高校科研团队 Microwulf 并行计算系统的搭建

刘新荣[1,2],胡元鑫[1],罗建华[3],葛 华[4]
(1. 重庆大学 土木工程学院,重庆 400044;2. 后勤工程学院 军事建筑工程系,重庆 401311;
3. 重庆南江水文地质工程地质队,重庆 401121;4. 成都地质调查中心,四川 成都 610081)

摘要:由于成本限制,高校科研团队的计算机多为普通型个人电脑(PC),无法执行涉及多节点的并行计算任务。基于 Beowulf 架构与可于大众电脑市场采购到的普通硬件设备搭建了一套四节点 8CPU 核 Microwulf 并行运算系统(命名为 LXR01)。该系统各节点均安装 Windows7 与 Linux 操作系统,其中服务器节点利用 NFS 建立共享文件系统,分别以 gcc、OpenMPI 为主要编译环境和并行计算环境。HPL 测试表明,LXR01 系统的最大运算能力为 39.93Gflops,其计算效率和成本效率分别为 81.6% 和 ¥583.8/Gflops。同时,SPECFEM3D 代码的计算实例证明 LXR01 系统能有效地解决并行计算问题。

关键词:并行计算;Microwulf;Linux;OpenMPI
中图分类号:G647 **文献标志码**:A **文章编号**:1005-2909(2011)02-0137-05

Microwulf 系统为低成本、高性能的个人型 Beowulf 系统的简称[1]。Beowulf 是一种用作平行计算的自有内存电脑集群架构,源于 NASA 的 Beowulf 高性能计算工程,通常由一台服务器和多台用户端并通过局域网进行信息传递的系统[2]。无论是国外还是国内,均有 Beowulf 系统的应用实例[3-8]。Microwulf 系统的硬件设备并非由特殊销售商提供,可通过大众电脑市场采购,这能有效降低成本;基于快速的电脑硬件发展速度及不断降低的硬件价格,该系统具有高度灵活的可扩展性和可升级性。

高校科研团队的计算机设备一般多为个人电脑(PC),除了普通的文档管理、撰写及一些基于单机的数值计算外,无法执行涉及多节点的并行计算任务,如流体力学、量子化学、大型数据库及各类电、磁、声、波场的计算等。部署商用高性能计算服务器或利用各高性能计算中心的大型机或巨型机能有效进行并行计算,但均需要高额成本。将科研工作室现有计算机资源组建为 Microwulf 系统是解决并行计算的最为经济高效的方案。本文利用电脑市场能购买到的普通硬件设备搭建了一套 Microwulf 系统(命名为 LXR01),并展示了用于该系统的基础软件和基础配置,并在 LXR01 系统上运行 SPECFEM3D 代码进行该系统的并行计算有效性验证。

收稿日期:2011-03-02
基金项目:国家自然科学基金创新群体基金(50621403);中国地质调查局项目"地震滑坡灾害编图方法示范研究"(1212010914011)
作者简介:刘新荣(1969-),男,重庆大学土木工程学院教授,博导,主要从事岩土工程及隧道工程研究,(E-mail) huanduh@gmail.com。

一、系统架构与搭建

(一)系统规划与硬件设备

根据科研工作室计算机既满足个人使用又能进行并行计算的要求,LXR01 系统被设计为一台服务器和数台 PC 用户端模式,而非传统的服务器加无盘工作站模式。这样整套系统可进行 Microwulf 模式与单 PC 的灵活转换。因此,LXR01 系统中服务器与用户端均为双系统平台,即同时安装 Windows 与 Linux 操作系统。文中的操作均属 Linux 下的操作。图 1 描述了 LXR01 系统的基本架构。

