新型高性能计算系统与技术

廖湘科*, 肖侖

国防科学技术大学计算机学院, 长沙 410000

* 通信作者. E-mail: xkliao@263.net

收稿日期: 2016-06-26; 接受日期: 2016-08-25; 网络出版日期: 2016-09-18

国家自然科学基金 (批准号: 61433019, 61402503, 61170288, 61332003) 资助项目

摘要 高性能计算技术是信息时代世界各国特别是发达国家激烈竞争的技术制高点. 本文针对未来新型高性能计算技术的挑战, 从微处理器、高性能计算机系统、高性能并行算法与应用等多个核心技术来分析高性能计算技术的国内外发展现状、趋势与挑战, 从基础研究的角度阐述我国未来高性能计算系统理论与技术领域需重点开展研究的领域 (或科学问题) 与相关政策建议.

关键词 高性能计算 微处理器 计算机体系结构 高性能计算应用

1 引言

高性能计算技术是国家战略高科技技术, 是解决经济建设、社会发展、科学进步、国家安全一系列重大挑战性问题的重要手段, 已经成为信息时代世界各国的战略技术制高点. 今天我国经济社会发展和国家安全对高性能计算具有迫切需求. 如解决能源短缺、环境污染、全球气候变化等重大挑战问题迫切需求高性能计算; 传统产业转型升级对以高性能计算为手段的模拟与优化设计有广泛需求; 个性化精准医疗、生物信息、国家突发恶性传染病预警迫切需要高性能计算支持; 以航空航天技术等为支撑的国家安全和战略领域离不开高性能计算技术支撑等. 高性能计算机正在国民经济各个领域发挥着不可替代的作用, 已经成为研究和解决各领域挑战性问题的重要手段.

目前, 美国、日本、欧洲等发达国家都在不断研制高性能计算机, 机器运算速度屡创记录^[1]. 美国作为高性能计算技术领头羊在使能技术储备、基础理论与方法、工程技术和应用等多个方面均处于世界领先地位. 尤其在 10/7 纳米先进半导体集成电路工艺、采用 3D 堆叠技术的混合存储立方体 (HMC)、硅基光互连技术方面具有巨大领先优势, 超级计算机研制能力领先; 另一方面美国超级计算机应用领域范围广, 应用开发能力强且生态环境成熟. 日本在高性能计算机研制和应用方面有着自己的发展规划和特色, 开发了结构独特的“地球模拟器”和“京”等多个超级计算机系统, 具有高性能计算应用的广度和深度. 欧洲尽管在高性能计算机研制方面已经落后于美国、中国和日本, 但在高性能计算应用方面在国际上首屈一指, 在计算模型、并行算法和并行应用等方面拥有相当多专利和知识产权.

引用格式: 廖湘科, 肖侖. 新型高性能计算系统与技术. 中国科学: 信息科学, 2016, 46: 1175–1210, doi: 10.1360/N112016-00147

我国在基金委、科技部的长期支持下,近年来在基础理论与方法、工程设计和制造技术、机器研制及人才培养等方面有了一个质的飞跃. 高性能计算机研发方面突破了异构协同并行计算体系结构、高性能处理器、高速互连、高密度组装、高效冷却、系统可靠性、并行系统软件等关键技术,成功研制了天河、神威、曙光、深腾等高性能计算机. 特别是自 2010 年以来,天河连续在世界超级计算机排行榜 TOP500 中七次占据第一,引起了国际高度关注;2016 年 6 月,采用国产处理器实现的“神威太湖之光”荣获 TOP500 第一,取得了以国产处理器研制万万亿次超级计算机的重要突破. 在世界超级计算机 HPCG 排行榜中(高性能共轭梯度, HPC 领域新的候选基准测试),天河二号仍旧高居榜首,获得“五连冠”. 尽管目前美国政府对高性能计算技术采取了新的限售措施,但中华人民共和国科学技术部国家高技术研究发展计划重大项目正开展的十亿亿次级高效能计算机研制进展顺利,以自主 CPU、互连、系统软件为基础的“十二五”二台十亿亿次高性能计算机系统目标仍将按期实现.

中国高性能计算技术异军突起,给世界高性能计算研究带来了一股新的力量. 为了应对高性能计算领域激烈竞争,世界主要发达国家已经开始更快计算机的各项研制计划. 2015 年美国发布“创建美国国家战略性计算计划 NSCT”总统令,旨在推动美国高性能计算效益最大化. 美国政府 DOD, DOE, NSA, NSF 等部门参与了高性能计算投资规划和应用引导, Intel, Nvidia, IBM, Cray 等企业已经开展了关键技术研制工作. 欧洲通过欧盟整体实施了 HPC 政策和资助项目,参与国家有德国、法国、英国等. 日本则由文部省规划,富士通公司、中国科学院理化技术研究所等参与研究. 俄罗斯由联邦原子能署牵头组织研究. 各国预期在 2020 年前后研制出 E 级超级计算系统. 我国科技部已开始启动 E 级超级计算机系统计划. 国家自然科学基金委已围绕着 E 级高性能计算技术发展在基础理论方法、技术和应用方面作出了战略规划和部署.

未来 E 级高性能计算机研制和应用面临着巨大挑战^[2]. 随着系统规模扩大,仅依靠已有传统技术来使高性能计算机性能大幅增长的做法不再可行;高性能计算机研制面临能耗墙、存储墙、编程墙、通信墙、可靠性墙 5 大问题的巨大挑战;高性能计算机应用也面临计算模型、大规模并行算法框架等严峻挑战. 与此同时,大数据处理、云计算、认知计算需求日益旺盛,这种巨大需求对传统高性能计算机技术提出了新的需求和技术挑战,如大数据处理应用需要高效支撑访存和 I/O 密集型的系统结构等. 针对上述未来新型高性能计算技术挑战,本文将从处理器体系结构、高性能计算机系统、并行算法与应用等多方面入手分析高性能计算技术国内外发展现状、困难和发展趋势. 论文的第 4 节将结合世界高性能计算技术发展和我国在这个领域发展的特殊需求和特点,从基础研究角度入手陈述高性能计算系统理论与技术领域亟需开展的科学问题与相关政策建议.

2 国内外研究现状

高性能计算关键技术涉及微处理器结构、高性能计算系统结构、存储结构、编程模型、系统软件、并行算法与应用软件等多个方面,本节将从研究与工程结合角度分析微处理器、高性能计算机系统、大规模并行算法等国内外研究现状.

2.1 微处理器体系结构

为了给更高性能计算系统提供计算能力的承载实体,研究人员从效能、实用性能、可靠性等方面不断推动着微处理器技术的持续发展. 在半导体领域,随着 MOS FET 工艺下登纳德缩放比例定律停滞,学术界工业界都在试图延续传统摩尔定律. 主要技术包括传统高性能通用处理器,受到领域专用

处理器影响的加速处理器, 以及异构混合高性能处理器等. 另外还包括能彻底突破功耗墙、访存墙或性能墙障碍的碳纳米管、低温超导量子忆阻器技术, 以及基于概率、认知等新型计算模型建立的新型处理器等.

2.1.1 通用高性能处理器体系结构

美国能源部从 2008 年开始联合工业界与学术界讨论 E 级高性能计算机系统, 美国国防部 DARPA 从 2009 年开始以泛在高性能计算 UHPC 项目为 E 级超级计算机进行预研. 针对 E 级 HPC 系统的需求, 目前微处理器领域研究主要集中在低功耗计算核心、众核处理器结构、异构加速计算核心、可靠性与可恢复性逻辑设计与软硬件协同设计等多个方面.

(1) 工艺支持两代以上高性能“众核”处理器可持续发展. 在新型器件和新计算模型处理器从实验室进入大规模商业量产之前, 传统众核处理器是未来 5 年高性能计算系统发展重点. 近年来, 超级计算机主处理器单芯片集成的核心数量稳步提升, 使用众核处理器的机器比例逐年增加. 2015 年 11 月发布的全球超级计算机 Top500 排行榜中, 最快的十台机器中已有 4 台使用众核处理器^[1]. Intel, IBM, nVidia 以及 AMD 等美国公司研制了多款众核处理器. 如 Intel 于 2015 年 6 月发布的最新 MIC 处理器 (Knight Landing) 采用 14 nm 工艺, 集成 72 个核心, 峰值性能达到 3TFlops^[3]. nVidia 公司 2016 年面向高性能计算的 Pascal 架构 GPGPU 产品 GP100, 双精度浮点性能达到 5.3TFlops^[4]. Intel 在 2016 的 Xeon 处理器核心数量也突破 22 核. 国防科学技术大学的飞腾高性能通用微处理器已经突破单芯片集成 64 核的目标.

目前, 最新量产的高性能通用众核处理器工艺为 14/16 nm. 按照 ITRS 预测, 10/7 nm 工艺在未来 5 年可以继续驱动众核核心数量提升. 由于封装、散热和功耗等实际技术工程障碍, 高性能众核处理器低功耗实现将成为关注的重点. 依托先进生产工艺, 在同构核心数量持续增加的情况下如何在结构设计中避免或减小工艺偏差与变化将变得更加重要. 面对 HPC 大规模部署和持续运行压力, 众核处理器的全局可靠性、热管理、动态任务迁移机制也是未来研究重点.

(2) 访存压力进一步恶化, 更多存储层次、非易失存储以及近数据处理吸引力增加. 众核体系结构设计趋势并未改善一直存在的计算存储鸿沟, 反而促使其进一步恶化. 当前高性能通用处理器普遍将片上高速缓存从 2 级提升到 3 级, 仍无法有效缓解存储墙问题. 工业界学术界已着手准备使用 DRAM 取代 SRAM 作为片上最外一级缓存的器件, 以进一步增加片上缓存容量^[5]. 除了已进行多年的如何在多个线程之间有效划分缓存容量之外, 如何缩减 DRAM Cache 由于刷新和数据不稳定带来的性能和功耗开销成为研究重点^[6,7]. 在核心数量持续增加的情况下, 传统缓存相关性模型与结构的实现开销也会成为瓶颈. 除了放松缓存一致性协议降低实现开销外, 也有以 MIT 基于时间标记的 Tardis 协议为代表的创新一致性方法^[8], 以更低开销支持 1000+ 以上核心互连的缓存一致性.

有研究表明处理器与高计算机系统结构中, 将数据从存储器移动到处理单元是能量耗费的关键. 同时对于越来越受到研究关注的并行型应用而言, 系统性能受限于从存储到处理器的通信带宽. 对于数据中心, 特别是进行巨量大数据运算的分析类应用, 采用低功耗非易失存储器件, 并将数据处理尽可能的拉近数据存储成为新的研究热点. 在三维封装以及兼容逻辑的存储生产工艺等技术支持下, 20 世纪的存储中计算 (PIM) 技术转变为近数据处理 (NDP) 技术^[9]. NDP 技术面临 PIM 技术类似的技术和实践问题, 即如何应用新型存储增加近数据的轻量计算能力, 保持当前存储计算模型的兼容性和正确性. 在这些问题得到解决之前, 受到应用开发限制, NDP 或仅作为一个专用领域解决方案, 或作为传统存储层次的一个加速解决方案.

(3) 片上与片间互连网络技术持续集成与融合. 片上网络已成为微处理器体系结构研究的热点

和难点问题。片上网络是处理器计算核访问存储部件的通路,也是多个计算核协同工作的基础,片上网络性能直接决定了系统性能。与片外网络不同,片上网络与计算核和存储部件竞争宝贵的芯片面积和功耗资源。因此片上网络设计必须具备较低硬件开销。围绕提升系统性能并降低硬件开销这一目标,国内外学者对片上网络设计空间各个要素开展了大量研究^[10]。片上网络拓扑结构研究围绕降低传输延迟和提高吞吐率展开,出现了 Flatten Butterfly 和 Multi-drop Express Channels 等高阶拓扑结构^[11,12],充分利用了片上丰富的连线资源。在路由器微结构设计方面,人们提出了使用猜测方式并行执行虚通道分配和交叉开关分配,减少路由器流水线级数。路由算法设计需要在保证网络无死锁的前提下尽量支持较高性能流控机制,直接影响到网络开销和性能。无缓存流控不需要在路由器上部署缓存有效地降低了硬件开销。

(4) 低功耗与可靠性设计仍是重要设计约束。 能量效率成为所有类型和尺寸的计算机系统设计中的关注点;可靠性设计受大规模系统部署时软错误影响在高性能处理器设计中也具有不可替代作用。由于较低的功耗开销,向量化、持续增宽的 SIMD 设计成为微处理核高效提升性能的有效方法。Intel 最新一代 SIMD 设计 AVX-512 已经在支持浮点/定点乘加运算的技术上将单个功能部件的数据宽度增加到了 512 位宽、8 个长整数或双精度向量^[13]。除微体系结构方面研究之外,低功耗与高可靠处理器的设计还需要系统级设计方法学和 EDA 工具^[14],学术界和工业界正在研究如何支持具有几十亿晶体管的片上系统的设计和验证^[15];研究如何解决片上系统的功耗和热的问题;如何在设计描述中考虑互连线延迟、噪声等纳米工艺电学特性;如何描述和考虑纳米工艺制造的物理效应和不确定性的影响;如何表示和描述可复用的 IP 模块;如何实现系统验证和测试等。工艺制造能力和设计方法、设计工具间的差距如不解决,会严重地限制 SoC 的进一步发展。国际上的 EDA 研究主要集中在面向 SoC 系统复杂性和纳米工艺的物理效应与可制造型方面,并在 SoC 的系统集成方法学、系统综合、系统验证与测试理论等 EDA 工具开发方面开展了全面、系统、深入的研究。

2.1.2 专用与领域特定的微处理器体系结构

在低功耗与高能效压力下,专用加速器硬件成为最具吸引力的高性能处理部件。但选择专用加速器通常意味着部分丧失通用编程特性,导致应用开发、调试和维护上额外工作量。同时对领域或者应用专用的加速器也受制于应用的快速演进和领域潜在变化,通常会带来重复设计和验证开销。面对先进工艺下大尺寸集成电路设计、逻辑设计、验证、测试和投产的技术、人力、物力和时间成本开销,即使在高性能计算领域,硬件专用加速器也不得不在通用性上妥协。

(1) GPGPU 为代表的专用处理器通用化结构趋势不减。 由于在 GPU 上实现通用计算是一个延续数十年的技术方向,通用图形处理器 GPGPU 是专用处理器中最典型的代表^[16]。Intel 以 Larrabee 显卡为基础的 Xeon Phi^[3] 产品线上峰值计算性能超过 3TFlops 的 Knight Landing^[13],以及 Nvidia 的 GPGPU 产品 Tesla GP100^[4] 仍是当前最具能效比、峰值性能最高的硬件加速器。这种分离加速器有足够的软件开发、应用调试和调优积累。但在系统结构和微体系结构上存在无法深度定制,受限于和主处理器间数据传输和通信瓶颈的问题。

高性能通用众核处理的体系结构也影响着这类通用化加速处理器。从微体系结构上看,专用加速处理器在流水线控制、线程管理、向量扩展、多级片上存储和片上网络上越来越接近通用处理器核心^[16]。以往多核处理器上常见的多任务并行隐藏延迟、预取与前瞻执行、DVFS 低功耗设计等也越来越多移植到 GPGPU 上^[17]。为解决与通用主处理器之间的数据共享问题,平坦一致性存储空间,基于运行时软件的协同访存一致性扩展等成为研究热点。近年来,诸多学者对片上与片外 GPU 核的微结构展开了一系列探索,主要解决了分支预测、大寄存器堆设计及高效访存机制等异构加速器共性

问题^[18].

除 GPGPU, 也有研究人员尝试将专用加速器应用描述特征通用化, 特别是并发性、计算、通信、数据重用和协作等要素. 通过标准化步骤, 采用大量低功耗和微小核构成面向多领域的加速器. 通常这样核心也呈现出通用化特征, 例如 LSSD 结构中每个核都有个可配置空间结构、暂存器和直接存储访问.

(2) 可重构加速设计仍是重要的技术候选. 可重构计算技术可带来性能和功耗收益, 美国国家自然科学基金委员会在 2007 年成立了可重构高性能计算研究中心 CHREC, 主要合作成员包括 IBM 等厂商, MIT Lincoln 实验室, Sandia 国家实验室及国防承包商 Lockheed Martin, NASA (National Aeronautics and Space Administration) 和美国空军等. FPGA (Field-Programmable Gate Array) 由于其本身较为成熟的开发工具链及灵活适应多种类型应用, 导致基于 FGPA 的可重构技术是流行应用的重要高性能加速方案. Intel 公司在收购 Altera 之后将以多模块封装的方式提供 Xeon+FPGA 的可重构高性能处理器方案.

学术界仍试图在现有 FPGA 阵列和通用众核接口间继续选择更优化的可重构和通用编程接口折中方案^[19], 表现为面向近数据处理或深度学习等应用进行适度通用化的可重构设计^[20]. 与此同时, 现在主要 FPGA 厂商 Xilinx 和 Altera 通常是最先进半导体工艺和封装工艺的主要应用厂商, 在 2.5D/3D 封装、高速 Serdes、片上出光互连等领先传统处理器厂商应用. 同时, 在商业可重构 FPGA 和软件环境基础上, 针对 OpenCL 等异构并行开发标准进行应用移植、调试调优.

2.1.3 异构融合高性能处理器体系结构

由于良好能效比、通用编程性与高加速性能的良好折中, 异构体系结构成为热门研究问题. 为解决不同结构处理器之间的任务调度、数据共享、通信带宽和延迟问题, 将异构核心集成设计在单个晶圆上, 即异构融合高性能设计是理想的技术方案. IBM 公司的 Cell 处理器就是这种异构融合处理器的代表. 基于 Cell 处理器的 Roadrunner^[21] 也成为最先突破 PFlops 性能的高性能计算系统. 针对未来 E 级计算, 在 DARPA 的支持下, 各个公司联合大学也开始了相应研究工作, 如 Intel 和 UIUC 的 Runnemedede^[22], NVidia 和 Stanford 的 Echelon^[16] 等, 目标是单芯片性能 8-16TFlops. 日本学术界和产业界联合开展的 E 级的研究计划中提出了 PACS-G 加速众核芯片的结构. 该结构采用了主从异构的超长 SIMD 结构, 集成 4096 个核心, 峰值性能达到 16TFlops^[23]. 国内江南计算技术研究所与上海高性能集成电路设计中心也提出 4 个主核与 256 个从核的异构融合设计, 在 1.5 GHz 主频下双精度浮点峰值达到 3TFlops.

虽然具有良好能效比、加速性能和可编程性等优势, 异构融合处理器芯片一旦定型量产, 其通用计算部分和加速部分的比例关系就固定了, 不能够根据应用需求自适应调整, 且访存带宽往往受限, 而分离加速器件设计具有较好的配置灵活性, 可根据应用需求在系统构建和使用过程中动态调整通用计算与加速部分的配比, 以适配应用需求. 因此, 在实际高性能计算系统构建工程中, 从适配应用需求灵活性等方面来看更倾向于分离加速器设计. 未来先进封装技术的广泛采用和降低成本有可能弥合两种方法的差别. 同时, 类似 Intel Xeon+FPGA 的灵活可重构方式也增强了异构融合设计的吸引力. 如果未来高性能处理器也趋向于采用类似移动处理器 SoC 方式, 3D/2.5D 封装进一步降低异构集成的难度, 异构融合设计将获得更多的关注.

