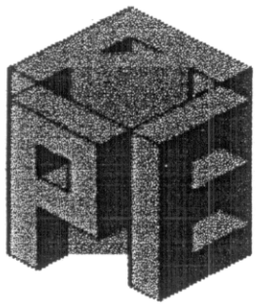


# Status of APE

**R. Tripiccione**

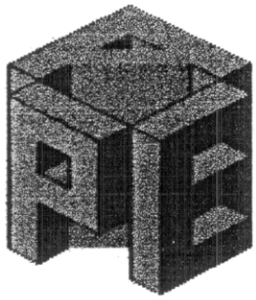
University and INFN, Ferrara

Talk at Lattice2001, Berlin, August 20th 2001

---

**Outline of the talk:**

- Where we start from (a short update of APEmille).
- Where we want to go (an introduction to apeNEXT).
- Status of apeNEXT.

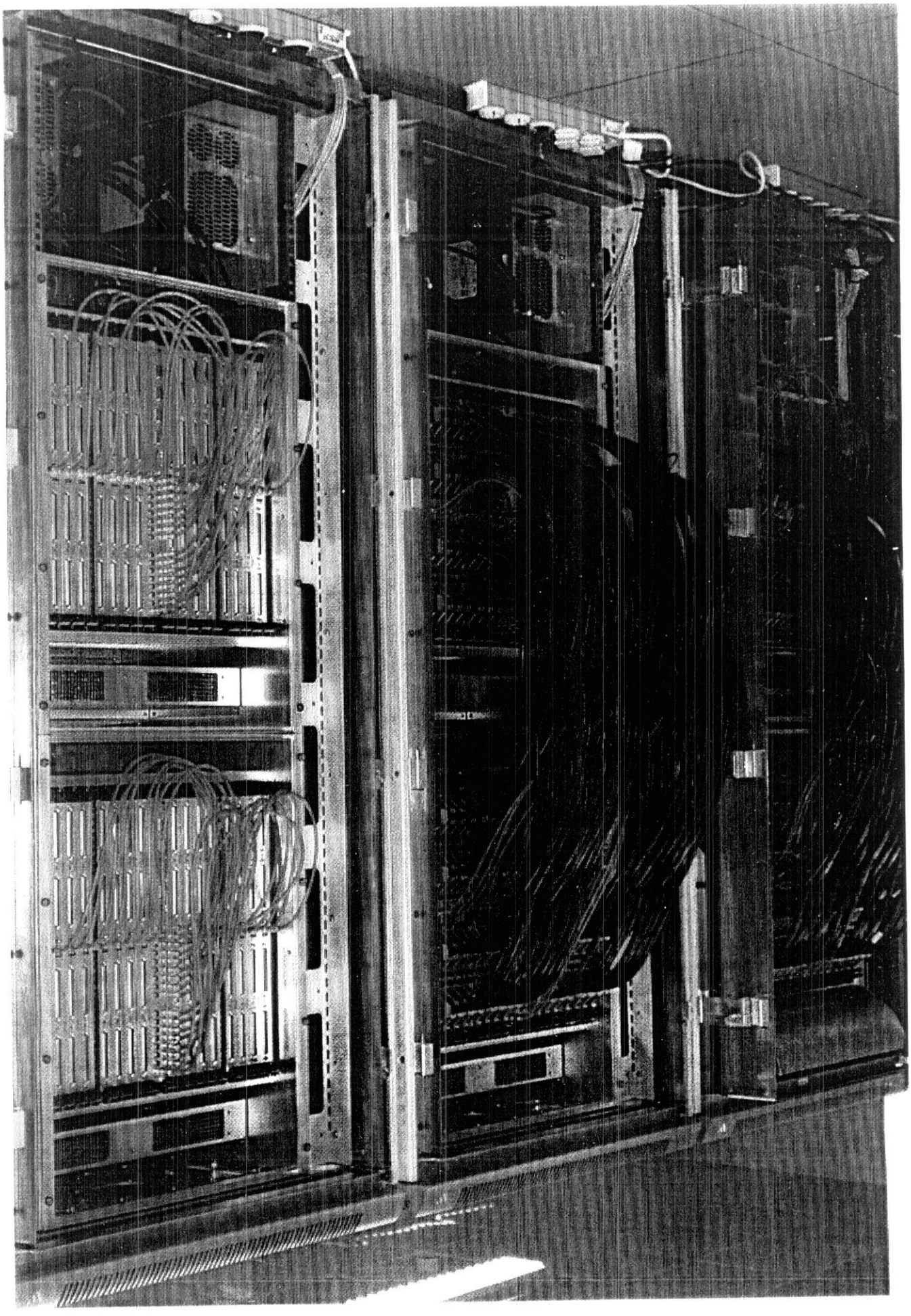


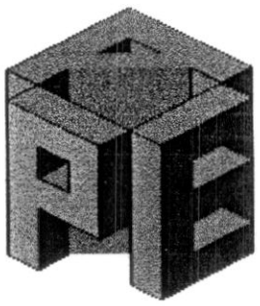
## APEmille (I)

APEmille has been commissioned at several sites, and provides a remarkably high overall number crunching performance.

| Site | Peak Gflops (now) | Peak Gflops (end 2001) |
|---|---|---|
| Rome I | 455 | 650 |
| DESY | 455 | 550 |
| Pisa | 130 | 260 |
| Rome II | 130 | 260 |
| ~~Rome I~~ | ~~455~~ | ~~650~~ |
| **Bielefeld** | 80 | 140 |
| Milano | 65 | 130 |
| Bari | 65 | 65 |
| **Swansea** | 65 | 65 |
| **Orsay** | 16 | 16 |
| Grand Total | 1450 | 2140 |

