网格化的动态自组织体系结构 DSAG

樊建平 陈明宇

(中国科学院计算技术研究所国家智能计算机研究开发中心 北京 100080)

(fan,cmy@ict.ac.cn)

摘 要 传统的高性能计算机系统面临着变革，特别是体系结构需要创新。应用程序的设计方法与运行模式、可重构计算、网格化技术和光互连技术将深刻影响未来高性能计算机系统体系结构的发展。本文提出了网格化动态自组织体系结构(DSAG),可以支持体系结构按需定制（Architecture on Demand）的新模式，并探讨了相应光互连及 DSAG 操作系统的设计以及 DSAG 思想在其他层次系统设计中的应用。

关键词 体系结构，高性能计算，光互连，可重构计算，网格化技术

中图法分类号 TP338.4

Dynamic Self-organized computer Architecture based on Grid-component(DSAG)

FAN Jian-Ping, CHEN Ming-yu

(National Research Center for Intelligent Computing Systems, Institute of Computing Technology,Chinese Academy of Sciences, Beijing 100080)

Abstract The traditional high performance computer (HPC) architecture is facing innovation. The design and running models of application, reconfigurable computing, grid and optical interconnection will impact the future architecture of HPC deeply. In this article the dynamic self-organized architecture based on grid component (DSAG) is presented. DSAG is based on the concept of de-clustering, that is to divide the different functional component of a traditional computer apart and reorganize them dynamically. The separated functional components (grid-components) provide service independently via optical interconnection network. Then the HPC system will be built with grid-components dynamically according to the application requirements . DSAG will enable very large scale HPC system to be built. Based on DASG a new HPC design model – "architecture on demand" is presented. In this model the architecture will be adjusted corresponding to the application characteristics instead of rewriting program to adapt to architecture change. The supporting optic technology and operating system for DSAG are discussed. Other levels system design that may apply the DSAG concept are presented program.

Key words architecture, high performance computing, optical interconnection, reconfigurable computing , grid

1、引 言[1]

以并行为主要特征的当代高性能计算机体系结构从 SMP（共享存储）、MPP（基于消息传递的大规模并行）到 Cluster（机群）以及现在的"后 Cluster 时代",其技术重点一直是解决以微处理器为核心的计算机系统的互连问题。以至于学术界曾有人提出过"体系结构就是互连问题"[1]。高性能计算机系统的性能提高有相当一部分是由 VLSI 和微处理器的技术进步带来的。单芯片的集成度按摩尔定律仍将增加，速度也将提高。但是单芯片的处理能力提高，并不意味着并

[1] 本课题得到国家高科技发展计划(863)基金支持(2003AA1Z2070)和中国科学院知识创新工程支持

行系统的整体性能也一定会相应地提高。一方面随着单机系统日趋复杂，功耗、可靠性等问题也越来越突出，多机并行系统中这些问题的累加效应已经无法回避[2]。另一方面，随着芯片内部信号频率进入 GHz 时代，传统芯片之间的铜线连接已经达到极限，成为系统的瓶颈。这些问题都阻碍了更大规模高性能并行计算机系统的研制，同时也引发高性能计算机体系结构的再研究。

针对高性能计算机体系结构来设计高性能算法依然是应用科学家今天必须面临的问题。对体系结构及系统软件详细的了解与理解是写出高效程序的关键，甚至有用户为提高应用程序效率自己重新开发操作系统的事例[3]。如何屏蔽高性能计算机的复杂性，使应用科学家集中于应用问题本身，摆脱应用系统及其算法跟随每一代新机器的研制而重新设计一遍的局面成为计算机设计师追求的目标之一 ，也是本文工作的研究目标之一。

近年来发展起来的可重构计算(Reconfigurable Computing)的主要思想是保持系统具有硬件一样运算效率的同时还具有软件解决方案的灵活性。各种可重构计算的支撑技术（FPGA、Multi-FPGA systems、Multi-Context FPGAs、DRFPGA、Embedded SRAMs、RISC/FPGA hybrid 等）的发展扩大了其应用范围（信号处理、机器人控制、并行处理、安全、人工智能、图像处理与压缩等）。硬件可重构技术的发展将逐步改变传统软件设计的模式（从软件向硬件靠到软硬件互动）[4]。而将可重构硬件的思想引入到高性能计算机系统设计中来是触发本文思想形成的关键之一。

