GEdge Platform

GEdge(Griffin-Edge) Platform

- 초저지연 지능형 클라우드 엣지 SW 플랫폼 -

Edge AI를 위한 딥러닝 실행 환경 최적화 / 작업 흐름 관리 기술 (GS-AI)

2020.12.10

GEdge Platform 코어 개발자

김성용(sykim@softonnet.com)

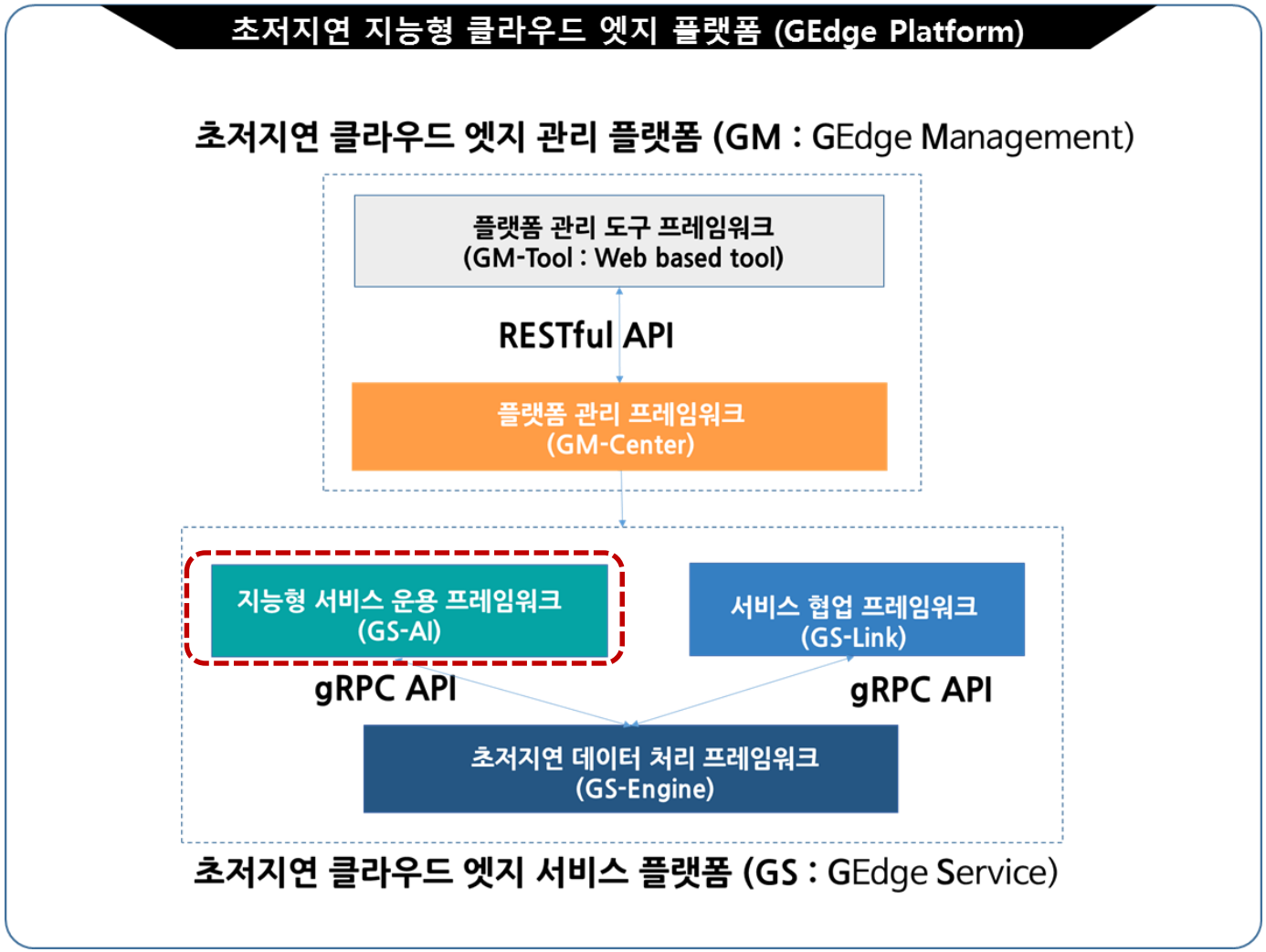
"The First talk of Edge Computing with Clouds"

