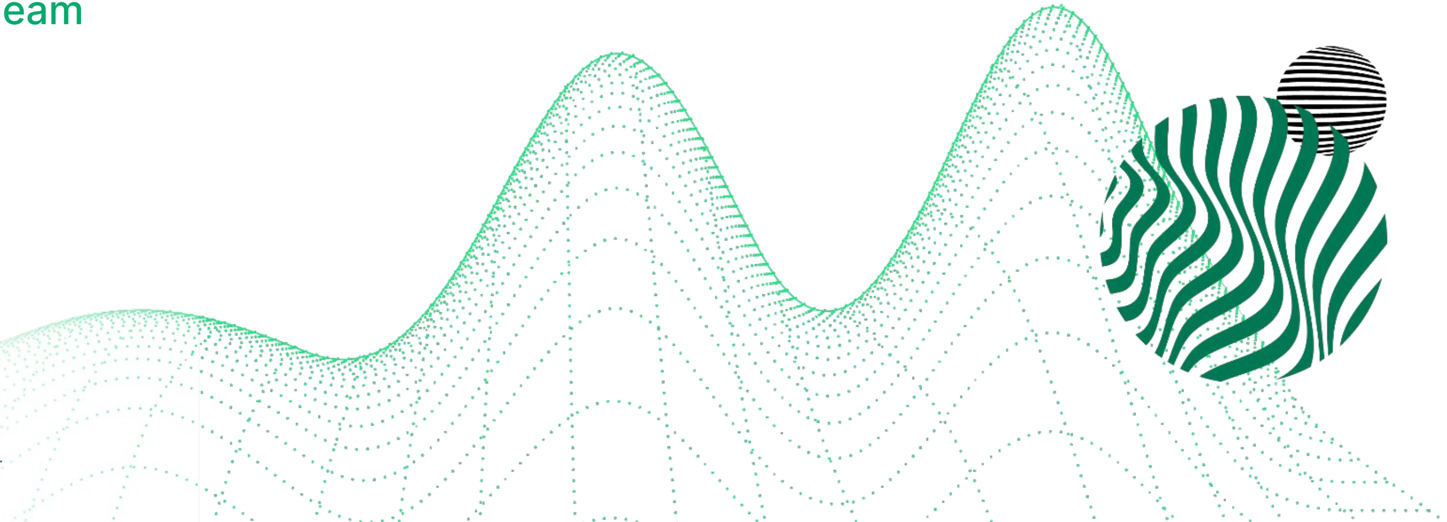
Improving Kafka Resilience - Gray Failures Mitigation

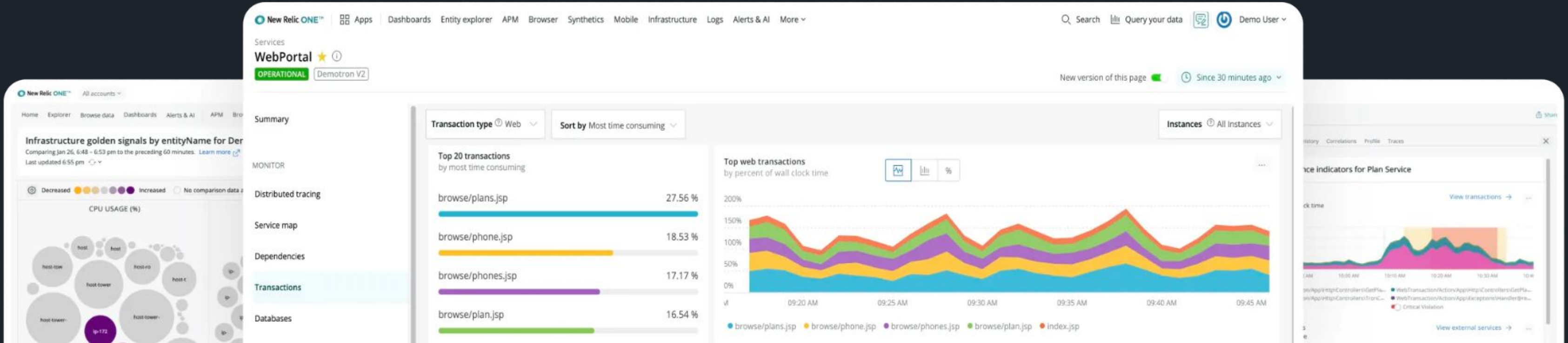
Michelle Valentinova, Senior Site Reliability Engineer

Kafka Platform Team



© 2023 New Relic, Inc. All rights reserved.

new relic®



Services

WebPortal

OPERATIONAL Demotron V2

Transaction type Web

Sort by Most time consuming

Instances All Instances

Top 20 transactions
by most time consuming

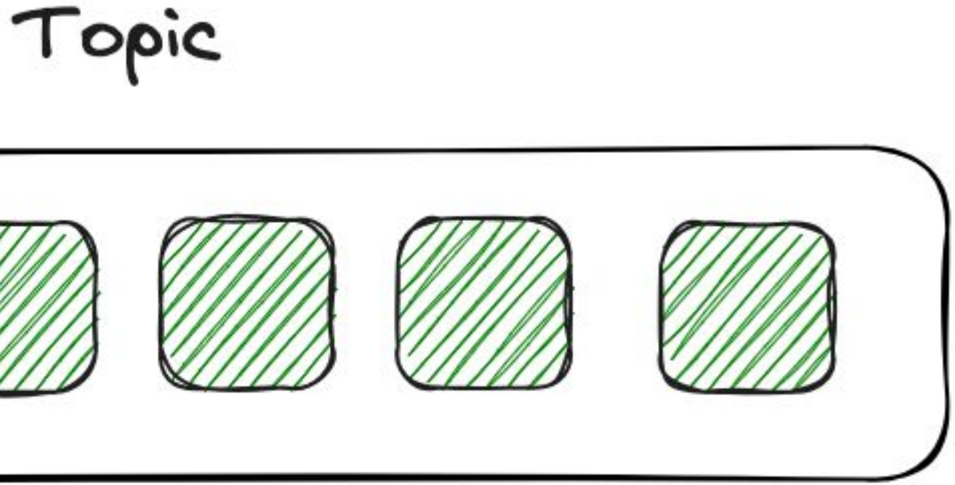
browse/plans.jsp	27.56 %
browse/phone.jsp	18.53 %
browse/phones.jsp	17.17 %
browse/plan.jsp	16.54 %

Infrastructure golden signals by entityName for Der

Comparing Jan 26, 6:48 - 6:53 pm to the preceding 60 minutes. Learn more

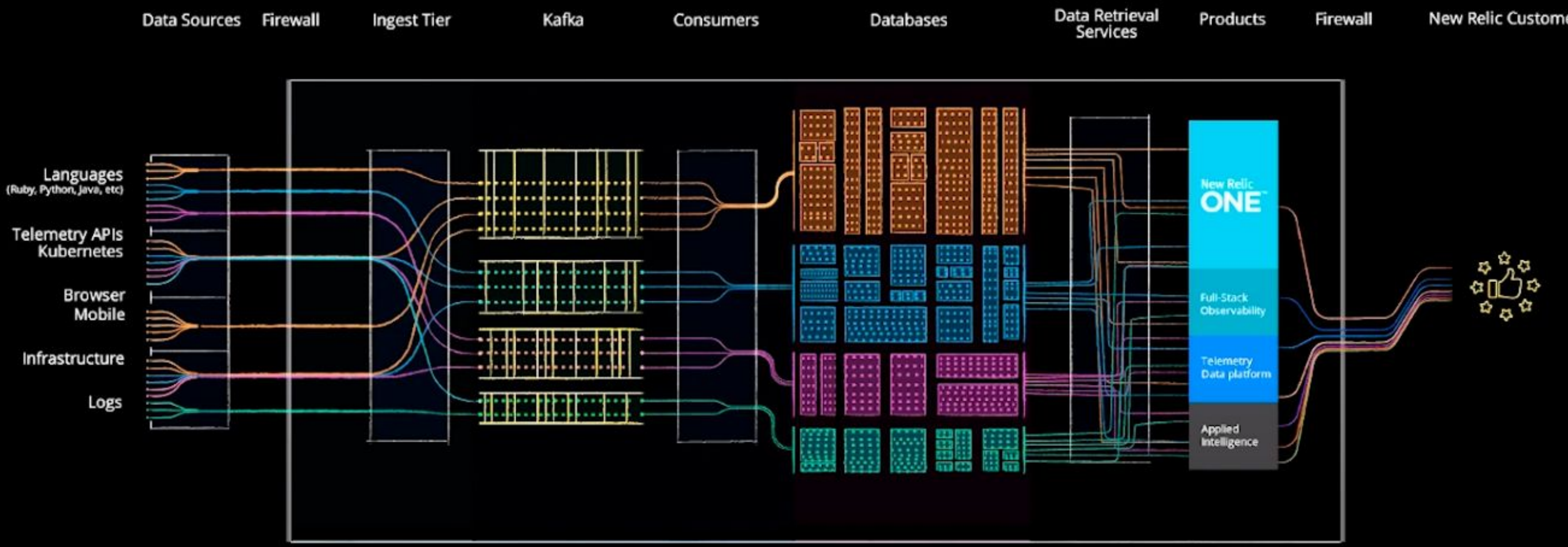
Last updated 6:55 pm

Apache Kafka - event streaming



© 2023 New Relic, Inc. All rights reserved

Kafka at New Relic



97 clusters
2000 brokers
180M messages/s

© 2023 New Relic, Inc. All rights reserved

Distributed systems

It works

It doesn’t work

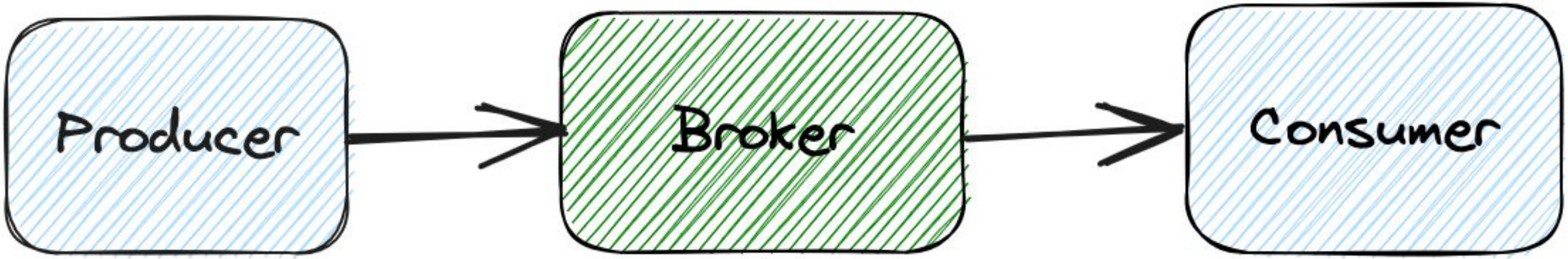
© 2023 New Relic, Inc. All rights reserved

