웹기반 대용량 계산환경 구축 및 응용연구

Development of Web-based High Throughput Computing Environment and Its Applications

정 민 중† 김 병 상*
Jeong, Min-Joong Kim, Byung-Sang
(논문접수일 : 2007년 2월 26일 ; 심사종료일 : 2007년 5월 24일)

요 지

공학 및 과학문제 해석을 위해 적용되는 전산 시뮬레이션은 다양한 변수 혹은 데이터의 변화를 통해 다수의 작업을 생성하고 계산함과 동시에, 생성된 결과를 비교 분석하기 위한 필수적인 기법이다. 본 연구에서는 그리드 컴퓨팅을 활용하여 웹상에서 대용량의 전산 시뮬레이션이 가능한 시스템을 개발하고, 이를 이용한 2가지 실제 응용사례를 제시한다. 첫 번째 응용사례는 e-AIRS(Aerospace Integrated Research Environment)라 명명된 연구포탈이다. e-AIRS는 수치해석 연구자가 대규모의 전산 해석을 실시하고, 실험 연구자가 원격지에서 실험을 요청하고 그 결과를 모니터링 할 수 있는 e-Science 연구환경을 제공한다. 두 번째 응용사례는 대규모 계산환경을 이용한 단백질 구조설계를 제시한다. 제안된 계산환경을 이용하여 생성된 단백질 전산 예측구조와 자연상태 구조를 비교하고, 제안된 계산환경의 유용성을 검토한다.

핵심용어 : 이사이언스, 대용량계산, 계산시뮬레이션, 이에어스, 단백질구조

Abstract

Many engineering problems often require the large amount of computing resources for iterative simulations of problems treating many parameters and input files. In order to overcome the situation, this paper proposes an e-Science based computational system. The system exploits the Grid computing technology to establish an integrated web service environment which supports distributed high throughput computational simulations and remote executions. The proposed system provides an easy-to-use parametric study service where a computational service includes real time monitoring. To verify usability of the proposed system, two kinds of applications were introduced. The first application is an Aerospace Integrated Research System (e-AIRS). The e-AIRS adapts the proposed computational system to solve CFD problems. The second one is design and optimization of protein 3-dimensional structures in structural biology.

Keywords *: e-Science, high throughput computing, computational simulation, e-AIRS, protein structure*

1. 서 론

현재 세계 각국에서 진행 중인 e-Science 환경구축은 분산된 컴퓨터 자원, 데이터저장 기술, 전문가들의 지적협업을 네트워크상에서 그리드 기술을 기반으로 제공하는 과학기술 연구의 새로운 전략이라 할 수 있다. 이러한 e-Science 환경 구축과 동시에 이를 이용한 천체물리 및 입자물리 연구, 생물학 및 화학 연구, 공학 및 지구환경 연구, 의료 분야 등의 연구가 전 세계적으로 활발히 진행되고 있다. 본 연구에서는 e-Science 환경 중 분산된 컴퓨터를 이용하여 웹상에서 대규모의 전산 시뮬레이션이 가능한 시스템을 개발하고, 이를 이용한 두 가지 실제 응용사례를 제시한다.

첫 번째 응용사례는 e-AIRS(Aerospace Integrated Research Environment)라 명명된 연구포탈(고순흠, 2006)이다. e-AIRS는 수치해석 연구자가 대규모의 전산 해석을 수행할 수 있는 환경과 실험 연구자가 원격지에서 실험을 요청하고

† 책임저자, 한국과학기술정보연구원 e-Science 응용연구팀 선임연구원 공학박사
Tel: 042-869-0632 ; Fax: 042-869-0789
E-mail: jeong@kisti.re.kr
* 한국과학기술정보연구원 e-Science 기술개발팀 연구원

• 이 논문에 대한 토론을 2007년 8월 31일까지 본 학회에 보내주시면 2007년 10월호에 그 결과를 게재하겠습니다.

