DOES YOUR INSTALLATION HAVE HALF-SECOND I/O RESPONSE TIME?[1]

A Closer Examination of Disconnect Time

Alan M. Sherkow
I/S Management Strategies, Ltd.
(414)332-3062
al@sherkow.com

Abstract

Modern DASD Subsystems, regardless of the vendor, while processing typical workloads, provide fast average response times. These response times are vastly improved over the response times of 3380 and 3390 type DASD. I/Os that are cache misses are often much longer than the average reported because of the manner in which disconnect time is computed. This paper will describe variability in the response time that is not typically observed in these subsystems by using disconnect time in a new manner. A proposal is being made to combine existing data sources to produce more meaningful data for performance analysis.

1 INTRODUCTION

Disconnect time is a component of DASD I/O response time:

Equation 1

$$\overline{Response} = \overline{IOSQueue} + \overline{Pend} + \overline{Disconnect} + \overline{Connect}$$

The disconnect time in the RMF74 record is the total disconnect time for the device in the interval. The RMF Direct Access Device Activity Report computes Average Disconnect Time with the formula:

Equation 2

$$\overline{Disconnect} = Disconnect\ Time / Measured\ Event\ Count$$

That's the best that RMF can do with the data available to it. But we can do better! I/Os that complete by using the cache, either reading or writing do not have a disconnect component. Dividing by all I/Os distributes the disconnect time over read hits and write hits. ***The disconnect time should only be divided by the misses.***

2 MODERN DASD SUBSYSTEM ARCHITECTURE

2.1 DASD Arrays

DASD Subsystems have changed in the last few years in two areas: large cache sizes and small form factor disk devices packaged in various forms. Most of the implementations use DASD arrays in one RAID (Redundant Array of Inexpensive Disk) scheme or another. Ac-

cording to the RAID Concepts introduced by Gibson, RAID refers to implementations where multiple devices are always used to record data. The actual implementation of RAID, or if the subsystem does not use RAID is immaterial in this analysis. The analysis presented below is based on RMF data and on DFSMS Data Set I/O Statistics data.

2.2 Large Cache Controllers

The controllers in DASD Subsystems use very large cache sizes and typically use the cache for all reading and writing activity. Read-misses, for data accessed only once, or rarely accessed will require I/O to the disk devices. Write-misses in some implementations also require access to the disk devices before the data is accepted into the cache. All writes must eventually store the data on the DASD devices.

3 RMF DASD RESPONSE TIME

Equation 1 showed that DASD response time includes four components: IOS Queue, Pending, Disconnect and Connect.

IOS Queue is the time that an I/O is queued in MVS[2] waiting for other, previous I/Os to the same device to complete. MVS does not know about other I/Os that other MVS systems may have started to the same device. Depending on the use of SRM parameters in your system, the I/O queue for each device may be ordered by priority or by first-in-first-out[3]. Once an I/O is started, it cannot be preempted by a more important I/O. IOS Queue is the responsibility of the installation. That is, each installation decides which data sets are stored together on a volume. Data sets sharing a volume are the primary cause of IOS Queue delay, though there is also queueing for I/Os within a single dataset. Someone once told me there are no inactive data sets … just data sets that are not active now.

There is good news about IOS Queue. IBM's Enterprise Storage Server introduced the concept of Parallel Access Volumes (PAV) that allow multiple I/O operations to a logical volume. While this will provide improvement in DASD response time I do not expect it to resolve the issues addressed in this paper.

Pending is the time required to establish a path between MVS and the device for data transfer. Pend delays include waiting for shared devices, ESCON directors ports, EMIF physical channels, and DASD subsystem directors. (Pend time due to shared devices and EMIF physical channel, is also an installation's decision, not the DASD subsystem vendor's decision.)

Disconnect time is the time when a device is processing an I/O request, but the device is not transferring data. For a cache read-miss this is the time positioning the actual disk device so that the data can be read into the cache and sent through the channel. This can include latency, seek, sibling pend (see Artis, 1996), RPS reconnect, and backend reconnect. Another source of disconnect is CCR, channel command retry. It is a forced disconnect by presentation of a particular status, which permits other I/Os to then share the channel while the reason for the CCR is resolved. Read-hits and write-hits are generally considered to have no disconnect time component in their response time.

Once pending time is complete some channel time is used for non-data transfer activities. Positioning and searching commands are handled during this time, which is often termed

---

[2] MVS is intended to represent the operating and could also be OS/390.

[3] With SRM in Goal Mode the changing of I/O Priority is dynamic.

protocol time. Some of the protocol time happens during RMF Connect time, while other protocol activities happen during RMF Disconnect time. Connect time includes some protocol time and the time actually transferring data. The data transfer time is dependent on the speed of the underlying disk device, the speed of the "backend" between the cache and the disk devices, and the speed of the channel.

The disconnect time reported with newer technologies such as Iceberg, IBM RVA, Hitachi 7700, EMC Symmetric may include time that is normally reported as part of PEND in a traditional subsystem. This is when, although all backend paths are busy with data transfer operations, the I/O is accepted by the front end for initial CCW processing. The I/O is then disconnected, and the subsequent wait for the storage path is reported in DISC.