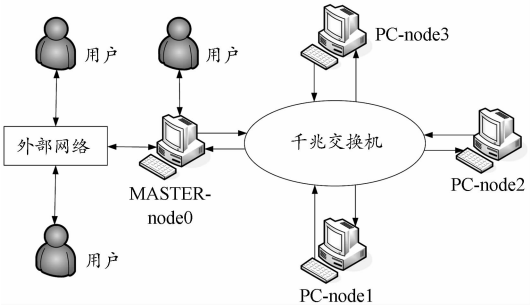

图 1 LXR01 系统的基本架构

由图 1 可知,该系统由一台服务器和三台计算节点组成,分别命名为 MASTER - node0、PC - node1 ~ 3(以下分别简写为 node0 ~ node3),其中 node0 也身兼计算节点的功能。用户可直接通过 node0 或通过外部网络登录到 node0 进行管理或计算。每节点采用 Dell Inspiron 灵越 580s 型台式机,其核心配置为 intel i3 - 540 双核 3.06GHz 处理器、4GB 内存与 500GB 硬盘。由于台式机主板上已有一块板载千兆网卡,为了保证每 CPU 核具有单独的数据传输通道,给 node1 ~ 3 分别添加了第二块千兆网卡;而给 node0 另行添加了两块千兆网卡,其中之一块与原板载网卡组成 CPU 核的传输通道,另一块用于与外部网络的联接。各节点之间的数据传输和交换通过千兆交换机进行。为了使 node0 有足够的磁盘空间保存数据和保证数据安全性,在 node0 添加了第二块 1T 硬盘。

(二)网络拓扑

通过在每节点添加额外网卡,系统内每 CPU 核均有单独的千兆网卡与其一一对应,其拓扑结构如图 2 所示。

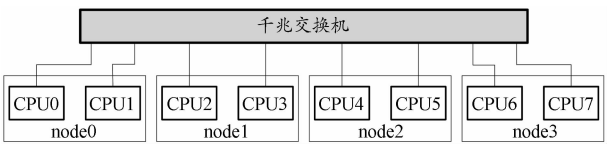

图 2 网络拓扑

LXR01 系统具有 4 节点 8CPU 核。从网络的角度,可将每 CPU 核视为一节点,因此可将板载网卡的 IP 地址设为私有静态地址 192.168.1.0/12,同理可将另行添加的 PCI 网卡的 IP 地址设为 192.168.2.0/12。表 1 列举了系统网卡设置,每节点每块网卡的子网掩码均为 255.255.255.0。

表 1 网卡 IP 地址设置

节点	板载网卡		PCI 网卡	
	IP 地址	代号	IP 地址	代号
node0	192.168.1.1	node0/0	192.168.2.1	node0/1
node1	192.168.1.2	node1/0	192.168.2.2	node1/1
node2	192.168.1.3	node2/0	192.168.2.3	node2/1
node3	192.168.1.4	node3/0	192.168.2.4	node3/1

为了通过主机名访问而非 IP 地址访问,需修改每一节点/etc/hosts 文件。其中 node0 的 hosts 文件如下:

```
127.0.0.1       node0/0
127.0.0.1       node0/1
192.168.1.2     node1/0
192.168.2.2     node1/1
192.168.1.3     node2/0
192.168.2.3     node2/1
192.168.1.4     node3/0
192.168.2.4     node3/1
```

对于其他节点,也应作类似修改,但要参照 node0 的/etc/hosts 文件与表 1 将该节点 IP 地址以 127.0.0.1 替换即可。

(三)软件环境及配置

LXR01 系统的软件环境除 64 位 Ubuntu10.04 操作系统外,还包括组建共享文件系统的 NFS 软件包、用于节点间访问和数据交换的 OpenSSH 软件包、gcc 编译环境、用于构建并行计算环境的 OpenMPI 或 MPICH2 软件包及集群作业管理系统 Torque,其中 gcc 的自定义编译安装可获得常用语言的编译环境。上述软件均为开源软件,可通过因特网免费获取和使用。

在安装 Linux 中,node0 除必要的分区外,另将第二块硬盘全部空间建立/workdir 分区,并将该分区共享给系统内每一节点。因此 node0 需安装 NFS 组件并启动其服务。安装 NFS 后,编辑/etc/exports 文件,在该文件中添加下列代码:

```
/workdir 192.168.1.0/12(rw,no_root_squash,sync,
no_subtree_check)
```

