Department of Computer Science

HPS

University of Reading

Tracking User-Perceived I/O Slowdown via Probing

Limitless Storage
Limitless Possibilities

https://hps.vi4io.org

Julian M. Kunkel, Eugen Betke

BoF: Analyzing Parallel I/O

2019-11-20

Motivation

University of Reading

- Performance of shared file system is load dependent
  - Also background activity may cause delays
- Difficult to judge: observed performance is slower/faster than normal
  - A subcomponent of a file system may be loaded (e.g., metadata)
  - Is it due to software updates/intermediate or permanent hardware issues?
- Users/staff may wonder for the cause of the experienced performance
  - “Is that caused by my application?” Can lead to support requests!
- Maybe a quantification of the file system load similar to `uptime` would help?

  Paper: Tracking User-Perceived I/O Slowdown via Probing (Julian Kunkel, Eugen Betke), In High Performance Computing: ISC High Performance 2019 International Workshops, Frankfurt/Main, Germany, June 20, 2019, Revised Selected Papers, Lecture Notes in Computer Science, Springer, HPC-IODC workshop, ISC HPC, Frankfurt, Germany

Probing Approach

University of Reading

- Many sites run periodic regression tests, e.g., nightly
  - Helps to identify performance regressions with updates
- Instead, we run a non-invasive benchmark (a probe) with a high frequency
  - Mimic the user-visible client behavior
  - Measuring latency for metadata and data operations
- Generate and analyze generated statistics
- Derive a slowdown factor (file system load)

Why not use server-sided information?

- Client perspective is different (involves network, too)
  - We need to compare standard values!
- Tracking response latencies for op type/size histograms would do
  - Vendors: integrate such a reporting (vendor neutral API!)

Performance Measurement

University of Reading

Preparation

- Data: Generate a large file (e.g., $>$ 4x main memory of the client)
- Metadata: Pre-create a large pool of small files (e.g., 100k+ files)

Benchmarks

- Repeat the execution of the two patterns every second
- DD: Read/Write a random 1 MB block
- MD-Workbench: stat, read, delete, write a single file per iteration
  - Allows regression testing, i.e., retain the number of files
  - *J. Kunkel, G. Markomanolis. Understanding Metadata Latency with MDWorkbench.*

Executed as Bash script or an integrated tool:
`https://github.com/joobog/io-probing`

Outline

University of Reading

Test Systems

University of Reading

- JASMIN, the data analysis facility of the UK
  - Precreation: 200k files, 200 GB data file
  - 60 days of data
  - Script runs exclusively on a node
- Archer, the UK national supercomputer service
  - Precreation: 200k files, 200 GB data file
  - 30 days of data
  - Script runs on a shared interactive node
- Mistral, the HPC system at the German Climate Computing Centre
  - Precreation: 100k files, 1.3 TB data file
  - 18 days of data
  - Tool runs on a shared interactive node

Understanding the Timeseries

University of Reading

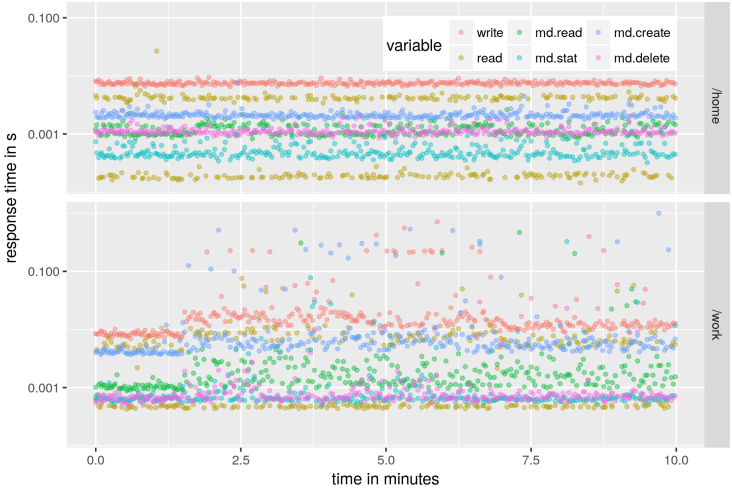

Figure: Jasmin every data point for 10 minutes of one node

- Every probe (1s) for 10 min
- For two file systems
- Home is stable
- Work shows irregularities

IO-500 Response Time on Archer

University of Reading

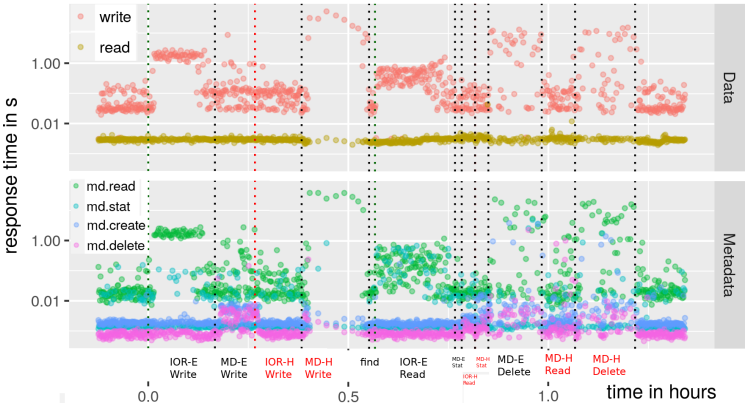

Figure: Response time (all measurements)

