

# HPP: 一种支持高性能和效用计算的体系结构

孙凝晖<sup>1)</sup> 李 凯<sup>2)</sup> 陈明宇<sup>1)</sup>

<sup>1)</sup>(中国科学院计算技术研究所 中国科学院计算机系统结构重点实验室 北京 100190)

<sup>2)</sup>(普林斯顿大学计算机科学系, 普林斯顿, 新泽西 08544, 美国)

**摘 要** 为了同时做到应对千万亿次高性能计算的技术挑战和满足数据中心(data center)未来的主要应用模式效用计算(utility computing)的需求,提出了一种称为 HPP(Hyper Parallel Processing)的高性能计算机体系结构. HPP 的主要特征是全局地址空间(global address space)和单一操作系统映像的超节点(hyper node). HPP 结合了 MPP 的可扩展性,DSM 的高效通信和机群的普及化的优点,为高性能计算和效用计算都提供了许多创新研究的机会. 基于 HPP 体系结构,实现了一个曙光 5000 高性能计算机的原型系统,初步验证了它的可行性.

**关键词** 高性能计算;效用计算;体系结构;超并行;千万亿次

**中图法分类号** TP303

## HPP: An Architecture for High Performance and Utility Computing

SUN Ning-Hui<sup>1)</sup> LI Kai<sup>2)</sup> CHEN Ming-Yu<sup>1)</sup>

<sup>1)</sup>(Key Laboratory of Computer System and Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

<sup>2)</sup>(Department of Computer Science, Princeton University, Princeton, New Jersey 08544, USA)

**Abstract** An architecture of high performance computer, called Hyper Parallel Processing (HPP), is proposed to satisfy the requirements of both High Performance Computing, and Utility Computing which will be the application model of data centers. HPP combines the benefits of the scalability of MPP, the communication efficiency of DSM, as well as the commodity of cluster. Comparisons of current main-stream high performance computer architectures show that none of them can satisfy both HPC and utility computing.

The main features of HPP are Global Address Space(GAS) and Hyper Node with single Operating System image. HPP supports the distributed global address space including both memory and I/O, but without hardware cache coherence. A Hyper Node consists of a set of application specific CPUs and a (or more) OS specific CPU. The OS CPU maintains the single system image, while the application CPUs run only lightweight run-time software. Besides the GAS interconnect network for applications, a standard SAN connects all OS CPUs and I/O devices providing system management and storage service.

HPP is able to provide many opportunities of innovative research in High Performance and Utility Computing areas, including communication, synchronization, programming model, node operating system, utility computing, fault isolation, CPU and system etc. According to HPP architecture, a prototype system of Dawning5000 HPC is implemented and the feasibility of HPP is proved.

**Keywords** high performance computing; utility computing; computer architecture; hyper parallel processing; petaflops

## 1 引言

每秒千万亿次浮点运算(petaflops,以下简称千万亿次)是当前高性能计算机研究的一个主要目标.学术界、企业界、用户都在为此而努力.达到这样的峰值,甚至某些应用达到这样的应用饱和性能,都不是难事,无论采用 Cray XT4 那样的大规模并行(MPP)体系结构、NEC 地球模拟器那样的并行向量机(PVP),还是 IBM BladeCenter 那样的工业标准机群(Cluster).困难的是能够解决高性能计算机面临的 4 个巨大挑战:(1) LPC 挑战,即低功耗(low power)、低占地(low proportion)、低价格(low price). IBM BlueGene<sup>[1]</sup>基于 SoC 的 MPP 系统是一种应对方法;(2) VLSP 挑战,即超大规模并行度(Very Large Scale Parallelism)带来的系统可靠性、互连网络的可扩展性、并行算法设计、并行编程模型和语言上的困难. IBM 基于 Cyclops64 CPU 的新型 HPC<sup>[2]</sup>是一种应对方法;(3) Productivity 挑战,即高效能,提高 HPC 作为超级计算中心的公共基础设施的产出能力,包括制约应用的实际效率的通信和同步性能、多用户使用环境下的 I/O 性能、故障的隔离和资源的虚拟化的需求,提高系统利用率和总拥有成本的管理技术,简化用户应用程序的开发和调试等. 美国 HPCS 计划(High Productivity Computing System)<sup>①</sup>中的技术需求覆盖了上述 3 个挑战.该计划中面向千万亿次的 IBM PERCS 系统和 Cray Cascade 系统对这些挑战都有所应对.(4) Commodity 挑战,即普及化,这是常常被忽视了的.高性能计算机的主流市场经历了向量机、MPP 系统、机群系统 3 个时代,机群能够取代以前以性能为导向的时代的主要原因,是它的 Commodity 特性<sup>[3]</sup>,即高性能计算机的主要部件(节点、互连网络、存储、操作系统、编译器)是数据中心市场的主流商品.我们看到,当前的大多数研究都是试图解决前 3 个挑战,面向军事应用、科学研究应用的最高端用户的需求.体系结构的创新不能解决所有的挑战性问题,但对许多问题的解决能够提供新的机会.我们的研究动机是,提出一种新型体系结构,面向这 4 个挑战,面向高性能计算机主流市场的下一个时代.

