesh 结构

规则 2 定义了 SMP 间的互连方式，s_k 的编码是从左到右，从上到下，按规则 2 递增，s_k 和 s_j 的互连端口号差 2 个。

按照规则 1 和规则 2，图 2(c)表示了本文高效能节点的互连方式。图 2(a)和图 2(b)示例了由龙芯 3A 组成的 2D mesh 结构的高效能云计算节点 $\Gamma(n,4,4)$ 的互连扩展过程。

性质 1 按定义 5，$\Gamma(n,4,4)$ 是以 $\Gamma(1,4,4)$ 为基元的 mesh 结构。

证明 定义 5 的规则 2 横向节点编号按 1 递增，纵向节点编号按 x 递增，构成典型的 2D mesh 结构。$\Gamma(n,4,4)$ 的每个节点都是 4 路 SMP，因此，$\Gamma(n,4,4)$ 是以 $\Gamma(1,4,4)$ 为基元的 mesh 结构。

证毕。

性质 2 定义 5 中，$\Gamma(n,4,4)$ 的处理器个数为 $4n$。

证明 按照定义 1，$\Gamma(n,r,m)$ 中的 n 为 SMP 个数，按照定义 5，$\Gamma(n,4,4)$ 中每个 SMP 都是 4 路，所以 $\Gamma(n,4,4)$ 的总处理器个数为 $4n$。证毕。

性质 3 定义 5 中，$\Gamma(n,4,4)$ 的处理器规模按 4 的倍数递增。

证明 按定义 5 和性质 1，$\Gamma(n,4,4)$ 无论在横向还是纵向，或者双向扩展，都将至少增加一个 $\Gamma(1,4,4)$，而 $\Gamma(1,4,4)$ 是一个 4 路的 SMP，因此，$\Gamma(n,4,4)$ 的规模按 4 的倍数递增。

证毕。

定理 1 $\Gamma(n,4,4)$ 中任意处理器间最长通信距离为 $x+y-2$。

证明 按性质 1，$\Gamma(n,4,4)$ 是 mesh 结构，mesh 结构的任意两点间最长距离等于 $x+y-2$。又因为 $\Gamma(n,4,4)$ 的每个节点都是 4 路 SMP，SMP 内的 4 路处理器通过内存共享进行通信，通信距离通常定义为 0。因此，$\Gamma(n,4,4)$ 中任意处理器间最长距离为 $x+y-2$。

证毕。

综上可见，本文所研制的高效能云计算节点总体具有如下性质：龙芯 3A 处理器间通过 HT 总线互连，在定义 5 的规则 2 下，$\Gamma(n,4,4)$ 的 SMP 间互连形成 2D mesh 结构，因此理论上具有 mesh 结构的所有特性，适用 mesh 结构的路由策略。在实际工程中，需根据具体需求和工艺制造特性确定节点的处理器规模。

4 通信机制

基于第 2 节的逻辑结构和第 3 节的互连结构，本文研制的高效能云计算节点总体包含 3 套网络，分别是板上 HT 总线网络、节点间的 InfiniBand 网络和吉比特以太网。本节将根据龙芯处理器 HT 总线特性重点介绍基于 HT 总线的通信机制。

