# 基于 HPC Profile 标准的网格批作业服务的研究与实现①

沈焱峰 (中国科学院 计算机网络信息中心 北京 100190;中国科学院 研究生院 北京 100049)

奚自立 (上海超级计算中心 上海 201203)

**摘　要：** 以上海超算中心目前急需解决的异构网格中间件之间的互操作问题为研究背景，深入研究了 HPC Profile 标准及实现的关键技术，设计和实现了基于 HPC Profile 标准的网格批作业服务，并将此服务集成到各网格中间件上，使之支持多种类型的作业调度器，并在此基础上实现了网格中间件之间的批作业服务互操作，提升现有网格应用系统的可扩展性和网格中间件的互操作性。

**关键词：** 网格；HPC Profile；批作业服务；中间件互操作

## Research and Implementation of Grid Batch Service Based on HPC Profile

SHEN Yan-Feng (Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China; Graduate University of Chinese Academy of Sciences, Beijing 100080, China)

XI Zi-Li (Shanghai Supercomputer Center, Shanghai 201203, China)

**Abstract:** In light of the problem of heterogeneous grid middleware interoperability in Shanghai Supercomputer Center, this paper designs and realizes batch grid operations services and grid middleware interoperation based on the HPC Profile standard. Then, it upgrades the scalability and interoperability of the grid middleware.

**Keywords:** grid; HPC profile; batch services; middleware interoperability

## 1　课题背景

目前上海超算拥有神威 cluster，曙光 4000A 以及实验集群等多个异构机器，上面安装部署了不同的作业调度系统。例如神威 cluster 装有 Open PBS，曙光 4000A 装有 LSF，而内部实验集群装有 MS CCS。以及三期工程即将安装的曙光 5000A 应用何种作业调度系统目前尚未明确。

在网格应用方面，上海超算参与了多个网格项目，包括基于 GOS 中间件的 CNGrid 以及基于 Globus Tookit 的上海网格，此外在内部实验集群上安装了支持 PBS 作业调度器的 eComputing 中间件，以及支持 LSF 作业调度器的 engineFrame 中间件。

目前主要存在以下两个问题：第一，由于作业调度器采用不同的接口标准，现有的网格中间件不能做到支持多种作业调度器；例如目前的网格中间件 GOS3 尚未能支持微软公司 CCS 作业调度器。

第二，由于采用不同的服务标准，各网格中间件互不兼容，互操作性差[1]。例如中心内部试验机群使用的是 eComputing 网格中间件，但是 eComputing 和 GOS 之间基于不同的服务开发标准，因而不能通过 eComputing 直接调用 GOS 网格服务实现互操作，因此在试验机群上进行实验操作和性能评测时，无法在 eComputing 上调用曙光 4000A 的软硬件计算资源，给实验操作造成很大不便。

针对上述问题，提出以下解决方法：在对 HPC Profile 标准深入研究的基础上，设计和实现基于 HPC

① 基金项目:国家高技术研究发展计划(863)(2006AA01A117)
收稿时间:2009-06-16

Profile 标准的网格批作业服务，并将此服务集成到各网格中间件上，使之支持多种类型的作业调度器[2]。

## 2 HPC Profile标准简介

PC Profile 是一种用于网格环境中的 Web 服务和作业管理的标准，主要针对高性能计算应用。

在下图 1 所示，在 HPC Profile 标准应用模型中，可以看到网格作业提交通过 JSDL 进行描述，上层应用支持 Web Services，胖客户端以及工作流引擎，通过 Web 服务接口 BES 同下层网格资源管理层进行交互[3]，此外 WS-I BP 机制用于用于安全控制和网格中间件之间的互操作。

