LLM 운영,
Red Hat으로 완성하기

OpenShift AI부터 vLLM까지

이명진
Team leader of Specialist Solution Architecture
Red Hat Korea

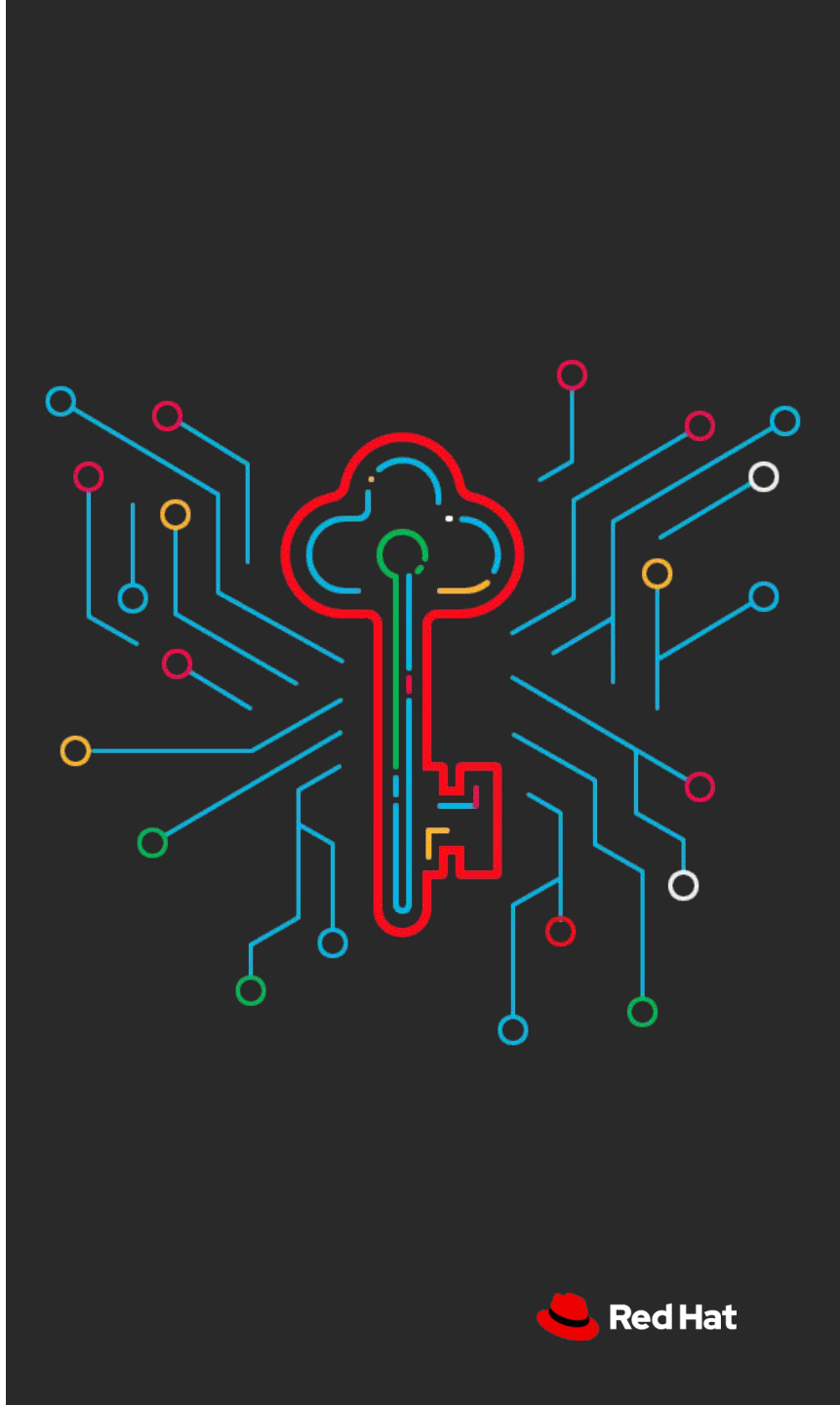
OpenInfra Days Korea
2025

Red Hat

발표 순서

- Red Hat이 AI 회사인가요?
- Red Hat AI를 통한 효율적인 LLM 운영: Inference
  : ①추론 효율 극대화 ②모델 추론 속도 향상 ③모델 양자화 ④추론 지원 도구
- 요약 및 결론

Red Hat이 AI 회사인가요?



1. Red Hat은 AI 회사인가요?

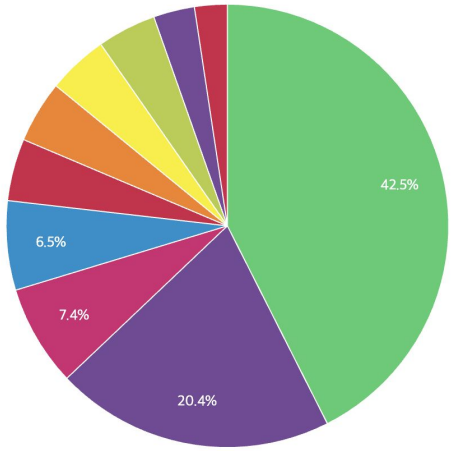
OpenStack에 대한 Red Hat의 기여

Commits by Company

Show 10 entries Search

#	Company	Commits
1	Red Hat	81333
2	Rackspace	29483
3	Canonical	25964
4	IBM	18111
5	Mirantis	18072
	*independent	17574
6	HP	17188
7	NEC	11992
8	Huawei	9531
9	Intel	8937

Showing 1 to 10 of 454 entries Previous Next



2010 NASA and Rackspace Hosting launches OpenStack
2012 Red Hat launches preview with “Essex”
2013 Red Hat launches commercial support with “Grizzly”
2015 Red Hat enables enterprise cloud with Cloud Suite
2017 Red Hat powers flexible, scalable hybrid clouds
2019 Red Hat launches Red Hat OpenStack Platform 15
2020 Red Hat brings long life support with version 16
2021 Red Hat extends OpenStack to the edge
2022 Red Hat helps enhance private cloud security
2025 More to come!

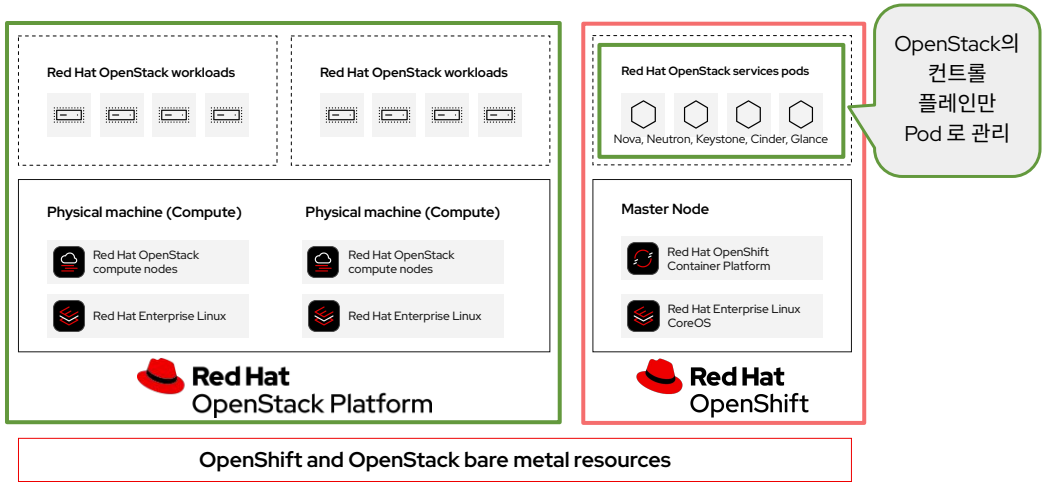
Red Hat OpenStack Services on OpenShift

Press releases > Red Hat OpenStack Services on OpenShift is Now G...

Red Hat OpenStack Services on OpenShift is Now Generally Available

Open source leader brings modern approach to OpenStack deployments, helping organizations build future-ready networks at massive scale

RALEIGH, N.C. - August 26, 2024 – Red Hat, Inc., the world's leading provider of open source solutions, today announced the general availability of Red Hat OpenStack Services on OpenShift*, the next major release of Red Hat OpenStack Platform. This is a significant step forward in how enterprises, particularly telecommunication service providers, can better unify traditional and cloud-native networks into a singular, modernized network fabric. Red Hat OpenStack Services on OpenShift opens up a new pathway for how organizations can rethink their virtualization strategies, making it easier for them to scale, upgrade and add resources to their cloud environments.



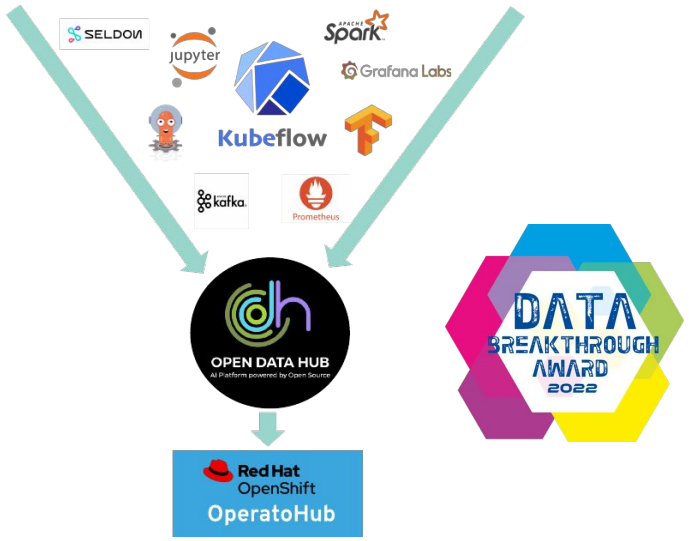
Source
- Stackanalytics, last updated on 19 August 2025 https://www.stackalytics.io/?release=all&metric=commits