- Run on 100 nodes
  score 8.45
- The IO-500 various phases
  Data and metadata heavy
- First, all measurements

Validating Slowdown on All Measurements

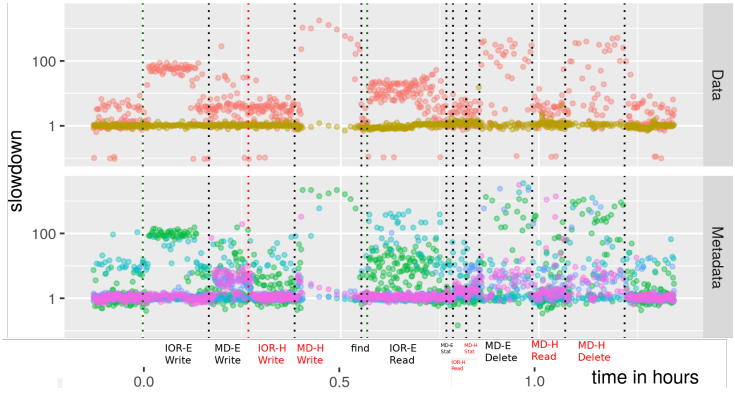

Figure: Slowdown (all measurements)

- Computed median slowdown
  Expected: median of 30 days
- Influence of phases is visible
- MDHard 1000x slowdown
  Influences data latency!
  10s of seconds latency
- IOREasy 100x slowdown
- IORHard not too much
- Data read is stable

Validating Slowdown: Reduced Data

University of Reading

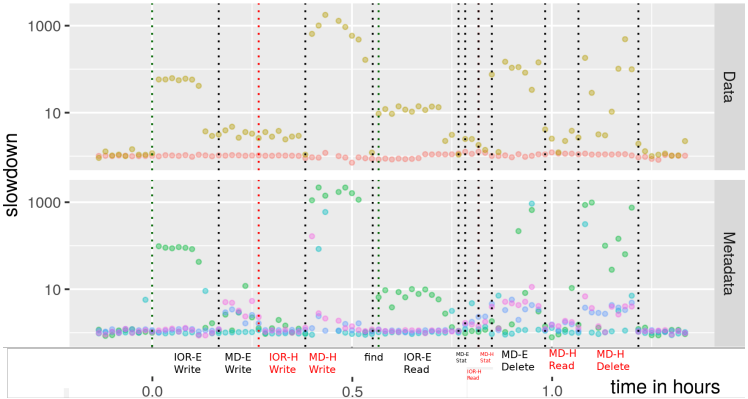

Figure: Slowdown (60s mean statistics)

- Data reduction: 60s mean
- More robust, clearer to see

Timelines of 4h Statistics

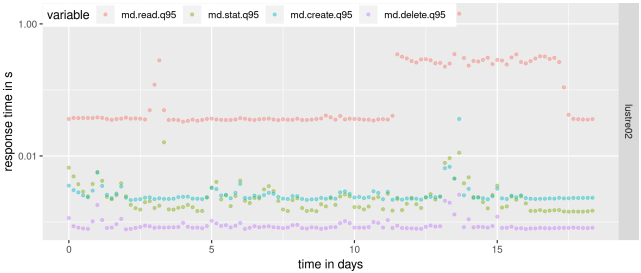

Figure: Mistral metadata timeline

- Use Q95, 5% ops are slower
- Change in behavior at day 12
  Reason: unknown

Slowdown for 4h Statistics

University of Reading

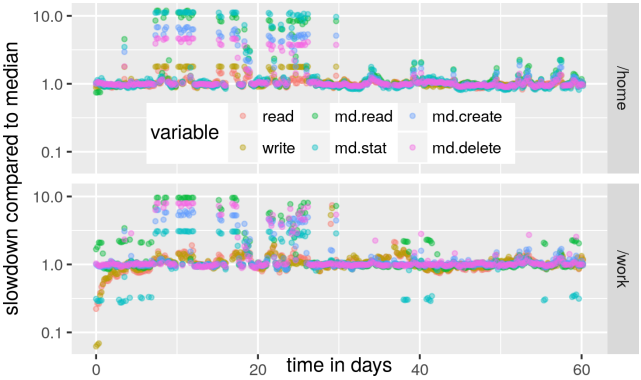

Figure: JASMIN, computed on 4 hour intervals

- Slowdown: Using the median
- Typically value is 1
- Sometimes 10x slower
- Values below 1, unusual (caching)
- Good to see long-term issues