OPENFABRICS ALLIANCE

10TH ANNUAL INTERNATIONAL DEVELOPER WORKSHOP

SRP and the scsi-mq Project

Bart Van Assche, FUSION-IO®

Overview

- Involvement with SRP.
- About storage API's.
- The Linux kernel, blk-mq and scsi-mq.
- SRP and the scsi-mq project.

Involvement with SRP

- Maintaining the open source Linux SRP initiator and the SCST SRP target drivers.
- Member of the Fusion-io ION team. ION is an all-flash H.A. shared storage appliance.
- Flash memory provides low latency and high bandwidth.
- The focus of RDMA is on low latency and high bandwidth.
- In other words, RDMA is well suited for remote access to flash memory.

About Storage API's (1/3)

- KVM = Kernel-based Virtual Machine, a hypervisor.
- KVM allows guests e.g. to access resources on the host system, e.g. block storage.
- KVM guests use paravirtualized drivers like virtio-blk and virtio-scsi.
- In 2007 the KVM virtio-blk driver was added to the Linux kernel [Ru08].
- virtio-blk provides a block device API to guests.
- Over time the KVM maintainers found themselves adding more and more SCSI features to the virtio-blk driver, e.g. disk identification and whether writeback is supported.
- In 2012 the virtio-scsi driver was merged in the Linux kernel.

About Storage API's (2/3)

Motivation for introducing the virtio-scsi driver:

> *The virtio-scsi HBA is the basis of an alternative storage stack for QEMU-based virtual machines (including KVM). Compared to virtio-blk it is more scalable, because it supports many LUNs on a single PCI slot), more powerful (it more easily supports pass-through of host devices to the guest) and more easily extensible (new SCSI features implemented by QEMU should not require updating the driver in the guest) [Bo12].*

About Storage API's (3/3)

- In other words ...
- A storage API must provide more functionality than only reading and writing blocks.
- There is a real need for the functionality present in the SCSI protocol.

