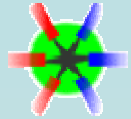


High Performance Event Building over **InfiniBand** Networks



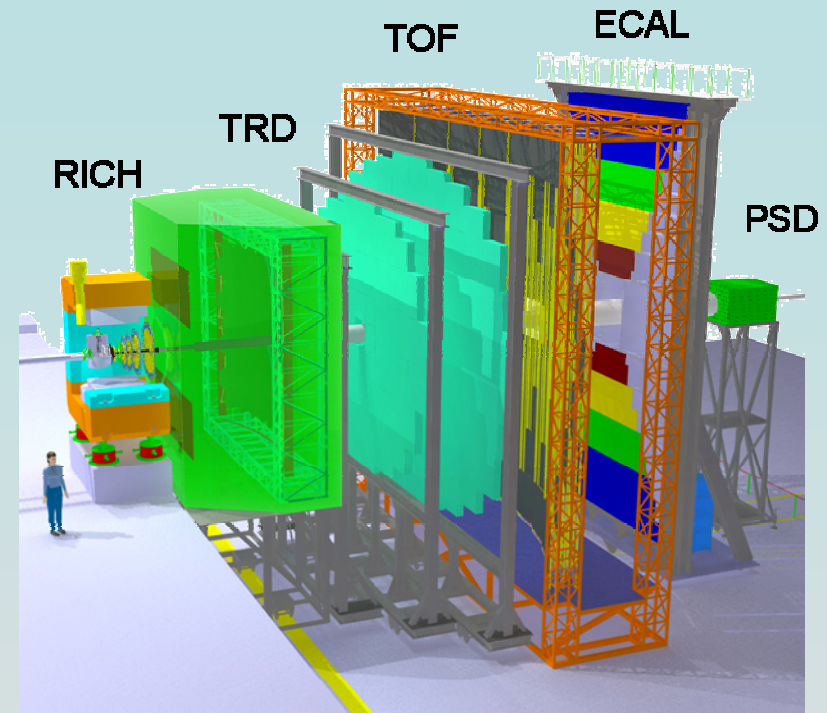
J.Adamczewski-Musch, H.G.Essel, N.Kurz, S.Linev

GSI Helmholtzzentrum für Schwerionenforschung GmbH
Experiment Electronics, Data Processing group

- **CBM data acquisition**
- **Event building network**
- **InfiniBand & OFED**
- **Performance tests**
- **Data Acquisition Backbone Core DABC**

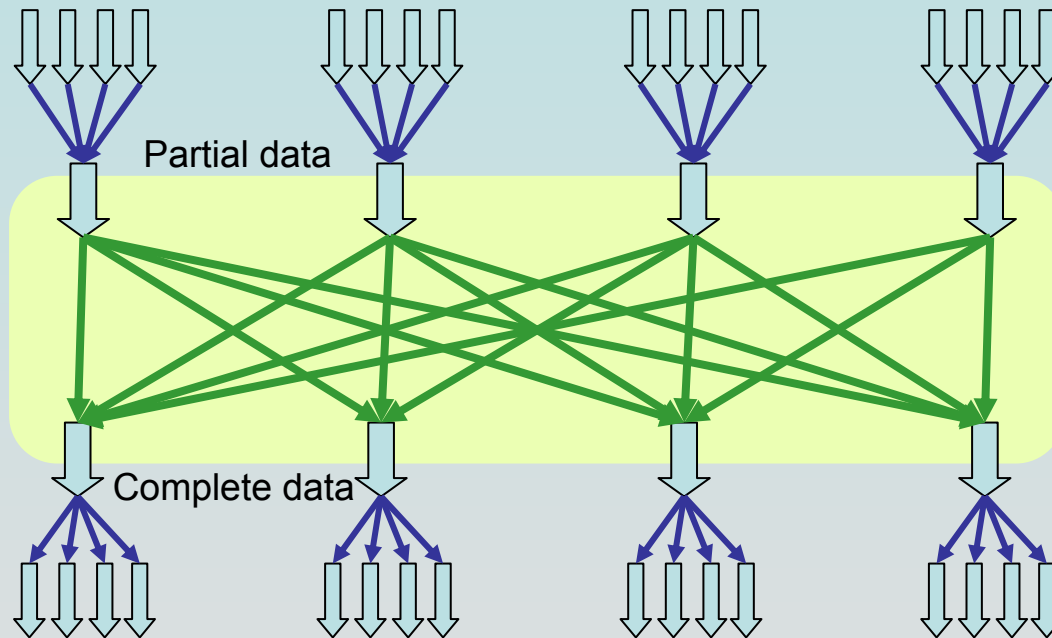
Work supported by [EU RP6 project JRA1 FutureDAQ](#) RII3-CT-2004-506078

- Complex trigger algorithms on full data:**
- ⇒ Self-triggered front-end electronics
 - ⇒ Time stamped data channels.
 - ⇒ Transport full data into filter farm
 - ⇒ Data sorting over switched network on full data rate of $\sim 1\text{TB/s}$
 - ⇒ Sorting network: ~ 1000 nodes



Is that possible (in 2012)?

