Storm on Multi-core

Mark Nemec

4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2015

Abstract

This is an example of `infthesis` style. The file `skeleton.tex` generates this document and can be used to get a "skeleton" for your thesis. The abstract should summarise your report and fit in the space on the first page. You may, of course, use any other software to write your report, as long as you follow the same style. That means: producing a title page as given here, and including a table of contents and bibliography.

Acknowledgements

Acknowledgements go here.

Table of Contents

1 Introduction **3**
- 1.1 Motivation . . . 3
- 1.2 Main Idea . . . 4
- 1.3 Structure of the Report . . . 4

2 Background **5**
- 2.1 Core Concepts . . . 5
- 2.2 Example Topology . . . 6
- 2.3 Additional Concepts . . . 9
- 2.4 Storm Architecture . . . 10
  - 2.4.1 Nimbus Node . . . 11
  - 2.4.2 Worker Nodes . . . 12
  - 2.4.3 Zookeeper Nodes . . . 12
- 2.5 Tuple Processing . . . 13
- 2.6 Serialisation . . . 13

3 Bringing Storm to Multi-core **15**
- 3.1 Apache Storm on Multi-core . . . 15
- 3.2 Storm-MC Architecture . . . 16
  - 3.2.1 Nimbus . . . 16
  - 3.2.2 Worker . . . 17
  - 3.2.3 State . . . 17
- 3.3 Tuple Processing . . . 17
- 3.4 Differences between Apache Storm and Storm-MC . . . 18

4 Evaluation **19**
- 4.1 Performance . . . 19
- 4.2 Challenges . . . 19

5 Conclusion **21**
- 5.1 Future Work . . . 21

Bibliography **23**

Chapter 1

Introduction

In recent years, there has been an explosion of cloud computing software. After Google published their paper on MapReduce [?], many new open-source frameworks for distributed computation have emerged, most notably Apache Hadoop for batch processing and Apache Storm for real-time data stream processing.

The main idea of these frameworks is to split the work that needs to be carried out and distribute it across multiple nodes of a cluster. Commercial companies and researchers have been able to utilise these frameworks and create distributed systems [?] which can accomplish things that would not be possible on a single computer. This has mostly been allowed by the low price of commodity hardware and good horizontal scaling properties.

This project is about taking the ideas from the distributed system Apache Storm and applying them in the context of multi-core machines instead of clusters.

1.1 Motivation

While the cost of a commodity hardware cluster might be lower than the price of a single computer with equal power there are certain limitations:

- The nodes of a cluster communicate through network. This limits the speed of communication between processes that live on different nodes.
- Distributed systems waste resources by replicating data to ensure reliability.
- Running a distributed computation on commodity hardware usually requires a data centre or renting out instances on cloud computing services such as Amazon EC2 or Rackspace. This is not ideal for some use cases which require full control over the system or a heightened level of security.

On the other hand, even though Moore's law still holds true, processor makers now favour increasing the number of cores in CPU chips to increasing their frequency. This trend implies that the "free lunch" of getting better software performance by

upgrading the processor is over and programmers now have to design systems with parallel architectures in mind. However, there are some limitations to this as well:

- It is generally believed that writing parallel software is hard. The traditional techniques of message passing and parallel threads sharing memory require the programmer to manage the concurrency at a fairly low level, either by using messages or locks.
- Apache Storm has become the de facto tool used in stream processing on a cluster and according to their "Powered By" page [?] there are tens of companies already using Storm to process their real-time streams. It would be nice if they could keep that code.

1.2 Main Idea

The solution proposed in this paper is to take the existing Apache Storm project and port it for multi-core machines. This is implemented in a library Storm-MC with an API compatible with Apache Storm. This allows us to take an existing application written with Apache Storm in mind and run it in a multi-core setting. This way, we can avoid network latency and enjoy the significant performance improvements of a shared-memory environment.