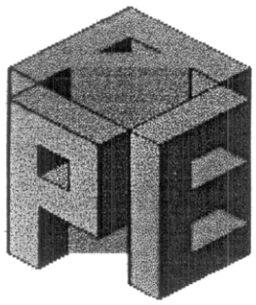
- good old TAO programming language
- good sustained performance (50 % of peak in real programs)
- significant bandwidth to disk (20 - 200 Mbytes/sec).
- hosted by small (20-30 units) clusters of Linux-based PC's





## apeNEXT: Basic ideas

- The architecture invented by this community about 15 years ago is still a very good choice.
- New ideas are being discussed. Still, it makes sense to stick to the basic APE architecture and boost its performance up to the levels allowed by current technology.
- Try to meet the requirements listed (for instance) in the ECFA report.
  - Dynamic fermions
    - L = 2 ... 4 fm
    - a = 0.1 ... 0.05 fm (lattice: 48^3 x 96)
    - M = 0.35 M
  - Quenched simulations on very large lattices
    - L = 1.5 ... 2.0 fm
    - a = 0.1 ... 0.02 fm
    - b-physics with little (??) extrapolation in the quark mass.

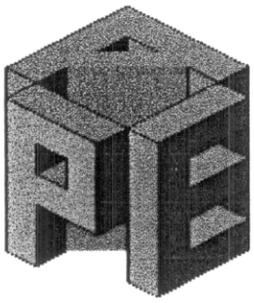


## APEmille (II)

Valuable physics is being churned out of these machines.

About 15 papers at this conference, that rely on APE-produced data.

- D. Becirevic
- M. D'Elia
- R. Frezzotti
- C. Gebert
- B. Gehrmann
- O. Kaczmarek
- G. Martinelli
- M. Papinutto
- J. Rolf
- Ch. Schmidt
- R. Sommer
- J. Heitger
- I.Wetzorke
- U. Wolff



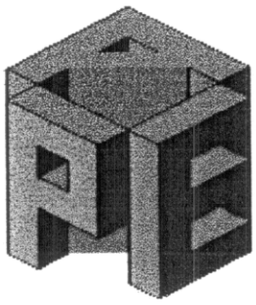
## apeNEXT: Basic ideas (II)

In computer terms, our requirements translate into:

- O(10 Tflops) peak computing performance.
- O(50%) sustained performance.
- The bulk of the processing power to be provided by a **small** number of **large** machines (3 - 5 TFlops each).
  - O(1 Tbyte) on-line memory for each system
  - ~ 1 Gbyte/sec bandwidth to disks.

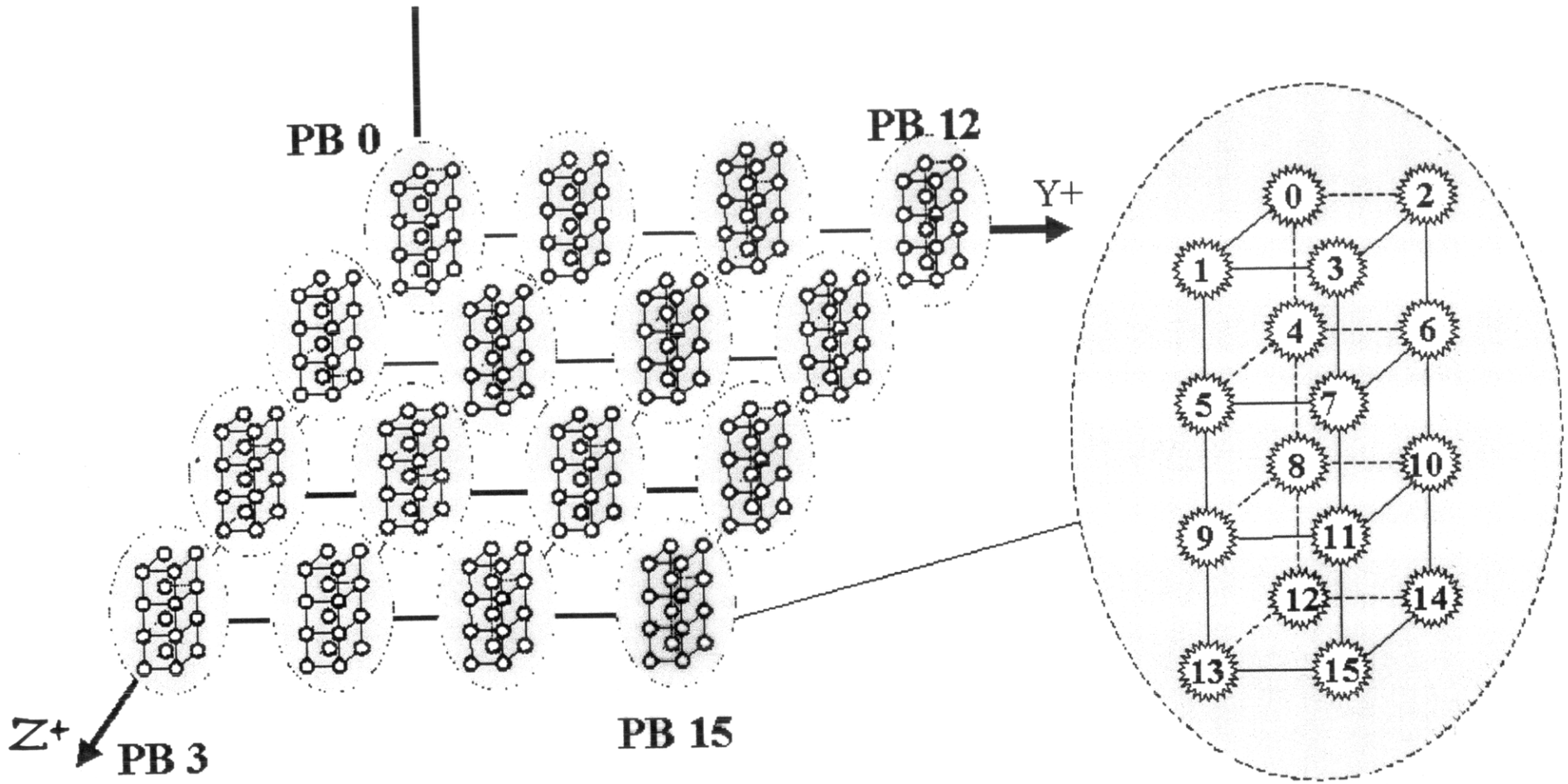
Our new project has:

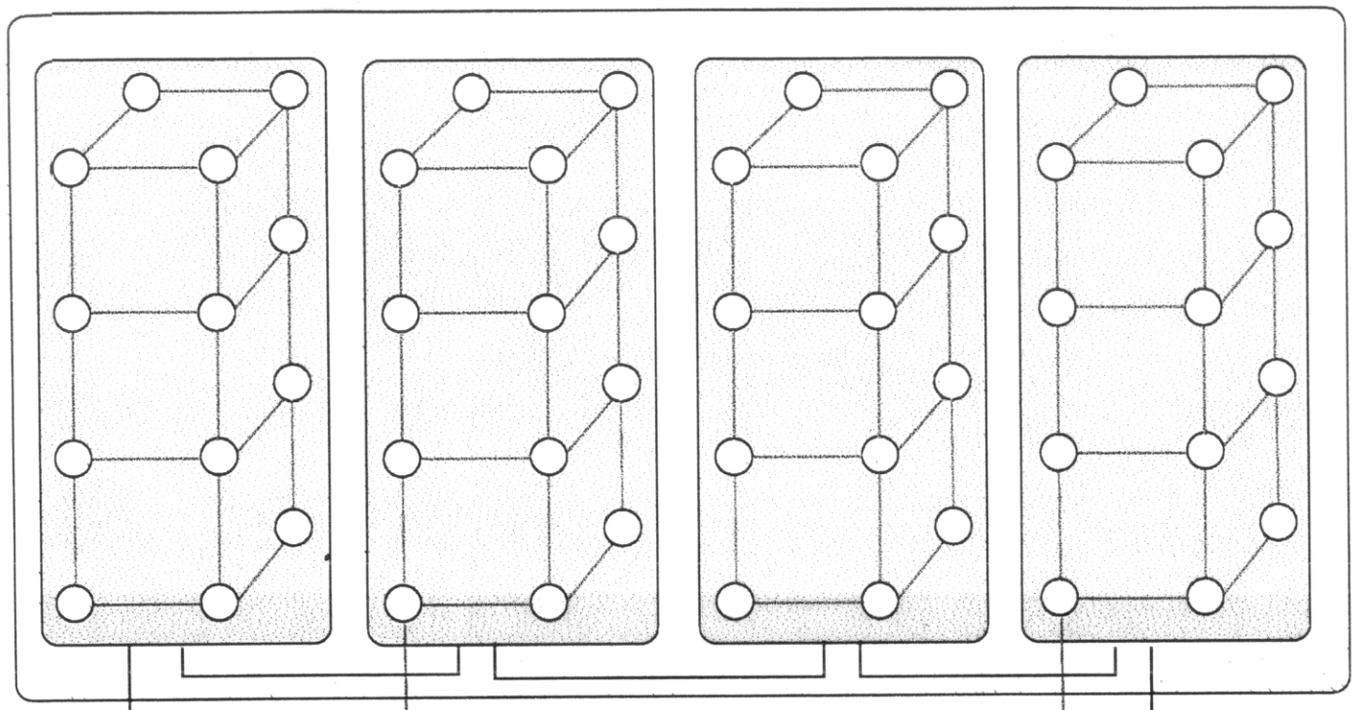
- striking similarities with a (scaled up) APEmille systems
- but also several new features.

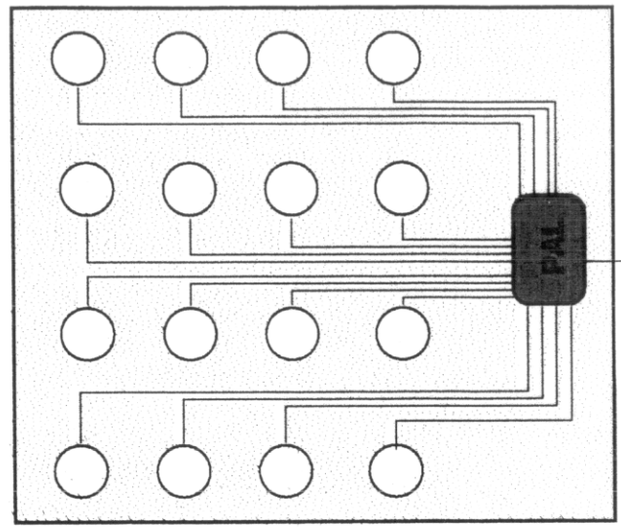


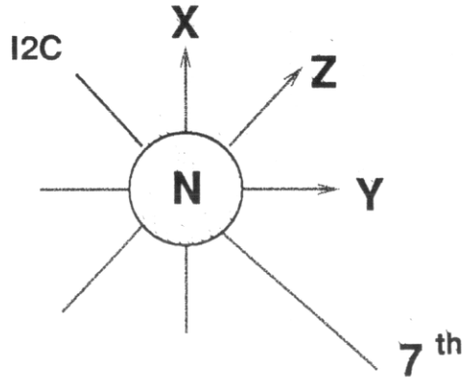
## apeNEXT: Good Old Features

- Three-dimensional array of processors with periodic boundaries.
- Data links between nodes optimized for nearest-neighbour communications.
- Fat arithmetic operators to achieve high performance at comparatively slow clock.
- Large register file as a replacement for data caches.
- Loosely coupled connection to a cluster of PC's for input/ouput.


