



GEdge(Griffin-Edge) Platform

- 초저지연 지능형 클라우드 엣지 SW 플랫폼 -

엣지 기반 AI/ML 서비스 흐름관리 및 실행환경 최적화 기술

2021.12.09

GS-AI 프레임워크 리더(GS-AIflow)

김성용(sykim@softonnet.com)

“GEdge Platform” 은 클라우드 중심의 엣지 컴퓨팅 플랫폼을 제공하기 위한
핵심 SW 기술 개발 커뮤니티 및 개발 결과물의 코드명입니다.

- Developer-Friendly

GEdge Platform Community 3rd Conference (GEdge Platform v2.0 Release) -

이번 발표의 기술적 포지셔닝

초저지연 지능형 클라우드 엣지 플랫폼 (GEdge Platform)

클라우드 엣지 관리 플랫폼 (GM : GEdge Management Platform)

플랫폼 관리 도구 프레임워크 (GM-Tool)

Framework I/F

플랫폼 관리 기능 프레임워크 (GM-Center)

Platform I/F

지능형 서비스 운용 프레임워크 (GS-AI)

엣지 AI 서비스 환경
(GS-Aiflow)

엣지 협업 학습 환경
(GS-Optops)

Framework I/F

서비스 협업 프레임워크 (GS-Link)

협업 게이트웨이
(GS-Linkgw)

협업 정책 생성
(GS-Linkhq)

Framework I/F

초저지연 데이터 처리 프레임워크 (GS-Engine)

엣지 전용 스케줄러
(GS-Scheduler)

엣지 메시지 브로커
(GS-Broker)

클라우드 엣지 서비스 플랫폼 (GS : GEdge Service Platform)

