SCI Europe'98

EMMSEC 98: European Multimedia, Microprocessor Systems and Electronic Commerce Conference and Exposition

A Comparison of Two Gigabit SAN/LAN Technologies: SCI versus Myrinet

Ch. Kurmann, T. Stricker

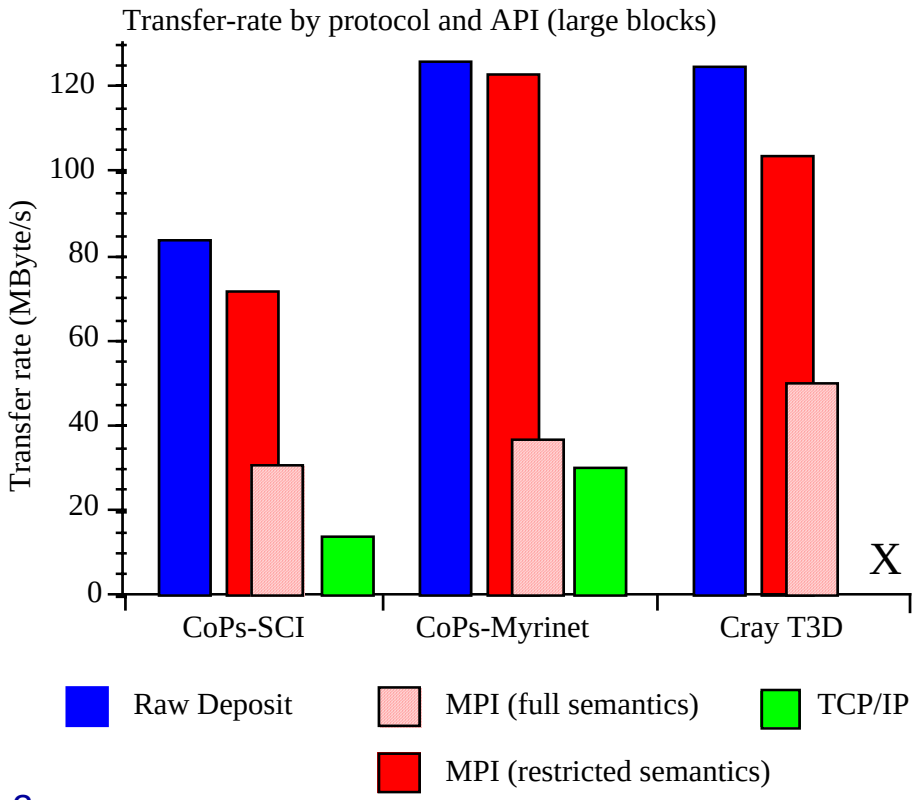
Laboratory for Computer Systems
ETHZ - Swiss Institute of Technology
CH-8092 Zurich

Eidgenössische Technische Hochschule Zürich

Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo
Swiss Federal Institute of Technology Zurich

Motivation

- Evaluation and comparison of Gigabit/sec interconnects need a *common architectural denominator*
- We propose three different levels:
  - highly optimized remote load/store operation
  - optimized standard message passing library
  - connection oriented LAN emulation



Overview

- Levels of comparison
- Previous work
- Technologies overview:
  - PC Platform
  - Myricom Myrinet
  - Dolphin CluStar SCI
  - SGI / Cray T3D
- Typical transfer modes
- Measurement results
- Conclusion

Levels of Comparison

- Three levels with different amount of support by the operating system:
  - DIRECT DEPOSIT:
    - simple remote load/store operation
    - performance is expected to be closest at actual hardware peak performance
  - MPI/PVM:
    - optimized standard message passing library
    - carefully coded parallel applications are expected to see this performance
  - TCP/IP:
    - connection oriented TCP/IP LAN emulation
  - ....
- Common architectural denominator

Previous Work

- Previous studies:
  - maximum bandwidth numbers
  - minimal latency numbers
  - performance results for an entire application
- Performance of application depends:
  - redistribution of data stored in distributed arrays
  - migration of data in fine grain object store
- We need a benchmark that covers data types beyond contiguous blocks of data (e.g. strided remote stores).

Direct Deposit

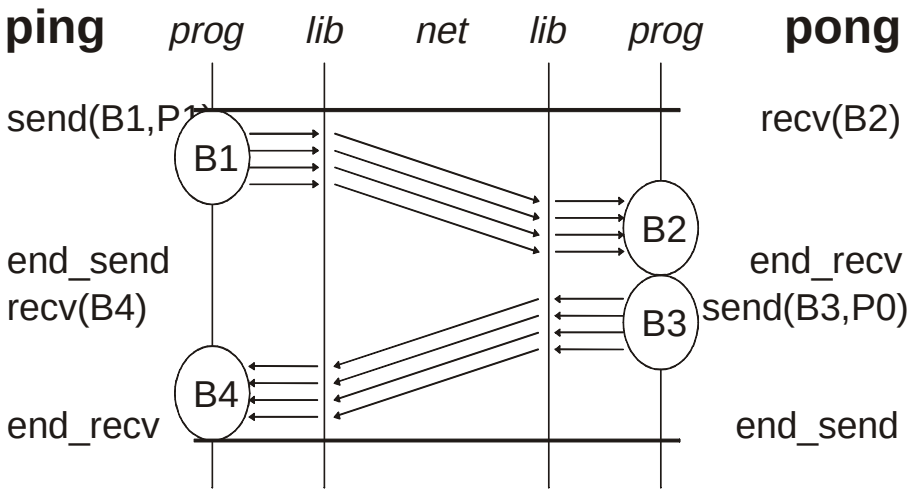
- The deposit model requires a clean separation and different mechanisms for:
  - control messages,
  - data messages.
- Data is “dropped” directly into the receivers address space by the hardware without active participation of the receiver process.
- Allows to copy fine grained data involving complex access patterns like strides.

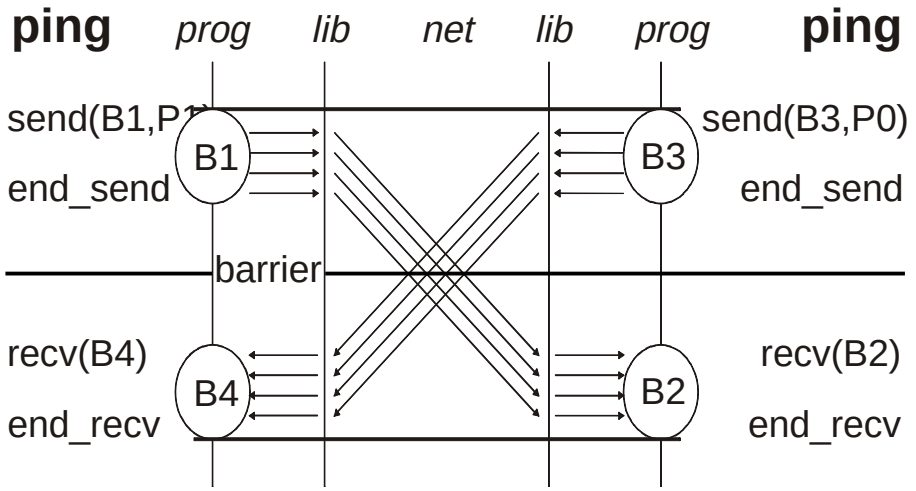
Message Passing Libraries

- Sender can send messages at any time, without waiting for the receiver to be ready
- Buffering is often done at a higher level and involves the memory system of the end-points
- Fine grain data accesses are implemented through buffer-packing / -unpacking

