Cluster Computers

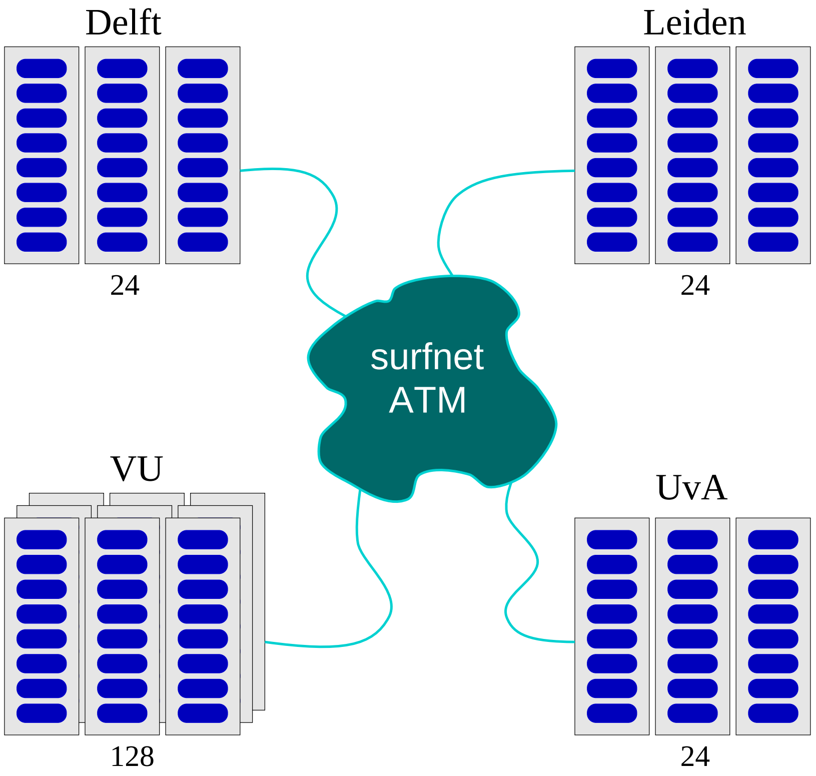
Introduction

- Cluster computing
  - Standard PCs or workstations connected by a fast network
  - Good price/performance ratio
  - Exploit existing (idle) machines or use (new) dedicated machines
- Cluster computers versus supercomputers
  - Processing power is similar: based on microprocessors
  - Communication performance was the key difference
  - Modern networks (Myrinet, ATM, SCI, Servernet) may bridge this gap