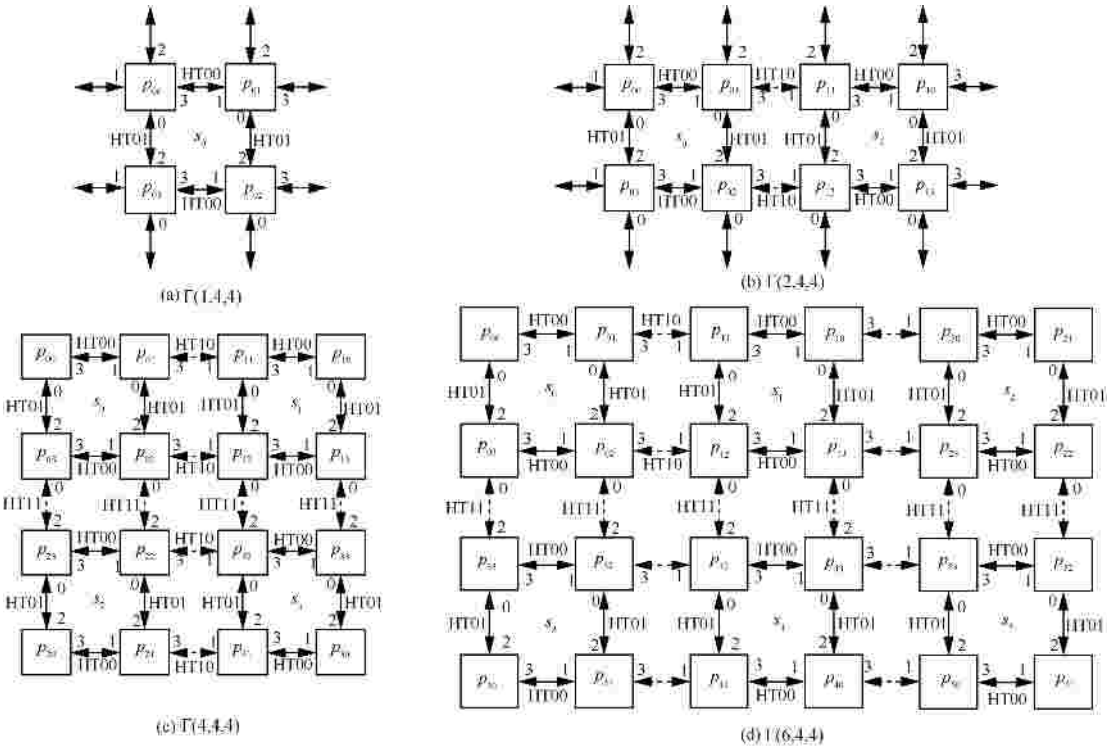


图 2 高效能云计算节点 $\Gamma(n,4,4)$ 互连扩展过程

如图 1 和定义 5 所示，高效能云计算节点的四组 s_i 都是 cc-NUMA 结构的 SMP，因此，路由分 SMP 内和 SMP 间两层。如图 2(a)所示，在 SMP 内，各处理器通过 HT 总线互连，4 个处理器具有共同的内存空间，因此，SMP 内的处理器间通过共享内存通信，故通常可认为 SMP 内的处理器间通信距离为 0。SMP 间可通过 HT 总线传递数据，通信规则基于 2D mesh 结构进行。

龙芯处理器的 4 个 SMP 间通过地址映射实现隐式通信，具体的通信过程为：处理器识别访存地址，如果地址在本地 SMP 的内存空间，则通过共享内存通信；如果地址在地址映射的范围内，则由处理器负责路由到相应的其他 SMP 内存空间去。具体地，通过地址映射将图 1 配成了内外 2 个闭环通路，外环按：

$$p_{0,1} \rightarrow p_{1,1} \rightarrow p_{1,3} \rightarrow p_{3,3} \rightarrow$$
$$p_{3,1} \rightarrow p_{2,1} \rightarrow p_{2,3} \rightarrow p_{0,3}$$

顺序形成顺时针环路，内环按：

$$p_{0,2} \rightarrow p_{2,2} \rightarrow p_{3,2} \rightarrow p_{1,2}$$

顺序形成逆时针环路，如图 3 所示。

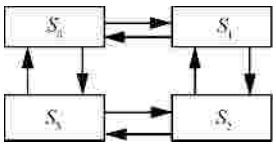


图 3　通信环路

在上述通信系统下，假定处理器 $p_{j,i}$ 要访问 $p_{k,h}$，并用 o_1 表示外环，o_2 表示内环，则路由算法伪码如图 4 所示。

```
算法 1　路由算法
Algorithm 1. Rooting Algorithm
//判断源节点与目标节点是否在同一个 SMP 内
If (InSameSMP( p_{j,i} , p_{k,h} ))
//通过共享内存通信
{ CommBySharedMemory( p_{j,i} , p_{k,h} );
Else //判断源节点与目标节点是否在同一个 SMP 内
{If (IsBusy( o_2 ))//优先选择通过内环通信
{ If (IsBusy( o_1 ))
//最后选择通过 InfiniBand 通信
{CommByIB( p_{j,i} , p_{k,h} );
Else
//其次选择通过外环通信
{CommByO2( p_{j,i} , p_{k,h} );
Else
//通过内环通信
{CommByO2( p_{j,i} , p_{k,h} )
```

图 4　路由算法伪代码

图 4 算法的基本原理为：如果源处理器与目标处理器在同一个 SMP 内，则通过共享内存通信；否则优先通过内环通信，因为内环相比外环和 IB 具有更短的通信路径。当内环路忙时，则选择通过外环通信；否则，绕道 InfiniBand 网络进行通信。

