KAFKA Consumer Redesign

Guozhang Wang

Cornell University

Email: guoz@cs.cornell.edu

What is Kafka?

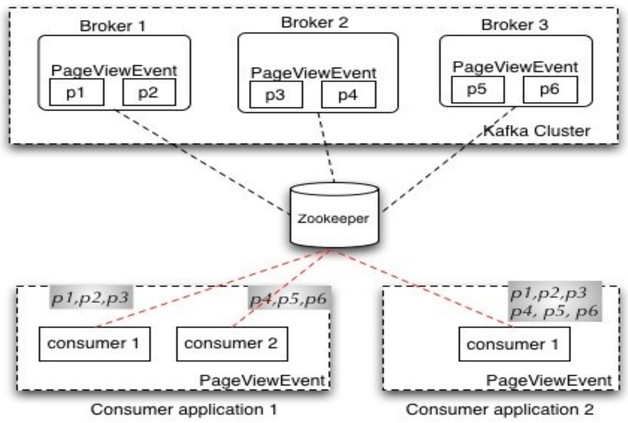
KAFKA is a distributed, persistent and high-throughput publish-subscribe messaging system mainly used for log processing.

At LinkedIn, KAFKA is used to deliver activity stream data and operational metrics. This powers various products like LinkedIn Newsfeed, LinkedIn Today in addition to offline analytics systems like Hadoop and DW.

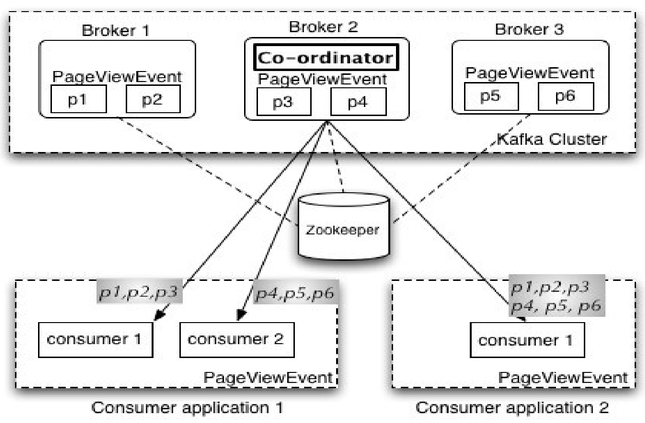
Motivation of Consumer Redesign

- Distributed co-ordination and load balancing on consumers is complicated, limits scalability, performance and is error-prone due to its split-brain nature.
- Distributed co-ordination also limits support for advanced consumer features like manual offset management and manual partition assignment.
- A thick consumer client with dependency on zookeeper makes it difficult to be reimplemented in other languages that lack stable ZK client library.

