과학 계산 실험을 위한 클라우드 자원을 활용한 실험 프로비넌스 모델 설계

안윤선 김윤희☆

숙명여자대학교 컴퓨터과학부

{ahnysun, yulan}@sookmyung.ac.kr

A Design of Provenance Model for Scientific Computation Experiments over Cloud Environment

Younsun Ahn Yoonhee Kim☆

Dept. of Computer Scinence, Sookmyung Women's University

요 약

과학 응용 실험 중 다양한 파라미터 값을 실행 파일에 적용하고 반복적으로 수행하여 결과를 얻어 파라미터 값에 따른 실험 결과를 비교 분석하는 파라메트릭 스터디(Parametric Study) 응용들이 다수 존재한다. 이러한 응용의 경우 복잡한 계산으로 인해 장시간의 고성능 컴퓨팅 환경을 독점적으로 사용해야 하는 등 실험에 많은 비용이 요구된다. 따라서 실험의 결과를 정규화하고 체계적으로 관리를 할 수 있다면 같은 실험에 대한 반복 실험을 배제하고 유사실험에 대한 효율적인 실험 환경 설정이 가능하다. 실험의 환경과 결과를 정규화하여 데이터를 재사용 할 수 있도록 구성하기 위해 실험 데이터에 대한 프로비넌스가 필요하다. 저장된 프로비넌스 데이터를 이용하여 같은 실험을 생략할 수 있고 비슷한 조건에서 수행수 해야할 때 요구되는 자원 환경의 예측도 가능하다. 본 논문은 프로비넌스 데이터 모델에 대해서 제안하고 실험 환경에 대한 응용 요구사항과 필요 자원, 실험결과를 효율적으로 관리 하도록 한다. 또한 프로비넌스 데이터에 대한 자원 온톨로지 정보를 활용하여 유사실험을 실시할 시 요구되는 자원에 대한 예측과 실험결과를 사전에 예측하여 효율적인 실험이 가능하도록 하고, 특히 장시간 실험에 치명적인 실패율이 적은 자원환경을 고려 할 수 있도록 하여 실험의 성공률을 높이도록 하였다.

1. 서 론

과학 응용 실험 중 다양한 파라미터 값을 실행 파일에 적용하고 반복적으로 수행하여 결과를 얻어 파라미터 값에 따른 실험 결과를 비교하고 분석하는 파라메트릭 스터디(Parametric Study) 응용들이 존재한다. 이러한 응용의 경우 효율적인 수행을 위해 실험의 결과를 정규화하여 기존의 실험을 재사용하도록 구성하는 프로비넌스가 필요하다.

프로비넌스는 실험 결과의 근원정보를 기록하는 것이며 이전 실험 결과의 출처를 이용하여 같은 결과를 얻기 위해서 사용한다[1]. 프로비넌스 데이터에 작업의 환경, 작업의 이력, 작업 수행 동안의 정보를 기록한다.

저장된 프로비넌스 데이터를 이용하여 같은 실험을 다시 수행할 때 비슷한 조건에서 수행할 수 있도록 하며 다음 실험을 효율적으로 수행 할 수 있게 한다.

클라우드 환경에서는 다양한 클라우드 자원을 필요한 만큼 빌려 쓸 수 있다. 기존에 수행한 실험의 프로비넌스 데이터를 이용하여 응용에 적합하고 효율적인 클라우드 자원을 선택하여 작업을 수행 할 수 있다.

이를 위해 작업 수행 환경인 클라우드 자원의 온톨로지를 구축하고 프로비넌스 데이터에서 사용된 클라우드 자원에 대한 부분을 가지고 온톨로지를 생성한다. 온톨로지는 지식 기반이 구축될 수 있는 기본 구조를 제공하는 도메인 개념들의 명확한 표현으로 정의된다[2]. 온톨로지를 기반으로 자원을 표현하면 자원의 특성을 개념화 하고, 의미 있게 그들 사이의 관계를 나타낼 수 있다.

