Scalability of Linux pNFS

Presentation by: Bill Hass

Outline

- Objective
- Testing Architecture
  - Clustered Back-end
  - Single iSCSI Back-end
- Initial Testing Using Linux Back-End
- Current Testing Using Windows Back-End
- Thoughts

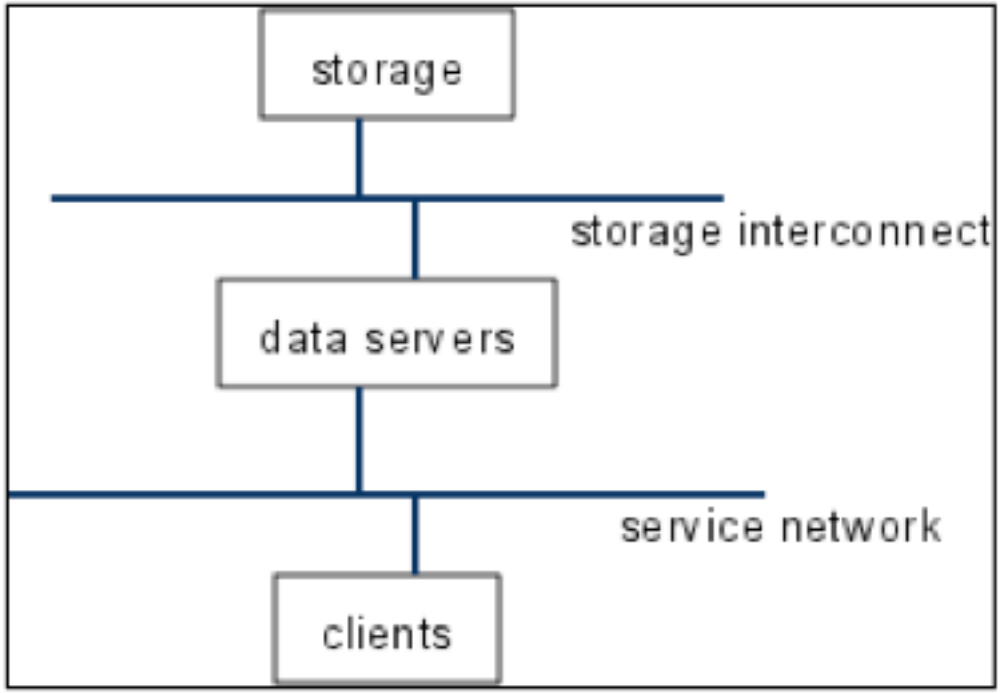
Objective

- To demonstrate the performance scalability of pNFS

Initial Testing Architecture
Clustered Back-End

The test platform is organized into three tiers:

- Storage servers
- Data servers
- Clients

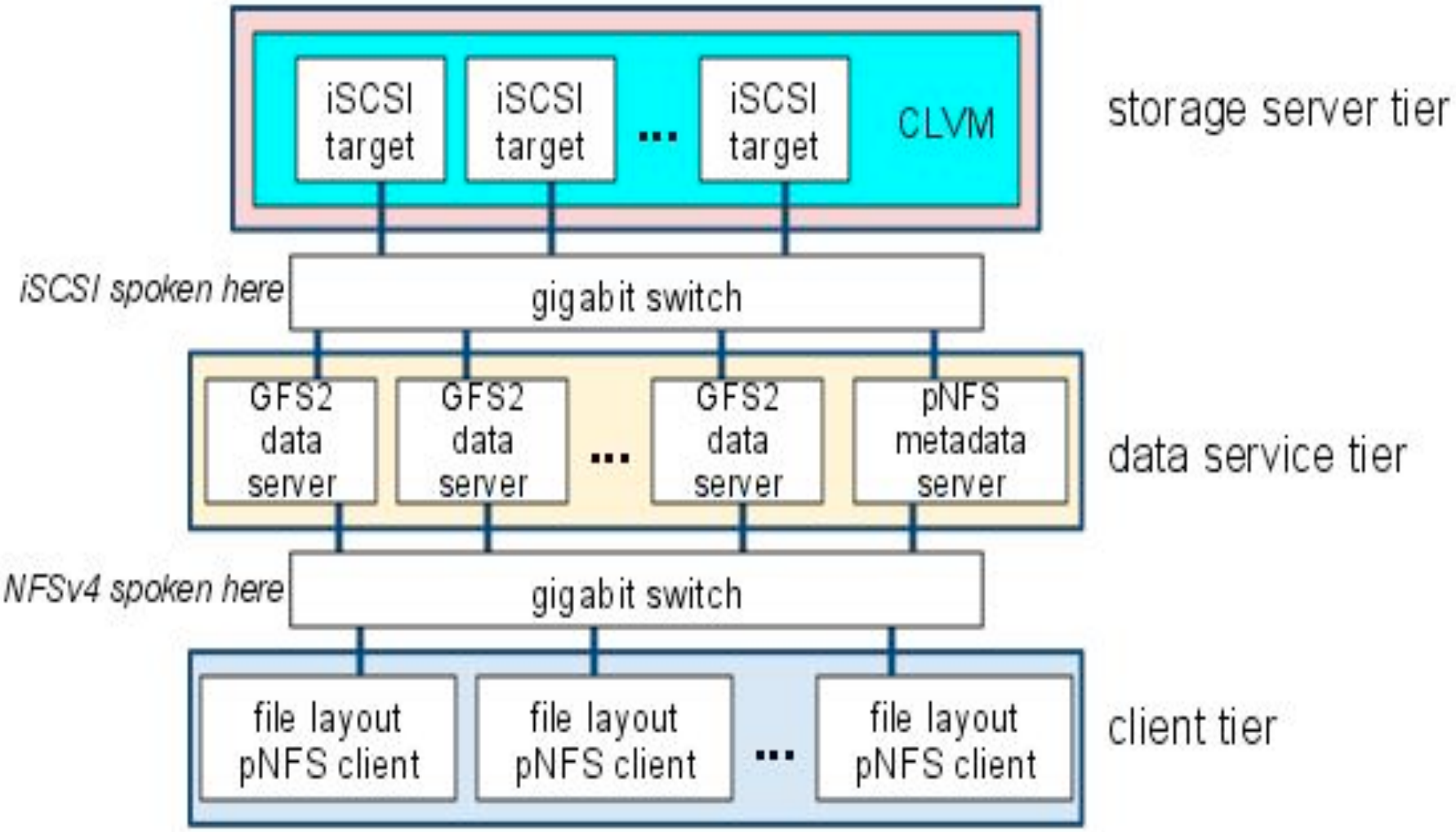


Storage servers are clustered using CLVM and the data servers are clustered using cman.

Data servers and storage servers communicate with the iSCSI protocol over 1 GbE (1 Gbps over Ethernet).

Data servers and clients communicate with the NFSv4.1 protocol over 1 GbE.

Initial Testing Architecture
Clustered Back-End



Problem?

- After running a few tests, Eric and Mike were unable to get good performance from the storage server back-end.
- We have a limited number of capable machines, using CLVM only allowed for up to 3 Data servers and 1 Metadata server.

Current Testing Architecture
Single iSCSI Back-End

The test platform is again organized into three tiers:

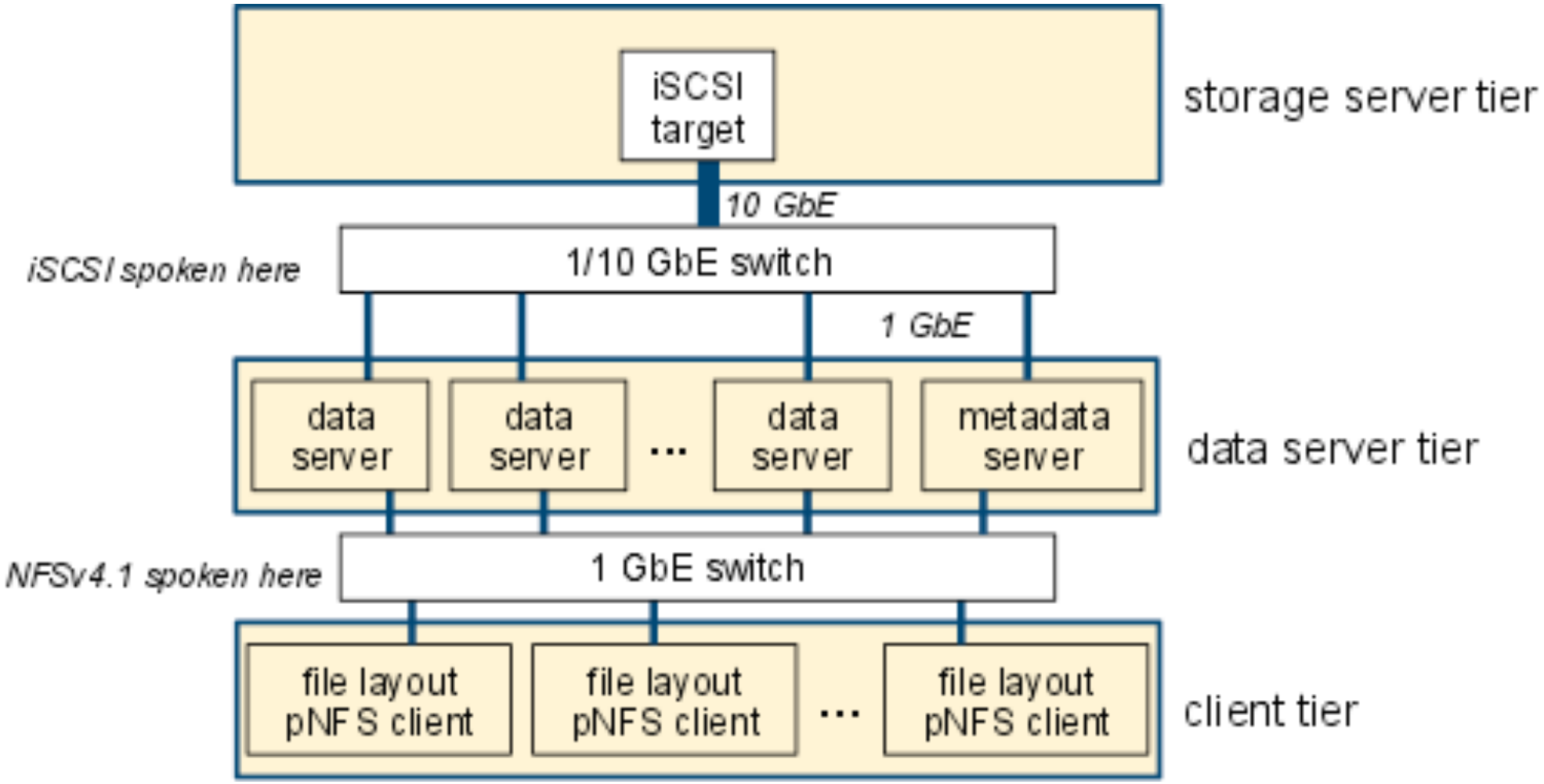
- Storage servers
- Data servers
- Clients

Data servers and storage servers communicate with the iSCSI protocol over 1 GbE and 10 GbE.

Data servers and clients communicate with the NFSv4.1 protocol over 1 GbE.