- GEdge Platform 커뮤니티 멤버들의 첫번째 이야기 -

GEdge Platform Community 1st Conference

Contents

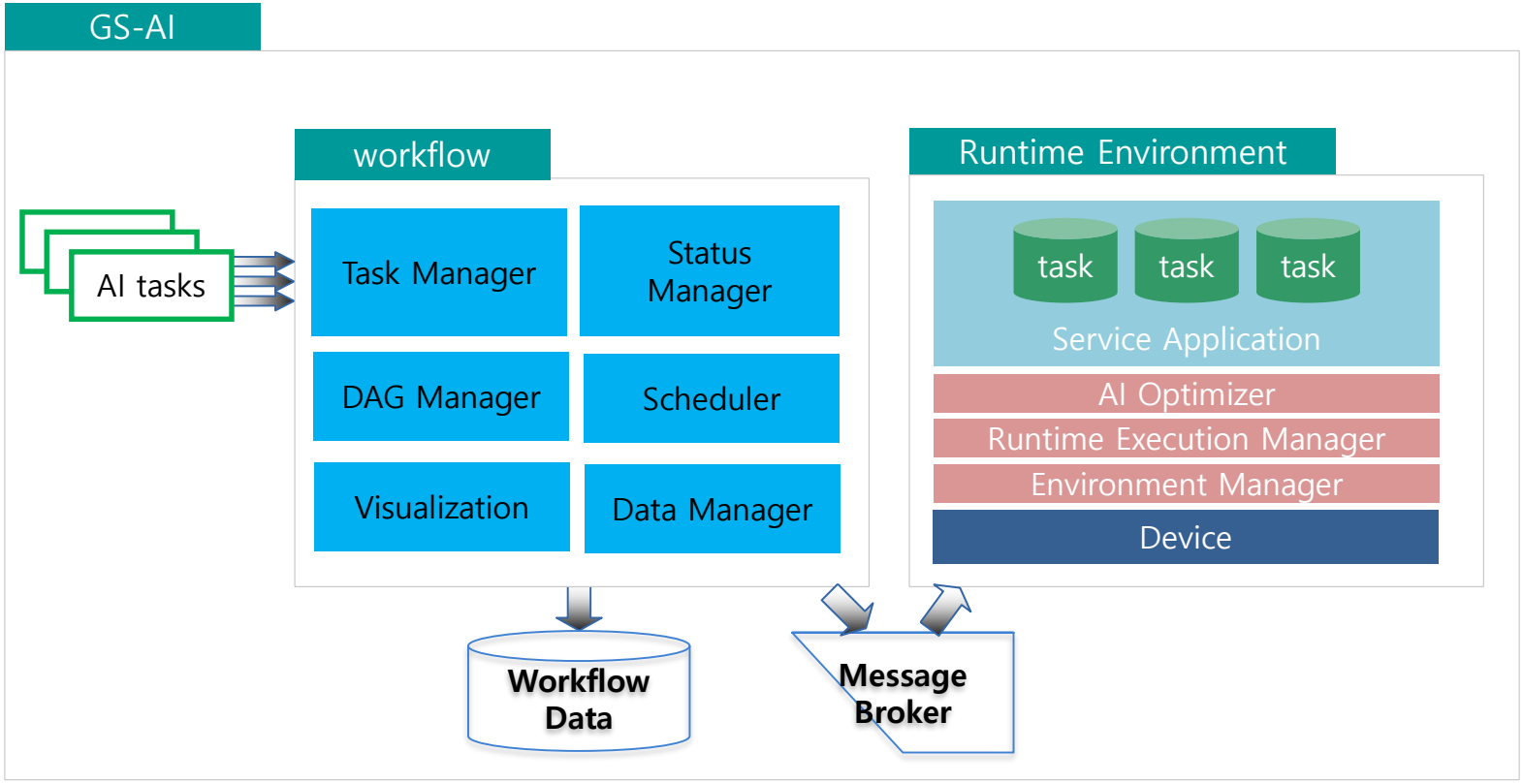
지능형 서비스 운용 프레임워크



