

# 小型通用高性能计算平台的设计与实现

陈红梅<sup>1</sup>, 郭伟<sup>2</sup>, 赖重远<sup>1</sup>

(1. 江汉大学 交叉学科研究院, 湖北 武汉 430056;

2. 江汉大学 数学与计算机科学学院, 湖北 武汉 430056)

**摘要:**随着高性能计算技术和应用的发展,计算模拟已成为科学研究中不可缺少的第三种方法。国内各大高校和科研院所纷纷建立了高性能计算平台,它们大多是针对重点学科大规模数据处理的特点,建立了比较有针对性的高性能计算平台。江汉大学作为一所综合性普通高等学校,其大数据处理的规模相对有限、类型多样,且时间具有阶段性。依据自身大数据处理的特点,本着开放共享、全体受益的原则,江汉大学设计并实现了小型通用高性能计算平台的建设。该平台创建了一个通用的计算环境,同时又考虑了不同用户和应用的特殊需求。平台提供多核CPU并行计算、GPU并行计算和SMP并行计算。平台建立后为各院系提供了高性能计算服务,支持了学校多个科研项目,大力支援了学校科研工作的发展。

**关键词:**大数据处理;高性能计算平台;多核CPU;GPU;SMP

**中图分类号:**TP399

**文献标识码:**A

**文章编号:**1673-629X(2019)10-0186-05

**doi:**10.3969/j.issn.1673-629X.2019.10.036

## Design and Implementation of Small Universal High Performance Computing Platform

CHEN Hong-mei<sup>1</sup>, GUO Wei<sup>2</sup>, LAI Zhong-yuan<sup>1</sup>

(1. Institute of Interdisciplinary Research, Jiangnan University, Wuhan 430056, China;

2. School of Mathematics and Computer Science, Jiangnan University, Wuhan 430056, China)

**Abstract:** With the development of high performance computing technology and application, computational simulation has become an indispensable third method in scientific research. Domestic universities and research institutes have established high performance computing platform, most of which are aimed at big data processing of key disciplines and have established a relatively high performance computing platform. As a comprehensive university, the big data processing in Jiangnan University is relatively limited, various types and its time is intermittent. According to the characteristics of its own big data processing, in line with the principle of open sharing and benefit for all, Jiangnan University has designed and implemented a small universal high performance computing platform. This platform creates a universal computing environment while taking into account the special needs of different users and applications. The platform provides multi-core CPU parallel computing, GPU parallel computing and SMP parallel computing. After the establishment of the platform, it provides high performance computing services for each department, supports many scientific research projects of the university, and strongly supports the development of scientific research work of the university.

**Key words:** big data processing; high performance computing platform; multi-core CPU; GPU; SMP

## 0 引言

长期以来,理论推导和科学实验是人类进行科学研究的两大方法,但随着高性能计算技术和应用的蓬勃发展,计算模拟已成为科学研究中不可缺少的第三种方法<sup>[1]</sup>。高性能计算已经在基础科学研究、工业设计等各个领域广泛应用,解决了一些重大科学和工程

问题<sup>[2]</sup>。高性能计算是前沿性的高技术,是各国争夺的战略制高点,是国家创新体系的重要组成部分。

高校的科学研究急需高性能计算技术的支持。综合性普通高等学校的大数据处理具有以下特点:

(1) 数据处理的规模相对有限;

(2) 数据处理的类型多样,有的需要复杂计算步

骤和复杂数据依赖,有的需要大量重复的数据集运算以及密集的内存存取,有的需要共享存储,等等;

(3)数据处理的时间具有阶段性,综合性普通高等院校的数据处理需求一般跟随项目情况变化。

## 1 高性能计算机发展现状

随着个人计算机和局域网技术的快速发展, Linux 操作系统的日趋稳定,以及基于消息传递的并行程序设计标准的发布,诞生了集群系统。

集群系统是使用高速通信网络将多台原本独立的微机或工作站连接在一起,构成一个统一的整体,使之可作为一种单一的资源来使用。以提高科学计算能力为目的的集群系统,称为高性能计算集群<sup>[3]</sup>,也称为集群式高性能计算机。

