

# 대용량 데이터 스트림을 처리하는 과학계산 응용을 위한 효율적인 데이터 이동 기법

변은규

한국과학기술정보연구원

e-mail : ekbyun@kisti.re.kr

## Efficient Data Movement for Scientific Application Processing Large Size Data Stream

Eun-kyu Byun

Korea Institute of Science and Technology Information

### 요약

대규모 실험장비에서 발생하는 아주 큰 사이즈의 데이터를 처리하기 위해서 기존에는 수집 및 저장, 계산 장비로의 원거리 전송, 데이터 분석 등의 단계를 따로 처리해 왔다. 데이터의 양이 폭발적으로 증가하고 있고 동시에 데이터의 실시간 처리 요구가 증가하는 상황이다. 이에 본 연구에서는 추상화된 입출력 계층을 이용하여 마치 로컬 저장소에 있는 데이터를 사용하는 것과 같은 인터페이스를 통해 원거리에서 생성된 데이터 스트림을 실시간으로 이동하고 처리할 수 있는 기법을 소개한다. 또한 데이터 전처리 계산 위치를 송신 측으로 변경하여 대용량 데이터를 효과적으로 전송하기 기법을 제안한다.

### 1. 서론

최근 HPC 성능이 향상됨에 따라 같은 시간엔 더 많은 연산을 수행할 수 있게 되고, 이는 처리할 수 있는 데이터의 크기 또한 크게 증가되는 결과를 낳았다. 이제 과학 응용은 고성능 컴퓨팅 노드의 메모리에 있는 데이터에의 연산을 반복하는 것뿐만 아니라, 아주 큰 데이터를 읽어오고 처리하고 저장하기 위한 고성능의 I/O 를 필요로하게 되었다. 그와 동시에 ITER(International Thermonuclear Experimental Reactor), KSTAR(Korea Superconducting Tokamak Advanced Research), NSTX(National Spherical Torus Experiment)등의 대규모의 과학계산 프로젝트들에서 사용하는 실험 기계와 측정기계들이 시간당 수~수백 테라 바이트(Tera Byte)에 이르는 아주 큰 데이터를 생성한다.

이러한 대규모 프로젝트에서 생성한 원시데이터의 특징은 한 명 혹은 한 그룹의 연구자들만이 사용하는 것이 아니라 지리적으로 멀리 떨어진 대륙 너머의 연구자들과도 공유된다는 점이다. 즉 데이터를 생성한 곳에서 분석하고 버려지는 것이 아니라, 데이터의 정제, 데이터의 저장, 데이터의 전송 등의 기능이 필요하다. 따라서 고성능의 I/O 및 네트워크 프레임워크가 필요하다.

또 다른 요구사항은 데이터의 실시간 분석을 필요로 한다는 점이다. 앞서 언급한 데이터의 경우 일련의 실험을 반복하는데, 앞서의 실험결과를 분석하여 다음 실험에 필요한 설정 값을 도출해야 효율적인 실험이 이루어지는 특징을 가지고 있다. 즉, 데이터가

생성됨과 동시에 데이터를 원거리에 있는 연구자의 위치 혹은 고성능의 컴퓨팅 자원이 있는 곳으로 데이터 분석을 빠른 시간에 끝내야 한다는 요구사항이 있다.

현재 수 페타바이트(Peta Byte)를 저장할 수 있는 스토리지 장치를 갖춘 연구기관들도 있고, 한국 미국간에 이론상 최대 10Gbps 로 데이터를 전송할 수 있는 네트워크가 구축되어 있는 등 이러한 요구사항을 해결하기 위한 최소한의 인프라는 구축되어 있다. 그러나 앞서 기술한 시나리오대로 데이터를 처리하기 위해서는 측정장비에서 측정한 원시데이터를 디지털화 및 포맷팅하여 로컬 스토리지에 저장하는 작업을 시작으로, 이 저장된 데이터를 분석을 하는 원거리 사이트에 전송하기 위한 서버/클라이언트 단계, 원거리 사이트의 스토리지에 저장하는 단계, 이 데이터를 분석 프로그램이 읽어오고 처리하는 단계 등의 여러 단계로 이루어지고 이를 위한 소프트웨어를 각각 구축해야 하는 문제가 있다. 단계가 많아짐에 따라 최적의 파이프라이닝이 이루어지지 않아 수행시간이 증가되는 문제가 생기고, 과학 계산 응용 연구자들이 신경 써야 하는 프로그래밍 및 설정 부하가 많다는 한계가 있다.

