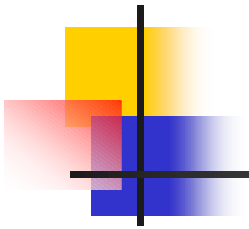
VIA and InfiniBand: Interconnects for High-Performance Computing

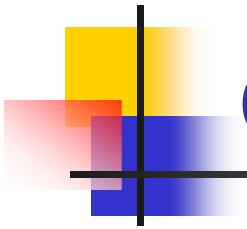
Odysseas Pentakalos, Ph.D.

odysseas@sysnetint.com



Introduction

- Moore's Law states that processing power doubles every eighteen months
- Network transfer rates have been increasing by orders of magnitude
- Most applications have some parallel component
- Clusters of COTS computers are increasingly the more appropriate solution
- Current shared bus technologies cannot keep up with increasing demands on data transfer rates

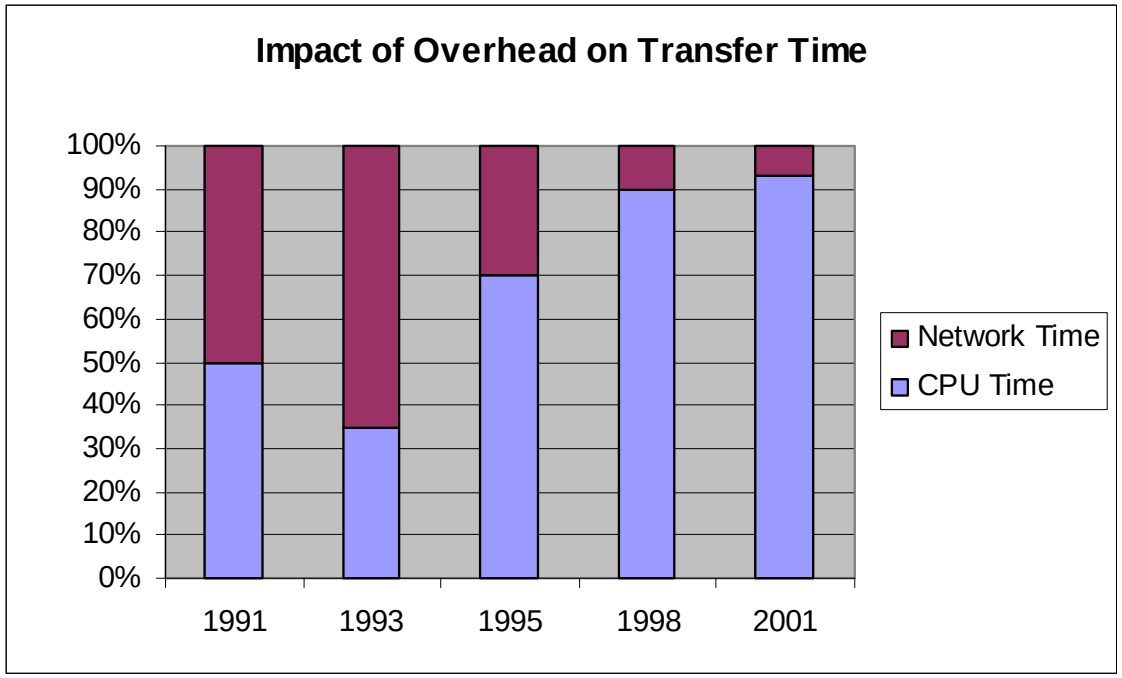


Outline

- Motivation
- Architectures for parallel computing
  - Shared memory
  - Distributed memory
  - Message passing
- Virtual Interface Architecture
  - Architectural stack
  - Operations
  - Applications
- InfiniBand
  - Architecture
  - Operations
- Future

Motivation

- CPU Processing Power increases have not kept up with Network Transfer rate increases
- TCP/IP may be too high-end a solution for the problem at hand



From “The Virtual Interface Architecture” by Don Cameron and Greg Regnier

Motivation

- As disparity between CPU processing time and network transfer time increases so will the overhead

