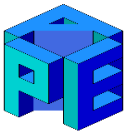
Das APEmille Projekt.

Stand und Perspektiven

H. Leich, P. Wegner

- **Einführung**
- **Übersicht über die APEmille Architektur**
- **Software**
- **Perspektiven**
- **Beiträge von DESY Zeuthen**



The APEmille Group

Roma

U. Alberti
A. Bartoloni
C. Battista
S. Cabasino
N. Cabibbo
M. Chiricozzi
M. Cosimi
P. De Riso
A. Lonardo
V. Mårdh
A. Michelotti
E. Panizzi
P. S.Paolucci
F. Rapuano
D. Rossetti
G. Sacco
M. Torelli
E. Valente
P. Vicini

Pisa

A. Cisternini
F. Laico
W.Errico
S. Giovannetti
G. Magazzu
R. Tripiccione [1])

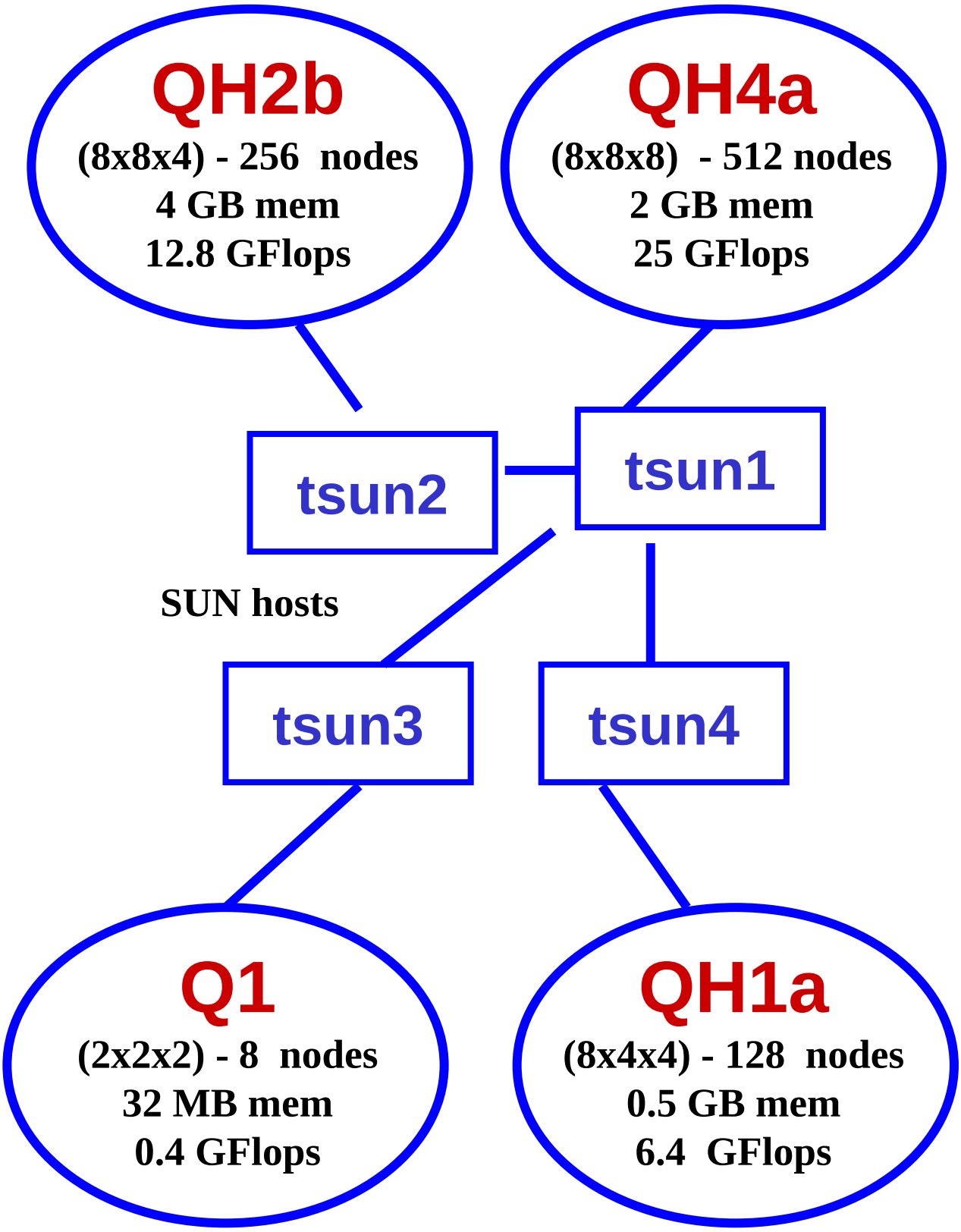
Zeuthen

U. Gensch [2])
H. Simma
W. Friebel
P. Wegner
H. Leich
U. Schwendicke
K. Sulanke
S. Menschikow

[1]) Projektverantw. INFN

[2]) Projektverantw. DESY

