* + 1. **Document VERSION 1.7**
       1. Define structures



* + - 1. Definitions:
         1. Completion Entry (CE)
         2. Completion Entry Size (CES)
      2. Define Q related structures
         1. struct cq\_track

U16 cq\_id

The ID of the CQ which is associated to an interrupt vector.

U8 isr\_fired

== 0; (false) isr did NOT fire.

== !0; (true) isr fired.

* + - * 1. struct irq\_track

U16 int\_vec;

What interrupt vector is this structure referring to.

Should allow MSI-X support since it will hold 2048 vector numbers

strcut cq\_track cq[]

linked list; keeps track of all the CQ’s which this int\_vec is responsible for processing.

* + - * 1. struct cmd\_track

U16 unique\_id

This is the driver assigned unique ID for a particular cmd. This allows processing a CQ to “lookup” a completed cmd in a list of possible cmds that were submitted to hardware to find out certain parameters about the cmd when it was sent. Like if the cmd has a PRP list or not.

U8 opcode

cmd opcode

enum {CMD\_ADMIN, CMD\_NVME, CMD\_AON} cmdSet

What cmd set does the opcode belong?

U16 persist\_q\_id

Filled in by IOCTL\_SEND\_64B\_CMD when a persistent queue is being created. It must be an admin cmd and a create CQ or SQ cmd for this to become populated correctly.

nvme\_prps prp\_nonpersist;

Points to the PRP list for this cmd if and only if there is a PRP list.

If a buffer of memory can be described within a single physical page of memory then a PRP list won’t be necessary.

* + - * 1. struct metrics\_sq

struct public\_sq

U16 sq\_id;

Even admin Q’s are supported here since q\_id = {0…}

U16 cq\_id

The CQ ID to which this SQ is associated

U16 tail\_ptr

Driver makes available the actual value residing within the appropriate controller space SQxTDBL register.

NOTE: Spec states reading SQxTDBL registers are undefined, thus the driver keeps track of the actual value in this variable

U16 tail\_ptr\_vir

Driver make available the value it will write to the appropriate SQxTBDL register as a result of copying new cmds to the SQ, but have not yet been told to “ring the doorbell”, i.e. uncommitted cmds residing within an SQ. Multiple cmds can be submitted to a SQ with the intent that the doorbell will be rung later. This allows submission of many commands at once. The driver must keep track of the cmds copied, but not yet committed via this parameter.

This value = tail\_ptr + <num\_cmds\_submitted\_B4\_ring\_doorbell>

if (tail\_ptr\_vir == tail\_ptr) then 0 uncommitted cmds exist within an SQ. A cmd is committed when it is placed in the SQ and the doorbell has rung to notify the hdw it is there.

U16 head\_ptr

Driver make available the last completion entry (CE) read from an appropriately linked CQ. Each CE contains a SQ head ptr value, this value must be tracked by the driver. The only way the driver can learn of the new SQ head ptr is if it is told to reap an element from a particular CQ. The user app will request items to be reaped and at this time only, while reaping only those which have been requested to be reaped will the driver learn of the head\_ptr since this value is provided within each completion element (CE) within every CQ.

U16 elements

Driver make available the total number of elements within this Q

This is a 1 based value

struct private\_sq

U8 \*vir\_kern\_add

physical address pointer to the Q’s allocated kernel memory, this is not pointing to a PRP list.

nvme\_prps prp\_persist;

Contains the possible persistent PRP lists for the SQ, this will hold PRP list for Admin Create IO SQ Commands only. Nothing else is really considered persistent. All other cmds can relinquish the PRP list after the command completes, but not the creation of CQ and SQ elements.

U8 contig

0 is false and indicates nvme\_prps is used, vir\_kern\_add is not used, and that the memory encompassing the SQ is not contiguous. !0 indicates true and means vir\_kern\_add is used, nvme\_prps is not used, and that the memory encompassing the SQ is contiguous.

U32 size

length in bytes of the allocated Q in kernel memory, may not be physically contiguous, but will be virtually contiguous.

U16 unique\_cmd\_id;

Global and unique to each SQ on a per device level. It is an incrementing counter to use for every cmd issued in this particular SQ. For the driver must over write CMD.DW3.CID during an IOCTL\_SEND\_64B\_CMD with this value, as it increments this number for each cmd being submitted into this SQ to guarantee uniqueness.

U32 \*dbs; points to the doorbell register in this particular SQ.

cmd\_track\_list

linked list of struct cmd\_track, possibly one for each cmd issue to hdw to this device. Not every cmd submitted to this Q will end up in this list, only certain requirements forces the logic to track a particular cmd. See description of IOCTL\_SEND\_64B\_CMD for details.

* + - * 1. struct metrics\_cq

struct public\_cq

U16 q\_id

Even admin Q’s are supported here since q\_id = {0…}

U16 tail\_ptr

Driver make available the value it thinks the respective tail\_ptr is. This value must be calculated. It is calculated by traversing the CQ from the point located by head\_ptr until the P-bit in the next higher Q location does not correlate to a new unreaped item. Thus if the head\_Ptr == 5, and the next 8 CQ locations have - bit set correctly, then the tail\_ptr = (head\_ptr + 8) = 13. This math must include Q wrapping anomalies.

See IOCTL\_REAP\_INQUIRY for details upon how and when this gets updated.

U16 head\_ptr

Driver make available the actual value residing within the appropriate controller space CQxTDBL register.

NOTE: Spec states reading CQxTDBL registers are undefined, thus the driver keeps track of the actual value in this variable

U16 elements

Driver passes back the total number of elements within this Q

This is a 1 based value

U8 irq\_enable

Driver passes back whether IRQ’s are used to learn of elements from this Q, if not, then it must be polled. This does not indicate how items are actually reaped, rather only how items are learned they need to be reaped. All items are reaped via IOCTL\_REAP when the user app requests items to be returned.

If disabled == 0; polling does not mean there is a thread waiting upon this CQ, rather the CQ is checked for new items during an IOCTL\_REAP\_INQUIRY. Updates to the tail\_ptr occur at this point in time.

If enabled == !0; an ISR must fire in order to allow IOCTL\_REAP\_INQUIRY to return the number of elements waiting within the spec’d CQ. If an ISR does not fire then the hardware did not notify the kernel of new elements, and regardless of whether or not there are actually elements within the CQ IOCTL\_REAP\_INQUIRY will return 0. The ISR must fire, then and only then, are we allowed to “see” elements within a CQ. Updates to the tail\_ptr occur at the time of IOCTL\_REAP\_INQUIRY.

U16 int\_vec

This is the interrupt for this which this CQ was registered when created. This is only has meaning if and only if irq\_enable == !0 (enabled).

struct private\_cq

U8 \*vir\_kern\_add

physical address pointer to the Q’s allocated kernel memory, this is not pointing to a PRP list.

nvme\_prps prp\_persist;

Contains the possible persistent PRP lists for the CQ, this will hold PRP list for Admin Create IO CQ Commands only. Nothing else is really considered persistent. All other cmds can relinquish the PRP list after the command completes, but not the creation of CQ and SQ elements.

U8 contig

0 is false and indicates nvme\_prps is used, vir\_kern\_add is not used, and that the memory encompassing the CQ is not contiguous. !0 indicates true and means vir\_kern\_add is used, nvme\_prps is not used, and that the memory encompassing the CQ is contiguous.