中国已成为继美国、日本之后的第三个具备高性能计算机研制能力的国家,成为了世界高性能计算机市场的“第三股力量”<sup>[4]</sup>。2013 年 6 月,由国防科技大学研制的“天河二号”超级计算机位列世界 TOP500 第一名,并连续 6 次蝉联冠军。2016 年 6 月,由国家并行计算机工程技术研究中心研制的“神威·太湖之光”取得世界 TOP500 冠军。截至 2017 年 11 月,“神威·太湖之光”和“天河二号”连续四次分列世界 TOP500 的冠亚军<sup>[5]</sup>。

基于高性能计算机的精确化研究和模拟分析手段,建设高性能计算平台对相关学科追踪国际科技前沿,提升科学研究水平,促进学科交叉和催生新兴学科具有积极意义。

鉴于此,国内各大高校和科研院所很多已经建立了自己的高性能计算平台。清华大学 2005 年底就建立了自己的高性能计算平台<sup>[6]</sup>。中国科技大学于 2003 年 10 月初步建成了中国科大-中国惠普高性能运算联合实验室,后经不断的发展扩建,于 2013 年建成了中国科技大学超级计算中心。上海交通大学于 2013 年建成了高性能计算中心<sup>[7]</sup>。武汉大学于 2015 年建立了超算中心<sup>[8]</sup>。北京大学也于 2016 年建成了高性能计算平台。建立了高性能计算平台的高校大多是依托校内的一两个重点学科,针对重点学科大规模数据处理的特点,建立了比较有针对性的高性能计算平台。

江汉大学作为一所综合性普通高等学校,校内各理工学科对于高性能计算也有着迫切需求。结合普通高校大规模数据处理的特点,本着开放共享、全体受益的原则,江汉大学于 2014 年建立了小型通用高性能计算平台。该平台为全校师生提供高性能计算服务,为学校的科研提供基础性平台支持。

## 2 小型通用高性能计算平台研究

### 2.1 几种类型的高性能计算对计算环境的需求

对于具有复杂计算步骤和复杂数据依赖的计算任务,如分布式计算、人工智能、物理模拟,以及其他通用应用程序等,适合使用多核 CPU (central processing unit) 进行并行计算。对于需要大量重复的数据集运算以及密集的内存存取的计算任务,如视频编码解码、矩阵运算、医疗应用、生命科学等研究应用,适合使用 GPU (graphics processing unit) 进行并行计算。对于共享存储的计算任务,如数据库、在线事务处理系统和数据仓库等,适合使用 SMP (symmetrical multi-processing) 进行并行计算。这是由 CPU、GPU 和 SMP 各自的设计目的和内部的结构差异所决定的。

(1)CPU 被设计成为一个“通才”,它要兼顾指令和数值的并行运算,大部分的晶体管用在了高速缓存和控制电路上,控制电路内部仅有少量的算术逻辑单元 (ALU),更多的是用于加速分支判断甚至更复杂的逻辑判断的硬件。CPU 的设计目的是指令执行的高效率,实现程序执行时的指令相关性和数据相关性等复杂逻辑<sup>[9]</sup>。CPU 擅长处理拥有复杂指令调度、循环、分支、逻辑判断以及执行等的程序任务。

(2)GPU 则被设计成为一个专注计算的“专才”,它内部的大部分晶体管被用来进行数据处理,只有少量的被用做数据缓存和指令流控制。GPU 的设计目的是面向矩阵类型的数值计算,它的众核架构(比如 NVIDIA Tesla K20M 有 2 496 个 CUDA 核心)<sup>[10-11]</sup>非常适合把同样的指令流并行发送到众核上,采用不同的输入数据执行。GPU 的优势是进行无逻辑关系数据的并行计算。

(3)SMP 体系结构<sup>[12]</sup>的特点是基于共享存储,具有多级高速缓存,通过高速监听总线实现处理器与共享存储器之间的连接。SMP 最重要的特性是系统是对称的,每个处理器可等同地访问共享存储器、I/O 设备和操作系统服务。正是对称,才能开拓较高的并行度。

综上所述,GPU 在并行计算上的优势无可厚非,但是 GPU 计算上的突出优势也仅仅体现在浮点运算上,在整数运算、逻辑运算和控制运算上,相较于 CPU 劣势十分明显。依据综合性普通高等学校大数据处理的特点,高性能计算平台合理的构建方式应该是主要配备 CPU 计算节点,然后配备少量的 GPU 计算节点和 SMP 计算节点。

