File Servers, Networking, and Supercomputers

Reagan W. Moore

San Diego Supercomputer Center
San Diego, California

Abstract

One of the major tasks of a supercomputer center is managing the massive amount of data generated by application codes. A data flow analysis of the San Diego Supercomputer Center is presented that illustrates the hierarchical data buffering/caching capacity requirements and the associated I/O throughput requirements needed to sustain file service and archival storage. Usage paradigms are examined for both tightly-coupled and loosely-coupled file servers linked to the supercomputer by high-speed networks.

Introduction

The file server capacity requirements are most strongly driven by the CPU power of the central computing engine. The workload that can be sustained on the supercomputer is ultimately limited by the ability to handle the resulting I/O. At the San Diego Supercomputer Center, the central computing resource is a CRAY Y-MP8/864 supercomputer with a peak execution rate of 2.67 Gflops, capable of generating up to

2.67E+9 operations/sec * 8 Bytes/operation * 86,400 sec/day

or 1.8 Petabytes per day. In practice, the actual data generation rate is determined by the workload characteristics. The two major sources of I/O are application disk I/O and job swapping to support interactive use. At SDSC, the batch load averages 8 GBytes of executable job images, while the interactive load peaks at 120 simultaneous users. The batch load is sufficiently large that the idle time on the supercomputer has averaged 1.5% of the wall clock time over the last 6 months. While maximizing CPU utilization has been an explicit goal at SDSC, this has also increased the total amount of data that must be manipulated.

Data Flow Analysis

Not all generated data are archived and not all archived data are saved forever. A data flow analysis is necessary to understand the characteristics of the I/O, including the amount of data actually generated, the length of time over which the data are accessed, and the rate at which the data are moved through multiple caching levels. A simple analysis of the data flow can be used to illustrate the results of changing data access methods, increasing processing power, or improving network bandwidth. In particular, the data flow patterns are expected to be different for loosely-coupled user-initiated file archiving than for automated file servers tightly-coupled to the supercomputer CPU.

Solid State Storage Device Cache

The current archival storage system in use at SDSC is DataTree which supports user-initiated file archiving. This system acts as the archival storage file server for the CRAY supercomputer and is accessed through a 100 Mbits/sec FDDI ring. Data generated on the CRAY Y-MP8/864 are ultimately stored on 3480 cartridge shelf tape. There are five levels of I/O buffering or caching, including a 1 GByte Solid State Storage Device, 42 GBytes of CRAY disk local to the supercomputer, 70 GBytes of archive disks, a 1.2 TByte tape robot, and 2 TBytes of manually mounted shelf tape. Table 1 illustrates the caching hierarchy. As expected, the amount of data moved towards the lowest archival storage level decreases as the required storage life of the data increases. Data resides on the SSD for periods on the order of minutes, on CRAY disk for up to two days, on archival storage disk for up to several weeks, in the tape robot for several months, and finally on shelf tape for years. The amount of data moved per day between each level varies from 1.5 TBytes/day through the SSD to CRAY disk, 14 GBytes/day through archival storage, 9 GBytes/day through the tape robot, and 2 GBytes/day to shelf tape. The residency time at any level may be estimated by dividing the size of the cache by the input I/O rate to the cache. This closely matches measured data residency times.

The SSD serves both as a data cache for the /root file system and the interactive swap space and as a data buffer for the large 42 GByte /usr/tmp file system. Caching versus data buffering depends on the amount of data reuse. The caching of /root to support interactive users is effective since a hit ratio exceeding 99% can be sustained when the cache size is set to 68 MBytes. Data caching for the interactive swap space is effective when about three MBytes of swap space is reserved per user. The actual interactive swap partition at SDSC is 320 MBytes on the SSD and is restricted to supporting job sizes less than 8 MBytes. Since the total SSD size is 1 GByte, there is not enough room to cache the 42 GByte /usr/tmp file system. Instead the /usr/tmp data effectively stream through the SSD with minimal reuse. The net effect is that the SSD buffers 196 kByte disk data reads for 32 kByte accesses by the application codes. This helps minimize the amount of time spent waiting on disk seek latencies. Buffering of /usr/tmp files dominates the I/O rate needed to support swapping of interactive jobs by a factor of 2.5. Although the SSD transfers data at over 1 GByte/sec, the steady state I/O rate needed to support streaming 1.5 TBytes of data per day through the SSD is only 17 MBytes/sec. Replacing with a slower speed communication channel would seriously degrade interactivity. Swapping jobs at the average transfer rate would require up to one/half second to load an interactive job into memory. Thus the dominant I/O support requirements for the SSD are split between providing a large storage area for data buffering and providing very high-speed access for the interactive job swapping data subset.

