面向 E 级计算的功耗管理技术

高剑刚 龚道永 吴 伟 郑 岩 朱 琪 王 飞 郑 方 金利峰

(国家并行计算机工程技术研究中心 北京 100190)

摘 要 E 级计算机的构建面临严峻的"功耗墙"问题.为了应对功耗挑战,本文面向神威 E 级系统提出了一套低功耗管理体系.该体系采用软硬件协同的多层次低功耗管理机制,主要技术包括高能效基础设施设计、低功耗编译优化和细粒度功耗运行时管理,在系统功耗量化监测技术支撑下实现软硬件协同能耗控制集成,具有功能层次多、覆盖面广、节能效果明显的特点.本文基于神威 E 级原型超级计算机进行了系统验证,实验结果证明本文提出的低功耗管理体系能够显著降低系统功耗,并且可扩展性良好,在大规模系统中具有广泛的适应性,能够支撑 E 级计算机绿色运行.

关键词 E 级计算机;异构众核处理器;功耗管理;编译优化;运行时优化

中图法分类号 TP302 **DOI 号** 10.11897/SP.J.1016.2022.01373

Power Management Technology for Exascale Computing

GAO Jian-Gang GONG Dao-Yong WU Wei ZHENG Yan ZHU Qi WANG Fei
ZHENG Fang JIN Li-Feng

(*National Research Center of Parallel Computer Engineering and Technology, Beijing* 100190)

Abstract As it is known to all, Power Wall is one of the biggest challenges that people have to face when they are building an Exascale computer system. Thousands of works have been done to reduce the power or energy consumption on all hierarchies of the computer system. However, the existing studies only focus on the traditional architecture. In the Exascale era, as the scale and complexity of systems increase, the practicability is not guaranteed. In order to overcome the aforementioned challenge, we propose a low power management system for the Sunway Exascale system. The low power management system has a several of innovative technologies, including the high energy-efficient infrastructure design, the low-power compilation optimization and the fine-grained power consumption runtime management technology. First, the high energy-efficient infrastructure design focuses on the power supply and cooling subsystem to achieve the high efficient power conversion by monitoring the power level of the whole system. Second, the low-power compilation optimization is proposed to reorder the instruction sequence elegantly which is based on the innovation design of the underlying architecture. Specifically, the memory addresses of the instructions are reassigned to reduce the data movement in the high speed LDM (Local Device Memory). The instructions are also scheduled to meet the constraints of the register bypassing requirements, and the loop structure is reshaped to let the instructions of the loop stay in the L0 cache as long as possible. We utilize this low-power compilation optimization

收稿日期:2020-12-12;在线发布日期:2021-09-16.本课题得到国家重点研发计划项目(2016YFB0200500)资助.**高剑刚**,硕士,正高级工程师,主要研究方向为计算机体系结构、高性能互连网络.E-mail:13701512205@139.com.**龚道永**,硕士,高级工程师,主要研究方向为并行操作系统、功耗管理、系统容错管理.**吴 伟**,硕士,工程师,主要研究方向为并行编译与优化.**郑 岩**(通信作者),学士,高级工程师,主要研究方向为并行操作系统、功耗管理、系统容错管理.E-mail:zyzhengyanzy@263.net.**朱 琪**,博士,助理研究员,主要研究方向为编译与运行时优化、功耗管理、异构计算.**王 飞**,硕士,副研究员,主要研究方向为并行编译与优化.**郑 方**,博士,副研究员,主要研究方向为计算机体系结构.**金利峰**,博士,高级工程师,主要研究方向为计算机体系结构.

by reducing some redundant operations in the processors in order to improve the power efficiency. Third, based on the Dark Silicon technology, the fine-grained power consumption runtime management consists of multilevel low power scheduling in the computer system. In the node level, with the help of the operating system, an automatic quick switch of the power supply for each node is introduced according to the status of the node. In the job level, the frequency adjustment and the sleep/run state management for the many-core array is proposed so that the many-core processors can be turned into a low power status when no workload needs to be executed. In the system level, we introduce a hierarchy parallel framework to achieve the power assignment for the high scale job management, in which the resources are grouped into several small sets, and the status of the resource group is adjusted individually. What is more, the proposed low power management system realizes the integration of hardware and software collaborative energy consumption control under the support of the system power consumption quantification monitoring technology. Our system shows some great advantages such as multiple hierarchies, wide coverage and energy saving. We do many experiments on the Sunway Exascale Prototype Supercomputer and the results show that the low-power management system proposed in this paper can significantly reduce the power consumption of the system. The system also has good scalability and can be widely adapted to various scenarios in the large-scale systems. The low power management system provides a practical way to alleviate the power constraint dilemma in the future Exascale computing field.

Keywords exascale computer; heterogeneous many-core processor; power management; compilation optimizing; runtime optimizing

1 引 言

超级计算机在提供极高性能的同时，也带来了巨大的系统功耗，面临着巨额的运行成本. 2020 年 11 月，TOP500 第一名的“富岳”超级计算机的峰值性能已经达到 537.2PFLOPS，相比 2008 年同档次系统，计算性能提高了 400 倍，能效提升了 30 倍，系统功耗大幅提升（参见图 1）. 从系统成本的角度考虑，美国能源部认为第一代 E 级系统的功耗应不超过 20 MW[1]. 但是，以现有技术水平进行推算，将目前系统（TOP1“富岳”功耗为 30 MW、TOP2 Summit 功耗为 10 MW）扩展到 E 级（Exascale，每秒百亿亿次）规模，整机运行功耗分别为 57 MW 和 72 MW，每年耗电约上亿度. 中、美、日、欧等国家和组织都计划在 2021～2023 年左右推出 E 级计算机，为了解决系统“功耗墙”问题，迫切需要在功耗管理技术方面进行突破和创新，构建功耗可承受的绿色节能系统.