5　基于龙芯 3A 多核处理器的云计算节点系统实现

5.1　主板

在主板设计上，高效能云计算节点系统具有 16 个处理器，组成 4 个 SMP，考虑到电路板制造和可调性等工艺问题，不能将系统设计成单块 16 路的大主板。因此，计算节点被设计成对称的两块主板，每板包含 8 路 2SMP，两板通过对插组成一个 1 U 的计算节点。 8 路主板逻辑图如图 5 所示。

在部件组成上，高效能云计算节点系统总体基于 16 个龙芯 3A 处理器，每个处理器支持 4 条内存条。16 个处理器放在 1 U 的标准高度和宽度的计算节点内。系统还支持 4 个 QDR Infiniband 高速网，4 个吉比特以太网，另外支持 VGA、串口和 USB 等。

在布局上，单板上有 8 个龙芯 3A 处理器，32 条内存条，2 个 IO 芯片，2 个吉比特以太网接口以及 VGA、USB 和串口。每颗龙芯处理器支持 4 个内存条。IO 芯片对外扩展 Infiniband、以太网、VGA、USB 和串口。

单主板的电子元器件多，但是主要的部件是处理器和内存，在布局上需重点考虑把这两类部件的布局和电子线路规划好。考虑到处理器之间的互连结构和各自连接内存条，因此，在布局上采用了规则的方阵式布局，即处理器放在中间，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。

在层数设计上，电路板采用 12~20 层的设计，以解决高速、低速数字信号和电源等模拟信号之间的干扰，从 PCB 印制板上首先保证信号的完整性。

在散热上，考虑到内存条和处理器的散热需求，因此采取如下布局方式：将处理器放在中间，较高的内存条放在两侧，所有的内存排列采取前后走向。其余的芯片如 IO 放置在不影响通道的位置。

5.2　硬件子系统

本节介绍其他关键子系统的设计方法。

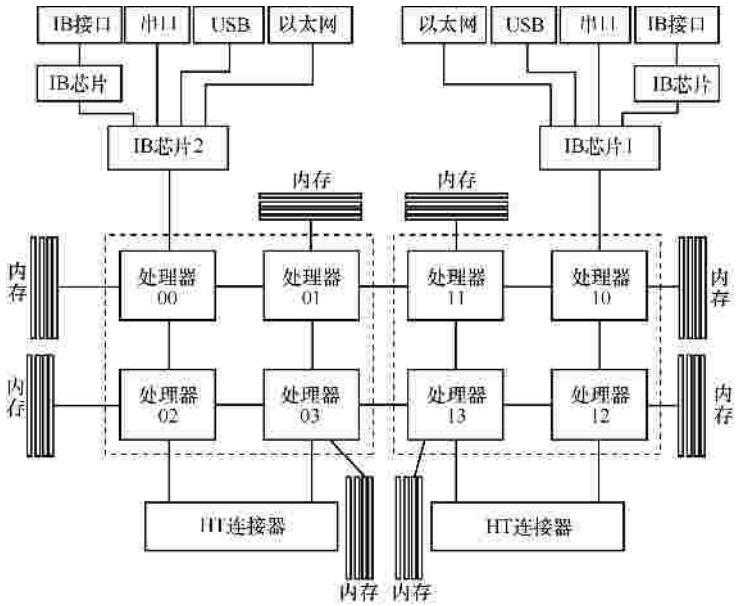


图 5 8 路主板逻辑图

高带宽 I/O 子系统：在系统架构和主板设计中充分考虑了较高的带宽性能和较高的扩展能力。高带宽 I/O 子系统包括集成在主板上的高速 Infiniband、吉比特以太网和 I/O 板卡槽。每个计算节点集成了 4 个 QDR 的高速 Infiniband 口，4 个吉比特以太网，还有 4 个 PCIE×8 的 I/O 卡槽，可用于安装 I/O 板卡。这些对外的系统接口保证了对外高 I/O 带宽，主板集成 4 个 QDR Infiniband 接口，综合传输带宽达到 160 Gbit/s。总体具有对外高带宽的网络，并可以通过外部的交换机连接成一个更大的系统，可见，高效能云计算节点具有很好的扩展性。

