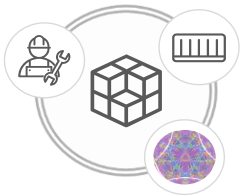
신한은행

Open Source를 활용한 금융권 On-Premise AI Infrastructure의 효율적 GPU 사용

신한은행 디지털혁신단
권오균 프로

Content



1

ML systems와
데이터 리소스 소개

AI서비스
도입 절차

2

3

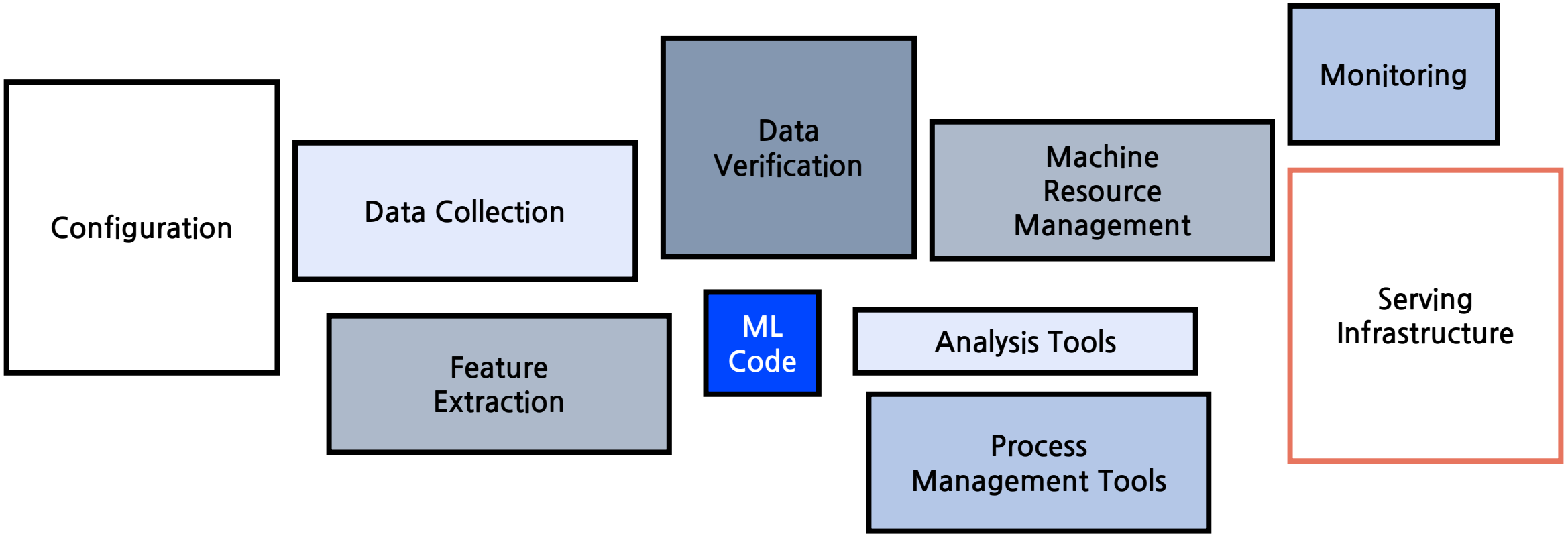
효율적인
MIG GPU
적용 사례

Hybrid Architecture
플랫폼

4

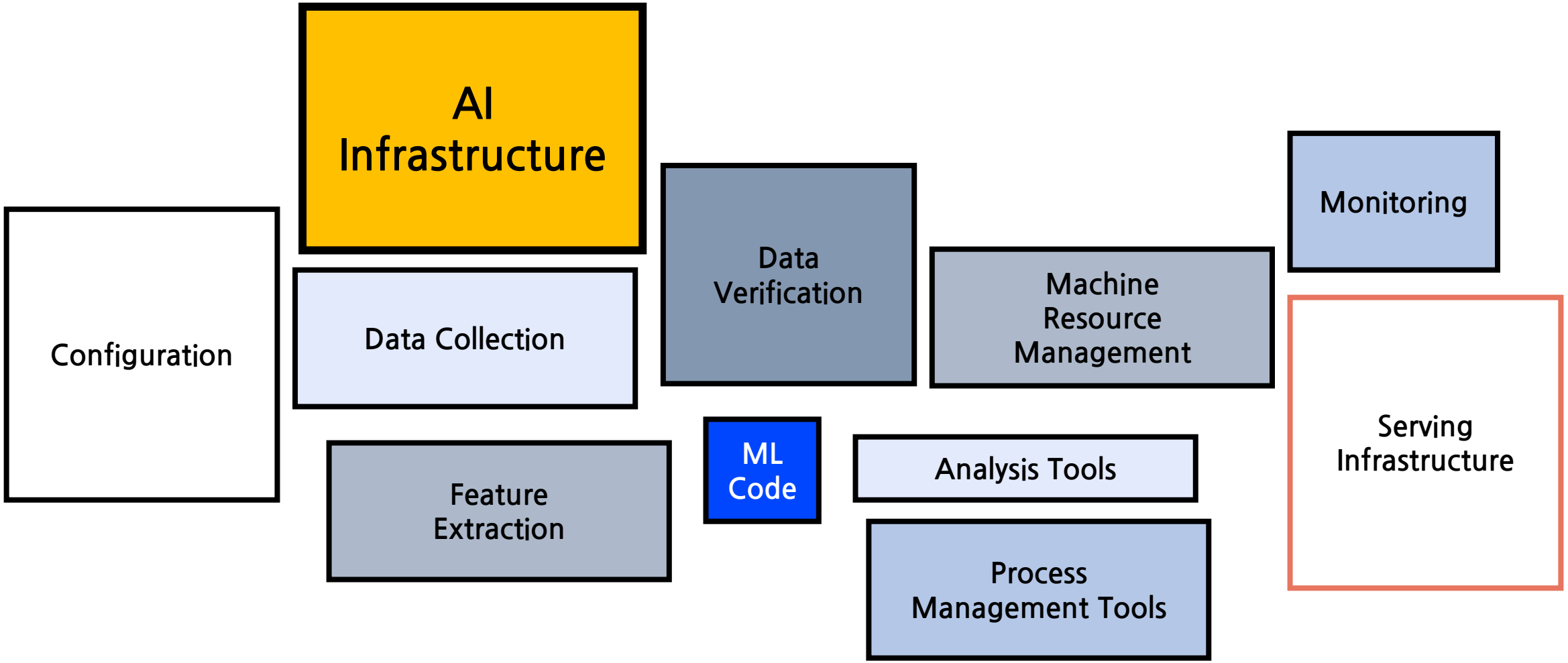
ML systems와 데이터 리소스 소개

1) Real-world ML systems

