

다양하고 복잡한 클라우드 환경에서 최적 인공지능 개발 환경 구성

AI 인프라 혁신센터 워크숍

한양대학교
데이터사이언스학부
분산데이터처리시스템연구실
이경용



SW STOR LAB
Software Technology Advanced Research

분산 데이터 처리 시스템 연구실



과학기술정보통신부
Ministry of Science and ICT

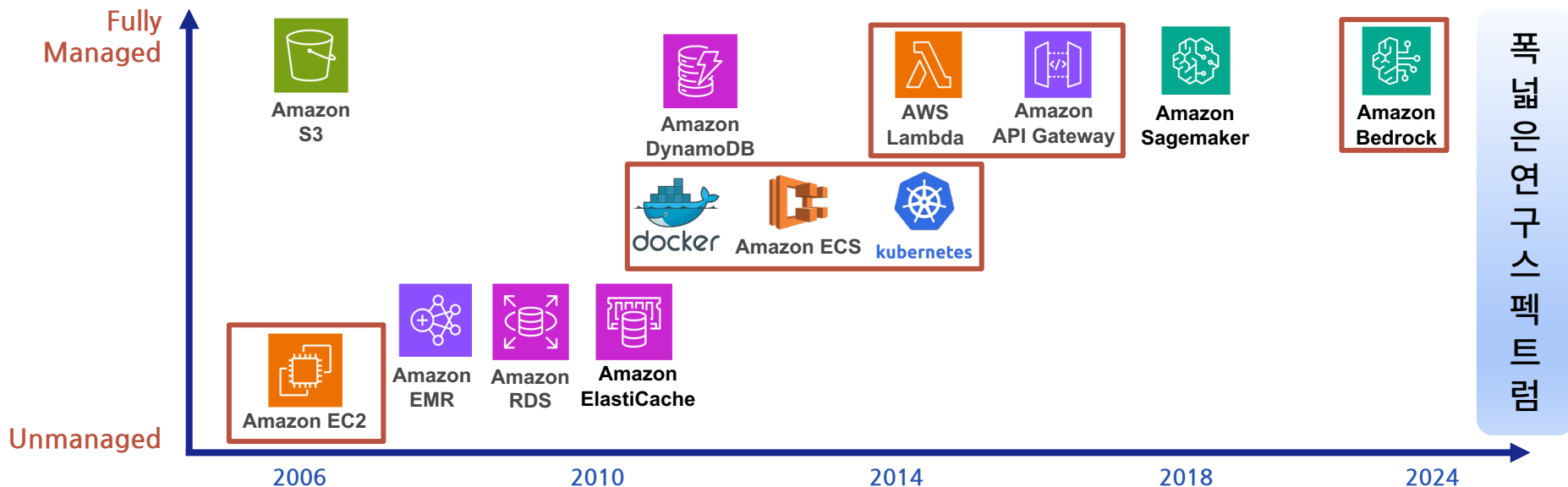


한양대학교



정보통신기획평가원
IITP

클라우드 컴퓨팅 서비스의 발전 및 다양한 연구



포스트 클라우드 컴퓨팅 연구 시스템

1 세대 클라우드 시스템 : Infrastructure-as-a-Service (IaaS) 기반 서비스 들 – Compute, Storage, ...

2세대 클라우드 시스템 : 컨테이너 기반 서비스 배포 (Docker & Kubernetes)

3세대 클라우드 시스템 : Function-as-a-Service (FaaS) 및 완전 관리형 서비스를 활용한 서버리스 컴퓨팅

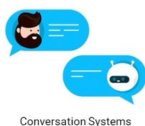
4세대 클라우드 시스템 : 생성형AI를 활용한 응용 및 서비스 개발을 위한 생태계

딥러닝 알고리즘을 활용한 인공지능 응용 서비스 보편화

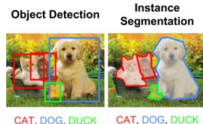
기반 기술 발전에 따른 다양한 분야에서 Neural Net 알고리즘 활용

- 병렬 작업 특화 하드웨어의 발전 (GPU)
- 모델 경량화 및 알고리즘 최적화 방안
- 사용하기 쉬운 개발 플랫폼 및 추상화 된 API

Deep NeuralNet



Conversation Systems

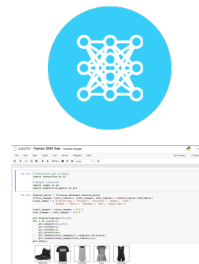


CAT, DOG, DUCK CAT, DOG, DUCK

딥러닝 알고리즘 개발 및 활용

학습 (Train)

- 입력 데이터의 출력값을 모델링 하는 구조 개발
- 반복적 수행을 통해 에러값을 최소화 하는 가중치 결정
- 입력 데이터 및 구조에 따라 오랜 시간 소요 (수 주/달)
- 높은 처리 능력이 중요 (throughput)



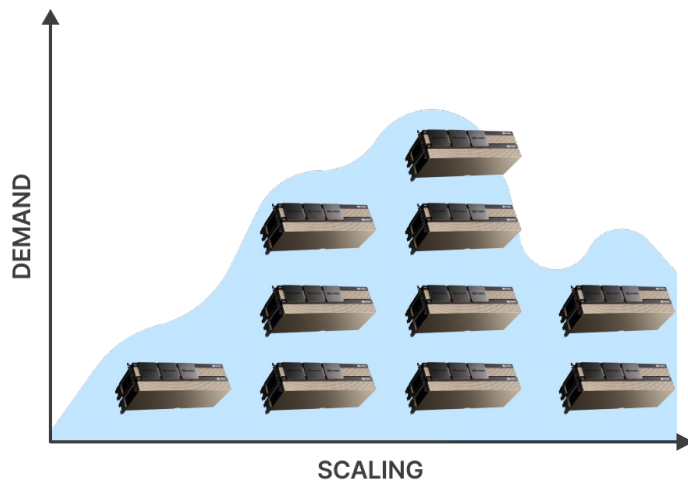
추론 (Inference)

- 학습이 완료된 모델을 활용한 예측 서비스를 제공
- 요청량에 따른 탄력적인 서비스 운용 체계 구축이 중요
- 입력 데이터의 빠른 처리 시간이 중요 (latency)



