Filesystems and I/O Balance on the NERSC T3E

Tina Butler, NERSC Systems Group

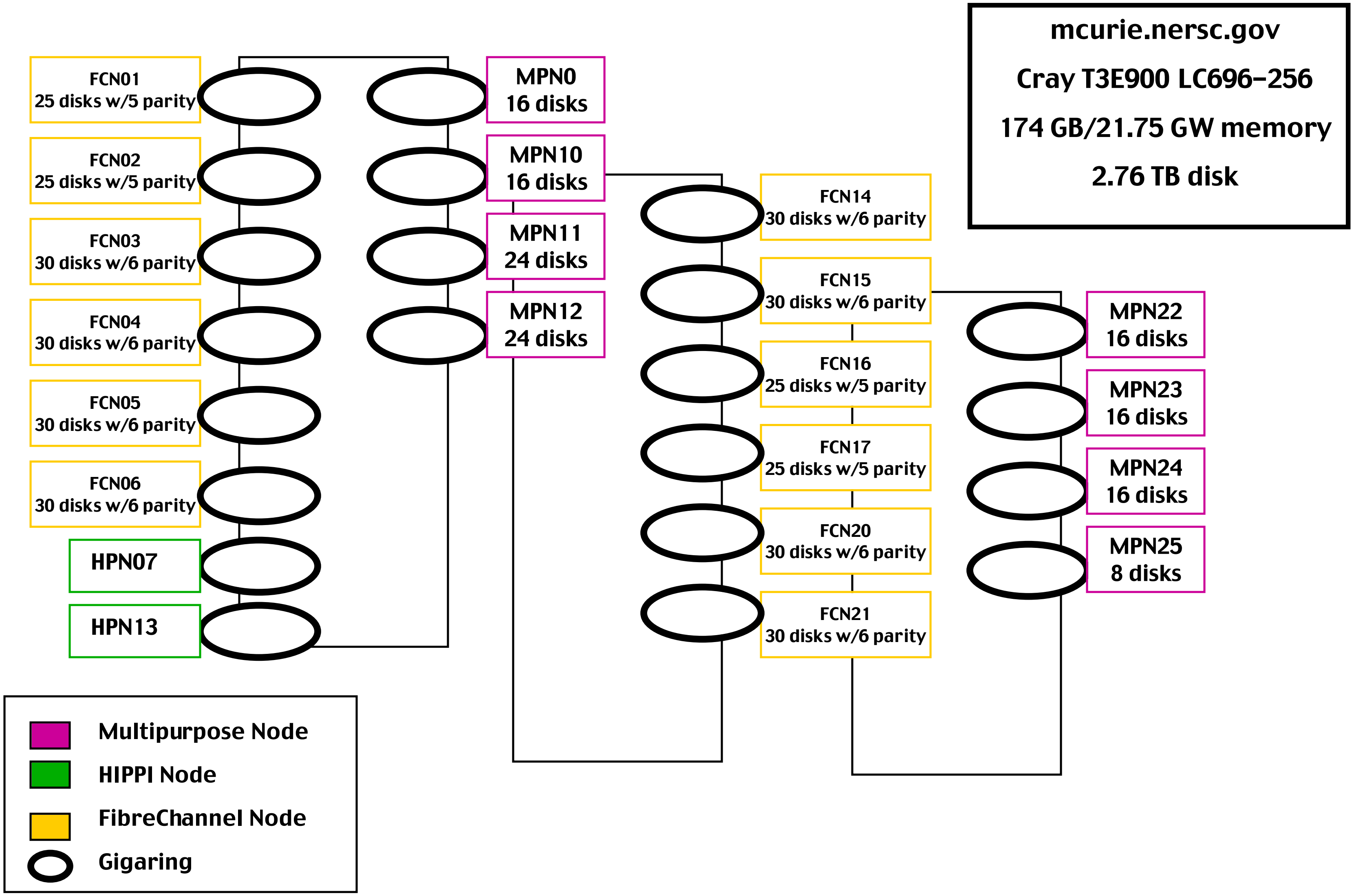
This work was supported by the Director, Office of Advanced Scientific Computing Research, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract number DE-AC03-76SF00098.

What is NERSC?

- National Energy Research Scientific Computing Center
  - Funded by DOE Office of Science
  - Located at Lawrence Berkeley National Lab
  - Provides Computational Resources to the following programs
    - Fusion Energy
    - High Energy and Nuclear Sciences
    - Basic Energy Sciences
    - Biology and Environmental Research
    - Computational and Environmental Research
  - Approximately 2500 Users from Major Universities and Government Labs
  - Hardware: 696 PE T3E-900, 1 J90 SE system (32 CPUs) & 3 SV1 (64 processors)