(1) 2.5D 与 3D 等先进封装技术下支持的异构方案. 2.5D/3D 封装技术并不能解决高性能微处理器严重的功耗和散热问题, 相反单位芯片面积上的功耗密度会增高. 2.5D/3D 封装能在不迁移到更先进工艺的条件下提高封装内集成度. 通过专门设计的 TSV 和分离晶圆内部互连, 可将用不同工

艺生产的晶圆封装在一起, 减少数据通信延迟提高带宽. 这使面向不同市场需求、采用不同生产工艺、面向低功耗和高性能加速不同目标的处理器有可能形成一个折中高性能处理器方案.

已有研究人员提出将多个芯片中的流水线部件抽象为功能单元池, 根据不同应用对不同并行性能的要求, 将三维集成处理器中功能单元分割组成不同类型的处理核心^[24]. 在芯片层面, 也可以有多种不同形式异构单元, 包括不同处理器核具有不同复杂度, 或者在同一片内部集成支持不同代次的同构指令集、不同功能扩展指令集, 乃至完全异构指令集的处理器核, 或者将通用处理器核和硬件加速器集成在同一芯片内部; 在微结构层面, 研究者关注如何针对特定应用和算法设计实现高效的硬件加速部件, 异构单元之间的高效通信机制, 以及数据一致性和同步问题. 三星基于 big.LITTLE 架构的处理器和 AMD APU 处理器是目前工业界比较有代表性的片上异构处理器^[25].

(2) 暗硅趋势与成本压力下选择 CPU+GPGPU/CPU+FPGA 单片 SoC 异构方案. 由于芯片散热技术和供电技术的限制, 以及成本限制, 可承受的微处理器功耗在近 10 年并没有太大变化. 这导致在未来 10 nm 或更先进工艺线下, 以摩尔定律增加的晶体管在动态运行期间必须通过关闭实现功耗控制, 即所谓暗硅 DarkSilicon 问题^[26]. 根据 University of California San Diego 研究人员较为悲观的预计, 如果不采用休眠等低功耗技术, 未来 10 nm 工艺结点附件 93% 以上的芯片资源在运行时必须关闭.

而先进工艺结点的生产成本越来越高, 因此未来高性能计算芯片的 SoC 设计可能仅能选择 CPU+GPGPU 或者 CPU+FPGA 这两种在通用性、生产成本和功耗下较为折中的方案. 只有在先进工艺生产成本降低, 并且大量标准化处理器 IP 成熟的情况下, 才会在单个 SoC 芯片中大量集成不同类型的加速器件.

2.1.4 基于创新器件、工艺、模型的新处理器体系结构

为在未来微电子技术突破上占据主动, 美国 DARPA 及半导体研究联盟 SRC 联合成立了先进半导体研究联合会 MARCO, 并在 2013 年建立先进半导体技术研究网络 STARnet, 主要支持在半导体集成电路领先的美国研究型大学如 UIUC (University of Illinois at Urbana-Champaign)、UCLA (University of California Los Angeles)、UC Berkeley (University of California Berkeley) 等建立 6 只研究队伍. STARnet 目标是为未来数十年的创新提供驱动力, 解决当前摩尔定律即将失效的技术障碍. STARnet 的 6 个研究中心囊括了美国在半导体、集成电路、体系结构以及应用方面的几乎全部著名大学, 研究涉及材料 FAME、纳米制造 SONIC、自旋电子 C-SPIN、低功耗系统 LEAST、未来体系结构 C-FAR 以及城市网络传感器和触发器分布蜂群 TerraSwarm. STARnet 计划在未来 5 年计划总投入 1.94 亿美元.

除了上述较为传统的计算机微体系结构研究, 认知 (类脑) 计算机和量子计算机是目前两类颠覆性的计算机体系结构. 世界主要国家都已启动了对类脑计算系统的大规模资助, 模拟人脑神经系统以复杂方式处理大量信息的原理, 用极低的功耗对信息进行异步、并行、低速和分布式处理, 并具备自主感知、识别和学习的能力. 同传统计算芯片相比, 类脑计算芯片能够突破传统体系结构“执行程序”和“处理数据”的局限. 目前该领域最重要的进展由 IBM 团队实现——IBM 团队 2008 年开始进行 SyNAPSE 项目的相关研究, 最终目标是建立一套内含 10 亿神经元、100 万亿突触的“类人脑”电脑, 而且其尺寸与功耗也要与人脑相当. 2014 年, 该团队采用三星公司 28 nm 工艺, 研制了含 4000 余个核的类脑计算芯片 TrueNorth^[27], 模拟 100 万个“神经元”、2.56 亿个“神经突触”. 美国加州大学和 Stony Brook University 的 Prezioso 等研制出了完全基于忆阻器的神经网络芯片^[28], 目前可基于该芯片感知和学习 3×3 像素的黑白图像. University of Manchester 的 Steve Furber 教授领导的起始于 2005

年的 Spinnaker 项目^[29], 计划用 5 万个 ARM 处理器互连模拟 10 亿个神经元, 能力相当于老鼠大脑. 该项目是欧盟脑计划的一部分, 目前已经有 48 芯片的单板系统, 多个这种单板系统可以通过以太网互连构建不同规模的大型系统. 德国海德堡大学提出的 BrainScaleS 项目也在类脑机制和高性能计算方面取得进展. 中国科学院计算技术研究所提出了针对学习的专门体系结构与指令集^[30], 并投产了原型处理器“寒武纪”.

2.1.5 差距

近年来国内处理器相关技术研究日益活跃. 不但有产品化的国产处理器出现, 杰出的学术研究成果也层出不穷. 飞腾、龙芯、申威、众志、兆芯、中晟宏芯等厂家的国产 CPU 研制使国际同行看到了中国在这一领域的努力, 同时提升带动了技术预研的热度. 许多大学和研究机构, 如国防科学技术大学、中国科学院、北京大学、清华大学、浙江大学、西北工业大学、上海交通大学、复旦大学等在多核/众核处理器、异构计算及处理器、高通量处理器、面向典型应用的专用体系结构、非易失存储处理器、类脑芯片、量子计算等领域作出了优秀的科研成果.

在众核体系结构方面, 我国在“核高基”重大专项、国家高技术研究发展计划、国家自然科学基金等项目支持下, 江南计算技术研究所、中国科学院计算技术研究所、国防科学技术大学等单位都深入开展了众核处理器相关的研究工作, 建立了完整的开发环境. 目前我国自主众核处理器的研究和开发水平已取得了长足进步, 部分型号已开始用于构建国产超级计算系统. 在异构体系结构研究方面也有了诸多的积累, 部分研究成果走在了国际前列. 有多家国内单位针对特定领域应用进行片上或片外加速器设计, 例如华为和百度在大数据处理场景下使用基于 FPGA 的压缩/解压缩方案进行加速.

在片上存储和片上网络方面, 我国也取得了一定的进展, 在国际顶级会议和期刊上发表了一些研究性论文, 也出现了一些原型验证芯片, 如复旦大学在 2012 年和 2013 年的国际固态电路会议 (ISSCC) 上连续公布了 16 核和 24 核的验证芯片. 但是总体而言, 我国在片上网络领域的研究依然落后于国外.

在机器学习芯片研制方面, 中国科学院计算技术研究所的基于人工神经网络算法的“电脑”系列学术研究取得了创新的成果, 引起了国际上广泛关注, 居于国际前沿地位^[30]. 基于相关研究的寒武纪系列神经网络处理器——寒武纪 1 号也已经流片. 国防科学技术大学、浙江大学以及百度公司等也在从事相关加速深度学习算法的硬件方面的研究. 作为当前加速处理器应用的热点, 国内在大规模机器学习芯片相关配套软件和应用移植方面仍有一定差距.

在类脑处理芯片方面, 国内布局也紧跟国际发展趋势. 国内研究多集中在神经形态器件的关键单元及所需新材料方面, 比如成功研制出全功能的 MB 级相变存储器件和阻变存储器件, 并提出了用神经形态存储器件辅助建立类脑计算机架构的初步设想. 浙江大学和杭州电子科技大学合作研发的类脑芯片——达尔文也于 2015 年底问世, 是我国首款支持脉冲神经网络的类脑芯片, 含 2048 个神经元和 400 万个突触, 后续将进一步增大规模^[31]. 此外, 清华大学、中国科学院自动化研究所、浙江大学、上海交通大学、国防科学技术大学、厦门大学等研究机构都成立了关于类脑计算的研究中心, 并在国家重大专项、国家重点基础研究发展计划、国家高技术研究发展计划项目的支持下建立起了开展研究所需的基本软硬件基础, 为类脑计算系统研究的实施提供了必要的技术支撑条件. 但是, 在类脑芯片的整体设计实现方面, 相关工作较少, 从事此方面的研究人员也不多, 特别是高水平成果少总体差距较大.

由于基础研究、制造体系及商业环境等方面原因, 我国在微处理器体系结构学术研究领域和工程实现领域有具体技术点和个别芯片的有限突破, 但在研究的深度、广度和持续性上与领先国家差距较大. 尤其尚未形成完整的学术研究、工业工程探索以及商业推广体系, 也没有引领国际研究潮流的突破性技术或架构出现, 多为跟踪性研究. 从十二五以来, 正在逐步形成以国家自然科学基金、国家高技

术研究发展计划以及国家科技重大专项“核心电子器件、高端通用芯片及基础软件产品”中 01/02 专项推动的前沿探索、学术研究、重点工程实现以及产业商业培育等相结合的技术发展生态环境。

2.2 高性能计算机系统技术

2.2.1 高性能计算机体系结构

过去 50 年中, 设计制造技术和应用需求不断发生变化, 高性能计算机主要经过了 3 个阶段的跨越. 1967 年 Amdahl 提出计算性能加速比公式, 强调采取程序串行部分并行化进行任务加速. 基于 Amdahl 定律, 1976 年 Cray 公司首次推出向量处理机 Cray-1, 采取向量指令和寄存器实现了向量并行结构, 性能比标量系统高一个数量级. 国际上将 Cray-1 的 1976 年称为“超级计算元年”. 此后向量计算机牢牢控制高性能市场 15 年. 80 年代并行向量处理机 PVP 也是靠并行处理提高运算速度, 包括 CDC Cyber205, NEC SX1/SX2 等. 1988 年 ACM 通讯上刊登的 Gustafson 定律开启了第 2 个阶段. Gustafson 在大量物理问题求解中发现程序无法并行化部分绝对量基本不变, 并且比例随问题规模扩大在不断减小. Gustafson 揭示出并程序可扩展本质, 促进了并行计算蓬勃发展. 20 世纪 90 年代大规模并行处理 MPP 技术本质上也是通过增加处理器数量来提升并行程序计算加速比, MPP 计算机有 Intel Paragon, IBM SP2, Intel TFLOPS 和银河曙光等. 同时代的对称多处理机 SMP 和集群系统遵循 Gustafson 定律对 MPP 路线进行了拓展延伸. 第 3 个阶段以异构计算为特征. 2002 年美国国防部意识到 MPP 和集群系统持续发展瓶颈, 引导高性能迈向高效能转变, 强调 HPC 实用效能、鲁棒性和产出率. 传统计算单元采取超标量或超长指令字, 过分依赖单个内核计算能力, 带来了复杂控制逻辑, 导致并行任务计算产出率低. 此时, 图形处理器 (GPU)、“众核”加速器为代表的新型结构步入舞台. 它们通过简化控制逻辑将更多晶体管贡献于计算功能, 带来了更多内核数量和并行计算能力, 性能面积和性能功耗比例明显提高. 2005 年 Sony 开发 Cell 处理器属于芯片异构计算^[32], Cray 开发 Rainier 系统属于结点异构计算^[33], 国际“十一五”期间主流 P 级系统均采取了异构方式, 如美国 RoadRunner 及 Cascade 等.

经历 3 个阶段, 高性能计算机系统仍在快速发展, 继续保持十年千倍增长趋势. 以美国为首的世界各国为占据科技制高点纷纷提出新一代超级计算机研制计划. 在美国“珊瑚”计划中, IBM 和 NVIDIA 公司将研制 100PF Sierra 系统; IBM 和 NVIDIA 公司将研制 150PF Summit 系统; Intel 和 Cray 将研制性能 180PF Aurora 系统. 目前高性能计算机研制处于 100P 量级, 异构计算仍属于主流. 国际后续发展计划预计在 2020 年后达到 E 级水平, 具体方案还未透露. 2010 年 4 月 IBM 公司《Some Challenges on Road from Petascale to Exascale》报告^[2]中指出, E 级系统 5 大挑战分别是功耗、访存、通信、可靠性和应用. 随着系统规模扩大, 这些挑战会严重影响系统可扩展性, 仅靠已有传统技术使高性能计算机性能大幅增长的做法不再可行, 必须有创新思路. 第 3 阶段后期将遇到功耗、成本和可靠性等严重约束条件, 亟需从体系结构、软件栈和应用方面进行探索, 寻求 E 级解决方案. 目前虽然还没有新技术明确成为 E 级主流路线, 但在系统结构层面需要关注以下几方面.

(1) 异构加速面向应用领域, 正向多态复合方向发展. 异构计算在 HPC 体系结构扮演越来越重要角色, 但形式越来越多样化. 异构计算起步的 CPU+GPU 模式具有浮点计算效率优势. 2007 年 NVIDIA 推出了统一的计算设备框架 CUDA, 标志 GPU 在通用计算领域开始成熟. 2009 年国防科技大学研制的“天河一号”在国际上首次大规模采用 CPU+GPU 异构融合体系结构, 开创了国际 HPC 体系结构新的发展方向. 如图 1 所示, 2015 年 12 月 Top500 上^[1]采用异构体系结构的千万亿次以上计算能力的系统共有 22 台. 其次, “大核 + 小核”异构计算方式也受业界青睐. Intel 推出了至强 + 众

图 1 Top500 上采用异构体系结构的千万亿次以上计算能力的系统

Figure 1 Petaflop systems with heterogeneous architecture on Top500 list

核 MIC 加速构架. 与至强处理器比较, MIC 具有更小的内核和硬件线程、更宽的矢量单元, 是提供更高性能满足高度并行化应用需求的理想之选. 理论上更简单专业的 GPU 有效计算资源更多, 能更有效提高 HPC 性能并减少能耗. 然而, Intel 具有统一编程模型和编程工具优势, 能在内核和线程之间、内存和高速缓存之间进行更多协同优化, 最大程度提高 MIC 性能. Aurora 计算机即采取了 Intel 72 核 Phi 加速构架. 异构计算还包括第 3 种 CPU+FPGA 模式. FPGA 浮点计算效能低于 GPU, 但定点计算优势显著. 随着大数据、人工智能、机器学习等新兴大数据需求不断涌现, FPGA 开始开辟新一代异构计算平台. 2013 年 IBM POWER 8 处理器提出了一致性加速处理器接口 (CAPI), 采取 PCIE 加速器方式使 FPGA 计算资源和 CPU 共享全局虚拟地址空间保持一致性, 开启了 CPU+FPGA 协同工作模式. 2015 年 Intel 收购 Altera 发布了 CPU 通过 QPI 直连 FPGA 方案^[34], 2016 年宣布将集成 Stratix 10 芯片的 Xeon E5 v4 云处理器, 预计每瓦性能提升 70%. 微软计划采用 FPGA 来代替原有处理器, 预计节省一半服务器数量. 此外 FPGA 厂商也各自提供 IDE 开发环境, 如 Xilinx SDAccel 开发环境提供了 FPGA 加速 OpenCL、C 和 C++ 内核的开发与部署等^[35]. GPU 在密集型数据和并行数据处理方面优势明显; MIC 适合于应用程序编程和性能调优; FPGA 灵活可编程且加速性能出色, 多种异构计算模式正向多态复合方向发展. 为此要结合 HPC 应用领域需求, 在多种异构、同构与异构、通用与专用、动态与静态之间获得异构计算最优平衡, 更高效的实现性能、功耗和可靠性等目标.

(2) 面向特定领域应用需求, 采取协同设计优化系统能耗效率及可用性. 针对 E 级系统挑战, 未来高性能结构针对用户需求在多重约束下进行协同设计, 通过多种结构协同适配应用, 通过软硬件协同实现系统性能、功耗、成本和可靠性方面平衡高效. 传统 HPC 软硬件设计独立, 系统构建后软件、网络与应用优化效果有限. HPC 软硬件协同设计在设计初期就结合应用领域需求, 采取软件模拟硬件仿真对系统效率、功耗、成本、可靠性进行量化分析, 采取循环迭代方式优化硬件资源配置, 并且指导硬件约束下应用软件和系统软件调优. 2011 年美国 Lawrence 实验室对美国能源部 E 级计算机开发了 CoDEx 软硬件协同设计环境^[36], 包括结点 FPGA 仿真器、访存通信特征分析工具 ROSE 及网络模拟器 SST/macro. 该环境采取结点 FPGA 平台反复迭代仿真, 利用处理器自动化工具 XPG 对指令集、通信接口及存储层次进行定制, 优化结点效能; 并采取 ROSE 通信轨迹, 利用网络模拟器 SST/macro 优化网络协议及拓扑, 降低互连能耗. 针对特定应用领域, CoDEx 优化 HPC 系统与同工艺下的 CPU 和 GPU HPC 系统比较提高了 7.6 倍和 5 倍效能. 在异构加速领域, CPU 和加速器异构架构潜力巨大, 涉及 GPU、APU、加速内核及 FPGA 多种结构协同, 采取模拟仿真平台对各种组合下的计算访存与通信指标进行评估, 不仅可以优化各种结构配置, 而且还可指导 CUDA, OpenCL 和 SDAccel 等并

行编程改善结点效能. 文献 [37] 针对 x86 异构多核处理器协同优化定制了大量精简内核改善了并行效率; 文献 [38] 结合应用任务映射对众核计算资源进行优化配置; 文献 [39] 提出了 FASTCUDA 编程环境在 FPGA 平台下对 CUDA 内核进行软硬功能划分进行加速.

