成果与应用

曙光 4000A 超级服务器的研制*

孙凝晖
(计算技术研究所 北京 100080)

摘要 曙光 4000A 是我国十五“863”计划支持的重大项目,总体上处于当前国际先进技术水平;在高组装密度的服务器设计、大规模机群的管理技术、网格路由器技术等方面达到国际领先水平。本文主要介绍曙光 4000A 的研制目标和总体思路、取得的成果、性能评测以及应用推广情况。

关 键 词 曙光 4000A,超级服务器

“曙光 4000A 超级服务器” 是计算技术研究所国家智能计算机研究开发中心承担的国家十五“863”计划“高性能计算机及其核心软件”专项重大课题。在科技部和中科院领导的关怀下,在专项专家组的指导下,经过近两年的艰苦努力,于 2004 年 4 月投入稳定运行,6 月完成系统鉴定,8 月作为国家“863”计划支持的“中国国家网格”(CNGrid)中的一个主结点落户上海超级计算中心。

1 研究目标和总体思路

曙光 4000A 研制目标是: 在 2004 年 6 月研制成一套“曙光 4000A”超级服务器系统,系统峰值浮点速度为每秒 4 万亿次,支持科学工程计算、事务处理和网络信息服务的通用高性能计算机,系统与主流商品化系统在部件、应用上二进制兼容,具有可扩展性、安全性、好用性、易管理性、高可用性和支持网格应用的特点。

此外,“863” 专项专家组还要求曙光 4000A 应成为“中国国家网格”的主节点。在产业化方面,必须能很快从科研性样机转换成曙光服务器产品系列,广泛地用于我国国民经济各个领域;产生的一批相关的关键技术,能为提高国产服务器的技术含量,促进我国高性能计算机的产业化提供帮助。

曙光 4000A 的总体思路是:面向上海超级计算中心的需求, 在深入分析技术发展趋势的基础上,重点对体系结构的选择、节点平台的选择、研制 10 万亿次系统的定位,以及工业标准机群技术增值点布局、开发大规模机群计算的关键技术突破、开发面向网格的技术创新等方面进行综合思考和战略决策,通过与曙光公司的分工协作及与 AMD 公司的战略联盟, 确保与产业紧密结合, 完成曙光 4000A 的各项研制任务。

2 取得的主要成果

2.1 主要技术指标

曙光 4000A 重要技术指标如下:①峰值计算速度:每秒 11.2 万亿次浮点运算,其中 Linpack 值每秒 8.06 万亿次浮点运算; ②节点:640 个 2U4P 节点,2 560 个 64 位 2.2GHz Opteron CPU; ③内存总容量:5TB; ④存储:42.5TB 存储, 含 20TB SCSI RAID 存储; ⑤互联网络:4 套,2Gbps Myrinet 计算网络,1Gbps Ethernet 存储网络,100Mbps Ethernet 管理网络,曙光专用大规模机群管理网络;⑥高速网络性能:在并行通信时单向带宽 494MB/s,单向延迟 6.72us;⑦曙光并行编程环境:BCL4 基本通信库,DPVM4.3.4 曙光 PVM 并行环境,DMPI1.2 曙光 MPI 并行环境;⑧曙光机群文件系统:DCFS2,曙光 IP SAN 网络存储软件; ⑨曙光机群操作系统:

* 收稿日期:2004 年 10 月 18 日

DCMS,DCIS,DCMM,MultiTerm,DSBS（系统管理,系统安装,系统监控,并行操作,作业管理);⑩网格零件:网格路由卡,网格面板,网格钥匙,网格视图,网格网关。

2.1 整体水平

到目前为止,曙光 4000A 已经申请并受理发明专利 17 项,取得的软件登记证书 7 项。

曙光 4000A 在高密度商用服务器主板设计上有突破,在大规模机群管理的软硬件设计方面有独到之处，在面向网格的支持技术方面有明显的特色。系统在大规模机群技术上进入世界领先行列。

曙光 4000A 在 2004 年 6 月的高性能计算机 TOP500 排名中位列全球第十，提高了国产服务器的国际影响。曙光 4000A LINPACK 持续性能达到 8 061GFLOPS,效率为 71.56%,是世界上所有采用 AMD Opteron CPU 的高性能计算机系统中速度最快、效率最高的。

ASCI White 在 NEC“地球模拟器”系统推出之前一直占据 TOP500 第一的位置，从 2001 年到 2003 年 11 月一直是 IBM 安装的最大的计算机系统，曙光 4000A 与 IBM ASCI White 相比在价格(1/10)、Linpack 速度、系统占地(60%)、功耗(1/3)等方面都好于它。

表 1 给出了曙光 4000A 和目前世界上已推出和即将推出的大规模高性能计算机系统在占用空间和功耗方面的具体指标。从中看出,曙光 4000A 无论是空间效益(单位空间所提供的性能)、还是功耗效益(单位功耗所提供的性能)都位列世界前茅,在目前工业标准机群上达到世界领先水平。

