面向异构计算的能效感知调度研究

王静莲[1,2],龚 斌[1,3]
(1. 山东大学计算机科学与技术学院,山东济南 250101;2. 鲁东大学信息与电气工程学院,山东烟台 264025;
3. 山东省高性能计算中心,山东济南 250101)

摘 要: 异构调度可使大规模计算系统采用并行方式聚合广域分布的各种资源以提高性能.传统调度目标追时限约束求高性能而忽视高效能,远不能适应绿色计算科学发展要求.因此,本文在理论上一方面建立融合能效感知的调度模型;另一方面提出适于超计算机混合体系的多学科背景的元启发式优化算法.从技术上解决了面向不同环境目标的调度实施条件界定及调度指标(时间、能耗)实时变化描述等问题.大量仿真实验结果表明:与三个元启发式调度器相比,论文方法在能效及可扩展等方面优势明显;对于高维实例,整体性能改善分别达到8%,15%和17%.

关键词: 异构调度; 绿色计算; 协同进化; 混合并行
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2016)04-0893-05
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.04.020

Research on Energy-Efficiency Aware Heterogeneous Scheduling

WANG Jing-lian[1,2], GONG Bin[1,3]
(1. *College of Computer Science and Technology, Shandong University, Jinan, Shandong* 250101, *China*;
2. *College of Information and Electrical Engineering, Ludong University, Yantai, Shandong* 264025, *China*;
3. *Shandong High Performance Computing Center, Jinan, Shandong* 250101, *China*)

Abstract: Enabled to provide pervasive access to distributed resources in parallel ways, heterogeneous scheduling is extensively applied in large-scaled computing system for high performance. Conventional real-time scheduling algorithms, however, disregard energy-efficiency in addition to stringent timing constraints. In recognition of green computing, an energy-aware model is firstly presented. Secondly, inspired by multi disciplines, the meta-heuristic is addressed based on the super-computer hybrid architecture. On the other hand, some technological breakthroughs are achieved, including boundary conditions for different heterogeneous computing and grid scheduling and descriptions of real-time variation of scheduling indexes (stringent timing constraints and energy-efficiency). Extensive simulator and simulation experiments highlight higher efficacy and better scalability for the proposed approaches compared with the other three meta-heuristics; the overall improvements achieve 8%, 15% and 17% for high-dimension instances, respectively.

Key words: heterogeneous scheduling; green computing; co-evolution; hierarchical parallelization

1 引言

目前,大规模计算系统采用并行技术可具高吞吐信息服务和海量数据处理能力,在科学计算和金融等领域需求迅猛增长[1].随着规模的不断扩大,异构计算聚合广域分布的各种同构与异构的计算机、工作站、机群、群集、数据库、高级仪器和存储设备等资源,可形成对用户相对透明的、虚拟的高性能环境[2].并行系统效能的高低很大程度上由部署在体系架构上的资源管理系统决定[3].任务调度是资源管理的核心,为了优化某个目标函数,其在一组具有任意特性的处理机中对任务集合进行排序和资源分配[4].当前,同构调度问题已被广泛研究[5];但异构调度,鉴于其复杂性、环境的多样性、应用的新需求和调度目标的折中性等,是一个亟待解决的高维多模优化难题[6].

与此同时,绿色计算因为与环境保护和人类可持续发展的密切关联引起越来越多的社会关注[7],而高性能领域的绿色计算成为数据和计算中心实际运行的关键问题[8].数据显示,一台不关闭的普通台式电脑每年耗电1270度,释放高达0.778吨二氧化碳;一次谷歌

收稿日期:2014-10-23;修回日期:2014-12-18;责任编辑:覃怀银
基金项目:国家自然科学基金(No.61272094);国家863高技术研究发展计划(No.2006AA01A113,No.2012AA01A306)

搜索消耗足以让一只 100 瓦的灯泡工作 1 小时的能量，而两次搜索就可释放与烧开一壶水相当的二氧化碳. 据调查，我国 IT 能源消耗约占全国每年 800 亿元政府能源消耗的 50%，而拥有大量服务器的大规模计算系统（如数据和计算中心）又占到 IT 能耗总开销的 40%. 并且，信息需求量的加大催生 IT 投资以每年超过 18% 的速度增长，IT 能耗也以每年 8% ~ 10% 的速度上升[9].

