Storm a Big Data

Marcin Cylke

Marcin Cylke

Marcin Cylke

what is storm

Web Images Maps Applications More Search tools

About 988,000,000 results (0.33 seconds)

storm

/stôrm/

Noun

A violent disturbance of the atmosphere with strong winds and usually rain, thunder, lightning, or snow.

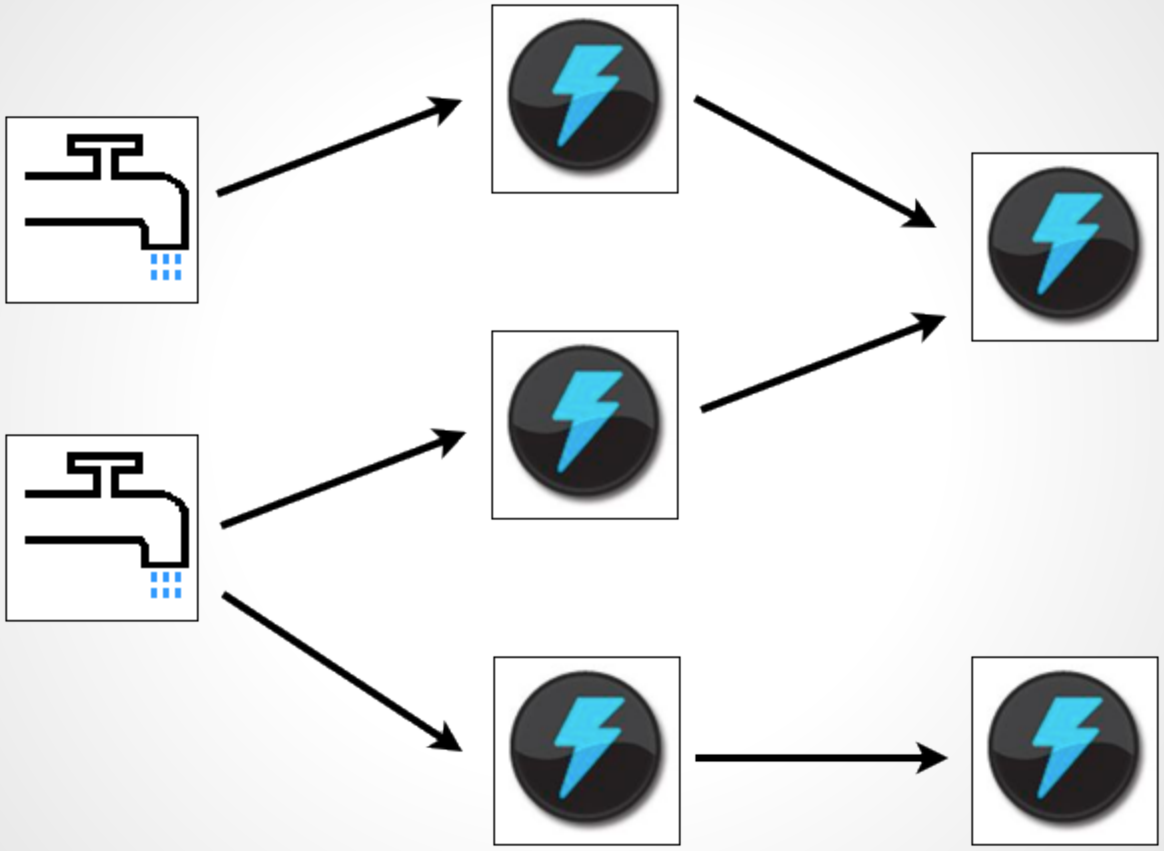
Verb

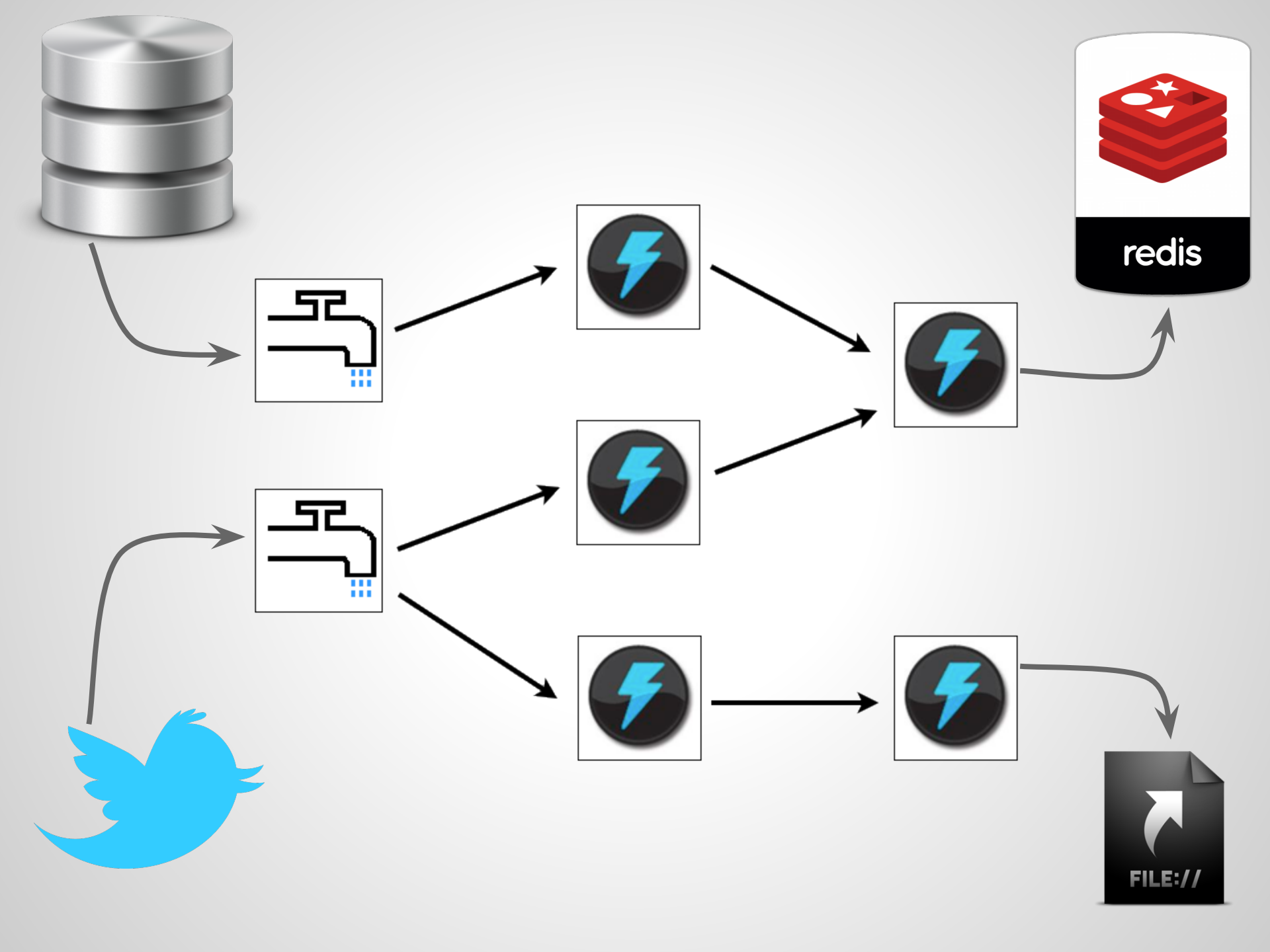
Move angrily or forcefully in a specified direction: "she stormed off".

Synonyms

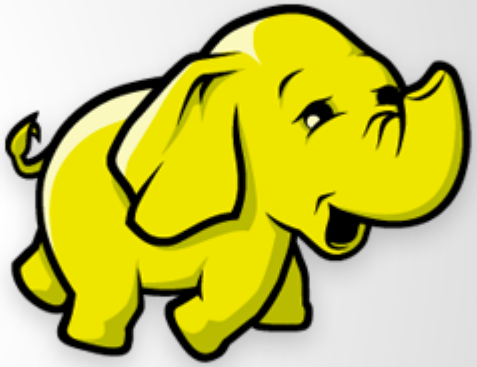
noun. tempest - gale - hurricane - squall - thunderstorm
verb. rage - assault - rave - bluster