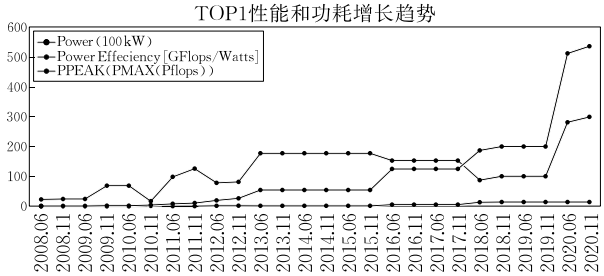


图 1 TOP1 系统性能、能效、功耗增长倍数变化曲线

2 背 景

低功耗技术研究范围广泛，涵盖了从系统架构、底层硬件到系统软件的各个层次.

（1）硬件低功耗设计技术. 该层次主要包括低电压设计、低功耗器件逻辑设计、互连网络低功耗优化、泄漏电流控制、芯片布局封装、多时钟域控制、先进制程工艺等技术[2]. 围绕软件控制需求，硬件低功耗设计提供资源睡眠、局部关闭、动态门控、频率调节等基础支撑技术. 硬件低功耗研究主要聚焦计算机底层硬件的低功耗设计，主要以芯片设计和制造为中心，通过硬件工艺和逻辑设计优化，为软件控制提供低功耗支持.

（2）基础设施低功耗技术. 该层次主要围绕系统运行所需的供电冷却等基础设施开展节能设计，降低基础设施的能耗需求. 基础设施低功耗研究主

要包括直流、分级变压等高效供电[3]，风冷、液冷、浸泡冷却、金属冷却、变频冷却等高效冷却[4]，以及储能、回收等能量重用等技术[5]等，在降低基础设施能量消耗的同时，力求提高系统的电源运行效率.

(3) 低功耗控制与优化技术. 该层次主要包括静态与动态电压调节(Dynamic Voltage Scaling, DVS)[6]技术、节点关闭及核心关闭[7]技术、动态电压频率调节(Dynamic Voltage and Frequency Scaling, DVFS)技术[8]、资源休眠/睡眠技术[9]、资源优化调度技术等. 根据使用情况，通过动态关闭空闲资源、调整资源工作频率/供电电压、优化使用模式、作业调度与指令调度等手段，在引入一定性能开销的前提下，尽可能降低资源功耗.

日本“富岳”超级计算机系统注重电源管理效率的提升，采用一系列的电源优化设计，比如时钟门控、电源旋钮和高效散热等技术，实现了高效能单处理器，在 HPL(High Performance Linpack)应用背景下达到 15 GF/W. 该技术实现了可控电源配置下的功耗调度，有效降低了系统运行功耗[10].

Intel 公司开发的全局可扩展开放电源管理器 GEOPM，是一个面向 E 级系统的开源电源管理运行时架构. 该架构基于应用程序的反馈来确定节点关键功耗路径，通过对处理器功率上限进行精细化调整，能够在给定功耗限制条件下最大化处理器性能，减少应用执行时间[11].

目前主流的低功耗技术方法还难以满足 E 级系统构建的需要，主要是因为：(1) 超级计算机的处理器设计、整机构建、系统运维等通常由不同厂商/组织负责，各自的利益诉求不同，很难围绕同一台机器的能效目标开展系统化、协同化设计，功耗管理存在一定的局限性；(2) 国内外的超级计算机主要基于商品化器件构建，以多核节能方法为主，缺少有效的面向众核结构的节能手段，与高性能处理器众核化的发展趋势不符；(3) 超级计算机高性能工作模式和服务器集群系统不同，其管理层次和模式相对复杂，单一的功耗管理策略无法满足超大规模 E 级系统需求.

针对以上问题，神威 E 级原型机重点围绕异构众核芯片的节能需求，通过基础硬件、操作系统、编译器、控制管理等多层次协同设计，建立了以国产异构众核处理器为核心的低功耗管理体系，取得了明显的节能效果，为 E 级系统设计开辟了新的技术路线.

本文主要贡献有 3 个方面：(1) 本文提出了面向国产异构众核系统的低功耗管理体系，开辟了一条新的异构众核系统节能设计的技术路线；(2) 本文提出了软硬件协同的多层次功耗管理机制，主要技术包括高能效基础设施设计、低功耗编译优化和细粒度功耗运行时管理三个方面的低功耗技术，从多个层面入手降低 E 级系统功耗；(3) 本文基于系统功耗量化监测技术，实现了软硬件协同能耗控制集成，在神威 E 级原型机平台上，对提出的低功耗管理体系进行了验证与评估，为 E 级系统的构建奠定了理论和实践基础.

3　神威 E 级原型机的低功耗管理体系

3.1　软硬协同的多层次功耗管理

众所周知，处理器功耗是影响高性能系统总功耗的主要因素. 现代处理器正朝着多核化、众核化、异构化方向发展，体系结构方面的创新已经成为处理器性能进一步提升的主要推动力. 面对日趋复杂的处理器结构，构建 E 级系统需要注重在硬件和软件协同层面应对“功耗墙”带来的挑战. 系统体系结构应更多地从软件应用需求出发，发挥定制化优势，提高硬件系统对特定应用领域支撑的效能；系统软件应充分利用新质硬件体系结构的定制化支撑，实现具备以功耗为中心的运行时/指令调度，从应用逻辑出发，精细化动态分配功耗资源，提升全系统运行效能.

在传统的高性能计算机上，通过在处理器架构中部署 DVFS(动态电压和频率缩放)、DCT(动态并发节流)和功率限制等节能技术，可以降低处理器运行功耗，节省大量电力成本. 但是，由于缺乏对软件应用逻辑与状态的有效感知，很难对系统运行性能/功耗比进行精细化描述与评估. 功耗受限的条件下，E 级计算需要在性能和功耗复杂度之间进行权衡，通过软硬件协同的监测与控制，在有限功耗下最大化系统性能.

