WSRF기반의 TIGRIS 그리드 MPI 서비스

TIGRIS Grid MPI Service based on WSRF

권오경, 박경랑*, 권오영**, 함재균, 이필우
KISTI 그리드컴퓨팅연구팀, 연세대학교*,
한국기술교육대학교**

Oh-Kyoung Kwon, Kyung-Lang Park*, Oh-Young Kwon**,
Jaegyoon Hahm, Pill Woo Lee
KISTI, Yonsei University*, KUT**

요약

본 논문에서는 WSRF기반의 TIGRIS 그리드 MPI 서비스를 기술하고자 한다. 본 서비스는 그리드 환경에서 MPI 작업을 실행하는 서비스이다. 다양한 계산 자원 및 MPI 라이브러리를 지원한다. 주요 기능은 다음과 같다. 첫째, 특정 MPI 라이브러리의 사용법을 알지 못해도 작업을 실행할 수 있다. 두 번째, 두 개 이상의 자원에서 동시에 MPI 작업을 실행할 수 있게 도와주는 그리드 MPI 라이브러리를 지원한다. 세 번째, 사용자가 컴파일을 하지 않아도 자동으로 실행할 수 있게 한다.

Abstract

In this paper, we describe TIGRIS Grid MPI Service, which is the WS-Resource Framework (WSRF) based services to enable an MPI job to be executed on Grid environments. It covers heterogeneous compute resources and diverse MPI libraries. The main functionalities are as follows. First, it allows an MPI user to seamlessly launch a job without knowing how to use the specific MPI library. Secondly, it executes an MPI job on the cross-site resources by supporting the Grid-enabled MPI library such as MPICH-G2. Thirdly, it enables the user to launch a job using the source code without compiling.

I. 서 론

그리드는 지리적으로 분산된 자원을 동시에 활용하여 거대 계산 문제를 해결하기 위한 방법을 제공한다[1]. MPI는 병렬 컴퓨터에서 효율적으로 빠른 시간 내에 수행되기 위한 응용프로그램에서 사용되어 왔다. 초기에 MPI는 여러 병렬 컴퓨터에서 효율적으로 수행되는 응용 프로그램을 작성하기 위해 개발되었으며, 대표적인 라이브러리 구현물로 미국 ANL(Argonne National Laboratory)에서 구현한 MPICH가 널리 이용되고 있다.

MPI와 같은 고성능 컴퓨팅이 필요한 환경이 그리드 환경으로 확장됨에 따라 그리드 상에서의 MPI라이브러리의 지원이 필요하게 되었다. 그래서 그리드 환경에서 동작하도록 한 대표적인 구현물이 MPICH-G2이다[2]. 하지만 MPICH-G2는 응용 사용자들 위한 저수준의 라이브러리이다. 다양한 MPI 라이브러리와 계산 자원에 대한 복잡성 및 이질성을 사용자가 모르게 하지만, 여전히 사용하기 어렵고 불편한 점이 있다. 즉, 사용자는 이기종의 자원에 대해서 새로 컴파일을 해야 하며, MPICH-G2가 지원하는 라이브러리만 사용가능하다. MPI 작업 2개 이상의 자원을 사용할 시에는 각 자원에 대해서 모두 로그인해서 실행파일에 대해서 컴파일을 다시 해야 한다.

그래서 위와 같은 사항을 만족시키기 위해, 우리는 고수준의 그리드 서비스인 TIGRIS 그리드 MPI 서비스를 제안한다. 다양한 종류의 MPI 라이브러리와 이기종의 계산 자원을 지원하기 위해 우리는 WSRF 기반의 그리드 서비스를 채택하였고, 기반 미들웨어는 글로버스 툴킷을 사용하였다. WSRF는 그리드에서 웹서비스를 적용하기 위해 필요한 여섯 개의 웹서비스의 집합체이다[3].

본 논문에서는 TIGRIS 그리드 MPI 서비스를 설계 및 구현하였다. 크게 두 가지 목적이 있다. 첫 째로 계산 그리드 환경에서 MPI 작업 수행을 위한 통일된 인터페이스를 제공한다. 두 번째로 그리드 환경에서 계산 작업에 대한 재사용 및 협업 환경을 제공한다.

그리드를 위한 병렬 프로그래밍 모델로서 MPI는 지속적으로 사용될 것이며, 특히 계산 과학을 위한 분야에 이용될 고성능 MPI 기술은 계속적으로 수요가 증가할 것으로 판단된다. MPI를 비롯한 고성능의 프로그래밍 모델은 응용분야의 지원과 발전을 위해 반드시 선행되어야 할 분야이므로 국내 그리드 환경의 구축과 활용을 위해 향후 지속적으로 연구될 전망이다.