Mcurie – The NERSC T3E

- T3E 900 with 696 PEs running UNICOS/MK 2.0.4.67
- 644 APP PEs
- 256 MB per PE
- 22 Gigarings
- 12 FCNs
- 8 MPNs
- 2 HPNs



mcurie.nersc.gov

Cray T3E900 LC696-256

174 GB/21.75 GW memory

2.76 TB disk

FCN01 25 disks w/5 parity

FCN02 25 disks w/5 parity

FCN03 30 disks w/6 parity

FCN04 30 disks w/6 parity

FCN05 30 disks w/6 parity

FCN06 30 disks w/6 parity

HPN07

HPN13

MPN0 16 disks

MPN10 16 disks

MPN11 24 disks

MPN12 24 disks

FCN14 30 disks w/6 parity

FCN15 30 disks w/6 parity

FCN16 25 disks w/5 parity

FCN17 25 disks w/5 parity

FCN20 30 disks w/6 parity

FCN21 30 disks w/6 parity

MPN22 16 disks

MPN23 16 disks

MPN24 16 disks

MPN25 8 disks

🟪	Multipurpose Node
🟩	HIPPI Node
🟨	FibreChannel Node
○	Gigaring

NERSC Job Mix – Application Mix

- Applications from the fields of
  - Chemistry
  - Materials Science
  - Fusion Energy
  - Geophysics
  - Biology
  - High Energy Nuclear Physics
  - Climate Modeling
  - Astrophysics
  - Computational Fluid Dynamics
- Mostly user-written codes

NERSC Job Mix – Diverse and Dynamic

App Size (PEs)	% of all Apps	% of PE Hours
2 - 16	56	6
17 - 64	38	56
65 - 128	5	29
129 - 512	1	9

App Run Time	% of all Apps	% of PE Hours
0 – 10 min	56	1
10 – 30 min	23	10
0.5 – 3.5 hr	17	49
3.5 – 12.0 hr	4	40

Mix of Development, Capacity and Capability computing

Mcurie Filesystems – performance

- 68 Fibre Channel disk arrays
- Striping of swap and checkpoint
- pcache for metadata optimization on root, usr, opt
- primary/secondary partitions
- remote mount file servers

Mcurie Filesystems – resiliency

- Mirroring of primary partitions for homes and usrtmp
- Alternate path for all arrays
- Sized for feasible dump/restore

Mcurie Filesystems – swap and checkpoint

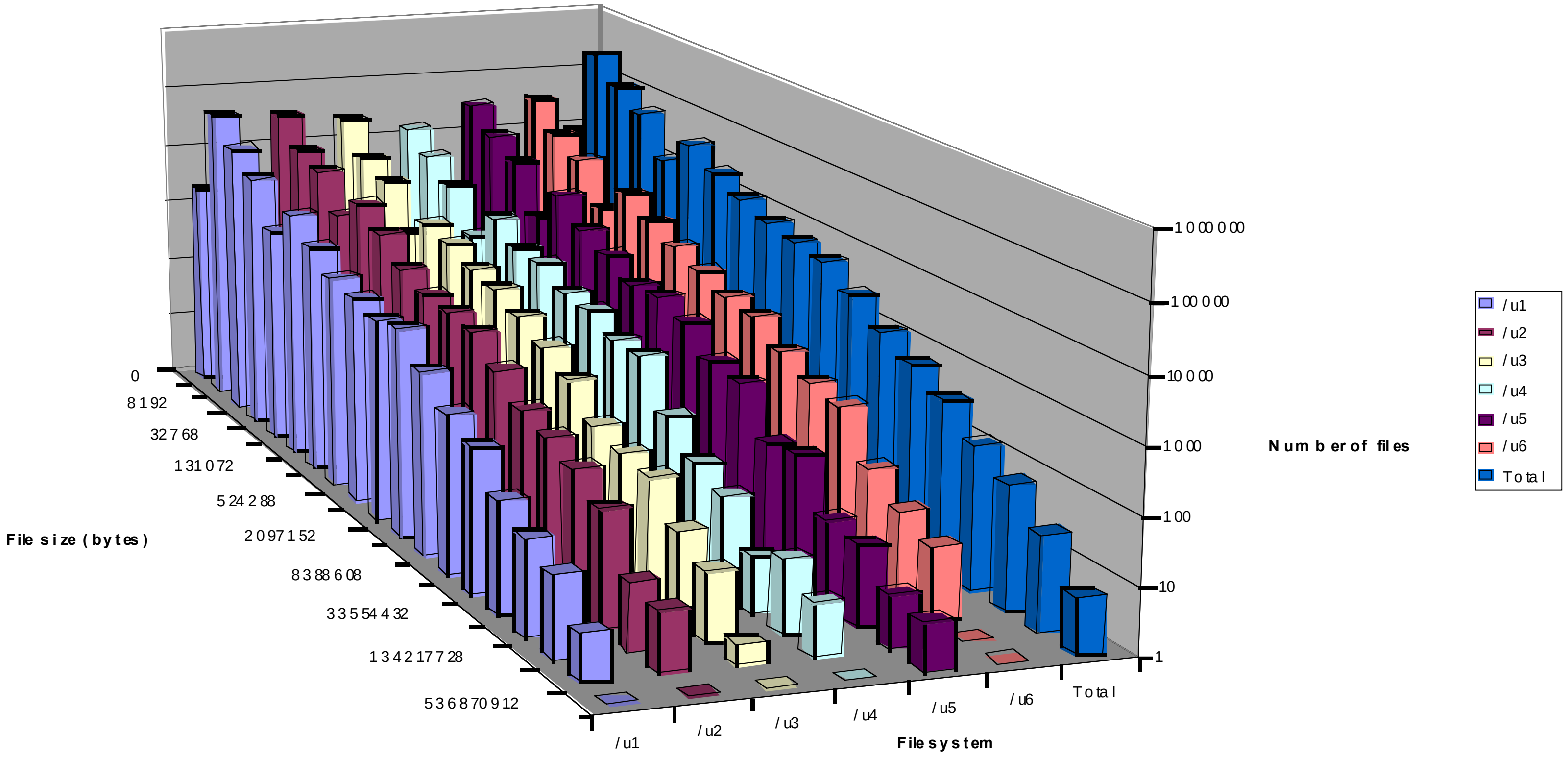
- NERSC uses both checkpointing and gang scheduling for system scheduling
- Swap – 383 Gigabytes – 2.4 times APP memory
- Checkpoint – 582 Gigabytes – 3.6 times APP memory
- Filesystems have 5 logical partitions that are 5 or 6-way striped on FCN disk
- 800 MB/sec observed on checkpoint
- Full machine checkpoint regularly under 5 minutes