응용 과학자 입장에서는 데이터를 읽고 분석하여 결과를 저장하는 논리적인 흐름에만 집중하고, 데이터가 어디에 어떤 방식으로 이동하며 저장되는지에 대한 자세한 내용을 신경 쓰지 않는 상황이 가장 최선이다. 즉 데이터 처리와 관련된 부분을 추상화하는 프레임워크의 제공이 필요하다. 이를 이용하여 응용

과학자는 어떤 데이터를 어떤 알고리즘을 사용하는지 만 기술하면, 하부의 실질적인 데이터 저장, 이동 등을 알아서 처리해 주게 된다. 이러한 구조는 호환성이라는 또 하나의 장점을 가지고 있다. 즉, 하드웨어의 구성이 다른 곳에서도 추가적인 프로그램을 고칠 필요 없이 응용의 실행이 가능하다는 점이다. 로컬 스토리지 저장된 데이터를 분석하는 응용의 변경 없이 네트워크 너머 원거리에 있는 데이터를 처리하는데 바로 활용도 가능하다.

앞서의 시나리오를 예를 살펴보면 원시데이터를 이 프레임워크 저장하는 부분, 그리고 그 데이터를 읽어와서 분석하고 결과를 저장하는 부분만을 작성하는 것으로 완료가 되고, 이 코드는 해당 프레임워크가 존재하는 어떤 설정에서도 활용이 가능하다.

본 연구에서는 기존의 I/O 미들웨어인 ADIOS[1]의 인터페이스를 기반으로 원거리 데이터 전송 기능을 추가로 구현하고, 이를 활용하여 KSTAR의 ECEI 분석 응용[2]을 개발하고 성능을 평가하였다.

## 2. 추상화 I/O API 를 통한 원거리 데이터 전송

추상화 I/O 인터페이스를 제공하는 프레임워크를 사용하는 것은 앞서 기술한 많은 장점을 가지고 있지만, 일반적인 경우처럼 만약 기존의 시스템에 맞추어 놓은 과학 응용 코드가 이미 있을 경우, 예를 들어 POSIX 나 MPIIO 를 이용하여 작성한 경우, I/O 관련 부분의 코드 수정이 불가피하다. 이러한 이유로 새로운 API 를 정의하여 사용하는 것 보다는 기존의 인터페이스를 기반으로 프레임워크를 작성하는 것이 훨씬 나은 접근성을 제공한다. 과학계산 응용에서는 HDF5, netCDF, ADIOS 등이 많이 쓰이고 있다. 본 연구에서는 이 중 ADIOS(Adaptive IO System)의 인터페이스를 기반으로 하여 추상화된 원거리 병렬 전송 기능을 제공하는 ICEE module 을 추가로 구현한 프레임워크를 개발하였다. ADIOS 는 미국 ORNL(Oak Ridge National Laboratory)에서 개발한 I/O 미들웨어로써, 과학계산 응용들이 데이터를 효율적이고 유연한 방법으로 다룰 수 있는 추상화된 인터페이스와 성능 향상 기술을 포함하고 있다. 사용자는 로컬 파일시스템, Lustre 와 같은 분산병렬파일시스템, MPIIO, DataSpace[4]나 FlexPath[4] 같은 분산공유메모리 등의 다양한 데이터 저장공간을 코드 수정없이 XML 을 이용한 간단한 환경 설정만으로 자유롭게 넘나들며 사용할 수 있게 한다. 또한 byte 단위의 관리가 아니라 데이터의 의미에 맞게 병렬화 하여 처리할 수 있는 고수준의 파일포맷 및 인터페이스를 제공한다.

본 연구에서 개발한 기능은 데이터 스트림을 병렬로 원거리 전송을 가능하게 한다. 즉 데이터 생산자가 기존의 ADIOS API 를 사용하면 데이터 쓰기를 수행하면 다른 기관 혹은 대륙에 있는 서버에 있는 데이터 소비자, 이 경우에는 분석을 하는 연구자의 응용 프로그램, 가 기존의 API 를 이용하여 마치 로컬 하드디스크에 있는 데이터에 접근하듯이 읽어서 데이터 처리가 가능하다.