(3) 柔性结构支持软件定义和硬件重构, 实现资源按需分配优化系统效能. 柔性结构指通过硬件资源逻辑解耦, 给客户提供软件可定义高性能计算体系结构. 它将硬件物理资源抽象为软件调度的虚拟资源, 将控制功能集中于上层软件堆栈中, 支持面向任务动态调度资源按需分配机制, 有效优化 HPC 实用效能. 如软件定义处理器、软件定义网络、软件定义存储等都属于柔性体系结构范畴, 它们都可有效应对极具弹性的业务需求. 柔性结构构建资源池可以从计算资源逐步扩展至 I/O 存储和内存资源, 支持资源池化灵活分配使用. 在计算资源方面, 文献 [40] 曾研究利用 HMC (Hybrid Memory Cube) 内存网络替代 PCIE 总线连接大量 GPU 资源, 解除 GPU 与 CPU 之间物理绑定, 实现了 GPU 资源池化, CPU 可以依据可扩展执行内核 (SKE) 并行度需求, 映射任务到数量可配的多个 GPU 上加速执行, 有效增加 GPU 利用率和效能. 软件可定义网络 (SDN) 将网络控制与网络转发解耦合, 基于 Openflow 模型通过控制平面^[41]对通信拓扑与交换机路由表进行软件定义, 通过合理分配网络资源达到流量均衡、拥塞避免及故障容错目的. IBM 鸮式网络也考虑了柔性结构支持, 它在网络接口中设置了多个 PCIE 根端口, 支持 CPU 采取高速网络对远程 PCIE 端口存储设备全局共享访问, 软件可以将分布式存储定义为存储资源池来灵活配置使用. 华为^[42]在构建以 PRAP 协议与云控制器为特征下一代数据中心硬件构架, 也是通过计算内存和 I/O 资源解耦来实现易扩展绿色节能目的. 未来 HPC 计算单元数量庞大、异构加速类型复杂且互连规模庞大, 具有成本、功耗和可靠性问题. 柔性结构适应不同业务场景, 可以通过软件定义及资源共享方式来支持计算弹性可扩展并提高系统效能.

2.2.2 高性能互连网络

高性能互连网络是 HPC 核心基础设施, 负责 HPC 所有结点连接和数据传输, 是 HPC 实现大规模并行计算的关键, 直接决定 HPC 性能和均衡扩展能力. 高带宽、低延迟、低能耗、高可靠、可扩展是未来互连网络设计主要目标. 高性能互连网络作为 E 级 HPC 重要组成部分, 面临挑战包括功耗、芯片吞吐率、可扩展性等. 为了有效应对上述挑战, 近年来学术界已在系统仿真、网络协议、芯片设计、光电集成、光交换等方面开展了大量工作.

(1) 仿真分析驱动网络整体设计, 属于探索新型技术的基础. 美国三大国家实验室、各国超级计算中心常年跟踪发布 HPC 应用通信特征分析报告, 并提出下一代 HPC 需求展望. IBM 联合 Barcelona 超级计算机中心开发了基于 OMNEST 模拟环境的 MARS 高性能系统互连网络模拟环境^[43], 将互连网络模拟器 Venus 与计算节点模拟器 Dimemas 整合, 支持 MPI 通信的交互式 Trace 执行驱动, 模拟精度达到微片级, 旨在验证最新一代 PERCS 互连系统. Menalox 公司基于 OMNeT++ 环境开发了 IB 协议网络模拟平台; 美国 Sandia National Laboratories 开发了 SST 并行模拟器^[44], 用于测试不同高性能计算机网络性能; IBM 面向 BlueGene 系列机器开发了 BigSim 等大规模并行模拟器^[45]; Cray 也研制了面向 Aries 系统的 Dragonfly 互连网络自适应模拟环境^[46]等. 针对高速网络特征行为分析目的在于在网络设计初期对网络协议、拓扑、芯片设计进行指导, 并辅助网络优化部署工作. Cray 通过模拟仿真验证了 Dragonfly 网络性价比优势, 带动了新型高阶网络拓扑和路由研究新浪潮, 研究者们相继提出了 SlimFly^[47], Diameter-Two^[48]等低直径新型拓扑. 国内国防科学技术大学、中国科学院计算技术研究所、清华大学、北京航空航天大学等单位也开展了超大规模网络模拟器相关工作. Intel 设计了处理器模拟模型 SIMICS^[49], 支持处理器计算与通信特征分析, 国防科学技术大学、西北工业大学等利用 SIMICS 模型开展了多核处理器系统级验证、分布式一致性协议仿真、多计算机系统通信

和大规模拓扑结构优化等工作.

(2) 高速 Serdes 保持持续发展, 交换芯片迈向更高阶数. HPC 每代互连网络都以 Serdes 速率为基本特征. 2013 年 Cray 公司 XC 系列 HPC 高速网络配置了 48 端口互连交换芯片 Aries^[50], 单个端口速率 3Lane×14 Gbps, 相对于上代 Gemini 互连端口 6.25 Gbps 速率明显改善; 2015 年日本“鲸”Tofu2^[51] 定制高速网络采取了 4Lane×25 Gbps 配置, Serdes 速率相对于 Tofu1 网络提高 2.5 倍; 2014 年 TH-2 高速互连网络配置了 24 端口交换芯片, Serdes 速率 14 Gbps, 相对 TH-1 网络交换芯片 16 端口和 10 Gbps Serdes 速率显著改善. 目前 EDR 标准 (单 Lane 25 Gbps, 单端口 100 Gbps) 是互连网络设计主流配置. Menalox 公司在 2014 年底推出了 36 端口 EDR InfiniBand 互连网络, 吞吐量 7.2 Tbps^[52]; Intel 公司收购扩展 Cray Aries 网络技术推出了 48 端口 EDR IB 网络 Omni-path, 吞吐量 9.6 Tbps^[53]; IBM 公司将高阶理念发挥极致, 推出了单端口仅 1Lane×25 Gbps, 集成了 136 端口的鸚式网络. 面向 E 级 HPC 计算机, 更高速 Serdes 可以改善带宽, 更高阶交换芯片可以减少跳步数并降低传输效能. 下一代网络将进入 HDR (单 Lane 56 Gbps) 时代, IBM, InPhi, Broadcom 及景略半导体、光梓科技等公司均在推出或在研发 56 Gbps Serdes, 为新一代网络设计提供了物理层基础. 光互连论坛 OIF 制定了 56 Gbps Serdes 规范^[54], 相对 25 Gbps 传输速率增加了 PAM4 编码格式及超短/甚短距离传输标准. 这些编码格式和标准的驱动能力和功耗将发生较大变化, 对交换芯片功能和结构设计产生了较大影响. 为此, 如何结合 Serdes 发展趋势, 充分利用路由器芯片晶体管及链路资源, 设计带宽均衡扩展、互连成本低的交换芯片是需要关注的问题.

(3) 互连通信与计算存储融合, 逐步深度协同发展. 计算与互连紧耦合是一种低延迟 I/O 架构, 缩短通信内存之间距离能显著降低延迟, 有利于优化工作负载并降低功耗. 2015 年甲骨文推出了一款面向企业级工作负载的低成本 SPARC 处理器, 整合了 SPARC M7 处理器及 InfiniBand 接口^[55]; 同年富士通公司在“鲸”超级计算机中升级了 Tofu2 互连技术, 相对 Tofu 第一代技术将互连接口、交换部件和 SPARC 处理器集成在单颗芯片中, 显著改善了密度能效, 并且提出 Cache 缓冲推送机制将延迟降低至 800 ns. 2016 年 Intel 下一代高速网络构架 Omni-Path 也宣布了计算与互连紧耦合框架, 分两个阶段实现多 die 封装及单芯片集成, 主要为降低延迟并改善传输效能. 在存储互连融合方面, 美光公司推出超立方体存储器 (HMC) 集成存储控制器设置了网络功能, 实现了内存互连紧耦合. 处理器利用 HMC 内存网络可访问远程内存资源, 支持数据远程内存之间直接拷贝, 回避了数据传输 I/O 路径上大量开销. 华为在混合内存硬件平台方面也开展类似工作, 尝试远程内存之间的 RDMA 直接传输. 针对对外存互连耦合, IBM 公司鸚式网络以 HPC 互连为基础, 融合了 100G 以太网、高速网络、PCI-E 网络、支持磁盘和 SSD 池化访问; 华为高通量数据中心计划采取 PRAP 协议和光互连网络技术来实现计算、内存和 I/O 资源池化整合^[42]. PLX 公司原本专注于 PCI-E Switch, 现在逐渐以多根 PCI-E 为基础, 向 PCI-E 网络化方向发展^[56,57]; Intel 也提出了基于光网络的机柜级资源池化方案 RSA (Rack Scale architecture), 准备实现以 PCI-E 网络为中心的分布式存储 I/O 构架^[58]. 中国科学院计算技术研究所国内率先开展了 PCI-E 网络化相关研究, 但研究局限在集中式资源池化, 难以满足未来高速存储设备性能及资源池规模可扩展需求.

(4) 硅基光互连技术发展迅猛, 高密度光电集成成为重要使能技术. 近些年工业界学术界对光互连技术关键理论和实现技术非常重视, 取得了辉煌成就. 2010 年 IBM 成功研制了硅混合激光源、MIT 激光源、Mach-Z 电光调节器、光子雪崩光电探测器^[59] 等核心光器件及 CMOS 光子单片集成, 宣布已经具备 Tbps 光互连商用化设计能力^[60,61]. 同年 DARPAR 资助 Oracle 公司提出了 macro-chip 方案^[62], 在 130 nm CMOS 工艺下验证了单路 20 Gbps 光收发芯片, 计划 2016 年实现 8×8 片上多核集成. 2012 年 Luxtera 公司推出了集成有光器件、模拟器件、数字器件的 100 Gbps 波分收发芯片, 计

划结合传统 CMOS 工艺与先进光 IP 研发新一代光互连芯片^[63]。2012 年 IBM 展出了一款光互连原型产品 Holey Opto-Chip^[64], 传输速率达 1 Tbps; 2013 年 6 月三星展出了光接口 DDR3 存储器; 2013 年 11 月, Compass EOS 公司推出了“硅光子”路由器, 每秒带宽达 TB 级。另一方面, 在光电集成微型化进程使光电集成技术产品带宽密度持续提高, 成本持续降低, 为 E 级 HPC 大规模应用创造了良好生态。Avago 和 Corning 发布 12×14 Gbps 光模块, 尺寸缩小至 7.8 mm×8.2 mm, 宣布了下代网络合作开发新的光连接业务; Molex 收购 Luxtera 公司 100 Gbps 光互连业务, 8×25 Gbps 光模块尺寸仅 22 mm×15 mm。Finisar 公司也宣布在 100G 及未来 400G 产品中采取硅光子技术提高速率和密度。工业界持续关注光电集成技术主要原因是提升集成密度使光学器件靠近电互连芯片, 有利于从器件、系统层面改善 HPC 密度功耗。2015 年国防科学技术大学自主研发了吞吐率 1200 Gbps 高密度板载光引擎, 通过近距离安置在处理器或交换芯片周围, 可以促进链路带宽和互连密度显著增长并降低能耗。未来光接口将被进一步集成到处理器或交换芯片内部。如 UC Berkeley 在 2015 年已宣布将硅光子器件光接口与处理器存储器集成在 CMOS 芯片上, 为光电集成指明了未来研究方向^[65]。

(5) 光交换技术趋于成熟, 全光交换在数据中心和 HPC 中持续演进。 MEMS 或 DBS 全光开关技术不断成熟, 表现出开关速度快、损耗低、矩阵数大、设备体积小等优势。2014 年 Xros 公司实现了 1152 端口光 MEMS 交换机, Polatis 公司推出了支持 SDN 的 384 端口完全无阻塞光交换机^[66], 加州大学 Davis 分校实现了 512 端口、尺寸仅 16 mm×11 mm 的 AWGR 光交换芯片^[67], 支持 ns 级光波长交换。上述器件逐渐成熟促使了光交换技术逐渐成为互连领域热点。目前数据中心和 HPC 领域正面临着数据流量爆发增长和动态任务特性等挑战, 通信负载容易超出数据中心或 HPC 互连能力, 引起网络拥塞和过载导致时延增加并带来服务和应用性能劣化。全光交换技术能有效应对上述挑战, 它根据用户需求可动态组建网络拓扑结构分配合理网络带宽, 不仅能均衡网络负载避免局部拥塞, 而且还可保证网络服务质量, 以支持大量应用对时延、抖动和可用性要求。特别是传统网络协议采取包交换机制, 设置了大量资源对突发流量进行优化设计, 对网络中较大持续数据流缺乏支持, 会造成网络效率偏低。如今数据中心和 HPC 中持续数据流已很常见, 如虚拟机迁移、负载均衡和数据存储及备份等。这些常用场景迫切需要增加全光交换创立更动态光纤连接, 有效处理持续数据流, 改善传输效果。此外 E 级 HPC 应用规模往往使互连全局流量低于局部流量, 以相同代价构建局部互连和全局互连会使 E 级 HPC 互连成本过高。利用全光交换技术及软件定义网络管理模式可根据当前应用需求动态重新分配光纤资源, 适合于 E 级 HPC 互连拓扑重构, 增加 HPC 可扩展性并节省互连成本。

2.2.3 基于新型存储介质的存储结构与技术

(1) 新型内存技术为内存计算带来了发展契机。 当今高性能计算机系统处于十年千倍增长的快速发展阶段, 随着多核处理器和众核处理器的普及以及系统规模的飞速增长, 无论是计算密集型还是存储密集型应用都对存储系统的容量、性能以及功耗不断提出更高的要求。然而, 由于传统 DRAM 内存工艺的发展落后于 CMOS 计算逻辑工艺, 基于 DRAM 的内存结构逐渐无法满足这些设计需求, “存储墙”问题日益突出^[68], 主要体现在: (1) 当 DRAM 的工艺超过 16 nm 后, 其容量扩展更加困难^[69], 越来越不能与下一代处理器的计算能力相匹配, 尤其很难满足数据密集型大规模科学计算、大数据处理等应用需求。(2) 系统功耗被认为是制约高性能计算技术未来发展的一大因素, 其中 DRAM 主存系统的功耗是系统功耗的重要组成部分, 每个计算结点中 DRAM 内存的功耗占整个计算结点功耗的 30%~50%^[70]。针对上述问题, 近年来学术界和工业界对计算机内存技术投入了大量研究, 取得了一系列研究成果。美光和三星公司开发了超立方体存储器 HMC 的全新内存架构。超立方体存储器 HMC 技术将多个 DRAM 芯片层层堆叠在一起, 增加存储密度与带宽并降低功耗, 对存储器性能带来革命

性影响, 预计未来 HMC 将比 DDR3 模块增加 15 倍的带宽, 而单位功耗将下降 70%. 同时, AMD 和海力士还合作开发了高带宽存储器 HBM 显存技术, 采用垂直堆叠和高速信息传输技术, 实现较小的芯片尺寸和超宽通信数据通路; 在性能方面相比 GDDR5 提升 65%, 功耗降低 40%. 2015 年 Intel 和 Micron 美光公司联合推出了 3D Xpoint, 一种全新的存储和内存技术, 采用交叉点阵列结构, 存储密度比 DRAM 提高 10 倍. 基于 3D XPoint 的 Intel DIMM 同 DDR4 在电气及物理特性上兼容, 支持下一代至强 Skylake 处理器平台, 相比 DRAM 内存容量提升 4 倍并显著降低成本.

内存计算^[71]是指采用高性能的内存来对数据进行实时计算和处理, CPU 直接从内存而不是外存上读取数据, 非常适合处理需要实时获得分析结果的数据. 近年来学术界已经有不少文献在研究相关的内存存储系统^[72]、内存检查点技术^[73]、内存数据库和内存文件系统技术. 随着内存硬件价格的下降、容量的提高, 加上 NVM 内存的非易失性优势, 为内存计算带来了发展契机.

(2) 新型非易失存储器件带来了存储层次的革新. 随着近几年微电子技术的飞速发展, 一些新型非易失存储器件 (NVM) 相继推出, 引起了学术界和产业界广泛关注, 适用于高性能计算机系统的新型 NVM 器件主要包括: 相变存储器 (PCM)^[73]、阻变随机存储器 (RRAM)^[74] 和自旋矩传输磁存储器 (STT-RAM)^[75] 等. 随着研究工作的深入, 这些新型 NVM 存储技术已逐渐从原型设计阶段走向产品产业化阶段, 为高性能计算机存储层次的变革带来了契机. 相关研究人员在缓存、内存和外存各个存储层次中尝试使用新型 NVM 器件, 达到提高存储性能和集成度、减少功耗等目的, 从各种 NVM 器件的特点和存储层次结构的不同需求考虑, 相关研究表明: 相变存储器 (PCM) 适用于设计内存和外存, 阻变随机存储器 (RRAM) 适用于设计缓存、内存和外存, 自旋矩传输磁存储器 (STT-RAM) 适用于设计缓存. 在已经广泛使用的闪存方面, 以 TLC 为基础的 3D NAND 将是超大容量、高性能外部存储系统的发展重点. 新型 NVM 存储器件带来了存储层次的变革, 给高性能计算机存储体系结构及相应的软件栈带来了变革与发展机遇, 为如何缓解“存储墙”问题提供了有益的探索途径.

(3) 新型存储技术为存储与计算的融合带来了契机. 目前已发现的新型存储材料中, 阻变随机存储器 (RRAM) 引起了学术界和产业界的广泛关注, 忆阻存储技术已经在实验室实现, 并正在向商品化的目标努力. RRAM 的典型代表为惠普实验室研究发现的忆阻器, 具有非易失性、存储密度高等优点; 与 PCM、MRAM 等新型非易失存储器相比, RRAM 具有器件结构简单、制备工艺简便、擦写速度快和存储密度高等诸多优势. RRAM 使用常规的薄膜制备技术和工艺即可, 器件结构简单、制备工艺简便; 半导体工艺兼容性好, RRAM 可以利用现有的半导体工艺技术生产, 从而大大缩减开发成本. RRAM 擦写速度快, 目前已有的研究中 RRAM 的擦写速度可达到 0.3 ns.

惠普实验室研究发现, 忆阻器兼备状态逻辑计算能力, 既是一种良好的新型存储器件, 又是一种颇具潜能的使能器件. 基于忆阻器构建的存储系统具备直接进行逻辑计算的能力, 可实现存储与计算的有机融合. 忆阻存储技术为存储与计算的融合带来了契机, 为高性能计算机体系结构带来了新的发展机遇.

2.2.4 并行编程模型与并行运行支撑系统

共享存储编程模型一般包含多个可以共享内存的并行任务, 通过隐式使用共享数据来完成任务间的数据交换. 共享存储模型的存储可扩展性好, 提供灵活的任务调度策略, 编程相对简单. OpenMP API^[76]是面向共享存储系统的主要并行编程模型和事实上的工业标准, 它基于编译指导命令的编程接口, 具有编程简单、可移植性好等优点. OpenMP API 从 4.0 版本开始支持异构计算, 目前已经更新到 4.5 版本, 大量新增的 API 体现在异构计算的支持方面, 以顺应目前异构计算的发展趋势. OpenMP 设计时将存储系统看作全局一致的, 并没有考虑访存不一致性 (NUMA), 因此在具有访存不一致性的

NUMA 系统上性能可扩展性较差; 模型也缺少面向层次式并行特征进行的针对性设计. 针对共享存储编程模型在层次式并行、局部性感知等方面的问题, 编程模型扩展和运行时优化是当前研究的热点.

