ORACLE CLOUD Infrastructure

AI 훈련과 추론 속도의 혁신적 향상, OCI의 NVIDIA GPU로 실현하기

생성형 AI와 도메인 특화 AI 모델이 우리 시대에 크나큰 혁신을 가져왔습니다. 폭발적으로 증가하고 있는 다양한 요구사항들과 용처에 부합하는 최적의 모델을 제공할 수 있도록, Oracle Cloud Infrastructure(OCI)와 NVIDIA는 지속적인 파트너쉽을 체결했습니다. 이를 통해 클라우드 환경에서 안정적인 GPU 가속 AI와 HPC 솔루션을 제공함으로써, 고객이 유의미한 인사이트를 도출하고, 실현 가능한 비즈니스 기회를 발굴하고, AI 기반 제품과 서비스를 개발할 수 있도록 지원하고 있습니다.

OCI 환경에서의 자연어 처리(NLP) 기술의 발전

대규모 언어 모델(LLM)은 번역, 챗봇, 개인 비서, 문서 요약, 기사 작성 등 자연어 처리(NLP) 기술을 사용하는 애플리케이션을 획기적으로 발전시키고 있으며, 나아가 의료, 금융, 에너지, 리테일 등의 산업에도 큰 변화를 가져오고 있습니다.이러한 AI 모델들을 활용하기 위해서는 수십억 개의 매개변수를 최적화해야 하고, 대규모 데이터 세트를 통한 학습 또한 필수적으로 진행되어야 하죠.한때 온프레미스 하드웨어에서만 사용 가능했던 대규모 NVIDIA GPU의 방대한 컴퓨팅 성능을 이제 OCI의 범용 GPU 인스턴스와 고성능 GPU 클러스터로 구성된 OCI AI 인프라 서비스를 통해서도 이용할 수 있게 되었습니다.

모델 학습 및 병렬 애플리케이션을 지원하는 OCI Supercluster

OCI의 NVIDIA GPU 기반 베어 메탈(BM) 및 가상 인스턴스(VM)는 스타트업이 더 빠르게 혁신을 달성할 수 있도록 지원합니다.

더 알아보기

해당 인스턴스들은 AI 기업이 필요로 하는 머신러닝, 이미지 처리, 모델 학습, 추론 연산, 물리 기반 모델링 및 시뮬레이션, 대규모 병렬 HPC 애플리케이션 실행 등을 위한 고성능 컴퓨팅 플랫폼을 제공합니다.

각 BM.GPU.A100-v2.8 인스턴스는 8개의 NVIDIA A100 GPU를 탑재하고 있으며, 마이크로초 단위의 지연 시간으로 RoCE(RDMA over Converged Ethernet) 기술을 사용하는 Oracle의 저지연 클러스터 네트워킹을 활용합니다. 현재 NVIDIA A100 GPU는 최대 32,768 개까지 확장할 수 있으며, 앞으로는 NVIDIA H100 GPU를 최대 16,384개까지 확장할 수 있는 신규 클러스터도 제공될 예정입니다.

전담 엔지니어링 지원

Oracle Cloud Infrastructure(OCI)는 세계적 수준의 기술 전문가들을 통해 비즈니스 계획부터 출시까지 클라우드 배포의 모든 단계에서 복잡한 기술적 장벽을 제거함으로써, 고객이 성공적으로 클라우드를 도입하고 활용할 수 있도록 보장합니다.

특히 AI 스타트업을 위해서는 Oracle의 전담 엔지니어링 지원팀이 자동화된 Terraform 배포를 사용하여 클러스터 설정을 지원하고 있습니다.

비용 효율성 개선

OCI는 각 산업의 복잡한 수학적, 과학적 문제를 AI 학습을 통해 해결하는 과정에 기여하는 강력하면서도 비용 효율적인 컴퓨팅 성능을 제공합니다.

Oracle과 타사의 AI 인프라 서비스 비용 비교해 보기 →

머신러닝, 이미지 처리, 대규모 병렬 고성능 컴퓨팅 작업 등을 활용하는 애플리케이션은 NVIDIA GPU를 활용하여 해당 기술들과 관련된 복잡한 문제를 해결하고 혁신을 가속화할 수 있습니다.

