

# The apeNEXT project

F. Rapuano  
INFN  
and  
Dip. di Fisica  
Università Milano-Bicocca

for the APE Collaboration



INFN Ferrara, Rome



DESY Zeuthen

Bielefeld  
University



Université de Paris-Sud, Orsay

# The Group

- **Italy**
  - Roma: N.Cabibbo, F. di Carlo, A. Lonardo, S. de Luca, D. Rossetti, P. Vicini
  - Ferrara: L. Sartori, R. Tripiccione, F. Schifano
  - Milano-Parma: R. de Pietri, F. di Renzo, F. Rapuano
- **Germany**
  - DESY, NIC: H. Kaldass, M. Lukyanov, N. Paschedag, D. Pleiter, H. Simma
- **France**
  - Orsay: Ph. Boucaud, J. Micheli, O. Pene,
  - Rennes: F. Bodin

# Outline of the talk

apeNEXT is completely operational  
and all its circuits are functioning perfectly  
Mass production starting

- A bit of history
- A bit of HW
- A bit of SW
- Large installations
- Future plans

# The Ape paradigm

- Very efficient for LQCD (up to 65% peak), but usable for other fields
  - The “normal” operation as basic operation  
 $a*b+c$  (complex)
- Large number of registers for efficient optimization, Microcoded architecture (VLIW)
- Reliable and safe HW solutions
- Large software effort for programming and optimization tools

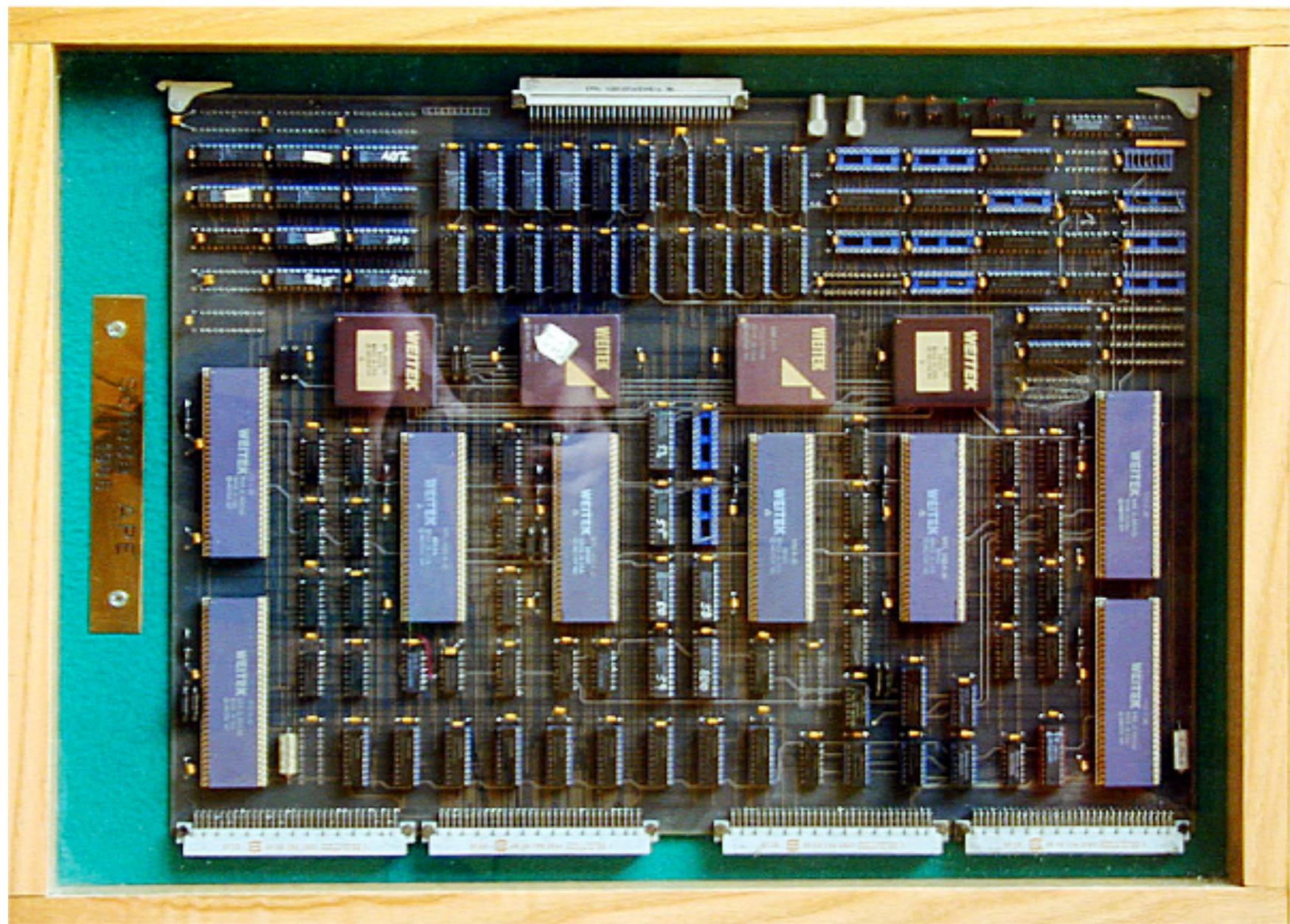
# “The APE family”

## *Our line of Home Made Computers*

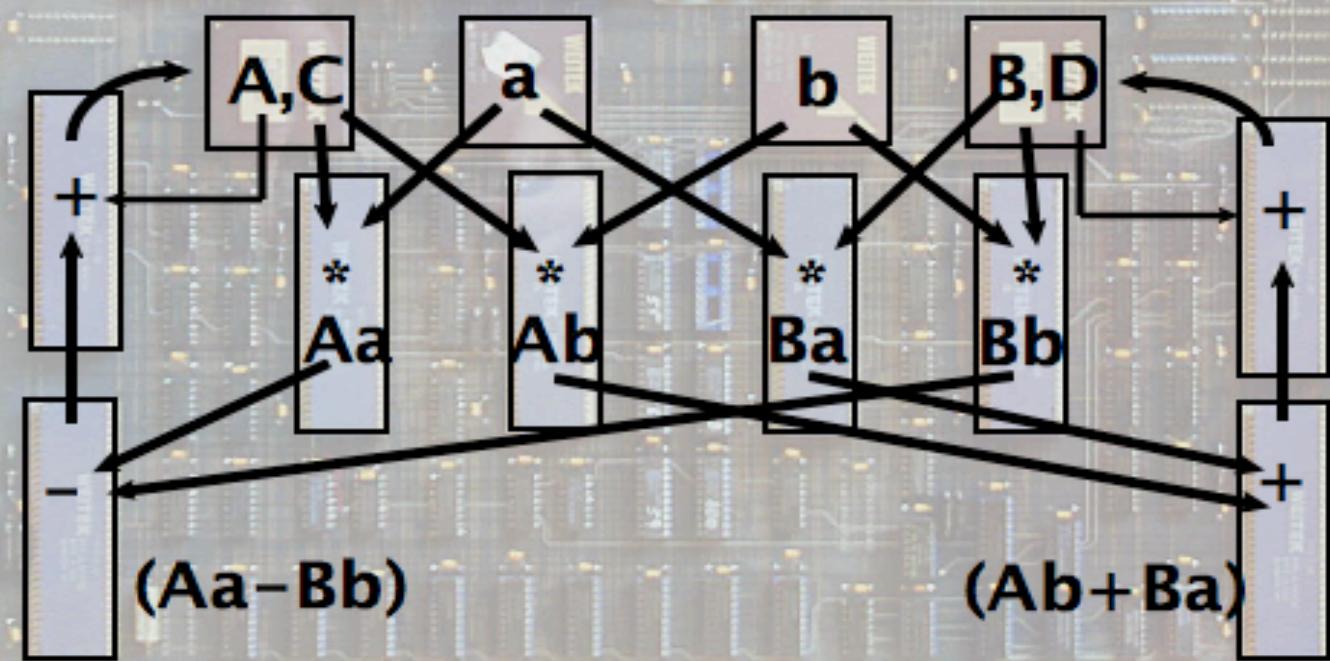
Once upon a time (1984) italian lattice physicists were sad ...

The diagram illustrates the evolution of the APE family. It starts with the text "Once upon a time (1984) italian lattice physicists were sad ...". Below this, two green arrows point downwards to specific rows in a table. The first arrow points to the row for "APE(1988)", which corresponds to the event "Home made VLSI begins". The second arrow points to the row for "apeNEXT(2004)", which corresponds to the event "EU collaboration". The table compares five generations of the APE family: APE(1988), APE100(1993), APEmille(1999), and apeNEXT(2004). The table includes columns for Architecture, # nodes, Topology, Memory, # registers (w.size), clock speed, and peak speed.

	APE(1988)	APE100(1993)	APEmille(1999)	apeNEXT(2004)
Architecture	SISAMD	SISAMD	SIMAMD	SIMAMD+
# nodes	16	2048	2048	4096
Topology	flexible 1D	rigid 3D	flexible 3D	flexible 3D
Memory	256 MB	8 GB	64 GB	1 TB
# registers (w.size)	64 (x32)	128 (x32)	512 (x32)	512 (x64)
clock speed	8 MHz	25 MHz	66 MHz	200 MHz
peak speed	1 GFlops	100 GFlops	1 TFlops	7 TFlops