More info - Wikipedia - Dictionary.com - Answers.com - Merriam-Webster




redis
FILE://

VS

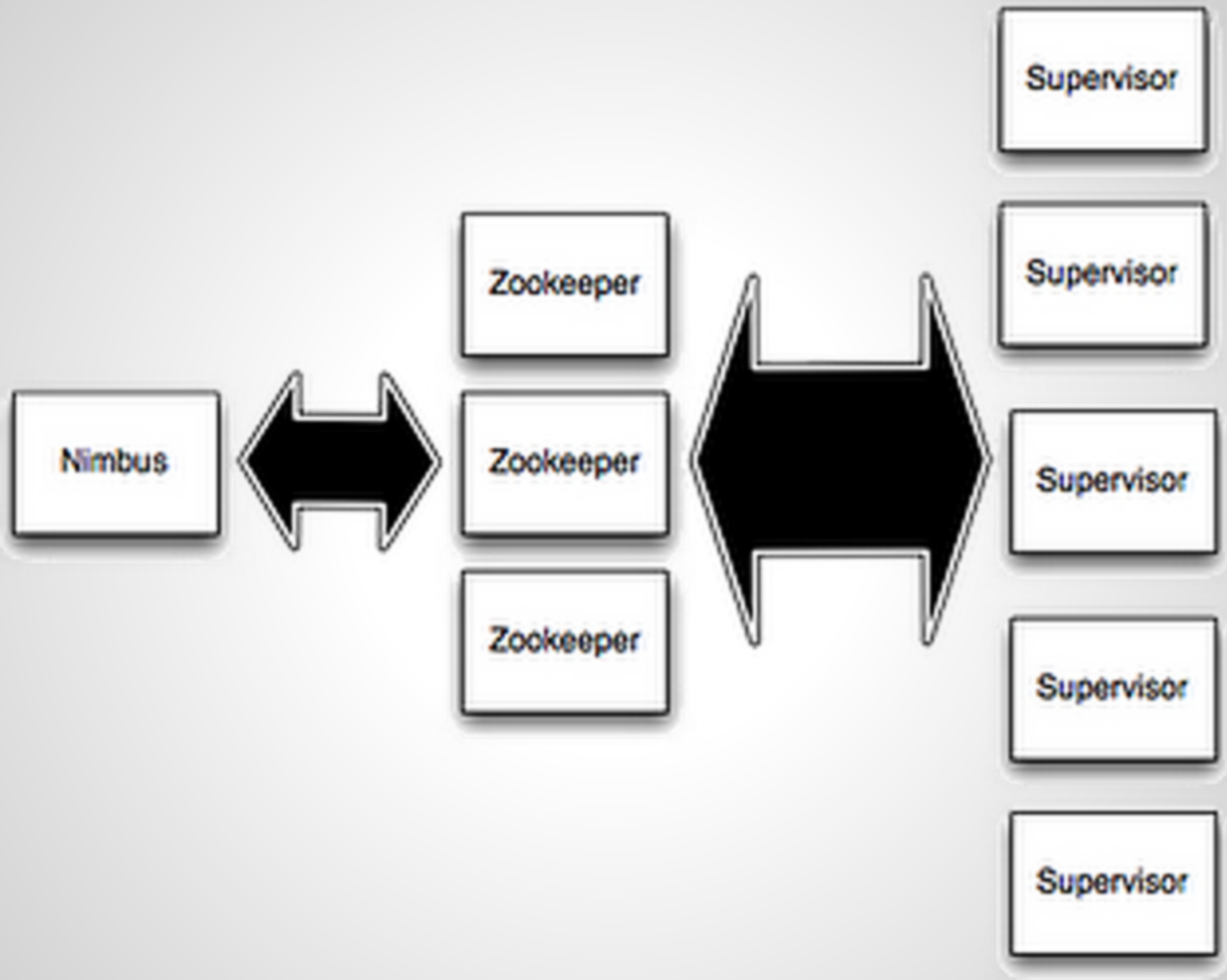


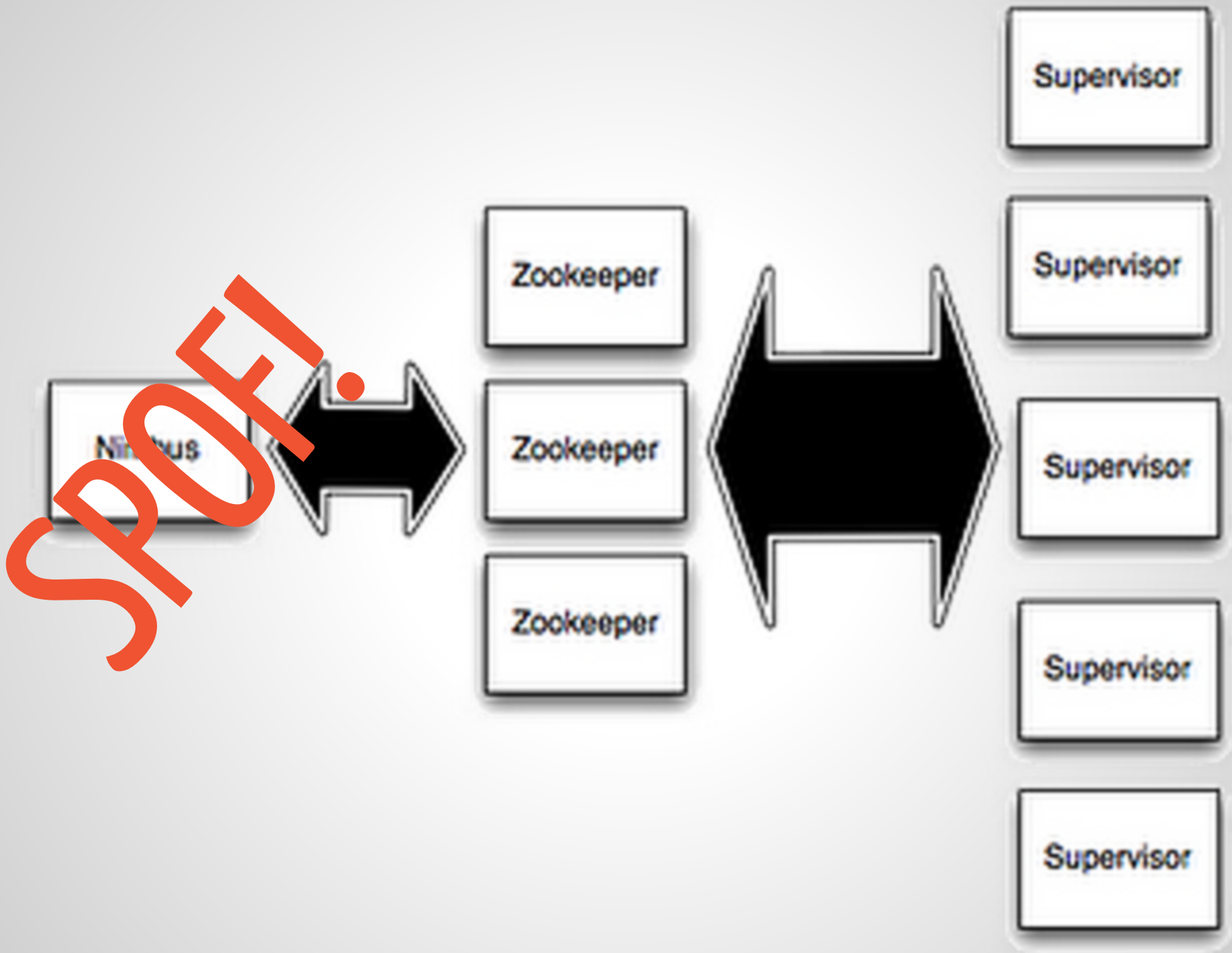
- guaranteed message delivery

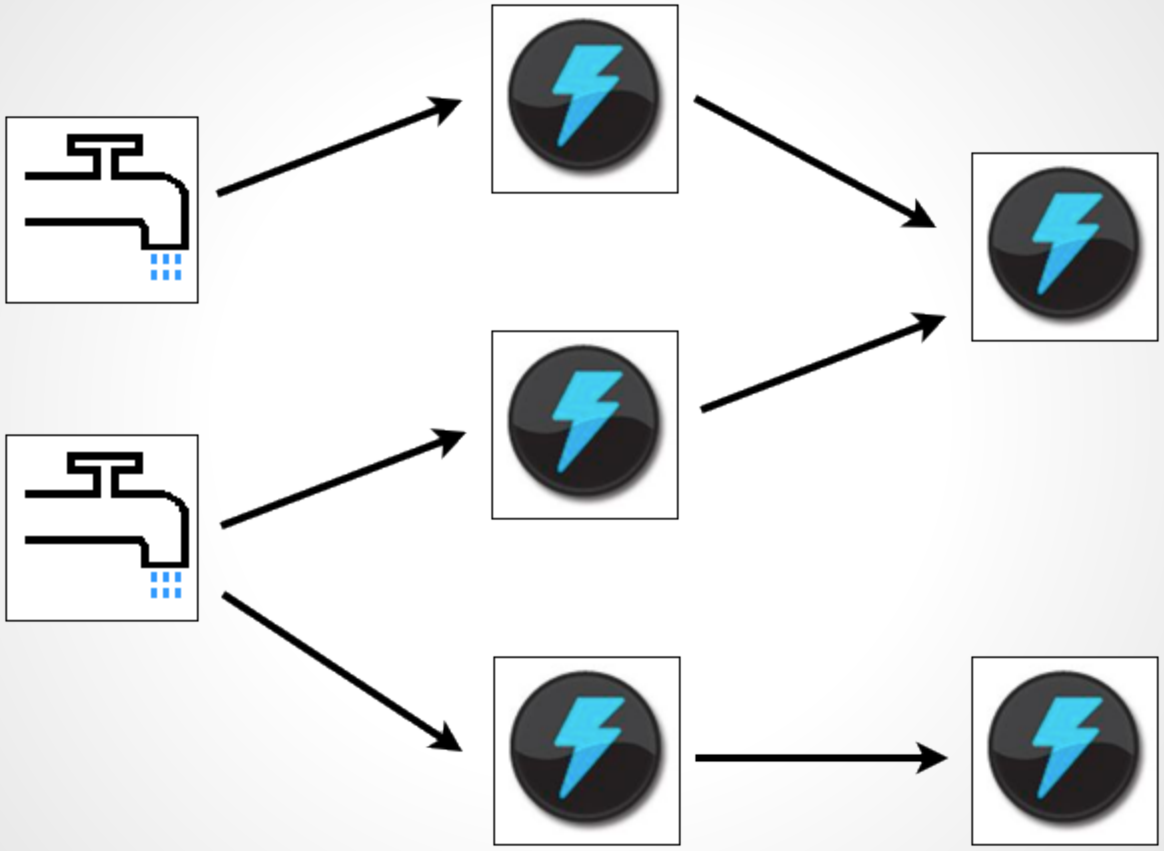
- guaranteed message delivery
- transactional

- guaranteed message delivery
- transactional
- ease of use

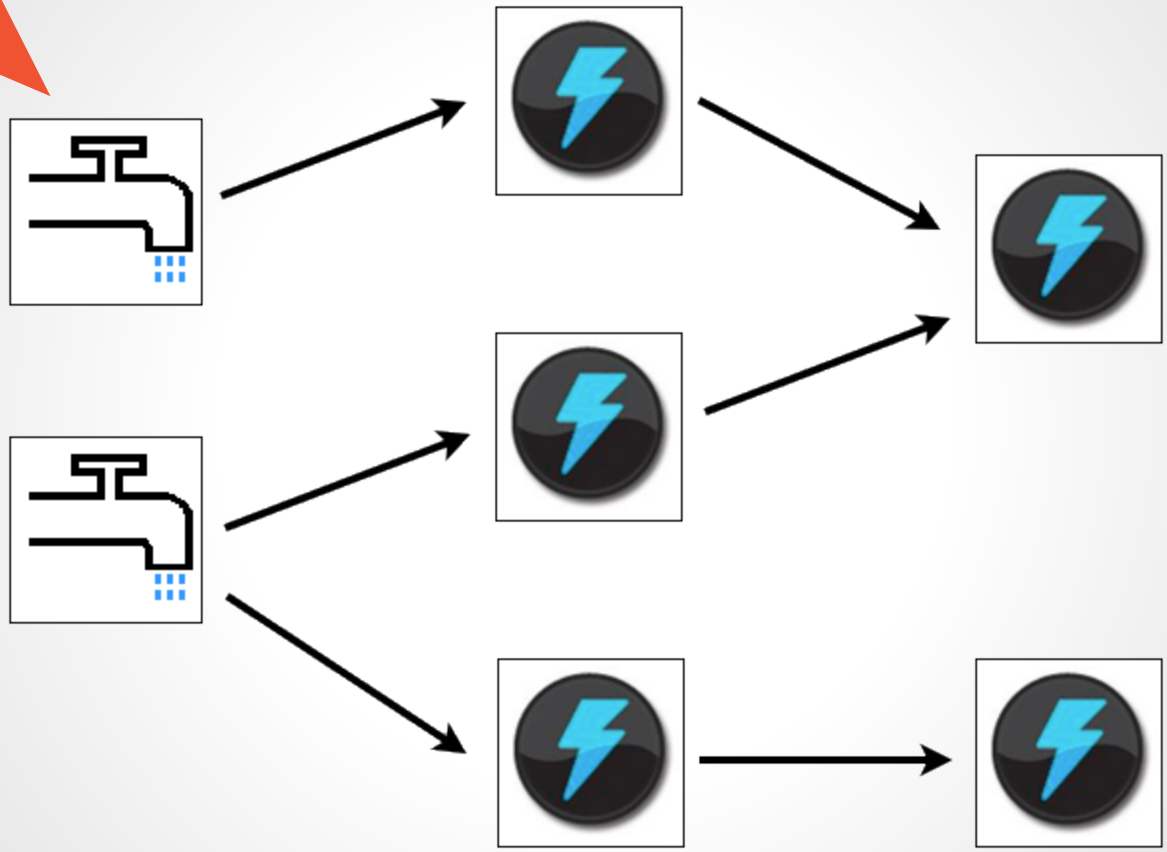
Architecture


Supervisor
Supervisor
Zookeeper
Nimbus
Zookeeper
Supervisor
Zookeeper
Supervisor
Supervisor

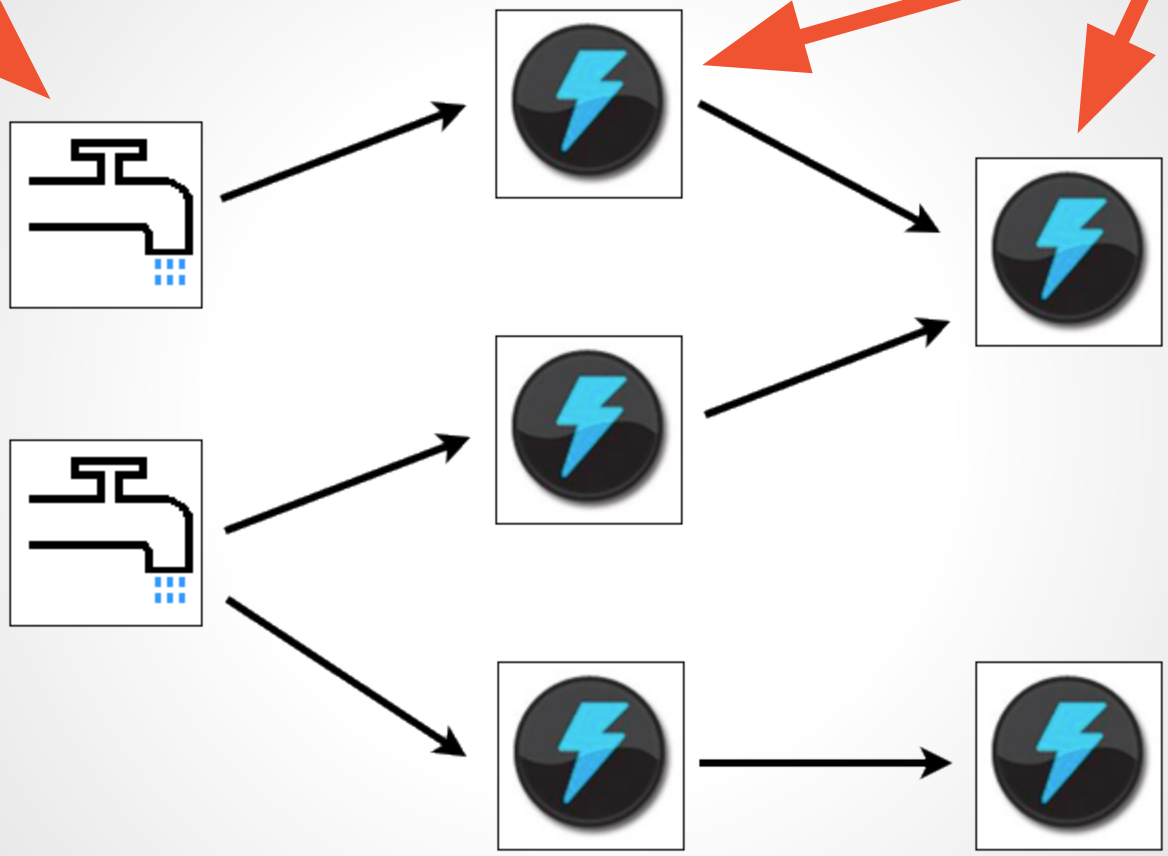

Supervisor
Supervisor
Zookeeper
Zookeeper
Supervisor
Supervisor
Zookeeper
Supervisor
SPOF!

