Performance Evaluation of RDMA over IP:
A Case Study with the Ammasso Gigabit Ethernet NIC

H.-W. Jin, S. Narravula, G. Brown,
K. Vaidyanathan, P. Balaji, and D.K. Panda

Network-Based Computing Laboratory
Department of Computer Science and Engineering
The Ohio State University

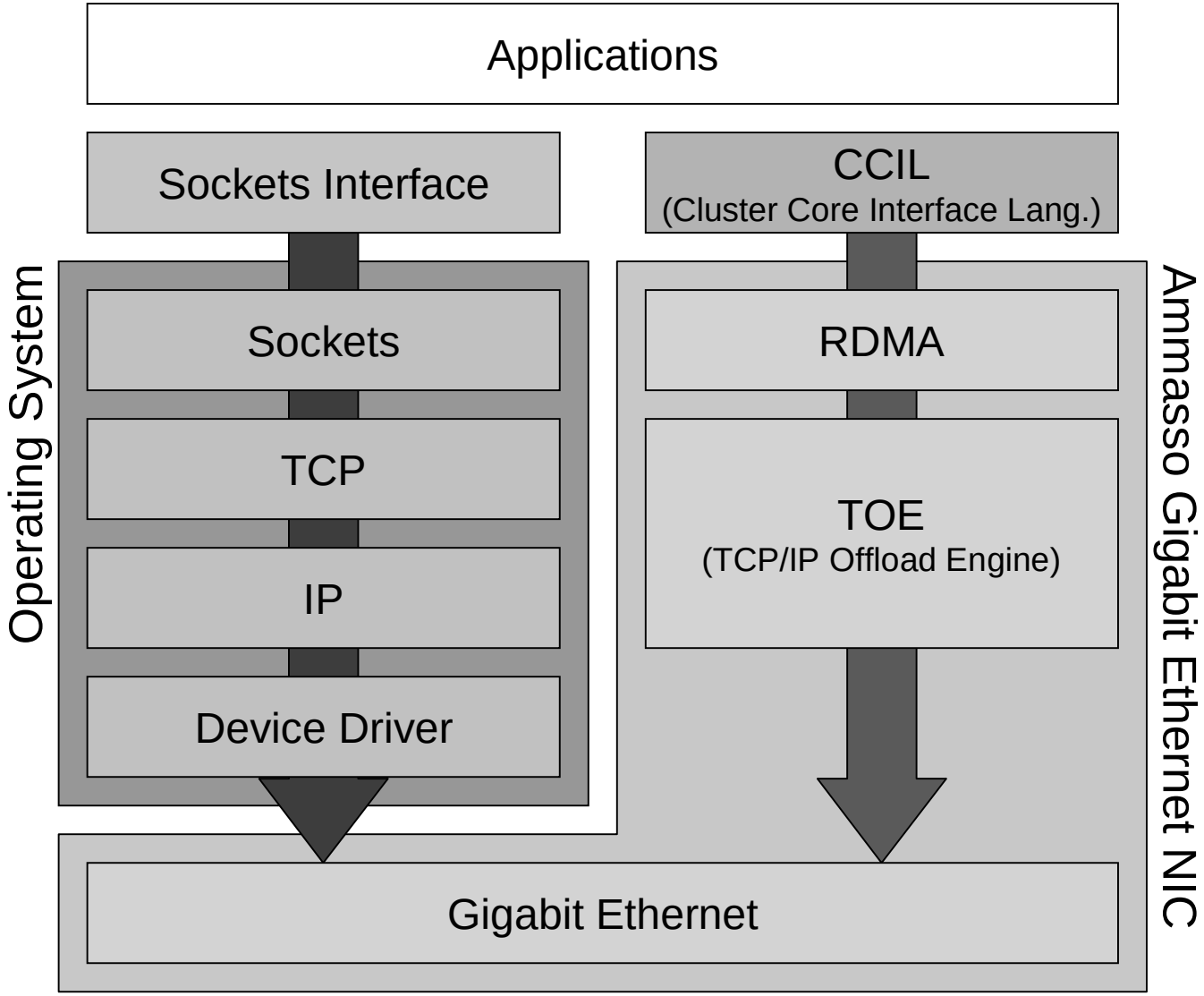
{ jinhy, narravul, browngre, vaidyana, balaji, panda}@cse.ohio-state.edu

Contents

Introduction

- Sockets over TCP/IP
- RDMA over LAN
  - InfiniBand, Myrinet, Quadrics
  - HPC middleware (MPI) and file systems (PVFS)
- RDMA over WAN
  - iWARP, RDDP
  - Grid and Internet applications
- RDMA-enabled Gigabit Ethernet NIC
  - Ammasso

