

# open



USE



IMPROVE



EVANGELIZE

開  
放  
的  
열린  
مفتوح  
libre  
मुक्त  
ಮುಕ್ತ  
livre  
libero  
ముక్త  
开放的  
açık  
open  
nyílt  
ᄇᄇᄇᄇ  
πᄇᄇ  
オープン  
livre  
ανοικτό  
offen  
otevřený  
öppen  
открытый  
வெளிப்படை

## xVM Server Networking

David Edmondson

[dme@sun.com](mailto:dme@sun.com)

Solaris Engineering



# Overview

- Challenges
- The Crossbow project
- xVM Server Networking
- Status
- Issues
- Further work



# Challenges

Shared access to resources is the only feasible mechanism to support networking

Ensuring that services are available, reliable and secure requires that we constrain virtual machines' use of the network:

- control access to physical resources
- control access to network services
- limit bandwidth use
- remove or limit the ability to damage other hosts
- remove or limit the ability to damage other virtual machines

Oh, and do all of this whilst providing great performance.



# Challenges: Implementation

Difficult to associate activity with “billable” entities:

- protocol processing in interrupt context
- anonymous packet processing in the kernel

Difficult to segregate traffic:

- common packet queues

Performance suffers:

- extra processing to ensure fairness, resource control, etc.

