Catching rogue jobs before they overload the storage: the importance of I/O profiling

Dr Rosemary Francis, CEO and Co-founder

Who we are

Ellexus is the I/O profiling company.

We work in HPC and wide-scale computing. Our latest tool Mistral has been developed with ARM.

This case study is taken from a real customer deployment using LSF and Spectrum Scale.
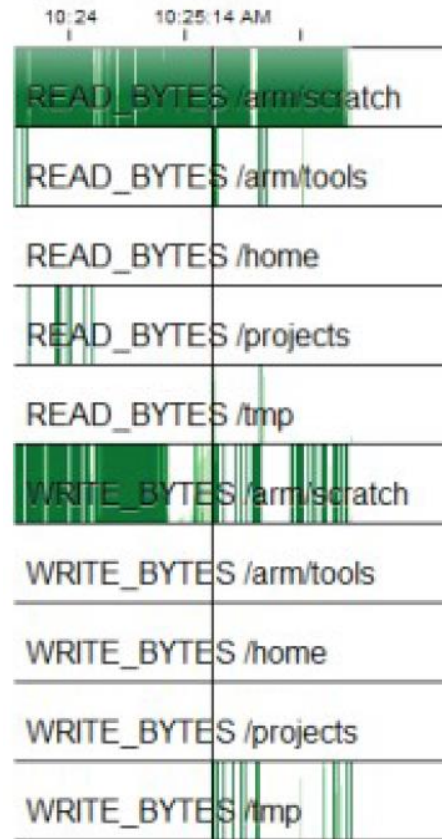
Customers include:

Mistral:
Solving the noisy neighbour problem

In a compute cluster it is possible for a small number of jobs to overload the shared file system. **Example of a rogue job from ARM:**

Lots of data is written to <u>remote storage</u> (/scratch).

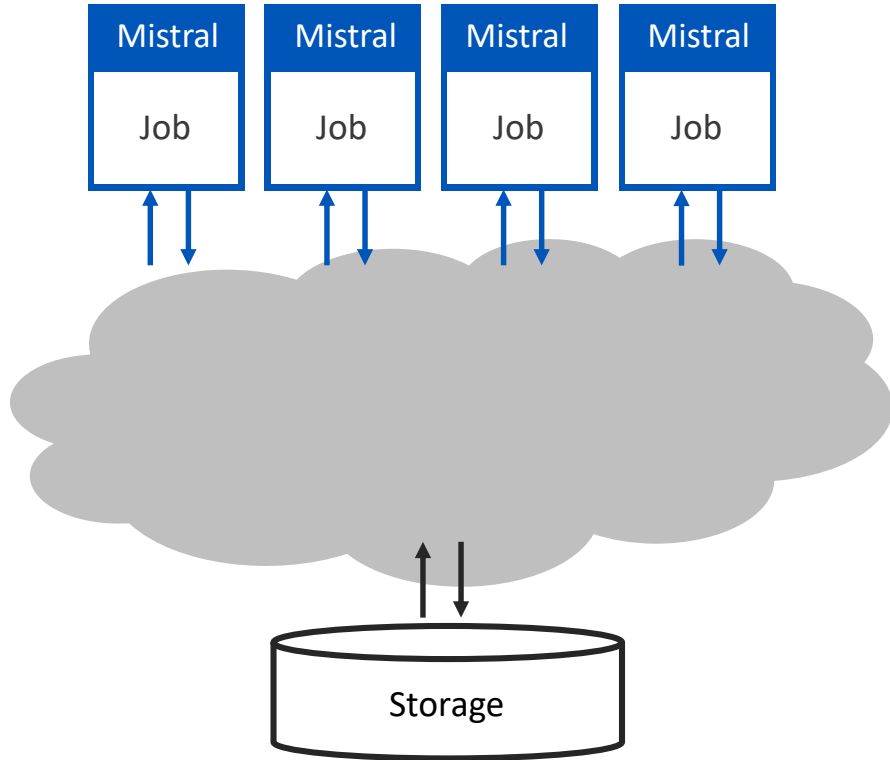But almost nothing is written to <u>local storage</u> (/tmp).

