Filesystems and I/O Balance on the NERSC T3E

Tina Butler, NERSC Systems Group

Abstract

NERSC's 696 processor T3E-900, mcurie, supports approximately 900 active DOE-affiliated users with an emphasis on large-scale, data-intensive applications. We will discuss the configuration of user file systems and archival storage that supports these computational and IO workloads, and examine usage and retrieval patterns on both the 150 GB DMF-migrated home file systems and 1.5 TB of non-migrated temporary storage areas.

1. Introduction

Although many people only look at the CPU speed of a processor to determine the "power" of a computing system, the CPU can only perform productive work when it is supplied with data in an expedient fashion. The true power of a system lies in the balance between its processor speed and its ability to move data efficiently to and from different levels of storage in the system as a whole. This balance can be particularly difficult to achieve in a general-purpose system that must support a wide variety of users that are using the system for a number of different tasks. The I/O capacity and bandwidth needs are very different for a user doing large-scale production climate modeling and a user in midst of code development.

2. NERSC and the DOE user community

NERSC is the National Energy Research Scientific Computing Center and is funded by the U.S. Department of Energy, Office of Science. It is located at Lawrence Berkeley National Laboratory in Berkeley, California. NERSC has a 25-year history of providing high performance computing to the DOE community. It was founded at the Lawrence Livermore National Laboratory in 1974 as the Magnetic Fusion Energy Computer Center and moved to Berkeley in 1996. The center provides computational resources to DOE programs in the fields of Fusion Energy, High Energy and Nuclear Physics, Basic Energy Sciences, Biology and Environmental Research and Computational and Technology Research. NERSC currently serves approximately 2500 users from major universities and government laboratories across the country on its Cray T3E and SV1 production systems. Approximately 900 NERSC users have active accounts on the NERSC T3E.

3. T3E history/configuration

NERSC acquired its first Cray T3E, mcurie, in June 1996. This original system has gone through a series of upgrades and configuration changes in the years since then. A brief history of NERSC T3E hardware is listed in Table 1. Pierre was a second, smaller system in general use until it was merged with mcurie in October of 1998.

Table 1: History of T3E Hardware at NERSC

System Name	Type	PEs	APP PEs	Date	Comment
Mcurie	T3E-600	136	128	9/96	Initial System
Pierre	T3E-600	104	96	12/97	Initial System
Mcurie	T3E-900	544	480-512	8/97	Phase II
Pierre	T3E-900	152	128	6/98	Upgrade
Mcurie	T3E-900	696	644	10/98	Merge with Pierre

In its current configuration (Figure 1), mcurie is a T3E-900 with 696 processing elements (PEs), each with 256 MB of memory and 22 Gigarings. The PEs are allocated with 644 application PEs, 36 command PEs, and 15 OS PEs. Mcurie is connected to the network through FDDI, Ethernet, and HIPPI interfaces. Mcurie has 12 Fibre Channel I/O nodes (FCNs) with 68 DA308 disk arrays providing 2,606 GB and 8 MPNs with 612 GB of non-RAID DD314 disks.