效用计算<sup>[4]</sup>(utility computing)是一种资源共享、按需服务的应用模式,是未来数据中心的发展方向.当前许多技术是面向这一目标的,例如,Web Service 是 Internet 应用领域的应用层按需服务技

术,网格计算是 HPC 应用领域的应用层按需服务技术,虚拟机是计算机系统的资源共享技术,分区(partition)是大型服务器的资源共享技术.目前,在计算机体系结构上还缺乏对效用计算的有效支持.所以,我们体系结构研究的技术路线是,提出一种同时支持高性能和效用计算的体系结构,它能够使系统达到千万亿次的性能,兼容机群的硬件、软件、应用等主要部件,高性能计算机的节点能够成为数据中心的效用计算服务器,并为高性能计算机的创新研究提供新的机会.

基于以上的研究目标,我们提出了超并行(HPP)体系结构.本文的第 2 节介绍了一些相关工作;第 3 节介绍了 HPP 体系结构的思想;第 4 节给出了与高性能计算机其它主流体系结构的对比和可能的创新研究机会;第 5 节介绍了一个案例,面向千万亿次计算的曙光 5000 的一台原型系统的实现,验证了 HPP 体系结构的可行性;最后是结论部分.

## 2 相关工作

本节简单介绍当前的一些主要的高性能计算机体系结构,分析它们应对本文的研究动机的不足之处,并对这些体系结构的主要特征进行了分析.

当前,能够达到千万亿次计算的体系结构包括:

(1) Cluster<sup>[3,5]</sup>: 机群体系结构,采用工业标准部件,如 IA 服务器、商品化互连网络、Linux 操作系统、MPI 并行编程环境、SAN 存储设备等,应用软件众多;

(2) Constellation: 星群体体系结构,采用 SMP<sup>[6]</sup>或 DSM<sup>[7-8]</sup>体系结构的 RISC 服务器,专用的互连网络,Unix 操作系统,从体系结构的角度看,与机群是属于同一类系统;

(3) MPP<sup>[9]</sup>: 大规模并行体系结构,采用可大规模扩展的专用互连结构,如 3D Torus,对超大规模时的通信和同步有特殊支持,采用专用的高密度系统组装方法、专用的节点操作系统;当前一个新的进步是 SoC MPP,将互连网络的路由功能集成在 CPU 芯片中,其中 CPU 是面向低功耗、高组装密度设计的;

(4) PVP: 分布式共享存储向量机,是采用向量处理器的 DSM 系统,采用专用的处理器、专用的互连网络、专用的系统软件;

① DARPA. High Productivity Computer System. <http://www.highproductivity.org/>

(5) Adaptive Supercomputing: 混合结构的自适应超级计算的典型代表是 Cray Cascade, 由通用多核、多线程、FPGA 可重构计算、专用众核结构的处理单元组成计算节点, 按应用特征动态分配任务;

