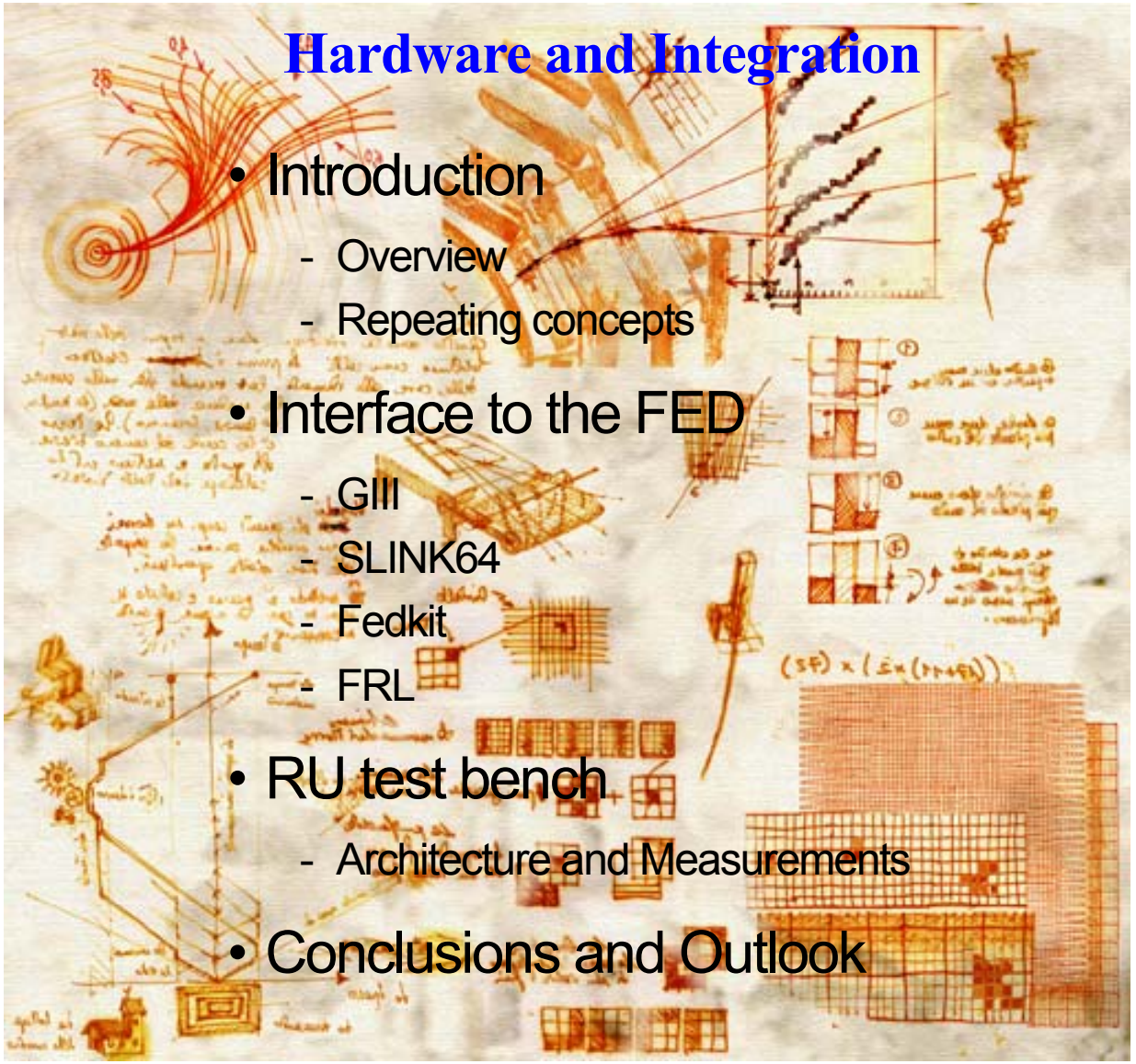
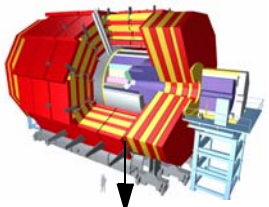
DAQ Column



Hardware and Integration

- Introduction
  - Overview
  - Repeating concepts
- Interface to the FED
  - GIII
  - SLINK64
  - Fedkit
  - FRL
- RU test bench
  - Architecture and Measurements
- Conclusions and Outlook

DAQ Column



200 MB/s

Front End Driver **FED**

SLINK64

FED Readout Link **FRL**

underground

surface

optical Link x 8

Trigger (Lvl 1) **TR**

FED builder switch

RU

Readout Unit

x 64

Event Manager **EM**

RU builder switch

x 64

BU

Builder Unit

DAQ slice (8x)

Filter Unit High Level Triggers **FU**

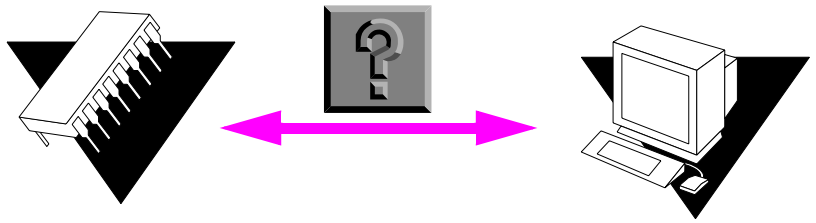
Computer Service Network tape writing **CSN**

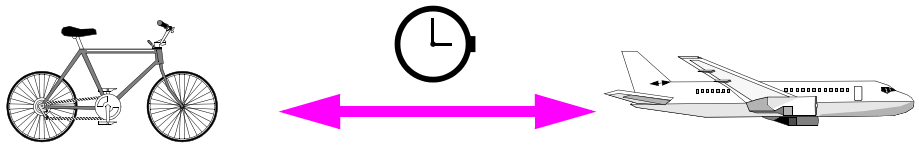
Event builder

covered in talk

Purpose of the Readout Column

- Integrate hardware and software
  - test interfaces between different units
  - test protocols
- Try out different implementation options for final design
  - test bench for various technologies
- Measure performance
  - this is one of the **primary goals** now
- Integrate FEDs into the system...
  - ...as soon as they are available

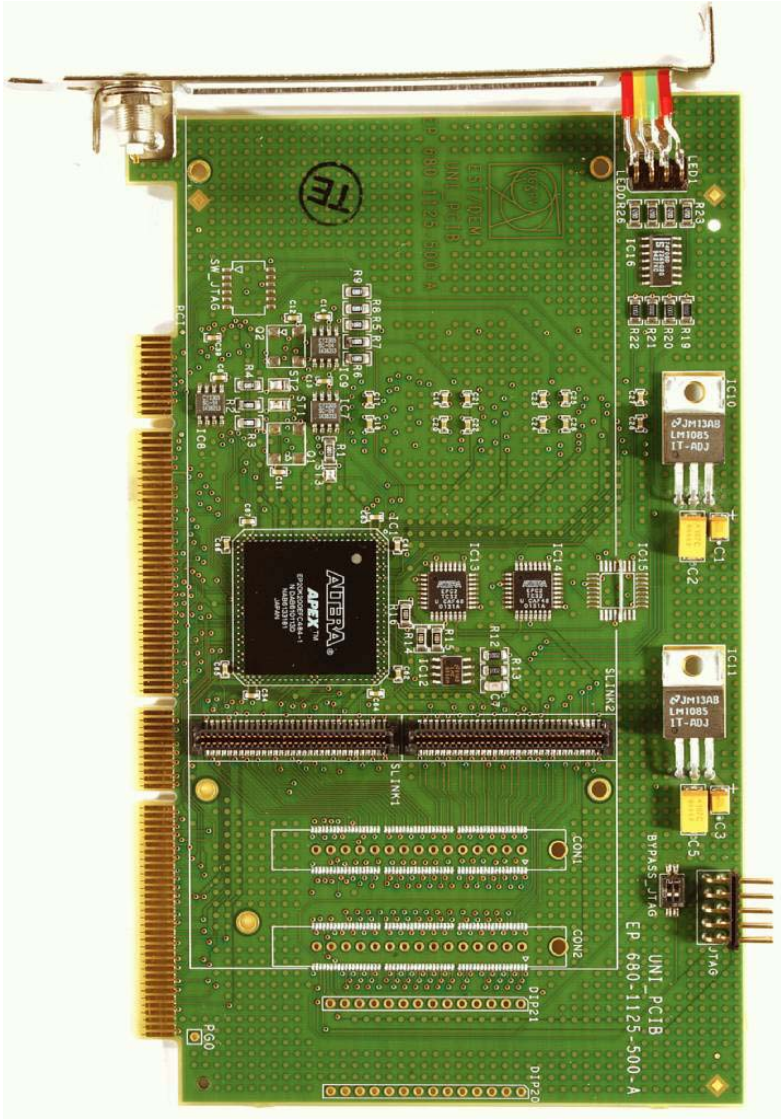




Develop a complete DAQ prototype

Reused Components / Concepts

- Hardware: Generic III platform
  - PCI 64 bit / 66MHz “universal card”
  - One FPGA (PCI interface + user code)
  - 32 MB SDRAM
  - Connectors: SLINK64 & multi purpose connector
  - Software Kit to download firmware “in situ”
- Available for everybody (see below)