본 논문은 다음과 같은 구조로 되어 있다. 2장은 TIGRIS 인프라스트럭쳐와 소프트웨어 스택에 대한 설명, 3장은 TIGRIS

그리드 MPI 서비스에 대한 전반적인 설명으로 되어 있다. 마지막으로 4장에 결론 및 향후 연구에 대해 기술한다.

Ⅱ. TIGRIS

1. TIGRIS 인프라스트럭쳐

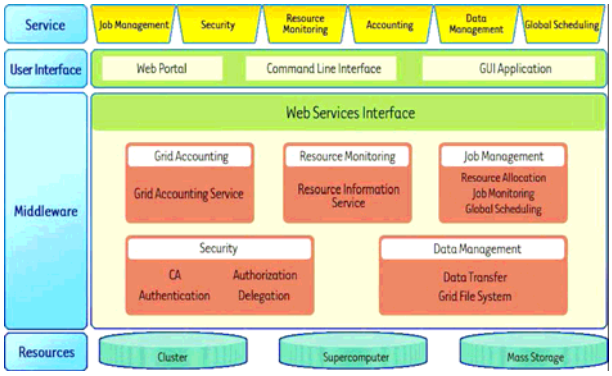
TIGRIS (Tera-scale Infrastructure for K*GRId Services)는 안정적인 국가 그리드 서비스 인프라를 제공하기 위해 진행되고 있다[4]. 현재 TIGRIS는 그림 1과 같이 한국의 서울대학교(서울), 부산대학교(부산), KISTI(대전)에서 그리드를 위한 계산 자원이 제공되고 있다. 각 계산 자원은 이기종의 프로세서, 운영체제, 배치 스케줄러를 사용하고 있다. 서울대학교는 파워 프로세서(PPC), 리눅스(Linux), Loadleveler, 부산대학교는 Itanium-II, 리눅스, PBS, 그리고 KISTI는 IA32, 리눅스, SGE를 보유하고 있다.

▶▶ 그림 1. TIGRIS 구조

2. TIGRIS 소프트웨어 스택

TIGRIS는 그림 2와 같은 소프트웨어 스택을 가지고 있다. TIGRIS 소프트웨어 스택을 통해 안정적인 서비스를 제공하고, 이기종의 자원과 풍부한 사용자 인터페이스 지원하는데 목적이 있다. TIGRIS 서비스는 글로버스 얼라이언스(Globus Aliance)에서 개발된 글로버스 툴킷(Globus Toolkit)기반으로 이루어져 있다[5].

글로버스 툴킷에서 제공하는 기본 서비스 기반에 TIGRIS는 다음과 같은 서비스를 포함하고 있다. TIGRIS에서 작업을 관리 및 제출하는 TIGRIS 작업 관리 서비스(TIGRIS Job Management Service), MPI 작업을 처리하기 위한 TIGRIS 그리드 MPI 서비스(TIGRIS Grid MPI Service), 파일 전송을 담당하고 있는 TIGRIS 파일 관리 서비스(TIGRIS File Management Service), TIGRIS 그리드 계정 서비스(TIGRIS Grid Account Service)이다. 본 논문에서는 TIGRIS 그리드 MPI 서비스에 대해서 논의하고자 한다.



▶▶ 그림 2. TIGRIS 소프트웨어 스택

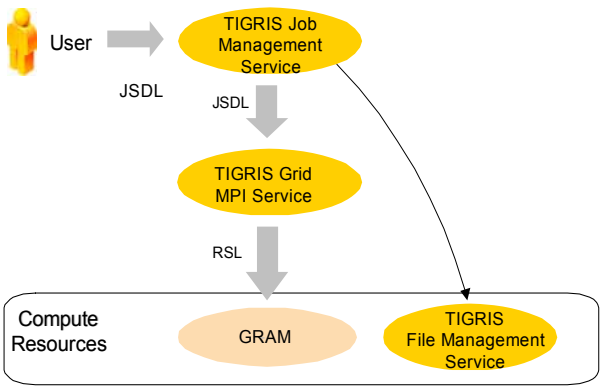
Ⅲ. TIGRIS 그리드 MPI 서비스

1. 개요

TIGRIS 그리드 MPI 서비스는 WSRF기반의 서비스로써, 그리드 환경에서 MPI 작업을 제출 및 실행할 수 있게 하는 서비스이다. TIGRIS 인프라스트럭처를 지원하기 위하여, 다양한 종류의 계산 자원과 MPI 라이브러리를 수행할 수 있다.

