



# The Path to DPDK Speeds for AF\_XDP

---

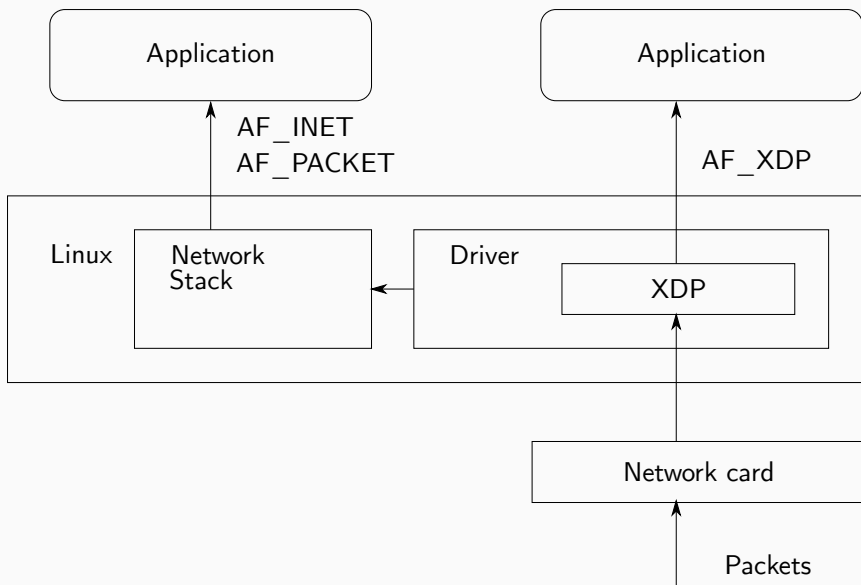
Magnus Karlsson, Björn Töpel

[magnus.karlsson@intel.com](mailto:magnus.karlsson@intel.com), [bjorn.topel@intel.com](mailto:bjorn.topel@intel.com)

Linux Plumbers Conference, Vancouver, 2018

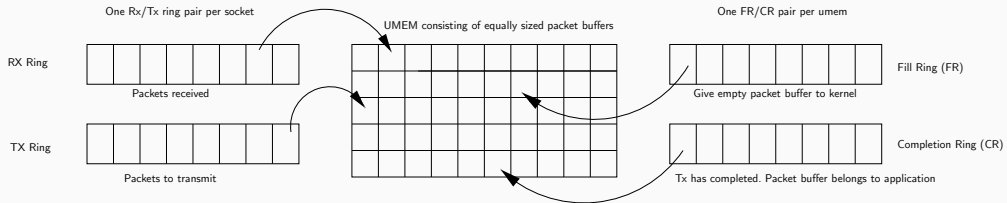
# Legal Disclaimer

- Intel technologies may require enabled hardware, specific software, or services activation. Check with your system manufacturer or retailer.
- No computer system can be absolutely secure.
- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit [www.intel.com/benchmarks](http://www.intel.com/benchmarks).
- Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.
- All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.
- Intel, the Intel logo, and other Intel product and solution names in this presentation are trademarks of Intel.
- Other names and brands may be claimed as the property of others.
- ©2018 Intel Corporation.



- Ingress
  - userspace XDP packet sink
  - XDP\_REDIRECT to socket via XSKMAP
- Egress
  - no XDP program
- Register userspace memory to kernel (UMEM)
- Pass packet buffer ownership via rings with descriptors
- Fill ring (to kernel) / Rx ring (from kernel)
- Tx ring (to kernel) / Completion ring (from kernel)
- copy mode (DMA to/from kernel allocated frames, copy data to user)
- zero-copy mode (DMA to/from user allocated frames)

# AF\_XDP 101

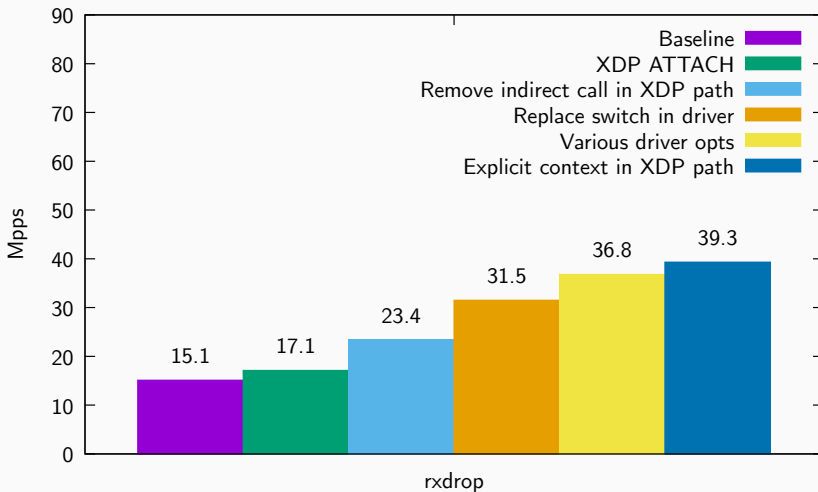


# Baseline and blueprint

- Baseline: 64B @ ~15-22 Mpps
- Blueprint
  - do less (instructions)
  - talk less (coherency traffic)
  - do more at the same time (batching, i\$)
  - Land of Spectres: fewer retpolines, fewer retpolines, fewer repolines

