



PCI Data Path Optimization with CQE Coalescing and New Ethtool Parameters

Haiyang Zhang
Principal Software Engineer
Microsoft / Linux Systems Group

Linux Plumbers Conference 2026, Prague
Networking Track



LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

Agenda

01 The problem

- Per-packet completions create interrupt pressure and PCI overhead on the MANA receive path.

02 The design

- Merging several receive completions into one CQE, bounded by a frame count and a timer.

03 The patch set

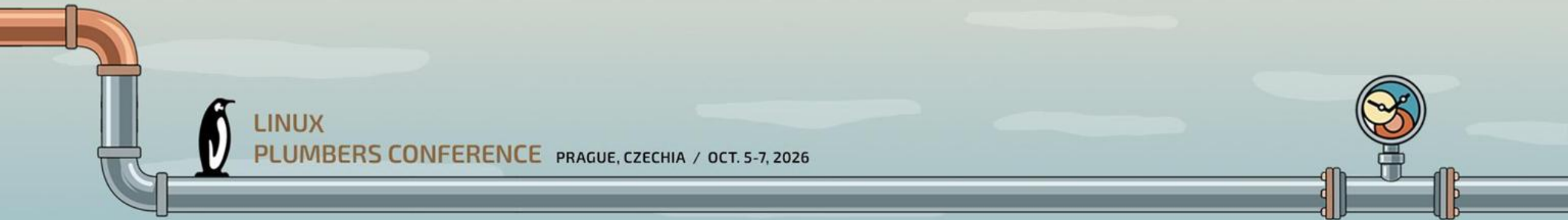
- Four net-next commits plus user-space ethtool support, upstreamed during 2026.

04 Results and tuning

- Significant throughput gain, more packets per interrupt, and guidance for tuning.

