네이버 스케일로 카프카 컨슈머 사용하기

이동진
NAVER Platform Labs

NAVER DEVIEW 2023

CONTENTS

About the Speaker (1)

- Naver, Platform Labs 소속
- 사내 Kafka 서비스 개발
  - Kafka 사용 관련된 문의 대응 및 troubleshooting
  - 내부 배포판 개발
    - Navercorp Kafka
    - Navercorp Cruise Control

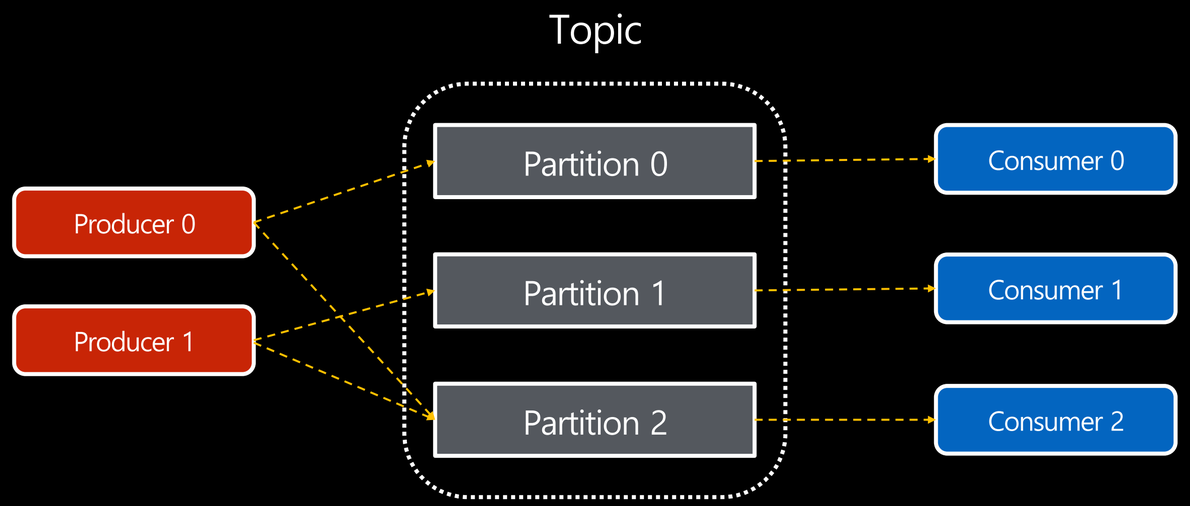
About the Speaker (2)

- Committer, Apache Software Foundation
  - Hadoop, Giraph, Hbase, Spark, Kafka, …
- Apache Kafka Contributor
  - 압축 관련 기능 개선 (KIP-110, KIP-390, KIP-780)
  - Log4j2 마이그레이션 (KIP-653, KIP-719)
  - Spark - Kafka Record Header 연동 기능 (SPARK-23539)
  - 그리고 그리고 …
- Kafka: the Definitive Guide 제 2판 역자

1. Kafka Consumer 동작 원리

1.1 Kafka Consumer: 초간단 소개

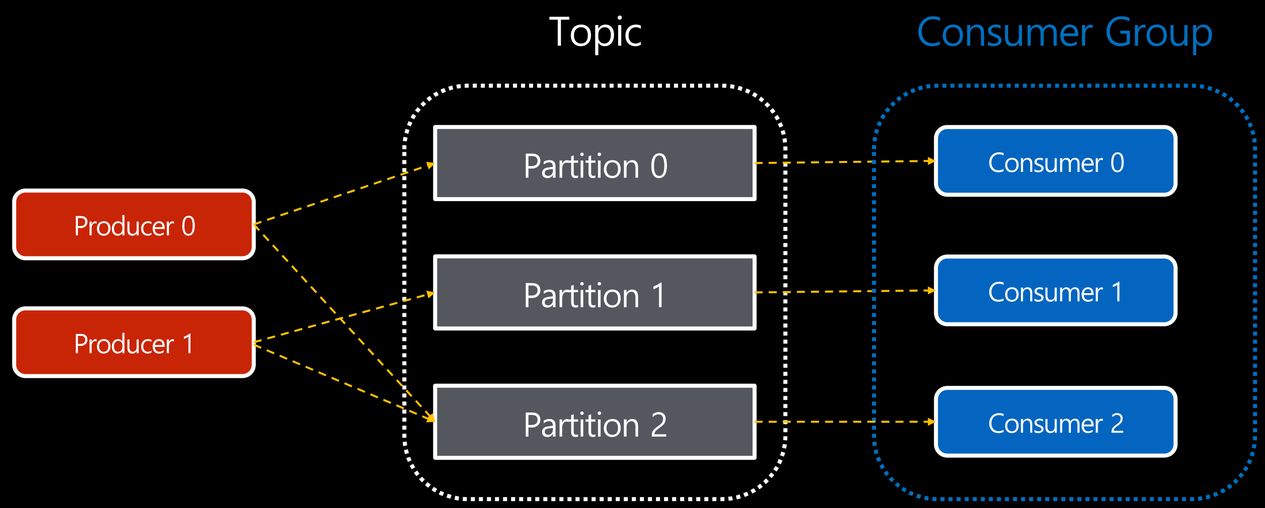
NAVER
DEVIEW
2023



- Topic
  - 1개 이상의 Partition으로 분할, 1개 이상의 Replica로 복제된 log 자료 구조
- Client
  - Producer: 쓰고자 하는 Topic Partition의 맨 끝에 record를 추가
  - Consumer: 읽어오고자 하는 Topic의 Partition에 저장된 record를 순차적으로 읽어 옴

