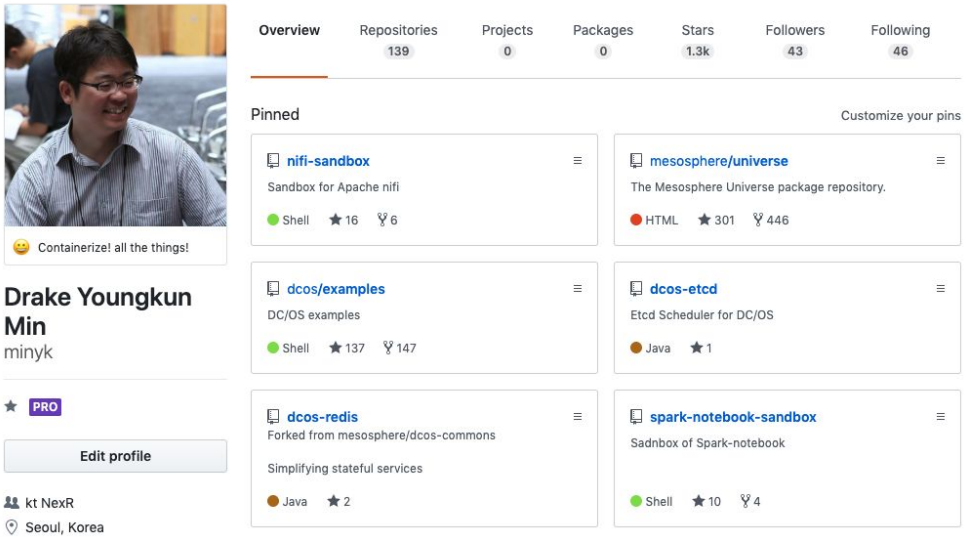
# Spark Everywhere: DC/OS & K8S

민영근 박사, DE팀

# 민영근

- 2019.4 ~ kt NexR
- 2016.6 ~ 2019.4 AJ IT센터
- 2013.8 ~ 2016.6 kt NexR
- 2011.2 단국대학교 공학박사