- Prices of high-end server have decreased and one can get a 32-core machine for 10,000 USD.

1.3 Structure of the Report

In chapter 1, blah blah.

Chapter 2

Background

In this chapter we give background information necessary to understand the design of Storm-MC. Apache Storm is developed in a mix of Java and Clojure. As mentioned by the author of Storm in [7], writing the Storm interfaces in Java ensured large potential user-base while writing the implementation in Clojure increased productivity.

To ensure API compatibility with Storm, Storm-MC was developed using the same languages. This allowed for great code reuse and not getting slowed down by re-implementing functionality already present in Storm in a different language.

2.1 Core Concepts

There are several core concepts used by Storm and hence by extension Storm-MC as well. These concepts are put together to form a simple API that allows the user to break down a computation into separate components and define how these components interact with each other. The three core concepts of Storm are:

Spout
: A spout is a component that represents the source of a data-stream. Typically, a spout reads from a message broker such as RabbitMQ [1] or Apache Kafka [2] but can also generate its own stream or read from somewhere like the Twitter streaming API.

Bolt
: A bolt is a component that transforms tuples from its input data stream and emits them to its output data stream. A bolt can perform a range of functions e.g. filter out tuples based on some criteria or perform a join of two different input streams.

Topology
: The programmer connects spouts and bolts in a directed graph called topology which describes how the components interact with each other. The

topology is then submitted to Storm for execution.

2.2 Example Topology

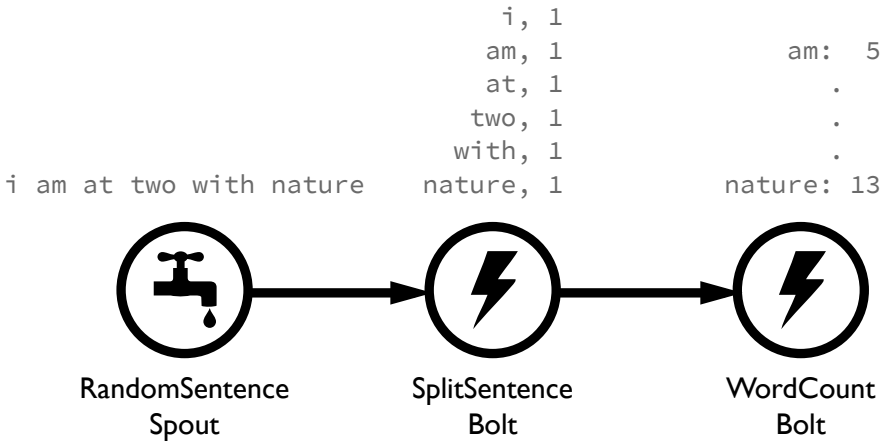


Figure 2.1: WordCount topology.

A classic example used to explain Storm topologies is the WordCount topology. In this topology, there is a spout generating random sentences, a bolt splitting the sentences on white space, and a bolt counting occurrences of every word. Figure 2.1 shows how we could represent this topology graphically.

This may seem as a simplistic example but it is useful when demonstrating how easy it is to implement a working topology using the Storm API.

```
public class WordCountTopology {
    public static void main(String[] args) throws Exception {
        TopologyBuilder builder = new TopologyBuilder();
        builder.setSpout("spout", new RandomSentenceSpout(), 5);
        builder.setBolt("split",
                new SplitSentence(), 8).shuffleGrouping("spout");
        builder.setBolt("count",
                new WordCount(), 12).fieldsGrouping("split", new Fields("word"));
        LocalCluster cluster = new LocalCluster();
        cluster.submitTopology("word-count", conf, builder.createTopology());
  }
}
```