Ammasso Gigabit Ethernet NIC

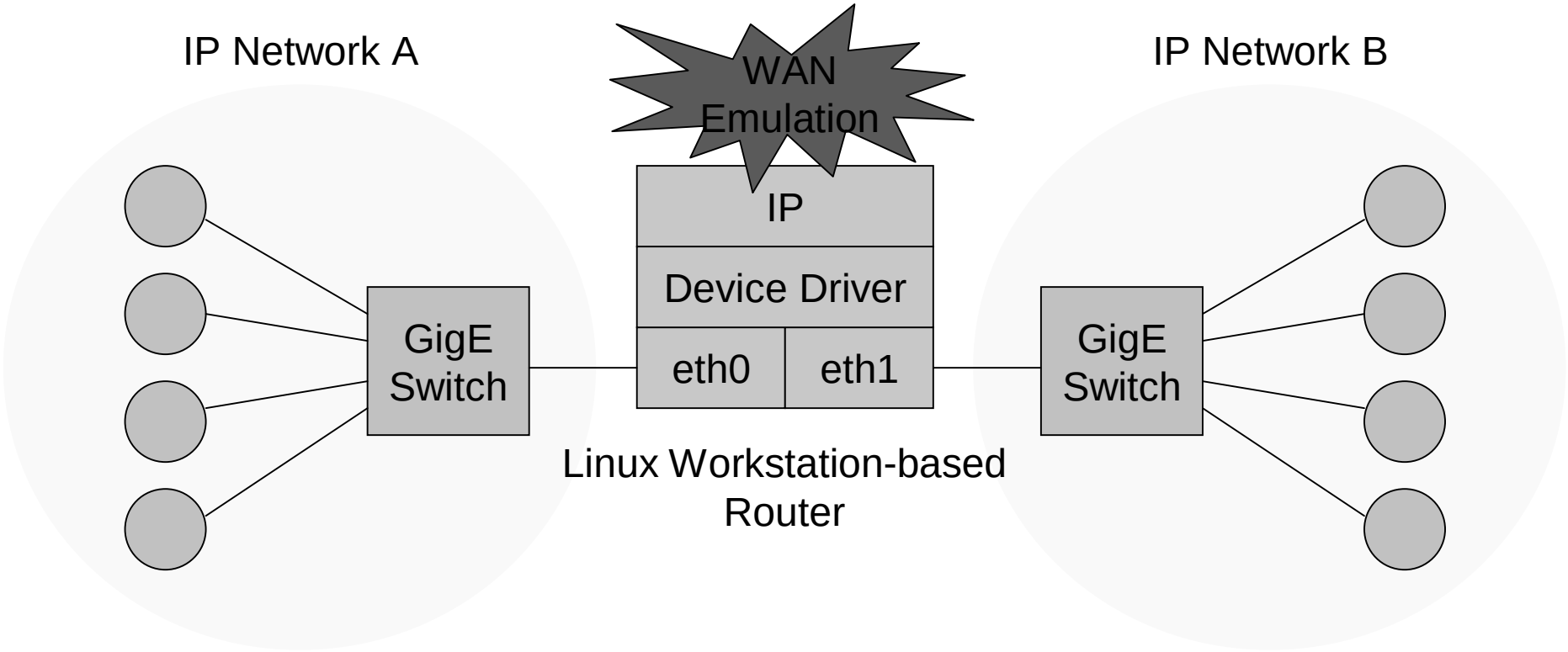


Problem Statement

- There have been no comprehensive quantitative evaluations of RDMA over WAN environment
- How to Emulate the WAN Environment?
- What Kind of Performance Metrics?
- Sockets vs. CCIL

Contents

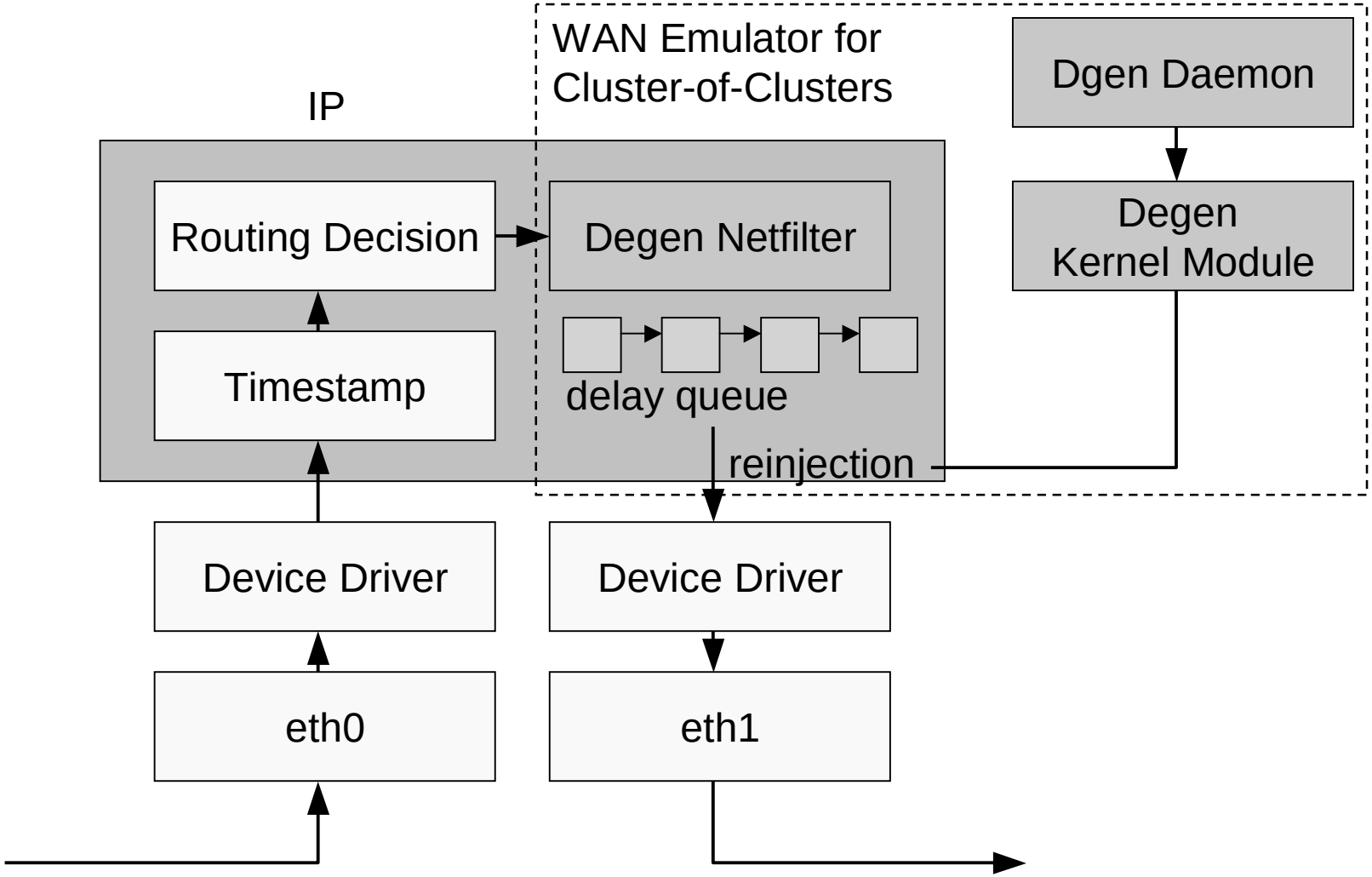
Experimental WAN Setup



WAN Emulator for Cluster-of-Clusters

- Characteristics of WAN Environments
  - High network delay
  - Packet loss
  - Etc.
- User-Level or Kernel-Level Emulator?
- Blocking or Queueing based Delay Adding?

