TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Macroarea di Ingegneria
Dipartimento di Ingegneria Civile e Ingegneria Informatica

Communication in Distributed Systems
Part 2

Corso di Sistemi Distribuiti e Cloud Computing
A.A. 2022/23

Valeria Cardellini

Laurea Magistrale in Ingegneria Informatica

Message-oriented communication

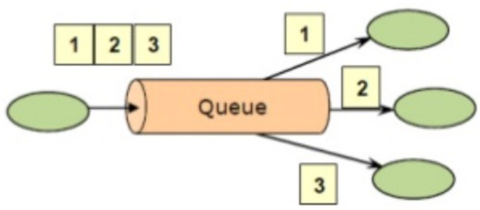
- RPC improves distribution transparency with respect to socket programming
- But still synchrony between interacting entities
  - Over time: caller waits the reply
  - In space: shared data
  - Functionality and communication are coupled
- Which communication models to improve decoupling and flexibility?
- Message-oriented communication
  - Transient
    - Berkeley socket
    - Message Passing Interface (MPI): see "Sistemi di calcolo parallelo e applicazioni" course
  - Persistent
    - **Message Oriented Middleware (MOM)**

Message-oriented middleware

- Communication middleware that supports sending and receiving messages in a persistent way
- Loose coupling among system/application components
  - Decoupling in time and space
  - Can also support synchronization decoupling
  - Goals: increase performance, scalability and reliability
  - Typically used in serverless and microservice architectures
- Two patterns:
  - **Message queue**
  - **Publish-subscribe** (pub/sub)
- And two related types of systems:
  - **Message queue system** (MQS)
  - **Pub/sub system**

Queue message pattern

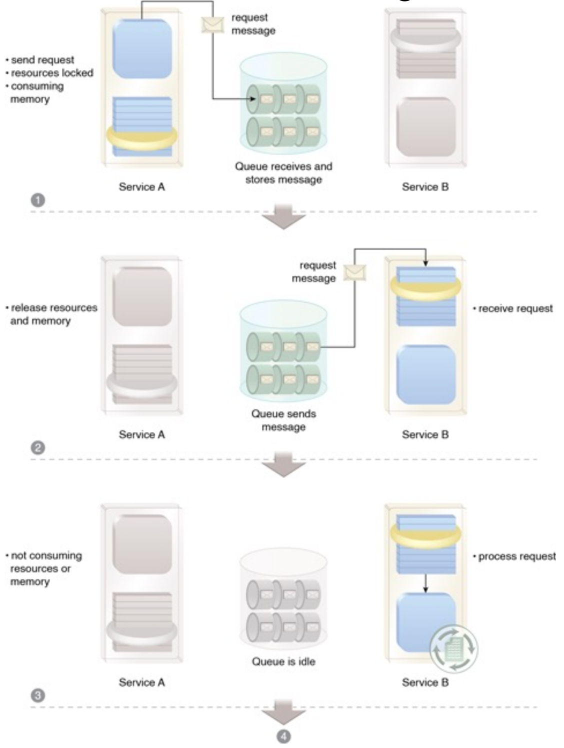
- Messages are stored on the queue until they are processed and deleted
- Multiple consumers can read from the queue
- Each message is delivered only once, to a single consumer



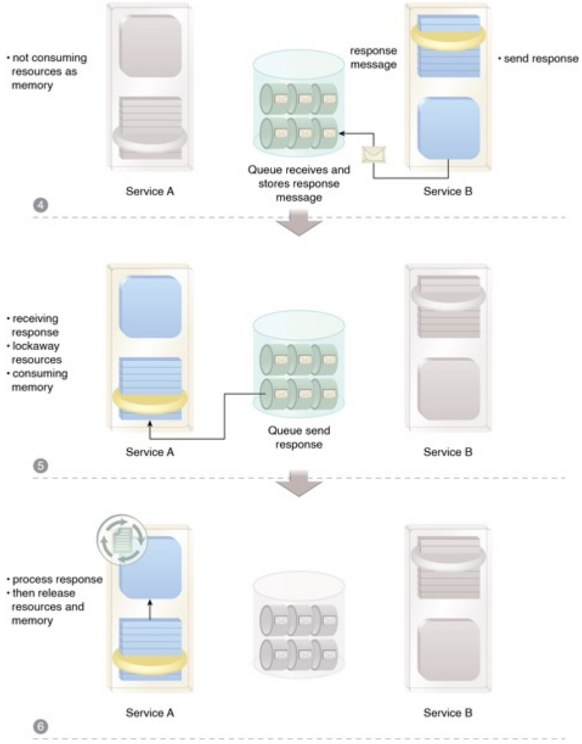
- Example of apps:
  - Task scheduling, load balancing, collaboration

Queue message pattern

A sends a message to B



B issues a response message back to A

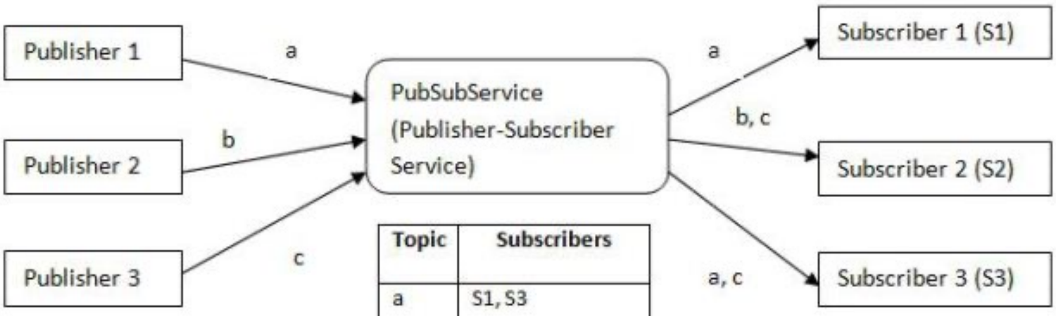


Message queue API

- Basic interface to a queue in a MQS:
  - **put**: nonblocking send
    - Append a message to a specified queue
  - **get**: blocking receive
    - Block until the specified queue is nonempty and remove the first message
    - Variations: allow searching for a specific message in the queue, e.g., using a matching pattern
  - **poll**: nonblocking receive
    - Check a specified queue for message and remove the first
    - Never block
  - **notify**: nonblocking receive
    - Install a handler (callback function) to be automatically called when a message is put into the specified queue

Publish/subscribe pattern

- Application components can publish asynchronous messages (e.g., event notifications), and/or declare their interest in message topics by issuing a *subscription*
- Each message can be delivered to multiple consumers



Topic	Subscribers
a	S1, S3
b	S2
c	S2, S3

Publish/subscribe pattern

- Multiple consumers can subscribe to topic with or without filters
- Subscriptions are collected by an *event dispatcher* component, responsible for routing events to all matching subscribers
  - For scalability reasons, its implementation is distributed
- High degree of decoupling among components
  - Easy to add and remove components: appropriate for dynamic environments

Publish/subscribe pattern

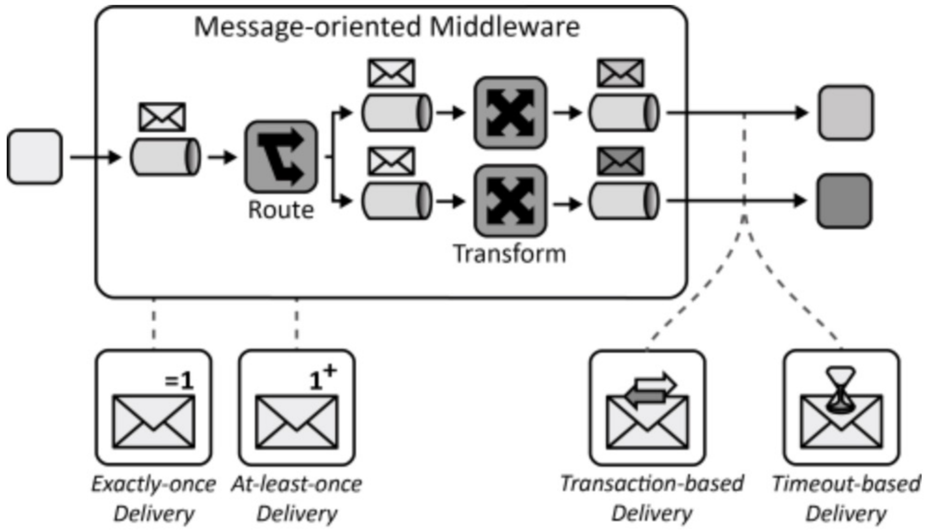
- A sibling of message queue pattern but further generalizes it by delivering a message to multiple consumers
  – Message queue: delivers messages to *only one* receiver, i.e., one-to-one communication
  – Pub/sub channel: delivers messages to *multiple* receivers, i.e., one-to-many communication

Publish/subscribe API

- Calls that capture the core of any pub/sub system:
  – publish(event): to publish an event
    - Events can be of any data type supported by the given implementation languages and may also contain meta-data
  – subscribe(filter expr, notify_cb, expiry) → sub handle: to subscribe to an event
    - Takes a filter expression, a reference to a notify callback for event delivery, and an expiry time for the subscription registration.
    - Returns a subscription handle
  – unsubscribe(sub handle)
  – notify_cb(sub_handle, event): called by the pub/sub system to deliver a matching event

MOM functionalities

- MOM handles the complexity of addressing, routing, availability of communicating application components (or applications), and message format transformations



Source: Cloud Computing Patterns
https://www.cloudcomputingpatterns.org/message_oriented_middleware

MOM functionalities

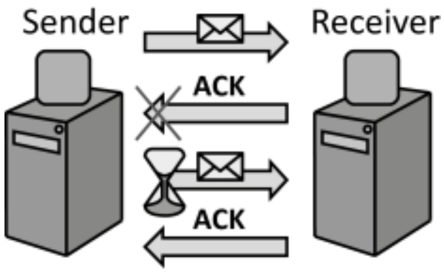
- Let us analyze
  - Delivery semantics
  - Message routing
  - Message transformations

Delivery semantics in MOM

At-least-once delivery

How can MOM ensure that messages are received successfully?

- By sending ack for each retrieved message and resending message if ack is not received
- Be careful, app should be tolerant to message duplications, i.e., it should be *idempotent* (not be affected adversely when processing the same message more than once)

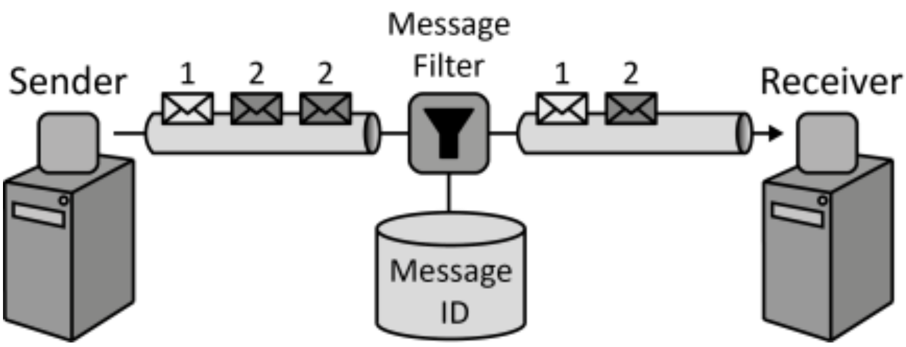


Delivery semantics in MOM

Exactly-once delivery

How can MOM ensure that a message is delivered only exactly once to a receiver?