(6) 基于专用处理器的系统: 设计一种专用处理器, 面向特定的应用领域, 例如, MDGrape3<sup>[10]</sup> 采用算法硬化的计算单元, 面向分子动力学应用; Cyclops64 采用细粒度显式存储的众核结构 (many-core), 面向情报分析应用; GPU 采用大规模运算单元和流式存储结构, 面向流式应用。

机群的最大优点是数据中心用服务器作为高性能计算机的节点, 它的主要不足是无法应对高性能计算机的前 3 个挑战, 在向高端发展时体系结构的限制尤显明显。例如, 点到点通信的延迟在商品化互连网络上很难低于 1 $\mu$ s, 另外在资源的有效利用

上也没有特别的支持。MPP 是为了 HPC 向上扩展设计的, 在向下扩展 (down-size) 时就没有优势了, 它的节点也无法支持数据中心应用。SMP 和 DSM 是非常适合效用计算的, 它的不足是可扩展性远远不够, 不能适合 HPC 的要求, 而且对故障的隔离性也不好。PVP、专用处理器系统、自适应超级计算等体系结构依赖于 HPC 应用专用的处理器, 很难适合通用的数据中心。

理想的体系结构首先要能满足千万亿次高性能计算机的需求, 具有 MPP 在性能和可扩展性上的优势, 能向上扩展; 同时它的节点能成为可量产的商品 (volume commodity), 在向向下扩展时, 具有 Cluster 的那些优势, 并提供一些创新研究的机会。