Degen: Delay generator



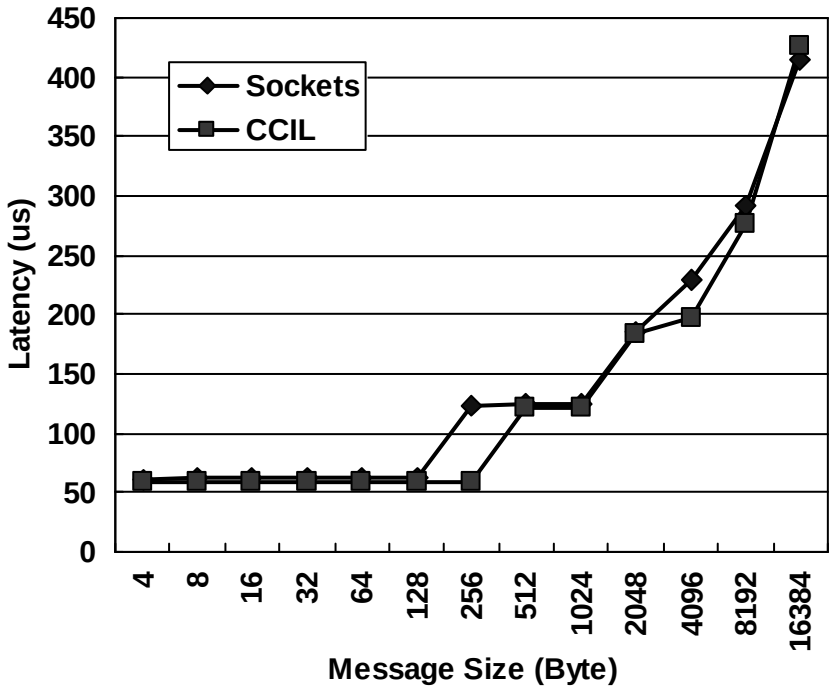
Kernel Patch for CCIL WAN Communication

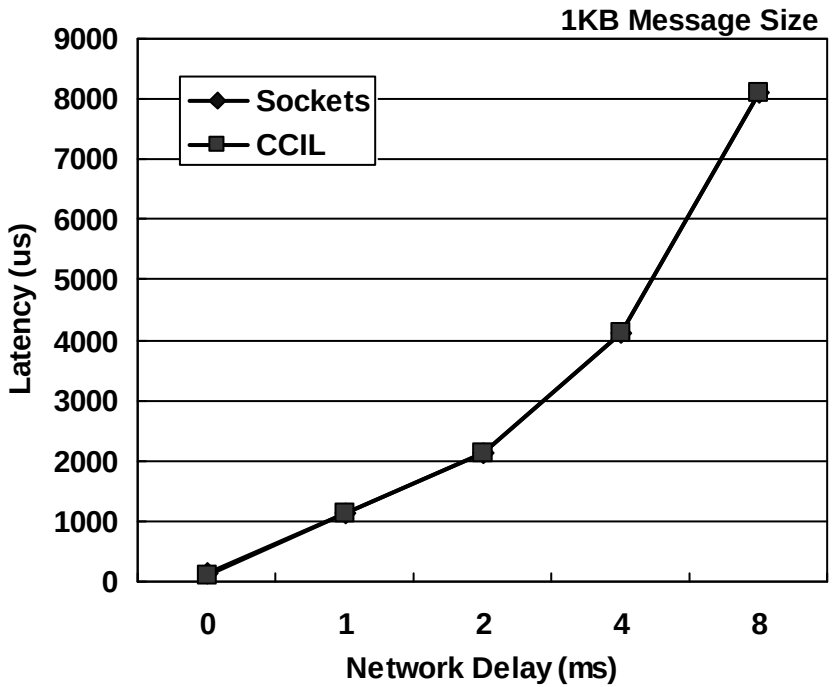
- Ammasso Setup
  - Ammasso 1100
  - Ammasso software version amso1100-1.2-ga2
- Packet Drops for CCIL WAN Communication
  - Timeout
  - Retransmission
- Kernel Patch on Router

Contents

- Introduction
- WAN Emulator for Cluster-of-Clusters
- Performance Evaluation of RDMA over IP
  - Basic communication latency
  - Computation and communication overlap
  - Communication progress
  - CPU resource requirements
  - Unification of communication interface
  - Bandwidth (throughput)
- Conclusions and Future Work