Distributed systems

It kind of works

© 2023 New Relic, Inc. All rights reserved

Kafka concepts



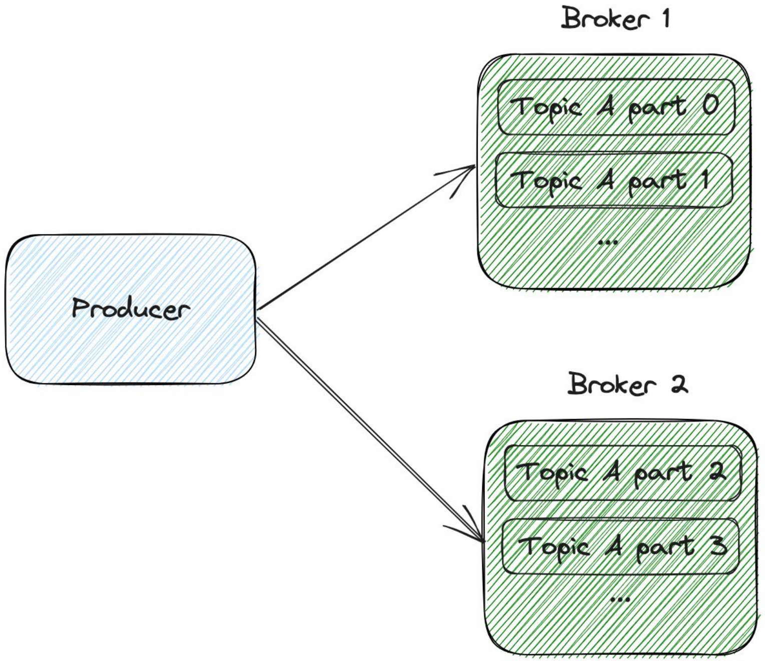
© 2023 New Relic, Inc. All rights reserved

Topics are partitioned

- Parallel processing
- High availability
  - Failover between partitions

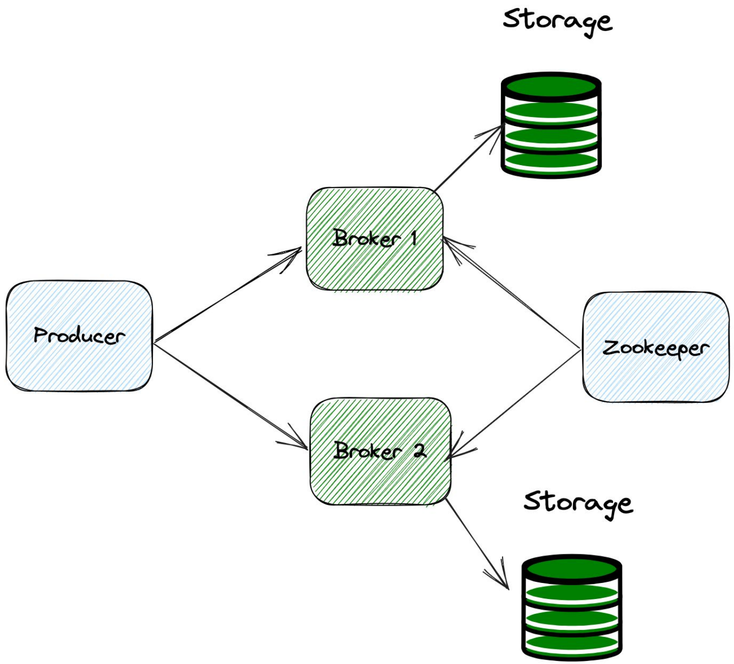
© 2023 New Relic, Inc. All rights reserved

Topics, partitions, and brokers in Kafka



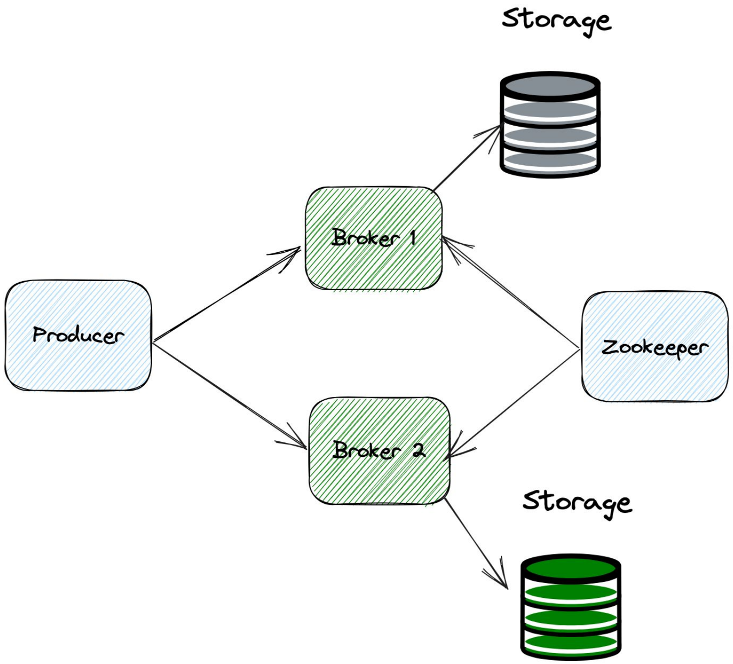
© 2023 New Relic, Inc. All rights reserved

Kafka in the cloud architecture



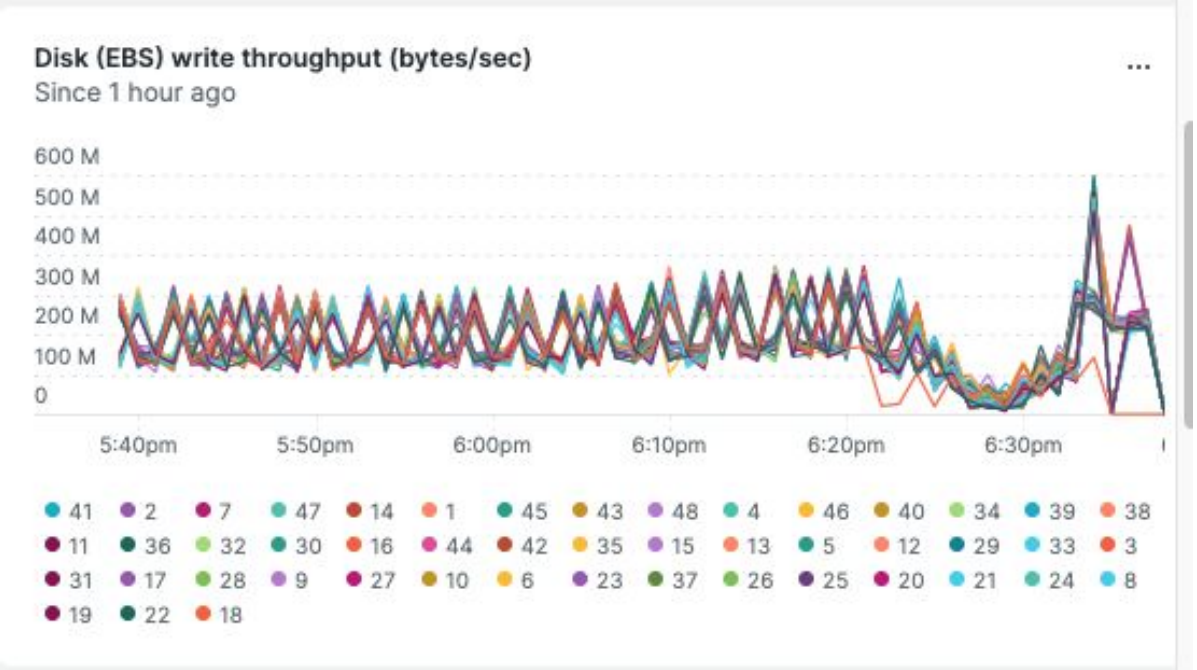
© 2023 New Relic, Inc. All rights reserved

Kafka broker's network storage very slow



© 2023 New Relic, Inc. All rights reserved

Kafka broker’s network storage very slow

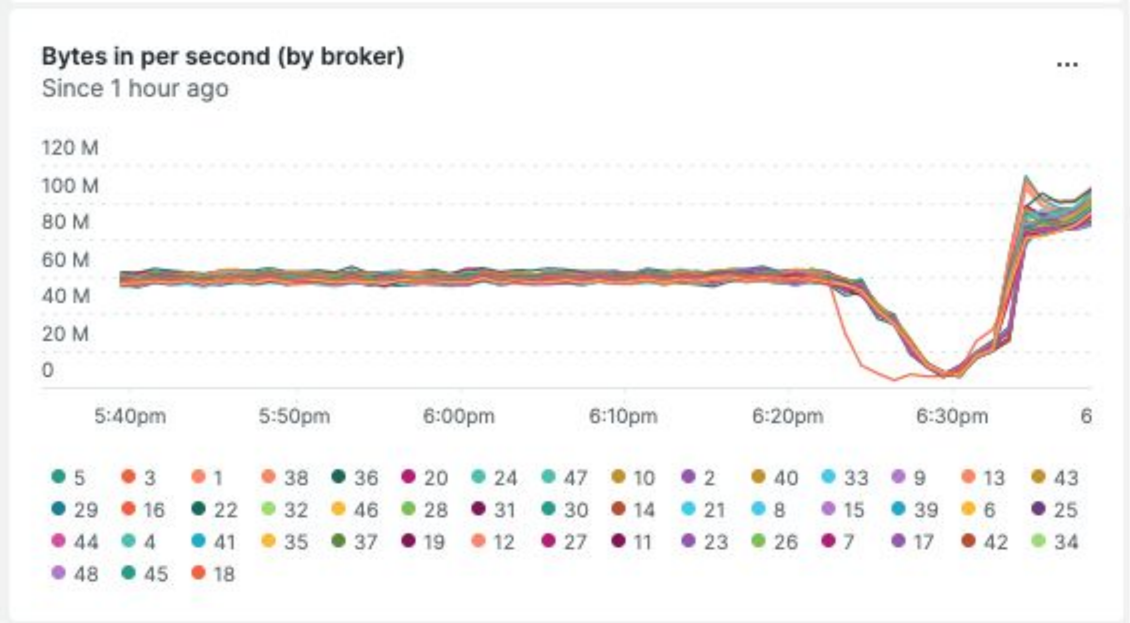


Disk (EBS) write throughput (bytes/sec)

