Performance of a Distributed Superscalar Storage Server

Arlan Finestead
University of Illinois, National Center for Supercomputing
605 East Springfield, Champaign, Il 61820
arlanf@ncsa.uius.edu

Nancy Yeager
National Center for Supercomputer Applications
152 Computer Applications Building
605 East Sprinfield, Champaign, IL 61820
nyeager@ncsa.uiuc.edu

Introduction

Traditionally, mass storage systems have been single centralized systems; however, a highly distributed mass storage server implemented on superscalar workstations may challenge the centralized model in terms of high file transfer rates and favorable price-performance characteristics. Additionally, a workstation based distributed mass storage server is scalable and may be hierarchically configured as a component of a larger more centralized mass storage system.

National Center for Supercomputing Applications offered a UniTree™ archival service to a select group of users for a trial period of time. The objectives of this trial period were to a) monitor distributed UniTree performance in a production environment under normal and high load conditions b) quantize archival transfer rates from supercomputer clients c) ascertain patterns of UniTree user access d) optimize system performance by tuning file migration from disk to tape.

The archive system architecture consisted of UniTree storage storage servers installed on an IBM RS/6000 Model 550 and an Amdahl model 5860. The UniTree archival software in conformance with the IEEE storage reference model supports a distributed architecture such that the disk operations and tape operations of the storage system may reside on physically separate hosts(see Appendix Figure I). The RS/6000 AIX machine which is fairly efficient at disk operations and protocol processing operations functioned as the Disk Server while the Amdahl UTS serviced tape operations.

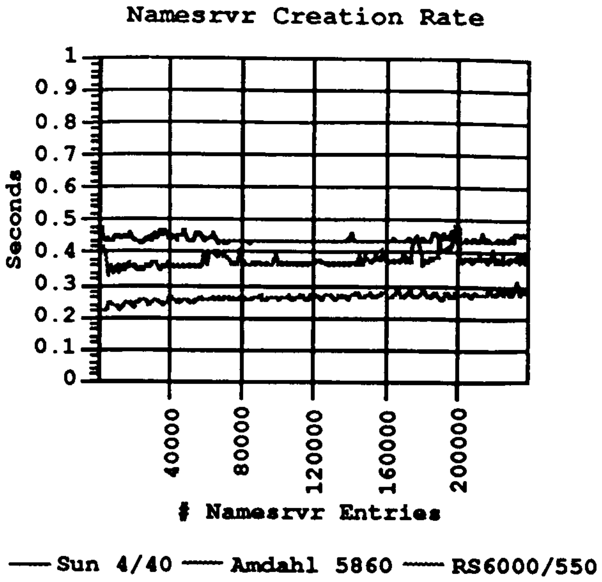
The mass storage serviced archive requests from a farm of loosely coupled IBM RS/6000 Model 550s running scalar computational chemistry codes such as Gaussian-90. Individual RS/6000's within the cluster are interconnected via ethernet; the UniTree Disk server RS/6000 is networked to the Amdahl tape server via FDDI and ethernet.

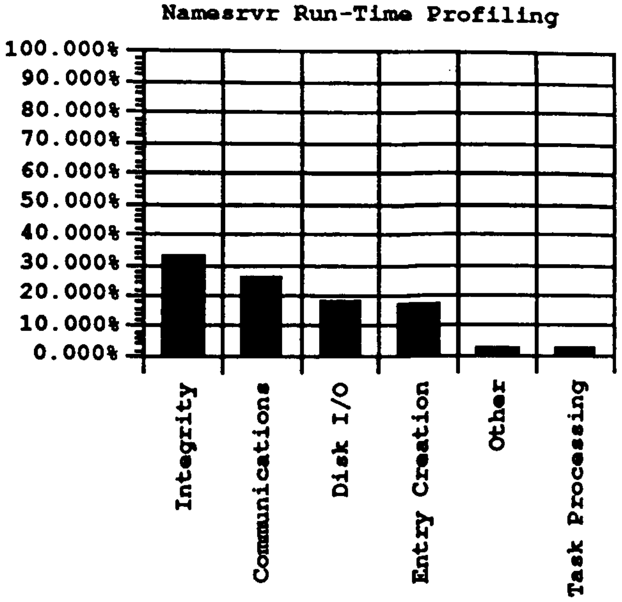
UniTree Performance Testing

Locally developed programs that interfaced directly with the various components of UniTree were used to ascertain the user-perceived performance of UniTree. The tests were varied to simulate a work load model (the load placed on a system by application users) and a system load model (the load according to system metrics such as CPU utilization, inter-server protocol processing, and network traffic). Each of the UniTree components were profiled to determine where potential bottlenecks might exist.

