IDG
STRATEGIC
MARKETING SERVICES

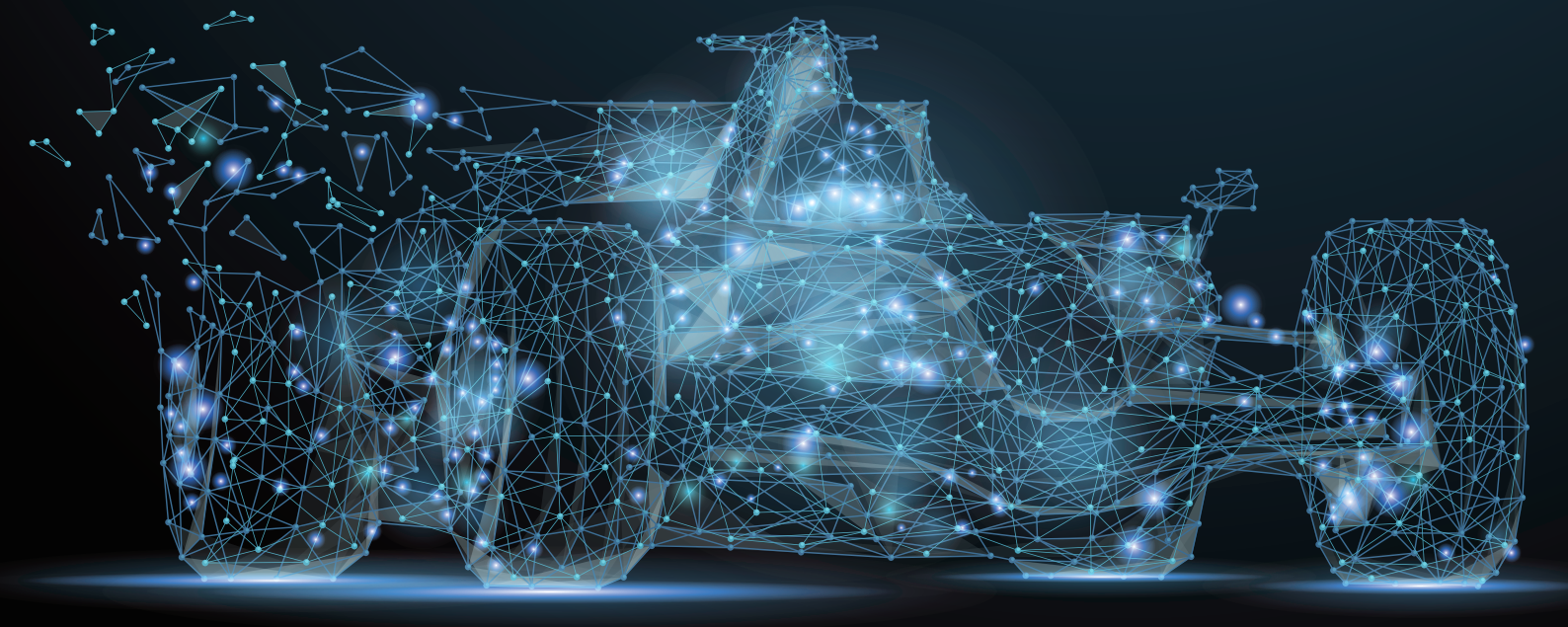
Hewlett Packard
Enterprise

SK 네트웍스 서비스

IDG Summary

자율 주행에서 머신 비전까지 'HPC+AI' HPE 글로벌 사례 집중 분석

점점 더 많은 기업이 인공지능(AI)과 머신러닝을 활용하면서 여러 가지 현장 적용 사례가 나오고 있다. 이때 AI의 효과를 극대화하려면 엣지에서 이루어지는 작업과 데이터센터 혹은 클라우드에서 처리하는 작업을 모두 고려해 인프라를 설계하는 것이 중요하다. 그렇다면 기업이 구체적으로 어떻게 준비해야 할까? 씨게이트의 엣지-투-코어 온프레미스 구축 사례, HPE 스위스 공장의 머신 비전 도입 사례 등을 통해 이상적인 AI용 인프라 구축 방안을 살펴보자. 또한, 비디오 분석과 감시를 활용한 안전 강화 사례, 자율주행차의 ADAS 시스템 지원 체계 구축 사례 등을 통해 엣지-투-코어 기술의 폭넓은 활용 가능성을 확인해 보자.



