A Presentation on

# Black-Box Problem Diagnosis in Parallel File System

Authors:
Michael P. Kasick,
Jiaqi Tan,
Rajeev Gandhi,
Priya Narasimhan

Presented by:
Rishi Baldawa

# Key Idea

Focus is on automatically diagnosing different performance problems in Parallel File Systems by  identifying, gathering and analyzing OS-level black-box performance metrics on every node in the cluster to identify the node(s) at fault and develop a root cause analysis procedure for the faults.

# Problem Diagnosis Techniques

- White Box testing incurs significant runtime overhead, requires code-level instrumentation and expert knowledge while Black Box Testing just needs identification of anomalies.
- SLO(Service Level Objective) violations
  - Hard to specify precise SLOs for HPCs
- Statistical/Machine Learning Algorithms
  - Fault Free Training Data
- PVFS and Lustre were probably chosen for as they are some of the most commonly used DFS.

# Background

- Target performance issues for HPC
- Black Box Tests are performed on two types of HPCs
  - PVFS 2.8.0
  - Lustre 1.6.6 (Linux + Cluster)
- Black Box performance analyzed at every node
- Low Overhead
- Low Data Requirement
- SLOs Avoided

# Lustre

- Currently under Oracle
- developed as a research project in 1999 by Peter Braam
- As of October 2010, 15 of the Top 30 super computers use it (including the 1st and the 2nd fastest super computers)
- Single metadata server, one management server (may be co-located with metadata server) and multiple object storage servers
- Implemented entirely in kernel space.
- User space client lib (`liblustre`) is also available
- Configurable striping across one or more object storage targets (`stripe_count, stripe_size`)
- Open

# Related Work

- Peer-Comparison
- Metric Selection
- Message-Based Problem Diagnosis

# Problem Statement

1. *Can we diagnose the faulty server in the face of a performance problem in a Parallel File System*, and (Fault Tolerance)
2. *If so, can we determine which resource is causing the problem?* (Root-Cause Detection)

# Goals

- Application transparency
- Minimal false alarms of anomalies
- Minimal instrumentation overhead
- Specific problem coverage [5]
  - Anomalous Behavior
  - Network Problems
  - Performance Faults
  - Non Fail-Stop Performance problems in Storage and Network.

[5] P. H. Carns, S. J. Lang, K. N. Harms, and R. Ross. Private communication, Dec. 2008.

# Non-Goals / Future Work

- Code-Level Debugging
- Dissimilar Requests Patterns
- Diagnosis of Non-peers
- General Design Flaws

