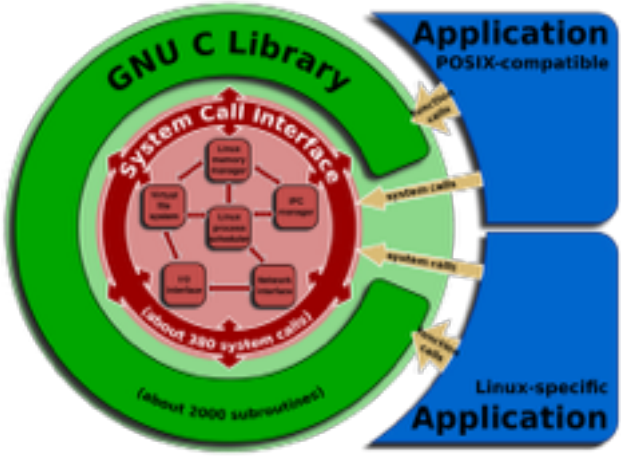
Software control issues of high bitrate data streams

Christoph Lameter, Ph.D.
cl@linux.com
Linux Core Kernel Maintainer for slab allocators and per cpu operations
R&D Architect Algorithmic Trading

Photonics for throughput and latency

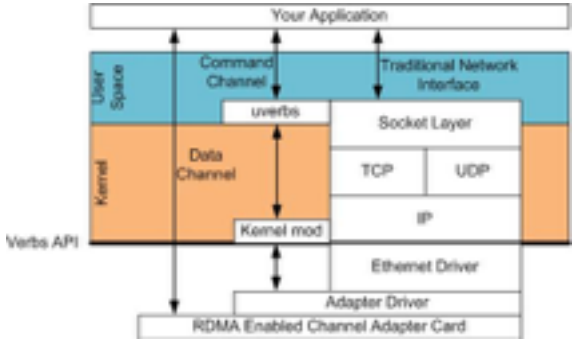
- HPC and AlgoTrading
  - Maximum Throughput
  - Minimal Latency
  - Rapid distribution of information via Multicast
- Our users want “bare metal” performance
  - OS noise issues
  - Reduce size of software
  - SDN? Uhhh….
- What we expect from Photonics
  - Radically lower latency
  - Extremely higher throughput

POSIX / System Calls



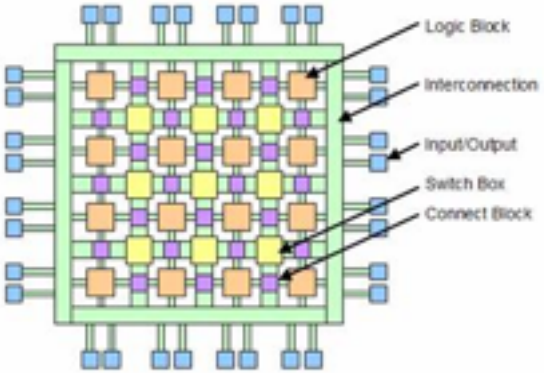
- *recvmsg*(), *sendmsg*() and TCP
- Made for 10Mbps networks
- Works well at 1Gbps.
- Data copied to and from process address space.
- Difficult at 10Gbps. Requires additional measures
  - Flow Steering
  - MultiQueue support
- Mostly kept away from the application but system needs to be tuned correctly.
- Single thread performance is limited. May have to distribute logic for performance reasons.
- Strong jitter (10-100 milliseconds normal, seconds possible)

Open Fabric API (RDMA)



- Direct Remote Memory Access (RDMA)
- Designed for 10Gbps networks.
- Kernel out of the data path. Zero Copy Control via sys calls.
- IBTA cross platform standardization of RDMA API
- Limited by PCI-E and memory speeds
- Works for 40G and 56G speeds. Additional measures
  - Flow control
  - Use cpu caches instead of memory
  - NIC connected to multiple sockets or NUMA nodes.
- Not sure how well this works at 100G and beyond

FPGA



- Works very well at 10Gbps
- Jitter free
- Current State of the Art in Electronic Trading
- Complex and difficult “coding” and debugging.
- Works also well at 40G and 56G.
- 100G problematic with todays FPGAs but newer technology that became available in 2015 will address that.
- Ability to process at line rate. Memory out of the critical path. Processes data while packet is being received.
- Striving to simplify what has to be done at the FPGA level.

Real Hardware

- Resource intensive to develop
- Line rate
- No jitter
- Deterministic
- Electronic Trading typically takes advantage of hardware developed for other industries because of limitations on the number of chips needed.

Controlling a high speed data stream 100G -> 1Tbps?

- Requires a layered approach
- All the presented techniques have their own means of control
- These are already found in most high end devices today
- Also present in latest paper by Harm Dorren on Optical switch (2014).
- The higher the speed of the data stream the more difficult the API is that is required to be used.
- POSIX and RDMA (OpenFabrics Stack) are the only standardized protocols.
- Control of methods are hardware/implementation specific.
- There is a project to standardize interaction with FPGA from the Linu Kernel developers.

OEO Layer 1 switches

- 4ns for passing through a packet. Latency of 1 meter of fiber.
- Packet is not modified
- Signal conversion to electronic, amplification, and replication
- Amplification and moderation of signal is controlled by FPGA
- FPGA is controlled by host processor
- Hots processor allows the use of RDMA APIs or POSIX APIs
- FPGA programming can be controlled from Host processor
- FPGA can shape the signal processing and distribution to multiple endpoints
- Multiple Vendors ExaBlaze, MetaMako etc.