DX 시대,
성공적인 AI 프로젝트를 완성하는 핵심 포인트 3

김형섭 차장 / 효성인포메이션시스템 HPC사업팀

DX 시대에 많은 기업은 '비즈니스 성과 달성'이라는 목표를 위해 클라우드를 선택한다. 그러나 클라우드를 통해 소기의 목적을 이루기 위해서는 적정 비용, 프라이빗과 퍼블릭 클라우드 간 유연한 상호 전환, 안정성·확장성·성능 관리 등이 뒷받침돼야 한다.

AI 프로젝트의 목표도 클라우드와 마찬가지로 '비즈니스 성과 달성'이지만, 프로젝트 성공을 위한 요건에는 다소 차이가 있다. AI 프로젝트를 추진하려는 기업들은 다음의 세 가지 사항을 특징적으로 요구한다. ▲클라우드, 컨테이너, VM(가상머신), GPU 연산, 빅데이터 구축 등 복잡한 인프라와 높은 솔루션 비용 문제 해결 ▲AI 모델에 대한 높은 기대치 충족 ▲실제 사용자를 위한 개발과 운영 관리 프로세스 가이드 등이다.

AI 프로젝트를 성공적으로 실행하기 위해서는 다음 사항들을 고려해 봐야 한다. 먼저 비용 측면에서는, AI 분석과 모델 연산을 위한 GPU 연산 자원을 어떻게 효율적으로 사용할 수 있는지, 그리고 늘어나는 데이터의 효과적인 저장과 성능 유지를 위한 방안을 고민해야 한다. 업무 효율을 위해서는 기존 낮은 성능의 인프라를 개선하기 위한 연산이나 I/O 성능 최적화 방안을 고민해야 한다. 마지막으로 효율적인 애플리케이션과 워크로드 운영을 위해 어떠한 플랫폼이 적합한지, 가상머신/컨테이너에서 동시에 운영하며 효율성을 높이는 방안, 그리고 AI 모델 개발 프로세스 및 절차에 대한 전반적인 전략이 제대로 수립되어야 한다.

↓ 프로젝트 추진 시 고려사항

DX 시대에 성공적인 AI 프로젝트 추진을 위한 세 가지 핵심 포인트를 알아보자.

01 | 비용효율성

연산 자원과 저장 자원의 효율화를 위한 방안을 수립해야 비용 효율적인 AI 프로젝트를 추진할 수 있다. GPU(Graphic Processing Unit) 연산 자원의 효율성을 향상시키기 위한 방식으로는 다음 네 가지가 대표적이다.

첫째로 실제 물리적인 GPU를 VM에 패스스루 방식으로 할당하는 '다이렉트패스(DirectPath) I/O' 방식이 있으며, 둘째로 엔비디아 vGrid 소프트웨어를 적용해 GPU 메모리를 원하는 크기로 분할해 VM에 할당하는 방식인 'vGPU'가 있다. 셋째로는 엔비디아 A100 GPU에서 사용할 수 있는 방식인 'MIG(Multi-Instance GPU)'를 통해 물리적 GPU를 메모리뿐 아니라 쿠다(CUDA) 코어까지 분할해서 활용하는 방안이 있다. 마지막으로 VM웨어의 비트퓨전(Bitfusion) 솔루션을 활용하면, 별도의 GPU 클러스터 풀을 구성해서 GPU가 없는 VM에서 GPU 연산이 필요할 때도 네트워크망을 활용하여 동적으로 GPU를 할당하는 방식이다.

위 방법 중 어떤 방식으로 진행할지는 고객이나 기관의 업무 특성에 따라 달라질 수 있으므로, 비즈니스 환경과 상황을 고려해서 선택해야 한다.

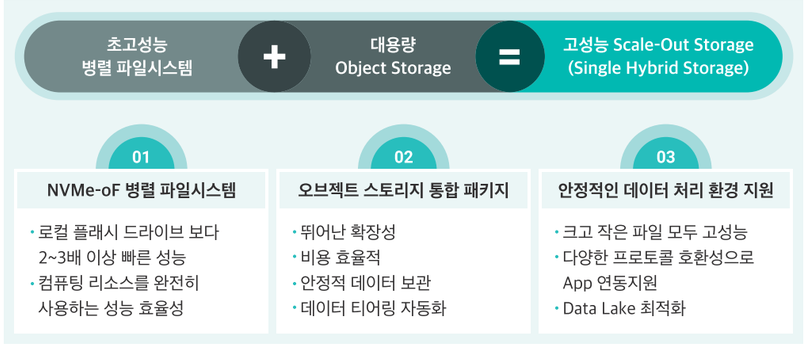
HCSF, 저장 자원의 효율성 향상

GPU 서버가 최대의 성능을 발휘한다고 해서 AI 분석 인프라도 최상의 성능을 보장하는 것은 아니다. AI 분석 인프라에 연계된 네트워크 I/O, 스토리지 I/O 등 스토리지 자원의 성능이 뒷받침돼야 하기 때문이다.

최근 많은 기업이 저장 자원과 관련해 초고성능 NVMe-oF 병렬 파일 시스템을 검토하고 있다. 로컬 플래시 드라이브에 비해 2~3배 이상 성능이 빠르고, 컴퓨팅 리소스를 완전히 사용해 성능 효율성도 탁월하다. 문제는 비용이 만만치 않다는 점이다.