- `XDP_ATTACH` and `bpf_xsk_redirect`, attach at-most one socket per netdev queue, load built-in XDP program, 2-level hierarchy
- remove indirect call, `bpf_prog_run_xdp`
- remove indirect call, XDP actions switch-statement ( $\geq 5 \implies$  jump table)
- driver optimizations (batching, code restructure)
- `bpf_prog_run_xdp`, `xdp_do_redirect` and `xdp_do_flush_map`: per-CPU struct `bpf_redirect_info` + struct `xdp_buff` + struct `xdp_rxq_info` vs explicit, stack-based context

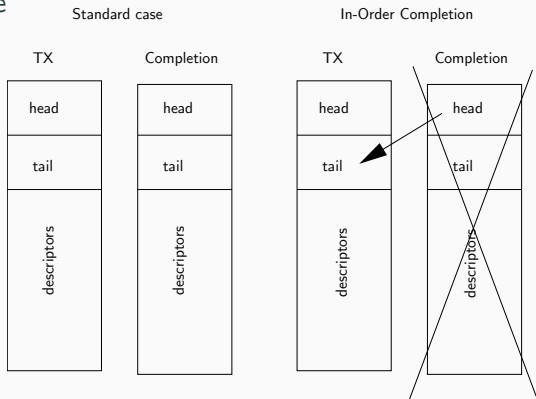
# Ingress, results<sup>1</sup>, data not touched



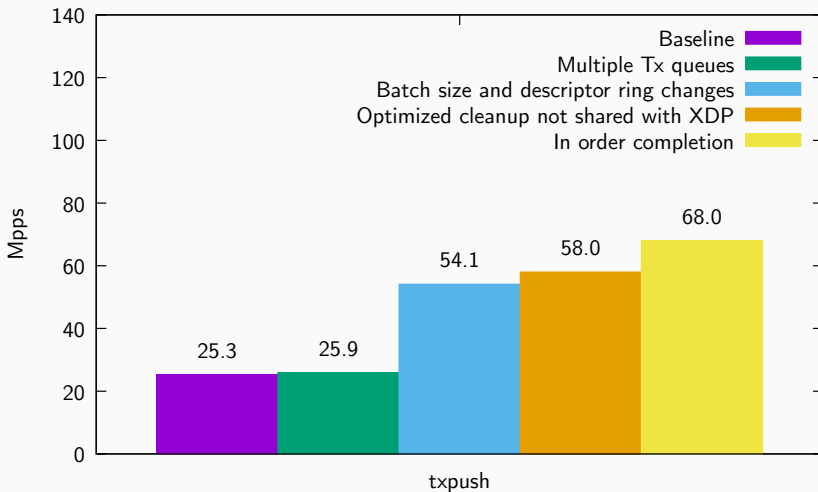
<sup>1</sup>Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to <http://www.intel.com/performance/datacenter>.



- Tx performance capped per HW queue  
⇒ multiple Tx sockets per UMEM
- Larger/more batching, larger descriptor rings
- Dedicated AF\_XDP Tx queues
- In-order completion, setsockopt XDP\_INORDER\_COMPLETION

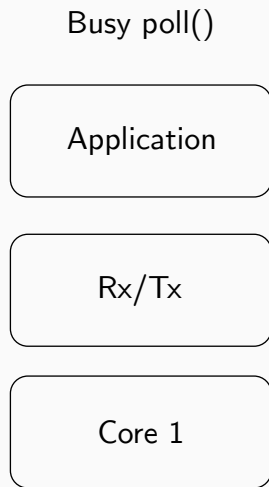
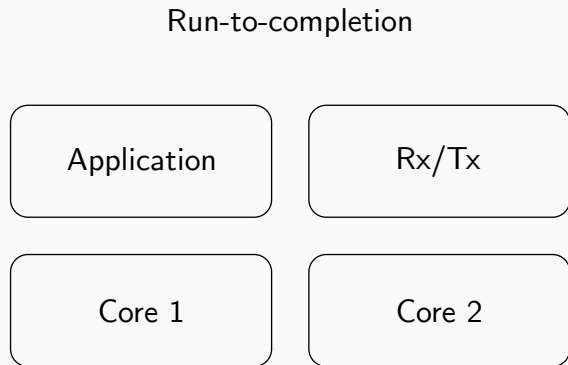