Contents



지능형 서비스 운용 프레임워크



워크플로우 및 런타임 실행관리 기술



AI/ML 최적화 기술



연구 내용

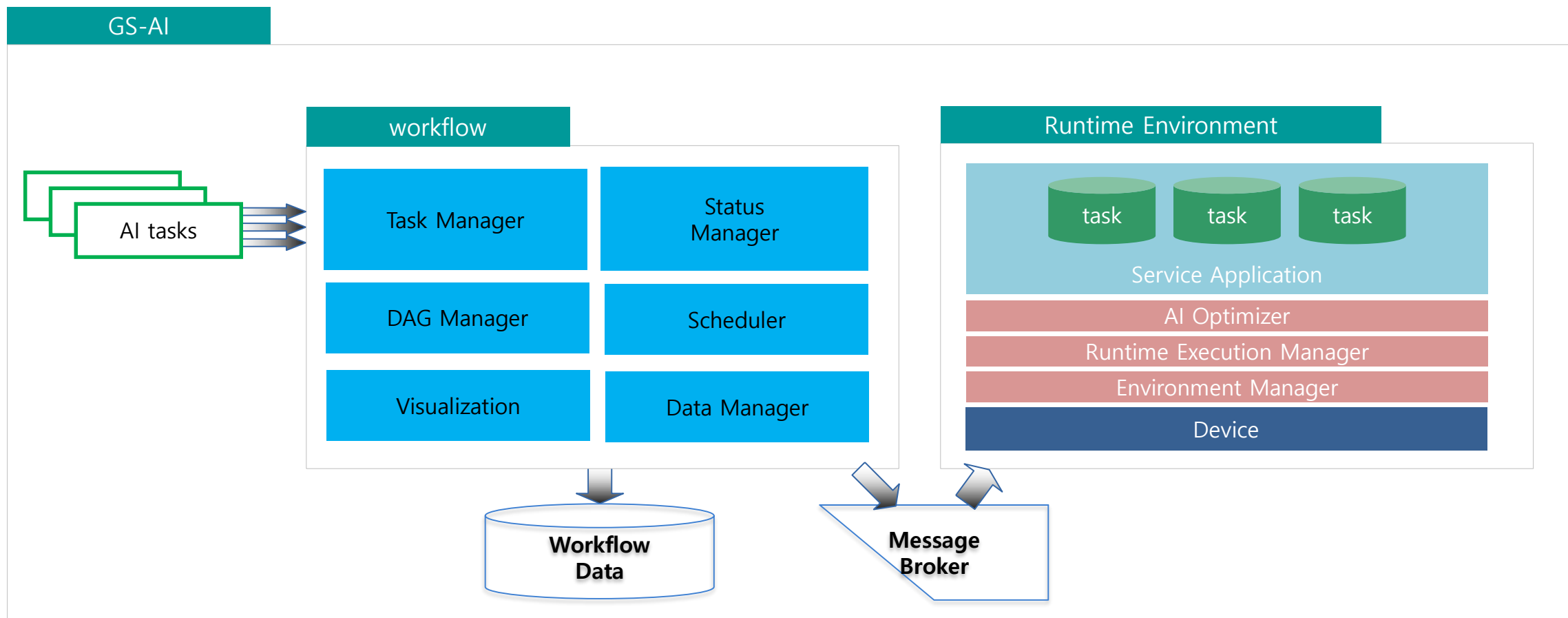


지능형 서비스 운용 프레임워크 개요



지능형 서비스 운용 프레임워크 구조

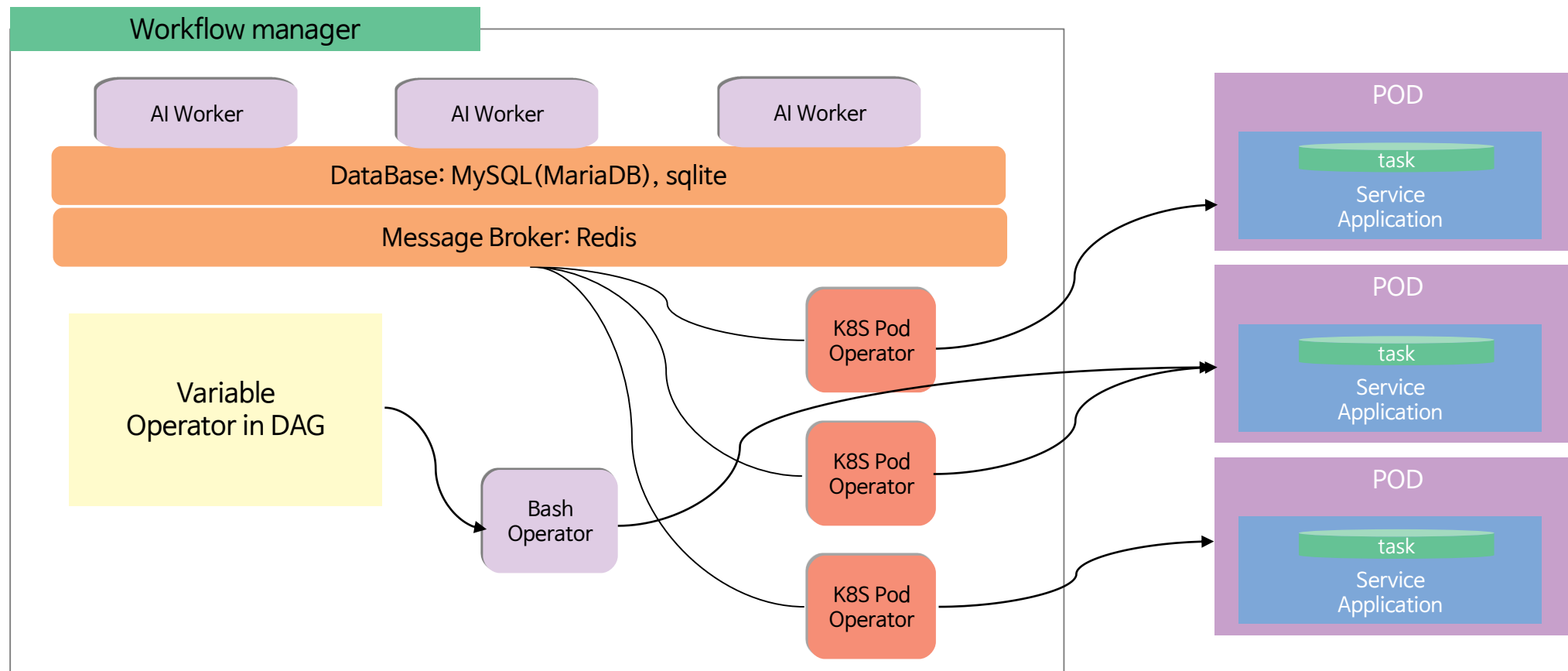
- 서비스를 담당하는 워크플로우 서비스 모듈과 실행 환경을 담당하는 런타임 관리 모듈로 구성



