

# Status and progress of apeNEXT

Hubert Simma  
APE Collaboration

26. September 2002



INFN Pisa  
INFN Rome



DESY Zeuthen



Université de Paris-Sud  
Orsay

## Overview:

- ➡ Next steps
- ➡ Architecture
- ➡ Software
- ➡ Benchmarks
- ➡ Summary



# Challenges on the Lattice

ECFA study [1999]:

## ✗ Hadron Spectroscopy

- benchmark for Lattice QCD
- effect of light “sea”-quarks, chiral perturbation theory
- glueballs, instable particles

## ✗ $\alpha_s$ , quark masses, hadronic matrix elements

- renormalised QCD parameters with few % error
- CKM parameters, CP violation, strong phases

$$X_{exp} = X_{th}(EW) \times X_{th}(QCD)$$

- B decays
- new physics in FCNCs
- structure functions

## ✗ QCD thermodynamics

- deconfinement phase transition
- quark-gluon plasma

## ✗ Theoretical questions

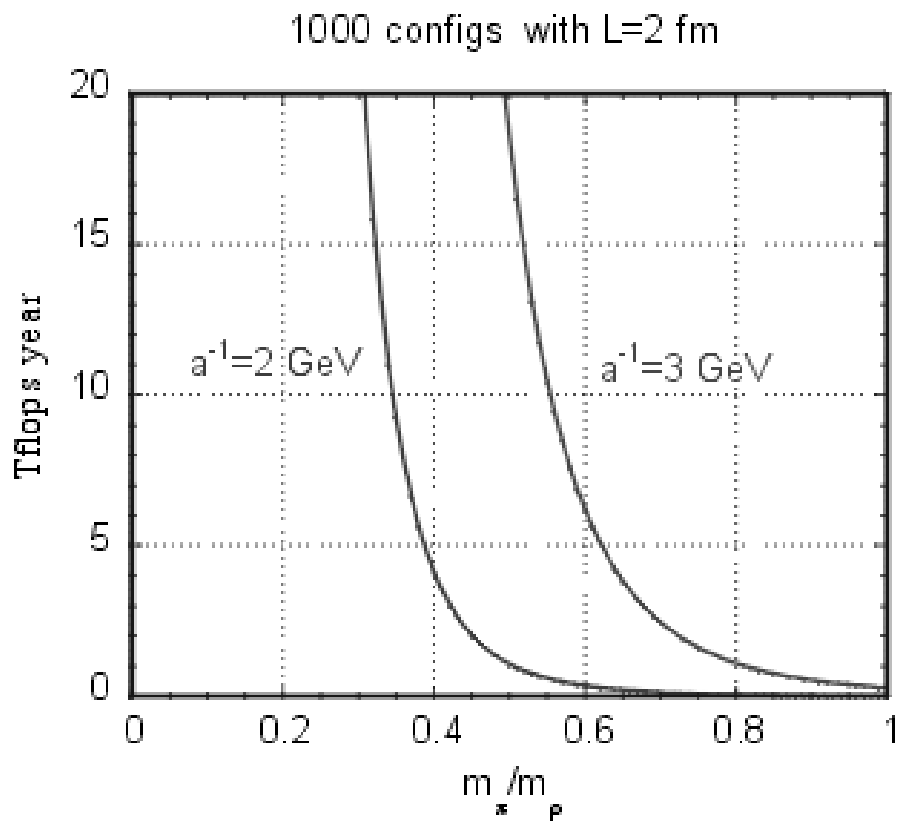
- electroweak physics
- symmetry breaking
- supersymmetry
- chiral fermions

➔ Requires  $O(10)$  TFlops compute power in 2003



## CPU Cost

[Ukawa]



Empirical estimate:

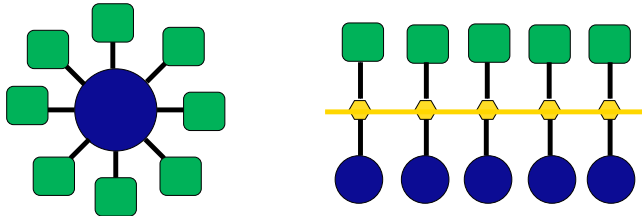
$$\begin{aligned}
 CPU \approx & \left( \frac{\#conf}{1000} \right) \times \left( \frac{m_\pi/m_\rho}{0.6} \right)^{-6} \times \left( \frac{L}{3fm} \right)^5 \times \left( \frac{1/a}{2GeV} \right)^7 \\
 & \times 2.8 \text{ Tflops} \cdot \text{year}
 \end{aligned}$$

➔ Need for dedicated compute engines



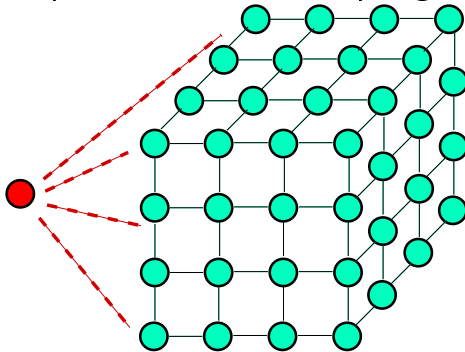
# Architectur Optimisation for Lattice QCD

