田　程,鲁绍坤. 基于树莓派 2 的微型农业大数据平台的可行性研究[J]. 江苏农业科学,2017,45(10):202 – 204.
doi:10.15889/j.issn.1002 – 1302.2017.10.056

基于树莓派 2 的微型农业大数据平台的可行性研究

田　程, 鲁绍坤
(云南农业大学基础与信息工程学院,云南昆明 650201)

摘要:探讨了一种建立于廉价低功耗的硬件平台上的农业大数据平台的可行性,搭建了 1 个廉价的由树莓派 2 组成的基于 Spark 的微型大数据平台,测试了其性能、功耗。结果表明,基于树莓派 2 的微型农业大数据平台是一种在农业环境下对计算性能要求不高时较为经济的选择。

关键词:Spark;树莓派 2;大数据;农业
中图分类号: S126　　**文献标志码**: A　　**文章编号**:1002 – 1302(2017)10 – 0202 – 03

随着信息时代的发展,各行各业每时每刻都在产生大量的数据,为了应对这些大量数据的存储及处理需求,建立了众多的数据中心,包括基于公有云服务的数据中心和私有的大数据中心。这些大数据中心的建立都耗资巨大,服务器和机房设备通常都价格昂贵,同时运行成本高昂,服务器及冷却设备都需要消耗大量的能源[1]。农业大数据作为大数据技术在农业领域的应用同样面临这一问题,寻求一种低成本、低功耗的设备来组建数据中心,让数据中心更加经济、绿色环保,变得越来越重要[2]。

国外研究者根据这一需求,搭建了一些基于 ARM 的低成本、低功耗的电脑组建的计算集群来研究低成本低功耗平台的计算集群应用的可行性。由于研究者大多使用上一代电脑和 Hadoop MapReduce 架构,研究结果多数因性能不足而不适合使用到生产环境,但人们仍然相信,使用新的 ARM 处理器和新的架构的计算集群可以满足低成本下大数据处理的需要[3]。新一代树莓派 2,性能较第 1 代有很大提高,同时比 Hadoop MapReduce 更快的 Spark 大数据计算框架的出现,让低成本的大数据应用成为了可能。

1　背景

1.1　树莓派 2

树莓派 2(Raspberry Pi 2 Model B)是由英国慈善组织“Raspberry Pi 基金会”开发的新一代基于 ARM 的卡片式电脑(表 1)。在 Sysbench 多核 CPU 测试中,树莓派 CPU 性能是第 1 代树莓派 Raspberry Pi 1 Model B + 的 6 倍,并且保持了和第 1 代树莓派相同的低廉售价($35)[4]。

1.2　Spark

Spark 是 1 个基于内存计算的开源大数据并行计算框架,于 2009 年诞生于加州伯克利分校 AMPLab,目的在于简单快速地处理大数据。目前由 AMPLab、Databricks 负责整个项目的开发维护,众多公司(如 Yahoo、Intel)和众多的开源爱好者都积极参与 Spark 的更新与维护。据 Spark 官方网站公布数据,Spark 运行于内存数据集时性能为 Hadoop MapReduce 的 100 倍,运行于磁盘数据集时也有其 10 倍的性能。Spark 使用 Scala 语言编写,同时提供多种编程接口,可以使用 Java、Python、R 语言编写程序,方便开发者自由选择;兼容 Hadoop 生态系统,能够运行在单机、Hadoop、YARN、Mesos 集群及多种云平台上[5]。

表 1　树莓派配置

部件	配置
CPU	900 MHz 4 核 ARM Cortex – A7
RAM	1 GB LPDDR2 SDRAM
以太网接口	板载 10M/100M RJ45 以太网接口
USB	4 个 USB 2.0 接口
存储	MicroSD 卡

1.3　国外的低成本计算集群

Iridis – pi cluster 由 64 个 Raspberry Pi Model B 组成,每个节点使用 16 GB SD 卡组成的 1 个低功耗、便宜、被动散热的可用于教育目的的集群[6]。基于 ARM 的低成本集群在合理的计算能力下,提供比传统 SATA 串行存储和 PCI 串行总线更大的扩展能力,是在严酷的、维护困难的应用环境和对可靠性要求较高、计算能力不是第一要求的情况下的一种选择。

