

蚁群:一种新型的高效多用计算平台

谢向辉 钱磊 吴东原 昊李祥

(数学工程与先进计算国家重点实验室 江苏无锡 214125)

(qian.lei@meac-skl.cn)

Ant Cluster: A Novel High-Efficiency Multipurpose Computing Platform

Xie Xianghui, Qian Lei, Wu Dong, Yuan Hao, and Li Xiang

(State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi, Jiangsu 214125)

Abstract Driven by the demands of scientific computing and big data processing, high performance computers in the world have been more powerful and the system scales have been larger than ever before. However, the power consumption of the whole system is becoming a severe bottleneck in the further improvement of performance. In this paper, after analyzing four types of HPC systems deeply, we propose and study two key technologies which include reconfigurable micro server (RMS) technology and cluster constructing technology with the combination of node autonomy and node cooperation. RMS technology provides a new way to make the performance, the power consumption and the size of computing nodes in balance. By combining the node autonomy and the node cooperation, a large amount of small-sized computing nodes can be aggregated to be a scalable RMS cluster. Based on these technologies, we propose a new high-efficiency multipurpose computing platform architecture called Ant Cluster and construct a prototype system which consists of 2,048 low-power ant-like small-sized computing nodes. On this cluster, we implement two actual applications. The test results show that, for real-time large-scale fingerprint matching, single RMS node can achieve 34 times speed-up compared with single Inter Xeon core and the power consumption is only 5W. The whole prototype system supports processing hundreds of queries on a database of 10 million fingerprints in real time. For data sorting, our prototype system achieves 10 times more performance per watt than GPU platform and obtains higher efficiency.

Key words high performance computing (HPC); cluster; reconfigurable computing; micro server; computing platform

摘要 在科学计算和大数据处理应用需求的推动下,高性能计算机的性能不断提升、系统规模日益扩大,系统功耗越来越成为制约能力提升的重要瓶颈.在深入分析现有4类高性能计算机的基础上探讨了2项关键技术:1)可重构微服务器(reconfigurable micro server, RMS)技术.解决单个计算节点在领域应用加速能力、系统功耗和体积间的平衡兼顾问题.2)自治与分治相结合的集群构造技术.解决基于微小型化计算节点的大规模计算平台构造与扩展性问题.在此基础上,提出了一种新型的高效多用计算平台架构——“蚁群”,构建了包含2048个低功耗、微小型化RMS计算节点的蚁群平台原型系统,并实现了大规模指纹实时比对和多RMS节点协同排序2个典型应用.测试表明,单个RMS节点的指纹比对性能是Xeon单核的34倍,功耗仅5W,整个原型系统可实现千万量级指纹库的数百并发实时查询;蚁群平台的数据排序性能功耗比是GPU平台的10倍以上,有效提升数据排序的效率.

关键词 高性能计算; 集群; 可重构计算; 微服务器; 计算平台

中图法分类号 TP391

科学计算是 20 世纪重要的科学技术进步之一, 与理论研究和实验研究相并列, 成为科学研究第 3 种方法. 作为体现国家科学技术核心竞争力的重要标志, 高性能计算技术发展迅速. 根据每年 TOP500 排名统计, 高性能计算机的性能以平均每年约 1~2 倍的速度增长. 然而, 随着系统峰值性能的提升, 尤其到 E 级系统规模时, 其功耗、并发性、可靠性以及可编程性等诸多方面将面临严峻挑战^[1-2]. 系统功耗将成为制约性能提升和规模扩展的最主要的瓶颈. 目前 TOP500 排名第 1 的天河 2 号系统的最大运行功耗达到 17.8MW, 预计未来 E 级系统功耗仍会在数十 MW 量级. 过高的系统功耗将直接影响系统运行的可靠性和运行成本. 同时, 随着互联网技术的发展, 互联网上的数据量呈指数增长. 据统计, 从 2002 年至 2009 年, 互联网数据流量每年增长 56 倍^[3]. Google、Facebook、百度、腾讯等国内外互联网公司不断扩建新的数据中心, 以满足日益增长的数据存储和大数据分析的需求. 然而, 随着数据中心规模的扩大和处理能力的提升, 系统功耗也已成为制约成本控制和业务发展的瓶颈.

面对上述问题, 学术界和工业界在降低功耗开销、提高系统效能方面不断探索新的解决方案. 尤其近年来, 随着 ARM 嵌入式处理器技术的发展, 采用低功耗嵌入式处理器构建高效能计算和数据处理平台逐渐被验证为一种可行的技术解决方案. 本文在长期从事高性能计算机体系结构研究的基础上, 面向领域应用的需求, 以性能、功耗、成本等关键因素为约束, 融合低功耗嵌入式技术和可重构加速技术, 提出了一种全新的高效多用计算平台架构——“蚁群”(下文简称: 蚁群平台), 探讨了 2 项关键技术, 并构建了具有 2048 个节点的原型系统.

1 相关工作

应用需求推动计算平台架构的演化, 其中, 核心计算单元的架构特点和性能决定了计算平台整体的应用适应性和领域应用加速能力. 本文根据核心计算单元类型的不同, 将当前的高性能计算机分为以下 4 种类型:

1) 基于通用多核/众核处理器的高性能计算机. 此类系统以通用的处理器架构、计算模型和编程模型为基础, 通用性强、应用适应面宽、灵活性好. 以 2014 年 11 月 TOP500 排名前 10 的系统为例, 有 5 台系统是基于通用多核处理器的计算机系统, 如美国 IBM 公司的 Sequoia, Mira, JUQUEEN 系统和日本理化研究所的 K 系统等^[4]. 另外 5 台系统则采用通用多核处理器和众核处理器的混合架构, 如国防科学技术大学的天河 2 号系统以及 Cray 公司的 Titan, Piz Daint 系统等^[4]. 目前, TOP500 中的高性能计算机仍以此类系统为主. 然而, 过于强调通用性使此类系统在处理数据密集或计算密集的应用时, 往往存在硬件资源利用率不高、系统效能低的问题.