PB 0
PB 12
Y+
Z+
PB 3
PB 15
0
2
1
3
4
6
5
7
8
10
9
11
12
14
13
15





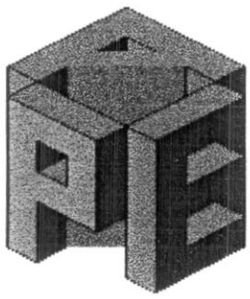




I2C board

7th link Board

7th link Board

LAN

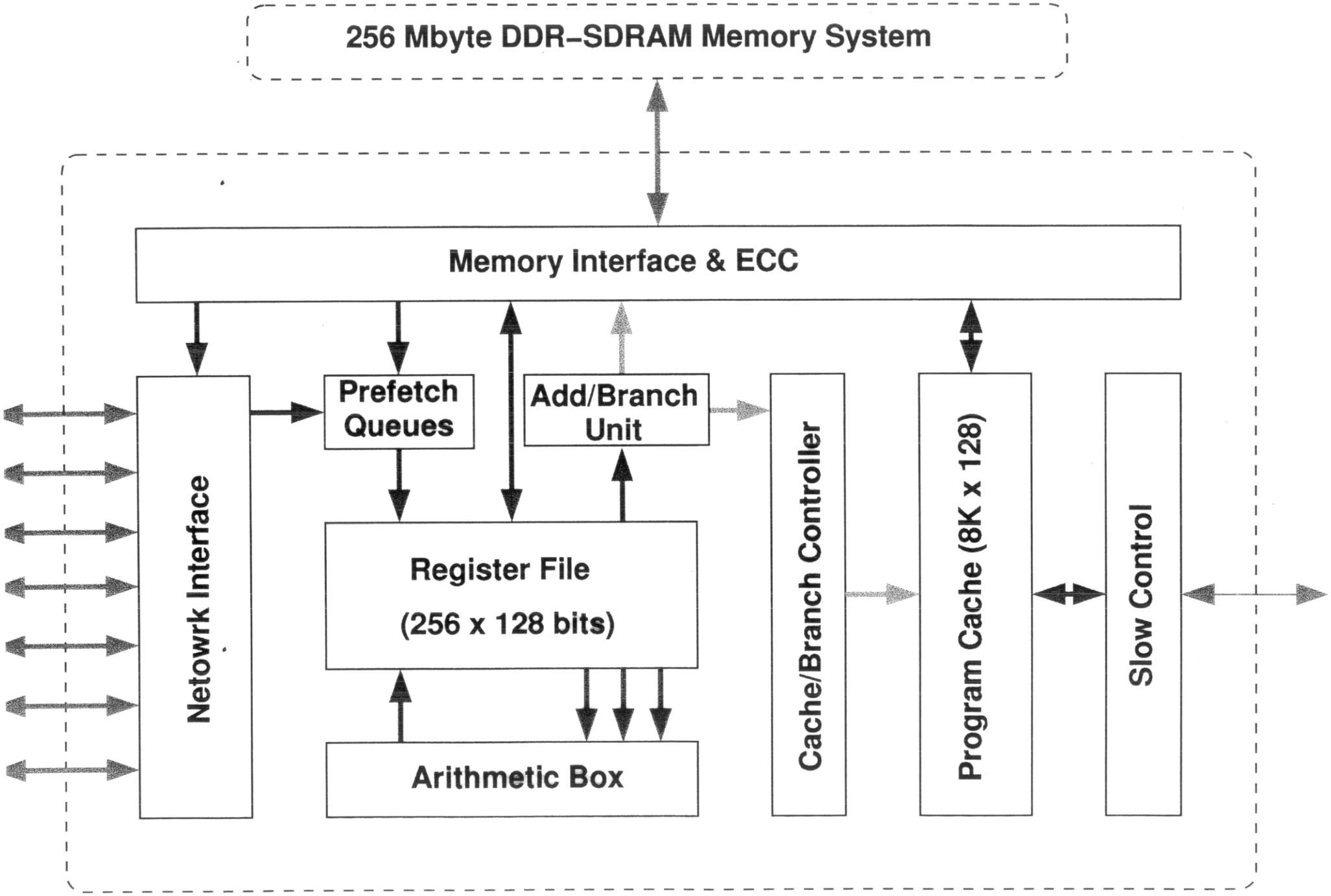


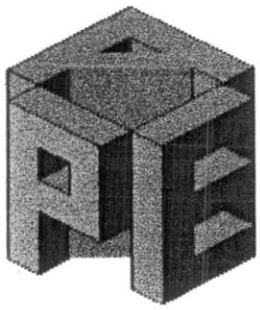
## apeNEXT: the J&T processor

**J&T** is the building block for apeNEXT. It is a real system on chip. It contains:

- An interface to DDR-SDRAM.
- A data prefetch-queue.
- A program cache.
- A Large multi-port register file.
- 8 Floating point operators (IEEE double precision everywhere).
- + Integer arithmetics + (stride 2) vector processing for real data.
- 6 + 1 fast communication channels (200 Mbyte/sec).
- 256 Mbytes memory per processor.
- 200 Mhz clock frequency --> 1.6 Gflops peak performance.