Service time is the sum of Pending, Disconnect and Connect. This is the "service" provided by the DASD subsystem and is the "measurement" that should be in your DASD subsystem performance clauses.

3.1 SMF74MEC is the Measurement Event Count

System/370 XA architecture introduced changes to how I/Os are processed and introduced special hardware to count events and record various measurements. The RMF 74 record has SMF74MEC, the measurement event count, which is the number of I/Os, or start subchannel commands.

3.2 SMF74DIS is the Number of Disconnect Counter Clicks

The disconnect time is kept as a counter and accumulated every 128 microseconds. Equation 3 computes the disconnect seconds by multiplying SMF74DIS by 128. Multiplying seconds by 1,000 as in Equation 4 provides the total disconnect time in milliseconds for the duration of the RMF record.

Equation 3

$$Disconnect\ seconds = SMF74DIS*128$$

Equation 4

$$Disconnect\ milleseconds = Disconnect\ seconds*1000$$

3.3 Reported Disconnect Time

The disconnect time per SSCH used in RMF reports, and most DASD performance analysis tools and reports is the average computed as:

$$\overline{DISms} = \frac{\left(\left(SMF74DIS*128\right)*1000\right)}{SMF74MEC}$$

4 RECOMPUTE DISCONNECT

SMF74MEC does not indicate if an I/O is a cache hit or not. It does not know if any disconnect time was actually incurred for any particular the I/O. We have the data to correct this. The data is recorded by the Cache RMF Reporter (CRR), and is now in RMF74 subtype 5[4]. The DASD subsystem control units have counters for total I/O, sequential I/O, Reads,

[4] RMF 4.3: OW19338, RMF 5.1: OW19337, RMF 5.2: OW18886, and all OS/390 RMF Releases.

Writes, etc. Total counter data is returned from the control unit, but the data in the RMF74-5 record is the difference of the counters between two consecutive intervals. We can use the read hit and write hit information to compute an average disconnect time for the I/Os that are not cache hits in Equation 5.

Equation 5

$$\overline{AdjDISms} = \frac{((SMF74DIS * 128) * 1000)}{SMF74MEC - rdHits - wrHits}$$

Now the various averages can be added together to approximate the response time of hits and misses. Using the formulas in this manner is probably not statistically pure. The denominators in computing the averages are not all the same. Some terms use SMF74MEC while the adjusted disconnect time term uses SMF74MEC less the cache hits. Further, the I/Os which are hits may be different from the I/Os that are misses. In particular if the block sizes are different then the connect time will be different; Equation 6 and Equation 7 assume the connect time, IOS Queue time and pending time are the same. Nevertheless, I propose that Equation 6 and Equation 7 are more meaningful than Equation 1.

Equation 6 Approximate HIT Response Time

$$\overline{Hit\ Rsp} = \overline{IOS\ Queue} + \overline{Pend} + \overline{Connect}$$

Equation 7 Approximate MISS Response Time

$$\overline{Miss\ Rsp} = \overline{IOS\ Queue} + \overline{Pend} + \overline{Connect} + \overline{AdjDISms}$$

One volume's RMF Direct Access Device Activity[5] report for one interval is below:

```
        D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y
                                        DEVICE   AVG  AVG    AVG  AVG  AVG
STORAGE   DEV  DEVICE   VOLUME  LCU ACTIVITY RESP IOSQ   PEND DISC CONN
 GROUP    NUM  TYPE     SERIAL       RATE     TIME TIME   TIME TIME TIME
          062B 33902    SLEEPY  003A  78.877     5  1.9    0.4  1.4  1.5
```

The volume's corresponding RMF Cache Subsystem Activity report is below:

```
                CACHE SUBSYSTEM DEVICE OVERVIEW
-------------------------------------------------------------------------------------
VOLUME     DEV  DUAL     %    I/O   ---CACHE HIT RATE--   ----------DASD I/O RATE----------
SERIAL     NUM  COPY   I/O   RATE    READ    DFW    CFW   STAGE  DFWBP    ICL    BYP  OTHER
SLEEPY    062B        21.5   78.9    52.9   20.1    0.0     5.9    0.0    0.0    0.0    0.0
```

The raw RMF 74 data for this device in this interval was analyzed to determine the additional data items needed:

SMF74MEC, I/O Count	70,989
Total disconnect time	00:01:38.99
Read Hits	47,587
Write Hits	18,128

[5] RMF reports average response time and IOSQ time as whole milliseconds. This table uses one decimal point precision by analysis of the raw data with MXG.

The number of misses is:

$$70{,}989\text{-}47{,}587\text{-}18{,}128 = 5{,}274$$

RMF had calculated the average disconnect time per as:

$$0{:}01{:}38.99 / 70{,}989 = 98.99 / 70{,}989 = 0.001394 = 1.4\ ms$$

Now, recalculate the disconnect time for the I/Os that are Cache Misses:

$$0{:}01{:}38.99 / 5{,}274 = 98.99 / 5{,}274 = 0.018769 = 18.8\ ms$$

We can calculate a response time for cache hits using Equation 6 as:

$$1.9 + 0.4 + 1.5 = 3.8$$

Equation 7 can be used to compute the response time of a miss:

$$1.9 + 0.4 + 1.5 + 18.8 = 22.6$$

The SSCH Rate for misses is calculated as

$$\left(Misses / IOs\right) * \left(IO\ Rate\right) = \left(5{,}274 / 70{,}989\right) * \left(78.877\right) = 5.860$$

This is summarized in the Table 1.

