FLink - a high speed PC network interface

Flink - standard PC-version

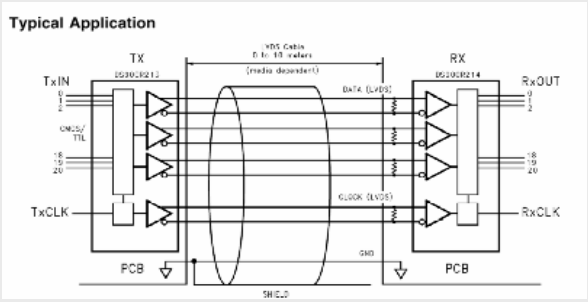
Contents

Introduction

- APEmille is based on a PC-cluster
- 32 processors á 0.5 GFLOP per PC
- 64 PC‘s for a 1TByte machine
- asynchronous communication network
- high speed... 1GBit/s
- low latency...< 1µs

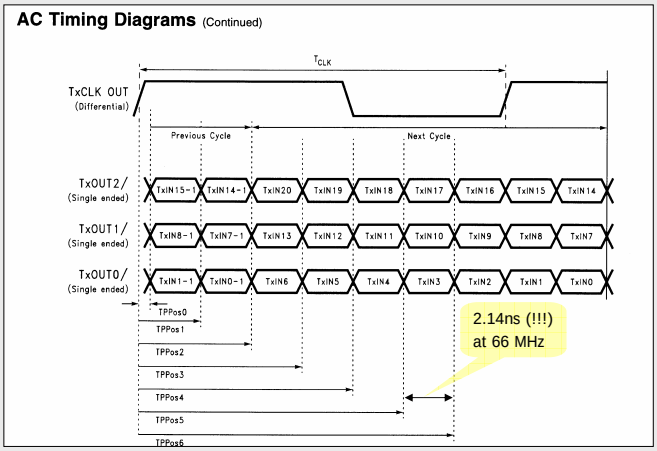
Physical Link Layer

- Chips, NSC : DS90CR213 / DS90CR214 (20..66 MHz)
- Compression, 21 bit parallel to 3x7 bit serial
- digital interface for flat panel displays, ...10 m Cable

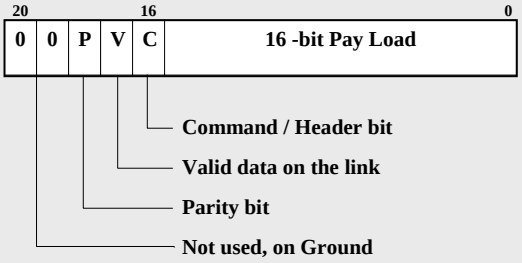


Physical Link Layer

- theoretical timing diagram



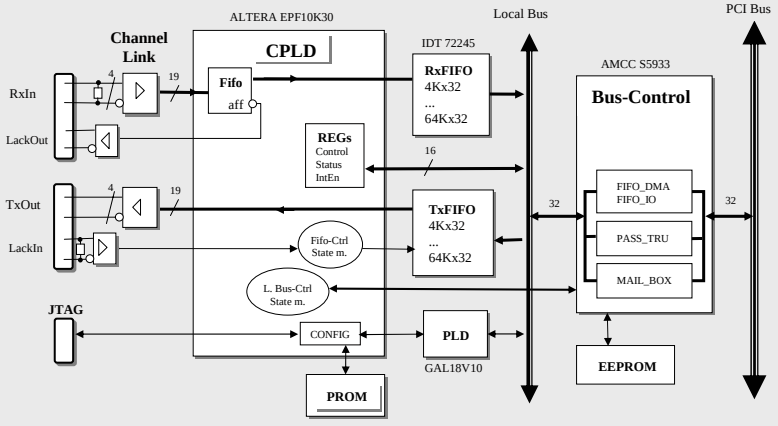
Channel link bits, usage



- Link clock is running all the time -> need of the Valid bit
- Parity bit can be used to recognize a disconnection
- Shortest message concists out of 2 channel link words

