Generic Disk Driver Enhancement

Jim Hawkins and Hari L., vCSY

Introduction

Historically MPE/iX HP e3000 customers have been advised to only use disk devices that were sold and supported by HP. This advice was based upon the design of the MPE/iX disk drivers which depend upon features of the device hardware and firmware to guarantee reliable storage of data. In order to assist HP e3000 customers in their desire to run their systems for as long as possible and, as result of the Interex System Improvement Ballot (SIB), the Generic Disk Driver Enhancement (GDD) has been designed to remove as many of this dependencies as possible while still retaining the data durability that the HP e3000 is famous for.

It is made available via patch MPENX24 for C.75.00.

Caution: HP advises customers to always seek to use HP branded disk devices as a first choice, then to try the OEM versions of these devices and, only after these options are exhausted, move to devices from other manufacturers. HP in no way guarantees that any "non-HP" SCSI disk devices will perform adequately with this patch. HP is not liable for any data durability issues or performance issues that result from the use of this patch or usage of any storage devices not previously certified by HP as HP e3000 compatible.

The remainder of this article describes:

- Requirements for MPE/iX Disks of any type
- MPE/iX SCSI Disk Drivers
- Description of changes to each SCSI disk driver

MPE/iX requirements for Disks

MPE/iX depends upon some basic requirements for any disk device;

- Disks subsystems should support 256 Bytes Sector access to data. Actually in the case of SCSI disks physical block/sector size is 512 Bytes. ALL disk access on MPE/iX is therefore done in at least 512 Byte "chunks", for memory mapped structures, I/O size is 8*512Bytes for 4096 Bytes per page. Various subsystems handle "conversion" from 256 Byte sectors to 512 Byte (or larger) I/O in their data access routines. For example in DEBUG Sub-sector reads like "dsec 1.100,20" are converted into a single I/O to LDEV 1 for 2048 Bytes. Upon the completion of that I/O Debug displays only the bytes requested, $100-$119, from the larger buffer.

- Disks must be durable. This may seem obvious but there is an issue with "modern" disks as they often employ RAM buffers or caches to speed up access. Ideally MPE/iX "write" operations should bypass disk caches to guarantee durability. Realistically it turns out that in many disks caches cannot be avoided so we have a requirement that disks be protected by UPS to avoid losing the data in cache due to a power failure.

- Disk access must be "atomic" – that is, all or nothing. Upper layers should be able to assume that all data has been transferred when "good" status is returned. "bad" status indicates an inability to be sure of the state of the data; it is expected that "bad" status should be consistently returned. (Disk DMs monitor "suspect" sectors and move/recover data to spare areas to avoid data loss.)

- Disks must guarantee ordered access. Simply put, if I write to a location on Disk first "A" next "B" and then "C" in three distinct I/Os then when I read the data back it must be "C" not "B" or "A." SCSI disks typically support the multiple "in flight" I/Os and often have RAM buffer/caches to store data; in both cases we want to be sure that I/Os are processed in the order that they are sent.

MPE/iX SCSI Disk Drivers

When a device is added using SYSGEN/IOCONFIG, based on the device ID supplied, a physical manager (pmgr, a.k.a Device Manager or DM) is assigned to the device. The DM used by a particular device can be seen when a 'lp' (list path) of the device's physical path is done. The following DMs (pmgrs) in MPE/iX are specifically meant for disk drives:

scsi_disc_dm (SDM, for stand-alone disks)*,*
scsi_disk_and_and_array_dm (SDARRAY, for disk arrays)*,*

We can also use certain generic IDs to configure devices. For instance, a disk can be configured with ID=LVDDISK, disk arrays with ID=HPDARRAY. The appropriate pmgr is set based on the ID.

SDM is the original SE-SCSI implementation for "disc" devices. It also allows interfacing with removable media in the form of CD-ROM and Magneto-Optic disc devices. This Device Manager is a single threaded driver, only one SCSI I/O may be "active" at a time. Most typically this DM is configured for devices that are connected via an HP28642A NIO SE-SCSI HBA (controlled by the scsi_dam) this implies a "narrow" (8bit) "slow" (5-Mega-Transfers/ Second) connection.