Since 1 hour ago

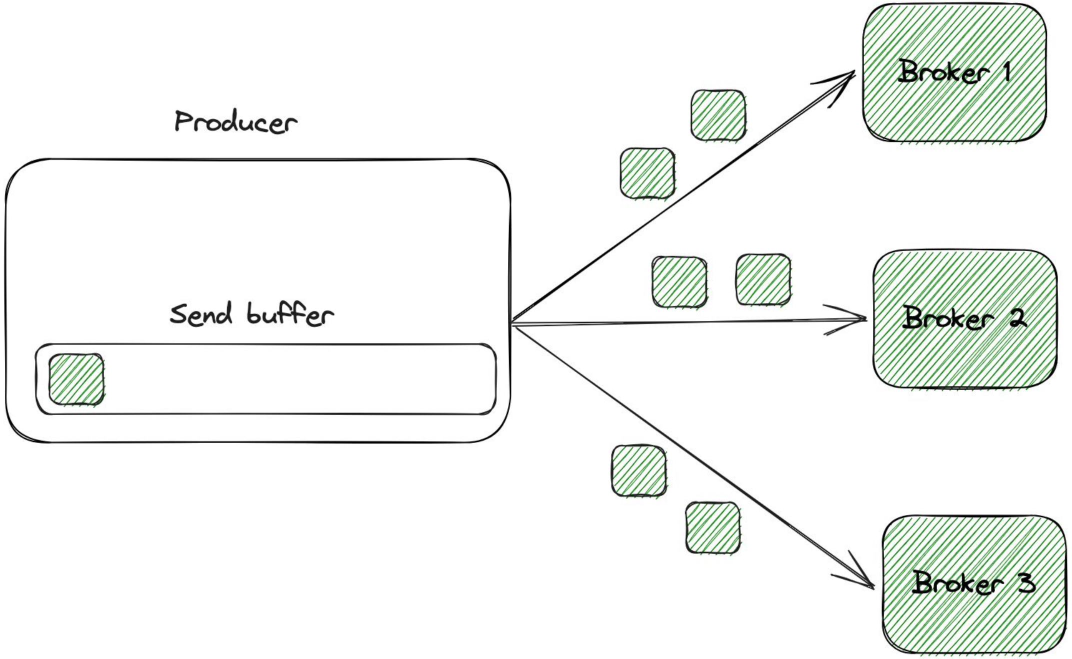
© 2023 New Relic, Inc. All rights reserved

Issues with a single broker affect all producers



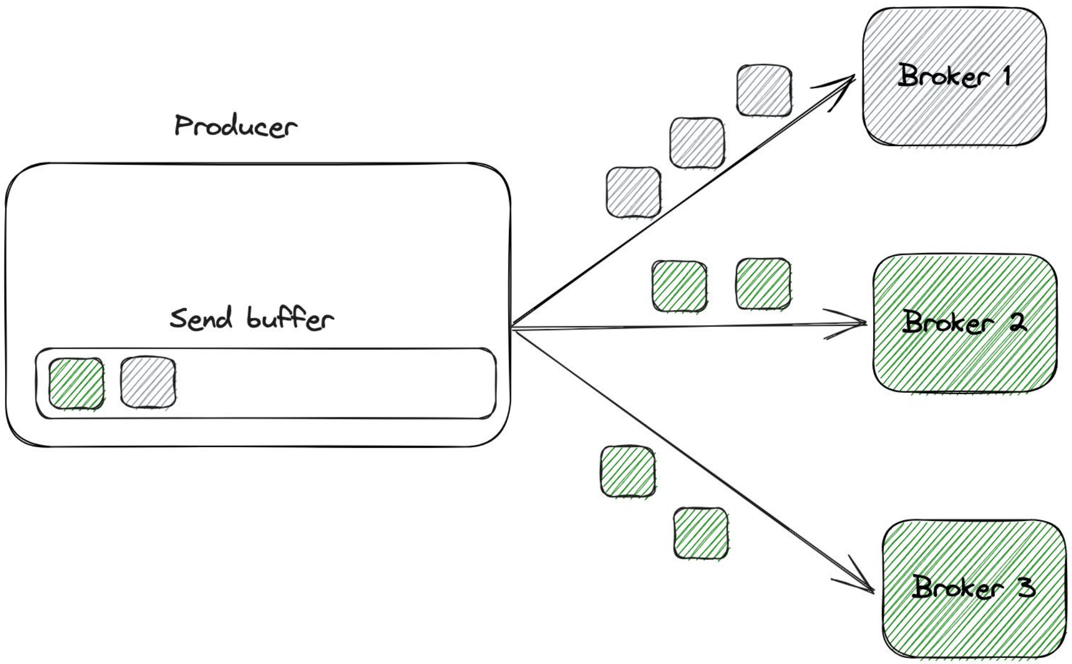
© 2023 New Relic, Inc. All rights reserved

Producers optimise for throughput (send buffer)



© 2023 New Relic, Inc. All rights reserved

One broker is slow



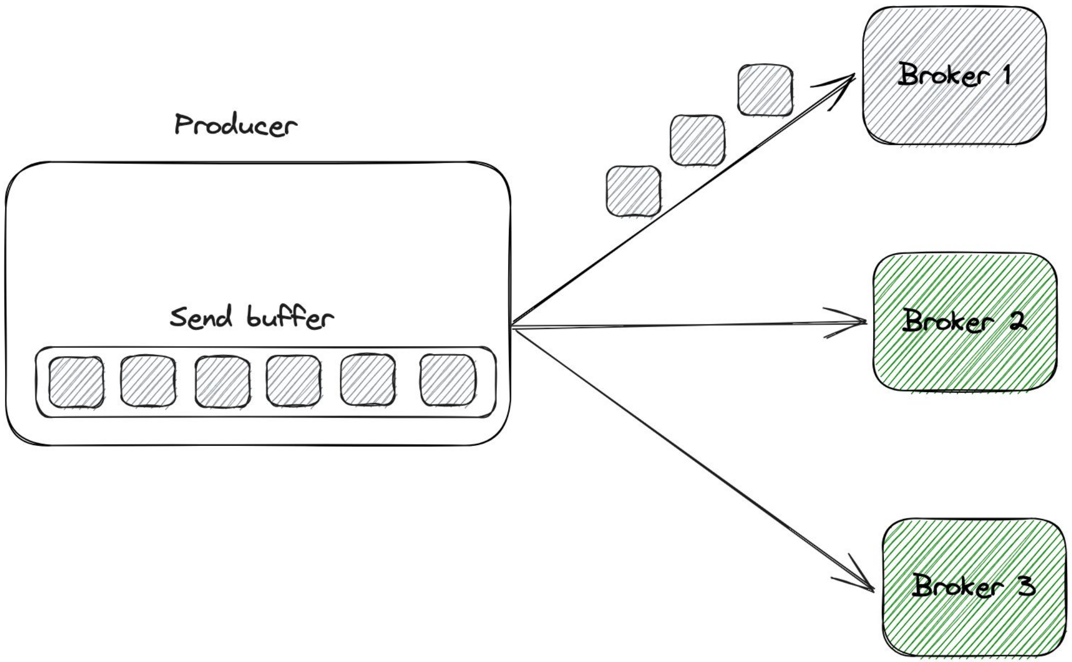
© 2023 New Relic, Inc. All rights reserved

Producers retry infinitely (configurable)

Don’t want to lose customer logs, metrics, miss alerts, etc.

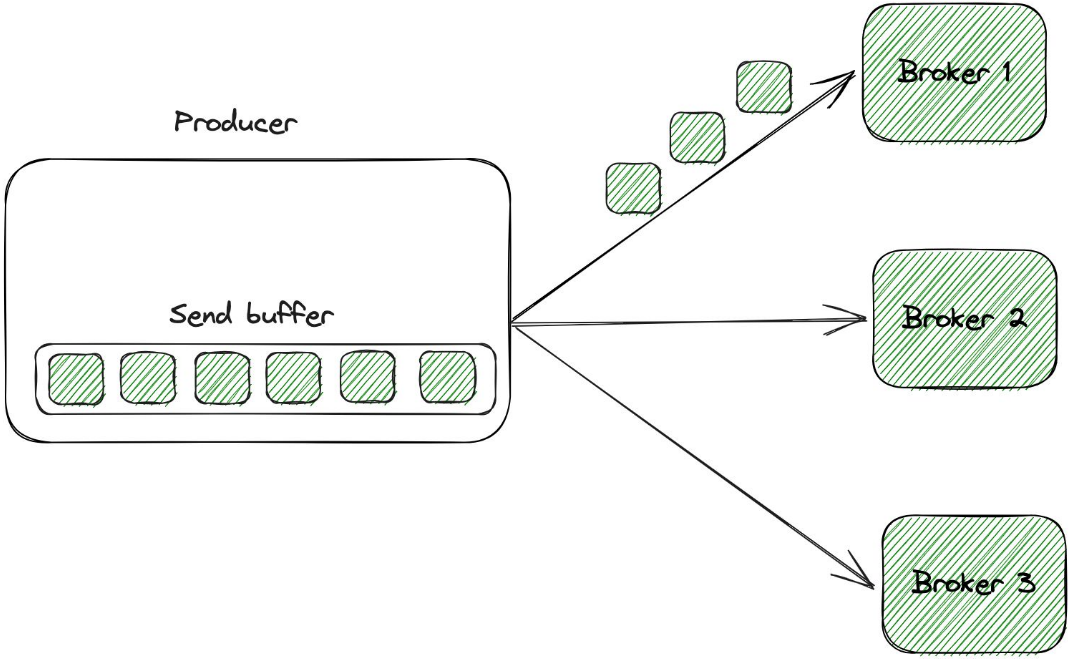
© 2023 New Relic, Inc. All rights reserved

