



GEdge(Griffin-Edge) Platform

- 초저지연 지능형 클라우드 엣지 SW 플랫폼 -

최적 자원 배치를 위한 글로벌 스케줄링 기술

2023.12.07

GS-Scheduler 프레임워크 코어 개발자

장수민(jsm@etri.re.kr)

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

- New Leap Forward of
GEdge Platform Community 7th Conference (GEdge Platform v4.0 Release) -

GEdge 플랫폼 내 GS-Scheduler의 역할

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)

서비스 협업 프레임워크 (GS-Link)

협업 게이트웨이
(GS-Linkgw)

협업 정책 생성
(GS-Linkhq)

Framework I/F

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



글로벌 스케줄러 개요



글로벌 스케줄러 핵심 기능



글로벌 스케줄러 고도화



향후 계획



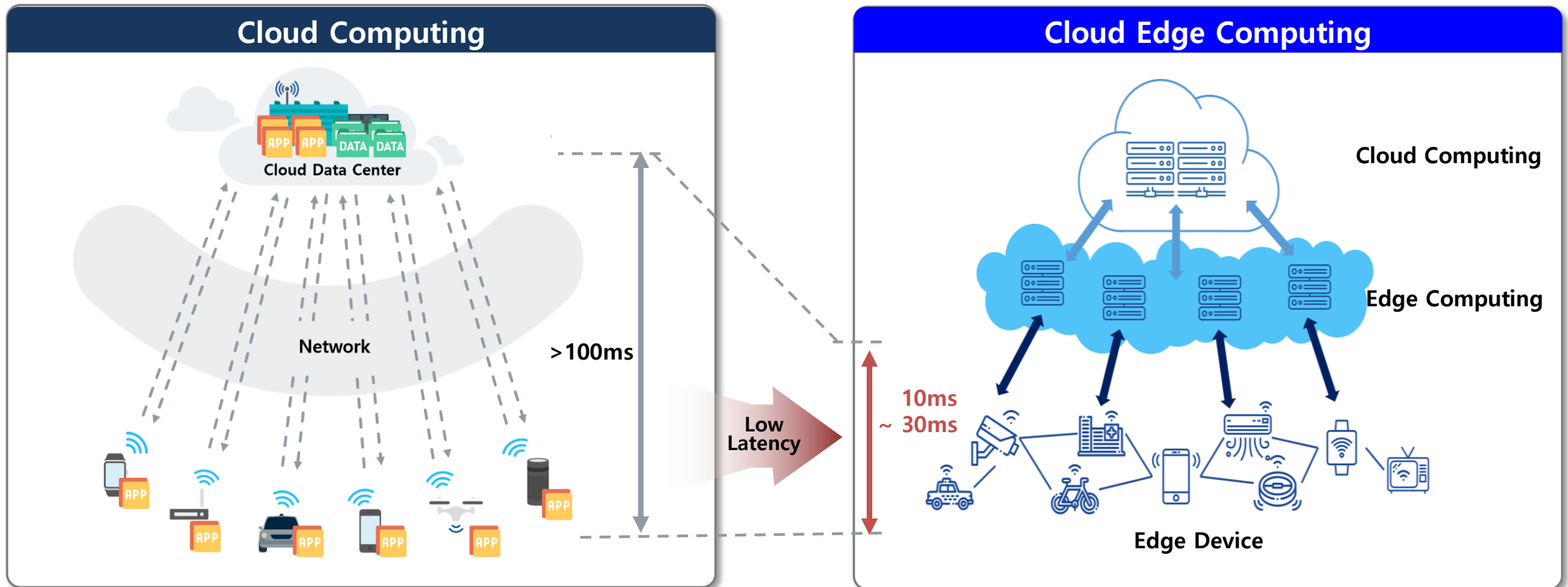
글로벌 스케줄러 개요



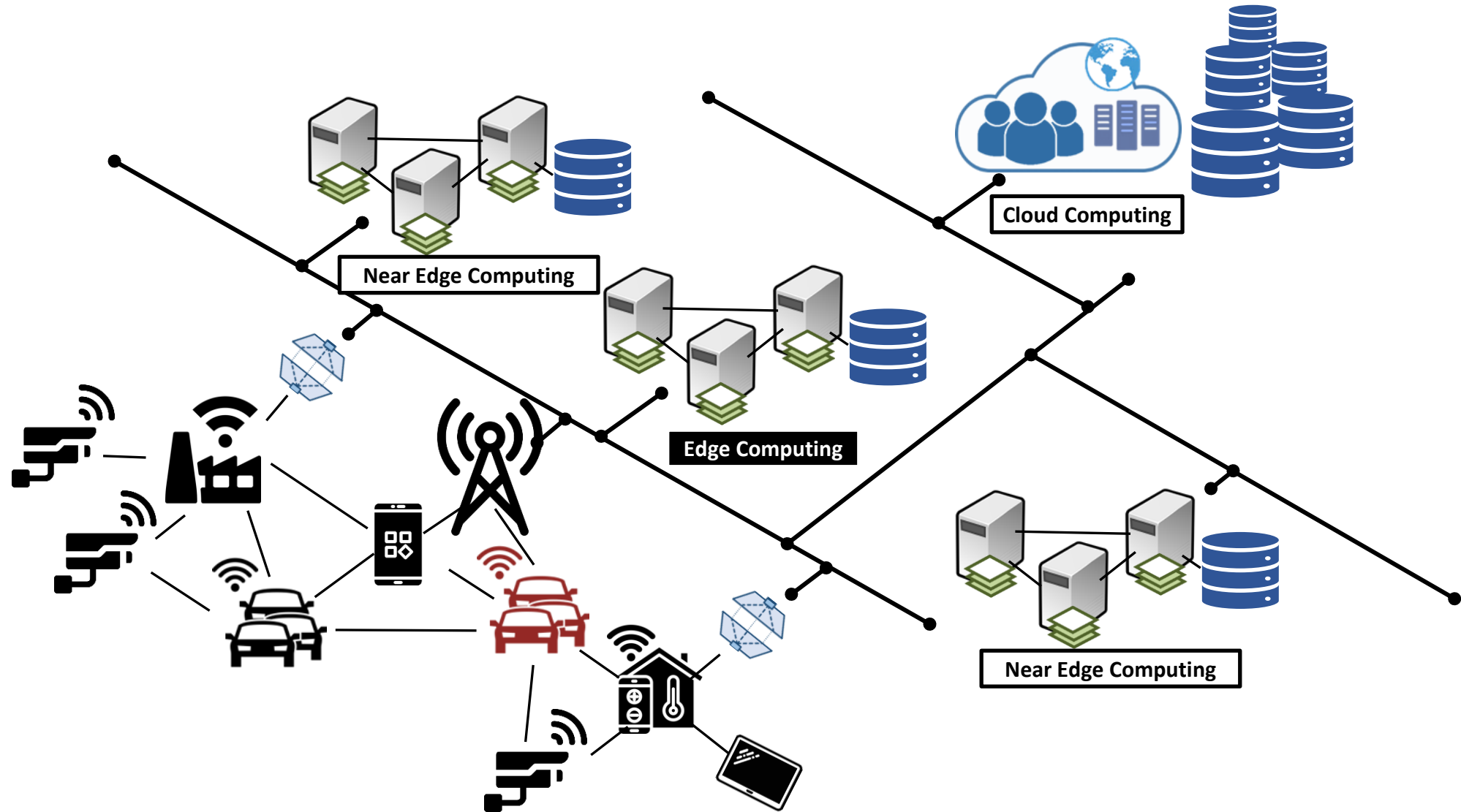
1 GEdge Scheduler 연구의 기술적 배경

클라우드 엣지 컴퓨팅

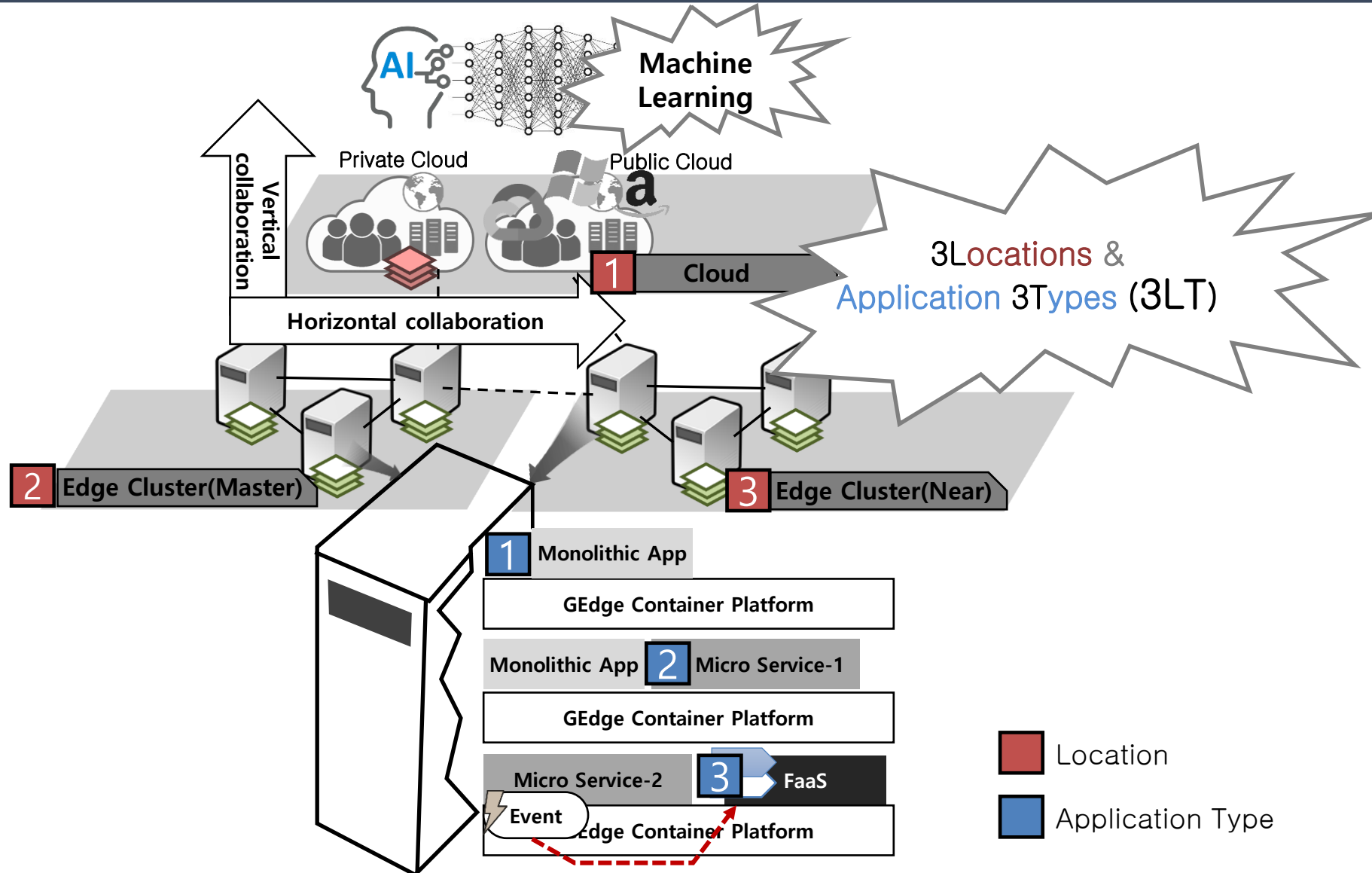
대규모 단말들이 생성하는 데이터를 **단말 근처에서 처리**하여 **불필요한 데이터 전송을 최소화**하고 엣지 서비스의 **응답 속도를 보장**하는 분산 클라우드 형태의 컴퓨팅 방식

