UNIVERSITY OF BIRMINGHAM

Birmingham's Experience with EOS

GridPP Meeting, 10th April, 2018

Mark Slater, Birmingham University

Disclaimer!

I'm not an expert!

This talk will present my experiences of setting up EOS from a general sys-admin point of view. In particular:

- **I almost certainly haven't got the most efficient setup**
- **I was focusing exclusively on functionality rather than performance**
- **I may have misunderstood some aspects**
- **Even the knowledge I have is patchy**

I would very much recommend going to http://eos.web.cern.ch for more information!

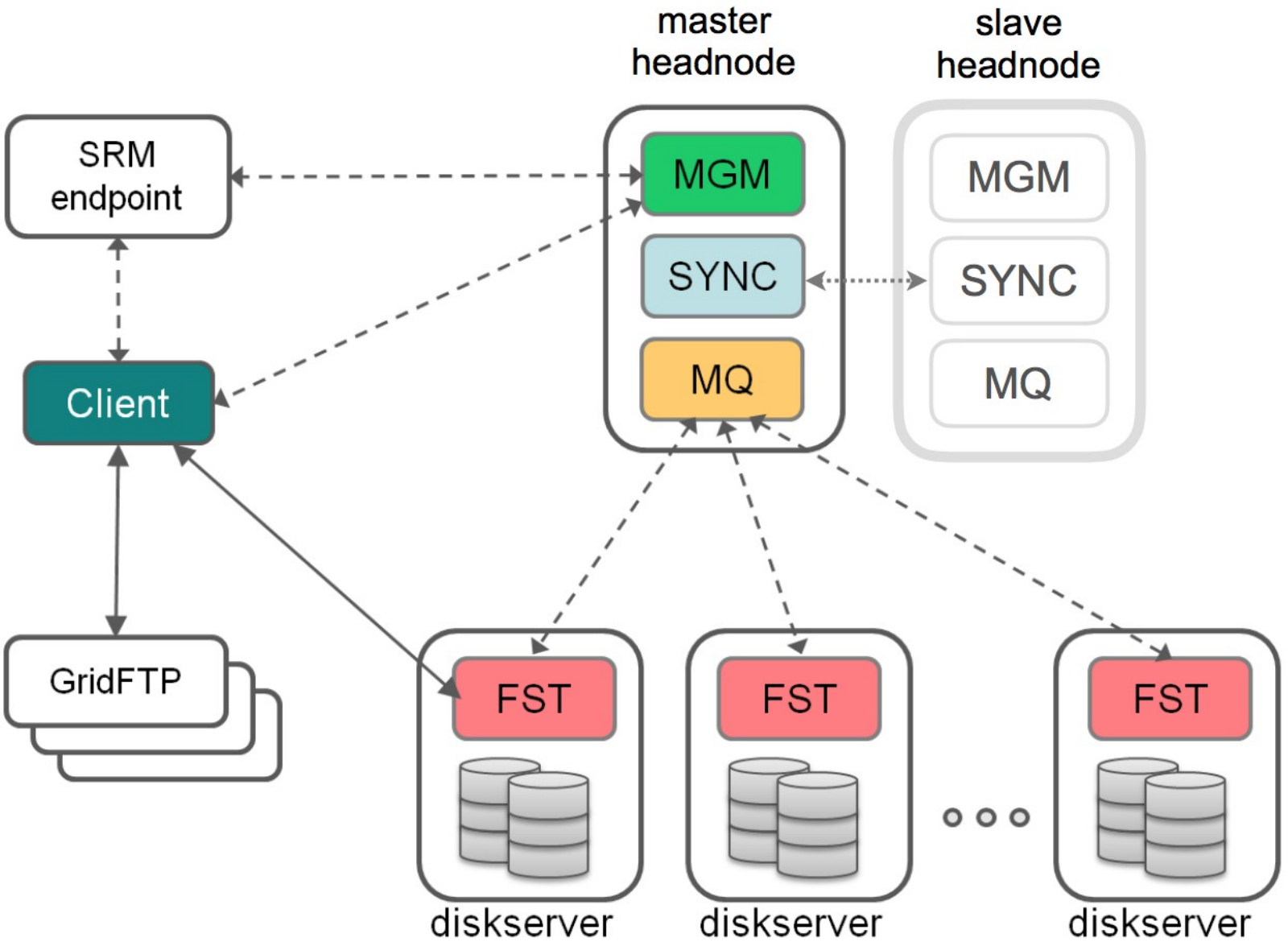
What is EOS?