U32 size

length in bytes of the allocated Q in kernel memory, may not be physically contiguous, but will be virtually contiguous.

U32 \*dbs; points to the doorbell register in this particular CQ.

* + - 1. struct nvme\_prps
         1. U32 npages

No. of pages inside the PRP List only. Pages of PRP1 and PRP2 are not counted in here.

* + - * 1. U32 type

Describes the combination of PRP’s. Following combinations are possible:

|  |  |  |
| --- | --- | --- |
| PRP1 | PRP2 | type |
| NULL | NULL | NO\_PRP |
| Points to a page | NULL | PRP1 |
| Points to a page | Points to a page | PRP1|PRP2 |
| Points to a list of PRP’s | NULL | PRP1|PRP\_List |
| Points to a page | Points to a list of PRP’s | PRP2|PRP\_List |

Note: PRP1, PRP2 and PRP\_List are enums having values 1,2 and 4 respectively.

* + - * 1. \_\_LE64 \*\*vir\_prp\_list

Pointer to a list of virtual kernel pointers to the PRP pages inside PRP List

* + - * 1. \_\_LE64 prp1

Physical address in PRP1 of command

* + - * 1. \_\_LE64 prp2

Physical address in PRP2 of command

* + - * 1. dma\_addr\_t first\_dma

Starting physical address of a PRPList. It can be either present in PRP1 or PRP2 field of the command.

* + - 1. struct metrics\_driver
         1. struct public\_driver

enum {INT\_PIN, INT\_MSI\_SINGLE, INT\_MSI\_MULTI, INT\_MSIX} irq;

Driver detects the active interrupt request scheme during driver load time and stores this away under driver only metrics area, this can be change later but will always represent the active IRQ scheme in place. All schemes are mutually exclusive.

U16 driver\_version;

Updated when the source code changes, do not change if only the API/IOCTL changes

U16 api\_version;

Resides in file nvme\_ioctl.h and describes the version of the API of the IOCTLs. It must be updated whenever the API changes, otherwise the user app could issue request which are not supported. This will not be updated when the code/logic changes, only when the IOCTL definition changes.

* + - * 1. struct private\_driver

TBD/unknown.

* + - 1. Define device related structures
         1. struct metrics\_device

metrics\_cq\_list

linked list of struct metrics\_cq

metrics\_sq\_list

linked list of struct metrics\_sq

MUTEX irq\_track\_mtx

Mutex must be taken any time anything in irq\_track\_list[] needs to be written, modified, removed, freed, deleted and even when read because logic may remove a node while someone is trying to remove it.

irq\_track\_list

linked list of isr\_track; Note this could also be an array of size 2048 since that is the maximum MSI-X interrupt vector there could be. But using a linked list or an array decision will be left up to the implementer.

* + - * 1. metrics\_device\_list

linked list of struct metrics\_device, one for each device controlled by this driver

* + - 1. Define nvme\_device structures
         1. struct pci\_dev \*pdev; // pointer to the PCI device. This should point to the correct device opened in the driver.
         2. struct nvme\_ctrl\_reg \*nvme\_ctrl\_space; //pointer to the nvme register space.
         3. U8 \*bar0mapped // points to the BAR0 of the PCI device.
         4. device \*dmadev // points to the device inside the pci structure.
         5. U8 device\_no; //Indicates the device no. of the nvme device currently in use.
    1. IOCTL\_GET\_DRIVER\_METRICS
       1. returns metrics\_driver.public\_driver
    2. IOCTL\_GET\_Q\_METRICS
       1. U16 q\_id
          1. Pass the Q ID for which you require the metrics to be returned. This works equally well for admin, (q\_id == 0) as well as any other Q.
       2. enum {METRICS\_CQ, METRICS\_SQ} type
          1. returns metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].public\_cq or metrics\_device\_list[this\_device].metrics\_sq\_list[q\_id].public\_sq
       3. u32 nBytes
          1. Number of bytes to copy to the user buffer. The size must be able to hold the metrics data requested.
       4. u8 \*buffer
          1. Buffer to copy the metrics data, which is either public\_cq or public\_sq based on the type of q.
       5. Notes:
          1. If nBytes < sizeof(data\_we\_are\_returning) then alert failure to copy.
    3. IOCTL\_DEVICE\_STATE
       1. enum {ST\_ENABLE, ST\_DISABLE, ST\_DISABLE\_COMPLETELY} new\_state
          1. ST\_DISABLE and ST\_DISABLE\_COMPLETELY; is actually a controller reset; ST\_DISABLE performs a reset but leaves the ASQ, ACQ memory but does reset the pointers into ASQ and ACQ. ST\_DISABLE\_COMPLETELY completely free’s all kernel memory and destroys all queues.

The ordering of the logic which follows is important.

set CC.EN = 0;

We must try to wait for the controller to go idle before starting cleaning up memory under its feet. Thus wait 1s. for CSTS.RDY == 0, if it doesn’t within this time, then fail this operation, and do not cleanup, exit early.

for (i=1; i < metrics\_device\_list[this\_device].metrics\_sq\_list[last\_element]; i++) // all IO SQ’s wiped out

for (j=0; j < metrics\_device\_list[this\_device].metrics\_sq\_list[i].cmd\_track[j]; j++)

free the prp list described by metrics\_device\_list[this\_device].metrics\_sq\_list[i].cmd\_track[j].prp\_non\_persist, do NOT free the memory which the list is pointing to.

free the cmd\_track node at metrics\_device\_list[this\_device].metrics\_sq\_list[i].cmd\_track[j]

if (metrics\_device\_list[this\_device].metrics\_sq\_list[i].contig == true) then

free the memory pointed to by metrics\_device\_list[this\_device].metrics\_sq\_list[i].vir\_kern\_addr

else

free the prp list described by metrics\_device\_list[this\_device].metrics\_sq\_list[i].cmd\_track[j].prp\_non\_persist, do NOT free the memory which the list is pointing to.

free the metrics\_sq node at metrics\_device\_list[this\_device].metrics\_sq\_list[i]

Wipe out all cmds within the ASQ

for (j=0; j < metrics\_device\_list[this\_device].metrics\_sq\_list[0].cmd\_track[j]; j++)

if the cmd being tracked by metrics\_device\_list[this\_device].metrics\_sq\_list[0].cmd\_track[j] is not (create CQ) or (create SQ) or (delete CQ) or (delete SQ) the do the following

free the prp list described by metrics\_device\_list[this\_device].metrics\_sq\_list[0].cmd\_track[j].prp\_non\_persist, do NOT free the memory which the list is pointing to.

free the cmd\_track node at metrics\_device\_list[this\_device].metrics\_sq\_list[0].cmd\_track[j]

if (new\_state == ST\_DISABLE\_COMPLETELY) then

if (metrics\_device\_list[this\_device].metrics\_sq\_list[0].contig == true) then

free the memory pointed to by metrics\_device\_list[this\_device].metrics\_sq\_list[0].vir\_kern\_addr

else

free the prp list described by metrics\_device\_list[this\_device].metrics\_sq\_list[0].cmd\_track[j].prp\_non\_persist, do NOT free the memory which the list is pointing to.

free the metrics\_sq node at metrics\_device\_list[this\_device].metrics\_sq\_list[0]

else // we are just ST\_DISABLE and will keep the actual ASQ memory, just reset pointers into that memory