Message Passing Model

- Different flavors for restricted and full postal semantics
  - non-buffering semantics: can be mapped directly to a fast direct deposit including synchronization
  - buffering semantics: non-blocking operation allows sending at any time and leads to an additional copy operation





Protocol Emulation

- Popular API → much software
  - UDP/IP - unreliable, connectionless network service
  - TCP/IP - allows reliable connection-oriented communication
  - NFS/IP - network file system
- Protocol stacks are provided by the OS
- Socket API, streams API are ubiquitous
- It is unrealistic to recode all commercial web servers, databases or middleware systems for message passing APIs like MPI.
- With IP support gigabit networks can speedup much more than just scientific applications!

PC Platform for this Talk

- Single/Twin Pentium Pro 200MHz
- Intel 440 FX Chipset
- 64-bit 66 MHz main memory interface, 0.5 GByte/s
- 32-bit 33 MHz PCI bus, 132 MB/s

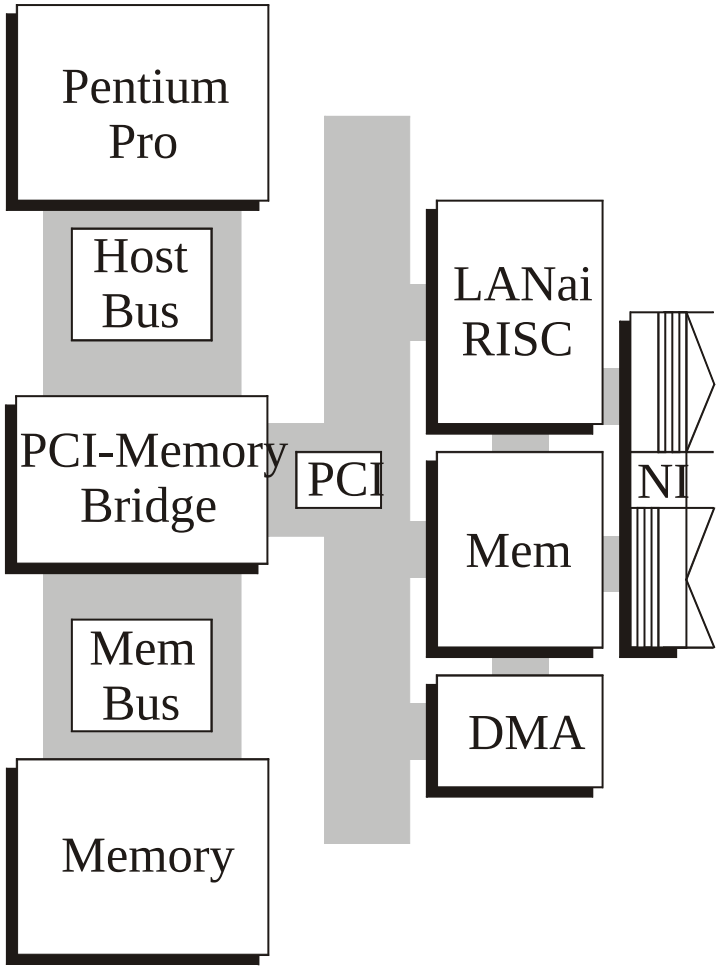
~ 3000 € per node

Myricom Myrinet

- Two 1.28 Gbit/s channels (duplex) connecting hosts and (4, 8, 16-port) switches point-to-point
- Supports any topology with switches, hot configurable
- Wormhole routing with link level flow control guarantees the delivery of messages
- Checksumming for error detection
- Packets of arbitrary length (unlimited MTU)
  → can encapsulate any type of packets

Myricom Myrinet Adapter

- RISC processor (LANai)
- 1MB SRAM to store MCP and to act as staging memory for buffering packets
- Bus Master DMA adapter-to-host (for the PCI)
- Two DMAs between memory and network FIFOs
- Concurrent operation of DMAs



Myrinet Control Program

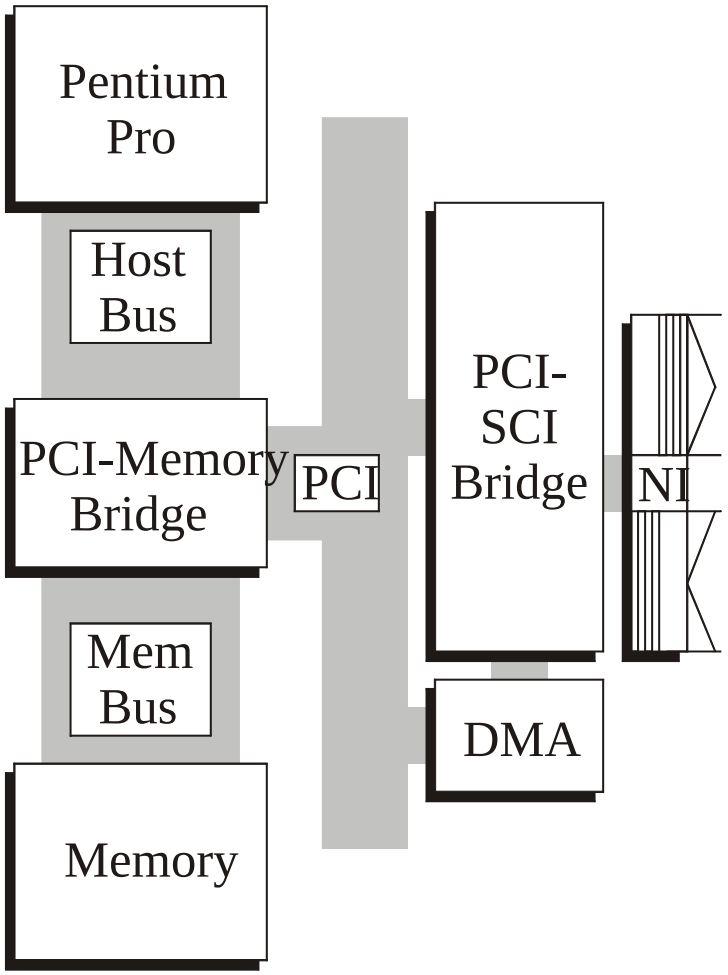
- The LANai is a 32-bit dual-context RISC Processor with 24 general purpose registers that runs the Myrinet Control Program (MCP)
- A typical MCP provides:
  - routing table management,
  - gathering operation,
  - checksumming,
  - send / receive operation,
  - control message generation,
  - scattering operation,
  - interrupts generation upon arrival