### 2.2 江汉大学通用高性能计算平台的设计与实现

依据综合性普通高等学校大数据处理的特点,江汉大学深入开展调研和前期论证研讨,遵照按需建设、适度超前的原则,创建了一个小型通用高性能计算平

台。此平台创建了一个通用的计算环境,同时又考虑了不同用户和应用的特殊需求。江汉大学小型通用高

性能计算平台的拓扑如图 1 所示。

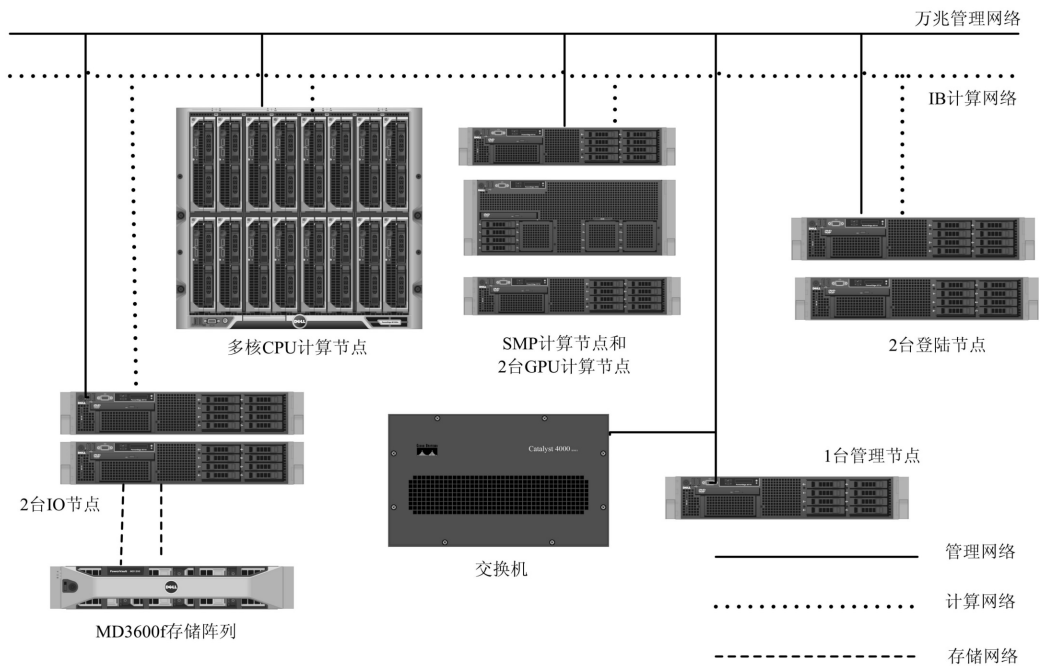


图 1 江汉大学小型通用高性能计算平台拓扑

(1) 计算节点。

江汉大学小型通用高性能计算平台拥有三种类型的计算节点:多核 CPU 节点、GPU 节点和 SMP 节点。

多核 CPU 节点:刀箱 1 台,含 16 个刀片服务器 DELL M620,每个刀片服务器含 2 个 CPU(E5-2650V2,8 核 2.6 GHz 20 M 缓存),64 GB 内存(8 \* 8 GB RDIMM,1 333 MHz)、300 GB 15 K 硬盘。CPU 节点的峰值计算能力是每秒 5.3 万亿次浮点运算。

GPU 节点 2 个:DELL R720 服务器,每台含 2 个 CPU(E5-2650V2,8 核 2.6 GHz 20 M 缓存),64 GB 内存(8 \* 8 GB RDIMM,1 333 MHz)、2 个 300 GB 15 K SAS 2.5 寸硬盘。每台 GPU 服务器配备 2 个 NVIDIA Tesla K20M GPU 卡。GPU 节点的双精度浮点性能是 1.17 万亿次每秒,单精度浮点运算能力是 3.52 万亿次每秒。

SMP 节点 1 个:DELL R910 服务器,4 个 CPU(E7-4820,8 核 2.0 GHz 18 M 缓存),1 T 内存(64 \* 16 GB 1 333 MHz)、2 个 300 GB 15 K SAS 2.5 寸硬盘。