Reset to default initial values all the contents of metrics\_device\_list[this\_device].metrics\_sq\_list[0].public\_sq

Do NOT touch the contents of metrics\_device\_list[this\_device].metrics\_sq\_list[0].private\_sq

for (i=1; i < metrics\_device\_list[this\_device].metrics\_cq\_list[last\_element]; i++) // all IO CQ’s wiped out

if (metrics\_device\_list[this\_device].metrics\_cq\_list[i].contig == true) then

free the memory pointed to by metrics\_device\_list[this\_device].metrics\_cq\_list[i].vir\_kern\_addr

else

free the prp list described by metrics\_device\_list[this\_device].metrics\_cq\_list[i].cmd\_track[j].prp\_non\_persist, do NOT free the memory which the list is pointing to.

free the metrics\_cq node at metrics\_device\_list[this\_device].metrics\_cq\_list[i]

Wipe out all cmds within the ACQ

if (new\_state == ST\_DISABLE\_COMPLETELY) then

if (metrics\_device\_list[this\_device].metrics\_cq\_list[0].contig == true) then

free the memory pointed to by metrics\_device\_list[this\_device].metrics\_cq\_list[0].vir\_kern\_addr

else

free the prp list described by metrics\_device\_list[this\_device].metrics\_cq\_list[0].cmd\_track[j].prp\_non\_persist, do NOT free the memory which the list is pointing to.

free the metrics\_cq node at metrics\_device\_list[this\_device].metrics\_cq\_list[0]

else // we are just ST\_DISABLE and will keep the actual ACQ memory, just reset pointers into that memory

Reset to default initial values all the contents of metrics\_device\_list[this\_device].metrics\_cq\_list[0].public\_cq

Do NOT touch the contents of metrics\_device\_list[this\_device].metrics\_cq\_list[0].private\_cq

* + - * 1. ST\_ENABLE

Sets CC.EN = 1

Waits CAP.TO for the CSTS.RDY == 1, if it doesn’t come ready within TO then driver fails this operation. This is not a logical bug, but rather a hdw spec violation.

NOTE: Before enabling the user app should have already sent requests to satisfy system setup of proper admin Q’s and other duties described by section 7.6.1 in rev 1.0b spec.

* + 1. IOCTL\_CREATE\_ADMIN\_Q
       1. enum {ADMIN\_SQ, ADMIN\_CQ} type
       2. U16 elements
          1. Allocate contiguous kernel memory for the Q, maybe must use kmalloc() for this.
          2. if (type==ADMIN\_SQ) then allocate contiguous kernel memory in (bytes = elements \* 64B)

if (elements > 4096) then fail this request

* + - * 1. if (type==ADMIN\_CQ) then allocate contiguous kernel memory in (bytes = elements \* 16B)

if (elements > 4096) then fail this request

* + - * 1. this value is specified as a 1-based value
        2. Create a node within at metrics\_device\_list[this\_device].metrics\_sq\_list[append] or at metrics\_device\_list[this\_device].metrics\_cq\_list[append] as appropriate to represent this new Q memory.

If a node already exists with ID=0, then cleanup and fail this request. The previous Q should have been destroyed before re-creating it. Basically we are not giving the ability to destroy any admin Q’s, rather you must do it by a controller reset, and then only if you specify ST\_DISABLE\_COMPLETELY.

* + - 1. NOTES:
         1. The driver must populate the appropriate ASQ or ACQ registers.
         2. The driver must update via read-modify-write sequence, the appropriate ACQS or ASQS field of AQA register
         3. Each Q ID must = 0
         4. The spec only allows admin Q creation or modification while the controller is disabled, See IOCTL\_DEVICE\_STATE, p.37 of spec rev/ 1.0b. This IOCTL does not check this requirement; rather it will be left up the user app to guarantee this is the situation.
    1. IOCTL\_PREPARE\_SQ\_CREATION
       1. U16 elements
          1. Total number of entries/elements that needs to be allocated in the kernel
          2. total memory allocated in ‘bytes’ = (elements \* 2^( CC.IOSQES))
          3. This is a 1-based value
       2. U16 sq\_id;
          1. The user specified unique SQ ID; the driver should fail this request if this ID is already in use.
          2. We could have CQ’s with the same ID and that will be allowed.
       3. U16 cq\_id
          1. Pass the existing CQ ID to which this SQ needs to be associated.
          2. Do NOT fail this request if the cq\_id is not in existence already. The user app may actually want to test this boundary case and so we must allow this illegal state to progress.
       4. U8 contig
          1. This value must be saved in metrics\_device\_list[this\_device].metrics\_sq\_list[append].private\_sq.contig
          2. 0 indicates false that the SQ will not be contiguous. In the non-contiguous buffers are allocated by user space applications via the usual malloc() or new operator. The user mode virtual buffer address will then be passed to the kernel when a Create SQ Cmd is issued via IOCTL\_SEND\_64B\_CMD and the dataBuf parameter is in fact the pointer to the non-contiguous SQ. In this case the kernel will not use metrics\_device\_list[this\_device].metrics\_sq\_list[append].private\_sq.vir\_kern\_addr but rather create a PRP list in metrics\_device\_list[this\_device].metrics\_sq\_list[append].private\_sq.nvme\_prps when the IOCTL\_SEND\_64B\_CMD is issued. In this case the CMD.DW11.PC must == 0, and bitmask MASK\_PRP1 must be set to indicate to the kernel that it is OK to populate the cmd with the PRP list which is created.

The use space allocates the actual memory for the SQ.

* + - * 1. !0 indicates true that the SQ will be contiguous. metrics\_device\_list[this\_device].metrics\_sq\_list[append].private\_sq.vir\_kern\_addr will be used when the user app issues a Create SQ Cmd. The IOCTL\_SEND\_64B\_CMD will peek into the cmd and if a Create SQ Cmd is being issued, it must assume a previous IOCTL\_PREPARE\_SQ\_CREATION was issue to actually create the memory which is being references within that cmd. This assumption is made only when the cmd being issued has its CMD.DW11.PC bit set to true indicating a contiguous buffer is referenced. Thus the metrics\_device\_list[this\_device].metrics\_sq\_list[append].private\_sq.vir\_kern\_addr will be the pointer to that memory.

The driver allocates the actual memory for the SQ now.

* + - 1. NOTE
         1. This is the mechanism to create IO SQ memory in the kernel, it is not used for creating ASQ objects.
         2. Duties

Create a node within metrics\_device\_list[this\_device].metrics\_sq\_list[append] to represent this new Q memory, fill out both private and public portions of this node completely.

If a node already exists, then cleanup and fail this request. The previous Q should have been destroyed before re-creating it.

* + 1. IOCTL\_PREPARE\_CQ\_CREATION
       1. U16 elements
          1. Total number of entries/elements that needs to be allocated in the kernel
          2. total memory allocated in ‘bytes’ = (elements \*2^( CC.IOCQES))
          3. This is a 1-based value
       2. U16 cq\_id;
          1. The user specified unique CQ ID; the driver should fail this request if this ID is already in use.
          2. We could have SQ’s with the same ID and that will be allowed.
       3. U8 contig
          1. This value must be saved in metrics\_device\_list[this\_device].metrics\_sq\_list[append].private\_sq.contig
          2. 0 indicates false that the CQ will not be contiguous. In the non-contiguous buffers are allocated by user space applications via the usual malloc() or new operator. The user mode virtual buffer address will then be passed to the kernel when a Create CQ Cmd is issued via IOCTL\_SEND\_64B\_CMD and the dataBuf parameter is in fact the pointer to the non-contiguous CQ. In this case the kernel will not use metrics\_device\_list[this\_device].metrics\_cq\_list[append].private\_cq.vir\_kern\_addr but rather create a PRP list in metrics\_device\_list[this\_device].metrics\_cq\_list[append].private\_cq.nvme\_prps when the IOCTL\_SEND\_64B\_CMD is issued. In this case the CMD.DW11.PC must == 0, and bitmask MASK\_PRP1 must be set to indicate to the kernel that it is OK to populate the cmd with the PRP list which is created.

