그리드를 이용한 바이오 인포메틱스 응용 클라이언트 설계

유승범, 신동규 ,신동일
세종대학교 컴퓨터공학과
bummy@gce.sejong.ac.kr

Design of Bioinformatics Application using Grid

Seung-Bum Yoo, Dong-Kyoo Shin, Dong-il Shin
Department of Computer Engineering, Sejong University

요 약

최근 생명공학 분야에서는 IT와 BT가 결합하는 새로운 패러다임의 컴퓨팅 환경이 구축되고 있다. 이에 게놈 프로젝트 결과 분석해야 하는 데이터의 양은 엄청나게 증가하고 있다. 그러한 데이터를 처리하기 위해서는 대규모 저장장치 외에 슈퍼컴퓨터 급의 고성능 컴퓨터가 필요하게 되었다. 그러한 데이터를 처리하기 위해서는 대규모 저장장치 외에 슈퍼컴퓨터 급의 고성능 컴퓨터가 필요하게 되었으며, 바이오 인포메틱스 분야를 지원하기 위해서는 대규모 하드웨어 뿐만 아니라 데이터베이스, 데이터 마이닝 등의 소프트웨어 기술로 인해 그리드 환경을 요구하게 되었다. 이에 본 논문에서는 그리드 환경에서 분산된 수많은 생물학 데이터베이스에 쉽게 접근할 수 있는 통합 환경으로 응용 클라이언트를 제시할 것이다.

1. 서 론

최근의 어플리케이션 프로그램 개발자들은 일반적으로 동질적으로 신뢰할 수 있으며 중앙에서 관리가 가능한 전산환경을 가정하면서 개발을 해왔다. 그러나, 실제 전산 개발환경은 분산 자원들과 관련이 있는 협업, 데이터 공유, 상호작용 등의 특징을 보이고 있다. 따라서, 기업들은 기업 내부와 외부에 존재하는 전산 시스템들의 상호연동 문제에 대해 점차 관심을 갖게 되었다. 이러한 전산환경의 급격한 변화에 의해 분산 어플리케이션 개발과 이용에 대한 새로운 요구가 발생했다. 오늘날 윈도우 NT, UNIX, J2EE, MICROSOFT .NET 등과 같은 특정 플랫폼용의 어플리케이션 운용을 위한 호스팅 환경을 제공하고 있다. 이러한 플랫폼들이 제공하는 기능들은 통합 자원관리 기능들로부터 데이터베이스 통합, 클러스터링 서비스, 보안, 작업 부하 관리, 문제점에 이르기까지 다양하지만 서로 다른 플랫폼 상에서는 이들 기능과 관련된 구현방법, 의미적 형태, API등은 서로 다르다. 이러한 다양성에도 불구하고 소프트웨어, 하드웨어, 인적자원들이 지속적으로 분산화 됨으로써 원하는 서비스품질(QoS)을 달성하는 것이 핵심적인 사항이 되었다. 따라서 분산화된 광역네트워크 환경에서 어플리케이션들이 자원과 서비스들에 효율적으로 액세스하여 공유할 수 있는 새로운 개념이 필요하게 되었고 이에 따라 그리드(Grid)기술이 개발되기 시작되었다[1]. 그리드는 전산분야에서 새롭게 부각되는 하나의 중요한 기술로서, 대규모 자원공유, 혁신적인 어플리케이션, 고성능 등에 초점을 맞추고 있다는 측면에서 전통적인 분산 전산환경과는 구분된다. 본 논문에서는 그리드의 대표적인 미들웨어인 *Globus*를 이용하여 분산된 생물학 데이터베이스 쉽게 접근하고 적절한 알고리즘을 통해 해석 할 수 있는 어플리케이션을 제시할 것이다.

2. 관련연구

2.1 그리드 정의 및 구조

그리드(Grid)는 지리학적으로 분산되어 있는 고성능 컴퓨팅 자원을 네트워크로 상호 연동하여 조직과 지역에 관계없이 사용할 수 있는 환경을 말한다. 그리드는 그림1과 같이 그리드 기반요소, 그리드 미들웨어, 그리드 개발환경, 그리드 애플리케이션으로 구성된다.

