J90 Performance and Tuning

Chad Vizino, Pittsburgh Supercomputing Center

ABSTRACT: *This tutorial will offer an introduction to J90 performance and tuning based on experiences that the Pittsburgh Supercomputing Center has had in operating its J90 computers. The tutorial will begin with an overview of the tuning process. System activity monitoring, workload management, memory performance, and disk performance will be discussed.*

Introduction

•Will share tips the PSC has gleaned in the operation of the current J90 supercomputers.
•This talk aimed at system administrators and analysts who are new to UNICOS and the J90 architecture.

Agenda

•Tuning Overview
•Overview of PSC
•Monitoring System Activity
•Managing the Workload
•Memory Performance
•Disk Performance (will spend most time here)
•Configuring UNICOS

Tuning Overview

•Why performance and tuning?
–Want to squeeze maximum performance out of machine.
–May be experiencing unacceptable delays in turnaround.
–Want to keep from having to buy more hardware.
•Performance problems are often not simple.

What to tune?

•IO
•Memory
•CPU
•User Codes

When to tune?

•Proactive: at configuration time.
•Later: when there is a problem.

What do we want?

•Goal: optimum performance
•How to get there?
–Tune individual resource pieces (cpu, memory, io, etc.)
–Educate users.

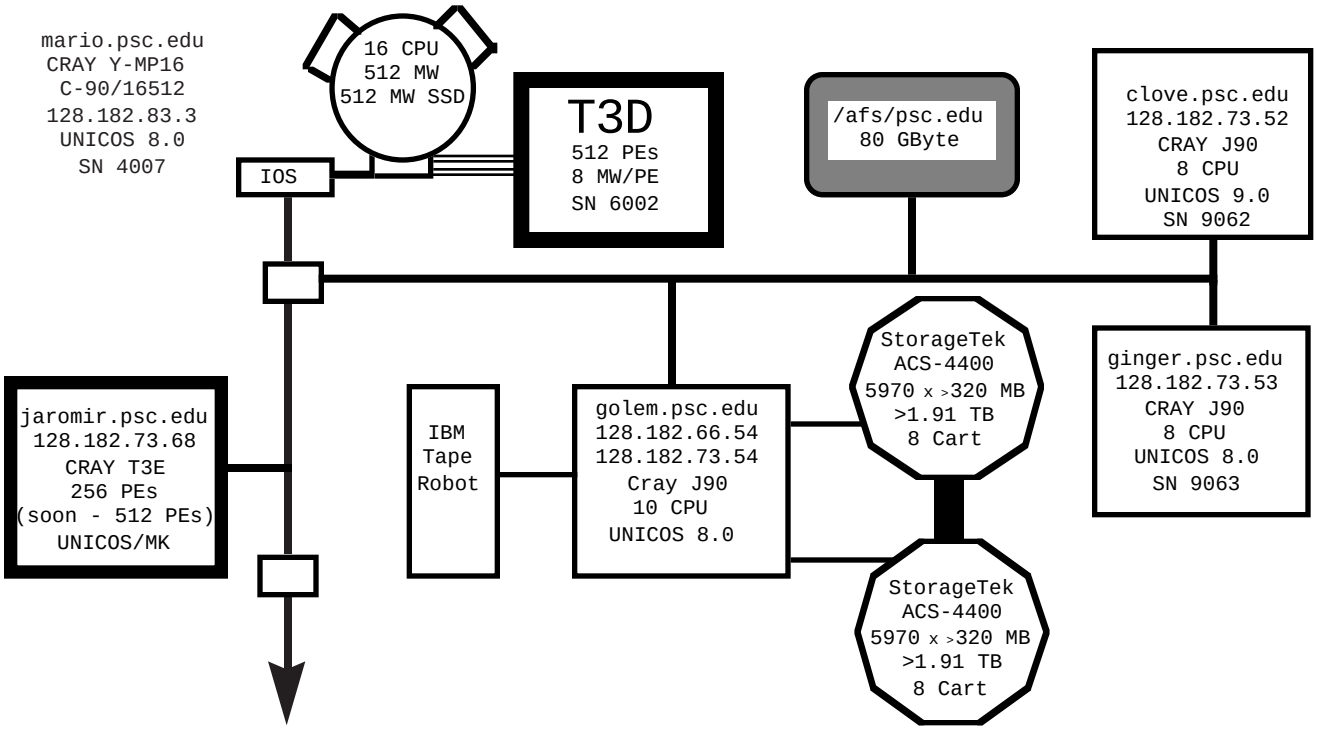
Overview of PSC

•Who we are
–One of four NSF funded supercomputing centers.
–Collaboration among Carnegie Mellon University, University of Pittsburgh and Westinghouse Electric Corp.
–Users are spread over the U.S.
–Users are academic researchers.
–Have a diversity of codes to run.
–Many needs arise with little notice.
•Network Map
•How did we learn what we did?
–Most of our experience came from operating a C90.

Four-fold Mission

•Enable solutions to important problems in Science and Engineering by providing leading-edge computational resources to the national community;
•Advance computational science, computational techniques and the National Information Infrastructure;
•Educate researchers in high performance techniques and their utility; and
•Assist the private sector in exploiting high performance computing for their competitive advantage.

Environment



Schematic Representation of the Cray Hardware Environment at the Pittsburgh Supercomputing Center

J90 Equipment

•Cray J916 Computer Systems
–ginger
–clove
•Each has 128mw memory and 8 cpus.