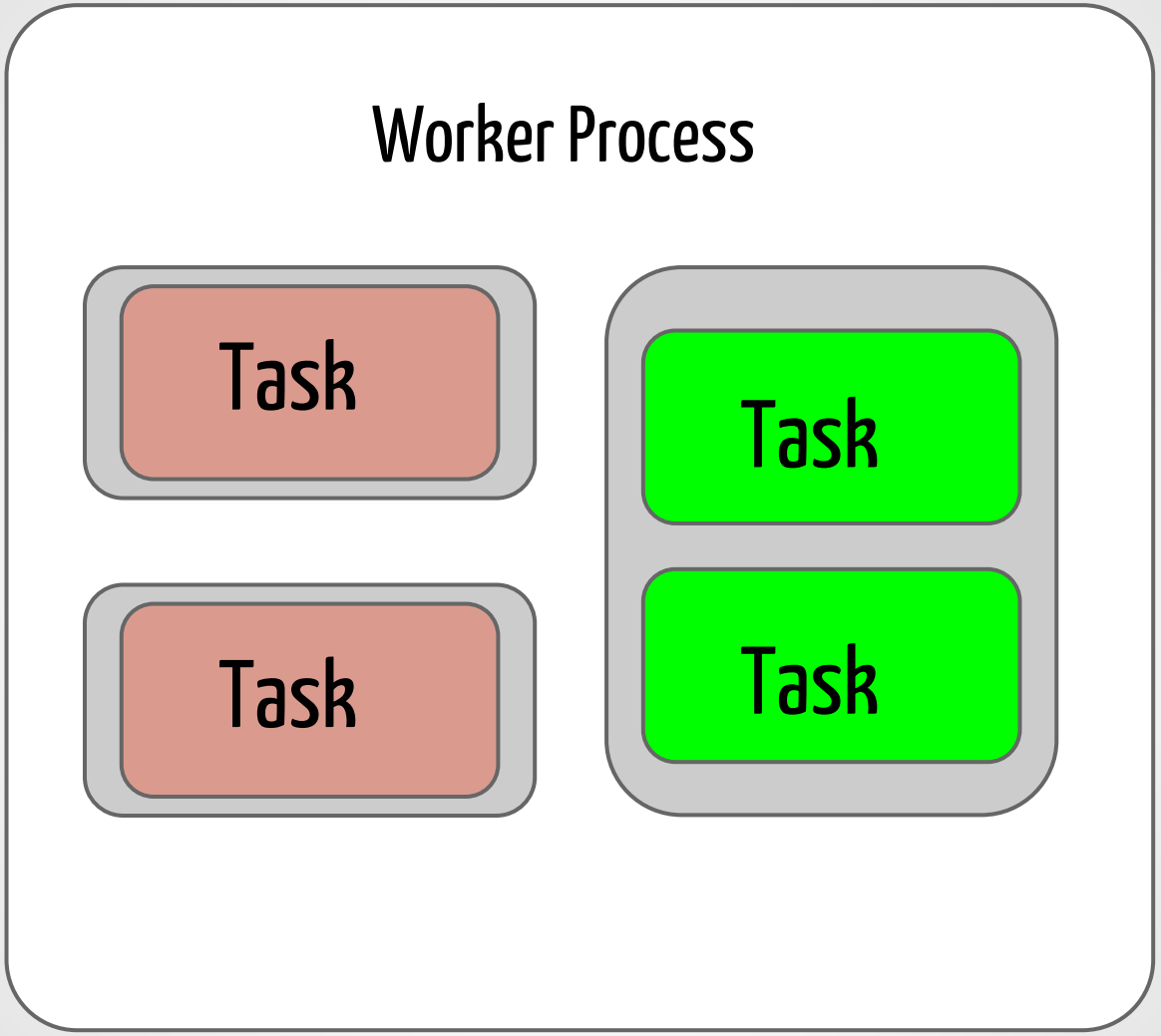


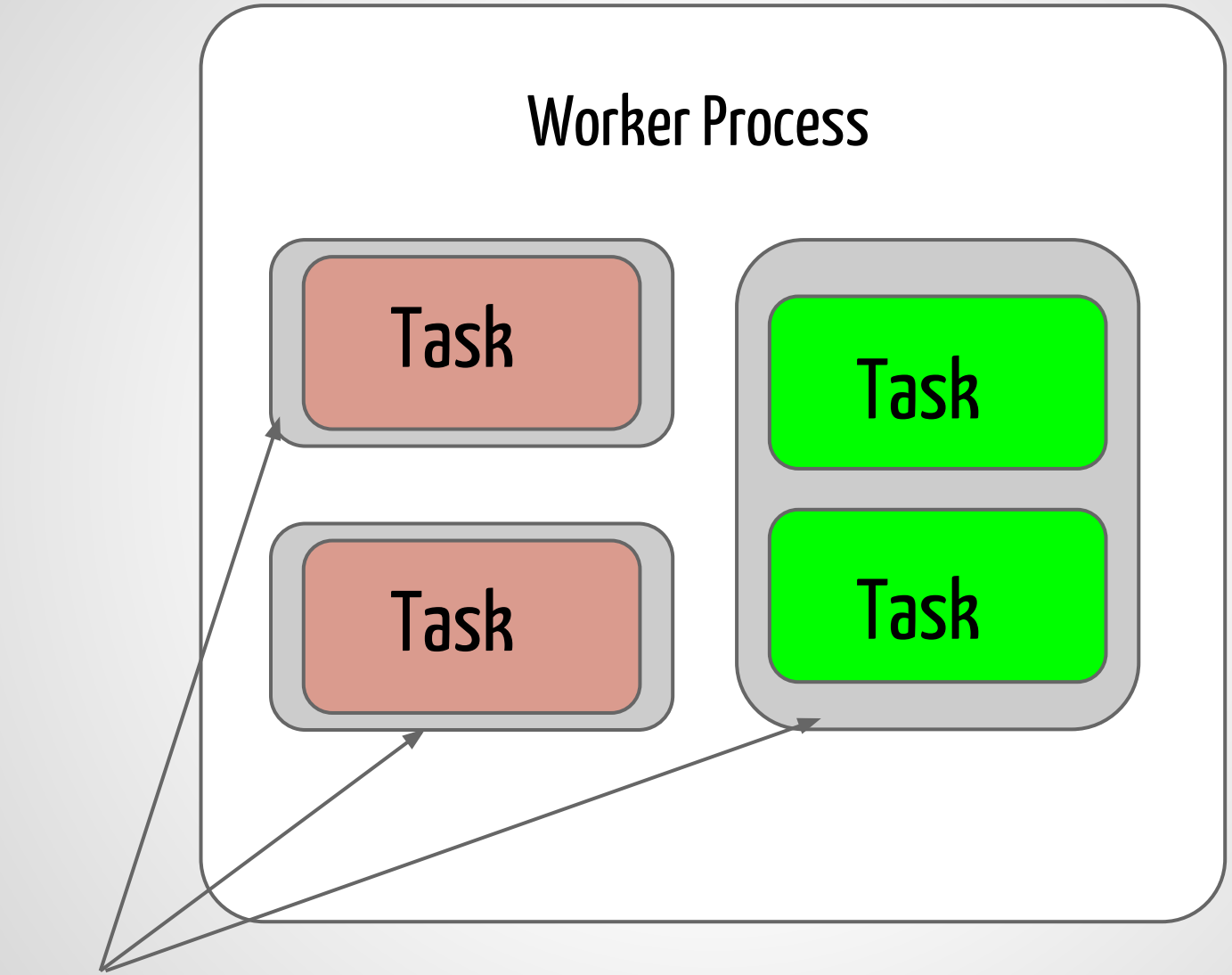
Spout (rynna)



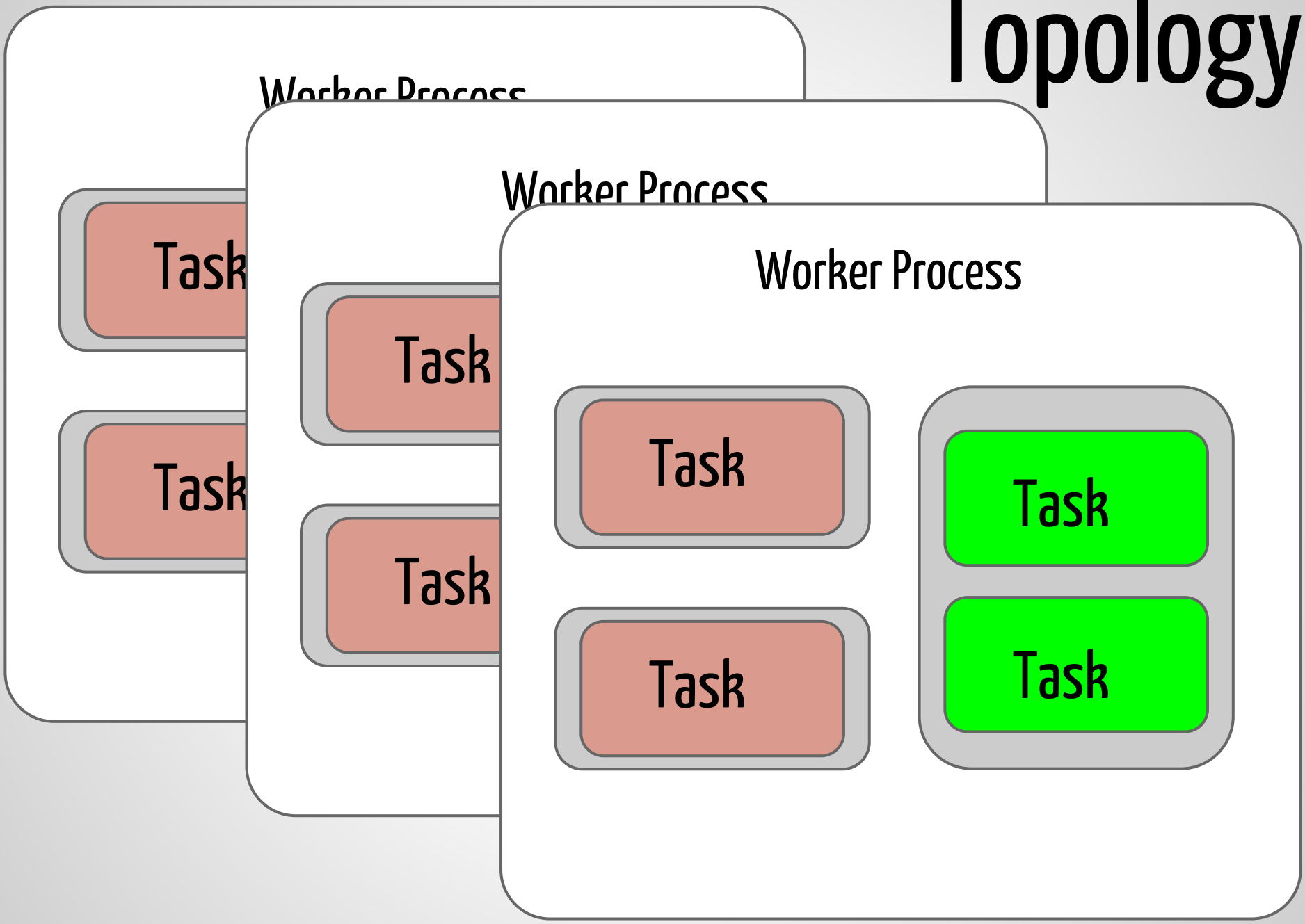
Spout (rynna) Bolt (rygiel)




