Hello, IX

a protected dataplane operating system for high throughput and low latency

May, 2017

Juhyun Choi (fenderstic@gmail.com)
Index

- Datacenter Invasion
- What Is The Problem and So Why?
- Conventional Wisdom
- IX Design
- Virtualization
- Control Plane and Data Plane
- Evaluation
- Discussion
- Conclusion
Datacentre Invasion

- Leading IT companies such as FANG (Facebook, Amazon, Netflix, Google) provide services by enormous datacentres.
- Topics regarding datacentres are populated in the research area.
- It costs astronomically.
- Networking requirements:
  - High packet rates
  - Microsecond tail latency
  - Protection
  - Resource efficiency
What Is The Problem?

- Software is the main cause of low performance in network
- Hardware is innocent : )
A Datacentre Scenario
So, Why?

- Because we have complex and mismatched design of OS.
  - Support various compatible devices
  - A flood of modules
  - Contention of resources
  - Scheduling and interrupts
Conventional Wisdom

- Bye, sorry, By pass kernel
  - mTCP that causes latency due to batch algorithm
- Alternatives to TCP
  - RDMA that requires reliable manner
- Alternatives to POSIX API
  - MegaPipe
- Tuning OS settings
  - SO_REUSEPORT, Core Affinity
IX Design

- Virtualization
- Control Plane
- Data Plane
Virtualization

- Instructions of x86 are categorized by Ring structure.
- VMM (Virtual Machine Monitor) is a hypervisor for serving physical devices to guest operating systems.
- VMM runs in VMX (Virtual Machine eXtensions) Root Mode which allows OS to execute privileged instructions.
- Guest operating systems run in VMX (Virtual Machine eXtensions) Non-Root mode which allows guest OS to execute privileged instructions by hypervisor.
Control Plane

- Control Plane is a kind of router in SDN (Software Defined Network)
- Centralized resource and polices
- De-centralized events
- State dissemination
- Responsible for resource configuration, provisioning, scheduling, and monitoring in coarse-grained manner
Data Plane

- A dataplane OS : ), library OS, exokernel
- Directly ACCESS to hardware by virtualization
- Exclusive large page
- Intel 82599 chipset VT-x virtualization
- lwIP for TCP/IP
- Batched API leads high locality
- OS Enhancement
- Flow-consistent hashing of incoming traffic to distinct queues
- A Data Plane controls exclusively queues and networking stack.
Back to the future

- Networking is just one of the feature which belongs to kernel, in terms of OS.
- Kernel allocates the resources once socket is created.
- During packet processing, context-switching is often occurred.
- Performance metrics that locality and cache-coherence are decreased.
- One of the biggest enemy is memory copy
  - Memory copy from NIC to kernel
  - Memory copy blah blah in kernel
  - Memory copy from kernel to Application
Back to the future

- The number of batch count is restricted by upper bound, and it guarantees latency.
- Run to completion prevents unnecessary switches and improve cache coherence.
- Dataplane directly access to NIC, so ZERO copy is exposed by API.
mTCP

- Make full use of extreme parallelism & batch processing
- Per-core resource management
- Cache-aware threading
- Eliminate system call overhead
- Reduce context switch cost by event batching
Evaluation for Goodput

- Short message connections are very often created among nodes.
- IX successfully obtains $\frac{1}{2}$ bandwidth at 20 KB message size while other gain under 1/5.
Evaluation for Connection Scalability

- Linux shows poor scalability by increasing number of connections.
- Thanks to core affinity and optimized data plane, IX maximizes the performance.
- mTCP is not invited due to absence of 40Gbps NIC support
Evaluation for Multicore Scalability

- 10Gbps IX needs only 3 cores for entering saturation while mTCP requires 8 cores.
- At 8 cores test, 40Gbps IX gets 40 times higher throughput than Linux.
Conclusions

- Control Plane and Data Plane are successfully separated by virtualization.
- Zero-copy APIs with flow control are exposed by Data Plane.
- Data Plane is directly access to resource of NIC such as tx, rx queues.
- Adaptive batch mechanism achieves high throughput as well as low latency.
- Even if all experiments shows that IX outperforms others, however there is a room for improvement by application optimization.
To Do list by Author

- IOMMU
- SR-IOV
- Controlling elastic threads in dynamic
- Microsecond level latency
- User-friendly API
Any questions?
End! Thank you for your attention : )

is here.