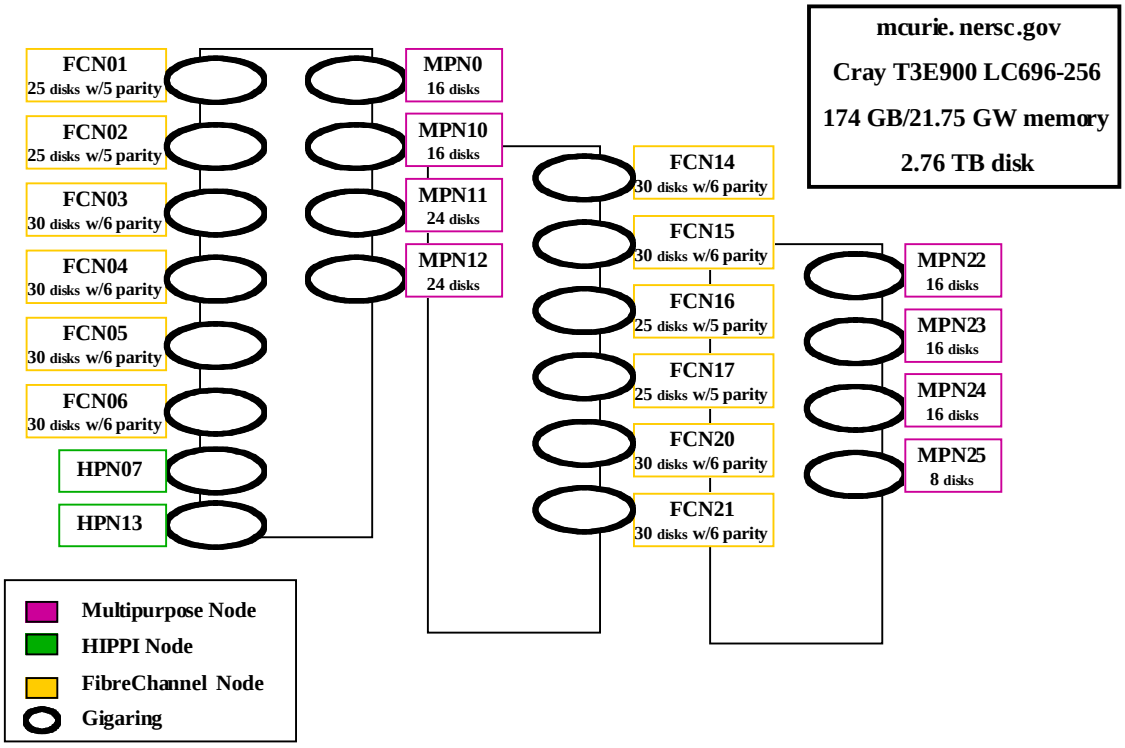


Figure 1.

4. Workload

The NERSC T3E is used for application development, medium-sized capacity-based computing and large-scale "capability" problems. User codes include applications from chemistry, materials science, fusion research, geophysics, high-energy nuclear physics, biology, climate modeling, astrophysics and fluid dynamics. The data shown in Table 2 indicate that the current workload is diverse and dynamic. A large number of small, short-running development applications are run on the system; it supports a steady flow of medium-sized applications; and 9 percent of the total machine resources have been consumed by applications requiring more than 128 processing elements.

Table 2: Workload Characterization

App Size (PEs)	% of All Apps	% of PE Hours
2-16	55.8	6.5
17-64	38.1	55.7
65-128	4.8	28.7
129-512	1.3	9.1

App Run Time	% of All Apps	% of PE Hours
0-10 min	55.7	1.1
10-30 min	23.2	10.4
0.5-3.5 hr	16.9	48.7
3.5-12.0 hr	4.1	39.8

5. Filesystem size and configuration

Filesystems on the T3E have been designed to provide capacity, performance and resiliency for the applications run at NERSC. Table 3 shows major filesystems on mcurie. System filesystems (i.e., root, usr, adm, spool, etc), swap, checkpoint, and the /usr/tmp secondary partitions reside on the high-performance FCN disk arrays. The FCN arrays are fully alternate-pathed for resiliency. The root, usr, local, and opt filesystems are pcached to help mitigate the effects of high metadata lookup rates and the larger block size (16K vs. 4K) for the RAID arrays. Root, usr, and opt typically maintain 99%+ hit rates for pcache. In order to distribute the OS workload, as well as shortening paths to disk servers, the large file systems, including /usr/tmp and checkpoint, are controlled by "remote mount" file servers.

Mcurie has 383 GB of swap space (2.38 times APP memory), and a 582 GB checkpoint filesystem (3.61 times APP memory). Each is composed of 5 logical disk partitions with each partition 5 or 6-way striped across DA308 disk arrays. This configuration was designed for speed when checkpointing, and swapping. NERSC uses both checkpointing and gang scheduling to implement scheduling policy, so checkpoint and swap speed during rank switches is an important consideration. Through utilizing parallel checkpoint processes, the entire application region of 644 PEs can be normally be checkpointed in less than 5 minutes. I/O transfer rates up to 800 MB/sec have been observed for checkpoint.

