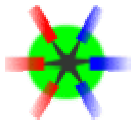
# High performance data acquisition with InfiniBand

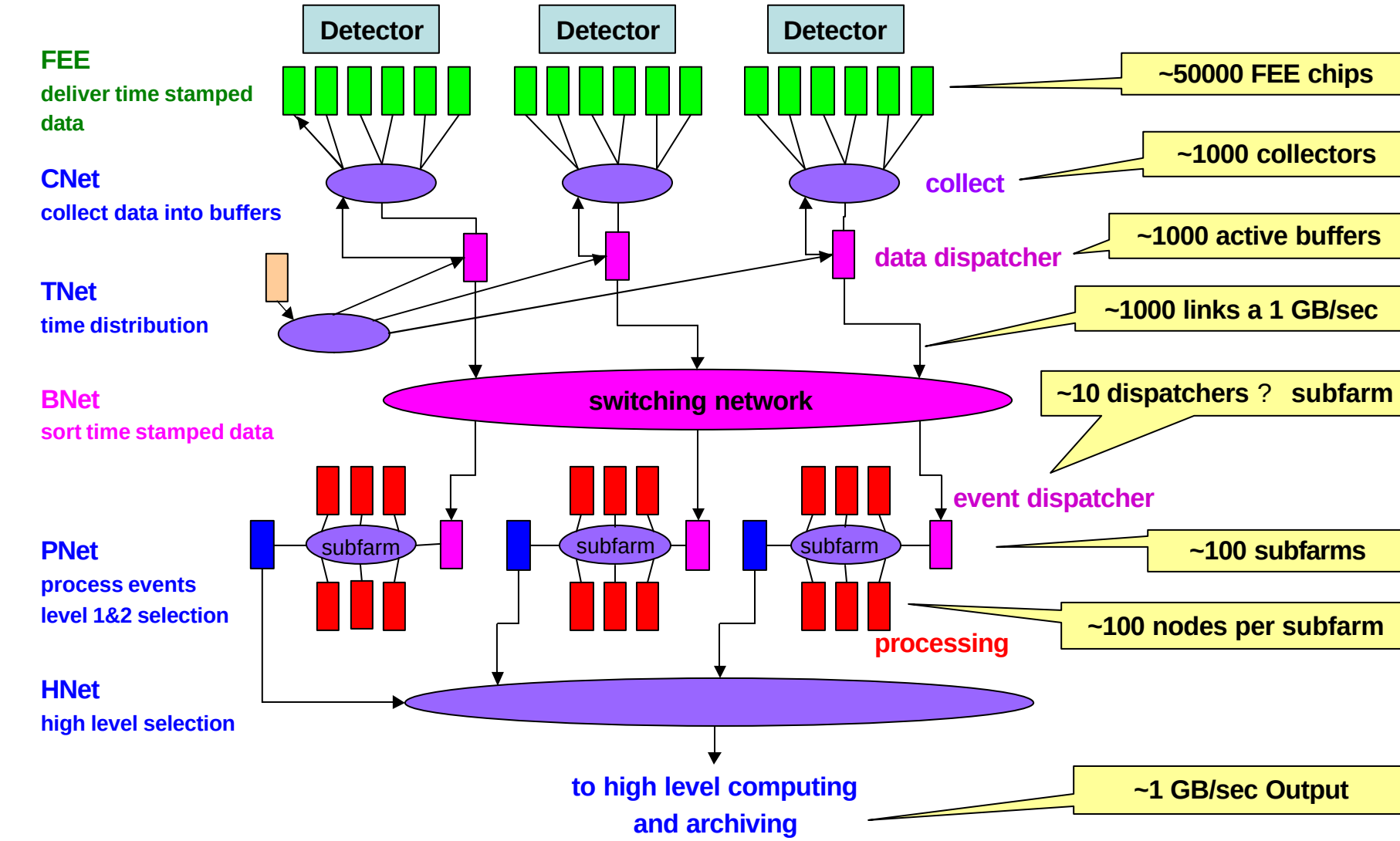
J.Adamczewski, H.G.Essel, N.Kurz, S.Linev
GSI, Experiment Electronics, Data Processing group



