Big Data for Genomics

SCALE

Scalable Computing Systems

Jim Dowling, Salman Niazi,
Mahmoud Ismail, Gautier Berthou,
Kamal Hakimzadeh
@ ICT, KTH
Ali Gholami, Ewrin Laure.
@ PDC, KTH
Jan-Eric Litton, Roxana Martinez
@ KI

Talk Outline

- Overview of Big Data
- Big Data for Whole Genome Sequencing
- Hadoop Open Platform-as-a-service (Hop)
- Funding Acknowledgements
  - SeRC Big Data Social Science
  - SeRC eCPC
  - EU FP7 BiobankCloud
  - EIT Europa (Cloud Computing Action Line)

What is Big Data?

Small Data

Big Data

Why is Big Data Important in Science?

- In a wide array of academic fields, the ability to effectively process data is superseding other more classical modes of research.

“More data trumps better algorithms”*

*“The Unreasonable Effectiveness of Data” [Halevey et al 09]

4 Vs of Big Data

- Volume
- Velocity
- Variety
- Veracity/Value/...

4 Vs of Big Data for Genomics

Whole Genome Sequencing Pipeline

30-60X genome

FastQ files (~250 GB) → **Align** → BAM file (~100 GB) → **Variant Calling** → VCF file (~10 MB) → **Analysis** → Results

http://blog.illumina.com/blog/illumina/2014/01/14/the-power-and-promise-of-population-scale-genomics

Population-Scale WGS: $1000 per Genome

HiSeq X Ten => 18,000 genomes/year^

Volume => 20 PB/year*

Velocity => 634 MB/sec*

Value => 634 MB/sec*

^Cost ~$10 million

*Assuming a replication factor of 3 and 30X coverage

Storage, Analysis, Administration Costs

180TB for $9,305

http://blog.backblaze.com/2014/03/19/backblaze-storage-pod-4

20PB for $1,033,888

http://blog.backblaze.com/2014/03/19/backblaze-storage-pod-4

But what about the Administration Costs...

Administration Costs

Facebook Operations staffers manage 20-26,000 servers each^

^ http://allfacebook.com/20000-servers_b127053

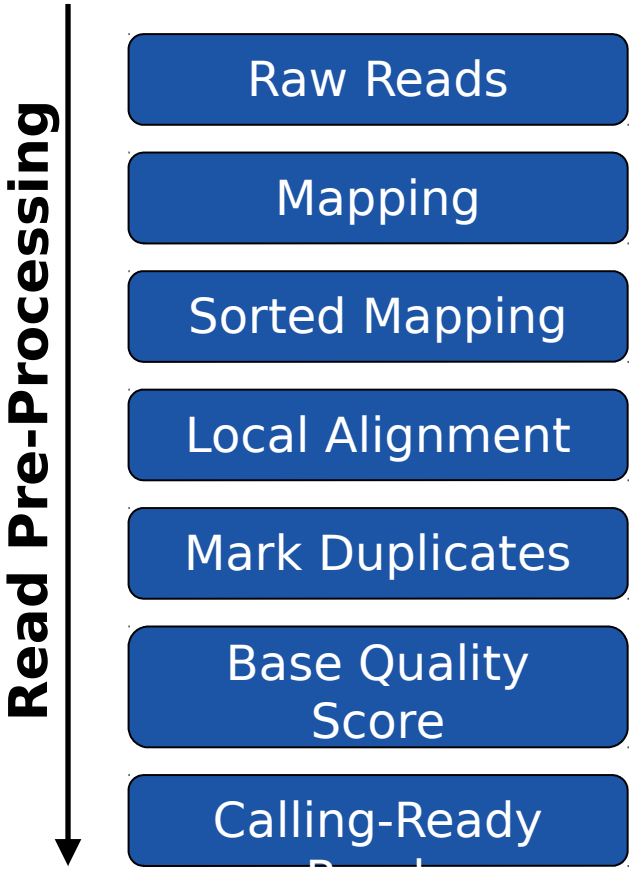
The Biobank Bottleneck: Genomics software

Pipeline Issues in popular NGS toolkits

- The time taken to get answers from reads is too long
- Population-level statistical analysis requires petabytes of data
- Standard analysis of genomes does not even scale to thousands of genomes

Single-Machine Genome Analysis using GATK

Stage	**GATK 2.7/NA12878**
Mark Duplicates	13 hours
BQSR	9 hours
Realignment	32 hours
Call Variants	8 hours
Total	**62 hours***



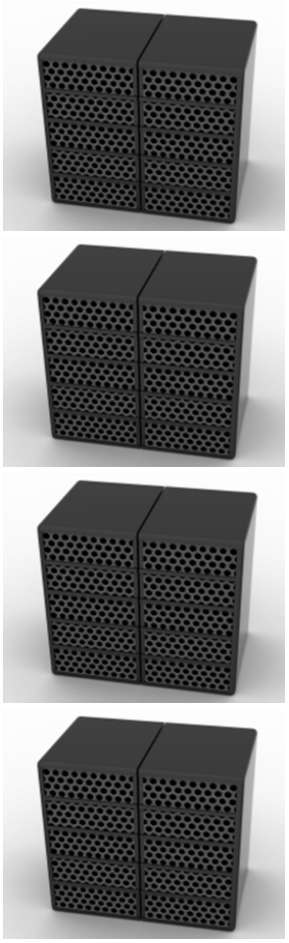
*http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-207.pdf

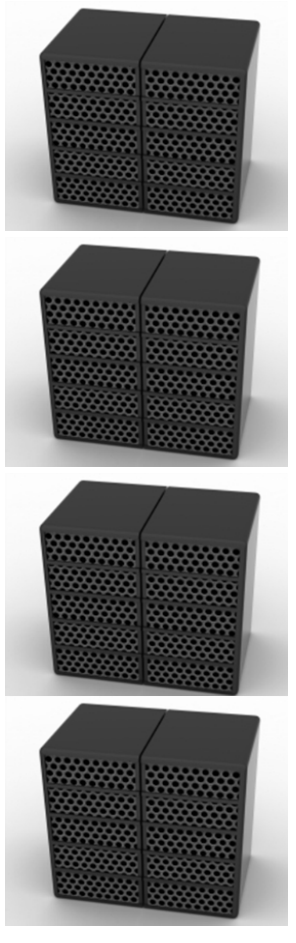
Bottleneck for serial analysis pipelines

Read genome on 1 machine:
~1000 secs

[Harddisk Image courtesy of sorapop / http://www.freedigitalphotos.net]