1.2 Consumer Group: 초간단 소개 (1)

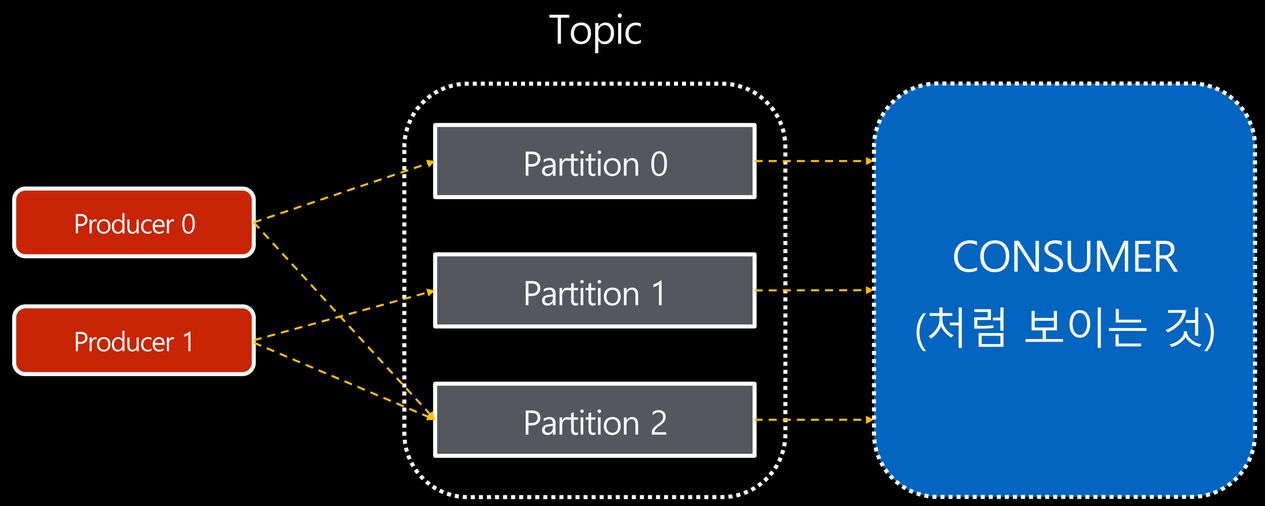
NAVER
DEVIEW
2023



- Consumer Group
  - 같은 'group.id' 설정값을 가진 Consumer들은 하나의 Consumer Group을 이룬다.
  - 같은 Consumer Group에 속한 Consumer들이 Topic에 속한 Partition들을 나눠서 읽는다.

1.3 Consumer Group: 초간단 소개 (2)

NAVER
DEVIEW
2023



- Consumer Group == “논리적인 Consumer”
- 거대한 Consumer 하나가 전체 Topic 내용을 읽어들이고 있는 것처럼 보인다.

1.4 Consumer Group에 필요한 것은?

NAVER
DEVIEW
2023

질문:

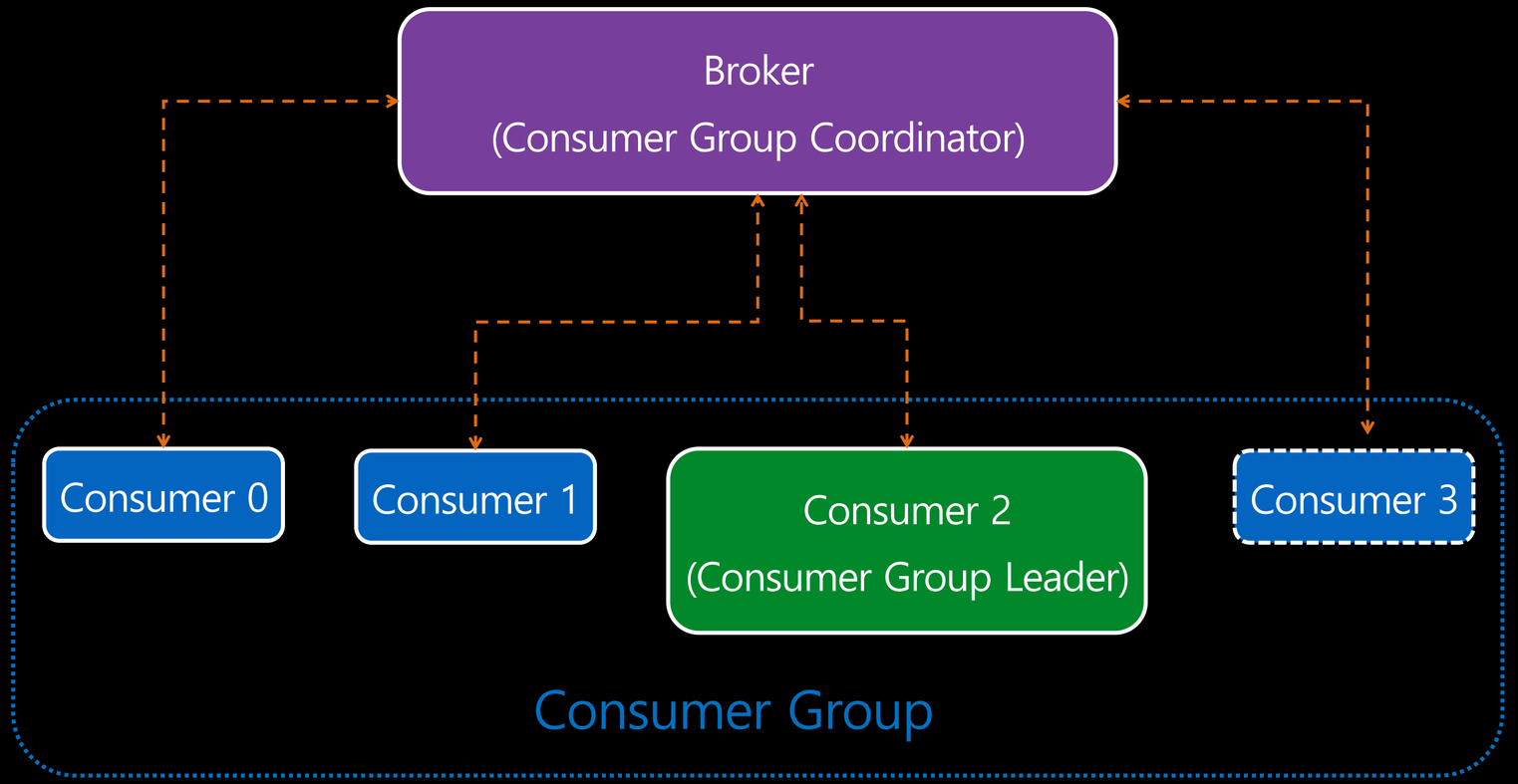
Consumer Group 기능이 제대로 동작하기 위해 필요한 것은?

정답:

1. Partition Assignment 기능
2. Offset Commit 기능

1.5 Consumer Coordination: 동작 원리 (1)

NAVER DEVIEW 2023



- Consumer Group Coordinator

  1. Consumer Group에 변경이 생겼는지 탐지

  2. TopicPartition에 변경이 생겼는지 탐지

  3. Consumer Group Leader와 나머지 Consumer들간의 communication 중개

- Consumer Group Leader

  - 현재 구독중인 topic의 파티션들을 consumer들에 할당

1.6 Consumer Coordination: 동작 원리 (2)

NAVER
DEVIEW
2023

- Q: 왜 이런 복잡한 구조를 택했나요?
- A: “Broker를 재시작할 필요 없이 더 유연하고 확장 가능한 파티션 할당을 지원하기 위해.” (출처)

1.7 Consumer Coordination: 관련 설정

NAVER
DEVIEW
2023

- max.poll.interval.ms
- session.timeout.ms
  - heartbeat.interval.ms
- partition.assignment.strategy
  - List of org.apache.kafka.clients.consumer.ConsumerPartitionAssignor class

1.8 partition.assignment.strategy: 동작 원리 (1)

NAVER DEVIEW 2023

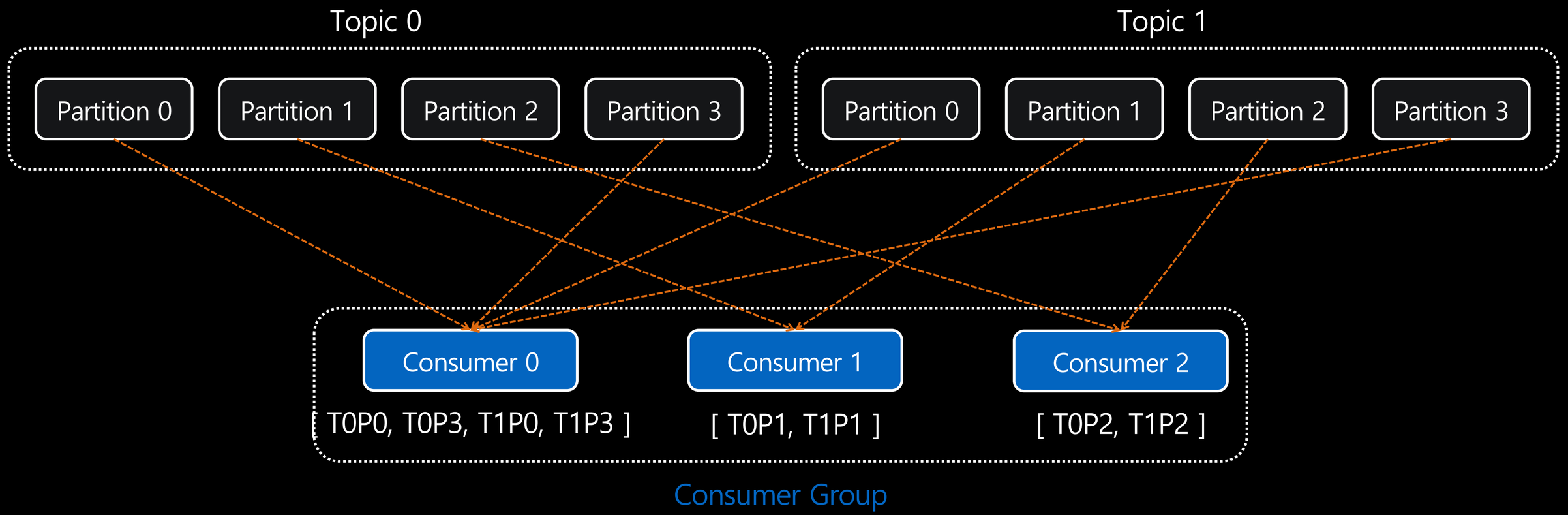
```
partition.assignment.strategy = [
    org.apache.kafka.clients.consumer.RangeAssignor.class,
    org.apache.kafka.clients.consumer.CooperativeStickyAssignor.class
]
```