另外，对高度数据密集型工作负载的支持正成为下一代计算和数据中心的关键技术. 这里，实时任务据其间的依赖性，分为独立任务或依赖任务应用；而数据密集型应用是指以数据为中心，存在海量数据传输的依赖任务应用. 计算资源的异构性、调度技术的局限性和调度指标的平衡性是面向数据密集应用调度研究所需考虑的重要因素[10].

再者，异构调度算法的研究，目前主要集中在基于需求建模的启发式算法[11]和基于进化理论的元启发式算法[12].

对任务的多项需求，启发式调度会将多目标聚合成单一目标函数处理. 这种方法简单而有效，因此被广泛研究和采用，但其自身也存在一些固有的缺陷：由于多目标优化问题的解并非唯一，而是存在一个最优解集合，称为非劣解. 而单目标优化算法仅能根据聚合函数得到决策空间的一个可行解，降低了最终解的质量，缺乏灵活性和扩展性.

更为合理的途径是采用多目标组合最优化算法来解决这个问题，基于进化理论的元启发式算法是较为有效的方法. 元启发式算法应用在异构调度问题已有十余年[13]. 但面对其复杂多样性，目前算法大多存在两个瓶颈：种群的收敛速度较慢或个体多样性不能保持. 并行与分布式元启发式算法（Parallel Distributed Metaheuristics，PDM）因在较大目标空间的搜索高效性及鲁棒性，近年来被广泛应用；具体分为四类：主从模型，细粒度模型，粗粒度模型和混合模型. 其中，粗粒度模型被广泛应用在超计算机体系结构上[14~16]. 但上述 PDM 在面向大规模实时实例时，运行算法的超计算机虽然能达到峰值需求，但很多时候效率不高；因此面向多核 CPU + GPU 混合的超计算机体系结构，元启发式算法的并行设计也是决定问题高效求解的重要元素之一.

2 数学建模

实时应用任务是由用户上传的若干子任务 $\alpha=\{\alpha_i\}(i=1,2,\cdots,m)$ 组成，具体可由一组属性集合表述：$\alpha_i=(b_i,\omega_i,\tau_i,\nu_i,\zeta_i,l_i,S_i)$. 其中，$b_i$、$\omega_i$及 τ_i表子任务的到达、估计执行和结束时间；ν_i表任务截止时间；ζ_i表计算量（周期数目）；l_i表需保护的数据量（单位：KB）；$S_i\rightarrow\kappa$（κ 是一个正实数集）表示实时任务对异构节点的安全需求集合；这里，通常是指保密、诚信和认证三种.

异构节点集合可表述为 $\phi=\{\phi_j\}(j=1,2,\cdots,n)$. 其中，每个节点 ϕ_j有不同的属性值，比如能耗 p_j、时钟频率 ζ_j和转换成本 ξ_j等.

本文前期研究[17]为弥补当前调度技术的局限和空白，解决不同约束与性能指标的冲突性等问题，并兼顾不同运行环境的差异性、不同应用的计算或者数据密集特性，有效量化了动态电压频率调整和动态电源管理，提出异构系统的能效优化函数（如式(1)所示）.

为不失一般性，假定同构系统存在某些常量，例如 $p_j=p,\zeta_j=f,\xi_j=e^d$ 以及任务间是按其 ν_i值的非递减顺序进行排序的，即如果 $j\leqslant j*$ 则 $\nu_j\leqslant\nu_{j*}$.

$$E(\boldsymbol{X})=\mathrm{Min}\left\{\sum_{i=1}^{m}\left\{(p/f)\sum_{j=1}^{n}x_i\zeta_j+e^d y_i\right\}\right\} \quad (1)$$

式(1)中，$\boldsymbol{X}$ 和 $E(\boldsymbol{X})$分别表示调度方案集及异构系统的总体能耗；对于任意 i，如果存在 $\sum_{j=1}^{n}x_i\neq 0$ x_i是 $\boldsymbol{X}$ 中的元素，则 $y_i=1$；否则，$y_i=0$.