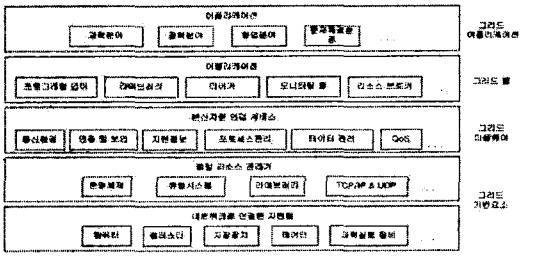

그림1 그리드 구조

그리드 기반요소는 개별적인 소프트웨어와 하드웨어를 하나의 통합된 자원으로 연결시키는 작업이 필요하다.

그리드 미들웨어는 사용 가능한 자원을 사용자에게 하나의

시스템처럼 보이도록 하는 미들웨어가 구현돼야 한다. 미들웨어는 프로세스 관리, 자원들의 동시 사용, 저장장치의 접근, 정보 보안, 사용자 인증, 네트워크 QoS, 자원예약 등과 같은 핵심 서비스를 제공한다. 미들웨어 구축이 가장 어렵고 핵심이라 할 수 있다. 현재 이뤄지는 그리드 관련 연구가 대부분 여기에 집중되고 있다. 주로 사용되는 미들웨어로는 Globus, Legion, MOL, Apples등이 있다.

그리드 개발환경은 그리드에서 수행되는 애플리케이션과 인프라를 관리하기 위한 도구들을 개발 또는 기존에 있는 것을 지원해야 한다. 프로그래머가 고성능 애플리케이션을 개발할 수 있게 해주는 서비스와 전체적인 자원 상에서 수행될 계산을 관리하고 계획하는 사용자 에이전트를 제공한다.

그리드 애플리케이션은 분산된 자원을 효율적으로 활용하는 애플리케이션 개발이 필요하다. 시뮬레이션이나 거대문제와 같은 애플리케이션은 엄청난 계산 능력이 필요하고 원격지에 있는 데이터를 쉽게 접근할 수 있어야 하고 고성능 실험장비와 연동할 수 있어야 한다[2].

2.2 그리드 미들웨어

2.2.1 글로버스 툴킷

글로버스 툴킷은 그리드 서비스를 제공하는 미들웨어로서 전 세계적인 그리드 개발 과제에서 가장 많이 사용되고 있다. 이렇게 글로버스 툴킷이 널리 사용되게 된 이유는 글로버스 툴킷이 분리될 수 없는 단일 시스템이 아니라 그리드에서 필요로 하는 다양한 독립적인 서비스를 제공하고 있다.

2.2.2 글로버스의 구조

글로버스 툴킷은 크게 그리드 보안, 정보 서비스, 자원 관리, 데이터 관리 등으로 나뉘어 진다.

보안을 담당하는 부분을 GSI라고 부르며, single-sign-on을 제공하고 globus proxy를 이용한다. 사용자는 그리드환경에 한번의 인증과정을 거침으로 사용이 허용된 자원들을 사용할 수 있고 분산된 각 자원에 대한 사용자 인증은 proxy가 대신 수행한다.

두 번째로, 정보 서비스를 수행하는 요소를 MDS라고 부른다. MDS는 그리드 내에 존재하는 자원들의 상태정보를 공유하고 사용자들에게 제공하기 위한 요소로써 인터넷의 DNS와 비슷한 것이다. 정보 서비스를 위해 글로버스에서는 두 개의 서버를 제공하는데, 각 자원의 정보를 수집하는 GRIS와 수집된 정보를 통합하는 GIIS이다.

세 째로, 글로버스에서는 데이터 관리를 위해 GASS, GridFTP, Replica catalog를 제공한다. GASS는 GRAM과 밀접한 관련이 있는 요소로써 원격지에 있는 파일을 사용하여 작업을 처리하기 원하거나 원격지에서 처리한 작업의 결과를 또 다른 저장장치에 저장하고 싶을 때 사용한다. GridFTP는 그리드 내의 데이터가 대규모 대용량이란 점을 고려하는 요소이다. Replica catalog는 데이터 그리드를 위해 개발된 것으로 데이터들을 분산 저장 및 관리함으로 필요할 때에 신속하게 데이터를 사용할 수 있게 하는 기술이다.