互连子系统：为满足系统的互连功能和良好的扩展性，每个计算节点设计了 4 个 QDR 的高速 Infiniband 口，4 个吉比特以太网，还有 4 个 PCIE×8 的 I/O 卡槽，可用于安装 I/O 板卡对外网络连接。Infiniband 网络接口可以连接到 Infiniband 交换机上，组成高速的计算网络。吉比特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。主板上的 I/O 板卡可以连接存储网络。

存储子系统：存储子系统设计包含本地存储和远程存储，其中，远程存储是通过 I/O 卡对外连接实现的，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，主板上的每个处理器支持 4 个内存条，作为板级的存储，用于计算数据的保存。这 3 部分组成了每个计算节点的存储系统。在由多个计算节点组成的系统中，I/O 系统和存储可合为一体，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。

电源子系统：电源系统包括主电源模块、电源分配板和主板上的 DC-DC 模块。系统采用宽电压设计，支持 110~240 V、50~60 Hz 宽电压输入。

高效能系统软件：高效能系统软件是高性能计算节点发挥其效能的关键。总体采用自下而上分 3 层来考虑系统软件：节点层系统软件、整机系统软件层、应用支撑软件层。

具体地，如图 6 所示，各层功能划分和详细设计如下所示。

图 6 高效能系统软件层次结构

节点系统软件层：单节点上承载的系统软件，主要包括基于龙芯 3A 的节点操作系统及其核心扩展以及节点提供的用户编程工具、环境和库。

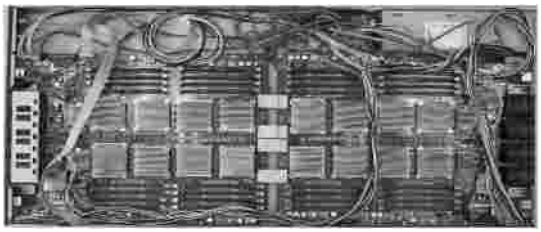
整机系统软件层：整机（多节点）层次所需的

系统软件。

应用支撑软件层：支撑应用编程的环境和工具软件。

6　原型实现与性能测试

笔者实现了基于龙芯 3A 多核处理器的高效能云计算节点，其整机与机箱实物图如图 7 所示。主板为对称的上下 2 个 PCB 板，由四对板间连接器互连，电源位于主板右侧，风扇位于主板后端，后面板提供 2 个 IB 接口，前面板提供网卡接口、USB 接口以及一块含电源开关。系统为每个龙芯 3A 处理器的每个 DDR2 通道各提供一个 DDR2 DIMM 内存，位于芯片同侧，因此板上每 2 个内存条为一个处理器提供服务。基于上述分析，所实现的高效能云计算节点的关键技术特征概括如下。



(a) 云计算节点机整机内部实现

(b) 云计算节点机整机外部接口实现

图 7　整机与机箱实物

板上机群：单计算节点包含两块完全相同的 PCB，每个 PCB 可分为左右近似对称的两部分，其中每部分运行一个操作系统。由于 PCB 间采用了高速接插件，因此从整个节点上看，对插的两块 PCB 可作为拥有 4 个操作系统的板上机群系统。

高密度高安全可控计算节点：每个节点包含国产龙芯 16 路处理器，处理器之间可通过 Hyper Transport 互连。单节点共包含 16 路处理器，虽然单节点上的 4 个操作系统每个都只包含其中的 4 路处理器，但龙芯 3A 处理器的可扩展性允许节点内部 16 路处理器互连，构成 4×4 的二维 mesh 结构，并可利用 BIOS 指定处理器间访问的路由配置。

多种互连方式：主板选用 InfiniBand 公司 MT25408 芯片实现系统间高速互连，选用 RTL8211B 芯片结合 *n*Vidia 南桥芯片的 Mac 层实现吉比特以太网互连，并且处理器的 HT1 总线可用于单节点内跨系统多处理器间互连，实现了多种互连方式。

7　性能测试

7.1　测试方法

为了测试基于龙芯 3A 多核处理器的高效能云计算节点的性能，如表 1 测试方法所示，其中“★”表示该行所对应的测试用例对该指标进行了测试，“—”表示该行所对应的测试用例对该指标未进行测试。本文对所实现的系统采用了 3 个测试用例，从功耗体积、通信性能和 Linpack 性能 3 个目标进行了实验。在测试用例和度量指标选取上，对测试用例 1，选取常用的功耗和体积 2 个常用指标。对测试用例 2，分别从点对点通信和 MPI 全局通信 2 个角度选取带宽和延迟进行了测试。对测试用例 3 分别从 Linpack 性能和可扩展性进行测试分析。本实验章节的部分数据是对论文成果[8]的数据扩展。