基于 FPGA 的可重构计算改变的是芯片内部功能单元和电路连接。而高性能计算机系统是一个更为庞大和复杂的系统，其可重构计算将依赖于网格化的部件和光互联。

网格计算的想法最早来源于通过 Internet 共享高性能计算资源的尝试，到现在网格已经被公认将成为下一代 Internet 应用的主要模式[5]。网格的核心思想是资源共享、互连互通和应用服务。网格应用不仅仅影响到最终用户，对于高性能计算机系统来说，资源的网格化是一种使能技术，为更大尺度的高性能计算机系统的设计提供了支持。网格的将来可能不仅仅是“ 计算机通过网络连接起来 ”，而是成为真正意义上的“ 网络连接起来的计算机 ”。

光互连相比于铜线连接具有高带宽、长距离、低损耗、无串扰等无可比拟的优点。这使得光互连首先在通信领域内迅速发展起来，其发展速度甚至已经超过了微电子技术的摩尔定律。而随着基于表面发射激光(VCSEL)技术将半导体和激光技术结合起来，光技术开始逐步走入计算机系统内部。随着成本的下降和技术的进一步成熟，光电技术的结合将逐步替代传统的电路技术，成为今后 20 年计算机制造的主流技术[6]。

本文将主要探讨一种新的网格化的可动态自组织的高性能计算机体系结构（DSAG）的设计思想，及基于光互连实现的考虑。同时对 DSAG 体系结构可能的应用范围进行探讨。

2、发展思路

2.1 “ 聚 ” (Clustering)—传统的发展思路

“聚”是传统的计算机系统设计的核心思想。通过将各种功能部件设计的尽可能的靠近，可以缩短部件之间的通讯距离，同时减少整体系统的体积。当前的一些研究方向如 SoC（片上系统）, SMP on Chip（片上多机 ）,Blade(刀 片) ， 以 及 Processor in Memory(计算存储一体芯片)等都是这一思想的延伸。聚的思想在某些应用场景下可以达到资源的最优组合。但是一旦聚在一起的资源组合不能满足特定应用的需求时，通信开销仍是不可避免的。当前 CPU 和 Memory 之间的“ 差距 ”（存取带宽与延迟）越来越大就是典型的例子。由于聚合系统产生的封闭性使得资源共享变得更为困难，系统的制造成本普遍高。这也是分布式共享内存的 CC-NUMA 计算机造价要远远高于机群系统的原因。

将不同的功能部件聚合在一起自然地

产生一种排斥性。不同的功能部件有不同的材料和生产工艺、不同的接口方式和不同的物理尺寸。将不同的功能部件集中在一起，不但必须增加各种结合的开销，同时也增加了系统设计和生产的复杂性，降低了可靠性。在并行高性能计算机系统的规模进一步增大的情况下，“聚”可能并不是一种有效的途径。

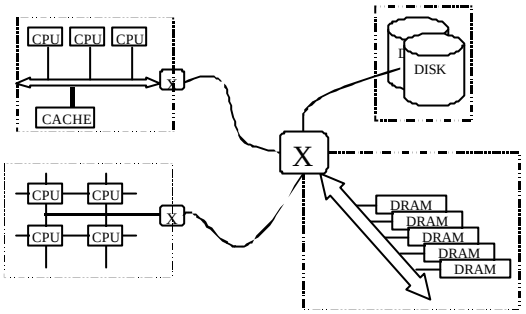

图 1 计算机系统的拆分和重组(De-Clusering)

2.2“拆”(De-clustering)— 我们的发展思路

随着高性能计算机系统规模的进一步扩大，单一的聚合系统必然无法满足所有应用的需求。系统之间的联接和共享将是不可避免的。既然如此，是否可以将原来聚合的系统拆分开，将不同的功能部件分离、同类的功能部件集中并相对独立、同时针对通信连接进行优化设计，从而构建更大规模的系统呢？这样拆分和重组将可能带来以下好处：

- 同类部件集中生产是很多大规模工业化生产的基本特征。同类部件聚合，将减少因不同部件规格差异带来的结合的开销，简化系统的生产过程，通过一种规模效应来降低生产的成本。
- 同类部件集中和独立，将便于其他部件对资源的共享，消除原来聚合系统带来的壁垒，提高资源的利用率。
- 同类部件的集中将便于可靠性系统的设计，特别是便于资源备份和故障切换的实现。
- 不同类功能部件分离，将减少部件之间的依赖关系。组成系统的不同部件之间不再是固有的联接，这将便于独立部件的升级和扩充。
- 不同类功能部件分离，将便于系统动态的组合，从而增加系统的灵活性和适应性。