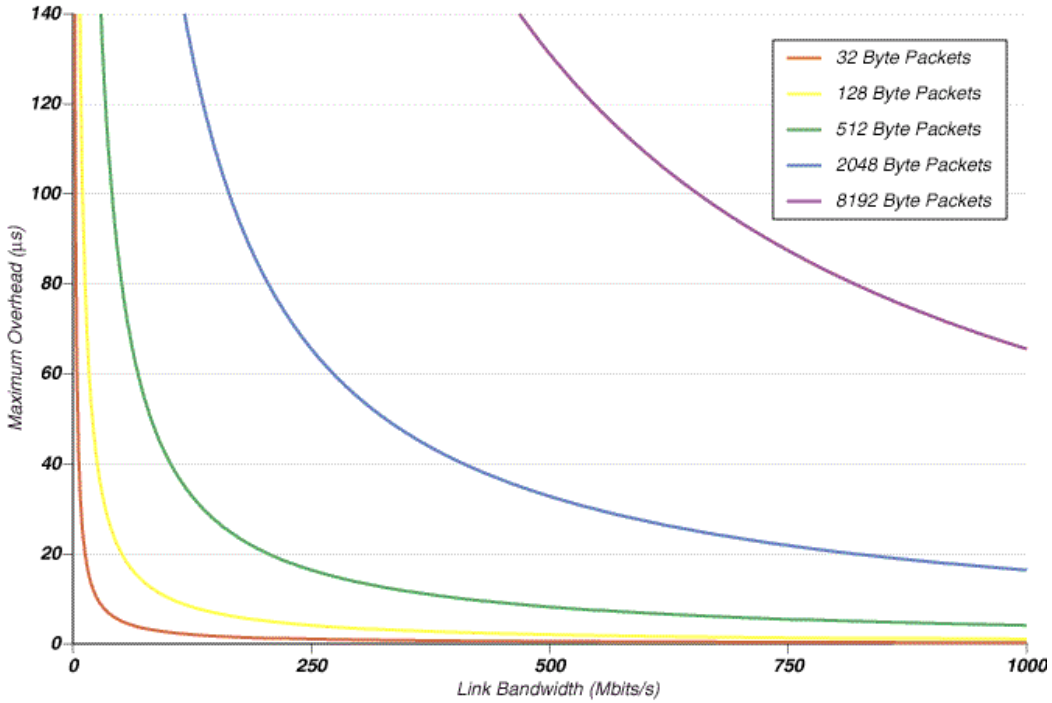


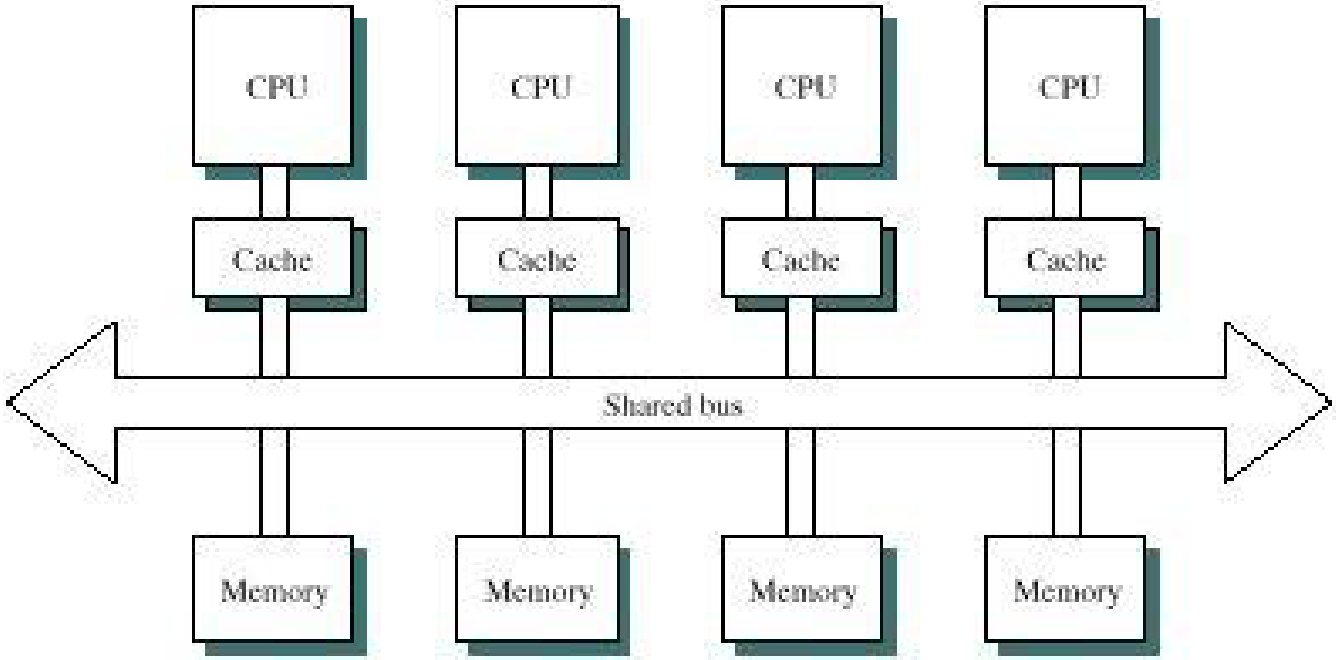
Figure 1: Maximum Allowable Overhead to Achieve a Throughput of One-Half the Link Rate For a Range of Average Message Sizes

From "An Implementation and Analysis of the Virtual Interface Architecture" by *Philip Buonadonna*, *Andrew Geweke*, and *David Culler* , Computer Science Department, UC Berkeley.

