

A Design Summary of the Oasis+ Distributed Shared Memory System¹

Karl Doering*, Eric Will*, Weiping Zhu², Eric Parsons³, Brett D. Fleisch*

**Department of Computer Science and Engineering, University of California, Riverside*

²Queensland University, Australia ³Nortel Networks, Ontario, Canada

{kdoering, ewill, weiping, brett} @cs.ucr.edu, eparsons@nortelnetworks.com

1. Introduction

The Oasis+ Distributed Shared Memory (DSM) system supports a reliable memory store for a small scale computing cluster. The software system is being implemented at the University of California, Riverside and uses a high-performance, high-availability page-based protocol called BR [1]. BR guarantees robust functionality despite multiple site failures that could arise. The system integrates range-based locking and eager release consistency (ERC) [2] for shared data synchronization. Reliability is achieved by replication and corrective cleanup recovery actions once a failure arises. The system uses a reliable multicast package called Spread [3] to disseminate updates to replicas.

This paper gives a brief overview of the salient design features of this system. In addition, we present results from measurements of low-level multicasts used to make updates in Oasis+. The results indicate the potential of this system to operate efficiently in a small cluster of servers where reliability and performance are essential.

2. Background

A high availability DSM system could replace persistent storage if robust methods are used in the design and implementation. A central concept in all these methods is the requirement for replicated data for high data availability. Our group has been studying protocols that not only provide replication, but also do so at low operation overhead. Further, the protocols provide configurable mechanisms for varying the level of replication, so that the system may be operated at the desired overhead cost. Our group has developed several DSM protocols with these features including BR [1], DBR [4] and DBRpc [5] and has performed extensive simulation of the impact of these systems in comparison to traditional DSM approaches. This work extends our previous work by designing and implementing a functional system that uses these protocols.

Coupled with our need for DSM protocols is the need for base primitives that can perform updates efficiently and can manage group membership and

associated topology changes. For this aspect we chose to use the Spread toolkit (v3.11 27 Jul 99) from Johns Hopkins University Computer Science Department [3]. Spread provides multicast and group communications support to applications across local and wide area networks. It consists of an easily snapped-in library with which user applications link and a daemon that runs on each system and manages group membership.

3. System Architecture

There were several changes necessary in the design of the protocols for a practical high availability implementation. BR and DBR originally were specified with a central server that stored a request queue. The queue stored the sequence of read and write operations issued by the client sites, ordered by time of receipt. However, Oasis+ eliminates the server since it is a central bottleneck and possible point of failure. The new system operates completely in a peer-to-peer fashion. Consequently, we have made several modifications to the BR protocol to support the new structure.

An important consideration is the coherency protocol, which influences the frequency of consistency operations in the cluster. BR was originally specified as using sequential consistency with a global queue for page requests. Oasis+ adopts ERC with address-range locking to reduce the frequency of updates to cluster replicas. ERC amortizes costs because write phases can be longer and consequently longer updates reduce the cluster-wide update frequency. Nonetheless, these changes do not alter the functional properties of BR because upon release of the lock the number of replicas is still reduced to the lower bound (LB) number of copies.

The Oasis+ design also considers site initialization and management, region management, and failure recovery. Global state information is maintained at each site instead of at a central server. Global information includes the status of participating sites (initializing/alive/dead) and region information (key, size, handle/region ID, region open site vector, and copyset). To ensure consistency of the information, we use Spread to atomically and

¹ This research was sponsored, in part, by NSF CCR-9704015, an equipment grant from HP Research Labs, and a grant from Nortel Networks. FastAbstract ISSRE Copyright 1999

reliably multicast state information during basic region operations such as create, open, close, and destroy. Spread also manages group membership.

Initialization poses a challenge for a number of reasons. Not only can multiple sites initialize simultaneously, but also multiple sites can fail before, during, and after initialization. Care must be taken to ensure that state information is consistent in all situations. When the first few sites initialize, there may be too few sites to guarantee that the LB replicas for a page exist. During this period, the system is temporarily vulnerable to failure, but as additional sites become available, additional replicas are created until LB copies exist. As a consequence, sites that do not have a region open may be forced to hold copies of pages for regions they do not otherwise use. The situation is self-correcting as additional sites open the region.

Finally, unlike the simulation studies we performed earlier to understand the protocols, our operational system must consider mundane issues such as site maintenance. Over time, sites generally either fail or are taken down for maintenance or upgrade. Provisions must be made for performing a clean shutdown of a single site, which differs from a hard shutdown in that the shutdown does not complete until additional replicas are created to assure the cluster does not store less than the requisite LB copies.

4. Performance of Multicast

The use of Spread has simplified our development task. However, Figure 1 shows that reliable ordered atomic multicast is a relatively expensive operation, particularly as the number of sites increases. The BR protocol bounds the number of replicas of a page between LB and UB. Because the number of replicas of a given page is bounded, the size of the multicast group is bounded by the coherence protocol, and consequently this bounds the cost of replica updates.

We carried out extensive tests to evaluate Spread's performance. These results were measured on HP9000 B160 PA-RISC workstations connected by HP's 100VG Ethernet technology. As shown in Fig. 1, as the number of receivers in the system increases, the time spent sending a message increases exponentially. Our tests also show there is little difference in sending a 16-byte message and sending a 1-KB message. When the message size increases to 4 KB, the difference becomes more pronounced.

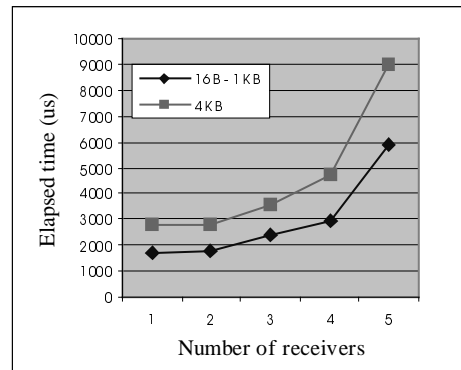


Figure 1: Elapsed time in microseconds for SP_multicast()

5. Conclusions

The design of a practical high availability DSM system has been a considerable challenge. Several aspects of the design needed substantial rethinking for an actual implementation. The BR protocol can bound the number of replicas that must be updated using multicast protocols that are often expensive. Snap-in protocols, such as Spread, are often convenient to use, but may impose high performance overhead, thus reducing their attractiveness.

6. References

- [1] O. E. Theel, B. D. Fleisch, The Boundary-Restricted Coherence Protocol for Scalable and Highly Available DSM Systems, *The Computer Journal*, Oxford Press, Vol. 39, No. 6, pp. 496-510, 1996.
- [2] D. Lenoski, J. Laudon, K. Gharachorloo, A. Gupta, and J. Hennessy, The Directory-Based Cache Coherence Protocol for the DASH Multiprocessor, *Proceedings of the Seventeenth International Symposium on Computer Architecture*, pp. 148-159, Seattle, WA, May 1990.
- [3] Y. Amir and J. Stanton, The Spread Wide Area Group Communication System, Technical Report CNDS-98-4, Johns Hopkins University, Center for Networking and Distributed Systems, 1998.
- [4] O. E. Theel, B. D. Fleisch, A Dynamic Coherence Protocol for Distributed Shared Memory Enforcing High Data Availability at Low Costs, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 7, No. 9, Sept. 1996.
- [5] J. Turk, B. D. Fleisch, DBRpc: A Highly Adaptable Protocol for Reliable DSM Systems, *19th IEEE International Conference on Distributed Computing Systems*, Austin, TX, May 31-June 4, 1999.