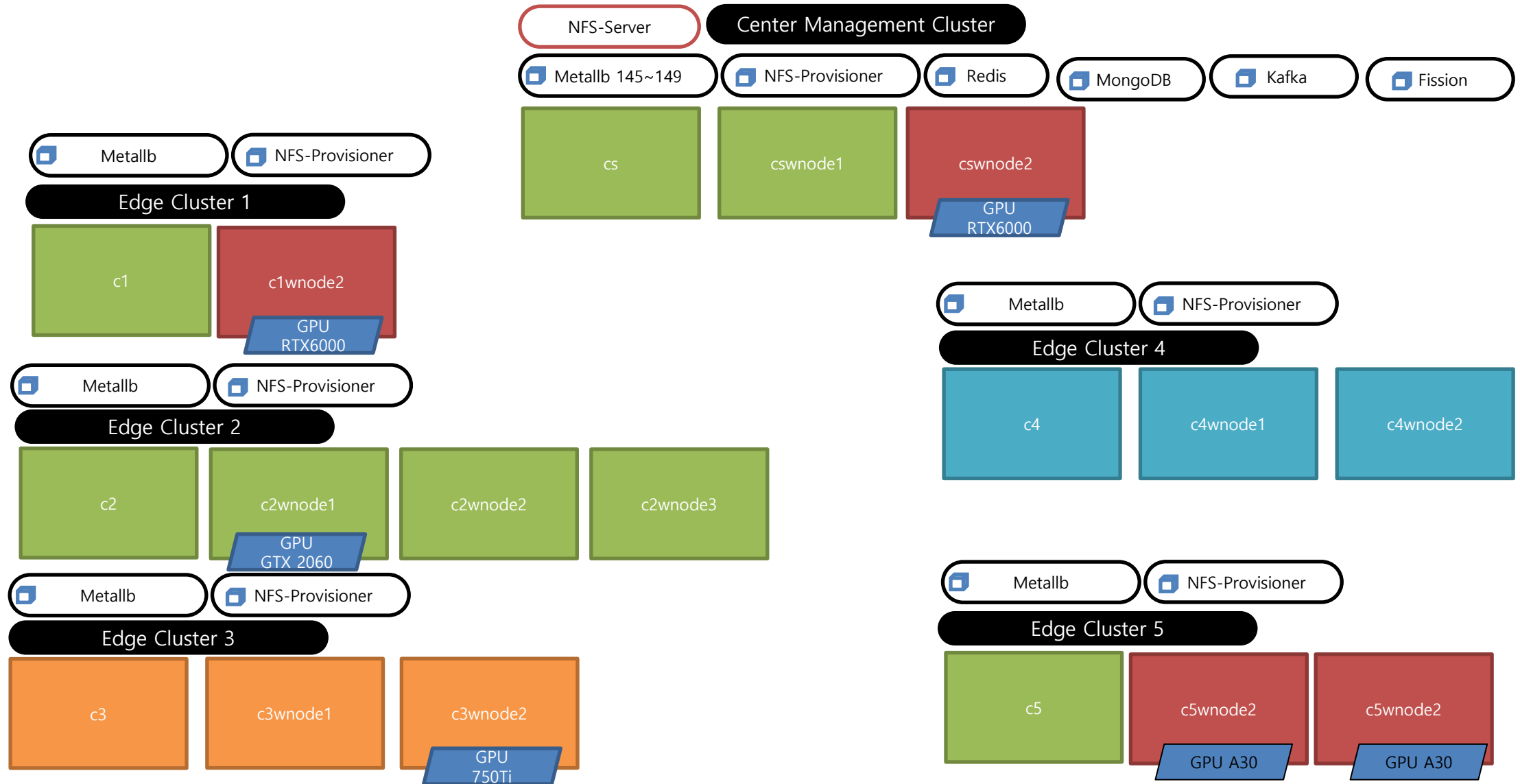


2 GEdge Scheduler 연구의 기술적 배경



3 GEdge-Platform System : 3LT 개념

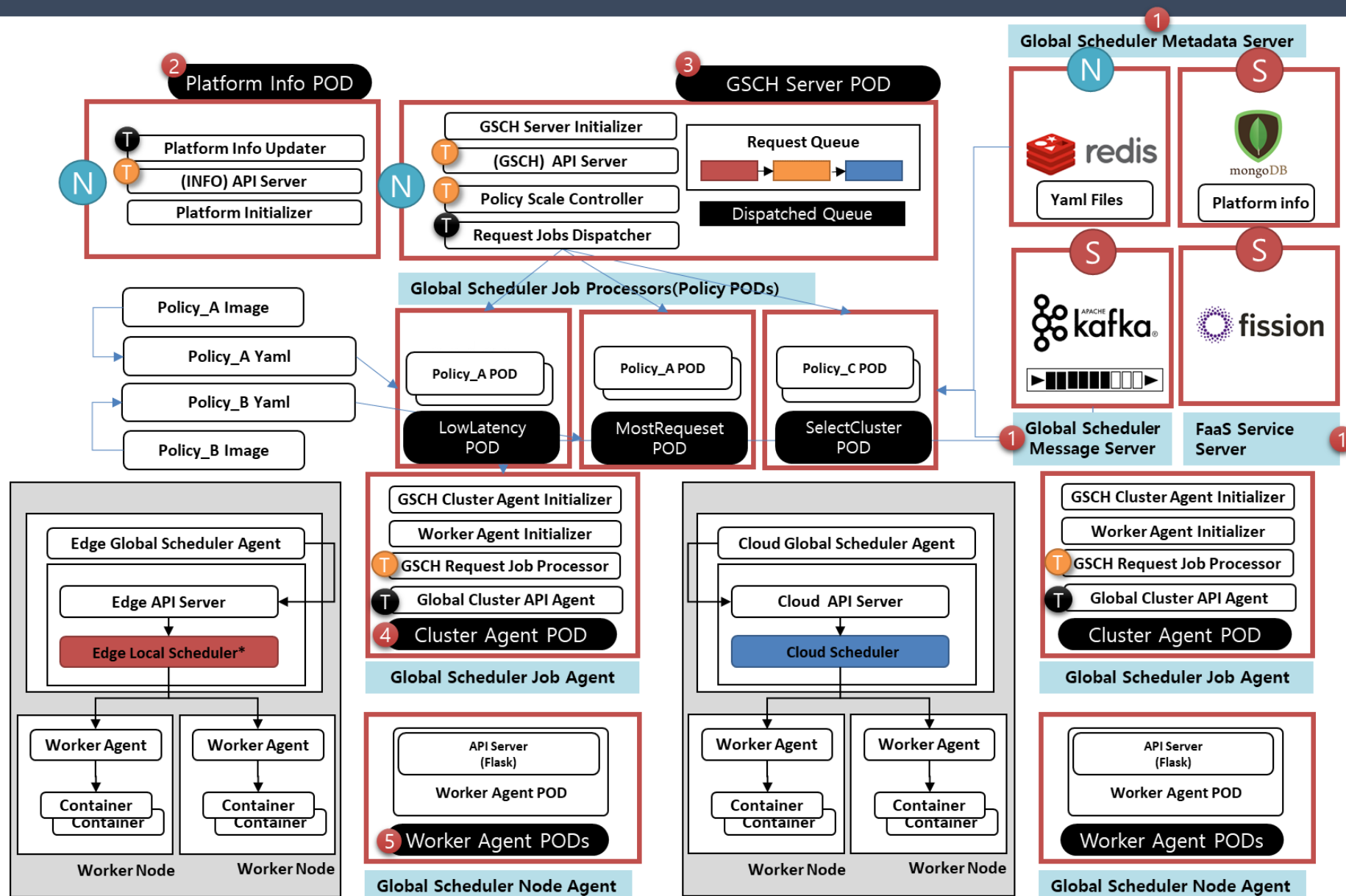


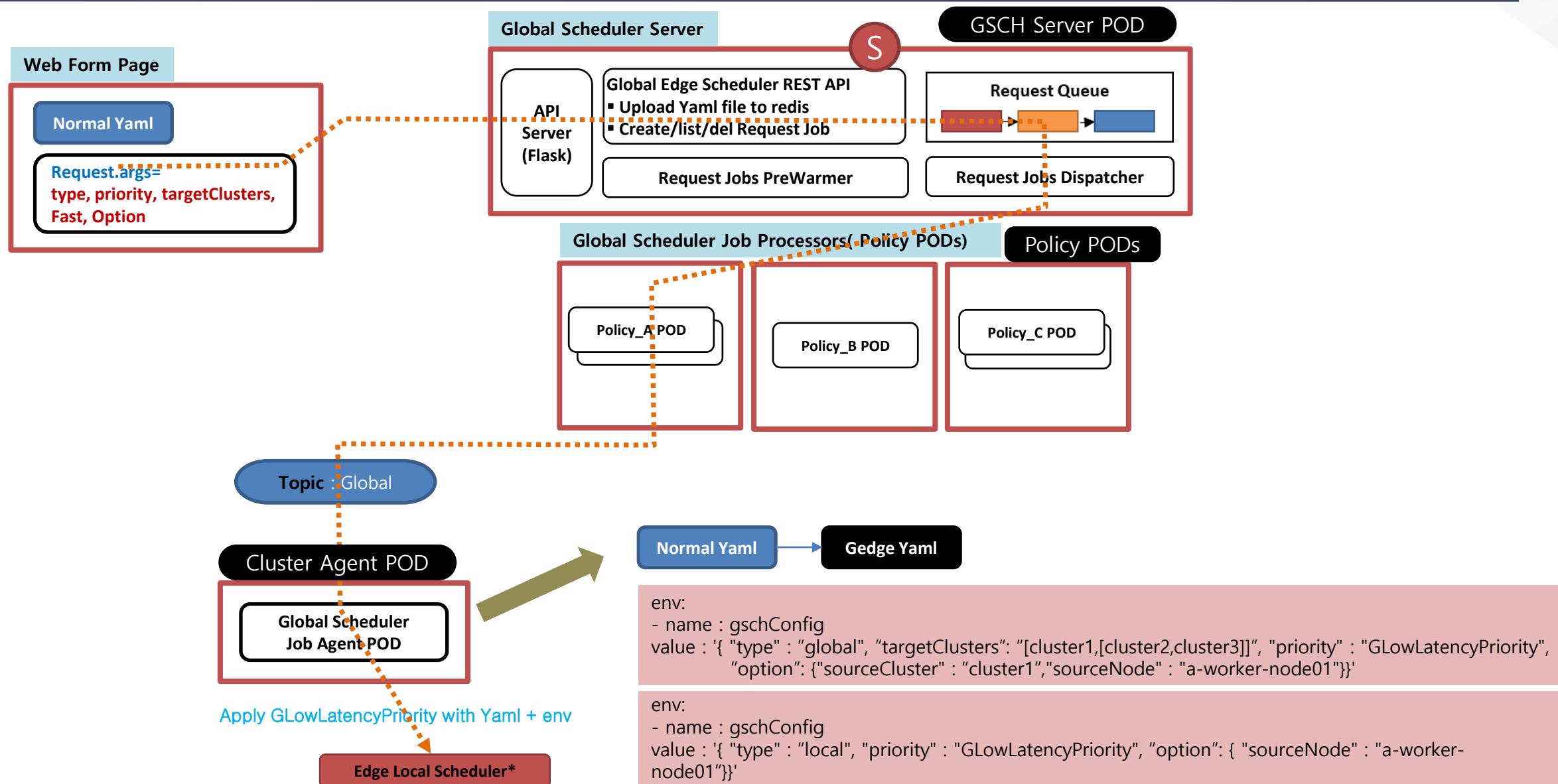




글로벌 스케줄러 핵심 기능

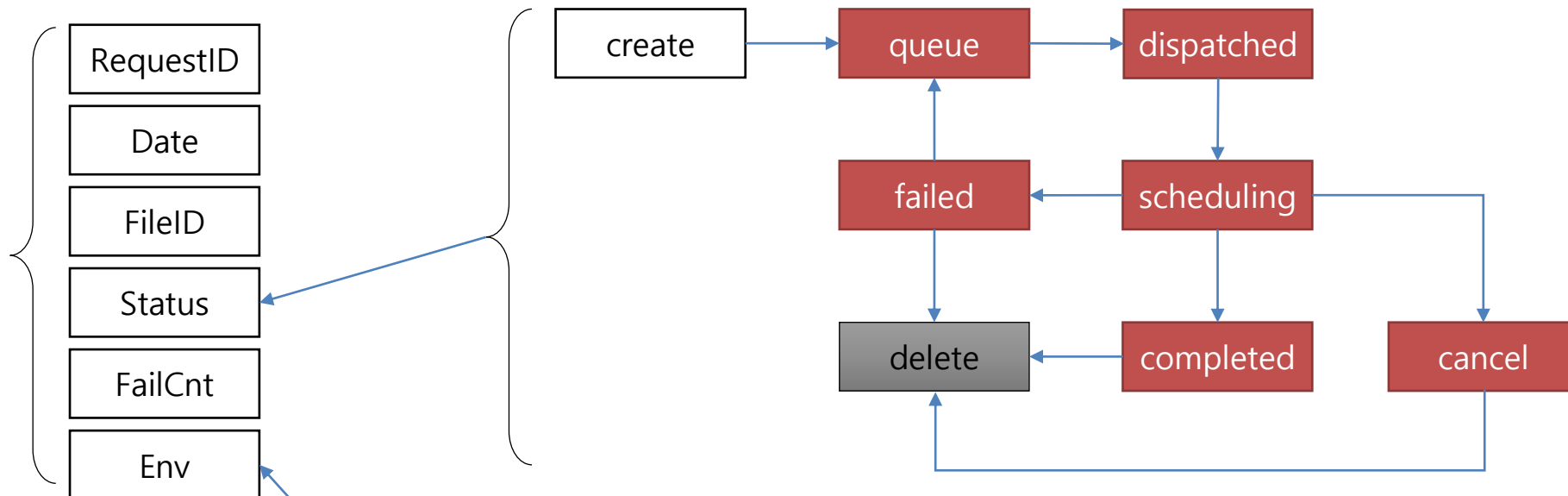






lifecycle

Request Job

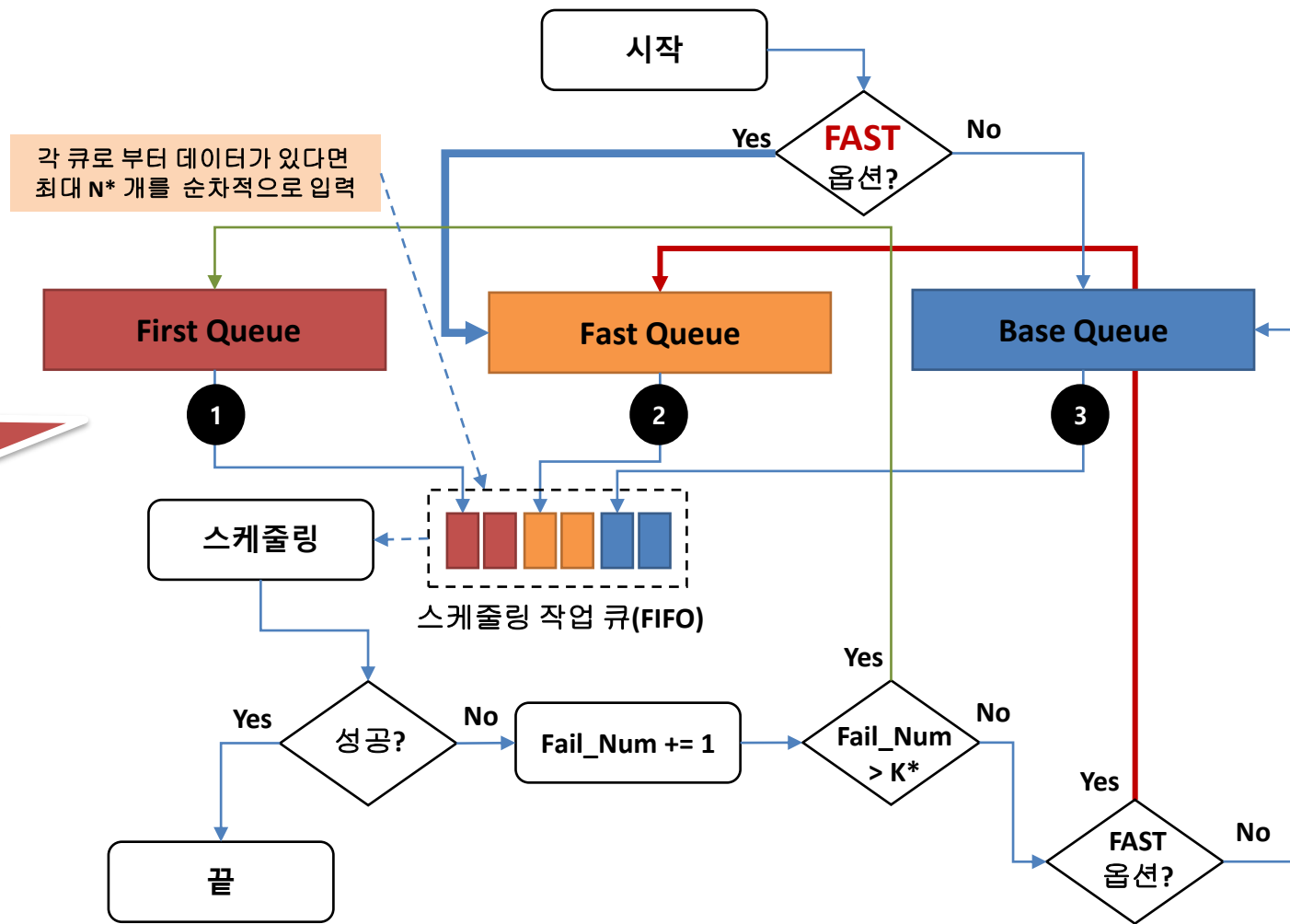


gschConfig Data

```
env:  
- name : gschConfig  
value : '{ "type" : "global", "select_clusters": "[cluster1,[cluster2,cluster3]]", "priority" : "GLowLatencyPriority",  
          "option": {"source_cluster" : "cluster1","source_node" : "a-worker-node01"} }'
```

```
env:  
- name : gschConfig  
value : '{ "type" : "local", "priority" : "GLowLatencyPriority", "option": { "source_node" : "a-worker-node01"} }'
```