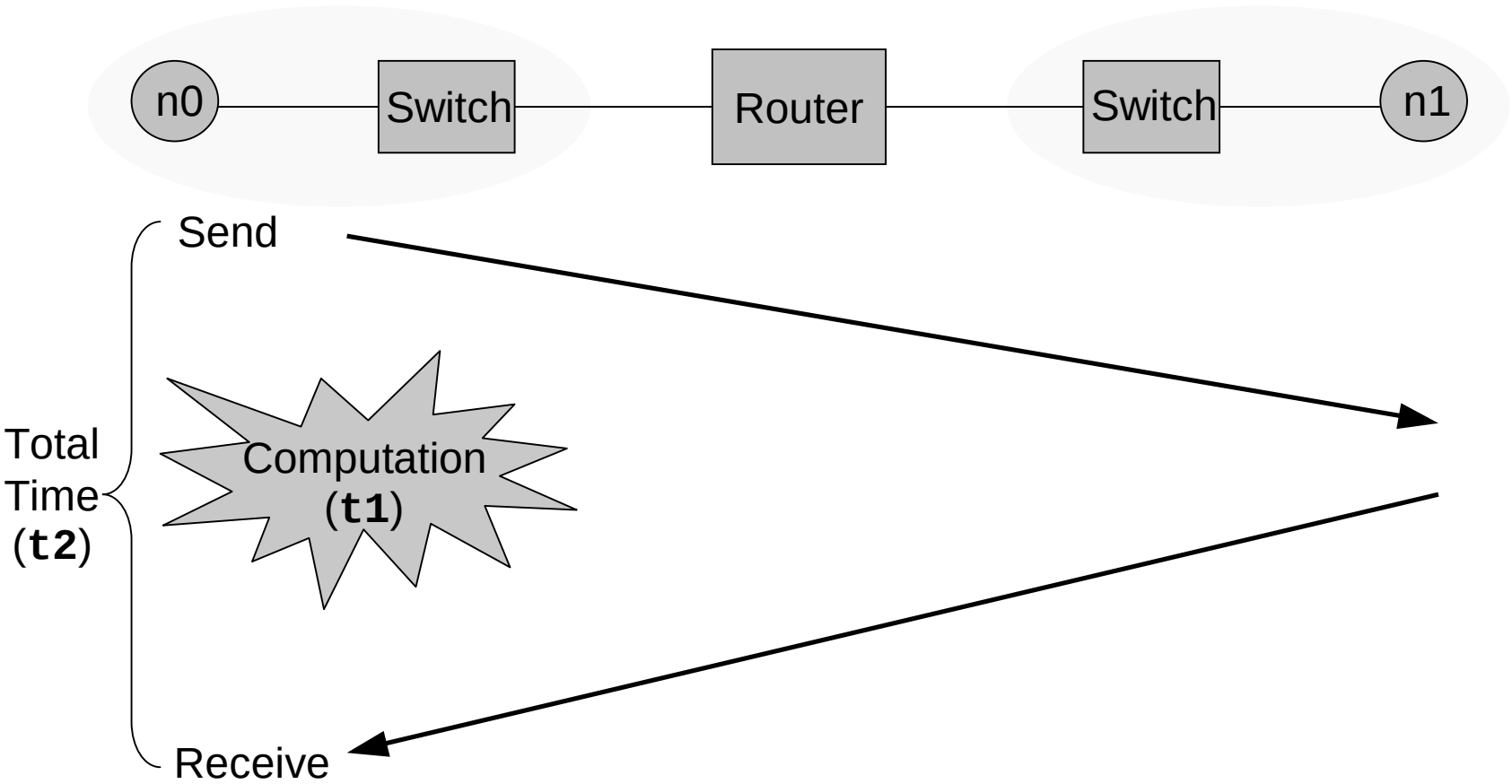
Basic Communication Latency





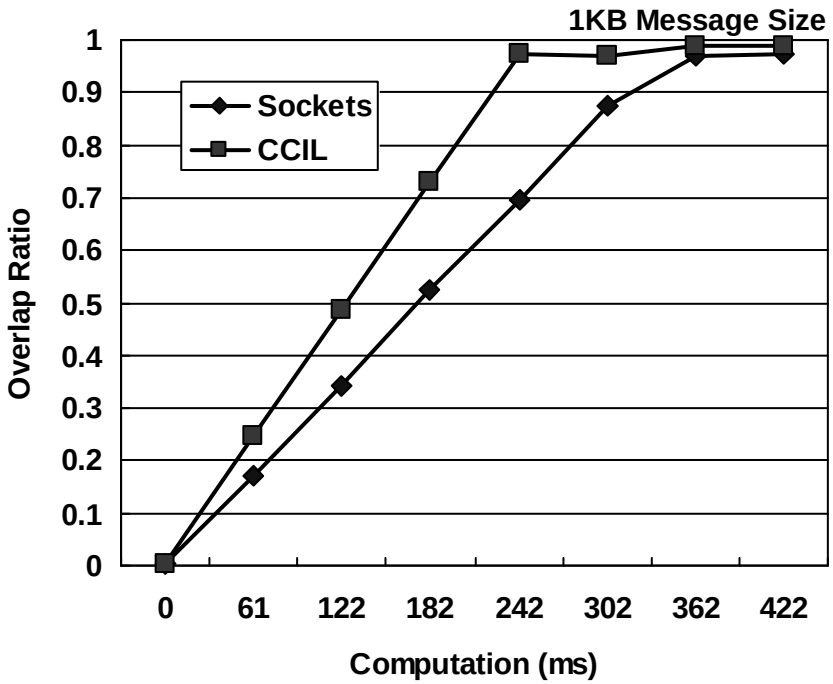
- No impact of zero-copy on the basic communication latency
- Basic communication is not an important metric

