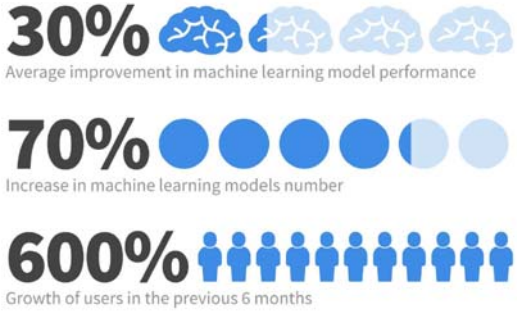
2019. 12. 17.

Testbed for AI Infrastructure

안 종 석
james@jslab.kr
JS Lab

AI Testbed 인프라 서비스 개요

- **AI 플랫폼 자동화 서비스**
  - 오픈소스 기반 등록/검색/분석 서비스 자동화
  - 융합 기술 개발에 집중하는 부가 서비스 제공
  - 학습 모델링 자동화로 비 데이터 분석 전문가의 사용 범위 확대
  - 학습 결과의 생성/배포 서비스로 활용 서비스 환경 지원
- **자원 활용의 극대화를 위한 클라우드 기반 ML/DL/AI 가속 서비스**
  - 클라우드 기반 ML/DL/AI 분석과 학습의 고속 처리를 위한 인프라 구성
  - 데이터 유통 관리의 무결성을 위한 클라우드 기반 블록체인 서비스 개발
  - 클라우드 서비스 기반 공유 자원의 활용을 극대화
  - 개인에게 독립적인 연구 환경 제공
  - 멀티 클라우드 기반 실시간 발생 데이터 분석 연계 서비스

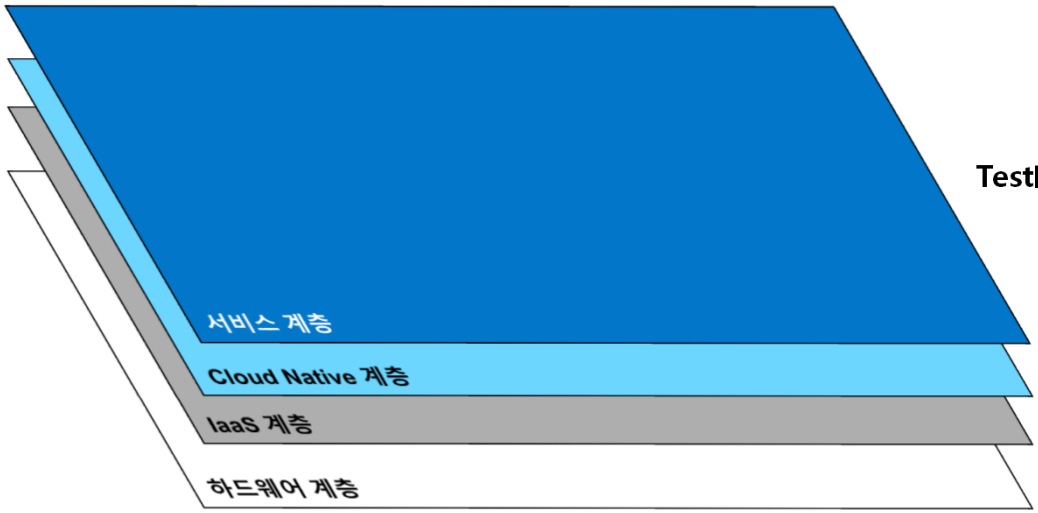


Rad Hat @ Kubecon 2019

인프라 구조

계층화 인프라 구조

- 클라우드 서비스를 위한 계층별 추상화로 활용 극대화 서비스 집중
- 계층간 격리/정책 기반 서비스 노출로 Compliance(준수)를 위한 보안 강화
- 성능 개선이나 연결 기관/서비스 등과 호환성을 위한 계층 Offload 구성 지원



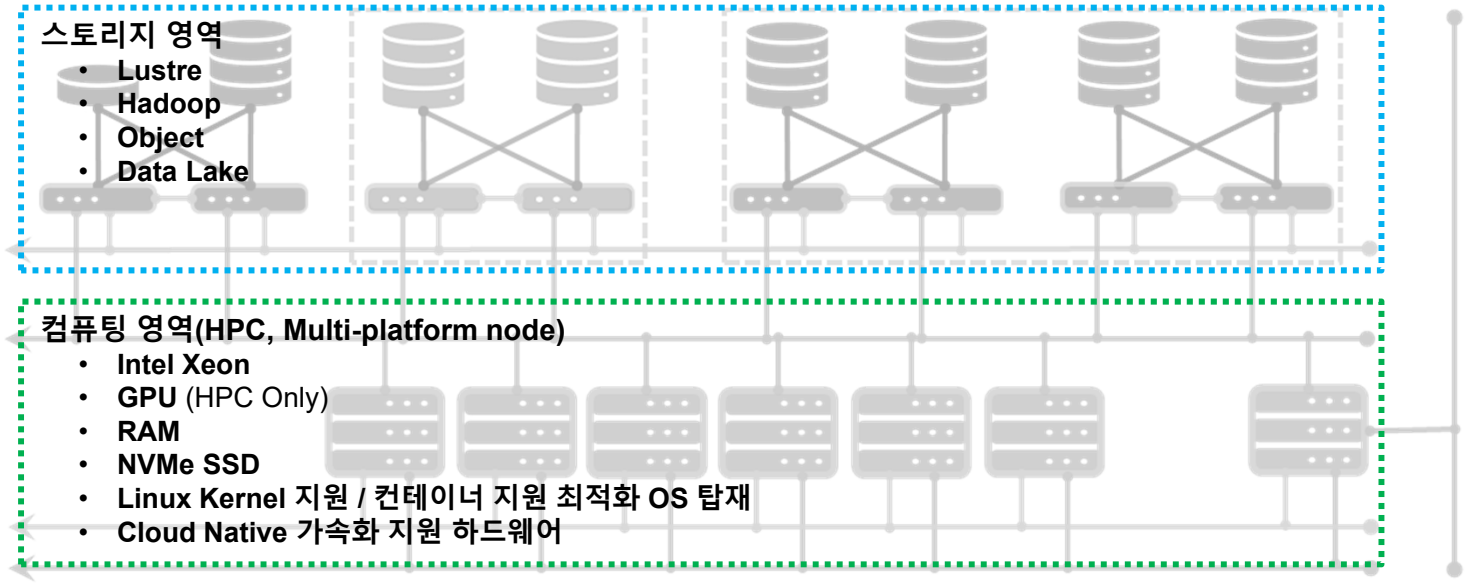
Testbed for AI Infrastructure

인프라 아키텍처의 계층 구조

- 하드웨어 계층
  - 협업과 융합 기술을 위한 고속의 클라우드 서비스용 오픈 하드웨어 아키텍처 구성
  - 기술 발전을 수용하고 클라우드 서비스 성능 향상을 지원하는 하드웨어 확장 구조
- IaaS 계층
  - vGPU 지원 등 SDDC 기반 클라우드 서비스의 유연성을 제공하는 가상화 인프라 계층
  - 클라우드 서비스를 위한 Compliance 수용 보안 강화
- Cloud Native 계층
  - 생산성 증대를 위한 클라우드 서비스 제공 (Hyperparameter Tuning, 메시지 제공 등)
  - 멀티 클라우드 기반 확대 서비스 (서비스 배포, Model Serving, Cloud bursting 등)
- 서비스 계층
  - 산/학/연 협동 연구와 상용화를 위한 플랫폼 서비스 제공
  - ML/DL/AI 개발과 상용화 과정의 자동화 제공과 지식의 자산화 Data Lake Ready