2) 基于定制芯片的专用计算系统. 如美国 DESRES(D. E. Shaw Research)公司研制的 Anton 系统和日本国家天文台主导的 GRAPE 系统等, 其中, Anton 系统采用了面向分子动力学仿真的 ASIC 芯片, 其加速能力是同时期 BlueGen/L 系统的 100 倍以上^[5]; GRAPE 系统专用于天体物理学中的 N 体问题仿真, 其最新的 GRAPE-8 芯片集成 48 条专用流水线, 理论性能功耗比达到 20.5GFlops/W^[6]. 此类系统的优点是应用算法针对性强、加速性能高. 但由于面向特定算法定制、软硬件系统深度优化、结构固定, 导致应用适应面窄. 此外, 算法与结构强相关, 系统研制难度大、周期长, 一定程度上限制了应用范围.

3) 基于 FPGA 芯片的专用计算系统. 此类系统具有面向单个或多个领域的重构加速能力, 在领域范围内具有一定的应用灵活性和远高于通用计算系统的性能. 典型系统如 PICO 公司研制的 SC 系列 FPGA 集群系统, 在一个 4U 机箱内集成 36 个 FPGA 插件模块, 针对密码学领域的应用, 加速性能是单个 GPU 系统的 500 倍, 是单个双核 CPU 的 7000 倍^[7]. SRC Computers 公司开发的可重构计算平台, 由通用处理器和可重构 MAP 处理器组成, 专用于信号处理、图像处理、石油天然气勘探方面的震波图分析等应用的计算加速^[8]. Convey 公司推出的 HC 系列可重构计算平台, 采用通用 Xeon 处理器和 FPGA 协处理器相结合的异构计算架构, 支持对 x86 指令集的扩展, 被应用于生物信息学、金融分析

算法以及大数据应用的加速^[9]. Maxeler 公司推出的可重构超级计算系统,由通用主处理部件和 FPGA 加速部件 2 部分组成,支持数据流驱动的大规模并行计算,能有效提高片上硅资源的利用率^[10]. 此外,还有安装在美国 Florida 大学 NSF 高性能可重构计算中心(CHREC)的 Novo-G 超级计算机系统,其在 2 个机柜规模的系统中集成了 192 片 FPGA. 系统主要应用于生物信息学领域,与单个 2.4 GHz 的 Opteron 处理器相比,能达到 5 个数量级的加速^[11]. 然而,受限于 FPGA 的结构特点,此类系统仅适合重构简单的定点和逻辑运算,难以适应具有复杂控制流的应用;同时,硬件加速部件的开发门槛高,软硬协同开发难度大等问题也阻碍了该类系统的广泛应用.

4) 近年来出现的基于微服务器的计算平台. 最典型的是欧洲于 2011 年启动的 Mont-Blanc 项目,计划采用嵌入式和移动设备器件设计一种新型的 HPC 架构,实现高能效的超算系统,为实现 E 级计算寻找全新的解决方案^[12]. 此外,学术界也以低功耗嵌入式处理器为基础,构建计算或数据处理平台,如 Chanwit 等人设计了一台由 22 个 ARM 开发板组成低功耗集群系统,部署了支持大数据应用的 Spark 和 HDFS 环境,在处理 34 GB 的维基百科数

据时,响应时间满足需求的同时系统功耗仅 0.061~0.322 kW^[13];Rajovic 等人采用 128 个 Tegra 2 开发板构建了集群计算平台,并完成了 96 节点规模的 HPL 测试^[14]; Balakrishnan^[15], Sukaridhoto^[16], Ou^[17]等团队也分别采用商用的 Pandaboard-ES, Raspberry Pi 开发板,构建小规模 of 的集群原型系统,完成 HPL, STREAM 等 benchmark 的测试,并与 x86 平台进行对比. 从上述研究可以看出,采用低功耗嵌入式处理器,拥有比传统服务器更高的效能,但其绝对计算性能仍难以满足高性能计算应用的需求.

2 蚁群平台架构与原型系统实现

本文针对现实计算需求,提出一种高效多用计算平台架构,它用低功耗的处理器部件紧密耦合可重构的加速部件,集成固态存储器件和多种互连网络接口,并配以标准操作系统、语言及编译系统、管理系统以及完整的应用开发软件包. 系统组成如图 1 所示. 由于极低的功耗、极小的体积、完整的功能、优越的应用性能、良好的扩展性,其节点称为可重构微服务器(reconfigurable micro server, RMS)——“蚂蚁”,集成的计算平台称为“蚁群”.

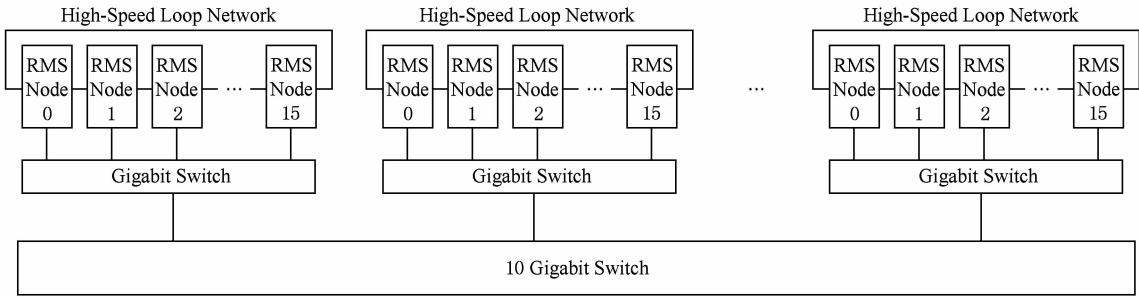


Fig. 1 Architecture of ant cluster.

图 1 蚁群高效多用计算平台架构

2.1 硬件系统

1) RMS 计算节点
蚁群平台包含 2 048 个 RMS 计算节点. 每个计算节点都是一个完整的可重构微服务器. 如图 2 所示,其核心部件是一个混合核心处理器,包含了通用嵌入式计算核心和基于 FPGA 的可重构计算核心. 2 种异构计算资源通过高速的互连总线紧密耦合,可以支持通用计算任务和加速计算任务的并行协同执行. 混合核心处理器外围集成了 1 GB 的低功耗 DDR 内存、32 GB 的 Flash 存储器、千兆以太网接口(GE Interface)、高速环形网接口. 单个计算节点的最高

功耗约为 10 W.

