소프트웨어 정의 기술 기반 실시간 HPC 응용 실행 구조 정의

*오지선, 김윤희
숙명여자대학교 컴퓨터과학과

*jsoh8088@gmail.com, yulan@sookmyung.ac.kr

Definition of Real-Time HPC Application Execution Blocks Based on Software-Defined Techniques

Ji Sun Oh, Yoonhee Kim
Dept. of Computer Science, Sookmyung Women's University

요 약

응용 실행의 최적화를 위해 소프트웨어 정의 기술을 활용한 IT 인프라를 동적으로 관리하는 적시적 자원 관리 연구가 활발하다. 데이터 집약적인 HPC 응용의 요구사항 중 실시간 분석 및 예측을 요구하는 경우 적시적 데이터 처리 및 성능과 결과 검증에 소프트웨어 정의 기술을 적용해야 한다. 본 논문은 고성능 데이터 전송 환경인 ScienceDMZ 에 대한 국내외 동향을 조사하고 이를 활용하는 실시간 HPC 응용 실행 구조를 정의한다.

1. 서론

소프트웨어 정의 기술은 IT 인프라를 동적으로 관리하여 응용 실행의 최적화를 위한 적시적 자원 관리 기술을 제공한다. 최근 서비스 특성에 따라 OpenFlow[1]와 같은 오픈소스 기반의 SDN(Software-defined Network) 기술을 적용한 동적 네트워크 관리 연구가 활발하다. 이와 함께 기기들의 고성능화에 따른 데이터의 폭증, 트래픽 동적 변이 심화와 더불어 클라우드 컴퓨팅의 확산과 서비스의 다양성을 고려한 소프트웨어 중심의 관리가 중요해지고 있다. 이러한 대규모의 데이터를 처리하는 과학 계산 응용의 요구사항을 충족시키기 위해 고성능 환경을 활용한 SDC(Software-defined Compute), SDS(Software-defined Storage) 기술이 새롭게 정의되고, 관련 연구가 활발히 진행 중이다. 기존의 HPC 연구는 필요 자원이 독점적으로 활용 가능한 시점에 분석을 진행하는 일괄 배치 작업 형태를 활용하고 있다. 기상 재난 예측과 같은 과학계산 응용은 실시간에 가까운 분석 및 결과 활용을 요구한다. 그러나 이러한 과학계산 응용은 자원 및 실행 인프라의 실시간 성능 처리와 결과 검증 문제에 대한 연구가 부족한 상황이다.

본 논문은 국내외에서 진행되고 있는 고성능 데이터 전송이 가능한 ScienceDMZ 환경 연구에 대한 조사와 실시간 예측 분석의 일례로 기상 재난 예측 분석 시나리오를 소개하고 소프트웨어 사용자 정의 기술인 SDC, SDN 을 적용하여 대용량 데이터 처리가 필요한 HPC 응용 실행 구조를 정의한다.

2. 관련연구

2.1 ScienceDMZ

미국 ESnet 에서 제안된 ScienceDMZ 는 데이터 집중형 과학을 위한 통합적 전송체계이며 개념적 용어이다. TCP 환경에서의 최적화된 전송 성능의 보장을 위하여 트래픽 손실 없는 네트워크 상태를 추구한다. 망 분리는 일반적인 데이터 트래픽과 ScienceDMZ 데이터 트래픽에 의해 이루어진다. 트래픽 손실이 없는 경우 LAN 과 international 구간에 대해 Line Rate 의 90% 이상의 전송 성능을 나타낸다. 또한 고성능 데이터 전송을 위해 저장, 캐싱, 보안 기능까지 감안한 전용 노드(DTN)가 필요하다[2][3]. 이를 기반으로 고성능 데이터 전송 서비스를 확충하여 고성능 연구 환경을 제안하였다. 그림 1 은 국내 ScienceDMZ 기반의 고성능 인프라이다.