1. Consumer Group에 참여한 모든 Consumer에 공통으로 설정된 Assignor 중에서
2. 우선순위가 가장 높은 것이 파티션 할당 전략으로 선택

1.9 partition.assignment.strategy: 동작 원리 (2)

NAVER DEVIEW 2023

- 예: RangeAssignor



2. Cloud 환경에서 Kafka Consumer 사용하기

2.1 Cloud 환경에서의 Consumer Group Coordination

NAVER
DEVIEW
2023

- 물리적 장비의 자원을 여러 pod가 나눠서 씀 (multitenancy)
  - “Noisy Neighbors” 현상
  - Network Hiccup
- Pod Rescheduling이 일상적임

2.2 새 설정: group.instance.id

NAVER
DEVIEW
2023

- “정적 그룹 멤버십” (2.3, KIP-345)
- 정상적 재시작 이전에 할당되어 있던 파티션들을 다시 할당
  - 같은 group.instance.id 설정을 가진 기존 컨슈머의 할당을 승계
  - Rebalance가 발생하지 않음
- 단순 pod 재시작 때문에 Partition Rebalance가 발생하는 사태를 방지
  - Kafka Streams가 내부적으로 이 설정을 사용

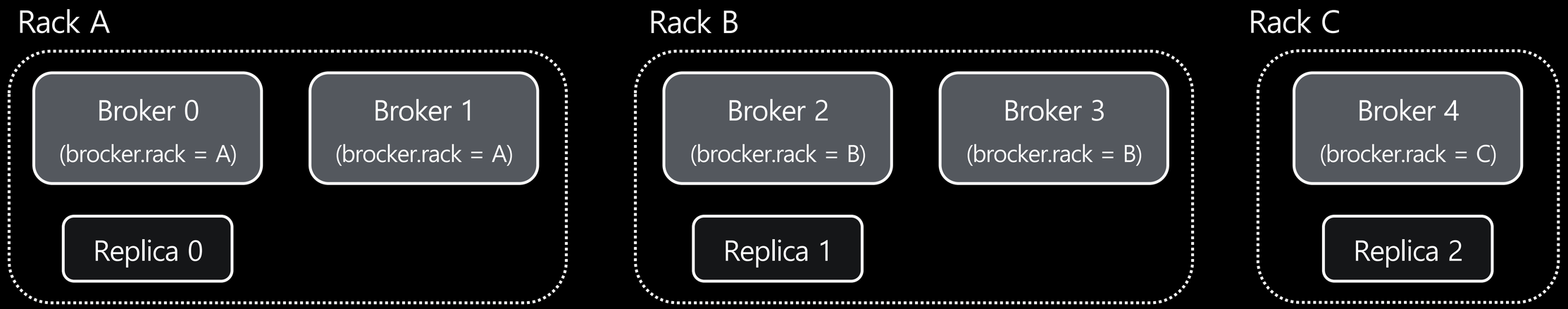
2.3 설정 변경: session.timeout.ms

NAVER DEVIEW 2023

- “Consumer 프로세스가 Broker와 신호를 주고받지 않고도 리밸런스를 발생시키지 않는 최대 시간”
- 기본값 변경
  - 3.0 이전: 10초 (10000)
  - 3.0 이후: 45초 (45000)
- 단순 network hiccup 때문에 Partition Rebalance가 발생하는 사태를 방지
  - Consumer 프로세스가 죽었는지 알아차리는 데 걸리는 시간은 증가

2.4 새 기능: Follower Replica로부터 읽기 (1)