Dolphin CluStar SCI

- Two unidirectional 1.6 Gbit/s links (CluStar: 3.2 Gbit/s)
- Multidimensional rings and switched ringlets
- Protocol uses data sizes of 16, 64, 256 Bytes
- Transparent PCI-to-PCI bridge operation through memory mapped load/store interface
- Possibility for fully coherent shared memory on high end implementations beyond PCI products
- Per word remote memory and block transfers for message passing operation

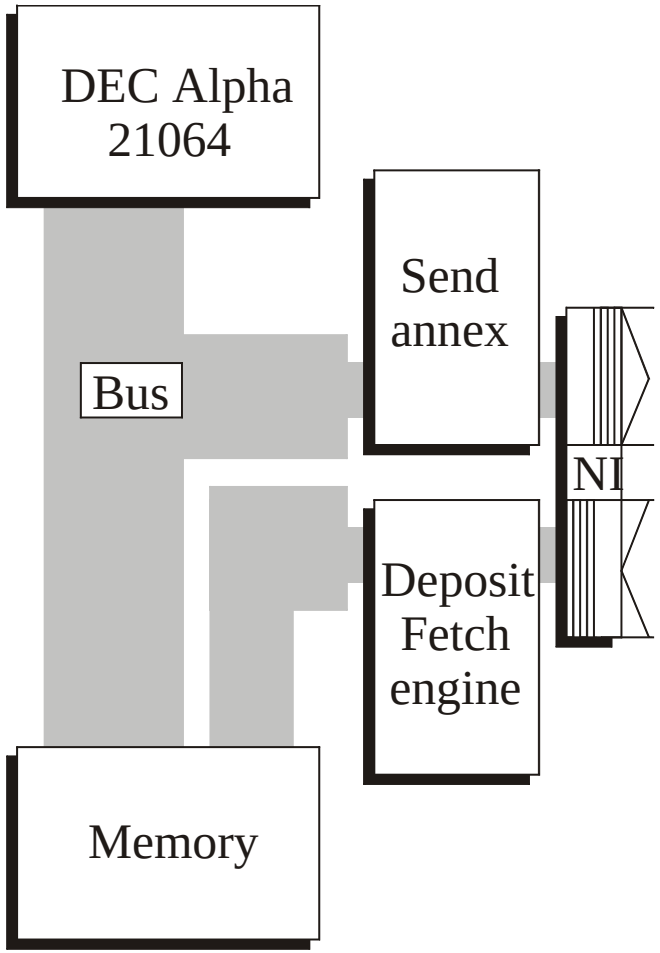
Dolphin CluStar SCI Adapter

- Protocol engine
  - 8 64Byte stream buffers
  - PCI-SCI memory address mapping by ATT
  - Busmaster DMA
- Link controller
  - Contains 3 FIFOs (TX, RX, Transit)
- The PCI-adapter supports a subset of IEEE-SCI without hardware cache coherency



SGI / Cray T3D as Reference Point

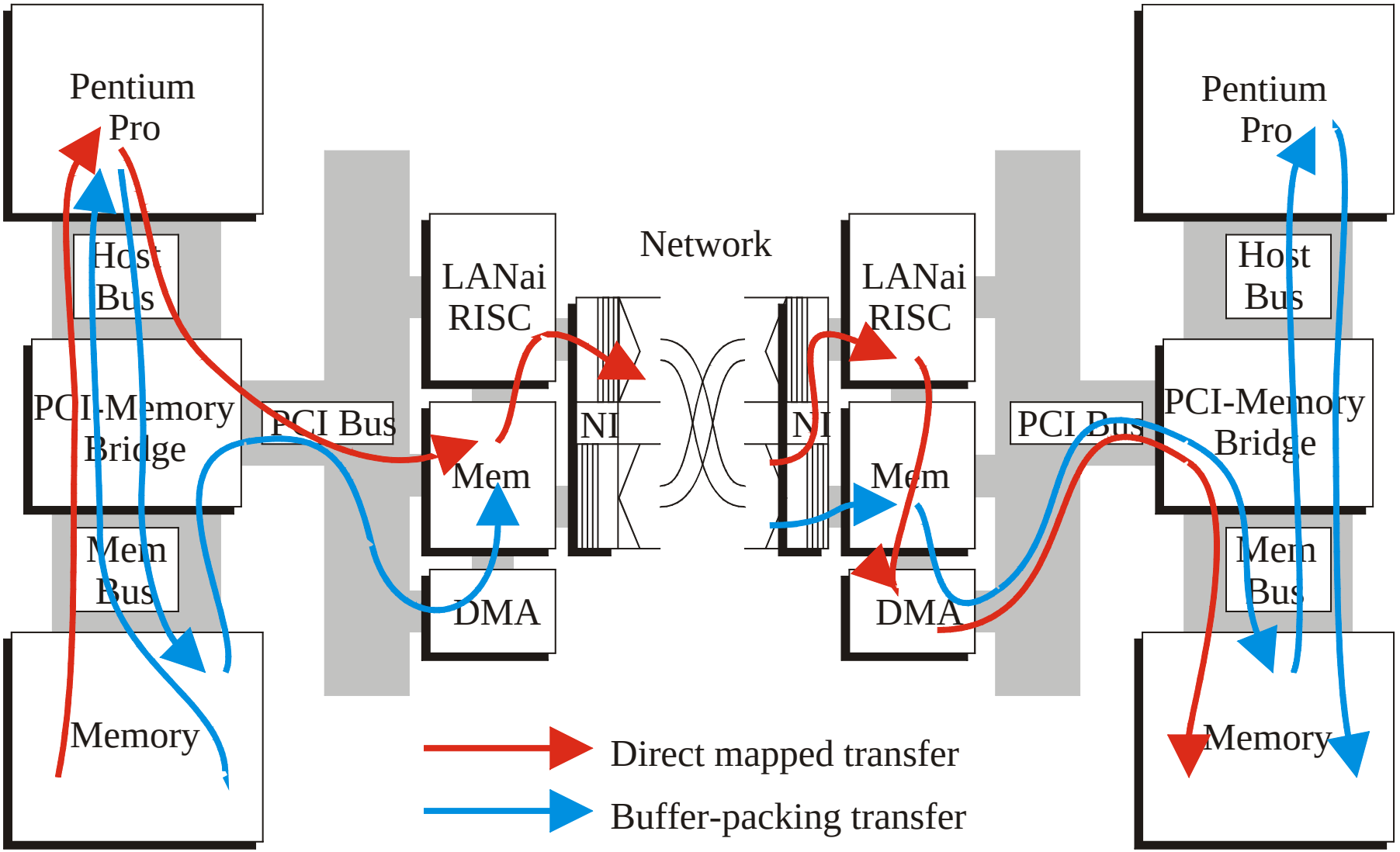
- 150 MHz 64bit DEC Alpha
- No virtual memory
- ca. 1.28 Gbit/link
- 3D torus topology
- Memory mapped network interface to send remote stores
- Fetch/deposit engine with separate memory bus (no involvement of processor)