Send buffer fills up



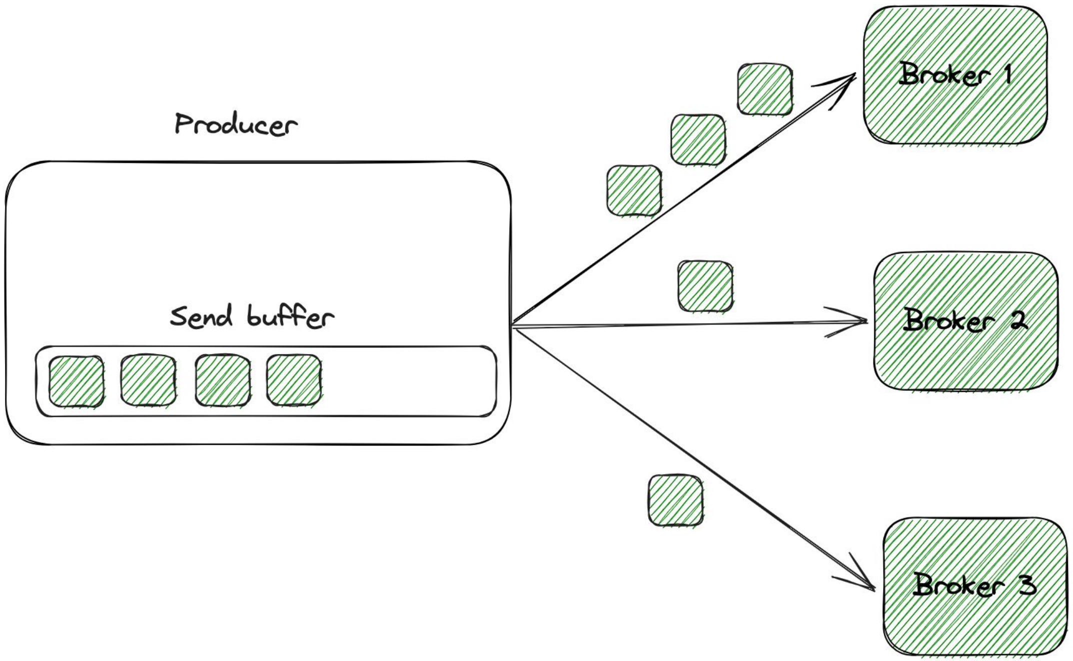
© 2023 New Relic, Inc. All rights reserved

The issue is resolved



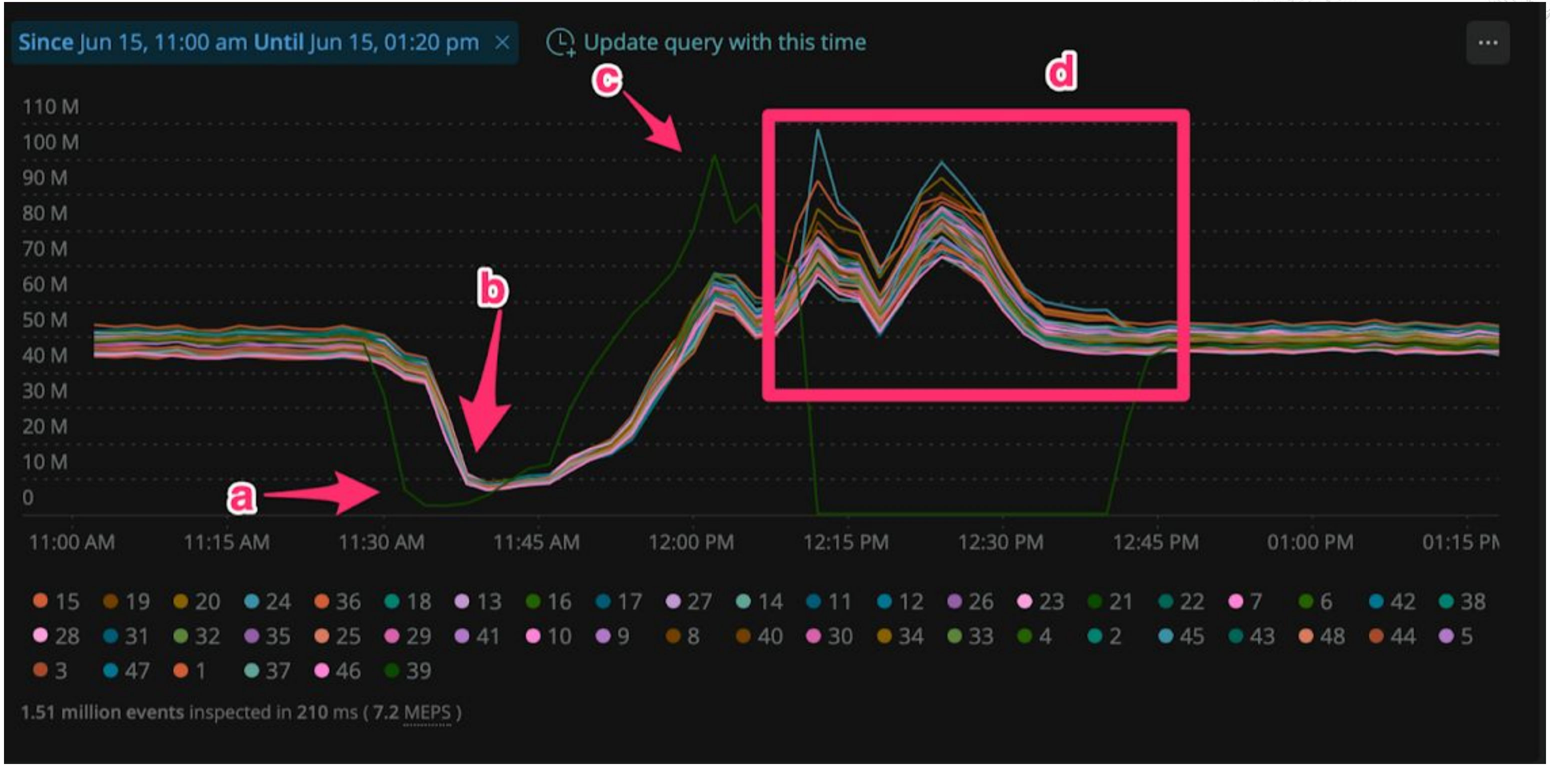
© 2023 New Relic, Inc. All rights reserved

The issue is resolved continued



© 2023 New Relic, Inc. All rights reserved

Issue with a single broker affects all producers



© 2023 New Relic, Inc. All rights reserved

Impact for all producers:
25 minutes

Customer metrics, logs, alerts, etc. delayed for 25 minutes

© 2023 New Relic, Inc. All rights reserved

Why do we even use network storage if it causes problems?

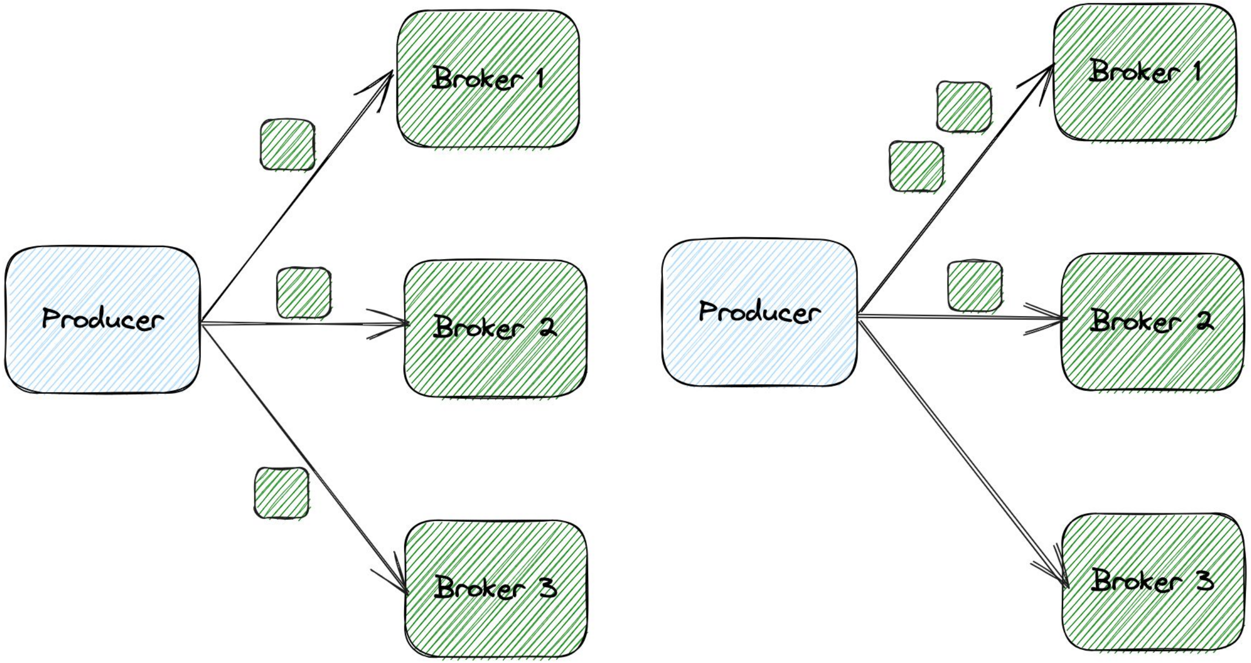
© 2023 New Relic, Inc. All rights reserved