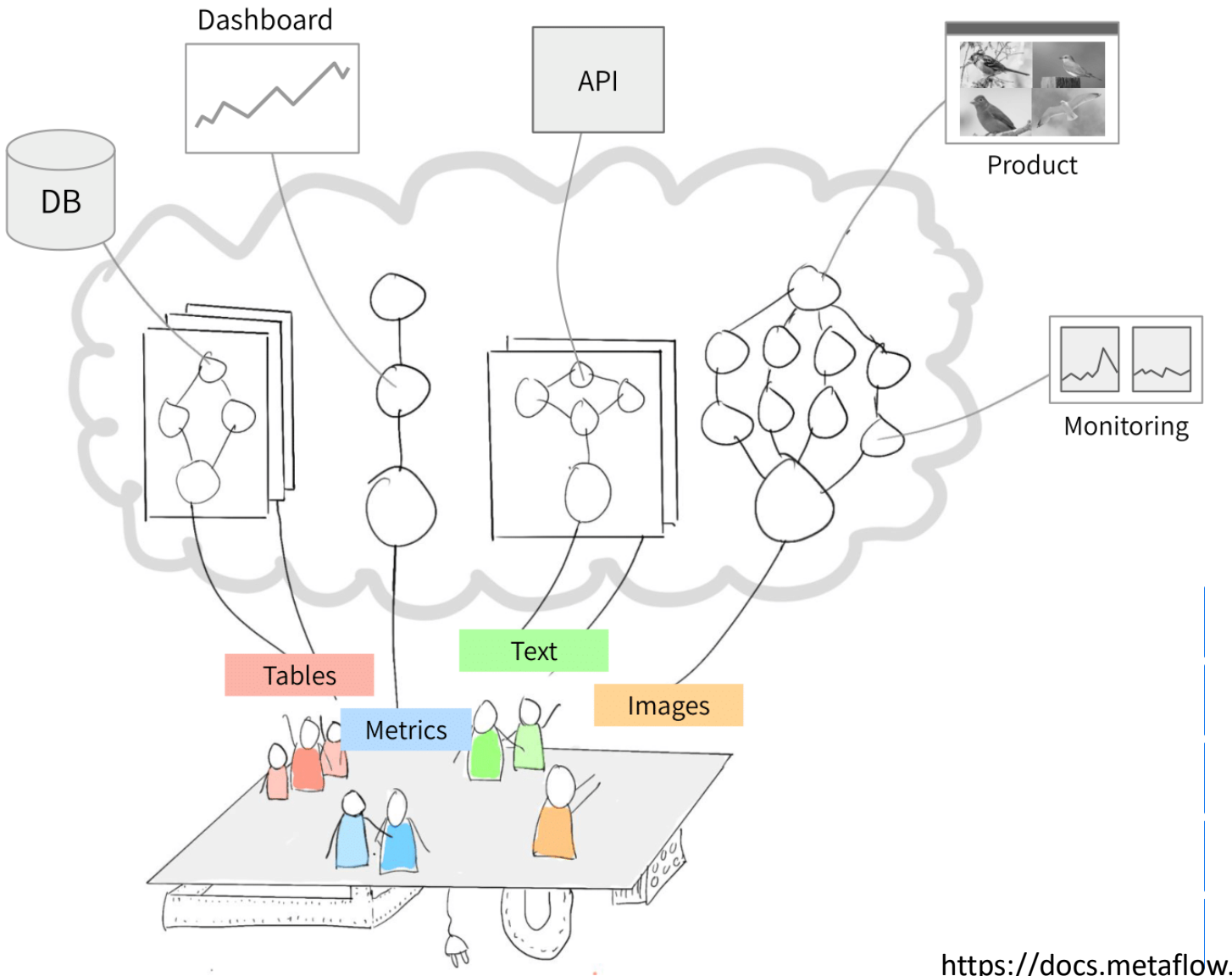


https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

1) Real-world ML systems



2) 다양한 채널의 데이터 리소스



https://docs.metaflow.org/introduction/why-metaflow

2 AI서비스 도입 절차

1) AI 프로젝트 도입 비용

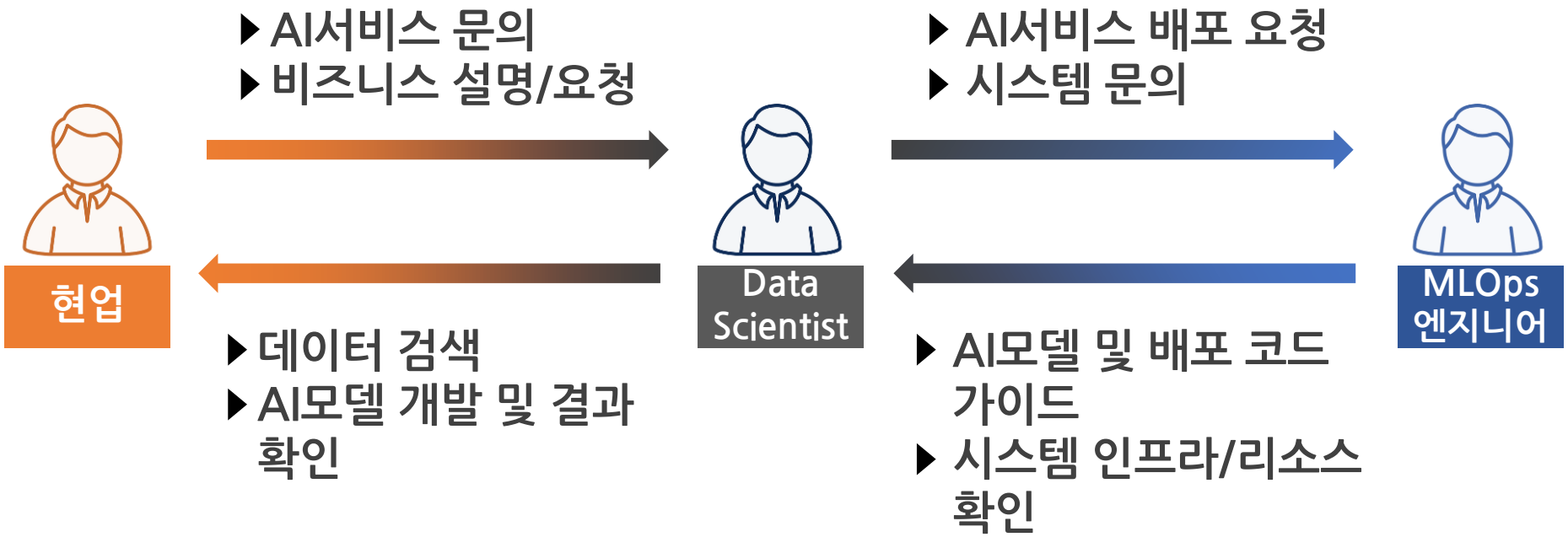
High → Low

AI Infra › 데이터 가공 › 인건비

[참고] GPU 도입 비용

2) AI 서비스 런칭 프로세스

" [현업, DS, MLOps엔지니어] 커뮤니케이션을 통해 GPU 도입 가능여부 검토 "