생성한 온톨로지로 다음 같은 실험을 수행할 때 클라우드 자원의 온톨로지와 비교하여 유사도가 높은 자원을 선택하여 작업을 보다 빠른 시간 내에 작업의 실패율을 낮추어 효율적으로 수행 할 수 있다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 관련 연구들을 살펴보고, 3장에서는 본 논문에서 제안하는 프로비넌스 데이터 모델에 대해 설명한다. 4장에서는 실험에 대해 살펴본다. 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

실험을 효율적으로 수행하기 위해서는 이전 실험의 결과를 가지고 실험을 모든 사람에게 의미 있게 일정한 형태로 만드는 것이 필요하다. 실험의 정규화를 위해

☆ 교신저자

"이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임"(NRF-2013R1A1A3007866)

필요한 프로비넌스에 대한 연구들이 이루어졌다.

[1]에서는 프로비넌스를 두 가지의 포맷으로 나누어 제시하였다. 프로스펙티브 프로비넌스(Prospective provenance)는 작업의 명세와 실험이 수행되는 단계를 나타내며, 레트로스펙티브 프로비넌스(Retrospective provenance)는 수행 중인 작업의 자세한 정보와 수행 환경에 대한 정보를 나타내고 있고, [3]에서는 프로비넌스 모델을 워크플로우 형태와 노드와 엣지 형태로 나타내었다. 하지만 정확히 어떠한 정보를 프로비넌스 데이터로 정의할지 나타내지 않았다.

온톨로지를 기반으로 자원을 선택한 연구는 [4]가 존재한다. [4]는 그리드 환경에서 온톨로지를 기반으로 자원 명세를 표현하고 온톨로지간 유사도를 비교하여 사용자가 요구하는 환경과 유사한 자원을 찾는 방법을 제안한다. 이를 바탕으로 클라우드 환경에서 온톨로지를 구축하고자 한다.

본 논문에서는 클라우드 환경에서 프로비넌스 데이터를 이용하여 실험을 정규화하고 자원에 대한 온톨로지를 바탕으로 이전 실험의 환경과 비슷한 자원을 선택하여 실험 결과를 예측 할 수 있고, 실험의 신뢰도를 높이고자 한다.

3. 제안한 연구

본 논문에서 제안한 프로비넌스 모델은 그림 1과 같다. 프로비넌스 모델은 반복적인 과학 응용의 효율적인 수행을 위해 필요한 데이터들을 나타내고 있다. 실험을 수행할 때마다 프로비넌스 데이터가 저장되어 축척된다.

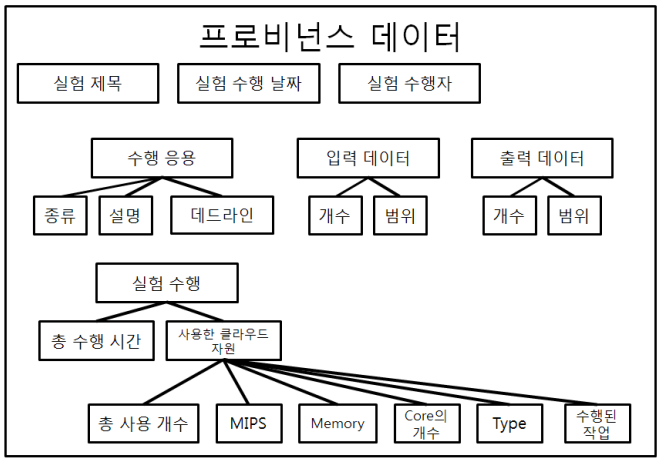

그림1 프로비넌스 데이터 모델

그림 1은 프로비넌스 데이터 모델을 나타내며, 프로비넌스 데이터에 저장되는 실험의 정규화에 필요한 정보들이 나타나있다. 실험에 대한 기본적인 제목, 날짜, 수행자의 정보가 필요하며, 수행 응용에 대한 종류, 설명, 데드라인 정보가 들어간다. 입력 데이터와 출력 데이터의 개수와 범위를 저장하며 실험 수행에서 필요한 총 수행 시간과 사용한 클라우드 자원에 대한 상세한 정보가 저장된다. 클라우드 자원의 온톨로지와 비교하기 위해 필요한, 클라우드 자원의 사용 개수, MIPS, Memory, Core의 개수, Type, 각각의 특정 클라우드 자원에서 수행된 작업 정보를 저장한다.