Red Hat

Red Hat의 오픈 소스 AI 히스토리

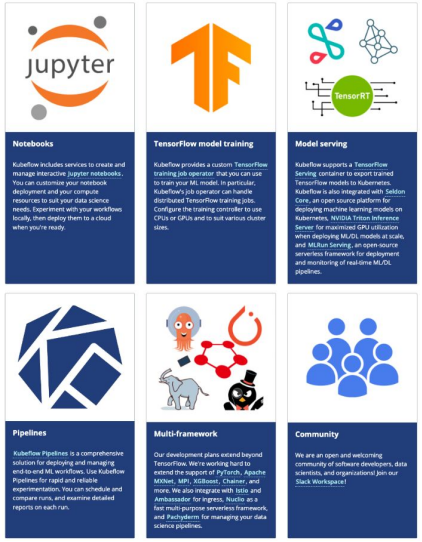
2015년

오픈 데이터 허브(Open Data Hub) 운영



2019년

쿠베플로우(Kubeflow) 컨트리뷰터



Training Operators

Training of ML models in Kubeflow through operators

- **TensorFlow Training (TFJob)** Using TFJob to train a model with TensorFlow
- **PaddlePaddle Training (PaddleJob)** Using PaddleJob to train a model with PaddlePaddle
- **PyTorch Training (PyTorchJob)** Using PyTorchJob to train a model with PyTorch
- **MXNet Training (MXJob)** Using MXJob to train a model with Apache MXNet
- **XGBoost Training (XGBoostJob)** Using XGBoostJob to train a model with XGBoost
- **MPI Training (MPIJob)** Instructions for using MPI for training
- **Job Scheduling** How to schedule a job with gang-scheduling

2024년

뉴럴 매직(Neural Magic) 인수

NEURAL MAGIC

vLLM

- **nm-vllm** Enterprise inference server for LLMs on GPUs.
- **Neural Magic Compress** Developer subscription for enterprises aiming to build and deploy efficient GenAI models.
- **DeepSparse** Sparsity-aware inference server for LLMs, CV and NLP models on CPUs.

Source:
- Driving Innovation with open source AI/ML, Red Hat Blog (April 22, 2022)
- Open source AI at Red Hat: Our journey in the Kubeflow community, Red Hat Blog (March 19, 2024)
- Red Hat Announces Definitive Agreement to Acquire Neural Magic, Red Hat Press releases (November 12, 2024)

Red Hat

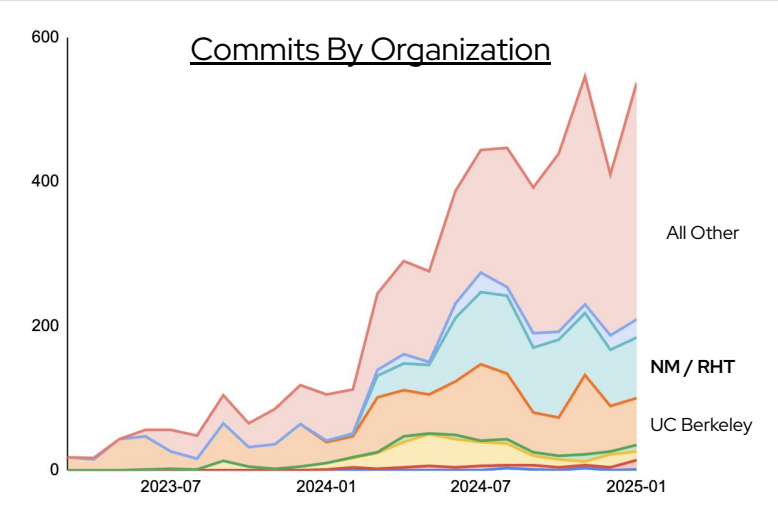
뉴럴 매직(Neural Magic): OSS GenAI Inference 리더

- 2018년 MIT 전기전자·컴퓨터공학(EECS) 그룹에서 스핀아웃
- 고성능 컴퓨팅(HPC) 팀을 구성해 딥러닝 CPU 추론 컴파일러를 개발
- 2023년 오픈소스 LLM의 등장과 함께 GPU 추론으로 개발 역량을 재집중
- 현재 vLLM의 TOP 컨트리뷰터로서, vLLM의 고성능 모델 실행과 관련된 핵심 이니셔티브를 주도

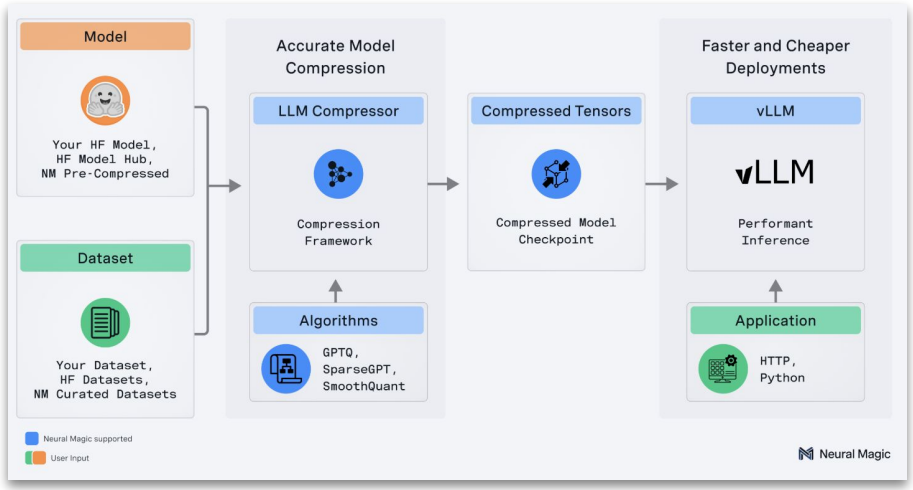
Optimized Model Hub

Llama, Qwen, Gemma, Ai2, IBM, Mistral, DeepSeek, Phi, Microsoft, NVIDIA

Neural Magic Community Contribution



LLM Compression Tools



Red Hat

1. Red Hat은 AI 회사인가요? Red Hat이 진정한 오픈 소스 AI-First 기업이 되려는 이유

Red Hat

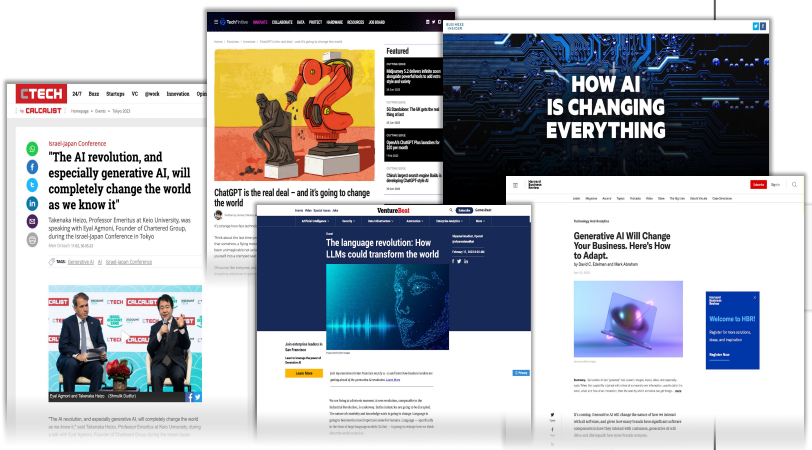
vLLM

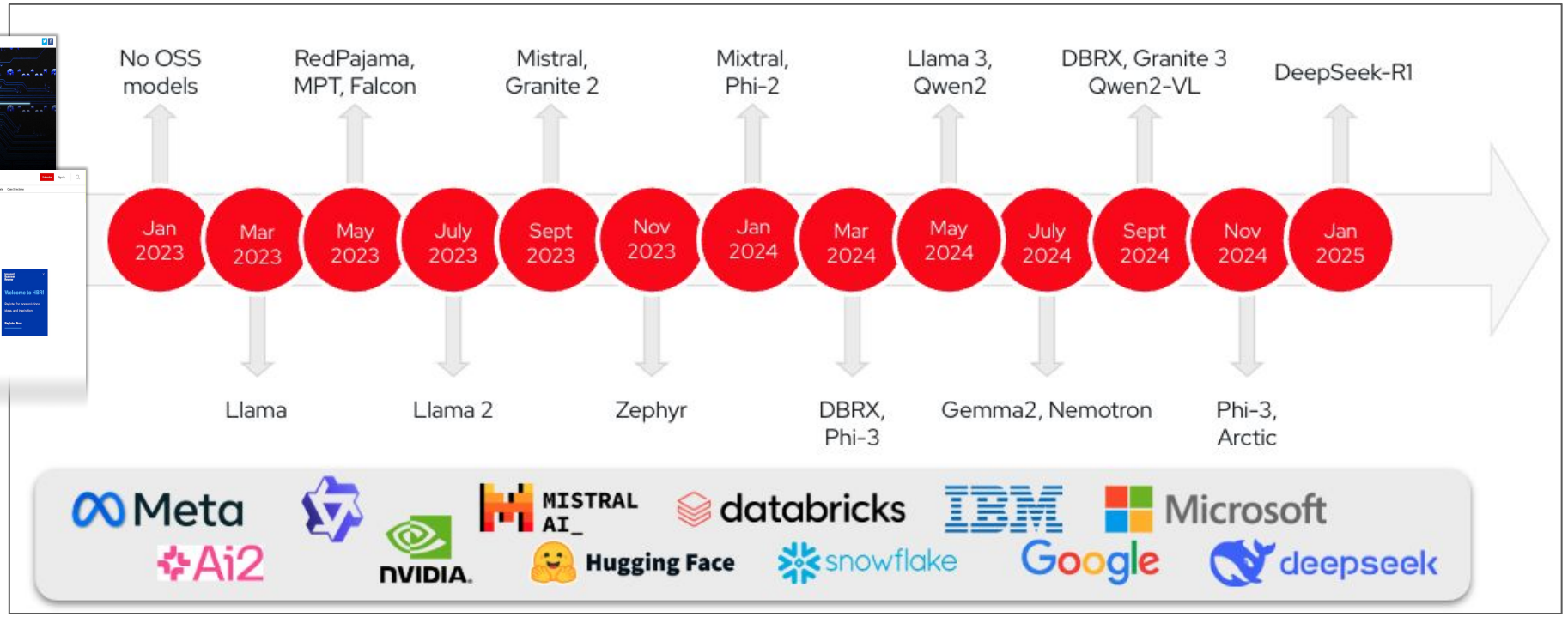
rebellions_

Together with SqueezeBits PyTorch Korea User Group

Red Hat

The Power of Open (1/2)

