OLCF

OAK RIDGE LEADERSHIP COMPUTING FACILITY

Testing methodology for large-scale file systems

Sarp Oral, PhD

Technology Integration Group

Lustre User Group Meeting, April 12, 2011

Large-scale FS testing

- Benchmarking and testing large-scale file and storage systems is not straight forward
  - Scale
  - Complexity
  - Limited toolset

- A black art
  - Tricks/tips/knowledge needs to passed down to new testers

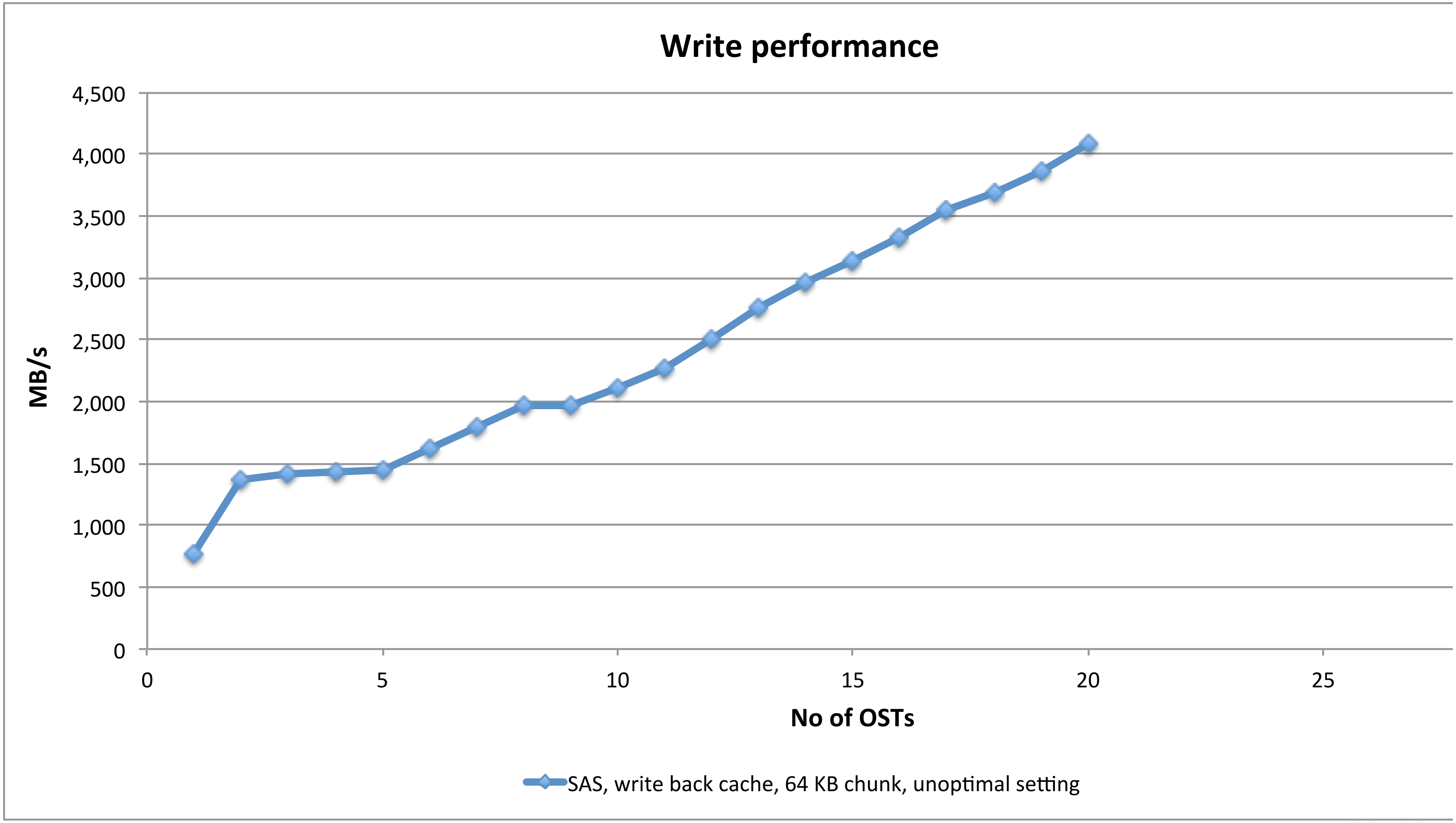
- Developing a complete and quantitative test methodology
  - Work in progress