神威 E 级原型机由 1024 个申威 26010＋众核处理器(简称 SW26010＋)组成，于 2018 年部署在国家超级计算济南中心[12]. 如图 2 所示，神威 E 级原型机基于软硬件协同的设计思想，引入了层次功耗管理体系. 如图 2 所示，该体系采用软硬件协同的多层次低功耗管理机制，主要技术包括高能效基础设施设计、低功耗编译优化和细粒度功耗运行时管理，在系统功耗量化监测技术支撑下实现软硬件协同能耗控制集成.

图 2 软硬件协同的多层次功耗管理体系

其中，高能效基础设施设计主要以供电和冷却系统的节能降耗为中心，采用系统功耗量化监测技术，通过 实现系统全负载状态下实现高效电能变换. 图 3 显示了某矩阵乘程序运行的实时功耗情况，通过功耗感知机制，可以准确统计系统的利用率和作业时间内的功耗差异分布，为整机系统以操控系统的电源配置调整提供了有力支持.

图 3 矩阵乘程序实时功耗

3.2 低功耗编译优化

系统能耗与处理器功耗有着直接的关系. 申威众核处理器采用片上融合的异构体系结构，为降低处理器取指和译码能耗，设计了基于指令窗口的指令缓冲、操作数锁存两种结构级能效优化技术[13]. 为了进一步优化国产众核处理器的功耗管理效果，神威 E 级原型机引入了性能与功耗平衡的低功耗编译优化. 基于硬件低功耗指令，通过指令调度/指令插桩，挖掘国产众核处理器功耗优化潜力，实现处理器的节能降耗.

(1) 面向容量约束的 LDM 空间重用优化技术. 通过在链接时获得的全局 LDM 空间使用情况，分析其时空关系；通过链接时地址分配优化实现空间重用，让用户更加方便直观地根据需求重用片上存储器空间. 该优化支持多段空间重用，减少不必要的主存访问和数据移动，更加灵活有效地提高运算核心局存空间的利用效率，提升应用能效.

(2) 面向寄存器旁路和总线翻转的低功耗指令调度技术. 如图 4 所示，面向寄存器旁路低功耗指令调度技术自动识别片上寄存器可旁路的场景，通过寄存器换名和指令调度(见算法 1)等优化技术，应用寄存器旁路优化，增加寄存器可旁路的机会，从而减少数据在寄存器间的移动，减少对寄存器的读写访问次数，降低处理器系统功耗.

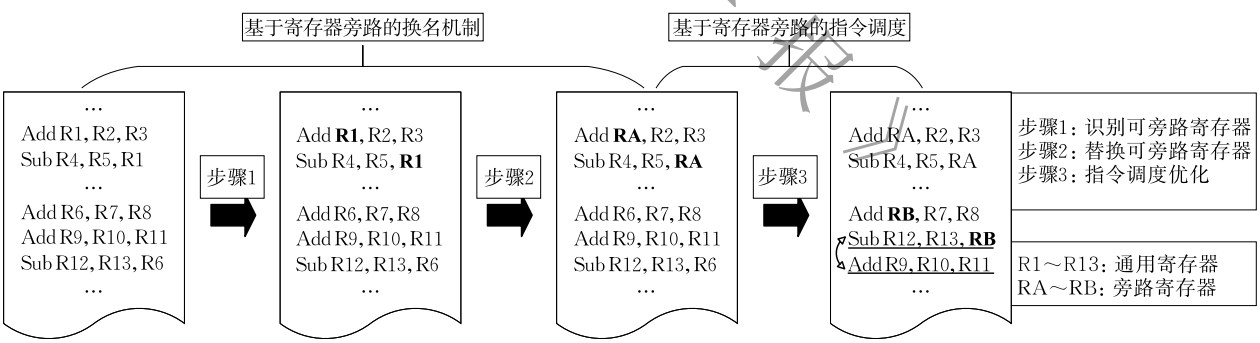


图 4 寄存器旁路优化

算法 1. 基于寄存器旁路编译优化的指令调度.

READY_LIST：*REDAY* 队列

ISSUE_LIST：*ISSUE* 队列

N：指令调度窗口大小

issue_list_tail(x)：*ISSUE* 队列中最新入队的第 x 条指令

is_RAW(x, y)：指令对 x 和 y 之间是否存在写后读依赖，x 写寄存器，y 读寄存器. 存在依赖，返回 1；不存在，返回 0.

update_priority(*insn_list*, x)：更新指令队列 *insn_list* 中指令的优先级，其中指令 x 的优先级设为最高.

1. for *insn_ready* in *READY_LIST*
2. for *index* in (1, …, *N*)
3. *insn_issue*=*issue_list_tail*(*index*)
4. if *is_RAW*(*insn_issue*, *insn_ready*)
5. *update_priority*(*READY_LIST*, *insn_ready*)

在算法 1 中，*READY* 队列保存满足发射条件的指令序列，*ISSUE* 队列保存已发射的指令序列. 根据已发射指令和未发射指令之间的数据依赖关系动态调整 *READY* 队列中指令的优先级. 具体来说，当一条指令发射后，指令进入 *ISSUE* 队列；*READY* 队列中的指令会逐一与上 n 条(n 为指令

调度窗口大小)已发射指令进行比对，判断是否存在对某个寄存器的写后读依赖：若存在，将该指令优先值设为最高；若不存在，则维持当前优先级不变.

使用面向总线翻转的低功耗指令调度技术以处理器特殊硬件特性和低功耗部件为基础进行优化. 与传统指令调度算法不同，面向总线翻转的指令调度算法是以指令序列编码翻转次数为启发函数，降低程序执行时流水线控制电路的开销，从而达到节省处理器功耗的目标.

从理论上对计算机基础部件的功耗进行分析，CMOS 的能量消耗包括动态能量消耗 P_d 和静态能量消耗 P_s 两部分，如式(1)所示.

$$P_{\text{total}}=P_d+P_s=CNV^2f+Vl_{\text{leak}} \tag{1}$$

其中，C 表示负载电容、N 表示电路每拍的信号翻转次数、V 表示电压、f 表示时钟频率、l_{leak} 表示泄漏电流. 流水线控制电路比特位翻转次数取决于指令序列的编码顺序，如果指令序列以最小化比特位翻转次数的方式重新排序，则可以降低控制电路的动态功耗.

