Storm @ Twitter: Discussant

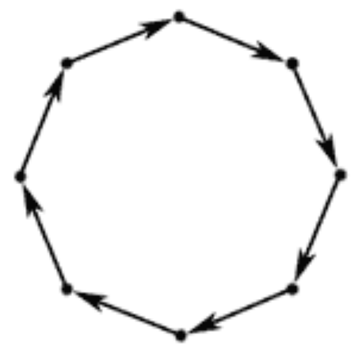
By William Truong

Overview of Storm

- Real-time, fault-tolerant and distributed stream data processing system.
  - Scalable, resilient, extensible, efficient, and easy to administer
  - Consists of streams of tuples flowing through topologies (logical query plan from a DBS perspective)
- “under active development”
  - Version 0.5.0 released in late 2011; current version (0.9.3) released in November 2014

Cycles in Topology

- Acker bolt keeps track of all tuples emitted by a spout running through topology.
- New tuples can be produced when processing a tuple; given a new message id and a timeout parameter.
  - If tuples do not complete within timeout param., tuples are marked as “failed”… then?!?!
- If a “complete” or “fail” message is not sent to acker bolt within a particular time (cycle), the tuples will be replayed back in the subsequent iteration.

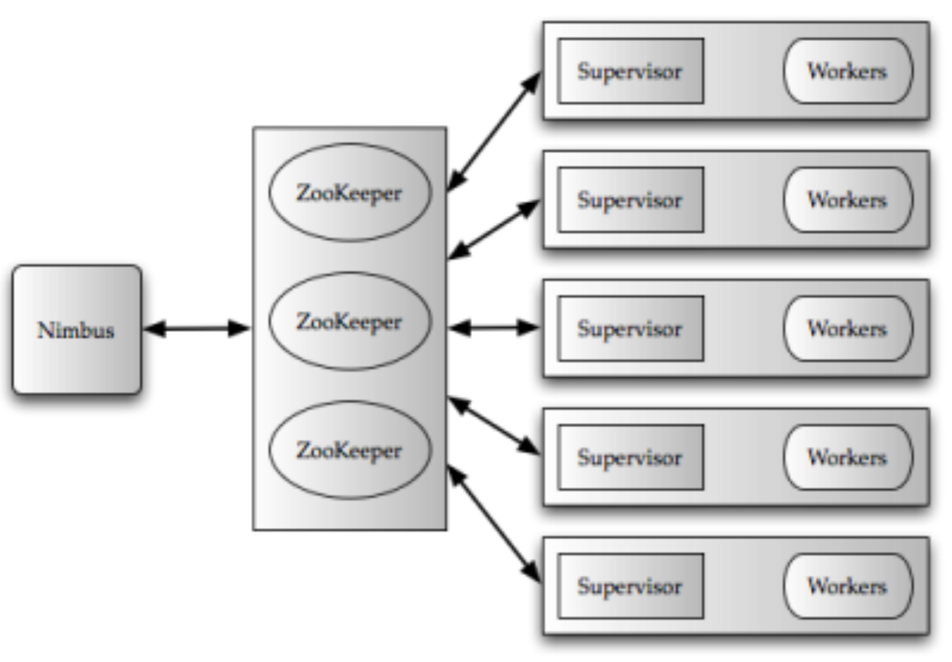


Static Topologies

- Topologies are static and un-optimized; “bad” topologies can still be run
  - “Each task is strictly bound to an executor because the assignment is currently static,” (pg. 152)
- Does not support declarative query language
  - Performance of topologies are dependent on the ability of the programmer
  - “programmer has to specify number of instances of each spout or bolt,” per topology
  - The user must specify a “limit on the number of tuples that are in flight in the topology,” (pg 153)

Nimbus Failure

- "If Nimbus is down, then users cannot submit new topologies," (pg 149)
- "If running topologies experience machine failures, then they cannot be reassigned to different machines until Nimbus is revived," (pg 149)



Other Remarks

- “Topologies are isolated on their own machines,” (pg 152)
  - Larger and more complex topologies are not possible to handle currently
- Reliance on ZooKeeper
- Questions?