Rules of thumb, #1

- Know thy hardware!
  - Identify all components of the disk backend system and how they are setup and configured
    - Setup/configuration directly impacts the expected/ observed performance

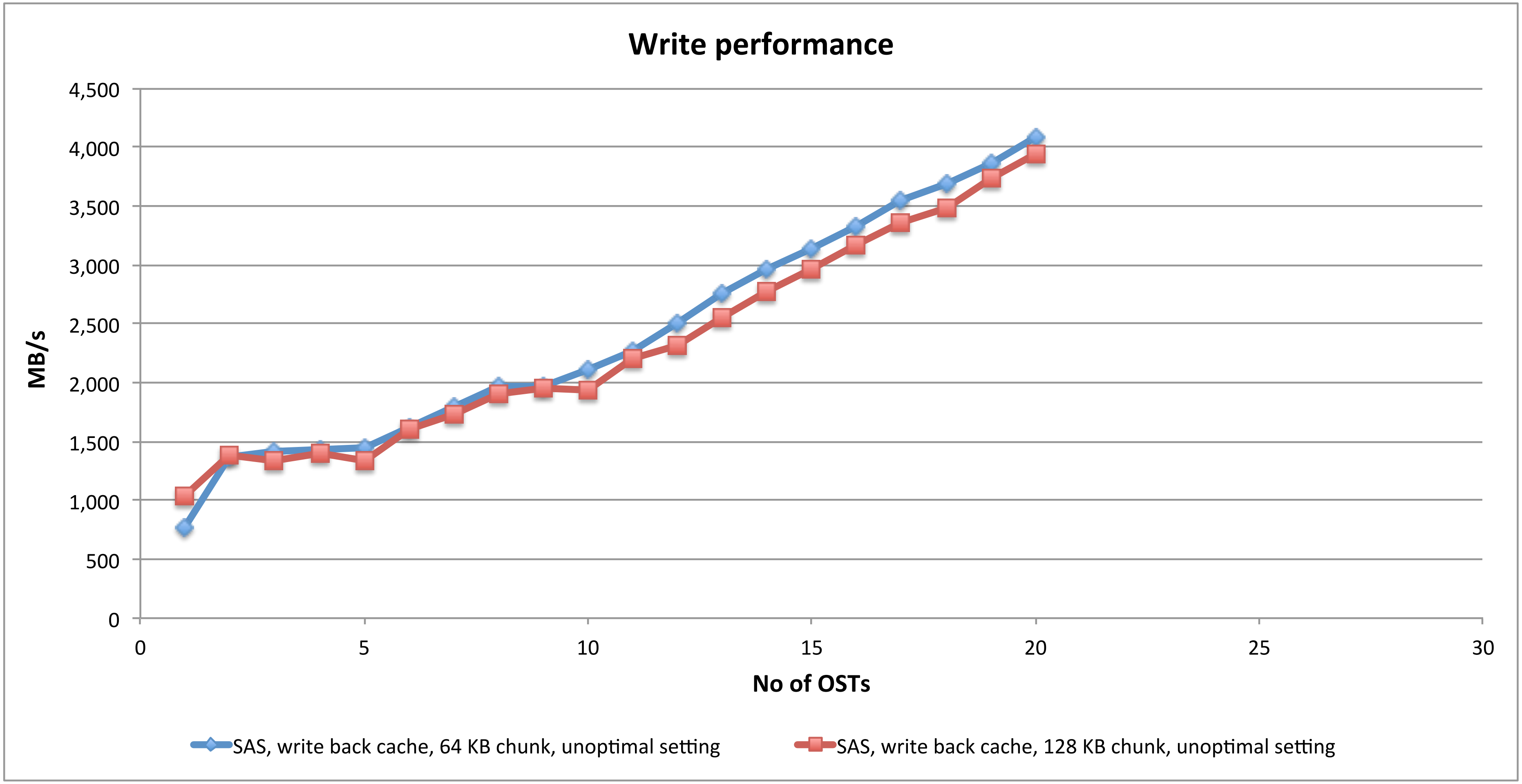
Understanding observed data

- 4 Hosts, QDR IB, 200 SAS disks, R6 (8+2), a pair of HW RAID controllers, obdfilter-survey