(1) **消息传递模型仍是大规模并行编程主流, 主要关注可靠性、可扩展性等方面.** 消息传递编程模型的并行任务间通过发送和接收消息完成数据交换和协同工作, 一般面向分布式存储系统的并行编程. MPI 作为消息传递编程模型事实上的工业标准, 设计符合分布式存储的特征, 具有程序性能高、扩展性好的特征. MPI 2.2 中提供的应用通信拓扑描述, 提供给用户优化日益层次化的并行计算系统; MPI 3.0 中进一步增加了非阻塞聚合通信和非阻塞单边通信的支持, 以便隐藏日益复杂的网络带来的通信延迟. 目前高性能计算领域基本达成共识, MPI+X^[77] 在相当长一段时间内仍是大规模并行计算系统的主流编程模式, 也是实现 E 级计算最可行的技术路线. 面对未来高性能计算的挑战, MPI 的相关研究重点关注可扩展性设计 (特别是存储可扩展)、异构计算、容错计算等技术^[31,78~80].

(2) **PGAS 模型借鉴共享存储和消息传递特征, 依赖编译优化制约其发展.** PGAS (partitioned global address space), 又称全局分割地址空间模型, 是一种介于共享存储和分布式存储之间的混合并行编程模型. PGAS 允许并发任务通过一个全局共享的数据结构访问远程数据, 共享数据结构可以在不同的任务之间进行逻辑上的划分, 任务访问本地数据空间比远程数据空间更加高效. 与消息传递的模型相比, PGAS 更容易组织通信并支持一些非规则数据访问; 与共享存储方式相比, 采用显式单边通信的远程数据访问方式, 避免了共享存储模型对存储一致性的依赖. 但是因为需要进行显式数据和任务划分, 采用 PGAS 编程并不简单; 底层的通信性能严重依赖编译优化, 可能影响程序性能扩展; 模型实现一般采用多任务方式, 存储资源占用较高; 对层次性、灵活性、异构性的支持都不佳. PGAS 的代表语言有 UPC^[81], Co-array Fortran^[82], GA^[83], Chapel^[84], X10^[85] 和 Fortress^[86] 等, 近几年研究热度有逐年下降的趋势.

(3) **异构编程发展迅猛, 标准化和性能优化是当前关键问题.** 随着异构体系结构在高性能计算领域的异军突起, 异构编程模型近年来发展非常迅猛, 研究主要是围绕支持加速器特别是当前流行的 GPGPU, Xeon Phi 等加速设备而展开的. 这种编程模型设计考虑了系统大量细粒度的线程和层次的存储结构, 暴露硬件特征使得用户可以显式控制管理底层硬件, 发挥众核性能. 由于体系结构的多样性以及不同编程风格的设计, 目前出现了 OpenMP 4.x^[76], Open-ACC^[87], CUDA^[88], OpenCL^[89] 等一系列异构编程接口, 其中大多是面向结点内异构计算设计的, 支持面向多级存储层次的编程, 设计针对性强, 因此在当前加速器系统上性能和存储扩展性较好. 缺点是不能很好地屏蔽硬件细节, 编程困难; 缺少对任务调度的灵活支持, 只能用于开发规则的数据级并行. 由于异构加速器设计模式的多种多样, 异构计算编程接口呈现碎片化的现状, 目前尚未有明显迹象表明哪种接口将成为最终的标准, 因此如何抽象出主流异构系统设计的关键特征, 推进异构编程接口的标准化, 从而使得编程模型不会成为程序员设计开发及性能优化的障碍, 是当前异构编程模型面临的巨大挑战.

(4) **国际上对并行编程发展路线没有清晰判定, 面临诸多挑战.** 2011 年美国能源部 ASCR 办公室组织的编程专题工作组对 E 级系统面临的编程模型挑战进行了分析, 对实现未来 E 级编程模型的技术路线持两种观点: 革命路线和进化路线^[90]. 在当前 MPI 编程方法编程难度不令人满意的情况下, 革命路线主张全面取代 MPI, 以克服多核、众核结构复杂性、多样性所带来的可移植性和性能屏障, 要在编程难度和程序效率方面全面超越 MPI, 成为学术界的一个重要研究方向. 但是目前没有哪种编程模型展示出这样的前景, 即使有这样的模型, 从研发到全面部署要很长的周期, 难以期望其在 E 级系统中得到全面部署. 更实际可行的路线是 MPI+X 的进化路线, 即结点间采用 MPI 支持并行开发, 而结点内采用统一的一种编程模型, 但目前这种模型是什么并不清晰. 采用进化技术路线的原因包括: 显然 MPI 在 10 年内不可能消失; 到 2016 年 MPI 作为分布系统编程标准已经达 20 多年历史, 大量

代码使用 MPI 开发; MPI 足以满足未来 E 级系统的结点数规模扩展. 尽管进化路线是更实际可行的技术路线, 但是即使采用进化路线, 仍然需要大量的研究和开发工作, 结点内编程模型是面临的主要挑战之一. 作为 MPI+X 混合编程模型一部分, E 级系统结点内编程模型不仅要满足传统混合编程模型的需求, 还要面临 E 级系统带来的包含能耗和可靠性等在内的新挑战.

2.2.5 差距

我国的高性能计算技术坚持创新发展, 天河二号成为世界超算史上首台连续 6 次位居 Top500 榜首的高性能计算机系统, “神威太湖之光”成为世界上首台达到十亿亿次的高性能计算机系统, 实现了高性能计算机研制的群体性突破, 但在关键器件基础方面仍落后于美国, 主要表现在我国基础器件后端生产、封装、测试等环节及设计工具长期受制于外, 国内半导体集成电路工艺普遍落后. 即使目前国产自主处理器领域进步显著, 但应对高性能计算领域的实际性能、稳定性和适用性仍有较大差距. 面临 E 级计算机功耗、可靠性、密度等挑战, 缺乏高端核心关键器件掌控能力必将难以开展体系结构、系统软件和应用软件一体化设计与优化, 无法有效提高整体计算效能. 在新原理计算器件和新型计算系统等方面基础性研究也显不足, 核心技术有较大差距, 对新型使能技术前瞻性研究比较落后. 从整体上看, 美国、日本在新型半导体器件上有巨大领先优势, 较多新技术已出现在高性能计算机发展道路之上. 采用美光公司开发的 3D 堆叠技术的混合存储立方体 (HMC)、AMD 和海力士合作开发的高带宽存储器 (HBM) 技术将对 E 级系统存储器性能带来巨大影响, 能大幅提升带宽并降低单位功耗; 采取 Intel 的 3D XPoint 新型存储有速度快、容量大、非易失特点有助于改善 E 级系统 I/O 存储性能; 光互连技术具有高带宽、低损耗、无串扰特点, 能有效解决通信墙问题, 硅基光互连技术更具有成本和互连密度上的优势. 尽管在大规模并行体系结构、自主 CPU 设计、系统软件和一些高性能计算应用等多方面取得令人瞩目的进展, 实现了历史性突破, 但国内目前在核心关键技术和基础新器件上与第一梯队相比仍有较大差距.

在高性能互连网络方面, 我国总体达到国际领先水平, 天河一号/二号基于自主设计的高阶路由器芯片和高性能网络接口芯片实现了高性能、高密度、均衡扩展互连网络, 取得多项创新性成果, 神威蓝光设计的大规模高速互连交换系统突破了一系列关键技术, 满足了大型超节点高性能通信和系统扩展要求. 但针对 E 级系统互连网络需求, 国内整合硬件体系结构能力以及吸收新型使能技术能力还显得不足. 目前众核处理器、加速器、专用 FPGA 等异构融合系统结构被广泛采用, 光互连光交换技术趋于成熟, 这些对系统互连网络提出了新挑战. 未来高速网络系统需要更强大的互连互通能力, 也需要在多网络融合、资源共享方面扮演重要角色. 为此, 往往强调体系结构层面计算与通信、网络与存储的深度融合, 各类存储协议及监控协议的融合. 但国内高性能处理器、高性能存储及各类存储协议与网络协议方面基础薄弱, 这必会影响硬件体系结构的整合协同创新能力. 比较而言国际上 Intel、甲骨文、美光等公司对体系结构资源具有较好的资源整合能力, 已取得部分创新成果. 另外, 在新型技术使能方面, 国外学术界工业界经过多年积累在高密度光电器件集成、硅基光互连芯片设计及系统光互连交换等方面发展迅速, 已经具有了初步产业化雏形, 有望应用于下一代互连网络中. 新型使能技术应用将会加速未来超高阶互连网络和动态可重构网络发展, 从而实现低功耗高密度互连网络. 但目前国内新型使能技术发展现状相对缓慢, 学术机构和企业在光互连光交换核心器件生产、封装、测试方面仍存在较大差距, 也必然会影响到下一步高速互连网络对新型使能技术吸收采用.

国内众多并行编程模型研究在支撑我国高性能计算技术整体达到国际先进水平发挥了重要作用, 但与国际主流的编程模型研究仍存在较大差距. 近些年来, 国内积极开展了并程序设计语言与模型、编译以及运行时系统等方面的研究. 代表性工作有清华大学面向 GPU 的编程模型 BSGP^[91]、中国科

学院计算技术研究所扩展 UPC 的数据分布层次支持异构 GPU 系统的研究^[92]、国防科学技术大学面向 CPU/GPU 异构计算结构的运行时协同优化^[93]等。但就当前现状而言, 研究人员倾向于支持开放的编程标准与接口, 包括 MPI, OpenMP, OpenACC, OpenCL 等。而这些编程标准设计者一般是来自工业界和学术界的研究人员, 主要是来自美国、欧洲、日本的研究机构, 他们一方面负责标准制定并根据需求适时进行修订, 另一方面提供在主流处理器平台上实现。如果将这些开放编程标准看作高性能计算并行编程未来, 那么积极参与这些标准的制定与修订工作变得异常重要。另一方面, 我国也在力推国产化高性能微处理器技术, 那么提供开放编程标准实现将更好地建立国产平台上的软件生态, 有利于将国产平台推向更广泛的应用。

2.3 高性能并行算法与应用

2.3.1 数值方法与基础并行算法

(1) 数值方法在科学和工程计算中得到广泛应用。领域应用分为基于连续模型建模和基于离散模型建模, 其中连续模型在使用并行手段进行求解时, 必须借助于数值方法对其进行离散, 进而映射为并行计算模型; 而离散模型则可以直接映射为并行计算模型。数值方法和数值模拟软件负责将领域应用问题模型映射为并行计算模型, 因而在高性能计算应用领域占据核心地位。当前基于连续模型的数值方法主要包括有限差分法、有限元法、有限体积法等, 随着问题复杂性和应用领域的扩展, 高分辨率和复杂外形适应成为该领域的热点和难点问题。因而相继开发出谱方法、谱元法、间断有限元方法等新型数值方法。基于离散模型的数值方法主要有分子动力学模拟方法、Lattice Boltzmann 方法、Monte Carlo 方法等。数值方法经过几百年的发展, 广泛应用到不同的专业领域, 成为认识和改造世界的有力工具。

(2) 基础并行算法研究成果颇丰, 细粒度数据级并行算法得到较多关注。基础并行算法主要是指使用并行手段求解问题时, 所需的共性通用并行计算过程, 主要分为数值类并行算法和非数值类并行算法。数值类并行算法主要集中在线性代数求解领域, 如 Gauss 消去法、基于矩阵分裂的 Jacobi 迭代法、基于 Krylov 子空间的共轭梯度法、广义小残差法, 这类并行算法存在固定操作模式; 非数值类并行算法主要包括抽样、搜索、排序、比较、非线性处理等问题, 较少有固定的操作模式。典型的对流扩散方程的并行数值求解算法被归纳成显式直接并行、隐式迭代并行、交替分组显式并行和 Monte Carlo 并行 4 类^[94], 这对其他领域的并行算法设计和开发具有指导意义。

国内外研究主要集中在细粒度数据级并行算法^[95,96]、分数阶并行算法^[97,98]、异构协同并行算法^[99~101]、波阵面并行算法^[102~105]、全局通信隐藏的共轭梯度法^[106,107]等方面。随着向量化、SIMD 功能部件的涌现, 越来越多并行应用关注到 GPU, MIC, DSP 上的细粒度数据级并行算法^[108~110]。

细粒度数据级并行算法主要针对类 SIMD 向量功能部件体系架构, 采用大规模细粒度并行计算技术, 结合多级访存优化, 设计与新型体系结构相适应的并行算法, 反应了现代体系架构演化趋势。由于分数阶算子的非局部性, 其计算复杂度比常规整数阶算子的复杂度高一个阶级, 分数阶问题的模拟求解必将越来越依赖于高性能并行计算。由于异构系统中主协处理器计算能力不均衡、通信能力不均衡, 异构协同并行算法主要聚焦于静态/动态任务划分、任务调度、主协处理器之间的通信设计及优化。波阵面并行算法与传统的矩阵分裂、Krylov 子空间方法有较大不同^[103,105,111], 波阵面扫描中的数据强依赖的解决方法对新型并行算法设计具有一定的借鉴意义。随着并行进程数增加, 全局归约带来的通信将是性能瓶颈, 流水化方法有助于提高 Krylov 子空间方法的并行性能。受益于 MPI-3 标准中的非阻塞聚合通信 (non-blocking collectives), 可以有效利用预条件算子的求解过程隐藏全局同步通

信开销, 从而提高算法的并行效率^[106]. MPI-3 标准中异步全局通信机制和全局同步通信开销隐藏的 Krylov 子空间方法对大规模并行应用有多大的影响, 仍然有待更多应用的验证.

2.3.2 高性能并行应用

(1) 美国引领高性能应用研究, 其研发模式值得借鉴. 长期以来, 美国引领了全球高性能计算应用的发展, 占据了高性能计算应用的战略科技制高点和全球较大比例的产业化市场份额. 总体而言, 美国的高性能计算应用呈现国家实验室提升能力、市场普适推广、国家超算中心普惠服务的局面, 形成了可持续发展的高性能计算应用生态环境.

美国国家实验室一直以高性能计算能力型应用为科技创新的强力引擎, 也是能力型应用软件的研发者和使用者. 例如, 美国能源部核武器实验室 LLNL (Lawrence Livermore National Laboratory)、LANL (Los Alamos National Laboratory) 和 SANL (Sandia National Laboratories) 是核武器研究、核能开发与利用、全球气候变化、复杂工程与重大装置优化设计、材料科学、生命科学等领域科技创新的实践者; 美国能源部 ORNL (Oak Ridge National Laboratory)、ANL (Argonne National Laboratory)、LLBL (Lawrence Berkeley National Laboratory) 等是核能开发与利用、复杂电磁环境、地球环境与资源勘探等领域先进技术的推动者; 美国宇航局 NASA 和气候研究中心 NCAR (National Center for Atmospheric) 是飞行器气动力学和全球气候变化与天气预报等领域科技创新的引领者. 2014 年 11 月, 全球 TOP500 排名前十名的高性能计算机系统中, 一半系统安装在这些实验室. 正是通过这些国家实验室在能力型应用和应用软件研发方面的卓越工作, 全球高性能计算才有突飞猛进的发展. 同时, 美国禁止能力型应用的技术出口, 确保其始终占据高性能计算应用的战略科技制高点. 欧洲在高性能计算应用方面拥有相当多的专利、知识产权和应用软件. 日本和欧洲在高性能计算应用方面也有很长发展历史, 在科学发现、数值模拟、设计优化等方面形成了一批重要应用系统.

(2) 典型并行应用研究成果显著, 国内研究有待加强. 2010 年美国能源部 DOE 指出 E 级计算的挑战包括 9 个方面应用: 航空航天飞行器和引擎上的湍流^[112]、天体物理、生物医药、天气和气候、材料科学、聚变能源、国家安全、核工程. 2006 年, 日本针对“鲸”超级计算机启动 ISLiM 项目, 主要想对生命物质进行模拟和新药开发^[113], 其中多尺度心脏模拟、分子动力学模拟、血液细胞流固耦合模拟分别可以获得 2, 3, 4 PFlops 的性能, 超声波模拟可以有效扩展到 10000 个计算结点. 欧洲启动 CRESTA 计划开发 E 级并行应用, 包括 ELMFIRE, GROMACS, IFS, HemeLB, NEK5000 和 OpenFOAM 等. 石油地震数据处理、飞行器流场模拟、生物分子动力学模拟、磁约束聚变数值模拟、湍流数值模拟、晶体硅分子动力学模拟、全球大气浅水波全隐式数值模拟、地球外核热流动数值模拟等在采用 CPU/GPU 异构融合架构的天河一号上进行了测试, 验证了应用和系统良好可扩展性^[114].

粒子输运数值模拟一直是推动高性能计算发展的主要动力之一^[95,102]. 2008 年, 粒子输运程序 Sweep3D 就是第一台千万亿次计算机系统 Roadrunner 上的第一个应用^[102], 在全系统 3060 个结点上测试结果表明, 使用 PowerXCell8i 增强型协处理器后, Sweep3D 单次迭代速度约为只使用 AMD Opteron 速度的 2 倍. 在 GPU 和 MIC 上, 粒子输运的数据强依赖被设计成波阵面并行扫描, Sn 递归方程采用循环拆分的方式进行并行处理, 提高整个应用的并行度, 具有一定的方法学意义, 在实际数值实验中获得较好的加速效果^[95,96]. 非结构化粒子输运问题, 一种有意思的物理模拟过程和 CUDA 并行编程模型之间的映射关系被提出^[103], 如图 2 所示. 这种的映射关系在其他应用普遍存在.

航空航天领域一直是国家战略支持的高科技领域, 引领人类文明发展方向, 与此相关的大规模并行应用算法、软件得到了国家的重视. 包为民院士等对航天领域高性能并行计算研究进展进行综述^[115], 指出国家科技部在“十二五”国家高技术研究发展计划“高效能计算机及应用服务环境”中涉

图 2 非结构粒子输运与 CUDA 编程模型之间的映射

Figure 2 Mapping between non-structured particle transport and CUDA programming model

及航天领域的重大应用软件系统所占比例达到 62.5%; 对气动力^[116,117]、气动热、化学非平衡、结构强度、热防护、Monte Carlo 方法和湍流研究中的大规模并行应用算法、软件和协同优化进行了分类总结; 指出航天领域高性能并行计算存在科学计算高并行效率和工程计算低实用价值, 并行应用的多样性和缺少科学的并行方法的矛盾. 这两种矛盾在科学工程研究中是普遍存在的问题.

地震波传播数值模拟不但可以用于预测、模拟地震灾害, 也可以用于石油勘探. 任意高精度间断有限元求解器在 SuperMUC 上可以获得 90% 的并行效率, 14.7 万核上可以获得 45% 的峰值性能和 1.09 PFlops 的持续性能^[118]. 这种 PFlops 持续性能的获得是很有挑战的. 在机器学习并行计算方面, Coates 等研究了 COTS 高性能计算系统上用三天时间得到 10 亿个参数的深度学习网络^[119]. 在基础科学研究方面, 异构并行计算取得丰硕成果. 分子动力学在 ASCI 定制的专用 Anton 2 上对 23558 原子系统可以达到 85 μ s 每天的速度, 比通用超级计算机快 180 倍. 这反映了计算机发展的定制化、软硬件协同设计、算法硬件化的趋势^[120]. 基础物理^[121]、生物医药^[122,123]、天体物理^[124]、气候气象^[125]、地球物理^[126,127] 等应用领域的大规模并行计算都得到验证.