- Software: the buffer loaning scheme
  - Used to transfer data from external card into PC (or vice versa)

Requirements for data transfer interface: hardware card - computer memory:

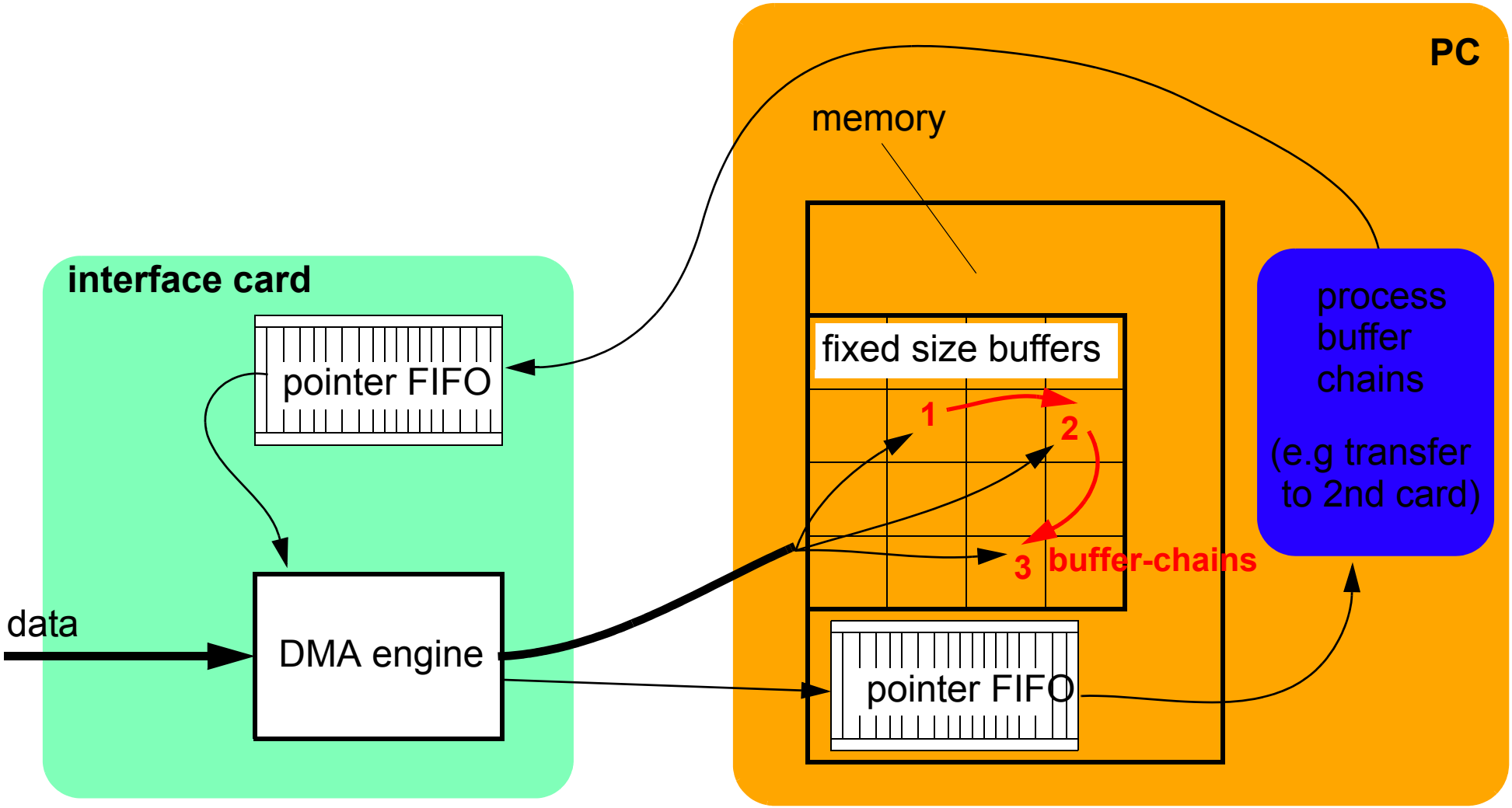
- CPU must not be loaded with data copying
- Length of data packet not include in header
- Hardware cards have DMA engines but small memory
- Robustness against fluctuations:
  - Data-volume
  - CPU availability (Linux is NOT Real Time)

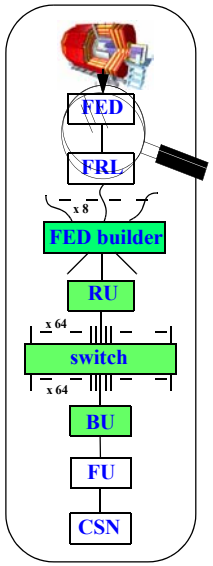
Solution: buffer loaning scheme (“zero copy”):

- CPU does not copy data (free for more important jobs)
- Data is moved by DMA engines
- In PC only pointers are handled around

This scheme is supported by / compatible with all hard and software components of DAQ (GIII, XDAQ, NICs)

Buffer loaning scheme: functional diagram for card to PC transfer





The Interface to the FED: SLINK64

Requirements:

- Easy to use for FED (FPGA friendly protocol)
- Data rate sustained >= 200 MB/s
- Data rate peak >= 400MB/s

Solution: SLINK64 protocol

- SLINK64 is based on SLINK; modification:

  32 bit ⇨ 64 bit

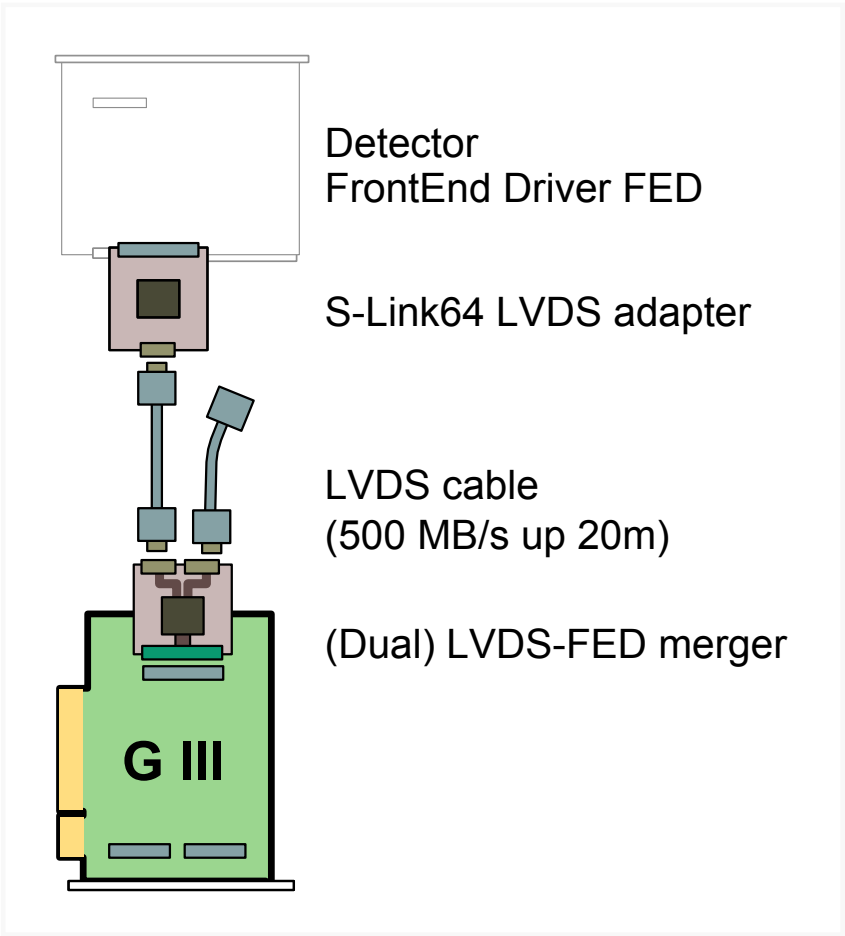
  40 MHz ⇨ max. 100 MHz

- Data format:

Header 1	BX, LVL1ID, SourceID
Header 2 (optional)	
fragment payload (arbitrary length)	
Trailer	Length, Status, Error detection code

Implementation and test bench of SLINK64

- LVDS implementation
  - Sender: FPGA, LVDS driver
  - LVDS cable (see below)
  - Receiver:
    - Merges up to two sources
    - FIFOs and FPGAs



- Measurement Setup
  - Data Generator (GIII)
  - Sender card (the card which will be plugged on the FED)
  - Receiver card
  - readout card based on GIII ⇨ PC
  - Sender card fully compliant to SLINK64

LVDS cable measurement results

vendor	length [m]	testing time	rate [MB/s] / ν [MHz]	result
AMP	2	1 month	800 / 100	no error
	7.5	8 hours	800 / 100	no error
	2+7.5+7.5	8 hours	528 / 66	no error (cables were connected via home made passive screened boxes)
Amphenol	15	4.5 hours	528 / 66	no error
	20	ε	264 / 33	errors
3M	10	1 hour	528 / 66	no error
	15	1 hour	528 / 66	no error

Conclusion for LVDS - SLINK64 prototype

- Fully functional SLINK64 has been built
- Easy to implement LVDS technology
- Throughput achieved is higher than needed

  528 MB / s at 17 m (400 MB/s required)
- More tests needed
  - Long term test with long cable
  - Use of “standard” test patterns

Remark: Cable paths in control room are not yet defined.

Fedkit

- Facility to test FED interface to DAQ.

Hardware components:

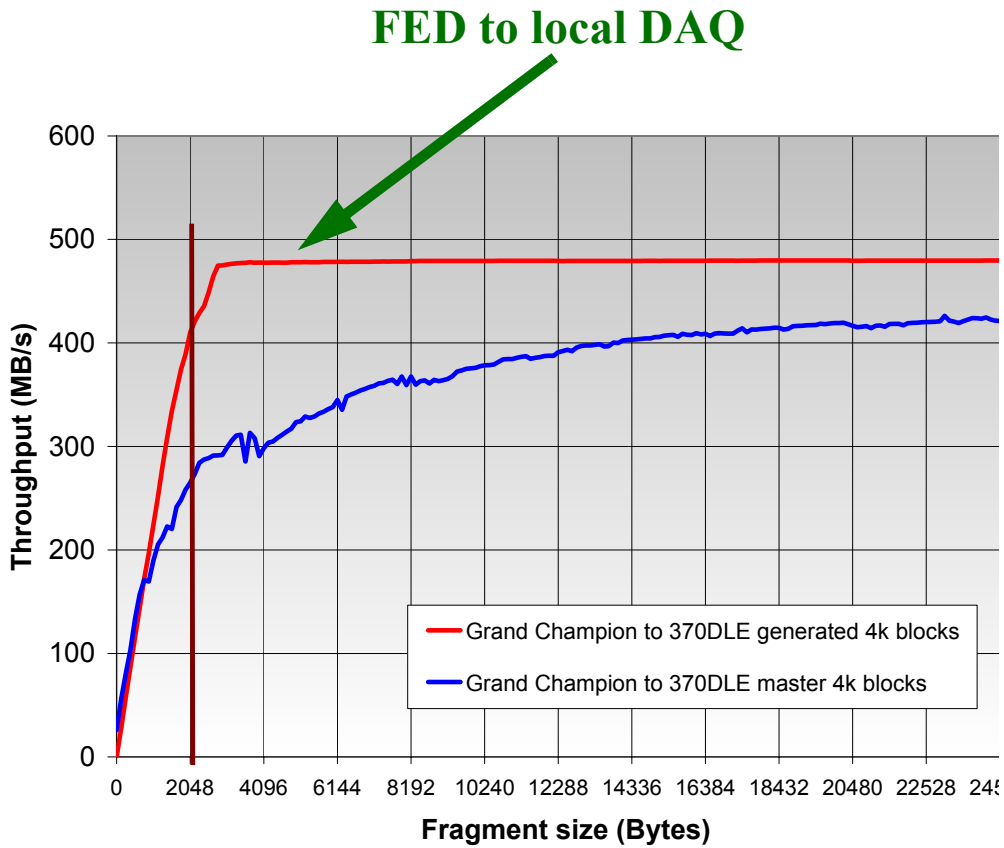
- SLINK64 sender
- Cable
- Receiver card
- GIII to be plugged into Linux PC

Software:

- Fedkit driver which handles hardware interaction
- XDAQ application with simple API

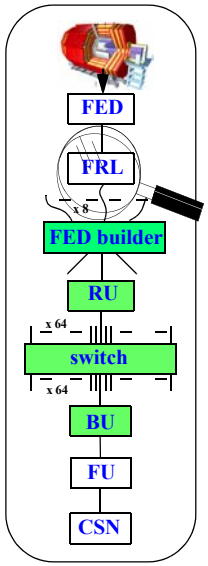
Performance

- Measurement: Fedkit driver only
- Various operation modes



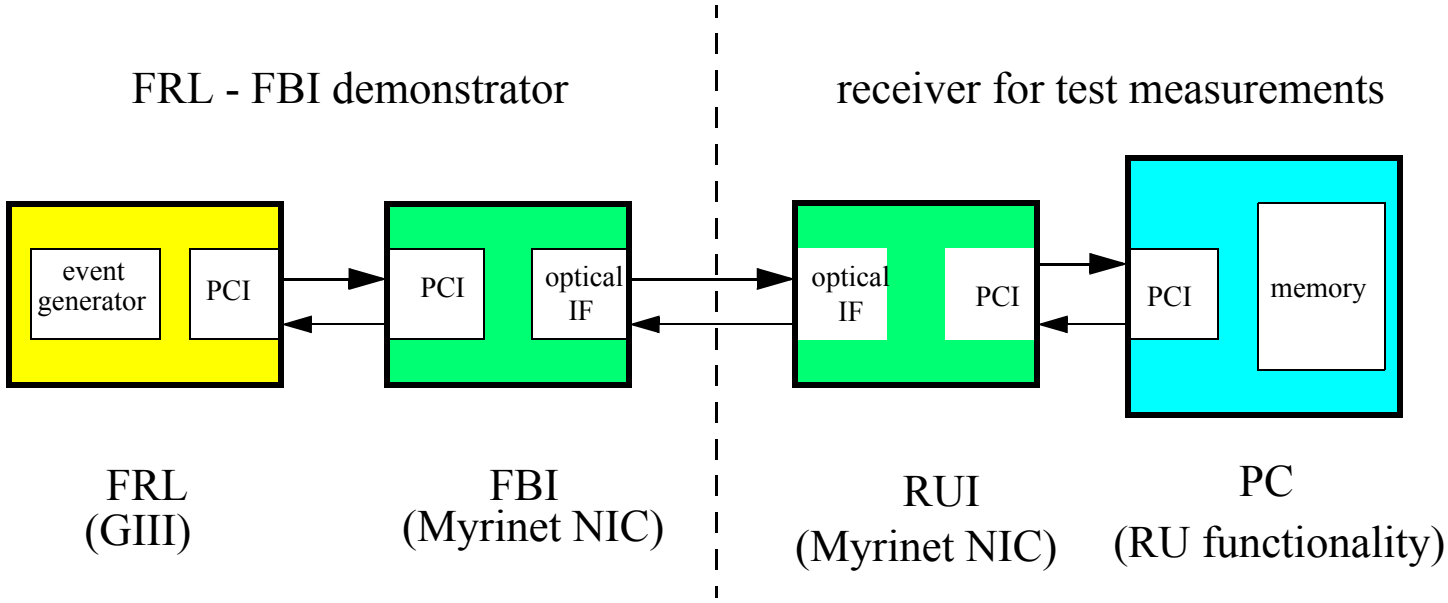
Fedkit in CMS

detector		pieces	date	remark
TriDAS		8 + 16 GIII	spring 2003	FED Builder demonstrator
subdetector Fedkits	Pixel	2	end 2003	
		2	end 2004	
	Tracker	2	12/2002	
	Preshower	3	10/2002	
	HCAL	2	9/2002	
	RPC	1	6/2002	
special	GTP emulator	1	-	
	ECAL	2	-	for link tests
total		23 + 16		