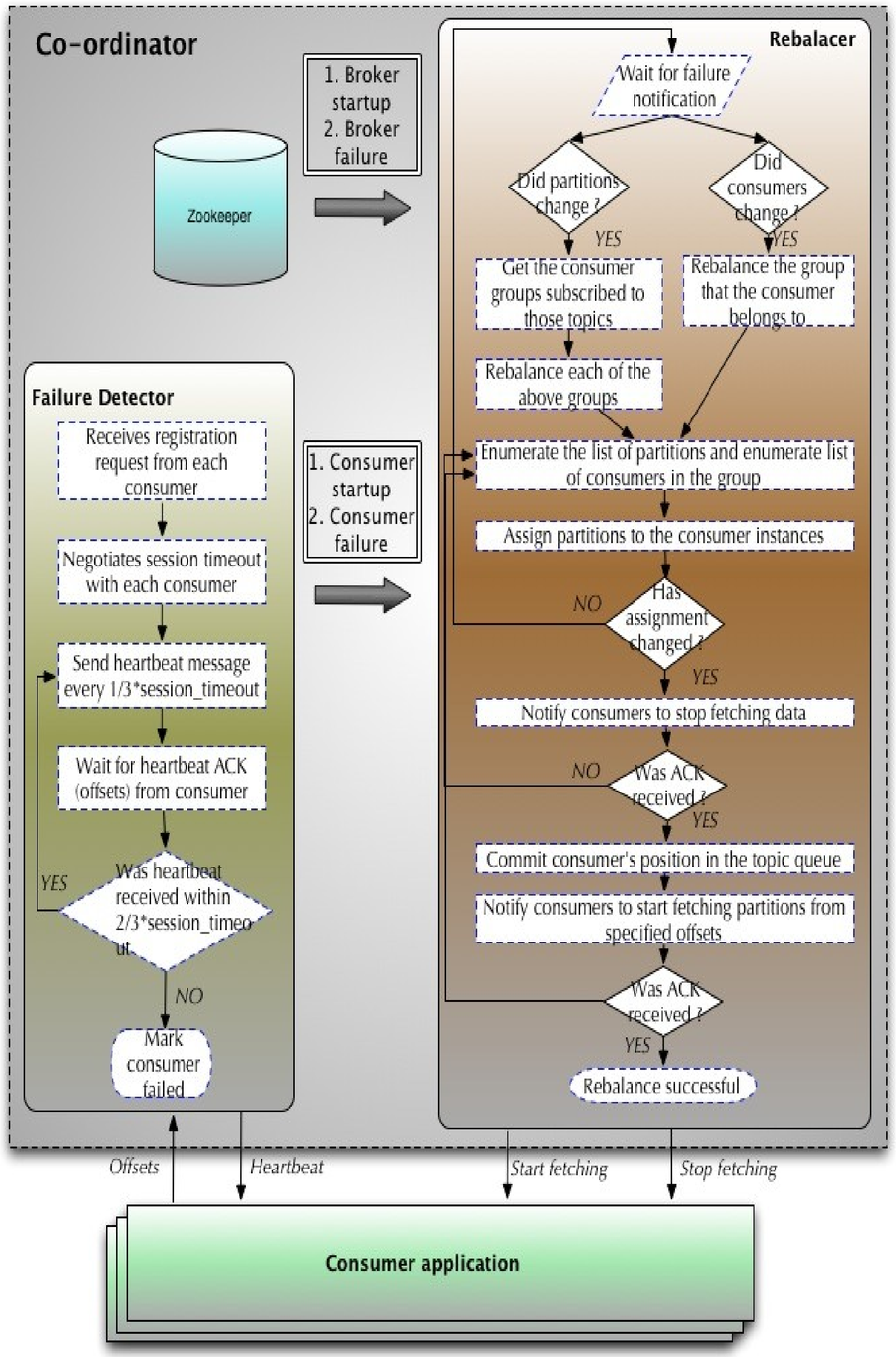
Original Kafka Architecture



New Kafka Architecture

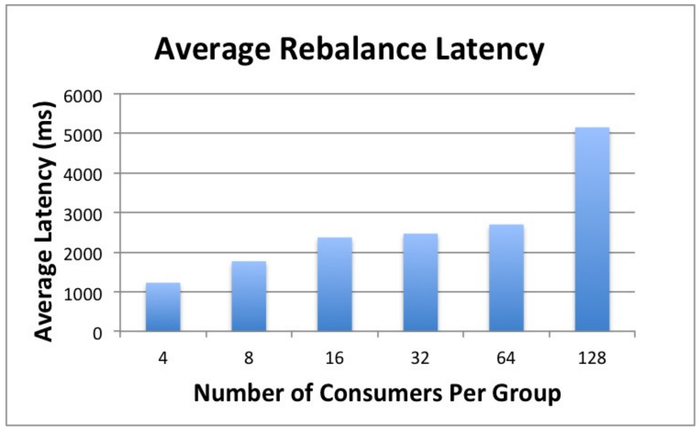


Co-ordinator-based Failure Detection and Rebalancing



- Use a centralized highly-available co-ordinator for load balancing topic partitions evenly amongst consumers.
- The centralized co-ordinator is elected amongst the brokers in a Kafka cluster.
- The co-ordinator detects failures on brokers as well as consumers.
- Upon detecting broker/consumer failure or newly added broker/consumer, thr co-ordinator will trigger the rebalance procedure for the affected groups.
- Once connected to the co-ordinator, consumers can simply listen on the channel for stop/start fetcher requests upon rebalancing.
- When a consumer suspects that the co-ordinator is no longer available, it will stop fetching until it discovers the new co-ordinator.

Results



Future Work

- Extend the thin consumer architecture to support manual partition assignment and manual offset management.
- Performance comparison between new and old consumer architecture.

Acknowledgements

Neha Narkhede, Jun Rao, Joel Koshy, Ye Yang, John Fung,

Jay Kreps, Mammad Zadeh