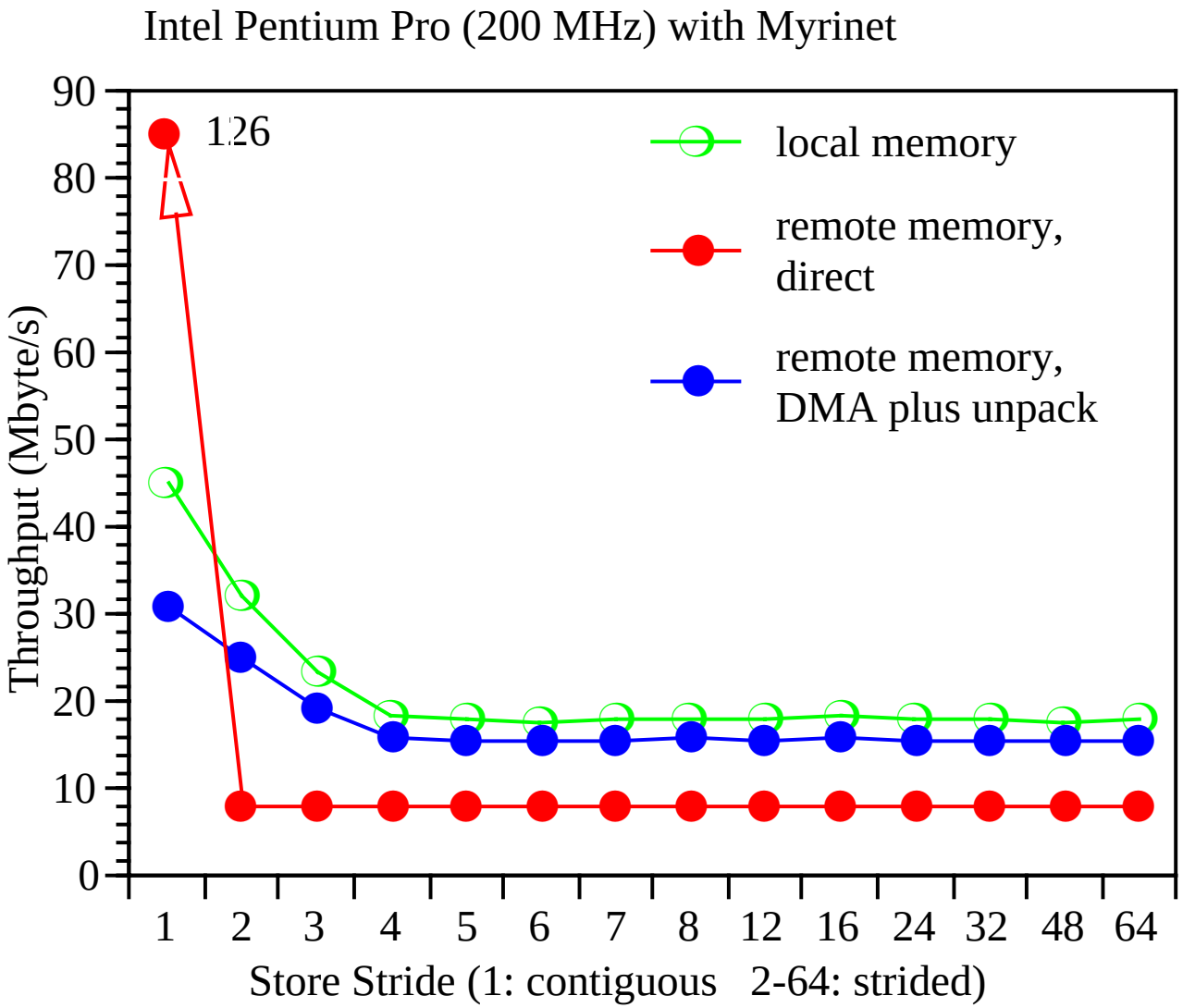
Typical Transfer Modes

- Peak bandwidth for large block transfers (zero-copy)
- Reduced bandwidth for remote memory operation including fine grain accesses to the memory system
- There are two modes for fine grain transfers: processor driven versus DMA driven:
  - Remote loads/stores by either the processor or the DMA (Direct Deposit Model)
  - Buffer-packing/-unpacking at the sender/receiver by either the processor or the DMA (Messaging Model)

