“神威·太湖之光”计算机系统大规模应用特征分析与 E 级可扩展性研究

刘 鑫[1)] 郭 恒[1)] 孙茹君[2)] 陈左宁[1)]

[1)](国家并行计算机工程技术研究中心 江苏 无锡 214083)

[2)](数学工程与先进计算国家重点实验室 江苏 无锡 214125)

摘 要 复杂应用系统面临着全系统、全物理过程、自然尺度的计算模拟,对计算机能力提出更高要求.该文介绍了“神威·太湖之光”系统半机以上超大规模并行应用的算法特点、体系结构适应性、计算复杂度、访存复杂度和通信复杂度的大规模实验分析结果,基于大规模应用计算和数据移动特征以及异构众核体系结构特点提出新的性能模型,得出影响大规模应用性能的关键因素,提出 E 级复杂应用对未来 E 级计算机系统的设计需求.

关键词 神威·太湖之光;大规模应用;复杂度分析;计算特征

中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2018.02209

The Characteristic Analysis and Exascale Scalability Research of Large Scale Parallel Applications on Sunway TaihuLight Supercomputer

LIU Xin[1)] GUO Heng[1)] SUN Ru-Jun[2)] CHEN Zuo-Ning[1)]

[1)](*National Research Centre of Parallel Computer Engineering and Technology*, *Wuxi*, *Jiangsu* 214083)

[2)](*State Key Laboratory of Mathematical Engineering and Advanced Computing*, *Wuxi*, *Jiangsu* 214125)

Abstract Complex application system is faced with large-computing simulation of the whole system, the whole physical process, true three-dimension and natural scale, which put forward higher requirements for the supercomputer's ability. Most large-scale applications of Sunway TaihuLight supercomputer are the largest scale of the correspondent field that partly represents the characteristics of the application. This paper mainly analysis the calculation characteristics and data migration behavior of the semi-scale and full-scale applications. First, we provide a brief introduction to the Sunway TaihuLight system, the architecture of the homegrown many-core SW26010 processor and some parallel programming methods and architecture-related optimization methods to supports large-scale parallel applications development. According to classification criteria of the University of California, Berkeley, we analysis the applications of ten computing themes such as dense linear algebra, sparse linear algebra, spectral methods, N-body methods, structured grids, unstructured grids, Map-Reduce, graph traversal and dynamic programming. Focusing on the characteristics of the algorithm, the adaptability of the architecture, the algorithm

收稿日期:2017-07-31;在线出版日期:2018-01-29.本课题得到“全球变化和应对”专项基金(2016YFA0602200)、国家“九七三”重点基础研究发展规划项目基金(2014CB744100)资助.**刘 鑫**,女,1979 年生,博士,副研究员,中国计算机学会(CCF)会员,主要研究方向为并行算法和应用.E-mail: yyylx@263.net.**郭 恒**,男,1993 年生,博士研究生,工程师,主要研究方向为并行算法和应用.**孙茹君**,女,1990 年生,博士研究生,工程师,主要研究方向为计算机体系结构和并行计算模型.**陈左宁**,女,1957 年生,高级工程师,博士生导师,中国工程院院士,主要研究领域为高性能计算机体系结构和操作系统.

complexity, the space complexity, the characteristics of memory access and the communication complexity, we get the bottlenecks of the application algorithms extended to the exascale. Based on the above analysis and the architecture characteristics of SW26010, we first propose a new performance model of one core group to help efficient algorithm designing on core group for different applications. From this model, the program designers should increase the execution efficiency of the CPU's computing units, improve the memory access bandwidth of real applications and reduce the amount and times of communication. For large-scale parallel applications, we also give a modified performance model of large problems. For most memory intensive applications, how to reduce the amount and times of discrete memory access, improving the bandwidth of direct memory access (DMA), making full use of local data memory of each CPU core is the key problem to get better performance. And the most important is how to take full advantage of the register communication mechanism of the many-core SW26010 processor on-chip mesh to improve the efficiency of memory accessing. For large full-scale applications, whether the amount of calculation and the cost of memory access and communication could increase linearly with the increasing of problem size is critical to extend to exascale. Based on this, we divide these applications into two categories, applications with regular computing and data migration and applications with irregular computing and data migration. The space and time complexity of regular applications increase with the expansion of the problem scale, and the patterns of memory access and communication are regular. And the computing to memory access ratio and computing to communication ratio are relatively high, which can get reasonable scalability and parallel efficiency. The irregular applications are usually with irregular memory access or communication and have low computing to memory access ratio and computing to communication ratio. The cost of memory access and communication increases sharply with the increasing of problem size, which results to complex data exchange and poor speedup. Based on these, the paper provides the suggestions for the exascale super computer system. For regular applications, polymorphic and multi-scale systems are needed to map different subsystem of complex systems and multi-level large-capacity storage hierarchy is required to support regular discrete access and efficient on-chip data sharing, and faster and lower -latency network is needed, which are some evolutional architecture changes. For irregular applications, innovative memory controller, on-chip inter-connect and on-chip cache are needed to increase memory access efficiency and some new architecture for efficient fine-grained parallelism should be explored, which are some radical architecture innovations.

Keywords Sunway TaihuLight supercomputer; large parallel application; complexity analysis; computing characteristics

1 引 言

"神威·太湖之光"系统[1-2]自投入使用以来,完成上百家用户单位,数百项大型复杂应用课题的计算,涉及天气气候、航空航天、海洋环境、生物医药、船舶工程等19个应用领域,实现了数百万核超大规模并行,其中整机应用17个,半机以上规模应用12个,百万核以上应用二十余个,基于该系统的三项应用[3-5]入围2016年度戈登贝尔奖,最终一项应用获奖;基于该系统的两项应用[6-7]入围2017年度戈登贝尔奖.从大部分应用可以看出,当前的实际复杂应用系统向着多时空尺度、强非线性耦合和三维真实构型的方向发展,包含着大量多尺度多模型的计算问题,存在多粒度、多维度、多层次的并行性,面临着全系统、全物理过程、真三维、自然尺度的计算

模拟，对计算机的能力提出更高要求.

"神威·太湖之光"系统诸多大型应用均是各应用领域的最大规模，具有一定代表性，本文主要针对该系统半机以上规模、计算密集的重大应用进行计算特征和数据迁移行为的分析，重点关注算法特点、体系结构适应性、算法的时间复杂度、空间复杂度、访存特点、通信复杂度[8]等特征[9]，分析各类应用算法扩展到 E 级时可能会遇到的瓶颈问题，并针对性能瓶颈问题提出了下一代 E 级高性能计算机系统体系结构需求和设计建议.

