地震灾情和地震动力学模拟系统①

廖凯宁, 郝永伟

(中国地震局地质研究所, 北京 100029)

摘 要: 中国地震局地质研究所地震灾情和地震动力学模拟系统承担着海量地震数据处理、地震动力学模拟等方面的计算任务. 从系统的建设目标、系统架构、软件环境、运维和业务应用等方面进行介绍, 总结了该系统在运维中遇到的问题和解决措施, 提出了优化并行程序在集群系统运行性能的办法, 并对下一步我所高性能计算集群系统的发展和改进方向进行分析.

关键词: 高性能计算集群系统; 地震动力学模拟; 系统运行环境优化

Earthquake Disaster and Earthquake Dynamics Simulation System

LIAO Kai-Ning, HAO Yong-Wei

(Institute of Geology, China Earthquake Administration, Beijing 100029, China)

Abstract: Institute of Geology, China Earthquake Administration's earthquake disaster and earthquake dynamics simulation system undertakes lots of computation tasks, such as processing of massive earthquake data and earthquake dynamics simulation. In this paper, we will introduce the construction objectives, system architecture, software environment, system operation and maintenance, system application etc., summarize the problems encountered during system operation and maintenance and corresponding solving measures, bring out a series of methods to optimize the parallel performance of HPC cluster system, and analyze the future development and improvement direction for our HPC cluster system.

Key words: high performance computing cluster system; earthquake dynamics simulation; optimization of the system running environment

中国地震局地质研究所依据《中央级科学事业单位修缮购置专项 2006 年——建立灾情与地震动力学模拟仿真系统》的要求, 构建了一个集仿真、高性能计算、存储能力共享的环境, 能完成高效率的地震动力学计算、设计并实现了以高性能集群计算为核心的地震灾情和地震动力学仿真模拟数据服务系统, 能为地震科学研究等相关领域的科研人员从事科研和模拟实验使用.

1 集群环境介绍

1.1 系统硬件架构

该系统采用惠普高性能刀片架构, 共 7 个机柜, 2 台管理节点和 2 台 IO 节点, 128 个计算节点, 每个计算节点配置两个 Intel(R) Xeon(R) E5405 四核 CPU, 主频 2.00GHz,单节点内存 8GB, 交换机 5 台, 三套 HPMSA 盘柜, 存储总量为 80TB.

最初将 HP HPC Cluster 划分为两套集群系统——Linux 业务区和 windows 业务区: 64 个 linux(操作系统为 Redhat AS 4.6)计算节点和 64 个 windows(操作系统为 Windows HPC Server 2008)计算节点. 两套集群系统的架构部署一致, 由管理节点、IO 节点、计算节点和存储四部分组成. 每套集群有四个内网: 千兆以太计算网络、IBA 网络、管理网络和基于 HP ILO 技术的监控网络. 计算网络由 HP5406 交换机连接所有计算节点、管理节点和 IO 节点; 管理网络由 hp2824

① 基金项目:国家自然科学基金青年基金(41104044)
收稿时间:2014-09-19;收到修改稿时间:2014-11-02

交换机连接所有计算节点、管理节点和 IO 节点; ILO 监控网络由 HP2626 交换机连接所有计算节点、管理节点和 IO 节点的 ILO. InfiniBand 网络由 IBA 交换机连接前 32 个计算节点(其中 16 个 Linux 计算节点, 16 个 windows 计算节点). 此外, 管理节点和 IO 节点各有一个网口连到行业专网, 用户可以在行业专网内直接访问. 磁盘阵列通过 SAN 光纤交换机挂接在 IO 节点下. (图 1)试运行情况: 对 16 台接入 InfiniBand 网络的 linux 计算节点进行 Linpack 测试, 结果为: 810.4 GFlops, 测试效率为 79%.