Though I'm sure everyone has at least a rough idea of what EOS is, I'll start with the PR spin:

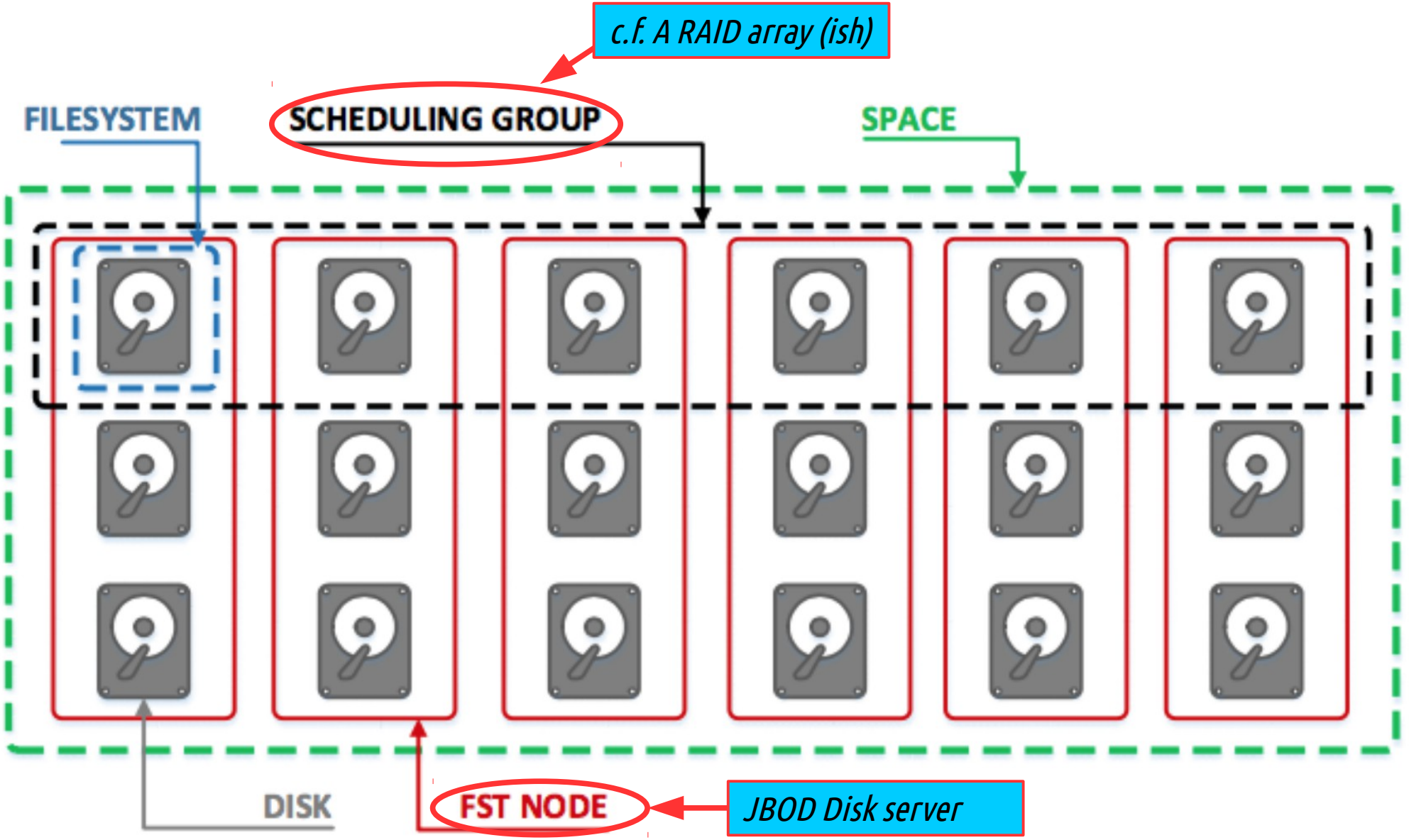
- EOS is an Open Source distributed storage solution built on XRootD
- Developed at CERN from 2010
- Wanted 'Organic' storage system: resilient, adaptive, self-healing (!)
- Based on disk-only JBODs
- Simple architecture with no relational DB or commercial dependencies
- Easily adaptable

For CERN's use case it's being used to manage 100s of PB of disk over 2(+?) sites and will be replacing AFS in the (near) future.