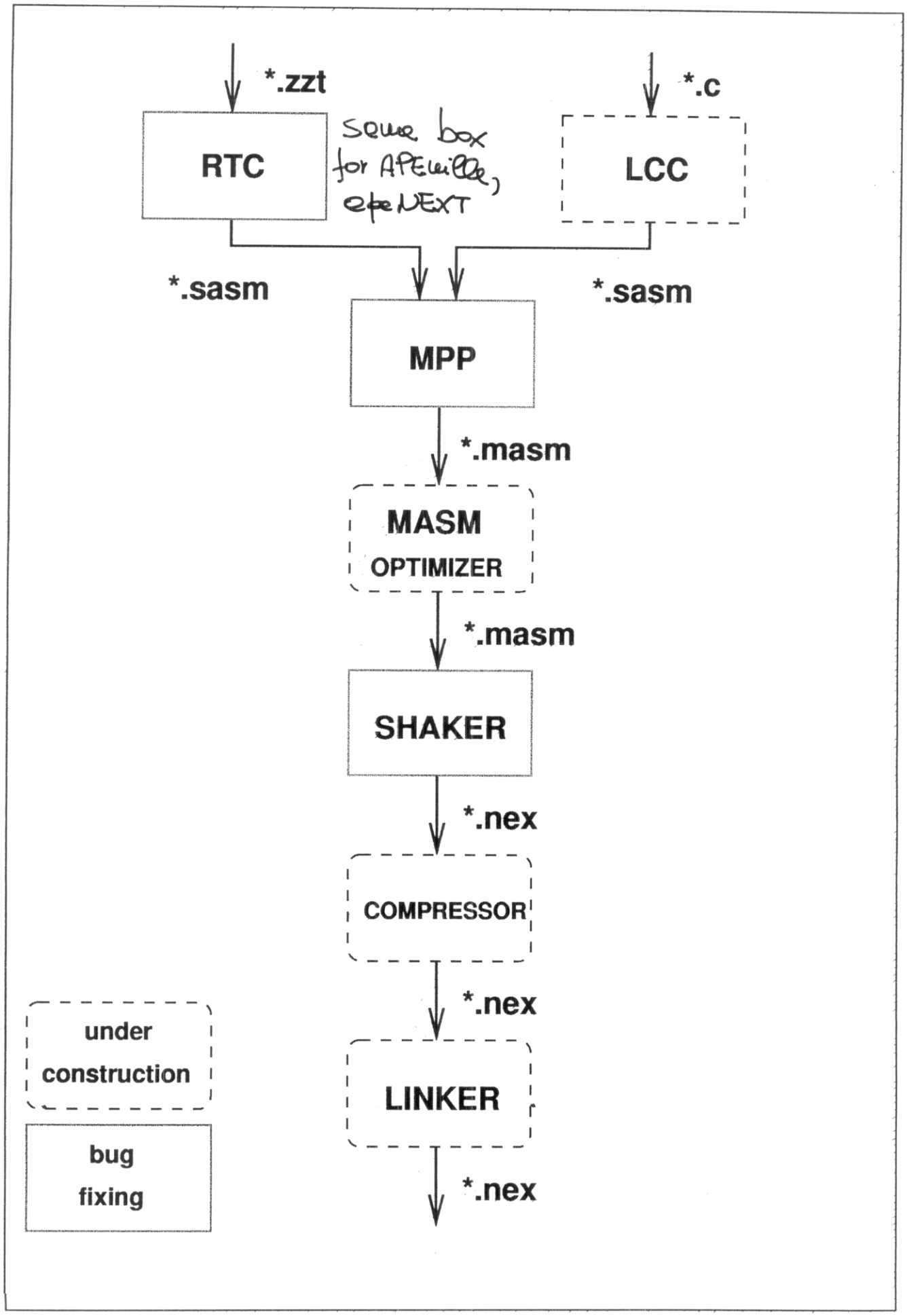
. Exactly matching needs with. systematic prefetching.

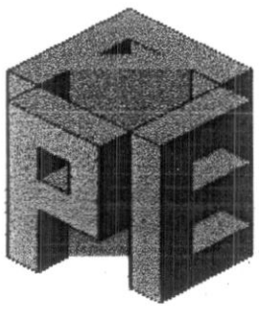

256 Mbyte DDR–SDRAM Memory System
Memory Interface & ECC
Netowrk Interface
Prefetch Queues
Add/Branch Unit
Register File
(256 x 128 bits)
Arithmetic Box
Cache/Branch Controller
Program Cache (8K x 128)
Slow Control



## apeNEXT: New architectural features

- Fully independent nodes: SPMD (as opposed to SIMD) programming.
- Program cache to reduce bandwidth needs.
- Program-driven prefetch queues to overlap computation with data load-store (local data).
- Register indexing.
- Concurrent and independent node and link operation, to hide remote comm. latencies and smear-out bandwidth requirements.
- TAO and C available on equal footing.


*.zzt
RTC
same box for APEmille, epeNEXT
*.c
LCC
*.sasm
*.sasm
MPP
*.masm
MASM
OPTIMIZER
*.masm
SHAKER
*.nex
COMPRESSOR
*.nex
LINKER
*.nex
under construction
bug fixing