设 I_n 为第 n 条指令，$S(I_k,I_{k+1})$ 为两条指令在处理器中比特翻转次数，BS 为总电路翻转次数，则

$$BS=\sum S(I_k,I_{k+1}),\ k=0,\cdots,n-1 \tag{2}$$

面向总线翻转的低功耗指令调度算法的目标是最小化 BS. 通过对编译流程进行修改，实时计算分析指令编码翻转率；利用贪婪算法对发射池中指令序列进行重排，最小化相邻指令对编码的差别.

(3) 面向硬件 L0 指令 Cache 的低功耗编译优化. 针对片上 L0 Cache 部件功耗低的特点，引入性能功耗平衡因子，执行多模式的循环展开和循环剥离优化；建立低功耗循环优化评估模型，实现循环优化的精细控制策略，保证 L0 Cache 的高命中率，以提高系统效能.

对于每一个循环优化模式，计算低功耗循环优化评估收益 $LOOP\_OPT\_BEN$，定义如式(3)所示.

$$LOOP\_OPT\_BEN=\frac{\sum_{i=1}^{n}\left(\left(INSN_i>\left(\frac{L0Size}{INSize}*PERF\_POW\_RATE\right)\right)?0:1\right)}{n} \tag{3}$$

其中，n 表示当前循环优化模式下的循环个数，$INSN_i$ 表示第 i 个循环的指令条数，$L0Size$ 表示 L0 Cache 的容量大小，$L0Size$ 表示单条指令的空间大小，$PERF\_POW\_RATE$ 表示性能功耗平衡因子，用于控制循环优化在功耗和性能指标上的倾向性.

针对所有循环优化模式的评估收益 $LOOP\_OPT\_BEN$，寻找所有值中的最大值. 最大评估收益 $LOOP\_OPT\_BENx$ 对应的循环优化模式即为最优模式. 根据最优模式执行循环优化的控制策略.

3.3　细粒度功耗运行时管理

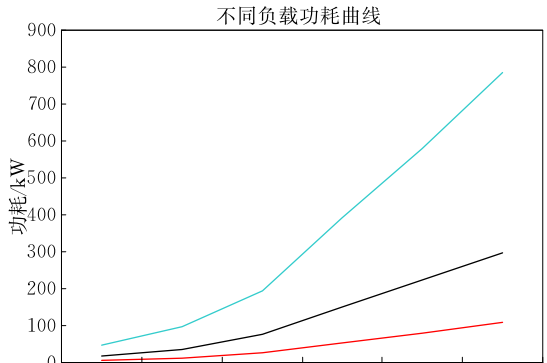
细粒度功耗运行时管理基于国产众核处理器暗硅(Dark Silicon)控制技术，从节点级、作业级和系统级等层面进行面向功耗的运行时调度，降低系统功耗，实现绿色计算.

(1) 节点级功耗管理

节点级功耗管理通过操作系统自动控制方式实现，根据众核处理器使用情况，按需进行自主节能功耗管理，实现基于节点资源微时间片忙闲状态的快速节能机制. 核心思想是利用节点内操作系统的细粒度资源分配、释放和进程加载管理机制，感知处理器核心空闲状态，控制空闲核心及阵列进行睡眠切换等方法节能；在需要使用时，及时控制处理器核心切换回正常运行状态，确保资源可用.

计算阵列是众核芯片功耗的主要来源. 当计算阵列空闲时，如果继续将阵列时钟维持在原有水平，将会损耗大量的无用功耗. 由于阵列上的所有任务加载、切换和退出都是操作系统可感知的，因此，当操作系统感知到阵列上无任务时，就断开时钟树上的阵列时钟，使阵列时钟处于断连状态以降低功耗. 当识别到需要重新使用时，则自动连接时钟树上被断开的阵列时钟，使阵列处于可用状态，加载并执行需要在阵列上执行的任务.

通过对神威 E 级原型机上的 4096 个处理器进行节点级功耗管理测试，并与基础功耗进行对比，可以发现，在节点空闲状态下，阵列睡眠可以节约节点功耗的 63%左右(图 5)，是一种比较快速、有效的异构众核资源节能方法.



	256	512	1024	2048	3072	4096
休眠/kW	6.56	12.76	27.24	53.48	80.72	107.96
空闲/kW	17.96	35.66	75.84	149.68	223.52	297.36
matmul/kW	48.12	98.34	195.38	390.6	579.56	785.92

图 5　节点级功耗管理效果

节点级功耗管理基于国产异构众核处理器结构设计，充分考虑了众核芯片的功耗分布及资源使用特点，能够跟随作业任务退出而自动生效. 由于每道作业都存在作业退出时机，采用基于时间阈值控制的芯片切换到睡眠状态前，总会存在一定时段无法睡眠，该时段内节点级功耗管理接管阵列状态，以进一步降低功耗.

(2) 作业级功耗管理

作业级功耗管理是从并行作业内多节点资源的角度，基于软硬件状态感知、资源调度管理、作业内多任务协同控制等机制的功耗管理方法. 其主要包括基于负载感知的阵列动态频率控制、作业驱动的多层次睡眠唤醒控制两个方面.

① 作业级功耗管理可以实现负载感知的阵列动态频率控制. 用户通过作业命令指示，引导系统调节计算阵列工作频率，实现降低功耗的目标. 比如 matmul 应用程序的计算核工作频率可以在 1.35G～1.45G 调整，通过动态调频，引入约 2%的性能开销，实现 5%的功耗降低，能够在作业运行时控制方面实现有效的功耗管理(表 1).

