NETWORK-BASED COMPUTING LABORATORY

Sockets Direct Protocol Over InfiniBand in Clusters: Is it Beneficial?

P. Balaji, S. Narravula, K. Vaidyanathan, S. Krishnamoorthy, J. Wu and D. K. Panda

Network Based Computing Laboratory

The Ohio State University

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Presentation Layout

- Introduction and Background
- Sockets Direct Protocol (SDP)
- Multi-Tier Data-Centers
- Parallel Virtual File System (PVFS)
- Experimental Evaluation
- Conclusions and Future Work

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Introduction

- Advent of High Performance Networks
  - Ex: InfiniBand, Myrinet, 10-Gigabit Ethernet
  - High Performance Protocols: VAPI / IBAL, GM, EMP
  - Good to build new applications
  - Not so beneficial for existing applications
    - Built around Portability: Should run on all platforms
    - TCP/IP based Sockets: A popular choice
    - Performance of Application depends on the Performance of Sockets
    - Several ***GENERIC*** optimizations for sockets to provide high performance
      - Jacobson Optimization: Integrated Checksum-Copy [Jacob89]
      - Header Prediction for Single Stream data transfer

[Jacob89]: "An analysis of TCP Processing Overhead", D. Clark, V. Jacobson, J. Romkey and H. Salwen. IEEE Communications

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

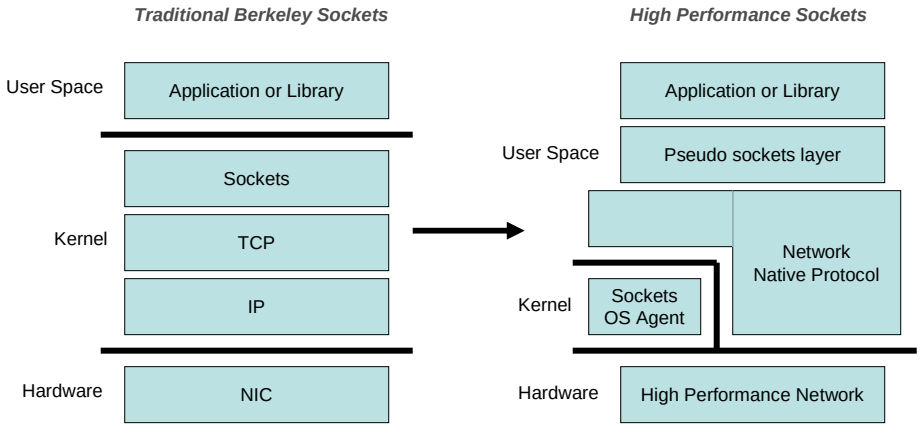
Network Specific Optimizations

- Generic Optimizations Insufficient
  - Unable to saturate high performance networks
- Sockets can utilize some network features
  - Interrupt Coalescing (can be considered generic)
  - Checksum Offload (TCP stack has to modified)
  - Insufficient!
- Can we do better?
  - High Performance Sockets
  - TCP Offload Engines (TOE)

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

High Performance Sockets



OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

InfiniBand Architecture Overview

- Industry Standard
- Interconnect for connecting compute and I/O nodes
- Provides High Performance
  - Low latency of lesser than 5us
  - Over 840MBps uni-directional bandwidth
  - Provides one-sided communication (RDMA, Remote Atomics)
- Becoming increasingly popular

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Sockets Direct Protocol (SDP*)

- IBA Specific Protocol for Data-Streaming
- Defined to serve two purposes:
  - Maintain compatibility for existing applications
  - Deliver the high performance of IBA to the applications
- Two approaches for data transfer: Copy-based and Z-Copy
- Z-Copy specifies *Source-Avail* and *Sink-Avail* messages
  - *Source-Avail* allows destination to RDMA Read from source
  - *Sink-Avail* allows source to RDMA Write to the destination
- Current implementation limitations:
  - Only supports the Copy-based implementation
  - Does not support *Source-Avail* and *Sink-Avail*