## apeNEXT: the system

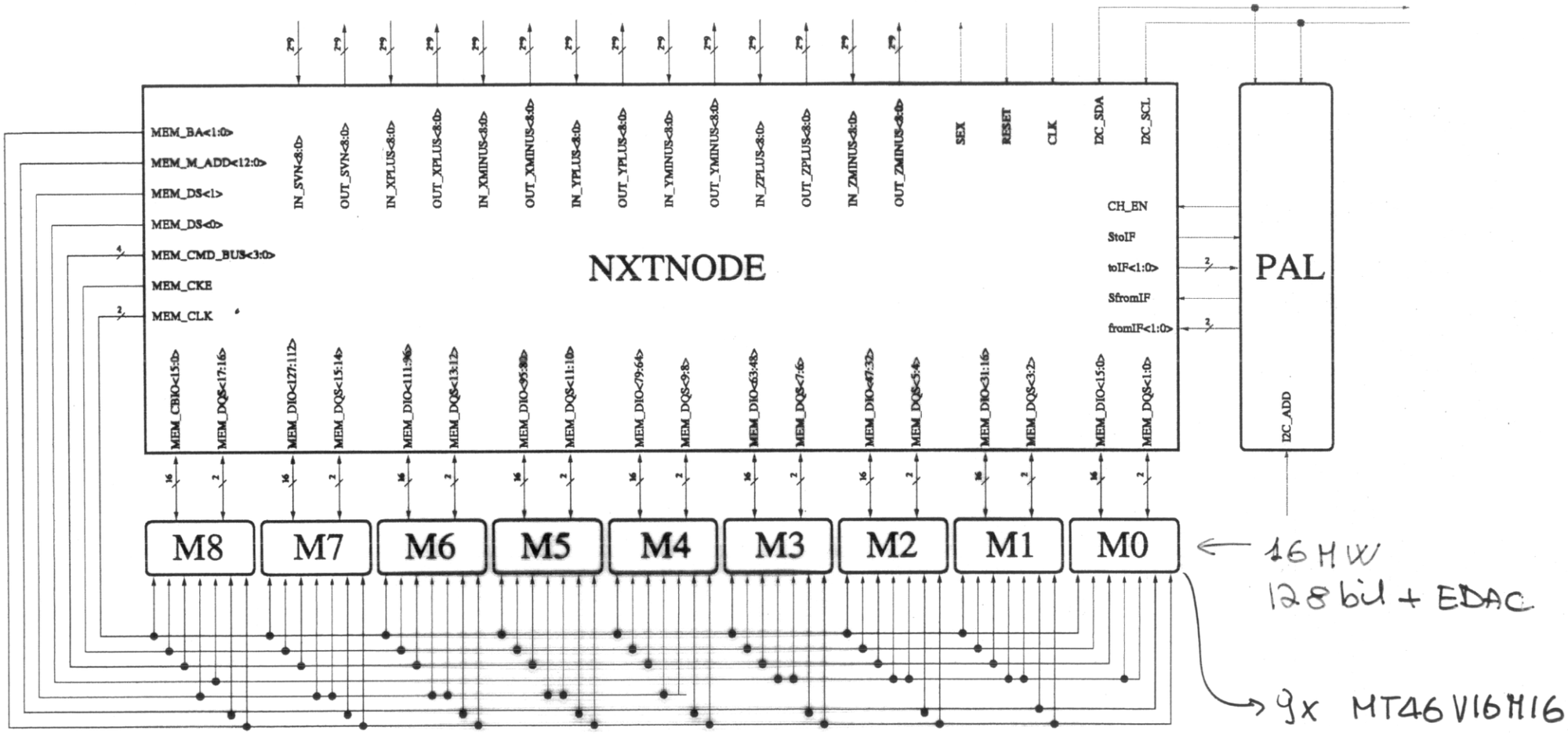
- Blocks of 16 processors are assembled onto a processing board (25 Gflops).

  ↗ processing node + memory.

- 16 processing boards are housed inside a system crate (400 Gflops).

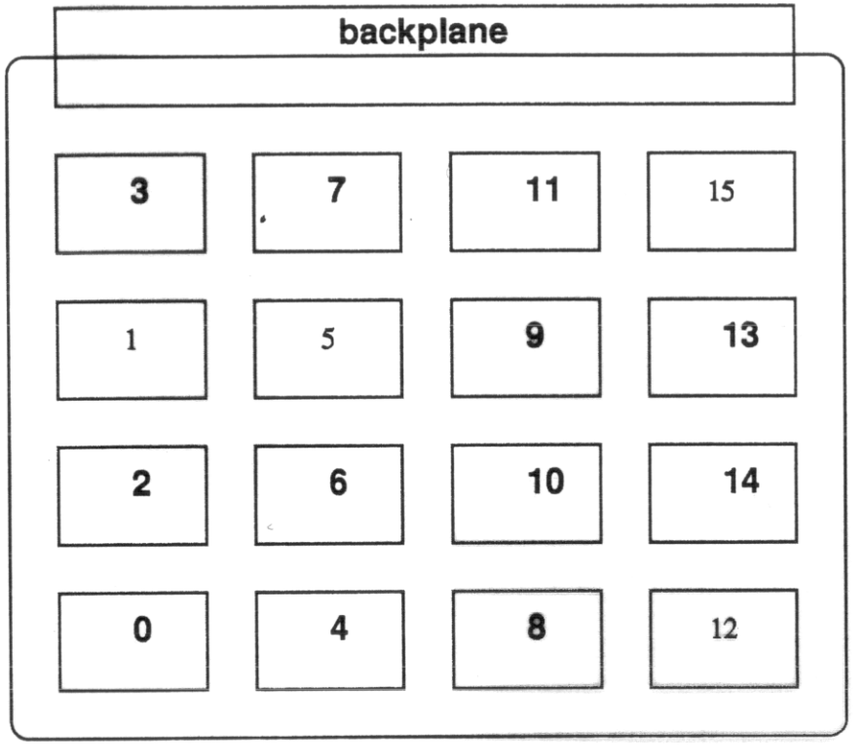
- 2 Crates are housed inside one rack (800 Gflops).

- Large systems are based on interconnected racks.



