面向神威众核超算系统的并行计算模型研究

高剑刚 刘 鑫 李 芳 刘 勇 彭达佳 陈 鑫 陈德训
(国家并行计算机工程技术研究中心 北京 100083)

摘 要 基于异构众核处理器的超级计算机已经成为 TOP500 高性能计算机的主流,BSP、LogP、PRAM 等已有并行计算模型均针对基于多核处理器的超级计算机设计,不能满足日益迫切的基于众核架构的超级计算机和应用发展需求.本文面向“神威·太湖之光”和神威 E 级原型系统的众核体系结构特点,提出 P-PALN(Parallel-Parallel Access via LDM & NOC)并行计算模型,对于计算节点间的并行,该模型沿用 BSP/LogP 模型描述;对于计算节点内的众核并行,该模型提供私有存储访问和片上阵列通信的众核并行架构的有效描述 PALN,能够协助用户进行众核并行算法设计,并在申威众核处理器硬件设计中指导参数的优化.实验结果表明,该模型可有效指导硬件设计和用户众核编程,从而提高系统和应用的性能.

关键词 众核处理器;并行计算模型;P-PALN;PALN;片上通信
中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2023.01339

Research on Parallel Computing Model for Sunway Many-Core Supercomputing System

GAO Jian-Gang LIU Xin LI Fang LIU Yong PENG Da-Jia CHEN Xin CHEN De-Xun
(*National Research Center of Parallel Computer Engineering and Technology, Beijing* 100083)

Abstract Supercomputers based on heterogeneous many-core processors have become the mainstream of TOP500 high-performance computers. Existing parallel computing models, such as BSP, LogP and PRAM, are designed for supercomputers based on multi-core processors, which cannot meet the increasingly urgent needs of supercomputers and application development based on many-core architecture. Aiming at the Sunway Taihulight system and Sunway exa-scale prototype supercomputing system, based on the characteristics of many-core architecture, this paper proposes a hybrid parallel computing model P-PALN (Parallel-Parallel Access via LDM & NOC). For parallelism mode between computing nodes, the model uses the traditional BSP and LogP to describe. For parallelism mode within many-core computing node, this model effectively describes many-core parallel architecture featuring parallel private storage access to LDM and on-chip communication (PALN), which can guide users to better design parallel programming algorithms based on many-core architecture and engineers to optimize the corresponding parameters directionally during hardware design of Shenwei many-core processor. The experimental results show that the proposed model can effectively guide the hardware design and user many-core parallel programming, thereby improving the performance of the system and application.

Keywords many-core processor; parallel computing models; P-PALN; PALN; on-chip communication

收稿日期:2022-03-22;在线发布日期:2022-10-08.本课题得到国家科技重大专项(2017-I-0004-0004)资助.**高剑刚**,硕士,正高级工程师,中国计算机学会(CCF)会员,主要研究方向为高性能互连网络和计算机体系结构.E-mail:13701512205@139.com.**刘 鑫**(通信作者),博士,研究员,主要研究领域为并行算法和应用.E-mail:yyylx@263.net.**李 芳**,博士,副研究员,主要研究方向为高性能计算应用.**刘 勇**,博士,副研究员,主要研究方向为并行算法.**彭达佳**,博士,工程师,主要研究方向为并行计算和调试.**陈 鑫**,硕士,工程师,主要研究方向为并行算法和应用.**陈德训**,博士,研究员,主要研究领域为高性能计算应用.

1 引 言

超级计算在国家安全和国民经济发展中起着越来越重要的作用,已成为国家科技实力和科研发展水平的重要标志.并行计算模型作为超级计算机的系统设计者和应用开发者之间的重要桥梁,从大规模应用并行算法设计需求出发,将真实并行计算机系统的计算、访存、通信等基本特征参数化抽象成计算模型,为并行计算提供硬件设计和应用开发的接口.在该接口的约定下,并行计算系统的软硬件设计者可以完成系统的并行支撑机制和并行软件环境开发,开展系统和应用的协同设计;应用开发者可以更充分发挥系统的并行计算特征,提高国防信息安全和国家重大工程等领域大规模并行应用的实际运行效率.可以看出,并行计算模型对并行计算系统的设计和并行应用的开发均发挥着至关重要的作用.

当前,基于异构众核处理器的超级计算机[1-2]已经成为 TOP500 高性能计算机榜单的主流,日益复杂的体系结构[3]带来新的编程墙问题.已有并行计算模型[4-13]以 PRAM[9]、BSP[12]、LogP[13]为代表,适用于基于传统多核处理器实现的并行计算机,不能满足日益迫切的基于众核架构的超级计算机和应用发展需求.本文针对神威 E 级原型系统[1]和"神威·太湖之光"系统[2],提出 P-PALN(Parallel-Parallel Access via LDM & NOC)并行计算模型,对于计算节点间的并行,该模型沿用 BSP/LogP 模型描述;对于计算节点内的众核并行,该模型提供私有存储访问和片上阵列通信的众核并行架构的有效描述 PALN,能够协助用户进行众核并行算法设计,并指导申威众核处理器[3]硬件设计参数的持续优化.实验结果表明,该模型可有效指导硬件设计和用户众核编程,提高系统和应用的性能.

2 研究背景

2.1 并行计算模型简介

PRAM[9](Parallel Random Access Machine,随机存取并行机器)模型面向共享存储的多核并行体系结构提出,旨在解决在单核计算性能有限的情况下,如何利用处理器内的核间并行提升计算性能的问题.BSP[12](Bulk Synchronous Parallel,块同步并行计算)模型和 LogP[12]模型面向分布式存储的多机并行体系结构提出,针对互连网络的带宽和延迟,描述了并行处理器间的消息传递和同步方式.BSP 模型中的通信集合为全体并行处理机,若采用通信子集同步方式进行优化(子集小到只包含成对的收发者),则变为异步的个体同步,即 LogP 模型.LogP[12]模型描述互连网络性能与处理机计算性能,用户可以通过该模型预估算法的执行时间.

BSP 和 LogP 模型有效描述了多计算节点间的并行,而 PRAM 模型描述了单个计算节点内共享存储体系结构下的多计算核心并行.以 BSP/LogP+PRAM 为代表的已有并行计算模型,适用于基于传统多核处理器实现的并行计算机,指导了现有众多大规模应用并行算法的设计和优化.

随着计算节点内处理器体系结构的发展,不断有新的面向共享存储体系结构的并行计算模型被提出.在处理器通过提高核心频率的方式来提升计算性能的效果不断降低的情况下,多处理核心并行成为提升单个处理器性能的有效方式.以 PRAM 为代表的计算模型有效描述了这种并行方式,支撑了算法的并行复杂度分析,APRAM[11]、分相 PRAM[14]等模型进一步支撑了处理核心间的异步计算.然而,随着芯片设计和制造工艺的发展,存储性能逐渐落后于处理核心的计算性能,访存带宽成为应用性能提升的瓶颈.利用程序对数据访问的局部性特征,层次化的存储体系结构被提出,以缓解访存瓶颈问题,块访问成为内存访存的重要特征,RAM(h)[15]模型有效描述了不同存储层次的访问开销以及分块大小对访存性能的影响.

