

Xen scsifront/back drivers

FUJITA Tomonori

tomof@acm.org

NTT Cyber Solutions Laboratories

Xen Summit 2006

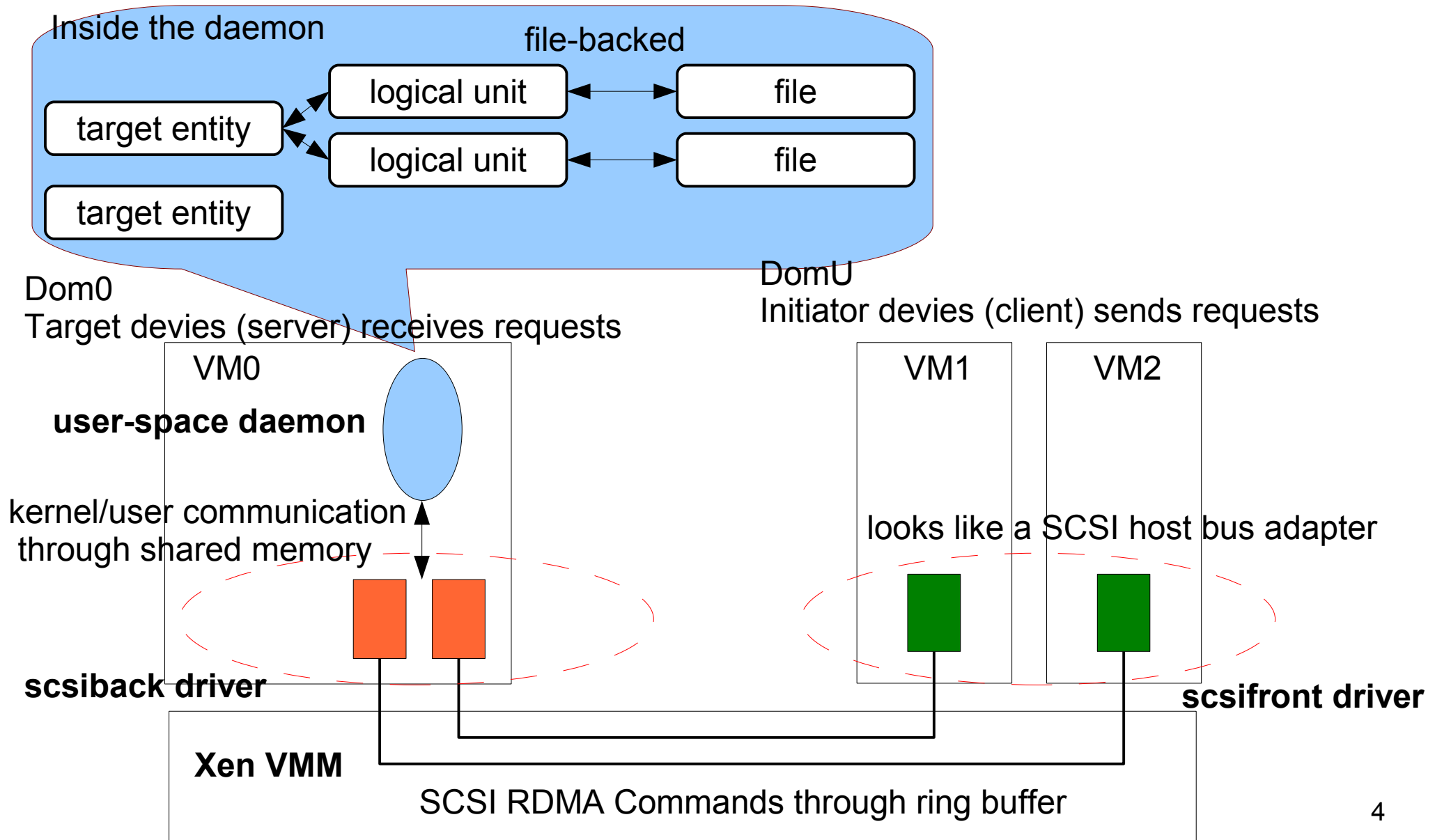
Current Block I/O: blkfront/back (or blktap)

