Lablup® Backend.AI®

복잡한 GPU 운영을 단순화시키는 AI 인프라 플랫폼

Backend.AI: 하이퍼스케일 딥러닝과 계산과학 분야를 위한 궁극의 솔루션

AI 시대에 엔터프라이즈 조직이 직면한 가장 큰 과제 중 하나는 막대한 AI 인프라를 효율적으로 운영하면서도 안정적인 성능을 동시에 확보하는 것입니다. Backend.AI는 AI 인프라에서 최대의 가치를 끌어낼 수 있도록 성능과 확장성, 신뢰성을 하나의 플랫폼에 담아 제공하는 AI 인프라 운영체제(OS) 입니다. 현재 Backend.AI는 전세계 110여개 이상의 사이트에서 16,000대 이상의 GPU를 관리하며, 다양한 산업분야의 고객을 지원하고 있습니다.

예를 들어 통신 사업자들은 Backend.AI를 이용하여 GPU 구독 서비스 (GPU-as-a-Service)를 운영하고, 금융 기업은 고객의 민감한 금융 정보 보호를 위해 완전히 폐쇄된 환경에서 자체적인 거대언어모델(LLM)을 구축하고 있습니다. 제조 분야 기업은 제품 개발 프로세스에 AI를 통합하기 위해 Backend.AI를 이용하고, 의료기관은 환자의 민감 의료 영상 데이터를 안전하게 분석할 수 있는 기반을 마련하고 있습니다.

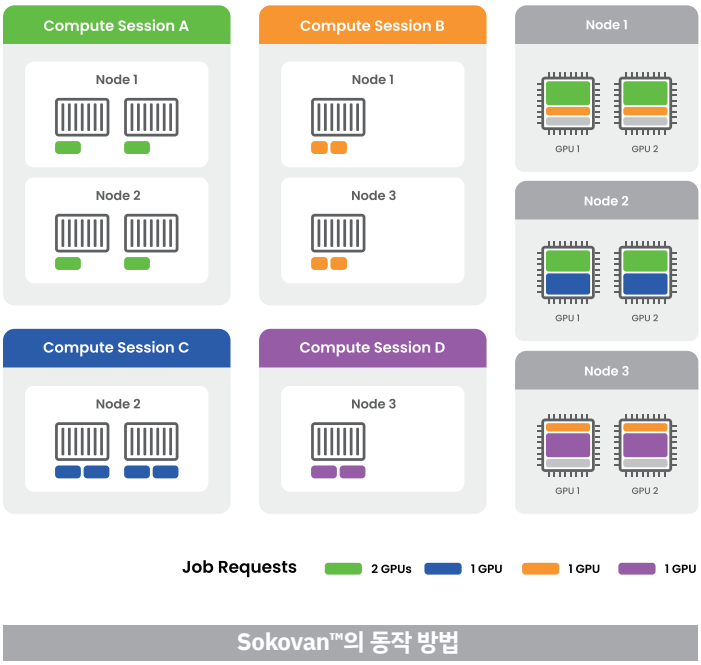
Backend.AI는 복잡한 인프라 관리의 부담을 덜어주고, 조직이 AI 가치 창출에 집중할 수 있도록 만들어주는 NVIDIA® DGX™-지원 소프트웨어입니다.

AI/HPC 워크로드를 위한 GPU 활용도 극대화

똑똑한 스케줄러와 Sokovan™, 최고의 소프트웨어가 선사하는 최상의 하드웨어 성능

Backend.AI Enterprise는 대규모 클러스터에서 GPU 활용을 최적화하는 고성능 스케줄러를 제공합니다. 컨테이너 수준 GPU 분할 가상화™와 정밀한 워크로드 수명 설정 기능으로 NUMA나 RDMA와 같은 고밀도 하드웨어를 효과적으로 관리하고 GPU 자원을 최적화할 수 있습니다.