TIGRIS 그리드 MPI 서비스는 앞 절에서 언급하였듯이 TIGRIS 소프트웨어 스택에서 제공하는 서비스 가운데 하나이다. 그래서 TIGRIS 다른 서비스와 함께 사용이 가능하다. 하지만 TIGRIS 서비스뿐만 아니라 다른 그리드 인프라 환경과도 사용가능하다.

그림 3은 TIGRIS의 다른 서비스와의 작업 흐름도를 나타내었다. 첫 번째 사용자는 MPI 작업을 TIGRIS 작업 관리 서비스에 제출한다. 제출된 작업은 TIGRIS 그리드 MPI 서비스를 통해 작업이 계산 노드에 할당된다. 이때 이동이 되어야 하는 파일은 TIGRIS 파일 관리 서비스에 의해서 수행한다.



▶▶ 그림 3. TIGRIS 그리드 서비스에서의 MPI 작업 흐름

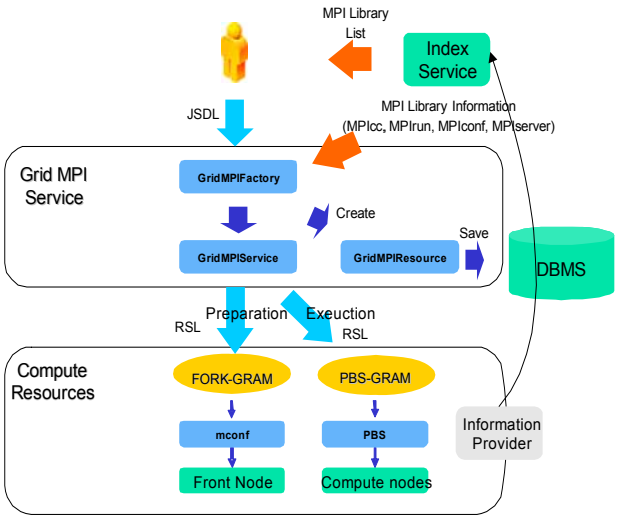
2. 기능

TIGRIS 그리드 MPI 서비스의 주요기능은 다음과 같다. 첫째, 특정 MPI 라이브러리의 사용법에 대해 알지 못해도 작업을 실행할 수 있다. 즉, 사용자는 특정 자원에서 MPI 라이브러리로 컴파일된 실행파일의 수행법을 알지 못해도, 서비스는 자동으로 실행 가능하게 한다. 두 번째, 두 개 이상의 자원에서 동시에 MPI 작업을 실행할 수 있게 도와주는 그리드 MPI 라이브러리를 지원한다. 현재 그리드 MPI 라이브러리는 글로버스 얼라이언스에서 개발한 MPICH-G2와 KISTI에서 개발한 MPICH-GX, 일본 AIST에서 개발한 GridMPI 등이 있다.

세 번째, 사용자가 컴파일을 하지 않아도 자동으로 실행할 수 있다. 다시 말해서, 사용자는 자원에 로그인을 해서 직접 컴파일을 수행하지 않아도 되고, MPI 라이브러리의 컴파일 방법 또한 알지 못해도 된다.

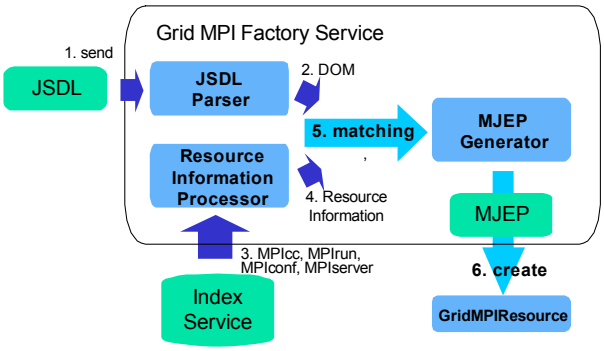
3. 구조

그림 4에 TIGRIS 그리드 MPI 서비스의 전체 구조도가 있다. 서비스는 두 개의 WSRF기반의 서비스(GridMPIFactoryService와 GridMPIService), 하나의 응용프로그램(mconf), 그리고 MPI 정보 제공자를 포함하고 있다. GridMPIFactoryService 서비스는 GridMPIService의 인스턴스(instance)를 생성하는 서비스이다. GridMPIService 서비스는 계산 자원에 로컬스케줄러를 통해 작업을 할당한다. mconf는 소스 코드를 컴파일하고 실행환경을 설정한다. 마지막으로 MPI 정보제공자를 통해 각 자원에 설치되어 있는 MPI 라이브러리들의 정보를 알 수 있다.