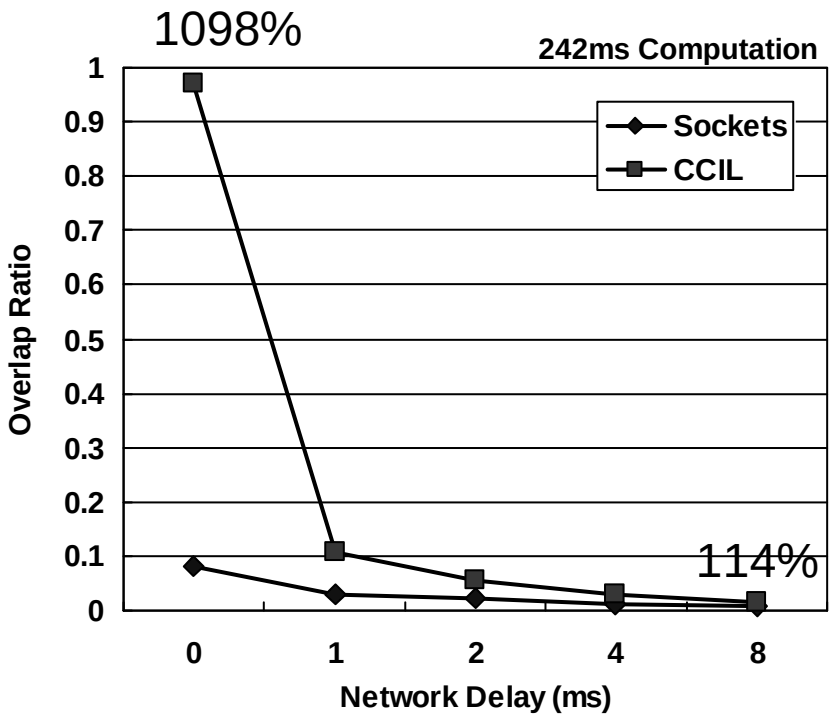
Computation and Communication Overlap



Overlap Ratio = **t1**/ **t2**

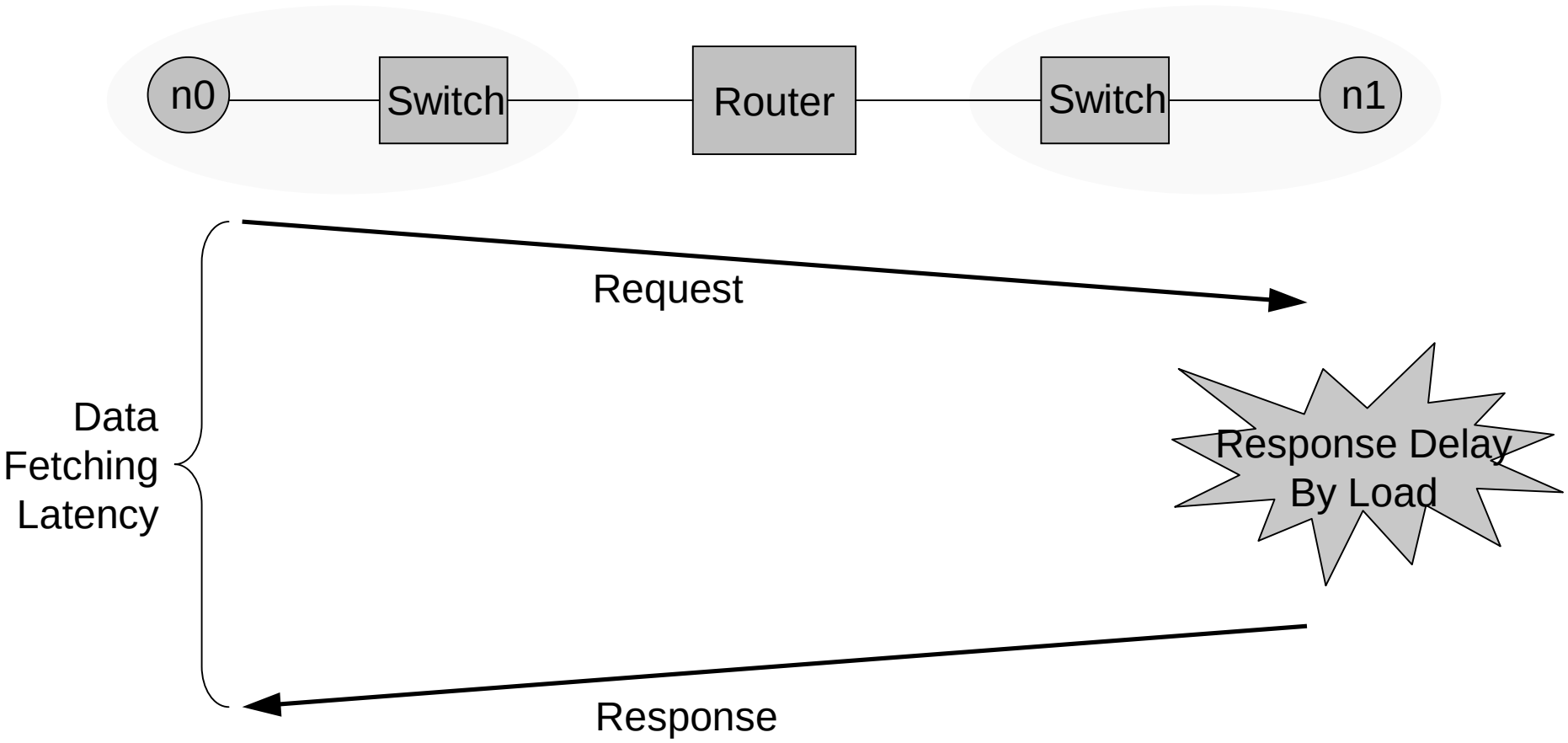
Computation and Communication Overlap



