NETWORK-BASED COMPUTING LABORATORY

Performance Characterization of a 10-Gigabit Ethernet TOE

W. Feng[¥] **P. Balaji[α]** C. Baron[£]

L. N. Bhuyan[£] D. K. Panda[α]

[¥]Advanced Computing Lab, Los Alamos National Lab

[α]Network Based Computing Lab, Ohio State University

[£]CARES Group, U. C. Riverside

OHIO STATE

Ethernet Overview

- Ethernet is the most widely used network infrastructure today
- Traditionally Ethernet has been notorious for performance issues
  - Near an order-of-magnitude performance gap compared to IBA, Myrinet, etc.
    - Cost conscious architecture
    - Most Ethernet adapters were *regular (layer 2)* adapters
    - Relied on host-based TCP/IP for network and transport layer support
    - Compatibility with existing infrastructure (switch buffering, MTU)
  - Used by 42.4% of the Top500 supercomputers
  - Key: Reasonable performance at low cost
    - TCP/IP over Gigabit Ethernet (GigE) can nearly saturate the link for current systems
    - Several local stores give out GigE cards free of cost ! ☺
- 10-Gigabit Ethernet (10GigE) recently introduced
  - 10-fold (theoretical) increase in performance while retaining existing features

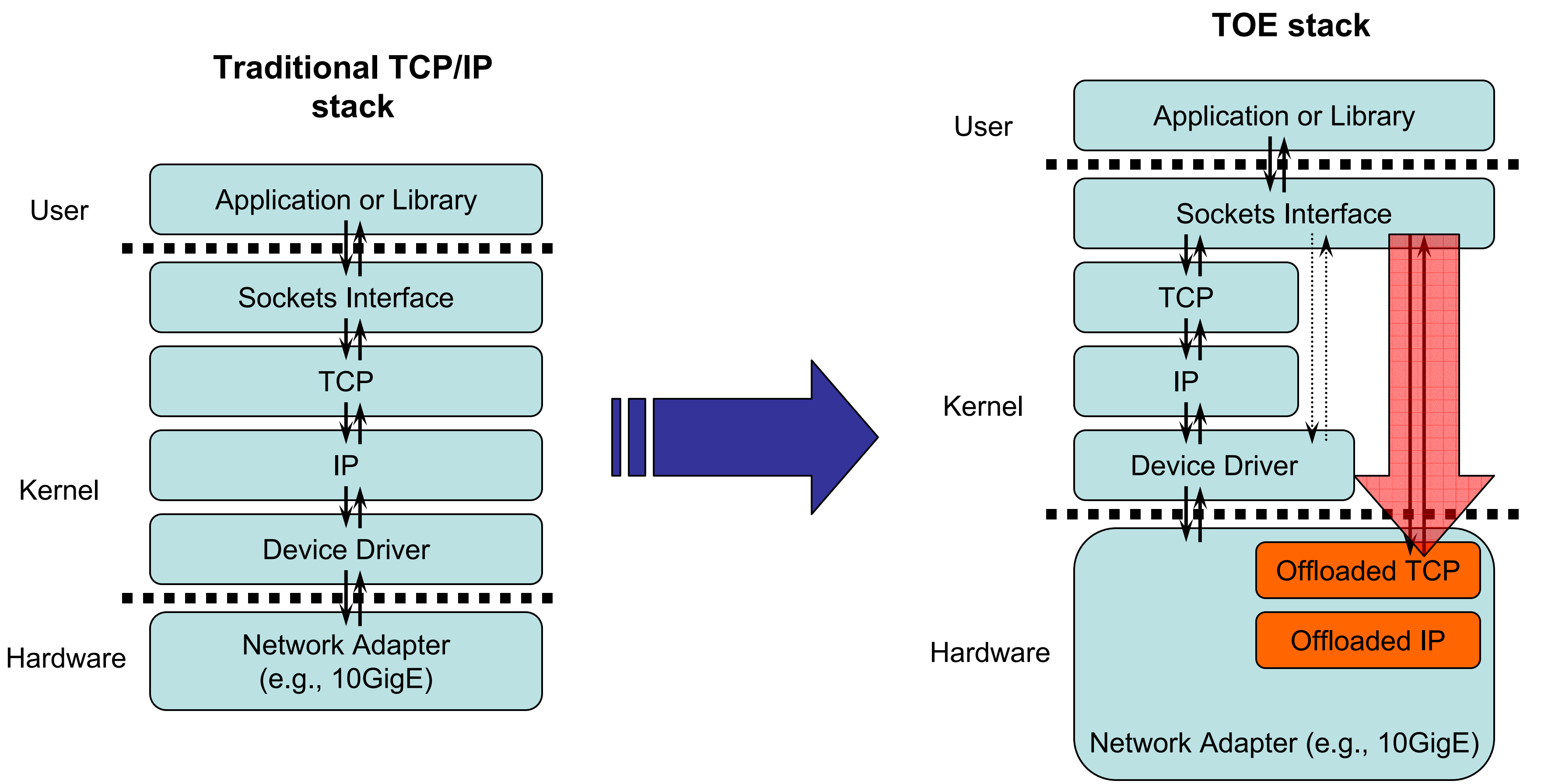
10GigE: Technology Trends

- Broken into three levels of technologies
  - Regular 10GigE adapters
    - Layer-2 adapters
    - Rely on host-based TCP/IP to provide network/transport functionality
    - Could achieve a high performance with optimizations ***[feng03:hoti, feng03:sc]***
  - TCP Offload Engines (TOEs) ***[Evaluation based on the Chelsio T110 TOE adapters]***
    - Layer-4 adapters
    - Have the entire TCP/IP stack offloaded on to hardware
    - Sockets layer retained in the host space
  - RDDP-aware adapters
    - Layer-4 adapters
    - Entire TCP/IP stack offloaded on to hardware
    - Support more features than TCP Offload Engines
      - No sockets ! Richer RDDP interface !
      - E.g., Out-of-order placement of data, RDMA semantics