Low-Latency
특화처리



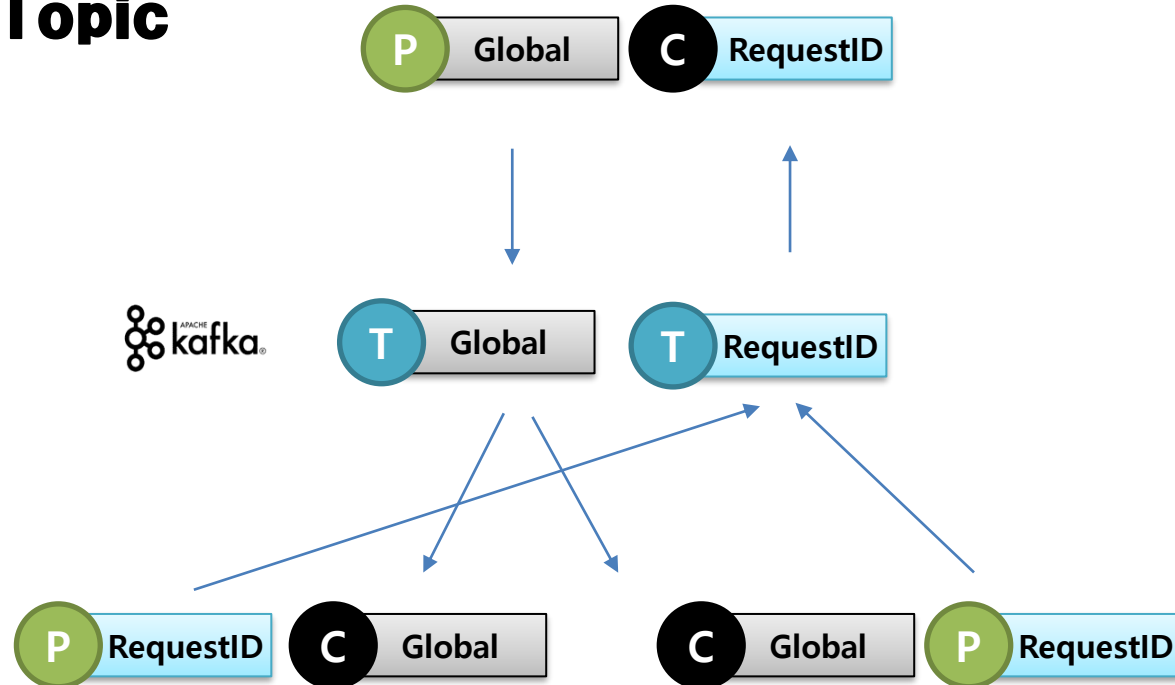
N*, K* Values are defined by System

P **Producer**

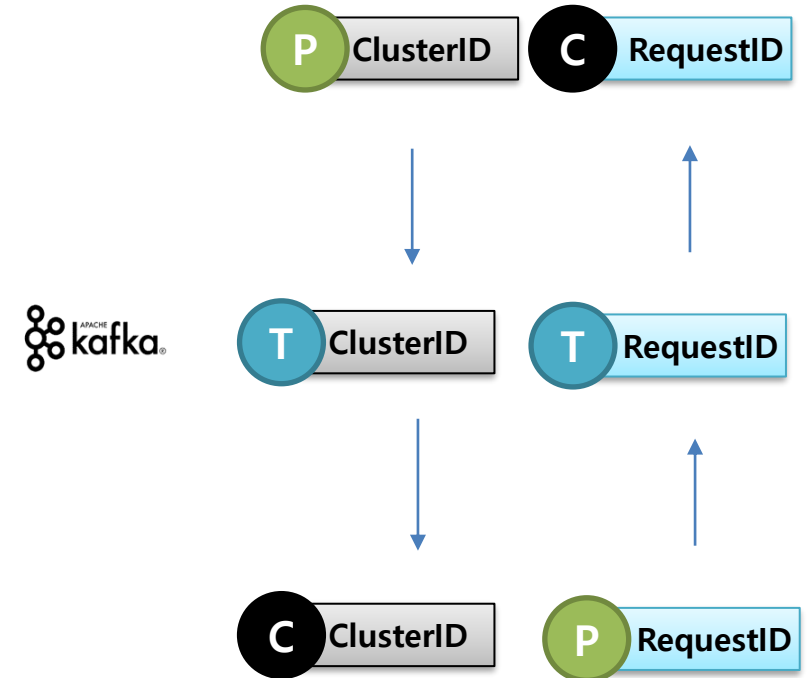
C **Consumer**

T **Topic**

1:N

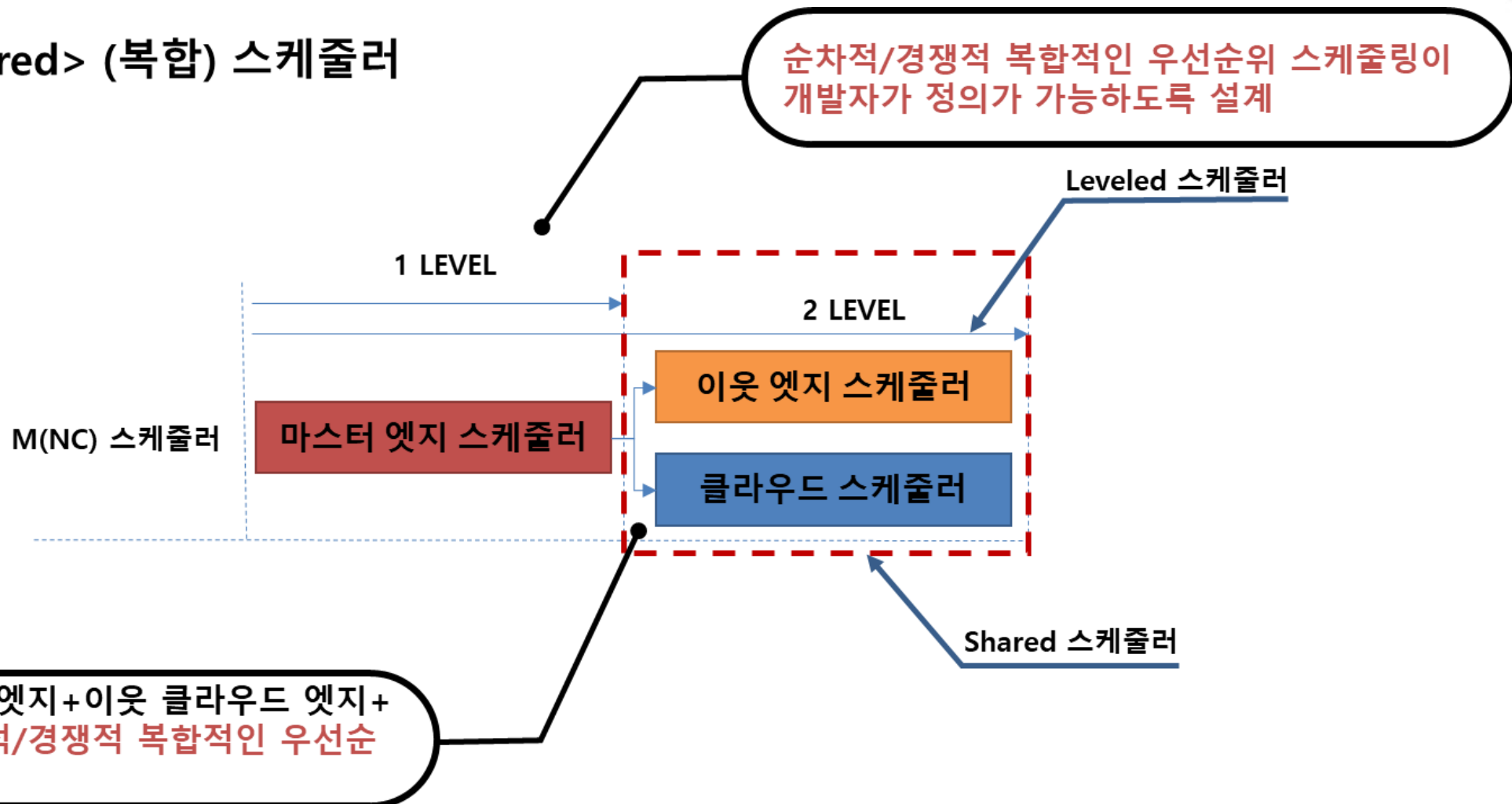


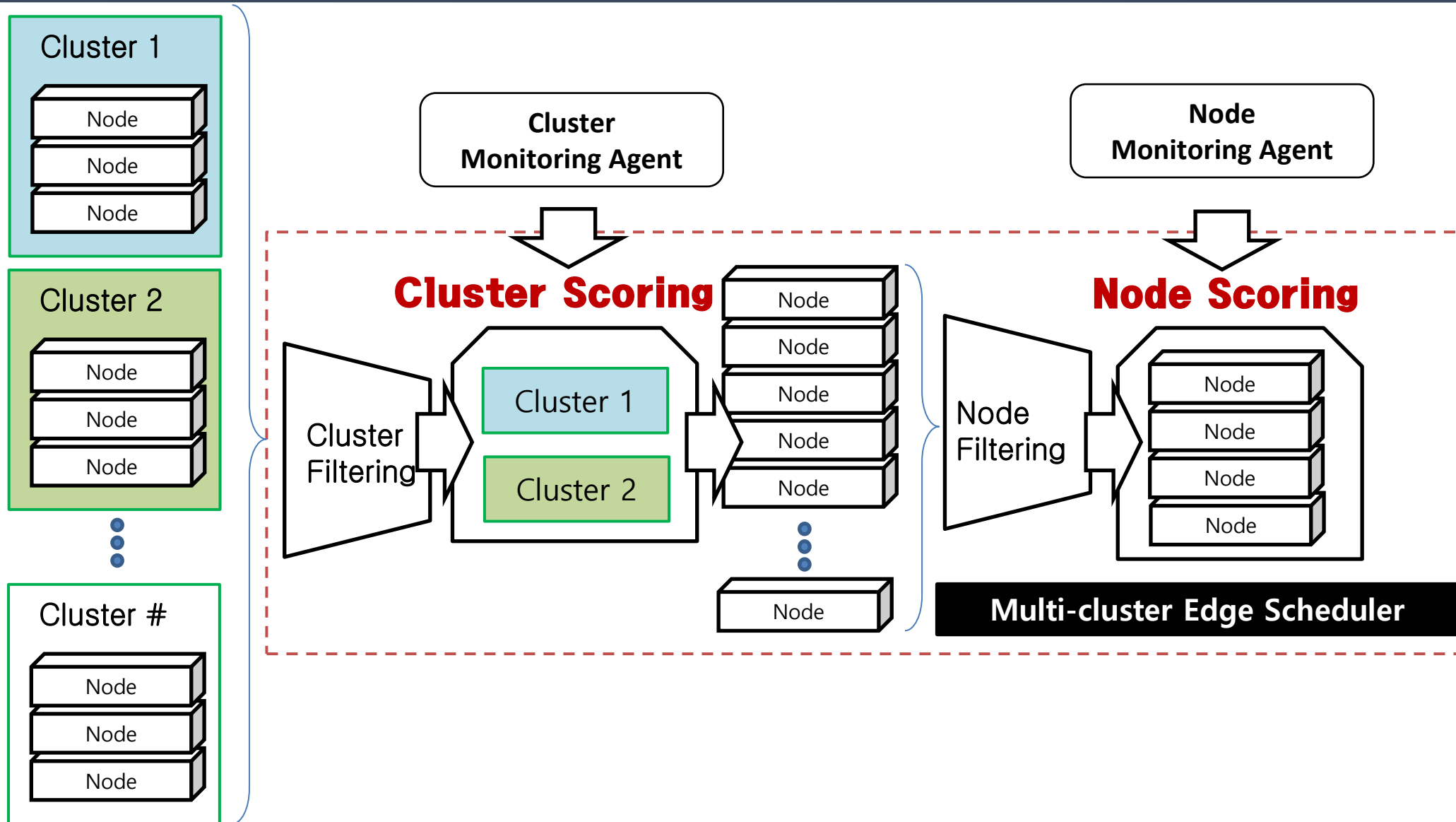
1:1



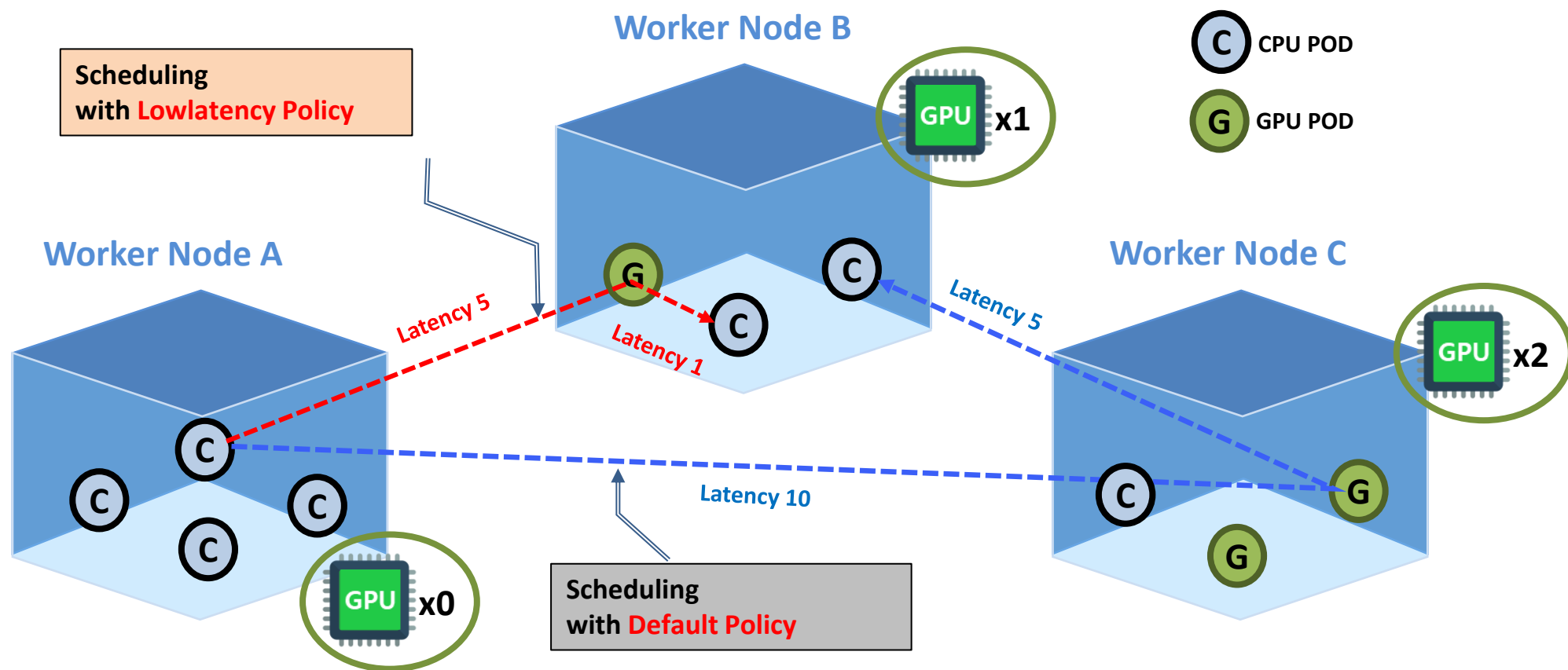
Source	Type	cluster/node/pod
	Object	"c1" {"cluster":"c1","node":"n1"} {"cluster":"c1","pod":"p1"}
Target	Type	cluster(s)/node(s)/pod(s)
	Object	c1/[c1,c2] [{"cluster":"c1","node(s)":n1/["n1"]}] [{"cluster":"c1","pod(s)":p1/["p1"]}]]
HCode		0x0001
LCode		0x0001
Msg		env: - name : gschConfig value : { "type" : "global", "targetClusters": "[cluster1,[cluster2,cluster3]]", "priority" : "GLowLatencyPriority" }

