A trace-driven analysis of disk working set sizes

Chris Ruemmler and John Wilkes

Operating Systems Research Department
Hewlett-Packard Laboratories, Palo Alto, CA

HPL–OSR–93–23, 5 April 1993

Keywords: UNIX, I/O disk access patterns

An analysis of the disk working set size from three different HP-UX computer systems is presented. The data analyzed covers a two-month period and represents every disk request generated over that period on each machine.

The results show that the working set size on all three systems is small by comparison to the total storage size (e.g., about 5% of the total disk storage is accessed over a 24 hour period), although the size exhibits considerable variability (up to up to 30% of the total storage in a 24 hour period), and can change rapidly over a short period of time (fifteen minutes to an hour). Most—but not all—of the large working sets are due to read activity at the disk as a result of system backup periods.

1 Introduction

Understanding the I/O system behavior of computer systems is a necessary prelude to improving their I/O performance. For example, an understanding of the sizes of working sets, and how they change with time, can provide insights into the value of caching data at various levels of the complete system, from disk drives up to and including host main memories.

A previous study of ours [Ruemmler93] was largely concerned with dynamic access characteristics and write traffic of three HP-UX systems over a two-month period. This study augments that analysis with a look at the disk working set sizes of the same three systems. We define the *disk working set* of a system to be the total number of different disk addresses accessed in a *window* of time (such as 1 hour, or 1 day). The size of the working set is measured in bytes. Similarly, the *read* and *write disk working sets* are the number of different addresses read or written respectively. (Notice that the joint working set size is usually smaller than the sum of the read and write sizes, thanks to overlaps between them.)

To maximize the amount of data we collected from a single run, we overlap the windows used to generate the analysis: a new window is started every so often (a time we call the step). With a step size equal to the window size, windows do not overlap; larger steps leave gaps, smaller ones cause windows to overlap. We report here on window sizes of 1 hour and 24 hours, using step sizes of 15 minutes and 1 hour respectively.

To perform this work, we used the same traces as the earlier study: these cover every single physical disk I/O done by the systems over a two month period. Two of these systems (cello and hplajw) were at Hewlett-Packard Laboratories, one (snake) was at UC Berkeley. Some salient attributes of the systems are presented in Table 1.

Table 1: the three computer systems analyzed.

Name	*Processor*[a]	*MIPS*	*HP-UX version*	*Physical memory*	*File buffer cache size*	*Fixed storage*	*Read/write ratio*	*Users*	*Usage type*
cello	HP 9000/877	76	8.02	96 MB	10/30[b] MB	10.4 GB	0.79	20	Timesharing
snake	HP 9000/720	58	8.05	32 MB	5 MB	3.0 GB	0.75	200	Server
hplajw	HP 9000/845	23	8.00	32 MB	3 MB	0.3 GB	0.72	1	Workstation

a. Each machine uses an HP PA-RISC microprocessor.
b. Cello's file buffer size changed from 10MB to 30MB on April 26, 1992.

This paper is organized as follows. First comes a description of our tracing method and some details of the systems traced. It is followed by an analysis of the I/O working sets observed on each of the systems. Finally, a concluding section describing the possible effects of the working set results on I/O system design is given.

The main results of this study are the following:

- the working set size is usually small (2.6–6.7% of the total storage space over 24 hours);
- maximum working set sizes can be quite large (16–34% of the total storage over 24 hours);
- the median write working set sizes are only about a third of the size of the joint working set size.

2 Trace gathering

We traced low-level I/O activity on the three different Hewlett-Packard computer systems described in Table 1. All were running release eight of the HP-UX operating system [Clegg86], which uses a version of the BSD fast file system [McKusick84].

All of our data were obtained using a kernel-level trace facility built into HP-UX. The tracing is completely transparent to the users and adds no noticeable processor load to the system. We logged the trace data to dedicated disks to avoid perturbing the system being measured (the traffic to these disks is excluded from the analysis). Channel contention is minimal: the logging only generates about one write every seven seconds.

Cello is a timesharing system used by a small group of researchers at Hewlett-Packard Laboratories to do simulation, compilation, editing, and mail. A *news* feed that was updated continuously throughout the day resulted in the majority (63%) of the I/Os in the system, and these I/Os have a higher-than-usual amount of writes (63%). Because of the large activity directed to the news partitions, the system as a whole does more writes (56%) than reads.