即将/workdir 分区共享给 192.168.1.0/12 网段内的任何节点。重启 node0 的 NFS 服务(/etc/init.d/nfs restart),并编辑 node1 ~3 的/etc/fstab 文件,添加下列代码以便 node1 ~3 在启动时自动挂载 node0 的/workdir 分区。这样在 LXR01 系统内获得了相同的/workdir 分区,满足了并行计算应当处于相同文件目录的要求。

192.168.1.1:/workdir /workdir nfs defaults 0 0

除共享分区外,每一节点均需安装 SSH 服务及客户端用于节点间的互访与信息传递。值得注意的是,节点间的 SSH 访问需配置为无密码访问以利于并行计算的顺利进行,但 node0 同外部网络之间的 SSH 登录却应配置为密码访问,阻止无关人员登录到 LXR01 系统进行操作。

并行计算环境的创建可应用 OpenMPI 或 MPICH2,安装时可采取编译安装,这样可获取多语言或多编译器下的并行计算环境,以便有不同需求的用户使用。本文建议使用 OpenMPI 软件包以避免 MPICH2 软件包创建并行计算环境时需先启动 mpd 环的方式。

作业管理系统 Torque 主要用于保证节点荷载均衡和计算作业的有序管理。该软件包的安装可参考安装手册,但应注意以下三点:(1)将 make packages 命令生成的 sh 脚本文件拷贝至每一节点并加 --install参数运行。(2)自行建立 node0 节点/var/spool/torque/server_priv/nodes 文件,在该文件的每一行加入系统内每一节点名与该节点 CPU 核数量(如"node0 np = 2")。(3)启动 pbs_server、pbs_sched 及 pbs_mom 并配置为开机自启动。

二、系统运算能力测试

计算机的运算能力用每秒执行的双精度浮点运算次数(flops)来衡量,而集群的理论峰值运算能力可通过下式确定:

$$R_{Peak} = N_{nodes} \cdot N_{cores/node} \cdot N_{FPU/core} \cdot Clockspeed$$

式中 R_{Peak} 为集群的理论峰值运算能力(Gflops),N_{nodes} 为节点数,$N_{cores/node}$ 为每节点的 CPU 核数,$N_{FPU/core}$ 为每 CPU 核的浮点运算单元数,$Clockspeed$ 为 CPU 的时钟速率(GHz),则本文所搭建的 LXR01 系统理论峰值运算能力为 $4 \times 2 \times 2 \times 3.06 = 48.96$Gflops。

但系统的理论峰值运算能力并不代表其真实的最大运算能力。通常集群运算能力利用 HPL 进行测试[9-10]。因此,本文在 LXR01 系统上安装 HPL、BLAS 数学库与 GOTOBLAS 软件包对其真实运算能力进行测试。测试参数如下:$P \times Q = 2 \times 4$, NB = {100, 120, 140, 160, 180, 200},N 从 1 000 开始递增,并采取双精度浮点计算。测试结果表明,当 NB = 160,N = 45 000 时,系统的真实最大运算能力为 $R_{Max} = 39.93$Gflops(图 3),则本文所搭建的 LXR01 系统的计算效率 $\eta = R_{Max}/R_{Peak} = 39.93/48.96 \approx 0.816 = 81.6\%$。通常对于集群而言,超过峰值运算能力的60%即可认为具有高计算效率。因此 LXR01 具有杰出的计算效率。

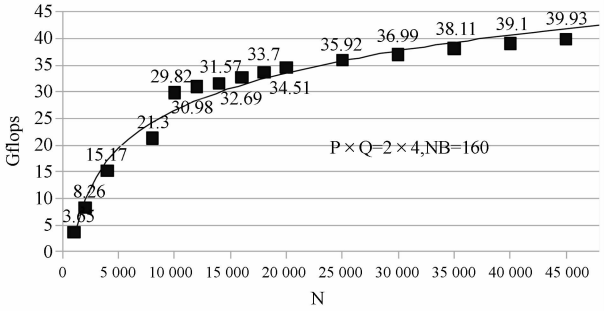

图 3　$P \times Q = 2 \times 4$, NB = 160 时的运算能力测试结果

三、SPECFEM3D 实例应用

SPECFEM3D 代码基于谱元法采用 Fortran90 语言开发,主要用于模拟地震波的传播及震源的反演[11],其核心运算为求解三维波动的弹性动力方程:

$$\rho(x) \cdot \frac{\partial^2}{\partial t^2} u(x,t) - \nabla \cdot \sigma(x,t) = f(x,t)$$

