Status of apeNEXT

Hubert Simma

9 April 2002

Inhalt:

- ✘ Physics
- ✘ APE Project
- ✘ apeNEXT Architecture
- ✘ Status and Challenges

Herausforderungen auf dem Gitter

ECFA Studie [1999]:

- ✘ Hadronen Spektrum
  - Benchmark für die Gitter QCD
  - Effekte leichter "See"-Quarks, chirale Störungstheorie
  - Gluebälle, instabile Teilchen
- ✘ α_s, Quark Massen, hadronische Matrixelemente
  - Renormierte QCD Parameter mit wenigen % Fehler
  - CKM Parameter, CP Verletzung, starke Phasen

    $$X_{exp} = X_{th}(EW) \times X_{th}(QCD)$$
  - B-Zerfälle
  - Neue Physik in FCNCs
  - Strukturfunktionen
- ✘ QCD Thermodynamik
  - Deconfinement Phasenübergang
  - Quark-Gluon Plasma
- ✘ Theoretische Fragen
  - Elektroschwache Physik
  - Symmetriebrechung
  - Supersymmetrie
  - Chirale Fermionen

➨ Bedarf $O(10)$ TFlops Rechenleistung in 2003

Theorie der starken Wechselwirkung – QCD

Teilchen/Felder	Eigenschaften	Parameter
Quarks	Spin 1/2, Pauli-Prinzip	m_q
Gluonen	Asymptotische Freiheit Confinement	α_s

✔ Bei hohen Energien (Störungstheorie) sehr genau verifiziert

✔ Große Vorhersagekraft da nur $1 + N_f$ fundamentale Parameter

Ziel: Ab-initio Berechnung von Größen bei (niedrigen) hadronischen Energieskalen

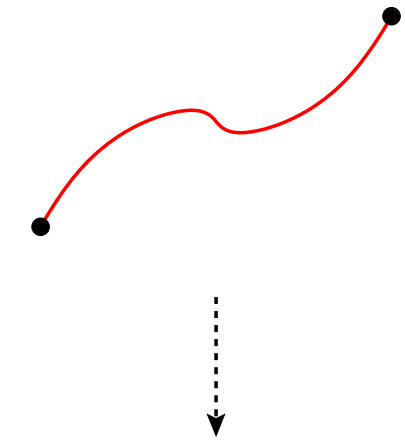
➠ Nicht-störungstheoretische Methoden

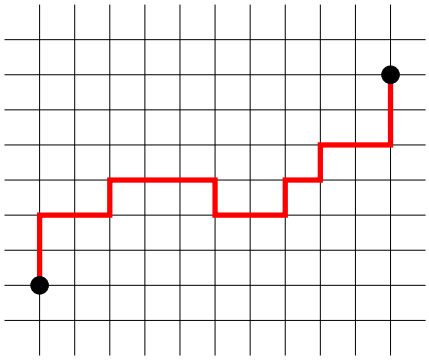
Ausgangspunkt ist die Pfadintegral-Formulierung:

$$\langle \ldots \rangle = \int_{\text{Felder}} D[\phi] \ \ldots \ e^{-S_E(\phi,P)}$$

wobei $S_E = \int s\left[\phi(x)\right] \, d^4x$

Feldtheorien auf dem Gitter





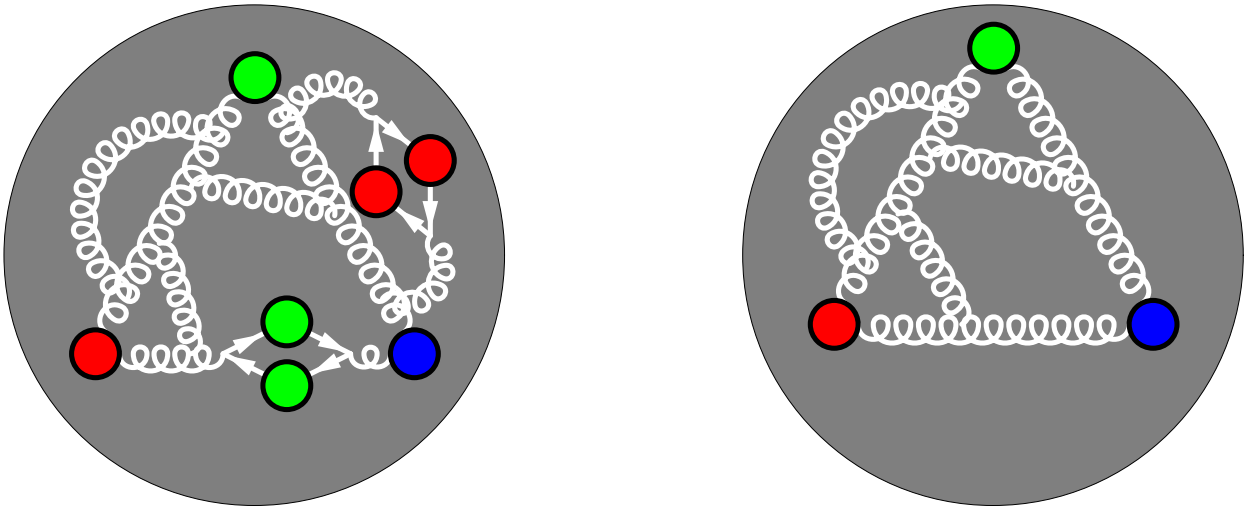
Felder "leben" in diskretisierter Raum-Zeit