Quadrics Installation at DESY Zeuthen



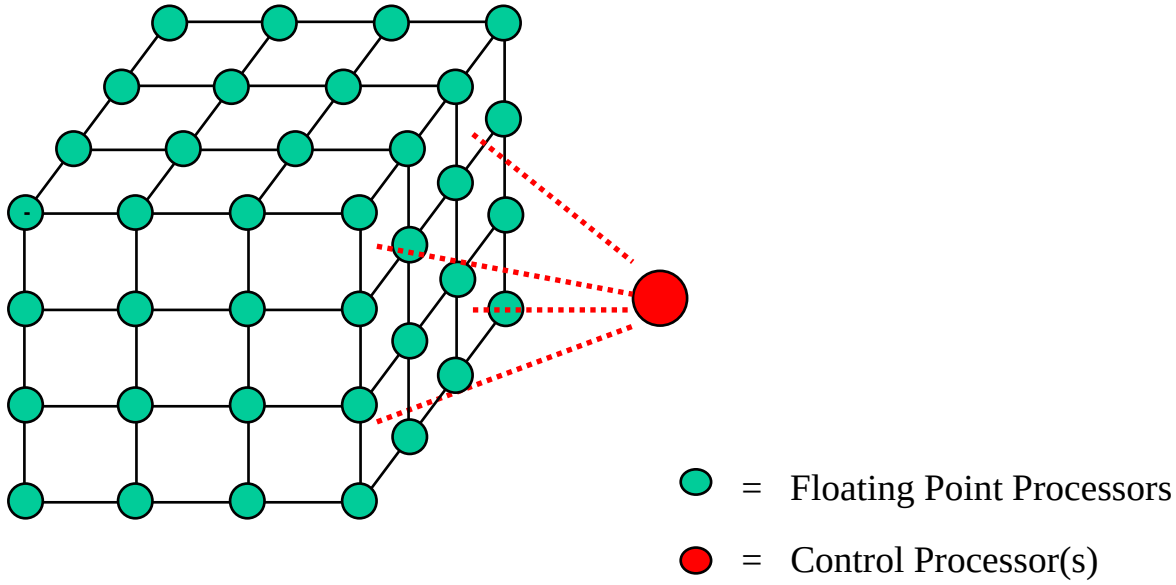
APE100 vs APEmille

	APE100	APEmille
Clock	25 MHz	66 MHz (100 MHz)
Performance (per FPU)	50 MFlops	528 MFlops
Register File (FPU)	128 x 32 Bit	256 x 32 Bit or 128 x 64 Bit
Prec. Of Arithm. Operations	single	single, double, complex, integer
Addressing	global (Z-CPU)	local (board controller)
Inter processor communication	nearest neighbour	nearest neighbour, broadcast
Communication bandwidth	12.5 Mbyte/s	200 Mbyte/s
Host I/O	Transputer links 500 kByte/s	PCI, 133 Mbyte/s

APEmille topology

•FPU‘s (8 per board)	528 MFlops
•Basic unit: board (2*2*2 FPU‘s)	4.22 GFlops
•Subcrate (2*2*8 FPU‘s)	16.9 GFlops
•Crate (2*8*8 FPU‘s)	67.5 GFlops
•Tower (4*8*8 FPU‘s)	135 GFlops
•APEmille (8 Tower 32*8*8)	1.081 TFlops