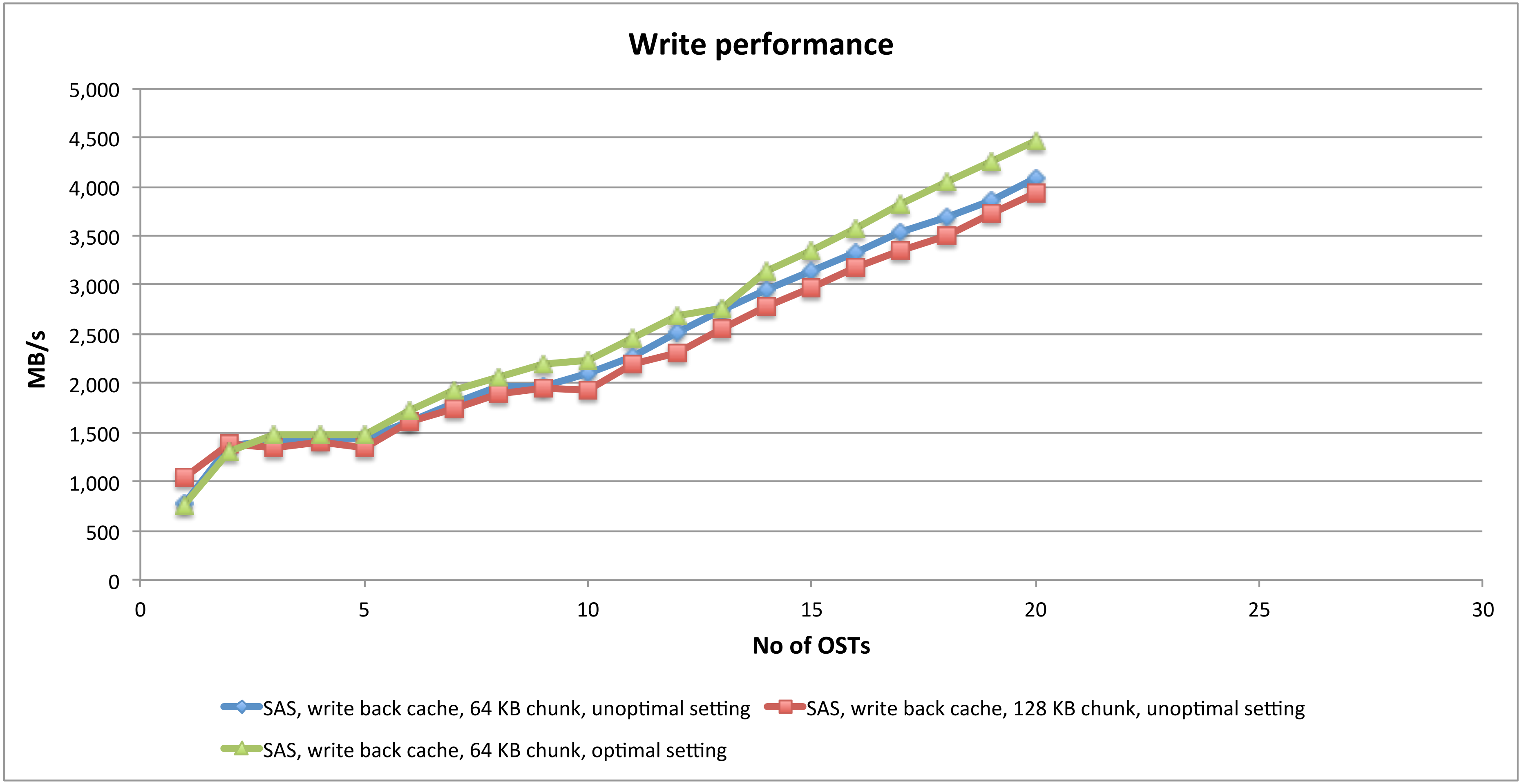
Understanding observed data

- 4 Hosts, QDR IB, 200 SAS disks, R6 (8+2), a pair of HW RAID controllers, obdfilter-survey