그 결과를 모니터링할 수 있는 시스템을 제공한다. 두 번째는 대규모 계산환경을 이용한 단백질 구조설계 사례를 제시한다. 제안된 계산환경을 이용하여 생성된 단백질 전산 예측구조와 자연상태 구조를 비교하고, 제안된 계산환경의 유용성을 검토한다.

2. 그리드 컴퓨팅과 대용량계산 환경

계산과학의 영역에서 과학적 현상의 분석을 위해 적용되는 전산 시뮬레이션은 하나의 파라미터 혹은 입력데이터를 가지는 단일 작업실행이 아닌 다양한 변수 혹은 데이터의 변화를 통해 다수의 작업을 생성하고 계산함과 동시에 생성된 결과를 비교 분석하기 위한 필수적인 기법이다. 이와 같은 대용량 데이터 혹은 작업의 처리(high throughput computing: HTC) 및 동시 시뮬레이션 환경을 지원하기 위하여 다수의 PC 클러스터를 이용한 계산환경은 매우 유용하게 사용되어져 왔다. 하지만 PC 클러스터기반의 계산환경은 고정된 계산노드에 의한 정적인 스케줄링만이 가능했으며, 특정 지역에 국한되어 확장성의 제한이 있다고 할 수 있다. 그리드 컴퓨팅(Foster, 1998) 혹은 그리드 서비스(Foster, 2001)는 공간적인 한계를 극복하고 인터넷을 통하여 지역적으로 떨어져 있는 다수의 슈퍼컴퓨터, 클러스터 등을 유기적으로 통합하는 것을 가능하게 하였다. 하지만 그리드 환경에서의 대용량 처리 환경(HTC)은 다수의 사용자와 다양한 목적에 따라서 작업을 동적으로 생성, 할당해야할 필요가 있기 때문에 기존의 클러스터기반의 작업 실행환경과는 많은 부분에 있어서 효율성의 재고가 필요하다(Berman, 2003).

Nimrod-G/O(Abramson, 1995), Condor(Raman, 1998) 등에서는 그리드 환경에서 제안된 파라미터 변형 및 작업 스케줄러를 위한 다양한 기법을 제안하고 있다. 하지만 상기의 시스템들은 주로 지역적으로 제한되어 있는 클러스터환경에서 진화되었기 때문에 그리드와 같은 원거리간의 작업을 실행하는 환경에서는 적절하지 않다. 또한 상기의 시스템은 일반적인 사용자들보다는 전문적인 지식을 통하여 복잡한 내부 구조를 이해함으로써 작업의 실행이 가능하기 때문에 사용자 친화적인 실행 환경을 제공하지 않는다. 이와 같은 한계를 극복하기 위해 표준화된 컴퓨팅(Jones, 2005)기법에서 제안하는 것과 같이, 표준 웹 서비스 프로토콜(SOAP)을 사용하여 서로 다른 목적의 사용자 혹은 다양한 클라이언트의 지원을 지향하는 서비스 지향 구조(Service Oriented Architecture))가 많이 제시되고 있다.

2.1 과학용 어플리케이션 가상 실행 환경

본 논문에서 제안하는 어플리케이션 가상 실행 환경(Virtual Computing Environment for Scientific Application:VCE)은 다양한 과학용 어플리케이션을 생성하고 각각의 파라미터 정보를 획득하여 목적에 부합하는 다수의 작업 생성을 가능하게 한다. 또한 작업 생성 및 실행에 필요한 작업 저장, 스케줄링, 모니터링 및 결과 데이터의 획득과 같은 모든 제반 사항은 하위의 그리드 인프라스트럭처의 기능을 통하여