3) GPU 도입

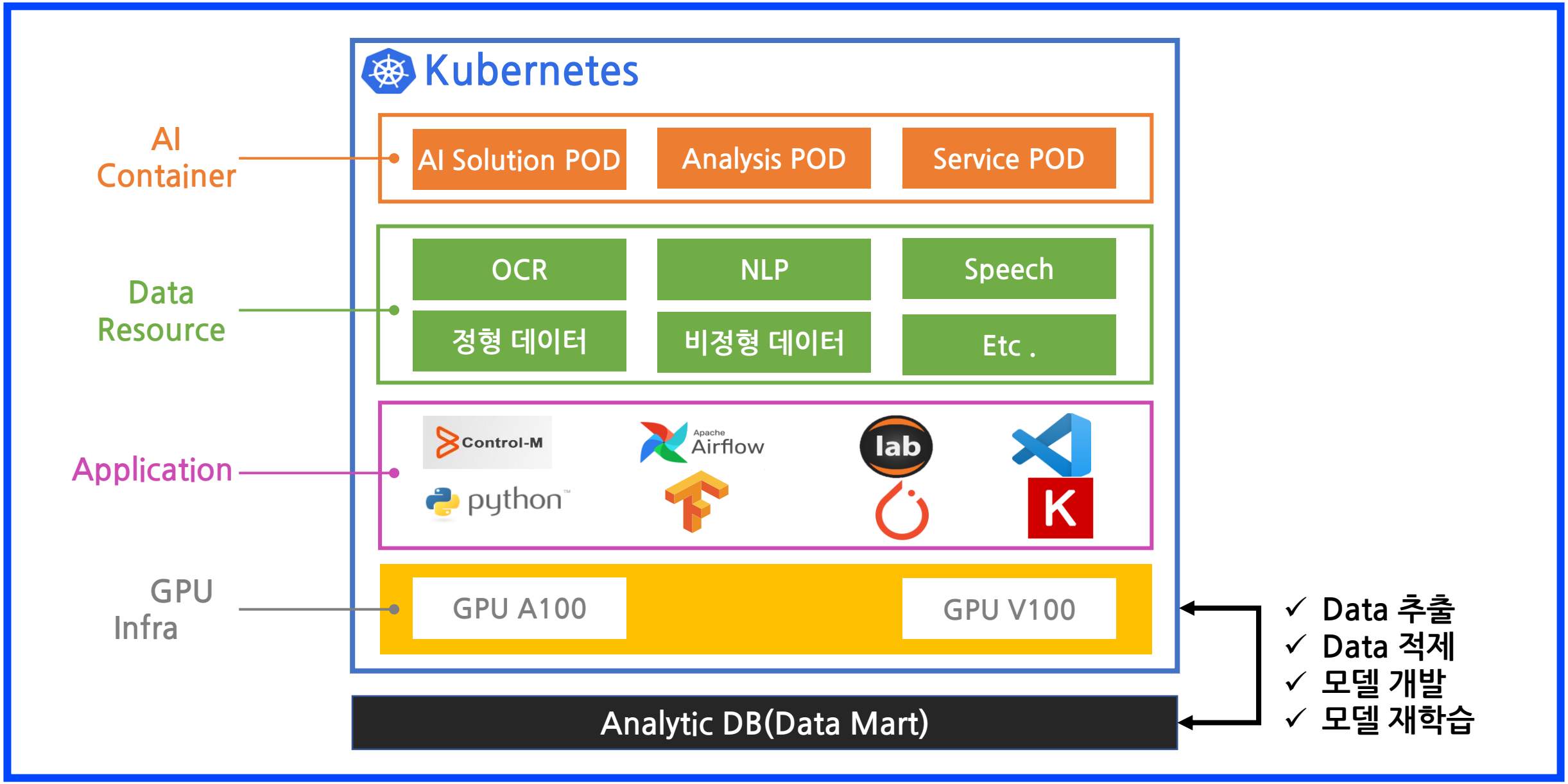
AI 서비스 런칭을 위한 GPU 도입 가능여부 검토 결과

- 다양한 AI서비스 증가
- 높은 GPU 도입 비용
- 탄력적 리소스 활용
- Public Cloud 연동 한계

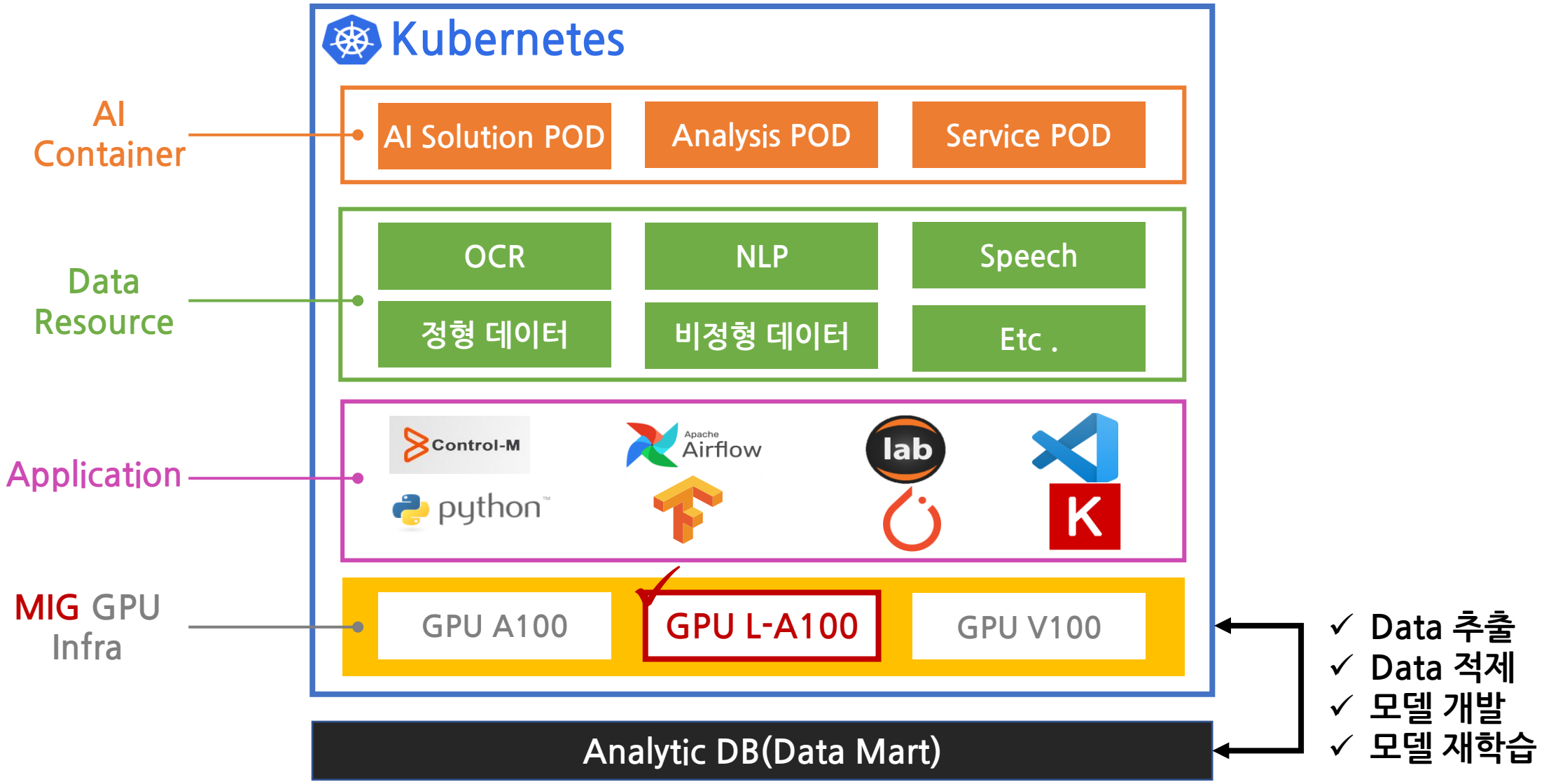
“효율적 GPU 활용을 위한 MIG 도입 결정”

3 효율적인 MIG GPU 적용 사례

1) AI 서비스 구성


Kubernetes
AI Container
AI Solution POD
Analysis POD
Service POD
Data Resource
OCR
NLP
Speech
정형 데이터
비정형 데이터
Etc .
Application
Control-M
Apache Airflow
lab
python
K
GPU Infra
GPU A100
GPU V100
Analytic DB(Data Mart)
✓ Data 추출
✓ Data 적제
✓ 모델 개발
✓ 모델 재학습

