PVFS Tuning

PVFS Development Team

April 14, 2008

Contents

1 Introduction

The default settings for PVFS (those provided and in the source code and added to the config files by `pvfs2-genconfig`) provide good performance on most systems and for a wide variety of workloads. This document describes system level and PVFS specific parameters that can be tuned to improve performance on specific hardware, or for specific workloads and usage scenarios.

In general performance tuning should begin with the available hardware and system software, to maximize the bandwidth of the network and transfer rates of the storage hardware. From there, PVFS server parameters can be tuned to improve performance of the entire system, especially if specific usage scenarios are targetted. Finally, file system extended attributes and hints can be tweaked by different users to improve individual performance within a system with varying workloads.

Some (especially system level) parameters can be tuned to provide better performance without sacrificing another property of the system. Tuning some parameters though, may have a direct effect on the performance of other usage scenarios, or some other property of the system (such as durability). Our discussion of performance tuning will include the tradeoffs that must be made during the tuning process, but the final decisions are best made by the administrators to determine the optimal setup that meets the needs of their users.

2 Cluster Partitioning

For users that have one use case, and a generic cluster, what's the best partition of compute/IO nodes? Is this section needed?

3 Storage

3.1 Server Configuration

How many IO servers? [1] How many MD servers? Should IO and MD servers be shared?

3.2 Local File System

- ext3
- xfs

3.3 Disk Synchronization

The easiest way to see an improvement in performance is to set the `TroveSyncMeta` and `TroveSyncData` attributes to "no" in the `<StorageHints>` section. If those attributes are set to "no" then Trove will read and write data from a cache and not the underlying file. Performance will increase greatly, but if the server dies at some point, you could lose data. At this point in PVFS2 development, server crashes are rare outside of hardware failures. PVFS2 developers should probably leave these settings to "yes". If PVFS2 hosts the only copy of your data, leave these settings to "yes". Otherwise, give "no" a shot.

Sync or not, metadata, data coalescing
distributed metadata

[1] The FAQ already answers this to some degree

3.4 Metadata

3.4.1 Coalescing

3.5 Data

4 Networks

4.1 Network Independent

1. Unexpected message size
2. Flow Parameters

- buffer size
- count

4.2 TCP

4.2.1 Kernel Parameters

4.2.2 Socket Buffer Sizes

4.2.3 Listening Backlog (?)

4.3 Infiniband

4.4 Myrinet Express

5 VFS Layer

5.1 Maximum I/O Size

5.2 Workload Specifics

5.3 Extended Attributes

5.3.1 Directory Hints

- Number of Datafiles
- Stripe Size
- Distribution

6 Workloads

6.1 Small files

6.2 Large Files

6.3 Concurrent IO

7 Benchmarking

- mpi-io-test
- mpi-md-test

8 References