Presentation of Diagnosing performance overheads in the Xen virtual machine environment

Jesper Kristensen

September 26, 2005

Outline

Statistical Profiling with OProfile

- Collecting statistical data from the system.
  - Performance monitoring via hardware.
  - Hardware counters.
  - TLB miss, cache miss, L2 miss.
- OProfile
  - Profile code on any privilege level of executing.
  - OS notifications with NMI (Non-maskable Interrupt).
  - Program counter (CPU register), the programs call stack.
  - Resource consumption.

OProfile and Xen

- System-wide profiling?
- Profiling of a single domain in Xen?
- Selecting a set of domains?
- Handling of NMI interrupts?

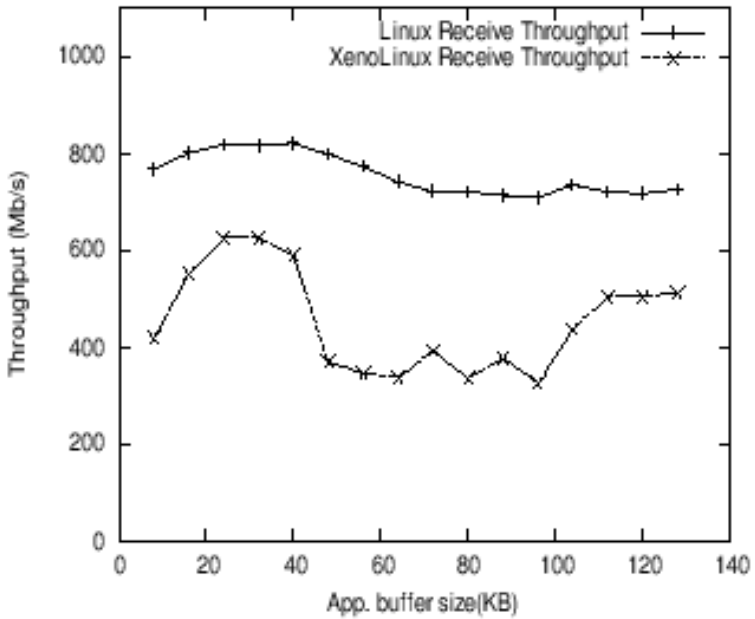
OProfile and Xen

- System-wide profiling?
- Profiling of a single domain in Xen?
- Selecting a set of domains?
- Handling of NMI interrupts?

Xenoprof

Outline

Example of Strange Behavior in Xen



What is Xenoprof

- System-wide profiling toolkit for Xen.
- Uses hardware performance monitoring.
- Determine the distribution of performance events.
- A Virtual Machine Monitor (VMM) layer in Xen.
  - Profiling through Hypercalls
  - Samples through events

Outline

Design Problems

- Centralized profiling not an option.
  - No centralized information (One kernel).
  - More the one domain running (More than one kernel).
- Coordination of profilers.
  - Each domain runs its own modified OProfiler.
  - Communication between domain level Profilers and Xenoprof.
  - Statistical distributed data.

Xenoprof Framework Interface

1. Performance event interface.
2. Register interrupts and sample buffers.
   - Virtual interrupts (event channels).
   - Xenoprof collects program counters.
   - Per-domain sample buffer.
3. Activation and deactivation of profiling:
   - for a set of domains (hypercall).
   - for a single domain (no coordination).

Xenoprof Framework Interface

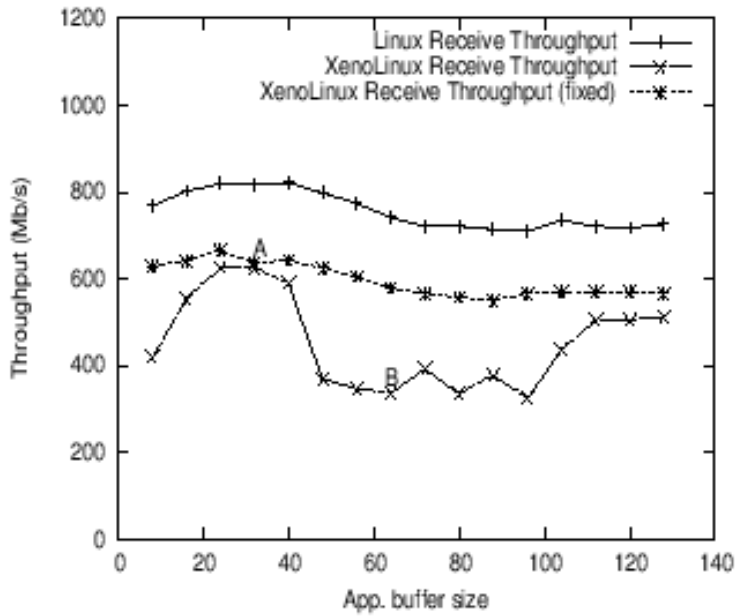
1. Performance event interface.
2. Register interrupts and sample buffers.
   - Virtual interrupts (event channels).
   - Xenoprof collects program counters.
   - Per-domain sample buffer.
3. Activation and deactivation of profiling:
   - for a set of domains (hypercall).
   - for a single domain (no coordination).