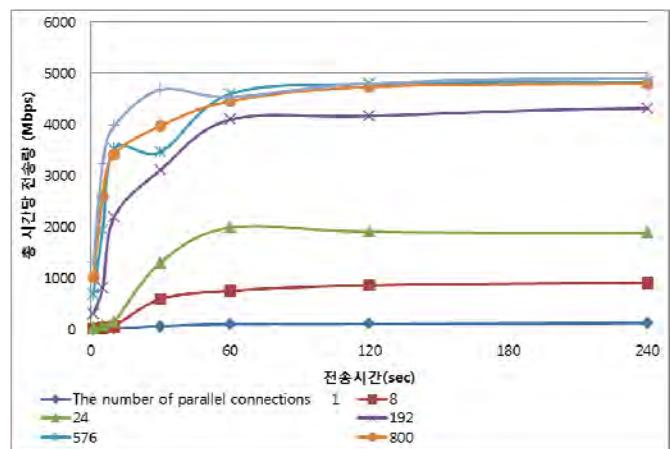
이때 데이터는 전체 데이터가 한꺼번에 전송되는

것이 아니라 순서대로 스트리밍의 형태로 보내는 것이 가능하고, 이에 따라 배치 작업의 파이프라이닝이 되어 성능향상이 가능하다.

본 연구의 큰 목적이 추상화를 통해 호환성을 최대화 하는 것이었기 때문에 ICEE module 은 다양한 네트워크 환경을 지원하기 위해서 단순한 TCP/IP 프로토콜을 사용하지 않고 유연한 네트워크 라이브러리인 EVPPath[5]를 사용하여 구현하였다. 이를 통한 당점은 Infiniband RDMA, TCP/IP 와 UDP 를 포함한 다양한 고속의 네트워크 전송 기법을 활용할 수 있다는 점이다.

ICEE module 의 또 다른 특징은 병렬 전송이다. 목표로 과학응용프로그램의 경우 병렬 프로세싱을 기본으로 하므로 기능적인 측면에서 병렬 전송이 필요하다. 보내는 쪽의 프로세스의 개수와 받는 쪽의 프로세스 개수를 자유롭게 변경하여 데이터를 나누고 각 프로세스가 필요한 부분만 전송 받을 수 있는 기능을 제공한다.

병렬화는 성능의 측면에서도 중요하다. 데이터를 보내는 쪽과 받는 쪽이 모두 국내에 있는 경우처럼 상대적으로 가까운 거리에 있을 때에는 상관 없으나 본 연구에서는 국제협력을 고려하고 있으므로 이러한 환경에서의 최적의 전송 방법이 필요하다. 이를 위해 KISTI 와 ORNL 간의 네트워크 특성을 분석하였다. KISTI 와 ORNL 사이에는 연구망인 KREONET 와 ESNET 으로 연결되어 이론상 최대 10Gbps 의 대역폭을 활용할 수 있다. 아래 그림 1 은 동시 접속 수와 연결지속시간에 따른 시간당 전송량을 iperf 를 이용해 측정한 결과이다. 동시에 여러 연결을 설정하여 수행하고 최소 수초 이상의 전송을 유지해야만 충분한 네트워크 성능이 보장되는 것을 알 수 있다.

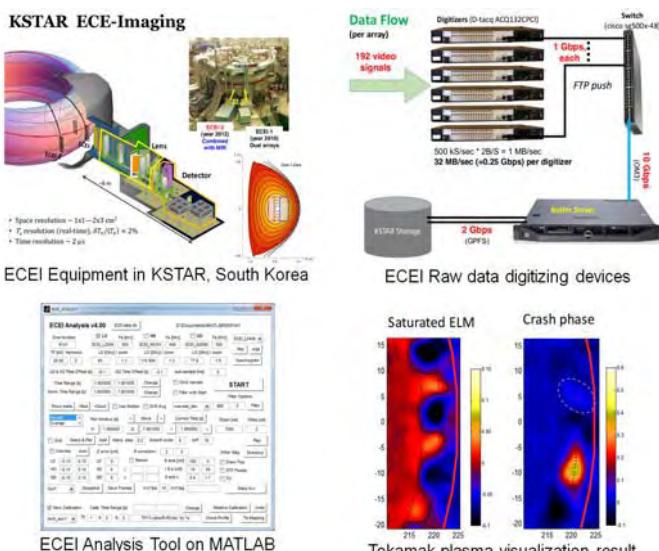


(그림 1) KISTI 와 ORNL 간 네트워크 전송 성능

## 3. ECEI 분석 응용

국가핵융합연구소 KSTAR 토카막에는 3 대의 ECEI(Electron Cyclotron Emission Imaging) 카메라가 설치되어 핵융합 실험시의 플라즈마의 온도를 측정한다. 이 데이터를 시각화하여 분석하여 핵융합의 안정성을