Table 1

D I R E C T A C C E S S D E V I C E A C T I V I T Y

DEVICE TYPE	VOLUME SERIAL	LCU	DEVICE ACTIVITY RATE	AVG RESP TIME	AVG IOSQ TIME	AVG PEND TIME	AVG DISC TIME	AVG CONN TIME
33902	SLEEPY	003A	78.877	5.2	1.9	0.4	1.4	1.5
MISS			5.860	22.6	1.9	0.4	18.8	1.5
HIT			73.017	3.8	1.9	0.4	0.0	1.5

Further we can separate the hits into read hits and write hits. The read hits are 47,587 leading to a read hit ratio of:

$$\left(47{,}587 / 70{,}989\right) * 100 = 67.0\%$$

A similar calculation produces write hits of 25.5%, and misses of 7.4%, leading to Table 2. Note that 92.6% of the I/Os take 3.8ms, while 7.4% take 22.6ms; 5.9 times as long.

Table 2

DEVICE TYPE	VOLUME SERIAL	% I/O	DEVICE ACTIVITY RATE	AVG RESP TIME	AVG IOSQ TIME	AVG PEND TIME	AVG DISC TIME	AVG CONN TIME
33902	SLEEPY	100.0%	78.877	5.2	1.9	0.4	1.4	1.5
MISS		7.4%	5.860	22.6	1.9	0.4	18.8	1.5
READ HIT		67.0%	52.875	3.8	1.9	0.4	0.0	1.5
WRITE HIT		25.5%	20.142	3.8	1.9	0.4	0.0	1.5

4.1 Does the Math Work?

First, the RMF total or accumulated response time in a second is 410.1604ms as computed in Equation 8.

Equation 8 RMF Cumulative Response Time

$$(SSCH\ Rate)\ (Resp) = 78.877 \times 5.2 = 410.1604\ ms$$

The similar calculation based on Adjusted Hit Response time and Adjusted Miss Response time is 409.9006 as in Equation 9. Though the result is not perfect, it is only off by 0.06%, which is most likely due to rounding.

Equation 9 Adjusted Cumulative Response Time

$$[(MissRate)\ (Miss\ Rsp)] + [(RdHtRate)\ (RdHt\ Rsp)] + [(WHtRate)\ (WHt\ Rsp)] =$$

$$[5.86 \times 22.6] + [52.875 \times 3.8] + [20.142 \times 3.8] =$$

$$132.436 + 200.925 + 76.5396 = 409.9006\ ms$$

Table 3 shows the results of Equation 6 and Equation 7 for another device.

Table 3

STORAGE GROUP	DEV NUM	DEVICE TYPE	VOLUME SERIAL	% I/O	DEVICE ACTIVITY RATE	AVG RESP TIME	AVG IOSQ TIME	AVG PEND TIME	AVG DISC TIME	AVG CONN TIME
	062B	33902	GRUMPY	100.0%	36.469	8.7	3.2	0.4	3.5	1.5
		MISS		7.4%	7.744	21.8	3.2	0.4	16.7	1.5
		READ HIT		67.0%	23.002	5.1	3.2	0.4	0.0	1.5
		WRITE HIT		25.5%	5.722	5.1	3.2	0.4	0.0	1.5

5 DFSMS DATA SET I/O STATISTICS (SMF 42-6)

We can further verify the data, and validate this analysis by applying the same concept to the DFSMS Data Set I/O Statistics that have been available since DFSMS/MVS 1.1.0. The data is recorded in the SMF 42 subtype 6 record. For each job, for each data set I/O statistics are gathered. The record is written at the end of each interval[6]. Initially the data was only collected for SMS managed data sets. PTFs, available in 1995, enabled collection of information for non-SMS managed data sets also.

---

[6] If the job is swapped out at the end of the interval it is not swapped in just to write the I/O statistics records. You cannot use only the "written to SMF" timestamp, but have to use start time also to determine if an SMF42-6 record has data for a device or data set you are interested in.

The response time components are the same as discussed in "3 RMF DASD Response Time". Additional information is also reported. Unlike the RMF74 data this data directly reports averages for response time, connect time, pend time and disconnect time. IOS Queue time is not reported and must be solved for using a variation of Equation 1. The number of I/Os is also recorded and information on the caching of individual data sets is recorded:

- Cacheable I/Os
- Cache Hits
- Write Cacheable I/Os
- Write Cache Hits
- Number of Sequential I/Os
- Record Cache Requests
- DCME Inhibit Misses

5.1 Recomputed Disconnect Time With DFSMS Data Set I/O Statistics

The sample data used in this paper is from an installation that is using SMF synchronization. This allows close matching of RMF74 data with SMF42-6 data. The RMF interval studied in "4 Recompute Disconnect", began at 14:14. Only one job is accessing the volume during this interval, and it is accessing 7 data sets. The I/O count during the 15-minute interval of SMF42-6 is 70,977 versus 70,989 for the RMF74 data above, less than a 0.25% difference. The calculations in Equation 10 are similar to Equation 5.