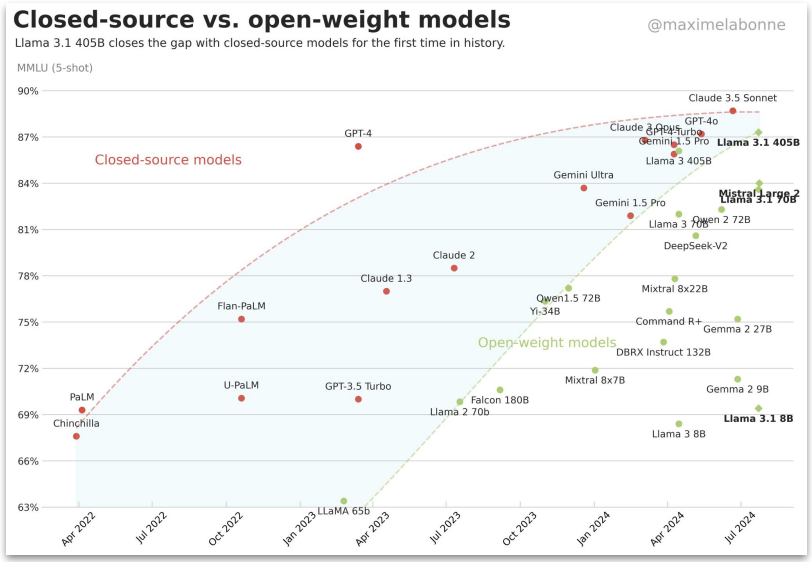




- 2023년과 2024년 초에는 클로즈드(Closed) 모델이 오픈(Open) 모델을 크게 앞서감
- 지난 2년(2024년-2025년) 동안 오픈소스 AI의 역량이 폭발적으로 증가함

Red Hat

The Power of Open (2/2)



Meta Llama

- 2024년도에 650M 다운로드
- 85,000 Llama 파생 모델
- 1B, 3B, 8B, 70B, 405B variants
- Multilingual, Multimodal, Mobile

deepseek

- OpenAI급 품질의 최초 추론 특화 모델
- 1-70B parameter distilled 버전
- Global market pandemonium?

온프레미스 기반 오픈 소스 모델 사용의 장점

비용(Cost)

- 셀프 관리형 인프라 (infrastructure)
- 1B-405B 규모 - 작업 난이도에 맞는 모델 선택

맞춤화(Customization)

- Task 특화 튜닝으로 정확도를 높이고 비용을 절감

제어(Control)

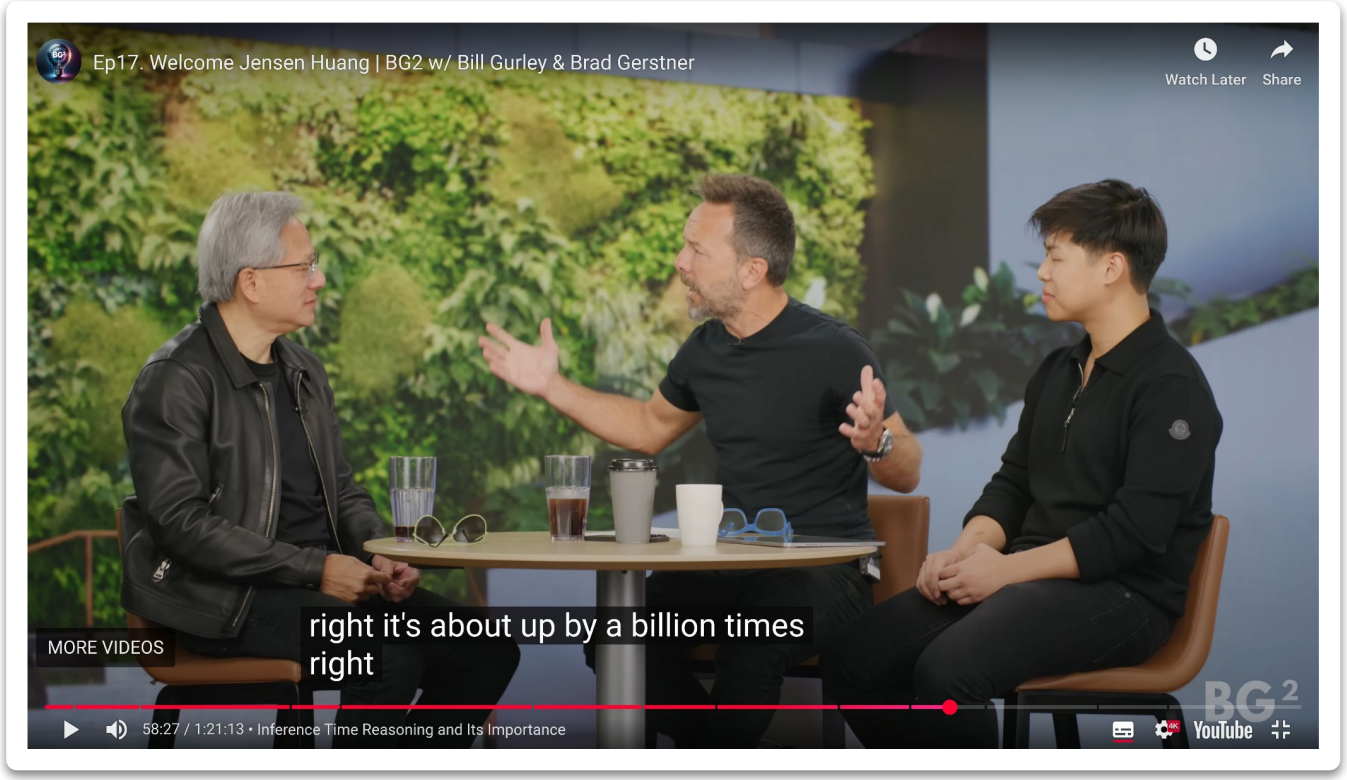
- 모델 라이프사이클 (기존 모델에 변경 없음)
- 리소스 (no rate limits / API downtime)

보안(Security)

- 완전한 데이터 프라이버시 (no 3rd party APIs)

Red Hat

In the beginning... It was all about
training and fine tuning

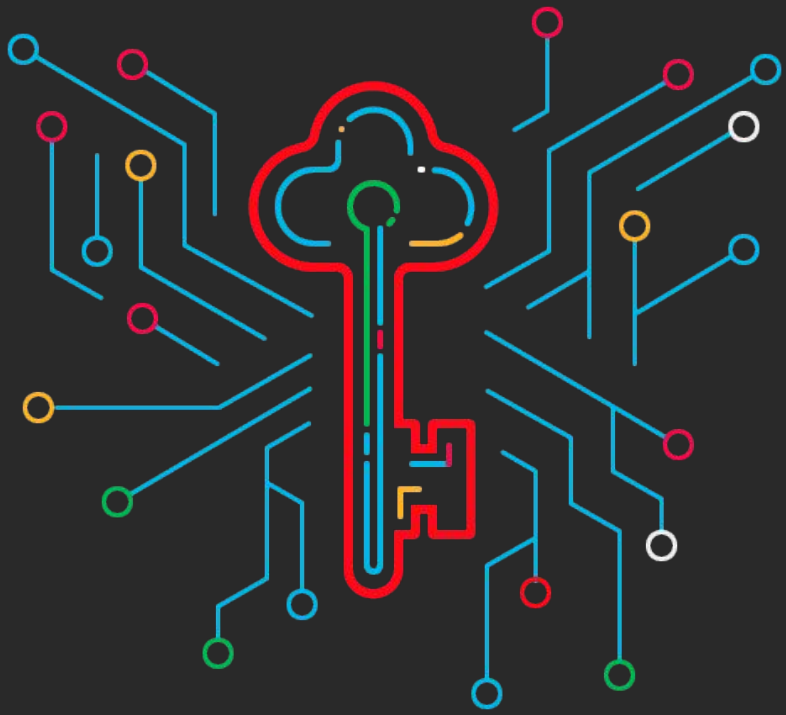


"[Inference] is about to go up by
a billion times."

https://www.youtube.com/watch?v=bUrCR4jQQg8

Red Hat

Red Hat AI를 통한 효율적인 LLM 운영



추론(Inference)이란?

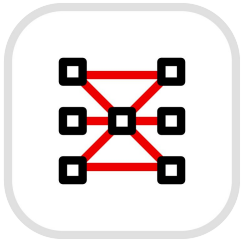
생성형 AI가 사용자 요청에 대해 빠르고 정확한 응답을 생성하는 핵심 과정

AI 앱 사용자

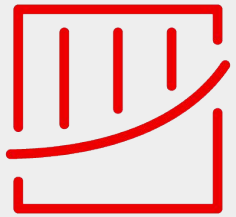
사용자가 request 전송

AI가 response 회신

AI Model은 inference server와 함께 answer를 처리



사용자의 기대를 충족시키고, 가장 정확한 응답을 최대한 빠르게 전달하는 것이 목표



추론 시장은 폭발적으로 성장 중

AI 추론 시장의 성장 추이:

2025년 현재 143조원에서 2030년 343조원에 달할 것으로 예상

Source:
AI Inference Market worth $254.98 billion by 2030, Feb 2025, PR Newswire

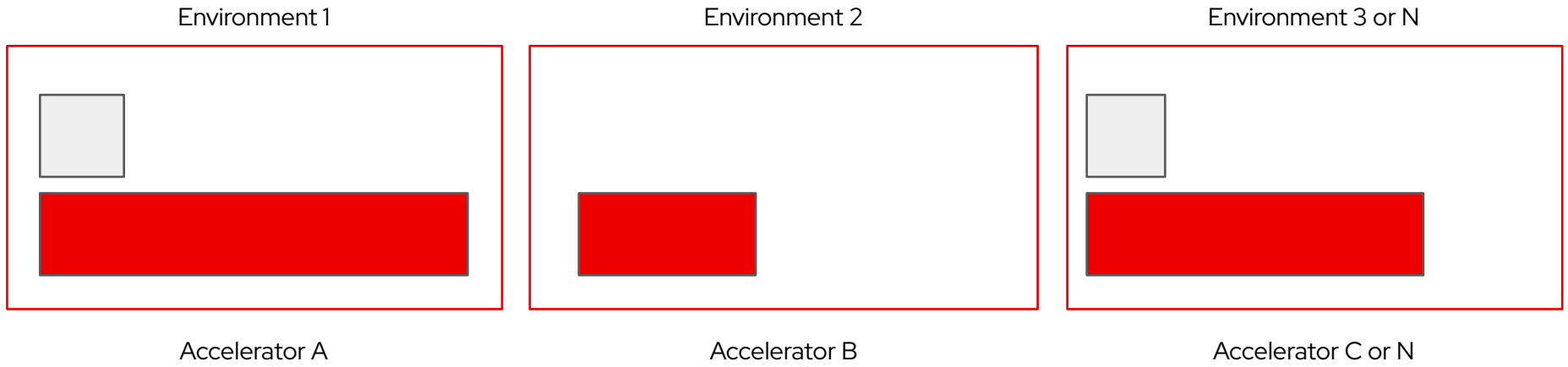
Red Hat

추론(Inference) 최적화는 매우 중요합니다.