<Leveled+Shared> (복합) 스케줄러

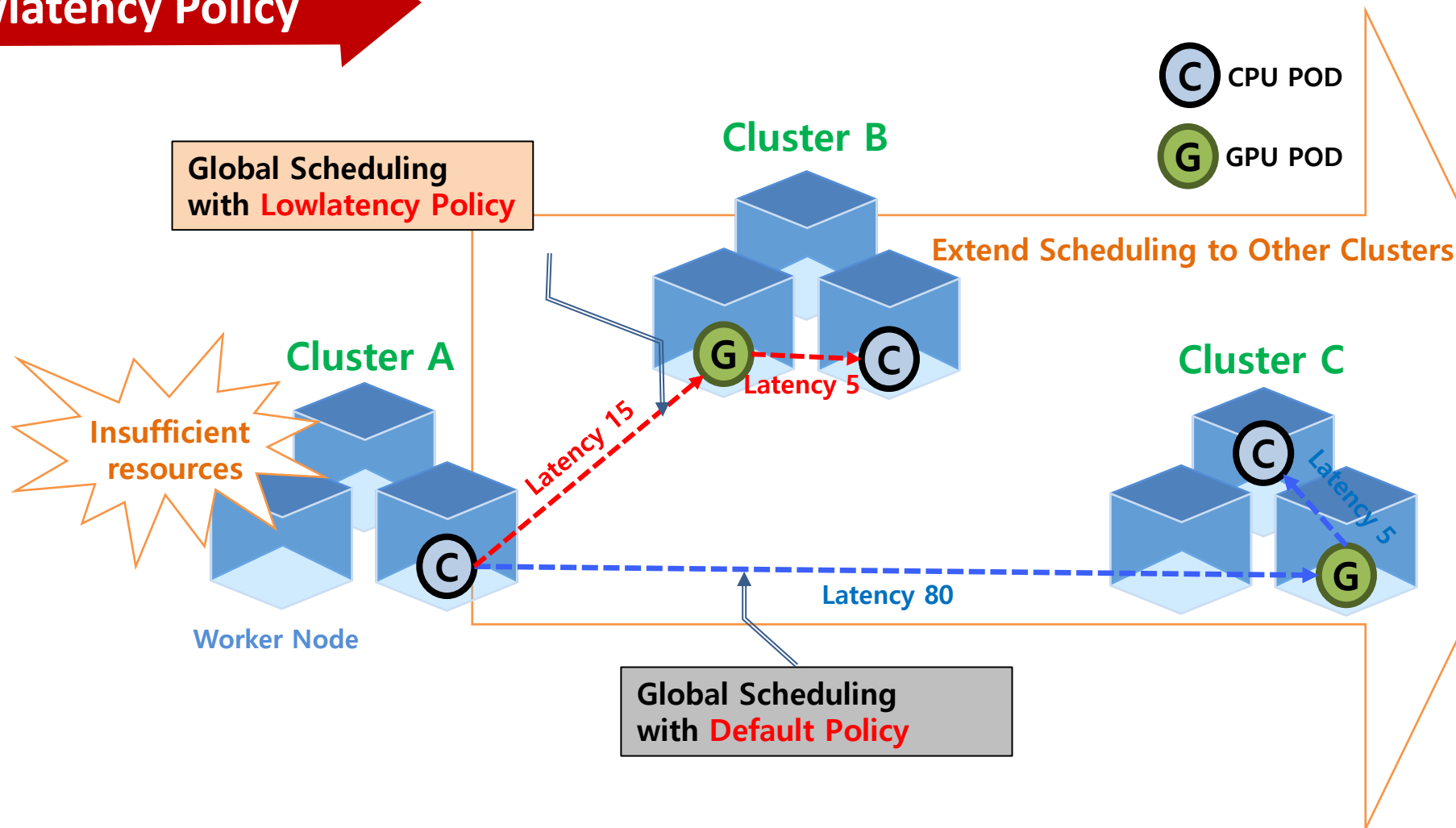




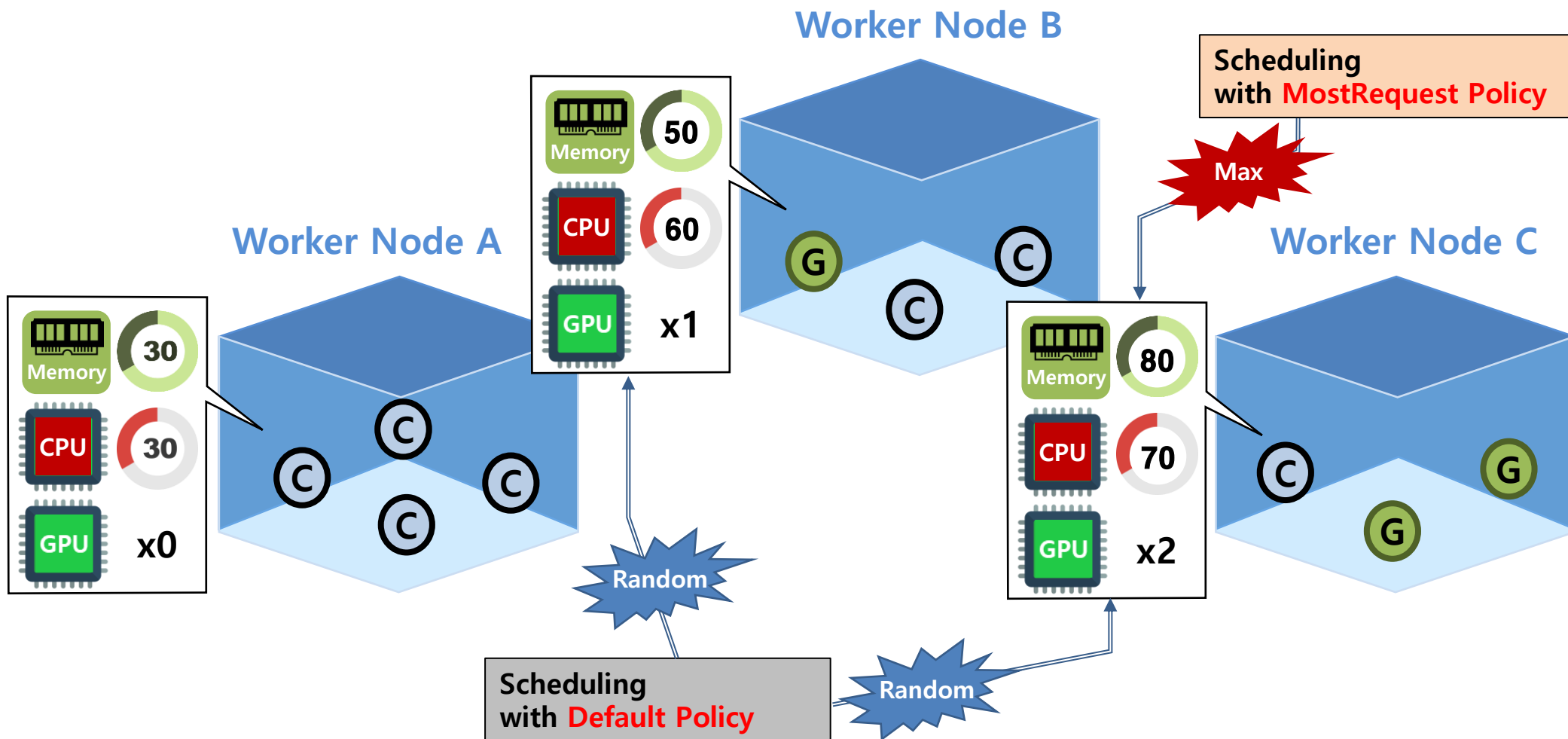
Lowlatency Policy



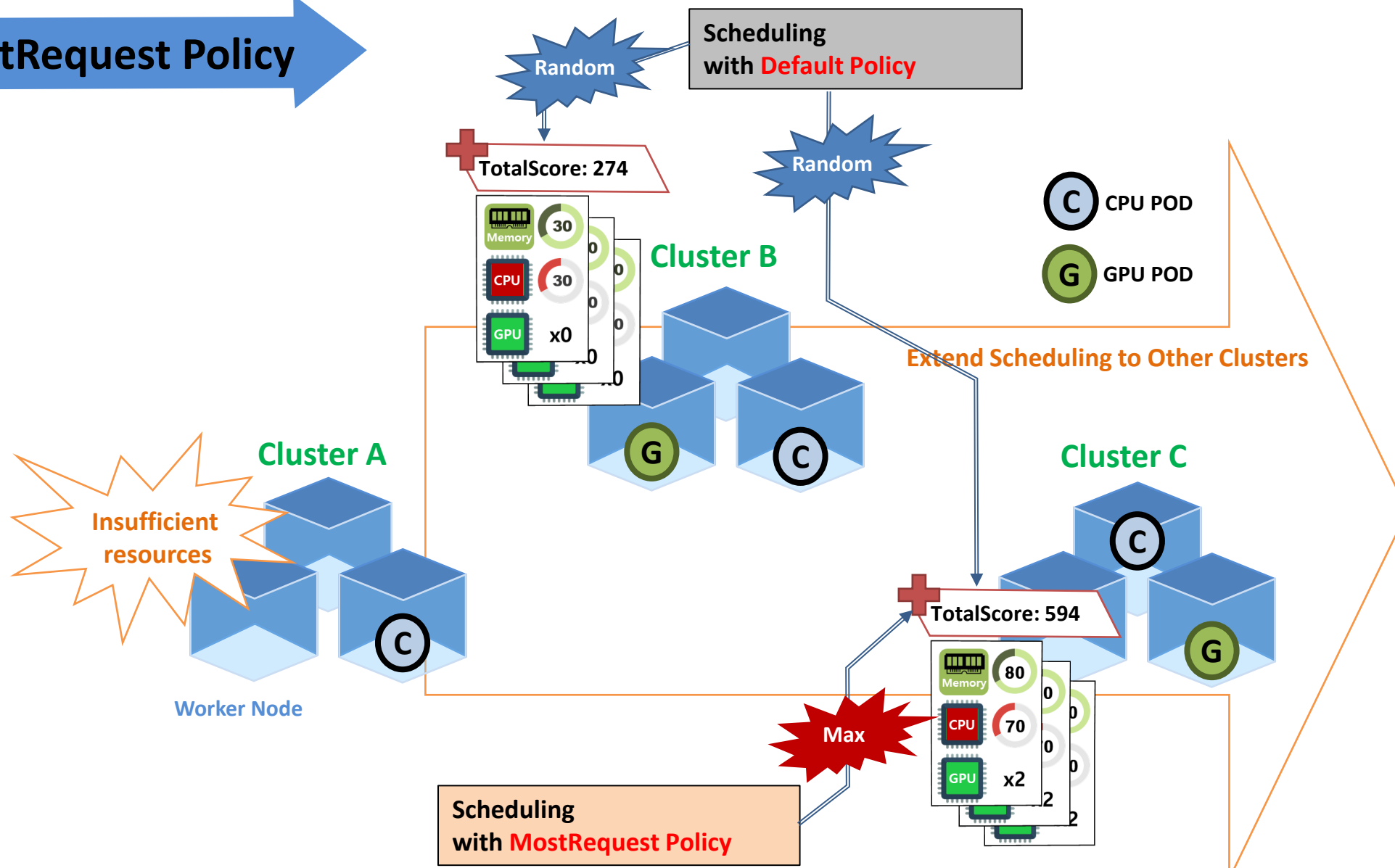
Lowlatency Policy



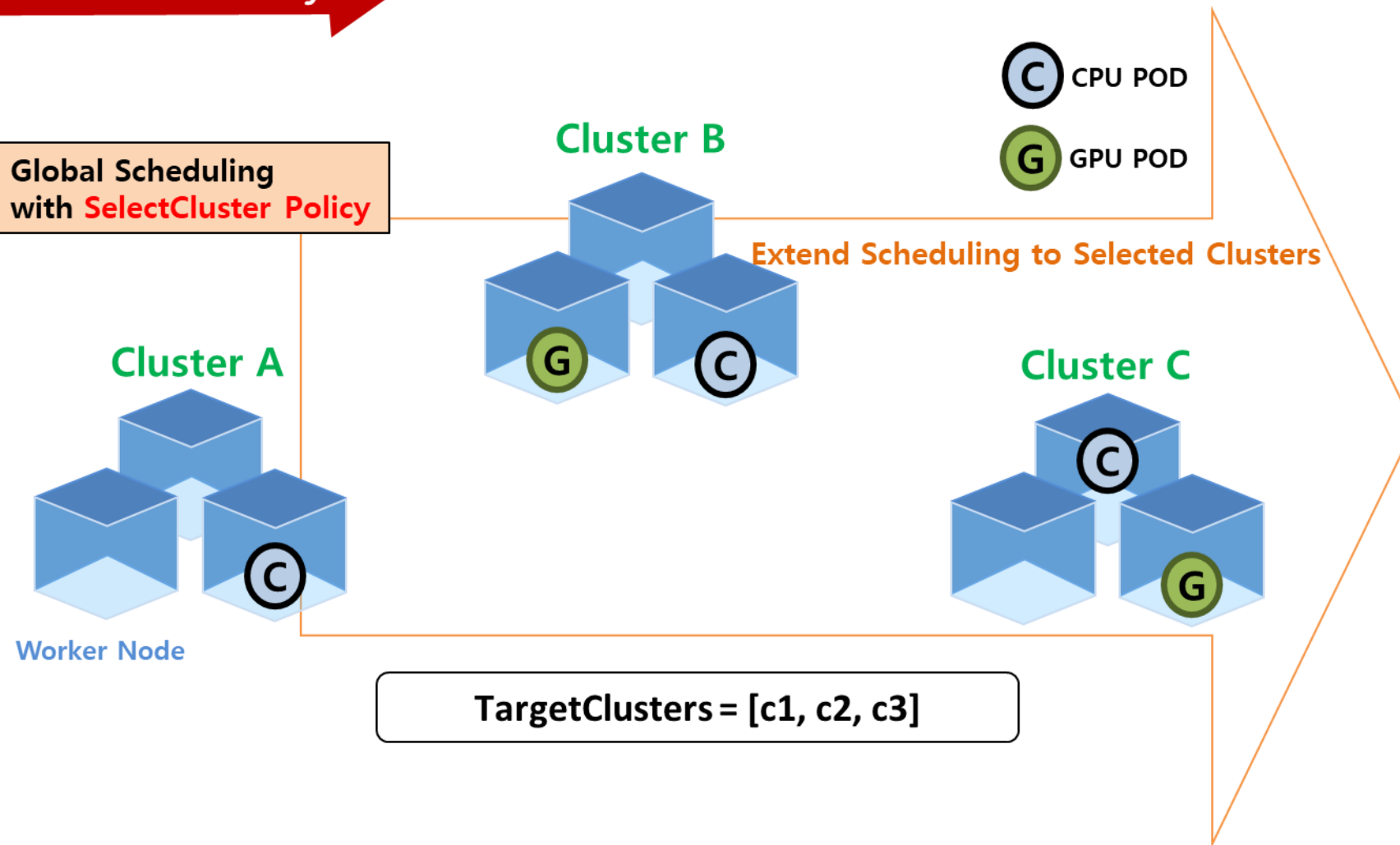
MostRequest Policy



MostRequest Policy



SelectClusters Policy



마이크로 서비스 배포정책의 성능평가

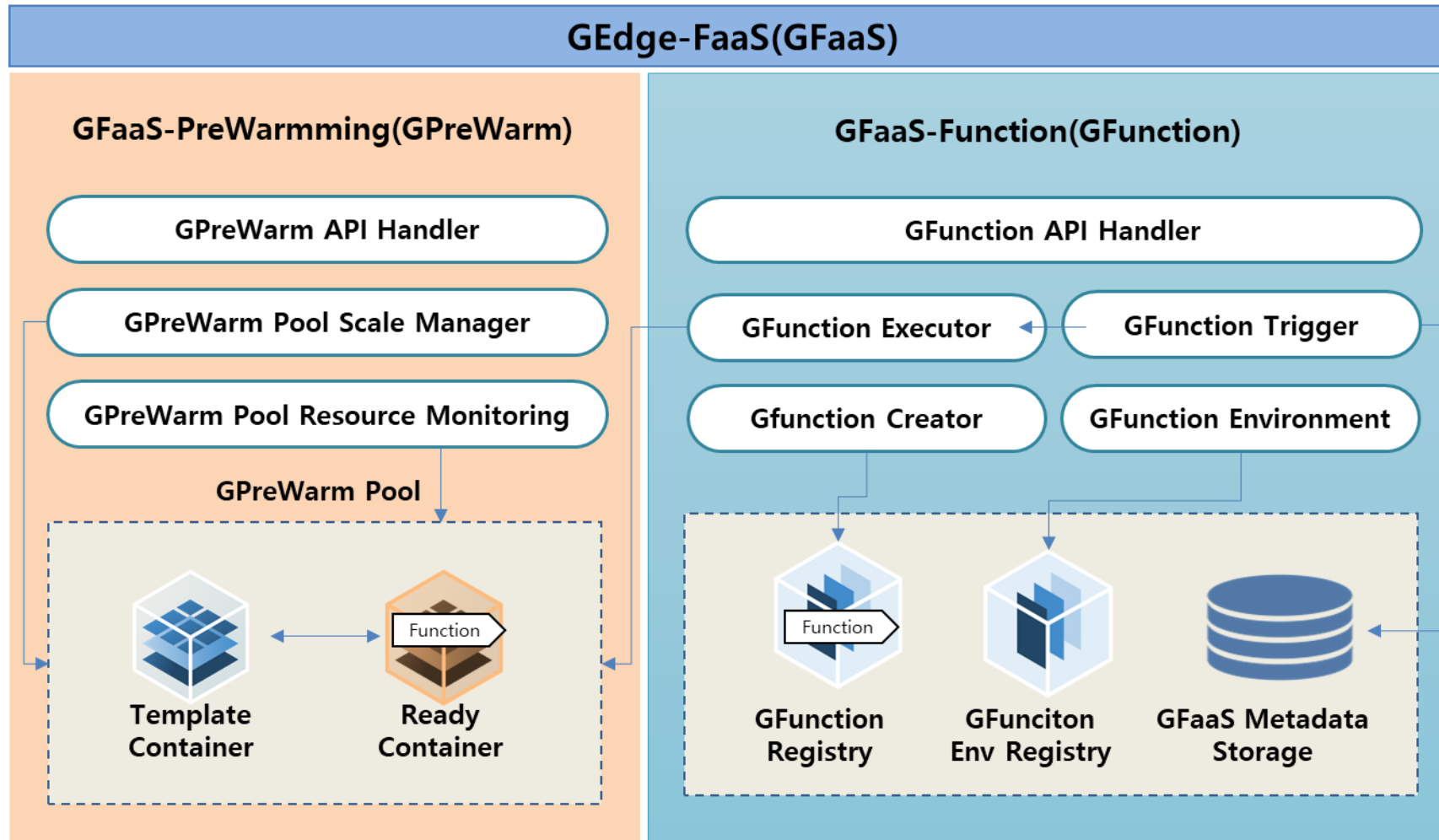
지능형 객체 인식 서비스 Yolo4로 검증

- ◆ 제안하는 마이크로 서비스 배포 정책의 성능 평가
 - 워커 노드가 4개, 2개인 클러스터 2개를 구성
 - 3개의 마이크로 서비스들로(선처리, 추론, 후처리) 구성된 지능형 객체 인식 서비스인 Yolo4 서비스를 자체 개발하여 평가
 - Yolo4 서비스의 주요 처리과정은 샘플 동영상 (30Mbyte)을 입력 값으로 받아서 추론을 통하여 객체인식 내용이 추가된 결과 동영상을 생성하는 방식
- ◆ 주요 테스트 서버 환경
 - Xeon 3.0Ghz, Memory 1.1TB, GPU V100가 장착된 워커 노드
 - 워커 노드별 네트워크 지연시간의 차등을 주기 위하여 네트워크 트래픽 부하 발생시켜서 특정 마스터 노드로부터 워커 노드 4개(약 5ms), 2개(약 1ms)되도록 설정