Equation 10

$$\overline{AdjDISms} = (\overline{DISms} * IOCOUNT) \Big/ (IOCOUNT - rdHits - wrHits)$$

For one data set, PTGUMSTR.DATA the components of the response time are shown in Table 4.

Table 4

	I/O Rate	Resp	Queue	Pend	Disc	Conn
Original Data	11.7	8.7	2.9	0.3	4.4	1.2
Recomputed Miss	2.9	22.2	2.9	0.3	17.8	1.2
Recomputed Hit	8.9	4.4	2.9	0.3	0.0	1.2

Remember the RMF reported average response time for the volume was shown in "4 Recompute Disconnect", to be 5.2ms. The recomputed RMF 74 response time for a miss was 22.6ms and the recomputed hit time was 3.8ms. For this data set, PTGUMSTR.DATA, the reported SMF42-6 response time is 8.7ms. Our recomputation shows 22.2ms for misses and 4.4ms for hits. Now we can recompute for all the data sets as shown in Table 5. The blocksize issue discussed in "4 Recompute Disconnect", is less of an issue here, because the data sets are being recalculated one at a time.

Table 5

I/O Count	Cache Hits	Resp-time	Orig-inal Disc	Recalced Disc	Avg ms for Hit	Avg ms for Miss	Block Size	Avg ms Connect	DSName
10,564	7,983	8.7	4.4	17.8	4.4	22.2	8,192	1.2	.PTGUMSTR.DATA
10,865	10,314	7.7	1.2	22.7	6.5	29.2	32,768	2.9	.AUDAFILE.DATA
14,745	13,979	4.7	0.9	17.2	3.8	21.1	1,024	0.9	.PTGUMSTR.INDEX
12,846	12,388	4.6	0.6	18.0	4.0	21.9	32,760	1.4	.DFHJ01A
5,210	4,962	7.3	0.8	16.1	6.5	22.7	512	1.8	.AUDAFILE.INDEX
15,938	15,362	4.0	0.4	10.6	3.6	14.2	8,192	1.2	.CHPR6A50.DATA
654	501	7.7	4.1	17.5	3.6	21.1	1,024	1.4	.CHPR6A50.INDEX
155	149	5.5	0.8	19.8	4.7	24.6	32,760	1.4	.CC01.DFHJ50B

6 DOES IT MAKE SENSE?

6.1 SMF 42-6 Disconnect Time When All I/Os are Cache Hits

Analyzing all SMF42-6 data where all I/Os are cache hits can validate the reasonableness of my proposed adjusting of disconnect time. Table 6 shows that 84.3% of the observations where I/O Counts = Cache Hits have 0.0ms of disconnect time. Why is there any disconnect time? Because the SMF42-6 data assumes an I/O is a cache hit if the disconnect time is less than about .5 milliseconds (see Berger). Note that these times: 0.128ms, 0.256ms, 0.384ms and 0.512ms correspond to 1, 2, 3 and 4 counts in the measurement block update facility.

Table 6 Average I/O Disconnect Milliseconds Per Ssch

AVGDISMS	NUM. OBSV	PERCENT
0.000	30925	84.3
0.128	5734	15.6
0.256	6	0.0
0.384	1	0.0
0.512	2	0.0

6.2 SMF42-6 Response Time When All I/Os are Cache Hits

We can also examine the response time when all I/Os are cache hits in Table 7.

Table 7 Response Time (In Milliseconds)

RESPTIME	NUM. OBSV	PERCENT
1	8587	23.4
2	17043	46.5
3	4172	11.4
4	1175	3.2
5	1170	3.2
6	1717	4.7
7	758	2.1
8	384	1.0
9	180	0.5
10	394	1.1
11	208	0.6

Though there are some outliers.

194	1	0.0
203	1	0.0
204	1	0.0
205	1	0.0
278	1	0.0

What are the characteristics of these long I/Os that should be fast?

Response Time	Max Service	Max Response	I/O Count	Avg IOS Queue	Avg Pnd	Avg Connect
278.4	1.4	278.4	2	277.0	.4	1.0
205.2	2.0	258.8	7	203.4	0.3	1.5

Most of these outliers all have long IOS Queue time as above, though some have long connect time. These provide an example of how one I/O is slowed down while waiting for other longer I/Os to complete. To further understand long I/Os that are hits a subsystem has been tested with a synthetic workload of 100% read hits. 256 devices were driven at 42 SSCH/Sec each. The results are in Figure 1's chart.

