# 面向科学计算的网格环境

迟学斌 肖海力 王小宁 曹荣强 卢莎莎 张宏海

（中国科学院计算机网络信息中心超级计算中心 北京 100190）

**摘 要** 为了充分整合分布的高性能计算资源，本文提出一种面向科学计算的网格环境，旨在形成一个可统一管理和运行维护的虚拟的超级计算机资源，面向用户提供统一、易用、可靠的科学计算服务。面向科学计算的网格环境通过轻量级网格中间件SCE汇聚资源，支持作业的全局调度、数据的统一管理视图，面向用户提供命令行和网格门户两种使用方式，并提供编程接口供专业社区和学科平台二次开发使用，满足不同层次的用户需求。目前，面向科学计算的网格环境已经在中国科学院超级计算环境（ScGrid）中得到应用和用户认可。

**关键词** 高性能计算；网格环境；网格中间件

## Scientific Computing Grid and SCE Middleware

CHI Xue-bin XIAO Hai-li WANG Xiao-ning CAO Rong-qiang LU Sha-sha ZHANG Hong-hai

( *Supercomputing Center of Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190 )

**Abstract** This paper introduces the scientific computing grid, ScGrid, and it's middleware SCE. ScGrid is built as one virtual supercomputer, integrating computing resource from more than 30 institutes. It provides unified, easy to use and reliable scientific computing services. SCE is a lightweight grid middleware, which supports global job scheduling and unified data view. It provides multiple user interfaces including command line, grid portal and APIs. At present, ScGrid has been very successfully used in Chinese Academy of Sciences and widely accepted by more than 200 users.

**Keywords** high performance computing; grid computing; grid middleware

## 1 引 言

计算科学已经成为继理论分析和实验观察之后的第三种科研工具，它在科研活动中的重要性日渐凸显，高性能计算及其应用水平也随之成为国家高科技发展和综合实力的重要指标。最近几年，高性能计算相关的新技术层出不穷，硬件成本不断降低，可用资源日益丰富。国内的相关软、硬件研究也紧追国际先进水平，从十一五初在北京和上海分别投入使用的联想深腾7000、曙光5000A国产百万亿次超级计算机，到天津、深圳、济南等地相继发布的天河-1A、曙光6000、神威蓝光等国产千万亿次超级计算机，相信很快会出现万万亿次甚至更大规模的计算资源。同时，一些高校和研究所也不断购买一些中小规模的计算集群。这些计算资源广泛分布在国内不同城市，依托于不同的科研教育机构和超级计算中心。

高性能计算可用资源日益丰富，带来的一个首要问题是如何充分、有效地组织和利用这些资源。由于科研活动的周期性，应用领域和地域等特点，每个高性能计算资源的利用程度不同，甚至出现有些计算资源作业排队现象很严重，而其他资源闲置的现象。资源整体利用不充分，造成一定程度的浪费。

另外，科研人员总是希望能及时得到优质、不间断的计算服务。面对诸多资源，如果只使用一台高性能计算机，当资源紧张引起作业排队时间过长，或者

**迟学斌**，研究员，博士生导师，研究方向为高性能计算方法与软件。E-mail: chi@sccas.cn。**肖海力**，高级工程师，研究方向为高性能计算与网格计算。E-mail: haili@sccas.cn。**王小宁**，副研究员，研究方向为高性能计算与网格计算。**曹荣强**，助理研究员，研究方向为高性能计算与网格计算。**卢莎莎**，助理工程师，研究方向为高性能计算与网格计算。**张宏海**，副研究员，研究方向为高性能计算与网格计算。

遇到日常维护、故障停机等不可用情况，都会严重影响科研工作进展。如果使用多台高性能计算机，可以在一定程度缓解上述问题。但是会造成使用上的不便：由于机器不同，账号申请过程、登录流程不同；提交计算任务的方法和流程也各有不同；在计算中遇到问题时获取技术支持的渠道和响应速度也各不相同；付费方式、账单交付也有所不同。这些问题都会对科研人员专注于科学问题产生极大的困扰。

同时，作为一门由数学、计算机科学、应用科学等多学科交叉的学科，同时也是备受关注的热点研究领域，高性能计算应用吸引了众多年轻研究者的关注与参与。如何降低初学者使用高性能计算的门槛，提高计算环境的易用性是一个非常急迫的问题。长期以来，高性能计算机难以使用一直都是制约其应用与发展的首要问题，美国NITRD委员会认为，高性能计算存在的主要问题之首，就是“难以使用”（hard to use）；而NSF也在制造高性能计算机的主要障碍中将“可用性”（availability）列在第一位，由此可见，提高高性能计算环境的用户友好性对推动高性能计算应用的发展至关重要。