2.2 体系结构发展对计算模型的挑战

应用对计算性能需求持续增长,处理器体系结构也在不断更新以满足应用需求,同时也需要更加合适的计算模型来指导应用算法设计与优化.随着 SIMD、众核和 Tensor Core 等新型部件的提出,处理核心计算性能进一步提升,访存瓶颈凸显,因此用户可控的片上高速缓存被提出,用以更加高效的数据复用,同时处理核心间的互连通信也被设计用于核心的私有高速缓存间的数据共享,从而增大处理单元的计算数据工作集,提高实际应用性能."神威·太湖之光"系统采用国产超高性能众核处理器构建,如图 1 所示,为缓解访存墙挑战,申威众核处理器[16]创新提出了具有片上通信功能的众核体系结构,单处理器集成 4 个核组,每核组包含 1 个运算控制核心和 1 个 8×8 的运算核心阵列;其中,运算核心可以直接离散或批量访问主存,也可以采用片上网络通信方式进行各自私有数据的交换.

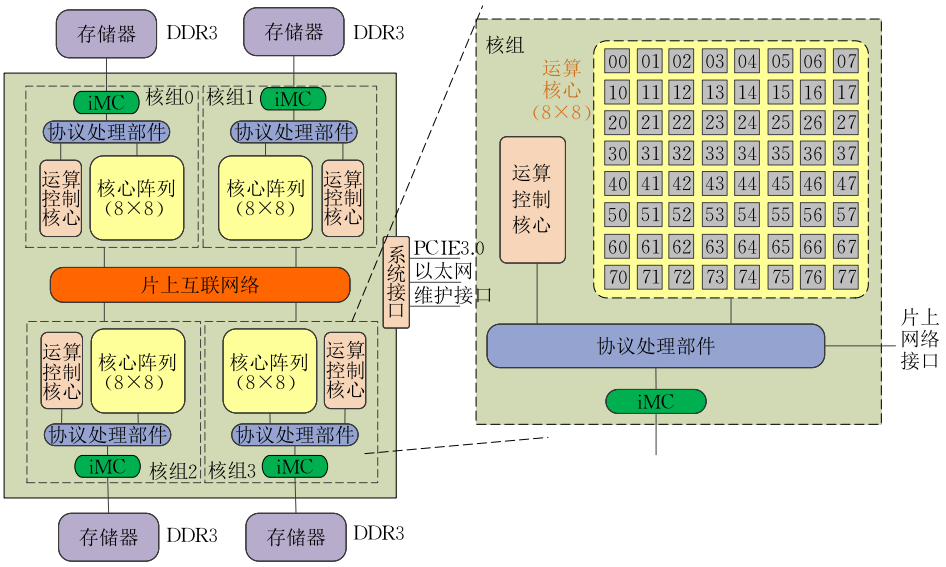


图 1　“申威 26010”异构众核处理器架构图

与此类似，为了提高执行效率，NVIDIA 最新发布的 H100 GPU 在一个 GPC（GPU Processing Cluster，GPU 处理簇）内的 SM（Stream Multiprocessor，流多处理器）间增加了 SM-To-SM 的网络，提供了面向 SM 上的高速数据缓存 SMEM（Shared Memory）间的快速数据共享能力，实现了比通过全局内存快约 7 倍的数据交换.

针对神威众核超算系统，众核处理器间的并行模型和传统模型类似，而采用何种并行模型对底层的全新众核处理器架构进行描述、指导处理器迭代优化和应用并行算法设计成为亟待解决的关键问题. 可以看出，现有并行模型已经不能充分匹配该类处理器的体系结构特点，需要根据最新众核架构，研发新的并行计算模型对申威众核处理器进行描述，同时指导用户的并行算法设计.

3　并行计算模型 P-PALN

面向神威系列众核超算系统，本文提出了P-PALN并行计算模型，如图 2 所示. 该模型分为两个层次：对于多计算节点间的并行，该模型沿用 BSP/LogP 模型描述，即模型的“P”部分；对于计算节点内的众核并行，基于申威众核处理器的硬件特征，抽象出基于处理单元私有空间直接访问和片上通信间接访问的混合并行计算模型，即模型的“PALN”部分，PALN 模型使用从核 LDM 访问和片上阵列通信的混合并行方式，能够对众核并行架构进行有效描述，协助用户进行众核并行算法设计，同时根据性能评估结果指导申威众核处理器硬件设计参数的持续优化. 因计算节点间并行模型主要沿用传统 BSP/LogP

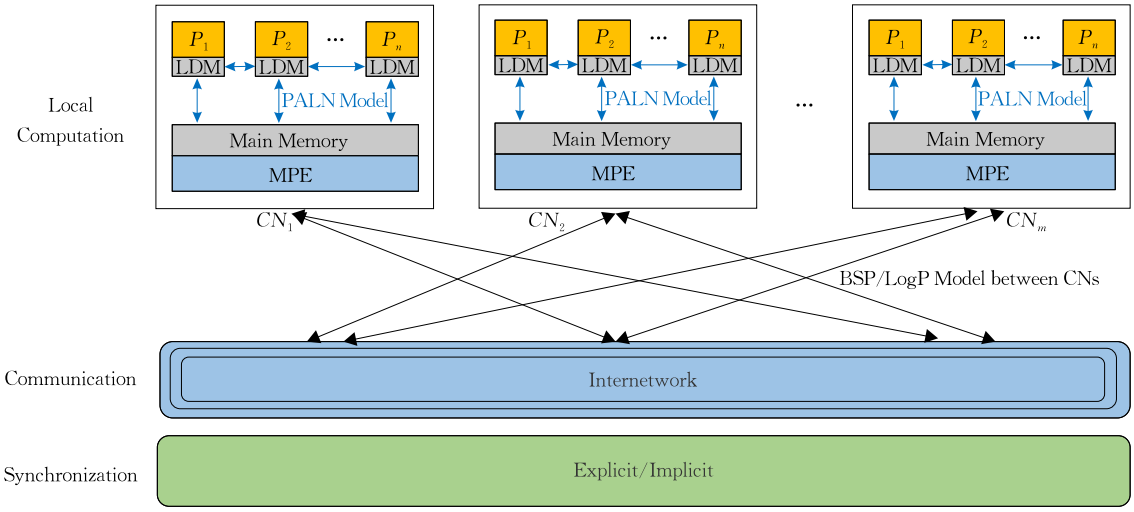


图 2　面向神威众核超算系统的并行计算模型 P-PALN

模型描述，本文重点对众核并行计算模型 PALN 模型进行分析.