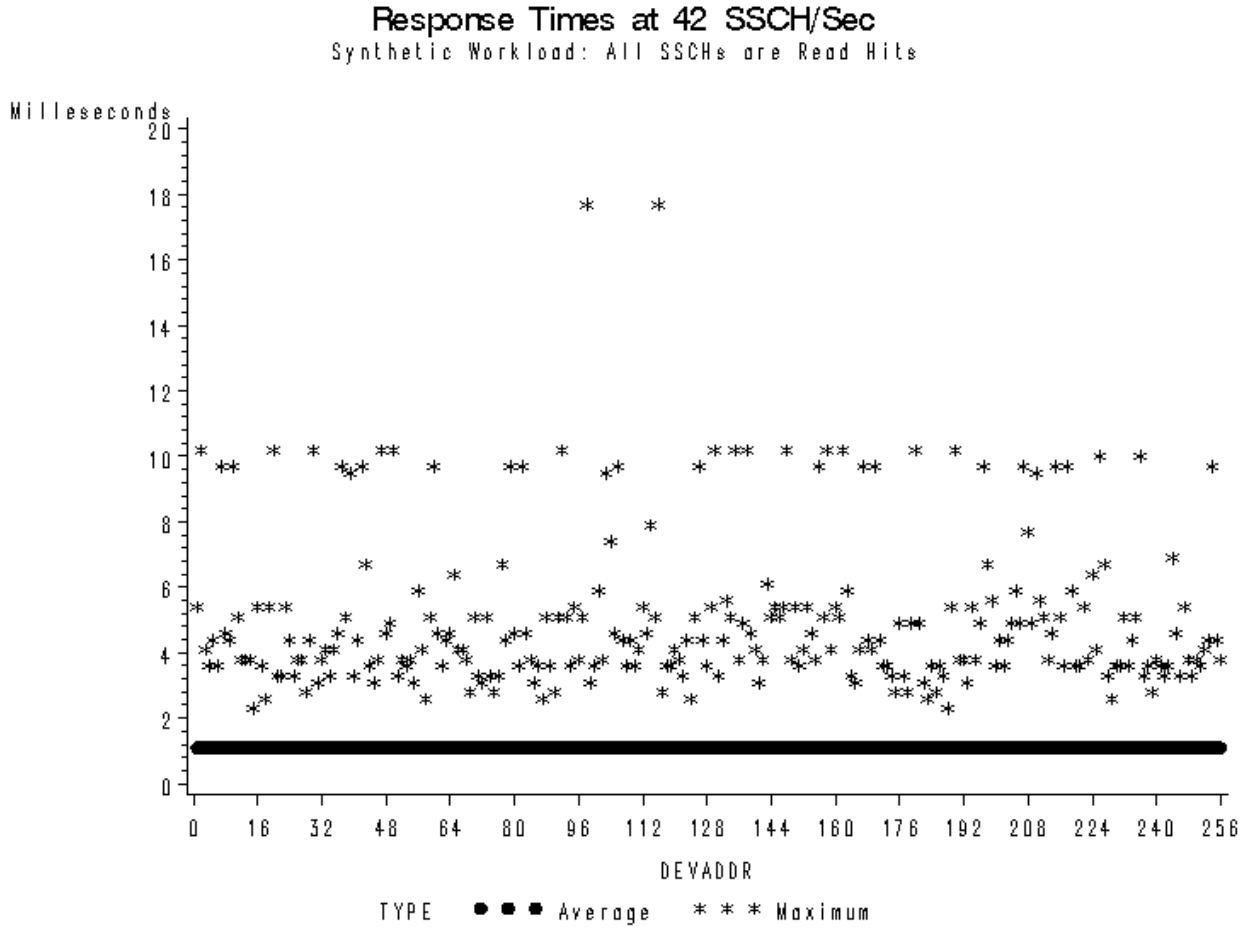

Figure 1

The average response time for each of the 256 devices was 1.1ms. The maximum response time were generally around 5ms. Even with a 100% read hit workload there were outliers. The outliers at 10ms and sometimes 18ms are not nearly as severe as the maximum response times in Table 8, but they reinforce the importance of looking beyond the averages. On single devices within a subsystem driven to 90 SSCH/Sec the maximum response time can exceed 130ms. What causes these delays? Protocol, cache management, queuing within the subsystem … the 'complexity'. How these long I/Os occur is an area for continuing study.

6.3 Analysis Of SMF42-6 Validates RMF 74 Above

Section 6.1 supported the assumption that Cache Hits typically do not have any disconnect time in their response time. Section 6.2 showed that for when all I/Os are hits 92.4% of the SMF42-6 data has response time of 6ms or less. The percent differences between the two, RMF74 and SMF42-6, and between the calculations in the table below validate this methodology.

Measure	RMF74	SMF42-6	Pct Difference
IO Count	70,989	70,977	-0.02%
Total Disconnect Time	0:01:38.99	0:01:32.84	-6.2%
Avg Disconnect Time	1.392ms	1.308ms	-6.0%
Recomputed Disconnect Time	18.8ms	18.5ms	-1.5%

7 LONG I/OS

The recalculations of "4 Recompute Disconnect" and "5.1 Recomputed Disconnect Time With DFSMS Data Set I/O Statistics" are still computing averages. I hate averages. I'd much rather have percentiles, or some other type of "bucket" data collection. But consider how much CPU time would be consumed if all the monitors collected all the data that anybody anywhere ever wanted! SMF42-6 data does collect maximums. We have two sets measures: average and maximum service time, and average and maximum response time. Recall that service time is pending time, disconnect time and connect time. Response time is IOS Queue plus service time. Earlier the response time for hits and misses were recomputed for 8 data sets on volume SLEEPY. The table below is for the same data sets with the maximums added.