根据美国加州大学伯克利分校的对科学与工程计算应用的分类标准[10]，我们对各应用分类如下：(1) 稠密线性代数[11-12]，如 LINPACK、大规模流固耦合和流声耦合计算、潜艇收发分置全向声散射特性等；(2) 稀疏线性代数[13]，如高超声速飞行器数值模拟、C919 大型客机失速特性模拟等；(3) 谱方法，如基于 FFT 的湍流直接数值模拟、BNU_ESM 地球系统模式等；(4) 多体问题，如 GROMACS[14]、NAMD[15]、MD 模拟等；(5) 结构网格，如高超声速飞行器数值模拟、可压缩边界层湍流直接数值模拟、地球系统模式[6,16]、地震模拟[5]等；(6) 非结构网格[13]，如航空发动机数值模拟、污染排放模拟、人体血流模拟等；(7) MapReduce，如蒙特卡罗模拟期权定价、托卡马克装置逃逸电子行为模拟[17]等；(8) 组合逻辑，如 AES[18]、MD5 等；(9) 图的遍历[19]，如社交网络分析等；(10) 动态规划，如精确基因序列比对分析[20]等；(11) 回溯和分支限界，如 SAT 代数攻击等；(12) 图的模型，如人工神经网络[21]、隐马尔可夫模型等；(13) 有限状态机，如网络协议分析等应用. 以上十三类应用均在"神威·太湖之光"计算机系统上完成计算.

2　"神威·太湖之光"系统体系结构

"神威·太湖之光"计算机系统[22]采用基于高密度弹性超节点和高流量复合网络架构和面向多目标优化的高效能体系结构，系统由 40 960 块"申威 26010"异构众核处理器组成，通过计算插件板、计算超节点和计算机仓等模式进行系统扩展，如图 1 所示，构成 125.436 PFLOPS 高速计算系统.

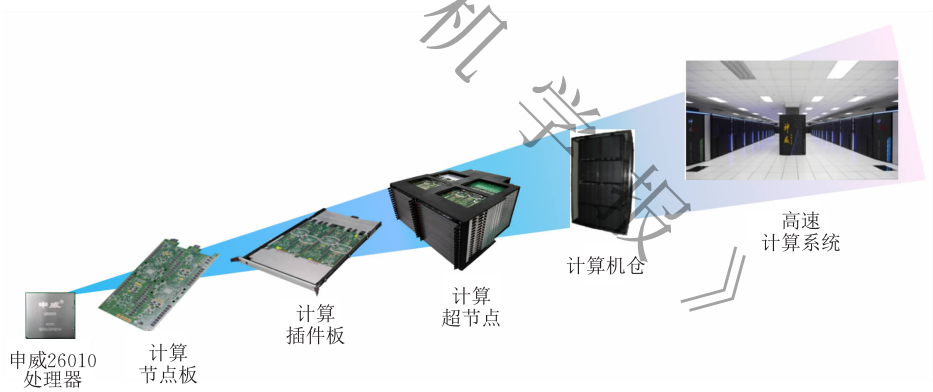

图 1　系统体系结构扩展示意图

"申威 26010"异构众核处理器采用片上计算阵列集群和分布式共享存储相结合的异构众核体系结构，单处理器芯片集成 4 个运算核组共 260 个运算核心，每个核组包含 1 个运算控制核心(主核)和 1 个运算核心阵列(从核阵列)，如图 2 所示. 采用寄存器级数据通信、多模式异步数据流传输和运算阵列快速同步等技术提高运算核心协同执行效率. 每个众核处理器配置 32 GB 内存，每核组本地内存为 8 GB；运算核心可以直接离散访问主存，也可以通过 DMA 方式批量访问主存，运算核心阵列之间可以采用寄存器通信方式进行通信；每个运算核心的局部存储空间大小为 64 KB，指令存储空间为 16 KB.

大部分科学与工程计算应用采用消息传递并行编程模型和共享变量并行编程模型的两级并行方式进行大规模并行，即进程级的 MPI 并行和线程级的 OpenACC[23]或加速线程库 Athread 并行. 应用性能优化方法主要有：利用众核处理器体系结构特点实现众核线程级的任务并行、数据并行和流水线并行的混合并行，提高众核并行效率；充分利用 DMA 批量访存和片上高效通信提高访存性能；利用指令流水、乘加优化和短向量优化等方法提高计算性能.

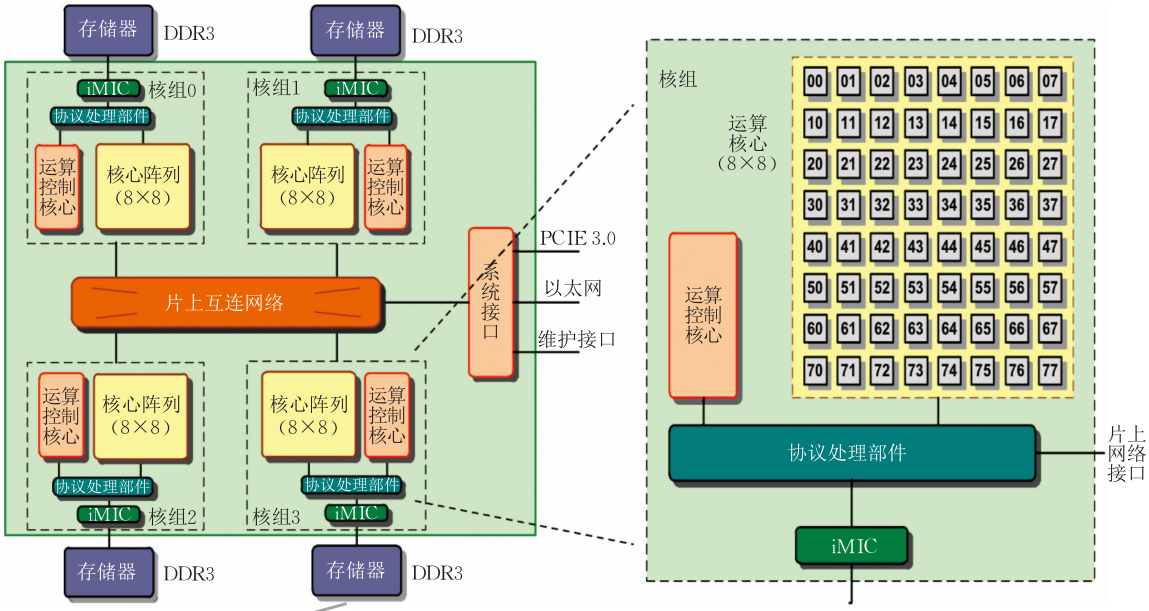

图 2 “申威 26010”异构众核处理器架构图

3 超大规模科学与工程应用分析

本文对十三类计算主题的半机以上规模应用进行计算特征和数据迁移行为分析(因组合逻辑、回溯分支限界和有限状态机类应用并行规模较小,不在分析范围内),具体如下.