因此，将这些分布的高性能计算资源汇聚在一起，统一运行、统一管理、统一调度，有利于科研活动的顺利进展，并且可以使高性能计算资源得到最大程度的利用。网格技术[1,2]经过最近十几年的发展，已经成为资源整合和汇聚方面较为成熟的技术之一。目前已经有一些网格中间件对分散的高性能计算机进行有效地组织和管理，包括美国的Globus、欧洲的gLite和中国的GOS，利用这些网格中间件建成了TeraGrid、WLCG、CNGrid等网格环境。然而这些软件，以及通过这些软件建立的网格环境往往存在不足。比如软件可靠性和可适应性不足、安装部署和运行维护成本高、不支持多种使用方式、对科研用户的使用习惯改变太大等问题。

基于以上的现状和分析，本文提出了构建面向科学计算的网格环境，以自主研发的网格中间件SCE，将分布在不同地理位置，自治管理的高性能计算资源整合和汇聚在一起，形成一个可统一管理和运行维护的虚拟的高性能计算机资源，面向用户提供统一、易用、可靠的科学计算服务。

## 2　针对科学计算特点的网格系统设计

### 2.1　要点

构建面向科学计算的网格环境，需要考虑以下几方面问题：

（1）面向用户需求提供易用的计算环境；

（2）有效汇聚资源是网格中间件的基本和核心任务；

（3）从实际情况出发减少网格软件的运维成本。

面对分布在不同地理位置的各种的高性能计算机，其计算机体系结构不同、操作系统不同、作业管理系统不同、应用软件版本和安装路径不同，种种不同给用户在使用方面造成一定的困扰，因此面向科学计算的网格环境用户界面易用性的最基本要求是提供统一的使用方式，允许用户以相同的方式使用网格环境中的不同计算资源。另外，在科学计算领域，不同学科软件有着不同的交互性要求，不同领域的用户对计算资源操作的接受和熟练程度也各有不同，因此面向科学计算的网格环境的用户界面易用性的另一个要求是提供多种不同的使用方式，用于满足不同需求的用户。

同时，如何保证用户的作业和数据在不同的计算资源之间被有效的调度和管理，也是面向科学计算的网格环境需要考虑的一个重要问题。在面向科学计算的网格环境中，高性能计算设备是自主管理的，它们往往有自己的用户，因此作业调度需要保证这些资源原有用户的使用优先级，在资源相对空闲的情况下接受来自其他节点的作业请求。在科学计算领域，计算和数据是密不可分的，因此需要提供一种简单有效的方式，管理分布在不同计算资源的用户数据。

低运行维护成本主要指：（1）对机群系统管理员在计算机管理方面的知识门槛和工作量要求低，这样不会给科学应用领域的研究人员增加额外的管理和运行维护的负担；（2）网格软件占用的机群资源少，这样可以保证有限的计算资源被最大化的利用于科学计算；（3）安装一套接入软件，可以方便接入多个计算集群。

基于以上分析，面向科学计算的网格环境要求网格中间件具有如下特点：

（1）稳定可靠

可持续稳定运行是对生产性软件的最基本要求，也是为用户提供优质的科学计算服务的最基本保障。高可靠的使用环境可以保证单个计算资源在停机维护期间或者故障期间不影响用户的科学研究活动，有效地提高科研人员的工作效率。

（2）轻量级软件

网格中间件部署在机群计算资源上模块或服务最少，占用的网络端口数量最少，并且保证其运行时占用的CPU和内存资源最少，易于大规模推广和快速部署。

（3）全局作业调度和数据管理

网格中间件支持作业的全局调度，允许作业被调度到任意一个可用的计算资源上执行，保证各计算资源设备负载均衡。同时，网格中间件提供全局统一的数据视图，方便用户对作业数据的管理维护。

（4）易用的用户界面

网格中间件要求屏蔽计算资源的各种异构性，提供统一的用户使用界面，并且面向不同需求的用户，提供命令行、网页门户和编程接口多种使用方式。命令行主要适用于传统的高性能计算用户，交互性较强；网页门户为成熟、非交互性的软件提供了易用的作业提交界面，降低用户的使用门槛；编程接口方式用于为专业社区和学科平台建设进行二次开发。