저장된 프로비넌스 데이터 중 사용한 클라우드 자원 부분의 정보를 가지고 온톨로지를 생성한다. 온톨로지를 생성하여 자원의 특성과 관계를 나타낼 수 있다. 클라우드 자원에 대한 명세를 온톨로지를 기반으로 표현하면 정확히 일치하지 않는 자원이라도 의미상 유사한 자원을 선택할 수 있다.

작업을 클라우드 자원에 스케쥴링 할 경우 성능이 좋고, 비싼 자원을 선택하지 않고, 프로비넌스 데이터를 가지고 구축한 온톨로지와 클라우드 자원의 온톨로지를 비교하여 자원을 선택한다. 이전에 수행한 실험의 환경과 비슷한 패턴, 비슷한 스펙, 유사한 결과를 얻을 수 있는 자원을 선택하여 작업을 클라우드 자원에 할당한다.

4. 실험

본 논문에서의 실험은 파라메트릭 스터디 응용 중 하나인 전산유체역학 (CFD, Computational Fluid Dynamics) [5] 응용을 대상으로 실험을 진행하였다. 전산 유체 역학은 복잡한 유동 현상에 대해 수치 해석 기법을 사용하여 유체 유동 문제를 풀고 해석하는 응용으로 같은 계산을 파라미터 또는 입력 파일을 바꿔가면서 반복적으로 수행한다. 전산 유체 역학 응용은 e-AIRS 2.0[6]을 이용하여 분석한 응용의 평균 길이를 실험에 사용하였다. 응용의 작업의 평균 길이는 660,000MI(Million Instruction)로, 작업의 길이를 정규분포를 이용하여 랜덤 하게 생성하여 수행하였다.

본 실험에서는 하이브리드 클라우드 환경에서 MIPS의 범위가 100~2000인 클라우드 자원을 사용하여 5000개의 전산유체역학 응용 작업들을 여러 번 수행하여 결과를 얻었다.

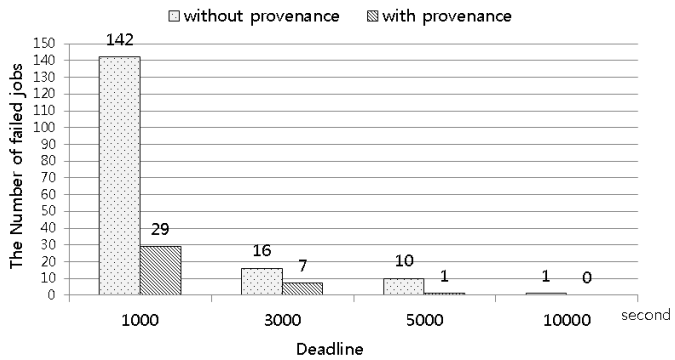

그림 2 프로비넌스 데이터를 이용한 결과와 프로비넌스 데이터를 이용하지 않았을 때 수행 실패한 작업 개수 비교

그림2는 전산유체역학 응용을 클라우드 자원에 할당하여 수행한 결과이다. 프로비넌스 데이터를 이용하지 않았을 때와 프로비넌스 데이터를 이용하여 여러 번 수행한 결과를 비교하였다. 데드라인의 변화에 따라 수행에 실패한 작업들의 평균 개수를 비교하였다.

프로비넌스 데이터를 이용하지 않았을 경우와 프로비넌스 데이터를 이용한 경우 모두 데드라인이 길어짐에 따라 작업의 실패 개수가 줄어든다.

