Dagger：一种散耦合的网格计算机体系结构

孙凝晖　樊建平
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
Email: {snh, fan}@ict.ac.cn

[摘要] 计算机系统面临着网格计算对其提出的资源共享、协同计算和应用服务的挑战，现有的计算机系统的体系结构不能够很好地适应这种需求。本文提出一种基于散耦合思想的网格计算机系统的体系结构，称之为 Dagger，采用八个基本概念构成网格计算机，从多个角度为网格应用提供支持。

[关键字] 网格计算机，体系结构

1.引言

网格计算（Grid Computing）被人们认为是未来 20 年计算机发展的技术方向。网格最早来源于人们对高性能计算（High Performance Computing）的追求，将 LAN/WAN 上的若干高性能计算机合成一个虚拟的大计算机，解决一个大问题或让许多用户共享这些昂贵的计算、存储、设备资源[2]。随后，人们将网格思想扩大化，用以解决 Web 服务、商业计算、计算机应用等领域存在的问题，我们认为，网格计算的主要目的是[1]：

- 资源共享：网格让计算机资源，包括计算能力、存储容量、大型设备、重要应用等，能有效地被许多用户所共享，提高资源的利用率和效能（Productivity）；
- 互连互通：网格让计算机在资源层面上互连，Internet 解决的是计算机之间的数据连接，Web 解决的是计算机之间的信息连接，需要在计算、存储、设备、应用、数据、信息、甚至知识之间建立连接；同时要实现互通，即协同计算，解决诸如信息孤岛、应用分布的问题；
- 应用服务：网格让计算机应用成为一种服务，像电信、电视、供水、供电这样的资源服务（Utility），这将改变应用的开发、部署、使用、管理、甚至销售模式；

总之，我们认为，网格研究的核心思想可描述为：基于网络，让现在位于一台计算机内的各种部件和资源都能独立上网(格)，共享资源、管理和服务，开创 WWW 后的 GGG（Great Global Grid）新型商业模式[1]。

面对网格计算，计算机系统的体系结构的挑战是什么？我们要区分需求，一种是“ XX Grid ”，如 Computational Grid（计算网格），Data Grid（数据网格），Access Grid（访问网格），Bioinfomacs Grid（生物信息应用网格），Web Service（Web 服务商业网格），这些需求的目的是基于资源共享解决大问题、提高应用效能、开拓新型应用，网格对计算机体系结构的要求较小，计算机系统被看成是网格的一个节点部件，重点在于应用基础软件。但是，其中一些软硬件部件如果加以改造，将能更好地适应网格的要求，如 GGF Grid Monitoring Architecture (GMA)，MPI-G2，支持网格的机群操作系统[3]等。

另一种是“ Grid XX ”，如 Grid Computer（网格计算机），Grid Storage（网格存储），Grid OS（网格操作系统），Grid Application（网格应用），Grid Commercial Computing（网格商业计算），这些需求的目的是用网格研究的思想解决存在的问题，使这些“ XX ”在网格环境下更好地服务于用户，这一思想与计算机发展过程中许多技术进步相同，如 Parallel Computer（并行计算机），Distributed OS（分布式计算机），Network Application（网络应用）。这种情况下，网格对计算机体系结构产生影响（impact）。

我们认为，网格计算的研究包括四个层次：

- 计算机系统：改进计算机系统，尤其是高性能计算机，的体系结构，研究将各种硬软件部件、资源变成网格零件，和组合成高效能计算机系统的方法与技术，提高系统的生产率，可扩展性，可管理性。
- 系统软件：改进操作系统的结构，适应网格零件即插即用（plug&play），应用和管理端到端的安全的要求，提供系统级的网格操作系统，管理网格资源。
- 应用基础软件：提供开发更加有效的传统应用、和新型应用所需的平台和工具，如 OGSA，Web Service，SUN ONE，VEGA Grid（应用级的网格操作系统）[1]，Globus，都属于这一范畴，研究如何开发、部署、使用、管理一个或多个应用。
- 网格应用：如 SETI@Home，网格应用应与非网格应用有一些本征（native）上的不同，例如，我们对网格高性能计算应用的描述是：首先是网络应用，即一个执行映像能支持多个客户端，用户的所有操作都在客户端完成，只能看到逻辑资源，使用时，只提供与应用相关的需求，如问题大小、输入数据、输出文件、期望运行时间、期望付出价格等。