Big Data means Parallelization





Read genome on 100 machines:
~10 seconds

Big Data Genomics with ADAM/Spark/HDFS

Speedup using ADAM/Spark/HDFS*

Sort (250 GB)

Picard	1 hs1.8xlarge	**17h 44m**
ADAM	100 m2.4xlarge	**21m**

Mark Duplicates (250 GB)

Picard	1 hs1.8xlarge	**20h 22m**
ADAM	100 m2.4xlarge	**29m**

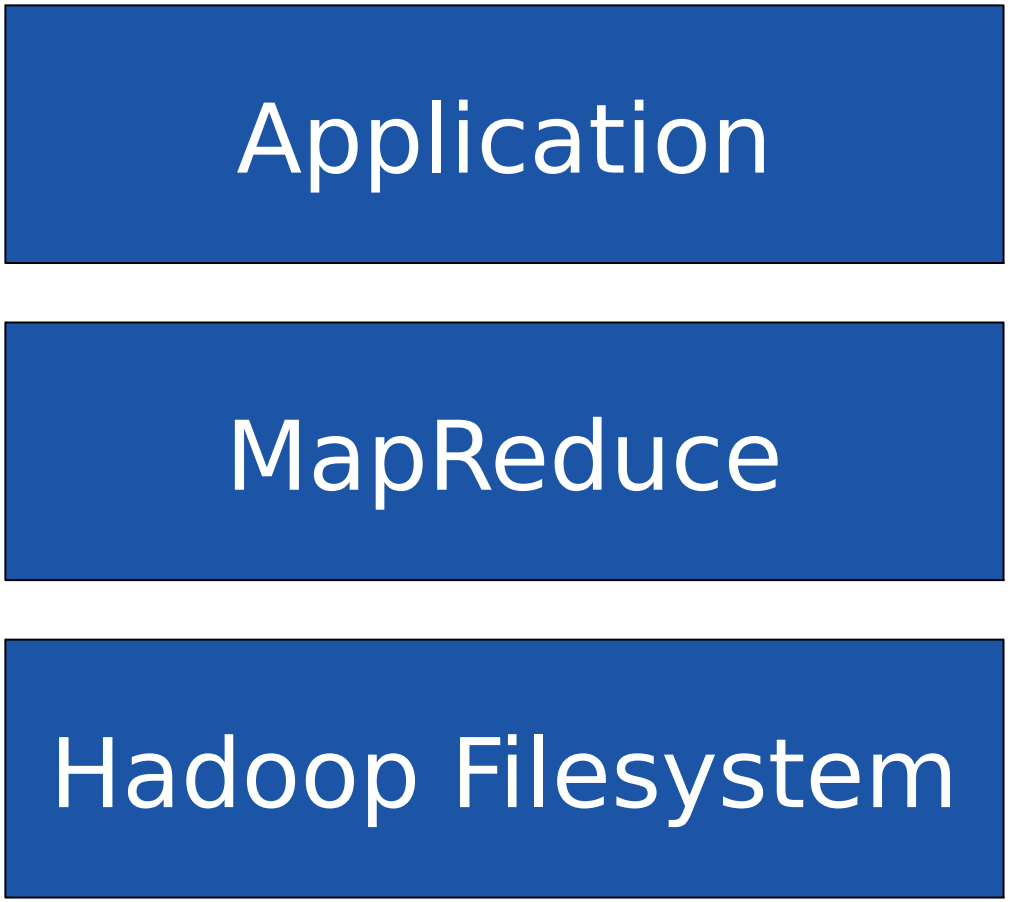
* http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-207.pdf

Storage and Processing of Big Data

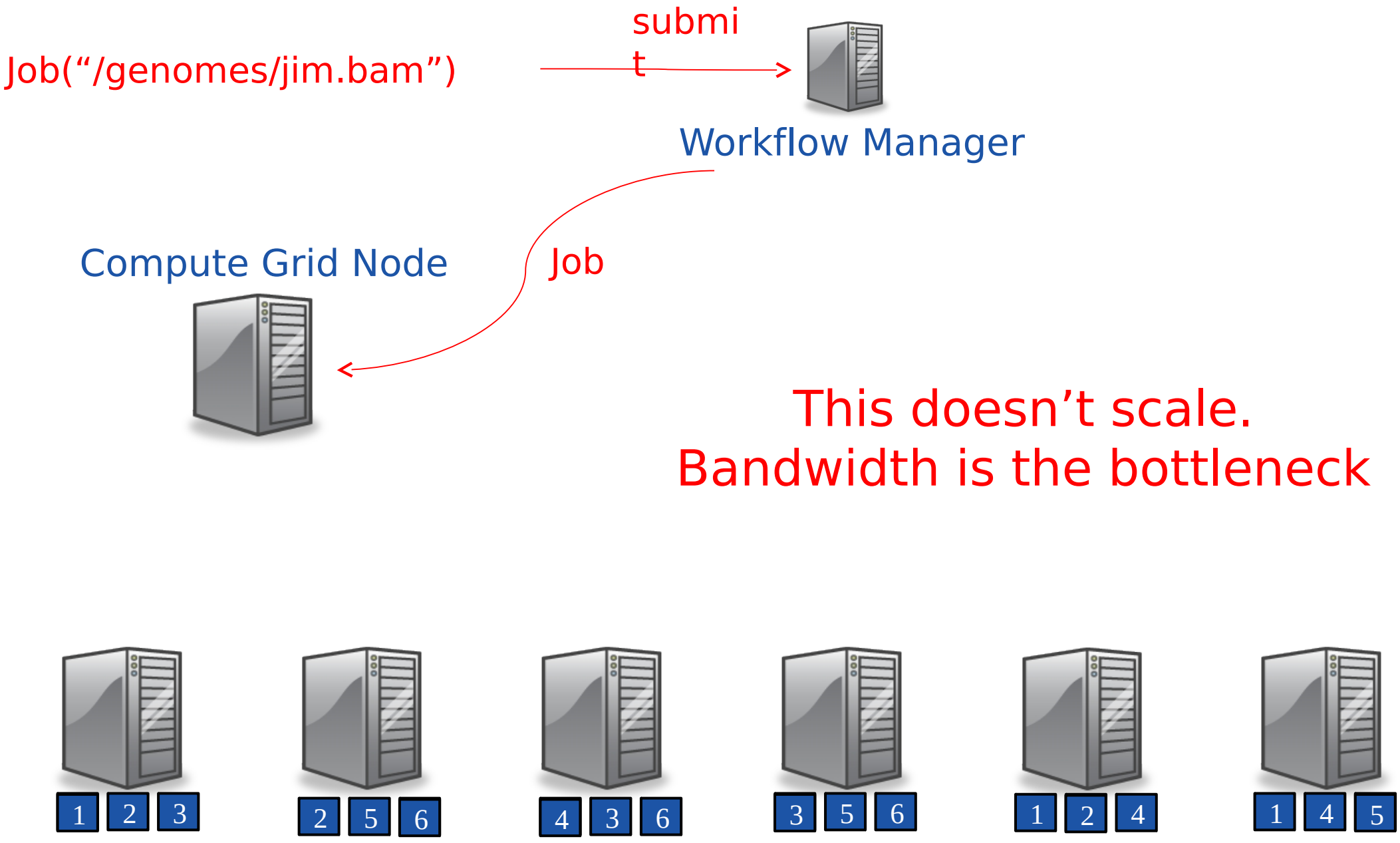
What is Apache Hadoop?