- By filtering possible message duplicates automatically
- Upon creation, each message is associated with a unique message ID, which is used to filter message duplicates during their traversal from sender to receiver
- Messages must also survive MOM components' failures

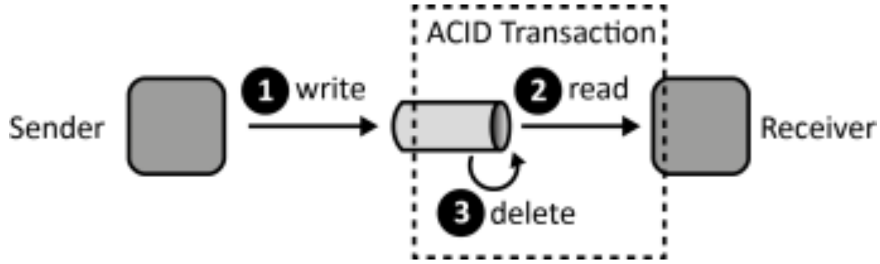


Delivery semantics in MOM

Transaction-based delivery

How can MOM ensure that messages are only deleted from a message queue if they have been received successfully?

- MOM and the receiver participate in a transaction: all operations involved in the reception of a message are performed under one transactional context guaranteeing ACID behavior

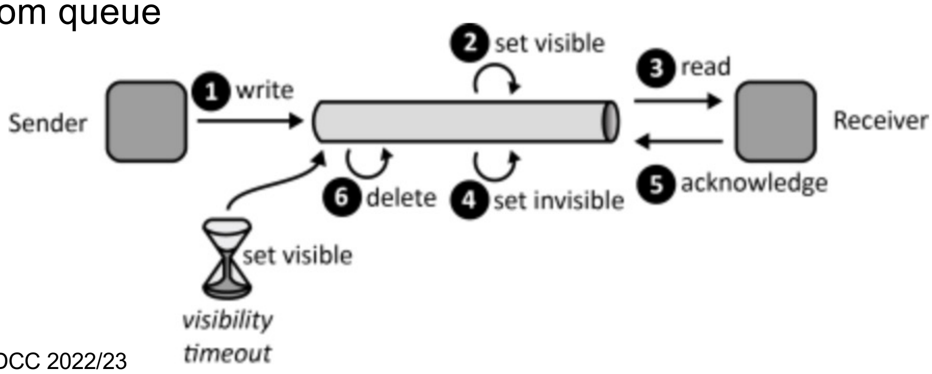


Delivery semantics in MOM

Timeout-based delivery

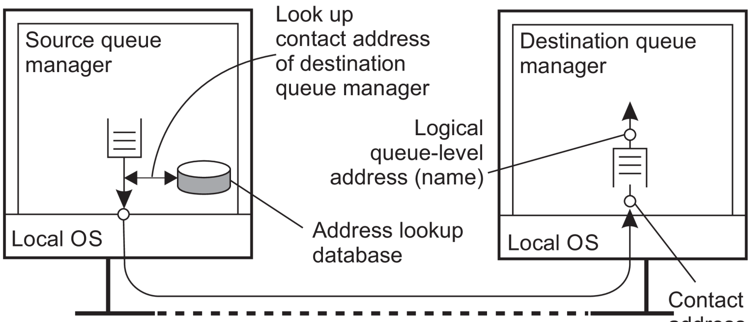
How can MOM ensure that messages are only deleted from a message queue if they have been received successfully at least once?

- Message is not deleted immediately from queue, but marked as being invisible until visibility timeout expires
- Invisible message cannot be read by another receiver
- After receiver's ack of message receipt, message is deleted from queue



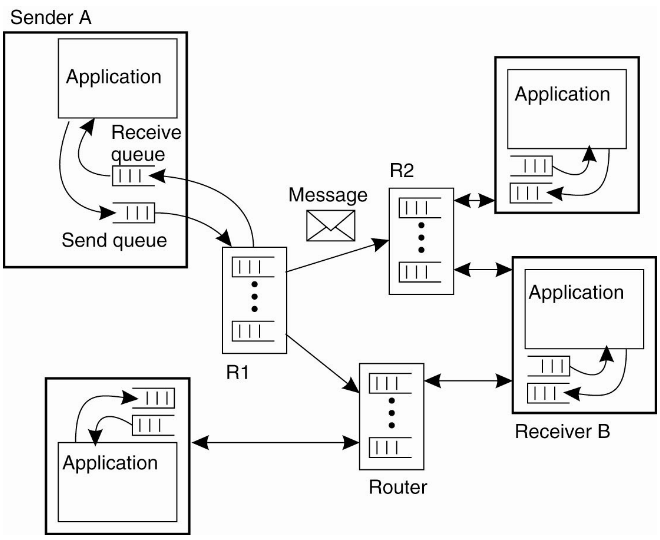
Message routing: general model

- Queues are managed by queue managers (QMs)
  - An application can put messages only into a local queue
  - Getting a message is possible by extracting it from a local queue only
- QMs need to route messages
  - Work as message-queuing "relays" that interact with distributed applications and each other
  - Realize an overlay network
  - There can also be special QMs that operate only as routers



Message routing: overlay network

- Overlay network is used to route messages
  - By using routing tables
  - Routing tables are stored and managed by QMs
- Overlay network needs to be maintained over time
  - Routing tables are often set up and managed manually: easier but …
  - Dynamic overlay networks require to dynamically manage mapping between queue names and their location

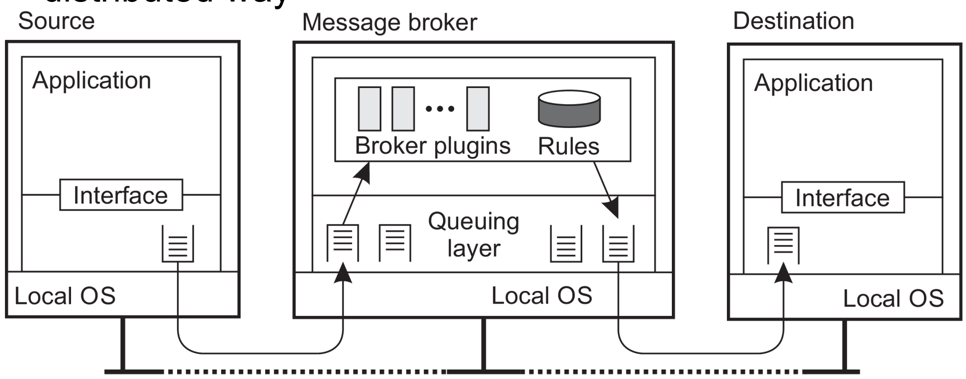


Message transformation: message broker

- New/existing apps that need to be integrated into a single, coherent system rarely agree on a common data format
- How to handle data heterogeneity?
  - We have already examined different solutions in the context of RPC
- Let's focus on ***message broker***
  - Message broker: component that usually takes care of application heterogeneity in a MOM

Message broker: general architecture

- Message broker handles application heterogeneity
  - Converts incoming messages to target format providing access transparency
  - Very often acts as an application gateway
  - Manages a repository of conversion rules and programs to transform a message of one type to another
  - May provide subject-based routing capabilities
  - To be scalable and reliable can be implemented in a distributed way



MOM frameworks

- Examples of MOM systems and libraries
  - Apache ActiveMQ http://activemq.apache.org
  - **Apache Kafka**
  - Apache Pulsar https://pulsar.apache.org
  - IBM MQ
  - NATS https://nats.io
  - Open MQ (JMS specification implementation)
  - **RabbitMQ** https://www.rabbitmq.com
  - ZeroMQ https://zeromq.org
- Clear distinction between queue message and pub/sub patterns often lacks
  - Some frameworks support both (e.g., Kafka, NATS)
  - Others not (e.g., Redis is pub/sub https://redis.io/topics/pubsub)

MOM frameworks

- Also Cloud-based products
  - **Amazon Simple Queue Service (SQS)**
  - **Amazon Simple Notification Service (SNS)**
  - CloudAMQP: RabbitMQ as a Service
  - Google Cloud Pub/Sub
  - Microsoft Azure Service Bus

RabbitMQ

- Popular open-source message broker
  https://www.rabbitmq.com

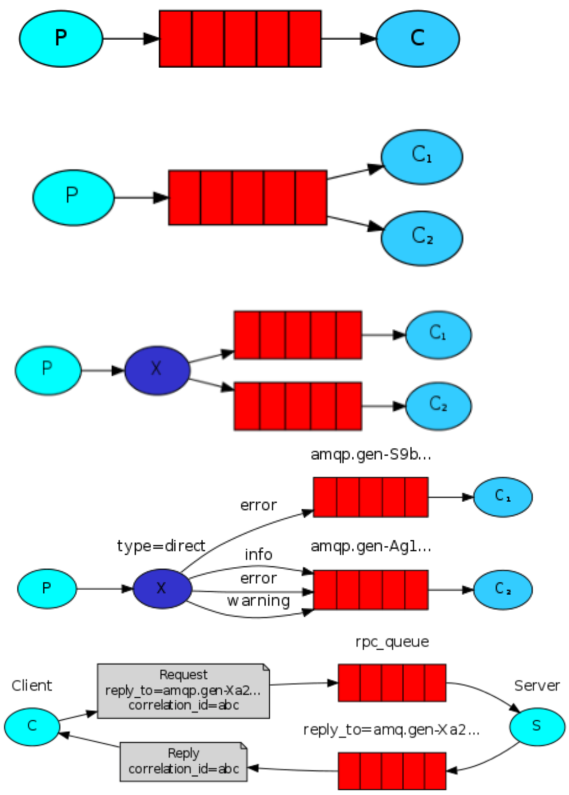
- Supports multiple messaging protocols
  - AMQP, STOMP and MQTT
- FIFO ordering guarantees at queue level
- Installation https://www.rabbitmq.com/download.html
- RabbitMQ CLI tool: `rabbitmqctl`

```
$ rabbitmqctl status
$ rabbitmqctl shutdown
```

- Also web UI for management and monitoring
- RabbitMQ broker can be distributed, for example forming a cluster https://www.rabbitmq.com/distributed.html

Using message queues: use cases

1. Store and forward messages which are sent by a producer and received by a consumer (message queue pattern)
2. Distribute tasks among multiple workers (competing consumers pattern)
3. Deliver messages to many consumers at once (pub/sub pattern) using a *message exchange*
4. Receive messages selectively: producer sends messages to an *exchange,* that selects the queue
5. Run a function on a remote node and wait for the result (request /reply pattern)



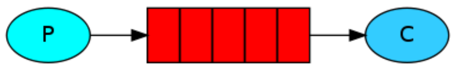
Source: RabbitMQ tutorial

Using message queues: RabbitMQ and Go

- Let's use RabbitMQ, Go and AMQP (messaging protocol, see next slides) to use a message queue for:

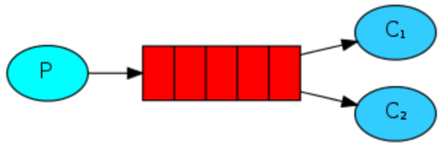
Ex. 1: Message queue pattern

https://www.rabbitmq.com/tutorials/tutorial-one-go.html



Ex. 2: Competing consumers pattern

https://www.rabbitmq.com/tutorials/tutorial-two-go.html



Code available on course site:
rabbitmq_1_hello.zip
rabbitmq_2_worker.zip

Using message queues: RabbitMQ and Go

- Preliminary steps:

1. Install RabbitMQ and start a RabbitMQ server on localhost on default port

```
$ rabbitmq-server
```

Some useful commands for `rabbitmqctl`

```
list_channels
list_consumers
list_queues
stop_app
reset
```

2. Install Go AMQP client library

```
$ go get github.com/streadway/amqp
```

See https://godoc.org/github.com/streadway/amqp for details on Go package `ampq`

Using message queues: RabbitMQ and Go

1. Message queue pattern

- Run with single producer/single consumers, multiple producers/multiple consumers
- Note that message is delivered to only one consumer
- Note that delivery is push-based

Using message queues: RabbitMQ and Go

2. Competing consumers (i.e., workers) pattern

- Version 1 (new_task_v1.go and worker_v1.go):
  - Use multiple consumers to see how queue can be used to distribute tasks among consumers in *round-robin* fashion
  - If consumer crashes after RabbitMQ delivers the message but before completing the task, the corresponding message is lost (i.e., cannot be delivered to another consumer)

    `auto-ack=true`: message is considered to be successfully delivered immediately after it is sent ("fire-and-forget")
- Version 2 (new_task_v1.go and worker_v2.go):
  - Let's set `auto-ack=false` in `Consume` and add explicit ("manual") ack in consumer to tell RabbitMQ that a particular message has been received, processed and that RabbitMQ can discard it
  - Shutdown and restart RabbitMQ: what happens to pending messages?
  - Which delivery semantics when using acks?

Using message queues: RabbitMQ and Go

2. Competing consumers (i.e., workers) pattern

- Version 3 (new_task_v3.go and worker_v3.go):
  - Let's use a durable queue so it is persisted to disk and survives RabbitMQ crash and restart
  - We need to use a new queue
  - Set `durable=true` in `QueueDeclare`
- Version 4 (new_task_v3.go and worker_v4.go):
  - Improve task distribution among multiple consumers by looking at the number of unacknowledged messages for each consumer, so to not dispatch a new message to a worker until it has processed and acknowledged the previous one
  - Use channel prefetch setting (`Qos`)

Amazon Simple Queue Service (SQS)

- Cloud-based message queue service based on polling model
  - Goal: decouple Cloud app components
  - Message queues are fully managed by AWS
  - Messages are stored in queues for a limited period of time
- Application components using SQS can run independently and asynchronously and be developed with different technologies
- Provides timeout-based delivery
  - Messages are only deleted from a message queue if they have been received properly
  - A received message is locked during processing (*visibility timeout*); if processing fails, the lock expires and the message is available again
- Can be combined with Amazon SNS
  - To push a message to multiple SQS queues in parallel

Amazon SQS: API

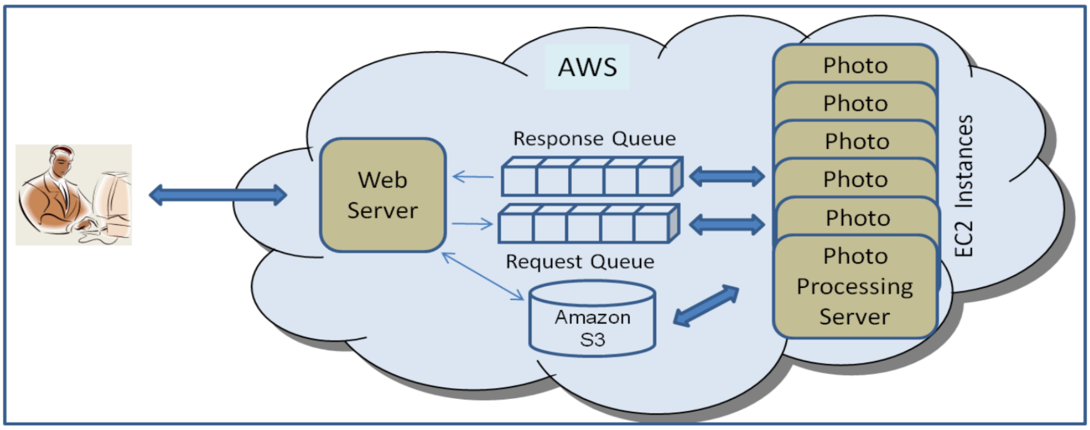
- **CreateQueue, ListQueues, DeleteQueue**
  - Create, list, delete queues
- **SendMessage, ReceiveMessage**
  - Add/receive messages to/from a specified queue (message size up to 256 KB)
  - Message larger than 256 KB?
    - Put in queue reference to message payload stored in S3
- **DeleteMessage**
  - Remove a received message from a specified queue (the component must delete the message after receiving and processing it)

Amazon SQS: API

- **ChangeMessageVisibility**
  - Change the visibility timeout of a specified message in a queue (when received, the message remains in the queue upon it is deleted explicitly by the receiver)
  - Default visibility timeout is 30 seconds
- **SetQueueAttributes, GetQueueAttributes**
  - Control queue settings, get information about a queue

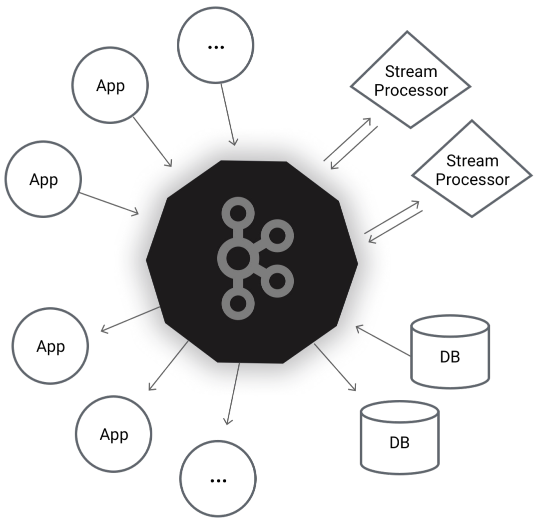
Amazon SQS: example

- Cloud app for online photo processing service
- Let's use SQS to achieve app components decoupling, load balancing and fault tolerance http://bit.ly/2gwJFBw



Apache Kafka

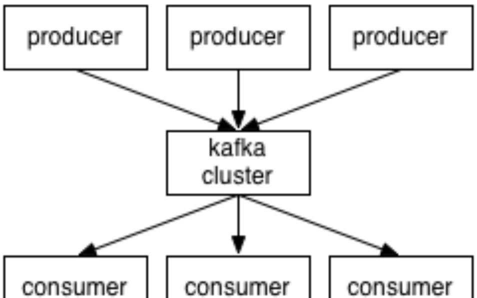
- General-purpose, distributed pub/sub system
- Originally developed in 2010 by LinkedIn
- Used at scale by tech giants (Netflix, Uber, LinkedIn, …)
- Written in Scala
- Horizontally scalable
- Fault-tolerant
- High throughput ingestion
  - Billions of messages
- Not only messaging, also data processing
  - We focus on messaging



https://kafka.apache.org/documentation/

Kreps et al., Kafka: A Distributed Messaging System for Log Processing, NetDB'11

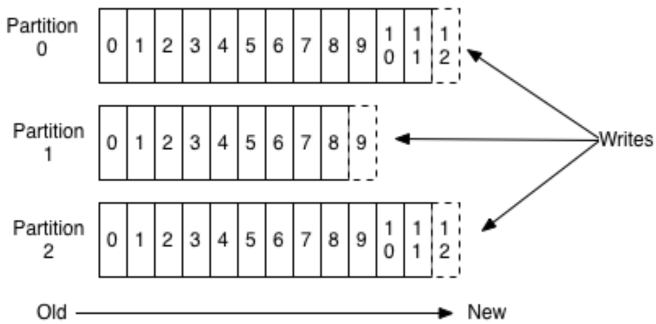
Kafka at a glance



- Kafka stores feeds of messages (or records) in categories called topics
  - A topic can have 0, 1, or many consumers subscribing to data written to it
- Producers: publish messages to a Kafka topic
- Consumers: subscribe to topics and process the feed of published messages
- Kafka cluster: distributed log of data over servers known as *brokers*
  - A broker is responsible for receiving and storing published data

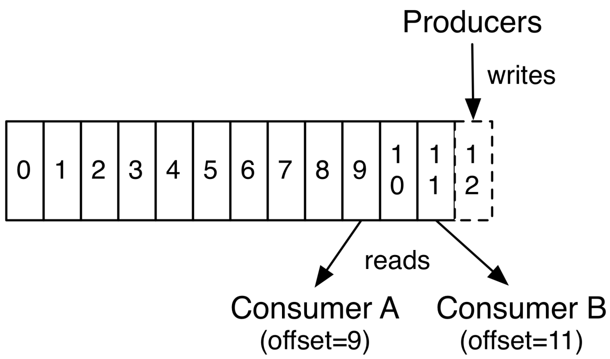
Kafka: topics and partitions

- Topic: category to which a message is published
- For each topic, Kafka cluster maintains a partitioned log
  - **Log** (data structure!): *append-only*, *totally-ordered* sequence of messages ordered by time
- Partitioned log: each topic is split into a pre-defined number of partitions
  - Partition: unit of parallelism for the topic (allows for parallel access)