融合能效感知的异构调度目标就是在满足任务间依赖关系同时，寻找任务需求模型与资源拓扑结构之间的映射调度方案$\{(\alpha_i,\phi_j)\}$，$i\in[1,m]$，$j\in[1,n]$，使异构计算任务的调度长度和系统节点能效收益达到最优，求尽可能多的（甚是是全部）的非劣解，可以描述为：

$$\mathrm{Min}\ Y=F(\boldsymbol{X})=(f_1(\boldsymbol{X}),f_2(\boldsymbol{X}),\cdots,f_\kappa(\boldsymbol{X})) \quad (2)$$

式(2)中，$\boldsymbol{X}\in\Phi$，$Y\in\Omega$. $\boldsymbol{X}$ 称为决策变量，Φ 是决策空间；Y 为目标函数值，Ω 是目标函数空间. $f_i(\boldsymbol{X}$い）表示异构计算任务的调度长度，$f_j(\boldsymbol{X})$表示异构系统能耗.

3 多学科启发的协同进化多目标优化算法

在免疫学中，抗原是导致免疫系统产生抗体的物质. 对于多目标优化问题，抗原被定义为目标函数（如式(2)）.

B 细胞、T 细胞及一些具抗原特异性的淋巴细胞通常称为抗体. 人工免疫系统中抗体代表抗原的一个候选解. 令 $\boldsymbol{X}$ 为抗体空间，抗体种群表示为 N-维抗体集合（N 是抗体种群规模）；且抗体由基因组成，表示为 $x_i=(G_1,G_2,\cdots,G_N)$. 在实际应用中，给定优化问题具有 N-维实数的目标搜索空间，一个候选解 x_i（$i\in[1,N]$）由 N 维实数组成；而每一维代表一个问题变量并被看作基因. 这里，基因的基本单位是基因座.

认知心理学认为模因（meme）是文化信息单位，是文化复制、传播和发展的“基因”. 模因作为一种选择与建构的创新思维和科学方法是社会进化原动力的一个表现形式. 在空间 $\boldsymbol{X}$ 中给定一个抗体 A_β，论文将抗体基

因的实时进化信息看作模因,并形式化为 $\boldsymbol{M}$. 矩阵中每一维数据都与相应的基因进化信息对应;Z、N 表示模因空间的维数.

论文多学科启发的基因-模因协同进化算法(IM_CP)较人工免疫算法有三方面改进,即除抗体个体,基因亲和度值的细粒度评估及模因的数学表述,基因-模因协同进化过程的高效模拟以及结合孤岛模型和主从模型的多层次并行化设计.

IM_CP 算法

```
1:Initializethe iteration (g) and the subpopulations {δ1,δ2,…,δC,…,δθ},each of s individuals;
2:While (g<gmax) and (other termination criteria are not satisfied)
3:  Do in parallel for each island  /* Obtain the coarse-grained model (also named as island model), one of the parallel and distributed models of metaheuristics */
4:    g=g+1;
5:    Do in parallel /* Obtain the master-slave model */
6:      Evaluate the affinity between the antibody and antigens (Eq.(2)) in the current population:Φ(Aβ);
7:      For (every couple of antibodies denoted as AU and AW)
8:        If(Φ(AU) > Φ(AW)
9:          For(C=1;C<=N;C++)
10:           For(K=1;K<=Z;K++)
11:             If(A'U GKC} <>AW' GKC)
12:               Update A'U meme vectors:
13:                 {MKC(t)=(1-ρ)×MKC(t-1)+ ΔMKC;
14:                 ΔMKC=Q(t)/Φ(AU);}
15:               Update A'U other meme vectors MJC(J<>K):MJC(t)=(1-ρ)×MJC(t-1);
16:             EndIf
17:           EndFor
18:         EndFor
19:       EndIf
20:     EndFor
21:   End Do in parallel
22:   Perform clonal selection operation;
23:   Perform clonal operation;
24:   Perform gene operations based on meme matrices;
25:   Save the best solution in the generation;
26:   If g=τ(migration interval) then
27:     Create λδ for the current subpopulation;
28:     Send λδ to the neighboring subpopulation;
29:     Receive λδ from the neighboring subpopulation;
30:     Construct the founding subpopulation σδ;
31:     Select s individuals in σδ;
32:     Replace the subpopulation λδ with λδ';
33:   End If
34:  End Do in parallel
35:End While
36:Output the best individual.
```

算法 IM_CP 第 7 步 ~ 第 20 步细粒度评估抗体基因的亲和度值并有效模拟种群自组织的模因更新;而第 22 步 ~ 第 24 步定义了模因传播并影响基因进化的过程.

