Power System AI 인프라 전략 리포트

: AI가 멈추면, 비즈니스도 멈춥니다

99.999% 가용성으로 실현하는 안전한 AI 추론 인프라

MetanetX | IBM Platinum Partner

AI 추론의 새로운 전장

AI가 기업의 핵심 업무로 들어오고 있습니다. 불과 2-3년 전만 해도 AI는 데이터 과학자들의 실험실에서 모델을 학습시키고 정확도를 개선하는 일이 전부였습니다. 그러나 이제 AI는 고객 응대, 여신 심사, 품질 검사, 재고 예측처럼 24시간 멈출 수 없는 비즈니스 프로세스의 중심에 서 있습니다.

이 변화는 인프라에 대한 요구사항을 근본적으로 바꾸고 있습니다. 프로덕션 환경의 AI 추론은 밀리초 단위 응답속도와 99.999%의 가용성을 동시에 요구합니다. 고객이 챗봇에 질문하는 순간, 대출 신청서가 제출되는 순간, 제조 라인에서 불량품이 감지되는 순간마다 AI 추론이 즉시 실행되어야 하기 때문입니다.

더 중요한 문제는 데이터의 위치입니다. 기존 접근법은 운영 중인 고객 데이터를 별도의 GPU 서버로 전송해 AI 추론을 실행했습니다. 하지만 이 과정에서 네트워크 레이턴시가 발생하고, 민감한 데이터가 이동하면서 보안 위험이 커집니다. 금융 거래 데이터, 의료 기록, 개인정보처럼 절대 외부로 나가서는 안 되는 데이터라면 이 방식은 애초에 불가능합니다.

이제 질문이 바뀌어야 합니다. "어떤 GPU를 쓸까?"가 아니라 "데이터가 있는 바로 그 자리에서 AI를 실행할 수 있을까?"입니다.

AI 추론, 속도와 안정성 사이

IBM Power10 프로세서는 각 코어에 MMA(Matrix Math Accelerator)라는 AI 전용 추론 엔진을 내장해, 데이터베이스 트랜잭션을 처리하는 바로 그 코어에서 동시에 AI 추론을 실행할 수 있습니다.

성능 차이는 명확합니다. Power E1080 모델은 이전 세대 대비 소켓당 AI 추론 처리량이 5배 향상되었습니다. BERT Large 같은 대형 자연어처리 모델도 실시간으로 처리할 수 있는 수준입니다. 코어당 4개의 MMA 엔진이 동시에 작동하며, PyTorch 같은 표준 AI 프레임워크를 그대로 사용할 수 있습니다.

더 중요한 것은 데이터 이동이 필요 없다는 점입니다. 고객의 계좌 정보로 이상거래를 탐지하든, 환자의 의료 영상으로 병변을 분석하든, 데이터는 원래 있던 보안 구역을 절대 벗어나지 않습니다. AI 모델이 운영 시스템에 직접 배포되고, 코어 내부에서 추론이 실행됩니다. 네트워크를 타고 GPU 서버로 가는 왕복 시간이 사라지니 레이턴시는 극적으로 줄어들고, 데이터가 이동하지 않으니 해킹이나 유출 위험도 원천 차단됩니다.

이 아키텍처는 AIX, Linux, IBM i 운영체제 모두에서 작동하며, Red Hat OpenShift와도 완벽하게 통합됩니다. 30년간 쌓아온 비즈니스 로직과 AI를 자연스럽게 결합할 수 있는 환경입니다.

멈출 수 없는 AI 비즈니스를 위한 설계

AI가 비즈니스 핵심으로 이동하는 순간, 가장 중요한 질문은 "얼마나 빠른가?"가 아니라 "얼마나 안정적인가?"입니다. 챗봇이 30초 응답하지 않으면 고객은 떠나갑니다. 품질 검사 AI가 5분 멈추면 불량품이 시장에 나갑니다.

IBM Power System은 14년 연속 ITIC 조사에서 가장 신뢰할 수 있는 서버로 선정되었습니다. Power10 시스템의 연간 계획되지 않은 다운타임은 평균 315밀리초입니다. 같은 조사에서 일반 x86 Linux 서버는 연간 59분의 다운타임을 기록했습니다. 99.9%와 99.999%는 실제로 연간 8시간과 5분의 차이입니다. 이 안정성은 메인프레임에서 40년 넘게 검증한 RAS 기술을 Power 시스템에 적용한 결과입니다.