- ◆ 성능평가 결과
 - 제안하는 Low-Latency 부분이 적용하지 않은 것에 비하여 약 20% 정도의 처리 속도가 향상
 - 노드들 간에 네트워크 지연시간 차이가 클수록 본 논문에서 제안하는 서비스 배포 정책이 더 의미가 있음

Deployment Policy Performance Comparison
with Yolo Microservice



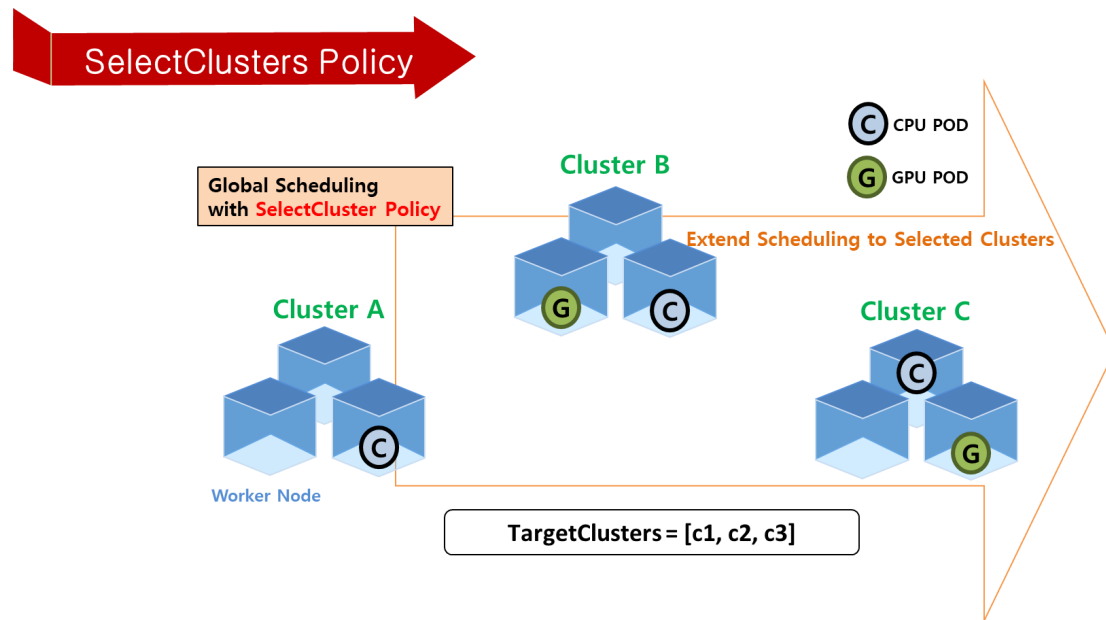
20% 처리속도 향상





글로벌 스케줄러 핵심 고도화





GSelectedCluster

- * 사용자가 현재 선택한 project에 속한 cluster 범위에서 선택
- * 단일 리스트만 가능
- * Level은 다수 개 설정될 수 있음
- * 하나의 Level에는 1개의 클러스터만 선택 가능
- * 순차적으로 할당을 시도하는 것으로 하나가 성공하면 다음 단계는 생략

사용자가 현재 선택한 project에 속한 cluster 범위에서 선택 하나의 클러스터만 가능

options

JSON

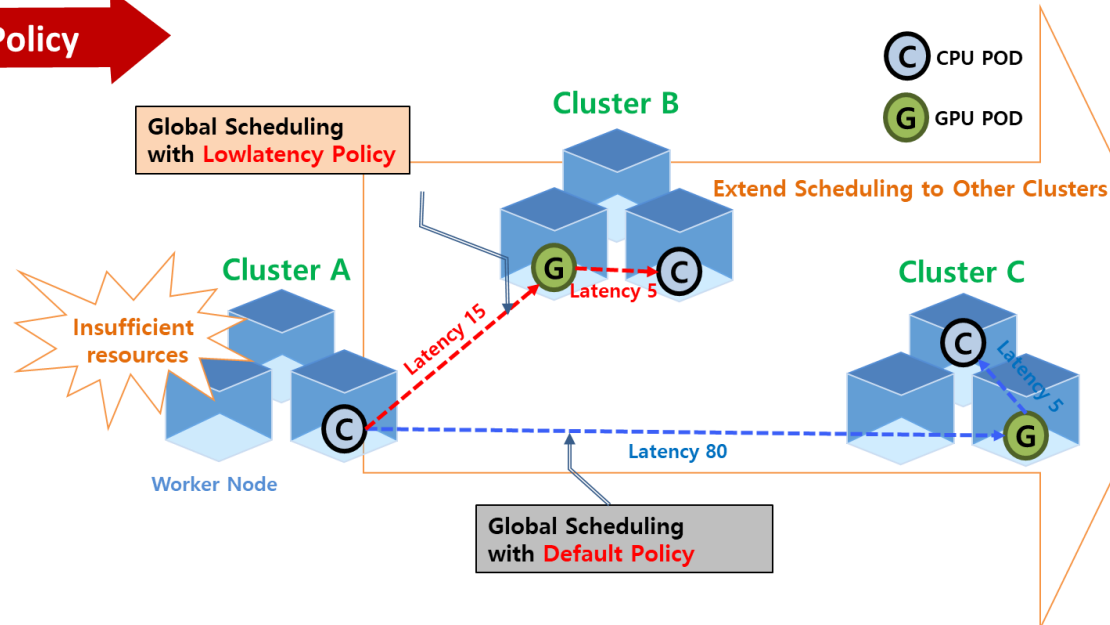
```
{
  user_name : user1,
  workspace_name: w1,
  project_name: p1,
  mode : cluster(default),
  parameters : {
    select_clusters: ['c1','c2']
  }
}
```

options

JSON

```
{