Potential High Performance Architectures

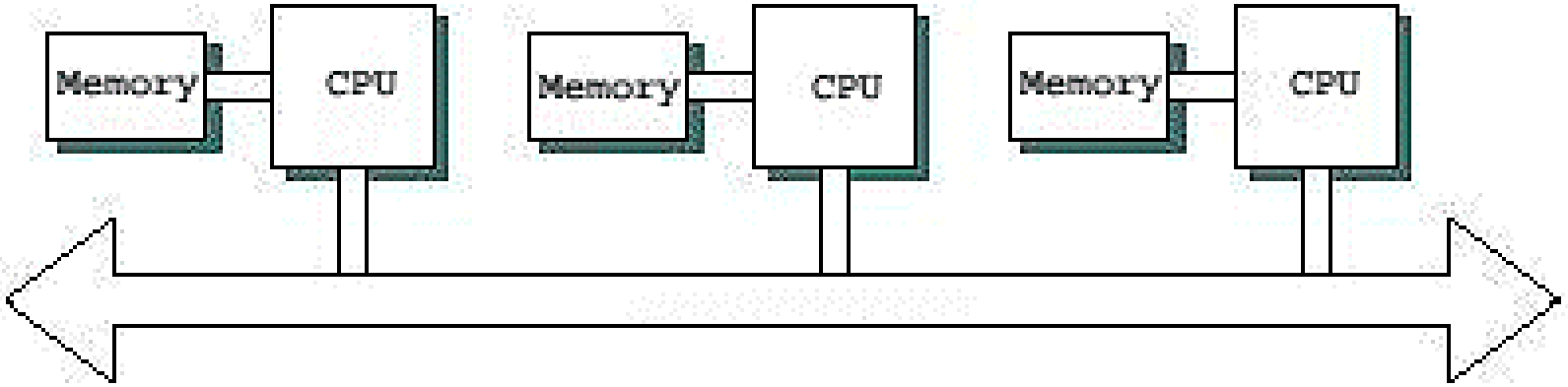
Shared Memory Architecture

- Symmetric Multiprocessing (SMP) share OS, memory and I/O bus



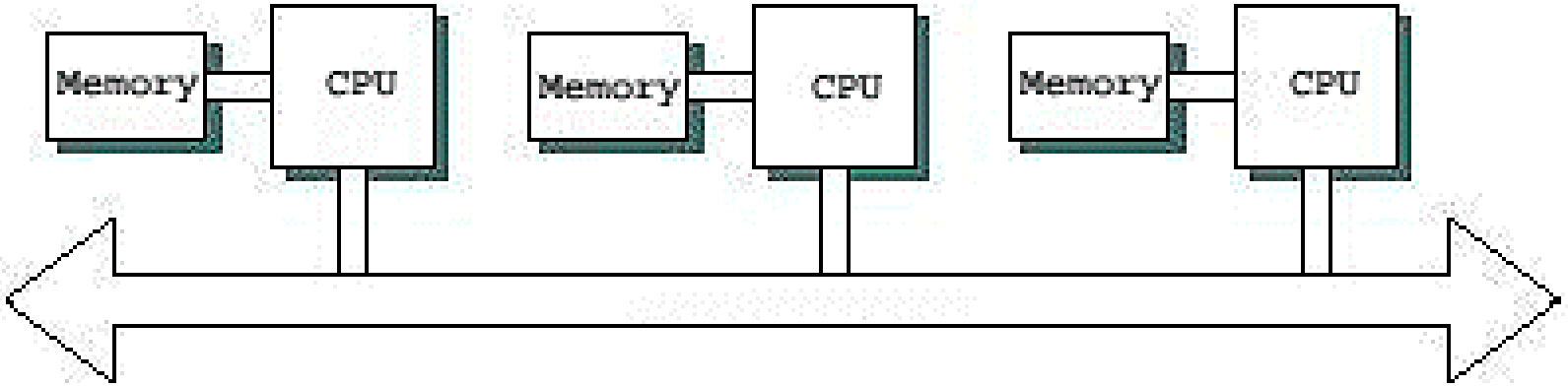
Distributed Shared Memory

- Provides the single shared memory abstraction on a physically distributed architecture