Worker Process
Task
Task
Task
Task


Worker Process
Task
Task
Task
Task
Executor

Topology

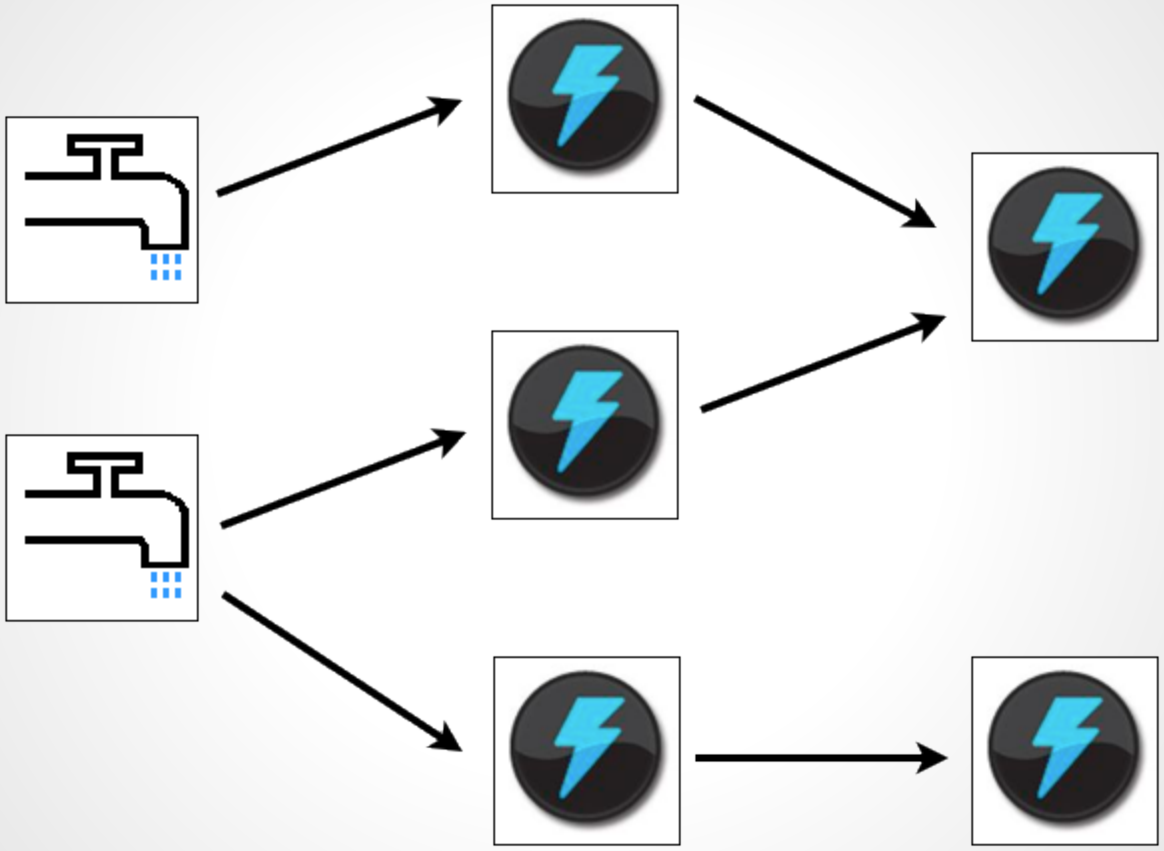

Worker Process
Task
Task
Worker Process
Task
Task
Worker Process
Task
Task
Task
Task

```
builder.setSpout("spout", new InputSpout(), 5);
builder.setBolt("split", new SplitSentence(), 8)
    .shuffleGrouping("spout");
builder.setBolt("count", new WordCount(), 12)
    .fieldsGrouping("split", new Fields
("word"));
```

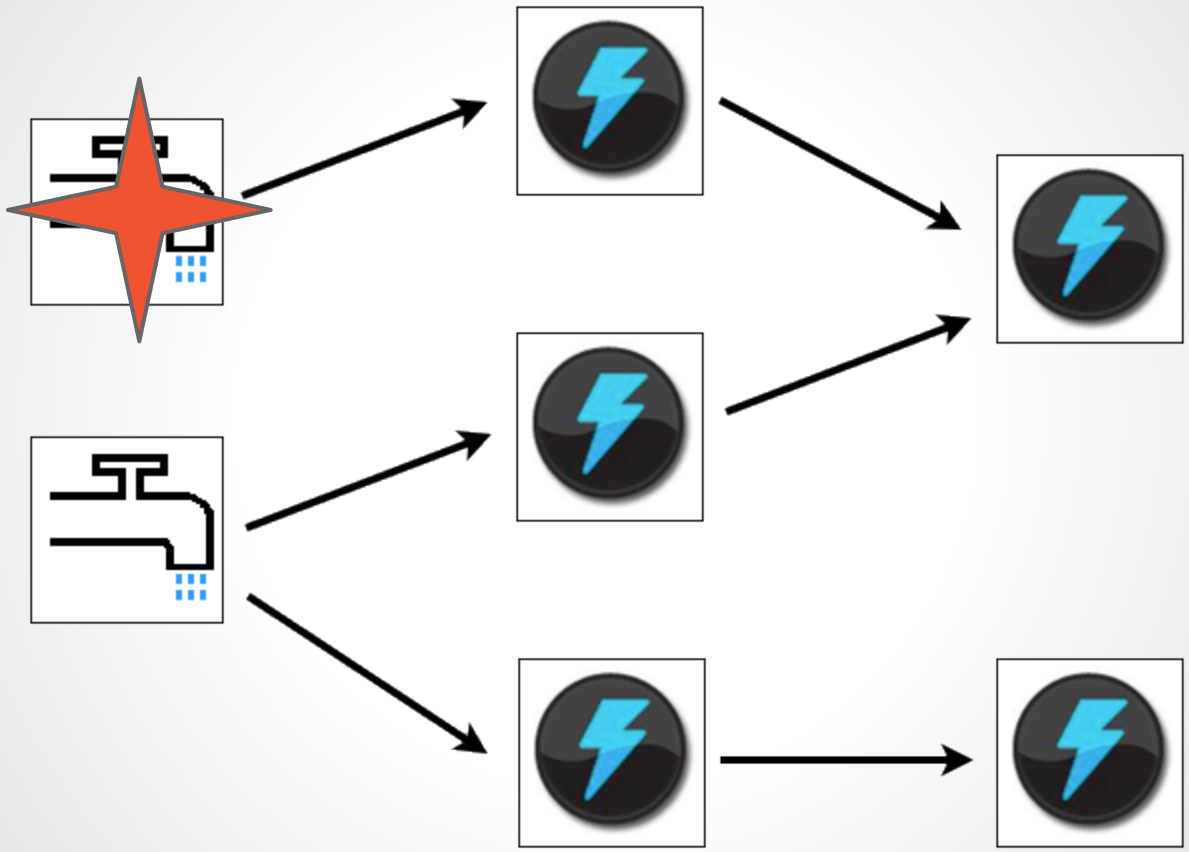
Bolts and Spouts

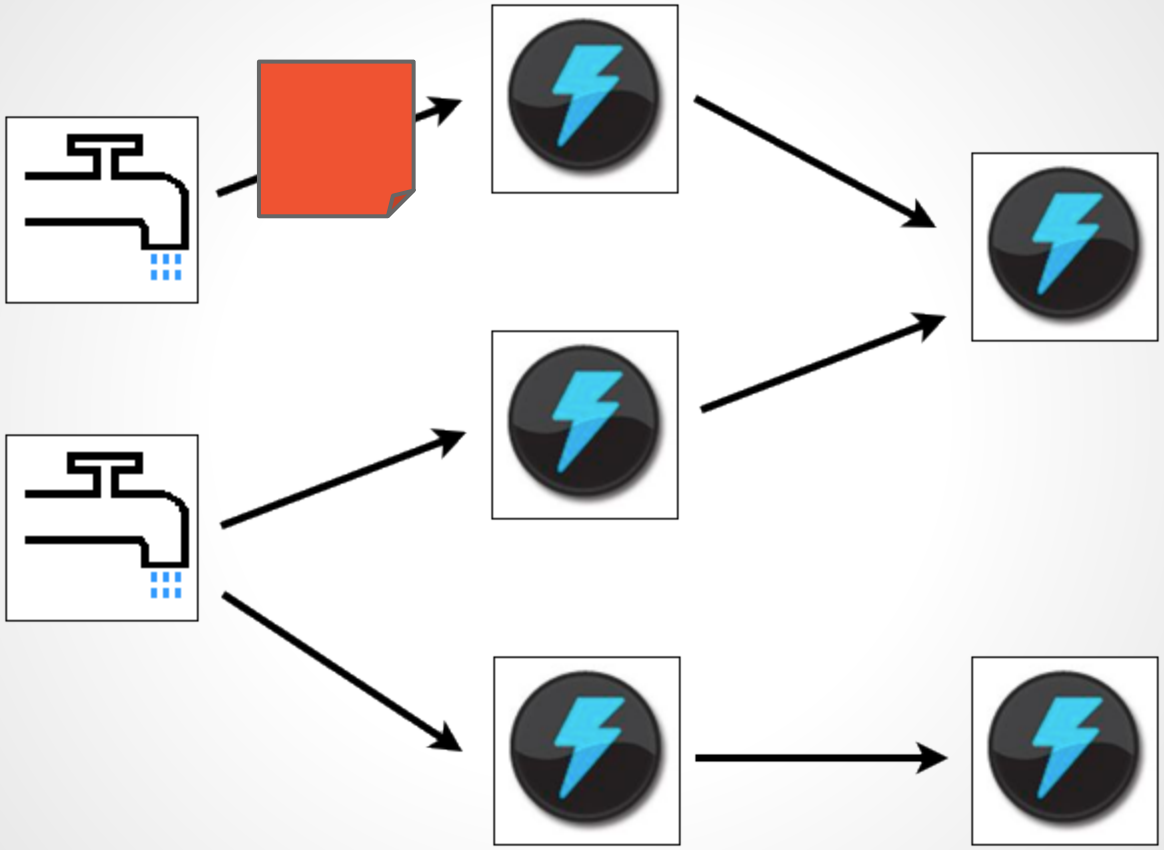
- storm-kafka
- storm-hbase
- storm-rdbms
- storm-jms
- storm-cassandra
- storm-state - write your own hbase!
- storm-ml