- Frequent communication with nearest neighbours



⇒ fast but simple network

- Simple parallelisation + program-control



⇒ SIMD

- Dominant arithmetic operation  $a \times b + c$  with **complex** operands
- Frequent access to entire data memory




$$R = \frac{\text{\#FP operations}}{\text{memory access}} \approx 4$$




⇒ limited efficiency of cache

- Scalability  
⇒ power consumption + stability

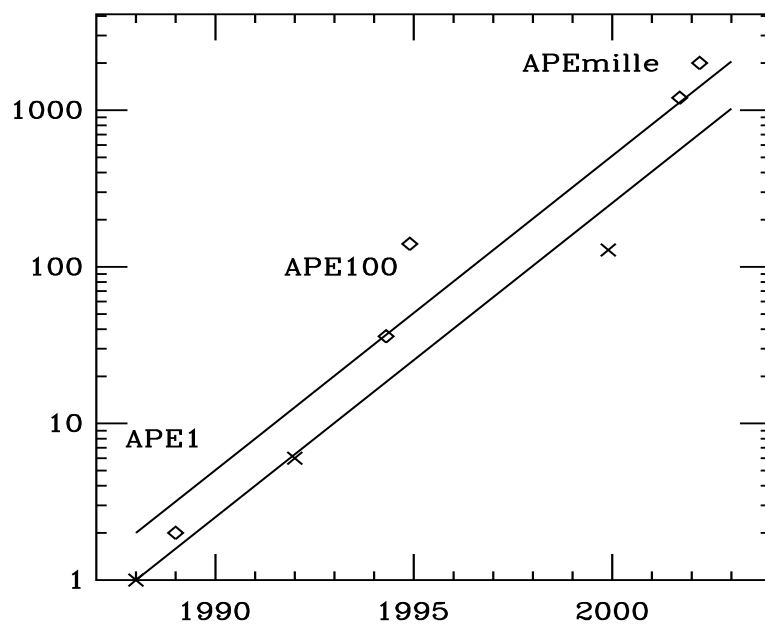


## Dedicated QCD Machines

	96:	CP-PACS:	600 GFlops	[Tokio Univ., Hitachi]
	98:	QCDSP:	1000 GFlops	[Columbia University, BNL]
	03:	QCDOC:	O(10) TFlops	[Columbia University, IBM]

		<u>Array Processor Experiment</u>	 
84-88:	APE1		[INFN]
89-92:	APE100		[INFN]
94-99:	APEmille		[INFN, DESY]
00-03:	apeNEXT		[INFN, DESY, Orsay]

APE and Moore's Law:



## Today's Workhorse: APEmille

### Large installations:

≈ 2 TFlops total in Europe

580 GFlops at DESY-Zeuthen  
(since Dec. 2001)

### Stabile operation:

- 👉 moderate effort  
(man-power + maintenance)
- 👉 some local know-how  
(hardware + software)

## Comparison

	SR8000	APEmille
Architecture:	SMP	SIMD
FP/proc [Mflops]	1500	528
Memory/FP [word/flops]	0.08	0.016
Comm. [word/flop]	1/96	1/64
Power [Watt/Mflops]	0.8	0.03
Density [Gflops/m <sup>3</sup> ]	10	100
Price [\$/Mflops sust.]	45	6



# The Future Workhorse: apeNEXT

apeNEXT Collaboration [4/2001]:



INFN



DESY



Orsay

Architectural goals:

- ✓ peak performance of large machines  $>5$  TFlops
- ✓ double precision arithmetics
- ✓  $O(50\%)$  sustained efficiency for key lattice gauge theory kernels
- ✓ large on-line data storage
- ✓ input/output channels able to sustain  $O(0.5)$  MByte/sec/GFlops
- ✓ programming environment which allows smooth migration from older APE systems
- ✓ costs of  $O(0.5)$  €/MFlops peak

Design challenges:

- ✓ all processor functionalities in single custom chip
- ✓ 200 MHz clock speed
- ✓ asynchronous communication



# apeNEXT Development Groups

## INFN Roma:

A. Lonardo	Functional Simulator
P. Vicini	PB+Backplane
(A. Michelotti – 1/2001)	VLSI



## INFN Pisa:

L. Sartori	VLSI
F. Schifano	Low-level SW, OS, Tests
R. Tripiccion	VLSI
(G. Magazzu – 3/2002)	VLSI (FPU)
(W. Errico – 1/2002)	VLSI
(T. Giorgino – 9/2001)	Tests
Parma/Milano	Tests, Libraries



## France:

Orsay	Tests, Benchmarks, Libraries
Rennes	Assembly Optimiser



## DESY-Zeuthen:

H. Kaldass	Tests
N. Paschedag	C-compiler, Tests
D. Pleiter	Tests, OS
H. Simma	Tests, TAO-compiler



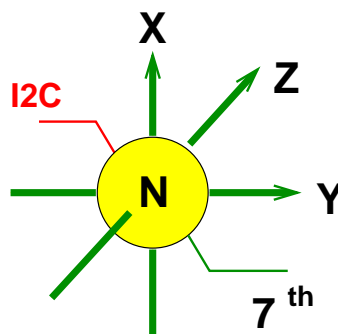
plus additional contributions by Dubna + Bielefeld