Name Server Performance

The UniTree Name Server daemon exhibited uniform, linear performance when directed to create 230,000 Name Server entries on a RS/6000 Model 550, on a Amdahl 5860, and on a SPARCstation IPC.





The average entry creation time on an RS/6000 Model 550 was .263 seconds, on an Amdahl 5860 was .377 seconds, and on a SPARCstation IPC was .442 seconds. Improvements in Name Server creation performance were realized when the testing program interfaced directly with the UniTree Name Server daemon, bypassing LibUnix altogether. Through optimized creates, entries could be created on the RS/6000 Model 550 in .07 seconds.

The UniTree Name Server was profiled to determine where the majority of execution time was being spent. The UniTree Name Server was categorized into six areas:

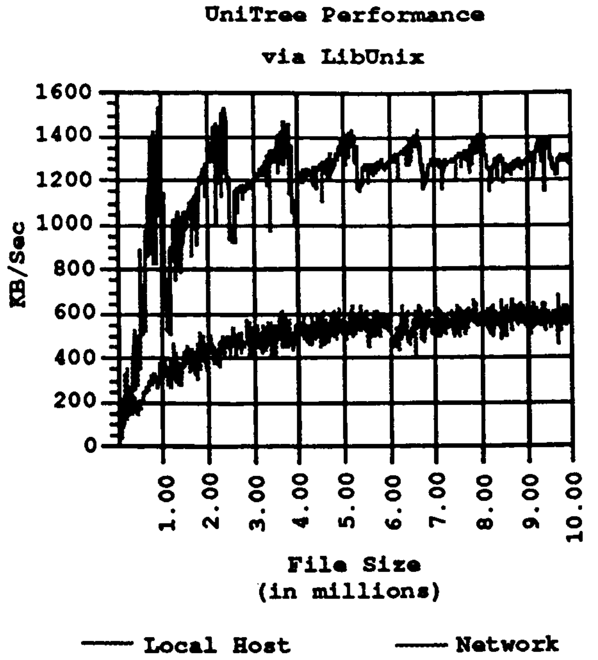
- Integrity - locking data structures, verifying Capabilities.
- Communications - sending, receiving messages via the UniTree APST communication mechanism.
- Disk I/O - performing actual disk operations such as reads and writes.
- Entry Creation - maintaining the Name Server btree structure.
- Task Processing - performing the UniTree task processing.
- Other - includes areas such as logging messages, opening configuration files.

The Name Server creation test program was used to gather the profiling data.

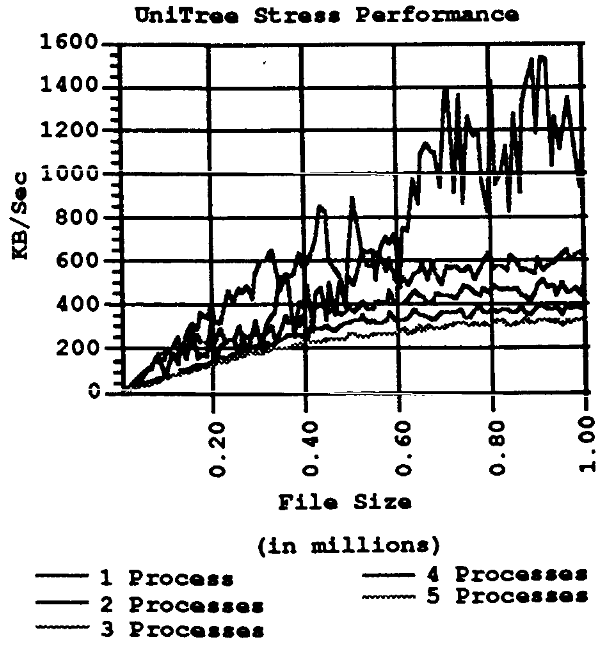
UniTree LibUnix Performance

A test program that interfaced with UniTree via the UniTree LibUnix library was used to determine the performance characteristics of the UniTree Name Server, Disk Server, and Disk Mover daemons. The test program generated increasingly large files in the UniTree archival system, recording the performance with each creation.