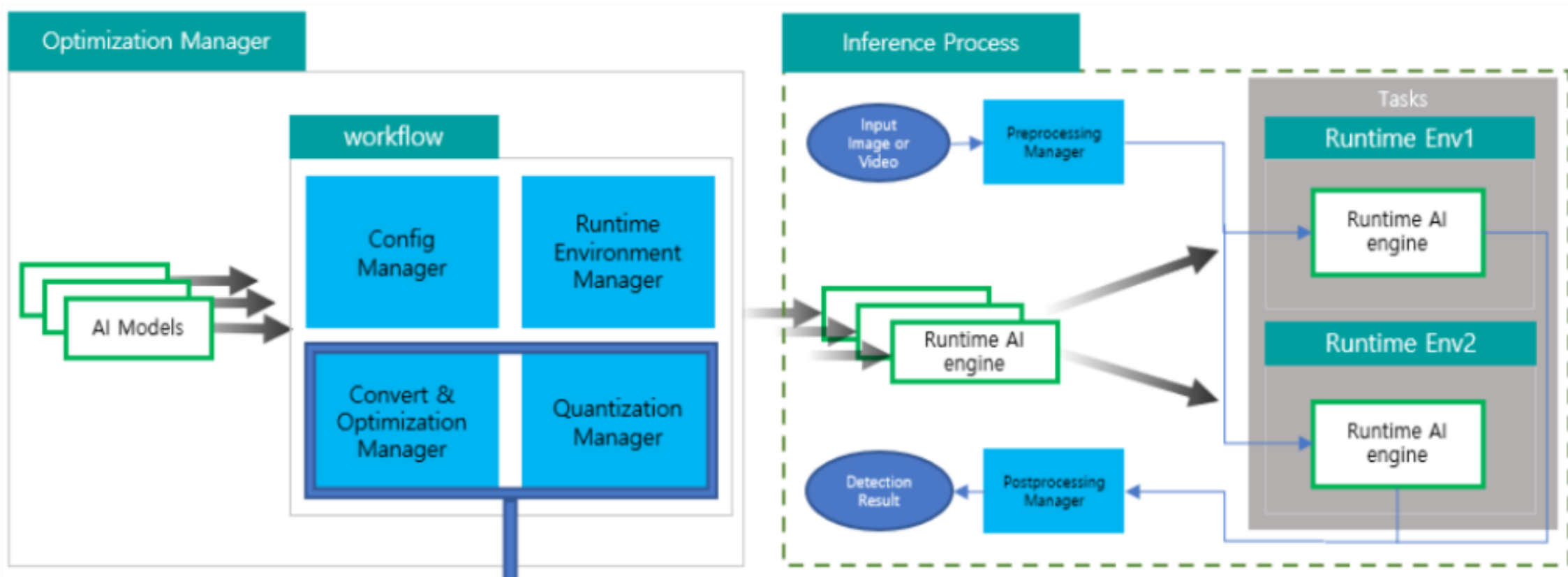
지능형 서비스 운용 프레임워크 동작

- 다양한 클라우드 엣지 환경에서 지능형 서비스를 실행하기 위한 워크플로우 관리

GS-AI



- 지능형 워크플로우 관리 기술
 - 최적화된 마이크로 서비스 실행 및 관리



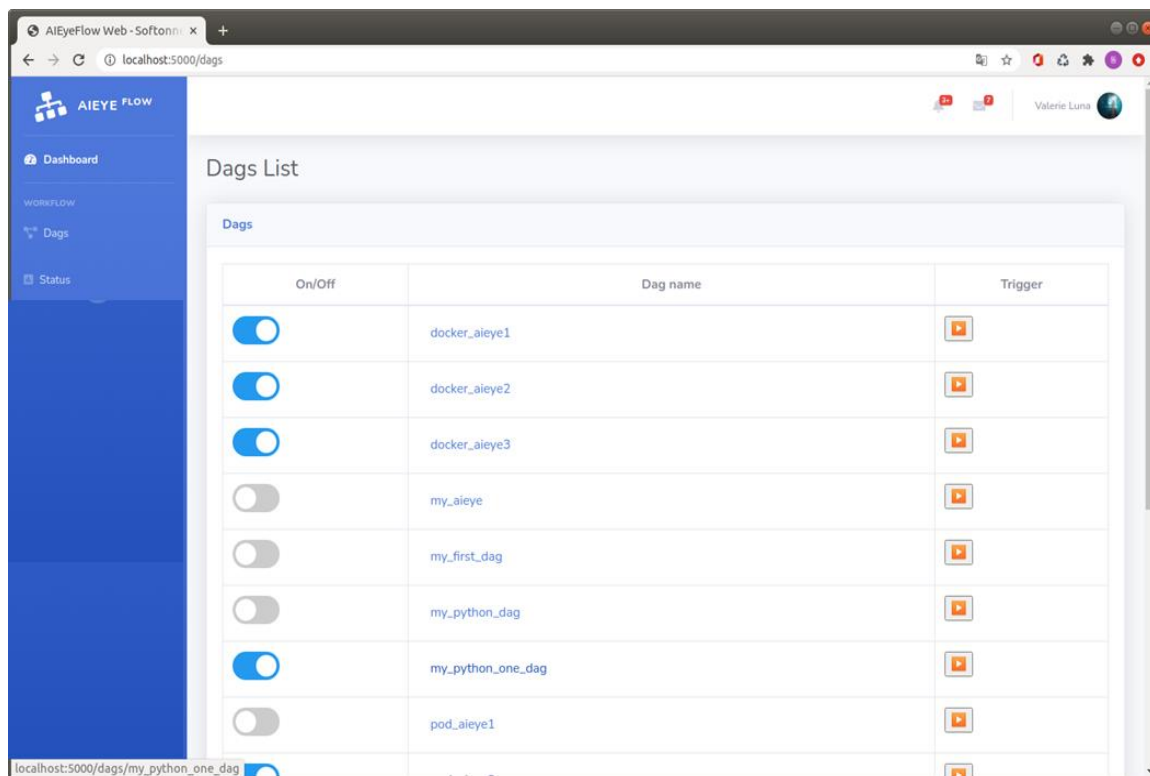


워크플로우 및 실행관리기술



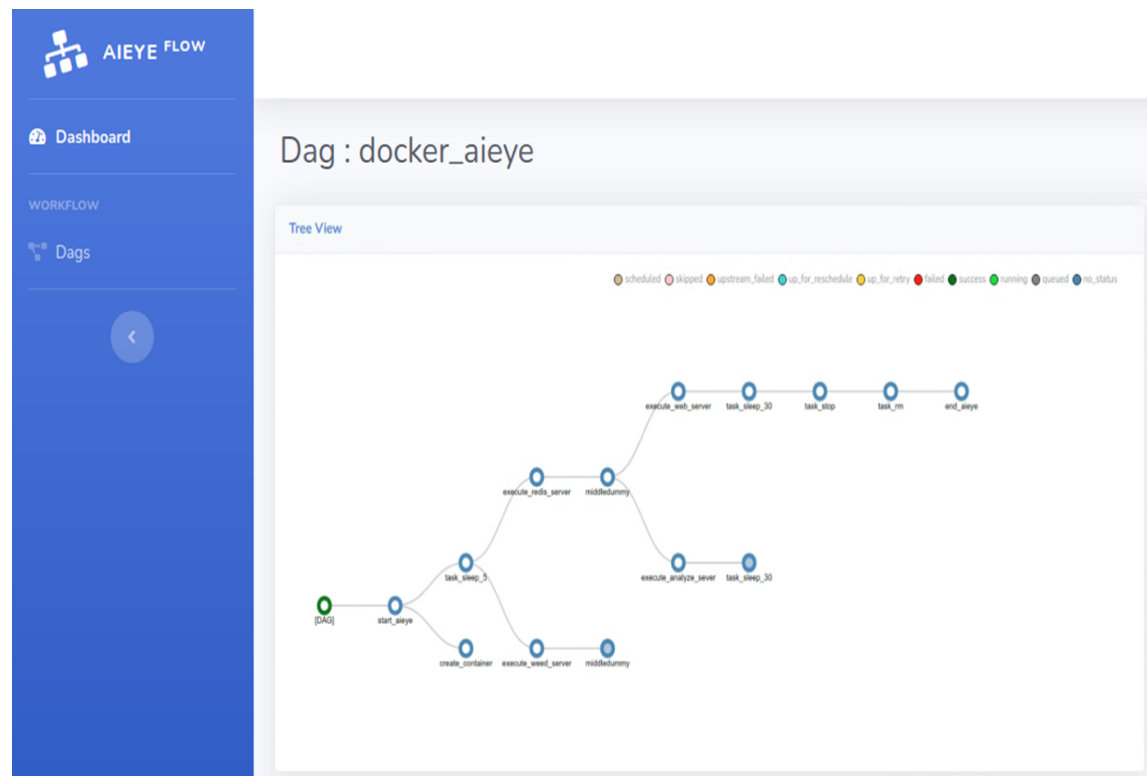
지능형 워크플로우 관리 개요

- 다양한 클라우드 엣지 환경에서 지능형 서비스를 실행하기 위한 워크플로우 관리 데시보드
- DAG의 형태를 Graph View, Tree View로 표현 Web UI 상에서 트리거 기능을 통해 스케줄링 수행
- 다양한 Executor 지원을 통해 병렬 및 분산 처리 지원



The screenshot shows the AIEYE FLOW web interface with a sidebar containing 'Dashboard', 'WORKFLOW', 'Dags', and 'Status'. The main area displays a 'Dags List' table with columns for 'On/Off', 'Dag name', and 'Trigger'.

On/Off	Dag name	Trigger
<input checked="" type="checkbox"/>	docker_aieye1	
<input checked="" type="checkbox"/>	docker_aieye2	
<input checked="" type="checkbox"/>	docker_aieye3	
<input type="checkbox"/>	my_aieye	
<input type="checkbox"/>	my_first_dag	
<input type="checkbox"/>	my_python_dag	
<input checked="" type="checkbox"/>	my_python_one_dag	
<input type="checkbox"/>	pod_aieye1	