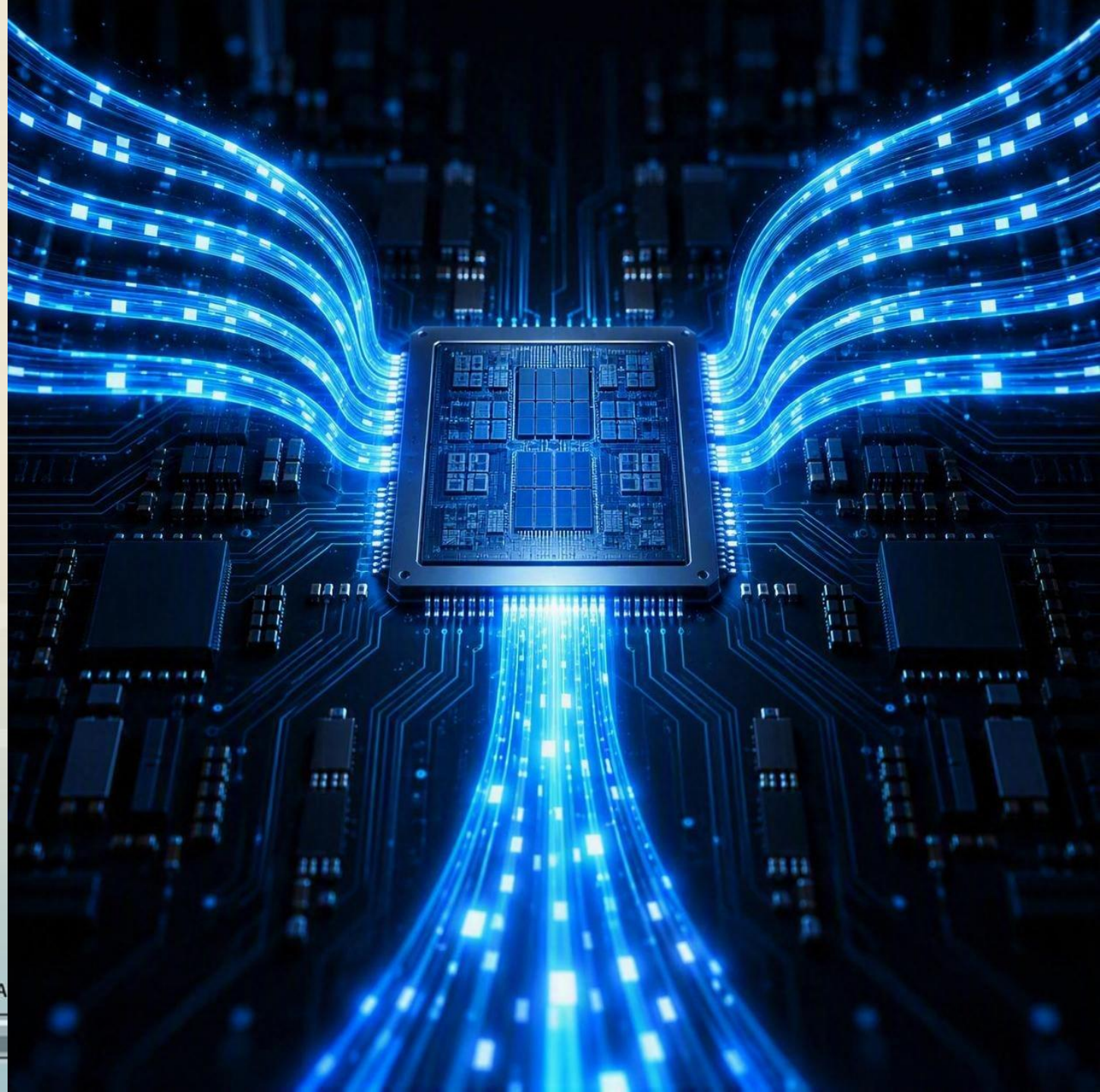


01 The problem: one CQE per packet



The Azure MANA NIC receive path

In the regular receive path, each packet generates its own completion entry for the driver to process.



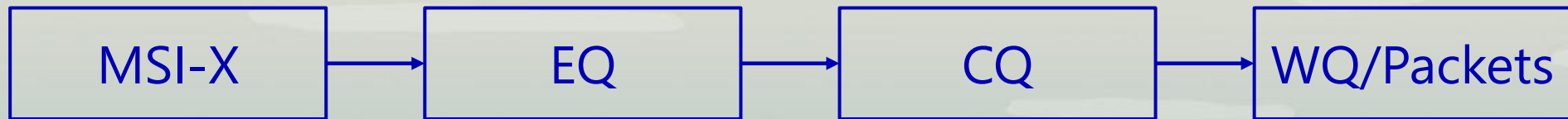
Per-packet completions are expensive

Cost on the PCI link

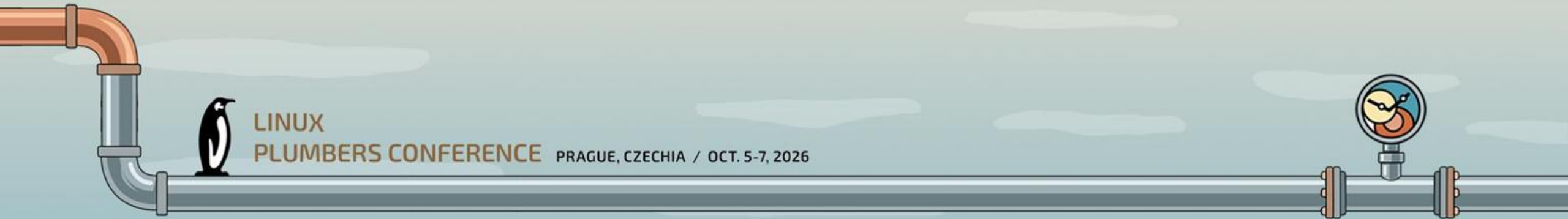
- Each received packet can produce a separate CQE.
- Every completion costs PCI transactions.
- The cost increases directly with packet rate.

Cost on the CPU

- Higher interrupt rate and more softirq work per second.
- Per-CQE bookkeeping repeats in the hottest part of the fast path.