Snake acted as a file server for an HP-UX cluster [Bartlett88] of nine clients at the University of California, Berkeley. Each client was an Hewlett-Packard 9000/720 workstation with 24MB of main memory, 66MB of local swap space, and a 4MB file buffer cache. There was no local file system storage on any of the clients; all the machines in the cluster shared the single common file system hosted by the server with complete single UNIX-system file operation semantics. The cluster had accounts for faculty, staff, graduate students, and computer science classes. The main use of the system was for compilation and editing. This cluster was installed in January 1992, so many of the disk accesses gathered in our traces were for the creation of new files. Over the tracing period, the /usr1 disk gained 243MB and /usr2 gained 120MB of data.

Finally, the personal workstation (hplajw) was used by a single user. The main uses of the system were electronic mail and editing papers.

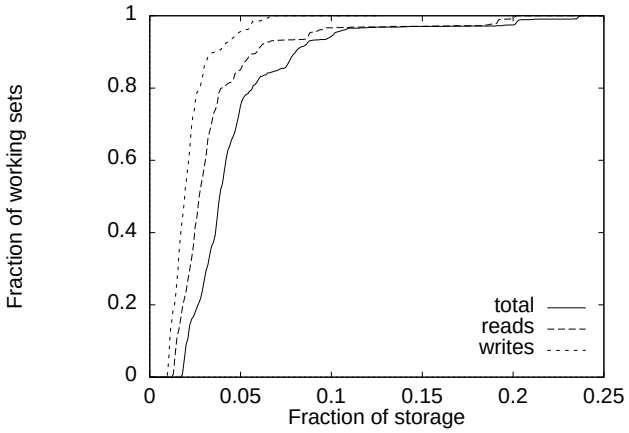
Cello and hplajw were traced from 92.4.18 to 92.6.20; snake from 92.4.25 to 92.6.27. Each trace started at 0:00 hours on a Saturday. The total numbers of I/O requests logged over the tracing period were: 29.4M (cello), 12.6M (snake) and 0.4M (hplajw).

3 Results

This section presents what we discovered about the sizes of the disk working sets.

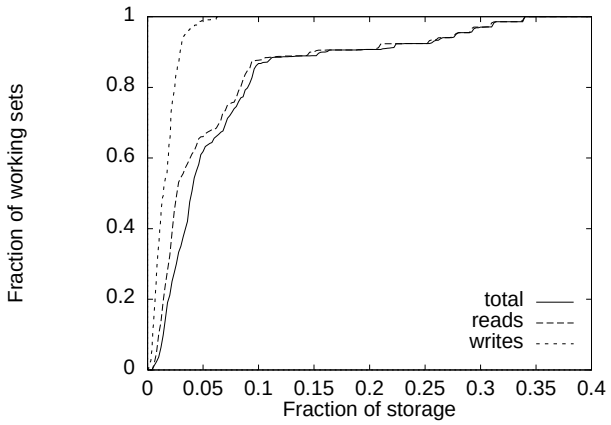
Figure 1 and Figure 2 show the distribution of working set sizes as a fraction of the total disk storage for each system. Figure 1 shows the data for a working set window of 24 hours with a step size of 1 hour, while Figure 2 shows the data for a working set window of 1 hour with a step size of 15 minutes.

The figures show a highly skewed distribution of working set sizes: most are small (medians of 3.9%, 4.0%, and 2.1% for cello, snake and hplajw respectively for the 24 hour windows), with 90th percentile sizes of 8.0%, 16.0% and 5.7% respectively. A small number of working sets were much larger: the maximum sizes were 24%, 34%, and 16% of the total. Table 2 shows the same results in terms of absolute working set sizes.