Glasgow Raspberry Pi cluster 使用 56 个 Raspberry Pi Model B 组成[7]。采用 LXC 作为 Container 搭建了 PiCloud 用于研究和教育目的。

Bolzano Raspberry Pi cluster 使用 300 个 Raspberry Pi Model B 组成,用于研究廉价绿色的云计算和作为移动数据中心在恶劣环境的可靠性[8]。

Kaewkasi 等使用 22 个基于 1 GHz ARM Cortex – A8 CPU 和 1GB RAM 的 Cubieboard 搭建了 1 个基于 Hadoop 架构的 Spark 低功耗计算集群,并测试了其在 SSD 和机械硬盘上的计算性能和功耗[3]。

Schot 使用 8 个树莓派 2 组建了基于 Hadoop 的微型数据中心,并与 University of Twente 的由 32 台 Dell R415 组成的运行于 Hadoop 的 CTIT cluster 做了性能与功耗对比[9]。

收稿日期:2016 – 03 – 10
基金项目:云南省教育厅科学研究基金(编号:2015J064)。
作者简介:田　程(1989—),男,云南曲靖人,硕士研究生,从事农业信息化研究。E – mail:tiancheng0093@126.com。
通信作者:鲁绍坤,博士,副教授,研究方向为农业信息化。E – mail:lsk999@126.com。

2　树莓派 2 大数据平台系统介绍

本研究搭建的微型计算包含 6 个节点，每个节点选用 1 个 Raspberry Pi 2 Model B 组成，使用 16 G UHS－I MicroSD 卡作为系统存储。使用 Spark 替代国外组建低功耗集群常用的 Hadoop 架构中的 MapReduce。系统架构选用 HDFS 作为分布式存储，Spark 作为计算引擎。节点操作系统选用树莓派官方开发的基于 Debian wheezy 的 Raspbian wheezy 于 2015 年 5 月 5 日发行版。Linux 核心版本 3.18.11，JDK 版本 1.8，Hadoop 版本 2.6.2，Spark 版本 1.5.0，运行于 Standalone 模式。Spark1 为主节点，Spark1 ~6 为从节点（图 1）。

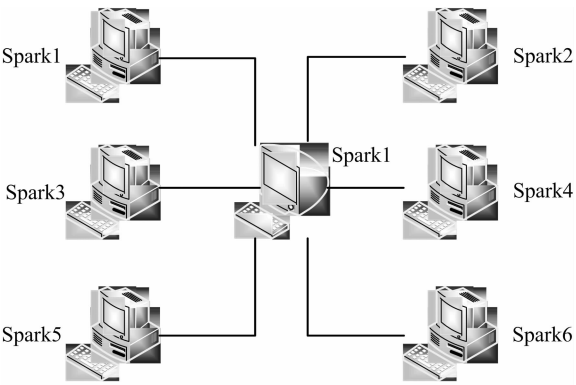

图1　Spark 计算集群

3　树莓派 2 大数据平台性能测试

由于 Spark 还处于快速开发中，目前还没有合适的针对 Spark 1.50 集群的基准测试标准。通过参考中国科学院计算技术研究所提出的大数据测试工具 BigDataBench 2.0，选取了“计数”“排序”和“查找”3 个操作作为 Micro benchmarks 的测试基准[10]。本研究编写了 1 个包含以上 3 种操作的程序来测试树莓派 2 集群的性能。为了对比集群的计算性能，使用云南农业大学基信学院计算机公共实验室的 1 台计算机搭建了伪分布式 Spark 集群进行性能对比。所用 CPU 为双核 4 线程 I3－2130。

3.1　单节点性能测试

使用 Sysbench 作为基准测试工具进行性能测试，分别测试了 1 台树莓派 2 和计算机实验室计算机的单线程 CPU 基准、多线程 CPU 基准、硬盘顺序读取速度、内存读取速度（表 2）。

表 2　Sysbench 测试结果

测试基准	CPU 基准（s）	内存读取速度（MB/s）	硬盘顺序读取速度（MB/s）
树莓派 2	175	2 807	18.445
计算机	19	5 004	124.560