计算机系统，尤其是高性能计算机，面临着网格计算对其提出的资源共享、协同计算和应用服务的挑战，现有的计算机系统的体系结构不能够很好地适应这种需求。第二节阐述当前计算机系统的问题和 Dagger 体系结构的设计思想，第三节解释了 8 个网格计算机的部件，第四节举例说明如何用 Dagger 的网格部件构造计算机系统，最后给出了一些讨论。

2．Dagger 设计思想

本文关于体系结构的研究动机是改进现有计算机系统的体系结构的若干缺陷，提出网格计算机的一种体系结构，使它能更好地构造在网格环境下的高性能计算机。

现在，计算机系统呈现的是各种硬件和系统软件，如服务器、磁盘阵列、显示器、交换机、操作系统、浏览器等，对于网格环境存在以下缺陷：

- 利用率低：“通用”计算机系统的体系结构在提供标准化和批量的同时，也导致使用效率和利用率低；几乎所有计算机外围设备都需要强大的通用 CPU 的驱动，通用 CPU 和系统的设计要满足各种不同应用的要求，使得计算机必须按照摩尔定律不断地提高 CPU 主频，不管是否需要，计算机系统必须连接 KVM（键盘/显示/鼠标）、存储、网络等接口；
- 静态资源：计算机系统的资源，如存储、管理部件、操作系统、应用，都是静态绑定的，称之为配置和安装，由此导致资源分散，使用管理不便；
- 可信安全：缺乏端到端的安全体系，系统的可信性在对使用者和外部资源未知的情况下去保证，如防火墙、login/passwd；

网格计算机体系结构的主要设计思想是，在冯-诺伊曼体系结构不变的前提下，将计算机系统中资源的进行重组，从资源、功能、服务的角度来定义网格计算机的部件。这种体系结构我们称之为 Dagger：De-coupled Architecture with Grid-key and Grid Enabling Resource，即具有“网格钥匙”和“网格使能资源”特征的一种散耦合的体系结构。下面的三个主要特征解决上述的三个缺陷：

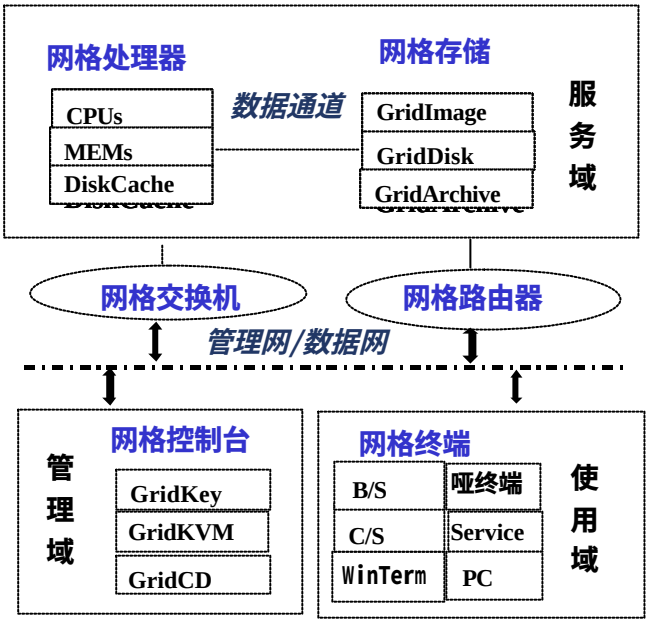
- De-coupled：即散耦合；将计算机系统拆分成可组合的网格部件和网格零件，进行资源重组，以提高资源使用效率、利用率、可管理性；
- Dynamic Deployment：即动态部署；将资源与资源的关系在创建应用运行环境时绑定，也称“late binding”；可以在三个层面将资源变成网格使能的，一是硬件，支

