IP Storage: The Challenge Ahead

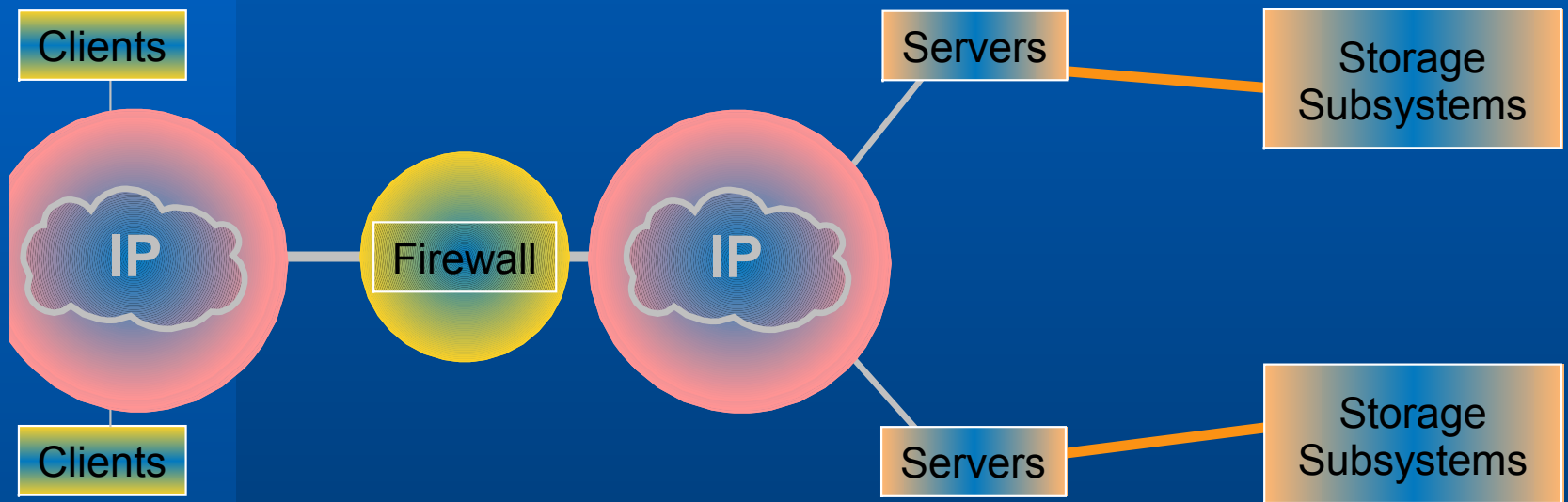
Prasenjit Sarkar

Kaladhar Voruganti

IBM Almaden Research

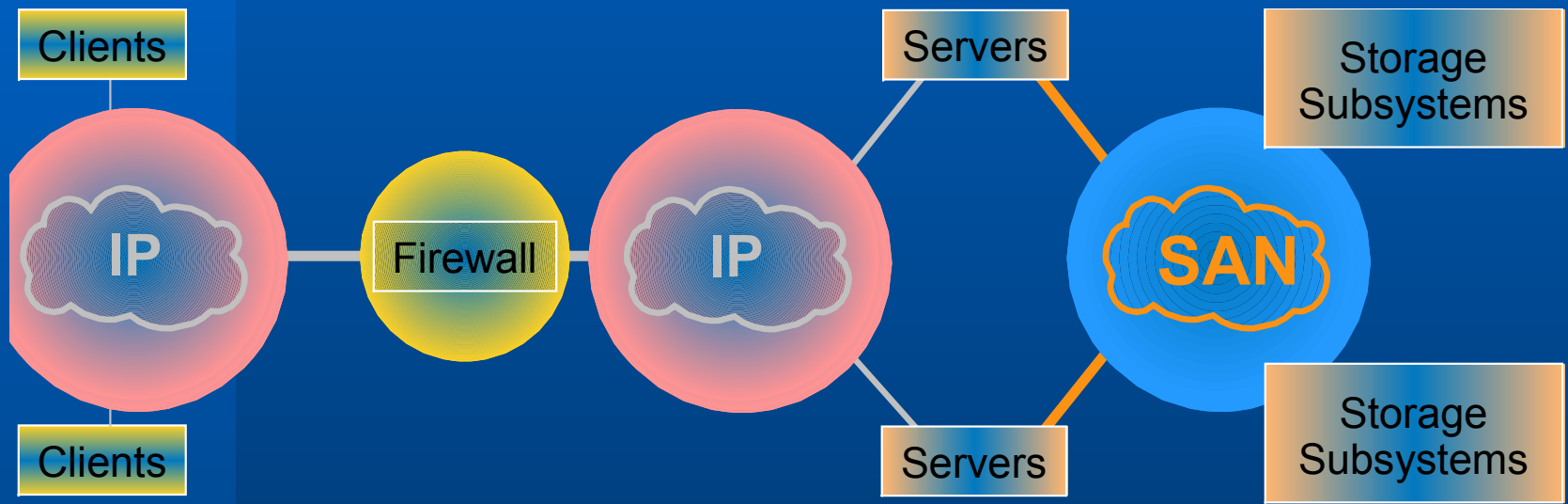
10th NASA Goddard Conference on Mass Storage Systems and Technologies

