HPE POINTNEXT

OpenVMS Performance & Availability Updates

OpenVMS Boot Camp 2017, SID 305

Keith Parris, with much help from Rob Eulenstein

Hewlett Packard Enterprise

Lock Value Block Sequence Number Overflow

- The Distributed Lock Manager uses Lock Value Blocks (LVBs) to pass crucial data around a cluster
  - Highest Sequence Number denotes most-recent data
  - Original design assumed 32-bit LVB sequence numbers would never, ever wrap
    - Then systems got much, much faster over decades of time
- If an LVB sequence number overflows, the most-recent data might now have a lower sequence number than older data, causing OpenVMS to make a wrong decision
  - Various possible symptoms can occur, including:
    - Directory entries not in alphabetical order
    - Files lost
    - Deleted files magically re-appear
    - Rare XQPERR bugchecks
- Fixed in:
  - HPE 8.4 Integrity: VMS84I_SYSLOA-V0300 kit released in August 2017
  - HPE 8.4 Alpha: VMS84A_SYSLOA-V0300 released in September 2017
  - VSI Integrity VMS842L1I_LCKMGR-V0100 for 8.4-1H1, 8.4-2, and 8.4-2L1 and undoubtedly a similar kit for Alpha

Lock Value Block Sequence Number Overflow

– How to see if you might be at risk for the problem:

```
$!
$!    Check_Value_Block_Sequence_Numbers.com
$!            Look for unusually-large lock value block sequence numbers
$     pid = f$getjpi("","PID")
$     temp_file := cvbsn_temp_'pid'.temp_file
$     open/write com 'temp_file'_com
$     write com "$ ANALYZE/SYSTEM"
$     write com "SET OUTPUT/NOHEADER ",temp_file,"_LIS"
$     write com "SHOW RESOURCE"
$     write com "EXIT"
$     write com "$ EXIT"
$     close com
$     @'temp_file'_com
$     delete 'temp_file'_com;
$     search 'temp_file'_LIS "Seqnum: 6","Seqnum: 7","Seqnum: 8","Seqnum: 9",-
"Seqnum: A","Seqnum: B","Seqnum: C","Seqnum: D","Seqnum: E","Seqnum: F"/window=(9,0)
$     delete 'temp_file'_LIS;*
$     exit
```

Corrections after VMS84x_SYSLOA-V0700 kit

- VMS84x_SYSLOA-V0700 kit had many valuable fixes, and also added support for Thin Provisioning
  - A couple of errors slipped in with the new Thin Provisioning support
- Symptoms:
  - Error: %SYSTEM-F-ILLIOFUNC, illegal I/O function code
  - On a volume marked for erase-on-delete, or a DELETE/ERASE command:
    - Error for non-privileged user: “-RMS-E-PRV, insufficient privilege or file protection violation” and the file is not erased
- Fixed in:
  - HPE 8.4 Integrity: VMS84I_F11X-V0400 & VMS84I_FIBRE_SCSI-V1100 and prerequisite VMS84I_CLIUTL-V0100
  - HPE 8.4 Alpha: VMS84A_F11X-V0400 and prerequisite VMS84A_CLIUTL-V0100

Lock Mastering average Lock Activity Value anomalies

- To make intelligent activity-based lock tree remastering decisions, OpenVMS keeps a running average of lock activity rates for each lock tree on each node in the RSB$W_OACT field of the root resource block for each tree
  - Rate is operations per 8-second remaster scan interval
    - You can see these values in the “Act” column in SDA> LCK SHOW ACTIVE output:

```
SDA> LCK SHOW ACTIVE

Active Resource Tree Information (Node XYZ)
-------------------------------------------
                         Total     Local
RSB Address              Locks     Locks   SubRSB     Act   Node      Resource Name
-----------------     -------   -------   ------   -----   ------    ------------------------------
FFFFFFFF.77F66BC0         18        18        7   65528   XYZ       RMS$.......XYZ_DATA1    ...
                                                             DISK$XYZ_DATA1:[APPL1]CUSTOMERS.DAT;1
FFFFFFFF.7BA5B840       1646      1646     1645   45678   XYZ       F11B$vXYZ_DATA1
```

  - Popular tuning strategy is to look at the top few locks listed, and try to reduce lock rates, by:
    - Adding RMS Global Buffers to RMS files
    - Reducing disk fragmentation, file open rates in applications, etc.
  - Field is only 16 bits wide
  - Rather than overflow, value “hits the peg” (like an analog automobile speedometer) at 65528
    - Implication: Rate is likely higher, but we can’t tell how much higher. If we make a tuning change, and the rate is still pegged at 65528, we can’t tell if we’ve made any improvement, unless perhaps we trace locks using the SDA extension LCK

