

The T-NET Hardware

Introduction

The bottleneck of most commodity supercomputers using MPI is the communication subsystem, which consist of the network hardware and the communication libraries.

In order to maximise the system performance a specialised network hardware (T-NET) and an own set of communication libraries (MPI/FCI) were developed. T-NET consists of Network interfaces (NIC) and switches. The NIC can move data to and from the network serving as a network port for a computational node. The switch forwards data from any input link to one (unicast) or several (multicast) output links. Each link can be either connected to a NIC or a switch forming any network topology.

The Link Protocol

The links of the network are fully bidirectional and have a raw bandwidth of 1.25 Gbit/s per direction. The usable bandwidth after deducting the protocol overhead is 100 MB/s per direction or 200 MB/s altogether.

Data is send over the network in micropackets (up to 128 Bytes). Each packet has a destination ID in it's header and, at the packet end a 16 bit CRC (cyclic redundancy check) which is used for error detection. After every transmission over a network link the packet is checked for correctness and automatically retransmitted if an error is detected.

The 16 bit CRC is generated in the NIC and not changed in the network, therefore bit errors in the switches themselves can be also detected.

The total bit error rate (BER) of the link is estimated to be 10-20 . So far, even during extensive testing and while using bad cables, no undetected and uncorrected errors where encountered.

For the cable connections between the switches and the NIC a standard Fibre channel (Dual twinaxial copper) cable (up to 20m) or multimode (up to 500m) optical fibre can be used.

The NIC

The NIC is a standard 32 Bit, 33 MHz PCI adapter providing an usable 80MB/s I/O bandwidth (133 MB/s theoretical peak) for the host computer.

Data can be transferred by Programmed I/O (the CPU is sending data directly to the NIC) or the DMA engine of the NIC can be used (NIC read data directly from the memory).

The data is packetized and send over the network.



Data received over the network is extracted from the packets and written directly to the memory. The virtual to physical address translation is performed with an on board page table.

The Switch

The T-NET switch has 12 ports. It consists of a active backplane with 12 slots and 12 link subunits (LSU). Packets are received from a link by the LSU and the header is extracted and send to the routing controller (RC). The RC uses it's 256 KB routing lookup table (RLUT) to find out to which output link(s) the packet needs to be switched. The packets are then switched by a 12x12 Crossbar switch.

The bisection bandwidth of the Crossbar is 1200 MB/s enabling 12 simultaneous transactions at full wire speed.



The Switch can be accessed via a serial interface and can be remotely configured.

The Hardware is FPGA based and all firmware can be changed over a service network.

Technical data

Bandwidth	
Raw link bandwidth (unidirectional)	156 MB/s
Usable link bandwidth (unidirectional)	100 MB/s
Usable link bandwidth (aggregate)	200 MB/s
Host to NIC bandwidth (unidirectional)	75 MB/s
Host to NIC bandwidth (aggregate)	80 MB/s
MPI bandwidth (measured)	72 MB/s
Switch bisection bandwidth	1200 MB/s

Latency	
NIC latency (input + output)	~3.5 µs
Switch latency	0.5 µs
MPI latency (measured)	< 20 µs

Error rate	
cable BER	< 10 ⁻¹²
retransmission protocol BER	~ 10 ⁻²⁰

Future Development Road map

In a first step, the NIC will be upgraded from 32 Bit PCI to 64 Bit PCI doubling the I/O bandwidth of the Host computer (160 MB/s instead of 80 MB/s).

In a second step, a larger switch (more links) will be build.

In a third step, the link speed will be doubled in order to double the unidirectional bandwidth and total network bandwidth.

Contact Person

Roland Paul
e-mail: roland.paul@scs.ch
Phone: +41 (0)1 445 16 00