Data-aware scheduling
Spark on Kubernetes with HDFS

Johannes M. Scheuermann

Karlsruhe, 18.10.2017
Johannes M. Scheuermann
IT Engineering & Operations @ inovex

› Software-Defined Datacenters
› Infrastructure as Code
› Cloud technologies
› High Availability & Scalability
› Want’s an IBM Z-Frame
› @johscheuer
Data-aware scheduling
Agenda

- Why data-aware scheduling
- Data-ware for non Big-Data application
- Data-ware scheduler
- Big Data on Kubernetes
  - Spark on Kubernetes
  - HDFS on Kubernetes
• I’m not a scheduling expert
• Concept is/was a PoC
• Share learnings/ideas
• Get feedback from the community
Why data-locality?
Data-aware scheduling for non Big-Data

• Databases
• (large) image processing
• Video encoding
• (Web)-Cache
Quobyte – What is Quobyte

- Distributed (parallel) POSIX file system
  - Any workload with high performance (incl. throughput, databases, small files)
- Can be deployed in containers, on kubelet hosts.
  - Linearly scalable performance.
- Fully fault-tolerant, split-brain safe
Quobyte - Architecture

1. open file
2. parallel read/write data
Quobyte - Placement

- Metadata servers make placement decisions against policies
  - on file level
  - tiering, isolation, ...
  - keep stripes of files on disks of same machine => enable local read
  - allow preferring writes to local storage servers => enable local write
- Locality information can be retrieved per file
  - that’s where the scheduler hooks in
Running multiple schedulers

1. Monitor unscheduled Pods
2. Get Nodes (cached)
3. Data-aware Scheduler
4. Bind pod to node (schedule)

SchedulerName == Myself?

Calculate best fit

InCluster or Out-of-Cluster
Scheduling data-aware (file-based)

- Specify wanted Data
- Lookup Data Placement
- Remapping if Storage runs in Containers
- Schedule Pod
Scheduler Architecture (4000ft)
Scheduler Architecture (1000ft)
Scheduler Architecture (containerized)
Benchmarks

Benchmark with fio (20GB)

Max Read (MB/s)

Avg Read (MB/s)

data-local  non data-local  data-local  non data-local
(Spark) Big-Data on Kubernetes
Spark on Kubernetes

- [https://github.com/apache-spark-on-k8s/spark](https://github.com/apache-spark-on-k8s/spark)
- Not the “faked” Spark on Kubernetes
- Still in development
- Still not in the official Apache Spark project
- Current: v2.2.0-0.4.0
  - Alpha/Beta ?!
Spark on Kubernetes

Spark Core

GraphX  SparkSQL  MLlib  Streaming

Kubernetes  Standalone  YARN  Mesos
Spark on Kubernetes

- Integrates with Kubernetes
  - RBAC
  - Resource Quotas
  - Audit logging
  - Etc.

- Only cluster-mode
Spark on Kubernetes (cluster-mode)
Spark + HDFS on Kubernetes

- Driver 1
  - Executor 2.1
- Executor 1
  - Driver 2
  - Executor 2.2
- HDFS NN
- HDFS DN 1
- HDFS DN 2

Pod Network

Host Network
Demo
Missing pieces

- Rack-locality
- Node preferences
- Priority-based scheduling (K8s 1.8 alpha)
- NameNode HA
- Kerberos support
Conclusions
Conclusions

• Good starting point
• Good integration
• Still some points open
• Work for better integration (more general)
• Play with it!
We are hiring!

Become an inovexpert

www.inovexperts.com
Further reading

- https://issues.apache.org/jira/browse/SPARK-18278
- https://www.youtube.com/watch?v=0xRHONrWwvU&feature=youtu.be
- https://www.youtube.com/watch?v=DxCDxi08HWo&feature=youtu.be