- CBM data acquisition
- Event building network
- InfiniBand & OFED
- Performance tests

Work supported by EU RP6 project JRA1 FutureDAQ RII3-CT-2004-506078

F.J.Müller, 2004

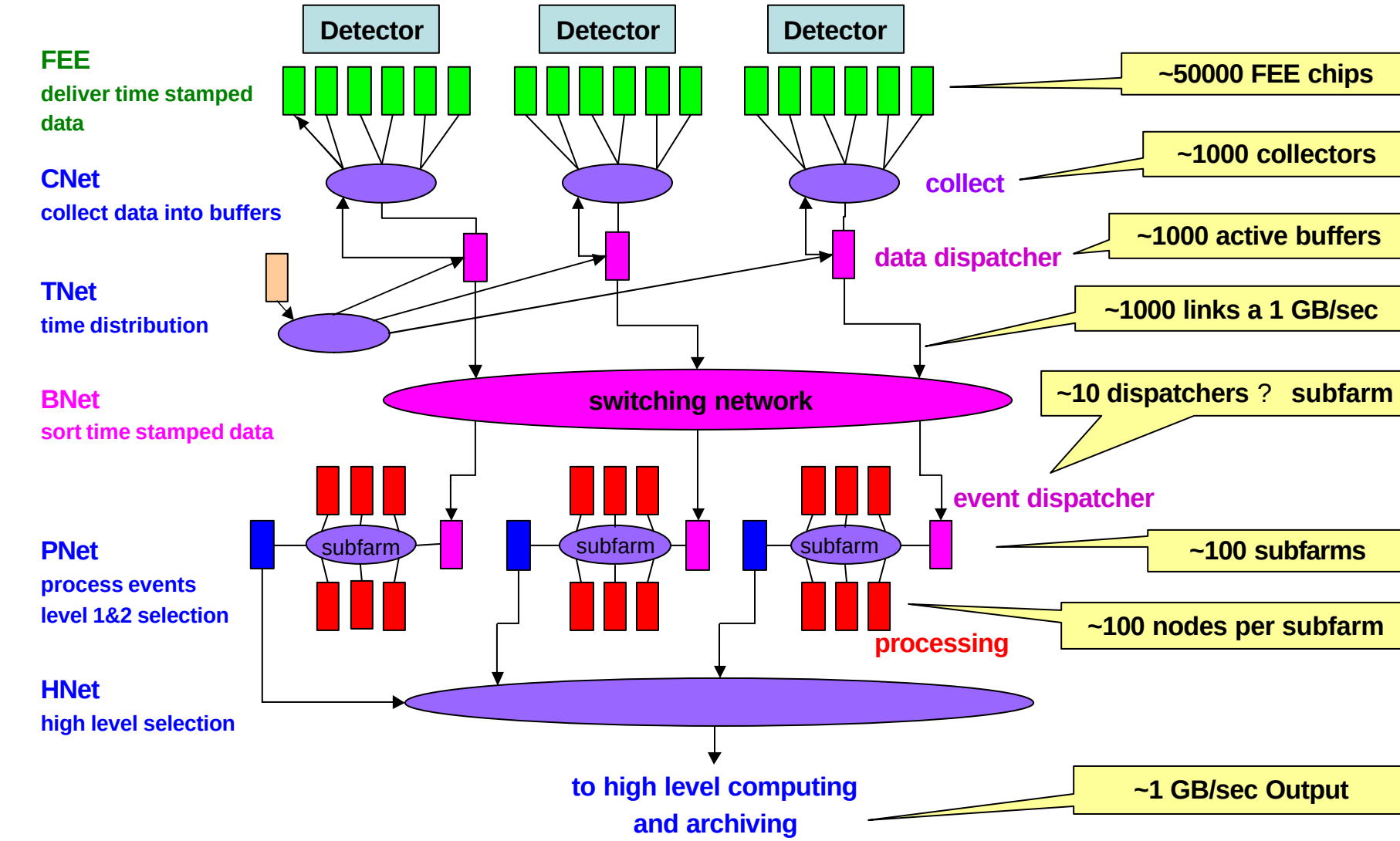


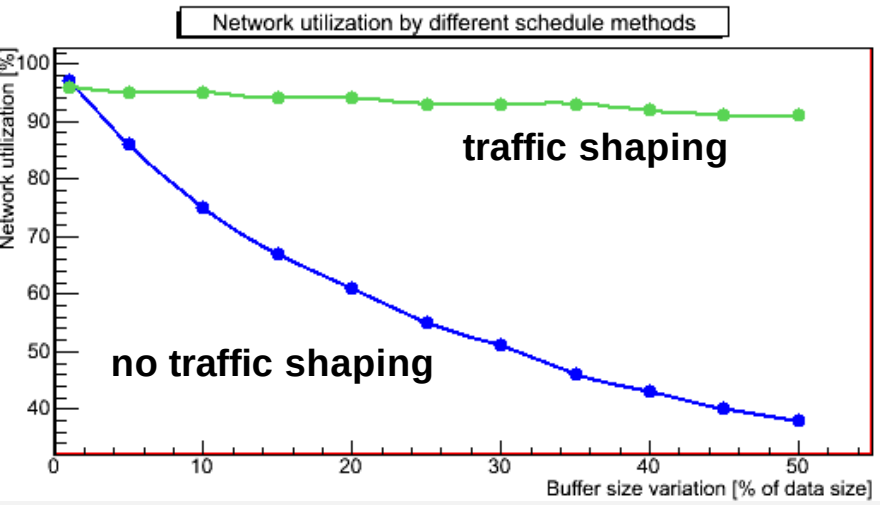
# CBM DAQ features summary

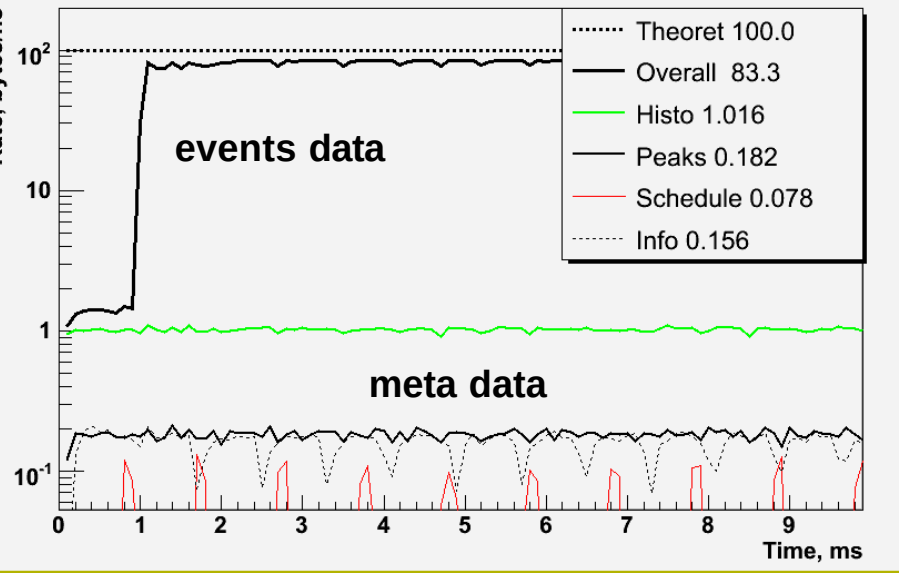
- Self-triggered time stamped data channels.
- Complex trigger algorithms ⇒ transport until filter farm
- FPGA controlled data flows
- Event building on full data rate ~1TB/s
- Event builder network BNet: ~1000 nodes, high-speed interconnections
- Linux may run on most DAQ nodes (even FPGAs)