2) 互连网络

蚁群平台的互连网络由多层交换网络和局部高速环网 2 部分组成. 其中,多层交换网络(如图 3 所示)采用基于标准以太网协议的树形网络结构. 该网络共分为 3 个层次,分别为接入层(access layer)、汇聚层(aggregation layer)和核心层(core layer). 接入层为计算节点提供千兆互连接口和局部 16 个节点的千兆全交换;汇聚层实现端口聚合和带宽裁剪,每 192 个节点聚合后共享 20 Gbps 的上行链路,连接到核心层交换机.

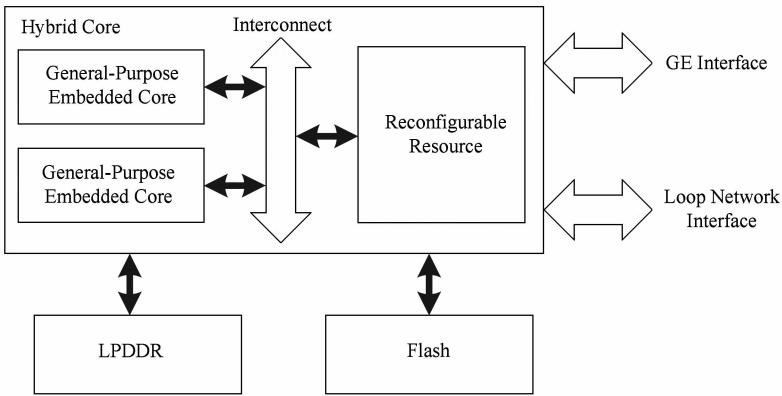


Fig. 2 Structure of RMS computing node.
图2 RMS 计算节点结构

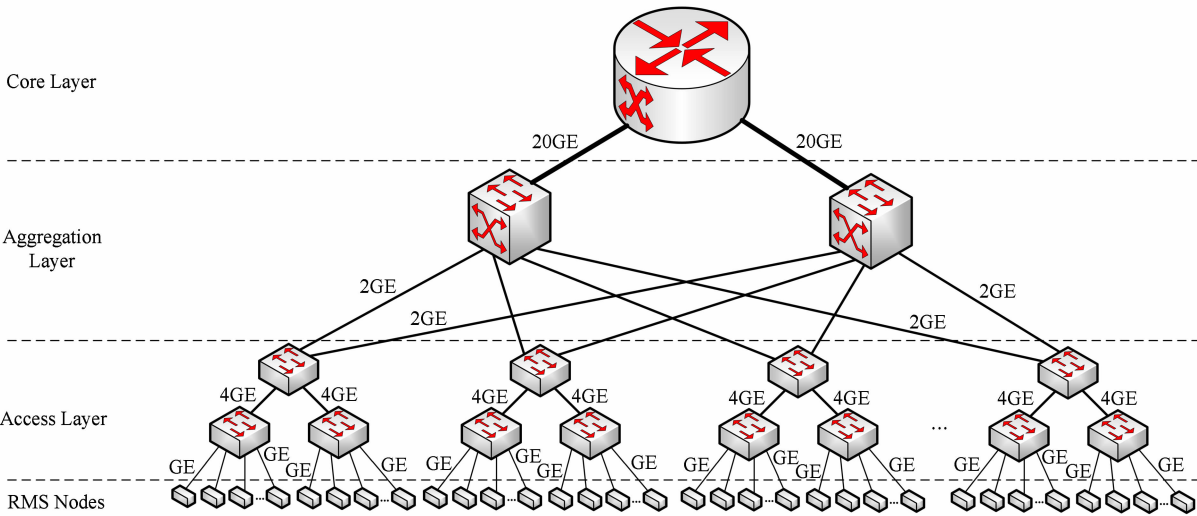


Fig. 3 Architecture of switching network.
图3 交换网络架构

蚁群平台还实现了一套支持节点局部自治的高速环网. 如图 4 所示, 每 16 个计算节点间连接组成一个局部高速环网. 环网直接将相邻节点的可重构计算单元连接在一起, 提供超过 6 Gbps 的高速数据通路, 支持多节点间的高效协同. 基于高速环网, 数据的生成和使用可以不经通用计算核心, 有效降低传输延迟, 提高多节点协同加速计算能力.

每个 RMS 计算节点通过千兆以太网接口连接到多层树形交换网络中, 构成可扩展的分治架构. 同时, 每 16 个 RMS 计算节点为一组, 通过局部高速环网构成自治的宏节点架构, 提供可重构计算核心间的高速数据交换能力, 提升面向领域应用的协同加速计算能力.

蚁群平台原型系统共有 4 096 个主频为 1 GHz

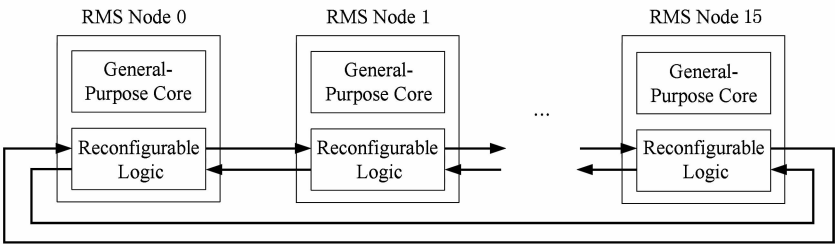


Fig. 4 Structure of loop network.
图4 局部环网结构

的 ARM 通用嵌入式计算核心、2 048 GB 的低功耗 DDR 内存,聚合访存带宽约为 8 TBps. 该系统还包含 64TB 的本地 Flash 存储器,用于操作系统、文件系统和应用数据的局部存储. 在可重构计算资源方面,该系统一共包含了 256×10^6 个可编程逻辑单元,等同于 3.9×10^9 个 ASIC 逻辑门,与 Xilinx 最高端的 Virtex 7 系列 XC7V2000T FPGA 相比,总资源量是其 135 倍. 在网络互连方面,多层交换网络的整系统聚合带宽约为 220 Gbps. 与传统高性能计算机相比,由于单个 RMS 计算节点的运行功耗低于 10 W,即使考虑外围系统开销和电源转换损耗等因素,总的系统功耗也不超过 35 kW.