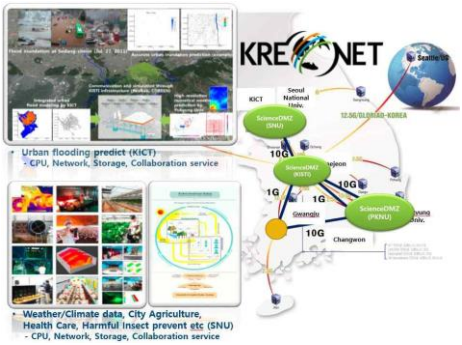

그림 1. 국내 대기과학분야 ScienceDMZ 기반의 고성능 인프라[2]

2.2 SDN, SDC

SDN 기술은 이종의 스위치와 라우터를 통해 손쉽게 네트워크를 제어할 수 있는 OpenFlow 에 기반하여 출현하였다. 소프트웨어 중심의 네트워킹 기술 개발을 위해 네트워크를 컴퓨터 시스템으로, OpenFlow 는 하드웨어와 네트워크 운영체제 사이를 연결하는 인터페이스로 하여 네트워킹 시스템 모델을 정의한다. SDN 기술은 네트워크 사용자가 하드웨어 기반 연구에서 벗어나 다양한 응용들을 손쉽게 개발할 수 있는 환경을 제공할 것이다[4].

SDC 기술은 SDN 의 개념에서 확장되었다. 데이터센터의 새로운 트랜드 기술로 데이터와 자원을 작업에 올바르게 배치하는데 있어서 작업의 동적인 수행을 고려한다[5]. 응용 실행의 최적화를 위한 자원관리 기술인 SDC 를 통해 데이터센터의 자원을 효율적으로 사용할 수 있다.

3. 기상 재난 예제 시나리오

고성능 전송, 보안, 인프라의 실시간 성능 처리 및 결과 검증 문제가 해결되지 않는 기존 환경에서 시간적인 한계를 가지고 결과를 도출해야 하는 고성능 연구는 ScienceDMZ 와 같은 인프라 구축이 필요하다. 응용의 예로는 국지성 태풍 예보와 같은 기상 재난 안전 관련 분석, 예측의 경우이다.

기상청은 정기적인 예보를 위해 과거 데이터를 바탕으로 다년간 개발된 기상모델을 이용하여 슈퍼컴퓨터와 같은 고성능 자원을 활용한 단기, 중기, 장기 예보를 실시하고 있다. 그러나 갑작스러운 기상 변화에 대한 분석 및 예측 서비스는 시간적 제약 하에 어려움이 있다. 그림 2 는 실시간 침수 예보를 가상화 하여 제공하는 인프라 구조를 나타낸다. 변동하는 기상 데이터를 위성 및 레이더로 측정하여 지형 및 광역, 국지 구름 모델을 분석한다. 분석 결과를 통해 강우 예상 데이터를 전처리하고 기상 모델에 적용하여 지역별 실시간 침수를 예측한다.

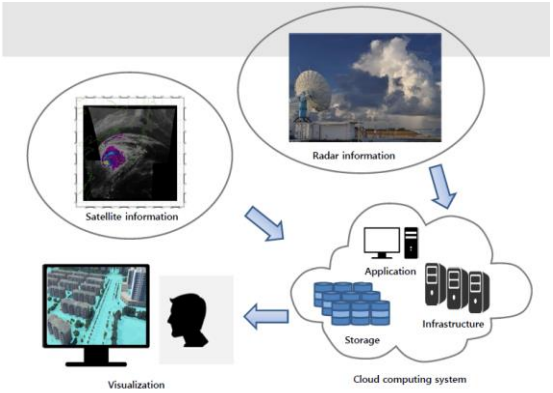

그림 2. 실시간 침수 예보 인프라[6]

이처럼 국지적 태풍 예보, 도로별 침수, 홍수 지역 예측 서비스 등과 같은 실시간 분석 및 적용 연구는 대용량 데이터의 이동, 충분한 고성능 자원, 새로운 분석 모델 개발 등을 필요로 한다. 고성능 데이터 전송이 가능한 ScienceDMZ 환경에서 데이터 저장과 계산 실험의 적시적 적용 및 통합적 연동이 가능한 상황을 제공해야 한다. 또한 응용 프로그램의 이해를 기반한 지능적 데이터 배치와 계산 배치가 우선적으로 필요하다.