Our busiest Kafka clusters have 1.7PB of storage

We use a managed service and can’t choose local disk

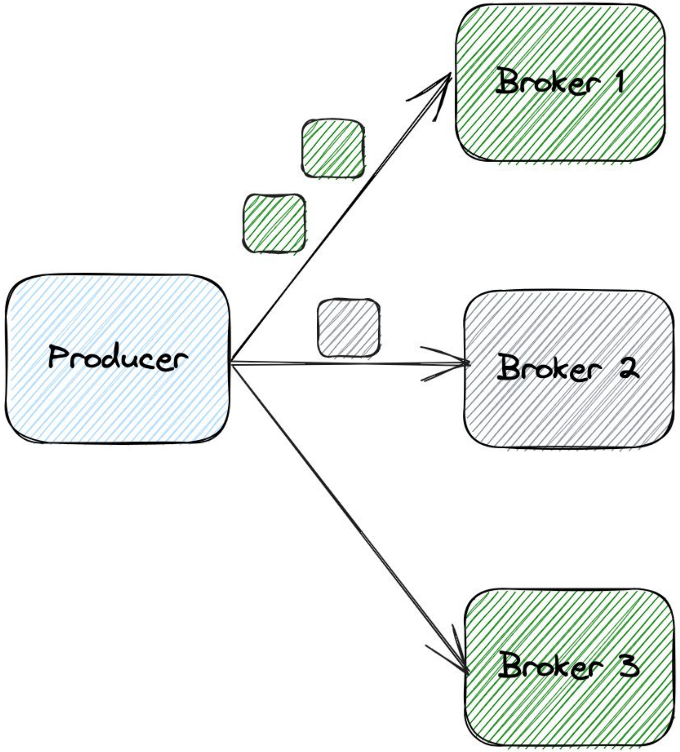
© 2023 New Relic, Inc. All rights reserved

Batching records using a partitioner



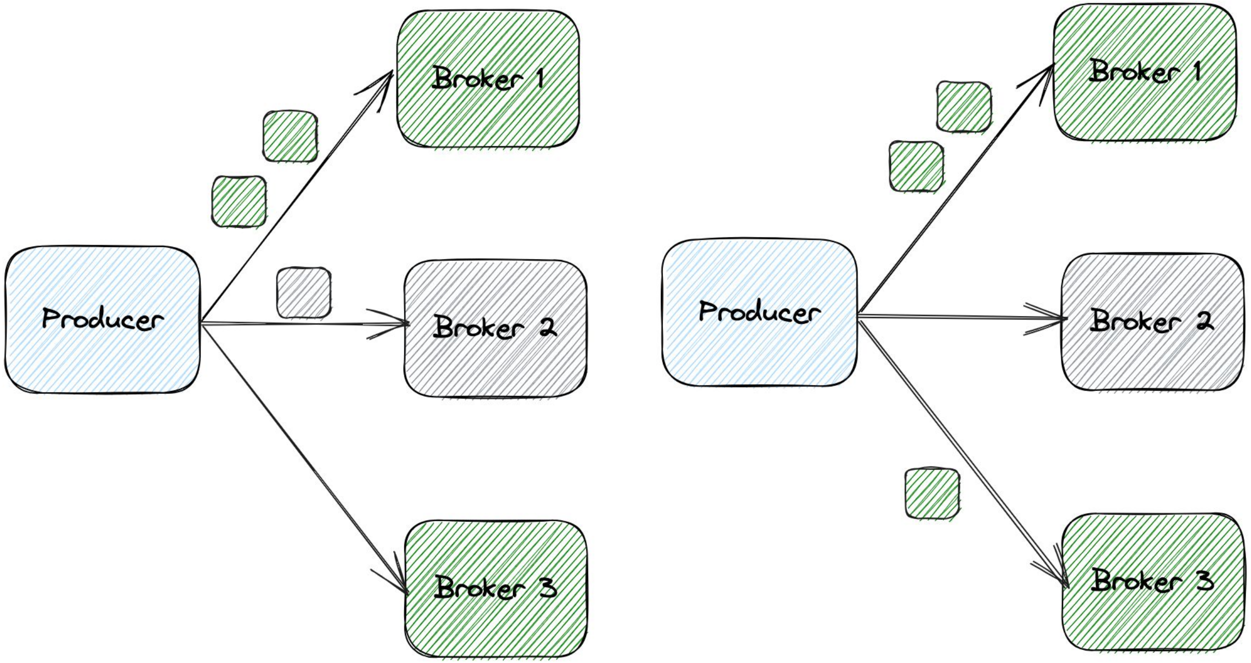
© 2023 New Relic, Inc. All rights reserved

What happens when there is a failure



© 2023 New Relic, Inc. All rights reserved

Adaptive random partitioning



© 2023 New Relic, Inc. All rights reserved

Strictly uniform sticky partitioner

- Available in Apache Kafka 3.3.1 - 1 year ago
  - Our busiest clusters have 280 topics
  - Rolling out changes can take some time
- Partition probability to get records is inverse of queue size
- The partitioner is configured on the Kafka producers

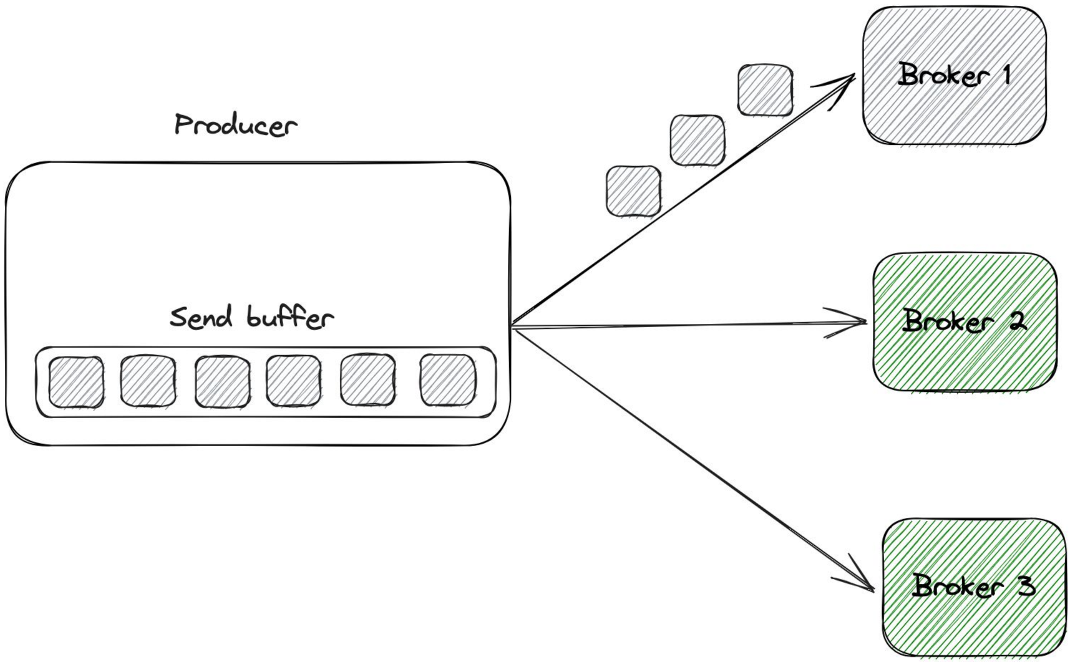
© 2023 New Relic, Inc. All rights reserved

Strictly uniform sticky partitioner

- Works well out of the box
- But can be configured to improve even further

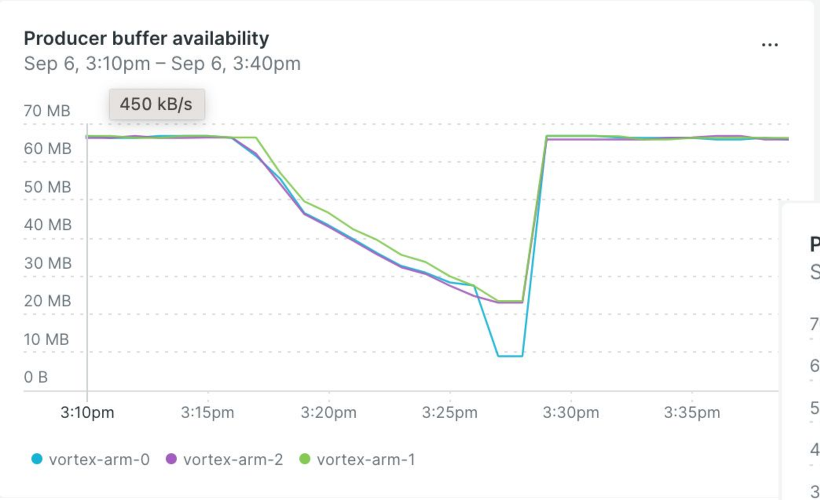
© 2023 New Relic, Inc. All rights reserved

Send buffer is shared



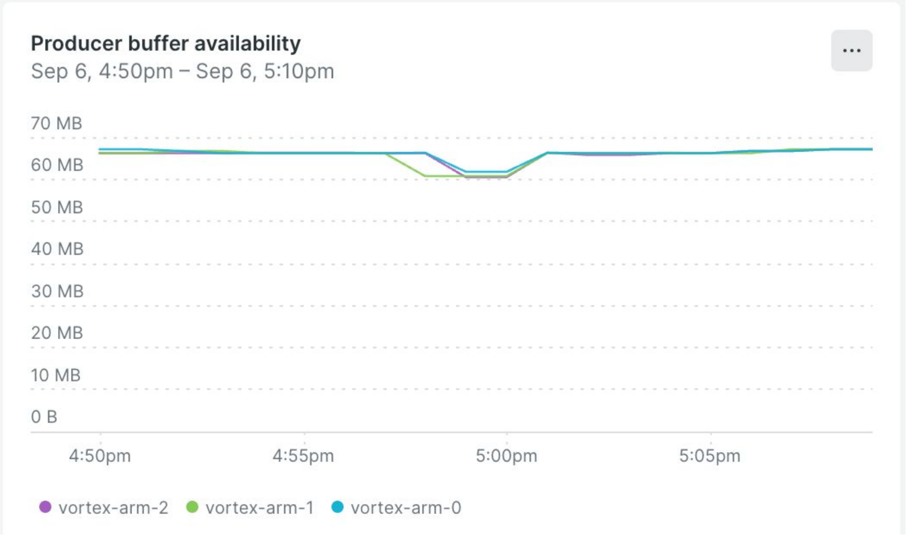
© 2023 New Relic, Inc. All rights reserved