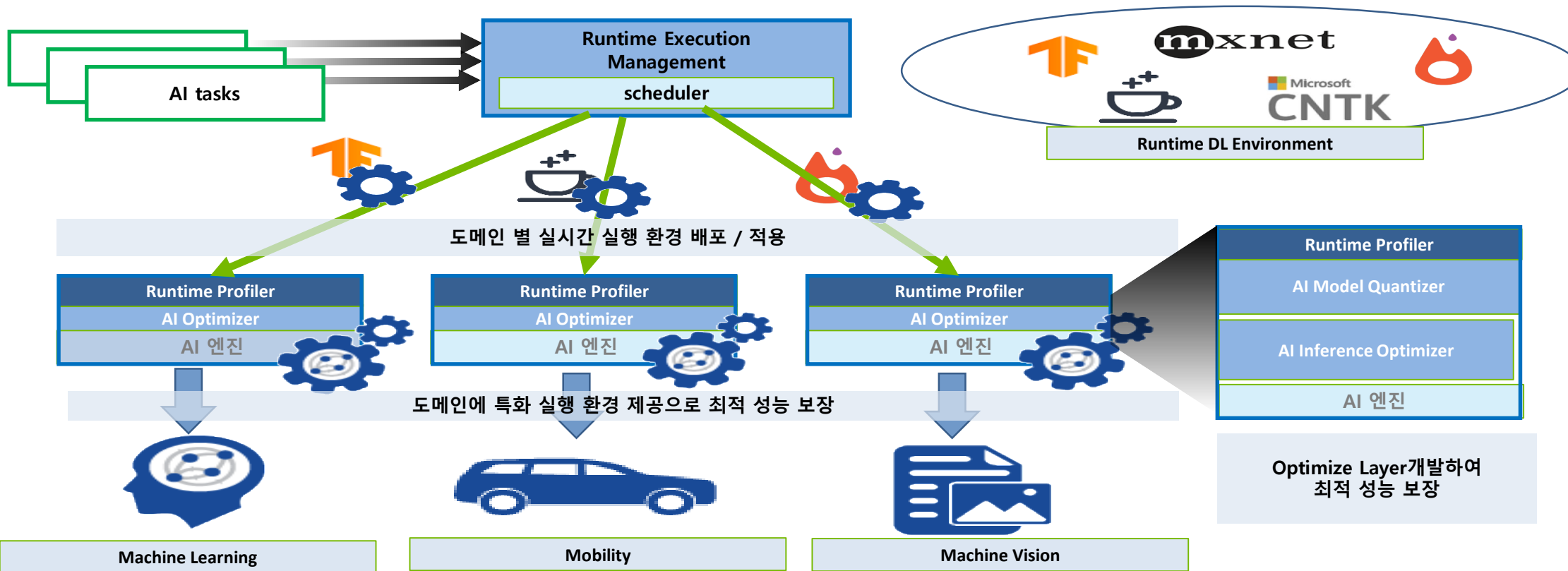
The screenshot shows the AIEYE FLOW web interface with a sidebar containing 'Dashboard', 'WORKFLOW', 'Dags', and 'Status'. The main area displays the 'Dag : docker_aieye' in 'Tree View'.

Legend: scheduled skipped upstream_failed up_for_retry failed success running queued no_status

```
graph LR; [DAG] --> start_aieye; start_aieye --> task_sleep_5; task_sleep_5 --> create_container; create_container --> execute_web_server; execute_web_server --> middledummy; middledummy --> execute_redis_server; execute_redis_server --> execute_web_server; execute_web_server --> task_sleep_30; task_sleep_30 --> task_stop; task_stop --> task_rm; task_rm --> end_aieye;
```

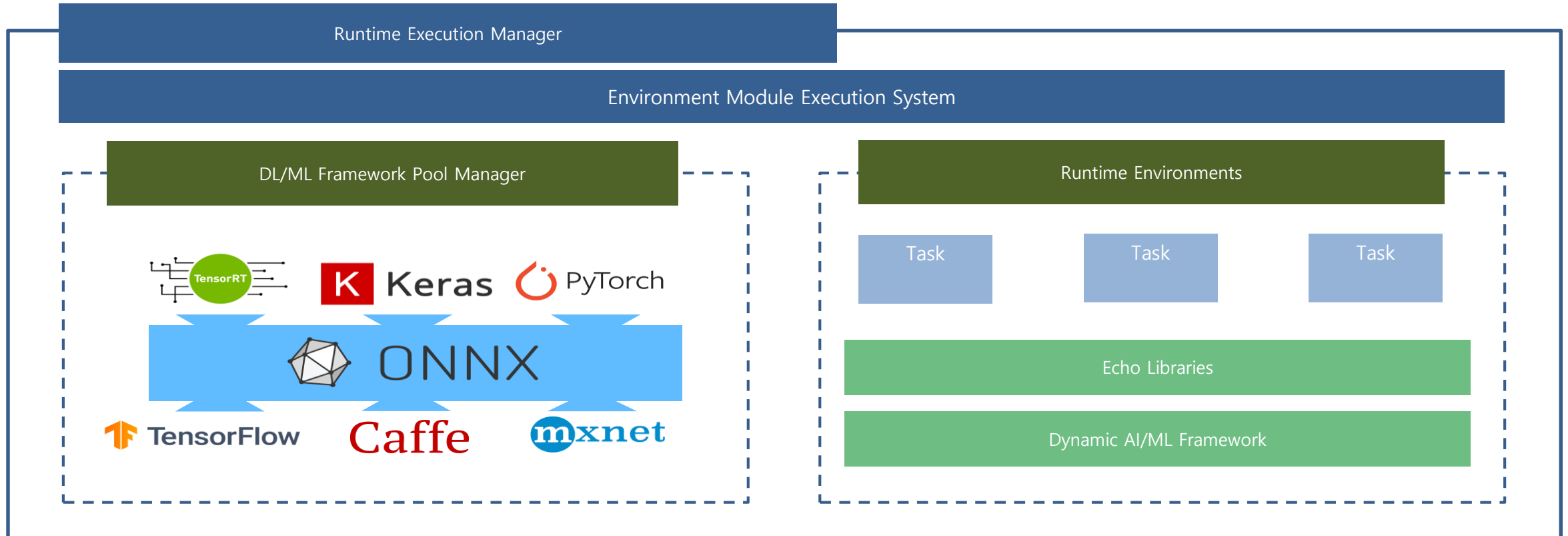
- 도메인 특화 AI 실행 런타임 관리 프레임워크
 - 다양한 도메인 지능형 서비스를 위한 실행 런타임 관리 기술