3.1 稠密线性代数

代表应用主要有大规模流固耦合和流声耦合计算、潜艇收发分置全向声散射特性、LINPACK 等. 以 LINPACK[①] 为例说明稠密线性代数类问题,“神威·太湖之光”系统 LINPACK 求解矩阵规模为 1228.8 万,持续运算速度 93.015 PFLOPS,浮点效率为 74.153%,E 级系统预计求解矩阵规模为 3000 万以上. 在体系结构方面,系统支持片上阵列寄存器通信、行模式 DMA,使访存带宽需求降为 1/4 以下;此外,可以利用运算控制核心和运算核心的异步设计隐藏大规模并行的通信开销,应用效率提高 10% 以上.

设矩阵规模为 N,进程数为 Np,则 LINPACK 求解时间复杂度为 $2/3\times N^3$,空间复杂度为 $O(N^2)$,计算过程中基本为规律内存访问,访存方式为连续访存和跨步访存;部分数组计算存在有规律的离散访存,可通过数组转置方法将离散访存转换为连续访存. 在通信复杂度方面,一般情况下若矩阵规模扩大为 $2\times N$,则处理器规模对应变为 $4\times Np$($Np=Nx\times Ny$,Nx、Ny 为行列方向进程数),行列方向进程数增加一倍变为 $2Nx$ 和 $2Ny$,列方向上通信步数由 $\log(Nx)$ 变为 $\log(2\times Nx)$,行方向上通信步数增加一倍(视算法优化程度定),进程间通信量变化不大. 可以看出,随着求解问题规模的扩大,算法时间复杂度呈立方增长,空间复杂度呈平方增长,通信复杂度线性增长;同时,该类应用的计算访存比和访存通信比相对较高,具有较好的可扩展性和并行效率,对计算能力需求较为突出.

3.2 稀疏线性代数

代表应用主要有高超声速飞行器数值模拟、C919 大型客机失速特性模拟等. 稀疏矩阵求解问题因矩阵存储方式、矩阵类型、求解方法千差万别,故这里我们以 HPCG[②] 为例说明稀疏线性代数求解类问题,HPCG 由 LINPACK 的设计者 Jack Dongarra 提出,作为 HPL 测试的补充,期望更全面地反映典型应用程序的实际性能. “神威·太湖之光”系统完成 HPCG 稀疏矩阵 3435 亿元的计算,浮点性能达 480 TFLOPS,采用基于寄存器通信机制的片上数据共享,计算、寄存器通信和 DMA 重叠的方式进行优化.

设矩阵非零元素个数为 N,进程数为 Np,HPCG 时间复杂度基本随问题规模的扩大接近线性(略超线性)增加,空间复杂度为 $O(N)$,计算过程中的矩阵向量乘存在大量离散访存(有规律但间隔较远接近随机访存),访存开销巨大. 在通信复杂度方面,HPCG 通信分邻居通信和全局规约两类,由于使用

① Top500. http://www.top500.org/
② HPCG-BenchMark. http://www.hpcg-benchmark.org/

27 点 stencil，所以每个进程邻居通信个数固定为 26 个，全局规约近似为 $\log(Np)\times$某常数. 随着求解问题规模的扩大，可能迭代步数会略有增加，单迭代步内部通信复杂度变化不大.

需要注意的是，稀疏矩阵求解问题因矩阵存储方式、矩阵类型、求解方法千差万别，CSR 存储方式(HPCG 使用)在大部分情况下性能不错，但对于对角化比较好的稀疏矩阵，DIA 存储方式将可能会获得更为优秀的性能(包括存储空间开销和访存开销). 当矩阵比较规则(每一行的元素个数基本相同，例如三对角对阵等)时，使用 ELL 压缩方式也将会比 CSR 获得更加优异的存储性能，访存开销也会减少，从而提高程序的整体性能. 因此，矩阵对角化很好时，选择 DIA 存储方式或 ELL 存储方式都将获得不错的效果；矩阵较为规则，每一行非零元个数差别不大时，选择 ELL 存储方式性能可能较优；一般的系数矩阵，选择目前使用最为普遍的 CSR 格式更为合适. 总体来说，不管采用哪种存储方式，该类应用计算过程中存在大量离散随机访存，计算访存比低，随着求解问题规模的扩大，亟需提高访存性能.

3.3　谱方法

本节我们以 FFT 为例说明该类问题，"神威·太湖之光"采用 16 384 个处理器求解了 16 384 立方规模的三维 FFT，每步计算时间为 97 s，是目前世界上最大规模的 FFT 类问题.

假设 FFT 基本长度为 N，则一维 FFT 的算法时间复杂度为 $O(N\log N)$，二维 FFT 的算法时间复杂度为 $O(N^2\log N)$，三维 FFT 的算法时间复杂度为 $O(N^3\log N)$. 在空间复杂度方面，二维 FFT 的空间复杂度为 $O(N^2)$，三维 FFT 的空间复杂度为 $O(N^3)$. 计算过程中访存较为规整，主要是读入和写出 FFT 输入输出数组，从核内部数据交换主要是蝶式计算时从核间的寄存器通信. 在通信复杂度方面，我们以通信最复杂的三维 FFT 为例进行分析，假设 FFT 长度为 N^3，处理器数为 Np(三个方向上的处理器数分别为 Nx、Ny、Nz)，若求解问题规模 N 扩大为原来的 2 倍，则处理器规模对应变为 $8\times Np$，三个方向处理器数变为 $2Nx$、$2Ny$ 和 $2Nz$，每个计算步骤的 AllToAll 通信规模增加一倍，每个进程相互通信个数增加到原来的 8 倍，进程间通信量变化不大.