1 지능형 서비스 운용 프레임워크 개요

지능형 서비스 운용 프레임워크 구조

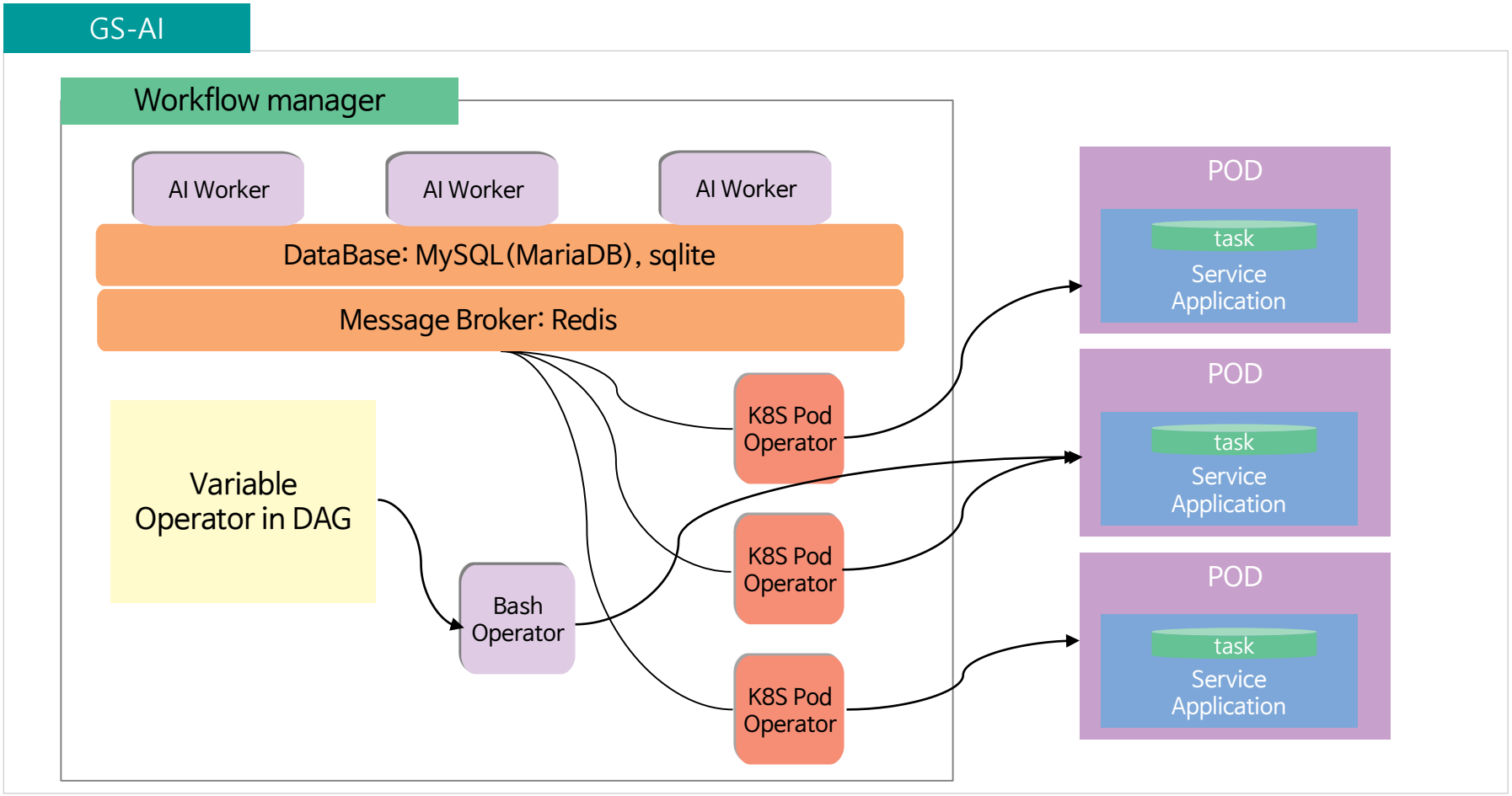
- 서비스를 담당하는 워크플로우 서비스 모듈과 실행 환경을 담당하는 런타임 관리 모듈로 구성