持硬件部件的功能动态组合和 plug&play；二是操作系统服务，支持操作系统映像、服务和核心模块的动态部署；三是应用服务，支持应用的服务化和协同；

- Dependable Security：即可信安全，通过在客户端增加表明身份、基于角色的网格钥匙，提高系统安全的可信度。

3.网格计算机部件

Dagger 网格计算机的体系结构框架如下图所示，由 6 个网格部件和 2 个网格概念组成。有些网格部件又由一些网格零件组成，所谓网格零件即在网格环境下可见的功能单元，它们具有明显的区别于普通计算机系统零件的网格特征。



设计网格部件的动机一是“拆分”，将传统的服务器拆分成计算部件、存储部件、管理部件，即“网格处理器”(Grid Processor)、“网格存储”(Grid Storage)、“网格控制台”(Grid Console)。带来的好处有：

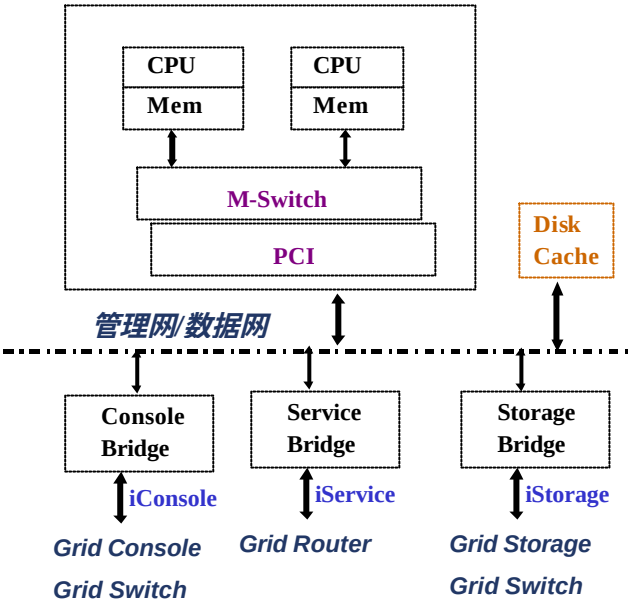
- 提高计算部件的资源利用率，不用被慢速设备所干扰；
- 计算部件的系统环境按需动态可变，支持硬件的动态部署；
- 管理部件在物理上靠近用户，方便管理；
- 存储资源、管理资源可被整个网格共享；
- 计算部件、存储单元可热插拔（plug&play）；

网格处理器由四个零件组成：

- CPU：传统的处理器，所有的计算能力都用在处理应用上；
- 内存：如果未来基于光纤的远程访问满足性能要求的话，内存也可以成为独立的网格部件；
- 连接件：如下图所示，网格处理器通过三个连接件（Connection）与控制、服务、存储部件相连，或通过数据交换设备实现共享，我们将这三种协议称为 iConsole、iService、iStorage，它们可以是硬件协议，也可以是软件协议；
- 硬盘缓存：本地硬盘不再存放永久数据，而成为硬盘缓存（DiskCache），用于网格本地临时存储，如 OS 映像缓存、操作系统 Swap 区、log 区、配置信息、应用临时数据（scratch）；这一改变是为了支持动态部署，同时避免远程访问的延迟，类似

于 CPU Cache 的作用；

- 在系统软件上，网格处理器需要一种类似于 BIOS 的网格软件，实现资源的动态部署和组合。



网格存储由三种网格零件分别组成：

- 网格映像（GridImage）：用于存放操作系统映像，动态部署到网格处理器上，并能自动设置除引导操作系统外的应用的运行环境，网格处理器只有操作系统的内存映像或内存映像的快照（snapshot）；
- 网格硬盘（GridDisk）：将网格连接的存储设备，动态地虚拟成网络存储设备，动态地为网格处理器提供数据和文件服务，支持海量存储设备和海量文件的虚拟管理，类似于 IP SAN 或 InfiniBand 存储的作用；
- 网格备份（GridArchive）：在网格内提供数据备份服务；
- 网格存储与网格处理器之间不限定物理连接方式，如 SCSI、FC、IP、iSCSI，或专有协议；