(2) 登录、管理和 I/O 节点。

登录服务器 2 台:DELL R720,每个登录服务器含 2 个 CPU(E5-2640V2,8 核 2.0 GHz,20 M 缓存),64 G 内存(8 \* 8 GB RDIMM,1 333 MHz)、2 个 300 G 15 K SAS 2.5 寸硬盘。

管理服务器 1 台:DELL R720,含 2 个 CPU(E5-2620V2,6 核 2.0 GHz,15 M 缓存),64 G 内存(8 \* 8 GB RDIMM,1 333 MHz)、2 个 300 G 15 K SAS 2.5 寸

硬盘。

I/O 服务器 2 台:DELL R720,每个 I/O 服务器含 2 个 CPU(E5-2640V2,8 核 2.0 GHz,20 M 缓存),64 G 内存(8 \* 8 GB RDIMM,1 333 MHz),2 个 300 G 15 K SAS 2.5 寸硬盘。

(3) 存储系统。

主存储是 MD3600f,含 12 个 600 GB 3.5" 15 K RPM,6 Gbps SAS 硬盘;配带 2 个存储盘柜 MD1200,每个盘柜含 12 个 600 GB 3.5" 15 K RPM,6Gbps SAS 硬盘。共 21.6 T。存储系统做完 RIAD5 之后,容量大概是 17 T。

备份存储是 DELL R720XD,含 12 个 2 T 的存储硬盘,配备一个存储盘柜,内含 8 个 2 T 的存储硬盘。

(4) 通信网络。

整个平台使用 CISCO WS-C4506 万兆以太网交换机,通过网络实现所有节点的互联。为满足高性能计算对数据传输的性能要求,计算网络使用 56 Gbps 速率的 Infiniband 网络<sup>[13-14]</sup>实现计算节点、登陆节点和 I/O 节点间的高速连接。

(5) 作业调度系统。

平台的操作系统为 Redhat Linux Server 6.4,应用开发环境软件为 Intel Cluster Studio 软件工具包。作业调度系统为 IBM Platform LSF,平台所有用户通过 Platform 作业调度系统<sup>[15]</sup>提交作业,所有作业统一排队等待系统分配资源运行。目前平台共提供三种队列:第一种是用于多核 CPU 计算的 normal 队列;第二