▶▶ 그림 4. 그리드 MPI 구조도

3.1 GridMPIFactoryService



▶▶ 그림 5. GridMPIFactoryService의 작업흐름도

그림 5는 GridMPIFactoryService의 작업흐름을 나타내고 있다. 첫 번째, 사용자는 JSDL을 통해 MPI 작업을 기술한다. JSDL은 OGF에서 표준화 작업을 거친 그리드 작업 제출언어이다[6]. 서비스는 JSDL을 통해 사용자의 요구사항을 처리한다. 이때 서비스는 사용자의 요구사항과 정보서비스로부터 받은 정보를 매치메이킹을 하여 실제로 작업을 던질 작업 스크립트를 작성한다. 작업 스크립트는 크게 두 단계로 작성이 된다. 하나는 MJEP(MPI Job Execution Plan)라는 언어를 통해 작성이 되고, 다른 하나는 글로버스 툴킷에서 사용되는 언어인 RSL로 나타낸다. MJEP는 실제 MPI 작업을 실행하기 전에 컴파일 및 환경설정을 하기 위해 사용이 되고, RSL은 실제 MPI 작업에 대한 명시다. 작업 스크립트가 작성되고 나면 마지막으로 WS-Resource가 생성되고 초기화 된다. 이때 생성된 정보는 DBMS에 저장이 됨으로써 서비스 컨테이너와 관련 없이 영구적으로 저장된다.

MJEP는 그림 6에 나타난 것처럼 컴파일과 환경 설정에 관련된 정보를 포함하고 있다. 컴파일은 MPI를 실행하기 위한 소스코드에 대한 정보이고, 환경 설정은 작업을 수행하기 전에 필요한 프로그램 등의 실행이다.

```
<mjep:JobDefinition ...>
   <mjep:RMs num="1" type="single">
      <mjep:RM>
         ...
          <mjep:ConfModule>
          ${GLOBUS_LOCATION}/libexec/mconf
          </mjep:ConfModule>
          <mjep:Execution>
            <mjep:Command>
            cd ~/
            </mjep:Command>
            <mjep:Command>
             /usr/local/mpich/bin/mpicc -o ~/cpi cpi.c
            </mjep:Command>
          </mjep:Execution>
      </mjep:RM>
   </mjep:RMs>
</mjep:JobDefinition>
```

▶▶ 그림 6. 컴파일 정보를 포함하는 MJEP 예제

3.2 GridMPIService

GridMPIService 서비스는 실제로 작업을 수행할 자원에 제출한다. 서비스는 미리 생성된 MJEP와 RSL을 이용해서 각 자원에 작업을 제출한다. MJEP는 글로버스 툴킷의 GRAM의 FORK작업 형태로 작업을 수행하고, RSL은 GRAM의 PBS와 같은 로컬 스케줄러 형태로 작업을 수행한다. TIGRIS는 GRAM을 사용하지만 glite와 같은 다른 그리드 미들웨어에도 작업 제출이 가능할 수 있게 어댑터(adapter)를 제공한다.

3.3 mconf

mconf는 다음과 같은 두 단계의 작업을 수행한다. 작업의 실행과 설정이다. 이때 입력은 그림 6과 같은 MJEP로 되어 있다. MJEP는 GridMPIService 서비스에서 작업을 수행하기 전에 각 자원의 상단 노드에 자동으로 이동한다. 실행 단계에서 작업에 명시된 순서대로 수행한다. 일반적으로 컴파일 명령이 포함되어 있다. 그리고 설정 단계에서 MJEP의 ConfModule 항목에 명시된 명령어를 수행한다. 보통 작업을 처리하는 데몬 프로세스 명령어다.

```
<MPIProbedInformation>

  <MPILibrary>
     <MPIname>MPICH</MPIname>
     <MPIcc>/usr/local/mpich/bin/mpicc</MPIcc>
     <MPIrun>/usr/local/mpich/bin/mpirun</MPIrun>
     <MPIMachineOption>
     -machinefile
     </MPIMachineOption>
     <MPINPOption>
     -np
     </MPINPOption>
  </MPILibrary>

  <MPILibrary>
     <MPIname>GridMPI</MPIname>
     <MPIcc>/usr/local/gridmpi/bin/mpicc</MPIcc>
     <MPIrun>/usr/local/gridmpi/bin/mpirun</MPIrun>
     <MPIconfig>
     /etc/MPI_service/gridMpi_conf.sh
     </MPIconfig>
    <MPIserver>
     /usr/local/gridmpi/bin/impi-server
    </MPIserver>
  </MPILibrary>

</MPIProbedInformation>
```