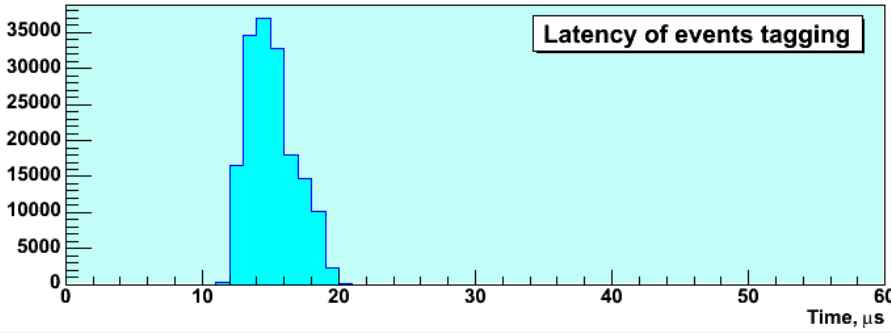
# Event building network - BNet

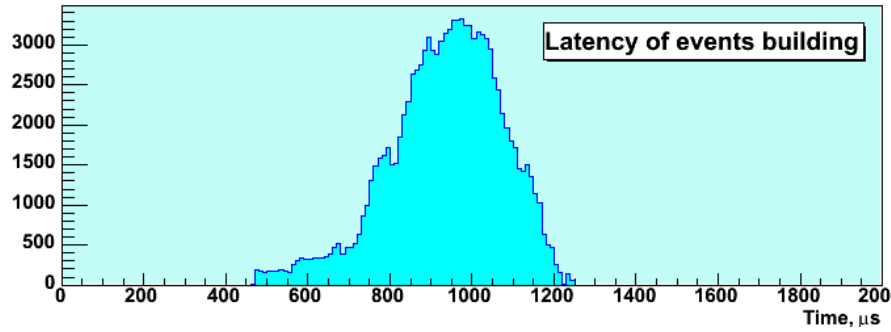
F.J.Müller, 2004




Network utilization by different schedule methods
Network utilization [%]
traffic shaping
no traffic shaping
Buffer size variation [% of data size]

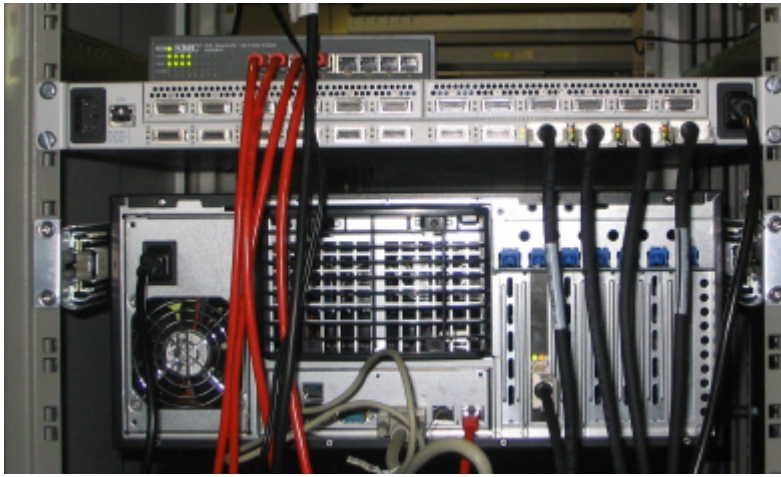

Theoret 100.0
Overall 83.3
Histo 1.016
Peaks 0.182
Schedule 0.078
Info 0.156
events data
meta data
Time, ms