Server-attached Storage



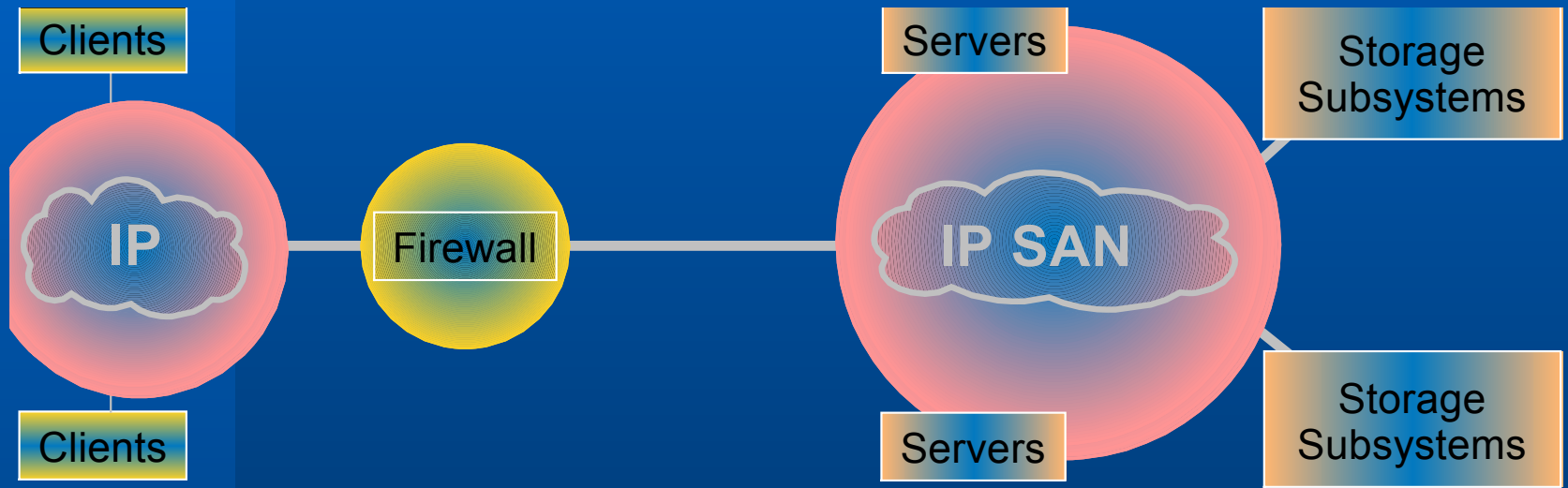
- **Restrictions on capacity, distance**
- **Management per server**
- **Limited sharing, availability**

Storage Area Networks



- **Storage as a service over gigabit networks**
- **Shared management**
- **Distance, capacity limitations largely removed**

