FLASHDA

Flash DASD Analysis

Jack Altman, IBM Poughkeepsie NY
altman@us.ibm.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

☆ Tool Objectives

"You can transform your knowledge of your DASD usage with just one run of FLASHDA!!!"

The purpose of the FLASHDA SAS® based tool is to provide users that are planning to exploit Solid State Drives (SSD) with a knowledgeable plan to manage the transition. The goal is to identify the volumes and datasets that will be most beneficial to reside on SSD.

FLASHDA will help make the best use of the new SSD feature in the IBM DS8000 storage subsystem with the IBM System z platform and z/OS operating system.

Some big advantages of the SSD drives compared to spinning disks are:

- Lower power consumption
- Higher I/O rates than traditional disks
- Lower latency than traditional disks
- Potentially higher reliability by eliminating moving parts.

An important fact to note is that even though the read and write latency for SSD is two orders of magnitude LOWER than traditional spinning disk, the latency to dynamic access random memory (DRAM) in the DS8000 cache is two orders of magnitude FASTER than SSD. Of course we would expect this fact that I/O requests directly from cache will provide the best service times. but the main point is, FLASHDA takes this into account to provide the best possible answers!

☆ Tool Reports

FLASHDA provides one report for dataset usage, another for DASD usage, and a final merged data report that shows final recommendations.

Since the I/O requests directly from cache will always be the fastest, we can reduce our service times caused by cache miss by moving those identified datasets to SSD.

The System z I/O architecture provides a detailed break down of the time spent executing I/O operations. One of these measurements is the Device Disconnect Time. This is a measure of the time spent resolving cache misses for READ I/O operations. Disconnect time also includes the time it takes to perform synchronous replication for WRITE I/O operations using, for example, IBM's Metro Mirror technology on the DS8000.

The FLASHDA report for datasets will show the highest READ ONLY Disconnect times on top, making it easy to find the datasets that had the highest cache misses! The report also has, for each captured dataset, the Disconnect time, response time, and much more. DFSMS APAR 0A25559 provides the new instrumentation to gather READ ONLY Disconnect time. Installing this SPE is recommended before using FLASHDA.

Another way we can look at it is with the FLASHDA report for DASD. We identify the volumes with the high I/O rates and a value for the "pseudo device" load is calculated in order to identify the devices that are causing the highest load on backend disks. The FLASHDA report for DASD will show the highest pseudo device load values on top, along with many other valuable statistics. The report will also show the number of cache to dasd transfers. This is a count of the number of of operations from DASD cache to the backend storage. The DASD report will also show Kbytes written, write response time, and more.

☆ Tool Input

The SAS code needs 2 types of SMF records to provide input for the calculations and the reports.
SMF type 42 subtype 6 are the records needed for dataset information.
SMF type 74 subtype 5 are the records needed for DASD information.

☆ Getting the Code

"Whenever you are asked if you can do a job, tell 'em, "Certainly I can!" Then get busy and find out how to do it." - Theodore Roosevelt (1859-1919)

FLASHDA is run with JCL using the SMF record input. It is written in SAS, so your system must have SAS installed to run it.

For users here in IBM Poughkeepsie, everything to run can be found on PLPSC or XRFMCL.
Both PLPSC and XRFMCL store the code in TOOLKIT.FLASHDA.SAS, and the JCL in TOOLKIT.FLASHDA.CNTL. XRFMCL has SAS installed, enabling a user to run directly on that system using their own copy of the code and JCL.

For users outside of IBM Poughkeepsie, a copy of the code and JCL will eventually be offered at the following web site:
http://www-03.ibm.com/systems/z/os/zos/downloads/

Code and Sample Reports

TOOLKIT.FLASHDA.CNTL ---------JCL
TOOLKIT.FLASHDA.LOAD ---------Load Module
TOOLKIT.FLASHDA.OUTPRT -----Merged Output
TOOLKIT.FLASHDA.OUTP42 ------Dataset Output
TOOLKIT.FLASHDA.OUTP74 ------DASD Output
TOOLKIT.FLASHDA.OUTREP -----Merged Report
TOOLKIT.FLASHDA.OUTREP2 ---Merged Report
TOOLKIT.FLASHDA.OUTR42 -----Dataset Report
TOOLKIT.FLASHDA.OUTR74 -----DASD Report
TOOLKIT.FLASHDA.SAS ----------SAS Code

Order of Execution

COPYSMF - If you would like to filter your SMF records by time or type.