Listing 1: WordCountTopology.java

Listing 1 shows how the topology is put together to form a graph of components. Storm uses the Builder design pattern to build up the topology which is then submitted to Storm for execution. The last argument to the setBolt/setSpout method is the number of parallel tasks we want Storm to execute for the respective component.

```
public class RandomSentenceSpout extends BaseRichSpout {
    SpoutOutputCollector _collector;
    Random _rand;

    public void open(Map conf, TopologyContext context,
            SpoutOutputCollector collector) {
        _collector = collector;
        _rand = new Random();
    }

    public void nextTuple() {
        Utils.sleep(100);
        String[] sentences = new String[]{
            "the cow jumped over the moon",
            "an apple a day keeps the doctor away",
            "four score and seven years ago",
            "snow white and the seven dwarfs",
            "i am at two with nature" };
        String sentence = sentences[_rand.nextInt(sentences.length)];
        _collector.emit(new Values(sentence));
    }

    public void ack(Object id) {}

    public void fail(Object id) {}

    public void declareOutputFields(OutputFieldsDeclarer declarer) {
        declarer.declare(new Fields("word"));
    }

}
```

Listing 2: RandomSentenceSpout.java

Listing 2 shows the definition of a spout that emits a randomly chosen sentence from a predefined collection of sentences.

```
public static class SplitSentence extends ShellBolt implements IRichBolt {

    public SplitSentence() {
        super("python", "splitsentence.py");
    }

    public void declareOutputFields(OutputFieldsDeclarer declarer) {
        declarer.declare(new Fields("word"));
    }

    public Map<String, Object> getComponentConfiguration() {
        return null;
    }
}
```

Listing 3: SplitSentence.java

```
import storm


class SplitSentenceBolt(storm.BasicBolt):

    def process(self, tup):
        words = tup.values[0].split(" ")
        for word in words:
            storm.emit([word])


SplitSentenceBolt().run()
```

Listing 4: splitsentence.py

Listings 3 and 4 show how a bolt defined in Python can be part of this Java-defined topology.

```
public static class WordCount extends BaseBasicBolt {
    Map<String, Integer> counts = new HashMap<String, Integer>();

    public void execute(Tuple tuple, BasicOutputCollector collector) {
        String word = tuple.getString(0);
        Integer count = counts.get(word);
        if (count == null)
            count = 0;
        count++;
        counts.put(word, count);
        collector.emit(new Values(word, count));
    }

    public void declareOutputFields(OutputFieldsDeclarer declarer) {
        declarer.declare(new Fields("word", "count"));
    }
}
```

Listing 5: WordCount.java

Finally, listing 5 shows how a bolt that counts the number of word occurrences can be implemented.

2.3 Additional Concepts

Stream
A stream is defined as an unbounded sequence of tuples. Streams can be thought of as edges of a topology connecting bolts and spouts (vertices).

Tuple
A tuple wraps named fields and their values. The values of fields can be of different types. When a component emits a tuple to a stream it sends that tuple to every bolt subscribed to the stream.

Stream Grouping
Every bolt needs to have a type of stream grouping associated with it. This grouping decides the means of distributing the tuples coming from the bolt's input streams amongst the instances of the bolt task. Following are the possible types of stream grouping:

Shuffle Randomly partitions the tuples among all the bolt tasks.

Fields Hashes on a subset of the tuple fields.

All Replicates the entire stream to all the bolt tasks.

Direct The producer of the tuple decides which task of the bolt will receive this tuple.

Global Sends the entire stream to a single bolt task.

Local or Shuffle Prefers sending to executors in the same worker process, if that is not possible it uses the same strategy as shuffle.

Users are also able to specify their own custom grouping by implementing the `CustomStreamGrouping` interface.

Explain why you would want to do that.