综上，随着求解问题规模的扩大，三维 FFT 算法时间和空间复杂度呈立方增长，通信复杂度呈线性增长；因应用本身是通信密集型问题，规模扩大后通信墙问题非常突出.

3.4　结构网格

结构网格类应用非常多，因基于结构网格的大多数常微分方程和偏微分方程求解最终归结为稀疏线性代数方程组和稠密线性代数方程组求解，这两类问题已详细讨论过，所以结构网格问题主要以 Stencil 计算为例. 利用"神威·太湖之光"整机系统，中国科学院软件研究所、清华大学联合开发的全球大气非静力全隐求解器，其中显式方法模块[24]完成网格数达 5152 亿算例的计算，浮点性能达 25.96 PFLOPS.

设 N 是网格点或网格单位总数，则三维问题的时间复杂度和空间复杂度均为 $O(N)$，因结构网格计算是规律内存访问，访存方式基本为连续访存和跨步访存；部分计算存在有规律的离散访存，可以通过数组转置方法将离散访存转换为连续访存. 在通信复杂度方面，随着网格规模的扩大，一般边缘通信的通信对数无变化，最大为 26(三维区域分解时，立方体的六个面、十二条棱和八个角点都需要通信)，通信长度随网格规模扩大线性增加.

综上，随着网格规模的扩大，结构网格 Stencil 计算的时间复杂度、空间复杂度线性增长，可以通过增大并行规模解决. 但值得注意的是，为保证物理上的收敛性，求解问题时间步长随着网格精细化程度的提高需相应减小，单纯通过提高并行规模对网格精细问题的整体求解速度提升不明显.

3.5　非结构网格

非结构网格在工程计算软件中使用越来越广泛，"神威·太湖之光"上该类问题的典型应用主要有航空发动机数值模拟、污染排放模拟等. 大部分非结构网格问题的计算方法与结构网格类似，不同的是非结构网格因数据存放的无序性导致内存访问的随机性. 以计算流体力学问题为例，"神威·太湖之光"上最大网格规模为燃烧问题数亿网格、完全气体问题百亿网格，系统级优化采用基于寄存器通信的数据重排、计算访存重叠、向量化等方法，预计 E 级求解规模约在千亿至万亿网格左右.

设 N 是网格点或网格单位总数，Np 是处理器个数，则某三维问题的复杂性分析如下：完全气体问题的计算复杂度一般为 $O(N)$，燃烧问题的计算复杂度涉及到多种化学反应计算，复杂度一般是化学反应组分数目的多项式复杂度. 空间复杂度一般为 $O(N)$，但因非结构网格存放的无序性，基本是内存离散访问，需要对网格单元和网格面进行重新排序

保证数据访问的连续性,或者增加冗余数据结构保证数据访问的连续性.通信复杂度取决于区域分解效果,一般来说,边缘通信的通信对数不会随并行规模和网格规模而变化,而通信长度随问题规模扩大线性增长.

随着网格规模的扩大,计算复杂度、空间复杂度线性增长,可以通过增大并行规模解决.但与结构网格类似,为保证物理上的收敛性,求解时间步长随着网格精细化程度的提高需相应减小,单纯通过提高并行规模对网格精细问题的整体求解速度提升不明显.

3.6 N-body 问题

多体问题类型较多,完整未经简化的多体问题如分子动力学领域的静电力计算和天体引力计算,需要计算所有粒子的相互作用力;大部分多体问题会针对不同的研究体系进行针对性的作用力计算简化或算法优化,比如只考虑一定范围距离内粒子间的相互作用[25].

中国科学院国家天文台在"神威·太湖之光"系统上完成了 11.2 万亿粒子宇宙演化的 N 体模拟解算,提出一种通过粒子网格方法(PM)和快速多极子方法(FMM)计算重力的混合方案,重力计算分为不同的尺度使全局通信被解耦,且能够实现计算和通信的灵活隐藏,最终平均性能达到 21.3 PFLOPS.日本理化学研究所计算科学研究机构(RIKEN AICS)在"神威·太湖之光"系统上完成了高达 1 万亿粒子的行星环模拟,基于粒子仿真框架 FDPS 使用 Barnes-Hut 树算法进行计算,采用域分解与自动负载平衡措施实现大规模并行,浮点性能约为理论峰值的 11%,即 13.75 PFLOPS.中国科学院过程工程研究所开展的非平衡分子动力学计算的模拟体系原子数目达到了 20 亿量级,单一方向空间特征尺度达到 500 μm 以上,浮点性能约为理论峰值的 15%,即 18.75 PFLOPS.

设 N 为粒子数,Np 为处理器数,完整未经简化多体问题的时间复杂度为 $O(N\times N)$,采用优化算法(如树状代码算法等)后时间复杂度降为 $O(N\times\log N)$;空间复杂度一般为 $O(N)$;通信复杂度为 $O(Np\times Np)$,采用优化算法后通信复杂度降为 $O(Np\times\log Np)$.

综上,简化多体问题随着网格规模和粒子数的扩大,计算复杂度、空间复杂度线性增长,可以通过增大并行规模解决;完整多体问题虽然通过算法改进能够适当降低通信复杂度,但问题规模扩大后仍需关注通信扩展难问题.

3.7 MapReduce

MapReduce 问题的算法特点是大量计算任务无相关性、可以并行执行.目前"神威·太湖之光"上该类课题主要有高通量药物虚拟筛选、中子输运过程模拟、托卡马克装置等离子体几何算法粒子模拟等.其复杂度与具体应用密切相关,这里以托卡马克装置等离子体几何算法粒子模拟为例,该应用采用几何算法计算每个粒子在电磁场中的运动,每计算核心负责一个粒子的计算,粒子间基本不需要通信."神威·太湖之光"系统完成聚变实验堆(ITER)逃逸电子 10^{18} 粒子时间步的模拟即 10^7 个粒子采样点、每个粒子迭代 10^{11} 步的计算,浮点性能约为理论峰值的 10%左右.

设粒子数为 N,时间迭代步数为 M,全局粒子信息收集次数为 L,算法的时间复杂度取决于粒子数和时间迭代步数,基本与问题规模呈线性关系.在空间复杂度方面,每个粒子有固定的临时内存空间,空间复杂度为 $O(N)$.除初始化过程外,粒子迭代所需内存常驻从核局部存储空间中,访存效率较高.在通信复杂度方面,该算法计算过程中需要在某些时刻收集粒子的全局信息,通信次数 L 一般远小于粒子数 N 和迭代步数 M.单次通信复杂度与粒子数 N 成正比;总通信开销与 N、L 线性相关,一般可忽略.

