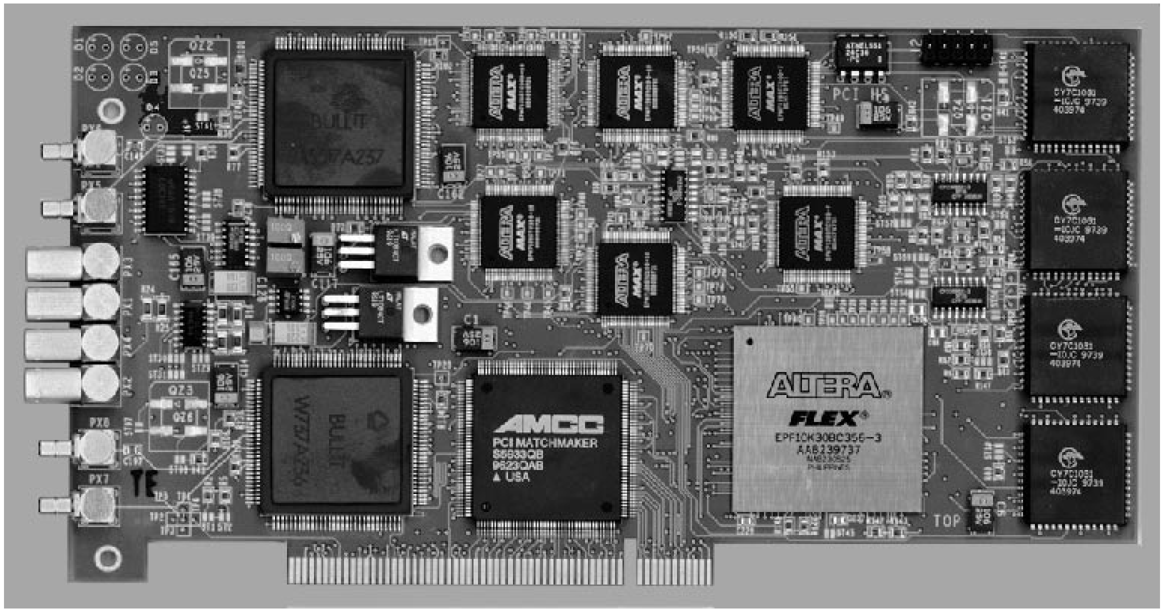
An HS–Link Network Interface Board for Parallel Computing

A. Cruz, J. Pech, A. Tarancón, C. L. Ullod, C. Ungil



RTNN

- Lattice Field Theory
  - Allows non-perturbative study of Quantum Field Theory
  - ...but Monte Carlo simulations require great computational power
- RTNN : A Beowulf cluster of 16 dual PentiumPro PCs
  - 16 nodes: 2 Intel PPro at 200 MHz, 64Mb memory, 2.3Gb disk, Linux OS
  - A front end machine manages the queue system (NQS)
  - Working since December 1996
    - has been used in 30 scientific papers and 6 Ph.D. thesis
- Used mainly as a task farm
  - MPI, PVM supported, but inefficient due to poor network performance
  - Both processors in a node can work in parallel through shared memory

Need for Parallelism

- Simulations carried out in RTNN

  $U(1)$ and $SU(2)$ Gauge–Higgs in d=4

  Z_2 and $O(N)$ Antiferromagnetic models in d=4

  $O(N)$ Antiferromagnetic models in d=3

  ISB in coupled scalar fields models in d=4

  Spin glasses in d=3

  High–T_c Superconductivity models

- Efficient parallelization would allow

  Larger lattice volumes in scalar models

  QCD: $SU(3)$ + dynamic quarks

  Condensed Matter: High-T_c Superconductivity with dynamic fermions

IEEE 1355 HS–Link

- Parallelization implemented using the PCI HS–Link interface
  - Developed at CERN to test HS–Link network technology
- HS–Links: 0.7–1GBaud, bidirectional, point to point connections
- Can be used to connect
  - Chips on a printed circuit
  - Printed circuits over a back plane
  - Racks by means of coaxial cable or optical fibre
- CERN involved in development and testing of the standard
  - Several boards has been developed to study different network topologies
  - A PCI HS–Link interface board can be used to generate network traffic
    - ...or to transfer data between PCs at high speed and low latency

