Aeron Transport Kernel Bypass

Aeron
Sponsored by Adaptive

10GigE network adapters are capable of tens of millions of packets per second. However, per core packet processing is limited to approximately 500,000 packets per second with Linux kernel implementations of BSD sockets.

Multiple cores and large packets are needed to utilise the full bandwidth, with each core reaching capacity at less than half the bandwidth of a 10GigE network, with the BSD sockets' API for UDP.

Since Meltdown and Spectre, mitigations have been applied to increase the system call costs, which has created further challenges.

These limitations can be addressed with kernel bypass technologies.

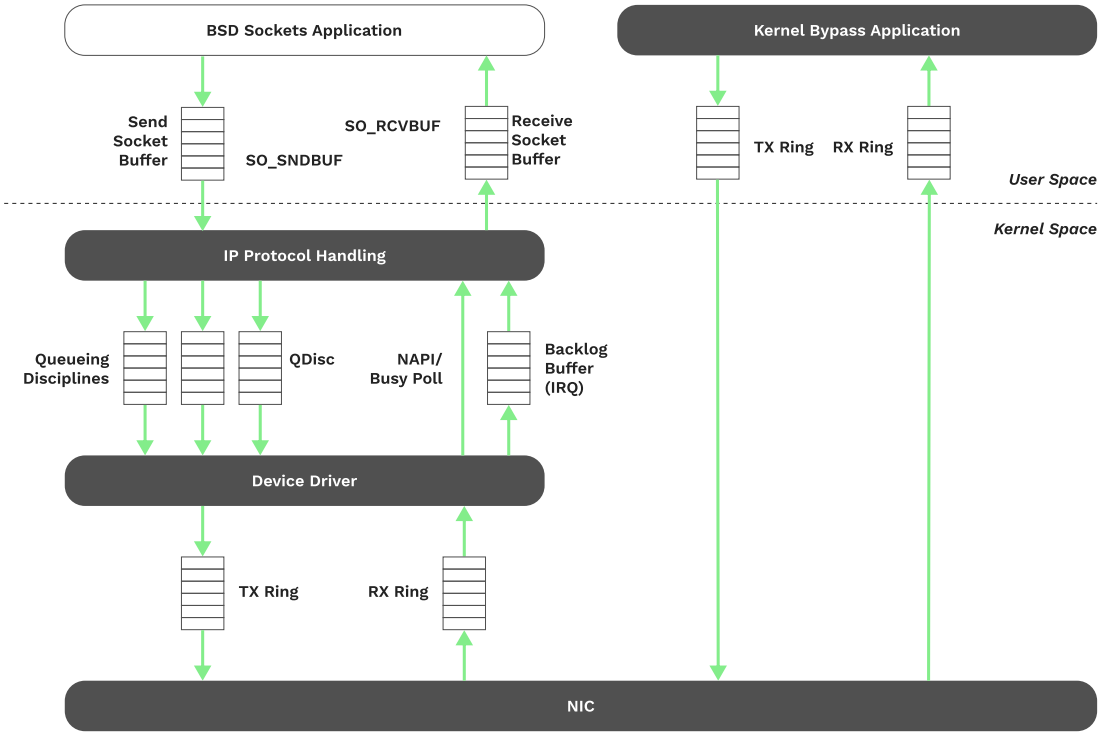
Introducing Aeron Kernel Bypass Extensions

Aeron leverages kernel bypass technologies which go directly to TX/RX ring buffers of network cards and provide APIs designed for performance, with zero copy semantics and real-time lock-freedom/ wait-freedom.

Aeron Kernel Bypass significantly reduces CPU overhead in the call path from user space to network card, yielding lower and more predictable latency. As the APIs are typically poll mode, they work well for Aeron, as all network interactions are non-blocking.

Sending and receiving Ethernet frames using raw kernel bypass APIs requires a complex programming approach to the entire network stack. Some libraries have helper functions for common tasks, but come with varying costs. Unlike other UDP messaging products, Aeron provides flow and congestion control on top of its UDP implementation, avoiding the need for a full implementation for WAN usage. Aeron extensions are loaded by the C media driver with no required changes to Aeron client APIs for usage.

Linux Network Stack



Aeron Transport Kernel Bypass

Aeron
Sponsored by Adaptive

The Extensions

ef_vi - Ethernet Frame Virtual Interface

Ef-vi is specific to Solarflare/Xilinx network cards.

- Flow steering of packets to user space ring buffers based on filters. This can co-exist with the kernel interface to allow select data to bypass the kernel.
- Efficient APIs to access ring buffers - zero copy and lock-free.
- Lowest possible latency and maximum throughput for a given network from a single thread.
- Single digit, microsecond, one-way latencies between servers on the same network segment.
- TCPDirect or Onload wrapping for ease of use at the expense of performance, compared to raw ef_vi.
- Onload enables a BSD sockets API, running in the user space, for a performance middle ground between kernel BSD sockets and ef_vi.

DPDK - Data Plane Development Kit

The DPDK is part of the Linux Foundation.

- Support for all major NICs and AWS Nitro instances.
- Implementations though similar, need to be coded - Aeron supports AWS Nitro cards.
- Using DPDK, the network interface is unbound from the kernel driver, to one specific to DPDK. This results in the interface disappearing from the standard Linux network tools and is instead dedicated to the application.
- DPDK gives the lowest possible latency and maximum throughput for a given network from a single thread. Latency can vary due to Nitro card wakeup, firewalls or other network latencies.

VMA - Mellanox Messaging Accelerator

VMA is specific to NVidia/Mellanox network cards. It includes:

- A BSD structured API that provides kernel bypass networking using custom drivers.
- VMA specific APIs for zero-copy receive operations to further increase performance.
- Lowest possible latency and maximum throughput for a given network from a single thread.
- Support for a LD_PRELOAD option to transparently switch over to kernel bypass (similar to OpenOnload)

Operational Considerations

- Available as an Aeron Premium feature.

Behind Aeron

Adaptive builds & operates bespoke trading technology solutions across asset classes for financial services firms wanting to own their tech stack to differentiate and compete in the long-term. Central to Adaptive's offering is Aeron, the global standard for high-throughput, low-latency and fault-tolerant trading systems - the open-source technology supported and sponsored by Adaptive.