SDP implementation from the Voltaire Software Stack

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

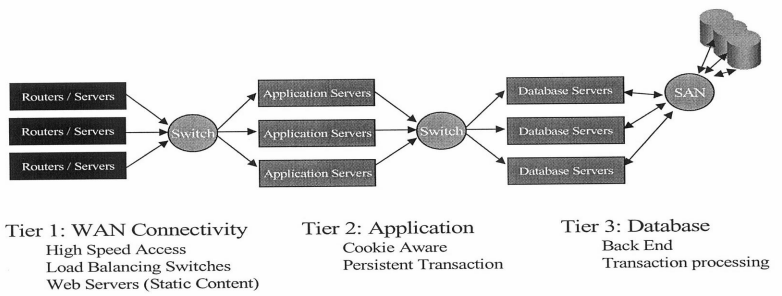
Presentation Layout

- ☞ Introduction and Background
- ☞ Sockets Direct Protocol (SDP)
- ☞ **Multi-Tier Data-Centers**
- ☞ Parallel Virtual File System (PVFS)
- ☞ Experimental Evaluation
- ☞ Conclusions and Future Work

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Multi-Tier Data-Centers



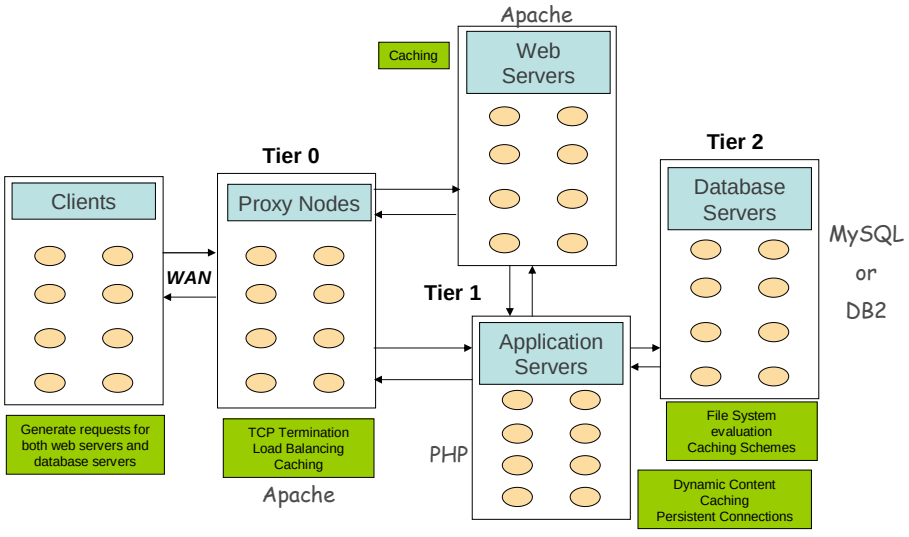
(Courtesy Mellanox Corporation)

- Client Requests come over the WAN (TCP based + Ethernet Connectivity)
- Traditional TCP based requests are forwarded to the inner tiers
- Performance is limited due to TCP
  - Can we use SDP to improve the data-center performance?
  - SDP is not compatible with traditional sockets: Requires TCP termination!

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

3-Tier Data-Center Test-bed at OSU



OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

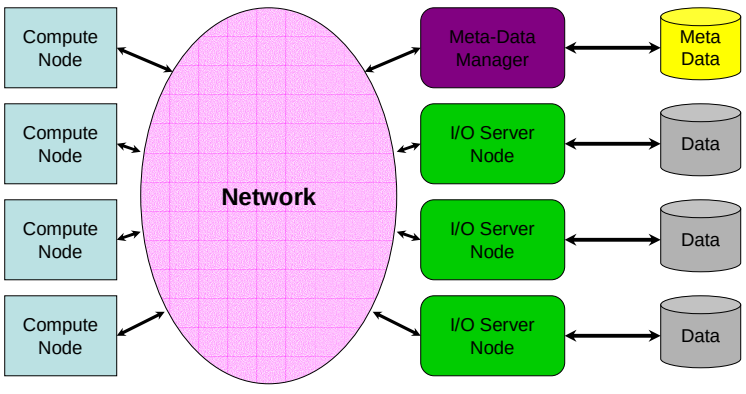
Presentation Layout

- Introduction and Background
- Sockets Direct Protocol (SDP)
- Multi-Tier Data-Centers
- **Parallel Virtual File System (PVFS)**
- Experimental Evaluation
- Conclusions and Future Work