Table 3. Mcurie Filesystems

Filesystem	Size (1k blocks)	Inodes
/dev/dsk/root	1280000	80000
/dev/dsk/usr	1440000	90048
/dev/dsk/usrtmp	1465085184	524384
j/dev/dsk/adm	2320000	131072
/dev/dsk/adm_sl	2400000	131072
/dev/dsk/dm	560000	35008
/dev/dsk/mail	208000	13056
/dev/dsk/local	10048000	262144
/dev/dsk/src	2920000	131072
/dev/dsk/spool	1760000	110016
/dev/dsk/spool_dm	560224	35072
/dev/dsk/dm_jrnl	960000	60032
/dev/dsk/opt	2960128	131072
/dev/dsk/dfscache	21014864	131072
/dev/dsk/chkpnt	610488192	32768
/dev/dsk/u	80000	5008
/dev/dsk/u1-u7	26448000	524288
/dev/dsk/dumps	26418048	65536
/dev/dsk/swap	401753088	

User home filesystems and /usr/tmp both make use of the primary/secondary feature of NC1 filesystems. This makes it possible to mirror the primary partitions where filesystem metatdata reside, as well as allowing different allocation strategies for large and small files. User home filesystems and primary partitions for the large /usr/tmp filesystem are on MPN disks, with space allocated for full mirroring in the near future. User home filesystems are designed to be a size that is manageable for filesystem dumps and restores, while allowing enough space to user needs. Home filespace is spread across seven filesystems which are each 25 GB; six, /u1 - /u6 are currently in use, with one in reserve. Currently each home filesystem has approximately 150 users assigned. Figure 2 shows the distribution of files (including offline migrated files) on /u1-/u6, with files sorted by size category. The /usr/tmp scratch file system is 1.5 TB, with seven primaries on MPN disk and 27 secondary partitions on FCN arrays.

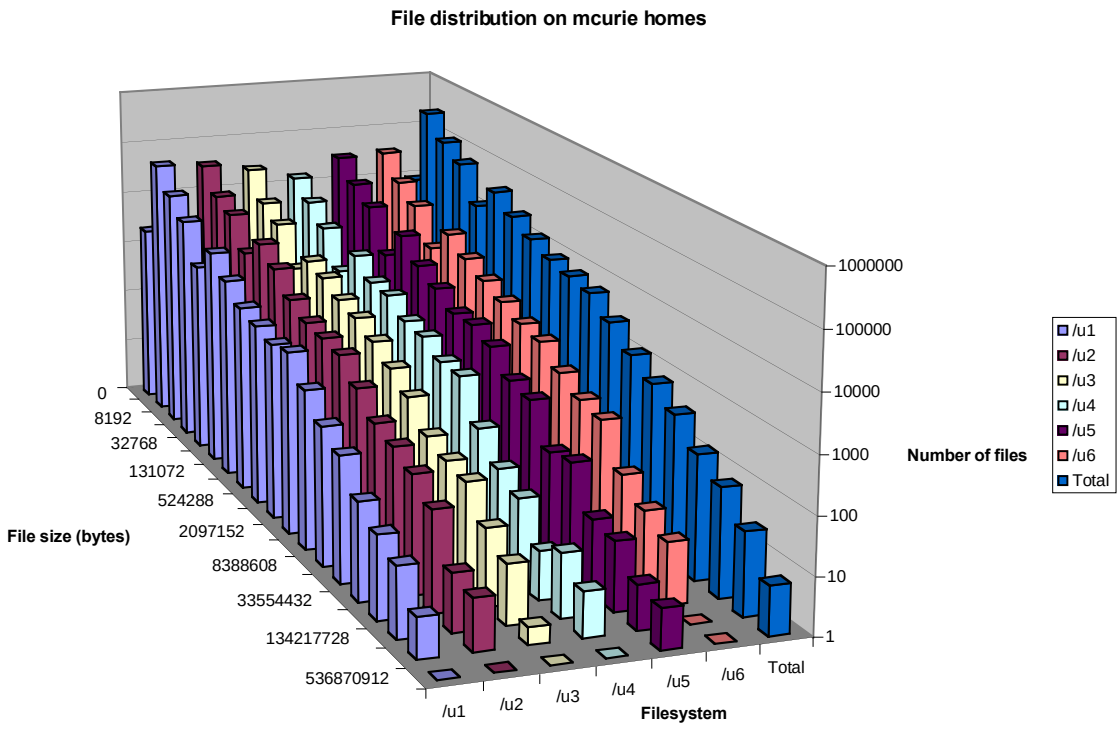


Figure 2.

