

BIG TCP for UDP tunnels

Alice Mikityanska

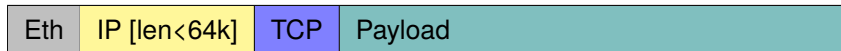
2026-10-07

What is this all about?

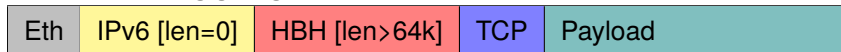
- ▶ BIG TCP extended GRO/GSO aggregation size beyond 64k
- ▶ ...but it only supported plain TCP.
- ▶ Production deployments of Kubernetes/Cilium often use VXLAN/GENEVE overlays rather than direct routing.
- ▶ Can we leverage BIG TCP with tunnels?

BIG TCP recap

Standard GRO/GSO:

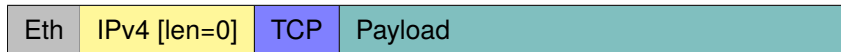


BIG TCP aggregate:



Driver strips HBH before TX.

Later extended to IPv4 using the len=0 hack:



Tunneled packet



Outer IP, UDP, and inner IP headers store len.

4 combinations of outer/inner IPv4/IPv6.

Even more variations of BIG/small len layouts.

The biggest problem: teach drivers to strip HBH correctly in all cases.

Step 1: Getting rid of HBH

- ▶ Length can be restored from `skb->len` (like IPv4).
- ▶ Unifies IPv4 and IPv6 BIG TCP representations.
- ▶ All drivers strip HBH before TX.
- ▶ No real use beyond tcpdump.
- ▶ Solves the indicated problems.



No HBH, use `skb->len`.

[PATCH net-next v5 00/12] BIG TCP without HBH in IPv6

<https://lore.kernel.org/r/20260205133925.526371-1-alice.kernel@fastmail.im/>

ipv6: Support new TSO without HBH - #1329

<https://github.com/the-tcpdump-group/tcpdump/pull/1329>

Addressing the remaining gaps

- ▶ Encoding large length in UDP header.

- ▶ Use len=0 and helpers to decode.

A few preexisting 16-bit overflows exposed.

```
l2tp: Drop large packets with UDP encap
net: ipv4: Fix UDP length overflow with PMTU discover and big MTU
net: ipv6: Fix UDP length overflow with PMTU discover and big MTU
net: ipv6: Clamp to IP6_MAX_MTU in ip6_dst_mtu_maybe_forward
geneve: Fix off-by-one comparing with GRO_LEGACY_MAX_SIZE
```

- ▶ Be careful about the checksum.

Full length goes into the pseudoheader.

- ▶ Be careful about packets on the wire with len=0.

```
udp: Validate UDP length in udp_gro_receive
```

Addressing the remaining gaps

► Implicit ceiling in `skb_segment`.

► `GSO_BY_FRAGS == 65535` was used as a magic value for `gso_size`.

Partial GSO is needed by NICs that only support inner csum offload.

`net: Enable BIG TCP with partial GSO`

```
- partial_segs = min(len, GSO_BY_FRAGS - 1) / mss;  
+ partial_segs = len / mss;
```

Addressing the remaining gaps

- ▶ Bumping `tso_max_size` in VXLAN and GENEVE.
- ▶ The easiest and straightforward.

```
+ netif_set_tso_max_size(dev, GSO_MAX_SIZE);
```

Turning on the feature

```
ip link set eth0 \  
    gso_max_size 196608 gso_ipv4_max_size 196608 \  
    gro_max_size 196608 gro_ipv4_max_size 196608  
  
ip link set vxlan0 \  
    gso_max_size 196608 gso_ipv4_max_size 196608 \  
    gro_max_size 196608 gro_ipv4_max_size 196608
```

Cilium auto-probes BIG TCP tunnels support.

```
helm install cilium cilium/cilium --version 1.20.2 \  
... \  
--set enableIPv4BIGTCP=true \  
--set enableIPv6BIGTCP=true \  
...
```

Performance evaluation

`netperf -t TCP_STREAM` on mlx5e, Gbit/s

Config	Baseline	BIG TCP	Gain
VXLAN, MTU=1500	34.44	39.56	14.9%
VXLAN, MTU=8192	55.68	61.47	10.4%
VXLAN, MTU=1500, software GSO	21.82	29.39	34.7%
GENEVE, MTU=1500	37.39	40.89	9.4%
GENEVE, MTU=8192	58.03	61.46	5.9%

Measurements by Daniel Borkmann.

Testing

- ▶ Selftest that covers all combinations of IPv4/IPv6 inner/outer with VXLAN and GENEVE.
- ▶ Check BIG TCP packets and almost no SACKs.
- ▶ Increase `tcp_min_tso_segs` and `initcwnd`.
- ▶ Also covers the software checksum path:
 - ▶ Disable TX CSUM offload.
 - ▶ Software GSO splits the inner packet before passing it to the lower device.
 - ▶ Checksum code is broken? $\Rightarrow$ Packets are dropped.

Future work and discussion

- ▶ Foo-over-UDP – a low-hanging fruit.
- ▶ Double GENEVE with GRO hints compatibility.
<https://lore.kernel.org/r/f65cb565-0899-4e0d-932a-28e15b3fe864@redhat.com/>
- ▶ Test coverage for "GSO after GRO".
<https://lore.kernel.org/r/d66c1cb5-6201-4390-a7e7-0a8461bc069b@redhat.com/>
- ▶ Bug in `skb_segment` with `frag_list` SKBs.
<https://lore.kernel.org/r/1cf2746f-f15b-406d-bb49-2c3f2dee1ec2@app.fastmail.com/>
- ▶ Early flush of malformed UDP packets.
<https://lore.kernel.org/r/CAD0BsJVntu3ENqpqQFf0oJ69gBZxoN8iMWRUnF-GdrZ2uaAy7A@mail.gmail.com/>
- ▶ Explore UDP GRO > 64k.
<https://lore.kernel.org/r/ad089459-361c-4209-9fdc-675cc4a658bc@app.fastmail.com/>
- ▶ Audit/fuzz all 16-bit boundaries.

Thanks for your attention!

Time for questions