Architecture of APEmille

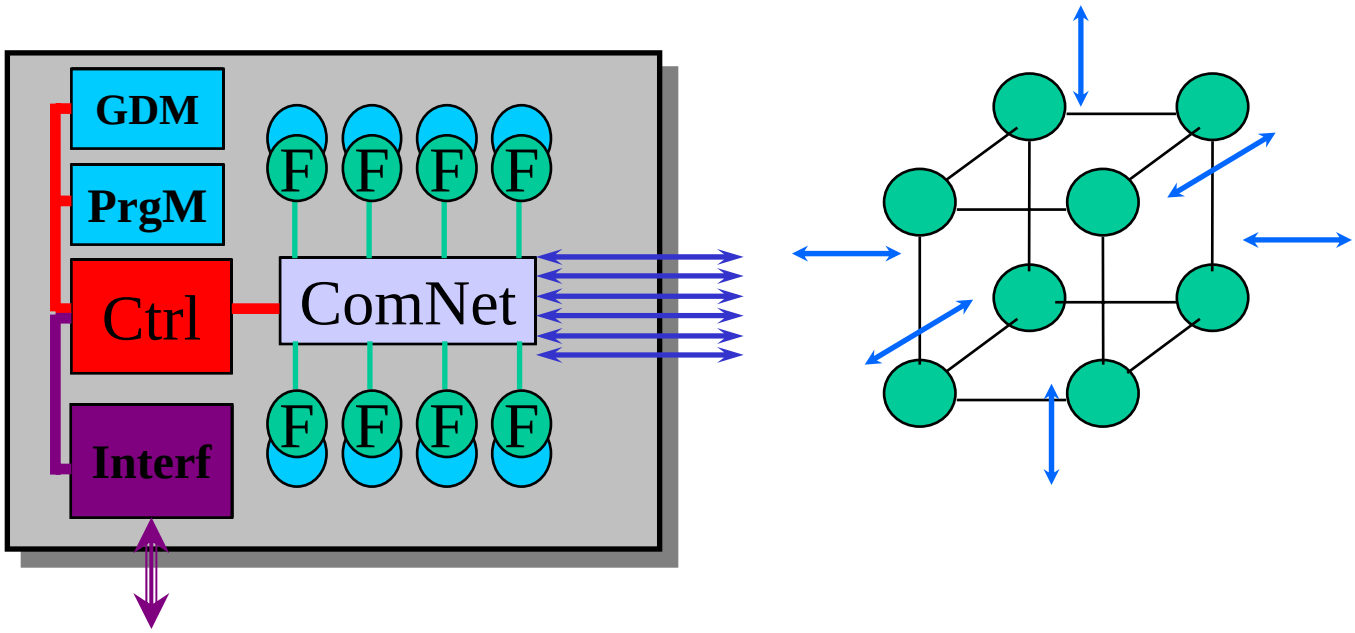


Characteristics:

- **3d Array of Floating Point processors**
  - large Reg. File with direct access to local and remote memory
  - normal operations a*b+c for single, double, complex, …
  - LUTs for inv, inv sqrt, exp, log …
- **SIMD with Local Addressing**
- **Very Long Instruction Word**
- **Flexible Communication Network**
  - arbitrary distances, automatic routing
- **Host = Cluster of Linux PC's**
- **Scalable and modular:** 4 Gflops - 1.1 TFlops

Processing boards

Board = 8 FP + Controller + Communication Network + Host Interface



Processors:

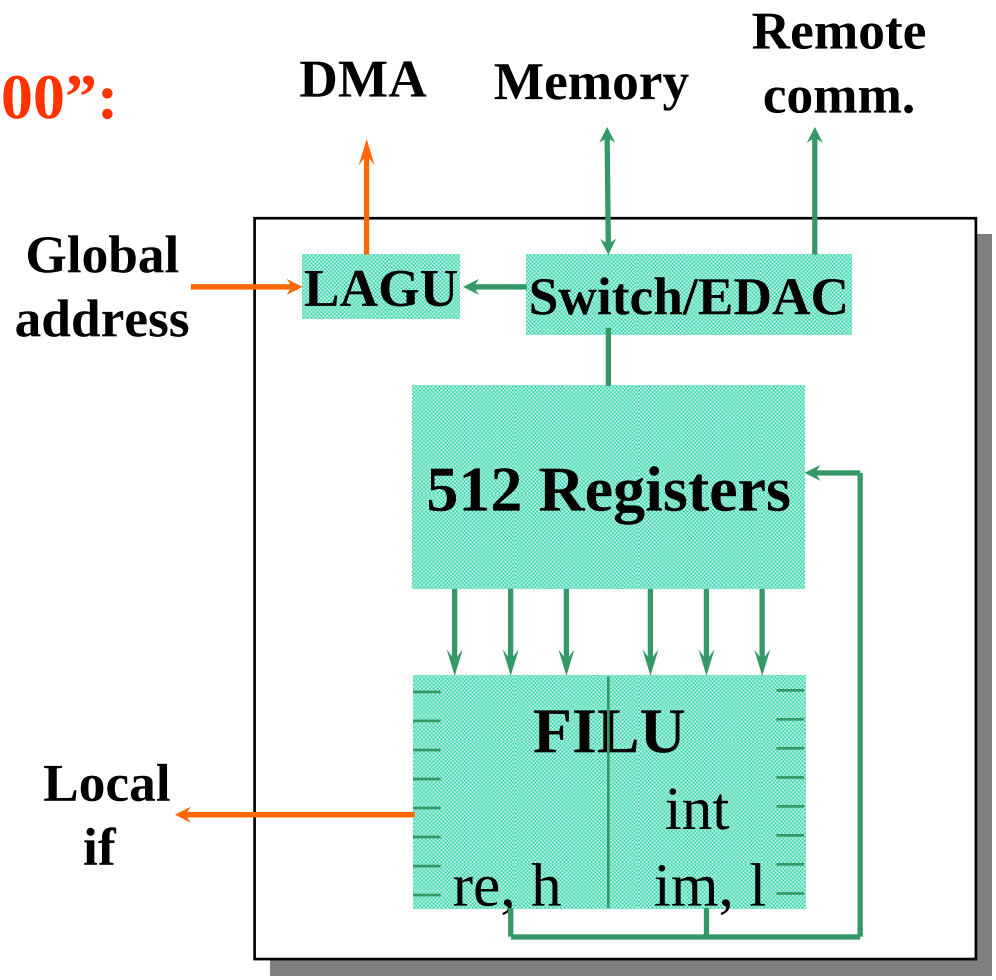
- ASIC design, 0.5 µ standard cell CMOS
- ca. 400 k gates per FP processor
- ca. 3 W power consumption per FP processor