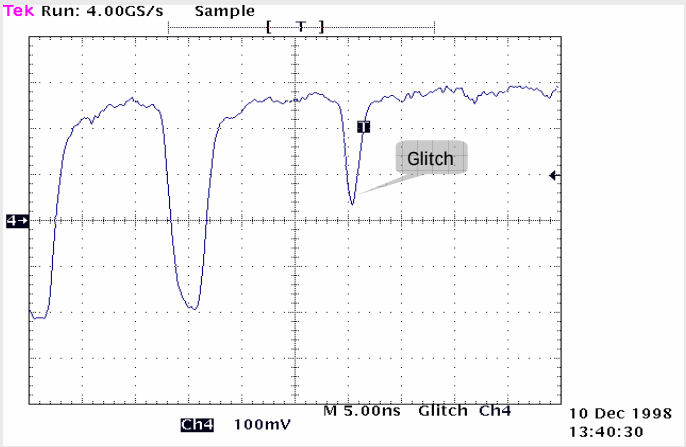
Blockdiagram - P2P protocol

- Conform to PCI2SHL software interface



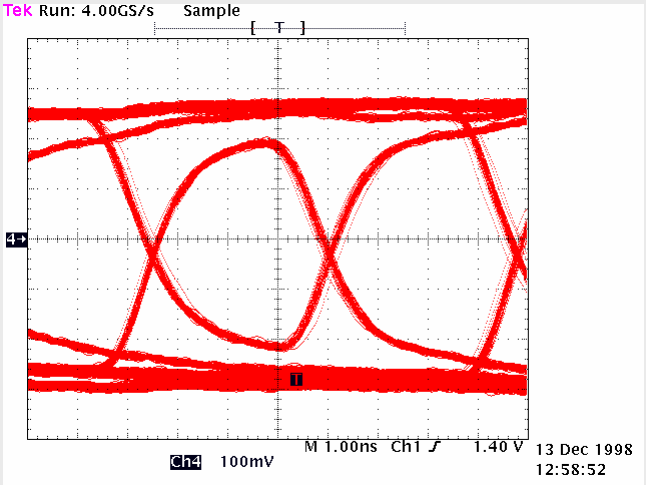
Glitch problem

- Glitch, due to low frequency power instability (100mV)



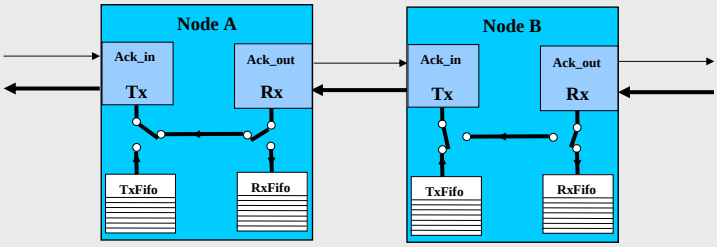
Eye diagram, 40 Mhz link clock

•At receiver input, 3m unshielded twisted pair cable, 100 Ω



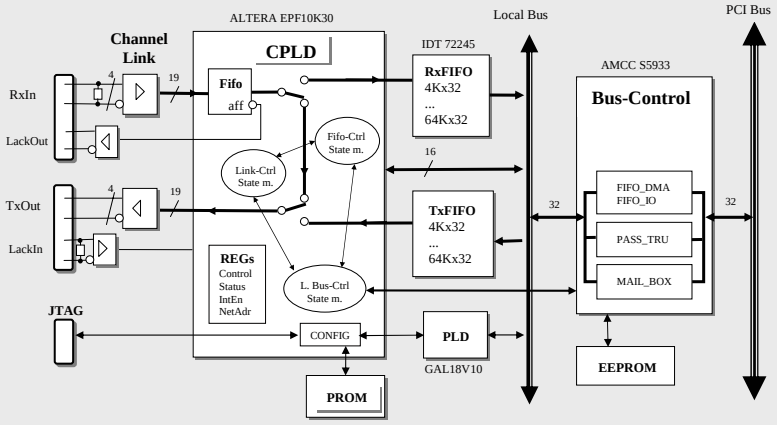
Principle of ring connection

- Point to point connection, max. 10 m distance
- link acknowledge signal for synchronisation
- „zipper principle“ for send / pass through - conflict