인공지능 서비스 개발 환경으로 클라우드 컴퓨팅

클라우드 환경에서 Auto-Scaling



필요에 따라 자유롭게 증감 가능

컴퓨팅 자원의 확장성 및 탄력성

딥러닝 학습

간헐적인 학습 진행

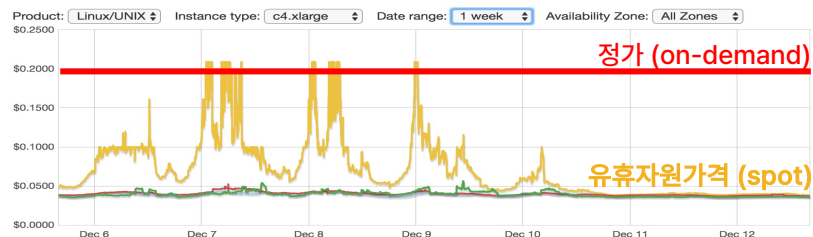
딥러닝 추론

가변적인 사용자의 요청 처리

유휴 클라우드 자원을 활용한 비용 절감

- 남는 자원을 정가 자원 (on-demand) 대비 **최대 90% 저렴한 가격에** 이용 가능
- 남는 자원량에 따라 **강제 종료** 가능

Spot Instance Pricing History

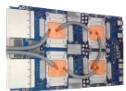




클라우드에서 최적의 딥러닝 개발 환경 구축의 어려움

학습을 위한 하드웨어 선택

Google TPU



AWS Tranium



AMD GPU



NVIDIA GPU



학습 환경 구축법

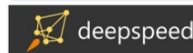
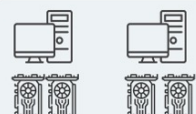
Single