2.3 网格化——小范围的拆与大范围的聚

网格化思想是小范围的拆与大范围的聚的结合。小范围的拆，是指将原来属于一个完备的计算机系统的功能部件和资源拆分开来，使得每个功能部件都可以独立的对外提供服务。大范围的聚,是指在 Internet 范围内，利用各种资源服务的聚合来完成特定应用的需求。网格化的资源的范围很广泛，既可以是 CPU、内存、硬盘等物理的功能部件，也可以是一台完整的计算机节点，或者是一个标准的应用程序如数据库服务等。网格化资源的共同特点是具有特定的功能并通过网络独立地对外提供服务。

光互联技术为大范围的聚提供了关键的使能技术。光互联的特点是支持数据低损耗长距离传输和高传输带宽,这就为大空间尺度上网格化计算部件之间的高速数据交换提供了支持。采用光互连网络连接的带宽理论上已经可以和计算机系统内部的互联带宽相比拟,大大降低了本地资源和远程资源之间的区别。这使得利用网格化部件动态构建高性能计算机系统真正成为可能。

3、DSAG 体系结构的描述

3.1、DSAG 的定义

网格化动态自组织体系结构(Dynamic Self-organized Architecture based on Grid-component)是一种新的高性能计算机体系结构。简单的说，DSAG 就是根据应用的计算模式和运行模式的需要，自动利用高速网络中独立的、网格化的功能部件动态组织成一个计算机系统，并有效运行应用程序。

网格化是指每种功能部件都形成一种独立的具有自我描述功能的网络服务。动态是指由于网格功能部件分布在网络的不同地方，将不再存在一个传统意义上一个固定的、物理的计算机系统。构成一个计算机系统的功能部件可以通过网络动态的申请和释放，根据应用的需求和可用的资源进行动态的调配。自组织是利用网格化的功能单元

具有自我描述的功能，自动在网络上搜索满足需求的功能单元服务，自动进行资源分配和协商，收集足够的计算资源并自动建立功能部件间的关联，以组织成一个完整的计算机系统。

传统的应用程序通常是针对固定的计算机体系结构进行优化设计的。由于各种科学计算问题的模型和算法的变化相对缓慢，而计算机体系结构与实现使能技术变化相对较快，形成为机器“配”软件的传统（应用适应体系结构）。DSAG 采用网格化的功能部件动态地构建适应于特定应用的计算机体系结构,可以实现一种“Architecture On Demand”(体系结构按需定制) 的计算模式。

3.2、资源组织方式与体系结构的动态映射

根据应用的计算模式的不同和运算阶段的不同，可以将网格化功能部件按不同的方式组织起来。通过改变功能单元之间的逻辑关系、互连模式，可以使应用程序看到的计算机系统具有不同的组织方式和程序运行模式，从而使应用程序以最优的方式执行。

例如采用独立的网格化 CPU、Memory 和 Disk 资源,通过动态通过动态映射可实现传统体系结构：

表 1 传统体系结构在 DSAG 下的映射

传统体系结构	功能单元映射
对称多处理 (SMP)	$\{CPU\}_n+\{Mem\}_m+\{Disk\}_k$
大规模并行处理(MPP)	$\{<CPU,Mem>\}_n+\{Disk\}_m$
机群(cluster)	$\{<CPU,Mem,Disk>\}_n$
流水线 (pipeline)	$<CPU>_n+\{Mem\}_m+\{Disk\}_k$
单指令流多数据流(SIMD)	MCPU+$<VCPU,Mem>_n+\{Disk\}_m$

说明：{} 对称关系, <> 顺序/绑定关系, +高速互连 MCPU 管理/控制节点, VCPU 虚拟机

还可以进一步根据需要建立这些基本的组织方式的各种组合模式，可以实现新的体系结构：

- SMP(0-1)+MPP(0-1)+Cluster(0-1)：将机器分割为 SMP、MPP、Cluster 三个区域，分别执行三类应用或一个任务的三类线程。
- 动态 D-SMP、动态 D-MPP、动态 D-Cluster：传统机器资源的动态加载。
- D-SMP(0-1)+D-MPP(0-1)+D-Cluster(0-1)：SMP、MPP、Cluster 三区的动态调整。

3.3、程序设计与运行模式

DSAG 结构的基本特点是计算机体系结构动态、灵活、可配置。改变过去程序（或编译系统）向体系结构靠(即通过优化程序结构来适应体系结构的特点)的局面，在机器硬件层面实现体系结构动态向应用程序靠或称体系结构动态适应应用程序的需求。从而更加有效的利用计算资源。