실행 런타임 관리기술 핵심 내용

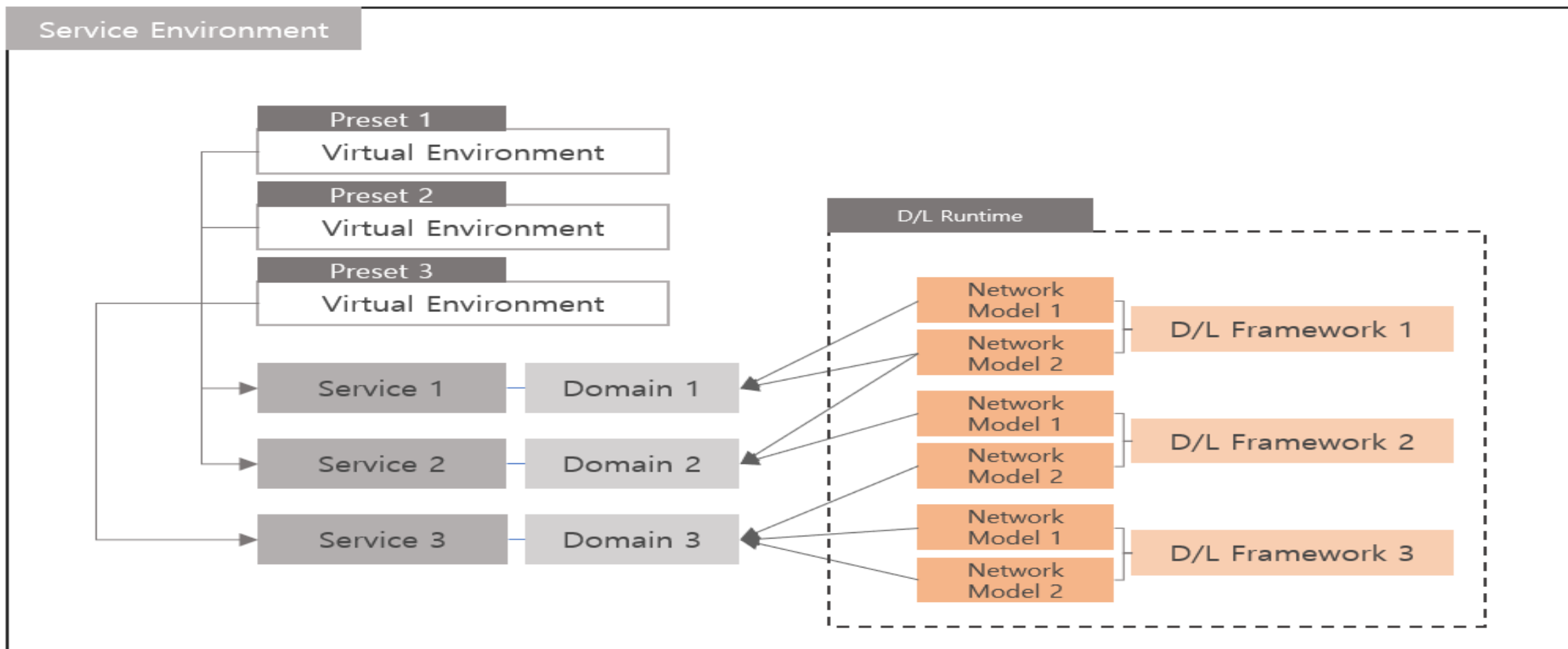


도메인 특화 AI 실행 런타임 관리 프레임워크

- 런타임 환경에 맞는 딥러닝 프레임워크 관리 기술
- 런타임 실행모듈에 맞는 딥러닝의 환경 실시간 배포
- 엣지 클라우드 환경에 최적화된 실행을 위한 최적화 모듈

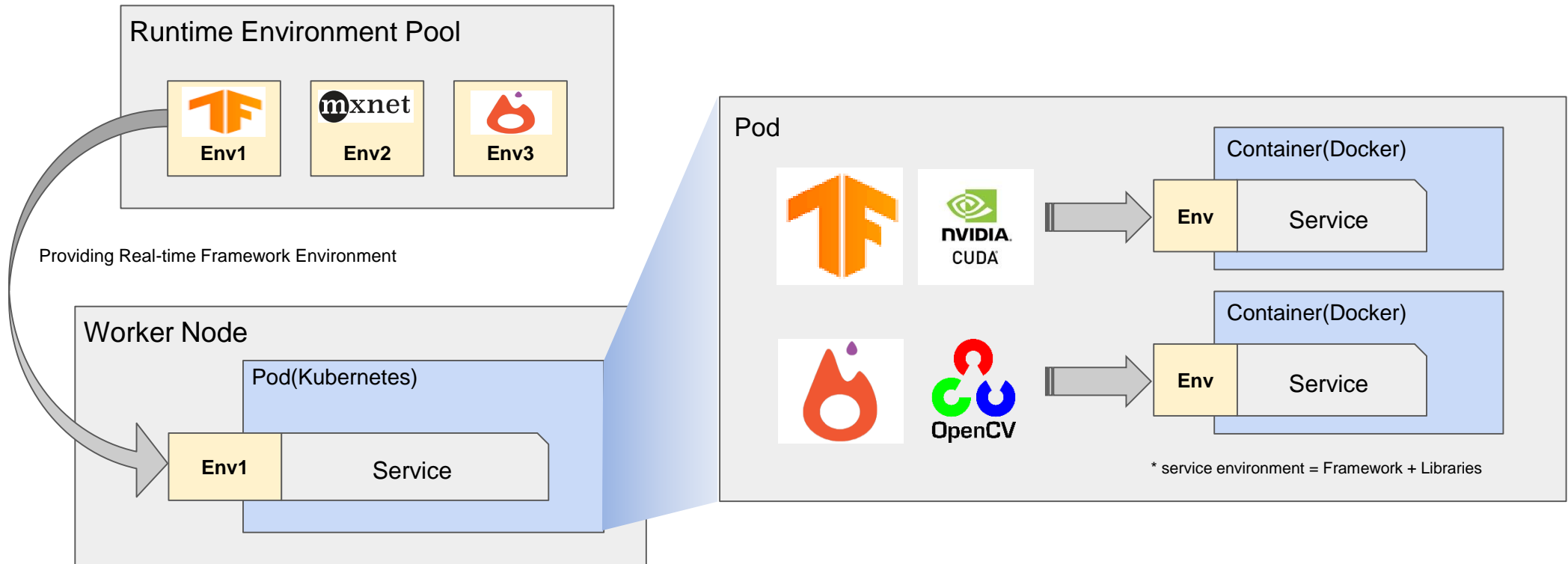


- 지능형 서비스 실행 관리 기술
 - 최적화된 마이크로 서비스 실행 및 관리



실행 런타임 환경 관리 기술

- 지능형 서비스의 요구사항에 맞는 환경을 실시간으로 구성



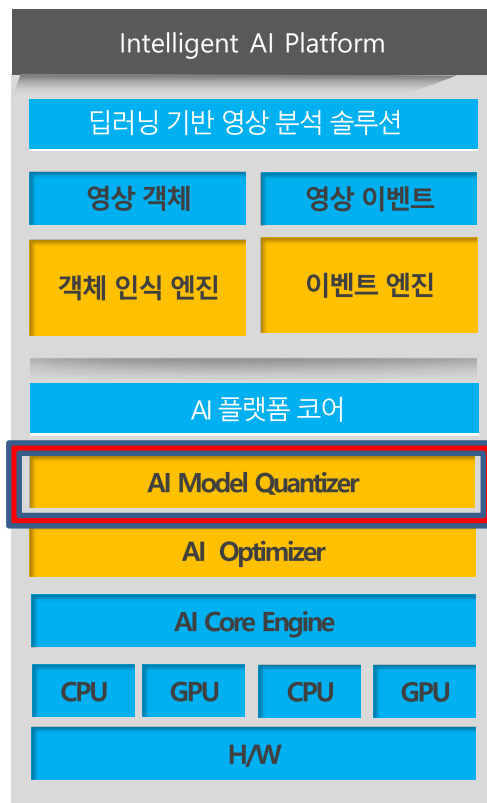


지능형 서비스 실행 최적화 기술



추론 모델 최적화 기술

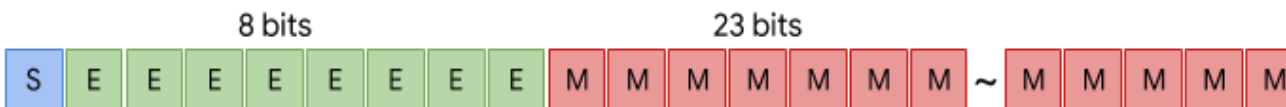
- FP32으로 되어 있는 Graph 연산값을 FP16, INT8등으로 Quantizing하여 연산의 속도를 높이는 기술