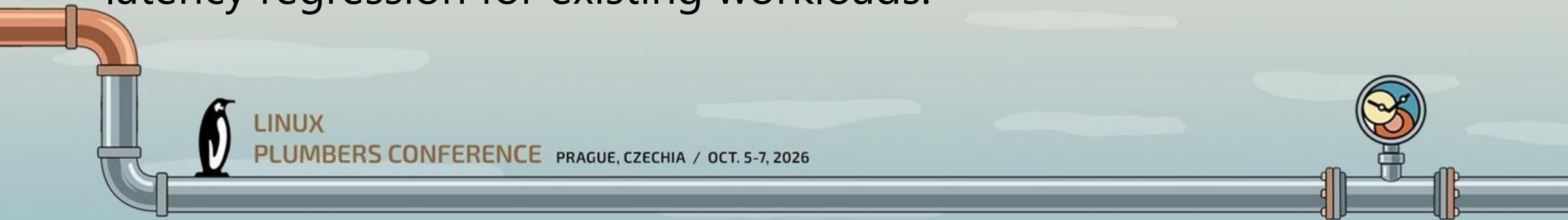


02 The design



CQE coalescing: one completion, multiple packets

- The NIC merges multiple receive completions into one aggregated CQE that represents multiple packets.
- The driver processes more packets per CQE and pays the per-CQE overhead far less often.
- Hardware advertises the capability through a PF flag, so the driver enables it only where it is supported.
- The feature is disabled by default in the MANA driver, to avoid any latency regression for existing workloads.



How a coalesced CQE is built



Disabled by default in the MANA driver; enabled only where the hardware PF flag advertises support, and user sets it with ethtool.



How a coalesced CQE is formed

Frame limit

- rx-cqe-frames controls how many packet completions the device may combine before it generates a coalesced CQE.

Time limit

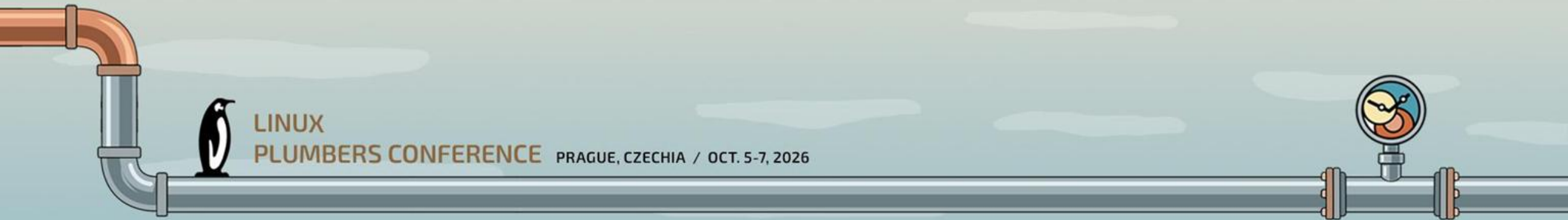
- rx-cqe-nsecs bounds how long the device waits after the first packet arrives before reporting the aggregated completion.

Driver processing

- The driver recognizes the coalesced CQE type and walks each packet the entry carries, then hands them up the stack.



03 The Implementation

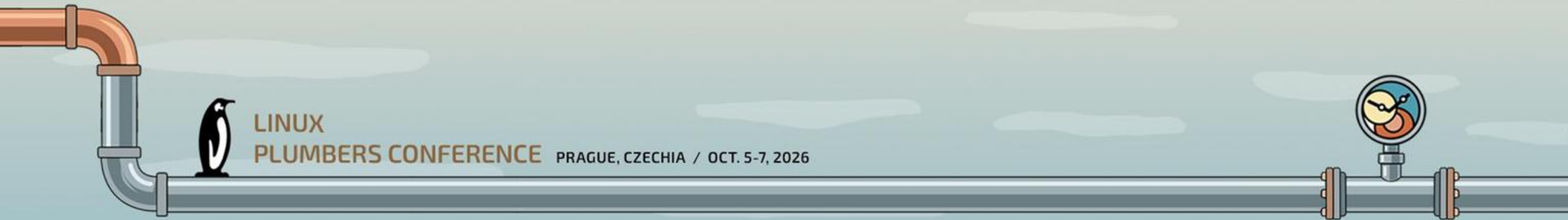


net: ethtool | commit
dc3d720e12f6

New ethtool parameters

ETHTOOL_A_COALESCE_RX_CQE_FRAMES
ETHTOOL_A_COALESCE_RX_CQE_NSECS

Two generic parameters,
configurable through the
standard Linux networking
model.