- ➡ Numerische Simulation ist möglich
- ➡ Methoden der statistischen Physik (Monte-Carlo Methode)
- ➡ Gitter-Formulierung ist einzige Methode für nicht-störungstheoretische Berechnungen in einer (nicht analytisch lösbaren) Feldtheorie

Praktische Schwierigkeiten der Gitterformulierung

- Grundsätzliche theoretische Probleme der Gitterformulierung (Rotationsinvarianz, chirale Symmetrie, ...)
- Extrapolation der Parameter in den physikalischen Bereich
- Numerische Simulationen sind äusserst aufwendig
  - ✘ Theoretische Konzepte
  - ✘ Gute Algorithmen
  - ✘ Immense Rechnerressourcen

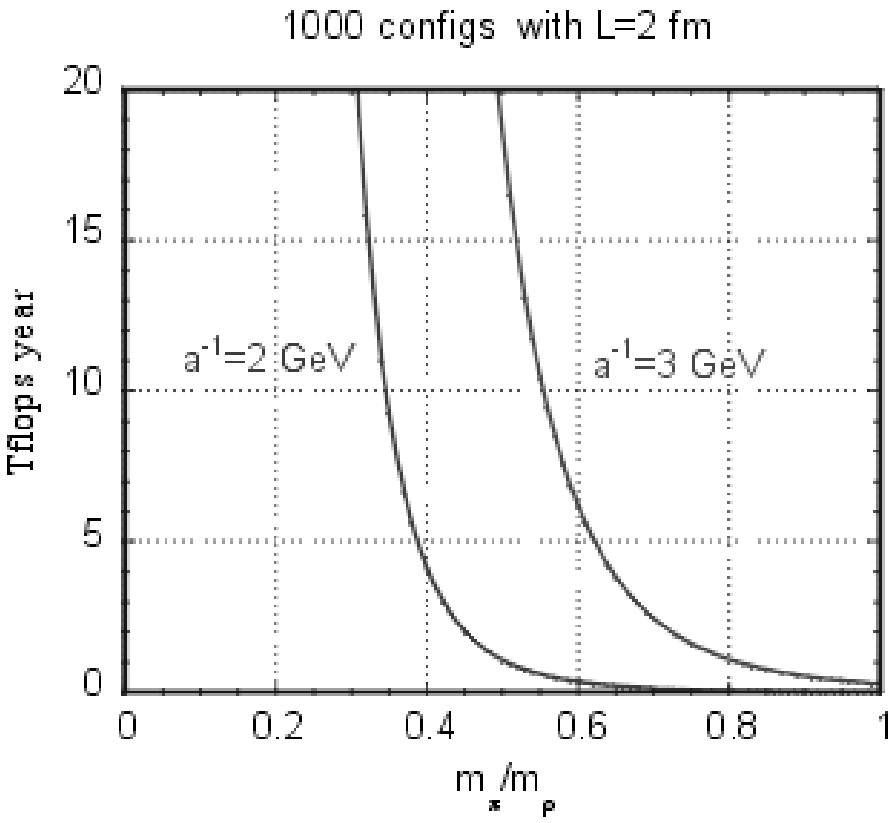
Dynamische Quarks vs. "Quenched" Näherung



$$\int D[\psi] \ldots e^{-\bar{\psi}(D+m)\psi} = det(D+m)^{N_f/2} \rightarrow const$$

CPU Kosten

[Ukawa]



Empirische Abschätzung:

$$CPU \approx \left(\frac{\#conf}{1000}\right) \times \left(\frac{m\pi/m\rho}{0.6}\right)^{-6} \times \left(\frac{L}{3fm}\right)^{5} \times \left(\frac{1/a}{2GeV}\right)^{7}$$

$$\times\ 2.8\ TFlops \cdot year$$

APE100

Entwicklung:	INFN
Betrieb:	1994-...
#CPU:	896
Max. Leistung:	45 GigaFlop/Sek.

⇓

APEmille

Entwicklung:	INFN + DESY
Betrieb:	2000-...
#CPU:	1024
Max. Leistung:	550 GigaFlop/Sek.

⇓

apeNEXT

Entwicklung:	INFN, DESY, Orsay
Betrieb:	2003-... ?
#CPU:	10.000 ?
Max. Leistung:	16 TeraFlop/Sek. ?