•Each has 2 IO subsytems (IOS) and 1 SI-3 controller on each IOS.
•Each use DD-6S disks exclusively.

Overview of PSC/Equipment

Ginger

•Running UNICOS 8.0.4
•Used to accommodate low-end C90 work.
•8 DD6S disks (72GB)

Clove

•Running UNICOS 9.0.2.1
•Used to accommodate low-end C90 work and T3E code development.
•Test platform
•12 DD6S disks (98GB)

Monitoring System Activity

•Two kinds of tools
–Real time
–After the fact

Real Time Monitoring Tools

•usage_mem

```
================================[ CPU Usage ] ========= (ginger) - 8.0.4 ===
Averages for 10/09/96 [ 84 %Usr, 2 %Sys, 14 %Idl, 77 %MemUse]
======================================== U[.] == S[s] == I[ ] == M[m] =======
 95%    ..............    . ..    ..
 85%   mmmmmmmmmmmmmmmm....m..m.....m..        .                   ...mm m
 75%  m..............m....mm.m...m.m.   m.mmmm....m.....m.. .......m..
 65%  ..............mmm.....mmm.... ...m....mmmm.mmmmm.mmmmmmmm....mmmmmmmm
 55%  ..............................m.....................................
 45%  ..............................m.....................................
 35%  ..............................m.m...................................
 25%  ......................................................................
 15%  ......................................................................
  5%  ....................................s...........................
======|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==
 Hour 12 1  2  3  4  5  6  7  8  9  10 11 12 1  2  3  4  5  6  7  8  9  10 11
```

•ps
–Ex. ps -lp pid or ps -wp pid to get job id.
•jstat -j jobid
–Helps to find out what system call is being called and how much system time jobid has accumulated.
•sam
–interesting graphs...of limited use

PDDSTAT(8) Cray Research, Inc. SR-2022 8.0

NAME
pddstat - Displays information about the IOS model E
SYNOPSIS
/etc/pddstat [-d device.unit] [-e] [-r rate] [-l] [-q] [-s]
IMPLEMENTATION
CRAY Y-MP systems with I/O subsystem model E
CRAY J90 systems
CRAY EL series systems

DESCRIPTION

The pddstat command gets information from the disk table, which controls disk input/output (I/O) in an IOS model E. The information is attained with the tabread (see tabinfo(2)) system call and is displayed on the caller's screen.

pddstat Sample Output

```
ginger% pddstat
            device                              # of sectors         errors
  name    type  unit state mode req  func     reads     writes  rd-ur wt-ur total
 -------- ----- ---- ----- ---- --- ------ --------- --------- ----- ----- -----
    24020  DD6S  0     up   rw   1  read     1664325   1774650     0     0     0
    24020  DD6S  1     up   rw   0  read      991558   3566693     0     0     0
    25020  DD6S  0     up   rw   0  read     2626838   3643985     0     0     0
    25020  DD6S  1     up   rw   0  read      647047   1314985     0     0     0
  1024030  DD6S  0     up   rw   0  read     3850580   3335577     0     0     0
  1024030  DD6S  1     up   rw   0  read      665311   2067900     0     0     0
  1024030  DD6S  2     up   rw   1  read    24617477   2634031     0     0     0
  1024030  DD6S  3     up   rw   0  write    7040256   3417370     0     0     0
```

•Other useful option

–-d (device)

sar

SAR(1) Cray Research, Inc. SR-2011 8.0

NAME

sar - Extracts operating system activity information

SYNOPSIS

```
sar [-a] [-b] [-c] [-d] [-g] [-h] [-j] [-k] [-l] [-o file] [-p] [-q]
[-t] [-u] [-v] [-w] [-x] [-y] [-z] [-A] [-B] [-H] [-L] [-M] [-P] [-T]    |
[-U] [-W] [-X] [-Z] seconds [integral]

sar [-a] [-b] [-c] [-d] [-e time] [-f file] [-g] [-h] [-i sec] [-j]
[-k] [-l] [-p] [-q] [-s time] [-t] [-u] [-v] [-w] [-x] [-y] [-z] [-A]
[-B] [-H] [-L] [-M] [-P] [-T] [-U] [-W] [-X] [-Z]                       |
```

•Two Modes

–Real time

–After the fact

sar -u

•Reports cpu usage.

ginger% sar -u 1 4

sn9063 ginger 8.0.4 viz.10 CRAY J90 10/08/96

```
15:08:14    %usr    %sys   %wsem     #locks   %idle  %guest
15:08:15      60       5       0         81      35       0
15:08:16      57       5       0         57      38       0
15:08:17      61       3       0         28      37       0
15:08:18      65       2       0         15      33       0

Average       61       4       0         45      36       0
```

sar -q

•Reports average queue length.

ginger% sar -q 1 4

sn9063 ginger 8.0.4 viz.10 CRAY J90 10/08/96

15:12:36 runq-sz %runocc swpq-sz %swpocc
15:12:37 6.0 96
15:12:38 5.0 96
15:12:39 14.0 96
15:12:40 5.0 96

Average 7.5 96

sar -t

•Reports system call information.