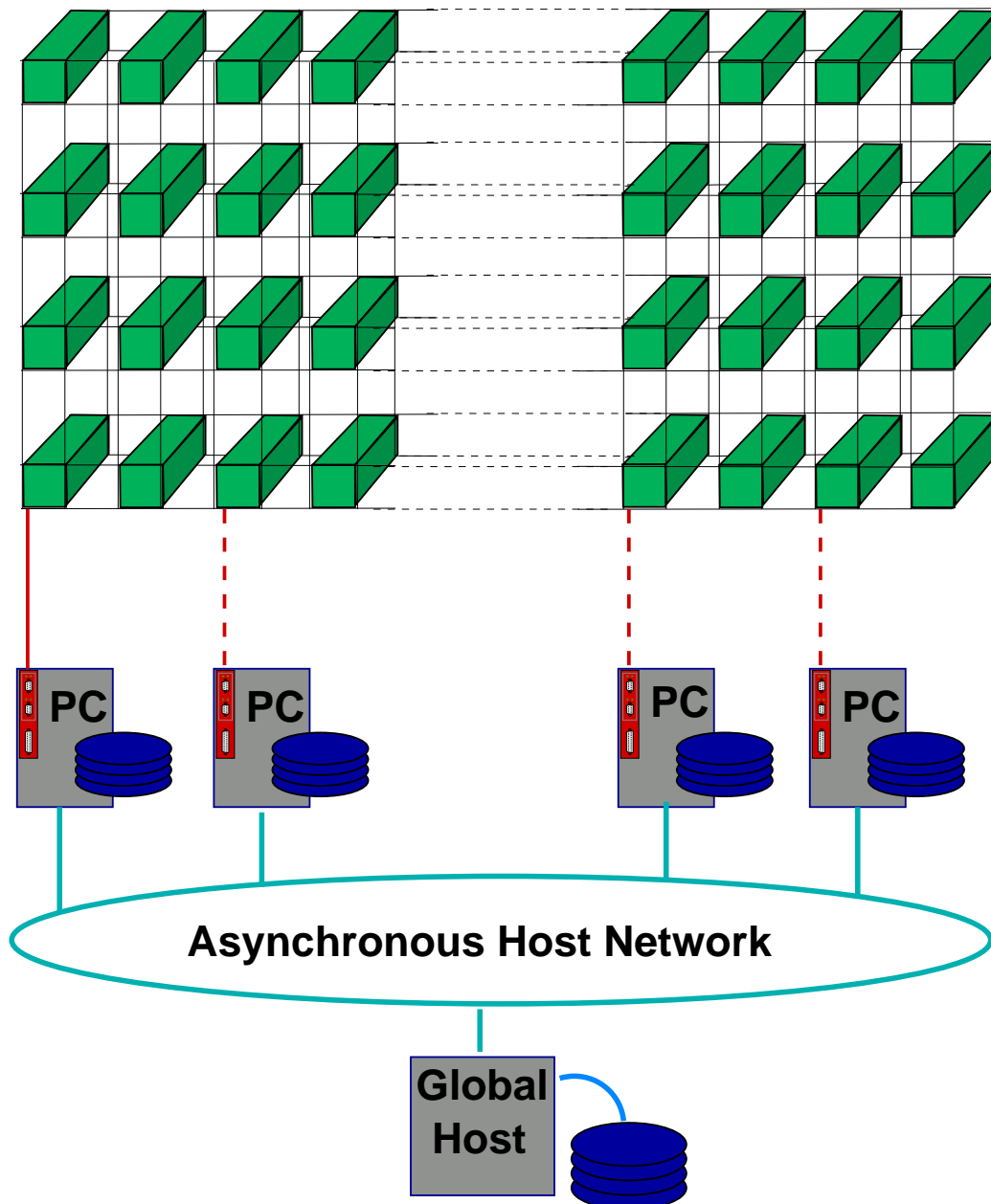


## apeNEXT Global Architecture

- 3-d array of  $O(2048)$  **autonomous** nodes  
→ all processor functionalities in single chip
- **asynchronous** operation (SPMD)  
→ synchronisation only at data exchange  
→ simpler technology upgrade
- distributed data and **program** memory
- **concurrent** communications via fast network
- data I/O via communication network and **7th link**
- low-level control via slow links (I2C)
- host = cluster of  $1 \dots N_{PB}$  Linux PCs