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Parallel Virtual File System (PVFS)

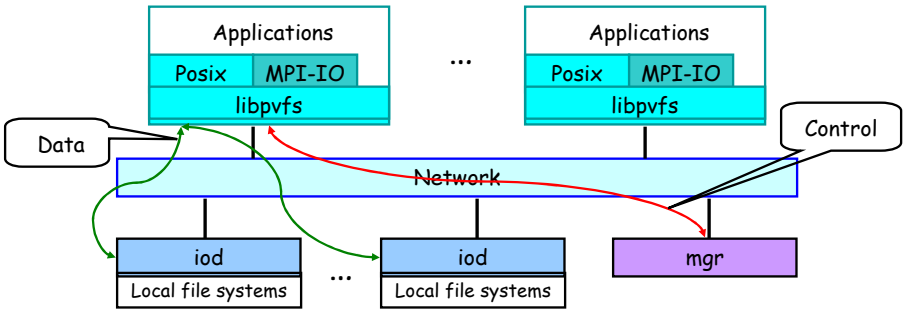


- *Relies on Striping of data across different nodes*
- *Tries to aggregate I/O bandwidth from multiple nodes*
- *Utilizes the local file system on the I/O Server nodes*

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Parallel I/O in Clusters via PVFS



- PVFS: Parallel Virtual File System
  - Parallel: stripe/access data across multiple nodes
  - Virtual: exists only as a set of user-space daemons
  - File system: common file access methods (open, read/write)
- Designed by ANL and Clemson

"PVFS over InfiniBand: Design and Performance Evaluation", Jiesheng Wu, Pete Wyckoff and D. K. Panda. International Conference on Parallel Processing (ICPP), 2003.

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Presentation Layout

- Introduction and Background
- Sockets Direct Protocol (SDP)
- Multi-Tier Data-Centers
- Parallel Virtual File System (PVFS)
- **Experimental Evaluation**
  - Micro-Benchmark Evaluation
  - Data-Center Performance
  - PVFS Performance
- Conclusions and Future Work

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

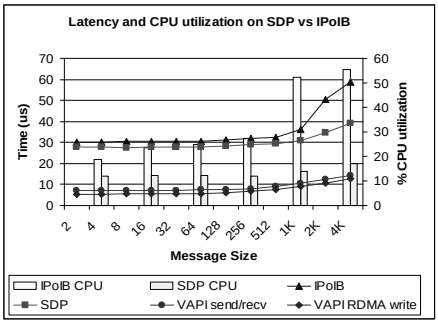
Experimental Test-bed

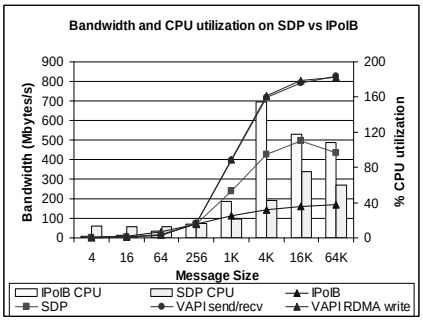
- Eight Dual 2.4GHz Xeon processor nodes
- 64-bit 133MHz PCI-X interfaces
- 512KB L2-Cache and 400MHz Front Side Bus
- Mellanox InfiniHost MT23108 Dual Port 4x HCAs
- MT43132 eight 4x port Switch
- SDK version 0.2.0
- Firmware version 1.17

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Latency and Bandwidth Comparison