FFDC(First Failure Data Capture) 기술은 CPU, 메모리, I/O에서 발생하는 모든 오류 신호를 실시간 감시합니다. 문제의 징후가 감지되는 순간, 서비스 프로세서가 오류 데이터를 저장하고 즉시 복구 프로세스를 시작합니다.

Chipkill Memory 기술은 메모리 칩 전체가 고장 나도 데이터를 복구합니다. 멀티비트 오류가 발생하면 문제의 칩 데이터를 여분의 메모리 공간에 자동으로 재배치하고, 시스템 운영은 중단 없이 계속됩니다.

Dynamic CPU Deallocation과 CPU Sparing은 코어 장애 시 대기 중인 여분의 코어를 자동 활성화합니다. 애플리케이션은 CPU 장애를 전혀 인지하지 못합니다.

각 파티션(LPAR)은 독립적으로 격리되어 있습니다. 한 파티션의 애플리케이션 오류나 OS 크래시가 다른 파티션에 전혀 영향을 주지 않습니다. 개발 환경에서 실험적인 AI 모델이 시스템을 과부하시켜도 운영 환경의 고객 서비스는 영향받지 않습니다.

전원, 냉각팬, I/O 어댑터는 모두 Hot Plug 방식으로 시스템을 끄지 않고 교체할 수 있습니다.

Dual Service Processor와 Dual Clock 설계로 단일 장애점을 제거했습니다. 결과는 명확합니다. AI 추론 서비스가 24시간 365일 멈추지 않고 돌아간다는 것입니다.

데이터부터 AI까지, 전방위 보안

AI 시스템의 보안은 네트워크 방화벽을 넘어섭니다. AI 모델 자체가 공격 대상이 되고, 학습 데이터가 유출되면 비즈니스 기밀이 노출되며, 추론 과정의 고객 데이터는 실시간으로 보호되어야 합니다.
Power10 프로세서의 투명한 메모리 암호화(Transparent Memory Encryption)는 프로세서와 메모리 사이를 오가는 모든 데이터를 AES 암호화로 자동 보호합니다. 애플리케이션 코드 수정이 필요 없고, 성능 저하도 없습니다. 메모리에 저장된 AI 모델 파라미터, 고객 데이터, 중간 연산 결과가 모두 항상 암호화된 상태입니다. 물리적으로 메모리 모듈을 빼내도 암호화된 데이터만 얻을 수 있습니다.
암호화 성능도 획기적입니다. Power10은 각 코어에 Power9 대비 4배 많은 암호화 엔진을 탑재했습니다. AES 암호화의 경우 코어당 성능이 2.5배 향상되었습니다. 데이터 Fetch와 저장 성능에 영향 없이 모든 데이터를 암호화할 수 있습니다.

보안은 여러 계층에서 작동합니다. 하드웨어 레벨에서는 Root-of-Trust와 TPM이 부팅 시점부터 시스템 무결성을 검증합니다. 운영체제 레벨에서는 Storage Keys 기술이 파티션 간 메모리 접근을 완벽하게 차단합니다. 워크로드 레벨에서는 서명된 이미지와 파일만 실행할 수 있도록 런타임 검증이 작동합니다.
Power10은 양자 내성 암호화와 완전 동형 암호화도 지원합니다. IBM PowerSC 솔루션은 클라우드 환경에서 실시간 컴플라이언스 모니터링, 보안 정책 위반 시 자동 경고, 감사 보고서 생성까지 통합 지원합니다.

결과적으로 Power System은 코어부터 클라우드까지 전방위 보안 아키텍처를 제공합니다.
AI 모델과 데이터가 처리되는 모든 계층에서 암호화, 격리, 무결성 검증이 작동합니다.