추론은 상시 운영되는 생성형 AI의 핵심 단계이기에, **성능과 비용을 동시에 최적화**해야 함

Fine-tuning은 새로운 데이터에 맞추기 위해 **주기적**으로 반복 수행됩니다.

Inference는 **상시 실행**되며, 사용자 요청에 따라 자동으로 확장됩니다.



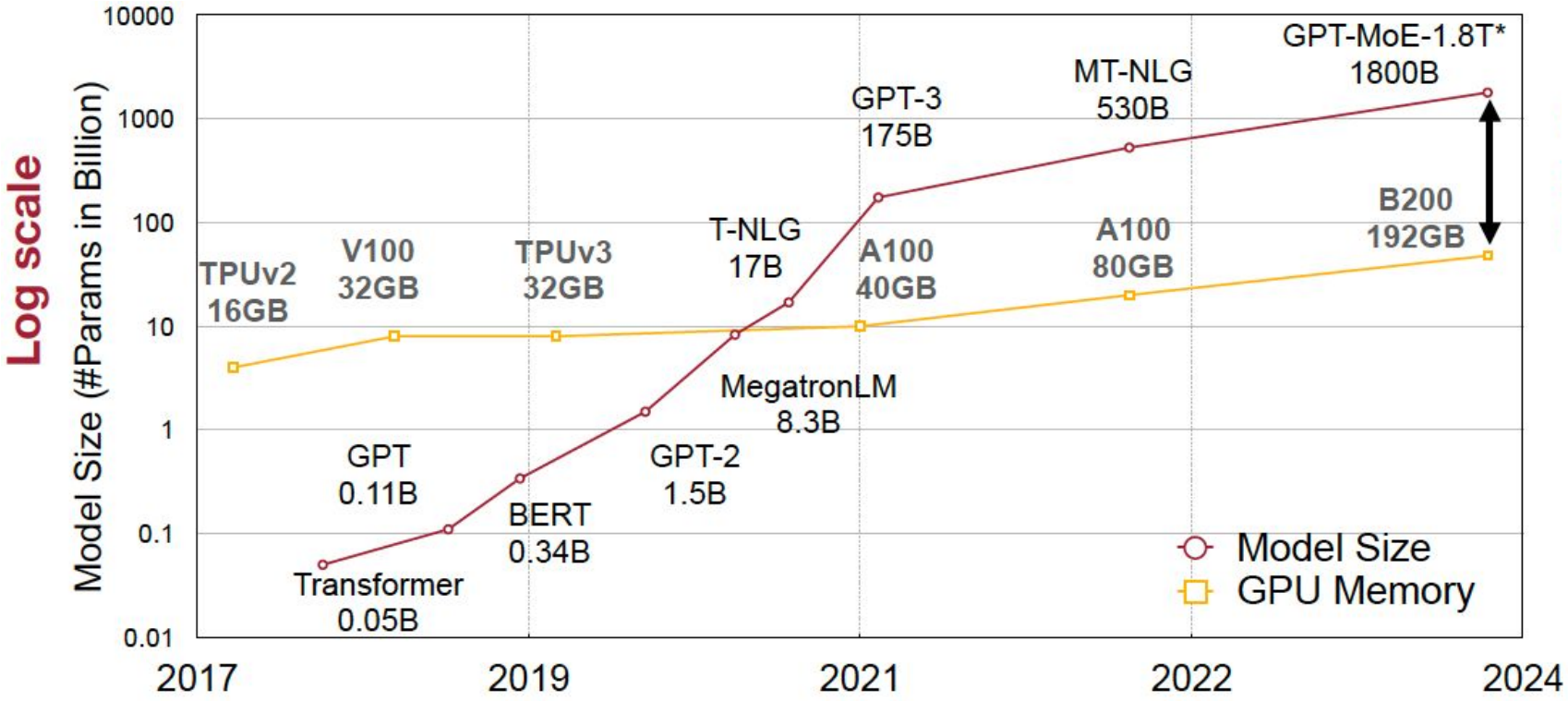
운영 단계에서의 추론의 어려움

- 전문적인 ML 엔지니어 부족: 모델 최적화·배포 어려움
- OS 모델 배포 지연: 혁신 속도 저하
- 비효율적인 GPU 사용: 추론 비용 증가
- 벤더 종속성: 유연한 배포 제한
- GPU 자원 낭비, 확장성 문제 발생
- 수동 양자화·압축으로 복잡성·리스크 증가

Source:
AI Inference Market worth $254.98 billion by 2030, Feb 2025, PR Newswire

추론(Inference) 최적화가 중요한 이유

대규모 모델을 효과적으로 처리하기 위한 메모리 효율성 향상 및 분산 처리 기술이 중요



- 모델 크기 증가 대비 GPU 메모리의 증가 비율 -

*: Jensen Huang, NVIDIA GTC 2024

모델 크기 성장에 비해 부족한 GPU 메모리 용량 문제를 해결하기 위해서는 **효율적인 알고리즘과 시스템이 요구**됩니다.

Source:
AI Inference Market worth $254.98 billion by 2030, Feb 2025, PR Newswire

Red Hat

모델 추론 최적화를 위한 요구사항과 해결방안

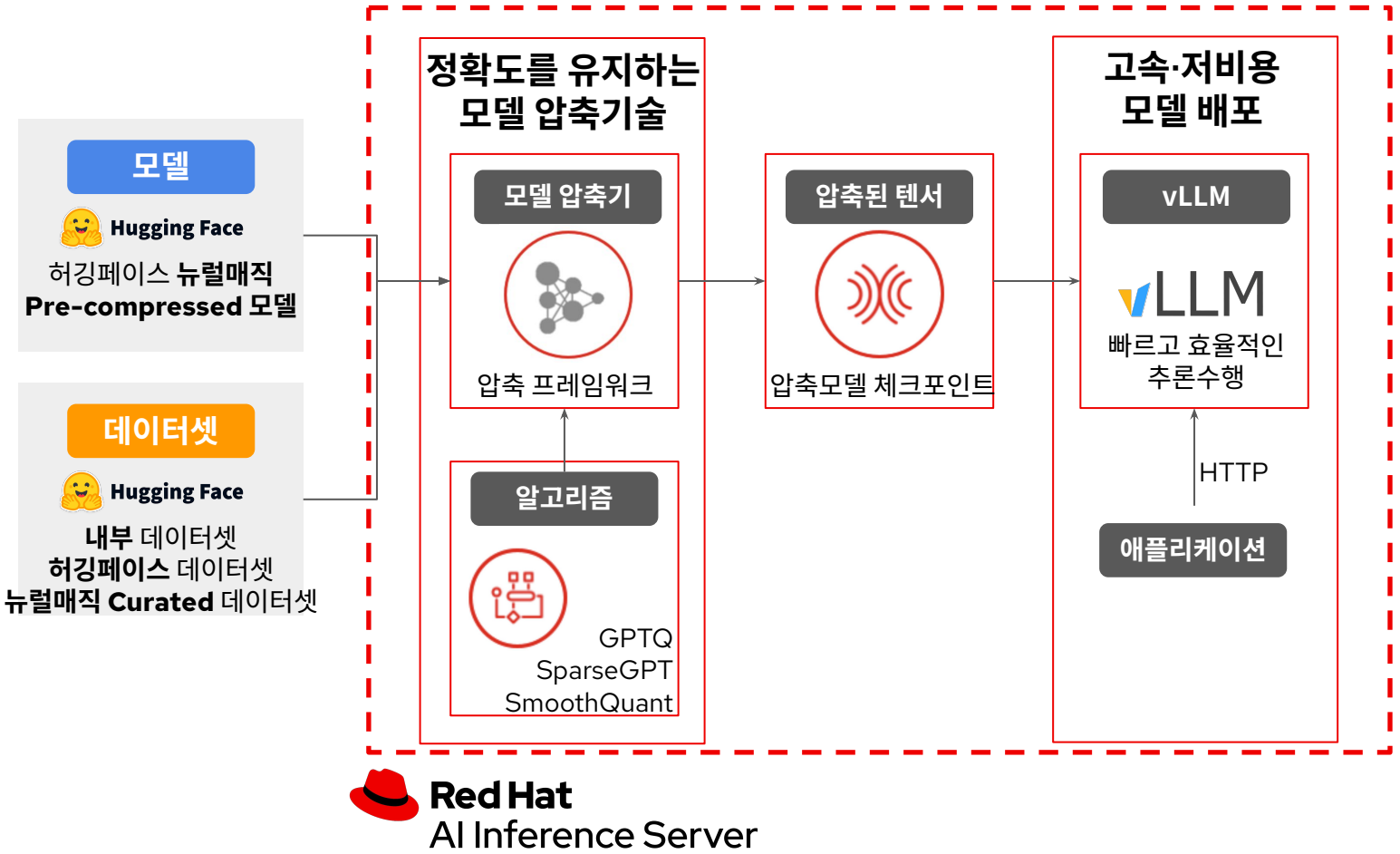
비용 효율적인 모델 / 추론 최적화를 위해서 vLLM과 함께 많은 도구들이 필요