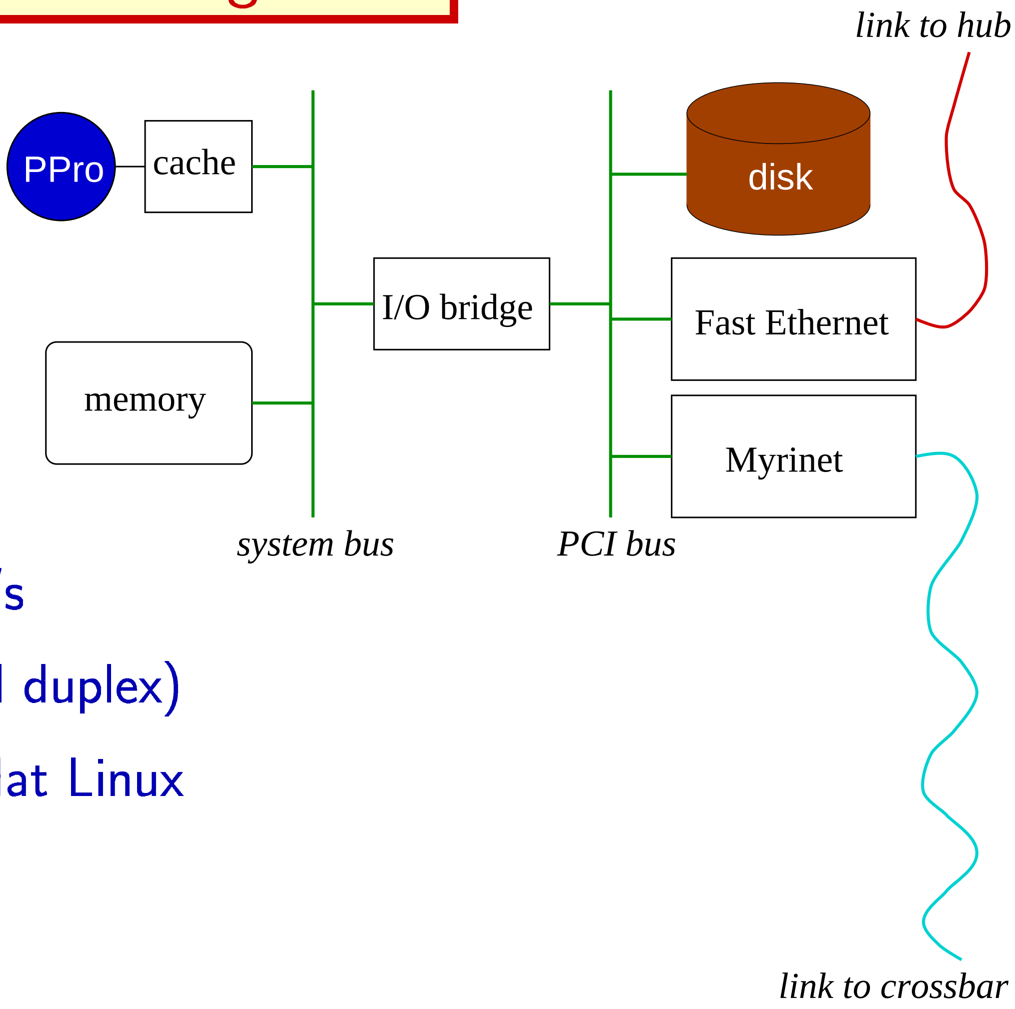
Overview

- Cluster computers at our department
  - 128-node PentiumPro/Myrinet cluster
  - 72-node dual-Pentium-III/Myrinet-2000 cluster
  - Part of a wide-area system: Distributed ASCI Supercomputer
- Network interface protocols for Myrinet
  - Low-level systems software
  - Partly runs on the network interface card (firmware)

Distributed ASCI Supercomputer



Node configuration



- 200 MHz Pentium Pro
- 128 MB memory
- 2.5 GB disk
- Fast Ethernet 100 Mbit/s
- Myrinet 1.28 Gbit/s (full duplex)
- Operating system: RedHat Linux

New DAS-2 cluster

- 72 nodes, each with 2 CPUs (144 CPUs in total)
- 1 GHz Pentium-III
- 1 GB memory per node
- 20 GB disk
- Fast Ethernet 100 Mbit/s
- Myrinet-2000 2 Gbit/s (crossbar)
- Operating system: RedHat Linux
- Part of wide-area DAS-2 system (5 clusters with 200 nodes in total)