GPU 가격 정책 및 사양 살펴보기 →

OCI는 클라우드 도입을 단순화하기 위해 전 세계의 모든 리전에 일관된 가격 모델을 적용합니다.

Oracle의 경쟁력 있는 글로벌 가격 정책 알아보기 →

OCI AI Infrastructure

소규모 AI 학습, 추론, 스트리밍, 게이밍, VDI

VM GPU.A10.1	VM GPU.A10.2	BM GPU.A10.4
A10	A10	A10
1	2	4
24 GB	48 GB	96 GB
$2.00 per GPU/hr	$2.00 per GPU/hr	$2.00 per GPU/hr

대규모 확장형 AI 학습

BM GPU4.8	BM GPU.A100-v2.8
A100	A100
8	8
320 GB	640 GB
1.6 Tb/sec RDMA	1.6 Tb/sec RDMA
$3.05 per GPU/hr	$4.00 per GPU/hr

다수의 노드로 처리해야 하는 긴밀하게 결합된 워크로드 지원

VM 가상머신 컴퓨팅

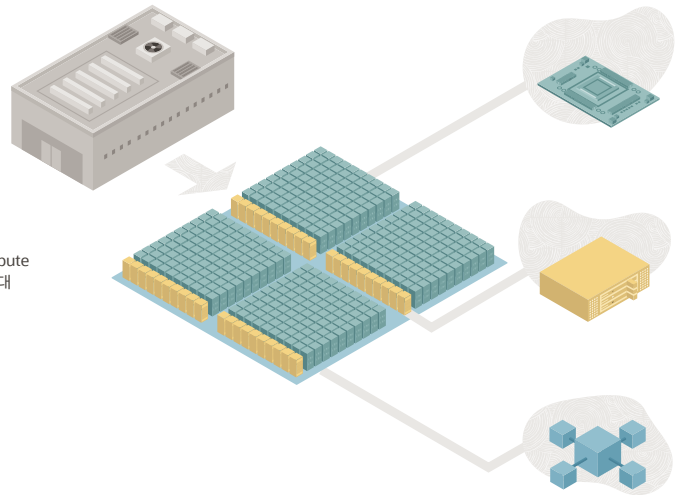
BM 베어메탈 컴퓨팅

GPU 코어

RDMA를 통한 RoCE 기반 노드 간 네트워크(마이크로초 단위 지연 시간)

* NVIDIA V100 및 P100을 탑재한 구형 인스턴스도 제공

OCI Supercluster



Oracle Cloud 리전

- 최대 4,096개 OCI Compute 베어메탈 인스턴스로 최대 32,768개 NVIDIA A100 GPU 지원

컴퓨트

- 노드당 8개 NVIDIA A100 80GB GPU
- 노드당 2개 64코어 3세대 AMD EPYC CPU
- 노드당 1개 2TB DDR4 메모리
- 노드당 8개 200Gb/초 클러스터 NIC

스토리지

- 블록 스토리지: 볼륨당 최대 32TB
- 객체 스토리지: 객체당 최대 10 TiB
- 파일 스토리지: 파일 시스템당 최대 8EB
- WEKA, BeeGFS, Lustre, IBM Spectrum Scale 등의 다양한 HPC 파일 시스템

네트워킹

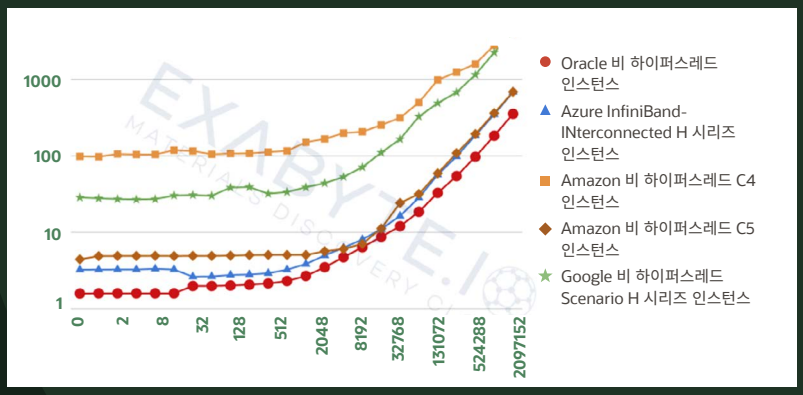
- 마이크로초 단위의 노드 간 지연 시간
- 초당 1,600 Gb의 OCI Supercluster 인터노드 대역폭, NVIDIA A100 GPU 지원