https://github.com/minyk



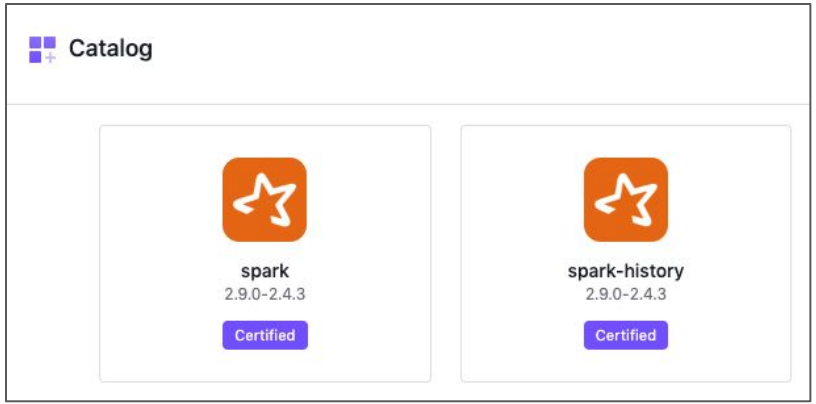
# Spark on DC/OS

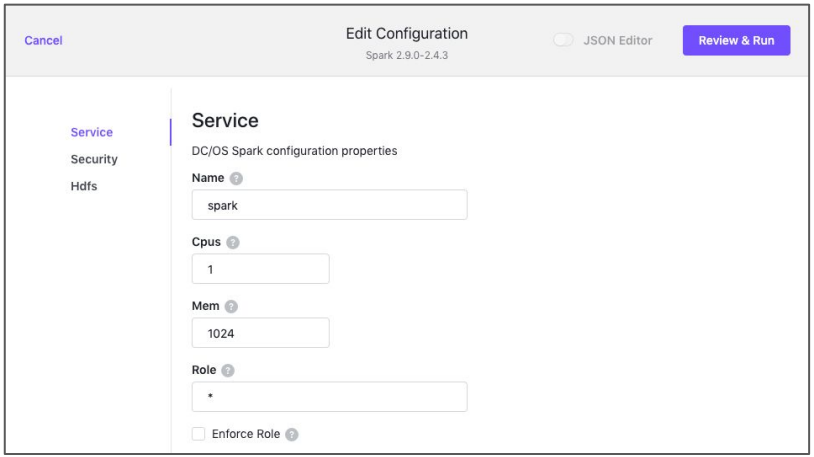
# D2IQ DC/OS

- D2IQ(전 Mesosphere 2013~), Open Source + Enterprise
- Apache Mesos(2011~) + Mesosphere Marathon(2013~)
  - Mesos-DNS, DC/OS-Net 등 30여가지 컴포넌트로 구성
- 규격화된 로깅, 메트릭 제공
- 서비스 카탈로그 제공
  - 배포하기 쉽도록 패키징된 형태
  - Web UI를 사용, 배포

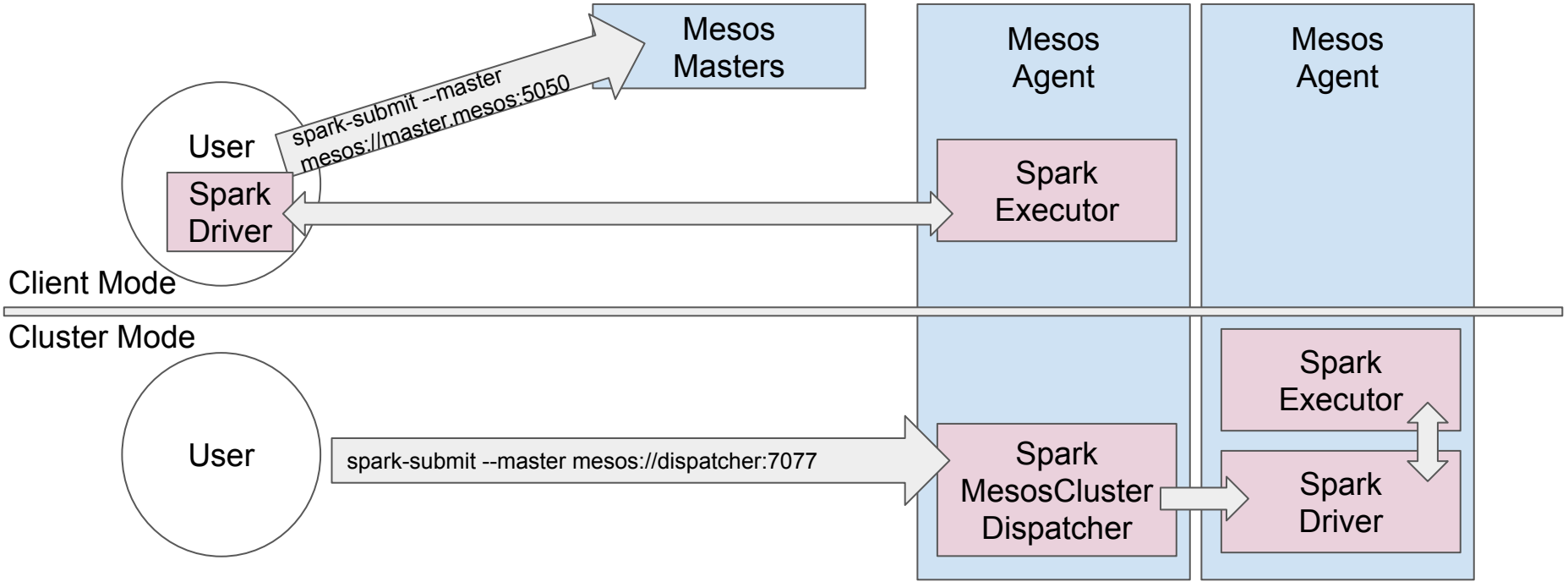
# Spark on DC/OS

- Spark on Mesos + 알파
- DC/OS Catalog 기능으로 쉽게 배포
- 동시에 다른 버전의 스파크 사용 가능
- CLI 도구 제공
- DC/OS 내의 Mesos 사용