## apeNEXT Host System



## apeNEXT Processor Features

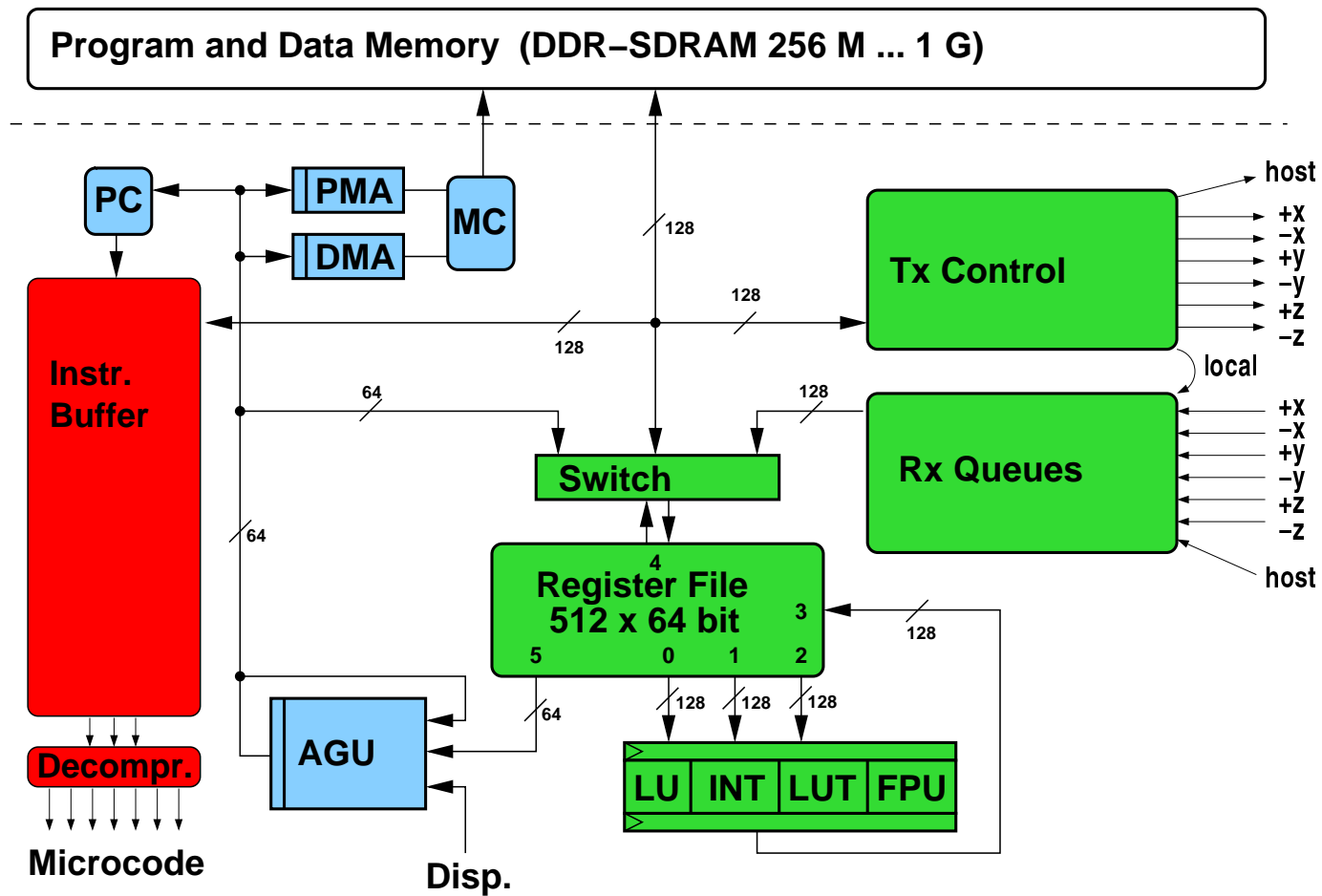
- 4 K  $\times$  128-bit Instruction cache and FIFO
- Microcode de-compression
- Integrated Memory Interface (DDR)
- Prefetch queues (1024+7\*128+32 128-bit words)
- Integrated Communication Interface (LVDS)
- Remote register-register communications
- Floating-Point “normal” Operations  $a \times b + c$   
Arithmetic throughput (64-bit IEEE):

Format	complex	vector	integer
Op. per clock	8	4	2

- 8-bit LUT for  $1/x$  and  $1/\sqrt{x}$
- 15 AGU instructions (including MULA and LSH)
- 6-port RF (256  $\times$  128 bit registers)
- Indirect register addressing (for LUTs or windowing)
- 50 configuration registers  
(accessible via I2C and in RUN-mode)
- . . .



# apeNEXT Processor Chip



- ASIC 0.18  $\mu m$  CMOS
- $9.2 \times 8.2 mm^2$ , 1.2 M gates
- 200 MHz clock
- 2 (proc) + 5 (mem) Watt power consumption

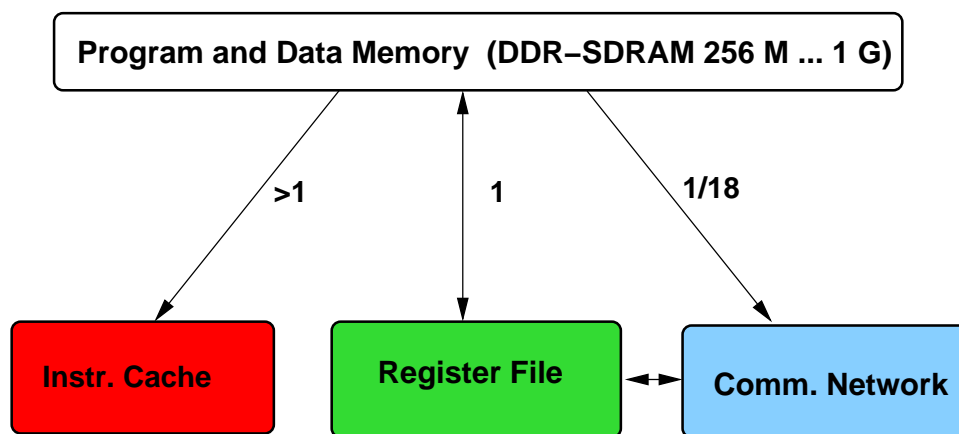


# Memory Interface

Memory Bandwidth: 128 bit (+ECC) / cycle

$$\Rightarrow R_{QCD} = \frac{\#FP \text{ operations}}{\text{memory access}} \approx 4$$

Problem:

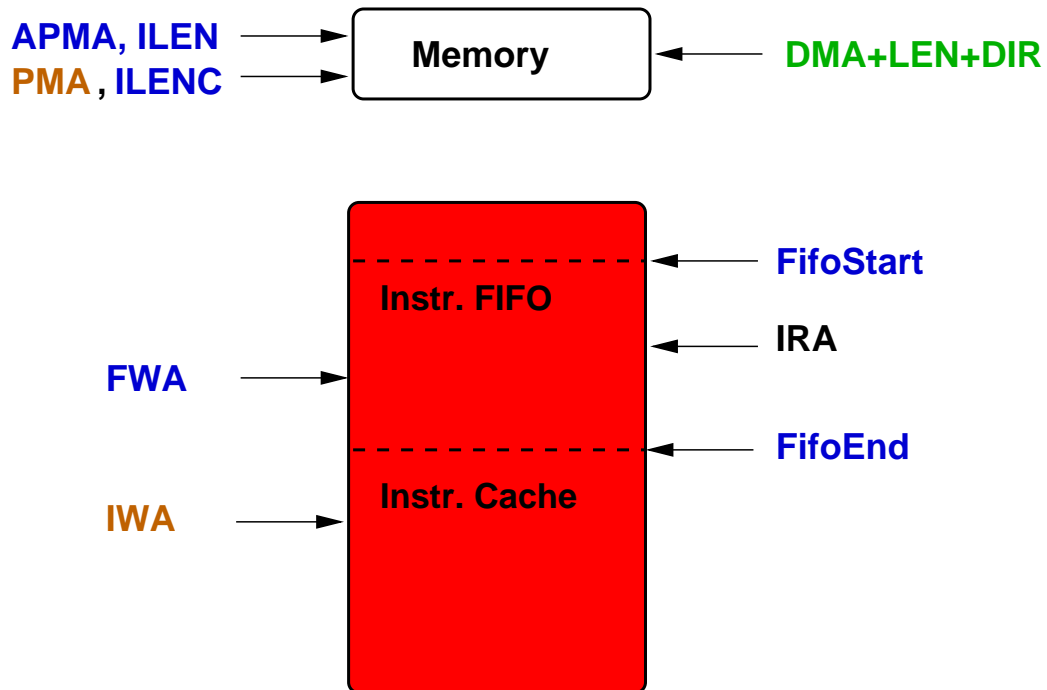


- Memory latency: > 16 cycles
- Cost of instruction loading
- Lower bandwidth of remote communications

☞ 6 different memory access types



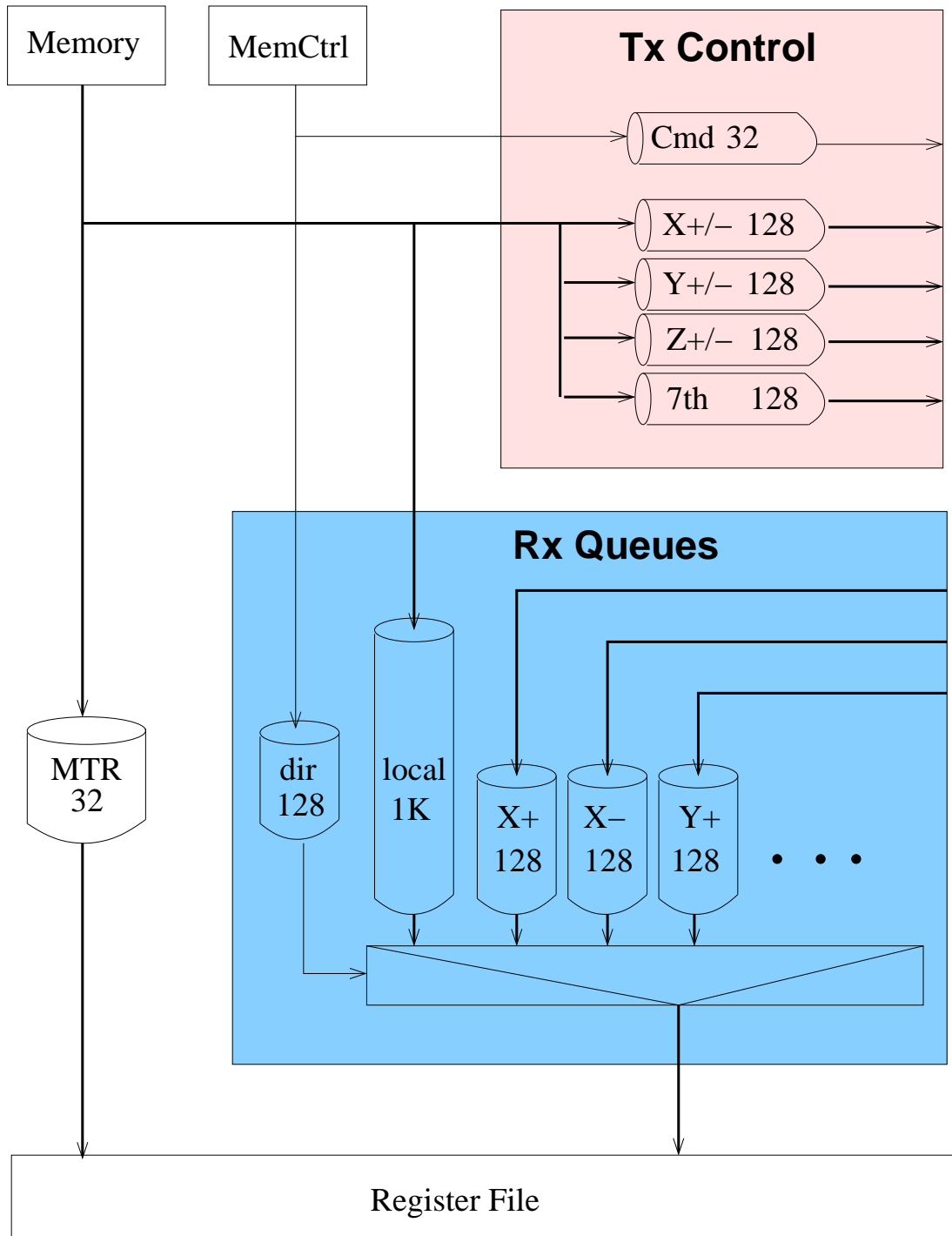
## Control of Memory Accesses



SW control (assembly)	HW Support (control registers)
MTF	APMA, ILEN, FWA, FifoStart/End
MTFC	PMA, ILENC, FWA, FifoStart/End
MTI	PMA, ILEN, IWA
MTR, RTM	DMA, LEN
MTQ	DMA, LEN, DIR