Presentation Overview

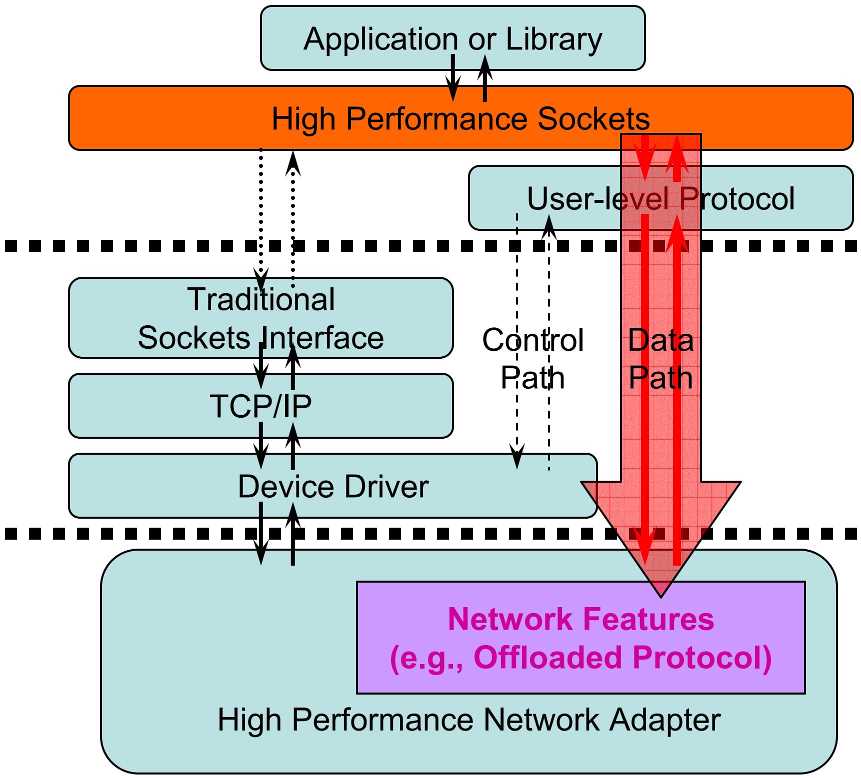
- Introduction and Motivation
- TCP Offload Engines Overview
- Experimental Evaluation
- Conclusions and Future Work

What is a TCP Offload Engine (TOE)?



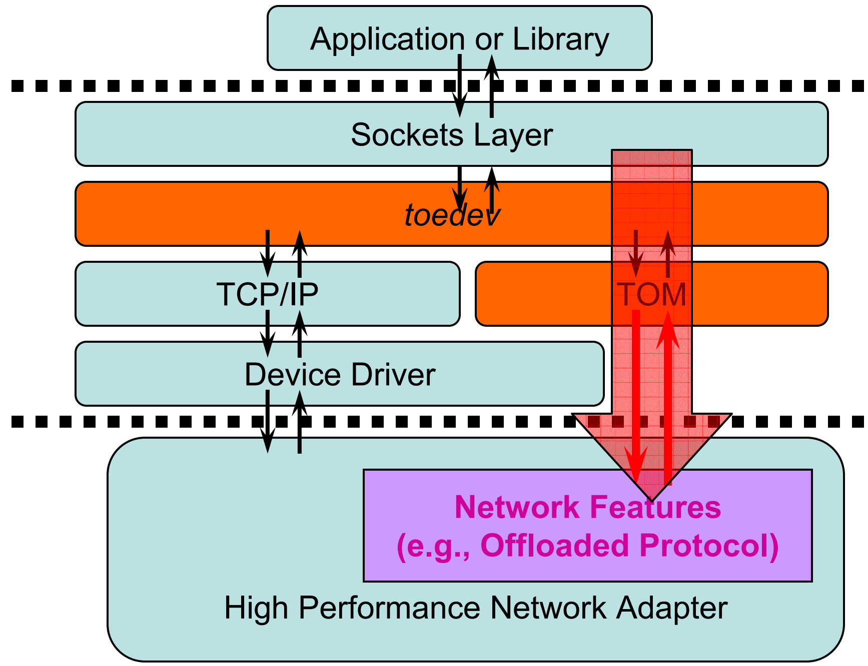
Interfacing with the TOE

High Performance Sockets



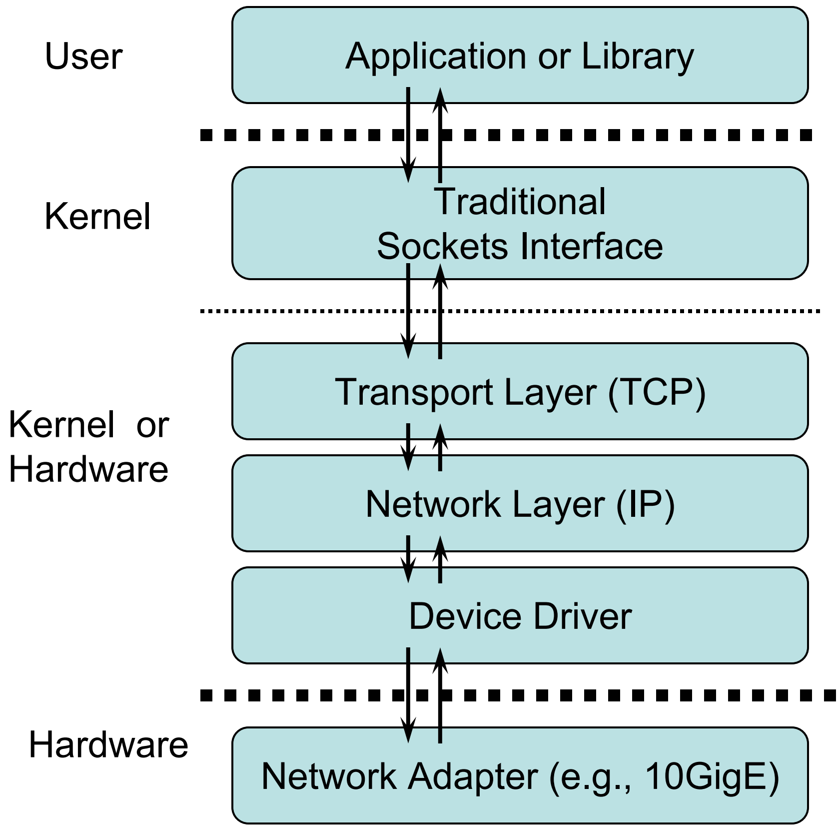
- No changes required to the core kernel
- Some of the sockets functionality duplicated

TCP Stack Override



- Kernel needs to be patched
- Some of the TCP functionality duplicated
- No duplication in the sockets functionality

What does the TOE (NOT) provide?