Local CRAY Disk Cache

The CRAY disks also sustain a total amount of I/O of about 1.5 TBytes per day, or an average of 17 MBytes/sec. Since the total /usr/tmp disk space is only 42 GBytes, the majority of this I/O is to scratch files which disappear at problem termination. This can be calculated using the average batch job execution time of one hour and the write rate to disk being one fourth the read rate. If all the generated data were saved, only about three hours of CRAY execution data could be stored on local CRAY disk before they would have to be migrated elsewhere. In practice, the files reside much longer on disk. Typically 60% of the disk files are up to one day old, and another 25% are up to two days old. The average residency time is about 30 hours, implying that only one tenth of the data written to disk survives application code termination. The CRAY disks therefore are serving as a cache for writing data from the supercomputer.

Archival Storage Disk Cache

The true long-term data generation rate is governed by how fast data are migrated to archival storage. On the DataTree archival storage system in use at SDSC, archiving of files is a user initiated process. Users explicitly choose which files to archive or retrieve. Typically 14 GBytes/day of data are transferred between the CRAY disks and the archival storage system of which one third is data written to storage. This amount of data flow is only 1/7 of that needed to migrate the data that survive on CRAY disk to archival storage. Thus about 1.4% of the total amount of data written to CRAY disk is archived. The archival storage disks form an effective cache between long-term storage on cartridge tape within the tape robot and the CRAY local disks. The hit ratio for archival storage data being retrieved from the archival storage disks is typically 92%.

Archival Storage Tape Caches

The average data transfer rate needed to support archival storage is 0.16 MBytes/sec. This should be compared with the observed sustainable archival storage data rates of 0.6 MBytes/sec supported by DataTree running on an Amdahl 5860 across 4.5 MBytes/sec I/O channels connected to a 12.5 MBytes/sec FDDI backbone network. During periods of heavy usage, the average transfer rate does approach the peak rate.

Long term archival storage to tape occurs both directly from the CRAY disk for large files (sizes greater than 200 MBytes) and by automatic data migration from the archival storage disks. The tape robot serves mainly as a data cache. Data currently reside about 15 months before migrating to shelf cartridges. Data caching attributes can be tracked by the fraction of tape mounts done manually. Typically the 1.2 TByte tape robot processes 85-90% of the tape mounts. The rate at which data are migrated from the tape robot to shelf tape is roughly 2/3 of the rate at which data are written to the robot. This ratio may approach one as data in the robot mature.

This data flow analysis demonstrates some interesting attributes of loosely-coupled user-initiated archival file storage systems.

> 10% of the generated data is stored temporarily on CRAY local disk, 1.4% of the generated data is written to archival storage, and 0.6% of the generated data is eventually transferred to long term shelf tape. Given the need to explicitly save files, users selectively store a fraction of their output.
>
> The multiple levels of the storage hierarchy serve mainly as caches with more data flowing into a given cache than flows out to lower caching levels.
>
> The amount of data read at each caching level is substantially higher than the amount written with the ratio varying from 4:1 for the highest speed cache on the SSD down to 2:1 for archival tape storage.

The above data flow analysis is typical only of user-initiated archival storage. If an automated archival storage scheme is used for supporting the CRAY disks, the amount of data that are archived could grow substantially. This can seriously impact the ability to adequately handle the I/O if the archival storage hardware environment is operating with relatively small safety margins. Pertinent safety factors are:

> cache residency time versus the latency that a data buffer is amortizing,

cache residency time of data files on local CRAY disks versus the time needed for the application to complete, and

sustainable I/O rate versus the peak I/O demand rate.

If any of these factors drop below one, the system will become severely congested and may even fail. At SDSC, all of these safety margins are relatively small. Due to the limited amount of CRAY local disk space, the residency time of files on CRAY disk is comparable to the wall clock time needed to complete an application run for large codes. The weekly average required I/O rate to access files on the archival storage system is 1/4 of the peak observed sustainable rate. Hourly averages of the required I/O rate approach the peak sustainable rate. A usage paradigm shift that increases the I/O load could seriously stress the archival storage system at SDSC.

