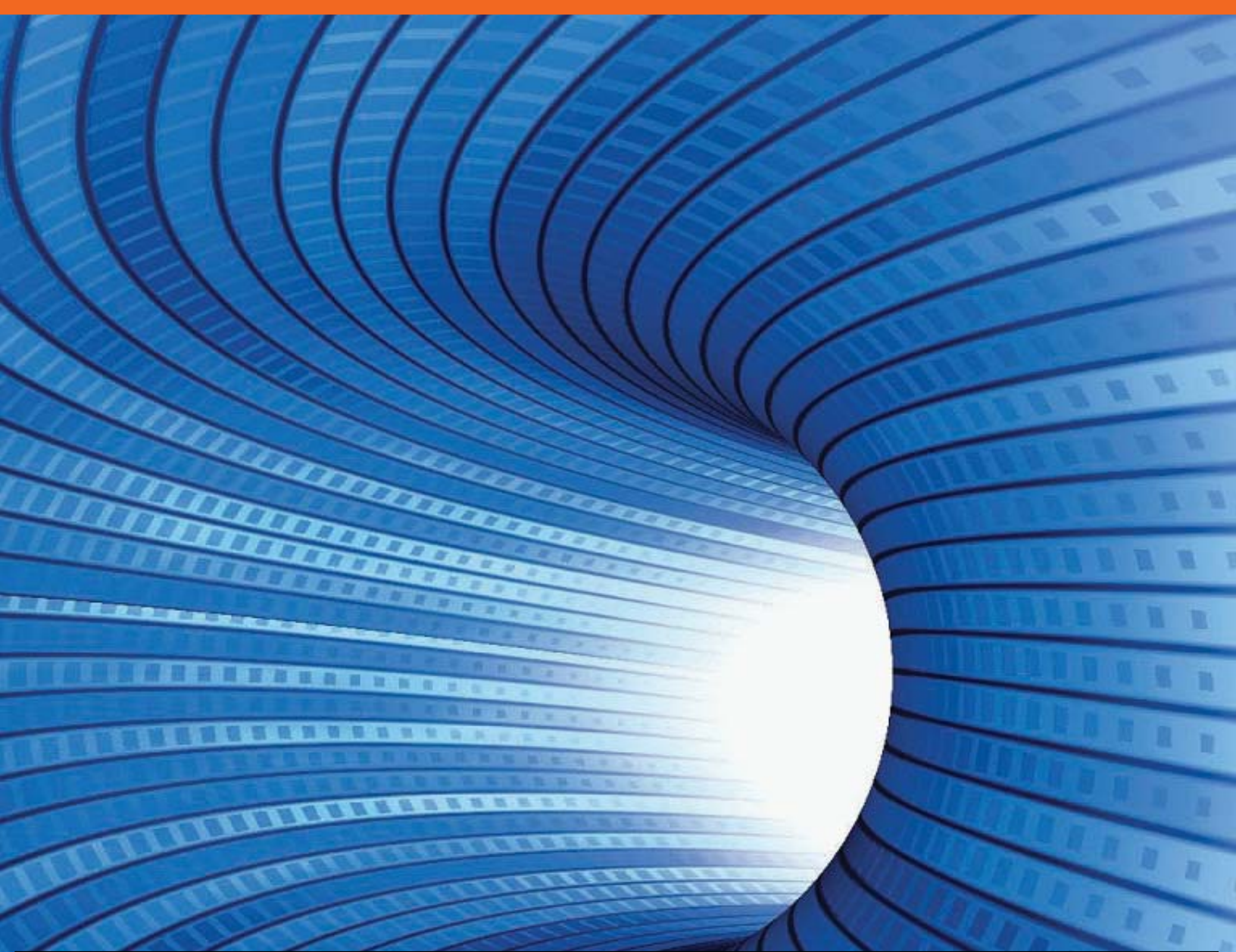


Local and Remote Metadata Replication



real testing | real data | real results



Signature Page, Release Acknowledgement

By signing below, the DICE collaborators certify that the contents presented in this document are accurate to the best of their knowledge.

Mark Roberts
Intivity Corporation

Tracey Wilson, TRB Chair

Al Stutz, Governance Board Chair

Local and Remote Metadata Replication Table of Contents

| | | |
|---|---|---|
| 1 | Project Executive Summary..... | 4 |
| 2 | Introduction..... | 4 |
| 3 | Project Overview/Description..... | 4 |
| 4 | Project Goals..... | 5 |
| 5 | Tests and Evaluations | 5 |
| 6 | Evaluation Results | 6 |
| 7 | Analysis of Deviations from Predictions | 7 |
| 8 | Conclusions..... | 8 |
| 9 | Proposed Next Steps | 8 |

1 Project Executive Summary

Intivity submitted this proposal to address the challenge areas of Distributed and Parallel File Systems & Global Data Sharing Technologies using Intivity's data management framework. Our goal was to deploy new features to the Intivity's DMcore framework to leverage distributed Hierarchical Storage Management (HSM) processes to manage the staging of large data sets between geographically disparate systems by replicating metadata in a federated namespace. Individual compute nodes will perform I/O to individual local filesystems, and HSM processes will manage the namespace consistency between them. We proposed to use RDBMS queues for metadata replication in a geographically disparate environment, to achieve highly reliable and easily monitored job queuing.