- **Clock:** 66 MHz
- **Memory:**

- Program:	512 k instructions (96+80 bit)	synDRAM
- Controller data:	128 k words (32 bit)	SRAM
- FP local data:	2-8 M words (32 bit)	synDRAM

Processors

FP "J1000":

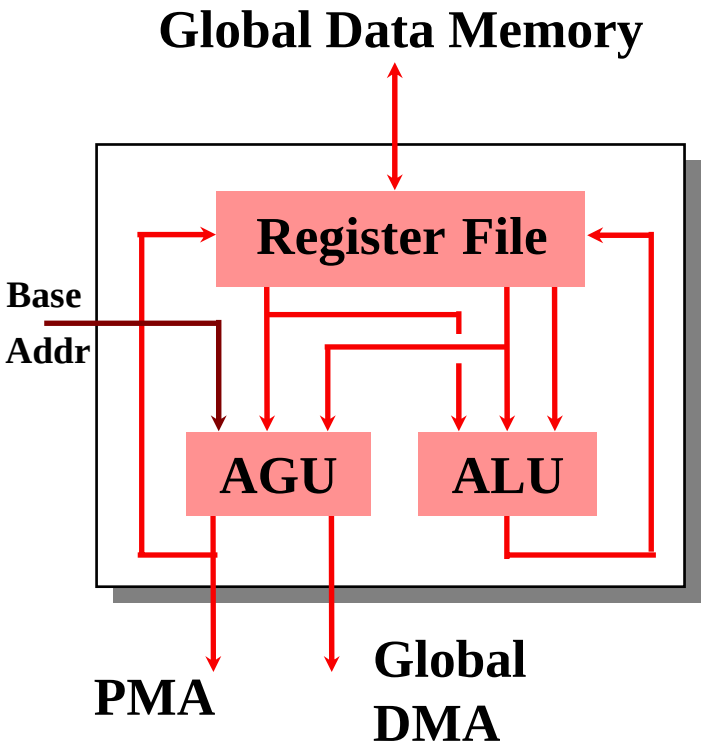


- **Normal operations: a*b+c**
  - single, double, complex, vector
  - IEEE with IBM precision enhancement
  - 1 pipelined normal operation per clock cycle
- **Local integers:**
  - bit-wise operations
  - independent local AGU
- **Large Register File:**
  - 512 words
  - independent memory access (1 per cycle)

Processors

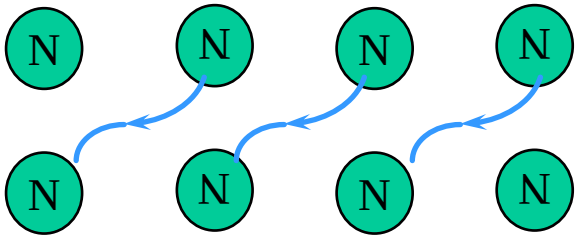
Control Processor “T1000”:

- ALU
- AGU
- 256 Registers
- efficient loop control
- page prediction



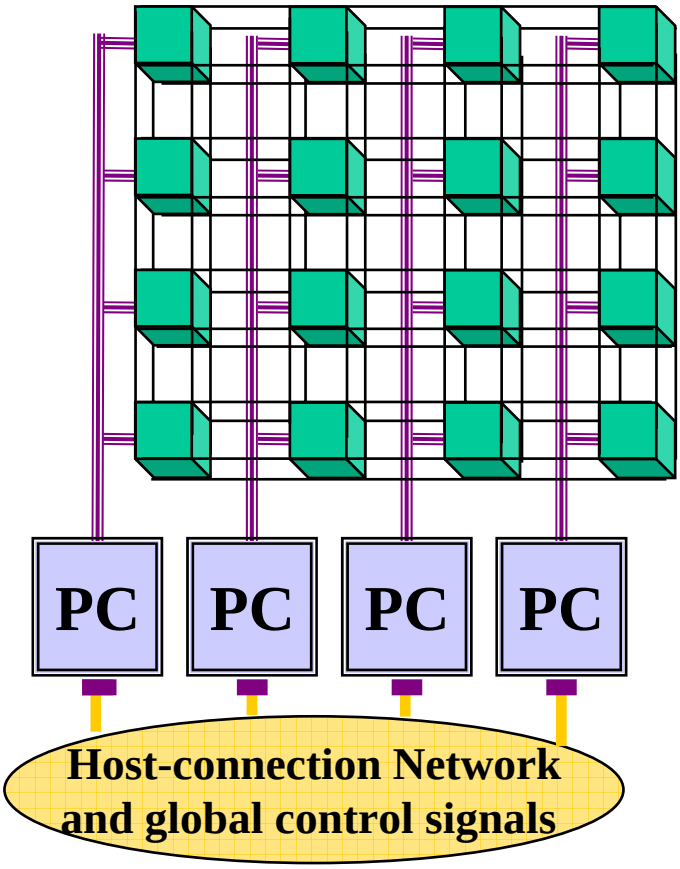
Communication Controller “COMM”:

- arbitrary homogeneous communications
  e.g.:



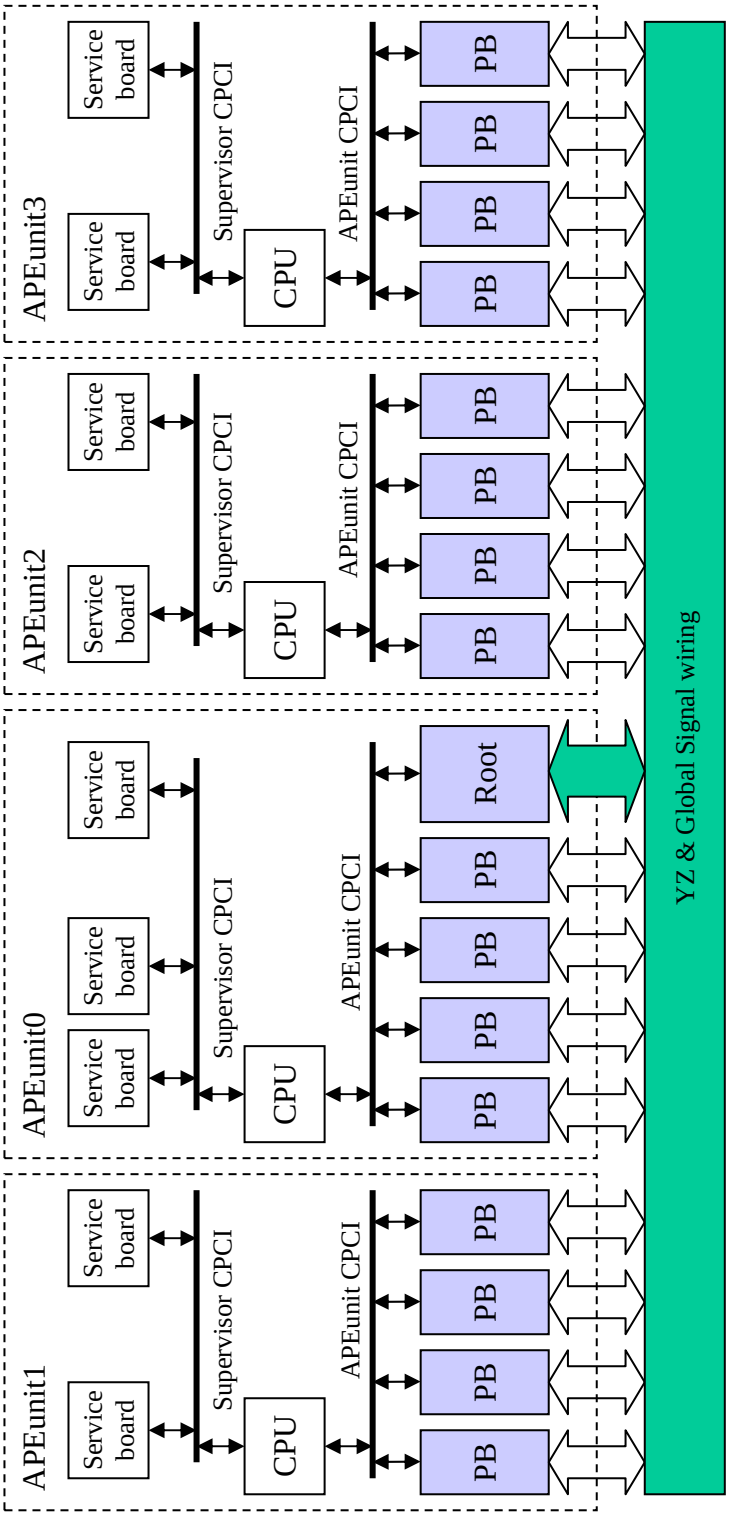
- automatic routing of multiple data paths
- 1, 2, and 3-dim broadcast
- next step: “soft communications”