支持 DSAG 体系结构的高性能计算机可高效运行（无需修改）目前已开发的几类并行程序：如适合于 MPP 机器的基于 MPI 库函数、适合 SMP 机器的基于 Pthread 库函数和适合 Cluster 机群结构的基于 PVM 库函数的三类程序。

为计算机系统（实现 DSAG 体系结构）编制应用程序时，应用科学家较传统并行系统具有更好的直接性和更大的实现灵活性。当他/她设计新算法解决新问题时，可以更加集中于问题本身的并行特征自由设计算法，无需过多考虑机器的特性。当设计完成编制程序时，可调用已存在的各种程序与库（无论是基于共享内存或者是消息传递模式），快速完成程序的开发（并行软件复用更加容易）。

针对 DSAG 机器将传统串行程序的并行化任务（可以是人工或者有自动并行编译完成）将更加有效与简单。寻找已有串行程序中的并行特征（粗、中、细粒度并行）并进行针对机器的有效映射是两个关键环节。DSAG 体系结构至少在后一个环节大大减少任务的复杂性。

目前采用可重构计算的应用领域均有可能是 DSAG 机器的应用范围。对 DSAG 应用特征的研究还需要继续深入下去。

4、支持 DSAG 体系结构的光互连实现及操作系统的考虑

DSAG 体系结构的硬件实现包括两个部分，一部分是可变的硬件结构，而另一部分是控制这种变化的硬件或软件，这种控制器也可以称为元处理器（Meta Processor）。元体系结构基本不变的，而通过元处理器管理可变硬件结构的动态变化。理想情况下，操作系统能够根据应用系统的特点动态的调整体系结构和操作系统，体现以用户为中心的思想。

4.1、光互连实现考虑

光互连是连接 DSAG 系统的核心部分。在 DSAG 系统中，光互连技术必须解决两个关键的问题：直接部件连接和动态拓扑改变。

DSAG 主要由网格化的功能部件组成，要利用分布的网格化部件构建高性能计算机系统，必须保证每个功能部件能够独立的、高速的被访问。这就需要光连接能够直接和功能部件相结合，尽量减少总线、I/O 通道等其他中间环节。基于表面发射激光的 VCSEL 技术可以将激光技术和 VLSI 技术较好的结合起来，为芯片一级直接进行光接收/输出提供了可能。目前 VCSEL 还限于 AsGa 工艺，一旦和 CMOS 工艺结合起来，其应用范围必将迅速发展。

实现 DSAG 的动态特性需要动态改变网格化部件之间的联接关系。传统的集中式电路交换技术受电信号之间交叉干扰和电信号频率的限制，其进一步大容量扩展受到限制。而光传输没有串扰和带宽限制，全光交换的潜力远远超过电交换的极限。目前全光交换技术已经开始在电信系统应用中实验，而基于微机电系统(MEMS)技术的自由空间光互连系统发展将进一步为动态系统互连提供支持。

光互联对 DSAG 的支持在理论上和实验上都已经成熟，目前欠缺的只是工业化规模生产阶段的到来。我们已经开始着手研究相关技术在高性能计算机系统中的应用。

4.2、支持 DSAG 体系结构的操作系统

操作系统作为联系计算机硬件和应用软件的纽带一直与计算机体系结构密切相联。传统计算机操作系统是一种与节点计算机绑定的小而全模式：操作系统只负责本地的资源分配，而不同计算机节点之间的协调则通过应用软件进行。在 DSAG 结构中，组成计算机的功能部件可以独立的上网并提供服务，传统意义上的本地计算机节点已经不存在，相应的本地操作系统也将不复存在。

网格化的计算机将是一个利用动态组合的资源构成的一个全局的计算机系统。相应传统单机操作系统的作业管理、内存管理、文件系统、I/O 处理等等必须在全局的角度重新进行考虑。支持 DSAG 的操作系统特征包括：

- 操作系统功能的分布化：操作系统的功能分解并分布到不同的服务节点上。各个操作系统模块本身也构成一种服务。各种不同的操作系统服务之间通过一定的方式进行动态的关联。各种计算资源不再由操作系统完全控制，而是自身构成独立的具有自我描述能力的网络服务，操作系统模块通过网络协议与各种资源服务器进行交互协商。
- 适合应用模式的按需组合的资源映射：操作系统可以根据应用模式的需要，将来自不同服务设施的各种功能部件组合成用户需要的计算机模式。同时操作系统本身的各种功能模块也可以按需组合。从而实现一种按需定制系统的计算模式。
- 资源的动态申请、分配和组合：网格计算环境下可用的计算资源将不再是固定的。随着应用运行模式的变化，资源随时可以动态的加入和退出。操作系统不仅是在一个任务开始前就完成资源的分配，而是要在任务执行过程中始终保持对资源的有效控制，保证任务能够根据资源的变动而进行规模和策略上的调整，实现子任务动态的分解和组合。
- 实现单一系统映像,隐藏系统结构差异：