综上,该应用计算时间复杂度随粒子数 N 和时间步数 M 的扩大线性增加,空间复杂度随粒子数 M 的扩大线性增加,通信开销基本可忽略.该类应用的访存和通信占比较低,具有较好的可扩展性,对计算能力需求较为突出.

3.8 图的遍历

典型应用如社交网络分析等,一般图算法主要包括深度优先搜索(DFS)和宽度优先搜索(BFS).BFS 是 Graph500① 中的重要算法,能够反映计算机计算和访存的综合效率,这里以宽度优先搜索算法为例说明图遍历问题.宽度优先搜索(BFS)算法在初始状态所有点标记未读,从一个起始点开始,访问当前点的所有邻居节点并标记已读,记录新标记的点以在下一轮访问中作当前点,直到所有点标记为已读,算法终止.

"神威·太湖之光"上 Graph500 的 BFS 图规模为 2^{40}(顶点数),使用 40 768 个节点实现 23 755.7GTEPS

① Graph500. http://www.graph500.org/

的性能. 系统级优化主要在消息聚合、转发以及以充分利用带宽为目标的多任务流水作业等方面.

设图的点数为$|V|$,边数为$|E|$,则 BFS 时间复杂度为$O(|V|+|E|)$,Graph500 中使用 Kroneckor 生成器生成的图,一般设置$|E|=4\times|V|$,而稠密图的$|E|$可以达到$|V|^2$的规模. 空间复杂度为$O(|V|+|E|)$,访存主要包括本地新顶点搜索、消息打/解包、写父顶点三部分:搜索新顶点一般为连续访问;对边列表的访问属于总体离散、局部连续,查找哪些点的边列表是离散访存,每个点的边列表连续存放(CSR 格式);写父顶点为离散访问,多个线程间需要使用原子操作竞争写父顶点. 在通信复杂度方面,设 BFS 算法在 D 轮之后结束(D 为图直径),总通信频次为$O(D\times Np^2)$,采用行列聚合策略后可使通信频次降低至$O(D\times Np)$.

综上,该类应用随着计算规模增大,计算复杂度线性增长,空间复杂度线性到平方增长且存在大量离散访存,通信复杂度与图的分布有关,若按照最大通信量估算,则通信复杂度为线性到平方增长. 可以看出,该类应用访存和通信密集且无规律,问题规模扩大后访存和通信将成为性能瓶颈.

3.9　动态规划

生物序列比对是生物信息学中最常见的问题之一,采用动态规划思想完成. 基于动态规划思想的序列比对并行算法一般采用分而治之的方法把参考序列划分为若干片段,并分配给相应的各个处理器,而后并行地按各具体算法与目标序列进行比对,再通过按一定规则的扩展过程求取序列的优化匹配. 算法过程分为索引阶段(一般预先建好),匹配阶段(占整体计算量 80%)和比对阶段(占整体计算量 20%).

"神威·太湖之光"上序列匹配问题的参考序列使用了 Hg8,Hg13,Hg19 等十三个人类基因,每一个基因包含了 24 条染色体,目标序列是实测数据 HG098. 系统级优化采用计算和访存的互相隐藏、向量化、指令流水优化等方法. 假设目标序列长度为 N,时间复杂度为多项式复杂度$O(p(N))$,空间复杂度为$O(N)$,计算过程中基本为连续访存. 在通信复杂度方面,除计算开始时主进程向从进程的任务分配以外,各进程间基本无通信,动态规划计算本身在进程内完成,与并行规模关系不大. 该类应用具有较好的可扩展性,对计算能力需求较为突出.

3.10　图模型

在图模型中,节点表示变量,边表示条件概率. 图模型包括贝叶斯网络、隐马尔可夫模型等,人工神经网络也划分为该类问题[17]. 单个图模型可以针对单个问题进行多次评估,或者可以为单个输入评估许多图模型. 例如,在语音识别中声音可能被分解成多个帧,可以针对许多模型来评估每帧,以导出帧匹配特定音素的概率分布. 因为可以独立地评估图模型或输入,所以图模型可以实现比较简单的并行化,但针对单个问题的图模型并行可能会由于更新图权值而变得非常复杂. 图模型通常用于人工智能和机器学习应用,如语音和图像识别. 这里我们以卷积神经网络为例进行复杂度分析.

系统级优化采用了面向本地存储资源优化的双缓存设计、数据分块设计和寄存器通信策略;面向寄存器优化的寄存器分块计算流程与重用策略;面向效率优化的循环展开和指令流重排等方法,核心矩阵向量乘计算效率达 94%.

设卷积神经网络的输入包含 Ni 个通道的特征图片,每个特征图片的行数和列数分别为 Ri 和 Ci,输出为 No 个通道的特征图片,每个特征图片的行列数为 Ro 和 Co. 每个输出特征图片通过一个 $K\times K$ 大小的卷积核与每个输入特征图片相连,保证了特征图片之间的全连接. 设并行规模为 Np,实际训练过程中每进程内需要分批对图片进行处理,批次为 B. 则算法复杂度分析如下:算法时间复杂度为$O(B\times Ri\times Ci\times Ni\times No\times K\times K)$,空间复杂度为$O(B\times Ri\times Ci\times Ni+B\times Ro\times Co\times No+Ni\times No\times K\times K)$,计算核心为矩阵向量乘,基本为连续访存. 在通信复杂度方面,单个问题卷积神经网络的并行计算过程中,每进程需要得到其他进程的图权值,存在大量 AllReduce 通信,通信量为$O(Ni\times No\times K\times K)$,通信次数为$O(Np\times\log Np)$. 可以看出,单个问题的图模型并行由于更新图权值而变得非常复杂,问题规模扩大后通信成为性能瓶颈.

4　应用分类和体系结构需求

设 N 代表问题规模,如稠密矩阵秩、稀疏矩阵非零元个数、FFT 长度、粒子数、网格数、计算任务数等;Np 代表处理器数;图顶点数为$|V|$,边数为$|E|$,图直径为 D;卷积神经网络输入输出为 Ni、Ri、Ci、No、Ro、Co;则各类问题复杂度分析结果如表 1 所示. 本小节将首先介绍大规模应用优化模型和可扩展分析,根据各主题计算特征和数据迁移行为进行分类,提出体系结构需求.