프로비넌스 데이터를 이용하지 않았을 경우, 데드라인 1000초에서 5000개의 작업 중 142개를 실패하여 2.84%의 실패율이 나타나지만 프로비넌스 데이터를 이용하였을 경우에는 5000개의 작업 중 29개가 실패하여 프로비넌스 데이터를 사용하지 않았을 경우보다 낮은 0.58%의 작업 실패율을 보인다. 데드라인이 3000초인 경우에는 99.68%의 성공률과 99.86%의 성공률을 보인다. 프로비넌스 데이터를 이용하여 실험한 경우 데드라인이 5000초에서는 99.8%의 성공률을 보이며, 10000초 데드라인 내에서는 100%의 성공률을 보인다. 프로비넌스 데이터를 사용하지 않았을 경우에는 5000초의 데드라인에서는 5000개의 작업 중 10개 실패하여 99.8%의 성공률을 보이고, 데드라인이 10000초인 경우에는 99.98%의 성공률을 보인다.

프로비넌스 데이터를 이용하지 않고 응용을 여러 번 수행 했을 경우에는 다양한 클라우드 자원을 사용하여 작업의 실패 개수를 줄이지 못한다. 반면 프로비넌스 데이터를 이용한 경우 각 작업이 수행되기 전에 해당 작업이 이전에 수행되었던 클라우드 자원의 정보를 바탕으로 작업의 수행을 예측하여 자원에 작업을 스케줄링 하게 된다. 기존에 수행했던 결과를 바탕으로 데드라인 내에서 수행을 완료 할 수 있고, 작업의 실패율을 낮출 수 있는 비슷한 스펙을 가진 클라우드 자원에 작업들을 할당하게 되어 작업의 실패율을 낮출 수 있다. 프로비넌스 데이터를 통해서 이전에 효율적으로 수행한 비슷한 환경에서 작업을 수행하여 실험의 안정도와 신뢰도가 높아진 것을 볼 수 있다. 데드라인이 10000초일 경우, 프로비넌스 데이터를 이용한 경우에는 실패한 작업 없이 모든 작업이 데드라인 내에서 수행을 완료하였다.

5. 결론 및 향후 연구

본 논문에서는 같은 실험을 다시 수행 할 때 비슷한 조건에서 수행할 수 있도록 프로비넌스 데이터를 이용하여 실험의 결과를 정규화하였다. 프로비넌스 데이터를 통하여 이전 실험의 데이터를 바탕으로 데드라인 내에서의 수행을 완료하며, 작업의 실패 개수를 줄였다. 이를 통해 가장 성능 좋고, 비싼 클라우드 자원에서 수행하는 것 보다 이전 실험을 바탕으로 예측 가능하고, 신뢰도를 높일 수 있는 효율적인 클라우드 자원을 선택하여 작업을 수행 할 수 있다.

향후에는 프로비넌스 데이터의 클라우드 자원 온톨로지와 클라우드 자원의 온톨로지의 유사도 계산 부분을 추가할 예정이다.

참고 문헌

[1]Davidson, Susan B, and Juliana Freire, "Provenance and scientific workflows: challenges and opportunities.", Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, pp.1345-1350, 2008.

[2] Swartout W, Tate A, “Guest editors’ introduction: Ontologies.”, IEEE IntelligentSystems, 14(1), pp.18-19, 1999

[3] Simmhan, Yogesh L, Beth Plale, and Dennis Gannon, "A framework for collecting provenance in data-centric scientific workflows.", In *Web Services,* ICWS'06. International Conference on IEEE, pp.427-436, 2006.

[4] 김지영, 김윤희, "온톨로지 기반 그리드 자원 선택 기법", 한국정보기술학회논문지, 7(4), pp.1-9, 2009.

[5]Computational Fluid Dynamics, http://www.cfd-online.com.

[6] Yoonhee Kim, Eun-kyung Kim, Jeu Young Kim, Jung-hyun Cho, Chongam Kim, and Kum Won Cho, "e-AIRS: An e-Science Collaboration Portal for Aerospace Applications.", HPCC LNCS (Lecture Note in Computer Science), vol.4208, pp.813-822, 2006.