Host Connection Network



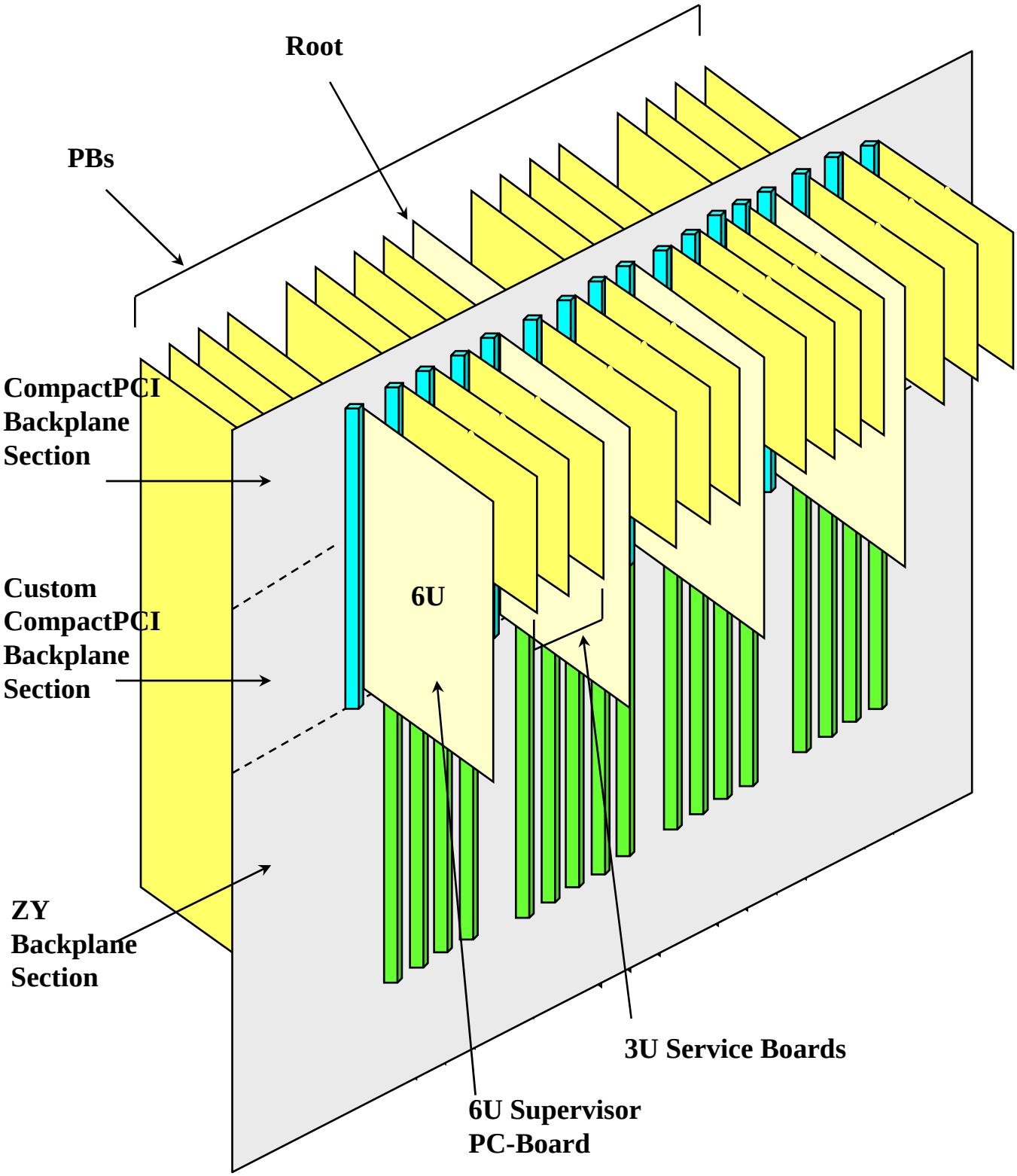
- **APE <--> Host connections:**
  - CompactPCI bus for a cluster of 4 boards:
    - 33 MHz / 32 Bit PCI Bus
    - 133 Mbytes/sec