1) AI 서비스 구성

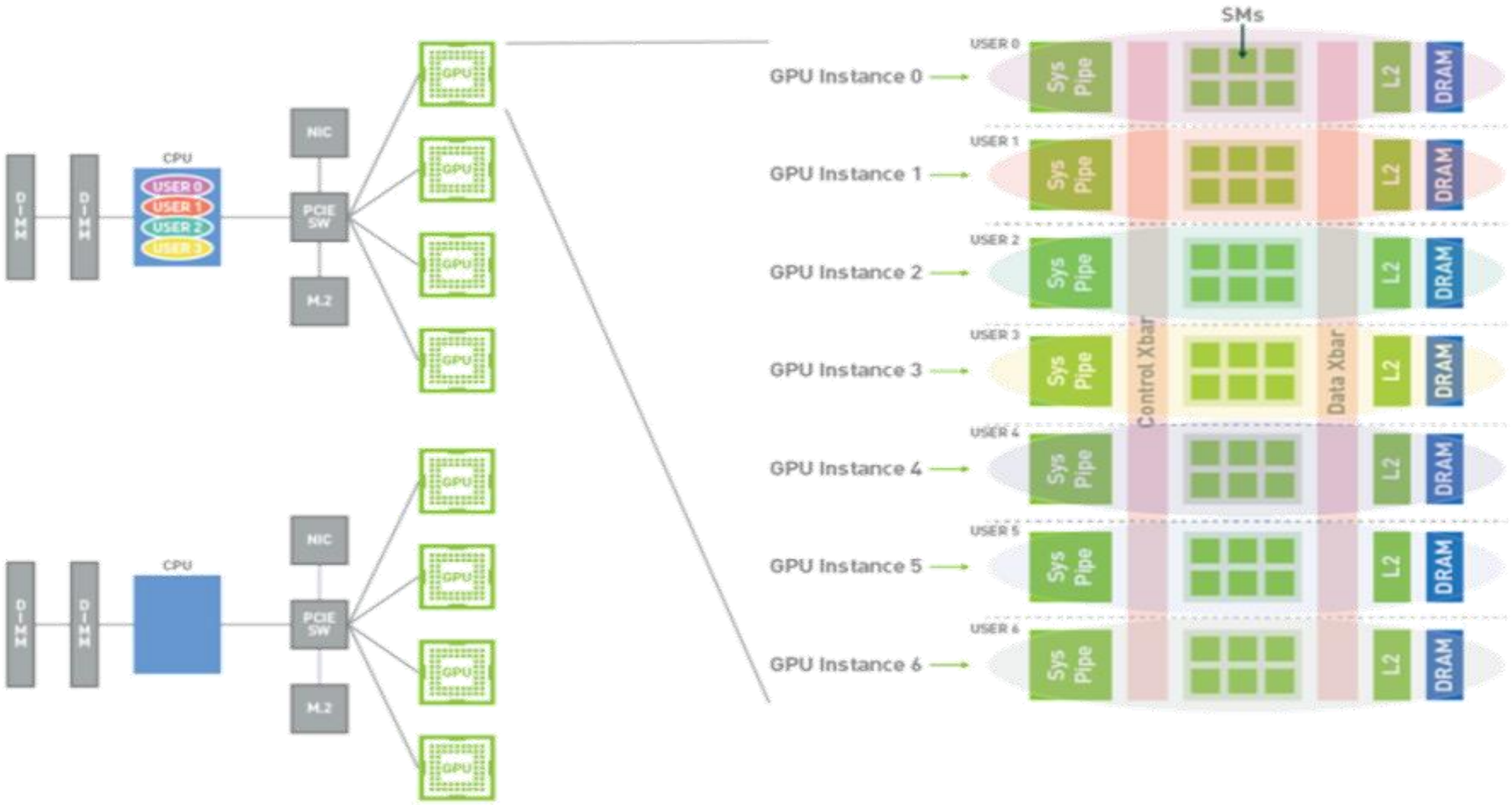

Kubernetes
AI Container
AI Solution POD
Analysis POD
Service POD
Data Resource
OCR
NLP
Speech
정형 데이터
비정형 데이터
Etc .
Application
Control-M
Apache Airflow
lab
python
K
MIG GPU Infra
GPU A100
GPU L-A100
GPU V100
Analytic DB(Data Mart)
✓ Data 추출
✓ Data 적제
✓ 모델 개발
✓ 모델 재학습

[참고] Nvidia MIG(Multi-Instance GPU) 소개

NVIDIA MIG: Nvidia 멀티 인스턴스 GPU A100의 기능

- 사용자는 여러 개의 작은 GPU가 있는 것처럼 **여러 GPU 워크로드를 동시에 실행**하여 단일 GPU의 활용도 극대화
- 단일 A100 GPU에서 여러 워크로드를 **병렬로 실행**하거나 **여러 사용자**가 하드웨어 수준 격리 및 서비스 품질을 통해 A100 GPU를 **공유**할 수 있도록 지원
- **Kubernetes 플랫폼**을 지원하며, k8s-device-plugin 및 gpu-feature-discovery 플러그인을 통해 확장 지원
- 각 k8s-device-plugin Kubernetes 컴퓨팅 노드에서 NVIDIA GPU의 가용성을 확인하고 nvidia.com/gpu리소스 유형 제공
- 노드의 gpu-feature-discovery 플러그인은 드라이버 버전, GPU 유형 등과 같은 GPU 장치의 메타 정보를 기반으로 노드 레이블을 생성, 적용

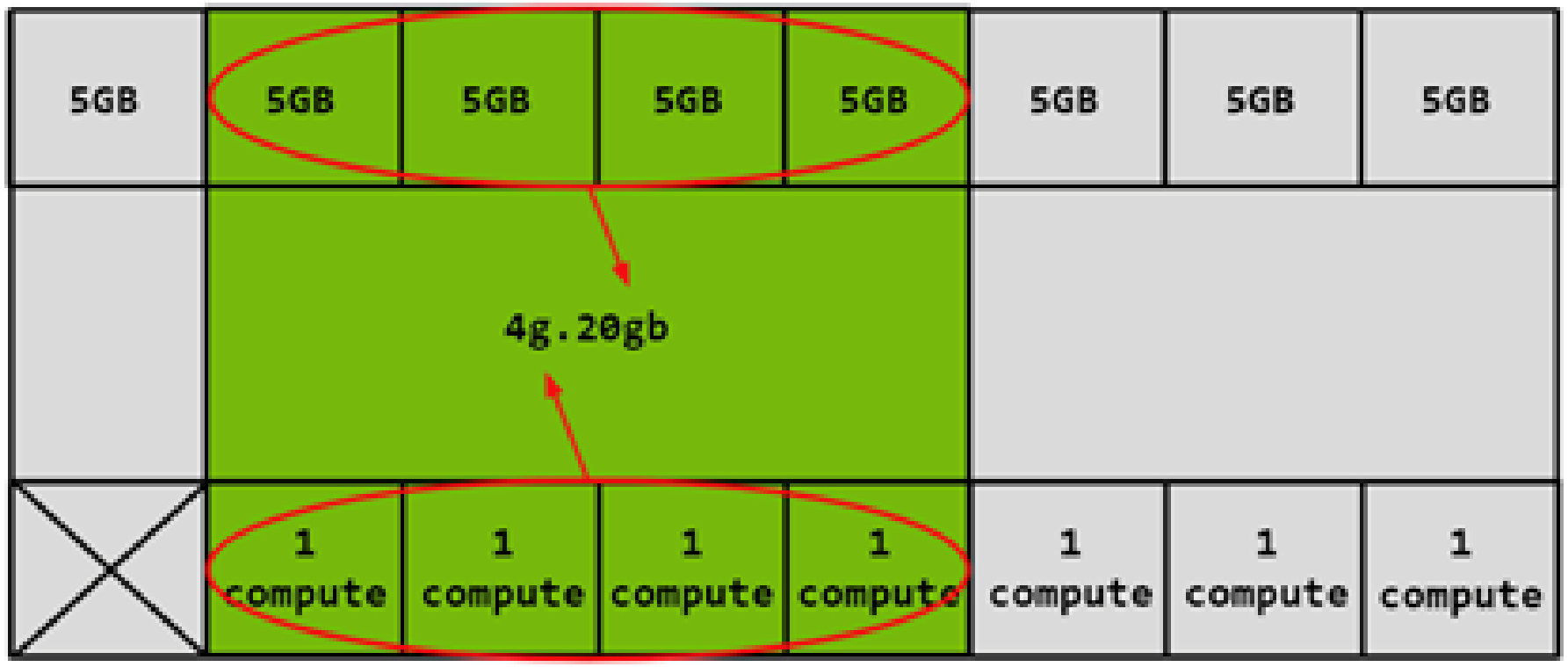
2) MIG의 주요 기능과 특징: MIG 1g.5gb의 7EA Instance 분할



2) MIG의 주요 기능과 특징: 분할 종류

MIG 이름	MIG Instance GPU 메모리	Max Instance GPU 수	MIG Instance Compute Units
1g.5gb	5GB	7	1
2g.10gb	10GB	3	2
3g.20gb	20GB	2	3
4g.20gb	20GB	1	4
7g.40gb	40GB	1	7
8g.40gb(MIG비활성화)	40GB	1	8