Parallel



Distributed training



추론을 위한 하드웨어 선택

Intel



ARM

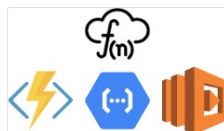


CPU

AWS Inferentia



Serverless Computing
& FaaS



추론모델 최적화 컴파일러

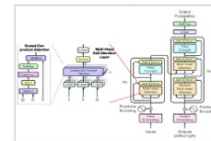


Compiler

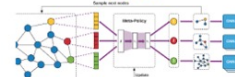


모델의 다양성

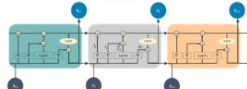
LLM



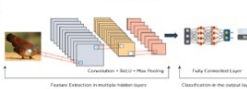
GNN



RNN



CNN



하이퍼파라미터 설정

Quantization

FP32



BF16



FP16



INT8



INT16



batch size



클라우드에서 제공되는 다양한 과금 정책

클라우드 서비스의 과금 정책

On-Demand

- 사용한 만큼 비용 지불
- 확장성 및 안정성 제공



Spot Instance

- 여유 자원을 싼 값에 제공 (~90% 할인)
- 인스턴스의 수요-공급 상황에 따라 언제든지 강제 종료 가능



AWS

Spot Instance



Azure

Spot Virtual Machines

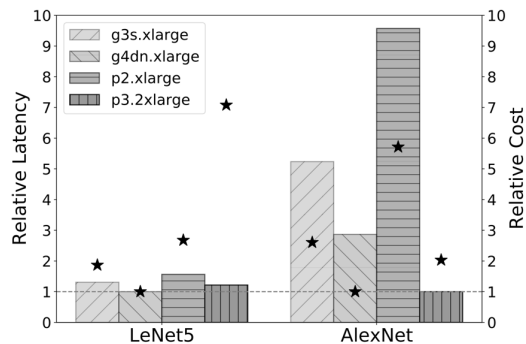


Google Cloud

Spot Virtual Machines

클라우드 자원을 활용한 인공지능 서비스 개발 시 성능 다변성

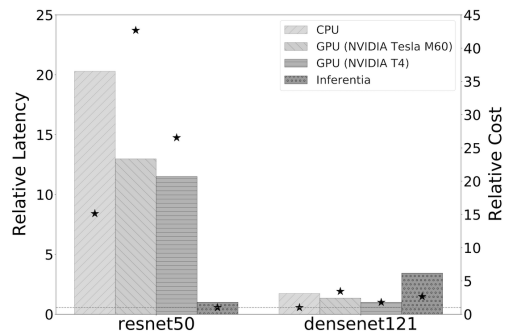
학습



딥러닝 학습 시 GPU에 따른 성능 변화

NVIDIA T4(G4DN)와 V100(P3)가 DNN 모델에 따라 처리속도, 비용에서 큰 차이를 보임

추론



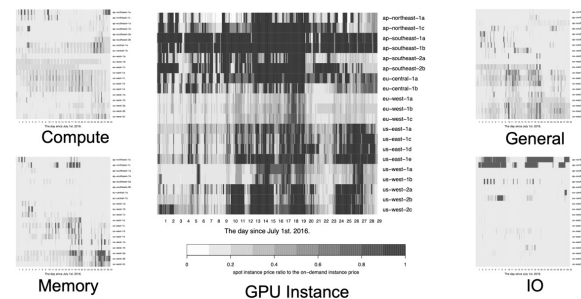
모델에 따라 최적의 추론 환경을
제공해주는 하드웨어가 다름

추론 전용 장치 AWS Inferentia

- resnet50 에서 최고 성능
- densenet121 에서 최저 성능

다양한 GPU 간 비슷한 처리 속도
에도 불구하고 큰 비용 차이

비용

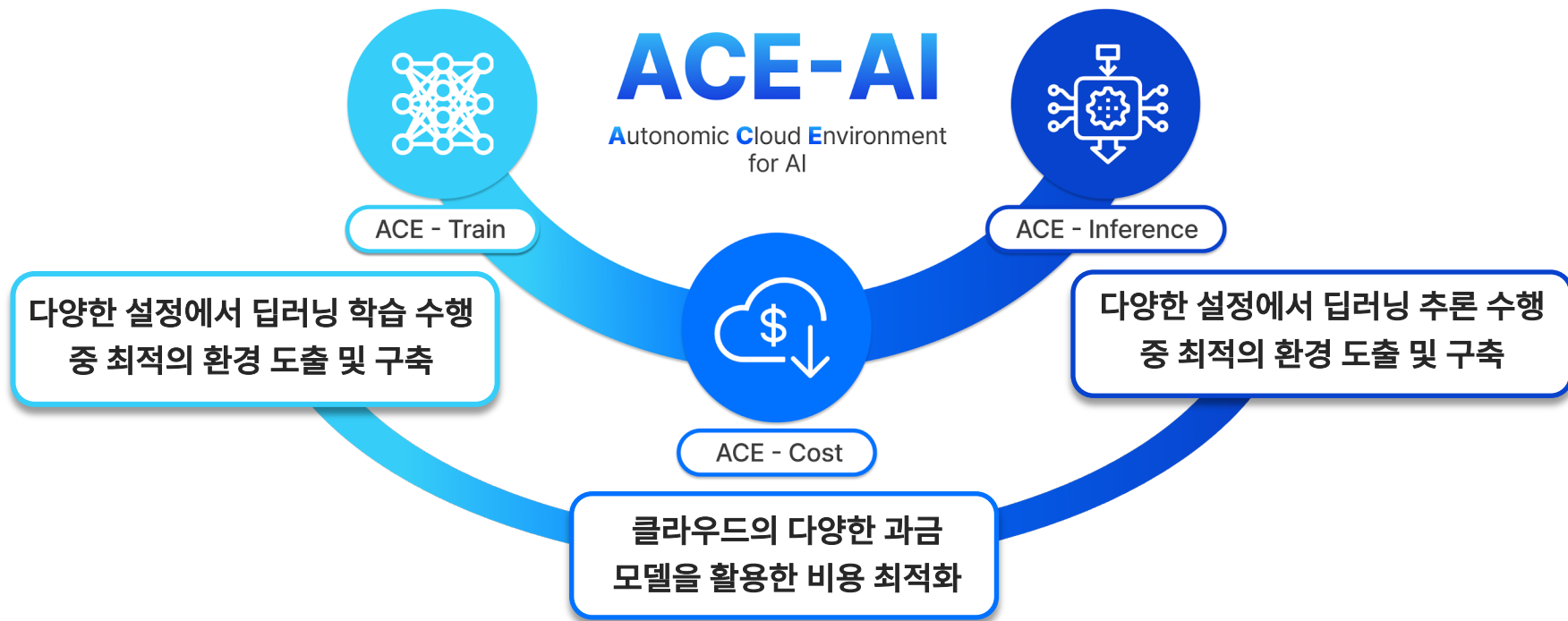


유휴자원을 활용한 비용 절감 가능

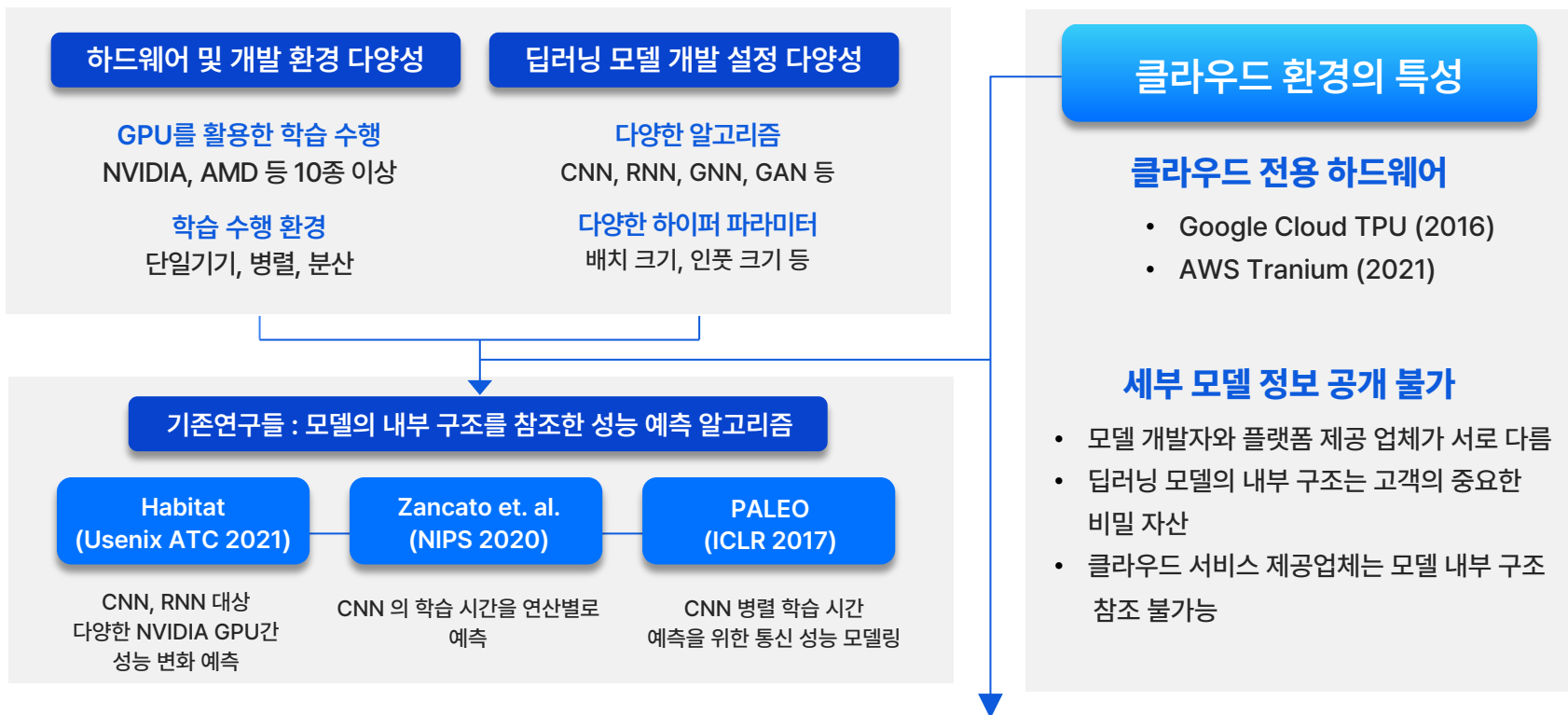
	요청실패율	강제 종료율
안정적	0%	12%
보통	26%	38%
불안정	48%	42%

하드웨어별 비용 절감 비율, 안정성에
매우 큰 편차를 보임

클라우드에서 비용 및 성능 관점 최적의 인공지능 개발 환경 구축



클라우드 환경에서 딥러닝 학습 설정의 다양성



ACE-Train : 모델의 내부 구조를 참조하지 않는 블랙박스 기반 딥러닝 학습 시간 예측 알고리즘 개발

블랙박스 기반 학습 소요 시간 예측 알고리즘

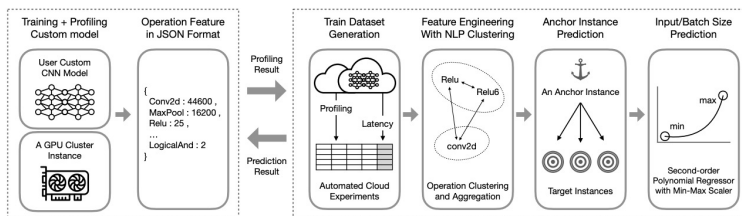
프로파일러 결과 활용

딥러닝 개발 플랫폼에서 제공되는
프로파일러 결과를 활용한 학습 시간 예측 모델링

Operation	Operation Details				Aggregated Time (ms)
	Layer Name	Output Tensor	MEM Request	Time (ms)	
Conv2D	conv2d_0	[32, 96, 224, 224]	616562 KB	50	286
	conv2d_1	[32, 256, 112, 112]	411041 KB	45	
	conv2d_2	[32, 512, 58, 58]	220463 KB	39	
	...	...	...	...	
ReLU	activation_0	-	-	11	26
	activation_1	-	-	7	
	activation_2	-	-	3	
	...	...	...	...	