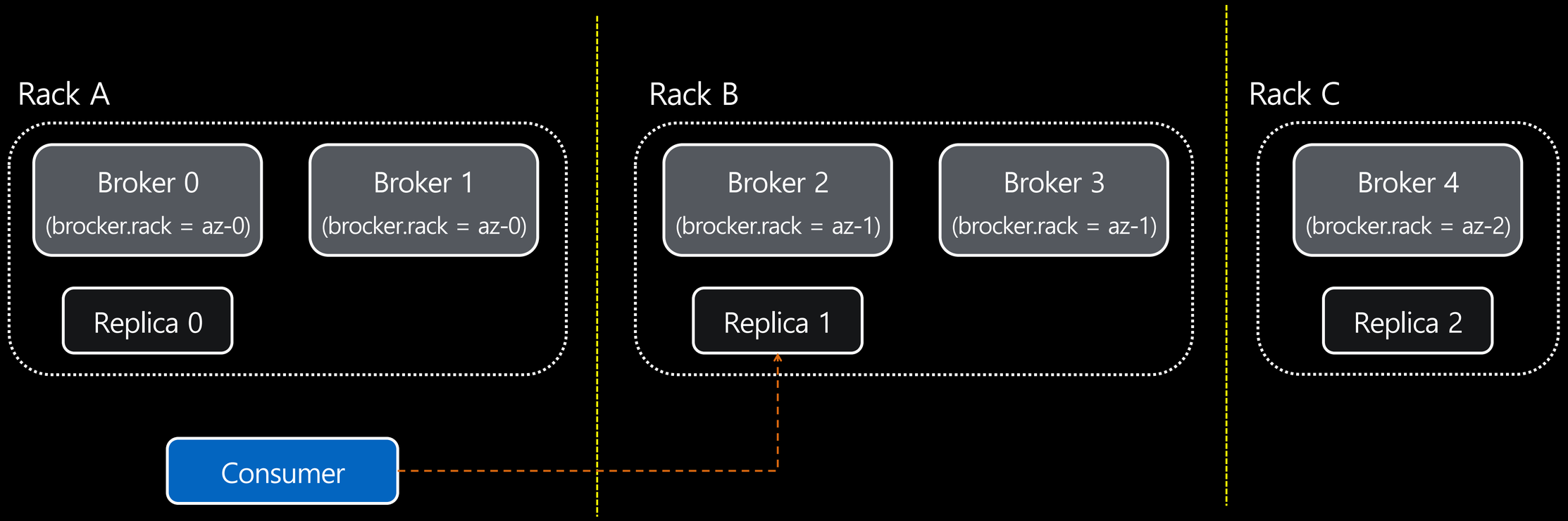
NAVER DEVIEW 2023



- broker.rack 설정 (broker 설정)
  - 새로 생성된 replica가 서로 다른 rack에 할당되도록 하기 위해 도입
  - "서버 랙 전체에 전력이 나가버리더라도 다수의 replica가 동시에 동작 불능에 빠지지는 않는다!"
  - 물리적 서버 시대의 유산

2.5 새 기능: Follower Replica로부터 읽기 (2)

NAVER DEVIEW 2023



- 클라우드 시대로 옮겨오면서 의미 변화
  - 물리적 서버 랙 → 가용 영역 (Availability Zone)
- 문제: Consumer와 Leader Replica가 서로 다른 AZ에 있으면?

2.6 새 기능: Follower Replica로부터 읽기 (3)

NAVER DEVIEW 2023

- “Consumer가 위치한 AZ를 알고 있고 해당 AZ에 leader replica와 동기화된 상태를 유지하고 있는 follower replica가 있다면, 여기서 읽어올 수 있게 하자!”
  - 2.4부터 추가된 기능 (KIP-392)
- client.rack (consumer 설정)
  - 클라이언트가 위치한 AZ를 정의
- replica.selector.class (broker 설정)
  - leader replica가 아니라 같은 AZ에 위치한 follower replica로부터 읽어올 수 있도록 해 주는 설정
  - org.apache.kafka.common.replica.RackAwareReplicaSelector

2.7 이걸로 문제 끝...?

NAVER
DEVIEW
2023

- 미봉책
  - Broker 쪽 replica.selector.class 를 일일이 업데이트 해줘야 함
  - broker.rack에 들어가는 값이 적을 때만 원하는 대로 동작

3. 네이버 스케일로 Kafka Consumer 사용하기

3.1 네이버의 문제

NAVER
DEVIEW
2023

- 기본적으로 제공되는 기능만으로는 해결이 불가능하다!
  - 엄청나게 많은 Kafka Cluster 수
  - 몇 개인지도 모르는 Consumer (Group) 수
  - 무수히 많은 개발조직
    - Mission Critical…?
  - 거대한 규모의 Datacenter (들)
    - Network 크기? Rack 수? Traffic?

3.2 … 한걸음 뒤에서 바라보면?

NAVER
DEVIEW
2023

- 본질적인 원인
  1. "'Rack'이 가리키는 바가 지나치게 애매모호하다."
     - '데이터센터'? '물리적 서버 랙'?
     - 의미하는 바가 '네트워크' 보다는 '전력' 에 기울어짐
  2. "Partition Assignor가 rack 정보를 고려하지 않는다."

3.3 그러니까 해법은?

