



# High Performance Communication Using a Commodity Network for Cluster Systems

S. Sumimoto, H. Tezuka, A. Hori,  
H. Harada, T. Takahashi, Y. Ishikawa

Real World Computing Partnership, JAPAN

HPDC-9

Real World Computing Partnership

RWC



## Cluster Systems on a Commodity Network

- Using a Commodity network (100BASE-T Ethernet, ATM) and commodity software (MPI on TCP/IP)
  - Easy to develop using an existing network and cheap
  - But far slower than parallel computers
- Characteristics of a Commodity Network
  - Many vendors provide many kinds of equipments.
  - The lifetime of hardware is relatively short because of vendor competition (cost and performance).

HPDC-9

Real World Computing Partnership

RWC



## Gigabit Ethernet

- A Next Generation Commodity Network
  - 125MB/s full duplex link speed
  - Getting cheaper and cheaper
- TCP/IP bandwidth is only 46.7MB/s
  - Pentium III 500MHz PC, Packet Engines G-NIC II
- Communication facility to achieve higher communication performance is needed.

HPDC-9

Real World Computing Partnership

RWC



## Outline

- Requirements and Approach of High Performance Communication on a Commodity Network
- Communication Processing Scheme Based on an Overhead Analysis of Existing Protocol Processing
- Implementation of PM/Ethernet on Gigabit Ethernet
- Evaluation

HPDC-9

Real World Computing Partnership

RWC



## Requirements of High Performance Communication on Commodity Network

- High Communication Performance
  - Low Latency and High Bandwidth
- Applicability to Many Kinds of Network Interface Card (NIC)
  - Easy to install to existing LAN environment
- Co-existence of a Cluster Communication with Existing Protocols (ex. TCP/IP)
  - Existing protocols are also used on commodity network.

HPDC-9

Real World Computing Partnership



## Approach and Design Objectives

- Hardware Independent Approach
  - Using existing Ethernet device drivers with no modification
  - Modifying kernel code, but minimizing changes
- Design Objectives: Reliable communication protocol
  - Where communication protocol should be processed?
  - Which protocol should be used?
- Analyzing TCP/IP Protocol Processing Cost
  - Designing communication scheme

HPDC-9

Real World Computing Partnership

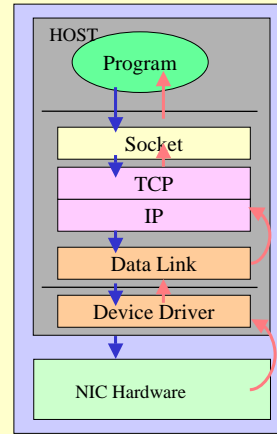




## Performance on Protocol Layers

- Communication Performance on Protocol Layers (Linux 2.2.12)
  - Pentium III 500MHz PC
  - Packet Engines G-NIC II Gigabit Ethernet NIC

	1/2 Round Trip	Bandwidth
TCP/IP	44.8 usec	46.7 MB/s
Data Link	18.8 usec	90.4 MB/s



 Transmit  
 Receive

HPDC-9

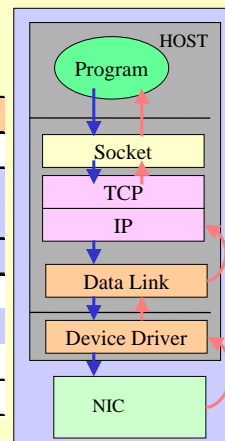
Real World Computing Partnership

RWC



## TCP/IP Protocol Overhead Analysis on Linux 2.2.12

Processing	Overhead	Rate %
System call and socket	1.6 usec	3.6
TCP	15.5 usec	34.6
IP	6.2 usec	13.8
Protocol Handler Invocation	3.2 usec	7.1
Device Driver	4.7 usec	10.5
Hardware Interrupt	5.9 usec	13.2
NIC+Media	7.7 usec	17.2
Total (1/2 Round Trip)	44.8 usec	100



 Transmit  
 Receive

- TCP/IP: 48.4%, Hardware Interrupt: 13.2 %  
 Protocol Handler Invocation: 7.1%

HPDC-9

Real World Computing Partnership

RWC



## Communication Handling Scheme

- TCP/IP Protocol Processing Overhead
  - GigaE PM network protocol (ICS-99): providing reliable high performance communication based on Go-Back N protocol
- Protocol Handler Invocation Overhead
  - Protocol processing on Data Link Layer
- Hardware Interrupt Overhead
  - **Interrupt Reaping technique**

HPDC-9

Real World Computing Partnership

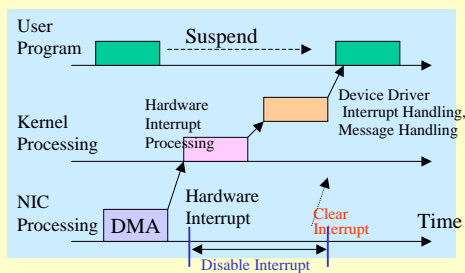
RWC



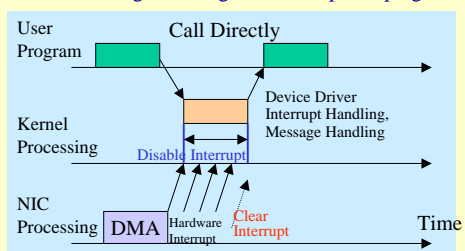
## Proposal of Interrupt Reaping Technique

- Existing Protocols
  - Protocol processing is triggered by hardware interrupt.
- Interrupt Reaping Technique
  - Protocol processing is triggered by the user program with disable interrupt.

