

# 국외 출장 결과보고서

|        |        |        |        |        |        |
|--------|--------|--------|--------|--------|--------|
| 결<br>재 | 담당자    | 팀장     | 팀장     | 실장     | 사업단장   |
|        | 12월 8일 | 12월 9일 | 12월 9일 | 12월 9일 | 12월 9일 |
|        | 노일석    | 김동일    | 심태현    | 김정한    | 이우진    |

## 1. 출장개요

|                   |                                                                                                                  |            |             |
|-------------------|------------------------------------------------------------------------------------------------------------------|------------|-------------|
| 출장목적              | SC 2022 (The International Conference for High Performance Computing, Networking, Storage, and Analysis 2022) 참석 |            |             |
| 출장일정              | 2022. 11. 12.~2022. 11. 20. (7박 9일) / 2022.11.12. ~ 2022. 11. 16. (3박 5일)                                        |            |             |
| 성명                | 소속                                                                                                               | 직급(직위)     |             |
| 김동일*              | 전산관리팀                                                                                                            | 책임급연구원(팀장) |             |
| 임수연               | 전산관리팀                                                                                                            | 원급연구원      |             |
| 노일석               | 모델체계팀                                                                                                            | 원급연구원      |             |
| 방문기관<br>(또는 회의장소) | 케이 베일리 허치슨 컨벤션 센터<br>(Kay Bailey Hutchison Convention Center Dallas, Texas, United States of America)            |            |             |
| 출발지               | 인천(Incheon)                                                                                                      | 도착지        | 댈러스(Dallas) |

※ 김동일: 2022 KIAPS 과학자문회의 참석으로 인한 조기 귀국

## 2. 업무수행내용

### ○ 개요

2022년 11월 13일부터 18일까지 미국 텍사스주 댈러스에서 열리는 2022 Super Computing Conference(이하, SC22)는 사업단에서 연구/개발/운영하는 수치예보모델의 구동에 필수적인 HPC 시스템을 주제로 전 세계 각 나라의 시스템과 엔지니어가 모여 의견을 나누며 보고 배울 수 있는 만남의 장소이다.

사업단의 KIM 모델은 거대한 자료를 기반으로 방대한 계산자원을 요구하기 때문에 HPC 및 Cluster 시스템은 KIM 모델 개발/운영에 필수적인 요소이며 이를 최적의 환경으로 구성하는 것은 KIM 모델 개발/운영에 중요한 축이다.

따라서, SC22에 참석하여 최신 HPC 및 Cluster 시스템과 연구 동향을 파악하고 직접 교육을 받음으로써 KIM 개발/운영에 적용할 수 있도록 하고자 한다. 그리고 세계적인 엔지니어들과 교류를 통해 견문을 넓히며 이를 바탕으로 현재 시스템의 문제를 해결하고 더 발전시키고자 한다.

### ○ 목적

- 김동일: SC22에서 고성능 슈퍼컴퓨터에 관한 최신 동향을 파악하고 배운 것을 참고하여 사업단의 현업모델 개발 및 운영하는 데 있어 시행착오를 줄이고 효율적으로 대응하고자 하며, 습

득한 최신 기술 정보를 활용하여 사업단 시스템 도입을 효율적인 방향으로 하기 위함

세계 각국의 엔지니어들과의 교류를 통하여 선진국들의 사례와 경험·노하우 등을 배우고 업무에 적극 활용하고자 함

- 임수연: 튜토리얼 및 워크숍 참석 등 교류 확대 기술 프로그램에서 나온 최신의 기술들을 습득하고 전 세계 기술 동향을 파악하며, 사업단 현업모델 개발과 운영을 더욱더 효율적으로 하기 위함

차세대 고성능 HPC 관련 기반기술(시스템 아키텍처, 페타급 스토리지, 고성능 네트워크, 병렬화 입출력, 병렬화 소프트웨어 등)을 경험하고, 현 시스템의 문제를 예측·진단하여 사업단 시스템의 문제 해결 및 튜닝 등의 기술을 발전시킬 수 있도록 자기개발을 위한 동기 부여