The use space allocates the actual memory for the CQ.

* + - * 1. !0 indicates true that the CQ will be contiguous. metrics\_device\_list[this\_device].metrics\_cq\_list[append].private\_cq.vir\_kern\_addr will be used when the user app issues a Create CQ Cmd. The IOCTL\_SEND\_64B\_CMD will peek into the cmd and if a Create CQ Cmd is being issued, it must assume a previous IOCTL\_PREPARE\_CQ\_CREATION was issue to actually create the memory which is being references within that cmd. This assumption is made only when the cmd being issued has its CMD.DW11.PC bit set to true indicating a contiguous buffer is referenced. Thus the metrics\_device\_list[this\_device].metrics\_cq\_list[append].private\_cq.vir\_kern\_addr will be the pointer to that memory.

The driver allocates the actual memory for the CQ now.

* + - 1. NOTE
         1. This is the mechanism to create IO CQ memory in the kernel, it is not used for creating ACQ objects.
         2. Duties

Create a node within metrics\_device\_list[this\_device].metrics\_cq\_list[append] to represent this new Q memory, fill out both private and public portions of this node completely.

If a node already exists, then cleanup and fail this request. The previous Q should have been destroyed before re-creating it.

* + 1. IOCTL\_SEND\_64B\_CMD
       1. U16 q\_id
          1. the queue ID to which ‘cmd\_buf’ must be copied into the next available location, queue id = 0 indicates this is an ASQ, all other values indicate this is an IO SQ.
          2. the next available location is known and updated by the kernel and is the metrics\_device\_list[this\_device].metrics\_sq\_list[q\_id].public\_sq.tail\_ptr\_vir
       2. enum {CMD\_ADMIN, CMD\_NVME, CMD\_AON} cmd\_set;
          1. User app passes in what cmd set this cmd belongs
       3. U16 bit\_mask
          1. MASK\_PRP1=0x01: indicates that PRP1 in this cmd can be overwritten by the driver.

if PRP1 is not needed then leave PRP1 alone.

see “Algorithm to populate PRP1/PRP2”

* + - * 1. MASK\_PRP2=0x02: indicates that PRP2 in the cmd can be overwritten by the driver.

if PRP2 is not needed then leave PRP2 alone.

see “Algorithm to populate PRP1/PRP2”

* + - * 1. MASK\_MPTR=0x04: indicates that MPTR in the cmd can be overwritten by the driver

if MPTR is not needed then leave MPTR alone

* + - * 1. Algorithm to populate PRP1/PRP2

The 'data\_buf\_size’ indicates the total, exact number of bytes being xfer’d by this cmd and the MASK\_PRP1 and/or MASK\_PRP2 indicate whether or not we can use those fields to describe the ‘data\_buf’ to the hdw. If a ‘data\_buf’ is passed and no MASK\_PRPx bits are set then obviously return with an error.

if (data\_buf\_size == 0) then skip this algorithm and do not touch PRP1 nor PRP2 of the cmd.

if (data\_buf\_size > CC.MPS) then we may have to create a PRP list for this cmd because PRP1 isn’t enough to describe this entire buffer

if (data\_buf\_size > (2\* CC.MPS)) then we definitely will have to create a PRP list for this cmd because PRP1 and PRP2 aren’t enough to describe this entire buffer

The exact details of converting the ‘user\_buf’ to PRP list will be left up the implementer but here are some details which cannot be missed in that algorithm

The allocation of memory to contain the PRP list needs to be remembered because eventually it must be freed back to the system.

Create a new node in metrics\_device\_list[this\_device].metrics\_?q\_list[q\_id].private\_?q.nvme\_prps if we are creating an IO SQ or an IO CQ, but only if there is a PRP list describing that Q and that means it will be a non-contiguous Q. Contiguous Q’s are described by a single PRP element and created by a call into the kernel via IOCTL\_PREPARE\_?Q\_CREATION.

Determine whether to use vir\_kern\_addr or nvme\_prps within metrics\_device\_list[this\_device].metrics\_?q\_list[q\_id].private\_?q by investigating if ((cmd\_set==CMD\_ADMIN) && (CMD.opcode==0x01)) then we are creating IO SQ. if ((cmd\_set==CMD\_ADMIN) && (CMD.opcode==0x05)) then we are creating IO CQ. Persistent PRP lists must remain until a controller reset occurs or a Delete IO Q Cmd is issued. To understand whether something is persistent or not we peek into the cmd and seek the value of CMD.DW11.PC

Create a new node in metrics\_device\_list[this\_device].metrics\_sq\_list[q\_id].private\_sq.cmd\_track[append].prp\_nonpersist[append] for all other cmds. Every other cmd except a Create IO Q Cmd.

Obviously if the system cannot provide enough dynamic kernel memory to make this PRP list this should be considered a failure

* + - 1. U8 \*cmd\_buf
         1. Virtual address pointer to 64B cmd, buffer alloc’d in user space. These values will be copied into the appropriate SQ which is reference by ‘q\_id’ It should be copied to the next available location as indicated by metrics\_device\_list[this\_device].metrics\_sq\_list.public\_sq.tail\_ptr\_vir. Then this pointer should be incremented to point to the next available location taking into account wrapping.
         2. if (cmd\_buf == NULL) then fail this request
      2. U32 data\_buf\_size
         1. size in bytes of data\_buf

if (data\_buf\_size!=0) && (cmd\_buf == NULL) then fail this request

if (data\_buf\_size ==0) && (cmd\_buf!= NULL) then fail this request

if (data\_buf\_size!= 0) and MASK\_PRP1 is missing then fail this request

* + - 1. U8 \*data\_buf
         1. Virtual user address space pointer to the allocated data buffer to associate with the cmd. Refer to section
         2. The user app has limited control of how many PRP entries will be generated by requesting the user buffer via normal user space allocation functions. Discontinuities are chunked up in pieces described by CC.MPS.
         3. Discontiguous CQ/SQ’s are considered user data buffers and this will be referenced here.
      2. U16 meta\_buf\_size
         1. size in bytes of meta\_buf

if (meta\_buf\_size!=0) && (meta\_buf == NULL) then fail this request

if (meta\_buf\_size ==0) && (meta\_buf!= NULL) then fail this request

if (meta\_buf\_size!= 0) and MASK\_MPTR is missing then fail this request

if (meta\_buf\_size > CC.MPS) then fail this request.

* + - 1. U8 \*meta\_buf
         1. Virtual user address space pointer to the user allocated meta data buffer to associate with the cmd.
         2. if MASK\_MPTR is set then convert this address to a physical address and write it into CMD.MTPR field of this cmd.

otherwise do not write into CMD.MPTR

* + - 1. NOTES:
         1. Does NOT ring the doorbell
         2. CMD.DW3.CID must be overwritten by the driver.

