OPENFABRICS ALLIANCE

SRP Update

Author: David Dillow and Bob Pearson
Date: March 28, 2012

What's SRP?

- SCSI RDMA Protocol
- Essentially a transport for SCSI Commands
- Wire protocol looks similar to SMB Direct and Portals 4 on IB
  - Initiator sends request with memory descriptors
  - Target performs IO and issues RDMA requests
  - Target sends response to Initiator

Outline

- Where we've been
- What's going on now
- What can we do in the future?

Past: IOPS Scaling Work

- Faster storage ==> more pressure on transports to improve
- Bart Van Assche noted heavy lock contention on the fast-path for initiator (target too)
  - SCSI host lock protected entire command path
  - Started work to break up the lock
  - Work stalled, kept in SCST tree

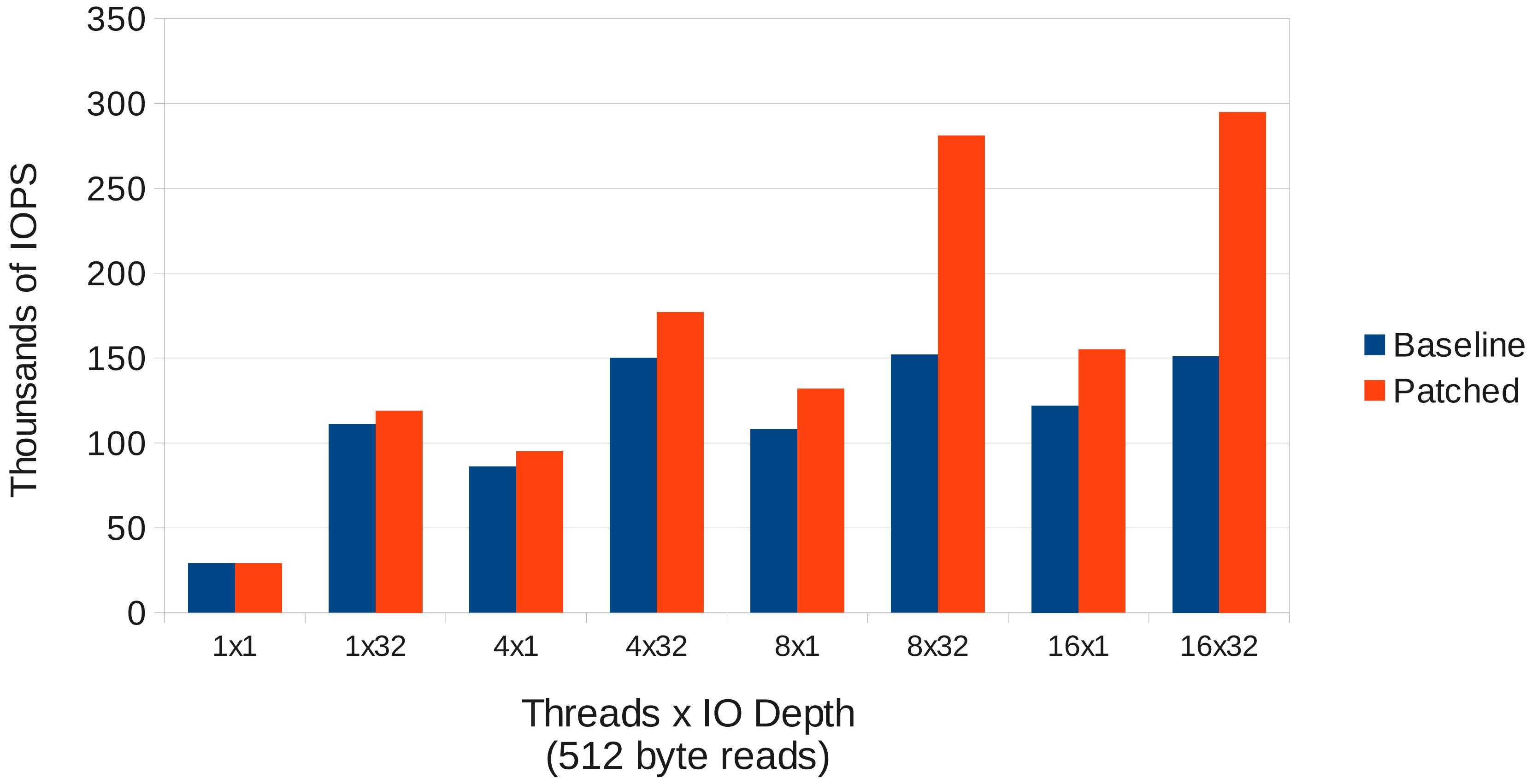
Past: IOPS Scaling Work

- SCSI maintainers pushed the host lock into the driver queuing function
- Picked up Bart's initial work
  - Broke it into digestible pieces
  - Cleaned up a few rough edges
  - Add a few optimizations on top

Past: IOPS Testing

- Custom SRP target designed to highlight lock overhead
  - Polling implementation in user space
  - Read-only LUNs
  - No data phase, just recv/send
- ConnectX 2, QDR
- Intel E5520 @ 2.27 GHz (quad-core, HT)

Past: IOPS Results



Past: Large IO Support

- Certain RAID systems work better with large IO requests
  - And most work better if you give them a full stripe
- Lustre prefers to send 1 MB requests
- Memory fragmentation often splits the 1 MB request into 256 4 KB pages

Past: Large IO Support

- Two SRP memory descriptor formats
  - Direct (1 contiguous region)
  - Indirect (Scatter/Gather list)
- SRP spec allows caching of Indirect table in request
  - However, it can be a partial cache!
  - Target must issue RDMA Read for remaining portion
- Most targets do not implement the spec!
  - Only support indirect descriptors that are fully cached in request
  - Limits us to 255 entries in S/G list