  user_name : user1,
  workspace_name: w1,
  project_name: p1,
  'mode' : 'node',
  'parameters' : {
    'select_cluster': 'c1',
    'select_node': 'n1'
  }
}
```

Lowlatency Policy



GLowLatencyPriority

options

JSON

```
{
  user_name : user1,
  workspace_name: w1,
  project_name: p1,
  mode : fromnode(default)
  parameters : {
    source_cluster: c1,
    source_node: node1,
    select_clusters: [['c1','c2'], ['c3','c4'], 'c5']
  }
}
```

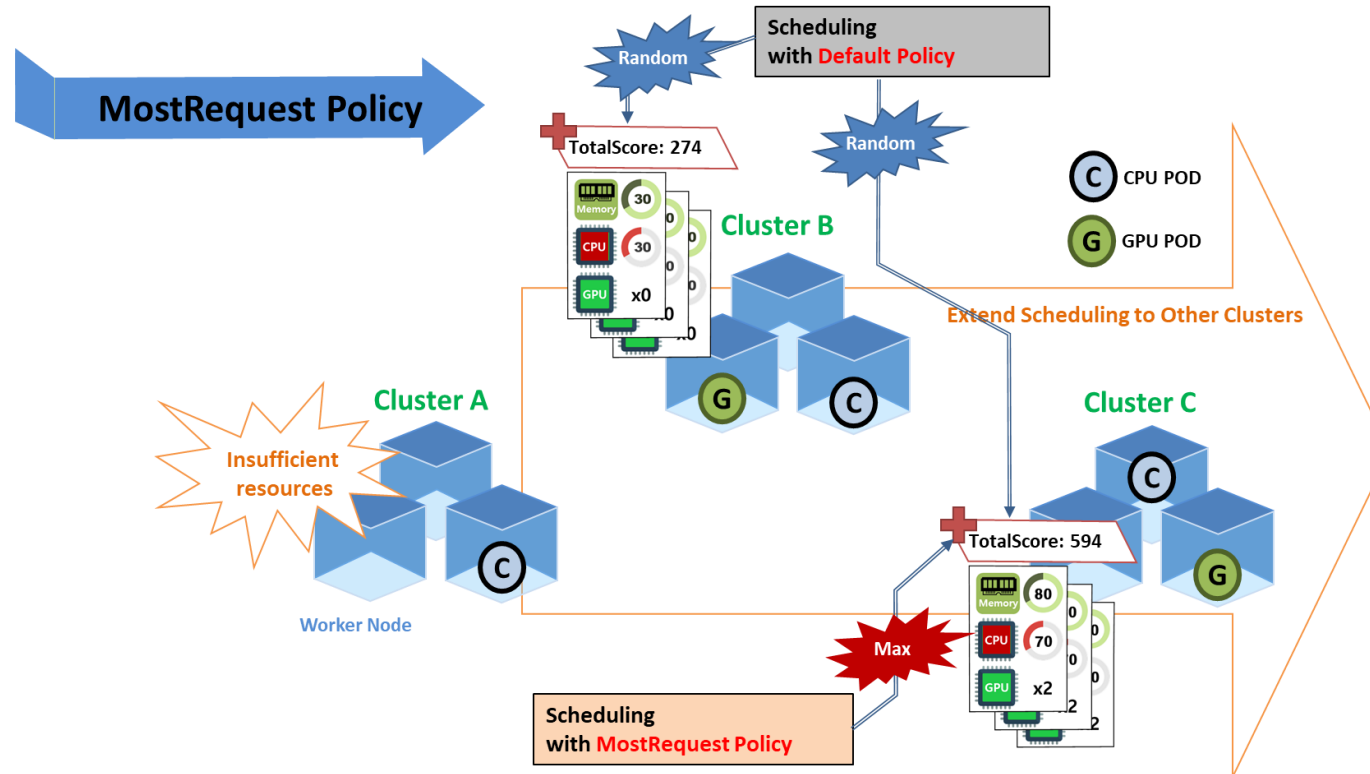
사용자가 현재 선택한 project에
속한 cluster 범위에서 선택
이중리스트(리스트안에 리스트) 가능

options

JSON

```
{
  user_name : user1,
  workspace_name: w1,
  project_name: p1,
  mode : frompod
  parameters : {
    source_cluster: c1,
    pod_name: p1,
    select_clusters: [['c1','c2'], ['c3','c4'], 'c5']
  }
}
```