The driver increments the variable in metrics metrics\_device\_list[this\_device].metrics\_sq\_list[q\_id].private\_sq.unique\_cmd\_id and then write that new value to CMD.DW3.CID for every cmd submitted to guarantee uniqueness for the entire device per each SQ.

* + - * 1. Creating IO Q’s

if ((cmdSet == CMD\_ADMIN) & (CMD.opcode == 0x01 or 0x05))

if ((CMD.DW11.PC==1)

Verify metrics\_device\_list[this\_device].metrics\_?q\_list[q\_id].private\_?q.contig states it is contiguous, else error out this IOCTL request.

Verify metrics\_device\_list[this\_device].metrics\_?q\_list[q\_id].private\_?q.nvme\_prps.prp\_list **==** NULL, else error out this IOCTL request.

if (metrics\_device\_list[this\_device].metrics\_?q\_list[q\_id].private\_?q.vir\_kern\_addr **==** NULL)

return error because the user never called the appropriate IOCTL\_ALLOCATE\_?Q to create the Q memory.

if (data\_buf != NULL)

return error because for the creation of an IO Q there will never be a buffer, rather the address stored at metrics\_device\_list[this\_device].metrics\_?q\_list[q\_id].private\_?q.vir\_kern\_addr is being used as the start of this Q’s memory.

The address metrics\_device\_list[this\_device].metrics\_?q\_list[append].private\_?q.vir\_kern\_addr must NOT be run through the algorithm to create a PRP list, rather it can be represented as a single PRP element since it is contiguous.

else //CMD.DW11.PC==0

Verify metrics\_device\_list[this\_device].metrics\_?q\_list[q\_id].private\_?q.contig states it is NOT contiguous, else error out this IOCTL request.

Verify metrics\_device\_list[this\_device].metrics\_?q\_list[q\_id].private\_?q.vir\_kern\_addr **==** NULL, else error out this IOCTL request.

if (metrics\_device\_list[this\_device].metrics\_?q\_list[append].private\_?q. nvme\_prps.prp\_list **!=** NULL)

return error because the user is specifying a non-contiguous buffer and one apparently already exists

if (data\_buf == NULL)

return error because data\_buf must point to the non-contiguous buffer for this cmd.

The data\_buf must be run through an algorithm to create a PRP list if the buffer is of sufficient size and thus requires one. If one is required then it must be placed within metrics\_device\_list[this\_device].metrics\_?q\_list[q\_id].private\_?q.nvme\_prps..

* + - * 1. Cmd tracking

Create a new node within metrics\_device\_list[this\_device].metrics\_sq\_list[i].private\_sq.cmd\_track\_list[append] only when…

if (cmdSet == CMD\_ADMIN) && (CMD.DW0.opcode == 0x00 or 0x01 or 0x04 or 0x05)) or…

The ‘data\_buf’ requires a PRP list.

How to create a new node

Create a new node within metrics\_device\_list[this\_device].metrics\_sq\_list[i].private\_sq.cmd\_track\_list[append]

Look into the ‘cmd\_buf’ and remove the opcode and copy that to metrics\_device\_list[this\_device].metrics\_sq\_list[i].private\_sq.cmd\_track\_list[append].opcode

copy ‘cmd\_set’ to metrics\_device\_list[this\_device].metrics\_sq\_list[i].private\_sq.cmd\_track\_list[append].cmd\_set

copy metrics\_device\_list[this\_device].metrics\_sq\_list[q\_id].private\_sq.unique\_cmd\_id to metrics\_device\_list[this\_device].metrics\_sq\_list[i].private\_sq.cmd\_track\_list[append].unique\_id

* + 1. IOCTL\_DUMP\_METRICS
       1. U16 flen;
          1. Length of the file name that is specified.
       2. U8 \*filename;
          1. Name of the file that need to be used for writing the metrics data.
       3. Notes:
          1. Dump all the metrics data structure in a structure way to the filename supplied.
    2. IOCTL\_RING\_SQ\_DOORBELL
       1. U16 q\_id;
          1. id of the SQ to ring its doorbell,
          2. seek for the SQ within metrics\_device\_list[this\_device].metrics\_sq\_list[]

Copy metrics\_device\_list[this\_device].metrics\_sq\_list[q\_id].public\_sq.tail\_ptr\_vir to:

metrics\_device\_list[this\_device].metrics\_sq\_list[q\_id].public\_sq.tail\_ptr

the appropriate SQxTDBL register

* + 1. IOCTL\_REAP\_INQUIRY
       1. U16 q\_id
          1. Pass the CQ ID for which queue needs to be reaped. q\_id=0 (ACQ) are supported, as well as all others.

if (q\_id == 0) then Completion Entry Size (CES) = 16B; it is OK to assume a constant

if (q\_id != 0) then Completion Entry Size (CES) = CC.IOCQES; this is variable so find its size

* + - 1. U16 num\_remaining

Driver returns the number of Completion Entry (CE’s) there are waiting to reap from the specified CQ. This number is explained in the NOTES below.

This is a 1-based number

* + - 1. NOTES:
         1. Only this IOCTL is allowed to modify metrics\_device\_list[this\_device].metrics\_sq\_list[q\_id].public\_sq.tail\_ptr. There literally should be no other software which modifies this value, except during driver load time.
         2. Algorithm

Traverse the CQ from metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].public\_cq.head\_ptr towards the tail of the queue and note the P-bit within the CE. Traverse until the P-bit indicates the element is NOT a new CE placed there by the hdw, this must therefore be the location of the metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].public\_cq.tail\_ptr, because the tail\_ptr indicates the next empty location to where an element will be placed by hdw.

Update the metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].public\_cq.tail\_ptr to this newly found CE

num\_remaining = metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].public\_cq.tail\_ptr - metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].public\_cq.head\_ptr

remember to take into account wrapping conditions for this calculation

* + 1. IOCTL\_REAP
       1. U16 q\_id
          1. Pass the CQ ID for which queue needs to be reaped. Q ID = 0 (ACQ) are supported, as well as all others.

if (q\_id == 0) then Completion Entry Size (CES) = 16B; it is OK to assume a constant

if (q\_id != 0) then Completion Entry Size (CES) = CC.IOCQES; this is variable so find its size

* + - 1. U16 element
         1. Pass the number of completion elements to reap. if (size < (element \* CES)) then fail this request, we must have enough room to return the data.
         2. If the number of ‘element’ requested can’t be satisfied, because there aren’t enough CE’s to be reaped, then fail this request.
         3. Allow partial reaping, where ‘element’ isn’t large enough to reap all CE’s in the CQ. In other words there are more CE’s to be reaped is allowed.
         4. num\_could\_reap = IOCTL\_REAP\_INQUIRY(q\_id);

We must effectively call IOCTL\_REAP\_INQUIRY internally to update the tail\_ptr and find out how many CE’s could be reaped at this time. Remember only IOCTL\_REAP\_INQUIRY is allowed to modify the tail\_ptr. This IOCTL\_REAP must use the values of tail\_ptr as set by IOCTL\_REAP\_INQUIRY’s internal calling scheme. Do this before reaping any elements.