▶▶ 그림 7. MPI 라이브러리 정보 제공자

3.4 MPI 정보 제공자

정보 제공자는 각 자원의 MPI 정보를 정보 서비스에게 제공한다. 정보 서비스는 모든 자원의 MPI 정보를 가지고 있으면서, TIGRIS 그리드 MPI 서비스가 수행 시 필요한 정보를 제공한다. MPI 작업을 컴파일하고 설정하고 수행하는 모든 정보을 포함한다. GRAM은 기본적으로 CPU, 호스트, 프로세스 정보를 제공하지만 MPI 라이브러리에 대한 상세 정보를 제공하지 않으므로 새로운 MPI 정보 제공자를 추가하였다.

정보는 XML형태로 표현이 된다. 다음과 같은 7가지 기본정보를 포함한다. MPI 라이브러리 이름 <MPIName>, 컴파일러 경로 <MPICC>, 실행기 경로 <MPIRun>, 환경 설정 파일의 경로 <MPIConfig>, mpirun 수행 시 머신 파일의 설정 옵션 <MPIMachineOption>, mpirun 수행 시 필요한 프로세스 개수에 대한 옵션 <MPINPOption>, 그리고 GridMPI 라이브러리를 사용 시 필요한 서버 명령어 <MPIServer>이다.

Ⅳ. 결론 및 향후 연구

TIGRIS는 안정적인 국가 그리드 서비스 인프라를 제공하기 위해 그리드 소프트웨어 스택을 제공하고 있다. TIGRIS 그리드 MPI 서비스는 TIGRIS 인프라에서 MPI 작업을 실행하기 위한 WSRF기반의 서비스이다. 다양한 MPI 라이브러리 및 이기종의 계산 자원을 지원하고 있다.

그리드 MPI 서비스는 다음과 같은 세가지 기능을 제공하고 있다. 첫째, 특정 MPI 라이브러리의 사용법에 대해 알지 못해도 작업을 실행할 수 있다. 즉, 사용자는 특정 자원에서 MPI 라이브러리로 컴파일된 실행파일의 수행법을 알지 못해도, 서비스는 자동으로 실행 가능하게 한다. 두 번째, 두 개 이상의 자원에서 동시에 MPI 작업을 실행할 수 있게 도와주는 그리드 MPI 라이브러리를 지원한다. 세 번째, 사용자가 컴파일을 하지 않아도 자동으로 실행할 수 있다. 다시 말해서, 사용자는 자원에 로그인을 해서 직접 컴파일을 수행하지 않아도 되고, MPI 라이브러리의 컴파일 방법 또한 알지 못해도 된다.

현재 TIGRIS 그리드 서비스는 MPI를 사용하는 모든 응용을 대상으로 하고 있다. 하지만 특정 응용을 대상으로 발생할 수 있는 경우에 대해 실험 및 성능 측정이 이루어지지 않았다. 앞으로 기상 분야의 수치 모델 및 CFD 등의 MPI 응용 프로그램을 TIGRIS 인프라에서 직접 수행해볼 예정이다.

▌참 고 문 헌▐

[1] I. Foster, C. Kesselman. Chapter 2 of "The Grid: Blueprint for a New Computing Infrastructure", Morgan-Kaufman, 1999.

[2] Nicholas T. Karonis, Brian R. Toonen, Ian T. Foster: MPICH-G2: A Grid-enabled implementation of the Message Passing Interface, Journal of Parallel and Distributed computing 63 (5) (2003) 551-563.

[3] Steve Graham, Anish Karmarkar, Jeff Mischkinsky, Ian Robinson, Igor Sedukhin, Web Services Resources specification 1.2, Web Service Resource Framework (WSRF) TC, OASIS, 2005

[4] J. Hahm, H. Myung, J. Kwak, O. Kwon, Design and Implementation of the Grid Services on TIGRIS, HPCAsia 2007, pp. 47-53.

[5] I. Foster, C. Kesselman, J.M. Nick, S. Tuecke, Grid services for distributed system integration, IEEE Computer 35 (6) (2002), pp. 37-46.

[6] A. Anjomshoaa, F. Brisard, M. Drescher, D. Fellows, A. Ly, S. McGough, D. Pulsipher, A. Savva, Job Submission Description Language (JSDL) Specification version 1.0, Job Submission Description Language Working Group, Open Grid Forum, 2002.