Current Testing Architecture

Single iSCSI Back-End



Testing Outline

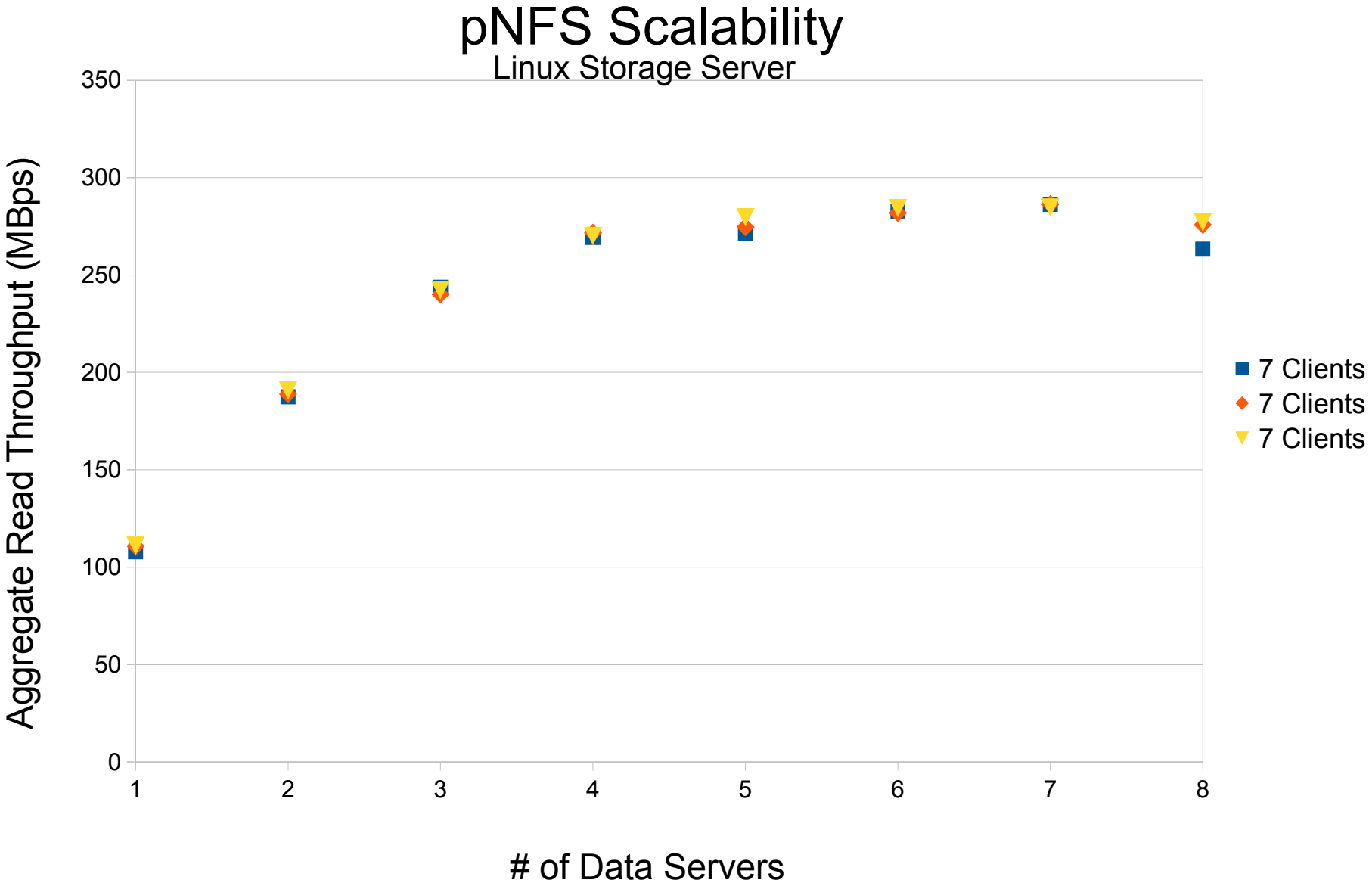
We designed a series of tests to isolate file system performance.

- Test each layer in isolation under different conditions to determine the actual performance versus expected performance.
  - Test storage tier local file system. Compare ext4 and gfs2 performance.
  - Test iSCSI performance with benchmarks running on data servers over ext4 and gfs2 file systems.
  - Test NFSv4.1 access from clients to local file systems on data servers.
- Measure performance of three-tier system with multiple clients and single DS, MDS, and SS.
- Incrementally add DSs until resources are depleted.