6. Space monitoring and administration

NERSC uses both quotas and the Cray Data Migration Facility for automated monitoring and management of user-writable filesystems on mcurie. User homes, /usr/tmp, /usr/mail and /usr/spool all have hard quotas in order to control usage of both inodes and data blocks. The default quotas for the home filesystems are 3500 inodes and 4GB of file space. This has proved to be a reasonable number for most users. Most requests for exceptions have been for an increase in inodes. Higher values on /usr/tmp accommodate the larger files necessary for application input, output and intermediate files. Table 4 shows filesystem default quota values.

Table 4. Mcurie filesystem quotas

Filesystem	File quota (4k blocks)	Inode quota
/u1 - /u6	488280	3500
/usr/tmp	18310545	6000
/usr/spool	18310	100
/usr/mail	100	

Space on user home filesystems is also managed by the Cray Data Migration Facility. DMF was put into service on the T3E in October of 1998. NERSC uses HPSS as its general archival mass storage system. HPSS is accessible from all NERSC production systems, and is used by both users and systems for long term storage. NERSC Cray systems use HPSS as the destination for files migrated through the Data Migration Facility, and send filesystem backups there as well. Users have direct manual access to HPSS for true archival storage, as well as DMF-automated migration and retrieval.

7. Operating philosophy

The operating philosophy for mcurie is to encourage large-scale applications and projects, in terms of both compute cycles and large amounts of data. Part of the implementation of that philosophy is the

persistence of user files on the /usr/tmp filesystem. Space on /usr/tmp is managed by hard quotas and an automated purging script. The default quotas for /usr/tmp are 6000 inodes and approximately 70 GB of data.
User files are not purged until they have been idle for 14 days. This allows users to preserve intermediate data files between runs, or to do a number of runs against the same input data set. Data persistence reduces the number of transfers to and from HPSS required, decreases turnaround time, and thereby increases a researcher's productivity. Exceptions for both quotas and access time have been put in place for users with special projects or requirements. Currently the largest exception on mcurie is a group quota of 360 GB on /usr/tmp.

8. Filesystem traffic and usage

Patterns of filesystem usage for the last twelve months show the success of these strategies. Figure 4 shows the aggregate (read and write) I/O volume for each of the six active home filesystems from October 1998 to mid-August 1999. Figure 5 shows the read and write volume for all home filesystems for the same period. Through most of the year, the I/O load is evenly distributed across the home filesystems. In June, one user repeatedly ran an application that was accessing data from his home directory on /u4 rather than using the larger /usr/tmp filesystem. The prominent peaks in Figure 4 and Figure 5 reflect prolonged, sustained transfers of about 3 MB/sec on /u4 over 2-3 days.

Figure 4.