- 노일석: 고성능 슈퍼컴퓨터를 활용한 대용량 병렬계산 등 최신 계산과학/공학 분야의 연구 동향을 파악하며, CPU 시스템의 연산 뿐만 아니라 거대 GPU 시스템을 활용한 과학계산에 대한 지식과 경험을 함께 습득하기 위함

튜토리얼 및 워크숍 등에 적극적으로 참여해 관련 기술을 습득하고, 이를 사업단 현업모델의 프레임워크를 고성능 컴퓨터의 자원을 최대한 활용할 수 있도록 최적화하는 데 활용 하고자 함

## 나. 주요 업무 수행 사항

### o 김동일

- Advanced MPI Programming, Efficient Distributed GPU Programming for Exascale

멀티 클러스터 HPC의 대규모 병렬 시스템 구성을 위해 대표적인 병렬 라이브러리인 MPI 튜토리얼과 최근 증가하는 이기종 시스템의 병렬 환경과 관련된 튜토리얼에 참석하였음. 튜토리얼을 통해 병렬 시스템에 대한 최근 동향과 연구자들의 시스템 성능 수요를 파악하였음. 특히, Efficient Distributed GPU Programming for Exascale 튜토리얼을 통해 이기종 시스템이며 클라우드 서비스를 채용한 올리히 연구소 JUWELS 시스템을 직접 경험해보았음. 사용자 입장에서 자체보유 HPC의 클라우드 서비스를 분석하고, 장단점을 파악할 기회가 되었음.

- Highlights of the 60th TOP500 List including Awarding

미국 테네시주 오크리지 국립연구소(ORNL)의 Frontier-System이 2022년 6월에 이어 Top500의 1위를 유지하였으며, 프론티어 시스템(Frontier System)은 1.6ExaFlops/s의 최고 성능을 가지고 있으며, 지금까지 1.102Eflops/s의 HPL 벤치마크 점수를 달성했다.

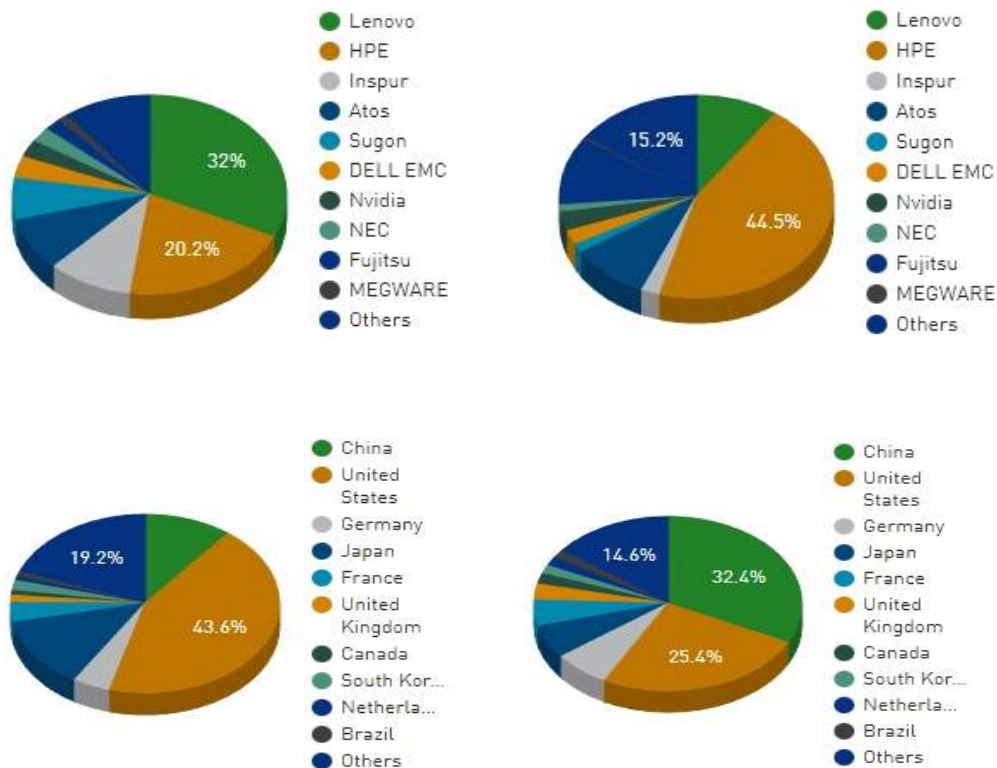
그뒤로 일본의 Fugaku, Finland의 LUMI HPC가 각각 2위 3위를 기록하고 있으며, 이번에 새로 기록된 Italy에 설치된 Atos 회사의 Leonardo 시스템이 새로이 4위로 기록되었다