- Gigabytes files, petabyte data sets
  - Scales to thousands of nodes on commodity hardware
- No Schema Required
- Fault tolerant
- Network topology-aware, Data Location-Aware
- Optimized for analytics: high-throughput file access

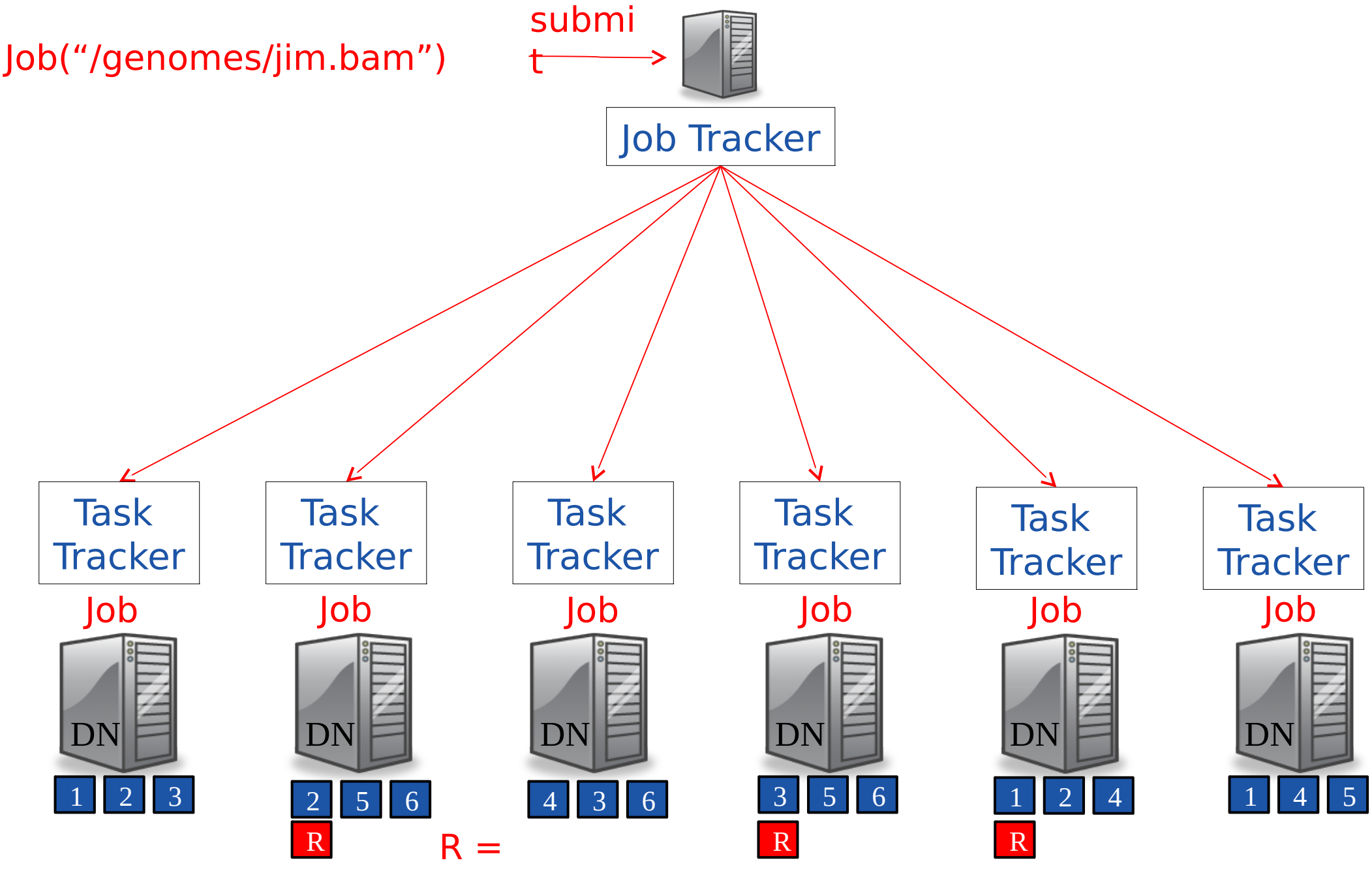
Hadoop (version 1)