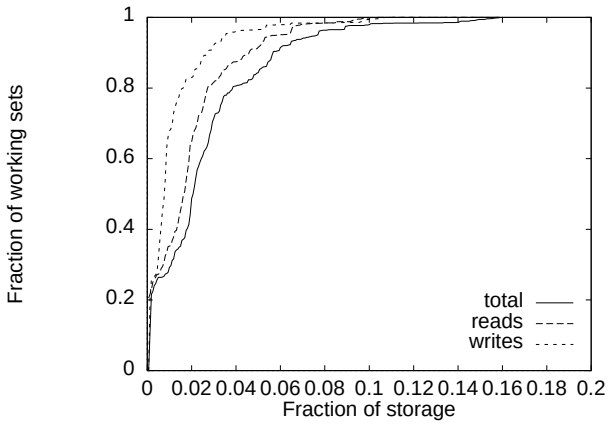
	reads	*writes*	*total*
mean	3.58%	2.19%	4.76%
mode	1.35%	1.58%	3.75%
standard dev.	3.20%	1.12%	3.52%
minimum	1.25%	0.96%	1.78%
10th percentile	1.48%	1.10%	2.06%
50th percentile	2.71%	1.92%	3.85%
90th percentile	5.90%	3.57%	8.02%
maximum	20.2%	6.82%	23.6%

a. Cello



	reads	*writes*	*total*
mean	6.02%	1.62%	6.76%
mode	2.10%	2.09%	3.70%
standard dev.	7.69%	1.06%	7.59%
minimum	0.42%	0.09%	0.45%
10th percentile	1.00%	0.46%	1.41%
50th percentile	2.65%	1.41%	3.96%
90th percentile	14.7%	2.93%	16.0%
maximum	33.9%	6.27%	33.9%

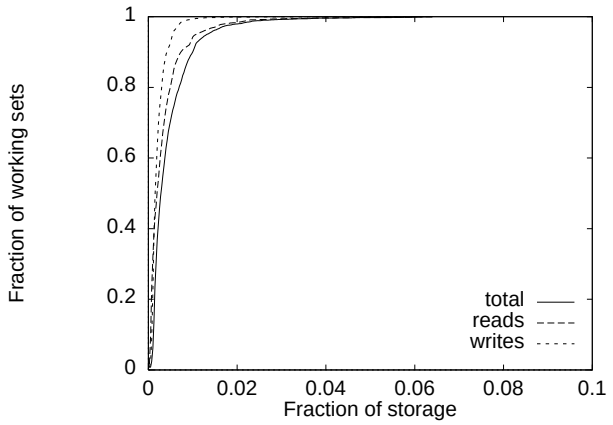
b. Snake



	reads	*writes*	*total*
mean	1.90%	1.17%	2.59%
mode	0.03%	0.09%	0.09%
standard dev.	1.89%	1.55%	2.64%
minimum	0.00%	0.06%	0.07%
10th percentile	0.03%	0.09%	0.12%
50th percentile	1.65%	0.77%	2.07%
90th percentile	4.59%	2.72%	5.67%
maximum	10.2%	10.7%	16.0%

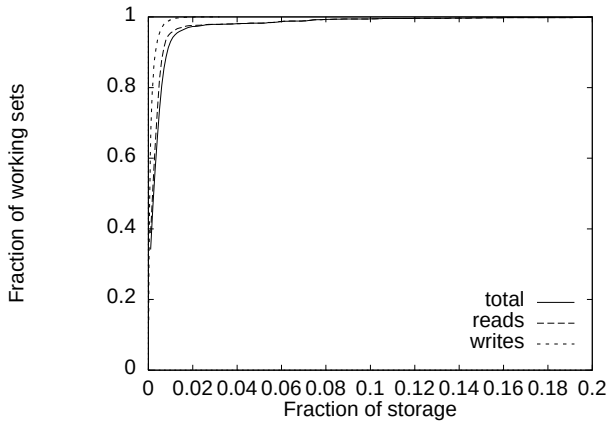
c. Hplajw

Figure 1. Distribution of working set sizes as a fraction of the total storage for all three systems. The working set window is 24 hours and it is moved in 1 hour steps. The numbers shown are the size of the working set as a percentage of the total storage space on each system.