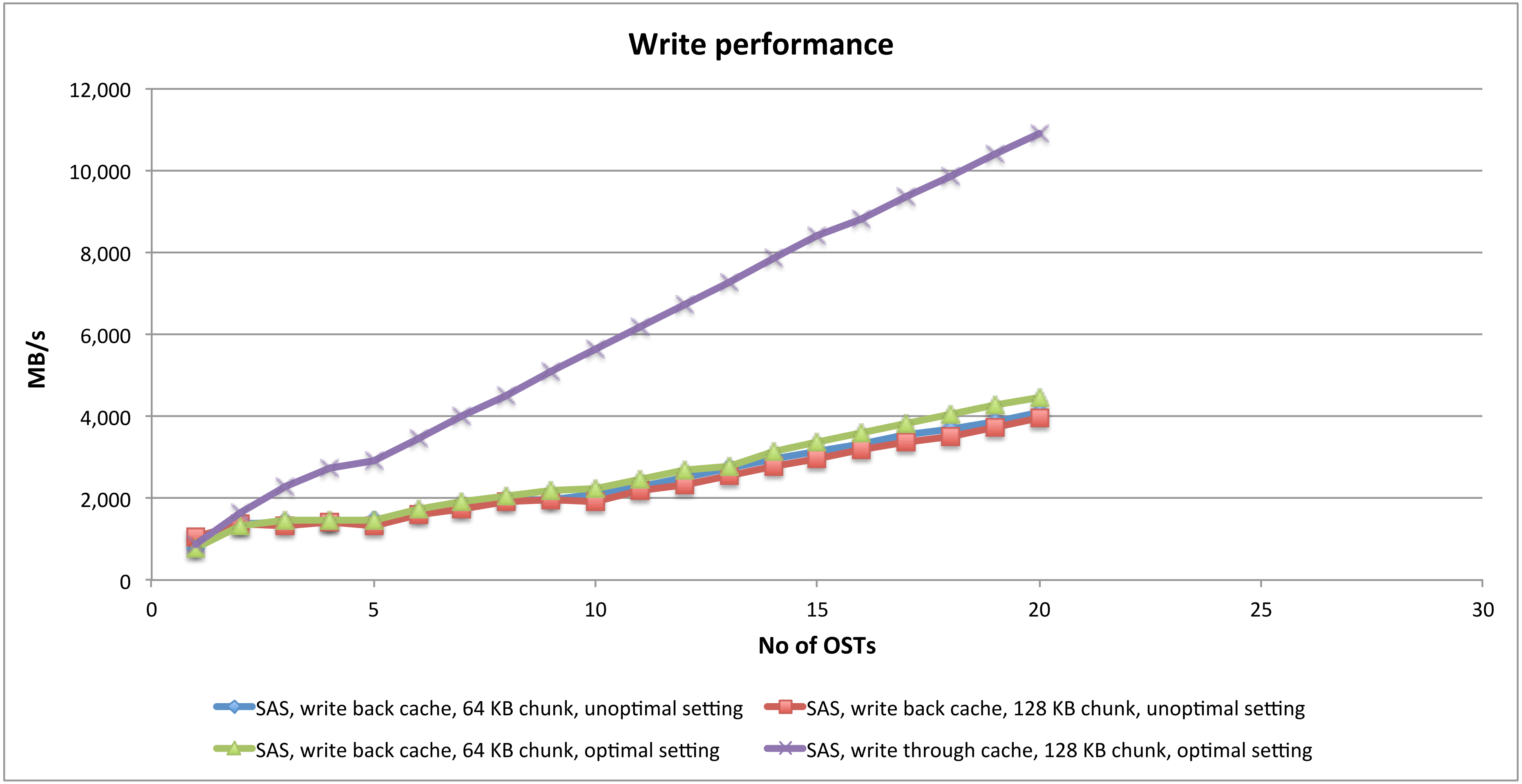
Understanding observed data

- 4 Hosts, QDR IB, 200 SAS disks, R6 (8+2), a pair of HW RAID controllers, obdfilter-survey