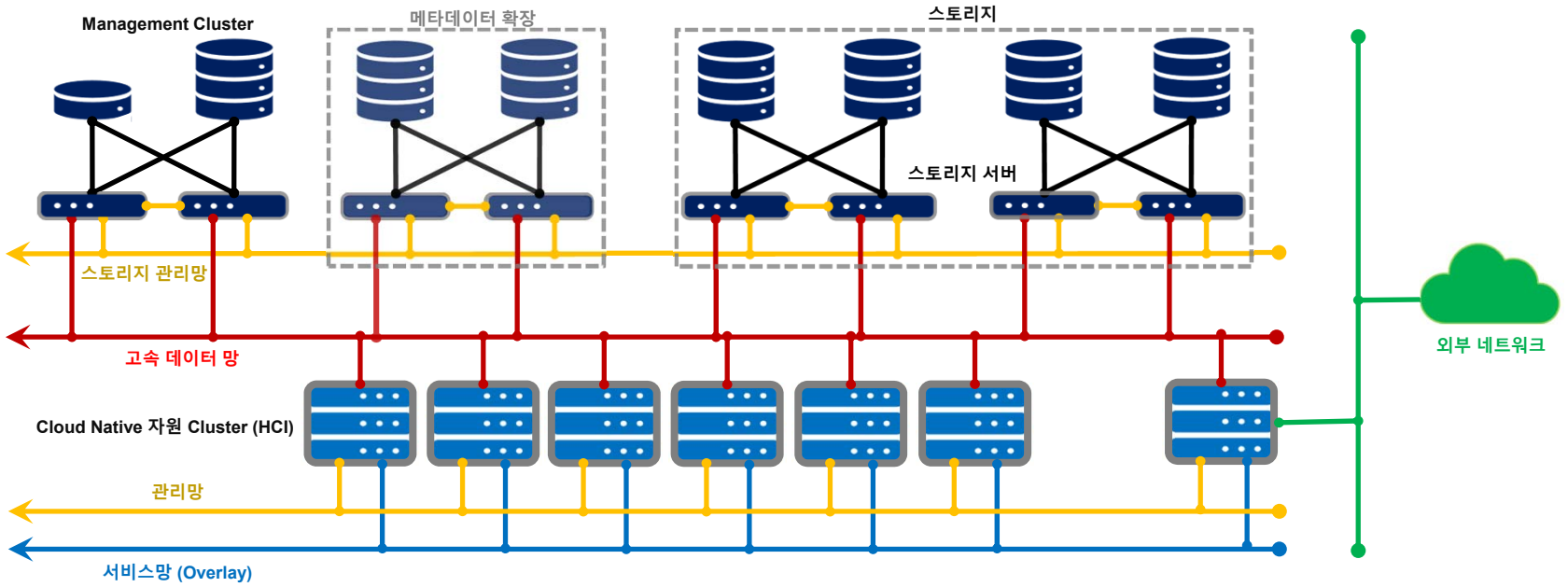
하드웨어 계층

- 스토리지 영역 구성: 객체 기반의 데이터를 분산 저장 (Data Lake화)
- 컴퓨팅 영역 구성: GPU 사용 노드 / DMZ 구성 등의 멀티플랫폼 노드로 구성
- Cloud Native 가속화 지원 하드웨어 계층 구조



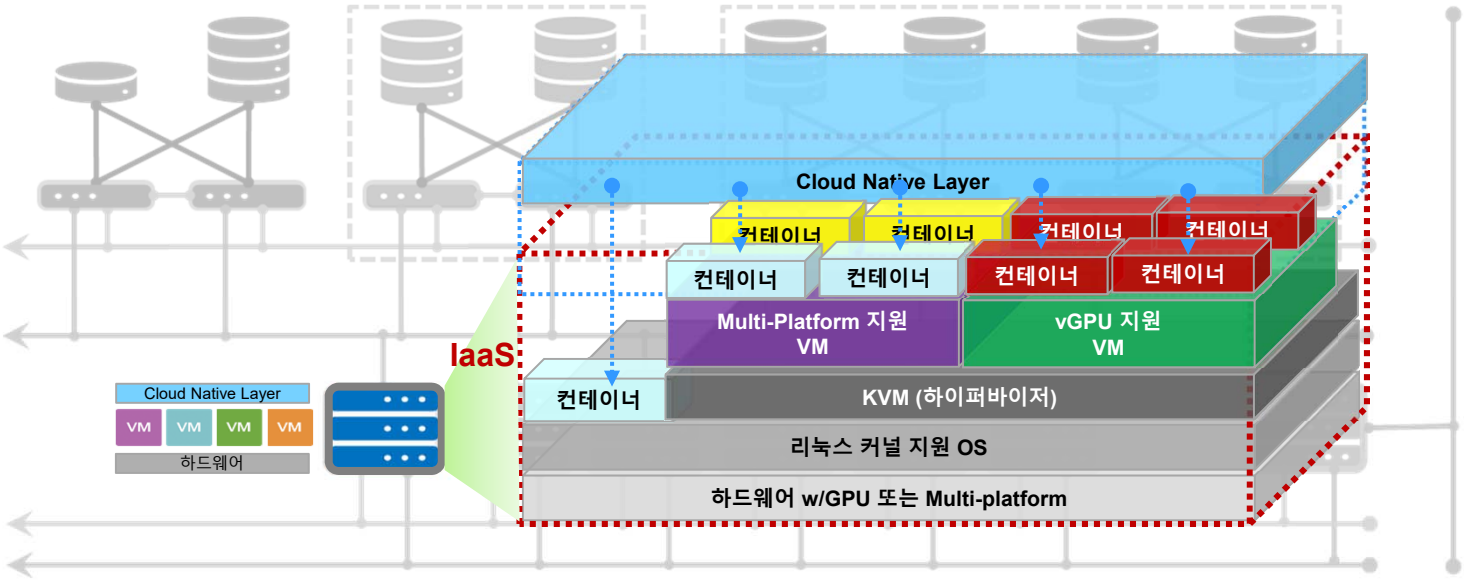
하드웨어 계층

- 고속 처리 이중화 기반 스토리지와 GPU 지원 HCI 기반 클러스터링 구성
- IaaS(Infrastructure as a Service)를 위한 가상화 OS(하이퍼바이저) 사용
- 클라우드 서비스 성능을 위한 스케일아웃(Scale-out) 확장 구조