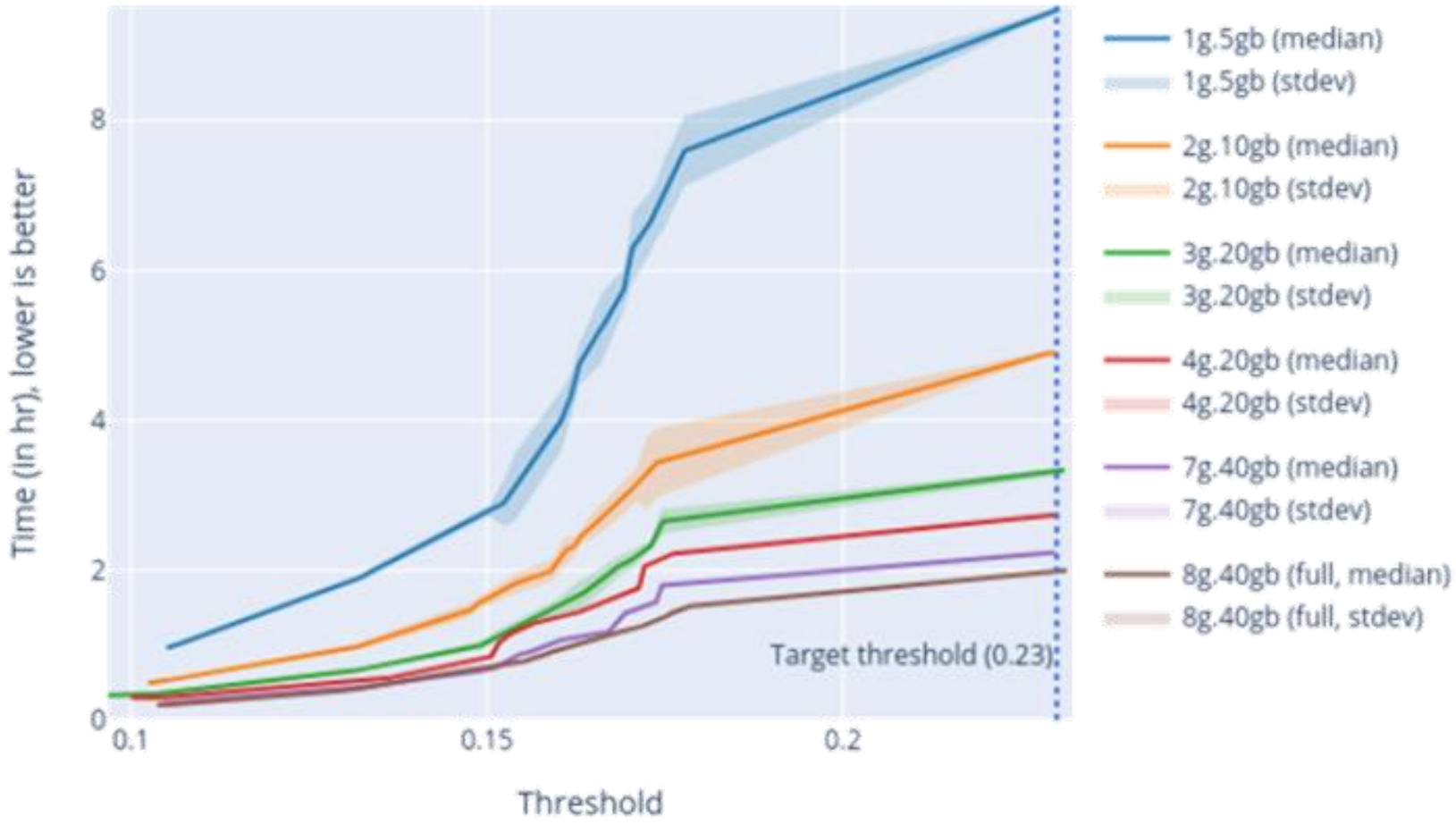
2) MIG 분할 명칭 & 구조(4g.20gb)



- 4g : 4 x Compute Instance
- 20gb : 5gb x 4 GPU Memory

2) MIG의 주요 기능과 특징: MIG 인스턴스 성능 지표

시간 경과에 따른 분류 임계값의 변화



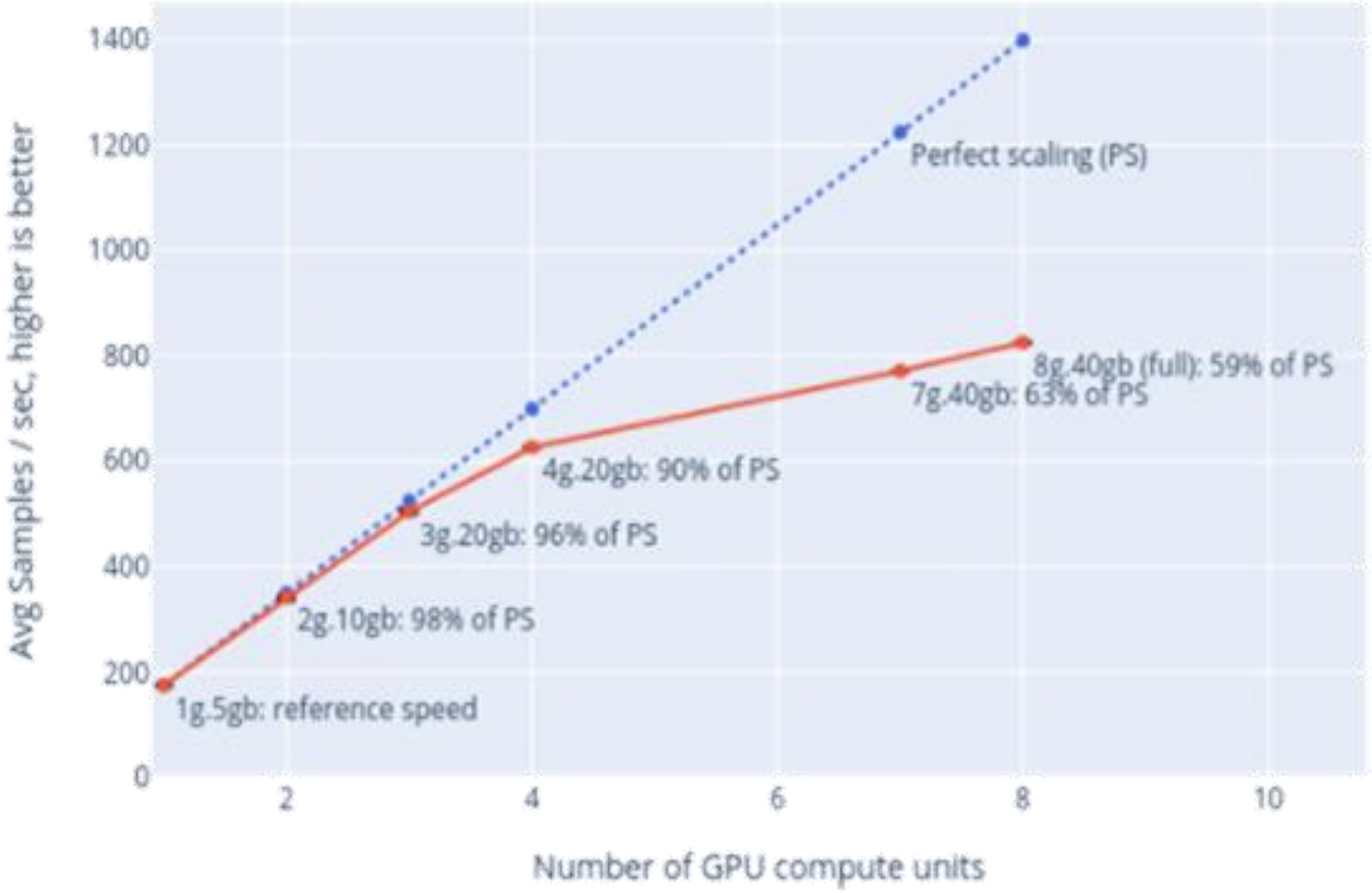
***이미지 분류 benchmark**

- Y축 시간 단위
- 세 번 실행 중앙값과 컬러 표면은 표준편차

https://cloud.redhat.com/blog/using-nvidia-a100s-multi-instance-gpu-to-run-multiple-workloads-in-parallel-on-a-single-gpu