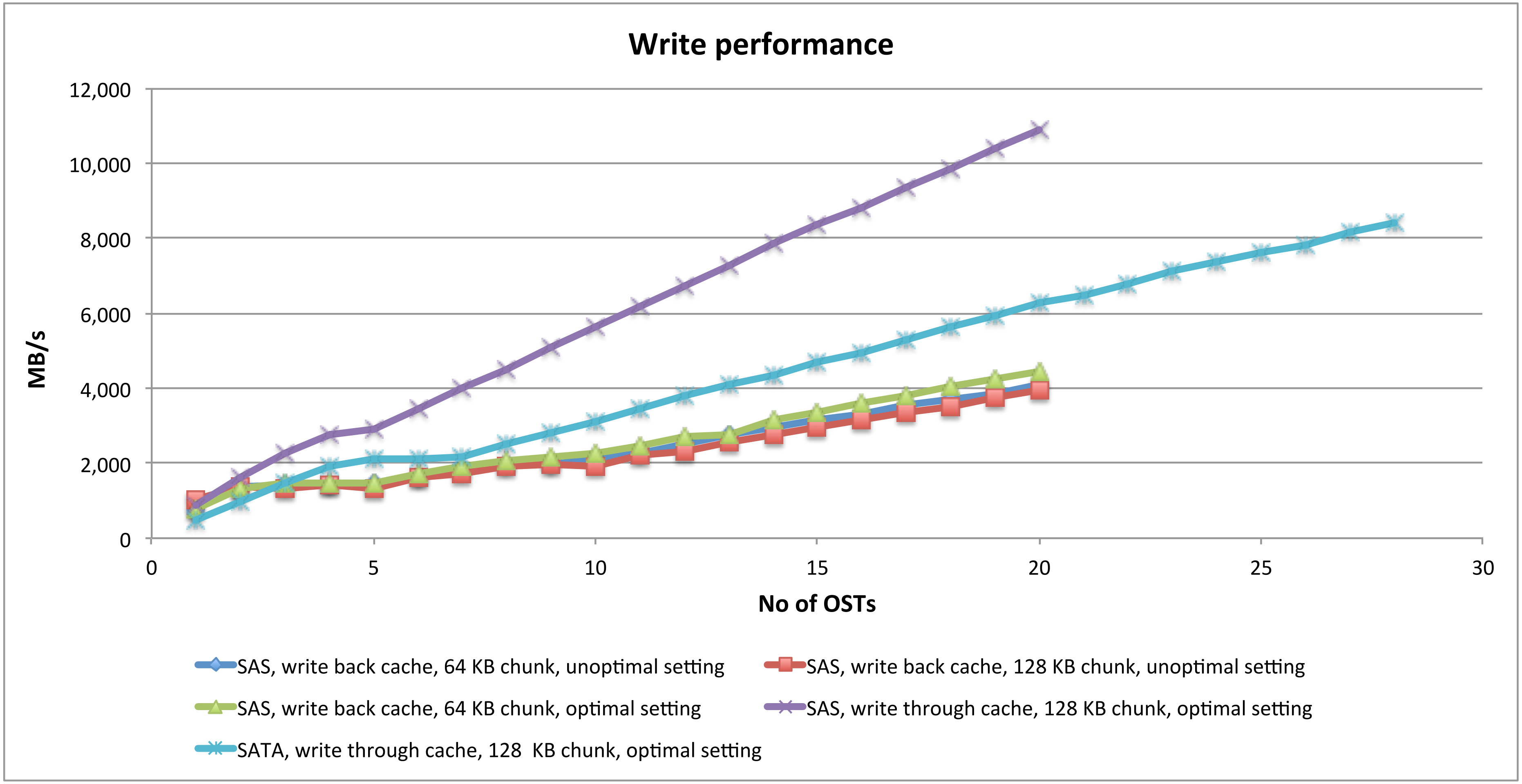
Understanding observed data

- 4 Hosts, QDR IB, 200 SAS disks, R6 (8+2), a pair of HW RAID controllers, obdfilter-survey