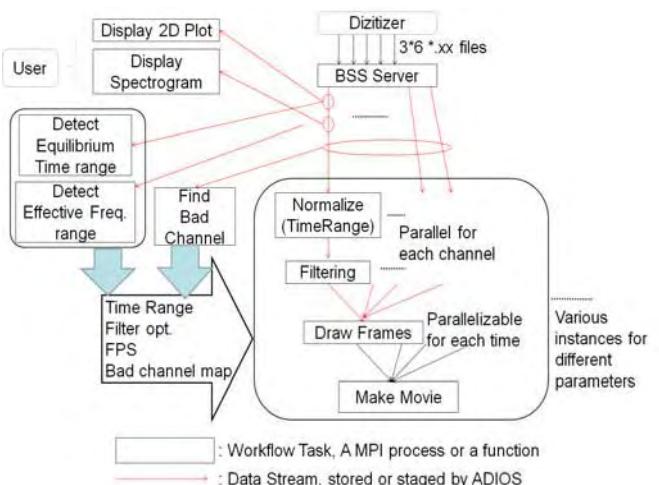
확보하기 위한 다음 데이터의 설정 값으로 활용되게 된다. 2014년의 측정의 경우 데이터의 분량은 다음과 같다. 3개의 카메라에서 가로 8대 세로 24대의 측정 장비에서 초당 50만 번의 측정으로 10여 초간 총 500만 개의 프레임을 측정하고 하나의 데이터의 크기는 2Byte이다. 원시 데이터의 총량은 하나의 실험 당 6GByte 정도이다. 이러한 실험을 하루에 10회 전후로 수십 일을 진행한다. 향후 전체 실험 시간도 늘어나고 프레임 간 간격도 들어날 예정이므로 한 실험에서의 데이터양은 수천 배 이상 늘어날 예정이다. 하나의 실험에서 다음 실험까지 15분가량의 시간 간격이 있는데 이 사이에 분석을 완료하여 적용하는 것을 목표로 하고 있으므로 빠른 데이터의 처리가 필수적이다. 그림 2는 ECEI 응용의 카메라 구성(왼쪽 위), 원시데이터 수집을 위한 Digitizer 구성(오른쪽 위), 분석 프로그램(왼쪽 아래), 시각화 결과(오른쪽 아래)를 각각 나타낸다.



(그림 2) ECEI 분석 응용 Overview

데이터를 최종 결과물인 동영상으로 만들기 위해서는 데이터의 유효성 검사, 필터링을 이용한 노이즈 제거 및 유효 데이터 부각을 위한 후처리, 시각화를 위한 평탄화 작업, 이미지로의 변환 및 동영상 압축의 단계를 거쳐야 한다. 기존에 포항공대 물리학과 연구팀에서 이와 같은 작업을 하기 위해서는, KSTAR 서버에 ftp로 접속하여 원시데이터를 다운로드 받고, 이를 이용하여 MATLAB으로 구현된 분석 프로그램으로 수동으로 작업을 진행하였다. 이 과정은 매우 오래 걸리고 분석 가능한 데이터량에도 상당한 제약이 있었다. 이에 본 연구에서는 MPI와 ADIOS를 이용하여 모든 작업을 자동화하고 병렬화하여 효율성과 성능을 획기적으로 향상 시켰다. 그림 3은 ECEI 분석 응용의 전체적인 workflow와 병렬화를 도식화한 그림니다. MPI와 ADIOS가 설치된 Linux 서버에서 수행 가능한 병렬 프로그램의 형태로 다음과 같은 기능들을 구현하였다.

- r2adios : 원시데이터의 유효성 검증 및 ADIOS 파일 포맷으로 변환, 원거리 전송
- makeframes : 데이터 normalization, filtering, 채널별로 병렬 수행
- spectrogram : 주파수별 밀도를 시각화
- drawchannel : 채널별 시간에 따른 데이터를 시각화
- makemovie : makeframes로 생성된 데이터를 동영상으로 제작
- findequil : 플라즈마 안정화 구간을 자동으로 탐색
- automake : 위 기능을 조합하여 실험 완료 후 모든 안정화 구간의 동영상 제작



(그림 3) ECEI 분석 응용 병렬화 Workflow

테스트에 사용한 9141번 실험의 데이터 이용하여 automake를 실행할 경우 6GB를 분석하여 각각이 300Mbyte 크기의 동영상 110개를 생성한다. 이 작업을 기준의 방법처럼 연구자의 PC에서 MATLAB 프로그램을 이용할 경우 30시간 이상이 소요된다. 이 작업을 병렬화된 프로그램을 이용하여 KISTI의 24core 클러스터에서 수행할 경우 9분이 소요되었다. 이는 15분의 실험 간격 안에 데이터 분석을 완료해야 되는 요구조건을 충족 시킬 수 있는 시간 간격이다. 또한 아직 병렬 최적화 등이 적용되지 않았음을 고려하면 추가적인 시간 단축도 가능할 것으로 기대된다.