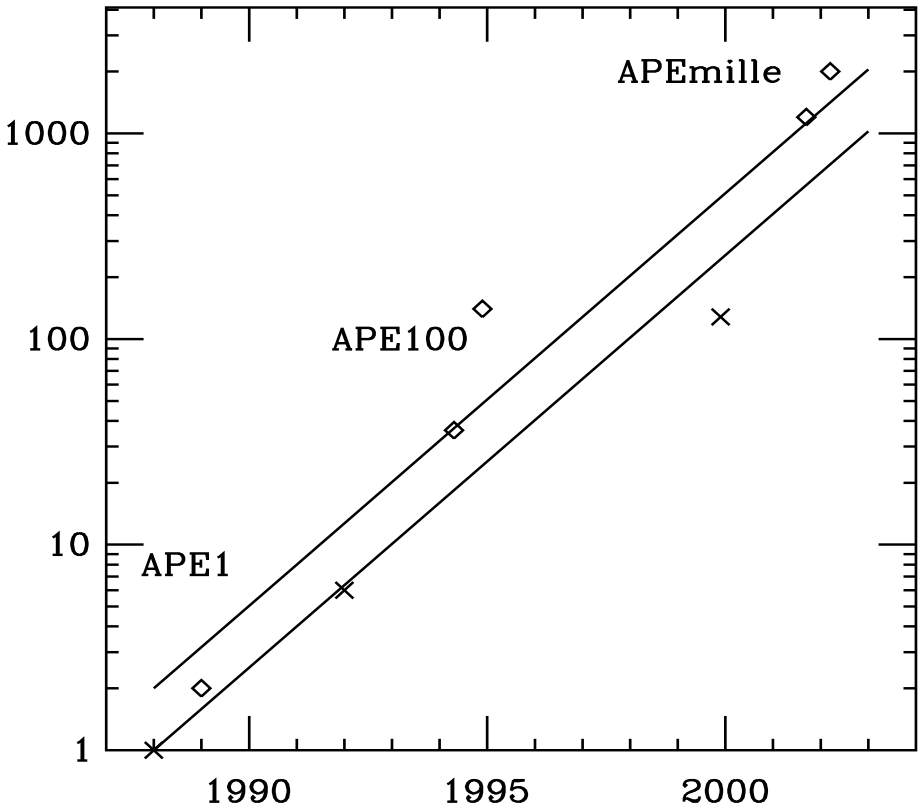
APEmille bei DESY-Zeuthen

- ☞ Installation von 550 GFlops im Dez. 2001 abgeschlossen
- ☞ Stabiler Betrieb bei geringem Aufwand (Personal+Unterhalt)
- ☞ Zugang durch NIC Rechenzeitkommission geregelt

Vergleich mit kommerziellem Supercomputer

	SR8000	**APEmille**
Architecture:	SMP	SIMD
FP/proc [MFlops]	1500	528
Memory/FP [word/Flops]	0.08	0.016
Comm. [Flop/word]	1/96	1/64
Power [Watt/MFlops]	0.8	0.03
Density [GFlops/m^3]	10	100
Price [$/Mflops sust.]	45	6

APE Entwicklung und Moore'sches Gesetz



Andere Eigenentwicklungen für Gitter QCD

96:	CP-PACS:	600 GFlops	[Tokio Univ., Hitachi]
98:	QCDSP:	400+600 GFlops	[Columbia University, BNL]
03:	QCDOC:	$O(10)$ TFlops	[Columbia University, IBM]

apeNEXT Project

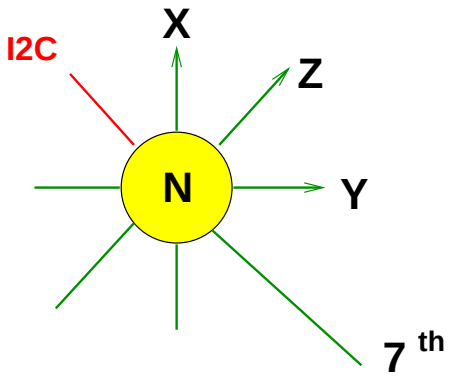
INFN DESY Orsay

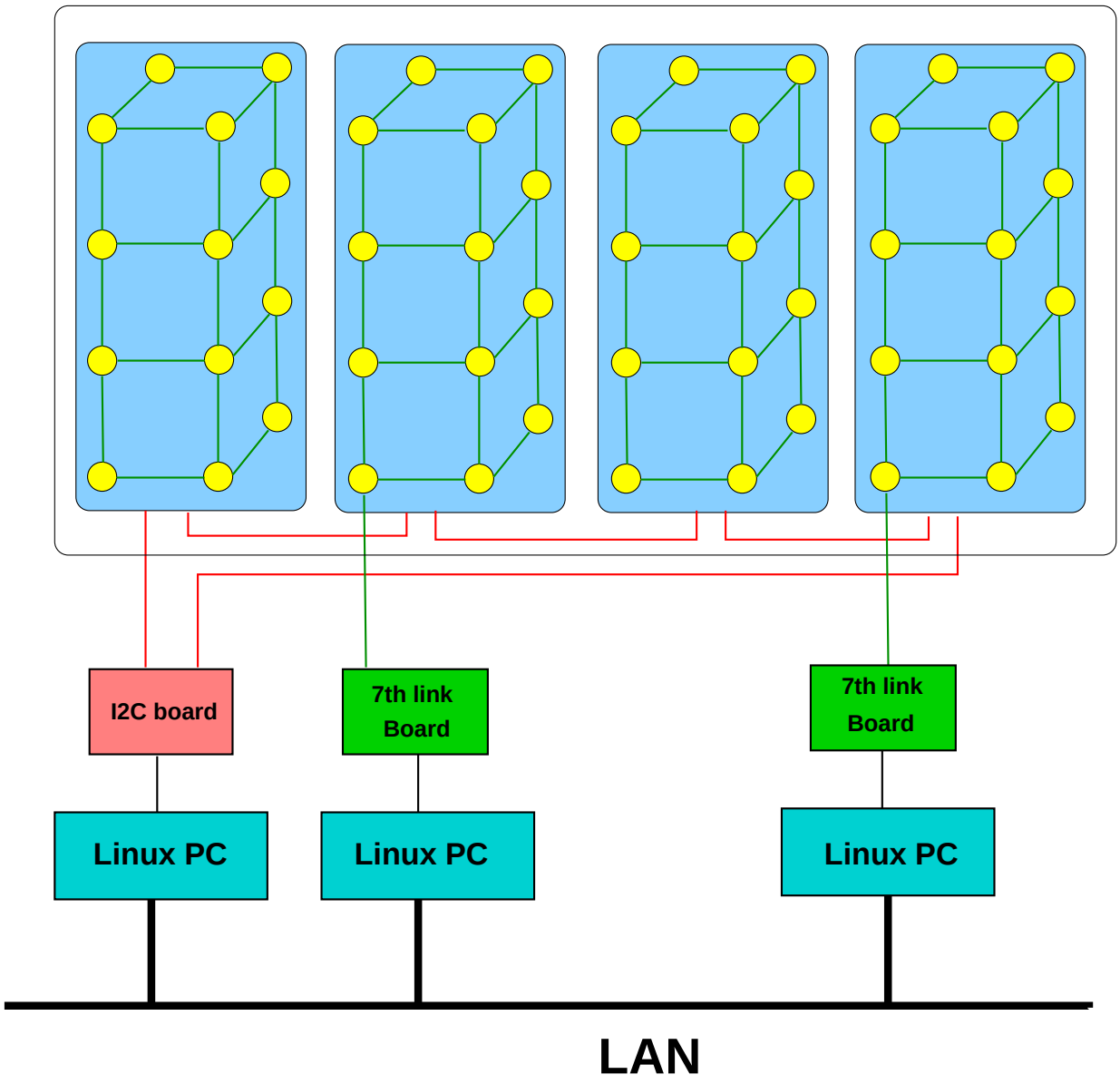
Collaboration Agreement:

- ➜ first prototype by end 2002
- ➜ shared development efforts:
  2070 kEuro (6:3:1), 280+140+50 FTE months
- ➜ large systems ($O(3)$ TFlops) at 0.5 Euro/MFlops peak

apeNEXT Global Architecture

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

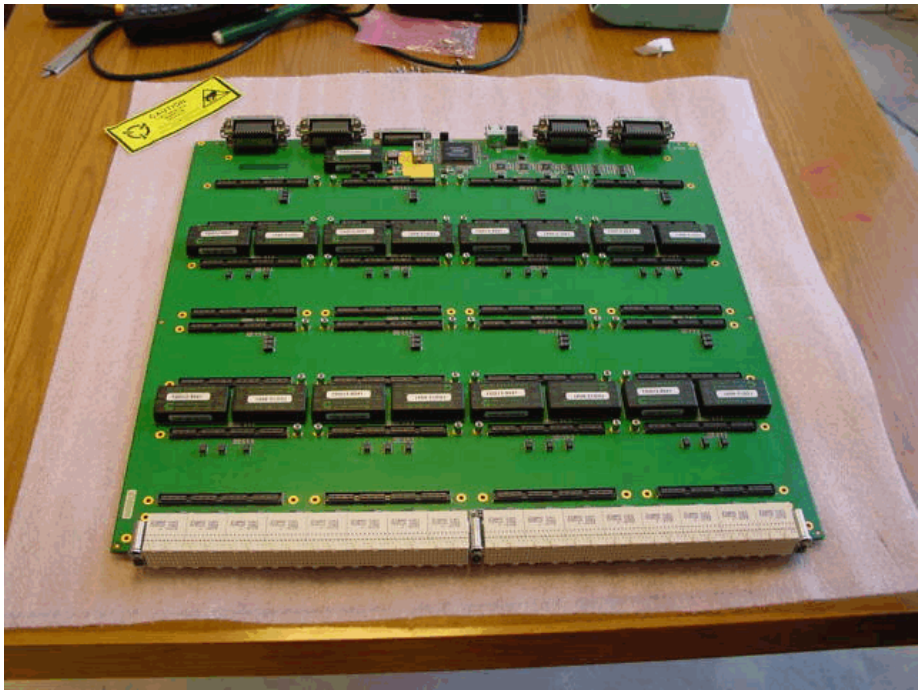



I2C board
7th link Board
7th link Board
Linux PC
Linux PC
Linux PC
LAN

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

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

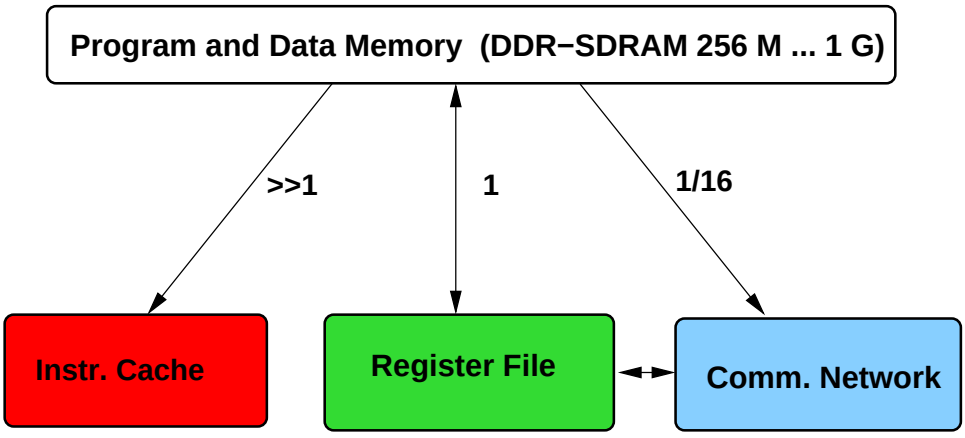
- 8-bit LUT for $1/x$ and $1/\sqrt{x}$
- 15 AGU instructions (including `MULA` and `LSH`)
- 6-port RF (256 $\times$ 128 bit registers)
- Indirect register addressing (for LUTs or windowing)
- 50 configuration registers
  (accessible via I2C and in Run-mode)
- ...