File Server Paradigm Shifts

Three possible usage paradigm shifts are being investigated at SDSC, two of which are related to file servers tightly coupled to the supercomputer CPU power. The first is a research project funded by the National Science Foundation and DARPA through the Corporation for National Research Initiatives. Prototypes of tightly coupled applications distributed across supercomputers connected by a gigabit/sec network are being developed, including the linkage of an application to the equivalent of a database interface to archived data. The second is a project to investigate the feasibility of incorporating the local CRAY disk and the SSD as caches directly controlled by the archival storage system. The third is the modeling of the impact on the archival storage system of an upgrade to a 100 Gigaflop/sec supercomputer.

High-speed Remote Access

The CASA Testbed is a collaborative effort between the California Institute of Technology, the Jet Propulsion Laboratory, the Los Alamos National Laboratory, and the San Diego Supercomputer Center. One objective is to demonstrate a distributed application efficiently utilizing two supercomputers while simultaneously using a substantial fraction of the gigabit/sec wide area network linking the computers. Simultaneously maximizing bandwidth utilization and CPU utilization requires minimizing the protocol overhead used for the data transmission[1]. The effective bandwidth for the optimal application is given by

$$B / (1 + O * B)$$

where B is the peak bandwidth (bits/sec) and O is the network protocol overhead measured in seconds of overhead per bit transmitted. For high speed networks, network protocol overhead becomes a critical limiting parameter. For present CRAY supercomputers, the network protocol overhead can require the execution power of an entire CPU to support TCP/IP at 700 Mbits/sec.

Given that a suitable file transport protocol is devised with a small enough protocol overhead, the issue of latency across wide area networks may be the next limiting factor. Since the speed of light is finite, data access delays between SDSC and LANL are as great as disk seek times. Efficient access of remote file systems must then cope with buffering data in addition to caching data. The amount of data shipped between an application and a remote database interface to archival storage must be large enough to amortize the data access delay. Depending on the protocol, the amount of data sent may need to be as large as

$$2 L * B$$

where L is the round trip latency measured in seconds. For a LANL/SDSC application running at 800 Mbits/sec, this is still feasible, requiring buffering on the order of 8 MBytes.

Integrated Local and Archival File Systems

Integrating the local file system into the archival storage file system will substantially increase the amount of data that must be processed by the archival storage software. As seen in the SDSC data flow analysis, the amount of data transferred between the supercomputer and the local disks is more than a factor of 1000 larger than the amount transferred to archival storage. Efficiently handling this increase in data rates will require differentiating between "reliable" local file transport and "unreliable" transport across a local network. By scaling the network protocol overhead needed to support TCP/IP at 700 Mbits/sec by the average CRAY local disk bandwidth derived in the data flow analysis, an estimate can be made of the protocol overhead increase. With no protocol enhancements, an additional 20% of a single CPU would be needed to support the archival and local file system integration. This indicates the need for the integrated system to recognize heterogeneous network environments.

An additional complication is that if all of the generated data stored temporarily on CRAY disk is automatically archived, the data flow from local CRAY disk to archival storage could increase by up to a factor of seven. Files written to the scratch /usr/tmp file system require different backup than files written to permanent home directories. An integrated local file system and archival storage file system must allow for a non-uniform usage pattern.

CPU Execution Rate Dependence

A possible ameliorating effect is that as supercomputers become faster, it may become more cost effective to recompute rather than save data. A supercomputer with a sustained execution rate of 100 Gigaflops is expected to be available by 1995. Assuming the data storage patterns remain the same, the I/O generated by such a machine can be estimated by scaling the results of the data flow analysis by the increase in the execution speed, which is roughly a factor of 3000. The cache sizes and I/O communication rates then become:

SSD	3 TBytes	50 GBytes/sec
Local disk	126 TBytes	50 GBytes/sec
Archive disk	210 TBytes	450 MBytes/sec
Shelf tape	6000 TBytes	60 MBytes/sec

The archival storage communication rates need to be decreased by a factor of 10 to become technically feasible. Thus a paradigm shift towards the dynamic regeneration of simulation output may become inevitable.

Acknowledgement

This work was funded in part by the National Science Foundation under Cooperative Agreement Number ASC-8414524 and Grant Number ASC-9020416.