Kafka: partitions

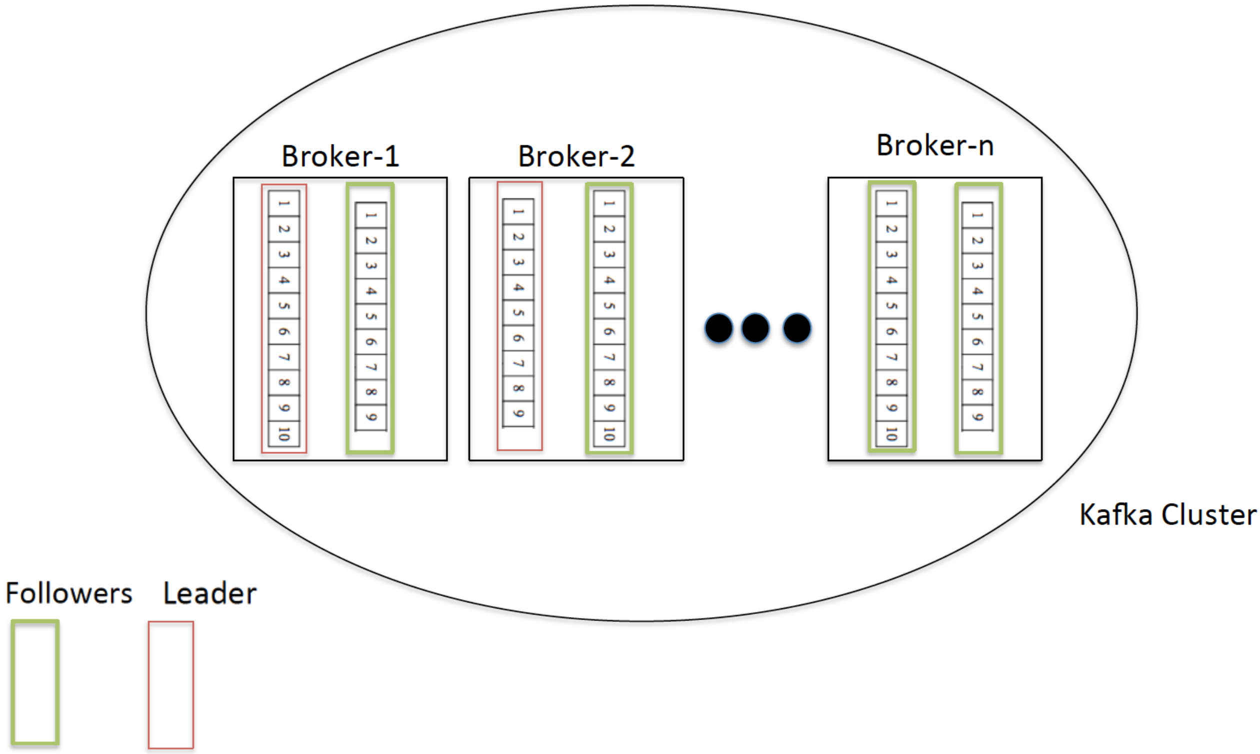
- Producers publish their records to partitions of a topic
- Consumers consume records published on a topic
- Each partition is an *ordered*, *numbered*, *immutable* sequence of records that is continually *appended to*
  - Like a commit log
- Each record is associated with a monotonically increasing sequence number, called offset



Kafka: partitions and design choices

- To improve scalability: partitions are *distributed* across brokers
  - By distributing partitions on multiple brokers, IO throughput increases
  - Parallel reads and writes on partitions of the same topic
    - Multiple producers can write in parallel
    - A single topic can be read by multiple consumers
- To improve fault tolerance: each partition can be *replicated* across a configurable number of brokers
  - Driven by *replication-factor*
- Each partition has one leader broker and 0 or more followers
  - followers > 0 in case of replication

Kafka: partition leader and followers



Kafka: partitions and design choices

- To simplify data consistency management: leader handles read and write requests
  - Producers read from leader, consumers write to leader
  - Followers replicates the leader and acts as backups
  - Followers can be *in-sync* (i.e., fully updated replica) with the leader or *out-of-sync*
- To share responsibility and balance load: each broker is leader for some of its partitions and follower for others
  - Brokers rely on Apache Zookeeper for coordination

Kafka: producers

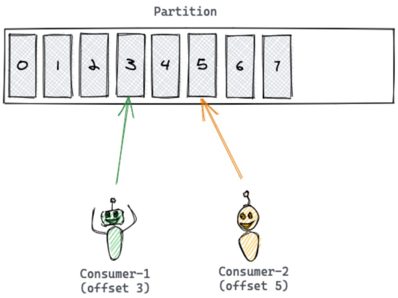
- Producers = data sources
- Publish data to topics of their choice
  – Producer sends data directly (i.e., without any routing tier) to the broker that is the leader for the partition
- Producer is responsible for choosing which record to assign to which partition within the topic: how?
  – *Key-based* partitioned, i.e., the producer uses a partition key to direct messages to a specific partition
    - E.g., if user id is the key, all data for a given user will be published in the same partition
  – Round-robin (default, if key is not specified)
- Multiple producers can write to the same partition

Design choice for consumers

- Push or pull model for consumers?
- **Push** model
  – Broker actively pushes messages to consumers
  – Challenging for broker to deal with different types of consumers as it controls the rate at which data is transferred
  – Need to decide whether to send a message immediately or accumulate more data and then send
- **Pull** model
  – Consumer is in charge of retrieving messages from broker
  – Consumer has to maintain an offset to identify the next message to be transmitted and processed
  ✓ Better scalability (less burden on brokers) and flexibility (different consumers with diverse needs and capabilities)
  ✗ In case broker has no data, consumers may end up busy waiting for data to arrive

Kafka: consumers

- Kafka uses a pull approach for consumers
  http://kafka.apache.org/documentation.html#design_pull
- Consumer uses the offset to keep track of which messages it has already consumed
- A partition can be consumed by more consumers, each reading at different offsets



- How can consumer read in a fault-tolerant way?
  – Once the consumer reads message, it stores its committed offset in a safe place (either Zookeeper or a special Kafka topic called `__consumer_offsets`)
  – After recovering from crash, consumer can replay messages using committed offset
  – By default, `auto-commit` is enabled

Hands-on Kafka

- Preliminary steps:
  – Download and install Kafka http://kafka.apache.org/downloads
    • Configure Kafka properties in `server.properties` (e.g., `listeners` and `advertised.listeners`)
  – Start Kafka environment

    Start ZooKeeper (default port: 2181)

    `$ zookeeper-server-start zookeeper.properties`

    Alternatively `$ zKserver start`

    Start Kafka broker (default port: 9092)

    `$ kafka-server-start server.properties`

- Let's use Kafka CLI tools to create a topic, publish and consume some events to/from the topic and delete the topic

Hands-on Kafka

- Create a topic named test with 1 partition and non-replicated

  `bootstrap_servers`: specify one broker to bootstrap initial cluster metadata

```
$ kafka-topics --create --bootstrap-server
  localhost:9092 --replication-factor 1 --partitions
  1 --topic test
```

- Write some messages into topic

```
$ kafka-console-producer --broker-list localhost:9092
  --topic test
> first message
> another message
```

- Read messages from beginning of topic partition

```
$ kafka-console-consumer --bootstrap-server
localhost:9092 --topic test --from-beginning
```


Hands-on Kafka

- Read messages using some offset (e.g., 2)

```
$ kafka-console-consumer --bootstrap-server
localhost:9092 --topic test --offset 2 --partition 0
```

- List available topics

```
$ kafka-topics --list --bootstrap-server localhost:9092
```

- Delete topic

```
$ kafka-topics --delete --bootstrap-server localhost:9092
--topic test
```

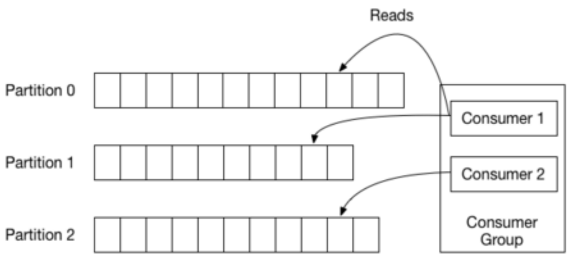
- Stop Kafka and Zookeeper

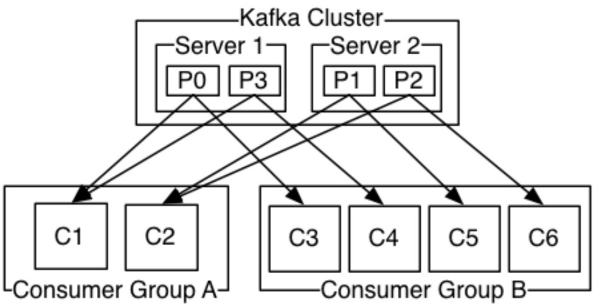
```
$ kafka-server-stop
$ zookeeper-server-stop
```

  Alternatively `$ zKserver stop`

Kafka: consumer group

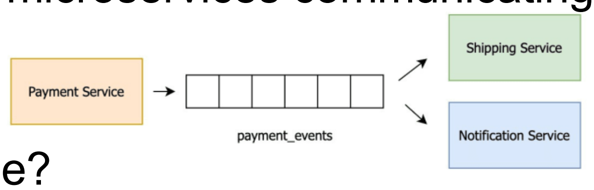
- Consumer Group: set of consumers which cooperate to consume data from some topics and share a group ID
  - A Consumer Group maps to a *logical* subscriber
  - Topic partitions are divided among consumers in the group for load balancing and can be reassigned in case of consumer join/leave
  - Every message will be delivered to *only one* consumer in group
  - Every group maintains its offset per topic partition



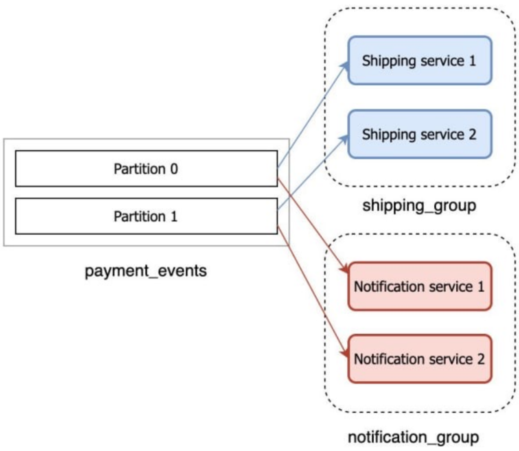


Kafka: consumer group

- How to have many consumers reading the same messages from the topic?
  - Need to use different group IDs
- Example: 2 microservices communicating using Kafka



- How to scale?



Kafka: ordering guarantees

- Messages published by producer to topic partition will be appended in the order they are sent
- Consumer sees records in the order they are stored in the partition
- Strong guarantee about ordering *only within a partition*
  - Total order over messages within a partition, i.e., *per-partition ordering*
  - Kafka cannot preserve message order between different topic partitions
- However, per-partition ordering plus ability to partition messages by key among topic partitions, is sufficient for most applications

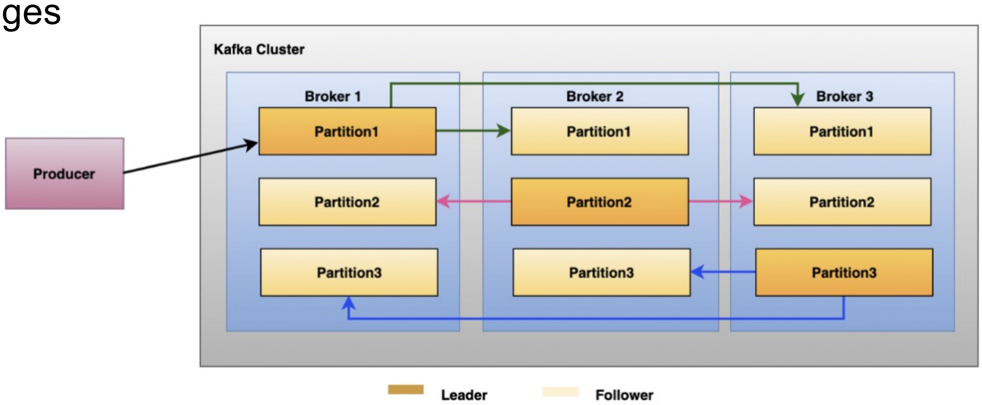
Kafka: delivery semantics

- Delivery guarantees supported by Kafka
  - **At-least-once** (default): guarantees no message loss but duplicated messages, possibly out-of-order
    - Which mechanism on producer side?
    - Set `acks=1` on producer
    - On consumer side: pull model and offset commit
  - **Exactly-once**: guarantees no loss and no duplicates, but requires expensive end-to-end 2PC
    - Set `acks=all` on producer
    - Not fully exactly-once
    - Support depends on destination system
  - User can also implement **at-most-once**: messages may be lost but are never re-delivered
    - By disabling retries on producer (i.e., `acks=0`) and committing offsets on consumer prior to processing a message

