Sales & Support
gate@slexn.com
Tel. 02-555-4847
hub.slexn.com | www.slexn.com

SLEXN
IT Solution & Service Provider

온프레미스 LLM 인프라 복잡성을 해결하는 단일 어플라이언스 Puteron AI

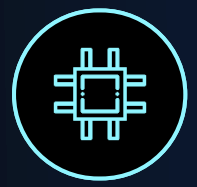
Puteron AI는 수많은 하드웨어 구성 요소와 소프트웨어 스택을 개별적으로 통합하고 최적화하는 복잡한 과정을 **단일 어플라이언스에서** 모두 해결할 수 있도록 설계되었습니다.

LLM 배포에 필요한 모든 필수 요소를 사전 구성·최적화하여 제공하며, 강력한 GPU부터 효율적인 데이터 관리 시스템, 직관적인 운영 소프트웨어까지 하나의 통합된 시스템 안에 담겨 있습니다.

단 한 번의 설치만으로 귀사의 AI 팀은 곧바로 LLM의 잠재력을 극대화할 수 있으며, 복잡한 인프라 관리 대신 모델 개발과 혁신에 집중할 수 있습니다.

스마트 AI 인프라 - 진화하는 "GPU Lego Box"

Puteron AI는 모듈형 AI 인프라를 통해 기업의 요구사항에 맞춰 LLM 운영 환경을 자유롭게 확장할 수 있으며, 초기 소규모 배포부터 대규모 엔터프라이즈 환경까지 지원합니다.



모듈형 Compute Bay:
최적화된 연산 유닛

다양한 폼팩터와 차세대 인터커넥트, 광범위한 하드웨어 호환성을 갖춘 최적화된 연산 환경으로 LLM 성능을 극대화합니다.

고성능 Scale-out 네트워크:
초저지연 LLM 분산 환경

초저지연, 고대역폭 네트워크와 간편한 확장 방식, NCCL-NVLink 가속으로 분산 LLM 학습 및 추론 효율을 높입니다.

고밀도 SSD 스토리지:
모델 로딩 및 데이터 처리 가속

NVMe Gen SSD와 지능형 데이터 압축으로 대용량 LLM 모델 및 데이터를 빠르고 효율적으로 로딩 / 저장합니다.

Puteron AI – 핵심 기능

LLM 서빙

완성된 LLM을 RESTful API 형태로 안정적으로 제공하여 실시간 추론과 서비스 운영을 지원합니다. AI Agent 기능을 실시간 제품 환경에서 바로 활용할 수 있습니다.

모니터링 & 분석

모델 품질, 리소스 사용량, 지연 시간 등을 실시간 시각화합니다. 대시보드 기반으로 시스템 상태를 추적하고 성능을 지속적으로 최적화할 수 있습니다.

리소스 관리

GPU·CPU·메모리 등 하드웨어 자원을 효율적으로 스케줄링합니다. Kubernetes 기반 운영으로 비용을 절감하고 실행 안정성을 보장합니다.

멀티모달 지원

텍스트, 이미지, 비디오, 오디오 등 다양한 모달 데이터 처리를 통합 지원합니다. 복합형 AI 모델 개발·운영의 기반을 제공합니다.

보안 & 접근 제어

조직 보안 정책에 맞춘 정교한 권한 관리 기능을 제공합니다.민감 정보 접근을 제어하고 데이터 유출을 원천 차단합니다.

하이브리드 검색 & RAG

Git·문서 저장소·내부 지식을 자동 수집해 즉시 검색과 RAG 응답을 제공합니다. 벡터 검색과 키워드 검색을 결합해 정확도를 크게 높입니다.

Puteron AI 는

"플러그 앤 플레이" 철학으로 설계되어
복잡한 AI 인프라 구축 과정을 단일 어플라이언스로 통합했습니다.

복잡한 DIY 접근 방식에서 벗어나 단일 어플라이언스로
엔터프라이즈급 AI 인프라를 즉시 활용해보세요.