Receiving a Message on an Existing Protocol



Receiving a Message on Interrupt Reaping



HPDC-9

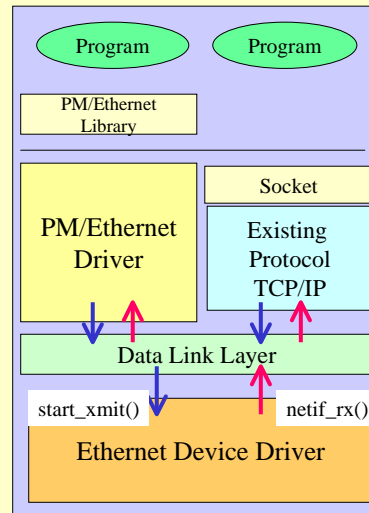
Real World Computing Partnership

RWC



## PM/Ethernet Implementation

- Implemented on Linux with minimal kernel modification
  - Using existing device driver interface
  - No modification to existing Ethernet device driver
  - GigaE PM network protocol is implemented on the PM/Ethernet device driver
- Modification to the Linux kernel
  - Ethernet frame handler on Data Link Layer
  - Registering interrupt handler information



HPDC-9

Real World Computing Partnership

RWC



## Evaluation of PM/Ethernet

- Comparison of PM/Ethernet, TCP/IP on Gigabit Ethernet and PM/Myrinet performance
- Evaluation Environment
  - 16 node cluster (Pentium III 500MHz)
  - Network Hardware:
    - NIC: G-NIC II, Myrinet on 32 bit 33MHz PCI Slot
    - Switch: 3Com 9300 for Gigabit Ethernet, OCT Switch for Myrinet
- Benchmarks
  - Application Level bandwidth, Round Trip Time
  - NAS Parallel Benchmark ver 2.3
    - PM/Ethernet, PM/Myrinet: MPICH/SCore
    - TCP/IP: MPI-LAM

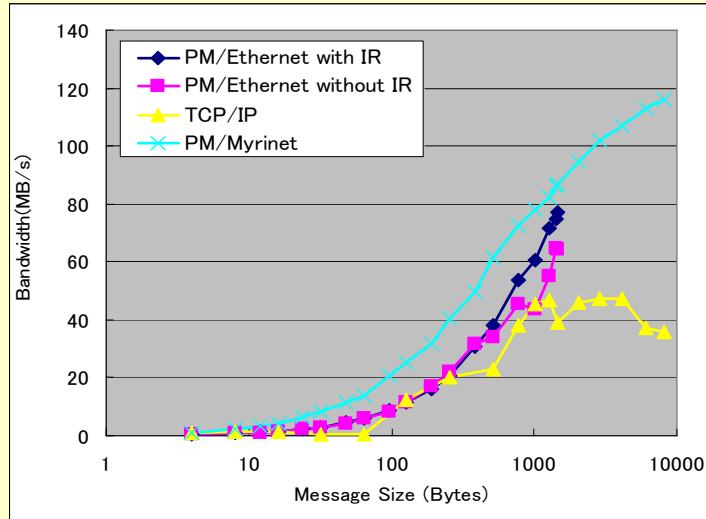
HPDC-9

Real World Computing Partnership

RWC



## Application Level Bandwidth



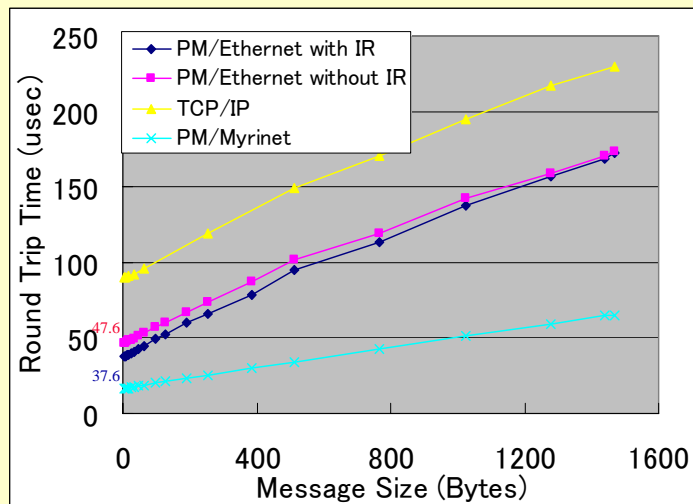
without Gigabit Ethernet Switch  
Real World Computing Partnership