마지막으로 글로버스 툴킷에서 자원 관리를 담당하는 부분을 GRAM이라 부른다. 그림2과 같이 Globus는 자원 중개자, 동시 할당자, 지역 자원 관리자로 이루어져 있다. 전체적인 요청을 간략히 살펴보면 응용 프로그램에서의 자원 요청은 중개자에게 전달되고, 각 특성에 맞는 중개자는 정보 서비스(Information Service)를 참고하여 자원 요청을 특정 지역 자원을 구별할 수 있을 때까지 세분화시킨다. 이 과정에서 다수의 중개자가 참여할 수 있으며 세분화를 위해 다른 중개자에게 요청을 전달하기도 한다. 이렇게 세분화된 요청이 다중자원을 요구할 경우, 자원 요청은 동시 할당자에게 전달되며 동시 할당자는 자원 요청에 명시된 지역 자원 관리자에게 전달한다. 지역 자원 관리자는 요청 받은 자원에 대해서 프로세스를 생성하고 그에 해당하는 job handler를 반환한다. 자원 요청을 명시하고 각 요소간 의사를 소통하는 수단으로는 RSL이 사용된다. 그 외 자원 중개자(Resource Broker)는 응용 프로그램으로부터 받은 추상적인 요청을 구체적인 요청으로 변환하는 일을 한다. 즉, 자원 중개자는 자원에 대한 정보를 요약하는 작업과, 자원을 선택하는 작업등을 수행하게되는데 이러한 역할을 수행하기 위해 정보 서비스의 한 형태인 MDS(Metacomputing Directory Service)에게 자원에 대한 정보를 요청한다. 또한 자원 동시 할당은 여러 자원을 동시에 요청할 경우에 자원 중개자에 의해서 다중요청이 만들어지고 동시 할당이 이루어진다. 마지막으로 지역 자원 관리자 GRAM (Globus Resource Allocation Manager)은 지역 사이트의 자원 관리자들과 광범위 메타 컴퓨팅 환경 사이에서 인터페이스 역할을 수행한다고 보며, 사용자와 자원과의 상호 인증, 지역 사용자 관리, Job Manager의 실행 등의 역할을 담당하는 Gatekeeper와 실제 프로세스를 생성하는 역할을 담당하는 Job Manager, 스케줄러의 현재 상황이나 정보를 나타내는 MDS의 구성요소로 구성되어 있다[3][4].

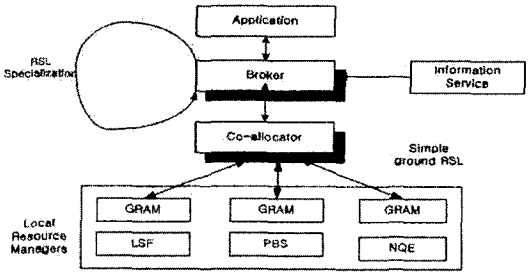

그림2 Globus 자원 관리 구조

3. 그리드 바이오 시스템의 설계

3.1 설계목적

이 시스템의 목적은 bio-data, 여러 알고리즘 프로그램 패키지 즉, Clustal W(sequence alignment), EMBOSS, GCG (sequence analysis)등을 그리드 서비스의 미들웨어인 Globus를 통해 무수한 정보 디렉토리들을 쉽게 찾을 수 있는 정보를 제공하고, 각각의 정보 및 데이터의 이동을 가능하게 하는 것이다[5][6].

3.2 그리드 인증

그리드 인증은 기본적으로 Globus 툴킷의 GSI(Grid Security Infrastructure)라는 부분에서 그리드 보안에 필요한 여러 기능을 구현하고 있다. GSI는 공개키 기반구조(PKI)와 SSL(Secure Socket Layer)프로토콜을 이용해 구현되었다. 그림3과 같이 사용자는 자신의 키와 인증서를 이용하여 사용자 프록시(user proxy)라는 것을 만들고, 이것을 통하여 원격지에 접근하게 된다. 이에 본 시스템에서는 myproxy를 이용하여 그리드 서버에 접근한다[3].