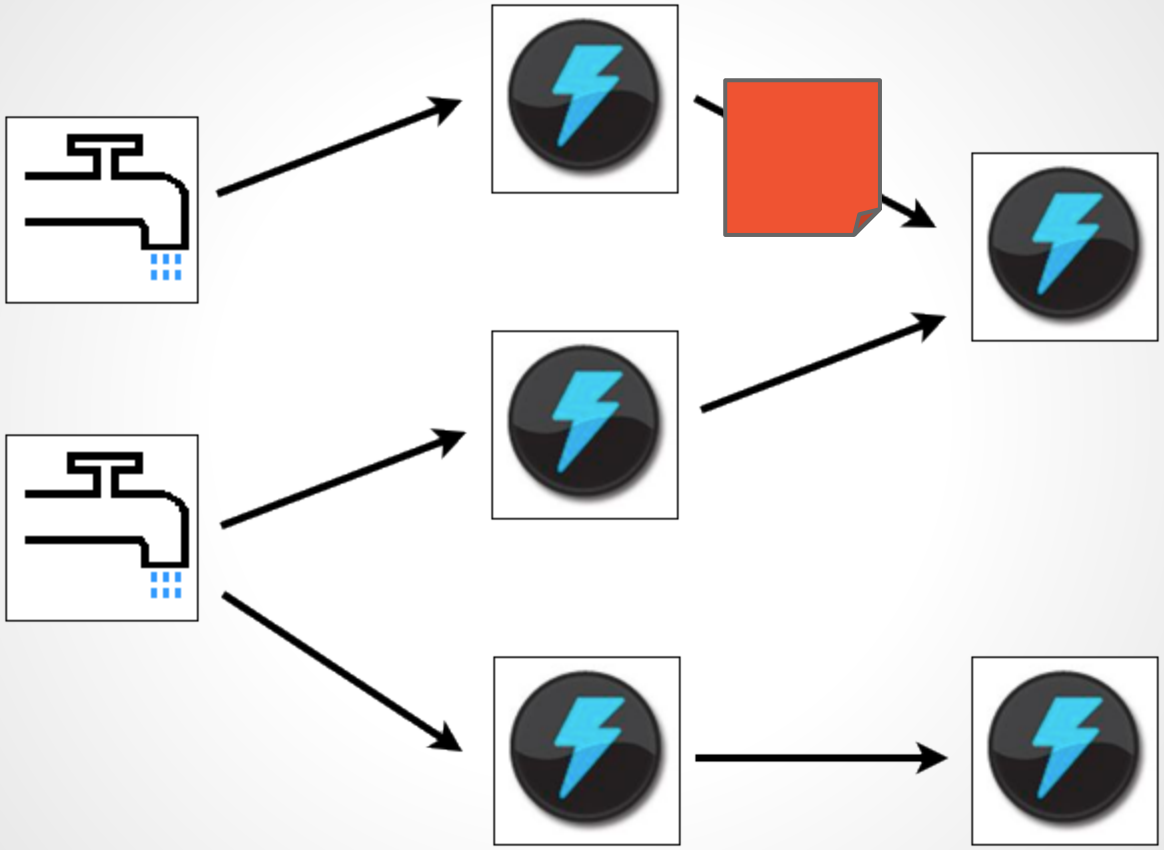
Reliability



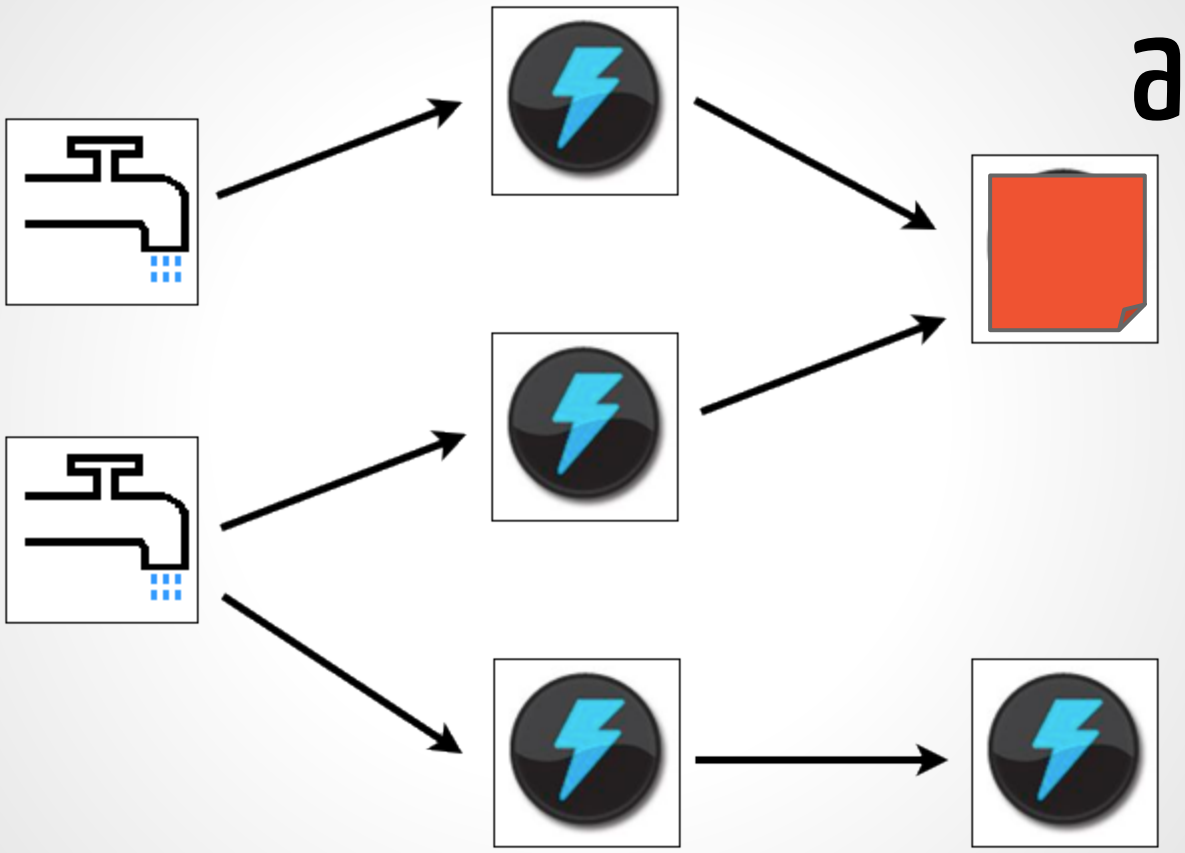
create







ack/fail



1 000 000 x 100 bytes

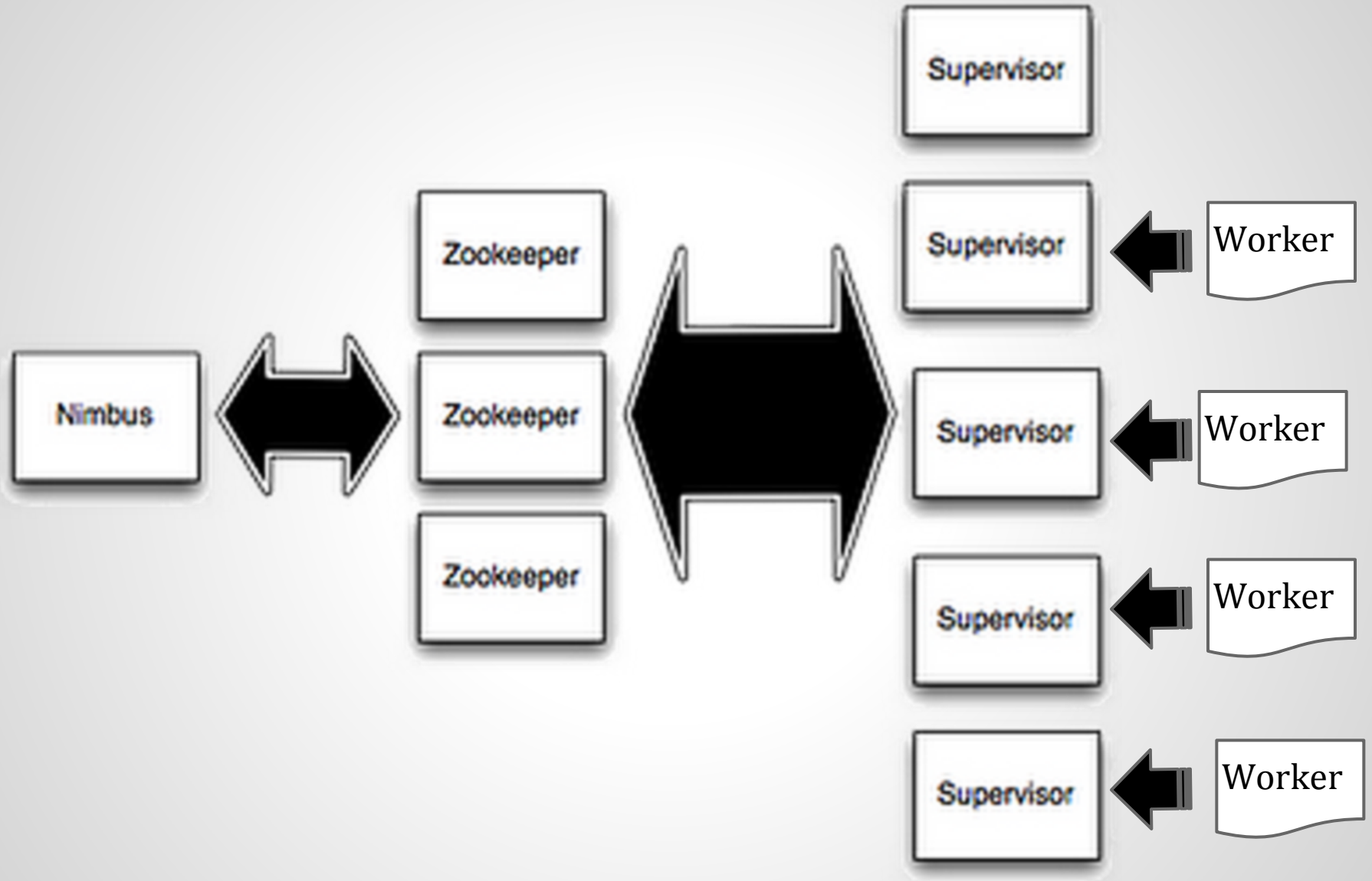
1 node in 1 sec

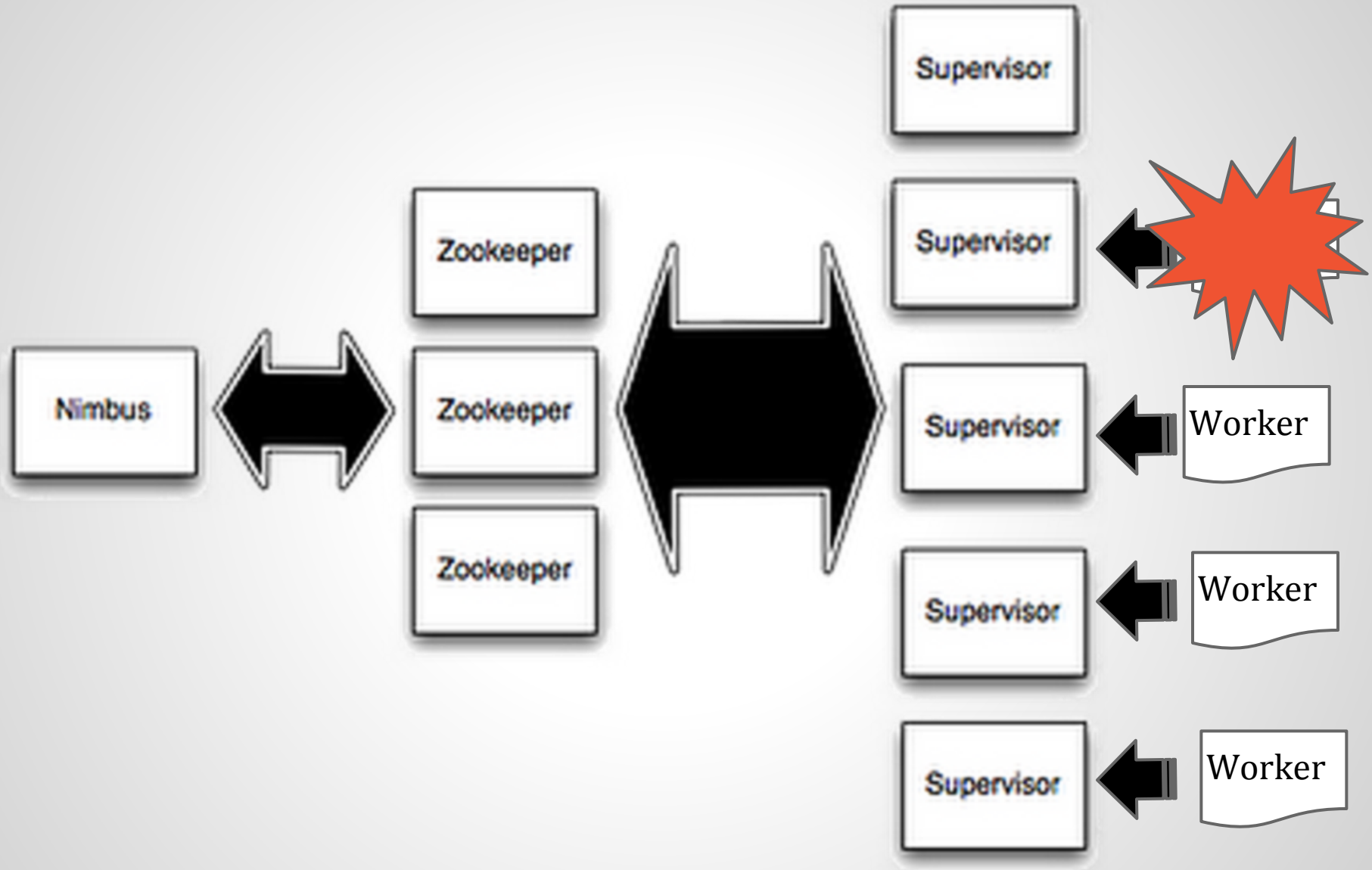
1 000 000 x 100 bytes

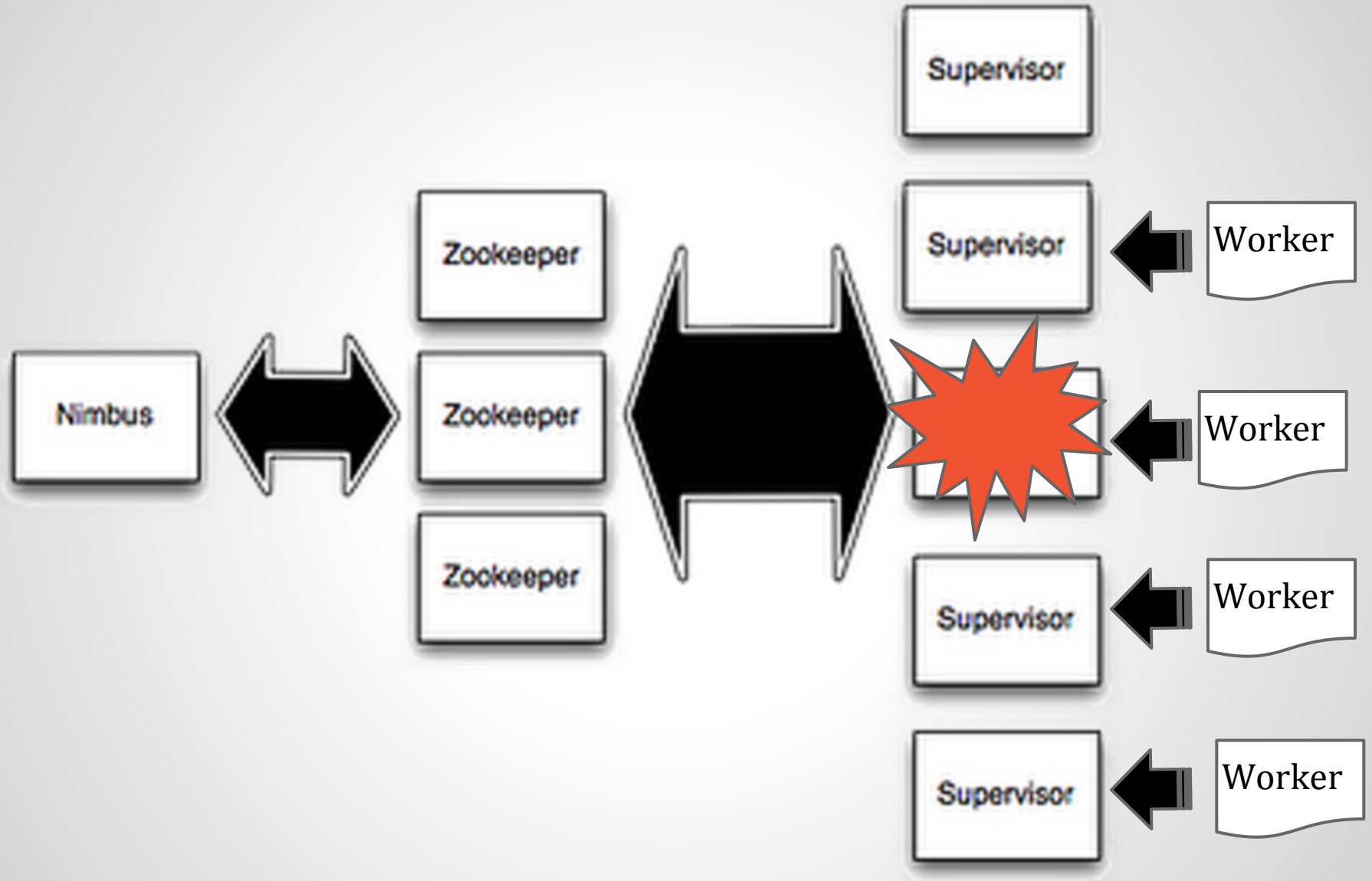
1 node in 1 sec

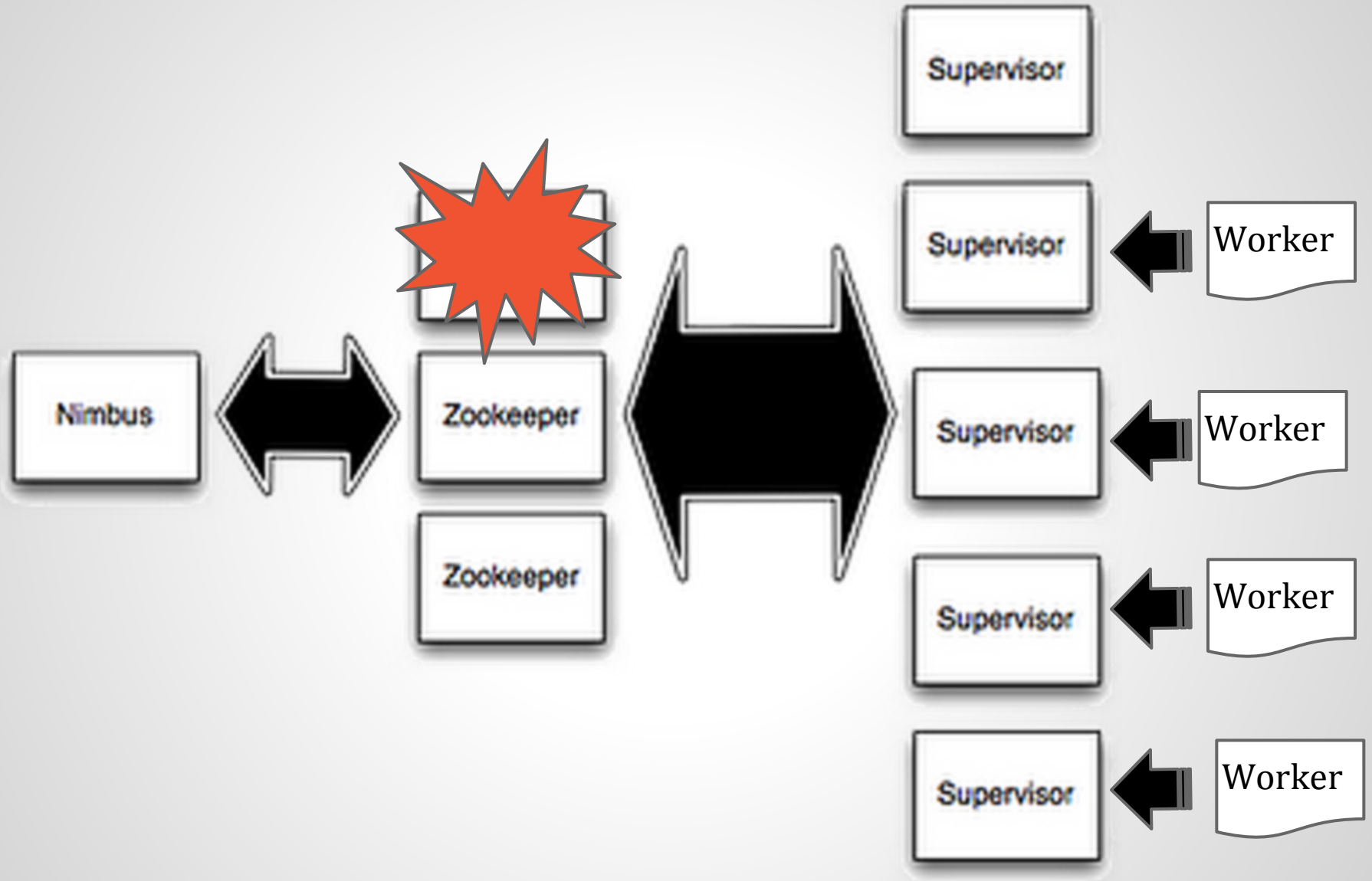
2x Intel E5645@2.4Ghz, 24GB RAM

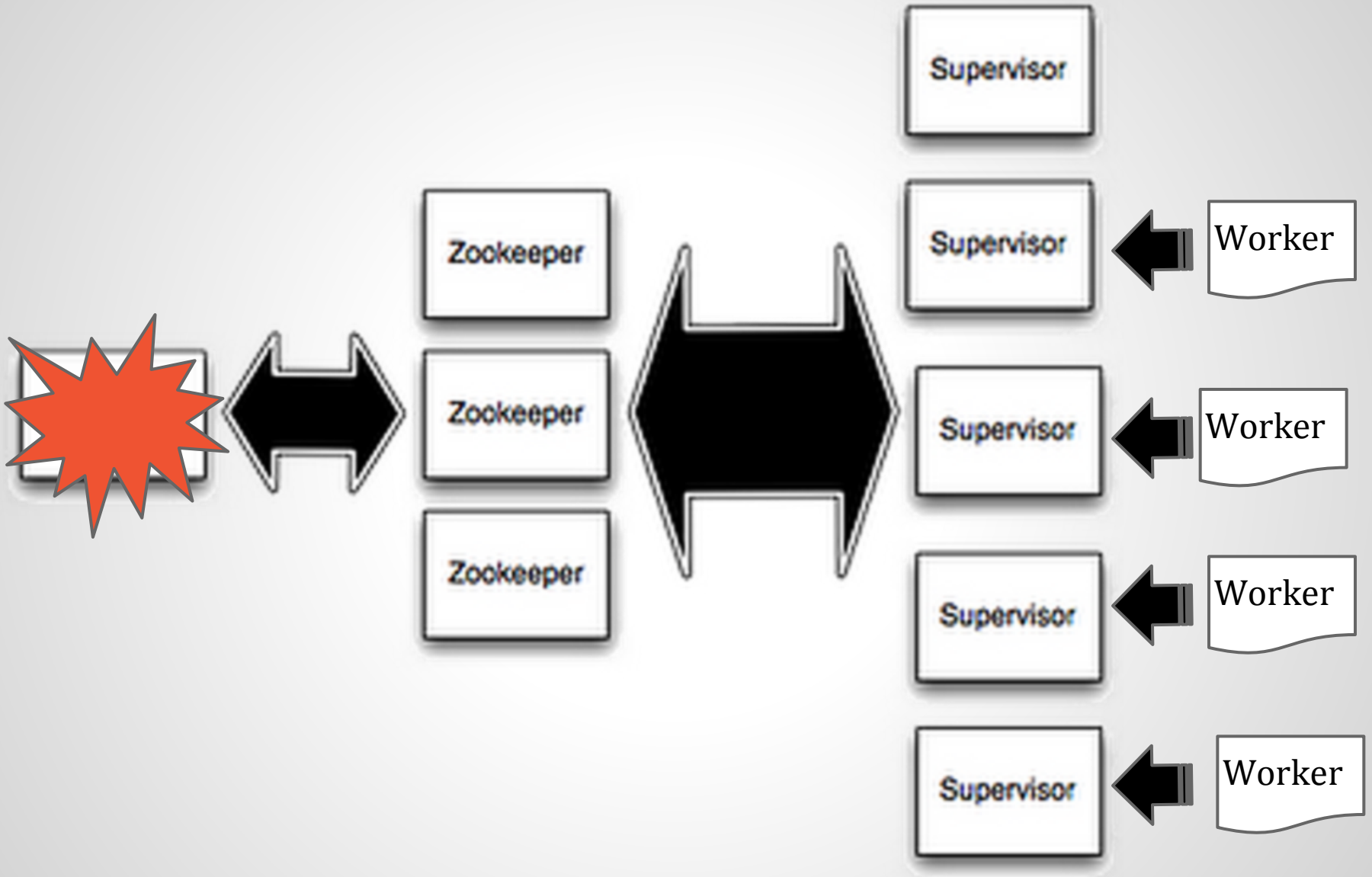

ØMQ
Netty


Supervisor
Zookeeper
Supervisor
Worker
Nimbus
Zookeeper
Supervisor
Worker
Zookeeper
Supervisor
Worker
Supervisor
Worker


Supervisor
Supervisor
Zookeeper
Nimbus
Zookeeper
Supervisor
Worker
Zookeeper
Supervisor
Worker
Supervisor
Worker


Supervisor
Zookeeper
Supervisor
Worker
Nimbus
Zookeeper
Worker
Zookeeper
Supervisor
Worker
Supervisor
Worker


Supervisor
Supervisor
Worker
Nimbus
Zookeeper
Supervisor
Worker
Zookeeper
Supervisor
Worker
Supervisor
Worker


Zookeeper
Zookeeper
Zookeeper
Supervisor
Supervisor
Supervisor
Supervisor
Supervisor
Worker
Worker
Worker
Worker

```
[program:storm-nimbus]
command=/home/mcl/storm-0.9.0/bin/storm nimbus
user=hdfs
autorestart=true
stopsignal=KILL

[program:storm-ui]
command=/home/mcl/storm-0.9.0/bin/storm ui
user=hdfs
autorestart=true
stopsignal=KILL
```