See https://kafka.apache.org/documentation/#semantics

Kafka: fault tolerance

- Kafka replicates partitions for fault tolerance
  - Leader coordinates to update followers with new messages



- In case of leader crash, a follower can be elected as new leader with the help of Zookeeper

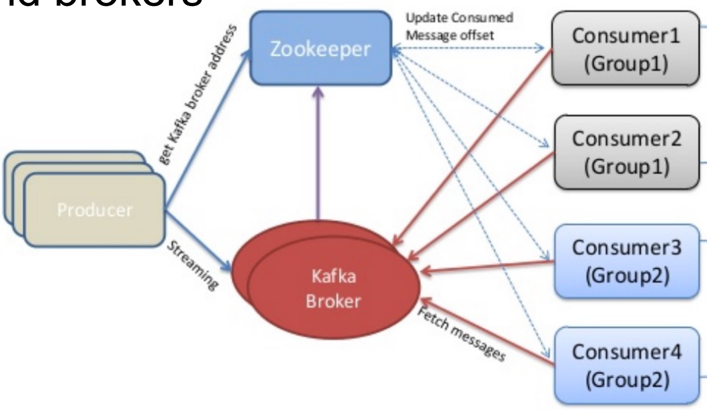
Kafka: fault tolerance

- Kafka makes a message available for consumption only after all the followers acknowledge to the leader a successful write
  - Messages may not be immediately available for consumption (tradeoff between consistency and availability)
  - This behavior can be relaxed if strong guarantee is not required (setting `acks=1`)

- Kafka retains messages for a configured period of time
  - Messages can be replayed in case of consumer crash
  - To free up disk space, messages have a TTL; upon TTL expiry, messages are marked for deletion

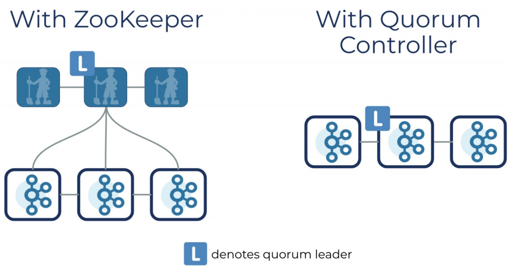
Kafka and ZooKeeper

- Zookeeper: hierarchical, distributed key-value store
  - Coordination and synchronization service for large distributed systems
  - Often used for leader election
  - Used within many open-source distributed systems besides Kafka (Apache Mesos, Storm, …)
- Kafka uses ZooKeeper to coordinate between producers, consumers and brokers
- ZooKeeper stores Kafka metadata
  - List of brokers
  - List of producers
  - List of consumers and their offsets



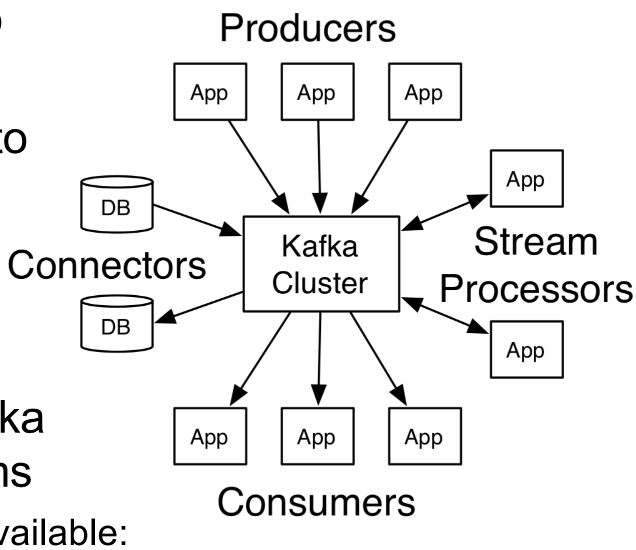
From ZooKeeper to KRaft

- Zookeeper cons
  - ✗ Different system for metadata management and consensus
  - ✗ Can become bottleneck as Kafka cluster grows
- New release: ~~Zookeeper~~ Apache Kafka Raft (KRaft)
  - Kafka cluster metadata is stored in Kafka cluster itself
  - ✓ Simpler architecture
  - ✓ Faster and more scalable metadata update operations
  - Metadata is also replicated to all brokers, making failover from failure faster
  - Consensus protocol based on Raft



Kafka: APIs

- Four core APIs https://kafka.apache.org/documentation/#api
- Producer API: allows apps to publish records to topics
- Consumer API: allows apps to read records from topics
- Connect API: allows implementing reusable connectors (producers or consumers) that connect Kafka topics to apps or data systems
  - Many connectors are already available: AWS S3, ActiveMQ, RabbitMQ, MySQL, Postgres, AWS Lambda, …



Kafka: APIs

- Streams API: allows transforming streams of data from input topics to output topics
  - Kafka is not only a pub/sub system but also a real-time streaming platform
    - Use Kafka Streams to process data in pipelines consisting of multiple stages
- Kafka APIs support Java and Scala only

Kafka: client library

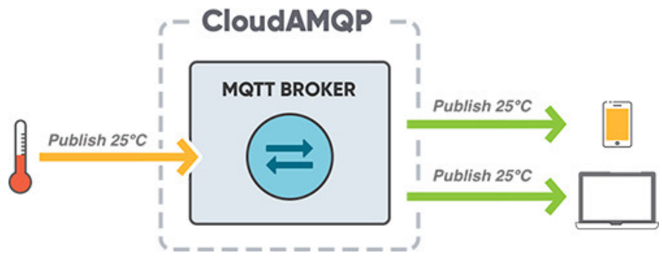
- JVM internal client
- Plus rich ecosystem of client library, among which:
  - Go
    https://github.com/Shopify/sarama
    https://github.com/segmentio/kafka-go
  - Python
    https://github.com/confluentinc/confluent-kafka-python/

Protocols for MOM

- Not only systems but also open standard protocols for message queues
  - **AMQP** Advanced Message Queueing Protocol
    - Binary protocol
  - MQTT Message Queue Telemetry Transport
    - Binary protocol
  - STOMP Simple (or Streaming) Text Oriented Messaging Protocol
    - Text-based protocol
- Goals:
  - Platform- and vendor-agnostic
  - Provide interoperability between different MOMs

Messaging protocols and IoT

- Often used in Internet of Things (IoT)
  - Use message queueing protocol to send data from sensors to services that process those data



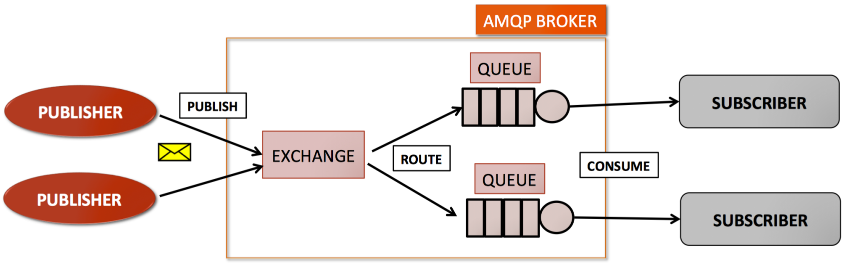
  - Exploit all MOM advantages seen so far:
    - Decoupling
    - Resiliency: MOM provides a temporary message storage
    - Traffic spikes handling: data will be persisted in MOM and processed eventually

AMQP: characteristics

- Open-standard protocol for MOM, supported by industry
  - Current version: 1.0 http://docs.oasis-open.org/amqp/core/v1.0/amqp-core-complete-v1.0.pdf
  - Approved in 2014 as ISO and IEC International Standard
- Binary, application-level protocol
  - Based on TCP protocol with additional reliability mechanisms (at-most once, at-least once, exactly once delivery)
- Programmable protocol
  - Several entities and routing schemes are primarily defined by apps
- Implementations
  - Apache ActiveMQ, RabbitMQ, Apache Qpid, Azure Event Hubs, Pika (Python implementation), …

AMQP: model

- AMQP architecture involves 3 main actors:
  - Publishers, subscribers, and brokers



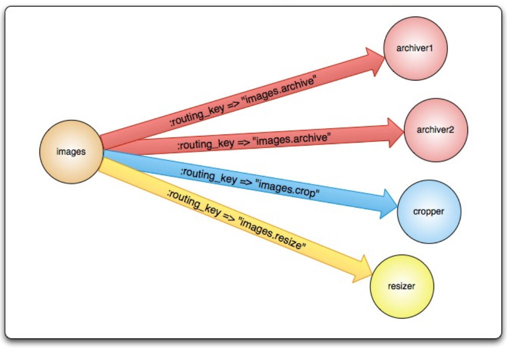
- AMQP entities (within broker): queues, exchanges and bindings
  - Messages are published to *exchanges* (like post offices or mailboxes)
  - Exchanges distribute message copies to *queues* using rules called *bindings*
  - AMQP brokers either push messages to consumers subscribed to queues, or consumers pull messages from queues on demand

https://bit.ly/2oP683F

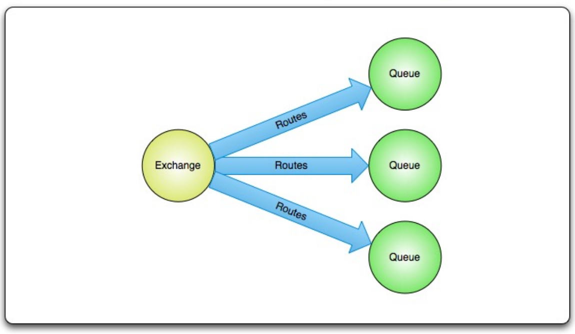
AMQP: routing

- Bindings:
  - Direct exchange: delivers messages to queues based on message routing key
  - Fanout exchange: delivers messages to all queues that are bound to it