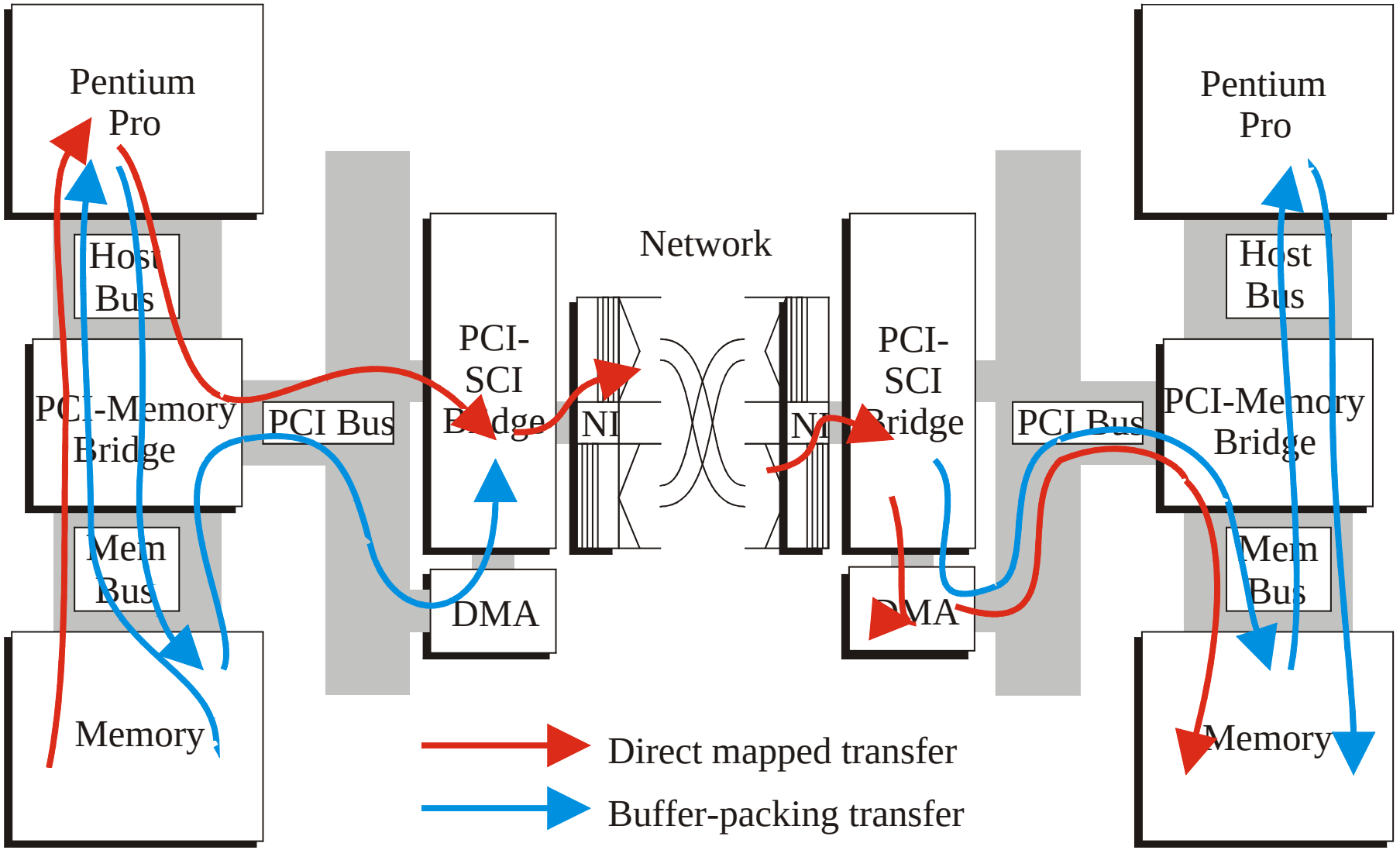
Myricom Myrinet



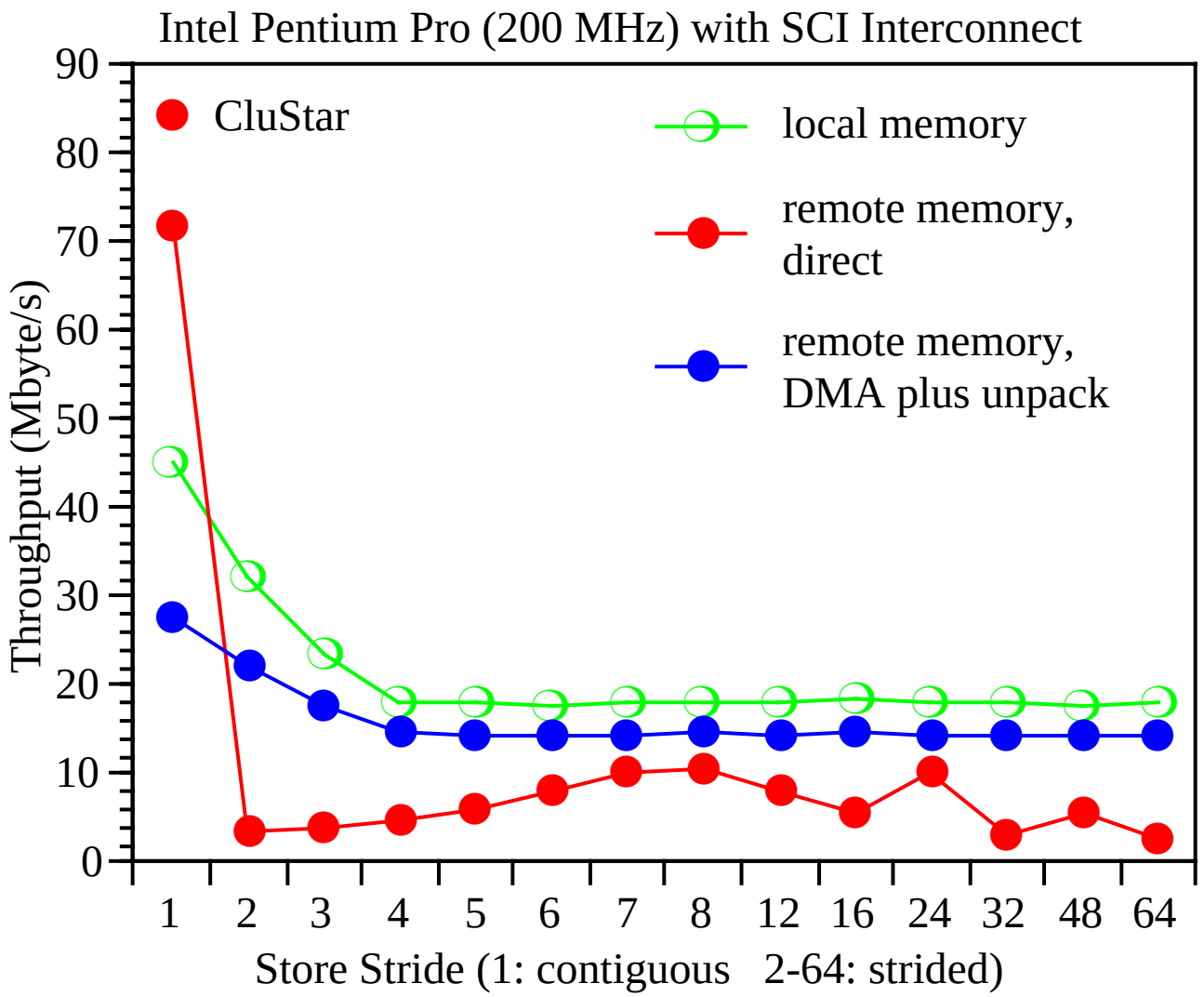
Deposit on Myrinet



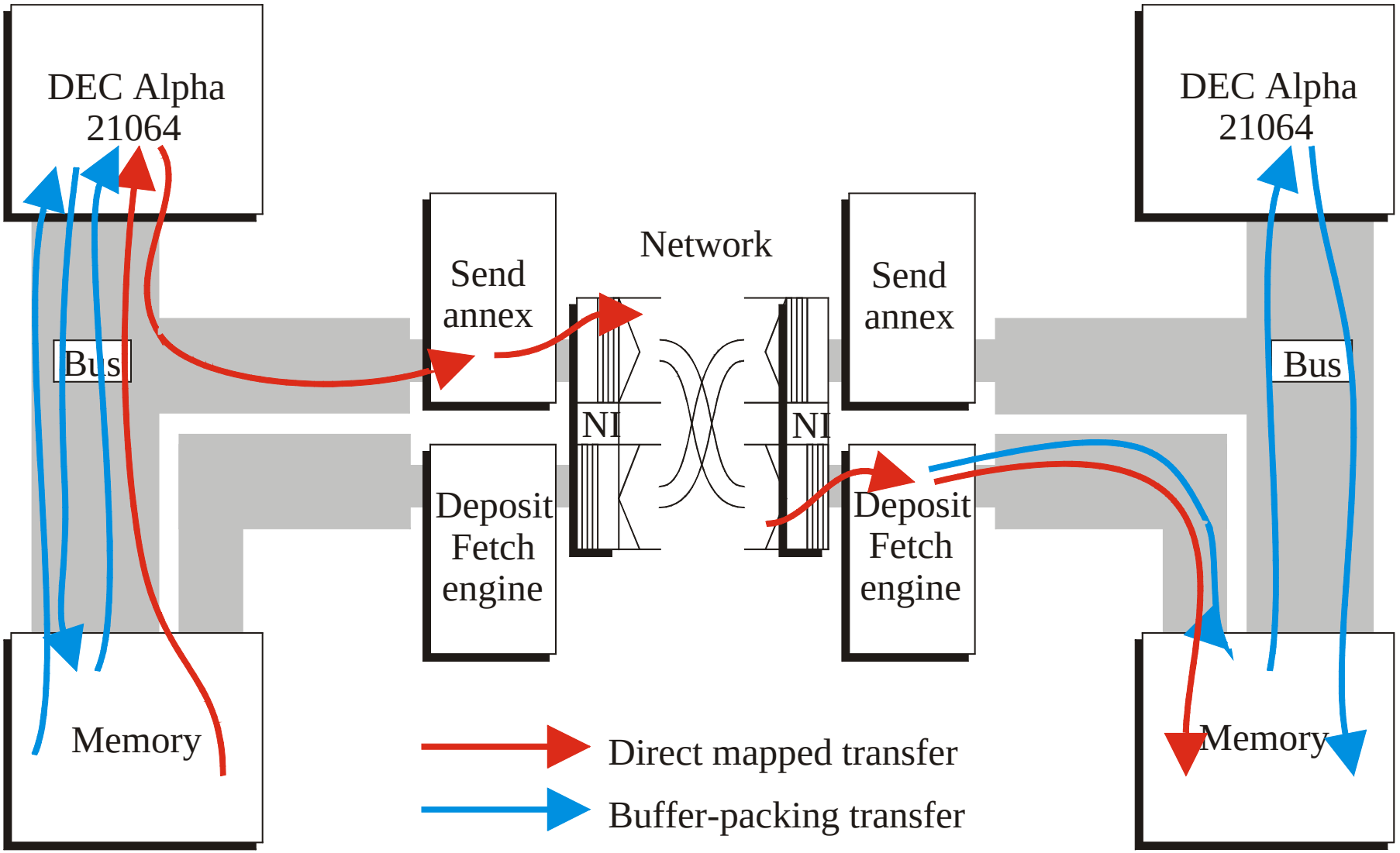
