A Scalable Distributed Datastore for BioImaging

R. Cai*, J. Curnutt*, E. Gomez*, G. Kaymaz*, T. Kleffel*, K. Schubert*, J. Tafas*
{egomez, schubert}@csci.csusb.edu
Department of Computer Science
California State University
San Bernardino, CA 92407

ABSTRACT: *Bioinformatics in general and bioimaging in particular are characterized by large, often distributed datastores. Datastores involve more than just a database, as relational databases cannot, for example, store and protect an image. This paper examines the effects of network performance on a bioimaging datastore when it is locally connected, LAN connected, and WAN connected. Issues of node configuration are also considered.*

KEYWORDS: *Bioinformatics, Distributed Datastore, Performance*

1. Introduction

Bioimaging is a special area of bioinformatics, where the information to be analyzed is an image, such as multi-dimensional microscope images and the associated meta-data. The Open Microscopy Environment (OME) provides a commonly used back end to store both microscope images and metadata, with links to analysis products like Matlab. Typically, bioimaging datastores are many terabytes in size and can be used by a consortium of institutions, so the data is of necessity distributed. Bandwidth and latency are well known problems in databases, [2, 3, 4, 5], so large distributed datastores have potentially large challenges to overcome. This presents the problem of what is the best way to distribute this library to maximize performance and availability to an entire bioinformatics collaboration, which exists at remote sites.

In this paper we will discuss the the performance of OME bioimage datastores and how to configure a bioimage datastore for maximum performance. In particular we will examine three major issues:

1. Effects of node configuration, such as system specs and network interface card (NIC) optimization.
2. Comparison of network file systems. NFS is a standard for this purpose and has a reasonable caching scheme to enhance performance. Lustre is a distributed network file system that uses journaling and supports network striping.
3. Effects of a wide area network (WAN) on database performance. This is crucial due to the need for remotely located researchers to interact in a timely manner.

*The authors gratefully acknowledge the support of the National Science Foundation under NSF ITR 0331697

2. Node Configuration

Performance tests were conducted using machines in CSUSB's Raven cluster and UCSB's Hammer and Nails cluster. Each of the six machines in the Hammer and Nail cluster is a new quad Xeon 3.2 Ghz 4GB RAM 140GB SCSI Dell server. Each of the thirteen machines in the Raven cluster is a 5 year old dual processor 1.4 GHz Pentium 3 256MB RAM 60GB SCSI Compaq Proliant DL-360 server. Raven is a loose cluster, to allow flexibility in configuration, and Raven has been optimized for small packets and low latency at the cost of some bandwidth operations. This optimization was done for other research conducted at the Institute of Applied Supercomputing at CSUSB on low latency networking and algorithms for parallel processing. Interestingly low latency can also benefit databases under the conditions below. In straight time Nail outperforms Raven due to it being a much faster machine, see Figure 1. The interesting part of the test data is in the ratios and relative performance. Nail takes a nearly consistent 2.5x more time on latency intensive operations like the aggregate tests and cross-section tests.

Raven's relative and ratio performances on the latency intensive operations are much better than Nail's, and suggest that a latency optimized system has advantages on systems which perform many small database operations from a few users, see Figure 2. In many cases this will be exactly what a bioinformatic collaboration will have, fewer power users. This suggests the best setup

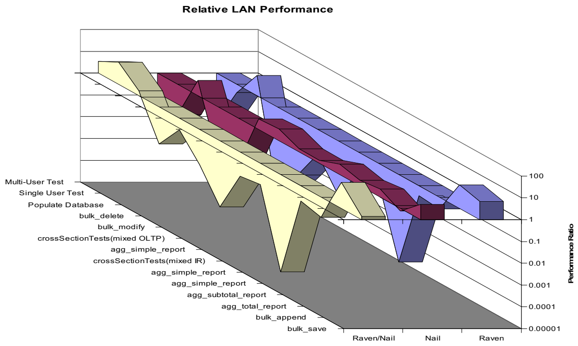

Figure 1: The relative performance of Hammer/Nail to Raven on database.