Partition availability timeout - disabled vs 5s



disabled

5s timeout



© 2023 New Relic, Inc. All rights reserved

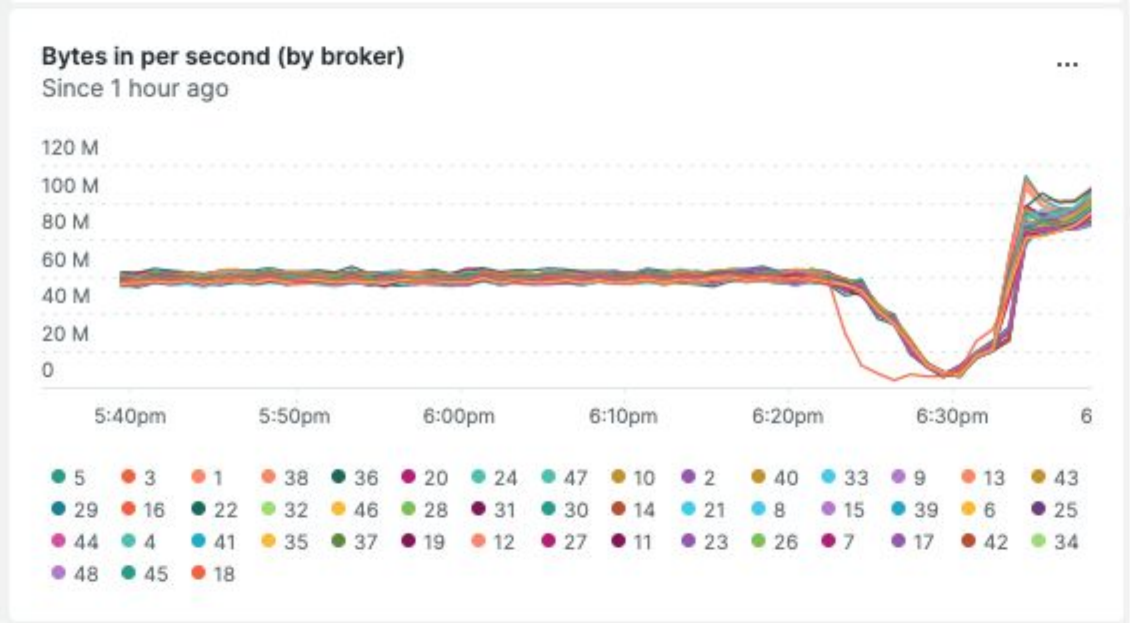
new relic

Partition availability configuration continued

- Value too low can result in too much flapping with usable brokers
  - Less throughput (Kafka specification is 10MB/s per partition)
- Value too high can result in still exhausting the send buffer

© 2023 New Relic, Inc. All rights reserved

Recovery requires even more throughput



© 2023 New Relic, Inc. All rights reserved

More throughput

- Analysed load
    - CPU had headroom - no need for more nodes
    - Disk write throughput could be improved

© 2023 New Relic, Inc. All rights reserved

More throughput continued

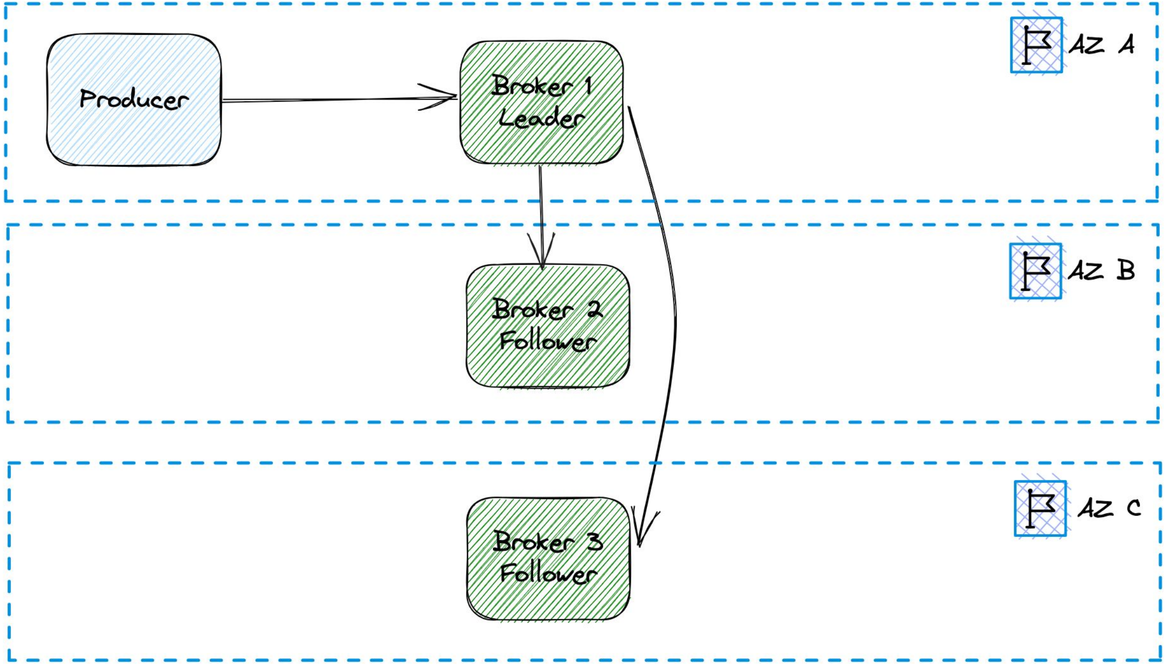
- Originally we could only use EBS gp2 volumes
  - Limited to 250MB/s write throughput
- New EBS gp3 provisioned throughput costs 30-40x less compared to extra EC2 instances with EBS gp2 volumes

© 2023 New Relic, Inc. All rights reserved

Does this solve all gray failures in Kafka?

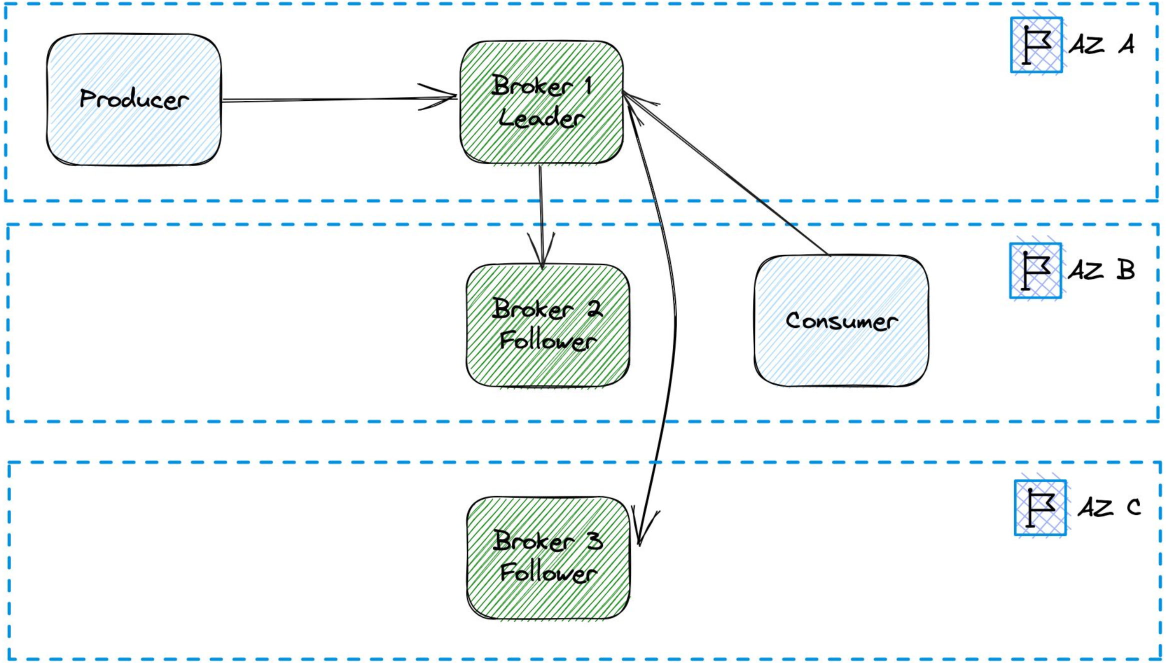
© 2023 New Relic, Inc. All rights reserved

High availability + durability with replication factor 3



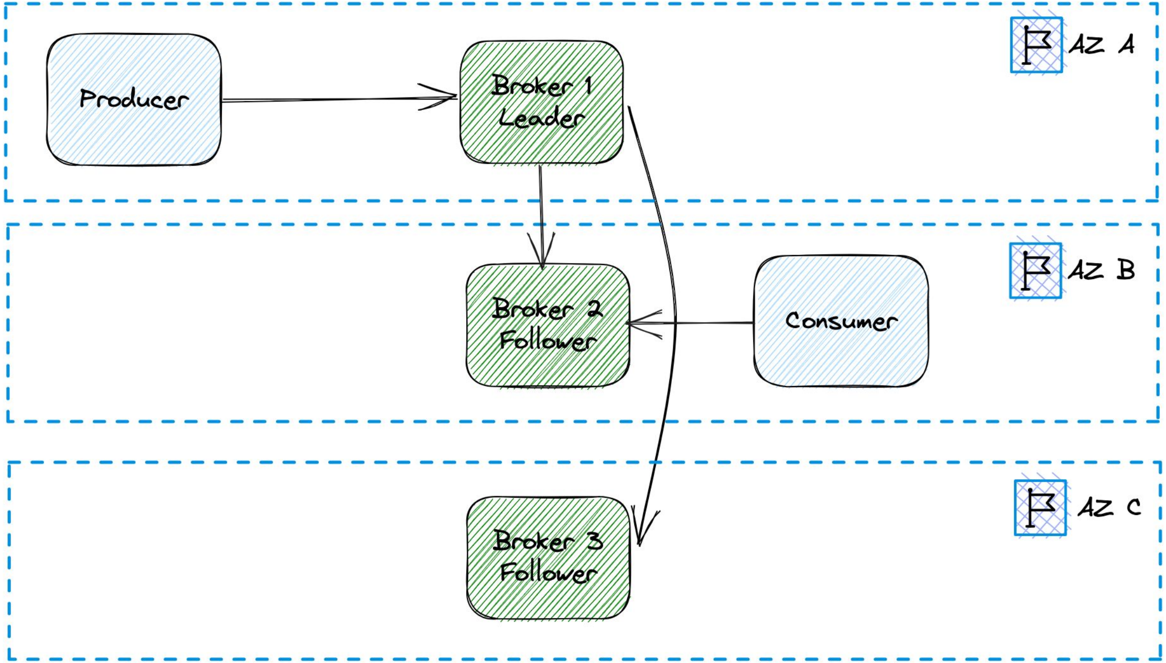
© 2023 New Relic, Inc. All rights reserved