Direct exchange routing



Fanout exchange routing

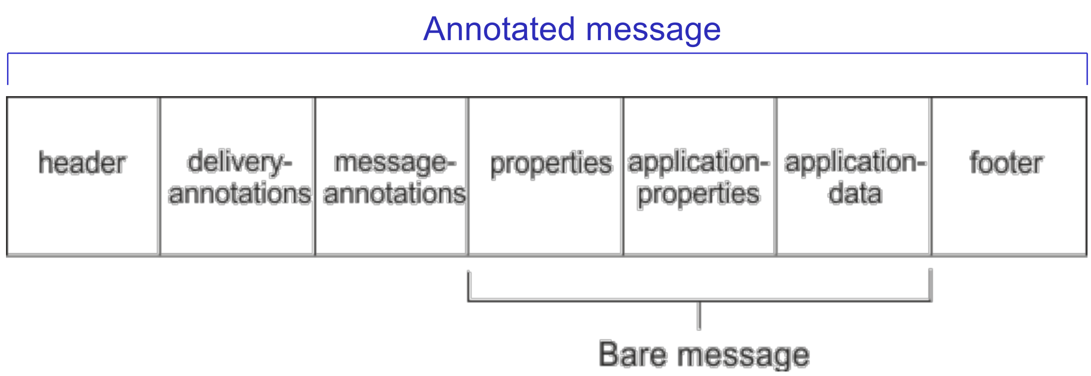


AMQP: routing

- Bindings:
  - Topic Exchange: delivers messages to one or many queues based on topic matching
    - Often used to implement various publish/subscribe pattern variations
    - Commonly used for multicast routing of messages
    - Example use: distributing data relevant to specific geographic location (e.g., points of sale)
  - Headers Exchange: delivers messages based on multiple attributes expressed as headers
    - To route on multiple attributes that are more easily expressed as message headers than a routing key

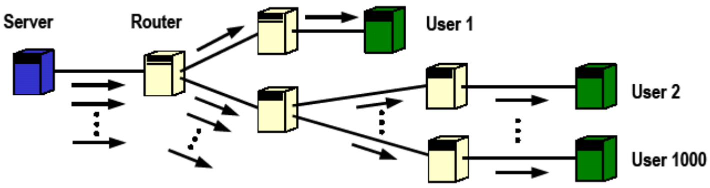
AMQP: messages

- AMQP defines two types of messages:
  - Bare messages, supplied by sender
  - Annotated messages, seen at receiver and added by intermediaries during transit
- Message header conveys delivery parameters
  - Including durability requirements, priority, time to live

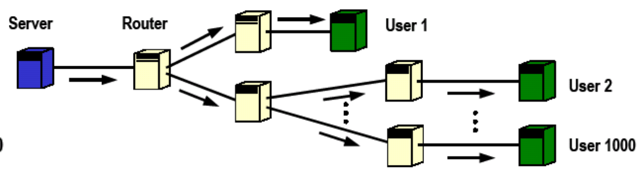


Multicast communication

- **Multicast communication**: group communication pattern in which data is sent to *multiple* receivers at once
  - Broadcast communication*:* special case of multicast, in which data is sent to *all* receivers
  - Examples of one-to-many multicast apps: video/audio resource distribution, file distribution
  - Examples of many-to-many multicast apps : conferencing tools, multiplayer games, interactive distributed simulations
- Traditional unicast communication does not scale



Unicast of video to 1000 users



Multicast of video to 1000 users

Types of multicast

- How to realize multicast?
  - **Network**-level multicast (IP-level)
  - **Application**-level multicast

Network-level multicast

- Packet replication and routing managed by routers
- IP Multicast (IPMC) based on groups
  - IPMC generalizes UDP with one-to-many behavior
  - Receivers use a special IP address which is shared among multiple hosts
  - *Group*: set of hosts interested in same multicast app; they are identified by the same multicast IP address
    - Multicast IP address from 224.0.0.0 to 239.255.255.255
  - IGMP (Internet Group Management Protocol) to join group
- Limited use:
  - Lack of large-scale support (only ~5% of ASs)
  - Difficult to keep track of group membership
  - Banned in some contexts, e.g., in Cloud data centers because of *broadcast storm* problem (exponential increase of network traffic with possible saturation)

Application-level multicast

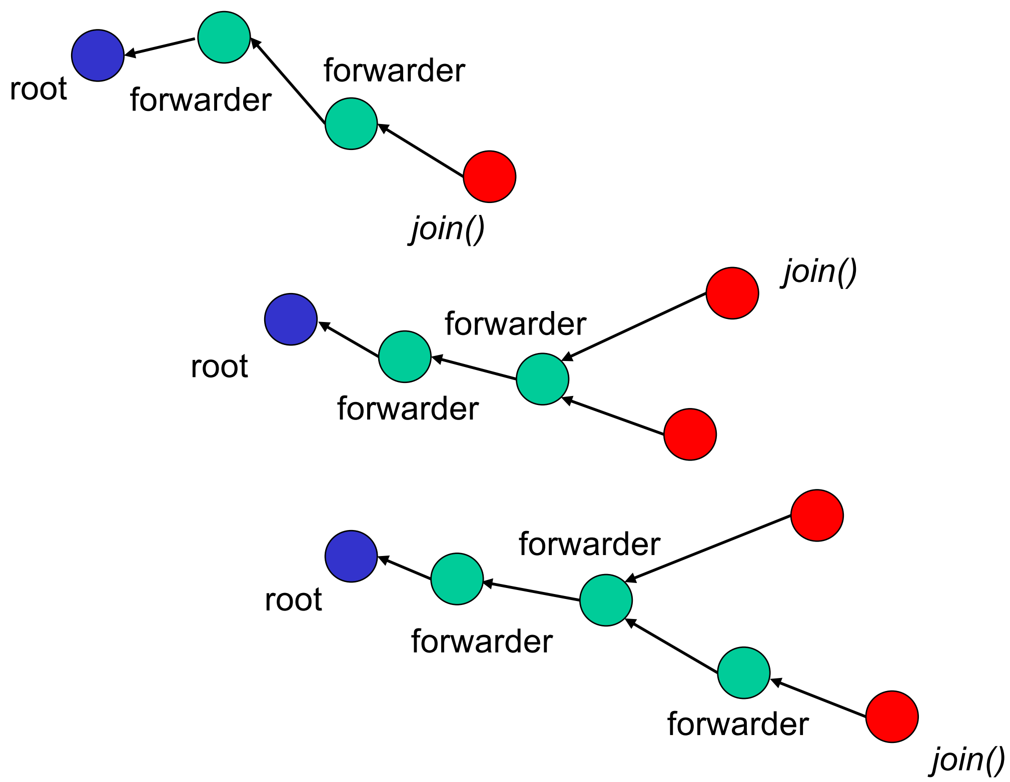
- Packet replication and routing managed by end hosts
- Basic idea:
  - Organize nodes into overlay network
  - Use overlay network to disseminate data
- Application-level multicast:
  - *Structured*
    - Explicit communication paths
    - How to build structured overlay network?
      - Tree: only one path between each node pair
      - Mesh: multiple paths between each node pair
  - *Unstructured*
    - Based on flooding
    - Based on gossiping

Structured application-level multicast: tree

- Let's consider how to build application-level multicast tree in Scribe
  - Scribe: pub/sub system with decentralized architecture and based on Pastry DHT (let's use Chord for request routing)
  1. Multicast initiator node generates multicast identifier *mid*
  2. Initiator lookups *succ*(*mid*) using DHT
  3. Request is routed to *succ*(*mid*), which becomes root of multicast tree
  4. If node *P* wants to join multicast, it sends join request to root
  5. When request arrives at *Q*:
     - *Q* has not seen a join request for *mid* before ⇒ *Q* becomes *forwarder*, *P* becomes Q's child; join request is forwarded by *Q*
     - *Q* knows about tree ⇒ *P* becomes Q's child; no need to forward join request

Castro et al., Scribe: A large-scale and decentralised application-level multicast infrastructure, *IEEE JSAC*, 2002

Structured application-level multicast: tree



Non-structured application-level multicast

- How to realize non-structured application-level multicast?
  - Flooding: already studied
    - Node *P* sends multicast message *m* to all its neighbors
    - In its turn, each neighbor will forward that message, except to *P*, and only if it had not seen *m* before
  - Random walk: already studied
    - With respect to flooding, *m* is sent only to one randomly chosen node
  - **Gossiping**

Gossip-based protocols

- Gossip-based protocols (or algorithms) are probabilistic; aka *epidemic* algorithms
  - Gossiping effect: information can spread within a group just as it would be in real life
  - Strongly related to epidemics, by which a disease is spread by infecting members of a group, which in turn can infect others
- Allow information dissemination in large-scale networks through random choice of successive receivers among those known to sender
  - Each node sends the message to a randomly chosen subset of nodes in the network
  - Each node that receives it will send a copy to another subset, also chosen at random, and so on

Origin of gossip-based protocols

- Gossiping protocols proposed in 1987 by Demers et al. in a work on data consistency in replicated databases composed of hundreds of servers
  - Basic idea: assume there are no write conflicts (i.e., independent updates)
  - Update operations are initially performed on one or a few replicas
  - A replica passes its updated state to only a few neighbors
  - Update propagation is *lazy*, i.e., not immediate
  - Eventually, each update should reach every replica

Demers et al., Epidemic Algorithms for Replicated Database Maintenance, *Proc. of 6th Symp. on Principles of Distributed Computing*, 1987.

Why gossiping in large-scale DSs?

- Several attractive properties of gossip-based information dissemination for large-scale distributed systems
  - Simplicity of gossiping algorithms
  - No centralized control or management (and related bottleneck)
  - Scalability: each node sends only a limited number of messages, independently from the overall system size
  - Reliability and robustness: thanks to message redundancy

Who uses gossiping? Some example

- AWS S3 "uses a gossip protocol to quickly spread information throughout the S3 system. This allows Amazon S3 to quickly route around failed or unreachable servers, among other things"
- Amazon's Dynamo uses a gossip-based failure detection service
- BitTorrent uses a gossip-based basic information exchange
- Cassandra uses Gossip protocol for group membership and failure detection of cluster nodes
- See gossip dissemination pattern
https://martinfowler.com/articles/patterns-of-distributed-systems/gossip-dissemination.html

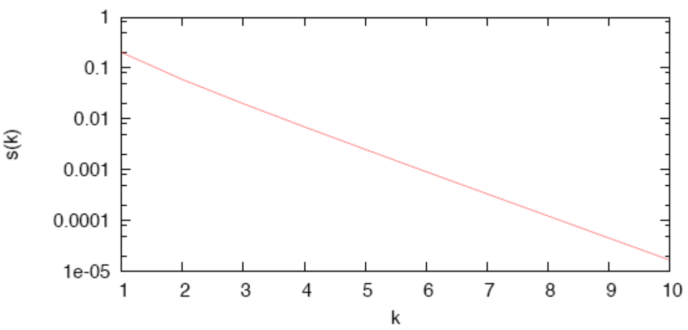
Propagation models

- Let's consider the two principle operations
  – Pure gossiping and anti-entropy
- Pure gossiping (or rumor spreading): a node which has just been updated (i.e., contaminated) selects *F* (*F* >= 1) other peers to forward the update message to (contaminating them)
- Anti-entropy: each node regularly chooses another node at random and exchanges state differences, leading to identical states at both afterwards

Pure gossiping

- Node *P* selects randomly node *Q* and forwards the update message to it
- If *Q* was already updated, *P* may lose interest in spreading the update any further and with probability $1/k$ stops contacting other nodes
- The fraction *s* of ignorant nodes (that have not yet been updated) is $s = e^{-(k+1)(1-s)}$