## 2.2 关键技术

本文基于以上对面向科学计算的网格环境的问题分析，设计并实现了网格中间件SCE，其体系结构图如下：

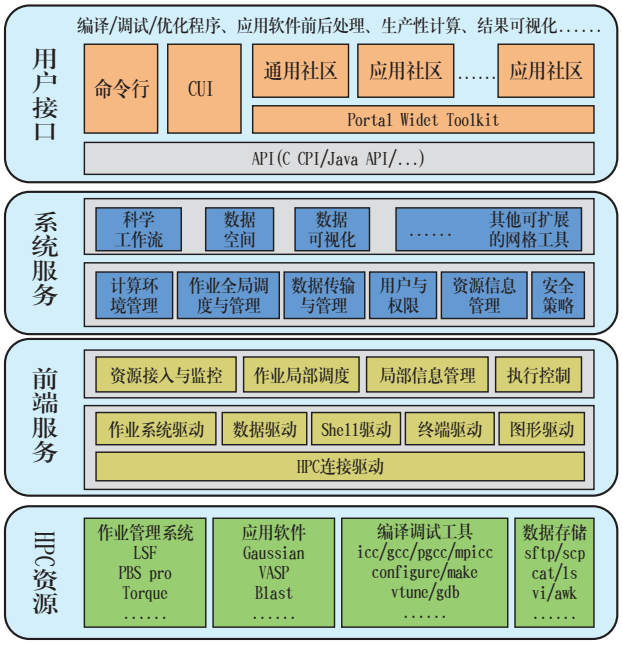


图1 SCE软件体系结构

前端服务FS主要用于资源接入与监控，作业局部调度，局部信息管理，以及一些计算资源的执行控制。通过定义高性能计算资源的各种驱动，可以有效地屏蔽不同的高性能计算资源在操作和信息方面的异构，以及连接方式上的差异。

系统服务层CS提供了用户使用超级计算环境所必需的最基本功能以及若干扩展功能。最基本功能主要包括作业全局调度与管理服务、数据传输与管理服务、用户与权限服务、资源信息管理服务、安全策略以及计算环境管理。通过对基本功能的组合，面向特定领域特定用户提供各类扩展的网格工具，比如科学工作流、数据空间、数据可视化等。

在前端服务FS下层分布着各种高性能计算资源，可分布在不同地理位置，并且自主管理。

### 2.2.1 全局作业调度与管理

SCE系统中作业采用二级调度机制。用户从客户机提交的所有作业，都将被提交到系统服务CS，由CS根据整个环境的使用情况以前端服务FS为单位进行调度。而FS收到作业请求之后则是根据其接入机群的情况，在局部范围内进行二次调度。

（1）系统服务CS端调度

整个SCE系统对于用户来说是相当于一台高性能计算机，接受用户作业提交请求，允许用户查看作业运行状况，当计算完成之后将计算结果放到指定位置提供给用户下载。用户完全不需要关心中间的处理过程。

在用户给定参数提交作业之后，客户端程序会自动生成相应的作业描述JSDL文件，并提交至系统服务CS处理，同时获取作业ID号。系统服务CS的第一次调度基本流程如下：

步骤1 当收到作业提交请求时，生成一个新的作业ID，同时，插入一条新的作业记录到作业信息库中；

步骤2 当收到作业的JSDL文件之后，解析JSDL作业描述，将其中关键信息填入到作业信息库中，将作业相关的文件存储在CS自带的可靠存储上，同时将作业状态置为NEW；

步骤3 系统服务CS定时访问作业信息数据库，根据作业提交的先后顺序和预设的调度策略查看并决定作业是否被调度。当满足条件时则将作业调度到相应的前端服务FS进一步处理，同时将作业状态设置为SCHEDULED；

步骤4 更新作业信息库中的资源项，同时根据系统的资源选择重新生成JSDL作业描述文件；

步骤5 将新产生的JSDL提交至选定的前端服务FS上进行二次调度，并处理响应结果。

（2）前端服务FS端调度

当前端服务FS收到来自系统服务CS的作业调度请求之后，完成作业的局部调度过程，具体处理流程如下：

步骤1 接收中央服务器发来的作业提交请求，获取该作业的ID和JSDL文件；

步骤2 根据作业的资源需求以及当前的调度策略，选定计算所在的机群资源和队列信息；

步骤3 通过选定机群资源配置的作业管理驱动，根据目标机群配置的作业管理系统和连接方式形成相应的作业提交脚本；

步骤4 根据作业ID，从系统服务CS处获取作业相关的输入文件，并传输到目标机群资源的工作目录中；

步骤5 根据目标机群资源配置的接方式，连接机群完成作业提交的准备工作，以及作业提交的请求，同时更新作业状态为PENDING。

SCE的二级调度机制，既保证了全局的调度机制，又给予局部管理范围一定的调度灵活性，可以很好的适应层次化管理需求。

SCE系统支持的作业全生命周期管理包括：

（1）作业提交

系统提供接口支持用户提交作业，用户在提交作业时只需选定作业相关的参数，比如应用名字，应用执行参数，作业执行估计时间长，并行规模等，无需关心具体应用的安装位置，需要设置的环境等细节。用户在提交作业时可以选择由系统自动调度选择资源，也可以自己指定计算的目标机群和队列名称。