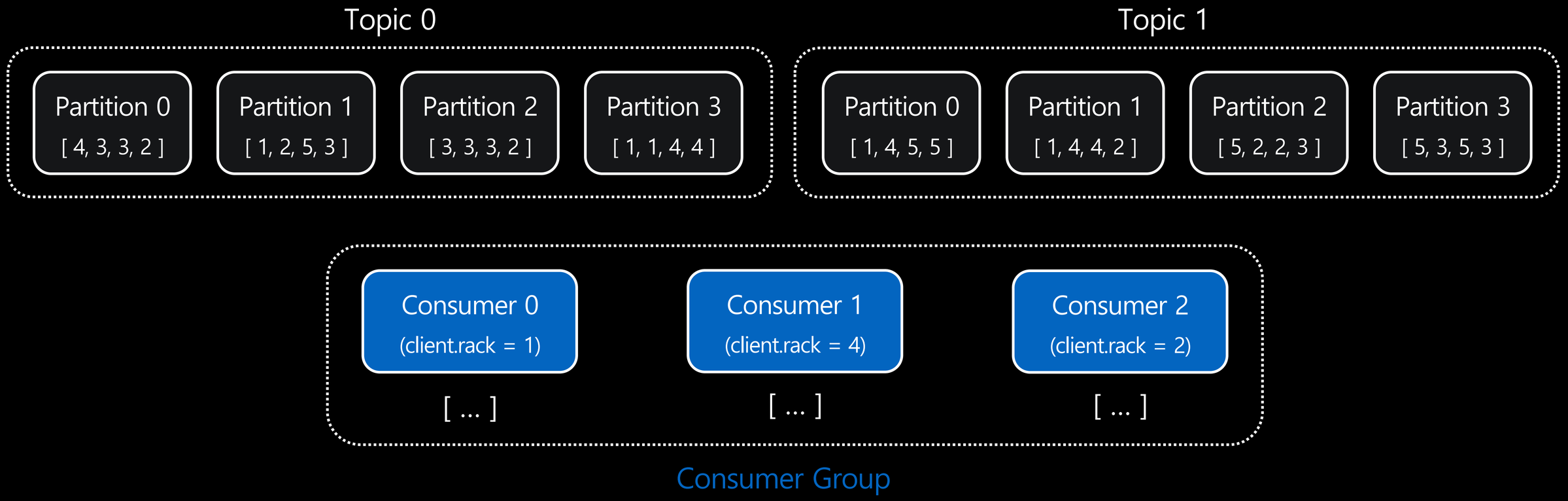
NAVER
DEVIEW
2023

- ‘Rack’의 의미 문제
  - Multilevel Rack 개념 (논의중, KIP-879)
- Partition Assignor
  - “Rack 설정을 고려하는 Partition Assignor를 개발해서 꽂아 넣는다.”

3.4 네이버의 해법: RackAwareRangeAssignor (1)

NAVER DEVIEW 2023

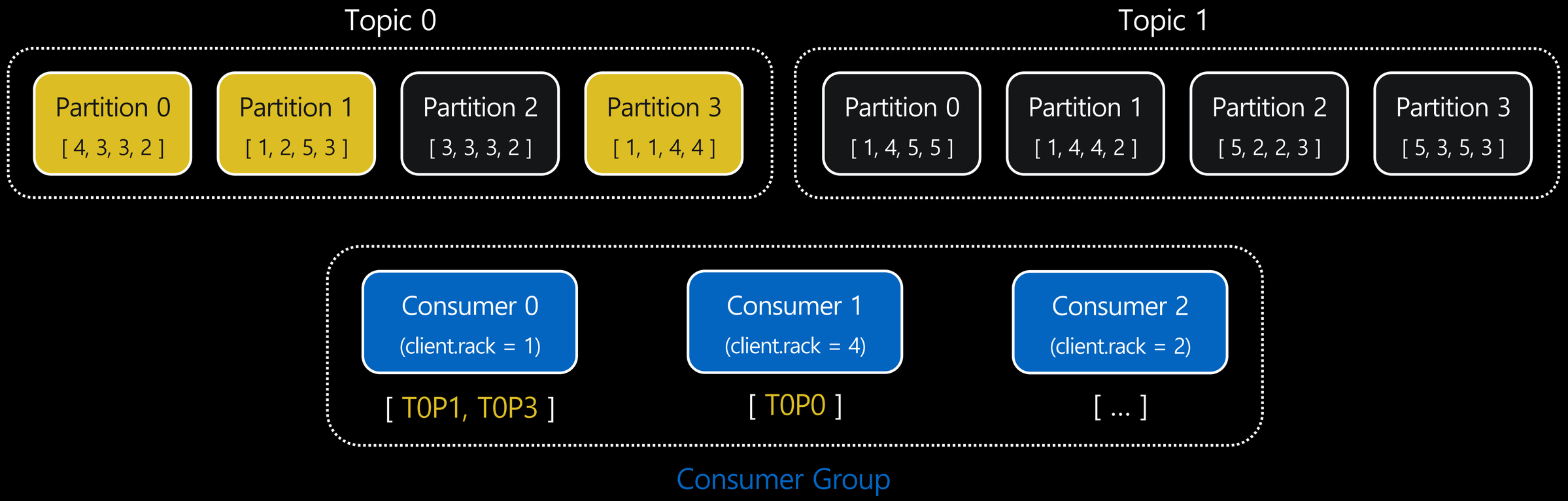
- Broker, Consumer에 설정된 rack 정보(broker.rack, client.rack)를 활용
  - Rebalance Protocol의 사용자 데이터 영역에 client.rack 설정을 실어 보낸다.