$$(a+ib)*(A+iB)+C+iD \\ =(Aa-Bb+C)+i(Ab+Ba+D)$$

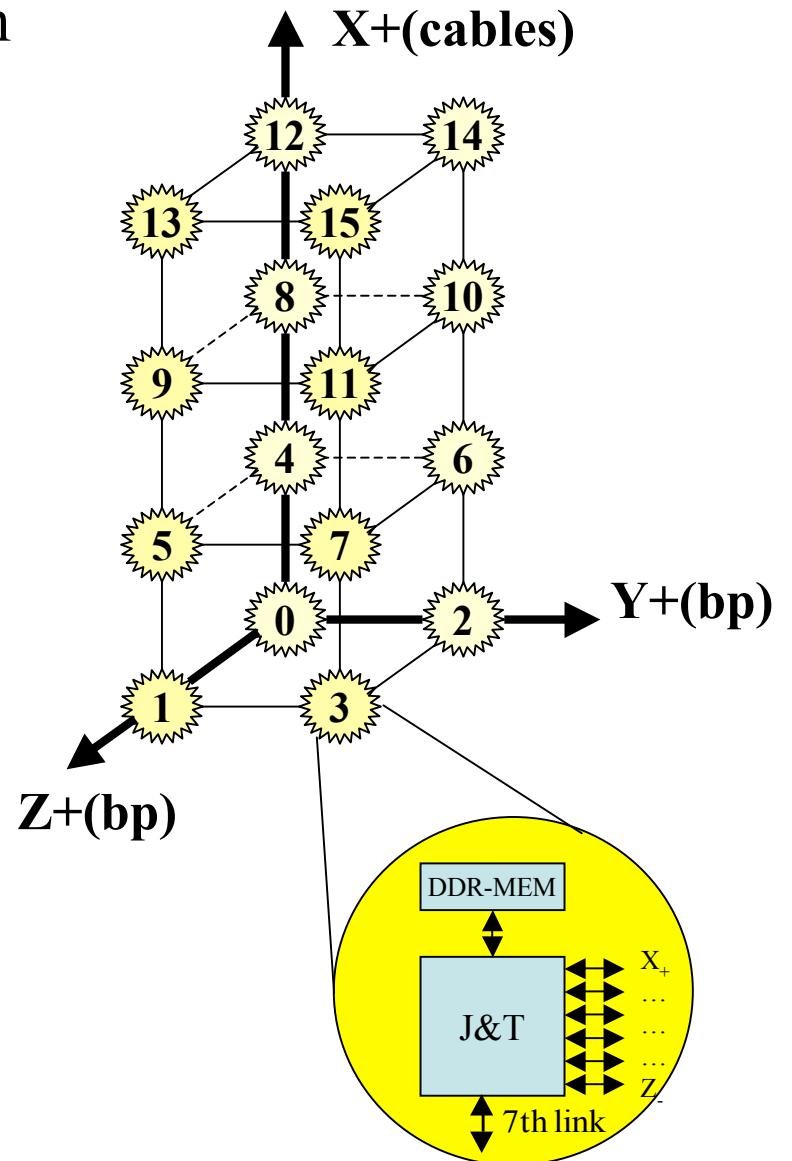


# Lattice conferences as status checkpoints

- Lattice 2000 (Bangalore) FR, general ideas
- Lattice 2001 (Berlin) R. Tripiccione, clear ideas, VLSI design started, simulator running
- Lattice 2002 (Boston) D. Pleiter, all designs complete, most HW prototypes ready, VLSI design complete
- Lattice 2003 (Tsukuba) no talk, 6 months delay in VLSI. Delivered in December

# apeNEXT Architecture

- 3D mesh of computing nodes, 64bit arithm
  - Custom VLSI processor - **200 MHz** (J&T)
  - **512 registers**
  - **1.6 GFlops** per node (complex “normal”)
  - **256 MB ÷ 1 GB** memory per node
  - **3.2 GB/s** memory bandwidth (128 bit chan)
  - **Prefetch queues**
- First neighbor communication network  
loosely synchronous (fifo based)
  - $\rho = 8 \div 16 \Rightarrow \text{200 MB/s}$  per channel
- Scalable  $25 \text{ GFlops} \div 7 \text{ Tflops}$   
 $16 \div 4096$  nodes
- Linux PCs as Host system



# Topology

- Two directions (Y,Z) on the backplane
- Direction X through front panel cables

## • System topologies:

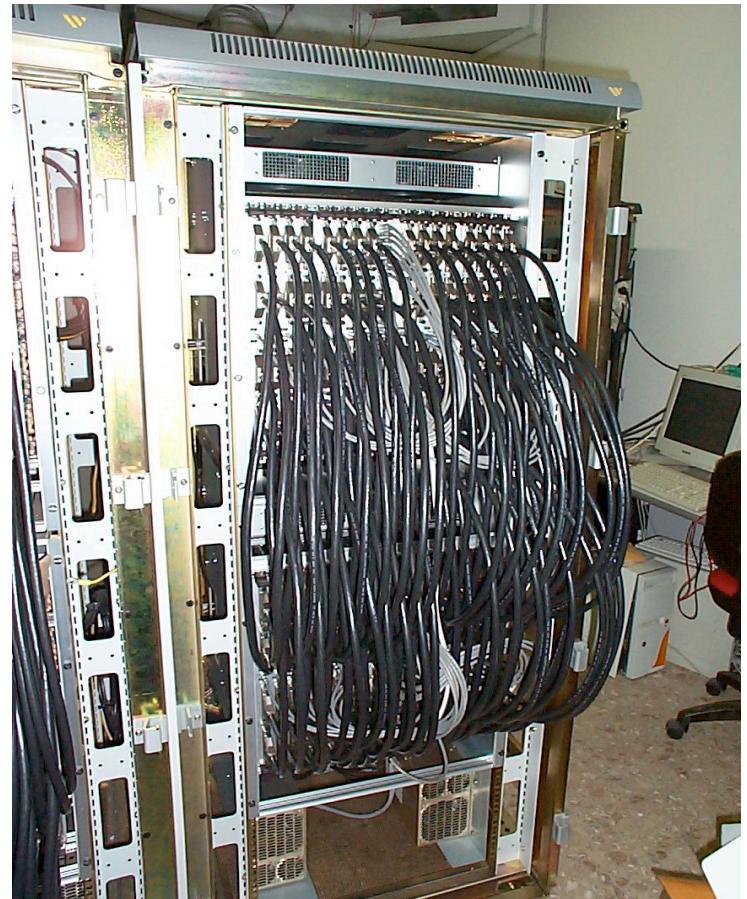
- Processing Board
- subCrate (16 PB)
- Crate (32 PB)
- Large systems

**4 x 2 x 2 ~ 26 GF**

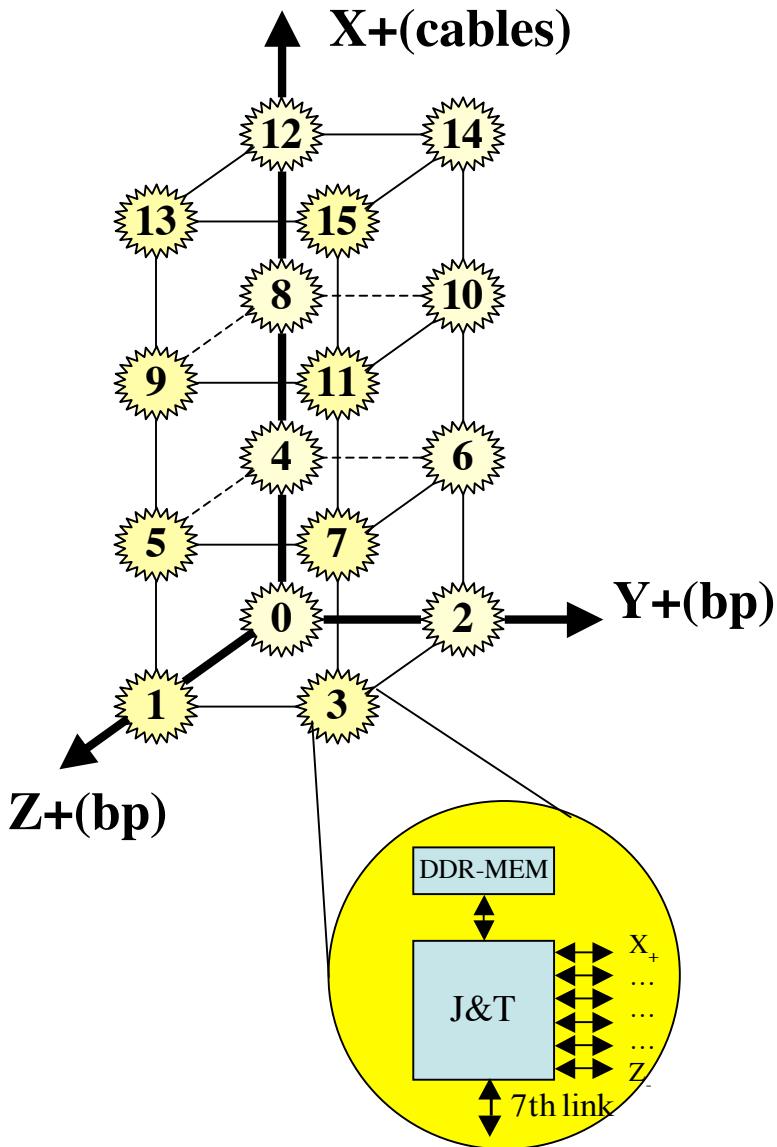
**4 x 8 x 8 ~ 0.4 TF**

**8 x 8 x 8 ~ 0.8 TF**

**(8\*n) x 8 x 8**



# PB

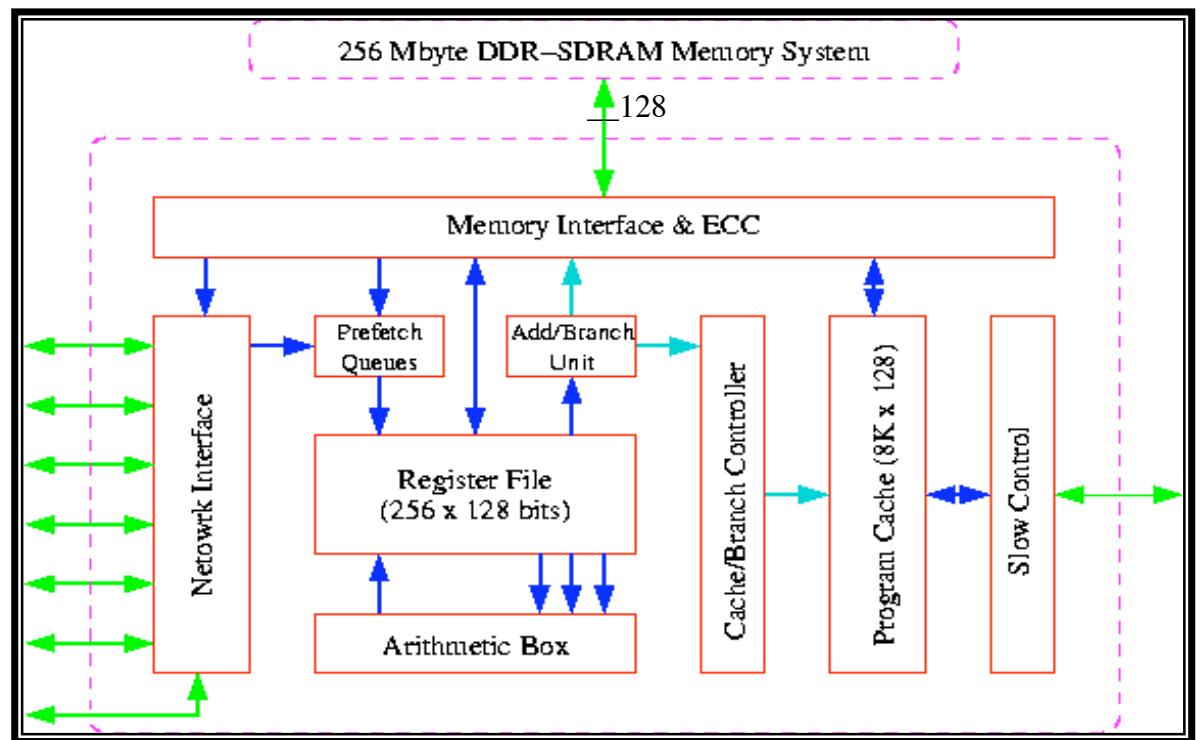


- 16 Nodes 3D-Interconnected
- 4x2x2 Topology **26 Gflops, 4.6 GB Memory**
- Light System:
  - J&T Module connectors
  - Glue Logic (Clock tree 10Mhz)
  - Global signal interconnection (FPGA)
  - DC-DC converters (48V to 3.3/2.5/1.8 V)
- Dominant Technologies:
  - LVDS: **1728** ( $16 \times 6 \times 2 \times 9$ ) differential signals 200MB/s, 144 routed via cables, 576 via backplane on 12 controlled-impedance ( $100\Omega$ ) layers
  - High-Speed differential connectors:
    - Samtec QTS (J&T Module)
    - Erni ERMET-ZD (Backplane)
- Collaboration with **NEURICAM spa**