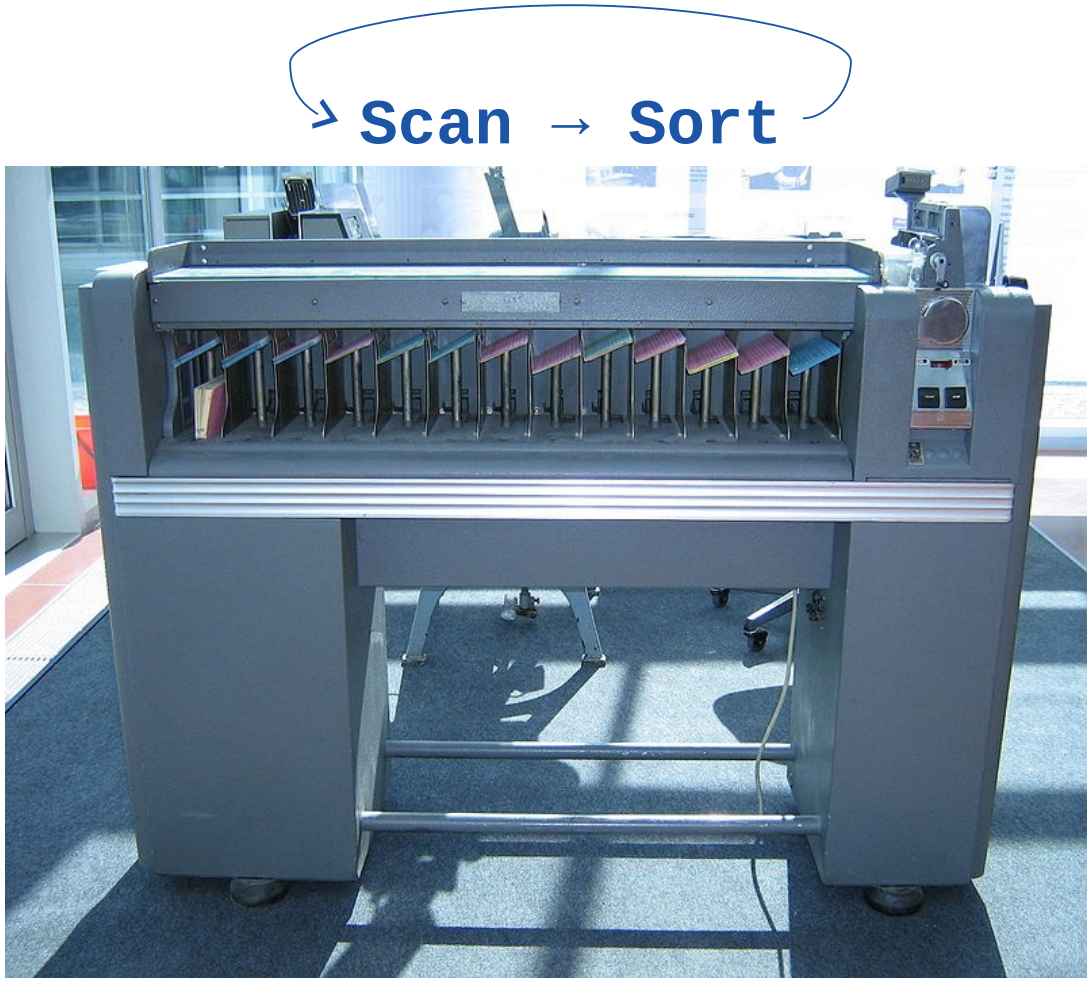
Big Data Processing with No Data Locality



MapReduce – Data Locality



MapReduce Programming Model – Batch Sequential Processing



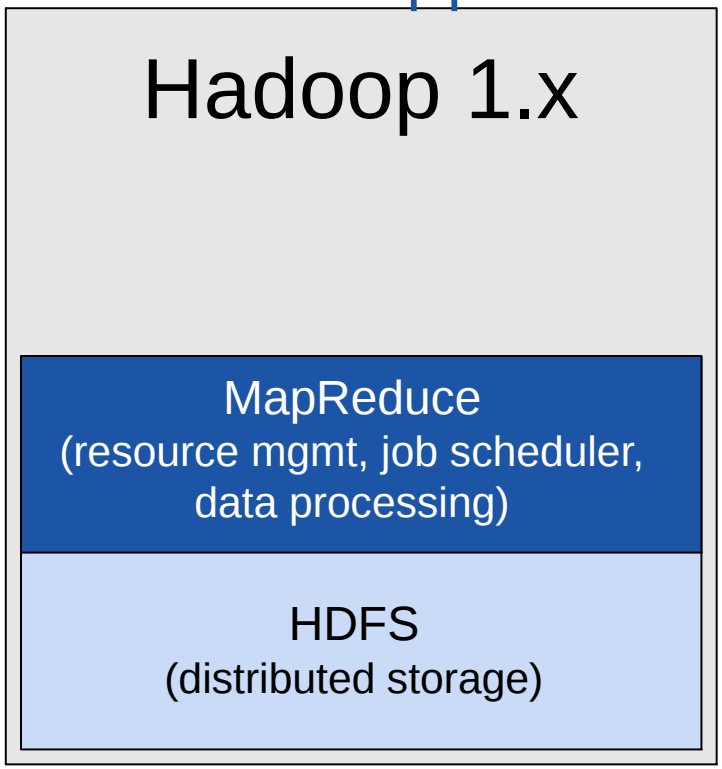
IBM 082 Punch Card Sorter

With Fault Tolerance

Hadoop 2.x

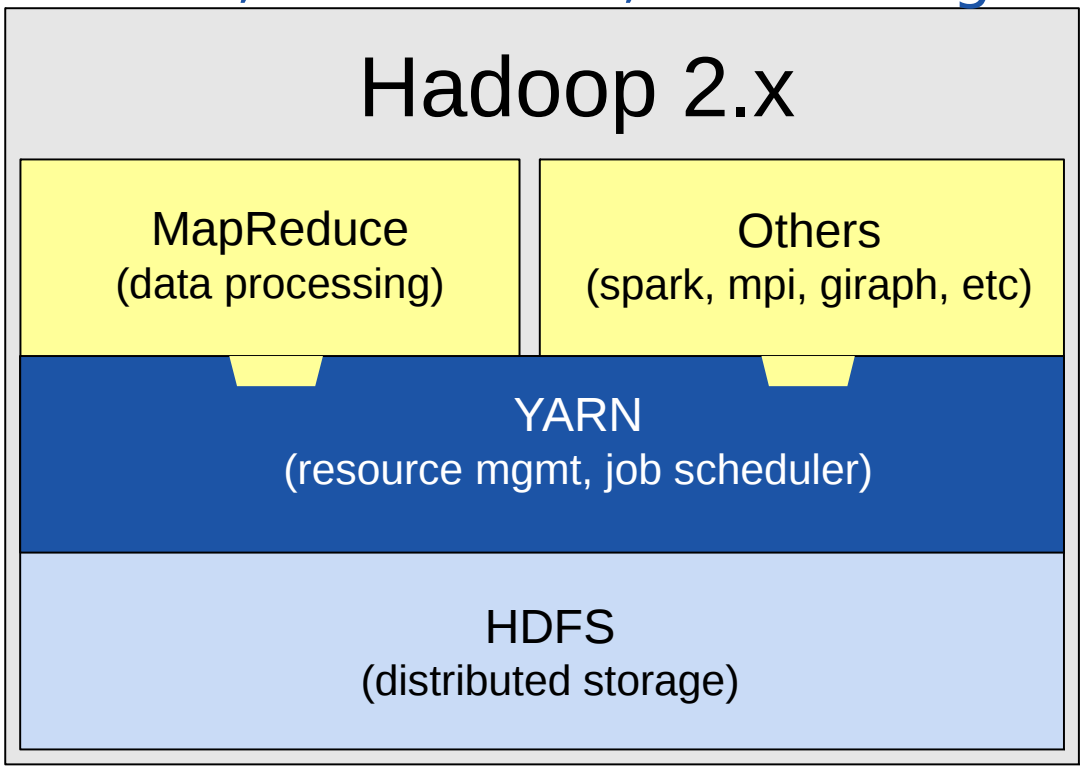
Single Processing Framework

Batch Apps



Multiple Processing Frameworks

Batch, Interactive, Streaming ...