Data path from producer to consumer



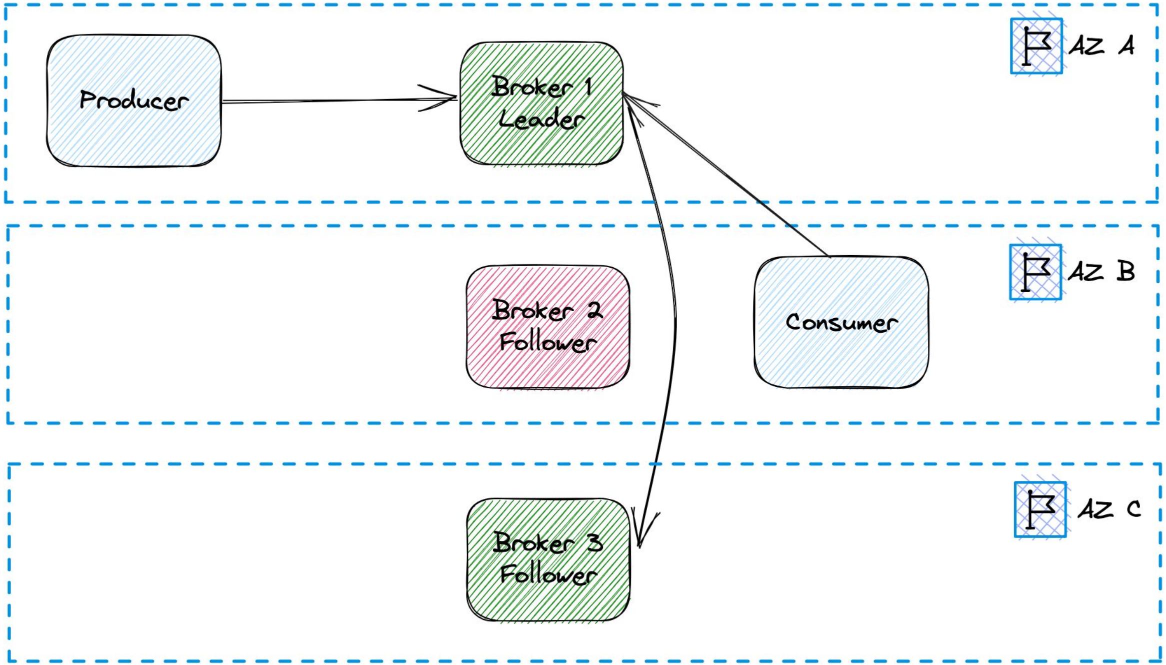
© 2023 New Relic, Inc. All rights reserved

Fetch from closest replica / less cross-AZ data



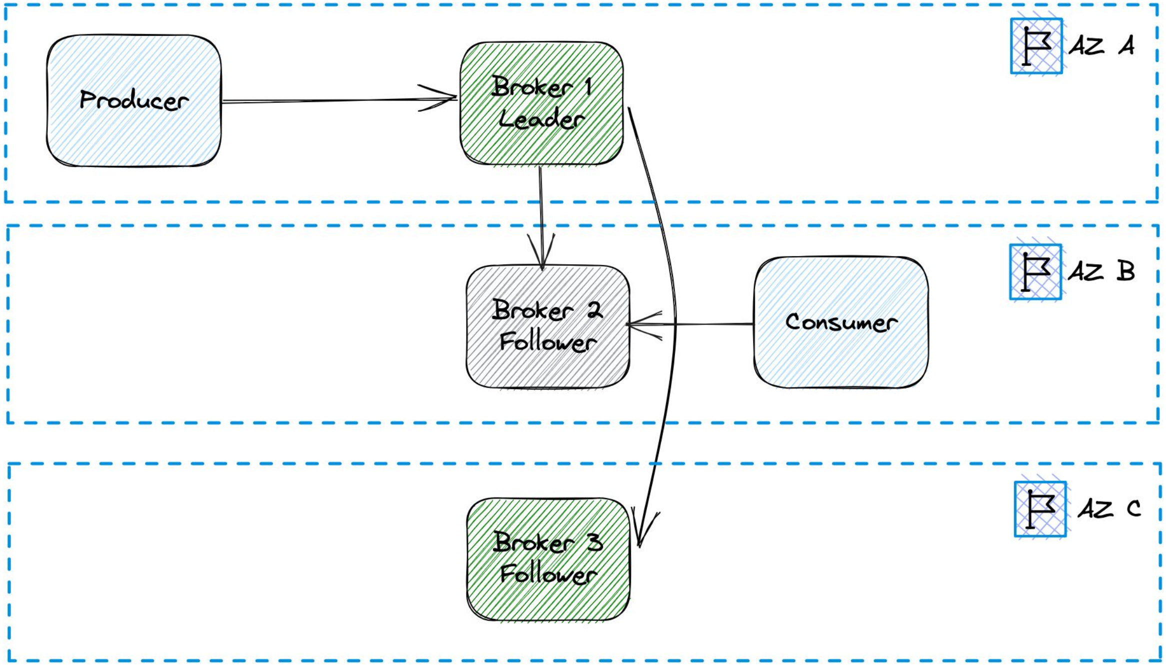
© 2023 New Relic, Inc. All rights reserved

Broker in AZ B has a problem



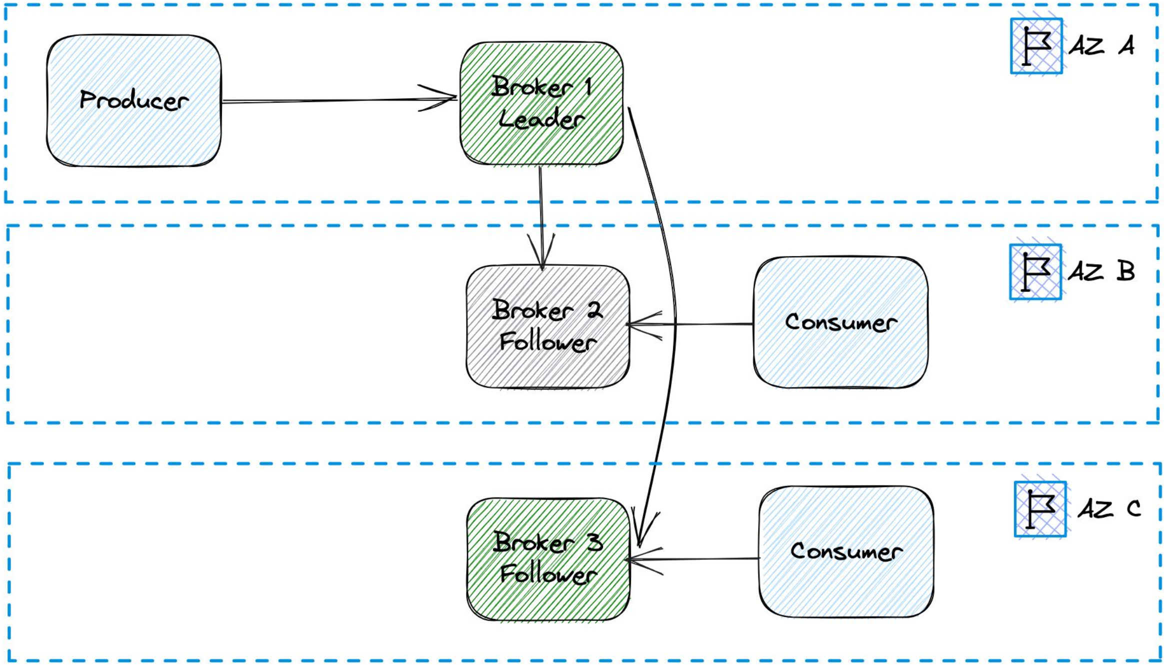
© 2023 New Relic, Inc. All rights reserved

Broker in AZ B recovers but it’s still out of sync



© 2023 New Relic, Inc. All rights reserved

Other AZs are NOT affected

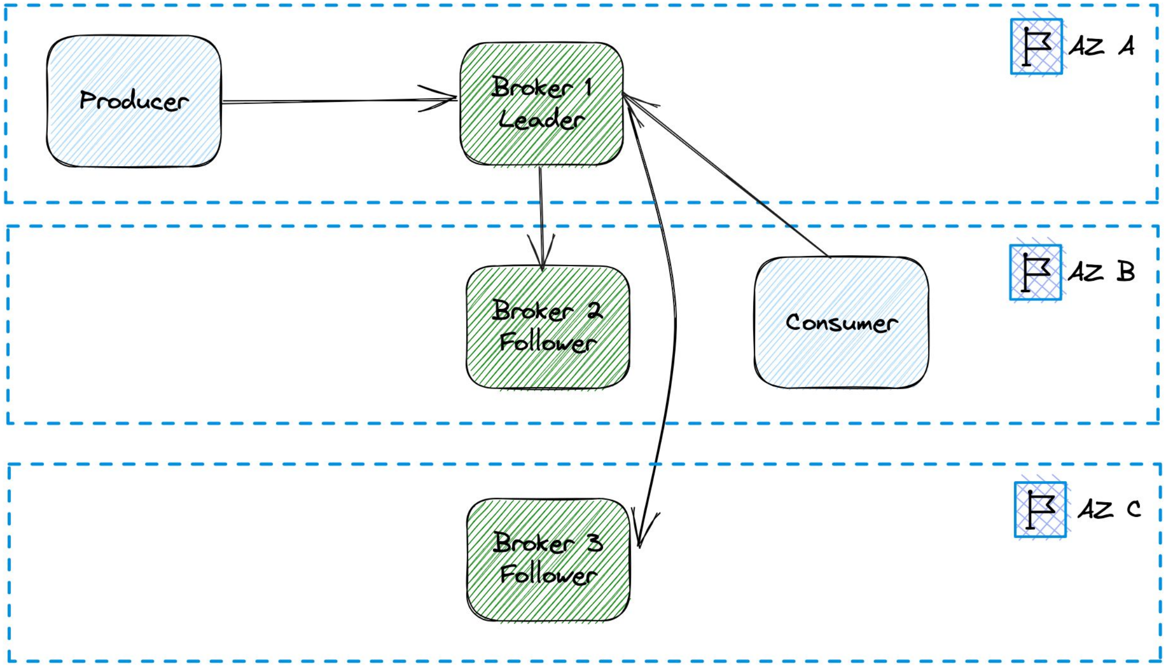


© 2023 New Relic, Inc. All rights reserved

Why not disable fetch from closest replica?