IP Storage



- **Single commodity networking infrastructure**
- **Leverage IP: connectivity, transport(TCP), management, security, advanced features (QoS)**

Challenges

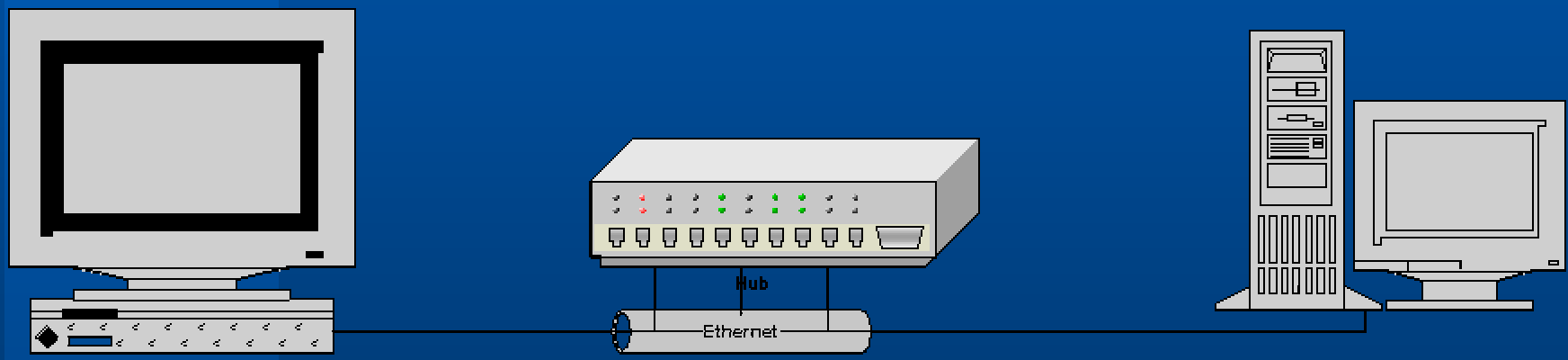
- Security
- Management
- Standardization
- Performance

Performance Issues

- **TCP/IP protocol overhead**
  - **Copy-and-checksum routine**
  - **Complexity relative to FC**
- **Interrupt overhead**
  - **One interrupt every TCP MSS**
    - **1.5K for Ethernet**
  - **Overhead significant for bulk data xfers**
- **Security overhead**