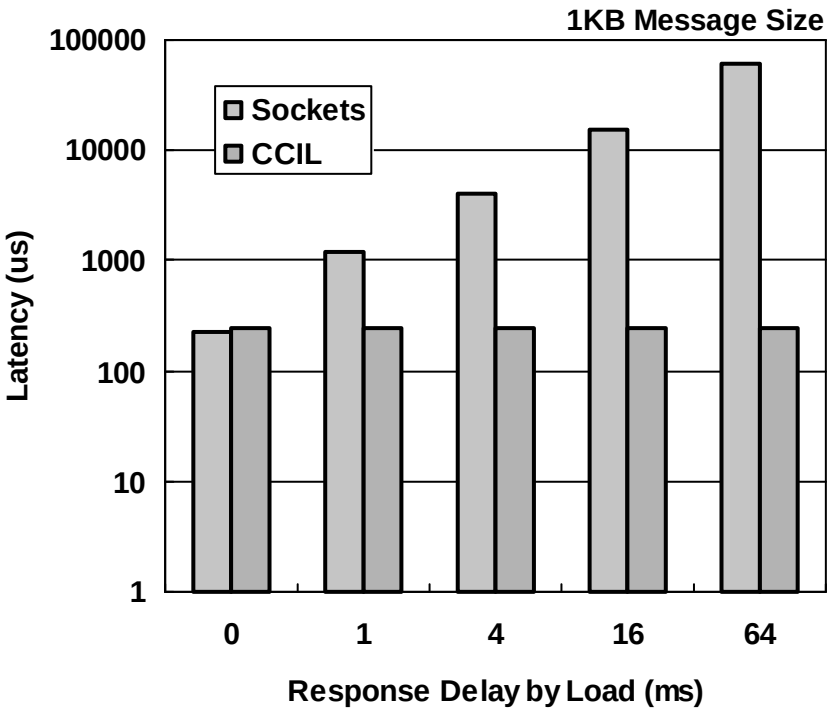


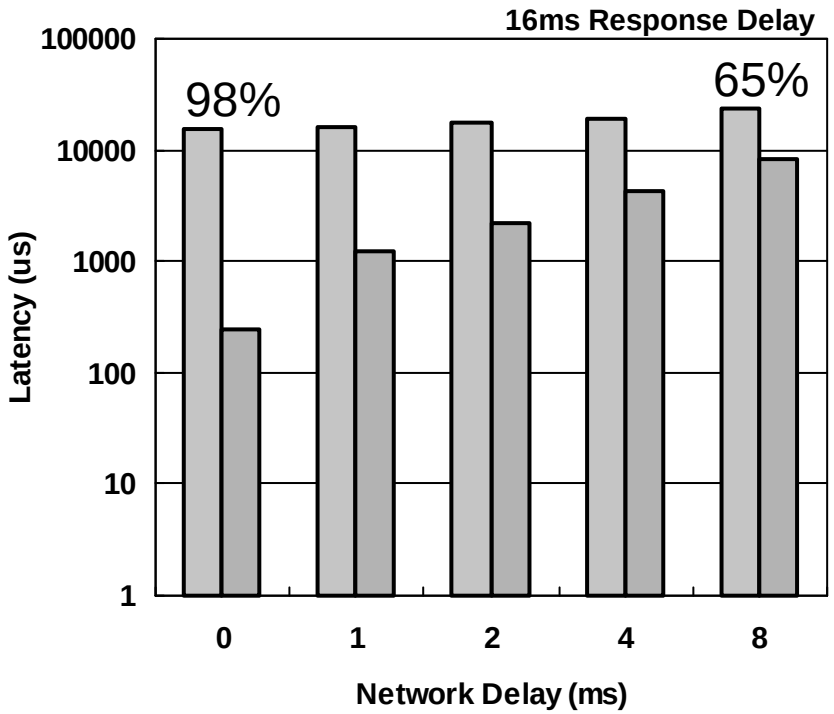
- RDMA can achieve a better computation and communication overlap
- Its benefit reduces as the network delay increases