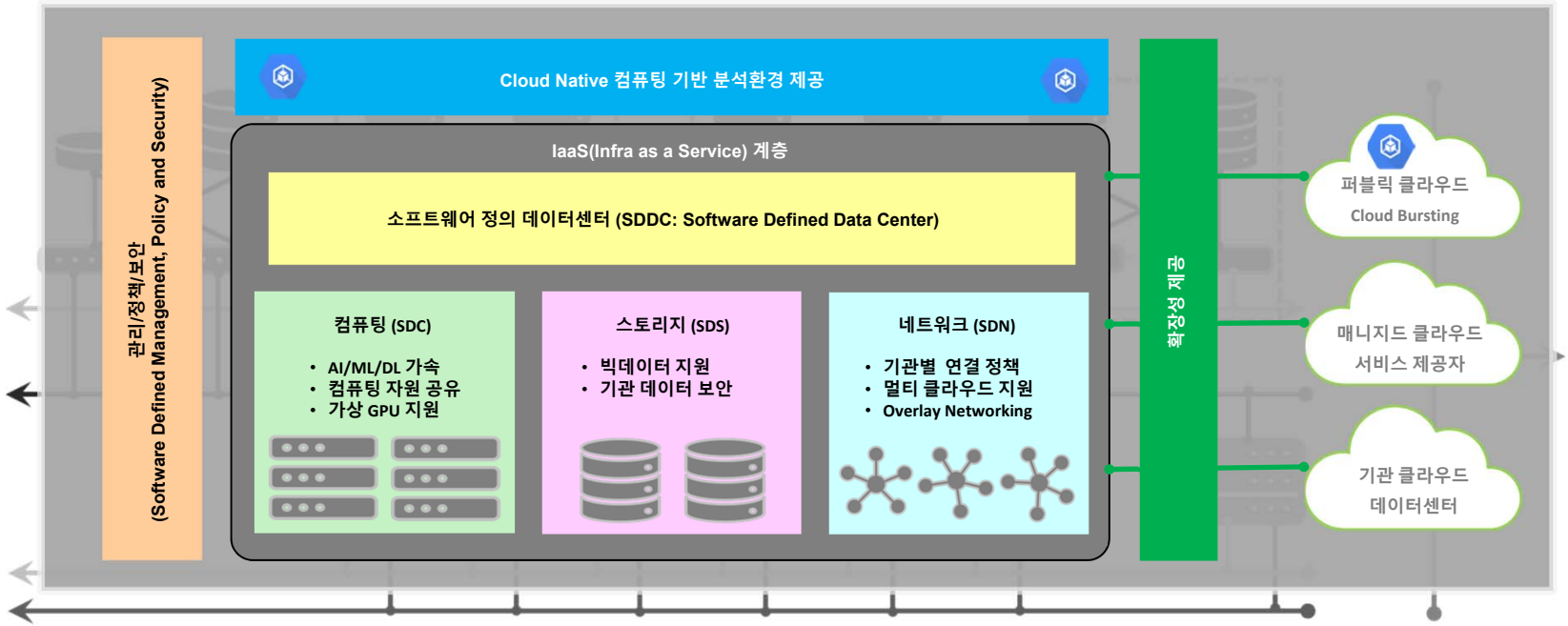
IaaS(Infrastructure as a Service) 계층 (1 of 2)

- 컴퓨팅 노드 가상화
  - NUMA(Non-Uniform Memory Access) Aware
  - vGPU 지원 VM 지원 제공
  - Container화 OpenStack의 Kubernetes 기반 제어



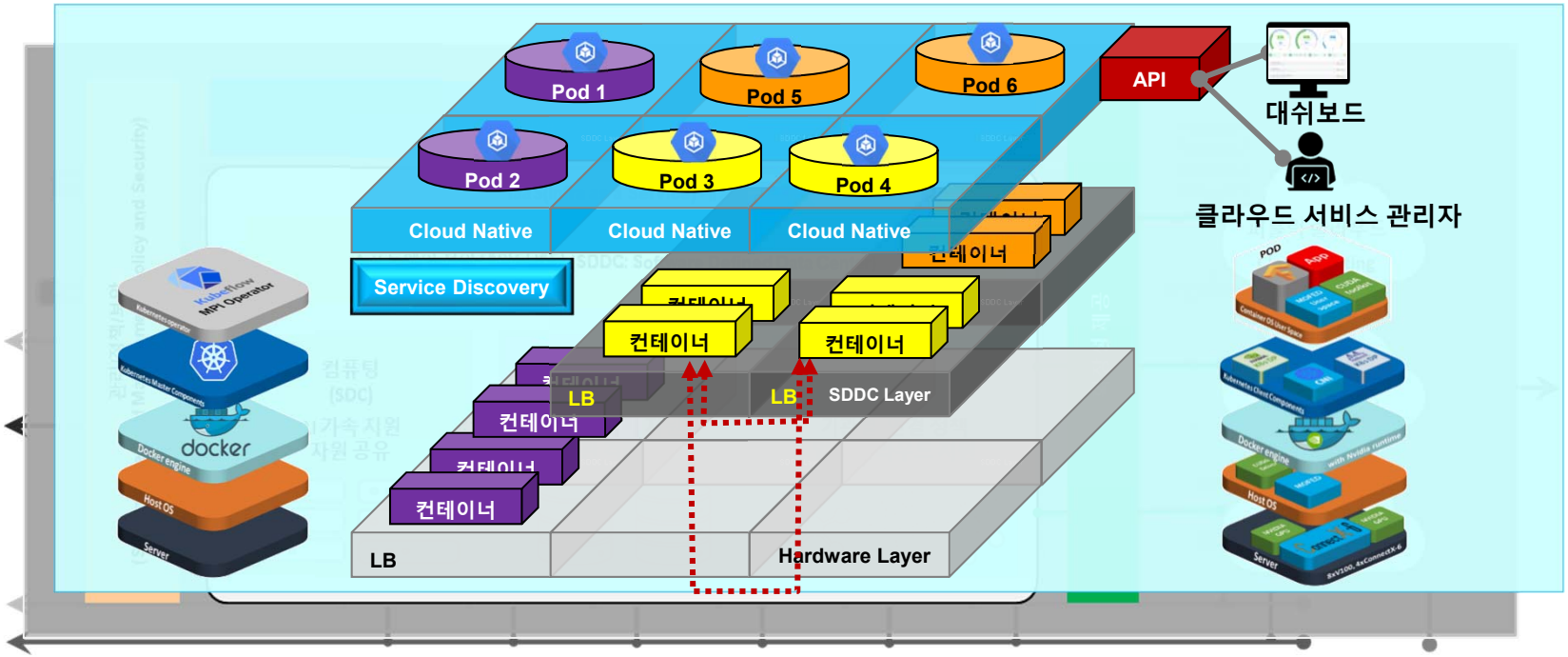
IaaS(Infrastructure as a Service) 계층 (2 of 2)

- 하드웨어 추상화 자원 제공으로 유연한 인프라 운영 확대
- SDDC 기반 컨테이너 보안 강화 클라우드 인프라 서비스 제공
- Cloud Native 지원



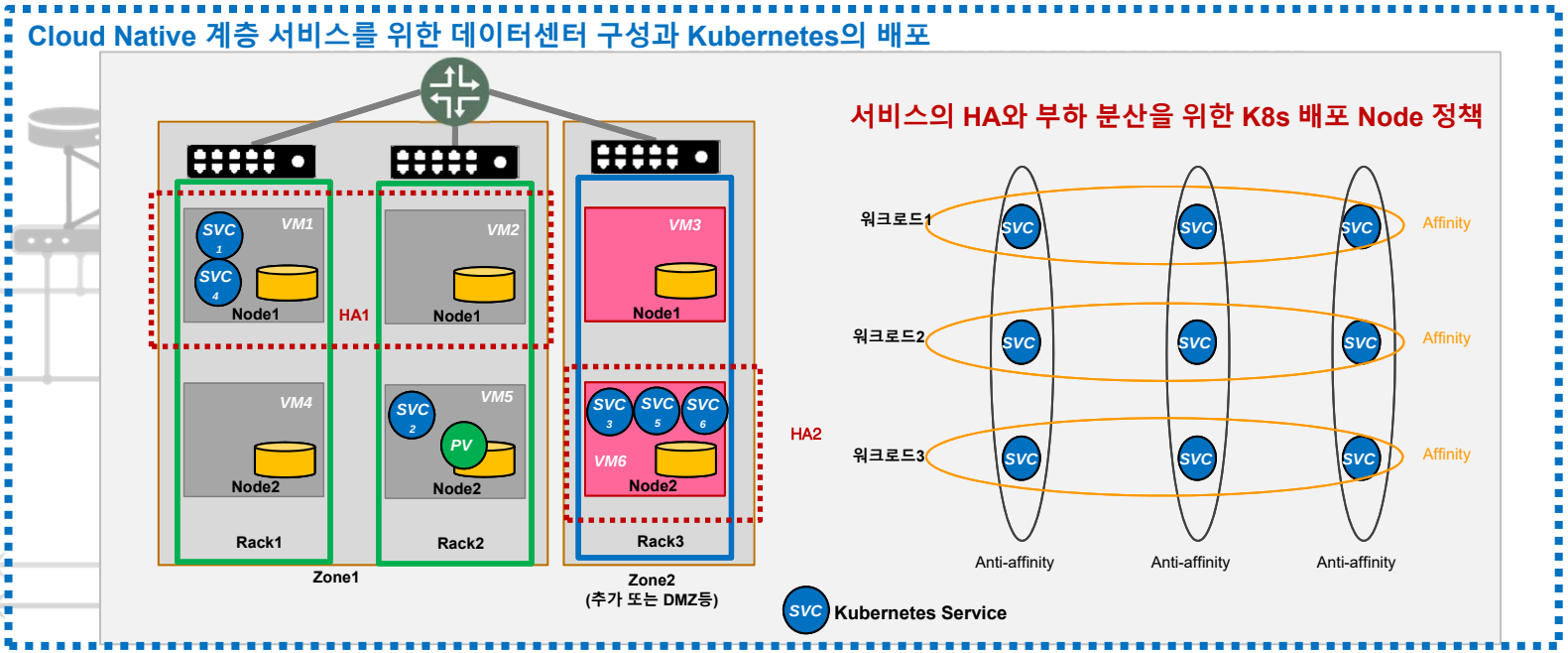
Cloud Native 계층 (1 of 3)