1. 모델 성능 유지와 **추론 효율 극대화**
   - **vLLM**
2. 모델 병렬 처리 및 분산 추론을 통한 대규모 환경의 **모델 추론 속도 향상**
   - **Model parallelism**
   - **llm-d**
3. 모델 크기를 줄여 **경량화**하기
   - 모델 양자화
   - **llm-compressor**
4. **추론 지원 도구**
   - **Red Hat AI Repository**
   - **GuideLLM**
   - 주요 하드웨어 가속기 지원

Red Hat

기업 환경에 맞는 모델 추론 최적화

Red Hat AI Inference Server: 고속·저비용 LLM 운영을 가능하게 하는 압축/추론 최적화 스택



- 최적화 알고리즘을 통합된 인터페이스로 제공 (GPTQ, AWQ, SmoothQuant 지원)
- FP8, INT8, INT4 등 다양한 저정밀도 포맷 지원
- Hugging Face AutoModel과의 매끄러운 연동
- vLLM과 호환되는 safetensors 기반 체크포인트 형식 지원
- Hugging Face accelerate를 통한 대용량 모델 지원 가능

Red Hat

vLLM 개요

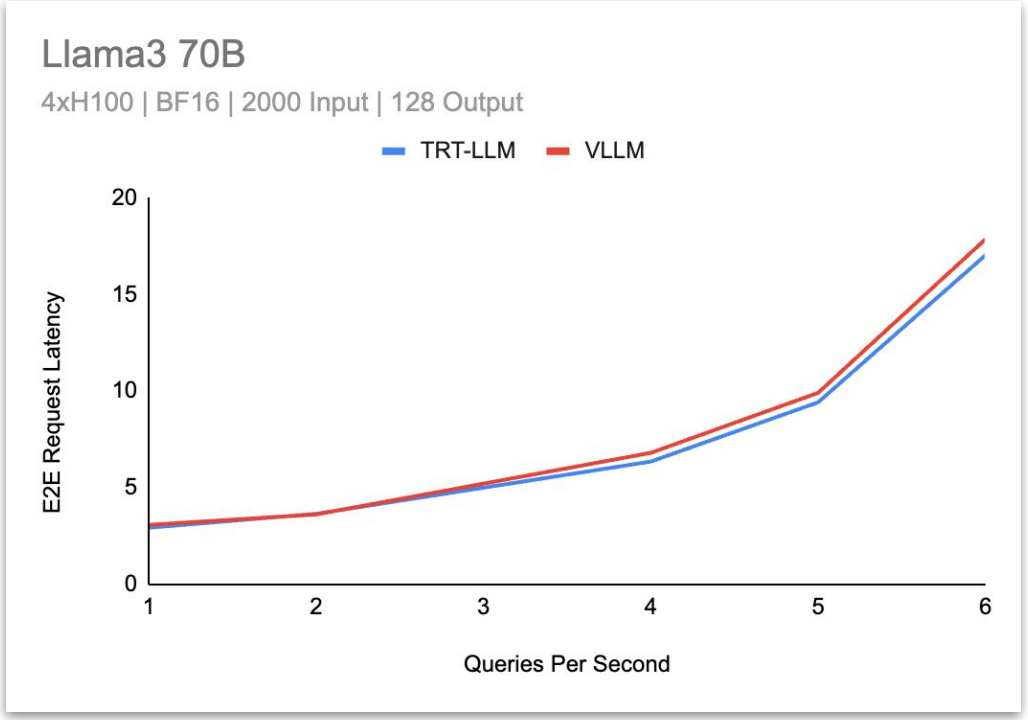
Gen AI에 필요한 LLM을 최적화하고 GPU 리소스를 절감해주는 오픈소스 추론 서버

- vLLM, 즉 Virtual Large Language Model은 vLLM 커뮤니티가 유지보수하는 오픈소스 코드 라이브러리

- **대규모 언어 모델(LLM)가 연산을 보다 효율적이고 대규모로 수행하게 할 수 있음**

- **GPU 유휴시간을 줄이고, 메모리 사용율을 향상**시키며, Gen AI 애플리케이션의 **결과 출력 속도를 높여 동시에 더 많은 요청을 처리하게 하는 추론 (Inference) 서버**

NVIDIA Stack과 vLLM 비교

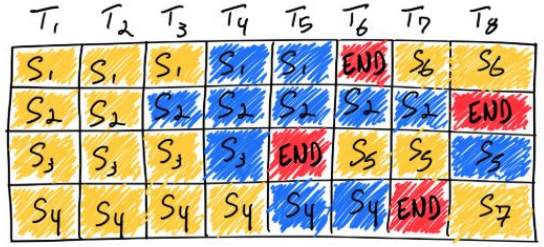


Red Hat

Continuous Batching

GPU 자원 활용을 극대화하고 지연을 최소화

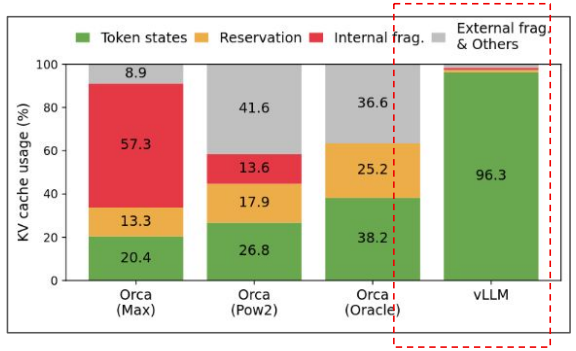
- **Batching 종류**



유형	설명	특징 및 장단점
Static Batching	클라이언트가 여러 요청을 미리 묶어서 서버에 전송	처리 완료 후 응답, 느리고 비효율적
Dynamic Batching	서버가 일정 시간(window) 또는 조건에 따라 요청을 묶음	짧거나 균일한 길이 요청에 적합
Continuous Batching	**진행 중인 배치와 새 요청을 동적으로 결합, 배치 크기 조절**	**자원 활용 극대화, 요청 일부 병렬 처리, 지연 최소화**

Paged Attention

GPU 메모리 단편화를 줄여 메모리 효율과 비용 절감을 돕는 가상 메모리 관리 기법

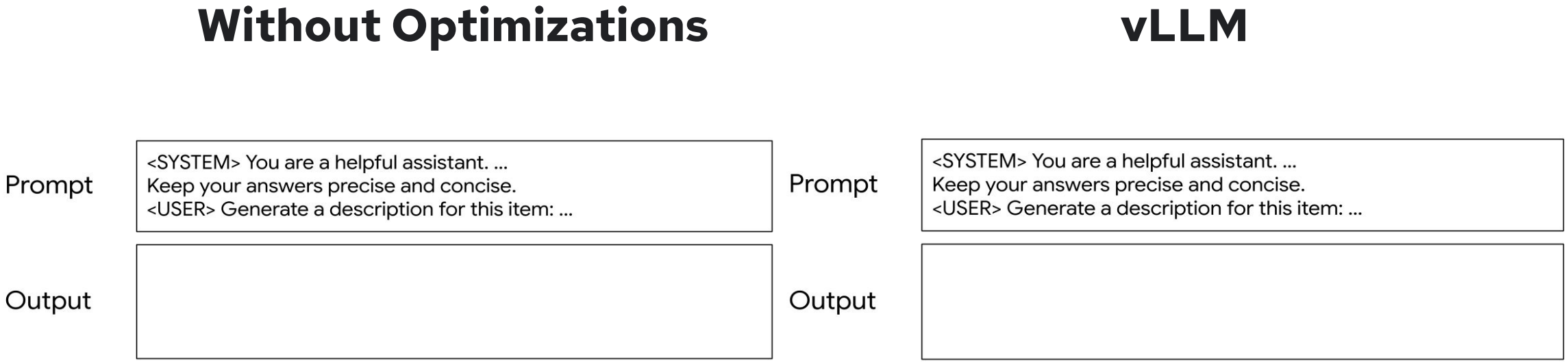


- **KV Cache**: Transformer 모델의 Self-Attention 과정에서 생성된 Key와 Value 벡터를 저장하여, 반복 계산을 줄이고 효율적으로 토큰을 생성할 수 있도록 돕는 메모리 캐시 기술
- **KV Cache 이슈**:
  기존 LLM 시스템들은 KV 캐시 저장 시 GPU 메모리의 60-80%를 단편화 및 불필요한 예약으로 낭비함
- Paged Attention은 메모리를 **페이지 단위로 관리해 필요한 부분만 GPU에 올려 효율적 메모리 사용을 구현**
- **개선효과:** GPU 메모리 단편화(fragmentation)와 과도한 예약(over-reservation)을 줄임
  - GPU 메모리 활용도 대폭 향상
  - **더 긴 컨텍스트(Context) length 지원 가능**
  - 비용 및 자원 낭비 감소

Red Hat

vLLM 성능 자료 (1/2)