References

1. Moore, Reagan, "Distributing Applications Across Wide Area Networks," General Atomics report GA-A20074, April 1990.

Table 1

Hierarchical Data Caching Levels

Caching Level	I/O per Day	Data Rate	Capacity	Utilization	Residency Period
SSD	1.5 TB	17 MB/s	1 GB	85-100%	minutes
CRAY Disk	1.5 TB	17 MB/s	42 GB	85-90%	days
Archive Disk	5 GB	0.05 MB/s	70 GB	98%	weeks
Tape Robot	9 GB	0.10 MB/s	1.2 TB	68%	months
Shelf Tape	2 GB	0.02 MB/s	2 TB	70%	years

File Servers, Networking, and Supercomputers

Reagan W. Moore

San Diego Supercomputer Center
San Diego, California

Archival Storage Systems as File Servers

- **Examine Hierarchical Caching systems**
  - Capacity requirements
  - I/O requirements
- **Based on Usage at SDSC**
  - Archiving supercomputer generated data

File System Usage Paradigms

- **Loosely-coupled to CPU**
  - User initiated file transfers to archival storage
- **Tightly-coupled to CPU**
  - NFS access
  - Integrated local and archival file systems

SDSC Archival Storage Environment

- **Data Generated by CRAY Y-MP8/864 Supercomputer**
- **FDDI 100 Mbits/sec backbone**
- **DataTree Archival Storage System on an Amdahl 5860**

Five Levels of Data Caching

- **Solid State Storage Device (SSD)**
  - 1 GB, 1.2 GB/s access from memory
- **CRAY local disk**
  - 42 GB, 10 MB/s access per disk
- **Archive storage disk**
  - 70 GB, 0.6 MB/s access across FDDI
- **STK tape robot**
  - 1.2 TB, 0.6 MB/s access across FDDI
- **Shelf cartridge tape**
  - 2 TB, 0.6 MB/s access across FDDI

SDSC Workload Characteristics

- **Application Disk I/O**
  - Generated by an average batch load of 8 GBs of executable jobs
- **Job Swapping**
  - Generated by up to 120 interactive users
- **User-initiated File Archiving**
  - Partial archiving of supercomputer data

Data Flow Analysis

- **Track Data Through the Multiple Caches**
  - Cache utilization
    - Hit rate
  - I/O throughput
    - Fraction of peak rate
  - File residency time
- **Identify Caching versus Data Buffering**

SDSC Data Flow

Cache Level	Capacity (GB)	Utilization
SSD	1	85%
CRAY disk	42	90%
Archive disk	70	98%
Tape robot	1200	68%
Shelf tape	2000	70%

SDSC Data Flow

Cache Level	Residency Time	Fraction saved of total I/O written from SSD
SSD	(seconds)	100%
CRAY disk	30 hours	10%
Archive disk	4 weeks	1.4%
Tape robot	15 months	1.4%
Shelf tape	5 years	0.6%

SDSC Data Flow

Cache Level	I/O per Day (GBytes)	Data Rate (MBytes/sec)
SSD	1500	17
CRAY disk	1500	17
Archive disk	5	0.05
Tape robot	9	0.10
Shelf tape	2	0.02

Data Caching Versus Data Buffering

- **SSD Cache Used for Both**
  - /root file system and Interactive swap space are cached
    - Hit rate for accesses is 99%
  - /usr/tmp file system is buffered
    - Hit rate for accesses is 75-85%

File Server Safety Factors

- **Cache Residency Time versus Latency Amortization Time**
- **Cache Residency Time versus File Usage Time**
- **Sustainable I/O Rate versus Peak I/O Demand Rate**