HS–Link Protocol

- Four protocol layers: bit, character, exchange and packet
- Characters encoded by a 8B/12B DC balanced scheme

Start	Parity	Data / Inverted Data								Invert	Stop
1	P	D0	D1	D2	D3	D4	D5	D6	D7	I	0

  Invert bit sometimes redundant → Control characters

  Used during startdown and shutdown procedures, to keep the receptor calibrated and for flow control to avoid loss of data by filling buffers

- Information transmitted in packets

  A header is used to route the packet through a switching fabric

  The packet ends with an end-of-packet control character

  There are no restrictions on the packet size

HS–Link Devices

- The Bullit chip provides a parallel interface to an HS–Link
  - Transmitter/receiver pair
  - Input/Output 80 character deep FIFO buffers
    - accessed through separated two character wide interfaces
  - Low level protocol engine
- The RCube is a 8 × 8 router for HS–Link networks
  - Based on a 8 × 8 non–blocking crossbar switch
    - and 8 bidirectional 1GBaud serial links
  - Wormhole routing allows packets of unlimited length to be routed
  - Adaptative routing enables efficient load balancing in multistage networks
  - Total data bandwidth: 640Mbyte/s
  - Latency: 150ns

HS–Link Network Construction

- Several boards have been developed using those devices
- The 4 × 8–way will be used to connect the nodes of RTNN

  Consists of 4 RCube switches with all their links brought to the front panel for connecting to other switch modules or network interface cards

  A T8 microcontroller is used to configure and monitor the switches

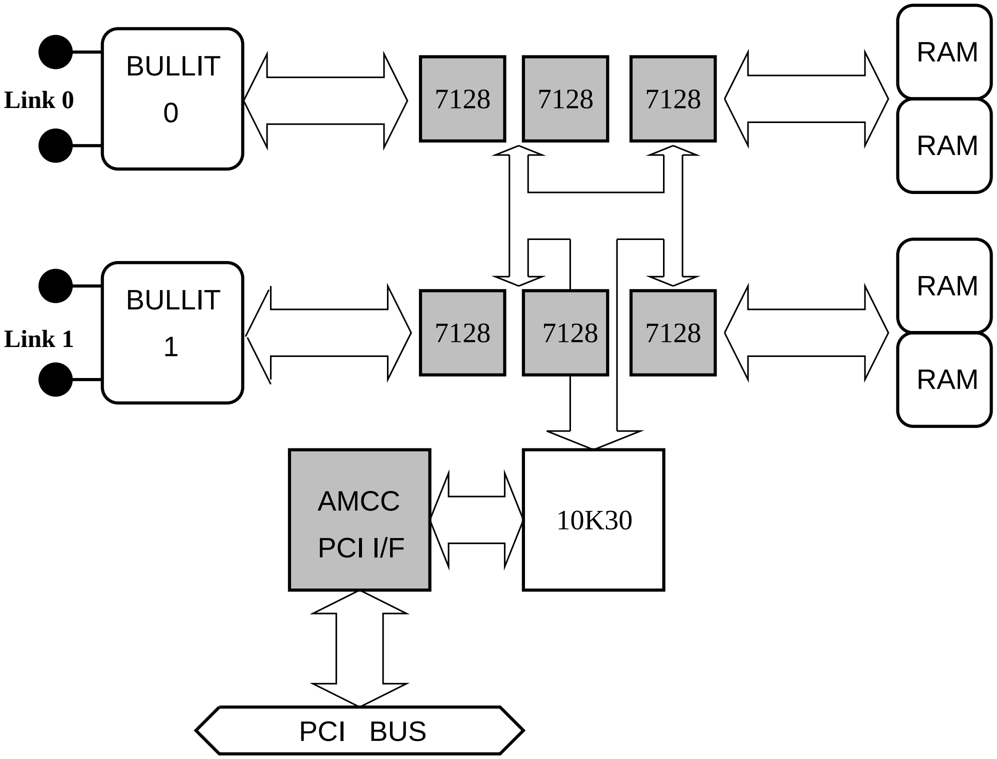
  IEEE 1355 DS–Links (or RS232) are used to control the module

  Several modules can be controlled with a daisy chain or star topology

- PCI HS–Link board used to access the HS–Link network

PCI HS–Link Interface

- Designed to serve two objectives:
  - Traffic generator for a network testbed
  - Interface between a PCI bus and a HS–Link network
    - (for network control or for high speed, low latency communication)