새롭게 발견되는 연산의 처리 기능 확보

연산 이름 클러스터링 기법 개발

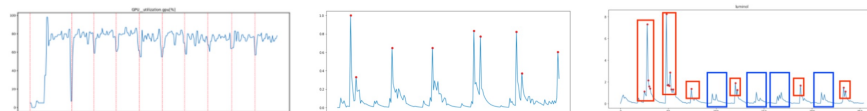


하드웨어 메트릭 활용

프로파일러를 통한 모델 사용 연산 유출 가능성 존재

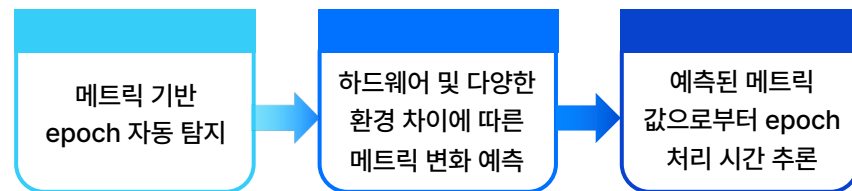
학습 중 발생하는 하드웨어 메트릭 활용

- 높은 수준으로 추상화 된 정보 제공
- GPU 활용율, CPU 활용율, 메모리 등 참조 가능
- 하드웨어 스펙 활용



AWS Compute Optimizer 에서 유사 정보 활용

하드웨어 메트릭 기반 학습 시간 예측에 필요한 단계



SpotLake : 클라우드 스팟 인스턴스의 안정적 활용을 위한 데이터 서비스

여러 벤더의 다양한 스팟 인스턴스 데이터 셋을 제공

- AWS, Azure, GCP 데이터 수집
- Price, Interrupt Ratio, Instant Availability

웹 콘솔을 통해서 현재 데이터만 제공되는 Azure ▶

▼ CLI 를 통한 과거 데이터 까지 제공해주는 AWS

```
kyungyong-IM-3:~$ kyungyong@aws ec2 describe-spot-price-history --instance-types c5.xlarge --start-time 2024-02-06T07:08:09 --end-time 2024-04-03T08:09:10
{
  "SpotPriceHistory": [
    {
      "AvailabilityZone": "us-west-2b",
      "InstanceType": "c5.xlarge",
      "ProductDescription": "Red Hat Enterprise Linux",
      "SpotPrice": "0.129600",
      "Timestamp": "2024-04-03T06:01:43+00:00"
    }
  ]
}
```

Find the Spot VM that is right for you

Enter a few parameters and find out which Spot VM has the lowest price, or costs less than the maximum price you are willing to pay. After you specify the requirements that your Spot VM should have, you can sort the results by price. Begin by applying filters to customize pricing options to your needs.

OS/Software: CentOS or U... Region: West US... Currency: United States... Display pricing: Hour... VM sizes: All...
 Min. vCPUs/Cores: 0... Min. RAM (GB): 0... Max. price: 0

Instance	vCPUs/Cores	RAM	Temporary storage	Pay as you go	Spot *
B3ts v2	2	1 GB	N/A	\$7.30/month	\$1.82/month 75% savings
B3ats v2	2	1 GB	N/A	\$8.18/month	\$2.04/month 75% savings
B3ts v2	2	1 GB	N/A	\$9.05/month	\$2.26/month 75% savings
A1 v2	1	2 GB	10 GB	\$31.39/month	\$3.14/month 90% savings

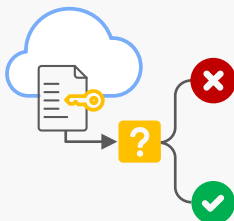
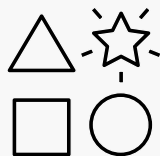
주요 해결 문제

각기 다른 수집 인터페이스

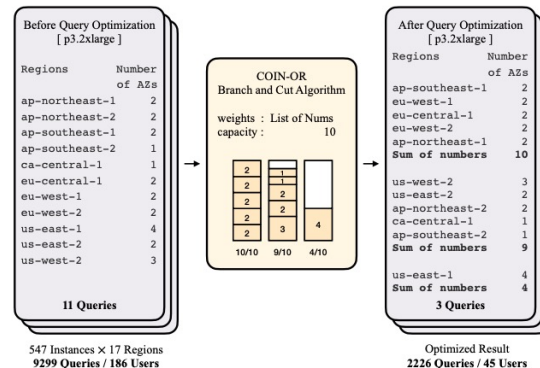
과거 데이터 제공 여부 상이

AWS SPS 수집 시 다양한 어려움 해결

웹 콘솔, SDK, CLI 등



- 계정 당 쿼리 수 제한
- 쿼리당 리턴 되는 응답 개수 다변화
- Bin-Packing with mixed integer programming solver
- 쿼리 부담 70% 이상 감소 시킴



SpotLake : 데이터 서비스 구현 및 연구 커뮤니티에 공유

SpotLake About Document Demo Contact

Amazon Web Services Google Cloud Platform Microsoft Azure

Instance: Region: AZ: Start date: 2024. 04. 05 End date: 2024. 04. 06 QUERY

FILTERS COLUMNS DENSITY EXPORT VISUALIZE

☐	InstanceType	Region	AZ	Availability	Interruption Ratio	SpotPrice (\$)	Savings (%)	Date
☐	c7gd.8xlar...	ap-south-1	ap...	3	3	0.292	68	2024-04-06 01:30:00
☐	c7gd.8xlar...	ap-south-1	ap...	3	3	0.375	59	2024-04-06 01:30:00
☐	c7gd.8xlar...	ap-south-1	ap...	3	3	0.296	68	2024-04-06 01:30:00
☐	r7i.4xlarge	ap-south-1	ap...	3	3	0.395	64	2024-04-06 01:30:00
☐	r7i.4xlarge	ap-south-1	ap...	3	3	0.549	50	2024-04-06 01:30:00
☐	r7i.4xlarge	ap-south-1	ap...	3	3	0.392	64	2024-04-06 01:30:00
☐	m6a.24xla...	ap-south-1	ap...	1	1.5	0.972	64	2024-04-06 01:30:00
☐	m6a.24xla...	ap-south-1	ap...	3	1.5	0.89	67	2024-04-06 01:30:00
☐	m6a.24xla...	ap-south-1	ap...	3	1.5	0.924	65	2024-04-06 01:30:00
☐	r5.16xlarge	ap-south-1	ap...	3	3	1.303	69	2024-04-06 01:30:00
☐	r5.16xlarge	ap-south-1	ap...	3	3	1.239	70	2024-04-06 01:30:00