System Overview



Disk Organisation in EOS



Features

EOS boasts many desirable features:

- Low latency for operations due to in-memory namespace
- No need for hardware RAID as redundancy handled in software
- Unix user, GSI or Kerberos access
- Rich Access Control Lists
- User, Group and Project Quotas
- Server on Linux but clients available for Linux and OSX
- SAMBA and WebDAV access available

As regards the day-to-day administration, EOS provides many easy and powerful tools:

- Comprehensive CLI tools to control all aspects of the system
- Automatic balancing between and within groups
- Draining of FSs and auto-repair from broken disks
- Highly configurable redundancy/striping on a per file/directory basis

Installation & Configuration

There is a decent quickstart guide for installation and basic configuration here:
http://eos-docs.web.cern.ch/eos-docs/quickstart.html

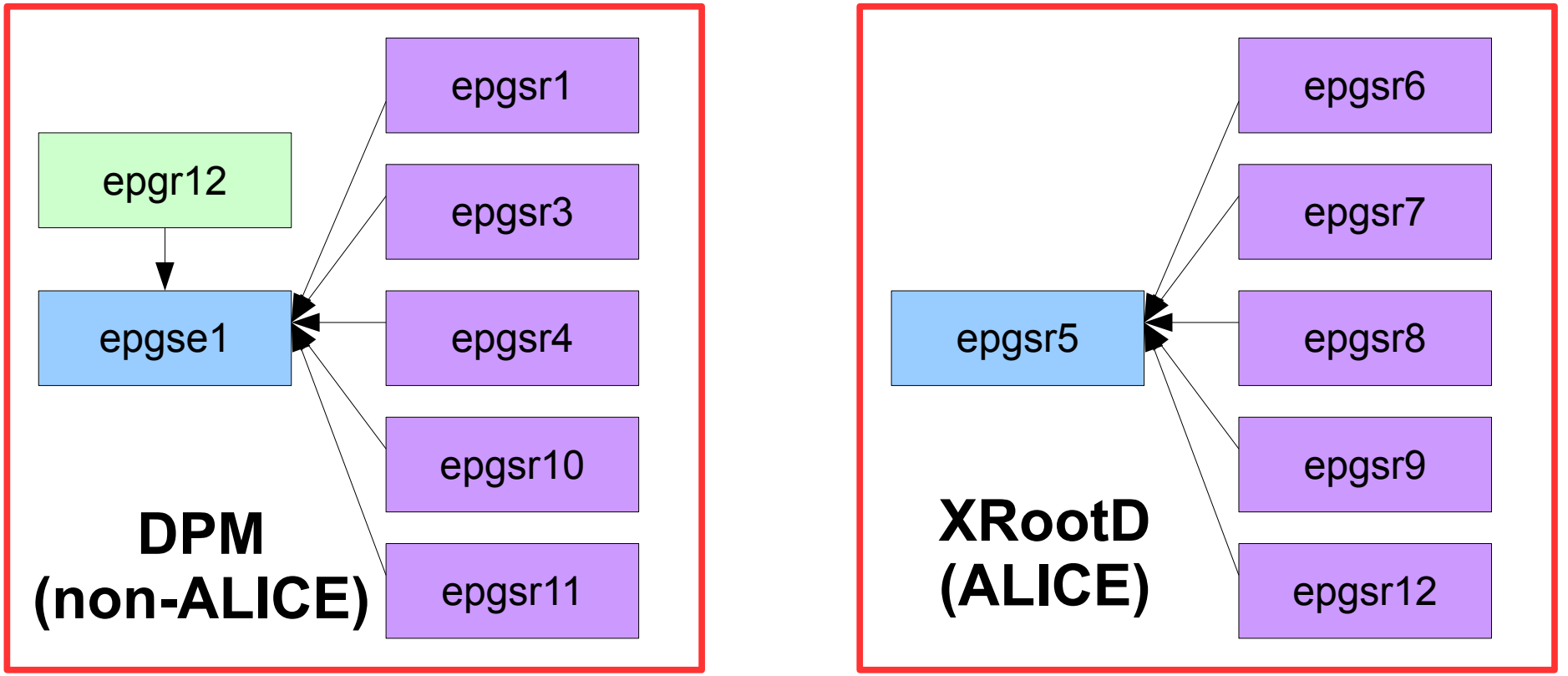
It comes down to setting up a yum repo, installing the appropriate packages and then applying a few config changes (out-of-the-box settings seem ~OK as well)