Detector electronics, time stamped data channels



Merge channels

Up to 1000 lines
1 GByte/sec each

Sort over
switched network,
**units are not events,
but time slice data!**

Distribute
complete data

Processor farms, event definition, filtering, archiving

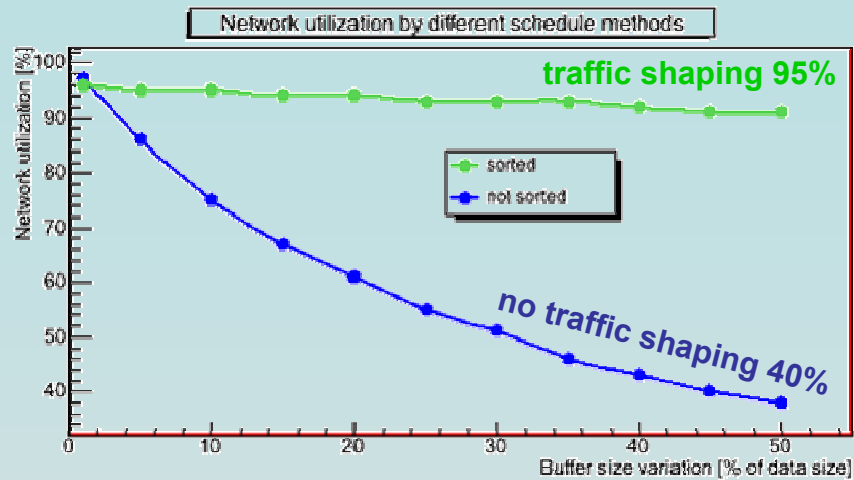
Software packages developed:

1. **Simulation** with SystemC (flow control, scheduling)
 - Meta data on data network
2. **Real dataflow core** (round robin, with/without sychronization)
 - Linux, InfiniBand, GB Ethernet
 - Simulates data sources
3. **Data Acquisition Backbone Core DABC** (includes dataflow core)
 - Controls, Configuration, Monitoring, GUI ...
 - Real data sources
 - General purpose DAQ framework

Poster N30-98: (Wednesday, 10:30)

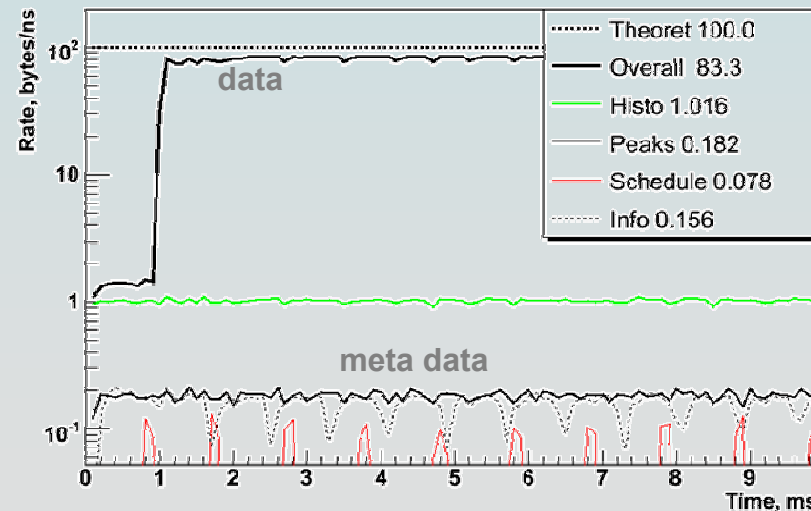
[DABC, a Data Acquisition Backbone Core Library.](#)