面向新近发展的混合多核 CPU + GPU 的高性能计算机集群体系结构,论文提出融合粗粒度模型和主从模型的层次并行模型. 即首先依据粗粒度模型,将种群划分成若干子群,并把每个子群分配到一个节点上. 而在每一个节点上,大量的个体适应度评估计算是适于 GPU 加速的主从式并行应用;这里 CPU 可看作主服务器,而在 GPU 多核上执行的若干线程就是客户端. 第 26 步 ~ 第 33 步具体描述了多层次模型之一的粗粒度模型(也称为孤岛模型)采用的基于模因库的并行迁移策略.

表 1　实验的相关参数设置

名称	参数值(Fixed) - (Varied)
CPU 速度	(100 million instructions/second or MIPS) - (100,200,…,800)
实时任务截止时间	(1000ms) - (1000,2000,…,100000)ms
节点数目	(64) - (8,16,32,64,96,128,256)
异构节点(Severs,f(GHz),P(W))	(IBM,2.13,675)(IBM,2.13,670)(IBM,2.66,1440)(IBM,2,1350)(IBM,1.86,1975)(IBM,2.33,310)(IBM,2.26,670)(IBM,3,400)(HP,2,460)(HP,2,750)(HP,2.4,460)(HP,2.4,300)(HP,2.4,920)(DELL,2.4,345)(DELL,2.4,305)(DELL,2.13,345)(DELL,2.26,1100)(DELL,2.33,345)

4　仿真实验及结果

实验在山东省高性能计算中心进行,采用浪潮天梭 TS10000 高性能集群系统,英特尔至强 5600 系列处理器(2.66GHz,12MB Cache),CPU + GPU 混合结构,共有 960 个计算核数,计算峰值达每秒 10 万亿次双精度浮点运算,内连 40Gb/s 带宽 1 ~ 2μs 超低延迟的高速网络. 算法的并行实现采用 MPICH-VMI (MPICH 1.2.7p1 版本)[18];表 1 总结了实验集群的相关参数设置.

4.1　整体性能比较

文献[15]提出了一些新的模拟中型规模网格的任务调度实例集,并可以通过网站 http://www.fing.edu.

uy/inco/grupos/cecal/hpc/HCSP 下载应用. 这些测试集是依据文献[19]所述的建模方法随机产生的,其目的是模拟复杂的异构计算环境;实例维数(任务 × 机器)包括 1024 × 32,2048 × 64,4096 × 128 及 8192 × 256,规模远大于文献[11]的经典十二个实例.

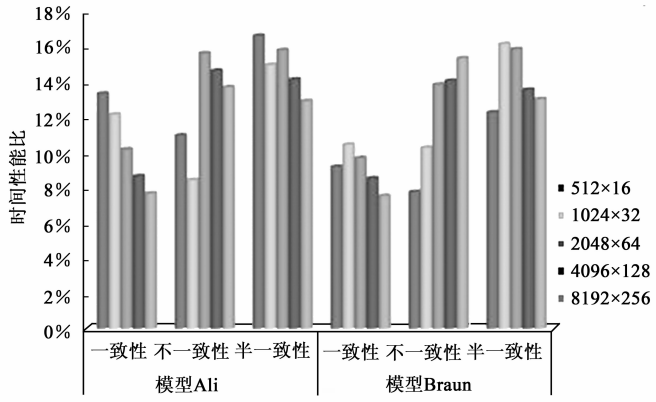

图1 与确定性启发式算法的时间性能比较

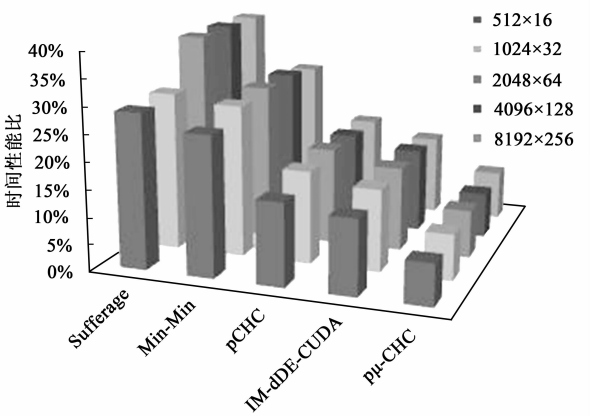

图2 不同元启发式算法的时间性能比较

首先,针对高维异构依赖任务调度问题,IM_CP 按实例的一致性、半一致性及不一致性分类与算法 Min-Min[11]和 Sufferage[11]进行性能比较. 这里一致性、半一致性及不一致性的定义依从文献[15].