자율 주행에서 머신 비전까지 'HPC+AI' HPE 글로벌 사례 집중 분석

정구형 | HPE Korea 차장

현재 테슬라는 미국 캘리포니아 팔로알토 본사 주차장에서 '모델 X'를 이용해 자율주행 테스트를 진행 중이다. 모델 X는 테슬라 최초의 SUV로, 3초 만에 100km/h까지 가속할 수 있는 전기차다. 테스트 중인 모델 X 운전석에는 사람이 없다. 자동차가 스스로 주변 상황을 인식하면서 사람이 나타나면 멈추고 지나간 후에 다시 출발한다.

그렇다면 '모델 X'가 사람을 파악해 정지 명령을 내리는 것은 인터넷을 통해 연결된 클라우드 또는 데이터센터 속 컴퓨팅 장비일까 아니면 모델 X 내부에 장착된 엣지 컴퓨팅 기기일까? 정답은 엣지다. 불과 수 밀리초(millisecond, ms) 시간 차이로도 사람의 생사가 갈릴 수 있기 때문에 가장 빨리 판단하기 위해 엣지에서 실시간으로 처리한다. 실제로 모델 X에는 엔비디아의 드라이브 PX2가 장착됐다. 콜라캔 높이의 미니 PC지만, 2개의 고성능 GPU가 들어가 있다.

사람을 인식하는 구체적인 작업은 이 기기에 설치된 인공지능(AI)이 담당한다. 자동차 앞에 새로운 물체가 나타났을 때 사람인지, 자동차인지, 나무인지 추론하고, 그 결과에 따라 멈추거나 방향을 바꾼다. 이와 같은 추론 작업을 위해서는 수많은 데이터를 이용해 사전에 학습한 모델이 필요하다. 사람과 자동차, 도로, 표지판 등을 구분해 인식할 수 있도록 하는 작업이다. 엣지에는 이 추론 모델이 업로드돼 있어 카메라와 센서를 통해 실시간으로 들어오는 주변 환경 데이터를 해석한다.

단, 모델을 이용해 추론하는 것과 모델을 만드는 것은 완전히 다른 작업이다. 모델이 여러 가지 경우의 수와 상황에 대처하려면 방대한 데이터를 이용해 학습하는 과정이 필요하다. 이는 엣지 컴퓨팅 능력을 넘어서는 강력한 연산력이 필요하므로 클라우드나 데이터센터에서 처리한다. 그래서 일반적으로 자율주행차의 AI 모델은 클라우드 혹은 데이터센터에서 학습해 정교한 추론 모델을 만든 후 주기적으로 이 모델을 엣지에 업로드한다. 결과적으로 실제 주행 과정에서는 엣지 컴퓨팅만으로 가장 정교한 모델을 사용할 수 있다.

이처럼 AI 기술을 잘 활용하려면 엣지에서 이루어지는 작업과 데이터센터 혹은 클라우드에서 처리하는 작업을 모두 고려해 인프라를 설계해야 한다. 그렇다면 실제 기업 현장에서는 어떻게 활용되고 있을까. 지금부터 여러 기업의 실제 사례를 통해 HPE가 제안하는 이상적인 AI용 인프라 구축방안을 살펴보자.