2 지능형 서비스 운용 프레임워크 동작

지능형 서비스 운용 프레임워크 동작

- 다양한 클라우드 엣지 환경에서 지능형 서비스를 실행하기 위한 워크플로우 관리

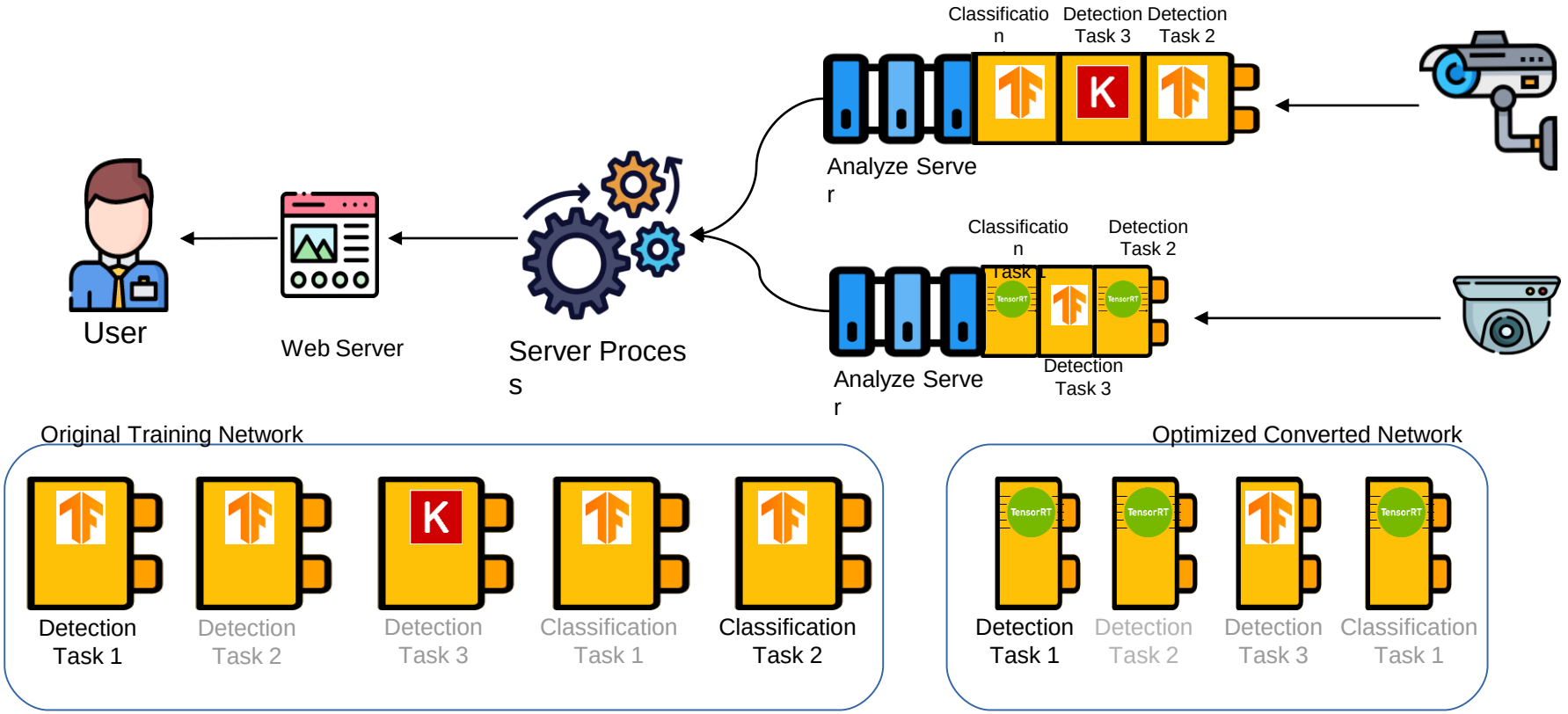


3 지능형 서비스 운용 프레임워크 동작

워크 플로우 관리 기술

최적화된 마이크로 서비스 실행 기술

- 최적화된 딥러닝 서비스 실행



4 지능형 서비스 운용 프레임워크 동작

지능형 서비스 운용 프레임워크 동작 UI

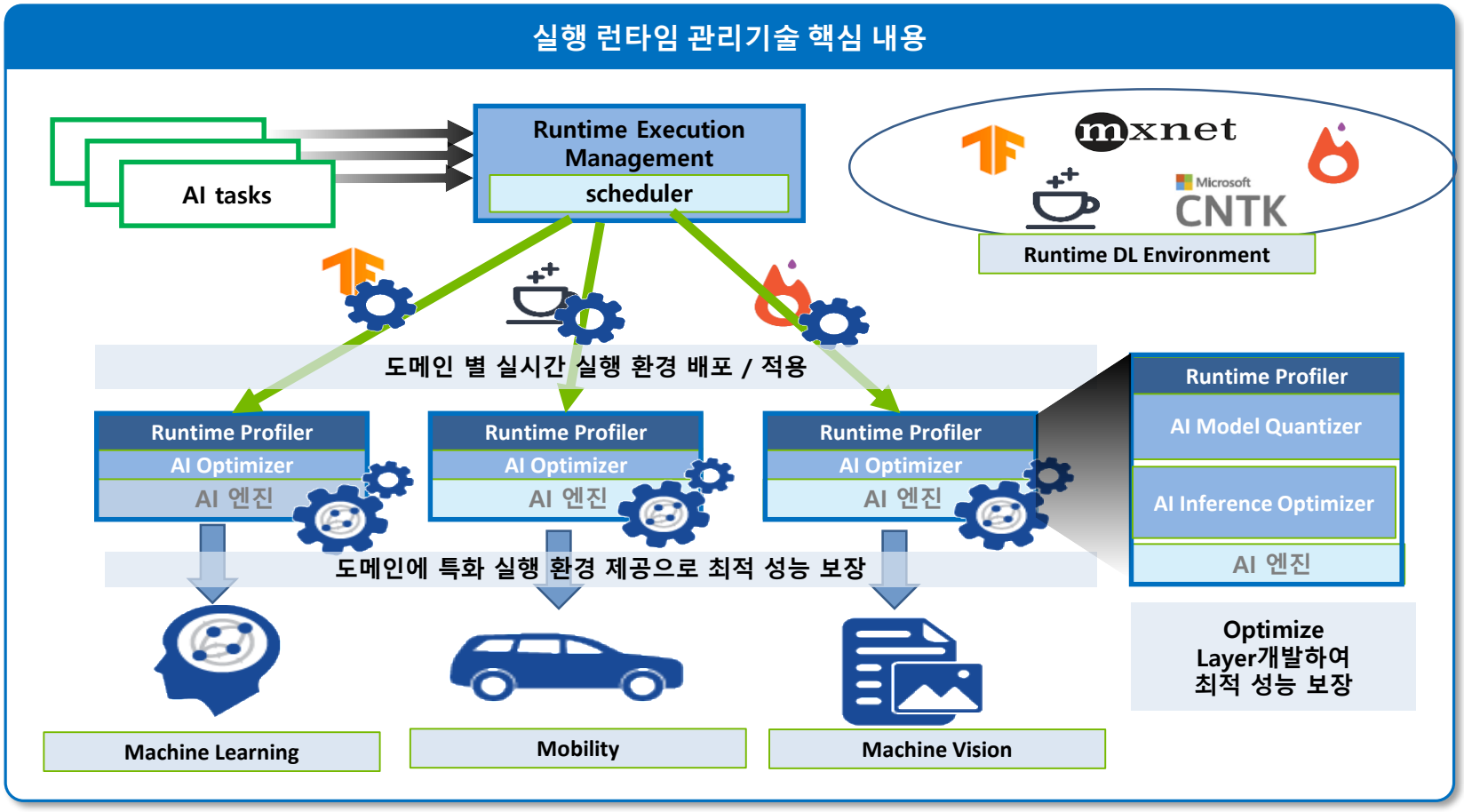
다양한 엣지 클라우드 환경에서 지능형 서비스를 실행하기 위한 워크플로우 관리 데시보드