TOP500의 상위 5개 시스템 순위는 아래와 같다.

| Rank | System                                                                                                                                                                              | Core      | Rmax<br>(PFlop/s) | Rpeak<br>(PFlop/s) | Power<br>(kW) |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|-------------------|--------------------|---------------|
| 1    | Frontier - HPE Cray EX235a, AMD Optimized 3 <sup>rd</sup> Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>DOE/SC/Oak Ridge National Laboratory<br>United States | 8,730,112 | 1,102.00          | 1,685.65           | 21,100        |
| 2    | Fugaku - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu<br>RIKEN Center for Computational Science<br>Japan                                                    | 7,630,848 | 442.01            | 537.21             | 29,899        |
| 3    | LUMI - HPE Cray EX235a, AMD Optimized 3 <sup>rd</sup> Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE<br>EuroHPC/CSC<br>Finland                                    | 1,110,144 | 151.90            | 214.35             | 2,942         |
| 4    | Leonardo - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband<br>EuroHPC/CINECA<br>Italy                                 | 1,463,616 | 174,700           | 255,751            | 5,610         |
| 5    | Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM<br>DOE/SC/Oak Ridge National Laboratory<br>United States        | 2,414,592 | 148.60            | 200.79             | 10,096        |

#### - TOP500 시스템의 주요 비교

(제조사별 점유율, 제조사별 성능 점유율, 국가별 점유율, 국가별 성능 점유율)



o 임수연

- Introduction to Networking Technologies for High-Performance Computing

HPC에 적용되는 네트워크는 기본적으로 속도가 중요하기 때문에 이더넷, ATM, Myrinet, Fibre Channel, InfiniBand 등 기술이 발전되어 왔다. 현재는 NVIDIA社에서 개발한 InfiniBand(IB) 성능이 400Gbit/sec까지 향상했으며, NVLink/NVLink2/NVSwitch, Slingshot, Tofu 및 Aquila 아키텍처를 기반으로 IB, HSE, RoCE를 캡슐화하는 OpenFabrics stack 아키텍처 기술이 관심받고 있음. 해당 기술은 Queue Pair Model을 적용하여 높은 처리량과 인접 HCA-스위치/스위치-스위치 간에 동작하는 링크레이어의 흐름 제어와 HCA 간에 end-to-end에서 동작하는 흐름 제어로 양단 버퍼의 빈용량을 상대방에게 통지함으로써 통신량을 조절할 수 있음. 두 종류의 흐름 제어로 폭주에 강한 네트워크를 구성하고, 처리량 한계에 가까운 성능을 쉽게 끌어내는 것이 가능함. 또한, 가속 엔진으로 슈퍼컴퓨터, 인공지능 및 하이퍼스케일 클라우드 데이터 센터에 적용할 수 있음.

- EduHPC-22: Workshop on Education for High Performance Computing

해당 워크숍은 하드웨어 및 소프트웨어 벤더, 정부 연구실 등에 소속되어 있는 개발자들과 함께 HPC 개발과정에서의 문제 및 이를 해결하기 위한 접근 방식을 공유하고, 다양한 아이디어를 의논하는 자리였음.

