SDC

STORAGE DEVELOPER CONFERENCE

SNIA SANTA CLARA, 2017

Past and present of the Linux NVMe driver

Christoph Hellwig

A driver..

Definition of DRIVER

: one that drives: such as

a : COACHMAN

b : the operator of a motor vehicle

c : an implement (such as a hammer) for driving

d : a mechanical piece for imparting motion to another piece

e : one that provides impulse or motivation • a *driver* in this economy

f : a golf wood with a nearly straight face used in driving

g : an electronic circuit that supplies input to another electronic circuit; *also* : LOUDSPEAKER

h : a piece of computer software that controls input and output operations

(from https://www.merriam-webster.com/dictionary/driver)

Two drivers..

	virtio_scsi	lpfc
.c files	1	16
.h files	1	19
LOC	1305	101,369

- Drivers can have orders of magnitude difference sizes
  - Types of supported hardware
  - Functionality

Three (NVMe) drivers..

	Linux 4.4	Linux 4.12	OFED 1.5 (Win)
.c files	3	1	7
.h files	3	0	13
LOC	7501	2350	29689

Four drivers..

	Linux 4.4	Linux 4.12 (nvme)	Linux 4.12 (nvme-core)	OFED 1.5 (Win)
.c files	*3*	1	2	7
.h files	*3*	0	3	13
LOC	*7501*	2350	6525	29689

The humble beginning

author	Matthew Wilcox <matthew.r.wilcox@intel.com> 2011-01-20 12:50:14 -0500
committer	Matthew Wilcox <matthew.r.wilcox@intel.com> 2011-11-04 15:52:51 -0400
commit	b60503ba432b16fc84442a84e29a7aad2c0c363d (patch)
tree	43dca7cd57965ce1a2b7b6f94437f0364fbc0034
parent	0b934ccd707ff33a87f15a35a9916d1d8e85d30e (diff)
download	linux-b60503ba432b16fc84442a84e29a7aad2c0c363d.tar.gz

NVMe: New driver

```
This driver is for devices that follow the NVM Express standard

Signed-off-by: Matthew Wilcox <matthew.r.wilcox@intel.com>
```

Diffstat

-rw-r--r--	Documentation/ioctl/ioctl-number.txt	1
-rw-r--r--	drivers/block/Kconfig	11
-rw-r--r--	drivers/block/Makefile	1
-rw-r--r--	drivers/block/nvme.c	1043
-rw-r--r--	include/linux/nvme.h	343

5 files changed, 1399 insertions, 0 deletions

The humble beginning

More than 1 month before the release of NVMe 1.0

author	Matthew Wilcox <matthew.r.wilcox@intel.com> 2011-01-20 12:50:14 -0500
committer	Matthew Wilcox <matthew.r.wilcox@intel.com> 2011-11-04 15:52:51 -0400
commit	b60503ba432b16fc84442a84e29a7aad2c0c363d (patch)
tree	43dca7cd57965ce1a2b7b6f94437f0364fbc0034
parent	0b934ccd707ff33a87f15a35a9916d1d8e85d30e (diff)
download	linux-b60503ba432b16fc84442a84e29a7aad2c0c363d.tar.gz

NVMe: New driver

This driver is for devices that follow the NVM Express standard

Signed-off-by: Matthew Wilcox <matthew.r.wilcox@intel.com>

LOC

Diffstat

-rw-r--r--	Documentation/ioctl/ioctl-number.txt	1
-rw-r--r--	drivers/block/Kconfig	11
-rw-r--r--	drivers/block/Makefile	1
-rw-r--r--	drivers/block/nvme.c	1043
-rw-r--r--	include/linux/nvme.h	343

5 files changed, 1399 insertions, 0 deletions

Early days

First version (Jan 2011) was very limited:

- Single SQ/CQ only
- Small data transfers (PRP1 only)
- Read and Write I/O commands and a few admin commands

Improved version merged into Linux 3.3 (Jan 2012):