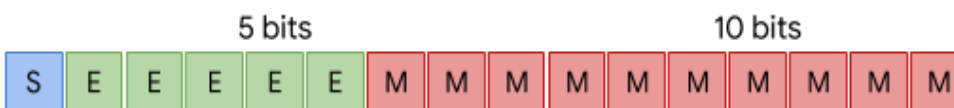
bfloat16
range: $\sim 1e^{-38}$ to $\sim 3e^{38}$



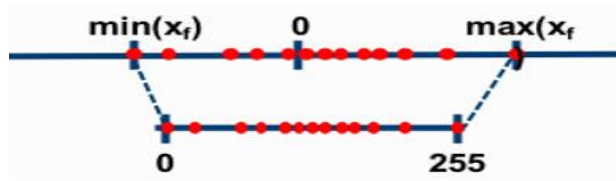
float32
range: $\sim 1e^{-38}$ to $\sim 3e^{38}$



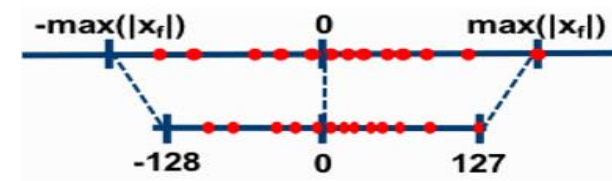
float16
range: $\sim 5.9e^{-8}$ to $6.5e^{-4}$



int8(uint8)



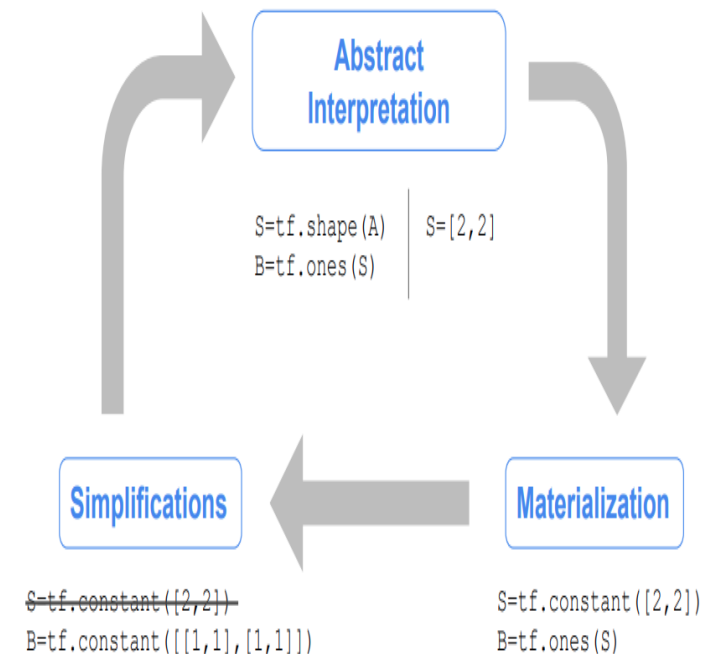
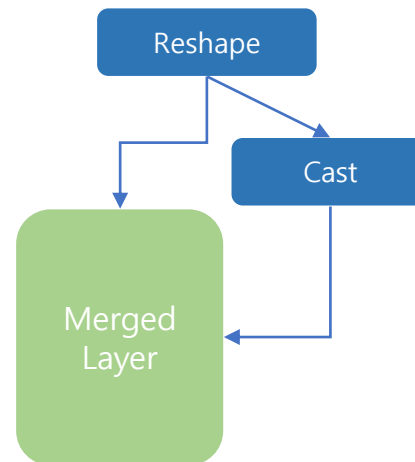
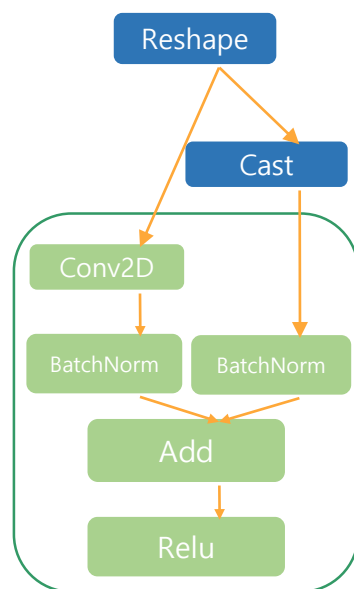
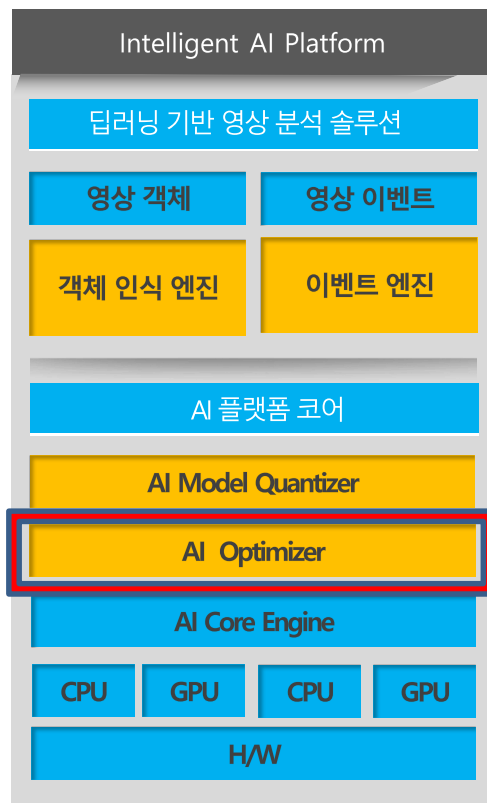
Asymmetric Mode