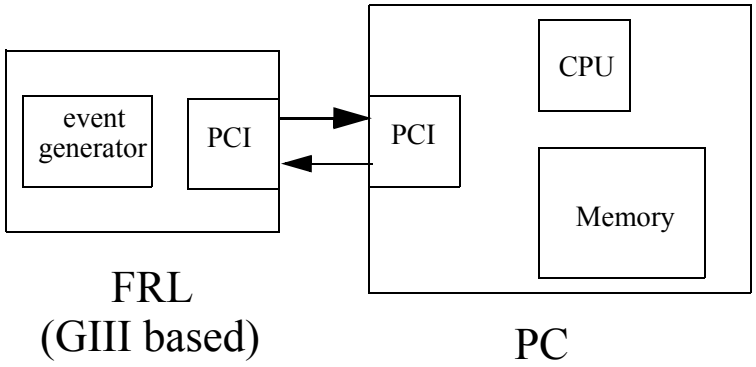
FRL Demonstrator

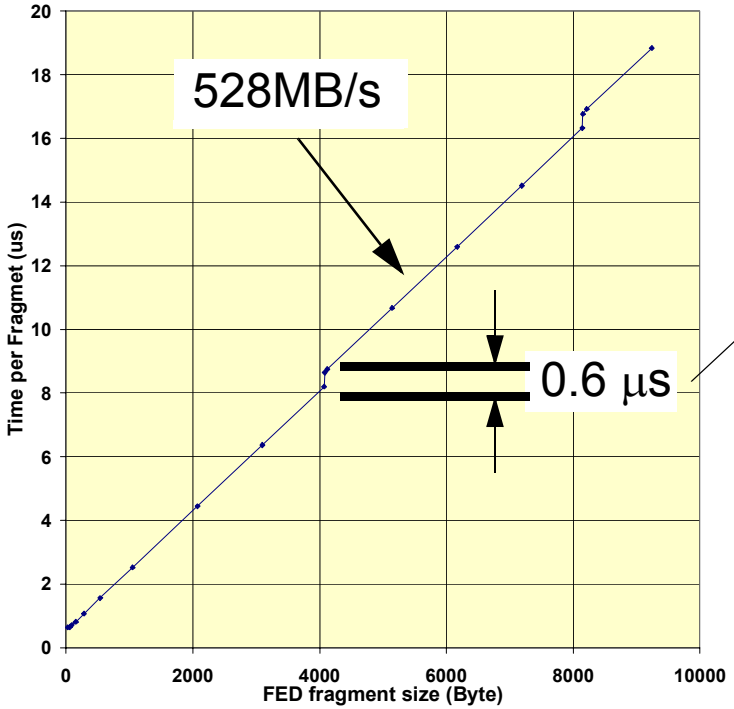
- Scope: Interface FRL, FBI (Front End Builder Input)
- FRL: GIII
- FBI: Myrinet NIC
- Data generator in GIII

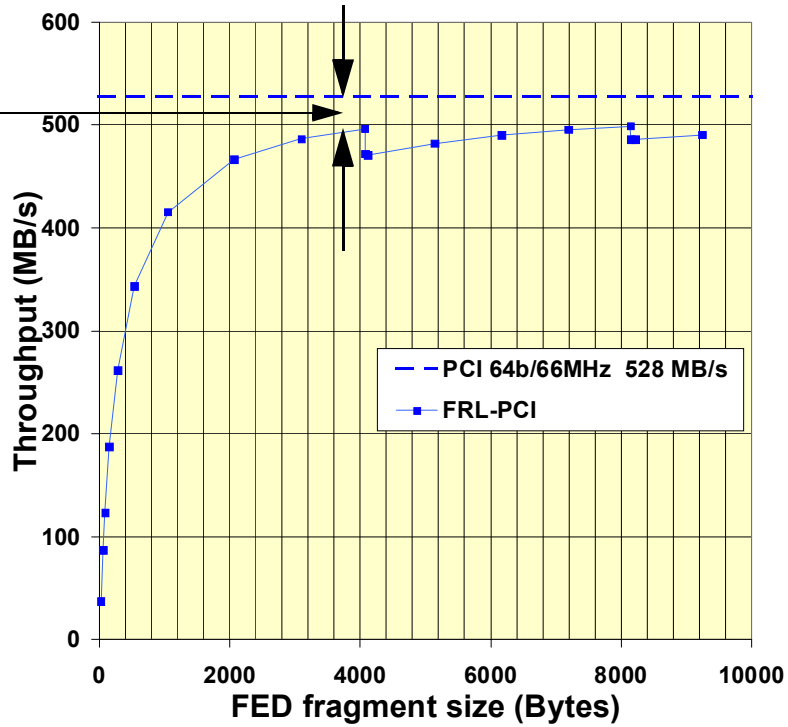


Measurements 1: FRL protocol

- PC emulates NIC card
  - Measures performance of FRL (PC is much faster than NIC)
  - Details: 4 kB pages, 66 Mhz 64 bit PCI-bus, 1024 page entries in GIII, fragment parameters for 512 fragments in GIII FIFO by PC

