Jan Weaver Hewlett Packard







Problem:

Customer upgraded from JFS 3.1 to JFS 3.3 or upgraded from HPUX 11.0 to HPUX 11i and he now has performance problems with his application and/or system.

He notices an increase in the disk activity.





Glance shows a high level of physical disk activity and a low buffer cache hit rate.



iz c2607iem File Edit Connection Setup Macro Window Help											
B3692A	Glan	cePlu	s C.03	8.70.00	14:4	2:17 bok	maai 90(10/820	Current	Avg Hig	h
<mark>CPU U</mark> Disk U	ltil Itil	SS F					F		7% 73%	7% 10% 78% 83%	~ %
Mem U Swap U	ltil til	ន ប		SL	J U <mark>b</mark> Ur r			В	90% 47%	90% 90% 47% 47%	%
					DI	SK REPORT			Us	ers= 3	-
Req Ty	pe	R	equest	: s %	Rate	Bytes	Cum	Req %	Cum Rat	e Cum Byte	e
Local	Logi	Rds	292	100.0	56.1	18.5mb	245	52 100.0	53.6	18.5mb	
	Logi	Wts	0	0.0	0.0	Okb		1 0.0	0.0	Okb	
	Phys	Rds	1003	99.6	192.8	22.1mb	844	15 99.2	184.7	184.6mb	
	Phys	Wts	4	0.4	0.7	5KD	1 04/)4 U.8	1.4	127Kb	
	User	Mom	1003	99.0	192.8	22.1MD 0kb	844	+/ 99.3 n nn	184.8	184.0MD	
	Queta		U 1	0.0	0.0	0KU 6kb	c	U U.U 20 07	U.U 12	UKU 115kh	- 1
	Dau	5111	4	0.4	0.7	0KD 0kb	L L	יב ט.י ח ח ח	1.5	113KD 0kh	
Remote		Rds	n N	0.0	0.0	0kb 0kh		0 0.0	0.0	0kb Nkh	
		Wts	Ő	0.0	0.0	Okh		0 0.0	0.0	Okh	
	Phys	Rds	Ō	0.0	0.0	Okb		0 0.0	0.0	0kb	
	Phys	Wts	0	0.0	0.0	Okb		0 0.0	0.0	Okb	
									Ра	ge 1 of 2	
Proce Lis	ss t F	CPU Report	Me : Re	mory port	Disk Report		Next Keys	Select Process	Help	Exit Glance	
LogI Wts 0 0.0 0.0 0kb 0 0.0 0kb Phys Rds 0 0.0 0.0 0kb 0 0.0 0.0 0kb Phys Wts 0 0.0 0.0 0.0 0kb 0 0.0 0.0 0kb Phys Wts 0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 Process CPU Report Memory Report Disk Report Next Keys Select Process Help Exit Glance 447.1 HP70092 - 15.31.49.132 via TELNET Memory Memor											

HP World 2003 Solutions and Technology Conference & Expo



i <mark>p c2607iem</mark> File Edit Connectio	on Setup	Macro Window	Help				
B3692A Glar	ncePlus	C.03.70.01	06:32:00	bokmaai	9000/820	Current A	vg High
CPU Util Disk Util Mem Util Swap Util	SS F SU	SI	J UB UR R	F	В	7% 72% 7 90% 9 47% 4	5% 79% 6% 91% 0% 90% 7% 47%
			DISK R	EPORT		User	s= 3
Req Type		Requests	Rate	Cum Req	Cum Rate	High Rate	-
Read Cache Write Cache DNLC Hits DNLC Longs	Hits Hits	2072 2 0 0	26.7 25.0 0.0 0.0	455822 635 0 0	46.9 27.8 0.0 0.0	100.0 0.0 0.0	
Process List	CPU Report	Memory Report	Disk Report	Nex Key	t Select s Process	Page Help	2 of 2 Exit Glance
88, 1	HP70092 -	- 15.31.49.132 via 1	TELNET				