- DAG의 형태를 Graph View, Tree View로 표현 Web UI 상에서 트리거 기능을 통해 스케쥴링 수행
- 다양한 Executor 지원을 통해 병렬 및 분산 처리 지원

5 런타임 실행 관리 기술 개요

도메인 특화 AI 실행 런타임 관리 프레임워크

다양한 도메인 지능형 서비스를 위한 실행 런타임 관리 기술

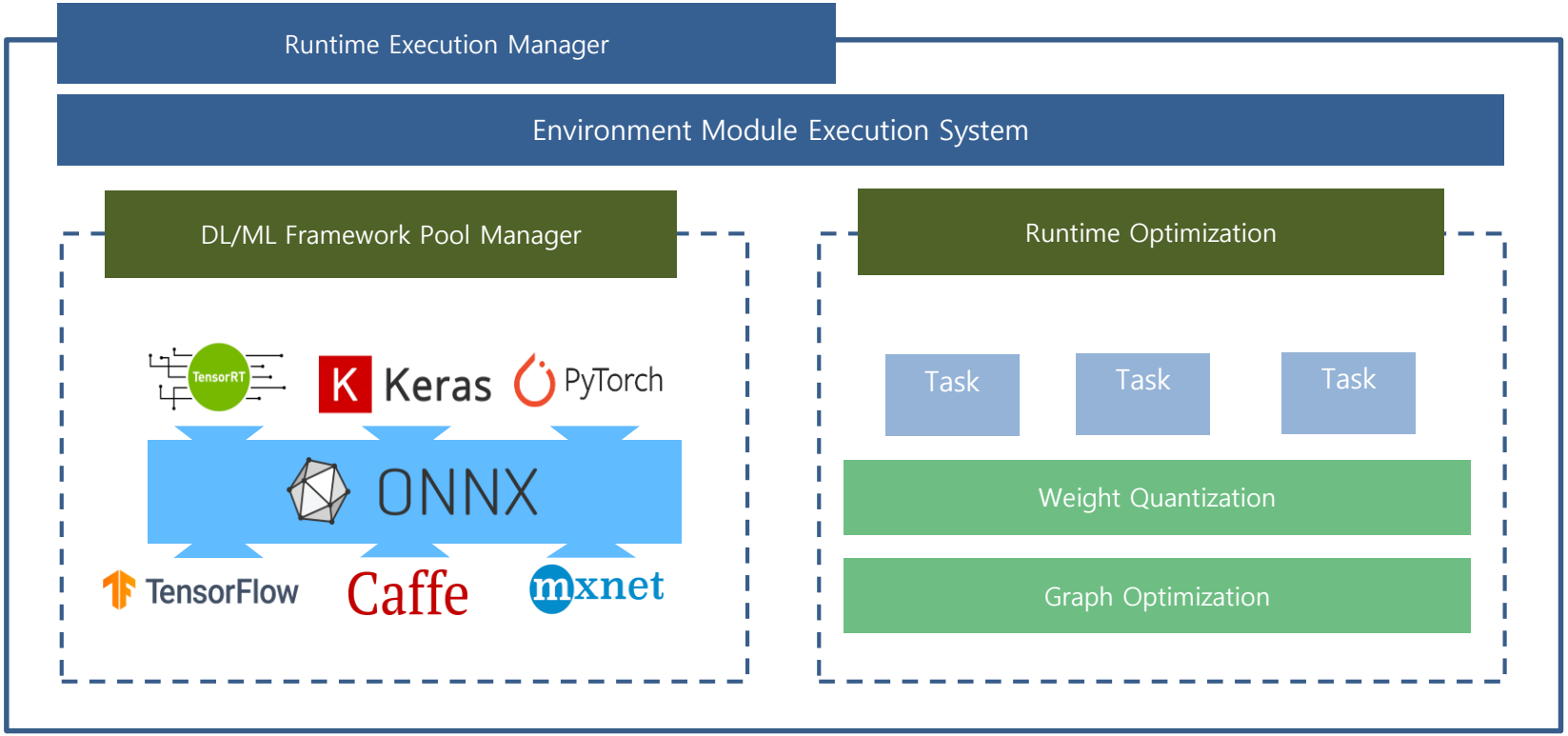


6 런타임 실행 관리 기술

도메인 특화 AI 프레임워크 관리 기술

런타임 환경에 맞는 딥러닝 프레임워크 관리 기술

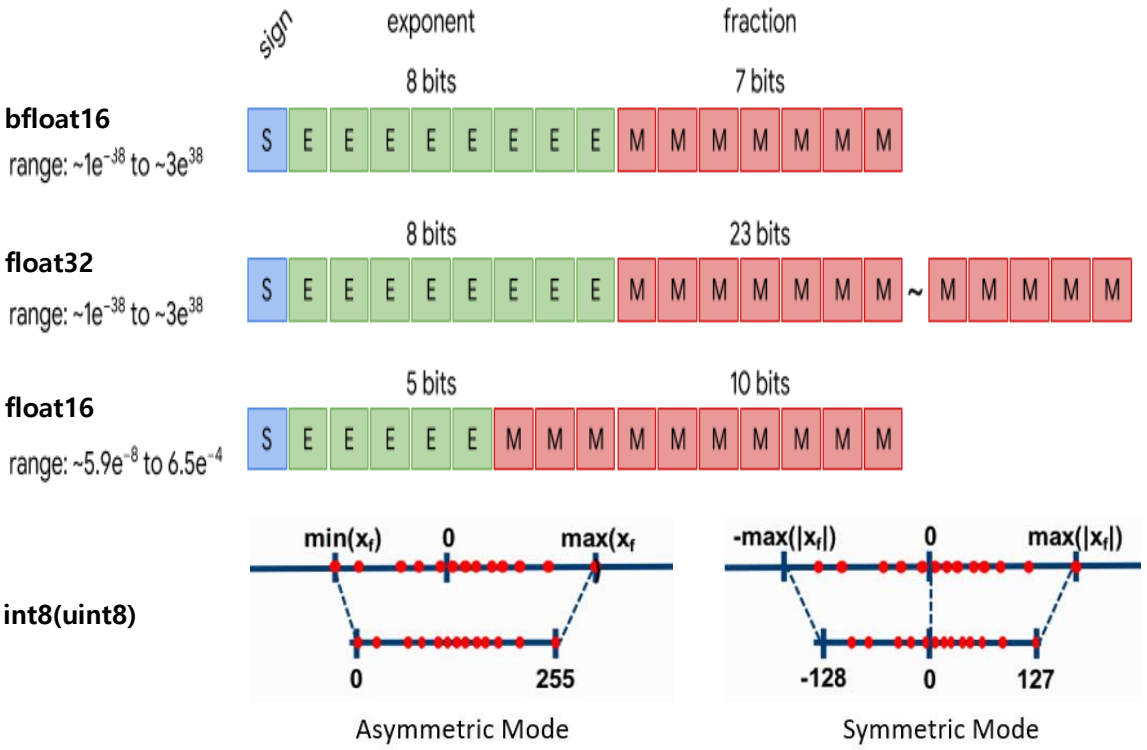
- 런타임 실행모듈에 맞는 딥러닝의 환경 실시간 배포