New Data Processing Frameworks

```
val input= TextFile(textInput)

val words = input
    .flatMap
      { line => line.split(" ") }

val counts = words
    .groupBy
        { word => word }
    .count()

val output = counts
    .write (wordsOutput,
        RecordDataSinkFormat() )

val plan = new ScalaPlan(Seq(output))
```

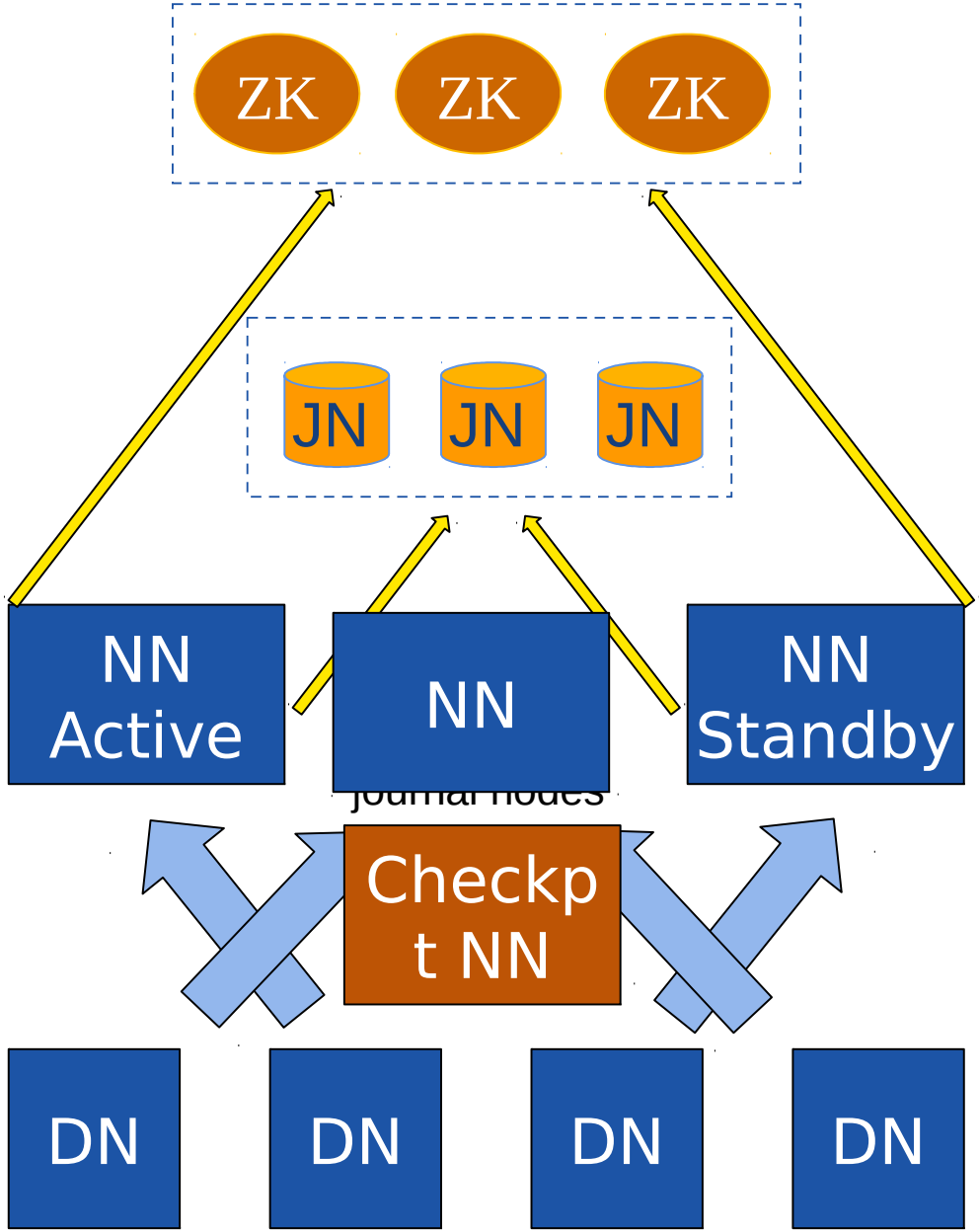