래블업의 독자 개발 컨테이너 오케스트레이터 Sokovan은 컨테이너화된 환경에서 리소스 집약적인 배치 워크로드의 문제점들을 해결합니다. 기존의 배치 스케줄링 도구인 Slurm과 비교할 때, Sokovan은 가속기 인식 기능, 멀티 테넌트 지원, 배치 중심 작업 스케줄링을 제공하며 하드웨어 가속을 통합하여 AI 워크로드의 성능을 크게 향상시킵니다. Backend.AI의 고급 스케줄러와 Sokovan의 시너지는 최신 하드웨어의 잠재력을 최대한 활용하면서 AI와 HPC 워크로드 모두에 탁월한 성능과 효율을 제공합니다.



Sokovan™의 동작 방법

NVIDIA GB200 Grace Blackwell의 능력을 극대화하는 Backend.AI

NVIDIA GB200 NVL72 시스템은 LLM 추론, 검색증강생성(RAG), 데이터 처리 분야에서 누구와도 비교할 수 없는 성능을 제공합니다. NVIDIA MGX™ 아키텍처 기반 단일 노드 확장 설계를 통해 기존에 고객이 보유하고 있는 데이터센터 인프라에 손쉽게 통합될 수 있는 다양한 시스템 구성과 네트워킹 옵션을 지원합니다.

Backend.AI는 소프트웨어 정의형 AI 인프라를 제공합니다. NVIDIA GB200 Grace Blackwell에서 검증된 Backend.AI는 규모에 관계없이 안정적인 성능을 보장합니다. NVIDIA GB200 Grace Blackwell 전체에 걸쳐 워크로드를 수평으로 확장하면서 시스템 통합을 유지할 수 있습니다. Backend.AI와 함께 인프라 병목을 제거하세요.

확장성의 한계를 넘어도 사용성은 그대로

Backend.AI는 단일 노드부터 수십에서 수만, 심지어는 수십만 대에 이르는 대규모 멀티 노드 환경에서도 일관적인 사용성을 제공하도록 설계되었습니다. 수십만 대의 GPU를 관리하는 경우에도 전체 시스템을 쉽게 조망하고, 각 리소스에 대한 제어를 유지할 수 있습니다. 유연성을 잃지 않으면서도 효율성을 극대화하는 Backend.AI의 설계 철학을 기반으로 래블업은 시장에 출시된 11종 이상의 GPU를 Backend.AI에서 이기종 구성으로 이용할 수 있도록 최적화하였고, AI 워크로드 특성에 맞춰 최적화된 운영을 제공합니다.

GPU 리소스 할당 및 스케줄링 최적화

Backend.AI는 멀티 노드 지원, 멀티테넌트 격리, 그리고 래블업이 특허를 보유하고 있는 컨테이너 수준 GPU 분할가상화 (Container-level GPU virtualization™)를 결합하여, 단일 GPU에서 다양한 크기와 개수의 워크로드를 동시에 실행할 수 있습니다. GPU 동적 할당 (Dynamic GPU allocation)을 통해 작업 스케줄링 시 세션을 동적으로 생성하고 삭제하며, 해당 동작이 이뤄지는 즉시 리소스가 할당되거나 회수됩니다. 이 과정에서 실행중인 서비스의 중단 없이 온디맨드 방식으로 GPU 노드의 가상화된 분할 영역에 대한 세분화된 접근 제어를 제공합니다. 이러한 아키텍처를 통해 Backend.AI는 다양한 AI 모델의 동시 실행을 가능하게 하면서도 클라우드와 온프레미스 모두 적용 가능한 확장성과 워크로드 격리를 보장합니다.

정밀한 스케줄링을 위한 논리적 리소스 그룹

Backend.AI의 리소스 관리 아키텍처는 클러스터를 독립적 스케줄러를 가진 논리적 리소스 그룹으로 분할할 수 있습니다. 뛰어난 제어 능력을 바탕으로 그룹 단위 맞춤형 정책을 적용하면서도 인프라 유연성을 유지하도록 설계된 Backend.AI는 사용자와 관리자의 요구사항을 동시에 만족시킵니다.

