Software Defined Networking for illumos

illumos Day, September 27th 2014
Robert Mustacchi
rm@joyent.com
Software Defined Buzzword

- Set up infrastructure once
- Modify it without connecting to every LOM or serial console in the building
- Extend and create OS abstractions as necessary
Problem Background

- Network Isolation (network multi-tenancy)
- Address space control
- VLANs
  - Annoying to Program
  - Limited Number
- Building blocks for network services
- Avoid magic hardware
Industry Trends - Consensus

- Encapsulation Formats
  - NVGRE
  - VXLAN
- Broadcom Trident II ASICs
- Intel 40 GbE Cards
Industry Trends - Anarchy

• 100 different ways to manage it
  • Openflow
  • VMware Nicira
  • Plumgrid
  • Userland software (Weave, etc.)
  • Hardware (Pluribus, Cumulus, etc.)

• Anarchy in management suggests illumos needs flexibility here
Enter Overlay Devices

• New dladm overlay device
• Created on top of zones L3 netstack (generally)
• Create VNICS and the like on overlay devices
• Two main pluggable components
  • Encapsulation (kernel)
  • Lookup (userland)
High level Architecture

Kernel

* ********

TCP/IP/VND

+------------+  +---------+  +---------+  +--------+  +--------+  +--------+  +--------+

| VNIC |          | Cache | Lookup | Cache | Lookup | Cache | Lookup | Cache | Lookup |

Overlay

+------------+  +---------+  +---------+  +--------+  +--------+  +--------+  +--------+  +--------+  +--------+  +--------+

| Overlay | Device | +----------| Encap | Plugin | Engine | +----------| Encap | Plugin | Engine | +----------| Encap | Plugin | Engine | +----------| Encap | Plugin | Engine | +----------| Encap | Plugin | Engine |

| GZ IP Stack |          | +----------| +----------| +----------| +----------| +----------| +----------| +----------| +----------| +----------|

| GZ (V)NIC, AGGR |          | +----------| +----------| +----------| +----------|

Userland

* ********

Virtual

+------------+  +---------+  +---------+  +--------+  +--------+  +--------+  +--------+

| Daemon | Cache | Cache | Cache | Cache | Cache | Cache | Cache | Cache | Cache |

Virtual Plugin

+------------+  +---------+  +---------+  +--------+  +--------+  +--------+  +--------+  +--------+  +--------+

| TCP/UDP Request | Static Request | Static Request | Static Request | Static Request | Static Request | Static Request | Static Request | Static Request | Static Request |

Async Upcall

* * * * *
Overlay driver

- GLDv3 driver
- Notion of an encapsulation id
- Pluggable encapsulation modules
  - Per-module properties
- Two destination modes
  - Point
  - Dynamic
- vxlan already implemented for this
Virtual ARP Daemon - varpd

- Userland side of an overlay device
- Door server for libdlnmgmtd
- Talks to kernel over an ioctl interface
  - Kernel treats varpd like IP does ARP, expects it can fail
- Implements lookups in pluggable backends
- Support for injecting packets into the devices
  - Useful for proxy ARP, NDP, etc.
  - Dropping packets
varpd backends

• Shared objects that implement an ops vector
• Define lookup mode
• Define supported destination types
• Define properties
• Two current prototype backend:
  • Direct – Point to point
  • Files – Glorified /etc/ethers
- VNIC MTU modification (already in the gate)
- vxlan decoding in snoop
- libidspace – user version of id_space interfaces
- librename – persistent atomic file renames
- re-entrant ethers(3SOCKET) functions
- Direct callbacks for ksockets
- UDP ksocket sendmblk support
- mpt_sas refhash made generic
Futures

- dlmgtmd persistence
- Performance
- Dynamic resize of the kernel target table
- Experimenting with distributed systems as plugin backing stores
- Better observability
  - DTrace, snoop, dumping target table
Take it Home Today

• Code at: https://github.com/joyent/illumos-joyent/tree/dev-overlay

Thanks

- Dan McDonald, Sebastien Roy, and Rich Lowe for enduring lots of annoying questions and design discussions
- Joyent, especially to those who slogged through the long design documents and discussions
- OmniTI
- illumos community