网格控制台实现对全网格的管理（administration）、控制（control）、监视（monitor），连接与管理相关的慢速设备，如传统南桥设备，至少包括三种网格零件：

- 网格显示器（GridKVM）：远程共享 KVM；
- 网格光驱（GridCD）：远程共享 CDROM，或提供其它南桥设备远程服务；
- 网格钥匙（GridKey）：体现可信安全的网格零件，它用在所有需要增强安全的管理域和使用域的部件上，包括网格用户身份认证，记费，数据加密，网格应用状态记录（session）等；
- 网格控制台可在全网格内内共享，在系统软件上，需要一种类似于 Firmware 的网格软件，实现网格部件的探查，和网格系统软件的运行；

设计网格部件的动机二是“互连”，一方面由拆分导致部件分离，难点在于互连；另一方面网格的重要思想就是互连互通。在互连上，我们提出“IntraGrid”、“InterGrid”两个概念，和网格交换机（Grid Switch），网格路由器（Grid Router），网格终端（Grid Terminal）三个网格部件。

IntraGrid（内部网格）实现在网格资源层的共享，用于连接多个服务域，使其成为一台网格计算机，主要功能是共享资源、应用集成、和资源的 Plug&Play。

InterGrid（外部网格）实现在网格应用层的共享，连接服务域和客户端，提供应用服务，和应用的 Plug&Play。IntraGrid 和 InterGrid 与 Intranet 和 Internet 提法类似，但不对应，同时 IntraGrid 和 InterGrid 相互独立，也可以共存。

网格交换机是实现 IntraGrid 的关键部件，通过资源层的数据交换，实现 IntraGrid 内资源共享和应用协同，主要功能包括：

- 连接共享的网格处理器
- 连接共享的网格存储
- 连接共享的网格控制台
- 可有管理网、数据网多种网格交换机，如 IP 交换机，交换网格元信息；iConsole 交换机交换管理信息，实现管理、控制、监视功能；iStorage 交换机交换存储数据，iService 交换机交换服务数据；

网格路由器是实现 InterGrid 内的应用服务和远地协同的关键部件，主要功能包括：

- 连接网格终端；
- 连接网格处理器；
- 网格部件间的网络连接，如网格服务器之间的路由，应用负载平衡器；从网格终端到网格服务器的防火墙，代理，安全认证，GSML（网格服务标记语言）[1]服务应用路由；
- 与传统路由器的不同是它是服务级的路由；

网格终端是用户使用网格计算机的客户端设备，应具有三大特性，即操作系统和应用零部署（Zero），用户可移动（Mobile），和应用会话可重构（Session），用户、应用、平台是动态绑定的。现在的 PC 做不到这一点，典型的场景是，我们先必须安装操作系统、浏览器或其它 Client 软件，用户不可以任选一台 PC 使用需要的应用，从一台 PC 转到另一台 PC，应用会话全部丢失。网格终端应能支持动态切换六种应用模式：Browser/Server，Client/Server，Windows 终端，哑终端，GSML 服务，Legacy PC。在系统软件上，需要一种类似于 Firmware 的网格软件，动态部署 Browser，操作系统，应用，WTerm 模拟器，KVM 模拟器等。