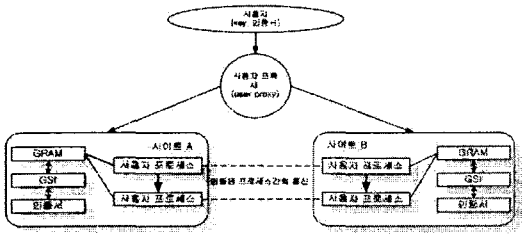

그림3 GSI 인증

3.3 Grid Job

윈도우 환경에서 글로버스에서 제공하는 그리드 컴퓨팅의 서비스를 접근할 수 있는 Cog(Commodity Grid) kit을 이용하여 인증된 그리드 서버에 접근하여 grid-job을 실행할 수 있다. 이를 통해 바이오 인포메틱스 (Clustal W, EMBOSS, GCG)등의 프로그램 명령을 실행 할 수 있다.

3.4 디렉토리 및 URL copy서비스

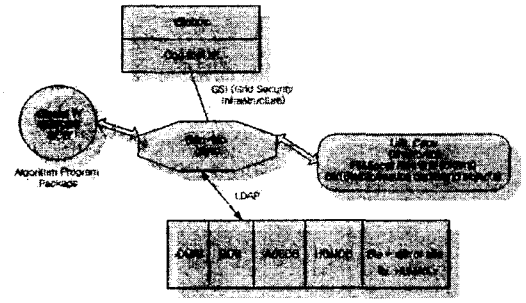

그림4 서비스 구조

그림4는 분산된 데이터를 FTP, GSIFTP(Grid Secure File Transport)등을 통한 LDAP(Lightweight Directory Protocol) 프로토콜을 이용하여 관심이 있는 바이오 데이터와 이것을 분석하기 위한 필요한 소프트웨어(Clustal W, EMBOSS, GCG)를 찾을 수 있도록 한다. 또한 URL copy 서비스를 통해 특정한 지역이나 온라인 상에 있는 데이터를 복사한다. 이는 Web(http://)과 directory(ldap://)의 서로간의 복사를 가능하게 하기 위함이다. 그 예로 biosequence data의 BLAST를 Bio-mirror나 다른 anonymous ftp 소스를 이용하여 est human.z 파일을 가져오게 되고 이 URL copy를 통하여 est human.z 파일의 주소와 소스를 볼 수 있다.

3.5 전체 구조도

Distributed grid computer(globus)는 Directory servers의 databank에서 필요한 데이터를 수집하고 Grid-bio Client는 리소스 자원 할당, Job 명령, User query를 Grid computers에게 요청하면 bio-sequence 프로그램을 통해 분석하여 결과를 얻을 수 있다. 전체적인 설계의 구조는 그림5와 같다.

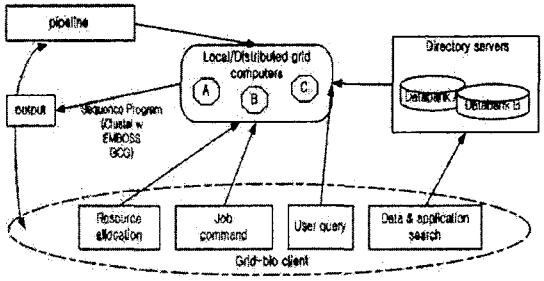

그림5 전체 구조도

4. 결 론

본 논문에서는 바이오 인포메틱스 분야에서 기하급수적으로 증가하고 있는 분산된 바이오 데이터들을 그리드 시스템을 통해 접근하고 적절한 바이오 데이터 소프트웨어를 통해 결과값을 얻을 수 있도록 설계하였다. 앞으로 바이오 인포메틱스 분야에서도 그리드 환경을 선호 할 것이며, 분산된 바이오 데이터베이스들의 일관된 포맷의 유지와 데이터의 접근성과 보안성의 보장, 효율적인 해석을 위한 통일된 방안이 제시되어야 할 것이다.

참고문헌

[1] I. Foster and C. Kesselman (eds.) "The Grid: Blueprint for a new Computing Infrasturcture" Morgan Kaufmann Publishers, 1998.

[2] Mark Baker, Rajkumar Buyya, and Domenico Laforenza, "Grids and Grid Technologies for Wide-Area Distributed Computing", Sub-mitted to the Software: Practice and Experience(SPE) Journal.

[3] The Globus Project, http://www.globus.org

[4] I. Foster and C. Kesselman, "Globus: A Metacomputing Infrastructure Toolkit," International Journal of Super computer Application, vol.11, no.2, 1997.

[5] Bioinformatics Web, http://bio.indiana.edu/