Let's have a look to these



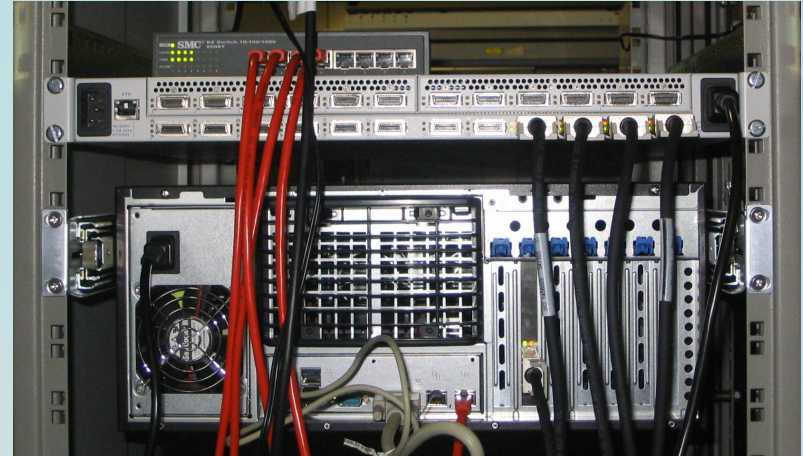
Optimization of network utilization for variable sized buffers.
Enhancement up to factor two.

Same network used for data and meta data.
Occupancy of network by different kind of data



High-performance interconnect technology

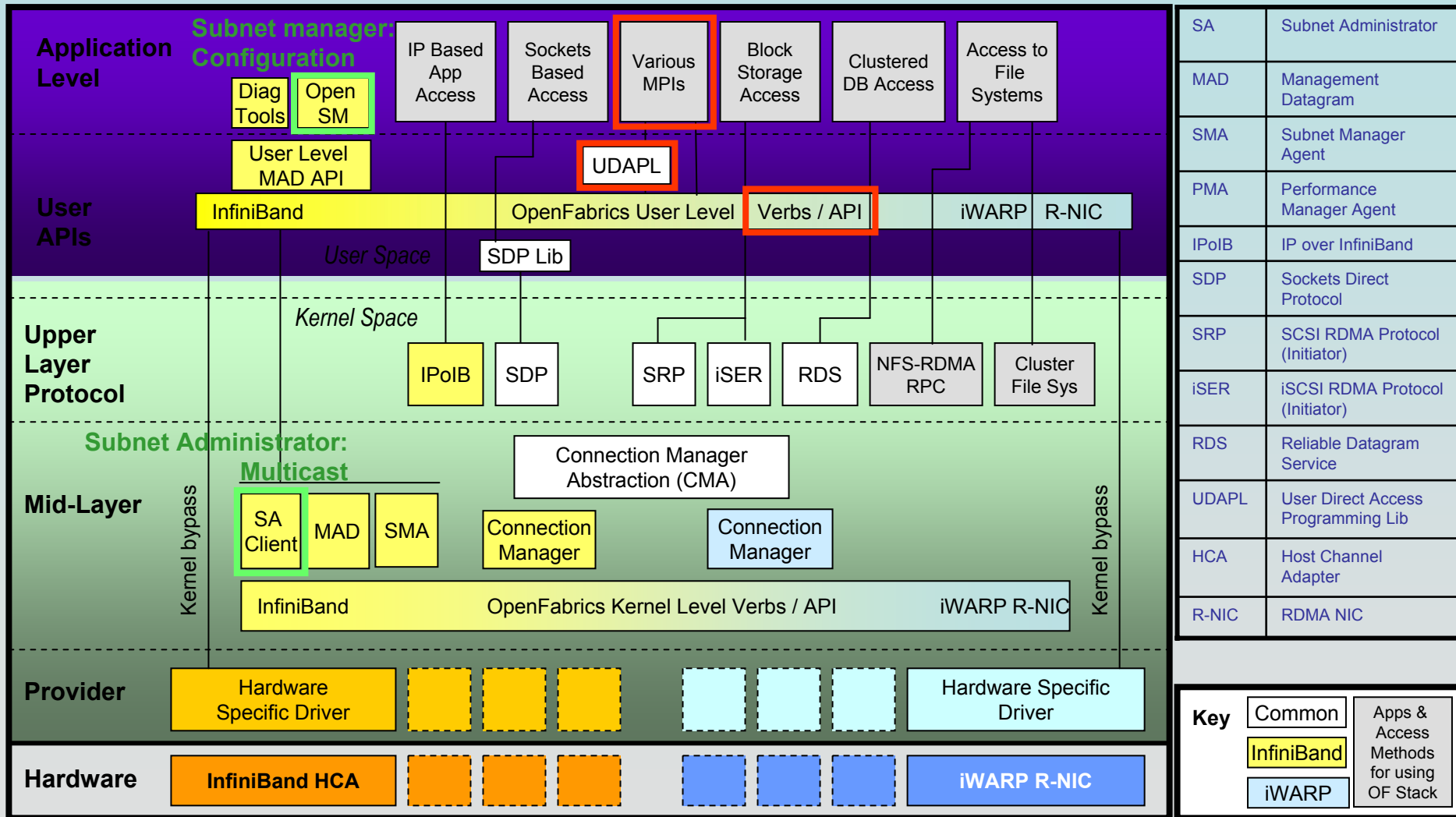
- switched fabric architecture
- up to **20 GBit/s** bidirectional serial link
- Remote Direct Memory Access (**RDMA**)
- quality of service
- **zero-copy** data transfer
- low latency (few microseconds)
- common in HPC