I/O Count	CacheHits	Max Service	Max Response	Resptime	Avg ms for Hit	Avg ms for Miss	DSName
10,564	7,983	230.66	232.19	8.7	4.4	22.2	.PTGUMSTR.DATA
10,865	10,314	187.78	187.90	7.7	6.5	29.2	.AUDAFILE.DATA
14,745	13,979	117.76	183.80	4.7	3.8	21.1	.PTGUMSTR.INDEX
12,846	12,388	98.05	153.86	4.6	4.0	21.9	.DFHJ01A
5,210	4,962	62.46	784.26	7.3	6.5	22.7	.AUDAFILE.INDEX
15,938	15,362	58.50	811.01	4.0	3.6	14.2	.CHPR6A50.DATA
654	501	47.49	89.86	7.7	3.6	21.1	.CHPR6A50.INDEX
155	149	32.00	59.65	5.5	4.7	24.6	.CC01.DFHJ50B

Note that "AUDAFILE.INDEX" had an original RMF average response time of 7.3ms. The adjusted response times for hits and misses are 6.5ms and 22.7ms. The maximum service time is 62.46ms, and the maximum response time is 784.26ms. The maximum service time and the maximum response time are not necessarily for the same SSCH. It they are from the same SSCH then the IOS Queue time can be computed from Equation 11.

Equation 11

$$IOS\ Queue = maximum\ response\ time - maximum\ service\ time$$

If they are not from the same SSCH then the IOS Queue time of the maximum response time must be greater than the result of. In this case the IOS Queue of the maximum response time for "AUDAFILE.INDEX" follows:

$$IOS\ Queue \geq 784.26 - 62.46$$

$$IOS\ Queue \geq 721.8$$

How can the IOS Queue time be so large? It is because of competition between data sets on volume SLEEPY. Using Equation 11 for "CHPR6A50.DATA" the IOS Queue time is 721.8 or greater. Which long IOS Queue time came first? We cannot tell from the data. What is known is that these are online databases and a user was waiting for these long I/Os.

The chart in Figure 2 plots Maximum Service Time versus SSCH Rate. It should not be surprising that service times of nearly 1 second cannot occur at high SSCH Rates. How many one second I/Os can you do in one second to one data set on one device? Where does the long service time come from? The installation used in this example does not have shared DASD, EMIF, or ESCON directors, so the pending time should not be extending. I propose that the most likely time delay is in disconnect time. Either sibling pend delays, or other back end delays.