All the components of a Storm topology execute in parallel. The user can specify how much parallelism he wants associated with every component and Storm spawns the necessary number of threads. This is done through a configuration file, defined in YAML, which is submitted along with the topology.

talk more about config

2.4 Storm Architecture

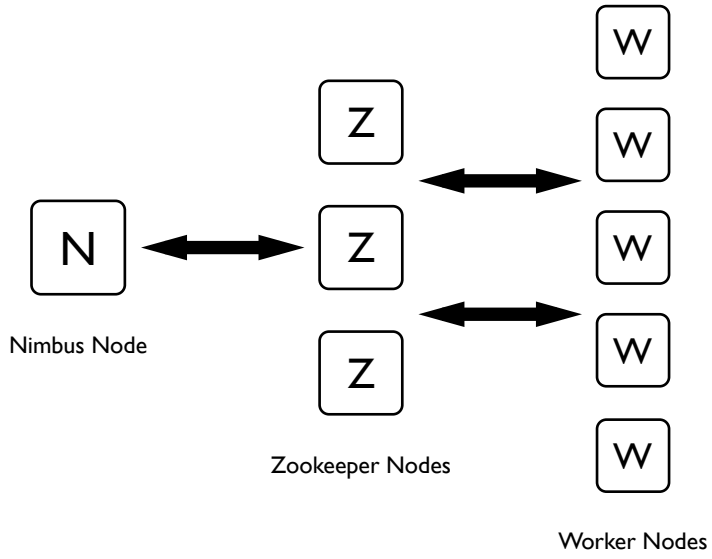


Figure 2.2: Apache Storm Architecture.

Maybe highlight similarities to Hadoop

A Storm cluster adopts the Master-Worker pattern. To set up a Storm topology, the user launches daemon processes on nodes of the cluster and submits the topology to the master node, also called Nimbus. The worker nodes receive task assignments from the master and execute on them. The coordination between the master node and the worker nodes is handled by nodes running Apache Zookeeper [3]. Figure 2.2 shows a graphical representation of Storm Architecture.

2.4.1 Nimbus Node

The master node runs a server daemon called Nimbus. The main role of Nimbus is to receive topology submissions from clients. Upon receiving a topology submission, Nimbus takes the following steps:

Validate the topology
The topology is validated using a validator to ensure that the submitted topology is valid before trying to execute it. The user can implement his own validator or use the default validator provided by Storm.

Distribute the topology source code
Nimbus ensures that the workers involved in the topology computation have the source code by sending it over the network.

Schedule the topology
Nimbus runs a scheduler that distributes the work among workers of the cluster. Similarly to validation, the user can implement his own scheduler or use the default scheduler provided by Storm. The default scheduler uses a simple Round-robin strategy.

Activate the topology
Nimbus transitions the topology to active state which tells the worker nodes to start executing it.

Monitor the topology
Nimbus continues to monitor the topology by reading heartbeats sent by the worker nodes to ensure that the topology is executing as expected and worker nodes have not failed.

Nimbus is an Apache Thrift [4] service (more on Thrift in section 2.6) that listens to commands submitted by clients and modifies the state of a cluster accordingly. Following are the commands supported by Nimbus:

Submit a topology
Clients can submit a topology defined in a Java Archive (JAR) file. The Nimbus service then ensures that the topology configuration and resources are distributed across the cluster and starts executing the topology.

Kill a topology
Nimbus can stop running a topology and remove it from the cluster. The cluster can still continue executing other topologies.

Activate/deactivate a topology
Topologies can be deactivated and reactivated by Nimbus. This could be useful if the spout temporarily cannot produce a stream and the user does not want the cluster to execute idly.

Rebalance a topology
Nimbus can rebalance a topology across more nodes. Thus if the number of

nodes in the cluster changes the user can increase or decrease the number of nodes involved in the topology.

2.4.2 Worker Nodes

The worker nodes run a daemon called Supervisor. There are 4 layers of abstraction which control the parallelism of a worker node.

Supervisor
A supervisor is a daemon process the user runs on a worker node to make it part of the cluster. It launches worker processes and assigns them a port they can receive messages on. Furthermore, it monitors the worker processes and restarts them if they fail. A worker node runs only one supervisor process.