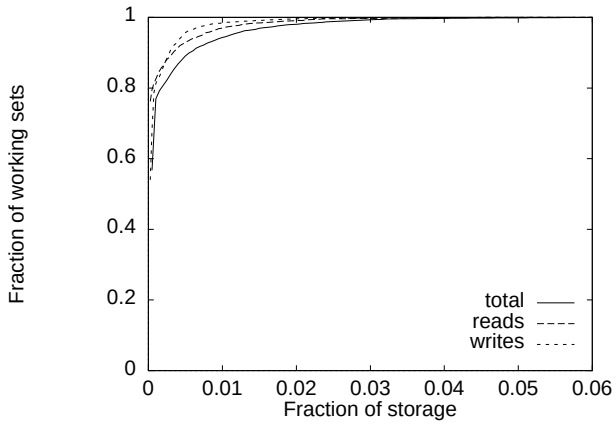
	reads	*writes*	*total*
mean	0.34%	0.21%	0.46%
mode	0.06%	0.11%	0.14%
standard dev.	0.50%	0.20%	0.55%
minimum	0.00%	0.01%	0.01%
10th percentile	0.05%	0.08%	0.11%
50th percentile	0.20%	0.15%	0.28%
90th percentile	0.73%	0.42%	1.00%
maximum	6.11%	4.29%	6.40%

a. Cello



	reads	*writes*	*total*
mean	0.44%	0.13%	0.51%
mode	0.05%	0.02%	0.05%
standard dev.	1.41%	0.19%	1.42%
minimum	0.00%	0.00%	0.01%
10th percentile	0.03%	0.01%	0.03%
50th percentile	0.19%	0.06%	0.24%
90th percentile	0.66%	0.32%	0.85%
maximum	20.2%	2.09%	20.2%

b. Snake



	reads	*writes*	*total*
mean	0.12%	0.10%	0.20%
mode	0.01%	0.01%	0.03%
standard dev.	0.36%	0.28%	0.52%
minimum	0.00%	0.00%	0.01%
10th percentile	0.00%	0.01%	0.01%
50th percentile	0.02%	0.02%	0.04%
90th percentile	0.33%	0.29%	0.55%
maximum	4.89%	4.49%	7.05%

c. Hplajw

Figure 2. Distribution of working set sizes as a fraction of the total storage for all three systems. The working set window size is 1 hour and it is moved in 15 minute steps. The numbers shown are the size of the working set as a percentage of the total storage space on each system.

Table 2: absolute total working-set sizes.

	24 hour windows		*1 hour windows*	
	50th %ile	*90th %ile*	*50th %ile*	*90th %ile*
cello	406 MB	834 MB	29 MB	104 MB
snake	120 MB	480 MB	7 MB	26 MB
hplajw	6 MB	17 MB	0.1 MB	2 MB

On all three systems, the write working set sizes were smaller than the read (1.9% to 2.7% for the median sizes for cello for 24 hour windows, 1.4:2.7% for snake, and 0.8:1.7% for hplajw). This is consistent with the high degree of temporal locality for write traffic found in [Ruemmler93].

Figure 3 and Figure 4 show how the working set size varies over time. In Figure 3, a 24 hour working set window is used, with 1 hour steps. The working set size is around 4% of the total storage space for most periods, but a few have working set size as large as 16–34% of the storage space. These large increases in the working set size result from full backups on cello and snake where all user data (modified or not) is backed up. The full backup occurs once a month on cello and on snake it occurs approximately weekly. A full backup is supposed to occur weekly on hplajw, but these periods are not real discernible from Figure 3c. Notice that the write working set size on each system does not exhibit as much variation as the total working set size.

Figure 4 corresponds to the 1 hour working set window with 15 minute steps. Most of the large bursts are caused once again by reads from system backups, except on snake, where other activities also generate many reads. Figure 2b indicates that the maximum write working set size is only 2.09% of the total storage for the snake machine, but there are many bursts in Figure 4b where the working set size is greater than 5% of the total storage. The working set size on cello tends to be relatively small (between 0.05% and 1.5% of the total storage), but it does fluctuate a lot and tends to be very periodic. This could be a result of running the news program on cello. Even on hplajw, the personal workstation, the working set tends to vary significantly (from almost 0% to about 3% of the total storage) over fifteen minute intervals.

4 Conclusion

This paper has presented an analysis of the disk working set sizes for three different HP-UX computer systems over a period of two months. The results show that the overall working set size corresponds closely to the read working set size, while the write working set is typically only half the size of the read one. There is considerable burstiness in the working set (up to a third of the total storage over a 24 hour window)—and although many of the larger working set sizes are the result of system backups, this is not always the case.