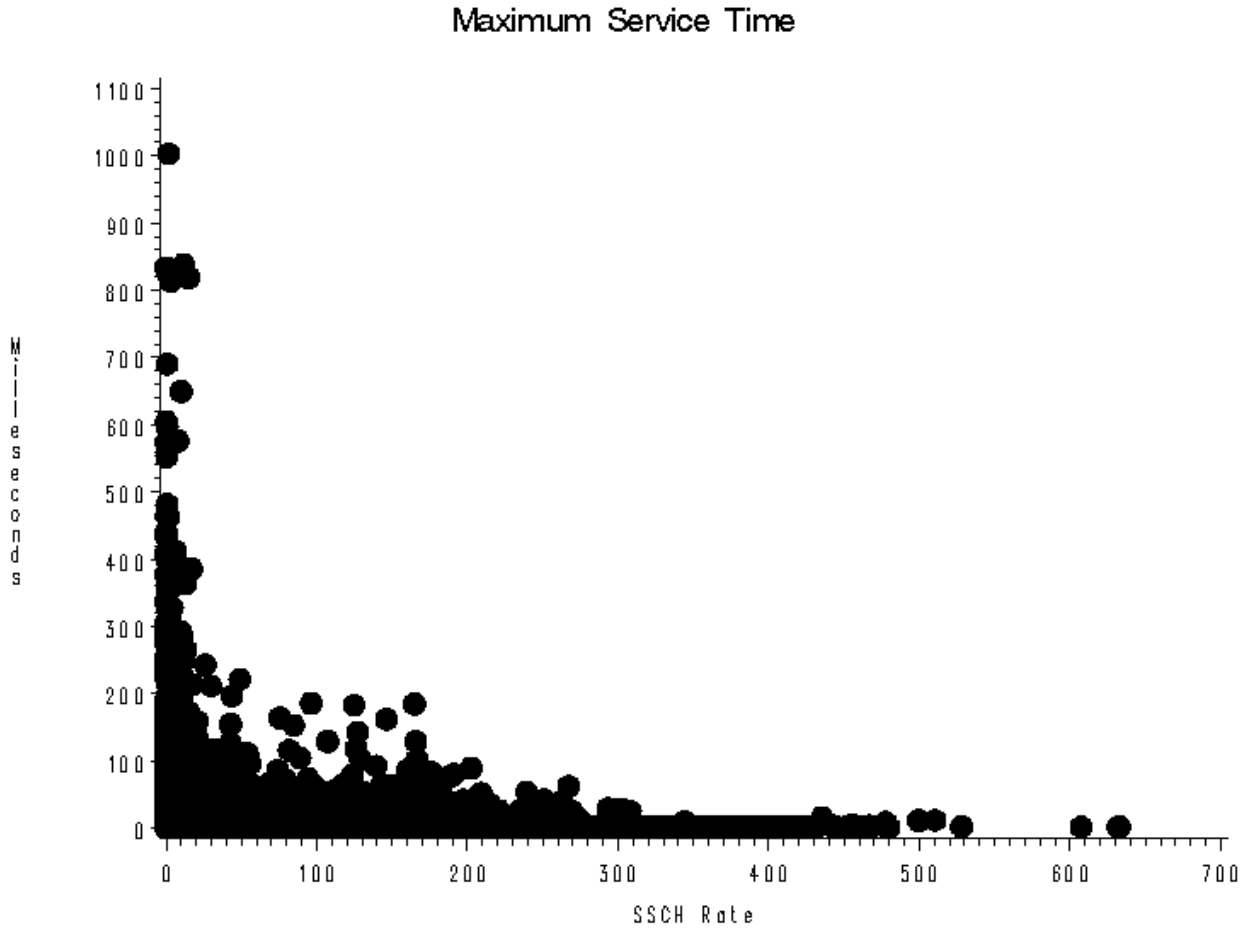

Figure 2

The maximum response times are in the following chart.

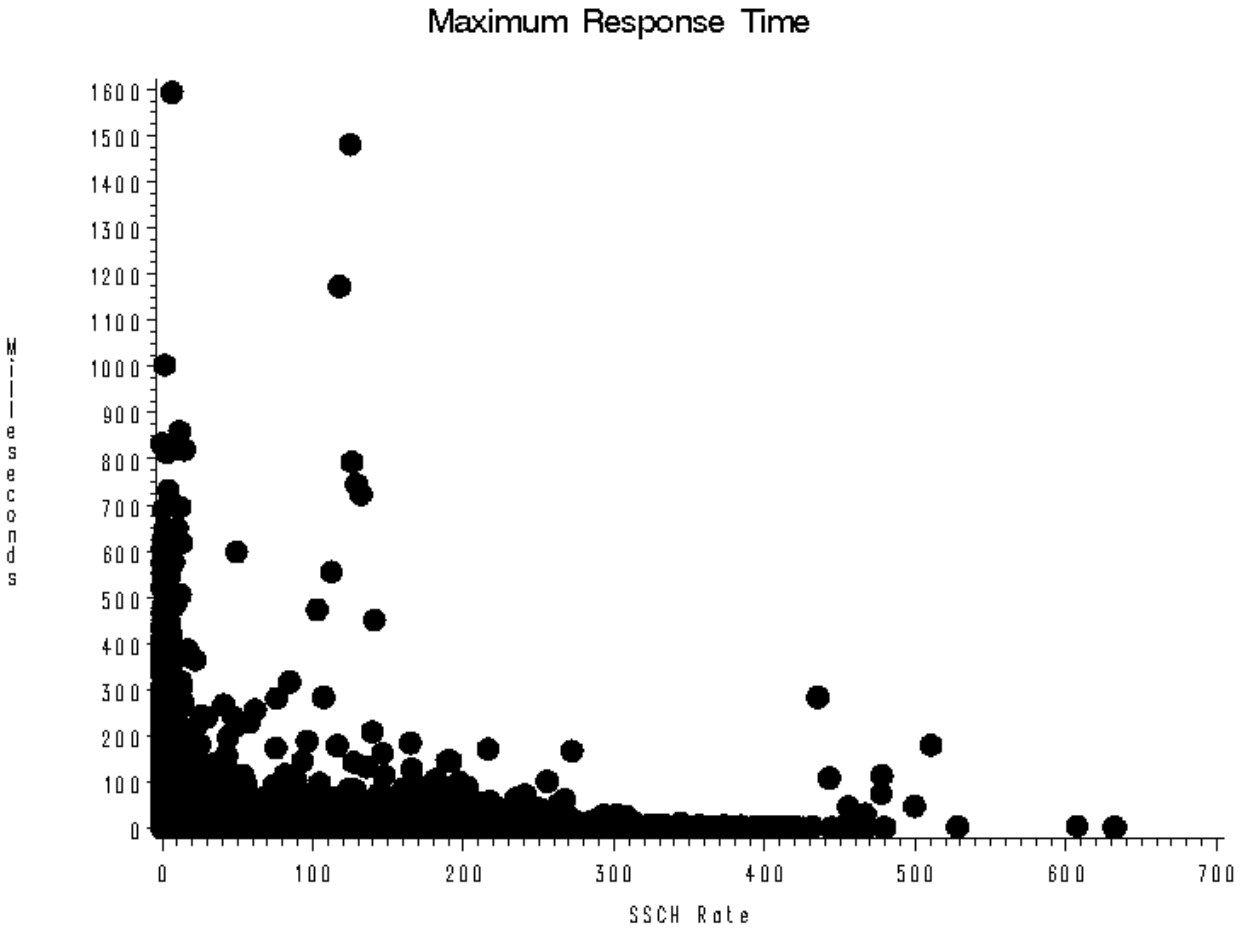

Figure 3

Let's examine the most obvious outlier points:

Table 8

SSCH Rate	MaxSrvTm	MaxRspTm	AvgMissResp	AvgHitResp
1.9	1003.65	1003.90	24.0	3.5
16.1	839.04	859.39	21.2	4.2
0.0	833.79	833.92	70.1	4.2
15.1	819.46	820.48	28.5	5.9
3.4	813.57	814.21	19.6	2.8

8 WHAT DOES THIS MEAN TO YOU?

Today's modern DASD subsystems do have some I/Os, even in normal operations[7], that do not have the 3-15 millisecond I/O time frequently associated with these devices. This paper has demonstrated that I/Os can be 500 milliseconds or more. The effect for online applications is occasional slowdowns in response time. Whether or not this is important to your installation depends on your desired service levels … but it probably is important. Remember the I/O queue discussed in "3 RMF DASD Response Time"? The queue may or may not be in priority order, but once an I/O has been started it cannot be preempted. The

[7] "Normal Operation" as compared to operations when recovery from an error is occurring.