All domain level profilers most be ported to the Xenoprof interface.

Outline

Xen network performance anomaly

Network performance anomalies in Xen virtual machines network throughput.



Performance at Points A and B

	Point A	Point B
XenoLinux Kernel	60	**84**
Network driver	10	5
Xen	30	11

Distribution of execution cost in point A and B

Function Breakdown at Points A and B

	Point A	Point B
skb_copy_bits	0.15	**28**
skbuff_ctor	absent	**9**
tcp_collapse	0.05	**3**
other_routines	99.8	60

Function Breakdown at Points A and B

	Point A	Point B
skb_copy_bits	0.15	**28**
skbuff_ctor	absent	**9**
tcp_collapse	0.05	**3**
other_routines	99.8	60

Roughly 40% more execution time spent in kernel routines.

Why is `skb_copy_bits` used so many times

- Simple data copying function (`skb_copy_bits`).
- XenoLinux routine to zero out pages (`skbuff_ctor`).
- Collapses the tcp socket's memory (`tcp_collapse`).
- Internal buffer fragmentation.
- Normal buffer size is equal to Maximum Transfer Unit (MTU).
- One page per received packet (4 KB/1500 Bytes).

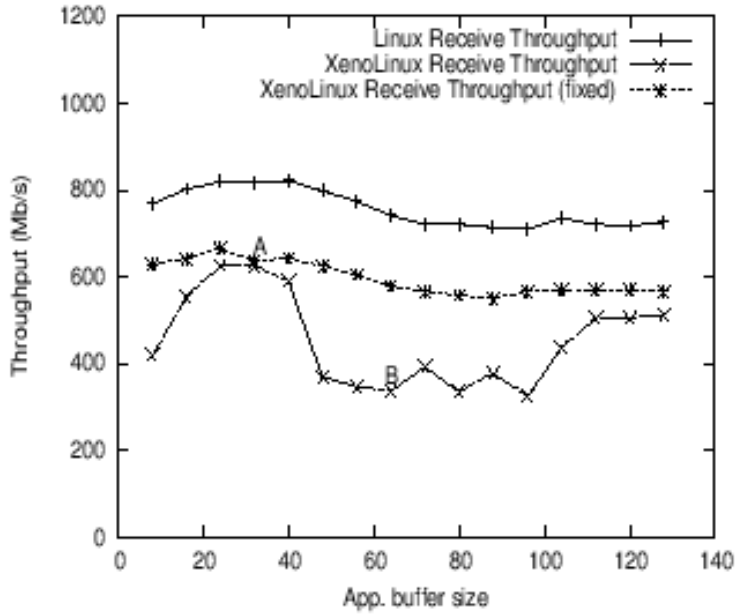
Why is `skb_copy_bits` used so many times

- Simple data copying function (`skb_copy_bits`).
- XenoLinux routine to zero out pages (`skbuff_ctor`).
- Collapses the tcp socket's memory (`tcp_collapse`).
- Internal buffer fragmentation.
- Normal buffer size is equal to Maximum Transfer Unit (MTU).
- One page per received packet (4 KB/1500 Bytes).

Which lead to internal fragmentation 3x1500 = 4500 bytes.