Worker
A worker process is assigned a port and listens to tuple messages on a socket associated with the port. A worker launches executor threads as required by the topology. Whenever it receives a tuple, it puts it on a queue where it is picked up by one or more executors of the worker process.

Furthermore, the worker has a transfer queue where its executors enqueue tuples ready to be sent downstream. There can be multiple workers processes running inside one supervisor.

Executor
An executor controls the parallelism within a worker process. Every executor runs in a separate thread. An executor's job is to pick up tuples from the receiver queue of the worker, perform the task of a component it represents, and put the transformed tuples on the transfer queue of the worker. There can be many executors running inside one worker and an executor performs one (the usual case) or more tasks.

Task
A task represents the actual tuple processing function. However, within an executor thread all the tasks are executed sequentially. The main reason for having tasks is that the number of tasks stays the same throughout the lifetime of a topology but the number of executors can change (by rebalancing). Thus if some worker nodes in the cluster go down, the topology can continue executing with the same number of tasks as before.

2.4.3 Zookeeper Nodes

The Storm cluster contains a number of Zookeeper nodes which coordinate the communication between the master and the workers. Storm does this by storing the state of the cluster on the Zookeper nodes where both Nimbus and worker nodes can access it.

The cluster state contains worker assignments, information about topologies, and heartbeats sent by the worker nodes back to Nimbus. Apart from the cluster state, Storm is completely stateless. Hence, if the master node or a worker node fail the cluster continues executing. The only time the cluster stops executing is if all the Zookeper nodes die.

2.5 Tuple Processing

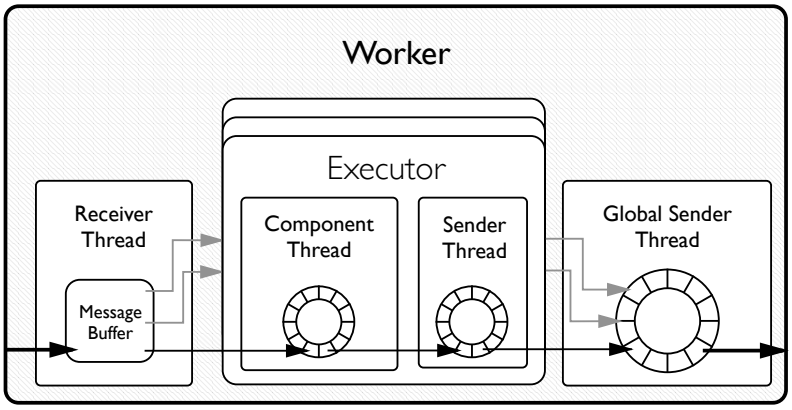


Figure 2.3: Tuple processing in Apache Storm.

Figure 2.3 shows how tuple processing is implemented inside a Storm worker process. The tuple is read from a message buffer by the receiver thread of the worker and put on a corresponding executor's receiver queue. The tuple is then picked up by the component thread of an executor.

After the component thread has executed the task it puts the tuple on the sender queue. There, it is picked up by the executor's sender thread which puts the tuple on the global sender queue of the worker. Lastly, the global sender thread of the worker serialises the tuple and sends it downstream.

Alternatively, if the tuple is forwarded to an executor in the same worker process it is put on the receiver queue of the corresponding executor directly after processing.

Storm uses the term queue freely but in fact the queues are implemented as ring buffers using LMAX Disruptor [?]. Detailed background of how Disruptor works and its performance benchmarks can be found in [8].

2.6 Serialisation

Since Storm topologies execute on a cluster all the components need to be serialisable. This is achieved with Apache Thrift. Components are defined as Thrift objects and Thrift generates all the Java serialisation code automatically.

Why is Thrift good?

Furthermore, since Nimbus is a Thrift service Thrift generates all the code required for remote procedure call (RPC) support. This allows defining topologies in any of the languages supported by Thrift and easy cross-language communication with the Nimbus service.