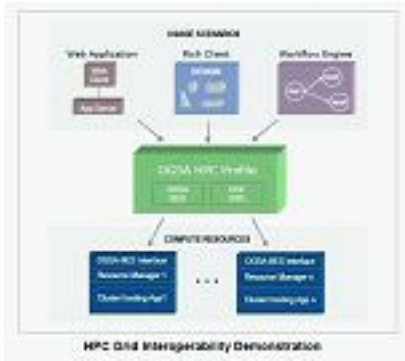

图 1 HPC Profile 标准应用模型

## 3 网格批作业服务系统设计及实现

### 3.1 系统功能及框架设计

设计基于 HPC Profile 标准的网格批作业服务系统如下图所示，包括如下三个核心部分：

(1) 网格批作业驱动程序

接受批作业服务的调用，生成辅助文件并调用批作业系统。

(2) 网格批作业服务端

通过对底层批作业系统(PBS 等)的服务化封装，为客户端提供统一、通用的批作业处理功能，包括：作业提交、状态查询、作业取消、获取标准输入/输出，数据的 stagein/stageout 等。

(3) 网格批作业客户端

通过网程与网格批作业服务之间进行交互，屏蔽底层服务的访问细节，为用户(应用)提供方便、易用的批作业处理接口，包括：作业提交、状态查询、作业取消、获取标准输入/输出等。

### 3.2 关键技术

拟解决的关键技术如下：

① 基于 JSDL 的网格作业描述，包括作业 ID，相关参数，输入输出等

② 对于轮询和通知相结合的作业查询机制的实现

③ 该服务的远程调用实现，包括远程地址解析和用户组安全认证

④ 对于资源预留机制的实现

⑤ 对于周期性资源回收机制的实现

### 3.3 系统开发实现

拟解决的关键技术如下：

① 基于 JSDL 的网格作业描述，包括作业 ID，相关参数，输入输出等

② 对于轮询和通知相结合的作业查询机制的实现

③ 该服务的远程调用实现，包括远程地址解析和用户组安全认证

④ 对于资源预留机制的实现

⑤ 对于周期性资源回收机制的实现

(1) 批作业服务端：

搜集调用信息，从 GOSContext 中获取 CallUser 创建本次调用工作目录并更改访问权限[4]；

如果有输入文件，则从全局空间下载到本地；

fn.origin 存放作业描述中的脚本部分

fn.stagein 存放作业描述中的 stagein 部分

fn.stageout 存放作业描述中的 stageout 部分

根据 gridmap，从 DN 映射到本地用户名，调用批作业驱动程序，获得作业 ID；

返回作业 ID 给客户端[5]。

(2) 批作业客户端：

记账：把作业信息放入作业描述表。作业描述表在批作业初始化的时候生成，存放所有提交过的作业信息，可被作业清除程序定期清除；

创建 StageoutWorke 线程,Stageout Worker 负责监控作业执行状态，当作业执行完毕时(status=Done)，此线程把结果数据上传到全局空间；

(3) 批作业驱动程序：

根据 fn.origin,fn.stagein,fn.stageo 生成 fn.pbs；

接受 Batch Service 调用，提交 fn.pbs 到后台

批作业系统;

返回结果文件以及 stdout 和 stderr，Setuid[6]。

## 4 网格批作业服务系统设计及实现

该网格批作业服务系统基于 HPC Profile 标准，相比原有网格批作业服务系统具有如下改进之处:

① 自动识别 Condor，LSF， SGE ， Maui ，Open PBS,PBS Pro 以及 MS CCS 等多种作业调度器。

② 作业监控采用轮询和通知相结合的方式

③ 采用 Exactly-Once 保障用户作业提交请求被获知

④ 支持同外部客户的实时交互

⑤ 支持客户和 SA 进行交互式作业管理

⑥ 支持 weighted-fair-share 等多种作业调度策略

⑦ 支持工作流

⑧ 资源预留机制以及周期性的资源回收机制

首先支持对 LSF，Open PBS 以及 MS CCS 等多种作业调度器自动识别机制。因此通过将基于此标准的网格批作业服务集成到相应的网格中间件中，可以在不需要进行二次开发的情况下，使该中间件支持更多类型的作业调度器[7]。

第二点，独特的安全控制和互操作机制 WS-IBP，主要用于网格中间件之间的批作业服务互操作。

此外，同原有网格批作业服务系统一样，该系统支持工作流引擎。在资源的分配方面支持资源预留机制，在资源的管理方面采取周期性的资源回收机制，并且支持元数据表达语义信息。