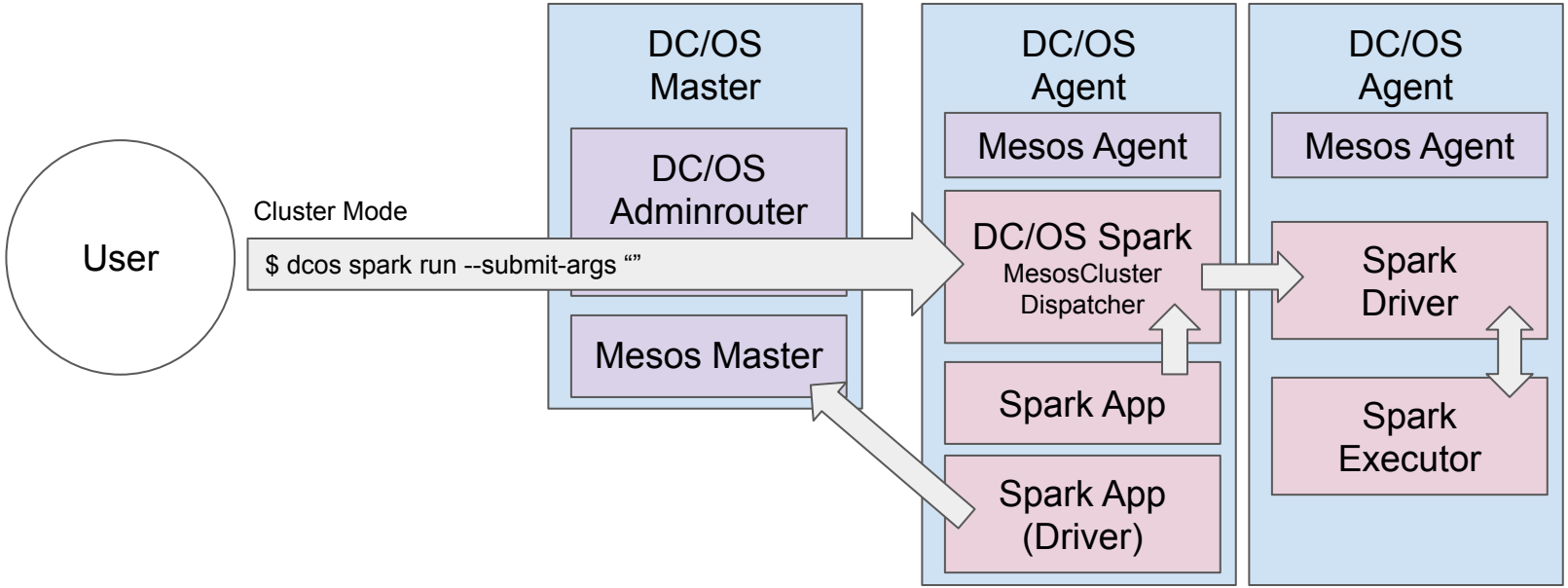


# Spark on Mesos 구조



# Spark on DC/OS 구조



# DC/OS Spark 자원 할당

- Coarse-Grained 사용 (정적 할당)
  - executor의 수를 spark.cores.max / spark.executor.cores 로 결정
  - spark.cores.max, spark.executor.cores를 --submit-args 로 전달

- 동적할당 사용 가능
  - spark-shuffle 서비스 설치
  - 네트워크를 “host” 로 사용: executor와 shuffle을 같은 IP로 접근
  - executor의 임시 디렉터리를 spark-shuffle 과 공유
  - spark.dynamicAllocation.maxExecutors 값까지 확장

# DC/OS Spark Image

- D2IQ Mesosphere Spark
  - 코드: https://github.com/mesosphere/spark/
  - custom-master, custom-branch-2.4.3 등
  - Apache Spark + DC/OS용 개선 코드