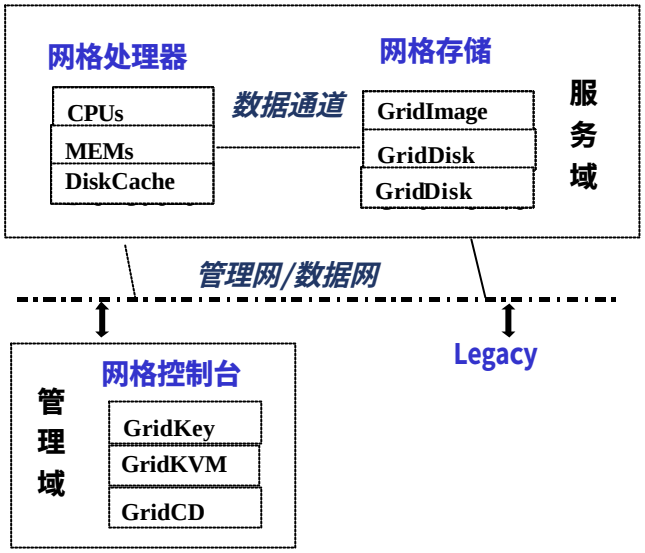
4.构成网格计算机系统

基于 Dagger 网格计算机体系结构和网格部件可以构造不同的融合传统体系结构的系统，使之具有新颖的特性。下面以普通的 PC 服务器和机群（Cluster）为例说明网格服务器和网格机群的特性。

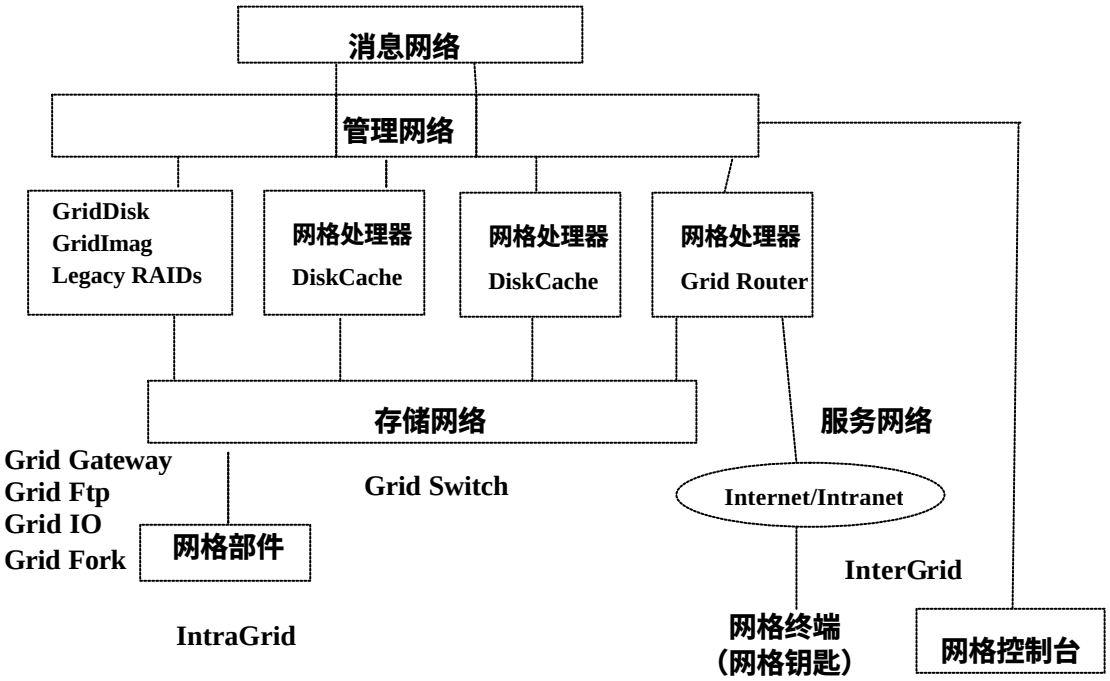
下图表示一个简单的网格计算机系统，在原来的 PC 服务器上进行 4 点改进：

- GridImage：操作系统映像通过 Ethernet 动态部署，可以同时支持多种包括操作系统在内的应用环境的快速改变；
- GridDisk：存储以 IP SAN 的形式提供，将设备和管理虚拟化，提供设备层的在线动态扩展功能；
- DiskCache：内置硬盘成为缓存，可以任意替换，提高了可管理性；
- Grid Console：将管理部件从服务端端剥离，管理员除机械动作外（如换热插拔硬盘），不需要走近服务器，提高了可管理性；由于 GridKey 的“管理角色”的支持，

提高了安全性；



曙光超级服务器使用机群体系结构[4]，机群是现在流行的成熟的高性能计算机的体系结构，我们按网格计算机的思想对它进行了重构，试图提升机群的技术，称之为网格机群系统，如下图所示。