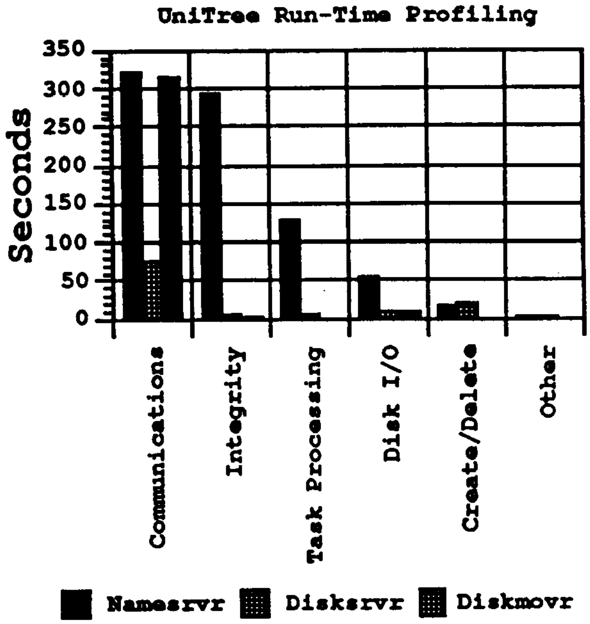
The UniTree performance of the Name Server, Disk Server, and Disk Mover daemons on the RS/6000 Model 550 showed performance at an average of 1311KB/sec when the test program was executed on the local host, and at an average of 594KB/sec when the test program executed on a remote host.



This test case was expanded further, and multiple processes were initiated on the UniTree local host to stress UniTree. Just the localhost was tested to eliminate the limitations of the network. There was a 46% drop in performance when the second process was added, and a 20% drop when each additional process was added.



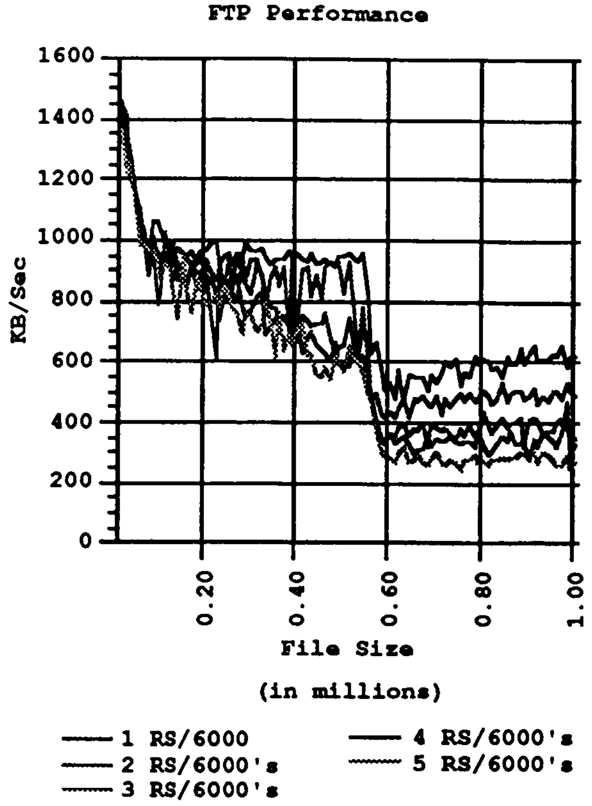
Using the above testing scenario with only one local process, the UniTree Name Server, Disk Server, and Disk Mover daemons were profiled. The daemons were categorized into the same six areas that the UniTree Name Server was categorized with the Entry Creation category broaden to include the functions the Disk Server uses to maintain the physical disk header map.