- 엣지 클라우드 환경에 최적화된 실행을 위한 최적화 모듈



7 런타임 실행 최적화

추론 모델 최적화 기술

- Graph 연산값을 FP16, INT8등으로 Quantizing하여 연산의 속도를 높이는 기술



8 런타임 실행 최적화

GEdge Platform

추론 모델 최적화 기술

Graph 연산값을 FP16, INT8등으로 Quantizing하여 연산의 속도를 높이는 기술

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%

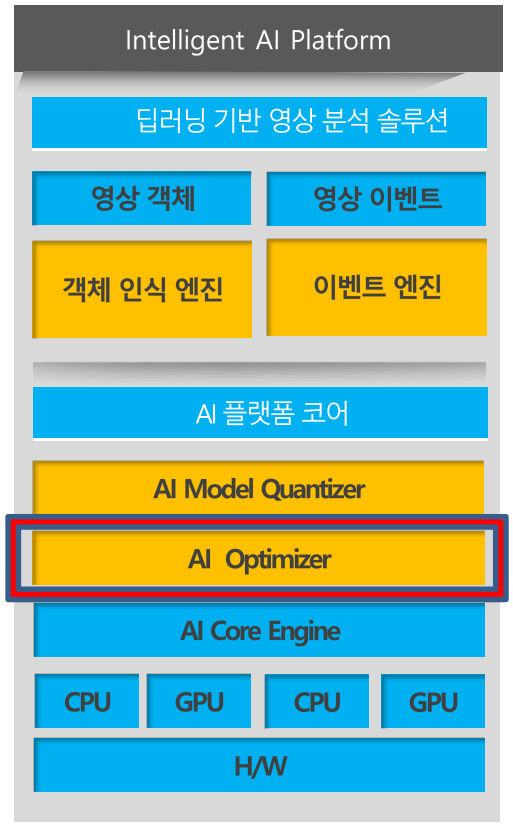
Latency and Accuray 비교 (tested on mobile device)