- D2IQ Mesosphere Spark Image
  - 코드: https://github.com/mesosphere/spark-build/
  - DC/OS Catalog 서비스에서 배포할 수 있도록 이미지 빌드

# D2IQ Mesosphere Spark 개선점들

- DCOS-45850: CNI support for docker containerizer
- DCOS-39150: Unique Executor ID
- DCOS-46389: Driver/Executor on the same virtual network
- DCOS-46585: Fix supervised driver retry logic
- DCOS-40974: Mesos checkpointing for spark driver
- DCOS-58386 Node draining support for supervised drivers

# D2IQ Mesosphere Spark Image

- DCOS-49019: Use DC/OS bootstrap for IP resolution
- COPS-3550, DCOS-39751: Named VIP fix
- DCOS-58390: Quota enforcement support
- DCOS-53535: StatsD Metrics Reporter for Spark
- DCOS-54557: Added SPARK_APPLICATION_ORIGIN and SPARK_INSTANCE_TYPE env variables propagation to StatsD Sink
- DCOS-52812: Replace Oracle Java with OpenJDK

# Spark on DC/OS: Logging & Metrics

- DC/OS의 규격화된 방법으로 제공
- 메트릭
  - UCR 사용 시 DC/OS Metrics API 를 통해서 수집 가능
- 로그
  - DC/OS Logging API 를 통해서 수집 가능
  - History Server 사용

# Spark on DC/OS

- Mesos와는 처음부터 절친
  - 초기 개발진이 중복
- D2iQ 의 개선점들이 사용을 편리하게 함
  - DC/OS CLI 플러그인 제공
  - Metrics/Logging
- 카탈로그에서 제공되는 다른 서비스들과 통합
  - HDFS
  - History Server
  - Kafka

# Spark on Kubernetes

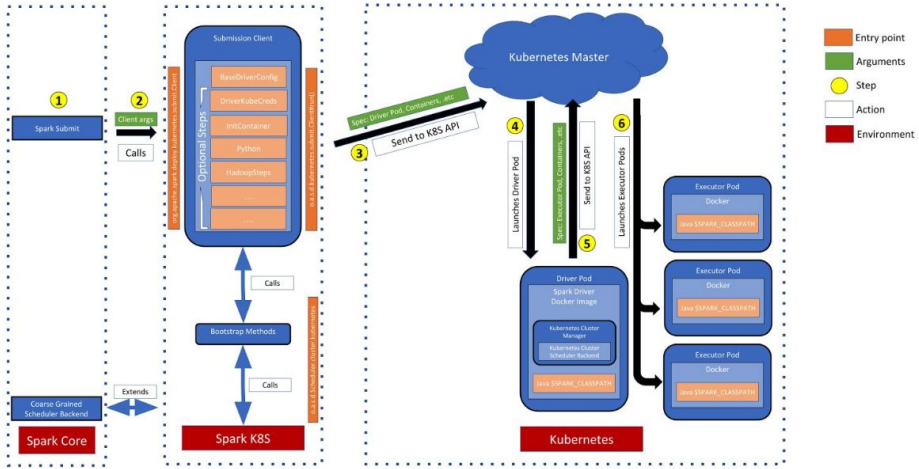
# Kubernetes

- Google에서 시작(2014)/기증(2015), 현재는 CNCF 주도
- 컨테이너 관리의 (거의) 업계 표준
- 다양한 Public/Private 서비스
  - EKS, GKE, AKS 등
  - Openshift, konvoy 등
- 높은 확장 가능성
  - CRD, Custom Controller
  - 핵심 서비스(apiserver/scheduler)의 확장 가능: webhook, scheduler extension
  - 네트워크/저장소 플러그인
- 100% Open Source