TYPE42R1 - List the dataset statistics from SMF type 42 subtype 6 records.

TYPE74R1 - List the DASD statistics from SMF type 74 subtype 5 records.

TYPE4274 - Merge to list the top DASD with the datasets that reside on them.

The Process

- Get your SMF records
- Update the JCL
- Run a report for data set stats
- Run a report for DASD stats
- Run a report to merge for final recommendations

It's that EASY!!!

☆COPYSMF

- Provides a way to filter your SMF records

The COPYSMF member of the JCL provides an example of how a user can filter their SMF records into another output file that will contain only the record types requested AND the sampling time requested. The DUMPIN DD card is the original SMF records. The DUMPOUT DD card is the output. The SYSIN DD card contains our options. For FLASHDA to work, we need to code the OUTDD option for the required type 42 subtype 6 and type 74 subtype 5.

This gets us the needed record types, but if we have a large amount of data, or we wish to analyze the data for a specific time period, we can use the DATE option with the START and END options. These options allow the user to focus in on specific times of the day. This is useful, for example, for identifying data sets or volumes needed for market open, after lunch, or the start of batch processing. The DATE option will filter the records by the starting and ending Julian dates.

Code the starting time in START and the ending time in END.

Example COPYSMF Filtering Options for July 28, 2008 between 7AM and 8AM

```
INDD(DUMPIN,OPTIONS(DUMP))
OUTDD(DUMPOUT,TYPE(42(6),74(5)))
DATE(2008210,2008210)
START(0700)
END(0800)
```

☆TYPE42R1

- List the top data set names

The TYPE42R1 member of the JCL will run the TYPE42R1 SAS code. Our goal is to find data sets that will be best to move to an SSD drive. Using the input SMF type 42 subtype 6 records, TYPE42R1 will collect and calculate the data to provide a good sorted list. The top of the final report has the best candidates.

The data is sorted by Total Read-Only Disconnect Time. The data sets on the top of the report will be the ones with the highest number of cache misses. These top data sets are the ones that are making us spin out of control on our traditional storage devices. These are the data sets that we would want to move to our DASD with SSD, so we can cut down this time and get them in a flash!

You can use z/OS Dataset Mobility Facility (zDMF) technology to relocate the data for use after an application restart. Information on zDMF can be found at the following web site:
http://www.ibm.com/services/datamobility/
DB2 online Re-Org can move table spaces to volumes mapped by SSD drives.

☆ TYPE42R1 JCL

The SMF DD card points to the SMF record input.
The SYSIN DD card points to the SAS code.
The OUTR42 DD card points to the output report. This is the main output to look at.
The datasets with the top Read-Only Disconnect Time totals are on top.
The OUTP42 output points to the SAS PROC PRINT output. This is basically the same output, and it also includes a small bar chart at the end which shows a more visual view of the volumes with the top Read-Only Disconnect Time totals.
The SASWRK DD card is our output SAS library. It saves the created database information for any following code to extract from.

☆ TYPE74R1

- List the top DASD

The TYPE74R1 member of the JCL will run the TYPE74R1 SAS code. Our goal is to find the most worked DASD that will be best to relocate to reside on SSD.
Using the input SMF type 74 subtype 5 records, TYPE74R1 will collect and calculate the data to provide a good sorted list. The top of the final report has the best candidates.

The data is sorted by a field called Spinning Pseudo Device Load. This field, and another called SSD Pseudo Device Load, are calculated values that the code makes using DASD to Cache transfers, Cache to DASD transfers, the maximum number of I/O operations per second, and more. The main point is that these calculated values represent the load on an SSD device compare to a spinning device. By sorting on this, we get the most worked DASD at the top of our output report. These are the volumes that we would want to relocate to our DASD with SSD.

You can use Transparent Data Migration Facility (TDMF) technology to relocate these volumes.