InfiniBand switch of the GSI cluster

Software library:

- **OpenFabrics** (OFED) package
see www.openfabrics.org for more information



* slide from www.openfabrics.org

Hardware for testing:

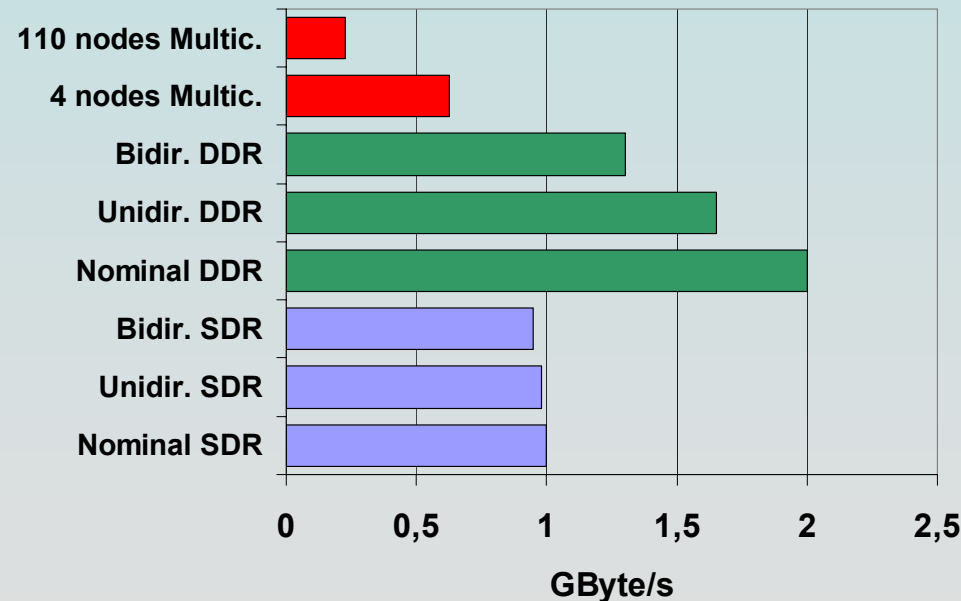
- GSI (November 2006) - 4 nodes, single data rate SDR
- Forschungszentrum Karlsruhe* (March 2007) – 23 nodes, double data rate DDR
- UNI Mainz** (August 2007) - 110 nodes, DDR