Latency of events tagging
Time, μs


Latency of events building
Time, μs

High-performance interconnect technology

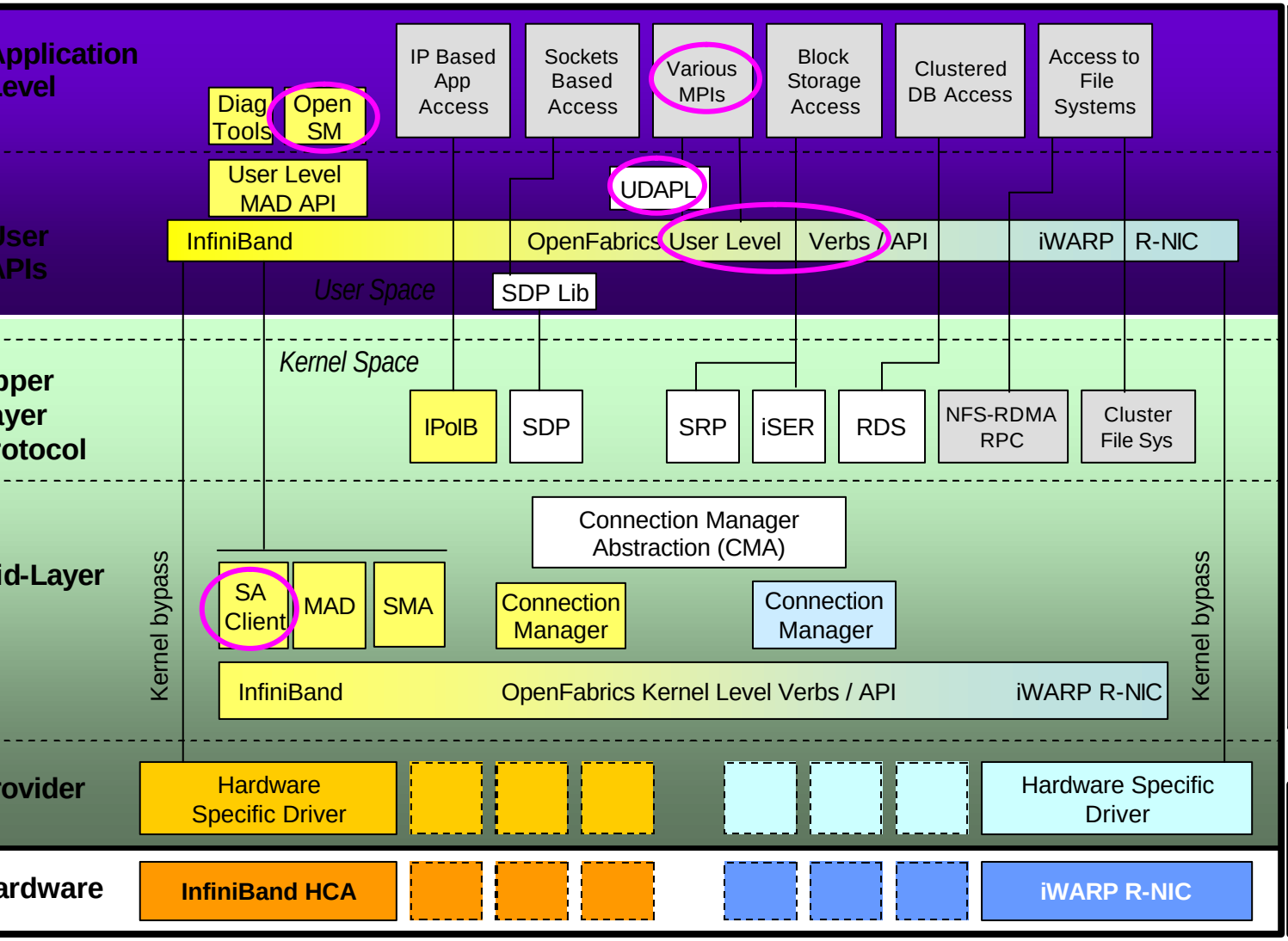
- switched fabric architecture
- up to 20 GBit/s bidirectional serial link
- Remote Direct Memory Access (RDMA)
- quality of service
- zero-copy data transfer
- low latency (few microseconds)