These data reinforce previous work suggesting that keeping much of the write working set in a fast-access memory (such as a non-volatile cache) is practical. The read working set sizes, however, are less easy to handle: their median sizes on the systems we measured were up to 406MB (cello)—on a system that its users felt was fairly generously endowed with 96MB of physical memory. The 90th percentile working set sizes were larger still: close to half a gigabyte for the snake server, and a gigabyte for cello. This result is somewhat at odds with at least one

earlier study of file access patterns [Ousterhout85a], which suggested that main memories of around 100MB might make disks obsolete.

References

[Bartlett88] Joel F. Bartlett. *Compacting garbage collection with ambiguous roots*. Research report 88 / 2. Digital Equipment Corporation Western Research Laboratory, Palo Alto, CA, February 1988.

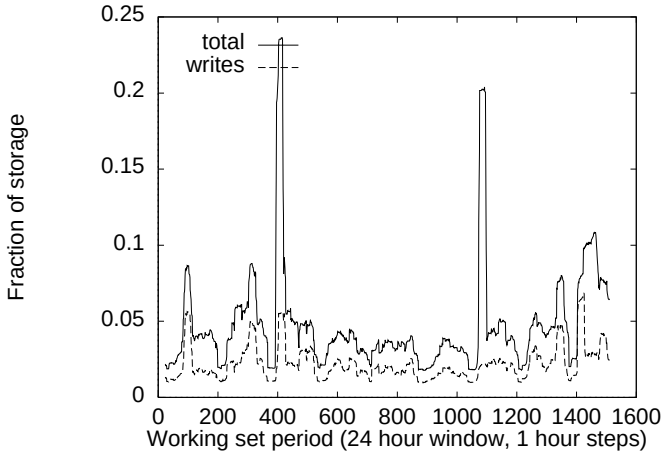
[Clegg86] Frederick W. Clegg, Gary Shiu-Fan Ho, Steven R. Kusmer, and John R. Sontag. The HP-UX operating system on HP Precision Architecture computers. *Hewlett-Packard Journal*, **37**(12):4–22, December 1986.

[McKusick84] Marshall K. McKusick, William N. Joy, Samuel J. Leffler, and Robert S. Fabry. A fast file system for UNIX. *ACM Transactions on Computer Systems*, **2**(3):181–97, August 1984.

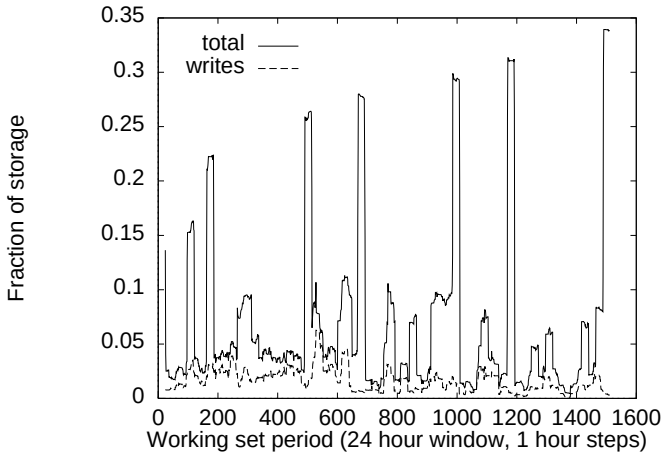
[Ousterhout85a] John K. Ousterhout, HervéDa Costa, David Harrison, John A. Kunze, Mike Kupfer, and James G. Thompson. A trace-driven analysis of the UNIX 4.2 BSD file system. *Proceedings of 10th ACM Symposium on Operating Systems Principles* (Orcas Island, Washington). Published as *Operating Systems Review*, **19**(5):15–24, December 1985.