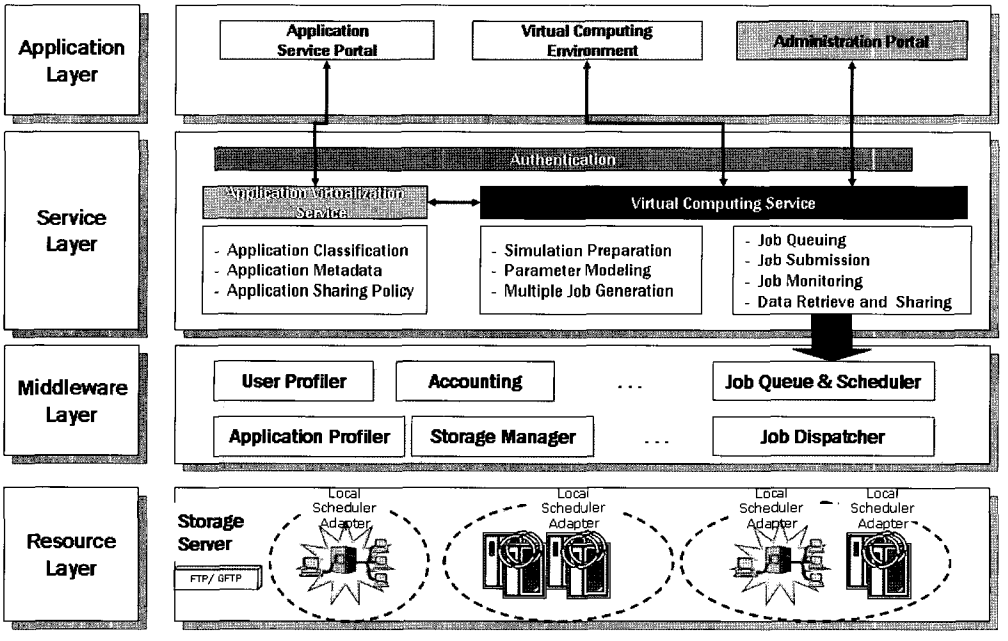

그림 1 Layered Architecture for Virtual Computing Environment.

일괄적으로 관리되도록 하고 있다.

그림 1은 VCE의 계층별 구조도를 보여주고 있다. 제안된 구조는 크게 네 개의 수평 구조도를 가지고 있으며, 상위의 계층은 하위의 계층과의 통신만을 통하여 보다 하위의 서비스들을 제공 받을 수 있도록 되어있다. 각 계층별 기능을 분류하면 아래와 같다.

- *Resource Layer* - 계산, 저장 및 장비와 같은 물리적 수준의 자원을 의미한다. 자원들은 인증 및 권한제약하에 미들웨어가 접근할 수 있는 기반을 제공해주며 작업의 실행 및 데이터의 전송이 행해지는 계층이라고 할 수 있다.
- *Middleware Layer* - 상위의 서비스들을 가능하게 하는 필수적인 기능들을 수행한다. 작업을 실행하기 위하여 하위의 물리적 자원을 관리하고 작업의 완료시점까지 작업의 상태를 모니터링한다. 또한 대용량 데이터의 전송을 가능하게 하는 GridFTP(Foster 2006)기능을 지원하고 있다.
- *Service Layer* - 서비스 계층에서는 상위의 어플리케이션 계층에서 요구하는 정보를 직접적으로 제공해주는 서버 혹은 서비스 역할을 담당한다. 서비스 계층은 어플리케이션 계층에서 필요로 하는 정보의 생성 및 작업 생성, 작업 실행 연계 등, 가장 핵심적인 부분을 담당하고 있다. 서비스 계층은 SOAP와 같은 웹 서비스 표준 통신 규약을 기반으로 하여 구현 및 서비스되며, 상위의 다양한 어플리케이션의 적용이 가능하다.
- *Application Layer* - 어플리케이션 계층은 수행하고자 하는 어플리케이션의 메타 정보를 제공하고 그 정보를 기반으로 작업을 실행 시킬 수 있는 가상 컴퓨팅 환경을 제공하고 있다. 사용자는 하위의 가상 컴퓨팅 서비스를 통해 작업 생성, 실행, 관리 및 데이터 획득과 같은 기능을 수행할 수 있다.