图 1 分别对高性能计算机的主要体系结构的重要特征进行了对比。

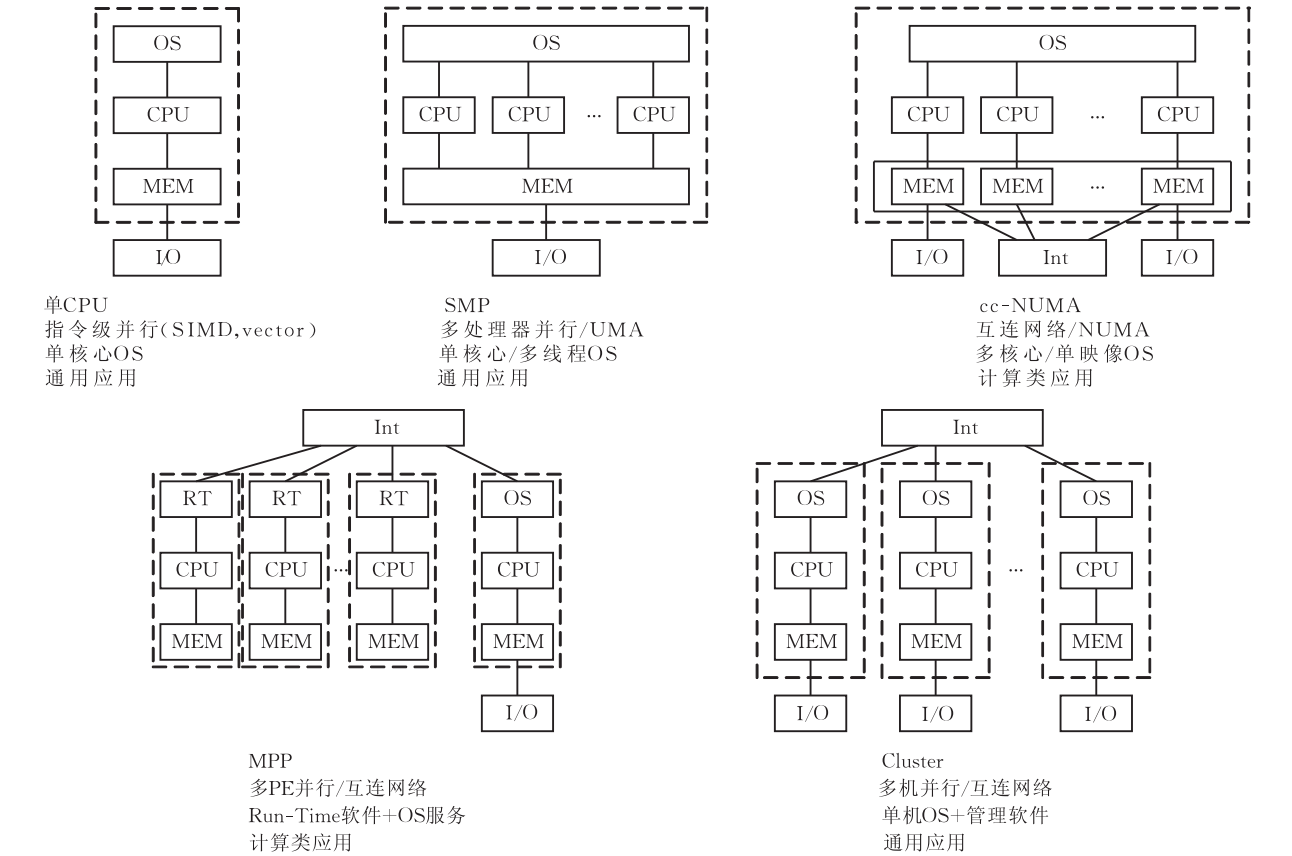


图 1 传统的高性能计算机体系结构特征图

### 3 超并行计算机体系结构

HPP 体系结构的思路是在系统硬件层面向 MPP、DSM 靠齐, 关注可扩展性和性能; 在系统软件层面向 Cluster 靠齐, 关注产业化和市场容量。从

用户和应用的角度看, 它应是一种改进的机群系统, 能够兼容已有的软硬件部件和应用。HPP 的主要特征包括:

(1) 全局地址空间 (Global Address Space). 它支持全系统的硬件分布式统一编址, 包括内存和 I/O, 但不要求硬件支持的 cache 一致性 (如 DSM 系

统);使系统保证较好的通信和同步性能和良好的可扩展性;这就要求与 CPU 和 NIC 连接的系统控制器(system controller)能够支持该特征;

(2) 功能分离的 CPU. CPU 分为应用 CPU 和操作系统 CPU, I/O 设备都连接在操作系统 CPU 上. 这种结构的好处是, 操作系统 CPU 可以采用工业标准部件, 通过互连网络连成机群结构, 能利用机群的所有好处, 如标准化部件、驱动程序、已有的第三方软件、应用等;而留给应用 CPU 更多创新设计的可能;

(3) 超节点(Hyper Node). 一组应用 CPU 和若干操作系统 CPU 构成一个超节点, 应用 CPU 运行轻量级运行时支持(Real-time)软件, 支持应用程序的运行;操作系统 CPU 提供所有的操作系统服务, 维持单一系统映像;超节点应能够成为数据中心的效用计算服务器, 成为批量的商品化部件;

(4) 功能分离的互连. GAS 互连网络连接所有应用 CPU, 支持高性能计算应用;操作系统互连网络连接所有操作系统 CPU 和 I/O 设备, 对存储、管理提供支持, 提供全局服务, 可以是标准的机群互连网络, 如 Infiniband;特别地, GAS 互连连接两类 CPU, 并对单一映像操作系统提供支持, 如细粒度锁.

图 2 对 HPP 体系结构的重要特征进行了说明.

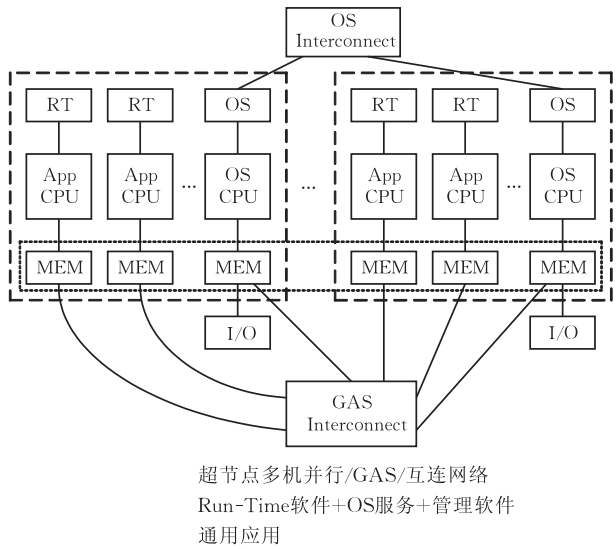


图 2 超并行计算机体系结构特征图

4 HPP 体系结构提供的研究机会

下面分析 HPP 体系结构给高性能计算和效用计算提供的新的研究机会.