Mcurie Filesystems – homes

- Multiple filesystems to distribute user load and risk
- Configured for full mirroring
- Six filesystems – 25 GB on MPN disks
- Approximately 150 users per filesystem

Mcurie Filesystems – homes

File distribution on mcurie homes



Mcurie Filesystems – /usr/tmp

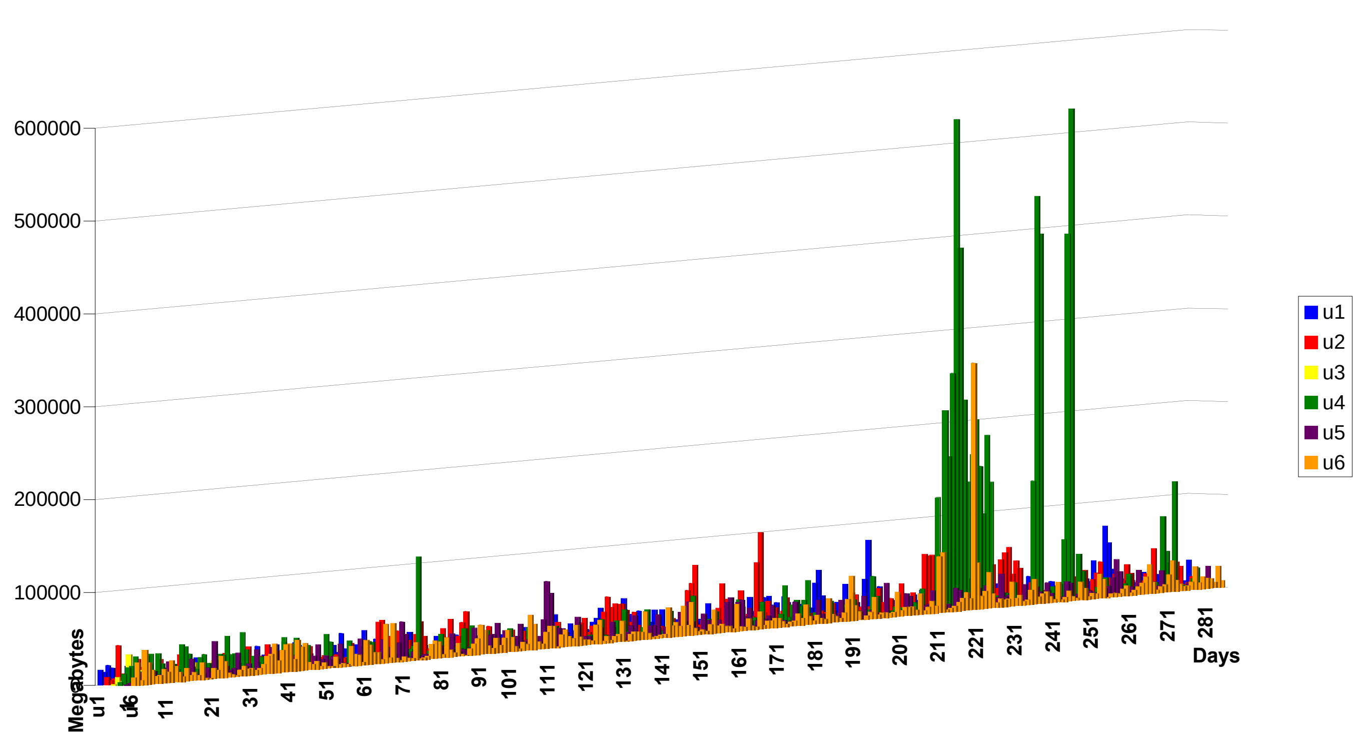
- Main area for user data files
- 1.5 TB of FCN disk arrays
- Primary/secondary partition configuration to allow mirroring of metadata