2) MIG의 주요 기능과 특징: MIG 인스턴스 성능 지표

인스턴스 처리 속도



- 초당 처리되는 평균 이미지 샘플 수
- 확장 속도
- 각 다른 GPU 엔진에서 작업 부하 병렬화 확인

https://cloud.redhat.com/blog/using-nvidia-a100s-multi-instance-gpu-to-run-multiple-workloads-in-parallel-on-a-single-gpu

2) MIG의 주요 기능과 특징: MIG 크기에 대한 NVIDIA 권장 워크로드 유형

MIG Instance	SMs per Instance	Memory per Instance	# Instances per GPU	Target Workload
MIG 1g.5gb	14	5GB	7	Jupyter Notebooks For Development, Model Tuning, Inference, Light HPC
MIG 2g.10gb	28	10GB	3	Inference, Light HPC
MIG 3g.20gb	42	20GB	2	Light Training, Inference, HPC
MIG 4g.20gb	56	20GB	1	Light Training, Inference, HPC
MIG 7g.40gb	98	40GB	1	Training, HPC

Note: Instance 이름의 'g' 앞의 숫자는 GPU compute 슬라이스의 숫자이며, 'gb' 앞의 숫자는 해당 instance에 할당된 GPU 메모리 사이즈임

3) 당행 상세 적용 방안 : 개발서버 MIG 분할 적용

개발서버 GPU 8EA에 MIG 적용

GPU 0 MIG 1g.5gb x 7	sudo nvidia-smi mig -i 0 -cgi 19,19,19,19,19,19,19 -C
GPU 1 MIG 1g.5gb x 7	sudo nvidia-smi mig -i 0 -cgi 19,19,19,19,19,19,19 -C
GPU 2 MIG 2g.10gb x 3	sudo nvidia-smi mig -i 0 -cgi 14,14,14 -C
GPU 3 MIG 2g.10gb x 3	sudo nvidia-smi mig -i 0 -cgi 14,14,14 -C
GPU 4 MIG 3g.20gb x 2	sudo nvidia-smi mig -i 0 -cgi 9,9 -C
GPU 5 MIG 3g.20gb x 2	sudo nvidia-smi mig -i 0 -cgi 9,9 -C
GPU 6 MIG 7g.40gb x 1	sudo nvidia-smi mig -i 0 -cgi 0 -C
GPU 7 MIG 7g.40gb x 1	sudo nvidia-smi mig -i 0 -cgi 0 -C

3) MIG 적용 상세 내용(nvidia-smi 내용)


```
$ nvidia-smi
+-----------------------------------------------------------------------------+
| MIG devices:                                                                |
+------------------+----------------------+-----------+----------------------+
| GPU  GI  CI  MIG |         Memory-Usage |        Vol|         Shared       |
|      ID  ID  Dev |                      |     SM    Unc| CE  ENC  DEC  OFA  JPG|
|                  |                      |        ECC|                      |
|==================+======================+===========+======================|
|  0    1   0   0  |     11MiB / 20224MiB | 42      0 |  3   0    2    0    0 |
+------------------+----------------------+-----------+----------------------+
|  0    2   0   1  |     11MiB / 20096MiB | 42      0 |  3   0    2    0    0 |
+------------------+----------------------+-----------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

```
$ nvidia-smi -L

GPU 0: A100-SXM4-40GB (UUID: GPU-5d5ba0d6-d33d-2b2c-524d-9e3d8d2b8a77)
  MIG 1g.5gb      Device  0: (UUID: MIG-c6d4f1ef-42e4-5de3-91c7-45d71c87eb3f)
  MIG 1g.5gb      Device  1: (UUID: MIG-cba663e8-9bed-5b25-b243-5985ef7c9beb)
  MIG 1g.5gb      Device  2: (UUID: MIG-1e099852-3624-56c0-8064-c5db1211e44f)
  MIG 1g.5gb      Device  3: (UUID: MIG-8243111b-d4c4-587a-a96d-da04583b36e2)
  MIG 1g.5gb      Device  4: (UUID: MIG-169f1837-b996-59aa-9ed5-b0a3f99e88a6)
  MIG 1g.5gb      Device  5: (UUID: MIG-d5d0152c-e3f0-552c-abee-ebc0195e9f1d)
  MIG 1g.5gb      Device  6: (UUID: MIG-7df6b45c-a92d-5e09-8540-a6b389968c31)
GPU 1: A100-SXM4-40GB (UUID: GPU-0aa11ebd-627f-af3f-1a0d-4e1fd92fd7b0)
  MIG 2g.10gb     Device  0: (UUID: MIG-0c757cd7-e942-5726-a0b8-0e8fb7067135)
  MIG 2g.10gb     Device  1: (UUID: MIG-703fb6ed-3fa0-5e48-8e65-1c5bdcfe2202)
  MIG 2g.10gb     Device  2: (UUID: MIG-532453fc-0faa-5c3c-9709-a3fc2e76083d)
GPU 2: A100-SXM4-40GB (UUID: GPU-08279800-1cbe-a71d-f3e6-8f67e15ae54a)
  MIG 3g.20gb     Device  0: (UUID: MIG-aa232436-d5a6-5e39-b527-16f9b223cc46)
  MIG 3g.20gb     Device  1: (UUID: MIG-3b12da37-7fa2-596c-8655-62dab88f0b64)
GPU 3: A100-SXM4-40GB (UUID: GPU-71086aca-c858-d1e0-aae1-275bed1008b9)
  MIG 7g.40gb     Device  0: (UUID: MIG-3e209540-03e2-5edb-8798-51d4967218c9)
GPU 4: A100-SXM4-40GB (UUID: GPU-74fa9fb7-ccf6-8234-e597-7af8ace9a8f5)
  MIG 1c.3g.20gb  Device  0: (UUID: MIG-79c62632-04cc-574b-af7b-cb2e307120d8)
  MIG 1c.3g.20gb  Device  1: (UUID: MIG-4b3cc0fd-6876-50d7-a8ba-184a86e2b958)
  MIG 1c.3g.20gb  Device  2: (UUID: MIG-194837c7-0476-5b56-9c45-16bddc82e1cf)
  MIG 1c.3g.20gb  Device  3: (UUID: MIG-291820db-96a4-5463-8e7b-444c2d2e3dfa)
  MIG 1c.3g.20gb  Device  4: (UUID: MIG-5a97e28a-7809-5e93-abae-c3818c5ea801)
  MIG 1c.3g.20gb  Device  5: (UUID: MIG-3dfd5705-b18a-5a7c-bcee-d03a0ccb7a96)
GPU 5: A100-SXM4-40GB (UUID: GPU-3301e6dd-d38f-0eb5-4665-6c9659f320ff)
  MIG 4g.20gb     Device  0: (UUID: MIG-6d96b9f9-960e-5057-b5da-b8a35dc63aa8)
GPU 6: A100-SXM4-40GB (UUID: GPU-bb40ed7d-cbbb-d92c-50ac-24803cda52c5)
  MIG 1c.7g.40gb  Device  0: (UUID: MIG-66dd01d7-8cdb-5a13-a45d-c6eb0ee11810)
  MIG 2c.7g.40gb  Device  1: (UUID: MIG-03c649cb-e6ae-5284-8e94-4b1cf767e06c)
  MIG 3c.7g.40gb  Device  2: (UUID: MIG-8abf68e0-2808-525e-9133-ba81701ed6d3)
GPU 7: A100-SXM4-40GB (UUID: GPU-95fac899-e21a-0e44-b0fc-e4e3bf106feb)
  MIG 4g.20gb     Device  0: (UUID: MIG-219c765c-e07f-5b85-9c04-4afe174d83dd)
  MIG 2g.10gb     Device  1: (UUID: MIG-25884364-137e-52cc-a7e4-ecf3061c3ae1)
  MIG 1g.5gb      Device  2: (UUID: MIG-83e71a6c-f0c3-5dfc-8577-6e8b17885e1f)
```

3) 당행 상세 적용 방안 : node describe 내용