- Cloud Native 계층의 Kubernetes는 자원 배포 단위로 Pod를 사용하여 추상화
- Pod는 Container로 구성, 제조사는 GPU나 스토리지 추상화 지원 가능



Cloud Native 계층 (3 of 3)

- HA를 위한 동일 하드웨어/OS/컨테이너 사용 구조의 Rack 구성
- Kubernetes(K8s)의 추상화 PV(Persistent Volume)에 물리 스토리지 매핑
- K8s는 vGPU, vCPU 제공 VM 사용, 또는 BareMetal의 CPU/GPU 사용 배포

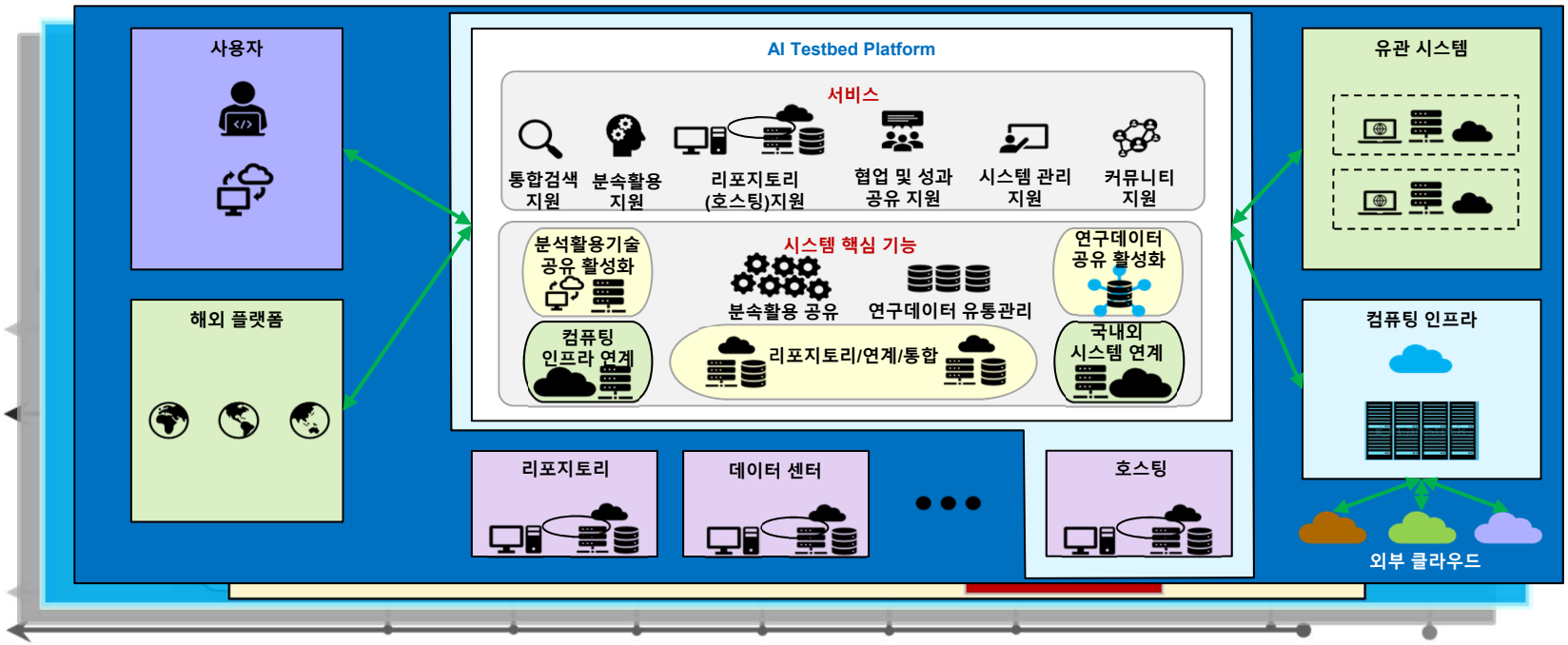


Cloud Native 계층 (3 of 3)

- 클라우드 네이티브화 지원/확장 플랫폼의 ML/DL/AI 서비스
  - 다양한 종류의 라이브러리 지원
    - TensorFlow, PyTorch, Caffe 등
  - 버전 별 라이브러리 동시 지원
    - TensorFlow 등의 버전 별 동시 운영 지원
    - PyTorch와 TensorFlow 등이 간섭 없이 동일 시스템에서 지원
  - 연산 자원 (CPU/GPU/RAM) 동적 할당
    - 요청 영역별 연산 자원 할당 (예: 4CPU + 2GPU + 64GB)
    - GPU 자원 할당 지원
    - GPU 램 파티셔닝 지원 (TensorFlow등)
  - GPU등 전문 하드웨어 지원 확장
    - Nvidia CUDA 기반 가속 (TensorFlow / PyTorch+Caffe)
    - AMD ROCm 기반 가속 (TensorFlow)
  - 라이브러리 자동 업데이트 지원

서비스 계층

- 산/학/연 협동 연구와 상용화를 위한 초고속 AI Testbed Platform 서비스 제공
- 연구/개발에 집중 가능한 ML/DL/AI 자동화 제공 플랫폼 서비스