would be a latency tuned system. This is counter to current wisdom which suggests a bandwidth tuned system. The bandwidth tuned paradigm comes from large commercial databases though, which have very different usages than a scientific data library. Figure 2 shows the relative performance

$$Perf_{relative} = 100\frac{T_{LAN} - T_{local}}{T_{local}}$$

for Nail, Raven and their ratio.

$$Perf_{ratio} = 100\frac{Perf_{relative}(Raven)}{Perf_{relative}(Nail)}$$

Note that Nail performs better for bandwidth intensive operations and Raven performs better for latency intensive operations.

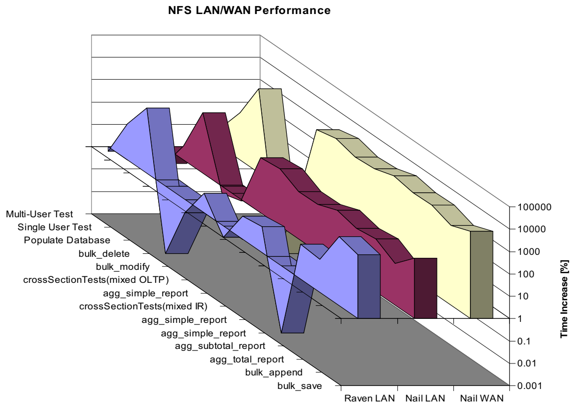

Figure 2: The performance ratio with respect to a local database as a percentage.

3. NFS/Lustre Performance

NFS is a well known standard for remote access to disks, and as such it was used to compare the performance of the Lustre file system. As Lustre's file system can be spread across multiple computers, tests were performed with stripes across 1, 2, and 3 computers.

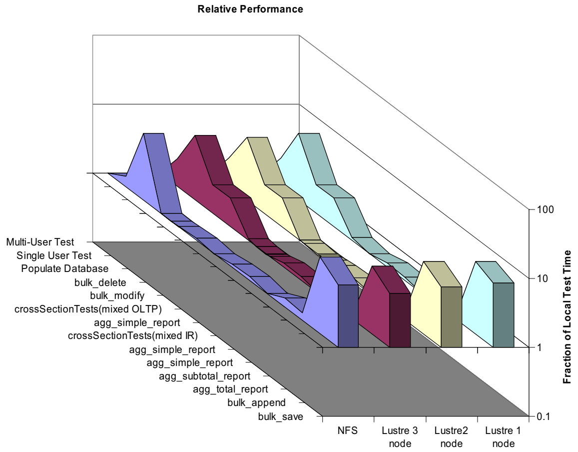

Figure 3: Relative performance of NFS and Lustre with 1, 2, and 3 computer striping on 1Gb LAN.

Several things stand out as interesting in Figure 3.

1. Lustre is consistent for 1, 2, or 3 computers in the clustered file system.
2. The bandwidth intense operations show the worst relative performance compared to a local database for both NFS and Lustre. Some of this is due to the Raven cluster being optimized for latency and not bandwidth, but most is undoubtedly the significantly smaller bandwidth of the network as compared to the internal bus.
3. The only significant NFS/Lustre performance differences are
   (a) the superior performance of NFS in bulk deletes and bulk modifies due to NFS having a delayed write cache scheme, which hides a single write. The effect is not noticed in the other tests because they would generate multiple writes and thus swamp the cache.
   (b) Lustre's ability to stripe across computers and thus increase bandwidth, noticeable in slightly lower times on the high bandwidth metrics. Lustre's performance improved slightly with more systems due to this.

Both NFS and Lustre perform well on latency tests due to Raven's latency tuned network. Unfortunately, NFS is known to have problems working with OME. We have successfully installed and tested OME over Lustre indicating Lustre is a viable possibility. Lustre is a journaling file system so it is important to have good backups, as journaling systems are prone to hanging and requiring a complete reformat to remove the error from the journal.