Experimental Setup

Benchmark: 100% read cache hit

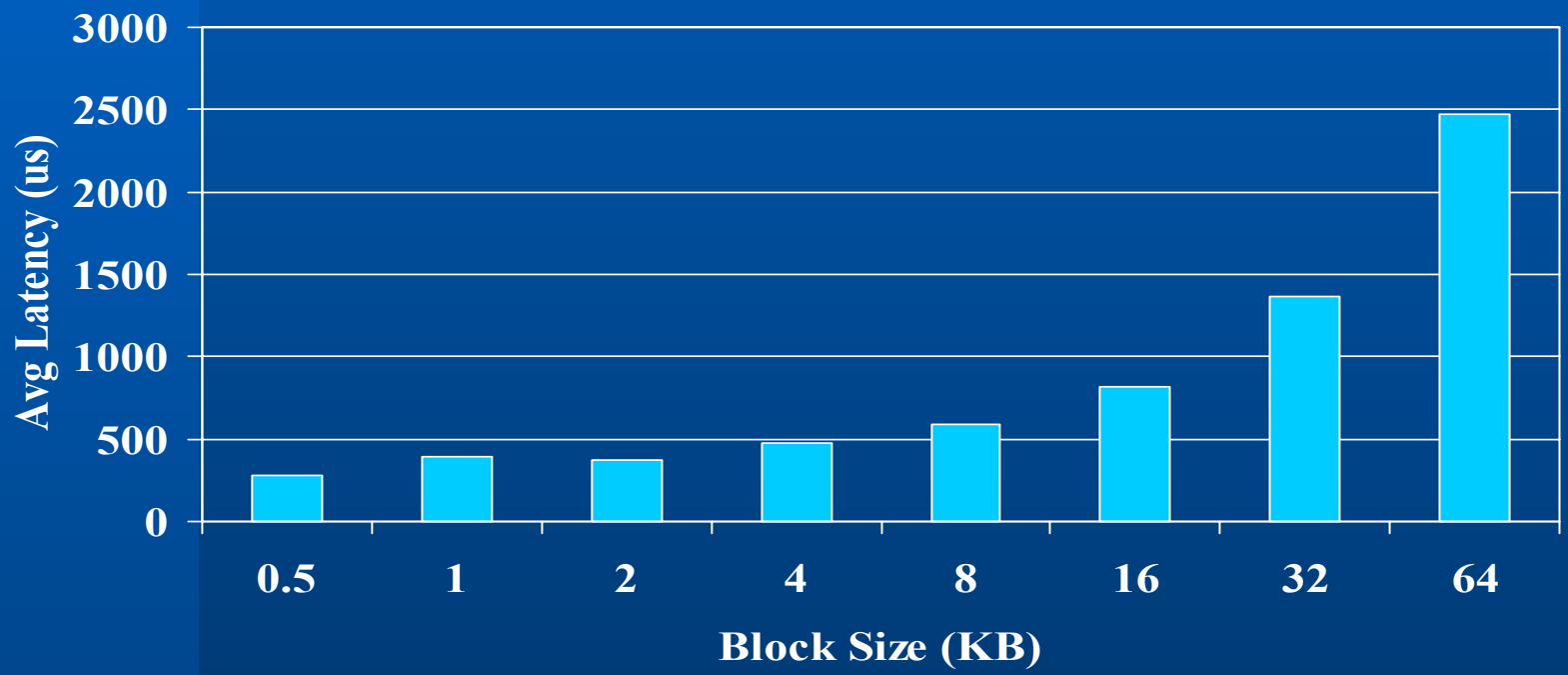


800 MHz PIII
Alteon NIC
iSCSI initiator

Alteon Gigabit
Ethernet switch

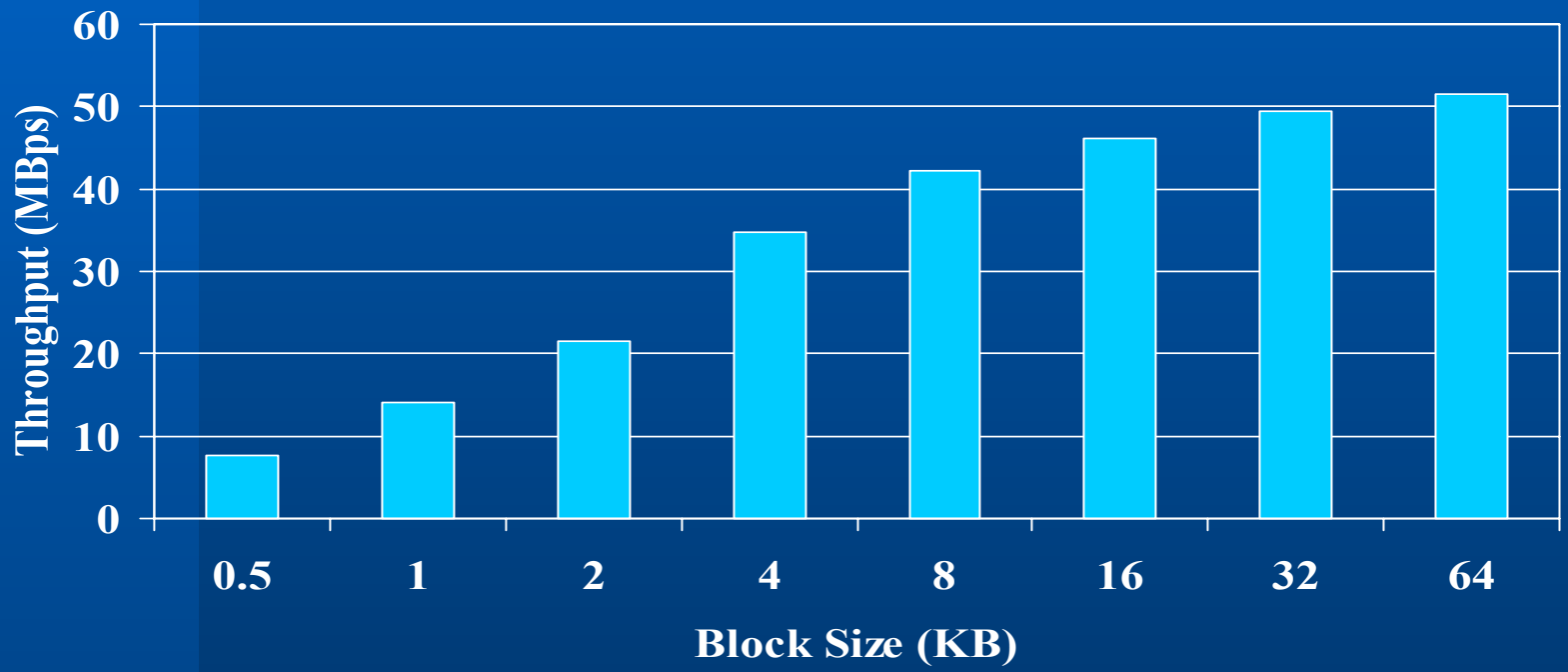
Dual 733 MHz PIII
Alteon NIC
iSCSI target