SDA is a second generation SCSI-2 compliant disk device manager. This DM handles both JBOD and Disk Array Products. A key feature of this DM is that it can track and manage up to 8 "active" I/O operations at a time. Originally this DM was designed to connect with the NIO F/W SCSI HBA HP28696A for fast (10 Mega-Transfers/Sec), wide (16 bit), HVD (high voltage differential) devices. This driver was later modified for SCSI devices connected through A/N-Class PCI-SCSI LVD&HVD HBAs (controlled by pci_scsi_dam/c720) as well as the A6795A 2Gbit FC HBA (controlled by fc_dam/fcp_nm). This DM also has embedded the majority of "High Availability Failover" code – this allows the DM to detect certain problems with an I/O that may be due to HBA or physical path failure and "retry" the I/O on an alternate HBA and physical path.

GDD Changes to the SCSI Disc Device Manager

There are two changes to this device manager during device initialization in the following areas:

a. Asynchronous Event Notification
b. PAGE 8 (cache settings)

a) Asynchronous Event Notification (AEN) – AEN is the ability of the discs to bring to the attention of the I/O driver on asynchronous events like a disc getting 'ready' or it changing states. As per SCSI-2 standards, Discs support setting of two fields to indicate if AEN is enabled, READY_AEN_PERMISSION and UNIT_ATTN_AEN_PERMISSION. To enable AEN on discs, the driver first checks if these fields can be set. If not, AEN is not supported on the discs and hence is not enabled. If these fields are modifiable, AEN is enabled.

Prior to this enhancement, AEN enabling was done only for HP devices. But since most of the other devices also support AEN, the enabling of AEN has been extended to these discs also provided the specific disc supports setting of the AEN fields.

b) PAGE 8 (Cache settings) – Page 8 allows setting of two fields that control the use of read and write caches during disc I/O. These fields are called ‘Read Cache Disabled (RCD)’ and ‘Write Cache Enabled (WCE)’. These fields are set by the disc driver if the disc supports the use of a cache, RCD is set to FALSE and WCE is set to TRUE.

WCE was being set unconditionally to TRUE for all discs. This has been corrected and is now set only if the discs support caching.

GDD Changes to the SCSI Disk Array Manager

Device initialization by this manager has been modified in the following areas:

a. The TUR (test Unit Ready) sequence
b. Device identification
c. Page A (control mode page)

a) TUR sequence: There are three TUR calls during device initialization. The first call clears stale sense information from the target, the second clears stale sense information from the LUN and the third call checks if the device is ready to continue. If the third TUR call in the TUR sequence returns a “Not Ready” status, the entire TUR sequence is restarted. This may result in an infinite loop of high-speed TURs and can potentially hang the system. With this enhancement, a polling mechanism has been introduced after a few quick trials of TURs. This will result in the DM slowing down to retrying at 3 second intervals.

b) Device Identification: The DM uses the SCSI INQUIRY command to identify the device. In some cases INQUIRY data can return a status which indicates the device has been "stopped” and requires a "Re-Start". This status is detected and a restart unit command has to be issued appropriately. Care has been taken to avoid high-speed Identify-Restart-Identify-Restart loops. As in the handling of the infinite sequence of TURs during initialization, there are a few attempts of restart units quickly and then the frequency of retrials is reduced to once in 3 seconds.

c) Page A (control mode page) – The setting of the PAGE A values depends on the Queue Algorithm Modifier (QAM) field. The QAM field specifies restrictions on the algorithm used for reordering certain I/O commands. A value of zero in this field specifies that the target shall order the actual execution sequence of these commands and ensure that data integrity is maintained.

A value of one in this field specifies that the target may reorder the actual execution sequence of the commands in any manner. Any data integrity exposures related to command sequence order are explicitly handled by the initiator through the selection of appropriate commands.

For device configuration to succeed after making the page A settings, the QAM field should be modifiable and should be set to zero. Patch MPEMXA7 for C.75.00 allows queuing but prevents overlapping writes for all devices so that the QAM field becomes less ‘critical’. Hence the dependency on QAM being modifiable and being set to zero is removed in this enhancement.