具体实现时,本文以 DDR2 DIMM 规范作为计算节点的物理结构,在一个 DIMM 结构上集成了 2 个 RMS 计算节点. 16 个节点组装成一块计算刀片,并提供千兆网络接口、背板高速互连和基于高速环网的多节点协同自治能力. 以计算刀片为基础,可以构建大规模的蚁群平台原型系统. 图 5 所示为蚁群平台原型系统实物:

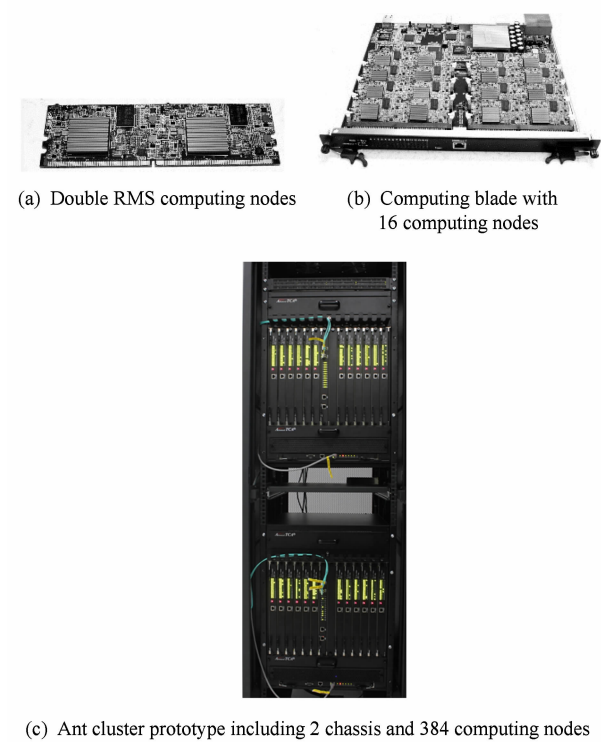


Fig. 5 Prototype of ant cluster system.

图 5 蚁群平台原型系统实物

2.2 软件系统

蚁群平台软件系统包括计算节点系统软件和分布式计算环境.

1) 计算节点系统软件

与 RMS 计算节点的结构特点相适应,本文在

开源的 Linux 系统的基础上对系统软件进行部分定制修改,构建了完整的、面向 RMS 的服务器系统软件栈. 经过定制修改后,RMS 计算节点的系统软件栈既包含了传统服务器系统软件的各种组件,如操作系统内核、文件系统、运行时库、GNU 工具链等,还包含了面向混合核心的定制驱动层,专用于异构核心间的计算协同和高效数据交换. 此外,为了便于构建基于 RMS 计算节点的大规模计算平台,本文在节点系统软件栈中集成了多个重要的功能模块,用于支持大规模的任务分发、调度和管理以及分布式系统资源管理和状态监测等.

2) 分布式计算环境

以 RMS 计算节点为基础构建的大规模并行计算平台具有节点数量多、系统规模大的特点,对分布式计算环境的可扩展性、可靠性提出了更高的要求. 为此,本文以开源的 Boinc 分布式计算平台为基础,构建面向蚁群平台的分布式作业管理系统. 系统不仅具备了任务分发、负载均衡、结果回收等基本功能,而且具有很好的扩展性和容错性,能够在节点出现软硬件故障时通过主动容错机制,将失效节点上的子任务重新分发给正常节点,以确保计算任务的完成. 此外,还构建了一套面向大规模计算平台的系统管理和资源监控系统,既可以实现全局的工作状态监控,也可以对每个单一计算节点进行细粒度管理.

3 蚁群平台关键技术

蚁群平台架构实现的关键在于可重构微服务器技术和自治与分治结合的集群构造技术的突破. 其中,可重构微服务器技术解决单个计算节点在领域应用加速能力、系统功耗和体积间的平衡兼顾问题;自治与分治相结合的集群构造技术解决基于微小型化计算节点的大规模计算平台构造与扩展性问题.

3.1 可重构微服务器(RMS)技术

为解决单个计算节点在领域应用加速计算能力与系统功耗、体积间的平衡兼顾问题,本文提出一种全新的 RMS 技术. 该技术的核心思想在于将低功耗嵌入式处理器与低功耗可重构器件紧密耦合,形成一种混合计算核心的新型异构计算单元. 该计算单元兼顾性能与功耗、性能与体积,具有较高的应用效能;同时集成低功耗内存、非易失固态存储器和多种网络接口,在极小的尺寸和功耗约束下提供完整的服务器软件栈环境和领域多用的加速能力.

基于 RMS 技术,通用微处理器与可重构加速部件直接通过处理器本地的高速数据总线相连,两者映射在统一的物理地址空间中,不仅解决了通用计算部分对可重构计算资源高效访问的难题,还极大地提高可重构加速部件的应用灵活性和适应面,实现对整个系统计算能力、内存能力和 I/O 能力等多方面能力的增强。

1) 计算能力增强. 通过在可重构部件中实现加速扩展指令、硬件线程、硬件加速单元等方式,在指令级、线程级、任务级 3 个逻辑层次实现系统计算能力增强. 由于可重构加速部件与微处理器紧密耦合,实现了一种通专结合、软硬协同的混合计算模型,使算法的复杂控制流逻辑和简单数据流逻辑在处理器和加速部件上合理高效地划分和部署,加速部件的运行频率可有效大幅提升,获得最佳的应用性能和资源利用率。