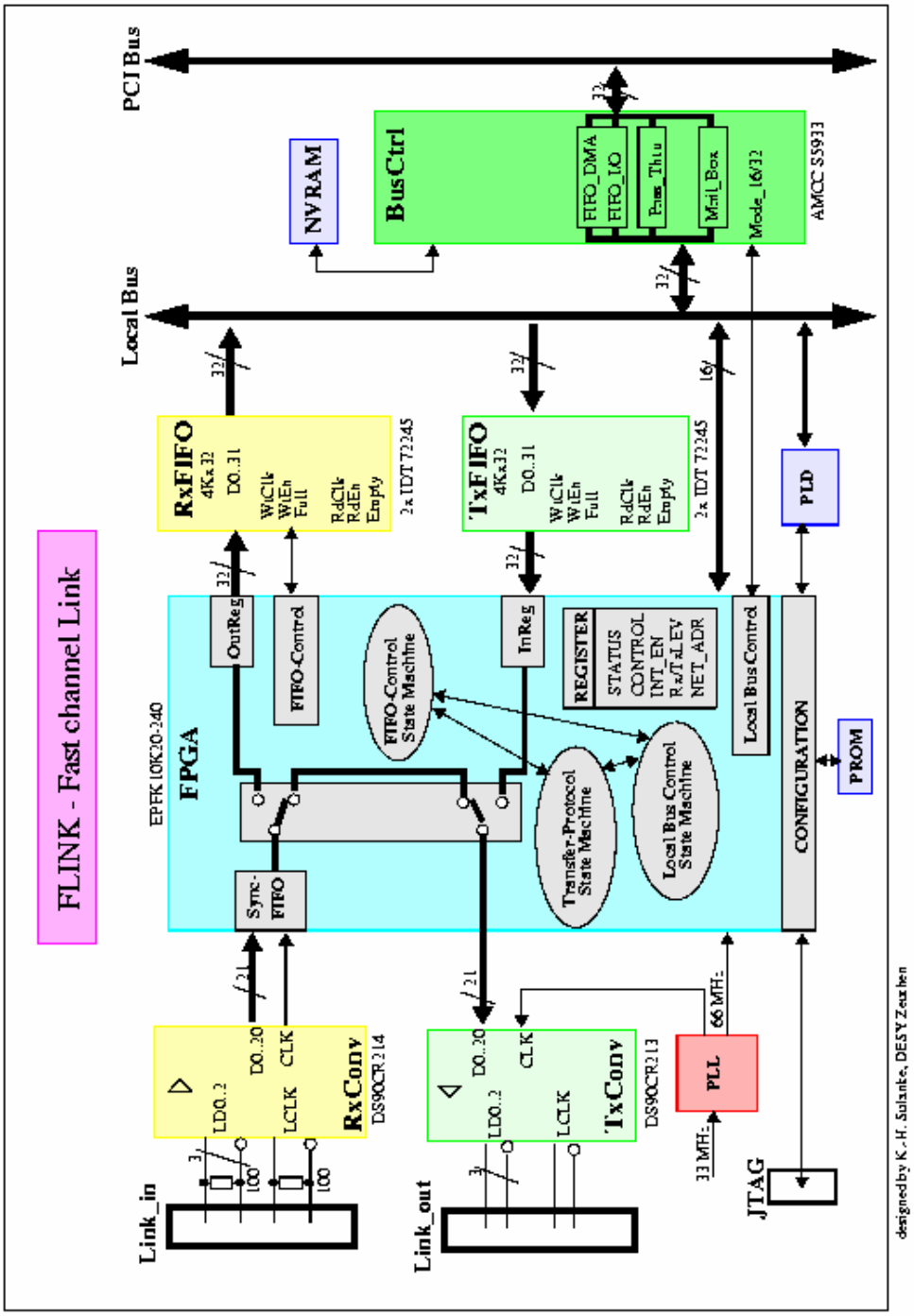
- **Host <--> Host connections:**
  - fast serial links (Channel-Link Technol.):
    - 132 Mbytes/sec
    - max. 10m distance



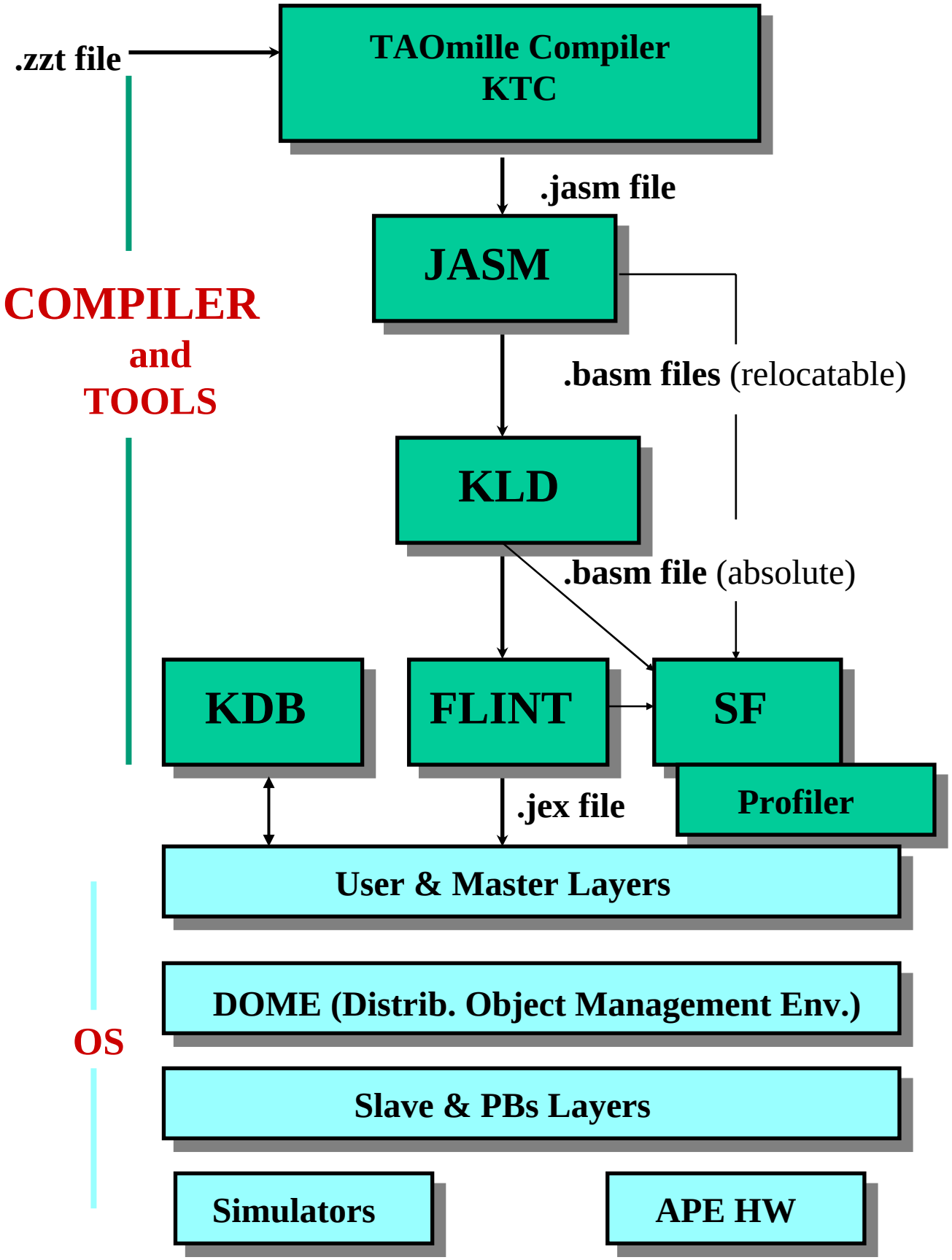
APEbackplane interconnectins.