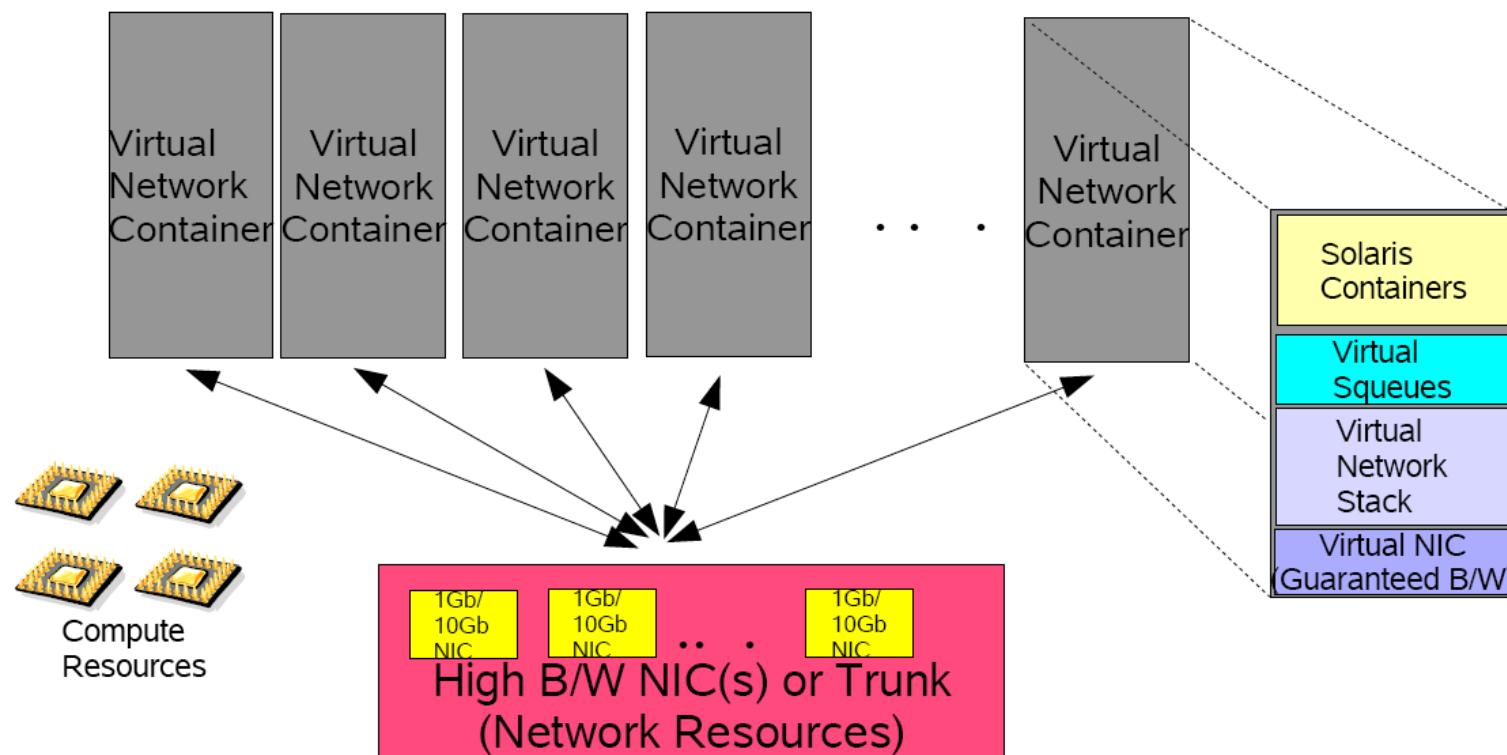


# Crossbow

An OpenSolaris project to improve network virtualisation:

- partition NIC memory, DMA channels, etc. into multiple “Virtual NICs”
- use a flow classifier to build a virtual stack on each VNIC
- independently switch individual VNICs between interrupt and polling mode
- control the rate of packet arrival for a VNIC independently of all others
- support a variety of hardware implementations