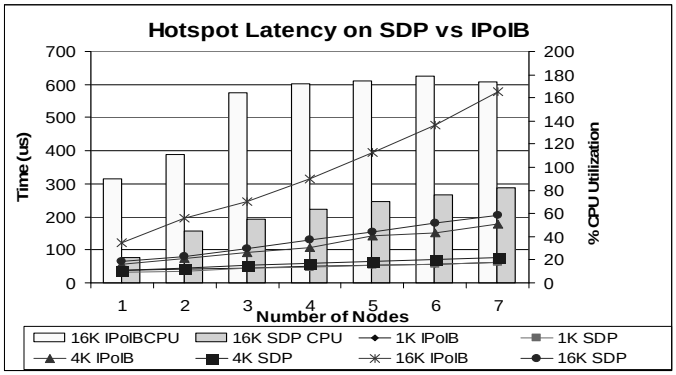


- *SDP achieves 500MBps bandwidth compared to 180MBps of IPoIB*
- *Latency of 27us compared to 31us of IPoIB*
- *Improved CPU Utilization*

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Hotspot Latency

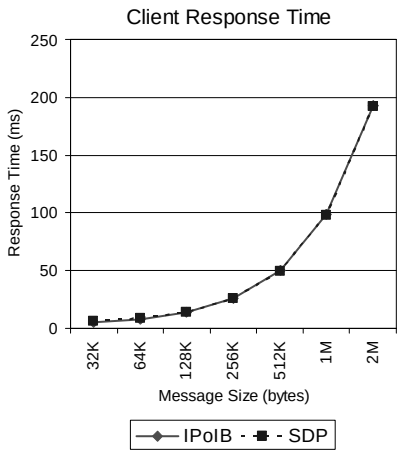


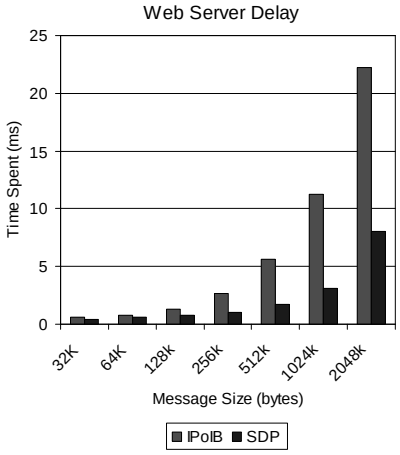
- *SDP is more scalable in hot-spot scenarios*

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Data-Center Response Time



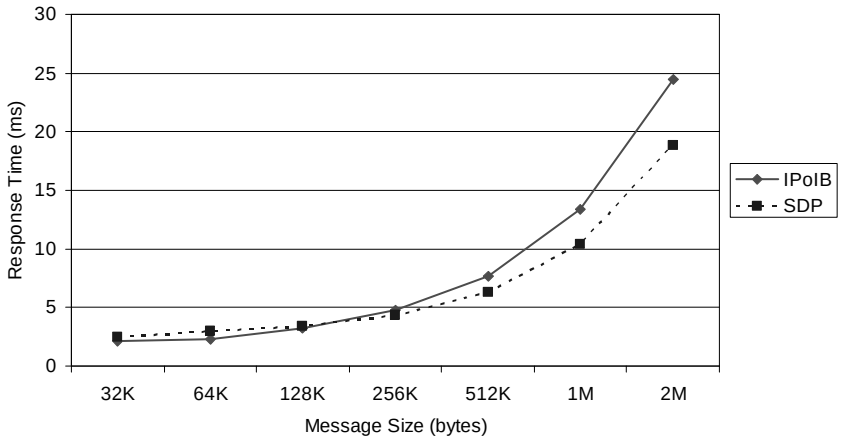


- *SDP shows very little improvement: Client network (Fast Ethernet) becomes the bottleneck*
- *Client network bottleneck reflected in the web server delay: up to 3 times improvement with SDP*

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Data-Center Response Time (Fast Clients)