Lock Mastering average Lock Activity Value anomalies

- OpenVMS fully intends to reset the lock new-activity counter accumulator RSB$W_NACT each 8-second interval, but
  - OpenVMS limits the lock-remastering recalculation activity each 8-second interval to about 16 milliseconds, then:
    - Keeps track of where it left off, and attempts to start over again at that point during the next 8-second interval
- This means on nodes with long root resource lists, the activity counter may not be reset as often
  - This tends to increase the average activity values
  - This may tend to keep lock trees on less-active nodes with long root resource lists, or even remaster them to there
  - This seems to happen at roughly a root resource list length size of ~55K to 65K on older Integrity machines
- To determine your root resource list length, use:

```
$ ANALYZE/SYSTEM
SDA> SET FETCH QUAD
SDA> VALIDATE QUEUE/QUAD @LCK$GQ_RRSFL !Count Root Resource List Length
Queue is complete, total of 193659 elements in the queue
```

Lock Mastering average Lock Activity Value anomalies

- The VMS84I_SYSLOA-V0300 and VMS84A_SYSLOA-V0300 kits added more counters to the SYS$CLUSTER execlet (visible only after issuing SDA> READ /EXEC) to assist in troubleshooting related issues:
  - LCK$Q_ENTRY_COUNT, the number of times the routine ADJ_COUNTERS has been called since boot to scan the root resource list
  - LCK$Q_RESTART_RSB, the number of times we timed out scanning the root resource list and saved the RSB address to start with in the next scan interval. ← Non-zero value here, continuing to increase over time, indicates the problem
  - LCK$Q_RESTART_AT_RSB, the number of times we tried to restart the root resource scan from a saved RSB address.
  - LCK$Q_RESTART_RSB_GONE, the number of times we tried to restart the root resource scan but the RSB we started on disappeared between time intervals and we had to start the scan from the beginning of the root resource list.
  - LCK$Q_COMPLETIONS, the number of times the root resource list scan has been completely scanned.
  - LCK$Q_LAST_COMPLETION, the quadword time of the last time we completed the scan of the entire root resource list. ← If this time is a long time ago (more than 8 seconds ago), this would indicate you have the problem
  - LCK$Q_PREVIOUS_COMPLETION, the quadword time of the previous time we completed scanning the entire root resource list.
  - LCK$Q_TIME_BTWN, the quadword time of the time between complete scans of the root resource list.
  - LCK$Q_MAX_TIME_BTWN, the quadword time of the maximum time between complete scans of the root resource list.
  - LCK$Q_RSBS_WALKED, the approximate number of RSBs scanned in the last time interval (within 1000 RSBs).

Lock Mastering average Lock Activity Value anomalies

- Remediation:
  - Faster CPUs can scan the root resource list faster
  - Lowering LOCKDIRWT on a node can reduce the length of its root resource list
  - HPE Technology Services may be able to help you examine the counters and perhaps adjust some new kernel tuning knobs available with VMS84{I|A}_SYSLOA-V0300 that can:
    - Raise the maximum amount of time allowed for root resource scanning, or
    - Enable a new feature which moves the already-scanned resources to the end of the root resource list

OpenVMS Cluster Performance in Less-than-ideal Networks

- PEDRIVER keeps track of how much time it should take to acknowledge a packet
- If an acknowledgement hasn't been received within what PEDRIVER considers a "reasonable" amount of time, PEDRIVER will assume it has been lost, and will proactively retransmit it, but:
  - When a retransmission occurs, PEDRIVER cuts the Current transmit window size to ½ of the Maximum value, and
  - If multiple retransmissions occur within a short period of time, PEDRIVER cuts the Current transmit window size all the way back to 1
  - Transmit window size grows (slowly) back up at a rate proportional to the rate of successfully-acknowledged packets
  - If we need to transmit without a free transmit window slot, we must wait -- counted as a Window Full (WinFull) event
- Some networks introduce a lot of jitter in packet delay, causing PEDRIVER to retransmit (and throttle back)
- Symptom: If this is happening a lot, in SCACP> SHOW VC /COUNTERS output you'll find the Retransmits counts and the Duplicates counts in the opposite direction are roughly equal numbers
- SYSGEN parameter PE2 allows you to tell PEDRIVER to be more patient before retransmitting
  - Units are number of 10-millisecond clock ticks, so value of 1 is 10 milliseconds extra; 2 is 20 milliseconds extra, etc.
- If most of the retransmitted packets are really getting lost (not just delayed), then raising PE2 would tend to hurt rather than help performance, because it would delay the retransmitting of lost packets

IPCI: TCP/IP Software vs. PEDRIVER: Who is smarter?