☆ TYPE74R1 JCL

The SMF DD card points to the SMF record input.
The SYSIN DD card points to the SAS code.
The OUTR74 DD card points to the output report. This is the main output to look at. The volumes with the highest device loads are on top.
The OUTP74 DD card output points to the SAS PROC PRINT output. This is basically the same output, and it also includes a small bar chart at the end which shows a more visual view of the volumes with the top Spinning Pseudo Device Load values.
The SASWRK DD card is our output SAS library. It saves the created database information for following code to extract from.

☆ TYPE4274

- List the top DASD with their top data sets

The TYPE4274 member of the JCL will run the TYPE4274 SAS code. Our goal is to find the most worked DASD's data sets that will be best to move to an SSD drive.
Using the data saved from the TYPE42R1 and TYPE74R1 runs, the code will merge and calculate the combined data to provide a good sorted list. Once again, the top of the final report has the best candidates.

There are 2 output reports, each with the same data, but sorted different ways.
The OUTREP report is sorted by Spinning Pseudo Device Load. The most worked DASD is at the top of our report, with multiple lines listing the datasets on those DASD.
The OUTREP2 report is sorted by Total Write I/O Count. This gives us a better view of which data sets are best to move.

☆TYPE4274 JCL

The SMF DD card points to the SMF record input.
The SYSIN DD card points to the SAS code.
The OUTREP DD card points to the output report sorted by Spinning Pseudo Device Load.
The OUTREP2 DD card points to the output report sorted by Total Write I/O Count.
These are the two main outputs to look at.
This is the final recommended datasets and volumes that we would want to move to our DASD with SSD.
The OUTPRT DD card output points to the SAS PROC PRINT output. This is basically the same output, but also includes output that displays how the input parms filtered the data.
The SASWRK DD card is our output SAS library.

3 parms are passed.
1) **SET TOPVOL=50** This example limits the output to the data sets residing on the top 50 DASD.
2) **SET TOPDS=10** This example limits the output to the top 10 data sets for each of those top 50 DASD.
3) **SET DSSRUN=N** This parm is a little different. While the other parms are mainly used to reduce the output size, this parm will add more information to the report if set to Y. There is a load module, called DSSIZE, in FLASHDA.LOAD. The STEPLIB DD card points to it. When DSSRUN=Y, and we are on the same machine that the SMF data was collected from, this load module will find the size of each dataset.

☆ TYPE42R1 Output Values

In the TYPE42R1 output, we see one line of output values for each dataset. The following is an explanation of the fields used in the SMF records to obtain these output values.

REPORTED DATASET VALUES

Device Number ----------> **S42DSDEV**
Volser ----------> **S42DSVOL**
Data Set Name ----------> **S42DSN**
Total Read Only Disconnect Time ------> **S42DSRDD * S42DSRDT**
Average Read Only Disconnect Time --> **S42DSRDD**
Total Disconnect Time ----------> **S42DSIOD * S42DSION**
Average Disconnect Time ----------> **S42DSIOD**
Average Response Time ----------> **S42DSIOR**
Average I/O Connect Time ----------> **S42DSIOC**
Average I/O Pending Time ----------> **S42DSIOP**
Average Control Unit Queue Time ----> **S42DSIOQ**
Storage Class ----------> **S42DSSC**
Block Size ----------> **S42DSBSZ**
Total Read I/O Count ----------> **S42DSRDT**
Total Write I/O Count ----------> **S42DSION - S42DSRDT**

SAMPLE VALUES

Device Number ----------> **3603**
Volser ----------> **MR0065**
Data Set Name ----------> **OMVSSPN.ZFS.DATA**
Total Read Only Disconnect Time ------> **33835.13**
Average Read Only Disconnect Time --> **3.133**
Total Disconnect Time ----------> **34382.72**
Average Disconnect Time ----------> **3.133**
Average Response Time ----------> **7.185**
Average I/O Connect Time ----------> **1.121**
Average I/O Pending Time ----------> **.128**
Average Control Unit Queue Time ----> **0.000**
Storage Class ----------> **SMSOE**
Block Size ----------> **4096**
Total Read I/O Count ----------> **8527**
Total Write I/O Count ----------> **138**

Reported Times Are In Milliseconds

S42DSRDD And S42DSRDT are available with version of IGWSMF, 0A25688

Sorted With Highest Total Read Only Disconnect Time On Top

☆ TYPE74R1 Output Values

In the TYPE74R1 output, we see one line of output values for each DASD. The following is an explanation of the fields used in the SMF records to obtain these output values.