ÂP.	c2607	iem	<u> </u>										l ×
File	Edit	Connection	Setup r	/lacro Window Hi	elp								
B	3692	2A Gland	ceP l us	C.03.70.00		14:46:41	bokmaai	9000/820) C	urren	t Avg	High	
	:PU	Util	S S							 I 6%	 6%	10%	
D	isk	Util	F					F		I 79%	78%	91%	
M	lem	Util	S	SU		UB		E	B	I 90%	90%	90%	
S	шар	Util	U		UR	R				47%	47%	47%	
						TO BY	DT9K				lsers=	 3	
1	dx	Device	9	U	til	Qlen	KB/	Sec	Log I	IO	Phys	IO	
		8/4 5 (1	1/	2	 N N	1 3/	32	 na/		 0 6/	13	
	2	8/4.11	, .0	0/	Ō	0.0	0.0/	0.0	na/	na	0.0/	0.0	
	3	8/4.8.0)	18/	22	4.8	487.2/	795.9	na/	na	32.7/	41.8	
	4	8/4.9.0)	79/	78	10.4	2163.6/	2481.1	na/	na	136.1/1	32.6	
	5	8/4.10	0	2/	3	0.0	45.4/	92.1	na/	na	2.2/	4.9	
	6	8/4.0.0)	0/	0	0.0	0.0/	0.1	na/	na	0.0/	0.0	
	- 7	8/4.0.3	3	0/	0	0.0	0.0/	0.0	na/	na	0.0/	0.0	
	8	8/16/5	2.0	0/	0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
	9	8/8.1.0)	07	0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
	10	8/8.2.0)	07	0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
	11	8/8.3.0)	07	0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
T	op o	disk use	er: PIC) 3665, ps	eudo)	171.1	IOs/sec	s -	Se l ec [.]	t a Dis Page 1	k of 1	
	IO File	l By : Sys	IO By Disk	IO By Logi Vol	Su Spa	іар Ісе	Nex Key	t CPU s Proce	By essr H	Alarm listor	y Se l	lect	
4	47.1		HP70092	15.31.49.132 via TE	LNET								

11/17/2003

HP World 2003 Solutions and Technology Conference & Expo





We need to focus on the IO – who is doing it and why.

Kitrace can be used to look at the individual IO's and the system calls made by the process

In this case kitrace shows mostly random IO – Iseek, read, Iseek, read

However, occasionally we see sequential IO – Iseek, read, read

11/17/2003



pid=3665 read ret1=8192
pid=3665 lseek ret1=365633536
pid=3665 read ret1=8192
pid=3665 read ret1=8192
pid=3665 lseek ret1=466845696
pid=3665 read ret1=8192
pid=3665 lseek ret1=262332416
pid=3665 read ret1=8192
pid=3665 read ret1=8192
pid=3665 read ret1=8192
pid=3665 read ret1=8192
pid=3665 lseek ret1=204439552
pid=3665 read ret1=8192
pid=3665 lseek ret1=229343232





When the sequential reads occur, we see lots of physical IO being launched to the disks.



pid=3665 lseek ret1=118677504 ENQUEUE pid=3665 wr=read len=8192 ENQUEUE pid=3665 wr=read len=8192 pid=3665 read ret1=8192 ENQUEUE pid=3665 wr=read len=8192 ENQUEUE pid=3665 wr=read len=16384 ENQUEUE pid=3665 wr=read len=8192 ENQUEUE pid=3665 wr=read len=8192 ENQUEUE pid=3665 wr=read len=8192 ENQUEUE pid=3665 wr=read len=8192 ENQUEUE pid=3665 wr=read len=8192

ENQUEUE pid=3665 wr=read len=24576 ENQUEUE pid=3665 wr=read len=65536 ENQUEUE pid=3665 wr=read len=40960 ENQUEUE pid=3665 wr=read len=8192 ENQUEUE pid=3665 wr=read len=57344 ENQUEUE pid=3665 wr=read len=32768 ENQUEUE pid=3665 wr=read len=8192 ENQUEUE pid=3665 wr=read len=8192 pid=3665 read ret1=8192





System is doing read ahead when the sequential IO is detected.