为表述方便，对后续的章节将用到的数学符号的含义说明如表 2 所示。

7.2　实验结果分析

实验 1　计算节点效能

单节点机的测试结果如表 3 所示。表 3 主要从

表 1　测试方法概览

测试用例	测试目标	度量指标				
		功耗	体积	带宽	延迟	可扩展性
测试用例 1	功耗和体积	★	★	—	—	★
测试用例 2	点对点通信	—	—	★	★	—
测试用例 3	MPI 全局通信	—	—	★	★	—
	Linpack 性能	—	—	—	—	★

表 2 简写数学符号含义

符号	含义	符号	含义
$ComNode_{Lg3a}$	基于龙芯 3A 的计算节点	$Node_{sys\#}$	HPC 系统中的节点数
$Core_{\#}$	单 CPU 的核数	$CPU_{sys\#}$	单 HPC 系统中的 CPU 数目
$Perf_{CPU}$	CPU 性能	Pow_{sys}	单 HPC 系统的功耗
$CPU_{n\#}$	单节点中 CPU 数目	$Cab_{\#}$	42U 机柜的数目
Pow_{node}	单节点功耗	$TCP_{mindelay}$	TCP 最小延迟
$Perf_{node}$	单节点性能	$TCP_{maxband}$	TCP 最大带宽
H_{node}	单节点高度		

表 3 单节点的效能

度量指标	CPU		单节点			
	$Core_{\#}$	$Perf_{CPU}$/GHz	$CPU_{n\#}$	Pow_{node}	$Perf_{node}$(GF)	H_{node}(U)
$ComNode_{Lg3a}$	4	1.0	16	~300	256	1（Rack）

表 4 基于不同处理器构建 1PF 高效能计算机系统的比较

性能	单 CPU	单节点			1PF 系统			
	Core#	CPUn#	Perfnode	Hnode(U)	Nodesys#	CPUsys#	Powsys(KW)	Cab#
Intel	8	2	192	0.5(Blade)	5 209	10 417	2 448	63
AMD	6	8	537.6	1(Rack)	1 861	14 881	2 382	47
Loongson3A	4	16	256	1(Rack)	3 907	62 500	2 872	98
Loongson3B	8	16	2 048	1(Rack)	489	7 813	477	13

单节点的 CPU 数目、功耗、性能、体积角度反映了基于龙芯 3A 的高效能云计算节点的效能。

基于龙芯 3A 的节点机的情况如表 4 所示，表 4 估算了分别用 Intel 8 核、AMD 6 核、龙芯 3A 和龙芯 3B 构建 1 Pflops 计算单元的功耗和体积。从表 4 中可以看出，如采用龙芯 3B 处理器构建 1 Pflops 计算单元，计算单元总功耗约为 477 KW，体积约占 13 个 42U 工业机柜。可见，相对于同期主流的 x86 处理器，用龙芯 3A(尤其是龙芯 3B)处理器构建计算单元在功耗和体积方面有明显优势，能较好地应对高效能目标中的低功耗和高密度挑战。

实验 2 通信性能

为了便于对基于 HT 互连的通信模块的时延性能进行对比分析，本文还在相同条件下使用 NetPIPE 软件测试了 IB 的通信时延和带宽。针对 IB 高速网卡的通信时延测试主要分为两部分：一部分是测试系统的 CPU 直接与 IB 网卡相连，一部分是测试系统的 CPU 与 IB 网卡间接相连；针对 IB 高速网卡的通信带宽测试主要分为两部分：分别是通信系统的 CPU 均直接与 IB 网卡相连以及通信系统的 CPU 均与 IB 网卡间接相连。

7.3 点到点通信性能

使用 NetPIPE 测试软件，在度量指标上测试不同计算单元间的点到点通信延迟与带宽。在测试互连场景上，对 3 种互连通信场景进行测试，具体包括同一虚拟子网内 2 个计算节点，同一 IB 子网内 2 个计算节点和相邻虚拟子网内 2 个计算节点间。测试结果如表 5 所示。

表 5 点到点通信延迟和带宽

测试场景\度量指标	$TCP_{mindelay}$/μs	$TCP_{maxband}$/(Mbit·s^{-1})
同一虚拟子网内 2 个计算节点	37.5	502.0
同一 IB 子网内 2 个计算节点	71.4	781.3
相邻虚拟子网内 2 个计算节点间	95.2	228.0