3.1 PALN 并行计算模型原理

私有访存和片上通信混合并行计算模型 PALN (Parallel Access via LDM & NOC)的基本原理如图 3 所示. 在该计算模型中，多个处理单元(Processing Element，PE)协同计算，PE 也称为并行加速计算核心. 其中，每个处理单元配置高速本地存储(Local Direct Memory，LDM)，从全局主存(Memory)中访问数据，并在 LDM 内重复使用，假定 PE 可以在单位时间内访问 LDM；若一个 PE 需要使用其他 PE 的 LDM 数据完成计算，则可以通过 PE 间的片上通信获取. 该模型的主要参数为

(1) PE 的个数：P；

(2) 全局主存的访存带宽(所有 PE 访问)：B；

(3) 片上通信带宽(单个 PE 从其他 PE 获取数据的聚合通信带宽)：b.

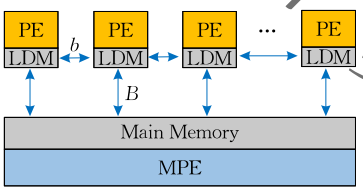


图 3 PALN 并行计算模型

该模型侧重于对访存性能的评估，在访存带宽受限的情况下，指导并行算法的映射和体系结构的设计. 考虑负载均衡算法的并行映射，令并行计算任务的工作数据集的数据量为 $P\cdot n$，且并行计算任务可以被均匀地切分为 P 份，然后加载到 P 个处理单元；其中，每个计算任务的工作数据集的数据量为 $\alpha P\cdot n$，α 为单个计算任务的工作数据集占并行任务工作数据集的比例，表示了数据复用度，即单个计算任务需要使用 $\alpha P\cdot n$ 的数据完成计算，且有

$$\alpha P\cdot n\geq n,\ \alpha P\geq 1.$$

在程序实现中，由于 LDM 容量和全局主存带宽的限制，每个 PE 首先获取 n 的数据到 LDM，然后在计算过程中通过片上通信从其他 PE 的 LDM 内获取 $\alpha P\cdot n-n$ 的数据完成单个计算任务的计算. 在该模型下，并行算法的访存开销为

$$M_{\mathrm{PALN}}=\frac{Pn}{B}+\frac{\alpha P\cdot n-n}{b} \tag{1}$$

在 PE 可以在单位时间内访问 LDM 的假设下，算法的计算时间复杂度分析于 RAM 模型相同，为

$$T=O(f(\alpha P\cdot n)) \tag{2}$$

即单个 PE 在 $\alpha P\cdot n$ 个数据上串行完成计算任务的时间复杂度，其中 f 与具体计算方法相关.

3.2 模型对并行算法映射的指导

与典型面向多核体系结构的并行计算模型(如 PRAM)相比，片上通信并行计算模型通过引入片上通信，有效提升 LDM 内数据重用率，减少访存时间. 典型 PRAM 多核并行计算模型如图 4 所示.

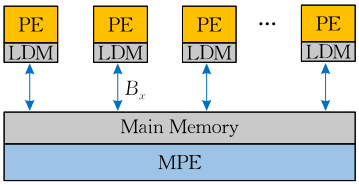


图 4 典型 PRAM 多核并行计算模型

在该计算模型中，PE 间无片上通信，无法实现 LDM 内数据的复用，因此各个 PE 需要各自从全局主存访问 $\alpha P\cdot n$ 的数据，完成被分配的计算任务. 按照上述的数据和任务划分方法，该模型完成并行核心算法的计算时间复杂度仍为 T，其访存开销为

$$M_{\mathrm{PRAM}}=\frac{\alpha Pn\cdot P}{B_x} \tag{3}$$

其中，B_x 为全局主存的访存带宽(所有 PE 访问).

若对以上两个模型，有 $B_x=B$，$Pb>B$，则

$$M_{\mathrm{PALN}}=\frac{Pn}{B}+\frac{\alpha Pn-n}{b}<\frac{Pn}{B}+\frac{\alpha Pn-n}{B/P}$$

$$=\frac{\alpha P^2n}{B}=M_{\mathrm{PRAM}} \tag{4}$$

即在主存访存带宽有限的情况下，若并行算法能够合理利用片上通信能力，则能够有效提升应用性能.

3.3 模型对体系结构设计的指导

在体系结构设计中，由于工艺等因素影响，提升访存带宽的技术手段有限，而引入片上通信，能够有效降低对访存带宽的需求. 针对以上两个模型，我们考虑在实现同等访存性能的情况下对访存带宽的需求，即 B 与 B_x 的关系. 令 $b=k\cdot B/P$，$k>0$，若

$$M_{\mathrm{PALN}}=\frac{Pn}{B}+\frac{\alpha Pn-n}{b}=\frac{Pn}{B}+\frac{\alpha Pn-n}{k\cdot B/P}$$

$$=\frac{\alpha P^2n}{B_x}=M_{\mathrm{PRAM}},$$

则有

$$B=B_x\cdot\frac{k+\alpha P-1}{\alpha Pk} \tag{5}$$

此时，判断 B_x，B 的大小关系，即判断 $(k+\alpha P-1)/\alpha Pk$ 与 1 的大小关系，进一步化为 $(k+\alpha P-1)-\alpha Pk$ 与 0 的大小关系. 其中，$k+\alpha P-1>0$，$\alpha Pk>0$ 且 $\alpha P\geq 1$. 由于

$$k+\alpha P-1-\alpha Pk=(1-k)(\alpha P-1).$$

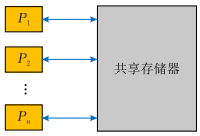
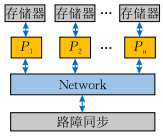
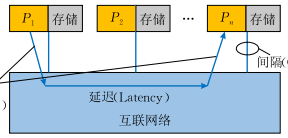
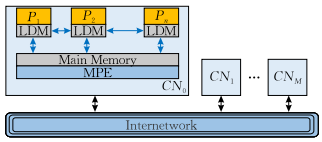
若 $k>1$ 且 $\alpha P>1$，则 $k+\alpha P-1-\alpha Pk<0$，$(k+$

$\alpha P-1)/\alpha Pk<1, B<B_x$. 即在体系结构设计中，在主存访存带宽 B 提升手段有限的情况下，通过增加带宽为 $b=k\cdot B/P, k>1$ 的片上通信，能够降低对访存带宽的需求，实现同等访存性能.

3.4　与其他并行模型的对比

并行计算模型 PALN 针对全新的具有片上通信功能的异构众核体系结构提出，表 1 列出了该模型与其他典型并行计算模型的比较.