다음 절에서는 어플리케이션 계층과 서비스 계층의 핵심 기능들을 구체화하여 기술한다.

2.2 어플리케이션 서비스 포탈 및 어플리케이션 가상화 서비스

어플리케이션 서비스 포탈은 어플리케이션 사용자들이 자신이 원하는 과학용 어플리케이션을 정의하고 등록하여 최종적으로 사용할 수 있도록 필요한 정보를 사전에 등록해 놓음으로서 시간과 장소에 구애 없이 쉽게 웹 환경에서 작업을 생성할 수 있도록 한다. 특히 어플리케이션의 파라미터를 변형함으로서 다수의 작업을 생성하고 파라미터의 변형을 통한 결과 데이터의 특성 분석이 가능하도록 한다.

어플리케이션 가상화는 사용자의 어플리케이션을 단일화된 스크립트 언어를 통하여 정의할 수 있도록 지원하며, 목적 및 영역에 따라 어플리케이션을 분류하고 보안 및 공유 정책에 따라 사용자들에게 선택적으로 제공해주도록 할 수 있다. 제안된 시스템은 그리드 환경의 표준 작업 기술 언어인 JSDL(Anjomshoaa, 2005)에 기반하며 어플리케이션의 메타 정보를 저장한다. 하지만 JSDL은 각각의 어플리케이션에 입력값이 되는 파라미터를 세부적으로 기술하고 파라미터의 변화를 통하여 다수의 작업을 생성할 수 있는 어휘를 제공하지는 않는다. 따라서 본 시스템에서는 PDL(Parameter Description Language)를 추가하여 어플리케이션의 정보를 획득하도록 하였다. PDL은 표준 JSDL의 확장을 통하여 구현되었으며 XML schema로 제공되고 있다. 파라미터 정의 언어(PDL)에 대한 이론적 전개는 참고문헌(Kim, 2007)에 구체적으로 기술되어 있다.

그림 2에 가상 컴퓨팅 서비스의 내부 모듈 및 기능성을 나타내었다. 가상 컴퓨팅서비스는 작업 실행 및 데이터 획득을 위하여 하위 미들웨어 계층의 글로버스 툴킷-4의 WSGRAM (Foster 2006) 및 GridFTP(Foster 2006)의 통신 규약을 사용하고 있다.

아래에 가상 컴퓨팅 서비스의 세부 모듈의 명칭 및 기능을 간략히 기술한다.

- *Virtual Computing Environment(VCE)* - 어플리케이션 호출을 통한 작업 생성을 하는 가상 컴퓨팅 클라이언트 환경을 제공
- *Application Virtualization Server(AVS)* - 어플리케이션의 실행에 필요한 메타 데이터 저장 및 제공
- *JP(Job Parser)* - 작업 생성 모듈로서 VCE에서 제공되는 JSDL을 전개(Parsing)하여 하위의 미들웨어가 이해할 수 있는 RSL(Resource Specification Language)로 변형
- *Parameter Sweeper* - JP의 내부 모듈로서 JSDL 전

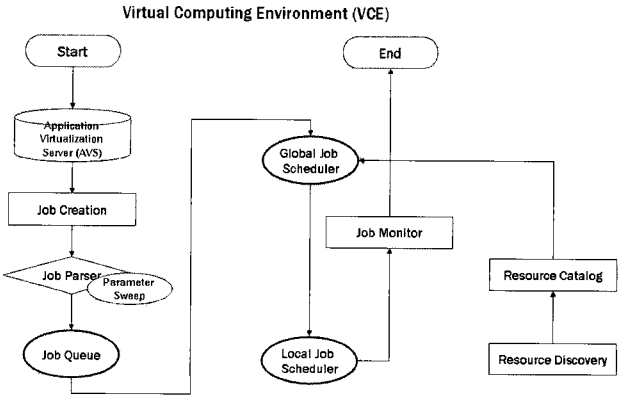

그림 2 Virtual Computing Service and its Functional Modules.