MANA driver support for coalesced CQEs

The driver checks for and processes the `CQE_RX_COALESCED_4` type, taking up to four received packets from a single completion entry and updating the receive path and structures to match.

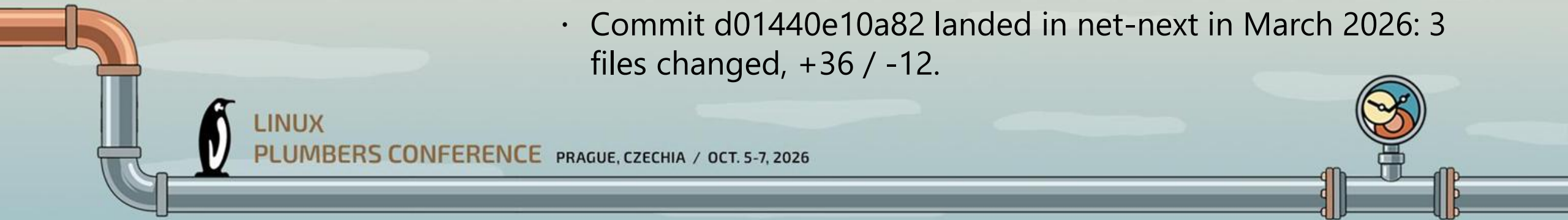
Commit `c2fe3ff3d66d`, net-next, March 2026 - 3 files changed, +113 / -29.

Coalescing ships disabled by default, and `rx-cqe-nsecs` is read-only on this NIC.



Counting what coalescing does

- New ethtool counters report how many RX CQEs arrived in coalesced form, per queue, so operators can confirm the feature is active and see how effectively completions are being merged in production traffic.
- Commit d01440e10a82 landed in net-next in March 2026: 3 files changed, +36 / -12.



```
net: mana | commit  
55d20f50a221
```

```
CQE_RX_COALESCED_4    /* up to 4  
packets */
```

```
CQE_RX_COALESCED_8    /* up to 8  
packets */
```

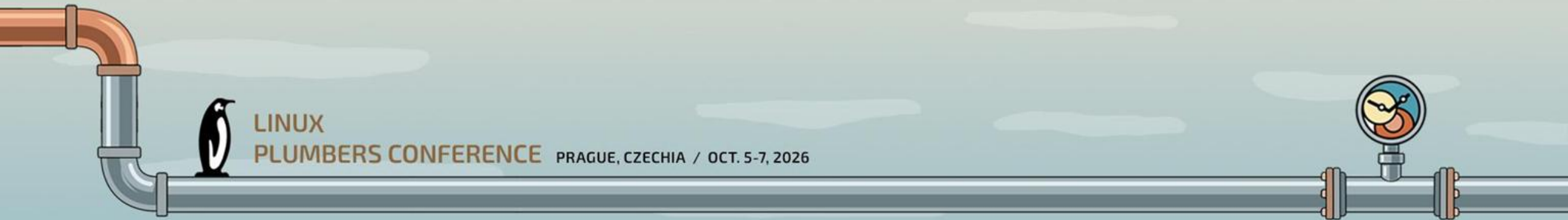
```
ethtool -C [VFnic] rx-cqe-frames 8
```

```
net-next, August 2026
```

```
4 files changed, +150 / -64
```

Extending to eight packets per CQE

Hardware advertising the relevant PF flag can pack up to eight packets into one completion entry, with CQE processing, ethtool handlers and per-queue statistics updated to match.



