第34卷 第21期 计 算 机 工 程 2008年11月
Vol.34 No.21 Computer Engineering November 2008
·开发研究与设计技术· 文章编号：1000—3428(2008)21—0255—04 文献标识码：A 中图分类号：TP303

超级计算机的多层架构抽象及描述

刘 辉，蒋烈辉，周武明，赵 钊，汪 莹
(解放军信息工程大学信息工程学院，郑州 450002)

摘 要：超级计算机系统发展迅猛，不仅速度上日新月异，而且其结构也日益复杂化、多样化。为了能够准确有效地描述超级计算机的结构，给出统一的描述方法，提出超级计算机的多层架构思想。为超级计算机的进一步逻辑抽象或编写其模拟器提供了基础，并对超级计算机系统的设计具有一定的指导作用。

关键词：超级计算机；多层架构；计算层次；计算部件

Multi-level Structure Abstract and Description of Supercomputer

LIU Hui, JIANG Lie-hui, ZHOU Wu-ming, ZHAO Zhao, WANG Ying
(College of Information Engineering, PLA Information Engineering University, Zhengzhou 450002)

【Abstract】 The development of supercomputer is very fast, not only the speed changes with each passing day, but also the structure becomes more and more complicated and diversiform. In order to describe supercomputer's structure rightly and availably, and give the uniform description method, this paper brings forward the multi-level structure and description of supercomputer. It can ground in farther abstracting or programming supercomputer simulator and help to design supercomputer system.

【Key words】 supercomputer; multi-level structure; computing level; computing component

1 概述

性能分析和预测是超级计算机研究的一个重要方面，而模拟器在性能分析和预测方面发挥着很重要的作用。目前虽然已有多处理器、多核处理器的全系统模拟器(如共享内存多处理器模拟器MulSim[1]、基于龙芯CPU的多核全系统模拟器SimOS-Goodson[2])，但对超级计算机模拟器的研究还很少。本文提出的超级计算机的多层架构思想和描述方法可为超级计算机模拟器提供超级计算机的结构信息，而且各种不同结构的超级计算机有相同的描述格式，这些格式化了的超级计算机结构信息为模拟器的可扩展性提供了良好的基础。

2 超级计算机的多层架构抽象

2.1 相关定义及分析

2007 年 6 月底，ISC07(国际超级计算机大会)公布了第29 届TOP500 超级计算机名单。这次排名的前 10 名和上次相比几乎全部洗牌，安装在美国能源部劳伦斯-利弗摩尔国家实验室(LLNL)的IBM蓝色基因/L系统继续保持第一名的位置，另外 2 套新的蓝色基因/L也进入了前十；2002 的榜首——地球模拟器则排名 20[3]。下面对这 2 个系统进行分析。

蓝色基因用 2 个嵌入式的IBM PPC440 处理核构成了一个节点，共有 65 536 个计算节点、1 024 个IO节点[4]，利用 5 个互联网络将这六万多个节点连接成一个完整的系统；地球模拟器采用的是共享内存的大规模并行向量系统体系结构，每个节点包括了 8 个向量处理器，而整个系统有 640 个处理器节点，16 个处理器节点构成一个集群，共有 40 个集群，40 个集群通过网络互连构成了一个由集群构成的集群，所以地球模拟器是一个两层集群系统，叫做超级集群系统[5]。地球模拟器中，第一级集群和第二级集群共用一个 640×640 的交叉开关进行互连。下面提出多层架构思想中的 2 个重要定义。

定义 1 计算部件

计算部件是能够执行一个完整程序的部件。

根据定义，现代多核处理器中的处理器核就是一个计算部件，因为如果将一个程序及其所有数据都存放到处理器核内部的 Cache 中，那么处理器核就能完整地执行这个程序。因此，在蓝色基因/L 中，其不同粒度的计算部件有：PowerPC440 处理核，计算节点和 I/O 节点，蓝色基因/L 系统本身。对于地球模拟器来说，其不同粒度的计算部件有：向量处理器，处理器节点，一级集群和超级集群。