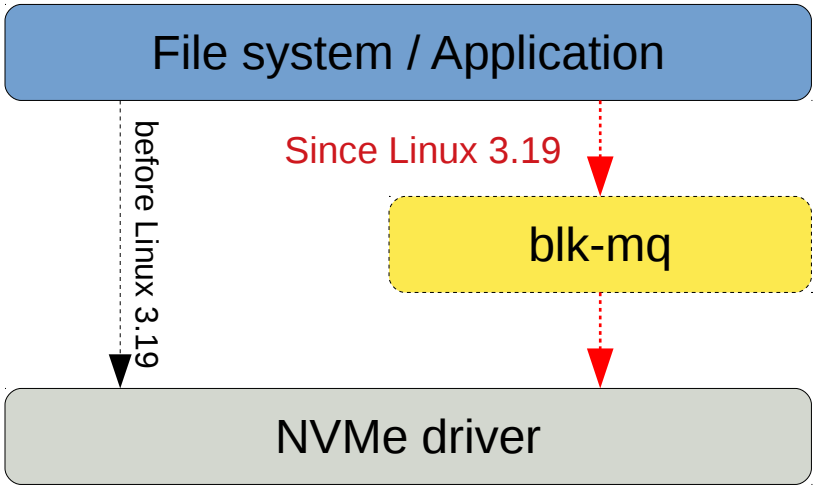
- Support for multiple queues
- Large data transfers using PRP chains
- Lots of fixes
- Drivers has grown by about 800 LOC

Junior years

Nothing too exciting until Nov 2015:

- Lots of bug fixes
- Support for deallocate (“discard”)
- Actually working flush support
- */dev/nvmeX* character devices
- Addition of a SCSI translation for ioctls

Using blk-mq in the NVMe driver



Linux 3.19 switch the NVMe driver to use blk-mq

- Allowed to remove hundreds of lines of code from the NVMe driver
- Very few modifications to the core blk-mq code were required
  - Most of that had been take care of for SCSI
- Building block for many future features

Blk-mq overview

What does blk-mq do?

- Split and merge I/O requests
- Manage multiple submission and completion queues
- Provide a command ID (tag) allocator
- Manage per-I/O data structures
- And much more

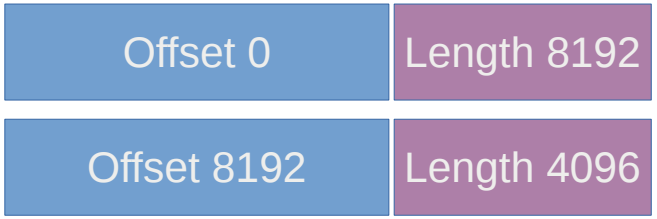
A bit of history:

- First prototyped in 2011
- Merged in Linux 3.13 (2014) for virtio
- Used by SCSI since 3.17 (2014)
- Used by NVMe since 3.19 (2015)
- And about a dozen other drivers now

PRPs and SGLs

PRPs describe each page:

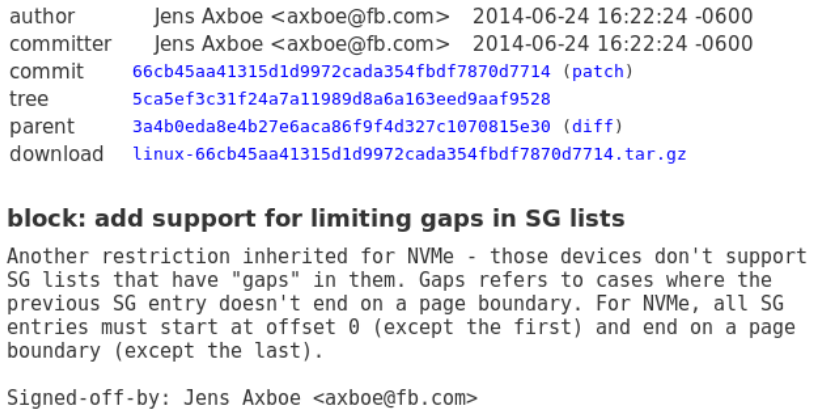
SGLs describe regions:



PRPs and Linux

The Linux I/O stack uses SGL-like structures as they are more flexible:

- They store large contiguous regions efficiently
- Allow arbitrary borders between segments
- But we could also support PRPs relatively easily
  - Also needed for RDMA, Hyper-V



author Jens Axboe <axboe@fb.com> 2014-06-24 16:22:24 -0600
committer Jens Axboe <axboe@fb.com> 2014-06-24 16:22:24 -0600
commit 66cb45aa41315d1d9972cada354fbdf7870d7714 (patch)
tree 5ca5ef3c31f24a7a11989d8a6a163eed9aaf9528
parent 3a4b0eda8e4b27e6aca86f9f4d327c1070815e30 (diff)
download linux-66cb45aa41315d1d9972cada354fbdf7870d7714.tar.gz

block: add support for limiting gaps in SG lists

Another restriction inherited for NVMe - those devices don't support SG lists that have "gaps" in them. Gaps refers to cases where the previous SG entry doesn't end on a page boundary. For NVMe, all SG entries must start at offset 0 (except the first) and end on a page boundary (except the last).