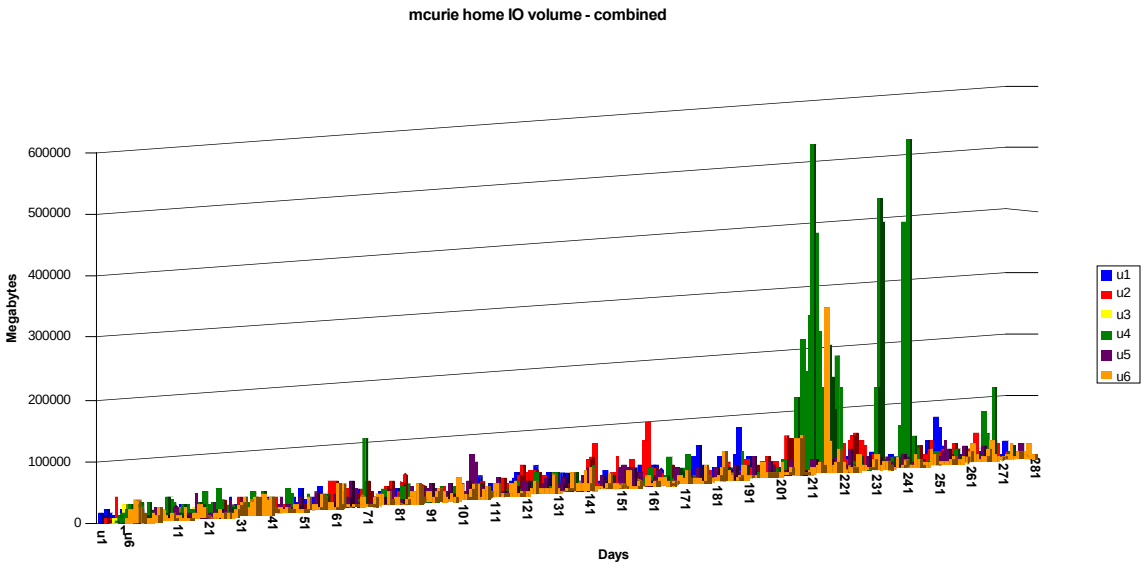


Figure 5.

