실시간 센서 데이터 처리와 분석을 위한 고성능 컴퓨팅 기술

박경석[1]
[1]한국과학기술정보연구원 슈퍼컴퓨팅기술개발센터
gspark@kisti.re.kr

High-Performance Computing Technologies for Real-Time Sensor Data Processing

Kyongseok Park[1]
[1]Center for Supercomputing Development, KISTI

요 약

과학기술의 발전으로 현대 과학은 대형 실험 장비를 기반으로 관측과 측정을 수행하고 이를 바탕으로 고성능 컴퓨팅 기술을 이용하여 시뮬레이션이 이루어진다. 실험 설비의 대형화와 복잡도 증가로 생성되는 데이터의 규모는 꾸준히 증가하고 있다. 데이터의 증가는 이를 효율적으로 수집하고 처리할 수 있는 기술을 필요로하며 연구자들은 이러한 대규모 데이터를 보다 체계적으로 관리하고 빠르게 분석할 수 있는 환경을 요구하고 있다. 발전 설비를 비롯한 대형 플랜트 시스템과 원자로, 핵융합로, 입자 가속기, 전파 망원경, 유전체 및 단백질 데이터 등 거의 모든 산업과 연구 분야에서 이같은 현상이 발생하고 있다. 본 연구에서는 대규모 센서 데이터를 실시간으로 처리하고 분석하기 위한 기술과 접근 방법을 논의하고자 한다.