# J&T

Computing & control  
integrated

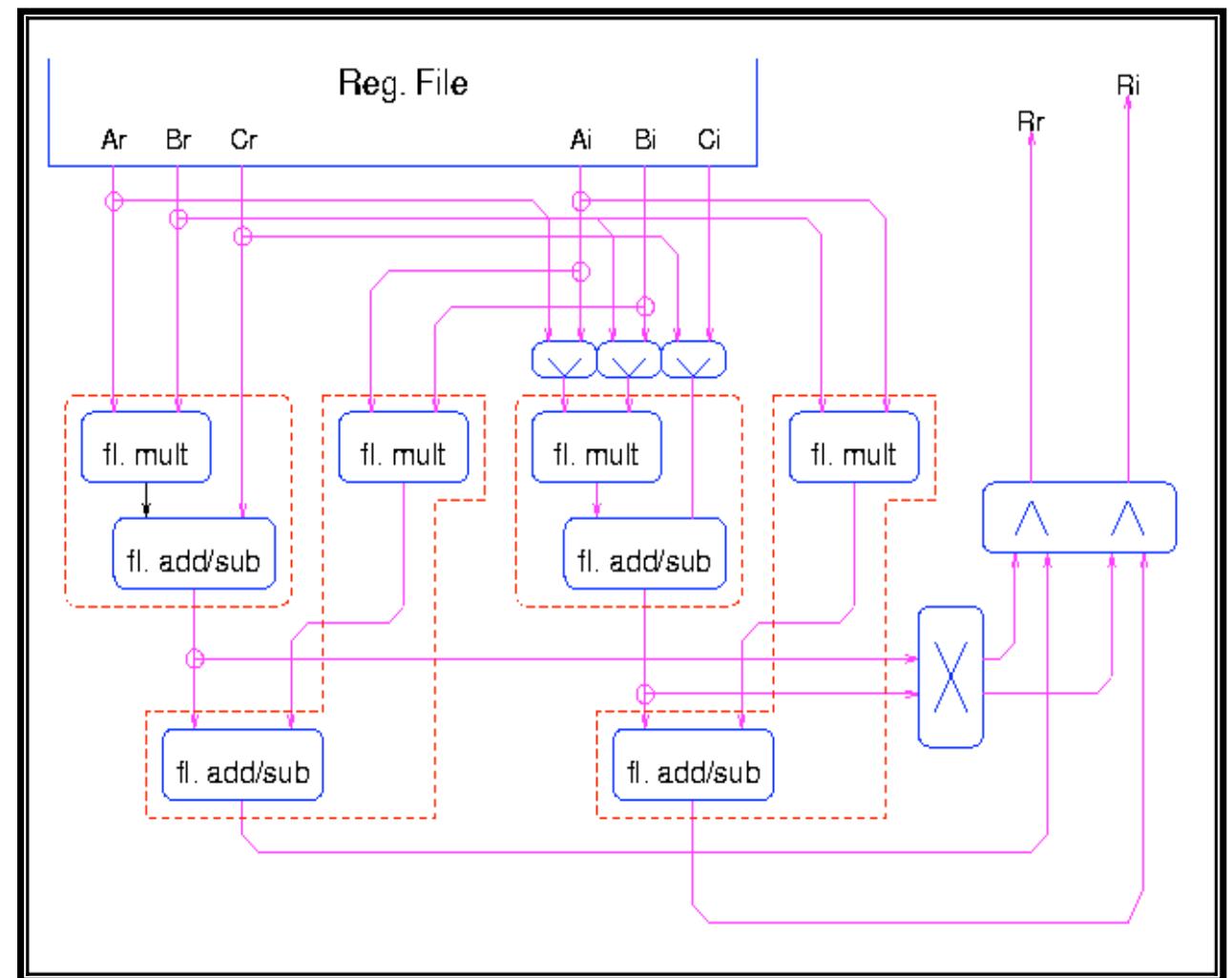
- no *glue logic*
- Reduced time for project, simulation and test of the prototype