（2）作业状态查看

SCE系统提供给用户一个全局统一的作业视图，这个视图给出了用户在整个系统中所提交的所有作业的信息和运行情况，其中作业状态通过系统定时汇报机制得到更新。

（3）作业计算中间文件查看

很多用户需要在计算过程中查看一些中间结果文件，以便及时判断作业是否执行出错，是否有必要终止作业而重新调整参数，因此SCE系统提供了接口支持用户查看作业计算过程中，在目标机群工作目录下产生的文件列表以及文件内容。

（4）作业后处理

作业运行完毕后，前端服务FS会发现作业运行完毕的状态，然后根据应用的描述，对作业结果进行相应的后处理，比如执行垃圾文件清理或者将一些大数据文件打包。

（5）作业终止

SCE系统提供了作业终止的接口供用户在发现某些问题之后及时终止作业的执行，避免不必要的资源浪费。该功能将最终被转换到目标机群上完成相应的终止动作。

2.2.2　文件管理

在SCE系统中，用户的文件数据是分布在网格家目录（HOME）、系统软件（CS/FS），以及计算目标机群资源的存储上的。其中系统软件存储的数据操作是系统软件行为，不显式由用户管理。机群存储的数据可认为是用户计算所需，并非长久存储需求。用户可以使用常用的Linux文件管理方式显式的管理网格家目录的文件数据，并通过SCE系统的接口查看并管理机群存储上的文件。

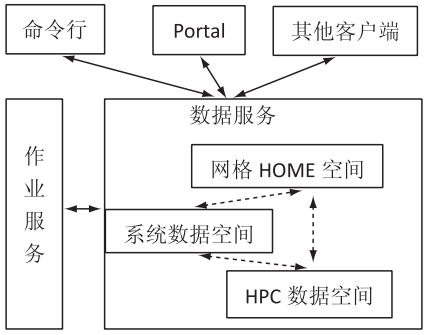


图2　SCE文件管理

（1）网格家空间

网格家空间建立在独立的存储设备上，用于存放用户数据，由用户自主管理。网格家空间在登录SCE命令行之后可直接访问，提供的操作主要是传统的Linux文件操作，包括文件上传、下载、查看、新建、删除、压缩等。

（2）系统数据空间

系统数据空间主要分布在系统服务CS和前端服务FS所挂载的存储上。CS主要存放用于作业调度和用户下载的作业文件，提供持久可靠的存储特性；FS主要存放用于作业提交产生的各种中间文件。系统数据空间中的文件管理由SCE系统根据作业情况自动管理，用户在提交作业时如果显式指定了需要下载的作业文件，可在作业执行结束之后通过系统数据空间提供的接口执行下载操作。

（3）机群数据空间

机群数据空间由各个机群的用户家目录数据空间构成，分布在各个机群的存储资源上，用于存放计算所需的输入数据和计算过程中产生的数据。根据用户的使用方式不同，机群数据目录的组织方式也不同。对于SCE命令行用户，机群家目录的组织与用户的网格家目录组织保持一致。对于计算门户用户，作业目录的位置由SCE系统自动确定，根据作业号组织。

机群数据空间的文件仅限于计算所需，并不提供

永久存储，对于超过一定期限（三个月）的数据文件，系统管理员具有删除的权利，因此网格用户无需对机群数据空间中的数据进行显式管理。

2.2.3 资源信息管理

SCE系统的资源信息管理主要采用资源主动汇报机制。对于静态资源信息，在机群资源接入的同时实现信息的接入，对于动态资源信息，根据信息变更频率，采用资源端定期汇报并更新系统信息的机制。监控的资源信息主要包括机群资源的连接信息、作业管理系统配置、队列信息、作业信息、应用安装信息以及用户映射信息。

（1）静态信息接入

静态信息主要包括机群的访问地址、连接协议、端口、访问账号、用户映射等，以及作业管理系统类型。这些信息相对固定，因此在机群资源接入的同时录入SCE系统。目前SCE系统支持的作业管理系统包括LSF、PBS和TORQUE，并为每一种作业管理系统配置了相应的驱动模块，用于提交作业、查看作业状态、删除作业、获取队列信息等。

（2）队列/作业信息汇报

队列信息和作业信息主要是通过作业管理系统的驱动模块获取的，由于队列的排队/运行信息，以及作业状态信息是动态变化的，因此这部分信息会有SCE系统定时自动获取会更新在系统中，以便用户可以查看到最新的资源状态，并对资源选择做出自己的判断。