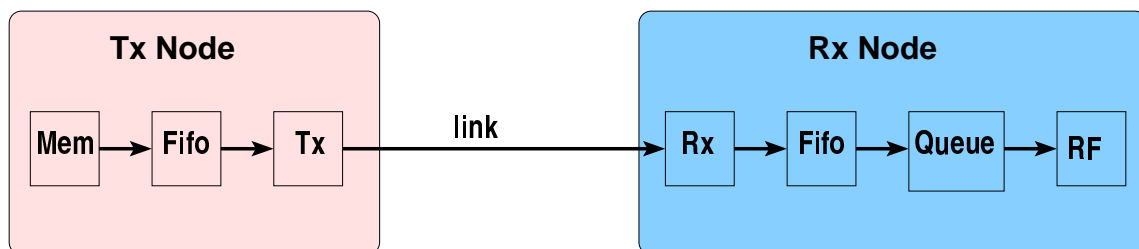


## Data Queues



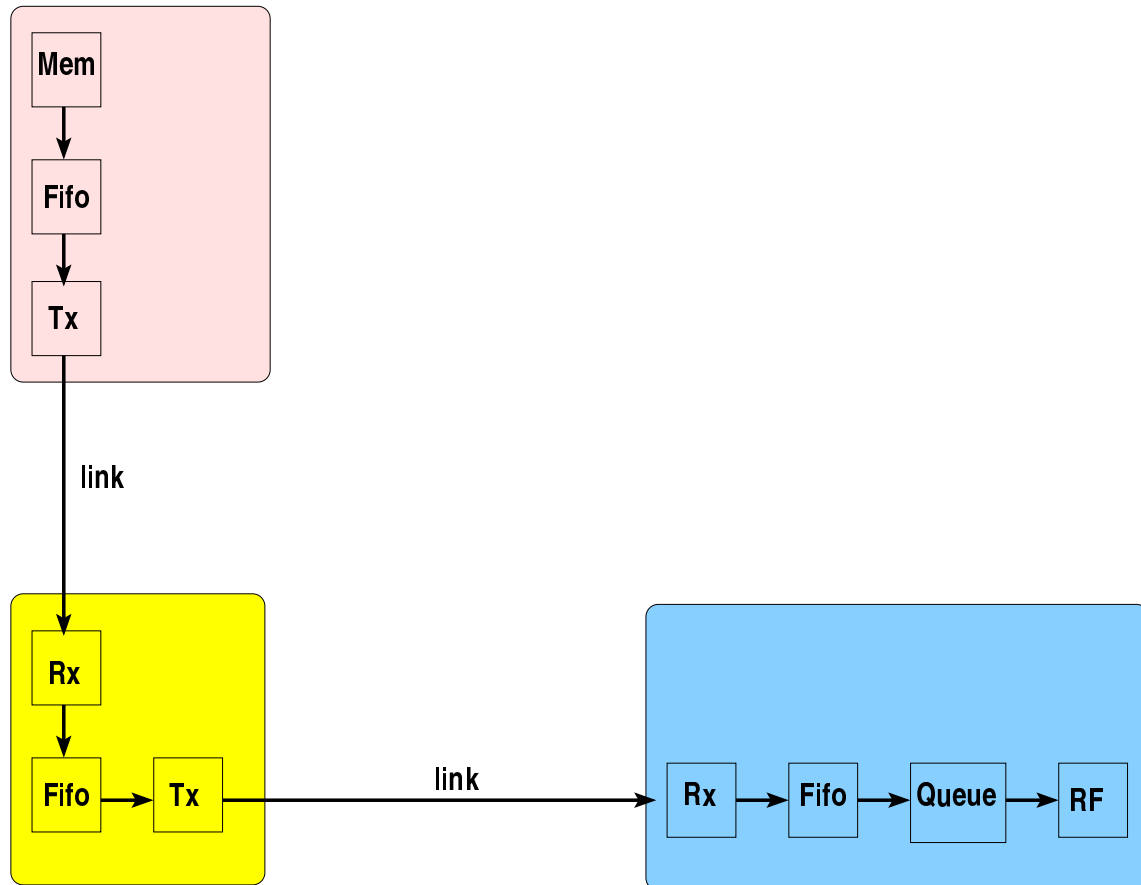
## Communication Network

- 7 bi-directional LVDS links
- 128 bit data + 16 bit CRC
- high bandwidth: 16 Byte/18 cycles (180 MByte/s)
- low latency:  $\approx 25$  cycles (125 ns)
- concurrent send and receive
- support for non-homogeneous communications
- configurable direction mapping
- concurrent transfer along orthogonal directions





## Two- and Tree-Step Communications

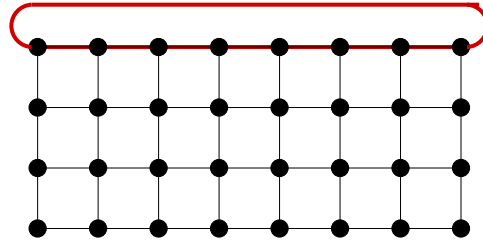


👉 26 HW-routed communication directions  
(all neighbors on  $3 \times 3$  hypercube)

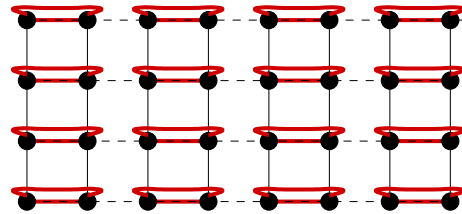


## Link Re-mapping

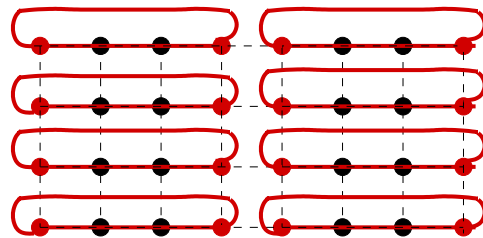
Full machine (e.g. along  $x$ ): No re-mapping