2) 内存能力增强. RMS 架构扩展了传统的微处理器架构下固定的内存连接和访问模式,通过可重构加速部件连接扩展内存,增加系统的内存容量和带宽,针对实际应用的内存访问模式及物理内存的特性,重构与优化访存控制逻辑,提高内存访问性能和效率,实现总体应用性能的有效提升。

3) I/O 能力增强. RMS 架构下的可重构加速部件可实现外部存储接口,实现数据在线实时压缩和加密等功能. 可重构加速部件实现的网络接口,直接处理网络协议,减轻处理器负载,提高整体的 I/O 处理能力。

3.2 自治与分治结合的集群构造技术

通常的集群构造技术以单一网络来支撑大规模计算节点的互连,系统的可扩展性受互连网络性能、成本和规模的制约. RMS 计算节点体积小,计算能力强,性能功耗比高. 以 RMS 计算节点为基础构建的大规模并行计算平台,节点数量多,系统规模大,单一千兆以太网无法提供与计算能力相匹配的网络数据传输能力,而万兆以太网、Infiniband 等高性能互连技术又存在技术实现复杂、功耗开销大、成本极高的问题. 为解决上述矛盾,本文提出了一种与 RMS 技术相适应的自治与分治相结合的集群构造技术. 通过采用通专结合的混合互连网络,使得蚁群平台在兼具高性能和低成本的同时具有良好的规模扩展性。

支持计算节点分治的标准互连网络和多节点协同自治的局部高速环网. 标准互连网络支撑大规模节点分治,网络中的每个计算节点独立而对等地参与

到计算任务执行和系统管理中,网络架构自身具有良好的伸缩性,支持计算平台规模的扩展. 局部高速环网则针对复杂领域应用加速需求,通过连接邻近的多个节点,构成可重构宏节点. 宏节点内部由应用自主发起和完成数据交换,协同实现高难复杂计算任务. 通过应用的自治与系统管理的分治,蚁群平台在高性能、低成本及高可扩展之间获得了最佳折中。

节点分治采用了基于千兆/万兆以太网协议的标准互连网络,数据传输带宽与计算节点的通用计算能力相匹配,具有技术成熟度高、成本低,兼容性好、支持协议完整及可扩展性强等技术优点,适合超大规模蚁群平台的构建. 该网络为各计算节点提供基础的数据交换通路,能支持大规模计算平台管理和通用计算任务分发执行. 但由于网络传输带宽较低、协议开销较大,使得计算节点的加速计算能力难以得到充分发挥。

基于局部性原理,本文实现了一套面向可重构加速部件的局部高速双向环网,连接邻近多个 RMS 计算节点的可重构加速部件. 该环网双向最高传输带宽为 6.6 Gbps,与节点的可重构逻辑资源规模相匹配. 网络接口协议和控制模块可根据应用的通信特点进行优化定制,降低传输延迟. 通过实现局部高速环网,支持多节点协同自治,形成多种形态的可重构宏节点,有效扩展了 RMS 计算节点的计算、存储和 I/O 资源。

本文利用该项技术实现了逻辑资源扩展和数据带宽增强这 2 种形态的可重构宏节点. 逻辑资源扩展型宏节点有效支撑了大数据分析、生物特征计算等所需逻辑资源超过单个 RMS 计算节点的复杂应用. 数据带宽增强型宏节点可以聚合内部多个节点数据通路,实现协同的数据加载,并通过局部高速互连网络实现节点间的数据分享和协同处理. 蚁群平台上实现的大规模数据集并行排序,性能功耗比是通用 CPU 和 GPU 平台的数十倍。

在构造大规模蚁群平台时,面对数千乃至数万 RMS 计算节点,系统管理和组装面临巨大挑战. 本文采用了一种基于空间坐标的集群网络自动配置及管理方法,以软硬结合的方式,根据每个节点在结构组装层次中的物理位置自动推算网络地址,消除手工编写网络地址与物理位置对应表的繁琐工作,实现节点的自适应配置、快速准确定位和电源管理. 在系统结构组装方面,采用模块化三维结构构造技术,使计算节点可以作为插件模块,以水平和垂直堆叠方式组装到背板上,最大限度利用组装空间,提高组装密度。

4 典型应用实现和分析

在蚁群平台原型系统上,针对不同应用领域进行探索,本文实现了大规模指纹比对、多层次混合排序应用,并与 CPU,GPU 等传统计算平台进行对比分析.

4.1 大规模指纹实时比对

指纹识别技术是当今应用最为广泛的生物特征识别技术,尤其在公共安全领域,指纹采样的人群范围广,指纹库规模大,对指纹比对的实时性要求高,传统通用服务器已经越来越难以满足要求.为此,本文在蚁群平台的基础上,研究并实现了一种软硬结合的

指纹快速比对算法,支持大规模指纹库上实时比对. 目前,最常用的一类指纹比对算法是基于指纹特征的点模式匹配算法.该类算法不仅包含大量的复杂操作,如平方根和超越函数的计算、指纹特征点集坐标平移和旋转等,而且算法控制流复杂、离散访存多,在通用计算平台上难以实现有效优化.本文基于 RMS 计算节点混合核心架构的特点,采用软硬协同的方法,将算法中指纹比对的计算密集部分从复杂控制流中分割出来,利用可重构逻辑实现指纹特征计算和比对过程的加速,而对于特征点集的旋转、平移这类控制流复杂、离散访存多的计算部分以及软硬件同步和数据传输控制部分,则交给通用计算核心处理.基于 RMS 的指纹比对算法结构如图 6 所示:

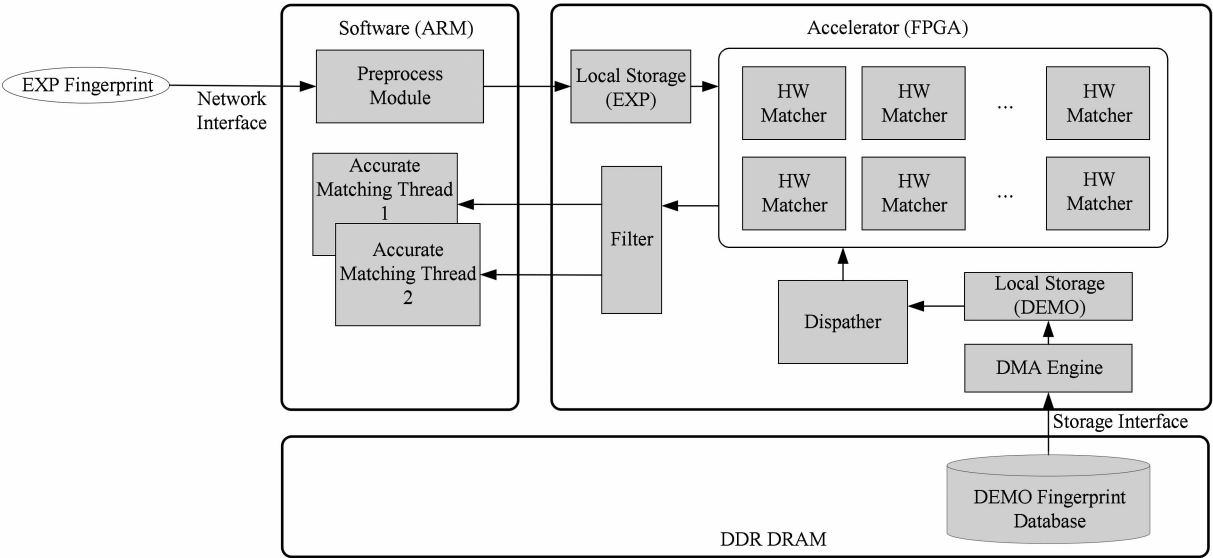


Fig. 6 Structure of RMS-based fingerprint matching algorithm. 图 6 基于 RMS 的指纹比对算法结构

在算法实现的基础上,本文将单个 RMS 计算节点上的指纹比对性能与通用 x86 服务器平台上的串行算法进行了对比分析.其中,RMS 计算节点的通用计算核心主频 1 GHz,内存 1 GB,指纹比对加速部件工作频率在 200 MHz;通用 x86 服务器采用 Intel Xeon E5620 处理器,主频 2.4 GHz,内存 40 GB.对比测试采用相同的指纹库作为输入,库中每个指纹的特征点个数在 20~45 个之间,完成整轮测试所执行的总指纹比对次数约为 223 万次.软件程序均采用-O3 优化.2 个平台的性能测试结果如表 1 所示.

从计算性能上看,单个 RMS 计算节点约为 Xeon 服务器上单核串行性能的 34 倍;但 RMS 计算节点在指纹比对过程中的实测功耗仅 5 W,其性能功耗比比通用服务器高出 2 个数量级.

Table 1 Comparison of Fingerprint Matching Performance 表 1 指纹匹配算法性能结果比较

Platform	Running Time/s	Performance/ (match/s)
RMS Node	2.12	1 050 000
Xeon Server (Single Core)	73.1	30 500

在大规模指纹比对应用时,本文采用了前端应用服务器和后端大规模蚁群平台相结合的架构.前端应用服务器处理来自网络并发指纹查询请求,并将请求分发给蚁群平台;蚁群平台作为后端处理系统,每个 RMS 计算节点独立响应请求,在局部指纹数据库上快速完成比对,并将比对结果返回给前端

服务器. 目前, 蚁群平台原型系统可以支持千万量级指纹库的数百并发实时查询.

4.2 多 RMS 节点协同排序

蚁群平台是顺应大数据应用发展趋势的新型计算平台架构, 具有高性能的数据处理能力和计算加速能力, 更具备并行排序算法所需的节点间数据通信能力, 能很好适应大规模数据排序问题. 基于蚁群平台的宏节点自治架构, 本文还提出并实现了一种多 RMS 节点协同的多层次混合排序算法.

本文通过对当前排序算法进行总结, 分析了各类算法的硬件化特性, 提出了将插入排序、FIFO 归并、归并树、RMS 计算节点间归并相结合的层次化

排序结构, 并有效改进了各层次排序算法. 图 7 所示为多节点协同排序硬件结构, 其核心思想是在每个 RMS 计算节点上, 通过有效利用插入排序流水线资源, 得到无空闲周期的部分有序流; 然后通过多个节点层层归并, 得到最终有序数据. 本文还在算法硬件实现上进行优化, 如通过提高 FIFO 归并结构存储利用率, 扩大了可排序规模; 通过拆分归并树排序中的控制依赖, 提高排序结构吞吐率等. 在多 RMS 节点协同上, 通过利用宏节点内的高速环网, 使多个 RMS 计算节点的可重构逻辑部分自主实现片间归并流水线, 从而实现在较大规模数据集上的并行排序.

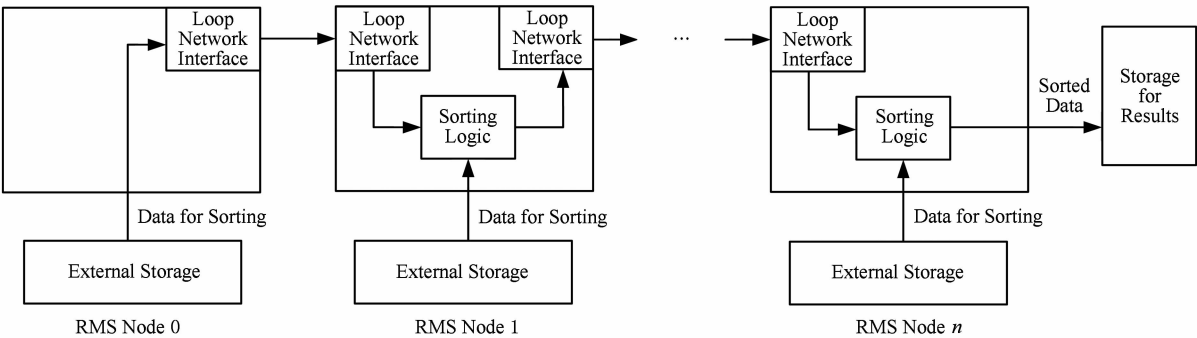


Fig. 7 Hardware structure of sorting algorithm based on node cooperation.