Prefix caching, FP8 KV Cache, Spec decoding 등 최적화로 출력 속도와 비용 효율을 크게 개선

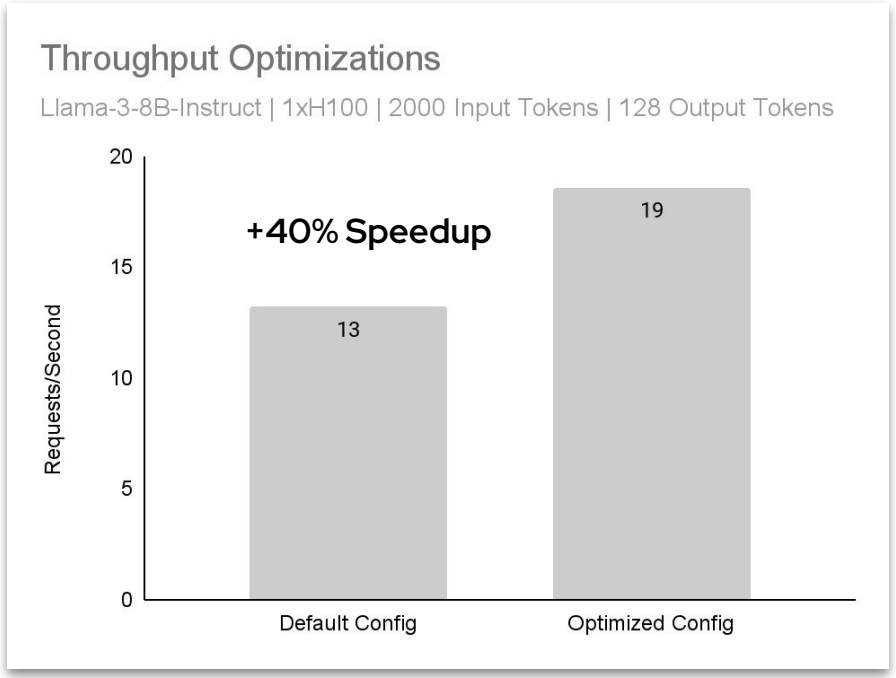


Red Hat

vLLM 성능 자료 (2/2)

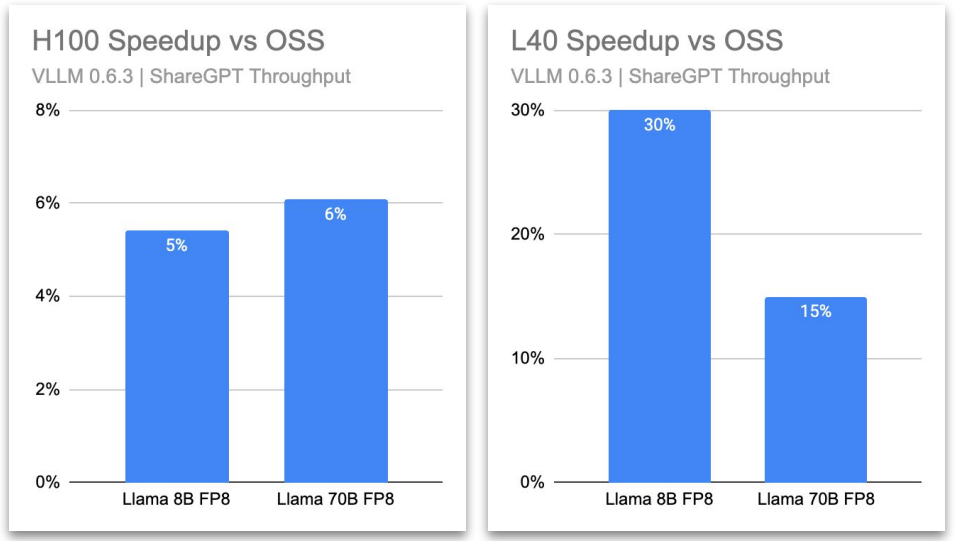
하드웨어·GPU·모델별 FP8 레이어 튜닝과 워크로드 최적화로 최대 40% 처리량 향상 달성

Offline Batch Use Case



워크로드 성능 튜닝을 통한 실질적 처리량 증가 효과가 큼

H100, L40 Speedup vs OSS



L40 GPU에서 경량 모델(Llama 8B)의 성능 향상이 두드러짐

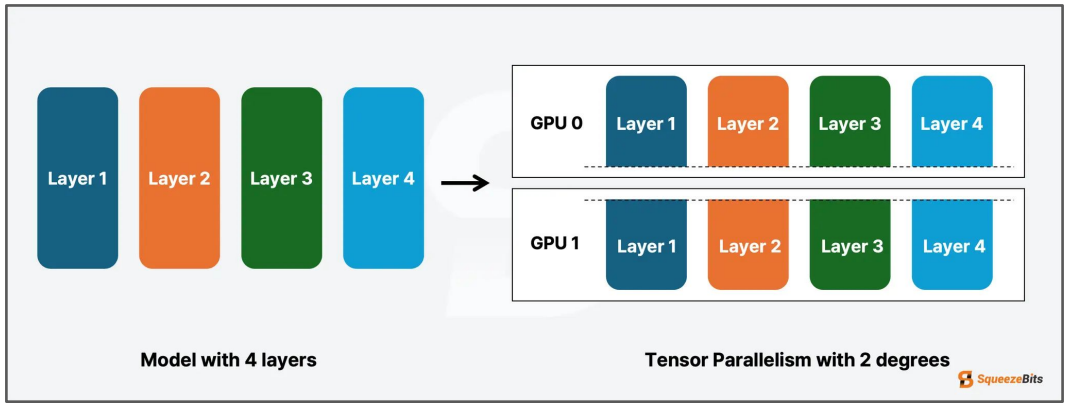
Red Hat

Model parallelism

대형 모델을 여러 GPU에 나눠 비용 효율적으로 대용량 연산을 가능

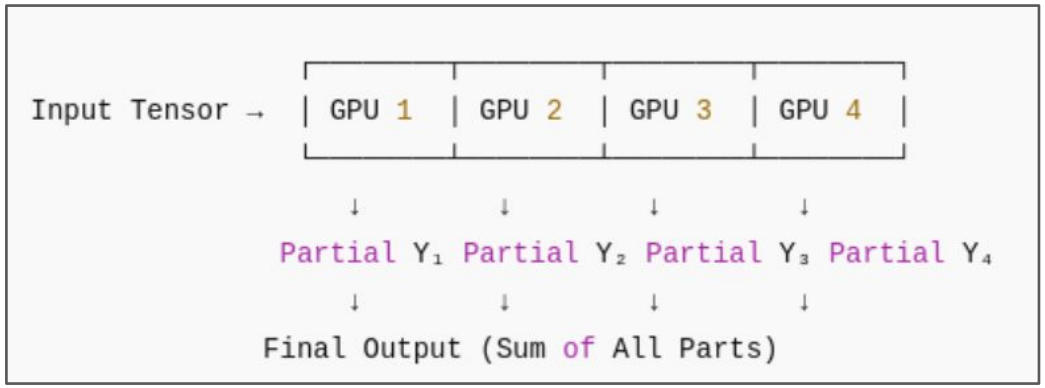
Model parallelism 이해

- 모델을 여러 GPU에 분산시켜 처리하는 방식
- 일반적인 방식은 모델의 레이어를 GPU에 분산하여 처리
- 이를 통해 대규모 모델을 효율적으로 배포할 수 있음
- 종류: Tensor, Pipeline, Expert, Data, Disaggregated Serving



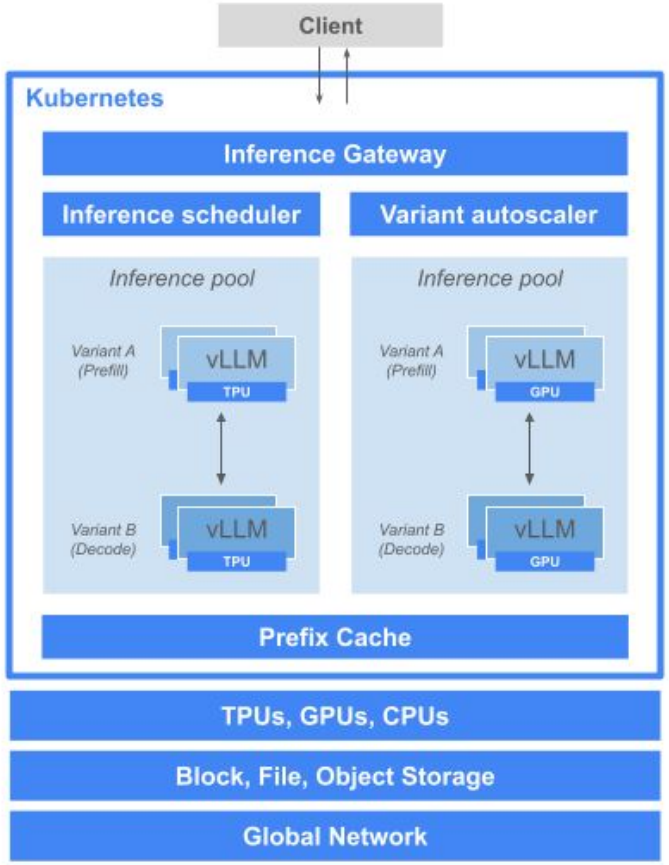
Tensor parallelism이 중요한 이유

- 하나의 GPU로 감당할 수 없는 대형 모델의 계산을 여러 GPU가 나눠서 동시에 수행함으로써, 메모리 한계를 넘고, 처리 속도를 높이며, 비용 효율성도 확보할 수 있게 해줌
- 모델 크기 한계 극복, 계산량 병렬처리, 비용절감



대규모 분산 추론을 위한 llm-d

Kubernetes 환경에서 고성능, 비용 효율적인 대규모 LLM 추론을 지원하는 솔루션



새로운 오픈소스 프로젝트인 llm-d

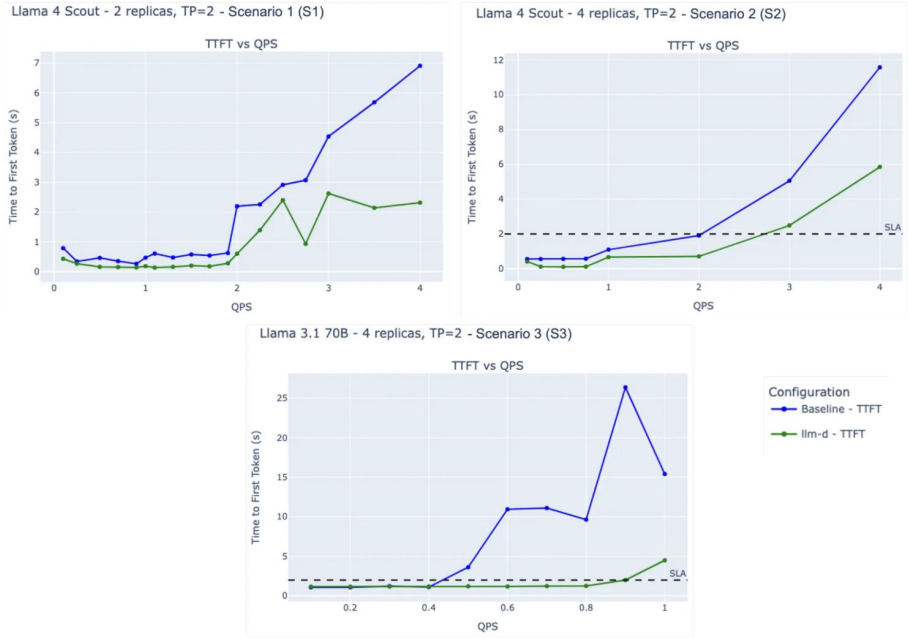
- Kubernetes 기반 분산 추론 환경 제공
- Reasoning 모델의 등장으로 추론 효율화 요구는 점점 증가
- 복수의 GPU 혹은 서버 간 KV Cache를 공유하거나 입력(Prefill)과 출력(Decode)을 분리
- 확장성 있는 추론(scalable inference)을 가능하게 하여 Latency와 Throughput을 크게 개선
- Google, AMD, NVIDIA, Hugging Face 등이 참여