（3）应用环境信息

应用环境信息主要指机群的软件安装配置信息，包括商业/开源软件的安装路径、执行环境配置信息，应用启动命令等。这些信息按照SCE系统给定的规范记录并维护，通过SCE系统的维护命令录入系统维护并管理，同时提供给用户一个可全局查询的统一视图，为用户屏蔽了很多可以不需要关心的细节。由于应用信息更新的动态性相对较弱，因此无需系统定时维护，当机群的应用软件安装情况或配置情况发生变化时，由系统管理员手工维护即可。

（4）用户映射信息

用户映射信息主要记录了网格用户与机群用户的映射关系。考虑到用户计算数据的隔离，目前主要采用一对一的映射关系。用户映射信息主要有SCE系统的用户管理模块自动更新并维护。

2.2.4 网格编译环境

为了屏蔽不同高性能计算机中编译环境的异构，包括安装路径、编译命令、链接库名字等不同，把其中安装的所有的软件和编程库的信息统一管理起来，SCE中提供了一套网络编译工具GCAide，其主要功能包括：网格环境中已安装的编译器和程序库的查看；根据用户需求自动设置网格编译环境；提供统一的网格编译脚本描述方式；自动替换并产生适应指定目标编译节点的编译脚本；支持远程执行编译过程。

基于网格编译工具GCAide，用户可以选择自己需要的编译器和编程库软件包作为自己的网格编译环境配置，当用户提交编译请求的时候，系统根据请求的高性能计算机提供的具体编译环境自动完成用户编译环境的配置，启动编译过程，使得编译生成的可执行文件可以在相应高性能计算机上运行。

网格编译工具GCAide为最终用户提供了统一的使用命令，scelib和scemake。scelib命令可以查看到当前网格环境中各节点的编译环境安装情况，包括编译器和程序库的软件包名称、版本、安装路径以及编译命令/链接选项说明。scemake命令设置必要的编译环境，实现程序的编译过程，可以兼容程序原有的编译方式，包括直接的编译命令方式、shell脚本方式、make方式、configure+make方式。

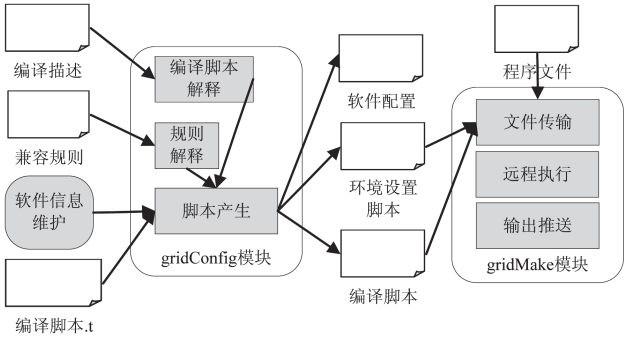


图3 网格编译工具流程

GCAide可以从三个不同层次为用户程序编译提供支持。

（1）支持用户手工查看软件安装信息并自行设置编译环境；

（2）支持用户提供的编译环境的软件列表，由系统自动设置编译环境的环境变量；

（3）支持系统自动替换并产生适应指定目标编译节点的编译脚本。

2.2.5 用户管理

SCE系统通过基于LDAP的全局分布式用户数据库实现全网格环境的用户管理子系统。LDAP分布式数据库负责账号信息在整个网格环境内的自动同步和更新功能。网格用户可以通过网格命令行、网格门户等多

种方式进行登录，这些最终从本地的LDAP数据库获取用户账号信息，从而实现全局统一的用户管理。不管用户通过何种方式和地点登录，都可以通过相同的用户名和密码登录SCE系统。并且登录成功后，可以获得相同的用户视图和权限，可以访问网格范围内具有权限的资源，由此实现网格环境的单一系统映像。

为了保证用户和计算资源提供者两个方面的权益，SCE系统中存在网格用户和主机用户两种用户形态。网格用户是指网格系统创建的账号，其信息存储网格用户子系统中，由网格软件管理。主机用户是指在高性能计算机上创建的账号，由机群管理员管理。网格用户和主机用户之间的映射关系存在于SCE系统中，每个高性能计算机资源拥有一份单独的映射文件。当网格用户请求具体的计算资源进行操作时，系统自动切换到主机用户，并完成用户指定的操作。SCE系统通过管理两类用户之间的映射，实现对网格用户的权限管理。机群管理员通过管理主机用户的权限，实现网格用户对其高性能计算机权限的管理，具有最高的管理权限。