씨게이트의 품질관리 장비(왼쪽)와 SEM 이미지 예시

씨게이트, 엣지-투-코어 HPE 제품 기반 온프레미스 구축

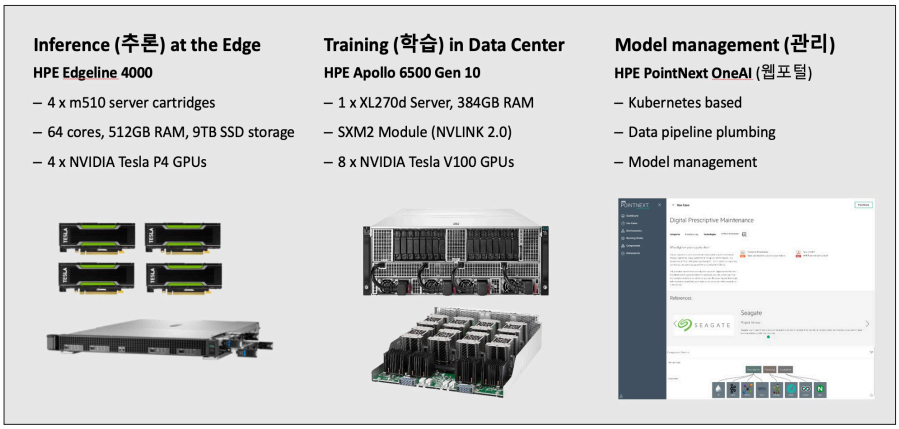
씨게이트(Seagate)는 세계적인 하드 디스크 제조기업이다. 하드 디스크의 품질을 결정하는 것은 웨이퍼이므로, 생산 과정에서 웨이퍼의 결함을 발견하는 것이 중요하다. 이를 위해 씨게이트는 카메라가 장착된 기계를 통해 SEM(Scanning Electron Microscope) 이미지를 찍는다. 1장당 800MB가 넘는 대용량 사진을 전자 현미경으로 분석해 불량이 있는지 검사한다. 이른바 '머신 비전(machine vision)' 기술이다.

씨게이트는 기존까지 이 시스템을 타사의 퍼블릭 클라우드 기반의 머신러닝 서비스로 처리했다. 사내에 전혀 인프라를 두지 않았으므로, 관리 측면에서 장점이 있었다. 문제는 24ms 수준의 레이턴시(latency)였다. 막대한 하루 처리량을 고려하면 치명적인 단점이었다.

결국, 씨게이트는 여러 고민 끝에 온프레미스 엣지 컴퓨팅 인프라를 구축하기로 했다. 이는 기본적으로 테슬라 사례와 같다. 공장 생산라인 바로 옆에서 실시간으로 머신러닝 모델을 이용해 이미지의 결함을 찾는다. 결함을 판단할 수 없는 이미지, 즉 아웃라이어(outlier)가 발견되면 데이터센터로 넘겨 새로 모델을 학습시킨 후 다음날 엣지에 수정된 더 정교한 모델을 업로드한다. 이를 통해 씨게이트는 레이턴시 문제를 해결하고 동시에 가장 정교한 모델로 검사 작업을 할 수 있다. 씨게이트가 도입한 엣지 컴퓨팅 제품은 HPE Edgeline EL4000 시스템이었다.

씨게이트 사례에서 또 눈여겨봐야 할 점은 아웃라이어를 학습하는 인프라를 온프레미스로 구축할지, 클라우드에 둘 것인지에 대한 판단이다. 결론적으로 씨게이트는 클라우드 대신 HPE Apollo 6000 제품을 이용해 온프레미스로 데이터센터를 구축했다. 이유는 3가지였다.

먼저 **비용**이다. 클라우드의 경제성과 유연성은 널리 알려졌지만, 일정 기간 머신러닝 용도로 사용할 때는 오히려 온프레미스보다 비용이 더 늘어날 수 있다. 예를 들어 한 번 학습하는 데 GPU가 8개씩 장착된 GPU 서버를 100대 사용해 10시간 돌려야 한다면 비용이 얼마나 들까? 타사의 퍼블릭 클라우드를 이용할 경우, GPU 1개당 1시간에 25달러, 약 3만 원이 필요하다. 서버 100대에는 800개



씨게이트의 머신러닝 프레임워크. 엣지와 데이터센터를 HPE 제품 기반의 온프레미스로 구축했고, 이 모두를 HPE Pointnext OneAI 웹 포탈로 관리한다.

의 GPU가 들어있으므로 1회 학습하는데 최대 2억 4,000만 원이 든다. 이를 24시간 3년 동안 운영한다고 가정하면 천문학적인 비용이 나온다.

이뿐만이 아니다. 네트워크 비용도 있다. 만약 전 세계 여러 곳에 공장을 두고 있다고 가정해 보자. 장당 800MB짜리 이미지를 매일 수백만 번, 수천만 번 클라우드로 전송하면 네트워크 비용만 많게는 월 수백만 원이 들 수 있다. 데이터센터가 실제로 위치한 지역의 로컬 네트워크 비용은 별도다. **씨게이트**는 이러한 비용 검토 끝에 최종적으로 온프레미스 구축을 선택했다.