I/O Profiling goals

Protect your storage from bad I/O patterns and rogue jobs.

Load balance your storage for good quality of service.

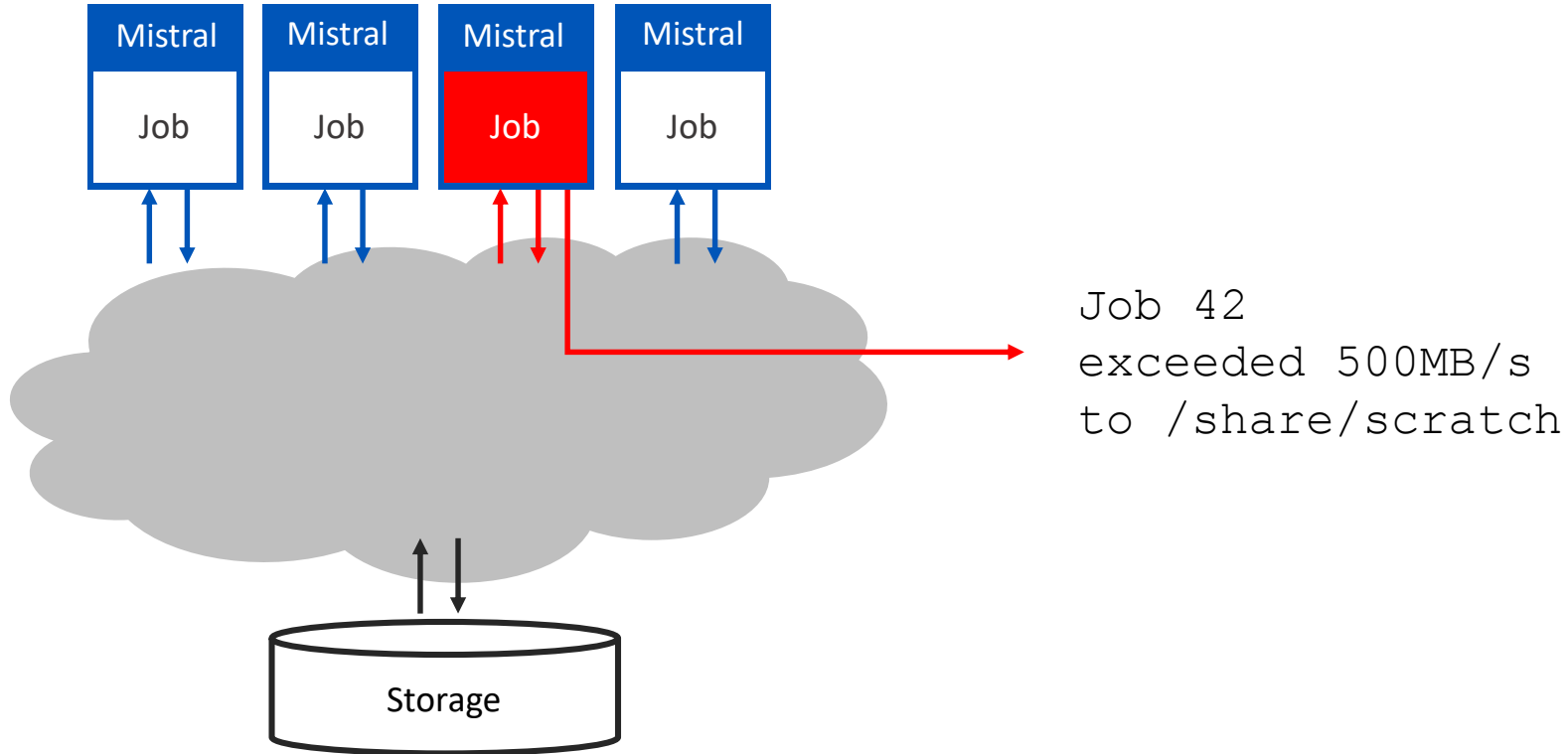**Optimise** your applications for the storage and your infrastructure for your jobs.