表 1 作业级功耗管理效果

优化选项	优化前/GHz	功耗/W
调频前	1.45	133.18
调频后	1.35	126.52
比率	0.93	0.95

基于负载感知的阵列动态频率控制是以异构众核芯片阵列级调频机制为基础. 高性能计算系统中运行的作业种类多种多样，对于 I/O、消息密集类型的并行作业或局部作业任务，可以将相应资源的计算阵列适当降频，在满足一定性能需求的同时，节约计算阵列运行时功耗；对于计算密集类型的作业或局部作业任务，在工作频率范围内适当超频运行，实现多任务间的工作进度匹配，避免因负载不均衡产生的不必要功耗损失，从而降低系统运行时功耗. 阵列频率控制的开销为节拍级，不同的调频幅度对应不同的节能效果，频率降幅越大，节能效果越明显.

② 作业驱动的多层次睡眠唤醒控制. 针对动态变化的作业资源分配/释放情况，采用可配置资源空闲时长阈值控制策略. 当资源空闲时间超过预设阈值时，切换空闲资源状态进入浅睡眠状态，从而降低静态功耗；在资源浅睡眠超过一定时间阈值时，切换节点状态进入深睡眠状态，进一步降低静态功耗. 当收到作业调度请求时，优先分配空闲资源；如果当前资源数量不满足需求，则继续唤醒并分配浅睡眠的资源；如果资源数量仍然不满足需求，则继续唤醒并分配深睡眠的资源，保证系统正常使用. 作业驱动的多层次睡眠唤醒控制流程如图 6 所示.

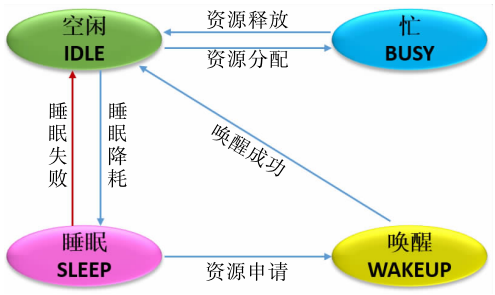


图 6 作业驱动低功耗控制示意图

(3) 系统级功耗管理

功耗管理需要对批量资源进行快速的状态切换和频率调节. 对于大规模资源的批量功耗控制，主流技术采用一对多控制方法，存在单次控制开销大，控制规模可扩展性差等问题. 在神威 E 级原型机中，系统级功耗管理采用分层并行的可扩展功耗控制管理架构，从系统批量控制模型角度进行性能/功耗优化. 其核心设计是采用分层并行、层间流水的控制思想，将传统单一层次的大规模一对多并发控制模型分解为多个并行化的小规模一对多控制模型. 同一并行层次的小规模一对多控制模型之间完全并行，不同层次的小规模控制模型之间并发流水，从而将单层的大规模一对多并行转换为分层并行的多层小规模一对多问题.

如图 7 所示，当多层并行控制模型树高为 X 层时，并行层次为 $X-1$ 层. 假设问题规模为 N(即树叶规模)，组大小为 S，树叶分组数为 G，并行层数为 L.

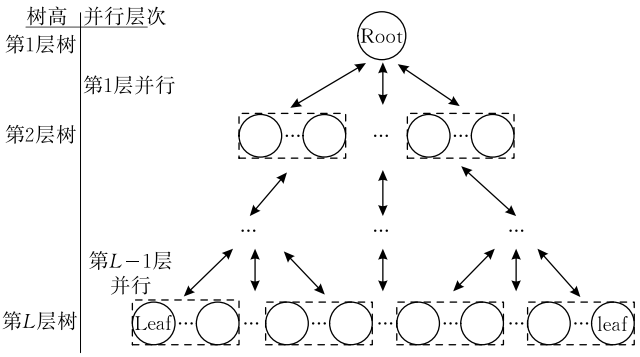


图 7 多层并行控制模型示意图

不分组时，模型为一个并行层数为 $L=1$、树叶组数 $G=1$、组大小 $S=N$ 的分组，组内并行度为 N.

分组大小为 S(每层分组大小相同)时，模型为组内并行度为 G 的一对多并行控制，其中并行层数为 $L=\log_S N$，树叶分组数为 $G=N/S$，组大小为 G.

可见，通过分层并行方法，可以将一个 $O(N)$ 规模的一个多并行控制问题，降维成一个 $O(N^{1/L})$ 规模的小规模并行控制问题，可显著降低并发控制中的单点压力，极大提升系统的可扩展性.

3.4　软硬件协同能耗控制集成

在国产众核处理器硬件方面，从结构设计到物理设计，综合应用多种低功耗设计技术，主要包括电路功耗和性能预算模型、基于门控时钟的触发器控制技术、基于多阈值部件的功耗控制、平滑过渡的片上时钟切换控制方法、支持运算核心时钟的动态调频和多粒度部件级低功耗设计，有效降低处理器动态功耗和静态功耗，提高芯片能效. 此外，硬件低功耗设计，为系统功耗控制提供了有效手段. 软硬件协同能耗控制集成为精细化性能/功耗资源的调度与决策提供了技术支撑.

通过在系统运行控制过程中采用多种软硬件协同控制集成策略，可以对符合一定条件的资源进行多策略、多状态、多粒度的能耗控制. 在当前系统状态感知、资源历史使用状况分析、资源未来使用需求预测、各种功耗控制措施的节能效果及相互关系的基础上，根据系统运行状况透明控制，实现自动节能，并支持按需人工干预.

针对国产异构众核处理器的特点，软硬件协同能耗控制集成将部分底层硬件功能封装成软件接口，由上层软件根据具体应用场景按需使用，进行能耗管理，主要包括：(1) 指令发射频度控制接口，支持系统软件按需调控处理器流水线的指令发射频度，控制芯片功耗；(2) 门控时钟软件控制接口，将底层硬件难以透明感知的切换控制开放给上层软件，由软件结合特殊应用场景进行时钟控制；(3) 动态频率控制接口，支持系统软件对计算阵列进行动态调频；(4) 部件级启停控制接口，在需要使用相应部件(如 SIMD)时动态使能，在不用时动态关闭，降低部件功耗；(5) 处理器核心关闭机制，可以按需关闭或使能空闲的处理器核心，控制处理器核心功耗；(6) 节点级睡眠唤醒接口，基于通用消息和定制协议实现处理器睡眠和唤醒控制，提供定制化睡眠及唤醒协议和接口，采用分级分层的睡眠唤醒策略，在收到外部请求时控制节点快速切换至睡眠状态或唤醒状态；(7) 计算阵列睡眠唤醒接口，处理器的控制核心支持快速控制指定阵列睡眠，并按需唤醒指定阵列.