# J&T

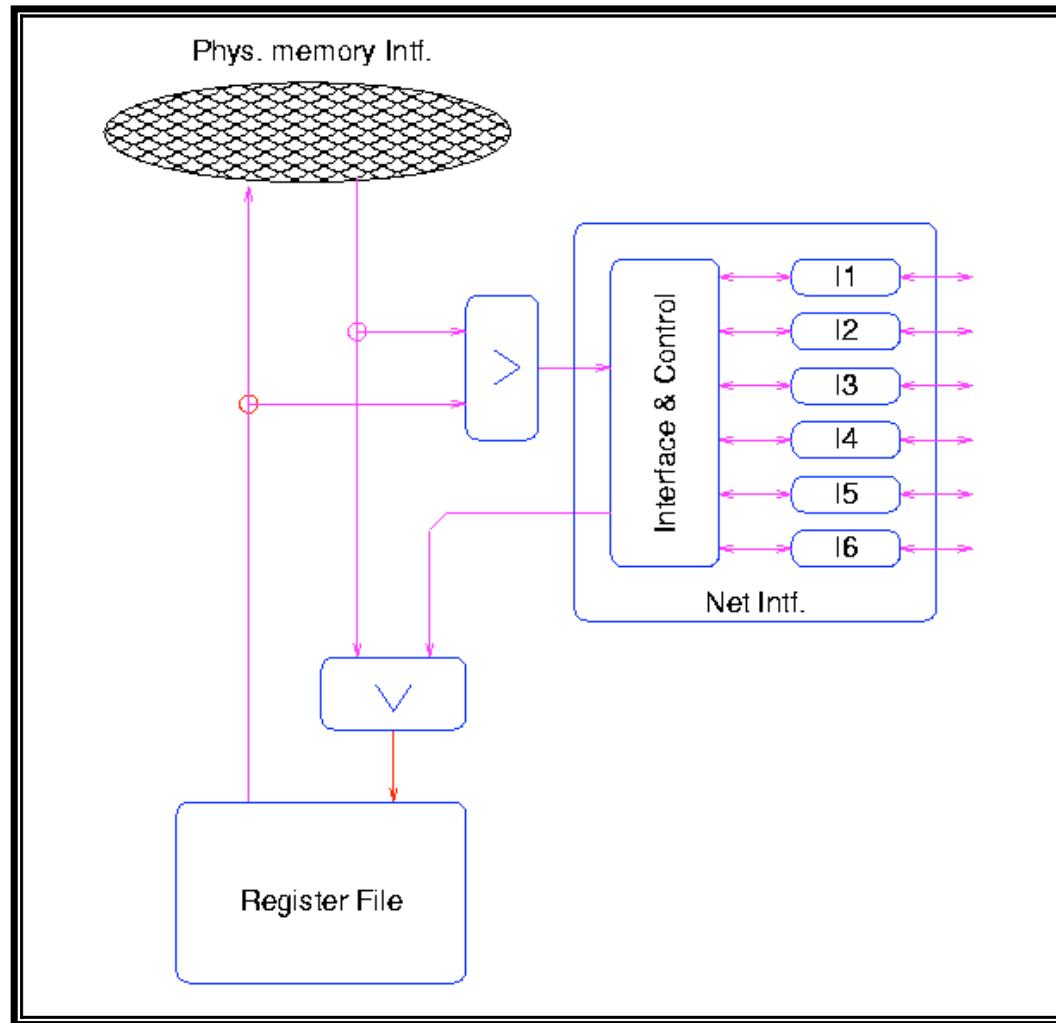
## the Arithmetic box

- Pipelined “normal”  $a \cdot b + c$   
(8 flops) per cycle



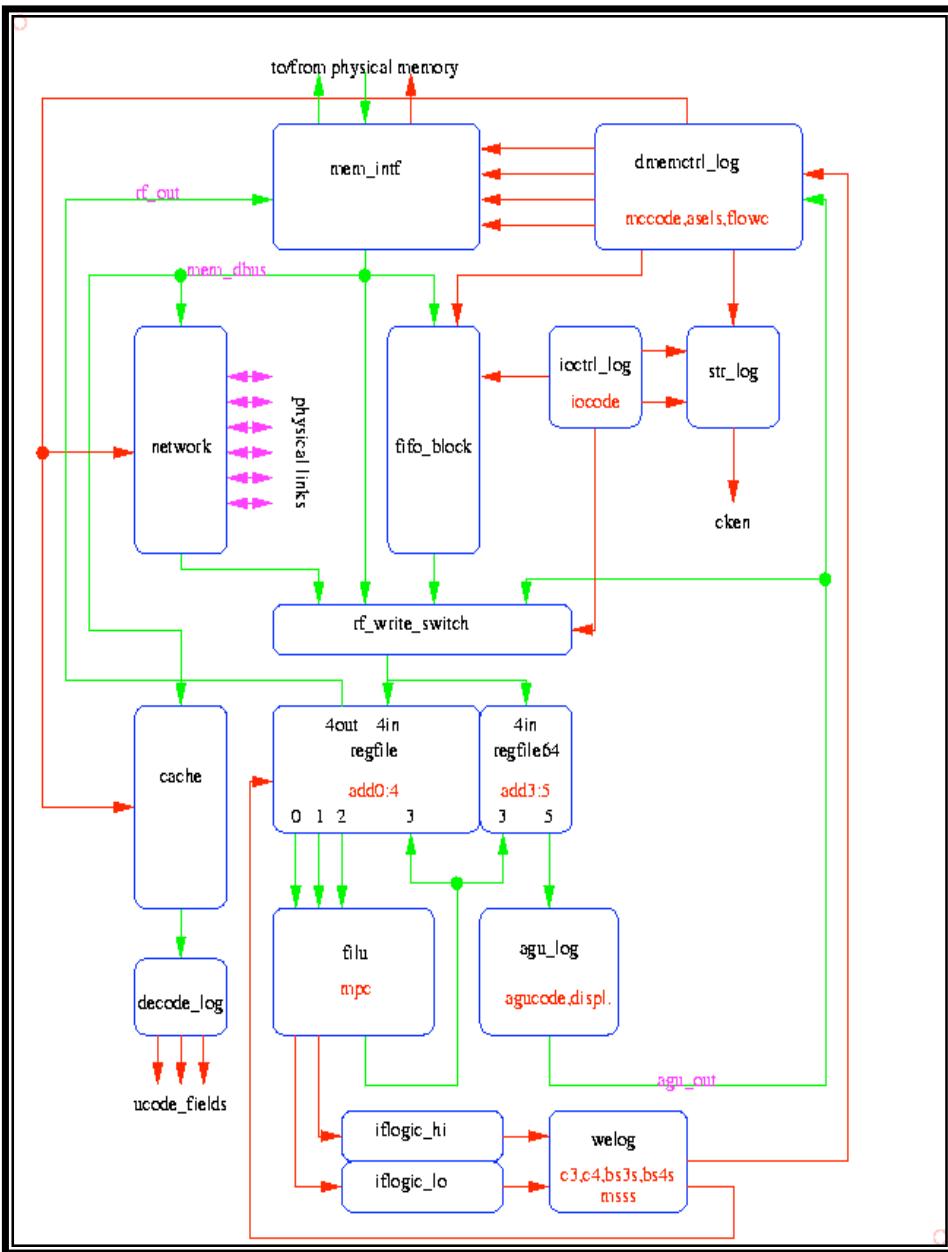
# J&T

## Remote I/O

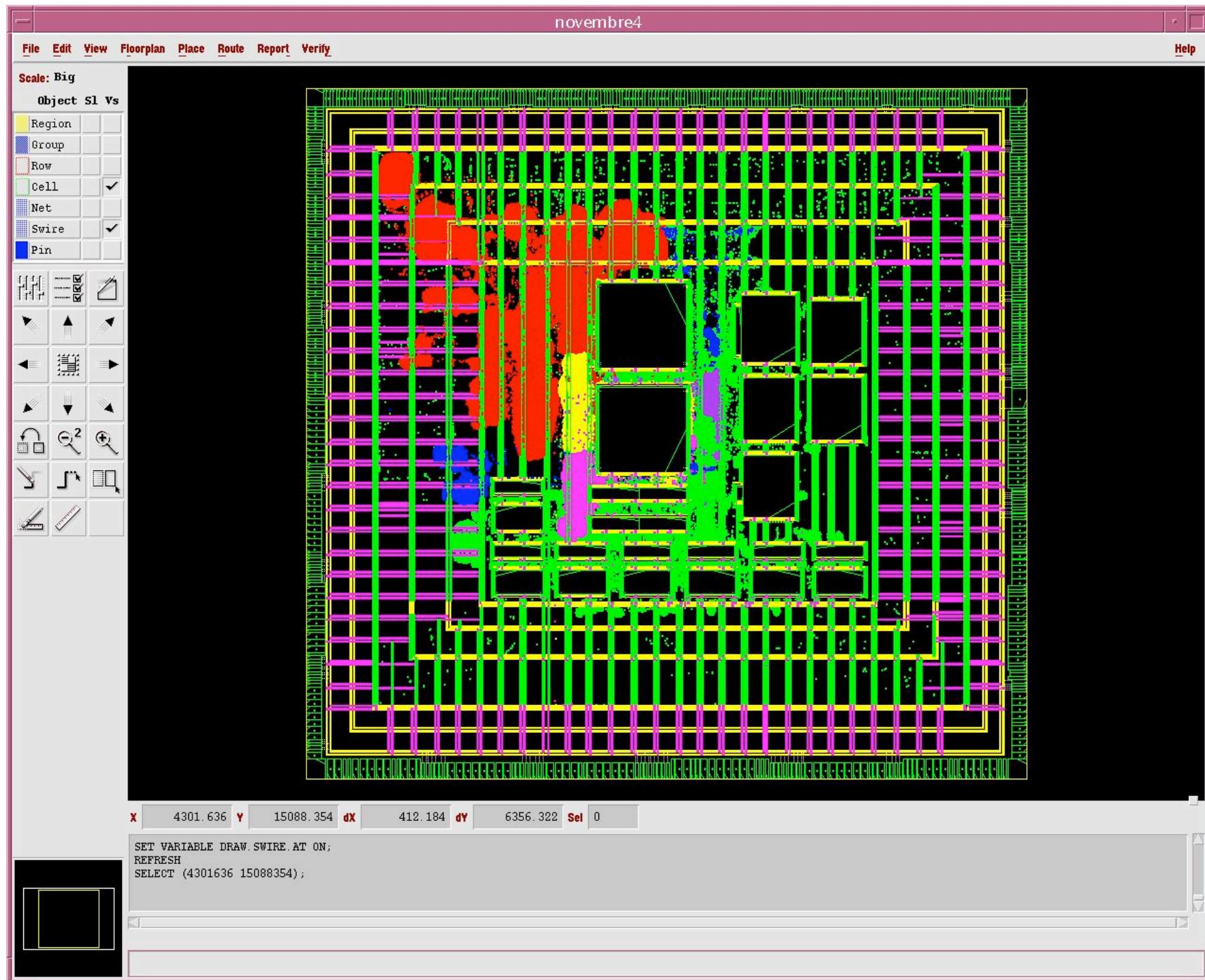


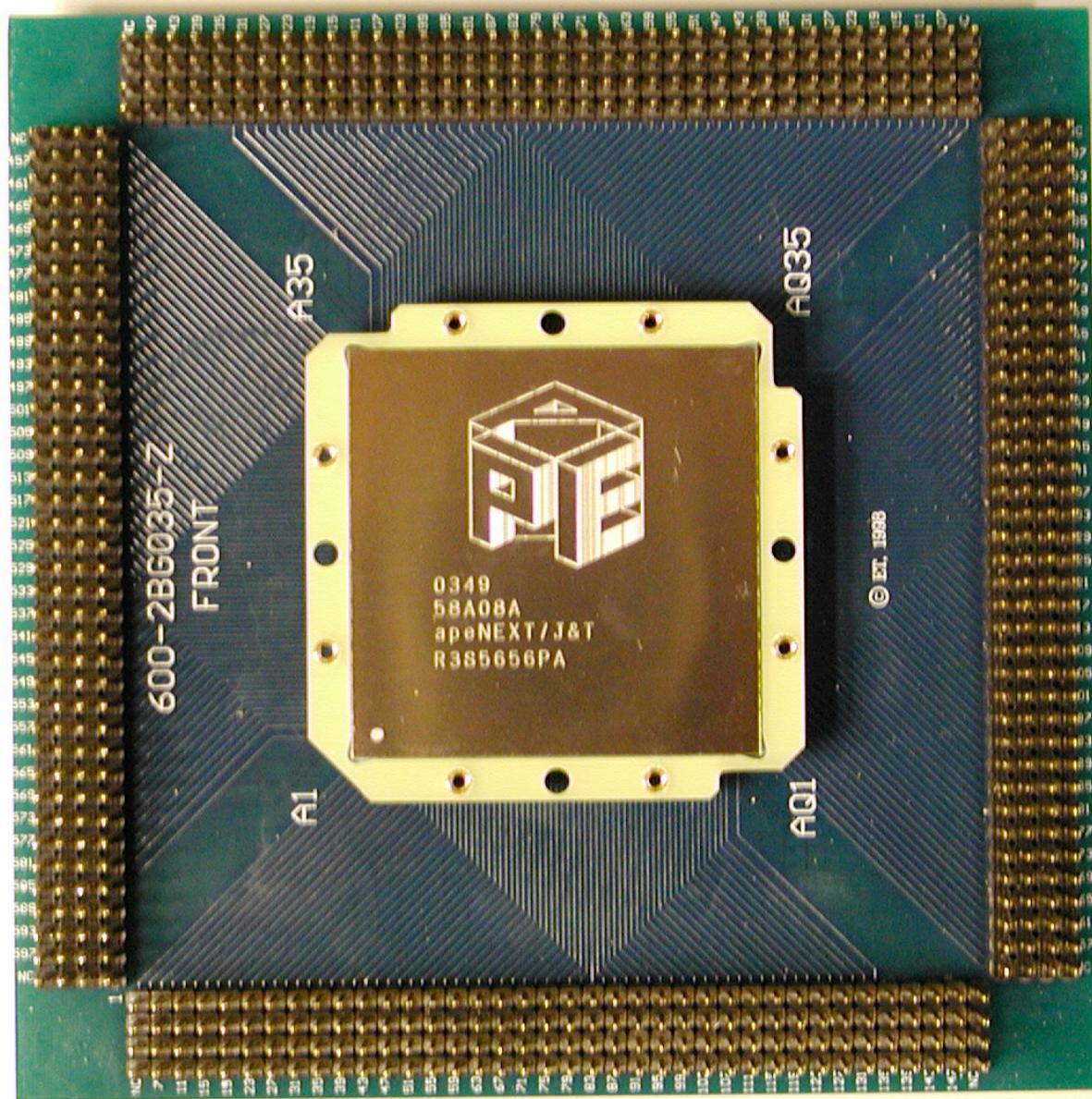
- fifo-based communication:
  - LVDS
  - 1.6 Gb/s per link  
(8 bit @ 200MHz)
  - 6 (+1) independent bi-dir links