Signed-off-by: Jens Axboe <axboe@fb.com>

NVMe and SGLs

Since version 1.1 NVMe has optional SGL support

- Useful for large contiguous transfers, but use PRPs otherwise
- Can not be used for admin commands
- Except for NVMe over Fabrics, where only SGLs can be used

Linux support for SGLs is pending

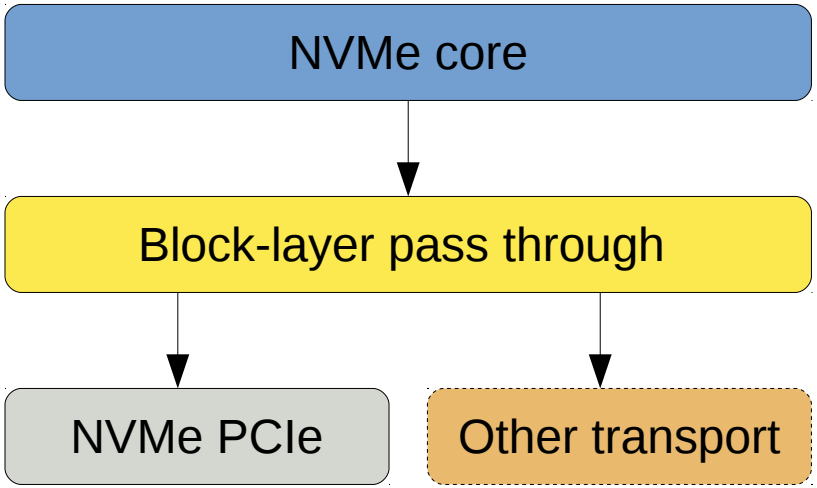
- Patches are out on the mailing list
- Need better detection of contiguous regions
- ~ 5% performance benefit for large transfers

Coming of age

Lots of feature work after the blk-mq switch:

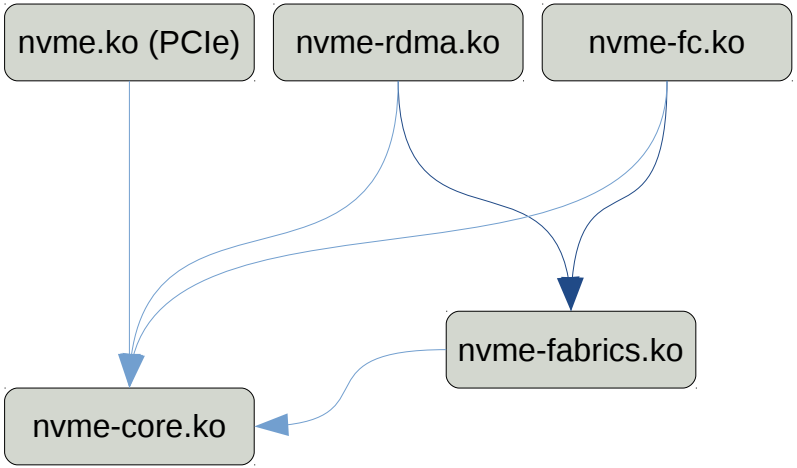
- T10 PI support (Feb 2015)
- CMB support, SQs only for now (Jul 2015)
- Persistent reservation support (Oct 2015)
- Support for weird Apple devices (Nov 2015)
- Basic SR-IOV support (Jun 2016)

From driver to subsystem, part 1



- Blk-mq allows passthrough requests that contain drivers specific raw commands
  - Initially used for SCSI CDBs
  - Generalized for NVMe commands
- Allowed us to split core vs PCIe to prepare for Fabrics
- Also supports multiple I/O command sets (e.g. LightNVM)