软硬件协同能耗控制集成对低功耗管理体系内的技术进行了系统融合，利用接口设计将高能效基础设施设计、低功耗编译优化、细粒度功耗运行时管理等多层级低功耗技术上下贯通，在系统功耗量化监测支撑下，可以实现多粒度多层次的软硬件协同功耗管理和优化.

4　实验结果与分析

4.1　实验平台

神威 E 级原型机采用申威 26010＋众核处理器，该处理器架构与“神威·太湖之光”系统的申威 26010 处理器[14] 架构类似，每个处理器包含 4 个运算控制核心和 256 个运算核心. 2 个处理器通过 PCIe3.0 连接到同一片网络接口芯片，构成一个运算节点(如图 8)，每个运算节点峰值运算能力为 6.12 TFlops. 512 个运算节点通过神威互连网络相

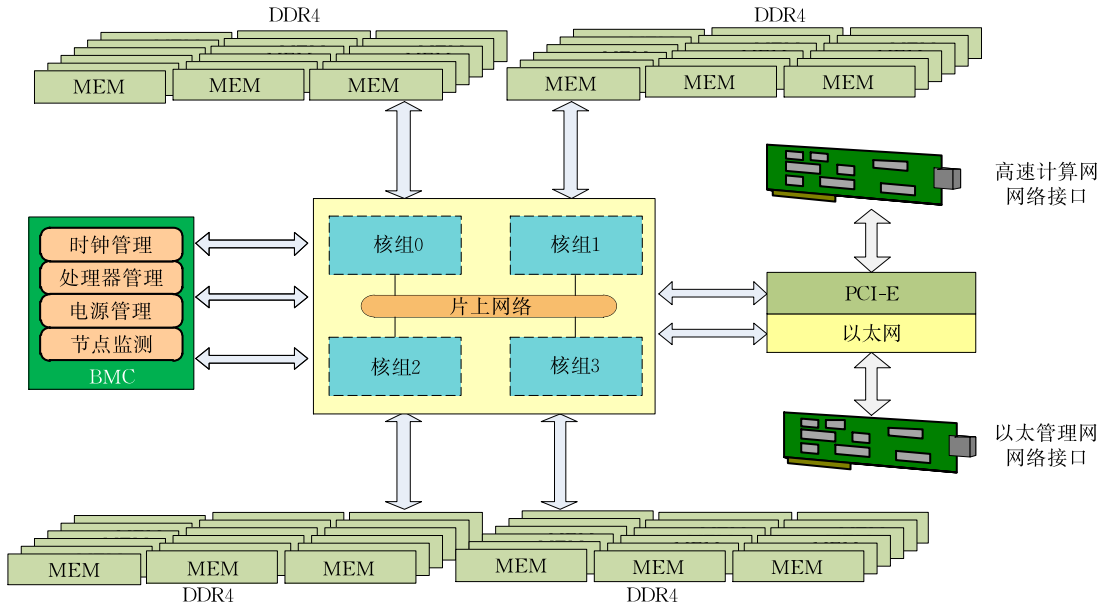


图 8　神威 E 级原型机运算节点架构

连,组成神威 E 级原型机.该原型机峰值性能达到 3.13 PFlops.

为了验证软硬件协同的多层次功耗管理技术有良好的系统节能效果,本文在神威 E 级原型机进行了相关功耗测试.另外随着计算资源规模的扩大,功耗管理技术与并行计算应用[15-17]一样,都面临可扩展性的挑战.为了验证本文提出的低功耗管理体系在 E 级计算环境下具有良好的可扩展性,本文基于"神威·太湖之光"系统进行了更大规模的可扩展性测试.

4.2 基础功耗测试

申威 26010+众核处理器由 260 个运算核心构成,每个运算核心频率配置为 1.45G,全片峰值性能为 3.068 TFLOPS.本文对单处理器在空闲状态和运行 matmul(矩阵乘)课题两种情况进行功耗测试,测试结果如表 2.

表 2 神威 E 级原型机基础功耗

CPU 数	空闲功耗/kW	运行 matmul 功耗/kW
1	0.058	0.151
4	0.274	0.677
16	0.967	2.768
64	4.124	11.320
256	17.960	48.120
1024	75.840	195.380

测试表明,系统功耗随处理器数量呈线性增长,且处理器空闲时系统静态功耗显著降低.

4.3 低功耗编译优化

基于神威 E 级原型机进行了低功耗编译优化技术测试.连续的乘加操作可以使得处理器持续处于高功耗运行状态,因此,本文采用基于矩阵乘核心段代码编写的测试用例进行功耗测试,以降低噪声对系统功耗波动的影响.测试从低功耗循环优化和低功耗调度优化两个角度入手,并对两种低功耗编译优化技术进行了集成测试,每种类型测试运行 50 遍并取均值,单个处理器测试结果如表 3、表 4.

表 3 基于低功耗循环优化的编译优化效果

优化技术	低功耗调度优化	
评估内容	功耗/W	性能/s
优化前	151.07	33.970
优化后	137.47	34.040
优化前后比率	0.91	1.002

表 4 基于低功耗调度优化的编译优化效果

优化技术	低功耗调度优化	
评估内容	功耗/W	性能/s
优化前	151.34	33.960
优化后	134.87	34.680
优化前后比率	0.89	1.021

测试表明,编译器具备性能与功耗指标均衡的编译优化支持.测试应用在性能开销最高 2.5%的情况下,采用低功耗循环优化能实现 9%的功耗下降,采用低功耗调度优化能实现 11%的功耗下降;同时采用低功耗循环优化和低功耗调度优化能够实现 12%的功耗下降,如表 5.