# Egress, results<sup>1</sup>, data not touched

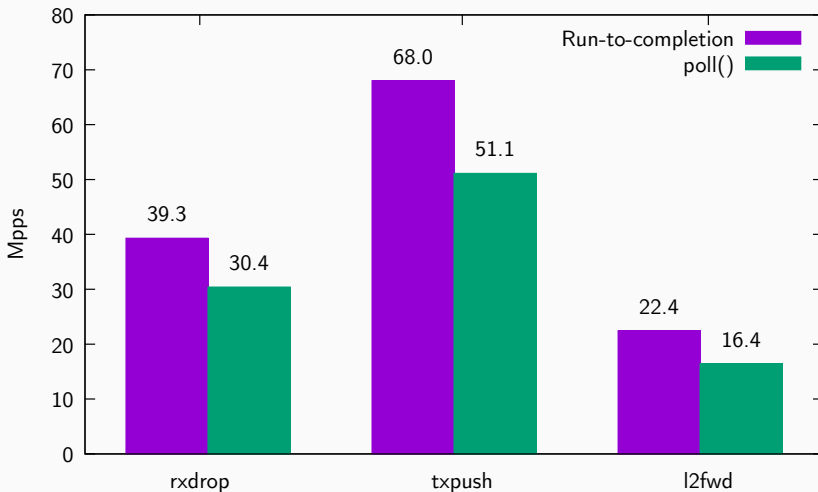


<sup>1</sup>Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to <http://www.intel.com/performance/datacenter>.

## Busy poll() vs run-to-completion



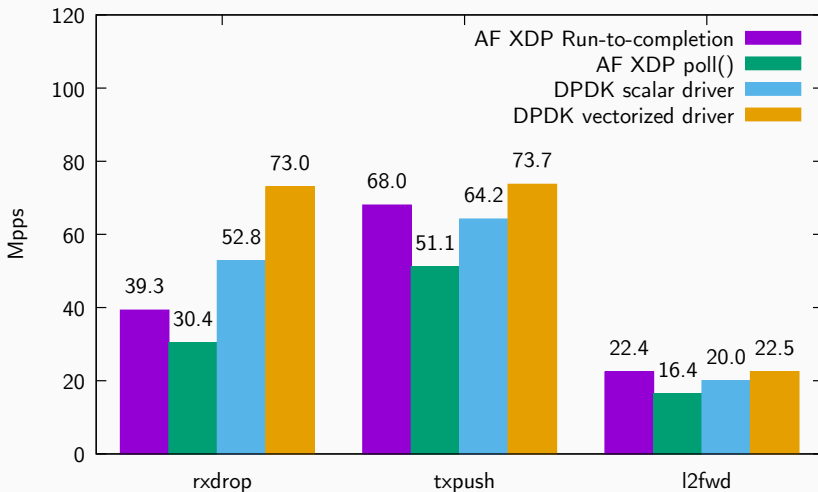
# Busy poll() vs run-to-completion, results<sup>1</sup>



<sup>1</sup>Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to <http://www.intel.com/performance/datacenter>.

- Userspace, vectorized drivers
- “Learning from the DPDK” [http://vger.kernel.org/netconf2018\\_files/StephenHemminger\\_netconf2018.pdf](http://vger.kernel.org/netconf2018_files/StephenHemminger_netconf2018.pdf)

# Comparison with DPDK, results<sup>1</sup>



<sup>1</sup>Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to <http://www.intel.com/performance/datacenter>.

Upstream!

- XDP: switch-statement
- Rx/Tx: drivers
- Rx: XDP\_ATTACH and bpf\_xsk\_redirect
- Tx: multiple Tx sockets per UMEM
- General leftovers still to-be-upstreamed: libbpf AF\_XDP support (easier to consume), selftest

## Future work

- hugepage support, less fill ring traffic (`get_user_pages`)
- fd.io/VPP work vectors (i\$, explicit batching in function calls)
- “XDP first” drivers
- collaborate/share code with RDMA (e.g. `get_user_pages`)
- Type-writer model (currently not planned)



# Thanks!

- Ilias Apalodimas
- Daniel Borkmann
- Jesper Dangaard Brouer
- Willem De Bruijn
- Eric Dumazet
- Alexander Duyck
- Mykyta Iziumtsev
- Jakub Kicinski
- Song Liu
- David S. Miller
- Sridhar Samudrala
- Yonghong Song
- Alexei Starovoitov
- William Tu
- Anil Vasudevan
- Jingjing Wu
- Qi Zhang

Questions?