One Platform for Enterprise AI

Puteron AI, LLM 운영의 모든 과정을 단일 플랫폼으로 혁신하다

1. One-Click LLM Foundry

- Llama-3-70B, Mixtral-8x22B, HyperClovaX-52B 등 100개 이상의 검증된 프리-빌트 모델 라이브러리 제공
- 웹 UI에서 한 번의 클릭으로 FP8, INT4, INT2 등 다중 정밀도 형식으로 모델 자동 변환 및 최적화
- 변환된 모델의 자동 벤치마크 테스트를 통한 상세한 성능 리포트(처리량, 지연 시간, 메모리 사용량 등) 즉시 생성
- QLoRA, LoRA 기반 경량 파인 튜닝 기능 지원

2. Auto-Ops Engine

- 선언적 운영을 위한 'Auto-Ops Engine'의 직관적인 웹 기반 드래그-앤-드롭 인터페이스로 손쉬운 YAML 관리 가능
- GitOps 파이프라인과 연동된 자동 Canary 업데이트 수행 및 새로운 버전의 모델을 안전하게 롤아웃
- 트래픽 분할을 통한 A/B 테스트 및 Shadow 배포 기능으로 실제 환경 영향 최소화하면서 모델 성능 검증
- 문제 발생 시 즉시 롤백을 통한 서비스 연속성 보장

One Platform for Enterprise AI

Puteron AI, LLM 운영의 모든 과정을 단일 플랫폼으로 혁신하다

3. Real-Time Telemetry

- LLM 운영 환경의 성능과 안정성을 실시간으로 모니터링 - GPU 메모리, 활용도, 온도, 전력 등 핵심 하드웨어 지표 초 단위 트래킹
- 'Hallucination Guard'로 모델의 비정상적인 응답 패턴 탐지
- 'Prompt-Efficiency Score'로 프롬프트 최적화 분석
- 통합 대시보드를 통한 실시간 시각화로 선제적 대응과 지속적인 LLM 품질 향상 지원

4. Secure Multi-Tenant Sandbox

- GPU SR-IOV 가상화 기술 활용, 1개 물리 GPU를 최대 7개의 독립 vGPU로 분할하여 안전하고 효율적인 리소스 공유 지원
- 각 vGPU에 QoS 설정을 적용해 보장된 대역폭과 컴퓨팅 자원을 할당하여 특정 사용자의 리소스 영향 방지
- NVIDIA, AMD, Intel 등 다양한 GPU 제조사 플러그인을 통합 관리하여 테넌트별 격리된 환경에서도 안전한 모델 학습 및 추론 보장

One-Click LLM Foundry | 프리-빌트 모델 라이브러리

웹 클릭 한 번으로 원하는 모델을 선택하고, 필요한 양자화 및 미세조정 방식을 적용하여 최적의 성능을 얻을 수 있습니다.

지원 모델

- GLM 4.6
- GLM 4.5-Air
- Mistral-Large-Instruct-2407
- Llama-4-Scout-17B-16E-Instruct
- GPT OSS-120B
- Qwen3-Next-80B-A3B-Instruct
- Google Gemma 3 27B
- Microsoft Phi-4 Reasoning Plus
- Llama-4-Maverick-17B-128E-Instruct
- IBM Granite 3.3 8B Instruct
- Qwen3-Coder-480B-A35B-Instruct
- Qwen3-Coder-30B-A3B-Instruct
- Qwen3-VL-235B-A22B-Instruct
- Codestral-22B
- IBM Granite 4.0 Tiny Preview
- Mistral NeMo Instruct-2407
- Yi Coder 9B Chat
- Seed-Coder-8B-Reasoning
- KAT-Dev-32B

Auto-Quantizer

GPTQ, AWQ, SpQR, FP8-MX 등 다양한 양자화 방식 지원

Fine-tune Templates

LoRA/QLoRA, DoRA, GaLore, ReLoRA 등 최신 미세조정 기법 지원

분산 전략

FSDP, DeepSpeed-ZeRO-3, Megatron-LM, µTransfer 지원

Auto-Ops Engine - 선언적 운영 자동화

모델 업로드 및 검증

- 개발자가 Git 저장소에 새 모델 파일을 푸시
- 자동으로 모델 무결성 검사 및 성능 벤치마크 실행
- 메모리 요구사항과 GPU 리소스 자동 계산