<https://spotlake.ddps.cloud/>

SpotLake Public Web Service

웹 기반 데이터 쿼리
인터페이스 개발

서버리스 컴퓨팅
아키텍처 도입

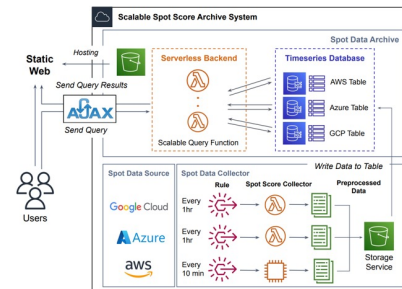
전체 데이터셋은
별도 요청 시 제공

유지보수 부담 획기적 감소
Lambda, Timeseries DB, S3 등 활용

Müller, Dominik; Dominko, Markus; ... - Wed, Jan 10, 2024 10:44 PM
Hello Prof. Kyungmin Lee,
Thank you very much for your quick reply.
I was able to successfully download the data from the links you provided.
I appreciate your team's work on the SpotLake project and am very
grateful for your decision to share the collected data openly.
It will be of great help with my research project.
Thank you again and all the best.
...
Dimitris Mueller
Department of Computer Science
Faculty of Engineering and Computer Science
Hamburg University of Applied Sciences

Sharma, Prateek; ... - Sat, Dec 3, 2022 12:44 AM
Hello Prof. Lee,
Thank you so much for providing the BigData data. I really appreciate
the data collection and your commitment to making
the data accessible. I have downloaded the data.
Hope to meet and discuss and cloud research with you some day.
Thanks again
-Prateek

Thank you note



서버리스 컴퓨팅 구조를 도입하여 시스템 구축

SpotLake Full Dataset Request

Send Query Results

Name *

Your answer

Organization *

Your answer

Purpose of use *

Your answer

데이터 요청 페이지

스팟 인스턴스의 인터럽트 비율 (안정성) 실험 결과

Kaplan-Meier Estimator

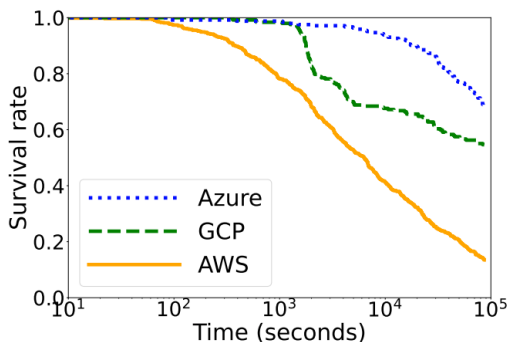
생존율 분석에서 사용되는 통계 기법으로
Non-parametric 기반 생존 함수 추정

$$\hat{S}(t) = \prod_{i: t_i \leq t} \left(\frac{n_i - d_i}{n_i} \right)$$

$S(t)$: t 이상 생존 확률

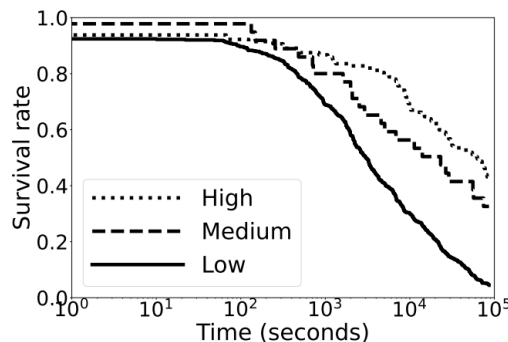
n_i : t_i 까지 생존한 갯수

d_i : t_i 까지 죽은 갯수

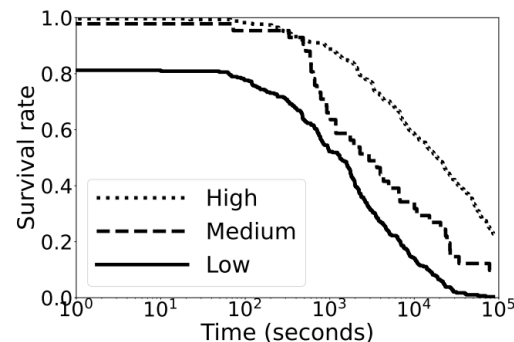


(a) Vendor

스팟 인스턴스 생존 확률
Azure > GCP > AWS



(b) Interrupt-free Score



(c) Instant Availability (SPS) Score

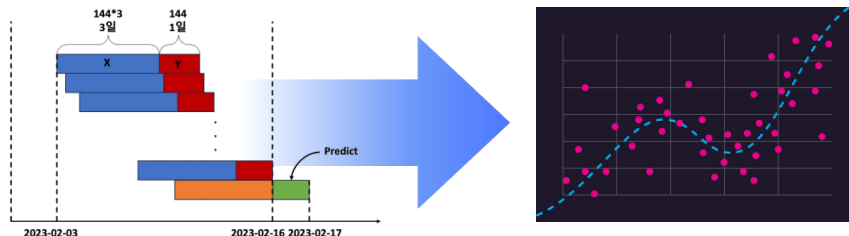
Instant Availability 데이터 셋

: Interrupt ratio, price 보다
인터럽트 이벤트 잘 표현

Instant Availability 데이터 셋의 예측이 안정성 향상에 도움이 될까?

주기적인 특성을 가지는 SPS 데이터 셋

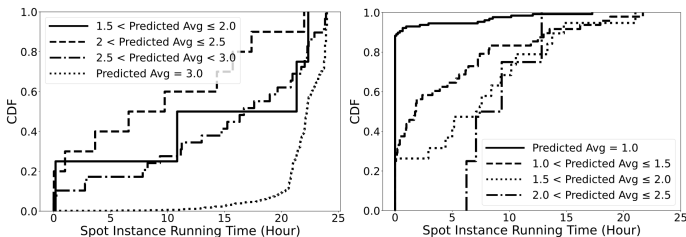
- 간단한 regressor 모델로 주기성 모델링
(XGBoost, Random Forest, ARIMA 등)



?

예측된 SPS 값은 스팟 인스턴스의
안정성 모델링에 도움을 줄까?