Storage in HDFS v2

High Availability in HDFS 2.0

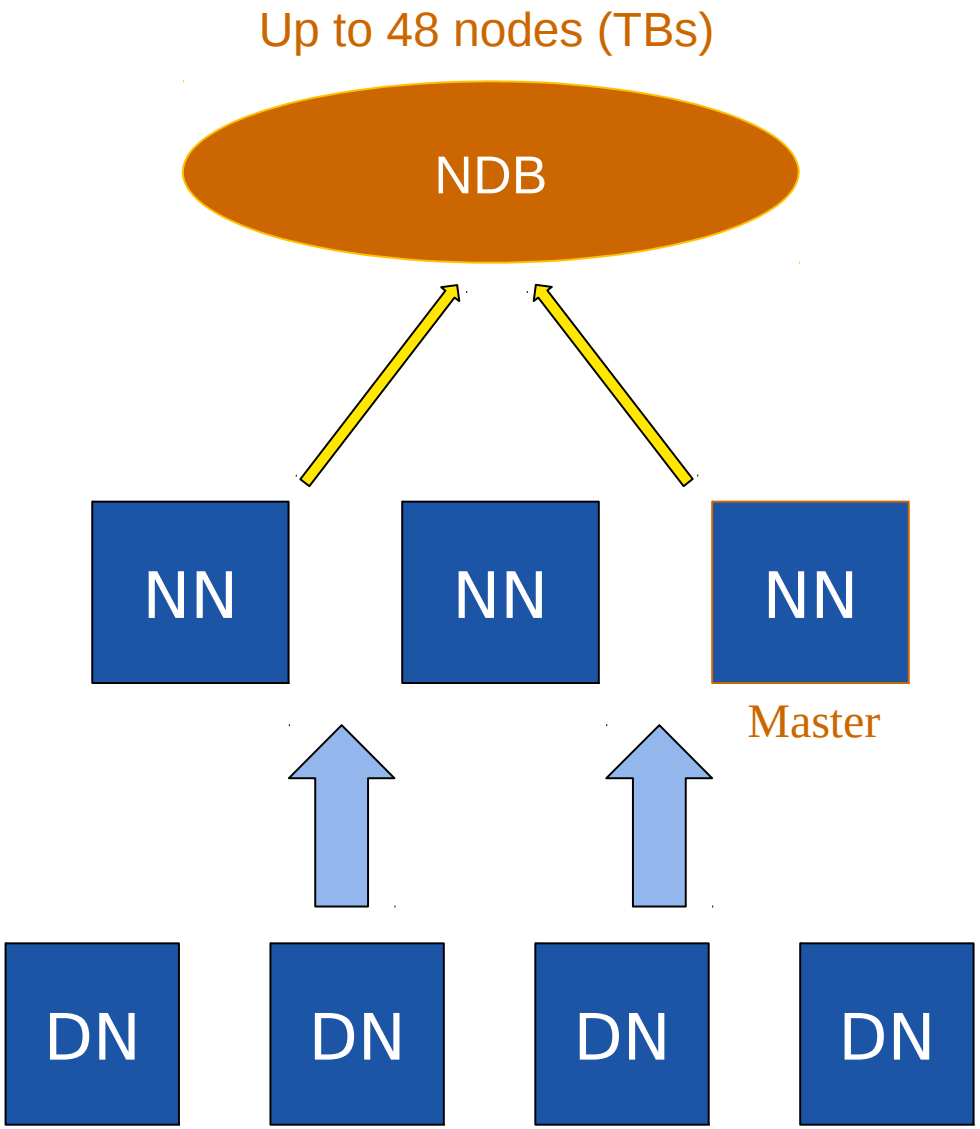
Agreement on the Active Master

Master-Slave Replication of NN State.

Faster Recovery, Cut Journal Log

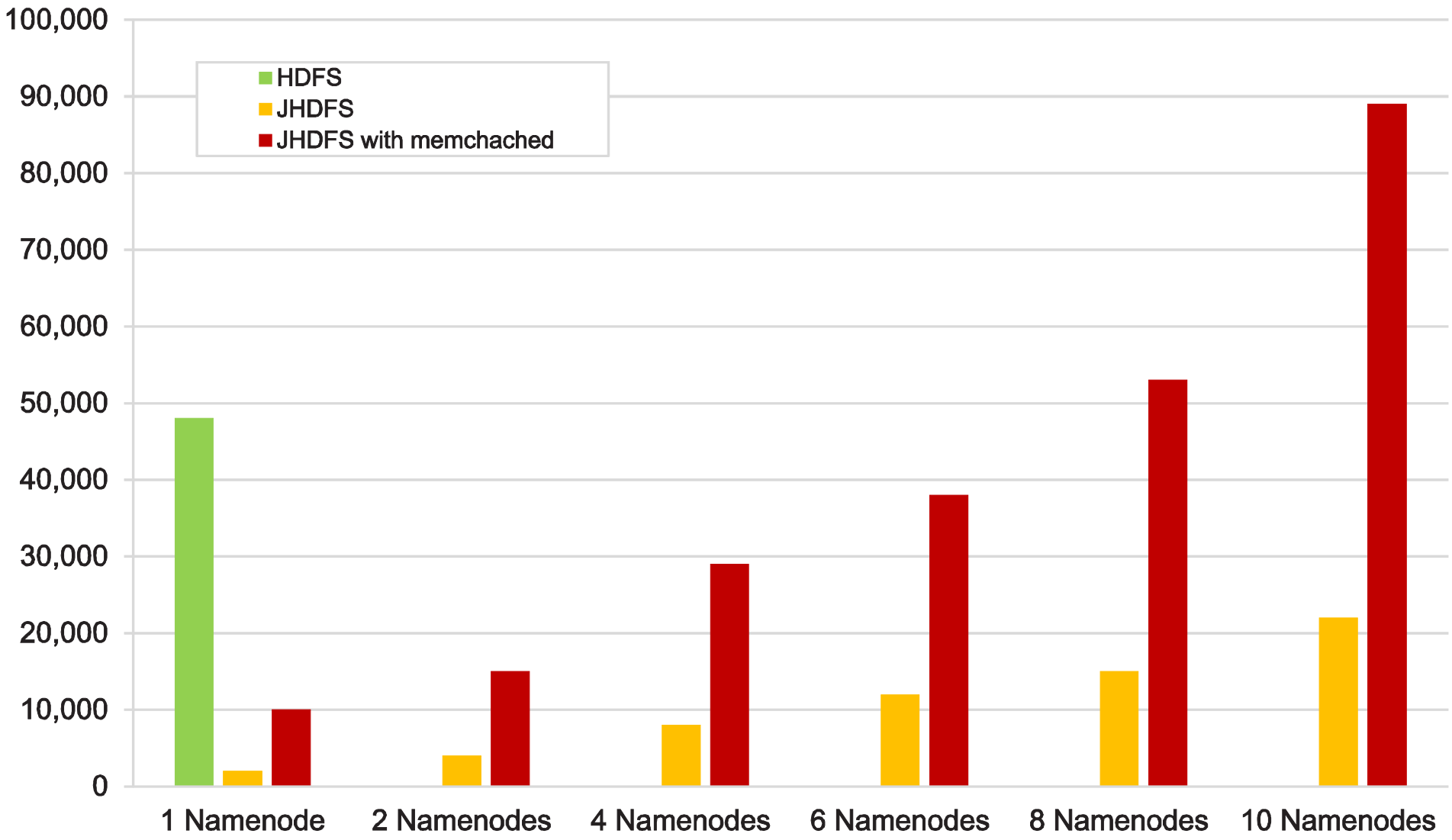


Hop-HDFS



Kamal Hakimzadeh, Hooman Peiro, and Jim Dowling, Scaling HDFS by moving Metadata to a Relational Database, DAIS, 2014.