1. **Compatibility:** Network-level compatibility with existing TCP/IP/Ethernet; Application-level compatibility with the sockets interface

- **Performance:** Application performance no longer restricted by the performance of traditional host-based TCP/IP stack

2. **Feature-rich interface:** Application interface restricted to the sockets interface ! *[rait05]*

[rait05]: Support iWARP compatibility and features for regular network adapters. P. Balaji, H. –W. Jin, K. Vaidyanathan and D. K. Panda. In the RAIT workshop; held in conjunction with Cluster Computing, Aug 26th, 2005.

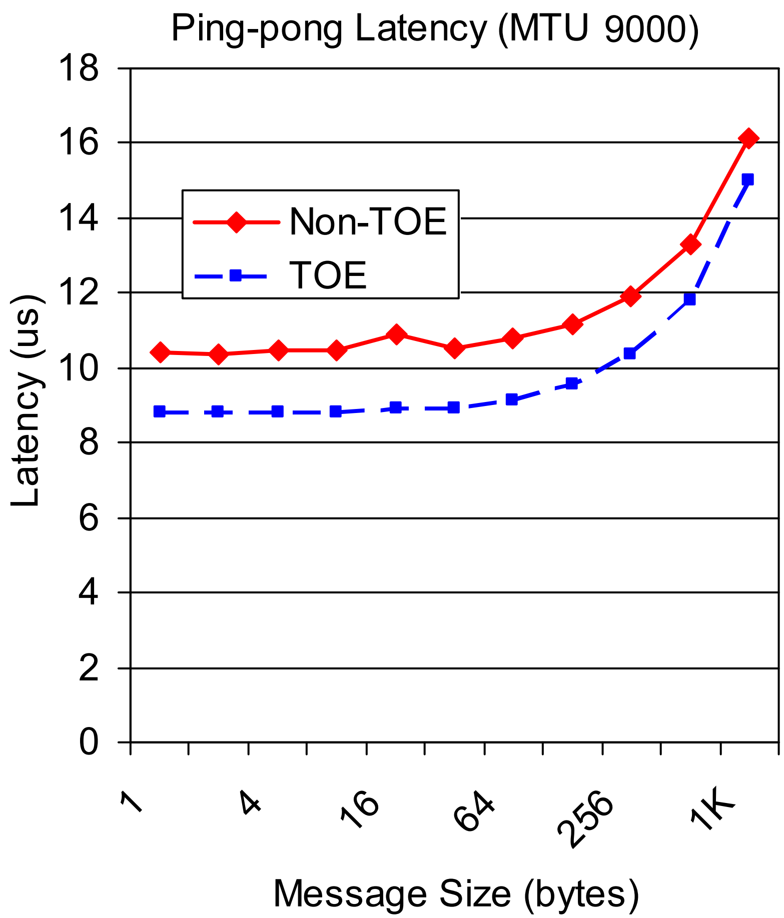
Presentation Overview

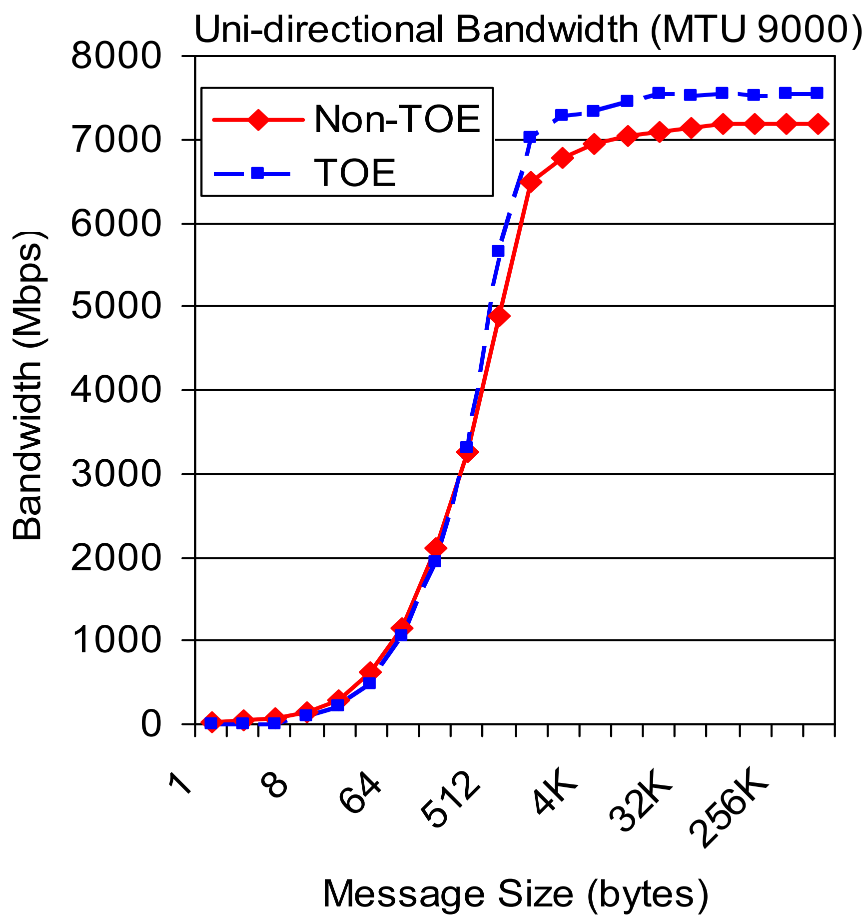
- Introduction and Motivation
- TCP Offload Engines Overview
- Experimental Evaluation
- Conclusions and Future Work

Experimental Test-bed and the Experiments

- Two test-beds used for the evaluation
  - Two 2.2GHz Opteron machines with 1GB of 400MHz DDR SDRAM
    - Nodes connected back-to-back
  - Four 2.0GHz quad-Opteron machines with 4GB of 333MHz DDR SDRAM
    - Nodes connected with a Fujitsu XG1200 switch (450ns flow-through latency)
- Evaluations in three categories
  - Sockets-level evaluation
    - Single-connection Micro-benchmarks
    - Multi-connection Micro-benchmarks
  - MPI-level Micro-benchmark evaluation
  - Application-level evaluation with the Apache Web-server