式中 $\rho(x)$ 为质量,$u(x,t)$ 为位移场,$\sigma(x,t)$ 应力场,∇ 表示 $\frac{\partial}{\partial x}$,$f(x,t)$ 系震源项。为了验证本文搭建的 Microwulf 系统有效性,本文建立的计算模型包含 $96 \times 96 \times 15$ 个谱单元,由于 SPECFEM3D 代码在每个谱单元内插 5^3 个 Gauss - Lobatto - Legendre 点,该模型网格共含有 138 240 个谱单元,9 088 756 个网格点,其总自由度为27 266 268(图 4),网格数据库文件大小为 4.1GB。同时 SPECFEM3D 代码需利用高阶 Lagrange 多项式对各谱单元进行内插及建立质量矩阵、刚度矩阵、震源项计算、全局聚合计算,计算过程中需进行大量的数据存储和交换。

将图 4 所示模型提交至搭建的 LXR01 系统,并分配四个节点 8CPU 核参与计算,所需内存约 14G,其计算时间为 12.3 小时;如分配两个节点 4CPU 核进行计算,则计算持续 26.8 小时。

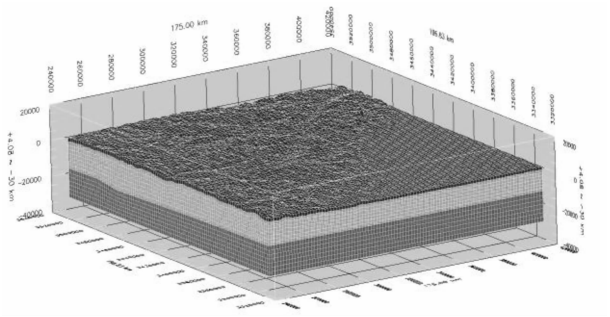

图 4 模型计算网格

为了进行比较,本文进行了另外两种方案的计算:其一采用一台安装有 4GB 内存的四核电脑对相同模型进行 4 进程并行计算(A 方案);其二将 SPECFEM3D 代码编译为非并行计算模式,并采用一台 2GB 内存电脑进行计算(B 方案)。其中 B 方案能生成网格,但在生成网格数据库时出现内存截断错误,无法正常计算。在计算过程中,无论是 LXR01 系统各节点或 A 方案计算的单机,均应用 top 命令监测系统(节点)负荷和 CPU 状态。监测结果表明,A 方案系统负荷远大于 LXR01 系统各节点,CPU 用于处理系统 I/O 的时间(%wa)也远大于其程序运行时间(%us),4 计算进程同时处于 R 状态(运行状态)的时间较短,多数为 1 计算进程呈 R 状态,其余 3 计算进程处于 D 状态(不可中断休眠状态),计算完成共耗时 304.8 小时。LXR01 系统各节点能保证各计算进程均同时处于 R 状态且绝大部分 CPU 时间用于程序运行而非处理系统 I/O。这说明单机 4 进程计算由于任务集中,系统需处理大量的数据交换而增加了计算耗时。

上述应用于不同平台的计算方案说明 LXR01 系统能有效处理具有较大运算量的并行计算任务。

四、讨论

(一)成本效率

了解 LXR01 系统的最大运算能力后,可利用该系统所有硬件的经济成本进行系统的成本效率分析。搭建该系统的所有硬件设备及其成本如表 2 所示。

利用表 2 所列硬件设备获得了 39.93Gflops 的最大计算能力,则本文所搭建的 LXR01 系统的成本效率为 ¥23 311/39.93Gflops ≈ ¥583.8/Gflops。该数据远低于近年来国外一些同样以 HPL 测评的 Beowulf 系统成本效率,如英国 Kentucky 大学于 2003 年搭建的 Beowulf 集群 KASY0 的成本效率为 $210/Gflop,2004 年 Virginia Tech 的 System X 集群的成本效率为 $465/Gflop,2007 年 SUN 公司的 Sparc Enterprice M9000 集群的成本效率为 $496/Gflop。上述比较说明随着硬件价格的不断下降,Microwulf 系统的成本效率还会进一步提高。