定义 2 计算层次

计算层次是一个计算部件，它由粒度更小的计算部件互连构成。

结合定义 1，可知计算部件是一个嵌套递归的概念——粒度小的计算部件可以构成粒度更大的计算部件，这个粒度更大的计算部件又称为计算层次。有了计算部件和计算层次这 2 个概念就可以对超级计算机的结构进行多层架构抽象了。对于蓝色基因/L 系统来说，它有 2 个计算层次：节点层次，它由 2 个粒度更小的计算部件——PowerPC440 共享第三级 Cache 耦合而成；蓝色基因/L 系统层次，它由 65 536 个计算节点和 1 024 节点互连构成。而地球模拟器则有 3 个层次：处理器节点层次，它由 8 个向量处理器构成；一级集群层次，它由 16 个处理器节点构成；超级集群层次，它由 40 个一级集群构成。

作者简介：刘 辉(1982－)，男，硕士研究生，主研方向：计算机系统结构，计算机应用技术；蒋烈辉，教授、博士；周武明、赵 钊、汪 莹，硕士研究生

收稿日期：2008-01-31 **E-mail：**zhanshen53@163.com

2.2 计算层次的结构抽象

根据计算层次的定义，计算层次由计算部件互连构成，因此在一个计算层次中应该包括计算部件和互连部件。但一个计算层次并非只有这 2 种部件，在实际设计中，互连部件往往还会连接相应的存储部件。因此，可以将一个计算层次内的部件抽象成 3 类：计算部件，存储部件和互连部件。这 3 类部件的关系如图 1 所示。

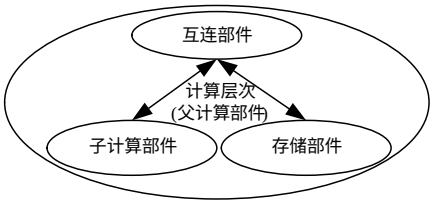


图 1　3 类部件之间的关系

由定义 1 和定义 2 可知，计算层次也是一个计算部件，它由更小粒度的计算部件构成，故称为父计算部件，同时称小粒度的计算部件为子计算部件。

通过多层架构抽象和计算层次的结构抽象，可以很容易地给出蓝色基因/L 和地球模拟各个计算层次的结构。

2.2.1　蓝色基因/L 各计算层次的结构

由前面的定义以及分析可知，蓝色基因/L 有 2 个计算层次：节点层次和系统本身。

图 2 是蓝色基因/L节点层次的内部结构[6]。其子计算部件是 2 个PowerPC440 处理核，互连部件是总线，存储部件是第三级 Cache。2 个 PowerPC440 通过共享第三级Cache耦合成 1 个节点。

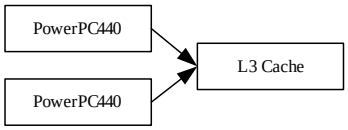


图 2　蓝色基因/L 节点层次的内部结构

图 3 是蓝色基因/L 系统层次的内部结构，其子计算部件是各个节点，图中的 DB 是存储部件，以太网、控制网络、JTAG、Torus 和 Tree 5 个网络是互连部件。互连部件连接所有的节点和存储设备构成一个完整的系统。