Mistral:
How does it work?

Mistral	Mistral	Mistral	Mistral
Job	Job	Job	Job

Storage

Mistral sits at the application level.

Mistral:
How does it work?

Job 42
exceeded 500MB/s
to /share/scratch

Catching rogue jobs before they overload shared storage
www.ellexus.com

Customer case study: Catching rogue jobs

… stories from a Silicon Valley software 'startup'

Customer case study: system architecture

Host

Leaf   Leaf   Leaf   Leaf
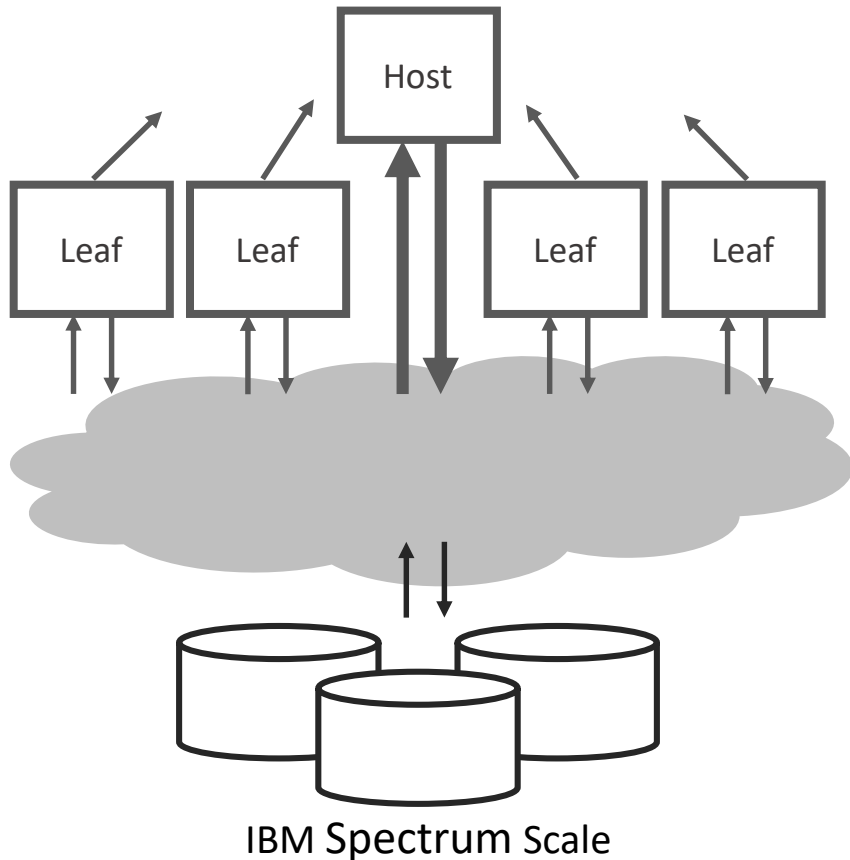
IBM Spectrum Scale