Chapter 3

Bringing Storm to Multi-core

The design of Storm-MC was ported over from Apache Storm. This enabled rapid progress while guaranteeing compatibility with Apache Storm API. Clearly, however, some differences had to be made to take advantage of multi-core machine performance implications. This chapter explains the design of Storm-MC.

3.1 Apache Storm on Multi-core

To begin, we discuss why Apache Storm does not perform optimally on a single multi-core machine. Storm can be ran in local mode where it emulates execution on a cluster. This mode exists so that it is possible to debug and develop topologies without needing access to a cluster. However, there are several reasons why the local mode is not as performant as it could be.

We described tuple processing in Storm in section 2.5. Clearly, there is a lot of overhead necessary to simulate sending tuples to executors in other worker nodes. This overhead comes not only from the tuple passing through several queues but also includes running a Netty [5] server used for message passing.

Storm runs many threads which are only useful in distributed context. Indeed, during our experiments we found that a topology with 8 executors was being executed with 64 threads.

Plot threads vs components

Maybe include Thread Dump and a graph of thread counts

Additionally, every worker has a heartbeat thread that simulates sending heartbeat messages to the Nimbus process. It does this by writing to a local cache which is persisted to a file by a blocking write on every heartbeat. While heartbeats are essential in cluster mode, there is no need for them in local mode.

Another overhead is running a local Zookeper server which emulates the Zookeeper nodes of a cluster. State is maintained in this server even though there is always only one topology executing at a time.

3.2 Storm-MC Architecture

Probably don't call this Architecture

The design we adopted for porting worker nodes is to only have one worker process running all the executor threads of a topology.

Additionally, the code for the Nimbus daemon was merged with the worker. This was done because there is no need to run the Nimbus and worker specific code at the same time. Once Nimbus sets up the topology, all the work is done by the worker. Hence they can be executed serially.

3.2.1 Nimbus

Unlike Nimbus executing on a Storm cluster, Nimbus in Storm-MC does not support running multiple topologies at the same time. However, to do that one only needs to run the topology in a separate process. This is because unlike when executing on the cluster different topologies do not need to share any state and it is more natural to execute them as separate processes.

This has the added benefit of each process having its own part of main memory thus reducing cache conflicts as shown in [6] and providing higher security by not having different topologies share memory space. Additionally, if a single thread of one topology is blocking it does not block other topologies.

Nimbus on Storm-MC does not support scheduling topologies. Since within one process there is only one topology running at a time and the hardware configuration of the machine does not change, the parallelism is clearly defined by the number of executors per component specified in the topology configuration.

One way to implement scheduling could be to pin threads to specific cores. Unfortunately, Java does not provide support for CPU affinity, the assignments are handled automatically by the JVM. Potentially, this could be achieved by using C or C++, both of which support CPPU affinity, but this was not implemented in Storm-MC.

Mention porting over from Thrift.

The role of Nimbus in Storm-MC has effectively been reduced to validating the topology and passing it along to the worker part of the process which handles the topology execution.

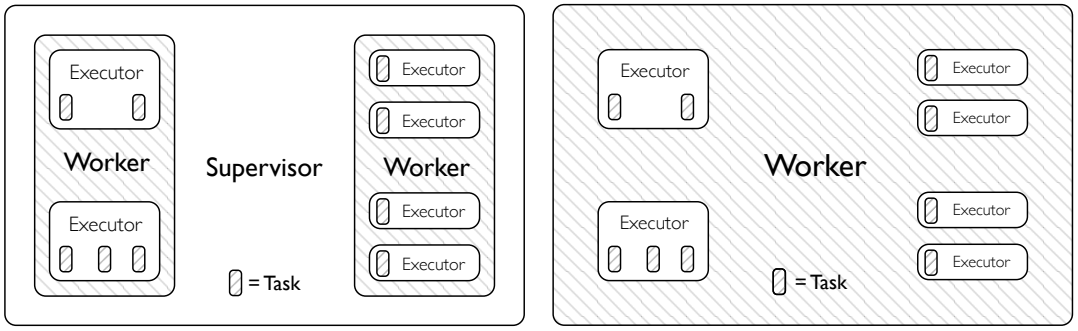
3.2.2 Worker

In Apache Storm, a worker node runs the supervisor daemon, which in turns launches worker processes which contain executor threads which contain tasks. In Storm-MC, however, there is only one worker process which contains all the executor threads and their tasks.