ginger% sar -t 1 2

sn9063 ginger 8.0.4 viz.10 CRAY J90 10/08/96

```
15:16:33 system-call       %time   calls/s   avetime     maxtime   mintime
15:16:34 write            42.452      87.3    2028.4    594940.7      15.8
         listio           20.074      53.7    1558.7    325464.6     193.2
...
15:16:35 write            31.686      99.5    2171.0    594940.7      15.8
         getdents         13.418      25.8    3541.4    159177.7     114.2
...
Average  write            35.766      93.4    2104.5    594940.7      15.8
         listio           14.143      50.3    1545.4    325464.6     193.2
```

sar -M

•Reports memory and swap usage.

ginger% sar -M 1 4

sn9063 ginger 8.0.4 viz.10 CRAY J90 10/08/96

15:22:55 umemtot umemuse memlock swaptot swapuse
15:22:56 202024 83628 1120 524288 1968
15:22:57 202024 83628 1120 524288 1968
15:22:58 202024 83628 1120 524288 1968
15:22:59 202024 83628 1120 524288 1968

Average 202024 83628 1120 524288 1968

sar -d

•Reports activity for each disk device.

ginger% sar -d 1 2

sn9063 ginger 8.0.4 viz.10 CRAY J90 10/08/96

```
15:25:18  path type unit    reads     writes    ncyls  avwait  avserv rerr uerr

15:25:19  5020 DD6S  0       1358          0     2253    0.03   13.36    0    0
          5020 DD6S  1         14          0     4682    0.00    0.00    0    0
...

15:25:20  5020 DD6S  0       1316          0     4364    0.03   13.17    0    0
          5020 DD6S  1        238          2    32680    0.03   20.42    0    0
...

Total     5020 DD6S  0       2674          0     6617    0.03   13.27    0    0
          5020 DD6S  1        252          2    37362    0.03   22.13    0    0
```

sar -v

•Reports status of text, process, nc1inode, and file tables.

ginger% sar -v 1 4

sn9063 ginger 8.0.4 viz.10 CRAY J90 10/08/96

```
15:44:05 text-sz proc-sz  inod-sz  file-sz text-ov proc-ov inod-ov file-ov
15:44:06   8/ 40 119/650 967/2500 321/2100       0       0       0       0
15:44:07   8/ 40 119/650 967/2500 321/2100       0       0       0       0
15:44:08   8/ 40 119/650 967/2500 321/2100       0       0       0       0
15:44:09   8/ 40 119/650 967/2500 321/2100       0       0       0       0
```

sar -b

•Reports buffer activity.

ginger% sar -b 1 5

sn9063 ginger 8.0.4 viz.10 CRAY J90 10/11/96

```
10:21:42 bread/s lread/s %rcache bwrit/s lwrit/s %wcache pread/s pwrit/s
10:21:43      12      33      65       7      20      67       2       0
10:21:44       0       1     100       0       0       0       0       0
10:21:45       1       5      80       1       3      67       0       0
10:21:46      48     151      68      34     101      67       2       0
10:21:47      46     144      68      33      99      67       2       0

Average       21      66      68      15      44      67       1       0
```

ldcache

LDCACHE(8) Cray Research, Inc. SR-2022 8.0

NAME

ldcache - Assigns and displays logical device cache

SYNOPSIS

/etc/ldcache -l dev [-h high[,low]] [-n units] [-r rate] [-s size]
[-t type] [-x max[,min]] [-p] [-w]
/etc/ldcache [-a] -b
/etc/ldcache [-a] -i
/etc/ldcache [-f file]
/etc/ldcache [-]

ldcache Display

```
ginger% ldcache
T Unit Size    Reads     Writes       Hits     Misses    Rate   Name
- ---- ----  ---------- ----------  ---------- -------  -----  --------
M  200   28       20605         75       17365     542  96.97  /dev/dsk/src4
M  200   28     1436837     107369      562683   53550  91.31  /dev/dsk/root4
M  200   28      216462      27482       89802    9373  90.55  /dev/dsk/usr4
M  450   28    45895237   19999584     9015956 2440107  78.70  /dev/dsk/tmp
M  200   28       36751      48400      102038    1725  98.34  /dev/dsk/spool
```

fsmon

FSMON(8) Cray Research, Inc. SR-2022 8.0

NAME

fsmon - Interfaces with the file system monitor fsdaemon(8)

```
ginger% fsmon
 n  File System    Status     Use   Warn   Crit  Time encountered
 1 /dev/dsk/root4 E-------   35.3%  90.0%  97.0%
 2 /tmp           E-------   26.0%  80.0%  90.0%
 3 /usr           E-------   57.7%  90.0%  95.0%
 4 /usr/adm       E-------   15.0%  90.0%  95.0%
 5 /usr/local     E-------   28.8%  91.0%  95.0%
 6 /usr/spool     E-------    9.8%  85.0%  95.0%
 7 /usr/users     E-------    8.8%  82.0%  90.0%
```

df (Display Free)

```
ginger% df -p /tmp
/tmp       (/dev/dsk/tmp     ):  8557102 sectors      0 trks  194029 I-nodes
                  total:  11568128 sectors     (0 trks)    229376 I-nodes

      Big file threshold:          32768 bytes
      Big file allocation minimum:     21 blocks

      Allocation Strategy:    round robin files
                              round robin all user data

           part   start     total         free (%)          frags (%)      device
           ----  --------  --------  -----------------  ----------------  --------
              0         0   1446016    870051 ( 60.2%)    195 (  0.022%)  tmp_1200
              1   1446016   1446016    749042 ( 51.8%)    823 (  0.110%)  tmp_1201
              2   2892032   1446016   1325030 ( 91.6%)    490 (  0.037%)  tmp_1202
              3   4338048   1446016   1333244 ( 92.2%)    459 (  0.034%)  tmp_1203
              4   5784064   2892032   2044997 ( 70.7%)    939 (  0.046%)  Stmp0
              5   8676096   2892032   2234738 ( 77.3%)    818 (  0.037%)  Stmp1
```

Monitoring/After the Fact Tools

•Accounting
•Logs
•Commands
Monitoring/After the Fact Tools ?Accounting
•CSA Accounting
–Provides much useful data.
–session files
»csajrep
–pacct files
»acctcom
•Daily Reports are helpful in locating performance problems.