Latency Measurements



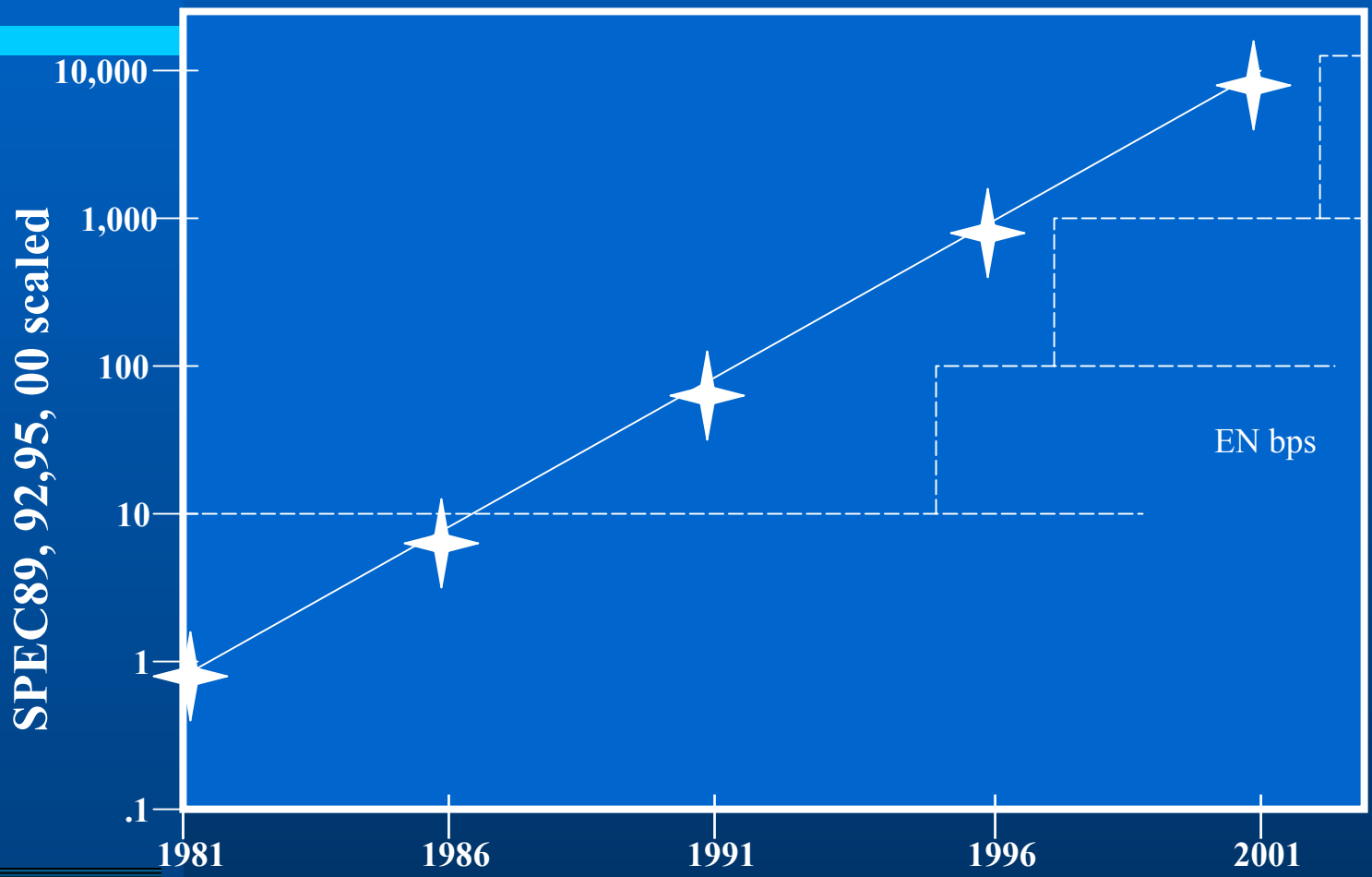
Excluding copy-checksum, latency 5% within FC