This design has several benefits:

- All the inter-thread communication is occurring within one Worker.
- Supervisor can be removed as there is no need to synchronise workers.
- There is no need to simulate over-the-network message passing.
- Message passing between executor threads within a worker stays the same as in Apache Storm.

A comparison of an Apache Storm worker node and its Storm-MC equivalent can be seen in figure 3.1.



(a) Worker in Apache Storm. (b) Worker in Storm-MC.

Figure 3.1: Comparison of a worker in Storm and Storm-MC

3.2.3 State

In light of previous subsections Storm-MC is completely stateless. The cluster state that was managed by Zookeeper in Apache Storm was completely stripped away. This state was only relevant when multiple topologies were sharing the cluster.

3.3 Tuple Processing

The implementation of tuple processing in Storm-MC can be seen in figure 3.2. As can be seen from the figure, the queues used for remote message sending were stripped away and there is only one Disruptor queue for every executor. Once an executor is done processing a tuple it simply puts it on the Disruptor queue of the bolts downstream.

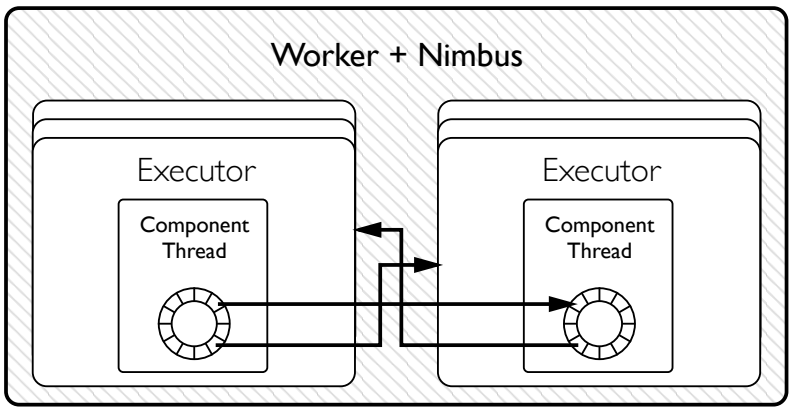


Figure 3.2: Tuple processing in Storm-MC.

There were several other options we considered when implementing tuple processing. However, the Disruptor shows superior throughput and latency compared to alternative solutions in [?].

find the citation for above claim.

3.4 Differences between Apache Storm and Storm-MC

Maybe present this as a table of features.

Chapter 4

Evaluation

In this chapter we evaluate Storm-MC. We do this by comparing its performance against the local mode of Apache Storm.

4.1 Performance

4.2 Challenges

In this section we are going to discuss challenges we encountered while porting Apache Storm to multi-core machines.

Chapter 5

Conclusion

5.1 Future Work

Bibliography

[1]

[2]

[3]

[4]

[5]

[6] Dhruba Chandra, Fei Guo, Seongbeom Kim, and Yan Solihin. Predicting inter-thread cache contention on a chip multi-processor architecture. In *Proceedings of the 11th International Symposium on High-Performance Computer Architecture*, HPCA '05, pages 340–351, Washington, DC, USA, 2005. IEEE Computer Society.

[7] Nathan Marz. History of apache storm and lessons learned - thoughts from the red planet - thoughts from the red planet, Oct 2014.

[8] Martin Thompson, Dave Farley, Michael Barker, Patricia Gee, and Andrew Stewart. Disruptor: High performance alternative to bounded queues for exchanging data between concurrent threads. May 2011.