Accounting Daily Report

•Important to keep up with how the system is dong.
•We try to include many different aspects of the system.
Accounting: Sample Daily Report

```
================================[ CPU Usage ] ========= (ginger) - 8.0.4 ===
Averages for 10/09/96 [ 84 %Usr, 2 %Sys, 14 %Idl, 77 %MemUse]
======================================== U[.] == S[s] == I[ ] == M[m] =======
 95%    ..............    . ..    ..
 85%   mmmmmmmmmmmmmmmm....m..m.....m..        .                   ...mm m
 75%  m..............m....mm.m...m.m.   m.mmmm....m.....m.. .......m..
 65%  ..................mmm.....mmm.... ...m....mmmm.mmmmm.mmmmmmmmm....mmmmmmmmm
 55%  ................................m.....................................
 45%  .................................m.....................................
 35%  .................................m.m...................................
 25%  ........................................................................
 15%  ........................................................................
  5%  .....................................s..............................
======|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==|==
 Hour 12 1  2  3  4  5  6  7  8  9  10 11 12 1  2  3  4  5  6  7  8  9  10 11
```

ginger Outages for 10/09/96
Time Duration(Min)

Total: 0 (100.00% uptime for 24 period)
Sample Daily Report: Backlog

```
NQS BACKLOG
Running: 5 jobs requesting 56.6 more hours
Queued (premium):     0 jobs requesting   0.0 hours
Queued (full-rate):   4 jobs requesting  68.1 hours
Queued (discount 1):  0 jobs requesting   0.0 hours
Queued (discount 2):  0 jobs requesting   0.0 hours
Queued (discount 3):  0 jobs requesting   0.0 hours
```

Total: 9 jobs 124.7 cpu hours (15.6 YMP8 hours == 0.6 YMP8 days)

Sample Daily Report: Process
CPU usage in YMP8 hours (ie 24 hours per day).

	User Mode	System Mode	Total CPU	Microtasking used	Microtasking charged	# of Processes	Avg Cpu per Process (secs)

Processes terminating between Wed Oct 9 00:30:14 and Thu Oct 10 00:30:12.
Report spans 23:59:58 wall-clock hours.

User	16.7	0.2	16.9	1.3	6.8	7868	62.036
Staff	5.3	0.1	5.4	0.0	0.0	8883	17.525
System	0.0	0.0	0.0	0.0	0.0	12820	0.093
TOTAL	22.1	0.3	22.4	1.3	6.8	29571	21.811

Sample Daily Report: NQS
Note:Avg CPU: seconds of user+system time for job
Avg Expansion: elapsed time from beginning of execution / cpu time
Max Expansion: largest elapsed time for a single job
Avg Memory: sum of memory integrals for all processes of a job
/ sum of cpu time for that job
Q_Delay: time (sec) spent waiting in queue to run

NQS Stats:

Queue Name	Limits			#	Avg	expansion		Memory	Q_Delay	
	Run	CPU	Mem	Jobs	CPU	avg	max	Avg	avg	max
q16_unl	15	unlimited	16mw	6	24011	1	7	9.50	6461	19595
q32_unl	15	unlimited	32mw	6	32745	0	2	16.48	7802	44765
q4_unl	15	unlimited	4mw	2	64986	1	1	1.96	22267	22366
q4_3600	15	3600sec	4mw	2	2147	1	1	0.33	2	2
qe16_600	3	600sec	16mw	1	9	1	1	1.95	17020	17020
TOTAL				17	27930	1				
interactive				44	4296	3	9572	0.75		
other_batch				5825	1	133	85023	2.57		

0 resumes (0 jobs) 0 reruns (0 jobs)
Sample Daily Report: Top Ten
Top Ten Users by CPU

User	Account	CPU (secs)	Avg Mem (MWords)	Wght. NCPUS	Wght. MFLOPS
enyedy	mcslvkp	196022	16.0879	1.20	40.76
oneal	staff	178264	0.6587	0.00	0.00
wchiu	ctt9gfp	129996	1.9146	0.00	40.46
fengj	cbs2ejp	103462	7.3919	0.00	29.47
hirst	syreu1p	37990	14.7428	0.00	118.28
johnsonb	trsc8ep	6319	2.5808	0.00	0.34
landman	mtt1mgp	5397	0.4188	0.00	106.09
mountzia	ctsvoep	5137	3.5907	1.09	6.80
brar	ssslvqp	3195	3.8737	0.00	0.46
kochmar	staff	1385	0.1160	0.00	0.00

Wght. NCPUS - Weighted CPU average (0.00 represents no-multitasking).
Wght. MFLOPS - Weighted MFLOPS (vector, 200 possible).