表 1　PALN 与其他并行计算模型的综合比较

	PRAM	BSP	LogP	PALN
模型结构				
体系结构	MIMD-SM	MIMD-DM	SIMD-DM	SIMD-SM
计算模式	同步计算	异步计算	异步计算	异步计算
同步方式	自动同步	路障同步	隐式同步	路障同步
计算粒度	细粒度/中粒度	中粒度/大粒度	中粒度/大粒度	细粒度/中粒度
通信方式	读/写共享变量	发送/接收消息	发送/接收消息	发送/接收消息
编程地址空间	全局地址空间	单/多地址空间	单/多地址空间	全局地址空间
参数	虚拟处理机个数 P	虚拟处理机个数 P 路障同步开销 L 全局通信网络带宽 g	通信延迟 L 消息发送/接收时间开销 o 消息收/发最小时间间隔 g 处理器集合 P	PE 的个数 P 主存访存带宽 B 片上通信带宽 b 计算任务的工作数据集 $P\cdot n$ PE 任务的工作数据集 $\alpha P\cdot n$
模型特征	处理器同步，通信无延迟，无限带宽	引入超步、网络带宽考虑同步开销	分布式存储、点对点通信 处理器间完全异步	多级存储 片上快速通信
模型优势	使用简单，利于并行复杂度分析	计算和通信分离，利于通信复杂性分析	突出访存瓶颈，处理机间异步工作，消息传送完成同步	提高数据重用率 缓解存储墙问题
访存开销	$O(1)$	$\frac{\alpha P^2\cdot n}{g}+L$	$\frac{\alpha P^2\cdot n}{g}+L+o$	$\frac{Pn}{B}+\frac{\alpha P\cdot n-n}{b}$
数据重用率 数据量/传输量	1	1	1	$\alpha P(\geq 1)$
适用性	多核架构 计算密集型应用	多机架构 计算与通信分离应用	多机架构 PE 间有一定数据交换量	异构众核架构 PE 间数据交换量高

在该模型中，并行算法的执行由数个包含以下两个阶段的计算核心构成：(1) 每个处理单元访问算法划分的数据，进行全局共享存储与高速私有存储间的数据传输；(2) 处理单元利用本地私有存储内的数据以及通过片上通信功能远程获取到的其他处理单元上私有存储内的数据，完成划分的计算任务. 在第二个阶段中，远程数据访问、本地数据访问和计算交错，该过程为分布式存储多处理机体系结构下的并行计算，数据的横向通信发生在多个处理单元利用消息传递数据和控制信息上，计算的优化需要考虑通信复杂度(即数据的通信和移动开销)，此时该模型退化为在有限数量的并行处理机及固定的互连网络下的分布式存储并行计算模型(如 BSP 或 LogP 模型).

然而，仅用分布式存储并行计算模型(如 BSP 或 LogP 模型)难以同时描述处理单元对全局共享存储的访问和处理单元间通过片上网络的数据共享. 分布式存储并行计算模型往往基于处理单元的本地存储的容量无限的假设，在处理单元执行的任务之间存储数据依赖或者控制依赖时，通过互连网络实现数据交换或同步；而对于本文描述的众核架构，处理单元本地的片上高速缓存(LDM)容量有限，因此能容纳的计算任务数据集有限，访问全局共享存储替换 LDM 内的数据集是必须的，而 BSP 或 LogP 模型仅能描述 LDM 间的横向数据交换. 通过 PALN 模型，并行程序设计人员能够有效平衡处理单元对全局共享存储和远程高速存储的数据访问，实现高效的数据复用.

表 2 将 PALN 模型与类 PRAM 的并行计算模型进行了比较. PRAM 模型面向典型多核和单层存储的体系结构设计，可被有效用于并行算法的复杂度分析. 针对 PE 间的同步时间，其衍生模型对 PRAM 模型进行了扩展，以支持不同的同步方式. 然而，这类模型基于单位访存时间的假设，并未考虑访存性能，因此更适用于处理核心计算性能与存储访存性能、延迟匹配的处理器的分析，计算密集型应用可以有效地分析并预测性能. 当访存带宽受限时，若应用的访存时间大于计算时间，即使通过计算访

表 2 PALN 模型与类 PRAM 模型的比较

	PRAM	APRAM	RAM(h)	PALN
体系结构特征	单层存储	单层存储	层次存储 块访问	层次存储、块访问 可控高速缓存 片上互连网络
参数	/	/	存储层数 h 数据长度 l 第 i 层分块大小 b_i 第 i 层的访存开销 c_i	PE 的个数 P 主存访存带宽 B 片上通信带宽 b 计算任务的工作数据集 $P\cdot n$ PE 任务的工作数据集 $\alpha P\cdot n$
访存开销	$O(1)$	$O(1)$	$\sum_{i=1}^{h}\lceil l/b_i\rceil c_i$	$\frac{Pn}{B}+\frac{\alpha P\cdot n-n}{b}$
模型核心	计算复杂度分析	计算复杂度分析 同步复杂度分析	纵向存储层次 数据复用	高速缓存的数据通信 PE 计算数据集的扩大
适用性	计算密集型应用 PE 同步执行	计算密集型应用 PE 异步执行	带宽受限体系结构 PE 任务数据交换量低	带宽受限体系结构 PE 任务数据交换量高

存重叠，计算也无法掩盖访存时间，访存性能决定了应用性能，此时访存时间成为应用优化的核心，这类模型难以有效指导算法的设计.

层次存储利用程序的数据局部性特征，通过块访问和数据重用，可以有效缓解访存瓶颈问题，提高应用性能，类 RAM(h)的层次存储模型考虑了该体系结构的变化，分析了不同存储层次的开销、分块大小对访存性能的影响. 这类模型基于 PE 通过全局共享存储进行数据交换的假定，适用于 PE 上的任务数据交换量低的应用. 对于本文讨论的通过片上网络实现高速缓存间的数据共享从而增大 PE 的计算数据工作集的体系结构，这类模型难以有效描述并提供高效的算法设计指导.

本文提出的从核私有和片上通信混合并行计算模型 PALN 拥有全局共享存储层次之上的高速私有存储层，并显式增加了片上处理单元间的数据通信功能，提供了处理单元间高效的数据共享能力. 与典型层次存储模型相比，该模型利用 PE 间的数据共享，增大了 PE 上任务的计算工作集，从而有效提高算法可实现的计算性能，详见 4.1 节.