SPS의 예측 유무에 따른 인터럽트 비율 계산



초기 SPS = 3

추후 SPS 값
감소 예상

✗ 낮은 가용성

추후 SPS 값
유지 예상

✓ 높은 가용성

초기 SPS = 1

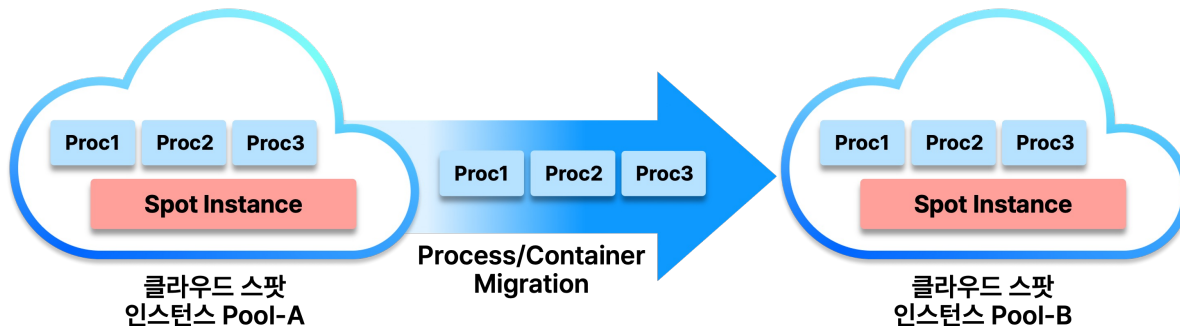
추후 SPS 값
증가 예상

✓ 높은 가용성

추후 SPS 값
유지 예상

✗ 낮은 가용성

클라우드 환경에서 런타임 실시간 작업 이관



응용 런타임의 실시간 이관 활용 사례

주피터 노트북 서버의 실시간 이관

Cunha. eScience 2021, Juric. arXiv 2021

응용 서비스 효율성 향상

Nadowda. ICDCS 2017, Ma. SEC 2017

자연재해대응

Kang. UCC 2014

마이그레이션 기법

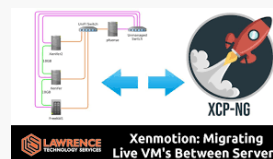
가상머신

Virtual Machine Monitor (VMM) 지원

컨테이너, 프로세스

CRIU

(Checkpoint & Restore in Userspace)



워크로드 특성을 고려한 이관 대상 클라우드 인스턴스 탐지

워크로드가 특정 CPU 의 모든 기능을 사용하는 것은 아님

AVX512 활용하여 성능 향상 기대



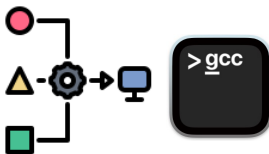
```
#include <x86intrin.h>
void foo(double s, double *b, double *c, int n)
{
    __m512d sv = _mm512_set1_pd(s); //
    for(int i=0; i<n; i++) {
        __m512d cv = _mm512_loadu_pd(&c[8*i]); //
        __m512d bv = _mm512_loadu_pd(&b[8*i]); //
        __m512d t1 = _mm512_mul_pd(bv, sv); //
        cv = _mm512_add_pd(t1, cv); //
        _mm512_storeu_pd(&c[8*i], cv); //
    }
}
```

ML 모델 학습 시 행렬 곱셈

AVX512 기능 미활용



DB Transaction 처리



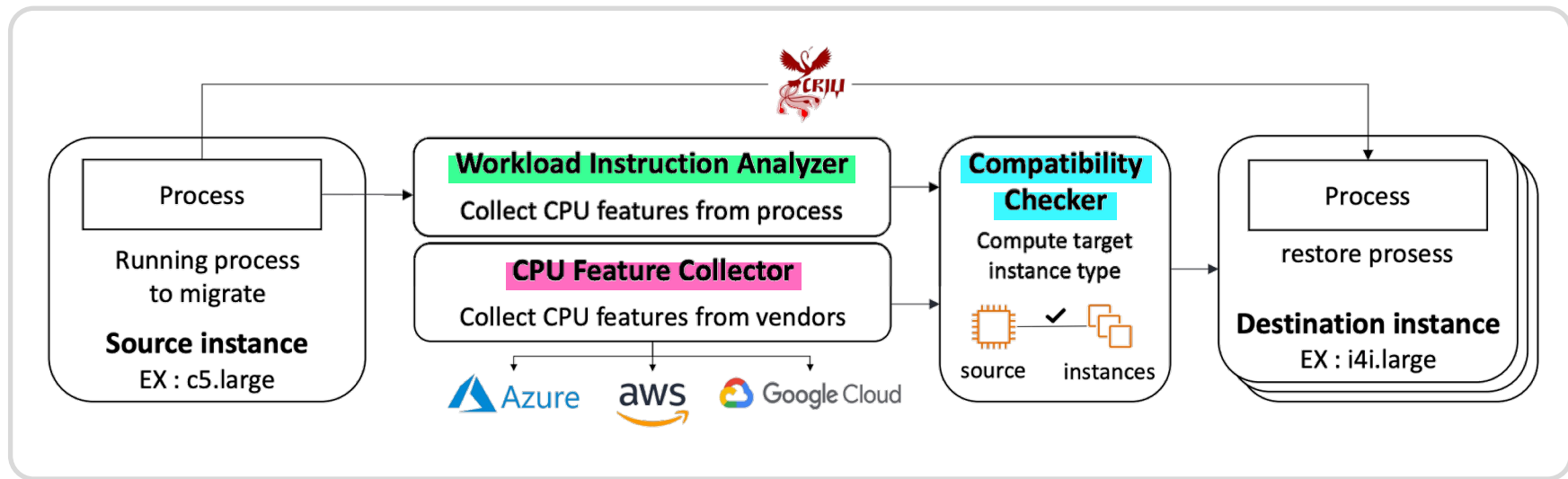
소스코드 컴파일



비디오 스트리밍

모든 워크로드가 CPU 가 지원하는 모든 기능을 사용하지 않으며,
사용하지 않는 기능은 호환성 검사 시 무시 가능

워크로드 특성을 고려한 호환성 여부 검사 시스템 구조도



CPU Feature Collector

- 클라우드 인스턴스들의 CPU 기능 수집 모듈

Workload Instruction Analyzer

- 워크로드에서 활용하는 CPU 기능 탐지 모듈
- Text Segment Full Scan
- Execution Path Tracking

Compatibility Checker

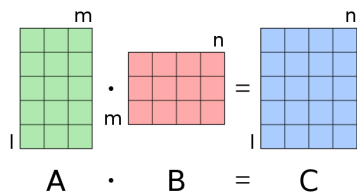
- 워크로드에서 사용하는 CPU 기능을 고려한 이관 가능 인스턴스 타입 탐지 모듈