2.3.3 并行算法库、框架与并行应用软件

(1) 多核异构的体系结构对并行算法库的研究提供了机遇和挑战. 在并行算法库领域, 国外有几十年的积累, 诞生许多优秀的开源并行算法库、软件包, 但是很多不适于多核和异构的计算机体系结构, 存在是并行性差、同步开销大、可扩展性不好等问题, 难于发挥机器峰值性能. 著名的经典算法库有基本线性代数库 BLAS^[128]、稠密数值代数库 LAPACK、可扩展线性代数计算软件包 ScaLAPACK^[129]、快速计算离散 Fourier 变换的标准 C 语言程序集 FFTW^[130] 等. 针对现代计算机的多核和异构体系结构, 目前正在开发的算法库有面向多核 CPU 的 PLASMA 以及分布式的 DPLASMA、面向 GPU 和 MIC 的 MAGMA 等, 这些算法库采用某个多任务调度的运行时系统, 目标是减少已有算法库的同步开销. 目前国内开源的并行算法库并不是很多, 据作者已知的有基本线性代数库 OpenBLAS、大规模对称特征问题并行求解器 HPSEPS. 计算机的多核、众核和异构 (CPU+GPU、MIC) 的体系结构给并行算法库的研制提出了重大的挑战, 许多算法需要重新设计和撰写, 从而提高并行性和可扩展性, 发挥机器的峰值性能. 国内在高效的并行算法库方面与国外的差距还很明显.

(2) 领域/并行计算框架是发展趋势但挑战巨大. 应用领域研究人员由于学科领域知识鸿沟, 难以将领域问题高效快速转化为并行程序模型, 这阻碍了并行计算手段在科学研究中的高效实施. 近年来人们开始研究和开发了面向特定领域的并行应用开发框架模式和工具, 其核心思想是针对特定应用领域对应用采取了高层抽象, 建立应用集成开发框架和平台, 应用领域专家仅通过近似自然语言的方

式对应用问题进行描述,屏蔽了更多的底层程序设计细节,用户无需了解太多计算机程序设计方面的细节就可以方便的开发出适合自身需求的应用;同时底层开发人员也可以方便的对平台各个方面进行优化,以提升系统性能.国外著名的领域应用框架包括科学计算可移植工具包 PETSC^[131]、求解大规模复杂科学与工程问题的 Trilinos、求解计算流体力学的 OpenFoam^[132].国内著名的框架包括三维并行结构自适应软件框架 JASMIN 和非结构网格 JAUMIN^[133],三维并行自适应有限元软件平台 PHG¹⁾,较好地解决了自适应方法并行实现中的负载平衡难题,在大规模计算机系统上可以获得较高的并行效率.这种面向领域的并行计算框架将会缩小系统开发人员和应用开发人员之间的鸿沟,实现了跨领域协作,最终实现并行应用的少编程或者零编程,并在框架内解决性能优化、可靠性等问题,是当前高性能计算软件研究中极具吸引力的一个发展趋势.但目前应用软件开发人员接受领域计算框架这种高层抽象仍然具有极大的挑战性,领域计算框架在函数接口、易用性、可移植性和功能模块设计与实现等方面要加强研究.

(3) 国外并行应用软件占据绝对优势地位,国内差距明显. 并行应用软件属于典型的普适型应用.在开源或者商业化的并行应用软件领域,与国外相比,国内的差距非常大.国外在科学与工程并行计算软件走在前列,并且进行商品化,但是国内使用的功能和并行度受到商业限制.基于连续模型的著名软件包括大型有限元分析软件 ANSYS (结构强度),有限体积数值模拟软件 Fluent、CFD++、通用有限元分析软件 Abaqus (电磁)、开源软件 OpenFoam^[132] 等;基于离散模型的著名软件包括分子动力学模拟软件 NAMD^[134]、LAMMPS^[135] 和 Gromacs^[136] 等,其中 NAMD 是当前在并行可扩展性、模拟规模方面最为成功的软件,而 LAMMPS 则在功能覆盖方面更为广泛.国内大规模商用的并行应用软件则相对匮乏,在特定领域,如中国空气动力研究与发展中心、国防科学技术大学、中国航空气动力技术研究院的 CFD 软件在国内有一定竞争力.中国空气动力研究与发展中心研发的面向流体工程的混合计算平台 PHENGLEI (风雷) 在产品商业化方面进行了有益的尝试.而在成功的开源应用软件方面,国内鲜有成功案例.

2.3.4 差距

从科学研究角度,我国缺乏原创性、具有普遍价值的高性能并行应用方法/算法;从工程实践角度,我国缺少具有国际竞争力的商业化并行应用软件产品.尽管近几年越来越多的应用领域采用高性能计算机开发和运行并行应用,我们在若干领域取得有限突破,但在并行算法理论研究水平、典型大规模并行应用、并行算法库/框架、开源应用软件和并行应用软件商业化方面都有相当大的差距,也极其缺乏跨领域的人才.尚未形成完整的理论研究、应用软件研发以及商业推广体系,产学研结合仍待加强,也没有引领国际研究方向的突破性算法或软件出现,多为跟踪性研究.对于 E 级计算,重要领域的高性能计算能力型应用、基础并行算法库仍然欠缺.与高性能计算系统上的进步相比,并行应用方面的差距就非常明显.到目前为止,我国很少有应用领域能够走在世界前列.应用是推动整个高性能计算发展的源动力之一,我们只有弥补应用上的差距,才能构建一个有竞争力的高性能计算生态系统,才能持续、健康向前发展.

3 发展趋势分析与展望

高性能计算技术的发展主要由器件技术和应用两个方面驱动.器件技术方面,半导体工艺线宽的

1) <http://lsec.cc.ac.cn/phg/>.

持续降低将在未来几年接近终结, 然而 3D 封装等技术使得摩尔定律仍然能够延续, 并使得人们可以建造更大和更小的计算机系统; 光互连技术有望缓解带宽瓶颈; PRAM, RRAM 等非易失内存技术的出现对传统计算机的结构和系统软件都提出了新的挑战. 应用方面, 以大数据处理为代表的访存密集型和 I/O 密集型应用近年来迅猛发展, 对以计算密集型为中心的传统计算机系统结构提出了挑战. 传统的科学与工程计算应用也在向着更大并行规模和更加复杂的计算模式方向演进. 认知计算则是刚刚兴起的一类重要应用.

下面从微处理器体系结构、高性能计算系统、高性能并行算法与应用这 3 个方面来分别说明相关的发展趋势.

3.1 微处理器体系结构

3.1.1 通用高性能处理器发展趋势与挑战

(1) **面临主要挑战.** 多核/众核高性能通用微处理器持续在 ILP、TLP 和 DLP 等多个维度上继续开发并行性. 但来自应用本身、底层器件、生产工艺和功耗/散热成本的实际限制也越来越严峻. 有限的功耗预算导致的功耗称为多核/众核高性能通用微处理器设计的首要限制和挑战. 保持其发展的主要动力来自于延续并保护 30 多年来以 OpenMP 和 MPI 为基础搭建并行应用价值, 能够降低并行应用移植代价. 以低延迟和高计算效率优化为主的核心仍然可以在更先进生产工艺下持续发展至少 2 代, 单芯片核心数量可以保持增长, 但能效比已经进入边际效应的末端. 核心数量的增长将对片上存储层次、片上互连、片间直连通信和一致性协议的扩展性都提出了更高要求. 持续使用最先进生产工艺也对整个处理器芯片的可靠性和容错设计提出更高要求.

(2) **技术发展趋势与解决思路.** 虽然目前多核高性能通用微处理器尚无法支持下一代 E 级高性能所需的能效比, 但在这类微处理器上的技术解决方案将可能使所有处理器受益.

(i) **借鉴移动处理器领域的低功耗设计, 追求整个处理器芯片的高能效比.** 由于功耗已经成为高性能计算系统整机到微处理器芯片的首要设计约束, 需要在处理器结构、逻辑、电路和器件工艺多个层次全面推进以往在移动处理器中常用的低功耗设计技术. 使用多阈值与多电压器件, 以及 DVFS 技术来改善动态和静态功耗已经成为必需的设计选择. 在进一步的功耗压力下, 选择单线程性能略低的高能效比核心, 通过更多核心数量来提升全芯片性能和能效是可能的设计选择.

(ii) **集成先进存储器件应对片上存储容量和延迟平衡需求.** 单处理器芯片性能提升需要与之对应的片上存储平衡设计, 以足够的缓存容量和延迟级别保证对高性能计算部件供应数据. 类似 eDRAM, NVM 等先进的存储器件能够保持最外一级高速缓存容量按照核心数量和芯片峰值计算能力同步增长, 同时还具有进一步降低面积和功耗需求等额外优势.

(iii) **进一步简化并集成高性能互连到处理器芯片内部.** 随着高性能处理器单芯片内部处理核心数量超过 20, 简单的多个环形互连已经很难满足处理核心之间, 以及核心与缓存之间的数据通信. Mesh、多维 Torus、Butterfly 或者 Dragonfly 等原用于高性能计算机系统的互连拓扑也被纳入到片上互连的设计层次中. 同时为了简化更大规模系统构建的复杂度, 片间可能包含一致性协议支持的直连机制也属于片上互连考虑的技术范畴.

(iv) **使用创新结构支持一致性、并行性、可靠性和容错需求.** 由于片上单元增多, 导致一致性协议开销增大, 传统基于目录的一致性协议可能会被基于时间标记的简单硬件实现替代. 事务存储已经成为软件并行实现的一种有效方案, 随着潜在事务粒度和回滚开销增大, 需要探索额外硬件支持和加速方式. 先进工艺的生产变化以及未来 E 级系统的极大规模, 导致软错误的影响进一步放大, 需要在

包括片上存储单元、流水线通路等各个方面考虑可靠性和容错技术的应用。

3.1.2 专用于领域特定微处理器发展趋势与挑战

(1) **面临主要挑战。** 由于较长时间和较为广泛的技术积累与应用移植经验, GPGPU 仍然是大规模系统部署情况下最常见, 并且是最具性价比的加速器发展方案。但 GPGPU 加速器仍然受制于加速板卡上数据存储容量和带宽限制, 以及加速器与主处理器之间的通信瓶颈。虽然大量销售的 GPGPU 有成本上的优势, 但并不如 FPGA 或者定制 ASIC 一样能够根据应用灵活调整计算部件、寄存器、片上存储和通信在最终处理器中的资源占用与实现方式。类似特定领域的 ASIP 以及可重构 FPGA 加速器仍然是小规模和特定领域加速的重要选择, 技术方案的适用性和可行性受限于可能收益应用的开发难度, 以及开发工具的完备性和有效性。

(2) **技术发展趋势与解决思路。** GPGPU 已经率先采用 HBM 为代表的三维封装存储来解决加速处理器所需的缓存带宽、容量与延迟问题。随着 GPGPU 等加速器峰值计算性能的进一步提高, 以 HBM2 为代表的堆叠存储也会遇到功耗墙等问题。需要通过类似 HMC 等简化存储控制协议, 或 NV-Link 等专用高速互连, 以及与 CPU 共享平坦存储空间, 并支持不同类型处理器间一致性协议扩展来提高整个计算系统的效率。

基于特定领域的定制加速处理器以及 FPGA 可重构方案则需要在兼顾 OpenCL 等标准化并行开发环境的基础上, 进一步结合目标应用提出更具能效比和扩展效率的解决方案。在软硬件协同设计 CoDesign 的早期, 使用通用、标准的描述方法来驱动软硬件协同优化是潜在的技术方向。

3.1.3 异构融合高性能处理器发展趋势与挑战

(1) **面临主要挑战。** 以 2.5D 和 3D 堆叠为代表的先进生产和封装工艺虽然能够以更多资源缓解高性能微处理器片上存储容量、数据访问延迟和片内通信带宽等瓶颈, 但导致单位面积上功耗密度增大, 使功耗墙问题恶化。在单一硅片和封装内融合多种定制加速器结构虽然能够以低动态功耗和平局功耗来缓解与暗硅相关的功耗问题, 但对各类处理核心晶圆设计和互连都提出额外要求。同时大量硅片资源以及动态热点区域的切换也可能导致片上电压、电流出现浪涌噪声, 并对器件有效寿命造成不良影响, 从而降低整个异构处理器的可用性和可靠性。

(2) **技术发展趋势与解决思路。** 将流水线, 特别是计算功能单元、缓存和路由器作为动态资源池, 依照应用类型的不同进行动态调度, 从而获得高峰值性能, 以及分散功耗热点, 延长器件老化寿命等优势是异构融合处理器可能的技术发展方向。单细粒度共享资源池的控制需要在微体系结构、逻辑设计和封装设计时就预先考虑, 同时也需要报告操作系统调度和运行时软件辅助等额外的软件优化配套工作。

在短期较近的发展时间内, 通过先进封装工艺将现有的异构加速器, 特别是 GPGPU 和 FPGA 与高性能多核处理器封装在一起, 提供统一的访存空间、一致性和数据通信视图是最容易实现的技术路径, 相关的软硬件技术储备也最完善。更多异构加速部件的集成融合有赖于软件系统的协同发展趋势。

3.1.4 基于创新器件、工艺、模型的新型处理器趋势与挑战

(1) **面临主要挑战。** 相变、阻变等非易失性新型存储器件受到器件可靠性、读写访问速度和生产工艺成本的限制, 在短期内可能只作为主存与传统磁盘之间的额外加速存储级别而存在, 并进一步

加速向近数据存储计算方向的发展. 但受限于计算模型和应用开发的配套推广水平, 尚难以完全以忆阻器的部件构建以存储为中心的高性能计算系统.

基于碳纳米管与单电子晶体管等创新器件的传统计算模型尚未完全脱离实验室规模制备的限制, 距离完全工业化生产尚有距离. 以拟神经/类脑、纳米光学、超导量子等全新器件为基础的新型计算模型尚在开辟和证明其在特定领域的有效性阶段, 在大多数现有应用上无法替代传统多核高性能微处理器.

(2) 技术发展趋势与解决思路. 随着信息化社会大量数据产生和采集, 以及大数据处理的流行, 从性能和能耗上考虑, 以存储为中心的数据处理都优于目前以计算为中心的体系结构设计. 例如以机器学习和深度学习为代表的加速处理器可以降低神经元的处理精度来获得更低的功耗, 并在单一芯片上实现更多层次神经触突与互连层次. 面向各种类型应用的近数据处理技术越来越受到学术界和工业界的重视. 但在短期内, 包括机器学习、量子处理和光计算等技术方案仍然处于应用加速的地位. 在完全替代传统处理器之前, 额外的近数据处理和新型存储器的缓冲机制, 会进一步提升高性能微处理器和整个计算系统的实现成本.

由于 CMOS 器件和工艺的广泛应用, 在新型器件和新计算模型使能技术上沿着 CMOS 工艺兼容方向发展非常重要. 一方面可以迅速降低新技术的推广成本, 另一方面也可以在兼容现有计算原理的前提下检验新型计算技术的有效性. 新型计算模型的推广与使用除了新微处理器结构设计和工艺实现的调整外, 也需要相关软件, 特别是应用开发、调试和优化工具软件的配套. 在这一领域协同设计的创新压力和工作量要远远超过当前的异构融合高性能处理器.

3.2 高性能计算机系统技术

3.2.1 高性能计算机体系结构发展趋势与挑战

(1) 面临主要挑战

长期以来 HPC 性能持续提升主要依赖 3 个方面进步: 1) HPC 系统结构持续创新; 2) 结点性能持续提高; 3) 系统规模不断扩大. 但随着 HPC 进入 E 级却遇到了以下情形: 1) 宏观系统结构创新开始不足, 异构融合系统结构在未来一段时间内仍将是 HPC 主流, 新型系统结构或颠覆性计算技术尚未出现或成熟; 2) 芯片工艺进步速度放缓, 芯片功耗、片内网络、芯片引脚数、芯片可靠性等因素影响日益凸显, 结点性能持续提升空间有限; 3) 系统规模扩展面临功耗、可靠性、应用可扩展等挑战.