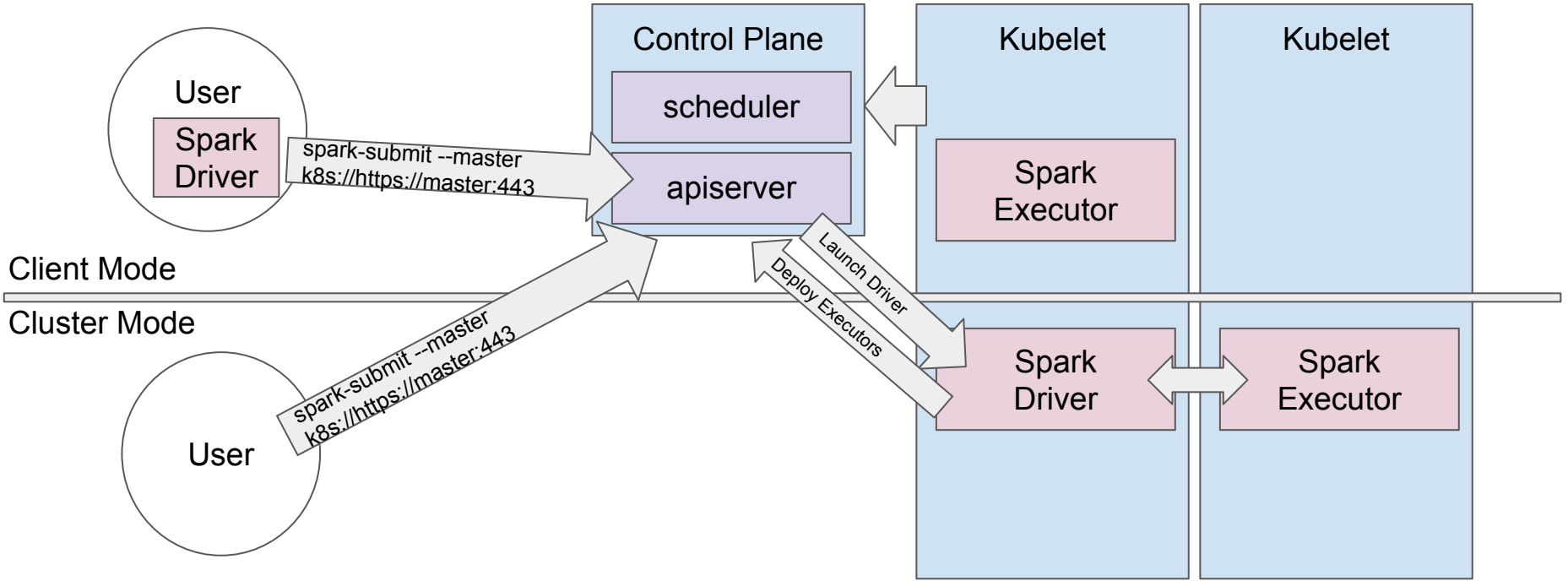
# Spark on Kubernetes

- 2016.11~2018.3 진행
  - Support Spark natively in Kubernetes #34377
  - SPARK-18278 SPIP: Support native submission of spark jobs to a kubernetes cluster
- 별도의 repo에서 개발되어 2.3.0에 병합