Communication Progress



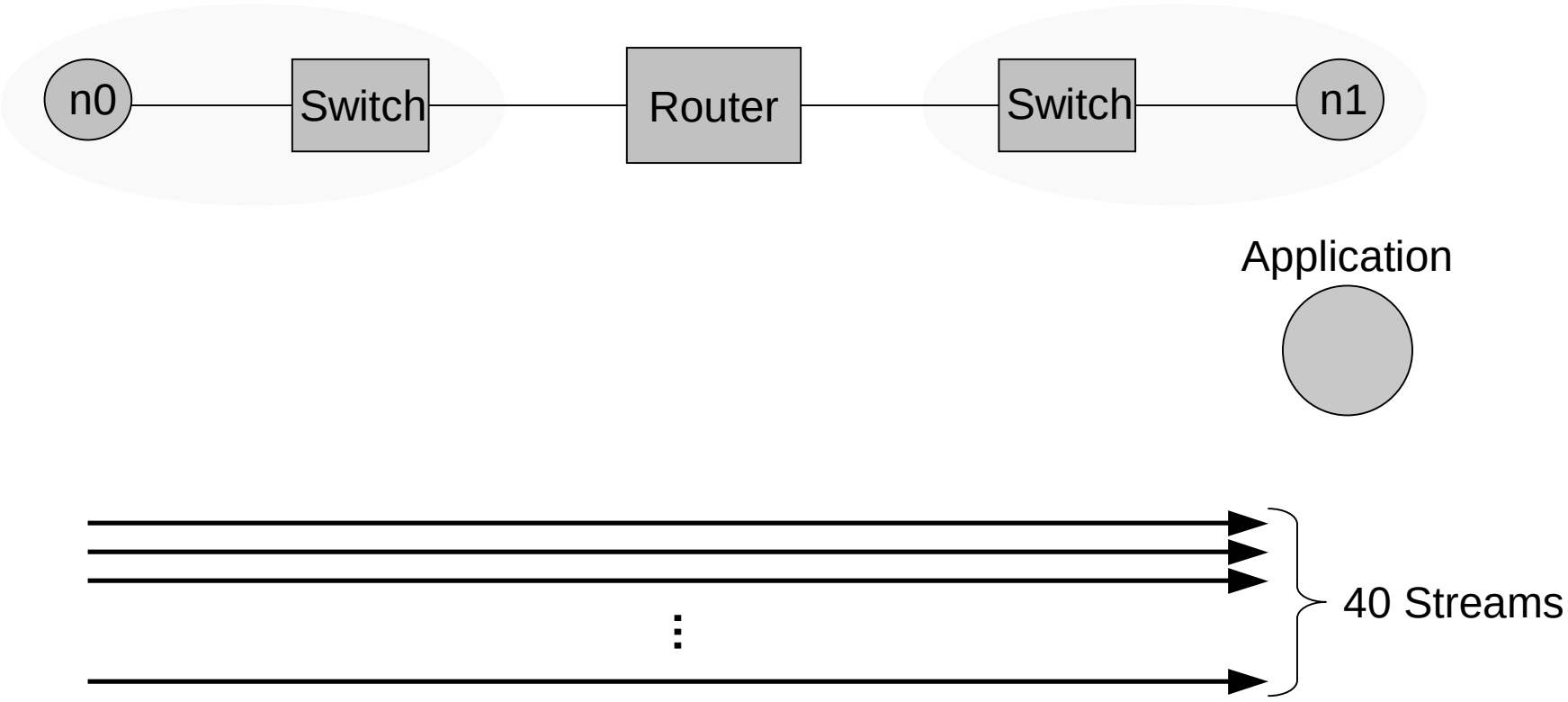
Communication Progress





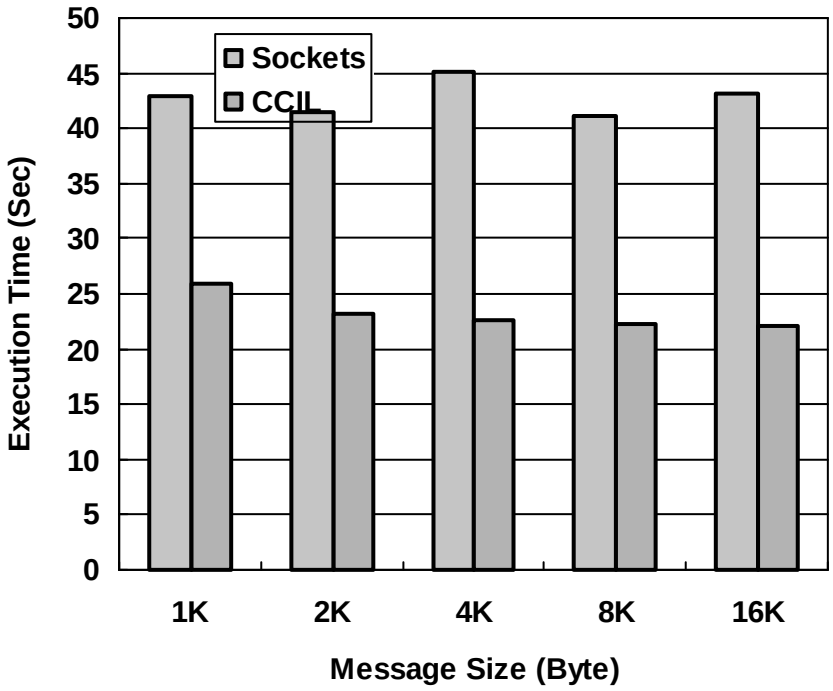
- RDMA can achieve a better communication progress
- Its benefit reduces as the network delay increases

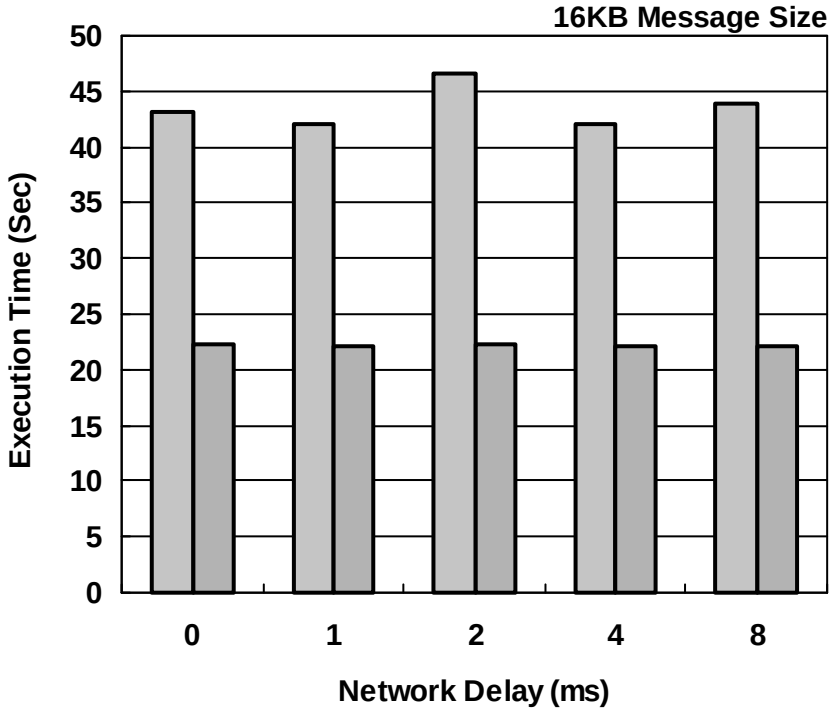
CPU Resource Requirements



Application Execution Time?