相较于典型 PRAM 模型通过读/写共享变量实现数据共享和通信的方式，该模型的片上通信功能允许程序设计人员更为精确地控制处理单元间的数据共享，提升数据重用率；同时，若要保证全局共享存储中共享变量在处理单元间的 Cache 一致性，硬件设计复杂度将随着处理单元数量的增加而提升，且作废或更新 Cache 副本的开销也急剧提高，对于配置了大量处理单元的众核体系结构，硬件设计难度和开销都是比较大的，而片上通信能够有效降低硬件设计复杂度，同时提供了等效的数据共享能力. 可以看出，当 $b=0$ 时，该模型退化为图 4 所示的 PRAM 模型.

4 应用情况

4.1 模型对硬件设计的支撑

4.1.1 片上网络通信拓扑的设计

结合应用特征和 PALN 并行计算模型，可以提出对片上网络通信拓扑的设计建议. 表 3 列出了十类科学与工程计算主题应用对片上网络通信的具体需求，结合本文提出的 PALN 并行计算模型，可以看出，大多数科学与工程计算应用中，采用区域分解、边缘通信的方式进行大规模并行，其计算任务相对从核阵列进行二维划分，数据共享一般在同行、同列的处理单元间进行，故而申威众核处理器的计算处理单元组织为 Mesh 网格的阵列.

表 3 典型应用算法对片上通信需求

应用类型	典型应用 代表算法	应用类型	片上通信需求
稠密线性代数	LINPACK	计算密集	LDM共享度越高，对 访存带宽要求越低
稀疏线性代数	HPCG	访存受限 (延迟主导)	降访存延迟或者 Cache 替代
谱方法	FFT	次计算密集	增加数据重用性
多体问题	宇宙演化	次计算密集	增加数据重用性
结构网格	Stencil 计算	访存受限 (带宽主导)	无需求
非结构网格	通量计算	访存受限 (延迟主导)	降访存延迟或者 Cache 替代
MapReduce	MapReduce	访存受限 (带宽主导)	无需求
图的遍历	BFS	访存受限 (延迟主导)	降访存延迟或者 Cache 替代
动态规划	序列比对	次计算密集	增加数据重用性
图的模型	神经网络	计算密集	增加数据重用性

4.1.2 处理器硬件参数选择

在确定片上网络拓扑后，结合 Roofline 性能模型，该模型可以指导多个硬件参数的平衡设计. 针对

选定的基线应用，设其核心算法为 A，令 LDM 容量为 C，按照算法的并行映射和数据划分，可以计算得到该算法在体系结构下的计算密度 $\varphi(A,C)$，同时可以确定数据复用度 α. 根据该模型的分析，通过片上通信实现数据共享取得的等效访存带宽为 $B_x=B\cdot\frac{\alpha Pk}{k+\alpha P-1}$，令处理器理论峰值性能为 F_{peak}，则该基线应用能够实现的理论最高性能为 $\min\{B_x\cdot\varphi(A,C),F_{\text{peak}}\}$

$$F_{\max}=\min\left\{B\cdot\frac{\alpha Pk}{k+\alpha P-1}\cdot\varphi(A,C),F_{\text{peak}}\right\}\quad(6)$$

此外，片上网络中单个 PE 从其他 PE 获取数据的聚合带宽 $b=k\cdot B/P$，$k>1$. 因此，在选择合适的基线应用后，通过对该基线应用能够实现的理论最高性能 $F_{\max}$ 中各个参数的分析，可以根据设计需求，平衡理论峰值计算性能 F_{peak}、LDM 容量 C、全局存储带宽 B、片上网络带宽 b 的设计.

以新一代申威众核处理器（申威 26010Pro）为例进行分析，该处理器配置了 6 个从核阵列，每个从核阵列由 64 个以 8×8 Mesh 网格互连的计算处理单元构成，即 PE 的个数 $P=64$，理论峰值计算性能 $F_{\text{peak}}=14.7$ TFlops. 综合考虑访存密集型应用对高访存带宽的需求以及设计工艺和生产成本的限制，单核组对全局主存的访存带宽按照最高性能设计，具体为 $B=50$ GB/s；在主存访存带宽有限的情况下，较大的 LDM 容量能够提高数据复用度，受限于芯片面积，每个计算处理单元配置的 LDM 容量为 256 KB. 在 LDM 容量 C 和全局存储带宽 B 给定的情况下，对于选定的基线应用，需要为理论峰值计算性能 F_{peak} 配置合适的片上网络带宽 b.

以矩阵-矩阵乘 $\boldsymbol{M}_C=\boldsymbol{M}_A\times\boldsymbol{M}_B$ 为例，矩阵元素以双精度浮点数存储，算法的并行计算任务为对矩阵 $\boldsymbol{M}_C$ 的计算，按照二维划分方式进行并行，如图 5 所示.

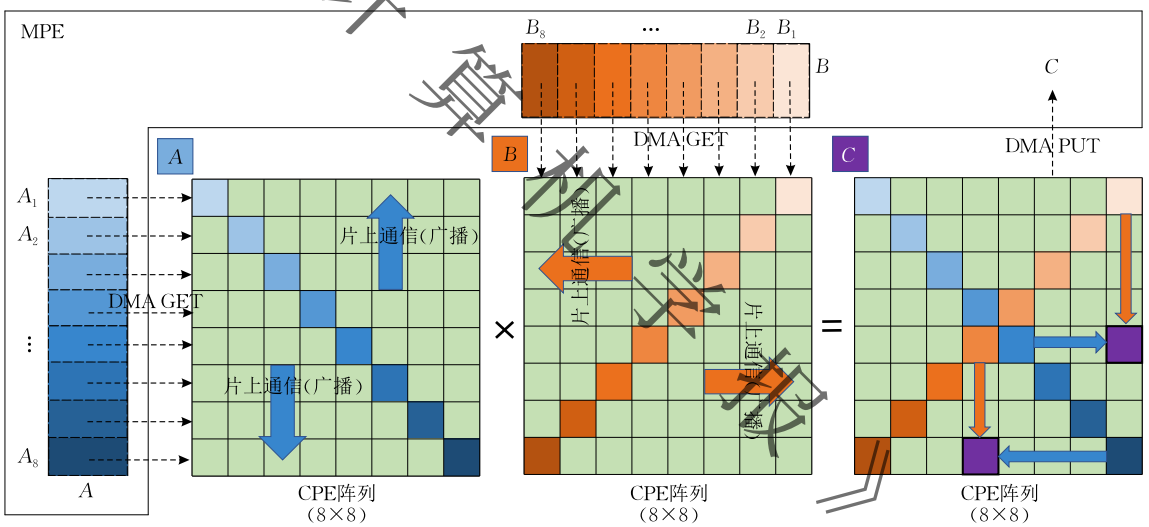


图 5　矩阵乘并行计算流程示意图

按照 8×8 的方式均匀切分 3 个矩阵，每个计算处理单元映射的计算任务为对矩阵 $\boldsymbol{M}_C$ 的对应分块（$\boldsymbol{M}_C/64$）的计算，且每个计算任务需要访问与该分块对应的矩阵 $\boldsymbol{M}_A$ 的同行和矩阵 $\boldsymbol{M}_B$ 的同列的分块数据来完成计算. 设矩阵分块的维度均为 $N\times N$，则 LDM 需要存储至少 3 个矩阵分块，$N=\sqrt{C/24}=104$. 在计算中，矩阵 $\boldsymbol{M}_C$ 常驻 LDM，矩阵 $\boldsymbol{M}_A$，$\boldsymbol{M}_B$ 不断更新，则算法的计算密度

$$\varphi=\frac{2N^3}{2N^2\times8(B)}=13.06\text{ op/B}.$$

此时，若无片上网络，该算法在典型多核计算模型下能够实现的理论最高性能为

$$\varphi\cdot B_x\cdot6=13.06\text{ op/B}\times50\text{ GB/s}\times6=3919\text{ GFlops}.$$