2.2.6　监控和记账服务

SCE监控服务采用三层结构。资源层通过传感器收集监控资源的性能数据，并封装成最原始的数据格式发布。服务层以服务形式封装监控功能，包括数据获取、缓存、过滤、存档、数据发布、数据展示等服务。监控系统的框架基于GMA，每个服务包含生产者接口和消费者接口，这种结构具有较好的功能可扩展性。目录层负责目录管理，包括目录查询，维护和激活，同时实现对本地服务请求的预处理。

SCE记账服务的主要功能是准确记录并有效统计用户和应用对资源的使用情况，从而支持网格环境资源的合理优化管理，为网格的服务质量和服务级别管理提供保障。常用的计费策略包括先计算后付费、先付费后计算和即时付费。

## 2.3　多种使用方式

面向科学计算的网格环境面向用户提供了两种使用方式：命令行方式和网页门户方式，同时面向专业社区和学科平台建设提供了用于二次开发的编程接口。

2.3.1　SCE 命令行

科学计算领域中很多用户习惯于传统的命令行操作方式。某些领域的计算前后处理程序具有较强的交互性，某些领域主要进行自主软件的研制，因此要求科研人员自行编写编译程序。根据这些需求，SCE提供了基于命令行方式的使用界面[3]。

通过SCE命令行，用户可以完成文件编辑、程序编译、全局资源查找、全局作业提交、作业管理、中间结果查看、文件上传下载等工作。网格命令行工具软件保持了传统的Linux操作习惯，使用简单，满足了交互处理的需求，同时基于应用软件封装保证了用户在操作不同的高性能计算机时通过相同的命令即可使用应用软件，包括作业提交和前后处理，基于网格环境提供了强大的全局资源管理和作业调度能力。

SCE命令包括网格作业管理命令bsub/bjobs/bkill/bqueues，网格资源管理命令listres/listnodes/listapps，机群资源文件管理命令scels/scecat/sceput/sceget，网格环境设置命令set/alias，网格编译命令scelib/scemake，同时SCE兼容部分Linux命令的cat/cd/cp/ls等，用于网格家目录数据的管理。

2.3.2　SCE 网格门户

网格门户是高性能计算近几年新兴的一种使用方式，具有简单易用、用户友好等特点，用户可以随时随地通过浏览器完成高性能计算的相关工作。网格门户提供了丰富的作业管理操作，支持灵活快速地定制新的应用，增强了网页表示层的响应灵活性和灵敏程

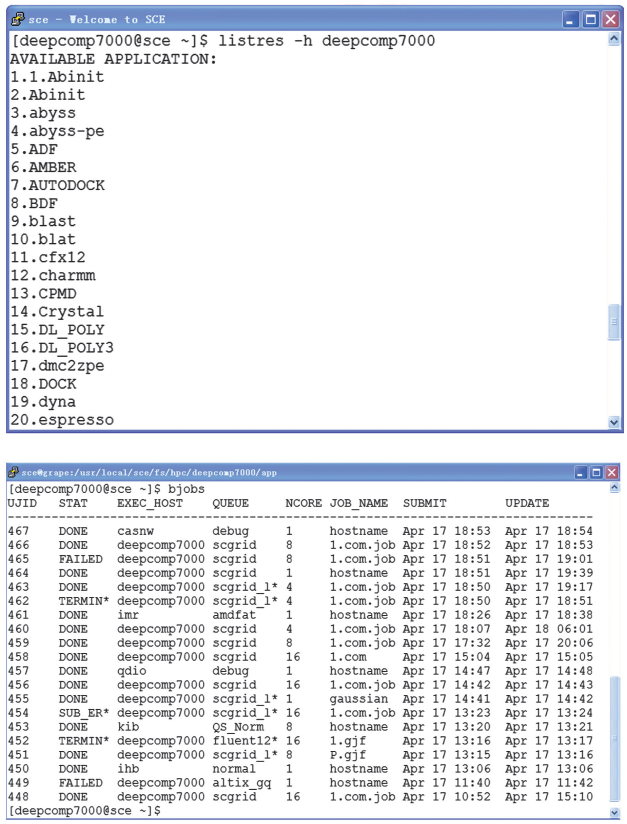


图4　SCE命令行操作界面

度，保障了数据传输的安全性，以及支持文件在线简单管理。

SCE网格门户[4,5]基于符合JSR-168规范的Portlet开发，并融合了Ajax技术[6,7]。Portlet 是门户的核心组件，负责接收浏览器端的请求并动态产生各种信息。Ajax能够有效增强Web表示层的响应灵活性和灵敏程度。网格门户由基础层、服务层和Portlets层构成。其中，Portlets的页面部分构成浏览器端，Portlets层的剩余部分、服务层和基础层构成网格门户的服务端。