The important files that you care about are:

Location	Significance
/etc/xrd.cf.mgm	xrootd server configuration file
/etc/sysconfig/eos	instance configuration file
/var/eos/md/files.<hostname>.mdlog	changelog file cont. file meta data of the name space
/var/eos/md/directories.<hostname>.mdlog	changelog file cont. directory meta data of the name space
/var/eos/config/<hostname>	directory cont. configuration files and configuration changelog
/etc/eos.keytab	keytab file for 'sss' authentication (shared secret)
/var/log/eos/mgm/xrdlog.mgm	eos/xrootd MGM log file
/var/eos/md/so.mgm.dump	Dump of the MGM shared object hash/queues

These are what need backing up on the MGM

Bham Situation



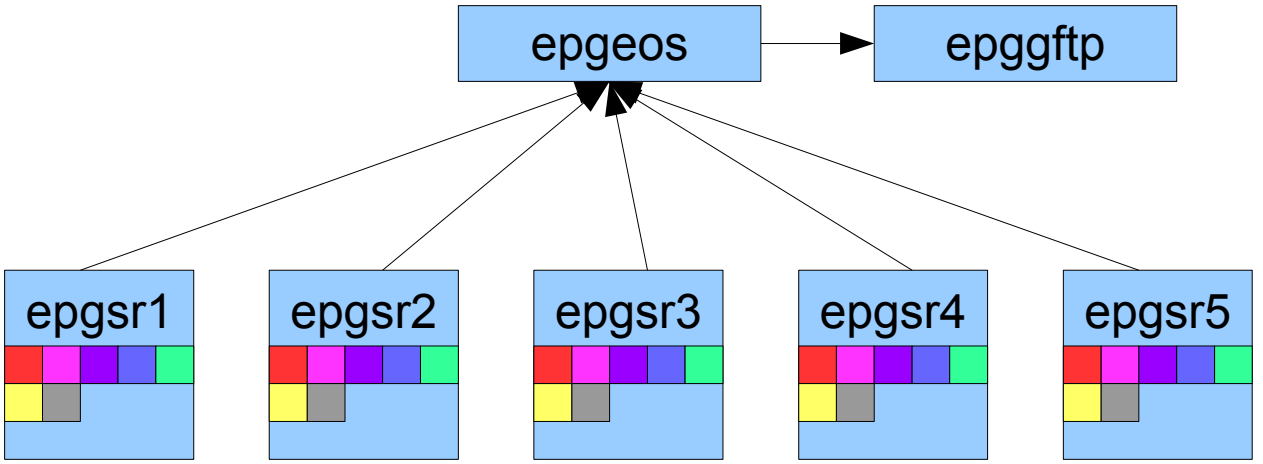
So why am I bothering with the move?

- Request from ALICE to try EOS over plain XRootD service
- Want to consolidate the two storage systems
- Would prefer something a bit easier to administrate than DPM
- Within <1 day, had a working EOS instance so was happy to continue

Disk Configuration

We have 16 disk servers available but they are of varying size so I've settled on a dual parity configuration with 8+2 stripes

This should allow us to be safe from a complete failure of (at least) a single server as well as up to two disks in the same sched group



Boxes are Hds, Colours represent different sched groups

Problems Encountered

Setting up EOS hasn't been without it's problems:

FST Conncetion Problems:

Found that FSTs would stay up for 1-2 hours under load and then connections would start backing up. Possibly solved in later releases, but for now set ….

Can only have one FS per node in a particular group:

You can't easily have two or more disks on a node in the same sched group. This can be got around by direct editing of the config but is rather hacky.

Documentation is a little lacking:

Though there is quite a bit of documentation available, it doesn't seem to spell everything out as plainly as I'd have liked. I spent sometime piecing together things from various sources.

Current Status and Plans

At present Bham has 240 TB on EOS with ALICE now using it exclusively

I plan to transfer another 400 TB from the old ALICE storage over to EOS in the next few weeks

After some investigation, I *think* I have figured out how to get Atlas to use it and (with help from the Atlas UK Computing team!) will hopefully transfer that over in the next couple of months

Finally, I will decommission our DPM instance, tentatively by Autumn this year.