在 PALN 模型下，矩阵 $\boldsymbol{M}_C$ 常驻 LDM，并行任务的工作数据集为 $\boldsymbol{M}_A$，$\boldsymbol{M}_B$，大小为 $2\times(8N)^2$，每个处理单元负责矩阵 $\boldsymbol{M}_C$ 对应分块的计算，每个计算任务的工作数据集的大小为 $2\times8\times N^2$，即 $\alpha=0.125$. 算法的计算过程为：第一步，通过全局访主存操作 DMA 和跨步 DMA 将矩阵 $\boldsymbol{M}_A$，$\boldsymbol{M}_B$ 分块传输到对应 PE 的 LDM 中；第二步，利用存于 PE 中 LDM 的数据进行矩阵 $\boldsymbol{M}_C$ 分块的部分运算；第三步，通过快速片上通信获取位于同行和同列的其他 PE 的 LDM 中的矩阵 $\boldsymbol{M}_A$，$\boldsymbol{M}_B$ 分块的数据，完成矩阵 $\boldsymbol{M}_C$ 分块的全部运算. 此时，期望算法理论性能与处理器理论峰值性能相当，即

$$B\cdot\frac{\alpha Pk}{k+\alpha P-1}\cdot\varphi\cdot6=F_{\text{peak}}\quad(7)$$

令 $\delta=\frac{F_{\text{peak}}}{6\cdot B}$，可得

$$k=\frac{\alpha P\delta-\delta}{\alpha P\varphi-\delta}=\frac{1}{\varphi}\left(\delta+\frac{\delta^2-\delta\varphi}{\alpha P\varphi-\delta}\right) \tag{8}$$

故有 $b_{\min}=k\cdot B/P=4.83\,\text{GB/s}$，即要实现矩阵-矩阵乘算法的最高理论性能，片上网络带宽需满足 $b\geq 4.83\,\text{GB/s}$.

由式(8)可见，k 随着的 α，φ 减小而增大，即可实现数据复用的算法对片上网络带宽的需求，随着数据复用度 α 和算法计算密度 φ 的降低而增加. 在实际中，许多应用的核心算法的数据复用度和计算密度往往低于矩阵乘法，因此需要高于 4.83 GB/s 的片上网络带宽. 同理，可以得到其他基线应用的量化分析结果，硬件可以根据选定的多种基线应用，设计更加合理的聚合片上通信带宽.

4.2 模型对应用的支撑

本小节将以实际课题中抽取的两个典型算子矩阵矩阵乘和稀疏矩阵向量乘为例，来验证本文提出的众核并行计算模型 PALN 的有效性和优越性；以大规模实际应用为例，验证大规模并行计算模型 P-PALN 的性能和有效性. 采用新一代神威超算系统进行测试，单核组访存带宽 $B=50\,\text{GB/s}$，PE 个数 $P=64$(单核组的从核个数).

4.2.1 矩阵-矩阵乘

基于 4.1 节的分析，针对矩阵乘算子，将系统和算法的实际参数代入模型的访存性能公式，可以算得 2 个模型下理论的访存时间. 从理论结果看，PRAM 并行计算模型的访存时间是 PALN 并行计算模型的 3.91 倍. 实际对比中，我们在新一代神威超算系统上分别运行基于 PALN 并行计算模型和 PRAM 计算模型的矩阵乘并行算法，测量其访存时间，结果分别为 0.402 ms 和 1.521 ms. PRAM 模型的访存时间是 PALN 模型的 3.78 倍，与理论分析较为吻合. 从而验证了本文所提出的 PALN 模型相比于传统的 PRAM 模型的模型优越性.

4.2.2 稀疏矩阵向量乘

以非结构网格典型算子稀疏矩阵向量乘(Sparse Matrix-Vector Multiply，SPMV)$\boldsymbol{A}\times\boldsymbol{x}=\boldsymbol{b}$ 为例，其并行计算流程如图 6 所示. 首先，对稀疏矩阵 $\boldsymbol{A}$ 按行进行并行划分，每个处理单元 CPE 负责划分到的矩阵块的计算；第一步，通过全局访主存操作 DMA 将划分好的稀疏矩阵块的非零元数组和向量 $\boldsymbol{x}$ 的小块导入到 CPE 的 LDM 中；第二步，通过快速片上通信得到所依赖其他 CPE 中存储的向量 $\boldsymbol{x}$ 的小块数据；第三步，每个 CPE 并行地进行各自部分的稀疏矩阵乘计算，并将计算结果通过 DMA 传回主存中结果向量 $\boldsymbol{b}$ 的相应位置. 上述过程中，第二步利用了片上通信传输，提高了数据复用率，对应本文提出的片上通信并行计算模型 PALN；若将片上通信传输替换成全局访主存操作 DMA，则对应 PRAM 模型.