2 Introduction

Intivity Corporation specializes in policy-driven data replication and distribution software that works at the file level. Intivity believes that data management must happen within the context of the needs of the business to provide the maximum ROI and alignment with direction of the data management marketplace. Intivity Active Data Management solutions provide reliable, scalable data migration, replication, storage tiering, and archival for any storage or distance. Intivity is headquartered in Austin, TX, USA.

Intivity DMcore is open systems software which serves as a platform to enable complex data management services such as transparent data migration, remote copy, archival, data mirroring and filesystem federation. The Intivity DMcore performs Active Data Management: policy-based data management linking related filesystems or file spaces (also known as tiers) together into alternative views, and optionally copies, of the same information.

This project involved federating a filesystem namespace both locally and remotely. As files were created or updated, their metadata was mirrored (or updated) to the other tiers within the federation. If a file was accessed from a tier where its data was non-resident, data was retrieved automatically. This amounts to a mirrored namespace with application-transparent, on-demand data retrieval.

3 Project Overview/Description

This project involved software/server installation and testing at the AVETEC/DICE facility in Springfield, OH and the Goddard Space Flight Center in Greenbelt, MD.

Intivity provided the DMcore software (formerly OpenSMS) configured to provide HSM functionality for the XFS filesystem with a Data Management API (DMAPI) enabled Linux kernel. SUSE Enterprise Linux 9 is a COTS Linux Distribution which includes all filesystem and kernel dependencies necessary to support the OpenSMS Data Management runtime environment.

Intivity also provided Intel/Linux servers for testing at the DICE and Goddard facilities.

4 Project Goals

The principal area of investigation in this deployment was the exploration of whether the efficiencies gained by local data caches for all writes, and most reads can overcome the latencies encountered when a cache miss is encountered, and the data must be staged in from another copy. We anticipated the processing environment to contain enough variables that we could only generalize the advantages and disadvantages of our approach after directly observing this system with several representative applications and system topologies. Our goal was to use the DICE processing environment to explore the implications of various SMP and/or MPP applications and their data access patterns, and to determine what optimizations should be pursued.

We believe our results have shown that the DICE processing environment will benefit not only from the deployment of this new data distribution and management architecture, but from being the reference implementation whose needs drive the next steps in the development of this architecture.

5 Tests and Evaluations

This exercise tested the Intivity DMcore (OpenSMS) support for filesystem federations with metadata replication. In a filesystem federation, files created or updated on any filesystem within the federation are replicated to the other filesystem(s) in the federation.

Test Configuration

1. A pair of file systems on separate nodes was configured as a federation.
2. Each filesystem in the federation was configured to receive metadata-only updates (data retrieval happens only if a non-resident file was accessed).
3. A “test set” of 1000 files was copied into each of the federated filesystems, at a different relative path, such that metadata was to be replicated both ways in order to build the aggregated filesystem view – e.g., a directory full of files was copied to `./set0` on the first filesystem, and `./set1` on the second filesystem. Replication resulted in both subdirectories appearing in both filesystems.

Test Measurements

1. Metadata replication time for the “test set” of files was repeatedly measured, with and without pre-existing replica files. This test will inherently measure file replication in “files per second” since it bypasses a file’s data and replicates only the attributes. Re-replicating a file involves a different and slightly lighter weight code path.
2. All files in both filesystems were compared (full data compare) for integrity. Since the files were replicated “in” as metadata only, this caused half the files to be “faulted in” on one filesystem, and the other half to be “faulted in” on the other filesystem – meaning all data comparisons triggered data retrieval on one filesystem or the other. Since the filesystems are resident on separate systems, the comparison test was accomplished by NFS mounting one of the filesystems on the other system to allow local filesystem semantics to be used for the comparison.

3. Metadata movement performance was measured, reported and analyzed.

6 Evaluation Results

Below is a complete breakdown of all test results at the AVETEC/DICE (local) facility and the NASA (remote) facility.