- With IPCI, each NIC has a different unique IP address
- PEDRIVER tracks paths on a per-endpoint (i.e. per-IP address) basis
- When PEDRIVER passed a UDP packet to the TCP/IP software for transmit, TCP/IP software sometimes thought it knew better and decided to send it out a different NIC than PEDRIVER intended, making it difficult for PEDRIVER to keep accurate path statistics.
  - TCP/IP software might even send a packet out through a NIC which has not been configured for IPCI or has been disabled for SCS traffic via:

    `SCACP> STOP IP_INTERFACE` *`xxn`*
- Fixed in VMS84I_DRIVER-V0400 plus TCP/IP HPE-I64VMS-NET_PAT-V0507-13ECO5-4 patch kit
- Must set SYSGEN parameter PE3 bit 5 (value of 32 decimal) to enable this fix

3PAR units may disappear from OpenVMS

– In a 3PAR remote copy group; adding a ***vv*** to or removing a ***vv*** from a remote copy group would cause OpenVMS to lose access to the ***vlun***
  – The 3PAR changed the Target Port Group (TPG) number for the device, and OpenVMS wasn't expecting the TPG to change without notice
– Fix is in VMS84I_FIBRE_SCSI-V1000

Non-paged Pool Corruption: Byte Order Reversed within Longwords

– Fibre Channel data standard is Big-Endian, but OpenVMS is Little-Endian

– Arguments to Endian conversion routine within PGQDRIVER were sometimes in error, causing OpenVMS to convert large portions of non-paged pool from Little-Endian to Big-Endian, corrupting data and causing crashes

– Data %xABCD1234 would be byte-wise reversed to %x3412CDAB

– Fix is available in VMS84I_FIBRE_SCSI-V1000

OpenVMS and Power Management

– Power management may be controlled at the firmware level (by the iLO console or Insight Power Manager software):

Power_Mode	Description	OpenVMS Implementation
Static high performance	The operating system makes no attempt to save power if there is any compromise in performance.	No power savings method used.
Static low power	The operating system saves power in every way it can, even to the detriment of performance.	On CPUs that support static low power, switch to the lowest p-state at all times. Also uses idle power savings on all CPUs.
Dynamic Power Savings	The operating system attempts to use lower power modes dynamically to save power while minimizing loss of performance.	Use idle power savings.
OS Control	The power savings mode is controlled by OS-specific mechanisms.	Enable the $POWER_ CONTROL system service and the CPU_POWER_MGMT SYSGEN parameter.

OpenVMS and Power Management

- A setting at the firmware level takes precedence over any setting at the OS level
- Only if “OS Control” is selected at the firmware level, then and only then, do the CPU_POWER_* SYSGEN parameters and the SYS$POWER_CONTROL system service take effect
- A setting of “Dynamic Power Savings” at the firmware level does allow the processors to operate in both high and low power modes, but the algorithm used to control this behavior (high power mode/low power mode) has nothing to do with the CPU_POWER_* SYSGEN parameter settings or the SYS$POWER_CONTROL system service
- For guaranteed maximum performance all of the time, select “Static High Performance” at the firmware level
- CPU_POWER_MGMT is used primarily to change the power state. The default for CPU_POWER_MGMT is 2. Its range of values and their meanings are:

CPU_POWER_MGMT Value	Mode Chosen
0	OpenVMS Static High Performance
1	OpenVMS Static Low Power
2	OpenVMS Dynamic Power Savings

- CPU_POWER_THRSH is used only in OpenVMS Dynamic Power Savings mode. It specifies the number of interrupts per 10-millisecond interval beyond which idle power savings will be turned off. The default value is 50. The higher this number, the more power is saved and the higher average interrupt latency the system will experience while processors are idle.

Problems with CSID ending in hex “FF”

– Classic “off-by-one” error. Day-1 bug in the cluster server process code (CSP.EXE) failed to initialize the last entry in an array of size 256.

– Symptoms: SSRVEXCEPT bugchecks with the CLUSTER_SERVER process current when another cluster node has a CSID ending in hex “FF”; for example 000100FF or 000200FF.