表 2 主要硬件设备及其成本

硬件设备	单价	数量	小计
Dell inspiron 580s 台式机	¥5 299	4	¥21 196
希捷 ST31000528AS 硬盘	¥450	1	¥450
金士顿 2GB DDR3 - 1333 内存	¥270	4	¥1 080
TP - LINK TG - 3269C 网卡	¥45	5	¥225
TP - LINK TL - SG1008 + 交换机	¥360	1	¥360
总计		¥23 311	

(二)系统优劣性

除成本优势外,Microwulf 系统首要优势在于其具有灵活的可升级性和可扩展性。

计算机硬件技术的发展日新月益,而 Microwulf 系统可跟踪硬件技术的发展,将不断出现的新技术或新硬件应用于系统或进行硬件升级以获取更强的运算性能。从这个角度看,没有相同的 Microwulf 系统,尽管它们都运行相同的软件。

科研团队一般都有一些旧电脑无人愿意使用,但弃之可惜。可将这类旧电脑添置 PCI 千兆网卡后连入 Mirowulf 系统作为计算节点,作为该系统运算能力的有效扩展。当系统的节点数发生改变后,需要修改服务器和各节点的/etc/hosts 文件,并配置新添加节点自动挂载服务器节点共享文件系统,同时也需将新添加的节点名称及 CPU 数加入服务器/var/spool/torque/server_priv/nodes 文件,以便集群作业管理 Torque 能识别新节点和管理计算资源。

Microwulf 系统的另一优势体现于柔性的可配置方案。不同的用户对系统有不同的需求,甚至有些需求看起来非常独特。因此,用户可从广泛的硬件销售商那里选择合适的子系统,并可从众多的开源软件里选择合适的软件包进行配置,如安装不同的操作系统(FreeBSD 或不同的 Linux 发行版本)以满足特殊需求、配置 CPU + GPU 计算环境、建立专门的存储节点和并行文件系统等。

Microwulf 系统的上述优势令其不属于任何一家硬件或系统销售商,但这也带来唯一的使用劣势,即系统得不到销售商的技术支持。因此,Microwulf 系统的使用者必须自己提供技术支持和系统维护。同

时,尽管 Microwulf 系统能提供低成本的高性能并行计算,但与专业的计算集群相比仍有差距。

(三)应用并行计算系统的先决条件

虽然 Microwulf 系统能提供并行计算服务,但并非所有程序都能利用该服务,如普通的文档撰写或计算即使提交到并行计算系统,其运行效率或稳定性或许不如单机运行环境。要想利用 Microwulf 的并行计算服务,必须要使程序或代码具有并行运行功能。一些商业计算平台如 Ansys、Matlab 等已具备并行运行能力,但对于崇尚自由与开源的 Microwulf 系统,自编代码或改编代码过程中对代码的并行化处理必不可少。所幸的是,这类并行运行代码的编写或改写能有效提高科研人员的专业素质,对于提高科研团队实力具有促进作用。因此,本文所搭建的 LXOR01 系统也可作为对研究生的教学和培训平台。

五、结语

本文的目的在于提供低成本的高性能并行计算环境。经过以上分析,可得到下列结论。

(1)利用 PC 和千兆网设备可搭建并行计算环境,对所需硬件设备无特殊要求,所需软件均为开源软件,无需额外购买。

(2)本文所搭建的 LXR01 系统既可作为单机使用,也可运行于并行计算模式,当处于并行计算模式下可获取 39.93Gflops 的最大运算能力,其计算效率为 81.6%,SPECFEM3D 代码在 LXR01 系统上的成功执行表明该系统能有效地解决并行计算问题。

(3)LXR01 系统的成本效率为 ¥583.8/Gflops,随着计算机硬件技术的发展,不仅可灵活地升级系统或搭建类似系统,还可利用已有旧电脑进行扩展。

(4)只有对程序或代码进行并行化处理后,才能最大限度地利用 LXR01 系统的并行运算能力。

参考文献:

[1]ADAMS J C, BROM T H. Microwulf: a beowulf cluster for every desk[C]//Proceedings of the 39th SIGCSE technical symposium on Computer science education. Portland, OR, USA: ACM, 2008: 121 - 125.