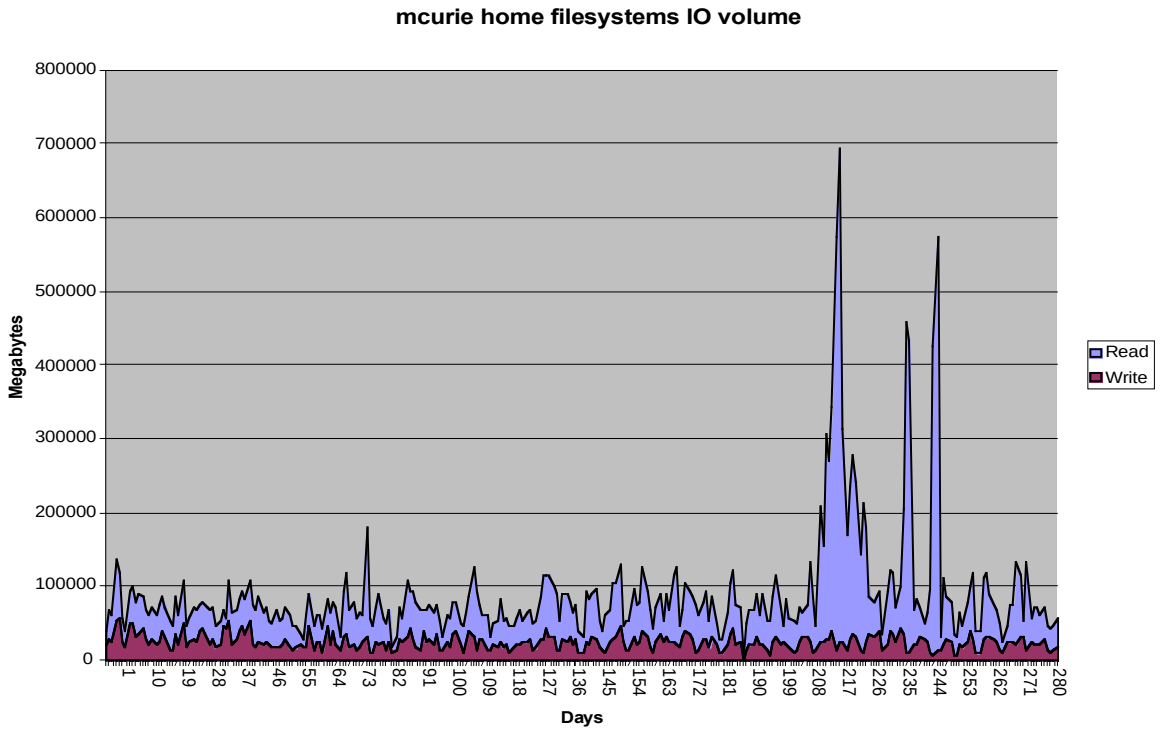
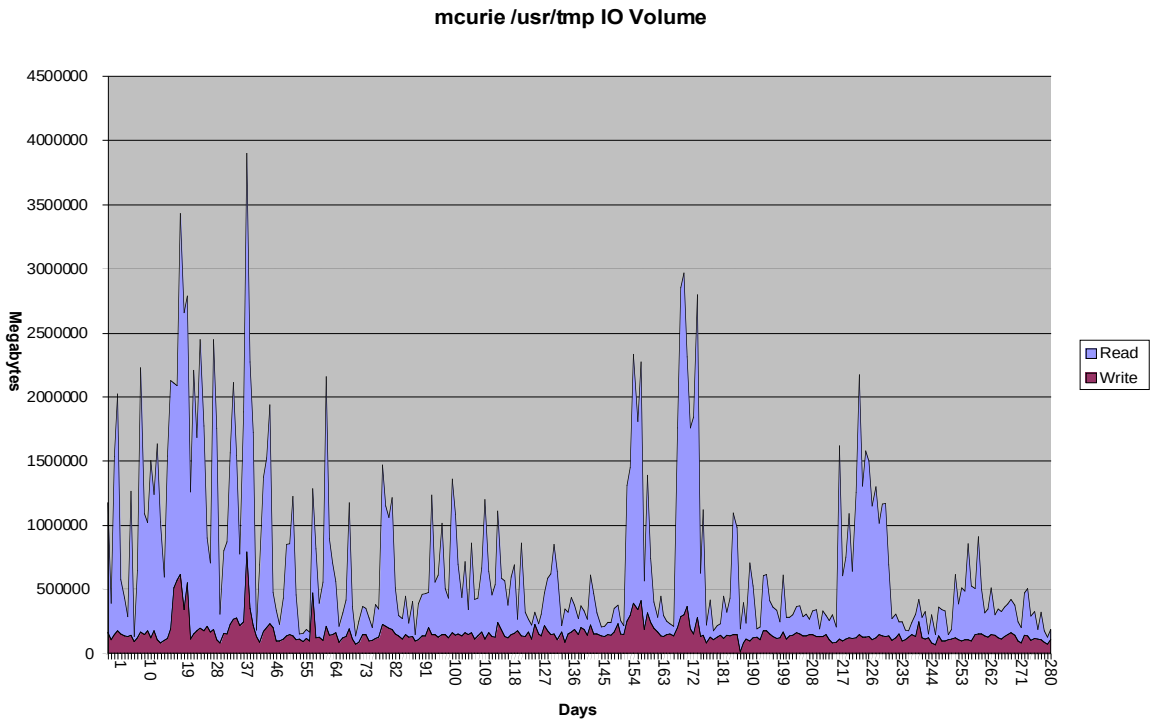