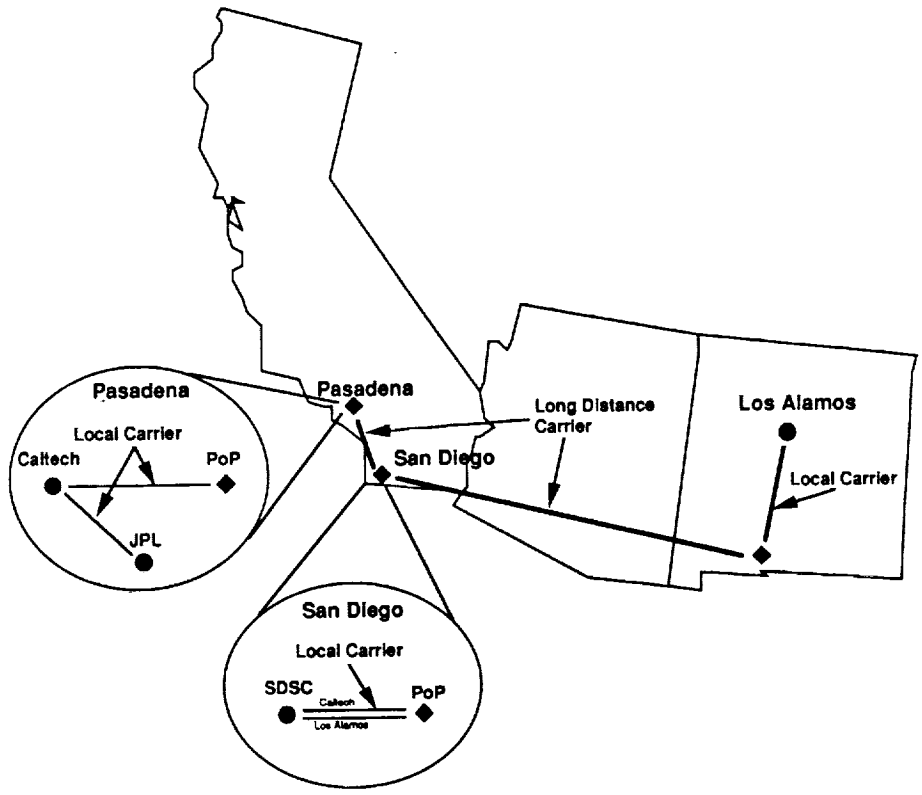
File Server Paradigm Shifts

- **Changes in Funtionality May Require Usage Paradigm Shift**
  - High-speed remote access
  - Integration of local and archival file systems
  - Very high-speed supercomputers

CASA Gigabit/sec Testbed

- **Collaboration between CalTech, JPL, LANL, SDSC**
- **Demonstrate Tightly Coupled Distributed Applications Linked by Gigabit/sec Wide Area Network**
  - Remote access of archived data through database interface

CASA GIGABIT WAN



Network Protocol Overhead Impact

- **Simultaneous Optimization of CPU and Bandwidth Utilization**
- **Effective bandwidth is given by**
  - B / (1 + O * B)
  - B = bandwidth (bits/second)
  - O = protocol overhead (seconds/bit transmitted)

Protocol Support Limitations

- **TCP/IP Protocol Can Require Execution Power of Entire CPU of Y-MP8/864 for 700 Mbits/sec Bandwidth**
- **Effective Bandwidth Reduced 45%**

Wide Area Network Latency Can Require Data Buffering in Addition to Data Caching

- **Finite Speed of Light Creates Latency Between SDSC and LANL Comparable to Disk Seek Latencies**
- **Amortize Latency by Shipping Large Files**
  - Size = 2 L * B
    - L = Round-trip latency (seconds)
  - For 800 MBits/sec network, ship 8 MB files

Integration of Local and Archival File Systems

- **Local CRAY Disk Supports 1000 Times as Much Data Transfers as Archival Storage at SDSC**
- **To Minimize Protocol Overhead**
  - Distinguish between
  - "Reliable" local file transport
  - "Unreliable" local network transport
  - Otherwise expect overhead to increase 20%

Integration of Local and Archival File Systems

- **User-initiated File Archival Storage Results In**
  - 1/7 of the data being archived
- **Automatic Migration of Local Files**
  - Allow non-uniform file migration across different file systems
    - /root versus /usr/tmp

Supercomputer I/O Scaling

- **For a 100 Gflops/sec Supercomputer**
  - Scale I/O by ratio of CPU speeds
  - Expect 3000 times as much I/O
- **Massive data generation may require dynamic regeneration of data rather than storage**

CPU Execution Scaling

Cache Level	Capacity (TB)	Data Rate (MB/s)
SSD	3	50,000
CRAY disk	126	50,000
Archive disk	210	450
Tape robot	6000	60
Shelf tape	12000	60

File Server Paradigm Shifts

- **Data Storage Requirements Will Be Increased by**
  - Integration of Local and Archival Systems
  - Higher Speed Supercomputers
- **Possible Shifts**
  - Local regeneration of data
  - Remote File Access