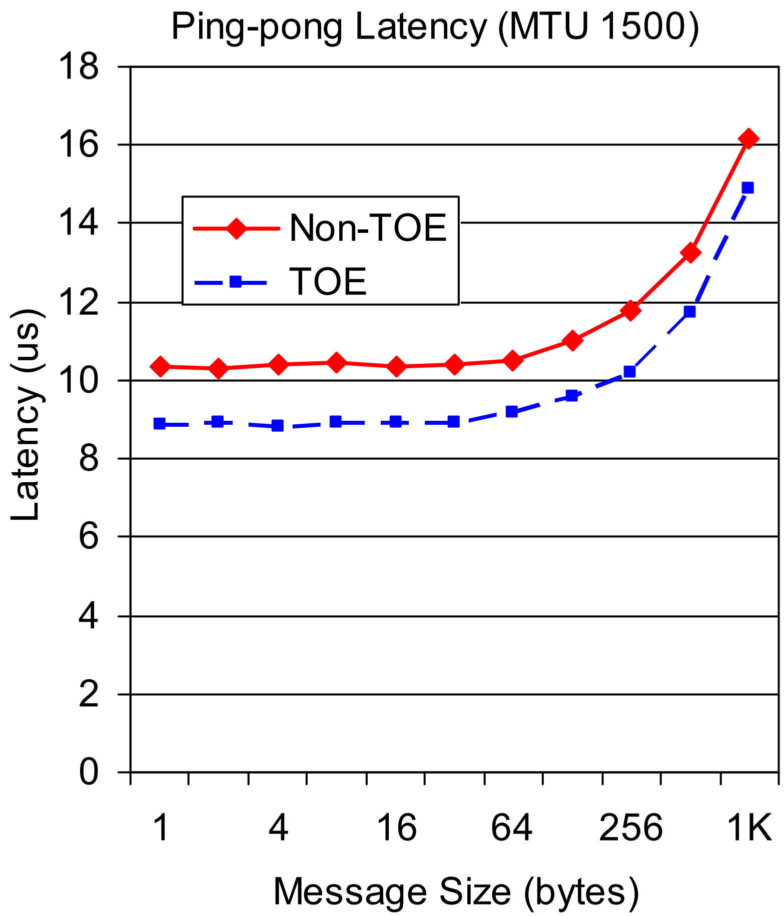
Latency and Bandwidth Evaluation (MTU 9000)

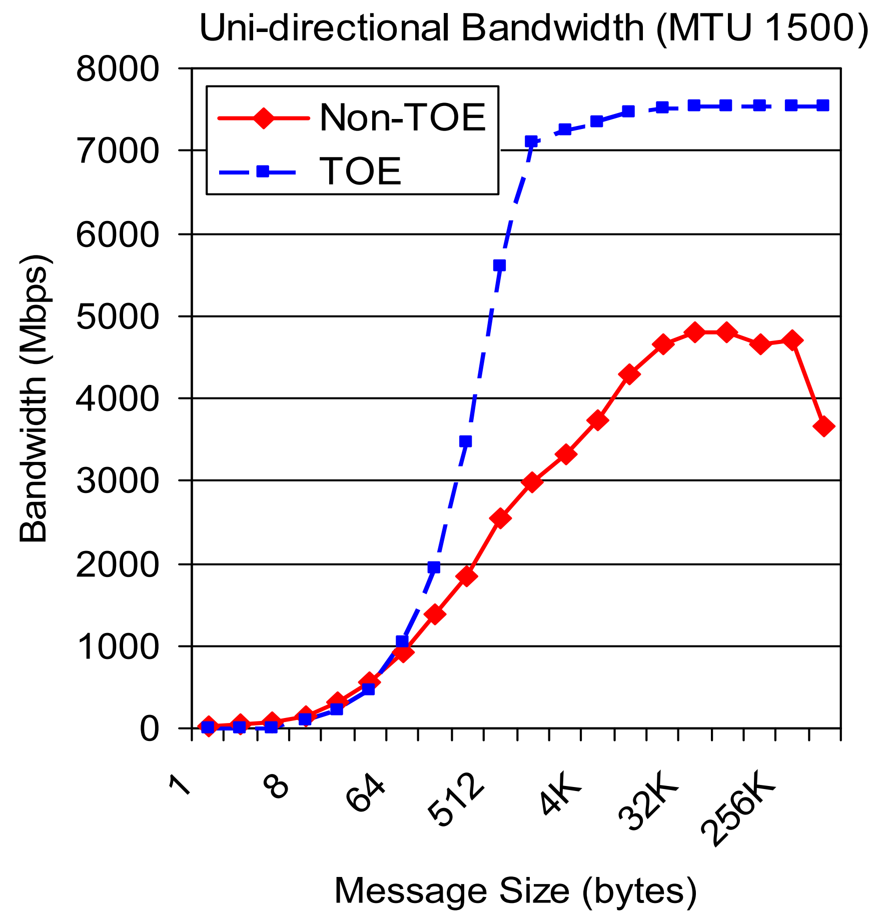




- TOE achieves a latency of about 8.6us and a bandwidth of 7.6Gbps at the sockets layer
- Host-based TCP/IP achieves a latency of about 10.5us (25% higher) and a bandwidth of 7.2Gbps (5% lower)
- For Jumbo frames, host-based TCP/IP performs quite close to the TOE