(1) 通信(communication). 由于有全局地址空

间, 可以在系统控制器中实现通信协议, 不需要像机群那样由专门的通信协处理器实现消息传递协议, 也可以比较容易地支持单边通信(single-side communication)、远程数据存取、用户空间 RDMA 等, 可提高大规模系统的通信性能.

(2) 同步(synchronization). 由于有全局地址空间, 可以实现可寻址的硬件同步机制、全局锁等, barrier、collective 操作也可以有新的实现方法.

(3) 编程模型(programming model). 除了 MPI 消息传递模型能够提高性能外, 全局地址空间对 OpenMP、PGAS(Partitioned GAS)语言如 Co-Array Fortran、Unified Parallel C、Titanium(Java 并行扩展), 适应多核 CPU 的“细粒度多线程+PGAS”编程模型等, 也能够提供硬件支持.

(4) 节点操作系统. GAS 互连网络为分布式操作系统提供了细粒度的同步机制, 可以高效地实现一个多核心(multi-kernel)的单一映像操作系统, 超节点可以做为高性能计算机的计算节点, 或数据中心服务器, 或高性能工作站;超节点的特点也使得千万亿次高性能计算机的操作系统映像数量大为减少, 提高了整机的可靠性和易管理性.

(5) 效用计算(utility computing). 全局地址空间对资源的虚拟化提供了很好的支持, 例如, 可以实现共享的全局内存或者全局 I/O 设备, 应用能够使用全系统的内存, 或者文件系统利用全系统的内存作为文件缓冲区, 或者应用利用全系统的 I/O 设备;可以实现一种分布式虚拟机, 实现系统资源在节点之间动态流动, 如内存容量在不同应用之间的动态分配.

(6) 故障隔离(isolation). HPP 体系结构提供了基于 CPU 或者节点的故障隔离的可能;两类互连网络的设计使得系统可以阻止故障的传播;全局地址空间还为一些创新的检查点机制提供了可能, 比如故障时检查点(crash-time checkpoint), 在应用出现故障时, 才由操作系统 CPU 切取检查点, 可以大为减少检查点开销.

(7) CPU. 分离的 CPU 设计为 CPU 创新打开方便之门, 同样的体系结构、互连网络、操作系统可以支持多种应用 CPU, 满足不同种类应用的需求, 比如通用多核应对机群应用, 专用众核应对计算密集应用, 低功耗 CPU 应对个人超级计算机等等, 大大减少了开发的周期;对应用 CPU 的要求仅是支持全局寻址, I/O 接口可以是 HT、PCI-E、Intel Front-Bus、QPI 等.

(8) 系统. HPP 体系结构的系统可以分成 3 个主要部件: 计算单元、超节点和互连网络. 计算单元可以是单 CPU 系统、也可以是 SMP 系统, 通过 HPP 系统控制器连接到 GAS 互连网络; 一个超节点包括多个计算单元、一个操作系统运行单元、I/O 设备、连接操作系统的互连网络, 运行一个单一映像操作系统; 超节点通过互连网络连接成系统; 其中互连网络可以有多种实现方式, 但应保持连接的灵活性和

对故障的隔离性. 超节点的设计可以有多种创新, 如面向计算密集的刀片式结构, 面向数据中心的以可靠性为主的结构, 面向个人超级计算的易管理性结构.

表 1 给出了 HPP 与其它高性能计算机主流体系结构的对比. HPP 与 MPP、NUMA 相比最大的好处还是对故障的隔离性, 它的节点是能够独立使用的单元; 而最大的不足是可扩展性不如 MPP 系统, 达不到 IBM BlueGene 那样的超大规模.