Memory Interface

Memory Bandwidth: 128 bit (+ECC) / cycle

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

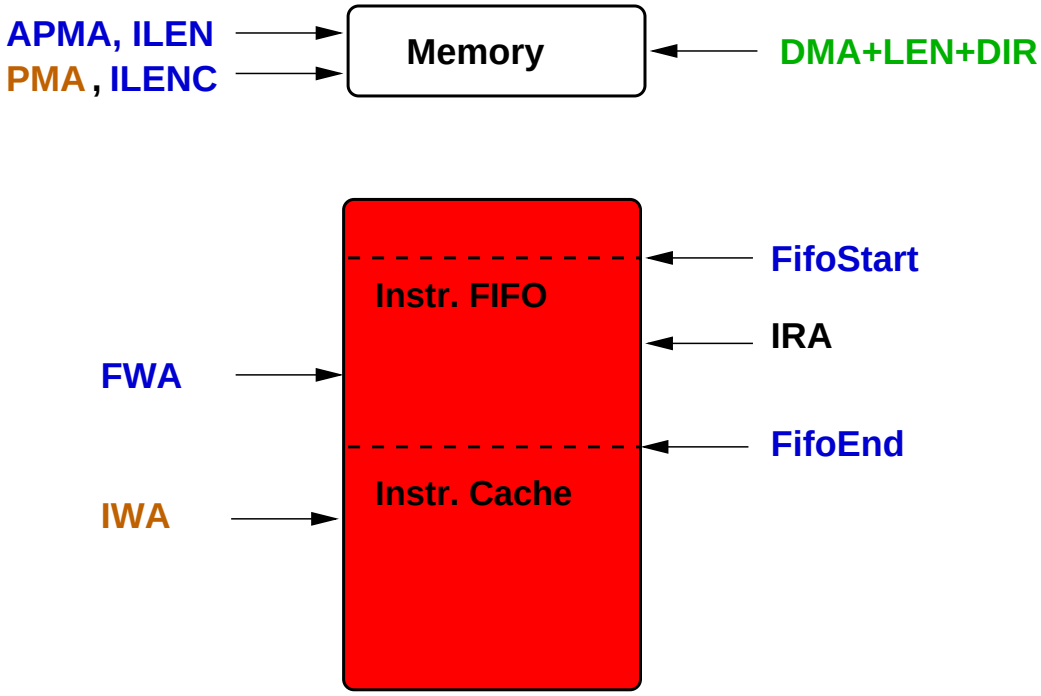
Problem:



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

☞ 6 different memory access types

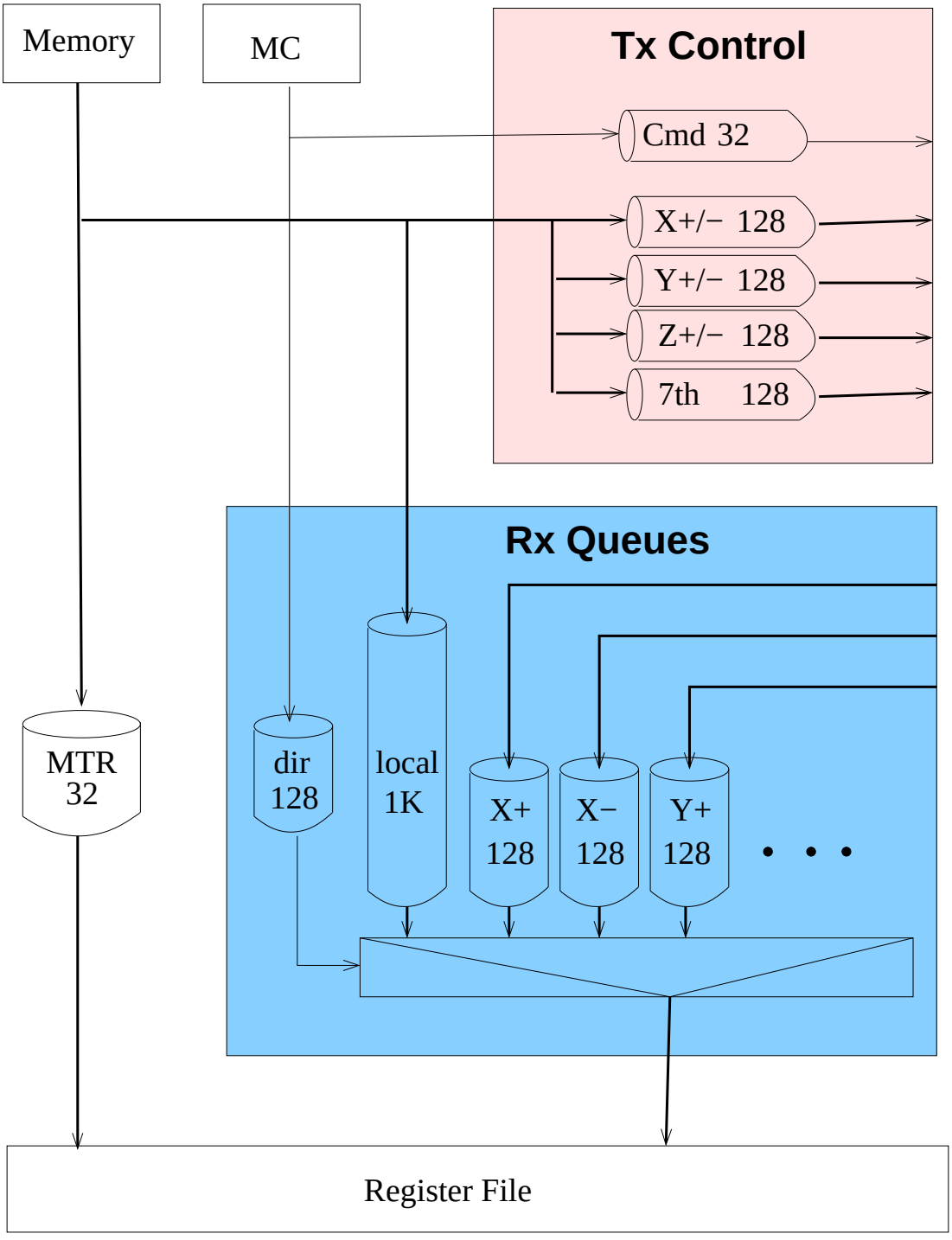
Control of Memory Accesses



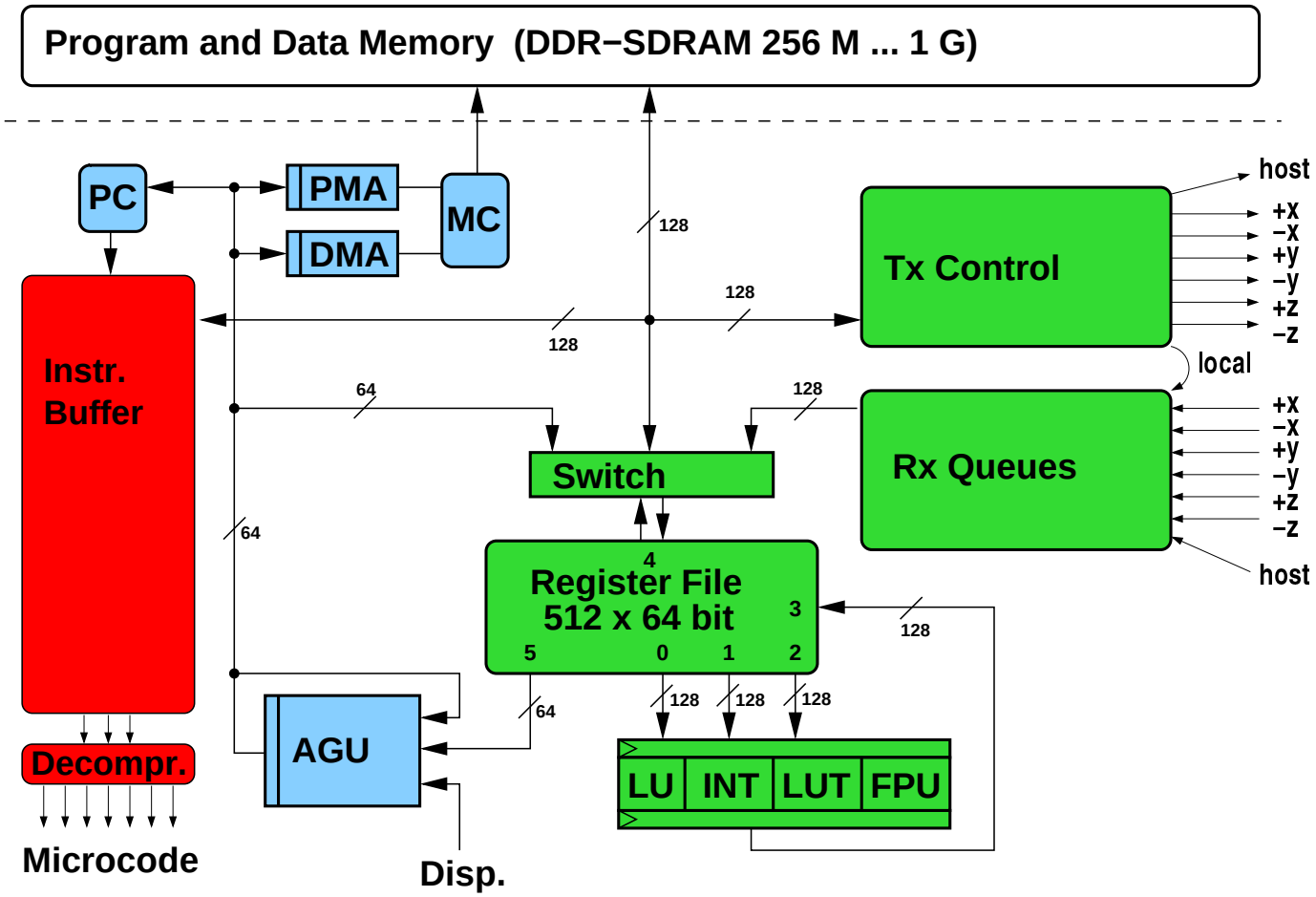
m

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

Data Queues


Memory
MC
Tx Control
Cmd 32
X+/− 128
Y+/− 128
Z+/− 128
7th 128
Rx Queues
MTR
32
dir
128
local
1K
X+
128
X−
128
Y+
128
Register File