```
Capacity:
  cpu:                256
  ephemeral-storage:  1843217020Ki
  hugepages-1Gi:      0
  hugepages-2Mi:      0
  memory:             1056648992Ki
  nvidia.com/gpu:     16
  nvidia.com/mig-1g.5gb:   0
  nvidia.com/mig-2g.10gb:  0
  nvidia.com/mig-3g.20gb:  0
  nvidia.com/mig-7g.40gb:  0
  pods:               110
Allocatable:
  cpu:                256
  ephemeral-storage:  1698708802820
  hugepages-1Gi:      0
  hugepages-2Mi:      0
  memory:             1056546592Ki
  nvidia.com/gpu:     16
  nvidia.com/mig-1g.5gb:   0
  nvidia.com/mig-2g.10gb:  0
  nvidia.com/mig-3g.20gb:  0
  nvidia.com/mig-7g.40gb:  0
  pods:               110
```

```
Allocated resources:
  (Total limits may be over 100 percent, i.e., overcommitted.)
  Resource                    Requests          Limits
  --------                    --------          ------
  cpu                         107500m (41%)     122200m (47%)
  memory                      472812Mi (45%)    480752Mi (46%)
  ephemeral-storage           0 (0%)            0 (0%)
  hugepages-1Gi               0 (0%)            0 (0%)
  hugepages-2Mi               0 (0%)            0 (0%)
  nvidia.com/gpu              4                 4
  nvidia.com/mig-1g.5gb       0                 0
  nvidia.com/mig-2g.10gb      0                 0
  nvidia.com/mig-3g.20gb      0                 0
  nvidia.com/mig-7g.40gb      0                 0
```

3) 당행 상세 적용 방안 : MIG 유형별 수행 테스트

MIG 적용 대상 모델 선정을 위한 유형별 테스트 수행

테스트 리소스: CPU 4core, Memory 24GB 이며 수행 결과는 epoch당 걸린 시간으로 측정

신한은행 NLP 모델(Bert)

MIG 이름	수행 결과
1g 5gb	메모리 allocation error
2g 10gb	메모리 allocation error
3g 20gb	메모리 allocation error
7g 40gb	470s

신한은행 이미지 분류 모델 (resnet50)

MIG 이름	수행 결과
1g 5gb	메모리 allocation error
2g 10gb	499s
3g 20gb	498s
7g 40gb	494s

3) 당행 상세 적용 방안 : MIG 유형별 수행 테스트

MIG 적용 대상 모델 선정을 위한 유형별 테스트 수행

테스트 리소스: CPU 4core, Memory 24GB 이며 수행 결과는 epoch당 걸린 시간으로 측정

신한은행 NLP 모델(Bert)

MIG 이름	수행 결과
1g 5gb	메모리 allocation error
2g 10gb	메모리 allocation error
3g 20gb	메모리 allocation error
7g 40gb	470s

➡ 대용량 모델로 GPU 메모리 사용량이 많아 MIG 적용 적합 X

✓ 신한은행 이미지 분류 모델 (resnet50)

MIG 이름	수행 결과
1g 5gb	메모리 allocation error
2g 10gb	499s
3g 20gb	498s
7g 40gb	494s

➡ 작은 GPU 메모리에도 사용 가능하므로 MIG 적용 적합 O

3) 당행 상세 적용 방안 : MIG 한계점 보완

MIG 적용 시 한계점	당행 도입 시 보완 방안
Memory/SM 성능손실	✓ 성능손실을 최소화한 3g.20gb 사용
NVLink 지원불가	
Multi GPU 사용 시, NCCL 사용불가	✓ Single 방식 적용으로 AI/ML 워크로드에 적합한 활용
서버 재부팅 시 MIG Profile 초기화	✓ 데몬셋 적용으로 서버 재기동 시, 자동 MIG Profile 적용

3) 당행 상세 적용 방안 : 운영환경 적용

적합한 MIG 유형 선정

- MIG 3g.20gb(각2EA로 분할)
- MIG 3g.20gb 적용 시, DGX 서버 GPU 8EA -> 16EA 확장

3) 당행 상세 적용 방안 : node describe 내용


```
nvidia.com/cuda.driver.major=450
nvidia.com/cuda.driver.minor=80
nvidia.com/cuda.driver.rev=02
nvidia.com/cuda.runtime.major=11
nvidia.com/cuda.runtime.minor=0
nvidia.com/gfd.timestamp=1650536401
nvidia.com/gpu=true
nvidia.com/gpu.compute.major=8
nvidia.com/gpu.compute.minor=0
nvidia.com/gpu.count=16
nvidia.com/gpu.engines.copy=3
nvidia.com/gpu.engines.decoder=2
nvidia.com/gpu.engines.encoder=0
nvidia.com/gpu.engines.jpeg=0
nvidia.com/gpu.engines.ofa=0
nvidia.com/gpu.family=ampere
nvidia.com/gpu.machine=DGXA100-920-23687-2530-000
nvidia.com/gpu.memory=20096
nvidia.com/gpu.multiprocessors=42
nvidia.com/gpu.product=A100-SXM4-40GB-MIG-3g.20gb
nvidia.com/gpu.slices.ci=3
nvidia.com/gpu.slices.gi=3
nvidia.com/mig.strategy=single
```

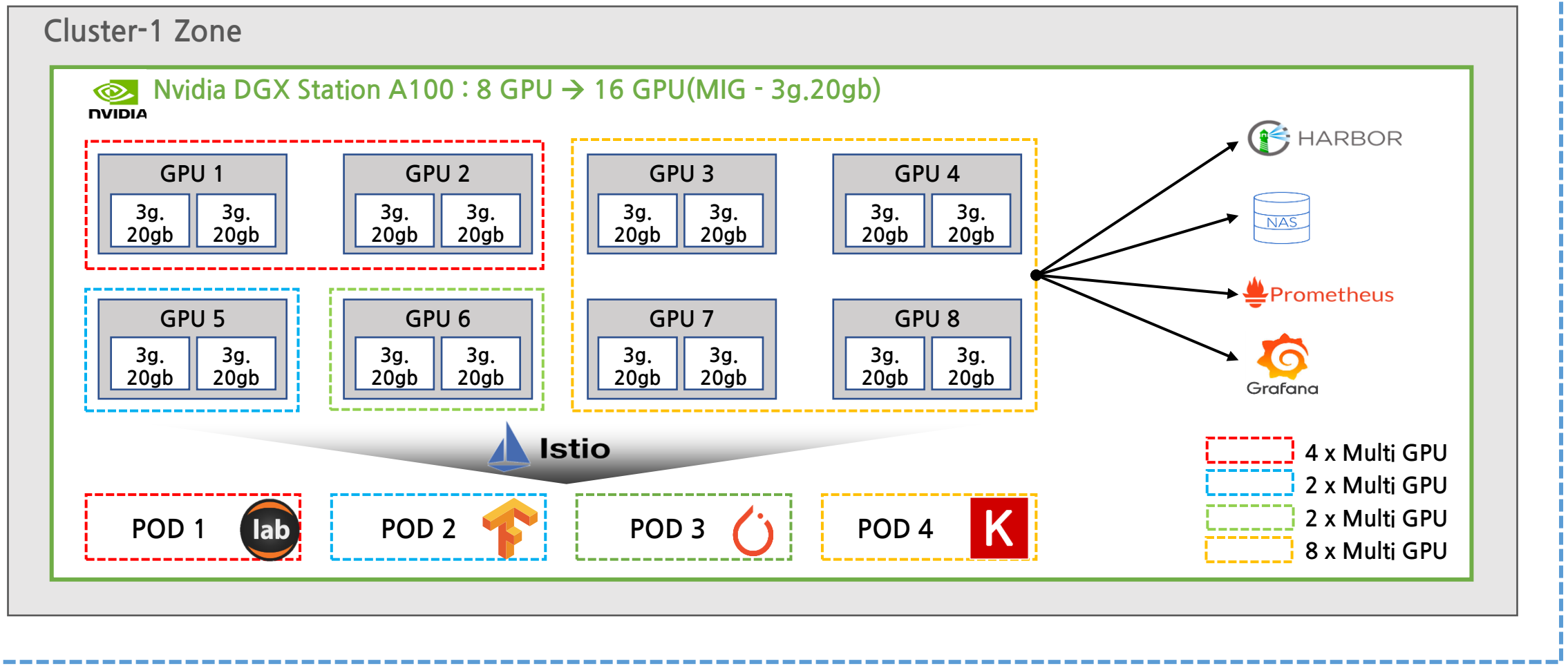
3) 당행 상세 적용 방안 : node describe 내용