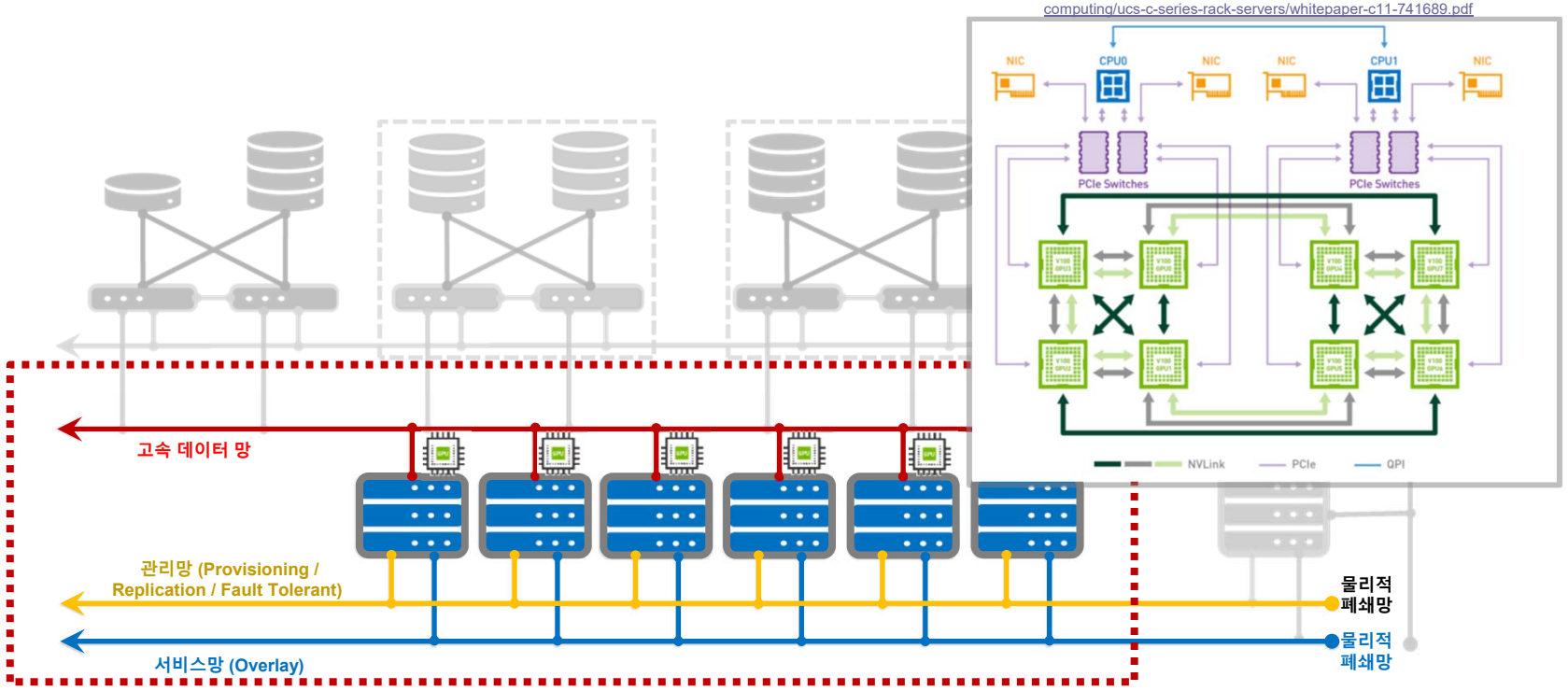


JS Lab

GPU Clustering 구성

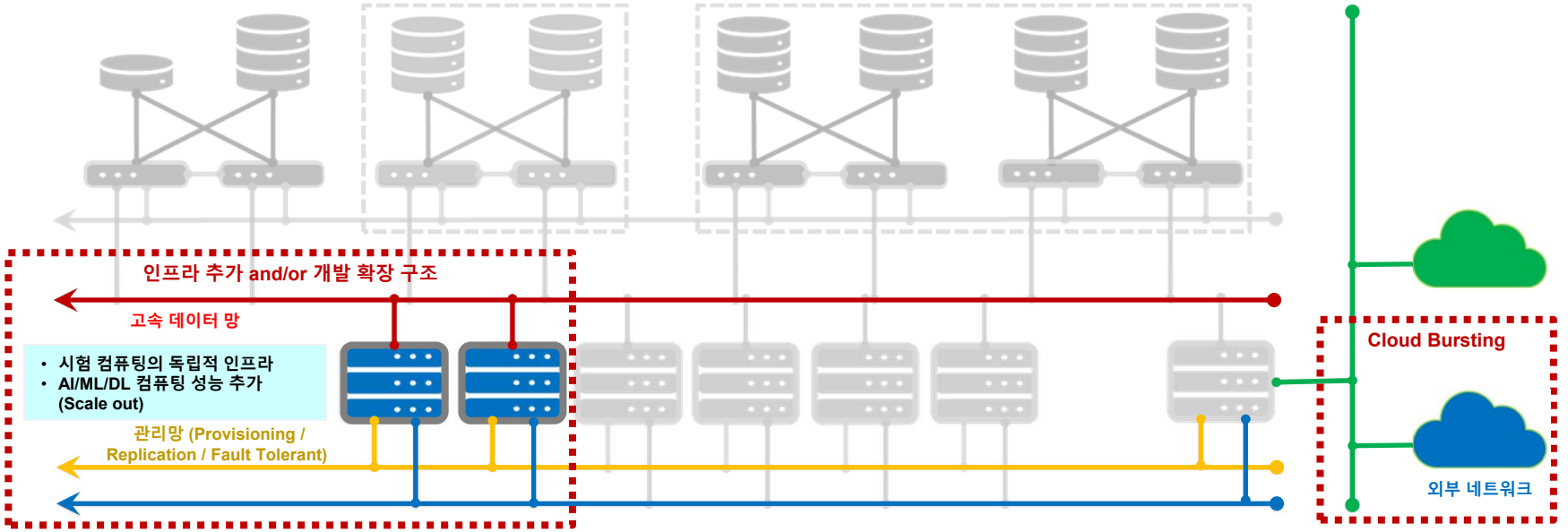
- 컴퓨팅 노드 내의 GPU 간 단축 경로 제공(CPU Bypass)
- GPU 클러스터 내의 컴퓨팅 노드간 단축 경로 제공(Offload)
- 컴퓨팅 노드 내의 구성은 전용 하드웨어 제조사 선택에 따라 다를 수 있음

https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/whitepaper-c11-741689.pdf



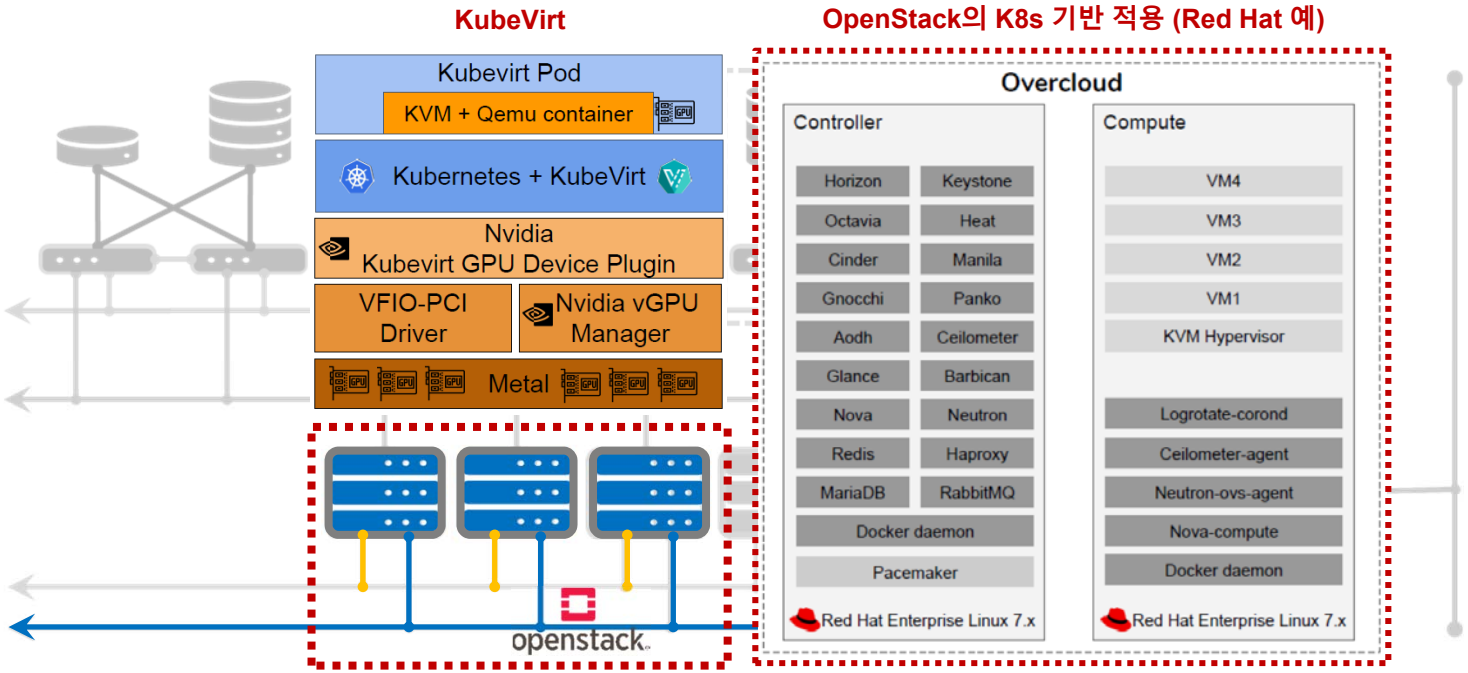
성능 확장 인프라

- Cloud Native Architecture(CNA) 기반 Scale out 확장 가능 구조
- 운영과 독립적인 ML/DL/AI 플랫폼 연구/개발 확장 환경
- Cloud bursting 확장



