TSVAREA Notes - Prague
July 17, 2017
Spender Dawkins & Mirja Kuelewind, ADs

Note-taker: David Black

Blue Sheets
Note Well

---- Agenda Review

ADs generously thank to the TSV Area Review Team - Reviews have helped the ADs a lot.

Review of TSV WGs - see slide.
Review of TSV document progress since last IETF - see slide.

BANANA BOF happened just before TSVAREA meeting - has some relationship to Transport, and especially to MPTCP.

---- Path Aware Networking Research Group (RG)[panrg] - proposed (Brian Trammell)

This is about path awareness, primarily from a transport perspective. Have noticed path-awareness surfacing repeatedly in recent IETF work and recent BOF proposals.

Initial meeting of proposed RG is Wednesday, organizers are particularly interested in Transport expert comments on appropriate scope of this activity.

---- PASTE: Network Stacks Must Integrate with NVMM Abstractions - Michio Honda (NEC Labs Europe)
[collaborators from UPisa and NetApp]

Networking API for Non-Volatile Main Memory (NVMM) - more details at https://goo.gl/Ssreei

This is about interaction of NVDIMMs (persistent, byte-addressable main memory, low to very low latency) with networking.

Review of how NVMM is added to current I/O stack. Application in user space is on data path between file in NVMM and NIC interface to network.

Use case: Application must ensure that data sent by client is persistent before responding to client.

Most of the time taken without NVMM is in Disk I/O. NVMM eliminates that, so data path overhead becomes significant.

Data path overhead - cache misses matter, more so than copies. Copy to a buffer that is repeatedly reused results in buffer staying in CPU cache, so copy is cheap, but data stores to MMAP-ed file miss in cache as writes move to new locations in the file. Result is that data copy needs to be avoided to optimize writing an MMAP-ed file that's stored in NVMM.

So, eliminate the data copy by:

- Locating packet buffers in NVMM
  (NIC does DMA transfers directly to/from NVMM).
- Using zero-copy APIs.

This is what has been implemented in PASTE.

Application must perform an explicit flush operation to ensure that cached data is stored on NVMM.

Implementation based on extending netmap framework in Linux 4.6 and above. Preliminary results show good improvements. Working on improving the implementation, especially increasing flexibility of packet buffer structure to span multiple files.

Q&A:

Jake Holland: What about encrypted traffic?
Michio: Works with data that was encrypted before client sent it. Application can encrypt/decrypt in place.

Colin Perkins: Relationship of this work to TAPS WG?
Michio: TAPS should consider NVMM in API design.
Colin: Suggest discussion of this with TAPS WG.

Tommy Pauly: Apple working on API layers for mobile devices. Approach uses shared memory with user-space network stack, which is similar to PASTE. How does NVMM affect netmap framework?
Michio: Protocol stack location is not important (kernel mode vs. user mode).
Tommy: TAPS WG has not been looking much at data path optimization, TAPS looking at PASTE may help.

Mirja K.(AD): TAPS WG looking at message-based vs. streaming API, is message-based likely to work better for this?
Michio: Message-based API is preferable, definitely interested in what TAPS is doing.

Q: DPDK instead of netmap? VPP project has TCP/IP stack.
A: No, DPDK runs device driver in user space, prefer device driver in kernel, as is the case for the netmap framework.
Q: Will this work with NVRAM on PCIe?
A: Not if copy to a main memory buffer cache is required. Will work if the PCIe NVRAM is DMA addressable.

Q: What is future role of stream-based APIs?
A: Stream- based API can be re-hosted on a fast message-based API.

Q: Any data comparing this to RDMA?
A: No, may look at this in future. RDMA bypasses everything, lacks some important transport protocol features.

---- Last words from ADs

Spencer: ADs are very happy with work being done in TSV Area. Transport reviews have moved from last surprises at IETF Last Call (MPLS/UDP) to getting early Transport review requests from WGs in other areas that are considering adopting drafts.

Mirja: No BOFs this time, new proposals encouraged, as AQM WG likely to be closed soon.

---- NOMCOM

Mirja is up for review this time, please think about good nominees, especially for TSV, provide feedback on willing nominees, and consider running yourself.

Think past this Nomcom cycle, too. Spencer is serving his 3rd term as AD, and won’t be serving a 4th term!