REPORTED DASD VALUES

Device Number ----------> **R745DEVN** or **R7451DVN**
Volser ----------> **R745DVOL**
Spinning Pseudo Device Load ---------> **R745DNTD,R745DCTD, R745CINT**
SSD Pseudo Device Load ----------> **R745DNTD, R745DCTD, R745CINT**
Cache to DASD Transfers ----------> **R745DCTD**
DASD to Cache Transfers ----------> **R745DNTD**
Device K Bytes Written ----------> **R7451CT2 * 128**
Write Response Time ----------> **R7451CT4 * 16**
Device MSECs Per K Byte ----------> **Write Response Time/Device K Bytes Written**
Device Write Sequential Requests ----> **R745DWSR**
Cache to Dasd/Write Ratio ----------> **R745DCTD/R745DWSR**
Device Type ----------> **R7451FLG**
Validity Flag ----------> **R7451INC**

SAMPLE VALUES

Device Number ----------> **3603**
Volser ----------> **MR0065**
Spinning Pseudo Device Load ---------> **7.095**
SSD Pseudo Device Load ----------> **0.116**
Cache to DASD Transfers ----------> **136991**
DASD to Cache Transfers ----------> **9**
Device K Bytes Written ----------> **9367552**
Write Response Time ----------> **264736**
Device MSECs Per K Byte ----------> **0.028**
Device Write Sequential Requests ----> **52910**
Cache to Dasd/Write Ratio ----------> **2.589**
Device Type ----------> **P**
Validity Flag ----------> **Y**

Sorted With Highest Spinning Pseudo Device Load On Top

Device Types

R=Raid Rank Data
P=Extent Pool and Storage Data
O=Other

Validity Flag

Y/N where Y = Valid Values (R745INC = 1)

☆ TYPE4274 Output Values

REPORTED VALUES

Volser ----------> **S42DSVOL, R745DVOL**
Data Set Name ----------> **S42DSN**
Total Write I/O Count ----------> **S42DSION - S42DSRDT**
Total Read I/O Count ----------> **S42DSRDT**
Average Response Time ----------> **S42DSIOR**
Average Disconnect Time ----------> **S42DSIOD**
Average Read Only Disconnect Time --> **S42DSRDD**
Total Disconnect Time ----------> **S42DSIOD * S42DSION**
Total Read Only Disconnect Time -----> **S42DSRDD * S42DSRDT**
Block Size ----------> **S42DSBSZ**
Data Set Size ----------> **DSSIZE Load Module**

Device Number ----------> **R745DEVN** or **R7451DVN**
Spinning Pseudo Device Load ---------> **R745DNTD,R745DCTD, R745CINT**
SSD Pseudo Device Load ----------> **R745DNTD, R745DCTD, R745CINT**
Cache to DASD Transfers ----------> **R745DCTD**
Device Write Sequential Requests ----> **R745DWSR**
Cache to Dasd/Write Ratio ----------> **R745DCTD/R745DWSR**

REPORTED VALUES

Volser ----------> **MR0065**
Data Set Name ----------> **OMVSSPN.ZFS.DATA**
Total Write I/O Count ----------> **138**
Total Read I/O Count ----------> **8527**
Average Response Time ----------> **7.185**
Average Disconnect Time ----------> **3.133**
Average Read Only Disconnect Time --> **3.133**
Total Disconnect Time ----------> **34382.72**
Total Read Only Disconnect Time -----> **33835.13**
Block Size ----------> **4096**
Data Set Size ----------> **0**

Device Number ----------> **3603**
Spinning Pseudo Device Load ---------> **7.095**
SSD Pseudo Device Load ----------> **0.116**
Cache to DASD Transfers ----------> **136991**
Device Write Sequential Requests ----> **52910**
Cache to Dasd/Write Ratio ----------> **2.589**

In the TYPE4274 output, we see one line of output values for each dataset. The merged DASD information is also included on each line.

Run time may greatly increase when we request data set sizes with **SET DSSRUN=Y**

OUTREP

Sorted With Highest Pseudo Device Load On Top

OUTREP2

Sorted With Highest Total Write I/O Count On Top

Data Set Size will be 0 when JCL options have SET DSSRUN=N