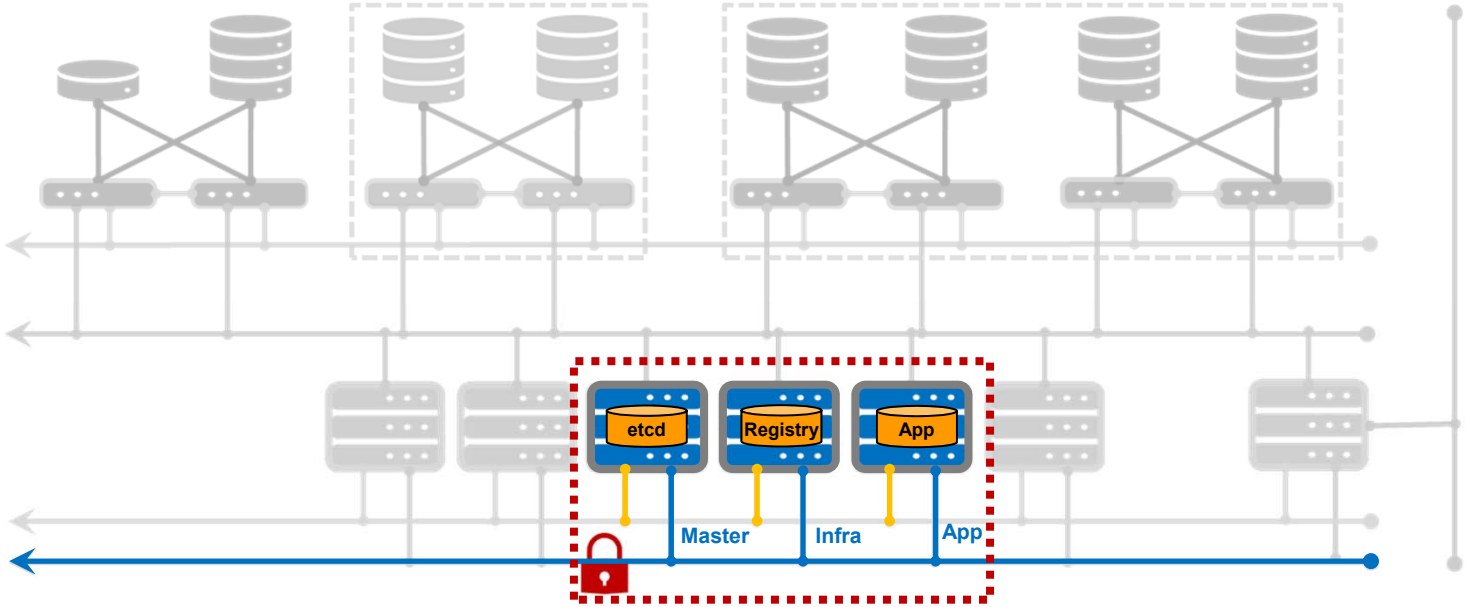
성능 확장 인프라

- **Kubernetes 상의 OpenStack 제어 기능 제공**
- **Kubernetes Clustering VM Node 배포** (OpenStack 또는 KubeVirt 사용 가능)
- **OpenStack 제어기능 들의 Packaging 설치를 위한 Helm 사용 가능**



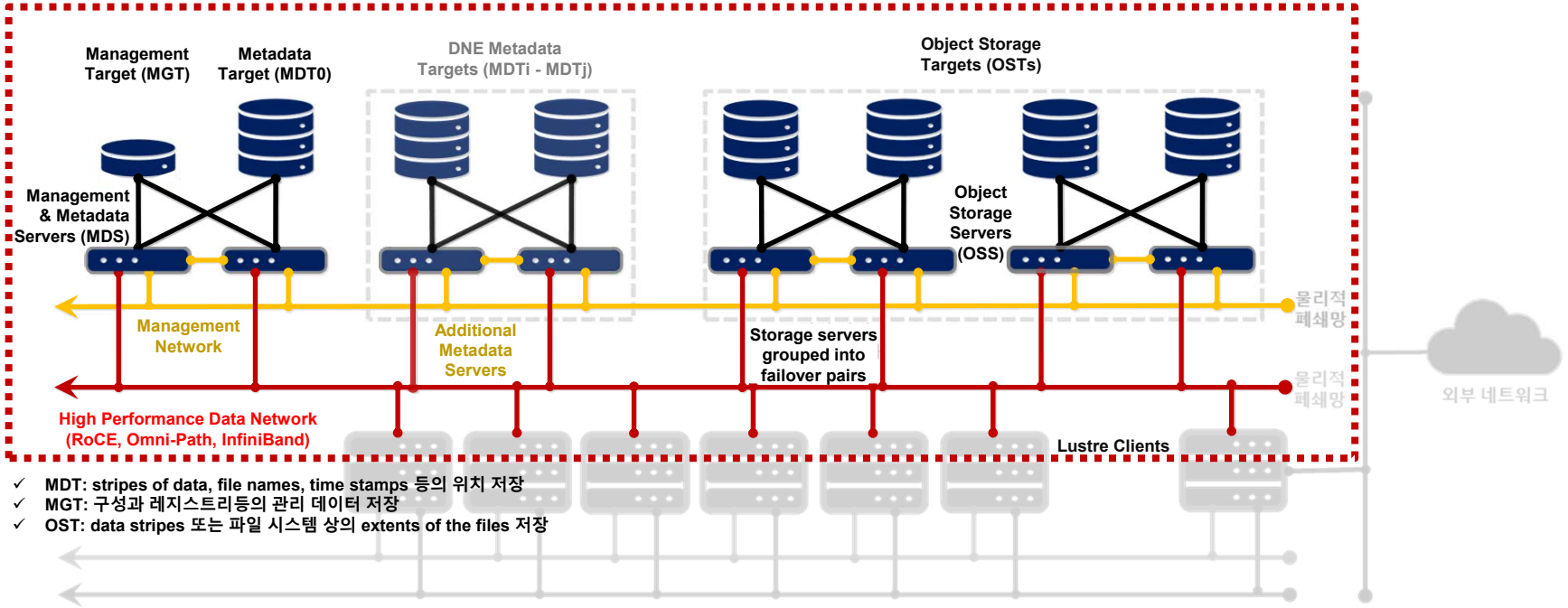
컨테이너 기반 보안 강화 플랫폼 구성

- Cloud Native Architecture(CNA)를 위한 개발/적용 환경 플랫폼 구성
- 클라우드 주요 요소 3중화 노드의 보안 관리와 네트워크의 노출 정책
- **'etcd'용 Master Node** (x3), **'Registry'용 Infrastructure Node** (x3), **'App'용 Application Node** (x3)



