Hortonworks

HDFS: Hadoop Distributed FS

Steve Loughran, Hortonworks
stevel@hortonworks.com
@steveloughran

ATLAS workshop, June 2013

What is a Filesystem?

- **Persistent store of data:**

  write, read, probe, delete

- **Metadata for organisation:**

  locate, change

- **A conceptual model for humans**
- **API for programmatic access to data & metadata**

Unix is the model & POSIX its API

Unix is the model & POSIX its API

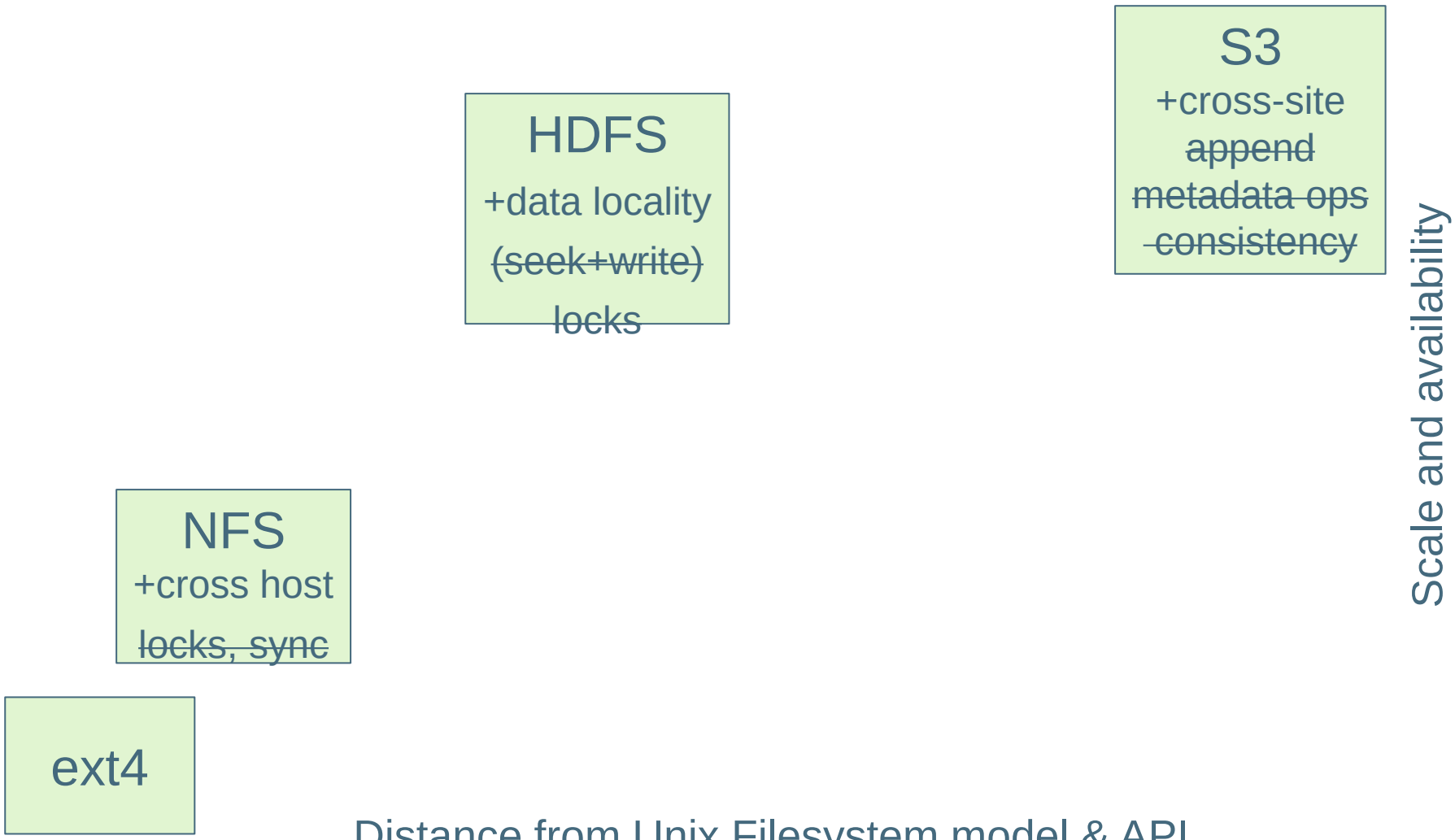
- **directories and files:**

  directories have children, files have data

- **API: open, read, seek, write, stat, rename, unlink, flock**
- **Consistency: all sync()'d changes are globally visible**
- **Atomic metadata operations: mv, rm, mkdir**

Features are also constraints

Relax constraints ➔ scale and availability



HDFS: what

- **Java code on Linux, Unix, Windows**
- **Open Source: hadoop.apache.org**
- **Replication rather than RAID**
  - –break file into blocks
  - –store across servers and racks
  - –delivers bandwith and more locations for work
- **Background work handles failures**
  - –replication of under-replicated blocks
  - –rebalancing of unbalanced servers
  - –checksum verification of stored files

Location data for the Job Scheduler

HDFS: why?

- **Store Petabytes of web data: logs, web snapshots**
- **Keep per-node costs down to afford more nodes**
- **Commodity x86 servers, storage (SAS), GbE LAN**
- **Accept failure as a background noise**
- **Support computation in each server**

Written for location aware applications -MapReduce, Pregel/Giraph & others that can tolerate partial failures

Some of largest filesystems *ever*

An emergent software stack

HDFS: what next?