Main problem for a long time – lack of common API for different HW vendors

- now there is OpenFabrics (OFED) package solves this problem
- see www.openfabrics.org for more information

# OpenFabrics software stack*




| | |
|---|---|
| SA | Subnet Administra |
| MAD | Management Datagram |
| SMA | Subnet Manager Agent |
| PMA | Performance Manager Agent |
| IPoIB | IP over InfiniBand |
| SDP | Sockets Direct Protocol |
| SRP | SCSI RDMA Proto (Initiator) |
| iSER | iSCSI RDMA Proto (Initiator) |
| RDS | Reliable Datagram Service |
| UDAPL | User Direct Acces Programming Lib |
| HCA | Host Channel Adapter |
| R-NIC | RDMA NIC |

**Key**

| | |
|---|---|
| Common | Apps Acce Metho for us OF St |
| InfiniBand | |
| iWARP | |

* slide from www.openfabrics.o

# Used components from OFED

- uDAPL – direct access library, was used in our first tests, but has reduced functionality (only peer-to-peer, no multicast) compared to available in InfiniBand network
- VERBS – standard user-level API for arbitrary InfiniBand devices, provides full access to InfiniBand functionality
- SA – subnet administrator client, allows creation/management of multicast groups
- OpenSM – InfiniBand compliant Subnet Manager, involved in configuration and controls of the network

# MPI evaluation

- MPI – Message Passing Interface, was designed for high performance on both massively parallel machines and on workstation clusters
- MVAPICH – MPI & MPI2 over InfiniBand project. Supports:
  - non-blocking send/receive operation
  - true hardware-based multicast, but only with blocking API
- Tests of data throughput and multicast performance were done. Good results for big (larger than 32K) packets, but difficulty to combine normal and multicast traffic.
- Can be as option in DAQ system, while it is supported on majority of modern massively parallel machines, where different interconnect technologies are used.

# VERBS versus DAPL

- Both have very similar functionality and API:
  - memory, queues, completion events;
  - message and RDMA transport.

- But, VERBS provides extra functionality:
  - reliable/unreliable data transfer;
  - multicast support.

- We decide to switch to use OFED VERBS while:
  - it supports full InfiniBand functionality;
  - it is new official development line for Mellanox products.

- The only significant verbs problem – lack of good documentation.

**Aim:** Prove of principal as event building network candidate for CBM

**Tests last year:**

- GSI cluster - 4 nodes, SDR
- Forschungszentrum Karlsruhe* (March 2007) – 23 nodes, DDR
- UNI Mainz** (August 2007) - 110 nodes, DDR