表 5 低功耗循环优化和调度优化的综合效果

优化技术	低功耗编译综合优化	
评估内容	功耗/W	性能/s
优化前	151.34	33.960
优化后	133.18	34.680
优化前后比率	0.88	1.025

4.4 细粒度功耗运行时管理

(1)基于负载感知的功耗管理

3.4 节介绍的节点级功耗管理和作业级功耗管理都是基于负载感知的功耗管理技术,通过功耗自动控制有效降低系统功耗.基于神威 E 级原型机 1024 个申威 26010+众核处理器规模,进行基于负载感知的功耗管理实验,对节点级功耗管理和作业级功耗管理效果进行评估.图 9 数据可以看出,通过综合运用上述管理技术,1024 处理器规模 6 h 正常运行课题平均节约功耗约 10%.

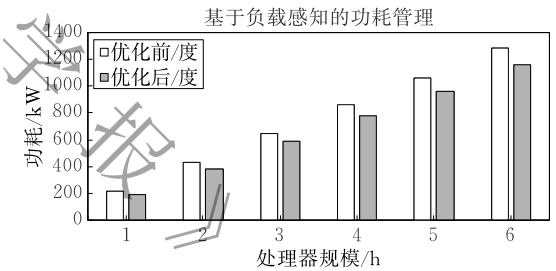


图 9 节点级功耗管理和作业级功耗管理效果

(2)系统级功耗管理开销测试

基于"神威·太湖之光"系统进行批量资源分级并行控制架构的系统级功耗管理实验.由于节点内的控制开销为常数 0.36 s,对批量控制部分的开销进行测试,实验结果如图 10 所示.

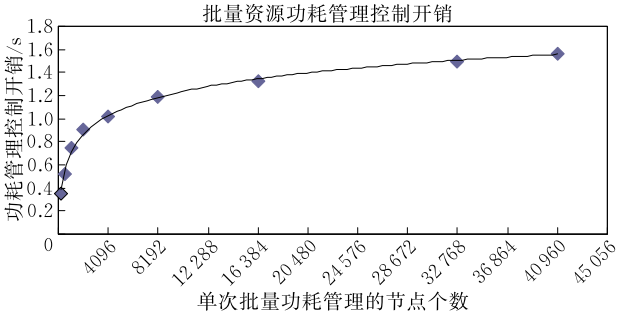


图 10 批量资源功耗管理控制开销

从 256 到 512 个节点规模，控制架构由单一层次增加到三层并行，开销呈指数级增加；从 512 到 40 960 个节点规模，控制开销随资源规模增加呈线性缓慢增长趋势. 得益于分层并行控制架构以及使用已建立的网络连接传递消息，大规模节点情况下总体开销较小. 通过理论测算，完成 10 万节点规模系统的功耗管理预计在 3 s 左右，能基本满足 E 级系统功耗管理的可扩展需要.

4.5　典型场景综合节能评估

图 11 模拟了在典型混合负载和不同机时利用率下，应用各种节能措施后神威 E 级原型机的综合节能效果. 其中，$X\%$ busy 表示机时利用率为 $X\%$（即：有 $X\%$的资源处于被各种典型应用混合使用的运行状态，有 $1-X\%$的资源处于空闲状态），机时利用率定义为

$$r=\frac{\text{系统中用户课题占用机时总和}}{\text{系统的总机时}} \tag{4}$$

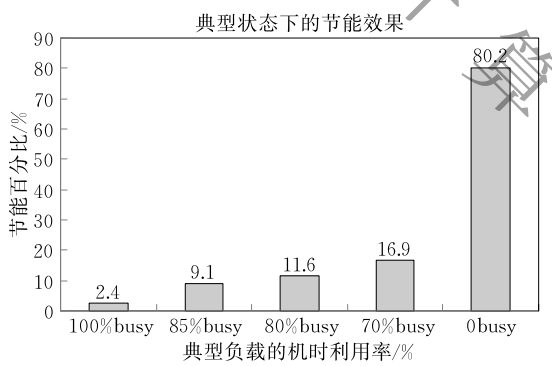


图 11　典型负载和不同机时利用率下的节能效果

由于作业运行中的编译优化、动态调频等幅度有限，而空闲状态下可调整的余地更大，因此，随机时利用率的增加，整体节能效果有所下降，即使在全机使用状态下，功耗管理手段仍有一定的节能效果. 机时利用率 85%时，平均节能效果可达 9%左右，效果较为明显.

5　结　论

目前主流的低功耗技术方法还难以满足 E 级系统构建的需要. 针对以上问题，神威 E 级原型机重点围绕异构众核芯片的节能需求，通过基础硬件、操作系统、编译器、控制管理等多层次协同设计，建立了以国产异构众核处理器为核心的低功耗管理体系.

E 级计算机的构建面临严峻的“功耗墙”问题. 主流的低功耗技术方法难以满足 E 级系统构建的需要，为了应对功耗挑战，以软硬件协同的低功耗管理思想为中心，本文围绕异构众核芯片的节能需求，通过基础硬件、操作系统、编译器、控制管理等多层次协同设计，建立了以国产异构众核处理器为核心的低功耗管理体系，该体系采用软硬件协同的多层次功耗管理机制，主要包括高能效基础设施设计、低功耗编译优化和细粒度功耗运行时管理三个方面的技术，在系统功耗量化监测支撑下实现软硬件协同能耗控制集成. 该体系具有功能层次多、覆盖面广、节能效果明显的特点.

本文基于神威 E 级原型超级计算机和“神威·太湖之光”系统进行了系统验证，实验结果证明本文提出的低功耗管理体系能够显著降低系统功耗，可扩展性良好，在大规模系统中具有广泛的适应性，能够支撑 E 级计算机绿色运行，为 E 级系统设计开辟了新的技术路线.