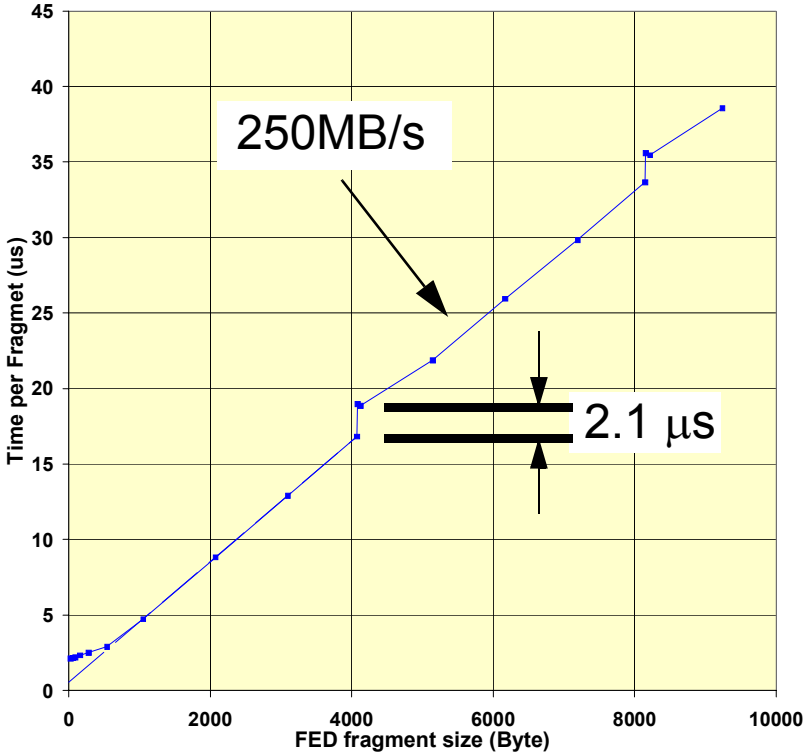


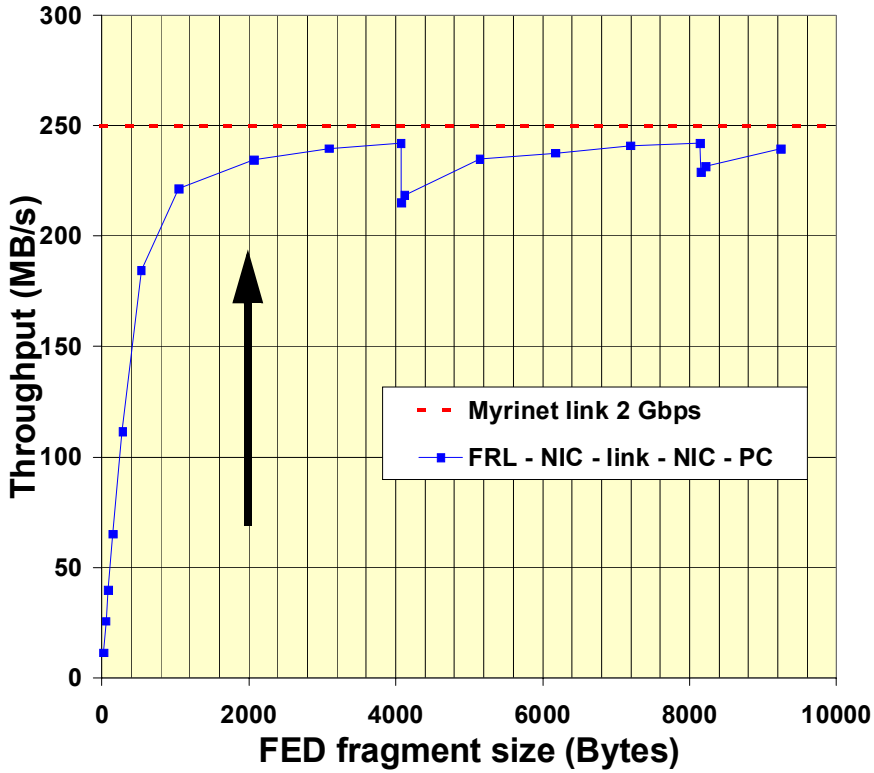




Measurements 2: full test bench

- Test 2: FRL - NIC - NIC - PC
  - Setup is Myrinet limited (theoretical max. 250 MB/s)
  - Offset 2.1 μs is partly shadowed (CPU works while DMA goes on)
  - 0.5 μs "irreducible" offset





Conclusions

- FRL - FBI interface performs close to theoretical limit
  - DMA performance of FRL is at theoretical limit: 528 MB/s
  - Small offset of 0.6 μs / DMA
  - With 2 kB fragments 460 MB/s throughput

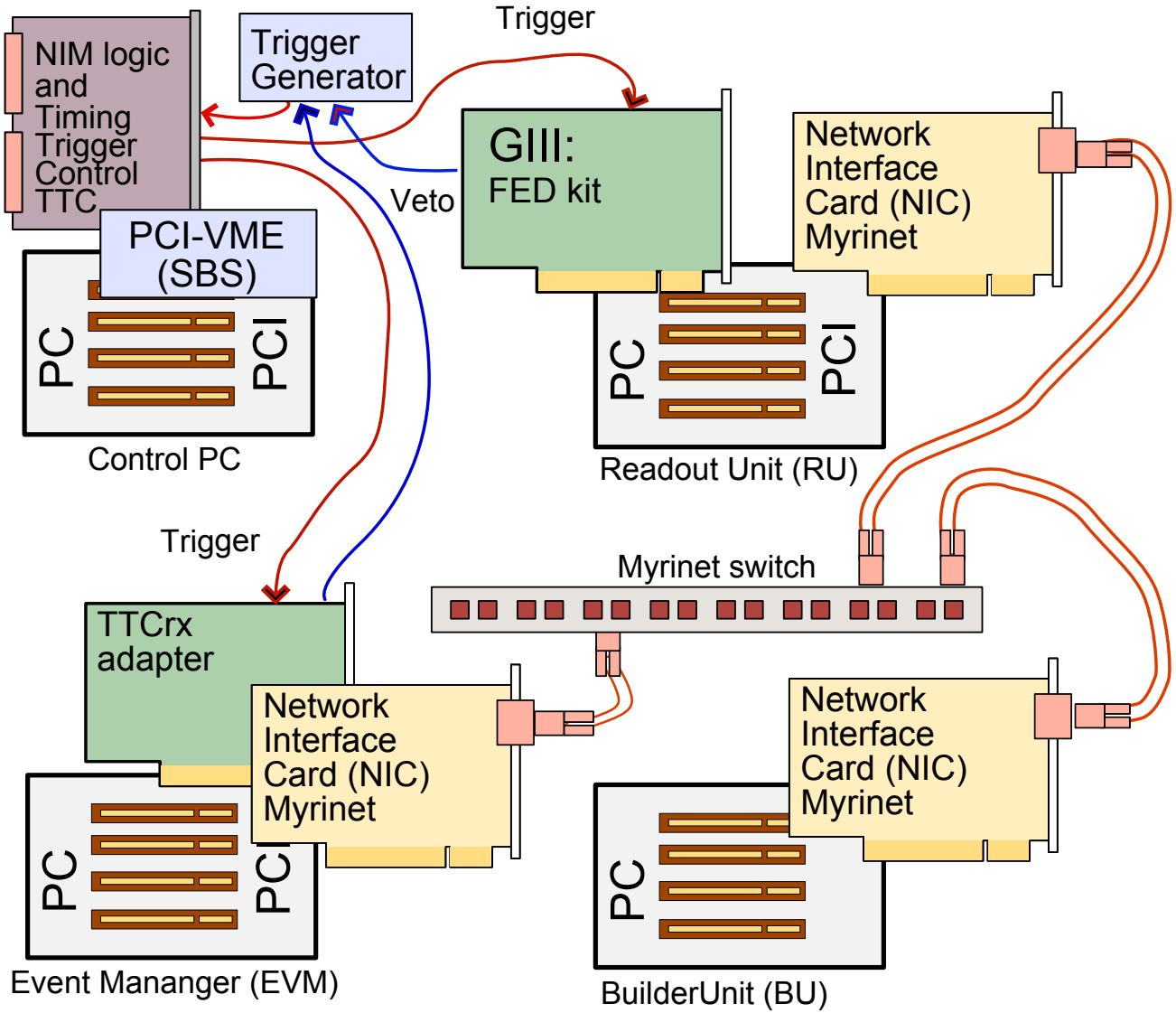
- FRL - FBI - NIC - PC chain
  - Limited by Myrinet performance
  - Bandwidth 230 MB/s for 2 kB fragments