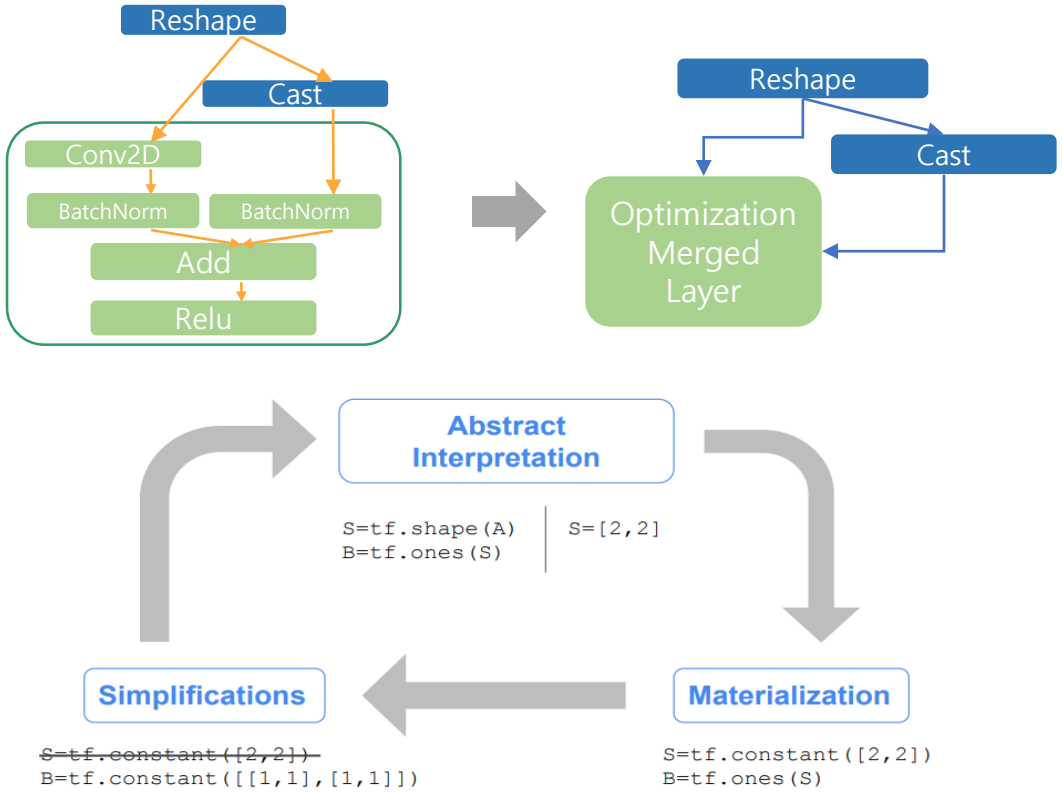
Model	Accuracy (Original)	Accuracy (Post Training Quantized)	Accuracy (Quantization Aware Training)	Latency (Original) (ms)	Latency (Post Training Quantized) (ms)	Latency (Quantization Aware Training) (ms)	Size (Original) (MB)	Size (Optimized) (MB)
MobilenetV1 224	0.709	0.657	0.70	124	112	64	16.9	4.3
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

9 런타임 실행 최적화 기술

추론 모델 최적화 기술

- Graph의 사용하지 않는 레이어 및 파라미터의 Operation을 최적화 하는 기술





10 실행 런타임 최적화 실적용 테스트

GEdge Platform

도메인 특화 AI 프레임워크 기술

실행 런타임 최적화 기술을 적용하기 전과 적용후의 성능 비교

Static Input shape

[Detection: Faster RCNN – Static HD, Static FHD]

Original Graph Operation Nodes: 3170
Segmentation Size: 10 / TensorRT Engine OP: 7
FP16 New Graph Nodes: 1219

VS

Dynamic Input shape

[Detection: Faster RCNN – Dynamic]

Original Graph Operation Nodes: 3702
Segmentation Size: 10 / TensorRT Engine OP: 6
FP16 New Graph Nodes: 2106

Input shape가 Static 인 경우 최적화 효율적

Object Detection(HD) + Tracking

	Detection	Tracking	총 소요 시간
Original	0.041~0.042	0.005~0.007	0.046~0.049
Optimization	0.027~0.028	-	0.032~0.035

VS

Object Detection(FHD) + Tracking

	Detection	Tracking	총 소요 시간
Original	0.041~0.045	0.005~0.007	0.046~0.052
Optimization	0.027~0.028	-	0.032~0.035