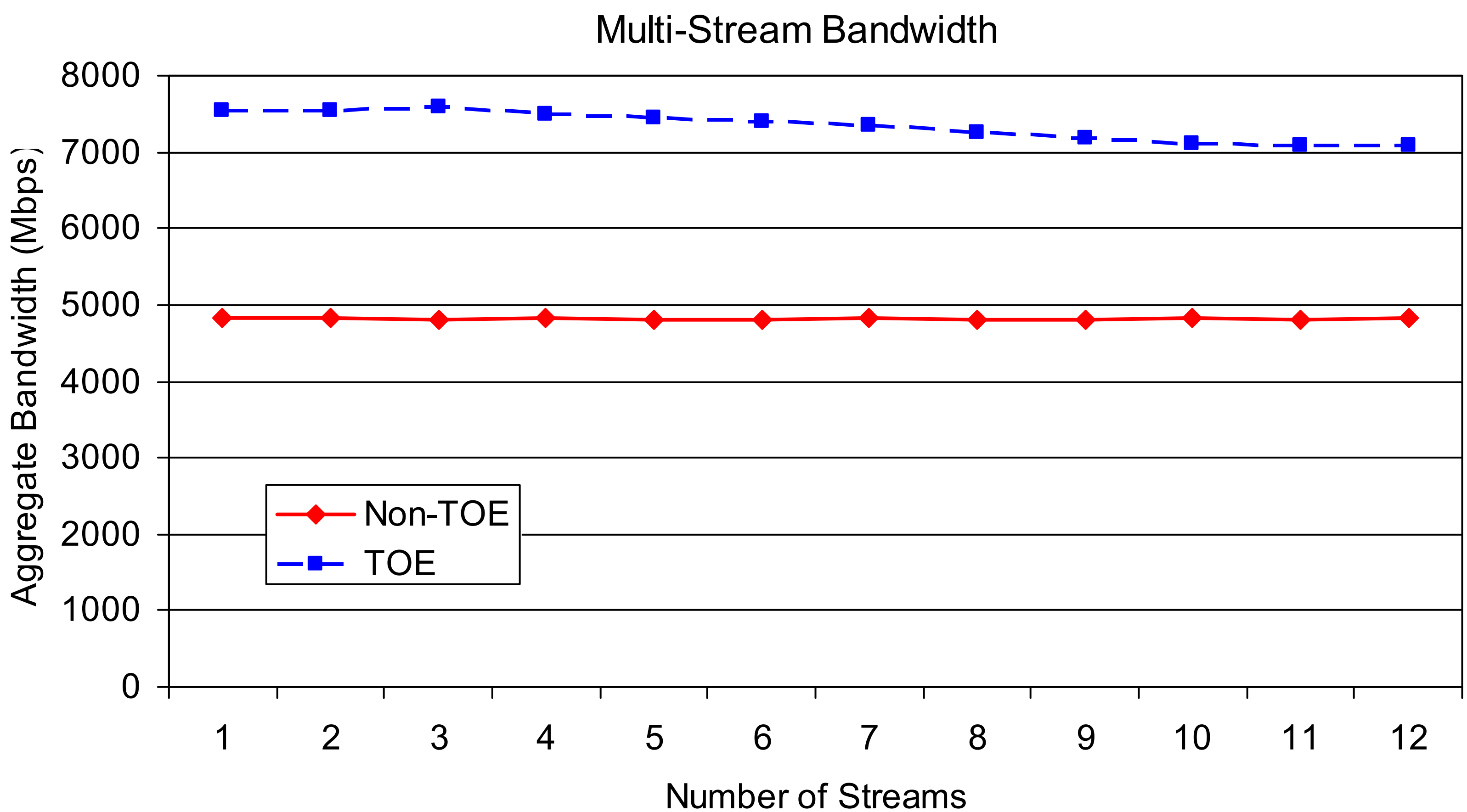
Latency and Bandwidth Evaluation (MTU 1500)





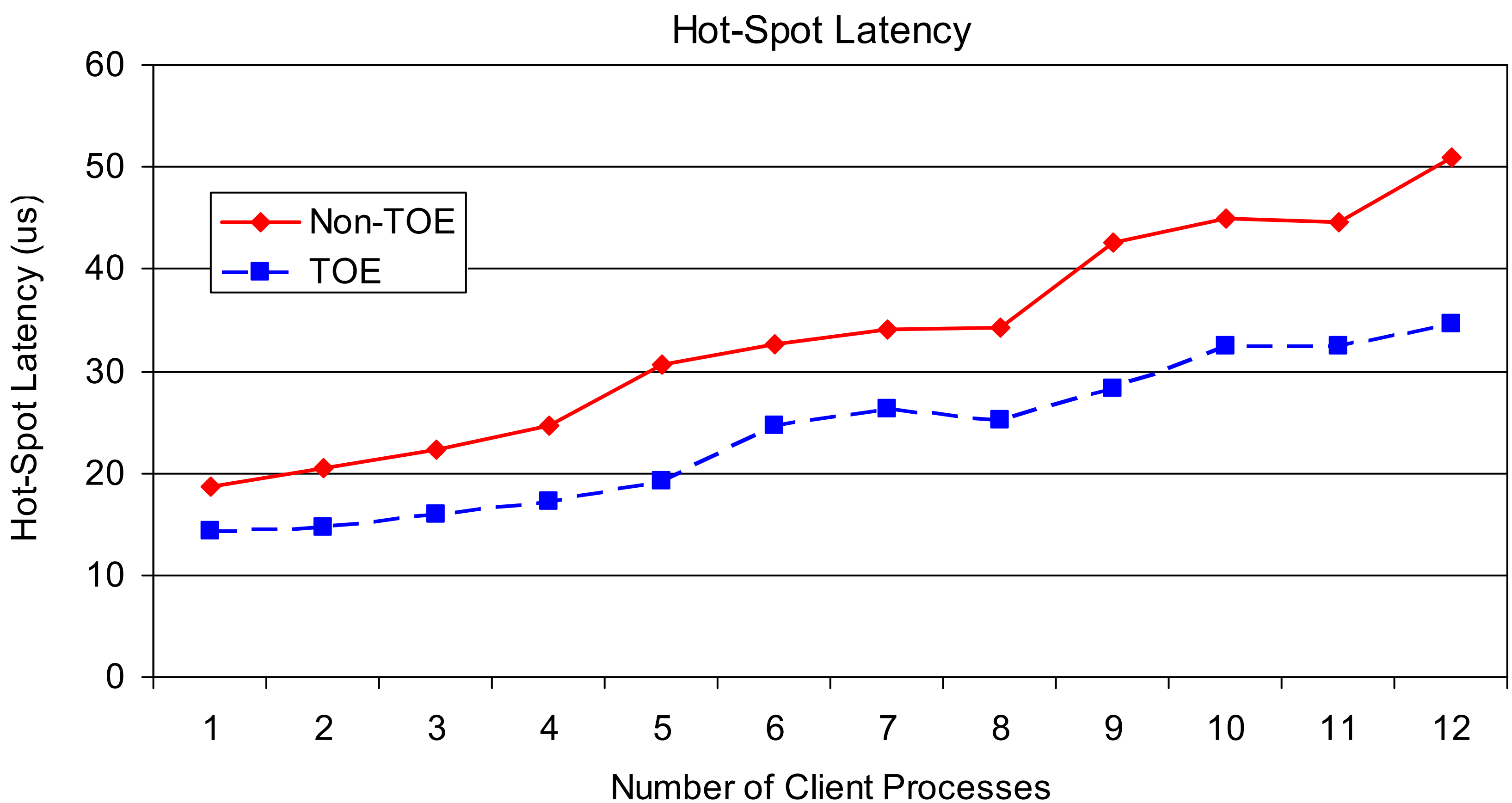
- No difference in latency for either stack
- The bandwidth of host-based TCP/IP drops to 4.9Gbps (more interrupts; higher overhead)
- For standard sized frames, TOE significantly outperforms host-based TCP/IP (segmentation offload is the key)