CRUDE: Column for Readout Unit Development

PC based RU demonstrator

- Layout
- Features
- Measurements
- Conclusions

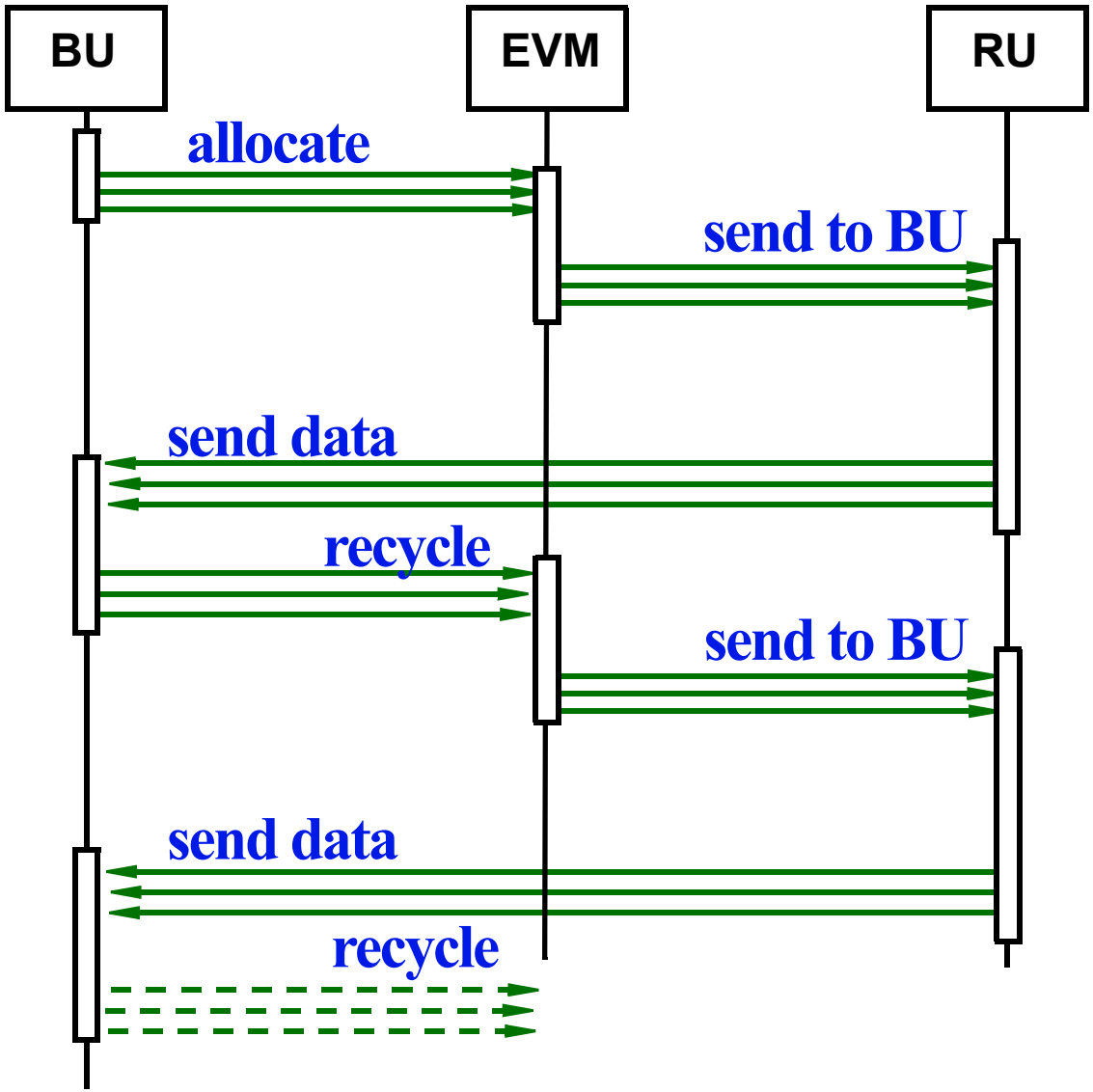
CRUDE layout



Test bench features for RU-PC

- XDAQ software environment
- XDAQ applications used and modified as needed
    - Event Manager Protocol: indirect mode
- Data source (RUI replacement): Fedkit with trigger and back pressure
    - Memory allocated by XDAQ
    - Backpressure latency compensated for with internal trigger counter
- Data output: BU
    - Myrinet card
    - GM transport layer (software from Myrinet)
- Implemented data checks:
    - Event numbers (always, in various places)
    - Data can be checked in BU (optional; not done if speed is measured)
    - No error recovery tried

Event Manager Protocol: Indirect Mode



Parameters to adjust

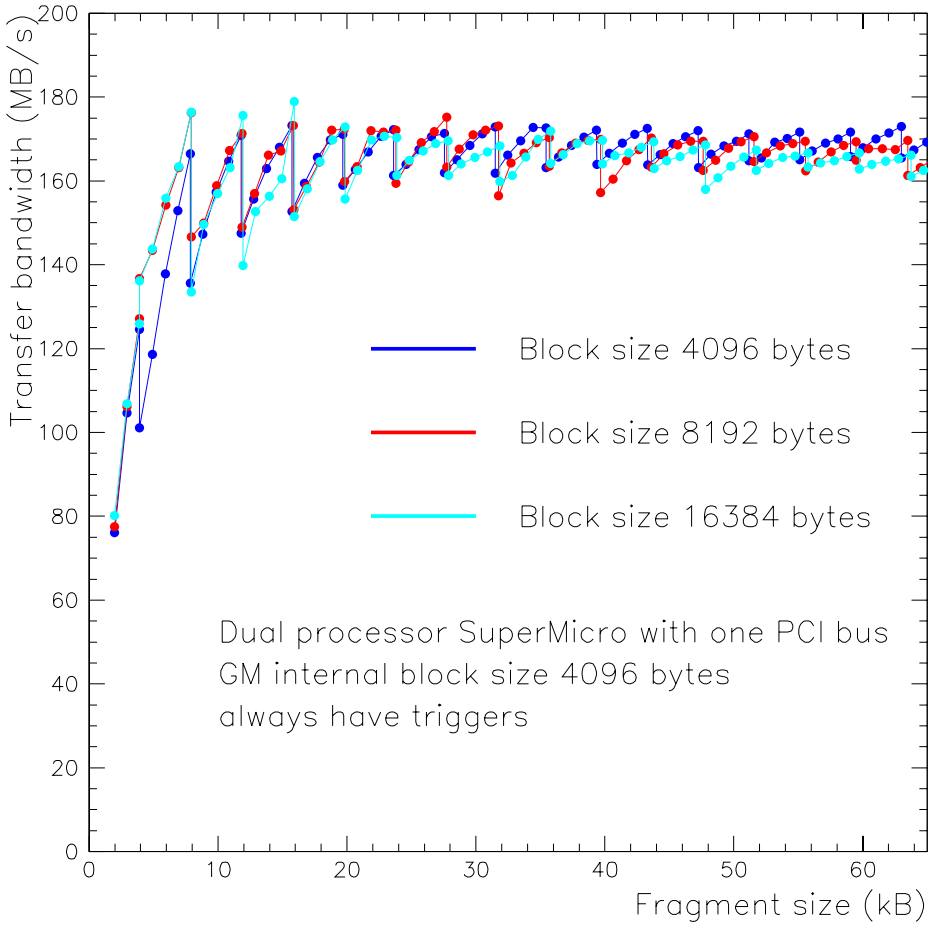
- PC type (chipset and motherboard)

Vendor	SuperMicro	SuperMicro	DELL	SuperMicro
Motherboard	DLE 370	DLE370	Powerdege 1550	P4DL6
Chip set vendor	ServerWorks	ServerWorks	ServerWorks	ServerWorks
Chip set	ServerSet III-LE	ServerSet III-LE	ServerSet III HE-SE-SL	Grand Champion LE
CPU	1 Pentium III	2 Pentium III	2 Pentium III	1 Pentium IV Xeon (hyperthreaded)
CPU clock	1 GHz	1 GHz	930 MHz	2.0 GHz
Memory	SDRAM (1.06 GB/s)	SDRAM (1.06 GB/s)	SDRAM (interleaved) (2.12 GB/s)	DDR SDRAM (interleaved) (3.2 GB/s)
System Bus	1.06 GB/s	1.06 GB/s	1.06 GB/s	3.2 GB/s
PCI 64 bit / 66 MHZ slots	2 slots on 1 bus	2 slots on 1 bus	2 slots in 2 independent buses	4 slots in 2 independent buses

- PC - system parameter
  - Bigphys size: DMA - capable RAM to buffer event fragments (is reserved at boot time; not available for OS)
- Buffer handling in XDAQ
  - Blocksize: blocks used in Fedkit (RUI), RU. (GM uses fixed blocksize of 4 kB)
  - Maximum number of blocks circulating for data transfer
- Event fragment generator
  - Fragment size distribution: constant / table driven
- EVM protocol
  - RU: bundling of send request
  - BU: bundling of requests

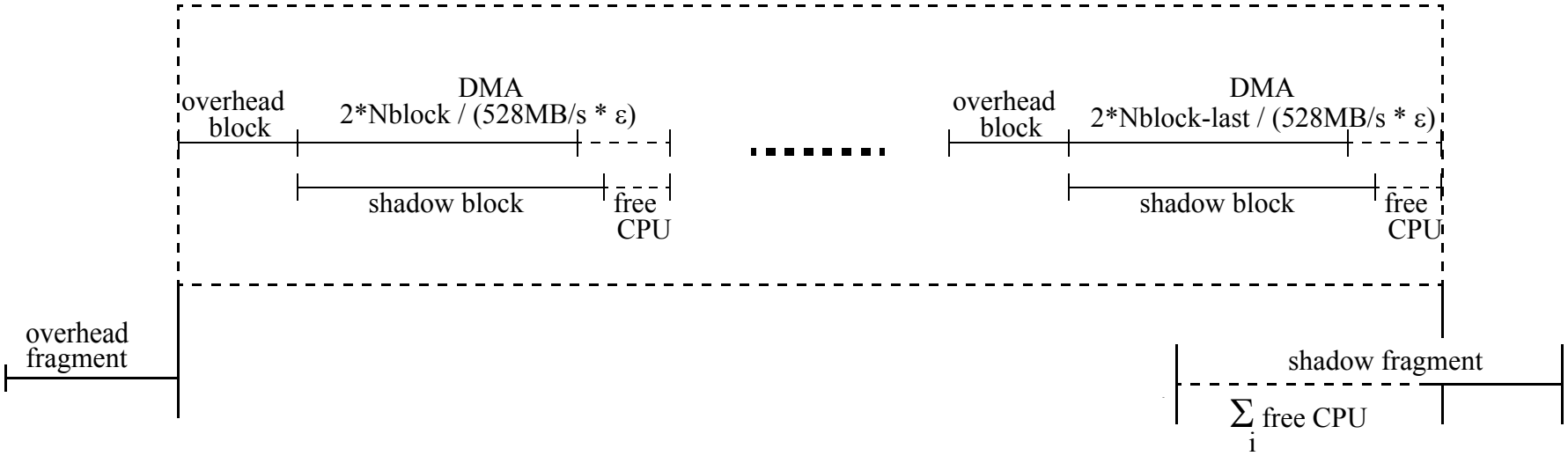
Results: Super Micro different block size