# Crossbow Architecture



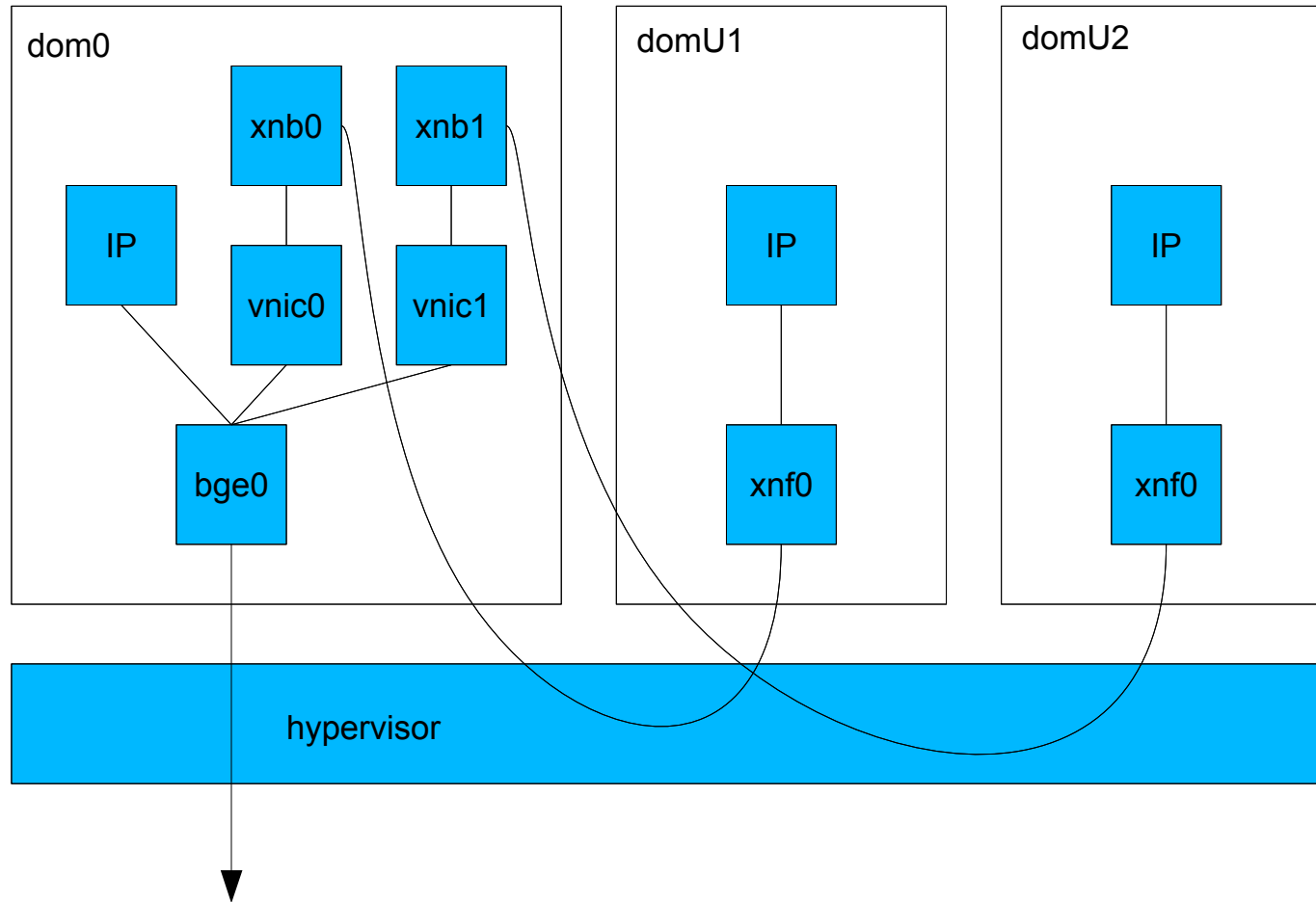


# Crossbow and the Network Backend

Provide network access to guest domains via a VNIC

- guest domain traffic is segregated from that of other domains
- hardware traffic classification

# xVM Server Network Backend





# Status

## OpenSolaris build 75a:

- VNIC infrastructure in place
- Network backend is VNIC based
- Multiple MAC addresses per physical NIC support

## Crossbow Project:

- Resource management and lots more hardware offload in place
- Currently merging with build 77
- OpenSolaris integration CQ1 2008



## Status: Performance

Performance characterisation under way with:

- ~2Ghz Opteron, four cores, 4G RAM
- No tuning:
  - dom0 uses all cores
  - domU uses all cores
- 1514 byte MTU, no TSO/LSO or jumbo frames
- hypervisor copy in dom0 -> domU path
- “back to back” connection with target system
- OpenSolaris build 77 + ongoing fixes



## Status: Performance with 1Gbit

1Gbit Broadcom 5704 derivative:

- Line rate transmit and receive for domU, single connection or multiple connections
- CPU overhead is high (up to 160% of metal)
- Latency is higher



# Status: Performance 10Gbit

## 10Gbit Sun Neptune:

- ~2Gbits/sec transmit/receive for domU, single connection
- Multiple connection throughput is lower
  - Lock contention in backend driver over ring access
- Driver issues related to holding buffers too long (causing transmit ring exhaustion)

Early days for this analysis, but already **lots** of scope for simple improvements.



# Issues

- Linux vs. Solaris architectural shear:
  - TSO, checksum offload, writable mappings, ...
- Lack of MSI to spread interrupt load in dom0
- Lack of parallelism due to single dom0 <-> domU channel
- No way to express ethernet multicast group interest
  