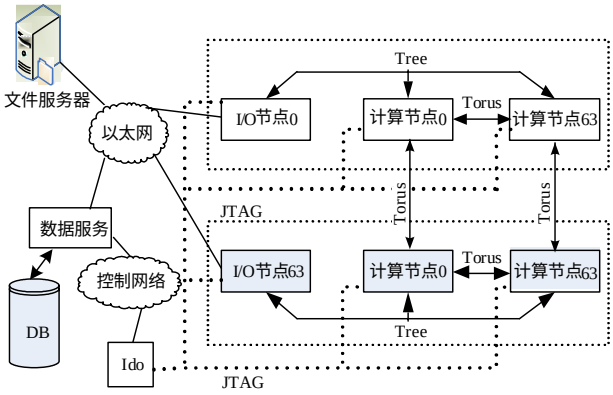


图 3　蓝色基因/L 系统层次的内部结构

2.2.2　地球模拟器各层次的结构

地球模拟器总共有 3 个层次：节点，一级集群，二级集群(超级集群)。

图 4 是地球模拟器节点层次的内部结构，其子计算部件是 8 个向量处理器，存储部件为内存，互连部件各类总线。总线连接所有子计算部件和存储部件构成一个完整的节点。

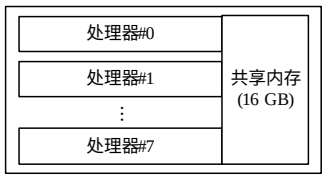


图 4　地球模拟器节点层次的内部结构

图 5 是地球模拟器的结构，虚框代表了一级集群，整个图代表了超级集群。

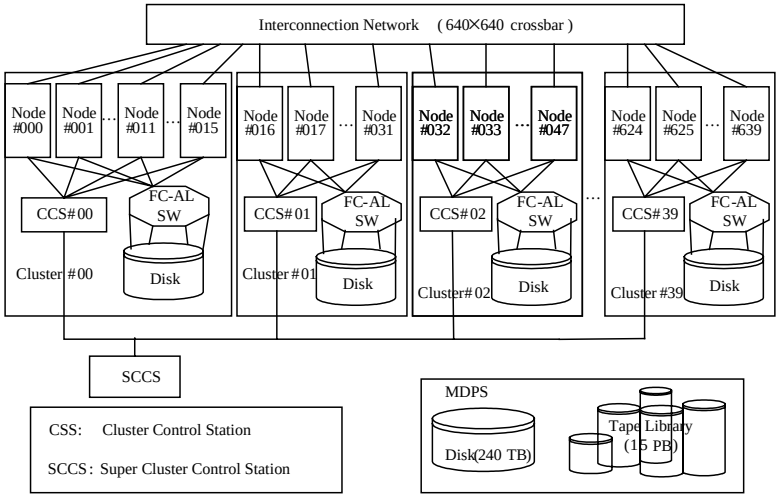


图 5　地球模拟器的结构

一级集群是地球模拟器的第 2 个计算层次，其子计算部件是 16 个节点和 1 个集群控制站(Cluster Control Station, CCS)，存储部件是图中虚框内的磁盘(Disk)，互连部件是 FC 网络和 640×640 的交叉开关。

超级集群是地球模拟器的第 3 个计算层次，其子计算部件是 40 个一级集群和 1 个超级集群控制站(Super Cluster Control Station, SCCS)，存储部件是图右下方的MDPS[7]，互连部件是存储网络和 640×640 的交叉开关。

3　超级计算机的结构描述

3.1　类 html 的标签语言

现在已经对超级计算机系统进行了多层架构抽象，对计算层次内部进行了部件级的抽象。

采用类 html 的标签来表达超级计算机的结构。

根据前面的多层架构思想和层次内部的部件级抽象，总共有 4 对标签：<LEVEL> </LEVEL>, <COMPUTER> </COMPUTER>, <STORAGE> </STORAGE>, <NETWORK> </NETWORK>。

所有的标签必须成对出现，标签对之间的所有内容描述了对应层次或者部件的结构信息：<LEVEL> </LEVEL>之间的内容描述一个层次的结构，<COMPUTER> </COMPUTER>之间的内容描述了一个计算部件的信息，<STORAGE> </STORAGE>和<NETWORK> </NETWORK>之间内容分别描述一个存储部件和互连部件的信息；<COMPUTER> </COMPUTER>标签对之间可以再有<COMPUTER> </COMPUTER>标签对，这也正好体现了计算部件是一个嵌套递归的概念。

利用标签可以基本确定超级计算机的结构框架，那么如何表示超级计算机各层次各部件的属性信息呢？根据关注的信息不同，在标签内部添加属性名，并以“属性名＝值”的形式给各种属性赋值。比如

```
<LEVEL>
Serial=1
</LEVEL>
```