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

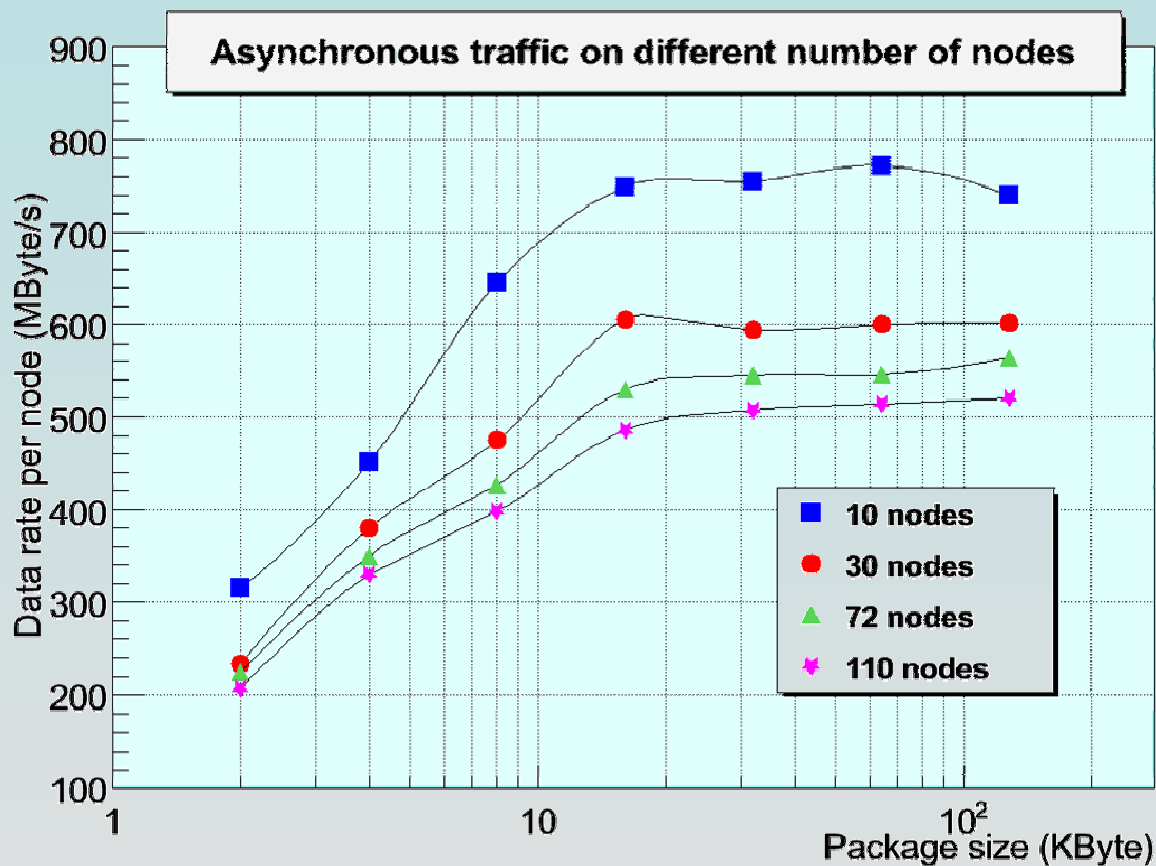
Multicast tests

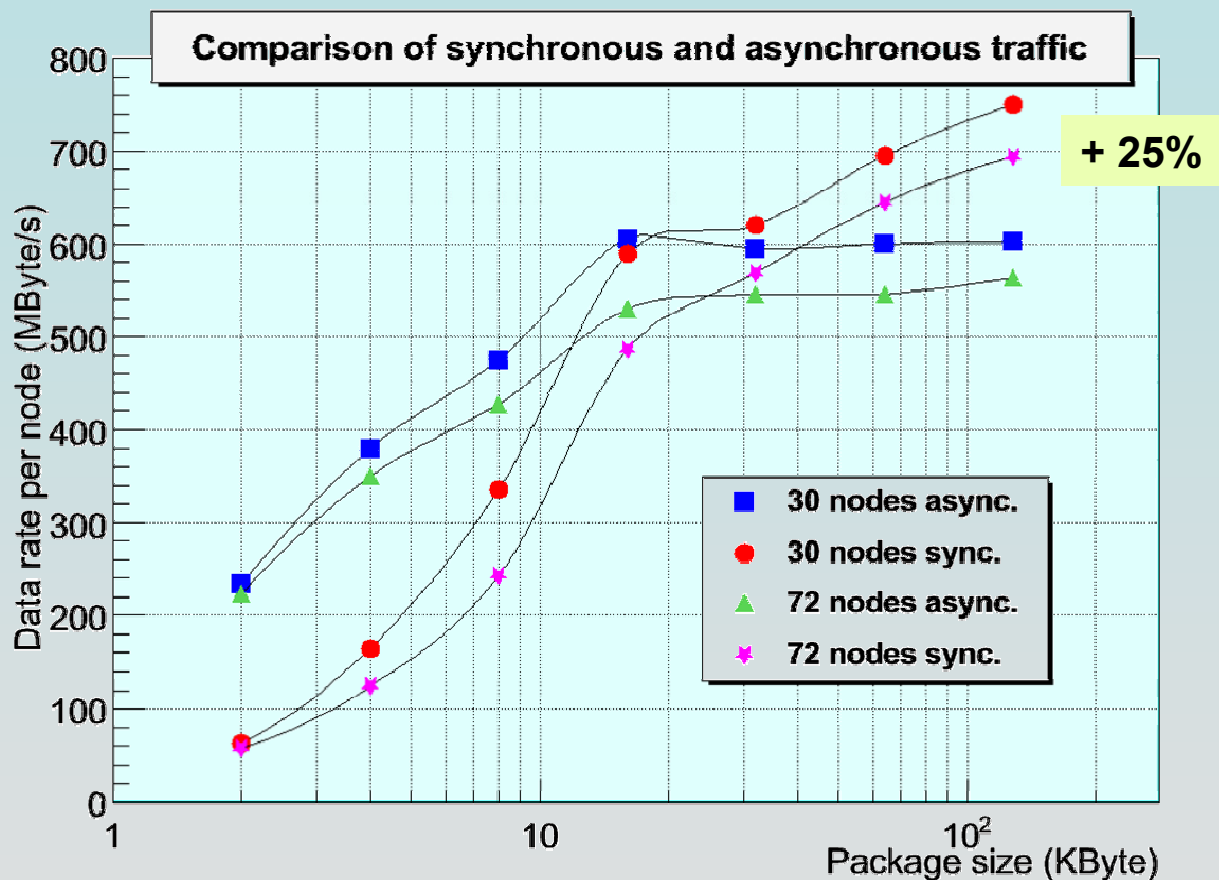
	Rate per node
GSI (4 nodes)	0.625 GB/s
Mainz (110 nodes)	0.225 GB/s