두 번째는 **성능**이다. 이미지를 캡처하는 현장과, 이를 분석하는 클라우드 혹은 데이터센터 간 거리가 멀수록 레이턴시가 늘어난다. CostOwl.com에 따르면, 60마일(96km)당 1ms가 증가한다. 따라서 전 세계에서 공장을 운영하는 씨게이트가 레이턴시를 낮추기 위해서는 클라우드가 아닌 온프레미스가 합리적인 선택이었다.

마지막은 **록인(lock-in)**, 즉 특정 업체나 서비스로의 종속이다. 현재 다양한 클라우드 업체가 머신러닝 서비스를 제공한다. 문제는 기존에 사용하던 서비스를 확장하거나 인하우스로 이전 혹은 다른 클라우드 서비스 업체로 마이그레이션 할 때 발생한다. 코드를 사실상 재개발해야 하므로 결과적으로 기존 서비스와 업체에서 벗어나기 힘들다. 반면 온프레미스로 구축하면 이런 작업을 최소화할 수 있다.

HPE 스위스 공장, 머신 비전으로 서버 불량 25% 감소

두 번째 사례는 HPE다. **HPE 스위스 공장**에서는 한 달 평균 4만 5,000대 정도 서버를 생산한다. 이 공장에는 씨게이트와 비슷한 머신 비전 AI 시스템이 구축돼 있다. 로봇과 카메라로 구성된 검사 기기 아래로 조립이 끝난 상태에서 덮개가 열린 서버 플레이트가 지나가고, 이를 촬영한 이미지를 분석해 흠집이 있거나 특정 부품이 누락됐는지 확인한다. 포트가 엉뚱한 곳에 연결되는 등 서버 설정 규칙에 맞지 않는 제품도 자동으로 찾는다. 심지어 손톱만 한 부품 속 돌기 방향이 제대로인지 사람이 맨눈으로 찾아낼 수 없는 불량까지 잡아낸다.

시스템 구성 역시 씨게이트와 비슷하다. 카메라와 연결된 현장에 있는 HPE Edgeline EL4000 서버가 추론을 맡고 백엔드에 있는 HPE Apollo 6500이 학습을 담당한다. 머신 비전을 도입한 후 이 공장에서는 생산 과정의 불량률을 25%P 줄였다. 이미지 1장당 600MB가 넘는데, 클라우드에서 처리하는 것보다 불량 검사 시간을 1/20로 단축했다. 또한 촬영한 이미지 데이터를 이용해 백엔드 데이터센터에서 트레이닝한 후 개선된 모델을 다음날 새롭게 현장 서버에 업로드해 정확성을 개선할 수 있었다.

비디오 분석과 감시를 활용한 안전 강화 사례

엣지에서 빠르게 추론하고 백엔드 데이터센터에서 모델을 개선하는 또 다른 분야가 바로 비디오 분석과 감시다. 예를 들어 규정된 안전 도구를 착용했는지 확인하는 시스템이 대표적이다. 상황실에서 사람이 모니터로 확인하는 기존 방식에서는 담당자가 모니터에서 눈을 떼는 순간 위험이 발생할 수 있다. 반면 AI를 이용하면 상시 모니터링이 가능하다.

이를 위해 HPE는 소프트웨어 업체 그레이매틱스(Graymatics)와 함께 PPE (Personal Protective Equipment) 확인 시스템을 개발했다. CCTV와 EL4000 서버, 비디오 분석 소프트웨어 등으로 구성된다. 안전모나 글러브 같은 규정된 안전 도구를 착용하지 않으면 모니터에 붉은색으로 경고가 나타난다. 같은 원리로 제한된 위험 구역에 사람이 들어가면 경고가 울리는 시스템도 있다. 쓰레기 수거장 같은 시설은 제한 구역에 사람이 들어가면 안전사고가 발생할 수 있다.

실제로 아시아권의 한 공항에는 이러한 시스템이 구축돼 있다. 활주로 등 위험 구역에 사람이 있는지 실시간 감시한다. 공항에서는 한 번 사고가 발생하면 많

HPE와 Graymatics가 공동 개발한 PPE 확인 시스템. 규정된 안정 도구를 착용했는지 자동으로 확인한다.