고속 RDMA 클러스터 네트워크

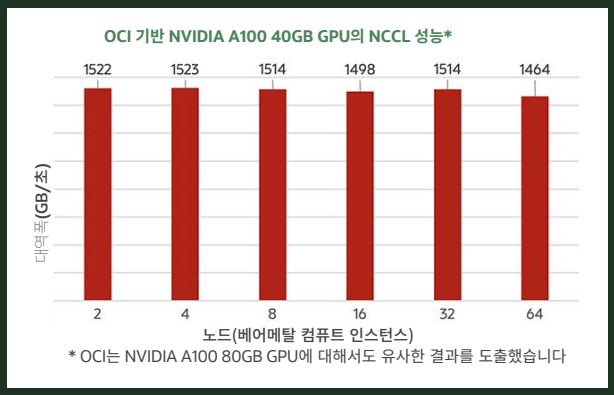
OCI의 베어 메탈 서버와 Oracle의 클러스터 네트워킹은 초저지연 RoCE v2(RDMA over Converged Ethernet)를 제공합니다.

1. Exabyte.io의 분석에 따르면, OCI는 1.7 마이크로초의 지연 시간을 제공하며, 이는 여타 어느 클라우드 공급업체들보다 빠른 속도입니다. OCI는 원격 직접 메모리 액세스(RDMA) 클러스터를 활성화하여 NVIDIA A100 GPU가 장착된 베어 메탈 서버에 대한 클러스터 네트워킹을 확장했습니다.

2. Oracle의 별도 분석에 따르면, OCI는 512개의 NVIDIA A100 GPU에 대해 91.5%의 대역폭 활용도를 제공합니다. 이는 네트워크 오버헤드가 거의 없어 HPC 및 AI 훈련 애플리케이션이 GPU 클러스터의 처리 능력을 거의 완벽하게 활용할 수 있음을 의미합니다.



1



2

ORACLE CLOUD Infrastructure | NVIDIA

OCI의 베어 메탈, 가상 머신, 클러스터 네트워킹, 스토리지로 구성된 AI 인프라와 결합된 NVIDIA 소프트웨어는 대규모 AI 학습 및 딥러닝 추론을 위한 폭넓고 접근하기 쉬운 포트폴리오를 제공합니다. OCI는 다음과 같은 다양한 NVIDIA 소프트웨어들을 지원합니다. OCI는 다음을 비롯한 다양한 NVIDIA 소프트웨어들을 지원합니다.

- **NVIDIA AI Foundations** – 생성형 AI 모델 제작 서비스
- **NVIDIA DGX Cloud** – 다중 노드 AI 훈련 서비스
- **NVIDIA AI Enterprise** – AI 워크플로를 위한 처리 엔진
- **NVIDIA RAPIDS** – Apache Spark 데이터 처리 가속화
- **NVIDIA Clara** – 헬스케어 AI 및 HPC 애플리케이션 프레임워크

"AI 기반 혁신이 제공하는 무한한 기회가 거의 모든 비즈니스의 전환을 지원하고 있습니다. NVIDIA와 Oracle Cloud Infrastructure(OCI)의 협업으로 모든 기업이 NVIDIA의 가속 컴퓨팅 플랫폼이 제공하는 탁월한 수퍼컴퓨팅 성능을 누릴 수 있게 되었습니다."

Manuvir Das
Vice President of Enterprise Computing
NVIDIA

더 알아보기

OCI Supercluster 및 AI Infrastructure 살펴보기

OCI Compute GPU 인스턴스 더 알아보기

OCI Marketplace에서 NVIDIA GPU 클라우드 머신 이미지 실행하기

NVIDIA NGC Catalog에서 GPU 최적화 소프트웨어 둘러보기

OCI 무료 체험하기

문의하기

한국오라클 대표번호 02-2194-8000, 또는 oracle.com/kr 웹사이트를 통해 Oracle 담당자에게 연락하실 수 있습니다.

북미 지역 외 국가인 경우 oracle.com/contact 에서 현지 지사를 찾을 수 있습니다.