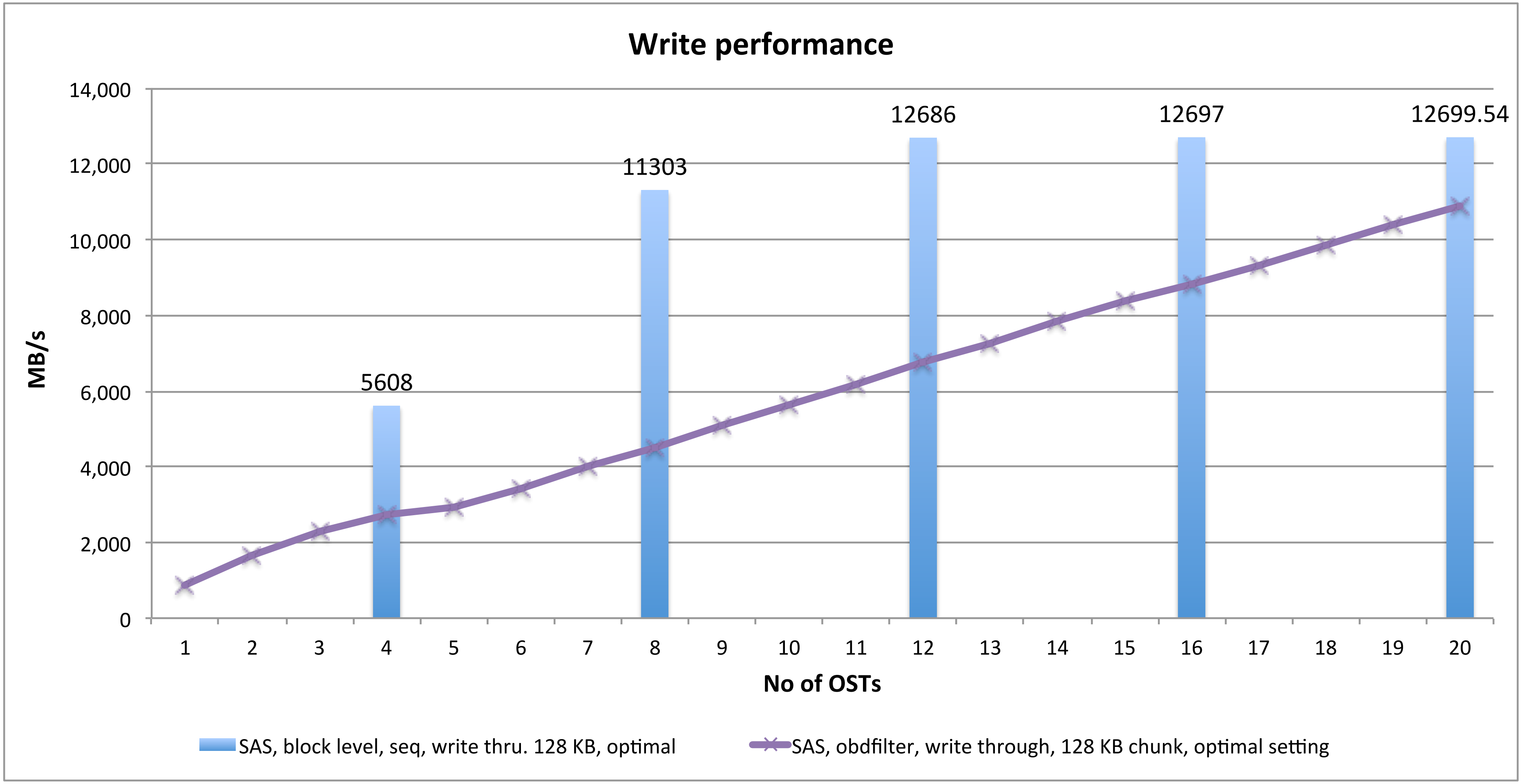
Understanding observed data

- 4 Hosts, QDR IB, a pair of HW RAID controllers, obdfilter-survey
  - 200 SAS disks, R6 (8+2)
  - 280 SATA disks, R6 (8+2)



Understanding observed data

- 4 Hosts, QDR IB, a pair of HW RAID controllers, 200 SAS disks, R6 (8+2)
  - Obdfilter-survey
  - In-house coded benchmark



Rules of thumb, #2

- Establish a clear understanding of the critical data path
  - From the disk backend host port to disk
  - Theoretical performance of all components on the critical data path
  - The system is a combination of serial and parallel (and combination of these two in some cases) connected set of devices
  - Lowest performing component will lower the overall performance

Rules of thumb, #3

- Understand how each component interacts with each other
  - A component's performance response may be different if it exercised differently
    - A RAID set might perform differently when there is background disk scrubbing going on
      - Contention on disks
    - A group of RAID sets may perform differently when only one of them is exercised compared to when all in the group exercised concurrently
      - Contention on RAID controller

Rules of thumb, #4

- Repeat the tests, verify the variability
  - An obtained performance may vary with time
    - Even under the same test conditions

Rules of thumb, #5

- Learn disk backend system's reporting and statistics mechanisms
  - Benchmarks will report the observed performance for a given test
  - Disk backend systems ***should*** provide some level of internal statistics and performance figures
  - Identify and familiarize with these and compare them with benchmark reported figures

Rules of thumb, #6

- Test from bottom-to-top