通过 Sysbench 测试可知，计算机实验室的计算机的 CPU 计算能力约为树莓派 2 的 9.2 倍，内存读取速度为 1.78 倍，硬盘顺序读取速度为 6.75 倍。

3.2　数据来源

本研究测试数据采用编写程序生成随机测试数据进行测试。模拟对温度的统计，生成了包含 2 亿条数据的 temperature.txt 文件，文件内容为 3 列，第 1 列为日期，第 2 列为区域，第 3 列为温度（℃）。测试文件大小为 3.06 GB。

3.3　测试方法

使用 Spark 的 count 函数计算 temperature.txt 文件中 A 地的记录数量，然后使用 sortby 函数对 A 地温度数据按温度高低排序，最后使用 first 函数找出 A 地的最高温度。分别测试运行于机房 1 台计算机的伪分布式集群和 6 台树莓派 2 组成集群的计算耗时，并使用 Spark WebUI 查看程序运行时间。

3.4　测试结果

运行测试程序测试 Spark 大数据平台性能结果显示，所用时间与单台机房计算机用时相当，树莓派 2 集群和机房计算机分别用时 517、527 s（图 2）。

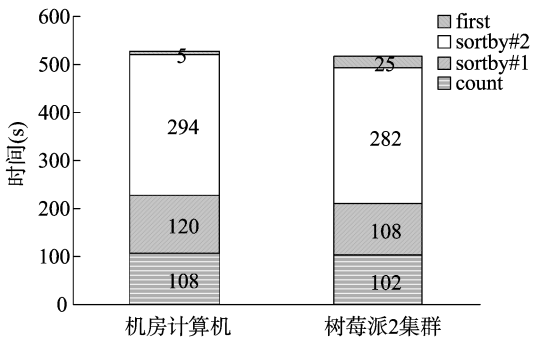

图2　Spark 集群处理用时

通过计算可知，6 台树莓派 2 集群组成的大数据处理平台处理速度为 0.353 GB/min，使用云南农业大学基信学院计算机公共实验室的 1 台计算机搭建的伪分布式 Spark 集群处理速度为 0.348 GB/min。两者运行 Micro benchmarks 时的性能相当。

4　树莓派 2 运行温度和功耗测试

4.1　温度测试

树莓派 2 设计为被动散热工作。为了测试树莓派 2 作为微型计算集群的稳定性，本研究测试了树莓派 2 在室温无风环境下，不使用任何主动散热设备的情况下，进行 100 min 满负载运行时的工作温度变化。通过每分钟读取 1 次 CPU 温度，记录了树莓派 2 工作时的 CPU 温度变化曲线（图 3）。树莓派 2 在满负载被动散热情况下最大温度仅 65.9 ℃，低于树莓派官方指导树莓派正常工作的最高温度（85 ℃）。本研究同时还进行了 12 h 空载待机温度测试，结果显示树莓派 2 空载时平均待机温度为 34.2 ℃。以上测试说明被动散热情况下树莓派 2 可稳定工作。

4.2　功耗测试

由于树莓派 2 使用 5 V 电源、USB 供电，采用睿登 OLED USB2.0 高精度测试仪测试节点满载功耗和待机功耗（表 3）。

通过测试结果可计算出 6 台树莓派 2 组成的集群待机功耗仅 7.32 W，满载功耗仅 15.90 W。而同样数据处理能力的计算机机房单台计算机仅 CPU 满载功耗就达 65 W。

5　树莓派 2 大数据平台成本分析

5.1　搭建成本

树莓派 2 官方统一售价为 35 美元，国内售价为人民币 235 元。6 台树莓派 2 集群成本共计 1869 元（表 4），搭建成

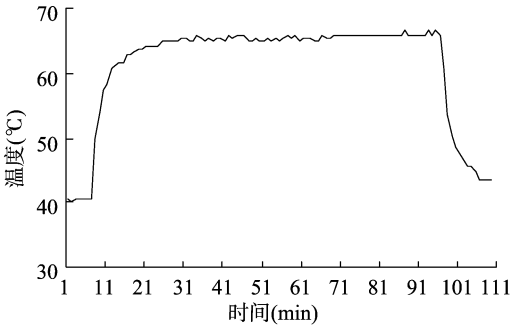

图3　满载温度测试

表 3　树莓派 2 集群单节点耗电

分别	电压（V）	电流（A）	功率（W）
待机	5.3	0.23	1.22
满载	5.3	0.50	2.65

表 4　树莓派 2 集群搭建成本

部件	单价（元）	数量（个）	总计（元）
树莓派 2	235	6	1 410
16 G SD 卡	50	6	300
6 口 USB 电源	120	1	120
microUSB 线	6.5	6	39