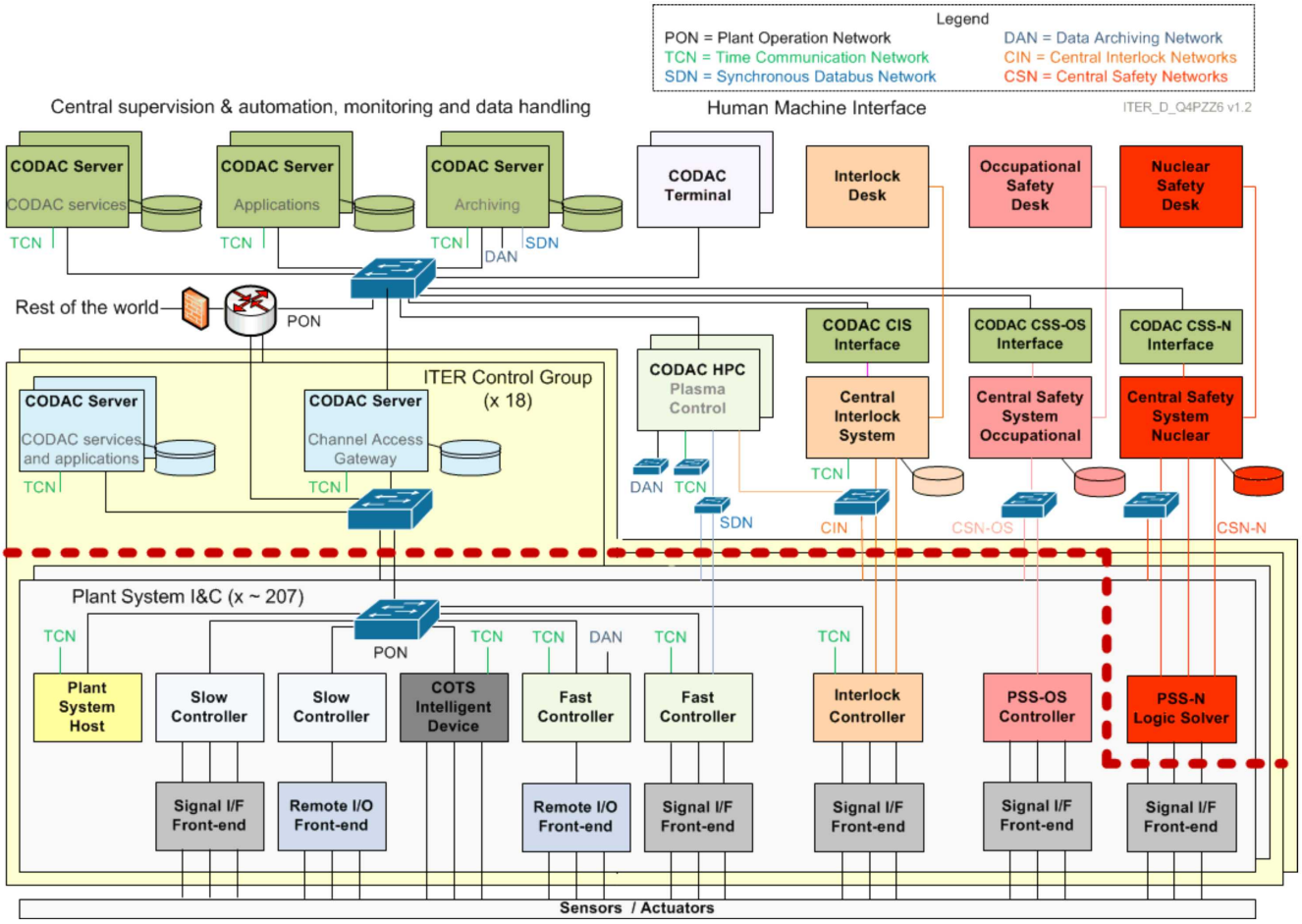
1. 연구 배경

고에너지물리(입자가속기) 분야에서는 TPC 연속 읽기(50 kHz)로 Run 2 대비 100배 이상 높은 데이터율을 GPU로 동시처리·보정·압축(온라인)하고, 무빔 기간에는 같은 GPU 팜으로 비동기(오프라인) 재구성까지 수행하는 '온라인 오프라인 통합' 체계, 실시간 처리와 저장비용 절감을 함께 달성하기 위한 ALICE O^2 (CERN, LHC Run 3) 연구가 수행되었다. 광학·전파를 기반으로 연구를 하는 천문학 분야에서는 Vera C. Rubin Observatory (LSST) 연구가 진행되었으며 30 초 노출마다 생성되는 차영상 기반 경보를 60 초 내 배포(평균 노출당 ≈10 천 경보, 밤당 ≈1천만 경보)하는 실시간 파이프라인과 추적 관측을 위해 지연을 엄격히 관리할 수 있는 기술을 요구하고 있다. SKA(Square Kilometre Array)에서는 원거리 관측소에서 중앙처리로 수 Tb/s급 데이터 전송과 실시간 연산을 위한 기술 요구사항이 존재하고 있다. 중력파 연구에서 핵심 역할을 담당하고 있는 LIGO Virgo KAGRA 시스템에서는 저지연 탐색과 스트림 매칭 필터 기반 파이프라인이 간섭계 스트레인 데이터 유입과 동시(online)에 이벤트 후보를 산출해 다중파장 추적용 신속 경보가 가능한 기술을 구현하였다. 핵융합 및 대형 플랜트 분야에서는 ITER CODAC Core System(EPICS 기반)과 다수 플랜트/진단 시스템의 분산 실시간 제어·계측 표준 툴킷인 JET/MAST 등 선행장치에 적용해 ITER 운영을 대비하고 있으며 Wendelstein 7X(W7 X) CoDaC에서는 장시간 정상상태 운전을 위한 제어 및 데이터 획득 체계를 초기 운전 단계에서 검증하기 위해 다양한 진단 장치 통합 프로세스를 설계하였다. DIII D Plasma Control System(PCS)에서는 수백 개 토카막 파라미터를 실시간으로 취득하고 피드백을 제어하는 디지털 제어 플랫폼의 기능 확장과 저지연화 연구가 진행 중이다.

2. 제안 기술

대형 실험 장비를 활용한 연구에서 고성능 컴퓨팅 기술과 실시간 빅데이터 분석 기술이 요구되는 이유는 첫째, 대형 실험·관측 시설의 데이터율이 저장·후처리 능력을 압도하고 있다. 예컨대 SKA는 원거리 링크만으로도 수 Tb/s 급 데이터를 전송하고, LHCb는 30 MHz 트리거리스 읽기로 수십 Tbit/s를 생성한다. 이처럼 모든 원시 데이터를 보관하는 것

은 물리·경제적으로 불가능하므로, 관측 현장에서의 실시간 축약·선별·압축이 필수이다. ALICE O²는 온라인 보정과 압축을 체계화한 대표적 사례이다. 둘째, 시간 임계적 과학 임무에서는 지연 시간을 최소화한 실시간 파이프라인 개발이 성과를 좌우한다. LSST는 각 노출 후 60초 내 경보를 전파해 신속한 후속 관측을 가능하게 하며, LIGO/Virgo/KAGRA와 지진 조기경보(ShakeAlert)는 분·초 단위 저지연 경보로 다중파장 추적이 가능하다. 셋째, 핵융합·원자로·대형 플랜트와 같은 설비에서는 실시간 피드백 제어와 상태 추정이 곧 시스템의 성공과 직결된다. ITER, W7 X, DIII D 등은 EPICS/CODAC 같은 표준화된 분산 제어 스택을 바탕으로 이기종 진단 장치를 초저지연으로 통합하여 안정 운전과 가동률 향상을 도모하고 있다. 넷째, 저장장치와 파일 기반 I/O의 병목 비용을 줄이기 위해 인 시투(in-situ), 인 트랜싯(in-transit) 분석의 중요성이 커지고 있다. ADIOS2 같은 고성능 스트리밍 I/O 프레임워크를 활용하면 데이터가 디스크에 쓰이기 전에 계산 중간 단계에서 분석·축약할 수 있다. 다섯째, 실시간 처리 환경에서도 재현 가능하고 공유 가능한 연구를 위해서는 FAIR 원칙에 부합하는 메타데이터, 영속적 식별자, 표준 스키마가 필수다. 이를 갖추면 경보·이벤트·환경 정보를 기계가 읽을 수 있는 형태로 연계할 수 있고, 후속 재분석·연계·자동화를 신뢰성 있게 수행할 수 있다. 여섯째, AI와 이상탐지 기법의 결합은 실시간 선택·요약의 품질을 높이고 저장 부담을 줄인다. GPU 트리거를 적용한 LHCb Allen이나