7.4 MPI 全局通信性能

使用 IMB 2.2 测试软件，对节点机不同规模的计算节点通过 16 384 byte 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 全局性能测试结果如表 6 所示。

表 6　　MPI 全局通信带宽和延迟

CPU#	SendRecv/(Mbit/s)	Reduce/μs	Allreduce/μs	Allgather/μs	Allgatherv/μs	Alltoall/μs	Bcast/μs	Barrier/μs
2	35.5	1 494.7	1 695.9	890.0	926.4	915.5	680.5	243.8
4	16.4	3 506.5	4 145.0	3 852.1	3 929.7	3 602.5	1 548.6	631.0
6	13.2	5 117.9	6 758.9	22 095.9	29 082.6	15 085.1	4 680.3	983.7
8	8.7	7 045.0	10 651.5	49 207.2	48 135.0	45 890.0	7 172.6	1 523.7

实验 3　系统 Linpack 性能

16 个 CPU（64 核）规模的系统和矩阵阶数从 10 000 增加到 48 000 时，Linpack 浮点运算性能如图 8 所示，在 64 核时，性能达到 44.5 Gflop。

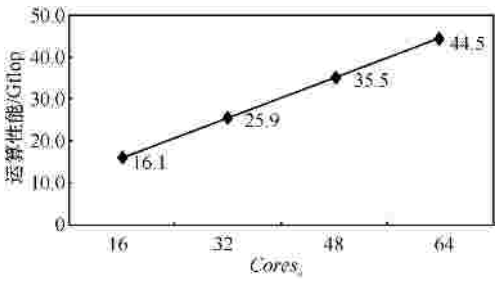


图 8　Linpack 性能

7.5　实验结果小结

对原型系统的实验和测试表明，基于龙芯 3 多核处理器的高效能节点机单节点具有每秒 0.256 万亿次浮点运算能力（Tflops），单一机柜可容纳 42 个 1 U 节点机箱，672 颗 CPU，2 688 个 CPU 核（672×4），总体具有高密度、高性能功耗比，基于自主知识产权处理器和安全可控性等优点。

7.6　相关研究

云计算节点是高性能计算机体系结构中除加速节点、主机单元、通信网络等外的核心部件之一，计算机节点的高效能已成为了整机“高效能”目标能否达成的关键。首先，目前在研和投入使用的一批大规模高性能计算机均十分注重计算节点高效能问题的突破，代表性的系统主要包括：Cray Titan[4]、日本富士通 K 计算机[9]、Tianhe-1A、Jaguar[10]、Cray 的 Cascade 和 Baker、IBM 的 Roadrunner 等[11~18]。其次，从处理器角度来看节点机，近年来随着多核处理器的兴起，目前主流的高效能计算机节点普遍呈现采用多路多核技术的趋势。但现有技术在路数上，尤以 4 路和 8 路居多，8 路以上高密度的比较缺乏。再次，在处理器选型上，现有高效能云计算节点机仍然主要采用国外处理器居多，在基于国产（如龙芯 3A 多核）处理器的高密度的计算节点的实践上相对较为稀少甚至缺乏。文献[8]中概要介绍了计算节点在设计和研制中的一些关键问题及其硬件方面的设计（包括系统总体结构以及处理器互连、时钟、上电与复位、内存、主板和结构等子系统）方法，但文献[8]更多是从高性能计算角度，并没有从云计算角度考察，尤其是并没有对云计算节点机的逻辑结构、互连规则、通信方法和软件系统等给出更为细致的理论建模和详细设计与实现等的介绍，本文工作是对文献[8]的进一步扩展和系统化。基于现有研究存在的问题，本文介绍了一种基于龙芯 3A 的高效能云计算节点机软硬件设计和实现方法。所实现的基于龙芯 3A 处理器的节点机是一款 16 路 1 U 机架式计算节点，占地 0.46 m^2，峰值性能 256 GFlops，峰值功耗不超过 300 W/U，计算/功耗比约 0.853 GFlops/W，如表 7 所示，与 IBM 等一些主流高性能计算节点相比，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，龙芯节点机的高性能、低功耗特点在实际运行中将体现得更为明显。