개시 사용자 어플리케이션 파라미터의 변위를 획득하여 자동 변환(증,감)을 통하여 다중 작업을 생성

- *JQ(Job Queue)* - 작업 저장을 위한 큐로서 실행전 모든 작업을 저장하는 기본 저장소
- *JM(Job Monitor)* - 실행중인 작업의 리스트를 저장하며 작업의 상태를 완료시점까지 모니터링
- *GJS(Global Job Scheduler)* - 제한된 자원의 범위내에 Job Queue에 작업 스케줄링 및 실행 요청
- *RC(Resource Catalog)* - 글로벌 작업 스케줄러에 가용한 자원의 상태 정보 제공
- *RD(Resource Discovery)* - 작업 실행을 위한 자원 확보

3. 대용량 계산시스템을 이용한 실제 응용 사례

본 장에서는 계산 환경을 이용한 실제 응용연구 사례를 소개한다. 첫 번째 응용연구 사례는 e-AIRS (Aerospace Integrated Research Environment)라 명명된 항공분야 연구포탈이고, 두 번째는 단백질의 구조설계 사례이다. 이러한 두 가지 사례는 매우 상이한 분야이나 다양한 파라미터에 대한 전산해석이 필요하고, 이에 따라 대용량 계산 환경에서 수행될 경우 더욱 연구생산성이 향상될 수 있다는 점에서는 공통점을 갖고 있다.

3.1 e-AIRS(Aerospace Integrated Research Environment)

슈퍼컴퓨팅 및 분산 컴퓨팅 기술의 발전은 전산 해석의 규모 증대로 이어져 통합된 항공우주 형상 전체에 대한 해석을 가능하게 하였고, 이는 항공우주 분야에서 전산 해석의 중요성을 높이는 결과로 이어졌다. 한편 실험 분야에서는 이와 같은 컴퓨터 및 네트워크 기술의 발전이 쉽게 접목되지 못했는데, 항공우주 실험 장비인 풍동의 경우 대상 물체의 거대성 등으로 인해 전자화 및 자동화의 여지가 상대적으로 부족했기 때문이다. 이러한 점들을 극복하기 위하여 e-AIRS(Aerospace Integrated Research Environment)라 명명된 연구 환경은 수치해석 연구자가 거대 규모의 전산 해석을 수행할 수 있는 환경과 실험 연구자가 원격지에서 실험을 요청하고 그 결과를 모니터링할 수 있는 시스템을 제공한다. 사용자의 수치해석