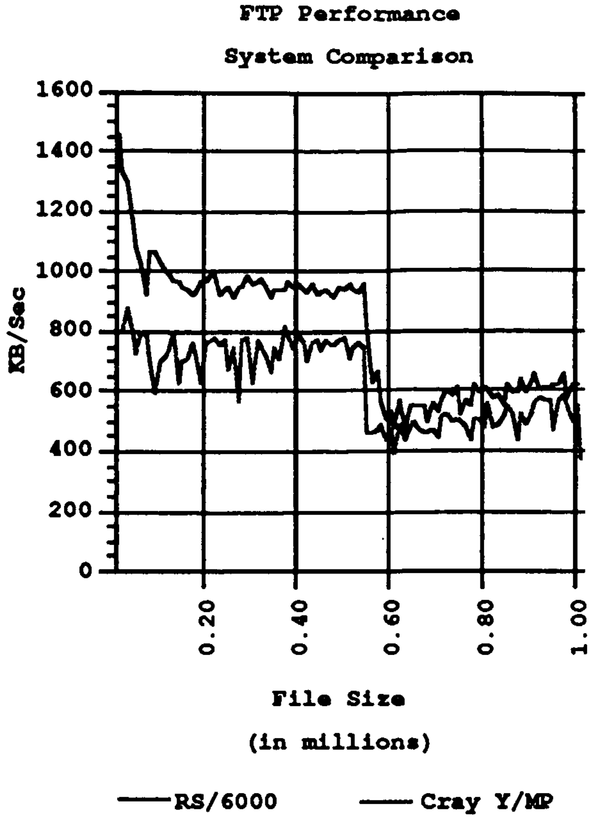


As with the UniTree Name Server profiling data, the categories Integrity and Communications show the highest execution usage.

FTP Performance

FTP clients were initiated on several RS/6000 Model 550 systems (connected via ethernet and on the same subnet) and directed to transfer increasingly large files into the UniTree system. Multiple instances of the FTP test programs were initiated and synchronized on separate systems to eliminate contention for system resources.





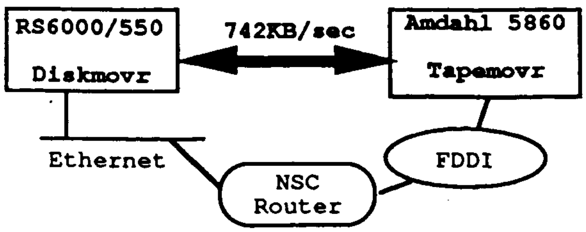
The performance data as cited by the FTP clients shows that there is a 15% degradation in performance as each additional client is added. However, the overall aggregate performance increases almost linearly with each additional client.

An FTP session was initiated on a Cray Y/MP to allow for a performance comparison between the RS/6000 and the Cray Y/MP.

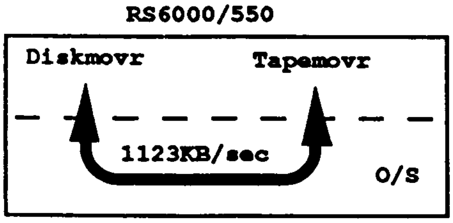
The Cray Y-MP shows comparable performance to the RS/6000 Model 550. The Cray Y-MP was tested while in a production, while the RS/6000 Model 550 was in a dedicated mode. The Cray Y-MP FTP session interfaced with UniTree through an FDDI and ethernet network.

Distributed UniTree Performance

In the NCSA distributed environment, the tape and the disk daemons of UniTree reside on physically separate hosts. The observed performance of the caching and migration of files between the disk daemons on the RS/6000 Model 550 and the tape daemons on the Amdahl 5860 was 742KB/sec.



Observed performance between the tape and the disk daemons when both reside on the same host was significantly faster - 1123KB/sec.



System Scalability

How well does the departmental server scale? Installation of multiple instantiations of the departmental Disk Server as seen in Appendix Figure II result in a disjoint namespace problem. Users do not have location independent file access capabilities under such a configuration. A user creating a file "foo" on archive server A would not be able to access "foo" if he or she were presently using server B for their archiving service. One method by which this problem could be circumvented would be to configure a global nameserver for use by both Disk Server A and Disk Server B (Appendix, Figure III). This configuration has been tested and was deemed functional. However, the UniTree servers lack some necessary intelligence when performing FTP operations and file attribute fetches. For example, the client must pass the address of it's Disk Server to the name server when requesting file attribute data such that the name server could fetch the information from the appropriate Disk Server. Disk Server addresses could be registered in a system configuration file. In summary, the servers would need non-trivial customized addressing enhancements in order to make this distributed system fully functional.

These customized enhancements are not , however, the correct approach to resolving the deficiencies in scalability. A scalable filesystem interface tightly integrated with the archive filesystem would be an effective way to solve the system scalability problem. This integration effort will be the focus of ongoing studies and software development efforts at NCSA.