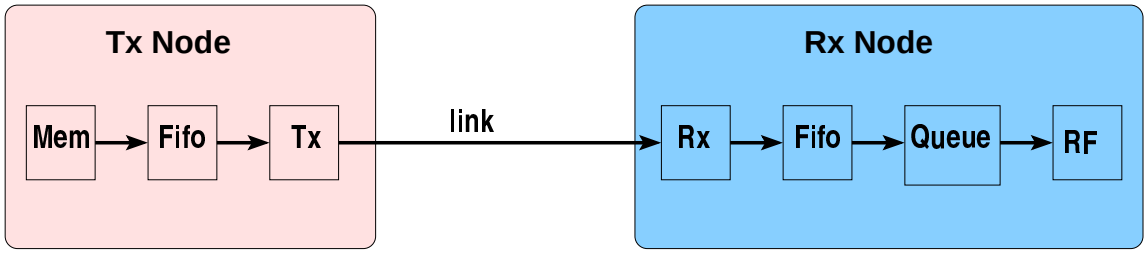
apeNEXT Processor Chip



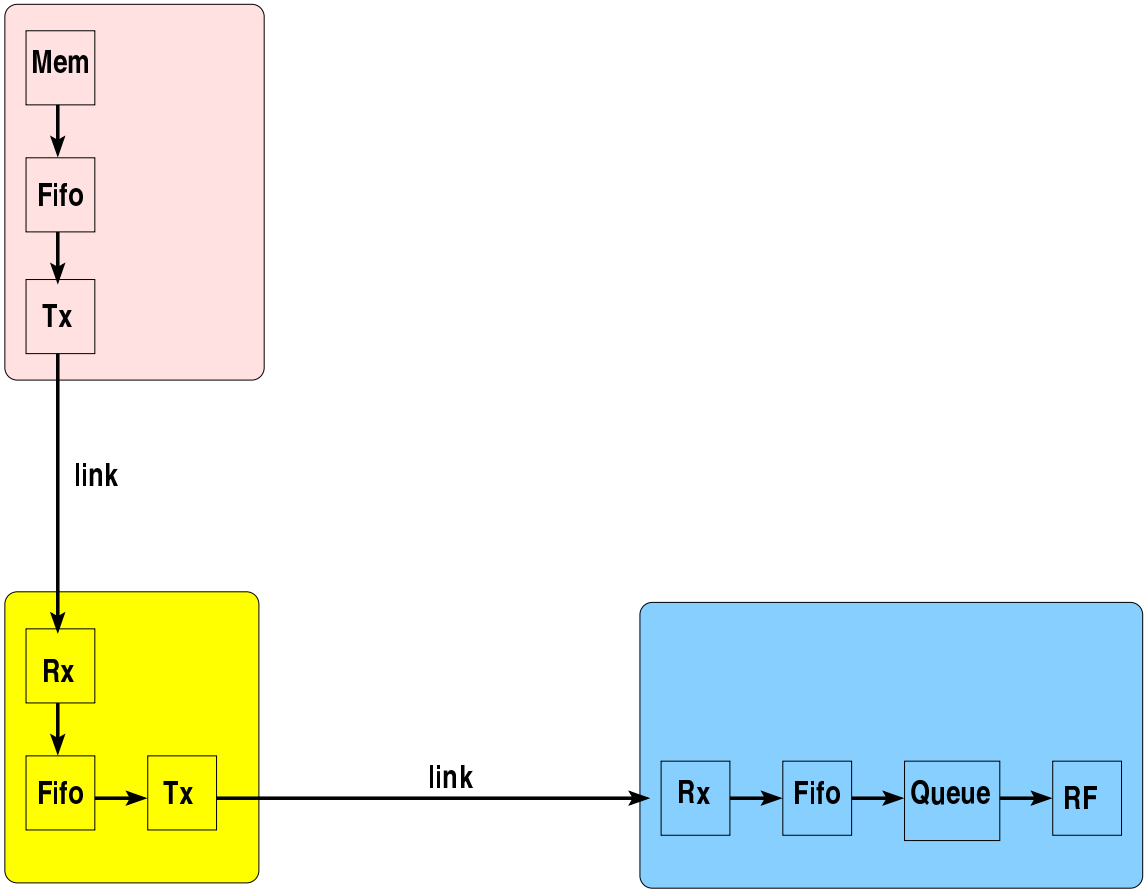
- 0.18 μm CMOS
- $9.2 \times 8.2 mm^2$, 1.2 M gates
- 200 MHz clock
- 2 (proc) + 5 (mem) Watt power consumption

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 hypercube)

apeNEXT Design Tests

Application codes:

- Dirac Operator (with/without SSOR)
- Gauge Update
- Linear Algebra (saxpy, Jacobi, ...)
- ...