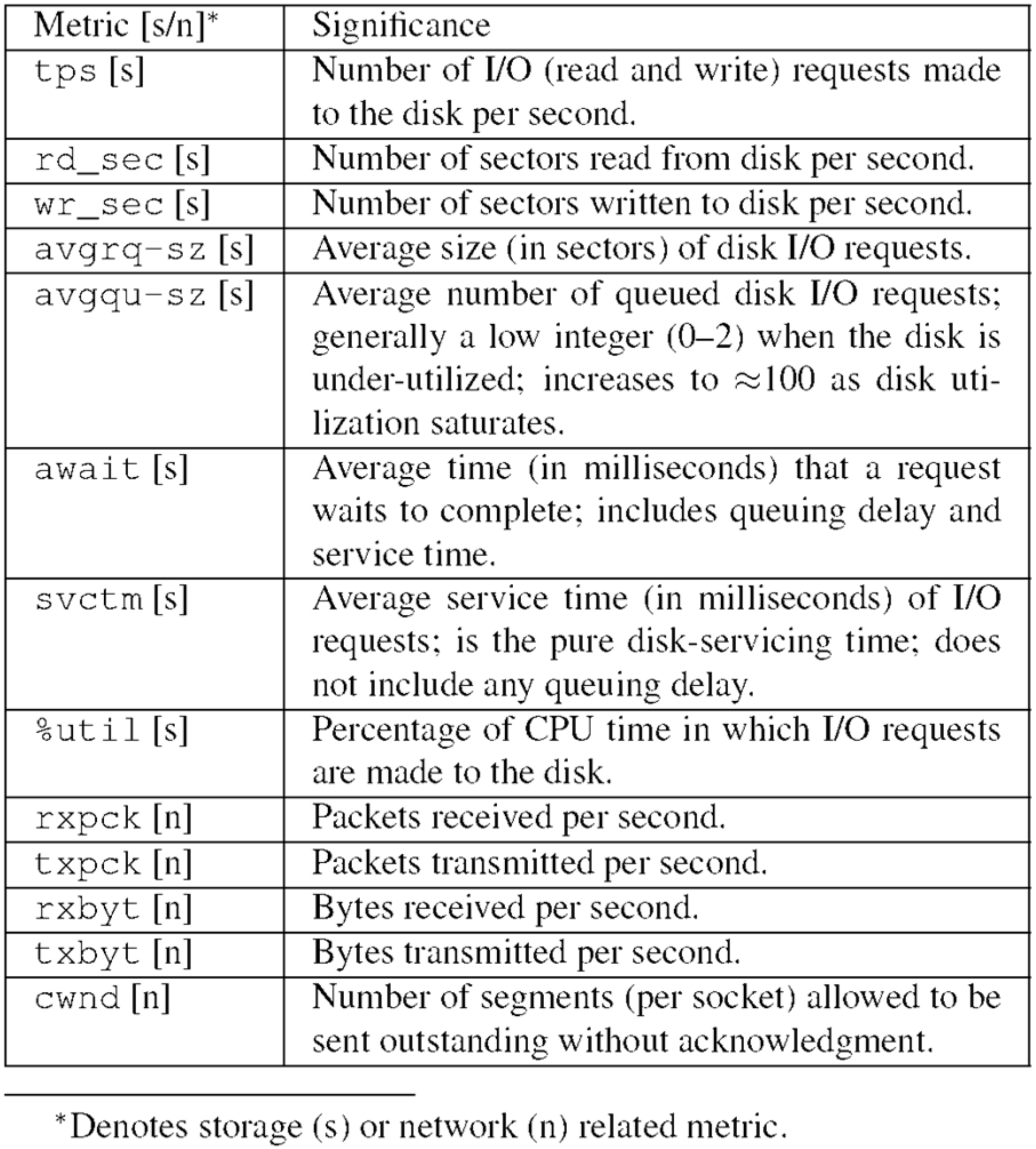
- Secondary Manifestation Realizations
- Design / Metric Flaw
- Heterogeneous Systems (Linux, PVFS/Lustre Independent)

# Parallel DFS Problems

- Hogs and Loss/Busy Faults
  - Disk Hogs
  - Disk Busy
  - Network Hogs
  - Packet Loss
- Workloads
  - DD (`ddr and ddw`)
  - Iozone (`iozoner and iozonew`)
  - Postmark (`v 1.51`)
- Packet Injections

10 experiments conducted for each workloads and fault injections using different fault combination

# Metrics



| Metric [s/n]* | Significance |
|---|---|
| `tps` [s] | Number of I/O (read and write) requests made to the disk per second. |
| `rd_sec` [s] | Number of sectors read from disk per second. |
| `wr_sec` [s] | Number of sectors written to disk per second. |
| `avgrq-sz` [s] | Average size (in sectors) of disk I/O requests. |
| `avgqu-sz` [s] | Average number of queued disk I/O requests; generally a low integer (0–2) when the disk is under-utilized; increases to ≈100 as disk utilization saturates. |
| `await` [s] | Average time (in milliseconds) that a request waits to complete; includes queuing delay and service time. |
| `svctm` [s] | Average service time (in milliseconds) of I/O requests; is the pure disk-servicing time; does not include any queuing delay. |
| `%util` [s] | Percentage of CPU time in which I/O requests are made to the disk. |
| `rxpck` [n] | Packets received per second. |
| `txpck` [n] | Packets transmitted per second. |
| `rxbyt` [n] | Bytes received per second. |
| `txbyt` [n] | Bytes transmitted per second. |
| `cwnd` [n] | Number of segments (per socket) allowed to be sent outstanding without acknowledgment. |

*Denotes storage (s) or network (n) related metric.