고성능컴퓨팅 하드웨어 기술로 이중 프로세서, 메모리와 스토리지, 연결망 기술과 소프트웨어 기술로 병렬 프로그래밍 모델, AI와 클라우드 기술이 접목된 사례들을 알 수 있었음. 하드웨어 기술은 프로세서, 메모리 스토리지 등 이중 시스템들이 더욱 분화되고 계층화되는 추세이며, 병렬 프로그래밍, AI, 빅데이터 분석 등 응용기술 분야의 관심이 높아지고 있음. 또한, 클라우드 기술이 고성능컴퓨팅 응용 개발자들 사이에서 필수 편의 서비스가 되어짐.

- Research Software Engineers in HPC: Creating Community, Building Career, Addressing Challenges(RES-HPC-2022)

현재까지 고성능 컴퓨팅에 적용되는 소프트웨어 연구 동향을 알아보고, 소프트웨어 개발과 관련된 패널들이 나와 앞으로의 소프트웨어 개발 산업의 확산 및 개발자들의 커리어를 향상시키기 위해 필요한 교육 과정 소개 및 피드백이 진행됨. HPC 수요가 증가함에 따라 효율적인 자원 공유 및 관리, 편의성, 워크플로우 자동화 등 관련 소프트웨어와 구축·서비스 사례를 소개함.

- First International Workshop on Cyber Security in High Performance Computing(S-HPC 2022)

많은 기업 및 기관이 고성능 컴퓨팅 클라우드를 도입하고 있으며, 기술 개발에 따라 사이버 공격 또한 진화되고 있는 추세임. 클라우드로 전환한 많은 기업 및 기관은 클라우드 환경에서

의 서버 자산에 대한 보안 운영 및 관리에 어려움을 겪고 있기 때문에 기존 물리적 환경에 최적화된 보안 체계를 그대로 클라우드 환경에 적용의 한계가 발생하고 있음. 이를 개선시키기 위해 NVIDIA社에서 간편한 설정만으로 보안 위협에 방지할 수 있는 DPU 서비스를 통해 네트워크 보안 기능을 DPU로 오프로드하여 애플리케이션 및 호스트 성능에 영향을 주지 않으면서 분산 방화벽 및 IDS/IPS 등 확장된 보안 기능을 제공하고 있음.

#### o 노일석

##### - 엑사스케일 컴퓨터 (Exascale computers)

Frontier 시스템의 등장 후, 엑사스케일(Exascale,  $10^{18}$  Flop/s) 컴퓨터 시대가 열렸음. 엑사스케일 컴퓨터의 핵심은 가속기(Accelerator)이며, 현재 대부분 이기종시스템(Heterogeneous System)의 가속기는 GPU (Graphics Processing Unit)를 채택하고 있음. 이기종시스템은 CPU 단일 시스템에 비해 월등한 성능을 내고 있으며, Top500 상위 10개 중 7개 시스템이 이기종 시스템임. 컨퍼런스에서 소개한 이기종시스템들은 대체로 CPU 2개와 GPU 4개 또는 CPU 2개와 GPU 8개로 계산 노드를 구성하고 있음.

가속기의 GPU는 NVIDIA사의 GPU 칩셋이 주류로 사용되고 있는데 그로 인해 CUDA 등 NVIDIA GPU 의존성을 갖는 언어와 라이브러리도 큰 커뮤니티를 이루고 있음. 또 다른 GPU 칩셋제조사인 AMD는 Frontier의 GPU를 구성하고 있지만 의존성을 갖는 HIP 언어와 파생 라이브러리 등의 개발이 CUDA를 비롯한 NVIDIA사의 소프트웨어보다 적게 이루어지는 추세임.

일본의 Preferred Networks 사는 독자개발한 GPU 칩셋 MN-Core를 소개하였음. 소개에 따르면 12nm 반도체를 사용한 1세대 MN-Core가 7nm 반도체를 사용한 NVIDIA사의 A100과 전력 대비 계산효율이 비슷한 수준이며, 추후 7nm 반도체를 사용한 2세대 MN-Core를 일본 슈퍼컴퓨터 후가쿠(Fugaku)의 후속 프로젝트에 포함하기로 함.