其中，Serial 是层次的一个属性名，它的值被赋为 1，说明是第一层。同样的，可以为各类部件添加所需要的属性信息。需要说明的是，所有的部件都有 Type 和 Name 属性，它们的值作为区分各部件的标识。

3.2 蓝色基因/L 的结构表示

下面利用这 4 个标签对和属性来表示蓝色基因/L 系统和地球模拟器系统的结构，其中使用到新的属性会进行说明。

蓝色基因/L 的配置信息如下：

```
<LEVEL>
  Serial=1
  <COMPUTER>
      Type=Processor
      Name=蓝色基因计算节点
      <COMPUTER>
          Member=PowerPC440
          Num=2
      </COMPUTER>
      <STORAGE>
          Type=Cache
      Name=蓝色基因第三级Cache
      </STORAGE>
      <NETWORK>
          Type=bus
          Name=共享Cache互连
      </NETWORK>
  </COMPUTER>
  <COMPUTER>
      Type=Processor
      Name=蓝色基因IO节点
      <COMPUTER>
          Member=PowerPC440
          Num=2
      </COMPUTER>
      <STORAGE>
          Type=Cache
      Name=蓝色基因第三级Cache
      </STORAGE>
      <NETWORK>
          Type=bus
          Name=共享Cache互连
      </NETWORK>
  </COMPUTER>
<LEVEL>
  <LEVEL>
  Serial=2
      <COMPUTER>
          Type=多节点互连
          Name=蓝色基因/L
  Member=蓝色基因计算节点
  Num=65536
  Member=蓝色基因IO节点
  Num=1024
      <STORAGE>
          Type=RAID
          Name=蓝色基因磁盘阵列
      </STORAGE>
      <NETWORK>
          Type=互连网络
          Name=蓝色基因环形网
      </NETWORK>
      <NETWORK>
          Type=互连网络
          Name=蓝色基因系统控制网
      </NETWORK>
      <NETWORK>
          Type=互连网络
      Name=蓝色基因系统中断和栏栅网
      </NETWORK>
      <NETWORK>
          Type=互连网络
          Name=蓝色基因全局集中网
      </NETWORK>
      <NETWORK>
          Type=互连网络
        Name=蓝色基因Gbit级以太网
      </NETWORK>
      </COMPUTER>
      <LEVEL>
```

地球模拟器的配置信息如下：

```
<LEVEL>
Serial=1
<COMPUTER>
    Type=Processor
    Name=地球模拟器节点
    <COMPUTER>
        Member=向量处理器
        Num=8
    </COMPUTER>
    <STORAGE>
        Type=Memory
        Name=节点内存
    </STORAGE>
    <NETWORK>
        Type=bus
        Name=共享内存互连
    </NETWORK>
</COMPUTER>
</LEVEL>
<LEVEL>
Serial=2
<COMPUTER>
    Type=集群
    Name=地球模拟器一级集群
    <COMPUTER>
        Member=地球模拟器节点
        Num=16
    </COMPUTER>
    <STORAGE>
        Type=Disk
        Name=地球模拟器一级集群磁盘
```

```
    </STORAGE>
<NETWORK>
        Type=存储网络
        Name=地球模拟器FC网络
    </NETWORK>
    <NETWORK>
        Type=交叉开关
        Name=640x640交叉开关
    </NETWORK>
</COMPUTER>
</LEVEL>
<LEVEL>
Serial=3
<COMPUTER>
    Type=集群
    Name=地球模拟器超级集群
    <COMPUTER>
        Member=地球模拟器一级集群
        Num=40
    </COMPUTER>
    <STORAGE>
        Type＝网络存储
        Name=地球模拟器MDPS
    </STORAGE>
    <NETWORK>
        Type=存储网络
        Name=地球模拟器存储网络
    </NETWORK>
    <NETWORK>
        Type=交叉开关
        Name=640x640交叉开关
    </NETWORK>
</COMPUTER>
</LEVEL>
```

配置信息中有 Serial, Name, Type, Member 和 Num 属性，其中前 3 项已经说明；Member 是子计算部件名称，而 Num 是子计算部件的个数。