API

```
public interface ISpout {

    void open(Map map, TopologyContext
              topologyContext,
SpoutOutputCollector
              spoutOutputCollector);

    void close();
    void activate();
    void deactivate();
    void nextTuple();
    void ack(Object o);
    void fail(Object o);
}
```

```
public class ExclIntegerShortSpout
                              extends BaseRichSpout {

 public void nextTuple() {
      ....
 }

 public void declareOutputFields
                (OutputFieldsDeclarer declarer) {
      ....
  }
}
```

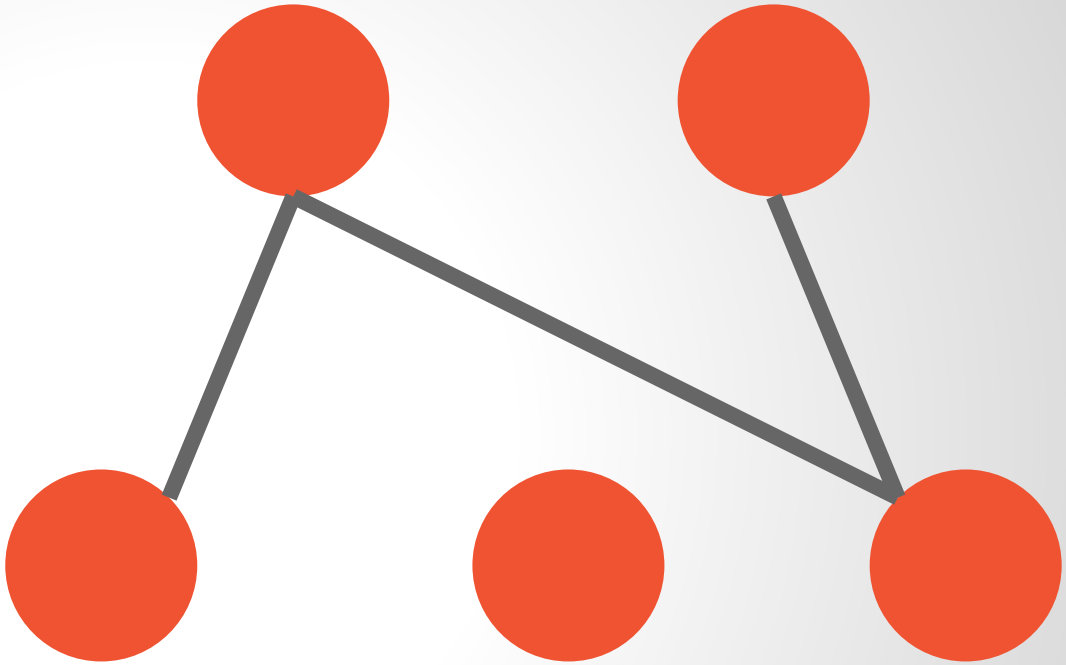
```
public interface IBolt extends {

    void prepare(Map map,
                 TopologyContext topologyContext,
                 OutputCollector
outputCollector);

    void execute(Tuple tuple);

    void cleanup();
}
```

```
public class ExclamationBolt extends BaseRichB

    public void prepare(Map map,
                        TopologyContext topologyContext,
                        OutputCollector outputCollector)
{
        ....
    }

    public void execute(Tuple tuple) {
        ....
    }

    public void declareOutputFields
       (OutputFieldsDeclarer outputFieldsDeclarer)
{
        ....
    }
```

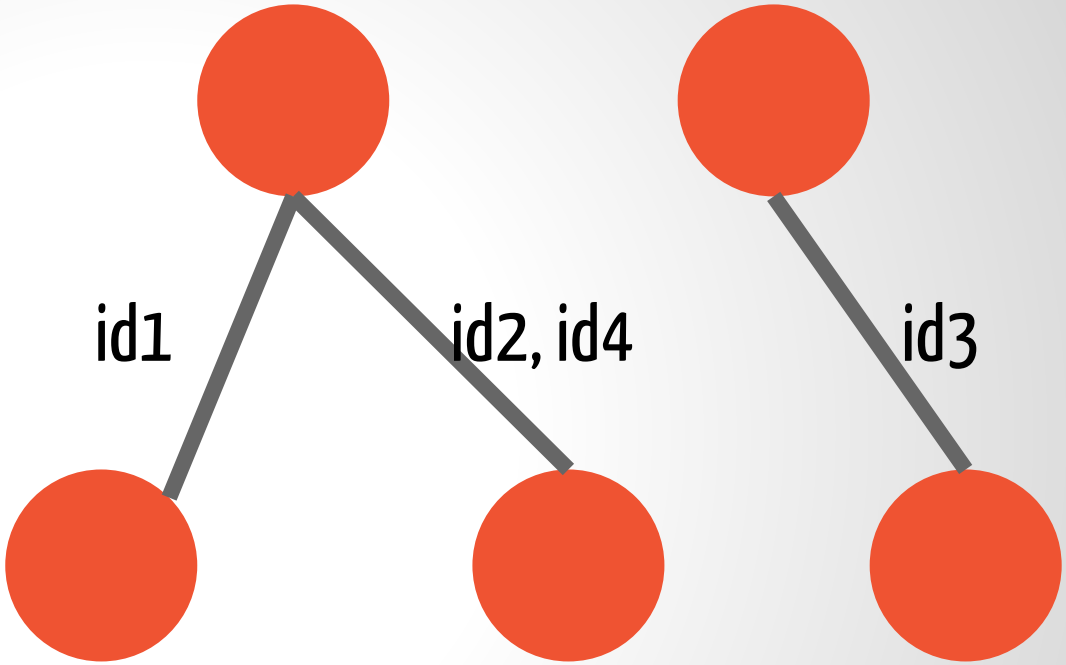
```
builder.setSpout("spout", new InputSpout(), 5);
builder.setBolt("split", new SplitSentence(), 8)
    .shuffleGrouping("spout");
builder.setBolt("count", new WordCount(), 12)
    .fieldsGrouping("split", new Fields
("word"));
```