网格计算环境下资源的分布更为不均匀，每个独立的功能模块所提供的性能各不相同，例如因距离不同而带来的不同的响应延迟等。操作系统必须为应用程序提供统一的系统映像，包括采用全局统一的虚拟地址空间等技术，避免为应用程序设计带来额外的复杂性。同时，也需要提供必要的接口以便需要时应用程序可以进行特别的优化。

- 自动故障屏蔽和恢复：网格化的计算机系统的规模将远大于现有的计算机系统。但随着系统规模的扩大，系统中发生局部故障的几率也增大。操作系统应该能自动屏蔽局部故障的影响，保证大型任务能够正常进行，包括故障监测、故障隔离、故障迁移、检查点恢复等技术。在功能部件级独立的情况下，传统的基于整机冗余的高可用技术将被细粒度的部件级冗余和检查点等技术取代。从而有效提高整体系统的可用性。

5、围绕 DSAG 体系结构的其他应用系统研究

DSAG 体系结构为高性能计算机系统设计中提出的，但 DSAG 本身的思想还可被应用其他的层次的系统设计中。

基于 Internet 支持 DSAG 体系结构的虚拟超级计算机研制(Virtual_HPC)：Internet 上拥有的数以亿计个人计算机资源，这些资源的聚合计算能力是无限的。但是目前这些计算资源的分散性和多样性使得很难通过某种规范来统一各种资源的调度和使用接口。可以通过定义虚拟机隐藏各种不同的计算资源的差异和细节，形成一种虚拟的网格化功能部件-VM。利用这些简化的、统一的虚拟资源来动态构建虚拟的超级计算机系统。

基于 DSAG 的 FPGA 芯片系统设计：结合 DSAG 的思想，可以将 FPGA 的可重构处理单元(RPU)的结构按功能单元进行划分和动态组织，从减少重新定制开发 RPU 的周期，甚至最终打破原有高性能计算中硬件/软件的分界线。

基于 DSAG 的 MultitheadingCPU 设计：现有 CPU 的发展趋势是在一个芯片内嵌入多个多核心和多线程。按照 DSAG 的思想，在芯片内部的资源组织模式不应该是固定的。可以加入内部模块连接网络和动态的资源划分功能，使得 CPU 内部组织可以根据特定的需求进行调整，便于实现功能流水、SIMD 等分工合作的模式。

6 总结

高性能计算机系统的结构正处在一个变革的时期。减少人机差距是我们追求的长远目标之一，高性能计算机也不例外。光互联技术和网格化技术是未来二十年重要的技术方向，将对高性能计算机系统的结构带来深刻的影响。网格化的动态自组织结构 DSAG 是我们综合相关思想与技术提出的一种新型的计算机体系结构。DSAG 具有动态、灵活、可配置等优点，不仅可以包含传统的 SMP、MPP、Cluster 等体系结构，而且可以实现一种新的体系结构按需定制的应用模式。DSAG 的思想也可以用于其他层次的系统设计。

DSAG 体系结构是我们创新的一种尝试，机器模型的研究尚需完善。目前器件级实验、模拟系统和操作系统的研究正在进行中。还有许多问题有待探讨和验证。欢迎感兴趣的同行与我们交流与合作。

参考文献

[1] “Computer architecture is all about interconnect”,William J. Dally, Keynote in Workshop on network processors along with HPCA 2002, February 3, 2002, Boston, Massachusettes

[2] David Patterson, Aron Brown etc. “Recover Oriented Computing (ROC): Motivation, Definition, Techniques, and Case Studies”, Computer Science Technical Report UCB/CSD-02-1175 U.C.Berkeley March 15,2002

[3] “Fifty Years of Computing at LLNL as a Lens to the Future”, Dona L. Crawford, 17th

International Supercomputer conference ,ISC2002, June 22, 2002
[4] K. Compton, S. Hauck, "Reconfigurable Computing: A Survey of Systems and Software", ACM Computing Surveys, Vol. 34, No. 2. pp. 171-210. June 2002
[5] I. Foster, C. Kesselman and S. Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. International Journal of High Performance Computing Applications, Vol 15, P.200-222, 2001.
[6] Neil Savage, “Linking with light”, IEEE Spectrum,Volumn 39, Number 8, PP32-36 August 2002