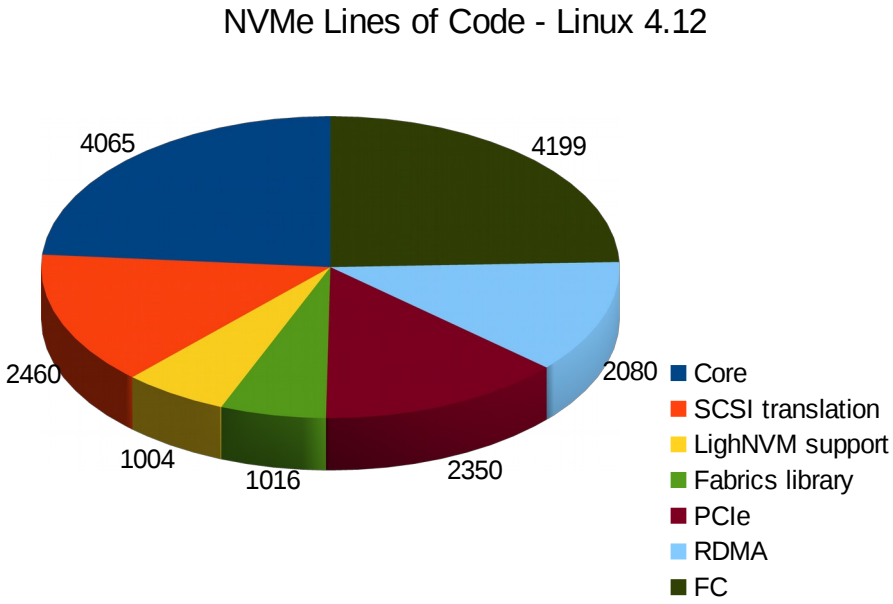
From driver to subsystem, part 2



Modularization for Fabrics:

- Move from ***drivers/block/*** to ***drivers/nvme/host/*** to prepare for a lot more NVMe related source files
- Split of the nvme-core module out of the existing nvme module also at the binary level
- Addition of new nvme-rdma and nvme-fabrics modules after the NVMe over Fabrics spec went public in June 2016

The NVMe (driver) subsystem



- 1 core NVMe module
  - Includes NVM I/O command set support
  - Including two optional features:
    - SCSI translation
    - LightNVM command set (Open Channel)
- NVMeOF library
- 3 transport drivers (PCIe, RDMA, FC)

Growing the family

- NVMe over Fabrics (RDMA) support (July 2016)
  - Including support for software defined NVMeOF controllers “target”
- Fibre Channel support (Dev 2016)

The chastity belt

author Scott Bauer <scott.bauer@intel.com> 2017-02-03 12:50:32 -0700
committer Jens Axboe <axboe@fb.com> 2017-02-06 09:44:21 -0700
commit a98e58e54fbd0c80b6a46a7cac6e231eed3b3efa (patch)
tree fa346839016a9667d47cf28d0744828d9db93006
parent 455a7b238cd6bc68c4a550cbbd37c1e22b64f71c (diff)
download linux-a98e58e54fbd0c80b6a46a7cac6e231eed3b3efa.tar.gz

nvme: Add Support for Opal: Unlock from S3 & Opal Allocation/Ioctls

```
This patch implements the necessary logic to unlock an Opal
enabled device coming back from an S3.

The patch also implements the SED/Opal allocation necessary to support
the opal ioctls.

Signed-off-by: Scott Bauer <scott.bauer@intel.com>
Signed-off-by: Jens Axboe <axboe@fb.com>
```