Deposit Dolphin CluStar SCI



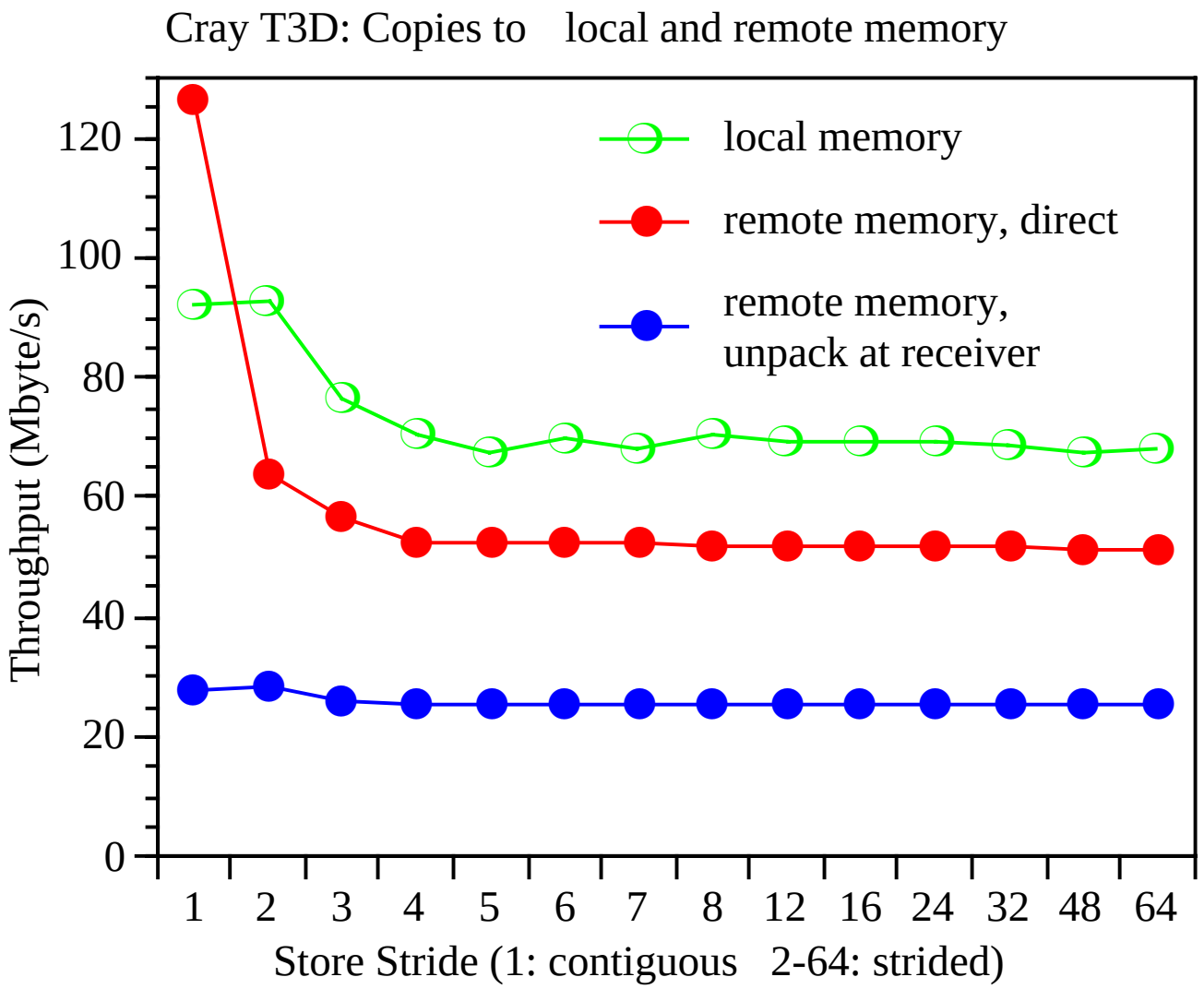
Deposit on SCI



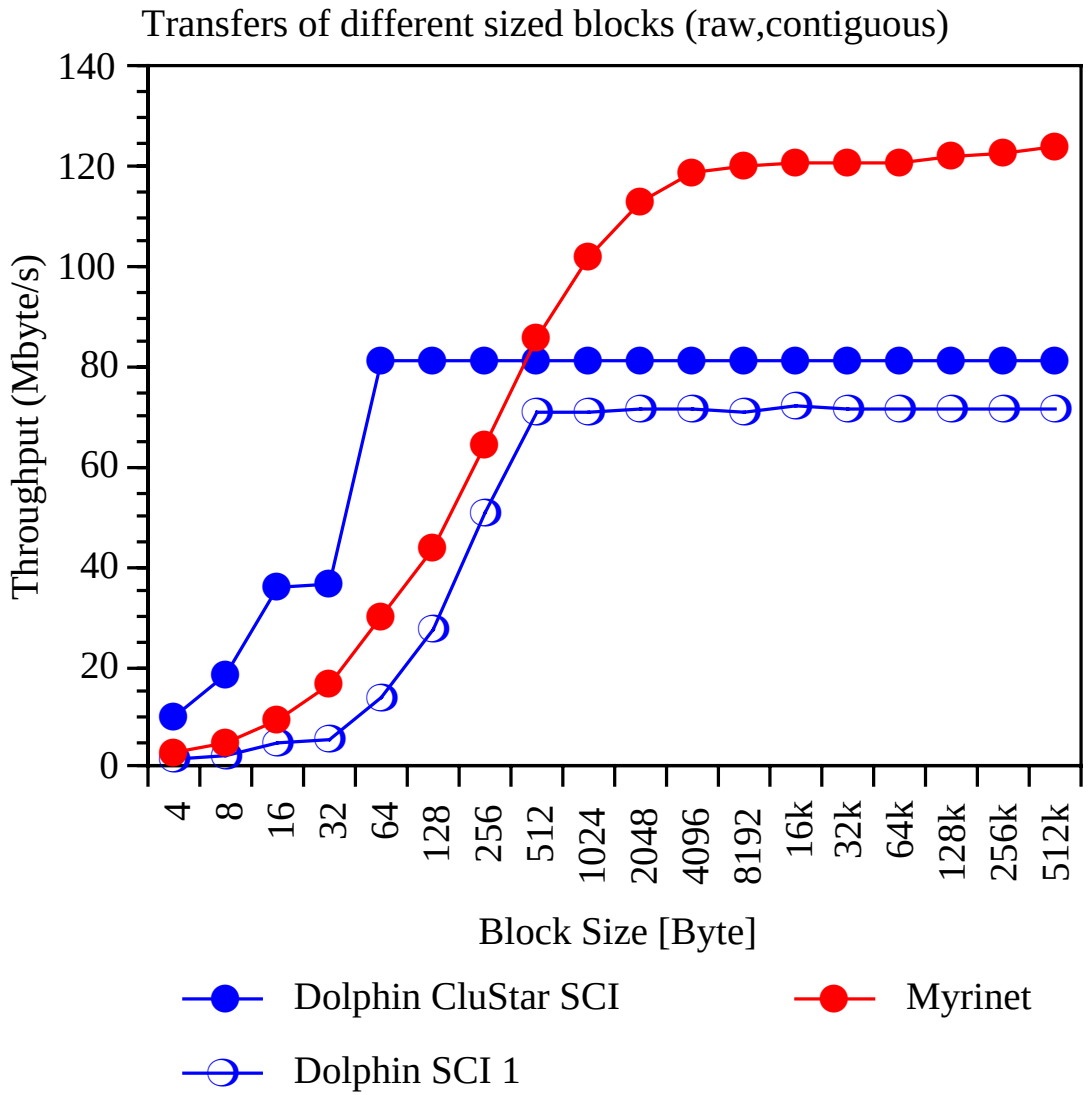
SGI / Cray T3D


DEC Alpha
21064
Bus
Memory
Send
annex
Deposit
Fetch
engine
Network
NI
NI
Send
annex
Deposit
Fetch
engine
DEC Alpha
21064
Bus
Memory
Direct mapped transfer
Buffer-packing transfer