Sample Daily Report: Top Ten
Top Ten Users by Average Memory

User	Account	CPU (secs)	Avg Mem (MWords)	Wght. NCPUS	Wght. MFLOPS
enyedy	mcslvkp	196022	16.0879	1.20	40.76
hirst	syreu1p	37990	14.7428	0.00	118.28
pilon	acsvo3p	900	8.5747	0.00	2.07
fengj	cbs2ejp	103462	7.3919	0.00	29.47
brar	ssslvqp	3195	3.8737	0.00	0.46
mountzia	ctsvoep	5137	3.5907	1.09	6.80
johnsonb	trsc8ep	6319	2.5808	0.00	0.34
wchiu	ctt9gfp	129996	1.9146	0.00	40.46
booker	ctsvorp	15	1.2227	0.00	0.00
oneal	staff	178264	0.6587	0.00	0.00

Wght. NCPUS - Weighted CPU average (0.00 represents no-multitasking).
Wght. MFLOPS - Weighted MFLOPS (vector, 200 possible).

Logs

•Log Files
–HPM data
»collected for every process over 20 cpu seconds
»stored in pacct file.
»similar functionality can be achieved by using hpmflop/global collection.
»dprx
•Daily Performance Report
–NQS logs

dprx output

ACCOUNT	USER	COMMAND	HI-MW	E-SEC	C-SEC	M	EXP	IO-SEC	IO%	MFLOPS	LPIO	MBYTIO	JID	PID
ctt9h1p	user3	rotlong	24.6	469	871	2	0.5	19	4%	162.2	1.0	129.3	575	10749
mcslvkp	user0	charmm24	19.3	60124	65334	8	0.9	5	0%	40.7	0.7	61.3	3162	63329
mcslvkp	user0	charmm24	19.3	60446	65472	8	0.9	4	0%	40.8	0.7	61.3	3161	63338
mcslvkp	user0	charmm24	19.3	59925	65187	8	0.9	4	0%	40.8	0.8	59.8	3244	6489
mtt1mgp	user1	mold	0.3	1151	1072	0	1.1	0	0%	116.6	2.0	1.3	9205	95128
syreu1p	user2	m94.exe	14.9	35671	21774	0	1.6	6159	17%	121.6	0.9	91886.7	3993	77979
syreu1p	user2	m94.exe	14.9	21290	15636	0	1.4	4491	21%	115.4	0.9	68866.3	3447	67608

•This information is reviewed on a daily basis.
•We do this on our C90 as well to help target codes that would be better suited to run on a smaller platform.

Commands

•sar -f file
–Use same options as previously discussed.
•fsmon (file system monitor)
•df (display free)
•setfs (change file system attributes)

Managing the Workload

•NQS (Batch System)
•Tools to help prioritize workload:
–renice
–FSS (Fair Share Scheduler)

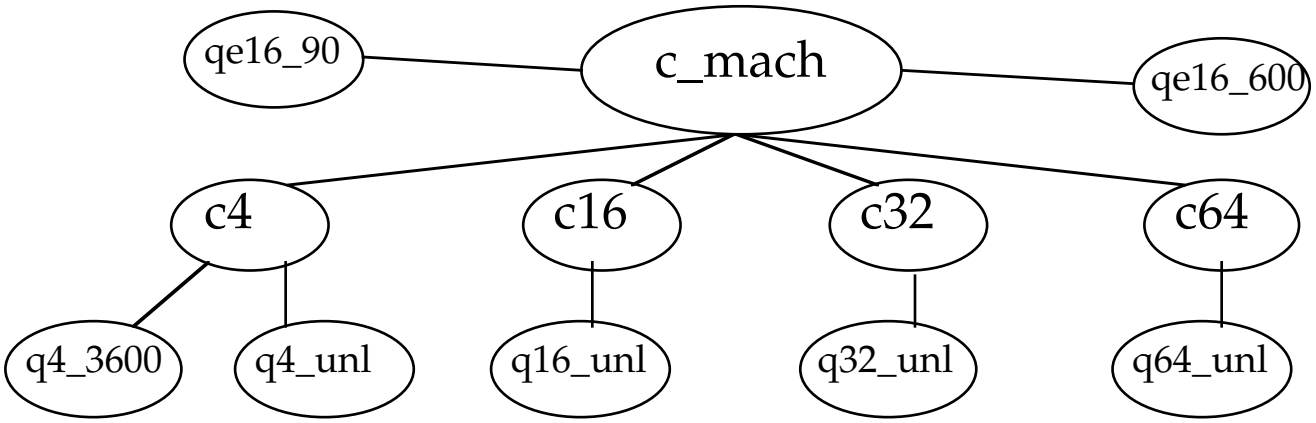
–Dedicated Runs

NQS

•Queues
–general purpose
–dedicated
–high priority
•Complexes
–Helpful in categorizing job classes.
•Limits
–Many to set: memory, cpu, disk, run, etc.

NQS at PSC

•Queues broken up by memory and time
•Memory
–4, 16, 32, 64mw
•Time
–90, 600, 3600, unlimited seconds
PSC NQS Layout



Managing the Workload/Prioritizing Workload

renice

•Sets system scheduling priorities of running processes.
•We don’t use this much.

Fair Share Scheduler

•Helpful for letting one or more users have a certain percentage of the cpu resources of the machine.
•Disadvantage: only schedules the CPU resource.
•Can be used to help setup a near dedicated run.