동일한 성능으로 비용 문제를 해결할 수 있는 솔루션으로는 효성인포메이션시스템의 HCSF(Hitachi Content Software for File)가 있다. HCSF는 대용량 오브젝트 스토리지와 연계해 핫 데이터는 초고성능 병렬 파일시스템에 두고, 콜드 데이터는 대용량 오브젝트 스토리지로 티어링하는 방식을 사용한다.

↓ 차세대 고성능 파일시스템, HCSF



02 | 성능 극대화

AI 프로젝트의 성능 효율적인 측면에서는 I/O 효율이 중요하다. 즉, I/O 성능을 높이는 것이 GPU 서버의 성능 최대치를 끌어올리는 방법이라고 보면 된다. 엔비디아는 네트워크와 스토리지 I/O를 극대화하는 매그넘(Magnum) I/O 기술을 선보이고 있다.

네트워크 I/O 최적화 기능인 'GPUDirect RDMA'는 GPU와 네트워크를 직접 연결한다. CPU와 시스템 메모리를 거치는 통상적인 경로를 줄여 성능을 극대화하는 방식이다. GPUDirect Storage는 스토리지 I/O 최적화 기능으로 스토리지와 GPU 메모리 간의 데이터 로드 프로세스를 개선해 성능 효과를 극대화한다. 이러한 구성은 대용량 데이터센터를 구축할 때, 특히 GPU 서버가 10~20대 이상일 경우라면 반드시 고려해야 하는 요소다.

DPU, 대규모 데이터센터에 필수

I/O 효율 향상을 위한 두 번째 방안은 CPU, GPU를 이을 새로운 데이터 처리 장치로 주목받고 있는 DPU(Data Processing Unit)다. 과거의 시스템들은 CPU가 VM, 컨테이너 등의 워크로드와 동시에 인프라 관리, 소프트웨어 정의 보안, 소프트웨어 정의 스토리지, 소프트웨어 정의 네트워크 등 관리 항목까지 함께 연산을 수행했다. 그러나 DPU 방식은 이러한 워크로드의 일부를 네트워크 연산 장치인 DPU에 할당해 CPU의 기능을 분산시킨다. 현재까지 국내에서 DPU 구축 사례는 없지만, 대규모 데이터센터를 구축한다면 데이터센터의 전체적인 성능 향상 방안으로 DPU를 검토하기를 제안한다.

03 | 개발/운영 전략

데이터베이스, ERP 등 과거의 주요 애플리케이션은 한 번 설치하면 변경이 거의 없었다. 그러나 AI 모델과 같은 최신 애플리케이션은 유동성이 크고 변경도 자주 발생하며, 최대의 안정성과 성능을 보장하는 환경도 기존 애플리케이션과 다르다.

과거의 애플리케이션은 VM에서 최고의 성능을 발휘하는 반면, 최신 애플리케이션은 컨테이너 환경에서 안정성과 성능이 극대화된다. 따라서 AI 프로젝트를 추진할 계획이라면 공존하는 두 종류의 애플리케이션을 효과적으로 배치할 수 있는 방법을 무엇보다 중요하게 고려해야 한다.

AI 모델은 개발부터 실제 운영하기까지 크게 4단계로 나눌 수 있다. ▲실시간·비정형·정형 데이터의 추출/정제 ▲데이터의 가공 ▲데이터 레이크(data lake) 또는 데이터 마트, 데이터 웨어하우스 환경에 저장 ▲AI 모델링 애플리케이션을 통한 모델링과 분석이 그것이다. 따라서 VM과 컨테이너의 구성이 달라야 하며, 이를 어떻게 효율적으로 운영할 것인지가 중요하다.

VM과 컨테이너를 하나로 관리, 애플리케이션 배치 효율화

효성인포메이션시스템은 하나의 하이퍼바이저 위에 컨테이너 엔진을 탑재하고, 그 위에 가상머신과 컨테이너를 싱글 포인트 매니지먼트로 관리할 수 있는 솔루션을 제공한다.

개발/운영과 관련해 두 번째로 고려할 사항은 AI 통합 플랫폼의 구축 방안이다. AI 통합 플랫폼의 AI 학습 프로세스는 데이터 준비, AI 알고리즘 개발, AI 모델 학습을 통한 실제 모델의 적용이라는 비교적 단순한 프로세스로 진행된다.

그러나 이를 수행하는 인프라 또는 관련 애플리케이션 소프트웨어는 상당히 복잡하다. 따라서 AI 통합 플랫폼을 구축할 때는 이러한 요소들을 전체적으로 아우를 수 있는 협력사를 선택하는 것이 현명한 방법이다.

효성인포메이션시스템은 기업들의 성공적인 디지털 전환을 지원하기 위해 다양한 솔루션 및 서비스 포트폴리오를 제공한다. 클라우드 인프라 관련 비즈니스는 크게 'SDDC&클라우드 솔루션'과 '빅데이터/AI 클라우드 솔루션'으로 나눌 수 있으며, 고객들은 자사의 비즈니스 환경에 맞는 솔루션을 선택할 수 있다.