이관 가능 인스턴스 탐지 Recall 성능 향상의 이점

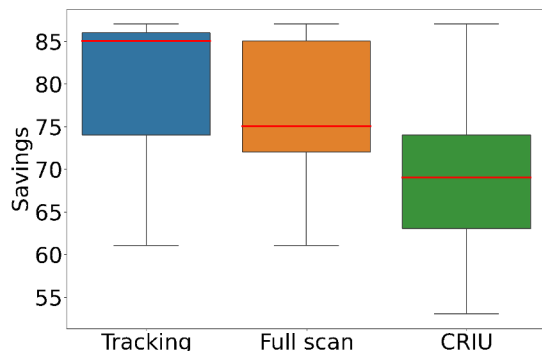
스팟 인스턴스는 타입별로 각기 다른
비용 절감율 및 안정성 특성을 가짐

인터럽트 발생 시 실시간 이관을 수행할 경우
대상 인스턴스 증가 효과

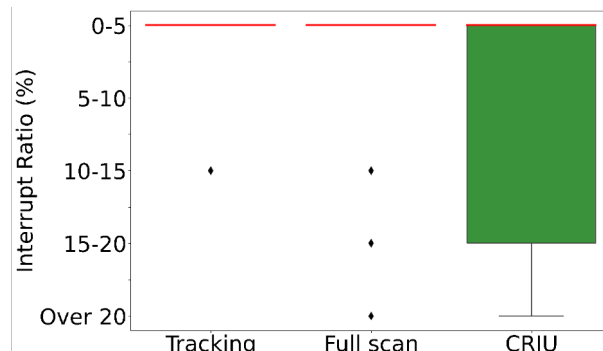
예제) AWS c6a.24xlarge
행렬 곱셈 수행



CRIU - 7 인스턴스 종류
제안 방법 - 449 종류



스팟 인스턴스 활용 비용 절감율 개선



스팟 인스턴스 안정성 개선

450 종류의 EC2 인스턴스에 7개 워크로드 수행 시

- Execution Path Tracking 적용 시 16%의 추가 비용 절감

- 인터럽트 관련 안정성 역시 향상 기대
- 데이터 시뮬레이션에는 SpotLake 의 Savings 및 Interrupt 데이터 셋 활용

Callisto : 클라우드 유휴 자원을 활용한 인공지능 서비스 개발 노트북

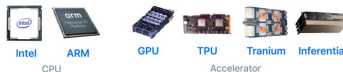
기존 연구 결과물을 활용 하여 서비스형 데이터/인공지능 서비스 개발 플랫폼

실시간
작업 이관 모듈



컨테이너 이관

다양한 하드웨어
추천 모듈



비용 최적화 모듈

SpotLake

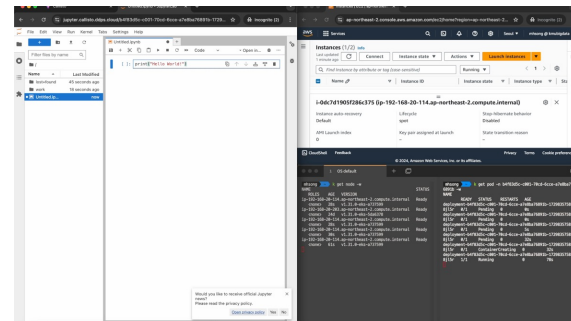
비용 및 성능 관점에서 최적의 환경 도출



kubernetes



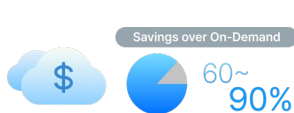
Karpenter



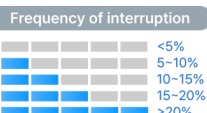
연구실의 과거 연구 결과물

서비스형 소프트웨어로 개발 하기 위한 플랫폼 및 응용 기술

Callisto : 주피터 노트북 기반 데이터/인공지능 서비스 개발 플랫폼



비용 효율성



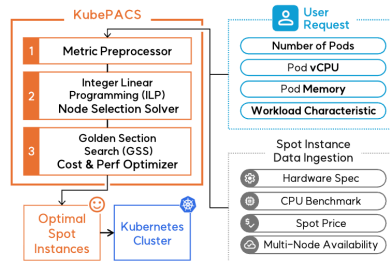
스팟 인스턴스 안정성

CoreMark-Pro	Platform Detailed Data									
	Baseline 1	Baseline 2	Baseline 3	Baseline 4	Baseline 5	Baseline 6	Baseline 7	Baseline 8	Baseline 9	Baseline 10
Baseline 1	2075.50	2185.71	2285.71	2385.71	2485.71	2585.71	2685.71	2785.71	2885.71	2985.71
Baseline 2	2185.71	2285.71	2385.71	2485.71	2585.71	2685.71	2785.71	2885.71	2985.71	3085.71
Baseline 3	2285.71	2385.71	2485.71	2585.71	2685.71	2785.71	2885.71	2985.71	3085.71	3185.71
Baseline 4	2385.71	2485.71	2585.71	2685.71	2785.71	2885.71	2985.71	3085.71	3185.71	3285.71
Baseline 5	2485.71	2585.71	2685.71	2785.71	2885.71	2985.71	3085.71	3185.71	3285.71	3385.71
Baseline 6	2585.71	2685.71	2785.71	2885.71	2985.71	3085.71	3185.71	3285.71	3385.71	3485.71
Baseline 7	2685.71	2785.71	2885.71	2985.71	3085.71	3185.71	3285.71	3385.71	3485.71	3585.71
Baseline 8	2785.71	2885.71	2985.71	3085.71	3185.71	3285.71	3385.71	3485.71	3585.71	3685.71
Baseline 9	2885.71	2985.71	3085.71	3185.71	3285.71	3385.71	3485.71	3585.71	3685.71	3785.71
Baseline 10	2985.71	3085.71	3185.71	3285.71	3385.71	3485.71	3585.71	3685.71	3785.71	3885.71

하드웨어 퍼포먼스 점수

$$E_{PerfCost} = \sum_i \frac{BS_i * x_i}{SP_i}, \quad E_{OverPods} = \sum_i \frac{Req_{pod}}{Pod_i * x_i} \quad \text{minimize} \sum_i \left(-\alpha \cdot \frac{Perf_i}{Perf_{min}} + (1 - \alpha) \cdot \frac{SP_i}{SP_{min}} \right) \cdot x_i$$

constraints : $x_i \leq 0.4 \cdot T3_i, x_i \in \mathbb{Z}^{0+}$



KubePACS
지능형 클라우드 스팟 인스턴스 Kubernetes 노드 스케줄러

Heterogeneous GPU LLM Serving System

분산 LLM 추론 환경 구성 시 고려 사항 (모델 관점)