Dedicated Runs

•Can be beneficial for efficient use of the machine.
•We use scripts to drive this automatically during the night.
•Useful for large memory codes (>50% of memory) that use multiple cpus well.

Memory Performance

•nschedv provides many “knobs” for tuning.
•Memory scheduling needs to be balanced for batch and interactive loads.
•Is very site specific.
•Need to ask: What is my tuning goal?

nschedv

•nschedv

ginger% nschedv

```
[-H] hog_max_mem....     170000 clicks (83.0M), limit 201960 clicks (98.6M)
[-h] memhog.........       1000 clicks (0.5M)
[-c] cpuhog.........        999 Secs   (99900000000 clocks)
[-f] fit_boost......  -2.000000
[-M] mfactor_in.....1000.000000        [-m] mfactor_out....-1000.000000
[-T] tfactor_in.....  -1.000000        [-t] tfactor_out....   1.000000
[-P] pfactor_in.....   0.000000        [-p] pfactor_out....   0.000000
[-N] nfactor_in..... 500.000000        [-n] nfactor_out....-500.000000
[-G] in_guarantee...   0.000000        [-g] out_guarantee..   0.000000
                   ,   0.000000                            ,   0.000000
[-K] constant_in....   0.000000        [-k] constant_out... 250.000000
[-R] thrash-inter...          0        [-B] thrash-blks....          0
[-C] compress_intv..         60        [-r] cpu_factor.....         20
[-L] big proc.......          0        [-x] max_outage.....          0
[-V] max sched runs.          4        [-i] intrctve prfrd.          1
[-y] small proc.....          0        [-Y] itime..........          0
[-X] MPX scheduling.not allowed
```

Disk Performance

•Poorly configured file systems can seriously affect system performance.

•Must decide among trade-offs:

–Performance

–Capacity

–Reliability

Striping

•Places successive blocks of a logical device on successive devices.

•Increases the max bandwidth available to the logical device in proportion in proportion to number of devices striped.

•Use on file systems that do large block IO (swap, /tmp).

Banding

•Spreading a file system across two or more physical devices.

•This can help spread the access load.

•We use this for /tmp.

Mirroring

•Can be used to increase data reliability for file systems.

•Sometimes useful for inode partitions.

•We use this on our archiver (a J90).

•Provides multiple read paths which can lead to faster completion.

Type of disk

•We use DD-6S (fast wide SCSI-2).

•Formatted capacity is 9.11 GB.

•Transfer rate

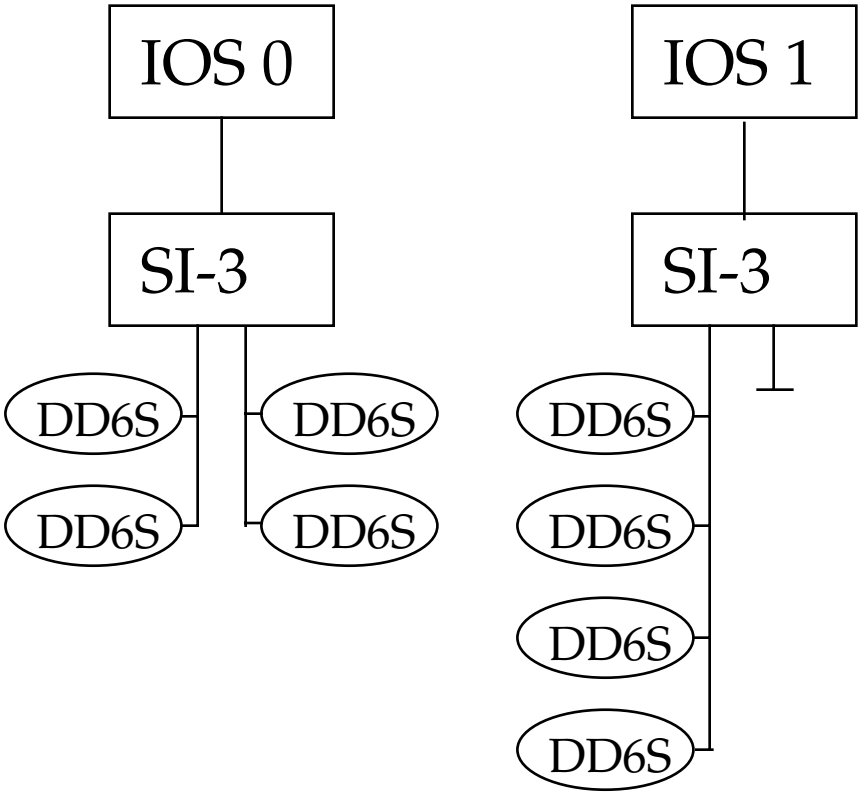
–7.2 Mb/s peak.

–4.2-6.2 Mb/s sustained.