Symmetric Mode

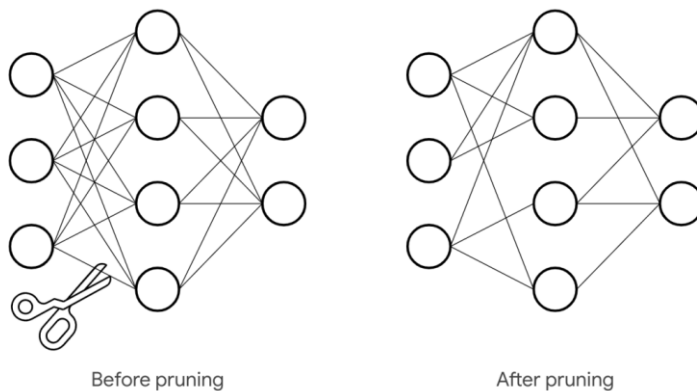
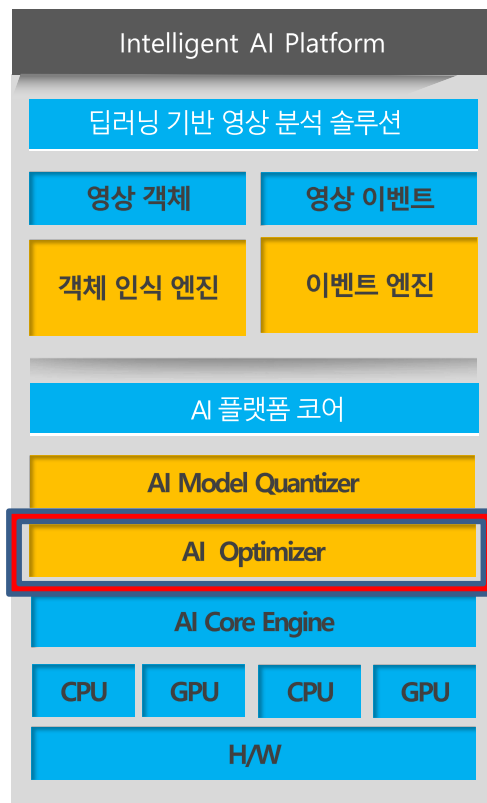
추론 모델 최적화 기술

- Graph의 사용하지 않는 레이어 및 파라미터의 Operation을 최적화 하는 기술

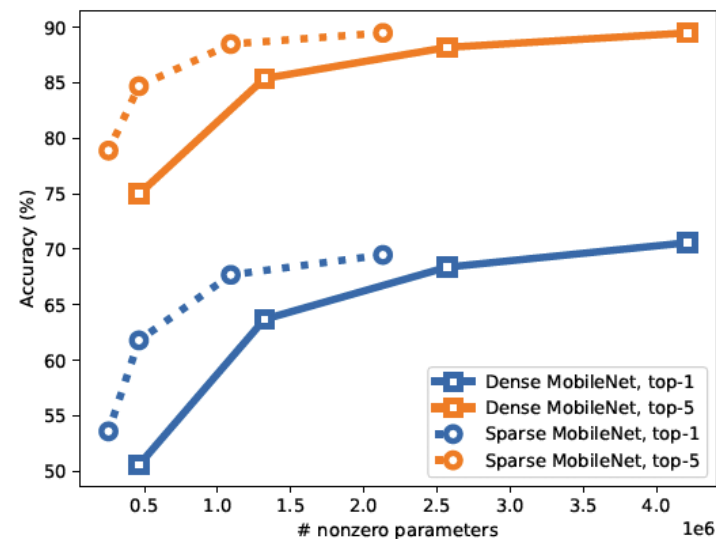


추론 모델 최적화 기술

- Graph의 가중치 중 작은 가중치 값을 모두 0으로 하여 네트워크의 모델 크기를 줄이는 기술



Pruning 개념



Pruning 전/후 변화 비교

- 추론 모델 최적화 성능 비교
 - 오리지널 그래프 및 최적화된 그래프 성능 비교

Accuracy 비교

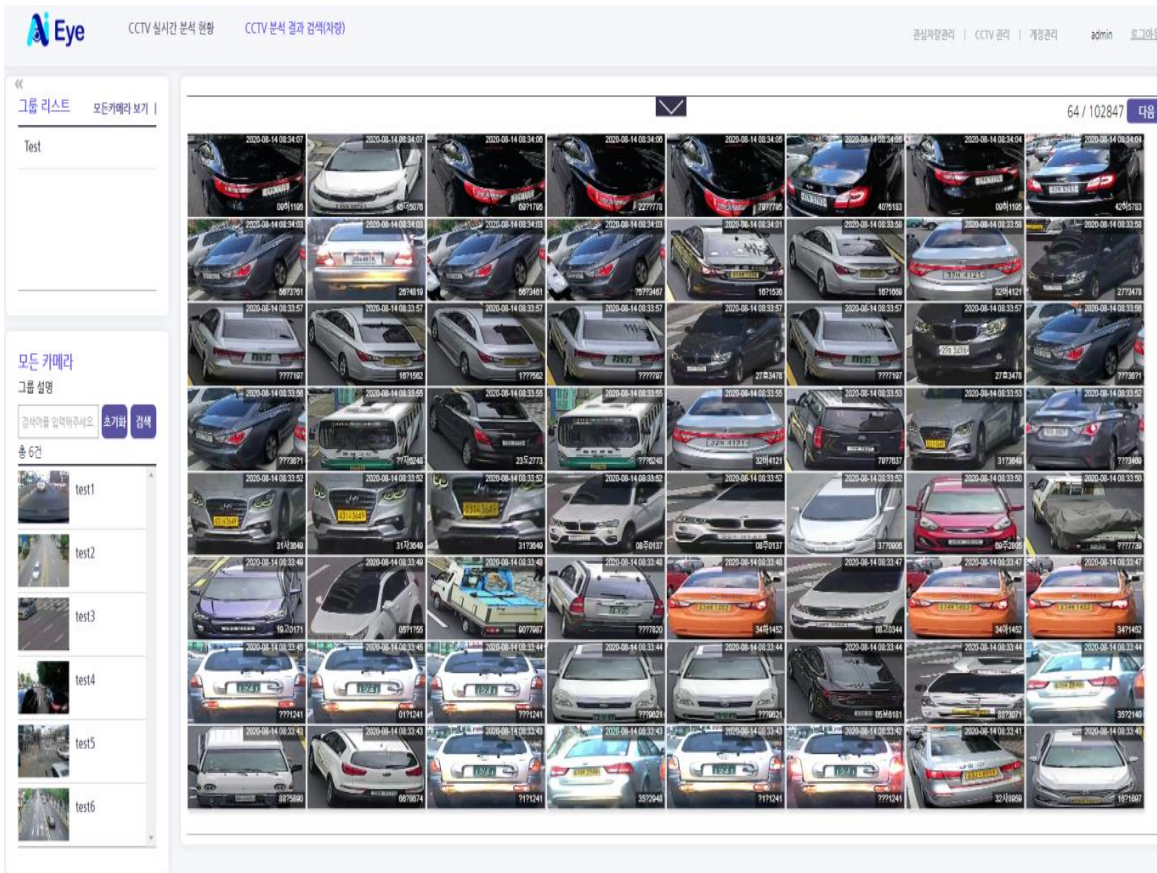
Model	Non-quantized Accuracy	8-bit Quantized Accuracy
MobilenetV1 224	71.03%	71.06%
Resnet v1 50	76.3%	76.1%
MobilenetV2 224	70.77%	70.01%