- joins
- batch processing
- in-memory cache

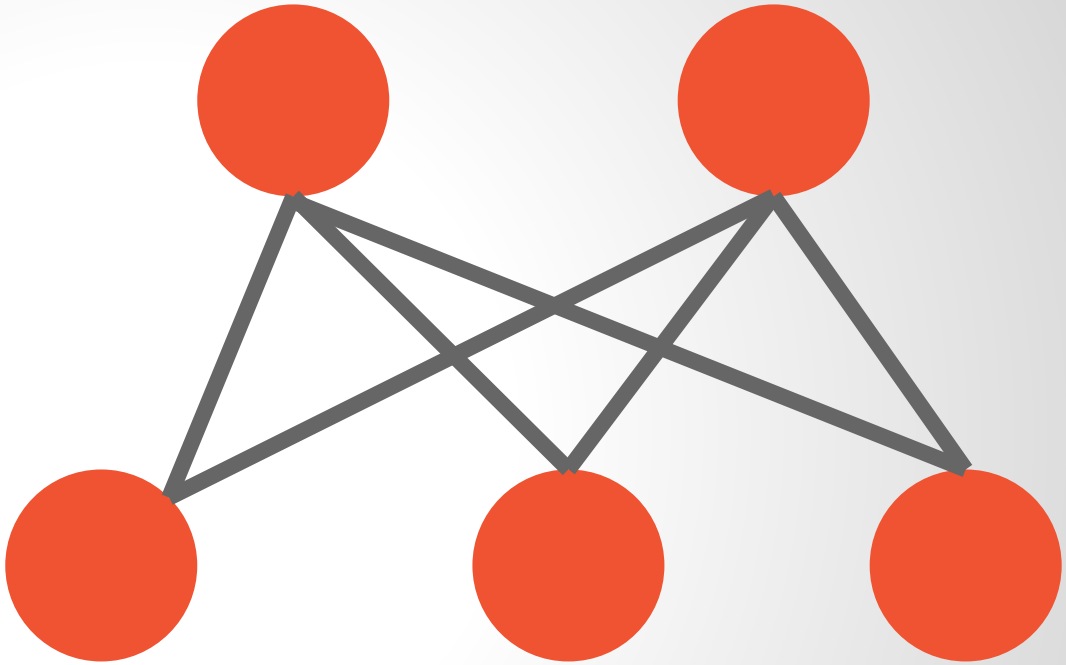
- shuffle



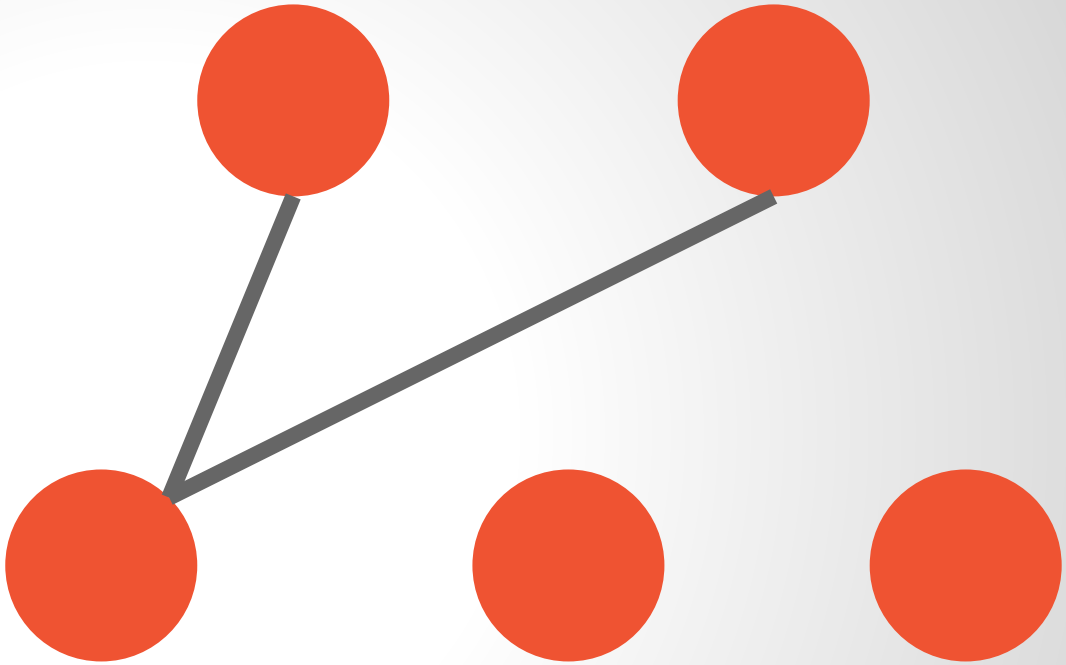
- shuffle
- fields



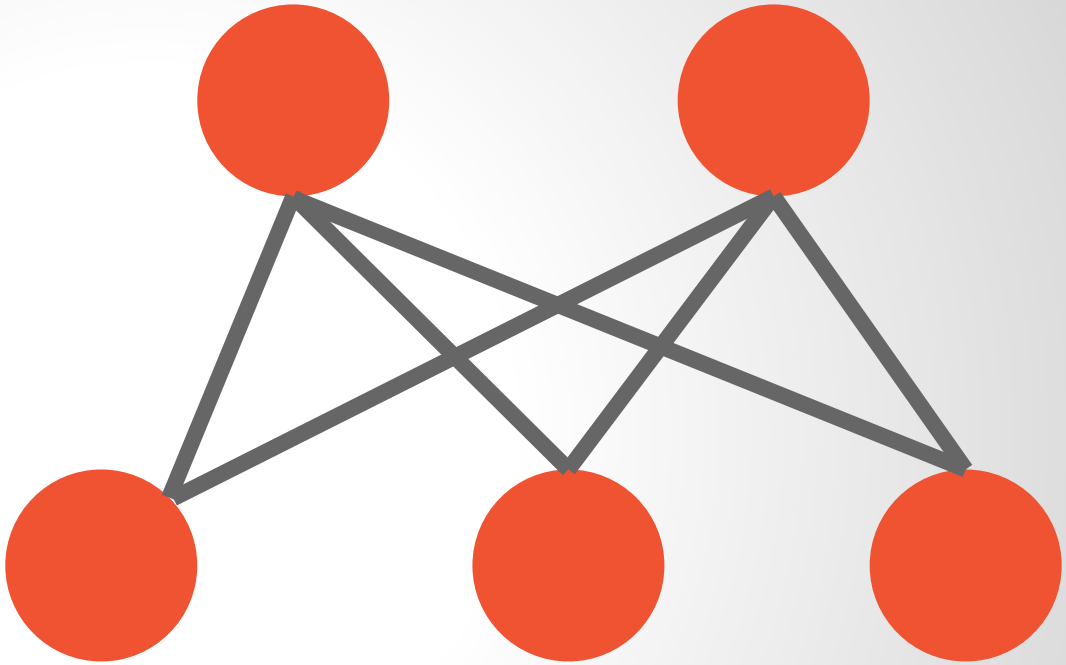
- shuffle
- fields
- all



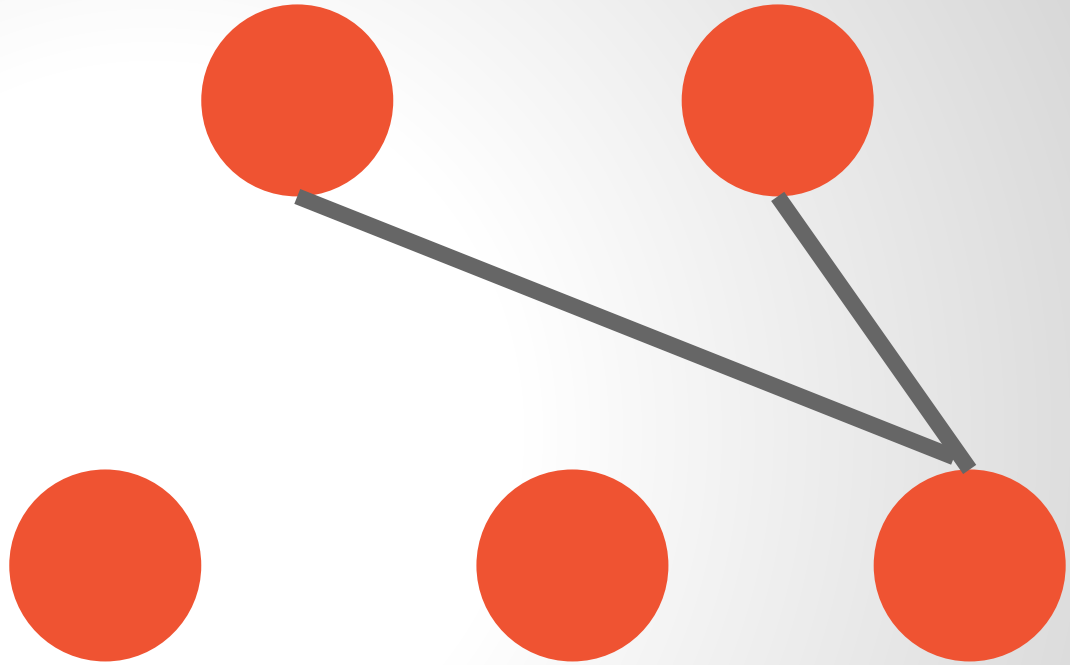
- shuffle
- fields
- all
- global



- shuffle
- fields
- all
- global
- none



- shuffle
- fields
- all
- global
- none
- direct



- Python
- Javascript
- Perl
- PHP

Setup

```
$ storm jar app.jar
          p.w.s.SimpleTopology

$ storm list

$ storm help
```

Problem #1

- dodanie pola w SplittingBolt
- odbieranie go w AckingBolt

Problem #2

- Wordcount

All work and no play makes Jack a dull boy. All work and no play makes Jack a dull boy.

All, work, and, no, play, makes, Jack, a, dull, boy, All, work, and, no, play, makes, Jack, a, dull, boy

All	work	play	makes
All	work	play	makes

All (2), work (2), play(2), makes(2), and (2) ...

File

Split

GroupBy

Join

Result

File

Split

GroupBy

Join

Result

File

Split

GroupBy

Join

Result

Problem #3

- WordCount & Trident

- WordCount & Trident
- PrintFilter

Trident API

- high level batch processing
- fully fault-tolerant, exactly-once processing semantics
- state querying
-

State

- transactional
- non-transactional
- opaque transactional

Blocks

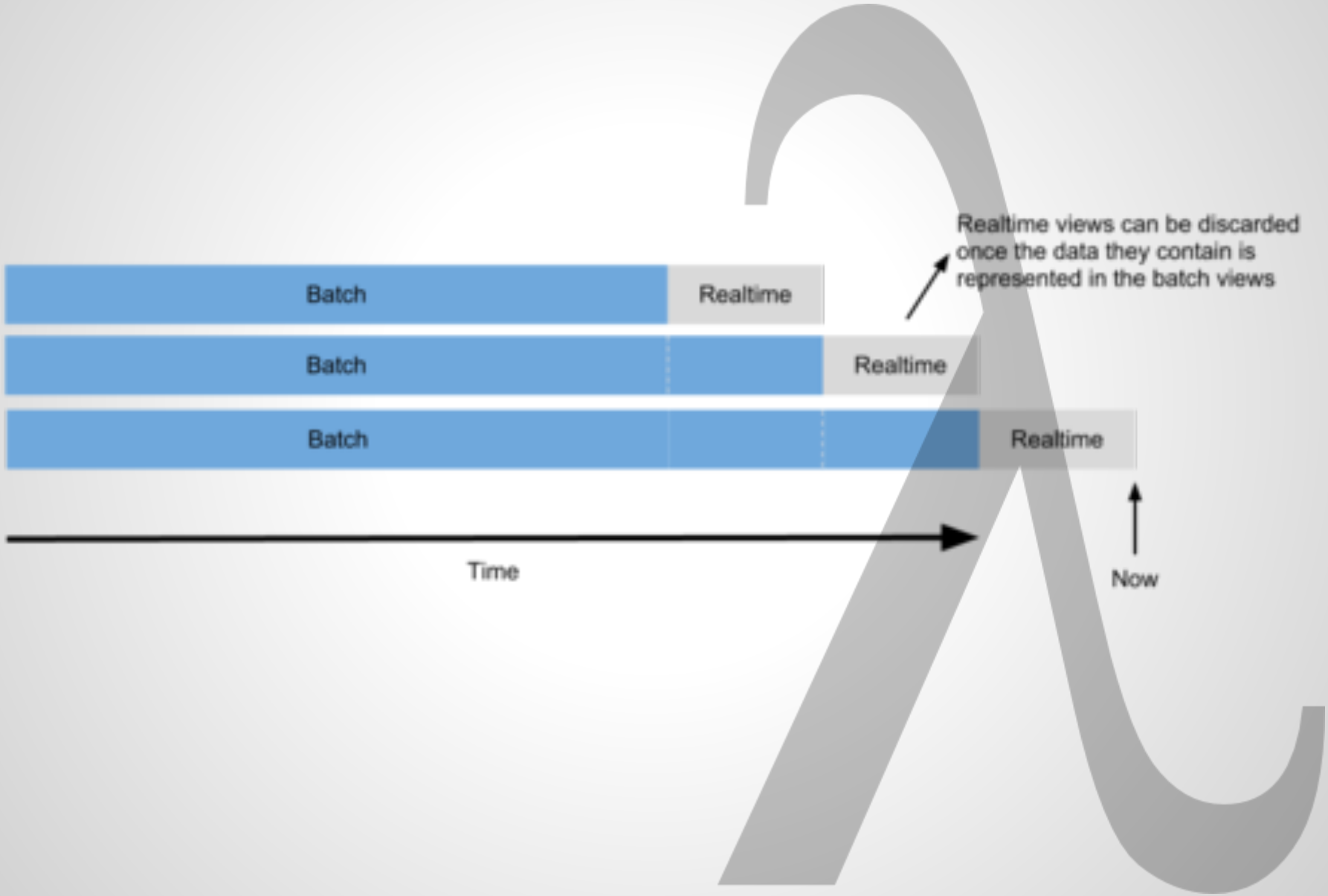
- functions
- filters
- partitionAggregator
- stateQuery and partitionPersist
- projection
- aggregates
- stream merging

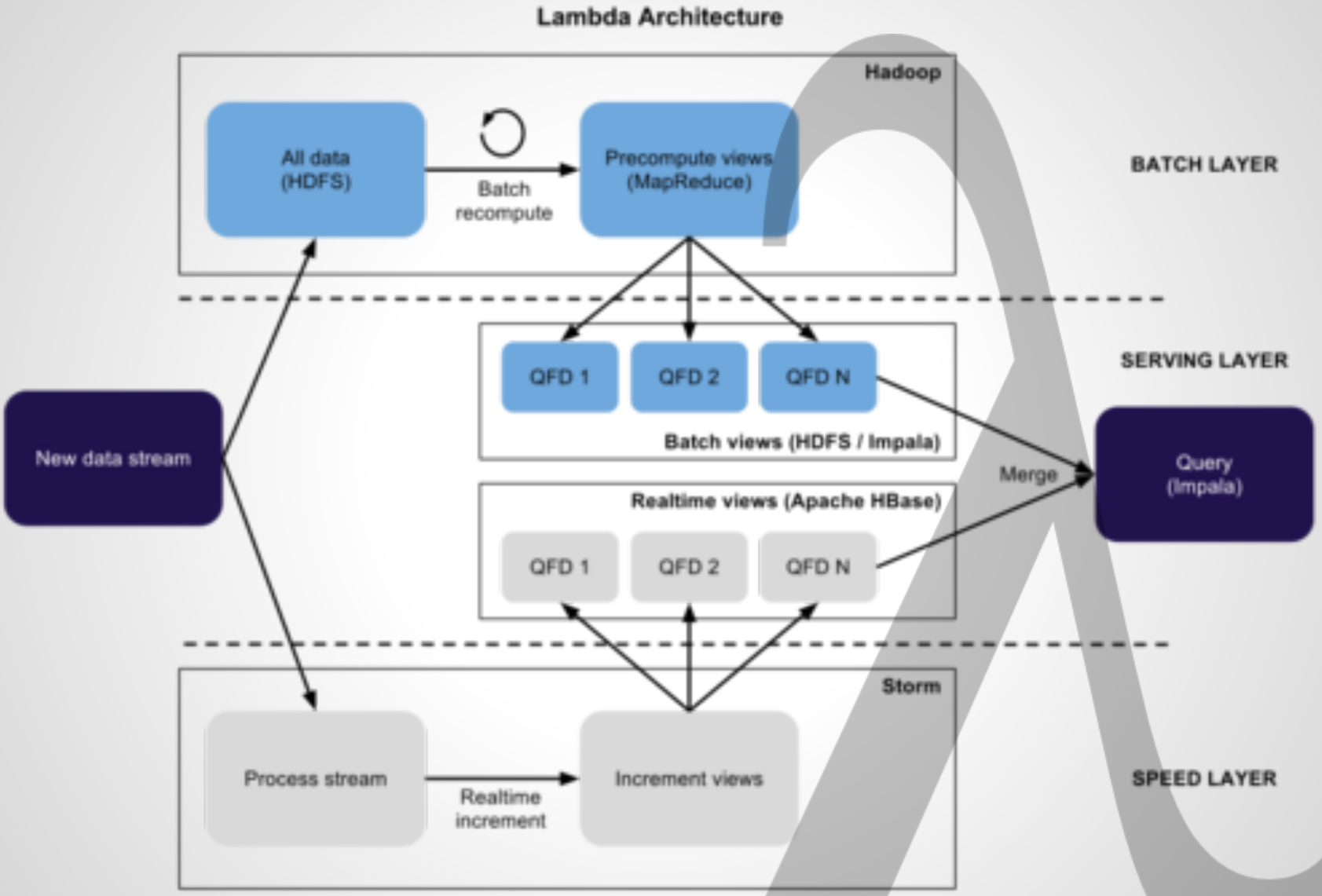
Problem #4

- SimpleCalculator

What's next?

- Document Classification and Ranking
- Malware classification, email/tweet/web spam classification
- Incident Management using sensors
- Risk Assessment - banking


Realtime views can be discarded once the data they contain is represented in the batch views
Batch
Realtime
Batch
Realtime
Batch
Realtime
Time
Now


Lambda Architecture
Hadoop
All data (HDFS)
Batch recompute
Precompute views (MapReduce)
BATCH LAYER
QFD 1
QFD 2
QFD N
Batch views (HDFS / Impala)
SERVING LAYER
New data stream
Merge
Query (Impala)
Realtime views (Apache HBase)
QFD 1
QFD 2
QFD N
Storm
Process stream
Realtime increment
Increment views
SPEED LAYER

storm-mesos

https://github.com/nathanmarz/storm-mesos

storm-yarn

https://github.com/yahoo/storm-yarn

storm-deploy

https://github.com/nathanmarz/storm-deploy

Storm @ Apache

http://wiki.apache.org/incubator/StormProposal