Throughput Measurements



Initiator CPU bottleneck: interrupt overhead

CPU versus Network Speeds



Software Optimizations

- **Jumbo Frames**
  - **Reduce interrupt overhead by 60%**
  - **Doubles throughput in some cases**
  - **Non-standard, not present in 10G**
- **Zero-copy TCP/IP stacks**
  - **Requires adapter checksum support**
  - **No RDMA capability**

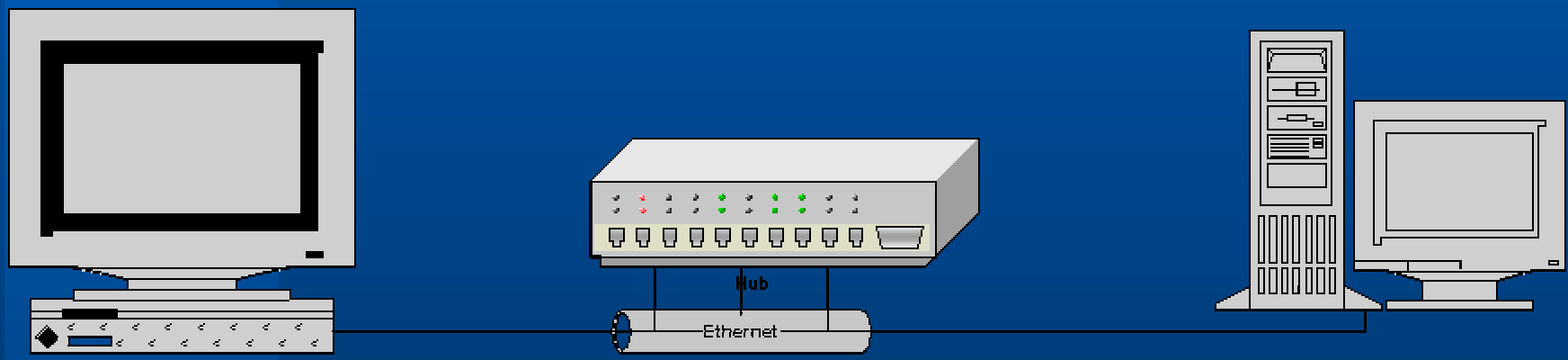
Advanced features costly

Hardware Optimizations

- **TOE**
  - **Minimizes interrupt overhead**
  - **No RDMA support**
  - **No advanced feature support**
- **HBA**
  - **Adds RMDA support + advanced features**
  - **No longer commodity**