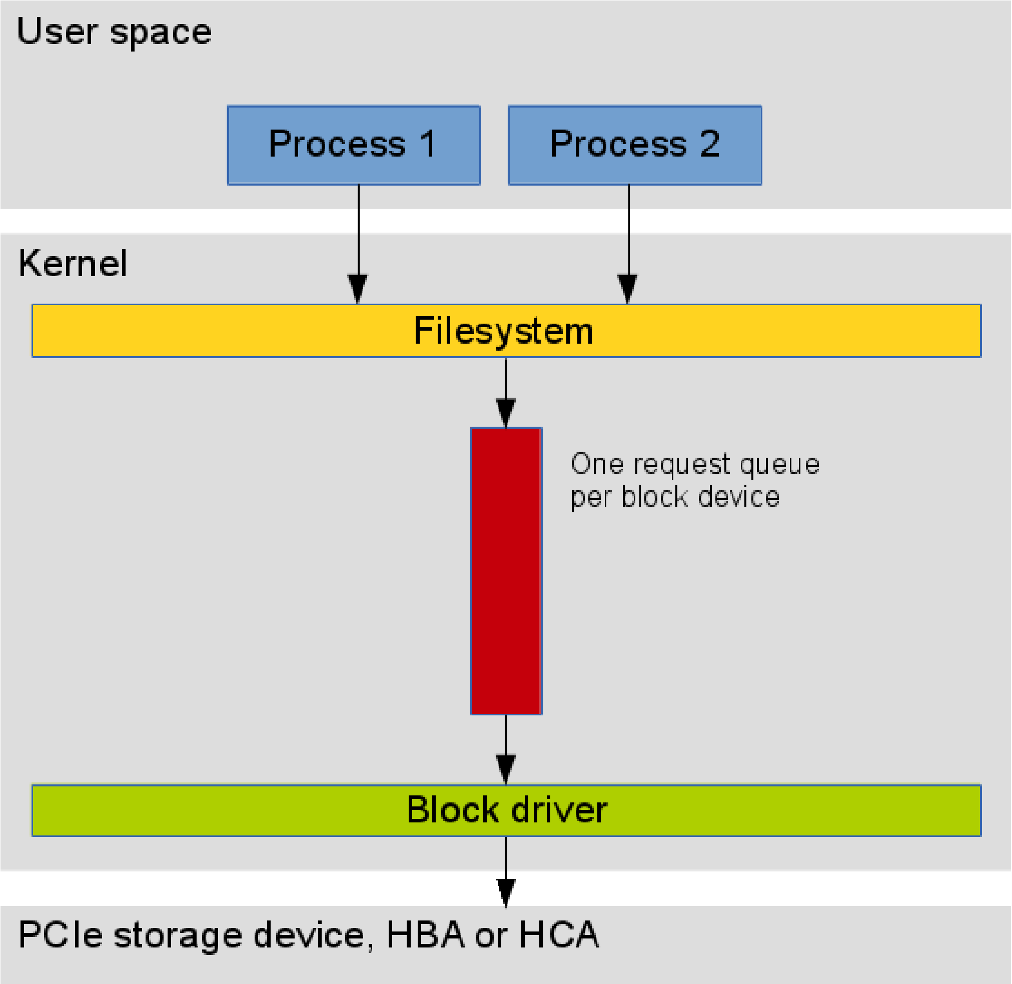
SRP and SCSI

- SRP defines a SCSI transport layer.
- Enables supports for e.g. these SCSI features:
  - Reading and writing data blocks.
  - Read capacity.
  - Command queueing.
  - Multiple LUNs per SCSI host.
  - Inquire LUN information, e.g. volume identification, caching information and thin provisioning support (a.k.a. TRIM / UNMAP).
  - Atomic (vectored) write - helps to make database software faster.
  - VAAI (WRITE SAME, UNMAP, ATS, XCOPY).
  - End-to-end data integrity (a.k.a. T10-PI).
  - Persistent reservations a.k.a. cluster support.
  - Asymmetric Logical Unit Access (ALUA).
- Fusion-io is actively involved in the ANSI T10 committee for standardization of new SCSI commands.

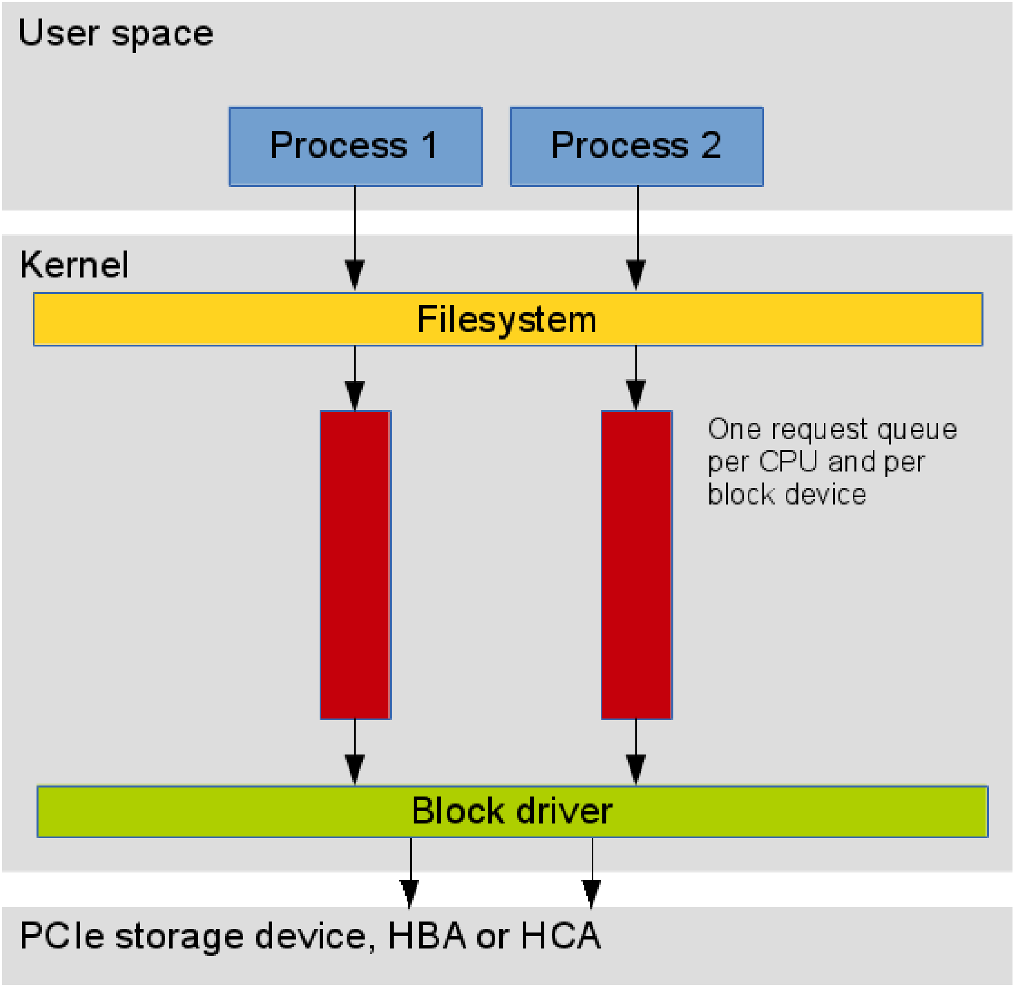
Linux Kernel and Storage Driver Performance