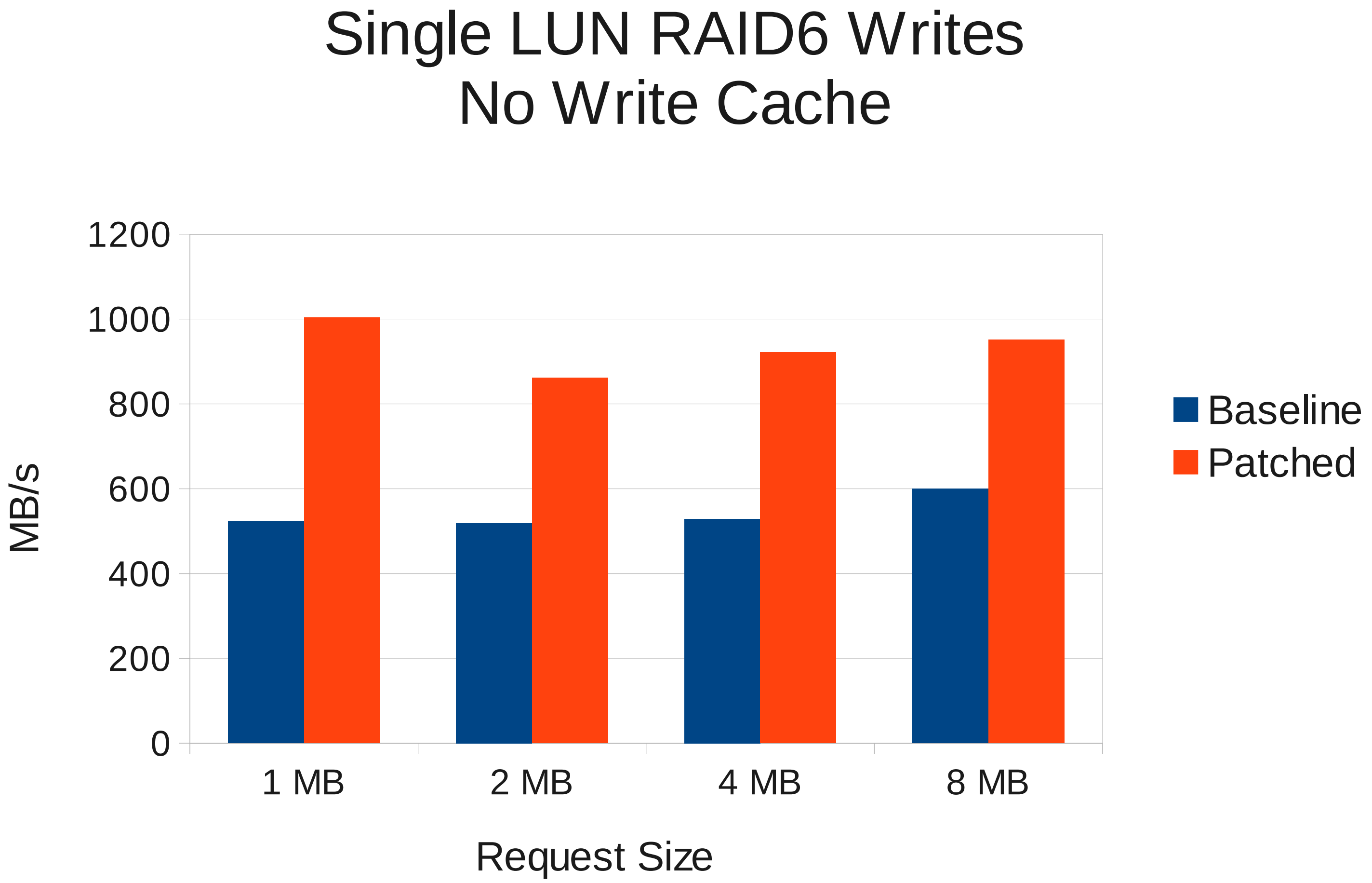
Past: Large IO Support

- FMR to the rescue!
  - But foiled by sg_tablesize
- Couldn't guarantee we be able to use FMR
  - Request had to be page aligned
  - Initiator asked for an FMR page size of 512 bytes
  - 128 KB request maximum (256 entry FMR)
    - (1 MB on older HCAs w/ minimum 4 KB FMR page size)
- Single attempt at FMR forced S/G limit
  - Had to be able to fall back to indirect table

Past: Large IO Support

- Use multiple FMR mappings
  - Just iterate over the list until we've mapped it all
  - Fall back to external indirect table on failure
  - Allows up to 2048 entries in S/G list
    - 8 MB requests can be guaranteed
- Requires target support for complete safety
  - FMR could fail (highly unlikely, maybe not possible at all)
  - No way to ask SCSI mid-layer to further break up the request
  - User must enable feature – no way to query target

Past: Large IO Support



Present: Error Handling/HA

- Fixes timeout inversion in initiator
- Allows disconnecting a single target without module unload
- Final result should be faster detection of failed target
- Currently working with Bart Van Assche to clean up patches
  - Probably too late for 3.4, but perhaps not

Future: Improved SCSI Support

- Bidirectional commands
  - Initiator only supports SCSI commands that send data in a single direction
  - Need bidirectional support to support Object Storage Device commands
  - Allows use of exofs over SRP
- Tagging support for initiator
  - Mostly there, just need to add hooks

Future: Data with Request/Reply

- Reducing latency in WAN environments, improve IOPS
- Limited to small data sizes (4-8 KB)
- Send data with command for write
  - Fairly straight-forward
  - Zero-copy possible
- Send data with response for read
  - More complicated
  - Currently requires a copy
    - Linux SCSI mid-layer expects to give us location to put data
    - Need new interface to tell mid-layer where the data is
- How to standardize?
  - SRP2 never finalized
  - No real interest in continuing that work?

Future: RDMA CM

- SRP currently uses IB CM to initiate the connection to the target
  - Locks us into InfiniBand
- Squeeze SRP_LOGIN_REQ into RDMA CM request
  - Opens up iWarp and RoCE
- Separate connection management from common path for command queuing
- Working code exists!

Future: Other Ideas?

- Multi-channel support
- T10-PI support

Questions?

dillowda@ornl.gov

rpearson@systemfabricworks.com