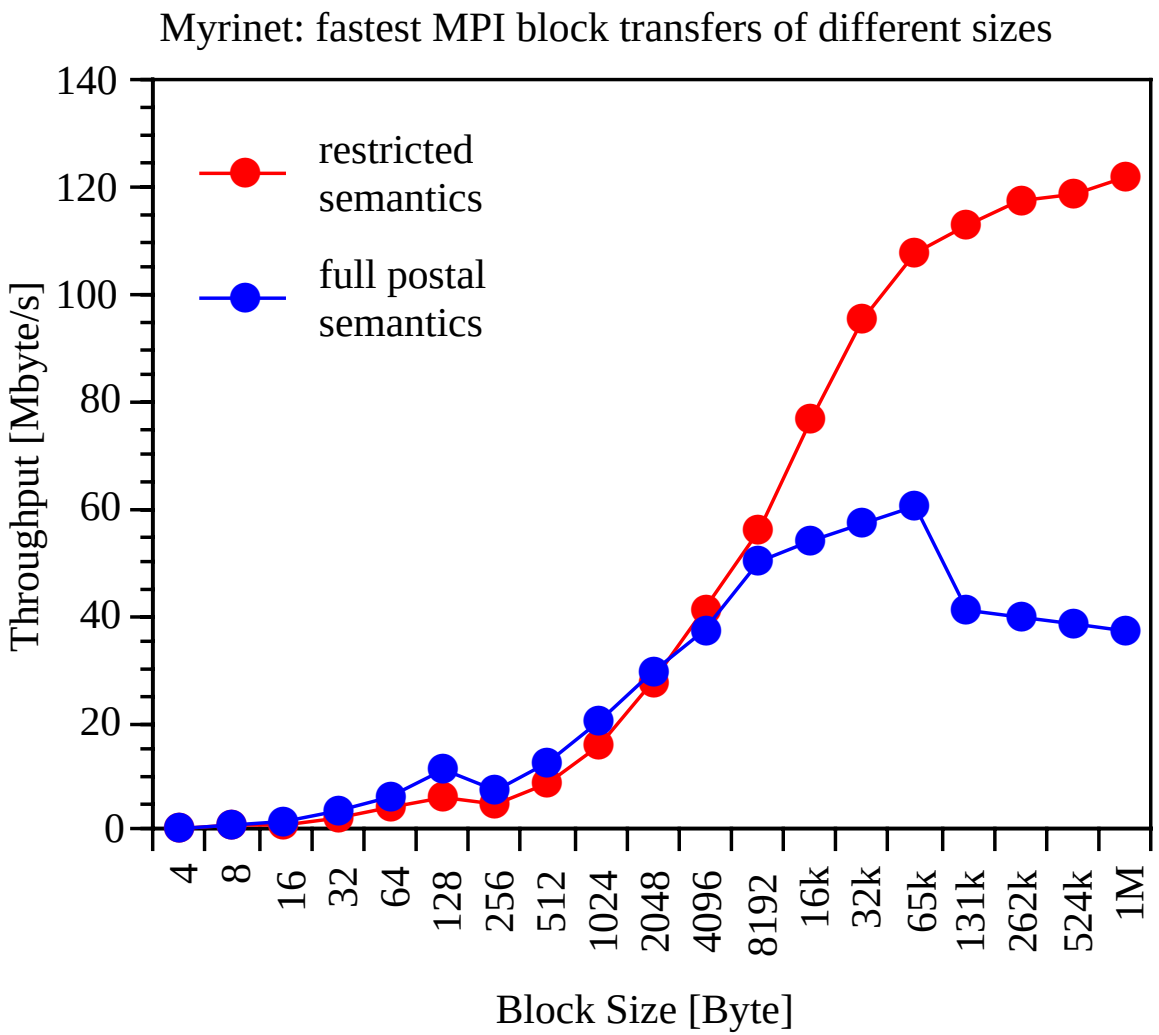
Deposit on SGI / Cray T3D



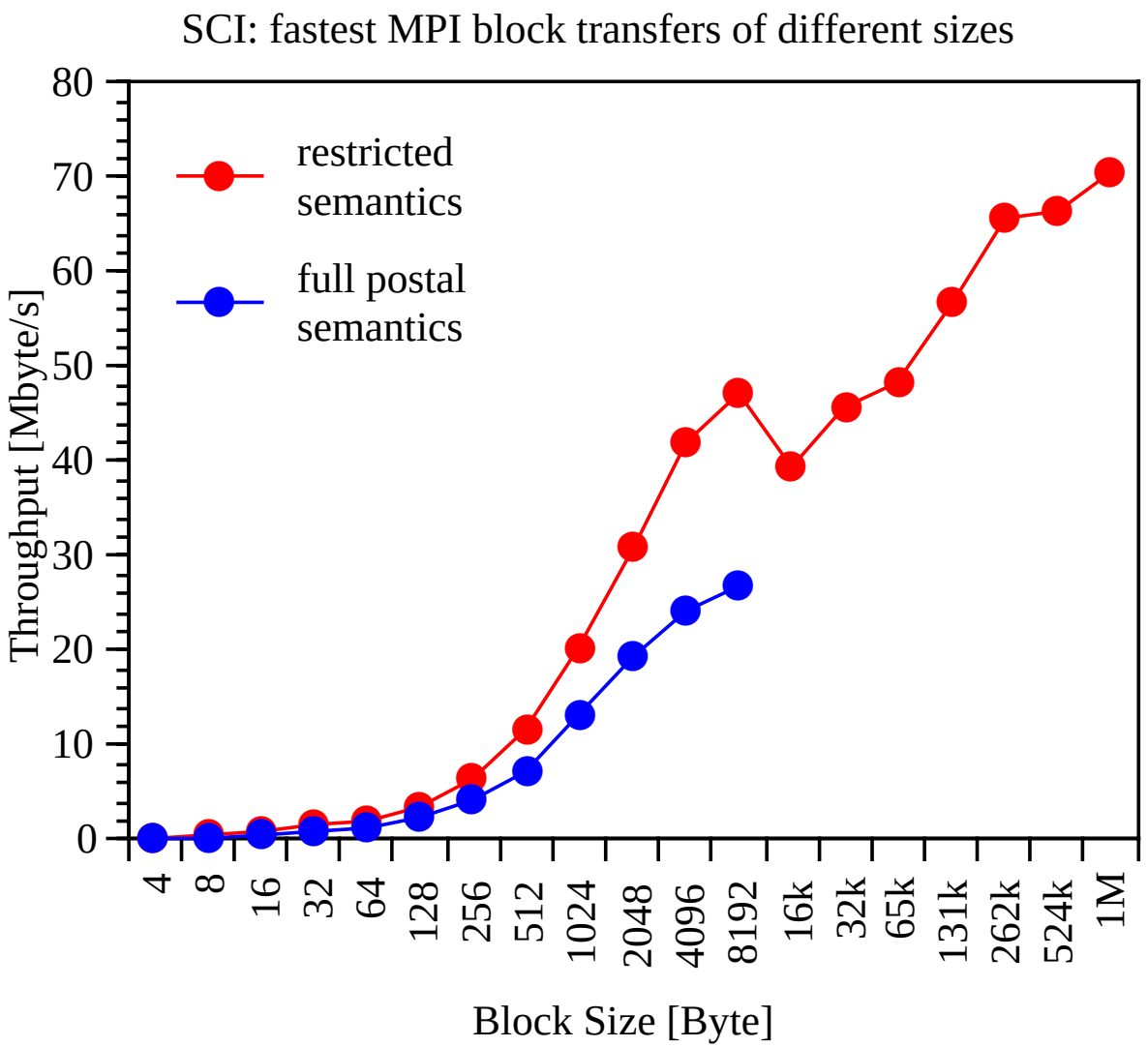
Raw Block Transfers Myrinet, SCI



MPI Transfer Myrinet

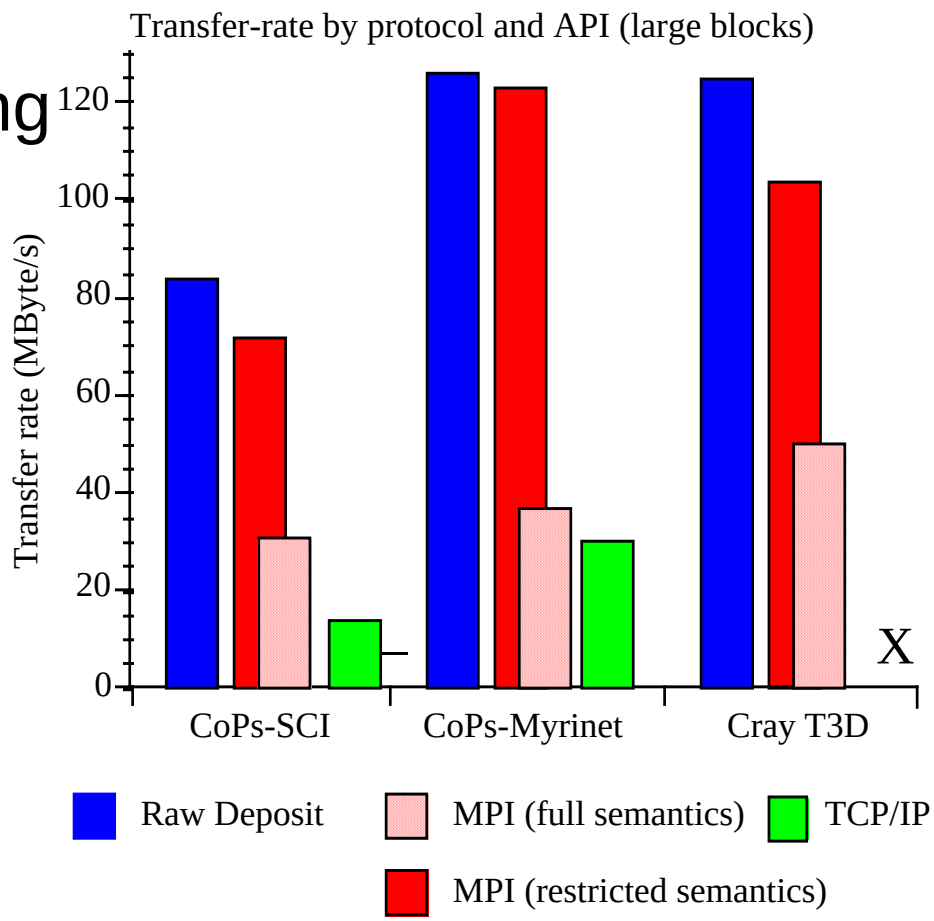


MPI Transfer SCI (Scali MPI)



Summary and Comparison

- Raw deposit bandwidth compared to MPI (blocking and non-blocking)
- Blocking MPI matches direct deposit bandwidth
- Non-blocking calls suffer from buffering
- TCP/IP performance results confirm the overhead of copies



Conclusion

- Three different levels of system software support for Gbit/s- networks permit a good comparison between different networking technologies based on micro-benchmarks.
- SCI, Myrinet and MPPs have excellent performance for contiguous blocks but for strided data the performance of PCI-adapters collapses.
- MPI with buffering semantics suffers from the poor memory copy performance whereas ‘zero copy’ MPI offer better speed.
- PCI card interconnects will get into difficulties with applications that require complex remote memory operations or high-level networking protocols.