下面以上海超算 CNGrid 网格中间件 GOS 和实验网格中间件 eComputing 为例，说明如何使之支持更多种作业调度器，并实现两者之间的互操作。

### 4.1 系统功能及框架设计

如下图 2 所示,将此服务以插件的形式集成到 GOS 网格应用平台上，如下图所示，由于采用基于 HPC Profile 的统一接口标准，使得 GOS 能够自适应支持 LSF，Open PBS 以及 MS CCS 等多种作业调度器。

### 4.2 网格中间件互操作

将此服务系统以插件的形式集成到 GOS 和 eComputing 上，通过远程调用彼此的批作业服务，实现两者之间的互操作。下面以一次跨平台作业请求说明如何实现网格中间件互操作。假设 GOS 平台上用户请求部署在 eComputing 平台上的该批作业服务。用户调用流程如下:

(1) 用户在 GOS 平台上提交作业请求，经过查询获得服务所在平台，在进行用户身份映射之后将作业请求提交给 eComputing 插件;

(2) eComputing 插件将参数格式转换为改平台执行管理模块调用接口可以识别的参数类型之后，再把作业请求提交 eComputing 执行管理部件;

(3) eComputing 执行管理部件在执行完作业之后，把作业结果返回给用户，或者根据用户调用方式(同步或者异步)将结果存放在网格系统中指定的存储空间;

(4) 用户再在 GOS 平台上提交作业请求，经上述步骤将请求依次转交给最底层的 eComputing 执行管理部件，由 eComputing 执行管理部件将保存的作业结果返回给用户，完成一次作业的提交，运行和返回结果。

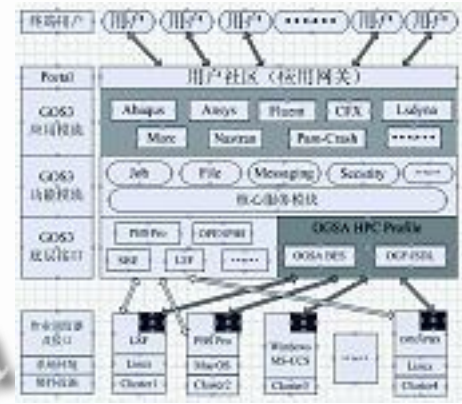


图 2　系统功能及框架设计

### 4.3 应用中发现的潜在问题

通过改批作业服务系统的实现和应用，发现 HPC Profile 标准在文件传输方面存在一些细微缺陷，文件传输速度过慢并且无法准确获取文件传输的进度信息有待进一步的修订和完善。

## 5 总结及展望

本文以上海超算中心目前急需解决的异构网格中间件之间的互操作问题为研究背景，深入研究了 HPC Profile 标准及实现的关键技术，设计和实现了基于

(下转第 239 页)

(上接第 212 页)

HPC Profile 标准的网格批作业服务，并将此服务集成到各网格中间件上，使之支持多种类型的作业调度器，并在此基础上实现了网格中间件之间的批作业服务互操作，提升现有网格应用系统的可扩展性和网格中间件的互操作性。

## 参考文献


1 郑然,金海,章勤.网格工作流资源层次模型与访问机制.华中科技大学学报(自然科学版), 2006,34(增刊 I):37－40.

2 许中清,李胜利,吴松,石宣化.基于插件的网格作业管理互操作策略.华中科技大学学报(自然科学版), 2006,34(增刊 I):141－143.

3 Theimer M, Parastatidis S, Hey T, Humphrey M, Fox G An Evolutionary Approach to Realizing the Grid Vision, February 2006.

4 Theimer M. HPC Use Cases- Base Case and Common Cases, April 2006.

5 Grimshaw A, WS-I Basic Security Profile Version 1.0, Final Material, 2007-03-30.

6 Smith MT. OGSA Basic Execution Service Version1.0.GFD-R-P.112,March2007.

7 Foster I. Web Services Security Username Token Profile, Working Draft 2, OASIS, 23 Feb 2003.