Host-Leaf architecture

Host compute nodes:
- Native Spectrum Scale
- 4 Gigabit/s Infiniband
- 32 Slots

Leaf compute nodes:
- Send data over RCP to host nodes
- Clustered NFS

Storage:
- 8 Spectrum Scale servers

Customer case study
Problem 1: Bad working practices
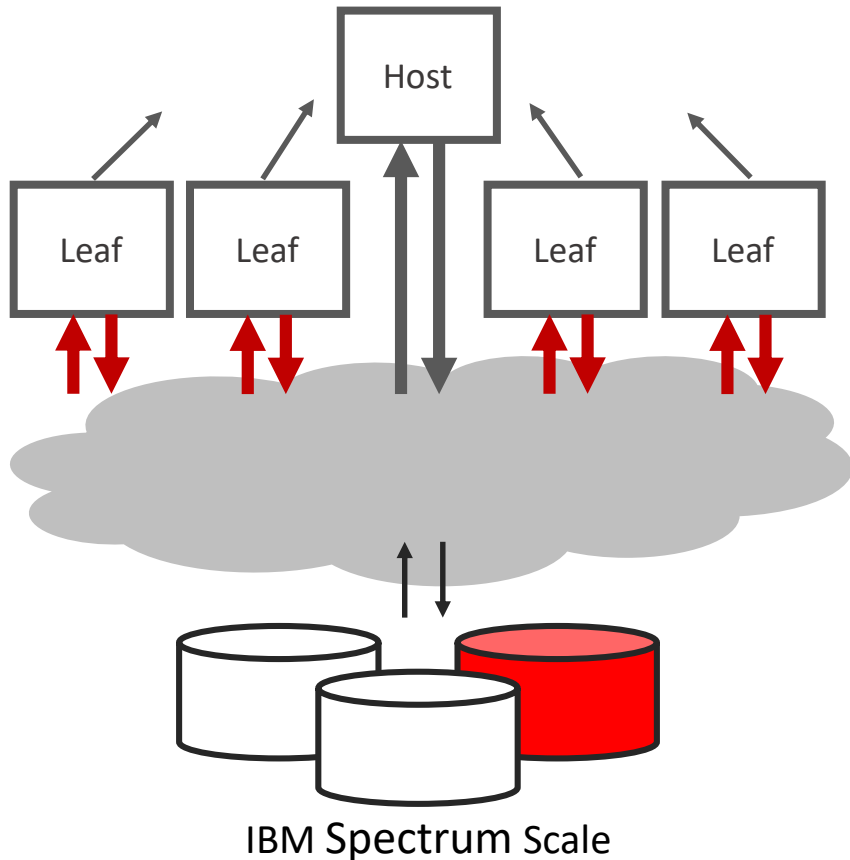
Jobs sometimes write from the leaf nodes instead of from the host.

If they hammer a particular mount point that can overload one of the filers.

IBM Spectrum Scale

Customer case study
Problem 2: Debug flag left on
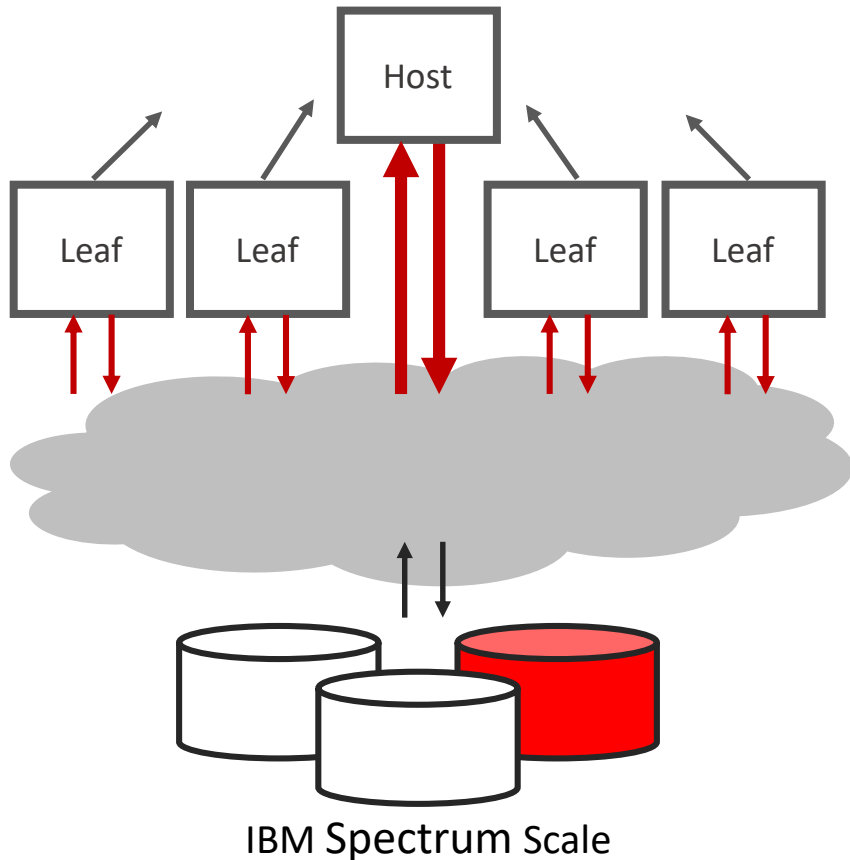
Job is debugged on one core then handed over to run at scale
– debug flag left on accidentally.