Point-to-point tests

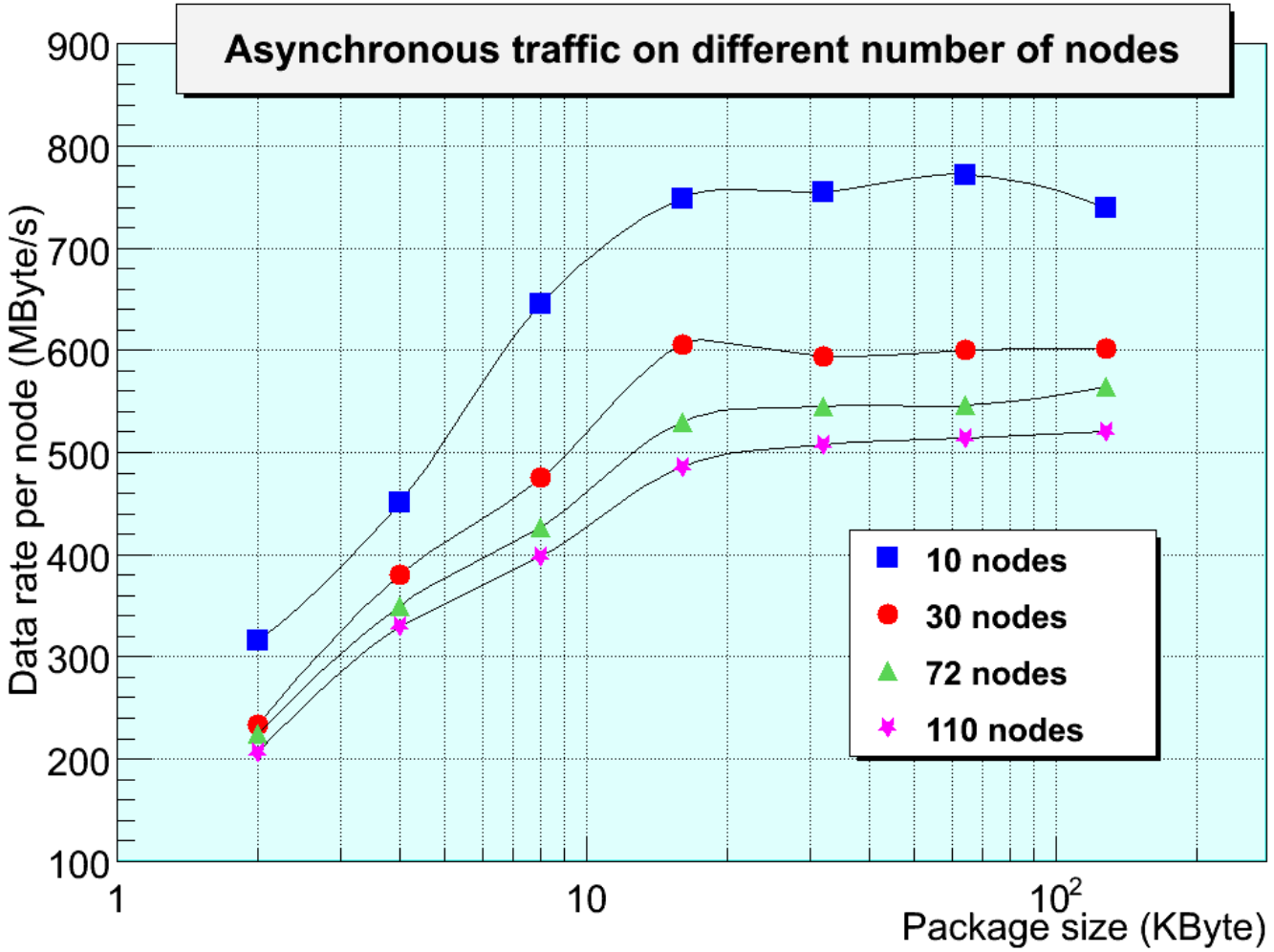
| | SDR (GSI) | DDR (Mainz) |
|---|---|---|
| Unidirectional | 0.98 GB/s | 1.65 GB/s |
| Bidirectional | 0.95 GB/s | 1.3 GB/s |