점진적 롤아웃

- 5분간 모니터링 후 이상 없으면 20%로 확장
- 토큰 생성 속도, 메모리 사용량, 응답 품질 실시간 추적
- 임계치 초과 시 자동 롤백 또는 운영팀 알림

배포 전략 설정

- 웹 UI에서 드래그 앤 드롭으로 배포 설정 구성
- Canary 배포: 전체 트래픽의 5%를 새 모델로 라우팅
- A/B 테스트 설정: 응답 품질과 지연시간 자동 비교

완전 배포 및 모니터링

- 모든 지표가 정상이면 100% 트래픽 전환
- 이전 모델 버전은 24시간 대기 상태로 유지
- 지속적인 성능 모니터링 및 최적화 제안

Real-Time Telemetry & Anomaly Brain

: 실시간 모니터링 및 이상 탐지

모니터링 지표

- GPU 메모리, SM 활용, NVLink 사용률
- TPU-MXU, PCIe 대역폭
- 온도, 전력 소비량
- 추론 지연 시간, 처리량

이상 탐지 기능

- 비지도 베이스라인 학습(1주간)
- μ±3σ 초과 시 즉시 알림
- Hallucination Guard
- Prompt-Efficiency Score

* 모든 지표는 Prometheus → VictoriaMetrics → Grafana Dashboard 파이프라인을 통해 실시간으로 시각화됩니다.

혁신적인 UI/UX 통합 환경 제공

LLM Studio

드래그 앤 드롭으로 배포 워크플로우를 설계하고, GPU 리소스와 실시간 메트릭을 시각화하여 최적의 운영 상태를 유지합니다.

Prompt Lab

실시간 대화형 테스트와 토큰별 성능 분석, API 코드 스니펫 자동 생성으로 모델 최적화와 개발 생산성을 극대화합니다.

Auto-Scheduler

GPU / 메모리 요구를 예측하고 리소스를 자동 배치 및 조정하여 안정적 운영과 비용 절감을 지원합니다.

업데이트 / 라이프사이클 관리

OTA Firmware

BMC + GPU VBIOS + NIC FW

→ 무중단 롤링 업데이트

Model Delta Sync

효율적인 증분 업데이트

→ 대용량 모델의 빠른 배포 지원

빠른 서비스 복구

최적화된 부팅 프로세스로

빠른 서비스 복구

콜드-스타트 최적화

모델 웨이트를 직렬화된

CUDA Graph로 미리 컴파일

Secure Multi-Tenant Sandbox - 안전한 리소스 공유

GPU SR-IOV 가상화

하나의 GPU를 최대 7개의 가상 GPU로 분할하여 여러 사용자나 워크로드가 독립적으로 안전하게 사용할 수 있습니다.

통합 쿠버네티스 플러그인

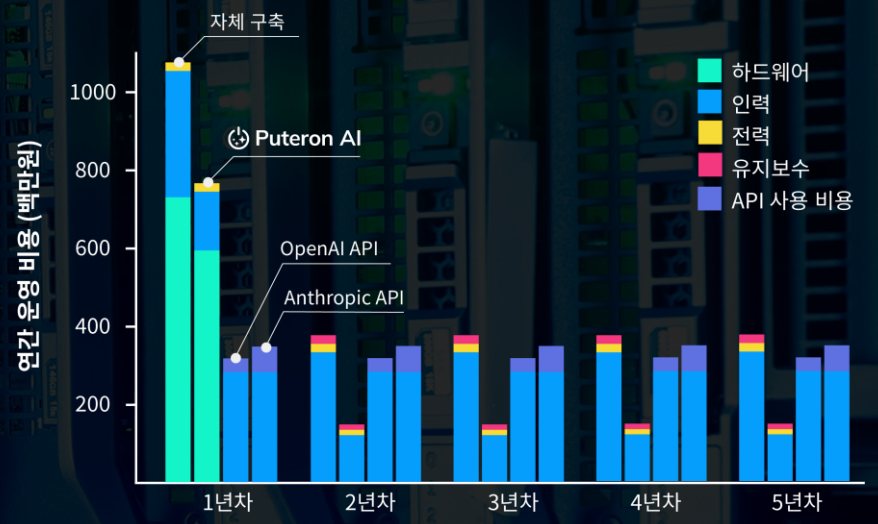
NVIDIA, AMD, TPU 등 다양한 하드웨어 플러그인을 통합 관리하여 멀티 벤더 환경에서도 일관된 관리 경험을 제공합니다.