Diffstat

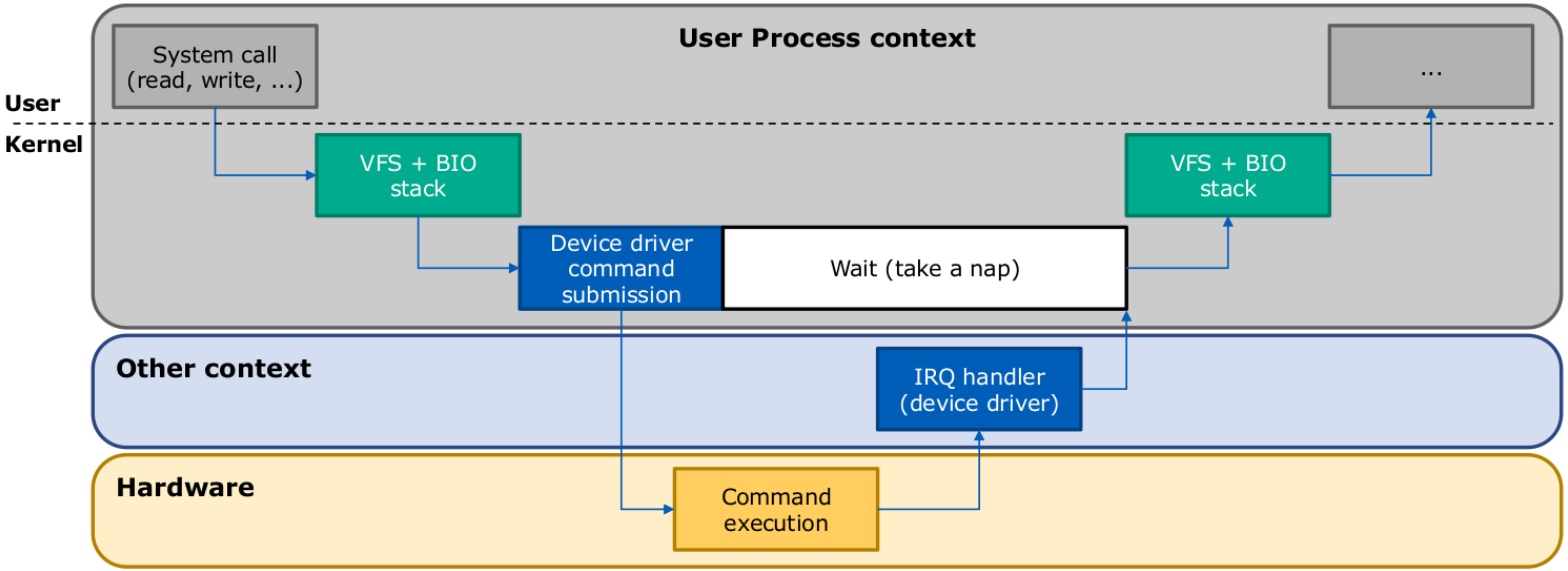
```
-rw-r--r-- drivers/nvme/host/core.c  25
-rw-r--r-- drivers/nvme/host/nvme.h 14
-rw-r--r-- drivers/nvme/host/pci.c   7
```

3 files changed, 46 insertions, 0 deletions

TCG Opal support in Linux 4.11:

- Disk encryption and access control
- Generic library
  - Less than 50 lines of code in NVMe
  - Now also supports for ATA