Local replication

Test systems at AVETEC:

“Green” - Athlon XP 1900, 512MB RAM, 100bt Ethernet

“Blue” - Xeon P4 2.4Ghz, 1GB RAM, 100bt Ethernet, Mysql database server

| | Blue to Green | Green to Blue |
|-----------------------|---|--|
| | (Average over 1476 tests of 501 files each) | (Average over 303 tests of 501 files each) |
| Replication | 32.1 files/sec | 26.6 files/sec |
| Re-Replication | 33.4 files/sec | 23.7 files/sec |

Metadata replication rates with added network load:

For these tests we ran flood pings in both directions between the nodes.

| | Blue to Green | Green to Blue |
|-----------------------|---|--|
| | (Average over 1476 tests of 501 files each) | (Average over 303 tests of 501 files each) |
| Replication | 25.3 files/sec | 26.5 files/sec |
| Re-Replication | 26.4 files/sec | 24.3 files/sec |

Data verification:

Blue to Green: good

Green to Blue: good

Remote Replication

Remote system setup:

“AVETEC (blue)” - Xeon P4 2.4Ghz, 1GB RAM, 100bt Ethernet, MySQL database server

“NASA (gill)” – Athlon XP2800+, 1GB RAM, Gigabit Ethernet

| | AVETEC to NASA | NASA to AVETEC |
|-----------------------|---|---|
| | (Average over 20 tests of 501 files each) | (Average over 20 tests of 501 files each) |
| Replication | 1.389 files/sec | 6.691 files/sec |
| Re-Replication | 1.805 files/sec | 8.462 files/sec |

No network loading was done during these tests, partly due to the minimal bandwidth to begin with.

Data verification:

AVETEC to NASA: good

NASA to AVETEC: good

7 Analysis of Deviations from Predictions

All tests were successful. Files were replicated correctly and data comparisons matched.

Performance for metadata-only replication was measured in files per second. Performance improvements will be possible by optimizing Intivity DMcore database access as well as local file operations.

Over the LAN, metadata replication performance was slower from Blue to Green than it was from Green to Blue. This is due to the following factors, though relative weighting is not available.

- Replication performance is primarily dependent on the performance of file operations and database queries on the *destination* node – enqueueing replication items appears to always be faster than dequeuing and acting upon them.
- Green, the system with the slower *destination*, performance is the slower of the systems.
- Green is also the node that must access the database via the network.

Over the WAN between Goddard and Ohio, performance was modest, although many improvements are in the works as a result of this testing. Although the limited bandwidth of that connection (about 450KB/sec) is surely a factor, a bigger factor still was the latency – ping latency averaged 80ms. Although performance was better by nearly a factor of five when the destination node contained the database, the limiting factor was the enqueue rate – which was primarily limited by the high latency connection. Work is underway to reduce or eliminate the effect of latency on replication performance.

The data comparison tests, although successful in terms of data integrity, exposed two performance issues that will be addressed in future releases of the software. Both relate to data fault handling (i.e. the retrieval of a file's data, triggered by a read of the file):

- Event triggered data faults were (as of the testing) subject to a 1-2 second polling latency which negatively affects response times and throughput, especially with single-threaded file access (such as our data comparison test). The next release of the software eliminates the polling latency, and is expected to take the pre-retrieval latency from seconds to a few milliseconds.
- If a queued migration is in progress, user read-triggered data faults take precedence as soon as any in-process file movement completes. If a large file is currently being moved, unreasonably long latencies can be introduced. The next release of the software contains separate multi-threaded data movers event-triggered and non-event-triggered data movement.

Enhancements:

After gathering results from the first test configuration, a number of performance enhancements were implemented:

- The mdhandler daemon – which reads metadata entries from a database table and creates or updates file metadata – was dramatically restructured to improve performance. Although the tests were not re-run on the DICE network, we observed an approximate doubling of performance on Intivity’s test network.
- The data mover logic was dramatically streamlined to eliminate a polling latency that affected all data transfers. This was not a factor in the test results, since data movement performance was not part of the protocol – but it made the data and metadata verification tests run dramatically faster.

The primary remaining bottleneck is that our current method for enqueueing metadata update entries amounts to blocking I/O (in the form of database inserts and queries) that is very sensitive to network and database latency. We are still in analysis mode on this issue, but our expectation is that the problem can be completely eliminated. The fix will be analogous to asynchronous, sliding window I/O to the metadata replication queue.

Overall, the test was a success and the DICE environment allowed Intivity to identify and fix several problem areas as well as meeting the Project goals.

8 Conclusions

This project proved that HSM processes can be easily extended to support metadata distribution in large-scale data environments. Our job queuing via RDBMS (MySQL) achieved highly reliable operation and extended the scalability of our technology under high-stress workloads.

We foresee that metadata distribution can be applied to support multiple applications of data management in large-scale environments. In particular, metadata distribution can support filesystem federations and complex distributed archive and HSM architectures.

9 Proposed Next Steps

Having established the mechanisms to efficiently replicate metadata between HSM systems and federations, we believe that the logical steps are to unify the name space between multiple environments, whether single instances or federations of filesystems. Also, as computational environments continue to grow in data volume, the number of servers, and overall geographical distances, Intivity DMcore’s context-aware policy engine can be extended to proactively move data to where it is needed in advance of application requests. Intivity currently has proposals to the TRB to address these areas.