参　考　文　献

[1] Sarkar V, Harrod W, Snavely A E. Software challenges in extreme scale systems. Journal of Physics: Conference Series, 2009, 180(1): 012045

[2] Amelifard B, Fallah F, Pedram M. Low-power fanout optimization using multiple threshold voltage inverters//Proceedings of the 2005 International Symposium on Low Power Electronics and Design. New York, USA, 2005: 95-98

[3] Kasper M, Chen Cheng-Wei, Bortis D, et al. Hardware verification of a hyper-efficient (98%) and super-compact (2.2kW/dm^3) isolated AC/DC telecom power supply module based on multi-cell converter approach//Proceedings of the 2015 IEEE Applied Power Electronics Conference and Exposition (APEC). Charlotte, USA, 2015: 65-71

[4] Wahlroos M, Syri S, Pärssinen M, et al. Utilizing data center waste heat in district heating — Impacts on energy efficiency and prospects for low temperature district heating networks. Energy, 2017, 140: 1228-1238

[5] Hsu C-H, Kremer U. The design, implementation, and evaluation of a compiler algorithm for CPU energy reduction//Proceedings of the ACM SIGPLAN 2003 Conference on Programming Language Design and Implementation. New York, USA, 2003: 38-48

[6] Alvarruiz F, de Alfonso C, Caballer M, et al. An energy manager for high performance computer clusters//Proceedings of the 2012 IEEE 10th International Symposium on Parallel and Distributed Processing with Applications. Madrid, Spain, 2012: 231-238

[7] Patil V A, Chaudhary V. Rack aware scheduling in HPC data centers: An energy conservation strategy. Cluster Computing, 2013, 16(3): 559-573

[8] Kim W, Gupta M S, Wei G Y, et al. System level analysis of fast, per-core DVFS using on-chip switching regulators// Proceedings of the International Symposium on High Performance Computer Architecture. Salt Lake City, USA, 2008: 123-134

[9] Mustafa R M, Nishkam R, Srihari C, et al. Power management for heterogeneous clusters: An experimental study// Proceedings of the 2nd International Green Computing Conference(IGCC' 11). Orlando, USA, 2011: 1-8

[10] FujitsuLimited. https://www.fujitsu.com/global/products/computing/servers/supercomputer/documents/White paper Advanced Software for the FUJITSU Supercomputer PRIMEHPC FX1000

[11] Eastep J, Sylvester S, Cantalupo C, et al. Global extensible open power manager: A vehicle for HPC community collaboration on co-designed energy management solutions. Springer International Publishing, Cham, 2017: 394-412

[12] Gao Jian-Gang, Lu Hong-Sheng, He Wang-Quan, et al. The interconnection network and message machinasim of Sunway Exascale prototype system. Chinese Journal of Computers, 2021, 44(1): 222-234(in Chinese)
(高剑刚, 卢宏生, 何王全等. 神威 E 级原型机互连网络和消息机制. 计算机学报, 2021, 44(1): 222-234)

[13] Zheng Fang, Zhang Kun, Wu Gui-Ming, et al. Architecture techniques of many-core processor for energy-efficient in high performance computing. Chinese Journal of Computers, 2014, 37(10): 2176-2186(in Chinese)
(郑方, 张昆, 邬贵明等. 面向高性能计算的众核处理器结构级高能效技术. 计算机学报, 2014, 37(10): 2176-2186)

[14] Dongarra J. Report on the Sunway TaihuLight system. UT EECS Technical Reports, UT-EECS-16-742, USA, 2016: 1-24

[15] Xiao G, Li K, Chen Y, et al. CASpMV: A customized and accelerative SpMV framework for the Sunway TaihuLight. IEEE Transactions on Parallel and Distributed Systems, 2021, 2(1): 131-146

[16] Xiao G, Chen Y, Liu C, et al. aeSpTV: An adaptive and efficient framework for sparse tensor-vector product kernel on a high-performance computing platform. IEEE Transactions on Parallel and Distributed Systems, 2020, 31(10): 2329-2345

[17] Hong Wen-Jie, Li Ken-Li, Quan Zhe, et al. PETSc's heterogeneous parallel algorithm design and performance optimization on the Sunway TaihuLight system. Chinese Journal of Computers, 2017, 40(9): 2057-2069
(洪文杰, 李肯立, 全哲等. 面向神威·太湖之光的 PETSc 可扩展异构并行算法及其性能优化. 计算机学报, 2017, 40(9): 2057-2069)

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

GONG Dao-Yong, M. S., senior engineer. His research interests include parallel operating system, power management, and system fault tolerance.

WU Wei, M. S., engineer. His research interests include parallel computing and optimization.

ZHENG Yan, B. S., senior engineer. His research interests include parallel operating system, power consumption management and system fault tolerance management.

ZHU Qi, Ph. D., assistant professor. His research lies in the broad filed of compiling and runtime optimization, power management and heterogeneous computing.

WANG Fei, M. S., associate professor. His research interests include parallel compilation and optimization..

ZHENG Fang, Ph. D., associate professor. His research interest is computer architecture.

JIN Li-Feng, Ph. D., senior engineer. His research interest is computer architecture.

Background

Many countries and organizations, including U. S., Japan, E. U. and China, have proposed their developing maps of Exascale Computing. Comparing with the current supercomputers, constraints on power efficiency, performance, programmability, reliability for the future exascale computers are much severer. Among these challenges, as we all know, power efficiency is the Achilles' heel. For example, DoE of U. S. proposed that the performance to power ratio of the Exascale computer should be at least 50GFLOPS/W, while the best system in the current GREEN500, *A64FX Prototype*,

can only achieve 16.8GFLOPS/W. Based on the architecture of the Sunway supercomputers, our team has proposed and implemented several state-of-the-art technologies to improve the power efficiency and the energy efficiency. Software and hardware are co-designed seamlessly to face the challenge of the Power Wall. Evaluations on the *Sunway TaihuLight* system and the *Sunway Exascale Prototype* system prove the effectiveness of the proposed technologies.

This project is supported by National Key R&D Project "Verification System of the Key Techniques for the Exascale System". The research achieves the goal of innovative design of the Exascale system by constructing the large-scale verification system, mastering the techniques of new interconnection network architecture, and testing based on domestic components and parts. 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.