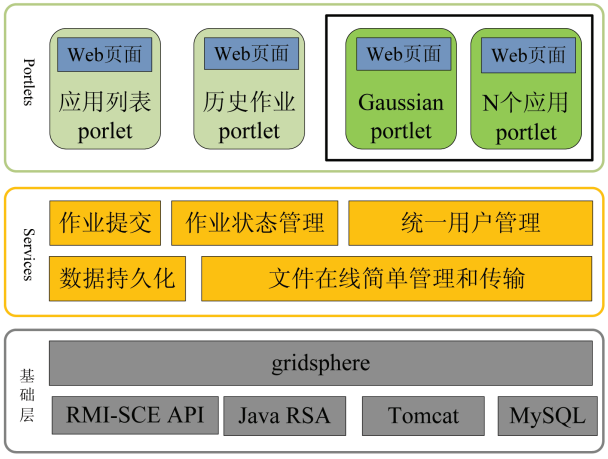


图5 SCE网格门户软件结构


图6 SCE网页门户操作界面

网格门户的核心服务包括作业提交服务、作业状态管理服务、文件在线简单管理和传输等几个部分。作业提交服务提供了JSDL描述标签处理、JSDL作业描述文件生成、文件上传、向SCE系统服务提交作业等功能。作业状态管理服务完成向用户展示实时或最终的作业状态信息，并提供终止作业的功能。文件在线简单管理和传输服务提供实时的作业工作目录列表、文件属性和文件内容查看、无缓存的大文件数据流下载功能。

除了命令行和网格门户两种用户使用界面，SCE系统还面向专业社区和学科平台建设提供了用于二次开发的编程接口和模型[8]。

## 2.4 SCE软件小结

SCE软件的设计及实现有以下4个方面的特点：

（1）稳定可靠

SCE软件通过资源汇报机制可以及时获取计算资源是否可用的信息，并反馈给用户，保证了资源不可用的时候可以及时给用户以提示，因此当单个计算资源在停机维护期间或者故障期间用户仍然可以选择其他可用的计算资源继续其科学研究活动。

SCE软件内部可以自行检测重要服务的失效性，并且自动重启失效的服务完成用户请求，这种机制在一定程度上保证了软件本身的稳定可靠。

同时，SCE软件结合自动测试框架[9]和第三方测试最大程度地保证软件代码质量和软件自身的稳定可靠性。

（2）轻量级软件

SCE软件总代码行数2万余行，所有模块和服务可部署于独立于高性能计算机群资源的网格服务器，因此部署在机群资源上的模块或服务数为0，占用的网络端口数量为1，并且只有在进行作业提交等某些资源操作时才与高性能计算资源交互，保证了其运行时占用的CPU和内存资源最少。

由于无需在计算资源上部署任何网格模块或服务，最大程度地减少了机群系统管理员在部署、维护和管理方面的工作量。

（3）全局作业调度和数据管理

SCE软件支持作业的全局调度，只要在作业提交

的时候选择A资源，系统会根据当前资源可用情况，自动将其调度到合适的的计算资源上执行。

SCE软件为用户提供了独立于计算资源的网格家空间，而机群数据空间只用于计算期间相关数据的非永久存储。这种方式下，用户只需自主管理其网格家空间数据即可，无需面对分布在不同计算资源上的临时存储空间的管理。

（4）易用的用户界面

SCE软件提供了统一的用户操作界面，屏蔽了各种计算资源的异构性，并且为了满足不同用户的需求，提供了命令行、网页门户和编程接口多种使用方式。

## 3　在科学院网格环境的应用

目前，面向科学计算的网格环境已经在中国科学院超级计算环境[10]中得到了实践和应用。中国科学院超级计算环境是由总中心、分中心、所级中心组成的三层架构科学计算网格环境（ScGrid）。它的计算资源在地理上分布甚广，计算能力和规模不同，使用的效率各不相同。通过SCE软件整合资源建成了统一的网格环境，实现了统一的运维管理与资源调度，提供了统一的用户使用界面，方便了用户使用，达到了资源整合与共享使用、提高资源利用率的目的。另外，通过聚合多个、多种计算资源，整体提高了计算服务的可靠性。实践证明，在单个超级计算系统因日常维护或故障停机时，网格环境可以发挥计算资源调配灵活的特点，由其他结点承担用户的计算任务，缓解停机对用户造成的不利影响。