Lustre Storage 구성

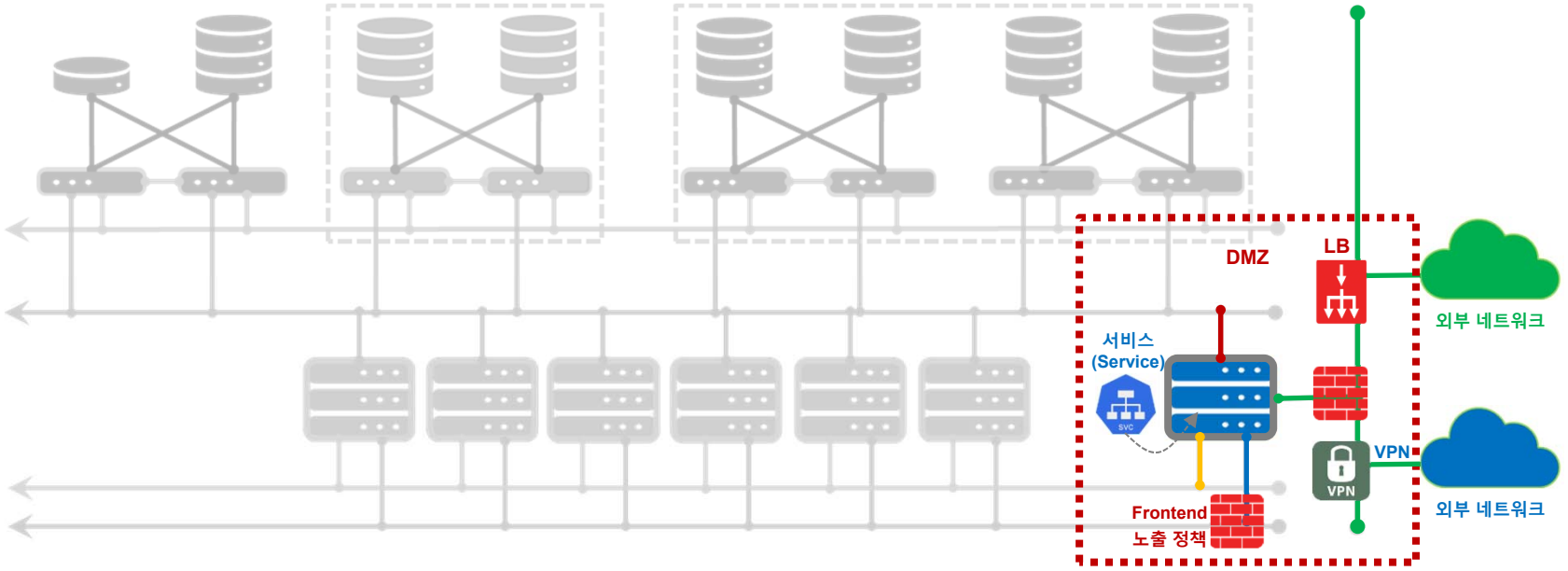
- 스토리지 영역을 위한 초고속 빅데이터 처리 구성
- Clients: 파일 위치 결정을 위한 MDS 접속과 데이터 R/W를 위해 OSS에 접속



- ✓ MDT: stripes of data, file names, time stamps 등의 위치 저장
- ✓ MGT: 구성과 레지스트리등의 관리 데이터 저장
- ✓ OST: data stripes 또는 파일 시스템 상의 extents of the files 저장

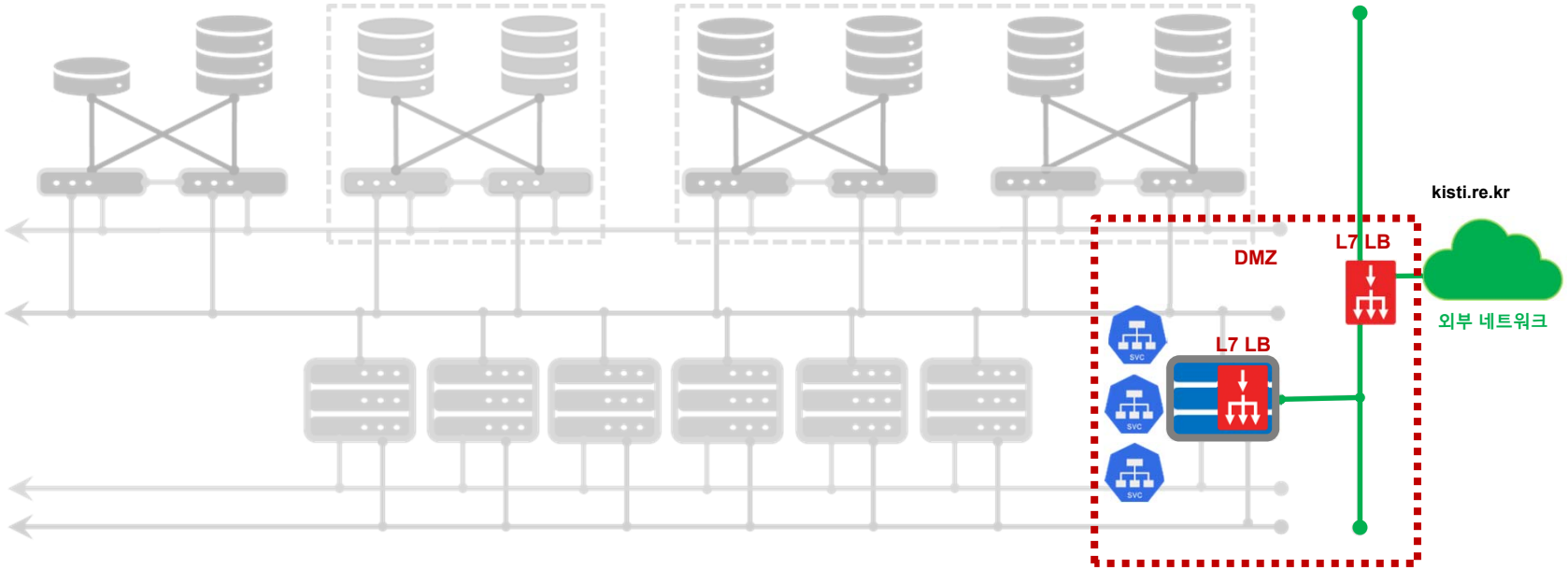
DMZ 구성

- 클라우드 포털 서비스의 가속 처리와 보안 강화 DMZ 구성을 위한 LB(Load Balancer)와 방화벽
- 멀티클라우드 수용 서비스 성능 확장 구조 (파트너 망 수용을 위한 VPN 연결)



L7 LB(Load Balancer) 구성

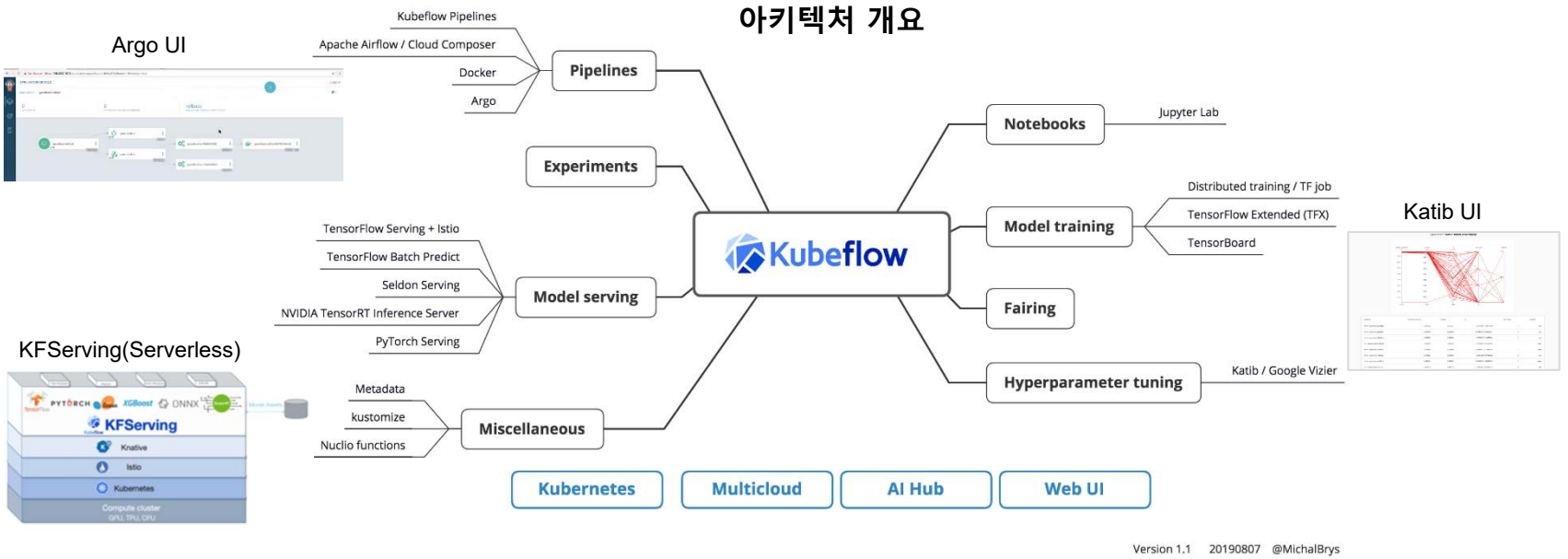
- Hash 기반 정책: Source IP/Port, Destination IP/Port, Cookie, Protocol
- 외부 요청 URL 주소를 클라우드 서비스 내부의 지정 Private IP에 연결
- L7 LB 제공 위치: Hardware , 서버의 IaaS VM, 서버의 Cloud Native Ingress