Multi-Stream Bandwidth



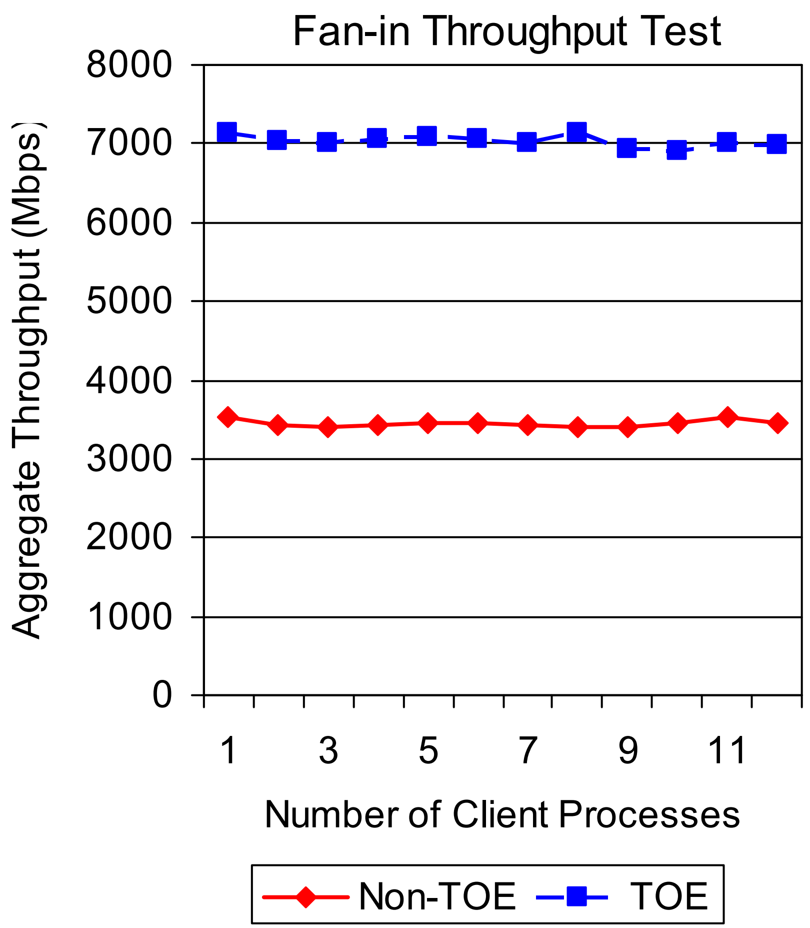
The throughput of the TOE stays between 7.2 and 7.6Gbps

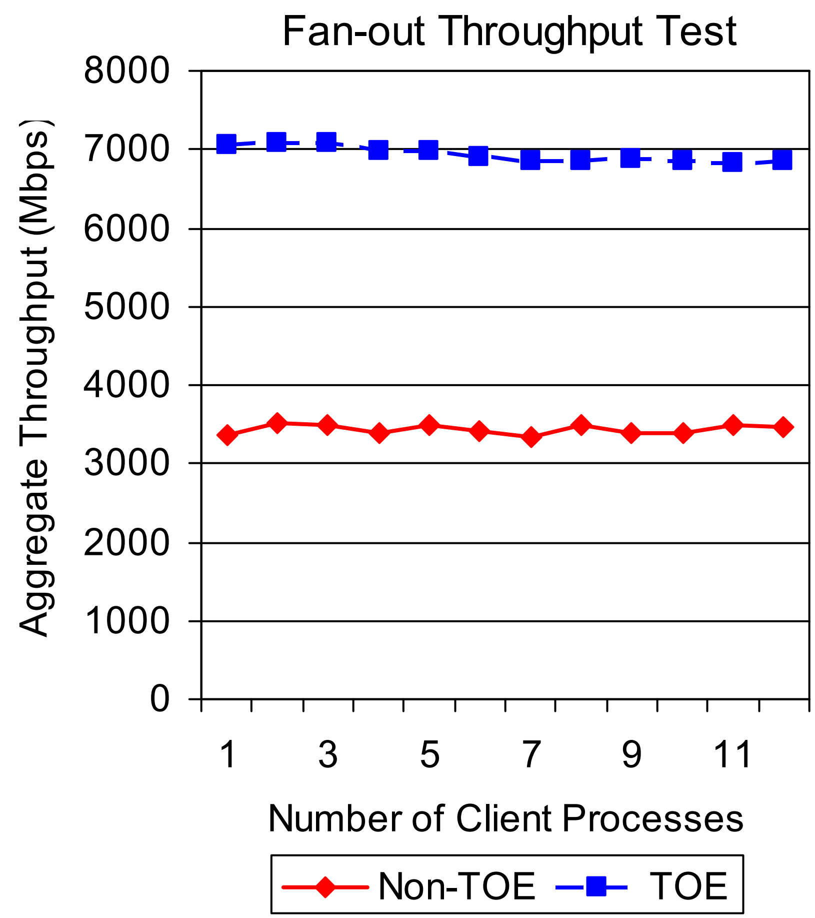
Hot Spot Latency Test (1 byte)



Connection scalability tested up to 12 connections; TOE achieves similar or better scalability as the host-based TCP/IP stack