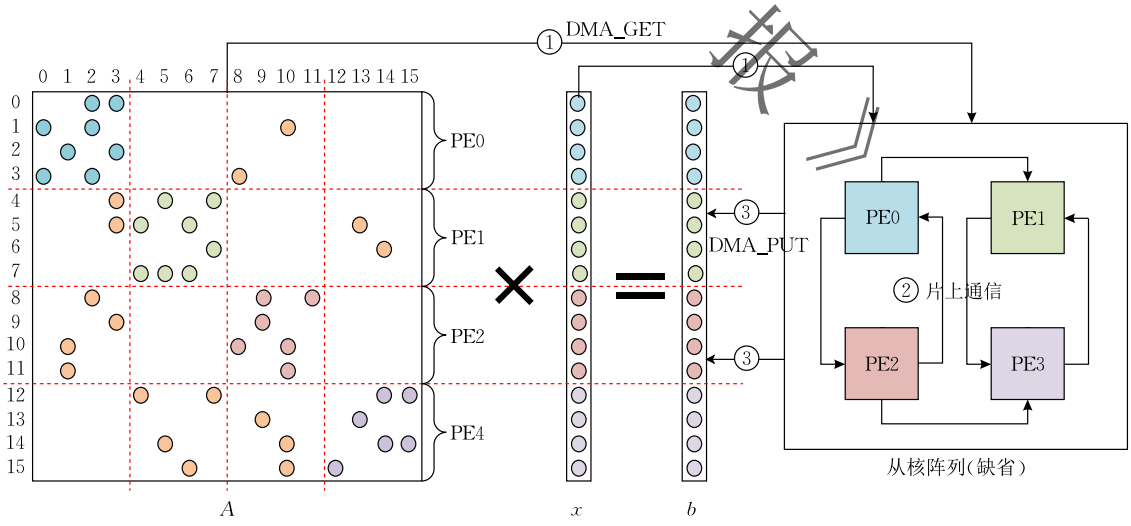


图 6 SPMV 并行计算流程示意图

采用某实际算例测试，单计算任务的工作数据集占并行任务工作数据集的比例 $\alpha=3/128$，矩阵维度为 69 147×69 147，非零元个数为 186 294，数据格式为双精度浮点型，并行计算任务的工作数据集的数据量 $Pn=186\,294\times(8+4)+69\,147\times8\times2=3263.6\,\text{KB}$，每个 PE 工作数据集的数据量 $n=51\,\text{KB}$，片上通信带宽 $b_2\approx1.688\,\text{GB/s}$(all to all 带宽在该应用场景下的调和平均值).

根据公式可计算出 PALN 并行计算模型下矩阵乘的访存时间为

$M_{\mathrm{PALN}}=Pn/B+(\alpha P\cdot n-n)/b_2=0.0687\ \mathrm{ms}$.

PRAM 并行计算模型下矩阵乘的访存时间为 $M_{\mathrm{PRAM}}=\alpha Pn\cdot P/B=4.78/(50*1024)=0.0934\ \mathrm{ms}$.

可以看出，针对 SPMV 算子，本文所提出的 PALN 并行计算模型的访存时间比 PRAM 并行计算模型低 26.4%. 在新一代神威超算系统上分别运行基于 PALN 并行计算模型和基于 PRAM 计算模型的稀疏矩阵向量乘算法，实测其访存时间分别为 0.0734 ms 和 0.0978 ms，性能提升了 24.9%，与理论分析相吻合，从而验证了本文所提出的 PALN 模型的有效性和相比于 PRAM 模型的优越性.

4.2.3　大规模应用

表 4 列举了新一代神威超算系统上十余个典型科学计算大规模应用在不同并行计算模型下的应用性能. 可以看出，在本文提出的并行计算模型 P-PALN 的支撑下，多个大规模应用的性能都实现了高效提升. 其中，神威量子模拟器[17-18]、拉曼光谱模拟[19-20]等应用均扩展到 4000 万个计算核心，并入围 2021 年“戈登·贝尔奖”(简称 GB 奖)，量子模拟器、拉曼光谱模拟、航空发动机燃烧室模拟、PageRank 图遍历、超大规模预训练框架“八卦炉”[21]、智能加速原子动力学蒙特卡罗方法 TensorKMC[22] 等应用中使用 P-PALN 模型比 BSP/LogP+PRAM 模型(简称 P-PRAM 模型)性能提升了 10%～60%，通过片上网络通信大幅提高了应用访存效率和整体性能.

值得注意的是，2021 年“戈登·贝尔奖”提名应用之一——托克马克等离子体模拟[23] 和分子动力学层状材料模拟 LMFF[24] 等应用的片上网络通信需求较小，因此使用 P-PALN 模型和 P-PRAM 模型模型的性能相同. 正如 4.4 节所述，当片上网络带宽 $b=0$ 时，P-PALN 并行计算模型等同于 P-PRAM 模型并行计算模型.

表 4　不同并行计算模型下典型课题性能差异

应用领域	应用课题	计算核心数	问题规模	P-PRAM 模型应用性能	P-PALN 模型应用性能
量子计算模拟	神威量子模拟器(2021 年 GB 奖)	4 千万	100 量子比特 40 层量子随机电路	1.08 EFlops	4.4 EFlops
第一性原理计算	新冠病毒拉曼光谱模拟(2021 年 GB 奖提名)	4 千万	3006 原子	232.4 PFlops	813.7 PFlops
电磁模拟	托克马克等离子体模拟(2021 年 GB 奖提名)	4 千万	111.3 万亿粒子，257 亿网格	201.1 PFlops	201.1 PFlops
人工智能	大模型预训练框架“八卦炉”	近 4 千万	174 万亿参数	311.9 PFlops	1.18 EFlops
材料计算	分子动力学用于层状材料的模拟 LMFF	近 3 千万	20 万亿粒子、1 纳秒	210 PFlops	210 PFlops
材料计算	智能加速原子动力学蒙特卡罗方法 TensorKMC	3 千万	50 万亿粒子	121 PFlops/Byte	509 PFLops/Byte
航空航天(结构网格类)	高超声速飞行器数值模拟	4 千万	1 千亿网格	192.75 PFlops	212 PFlops
航空航天(非结构网格类)	航空发动机燃烧室模拟 Saturne	1 千万	10 亿网格	0.745 PFlops	1.05 PFlops
生物药物类	药物筛选 Vinado	1 千万	1000 万分子对接	3.12 PFlops	3.12 PFlops
天气气候	海洋环流模式 SWNEMO	3 千万	分辨率 500 m	1.88 PFlops	1.97 PFlops
材料应用	纳米材料高分辨率模拟	4 千万	86016 个粒子	1.4 EFlops	1.4 EFlops
大规模图计算	PageRank	3 千万	2^{43} 个顶点	6891.5 GTEPS	10888.1 GTEPS

5　结束语

本文针对“神威·太湖之光”系统和新一代神威 E 级原型系统，提出 P-PALN(Parallel-Parallel Access via LDM & NOC)并行计算模型，对于计算节点间的并行，该模型沿用 BSP/LogP 模型描述；对于计算节点内的众核并行，该模型提供私有存储访问和片上阵列通信的众核并行架构的有效描述，能够协助用户进行众核并行算法设计，并在申威众核处理器硬件设计中指导参数的持续优化. 实验结果表明，该模型可有效指导硬件设计和用户众核编程，提高系统和应用的性能. 该模型同样适用于其他具有片上网络通信功能的众核架构处理器，如 NVidia H100 处理器等，具有较好的应用前景.

参　考　文　献

[1] Gao J, Zheng F, Qi F, et al. Sunway supercomputer architecture towards exascale computing: Analysis and practice. Science China Information Sciences, 2021, 64(4): 1-21

[2] Fu Hao-Hua, Liao Jun-Feng, Yang Jin-Zhe, et al. The Sunway Taihulight supercomputer: System and applications.