另外,通过与 AMD 进行战略合作,普及 64 位计算,在形成我国服务器和高性能计算机产业的良好产业格局上,发挥了有益的作用。

2.2 领先方面

曙光 4000A 使中国成为继美国、日本之后第三个能制造和应用 10 万亿次商用高性能计算机的国家,也表明我国生产、应用、维护高性能计算机的能力达到世界先进水平,得到广泛的国际关注。曙光 4000A 在技术上处于领先水平的独创之处包括:

(1) 高密度商用服务器主板。在工业标准的主板尺寸内实现了 4 个 64 位 Opteron 的 SMP 系统,2U 服务器采用标准的机箱、电源、风扇等部件,有独到的通风设计和部件布局。

(2) 网格部件。通过网格路由器、网格网关、网格钥匙、网格视图等网格部件的研制，使曙光 4000A 在网格环境下能更好地服务于具有多样性的用户需求。

(3) 主板集成的大规模机群管理网络。通过在主板上集成管理接口,开发大规模机群专有的管理网络,使得大规模机群能够被有效地管理、控制和操作,系统管理员不需要走近高性能计算机。

(4) 机群操作系统核心。通过合理地划分机群软件栈(software stack),将公共支撑部分提取成机群操作系统的核心,改变了机群上系统软件缺乏统一框架的情况。

2.3 技术创新点

(1) 面向网格的大规模机群结构。该结构综合了 MPP 系统和传统机群结构的特点，将存储从节点机中分离出来，实现存储（系统存储和数据存

表 1 曙光 4000A 和世界上大规模高性能计算机系统在占用空间和功耗方面的具体指标

系统	峰值速度 (TFLOPS)	占地面积 (平方英尺)	功耗 (KW)	GFLOPS/ 平方英尺	GFLOPS/KW	推出时间
地球模拟器	40	34 000	5 000	1.18	8	2002
ASCIQ	30	43 500	7 100	0.69	4.23	2004
Virginia Tech	17.6	3 000	3 000	5.87	5.87	2003
ASCI Purple	100	43 560	7 500	2.29	13.33	2004
曙光 4000A	11.2	1 528	760	7.37	15	2004

储)、键盘鼠标显示器(KVM)、电源开关控制的网络化和单一系统映像。支持系统环境的动态调整。

(2) 高密度服务器主板。将节点机的密度提高到每 U 高度 2 个 CPU，重点解决了高密度条件下风冷散热问题,同时在服务器主板上集成与管理网络的接口,提供对系统管理的支持。

(3) 高性能系统域网络和并行通信协议。硬件交换机芯片设计采用源址路由方法、缓冲虫洞路由交换和流量控制、PECL 接口连接、双路双沿源时钟同步数据传输;物理带宽达到单向 5Gbps,软件通信协议支持多套网络上的消息分片和并行传输,提高了通信系统的可扩展性和可用性。

(4) 大规模机群管理网络。通过一套网络实现对机群中不同节点的 KVM 切换和电源开关控制、硬件状态信息采集,大大方便了系统管理。

(5) 一体化、构件化、高可用机群操作系统核心。基于分布式构件平台 TAO,实现主要机群系统软件的一体化设计。机群操作系统核心通过元组机制为所有上层系统和应用软件提供高可用基础设施,保证大规模机群系统的高可用性。一体化设计可以降低机群系统软件的复杂性,保证一致性和高效性。

(6) 高可用性的机群文件系统。支持协作式多元数据服务器和基于日志及分布式事务处理机制的高可用性。

(7) 轻核心节点操作系统。针对科学计算类应用的特点,对通用核心中原有的内存管理策略进行优化,保证物理内存的连续性,并在此基础上优化通信协议。这样有助于提高应用的真实性能。

(8) 操作系统动态部署。各节点上取消系统盘,所有节点上的操作系统无需安装,统一从一安有操作系统映像的控制结点进行网络引导,这便于对操作系统映像的管理。同时,在系统运行过程中,可以动态改变部分或全部节点的操作系统配置,调整运行环境,支持系统中同时运行不同的操作系统。

(9) 系统自主管理技术。通过故障节点的自动侦测和重启恢复、系统在线扩展和配置自动调整、资源在不同分区间自动租借、管理策略定制,使系统具有自主管理能力,降低了系统管理的复杂性。

(10) 核心级高速 Socket。通过在操作系统核心中裁减 TCP/IP 协议栈，使得通过 Socket 的通信延迟降低。由于是在核心中裁减,所以能够实现应用程序的二进制代码兼容。

(11) 多种网格零件。支持面向网格的软硬件子系统,包括网格路由器、网格钥匙、网格视图、网格网关,作业管理系统支持通过“织女星网格操作系统”VEGA OS 提交并运行网格作业。这些网格零件对高性能计算机在网格中高效、安全地使用提供了很好的支持。

(12) 面向用户的服务器性能评测方法,实现工作负载的适度定制。

3 应用和产业化情况

曙光 4000A 能同时支持商业应用软件和有源码的应用软件,已成功运行证券指数计算、电力安全评估、建筑工程抗震性评估、天气预报、石油地震资料处理、核能开发利用、汽车碰撞、电磁辐射、计算流体力学、基因匹配与拼接、蛋白质结构分析和材料科学等领域的 20 多项应用。