AMD cisco CoreWeave Google Hugging Face
IBM intel Lambda Mistral AI NVIDIA

Source:
- Red Hat Launches the llm-d Community, Powering Distributed Gen AI Inference at Scale
- llm-d: Kubernetes-native distributed inferencing | Github - llm-d

대규모 분산 추론을 위한 llm-d

Kubernetes 환경에서 고성능, 비용 효율적인 대규모 LLM 추론을 지원하는 솔루션



핵심 관찰사항(Key Observations):

- **S1:** 4 QPS에서 llm-d는 기준 대비 약 3배 낮은 평균 TTFT(최초 토큰까지의 시간)를 달성함
- **S2:** llm-d는 SLO(서비스 수준 목표)를 만족하면서 기준 대비 **50% 더 높은 QPS**를 제공함
- **S3:** llm-d는 SLO 제약 조건 하에서 **기준 대비 2배의 QPS**를 지속적으로 유지함

- **IGW와 llm-d의 vLLM**이 함께 KV 캐시를 인식하는 **부하 분산(로드 밸런싱)**과 prefix cache를 인식하는 **라우팅 방식을 공동 개발**
- 이 기능을 테스트하기 위해 **LMbenchmark**라는 도구로 2대의 8 GPU(H100) 노드(총 16 GPU) 환경에서 입출력 워크로드를 기반으로 평가
- KV 캐시의 **재사용,** 라우팅(분배) **품질** 테스트

	Model	Configuration	ISL	OSL	Latency SLO
S1	LlaMA 4 Scout FP8	TP2, 2 replicas	20,000	100	None
S2	LlaMA 4 Scout FP8	TP2, 4 replicas	12,000	100	P95 TTFT <= 2s
S3	Llama 3.1 70B FP16	TP2, 4 replicas	8,000	100	P95 TTFT <= 2s

LLM Compressor

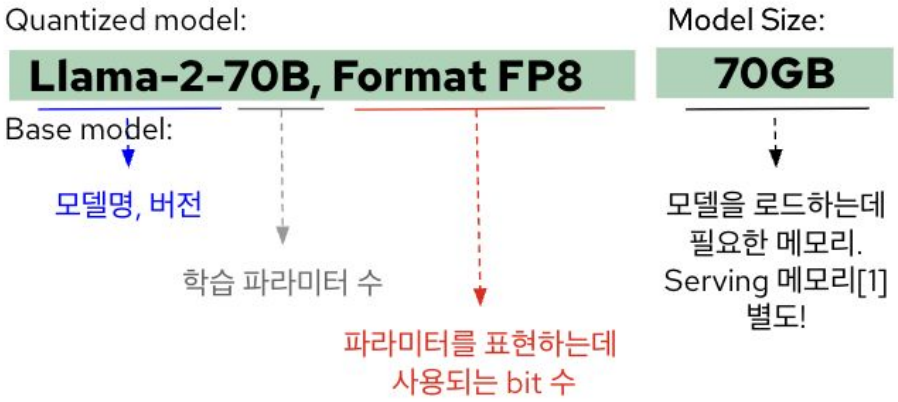
Quantization

LLM을 저정밀도로 변환해 연산과 메모리를 줄이고 비용과 속도를 개선

양자화(Quantization):

- 디지털 신호의 정밀도를 낮춰 연산 효율을 높이는 **양자화(Quantization)**는 AI/ML, 신호 처리, 압축 등에 활용
- LLM에서는 수치 연산을 간소화해 **메모리 절감**, **추론 속도 향상**, **전력 소모 감소** 효과
- 주요효과
  - **Weight 양자화**: 저장 공간 및 메모리 절감

    e.g. 100B 모델 – BFloat16(200GB) → FP8(100GB)
  - **Activation 양자화**: 연산 속도 향상
  - **KV Cache 양자화**: 캐시 용량 및 Attention 속도 최적화



Source:
- What is quantization?

Red Hat

LLM Compressor

LLM Compressor를 통한 양자화

추론 전 모델을 LLM Compressor를 통해 양자화 하여 모델 사이즈를 줄이고 속도와 정확도를 높여줌

	모델	포맷	모델 사이즈*	Serving 메모리**	속도	비용	정확도
Llama 3.1 압축 전	Llama 3.1 **8B** (Lean Llama)	FP16	16GB	~40GB	Fast	💸	**Good**
	Llama 3.1 **70B** (Bulky Llama)	FP16	140GB	~280-300GB	Slow	💸💸💸	SOTA
Llama 3.1 압축 후	**Llama 3.1 70B-FP8 (Optimized Llama)**	**FP8**	**70GB**	**~120-140GB**	**Fast**	💸💸	**SOTA-equivalent**

**Model size* = *just model weights*

Serving memory* = *model weights + KV cache + activations + runtime (vLLM) overhead

Source:
- What is quantization?

Red Hat

Hugging Face

최적화된 모델 제공을 위한 Red Hat AI Repository

vLLM으로 사전 최적화된 모델 3rd party 모델 컬렉션을 제공 (on Hugging Face)

다양한 컬렉션

Llama
Qwen
Gemma
Mistral
DeepSeek
Phi
Molmo
Granite
Nemotron

Validated models

- 현실성 있는 시나리오로 검증
- 다양한 하드웨어에서 성능 평가
- GuideLLM 벤치마크와 LM Eval Harness를 사용하여 수행

Optimized models

- LLM을 압축하여 속도, 효율성 달성
- 더 빠르고, 리소스는 적게 사용하면서 정확도는 유지
- 최신 알고리즘이 적용된 LLM Compressor를 사용하여 수행