Blockdiagram - Ring protocol

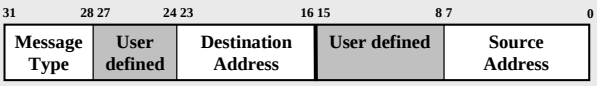
- sophisticated CPLD design



Ring protocol

- 2 data types: single commands and packets
- single command: 32 bit word
- packet: 32 bit header
  32 bit packet length (16 bit used)
  1..16K-1 32 bit words
  32 bit checksum (XOR over all)

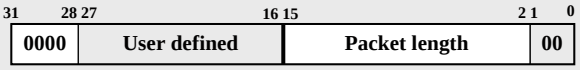
Command / Packet header



- *Message Type* kind of data following
- *Destination Address* 8 bit node address, decoded by the Flink link arbiter
- *Source Address* 8 bit, copied from the ADR-register, when sending
- *User defined* these bits are not relevant for the Flink hardware

Message types

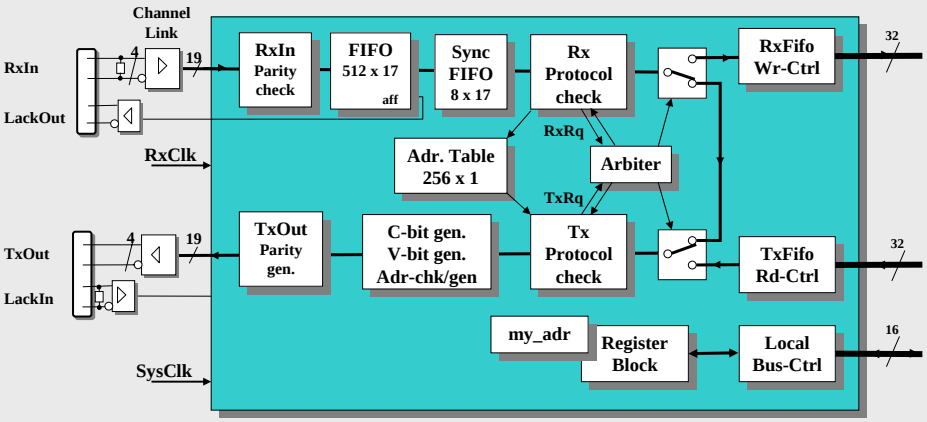
- *OPN* - Open, modifies the address table (RAM 256x1) of every node, when passing through, RAM[source_adr] = 1
- *CLS* - Close, modifies the address table of every node, when passing through, RAM[source_adr] = 0
- *LWR* - Local Write, packet header, exchange of data between two nodes
- *GWR*- Global Write, packet header of a broadcast message
- *PLE* - Packet Length in multiples of 4 bytes, must follow LWR or GWR



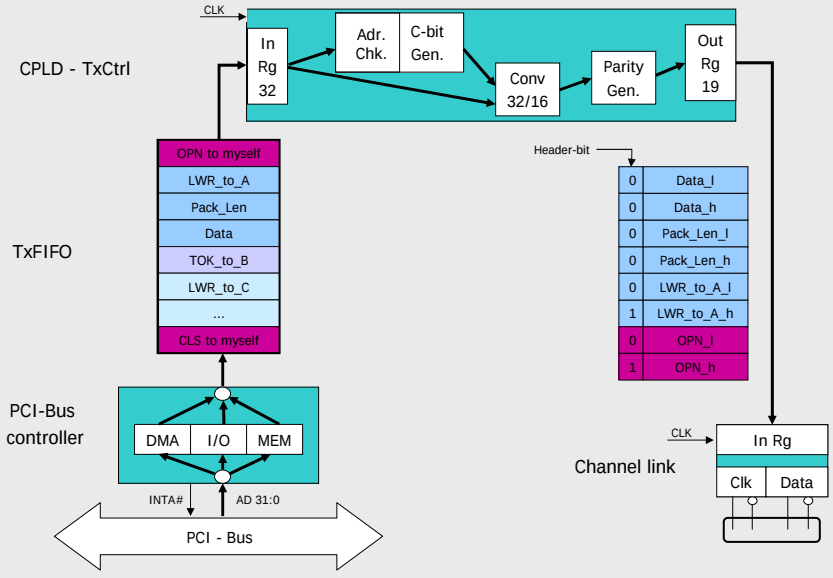
Message types, cont.