曙光 4000A 是完全兼容 32 位应用软件的 64 位平台,为商业应用软件的平滑过渡奠定了坚实的基础。Fluent、LS-DYNA、NASTRAN、MSC.MARC、PAM-CRASH、ANSYS、FEKO 等应用软件都可以在曙光 4000A 上进行安装、运行,并实现并行计算。针对 AMD64 的软件版本的计算效率几乎是同类型 32 位软件的 2 倍。

隧道建设是一项投资大、工期紧、施工风险大的工程领域,隧道设计论证、隧道安全性评估等涉及到平方公里量级的计算域，有限元 / 有限差分网格数达到近千万量级,只能在高性能计算机系统上求解。曙光 4000A 良好的兼容性,为满足工程实践对高性能计算的迫切需求创造了条件。曙光 4000A 运用商业化软件 LS-DYNA，求解了一个 400 多万单元的过江隧道抗震安全性评估问题，使用曙光 4000A 的 32 计算节点 128 CPU,22 小时完成计算。中国气象局北京城市气象研究所运行了 MM5 中尺度气象模式，预测 2008 年奥运会 36 小时气象预

报,时间步长 81 秒,4 层嵌套区域面积为 1 661 万、211 万、30 万、4.5 万 km^2, 网格精度相应为 27km、9km、3km、1km。使用曙光 4000A 的 256 个 CPU,可以在 2 小时左右完成计算;使用 800 个 CPU,可以在 1 小时左右完成预报计算。

曙光 4000A 在研制过程中就从科研性样机转换成曙光天潮产品系列,形成批量生产,广泛地用于我国国民经济各个领域,销售势头良好。

曙光 4000A 产生的一批相关的关键技术,已应用于曙光公司 64 位服务器产品中, 并实现 ODM 出口,提高了曙光服务器的技术含量,为促进我国高性能计算机产业的发展发挥重要作用。曙光 4000A 高性能计算机的性能价格比与国际企业的主流系统相比有优势,已经具有一定的市场占有率和用户群,技术上达到国际同类产品的先进水平。

2U4P 服务器的研制成功和投入市场, 标志着我国在商用服务器主板和整机领域取得新突破,使得曙光公司领先 IBM 一年将 4 路服务器推向市场,赢得市场先机。一家美国公司已经和曙光公司签署了该服务器的 ODM 协议, 产品销往美国、欧洲,首批订单的 200 台产品已在天津保税区进行生产。同时,该系统在香港、韩国广受欢迎,并有国际上的多家服务器厂商基于该主板提供服务器产品。

曙光 4000A 是一个稳定可靠、使用方便的超级服务器产品,达到了当前国际先进水平,很好地满足了上海超级计算中心的要求。

另外, 我们认识到 “网格不等于高性能计算机”,未来高性能计算存在两种服务形态:一种是不断研制高性能的超级计算机;另一种是通过网络技术发展网格计算。高性能计算机是网格中的重要计算资源, 网格是高性能计算机系统的良好应用环境,二者是互补而不是替代的关系,采用传统紧耦合结构的高性能计算机是基础,而分布式结构的网格计算将成为主流应用形态。

主要参考文献

1 Zhou X C, Huo Z G, Ma J *et al* . The Parallel Communication Protocol in BCL-4. In Proceeding of HPCAsia, July 2004.

2 Tang R F, Meng D, Wu S N. Optimized Implementation of Extendible Hashing to Support Large File System. Directory2003 IEEE International Conference on Cluster Computing (Cluster2003)Dec.1-4, 2003, Hong Kong.

3 Xiong J, Wu S N, Meng D *et al*. Design and Performance of the Dawning Cluster File System2003 IEEE International Conference on Cluster Computing (Cluster2003)Dec.1-4, 2003, Hong Kong.

4 孙凝晖, 樊建平. Dagger:一种散耦合的网格计算机体系结构. 计算机研究与发展, 2003, (12).

Dawning4000A Superserver

Sun Ninghui

(Institute of computing Technology, CAS, 100080 Beijing)

Dawning4000A is a key project supported by 863 plan of China, and finished by National Research Center for Intelligent Computing Systems (NCIC). Its performance and functionality is in the leading position in the world, especially on high density server designing, massive cluster management technology, and grid router. This article briefly introduces the aim and strategy of developping Dawning4000A superserver, presents the testing and performance evaluation results and outlines the application areas of the superserver.

Keywords dawning 4000A, superserver

孙凝晖 男,计算技术研究所研究员,博士生导师。国家智能计算机研究开发中心主任、曙光公司首席科学家。自 1992 年以来一直从事高性能计算机的研制工作。参加过曙光一号 SMP 系统、曙光 1000MPP 系统、曙光 2000-I、曙光 2000-II,曙光 3000、曙光 4000L,曙光 4000A 机群系统等高性能计算机的研制,从曙光 2000-II 开始担任项目负责人和曙光高性能计算机的总体设计师。多次获国家科技进步奖一、二等奖。