表 1 HPP 与其它主流高性能计算机体系结构的对比

	高性能计算							效用计算				低能耗
	粒度	通信性能	I/O	可靠性	可管理性	可扩展性	可编程性	商品化	虚拟化	资源共享	隔离性	
MPP	应用	好	好	好	好	超大规模	消息传递	差	差	差	不确定	好
CC-NUMA	线程	好	好	不确定	好	小规模	共享内存	差	Fine grain	好	差	差
Cluster	进程	中	差	差	差	中等规模	消息传递	好	Node level	差	好	差
HPP	进程	好	好	好	好	大规模	消息传递, PGAS	好	Fine grain	好	好	不确定

5 曙光 5000 原型系统的实现

曙光 5000 是面向千万亿次计算的高性能计算机, 它将采用 HPP 体系结构和龙芯 3 16 核 CPU. 下面描述曙光 5000 中与 HPP 体系结构相关的主要设计.

(1) 计算单元. 曙光 5000 的计算单元包括 2 颗 16 核龙芯 3 CPU, 连接到一个 HPP 系统控制器上, 控制器包括连接 CPU 的 HT 接口、对全局地址空间的支持、可寻址的硬件锁、具有 4 个独立通道的互连网络接口;

(2) 超节点. 曙光 5000 的节点由 8 个计算单元和 1 个操作系统单元组成, 它们之间由一个 16 端口的互连芯片连接, 其中 9 个端口内连, 7 个端口连接到骨干交换机上. 操作系统单元采用标准的 X86 处理器和各种 I/O 设备, 连接 Infiniband 互连网络; 曙光 5000 的全局存储设备连接到 Infiniband 网络上;

(3) 互连网络. 曙光 5000 的 GAS 网络由 16 端口互连芯片按照多级网络(multi-stage)拓朴连接而成, 3 级可扩展到 1024 个计算单元, 4 级可扩展到 8192 个计算单元, 互连网络内置硬件实现的 barrier 和 bcast 操作. 曙光 5000 的通信协议既支持消息传递, 又支持远程内存读写(ld/st)和 RDMA, 同时支持标准 TCP/IP 协议;

(4) 节点操作系统. 曙光 5000 的节点操作系统由 run-time 软件和扩展的标准 Linux 组成. Run-time 软件支持最基本的功能, 包括应用程序的加载和运行管理, 支持龙芯 3 CPU 的 x86 虚拟机的运

行, 能够实现 X86 应用的二进制翻译. 所有的系统调用和操作系统服务都转发到扩展的标准 Linux 上, 它运行在 X86 CPU 上, 实现用户需要的所有操作系统功能; 从用户的角度看, 节点就是一台支持 16 个多核 CPU 的服务器. 在节点操作系统之上可运行标准的机群管理软件, 用户从全系统的角度看, 还是一个机群系统;

(5) 虚拟机. 曙光 5000 的虚拟机包括两类, 一是基于龙芯 3 CPU 的指令级虚拟机, 支持 X86 应用程序, 所以从用户的角度看, 使用曙光 5000 和使用 X86/Linux 机群一样, 应用软件能全兼容; 另一个虚拟机是系统级的分布式虚拟机, 利用 HPP 的特性, 实现内存和 I/O 设备在全系统的共享和动态流动;

(6) 编程模型. 曙光 5000 的编程模型包括标准的 MPI, 还有利用 HPP 特征优化的 UPC<sup>①</sup> 以及细粒度多线程和它们的混合模式;

(7) 管理软件. 曙光 5000 的管理软件包括利用 HPP 特性的故障时检查点, 即只在故障发生时由操作系统单元切取检查点, 还有支持效用计算模式、利用分布式虚拟机的资源动态管理软件.

到目前为止, 我们应用 HPP 体系结构开发了一个 HPP 超节点, 作为曙光 5000 原型系统, 以验证其可行性. 我们实现了对龙芯 2E SYSAD 总线、AMD Opteron 的 HT 总线的支持, 实现了计算单元、超节点、互连网络芯片、节点操作系统. 详细的数据分析正在进行中.