- **Exabytes in a single cluster**
- **Cross cluster, cross-site**
  what constraints can be relaxed here?
- **Data Provenance, tainting**
- **Evolving application needs.**
- **Power budgets**

HDD ➔ HDD+ SSD ➔ SSD

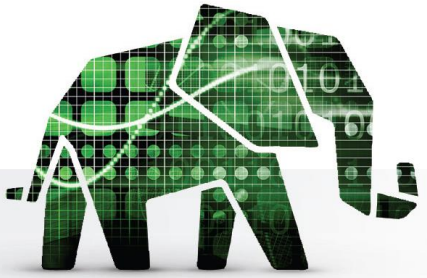
- **New solid state storage technologies emerging**
- **When will HDDs go away?**
- **How to take advantage of mixed storage**
- **SSD retains the HDD metaphor, hides the details (access bus, wear levelling)**

We need to give the OS and DFS control of the storage, work with the application

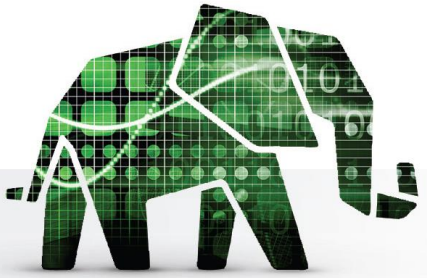
Download and Play!

http://hadoop.apache.org

http://hortonworks.com

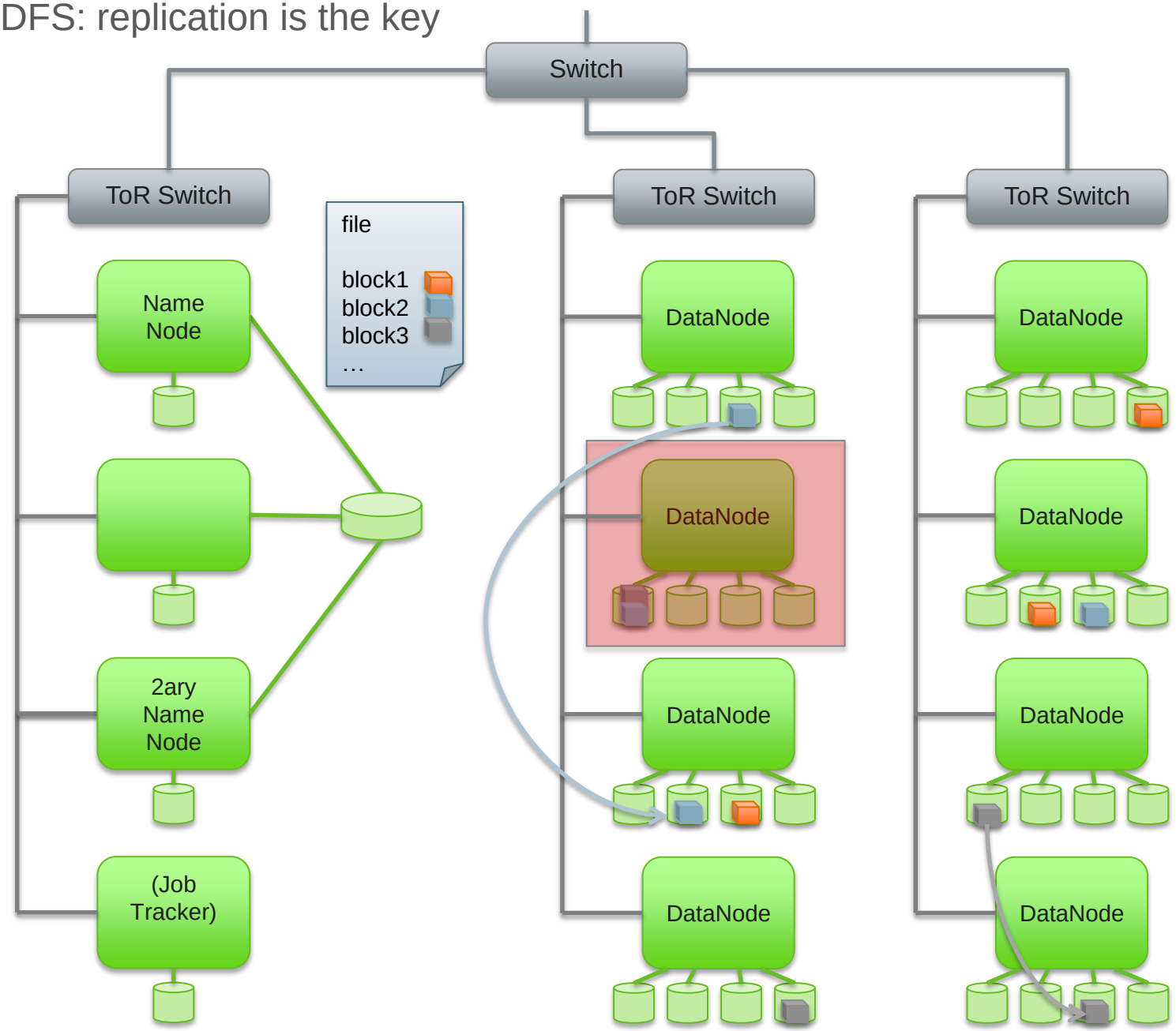


P.S: we are hiring



http://hortonworks.com/careers/

Hadoop HDFS: replication is the key



Replication handles data integrity

- CRC32 checksum per 512 bytes
- **Verified across datanodes on write**
- **Verified on all reads**
- **Background verification of all blocks (~weekly)**
- **Corrupt blocks re-replicated**
- **All replicas corrupt ➔ operations team intervention**

2009: Yahoo! lost 19 out of 329M blocks on 20K servers –bugs now fixed

Rack/Switch failure