쿠베플로우 (1 of 2)

쿠베플로우(Kubeflow) 수용 오픈소스 프로젝트

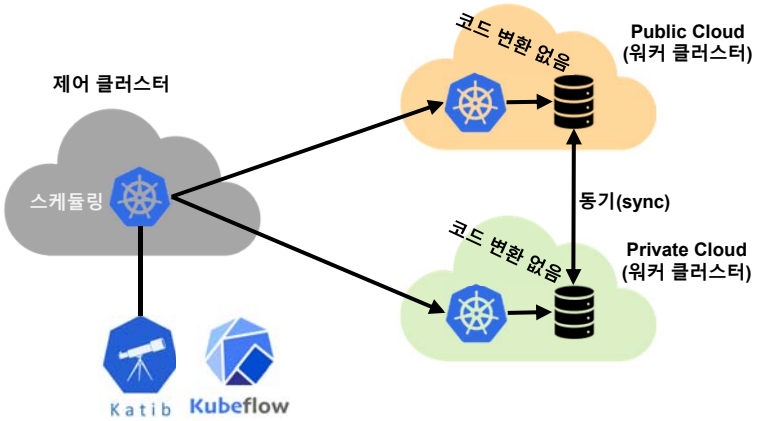
- Jupyter Notebook: Easy GPU provisioning
- Katib: 하이퍼파라미터 튜닝
- Argo: Pipeline / ML Workflow 작성 (자동화)
- Fairing: building, training, and deploying ML models in a hybrid cloud
- TFJob, PyTorchJob: CRD
- Seldon, TF Serving: Model serving



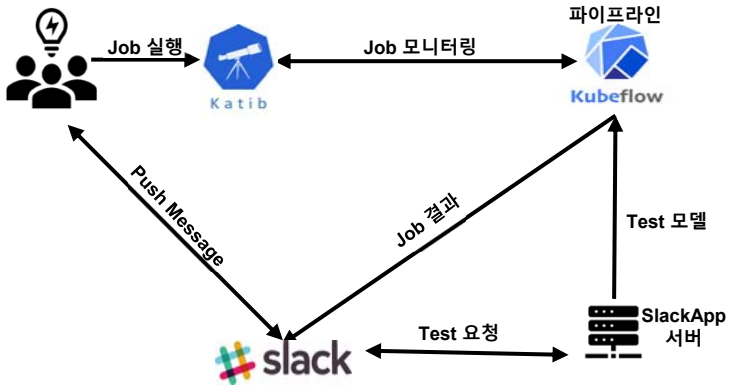
쿠베플로우 (2 of 2)

- **Cloud Bursting 지원 운영/관리**
  - 제어 및 관리, 서비스 디스커버리, 로드 밸런싱, 서비스 호출/응답 관련
- **Recurring Run (반복 실행) 서비스**
  - Kubernetes 기반 Katib, Kubeflow
  - Job 실행/모니터링/결과, 데스트 요청/모델 실행
  - 엔터프라이즈급 메신저 Slack 이용

Cloud Bursting 지원



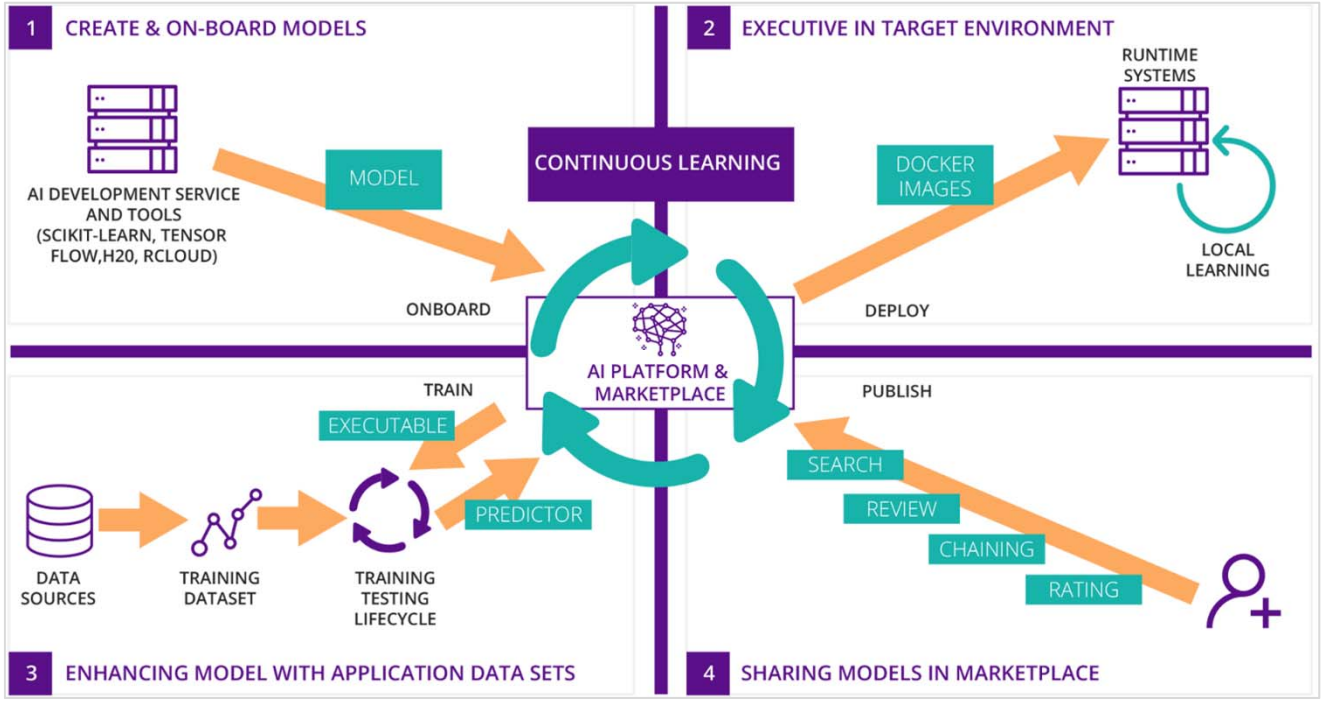
Recurring Run (반복 실행) 서비스



Acumos AI - LFAI

Acumos AI :

- 홈페이지: https://www.acumos.org/
- LFAI 내에서 졸업한(Graduated) 프로젝트
- 주요 멤버는 통신사와 서비스사를 포함한 관련 제조사
- 모델을 생성하여 지속적인 학습 결과를 시스템에 적용하는 프레임워크 제공



감사합니다