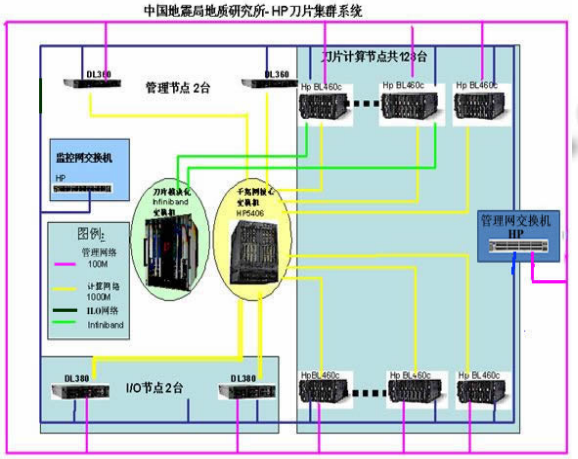

图 1 地震灾情和地震动力学模拟系统的网络拓扑结构

根据前期的运行情况和实际业务需求, 对 HP HPC Cluster 进行了重新部署, 更换了管理节点和 IO 节点设备, 重新划分集群系统的业务分区——80 个 linux 计算节点和 48 个 windows 计算节点, IBA 网络的 32 台计算节点全部为 linux 系统. 调整了 linux 业务区的网络拓扑结构, 增加了 70T 的存储空间.

1.2 软件环境

(1) 操作系统

windows 业务区所有节点安装的操作系统为 Windows Server 2008 HPC Edition. Linux 业务区最初安装的操作系统为 Redhat Linux 4 Update6, 后升级为 Redhat Enterprise Linux 5.5.

(2) 集群管理软件

Windows 业务区的管理节点和计算节点安装了 Windows Server 2008 HPC Edition 和 Microsoft HPC Pack 2008 管理集群系统.

最初 Linux 业务区安装的集群管理软件为 HP 的 CMU V3.2, 在管理节点安装 CMU server, 在计算节点、IO 节点安装 CMU 的监控模块. 通过 CMU 可以完成各节点软件的安装、文件的分发、备份、系统实时监控和远程开关机等功能.

系统重新部署后, 安装了 CHESS3.0(Clustertech HPC Environment Software Stack)联科高性能计算环境管理软件. CHESS3.0 基于 B/S 架构, 管理员和用户通过 Web 界面就可对集群环境和作业任务进行查看并执行操作(有权限限制).

(3) 应用环境及软件

目前系统安装了 GNU(C、C++、Fortran)、Intel、PGI 编译器. 应用软件主要有数值模拟软件 Ansys12, 三维电磁反演软件 WSINV3DMT、ModEM, GPS 数据处理软件、CFD 商业软件 fluent12.0、有限元计算分析软件 ABAQUS 等. 该套系统的运行环境如图 2 所示.

图 2 系统的运行环境

2 系统优化措施

经过两年多的运行实践, 根据科研人员对高性能计算的使用需求与预期目标, 对该套系统进行了针对性的调整完善.

2.1 调整了集群系统架构

针对 linux 环境下的应用软件和高性能计算用户的数量不断增多情况, 重新规划了业务分区、优化了系统. 一是将 IBA 网络内的计算节点全部划为 Linux 系统, 总计 80 计算节点的 linux 业务区, 重点保证 linux 用户的业务需求; 二是更换 Linux 业务区的管理节点和 IO 节点设备, 调整网络拓扑结构、增加磁盘阵列; 三是将系统版本升级为 Redhat Enterprise Linux 5.5; 四是制定集群管理策略, 从整体上对集群资源进行合理的分配和管理.

2.2 采用了符合应用特色的集群管理软件

最初 Linux 业务区安装的是 HP 的 CMU V3.2 集群管理软件, 该软件的功能简单, 仅具有节点设备增减、系统安装备份、远程开关机的功能, 类似增删集群用户等简单的工作需要执行大量指令才能完成, 系统管理员无法从常规系统管理中解放出来; CMU 的性能不太稳定, 诸如 Java 版本升级会导致 CMU 服务停止, 系统稳定性差; CMU 也未能解决高性能集群管理员最关心的问题——作业调度和资源配置策略, 如何合理优化系统环境、提高集群系统的使用率.

随着用户和计算量的不断增多, 如何合理分配计算资源、让更复杂更大规模的计算任务以更高的性能和准确性运行在计算平台, 成为高性能集群计算系统迫切需要解决的问题. 通过综合分析和运行实践, 采用 CHESS3.0 高性能计算环境管理软件, 可以满足高性能计算系统的管理和使用需要, 确保系统在多用户多程序状态下有条不紊的工作.

2.3 优化了系统运行环境

地震灾情和地震动力学模拟系统是面向多用户多任务的计算平台, 在硬件条件基本确定的情况下, 只能通过优化运行环境来提高软件程序运行性能. 我们的改进方案是: 在资源分配时, 一方面能够根据用户程序特征将更为优化的资源集合分配给用户作业使之获得更好的运行状态, 另一方面从整体上保障资源分配的合理性和系统运行的稳定性.