根据标签对和标签对内部的属性，可以很容易得到蓝色基因/L 和地球模拟器的结构，说明多层架构的抽象及其描述方法是有效可行的。

3.3 应用说明

可以通过标签对刻画超级计算机的结构特征，而各个具体部件的具体特征可以通过属性来描述。但是到底需要哪些属性并不是确定的，这需要根据具体的应用来确定。如果用户关心超级计算机的性能，那么在描述超级计算机的信息时就特别关注各个部件和性能相关的参数和性能指标，此时用户就应该添加和这些参数和性能指标相对应的属性，用户自己负责管理这些对应关系。这虽然给用户带来了不便，但是提高了描述语言的灵活性。

4 结束语

本文通过跟踪当今超级计算机的发展方向，并以蓝色基因/L 系统和地球模拟器为例，分析了超级计算机的结构特征，提出了超级计算机的多层架构思想，并对超级计算机各个层次内部的结构进行了统一的抽象，在此基础上给出了对超级计算机结构信息的描述方法。最后又以蓝色基因/L 系统和地球模拟器系统为例对这种描述方法进行了阐述，也验证了这种方法能够很好地描述超级计算机的结构信息。

参考文献

[1] Matloff N, Rich K. MulSim Multiprocessor Simulator[D]. California, USA: Department of Computer Science, University of California at Davis. 2000.

[2] 高　翔, 张福新, 汤　彦, 等. 基于龙芯 CPU 的多核全系统模拟器[J]. Journal of Software, 2007, 18(4): 1047-1055.

[3] TOP500 List-June 2007(1-100)[Z]. [2007-06-27]. http://www.top500.org/list/2007/06/100.

[4] Almási G, Archer C. Design and Implementation of Message-passing Services for the Blue Gene/L Supercomputer[J]. IBM Journal of Research and Development, 2005, 49(2/3): 393-406.

[5] Yanagawa T. Operating System for the Earth Simulator[J]. NEC Res. & Dev., 2003, 44(1): 43-46.

[6] 张跃冬, 吕　涛, 樊建平. IBM Blue Gene 项目介绍及 BG/L 的体系结构和系统软件[EB/OL]. (2004-02-17). http://lib.ict.ac.cn/ITL/data/2004/1/IBM%20Blue%20Gene%E9%A1%B9%E7%9B%A%E4%BB%8B%E7%BB%8D%E5%8F%8ABG.pdf.

[7] Itakura K. MDPS: The New Mass Data Processing Storage System for the Earth Simulator[J]. Journal of the Earth Simulator, 2006, 5(20): 20-24.

（上接第 251 页）

参考文献

[1] 贾云得. 机器视觉[M]. 北京: 科学出版社, 2000.

[2] 袁宝民, 金一粟, 于万波, 等. 基于移差扫描和窗口搜索的车牌定位方法[J]. 计算机工程, 2003, 29(14): 103-105.

[3] 关　鹏, 顾晓东, 张立明. 一种基于图像处理的铁轨自动检测方法[J]. 计算机工程, 2007, 33(19): 207-209, 212.

[4] Doiron T D. Computer Vision Based Station for Tool Setting and Tool Form Measurement[J]. Precision Engineering, 1989, 11(4): 231-238.

[5] Weng J, Cohen P, Herniou M. Camera Calibration With Distortion Model and Accuracy Evaluation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(10): 965-980.

[6] Lenz R K, Tsai R Y. Technique for Calibration of the Scale Factor and Image Center for High Accuracy 3-D Machine Vision Metrology[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1988, 10(5): 713-720.

[7] 田涌涛, 王有庆, 李从心, 等. 基于共线点的透镜畸变系数标定[J]. 上海交通大学学报, 2003, 37(11): 1739-1742.

[8] Tian Yongtao, Li Juan, Wang Youqing, et al. A Simple Centroid-based Method to Find Principal Point of Image Plane for Pincushion Distortion[J]. International Journal of Advanced Manufacturing Technology, 2004, 23(1): 110-115.