- Beauty
 - Simple and fast (Xen original protocol)
- Issues
 - Extra effort for some software
 - Journaling file systems, udev, installer, etc
 - Possible improvement
 - Error handling
 - Dynamic device management
 - Other virtual storage drivers: tape, CD-ROM/DVD, etc

scsifront/back drivers

- SCSI-level device channel protocol
 - The frontend and backend drivers use SCSI RDMA Protocol (SRP) through the ring buffer
- SCSI protocol processing in user space
 - User-space daemon in dom0 does SCSI protocol processing and I/O executions (similar to blkmap)
- SCSI protocol saves lots of work
 - The existing software just works
 - Linux SCSI mid-layer provides error handling, dynamic management, various device support, etc

scsifront/back overview



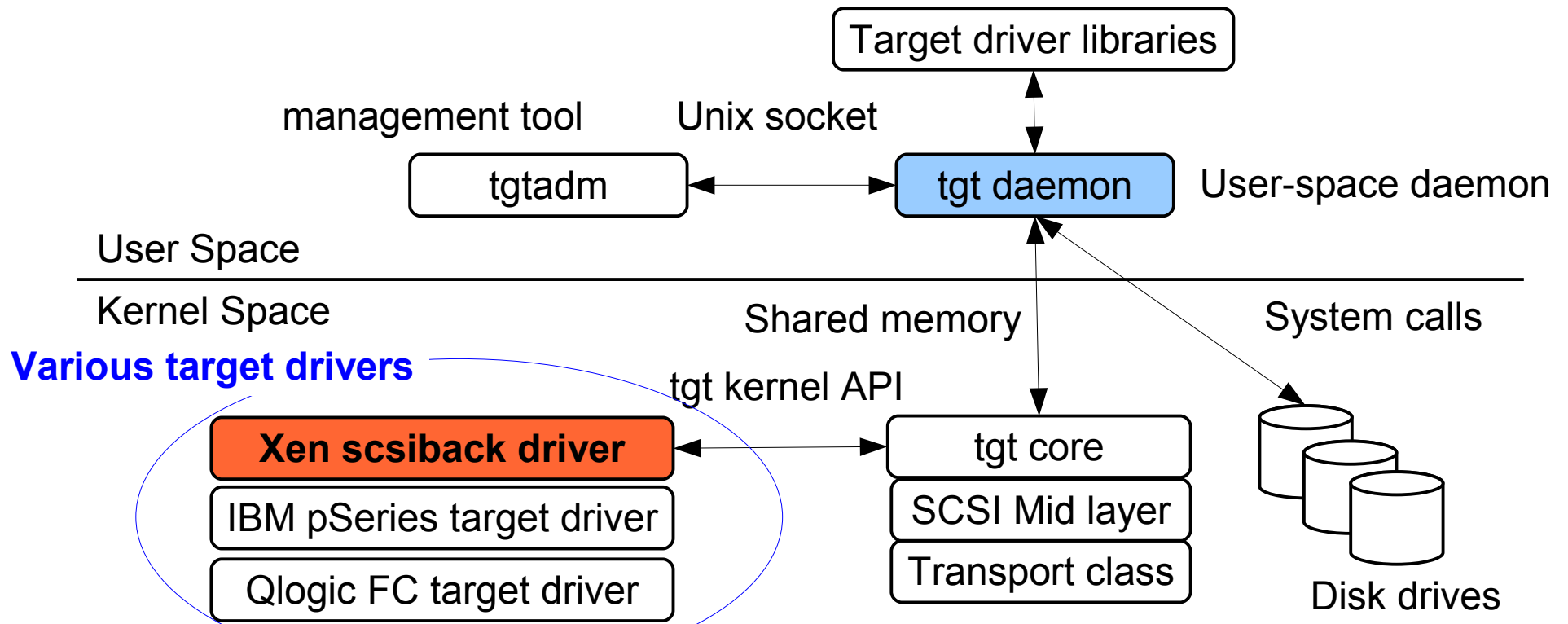
scsifront/back design

- I/O execution in user space (like blktp)
 - Exports whatever you want to VMs: raw partitions, regular files, LVM, files over network, etc
 - Easily nice tricks like metadata disk format
- Zero-copy of data pages (like blktp)
 - scsifront grants pages
 - scsiback maps the pages to user space
- Fits well in the existing Linux facilities
 - scsifront is a simple SCSI Virtual HBA driver
 - scsiback exploits SCSI target framework (tgt)

tgt: framework target drivers

tgt can simplify SCSI target drivers by providing SCSI protocol processing, target and logical unit management, etc

Xen scsiback only handles low level details



tgt support various SCSI transport protocol target drivers, fibre channel, SRP, iSCSI, etc ⁶

How scsiback works?

Write command case

- scsiback gets a SCSI command through ring buffer and maps the pages to tgt's address space
- scsiback passes the command to tgt core via tgt kernel APIs
- scsiback sends the mapped address information to tgt via scsiback and tgt
- tgt core sends them to tgt through shared buffer between tgt-core and tgt
- tgt performs the command, does AIO write, and sends the results to scsiback via tgt core
- scsiback sends the result to scsifront

tgt status

- tgt, ibmvstgt driver, and libsrp are in the -mm tree
 - ibmvstgt driver
 - SCSI target driver for IBM pSeries VM environments
 - Plays the same role as scsiback driver