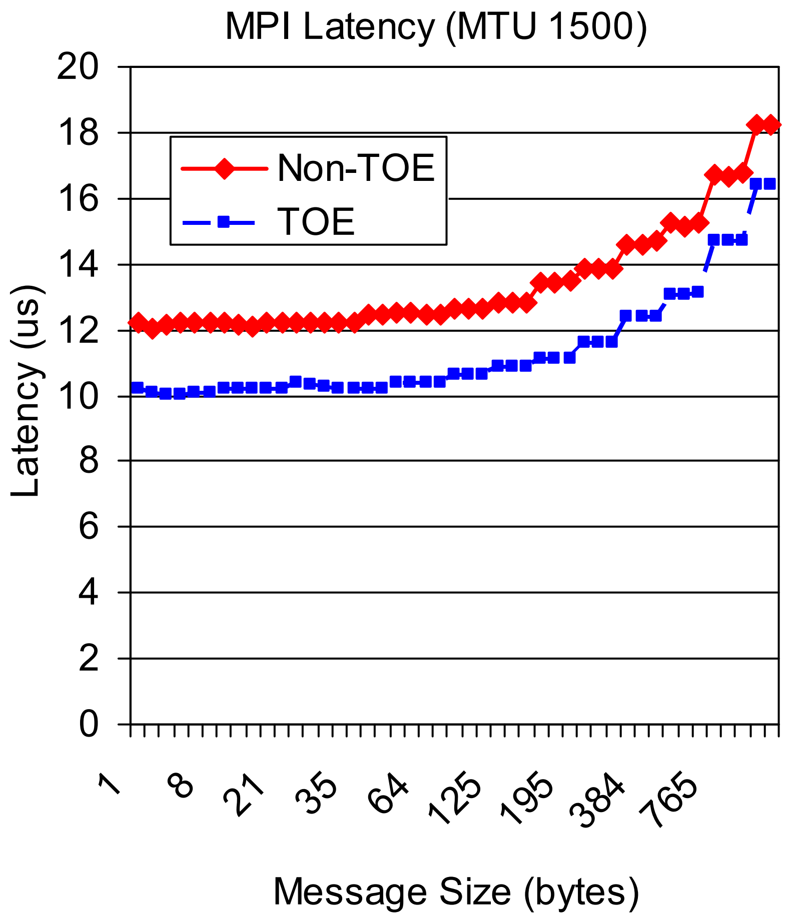
Fan-in and Fan-out Throughput Tests

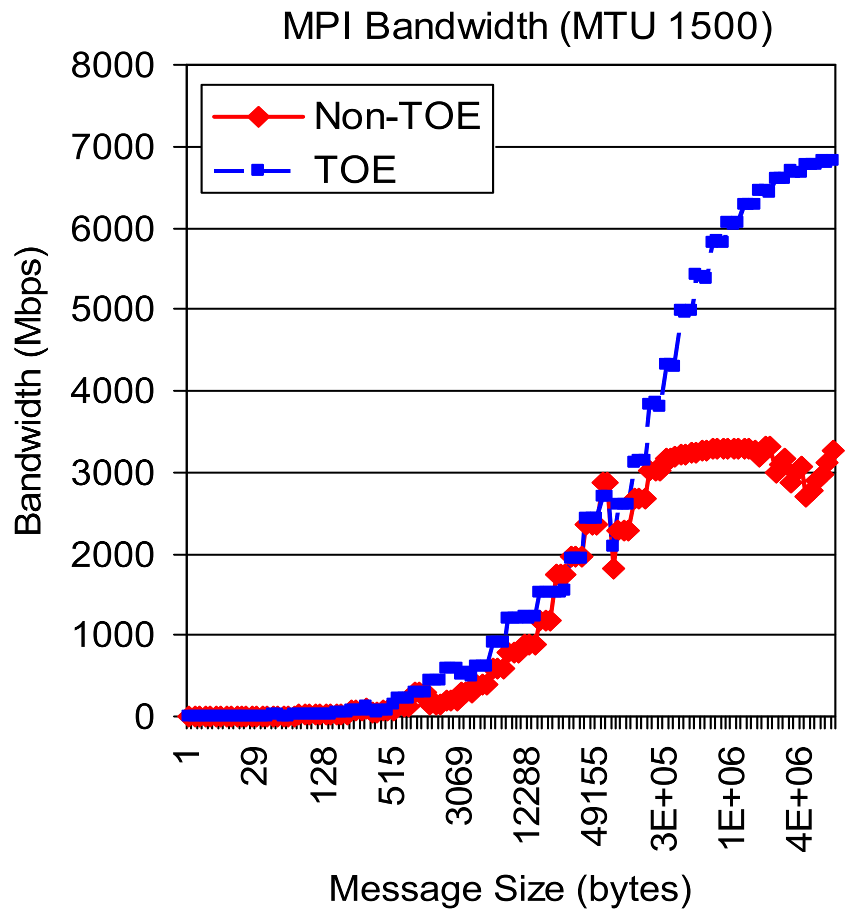




Fan-in and Fan-out tests show similar scalability

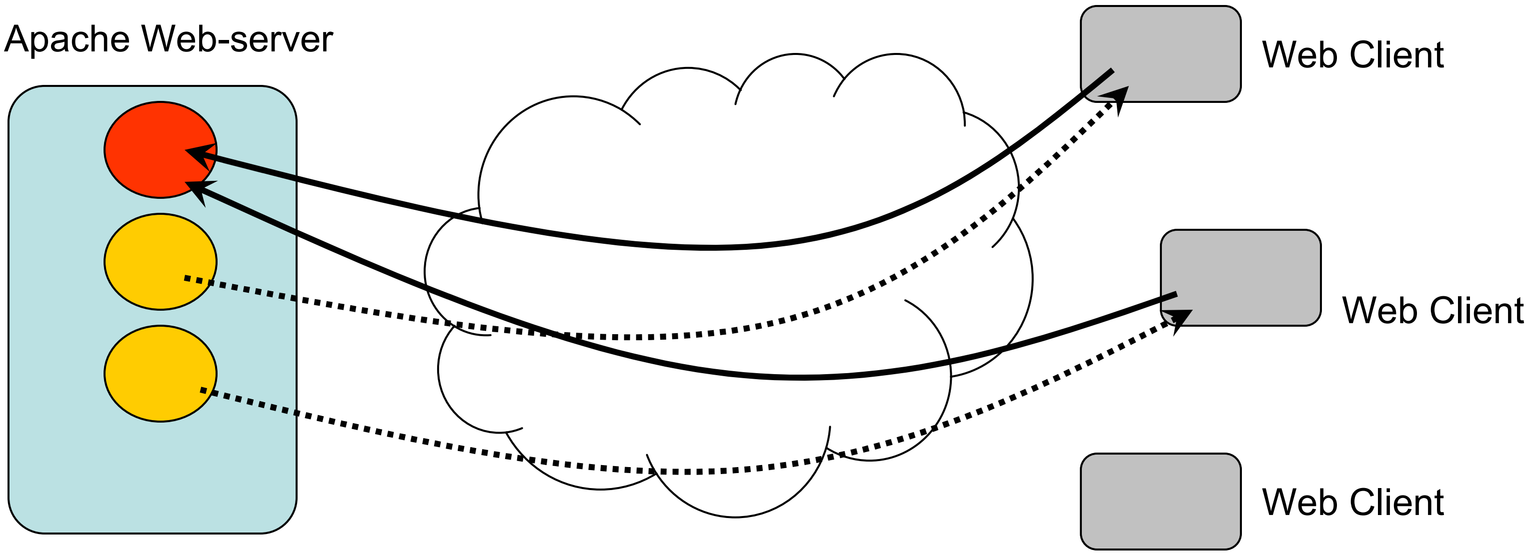
MPI-level Comparison





MPI latency and bandwidth show similar trends as socket-level latency and bandwidth