3.5 네이버의 해법: RackAwareRangeAssignor (2)

NAVER DEVIEW 2023

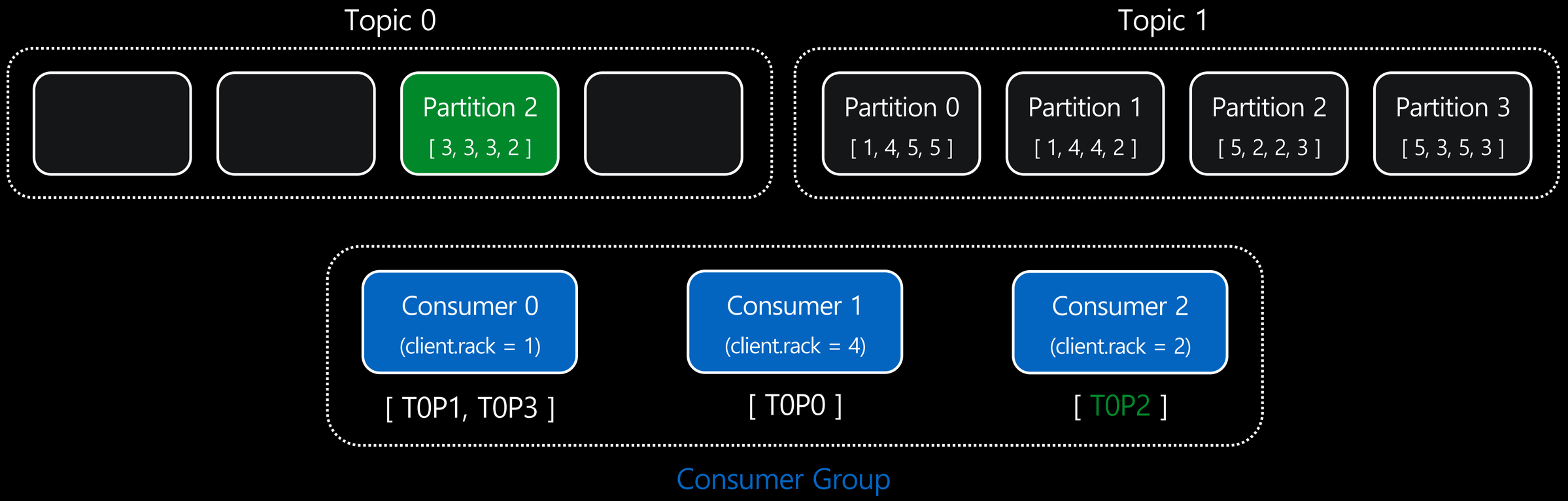
- 1단계: leader replica의 rack에 따라 할당



3.6 네이버의 해법: RackAwareRangeAssignor (3)

NAVER DEVIEW 2023

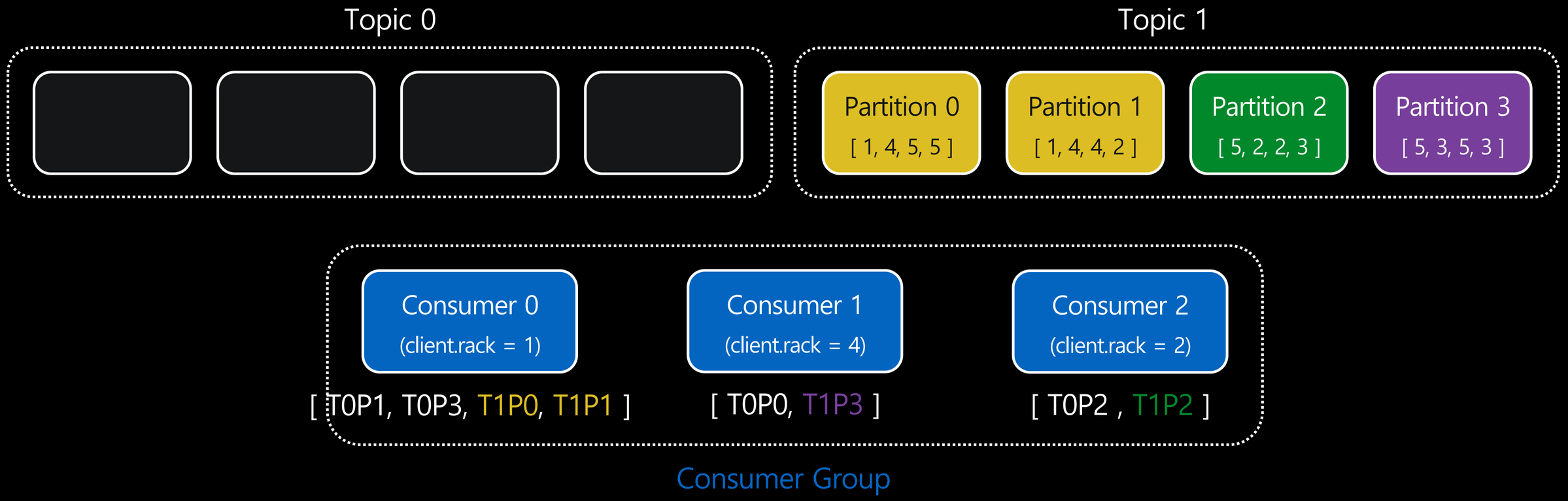
- 2단계: follower replica의 rack에 따라 할당



3.7 네이버의 해법: RackAwareRangeAssignor (3)