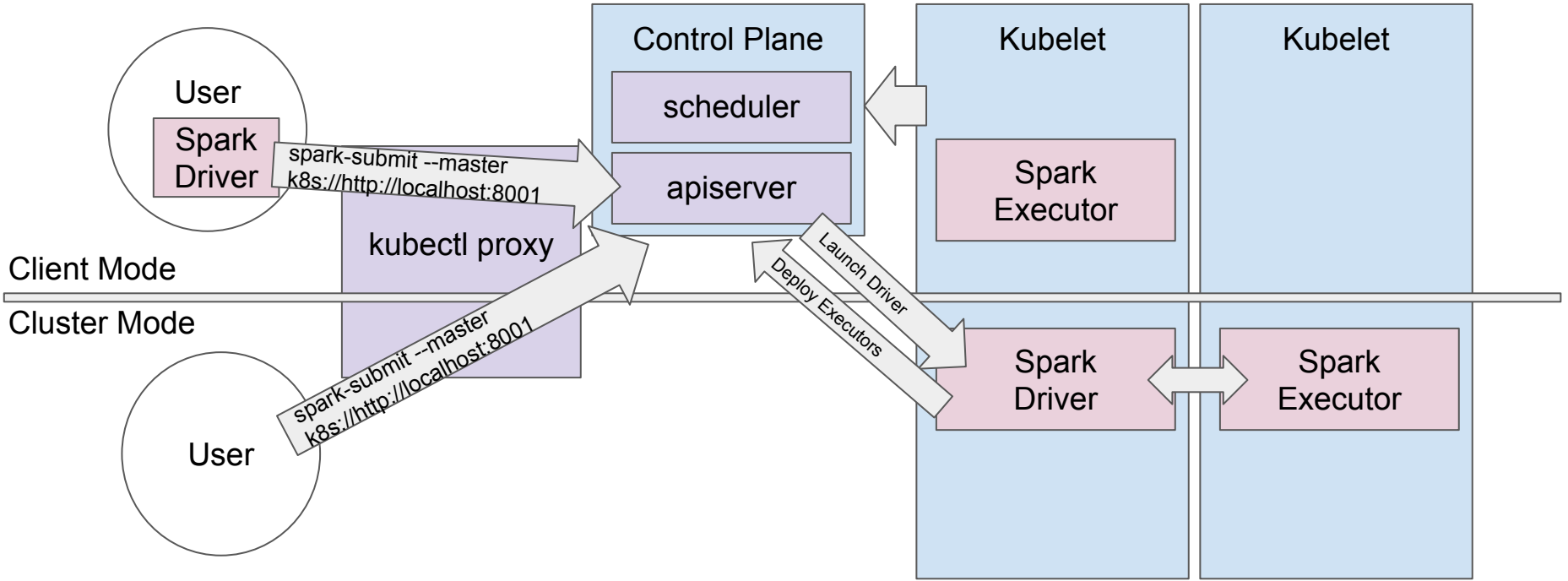
Summary Architecture Diagram



# Spark on Kubernetes 구조



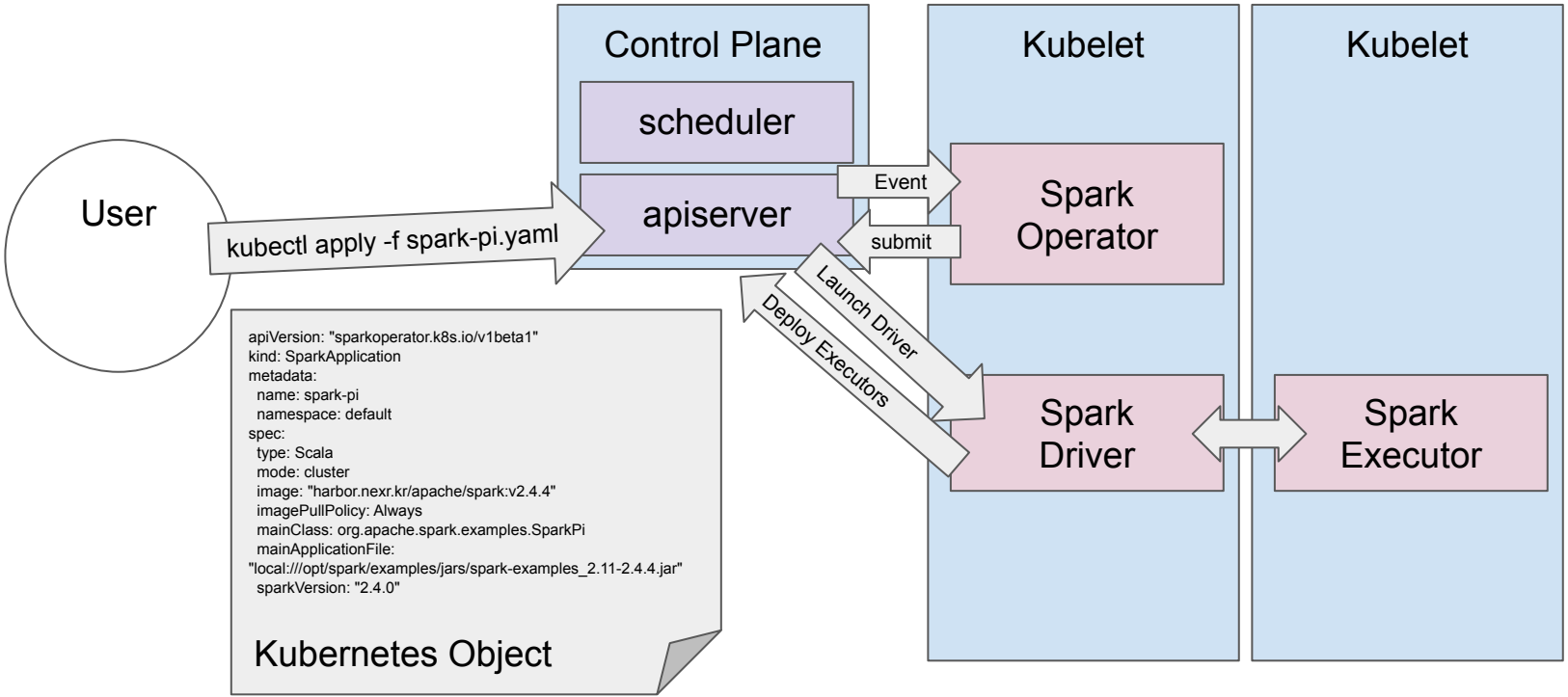
# Spark on Kubernetes 구조



# K8S Spark Operator

- googleCloudPlatform/spark-on-k8s-operator
- spark-submit 이 아니라 kubectl 을 사용
- Spark Job을 정해진 YAML 형식으로 작성
- Spark Operator가 대신 submit을 수행
- 스케쥴러등 부가 기능을 사용할 수 있음

# Spark Operator 구조



# K8S Spark Advanced Scheduling

- kubernetes-sigs/kube-batch
  - K8S 커뮤니티 개발
  - 우선순위, 갱, DRF 스케쥴링
- volcano-sh/volcano
  - 화웨이 주도, kube-batch 기반
  - GPU 스케쥴링, Job 관리
- palantir/k8s-spark-scheduler
  - 갱 스케쥴링

# K8S Spark: volcano-sh/volcano

- 스케쥴러 확장으로 설치
- kubernetes-sigs/kube-batch + 추가 기능 제공
  - GPU 스케쥴링
  - Job Queue 관리
  - Singularity CRI
- 중요 오퍼레이터에서 지원
  - kubeflow/tf-operator
  - googleCloudPlatform/spark-on-k8s-operator

# 번외: hub.helm.sh/charts/microsoft/spark

- 구성
  - Spark 단독 클러스터: 1 Master, 3 Workers