4. HPC 응용 실행을 위한 소프트웨어 정의 컴퓨트(SDC) 구조

본 논문에서는 앞 장의 응용 시나리오에 적합한 소프트웨어 정의 기술을 활용하여 응용의 특성을 고려한 실시간 HPC 응용 실행 구조를 정의하였다.

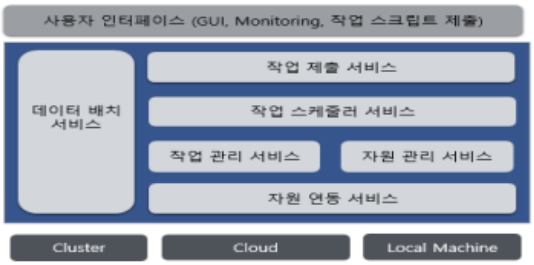

그림 3. 소프트웨어 정의 컴퓨트(SDC) 구조

그림 3 은 본 논문에서 제안하는 기능 구조도이다. 이 구조는 차세대 고성능 인프라에서 발생하는 데이터 전송 및 계산 배치의 문제를 효과적으로 해결할 수 있도록 데이터 집약적인 응용을 위한 데이터 배치 및 작업 스케줄러를 포함한다.

플랫폼의 작업 제출 및 작업 스케줄러 서비스는 응용의 특성 분석을 통해 찾아낸 각각의 특징 정보를 이용하여 세밀한(fine-grained) 자원 요구를 만족시키는 SDC 기술 기반의 작업 스케줄링 기법을 제공한다.

기상 재난과 같은 HPC 응용은 대규모의 입력 데이터, 중간 데이터와 출력 데이터를 생성한다. 데이터의 생성 시점 및 종류에 따라 사용되는 용도와 사용 빈도가 다르다. 따라서 데이터의 생성 시점과 종류를 분류하고 그 결과에 따라 관리, 사용한다. 이러한 데이터 배치 기법을 통해 응용에 사용되는 데이터 특성 이해를 기반으로 하여 지능적으로 데이터를 분배, 배치해야 한다. 데이터 집약적인 HPC 응용의 데이터와 작업 간의 관계 정보를 통해 데이터 이동을 최소화하는 데이터 배치 서비스를 제공하는 것이 한 예이다. 또한 데이터 배치 전략에서 성능 하락의 큰 요인인 네트워크 병목현상으로 인한 지연을 최소화 한다. SDN 환경에서의 데이터 움직임을 최소화하기 위해 네트워크 전송 속도 및 대역폭을 고려한 서비스를 제공한다.

5. 결론

본 논문에서는 HPC 응용 실행을 위한 소프트웨어 정의 컴퓨트 구조 대해 정의하였다. 응용의 특성과 데이터의 이동을 고려하면서 효과적인 스케줄링을 제공하여 자원의 효율성 및 활용성을 극대화 할 수 있다. 또한 ScienceDMZ 와 같은 국가적 차원의 고성능 데이터 전송체계에서 데이터배치 기법은 데이터 집약적인 HPC 응용 과학 연구를 위한 기술 발전으로 이어진다. 현재 그리고 향후 국가적으로 추진되고 있는 다양한 과학 응용 개발 사업에도 원천기술로써 적극적으로 활용이 가능하다.

ACKNOWLEDGMENT

이 논문은 2017 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017R1A2B4005681)

참 고 문 헌

[1] http://www.openflow.org/videos

[2] 문정훈, and 이민선. "데이터 집중형 (Data-Intensive Science) 과학을 위한 ScienceDMZ 기반 빅데이터 전송 연구." *한국정보과학회 학술발표논문집* (2016): 1359-1361

[3] Dart, Eli, et al. "The science dmz: A network design pattern for data-intensive science." *Scientific Programming* 22.2 (2014): 173-185.

[4] ETRI, " 미래 네트워킹 기술 SDN," 2012 Electronics and Telecommunications Trends.

[5] Nam, Yoonsung, et al. "Workload-aware resource management for software-defined compute." *Cluster Computing* 19.3 (2016): 1555-1570.

[6] Seungsoo Lee, "Development and Application of 1D-2D Coupled Urban Inundation Model for Real-time forecasting based on Remote Sensing Technique", KNOM tutorial 2016