4. Wide Area Network

A significant issue for distributed libraries is the effect of the connection network, including latency and bandwidth [1]. To test the effects of WAN connections we compared three types of tests:

1. Local- database on the local hard drive.
2. LAN- database on a computer connected by a switched, high speed local area network.
3. WAN- database on a remote computer connected by the wide area network between UCSB and CSUSB.

The most significant effects can be seen not in the bandwidth intensive operations, but rather in the latency intensive operation. This suggests that even a high-bandwidth, next-generation WAN will not solve the problem. Any data that is used frequently must be kept physically nearby. To have large data sets and remote collaborators thus requires either:

- data replication with frequent updates and method to maintain coherence and consistency (like preset exclusive ownership of data categories by location) or
- a database that supports network distributed library by specifying the storage location of each table which can be localized by the network distributed library or
- a database cluster.

5. Conclusions

The major performance results are:

- Latency tune network to support scientific researchers, don't bandwidth tune
- Latency of WAN is too large, so replicate data and update

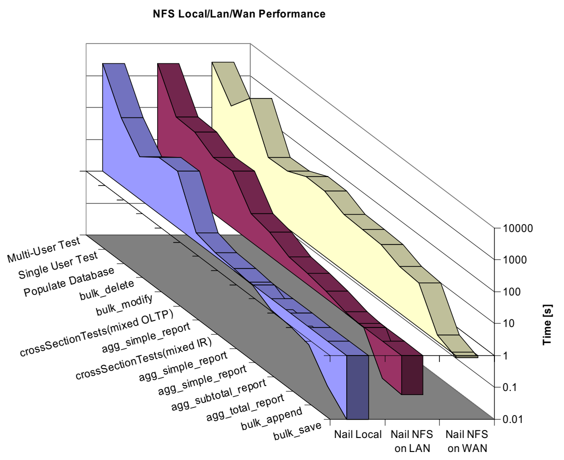

Figure 4: NFS performance on local disk versus 1Gb LAN versus approximately 50Mb WAN.

- Lustre is the only network file system supporting Bisque/OME
- Stripe Lustre across systems can help bandwidth applications

The three main areas studied in this research project yielded several key design rules for a well performing bioinformatic library.

Data should be stored relatively close to the frequent users to reduce latency, which will be a major performance issue for the frequent smaller accesses a single researcher will do. At the moment only data replication is supported. Postgres 8 supports clustering but is not supported by OME at this time. The network aware database would be particularly useful in object oriented databases which would also allow the images and associated data to be stored in the database instead of just being linked.

A latency tuned network will better support a small number of scientific researchers, instead of a bandwidth tuned network, which assumes a larger number of concurrent accesses from typical users. Additionally since the latency of the WAN is too large, data should be replicated with regular updates. To avoid consistency and coherence problems an exclusive ownership of data categories for each location should be used.

Lustre is the only current alternative for networked file systems, and its ability to stripe across systems can help bandwidth applications. A superior caching scheme would be very helpful to hide network delays.

References

[1] Gerco Ballintijn, Maarten van Steen, and Andrew S. Tanenbaum. Characterizing internet performance to support wide-area application development. *Operating Systems Review*, 34(4):41–47, 2000.

[2] Sujata Banerjee, Victor O. K. Li, and Chihping Wang. Distributed database systems in high-speed wide-area networks. *IEEE Journal on Selected Areas in Communications*, 11(4):617–630, 1993.

[3] L. Kleinrock. "the latency/bandwidth tradeoff in gigabit networks". *IEEE Communications Magazine*, 30:36–40, April 1992.

[4] N. Knafla. A prefetching technique for object-oriented databases. In C. Small, P. Douglas, R. Johnson, P. King, and N. Martin, editors, *Advances in Databases, 15th British National Conference on Databases, BNCOD 15*, pages 154–168, London, United Kingdom, 1997. Springer Verlag.

[5] Nils Knafla. Speed up your database client with adaptable multithreaded prefetching. In *Proc. of the Sixth IEEE International Symposium on High Performance Distributed Computing*, pages 102–111, Portland, Oregon, 1997. IEEE Computer Society.