表 7　　相关云计算节点机的比较

云计算节点机	CPU 数量和类型	CPU 时钟频率	节点功耗	节点体积	基于龙芯 CPU
基于龙芯 3 多核 CPU 的云计算节点	16 路 4 核 Loongson 3A	1.0 GHz	约 300 W	1 U Rack	Y
IBM Blade JS20 服务器	2 路，IBM PowerPC970	2.2 GHz	约 395 W	约 1/2 U Blade	N
IBM BladeCenter JS22 Express 服务器	4 路，POWER6	4.0 GHz	约 350 W	约 1/2 U Blade	N
TYAN GT62B8230-L 服务器	2 路，AMD 12-Core Opteron 6100	2.1 GHz	约 350 W	1U Rack	N

8 经验和教训

本节将重点对基于龙芯多核处理器的高效能云计算节点机在软硬件设计与实现方面的经验和教训，以及与现有系统的优势与差距方面的具体原因进行一些讨论与分析。

系统架构方面：相对于现有系统，由于本系统采用 mesh 结构互连的 16 路处理器，由每 4 路处理器构成一个 CC-NUMA 单元，该架构的优势是能够充分利用并行性，但受系统总线带宽以及龙芯处理器开放接口数的限制，所能支持的处理器数目会受到限制，再加上软硬件设计工艺复杂，对于商用化目标来说，相对造价成本会较高，目前笔者正在进一步优化该架构，正在研究一种“多胞胎”架构的节点机。

处理器的性能和稳定性方面：由于龙芯处理器是核心，处理器的性能和稳定性直接关系到系统的性能、设计和实现的难度和效果。

器件布局、高速型号处理和散热方面：由于单主板的电子元器件数量和种类都比较多，需要处理的核心部件是处理器和内存，在布局上尤其需重点考虑把这两类部件的布局、高速型号处理、电子线路和散热规划好。基于对处理器之间的互连结构，及其各自连接内存条以及工艺难度等方面的综合考虑，因此，在布局上需要综合考虑好处理器和内存条互连，尤其要处理好高速信号的连接问题，笔者最终采用了规则的方阵式布局，即处理器放在中间，内存条放置在两侧。这种规则布局的方式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。在满足内存条和处理器的散热需求时，需要综合考虑分冷效果，笔者采取将处理器放在中间，较高的内存条放在两侧，所有的内存排列采取前后走向，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 I/O 放置在不影响通道的位置。

互连方式方面：多种方式的支持目标要求重点考虑如何支持系统间的高速互连、吉比特以太网互连以及单节点内跨系统处理器互连。笔者的主板选用 InfiniBand 公司 MT25408 芯片实现系统间高速互连，选用 RTL8211B 芯片结合 *n*Vidia 南桥芯片的 Mac 层实现吉比特以太网互连，并且处理器的 HT1 总线可用于单节点内跨系统多处理器间互连，这样就实现了多种互连方式。

软硬件配合方面：软硬件总体采用分层的设计思路，但底层软件（如操作系统等）对龙芯处理器的指令集等的依赖性还是比较大，在软件实现方面也需要针对性地定制和优化，以取得较好的性能效果。

在商用化及云计算节点经济性能方面：当前系统总体处于原型系统研制阶段，尚没有价格方面的数据。笔者基于材料成本进行了初步估算，除龙芯 CPU 外，系统的其他材料都是市场通用的材料。一颗龙芯 3A CPU 约 1 500 元，而目前性能和龙芯 3A 最接近的 x86 服务器 CPU 的成本约为 500 元。总体看来，云计算节点的经济性能和国外同类产品相比而言仍然存在比较大的差距，而这方面差距的核心是由国产 CPU 的性能和成本等核心元器件决定的。随着国产处理器生态环境的逐步改善，基于国产处理器的高效能云计算节点机预期或能取得更大的商业应用上的突破。

9 结束语

适于云计算的高效能计算系统是新一代高性能计算机重要研制目标。高效能云计算节点是 HPCS 的核心部件，是实现高效能目标的关键。性能功耗比、高集成度等高效能目标的实现是当前高效能计算节点面临的关键挑战。基于国产处理器的高效能云计算节点突破更是我国自主知识产权和安全可控高效能计算机战略的关键。本文关注于现有研究中基于龙芯处理器的高效能节点机实际系统相对缺乏等问题，提出了一种基于龙芯 3A 多核处理器的高效能云计算节点的软硬件设计和实现方法，并研制成功了相关的原型系统。测试数据、已实现的原型系统以及同现有系统的对比，验证了本文所研制的系统在基于国产龙芯处理器、性能功耗比可集成度和自主知识产权等方面具有优势，是基于国产处理器的一个重要尝试。本文的工作更是下一代基于龙芯 3B 高效能云计算节点，以及基于龙芯处理器的高效能计算机设计与研制的重要基础。

