



TECHNISCHE
UNIVERSITÄT
DRESDEN

Fakultät Informatik, Institut Systemarchitektur, Professur Systems Engineering

Crawling the DNS

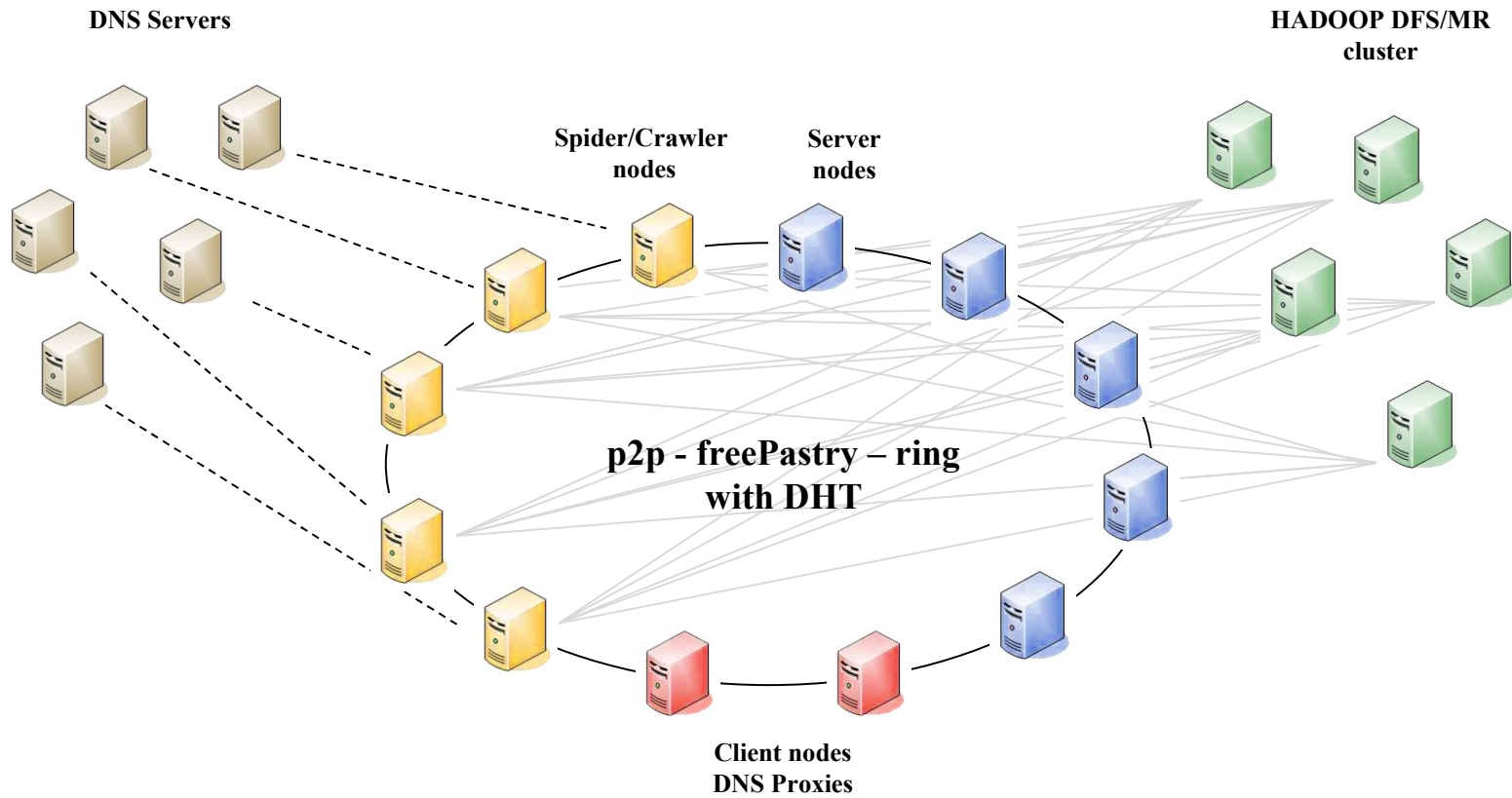
Experiences with Hadoop

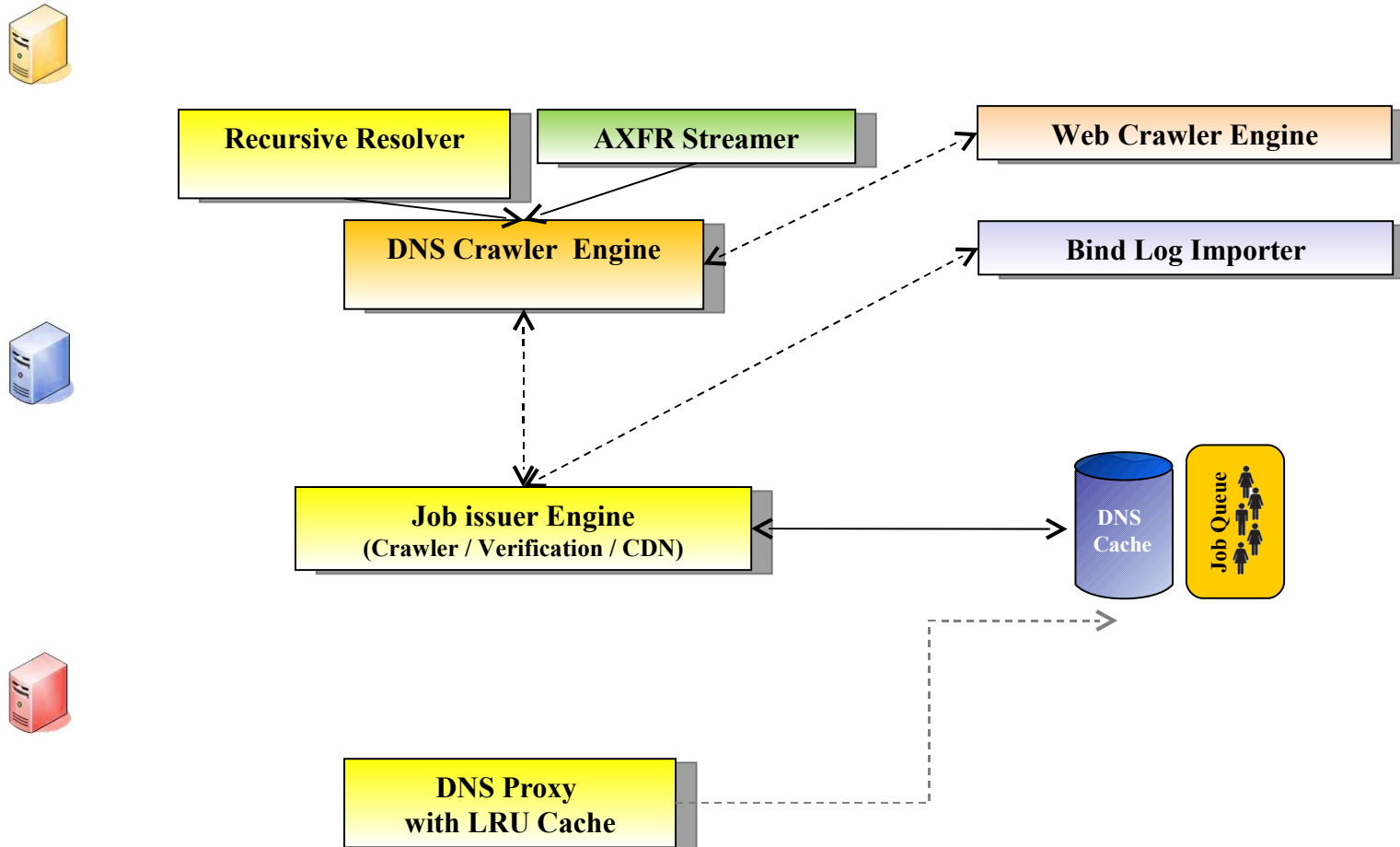
Gert Pfeifer

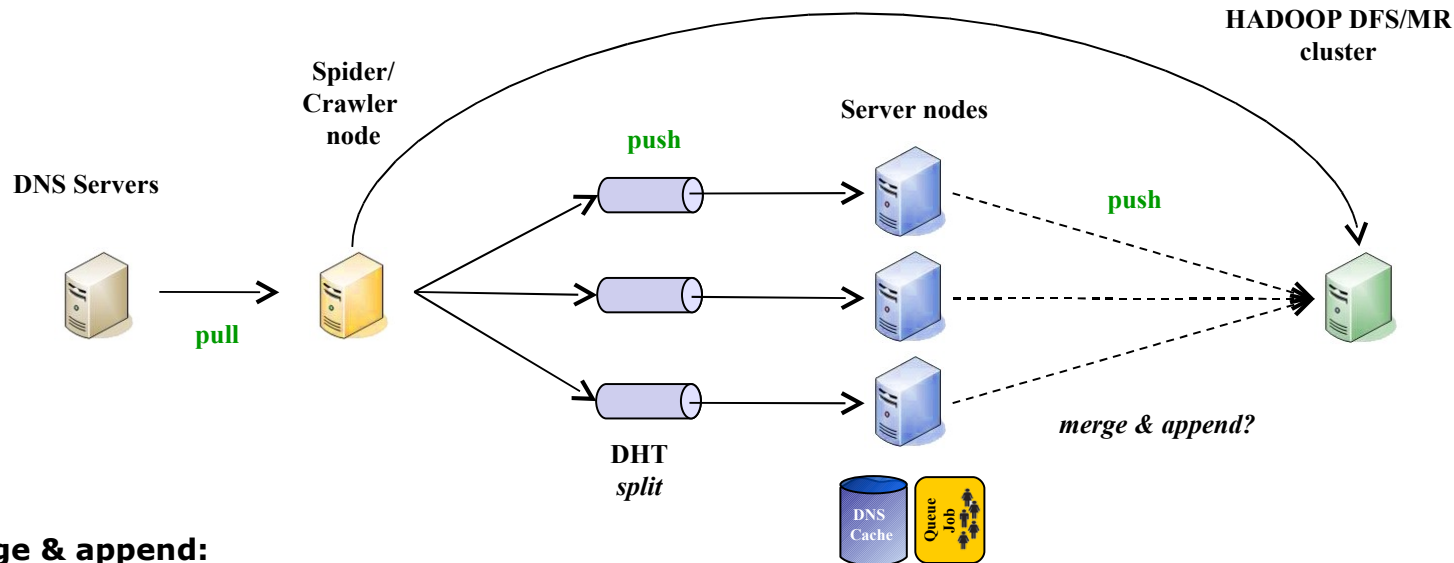
Dresden, 24.06.2008

Motivation

- we want to store DNS data on a central data base
 - Why? ... many reasons
- tried before with MySQL
 - failed
 - scheme of the DB was never sufficient for new kinds of data mining
 - with each new idea a new index was needed
- now storing raw data with Hadoop file system
- using Map/Red for data mining
 - easy to use, quite scalable







Merge & append:

- Hadoop Map/Red is efficient with big files
- Spiders often deliver small amounts of data
- Files can be written only once on Hadoop FS
- needs additional consolidation

Typical case:

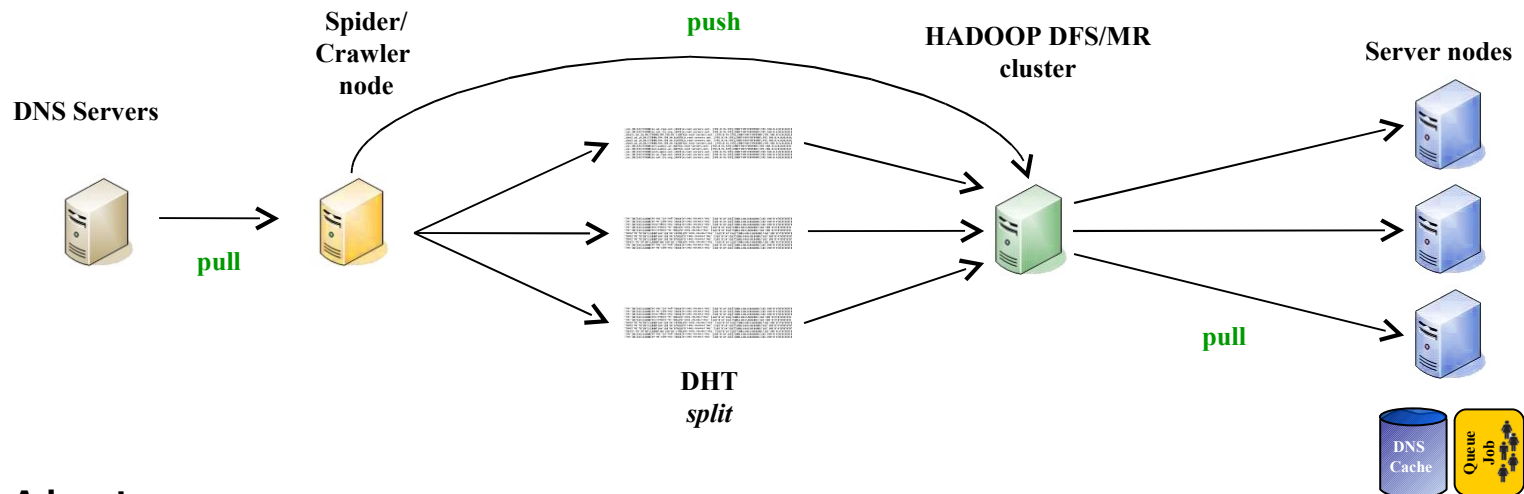
```
java.io.IOException: Could not get block locations. Aborting...
```

```
at org.apache.hadoop.dfs.DFSClient$DFSOutputStream.processDatanodeError(DFSClient.java:1832)
```

```
at org.apache.hadoop.dfs.DFSClient$DFSOutputStream.access$1100(DFSClient.java:1487)
```

```
at org.apache.hadoop.dfs.DFSClient$DFSOutputStream$DataStreamer.run(DFSClient.java:1579)
```

- No APPEND available for Hadoop (JIRA ISSUE HADOOP-1700)



Advantages

- DFS as "buffer" for low bandwidth PL nodes
- Once successfully saved on DFS, the data is "save"!

Disadvantages

- DFS unable to catch up with deletion of tmp files (only 100 blocks per node / 3 sec)
- ⇒ 8 server nodes x 2 (jobs list + split) x 50 spiders x 5 crawler threads x 3 replicates = 12,000 files/sec (worst case)
- ⇒ Would work with a 360 (data-)nodes DFS cluster if job processing length = 1 sec

- Hadoop DFS/MapRed cluster
 - 5 storage nodes - running 5 Hadoop DFS datanodes and tasktrackers
 - each equipped w/ 400 GB HDD (formerly 50 GB)
 - 8 additional (temporary) storage nodes - running 8 Hadoop DFS datanodes and tasktrackers
 - **Dell PowerEdge SC1435** - running Hadoop's namenode & jobtracker
- Server nodes
 - 4 instances running on a **Dell PowerEdges 1950** - 16 GB RAM
- Spider nodes
 - ~50-70 nodes with various hardware configuration // planet-lab.org
 - only 128 MB for JVM usable



Future work

- We keep crawling the DNS
- Try to come up with a streaming system to do it
- Replace hadoop after thorough comparison
- Advantages of streaming solutions are:
 - In-stream processing -> low latency
 - High level query language
 - Scalability
 - Allows immediate results for data mining applications
 - Deals with infinite sequence of events