# Parallel DFS Behavior/Observations

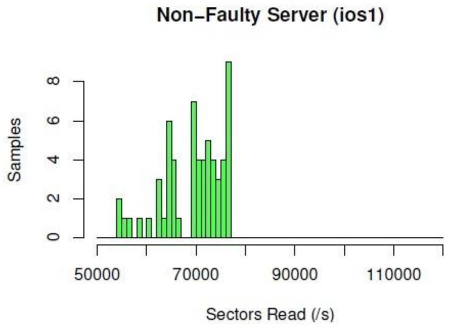
- *In a homogeneous (i.e., identical hardware) cluster, I/O servers track each other closely in throughput and latency, under fault-free conditions.*
- *When a fault occurs on at least one of the I/O servers, the other (fault-free) I/O servers experience an identical drop in throughput.*
- *When a performance fault occurs on at least one of the I/O servers, the other (fault-free) I/O servers are unaffected in their per-request service times.*

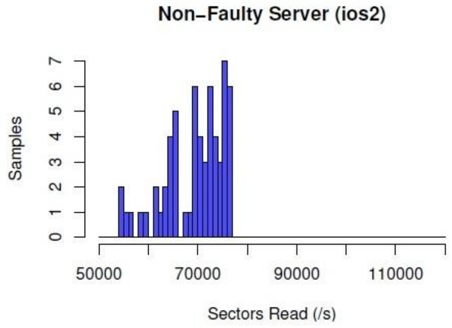
# Parallel DFS Behavior/Observations

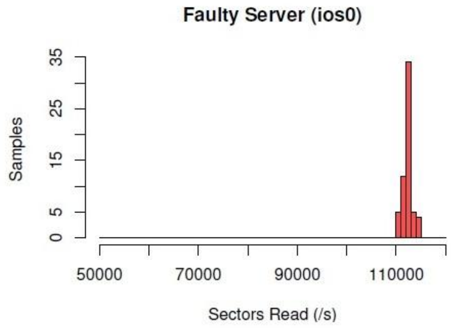
- *For disk/network-hog faults, storage/network-throughput increases at the faulty server and decreases at the non-faulty servers.*
- *For disk-busy (packet-loss) faults, storage (network) throughput decreases on all servers.*
- *For disk-busy and disk-hog faults, storage-latency increases on the faulty server and decreases at the non-faulty servers.*
- *For network-hog and packet-loss faults, the TCP congestion-control window decreases significantly and asymmetrically on the faulty server.*

# Diagnosis

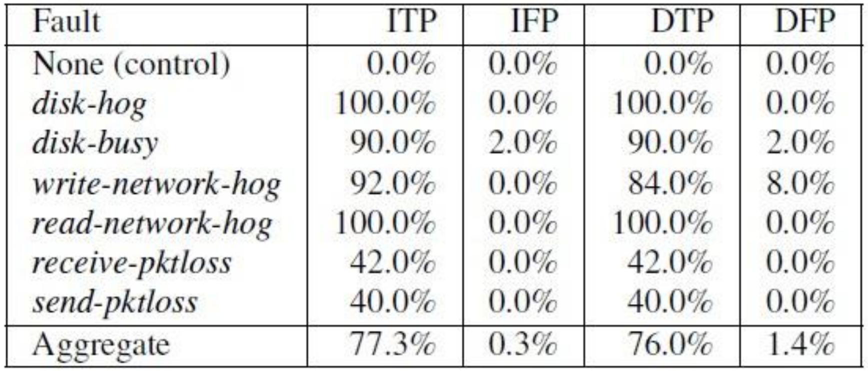
- Finding the faulty server
  - Histogram Based Approach
    - KL Divergence
  - Time Series Based Approach
    - Cwnd
  - Threshold Selection
- Root Cause Analysis
  - Storage Throughput
  - Storage Latency
  - Network Throughput
  - Network Congestion