This happens weekly.

Host

Leaf   Leaf   Leaf   Leaf

IBM Spectrum Scale

Phase 1: Mistral deployment

IBM Spectrum Scale
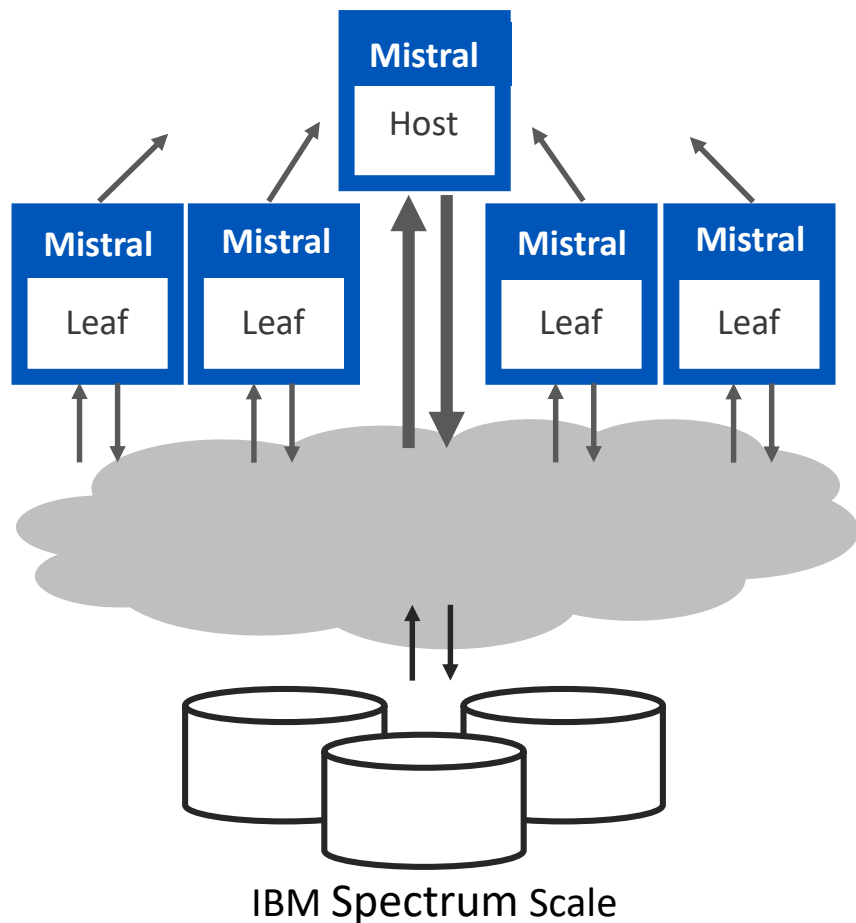
Production jobs are wrapped in Mistral via an LSF job starter.

Mistral generates an alert if:
- Leaf nodes write to the file system
- Host nodes write too much data

Two use cases:
- My house is on fire
- My house was on fire, but I missed it

Integration with RTM makes this possible.

