Pick a Volume, Pick Any Volume

SMS Volume Selection Algorithms

BY STEVE PRYOR

Back in the "good old days" of green screens and Buick-sized CPUs it was easy to determine where a dataset belonged. Users ruled. One merely looked at the JCL or IDCAMS control statements to see what volume the user had coded. If a different volume was needed, the VOL parameter was changed to the desired value.

Of course, user control of dataset placement had its disadvantages, too. Without centralized control, balancing the use of space across the system was impossible, leading to over-utilization and out-of-space errors on some volumes, while others remained mostly empty. Additionally, as leases expired or equipment was updated, it was difficult to remove old volumes from the system or add new ones because all of the JCL that referenced the old volumes had to be manually updated.

System-managed storage changed all that. Nowadays, except in the case of a dataset with a "guaranteed space" storage class, the user's volume serial number is ignored, and the storage management subsystem decides where the "best" place for the dataset is, in accordance with system-wide policies that aim to provide maximum utilization of disk space, proper performance, and assured backup and recoverability. These policies, implemented via the Automatic Class Selection (ACS) routines and the SMS constructs (data class, storage class, management class, and storage group) are under the control of the storage administrator. The details of dataset placement are hidden from the user, who is free to concentrate on more important work.

The process of introducing new hardware has been eased considerably with the advent of dynamic UCBs, HCDs (Hardware Configuration Dialogs) and system-managed placement of datasets. Automating the process of dataset allocation allows installations to use scarce DASD resources more efficiently. When problems arise, however, the storage administrator must be prepared to answer questions such as, "Why was this volume not chosen to contain this dataset?" and "Why was the 'wrong' volume chosen instead?" When utilization across volumes in a storage group becomes unbalanced, or when "small" datasets wind up in the "large" pool, the storage administrator must look to the volume selection process used by SMS for answers.

Volume selection begins once the ACS routines have been run and a set of SMS constructs has been associated with the dataset. As a result of ACS processing, one or more storage groups will have been assigned. Only volumes belonging to these storage groups, of course, are candidates to receive the dataset. Allocation then proceeds along one of two paths, depending upon whether the dataset will be striped (storage class SUSTAINED DATA RATE greater than zero) or non-striped. Most datasets follow the non-striped path. The eligible volumes are divided into two or three groups, depending upon their ability to satisfy the needs of the allocation.

Two sets of characteristics intersect to determine whether volumes belong to the "primary" list or the "secondary" list – those of the volume, and those of the storage class assigned to the dataset. Volumes in the primary list are those which can meet all of the dataset's requirements for performance and availability, which are ENABLEd for allocation, and which will not exceed the high-utilization threshold set for the storage group after the dataset has been created. An attempt is first made to place the dataset on one of the volumes in the primary list. The list of primary volumes is passed to the System Resource Manager (SRM) and the volumes are tried in order of channel path utilization. If no primary volume turns out to be suitable, then volumes in the secondary list must be tried.

What makes a volume a primary volume? The answer lies in the SMS storage class and storage group constructs. The storage group and volume status must be ENABLEd (volumes in QUINEW status are automatically placed in the secondary list) and the storage group utilization must still remain below the threshold once the allocation is complete. If these tests are met, each volume is examined to see if it can meet the requirements imposed by the storage class attributes, including:

- **IART (Initial Access Response Time)** — for ordinary DASD devices, this value must be zero. Non-zero values are usually used to select optical jukeboxes and the like.
- **ACCESSIBLITY and AVAILABILITY** — these attributes determine what types of devices (Concurrent-Copy capable, Dual-Copy, RAID-5, or RAMAC Virtual Array (RVA) can be included in the primary list. An important distinction here was introduced in DFSMS version 1.2, when NOPREFERENCE was introduced as the new default, instead of the previous value of STANDARD. STANDARD availability now indicates a simplex, non-RAID volume is needed, while NOPREFERENCE considers all types of volumes equally.
- **DIRECT and SEQUENTIAL MSR (Millisecond response time) and BIAS** — these attributes, which are used to select devices based on performance, are less important than they used to be, since virtually all devices are cached now. If an MSR value is available from the storage class, it is used to choose devices capable of that performance according to a table (the values in the table are documented in APAR OW08472). Devices with capabilities near those of the requested MSR (or, for DFSMS 1.1

and lower, any device with at least that MSR) and which have the necessary caching and DASD Fast-Write capabilities included in the primary list.

Once the list of primary volumes has been determined, the System Resource Manager (SRM) is called to order the list based upon channel path and device utilization. Each volume is then passed to DADSM, the Direct Access Device Space Manager, to actually allocate space on the device and create the necessary DSCBs. If DADSM rejects a volume, say because of excessive fragmentation or a non-zero return code from IGGPRE00, the next volume in the list is tried.

Probably the most common cause of datasets ending up on the "wrong" volume is that the primary list contains no volumes at all. This might happen, for example, if the storage class ACCESSIBLITY attribute specifies CONTINUOUS PREFERRED and LEVEL=VERSIONING, but no RVA devices with sufficient space below the storage group high-allocation threshold are available. Another instance can occur if multiple volumes are requested for the dataset, and a storage group does not contain the required number of volumes. In this case, all volumes in the storage group go to a "tertiary" list, which is only considered when both the primary and secondary lists are exhausted. This is particularly likely if a Tape Mount Management (TMM) request redirects a tape file to DASD, but the JCL still specifies VOL=(,,,255).

If the secondary volume list must be used, SRM is not called to order the volume list. Instead, the list is ordered in a hierarchy of space utilization, performance attributes and space availability. Usually, the determining factors are how close the volume's utilization is to the storage group limit and how much free space is available on the volume. Unfortunately, this often means that whenever it is necessary to use the secondary list, all allocations go to the volume with the most free space, leaving the utilization levels out-of-balance. APAR OW23333 was introduced for DFSMS 1.2 to address this problem. This APAR replaces the selection of the volume with the most free space with a randomization algorithm that makes it more likely that datasets will be spread evenly across the storage groups.

For striped datasets, the volume selection process is quite different. Additional restrictions are imposed, such as the requirement that all volumes remain below the high-allocation threshold, and that no more than 16 volumes (stripes) are allowed. An attempt is made to spread the allocation across controllers, choosing only one from each controller for the primary list.

SUMMARY

An understanding of the volume selection algorithms and their relationship to the SMS constructs is often the key to determining why good datasets end up in bad places.

A good explanation of the volume selection process can be found in the *DFSMSdfp Storage Administration Reference* (SC26-4940-04), while APARS II08004, II08987, and II08618 provide additional details.

NaSPA member Steve Pryor has more than 15 years of experience in storage management, disaster recovery, software development, and technical support. Steve can be contacted via the Internet at pryor@atlanta.com.