* + - 1. U16 num\_remaining

Driver returns the number of Completion Entry (CE’s) there are waiting to reap from the specified CQ after reaping the specified ‘element’. Thus there are more waiting in the CQ and the caller may want to know this.

num\_remaining = (num\_could\_reap – element);

This is a 1-based number

* + - 1. U16 size
         1. Pass number of bytes of the ‘buffer’
         2. As long as there are enough ‘size’ bytes to copy all the number of ‘element’ entries this will be allowed, even if there is more buffer space that is needed.
      2. U8 \*buffer;
         1. Pass the user mode virtual buffer to which the completion elements will be copied starting from index = 0, byte = 0.
      3. NOTES:
         1. Only this IOCTL is allowed to modify metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].public\_cq.head\_ptr. There literally should be no other software which modifies this value, except during driver load time.
         2. Purpose is to copy as many CE’s as indicated by the number ‘element’ into ‘buffer’ for the CQ with ID==‘q\_id’.
         3. ALGORITHM: reaping on CQ with q\_id=0, i.e algorithm specific to ACQ

Internally generate a call to IOCTL\_REAP\_INQUIRY to find out num\_could\_reap value.

For each CE which must be reaped do the following:

Allow metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].public\_cq.head\_ptr to point to the next items in the list because we are reaping this element.

Using CE.sq\_id and CE.cmdId perform a lookup to find the cmd\_track node within metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[x]. Call this node metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked] for future reference.

if (we didn’t find any node during this lookup) then

Update metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].public\_sq.head\_ptr with value from CE.head\_ptr

Copy this CE into the next free element location of ‘data\_buf’, there may be more CE’s to copy thus an indexer will have to be maintained for the next possible copy forthcoming.

Continue to the next for loop CE element we may have to process.

else continue to the next step on the next line below.

if (metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked].cmdSet == CMD\_ADMIN) then

if (metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked].opcode == 0x05) //creat CQ

if (CE.status == SUCCESS(0)) then follow ALGO\_CREATE\_CQ\_SUCCESS described below.

else follow ALGO\_CREATE\_CQ\_FAILED described below.

else if (metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked].opcode == 0x01) //creat SQ

if (CE.status == SUCCESS(0)) then follow ALGO\_CREATE\_SQ\_SUCCESS described below.

else follow ALGO\_CREATE\_SQ\_FAILED described below.

else if (metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked].opcode == 0x04) //del CQ

if (CE.status == SUCCESS(0)) then follow ALGO\_DELETE\_CQ\_FAILED described below.

else follow ALGO\_DELETEE\_CQ\_SUCCESS described below.

else if (metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked].opcode == 0x00) // del SQ

if (CE.status == SUCCESS(0)) then follow ALGO\_CREATE\_SQ\_SUCCESS described below.

else follow ALGO\_CREATE\_SQ\_FAILED described below.

else the follow ALGO\_GEN\_PROCESSING described below.

else // this is not an admin cmd at all

Lookup CC.CSS and follow the appropriate algorithm based upon the cmd set chosen.

NVM cmd set

follow ALGO\_GEN\_PROCESSING described below.

AON cmd set

follow ALGO\_GEN\_PROCESSING described below.

The rest is TBD

Update the appropriate register CQxHDBL as indicated by ‘q\_id’ with the value residing within with metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].public\_sq.head\_ptr.

* + - * 1. ALGO\_CREATE\_CQ\_SUCCESS and ALGO\_DELETE\_CQ\_FAILED:

free cmd\_track node from the ASQ, i.e. the ASQ is metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[]. Do NOT free and child elements within this node like a prp list.

Update metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].public\_sq.head\_ptr with value from CE.head\_ptr

Copy this CE into the next free element location of ‘data\_buf’, there may be more CE’s to copy thus an indexer will have to be maintained for the next possible copy forthcoming.

Continue to the next for loop CE element we may have to process

* + - * 1. ALGO\_CREATE\_CQ\_FAILED and ALGO\_DELETE\_CQ\_SUCCESS:

Let XXXX = metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked].persist\_q\_id

Perform a lookup of within metrics\_device\_list[this\_device].metrics\_cq\_list[XXXX].private\_cq.contig to see if the Q’s memory is contiguous or not.

if (Q memory is contiguous) then

free the memory pointed to by metrics\_device\_list[this\_device].metrics\_cq\_list[XXXX].private\_cq.vir\_kern\_addr

else // memory is not contig

free the memory pointed to by metrics\_device\_list[this\_device].metrics\_cq\_list[XXXX].private\_cq.prp\_persist

free the prp list described by memory pointed to by metrics\_device\_list[this\_device].metrics\_cq\_list[XXXX].private\_cq.prp\_persist

free the cmd\_track node at metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked]

free the metrics\_cq node at metrics\_device\_list[this\_device].metrics\_cq\_list[XXXX].

Update metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].public\_sq.head\_ptr with value from CE.head\_ptr

Copy this CE into the next free element location of ‘data\_buf’, there may be more CE’s to copy thus an indexer will have to be maintained for the next possible copy forthcoming.

Continue to the next for loop CE element we may have to process

* + - * 1. ALGO\_CREATE\_SQ\_SUCCESS and ALGO\_DELETE\_SQ\_FAILED

free cmd\_track node from the ASQ, i.e. the ASQ is metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[]. Do NOT free and child elements within this node like a prp list.

Update metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].public\_sq.head\_ptr with value from CE.head\_ptr

Copy this CE into the next free element location of ‘data\_buf’, there may be more CE’s to copy thus an indexer will have to be maintained for the next possible copy forthcoming.

Continue to the next for loop CE element we may have to process

* + - * 1. ALGO\_CREATE\_SQ\_FAILED and ALGO\_DELETE\_SQ\_SUCCESS:

Let XXXX = metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked].persist\_q\_id

Perform a lookup of within metrics\_device\_list[this\_device].metrics\_sq\_list[XXXX].private\_sq.contig to see if the Q’s memory is contiguous or not.

if (Q memory is contiguous) then

free the memory pointed to by metrics\_device\_list[this\_device].metrics\_sq\_list[XXXX].private\_sq.vir\_kern\_addr

else // memory is not contig

free the memory pointed to by metrics\_device\_list[this\_device].metrics\_sq\_list[XXXX].private\_sq.prp\_persist

free the prp list described by memory pointed to by metrics\_device\_list[this\_device].metrics\_sq\_list[XXXX].private\_sq.prp\_persist

free the cmd\_track node at metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked]

free the metrics\_sq node at metrics\_device\_list[this\_device].metrics\_sq\_list[XXXX].

Update metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].public\_sq.head\_ptr with value from CE.head\_ptr

Copy this CE into the next free element location of ‘data\_buf’, there may be more CE’s to copy thus an indexer will have to be maintained for the next possible copy forthcoming.

Continue to the next for loop CE element we may have to process

* + - * 1. ALGO\_GEN\_PROCESSING

free the PRP list described by metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked].prp\_persist

free the cmd\_track node from metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[cmd\_we\_tracked]

Update metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].public\_sq.head\_ptr with value from CE.head\_ptr

Copy this CE into the next free element location of ‘data\_buf’, there may be more CE’s to copy thus an indexer will have to be maintained for the next possible copy forthcoming.