Min-Min 和 Sufferage 是能在合理时间内求解低维异构调度问题的两种较好启发式算法. 从图 1 看,对于每维的一致性实例,算法 IM_CP 相较 Min-Min 和 Sufferage 的时间性能改善约为 9%,而对半一致性实例,时间性能改善上升为 12%. 另外,虽然对于低维的不一致性实例,IM_CP 相较 Min-Min 和 Sufferage 的时间性能改善不明显,但面向高维的不一致性实例,其时间性能改善超过 14%.

然后,针对上述同样实例,IM_CP 与目前面向计算机集群体系结构设计的较好的三个元启发式算法(pCHC[14],pμ-CHC[15]和 IM-dDE-CUDA[16])进行性能比较. 如图 2 总结所示,随着异构调度问题的规模增大,IM_CP 表现的性能改善越大. 值得注意的是,对于高维实例 4096 × 128,IM_CP 相较 pμ-CHC,pCHC 和 IM-dDE-CUDA 的性能改善分别达到 8%,15% 和 17%.

4.2 融合能效感知模型的影响

以高性能计算机集群节点个数为自变量,以能源效率作为函数值的四种算法的比较实验结果如图 3 所示. 图 3 表明在高性能计算机集群节点个数从 8 到 256 递增过程中,IM_CP 较其余三种算法,能源节约优势明显加强.

模型中任务最后时间期限参数对算法解的能效影响如图 4 所示. 图 4 清楚表明:模型中任务截止时间参数值 1 ~ 100 秒的递增过程中,IM_CP 求得解的能源消耗急剧下降;而在这一过程中,pCHC 的能源节约能力表现次优. 这一结果再次印证,面向高性能计算机集群算法 IM_CP 求解异构调度问题的能效感知优势,且随着异构任务的截止时间参数值的不断增大,这种优势更明显.

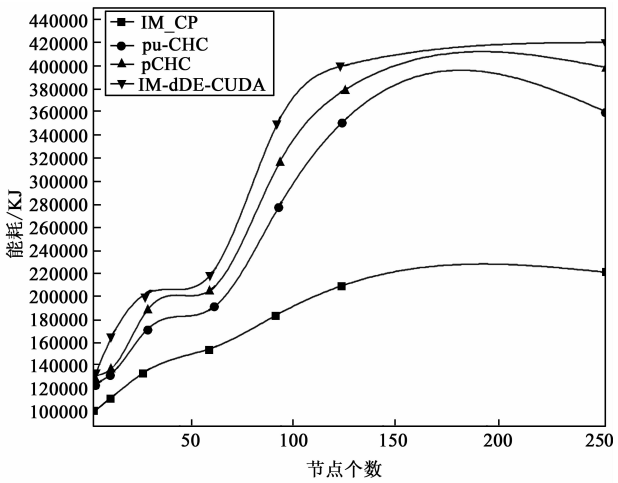

图3 不同节点数目的能效比较

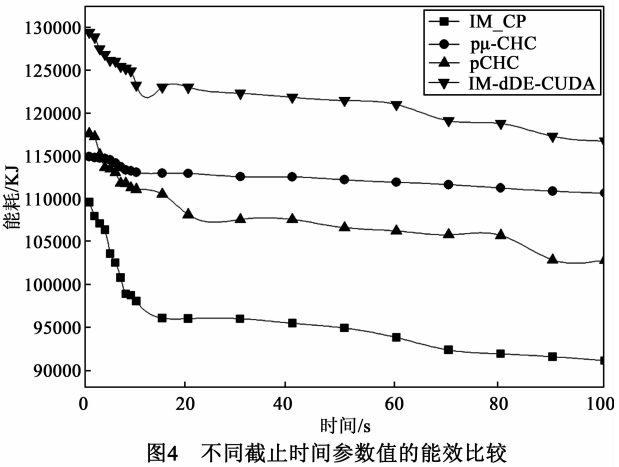

图4 不同截止时间参数值的能效比较

5 结论

本文对融合能效感知的异构调度进行研究,有效降低数据密集应用的通信开销、兼顾提供者和消费者双方的利益并保证系统双层负载均衡性. 随着新应用(数据密集应用、计算密集应用)、新环境(计算网格、云计算、多集群环境、异构环境)和新性能指标(能效)的出现,分布式计算资源管理核心技术之调度研究具有重大的理论和应用价值.