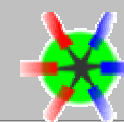
* thanks to Frank Schmitz, Ivan Kondov and Project CampusGrid in FZK

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





Topology optimization for 110 nodes would have required different cabling of switches!



2004 → EU RP6 project JRA1 FutureDAQ*

2004 → CBM FutureDAQ for FAIR

1996 → MBS future
50 installations at GSI,
50 external
<http://daq.gsi.de>

Intermediate
demonstrator

Use cases

- Detector tests
- FE equipment tests
- Data transport
- Time distribution
- Switched event building
- Software evaluation
- MBS event builder
- General purpose DAQ

Requirements

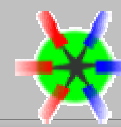
- **build events over fast networks**
- handle triggered or self-trigger front-ends
- process time stamped data streams
- **provide data flow control (to front-ends)**
- connect (nearly) any front-ends
- provide interfaces to plug in application codes
- connect MBS readout or collector nodes
- be controllable by several controls frameworks

**Data
Acquisition
Backbone
Core**

* RII3-CT-2004-506078

Poster N30-98: (Wednesday, 10:30)

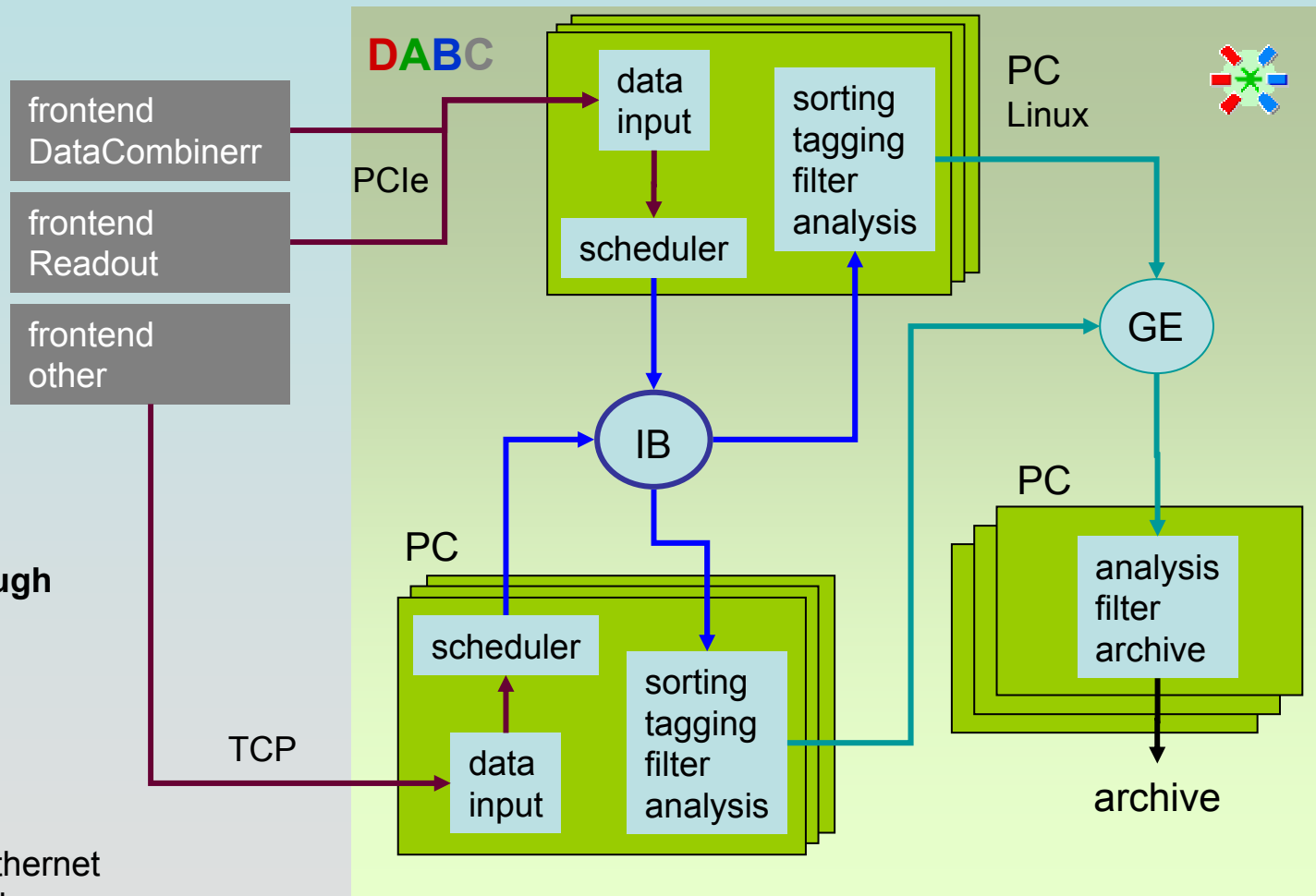
[DABC, a Data Acquisition Backbone Core Library.](http://dabc.gsi.de)