具体来说, E 级系统主要面临能效、访存 I/O 带宽、可靠性及可扩展性等方面挑战. 在能效方面, HPC 功耗已攀升至数十万亿瓦, 对供电设计、散热设计产生重大影响, 尤其是持续运营成本过高, 严重影响和制约 E 级高性能计算机研制和应用推广. 目前业界共识, E 级高性能计算机功耗将限制在 20 MW 左右, 需要在天河二号同等功耗预算下提高整体性能约 20 倍, 意味着需要提高系统整体能效比约 20 倍. 功耗约束下系统整体优化设计是面临的最重大的挑战. 在存储 I/O 带宽方面, 随着芯片设计工艺进步, 单芯片内集成资源更加丰富, 芯片处理能 (计算、交换等) 持续快速提升, 但芯片引脚 I/O 能力或芯片吞吐率提升较为缓慢. 一方面表现为芯片有效引脚数增加非常缓慢, 甚至呈下降趋势, 另一方面表现为单引脚带宽提升空间有限. 对处理器而言, 表现为存储墙问题; 对互连交换芯片而言, 表现为芯片吞吐率提升缓慢. 在 RAS 方面, E 级高性能计算机规模仍将持续扩展, 可能达到 10~20 万结点规模, 超大规模系统可靠性、可用性、可维修设计将是 E 级高性能计算机设计中重要挑战. 在可扩展性方面, E 级计算的访存墙、通信墙、可靠性墙、能耗墙严重制约并行计算的可扩展性, 即并行应用运行效率并不随着计算机规模扩大而线性增长, 因此, 可扩展性是 E 级计算的重要挑战之一, 也是

业界关心的核心问题之一。

(2) 技术发展趋势与解决思路

(i) 多态融合异构着力于改善高性能计算及大数据处理能效。异构计算体系结构较长时间之内仍是高性能计算机主流, 可能出现采用 CPU/多类加速器协同计算、单独使用加速器或者多种加速器组合方式来满足多种不同领域应用需求。未来趋势将根据特定应用需求构建不同性能和功能的计算环境, 高效的实现不同性能和效能目标, 涉及到通过统一节点规范和互连接口, 灵活组织异构计算体系结构, 实现多种计算资源比例灵活调整和性能功耗的平衡。系统层面还将会利用可扩展紧耦合互连网络聚合多态异构资源以达到较高的应用饱和性能。此外, 高性能计算出现与大数据问题融合趋势, 采取高效的异构模式处理大数据应用丰富的并行性, 也将成为未来发展重要趋势。

(ii) 体系结构与应用协同设计及新技术软硬件适配设计。E 级计算系统需要围绕负载特征开展体系结构软硬件协同设计。美国能源部明确提出 E 级计算机主要瞄准全球变化气候与资源环境、国家核武库安全等 7 个领域, 为 Intel, IBM, Cray 等系统设计商提供了明确指导; 欧洲和日本拟研制部署的 E 级计算系统也具有各自重点应用领域。从应用软件系统角度看, 计算模式和通信模式都将远远超过现阶段 P 级系统水平, 在存储、计算、通信中都会呈现特别复杂的层次性和异构性。未来发展趋势需要大规模软硬件协同设计环境指导 E 级系统设计, 或者通过现有 HPC 系统上提取应用特征推演至 E 级规模, 涉及应用特征数据采集、分析与建模, 对 E 级弹性可扩展行为特征预测, 给 E 级计算机系统设计提供参考。从系统评测角度看, 针对未来复杂多样应用, 设计选择体现应用特点的基准测试程序集合, 完成对 E 级高效能计算机系统的全面评测也是重点研究领域。另外 E 级系统为了满足多目标指标需求, 必须有新的创新思路, 可能出现异构加速部件集成、3D 处理器结构、非传统处理器结构、非易失新型存储器件、新型互连器件以及光电集成等等。融合新概念器件特性的计算机体系结构需要进行软硬件协同优化设计才能实现性能、功耗、成本和可靠性等方面的综合平衡高效。

(iii) 故障监控机制研究有助于提高大规模系统可靠性。E 级计算机将不可避免地呈现出系统架构多态复合、资源类型异构多样、资源数量急剧增多、巨量资源故障频发、可靠性急剧降低等显著特点。为保证大规模系统的可用性, 需从系统体系结构层面开展并部署大规模系统的监控管理技术, 主要包括能效智能控制管理、基于深度学习的故障发现与预测机制及基于无线网络的高效监控管理等技术。

3.2.2 高速网络发展趋势与挑战

(1) 面临主要挑战。以 E 级新一代高性能高密度互连网络为总体目标, 对高性能互连网络进行性能指标分解。E 级互连网络规模多达 10 万个结点, 为天河二号规模 6 倍, 每个结点需要升级至 400 Gbps 速率, 整个互连功耗预算 4~6 MW。采取传统互连技术实现 E 级系统互连网络难度较大, 主要面临 3 个方面技术挑战。1) 芯片功耗和系统功耗难以承受。近年来 Serdes 传输效能提升缓慢, 使得交换芯片功耗与吞吐率同步保持增加。2015 年 OIF 报告指出每代 Serdes 速率翻番对应的每比特功耗仅仅下降 20%。随着 HDR Serdes 速率即将进入 50 Gbps, 各种复杂均衡放大及 DSP 电路还会继续恶化 Serdes 功耗。现在交换芯片已接近 200 W 功耗上限, 继续提升交换芯片吞吐率会面临芯片功耗上限问题。在系统层次由于芯片总吞吐率受限, 提升 Serdes 速度则会使交换芯片阶数下降, 最终导致 E 级网络规模急剧增加, 互连系统功耗恶化。2) 互连网络密度严峻。HPC 处理器性能每隔 18 个月翻一倍, 而 Serdes 端口速率每次翻番约 3~4 年, 造成处理器性能和网络速率剪刀差发展。高性能计算机为维持全局平衡的性能通信比会增加交换机数量换取聚合带宽, 势必造成网络密度工程化瓶颈。如果 E 级 HPC 继续采用现有密度扩展至 10 万结点可能需要 130 个通信机柜, 接近计算机柜数量。交换机密度

与光纤铜缆密度尤其突出, 需要互连网络协议和结构方面统筹考虑. 3) 可靠性问题突出. Serdes 速率提升后信号完整性要求越来越高, 不利于高速信号质量控制, 会导致 PCB 良率下降, 将对 E 级 HPC 互连网络可靠性提出更大挑战. 为了在低可靠性链路上实现高可靠传输, 允许在链路层采取更强纠错能力的编码技术, 如前向纠错编码, 但会显著恶化网络延迟. 下一代互连网络可靠性将对互连成本性能等产生不利影响.

(2) 技术发展趋势与解决思路. 以上述具体挑战为牵引, 高性能互连网络可从高速网络系统结构、融合互连网络设计实现、新型使能技术等多个层面出发探索技术发展趋势和可能解决思路.

(i) 处理器与通信接口片上集成显著减小延迟并改善密度功耗. 传统网络架构通常需设计由两款芯片构成的互连芯片组, 接口芯片实现各类结点到网络的接入, 交换芯片构建大规模网络实现结点间互连和交换. 下一代互连网络设计将实现计算与通信融合设计, 将互连通信接口嵌入到处理器中. 目前在摩尔定律驱动下, 几家芯片设计公司包括 Intel, AMD, NVIDIA 等都在逐步将通信接口集成到处理器内部, 能显著降低通信延迟, 提高系统密度并减小系统功耗.

(ii) 高密度光电集成和系统光交换是实现低功耗高密度网络重要手段. 互连网络功耗较程度上归结于 Serdes. 高密度光电集成技术能显著缩短芯片高速 Serdes 驱动距离, 提高信号完整性, 减小 Serdes 面积, 使单个芯片可以集成更多 Serdes, 设计更高阶网络芯片成为可能, 是当前提高芯片集成密度和系统互连密度最为重要的使能技术, 也是当前国际研究发展趋势. 目前国际光器件厂商推出板载光引擎也遵循这一技术思路, UC Berkeley 推出处理器和硅光器件单芯片集成更是引领了该研究方向. 另外基于波长交换、光交换开关等技术已成熟可以实现系统级链路光交换, 以此实现系统级可重构光互连拓扑, 实现网络资源弹性分配也是降低 E 级网络资源配置有效手段. ISCA、SIGCOM 和 NSDI 等顶级国际会议已相继发表相关论文, 已成为本领域研究热点.

(iii) 实现网络与 I/O 存储融合设计来提高资源利用率. 系统资源动态分配与资源共享是 E 级 HPC 改善系统效能的有效手段. 除了计算资源, 实现内存资源和存储资源分配调度和管理访问能大幅节省成本. IBM 的 Harrier 网络支持磁盘和 SSD 分布式全局访问; Intel 和 Avago 的 PCI-E 网络主要面向机柜级资源池化共享; 华为数据中心部署提出 PRAP 协议, 利用精简协议栈和内存地址路由目标实现计算结点资源解耦和按需分配. 传统分布式计算架构中 I/O 存储依附于结点, 只能静态分配资源. 未来依托光互连技术将内存资源、存储资源、I/O 资源依附在高速互连网络之上, 实现网络与存储深度融合, 实现分布式 I/O 存储资源池化, 能在保证系统扩展性同时提高系统资源利用率.

(iv) 实现 HPC 多网络融合简化 HPC 网络架构支持高密度低成本网络. HPC 高性能互连网络、存储互连网络、以太网监控管理网络设置了多套网络, 影响了 HPC 系统密度, 带来了更高设计成本、使用成本和维护成本. 实现多网络融合设计将是未来网络重要发展趋势. 高速网络和存储网络融合在协议栈及数据通路层面的融合统一, 建立网络与存储 I/O 的协同性能优化机制, 有助于发挥新型存储介质性能; 采取高速互连网络代替以太网管理 HPC 系统, 有助于保持管理通道与数据通道故障一致性, 降低管理维护成本.

(v) 高阶路由器结构、大规模低直径拓扑、容错路由算法及网络管理方式持续优化. E 级 HPC 网络规模扩大造成的功耗密度可靠性问题带来了新的系统设计约束, 路由器结构、网络拓扑、路由策略、网络监控等问题将持续发展演化. 高阶路由器结构需要解决端口数量增加引起的资源聚增问题; 大规模网络拓扑需要关注维度增加和流量均衡之间的协同问题; 容错路由需要区分各种故障实现软件故障无感知的通信服务; 网络管理技术需要实现网络在线故障检测、诊断和隔离等.

3.2.3 于新型存储介质的存储结构与技术

(1) 面临主要挑战. 随着高性能计算机系统规模扩展到 E 级, 数据存储规模达到 EB 甚至 ZB 规模, 存储与计算的矛盾日益突出, 现有存储系统架构和 I/O 软件栈技术无法支撑未来 E 级系统的要求, 存储系统结构对计算机系统整体性能的制约日益突出, 主要表现在传统存储层次的各个方面: 一是传统的静态随机存取存储器 SRAM 和动态随机存取存储器 DRAM 工艺的存储密度限制了存储容量的增长, 制约了缓存 (CPU 片上存储) 和内存容量的扩大; 二是传统的 SRAM 和 DRAM 存储器工艺静态功耗较高, 内存系统功耗是全系统功耗的重要组成部分, 是制约高性能计算机向 E 级规模发展的一大因素; 三是以传统磁盘为主要存储介质的外存系统, 受限于磁盘本身的性能劣势, 一直为制约 I/O 性能的主要技术瓶颈; 四是随着计算机系统规模扩展到 E 级, 数据存储规模扩展到 EB 级, 而容量不断扩大的存储器件错误率越来越高, 给 E 级计算机系统的数据可靠性带来了严峻挑战.

(2) 技术发展趋势与解决思路. 针对超大规模并发度以及数据密集型应用对 E 级计算机存储系统的现实需求, 可从基于新型存储器件的存储系统结构、系统容错技术、存储与计算融合技术等方向展开研究, 探索技术发展趋势和“存储墙”可能的解决思路.

(i) 利用新型非易失存储器件设计混合多层次存储结构. 随着近几年微电子技术的飞速发展, 一些新型存储器件相继推出, 具有存储密度高容量大、功耗低、非易失、性能高等优点, 不仅能够显著降低引入大量 DRAM 内存产生的静态功耗, 同时有望填补 SRAM 缓存、DRAM 内存与磁盘外存之间不断扩大的性能鸿沟, 为高性能计算机存储层次带来变革, 存储领域在未来相当长时间里将会是多种存储介质共存, 各自拥有独特的应用领域或者存储层次, 更加多样性的混合多层次存储将是未来高性能计算机系统的一个重要特征. 如何有效利用新型存储介质进行存储结构设计, 针对新型非易失存储器件设计新型存储层次, 是为 E 级计算机系统设计超大规模高性能存储系统的有效技术途径.

需要研究解决的具体问题及技术途经主要包括以下几个方面:

- 有效利用新型存储介质设计混合多层次存储结构, 通过利用 I/O 局部性, 在多个存储层次处理响应 I/O 访问请求, 以提供高带宽、可均衡扩展的 I/O 能力. 由于新型非易失存储器件本身固有的一些特性, 例如读写性能不对称、写次数有限等, 这些特性为新型非易失存储器件在高性能计算机存储系统中的应用提出了挑战. 因此需要研究如何在各个存储层次选择合适的新型存储介质. 如利用 DRAM 和 3D Xpoint 存储器件设计混合的非易失内存; 利用 3D NAND 的 TLC 闪存构建超大容量、高性能外部存储系统. 同时, 采用软件定义存储理念, 利用与计算结点基本同构的服务器平台, 实现面向存储的功能定制、并发 I/O 性能优化, 如实现 I/O 转发、协议转换、数据缓存、特定运算和存储加速功能.

- 传统操作系统、文件系统和数据库系统等软件都是针对传统基于磁盘的存储系统而设计的, 由于新型存储介质模糊了内存与外存的界限, 不加变革的直接沿用将无法充分发挥新型存储器件的技术优势, 因此需要研究相应的软件栈和存储软件优化方法, 研究如何实现内存与外存的统一管理, 针对新型 NVM 器件设计新的数据结构和缓存优化算法, 减少不必要的写操作, 降低其写能耗和写延迟.

- 研究基于新型非易失存储系统的数据可靠性优化方法. 研究基于新型非易失内存的检查点技术, 减少检查点的保存与恢复开销, 高效支持应用续算, 提高 E 级系统的鲁棒性. 研究针对应用场景适配的耗损均衡算法, 提高新型非易失存储器件的寿命.

(ii) 基于新型存储技术研究存储与计算融合的方法. 在高性能计算系统中, 大量的能耗和操作延时是数据在系统中不同部件间进行搬运导致的, 计算与存储融合是大数据背景下计算机体系结构发展的一个重要趋势. 为了消除存储瓶颈, 研究面向高性能计算的宏观融合存储体系结构, 以及计算与存储微观集成的体系结构. 以新型存储器件技术、三维封装技术和互连技术为基础, 根据技术成熟度

和应用需求, 在宏观和微观等不同层次上实现计算与存储相融合就近处理数据。

忆阻器 RRAM 兼备逻辑计算能力, 既是一种良好的新型存储器件, 又是一种颇具潜能的新型使能器件, 为存储与计算的融合带来了契机。基于新型存储技术, 面向特定应用研究存储与计算的融合方法。如我们针对图像处理, 研究提出了基于忆阻器的图像处理运算与存储融合结构设计, 在图像数据的存储位置上即可进行简单的图像处理操作, 从而大大减少图像数据的访存操作, 达到提高图像处理性能和效率的目的。面向大数据处理, 研究设计数据查询、排序、聚合、谓词过滤等典型数据处理操作。在自治容错方面, 研究设计了基于忆阻器的自治容错结构和基于水平垂直一致校验码的自治容错技术。

3.2.4 并行编程模型与并行运行支撑系统

(1) 面临主要挑战。 从高性能计算的发展趋势看, 虽然不排除使能技术的突破导致单芯片的性能得到飞跃性的提升, 但开发并行性仍将是提升性能最主要的技术手段。在从现阶段的 P 级计算向 E 级计算演进的过程中, 系统的并行度仍将得到大幅扩展, 并且包含多个层次、形态多样的并行。PGI 公司的编译器工程师、著名的计算机专家 Michael Wolfe 将并行总结为 6 个层次: 结点级并行、Socket 级并行、核级并行、向量级并行、流水线级并行和指令级并行, 并指出, 要实现 E 级计算, 在编程框架中必须同时充分发掘这 6 个层次的并行, 缺一不可。这对编程模型和运行时支撑系统的研究提出了明确的要求: 必须能高效、简洁地开发各种层次的并行。

目前的编程框架中, 各个层次的并行往往由不同的编程模型或接口开发, 例如结点级并行通常依赖 MPI 通信库, Socket 级、核级并行通常使用 OpenMP 模型, 向量级并行主要依赖程序员或程序员辅助的编译器开发, 流水线和指令级并行的开发则主要依赖芯片微体系结构。这种复杂多态的编程体系对应用开发造成很大困扰, 以至于很难对各个层次的并行同时进行把握和开发。另一方面, 异构系统的出现加剧了编程的复杂性。目前异构并行主要出现在结点内, 由通用多核处理器和加速器, 如 GPU、MIC 等加速芯片构成, 虽然这可以归于 Socket 级并行, 但由于指令集的不同、存储空间相互独立, 这种异构并行开发和同构多处理器有着很大不同, 仍存在很多亟待解决的问题。最后, 无论是使用 MPI 还是 OpenMP, 都需要程序员具有很强的计算机领域背景和对目标体系结构较深的理解, 才能开发出高效的并行程序。而计算机系统的最终用户不一定是计算机领域的专家, 更多的是来自物理、化学、生物、图形图像处理等各个领域的研究者, 让他们使用计算机属性很强的编程模型和框架去表示应用中的并行性本身就具有很大的局限性, 这也很大程度上困扰了高性能计算机的普适应用。

(2) 技术发展趋势与解决思路。 从技术发展趋势看, 以消息传递为基础的大规模并行编程模型 MPI+X 仍旧是未来高性能计算机的主流发展趋势之一。伴随着最近几年并行计算机系统体型结构的新特性, 以及大规模并行应用的需求, MPI 标准接口的扩展和更新也趋于活跃, 出现了一些新的标准接口定义以及功能扩展的探讨, 以便提高并行应用的通信与计算重叠执行能力、可扩展性、面向多核处理器的优化、面向应用通信模式的单边通信优化, 以及容错支持等。如何降低此类编程模型的复杂性和提高效率是关键。海量并发编程模型、异构众核平台编程模型等新型编程语言、工具和并行编译优化技术的研究将是未来的发展重点。一方面是为了更好地利用体型结构内多层并行特性, 另一方面也是为了减少并行应用中的并行任务数, 降低系统资源需求, 从而提供全系统并行应用的可扩展性。以 MapReduce 为代表的易于编程的云计算并行编程模型将会异军突起, 在面向大数据处理应用中将很可能诞生出各种诸如面向图处理、流处理等不同新型领域的编程模型, 这些并行编程模型都需要高性能计算机系统提供相应的高效并行运行系统来支撑其并行应用的高性能运行, 也是研究的重点之一。

3.3 高性能并行算法与应用

(1) 面临主要挑战. 高性能并行应用主要挑战分为 4 个层次: 1) 方法学层面, 原创性、基础性且被广泛应用的并行计算方法较少, 基本上处于一种跟随研究、应用的状态, 没有开创性、引领性研究方向. 2) 应用层面, 典型高性能应用取得长远进步, 但国际认可度仍有待加强. 3) 软件层面, 在专业应用软件和开源软件研发方面差距明显. 我国高性能计算应用软件开发起步晚, 整体力量弱, 目前仍大量采用从案例到案例的模式, 基于基础并行支撑环境开发单个并行应用, 整个开发过程缺乏框架化、系统化、可继承性设计, 存在大量重复设计工作. 目前国际主流方向是针对某一领域开发可解决一类问题的普适性并行软件. 我国绝大多数重要行业的能力型应用软件开发和实际应用水平还处于起步或跟跑阶段. 在普适应用软件开发和应用方面, 我国高性能计算部署较少, 总体力量薄弱. 绝大多数高性能计算应用的市场公司是美国等发达国家商业软件代理商, 不具备自主创新产品研发能力. 4) 协同创新层面, 学科领域分割严重且协同创新能力不足. 高性能计算应用软件开发面临严峻的跨领域学科鸿沟问题, 在开发过程中计算机学科专家与应用领域专家缺乏有效沟通, 计算机专家只能从计算片段优化入手参与并行应用软件开发, 对整个并行应用软件开发全貌缺乏宏观系统掌握. 当前国际主流并行应用软件开发领域主要采用面向领域的应用开发框架模式, 由多领域专家协同开发和优化已成为业界共识.

具体针对 E 级系统并行应用方面, 面临的主要挑战包括: 1) 面向新型体系结构的细粒度数据级并行算法, 必须开发面向量化功能部件的并行性, 这是 E 级并行应用基础. 2) 适应异构体系结构的动态任务分配调度、通信优化方法, 这是应用扩展到 E 级规模和实现 E 级异构协同计算的关键. 3) 面向多级存储结构的访存优化方法, 应用才能够获得较高的运行性能, 这是 E 级并行应用性能优化的核心. 4) 面向 E 级计算的国家重大行业应用领域和前沿基础研究领域, 针对普遍适应的共性问题, 凝练和分析并行算法的可扩展性能瓶颈, 研究适应于百亿亿次计算的可扩展并行算法, 研制相应的并行算法库, 并进行典型应用验证是实现 E 级能力型并行应用和并行软件代码重用的必由之路.