Figure 6 shows similar data on the amount of user data read and written on /usr/tmp.

Figure 6.



9. Data Migration traffic

Figures 7 and 8 show DMF traffic for the last 12 months on mcurie. Although there has definitely been a trend of increasing usage, DMF is not stressed in any way. The disproportion between puts and gets indicates that DMF has been a useful relief valve for the home filesystems, but homes are still well-sized for existing NERSC users.

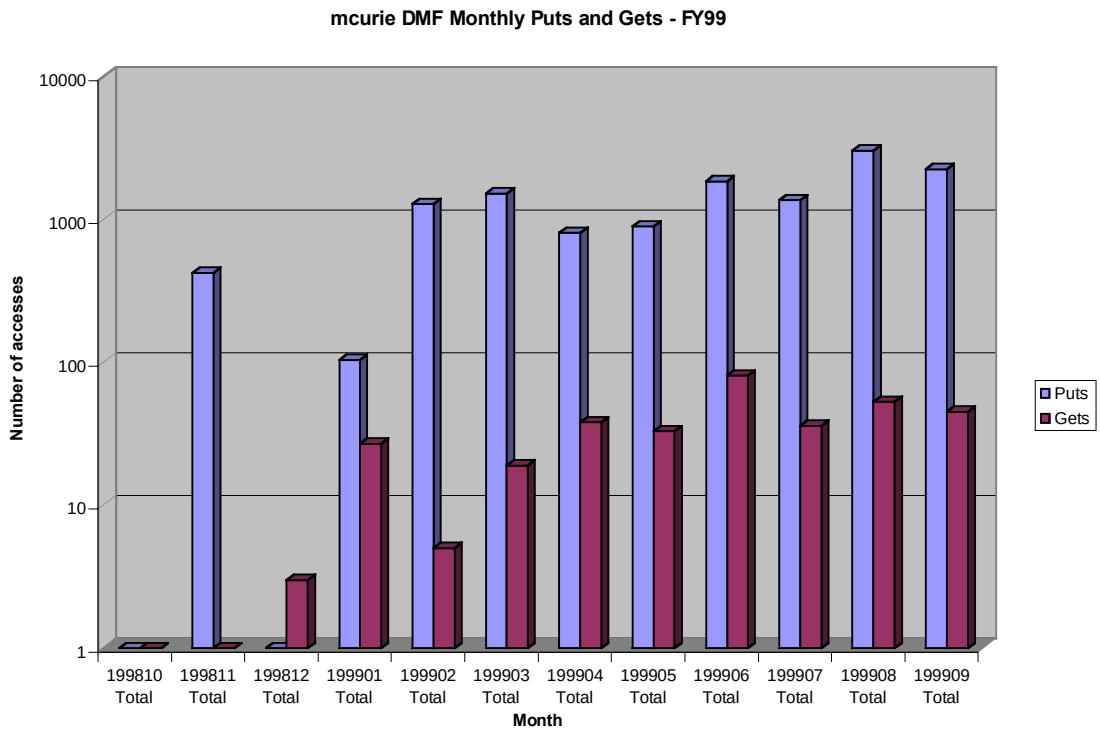


Figure 7.