表 1 10 类计算主题复杂度分析简表

类型	典型应用和代表算法	规模	时间复杂度	空间复杂度	通信复杂度
稠密线性代数	LINPACK	1228.8 万元	$2/3\times N^3$	$O(N^2)$	线性增长
稀疏线性代数	HPCG	3435 亿非零元	$O(N)$	$O(N)$,离散访存,近似随机	归约通信数线性增长
谱方法	FFT	16384^3	$O(N^3\times\log N)$	$O(N^3)$	$O(Np\times\log Np)$
多体问题	宇宙演化	11.2 万亿粒子	$O(N\times\log N)$	$O(N)$	$O(Np\times\log Np)$
结构网格	Stencil 计算	5.1×10^{11} 网格	$O(N)$	$O(N)$,规则访存	通信数不变,长度线性增长
非结构网格	通量计算	10^{10} 网格	$O(N)$	$O(N)$,随机访存	通信数不变,长度线性增长
MapReduce	MapReduce	10^7 任务数	$O(N)$	$O(N)$	基本无通信
图的遍历	BFS	2^{40} 顶点	$O(\|V\|+\|E\|)$	$O(\|V\|+\|E\|)$,随机访存	$O(D\times Np)\sim O(D\times Np^2)$
动态规划	序列比对	800GB 基因序列	$O(p(N))$	$O(N)$	基本无通信
图的模型	卷积神经网络	—	$O(B\times Ri\times Ci\times Ni\times No\times K\times K)$	$O(B\times Ri\times Ci\times Ni+B\times Ro\times Co\times No+Ni\times No\times K\times K)$	$O(Np\times\log Np)$

4.1 应用性能优化模型和可扩展性分析

大规模并行应用的性能优化方法主要分为以下几类:(1)计算方法优化:利用众核处理器体系结构特点实现应用程序在众核线程级的任务并行、数据并行和流水线并行的混合并行,提高众核并行效率;(2)访存优化:充分利用访存带宽和片上高效通信提高访存性能;(3)计算优化:利用指令流水、乘加优化和短向量优化等方法提高计算性能;(4)通信优化:利用数据打包、计算通信重叠、通信与网络拓扑结构的映射等方法提高大规模并行通信性能.

首先从单核组角度考虑影响计算性能的相关因素,单核组峰值性能为 742.4GFLOPS,假设单从核执行效率为 ee,应用计算量为 NN FLOPS(对应算法时间复杂度),访存量为 M Bytes(对应算法空间复杂度),应用实测访存带宽为 MBW GB/s,通信对数为 $Ncomm$、通信总量为 C Bytes(对应算法通信复杂度),网络带宽为 NBW GB/s,消息延迟为 N_L,则实际应用运行时间 T 为

$$T=\frac{NN}{742.4\times ee}+\frac{M}{MBW}+\frac{C}{NBW}+N_{\mathrm{LT}}\times Ncomm \tag{1}$$

考虑到整机大规模应用已实现计算和访存的互相隐藏以及计算和通信的互相隐藏,则优化后实际应用单核组性能为

$$T=\mathrm{Max}\left(\frac{N}{742.4\mathrm{G}\times ee},\frac{M}{MBW},\frac{C}{NBW}+N_{\mathrm{LT}}\times Ncomm\right) \tag{2}$$

从式(1)、(2)可以看出,针对不同类型的应用,性能优化方法应分别以增加执行效率 ee、提高实际应用访存带宽 MBW、减少通信量 C 和通信次数 $Ncomm$ 为主要目标.

从十类计算主题的 E 级需求可以得出,对于大部分访存受限课题,需要基于处理器的多级存储资源进行访存优化,目前单核组实测离散访存带宽 $MBW_{\mathrm{gld}}<1$GB/s,DMA 批量访存带宽 $MBW_{\mathrm{DMA}}\approx30$GB/s,从核访问 LDM 带宽 $MBW_{\mathrm{LDM}}=46.4$GB/s,片上阵列寄存器通信网络对分带宽为 $MBW_{\mathrm{reg}}=750$GB/s,应用实测访存带宽的具体组成如下:

$$MBW=M\Big/\left(\frac{M_{\mathrm{gld}}}{MBW_{\mathrm{gld}}}+\frac{M_{\mathrm{DMA}}}{MBW_{\mathrm{DMA}}}+\frac{M_{\mathrm{LDM}}}{MBW_{\mathrm{LDM}}}+\frac{M_{\mathrm{reg}}}{MBW_{\mathrm{reg}}}\right) \tag{3}$$

式中 M_{gld} 为单核组离散访存总量,M_{DMA} 为单核组 DMA 访存总量,M_{LDM} 为单核组访问 LDM 总量,M_{reg} 为单核组寄存器通信总量,受限于 LDM 容量访存有部分重叠,即 $M_{\mathrm{gld}}+M_{\mathrm{gld}}+M_{\mathrm{gld}}+M_{\mathrm{gld}}>M$. 从式(3)中可以看出,减少离散访存、提高 DMA 访存带宽、提高 LDM 使用率、充分利用片上阵列的高效通信机制等方法是大多数访存受限类应用性能提高的关键.

在大规模整机应用的并行效率和可扩展性分析方面,随着问题规模的增加,计算量、访存开销和通信开销是否线性增长,成为实际应用能否扩展到 E 级的关键问题之一. 从式(1)、(2)的性能优化模型可以看出,影响应用整体并行效率和可扩展性的重

要指标是计算量、访存开销和通信开销是否随着问题规模的扩大而线性增长.

4.2　应用分类

根据对以上应用的计算特征和数据迁移行为进行分析,我们将以上应用分为两大类,即计算和数据迁移规则型应用、计算和数据迁移不规则型应用.随着问题规模的扩大,这两大类应用扩展到 E 级可能会遇到的瓶颈问题不同,对超级计算机系统体系结构和软件环境也提出不同的需求,具体如下.

4.2.1　计算和数据迁移规则型应用

从计算核心的算法和访存分析可以看出,该类应用程序规则、计算量大、并行性好,访存规律(连续访存或跨步访存),通信模式上具有以下特点之一:完全并行无通信;通信量少或固定;通信模式固定;随问题规模增大通信量变化不大或线性增长.该类应用随着问题规模的扩大,应用时空复杂度增加,但计算通信比和计算访存比相对较高,具有较好的可扩展性和并行效率.上述十三类计算主题的稠密线性代数、简化多体问题、结构网格、MapReduce、组合逻辑、动态规划等均属该类应用.