Myrinet

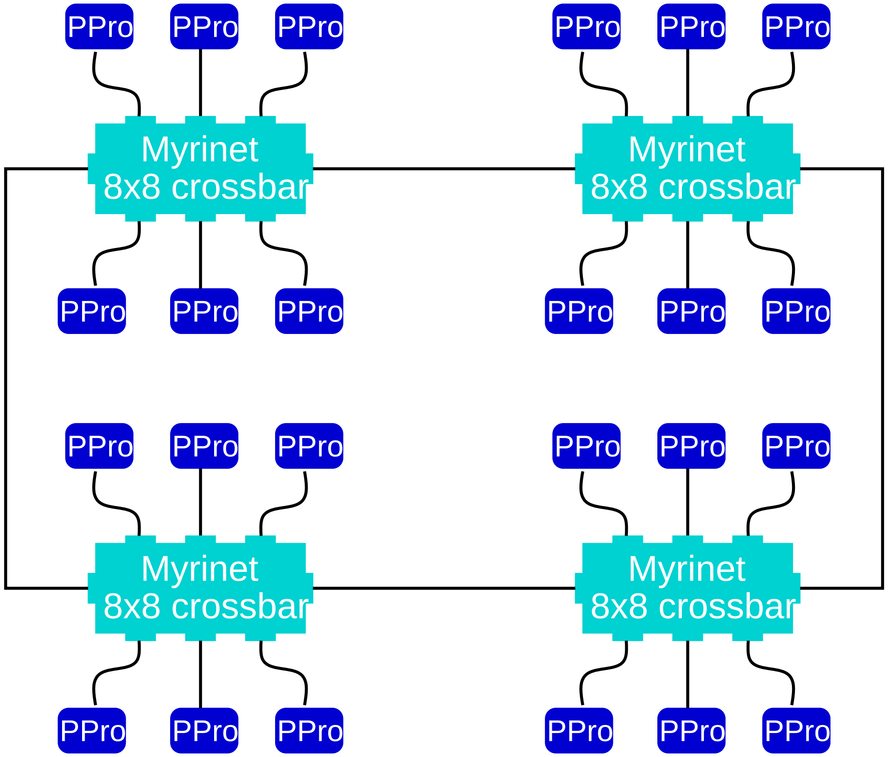
Components:

- 8-port switches
- Network interface card for each node (on PCI bus)
- Electrical cables: reliable links

Myrinet switches:

- 8×8 crossbar switch
- Each port connects to a node (network interface) or another switch
- Source-based, cut-through routing
- Less than 1 microsecond switching delay

24-node DAS-1 cluster



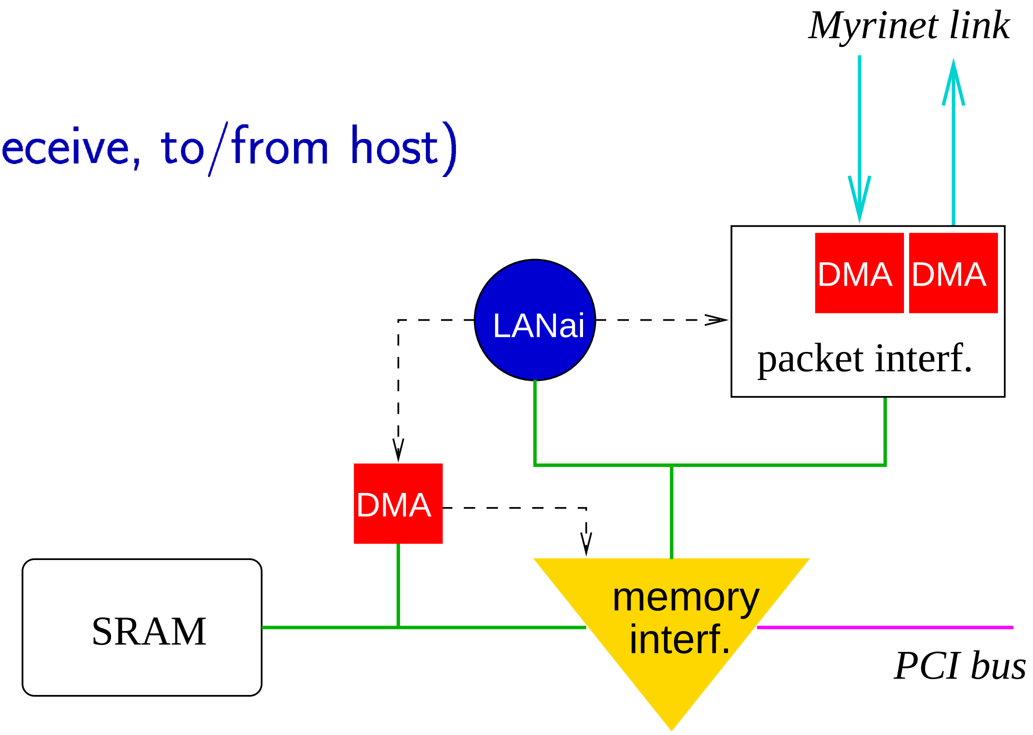
128-node DAS-1 cluster

- Ring topology:
  - 22 switches
  - Poor diameter: 11
  - Poor bisection width: 2
- Our cluster uses a grid with wrap-around
  - Each switch is connected to 4 other switches and 4 hosts
  - Need 32 switches (128/4) → 4×8 grid
  - Diameter: 6
  - Bisection width: 8