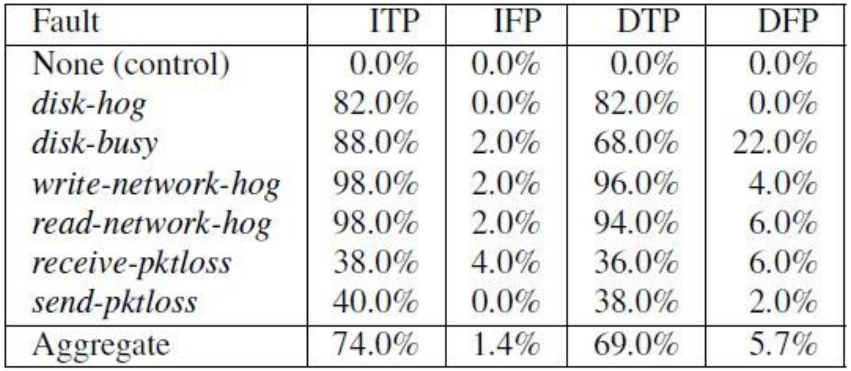


# Results



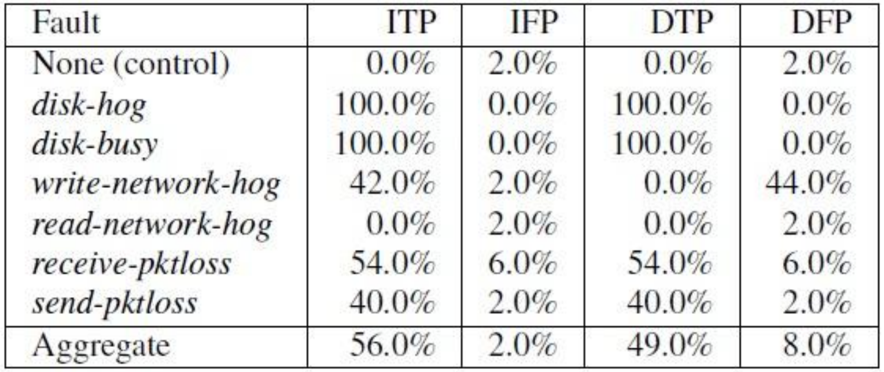
| Fault | ITP | IFP | DTP | DFP |
|---|---|---|---|---|
| None (control) | 0.0% | 0.0% | 0.0% | 0.0% |
| *disk-hog* | 100.0% | 0.0% | 100.0% | 0.0% |
| *disk-busy* | 90.0% | 2.0% | 90.0% | 2.0% |
| *write-network-hog* | 92.0% | 0.0% | 84.0% | 8.0% |
| *read-network-hog* | 100.0% | 0.0% | 100.0% | 0.0% |
| *receive-pktloss* | 42.0% | 0.0% | 42.0% | 0.0% |
| *send-pktloss* | 40.0% | 0.0% | 40.0% | 0.0% |
| Aggregate | 77.3% | 0.3% | 76.0% | 1.4% |

Results of PVFS diagnosis for the 10/10 cluster.



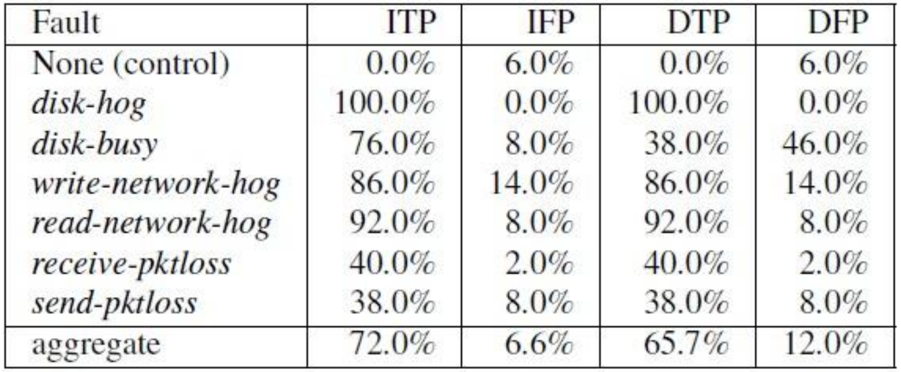
| Fault | ITP | IFP | DTP | DFP |
|---|---|---|---|---|
| None (control) | 0.0% | 0.0% | 0.0% | 0.0% |
| *disk-hog* | 82.0% | 0.0% | 82.0% | 0.0% |
| *disk-busy* | 88.0% | 2.0% | 68.0% | 22.0% |
| *write-network-hog* | 98.0% | 2.0% | 96.0% | 4.0% |
| *read-network-hog* | 98.0% | 2.0% | 94.0% | 6.0% |
| *receive-pktloss* | 38.0% | 4.0% | 36.0% | 6.0% |
| *send-pktloss* | 40.0% | 0.0% | 38.0% | 2.0% |
| Aggregate | 74.0% | 1.4% | 69.0% | 5.7% |

Results of Lustre diagnosis for the 10/10 cluster.