# J&T

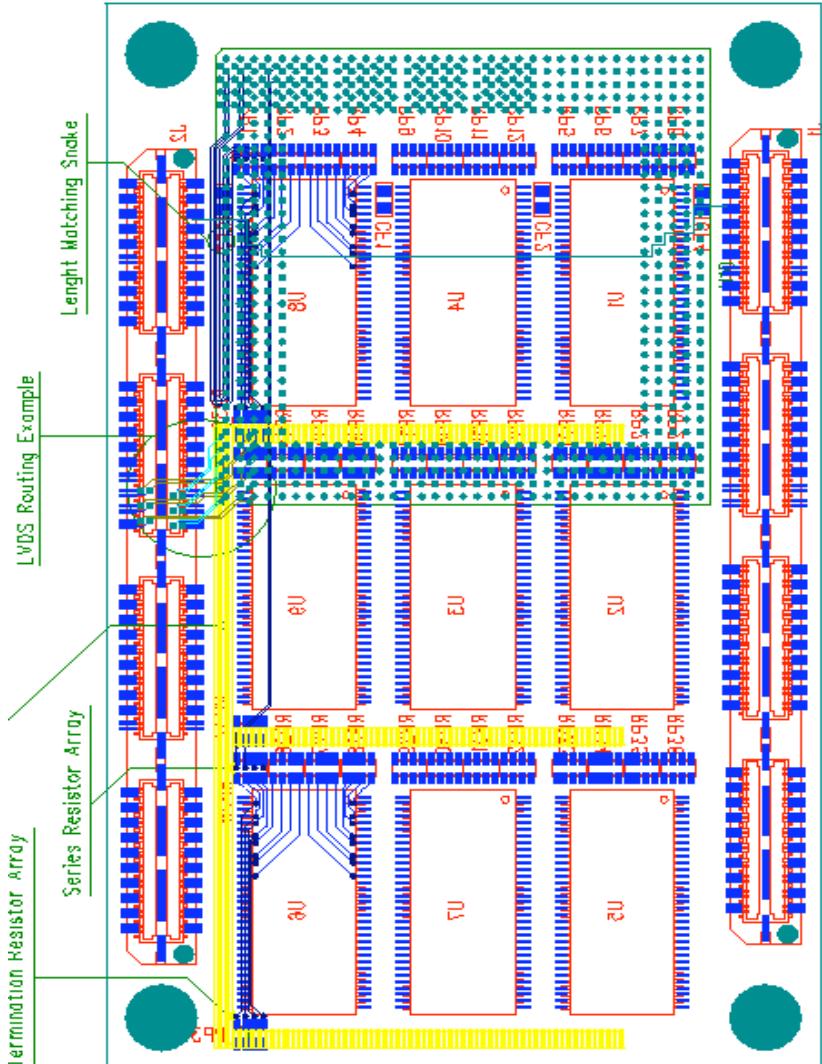


- CMOS 0.18 $\mu$ , 7 metal (ATMEL)
- 200 MHz
- Double Precision Complex Normal Operation
- 64 bit AGU
- 8 KW program cache (user-controllable)
- 128 bit local memory channel
- 6+1 LVDS 200 MB/s links
- BGA package, 600 pins

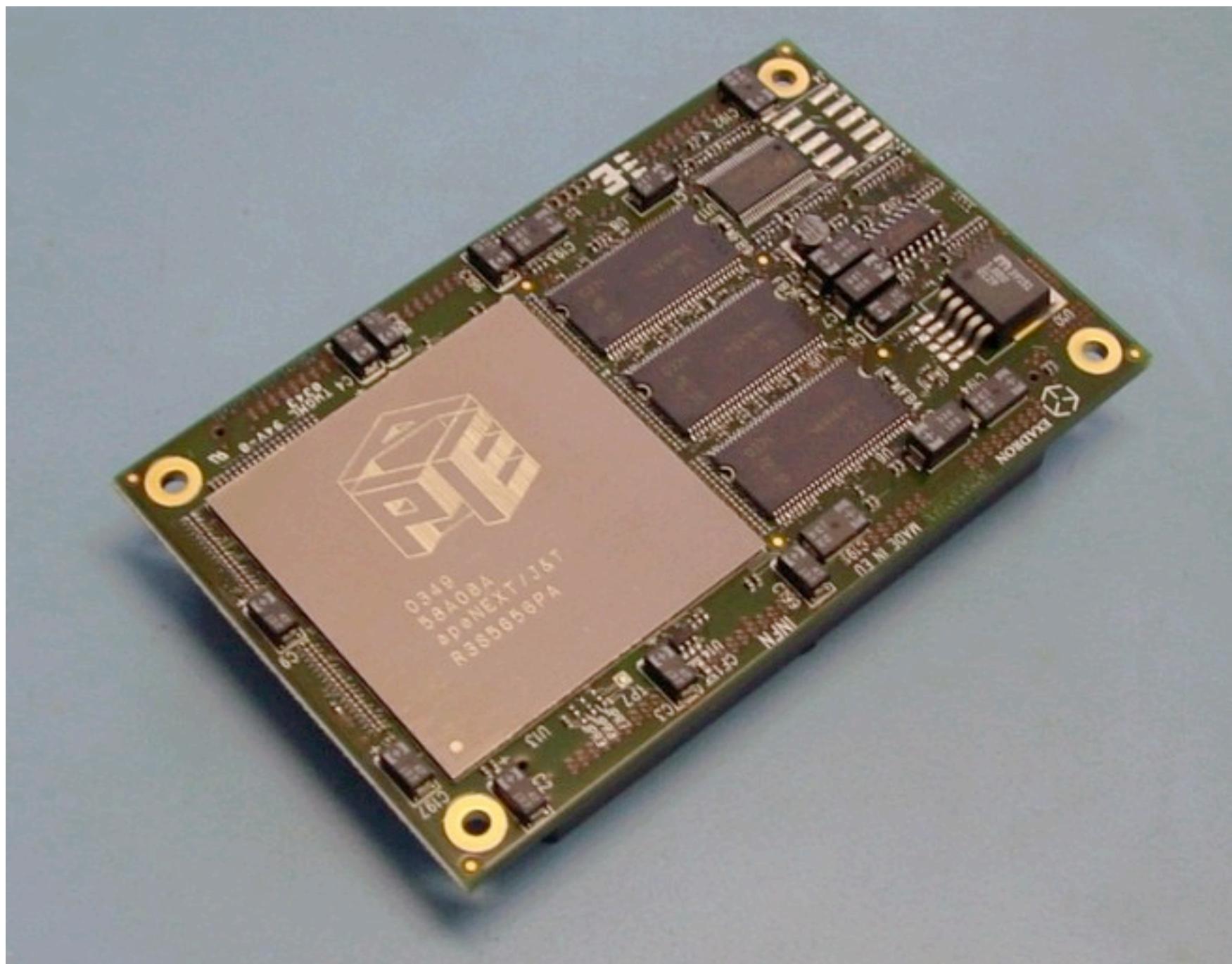




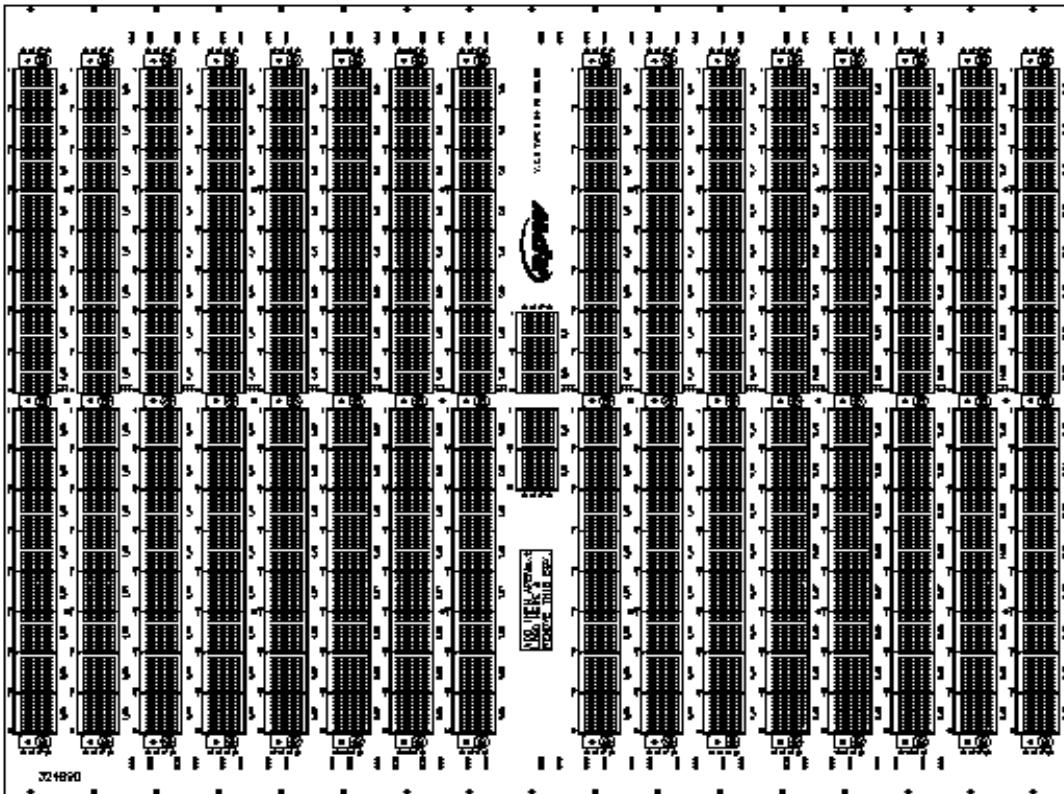
# J&T Module



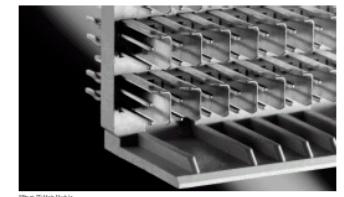
- J&T
- 9 DDR-SDRAM, 256Mbit (x16) memory chips
- 6 Link LVDS up to 400MB/s
- Host Fast I/O Link (7th Link)
- I2C Link (slow control network)
- Dual Power 2.5V + 1.8V, 7-10W estimated
- Dominant technologies:
  - SSTL-II (memory interface)
  - LVDS (network interface + I/O)



# NEXT BackPlane



- 16 PB Slots + Root Slot
- Size **447x600 mm<sup>2</sup>**
- **4600** LVDS differential signals, point-to-point up to **600 Mb/s**
- **16** controlled-imp. layers (32 Tot)
- Press-fit only
- Erni/Tyco connectors
- **ERMET-ZD**
- Providers:  
**APW** (primary)  
**ERNI** (2nd source)