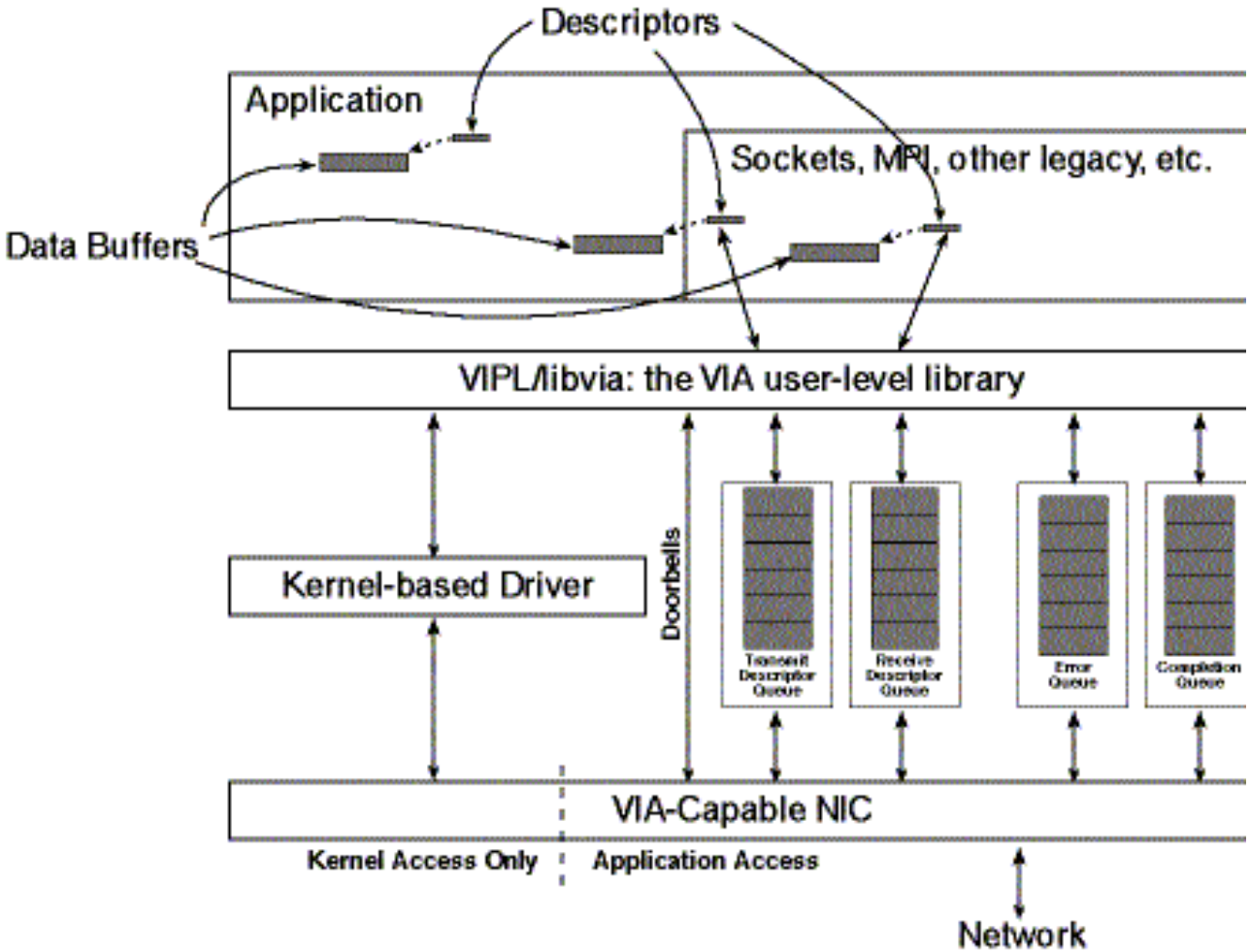
Message Passing

- Same physical architecture but the shared memory abstraction is no longer there

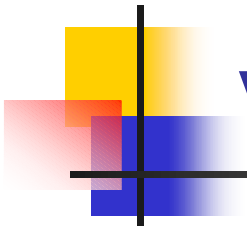


Virtual Interface Architecture

Virtual Interface Architecture

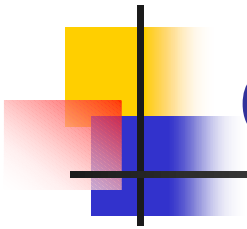


From "An Implementation and Analysis of the Virtual Interface Architecture" by *Philip Buonadonna*, *Andrew Geweke*, and *David Culler* , Computer Science Department, UC Berkeley.



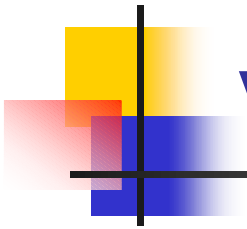
VIA Operations

- **Send/Receive**: transfers sequence of bytes using scatter/gather capabilities between applications
- **RDMA-Write**: copies data to a remote buffer using zero-copy semantics. Supports gather but not scatter semantics
- **RDMA-Read**: reads data from a remote buffer using zero-copy semantics. Supports scatter but not gather semantics.



Other VIA Concepts

- Descriptors are used for specifying the operation to be performed
- Work Queues are used for queuing operations
- Doorbells are used for notifying the VI NIC that work is available
- Work Queue Completion
- Memory registration operations