사용자가 현재 선택한 project에
속한 cluster 범위에서 선택
이중리스트(리스트안에 리스트) 가능



GMostRequestPriority

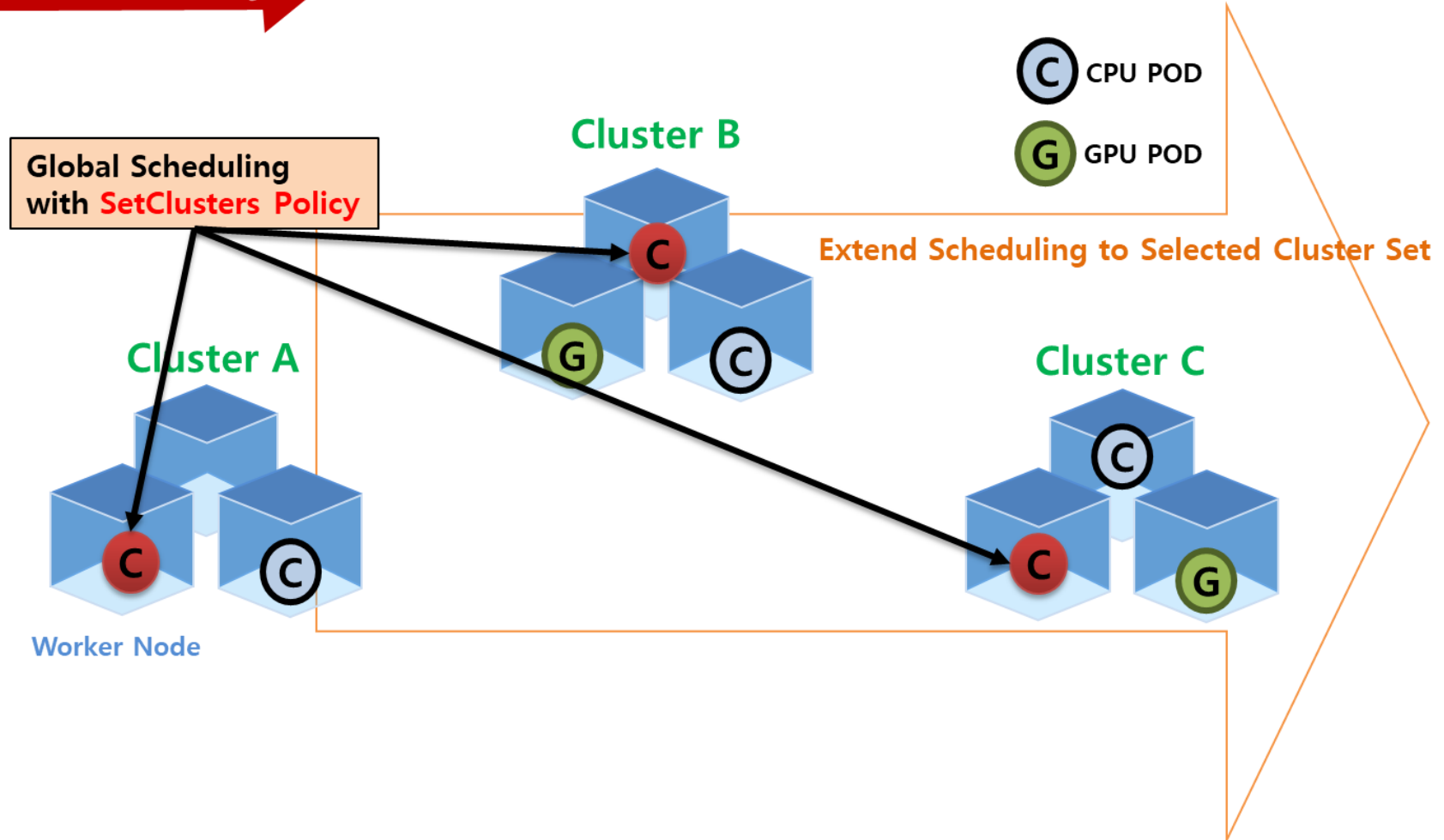
- * 사용자가 현재 선택한 project에 속한 cluster 범위에서 선택
- * 이중리스트(리스트안에 리스트)가능
- * 이중리스트는 서로 경쟁을 통하여 우선순위가 높은 클러스터부터 처리

options

JSON

```
{
  user_name : user1,
  workspace_name: w1,
  project_name: p1,
  mode : default/gpu/cpu/memory
  parameters : {
    select_clusters: [['c1','c2'],['c3','c4'],'c5']
  }
}
```

SetClusters Policy



IV

향후 계획



1 GS-Scheduler 안정화 및 성능 개선

GS-Scheduler 안정화

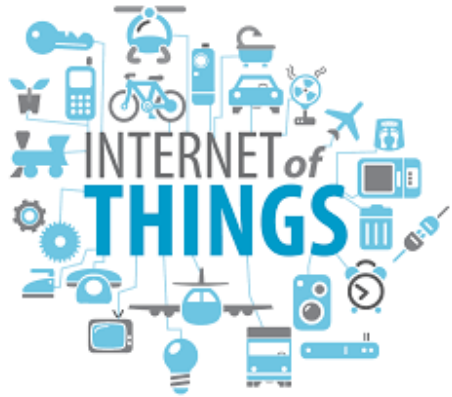
- * GS-Scheduler 고도화로 개발된 핵심 모듈 기능 테스트 및 안정화 작업
- * 다양한 서비스 응용을 적용하여 안정화에 주력

GS-Scheduler 성능 극대화

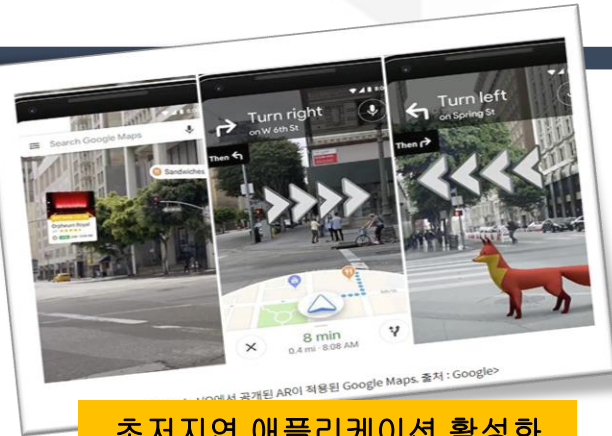
- * 다양한 성능 평가를 통한 핵심 모듈의 최적화 작업을 통하여 GS-Scheduler 성능 극대화
- * 주요 엣지 모의 테스트 응용 선별 및 개발 응용을 통한 성능 개선 작업

GS-Scheduler 활용영역 확대

- * 로봇 응용 서비스를 위한 엣지 서버 적용, 다양한 IOT 응용 분야의 게이트 서버 적용
- * 엣지 서비스 관련 업체의 리즈 분석 및 활용 영역 확대



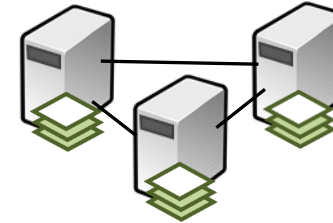
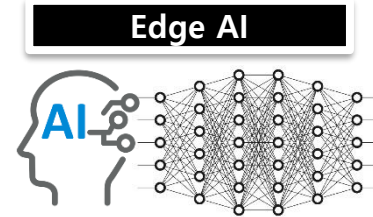
IoT



초저지연 애플리케이션 활성화



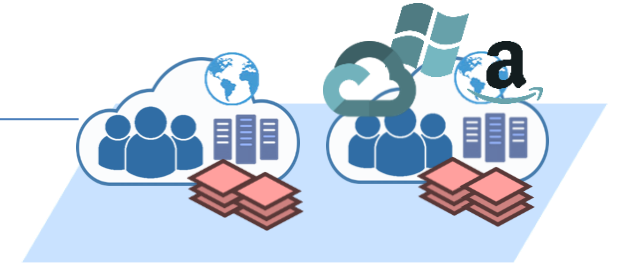
5G Network



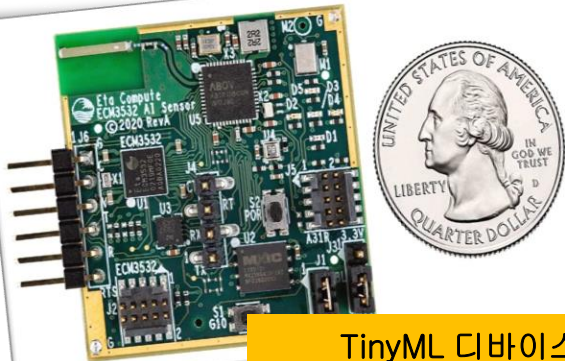
Edge Platform



클라우드에서 엣지로 AI 역할 이동



Cloud



TinyML 디바이스 확대



다양한 형태의 에지 플랫폼 활성화

감사합니다.

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



GS-Scheduler 프레임워크 코어 개발자
장수민(jsm@etri.re.kr)

Welcome to GEdge Platform

An Open Cloud Edge SW Platform to enable Intelligent Edge Service

GEdge Platform will lead Cloud-Edge Collaboration