[2]STERLING T, BECKER D J, SAVARESE D, et al. Beowulf: A Parallel Workstation For Scientific Computation [C]//In Proceedings of the 24th International Conference on Parallel Processing. CRC Press, 1995: 11 - 14.

[3]丁海平, 刘启方, 黄勇, 等. 三维地震动场数值模拟并行计算系统[J]. 地震工程与工程振动, 2004, 24(02): 19 - 22.

[4]刘宾, 刘广钟. 基于 Linux 集群的高性能计算环境[J]. 辽宁工程技术大学学报, 2006, 25(S2): 254 - 256.

[5]SONZOGNI V E, YOMMI A M, NIGRO N M, et al. A parallel finite element program on a Beowulf cluster[J]. Advances in Engineering Software, 2002, 33(7 - 10): 427 - 443.

[6]DMITRUK P, WANG L -, MATTHAEUS W H, et al. Scalable parallel FFT for spectral simulations on a Beowulf cluster[J]. Parallel Computing, 2001, 27(14): 1921 - 1936.

[7]李贵明, 俞国扬, 罗家融. 基于 Linux 的 Beowulf 集群的实现[J]. 计算机工程, 2003, 29(11): 49 - 51.

[8]祝永志, 赵岩, 魏榕晖. 基于 MPICH 的 Beowulf 集群系统构建与性能评测[J]. 计算机工程与应用, 2006(14): 132 - 133.

[9]肖明旺, 许坚, 车永刚, 等. 一个实用高性能 PC 集群的 Linpack 测试与分析[J]. 计算机应用研究, 2004(09): 183 - 184.

[10]王晓英, 都志辉. 基于 HPL 测试的集群系统性能分析与优化[J]. 计算机科学, 2005, 32(11): 231 - 234.

[11]KOMATITSCH D, RITSEMA J, TROMP J. The Spectral - Element Method, Beowulf Computing, and Global Seismology[J]. Science, 2002, 298(5599): 1737 - 1742.

Microwulf parallelism system in university research team

LIU Xin-rong[1,2], HU Yuan-xin[1], LUO Jian-hua[3], GE Hua[4]

(*1. College of Civil Engineering, Chongqing University, Chongqing 400045, P. R. China*;
2. Department of Architecture and Civil Engineering, Logistical Engineering University, Chongqing 401311, P. R. China;
3. Nanjiang Hydrogeology and Engineering Geology Brigade, Chongqing 401121, P. R. China;
4. Chengdu Center of China Geological Survey, Chengdu 610081, P. R. China)

Abstract: On account of cost restriction, the computers of university research team are almost personal computers, which can not perform parallel computational tasks related to multiple computer nodes. Based on architecture of Beowulf and common hardware components bought from mass PC market, a Microwulf parallelism system with four nodes and eight CPU cores, named LXR01, was built. Windows7 and Linux operating system were installed on each node of the LXR01 system. The master node of LXR01 built shared file system with NFS, and adopted gcc and OpenMPI as main compiler and parallelism environment, respectively. The LXR01 system was benchmarked using HPL, from which 39.93 Gflops was max performance and the computational efficiency and cost efficiency of LXR01 system was 81.6% and ¥583.8/Gflops respectively. Meanwhile, it was concluded from parallel computational example of SPECFEM3D codes on LXR01 system that the LXR01 system can solve parallel computational problems effectively.

Keywords: parallel computation; Microwulf; Linux; OpenMPI

(编辑 欧阳雪梅)