Continue to the next for loop CE element we may have to process

* + 1. ISR Design: ISR Routine
       1. There will be one global ISR routine connected to all possible interrupt vectors using request\_irq().
       2. Since admin Q’s only work with interrupts, there must be an ISR at all times.
       3. There will be top half and bottom half ISR processing. The top half is the global ISR routine and it should do the following:
          1. irq\_return\_t TopHalfIsr(int irq, void \*dev\_id, struct pt\_regs \*regs)

Do NOT take the metrics\_device\_list[this\_device].irq\_track\_mtx mutex. We never want the top\_half to have to gain access to the mutex, because gaining access to a mutex during an ISR is very bad, we will not take out this mutex during the top half.

Disable this ‘irq’ in the appropriate manner. Must consider all ways of disabling irq’s including INT\_PIN, INT\_MSI\_SINGLE, INT\_MSI\_MULTI, INT\_MSIX.

Schedule a work item in the system workqueue. Pass to this work item the values of ‘irq’ and ‘dev\_id’. This work item is the bottom half.

* + - * 1. BottomHalfIsr(int irq, void \*dev\_id)

Must grab the metrics\_device\_list[this\_device].irq\_track\_mtx mutex for the entire duration of this processing to guarantee coherency with the IOCTL\_REAP processing. Release the mutex upon exit of this routine.

Seek for the node within irq\_track\_list[x] where irq\_track\_list[x].int\_vec == irq.

Call this node irq\_track\_list[current\_irq] for future reference.

for each node in irq\_track\_list[current\_irq].cq\_track[x] do the following

irq\_track\_list[current\_irq].cq\_track[x].isr\_fired == TRUE

Re-enable the ‘irq’ which was disabled in TopHalfIsr().

* + - 1. Notes:
         1. All an ISR is doing for the test environment is validating that an ISR was generated by the hardware and this action simply notifies/allows the kernel driver to process cmds within the associated CQ. The actually processing/reaping of those completion elements (CE) will be done by IOCTL\_REAP in a polling fashion regardless.
         2. The IOCTL\_REAP\_INQUIRY will return that no elements are in a CQ, unless the associated ISR has fired. This is enough to prove interrupts are working and it also allows ISR’s to work in a user space vs. kernel space polling API. IN other words the general interface into the kernel for reaping CE’s is polling regardless if interrupts are enabled for a particular CQ or not.
         3. The IOCTL\_REAP\_INQUIRY updates the metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].public\_cq.tail\_ptr if and only if the associated ISR has fired.
         4. The IOCTL\_REAP won’t be told there is anything to reap and thus won’t reap unless IOCTL\_REAP\_INQUIRY states there are items to reap. The IOCTL\_REAP\_INQUIRY is called from within IOCTL\_REAP to learn if there are any CE’s to reap. Thus IOCTL\_REAP won’t be able to reap unless an ISR has fired. This proves interrupts are working even though the interrupt didn’t do any processing of the CE’s in a CQ.
    1. ISR Design: IOCTL\_SET\_IRQ
       1. u8 enum {INT\_PIN, INT\_MSI\_SINGLE, INT\_MSI\_MULTI, INT\_MSIX} new\_irq;
          1. What is the new interrupt scheme which is desired, regardless of what it is now when this IOCTL has finished this new scheme will be active.
       2. U16 num\_irqs;
          1. Pass the number of interrupt vectors we wish to reserve and request the OS to grant to us at this time.
          2. Algorithm to verify validity of this number

if (new\_irq == INT\_PN) then

if (num\_irqs > 1) return failure

else if (new\_irq == INT\_MSI\_SINGLE) then

if (num\_irqs > 1) return failure

else if (new\_irq == INT\_MSI\_MULTI) then

if (num\_irqs > MSICAP.MC.MME) return failure

if (num\_irqs > 32) return failure

Notes

These interrupt vectors will/must range from (0 - (num\_irqs-1))

else if (new\_irq == INT\_MSIX) then

if(num\_irqs > MSIXCAP.MXC.TS) return failure

if (num\_irqs > 2048) return failure

Notes

These interrupt vectors will/must range from (0 - (num\_irqs-1))

* + - 1. Notes:
         1. At driver load time, or sometime really early before any devices are actually opened

the driver must fill/find/discover the value of metrics\_driver.public\_driver.irq

The effectively call IOCTL\_SET\_IRQ with the value learned for metrics\_driver.public\_driver.irq to setup the system with some initial interrupt scheme. Remember admin queue’s must work with ISR’s, this will guarantee they are working before any device is opened.

* + - * 1. IOCTL algorithm is as follows:

Disable all ISR’s for all interrupt vectors until this driver for this device. Disabling ISR’s will be unique to each interrupt scheme, i.e. INT\_PIN, INT\_MSI\_SINGLE, INT\_MSI\_MULTI, INT\_MSIX.

Any access to metrics\_device\_list[this\_device].irq\_track\_list[] must grab the metrics\_device\_list[this\_device].irq\_track\_mtx mutex

Generate an internal call to effectively do the work of calling IOCTL\_DEVICE\_STATE with newState = ST\_DISABLE\_COMPLETELY. All hardware is stopped. All kernel memory is returned back to the system, basically this is as a soft reset.

Delete all nodes within metrics\_device\_list[current\_device].irq\_track\_list[] because we will be rebuilding a new one as follows.

if (new\_irq == INT\_PIN) then

Create 1 node within metrics\_device\_list[current\_device].irq\_track\_list[]

metrics\_device\_list[current\_device].irq\_track\_list[0].int\_vec = INTR.ILINE;

this is a PCI space register at address 0x3C.

This is how we probe the hardware, assigned by BIOS.

else if (new\_irq == INT\_MSI\_SINGLE) then

Create 1 node within metrics\_device\_list[current\_device].irq\_track\_list[]

call pci\_enable\_msi()

metrics\_device\_list[current\_device].irq\_track\_list[0].int\_vec = dev.irq\_number

this is a kernel structure and will be set by the kernel after calling pci\_enable\_msi()

else if (new\_irq == INT\_MSI\_MULTI) then

Create ‘num\_irqs’ nodes in metrics\_device\_list[current\_device].irq\_track\_list[]

call pci\_enable\_msi\_block()

if we don’t get at least what we ask for then return failure.

This call may allocate > we ask for because it has power of 2 requirements

metrics\_device\_list[current\_device].irq\_track\_list[0].int\_vec = dev\_irq\_number

metrics\_device\_list[current\_device].irq\_track\_list[1].int\_vec = dev\_irq\_number+1

metrics\_device\_list[current\_device].irq\_track\_list[2].int\_vec = dev\_irq\_number+2

do until num\_irqs is satisfied

else if (new\_irq == INT\_MSIX) then

Create ‘num\_irqs’ nodes in metrics\_device\_list[current\_device].irq\_track\_list[]

call pci\_enable\_msix()

if we don’t get at least what we ask for then return failure.

metrics\_device\_list[current\_device].irq\_track\_list[0].int\_vec = msix\_entry[0].vector;

metrics\_device\_list[current\_device].irq\_track\_list[1].int\_vec = msix\_entry[1].vector;

metrics\_device\_list[current\_device].irq\_track\_list[2].int\_vec = msix\_entry[2].vector;

do until num\_irqs is satisfied

Call request\_irq() for all elements of the newly created metrics\_device\_list[current\_device].irq\_track\_list[] and assign the top half callback defined earlier by TopHalfIsr(int irq, void \*dev\_id, struct pt\_regs \*regs)

The driver must fill/find/discover the value of metrics\_driver.public\_driver.irq because we just change the scheme.