  - Zeppelin: 단독 클러스터 사용
  - Livy: 단독 클러스터 사용
- 두 가지 방법으로 Spark 작업 실행
  - k8s apiserver 직접 호출: --master k8s://https://apiserver:443
  - zeppelin/ livy: 같이 설치된 단독 클러스터 사용
- 자동 확장 기능 제공
  - Spark Worker Pod에 대해서 K8S의 HorizontalPodAutoscaler 객체 적용
  - CPU 평균 사용률

# Spark on K8S: Logging & Metrics

- 규격화된 방법은 없음.
- 메트릭
  - operator를 사용하면 jmx-to-prometheus 로 수집 가능
  - operator를 사용하지 않으면 별도 플러그인을 직접 설정
- 로그
  - K8S Logging Operator 사용
  - History-Server 사용

# Spark on Kubernetes

- 새로운 친구
  - Spark 2.3.0에서 실험적 기능으로 처음 소개
  - 커뮤니티간의 불협화음
- 아직 해결해야 할 문제들이 있음
  - SPARK-24434 Support user-specified driver and executor pod templates
  - SPARK-24432 Add support for dynamic resource allocation
  - Cgroup leaking, no space left on /sys/fs/cgroup #70324
- 주목해야 할만한 진행 사항
  - Spark 3.0, 2020년 초(?)
  - Apache Bigtop BIGTOP-3225 Cloud Native Bigtop

끝내며



kubernetes

# Q&A