Read ahead is more aggressive on JFS 3.3 than it was on JFS 3.1.

It is controlled by the vxtunefs parameters read_nstream and read_pref_io.



>vxtunefs /data Filesystem i/o parameters for /data read_pref_io = 65536 read_nstream = 10 read unit io = 65536write_pref_io = 65536 write nstream = 1write unit io = 65536 pref strength = 10buf breakup size = 131072 discovered direct iosz = 262144 max direct iosz = 655360 default indir size = 8192 qio_cache_enable = 0 $max_diskq = 1048576$ initial_extent_size = 4 max seqio extent size = 2048 max buf data size = 8192





Due to the generally random IO of the application the read ahead was unnecessary and in fact was likely harmful.

Filesystem parameters read_nstream and/or read_pref_io can be tuned to reduce the amount of read ahead that is performed.

Note that the application could also be changed to include code to advise the filesystem that the IO is random.



ile Edit Cor	nnection Setup	Macro	Window H	lelp						
B3692A (GlancePlu	s C.03	8.70.00	07:05	5:21 bok	maai 9000/0	B20	Current	Avg	High
CPU Ut: Disk Ut: Mem Ut: Swap Ut:	il SS il F il S il U		ទប	UR R		F	В	4% 79% 90% 47%	4% 79% 90% 47%	9% 86% 90% 47%
				DI	SK REPORT			 Us	 ers=	2
Req Type	e R	equest	: s %	Rate	Bytes	Cum Red	q %	Cum Rat	e Cum	Byte
Local I	Logl Rds Logl Wts	2077 0	100.0 0.0	185.4 0.0	234.1mb 4kb	30292 32	99.9 0.1	171.9 0.1	234.	4mb 4kb
	Phys Rds	1313	99.0	117.2	13.0mb	21 40 9	98.9	121.5	211.	6mb
l l	Pnys wis User	1312	1.U 98.9	117.1	22ко 13.0mb	241	1.1 99.0	121.6	211.	18KD 8mb
,	Virt Mem	0	0.0	0.0	Okb	3	0.0	0.0	2	2kb
1	System Dau	13	1.0	1.1 0 0	22kb 8kb	204 3	0.9 0 0	1.1 0 0	35	i2kb Mkh
Remote I	Logi Rds	0	0.0	0.0	Okb	5 0	0.0	0.0	Ľ	Okb
I	LogI Wts	0	0.0	0.0	Okb	0	0.0	0.0		Okb
I	Phys Rds Phys Wts	0 0	0.0 0.0	0.0 0.0	0kb 0kb	U 0	0.0 0.0	0.0 0.0		0kb 0kb
								Pa	ge 1 o	f 2
Process List	; CPU Repor	Me t Re	mory port	Disk Report		Next S Keys Pr	elect ocess	Help	Ex: Gla	it nce



<mark>.н́р</mark> с26	07iem	en Cabun	Manua III indone	u-h-						×
B36	92A Gla	ncePlus	C.03.70.0	0 07:05:5	i1 bokmaai	9000/820	Current	Avg	High	
СРU Dis Mem Sµa	Util k Util Util p Util	SS F S U	S	U UB UR R		F	3% 78% 90% 47%	4% 78% 90% 47%	9% 86% 90% 47%	
				 DTSK	 RFPORT		 lls	ers=	2	
Req	Туре		Requests	Rate	Cum Req	Cum Rate	High Rat	e 		
Rea Wri	d Cache te Cach	Hits e Hits	2030 1	75.3 16.7	82470 35	72.1 13.9	75.	3		
DNL	C Hits		0	0.0	0	0.0	0.	0		
DNL	C Longs		0	0.0	0	0.0	0. Pa	0 ge 2	of 2	
Pro	ocess List	CPU Report	Memory Report	Disk Report	Next Keys	t Select s Process	Help	Ex Gla	kit ance	
	1	HP70092	- 15.31.49.132 via							J



Customer sees a similar performance slowdown after adding online JFS.