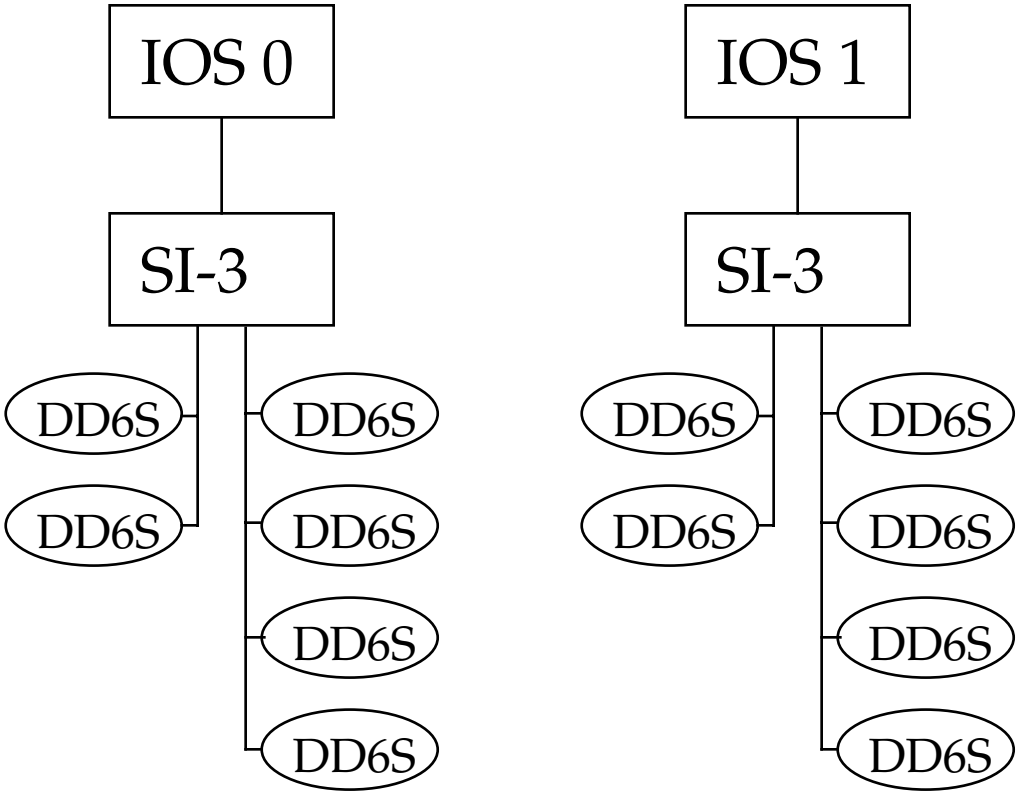
•Track size varies by zone.

•Cylinder size varies by zone.

Ginger IOS Configuration



Clove IOS Configuration



Daisy Chaining

- •Two or more drives on one channel.
- •Increases storage capacity.
- •Can decrease bandwith per disk.
- •ldcache can help.
- •All our disks are daisy chained.

Fragmentation

- •Noncontiguous allocation of data blocks for a file.
- •Increases seek time to access data.
- •Slows file system checks using fsck.
- •Use df -p fs to examine.

mkfs, setf

- •Need to wisely choose -A, -B options with mkfs.
- •Users can use setf to allocate contiguous blocks.

Partition Placement

- •Need to consider where partitions will be placed.
- •Important to consider channel activity and file systems on partition.
- •A poor choice can lead to horrible system performance.

mkfs

- •Useful to do a file size distribution on a key file system (ex. /tmp).

–ff is a usefule tool for this.

- •What size is a small file? Large file?
- •What percentage of all files are small? Large?
- •What percentage of space do all small files consume? Large files consume?

File Size Distribution (1/2)

ginger File Distribution for /dev/dsk/tmp
38240 data points, 38240 fits, 0 no fits

```
Mean:                44.1985
Min:                  0.0000  Max:             213876.0000
Pop.   Variance: 1872118.5826  Sample Variance: 1872167.5409
Pop.   Std. Dev:  1368.2538  Sample Std. Dev:  1368.2717
Pop.   C.O.V.  :    30.9570  Sample C.O.V.  :    30.9575
Num. of Buckets:         21
```

File Size Distribution (2/2)

BUCKET

LO	HI(incl)	COUNT	CUMCNT	PERCENT	CUMPCT	SUM	CUMSUM	PERCENT	CUMPCT
**********	1.0	29521	29521	77.20	77.20	10623.05	10623.05	0.63	0.63
1.0	2.0	1665	31186	4.35	81.55	2397.20	13020.25	0.14	0.77
2.0	4.0	1586	32772	4.15	85.70	4583.78	17604.03	0.27	1.04
4.0	8.0	2092	34864	5.47	91.17	12135.20	29739.22	0.72	1.76
8.0	16.0	1842	36706	4.82	95.99	20534.05	50273.27	1.21	2.97
16.0	32.0	604	37310	1.58	97.57	14146.11	64419.38	0.84	3.81
32.0	64.0	372	37682	0.97	98.54	18283.48	82702.86	1.08	4.89
64.0	128.0	138	37820	0.36	98.90	11169.64	93872.50	0.66	5.55
128.0	256.0	79	37899	0.21	99.11	14788.09	108660.59	0.87	6.43
256.0	512.0	103	38002	0.27	99.38	33189.54	141850.14	1.96	8.39
512.0	1024.0	47	38049	0.12	99.50	33705.20	175555.34	1.99	10.39
1024.0	2048.0	29	38078	0.08	99.58	42367.13	217922.47	2.51	12.89
2048.0	4096.0	52	38130	0.14	99.71	169408.17	387330.64	10.02	22.92
4096.0	8192.0	64	38194	0.17	99.88	359111.39	746442.03	21.25	44.16
8192.0	16384.0	41	38235	0.11	99.99	545764.18	1292206.21	32.29	76.46
16384.0	32768.0	3	38238	0.01	99.99	64800.00	1357006.21	3.83	80.29
32768.0	65536.0	0	38238	0.00	99.99	0.00	1357006.21	0.00	80.29
65536.0	131072.0	1	38239	0.00	100.00	119267.00	1476273.21	7.06	87.35
131072.0	262144.0	1	38240	0.00	100.00	213876.00	1690149.21	12.65	100.00
262144.0	524288.0	0	38240	0.00	100.00	0.00	1690149.21	0.00	100.00
524288.0	**********	0	38240	0.00	100.00	0.00	1690149.21	0.00	100.00