From this, **DABC** evolved as a general purpose DAQ framework.
The main application areas of **DABC** are:

1. **High speed event building over fast networks**
2. **Front-end readout chain tests (CBM, September 2008)**
3. **DABC as MBS* event builder (Ready)**
4. **DABC with mixed, triggered (MBS) and time stamped, data channels**
 - Needs Synchronization of between both
 - Insert event number from trigger to time stamped data stream
 - Read out time stamp from MBS via VME module (to be built)

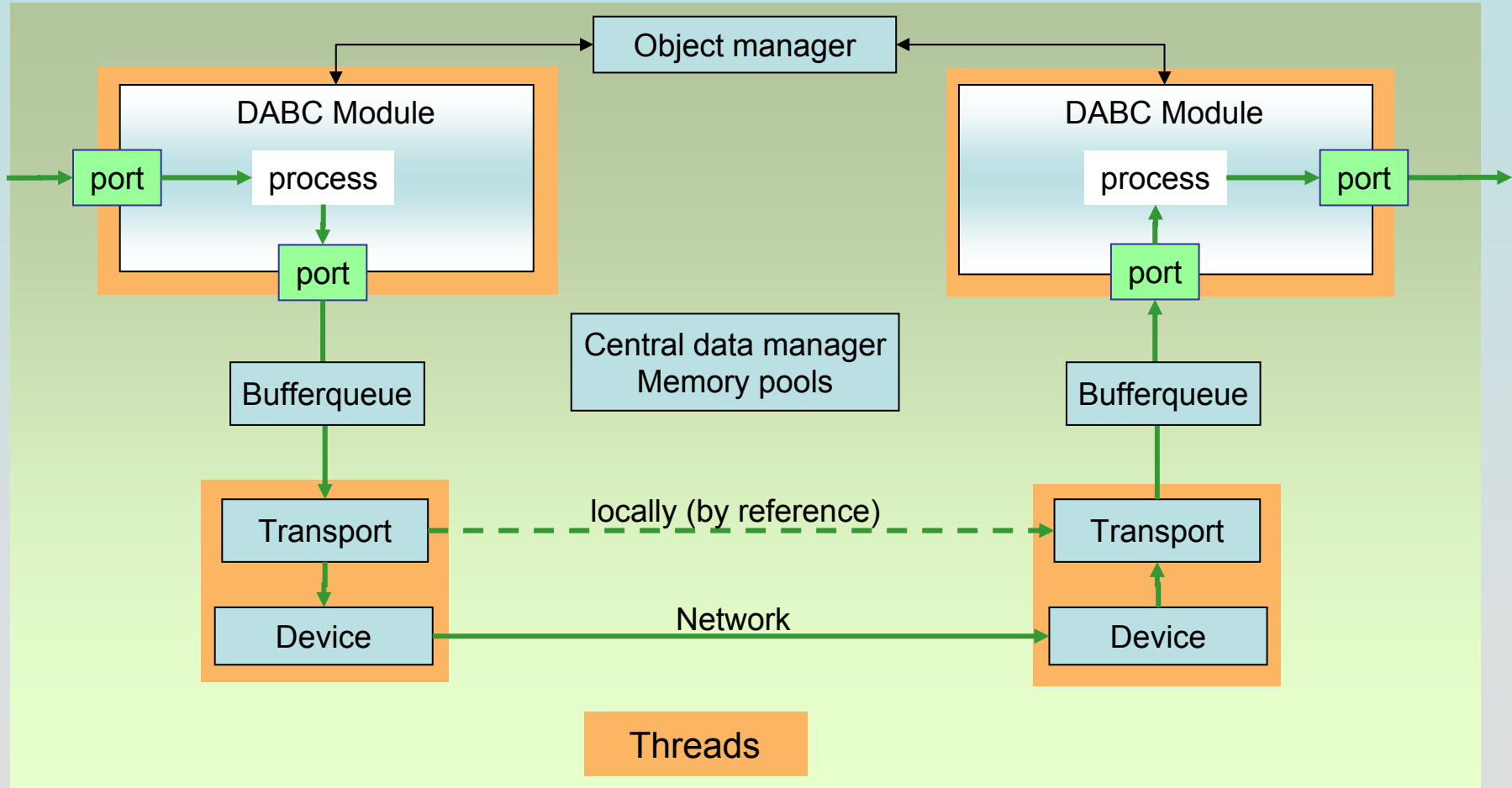
* **MultiBranchSystem**: standard DAQ system at GSI