개발된 ECEI 프로그램은 KSTAR 내부 서버를 이용하여 로컬로도 수행 가능하지만 원거리 전송 기능도 포함한다. KISTI-포항공대 간 KISTI-ORNL 간에 실험을 수행하였다. 동영상 렌더링 시간을 제외한 데이터 변환 및 전송에 소요된 시간이 표 1에 나타나 있다.

&lt;표 1&gt; ICEE module 이용 6GB 데이터 전송 소요시간

네트워크	소요시간
KISTI-ORNL	180초
KISTI-포항공대(1Gbps)	37초
Local loopback	78초

KISTI-ORNL 망의 경우 병렬 연결 수를 늘리면 성능이 좋아지는 특성을 활용하면 국내 망과 넓은 대역폭을 활용하여 보다 나은 전송성능을 활용 대용량 데이터 처리가 가능할 것으로 판단된다. 현재의 환경에서도 전송시간과 계산시간을 합쳐도 12 분 가량이 소요된다. 이를 모두 종합하여 볼 때, ADIOS-ICEE module 프레임워크를 이용하여 병렬 처리 및 전송 기술을 활용함으로써 기존에 불가능하였던 실시간 요구사항인 15 분 이내의 분석을 가능하게 함을 확인할 수 있었다.

#### 4. 결론

본 연구에서는 대용량의 데이터를 다루며 국제 협력 등의 이유로 데이터를 이동하고 분석해야 하는 환경에서 연구자들의 프로그래밍 난이도를 낮추고 호환성을 향상시킬 뿐 아니라 고성능을 제공하여 실시간 분석을 가능하게 하는 추상적이고 유연한 I/O 인터페이스 계층의 필요성을 제기한다. 그리고 그 해답으로 ADIOS 기반의 인터페이스에 대규모데이터의 병렬 원거리 전송기능인 ICEE module 을 설계하고 구현하였다. 또한 이 기술을 바탕으로 KSTAR 토카막 실험에서 쓰이는 ECEI 응용을 MPI 기반으로 병렬화하여 수십 시간 소요되던 작업을 15 분 이내로 단축하여 실시간 반영이 가능함을 증명하였다.

앞으로의 연구를 통해서 데이터의 크기가 더욱 커지는 상황에서 효율적인 전송을 위해 1) 데이터의 우선순위를 정하여 보다 중요하고 빨리 분석되어야 하는 데이터를 미리 전송하고 분석을 시작할 수 있는 프레임워크, 2) 데이터 분석 코드를 전송자 쪽으로 이동하여 실제 전송되는 데이터의 양을 줄여서 성능을 향상시키는 기법, 3) 병렬 전송 및 데이터 공유로 인해 발생하는 데이터 중복을 최대한 활용한 데이터 라우팅 기법 등의 기능을 추가 개발할 계획이다.

#### 참고문헌

- [1] Qing Liu, Jeremy Logan, Yuan Tian, Hasan Abbasi, Norbert Podhorszki, Jong Youl Choi, Scott Klasky, Roselyne Tchoua, Jay Lofstead, Ron Oldfield, et al., “Hello ADIOS: the challenges and lessons of developing leadership class I/O frameworks”, Concurrency and Computation: Practice and Experience, 26(7):1453{1473, 2014.
- [2] GS Yun, W Lee, MJ Choi, J Lee, HK Park, B Tobias, CW Domier, NC Luhmann Jr, AJHDonne, JH Lee, et al., “Two-dimensional visualization of growth and burst of the edge-localized filaments in kstar h-mode plasmas”. Physical review letters, 107(4):045004, 2011
- [3] Fang Zheng, Hongbo Zou, Greg Eisenhauer, Karsten Schwan, Matthew Wolf, Jai Dayal, Tuan-Anh Nguyen, Jianting Cao, Hasan Abbasi, Scott Klasky, et al., “Flexio: I/o middleware for location-exible scientific data analytics”, In Parallel & Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on, pages 320-331. IEEE, 2013.
- [4] Ciprian Docan, Manish Parashar, and Scott Klasky. “Dataspaces: an interaction and coordination framework

for coupled simulation workflows”, Cluster Computing, 15(2):163-181, 2012.

- [5] Georgia Tech Research Corporation, “EVPath”, <http://www.cc.gatech.edu/systems/projects/EVPath/>