Example contracts

Example contract:

```
monitoringtimeframe,60s
NAME     PATH/MOUNT POINT     IO TYPE          LIMIT
M1,      /gpfs/,              read,            500MB
M2,      /gpfs/,              write            500MB
M3,      /gpfs/,              create+delete,   2000
```

Example alert:

TIME	RULE	VIOLATION	PID	PROGRAM	FILE	JOB ID
17:30:05	M1	643MB/60s,	2784,	/tools/bin,	/gpfs/tools/config,	12
17:30:05	M2	505MB/60s,	2784,	/tools/bin,	/gpfs/data/file01,	12
17:30:06	M3	7405/60s,	2784,	/tools/bin,	/gpfs/data/set05,	12

Mistral RTM integration: Who is using the storage?

Table showing:
- Cluster name
- Job ID
- Host Queue
- Project
- User
- Mistral I/O rule
- PID
- I/O observed (MB)
- Time frame
- Log time

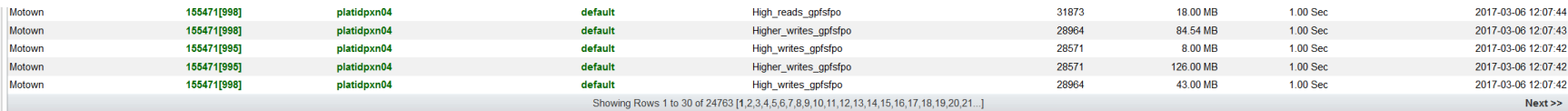Catching rogue jobs before they overload shared storage
www.ellexus.com

Mistral RTM integration: Who is using the storage?

After 12 months
(Almost) No more rogue jobs!

Catching rogue jobs before they overload shared storage
www.ellexus.com

Phase 2: Performance
Catching rogue jobs before they are run
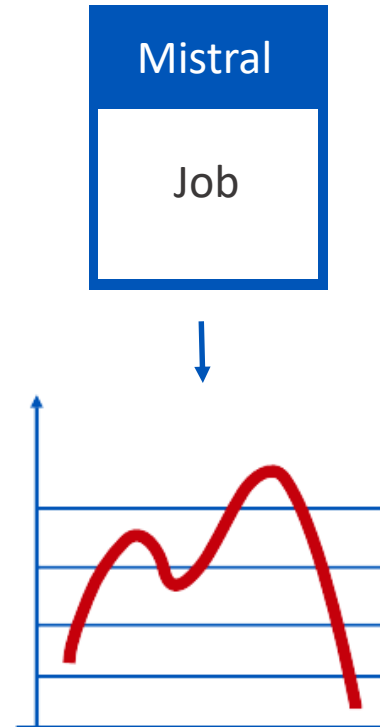
R&D jobs often seg fault and this exposes Mistral.
How to protect the storage without changing the R&D environment?

Solution

1. Guess which job is causing the problem the old way
2. Check problem jobs by profiling I/O in detail

Set up a special LSF queue with a detailed Mistral I/O profiling contract.

Make every user an I/O expert

Mistral
Job

Phase 2: I/O Profiling-as-a-Service

Introduce methodologies for finding and fixing I/O issues

Step 1: Qualify new jobs on test queue

Step 2: Run at scale

Step 3: Return to profiling queue if there is a problem

Mistral: IT and QA raise I/O tickets
Breeze: R&D to fix the problem.

Summary

Profile profile profile.

Protect your storage by monitoring the whole cluster for rogue jobs. Prevent bad I/O patterns by running health checks on new work flows.

Optimise applications for shared storage and optimise the compute and storage infrastructure for the jobs.

Thank you

Please get in touch for more information.

Dr Rosemary Francis
CEO and Co-founder

Ellexus Ltd
St John's Innovation Centre, Cowley Road, Cambridge CB4 0WS, UK
info@ellexus.com
01223 123456