Mcurie filesystems – space management

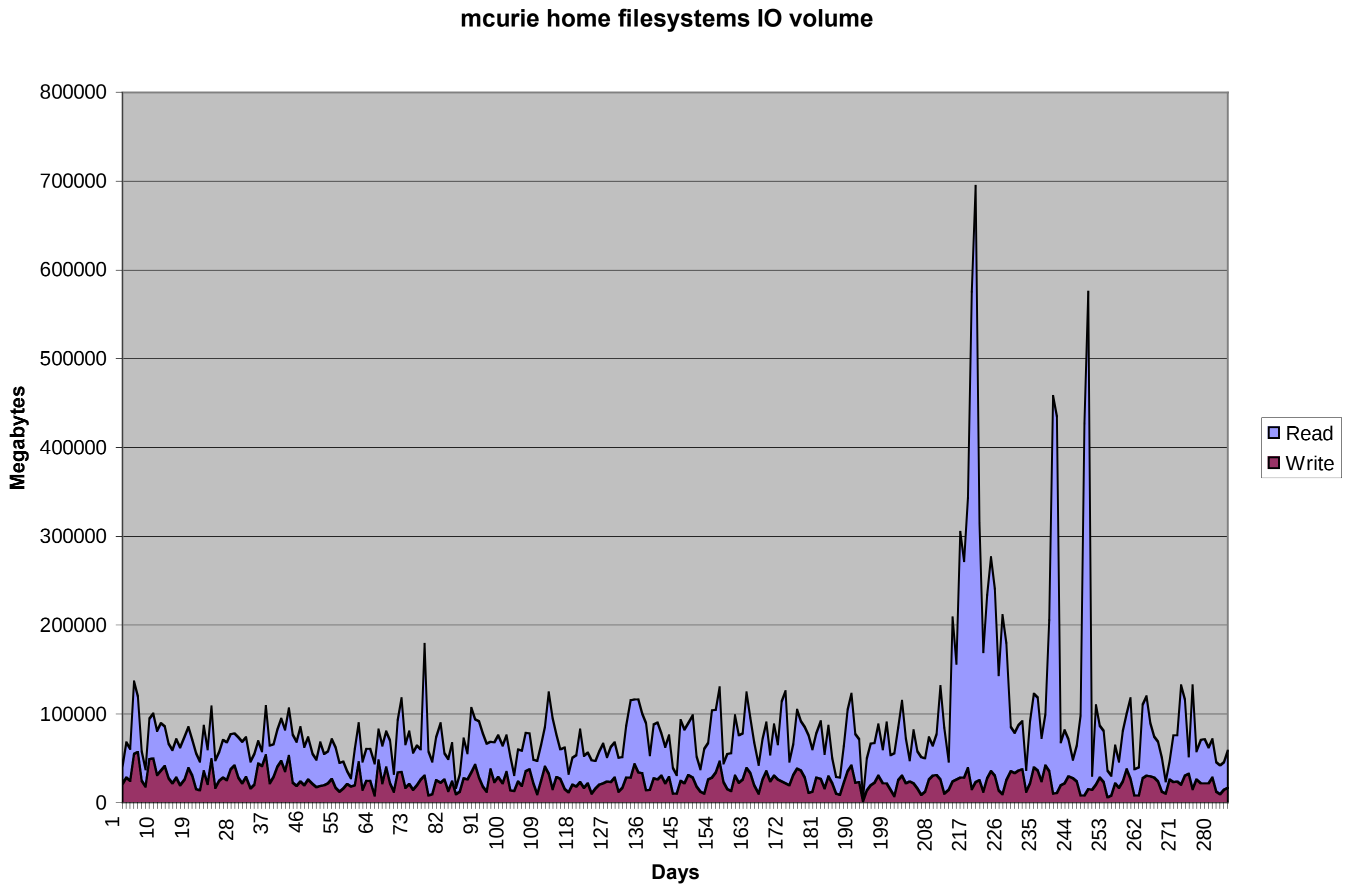
- Hard quotas on user-writable filesystems
- Home filesystems – 4 GB and 3500 inodes
- /usr/tmp filesystem – 70 GB and 6000 inodes
- Homes migrated to HPSS under Cray DMF control
- /usr/tmp – purging of files inactive for 14 days