Applications run slower and there is more physical IO than seen previously.



15.31.49.	.123.r1w - <u>Re</u> f	lection for I	HP						_	
File Edit C	ionnection Set	up Macro	Window H	Help						
🗅 🚅	. 4	•		• •						
B3692A	GlanceP	lus C.O3	3.70.00	15:48	3:22 c260	7ied 9000/8	89	Current	Avg Hig	h
CPU U Disk U Mem U Swap U	til S til F til S til Ul	SU U JR R	JB B				F	2% 98% 28% 18%	2% 25 58% 100 28% 29 17% 18	~ % % %
Reg Ty		Request	t s %	DIS Rate	3K REPORT Bytes	Cum Req	*	Use Cum Rate	ers= 3 e Cum Byt	e
Local	Logi Rds Logi Wts	5 699 5 0	100.0 0.0	35.8 0.0	5.75gb 20.3mb	43346 2501	94.5 5.5	 70.1 4.0	5.76gb 20.3mb	-
	Phys Rds Phys Wts	5 1235 5 32	97.5 2.5	63.3 1.6	308.5mb 97kb 1kb	23373 1862 66	92.6 7.4	37.8 3.0 0 1	5.70gb 24.1mb 203kb	
	Virt Men System	n 1 1264	0.1 99.8	0.0 64.8	1kb 308.6mb	19 25139	0.1 99.6	0.0 40.7	19kb 5.73gb	
Remote	Raw Logi Rds Logi Wts	1 5 0 5 0	0.1 0.0 0.0	0.0 0.0 0.0	8kb Okb Okb	11 0 0	0.0 0.0 0.0	0.0 0.0 0.0	88kb Okb Okb	
	Phys Rds Phys Wts	6 0 6 0	0.0 0.0	0.0 0.0	0kb 0kb	0 0	0.0 0.0	0.0 0.0	0kb 0kb	
								Paç	je 1 of 2	
Proces List	ss CPU t Repo	nt Re	mory port	Disk Report		Next Se Keys Pro	lect cess	Help	Exit Glance	
44, 1	HP700)92 15.31.4	19.123 via T	ELNET						Ē



<mark>P</mark> 15.31.49.123.r1w - R	Reflection for HP						_ □
ile Edit Connection !	Setup Macro Window	Help					
D 🖻 🖥 🎒	Þ 🛍 🗣 😰	▶ ● <u></u> ₩?					
B3692A Glance	Plus C.03.70.0	0 15:48:	37 c2607ied	9000/889	Current	Avg	High
CPU Util S Disk Util F Mem Util S Swap Util U	SU UB B UR R			F	2% 98% 28% 18%	2% 59% 28% 17%	25% 100% 29% 18%
Rea Tyne	Requests	DISK Bate	REPORT Cum Rea	Cum Rate	eu Binn Rat	ers=	3
 Read Cache Hi Write Cache H DNLC Hits DNLC Longs	ts 11 its 0 0 0	100.0 0.0 0.0 0.0	61748 3741 0 0	100.0 51.3 0.0 0.0	100. 0. 0.	0	
Process Cl List Rei	PU Memory nort Renort	Disk Renort	Nex1	t Select	Pa Help	ige 2 Ex Gla	of 2 xit
44, 1 HP	70092 15.31.49.123 via	TELNET					