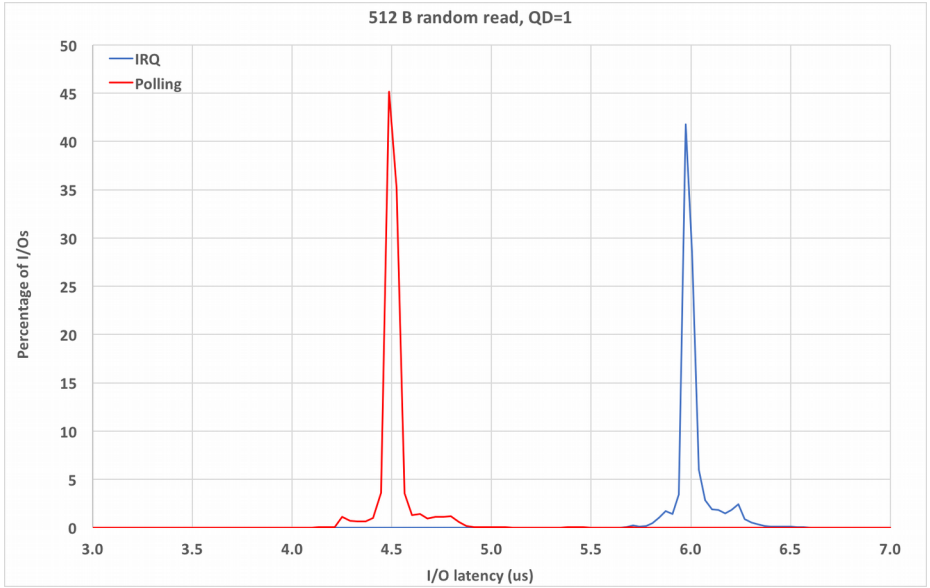
I/O flow – traditional IRQ path



Hitting it hard

Linux 4.4 introduced a polled I/O mode

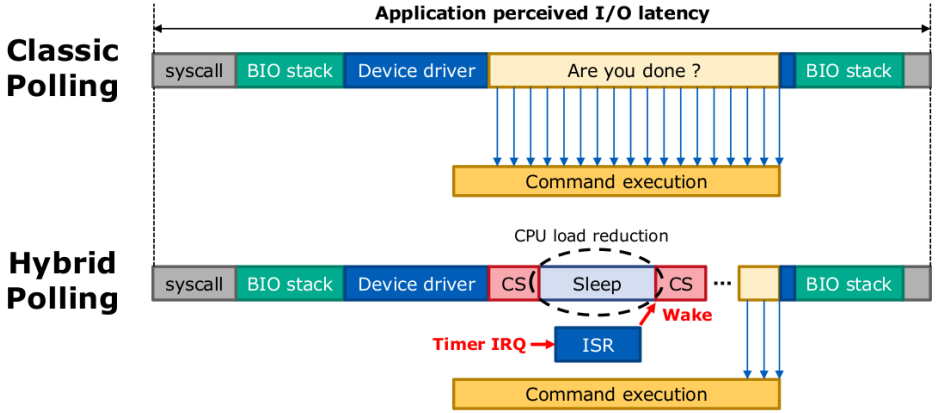
- Controlled by new RWF_HIPRI flag to the new preadv2/pwritev2 system calls
- Polling starts after I/O submission
- 100% CPU usage



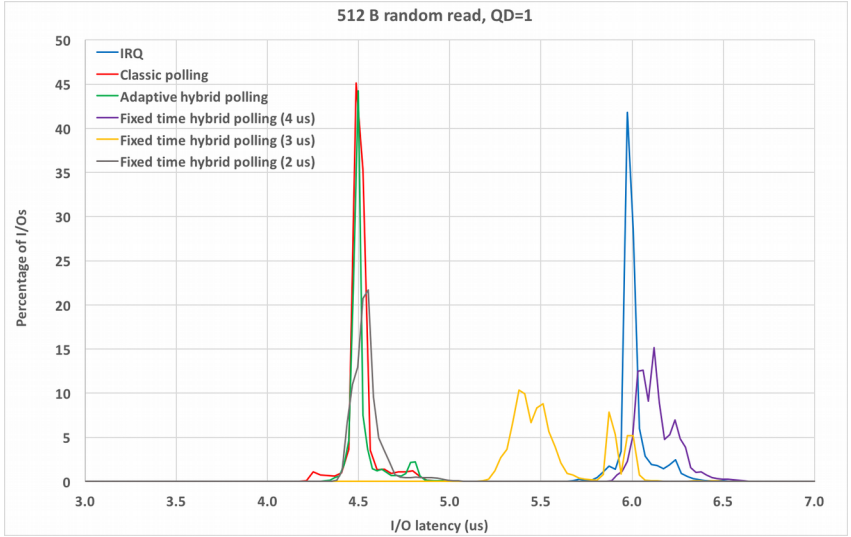
Going hybrid

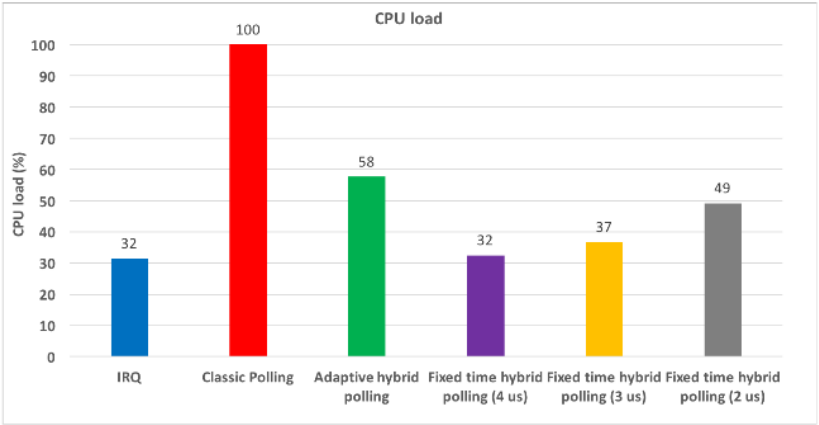
Linux 4.10 added hybrid polling

- Don’t start polling after submission – wait for half the average completion time
- Needs a good time estimate (especially for different I/O sizes)
- Still waists a lot of CPU – new patche to only start polling at the expected completion time