Myrinet interface board

Hardware

- 40 MHz custom cpu (LANai 4.1)
- 1 MByte SRAM
- 3 DMA engines (send, receive, to/from host)
- full duplex Myrinet link
- PCI bus interface



Software

- LANai Control Program (LCP)

Properties of Myrinet

- Programmable processor on the network interface
  - Slow (40 MHz)
- NI on the I/O bus, not the memory bus
  - Synchronization between host and NI is expensive
- Messages are staged through NI memory

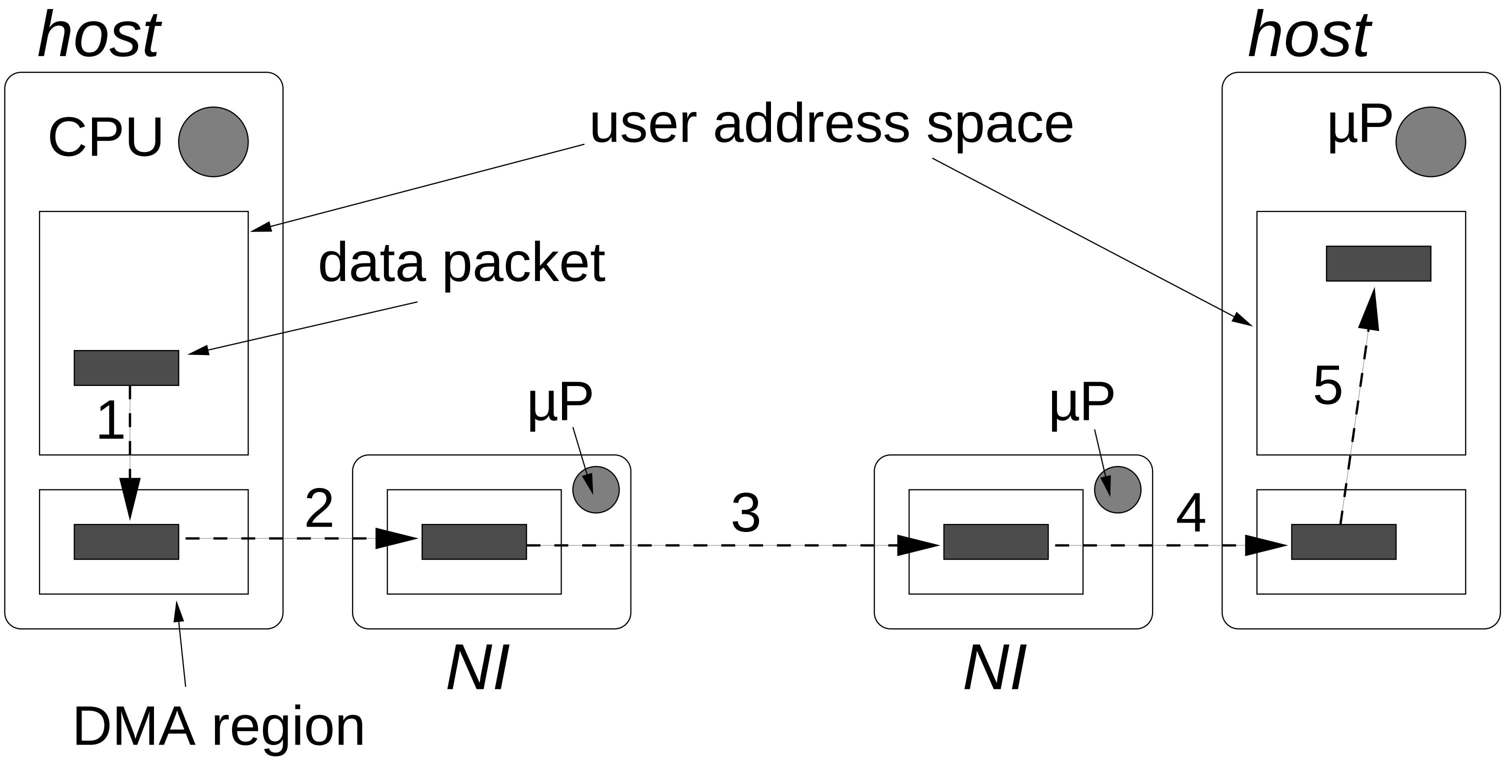