setfs

SETFS(8) Cray Research, Inc. SR-2022 8.0

NAME

setfs - Changes dynamic information in file system super block

SYNOPSIS

/etc/setfs [-B bf] [-A bu] [-L sl] [-U sl] [-a al] [-b flaw_list] [-c] |
[-i] [-s arbiter:semaphore_count] [-z] special

IMPLEMENTATION

...

DESCRIPTION

The setfs command makes changes to the file system super block without requiring you to make an entire new file system. You must unmount the file system before using this command to make alterations.

...

setfs Information

- Very handy for tweaking a file system.
- Most often use for resizing the bigfile threshold and bigfile allocation unit.
- Example:

ginger% setfs /dev/dsk/tmp

setfs: File system on /dev/dsk/tmp
*** LOWER security level = 0 UPPER security level = 16
*** Minimum allocation unit: 1 block
*** big file: 32768 bytes big allocation unit: 21 blocks
*** Allocation strategy: Round robin all files(rrf)
*** Panic on error
*** Inode allocation strategy: Enabled

Making Logical Devices

•Physical and logical devices can be made without rebooting using econfig -d on an edited Configuration file.
•This can help speed the file system parameter testing phase of setting up an optimal file system for your site.

df output of /tmp

/tmp (/dev/dsk/tmp): 9835697 sectors 0 trks 191121 I-nodes
total: 11568128 sectors (0 trks) 229376 I-no
des

Big file threshold: 32768 bytes
Big file allocation minimum: 21 blocks

Allocation Strategy: round robin files
round robin all user data

part	start	total	free (%)	frags (%)	device
0	0	1446016	1183189 (81.8%)	283 (0.024%)	tmp_1200
1	1446016	1446016	1120044 (77.5%)	497 (0.044%)	tmp_1201
2	2892032	1446016	1124327 (77.8%)	517 (0.046%)	tmp_1202
3	4338048	1446016	1260072 (87.1%)	491 (0.039%)	tmp_1203
4	5784064	2892032	2450256 (84.7%)	956 (0.039%)	Stmp0
5	8676096	2892032	2697809 (93.3%)	274 (0.010%)	Stmp1

mkfs example: tmp configuration

•Stmp0 and Stmp1 are two-way striped slices of /tmp.
•These can be accessed by a user doing a setf.
–Example: setf -p 4-5 ...

Configuring UNICOS

•ldcache
–Need to decide if using main memory for ldcache is desirable.
–Trickle Sync
•Kernel Tuning
–System buffers

ldcache Assignment

•Assignment

ldcache -l /dev/dsk/tmp -tMEM -s28 -n 450 -x90,60 -h350,300

•Trickle sync can smooth cache flushes.
•Options
-s size in 4k blocks of each cache unit.
Choose size that is used in the mkfs command to build file system.
-n number of cache units to assign.

ldcache Assignment (-x)

ldcache -l /dev/dsk/tmp -tMEM -s28 -n 450 -x90,60 -h350,300

-x max,min When age of any dirty cache unit exceeds max, flush all dirty units older than min.

•Disable LDSYNCTM

•Consider using if applications do frequent writes.

ldcache Assignment (-h)

ldcache -l /dev/dsk/tmp -tMEM -s28 -n 450 -x90,60 -h350,300

-h high,low Specifies threshold values for dirty units in cache. high specifies the maximum number of dirty units that may be in cache at any one time. If the number of dirty units exceeds high new requests to dirty units will sleep until the number falls below the threshold. When the number of dirty units in cache exceeds low, the system automatically starts flushing the oldest dirty units.

ldcache Assignment (-h)

•Provides stable read cache within larger read/write cache.

•Limits the number of dirty blocks that can be in the cache.

•Cache size - -h param = size of read cache.

•Need to determine relative amounts of IO made to ldcached file system.

Kernel Tuning

•System buffer cache

–too little

»degrades performance

–too big

»wastes memory

–Can increase throughput for IO-intensive applications that do not use raw IO.

–Experiment: use sar -b or -B to monitor.

Summary

•When tuning one part of system, important to consider all other parts.

•Tuning can be an art not a science.

•CRI provides many good resources in the tuning process.

Where to get more information

•CUG Talks

–UNICOS File Managment, Howard Mundy

–CrayTools: Perf. & Debug. Tools, Koushik Ghosh

–Mass Storage at the PSC, Phil Andrews

•Manuals

–UNICOS Tuning Guide (SR-2099)

–UNICOS Basic Admin. Guide for J90 (SG-2416)

•CRI Training

–UNICOS Performance Evaluation and Tuning (UPT)

–J90 System Administration (USAJEL)

–UNICOS Accounting (UACC)

•Books

–System Performance and Tuning, Loukides

–The Art of Computer Systems Performance Analysis, Jain