[Ruemmler93] Chris Ruemmler and John Wilkes. UNIX disk access patterns. *Proceedings of Winter 1993 USENIX* (San Diego, CA, 25–29 January 1993), pages 405–20, January 1993.

a. Cello



b. Snake



c. Hplajw

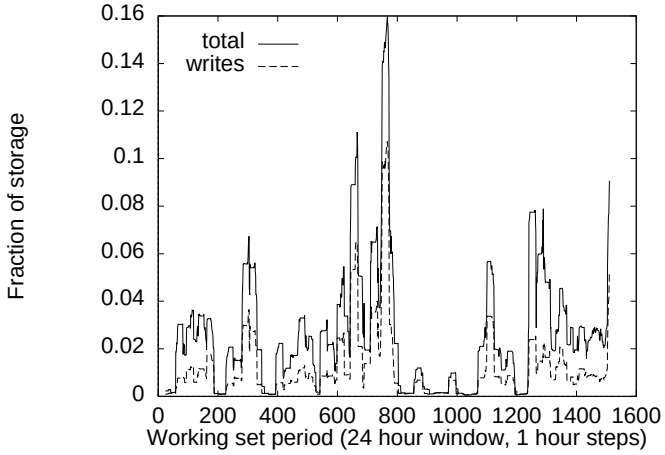
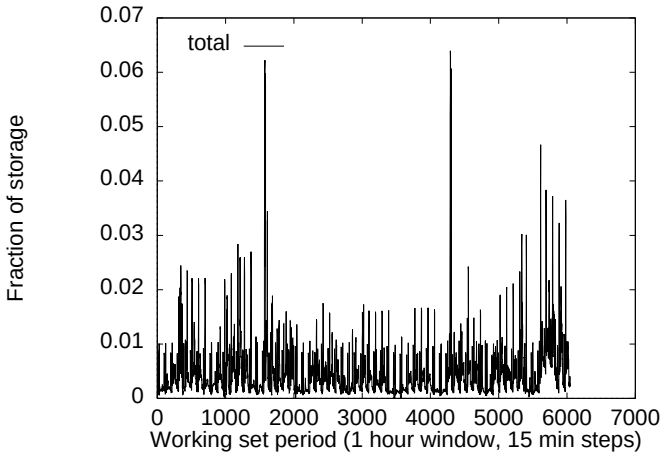
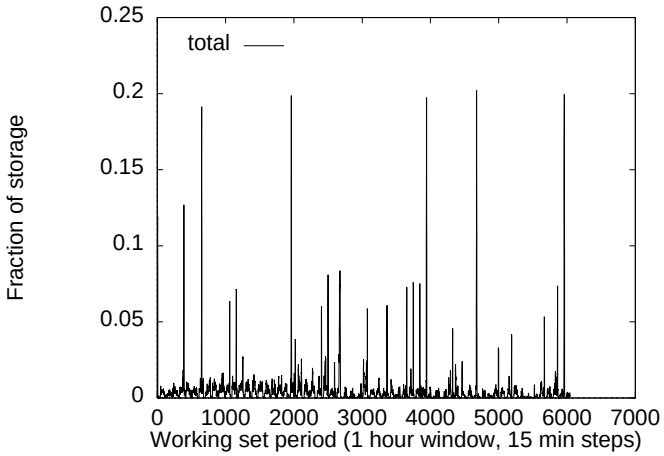


Figure 3. Working set size variation over the entire trace period. This graph shows the variation for a twenty-four hour working set window with one hour steps. Since the read working set size corresponds almost directly to the total working set size for all systems, it is not plotted here.

a. Cello



b. Snake



c. Hplajw

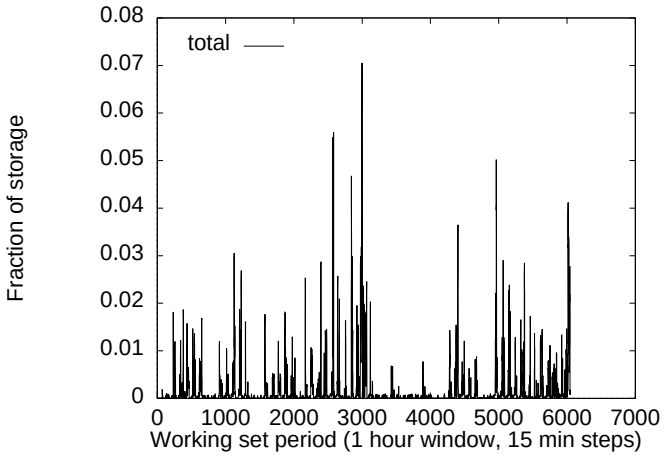


Figure 4. Working set size variation over the entire trace period. This graph shows the variation for a one hour working set window with 15 minute steps.