- *TOK* - Transfer OK, software initiated acknowledge, can cause an interrupt at the destination
- *RRS* - Remote Reset, to reset or synchronise nodes by a ring master, can cause an interrupt at the destination
- bits 28..30 are used only, protocol extension are possible

CPLD - Blockdiagram



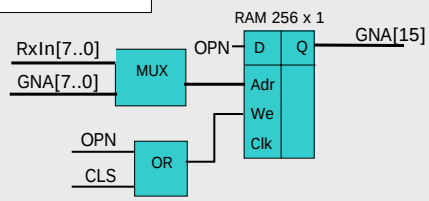
Sending Packets / Commands



OPN / CLS - Command

- Every node „knows“ the addresses from all other nodes
- interrupt on change of the RAM contents (OPN/CLS event)
- TxData can be flushed, if a non existing address is in use

GNA		
	name:	Get Node Addresses
	address offset:	24h
	load-up value :	0000 0000h
	attribute:	RW, RO



Special Effects

- when sending, the contents of the ADR-register (own net adr.) is copied to the source portion of the header
- the destination portion of the header is verified to the address RAM to prevent infinite loops of data
- Auto_Clear, in case af any error a global Ring_Clear is actived
  - bypassing the internal data channel of the CPLD
  - asserted until the cause of error disappears
- Interrupt on Ring_Clear

Ring mode

- Everyone may send to everyone
- to prevent a ring blocking (LACK-signal) the following rules have to take into account:
  - if more than two nodes are active
    - max_pack_len = CPLD_fifo_size / count_of_active_nodes
    - CPLD_fifo_size = 1 Kbyte -> poor performance, if more than 2 nodes
    - workaround: bigger fifo or dynamically OPN / CLS usage
    - implementation of the token ring principle (changing the CPLD design)
    - redesign with bigger fifo, outside of the CPLD, ...128 Kbyte possible
  - never send twice to the same node without acknowledge
  - Receive before Sending

Master / Slave (APE) mode

- Slaves may send only to the master, including OPN / CLS
- following rules have to take into account:
  - max_pack_len = external_fifo_size, independent on the count of active nodes
  - external_fifo_size = 16 Kbyte -> good performance
  - Receive before Sending

Test results

- Prototype runs not with full link speed (132 Mbytes/sec) due to:
  - not optimal PCB design+power/clock supply of the channel link
- up to 100 Mbytes/sec o.k. (loop back mode checked only), but to much PCB changes ->easier per jumper 66 Mbytes/sec
- Ring - and Master / Slave mode has been succesfully tested under the following conditions:
  - ...4 different PC's (Pentium 90MHz...233MHz) connected
  - dynamically adaption of the maximum packet length (Ring mode)
  - random packet length within possible limits
  - random addressing
  - noise causing data pattern
  - Parity ,Protocol and Data check
  - random sequence of GWR / LWR (Master / Slave mode)

Conclusions

- reliability of Channel Link is proofed:
  - long term tests, 20 TByte of data, without errors
  - low pass through latency of 300ns
  - link speed of 66 Mbytes / sec
  - DMA speed of 120 Mbyte / sec
- redesign with:
  - improved PCB layout
  - newest channel link chipset DS90CR217 / 218A (...170 Mbytes/sec)
  - faster ALTERA CPLD (10KE series)
  - AMPHENOL „skew clear" cable (350 ps / 10m , pair to pair)
  - bigger Fifo for Ring mode