Upstreaming timeline

Mar 2026
net: ethtool API
dc3d720e - 5 files, +39 / -2
MANA coalesced CQEs
c2fe3ff3 - 3 files, +113 / -29
Per-queue counters
d01440e1 - 3 files, +36 / -12

Mar 2026
ethtool netlink (user space)
d35d87fb - 5 files, +25

Aug 2026
Up to 8 packets per CQE
55d20f50 - 4 files, +150 / -64

Four kernel patches in net-next plus user-space ethtool netlink support, all upstream in 2026.



**LINUX
PLUMBERS CONFERENCE** PRAGUE, CZECHIA / OCT. 5-7, 2026



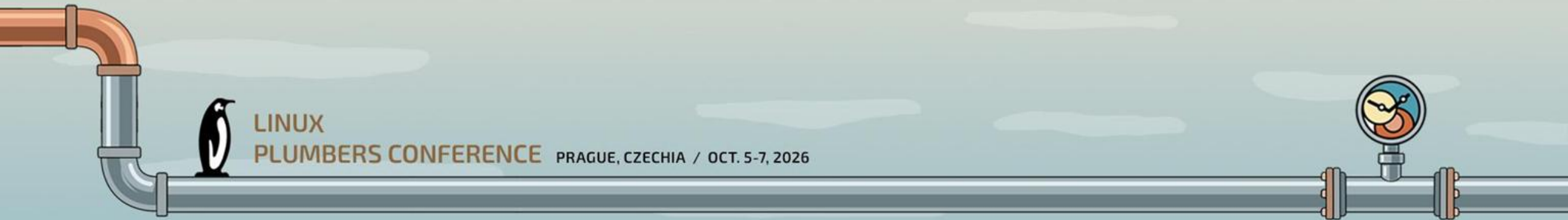
How other NIC drivers do it

Mellanox

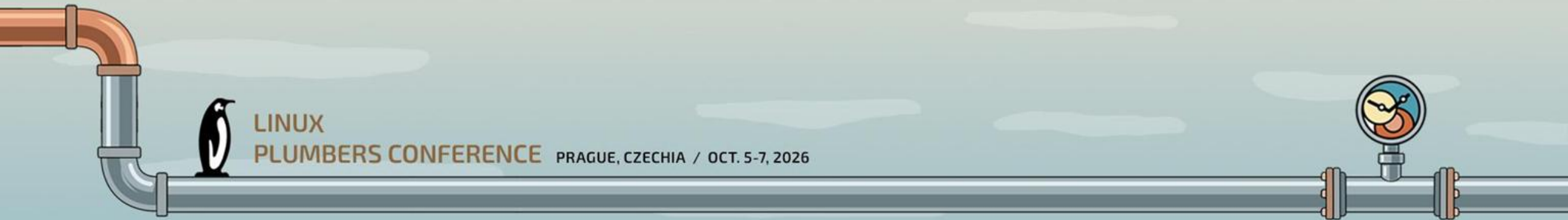
- Mlx nic merges receive completions in similar way (CQE compression).
- Available before this generic ethtool API.
- Tunable through an ethtool private flag, **rx_cqe_compress**.

Meta fbnic

- fbnic also coalesces receive completions in hardware.
- Recently made the timer configurable through this generic API: **ETHTOOL_COALESCE_RX_CQE_NSECS**



04 Results and tuning



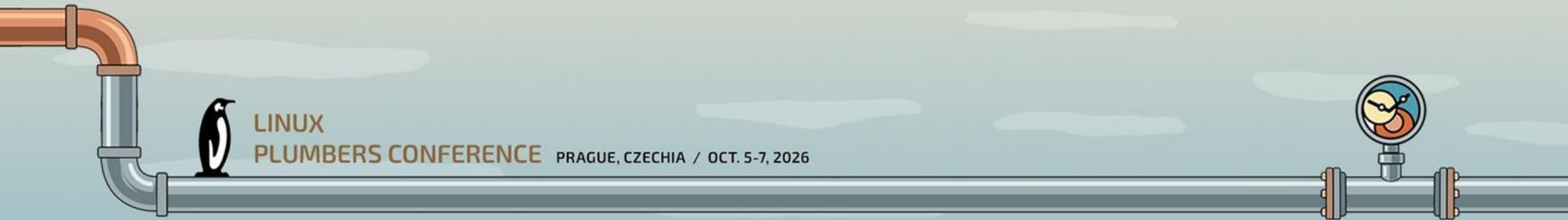
Tuning for your workload

Favor throughput

- Raise rx-cqe-frames toward the hardware maximum.
- Fewer interrupts and less repeated per-packet work.
- Best for bulk transfer and high throughput workloads.