ACM 튜링상 수상자 Jack Dongarra 교수는 초청 강연에서 엑사스케일 컴퓨터 이후 제타스케일(Zettascale,  $10^{21}$  Flop/s) 컴퓨터의 등장 가능성에 대한 질문에 현재 CPU+GPU 구성에 GPU가 아닌 다른 가속기(예를 들면, 인공지능, 양자 컴퓨팅 등)를 추가로 결합한 시스템에 대해 상상할 수 있다고 답함.

##### - 클라우드 시스템 (Cloud systems)

전시 부스에서 다양한 업체가 독자적인 클라우드 시스템을 개발해왔고 각자의 장점들을 소개함. 통상적으로 Amazon사의 AWS로 대표되는 클라우드 서비스는 Google, Intel, Dell, Lenovo 등 대형 업체도 클라우드 서비스를 운영하고 있음. 또한, WEKA사로 대표되는 고성능 클라우드 데이터 스토리지 관련 분야가 최근 급성장하고 있음. 클라우드 데이터 스토리지 솔루션은 AI 분야에서 클라우드 DB 관리와 데이터 입출력을 목적으로 주로 사용하고 있음.

서비스 업체에 클라우드 서버를 두는 클라우드 서비스뿐만 아니라, 기관이 자체 보유한 서버에 클라우드 솔루션을 제공하는 업체도 다수 전시 부스를 구성하였음. 대표적으로 독일의 율리히 연구소(Forschungszentrum Julich)는 자체 보유한 JUWELS 시스템 등을 클라우드 시스템으로 사용하고 있으며, 사용자는 웹기반 클라우드 서버에 접속해 터미널/노트북 등으로 코드를 작성하고 작업을 제출함.

– 가속 라이브러리 (Accelerate Libraries)

NVIDIA 사는 CUDA 언어를 기반으로 자사 GPU에 의존성이 있는 더 많은 라이브러리를 개발하고 이를 소개하였음. 대표적으로 GPU 간 직접 MPI 통신을 수행할 수 있는 NCCL 라이브러리, GPU 클러스터 간 공유메모리(Shared memory) 사용을 위한 OpenSHMEM 라이브러리 기반의 NVSHMEM 등이 있음.

MPI 커뮤니티도 대규모 멀티 GPU 병렬화에 대한 수요를 느끼고 추후 MPI 5.0 업데이트에 GPU 간 MPI 직접통신 기능 추가를 고려하고 있음.

이처럼 멀티 GPU 또는 멀티 노드 GPU 프로그래밍의 경우 GPU 간 직접통신에 대한 이슈가 있고 연구자 및 개발자들은 하이브리드 병렬화(MPI+X)에서 해답을 찾고 있음. 하이브리드 병렬화에서 X에 해당하는 가속 라이브러리는 대표적으로 OpenACC, OpenMP, OpenCL, SYCL, Kokkos 등이 대표적임. 이중 OpenMP는 기존 CPU Threading 기능으로 많은 사용자를 보유하고 있는데 GPU kernel 생성 기능이 추가되며 기존 사용자들이 이기종 가속 전략으로 많이 채택할 것으로 보임. 반면 OpenACC는 워크숍과 논문 발표 등에서 현재 대부분 컴퓨팅 환경에서 OpenMP보다 나은 성능을 보이는 것으로 나타남.

### 3. 출장성과 · 시사점 및 향후 업무 활용 계획

#### 가. 출장성과 · 시사점

##### o 김동일

– 최근 해외에 다양한 기관에서 이기종 슈퍼컴퓨터를 도입하였으며, 일부 시스템은 클라우드 솔루션을 적용해 서버를 관리하는 동향을 파악할 수 있었음. 해외 기관의 최신 이기종 클라우드 서버에 직접 접속해 사용해 보며, 자원관리, 네트워크 보안, 스토리지 정책들을 직접 둘러볼 수 있었음. 또한, 이기종시스템에서 GPU 프로그래밍을 할때 GPU간 통신을 PCIe와 infiniband를 통해서도 현재 사용자의 수요 수준을 충족할 수 있을 것으로 보임.