클라우드처럼 유연하게, 온프레미스처럼 안전하게

AI 워크로드는 예측하기 어렵습니다. 신제품 출시 직후 고객 문의가 폭증하면 챗봇 AI의 부하가 10배 늘어납니다. 월말 결산 시즌에는 이상거래 탐지 AI가 평소보다 5배 많은 트랜잭션을 처리합니다. Peak 워크로드에 맞춰 인프라를 구축하면 평상시 80%가 유휴 상태로 낭비되고, 평균 워크로드에 맞추면 중요한 순간에 시스템이 멈춥니다.
Power System의 공유 유틸리티 용량(Enterprise Pool 2.0)은 이 딜레마를 해결합니다. 시스템 총 용량 중 최소한의 기본 용량만 선불로 구매하고, 나머지는 실제 사용한 만큼만 분단위로 과금됩니다.
퍼블릭 클라우드처럼 사용량 기반 소비 모델이지만, 데이터는 절대 온프레미스를 벗어나지 않습니다.
예를 들어 3대의 Power 서버로 Enterprise Pool을 구성하면, 각 서버가 평상시 기본 용량만 사용하니 고정 비용만 발생합니다. 월말에 특정 서버의 배치 작업이 폭증하면 자동으로 용량을 늘립니다.
풀 전체의 기본 용량을 초과한 만큼만 분단위로 과금됩니다. 관리자가 코어를 활성화하거나 라이선스를 추가할 필요가 없습니다.

더 강력한 것은 자원 공유입니다. 메인 센터와 DR 센터 서버를 하나의 풀로 구성하면, 평상시 DR 센터는 최소 용량만 유지합니다. 메인 센터에 장애가 발생하면 DR 서버가 즉시 풀의 기본 용량 전체를 사용합니다. 자동으로 용량이 확장됩니다.
가상화 기술도 이 유연성을 뒷받침합니다. Dynamic LPAR 기능으로 온라인 중에 CPU, 메모리, I/O를 파티션 간에 재할당할 수 있습니다. 공유 프로세서 풀은 유휴 CPU 자원을 실시간으로 필요한 파티션에 자동 배분합니다. 마이크로파티셔닝 기술은 CPU 1개를 20개 파티션까지 세밀하게 나눠 리소스 낭비를 최소화합니다.

Live Partition Mobility는 작동 중인 VM을 서버 간에 무중단 이동시킵니다. 이 모든 기능은 IBM Cloud Management Console이라는 단일 대시보드에서 관리됩니다. 여러 데이터센터에 흩어진 모든 Power 서버를 통합해서 볼 수 있습니다.

결과적으로 Power System은 퍼블릭 클라우드의 탄력성과 온프레미스의 보안성을 동시에 제공합니다.
AI 워크로드가 폭증해도 즉시 대응하고, 비용은 실제 사용량만큼만 지불하며, 데이터는 절대 외부로 나가지 않습니다.

AI 시대, 인프라 선택의 기준

AI는 이제 선택이 아닌 필수입니다. 문제는 AI를 "어떻게" 실행하느냐입니다. 실험실의 AI는 정확도와 속도만 중요합니다. 하지만 고객 앞에서 돌아가는 AI는 한 순간도 멈춰서는 안 되고, 데이터는 철저히 보호되어야 하며, 비용은 예측 가능해야 합니다.
많은 기업이 "성능 아니면 안정성" 중 하나를 선택해야 한다고 생각합니다.
그러나 이것은 잘못된 이분법입니다.

IBM Power System은 이 두 요구사항을 하나의 플랫폼에서 해결합니다. 코어에 내장된 AI 엔진은 5배 빠른 추론 속도를 제공하고, 메인프레임 DNA는 연간 315밀리초 다운타임으로 24시간 가용성을 보장합니다. 투명한 메모리 암호화는 성능 저하 없이 모든 데이터를 보호하며, 분단위 과금 모델은 클라우드 수준의 비용 효율성을 온프레미스에서 구현합니다.

더 중요한 것은 데이터의 위치입니다. Power System은 운영 데이터가 있는 바로 그 자리에서 AI를 실행합니다. 30년간 쌓인 비즈니스 로직과 AI가 같은 코어에서 자연스럽게 결합됩니다.
엔터프라이즈 AI의 기준은 명확합니다. 실험실이 아닌 현장에서 검증되었는가? 한 순간의 중단도 허용되지 않는 미션크리티컬 업무를 감당할 수 있는가? 고객 데이터와 AI 모델을 동시에 보호할 수 있는가?
예측 불가능한 워크로드 변동에 즉시 대응할 수 있는가?

IBM Power System은 이 모든 질문에 "예"라고 답합니다. 14년 연속 업계 최고 신뢰성, 99.999% 가용성, 코어 내장 AI 가속, 투명한 메모리 암호화, 분단위 탄력적 용량 관리. 이것이 AI 시대 엔터프라이즈 인프라의 새로운 표준입니다.

MetanetX | IBM Platinum Partner

AI가 멈추면 비즈니스도 멈춥니다.

멈추지 않는 AI를 선택하십시오.