* + 1. ISR Design : Additions to IOCTL\_CREATE\_ADMIN\_Q
       1. if (type == ADMIN\_Q) then
          1. Must grab the metrics\_device\_list[this\_device].irq\_track\_mtx mutex
          2. switch(metrics\_driver.public\_driver.irq)
          3. case INT\_MSI and case INT\_PIN:

Create a new cq\_track node and add it to metrics\_device\_list[current\_device].irq\_track\_list[0].cq[append]

metrics\_device\_list[current\_device].irq\_track\_list[0].cq[append].cq\_id == 0 // this is a constant on purpose

metrics\_device\_list[current\_device].irq\_track\_list[0].cq[append].isr\_fired == FALSE

* + - * 1. case INT\_MSI\_MULTI and case INT\_MSIX

Search for node where this is true: metrics\_device\_list[current\_device].irq\_track\_list[x].int\_vec == 0; call this device\_list[current\_device].irq\_track\_list[one\_we\_found] for future reference

Create a new cq\_track node and add it to metrics\_device\_list[current\_device].irq\_track\_list[one\_we\_found].cq[append]

metrics\_device\_list[current\_device].irq\_track\_list[one\_we\_found].cq[append].cq\_id = 0 // this is a constant on purpose

metrics\_device\_list[current\_device].irq\_track\_list[one\_we\_found].cq[append].isr\_fired = FALSE

* + - 1. Notes:
         1. interrupt vector 0 is associated with admin CQ’s as stated in the spec.
    1. ISR Design : Additions to IOCTL\_SEND\_64B\_CMD
       1. if ((cmd\_set == CMD\_ADMIN) && (CMD.DW0.opcode == 0x05)) then // Create CQ cmd
          1. if (CMD.DW11.IEN == true) then

Must grab the metrics\_device\_list[this\_device].irq\_track\_mtx mutex

switch(metrics\_driver.public\_driver.irq)

case INT\_MSI and case INT\_PIN:

Create a new cq\_track node and add it to metrics\_device\_list[current\_device].irq\_track\_list[0].cq[append]

metrics\_device\_list[current\_device].irq\_track\_list[0].cq[append].cq\_id == CMD.DW10.QID;

metrics\_device\_list[current\_device].irq\_track\_list[0].cq[append].isr\_fired == FALSE

case INT\_MSI\_MULTI and case INT\_MSIX

Search for node where this is true: metrics\_device\_list[current\_device].irq\_track\_list[x].int\_vec == CMD.DW11.IV; call this device\_list[current\_device].irq\_track\_list[one\_we\_found] for future reference

Create a new cq\_track node and add it to metrics\_device\_list[current\_device].irq\_track\_list[one\_we\_found].cq[append]

metrics\_device\_list[current\_device].irq\_track\_list[one\_we\_found].cq[append].cq\_id = CMD.DW10.QID;

metrics\_device\_list[current\_device].irq\_track\_list[one\_we\_found].cq[append].isr\_fired = FALSE

set metrics\_device\_list[current\_device].metrics\_cq\_list[CMD.DW10.QID].public\_cq.irq\_enabled = TRUE

this notifies that interrupts are now enabled for this CQ which is being created

set metrics\_device\_list[current\_device].metrics\_cq\_list[CMD.DW10.QID].public\_cq.int\_vec = CMD.DW11.IV

this notifies what interrupt vector is associated to this CQ.

* + 1. ISR Design: Additions to IOCTL\_REAP\_INQUIRY
       1. For each cmd we must process from CQ with id=q\_id do the following
          1. if (metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].public\_cq.irq\_enabled == true) then

Must grab the metrics\_device\_list[this\_device].irq\_track\_mtx mutex

Use metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].public\_cq.int\_vec to find the irq\_track within irq\_track\_list[]. Call this node irq\_track\_list[int\_we\_found] for future reference

seek for the cq\_track node within irq\_track.cq[] using ‘q\_id’. Call this node irq\_track\_list[int\_we\_found].cq[cq\_we\_found] for future reference.

if (irq\_track\_list[int\_we\_found].cq[cq\_we\_found].isr\_fired == true) then

We are allowed to perform the same algorithm we do for non-isr based reaping and return the number of elements within CQ with q\_id.

If we find there a no items to reap, then reset irq\_track\_list[int\_we\_found].cq[cq\_we\_found].isr\_fired = false; and return 0 to the caller.

else return 0

Nothing is in the request CQ even if there really are elements because the ISR has not fired as of yet. We may have to deal with aggregating ISR’s and thus we the host are not yet suppose to be notified of any complete elements. So we return 0.

* + 1. ISR Design: Additions to IOCTL\_REAP
       1. We must not have the bottom half processing making changes under our feet otherwise we will miss CE’s being placed into a CQ with this design. Thus do the following as the 1st thing which is done in this IOCTL\_REAP coding.
          1. if (metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].private\_cq.irq\_enabled == true) then

Must grab the metrics\_device\_list[this\_device].irq\_track\_mtx mutex for the entire duration of this IOCTL processing to guarantee coherency with the bottom half processing. Release the mutex upon exit of this IOCTL.

* + - * 1. **NOTE: WE must have recursive mutex otherwise this logic will not work. If we take the mutex here and then later when this logic internally generates a call to IOCTL\_REAP\_INQUIRY it will be grabbed again, thus recursive mutexs are needed in the current design. Verify this will be possible.**
      1. For each cmd we must process from CQ with id=q\_id do the following
         1. if (metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].private\_cq.irq\_enabled == true) then

Lookup the cmd in the appropriate metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[x]; and if a cmd is found, and if that cmd is a Create CQ cmd, and the CE.status indicates that it FAILED, then do the following:

Using metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].private\_cq.int\_vec find the irq\_track node within metrics\_device\_list[this\_device].irq\_track\_list[x] and call that node metrics\_device\_list[this\_device].irq\_track\_list[one\_we\_found] for future reference.

Remove the cq\_track node from metrics\_device\_list[this\_device].irq\_track\_list[one\_we\_found].cq[x] which matches ‘q\_id’; we did not create a CQ successfully, thus this CQ should no longer be associated with an interrupt vector.

if (metrics\_device\_list[this\_device].irq\_track\_list[one\_we\_found].cq[] list is completely empty then

free the node metrics\_device\_list[this\_device].irq\_track\_list[one\_we\_found]

else we are done with the extra interrupt processing for this cmd.

Lookup the cmd in the appropriate metrics\_device\_list[this\_device].metrics\_sq\_list[CE.sq\_id].private\_sq.cmd\_track\_list[x]; and if a cmd is found, and if that cmd is a Delete CQ cmd, and the CE.status indicates that it SUCCESS, then do the following:

Using metrics\_device\_list[this\_device].metrics\_cq\_list[q\_id].private\_cq.int\_vec find the irq\_track node within metrics\_device\_list[this\_device].irq\_track\_list[x] and call that node metrics\_device\_list[this\_device].irq\_track\_list[one\_we\_found] for future reference.

Remove the cq\_track node from metrics\_device\_list[this\_device].irq\_track\_list[one\_we\_found].cq[x] which matches ‘q\_id’; we did not create a CQ successfully, thus this CQ should no longer be associated with an interrupt vector.

if (metrics\_device\_list[this\_device].irq\_track\_list[one\_we\_found].cq[] list is completely empty then

free the node metrics\_device\_list[this\_device].irq\_track\_list[one\_we\_found]

else we are done with the extra interrupt processing for this cmd.