##### o 임수연

– 엑사스케일 환경에서의 EDR/HDR/NDR 등 Infiniband Networking 시스템과 Lustre 등 스토리지 시스템으로 저지연, 광대역 낮은 CPU 사용률 조건 및 고성능 I/O 성능을 만족 유무와 보안을

강조하고 있음. 앞으로의 HPC 발전 방향을 참고하여 사업단에서도 현재 연구하고 있는 모델들이 동시에 작업함에 따라 고속의 데이터 전달이 가능한 네트워크 환경을 최적화함으로써 엑사스케일 슈퍼컴퓨터 등 최신의 HPC 시스템에서도 적합하게 활용할 수 있도록 개선하는 작업이 필요하다고 할 수 있음.

#### ○ 노일석

- KIM GPU 가속 전략으로 기존 MPI와 더불어 OpenACC를 함께 하이브리드 병렬화를 채택한 것에 대해 적절한 방향으로 전략을 설정했음을 확인할 수 있었으며 OpenACC가 NVIDIA HPC SDK 컴파일러에 의존성을 갖고 있다고 생각하고 있었는데 GNU 컴파일러에서도 호환성을 가진 것을 학회를 통해 알게 되었음. 그리고 현업화까지 고려한다면 Intel 컴파일러에서 호환성을 지원하는 OpenMP에 대한 가능성도 확인함(현재 Intel 컴파일러는 OpenACC를 지원하지 않고 있음).
- 엑사스케일 컴퓨터인 Frontier를 비롯해 향후 더 많은 이기종으로 구성된 슈퍼컴퓨터들이 등장할 것으로 예상되며, GPU 가속기를 장착한 이기종시스템이 성능과 활용도 측면에서 기존 CPU 단일 시스템보다 좋은 평가를 받고 있음. 유럽의 ECMWF는 이미 이기종시스템인 Leonardo로 연구에 활용하고 있으며, Destination Earth 이란 비전을 세우고 AI등을 접목해 1~4km 수평해상도의 Digital twin을 계획함. 향후 도입될 기상청 6호기도 세계적인 추세에 맞춰 이기종시스템으로 도입될 가능성이 높을 것으로 보이며, 우리 사업단도 이기종시스템을 충분히 활용하는 연구가 필요할 것으로 보임.

### 나. 향후 업무 활용 계획

#### ○ 김동일

- 여러 해외 기관의 HPC 시스템의 도입 현황과 사용 목적과 최근의 흐름을 파악할 수 있었으며 이를 사업단에서 운영중인 정책과 비교해보고 추후 사업단 개발 서버와 기상청 슈퍼컴퓨터 사용 정책 수립하고 개선하는 데 참고할 수 있을 것으로 보임.

#### ○ 임수연

- 세계적인 추세에 따라 사업단에서 연구개발 중인 데이터, 모델 등이 엑사스케일의 HPC와 같은 최신의 시스템에서도 동작할 수 있도록 사업단의 GAON2 연구개발 클러스터를 업데이트할 필요가 있음. 적용할 수 있는 최신의 네트워크와 최신의 스토리지 등을 설치하여 운영함으로써, 사업단에서 생산하는 데이터와 모델 등이 빠르고 최적의 환경의 HPC에서도 동작할 수 있도록 설계를 추진할 예정임.

#### ○ 노일석

- KIM 물리과정 GPU 가속 코드 이식 연구에서 현재 적용한 물리과정 외 더 많은 물리과정으로 GPU 가속 코드를 확장할 예정이고, 튜토리얼에서 습득한 프로파일 기술을 적용하고자 함.
- 가속화 전략인 OpenACC 라이브러리가 현재 기상청 현업 컴파일 환경인 Intel 컴파일러에서 지원하지 않기 때문에 OpenMP 라이브러리를 활용한 가속코드 실험이 필요할 것으로 보임.
- 현재 KIM 병렬통신방식인 양방향 통신(Two-side communication)을 대신해 속도에 장점이 있는 공유메모리를 사용한 단방향 통신(One-side communication)을 적용하는 것에 대해 논의하고자 함.