If *k* increases, the probability of spreading the update increases as well

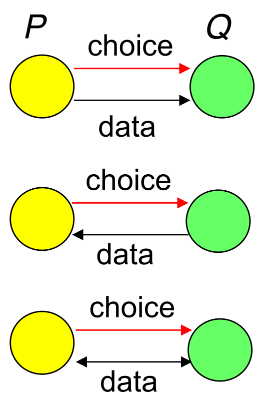


Consider 10,000 nodes		
k	*s*	N_s
1	0.203188	2032
2	0.059520	595
3	0.019827	198
4	0.006977	70
5	0.002516	25
6	0.000918	9
7	0.000336	3

- To improve message spreading (especially when $1/k$ is high), let's combine pure gossiping with anti-entropy

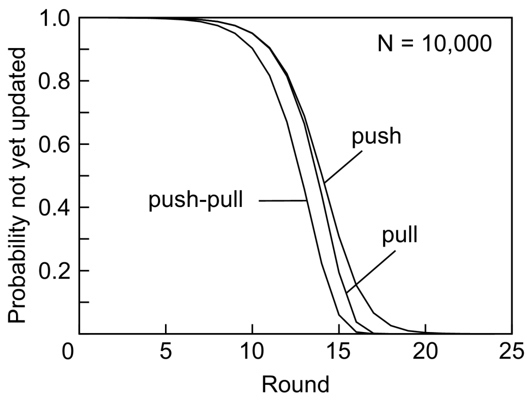
Anti-entropy

- Goal: increase node state similarity, thus decreasing "disorder" (reason for name!)
- Node *P* selects node *Q* randomly: how does P update *Q*?
- 3 different update strategies:
  - **push**: *P* only pushes its own updates to *Q*
  - **pull**: *P* only pulls in new updates from *Q*
  - **push-pull**: *P* and *Q* send updates to each other, i.e., *P* and *Q* exchange updates



Anti-entropy: performance

- Push-pull
  - Fastest strategy
  - Takes O(log(*N*)) rounds to disseminate updates to all *N* nodes
  - *Round* (or *gossip cycle*): time interval in which every node takes the initiative to start an exchange



General schema of gossiping protocols

- Two nodes *P* and *Q*, where *P* selects *Q* to exchange information with
  - *P* runs at each new gossip round (every Δ time units)

Active thread (node P):		Passive thread (node Q):
(1) **selectPeer**(&Q);		(1)
(2) **selectToSend**(&bufs);		(2)
(3) sendTo(Q, bufs);	----->	(3) receiveFromAny(&P, &bufr);
(4)		(4) **selectToSend**(&bufs);
(5) receiveFrom(Q, &bufr);	<-----	(5) sendTo(P, bufs);
(6) **selectToKeep**(cache, bufr);		(6) **selectToKeep**(cache, bufr);
(7) processData(cache);		(7) processData(cache)

selectPeer: randomly select a neighbor
selectToSend: select some entries from local cache
selectToKeep: select which received entries to store into local cache; remove repeated entries

Kermarrec and van Steen, Gossiping in Distributed Systems, *ACM Operating System Review,* 2007

Framework of gossip-based protocols

- Simple? Not quite getting into the details…
- Some crucial aspects
  - Peer selection
    - E.g., *Q* can be uniformly chosen from set of currently available (i.e., alive) nodes
  - Data exchanged
    - Exchange is highly application-dependent
    - Choice of update strategy
  - Data processing
    - Again, highly application-dependent

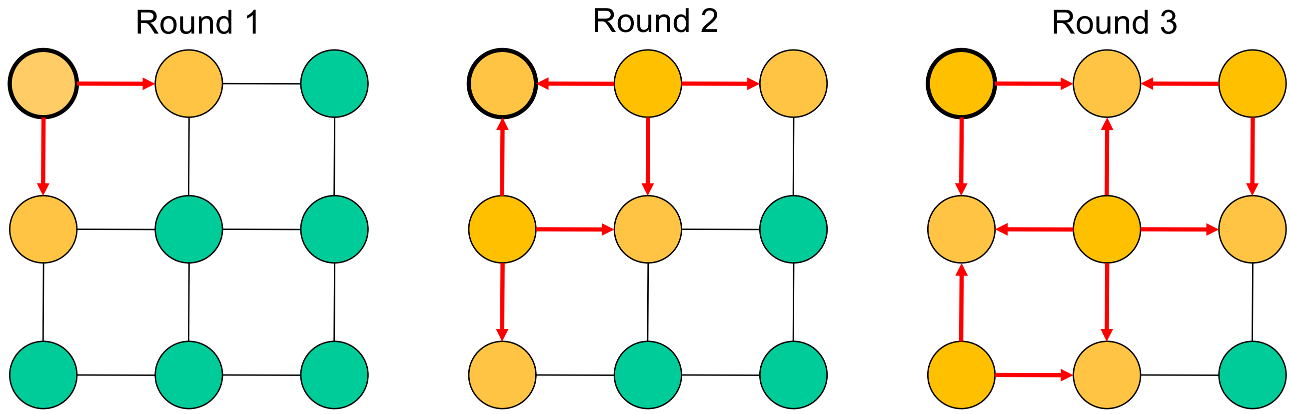
Implementing a gossiping protocol

Which issues to address when implementing a gossiping protocol?

- Membership: how nodes can get to know each other and how many acquaintances to have
- Network awareness: how to make node links reflect network topology for satisfactory performance
- Cache management: what information to discard when node's cache is full
- Message filtering: how to consider nodes' interest in update message and reduce the likelihood that they will receive information they are not interested in

Gossiping vs flooding: example

- Information dissemination is the classic and most popular application of gossiping protocols in DSs
  - Gossiping is more efficient than flooding
- Flooding-based information dissemination
  - Each node that receives message forwards it to its neighbors (let's consider *all* neighbors, including the sender)
  - Message is eventually discarded when TTL=0

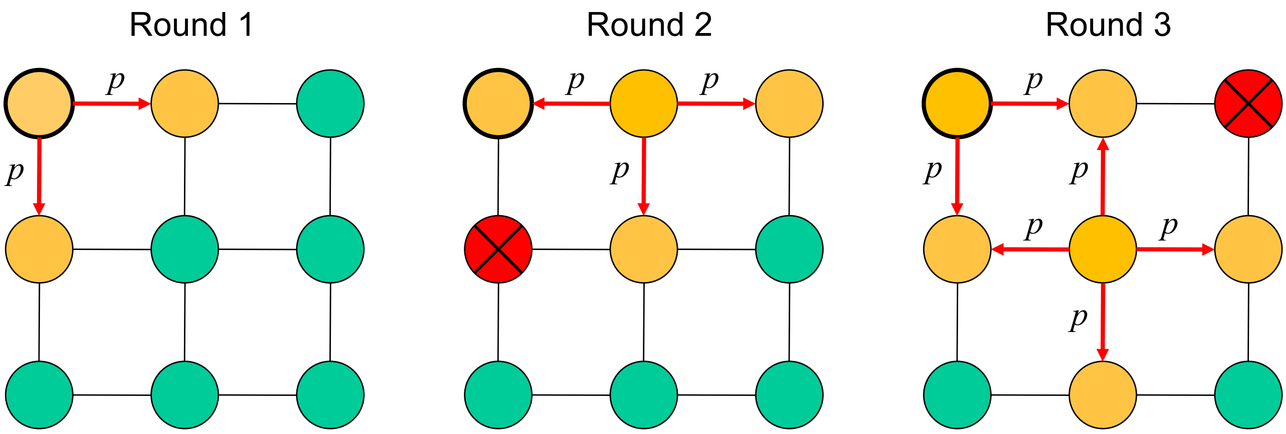


Sent messages: 18
Reached nodes: 8 out of 9

Gossiping vs flooding: example

- Let's use only pure gossiping
  - Message is sent to neighbors with probability *p*

for each msg *m*
    if random(0,1) < *p* **then** send *m*



Sent messages: 11
Reached nodes: 7 out 9

Gossiping vs flooding

- Gossiping features
  - Probabilistic
  - Takes a localized decision but results in a global state
  - Lightweight
  - Fault-tolerant
- Flooding has some advantages
  - Universal coverage and minimal state information
  - … but it floods the networks with redundant messages
- Gossiping goals
  - Reduce the number of redundant transmissions that occur with flooding while trying to retain its advantages
  - … but due to its probabilistic nature, gossiping cannot guarantee that all the peers are reached and it requires more time to complete than flooding

Other application domains of gossiping

- Besides information dissemination…
- Peer sampling
  - How to provide every node with a list of peers to exchange information with
- Resource management, including monitoring, in large-scale distributed systems
  - E.g., failure detection
- Distributed computations to *aggregate* data in very large distributed systems (e.g., sensor networks)
  - Computation of aggregates e.g., sum, average, maximum and minimum values
  - E.g., to compute average value
    - Let $v_{0,i}$ and $v_{0,j}$ be the values at time t=0 stored by nodes i and j
    - Upon gossip, i and j exchange their local value v_i and v_j and adjust it to

      $v_{1,i}, v_{1,j} \leftarrow (v_{0,i} + v_{0,j})/2$

Two gossiping protocols

- Let's consider two examples of gossiping protocols
  - **Blind counter rumor mongering**
  - **Bimodal multicast**

Blind counter rumor mongering

- Why such name for this protocol?
  - *Rumor mongering* (def: "the act of spreading rumors", also known as gossip): a node with "hot rumor" will periodically infect other nodes
  - *Blind*: loses interest regardless of message recipient (*why*)
  - *Counter*: loses interest after some contacts (*when*)

- Two parameters to control gossiping
  - *B*: maximum number of neighbors a message is forwarded to
  - *F*: number of times a node forwards the same message to its neighbors

Portman and Seneviratne, The cost of application-level broadcast in a fully decentralized peer-to-peer network, ISCC 2002

Blind counter rumor mongering

- Gossip protocol

A node *n* initiates a broadcast by sending message *m* to *B* of its neighbors, chosen at random

When node *p* receives a message *m* from node *q*

If *p* has received *m* no more than *F* times

p sends *m* to *B* uniformly randomly chosen neighbors that *p* knows have not yet seen *m*

– Note that *p* knows if its neighbor *r* has already seen the message *m* only if *p* has sent it to *r* previously, or if *p* received the message from *r*

Blind counter rumor mongering: performance

- Difficult to obtain analytical expressions to analyze the behavior of gossiping protocols, even for relatively simple topologies ⇒ simulation analysis
- Assume Barabási network topology (i.e., power-law)
  - 1000 nodes with average node degree equal to 6
  - Cost: measured as number of forwarded messages

	Flooding				Rumor Mongering (F=2, B=2)			
	mean	min	max	std dev	mean	min	max	std dev
Cost	4990	4990	4990	0	2555.42	2471	2638	25.02
Reach	1000	1000	1000	0	918.44	894	945	8.52
Time	8.94	8	10	0.36	21.33	19	30	1.32

  - Rumor mongering vs flooding scalability (F=2, B=2)