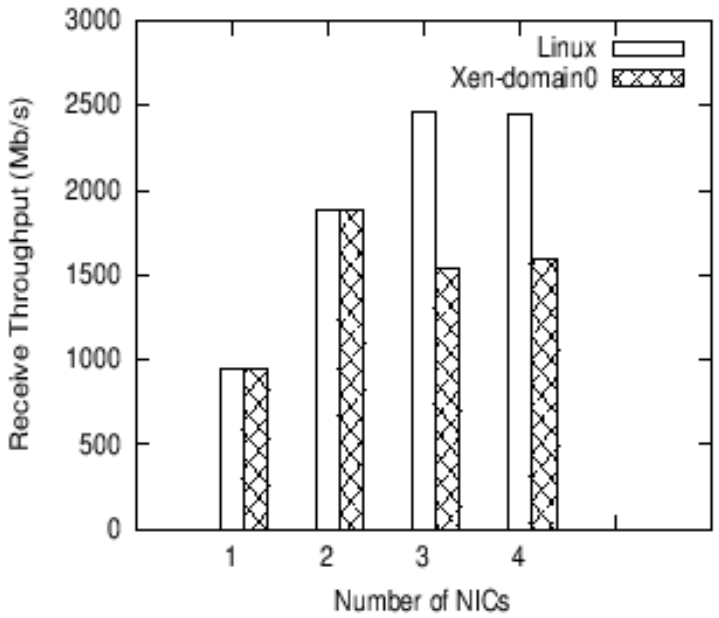
Xen Network Performance Anomaly

Network performance in Xen virtual machine with kernel parameter `tcp_adv_window_scale`.



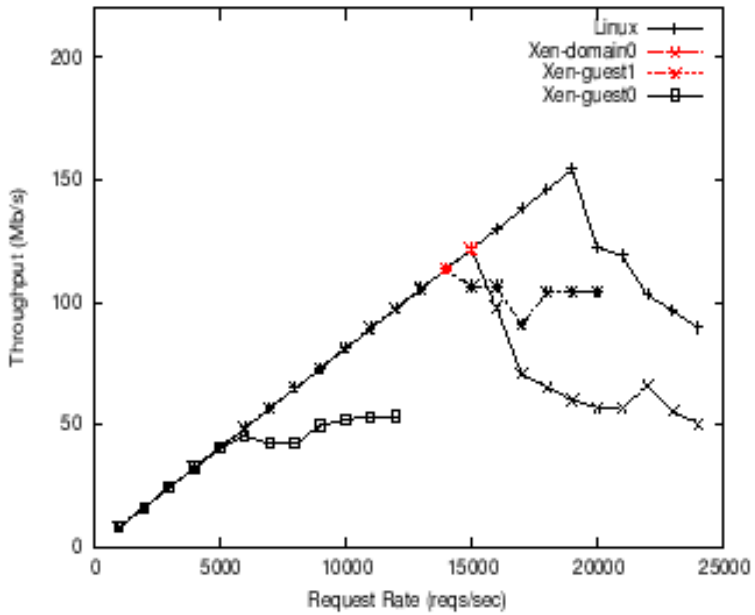
Outline

Receive throughput for 4 NIC



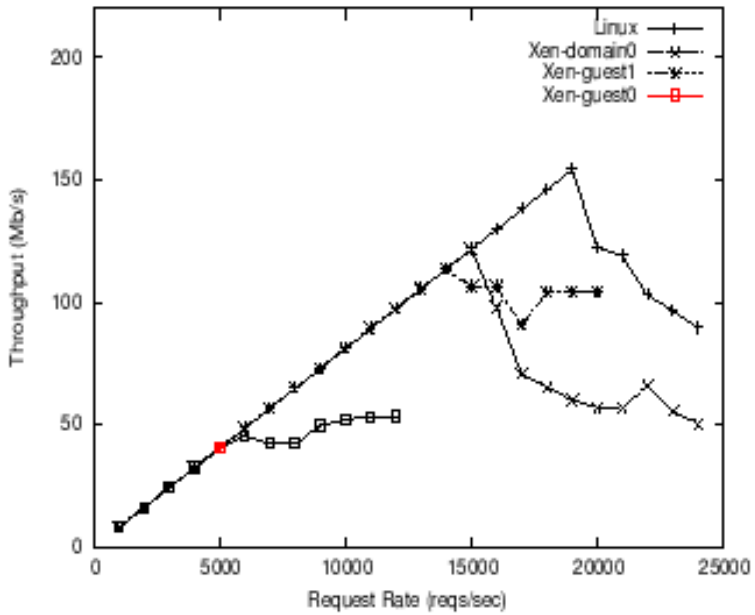
The throughput is roughly 75% of the throughput in Linux and this is because a overhead in virtual interrupts.

Web-server performance



The throughput is less than 80% of the throughput in Linux.

Web-server performance



The throughput is only around 34% of the throughput in Linux.

Outline

Questions

Questions?