参考文献

[1] ZHONG L,et al. A task assignment algorithm for multiple aerial vehicles to attack targets with dynamic values[J]. IEEE Transactions on Intelligent Transportation Systems, 2013,14(1):236 - 248.

[2] ZHENG X H,et al. A task operation model for resource allocation optimization in business process management[J]. IEEE Transactions on Systems, Man and Cybernetics, Part A:Systems and Humans,2012,42(5):1256 - 1270.

[3] JING T, et al. Diversity-adaptive parallel memetic algorithm for solving large scale combinatorial optimization problems[J]. Soft Computing,2007,11(9):873 - 888.

[4] SALCEDO-SANZ S, et al. Hybrid meta-heuristics algorithms for task assignment in heterogeneous computing systems[J]. Computers & Operations Research, 2006, 33(3):820 - 835.

[5] GONG Y J,et al. An efficient resource allocation scheme using particle swarm optimization[J]. IEEE Transactions on Evolutionary Computation,2012,16(6):801 - 816.

[6] CHITRA P,et al. Application and comparison of hybrid evolutionary multiobjective optimization algorithms for solving task scheduling problem on heterogeneous systems [J]. Applied Soft Computing,2011,11(2):2725 - 2734.

[7] LI M,et al. Min-energy voltage allocation for tree structured tasks[J]. Journal of Combinatorial Optimization, 2006,11(3):305 - 319.

[8] LI M,et al. An efficient algorithm for computing optimal discrete voltage schedules[J]. SIAM Journal on Computing,2005,35(3):658 - 671.

[9] IBM Blue Gene team. Overview of the IBM Blue Gene/P project[J]. IBM Journal of Research and Development, 2008,52(1):199 - 220.

[10] HAN X, et al. Deadline scheduling and power management for speed bounded processors[J]. Theoretical Computer Science,2010,411(40 - 42):3587 - 3600.

[11] BRAUN T D,et al. A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems[J]. Journal of Parallel and Distributed Computing,2001,61 (6):810 - 837.

[12] GONCALVES J F, et al. A hybrid genetic algorithm for the job shop scheduling problem[J]. European Journal of Operational Research,2005,167(1):77 - 95.

[13] WEBER M,et al. Distributed differential evolution with explorative-exploitative population families[J]. Genetic Programming and Evolvable Machines,2009,10 (4):343 - 371.

[14] NESMACHNOW S,CANCELA H,ALBA E. Heterogeneous computing scheduling with evolutionary algorithms [J]. Soft Computing,2011,15(4):685 - 701.

[15] NESMACHNOW S,et al. A parallel micro evolutionary algorithm for heterogeneous computing and grid scheduling [J]. Applied Soft Computing,2012,12(2):626 - 639.

[16] DE FALCO I,et al. Biological invasion-inspired migration in distributed evolutionary algorithms [J]. Information Sciences,2012,207(10):50 - 65.

[17] 马艳,龚斌,邹立达. 基于平衡定价和成本梯度的科学工作流调度策略[J]. 电子学报,2010,38(10):2416 - 2421.
MA Y,GONG B,ZOU L D. Equilibrium pricing and cost gradient based scheduling strategy of scientific workflow [J]. Acta Electronica Sinica,2010,38(10):2416 - 2421. (in Chinese)

[18] PANT A, et al. Communicating efficiently on cluster based grids with MPICH-VMI[A]. Proceedings of IEEE International Conference on Cluster Computing[C]. San Diego,California,USA:IEEE,2004. 23 - 33.

[19] ALI S,et al. Task execution time modeling for heterogeneous computing systems [A]. Proceedings of the 9th Heterogeneous Computing Workshop [C]. Washington, USA:IEEE,2000. 185 - 199.

作者简介

王静莲　女,1979 年 5 月出生于山东省莱州市. 现为山东大学博士研究生. 主要从事并行与高性能计算、绿色计算和多目标全局优化算法等方向的研究.
E-mail:wjljing@163.com

龚　斌　男,1964 年 10 月生于山东省济南市. 教授、博士生导师. 现为山东大学计算中心主任、山东省高性能计算中心副主任,主要从事网格与高性能计算、机群计算方面的研究工作.
E-mail:gb@sdu.edu.cn