에이전트 셀렉터 (Agent Selector) 및 작업 우선순위 (Job Priority)를 조합하면 워크로드에 따른 세분화된 스케줄링 정책을 구현할 수 있습니다. Backend.AI는 배치 작업, 추론 세션, 대화형 컴퓨팅 세션과 같은 다수 워크로드 타입에 대해 FIFO, DRF (자원 공정성 알고리즘), 사용자 정의 스케줄링 알고리즘을 리소스 그룹별로 선택 적용, 각 그룹 내에서 최적화된 워크로드 분산을 통해 운영 효율성을 극대화합니다.

컨테이너와 스토리지의 스마트한 고속 연결

작은 개발환경 트래픽부터 대규모 학습 데이터 전송까지 다룰 수 있는 방화벽 규칙과 네트워크 QoS 설정의 복잡한 작업도 Backend.AI를 사용하면 간편해집니다. Backend.AI는 API 서버와 별도로 실행되는 전용 프록시 서비스를 통해 컨테이너 내 앱(Jupyter, SSH, Visual Studio Code 등)과 네트워크 공유 스토리지 파일 시스템에 접근함으로써 I/O 병목 현상을 획기적으로 감소시킵니다.

Backend.AI는 고성능 스토리지 솔루션의 가속 기능을 통합하여 초고속 스토리지를 고객에게 더 쉽게 제공, Dell PowerScale, VAST Data, WekaFS, NetApp 등의 통합 스토리지 솔루션으로부터 최대한의 성능을 이끌어낼 수 있습니다. 또한 Backend.AI는 NVIDIA® Magnum IO GPUDirect® Storage를 지원하여 GPU가 CPU를 거치지 않고 네트워크 스토리지에서 데이터를 메모리로 가져올 수 있게 합니다.

NGC™ 통합 및 사전 구성된 환경으로 손쉬운 AI 솔루션 제공

Backend.AI를 통해 지속적으로 발전하는 소프트웨어 스택을 쉽게 관리할 수 있습니다. Backend.AI는 엔터프라이즈 수준의 호환성 테스트를 거친 다양한 사전 구성 환경을 제공하여 사용자에게 편리한 AI 환경을 제공합니다. 여기에는 래블업의 AI 소프트웨어 스택, NGC™ 카탈로그 이미지, NVIDIA NIM™ 컨테이너에 대한 완전한 지원이 포함됩니다.

Backend.AI는 NIM 지원을 통해 추론 마이크로서비스를 원활하게 실행, 관리 및 확장하며 리소스 최적화를 완전 자동화합니다. 이를 통해 사용자는 최신 딥러닝 및 추론 환경에 즉시 접근할 수 있으며, 각자의 필요에 맞게 맞춤화된 환경을 이용할 수 있습니다.

Backend.AI에 대해 더 알아보고 싶으시다면, backend.ai를 방문하세요.

Backend.AI 설치 요구 사항

설치 요구사항은 상황에 따라 변경될 수 있습니다. 최신 설치 요구 사항은 bnd.ai/requirements 페이지에서 확인하실 수 있습니다.

소프트웨어	최소 사양	권장 사양
Operating System	Ubuntu 22.04 RHEL 8	Ubuntu 24.04 RHEL 9+ / Alma Linux 9+
Docker Engine	20.10	25.0+
CUDA	11.0	12.8*
PostgreSQL	12.0	16.0
Redis	6.2	7.2

* 오픈 소스 버전 드라이버 사용을 권장합니다.

서비스 노드 유형	CPU**	RAM**	Disk**
Manager	2-16	2 GiB – 16 GiB	100 GB – 500 GB
Agent	1	512 MiB – 1 GiB	20 GB – 2 TB
App Proxy, Web Server	1-8	2 GiB – 16 GiB	10 GB
Storage Proxy	1-8	4 GiB – 16 GiB	10 GB (excl. storage volume)
Container Registry	1-4	2 GiB – 8 GiB	500 GB – 10 TB

** 범위 값은 최소 및 권장 용량을 나타냅니다.