截至2011年底， ScGrid已经在总中心、分中心、所级中心及GPU集群的31个网格节点部署SCE软件，聚合了近300万亿次的通用计算能力和近3000万亿次的GPU计算能力，实现了全环境的统一运行管理、技术支持与服务。已开通用户账号近200个，其中包括来自国外的用户申请，支持了一批国家863计划、973计划、国家自然科学基金等国家重点科研项目，用户累计提交的网格作业超过10万个，使用机时超过2000万CPU小时。据不完全统计，已有近20篇用户论文中明确标注受网格环境支持，这表明面向科学计算的网格环境已经得到用户认可。

在日常运行维护中，根据实际需要还进一步加强了环境的高可用性和容灾能力建设，采取的措施如网格服务器计算与存储分离，加强数据备份、实施数据库高可用方案，尝试机房内备份、机房间备份、异地备份的多级容灾方案，保证了优质的高性能计算服务，用户的科研工作并不中断，充分体现出网格环境可靠性的优势。

## 4　结　论

本文提出了面向科学计算的网格环境，旨在将分布的高性能计算资源汇聚和整合为一台虚拟的超级计算机，为用户提供统一、易用、可靠的科学计算服务。本文详细分析了在面向科学计算领域，网格中间件设计的难点和需要具备的特点，设计实现了网格中间件SCE，具有可靠稳定、轻量级的特点，支持全局作业调度和数据管理，并且提供了易用的用户界面。目前，面向科学计算的网格环境已经在中国科学院超级计算环境中得到了实践与应用，并且得到了广大用户的认可和好评。研究成果可以在相关科学计算环境中得到进一步推广和借鉴。

### 参 考 文 献


[1] Grid computing [EB/OL]. http://www.gridcomputing.com/.

[2] Foater I, Kesselman C. The grid: blueprint for a new computing infrastructure [M]. Morgan Kaufmann, 2004.

[3] 龙斌, 迟学斌, 肖海力. 基于命令行客户端的网格软件SCE设计与实现 [J]. 计算机系统应用, 2010, 19(9): 64-68.

[4] 曹荣强, 迟学斌, 武虹, 等. 基于Portlet的高性能计算Portal [J]. 计算机工程, 2009, 35(15): 1-3.

[5] Cao R Q, Chi X B, Cao Z Y, et al. USGPA: a user-centric and secure grid portal architecture for high-performance computing [C] //IEEE International Symposium on Parallel and Distributed Processing with Application. 2009: 432-438.

[6] Yang X, Allan R. Bringing AJAX to grid portals [C] // Collaborative Technologies and Systems. 2007 CTS International Symposium. Orlando, FL, 2007: 257-264.

[7] Zhu Y, Shen P. Bring AJAX to web application based on grid service [C] //Proceedings of the 2009 International Conference on Intelligent Human-Machine Systems and Cybernetics. Hangzhou(China), 2009: 162-165.

[8] 曹荣强, 曹宗雁, 迟学斌, 等. 基于RMI的高性能计算网格二次开发模型 [J]. 计算机应用, 2010, 30(9): 2526-2529.

[9] 卢莎莎, 迟学斌, 肖海力, 等. 基于科学计算网格软件SCE的自动化测试 [J]. 计算机应用研究, 2011, 28(增刊) : 591-593.

[10] 金钟, 朱鹏, 肖海力, 等. 中国科学院超级计算环境及其应用 [J]. 中国科研信息化蓝皮书, 2011: 144-155.

[11] Xiao H L, Wu H, Chi X B. SCE: grid environment for scientific computing [C] //Proceedings of GridNets'. 2008: 35-42.

[12] Catlett C, Allcock W, Andrews P, et al. Teragrid: analysis of organization, system architecture, and middleware enabling new types of applications [C]. HPC and Grids in Action, Amsterdam. 2007.

[13] EGEE [EB/OL]. http://public.eu-egee.org/

[14] WLCG [EB/OL]. http://lcg.web.cern.ch/

[15] Depei Q. CNGrid: A test-bed for grid technologies in china [C]. 10th IEEE International Workshop on Future Trends of Distributed Computing Systems (FTDCS'04). 2004.

[16] 钱德沛. 国家高性能计算环境及其应用 [J]. 中国科研信息化蓝皮书, 2011: 130-135.

[17] 查礼, 韦海亮, 程伯群, 等. 中国国家网格软件CNGrid GOS [J]. 中国科技成果, 2009, 10(11): 7-13.

[18] Jin H. Chinagrid: making grid computing a reality [J]. Digital Libraries: International Collaboration and Cross-Fertilization, 2005: 13-24.

[19] 金海, 吴松. 中国教育科研网格及其应用 [J]. 中国科研信息化蓝皮书, 2011: 136-143.