게는 수백억 원까지 손실이 발생하므로 이러한 비디오 분석 시스템은 비용 효율적인 안전장치다. 이밖에 작업장 바닥의 물고임을 모니터링하는 시스템도 있다. 폐기물 처리 업체의 경우 보통 무게에 따라 가격을 책정하는데 물에 젖은 폐기물은 더 큰 비용이 발생한다. 센서를 설치해 일정 높이 이상 물이 차면 경고하도록 하면 이런 비용을 절감할 수 있다.

비디오 분석 중 가장 어려운 것이 화재와 연기 감지다. 불꽃은 그 모양과 형태, 색깔이 천차만별이고 기존의 온도 감지 카메라는 화재만 탐지할 수 있을 뿐 연기는 잡아내지 못하는 한계가 있었다. 이에 따라 HPE는 수많은 화재 케이스를 학습시킨 후 그 모델을 EL4000에 올려 실시간 추론하는 시스템을 개발했다. 이를 통해 정확도를 개선한 것은 물론 화재를 탐지하는 시간도 단축했다.

실제로 연기 센서가 연기를 감지하는 데 일반적으로 30초 정도 걸리는 반면 이 시스템은 이 시간을 5~10초로 줄일 수 있었다. 화재 상황에서는 이러한 작은 시간 차이만으로도 피해를 크게 줄일 수 있다. 비디오 분석과 감시 사례에는 앞서 살펴본 것과 비슷한 아키텍처가 적용된다. 현장에는 카메라와 엣지 서버를 두고, 백엔드 데이터센터에서 GPU 서버가 지속적으로 트레이닝하면서 더 정교한 모델을 다시 엣지 서버로 보내준다.

자율주행, AI-HPC-빅데이터의 결합

마지막 사례는 **자율주행**이다. 미국에 있는 한 기업의 자율주행 테스트를 위해 HPE의 전문 컨설턴트가 AI와 관련 인프라를 구축했다. 카메라와 센서로 수집한 정보를 분석해 운전자에게 경고하거나, 주행 자체를 제어하는 자동차 내부 시스템, 즉 ADAS(Advanced Driver Assistance Systems)용 지원 체계를 만드는 것이 핵심이었다.

이 자율주행 사례는 2가지 점에서 흥미롭다. 첫째, 빅데이터와 결합했다는 점이다. 이 기업은 자율주행차 여러 대를 테스트 중인데, 이들 차량은 정해진 시험장 내에서 충돌 사고 등 다양한 시행착오를 겪는다. 이러한 모든 기록이 영상으로 저장되므로 매일 엄청난 용량의 파일이 생성된다. 빅데이터 시스템이 필요한 것도 이 때문이다. 주행을 통해 얻은 다량의 비디오 데이터를 분석해 밤사이에 학습하고 이렇게 만든 더 정교한 모델을 다시 자동차의 엣지 컴퓨팅에 주입한다.

앞서 살펴본 다른 사례에서는 GPU 노드에서 쏟아지는 데이터를 저장하기 위해 러스터(Lustre), WEKA IO 같은 NAS 기반의 고성능 병렬 파일 시스템을 사용했다. 그러나 자율주행 사례에서는 대용량 비디오 데이터가 쏟아지므로 이런 데이터에 적합한 스토리지 시스템이 필요하다. HPE가 하둡 기반 빅데이터 시스템과 오브젝트 스토리지를 선택한 것도 이 때문이다.

여기서 끝이 아니다. 빅데이터 구조에서 중요하게 고려해야 할 것이 바로 데이터의 온도(temperature)다. 빈번하게 접근하는 파일은 핫(hot) 데이터, 하루에 한 번 정도 접근하거나 아카이빙하는 것은 콜드(cold) 데이터라고 하는데, 일반적으로 핫 데이터는 구조화된 스토리지에, 콜드 데이터는 오브젝트 스토리지에 저장하는 것이 효율적이다. 핵심은 들어오는 파일이 핫 데이터인지 콜드 데