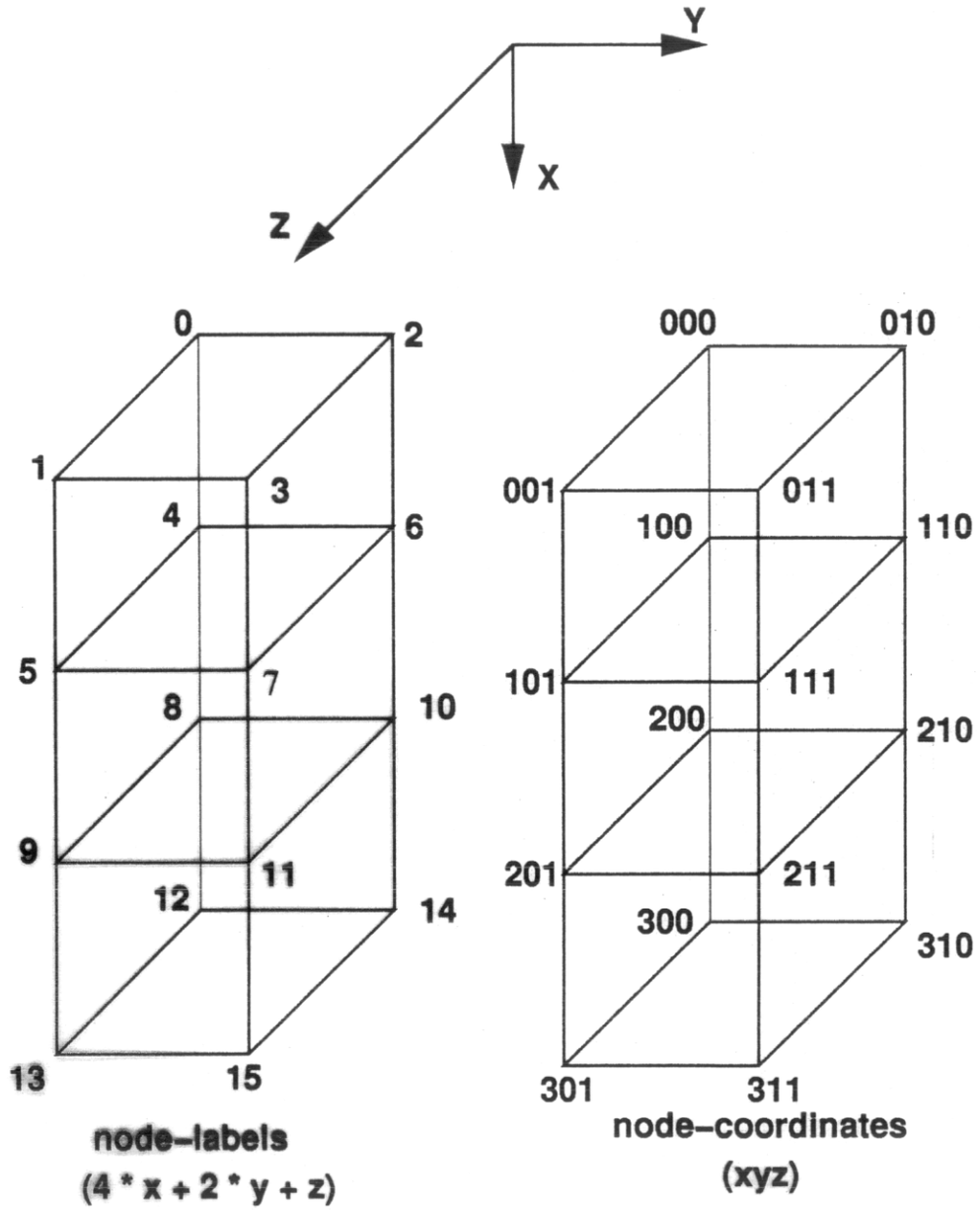

← 16 HW
128 bit + EDAC

→ 9x MT46V16M16

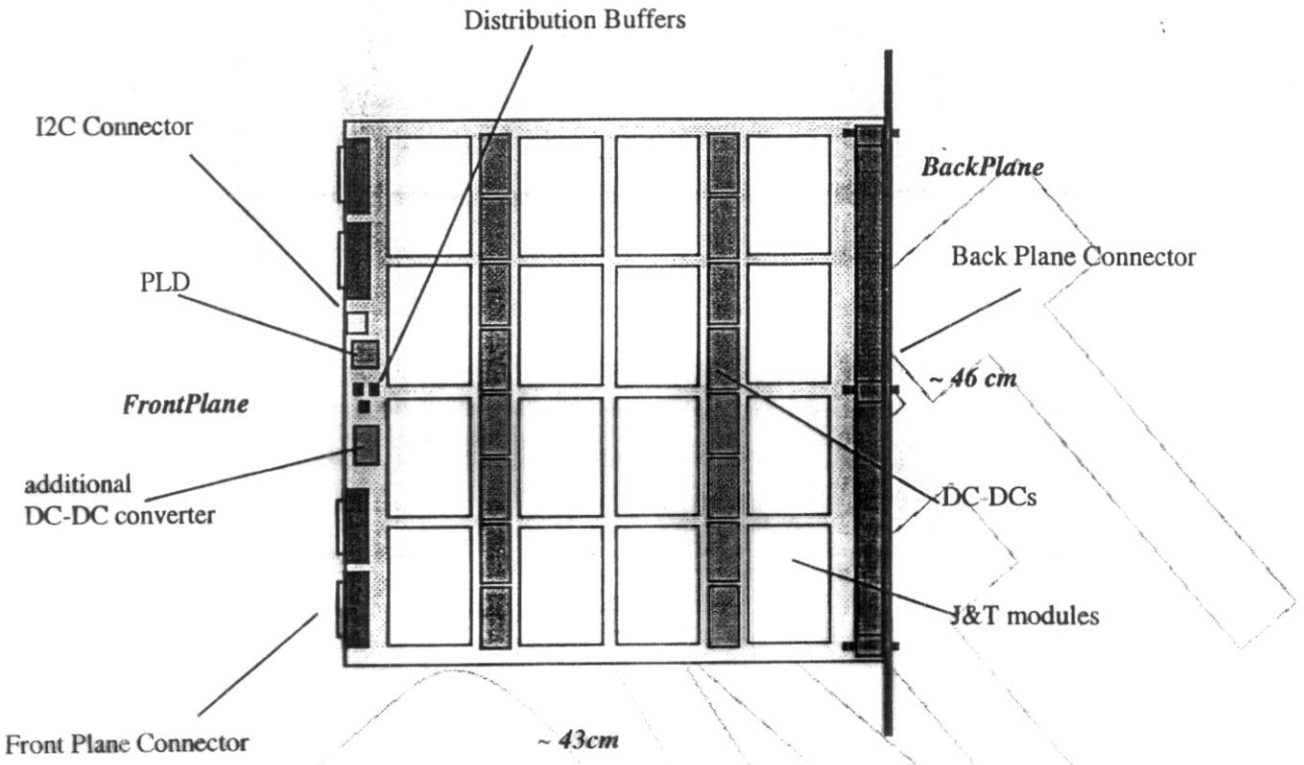



node labelling, axes orientation and all that

July 18th, 2001

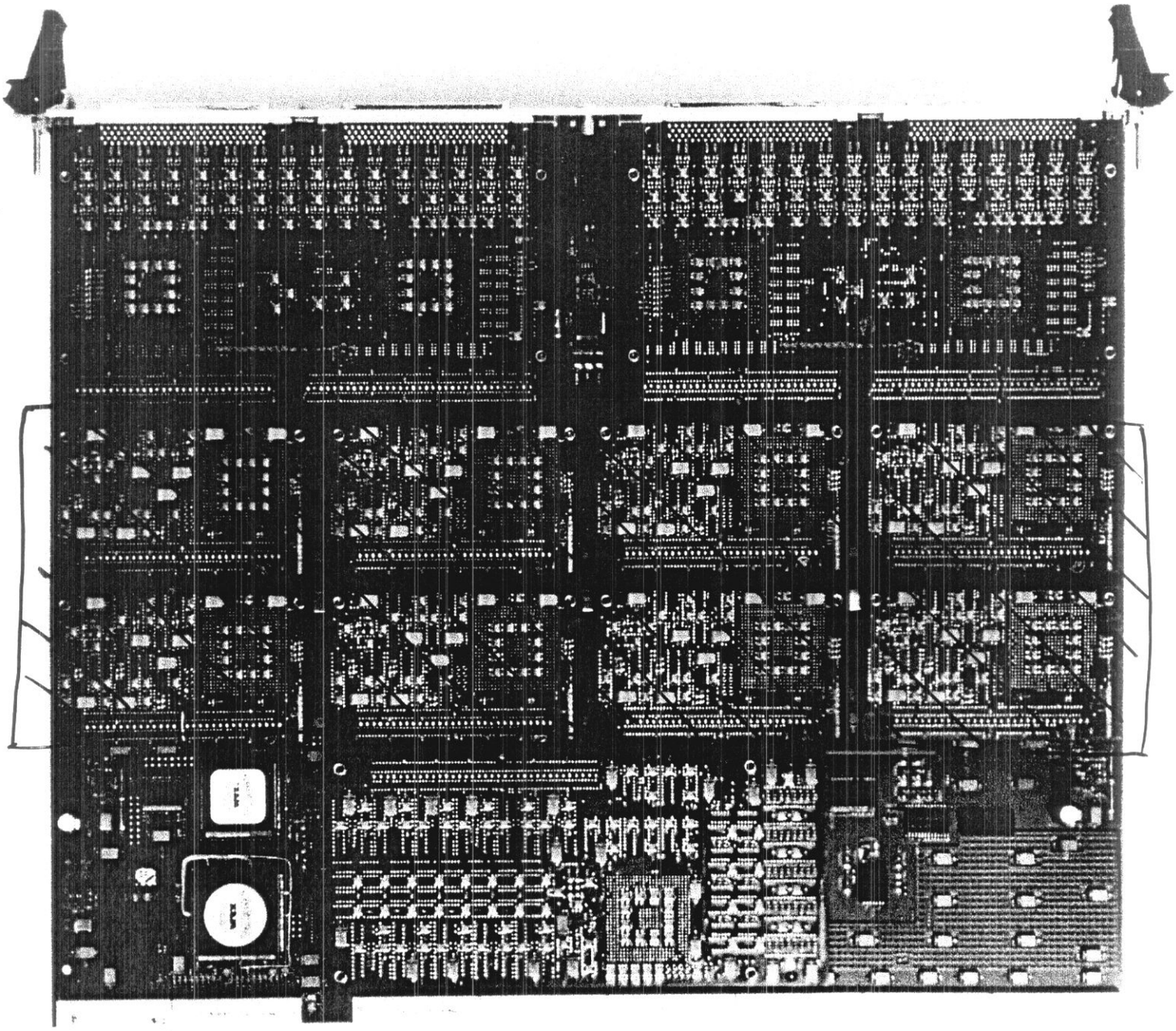


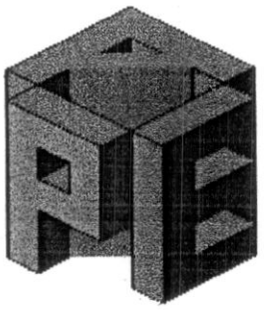
*PB Components Area Evaluation* - **ANNEX C**




*FrontPlane* Connectors *LVDS Busses Pinout* - **ANNEX D**

| | I | | | | II | | | | III | | | | IV | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0-8 | 0X- | 8X- | | | | | | | | | | | | | 11X+ | 3X+ |
| 9-17 | 0X- | 8X- | | | | | | | | | | | | | 11X+ | 3X+ |
| 0-8 | 4X- | 12X- | | | | | | | | | | | | | 15X+ | 7X+ |
| 9-17 | 4X- | 12X- | | | | | | | | | | | | | 15X+ | 7X+ |





## apeNEXT: Status of the Project

Our goals:

- At least one large prototype (400 Gflops) in late 2002
- Tflops for physics in late 2003 (0.5 €/Mflops)

Where we are today:

- All hardware bits and pieces designed.
- Complete system simulated in all details
- First hardware prototypes in early october.
- J&T prototypes in February-March 2002.
- Basic versions of the software chain already developed.
  - Wilson-Dirac operator extensively tested (66 % efficiency)
  - Jacoby solver tested
  - HMC forces/ HMC determinant under test.
  - .....