- Today some storage drivers are capable of more than one million IOPS: high-end SSDs and storage over fast networks.
- Some Linux kernel block drivers achieve up to 3 million IOPS.
- Linux SCSI kernel drivers achieve up to 1 million IOPS.
- Dilemma for high-end storage device driver developers: high performance and limited functionality (block driver) or limited performance and full functionality (SCSI driver) ?
- Traditional Linux block layer triggers lock contention on multicore systems.
- Multi-queue block layer (blk-mq) eliminates lock contention.
- Has been merged in Linux kernel version 3.13 [Bj13].
- Fusion-io has asked Christoph Hellwig to rewrite the Linux SCSI mid-layer as a multi-queue block driver (scsi-mq).

Traditional Linux Block Layer


User space
Process 1
Process 2
Kernel
Filesystem
One request queue per block device
Block driver
PCIe storage device, HBA or HCA

Multi-queue Block Layer


User space
Process 1
Process 2
Kernel
Filesystem
One request queue per CPU and per block device
Block driver
PCIe storage device, HBA or HCA

Advantages of the blk-mq approach

- One request queue per CPU eliminates lock contention.
- Certain SSD's and RDMA HCA's support multiple hardware queues and multiple MSI-X vectors.
- Using multiple hardware queues reduces contention and allows to spread interrupt load over multiple CPU cores.
- An example of multiple MSI-X vectors allocated for one IB port:

```
# sed -n 's/^\([^:]*:\).*\(mlx4-ib-1-.@PCI Bus 0000:21\)/\1 \2/p' /proc/interrupts
175: mlx4-ib-1-0@PCI Bus 0000:21
176: mlx4-ib-1-1@PCI Bus 0000:21
177: mlx4-ib-1-2@PCI Bus 0000:21
178: mlx4-ib-1-3@PCI Bus 0000:21
179: mlx4-ib-1-4@PCI Bus 0000:21
180: mlx4-ib-1-5@PCI Bus 0000:21
181: mlx4-ib-1-6@PCI Bus 0000:21
182: mlx4-ib-1-7@PCI Bus 0000:21
```

Current scsi-mq Status

- Traditional SCSI core is implemented as a block driver.
- scsi-mq = SCSI core based on the multiqueue block layer (blk-mq).
- One request queue per CPU and per LUN.
- Preliminary results for multi-queue support in the SRP initiator driver:
  - Very significant CPU usage reduction - up to 250%.
  - Higher IOPS when using multiple RDMA channels.
  - Higher bandwidth when using multiple RDMA channels.
- Latest scsi-mq patches have been posted on March 17 on the linux-scsi and linux-kernel mailing lists [Ch14].
- Open issues:
  - Implementing multiple hardware queues in a SCSI driver is possible but is not yet integrated with the blk-mq layer.
  - Hardware queues are per LUN instead of per SCSI host. This means "queue full" detection is done by the SCSI layer instead of the block layer.
  - There is one tag pool per hardware queue so the "one hardware queue" model is a contention point on NUMA systems.

References

- [Ru08] Rusty Russell, *virtio: towards a de-facto standard for virtual I/O devices*, ACM SIGOPS Operating Systems Review 42.5 (2008): 95-103.
- [Bo12] Paolo Bonzini, *virtio-scsi: SCSI driver for QEMU based virtual machines*, Linux kernel tree, February 2012.
- [Ta12] Nisha Talagala, *Under the Hood of the ioMemory SDK*, Fusion-io blog, April 2012.
- [Bj13] Matias Bjørling, et al., *Linux block IO: introducing multi-queue SSD access on multi-core systems*, Proceedings of the 6th International Systems and Storage Conference. ACM, 2013.
- [El14] Robert Elliott e.a., *SBC-4 SPC-5 Atomic writes and reads, ANSI T10 committee*, February 2014.
- [Ch14] Christoph Hellwig, *[WIP] scsi multiqueue*, Linux SCSI mailing list, March 17, 2014.

Thank You

OPENFABRICS ALLIANCE

10TH ANNUAL INTERNATIONAL DEVELOPER WORKSHOP