Polling latency





Coming of age

Catching up:

- Ranged deallocate support (Feb 2017)
- Autonomous Power State Transitions (Feb 2017)
- Host Memory Buffer support (May 2017)

And leading the pack with new NVMe 1.3 features:

- Set Doorbell Buffer, aka paravirtualized NVMe (Apr 2017)
- UUID identifiers (Jun 2017)
- Hot/Cold separation by (ab)using streams (Jun 2017)

Multipathing in NVMe

NVMe 1.1+ supports multiple controllers per subsystem

- Can be used to access shared storage from multiple systems
- Or to access data from the same system through different paths

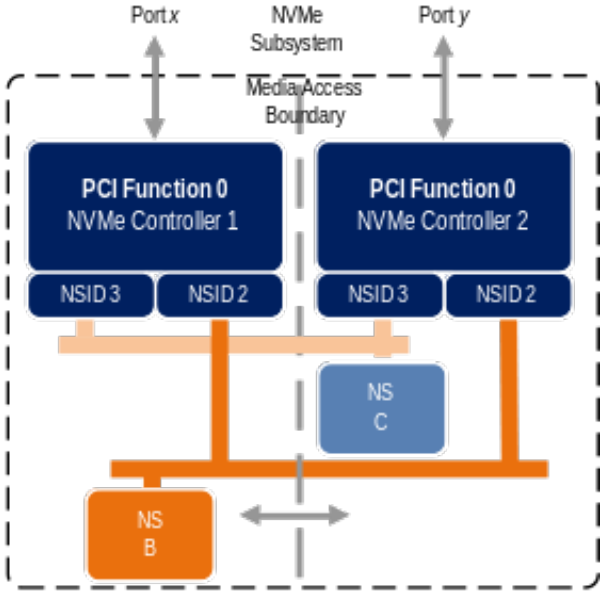
Use cases for multi-path access

- Aggregate bandwidth over multiple connections
- Redundancy
- Locality of access

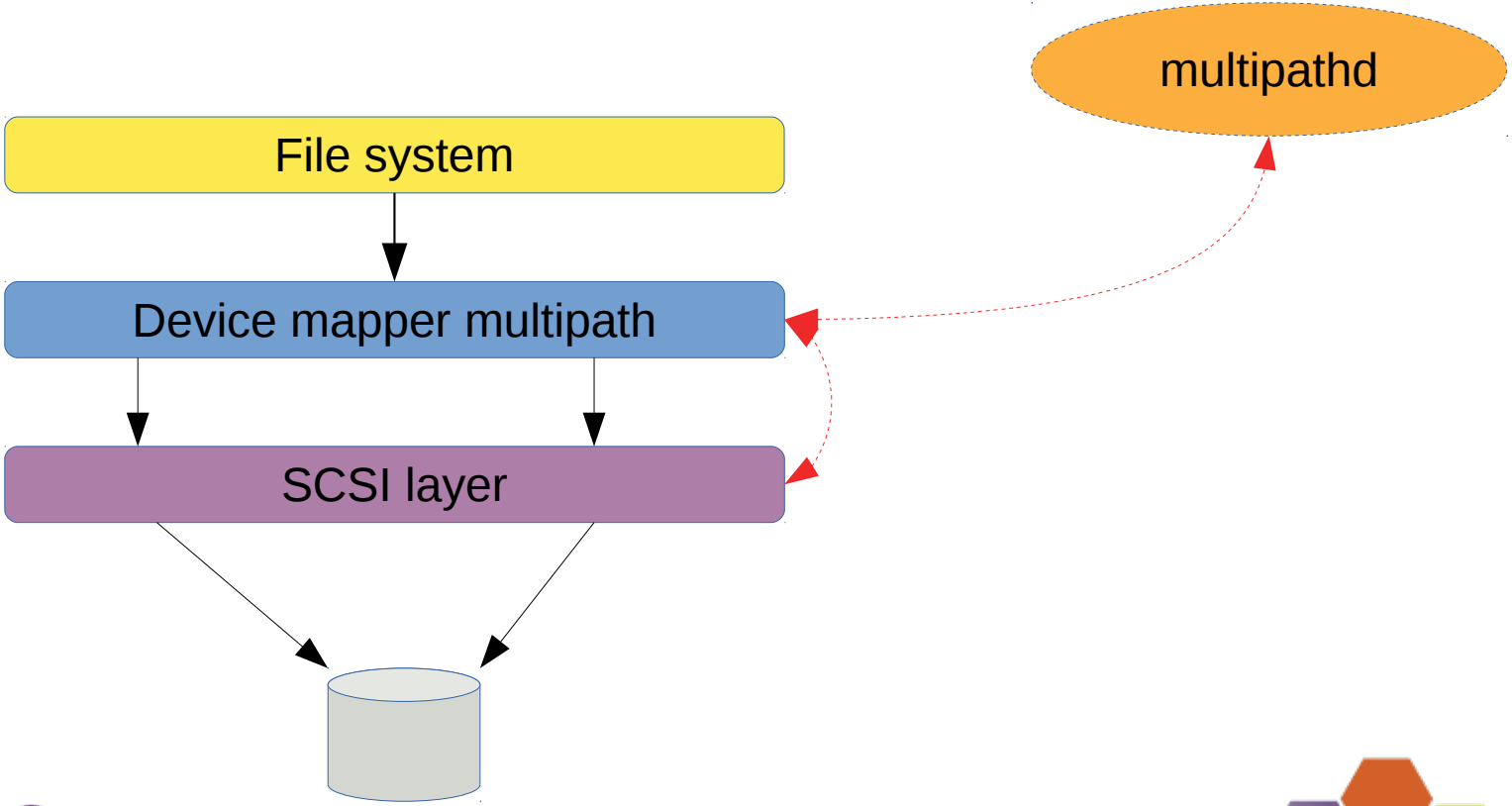
All roads lead to Rome..