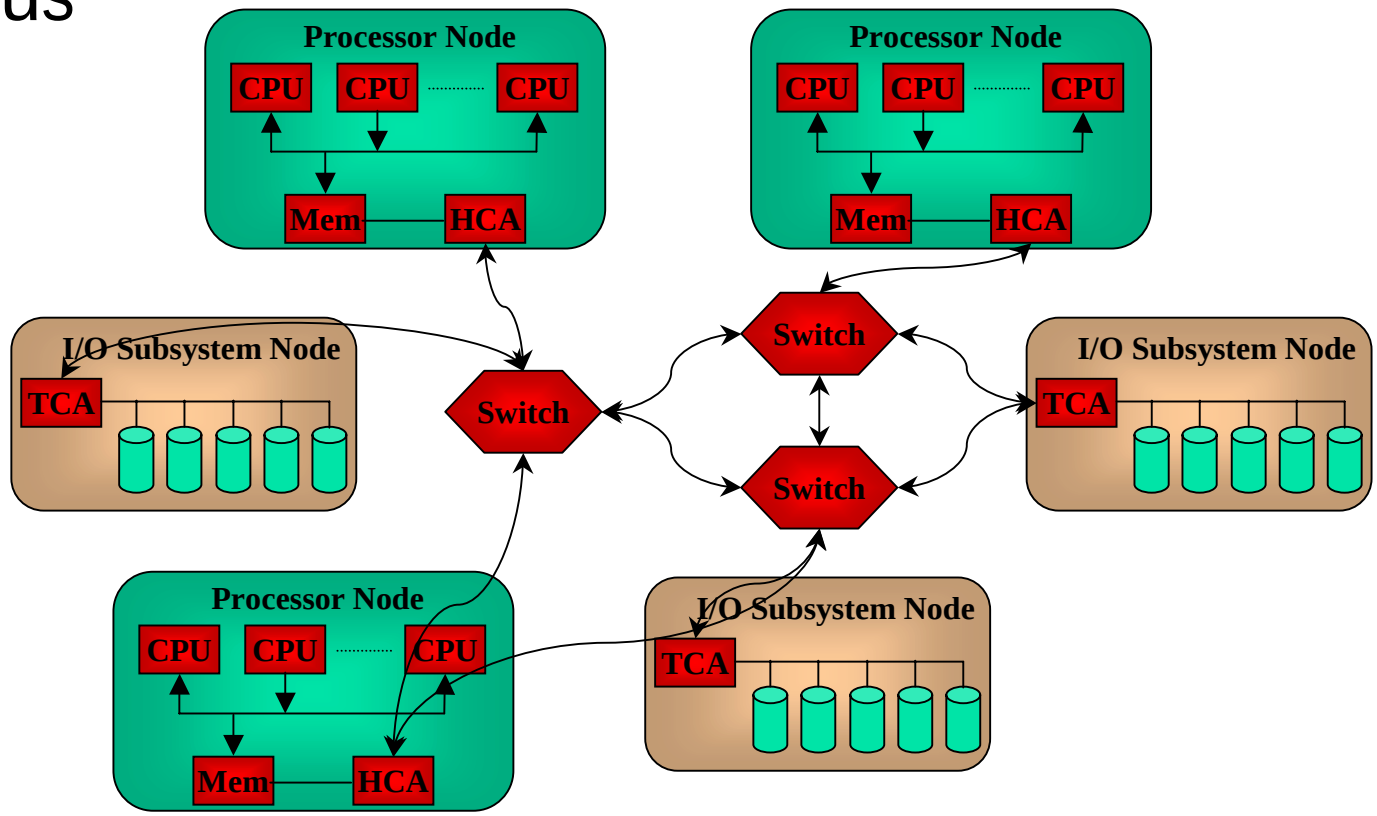
VIA Applications

- IBM DB2 Universal Database - Enterprise Edition
- Microsoft SQL Server Enterprise Edition on Windows 2000 Datacenter Server
- Microsoft Winsock Direct
- DAFS Filesystem
- FC-VI – VI Architecture over Fibre Channel (supported by Emulex and Qlogic)