NAVER DEVIEW 2023

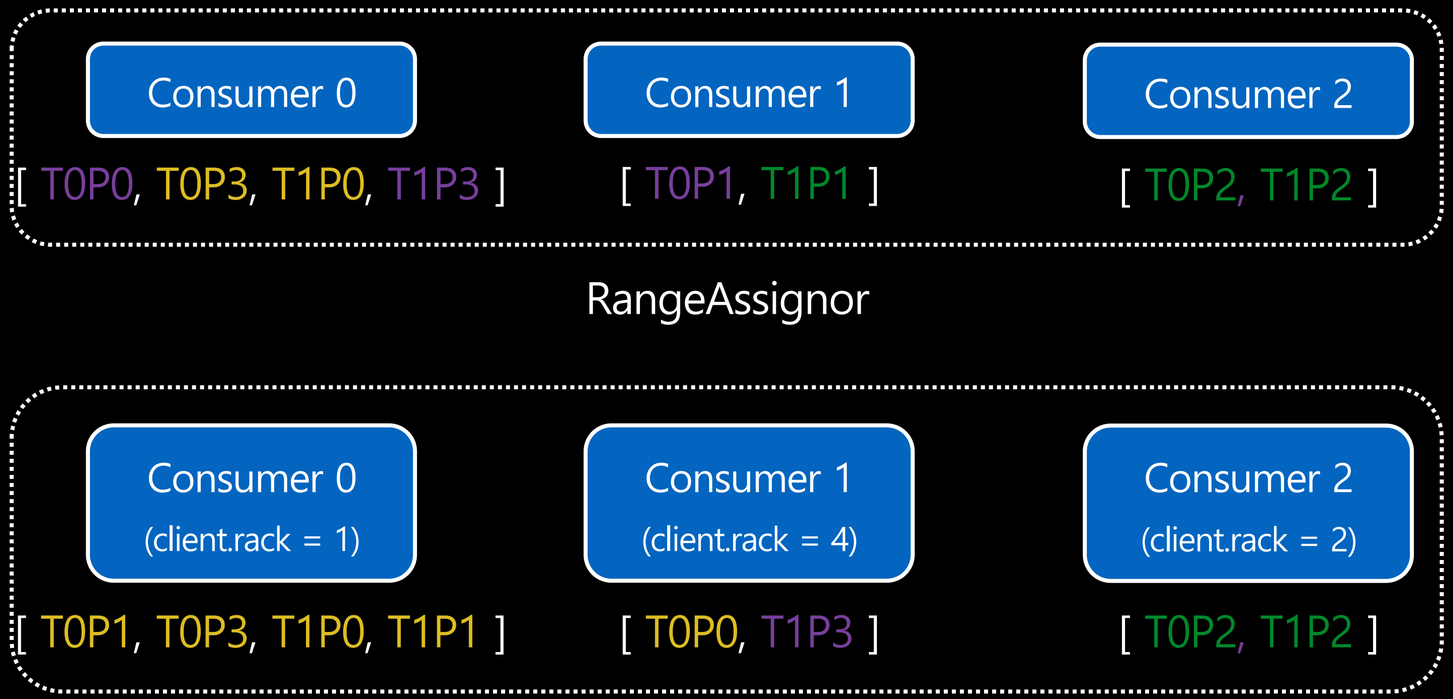
- 3단계: 남은 Partition들을 순차적으로 할당
  - Topic별로 위 과정을 반복



3.8 네이버의 해법: RackAwareRangeAssignor (4)

NAVER DEVIEW 2023

- 비교
  - Consumer와 같은 rack에 배치된 leader replica 수: 2:5
  - Consumer와 같은 rack에 배치된 follower replica 수: 3:2



3.9 Apache Kafka 3.4 업데이트

NAVER
DEVIEW
2023

- 기존 Consumer Protocol이 확장되어 rack 정보를 담을 수 있게 됨 (KIP-881)
  - 아직 이 정보를 활용하는 Partition Assignor 구현체가 탑재되지는 않음
  - 차기 Consumer Protocol (KIP-848) 부터는 처음부터 rack 정보를 고려할 예정

4. 요약

4.1 요약 (1) – Kafka Consumer

NAVER
DEVIEW
2023

"Kafka Consumer에는 Consumer Group이라는 개념이 있으며, 같은 Consumer Group에 속한 Consumer들은 Topic을 읽어올 때 여기 속한 Partition들을 자동으로 나눠 가진다."

4.1 요약 (2) – Cloud에서의 Kafka Consumer

NAVER
DEVIEW
2023

"클라우드 환경에서 Consumer Group 기능을 사용하는 것이 쉽지만은 않으며, 2.x 이후 업데이트된 기능과 설정들을 적절히 활용함으로써 문제 발생을 막을 수 있다."

4.1 요약 (3) – 네이버에서의 Kafka Consumer

NAVER
DEVIEW
2023

“Rack 관련 최적화 기능이 추가된 Partition Assignor는 가능하며, 멀지 않은 미래에 보편적인 기능이 될 것이다.”

질문 받습니다.

감사합니다. 🙏