Favor latency

- Set rx-cqe-frames to 1 to disable coalescing.
- Completions are reported as packets arrive.
- This is the default, so latency-sensitive workloads are unaffected.



Turning coalescing – the command lines

ethtool command line

merge up to 4 or 8 packet completions per CQE

```
ethtool -C [vfnic] rx-cqe-frames 4 //or 8
```

back to one packet per CQE

```
ethtool -C [vfnic] rx-cqe-frames 1
```

read the current settings

```
ethtool -c [vfnic]
```

.....

```
rx-cqe-frames:          8
```

```
rx-cqe-nsecs:          2048
```



Tuning the rx-cqe-frames dial

Favor latency

Favor throughput

1

rx-cqe-frames 1 (default)

4

rx-cqe-frames 4

8

rx-cqe-frames 8 (max)

Completions reported as packets arrive.
No added delay; existing workloads unaffected.

Fewer interrupts, less repeated per-packet work.
35-60% throughput gain observed on ARM64.

rx-cqe-frames	Throughput
1 (default)	112 Gbps
4	155 Gbps
8 (max)	181 Gbps

Test conditions

256 concurrent threads,
64-vCPU VMs,
200G link,
Arm64 CPU



**LINUX
PLUMBERS CONFERENCE** PRAGUE, CZECHIA / OCT. 5-7, 2026



Reference – the patch list

- Four kernel patches
- One user-space ethtool patch.

The patches - git.kernel.org/pub/scm/linux/kernel/git/netdev/net-next.git/commit/?id=<commit> (ethtool.git for the user-space patch)

net: ethtool - add COALESCE_RX_CQE_FRAMES/NSECS [dc3d720e12f602059490c1ab2bfee84a7465998f](https://git.kernel.org/pub/scm/linux/kernel/git/netdev/net-next.git/commit/?id=dc3d720e12f602059490c1ab2bfee84a7465998f)

net: mana - add support for RX CQE coalescing [c2fe3ff3d66d6f53ec5857c277fae5b3ff9881c1](https://git.kernel.org/pub/scm/linux/kernel/git/netdev/net-next.git/commit/?id=c2fe3ff3d66d6f53ec5857c277fae5b3ff9881c1)

net: mana - ethtool counters for coalesced RX CQEs [d01440e10a82cae2c4a28c76e46e6a8b94b27a84](https://git.kernel.org/pub/scm/linux/kernel/git/netdev/net-next.git/commit/?id=d01440e10a82cae2c4a28c76e46e6a8b94b27a84)

net: mana - extend RX CQE coalescing up to 8 packets [55d20f50a221bdd7b19befc6e3685a2ef5d4c07b](https://git.kernel.org/pub/scm/linux/kernel/git/netdev/net-next.git/commit/?id=55d20f50a221bdd7b19befc6e3685a2ef5d4c07b)

ethtool (user space) - netlink support for RX CQE params [d35d87fbcda97fe31df79d62277743214641892a](https://git.kernel.org/pub/scm/linux/kernel/git/netdev/net-next.git/commit/?id=d35d87fbcda97fe31df79d62277743214641892a)



LINUX
PLUMBERS CONFERENCE PRAGUE, CZECHIA / OCT. 5-7, 2026



Thank you

Haiyang Zhang, Microsoft

haiyangz@microsoft.com | Linux Plumbers Conference
2026, Prague



**LINUX
PLUMBERS CONFERENCE**

PRAGUE, CZECHIA / OCT. 5-7, 2026