Science China Information Sciences, 2016, 59(7): 072001

[3] 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

[4] Chen Guo-Liang, Miao Qian-Kun, Sun Guang-Zhong, et al. Layered models of parallel computation. Journal of University of Science and Technology of China, 2008, 38(7): 481-487 (in Chinese)
(陈国良，苗乾坤，孙广中等. 分层并行计算模型. 中国科学技术大学学报，2008，38(7)：481-487)

[5] Wang Huan, Du Zhi-Hui. Contrastive analysis of parallel computation model. Computer Science, 2005, 32(12): 142-145(in Chinese)
(王欢，都志辉. 并行计算模型对比分析. 计算机科学，2005，32(12)：142-145)

[6] Mattson T G, Sanders B, Massingill B. Patterns for Parallel Programming. Boston, USA, 2005

[7] Chen Guo-Liang. The Design and Analysis of Parallel Algorithms. Beijing: Higher Education Press, 2002(in Chinese)
(陈国良. 并行算法设计与分析. 北京：高等教育出版社，2002)

[8] Zhang Y Q, Chen G L, Sun G Z, et al . Models of parallel computation: A survey and classification. Frontiers of Computer Science in China, 2007, 1(2): 156-165

[9] Fortune S, Wyllie J. Parallelism in random access machines// Proceedings of the 10th Annual Symposium on Theory of Computing. San Diego, USA, 1978: 114-118

[10] Goldschlager L M. A universal interconnection pattern for parallel computers. Journal of the ACM, 1982, 29(4): 1073-1086

[11] Cole R, Zajicek O. APRAM: Incorporating asynchrony into the PRAM model//Proceedings of the 1st Annual ACM Symposium on Parallel Algorithms and Architectures. Santa Fe, USA, 1989: 158-168

[12] Valiant L G. A bridging model for parallel computation. Communications of the ACM, 1990, 33(8): 103-111

[13] Culler D, Karp R, Patterson D, et al. LogP: Towards a realistic model of parallel computation//Proceedings of the 4th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. San Diego, California, USA, 1993: 1-12

[14] Gibbons P B. A more practical PRAM model//Proceedings of the 1st Annual ACM Symposium on Parallel Algorithms and Architectures. Santa Fe New Mexico, USA, 1989: 158-168

[15] Zhang Yun-Quan. DRAM(h): A parallel computation model for high performance numerical computing. Chinese Journal of Computers, 2003, 26(12): 1660-1670(in Chinese)
(张云泉. 面向高性能数值计算的并行计算模型 DRAM(h). 计算机学报，2003，26(12)：1660-1670)

[16] Liu Xin, Guo Heng, Sun Ru-Jun, et al. The characteristic analysis and exascale scalability research of large scale parallel applications on Sunway TaihuLight supercomputer. Chinese Journal of Computers, 2018, 41(10): 2209-2220(in Chinese)
(刘鑫，郭恒，孙茹君等. “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. 计算机学报，2018，41(10)：2209-2220)

[17] Liu Yong, Liu Xin, et al. Closing the “quantum supremacy” gap: Achieving real-time simulation of a random quantum circuit using a new Sunway supercomputer//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'21). City of Saint Louis, USA, 2021: 1-12

[18] Li Fang, Liu Xin, et al. SW_Qsim: A minimize-memory quantum simulator with high-performance on a new Sunway supercomputers//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis(SC'21). City of Saint Louis, USA, 2021: 1-13

[19] Extreme-scale ab initio quantum raman spectra simulations on the leadership HPC system in China//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'21). City of Saint Louis, USA, 2021: 1-13

[20] Shang H, Li F, Zhang Y, et al. Accelerating all-electron *Ab initio* simulation of raman spectra for biological systems//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'21). City of Saint Louis, USA, 2021: 1-15

[21] Ma Zixuan, et al. BaGuaLu: Targeting brain scale pretrained models with over 37 million cores//Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. Seoul, Republic of Korea, 2022: 192-204

[22] Shang H, Chen X, Gao X, et al. TensorKMC: Kinetic Monte Carlo simulation of 50 trillion atoms driven by deep learning on a new generation of Sunway supercomputer//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis(SC'21). City of Saint Louis, USA, 2021: 1-14

[23] Xiao Jianyuan, et al. Symplectic structure-preserving particle-in-cell whole-volume simulation of tokamak plasmas to 111.3 trillion particles and 25.7 billion grids//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'21). City of Saint Louis, USA, 2021: 1-13

[24] Gao P, et al. LMFF: Efficient and scalable layered materials force field on heterogeneous many-core processors//Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis(SC'21). City of Saint Louis, USA, 2021: 1-14

GAO Jian-Gang, M. S., senior engineer. His research interests include high-performance interconnection network and computer architecture.

LIU Xin, Ph. D., researcher. Her research interests include parallel algorithms and applications.

LI Fang, Ph. D., associate researcher. Her research interest is high-performance computing applications.

LIU Yong, Ph. D., associate researcher. His research interest is parallel algorithms.

PENG Da-Jia, Ph. D., engineer. His research interests include parallel computing and debugging.

CHEN Xin, M. S., engineer. His research interests include parallel algorithms and applications.

CHEN De-Xun, Ph. D., researcher. His research interest is high-performance computing applications.

Background

Parallel computing model is an important bridge between supercomputer designers and application developers. Based on the requirements of parallel algorithms in large-scale applications, parallel computing model abstracts the basic features of real parallel computer architecture, including computing, memory access and communication, and provides an interface for hardware design and application development. As supercomputers based on heterogeneous many-core processors have become the mainstream of TOP500 high-performance computers, existing parallel computing models designed for supercomputers based on multi-core processors, such as BSP, LogP and PRAM, cannot meet the increasingly urgent development needs of many-core architecture-based supercomputers and applications. Aiming at the Sunway TaihuLight system and the Sunway exa-scale prototype supercomputing system, a hybrid parallel computing model P-PALN(Parallel-Parallel Access via LDM & NOC) is proposed based on the characteristics of many-core architecture, which uses a hybrid parallel mode of the access to the CPE LDM and the on-chip communication to describe the many-core parallel architecture. The PALN model effectively assists users in the design of many-core parallel algorithm, and guides the continuous parameters' optimization of Shenwei many-core processor.

This project is supported by the National Science and Technology Major Project "Research on method and software of high-fidelity numerical simulation of aero-engine and gas turbine combustor". The research achieves method design and software implementation of high-fidelity numerical simulation of aero-engine and gas turbine combustor. The research team has long-term technology accumulation in the high-performance computing field and has accomplished Sunway TaihuLight supercomputer in 2015, supported by the 863 project. Sunway TaihuLight won the first place of Top500 for 4 times.