Hop-HDFS Read Ops/Sec



HDFS' NNThroughput Benchmark

Terabytes of Meta-Data in Hop-HDFS

Meta data analysis

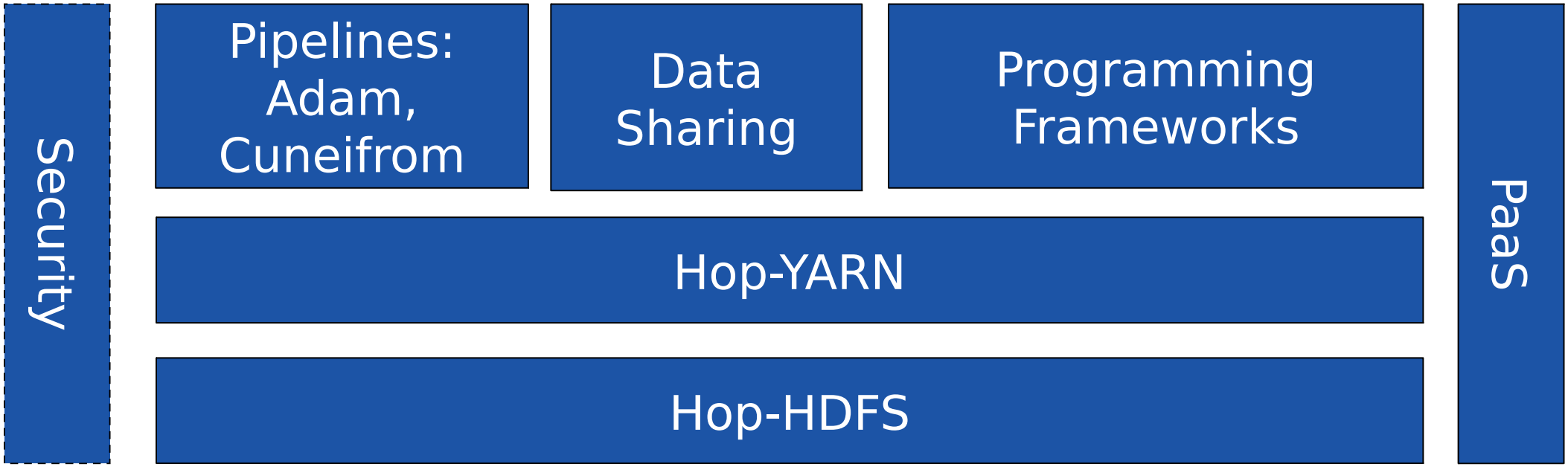
	MBs	GBs	No of NDB Nodes with 60 GB RAM
1 file	0	0	1
100 million files	127780	125	2
500 million files	638902	624	10
1 billion files	1277804	1248	21
2 billion files	2555609	2496	42

	Hop-HDFS	HDFS
Max Metadata capacity	4 TB	64 GB
Max files	3 Billion	100 million

Hadoop Open Platform-as-a-service (Hop)

- Platform-as-a-Service Support
  - Installation, Management and Monitoring
- Erasure-Coded Replication (~50% less storage)
- Block-level Indexing
  - Efficient statistical analysis of genomes
- Secure
  - Identity management and multi-tenancy

BiobankCloud

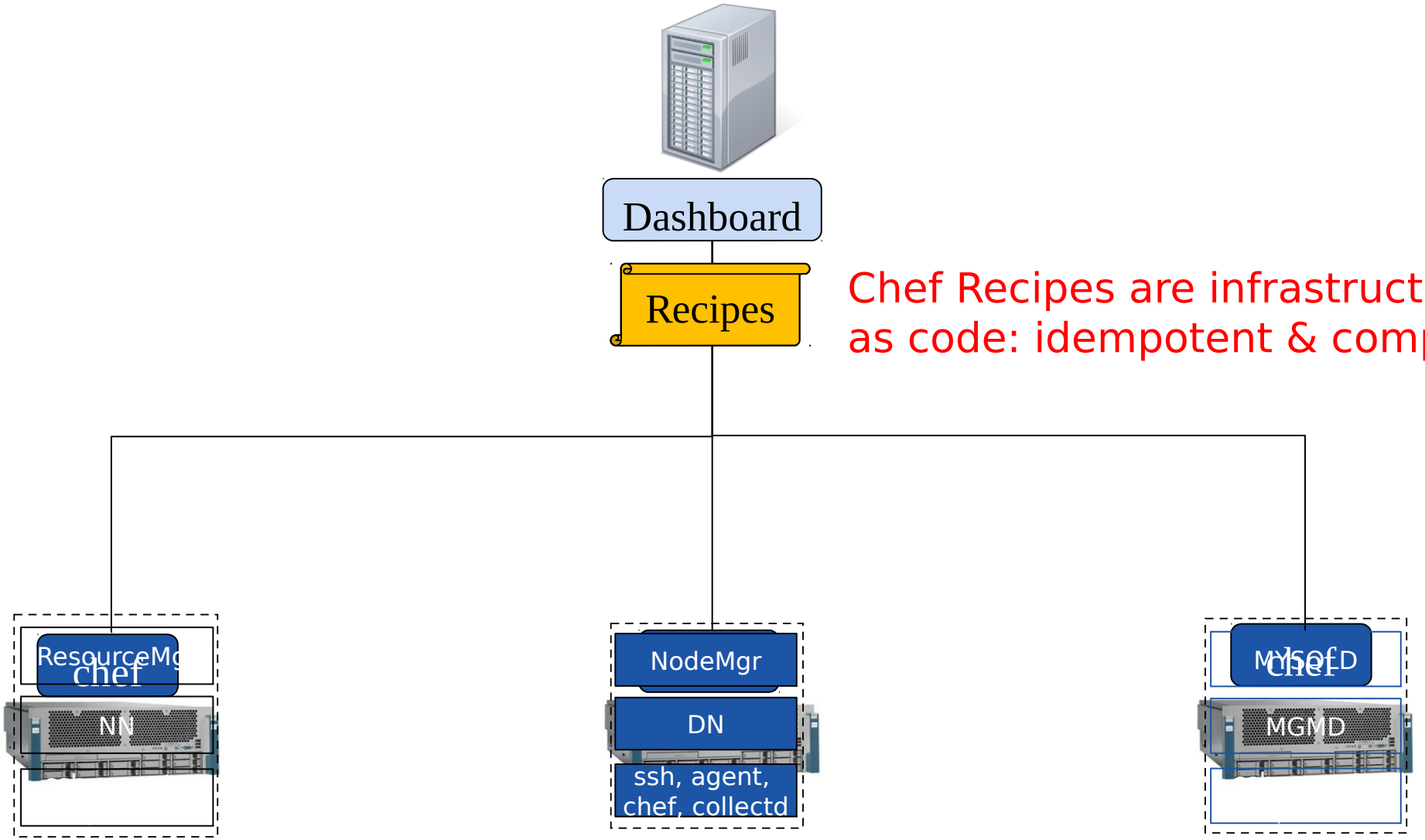


Configured stack of servers, dependencies, and firewalls with installed apps.

www.biobankcloud.eu

Running on lots of machines...