– Fix: No “official” patch kit yet, but a side-build CSP.EXE image with the fix is available through HPE Technology Services

– Workaround: Reboot the cluster node with a CSID ending in hex “FF”

Relationship between RECNXINTERVAL and QDSKINTERVAL

- When clusters were first introduced, default values were:
  - RECNXINTERVAL = 20
  - QDSKINTERVAL = 10
- Years ago, presumably with the intent of reducing the amount of time quorum is lost following the unexpected exit of a node from the cluster, the default value for QDSKINTERVAL was lowered to 3 seconds. The default value of RECNXINTERVAL was left at 20 seconds.
- Since then, we've observed multiple cases where a node may lose network connectivity with the rest of the world, including the rest of the cluster, but before the set of healthy nodes can remove it (at the expiration of the RECNXINTERVAL timer), it grabs ownership of the quorum disk (and acquires its QDSKVOTES number of votes) after two quorum-disk scans (2 x QDSKINTERVAL seconds) and achieves quorum, and all the rest of the nodes go down with CLUEXIT bugchecks when they get the bad news via the quorum disk. As a result, it may be only the node with failed network hardware which stays up.
- Conclusion: Let the cluster software decide the fate of a node before any node is allowed to grab ownership of the quorum disk. To implement this:
  - Ensure that QDSKINTERVAL is always at least half of RECNXINTERVAL, or lower QDSKVOTES such that a node can't achieve quorum with only itself and the votes supplied by the quorum disk.

Random File Header looks like a Home Block, preventing Mount

- The MOUNT utility scans likely areas of the disk for home blocks
- Because the checksum for a file header is in the same location as the checksum for a home block, in very rare circumstances a file header could look like an ODS-1 Home Block
- Symptoms: Mount fails with the error “'%MOUNT-F-FILESTRUCT” and the disk cannot be mounted, even with the /FOREIGN qualifier
- Fix is available in:
  - HPE 8.4 Integrity: VMS84I_MOUNT96-V0300
  - HPE 8.4 Alpha: VMS84A_MOUNT96-V0300

Problem using storage volume exactly 2 TB in size

- Customer tried to create an EVA volume of exactly 2 TB in size, specifying 2048 GB as size on the EVA
  - Received error on $INITIALIZE of “%INIT-F-BADPARAM, bad parameter value”
- Largest size volume OpenVMS can actually take advantage of is 2032 GB
- Sizes from 2033 GB to 2047 GB work, but OpenVMS can’t take advantage of the additional space

P400 RAID Controller problems after update

– Customer reported problems deleting a spare from a RAID unit on a P400 RAID Controller after installing the latest FIBRE_SCSI patch kit. Error was “Adding or modification of Raid Unit failed.”

– Patch kit introduced new MSA$UTIL.EXE program (linked 4-MAR-2016)

– HPE testing showed same command would work using old MSA$UTIL.EXE (linked 12-NOV-2012)

– OpenVMS Engineering has been able to duplicate the problem and is working on a fix

Summary: Recommended List of Patches for HPE OpenVMS 8.4

Integrity

Kit	Description
VMS84I_PCSI-V0400	PCSI Fixes
VMS84I_UPDATE-V1200	Consolidation Of Previously Released OpenVMS V8.4 Patch Kits, New Fixes, And New Functionality
VMS84I_SYS-V0700	OpenVMS Executive Fixes
VMS84I_SYSLOA-V0300	Fixes an issue in distributed lock manager code which could result in directory corruption or an occasional XQPERR bugcheck
VMS84I_CLIUTL-V0100	Fixes various issues with SET and SHOW commands after the installation of the SYS-V0700 kit
VMS84I_FIBRE_SCSI-V1100	Fibre-Channel and SCSI Driver Fixes
VMS84I_F11X-V0400	Fixes various RMS-E-PRV issues with DELETE, CREATE and PURGE commands after the installation of the SYS-V0700 kit
VMS84I_MOUNT96-V0300	Fixes a rare %MOUNT-F-FILESTRUCT' error condition which prevents the mounting of an otherwise perfectly valid ODS2/ODS5 volume

Summary: Recommended List of Patches for HPE OpenVMS 8.4

Alpha

Kit	Description
VMS84A_PCSI-V0400	PCSI Fixes
VMS84A_UPDATE-V1200	Consolidation Of Previously Released OpenVMS V8.4 Patch Kits, New Fixes, And New Functionality
VMS84A_SYS-V0700	OpenVMS Executive Fixes
VMS84A_SYSLOA-V0300	Fixes an issue in distributed lock manager code which could result in directory corruption or an occasional XQPERR bugcheck
VMS84A_CLIUTL-V0100	Fixes various issues with SET and SHOW commands after the installation of the SYS-V0700 kit
VMS84A_FIBRE_SCSI-V1000	Fibre-Channel and SCSI Driver Fixes
VMS84A_F11X-V0400	Fixes various RMS-E-PRV issues with DELETE, CREATE and PURGE commands after the installation of the SYS-V0700 kit
VMS84A_MOUNT96-V0300	Fixes a rare %MOUNT-F-FILESTRUCT' error condition which prevents the mounting of an otherwise perfectly valid ODS2/ODS5 volume

Questions?

HPE
POINTNEXT

Thank you

keith.parris@hpe.com and
rob.eulenstein@hpe.com

Hewlett Packard
Enterprise