机群的主要部件成为网格计算机部件（Grid-enabling Components），节点不再是能独立工作的计算机系统，即节点在硬件上是不完整的服务器，运行不完整的操作系统映像，网格机群之间在管理上、应用上能够实现 Plug&Play。资源的改变包括：

- 计算节点：成为网格处理器，可以任意替换；
- 存储：分成本地磁盘缓存、OS 映像、和共享数据区；
- 网格控制台：集中了 KVM、CDROM、Panel（节点前面板）等管理功能；
- 互连网络：分成消息网络、管理网络、存储网络、服务网络；通过网格交换机实现

资源的 Plug&Play；

- 资源动态部署：OS 能动态部署，使系统的运行环境动态改变，如 Cluster/Linux、Farm/Windows、MPP/LWK（轻核心）；OS 模块、OS 服务、存储能动态部署，应用能动态部署
- 机群系统之间资源共享：通过网格交换机共享存储，和网格控制台；
- 客户端：成为网格终端，通过网格路由器接入，网格钥匙提供认证。

5.结论

针对网格计算对计算机系统的要求，我们提出了一种散耦合的网格计算机体系结构，其本质上是一种计算机系统组织的方法，没有改变应用的编程模式，这一点与 SMP、MPP、NUMA、Cluster 不同。在国家 863 重大项目“面向网格的高性能计算机－曙光 4000”的设计中部分采用它的设计思想，开发了操作系统映像加载器、网格钥匙、网格控制台、网格显示器等若干网格零件，并正在研制一台验证这一思想的网格服务器。

在网格计算机体系结构方面，研究刚刚开始，还有许多问题待探讨和验证。在网格计算机上层的网格系统软件、应用基础软件的体系结构尚需要大量研究工作：

- GridFirmware：应有一种“网格固件”软件，支持动态部署硬件、部署操作系统、引导操作系统，BIOS 设置，网格计算机配置信息的设置，运行网格零件控制软件，进行网格资源的监视，网格部件的探查（probe）和例外通知（exception）等功能。
- GridOS：是面向网格的操作系统功能的重组，即系统层的“网格操作系统”，支持基于管理员角色的 OS 管理，并能动态部署到网格控制台运行；基于网格用户角色的 OS 使用功能，并动态部署到网格终端运行；支持 GridSyscall（网格系统调用），提供远地服务；支持管理远地资源的网格驱动，如 iConsole、iStorage、iService 驱动。
- GridDeploy：应有一种“网格部署器”软件，支持根据网格计算机配置信息和当前应用，动态部署 OS 模块、OS 服务、管理软件、应用，支持生成 OS 快照。
- 网格应用基础软件体系结构：这里存在许多概念，网格服务标记语言（GSML），网格社区（Community），网程（Grip）[1]，Globus/OGSA，XML/Soap/UDDI，需要一种体系结构整和在一起,并提出对网格计算机体系结构的要求,如网格路由器。

参考文献：

[1] 徐志伟，织女星网格的总体思路，计算所技术报告，织女星网格文档 VGD-1，2001 年 11 月 3 日。

[2] 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.

[3] 孙凝晖，刘淘英，支持网格的机群操作系统的设计，计算机研究与发展学报，第 39 卷，第 8 期，2002 年 8 月。

[4] 孙凝晖，孟丹，曙光 3000 超级服务器设计的关键问题研究，计算机学报，Vol.25, No.11, 2002 年 11 月。

Dagger: A Decoupled Architecture of Grid Computer

Sun Ning-hui, Fan Jian-ping

Institute of Computing Technology, Chinese Academy of Sciences

National Research Center for Intelligent Computing Systems
E-mail: {snh, fan}@ict.ac.cn

Abstract：Grid computing issues new challenges as resource sharing, coordinate computing and application service to computer system. The current architecture of computer system can not meet the requirements well. In this paper, we propose a decoupling-based architecture of grid computer, called Dagger, which adopts eight conceptions to organize the grid computer system, and supports the grid applications in some aspects.

Keywords: Grid Computer, System Architecture