(2) 技术发展趋势与解决思路. 以上述具体挑战为牵引, 并行应用主要从基础并行算法设计、能力型和普适型应用研究、并行算法库和并行框架等多个层面进行技术探索.

(i) 面向 E 级计算和理论前沿开展基础并行算法研究. 面向 E 级计算的基础并行算法主要是与 E 级系统体系结构相适应的细粒度数据级并行算法、任务分配和通信优化方法、多级访存优化方法等. 针对量子计算、分数阶问题、计算物理/化学、计算生物等理论前沿问题, 开展高效并行算法是实现相应问题验证与确认的必由之路. 在第一性原理计算中, 就存在两种截然不同的模式: 一种是计算效率高, 但是可扩展性极差; 另一种是计算效率极低, 但是可扩展性好. 在分数阶微积分模型并行模拟中, 除了传统整数阶微积分模型并行模拟所面临的问题外, 其存在时空全局依赖性和时空复杂度提高一个阶次. 所以, 探索并行性好、计算效率高、存储量少的高效并行算法显得尤为必要.

(ii) 面向重大应用领域、重要行业开发能力型和普适型应用. 针对重大应用领域开发能力型应用, 围绕重要行业或多个行业中特定领域的挑战性应用问题展开, 通常要求高性能计算机系统具有较大规模的计算能力, 侧重于核心关键技术的突破, 可以体现高性能计算作为国家战略科技创新强力引擎的核心价值. 普适型应用围绕国家重要行业或多个行业中普遍存在的科学与工程应用问题展开, 侧重于核心关键技术的产品化和市场化, 对高性能计算机系统计算能力的需求明显低于能力型, 使用高性能计算机系统的模式通常为容量型, 其高性能计算能够为新型战略性新兴产业升级、科技创新提供强力支持. 能力型应用面向 E 级计算, Gordon Bell 奖是一个衡量的标志之一. 普适型应用面向工程计算, 探索商业化软件产品.

(iii) 面向领域开展并行算法库和并行应用开发框架研究。并行算法库可以软件接口形式提供给第三方并行应用软件调用, 提高了代码的重用率和应用软件开发效率, 减少了不必要的重复开发工作。领域并行框架通过框架解决领域用户所面临的编程墙、多学科协同合作问题, 着重于解决两个问题: 一是领域用户友好的开发接口和使用接口, 使得领域用户少编程或者无编程, 降低并行手段使用门槛; 二是框架内的各领域专家的协作分工, 并实现跨领域的面向应用和面向体系结构的协同优化, 提高并行手段使用效率。

4 重点开展研究的领域 (或科学问题) 与相关政策建议

本文将根据技术的发展趋势与挑战性问题, 以及我国的实际情况和国家在自主可控和自主创新方面的发展战略, 建议相关的优先资助领域。

4.1 众核和异构处理器体系结构

未来十年, 处理器上的晶体管数仍然会持续增加, 从而给处理器的功耗、访存带宽、可靠性等提出重大挑战性问题。众核和异构体系结构是缓解上述挑战的重要方向, 研究工作主要包括: 面向高性能计算或数据密集型应用的多目标平衡的可扩展微处理器结构; 可扩展片上存储层次与片上互连网络; 众核处理器低功耗管理与设计; 众核处理器可靠性和自愈技术; 异构处理器系统的设计和资源管理调度等。

4.2 面向应用的低功耗高性能定制处理器与计算机系统

面向应用的定制处理器在特定应用领域可以提供显著高于通用处理器的性能价格比和性能功耗比, 研究工作主要包括: 应用特征提取技术、定制处理器体系结构设计方法; 机器学习共性运算和存储特性提取和归纳、机器学习处理器结构与设计; 类脑计算模型、类脑计算芯片结构、类脑计算系统高性能模拟验证相关支撑软件; 高效图计算机结构模型、图计算机系统。

4.3 面向先进工艺、器件和计算模型的模拟、验证和测试等工具链

下一代深亚微米工艺下传统器件, 以及新型的单电子晶体管、忆阻器、碳纳米管等, 特别是量子、神经元等新计算模型的机理尚未完全明确。在这些技术大规模应用于高性能计算系统实现前, 首先主要挑战问题是从器件到电路、模块、子系统和全系统的功能、性能以及可靠性和可信性的模拟、验证和最终测试方法的缺失。现有软件工具的发展仅能满足传统工艺未来两代的发展, 尚无法有效支撑新型计算技术的全面发展。研究工作主要包括: 同步/异步混合芯片的设计流程及方法学; 相变存储等忆阻器和碳纳米管等新型器件的电路、子系统和全系统级建模与快速模拟验证; 与传统设计和结构组成并行模拟、接口融合以及灵活调用的协同设计框架与工具集。

4.4 可扩展高能效计算机体系结构

面向 E 级以上高性能计算机系统如何在低功耗的情况下保持系统的高性能成为其系统结构的首要挑战。研究工作主要包括: 支持传统科学工程计算和科学大数据等多类型应用的低功耗高性能计算系统结构; 异构协同计算方法和结构; 系统的计算、存储和通信的均衡设计和实现技术, 软硬件系统的协同设计与优化技术; E 级以上高性能计算机系统新的基准测试方法和程序。

4.5 面向 E 级 HPC 的融合网络体系结构

当前 HPC 正迈向 E 级计算时代, 系统规模不断扩大, 应用领域不断拓展, 包括众核处理器、加速器、专用 FPGA 的异构融合系统结构被广泛采用, 光互连光交换等新型技术趋于成熟, 都对系统互连网络提出了诸多新挑战. 为此需要面向 E 级 HPC 需求, 研究新的互连网络架构及其实现技术. 研究工作主要包括: 深入分析 E 级 HPC 应用通信行为特征, 研究高密度光电子集成和系统级光交换等新型使能技术, 探索高性能互连网络与计算系统、存储系统的深度融合实现技术, 实现高性能网络与国产处理器的深度融合, 支持构建系统弹性可扩展、与应用特征相匹配的融合多协议高性能高密度新型互连网络.

4.6 高性能网络计算技术

从网格计算到云计算以及未来可能的边缘计算等多种网络计算都是在探索如何构建国家的未来数字计算基础设施, 这个过程一直面临许多重大技术挑战, 其研究工作主要包括: 网络计算的核心结构、高性能云计算中心的架构、云操作系统与一系列支撑技术、新型的边缘计算模型和分布系统结构、云边协作方法与技术 and 系统支撑技术等.

4.7 面向大数据处理的计算系统

大数据的出现对计算机系统结构提出了挑战. 研究工作主要包括: 以数据为中心的计算和存储一体化的新型计算机系统; 基于大内存的数据组织和处理方法, 基于大内存和新型存储级内存的计算系统; 面向大数据的高时效分布式处理与存储系统架构与技术, 面向大数据挖掘、数据关联性分析、图处理与流处理等一系列大数据处理应用的并行分布系统软件技术.

4.8 并行编程模型和并行运行支撑系统

并行编程一直是高性能计算机发展的重大挑战性问题, 其研究工作主要包括: 海量并发编程模型; 面向异构众核的程序设计语言及其编译与并行运行时系统; 面向性能、资源利用率和功耗等指标, 研究并行编译优化和自动调优技术; 领域编程语言的平台研究: 支持高性能的领域编程语言的快速研制, 支持语言的互操作; 面向大规模 MPMD 应用程序和数据并程序的性能调优技术; 有效管理海量资源和进程的高效并行运行与交互机制等大规模并行运行系统软件技术.

4.9 高性能并行算法与应用

高性能计算科学与工程应用领域模型向着多尺度、多物理和百亿亿次性能需求发展, 计算模型的复杂性越来越高, 其研究工作主要包括: 与体系结构相适应的细粒度数据级并行算法; 极大规模可扩展并行算法; 分数阶并行算法等理论前沿问题; 高精度数值计算格式; 开源并行算法库和共性算法库; 领域框架、并行框架及性能优化技术; 重大应用领域典型能力型并行应用和商业化应用软件等.

4.10 新型存储介质技术和新型主存级存储技术

新型存储介质与技术为存储带来革命性的变革的同时也带来了巨大的技术挑战, 其研究工作主要包括: 基于新型存储介质的存储芯片与系统构建技术; 大规模新型存储阵列系统与可靠性技术; 面向新型存储介质的 I/O 系统优化技术和文件系统技术; 新型非易失主存级存储系统的存取接口、访问与管理技术; 基于新主存技术的操作系统等系统软件的优化和适应性技术; 新型存储介质与 DRAM 的

混合主存系统技术; 新型存储介质和磁盘的混合存储技术; 研究基于新型存储介质的新型存储网技术; 基于忆阻器的计算与存储耦合结构和方法。

4.11 大规模存储系统和软件定义存储技术

大数据和云计算环境的存储面临巨大挑战, 研究工作主要包括: 面向大数据的大规模存储系统技术与分布存储技术; 云存储模式与构建方法、基于互联网高效数据传输和易访问管理、支持移动环境等云端协同存储管理访问技术; 超大规模云存储纠删码技术与可靠性技术; 重复数据删除技术; 适应大规模虚拟化计算环境下的虚拟存储技术; 软件定义存储的体系结构、存储软件栈优化、存储硬件虚拟化与标准化、面向应用的存储自动化等技术。

5 总结

高性能计算技术是国家战略高科技技术, 是解决经济建设、社会发展、科学进步、国家安全方面一系列重大挑战性问题的重要手段。我国在高性能计算方面的研究和工程方面都取得了很大的进步, 已经具备了世界一流水平的地位和势态, 但是我们也应当看到这个高科技领域的竞争非常激烈, 我们还有许多方面的工作和核心基础存在较大的问题, 因此我国对这个方面的长期投入是必须坚定不移的, 而且应当采取基础研究和工程与应用的互动和结合的方式, 对世界开放、包容和共赢的方式开展这个领域的未来工作, 这对下一步我国自主创新的发展战略具有重要的意义。以上内容是在国家自然科学基金委的大力支持下, 由许多专家共同参与撰写, 参考了中华人民共和国科学技术部高性能计算专家团队撰写的稿件等, 在此表示感谢, 以上内容和建议仅供学者和机构参考。

参考文献

- 1 Meuer H, Strohmaier E, Dongarra J, et al. TOP500 Supercomputer Sites. <http://www.top500.org>. 2016
- 2 Lucas R, Ang J, Bergman K, et al. DOE advanced scientific computing advisory subcommittee (ASCAC) report: top ten exascale research challenges. <http://www.osti.gov/scitech/biblio/1222713>. 2014
- 3 Reinders J. Knights corner: your path to knights landing. Intel Developer Zone, 2014
- 4 Nvidia. Nvidia Tesla P100—the most advanced Datacenter Accelerator ever built featuring pascal GP100, the world's fastest GPU. Nvidia Whitepaper. Nvidia WP-08019-001_v01.1. 2016
- 5 Lee Y J, Kim J, Jan H, et al. A fully associative, tagless DRAM cache. In: Proceedings of the 42nd International Symposium on Computer Architecture (ISCA2015). New York: ACM, 2015. 211–222
- 6 Jang H, Lee Y, Kim J, et al. Efficient footprint caching for tagless DRAM caches. In: Proceedings of the 22nd IEEE Symposium on High Performance Computer Architecture (HPCA2016), Barcelona, 2016. 237–248
- 7 Bhati I, Chishti Z, Lu S-L, et al. Flexible auto-refresh: enable scalable and energy-efficient DRAM refresh reductions. In: Proceedings of the 42nd International Symposium on Computer Architecture (ISCA2015). New York: ACM, 2015. 235–246
- 8 Yu X Y, Davadas S. Tardis: time traveling coherence algorithm for distributed shared memory. In: Proceedings of the 24th International Conference on Parallel Architecture and Compilation Techniques (PACT2015), San Francisco, 2015. 227–240
- 9 Balasubramonian R, Grot B. Near-data processing. IEEE Micro, 2016, 36: 4–5
- 10 Balfour J, Dally W J. Design tradeoffs for tiled cmp on-chip networks. In: Proceedings of the 20th International Conference on Supercomputing (ICS'06). New York: ACM, 2006. 187–198
- 11 Kim J, Balfour J, Dally W J. Flattened butterfly topology for on-chip networks. In: Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'07), Chicago, 2007. 172–182

- 12 Demir Y, Hardavellas N. SLAC: stage laser control for a flattened butterfly network. In: Proceedings of the 22nd IEEE Symposium on High Performance Computer Architecture (HPCA2016), Barcelona, 2016. 321–332
- 13 Sodani A, Gramunt R, Corbal J, et al. Knight landing: second-generation Intel Xeon Phi product. IEEE Micro, 2016, 36: 34–46
- 14 Amaru L G. New data structure and algorithms for logic synthesis and verification. Dissertation for Ph.D. Degree. Lausanne: Ecole Polytechnique Federale de Lausanne, 2015
- 15 Borkar S. Thousand core chip—a technology perspective. In: Proceedings of the 44th Design Automation Conference (DAC2007), San Diego, 2007. 746–749
- 16 Keckler S W, Dally W J, Khailany B, et al. GPUs and the future of parallel computing. IEEE Micro, 2011, 31: 7–17
- 17 Yuffe M, Knoll E, Mehalel M, et al. A fully integrated multi-CPU, GPU and memory controller 32nm processor. In: Proceedings of IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, 2011. 264–266
- 18 Smith R, Goyal N, Ormont J, et al. Evaluating GPUs for network packet signature matching. In: Proceedings of IEEE International Symposium on Performance Analysis of Systems and Software, Boston, 2009. 175–184
- 19 Wang Z K, He B S, Zhang W, et al. A performance analysis framework for optimizing OpenCL applications on FPGAs. In: Proceedings of the 22nd IEEE Symposium on High Performance Computer Architecture (HPCA2016), Barcelona, 2016. 114–125
- 20 Gao M Y, Kozyrakis C. HRL: efficient and flexible reconfigurable logic for near-data processing. In: Proceedings of the 22nd IEEE Symposium on High Performance Computer Architecture (HPCA2016), Barcelona, 2016. 126–137
- 21 Komornicki A, Mullen-Schultz G, Landon D. Roadrunner: Hardware and software overview. USA: IBM Redbooks, 2009
- 22 Carter N P, Agrawal A, Borkar S, et al. Runnemed: an architecture for ubiquitous high-performance computing. In: Proceedings of the 19th IEEE International Symposium of High Performance Computer Architecture (HPCA2013), Shenzhen, 2013. 198–209
- 23 Mitsuhsa S. Feasibility Study on Future HPC Infrastructure. <http://www.ccs.tsukuba.ac.jp/files/ex-review/FS-ccs-eval-2014.pdf>. 2014
- 24 Homayoun H, Kontorinis V, Shayan A, et al. Dynamically heterogeneous cores through 3D resource pooling. In: Proceedings of the 18th IEEE International Symposium of High Performance Computer Architecture (HPCA2012), New Orleans, 2012. 277–288
- 25 Branover A, Foley D, Steinman M. AMD fusion APU: llano. IEEE Micro, 2012, 32: 28–37
- 26 Taylor M B. A landscape of the new dark silicon design regime. IEEE Micro, 2013, 33: 8–19
- 27 Merolla P A, Arthur J V, Alvarez-Icaza R, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science, 2014, 345: 668–673
- 28 Prezioso M, Merrih-Bayat F, Hoskins B D, et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. Nature, 2015, 521: 61–64
- 29 Khan M M, Lester D R, Plana L A, et al. SpiNNaker: mapping neural networks onto a massively-parallel chip multiprocessor. In: Proceedings of IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008. 2849–2856
- 30 Chen T, Du Z, Sun N, et al. Diannao: a small-footprint high-throughput accelerator for ubiquitous machine-learning. ACM Sigplan Notices, 2014, 49: 269–284
- 31 Shen J C, Ma D, Gu Z H, et al. Darwin: a neuromorphic hardware co-processor based on Spiking Neural Networks. Sci China Inf Sci, 2016, 59: 023401
- 32 Kahle J. The cell processor architecture. In: Proceedings of the 38th Annual IEEE/ACM International Symposium on Microarchitecture. Washington: IEEE Computer Society, 2005. 3
- 33 Nyberg P. The Cray rainier system: integrated scalar/vector computing. <http://www.ecmwf.int/sites/default/files/elibrary/2004/14161-the-cray-rainier-system-integrated-scalar-vector-computing.pdf>. 2004
- 34 Intel Xeon+FPGA Platform for the Data Center. The 4th Workshop on the Intersections of Computer Architecture and Reconfigurable Logic. <https://www.ece.cmu.edu/calcm/carl/lib/exe/fetch.php?media=carl15-gupta.pdf>. 2015
- 35 Wirbel L. Xilinx SDAccel—a Unified Development Environment for Tomorrow's Data Center. Technical Report, the Linley Group Inc, 2014