15.31.49.123.r1w - Reflection for HP									1)>
File Edit Connection Setup Macro Wir	ndow Help	10							_
		R?							
B3692A GlancePlus C.03.7	70.00	15:49:01	c2607ied	9000/8	89	Curren	t Avg	High	
CPU Util 8						I 2%	 2%		
Disk Util F					F	1 98%	60%	100%	
Mem Util S SU UB	B					1 28%	28%	29%	
Swap Util <mark>U U</mark> R R						I 18%	17%	18%	
		IO BY	′DISK				 Users=	3	
Idx Device	Util	Qlen	KB/8	Gec	Log	I IO	Phys	3 IO	
1 10/0.5.0	2/2	 0.0	5.1/	 5.6	0.6/	48.3	1.7/	1.7	
2 10/8.8.0.255.0.1.3	0/ 0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
3 10/8.8.0.255.0.1.2	98/ 61	0.0	16213.3/10	0102.2	31.7/	23.4	63.3/	40.5	Ē
4 10/8.8.0.255.0.1.0	0/ 0	0.0	0.0/	0.1	0.0/	0.0	0.0/	0.0	
5 10/12/5.2.0	0/ 0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
6 10/0.3.0	0/ 0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
7 10/0.4.0	0/ 0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
8 10/0.6.0	0/ 0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
9 10/8.8.0.255.0.1.1	0/ 0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
10 10/8.8.0.255.0.1.4	0/ 0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
11 10/8.8.0.255.0.1.7	0/ 0	0.0	0.0/	0.0	0.0/	0.0	0.0/	0.0	
Top disk user: PID 12131	l, direct	t	73.7]	[Os/sec	S -	Selec	t a Dis Page 1	sk of 1	
	By Su	ian	Next	E CP		Alarm	Se	lect	
File Sys Disk LogI	Vol Spa	ice	Keys	s Pro	cessr	Histor	y		
44, 1 HP70092 15.31.49.1	23 via TELNET								-

HP World 2003 Solutions and Technology Conference & Expo



Again we can use Kltrace to see the characteristics of the IO.

Kparse will take the Kltrace output and extract such things as disk service times, queue lengths and disk block frequency.



From the Kparse report:

Disk block frequency...

Freq	Dev	Block	
597	dev_t=31/0x031200	blkno=0xb37340	wr=read
597	dev_t=31/0x031200	blkno=0xb37240	wr=read
4	dev_t=31/0x031200	blkno=0x538	wr=write
2	dev_t=31/0x031200	blkno=0xbee378	wr=write
2	dev_t=31/0x031200	blkno=0xb4f2ec	wr=write
2	dev_t=31/0x031200	blkno=0xb4f2cc	wr=write
2	dev_t=31/0x025000	blkno=0x3fb2b4	wr=write
2	dev_t=31/0x025000	blkno=0x30badc	wr=write



We see the same physical blocks being read from the disk multiple times during the short (20 second) data collection.

Why are these blocks being continuously read from the disk when the file system should be using the buffer cache and therefore the block should be available in the buffer cache?



If we look at a particular pid doing IO we can see what the IO looks like:

```
pid=12131 ktid=13338 lseek err=0 ret1=0
ENQUEUE dev_t=31/0x031200 pid-u/a=12131/12131 wr=read blkno=0xb37240
b flags=call/ndelay/busy/read/pftimeout/phys/
ENQUEUE dev t=31/0x031200 pid-u/a=12131/12131 wr=read blkno=0xb37340
b flags=call/ndelay/busy/read/pftimeout/phys/
 pid=12131 ktid=13338 read err=0 ret1=524288
 pid=12131 ktid=13338 lseek err=0 ret1=0
b flags=call/ndelay/busy/read/pftimeout/phys/
ENQUEUE dev t=31/0x031200 pid-u/a=12131/12131 wr=read blkno=0xb37340
b flags=call/ndelay/busy/read/pftimeout/phys/
 pid=12131 ktid=13338 read err=0 ret1=524288
 pid=12131 ktid=13338 lseek err=0 ret1=0
ENQUEUE dev t=31/0x031200 pid-u/a=12131/12131 wr=read blkno=0xb37240
b flags=call/ndelay/busy/read/pftimeout/phys/
ENQUEUE dev t=31/0x031200 pid-u/a=12131/12131 wr=read blkno=0xb37340
b flaqs=call/ndelay/busy/read/pftimeout/phys/
 pid=12131 ktid=13338 read err=0 ret1=524288
```