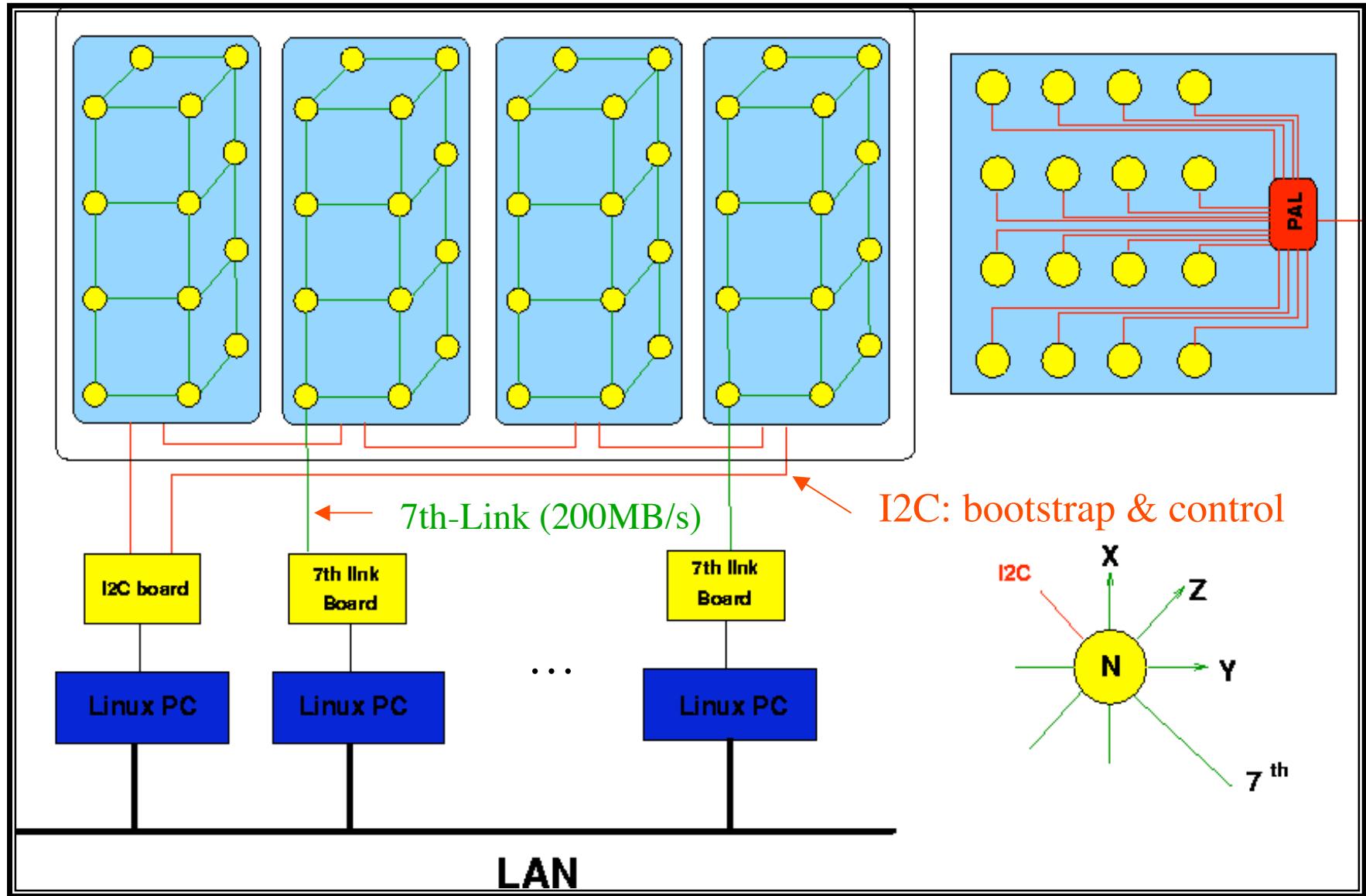


Activity	Status	Who	Cost	Note
BP development	Done	APW(ERNI)	32 KEuro	
BP prototypes (3)	Done	APW	41 KEuro	

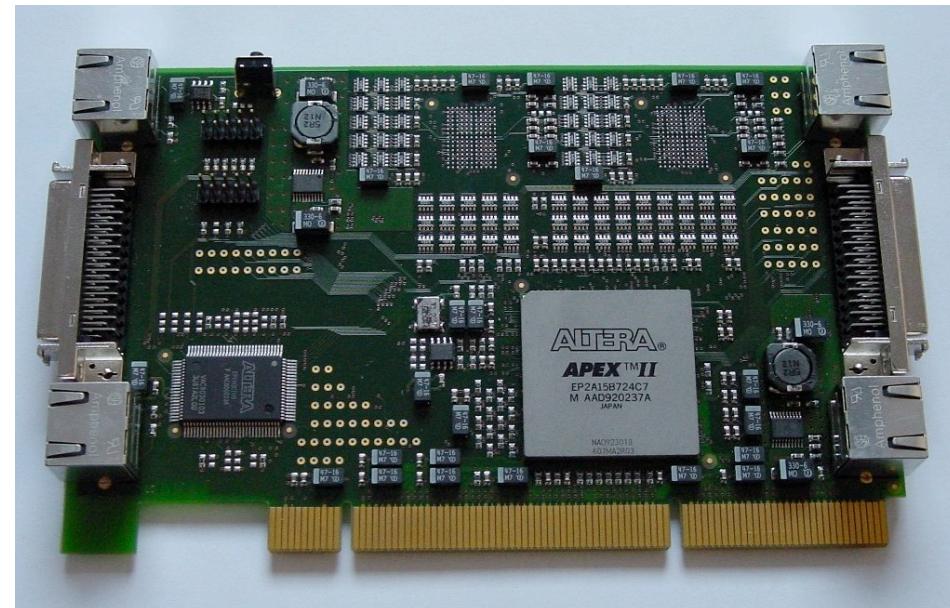
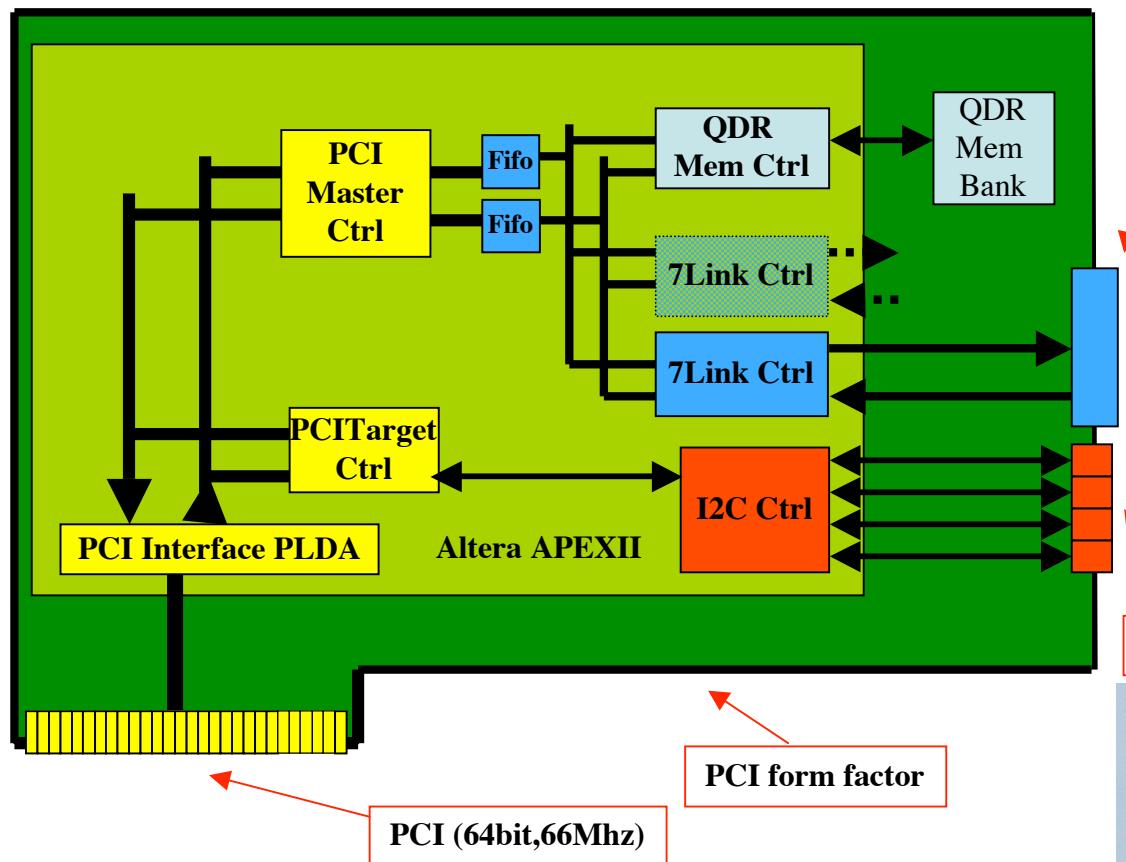
connector kit cost: **7KEuro** (!)

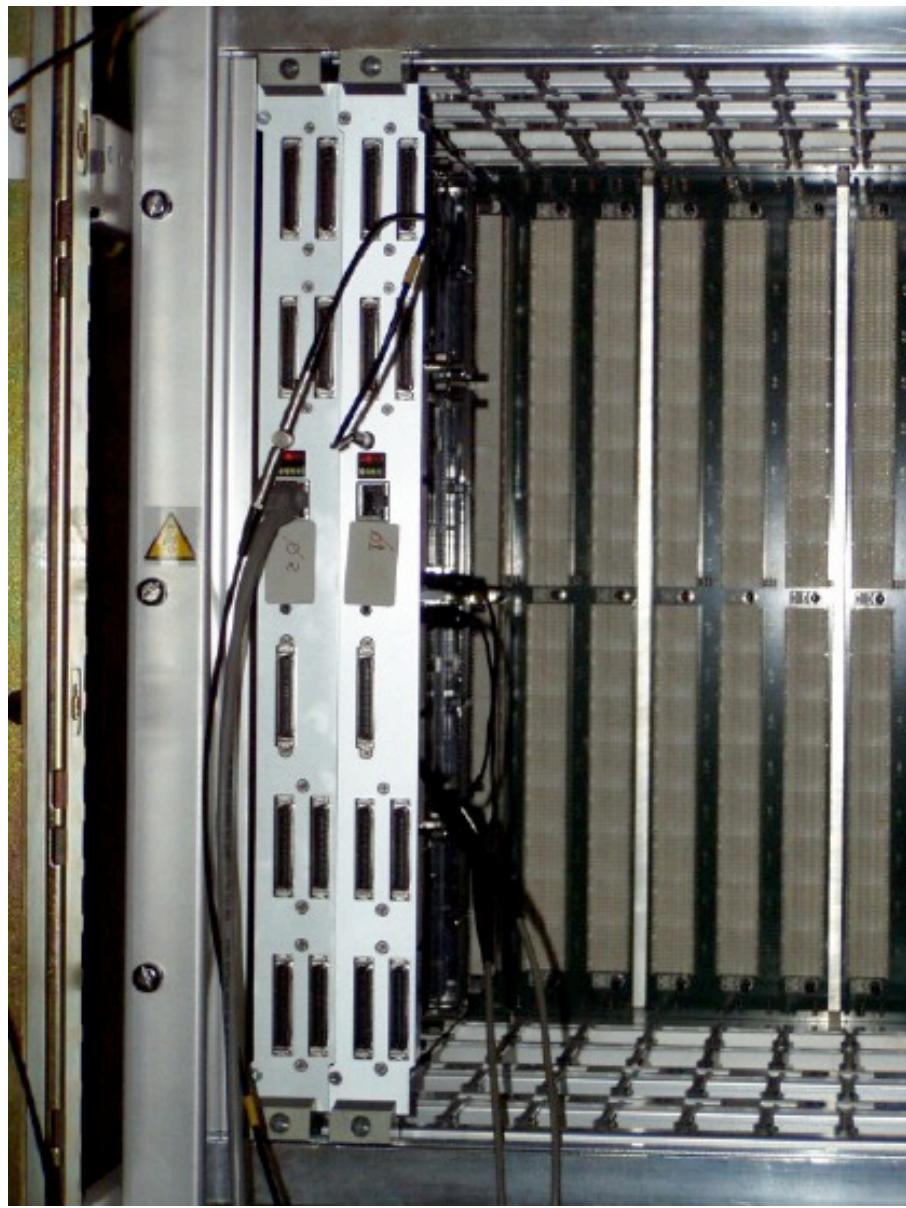
PB Insertion force: **80-150 Kg** (!)