Latency and Accuray 비교 (tested on mobile device)

Model	Accuracy (Original)	Accuracy (Post Training Quantized)	Accuracy (Quantization Aware Training)	Latency (Original) (ms)	Latency (Post Training Quantized) (ms)	Latency (Quantization Aware Training) (ms)	Size (Original) (MB)	Size (Optimized) (MB)
MobilenetV1 224	0.709	0.657	0.70	124	112	64	16.9	4.3
MobilenetV2 224	0.719	0.637	0.709	89	98	54	14	3.6
InceptionV3	0.78	0.772	0.775	1130	845	543	95.7	23.9
ResnetV2 101	0.770	0.768	N/A	3973	2868	N/A	178.3	44.9

스마트시티 / 스마트공항 보안관제 활용

- 대량의 CCTV 영상을 분석하여 지연없이 사건 / 사고를 인식하고 예방할 수 있는 서비스




스마트시티 / 스마트공항 보안관계 활용

- 대량의 CCTV 영상을 분석하여 지연없이 사건 / 사고를 인식하고 예방할 수 있는 서비스

<<
그룹

- 전체 그룹
 - car
 - x dd
 - x v121
 - group1
 - x c1
 - cctv
 - ✓ 침입 감시(event)
 - x congestion
 - ✓ 혼잡도 측정(head)
 - x direction
 - x fire
 - ✓ 연기 감시(123)
 - x invasion
 - ✓ 침입 감시(e1)
 - x invasion2
 - ✓ 침입 감시(event1)
 - x neglected
 - x parking
 - ✓ 주차장 감지(parking1)
 - x ttee
 - test
 - x 45
 - x aa
 - x aisin
 - x dd
 - ✓ 침입 감시(21)
 - x v1

실시간 영상 ■



[GROUP : group1] - CAM : cctv

감시영역1

실시간 이벤트 결과

시간	CCTV	이벤트 종류	객체 종류	객체 색상	영상 보기
2020-11-09 15:36:51 ★	cctv	침입 감시	사람	알수없음	▶ 🔍
2020-11-09 15:36:51 ★	cctv	침입 감시	사람	알수없음	▶ 🔍
2020-11-09 15:36:51 ★	cctv	침입 감시	사람	알수없음	▶ 🔍
2020-11-09 15:36:50 ★	cctv	침입 감시	사람	알수없음	▶ 🔍
2020-11-09 15:36:50 ★	cctv	침입 감시	사람	알수없음	▶ 🔍
2020-11-09 15:36:49 ★	cctv	침입 감시	사람	알수없음	▶ 🔍
2020-11-09 15:36:47 ★	cctv	침입 감시	사람	알수없음	▶ 🔍
2020-11-09 15:36:47 ★	cctv	침입 감시	사람	알수없음	▶ 🔍

■ 그룹 ■ CCTV 동작중 x CCTV 비동작중
✓ 이벤트 동작중 x 이벤트 비동작중

중공업 및 건설등 산업 현장 중대 재해 보호를 위한 AI산업안전 솔루션

- 주요 산업 현장에서 발생하는 중대 재해 사고 예방을 위한 AI기반 초저지연 중장비 및 작업 안전 모니터링 서비스 제공



SmartEye_home x +

← → ↺ ⚠️ 안전하지 않음 | 172.16.16.79:5000/realtime.html

Eye CCTV 실시간 분석 현황 CCTV 분석 결과 검색(사람 및 사물)

CCTV 관리 | 계정관리 admin 로그아웃

그룹 리스트 모든카메라 보기 |

group1

모든 카메라 그룹 설명

검색어를 입력해주세요. 초기화 검색

총 1건

camera

camera : 2018-01-29 01:23:06

person (7.5m) person (4.4m)

사람 접근 감지 (2018-01-29 01:23:06)
카메라(camera)에서
1.5m 접근한 사람이 감지되었습니다.

사람 접근 감지 (2018-01-29 01:22:52)
카메라(camera)에서
2.8m 접근한 사람이 감지되었습니다.

사람 접근 감지 (2018-01-29 01:22:23)
카메라(camera)에서
2.9m 접근한 사람이 감지되었습니다.

사람 접근 감지 (2018-01-29 01:22:06)
카메라(camera)에서
0.7m 접근한 사람이 감지되었습니다.

사람 접근 감지 (2018-01-29 01:22:01)
카메라(camera)에서
2.7m 접근한 사람이 감지되었습니다.

사람 접근 감지 (2018-01-29 01:21:52)
카메라(camera)에서
2.7m 접근한 사람이 감지되었습니다.

사람 접근 감지 (2018-01-29 01:21:14)
카메라(camera)에서
2.7m 접근한 사람이 감지되었습니다.

사람 접근 감지 (2018-01-29 01:21:09)
카메라(camera)에서
2.8m 접근한 사람이 감지되었습니다.

사람 접근 감지 (2018-01-29 01:20:37)
카메라(camera)에서
2m 접근한 사람이 감지되었습니다.

사람 접근 감지 (2018-01-29 01:20:25)
카메라(camera)에서
2.9m 접근한 사람이 감지되었습니다.

사람 접근 감지 (2018-01-29 01:20:11)

무엇이든 찾아보세요

오후 4:28 2021-05-27

IV 2022 연구 계획



2022 지능형 서비스 운용 프레임워크 연구 및 개발 방안

인텔리전스 프레임워크에 배포되는 ML/DL 플랫폼 동작 환경 최적화 기술 연구 개발

인텔리전스 프레임워크에 동작하는 ML/DL 모델 최적화 기술 연구 개발

인텔리전트 프레임워크의 워크플로우 통합 대쉬보드 개발
인텔리전트 프레임워크의 태스크별 동작 개발 고도화

2022
1Q

2022
2Q

2022
3Q

2022
4Q

지능형 엣지 클라우드 지능형 서비스 통합 / 테스트

도커를 활용한 ML/DL 플랫폼별 실행 런타임 프로비저닝 기술 연구 개발

실시간 런타임 AI/ML 프레임워크 환경 제공 기술 연구 개발



초저지연 지능형 엣지 클라우드 플랫폼

감사합니다.

<http://gedge-platform.github.io>



GS-AI 프레임워크 리더(GS-AIflow)

김성용(sykim@softonnet.com)

Welcome to GEdge Platform

An Open Cloud Edge SW Platform to enable Intelligent Edge Service

GEdge Platform will lead Cloud-Edge Collaboration