Here we see the same blocks being read repeatedly by the application (Iseek to position 0, read), the reads rather large (524288 bytes) and the IO bypassing the buffer cache (b_flags=phys)

pid=12131 ktid=13338 lseek err=0 ret1=0 ENQUEUE dev_t=31/0x031200 pid-u/a=12131/12131 wr=read blkno=0xb37240 b_flags=call/ndelay/busy/read/pftimeout/phys/ ENQUEUE dev_t=31/0x031200 pid-u/a=12131/12131 wr=read blkno=0xb37340 b_flags=call/ndelay/busy/read/pftimeout/phys/ pid=12131 ktid=13338 read err=0 ret1=524288



This is the discovered_direct_io feature of Online JFS.

Large reads typically are done once (backups or copies) and do not need to be kept in the buffer cache.

However, in this case the reads were repeated. The discovered_direct_io parameter should be tuned for this application.



vxtunefs /home/jan Filesystem i/o parameters for /home/jan read_pref_io = 65536 read nstream = 1read_unit_io = 65536 write pref io = 65536write nstream = 1write_unit_io = 65536 $pref_strength = 10$ buf_breakup_size = 262144 discovered_direct_iosz = 262144 max direct iosz = 1048576default indir size = 8192 qio cache enable = 0 $max_diskq = 1048576$ initial_extent_size = 2 max_seqio_extent_size = 2048 max buf data size = 8192



Ŕ	15.31.49.	123.r1w	- Refle	ction for I	HP							_ 🗆	×
Fi	e Edit C	onnectior	n Setup) Macro	Window H	Help							
	🗅 😅 (8 6	Ba I	B		• •							
	B3692A	Glan	cePlu	is C.03	3.70.00	16:3	8:19 c260	7ied 9000)/889	Current	Avg	High	
	 רסוו ווי	 + i	g		S.					1 27%	20%	27%	
	Disk II [.]	til	FF		00					1 3%	3%	3%	
	Mem U	til	S	ՏՍԼ	ĪB E	1				1 30%	30%	30%	
	Swap U	til	ŪU	R F	5	•				I 19%	19%	19%	
						DI	SK REPORT	•		Us	ers=	3	
	Req Ty	pe	R	lequest	t s %	Rate	Bytes	Cum F	Red %	Cum Rate	e Cum	Byte	
		Logi		1000	100 0	270 4					 ว		
	LUCAI	Logi	KUS Wte	1900		370.4 N N	3.4390 ՈԽհ	1	r 100.0 I N N	273.9	5.	43YD NVh	
		Phys	Rds	1	77	0.0	8kh	43	3 41 7	16	2	0 KD Omh	
		Phys	Wts	12	92.3	2.3	39kh		58.3	2.2	2	N9kh	
		User		0	0.0	0.0	Okb		4 3.9	0.1	_	9kb	
		Virt	Mem	1	7.7	0.1	1kb	1	1.0	0.0		1kb	
		Syst	em	11	84.6	2.1	38kb	97	7 94.2	3.7	2	. 2mb	
		Rau		1	7.7	0.1	8kb	1	I 1.0	0.0		8kb	
	Remote	Log I	Rds	0	0.0	0.0	Okb	() 0.0	0.0		Okb	
		Logi	Wts	0	0.0	0.0	Okb	() 0.0	0.0		Okb	
		Phys	Kds	U	0.0	U.U	Ukb	l	J U.U	U.U		Ukb	
		Phys	Wts	U	0.0	0.0	UKD	l	J U.U	0.0		UKD	
										Pa	ge 1	of 2	
	Proces	55	CPU	Me	mory	Disk		Next	Select	Help	Ex	it 📗	
	List	: F	Repor	t Re	port	Report		Keys	Process		Gla	ince	
1													-
	500, 1		HP70092	2 15.31.4	19.123 via T	ELNET							

HP World 2003 Solutions and Technology Conference & Expo



Interex, Encompass and HP bring you a powerful new HP World.