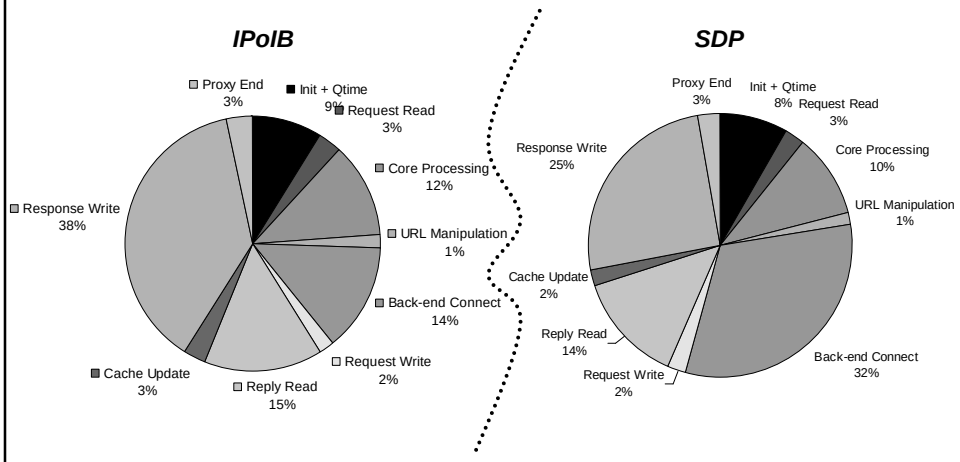


• *SDP performs well for large files; not very well for small files*

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

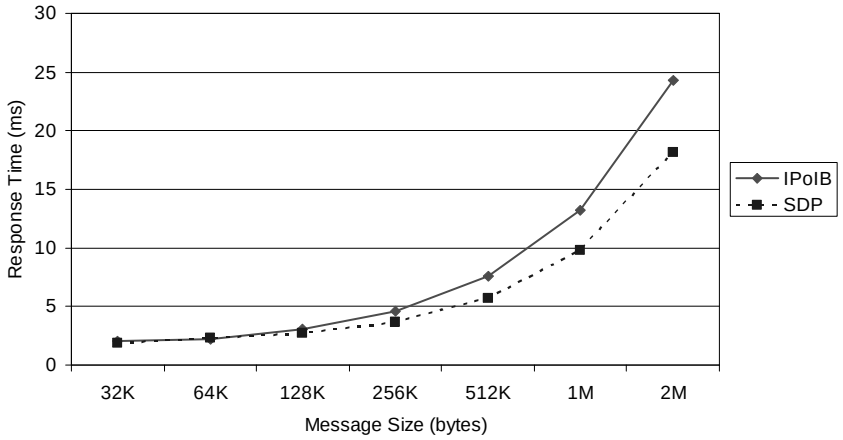
Data-Center Response Time Split-up



OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Data-Center Response Time without Connection Time Overhead