| Fault | ITP | IFP | DTP | DFP |
|---|---|---|---|---|
| None (control) | 0.0% | 2.0% | 0.0% | 2.0% |
| *disk-hog* | 100.0% | 0.0% | 100.0% | 0.0% |
| *disk-busy* | 100.0% | 0.0% | 100.0% | 0.0% |
| *write-network-hog* | 42.0% | 2.0% | 0.0% | 44.0% |
| *read-network-hog* | 0.0% | 2.0% | 0.0% | 2.0% |
| *receive-pktloss* | 54.0% | 6.0% | 54.0% | 6.0% |
| *send-pktloss* | 40.0% | 2.0% | 40.0% | 2.0% |
| Aggregate | 56.0% | 2.0% | 49.0% | 8.0% |

Results of PVFS diagnosis for the 6/12 cluster.



| Fault | ITP | IFP | DTP | DFP |
|---|---|---|---|---|
| None (control) | 0.0% | 6.0% | 0.0% | 6.0% |
| *disk-hog* | 100.0% | 0.0% | 100.0% | 0.0% |
| *disk-busy* | 76.0% | 8.0% | 38.0% | 46.0% |
| *write-network-hog* | 86.0% | 14.0% | 86.0% | 14.0% |
| *read-network-hog* | 92.0% | 8.0% | 92.0% | 8.0% |
| *receive-pktloss* | 40.0% | 2.0% | 40.0% | 2.0% |
| *send-pktloss* | 38.0% | 8.0% | 38.0% | 8.0% |
| aggregate | 72.0% | 6.6% | 65.7% | 12.0% |

Results of Lustre diagnosis for the 6/12 cluster.

**ITP** is the percentage of experiments where all faulty servers are correctly indicted as faulty, **IFP** is the percentage where at least one non-faulty server is misindicted as faulty. **DTP** is the percentage of experiments where all faults are successfully diagnosed to their root causes, **DFP** is the percentage where at least one fault is misdiagnosed to wrong root cause.

# Experiences

- Heterogeneous Hardware
- Multiple Clients
- Buried ACKs
- Delayed ACKs
- Cross-Resource Fault Influences
- Metadata Request Heterogeneity
- Network Metric Diagnosis Ambiguity

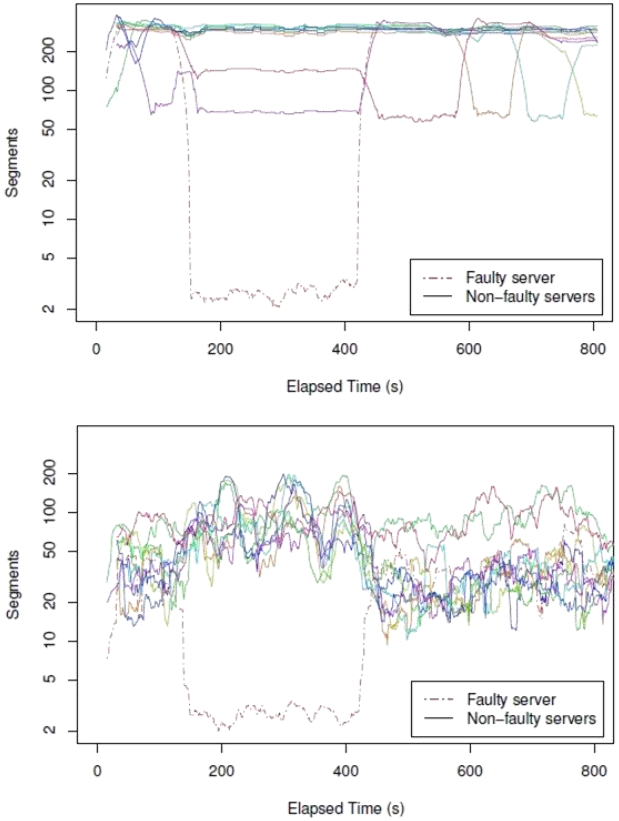


Figure : Single (top) and multiple (bottom) client cwnds for ddw workloads with *receive-pktloss* faults.

# Conclusion

- Black Box Testing strategies based on empirical insights have be used for recognizing faults and the resources that create them in PVFS and Lustre Parallel File Systems

# Thank You

Questions?