当前该类问题的实际复杂应用系统向着多模式、多尺度和三维真实构型的方向发展,包含着大量多尺度多模型的计算问题,存在多粒度、多维度、多层次的并行性,面临着全系统、全物理过程、真三维、自然尺度的计算模拟,对计算机体系结构和软件环境提出更高要求,需要对当前计算机体系结构进行提升和改进.

在体系结构需求方面,该类应用需要多态、多尺度系统以实现复杂系统不同子问题的映射,满足复杂应用多种计算形态平滑无缝耦合;需要将不同类型核心集成在一个芯片内,并与不同特征的代码段进行匹配,以期达到最优的性价比;需要更高性能的多级多层次大容量存储、支持离散访存、高效片上数据共享,缓解复杂系统真三维模拟的访存墙问题;需要更快的网络带宽、更低的通信延迟和更通畅的 I/O 吞吐能力.

在编程环境需求上,单一的编程模型很难高效满足应用的多态多样性需求,需要多级多模式并行计算模型;针对不同物理过程,需要支持不同的网格剖分方案和并行算法以获得理想的并行效率;此外,在不同尺度和物理过程耦合计算中需要高效的并行耦合方法,保证数值模拟精度;需要支持动态负载平衡和自适应网格构建模型,提高湍流燃烧等瞬间负载不平衡问题的并行效率;需要研究复杂应用整体性能多目标优化方法,以提高复杂应用系统的整体运行效率.

4.2.2　计算和数据迁移不规则型应用

该类应用计算访存比低、访存不规则,主要是离散访存或完全随机访存,大规模应用的数据量远大于第一类应用.通信以动态的不规则通信为主.随着问题规模的扩大,数据量与计算量相伴增大,通信量可能出现超线性,甚至多项式增长:伴随产生的额外内存开销增速快于数据量增速;数据交互多样且复杂,导致网络需求较高且不固定.因此,在现有体系结构下其并行可扩展性较差.上述十三类计算主题的稀疏线性代数、谱方法、非结构网格、完全多体问题、图的遍历、回溯和分支限界、图模型等均属计算和数据迁移不规则型应用.

在体系结构需求上,需要匹配计算量和访存量,设计更大容量的片上存储,更高的访存带宽;该类问题求解过程存在的大量离散访存需要创新的内存控制器、片上互联、片上缓存等,以缓解离散访存带宽导致的性能瓶颈;部分应用的并行模式高度依赖于数据,且与体系结构的并行程度粒度不同,具有不确定性,在提高访存容量和带宽的同时,需要设计新的体系结构实现细粒度并行机制,使应用层面的小规模离散数据计算能得到硬件层面的多线程、并发访存、原子操作等支持;此外,在此类应用优化中广泛采用的消息合并等手段不能从本质上解决通信墙问题,需要提升网络能力、提高网络在系统中地位,根据应用特征提供更匹配的网络互联结构、高效的聚合通信支持;当前高性能计算机体系结构的设计思路和方式严重影响图的遍历等不规则应用的性能,需要以数据为中心的体系结构设计,突出存储和网络的作用.

在编程环境需求方面,该类问题需要编程模型支持多种并行粒度的抽象,便于描述数据不同类别的复杂问题;编程环境支持优化的数据分布方式,利用创新的内存控制器、片上互联、片上缓存机制减少计算过程中的大量离散访存;使用数据相关性分析和自动编译优化方法,提高部分离散数据的读写效率;并行环境需要增加支持细粒度并行的描述,提高

应用并行效率;支持同步及异步迭代执行,提高该类应用的整体效率.

5 结束语

本文中,每类科学工程计算问题在"神威·太湖之光"上的大规模实验和分析结果基本代表了该类问题的最大规模,分析数据具有一定代表性,但目前的实验分析未能说明随着应用规模的扩大,结果误差是否增大或计算是否收敛.此外,实际应用问题可能包含了多个计算主题,而我们的分析仅选取了其中的一部分;而且随着应用本身的发展,应用问题的分类随着多模式的加入会发生新的变化.最后,本文并未对机器学习类问题展开深入讨论,其计算核心在现有计算模型规模下已经取得较理想效果,但受限于计算模型本身的可扩展性,或将需要革命性的体系结构变革.

致 谢 非常感谢清华大学杨广文教授、陈文光教授、薛巍副教授、付昊桓副教授、林恒博士、方佳瑞博士,中国科学技术大学秦宏教授、刘建副教授、安虹教授,中国科学院过程工程研究所葛蔚研究员、侯超峰副研究员,中国科学院软件研究所杨超研究员、敖玉龙博士,山东大学刘卫国教授、段晓辉博士在应用分析方面提供的宝贵帮助,在此表示诚挚谢意!

参 考 文 献

[1] Zheng F, Li H L, Lv H, et al. Cooperative computing techniques for a deeply fused and heterogeneous many-core processor architecture. Journal of Computer Science and Technology, 2015, 30(1): 145-162

[2] Fu Hao-Huan, Liao Jun-Feng, Yang Jin-Zhe, et al. The Sunway Taihulight supercomputer: System and applications. Science China Information Sciences, 2016, 59(7): 072001

[3] Yang C, Xue W, Fu H, et al. 10M-core scalable fully-implicit solver for nonhydrostatic atmospheric dynamics// Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Salt Lake City, USA, 2016: 57-68

[4] Qiao F, Zhao W, Yin X, et al. A highly effective global surface wave numerical simulation with ultra-high resolution// Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Salt Lake City, USA, 2016: 46-56

[5] Zhang J, Zhou C, Wang Y, et al. Extreme-scale phase field simulations of coarsening dynamics on the Sunway Taihulight supercomputer//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Salt Lake City, USA, 2016: 4

[6] Fu H, Liu W, Wang L, et al. Redesigning CAM-SE for peta-scale climate modeling performance and ultra-high resolution on Sunway TaihuLight//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Denver, USA, 2017: Article No. 1

[7] Fu H, Yin W, Yang G, et al. 18.9-Pflops nonlinear earthquake simulation on Sunway TaihuLight: Enabling depiction of 18-Hz and 8-meter scenarios//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Denver, USA, 2017: Article No. 2

[8] Shalf J, Kamil S, Oliker L, et al. Analyzing ultra-scale application communication requirements for a reconfigurable hybrid interconnect//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Seattle, USA 2005: 17-20

[9] Sreepathi S, Grodowitz M L, Lim R, et al. Application characterization using Oxbow toolkit and PADS infrastructure// Proceedings of the 1st International Workshop on Hardware-Software Co-Design for High Performance Computing. New Orleans, USA, 2014: 55-63