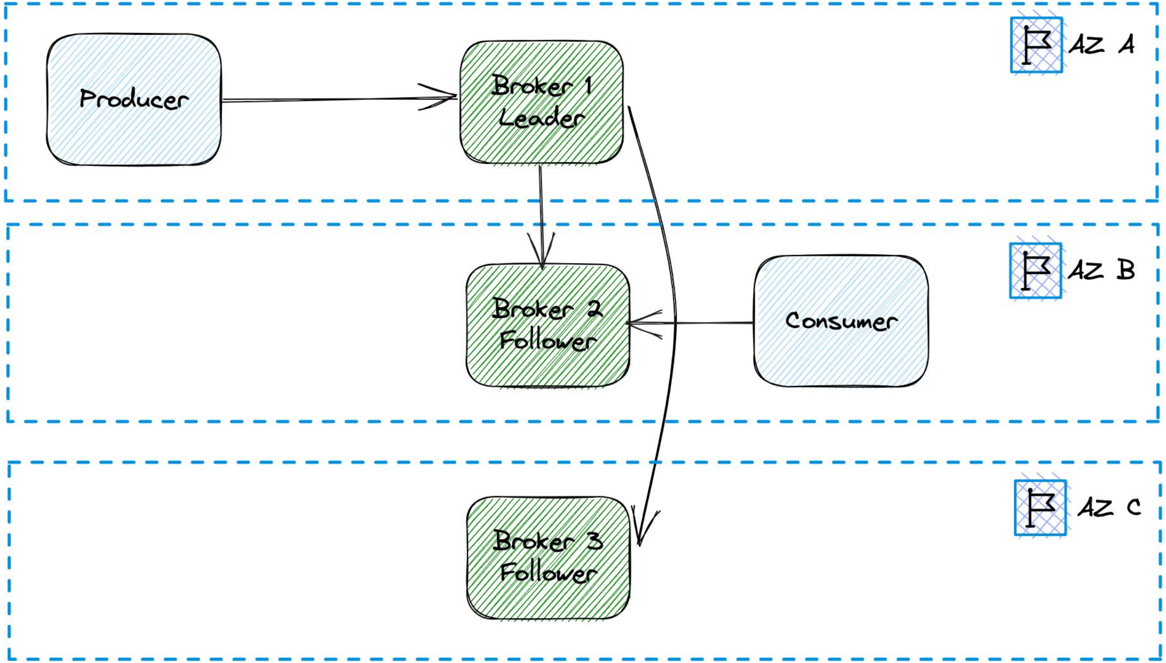
© 2023 New Relic, Inc. All rights reserved

Data path from producer to consumer



© 2023 New Relic, Inc. All rights reserved

Fetch from closest replica / less cross-AZ data

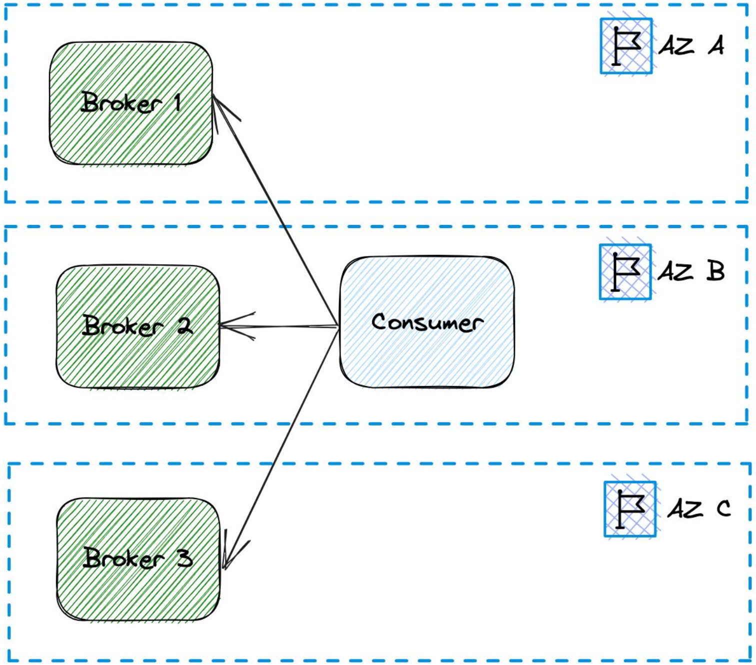


© 2023 New Relic, Inc. All rights reserved

Peak daily ingest 85GB/s

© 2023 New Relic, Inc. All rights reserved

Cross AZ traffic is only about ⅔ of all



© 2023 New Relic, Inc. All rights reserved

Why we didn’t disable fetch from closest replica

- Finance would still be unhappy even with “just” 57GB/s
- The fix was already in the works
- Can be disabled either in client or server
  - But 95 clusters
  - And close to 300 topics in busiest clusters

© 2023 New Relic, Inc. All rights reserved

What’s needed

- Kafka clients 3.3.2 / 3.4.0 - 8 months ago
- Kafka server 3.3.2 / 3.4.0
- Either helps, but full fix requires both

© 2023 New Relic, Inc. All rights reserved

Alternative - if you can’t upgrade

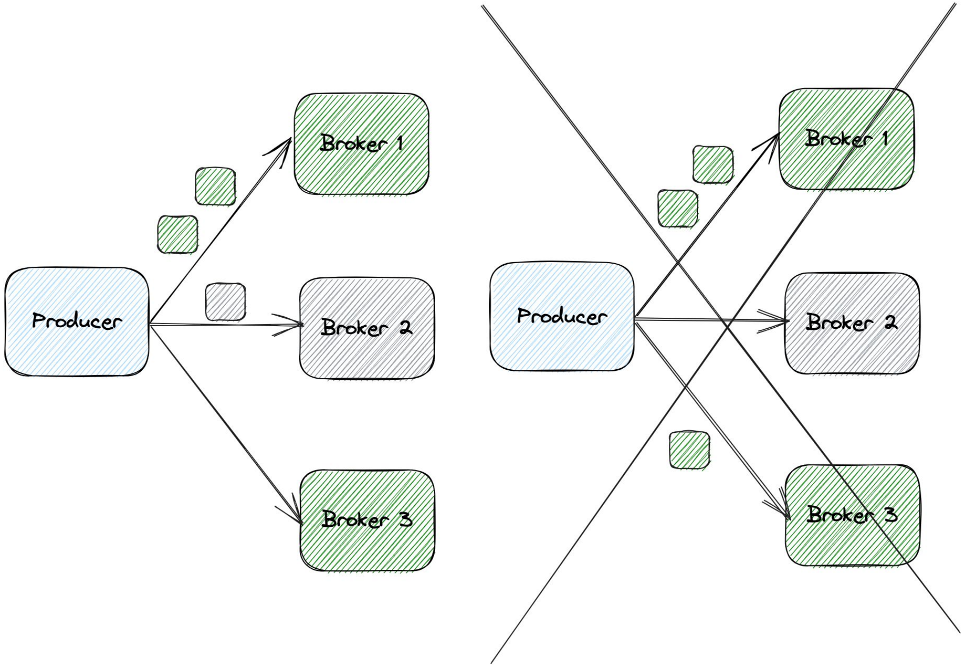
- If you are OK with more cross-AZ traffic
  - disable fetch from closest replica

© 2023 New Relic, Inc. All rights reserved

Does this solve all problems?

© 2023 New Relic, Inc. All rights reserved

Producing with a partition key



© 2023 New Relic, Inc. All rights reserved

Takeaways - producers with random partitioning

- Use Kafka clients 3.3.1 - 1 year ago
- Use the Strictly Uniform Sticky Partitioner
  - It’s the default - don’t override the partition class
  - Improves send buffer exhaustion on its own
  - Even better with partitioner.timeout.availability.ms

© 2023 New Relic, Inc. All rights reserved

Takeaways continued

- Use Kafka clients 3.3.2 / 3.4.0 for consumers
  - 8 months ago
- Use Kafka server 3.3.2 / 3.4.0

Alternative:

- If you can’t upgrade but you are OK with more cross-AZ traffic - disable fetch from closest replica

© 2023 New Relic, Inc. All rights reserved

Takeaways SRE

- Gray failures are hard to deal with
- Gamedays to reproduce gray failures
- Analyse system architecture
- Plan capacity according to your needs

© 2023 New Relic, Inc. All rights reserved

Special thanks

- Alex Thengumpalli
- Anton Rodriguez
- Luke Kirby
- Christopher Wildman
- Alex Lindeman
- Kafka Platform Team
- Streaming SRE team

© 2023 New Relic, Inc. All rights reserved

new relic

Thank you.

https://www.linkedin.com/in/michellevalentinova/

© 2023 New Relic, Inc. All rights reserved