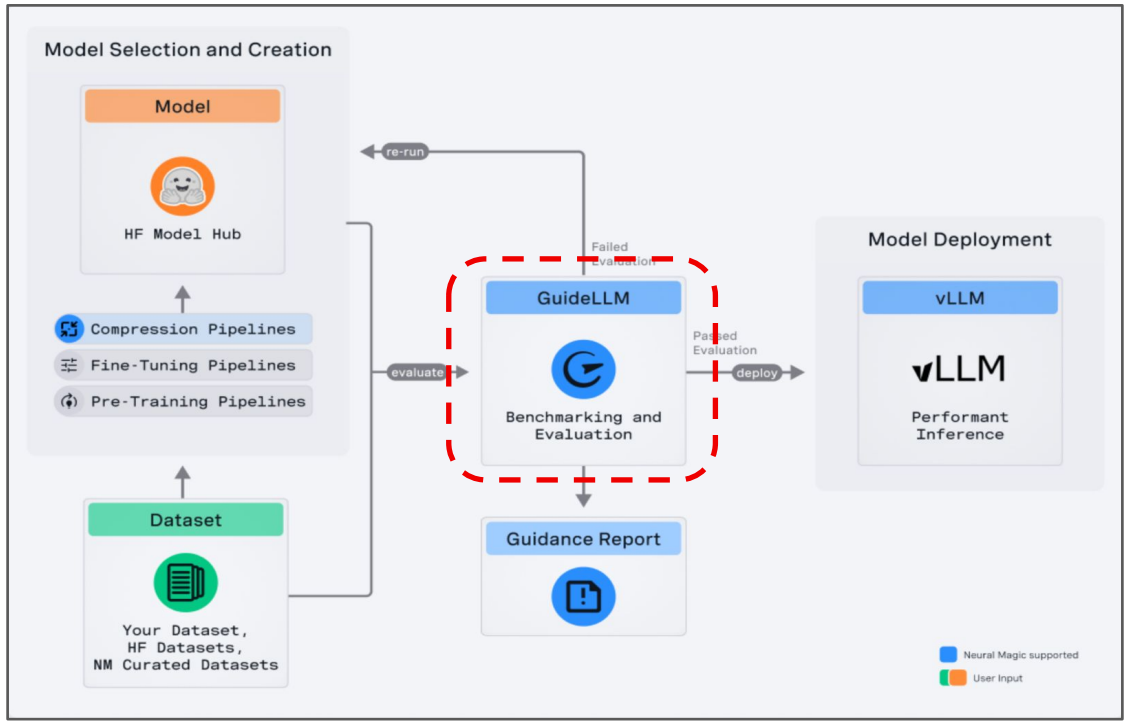
vLLM으로 최적화한 모델을 Hugging Face에 공개

Red Hat

GuideLLM

LLM 성능 측정을 위한 GuideLLM

다양한 LLM 성능을 평가하고 비교할 수 있는 오픈소스 지표 기반 평가 프레임워크

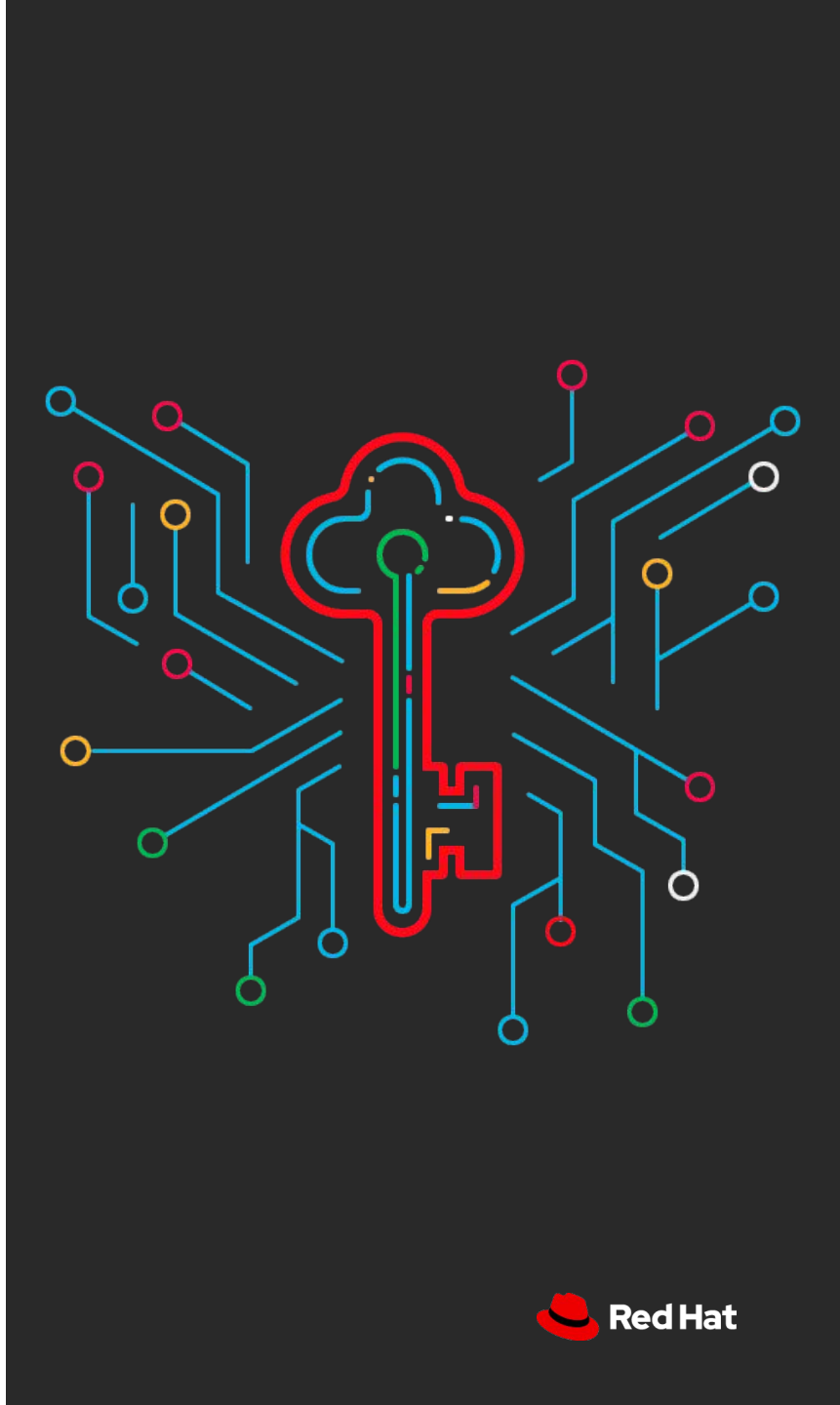


지표 종류	내용 요약	상황별 중요도
1. 추론 성능 지표	TTFT, TPOT, TPS 등 모델 응답 속도 관련 지표	대화형 챗봇, **실시간 응답**이 중요한 경우 최우선
2. 품질 및 정확도 지표	답변의 사실성, 논리성, 허위정보 생성 여부, 작업 성공률 등	고객 요구가 '**정확한 답변**' 혹은 '비즈니스 맞춤형'일 때 최우선
3. 시스템/인프라 지표	메모리 사용량, 동시 처리량, 실패율 등 인프라 효율성 관련 지표	**인프라 비용과 확장성, 안정성**이 중요할 때 주로 집중
4. 비용 지표	토큰당 비용, 요청당 비용 등 운영 비용	**비용 절감이** 고객의 핵심 요구일 때 중요

- **GuideLLM**은 대규모 언어 모델(LLM)의 배포를 평가하고 최적화하는 강력한 도구입니다.
- **GuideLLM**은 실제 추론 워크로드를 시뮬레이션하여 사용자가 다양한 하드웨어 구성에 **LLM을 배포할 때의 성능, 리소스 요구 사항 및 비용 영향을 측정**할 수 있도록 지원합니다.

Red Hat

요약 및 결론



LLM 운영, Red Hat으로 완성하기

Red Hat Inference Server를 통한 모델 추론 최적화

1. **효과적인 LLM 추론** — vLLM 기반 추론 최적화로 GPU 및 메모리 활용 극대화
2. **추론 속도 향상** — Model parallelism, llm-d 통한 대규모 분산 추론
3. **모델 압축** — llm-compressor를 통한 모델 양자화와 경량화
4. **추론 지원 도구** — 최적화된 모델, GuideLLM, 주요 하드웨어 가속기 지원

Red Hat AI Inference Server

모델 및 추론 최적화로 비용 효율적인 추론 가능

Red Hat Enterprise Linux AI

Red Hat OpenShift AI

Red Hat

Red Hat AI Inference Server

일관되고 빠르며 비용 효율적인 대규모 추론을 제공

Inference runtime for the hybrid cloud
선택한 모델을 어떠한 액셀러레이터나 어떠한 환경에서든 실행할 수 있습니다.

Red Hat AI Hugging Face repository
추론에 필요한 타사 검증 또는 최적화된 모델 컬렉션에 액세스 가능합니다.

Compress Models
정확성을 유지하면서 연산량과 비용을 절감합니다.

Certified for all Red Hat products
Red Hat이 아닌 리눅스 및 쿠버네티스 플랫폼에서도 배포가 가능합니다.

Select an optimal LLM
Red Hat AI의 서드파티에서 검증하거나 압축한 모델들은 바로 사용 가능합니다.

An efficient inference runtime
vLLM

Broad support for hardware
AMD NVIDIA IBM Google intel aws

Red Hat AI Inference Server는 OpenShift AI와 RHEL AI에 포함되어 있습니다.

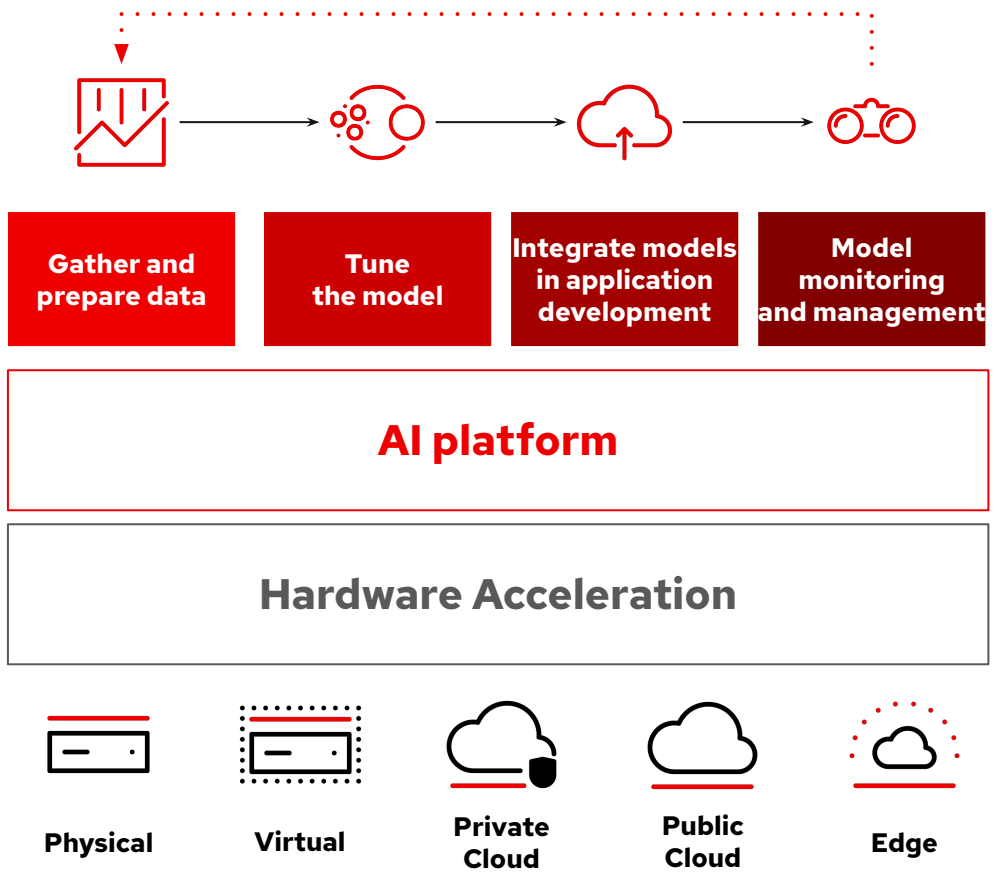
Red Hat

Red Hat는 오픈 소스 AI 여정의 동반자입니다.

Red Hat AI는 하이브리드 클라우드 전반에서 AI 모델, AI 기반 애플리케이션, 그리고 AI 에이전트를 대규모로 일관되게 구축·배포·운영할 수 있는 AI 플랫폼을 제공합니다.

It provides:

- 효율적인 추론 런타임 (vLLM)
- 검증 및 최적화된 서드파티 모델
- 맞춤화를 위한 InstructLab 및 RAG
- MLOps 및 LLMOps 기능
- 모니터링, bias 감지 및 가드레일



Red Hat

감사합니다.

Red Hat은 세계 최고의 엔터프라이즈 오픈 소스 소프트웨어 솔루션 제공업체입니다. 수상 경력에 빛나는 지원, 교육 및 컨설팅 서비스를 통해 Red Hat은 Fortune 500대 기업의 신뢰받는 자문 기업으로 평가받고 있습니다.

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

twitter.com/RedHat

Red Hat