1) 结合用户作业运行的需求, 科学制定管理策略、任务调度和资源配置策略以保证系统的稳定性和系统资源分配的合理性.

保证整个系统安全稳定的前提下, 充分赋予了用户的使用权限. 例如, 为确保并行计算的顺利运行和性能最大化, 各计算节点间采用互相信任的访控机制, 同时为保证系统安全禁止普通用户直接登录计算节点. 根据用户的使用情况, 可以通过 chess 软件的作业调度模块制定任务调度, 如设计多种队列、为用户(用户组)制定优先级, 并不断完善资源配置策略, 如将某些计算资源优先分配给某个用户(组), 系统优先选择分配空闲计算资源或负载量小的计算资源等等.

2) 根据不同程序的特性, 不断改善应用程序在系统中的运行环境以提高计算速度和系统效率.

地震灾情和地震动力学模拟系统承担着海量地震数据处理、地震动力学模拟等方面的科研任务. 这类程序运算的特征是进程间频繁的数据交换、巨大的磁盘 I/O 访问量, 因此当作业量巨大时容易导致网络停滞、计算迟缓. 解决这一情况仅依赖作业管理系统是不够的, 还需优化系统环境, 让应用程序运行在符合其特征的环境下, 从而减少运行时间提高系统效率. 可以根据网络现状进行资源划分, 将不同队列、不同作业尽量划分到整个网络中的某一小部分子网中, 尽可能使通信局部化, 不影响其他作业或队列的运行[1]. 将进程间消息交换频繁的一类应用程序优先分配 InfiniBand 网络资源, 降低进程间通信延迟对程序性能的影响.

3) 保证计算节点的负载均匀, 提高系统整体的运行效率.

该套系统中的每个计算节点配置两个 Intel(R) Xeon(R) E5405 四核处理器、8GB 内存. 实验发现, 通过选取尽可能少的计算节点而占满这些节点的处理器来完成一项作业的效率往往不是最高的, 而是选取相对多的计算节点以保证计算节点的内存不被占满、处理资源具有一定的空闲比所获得的性能更高些. 以一个 24 个进程的三维电磁反演软件 WSINV3DMT 并行程序为例, 该程序的特点是占用大量内存. 分配不同的计算资源进行计算得出运算所需的时间见表 1.

表 1 分配不同计算资源所需运算时间

计算资源的分配	计算时间(IBA)(小时:分钟:秒)
Nodes=3:ppn=8	41:40:19
Nodes=4:ppn=6	31:40:06
Nodes=6:ppn=4	21:06:10
Nodes=8:ppn=3	18:39:56

因此, 可以在运行环境允许的情况下, 根据用户作业的特征合理调配计算资源; 在作业高峰时还可对用户并发运行的作业个数进行限制, 保证计算节点的负载均衡.

4) 通过调整网络拓扑结构增加带宽、提高网络传输性能.

实际运行中存在因作业数量和运算规模的激增而导致系统出现网络停滞、系统运行缓慢的现象. 经过仔细查看用户作业情况、计算节点根分区、内存、pbs 系统、NFS 系统等发现, 这一现象发生时总有某用户的大量作业在运行, 该类程序的特点是大数据量的并发读写, 计算节点在与 IO 节点通信时超过 100 多兆/秒. 正是计算节点通过千兆以太网络访问 IO 节点的这

种传统共享网络存储结构导致了海量数据和并发网络访问时的I/O等待，集群系统效率低下. 在现有条件下，采取将IO服务器不常用的S3网段、外网网段取消，利用网卡绑定技术将 IO 服务器的三块网卡绑定当作一块网卡使用，计算节点和IO节点间的数据传输带宽能提高至3000兆. 同时将磁盘I/O的挂载点由S2网段更换至S1网段(即做了bonding的网卡下)，经测试，数据与 I/O 交换的带宽明显提高(如图 3：蓝色虚线为原先网络拓扑结构，红色为绑定网卡). 同时，修改系统设置、优化作业调度系统、调整用户使用权限，有效地缓解了数据交换带宽的I/O瓶颈. 实际运行证明，较好地解决了问题，有效提高了网络传输性能.