PCI HS–Link Design

- Interface chip to PCI bus (AMCC S5933)
- Glue logic implemented in a CPLD (Altera FLEX 10K30)
  - Handles multiplexing of the PCI interface between the HS–Link channels
- Two HS–Link channels
  - One Bullit
  - Supporting logic for reading/writing emission and reception FIFOs and accessing inner registers in the Bullit (Altera MAX 7K)
  - Two memory banks
- Functionality depends on the logic actually programmed
  - FLEX 10K30 programmable through the AMCC
  - MAX 7128s linked by a JTAG programming chain controlled from the FLEX 10K30 or from a specific connector

PCI HS–Link as Traffic Generator

- Boards generate traffic according to certain pattern
  - Packet descriptors (destination, length, launching time) are stored in the transmission memories
  - During the simulation the implemented logic interprets the descriptors and generates the packets
  - Similar descriptors are stored in the reception memories registering the incoming packets, and later downloaded through the PCI bus
- Used to study the behavior of different network topologies
  - MACRAMÉ, ARCHES projects, ATLAS experiment

PCI HS–Link as Network Interface Card

- Glue logic modified for the network interface use
  - AMCC PCI interface acts as the master of the transfers

    (to enable burst mode in both reading and writing)
  - Support for Direct Memory Access (DMA) added to the logic
  - Outgoing data sent directly through the links
  - Incoming data stored in the memories if the PCI bus is busy

    and downloaded later when becomes available
- PCI bandwidth 132 Mbyte/s, unidirectional
- Each HS–Link bandwidth 66 Mbyte/s, bidirectional
- PCI HS–Link board peak transmission rate 132 Mbyte/s
  - Preliminary tests: 79 Mbytes/s (limited by the chipset in RTNN)

PCI HS–Link Software Support

- A Linux device driver is used to control the boards

  Supports standard UNIX calls: read( ), write( ), ioctl( )

  Provides access to the inner registers in the Bullits and the glue logic

  Multiple boards are supported and detected automatically

  select( ) function, timers, task queues and interrupt handlers
  added to manage communication tasks

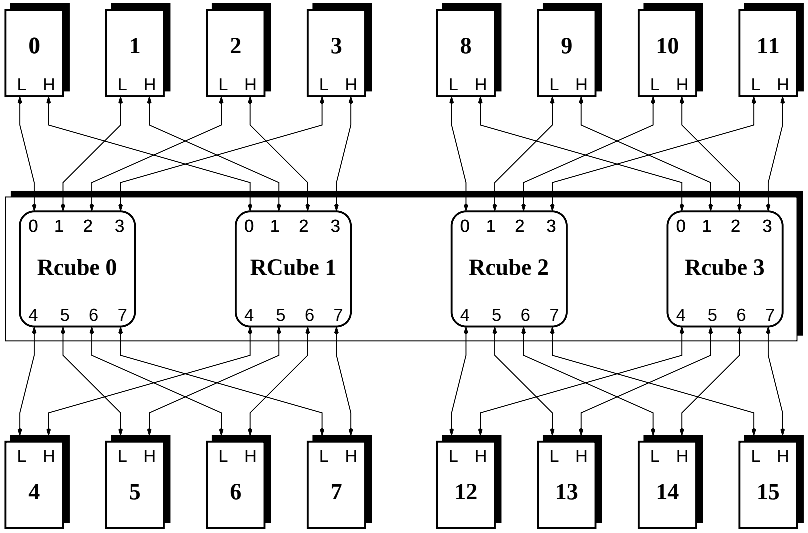
  Internal memory mapped to PCI memory and accesed using mmap( )

- Direct control by the user increases performance

  Communication management done by the application →
  context switching and data duplication avoided→
  communication overhead reduced

RTNN Parallelization

- Two subclusters of eight nodes can be fully parallelised using a 4×8–way switch



- Parallelization of the whole cluster for frequently used topologies can be achieved using three switches

Conclusions

- Technology for high speed communication has been adapted to parallelize a 16–node PC cluster
- PCIHS board main purpose: Traffic Generator
- On–board reprogrammability allows changes of functionality
- The device driver has been modified to manage communication
- Two full parallel 8–node clusters will be achieved with a switch
- The boards will be installed in a new cluster with faster processors
- Any question? (Jaroslav.Pech@cern.ch)