① UPC Consortium. UPC Language Specifications 1.2, 2005. [http://crd.lbl.gov/UPC/images/f/f6/Upc\\_specs\\_1.2.pdf](http://crd.lbl.gov/UPC/images/f/f6/Upc_specs_1.2.pdf)

## 6 结 论

面向高性能计算和效用计算,我们提出了一个 HPP 体系结构,它与其它面向千万亿次高性能计算的技术路线不同的是,考虑可扩展性的同时也考虑了它的主要部件能用于未来的数据中心的效用计算,从而结合了 MPP 的可扩展性,DSM 的高效通信和机群的普及化的优点. 我们还没有通过实际系统验证它的好处,现在正在开发软件模拟器和原型系统进行一些主要特征的验证. 采用 HPP 结构的曙光 5000 高性能计算机预计在 2010 年最后完成.

不存在一种完美的体系结构,不同的体系结构有它适合的应用和市场定位. HPP 也是一样,它不能解决所有的挑战性问题,期望它能成为解决千万亿次计算的体系结构中适用面最广的一个.

## 参 考 文 献

- [1] IBM BG/L Team. An overview of BlueGene/L supercomputer//Proceedings of the ACM Supercomputing Conference, 2002
- [2] del Cuvillo J B, Hu Ziang, Zhu Weirong, Chen Fei, Gao G R. Toward a software infrastructure for the cyclops64 cellular architecture. Department of Electrical and Computer En-

gineering, University of Delaware, Newark, DE: CAPSL Memo55, 2004

- [3] Sterling T, Becker DJ, Dorband J E, Savarese D, Ranawake U A, Packer C V. Beowulf: A parallel workstation for scientific computation//Proceedings of the 24th International Conference on Parallel Processing, 1995
- [4] Ross J W, Westerman G. Preparing for utility computing: The role of IT architecture and relationship management. IBM Systems Journal, 2004, 43(1): 5-19
- [5] Sunderam V S. PVM: A framework for parallel distributed computing. Concurrency: Practice and Experience, 1990, 2(4): 315-339
- [6] Conway M E. A multiprocessor system design//Proceedings of the 1963 Fall Joint Computer Conference, AFIPS Conference Proceedings, 1963, 24: 139-146
- [7] Wulf W A, Harbison S P. Reflections in a pool of processors: An experience report on C.mmp/Hydra//Proceedings of the AFIPS National Computer Conference, 1978: 939-951
- [8] Lenoski D, Laudon J, Gharachorloo K, Weber W, Gupta A, Hennessy J, Horowitz M, Lam M. The stanford DASH multiprocessor. IEEE Computer, 1992, 25(3): 63-79
- [9] Hillis W D, Tucker L W. The CM-5 connection machine: A scalable supercomputer. Communications of the ACM, 1993, 36(11): 31-40
- [10] Fukushima T, Taiji M. A highly-parallelized special-purpose computer for many-body simulations with an arbitrary ventral force: MD-GRAPE. The Astrophysical Journal, 1996, 468: 51-61



**SUN Ning-Hui**, born in 1968, Ph.D., professor, Ph.D. supervisor. His major research area is high performance computer.

**LI Kai**, born in 1954, Charles Fitzmorris professor. His research interests include operating systems, parallel and distributed systems, scalable display systems and content-based search.

**CHEN Ming-Yu**, born in 1972, Ph.D., professor. His research interests include high performance computer architecture, operating systems and parallel application.

## Background

Today high performance computer architecture has developed into a transforming age. The main-stream cluster architecture is facing more power and reliability problem. Many new technologies such as multi-core are emerging. Researchers in the world are doing research on the new architecture. An ideal architecture should solve the petaflop-scale problem as well as the commercialization issue of middle-to-small scale system. There are many technical and application challenges.

The Institute of Computing Technology has been working on high performance computer for a long period. According to the trend, the authors propose a new architecture, hyper parallel processing, to satisfy the requirements of both

High Performance Computing and Utility Computing. This work is supported by the National Science Foundation of China and the Hi-tech research and development program of China. The International Partnership Project of Chinese Academy of Sciences also supported the cooperation with Professor LI Kai of Princeton University.

This paper is a progress report of this work. The motivation, main issue, basic definition and research chances of HPP are given in this paper. Now the simulator and a FPGA prototype of HPP are already finished. Test and analysis result will be presented later.