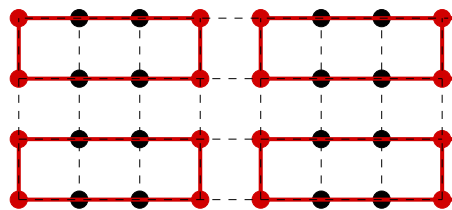
Period 2 (e.g. along  $x$ ):  $x_{\pm} \rightarrow x_{\mp}$



Period 4 (along  $x$  only):  $x_{\pm} \rightarrow 7th$



2-d topologies (e.g.  $x$  mapped into  $xy$ -plane):  $x \rightarrow x$  or  $y$



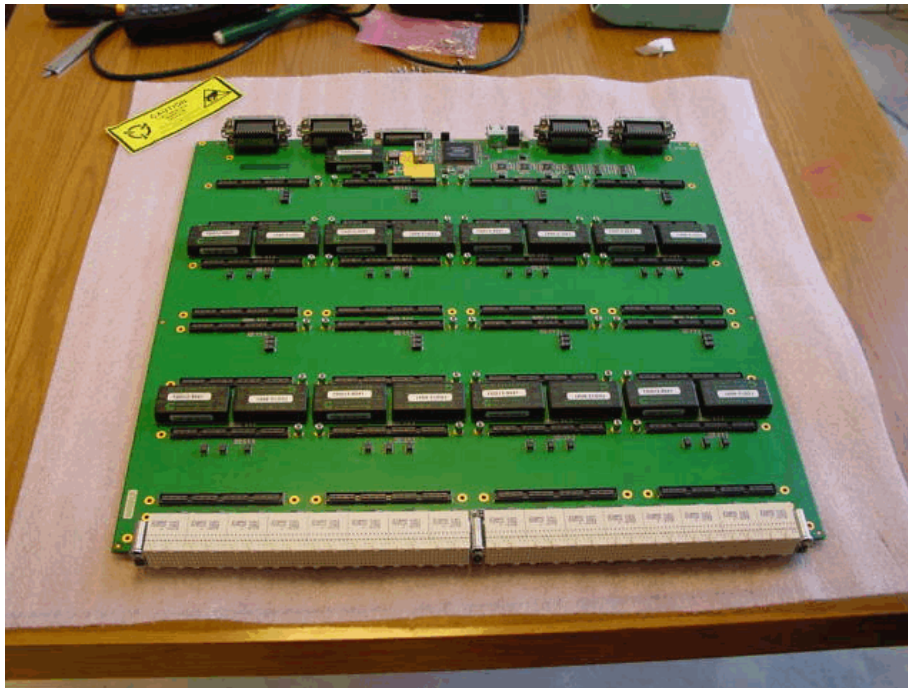
1-d topologies (e.g. for routing of I/O)



## apeNEXT Boards and Backplane

### Processing Board:

- 16 processor piggy-backs per PB
- 1 FPGA (Root logic + I2C) + DC-converters
- 7 LVDS signal layers, 8 GND and PWR layers
- Front connectors: 9 ( $\pm x$ , 7th) links à 18 LVDS pairs
- Back connectors: 32 ( $\pm y$ ,  $\pm z$ ) links à 18 LVDS pairs



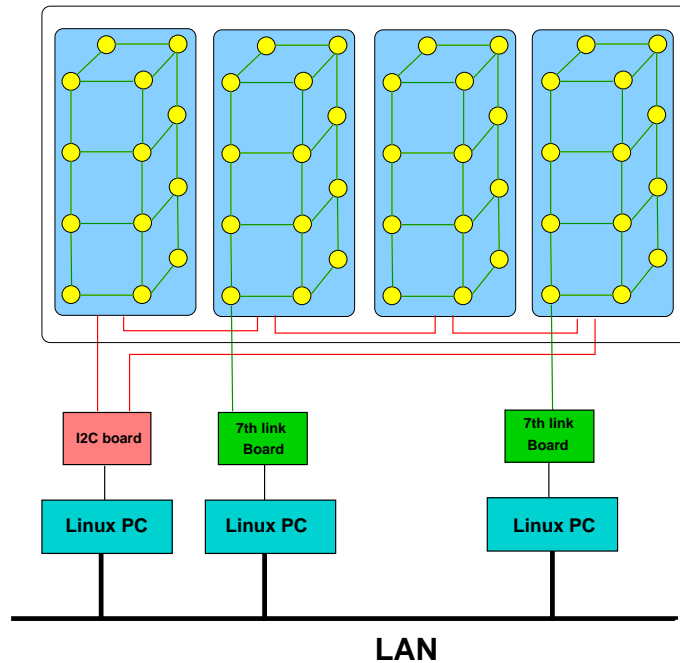
### Backplane:

- 34 LVDS layers for  $y$  and  $z$  channels
- no PCI

👉 Prototypes of PB and Backplane delivered end 2001



# Operating System



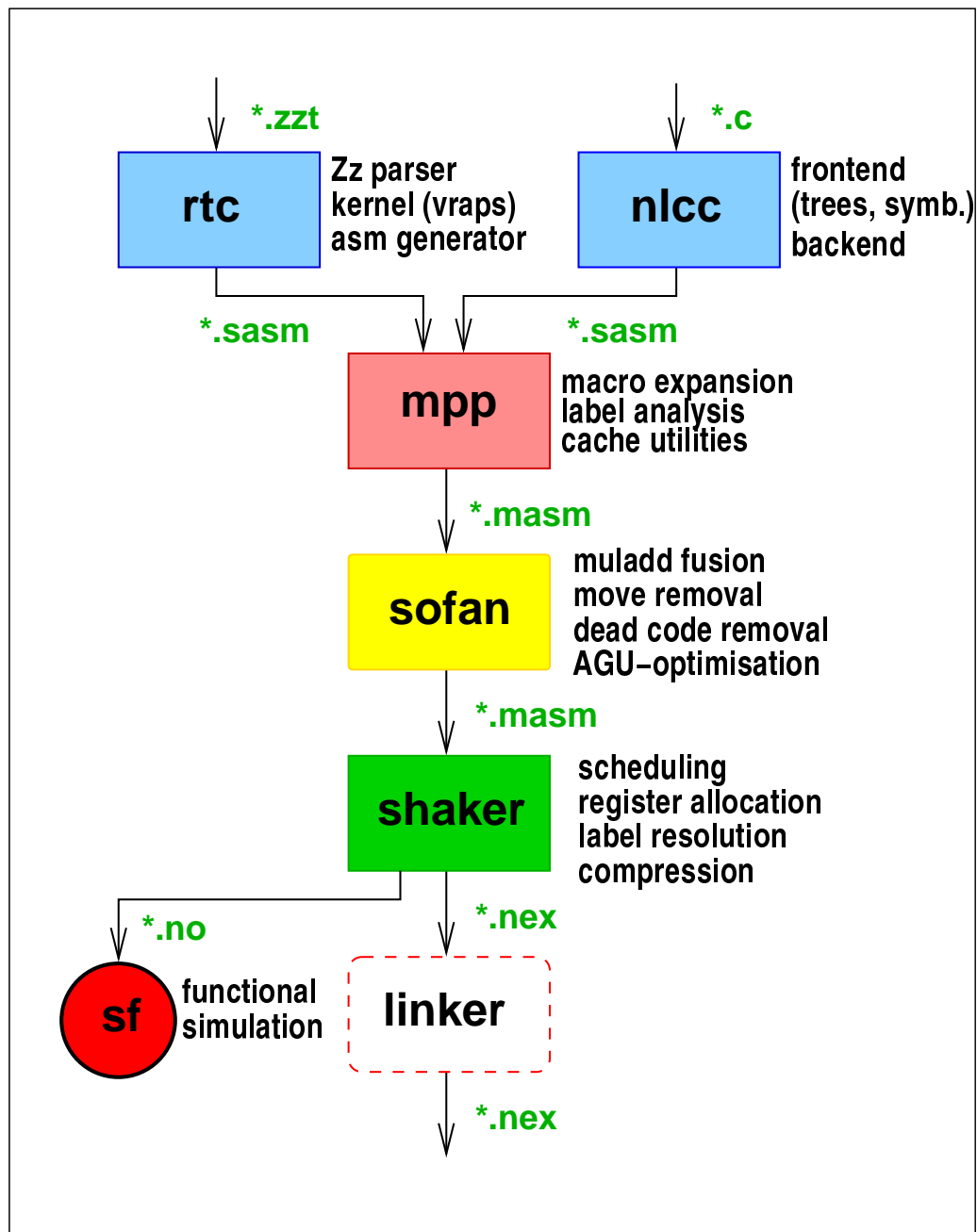
Main elements:

- Host: interface for slow control channel (I2C)  
bootstrap, configuration, exception handling, debugging
- Host: interface for fast data channel (7th link)  
fast data and program I/O
- Node: service routines  
data moving and routing to/from 7th link

☞ Simple high-level structure as in APEmille



# Compilation Chain



✗ stable and unified TAO and C environment

✗ improved low-level optimizations (assembly, microcode)



## Performance: Linear Algebra

### Vector Norm

$$c = \sum_{i=1}^N a_i^* a_i \quad a_i \dots \text{complex numbers}$$

Memory bandwidth: 1 word/cycle

Maximum sustained performance: 50% (FP saturation)

Measured performance ( $N = 2^3 \times 16 \times 12 = 1536$ ):

burst len	unroll	% peak
16	1	13
16	8	37
64	8	44

### Scalar Product

$$c = \sum_{i=1}^N b_i a_i \quad a_i, b_i \dots \text{complex numbers}$$

Maximum sustained performance: 50% (memory limited)

Measured performance ( $N = 2^3 \times 16 \times 12 = 1536$ ):

burst len	unroll	% peak
16	1	21
16	8	41
64	8	47



## Performance: QCD

Dirac Operator

$$\psi_e = D_{eo}\phi_o$$

where  $D$  is hopping term of the (unimproved) Wilson-Dirac operator.

For this operation we have

$$\frac{\text{\#normal operations}}{\text{\#memory operations}} \approx 1.4$$

Use prefetch by 2 sites and local/remote gauge fields:

$V_{txy} \times L_z$	gauge	unroll	% stretch	% peak
$2^3 \times 16$	remote	2	28	42
$4^3 \times 16$	remote	2	11	53
$2^3 \times 16$	local	2	2	57
$2^3 \times 16$	local	4	4	59



## Summary

- ☞ Design and performance goals reached
- ☞ Schedule:
  - 12/2001 Prototype of board and backplane delivered
  - 1/2002 VHDL design of processor completed
  - 10/2002 Chip sign-off
  - 1/2003 Running Board with 16 Processors
  - ≤6/2003 1–2 running prototype crates (400–800 GFlops)
- ☞ Integration of (centralised) large systems:  
German LQCD community requires  $O(20 \text{ Tflops})$