Network interface protocols for Myrinet

- Myrinet has programmable Network Interface processor
  - Gives much flexibility to protocol designer
- NI protocol: low-level software running on NI and host
- Used to implement higher-level programming languages and libraries
- Critical for performance
  - Want few μsecs latency, 10s MB/sec throughput
- Goal: give supercomputer communication performance to clusters

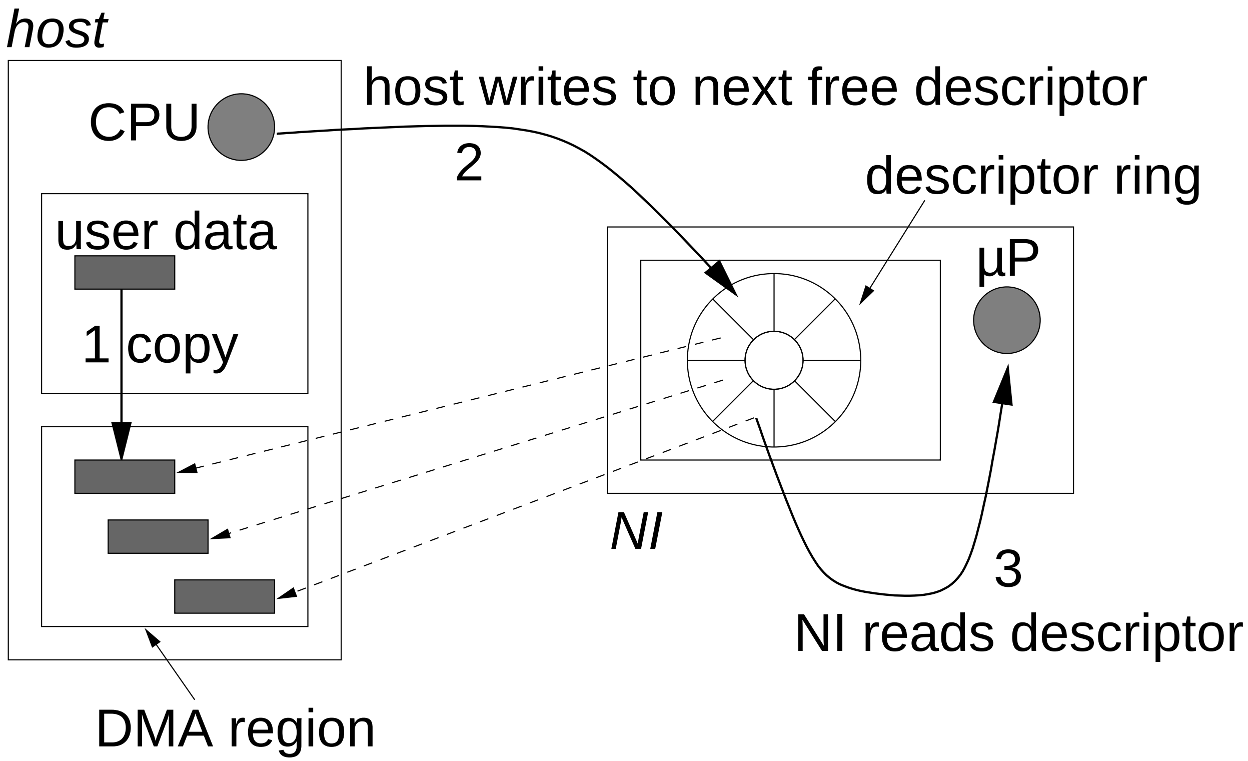
Basic Network Interface protocol for Myrinet