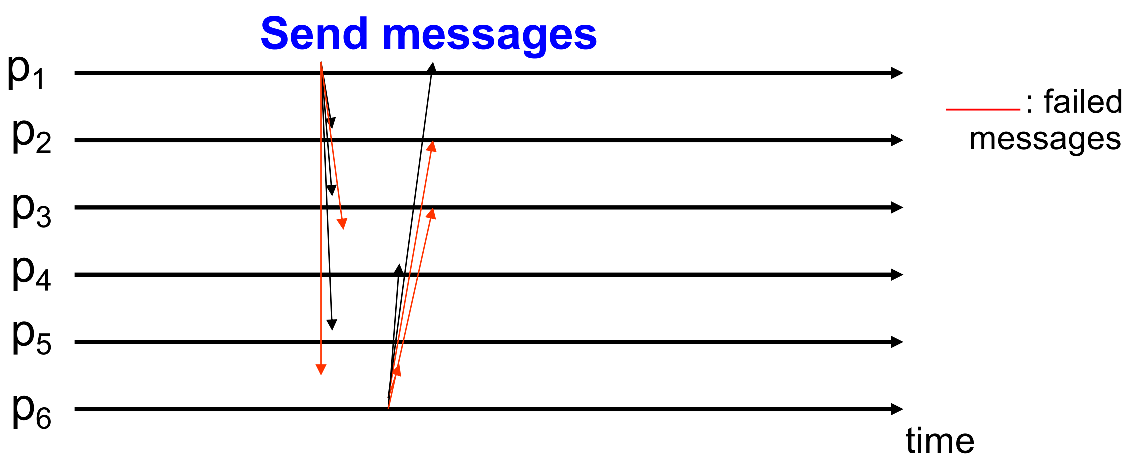
N	cost	reach	time
100	52.60%	93.70%	198%
1000	51.20%	91.10%	240%
10'000	50.00%	92.30%	290%

Bimodal multicast

- Aka pbcast (probabilistic broadcast)
- Composed by two phases:
  1. *Message distribution*: a process sends a multicast message with no particular reliability guarantees
     - Using IP multicast if available, otherwise application-level multicast (e.g., Scribe tree)
  2. *Gossip repair*: after a process receives a message, it begins to gossip about the message to a set of peers
     - Gossip occurs at regular intervals and offers the processes a chance to compare their states and fill any gaps in the message sequence

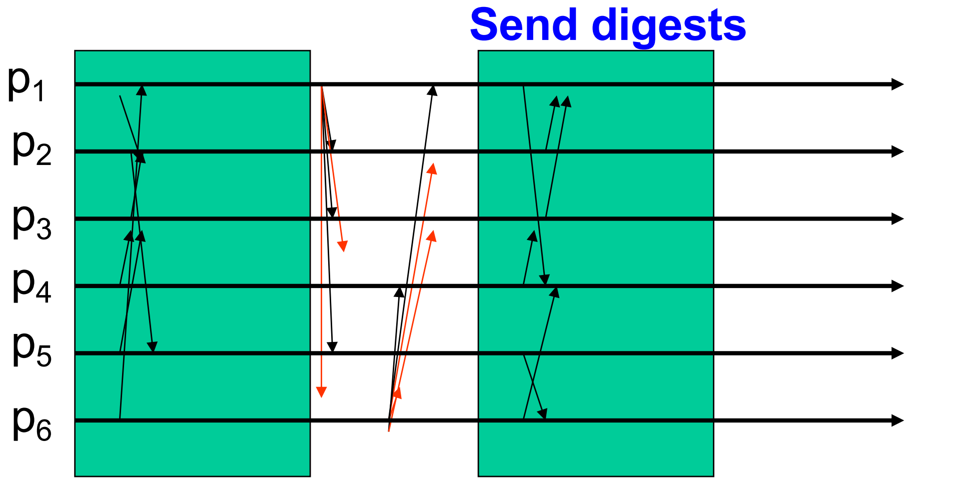
Birman et al., Bimodal multicast, *ACM Trans. Comput. Syst., 1999*

Bimodal multicast: message distribution



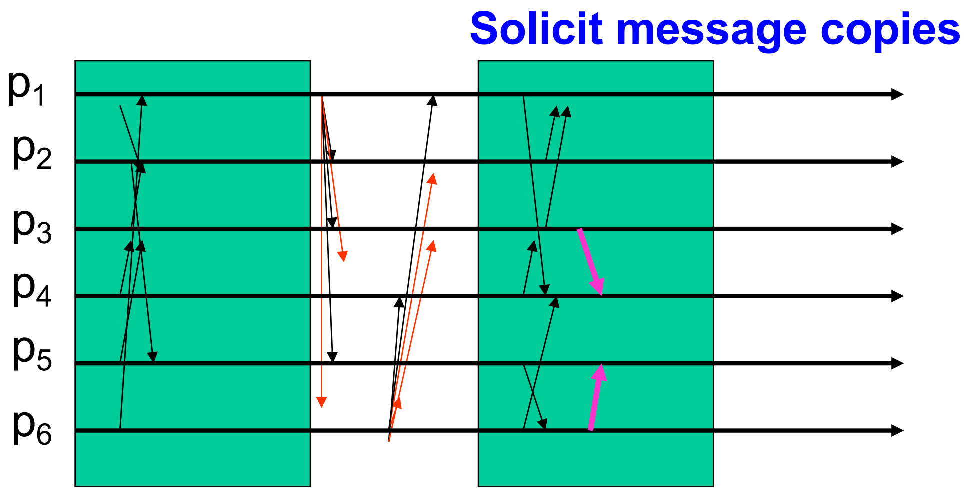
- Start by using *unreliable* multicast to rapidly distribute messages
- Partial distribution of multicast messages may occur
  - Some message may not get through
  - Some process may be faulty

Bimodal multicast: gossip repair



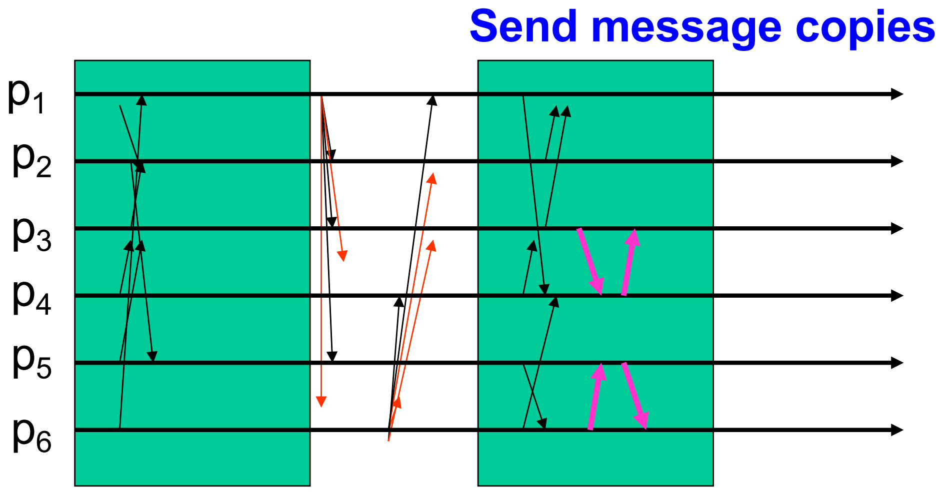
- Periodically (e.g., every 100 ms) each process sends a *digest* describing its state to *some randomly* selected process
- Digest only identifies messages, without including them

Bimodal multicast: gossip repair



- Recipient checks gossip digest against its own history and *solicits* a copy of any missing message from the process that sent the gossip

Bimodal multicast: gossip repair



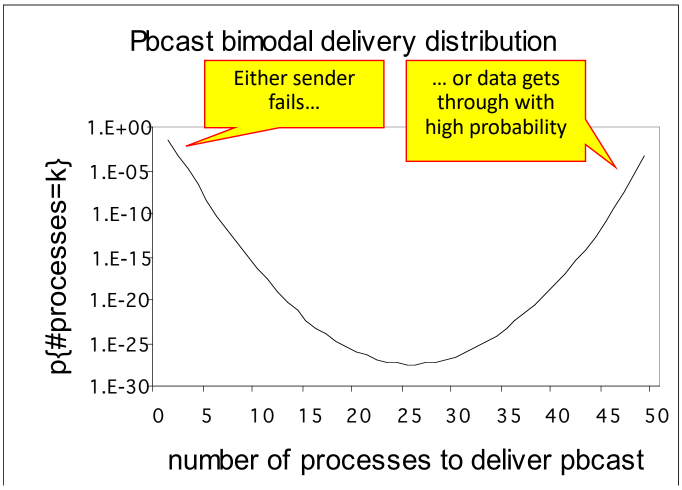
- Processes reply to solicitations received during a gossip round by retransmitting the requested message
- Some optimizations (not examined)

Bimodal multicast: why "bimodal"?

- Are there two phases?
- Nope; description of dual "modes" of result

1. pbcast is almost always delivered to most or to few processes and almost never to some processes

   Atomicity = almost all or almost none

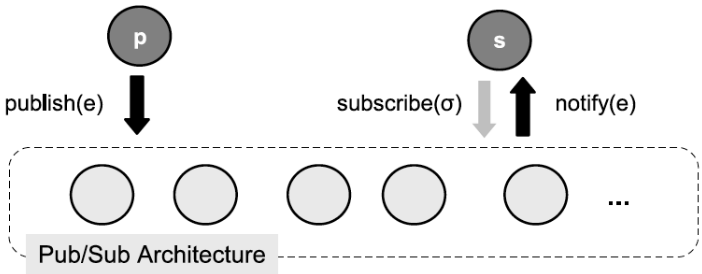


2. A second bimodal characteristic is due to delivery latencies, with one distribution of very low latencies (messages that arrive without loss in the first phase) and a second distribution with higher latencies (messages that had to be repaired in the second phase)

Distributed event matching

- **Event matching** (aka notification filtering): core functionality of pub/sub systems

- Event matching requires:
  - Checking subscriptions against events
  - Notifying subscribers in case of match



Distributed event matching: centralized architecture

- First solution: centralized architecture
- Centralized server handles all subscriptions and notifications
- Centralized server:
  - Handles subscriptions from subscribers
  - Receives events from publishers
  - Checks every and each event against subscriptions
  - Notifies matching subscribers

✓ Simple to realize and feasible for small-scale deployments

✗ Scalability

✗ SPOF

Distributed event matching: distributed architecture

- How can we address scalability through distribution?
- Simple solution: partitioning
- Master/worker pattern (i.e., hierarchical architecture): master divides work among multiple workers
  - Each worker stores and handles a subset of subscriptions
  - *How* to partition?
    - Simple for topic-based pub/sub: use hashing on topics' names for mapping subscriptions and events to workers
  - ✗ Single master
- Alternatively, avoid single master and use a set of distributed servers (brokers) among which work is spread
  - Organized in a flat architecture, hashing can still be used
  - Example: Kafka

Distributed event matching: distributed architecture

- Other alternatives: decentralized solutions based on P2P overlay networks
- P2P unstructured overlay: use flooding or gossiping to disseminate information
  - Trivial solution for flooding: propagate each event from publisher to all P2P nodes
  - To reduce message overhead of flooding, use selective event routing
- Alternatively, can also exploit P2P structured overlay
  - Example: Scribe