Asynchronous Namespace Access

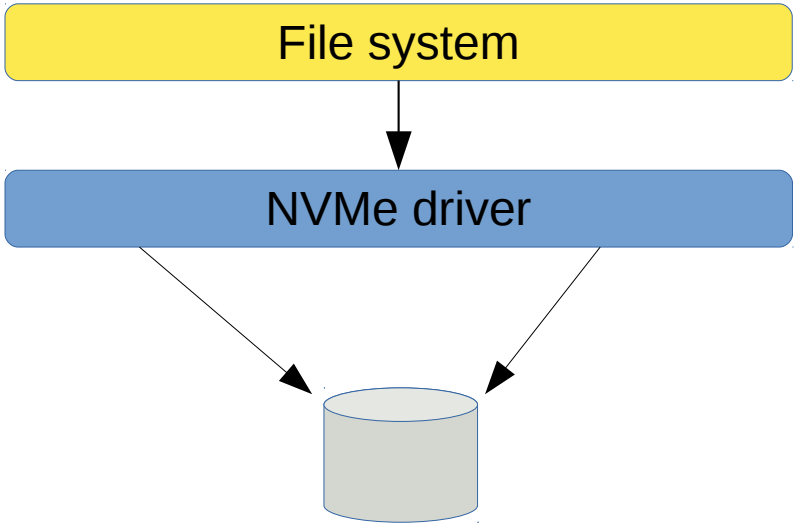
- Allows NVMe controllers to report access status per *(namespace, controller)* tuple
- Logical equivalent of ALUA in SCSI



(SCSI) Multipathing in Linux



Plan for NVMe Multipathing in Linux



Get the middle man out of loop

- NVMe already manages discovery of namespaces, and reporting ANA states
- Allows for automatic discovery and set up
- Allows for no added latency in NVMe vs additional 5-6 microseconds with device mapper
-

References

I/O Latency Optimization with Polling

http://events.linuxfoundation.org/sites/events/files/slides/lemoal-nvme-polling-vault-2017-final_0.pdf

Improving Block Discard Support throughout the Linux Storage Stack

http://vault2017.sched.com/event/9WQW/improving-block-discard-support-throughout-the-linux-storage-stack-christoph-hellwig

Increasing SCSI LLD Driver Performance by using the SCSI Multiqueue Approach

http://events.linuxfoundation.org/sites/events/files/slides/Vault%20-%20scsi-mq%20v2.pdf

Question?