Multicast tests

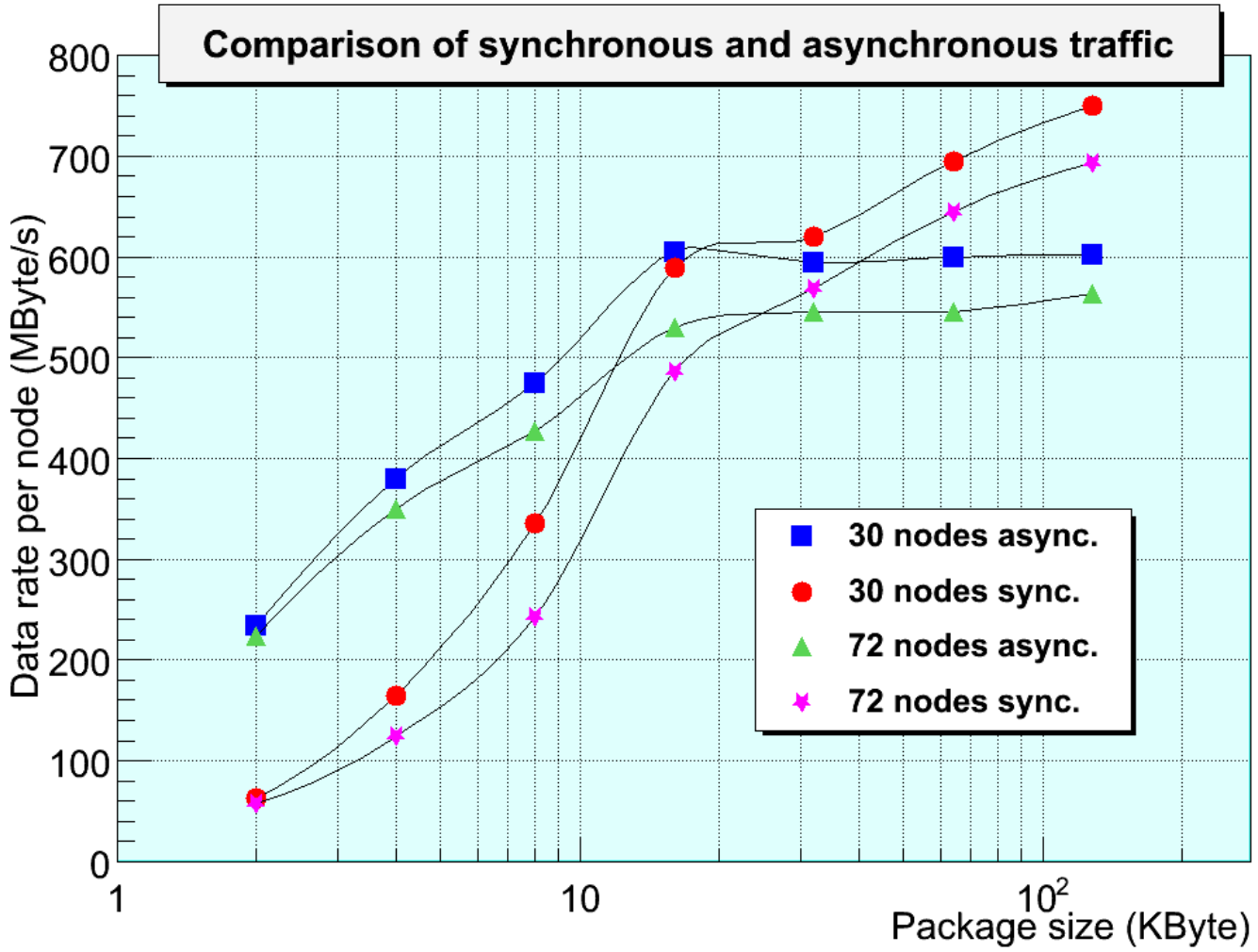
| | Rate per node |
|---|---|
| GSI (4 nodes) | 625 MB/s |
| Mainz (110 nodes) | 225 MB/s |

* thanks to Frank Schmitz, Ivan Kondov and Project CampusGrid in FZK

** thanks to Klaus Merle and Markus Tacke at the Zentrum für Datenverarbeitung in Uni Mainz


Asynchronous traffic on different number of nodes
Data rate per node (MByte/s)
900
800
700
600
500
400
300
200
100
1
10
$10^2$
Package size (KByte)
10 nodes
30 nodes
72 nodes
110 nodes

# Chaotic (async.) versus scheduled (sync.)



- DABC – Data Acquisition Backbone Core, new software development for general-purpose DAQ in GSI

- DABC fully supports InfiniBand as data transport between nodes
  - connection establishing
  - memory management for zero-copy transfer
  - back-pressure
  - errors detection

# Conclusion

- InfiniBand is a good candidate for CBM event building network
- Up to 700 MB/s bidirectional data rate is achieved on 110 nodes cluster
- Mixture of point-to-point and multicast traffic is possible
- Further investigation of scheduling mechanisms for InfiniBand is required