- Fixed parameters
  - 200 MB total buffer (bigphys)
  - 1000 buffers for Fedkit
  - Fixed fragment sizes
  - BU and RU request bundling: 50
- GM works with fixed blocksize of 4 kB -> saw tooth
  - Block sizes multiples of 4 kB
- At 16 kB:
  160 MB/s +/- 13 MB/s



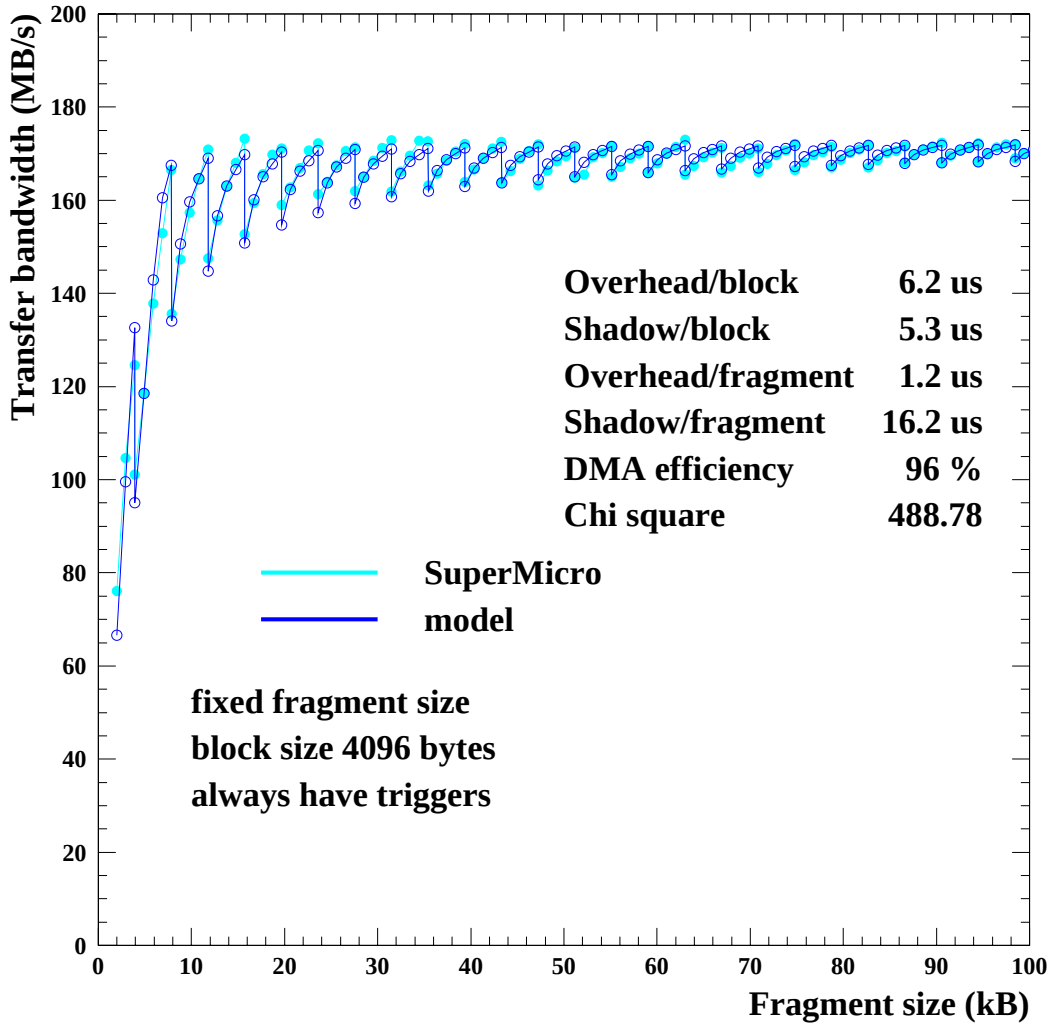
Results: Comparison with a simple model

- Only model PCI bus
  - does not work if link becomes limiting factor:



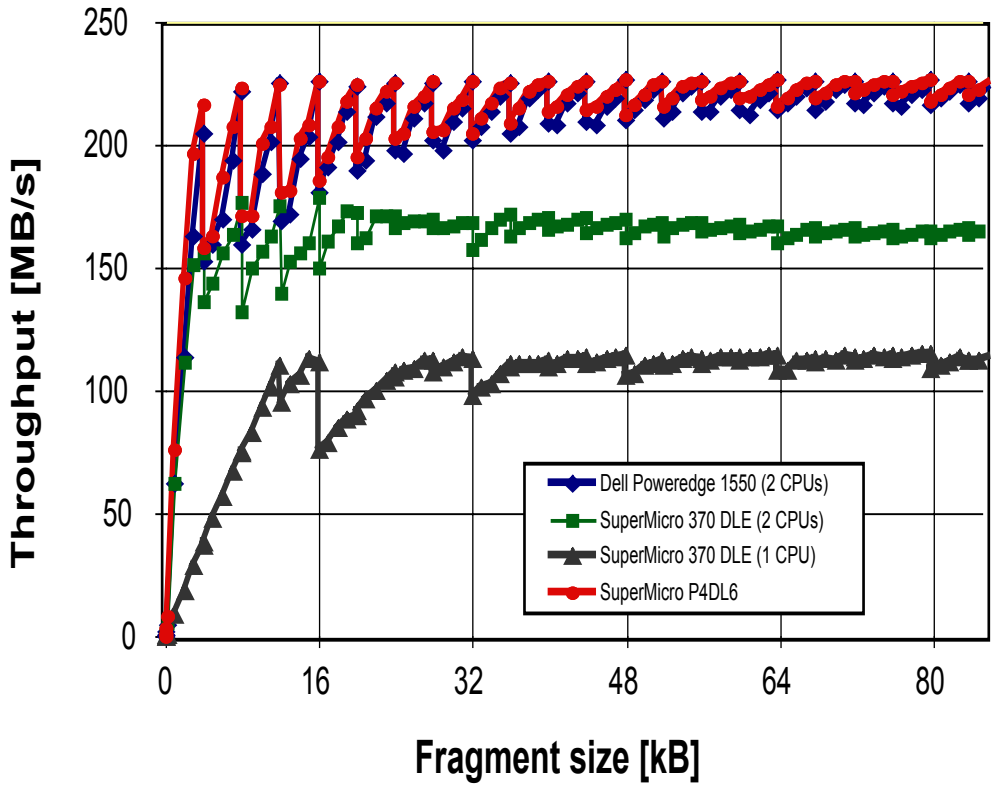
- Try to fit to this model with 5 parameters...