图 7 多节点协同排序硬件结构

通过测试, 如图 8 所示, 单个 RMS 计算节点上, 453.75 MB 数据可在 859.5 ms 内完成排序, 吞吐率可达 527.9 MBps; 2 个节点协同进行并行归并, 可在 2144.4 ms 内完成 907.5 MB 数据排序, 吞吐率达 423.2 MBps; 16 节点构成的宏节点内, 利用高速环网, 7.09 GB 数据可在 11072.8 ms 内完成排序, 吞吐率达 655.7 MBps. 蚁群节点在排序数据规模、吞吐率上达到良好的性能指标, 而单节点功耗仅

为 10 W, 其性能功耗比具有显著优势.

与通用 CPU 相比, 单个节点的性能功耗比有近 10 倍提升; 与 Cell 处理器相比, 无论是排序规模、吞吐率还是性能功耗比都有巨大优势; 与 GPU 相比, 蚁群平台以很小的内存开销实现了较大数据规模排序, 吞吐率在 2 倍差距之内, 考虑到功耗因素, 蚁群节点性能功耗比相对 GPU 有 10 倍以上提升.

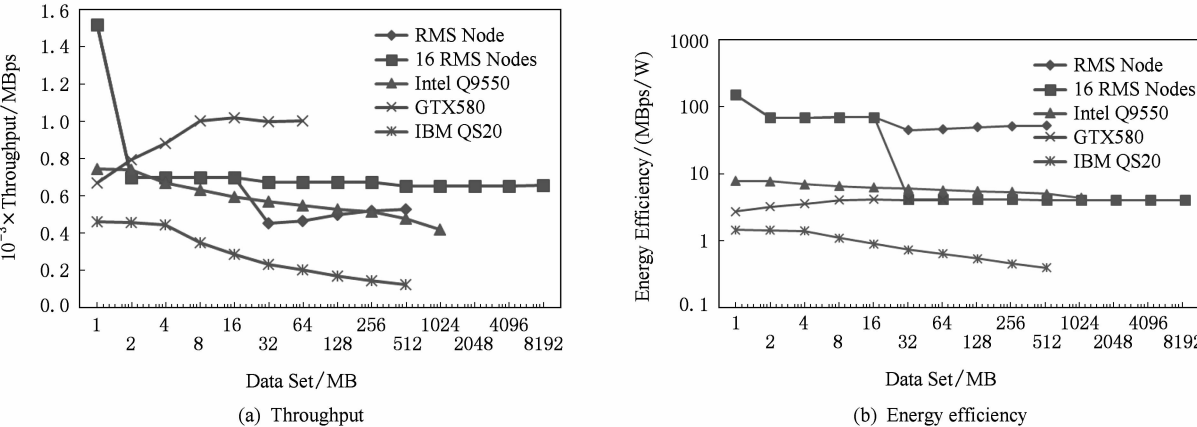


Fig. 8 Comparison of throughput and energy efficiency among different computing platforms.

图 8 在多种计算平台上排序算法的吞吐率和能效对比

5 总 结

面向传统高性能计算机存在的问题和挑战,本文研究了2项关键技术:1)可重构微服务器技术,解决了单个计算节点在领域应用加速能力、系统功耗和体积间的平衡兼顾问题。2)自治与分治相结合的集群构造技术,解决了基于微小型化计算节点的大规模计算平台构造与扩展性问题。在此基础上,提出一种新型的蚁群高效多用计算平台架构,实现了包含2048个低功耗、微小型化 RMS 计算节点的蚁群平台原型系统。

基于蚁群平台实现的多个典型应用获得高效的结果。其中,指纹比对算法在单个 RMS 节点上的计算性能是 Xeon 单核的 34 倍且功耗仅 5W,整个原型系统可实现千万量级指纹库的数百并发实时查询;基于宏节点自治架构实现的多 RMS 节点协同混合排序,利用宏节点内高速环网,支持较大规模数据集上的并行排序,与 CPU 和 GPU 平台相比性能功耗比均在 10 倍以上。

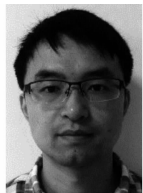
作为一种新型的高效多用计算平台,蚁群平台原型系统已经推广应用到多个应用领域,体现出良好的应用适应性;但同时,蚁群计算平台也存在一些局限性。如在多精度浮点计算方面,存在 ARM 处理器的多精度浮点能力较弱、FPGA 实现多精度浮点运算资源开销大的问题,限制了平台的通用性。本文后续也会在逻辑高层综合技术、混合核心的指令集扩展技术等高效应用开发技术方面展开深入研究,进一步提高蚁群平台的应用开发效率。