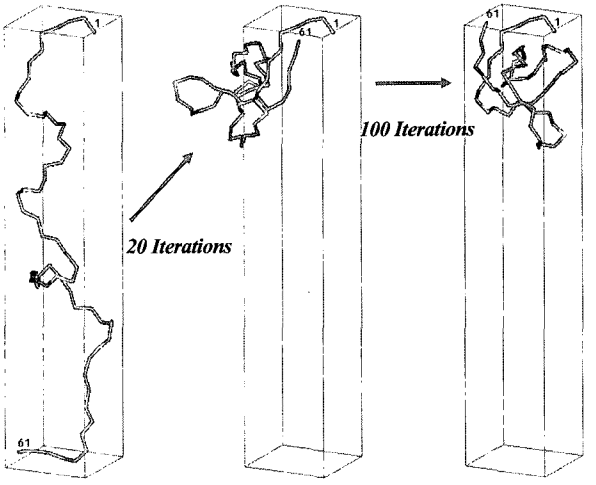

그림 4 An example of protein structure optimization

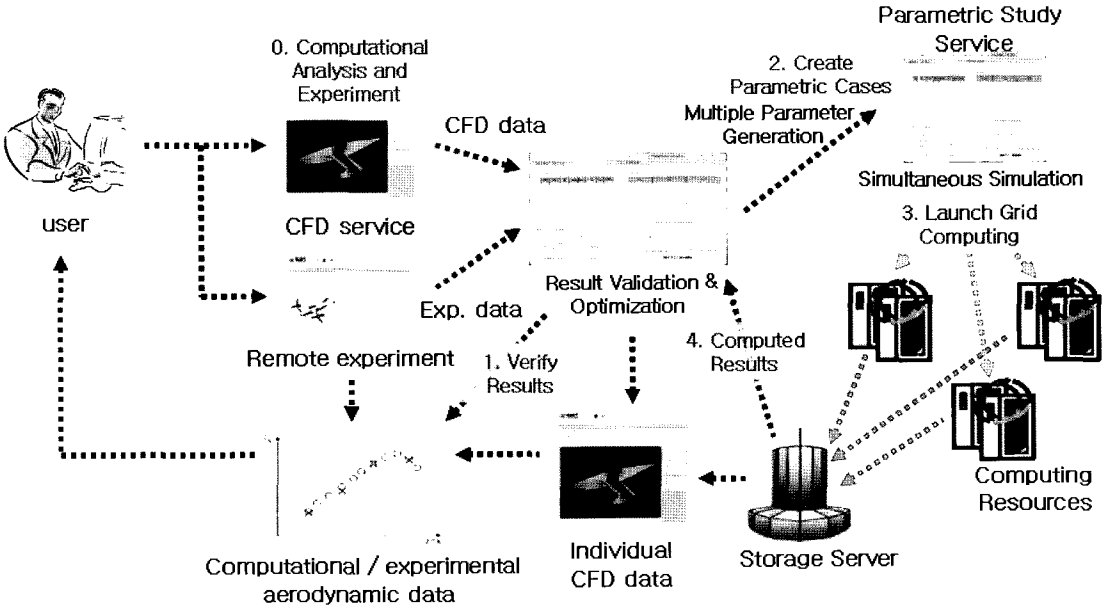

그림 3 An execution environment with parameter sweep engine for parametric applications

과 관련하여 e-AIRS에서는 격자 생성 서비스, 자동 병렬 해석 서비스, 해석 모니터링 및 가시화 서비스를 제공한다. 이로부터 사용자는 웹에 접속하는 것만으로 거대 규모의 컴퓨팅을 수행할 수 있다. 그림 3에 개발된 e-AIRS에서의 대용량 계산처리를 나타내었다. 그림 3과 같이 대용량계산 환경은 원격실험 환경과 같이 통합화 되어 개별 연구자들에게 e-AIRS 웹포탈 상에서 제공된다. 또한 연구자들은 본 연구 환경과 연계된 원격 회의 시스템을 활용함으로써 다른 연구자 및 그룹과 토론 및 공유가 가능하고, 이를 통해 연구 범위의 확장과 효율성 증대를 이룰 수 있다.

3.2 단백질 3차원 분자구조 설계

정민중(2006)은 단편조립법과 접힘최적화를 이용하여 자연 상태의 단백질 3차원 구조에 근접한 시뮬레이션 기반의 구조를 제안한 바 있다. 그러나 이러한 방법은 수만개에 이르는 초기구조 중에서 최적설계의 대상 구조를 일부 선택하여 구조 시뮬레이션을 수행한다. 그러나 계산자원에 대한 자유로운 접근 및 조작이 가능하여 수만개의 초기구조 전체를 최적화 할 경우, 보다 자연 상태에 유사한 분자구조를 설계할 수 있을 것이다.