# Host I/O interface



# Host I/O Interface

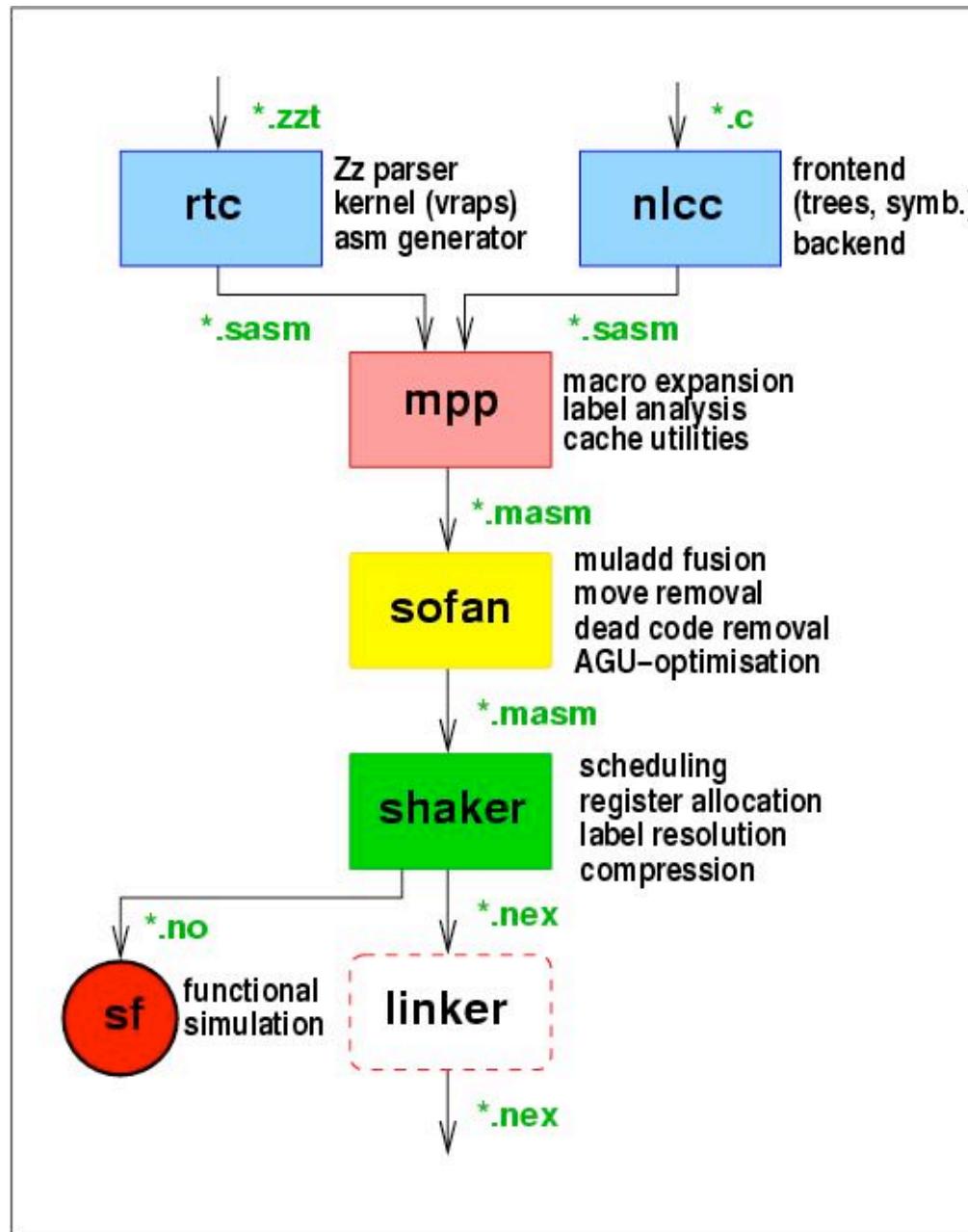




# Programming Languages

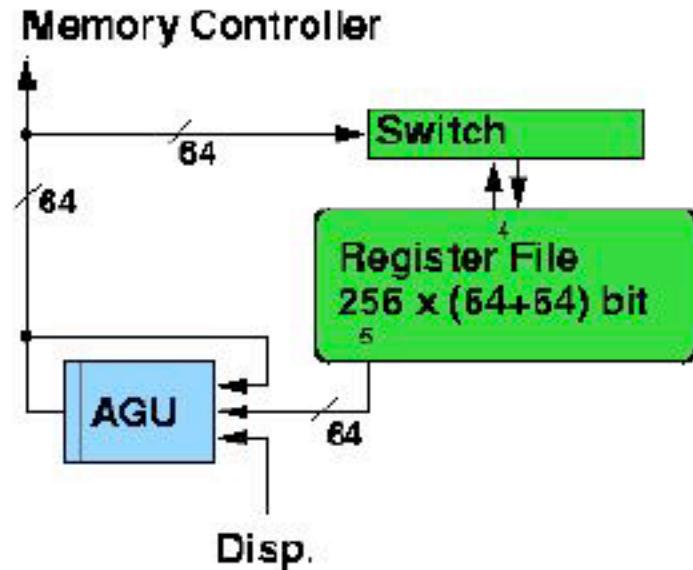
- Tao (was Apese)
  - Fortran-like, very simple to learn
  - Dynamical grammar, OO-style programming, QCDlib
  - Many tens thousand lines of codes existing all over Europe
  - All APEmille code compiles with no changes
- C
  - Based on lcc
  - Language extensions (complex vector, ~, where (), all() ...)
- SASM
  - High level assembly (should never be needed!!)

# Software Overview



## Assembler Optimizer: Sofan

- ❑ Optimization operating on [low-level assembly](#)
- ❑ Based on optimization toolkit [SALTO](#) (IRISA, Rennes)
- ❑ Optimization steps:
  - merging APE-normal operations
  - removing dead code
  - eliminating register moves
  - optimizing address generation:
  - instruction pre-scheduling
  - ...



## Benchmarks: Linear Algebra

operation	IO-Op	Flop	sustained performance “maximum”	measured
vnorm	1	4	50%	37%
zdotc	2	8	50%	41%
zaxpy	3	8	33%	29%
$U V$	27	202	92%	65%

“maximum” sustained performance ← ignoring latency of floating point pipeline and loop overhead ■

### Optimization “tricks”:

- loop unrolling
- burst memory access ■
- instructions kept in buffer ■

From Pleiter, Simma,...

### Performance limitations:

- start-up latency
- loop overhead

## Benchmarks: Results from C

operation	assembler	C	C + Sofan
vnorm	37%	31%	34%
zdotc	41%	28%	40%

→ Assembler programming not required

## Benchmarks: Wilson-Dirac Operator

$$\Psi_x = D_{xy}[U] \Phi_y$$

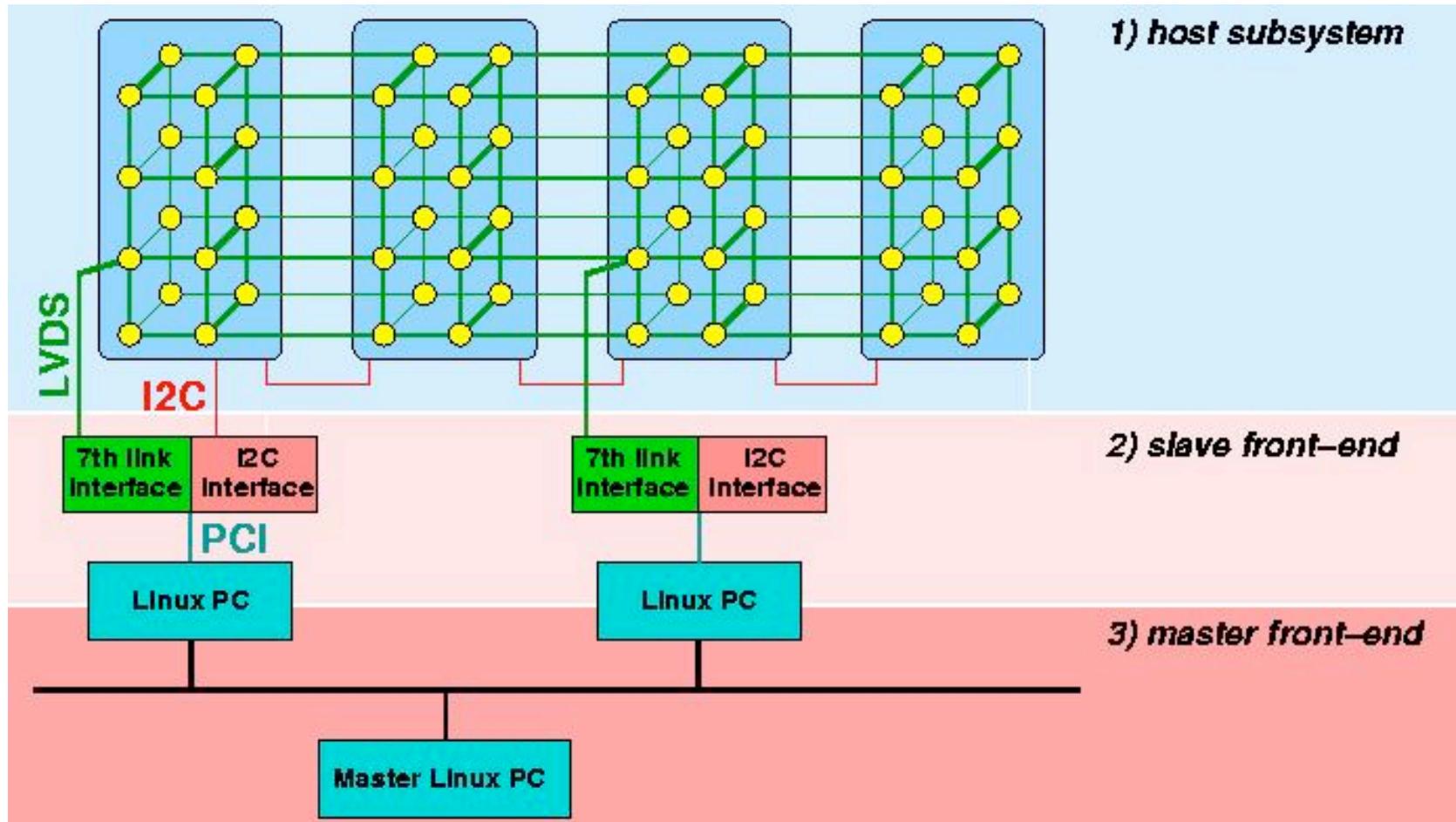
Consider worst case: local lattice size  $16 \times 2^3$