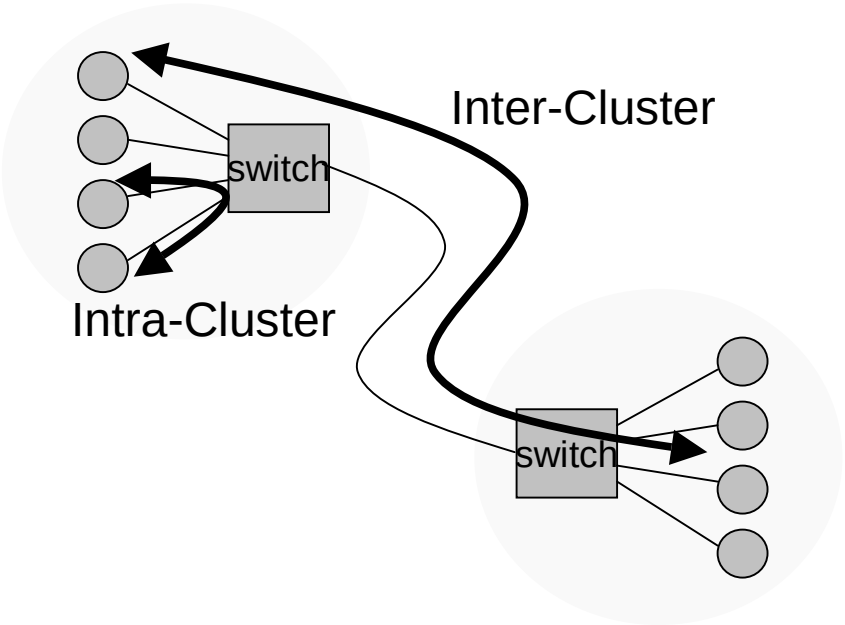
CPU Resource Requirements

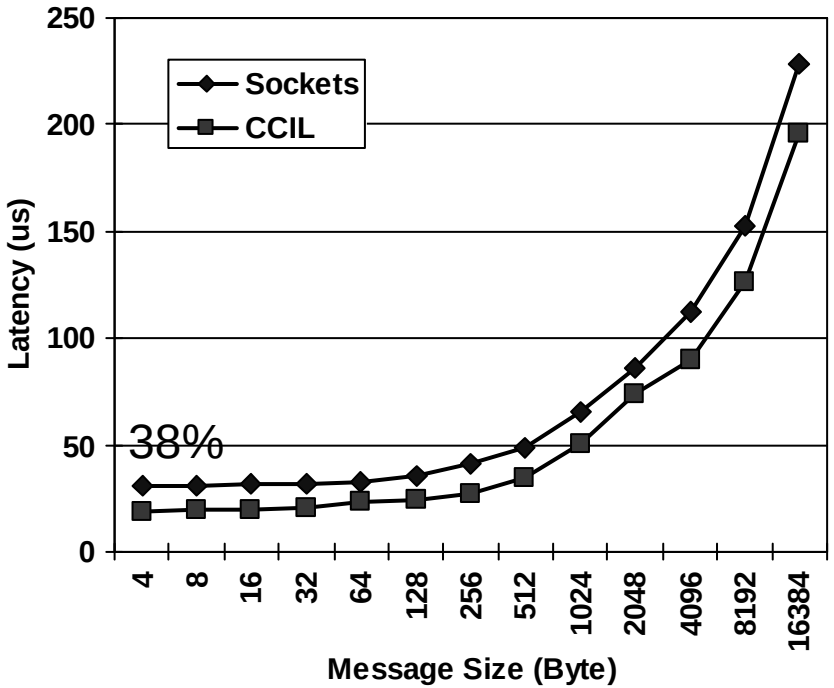




- RDMA-based communication does not affect to the application execution time
- RDMA has a strong potential of saving the CPU resource

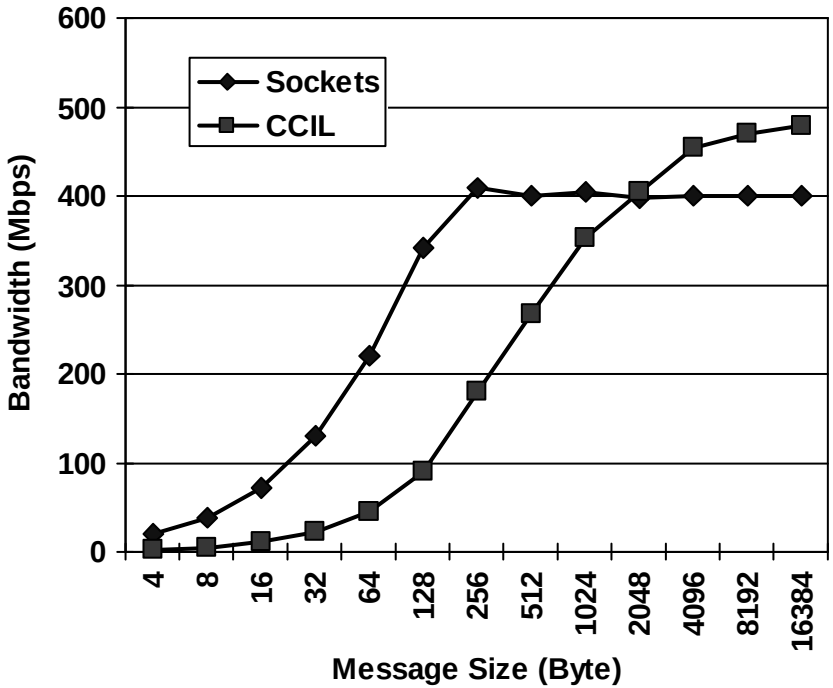
Unification of Communication Interface

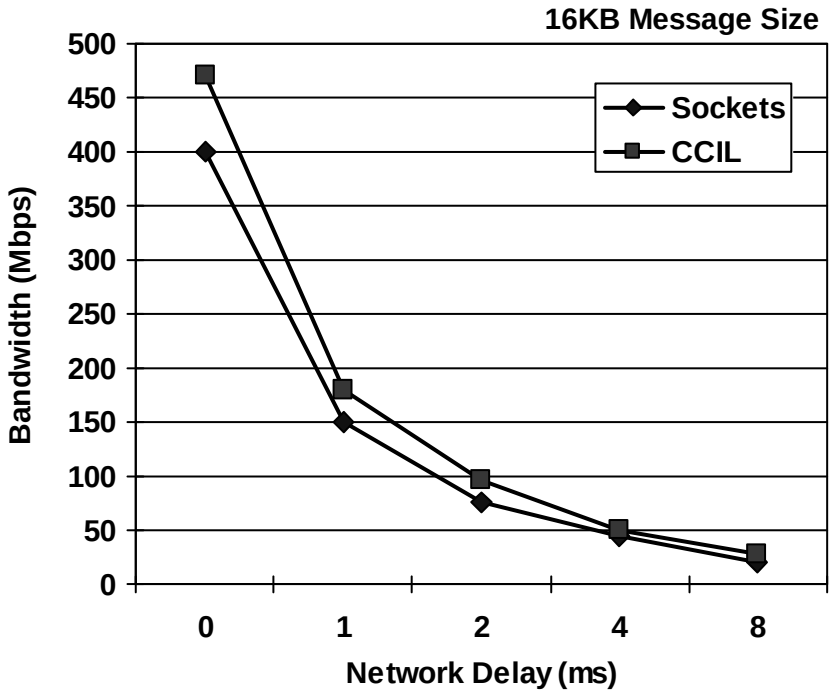




- RDMA over IP can provide a unified communication interface
- RDMA can achieve lower latency for intra-cluster communication

Bandwidth





- Where is the bottleneck?
  - Ethernet devices on the router
  - TCP window size

Contents

Conclusions

- The first quantitative study of RDMA over IP on a WAN setup
- WAN Emulator for Custer-of-Clusters
  - Degen
- RDMA over IP Can
  - Save CPU resource on the server side even on a high delay WAN environment
  - Achieve better
    - computation and communication overlap
    - communication progress
    - peak bandwidth
  - Provide unified interface

Future Work

- Performance Evaluations
  - Other performance factors
    - impact of address exchange
    - bandwidth
  - Application-level performance
- WAN Emulator for Cluster-of-Clusters
  - Delay model
  - Other components
- RDMA-aware Middleware for Widely Distributed Systems over WAN

Acknowledgements

Our research is supported by the following organizations:

- Current Funding support by

- Current Equipment donations by

Thank You

{ jinhy, narravul, browngre, vaidyana, balaji, panda}@
cse.ohio-state.edu

Network-Based Computing Laboratory

http://nowlab.cse.ohio-state.edu/