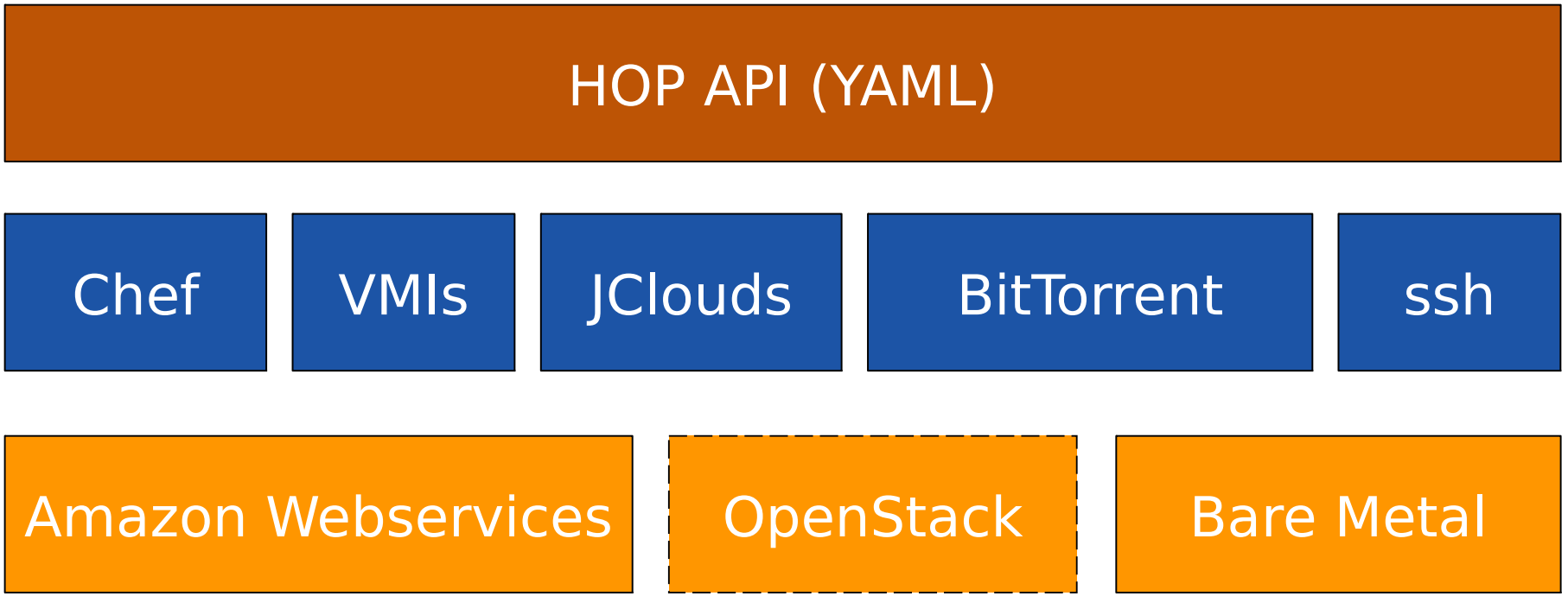
Automate Installation using Chef



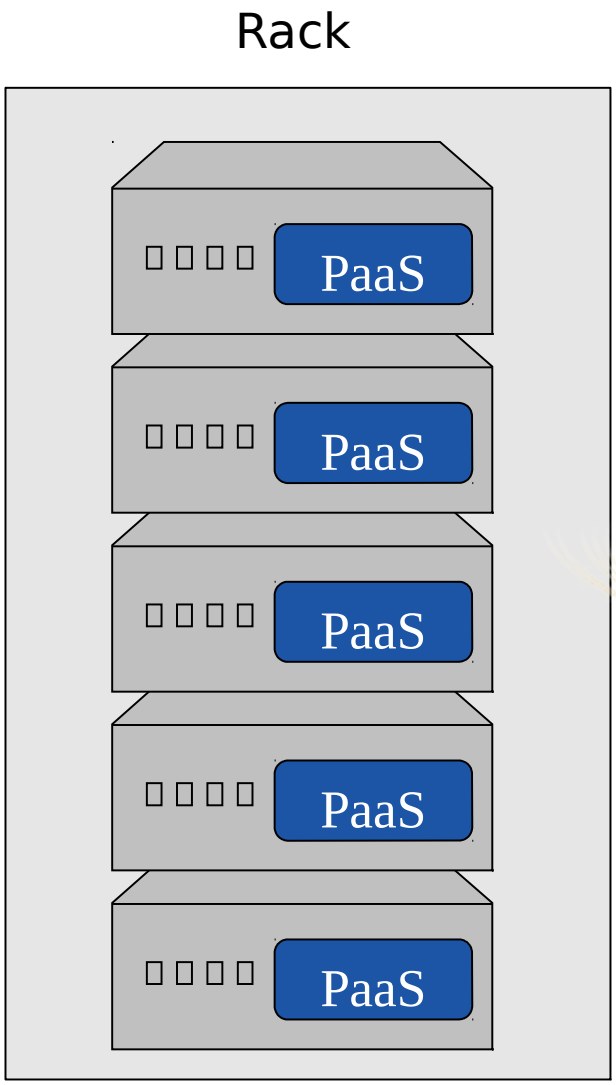
Described in a YAML file

```
name: biobankCloud

provider:
    name: aws-ec2

nodes:
    - services: [ndb::dn, hop::nn]
    number: 2
    - services: [ndb::mgm, ndb::mysqld, hop:dash]
    number: 1
    - services: [hop::rm]
    number: 1
    - services: [hop::dn, hop::nm, spark, adam, avokado]
    number: 20
```

Hadoop Open Platform-as-a-Service (HOP)



BiobankCloud in the near future



BIOBANKCLOUD

Collaboration with Industry

Conclusions

- Big Data for Genomics is in an embryonic phase of development
- Hadoop is currently the dominant paradigm for Big Data
- Hadoop Open Platform-as-a-service
  - Scalable
  - Easy to install and manage
  - Support for security and biobanking on its way