Summary

The RS/6000 performed well in our test environment. The potential exists for the RS/6000 to act as a departmental server for a small number of users, rather than as a high speed archival server. Multiple UniTree Disk Server's utilizing one UniTree Name Server could be developed that would allow for a cost effective archival system.

Our performance tests were clearly limited by the network bandwidth. The performance gathered by the LibUnix testing shows that UniTree is capable of exceeding ethernet speeds on an RS/6000 Model 550. The performance of FTP might be significantly faster if asked to perform across a higher bandwidth network.

The UniTree Name Server also showed signs of being a potential bottleneck. UniTree sites that would require a high ratio of file creations and deletions to reads and writes would run into this bottleneck. It is possible to improve the UniTree Name Server performance by bypassing the UniTree LibUnix library altogether and communicating directly with the UniTree Name Server and optimizing creations.

Although testing was performed in a less than ideal environment, hopefully the performance statistics stated in this paper will give end-users a realistic idea as to what performance they can expect in this type of setup.

UniTree Archive Server

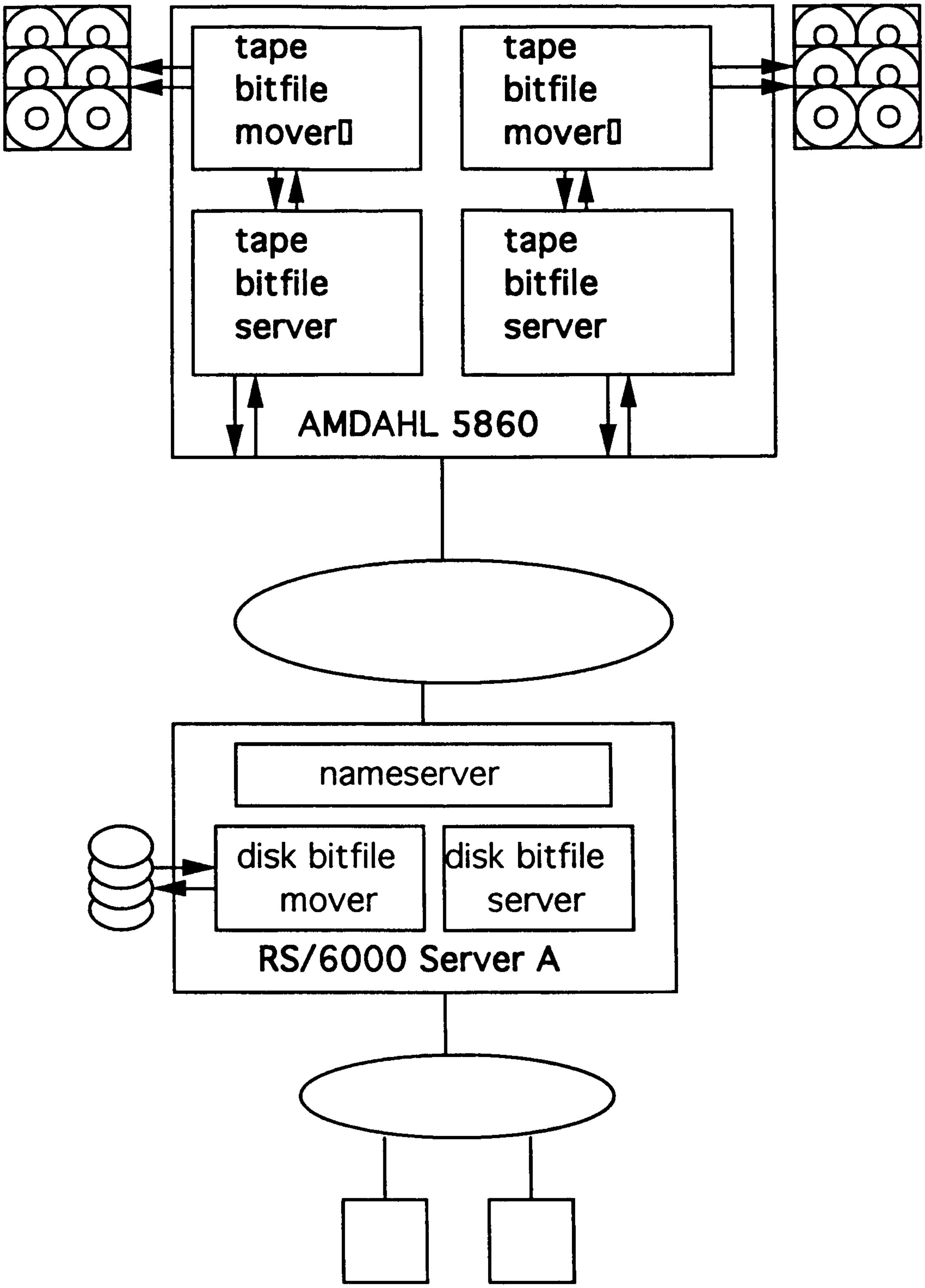


Figure I

UniTree Archive Server

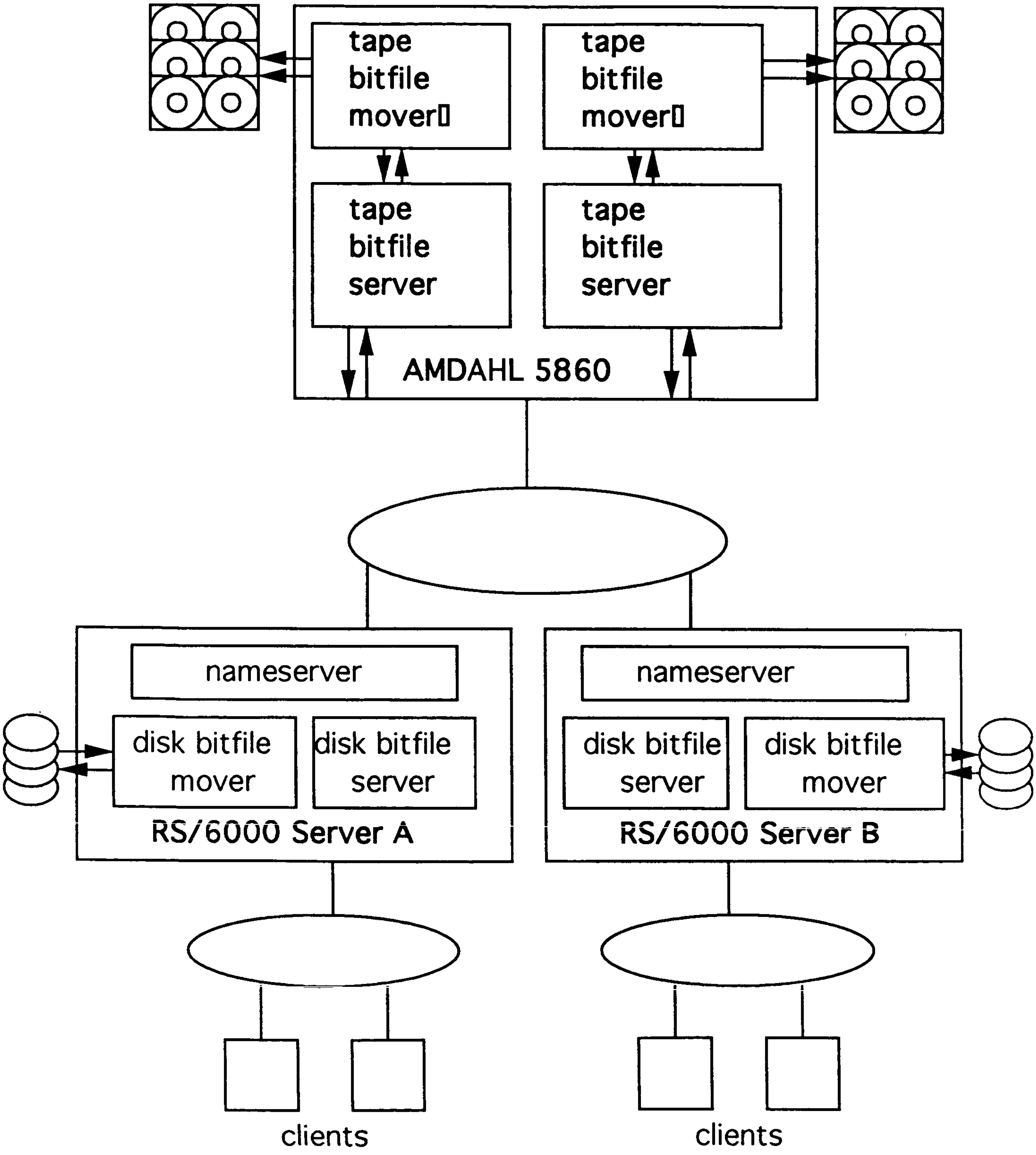


Figure II

UniTree Archive Server

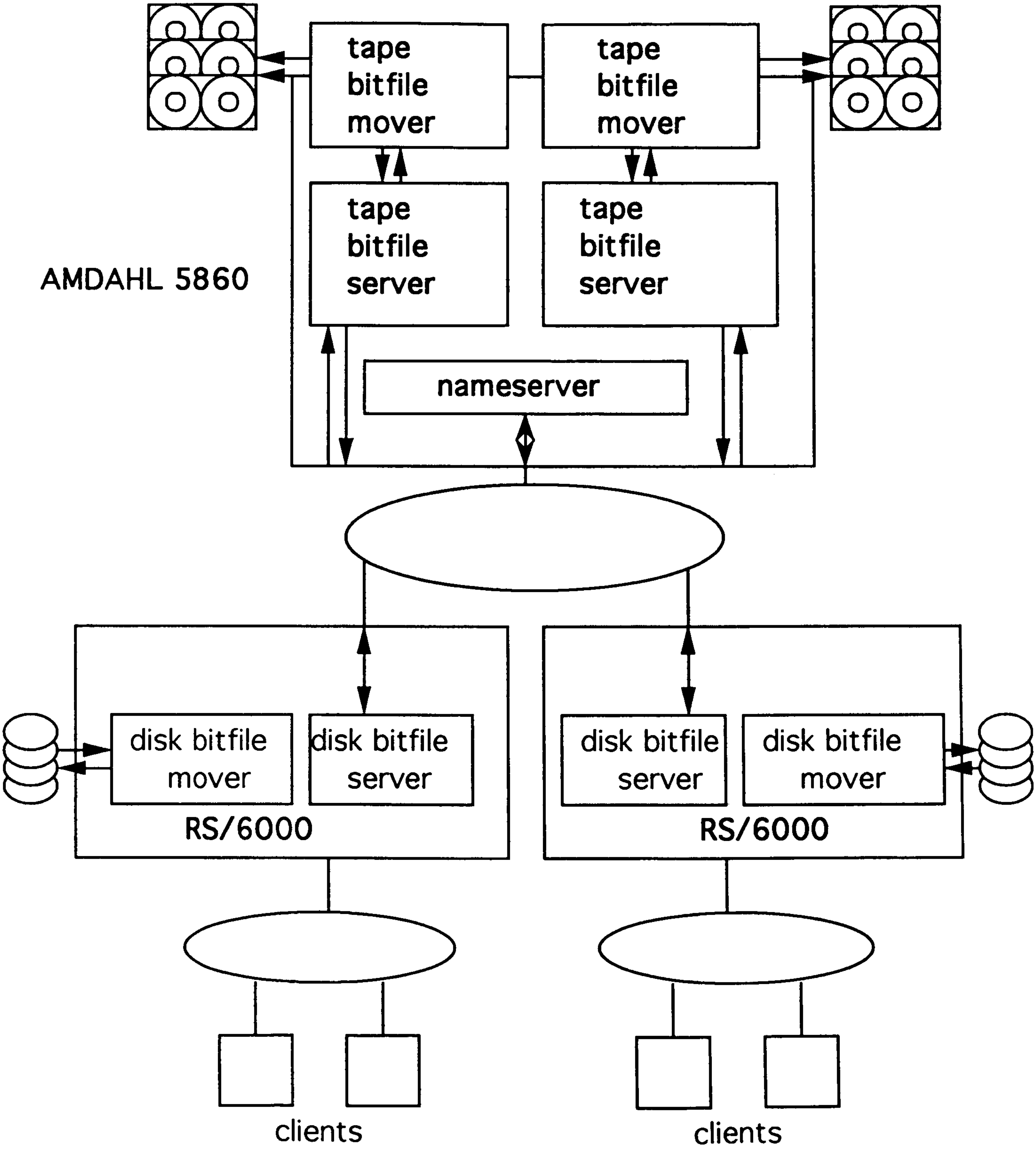


Figure III