- 36 Shalf J, Quinlan D, Janssen C. Rethinking hardware-software codesign for exascale systems. *Computer*, 2011, 44: 22–30
- 37 Bertels K, Sima V M, Yankova Y, et al. HArtes: hardware-software codesign for heterogeneous multicore platforms. *IEEE Micro*, 2010, 5: 88–97
- 38 Kinsy M A, Devadas S. Heracles: a tool for fast RTL-based design space exploration of multicore processors. In: *Proceedings of the ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, Monterey, 2013. 125–134
- 39 Mavroidis I, Mavroidis I, Papaefstathiou I, et al. FASTCUDA: open Source FPGA Accelerator Hardware-Software Codesign Toolset for CUDA Kernels. In: *Proceedings of Euromicro Conference on Digital System Design*, Funchal, 2012. 343–348
- 40 Kim G, Lee M, Jeong J, et al. Multi-GPU system design with memory networks. In: *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*, Cambridge, 2014. 484–495
- 41 Kobayashi M, Seetharaman S, Parulkar G, et al. Maturing of OpenFlow and software-defined networking through deployments. *Comput Netw Int J Comput Telecommun Netw*, 2014, 61: 151–175
- 42 Huawei. High Throughput Computing Data Center Architecture—Thinking of Data Center 3.0. Technical White Paper, 2014
- 43 Minkenberg C, Rodriguez G. Trace-driven co-simulation of high-performance computing systems using OMNeT. In: *Proceedings of the 2nd International Conference on Simulation Tools and Techniques*, Belgium, 2009. 65
- 44 Mingyu H, Kevin P, Jie M, et al. SST + gem5 = a scalable simulation infrastructure for high performance computing. In: *Proceedings of the 5th International ICST Conference on Simulation Tools and Techniques*, Desenzano, 2012. 196–201
- 45 Zheng G, Kakulapati G, Kalé L V. BigSim: a parallel simulator for performance prediction of extremely large parallel machines. In: *Proceedings of the 18th International Symposium on Parallel and Distributed Processing*, Santa Fe, 2004. 78–87
- 46 Bhatele A, Jain N, Livnat Y, et al. Evaluating System Parameters on a Dragonfly using Simulation and Visualization. Technical Report, 2015
- 47 Besta M, Hoefler T. Slim fly: a cost effective low-diameter network topology. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC2014)*, New Orleans, 2014. 348–359
- 48 Kathareios G, Minkenberg C. Cost-effective diameter-two topologies: analysis and evaluation. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC2015)*, Austin, 2015. 36–46
- 49 Magnusson P S, Christensson M, Eskilson J, et al. Simics: a full system simulation platform. *IEEE Comput*, 2002, 2: 50–58
- 50 Alverson B, Froese E, Kaplan L, et al. Cray XC @Series Network. White Paper WP-Aries01-1112. <http://www.cray.com/Assets/PDF/products/xc/CrayXC30Networking.pdf>. 2012
- 51 Ajima Y, Inoue T, Hiramoto S, et al. Tofu interconnect 2: system-on-chip integration of high-performance interconnect. In: *Supercomputing*. Berlin: Springer, 2014. 498–507
- 52 Mellanox Technologies. Switch-IB™EDR Switch Silicon. Product Brief, 2014
- 53 Birrittella M S, Debbage M, Huggahalli R, et al. Intel®Omni-path architecture enabling scalable, high performance fabrics. In: *Proceedings of IEEE Symposium on High-performance Interconnects*, Santa Clara, 2015. 1–9
- 54 Tracy N, Wuth T. OIF Next Generation Interconnect Framework. OIF-FD- Client-400G/1T-01.0. <http://www.oiforum.com/public/documents/OIF-FD-Client-400G-1T-01.0.pdf>. 2013
- 55 Vinaik B, Puri R. Oracle's Sonoma Processor: Advanced Low-cost SPARC Processor for Enterprise Workloads. In: *Proceedings of IEEE Hot Chips 27 Symposium (HCS)*, Cupertino, 2015. 1–23
- 56 PLX. PCI Express Gen3 Switch. Product Brief. https://www.synopsys.com/dw/doc.php/ss/plx_ss.pdf. 2012
- 57 Regula J, Subramaniyan M, Dodson J. Integrating rack level connectivity into a PCIe switch. In: *Proceedings of International Symposium on High Performance Chips*, Stanford, 2013. 259–266
- 58 Kumar M J. Rack Scale Architecture for Cloud Keynote of Intel Developer Forum (IDF). <http://blog.scottlowe.org/2013/09/11/idf-2013-rack-scale-architecture-for-cloud/>. 2013

- 59 Assefa S, Xia F. Reinventing germanium avalanche photo-detector for nanophotonic on-chip optical interconnects. *Nature*, 2010, 464: 80–84
- 60 Lee B G, Rylyakov A V, Green W M J, et al. Monolithic silicon integration of scaled photonic switch fabrics, CMOS logic, and device driver circuits. *J Light Wave Tech*, 2014, 32: 743–751
- 61 Barwicz T, Taira Y, Lichoulas T W, et al. Photonic packaging in high-throughput microelectronic assembly lines for cost-efficiency and scalability. In: *Proceedings of Optical Fiber Communications Conference and Exhibition (OFC)*, Los Angeles, 2015. W3H4
- 62 Krishnamoorthy A V, Ho R, Zheng X, et al. Computer systems based on silicon photonic interconnect. *Proc IEEE*, 2009, 97: 1337–1361
- 63 Dobbelaere P D. Silicon photonics technology platform for integration of optical IOs with ASICs. In: *Proceedings of International Symposium on High Performance Chips*, Stanford, 2013. 115–123
- 64 Doany F E, Lee B G, Kuchta D M, et al. Terabit/Sec VCSEL-based 48-channel optical module based on holey CMOS transceiver IC. *J Light Wave Tech*, 2013, 31: 672–680
- 65 Sun C, Wade M T, Lee Y, et al. Single-chip microprocessor that communicates directly using light. *Nature*, 2015, 528: 534–538
- 66 Polatis. SERIES 7000-384×384 port software-defined optical circuit switch. Product Brief, 2016
- 67 Cheung S, Su T, Okamoto K, et al. Ultra-compact silicon photonic 512×512 25-GHz arrayed waveguide grating router. *IEEE J Sel Topics Quant Electron*, 2014, 20: 310–316
- 68 Yang X J. Sixty Years of Parallel Computing. *Comput Eng Sci*, 2012, 34: 1–10 [杨学军. 并行计算六十年. *计算机工程与科学*, 2012, 34: 1–10]
- 69 International technology roadmap for semiconductors (itrs). <http://www.itrs2.net/2013-itrs.html>
- 70 Tolentino M E, Turner J, Cameron K W. Memory MISER: improving main memory energy efficiency in servers. *IEEE Trans Comput*, 2009, 58: 336–350
- 71 Shinnar A, Cunningham D, Saraswat V, et al. M3R: increased performance for in-memory Hadoop jobs. *Proc VLDB Endowment*, 2012, 5: 1736–1747
- 72 Mittal S, Vetter J S. A survey of architectural approaches for data compression in cache and main memory systems. *IEEE Trans Parall Distrib Syst*, 2016, 27: 1524–1536
- 73 Lee B C, Zhou P, Yang J, et al. Phase-change technology and the future of main memory. *IEEE Micro*, 2010, 30: 143
- 74 Zhang L, Zhu M, Huang R, et al. Forming-less unipolar TaO_x-based RRAM with large CC-independence range for high density memory applications. *ECS Trans*, 2010, 27: 3–8
- 75 Li H, Xi H, Chen Y, et al. Thermal-assisted spin transfer torque memory (STT-RAM) cell design exploration. In: *Proceedings of IEEE Computer Society Annual Symposium on VLSI*, Tampa, 2009. 217–222
- 76 Corporation I. OpenMP Application Program Interface, version 4.5. OpenMP Architecture Review Board. <http://www.openmp.org/mp-documents/openmp-4.5.pdf>. 2015
- 77 Carter N P, Agrawal A, Borkar S, et al. Runnemed: an architecture for ubiquitous high-performance computing. In: *Proceedings of High Performance Computer Architecture (HPCA2013)*, Shenzhen, 2013. 198–209
- 78 Vetter J S, Mittal S. Opportunities for nonvolatile memory systems in extreme-scale high-performance computing. *Comput Sci Eng*, 2015, 17: 73–82
- 79 Xu W, Lu Y, Li Q, et al. Hybrid hierarchy storage system in MilkyWay-2 supercomputer. *Front Comput Sci*, 2014, 8: 367–377
- 80 Lee B C, Zhou P, Yang J, et al. Phase-change technology and the future of main memory. *IEEE Micro*, 2010, 30: 143
- 81 Carlson W W, Draper J M, Culler D E, et al. Introduction to UPC and Language Specification. Technical Report CCS-TR-99-157. 1999
- 82 Numrich R W, Reid J. Co-array fortran for parallel programming. *ACM Sigplan Fortran Forum*, 1998, 17: 1–31
- 83 Nieplocha J, Harrison R J, Littlefield R J. Global arrays: a portable shared-memory programming model for distributed memory computers. In: *Proceedings of the ACM/IEEE Conference on Supercomputing*, Los Alamitos, 1994. 340–349
- 84 Chamberlain B L, Callahan D, Zima H P. Parallel programmability and the chapel language. *Int J High Perfor*

- Comput Appl, 2007, 21: 291–312
- 85 Charles P, Grothoff C, Saraswat V, et al. X10: an object-oriented approach to non-uniform cluster computing. In: Proceedings of the 20th Annual ACM SIGPLAN Conference on Object Oriented Programming Systems Languages and Applications, New York, 2005. 519–538
- 86 Steele Jr G L, Allen E, Chase D, et al. Fortress (Sun HPCS language). In: Encyclopedia of Parallel Computing. Berlin: Springer, 2011. 718–735
- 87 Wienke S, Springer P, Terboven C, et al. OpenACC—first experiences with real-world applications. In: Proceedings of Euro-Par 2012 Parallel Processing, Pittsburgh, 2012. 859–870
- 88 David K. NVIDIA CUDA software and GPU parallel computing architecture. In: Proceedings of the 6th International Symposium on Memory Management, Vancouver, 2007. 103–104
- 89 Gaster B, Howes L, Kaeli D R, et al. Heterogeneous Computing with OpenCL: Revised OpenCL 1. London: Newnes, 2012
- 90 Amarasinghe S, Hall M, Lethin R, et al. ASCR programming challenges for exascale computing. In: Report of the 2011 Workshop on Exascale Programming Challenges, Marina del Rey, 2011
- 91 Hou Q, Zhou K, Guo B. BSGP: bulk-synchronous GPU programming. ACM Trans Graph, 2008, 27: 19
- 92 Chen L, Liu L, Tang S, et al. Unified parallel C for GPU clusters: language extensions and compiler implementation. In: Proceedings of International Workshop on Languages and Compilers for Parallel Computing. Berlin: Springer, 2010. 151–165
- 93 Yang C, Wang F, Du Y, et al. Adaptive optimization for petascale heterogeneous CPU/GPU computing. In: Proceedings of 2010 IEEE International Conference on Cluster Computing, Chemnitz, 2010. 19–28
- 94 Gong C Y, Bao W M, Liu H D, et al. A survey of the parallel solutions to convection-diffusion equation. Comput Eng Sci, 2015 37: 628–633 [龚春叶, 包为民, 刘海东, 等. 对流扩散方程并行求解方法研究综述. 计算机工程与科学, 2015, 37: 628–633]
- 95 Gong C, Liu J, Chi L, et al. GPU accelerated simulations of 3D deterministic particle transport using discrete ordinates method. J Comput Phys, 2011, 230: 6010–6022
- 96 Wang Q L, Liu J, Gong C Y, et al. Scalability of 3D deterministic particle transport on the Intel MIC architecture. Nucl Sci Tech, 2015, 26: 050502
- 97 Gong C, Bao W, Tang G. A parallel algorithm for the Riesz fractional reaction-diffusion equation with explicit finite difference method. Fract Calc Appl Anal, 2013, 16: 654–669
- 98 Gong C, Bao W, Tang G, et al. An efficient parallel solution for Caputo fractional reaction–diffusion equation. J Supercomput, 2014, 68: 1521–1537
- 99 Yang B, Lu K, Gao Y, et al. GPU acceleration of subgraph isomorphism search in large scale graph. J Central South Univ, 2015, 22: 2238–2249
- 100 Wu Q, Yang C, Tang T, et al. Exploiting hierarchy parallelism for molecular dynamics on a petascale heterogeneous system. J Parall Distrib Comput, 2013, 73: 1592–1604
- 101 Xu C, Deng X, Zhang L, et al. Collaborating CPU and GPU for large-scale high-order CFD simulations with complex grids on the TianHe-1A supercomputer. J Comput Phys, 2014, 278: 275–297
- 102 Barker K J, Davis K, Hoisie A, et al. Entering the petaflop era: the architecture and performance of Roadrunner. In: Proceedings of the ACM/IEEE Conference on Supercomputing. Piscataway: IEEE Press, 2008
- 103 Gong C, Liu J, Huang H, et al. Particle transport with unstructured grid on GPU. Comput Phys Commun, 2012, 183: 588–593
- 104 Yan J, Tan G M, Sun N H. Optimizing parallel S n sweeps on unstructured grids for multi-core clusters. J Comput Sci Tech, 2013, 28: 657–670
- 105 Gong C, Bao W, Liu J, et al. An efficient wavefront parallel algorithm for structured three dimensional LU-SGS. Comput Fluids, 2016, 134: 23–30
- 106 Ghysels P, Vanroose W. Hiding global synchronization latency in the preconditioned conjugate gradient algorithm. Parall Comput, 2014, 40: 224–238
- 107 Liu Y, Yang C, Liu F, et al. 623 Tflop/s HPCG run on Tianhe-2: Leveraging millions of hybrid cores. Int J High Perfor Comput Appl, 2016, 30: 39–54
- 108 Liu J, Gong C, Bao W, et al. Solving the Caputo fractional reaction-diffusion equation on GPU. Discrete Dyn Nat

- Soc, 2014, 2014: 1–7
- 109 Jia W, Fu J, Cao Z, et al. Fast plane wave density functional theory molecular dynamics calculations on multi-GPU machines. *J Comput Phys*, 2013, 251: 102–115
  - 110 Liu Y Q, Li Y, Zhang Y Q, et al. Memory efficient two-pass 3D FFT algorithm for Intel®Xeon Phi™ coprocessor. *J Comput Sci Tech*, 2014, 29: 989–1002
  - 111 刘鑫. 面向化学非平衡流的 CFD 并行计算技术和大规模并行计算平台研究. 博士学位论文. 郑州: 解放军信息工程大学. 2006
  - 112 Lee M, Malaya N, Moser R D. Petascale direct numerical simulation of turbulent channel flow on up to 786k cores. In: *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, Denver, 2013. 61
  - 113 Himeno R. Grand challenge in life science on K computer. In: *Proceedings of International Conference on High Performance Computing for Computational Science*, Kope, 2013. 17–22
  - 114 Lei X L, Zhang T, Zhao Y, et al. The Application on Large-Scale Scientific Computing and Complex Engineering Based on TH-1A. *Comput Eng Sci*, 2012, 34: 176–183 [雷秀丽, 张婷, 赵洋, 等. “天河一号” 大规模科学与工程计算应用. *计算机工程与科学*, 2012, 34: 176–183]
  - 115 Gong C Y, Bao W M, Tang G J. Recent progress in high-performance parallel computing of the aerospace area. *Comput Eng Sci*, 2014, 36: 1629–1636 [龚春叶, 包为民, 汤国建, 等. 航天领域高性能并行计算研究进展. *计算机工程与科学*, 2014, 36: 1629–1636]
  - 116 Bermejo-Moreno I, Bodart J, Larsson J, et al. Solving the compressible Navier-Stokes equations on up to 1.97 million cores and 4.1 trillion grid points. In: *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. New York: ACM, 2013. 62
  - 117 Che Y, Xu C, Fang J, et al. Realistic performance characterization of CFD applications on Intel many integrated core architecture. *Comput J*, 2015, 58: 3279–3294
  - 118 Breuer A, Heinecke A, Rettenberger S, et al. Sustained petascale performance of seismic simulations with seissol on supermuc. In: *Proceedings of Supercomputing Conference*, New Orleans, 2014. 1–18
  - 119 Coates A, Huval B, Wang T, et al. Deep learning with COTS HPC systems. In: *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, 2013. 1337–1345
  - 120 Shaw D E, Grossman J P, Bank J A, et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, New Orleans, 2014. 41–53
  - 121 Staar P, Maier T A, Summers M S, et al. Taking a quantum leap in time to solution for simulations of high-Tc superconductors. In: *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, Denver, 2013. 1
  - 122 Bernaschi M, Bisson M, Fatica M, et al. 20 petaflops simulation of proteins suspensions in crowding conditions. In: *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, Denver, 2013. 2
  - 123 Randles A, Draeger E W, Oppelstrup T, et al. Massively parallel models of the human circulatory system. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, Austin, 2015. 1
  - 124 Berczik P, Spurzem R, Zhong S, et al. Up to 700k GPU cores, Kepler, and the Exascale future for simulations of star clusters around black holes In: *Supercomputing*. Berlin: Springer, 2013. 13–25
  - 125 Rossinelli D, Hejazialhosseini B, Hadjidoukas P, et al. 11 PFLOP/s simulations of cloud cavitation collapse. In: *Proceedings of High Performance Computing, Networking, Storage and Analysis (SC)*, Denver, 2013. 1–13
  - 126 Bussmann M, Burau H, Cowan T E, et al. Radiative signatures of the relativistic Kelvin-Helmholtz instability. In: *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, Denver, 2013. 5
  - 127 Bédorf J, Gaburov E, Fujii M S, et al. 24.77 pflops on a gravitational tree-code to simulate the Milky Way galaxy with 18600 GPUs. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, New Orleans, 2014. 54–65
  - 128 Dongarra J J, Du Croz J, Hammarling S, et al. A set of level 3 basic linear algebra subprograms. *ACM Trans Math*

- Softw, 1990, 16: 1–17
- 129 Balay S, Abhyankar S, Adams M, et al. PETSc Users Manual Revision 3.5. Technical Report, ANL-95/11 Rev. 3.5 108682. 2014
- 130 Blackford L S, Choi J, Cleary A, et al. ScaLAPACK users' guide. Siam, 1997
- 131 Frigo M, Johnson S G. FFTW: an adaptive software architecture for the FFT. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal, Seattle, 1998. 3: 1381–1384
- 132 Mo Z, Zhang A, Cao X, et al. JASMIN: a parallel software infrastructure for scientific computing. Front Comput Sci China, 2010, 4: 480–488
- 133 Jasak H, Jemcov A, Tukovic Z. OpenFOAM: a C++ library for complex physics simulations. In: Proceedings of International Workshop on Coupled Methods in Numerical Dynamics, Dubrovnik, 2007. 1000: 1–20
- 134 Nelson M T, Humphrey W, Gurosoy A, et al. NAMD: a parallel, object-oriented molecular dynamics program. Int J High Perfor Comput Appl, 1996, 10: 251–268
- 135 Plimpton S, Crozier P, Thompson A. LAMMPS-large-scale atomic/molecular massively parallel simulator. Sandia National Laboratories, 2007, 18
- 136 Berendsen H J C, van der Spoel D, van Drunen R. GROMACS: a message-passing parallel molecular dynamics implementation. Comput Phys Commun, 1995, 91: 43–56

Emerging High-Performance Computing Systems and Technology

Xiangke LIAO* & Nong XIAO

School of Computer, National University of Defense Technology, Changsha 410000, China

*E-mail: xkliao@263.net

Abstract High-performance computing (HPC) technology is the strategic height of modern computing among countries in the current information era. This paper focuses on the challenges of emerging HPC technology in the future, and discusses the developing status and trends of HPC architecture, algorithms, and applications. From the viewpoint of foundation research, we propose significant research fields (or scientific issues), as well as related policies for HPC systems and technology in China.

Keywords high-performance computing, microprocessor, computer architecture, high-performance computing application

is the principal investigator and chief designer of the Tianhe-2 supercomputer.

Xiangke LIAO received his B.S. degree from Tsinghua University, China, in 1985, and his M.S. degree from the National University of Defense Technology (NUDT), China, in 1988, both in Computer Science. Currently, he is a professor and the dean of the College of Computer, NUDT. His research interests include high performance computing systems, operating systems, and parallel and distributed computing.

rent research interests include network computing, large-scale storage systems, and computer architecture.

Nong XIAO received his B.S., M.S., and Ph.D. degrees in Computer Science from the National University of Defense Technology (NUDT), China. He is now a professor in the State Key Laboratory of High Performance Computing (HPCL), NUDT. He was named the "Chang Jiang Scholar" by the Chinese Ministry of Education, and the "Distinct Young Scholar" by the Chinese National Science Foundation. His current research interests include network computing, large-scale storage systems, and computer architecture.