  - Start testing with the very basics and from the bottom of the disk backend system
  - Compare observed results with theoretical/expected result
  - Identify performance gaps, if any and explain them
  - Keep testing by adding one more component to the test setup and move towards the top
  - Bottom, naturally is the disks and top is the host ports for a given disk backend system

Rules of thumb, #7

- Have an expected result (hypothesis) before actually running any test
  - This will help analyzing the results and tester's understanding of the underlying hardware

Rules of thumb, #8

- Use the right tool for the right job
  - Do not use a cannon to kill a fly!
    - If you want to exercise the disks at the very basic level (e.g. block level), do not use a user-level MPI-based benchmark (e.g. IOR)

Rules of thumb, #9

- Be a good recorder; document everything!
  - You will forget
    - *What you did*
    - *How you did it*
    - *What were the results*
  - Down the road you will need to revisit the results or the tests

Rules of thumb, #9 (continued)

- Document everything after a given test
- Keep notes of
  - Benchmark command lines
  - Configurations settings
  - Client configurations
  - Storage and file system settings
- If possible, write a few sentences about the test and the results
- Archive and time stamp

Putting all together

- Know thy hardware
- Understand the critical data path
- Understand how component interacts with each other
- Repeat the tests, verify the variability
- Gather stats and performance data from disk backend
- Test bottom-to-top
- Have a hypothesis before running the test
- Use right tools
- Record and document everyhting

Thank You

- Questions?

Sarp Oral
Research and Development Staff Member
865-574-2173
oralhs@ornl.gov