Receive Descriptor Recycling
for
Small Packet High Speed Ethernet Traffic

IEEE Region 8 Student Paper Contest 2006

Cedric Walravens
student
K.U.Leuven, Belgium

16th May 2006
Contents

• Introduction to Ethernet and the Problem of Small Packet High Speed Traffic
  – Throughput Measurements

• Identification of cause by :
  – PCI/PCI-X Analysis
  – Kernel Profiling

• Proposed solution :
  Receive Descriptor Recycling

• Conclusions, Applications and Further Work
IEEE 802.3 Ethernet

MELECON 2006, Málaga

The diagram shows a network topology with a switched hub connecting multiple devices (A, B, C, D, E). The header of each packet is depicted with variable length from 46B to 1500B. The payload follows the header, and there is a gap after each packet. The network uses Ethernet packets for data transmission.
Raw Throughput = \frac{pay + hdr}{pay + hdr + gap} \times \text{Bandwidth (1 Gbps)}
Throughput Measurements

\[ \text{Throughput} \quad \text{[Gbps]} \]

\[ \frac{p + 18}{p + 18 + 20} \]

\[ \text{Payload} \ p \ [\text{bytes}] \]
Throughput Measurements

Throughput [Gbps]

Payload $p$ [bytes]

send

receive

1
Small Packet Traffic

Small packet = Small payload
(46…500 byte)

- SNMP, ICMP
- DHCP, DNS, NTP
- SSH, Telnet, FTP commands
- IRC, IM (e.g. MSN, ICQ, …)
- VoIP, RDP
- HTTP, SMTP

Some more frequently than others, but almost every protocol uses small packets at some point!
Setup for Throughput Measurements

MELECON 2006, Málaga

Network Processor
IBM PowerNP

Gigabit Ethernet
w/ CSMA/CD disabled
(cat 5e copper UTP)

Device Under Test
Dell SC1425 w/ Intel Pro/1000MT
Small Packet Performance

(b) Quad card, two ports simultaneously, plotted with theoretical Gigabit Ethernet limit.
Small Packet Performance

(b) Quad card, two ports simultaneously, plotted with theoretical Gigabit Ethernet limit.
Contents

• Introduction to Ethernet and the *Problem* of Small Packet High Speed Traffic
  – Throughput Measurements

• *Identification of cause* by:
  – PCI/PCI-X Analysis
  – Kernel Profiling

• Proposed *solution*:
  Receive Descriptor Recycling

• Conclusions, Applications and Further Work
Setup for PCI/PCI-X Analysis

Device Under Test
Dell SC1425 w/ Intel Pro/1000MT

Gigabit Ethernet
w/ CSMA/CD disabled
(cat 5e copper UTP)

Network Processor
IBM PowerNP

Packet
Flood

Analyzer
Receive Descriptors

Memory

PCI bus

Intel e1000
Linux driver

NIC
Receive Descriptors

Memory

RD Transfer

NIC

RD’s

Buffer of Receive Descriptors (RD’s)

driver
Receive Descriptors
Receive Descriptors

Memory

NIC

driver

Incoming packets
Receive Descriptors

MELECON 2006, Málaga

Memory

Received packets

driver

Incoming packets

NIC
Receive Descriptors
Receive Descriptors

Memory

Received packets

NIC

Incoming packets

driver

And so on…
Receive Descriptors

Memory

Received packets

…until all RD’s are used

driver

……

Incoming packets

NIC
Receive Descriptors

Memory

Received packets

driver

Fetch new RD’s from driver

NIC

RDT of new RD’s

Incoming packets
Setup for PCI/PCI-X Analysis

**Network Processor**
IBM PowerNP

**Device Under Test**
Dell SC1425 w/ Intel Pro/1000MT

**Gigabit Ethernet**
w/ CSMA/CD disabled
(cat 5e copper UTP)

**Analyzer**
### PCI/PCI-X Analysis

#### Continuous flood of packets

<table>
<thead>
<tr>
<th>Signal Name</th>
<th>Value (X)</th>
<th>T</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Gap</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addr[64]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data[64]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYCLE</td>
<td>WAIT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FRAME#</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRDY#</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEVSEL#</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRDY#</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STOP#</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**KATHOLIEKE UNIVERSITEIT LEUVEN**

**MELECON 2006, Málaga**

---

23
But the bus is not to blame

(a) Dual card, one port, plotted with theoretical Gigabit Ethernet limit, PCI bandwidth and slope.
**PCI/PCI-X Analysis**

**Continuous flood of packets**

<table>
<thead>
<tr>
<th>Signal Name</th>
<th>Value (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock</td>
<td></td>
</tr>
<tr>
<td>Time Gap</td>
<td></td>
</tr>
<tr>
<td>Addr[64]</td>
<td></td>
</tr>
<tr>
<td>Data[64]</td>
<td></td>
</tr>
<tr>
<td>CYCLE</td>
<td><strong>WAIT</strong></td>
</tr>
<tr>
<td>FRAME#</td>
<td>0</td>
</tr>
<tr>
<td>IRDY#</td>
<td>1</td>
</tr>
<tr>
<td>DEVSEL#</td>
<td>1</td>
</tr>
<tr>
<td>TRDY#</td>
<td>1</td>
</tr>
<tr>
<td>STOP#</td>
<td>1</td>
</tr>
</tbody>
</table>
received packet transferred to memory

descriptor written back to driver
new RDT from driver to NIC

RTD of new RD's

Memory

Received packets

driver

NIC

Incoming packets
Contents

• Introduction to Ethernet and the Problem of Small Packet High Speed Traffic  
   – Throughput Measurements

• Identification of cause by:
   – PCI/PCI-X Analysis
   – Kernel Profiling

• Proposed solution:
  Receive Descriptor Recycling

• Conclusions, Applications and Further Work
Setup for Kernel Profiling

**MELECON 2006, Málaga**

**Network Processor**
IBM PowerNP

**Gigabit Ethernet**
w/ CSMA/CD disabled
(cat 5e copper UTP)

**Device Under Test**
Dell SC1425 w/ Intel Pro/1000MT

---

Running Profile

---
Memory Management is to blame!
Bad Memory Management

- Allocate memory space for RD
- Initialize RD
- Provide RD to NIC

- Process the returned RD (& associated packet)
- Release memory space of RD
- Allocate (other) memory for RD
- Initialize RD
- Provide RD to NIC
Receive Descriptor Recycling

- Allocate memory space for RD
- Initialize RD
- Provide RD to NIC

- Process the returned RD (& associated packet)
  - *Reset* memory space of RD

- Provide RD to NIC
Receive Descriptor Recycling

NP->SRV06 - Intel dual card, two ports to NP - 2.6.12
Receive Descriptor Recycling

NP->SRV06 - Intel dual card, two ports to NP - 2.6.12

+ 40 %
Conclusions

• A complete study of both soft- and hardware performance of Gigabit Ethernet networking

• Low level hardware measurements for small packet traffic performance

• A solution by means of the presented Receive Descriptor Recycling (RDR) concept

→ 40% performance increase in terms of throughput and reduced packet loss
CERN – An International Collaboration

20 Member States and 8 Observers (UNESCO, EU, Israel, India, Turkey, USA, Japan, Russia)

- 2500 staff
- 6500 users
- 500 Fellows and associates
Acknowledgements

• K.U.Leuven – ESAT

• CERN Summer Student Committee and the LHCb team

• IEEE Region 8