参 考 文 献

- [1] Kogge P, Bergman K, Borkar S, et al. Exascale computing study: Technology challenges in achieving exascale systems, TR-2008-13 [R]. South Bend, Indiana, USA: Department of Computer Science and Engineering, University of Notre Dame, 2008
- [2] Simon H D. Barriers to exascale computing [G] //High Performance Computing for Computational Science(VECPAR 2012). Berlin: Springer, 2013: 1-3
- [3] Chong F T, Heck M J R, Ranganathan P, et al. Data center energy efficiency: Improving energy efficiency in data centers beyond technology scaling [J]. IEEE Design & Test of Computers, 2014, 31(1): 93-104
- [4] TOP500 Orgnization. HPC Top500 list November 2014[EB/OL]. (2014-11-05) [2015-04-27]. <http://www.top500.org/lists/2014/11>
- [5] Shaw D E, Deneroff M M, Dror R O, et al. Anton, a special-purpose machine for molecular dynamics simulation [J]. Communications of the ACM, 2008, 51(7): 91-97
- [6] Makino J, Daisaka H. GRAPE-8—An accelerator for gravitational N-body simulation with 20.5 GFlops/W performance [C] //Proc of 2012 Conf on High Performance Computing, Networking, Storage and Analysis (SC). Piscataway, NJ: IEEE, 2012: 1-10
- [7] Pico Computing, Inc. HPC systems [EB/OL]. (2014-08-20) [2015-04-27]. <http://picocomputing.com/products/hpc-systems>
- [8] SRC Computers, LLC. SRC reconfigurable MAP processors[EB/OL]. (2015-01-10) [2015-04-27]. <http://www.srccomputers.com/reconfigurable-map-processors>
- [9] Nagar K K, Bakos J D. A sparse matrix personality for the convey HC-1 [C] //Proc of the 19th IEEE Annual Int Symp on Field-Programmable Custom Computing Machines (FCCM). Piscataway, NJ: IEEE, 2011: 1-8
- [10] Pell O, Mencer O. Surviving the end of frequency scaling with reconfigurable dataflow computing [J]. ACM SIGARCH Computer Architecture News, 2011, 39(4): 60-65
- [11] George A, Lam H, Stitt G. Novo-G: At the forefront of scalable reconfigurable supercomputing [J]. Computing in Science and Engineering, 2011, 13(1): 82-86
- [12] Puzovic N. Mont-Blanc: Towards energy-efficient HPC systems [C] //Proc of the 9th Conf on Computing Frontiers. New York: ACM, 2012: 307-308
- [13] Kaewkasi C, Srisuruk W. A study of big data processing constraints on a low-power Hadoop cluster [C] //Proc of 2014 Int Computer Science and Engineering Conference (ICSEC). Piscataway, NJ: IEEE, 2014: 267-272
- [14] NRajovic N, Rico A, Vipond J, et al. Experiences with mobile processors for energy efficient HPC [C] //Proc of the Conf on Design, Automation and Test in Europe. EDA Consortium. New York: ACM, 2013: 464-468
- [15] Balakrishnan N. Building and benchmarking a low power ARM cluster [D]. Edinburgh, UK: University of Edinburgh, 2012
- [16] Sukaridhoto S, KHalilullah A, Pramadihato D. Further investigation of building and benchmarking a low power embedded cluster for education [C/OL] //Proc of Int Seminar on Applied Technology, Science and Art. (2013-12-10) [2015-01-15]. https://lecturer.eepis-its.edu/~dphoto/publications/conference/2013/22-APTECS2013-surabaya/Dphoto-Bejo-Dadet_Full-Paper-APTECS-2013.pdf
- [17] Ou Z, Pang B, Deng Y, et al. Energy-and cost-efficiency analysis of arm-based clusters [C] //Proc of the 12th IEEE/ACM Int Symp on Cluster, Cloud and Grid Computing (CCGrid). Piscataway, NJ: IEEE, 2012: 115-123



Xie Xianghui, born in 1958. PhD supervisor and professor. Senior member of China Computer Federation. His current research interests include high performance computer architecture, reconfigurable computing and distributed computing.



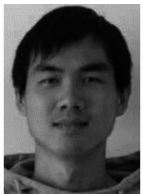
Qian Lei, born in 1982. PhD candidate and engineer. His current research interests include high performance computer architecture and reconfigurable computing.



Yuan Hao, born in 1984. Master and engineer. His current research interests include computer architecture and embedded system.



Wu Dong, born in 1972. PhD and senior engineer. His current research interests include high performance computer architecture and reconfigurable computing.



Li Xiang, born in 1990. Master and engineer. His current research interests include computer architecture and network.

2015 年《计算机研究与发展》网络与信息安全专辑(正刊)征文通知

——“网络安全与隐私保护研究进展”

近年来,随着计算机网络的不断发展,数据量的井喷式增长催生了大量新型网络应用,随之产生的网络安全问题也更加突出. 网络安全除了网络密码安全、网络系统安全外,由于云计算、大数据等新型应用模式需求,产生了大量的安全与隐私保护问题,这些问题的研究不仅关乎国家安全、经济发展和社会稳定,也关乎网络安全学科本身的进步. 为推动我国学者在这方面的研究、及时报道我国学者的最新研究成果,《计算机研究与发展》将于 2015 年 10 月出版网络与信息安全专辑——网络安全与隐私保护研究进展,欢迎相关领域的专家学者和科研人员踊跃投稿.

征文内容

本专辑包括(但不限于)下列网络安全与隐私保护的相关主题:

- 1) 网络密码安全(面向网络环境的密码算法、协议及应用密码技术解决网络安全问题);
- 2) 网络系统安全(网络架构安全、网络空间安全、网络设备安全);
- 3) 新型应用模式下的安全与隐私保护(云计算、大数据等新型应用模式下的安全与隐私保护).

投稿要求

- 1) 论文应属于作者的科研成果,数据真实可靠,具有重要的学术价值与推广应用价值,未在国内外公开发行的刊物或会议上发表或宣读过,不存在一稿多投问题. 作者在投稿时,需向编辑部提交投稿声明.
- 2) 论文应包括题目、作者信息、摘要、关键词、正文和参考文献,论文一律用 word 排版,论文格式体例格式请参考《计算机研究与发展》近期文章.
- 3) 论文需附通信作者的联系地址、电话或手机及 E-mail 地址.
- 4) 论文请通过期刊网站(<http://crad.ict.ac.cn>)进行投稿,并在留言中注明“网络与信息安全 2015 专题”. (否则按自由来稿处理)

重要日期

征文截止日期: 2015 年 6 月 15 日

录用通知日期: 2015 年 7 月 25 日

作者修改稿提交日期: 2015 年 8 月 25 日

出版日期: 2015 年 10 月

特邀编委

徐秋亮 教授 山东大学 xql@sdu.edu.cn

张玉清 教授 中国科学院大学、国家计算机网络入侵防范中心 zhangyq@ucas.ac.cn

董晓蕾 教授 华东师范大学 dongxiaolei@sei.ecnu.edu.cn

联系方式

联系人: 徐秋亮 xql@sdu.edu.cn 13969133581

编辑部 crad@ict.ac.cn 010-62620696, 010-62600350

通信地址: 北京 2704 信箱《计算机研究与发展》编辑部

邮政编码: 100190