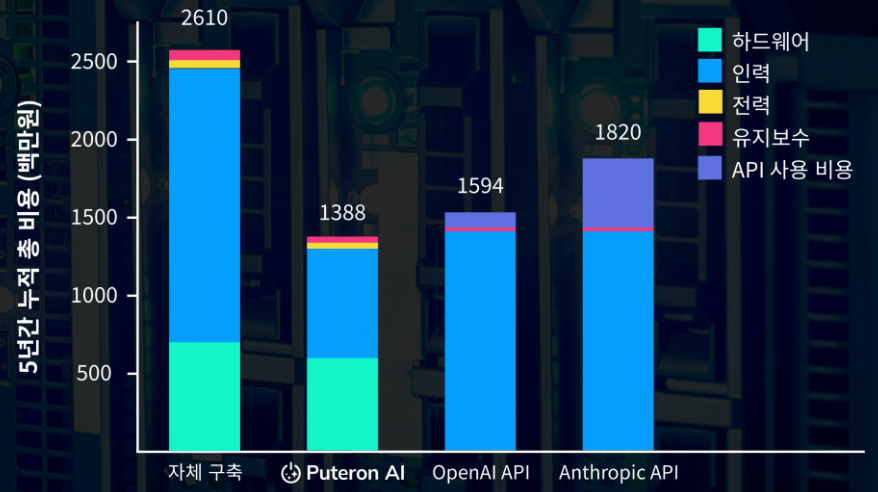
vGPU QoS(Quality of Service)

각 가상 GPU에 보장 대역폭(예: 10 TFLOPS)을 설정하여 중요 워크로드의 성능을 보장합니다.

Puteron AI가 제공하는 차세대 AI 인프라 비용 최적화 전략

- 중복 구축되는 GPU/CPU 리소스를 통합 운영해 자원 활용률을 극대화합니다.
- 자동화된 모니터링·스케줄링을 통해 전력/유지비/인력 비용을 지속적으로 절감합니다.
- RAG·서빙·파인튜닝을 하나의 환경에서 제공해 별도 시스템 유지비를 제거합니다.
- 그 결과, 자체 구축 대비 최대 40~60% 수준의 TCO 절감 효과를 제공합니다.





Puteron AI는 초기 투자와 운영 인건비까지 고려했을 때 장기적으로 가장 경제적인 선택입니다.

Puteron AI

SLEXN IT Solution & Service Provider

온프레미스 LLM의 새로운 표준

Puteron AI

Puteron AI는

"온프레미스 LLM 인프라 = 복잡한 DIY"라는 공식을 뒤집어,

"GPU가 달린 고성능 서버에 모델·운영·모니터링까지 한 번에 끝“

이라는 단일 어플라이언스로 재정의합니다.

귀사의 LLM 여정을 가속화하고 AI 시대의 경쟁 우위를 확보할 수 있는 Puteron AI를 경험해보세요.

온프레미스 LLM의 새로운 표준

Puteron AI

귀사의 LLM 여정을 가속화하고 AI 시대의 경쟁 우위를 확보할 수 있는 Puteron AI를 경험해보세요.

홈페이지

02-555-4847

www.slexn.com

bizops@slexn.com

문의하기