- Implement simple interface:
  - send(dest, buf);
  - poll();
  - handle_packet(buf);
- Map network interface (NI) into user space to avoid OS overhead
  - No protection (or sharing)
- No flow control
  - Drop messages if buffers overrun
  - Unreliable communication

Basic NI protocol – Overview



Basic NI protocol – Sending packets



Issues

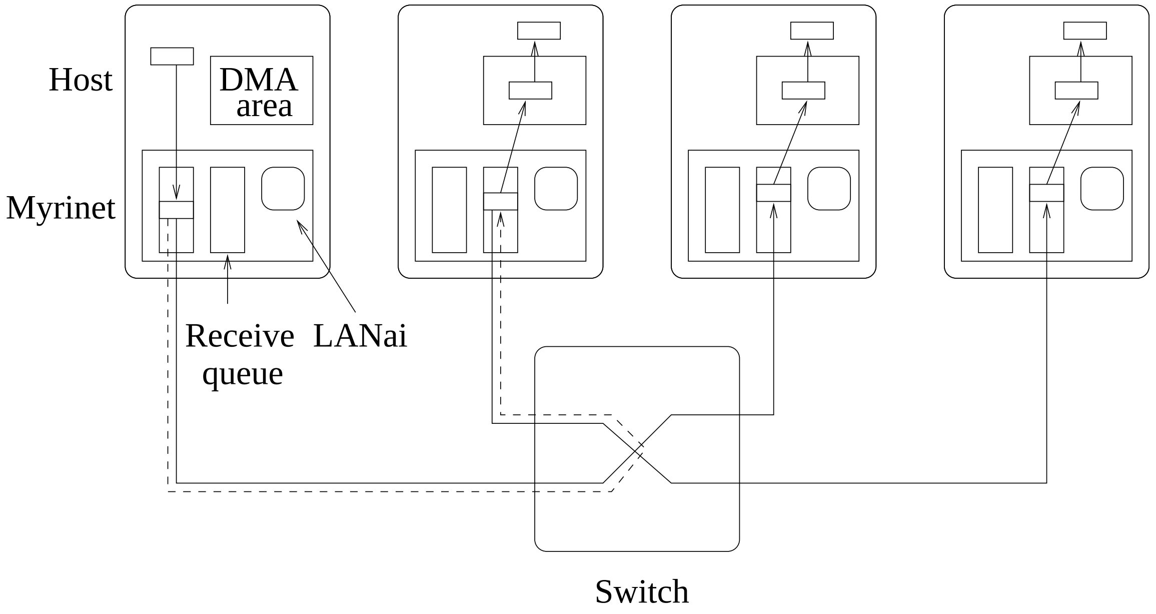
- Optimizing throughput using Programmed I/O instead of DMA
- Making communication reliable using flow control
- How to receive messages: polling overhead → Interrupts vs. polling
- Efficient multicast communication

Control transfers: polling versus interrupts

- Interrupts
  - User-level signal handlers are very expensive (24 μsec on BSD/OS)
- Polling
  - Hard to determine optimal polling rate
  - Burdon on programmer or compiler
- Combine polling and interrupts
  - Host polls when idle, else it enables interrupts
  - Requires integration with thread scheduler
- Polling watchdog (LFC)
  - Generate interrupt only if host does not poll within T μsec
  - Implemented using timer on NI

Multicast

- Implement spanning tree forwarding protocol on NIs
  - Reduces forwarding latency
  - No interrupts on hosts



Performance on DAS-1

- 9.6 μsec 1-way null-latency
- 57.7 MB/sec point-to-point throughput
- 48.0 μsec multicast null-latency
- 11.0 MB/sec multicast throughput (1 sender)