种是用于共享存储计算的 bigmem 队列;第三种是用于 GPU 计算的 owerns 队列。

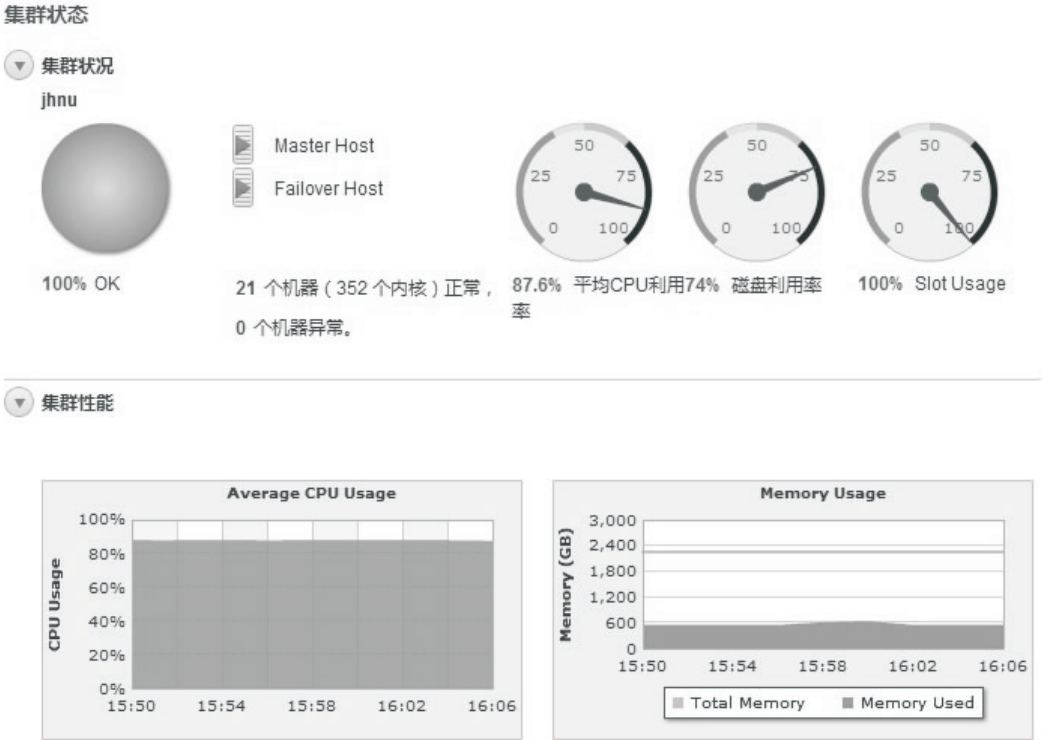
3 江汉大学通用计算平台对科研的支持

小型通用高性能计算平台建立后,通过为校内科研工作者提供高质量的计算服务,对支援学校科研发展起到了积极的作用。从 2014 年 9 月建立,至 2018 年 10 月,该计算平台为交叉学科研究院、数学与计算

机科学学院、护理与医学技术学院、医学院、物理与信息工程学院、工程训练中心等院系提供了高性能计算服务,专业及研究方向涵盖凝聚态物理、计算机应用、模式识别、软物质物理、表观遗传学、有机高分子材料、医学等。使用高峰时多核 CPU 计算和共享存储计算队列需要排队等待,图形计算的 GPU 队列使用较少。平台使用情况如图 2 所示。

<input type="checkbox"/>	作业ID	作业类型	作业名	状态
<input checked="" type="checkbox"/>	184322	作业	JobName	等待
<input type="checkbox"/>	184323	作业	JobName	等待
<input type="checkbox"/>	184324	作业	JobName	等待
<input type="checkbox"/>	184325	作业	JobName	等待
<input type="checkbox"/>	184326	作业	STM	等待
<input type="checkbox"/>	184327	作业	STM	等待
<input type="checkbox"/>	184251	作业	STM	运行
<input type="checkbox"/>	184276	作业	STM	运行
<input type="checkbox"/>	184277	作业	STM	运行
<input type="checkbox"/>	184308	作业	WeMC7_RE	运行
<input type="checkbox"/>	184313	作业	JobName	运行
<input type="checkbox"/>	184317	作业	JobName	运行
<input type="checkbox"/>	184318	作业	JobName	运行
<input type="checkbox"/>	184319	作业	JobName	运行
<input type="checkbox"/>	184320	作业	JobName	运行
<input type="checkbox"/>	184321	作业	.JobName	运行

(a)通用计算平台作业列表



(b)通用计算平台资源使用情况

图 2 平台使用情况

江汉大学小型通用高性能计算平台建立之后的四年时间里,共支持了七项国家级项目,分别是“聚合物



材料的光电性质及杂原子微观动力学机制”、“基于聚苯胺金属纳米线复合透明对电极的双面进光燃料敏华太阳能电池研究”、“亚分子级分辨原子力显微镜中分子修饰针尖的第一性原理研究”、“步行分子马达在金属表面的定向扩散机制的理论研究”、“二维平面上混合自组装单层有序微相的结构调控和性质研究”、“物体形状部分视觉显著性度量及其应用”、“基于分子内电荷转移的 BODIPY 类荧光衍生物的机制研究”。支持了四项省部级项目,分别是“基于随机优化的公共服务设施选址问题研究”、“突发事件下城市交通拥堵传播规律及控制策略研究”、“太阳能光伏发电智能化监测及信息管理系统”、“基于第二代高通量测序技术研究干旱胁迫条件下拟南芥去乙酰化酶 HDA9 负调控和 HDA6 正调控作用机制”。支持了一项校级项目“混合溶液中核酸单链的柔性”。

#### 4 结束语

江汉大学小型通用高性能计算平台的建立,缓解了江汉大学科研工作者计算资源短缺的问题,为他们提供了一个稳定、可靠的计算环境。高性能计算平台的良好运行,提高了江汉大学的科技成果产出效率,使得江汉大学在计算方面的科学研究得到了进一步的发展。

#### 参考文献:

- [1] 陈志明. 科学计算:科技创新的第三种方法[J]. 中国科学院院刊,2012,27(2):161-166.
- [2] 臧大伟,曹政,孙凝晖. 高性能计算的发展[J]. 科技导报,2016,34(14):22-28.
- [3] 万晓姣. 基于 linux 系统集群的架构与实现[J]. 电子世界,2012(10):94-95.