5 parameter fit to the simple model



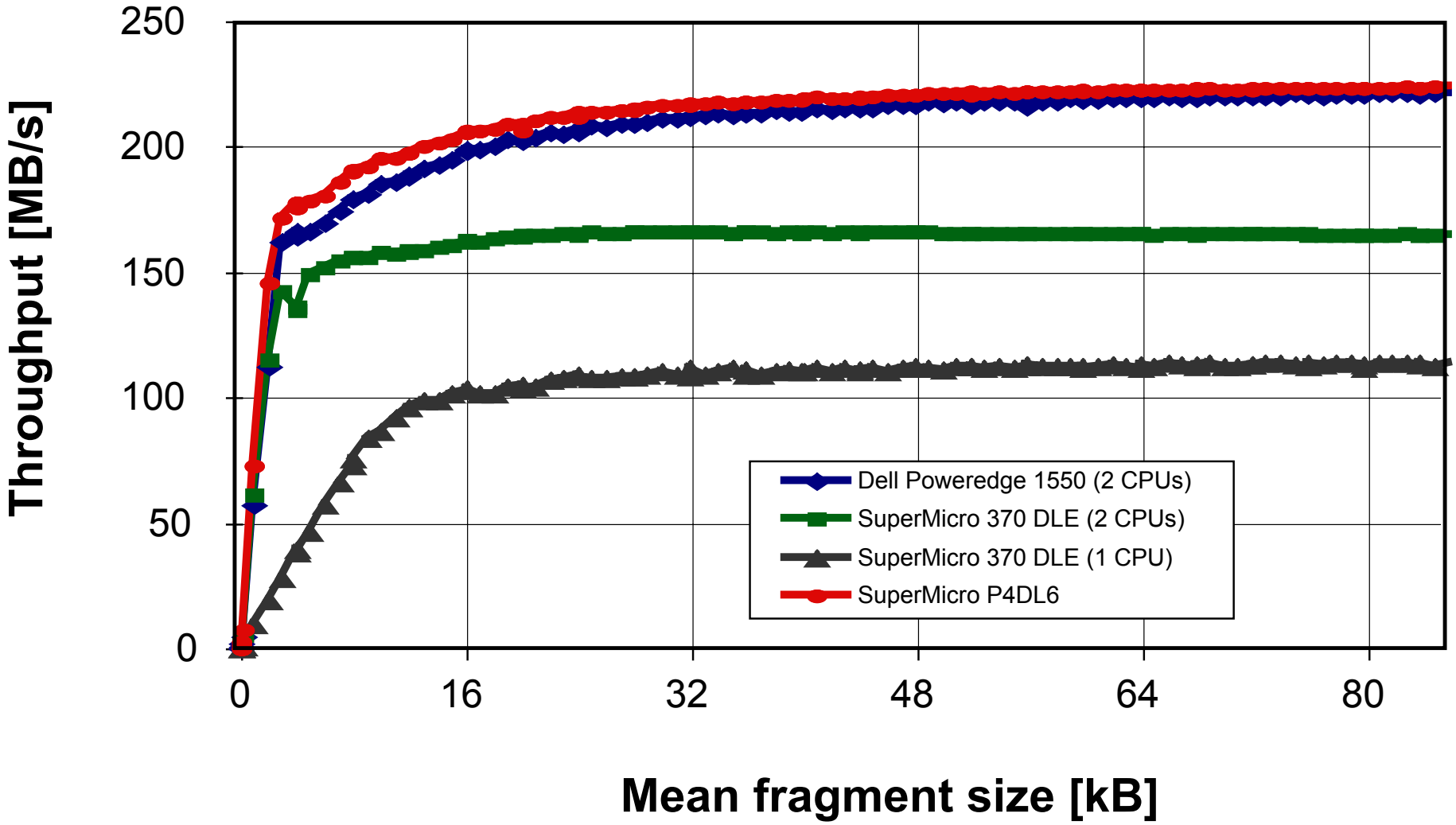
Results: Comparison of various PCs

- High throughput for PCs with independent PCI buses
  - DELL and P4DL6 servers have independent PCI buses
  - throughput starts to be limited by Myrinet link
- Simple Model fit cannot describe the data
  - takes only PCI bus into account



Results: variable fragment size

- fragment size is varied with rms = mean fragment size / sqrt(8)



RU test bench summary

- Different PCs have been tested
- Good Performance with independent PC buses
  - P4DL6 206 MB / s with 16 kB mean fragment
  - DELL Poweredge 198 MB / s with 16 kB mean fragment

- Required bandwidth of 200 MB/s has been reached

Possible improvements:

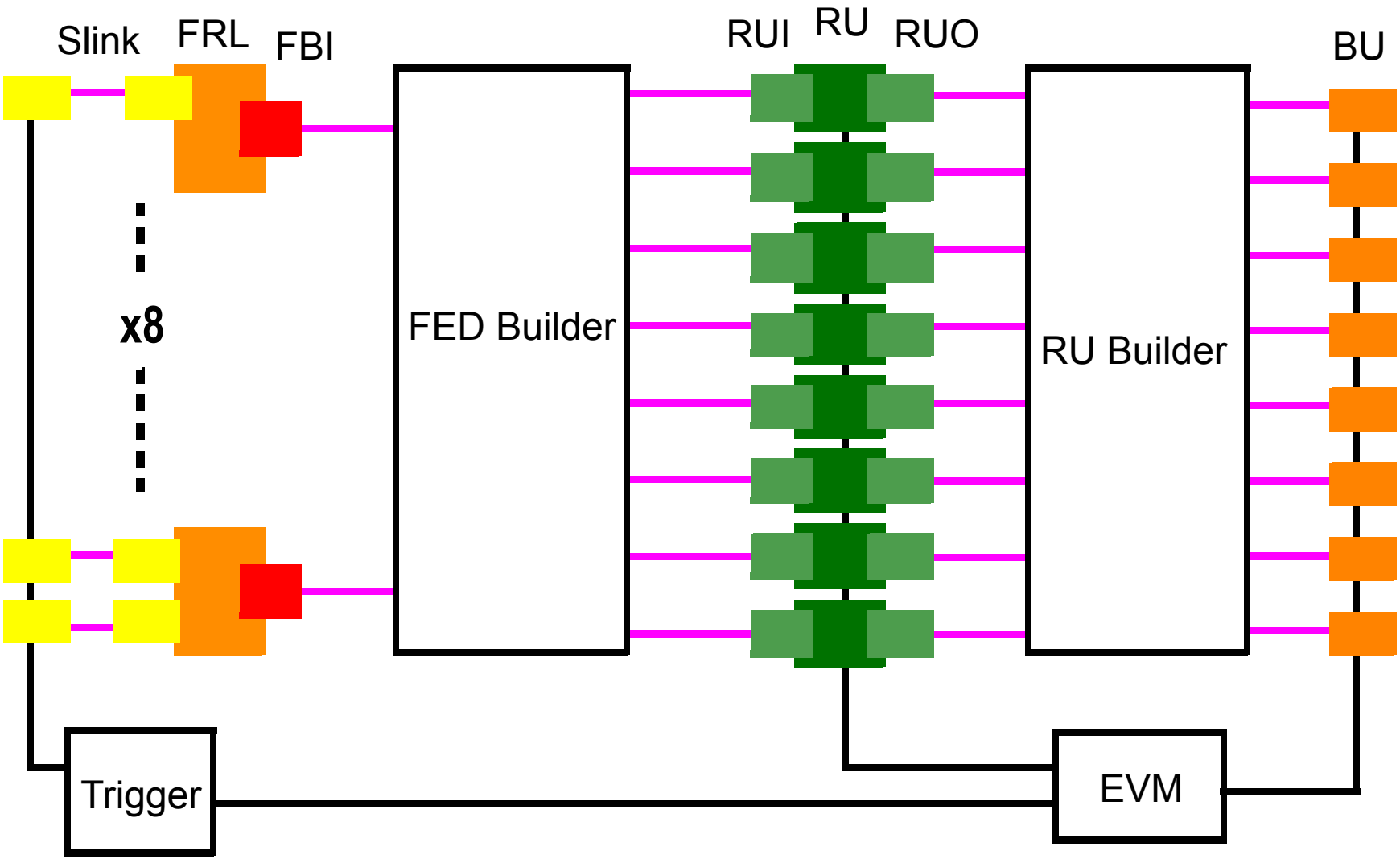
- Make sure that link is not the limitation (LANai10 has two 2.5 Gbit/s links)
- Code optimization
- Substitute GM by in house Myrinet firmware (MAZE)

- Next steps (until spring 2003):
  - Solve one pending issue for 32kB block sizes
  - Implement direct EVM protocol
  - Merge RU-bench with FRL bench
  - Complete system with SLINK - Merger - FRL - Myrinet FED-Builder / RUI - RU - BU
- Merge event builder Demonstrators with DAQ Column (until September 2003)

    2 pseudo FEDs (4 will be merged in FRL) with Slink
    8 FRLs
    8 x 8 FED Builder (possibly with LANai 10)
    8 RUI / RUs
    RU-Builder 8 x 8
    8 BUs
    Trigger Emulator, Event Manager

  - Later : merge with Filter Unit

DAQ Prototype



Conclusions

Achievements so far:

- SLINK64
  - Designed, built, tested
  - Fully functional
- FEDkit
  - SLINK64 sender, receiver and XDAQ software kit developed
  - Production for FED designers started
- FRL demonstrator
  - Meets bandwidth specifications
- RU demonstrator (PC based)
  - Meets requirements of 200 MB/s data throughput
- For next year a complete DAQ prototype is planned