本远低于单台计算机成本。

5.2　运行成本

通过计算 6 台树莓派 2 集群满载 10.59 W 运行和单台计算机以 65 W 各运行 3 年电量消耗可知树莓派 2 集群共消耗电量 412.128 kW·h，单台计算机耗电量 1 684.8 kW·h，树莓派 2 集群耗电量约为单台台式机的 1/4（图 4）。运行 3 年树莓派相比单台计算机共节约 1 272.672 kW·h 电量。

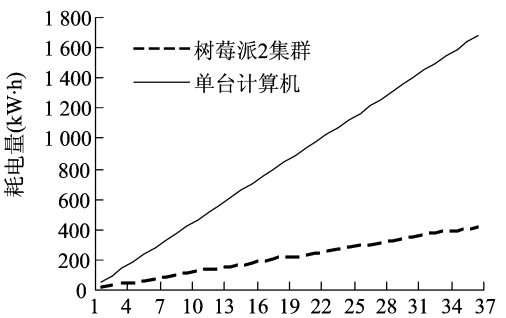

图4　耗电量曲线

6　总结与展望

由 6 台树莓派 2 组成的大数据平台具有和 1 台双核 4 线程 CPU 组成的台式机相当的数据处理速度。其低廉的节点价格，无需主动散热设备的投入，使得集群初装快速简单，成本低廉。其极低的耗电量，使得集群后期运行成本远低于台式机，同时节约大量电能。树莓派 2 适用于在农业大数据应用中不严格要求计算性能的情况下组建具有一定运算能力的低功耗、廉价的绿色计算集群，是一种降低农业大数据数据处理成本的可行方案。

随着性能更强的低功耗 ARM CPU 的不断研发，更新更快的大数据处理架构的出现以及更多针对低功耗 ARM 平台计算集群的优化方案，组建廉价绿色的计算集群将更加可行，将为农业大数据的普及和发展提供有力的支持。

参考文献：

[1]邓　维，刘方明，金　海，等. 云计算数据中心的新能源应用：研究现状与趋势[J]. 计算机学报，2013，36(3)：582－598.

[2]王文生，郭雷风. 农业大数据及其应用展望[J]. 江苏农业科学，2015，43(9)：1－5.

[3]Kaewkasi C，Srisuruk W. A study of big data processing constraints on a low－power Hadoop cluster[C]//International Computer Science and Engineering Conference，2014.

[4]UPTON Upton E. Raspberry Pi 2 on sale now at $35[EB/OL]. [2016－02－10] https://www.raspberrypi.org/blog/raspberry－pi－2－on－sale/.

[5]Apache. Spark：lightning－fast cluster computing[EB/OL]. [2016－03－06]. http://spark.apache.org.

[6]Cox S J，Cox J T，Boardman R P，et al. Iridis－pi：a low－cost，compact demonstration cluster[C]. The 2013 IEEE 33rd International Conference on distributed computing systems workshops，2013.

[8]Abrahamsson P，Helmer S，Phaphoom N，et al. Affordable and energy－efficient cloud computing clusters[C]//The 2013 IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom)，2013.

[9]Schot N. Feasibility of raspberry Pi 2 based micro data centers in big data applications[EB/OL]. [2016－03－06]. http://referaat.cs.utwente.nl/conference/23/paper/7509/feasibility－of－raspberry－pi－2－based－micro－data－centers－in－big－data－applications.pdf.

[10]Wang L，Zhan J，Luo C，et al. Big data bench：a big data benchmark suite from Internet services[C]. The 2014 IEEE 20th International Symposium on high performance computer architecture (HPCA)，2014.