Experimental Setup

Benchmark: 100% read cache hit



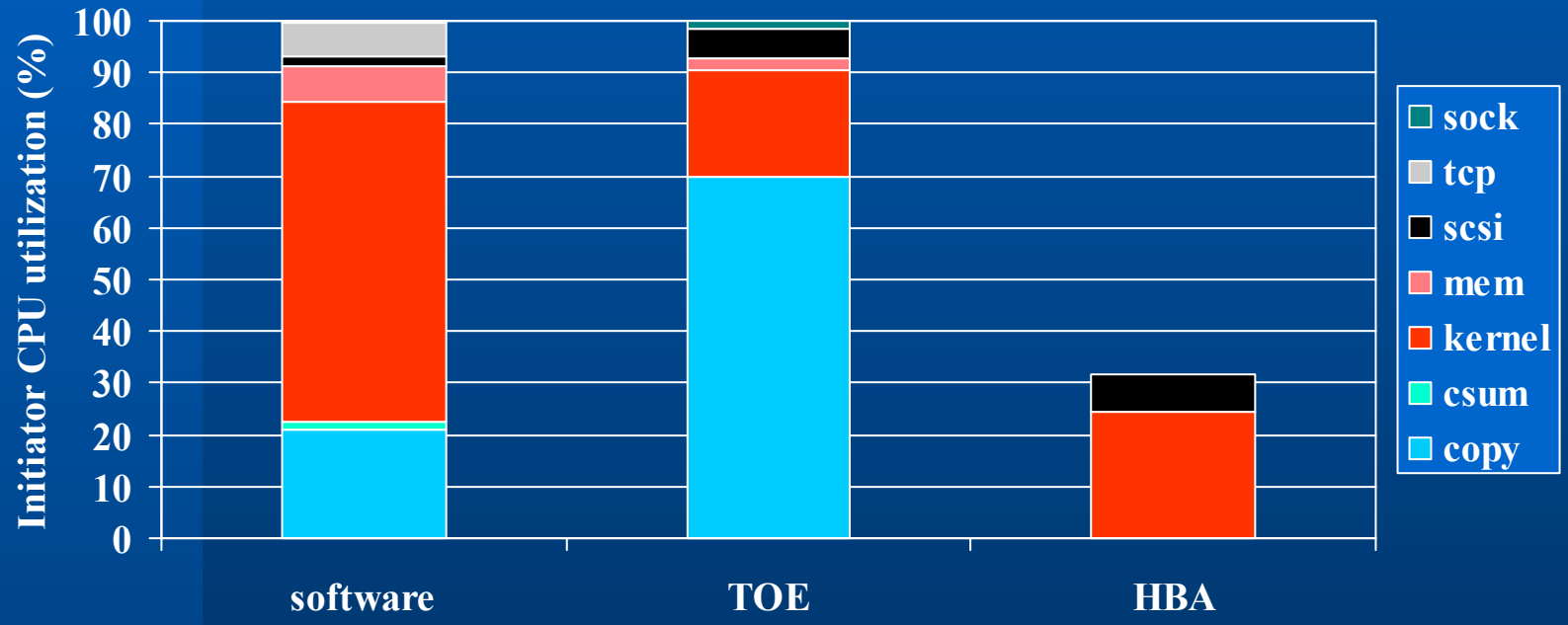
450 MHz PIII
HBA/TOE/NIC
iSCSI initiator

Alteon Gigabit
Ethernet switch

450 MHz PIII
HBA/NIC/TOE
iSCSI target

HBA & TOE Performance *projected*

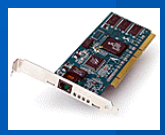
Mbps	26	85	104



TOE offloads copy-checksum, HBA adds RDMA

Matches Alacritech results

Wire-speed iSCSI: Alacritech, Nishan, Hitachi
http://www.alacritech.com/iscsi/pr



Alacritech
Server and Storage Accelerator

Server

iSCSI / Gigabit Ethernet

Nishan
IP Storage Switch

Switch

Fibre
Channel

HITACHI
DATA SYSTEMS

Freedom Storage™
Fibre Channel
Storage System

Additional
Fibre Channel
Drives (JBODs)

Storage

Bi-directional Throughput of 219.64 MBps

Future

- **IOPS also matters**
  - **4-8KB block xfers**
- **Security overhead**