```
Capacity:
  cpu:                256
  ephemeral-storage:  1843217020Ki
  hugepages-1Gi:      0
  hugepages-2Mi:      0
  memory:             1056643388Ki
  nvidia.com/gpu:     16
  pods:               110
Allocatable:
  cpu:                256
  ephemeral-storage:  16987088802820
  hugepages-1Gi:      0
  hugepages-2Mi:      0
  memory:             1056540988Ki
  nvidia.com/gpu:     16
  pods:               110
```

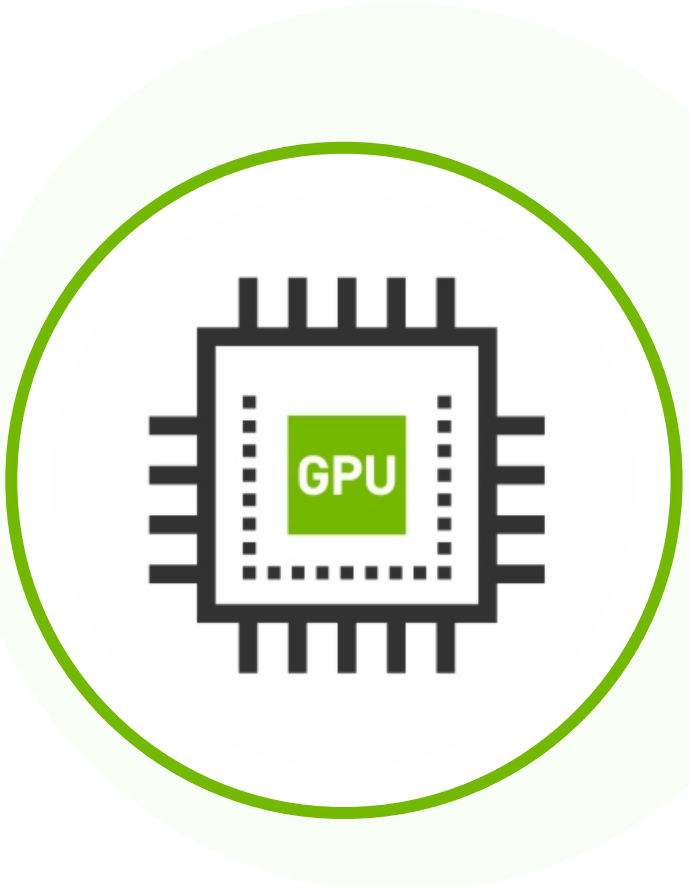
4) 서비스 아키텍처

Private Cloud

Kubernetes



5) 기대효과



01 AI모델 분석 시, GPU 사용성 향상

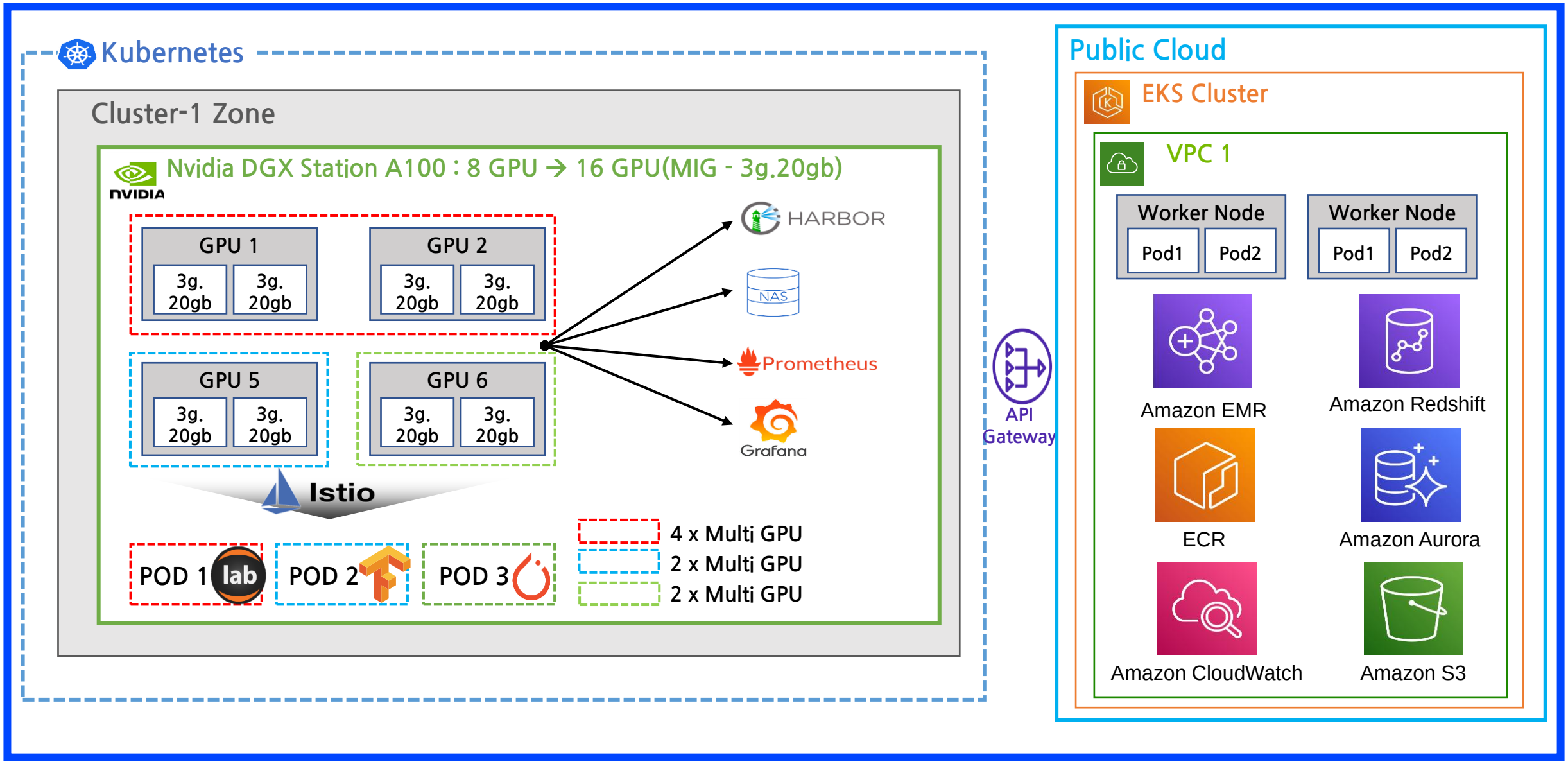
02 Multi-GPU 사용 장려 및 활용

03 AI 실시간/ 배치 서비스 증가

04 탄력적 GPU 활용

4 Hybrid Architecture 플랫폼

1) Hybrid Architecture (To-Be)



2) Hybrid Architecture 필요성

✓ Public Cloud의 높은 GPU 비용

- Public Cloud에서의 GPU 비용은 너무 높음
  - 서빙의 경우 24시간 GPU 사용됨

✓ Public Cloud의 가격적 효율성

- GPU는 On-Premise에서 적극 활용
- Infra 환경의 Data 분석, 학습, 서빙의 적절한 분배
- 다양한 SaaS 활용

3) 결론

MIG를 활용한 효율적 GPU 구성 환경을 통해…

- AI/ML 워크로드의 성능 손실 최소화
- L-A100 GPU 이용, oo개 배치/실시간 서비스를 통한 안정성 입증
- 인프라 도입 비용 절감
- Hybrid Architecture 구성으로 탄력적 환경 구성

End of Document