- Parallelism focus has always been for layer 3+



## Further Work (1)

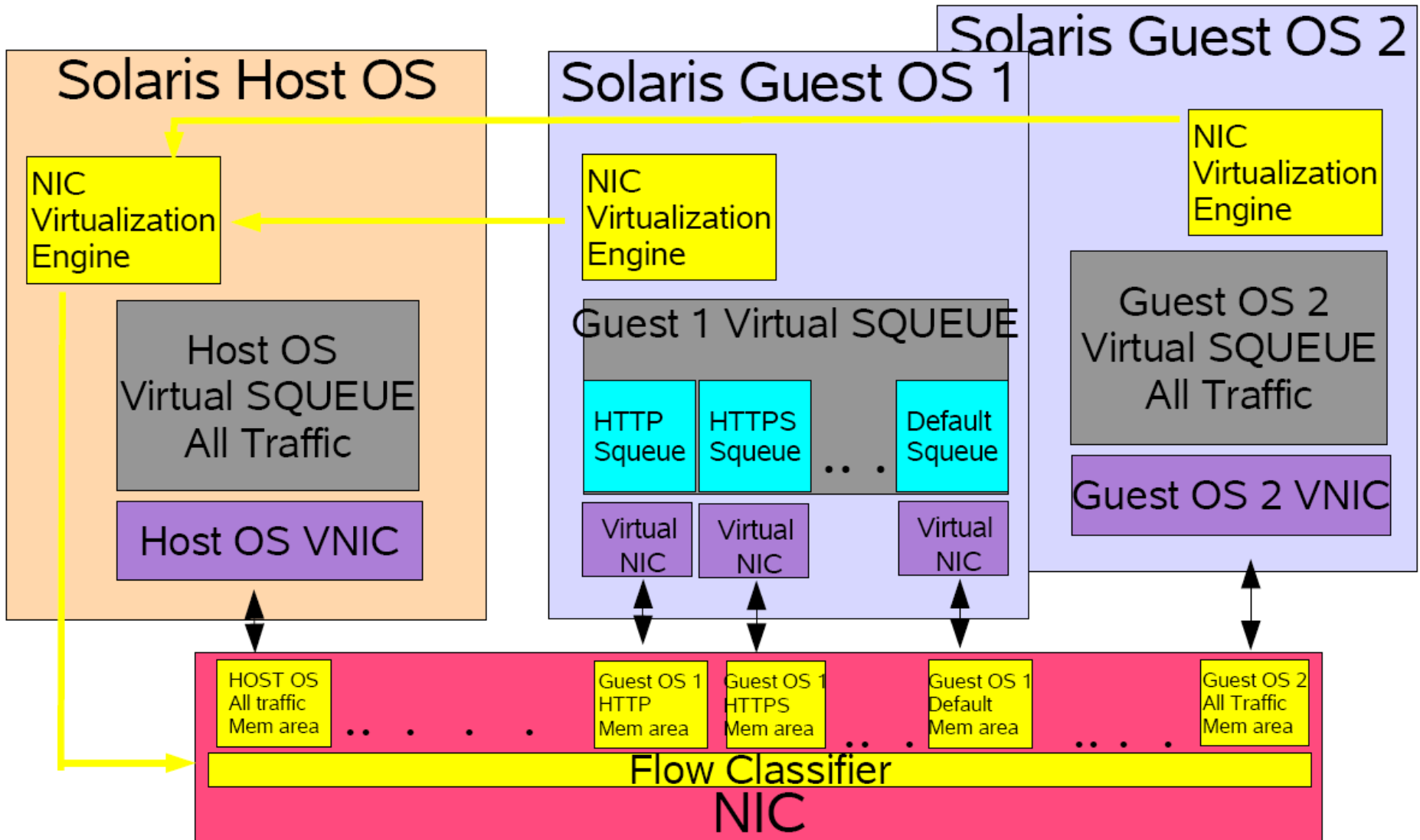
- 10Gbit analysis is only just getting started
  - look at other NICs
  - 16 core performance testing begun
- Implement the NetChannel2 proposals and provide feedback
- Decide whether TSO/LSO is worth the bother
  - Get to ~3Gbit/sec without, then use IOV for anything more?



## Further Work (2)

- Direct domU to domU path rather than via backend
  - distributed switching across a physical machine (LDomS has this today)
  - distributed switching across many machines (multiple ingress/egress points, ...)
- PCI IOV
  - maybe Neptune pre-IOV approach
  - runs contrary to some requirements (e.g. security), though some NIC implementations will provide options
- Support for more complex network topologies
  - finish our “routed” and “NAT” glue
  - libvirt virtual networks

# Solaris, Crossbow and PCI IOV





# Finding out more

- OpenSolaris xVM Server community
  - [xen-discuss@opensolaris.org](mailto:xen-discuss@opensolaris.org)
  - <http://opensolaris.org/os/community/xen>
  - <http://openxvm.org>
  - <irc://irc.oftc.net/solaris-xen>
- Crossbow community
  - [crossbow-discuss@opensolaris.org](mailto:crossbow-discuss@opensolaris.org)
  - <http://opensolaris.org/os/community/crossbow>

# open



USE



IMPROVE



EVANGELIZE

## Thank you.

David Edmondson  
Solaris Engineering  
dme@sun.com  
<http://dme.org>

“open” artwork and icons by chandan:  
<http://blogs.sun.com/chandan>

開  
放  
的  
열린  
مفتوح  
libre  
मुक्त  
ಮುಕ್ತ  
livre  
libero  
ముక్త  
开放的  
açık  
open  
nyílt  
:::  
πικρ  
オープン  
livre  
ανοικτό  
offen  
otevřený  
öppen  
открытый  
வெளிப்படை