I/Os with long response time, cause the I/Os following them to have long IOS queue time. The I/Os that queue behind a 500ms I/O wait half a second before they even start, regardless of their service time. The I/O load discussed in "4 Recompute Disconnect" had an I/O rate of 78.877. If the I/O arrival was truly uniform then during a 500ms I/O another 39 I/Os would queue on the UCB.

8.1 Does This Exist In Your Shop?

8.1.1 How To Check

Use equations 5, 6 and 7, presented in this paper to prepare reports for your site. First find outliers in the RMF74 data. Once you have identified specific volumes and time periods then begin to look at the SMF42-6 data to identify specific datasets.

8.1.2 What to do?

Now that you have seen that this does occur in your shop what should you do? First, determine if the occurrences require fixing. Are you meeting your service levels? The problem has been there a long time; do you even need to fix it? This is the same problem we faced in the 1980s early 1990s with DASD subsystems that had small cache sizes. As stated by J.P. Burg "the key indicator of effective cache performance is low disconnect time". You can move the data sets that are performing poorly to a few DASD volumes, or to their own DASD volumes. This gives you control over which volumes have long response time and the long I/O queue time, but you may not be making changes of any significance when viewed from the physical perspective of your subsystems back end. Some DASD subsystems allow you to define volumes of various sizes without loss of "real" DASD storage capacity. This capability can be used to effectively define an I/O queue per data set, for these particular data sets.

These recommendations seem quite contrary to using DF/SMS to control data set allocation and performance. Your service level objectives should control whether or not you do anything!

9 CONCLUSION

This paper has proposed recomputing disconnect time for RMF data and for DFSMS Data Set I/O Statistics data such that the SSCHs that are cache hits do not have a disconnect component. The disconnect time should only be attributed to the SSCHs that are cache misses. The RMF74 data has been validated by the SMF42-6 data.

You may not be able to do anything to resolve these very intermittent long I/O incidents, but hopefully you will better understand them, and be able to look for them in your installation.

10 REFERENCES

Aman, J.and C. K. Eilert, D. Emmes, P. Yocom, D. Dillenberger, "Adaptive Algorithms For Managing A Distributed Data Processing Workload," IBM Systems Journal 36, No. 2, 242-283, 1997.

Artis, H.P., MVS DASD Subsystems: Understanding, Evaluating, and Acquiring New Technologies, Performance Associates, 1995.

Artis, H.P., "Sibling PEND: Like a Wheel Within a Wheel," Proceedings CMG'96 Conference, December, 1996.

Berger, Jeffrey A., "DFSMS Data Set I/O Statistics", CMG Transactions, Winter 1993, pg. 75.

Burg, John P., "Cache Performance Management", GG66-3214, IBM, 1992.
Burg, John P., "Performance in a SMS World", SHARE Session 3242, March 1995.
EMC Corporation. *Symmetrix 3000 and 5000 ICDA Product Description Guide*. 1997
Gibson, G.A., "Performance and Reliaility in Redundant Arrays of Inexpensive Disks," Proceedings CMG'89 Conference, Reno, NV, December, 1989.
Grossman, C. P., IBM 3990 ESCON Function, Installation and Migration, IBM Washington Systems Center Technical Bulletin, GG66-3213.
Holtz, Jürgen M., "RMF Storage Subsystem Support", SHARE Session 2555, March 1997.
Houtekamer, Gilbert E. and H.P. Artis, MVS I/O Subsystems: Configuration Management and Performance Analysis, McGraw-Hill, Inc., 1993.
IBM, Analyzing Resource Measurement Facility Reports, LY28-1007.
IBM,Enterprise Storage Server,
http://www.storage.ibm.com/hardsoft/products/ess/ess.htm
IBM, OS/390 V2R5.0 MVS System Management Facilities (SMF), GC28-1783.
IBM,Enterprise Storage Server,
http://www.storage.ibm.com/hardsoft/products/ess/ess.htm
McAuley, Dave and Alison Pate, Gillian Docherty, Daniel Leplaideur, Baljinder Chana, Volker Bueffel, Willem Johannes, Nel Ian Black. IBM RAMAC Virtual Array. Document Number: SG24-4951-00. July, 1997.
Merrill, H.W. "Barry", Merrill's Expanded Guide to Computer Performance Evaluation Using the SAS System. Cary, North Carolina. SAS Institute, Inc. 1984.
Merrill, H.W. "Barry", Merrill's Expanded Guide Supplement. Cary, North Carolina. SAS Institute, Inc. 1987.
Samson, Stephen L., MVS Performance Management OS/390 Edition with MVS/ESA SP Version 5, McGraw-Hill, New York, 1997.