Mcurie Filesystems – homes

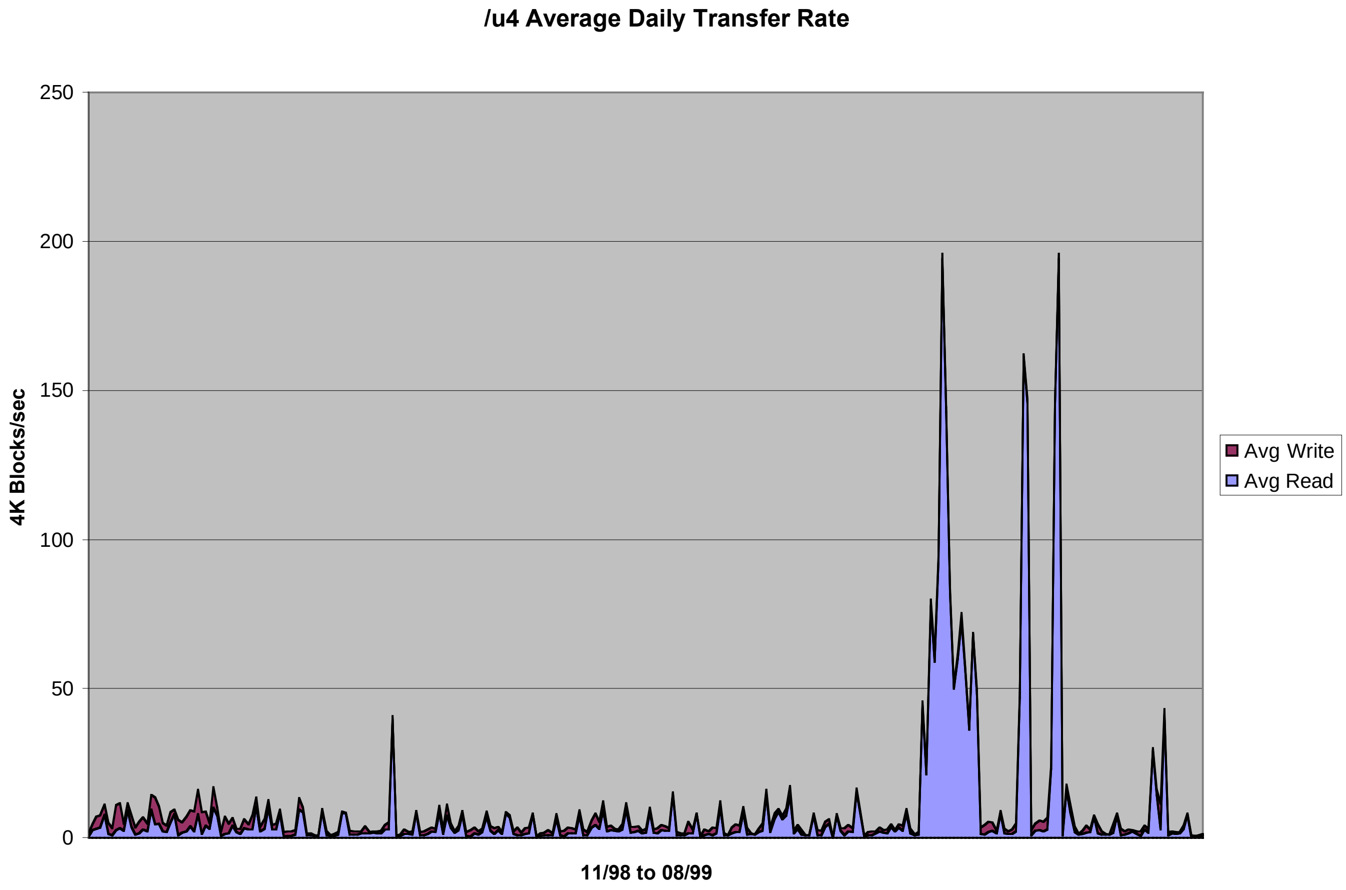
mcurie home IO volume - combined



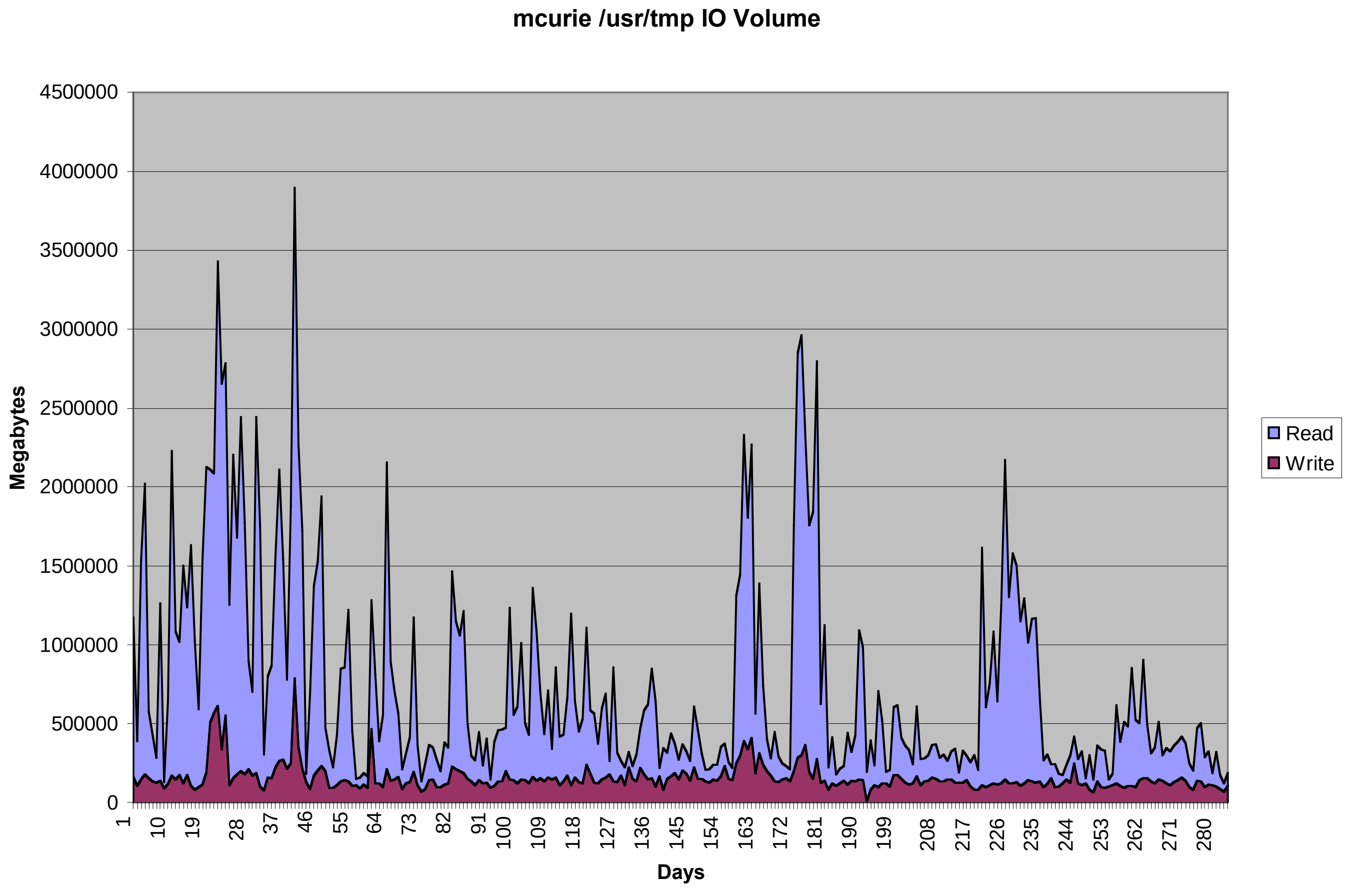
Mcurie Filesystems – homes



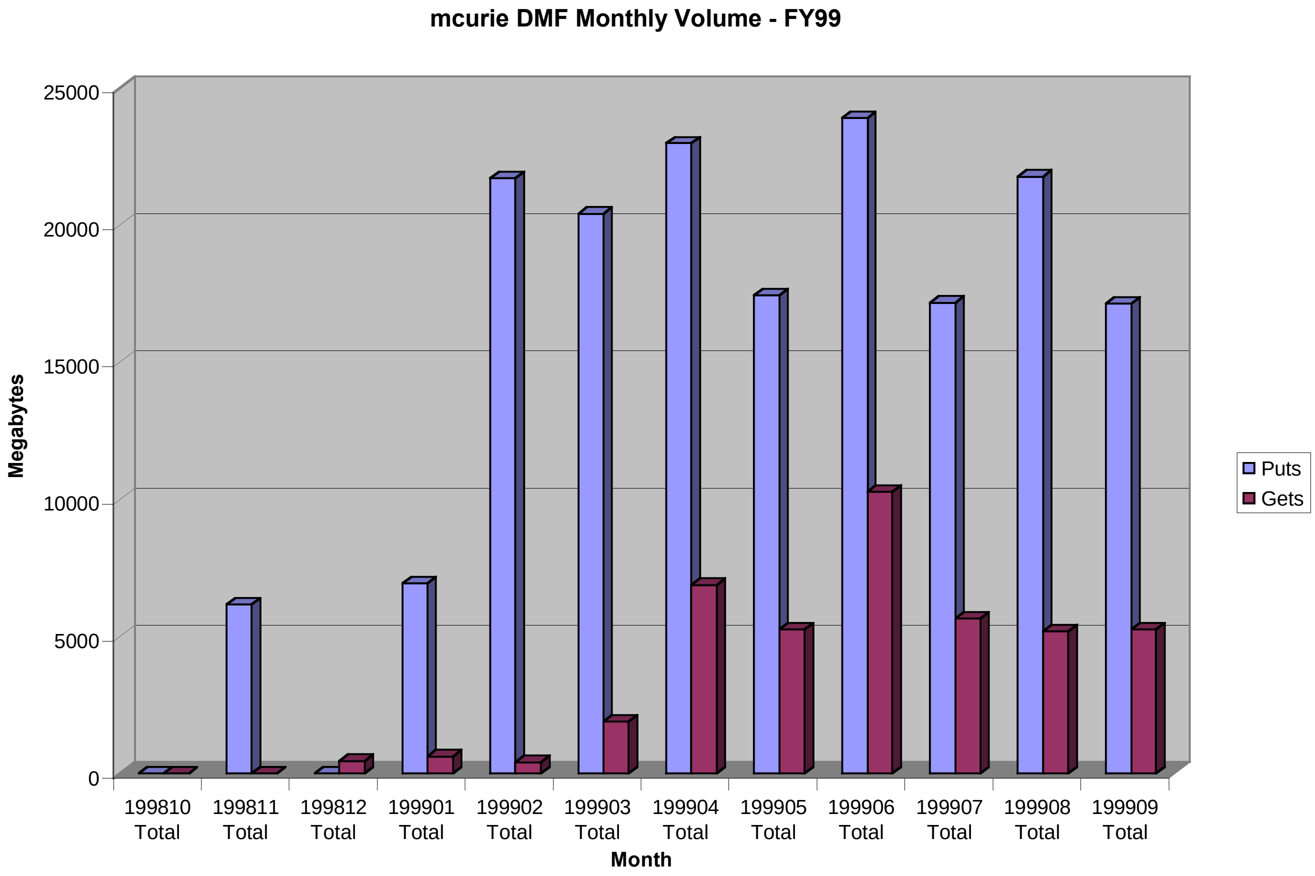