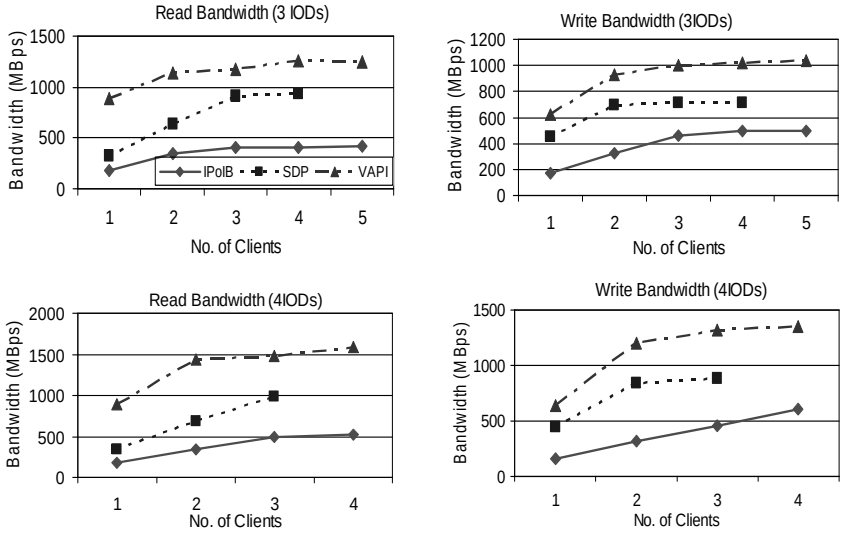


• *Without the connection time, SDP would perform well for all file sizes*

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

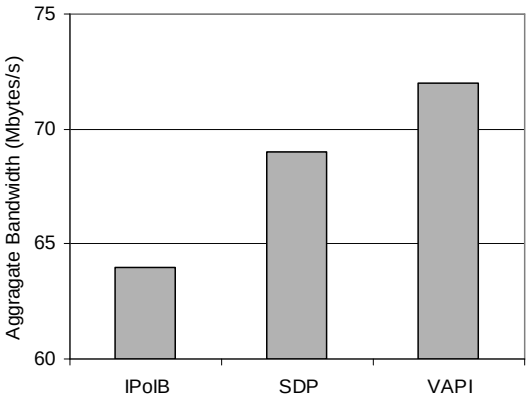
PVFS Performance using *ramfs*



OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

PVFS Performance with sync (ext3fs)



• *Clients can push data faster to IODs using SDP; de-stage bandwidth remains the same*

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Conclusions

- User-Level Sockets designed with two motives:
  - Compatibility for existing applications
  - High Performance for modern networks
- SDP was proposed recently along similar lines
- Sockets Direct Protocol: Is it Beneficial?
  - Evaluated it using micro-benchmarks and real applications
    - Multi-Tier Data-Centers and PVFS
  - Benefits in environments it's good for
    - Communication intensive environments such as PVFS
  - Demonstrate environments it's yet to mature for
    - Connection overhead involving environments such as Data-Centers

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Future Work

- Connection Time bottleneck in SDP
  - Using dynamic registered buffer pools, FMR techniques, etc
  - Using QP pools
- Power-Law Networks
- Other applications: Streaming and Transaction
- Comparison with other high performance sockets

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

Thank You!

For more information, please visit the

Home Page

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

Network Based Computing Laboratory,

The Ohio State University

OHIO STATE

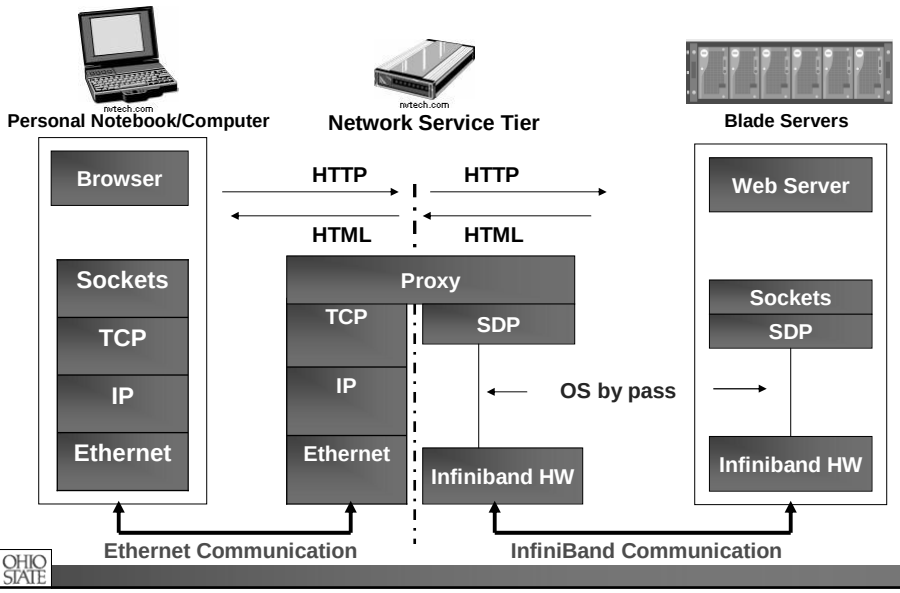
NETWORK-BASED COMPUTING LABORATORY

Backup Slides

OHIO STATE

NETWORK-BASED COMPUTING LABORATORY

TCP Termination in SDP



OHIO STATE