Multipath TCP integration in Linux kernel

Florian Westphal
4096R/AD5FF600 fw@strlen.de
80A9 20C5 B203 E069 F586
AE9F 7091 A8D9 AD5F F600

Sep 2019

Current status

- no mptcp support in Linux
- Fork available at `http://multipath-tcp.org`
  - started in 2009
  - adds MP-TCP to Linux network stack
    - deemed non-upstreamable
- on-going "mptcp-next" development effort
  - "complete rewrite" aimed towards merge into mainline kernel
  - push work to userspace where possible (esp. path management)

Initial feature set for merging

"server use case"

- MPTCPv1 (rfc6824bis)
- Active-Backup only
- netlink based path manager, shared with multipath-tcp fork
- Handle incoming joins only (as opposed to initiating multiple subflows)

Current status: mptcp-next

- `bool CONFIG_MPTCP` – no code changes with MPTCP=n in kernel config
- TCP is TCP
- for MPTCP: `socket(AF_INET, SOCK_STREAM, IPPROTO_MPTCP)`
  - `MP_JOIN` work in progress
  - single flow (with Data Sequence signal mapping)
  - doesn't announce any extra addresses so far by default
- patch adds roughly 4000 LOC
- very few changes in core TCP stack:
  `7 files changed, 114 insertions(+), 9 deletions(-)`

Current status: mptcp-next (2)

- MPTCP meta socket
  - created on behalf of userspace via `socket`, `accept`, etc.
  - contains MPTCP state: logical sequence numbers, keys, token, ...
  - subflows (tcp sockets) are kept on a list via this mptcp meta-socket
- Userspace doesn't interact with TCP subflows directly (file descriptor identifies MPTCP meta socket)
- ULP is used to plumb tcp sockets to the mptcp parent socket

ULP: Upper Layer Protocol

- Kernel infrastructure to add protocol on top of TCP
  - added 2017 for kTLS

    ```
    ret = setsockopt(tcpsockfd, IPPROTO_TCP, TCP_ULP, "tls", sizeof("tls"));
    setsockopt(tcpsockfd, SOL_TLS, TLS_TX, &tls12, sizeof(tls12));
    ```
  - allows to attach blob of data to the tcp socket
  - allows to override/replace socket function pointers, e.g. call different function if userspace writes or reads from such a socket
- mptcp-next adds "hidden" `"mptcp"` ULP
  - overrides a few tcp socket functions, e.g. `sk_data_ready`
- MPTCP ULP blob added to all MPTCP subflow (tcp) sockets
  - contains backpointer to parent mptcp meta-socket

Completed/Merged work

Parts of infrastructure needed has been upstreamed already

- socket buffer extensions
  - will be used to carry dss mapping from mptcp layer to tcp
  - merged with ipsec and bridge netfilter as first users
- ULP inet diag support
  - will be used to export MPTCP information to userspace for statistics/troubleshooting etc.
  - merged with kTLS as first user

Ongoing work

- make MPTCP `join` work for real
  - somehat works right now, depending on who initiates
  - still suffers from bugs (e.g. resource leaks)
- ULP inet diag support
- integrate netlink path manager
  - to export MPTCP information to userspace for statistics/troubleshooting etc.
- add MIB stat counters
- IPv6 support

Future work

- switch to MPTCPv1 (prototype is still v0)
- client use case – active opening of new subflows
- MPTCP support in upstream packetdrill for testing
- API: see what might be needed
  - `connectx`
  - add support for normal tcp setsockopts?
    - might not even make sense to begin with
    - would have to "remember" to replay settings for new subflows
- Performance optimizations
- ability to "join" regardless of tcp ports – "token only"