**Measured sustained performance: 55%**  
**Measured number of stretch cycles: 4%**

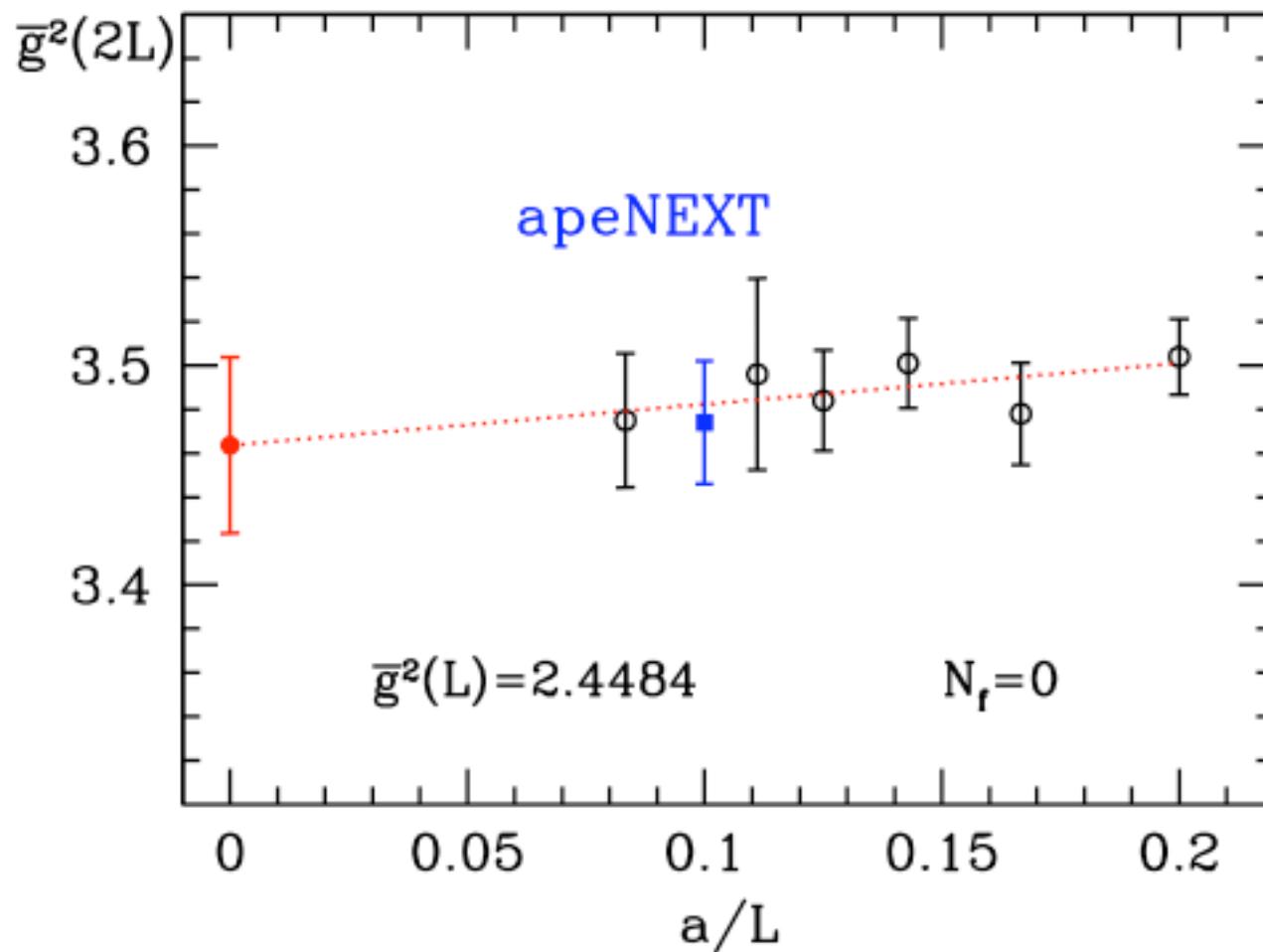
Optimization “tricks”:

- keep gluon fields local
- pre-fetching 2 sites ahead
- orthogonal communication directions
- some unrolling

# Operating system



Step Scaling Function for the running coupling constant in SU(3), 16 node apeNEXT  
Non ape data from S.Capitani et al. Nucl.Phys. B544 (1999) 669



# Costs

- 1700 KEuro developments  
550 KEuro + 1050 KEuro  
    Non VLSI      VLSI
- NO SALARIES
- Prototype production cost  $\sim$  0.6-0.7 Euro/Mflops  
Large scale as low as  $\sim$  0.5, see next

- Like APEmille, apeNEXT will be commercially available.
- Slow EU procedure for official tender (start 03/04, end ~ 08/04) to choose the company
- Committee (Vicini, Simma, FR, INFN administratives) at work
- Machines by Nov-Dec 2004 at a rate of 2x512-node/Month

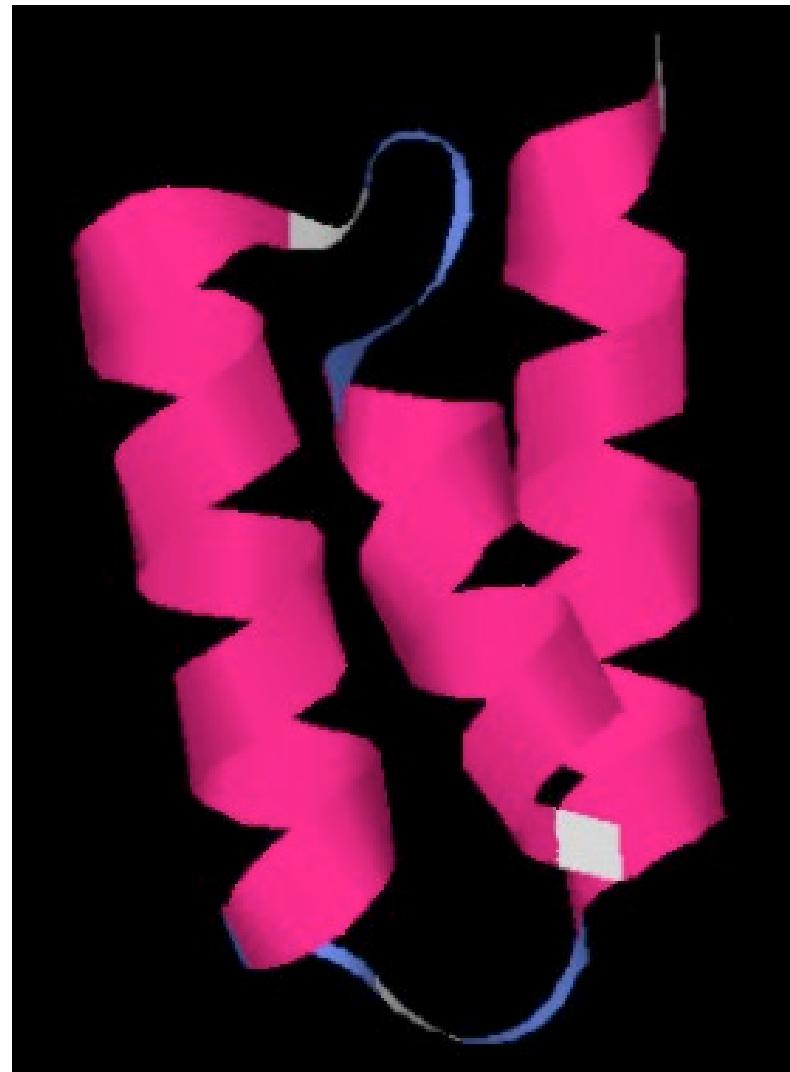
Bielefeld	130 GFlops	(2 crates)
Zeuthen	520 GFlops	(8 crates)
Milan	130 GFlops	(2 crates)
Bari	65 GFlops	(1 Crates)
Trento	65 GFlops	(1 Crates)
Pisa	325 GFlops	(5 Crates)
Rome 1	520 GFlops	(8 Crates)
Rome 2	130 GFlops	(2 Crates)
Orsay	16 GFlops	(1/4 crates)
Swansea	65 GFlops	(1 crates)

APEmille in Europe

- INFN has already funded apeNEXT per un totale di circa 10 Tflops in Italy to be installed at “la Sapienza”. More funds may come
- Germany and France are still contracting with their funding agencies

# Plans

- Physics
  - LQCD of course (so many groups), see Lattice 2005
  - Turbulence (Fe)
  - Complex System (Rm)
- apeNEXT<sup>2</sup>
  - Activity will continue
  - Intermediate 2-4 x machine?
  - 100TF project???
- QBIO
  - Protein (mis)folding
  - Drug docking
  - See QBIO archive @ LANL



## Structure Explorer - 1BDD



<i>Title</i>	Staphylococcus Aureus Protein A, Immunoglobulin-Binding B Domain, NMR, Minimized Average Structure
<i>Classification</i>	Immunoglobulin-Binding Protein
<i>Compound</i>	Mol_Id: 1; Molecule: Staphylococcus Aureus Protein A; Chain: Null; Fragment: B Domain; Engineered: Yes
<i>Exp. Method</i>	NMR, Minimized Average Structure