 - VMs communicate each other by using RDMA
 - libsrp
 - Provides common features for SRP target drivers like Xen scsiback and ibmvstgt drivers

scsifront/scsiback status

- Previous version (submitted 2006/08/02)
 - DomU can fdisk, mkfs, read/write with virtual disk
- Current
 - Added AIO support to user-space daemon (tgt) for better performance
 - The majority of the target drivers doesn't need AIO
 - Modified libsrp for scsiback driver
 - It was designed for ibmstgt driver initially

I'll release the next version soon
(and performance results)

Next steps

- Add more support for scsiback to tgt
 - Use tgt's shared buffer to send the mapped page info and remove scsiback's own shared buffer
- SRP transport library
 - scsifront is the third SRP initiator driver in Linux
 - ibmvscsi (pSeries) and srp_ib (infiniband)
 - Needs SRP transport class (scsi_transport_srp) that provides common features for SRP initiator drivers
 - SCSI-ml already has SPI, FC, iSCSI, SAS transport classes
 - libsrp provides common features SRP target drivers
 - Merge scsi_transport_srp and libsrp?

Next steps (cont.)

- AIO and non-AIO event notification facility
 - tgttd needs to wait on AIO and non-AIO file descriptors, however, no handy interface for that.
 - blkctap daemon needs it too and uses a kernel patch to add AIO event support to select interface
 - But probably the patch would not go into mainline
 - Short term solution is adding non-AIO event support to AIO interface (under development)
 - Long term solution is unified event notification facility such as kevent (still no agreement on the design)

Next steps (fun stuff finally)

- Add metadata disk format support
 - easily takes blkmap code
- tgt only supports disk now, but there are many SCSI virtualization possibilities
 - A tape drive by using a file
 - A cdrom drive by using an iso image file
- Direct access to SCSI hardware (passthrough)
 - tgt executes SCSI commands via SG_IO

Further information

- scsifront/back drivers (2006/08/02)
 - <http://lists.xensource.com/archives/html/xen-devel/2006-08/msg00096.html>
- tgt website
 - <http://stgt.berlios.de/>
 - OLS2006 paper and slides are available

Thanks !