[10] Asanovic. The Landscape of Parallel Computing Research: A View from Berkeley. Vol. 2. Technical Report UCB/EECS-2006-183. EECS Department, University of California, Berkeley, USA, 2006

[11] Lin J, Xu Z, Nukada A, et al. Optimizations of two compute-bound scientific kernels on the SW26010 many-core processor// Proceedings of the International Conference on Parallel Processing. Bristol, UK, 2017: 432-441

[12] Xu Zhi-Geng, Lin J, Matsuoka S. Benchmarking Sunway SW26010 many-core processor//Proceedings of the 7th International Workshop on Accelerators and Hybrid Exascale Systems (AsHES) (IPDPS Workshop). Orlando, USA, 2017: 743-752

[13] Meng De-Long, Wen Min-Hua, Wei Jian-Wen, Lin Xin-Hua. Porting and optimizing OpenFOAM on Sunway TaihuLight system. Computer Science, 2017, 44(10): 64-70(in Chinese) (孟德龙, 文敏华, 韦建文, 林新华. 神威·太湖之光上 OpenFOAM的移植与优化. 计算机科学, 2017, 44(10): 64-70)

[14] An Hong, et al. Pipelining computation and data reuse strategies for scaling GROMACS on the Sunway Many-Core Processor//Proceedings of the 18th International Conference

on Algorithms and Architectures for Parallel Processing (ICA3PP-2018), 2018(to be accepted)

[15] Yao Wen-Jun, Chen Jun-Shi, Su Zhi-Chao, et al. Porting and optimizing of NAMD on Sunway TaihuLight system. Computer Engineering & Science, 2017, 39(6): 1022-1029 (in Chinese)
(姚文军，陈俊仕，苏志超等. 基于神威·太湖之光的 NAMD 软件的移植与优化. 计算机工程与科学，2017，39(6)：1022-1029)

[16] Fu H, Liao J, Xue W, et al. Refactoring and optimizing the community atmosphere model (CAM) on the Sunway TaihuLight supercomputer//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Salt Lake City, USA, 2016: 969-980

[17] Liu J, et al. Largest Particle Simulations Downgrade the Runaway Electron Risk for ITER. arXiv preprint arXiv: 1611.02362, 2016

[18] Chen Y, Li K, Fei X, et al. Implementation and optimization of AES algorithm on the Sunway TaihuLight//Proceedings of the 17th International Conference on Parallel and Distributed Computing, Applications and Technologies. Guangzhou, 2016: 256-261

[19] Lin H, Tang X, Yu B, et al. Scalable graph traversal on Sunway TaihuLight with ten million cores//Proceedings of the 31st IEEE International Conference Parallel and Distributed Processing Symposium. Orlando, USA, 2017: 635-645

[20] Duan X, Xu K, Chan Y, et al. S-aligner: Ultrascalable read mapping on Sunway TaihuLight//Proceedings of the IEEE International Conference on Cluster Computing. Hawaii, USA, 2017: 36-46

[21] Fang J, Fu H, Zhao W, et al. swDNN: A library for accelerating deep learning applications on Sunway TaihuLight//Proceedings of the 31st IEEE International Conference Parallel and Distributed Processing Symposium. Orlando, USA, 2017: 615-624

[22] Qi Feng-Bin. Sunway TaihuLight super computer. Communications of the CCF, 2017, 13(10): 16-22(in Chinese)
(漆锋滨. "神威·太湖之光"超级计算机. 中国计算机学会通讯，2017，13(10)：16-22)

[23] He Cang-Ping. OpenACC Parallel Programming. Beijing: China Machine Press, 2016(in Chinese)
(何沧平. OpenACC 并行编程实战. 北京：机械工业出版社，2016)

[24] Ao Y, Yang C, Wang X, et al. 26 PFLOPS stencil computations for atmospheric modeling on Sunway TaihuLight// Proceedings of the 31st IEEE International Conference Parallel and Distributed Processing Symposium. Orlando, USA, 2017: 535-544

[25] Dong W, Kang L, Quan Z, et al. Implementing molecular dynamics simulation on Sunway TaihuLight system// Proceedings of the 18th IEEE International Conference on High Performance Computing and Communications. Chengdu, 2016: 443-450

LIU Xin, born in 1979, Ph. D., associate professor. Her research interests include parallel algorithms and parallel application software.

GUO Heng, born in 1993, Ph. D. candidate, engineer. His research interests include parallel algorithms and parallel application software.

SUN Ru-Jun, born in 1990, Ph. D. candidate, engineer. Her research interests include computer architecture and computing models.

CHEN Zuo-Ning, born in 1957, senior engineer, Ph. D. supervisor, member of Chinese Academy of Engineering. Her research interests include high performance computer system architecture and operating system.

Background

This research belongs to the project of "The Research and Development of Coupler Platform of Large-Scale, Multi-Model, Multi-Process Earth System Model" as the part of the National Major Project—"The Global Climate Change and Response Project" granted by No. 2016YFA0602200.

Sunway TaihuLight supercomputer system has supported several hundreds of users and one hundred more large complex applications of the calculation, involving weather, aerospace, marine environment, bio-medicine, ship engineering and other 19 application fields since put into practical use. Twenty more applications achieved ultra-large-scale parallel scale of one million cores, which involved 17 full-scale applications and 12 semi-scale applications. Five full-scale applications entered the finalists of Gordon Bell Awards.

It can be seen from most applications that the current real application problem is oriented toward multi-scale, strong nonlinear coupling and three-dimensional, including a large number of multi-scale and multi-model computing problems. There are multi-granularity, multi-dimension and multi-level parallelism in these applications, faced with the large computing simulation of the whole system, the whole physical process, true three-dimensional and natural scale, which put forward higher requirements for the supercomputer's ability.

Most large-scale applications of Sunway TaihuLight supercomputer are the largest scale of the correspondent field that partly represents the characteristics of the application. This paper mainly analysis the calculation characteristics and data migration behavior of the semi-scale and full-scale computing-intensive applications. Focusing on the characteristics of the algorithm, the adaptability of the architecture, the algorithm complexity, the space complexity, the characteristics of the memory access and the communication complexity, we got the bottlenecks of the application algorithms are extended to the exa-scale. Based on the performance bottlenecks, this paper proposes the computer architecture requirements and design recommendations about the next generation exa-scale supercomputer.