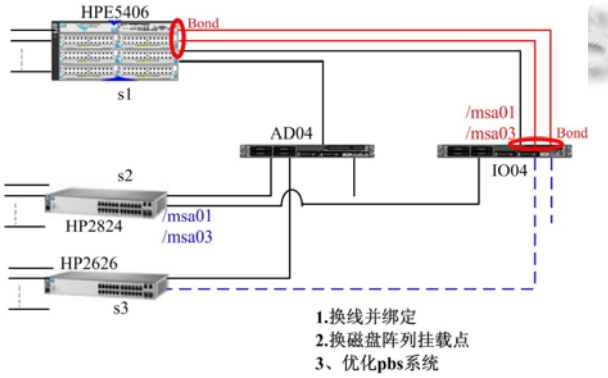

图 3　作业量过大导致系统运行缓慢现象的物理解决方案

3　应用情况

系统优化完善后，运行稳定，便于使用和管理，用户数量明显增多，作业任务激增，保持了较高的使用率. 目前，常用用户 30 多名，2013-2014 上半年度完成 5678 例应用作业，作业累计运行时间达 161730 小时.

地震灾情和地震动力学模拟系统可以完成以下功能应用：一是进行海量地震波数据的反演计算；二是利用震源及区域应力场资料计算周围区域的应力变化；三是借助于 Ansys、Abaqus 软件进行地震动力学与地球动力学的模拟；四是基于陆态网络连续 GPS 观测站数据处理以及 IGS 核心站 GPS 数据处理计算地壳形变；五是利用地震环境噪声成像方法研究地壳的断层精细速度结构；六是震后形变研究及震后模型计算；七是形变场时空演化模拟研究；八是中国及邻区地壳以及地幔处的深部构造研究；九是深部电性结构与地力学研究；十是地震震源机制的研究等.

由于地震灾情和地震动力学仿真模型演算的时长都相当长，可达数十万年，普通计算机无法进行如此大时间尺度的运算，并且每次运算的结果数据可达千G，也是一般配置计算机不能满足的. 该系统的运用有效地提高了运算速度，节省了计算时间. 以计算地壳网络的GPS站点1998-2010年间数据为例，使用32个计算节点、每节点2进程同时运算约需3-4天. 相比较一般的塔式服务器，运算时间缩短约6倍. 同时，该套系统由最初的 80T 扩容至 150T 的存储能力可以轻松解决运算中产生的临时文件和结果数据等.

截至目前，本系统已承担了二十多项国家自然科学基金项目、国家科技重大专项、地震行业专项和所长基金等项目的计算任务，完成大量数字图像处理计算 、三维正反演模拟计算.

4　系统的改进方案和进一步的研究方向

从该系统的实际运行情况，以及近年来高性能计算机系统发展和建设情况来看，系统可以在以下方面进一步优化和完善：一是提高硬件的配置. 如更换计算节点设备提高CPU的主频，增加计算节点的内存以适应网格较多的模型计算. 二是建立设备备份. 考虑集群系统的稳定可靠性，最好为每个业务区配置两台管理节点和两以上的 IO 节点以避免单点故障的发生. 三是改变系统架构以消除 I/O 瓶颈. 部署并行文件系统取代传统共享网络存储系统，消除 I/O 瓶颈提高地震资料处理效率. 四是加强安全防护. 随着我所高性能计算的任务和存储数据的增多，超算集群需要兼顾内部和外部的安全防护[2]. 五是继续开展应用软件并行优化策略和系统运行环境优化策略的研究. 六是绿色节能. 高性能计算集群设备和周边配套设备的能耗极高，如何提高智能化管理、避免资源浪费是维管始终需要思考的课题.

5　结语

由于高性能计算集群系统拥有很高的计算能力、I/O 性能和存储空间，因此被越来越多地运用于大规模科学计算和复杂问题求解当中. 作为地球科学基础理论的地震科学研究需要获取分析大量的地球观测数据和开展以震源环境、地震过程和震源破裂机理等地震科学基础研究为理论依据的地震动力学模拟仿真实验，随着研究任务的加重、研究手段和应用软件的升

级, 地震行业对高性能计算集群系统的运算性能和存储空间的要求和依赖程度越来越高. 因此, 及时总结当前系统的运行现状、规划下一步我所高性能计算集群的发展和研究方向十分必要, 对于地震行业系统的高性能计算集群的管理和规划也具有重要的参考意义.

致谢

感谢固体地球物理与深部构造研究室提供的算例; 感谢联智科技有限公司李工在系统软件测试中提供的技术支持.

参考文献

1 曹宗雁.高性能计算集群运行时环境的配置优化.科研信息化技术与应用,2011,2(6):52–61.

2 牛铁,朱鹏,曹宗雁,刘飞.超级计算机群的安全防护.科研信息化技术与应用,2011,2(6):45–51.