본 연구에서는 대용량 계산의 사례로 특정 단백질(PDB ID 1LEB) 전체 초기구조에 대한 최적설계를 수행하였다. 그림 4에 단백질 3차원 분자구조의 최적설계 사례를 나타내었다. 총 15,081개의 최적설계 시뮬레이션을 수행하였고, 최종적으로 설계된 3차원 분자구조들은 고정된 저장디렉터리로 취합되었다. 최적설계의 우수성를 검토하기 위하여, 단백질 1LEB에 대한 최적설계 구조와 자연 상태 1LEB의 3차원 구조편차 RMSD(Root Mean Square Deviation)가 조사되었다. 계산된 구조들은 자연 상태와 비교하여 최저 RMSD 5.47Å을 나타내었다. 1LEB와 같은 구형 단백질의 경우 전산설계구조와 자연 상태 구조간의 RMSD가 약 6.5 Å 이하일 경우, 전산 예측된 구조는 자연 상태의 구조와 유사한 것으로 판단된다. 정민중(2006)이 제안한 일부 초기구조에 대한 최적설계 시뮬레이션 결과가 RMSD 6.08Å임을 고려할 때, 제안된 계산환경을 이용할 경우 보다 우수한 구조를 찾아낼 수 있음을 알 수 있다.

4. 결 론

본 연구에서는 새로운 컴퓨팅 서비스로서 웹기반 대용량 계산환경을 제안하였다. 제안된 계산환경은 High Throughput Computing(HTC)이 가능한 서비스 형태로, 다양한 전산 시뮬레이션에 사용될 수 있도록 개발되었다. 실제 개발된 계산환경을 이용하여 항공우주분야의 공력해석과 생명과학분야의 단백질 3차원 분자구조 설계를 지원하였다. 제안된 계산환경은 기존의 대용량계산을 위한 복잡한 전산처리 과정을 피할 수 있게 해주며, 보다 효율적인 전산해석을 기대할 수 있도록 하는 기법이다.

제안된 계산환경은 다양한 공학적 및 과학 분야의 시뮬레이션, 특히 반복적이고 많은 양의 계산이 필요한 변수해석, 최적화의 분야에 사용될 수 있을 것이다. 현재 시험가동 및 보완을 거듭하고 있는 대용량 계산환경은 향후 KISTI(한국과학기술정보연구원)의 공공서비스 형태로 일반연구자들이 자유롭게 사용할 수 있도록 제공될 예정이다.

참 고 문 헌

고순흠(2006) e-Science 기반의 항공 우주 연구 환경 개선, 한국정보과학회 HPC 학술대회 논문집, 17(2), pp.49~56.

정민중(2006) 최적설계 기법을 이용한 단백질 3차원 구조예측, 대한기계학회 논문집 A, 30(7), pp.841~848.

Abramson, D., Sosic, R., Giddy, J., Hall, B.(1995) Nimrod: A Tool for Performing Parametised Simulations using Distributed Workstations, The 4th IEEE Symposium on High Performance Distributed Computing, Virginia.

Anjomshoaa, A.(2005) Job Submission Discription Language (JSDL) Specification, Version 1.0. Global Grid Forum, GFD-R.056.

Berman, F., Fox, G., Hey, T.(2003) Grid Computing-Making the Global Infrastructure a Reality, Willey, Hoboken, p.1060.

Foster, I., Kesselman, C.(1998) The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, San Francisco, p.701.

Foster, I., Kesselman, C.(2001) The Anatomy of the Grid:Enabling Scalable Virtual Organizations, Int. *J. Supercomputer Applications*, pp.200~222.

Foster, I.(2006) Globus Toolkit Version 4: Software for Service-Oriented Systems, Springer-Verlag LNCS 3779, pp.2~13.

Jones, S.(2005) Toward an acceptable definition of service[service-oriented architecture], Software, IEEE, 22(3), pp.87~93.

Kim, B.S.(2007) Parametric Study Service for Large-scale Scientific Simulations on the Grid, *International Conference on Advanced Communicaiton Technology*, Korea Feb. pp.12~14.

National e-Science Project.. Application Development, http://www.escience.or.kr/

Raman, R. Livny, M., Solomon, M.(1998) Matchmaking: Distributed Resource Management for High Throughput Computing, Proceedings of the Seventh, IEEE, International Symposium on High Performance Distributed Computing, pp.28~31.