HPDC-9

RWC



## Application Level Round Trip Time



without Gigabit Ethernet Switch  
Real World Computing Partnership

HPDC-9

RWC



## PM/Ethernet vs. TCP/IP on G-NIC II

Processing	TCP/IP	PM/Ethernet
System call and Socket	1.6 usec	1.6 usec
TCP	15.5 usec	-
IP	6.2 usec	-
GigaE PM Protocol	-	4.8 usec
Protocol Handler Invocation	3.2 usec	-
Device Driver	4.7 usec	4.7 usec
Hardware Interrupt	5.9 usec	-
NIC+Media	7.7 usec	7.7 usec
Total (1/2 round trip)	44.8 usec	18.8 usec

HPDC-9

Real World Computing Partnership

RWC



## PM/Ethernet on Other Platforms

	RTT w/ IR	RTT w/o IR	BW w/ IR	BW w/o IR
Alpha 21264	44.6	49.8	96.7	82.5
G-NIC II	usec	usec	MB/s	MB/s
Pentium III	70.1	108	71.0	71.2
Intel E1000	usec	usec	MB/s	MB/s
Pentium III	106.4	128.2	11.9	11.8
EEPRO 100	usec	usec	MB/s	MB/s

- Other tested NICs:
  - Gigabit Ethernet: Syskonnect SK-9843-SX, Lanced LD-1000/SX, Planex GN-1000SX, Alteon ACENIC
  - 100BaseT: Tulip(2124x), 3Com 908B, VIA chipset NIC

No modification to the Ethernet device drivers

HPDC-9

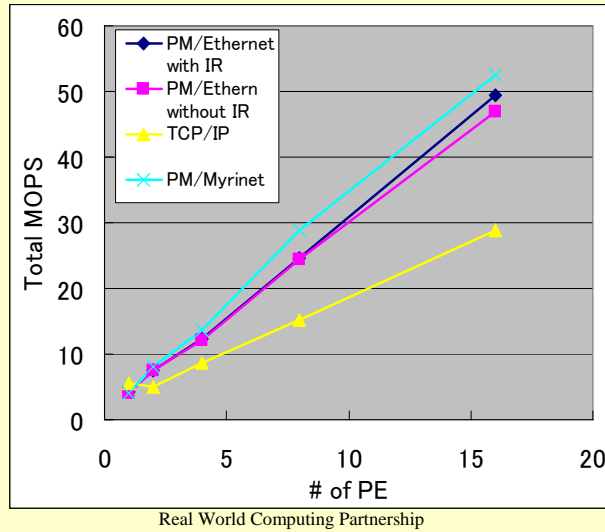
Real World Computing Partnership

RWC



## Application Performance

- NAS Parallel Benchmarks IS Class A



HPDC-9

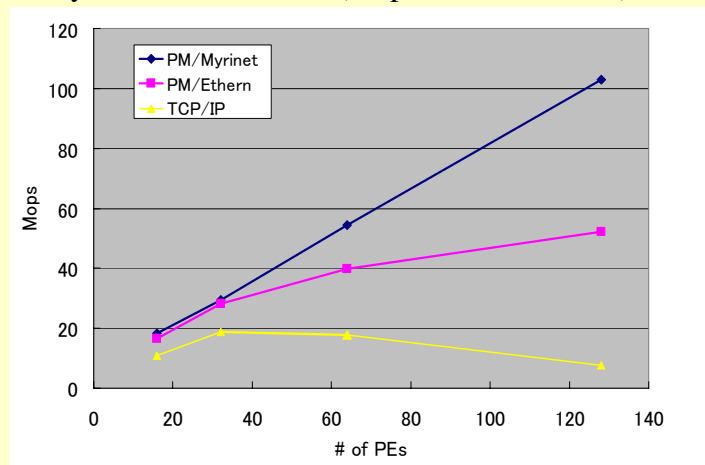
Real World Computing Partnership

RWC



## NPB IS (Class B) on PCC2

- 128 node Pentium PRO 200MHz Cluster  
– Myrinet and 100BaseT(tulip, 3Com 3900 x 4)



HPDC-9

Real World Computing Partnership

RWC



## Communication Facility on a Commodity Network

- U-NET: ATM and Fast Ethernet (tulip)
  - User level communication
    - Eliminating system call and hardware interrupt overhead.
  - Hardware depended implementation
- GAMMA: Fast Ethernet and Gigabit Ethernet
  - Kernel level communication
  - Hardware depended implementation:
    - Modification to Kernel and device driver
  - No reliable communication protocols: Program fails when a frame is lost.

HPDC-9

Real World Computing Partnership

RWC



## Summary

- High Performance Communication Processing Scheme Using a Commodity Network
  - Hardware independent approach
  - Protocol processing on Data Link Layer
  - Interrupt Reaping technique
- PM/Ethernet: an Instance of the Scheme
  - Implemented on Linux: working on different NICs, Intel and Alpha.
  - Communication Performance on G-NIC II:  
Bandwidth 77.6MB/s, RTT 37.6 usec
  - Application Performance: NPB IS Class A  
1.75 times faster than TCP/IP, 7.8% Slower than Myrinet
- URL: <http://www.rwcp.or.jp/lab/pdslab>

HPDC-9

Real World Computing Partnership

RWC