- inference time만 고려 시 약 60% 향상

- inference time만 고려 시 약 60% 향상

최적화를 통해 약 60%성능 향상

11 실행 런타임 최적화 적용 테스트

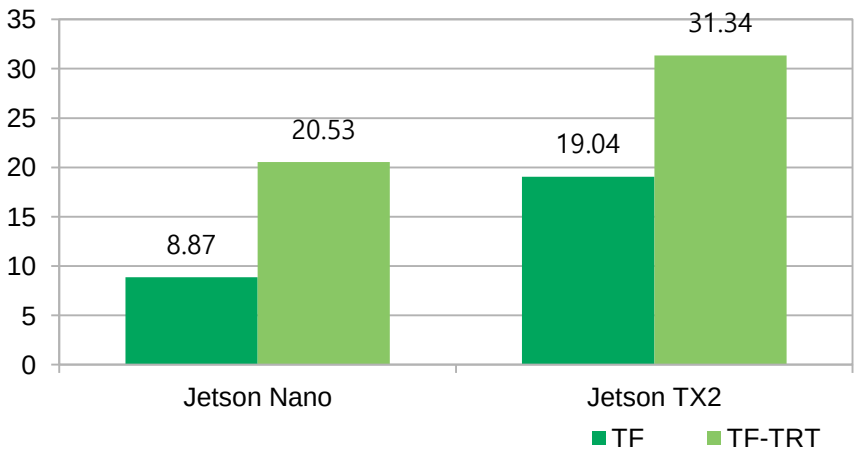
도메인 특화 AI 프레임워크 기술

실행 런타임 최적화 기술을 적용하기 전과 적용후의 성능 비교(엣지 환경)

- 라이브러리 별 inference 그래프 벤치마크
  - 단일 이미지 추론 테스트 : TF(.pb) vs TF-TRT(.pb)
  - Jetson NANO, TX2 - Jetpack 4.4 / TensorRT 7.1.3 / TF 1.15.2
  - test model: SSD Mobilenet v2 coco (object detection)

Jetson Nano(Object Detect)	TF	TF-TRT
Model Upload time(sec)	38.09	8.50
1 frame Inference time(sec)	0.1127	0.0487
FPS	8.87	20.53
Memory Use(def: 0.45GB)	3.8+1.2	2.55

Jetson TX2(Object Detect)	TF	TF-TRT
Model Upload time(sec)	14.85	9.43
1 frame Inference time(sec)	0.0525	0.0319
FPS	19.04	31.34
Memory Use(def: 0.3GB)	3.9	2.6



엣지 디바이스 별 추론 성능(FPS)

12 21년도 개발 계획

클라우드 엣지 기반의 지능형 서비스 운용 최적화 지원 기술 개발

클라우드 엣지 기반의 지능형 서비스를 위한 엣지 인텔리젼스 프레임워크 및 경량화 기술 개발

- AI 엣지 단말 및 프레임워크 개발
- AI 임베디드 엣지 디바이스 레이어를 통한 워크로드 분석 시스템 개발

도메인 특화 지능 서비스의 실행 런타임 및 관리 기술 개발

- 도메인별 지능 서비스 환경 배포를 위한 딥러닝 프레임워크 가상환경 서비스 관리기술 개발

클라우드 엣지 기반의 지능형 서비스의 운영 관리 기술 개발

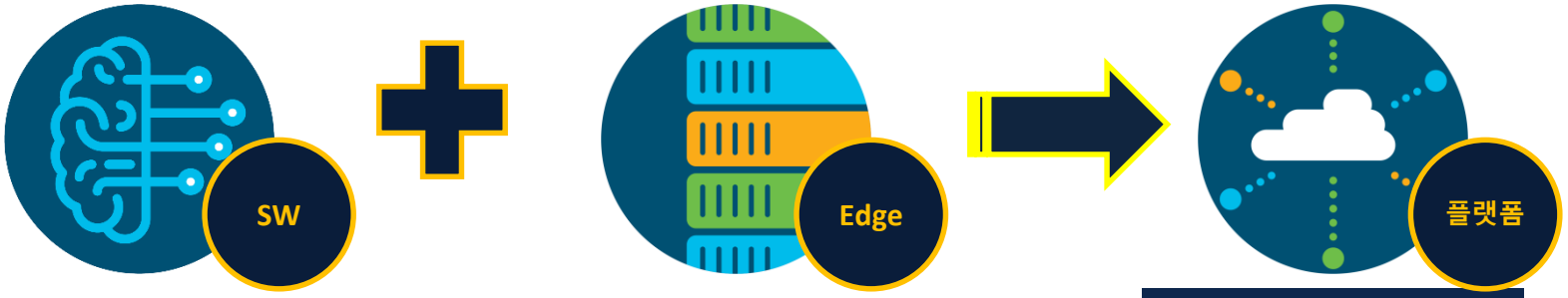
- 인텔리전트 프레임워크의 워크플로우 UI 개발
- 인텔리전트 프레임워크의 태스크 실행/관리 기능 개발

초저지연 지능형 서비스 PoC 시나리오 및 시스템 설계

지능형 클라우드 엣지 SW 플랫폼을 위한 PoC 서비스 분석 및 설계

- 차량 객체 분석 PoC 시나리오 개발
- 사람 객체 혼잡도 및 이상행동 분석 PoC 시나리오 개발

13 고성능의 AI 엣지 플랫폼이 필요하다면?



인공지능 최적화 솔루션
SOFTonNET AI

신뢰성 있는 DL/ML 플랫폼
지능형 엣지 클라우드

최고의 DL/ML 성능
초저지연 플랫폼

GEdge Platform

초저지연 지능형 엣지 클라우드 플랫폼

감사합니다.

http://gedge-platform.github.io

GEdge Platform 코어 개발자
김성용(sykim@softonnet.com)