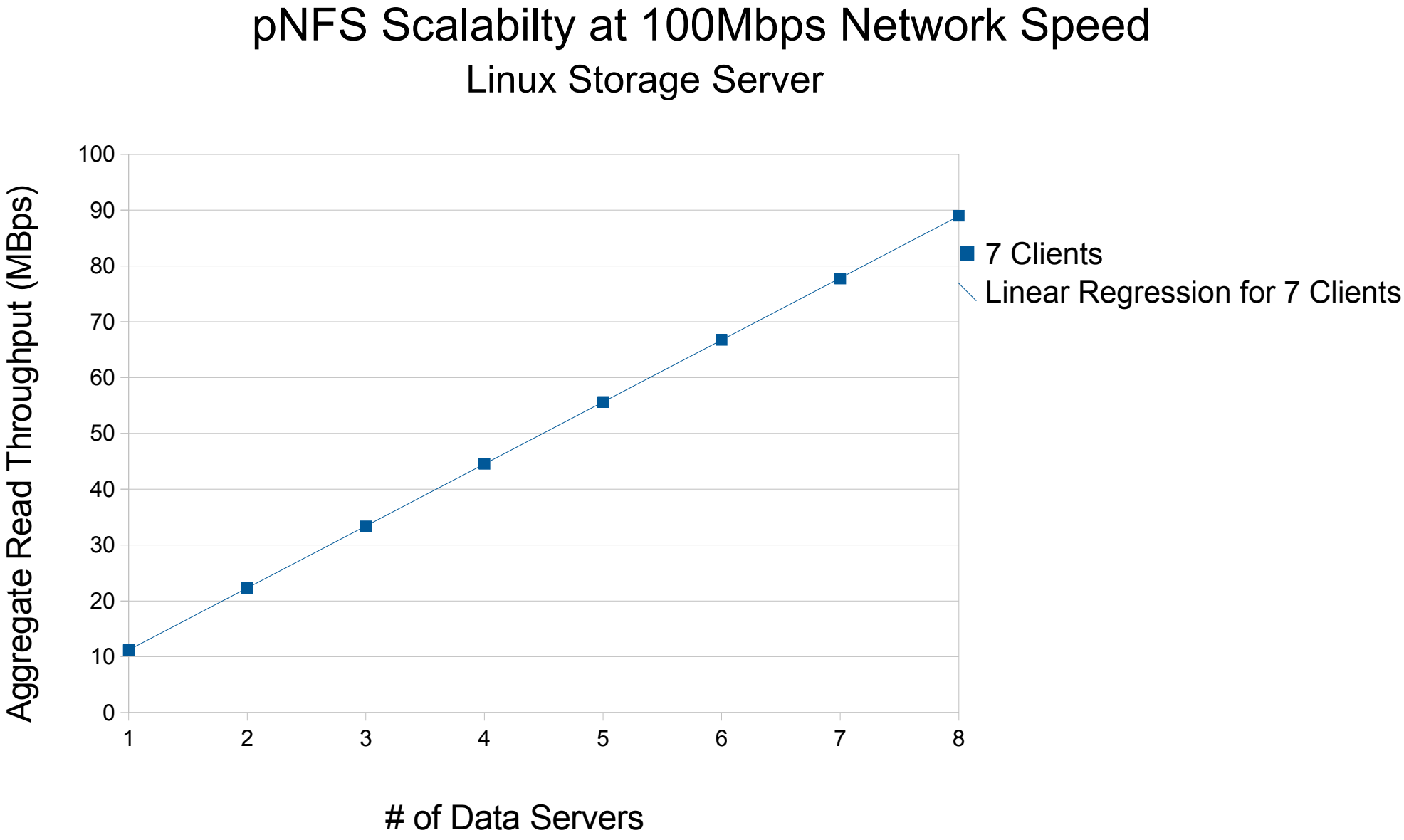
Linux SS Calibration Results

- Isolated test cases shows **storage server local read** throughput performance of **1300 MiBps**.
- NFSv4.1 throughput between **one data server** and **one storage server** (over iSCSI) yields **109 MiBps**.
- Throughput between **one pNFS client** and **one data server** measures **100 MiBps**.
- Aggregate throughput between **multiple clients** to **multiple data servers** exceeds **100 MiBps**.

Linux SS Results



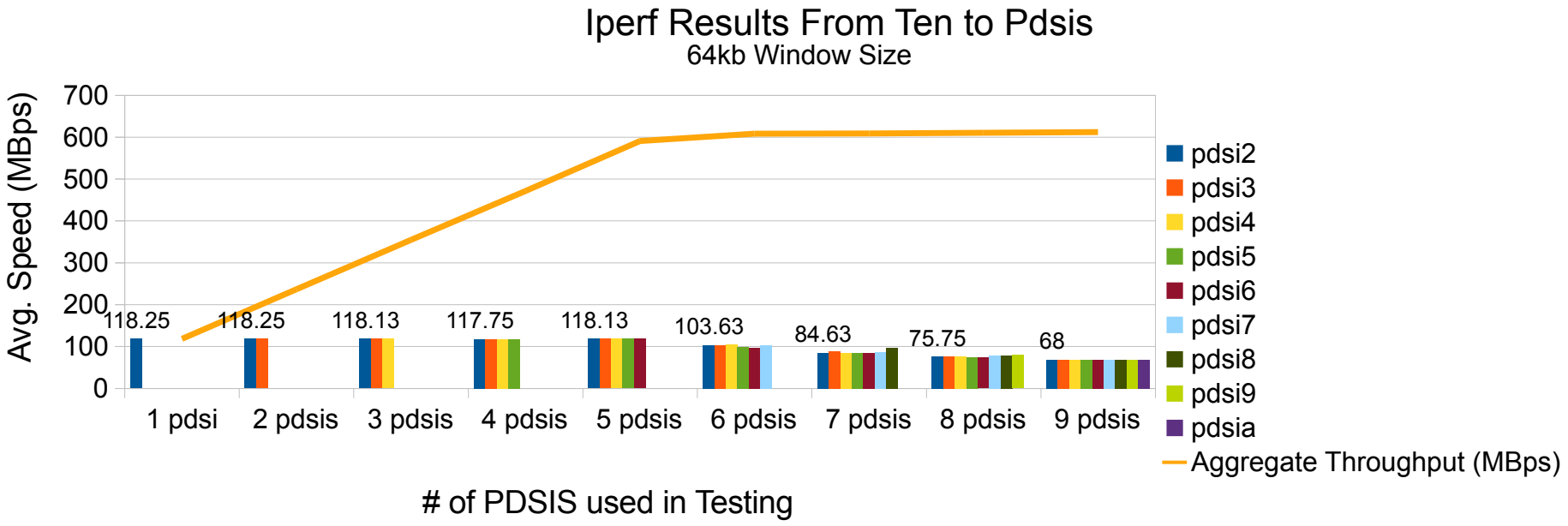
Linux SS Results (cont.)