基于本文工作成果，目前正在开展基于龙芯 3B 国产处理器的高效能云计算节点，以及适于云计算的基于龙芯处理器的高端服务器的软硬件设计与研制等工作。

参考文献：

[1] SOTOMAYOR B, MONTERO RUBEN S, LIORENTE I M, *et al.*

Virtual infrastructure management in private and hybri clouds[J]. IEEE Internet Computing, 2000, 13(5):14-22.

[2] ZHANG W, SONG Y, RUAN L, *et al*. Resource management in internet-oriented data centers[J]. Journal of Software, 2012,23(2):179-199.

[3] Hpcs-high productivity computer systems[EB/OL]. http://www.highproductivity.org/.Top500 Supercomputer 2012 Rank List.

[4] http://www.top500.org/lists/2013/11/.

[5] CUI Y F, OLSEN KIM B, JORDAN T H. Scalable earthquake simulation on petascale supercomputers[A]. Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis[C]. Washington DC, USA, 2010. 1-20.

[6] RUAN L, XIAO L M, ZHU M F. Content addressable storage optimization for desktop virtualization based disaster backup storage system[J]. China Communications, 2012,9(7):1-13.

[7] LIU Y H, ZHU M F, XIAO L M, *et al*. Design and implementation of loongson 3A CPU based high productivity computing nodes[J]. High Performance Computing Technology, 2010, 6:46-53.

[8] KUROKAWA M. The K computer: 10 Peta-FLOPS supercomputer[A]. The 10th International Conference on Optical Internet(COIN)[C]. Yokohama, Japan, 2012.29-31.

[9] XIE M, LU Y T L, LIU L, *et al*. Implementation and evaluation of network interface and message passing services for tianHe-1A[A]. 19th Annual Symposium on High Performance Interconnects (HOTI)[C]. Santa Clara, CA, 2011.78-86.

[10] NCCS[EB/OL]. http://www.nccs.gov/jaguar, 2010.

[11] MCCURDY C W, STEVENS R. Creating Science-Driven Computer Architecture: a New Path to Scientific Leadership[R]. NERSC Technical Report, 2002.

[12] VECCHIOLA C, PANDEY S, BUYYA R. High-performance cloud computing: a view of scientific applications[A]. The 10th International Symposium on Pervasive Systems, Algorithms, and Networks[C]. 2009.4-16.

[13] HABATA S, YOKOKAWA M, KITAWAKI S. The earth simulator system, the earth simulator system[J]. NEC Res & Develop, 2003, 44(1):21-26.

[14] LAUDON J, LENOSKI D. The SGI origin: a ccNUMA highly scalable server[A]. International Symposium on Computer Architecture[C]. Colorado, USA, 1997.241-251

[15] ABTS D. The cray XT4 and seastar 3-D torus Interconnect[M]. David Padua,ed: Encyclopedia of Parallel Computing, 2011.470-477.

[16] SCOTT S. Thinking ahead: future architectures from cray[EB/OL]. http://nccs.gov/news/workshops/cray/pdf/Cray_Tech_Workshop_sscott_2_26_07.pdf, 2007.

[17] FELDMAN M. ORNL gears up for new leadership computing ystems[EB/OL]. http://www.hpcwire.com/ hpc/1356225.html,2007.

[18] TURNER JOHN A. Roadrunner: Heterogeneous Petascale Computing for Predictive Simulation[R]. Los Almos Unclassified Report LA-UR-07-1037, 2007.

作者简介：

阮利（1978-），女，四川成都人，博士，北京航空航天大学硕士生导师，主要研究方向为虚拟化与云计算、分布式与并行存储、高效能计算机与高端服务器。

秦广军（1977-），男，河南安阳人，北京航空航天大学博士生，主要研究方向为计算机系统结构、网络与通信协议、高性能计算机。

肖利民（1970-），男，江西南康人，博士，北京航空航天大学教授，主要研方向为计算机系统结构、虚拟化与云计算、高效能计算机与高端服务器。

祝明发（1945-），男，四川岳池人，博士，北京航空航天大学教授，主要研究方向为计算机系统结构、人工智能、虚拟化与云计算、高效能计算机。