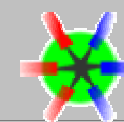


GE: Gigabit Ethernet
IB: InfiniBand

A *module* processes data of one or several data streams.

Data streams propagate through **ports**, which are connected by **transports** and **devices**

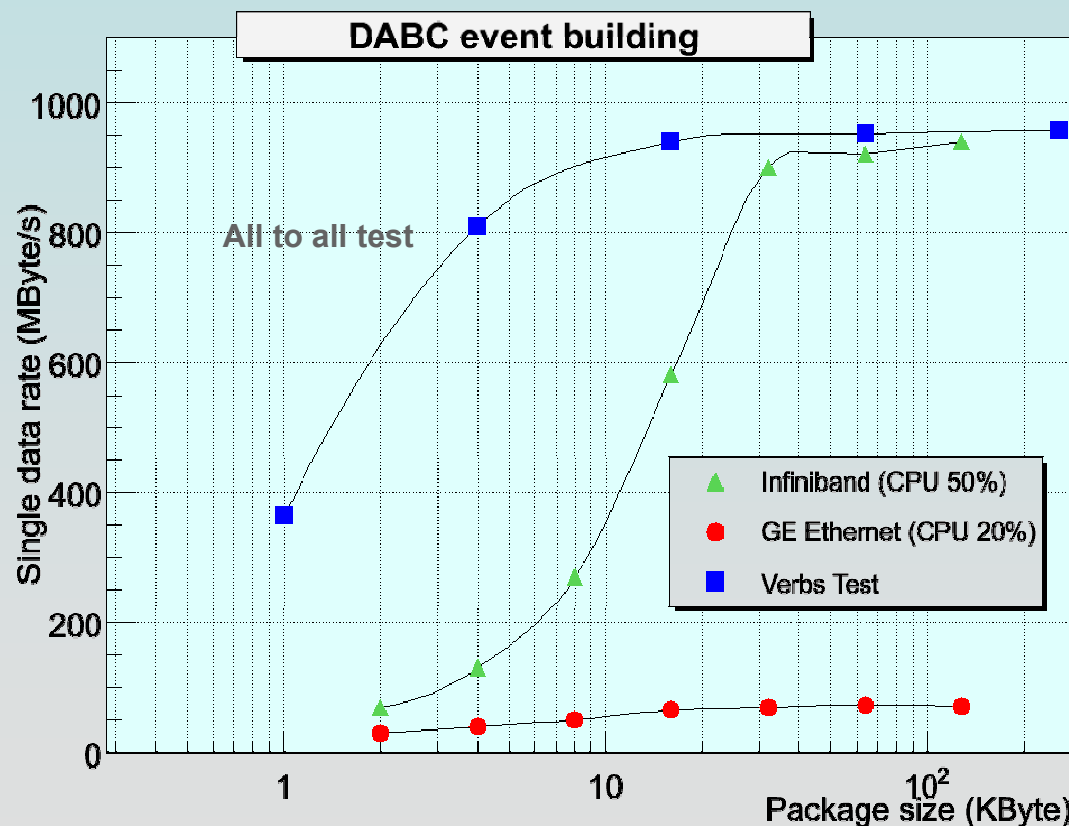




DABC fully supports InfiniBand as data transport between nodes

- connection establishing
- memory management for zero-copy transfer
- back-pressure
- errors handling
- thread sharing

GSI InfiniBand cluster
with four nodes SDR



- **InfiniBand** is a good candidate for fast event building network
- Up to **520 MB/s** bidirectional data rate is achieved on **110** nodes without optimization.
- Mixture of point-to-point and **multicast** traffic is possible
- Further investigation of **scheduling** mechanisms is required
- Further investigation of network **topology** is required

Thank You!