Mcurie Filesystems – home



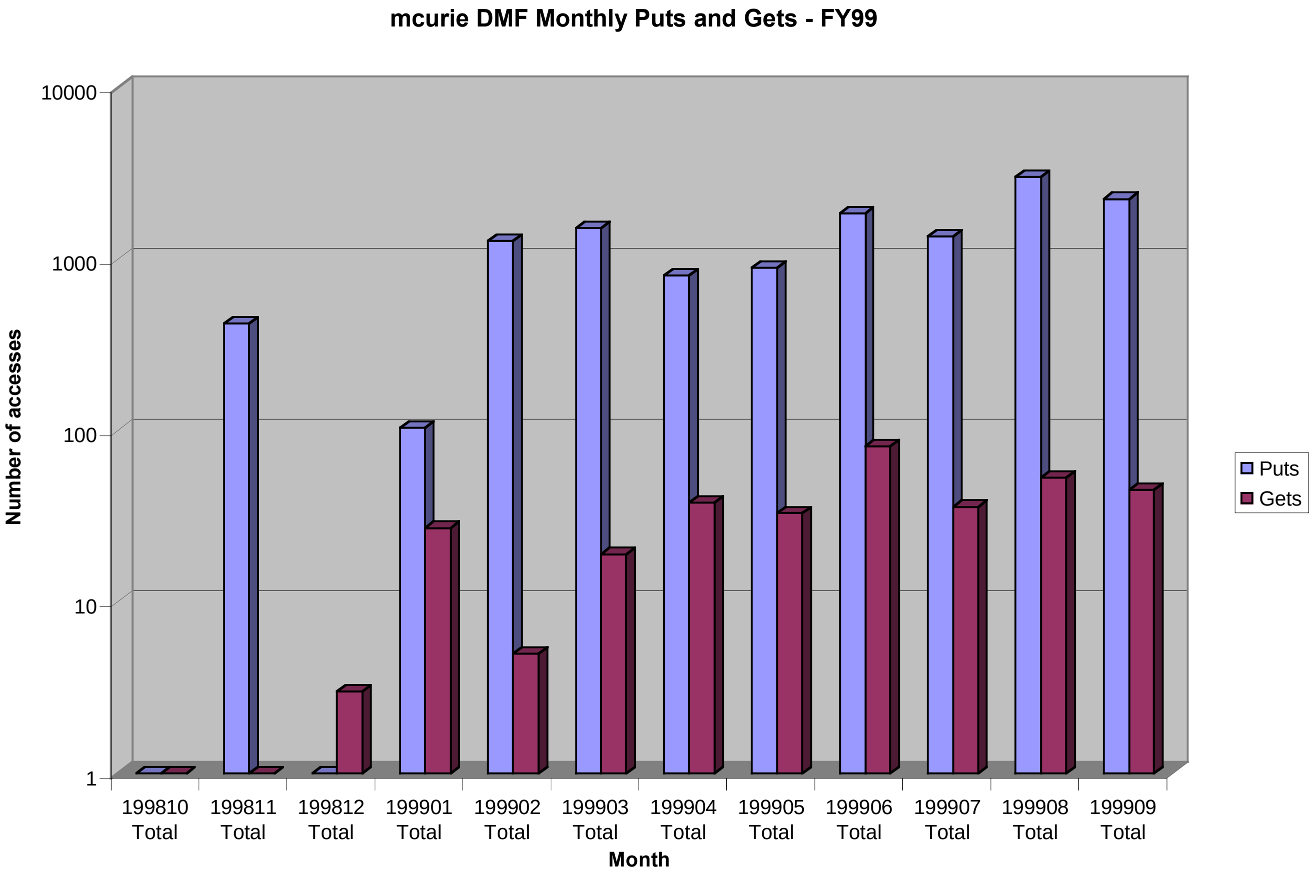
Mcurie filesystems – /usr/tmp



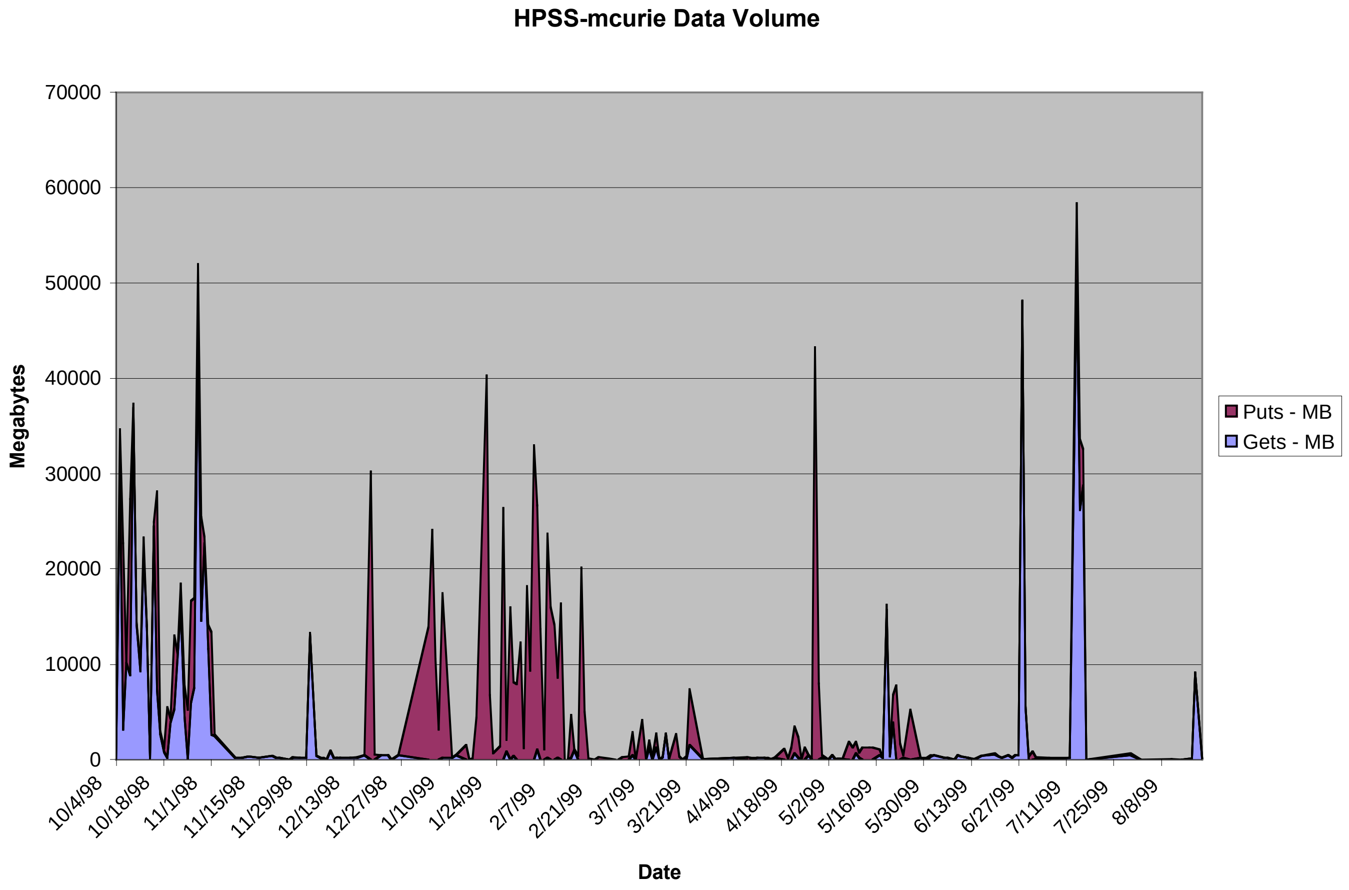
Mcurie Filesystems – DMF traffic



Mcurie Filesystems – DMF traffic



Mcurie Filesystems – HPSS traffic



Mcurie Filesystems – Conclusions

- User home filesystems are well balanced in file distribution and transfer load
- Data migration is a relief valve for homes, but not a critical resource yet
- /usr/tmp filesystem buffers user intermediate data
- HPSS is being used as a long-term archive resource for user data
- NERSC's T3E storage resources are successful in supporting the growing utilization of the system