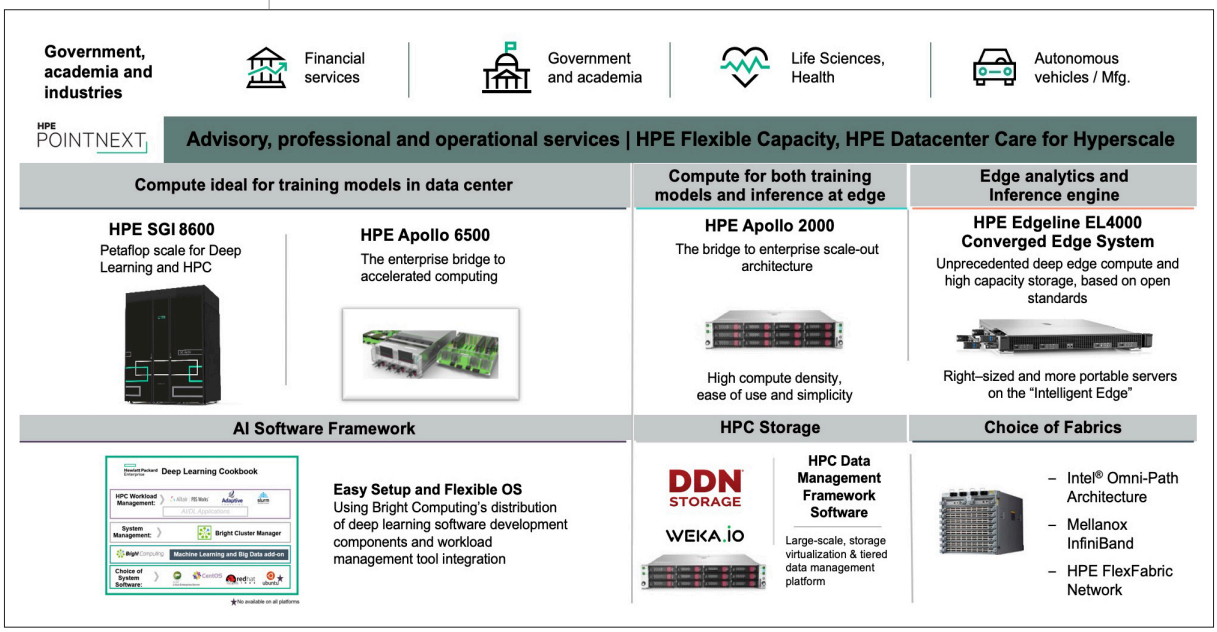
이터인지 추론해 더 적합한 스토리지로 분기하는 것이다. 이 자율주행 사례에서는 HPE의 데이터 매니지먼트 프레임워크(DMF)가 이 역할을 담당했다. 파일 시스템과 통신하는 정책 엔진이 내장돼 있어 조건에 따라 알맞은 저장소에 자동으로 티어링한다.

자율주행 사례의 두 번째 특징은 오픈소스 위주로 구축했다는 점이다. 이번 사례의 전체 시스템 구성을 보면, 일단 자동차가 있는 현장에는 여러 가지 Edgeline 서버가 사용됐다. 그 위에 사물인터넷 데이터를 처리하는 소프트웨어가 올라가고 이를 스트리밍해 현장에서 실시간 추론한다. 즉 장애물이 있을 때 정지할지 피할지 등을 현장에서 바로 판단한다. 하루 동안 주행을 마치고 돌아오면 대량의 비디오 파일을 빅데이터 시스템을 통해 하둡에 저장하고 학습을 시작한다. 이 모든 과정을 대시보드나 시각화 툴로 볼 수 있으며 모니터링도 가능하도록 구축했다.

실제 쓰인 장비를 보면 현장에서는 사물인터넷에 특화된 Edgeline 서버를 사용했다. 하둡 시스템으로는 x86 기반의 스토리지 서버인 Apollo 4200, Apollo 4510을 썼다. WEKA IO 부분에서는 HPE가 이미 레퍼런스 아키텍처를 보유한 Apollo 2000과 ProLiant DL360 서버를 이용해 병렬 파일 시스템을 구축했다. 백엔드의 HPC 클러스터에는 Apollo 6500 서버가, 나머지 대시보드와 모니터링 부분에는 웹서버인 ProLiant DL360과 DL380을 사용했다.