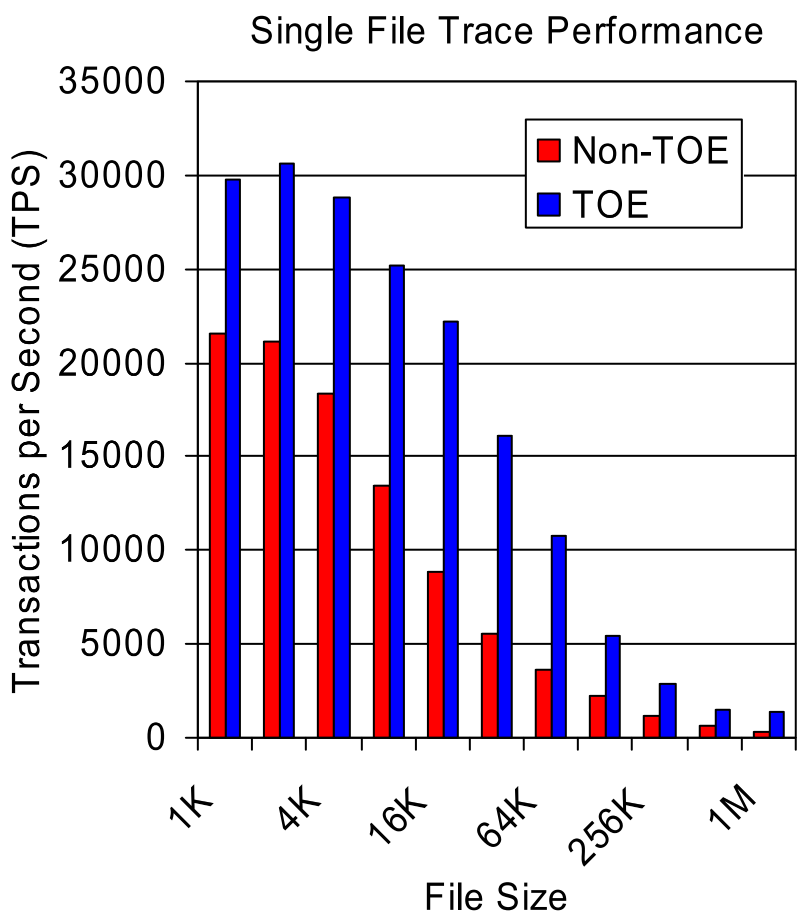
Application-level Evaluation: Apache Web-Server

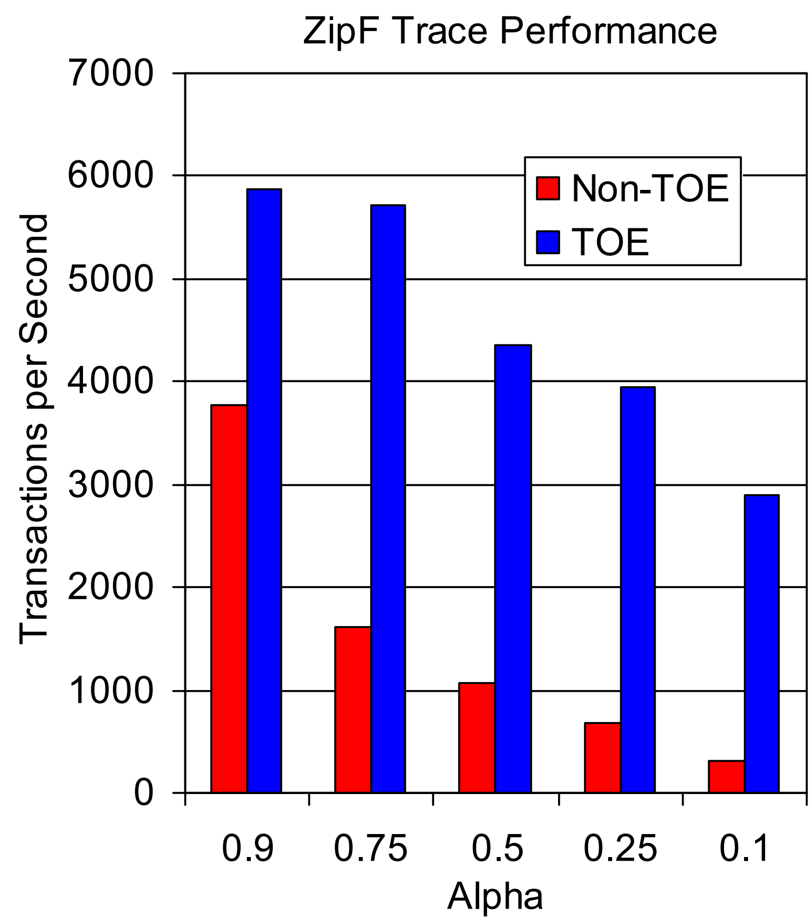


We perform two kinds of evaluations with the Apache web-server:

1. Single file traces
   - All clients always request the same file of a given size
   - Not diluted by other system and workload parameters
2. Zipf-based traces
   - The probability of requesting the I^{th} most popular document is inversely proportional to I^{α}
   - α is constant for a given trace; it represents the temporal locality of a trace
   - A high α value represents a high percent of requests for small files

Apache Web-server Evaluation





Presentation Overview

- Introduction and Motivation
- TCP Offload Engines Overview
- Experimental Evaluation
- Conclusions and Future Work

Conclusions

- For a wide-spread acceptance of 10-GigE in clusters
  - Compatibility
  - Performance
  - Feature-rich interface
- Network as well as Application-level compatibility is available
  - On-the-wire protocol is still TCP/IP/Ethernet
  - Application interface is still the sockets interface
- Performance Capabilities
  - Significant performance improvements compared to the host-stack
    - Close to 65% improvement in bandwidth for standard sized (1500byte) frames
- Feature-rich interface: Not quite there yet !
  - Extended Sockets Interface
  - iWARP offload

Continuing and Future Work

- Comparing 10GigE TOEs to other interconnects
  - Sockets Interface [cluster05]
  - MPI Interface
  - File and I/O sub-systems
- Extending the sockets interface to support iWARP capabilities [rait05]
- Extending the TOE stack to allow protocol offload for UDP sockets

Web Pointers

RADIANT — RESEARCH & DEVELOPMENT IN ADVANCED NETWORK TECHNOLOGY

Network Based Computing Laboratory

NOWLAB

http://public.lanl.gov/radiant

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

feng@lanl.gov

balaji@cse.ohio-state.edu