병렬처리 기법 – Tensor/Pipeline

모델 파티셔닝 전략

단계별 이질적 작업 특성 – Prefill/Decode

입력 배치 사이즈 결정

배치처리/실시간처리

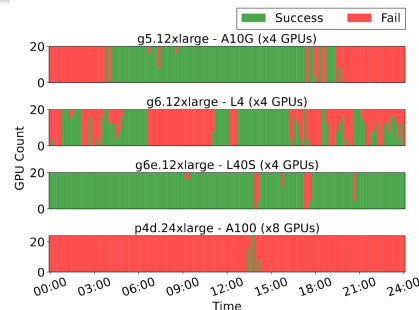
클라우드 환경에서 제공되는 다양한 종류의 GPU 인스턴스의 가용성 차이

Instance Type	GPU	On-Demand	Spot	Savings
g5.12xlarge	A10G (4x)	\$5.67/hr	\$2.26/hr	60.2%
g6.12xlarge	L4 (4x)	\$4.60/hr	\$1.90/hr	58.6%
g6e.12xlarge	L40S (4x)	\$10.49/hr	\$4.35/hr	58.6%
p4d.24xlarge	A100 (8x)	\$21.96/hr	\$9.68/hr	55.9%
p5.48xlarge	H100 (8x)	\$55.04/hr	\$20.72/hr	62.4%

AWS GPU 인스턴스 on-demand 및 spot 가격

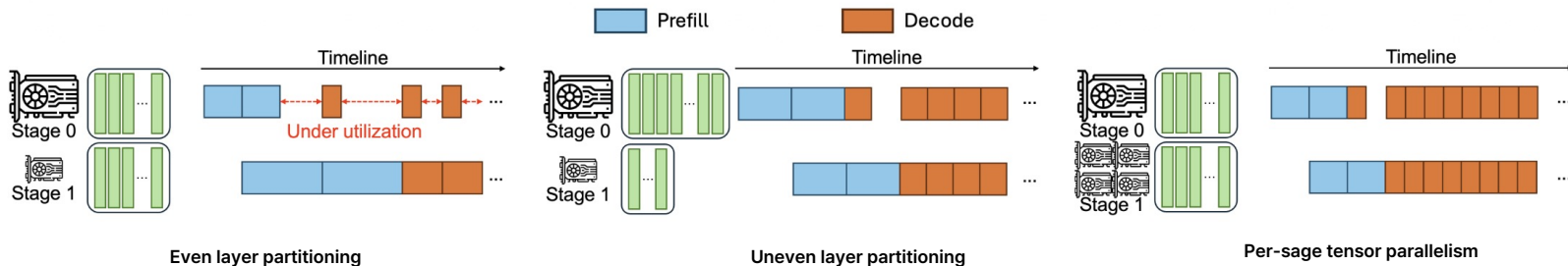
GPU	Mem. (GB)	FLOPS (TFLOPS)	Mem. BW. (GB/s)
L4	24	121	300
A10G	24	70	600
L40S	48	362	864
A100	40	312	1,555
H100	80	989	3,350

다양한 GPU 별 스펙



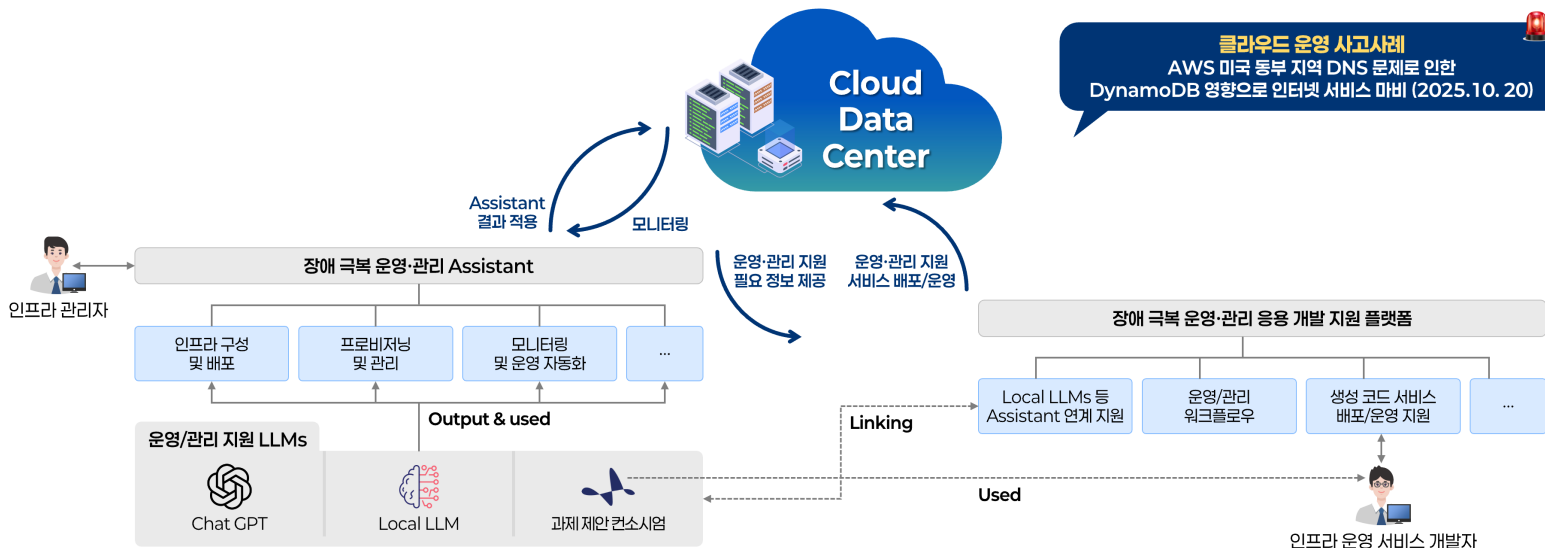
GPU 장착 스팟 인스턴스의 가용성

각기 다른 성능을 가지는 Heterogeneous GPU 로 추론 시스템 구성 시 최적 구성 방안은?



AIOps for Cloud

클라우드 장애극복을 위한 AI 어시스턴트 기반 운영관리 자동화 기술개발 by IITP (~2028)



클라우드 운영 사고사례
AWS 미국 동부 지역 DNS 문제로 인한
DynamoDB 영향으로 인터넷 서비스 마비 (2025.10. 20)

AS-IS

- 인적오류, 느린 대응 속도 등 비효율적인 클라우드 운영·관리
- 과도한 클라우드 운영·관리 비용
- 사람의 개입을 통한 성능 및 확장성 저하

TO-BE

- 자동화된 운영·관리로 효율적인 클라우드 서비스 제공
 - AIOps의 단점을 보완하는 AI 어시스턴트(LLM) 개발
- 클라우드 자원, 워크로드 등의 효과적인 분산을 통한 성능 및 확장성 증대
 - 고도화된 RCA(Root Cause Analysis) 기반 지능형 추론

Discussions