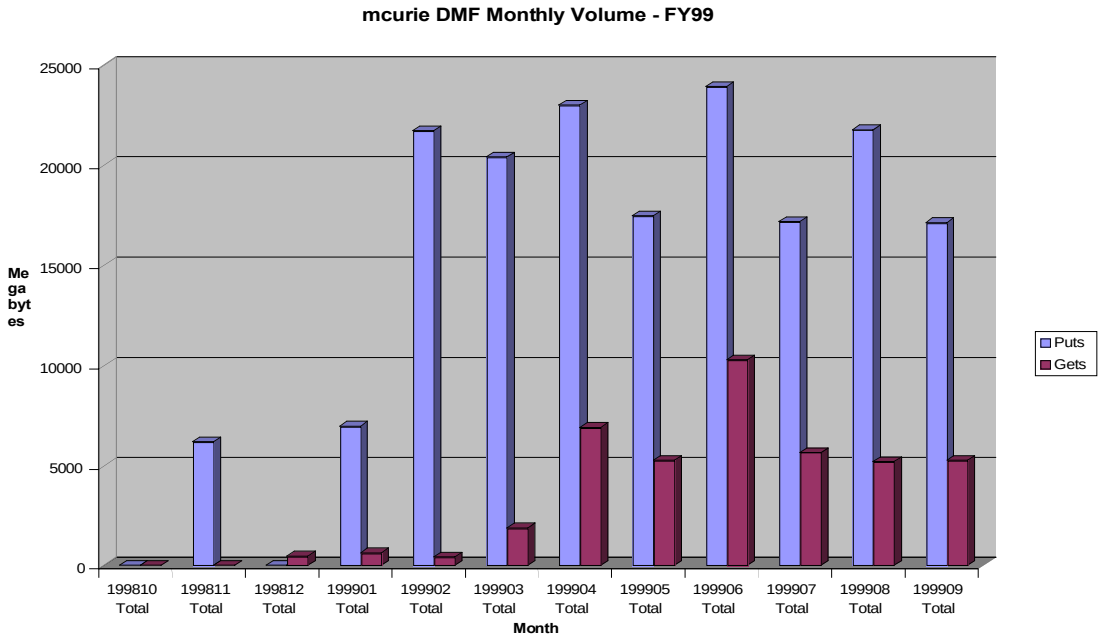
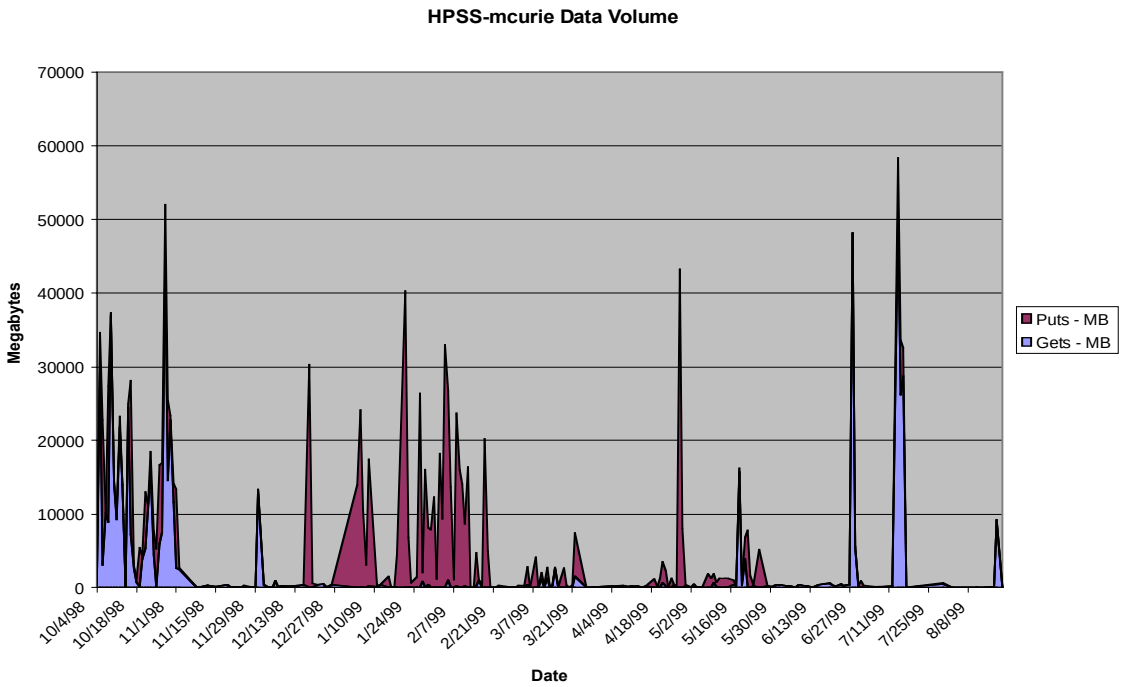


Figure 8.

10. HPSS traffic

Figure 9 shows the daily volume of data transferred between mcurie and HPSS from October 1998 to mid-August 1999. System backups have been excluded. The amount of data transferred is, overall, surprisingly low. This indicates that user needs for intermediate data storage are being met by /usr/tmp. If HPSS were being used for storage of intermediate datasets, HPSS traffic should be much higher, and not nearly as bursty.

Figure 9.



11. I/O utilization and computation

Over the last six months, NERSC has enjoyed increased system utilization and throughput on mcurie because of improved scheduling using the UNICOS/mk political scheduler, PScheD. The existing filesystem configuration has worked to support a high level of system utilization without being swamped by I/O wait time. Most DMF traffic is puts; the much lower number of gets shows that most jobs are not dependent on recalls. HPSS traffic is surprisingly low; this tends to support the conclusion that allowing medium time residency on /usr/tmp of large files is sufficient to support interim file requirements of most NERSC applications. Most traffic to HPSS appears to be long-term archival traffic. Some users have requested and received dispensation to retain large intermediate files on /usr/tmp and exempt them from the normal purging cycle, in order to facilitate making a large series of runs against a particular data set. Having a large enough scratch area to be able to accommodate this type of request has made it possible to enhance users' productivity.

The existing configuration of disks, filesystems, and backing storage appears to be well balanced for NERSC's T3E user community and their applications.

This work was supported by the Director, Office of Advanced Scientific Computing Research, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract number DE-AC03-76SF00098.