HPC 및 AI에 최적화된 HPE 통합 플랫폼

지금까지 살펴본 것처럼 HPE는 방대한 엣지-투-코어(Edge-to-core) 포트폴리오를 통해 기업이 자사 환경에 꼭 맞는 AI 플랫폼을 구축할 수 있도록 지원하며 시장과 기술을 이끌고 있다. 기업은 다양한 사양의 서버부터 스토리지, 소



HPE의 통합 딥러닝 제품군

프트웨어까지 실제 환경과 필요 요건에 따라 자유롭게 선택할 수 있다.

페타플롭 규모의 딥러닝이 필요하다면 HPE SGI 8600이 안성맞춤이다. 대규모 트레이닝 요건에 적합한 성능과 집적도, 효율성을 제공하는 액체 냉각 플랫폼이다. 다양한 기업에서 일반적인 트레이닝 작업을 처리하기에 가장 적합한 제품은 HPE Apollo 6500이다. HPE Apollo 6500 섀시에는 GPU 트레이와 CPU 트레이가 분리되어 장착된다. HPE Apollo 2000은 섀시 안에 최대 4개의 서버를 장착할 수 있어 집적도가 뛰어난 제품이다.

HPE Edgeline EL4000도 있다. 다양한 엣지 환경에서 추론 작업을 할 수 있도록 개발됐다. 업계 표준 x86 딥 컴퓨트 플랫폼을 기반으로 데이터를 수집해 분석하는 것은 물론, 엣지의 제어 작업을 관장하고 데이터 전송 문제를 해결한다. 이를 통해 더 빠르게 인사이트를 발굴하고 비즈니스 민첩성을 높인다. HPC와 딥러닝을 위해 개발한 포괄적인 소프트웨어 스위트도 활용할 수 있다.

고객의 요구사항에 맞는 엣지-투-코어 아키텍처 구성

점점 더 많은 기업이 AI와 머신러닝을 활용하면서 다양한 현장 적용 사례가 나오고 있다. 그러나 머신 비전, 비디오 분석과 감시, 자율주행 등 구체적인 예를 보면 기업의 개별 요건에 따라 다양한 형태로 인프라를 구성할 필요가 있음을 알 수 있다. 이런 상황에서 만능의 솔루션은 없다. 특히 클라우드에 대한 맹신은 오히려 위험할 수도 있다. 하드디스크 불량 검사나 자율주행차의 ADAS 등 레이턴시가 중요한 현장에 클라우드를 고집하면 정작 엣지에서 필요한 성능 수준을 맞출 수 없다. 최악의 경우 사람의 안전까지 위협할 수 있다.

엣지-투-코어 인프라를 둘러싼 전체 생태계도 잘 살펴야 한다. 예를 들어 비디오 분석과 감시 시스템을 도입한다면 구축 경험이 많은 소프트웨어 업체와의 협업이 중요하다. HPE는 다양한 비디오 분석 구축 사례에서 어떤 카메라와 서버, 소프트웨어를 어떻게 사용했는지 백서(https://idg.me/2Iwa4P0)로 자세히 제공한다. EL4000 서버 몇 대로 카메라 몇 대까지 운영할 수 있는지 등 상세히 기술했으므로, 비디오 분석 감시를 구축하는 기업에 실질적인 도움이 된다.

마지막으로 엣지 이외에 코어 데이터센터를 클라우드와 온프레미스 중 어떻게 구축할지 신중하게 고민해야 한다. 클라우드는 뚜렷한 장점이 있다. 이제 사업을 시작하는 스타트업이거나 1년에 1~2번 정도 머신러닝을 이용해 분석하는 기업이라면 최적의 선택일 수 있다. 그러나 본격적으로 머신러닝, AI 환경을 구축하고 싶다면 장기적인 관점에서 비용과 성능을 따져봐야 한다. 업체 종속 가능성도 꼼꼼히 점검할 필요가 있다. 이런 고민에 충분한 시간을 들인다면, 기업에 가장 적합하면서 비용 효과적인 AI 인프라를 구축할 수 있을 것이다.

SK네트웍스서비스 **최현석** 매니저 e-mail tuenrin@sk.com Tel 070-4755-9288
SK네트웍스서비스 **손승진** 매니저 e-mail sjson@sk.com Tel 070-4755-9209