InfiniBand Architecture

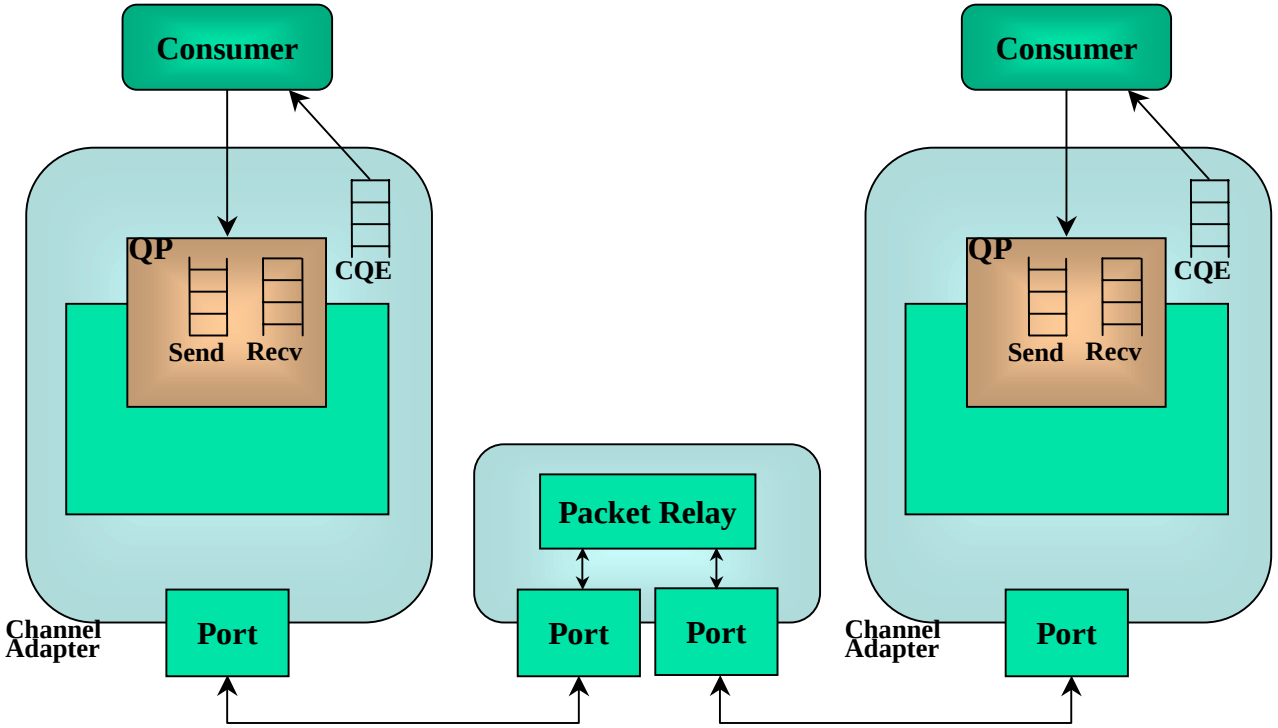
InfiniBand Architecture

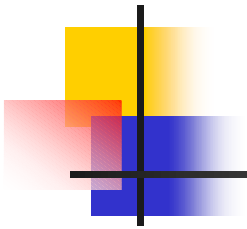
- Introduces a high-speed fabric in place of a PCI-bus



InfiniBand Architecture

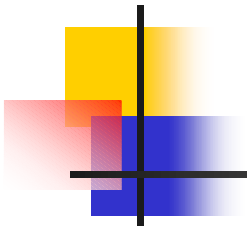
- Borrows heavily from VIA to provide a low-latency interconnect





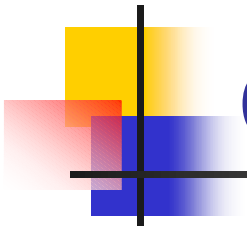
InfiniBand Operations

- Send/Receive
- RDMA-Write
- RDMA-Read
- RDMA-Atomics: Provides two additional operations for synchronization: Compare & Swap and Fetch-Add.



InfiniBand Applications

- SRP – SCSI RDMA protocol
- DAFS – Direct Access File System
- SDP – Socket Direct Protocol
- IPoIB – IP over InfiniBand



Conclusion

Highly Recommended References:

- “The Virtual Interface Architecture” by Don Cameron and Greg Regnier; excellent reference on everything about VIA
- “InfiniBand Architecture: Development and Deployment” by William T. Futral; excellent reference on the InfiniBand