(그림 1) The global I&C Logical Architecture (Source: ITER)

CMS의 실시간 이상탐지 시도처럼, 스트림 상에서의 추론 기반 선별은 신호 민감도를 높이면서 필요한 데이터만 남기는 방향으로 진화하고 있다. 본 연구에서는 ITER의 대규모 센서 데이터를 신속하게 저장하고 연구자가 원하는 시점의 데이터를 필요한 해상도로 추출하고 저해상도 데이터부터 고해상도 데이터에 이르기까지 효율적으로 처리하고 분석할 수 있는 시스템을 위해 기존 ITER 시스템을 분석하고 기술적 요구사항을 제안하였다.

Components	Descriptions
High performance large-scale sensor data management	• Development of technology to systematically manage and produce large-scale sensor data using open-source libraries. • Development of technology to read and write large-scale sensor data at high-speed using parallel I/O technology. • Applying Parallel I/O technology to high-performance file systems
Large-scale multi-channel sensor data management	• Development of large-scale multi-channel time-series data management technologies capable of systematically managing and processing the data • Development of file system and database technologies for large-scale multi-channel data management and processing. • Development of distributed processing and parallel processing technology for large-scale multi-channel data
Real-time big data processing and management	• Development of message management and query processing technologies for real-time big data processing. • Real-time big data processing technology and distributed system integration • Implementation of distributed message processing technology for large-scale stream data processing.

(표 1) Core Reqirements for ITER System

3. 결론 및 추가 연구

데이터 폭증에 따른 I/O 지연과 성능 제약, 재현 가능성 및 AI 결합이라는 다양한 요인이 상호 맞물리며, 데이터가 생산되는 영역에서부터 효율적 처리와 분석이 가능한 기술이 필요하고 이를 기반으로 실시간 처리와 분석이 가능한 시스템은 다양한 영역에서 핵심 성공 요인으로 인식되고 있다. 본 연구에서는 차세대 에너지 기술로 관심을 받고 있는 핵융합 기술과 관련하여 전 세계 연구자의 관심과 참여가 이루어지고 있는 ITER의 향후 시스템 개발을 위한 기술적 제안을 하고 있다. 향후 ITER 시스템의 개발과 운영에 따른 시스템의 새로운 요구사항이 반영된 시스템을 제안해 나갈 예정이다.

사사(Acknowledgement)

이 논문은 IITP(No. RS-2024-00397359)와 KISTI(No.K25L1M2C2) 지원으로 수행되었습니다.

참고문헌

[1] ALICE Collaboration, “The ALICE Run 3 Online/Offline processing,” arXiv:2208.07412, 2022.

[2] M. Eulisse and D. Rohr, “The ALICE O^2 computing system for LHC Run 3,” arXiv:2402.01205, 2024.

[3] R. Aaij et al., “Allen: A High-Level Trigger on GPUs for LHCb,” Computing and Software for Big Science, vol. 4, no. 1, 2020, doi: 10.1007/s41781-020-00039-7.

[4] ATLAS Collaboration, “The ATLAS Trigger and Data Acquisition (TDAQ) system for Run 3,” JINST, vol. 19, p. P06029, 2024

[5] J. Bosch et al., “An Overview of the LSST Image Processing Pipelines,” in ADASS XXVIII, ASP Conf. Ser., vol. 523, pp. 521 - 530, 2019.

[6] B. A. Ewing et al., “Performance of the low-latency GstLAL inspiral search towards LIGO - Virgo - KAGRA observing run O4,” Phys. Rev. D, vol. 109, 042008, 2024.

[7] O. Martins et al., “An integrated approach to the development of the ITER control system,” in Proc. ICALEPCS 2011, 2011.

[8] E. Davis et al., “The ITER Controls are approaching one million integrated EPICS Process Variables,” in Proc. ICALEPCS 2023, 2023.

[9] ITER Organization, “CODAC Core System overview,” ITER Official Paper, 2025.

[10] Bluesky Project, “Bluesky Data Collection Framework – Documentation (v1.6.x),” 2024.