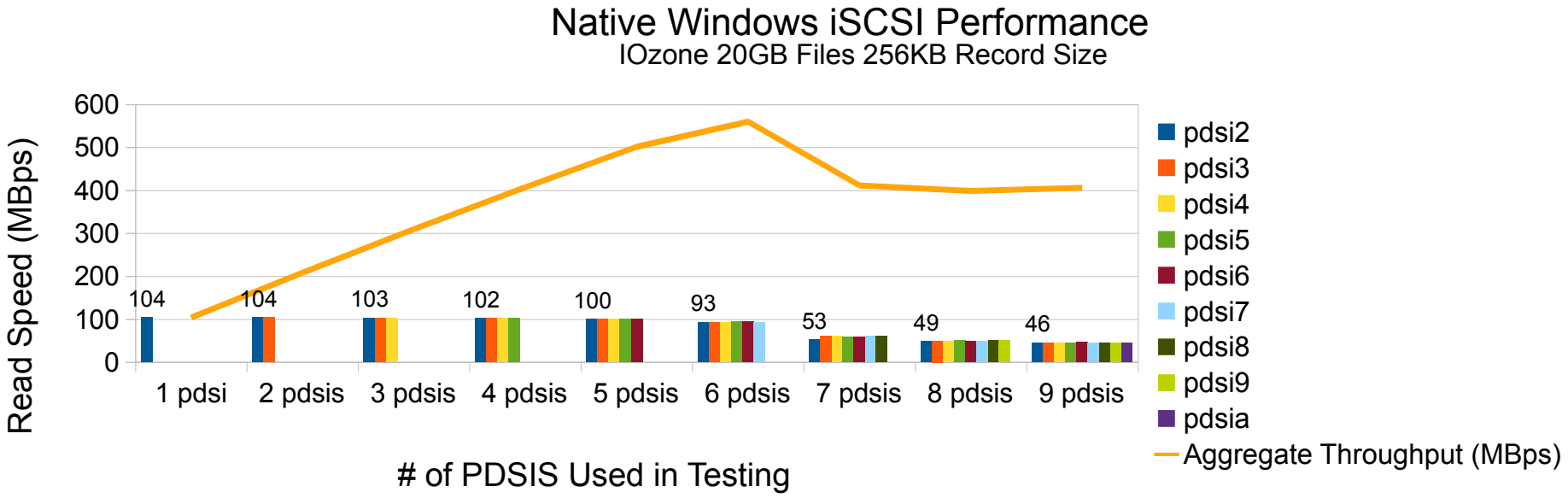


Windows SS Calibration Results

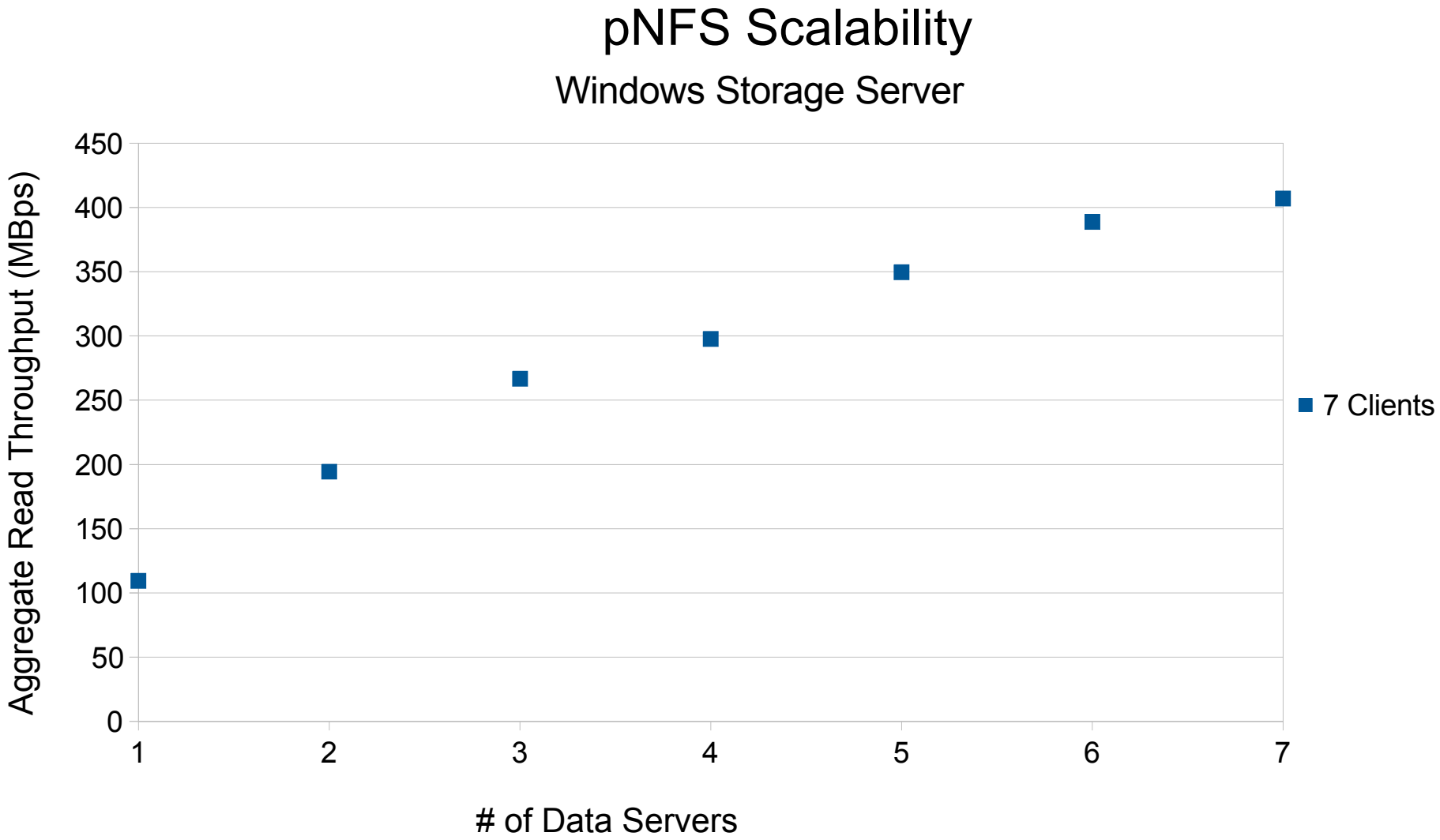
- Isolated test cases show **storage server local read** throughput performance of ~**1100 MiBps**.
- **pNFS client to data servers** performance remained **unchanged**. (We didn't touch that layer.)
- However, communication between our storage server and data servers was different than in Linux.
- **Iperf results** show **linear performance scaling** with number of nodes until the sixth node is introduced.
- **ISCSI results** also show **linear performance scaling** with number of nodes until the sixth node is introduced.

Calibration Graphs





Windows SS Results



Other Things We've Tried

- Linux SS:
  - AoE: ATA over Ethernet
    - Terrible read performance
  - NBD: Network Block Device
    - Equally as good as iSCSI between data servers and storage server, but couldn't reproduce results.
- Windows SS:
  - Starwind iSCSI
    - Propietary 30-day trial, CPU bottleneck

Thoughts

- Stripe Size testing
- Suggestions?