(上接第 169 页)

- [4] [C]//Proceedings of 2014 Asia-Pacific conference on computer science and applications. Boca Raton, USA: CRC Press,2015:187-190.
- [5] CARDELLINI V, GRASSI V, PRESTI F L, et al. Distributed QoS-aware scheduling in storm [C]//ACM international conference on distributed event-based systems. Oslo, Norway: ACM,2015:344-347.
- [6] ANIELLO L, BALDONI R, QUERZONI L. Adaptive online scheduling in storm [C]//ACM international conference on distributed event-based systems. Arlington, Texas, USA: ACM,2013:207-218.
- [7] 熊安萍,王贤稳,邹洋. 基于 Storm 拓扑结构热边的调度算法[J]. 计算机工程,2017,43(1):37-42.
- [8] ESKANDARI L, HUANG Zhiyi, EYERS D. P-Scheduler: a-daptive hierarchical scheduling in apache storm [C]//Australasian computer science week multiconference. [s. l.]:

- [4] 赵毅,朱鹏,迟学斌,等. 浅析高性能计算应用的需求与发展[J]. 计算机研究与发展,2007,44(10):1640-1646.
  - [5] 郑晓欢,陈明奇,唐川,等. 全球高性能计算发展态势分析[J]. 世界科技研究与发展,2018,40(3):249-260.
  - [6] 林皎,陈玉洁,张武生,等. 高性能计算平台建设的探索与实践[J]. 实验技术与管理,2012,29(5):217-220.
  - [7] 林新华,顾一众. 上海交通大学高性能计算建设的理念与实践[J]. 华东师范大学学报:自然科学版,2015(S1):298-303.
  - [8] 黄建忠,张沪寅,程媛. 开放式高性能计算平台的建设与研究[J]. 计算机教育,2012(22):55-59.
  - [9] PRODROMIDIS A L, STOLFO S J. Cost complexity-based pruning of ensemble classifiers[J]. Knowledge and Information Systems,2001,3(4):449-469.
  - [10] 苏华友. 面向应用的 GPU 并行计算关键技术研究[D]. 长沙:国防科学技术大学,2014.
  - [11] SEILER L, CARMEAN D, SPRANGLE E, et al. Larrabee: a many-core x86 architecture for visual computing[J]. ACM Transactions on Graphics,2008,27(3):1-15.
  - [12] RABENSEIFNER R, HAGER G, JOST G. Hybrid MPI/OpenMP parallel programming on clusters of multi-core SMP nodes [C]//17th Euromicro international conference on parallel, distributed and network-based processing. Weimar, Germany: IEEE,2009:427-436.
  - [13] DESHMUKH V D. InfiniBand: a new era in networking [C]//National conference on innovative paradigms in engineering & technology. [s. l.]: IEEE,2012:15-18.
  - [14] SHANLEY T. InfiniBand network architecture [M]. USA: Addison-Wesley Professional,2002.
  - [15] BHATELE A, KALE L V, KUMER S. Dynamic topology aware load balancing algorithms for molecular dynamics applications [C]//International conference on supercomputing. Yorktown Heights, NY, USA: ACM,2009:110-116.
- 
- ACM,2016:26-31.
  - [9] 鲁亮,于炯,卞琛,等. Storm 环境下基于权重的任务调度算法[J]. 计算机应用,2018,38(3):699-706.
  - [10] YANG Xinshe, DEB S. Engineering optimisation by cuckoo search[J]. International Journal of Mathematical Modelling & Numerical Optimisation,2010,1(4):330-343.
  - [11] 施文章,韩伟,戴睿闻. 模拟退火下布谷鸟算法求解车间作业调度问题[J]. 计算机工程与应用,2017,53(17):249-253.
  - [12] 杨辉华,张晓凤,谢谱模,等. 基于布谷鸟搜索的多处理器任务调度算法[J]. 计算机科学,2015,42(1):86-89.
  - [13] 赵莉. 基于改进布谷鸟搜索算法的云计算资源调度[J]. 南京理工大学学报:自然科学版,2016,40(4):472-476.
  - [14] 孙大为,张广艳,郑纬民. 大数据流式计算:关键技术及系统实例[J]. 软件学报,2014,25(4):839-862.
  - [15] 熊安萍,段杭彪,蔺亚雄. Storm 下基于最佳并行度的贪心调度算法[J]. 计算机应用研究,2019,36(5):1068-1071.