HW specific tests:

- Communication Network (incl. errors and delays)
- Memory Controller
- Arithmetics
- Cach Handling
- Register File
- ...

Coverage:

- ca. 7'000 TAO or C lines
- ca. 700'000 assembly lines
- ca. 1'000'000 microcode lines

Logistics:

☞ ca. 11'000'000 clock cycles

☞ $>$ 500 h VHDL simulation time

apeNEXT Schedule

4/2001 Collaboration agreement INFN-DESY-Orsay

12/2001 Prototype of board and backplane delivered

1/2001 VHDL design of processor completed

?/2002 Chip sign-off

?/2002 Test of Board with 16 Processors

?/2002 Test of Prototype Crate (400 GFlops)

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. Tripiccione	VLSI
(G. Magazzu – 3/2002)	VLSI (FPU)
(W. Errico – 1/2002)	VLSI
(T. Giorgino – 9/2001)	Tests
Parma/Milano	Tests, Libraries

France:

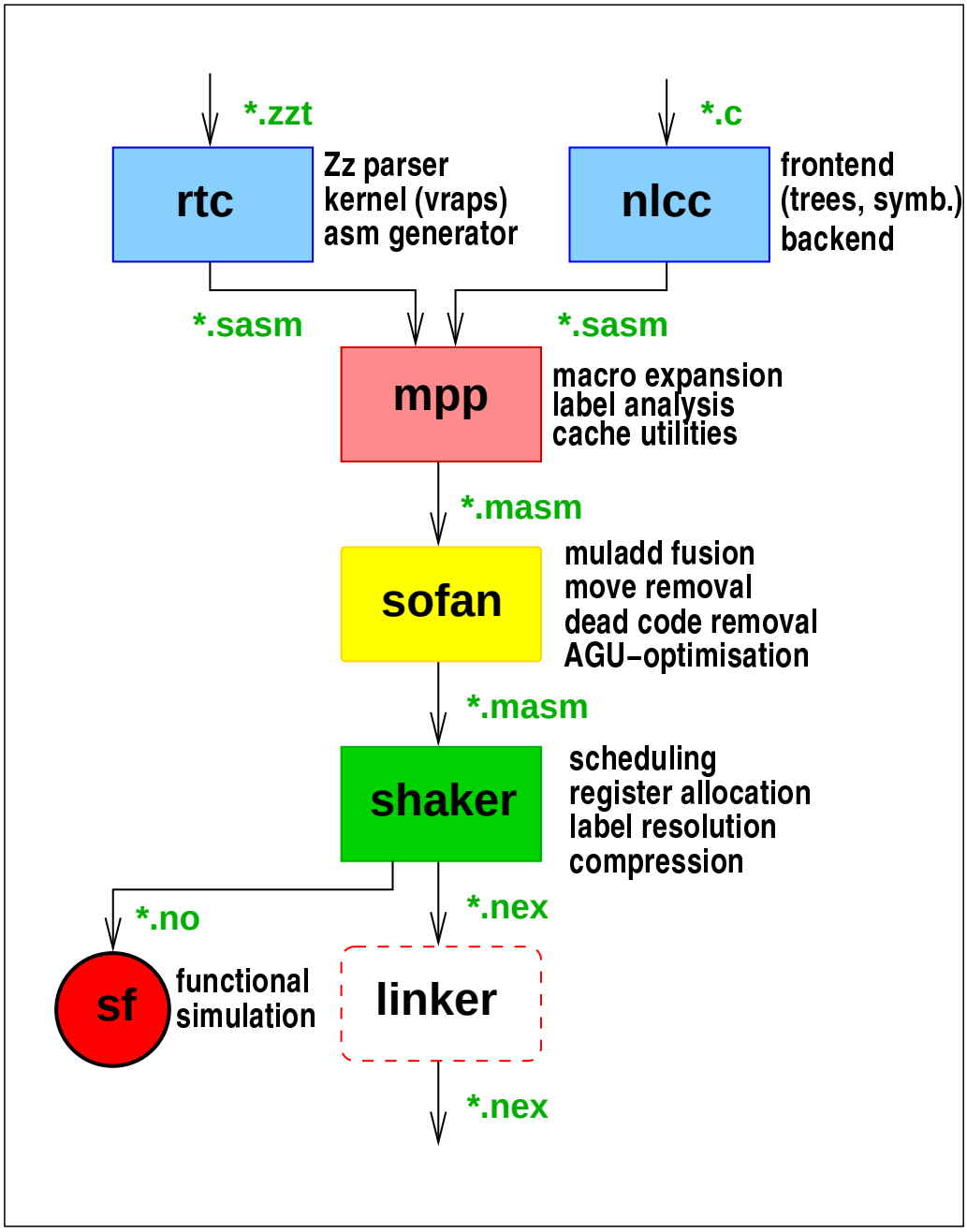
Orsay	Tests, Benchmarks, Libraries
Rennes	Assembly Optimiser

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

Compilation Chain



- stable and unified TAO and C environment
- improved low-level optimizations (assembly, microcode)

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

☞ Keep simple high-level structure as in APEmille

Final Challenge

Requirement of German LQCD community in 2003:

> $\approx$ 16 TFlops (peak)
> (= 20 apeNEXT Racks)

☞ apeNEXT is a suitable platform

☞ DESY-Zeuthen is an ideal site