Crate / Back-plane




FLINK - Fast channel Link
PCI Bus
Local Bus
NVRAM
BusCtrl
FIFO_DMA
FIFO_IO
Bus_Thru
Mail_Box
Mode_16/32
AMCC S5933
RxFIFO
4Kx32
D0..31
WrClk
WrEn
Full
RdClk
RdEn
Empty
2x IDT 72245
TxFIFO
4Kx32
D0..31
WrClk
WrEn
Full
RdClk
RdEn
Empty
2x IDT 72245
PLD
EPFK10K20-240
FPGA
OutReg
FIFO-Control
FIFO-Control State Machine
InReg
REGISTER
STATUS
CONTROL
INT_EN
RxTxLEV
NET_ADR
Local Bus Control
CONFIGURATION
PROM
Sync-FIFO
Transfer-Protocol State Machine
Local Bus Control State Machine
RxConv
DS90CR214
TxConv
DS90CR213
LD0..2
D0..20
CLK
LCLK
PLL
66 MHz
33 MHz
JTAG
Link_in
Link_out
designed by K.-H. Sulanke, DESY Zeuthen

APEmille software architecture



Compiler and Optimizer

- **TAOmille (ktc):**

- fully compatible with TAO
- new features,

  e.g.:
  - localint
  - broadcast, remote communications
  - type conversions
  - strings
  - functions

- **High-level optimizer:** (before Assembler generation)
  - common subexpression elimination
  - redundant load/store elimination
  - dead code removal
  - normal insertion
  - optimization of burst memory access
  - automatic loop unrolling

- **Assembler:**

- one-to-one correspondence to HW instructions
- direct cross-compilation from APE100 for tests

Performance Considerations

- **Performance enhancements:**
  - free temporization of Controller operations
  - shorter latencies for address calculations
  - efficient flow control instructions
  - Normals: Single Vector Complex Double
    Flop/cycle: 2 4 8 2

- **Algorithmic enhancements:**
  - local addressing: Preconditioning, FFT
  - communications beyond next neighbours
  - local remote-addressing only with “soft communications”

- **Performance estimates:**

	cycles	pipeline	perf.
Dirac operator (Snorm):	1080	88 %	22 %
Dirac operator (Cnorm):	760	88 %	60 %
Dirac operator (Dnorm):	1150	82 %	21 %
Clover matrix (Cnorm):	745	87 %	80 %

Operating System

- object oriented: C++ and RPC
- transparent communications
- support of multicast for host-host communication
- layered client/server

- **Master (=client):**
  - **ApeMaster:**
    Load, Init, Run, Stop, Restart, Step, ...
  - **MachineInterface:**
    SysServer, Topology, Resources, RootBoard, ...

- **Slave (=server):**
  - **ApeSlave:** analog to ApeMaster for single ApeUnit
  - **ApeInterface:**
    Read/WriteXxxMem, Wait, SlaveSysServer

- **Net Layer:**
  - **Stub generation for remote objects**
  - **Driver for Linux kernel**

Simulation and Tests

- **VHDL Simulator:**
  - incorporates design of all HW components
  - electrical simulation and timing estimates
- 5 Hz equivalent clock

- **C++ Simulator (“sf”):**
- full behavioral simulation
- freely configurable
- test of compiler and operating system SW
- generation of test vectors for VHDL simulator
- runs realistic portions of application code
- 100 Hz equivalent clock

Simulation and Tests (sqrt assembler)

! SQRT, Newton Raphson Method: $x_{k+1} = (½)x_k (3.0 – a\ x_k^2)$

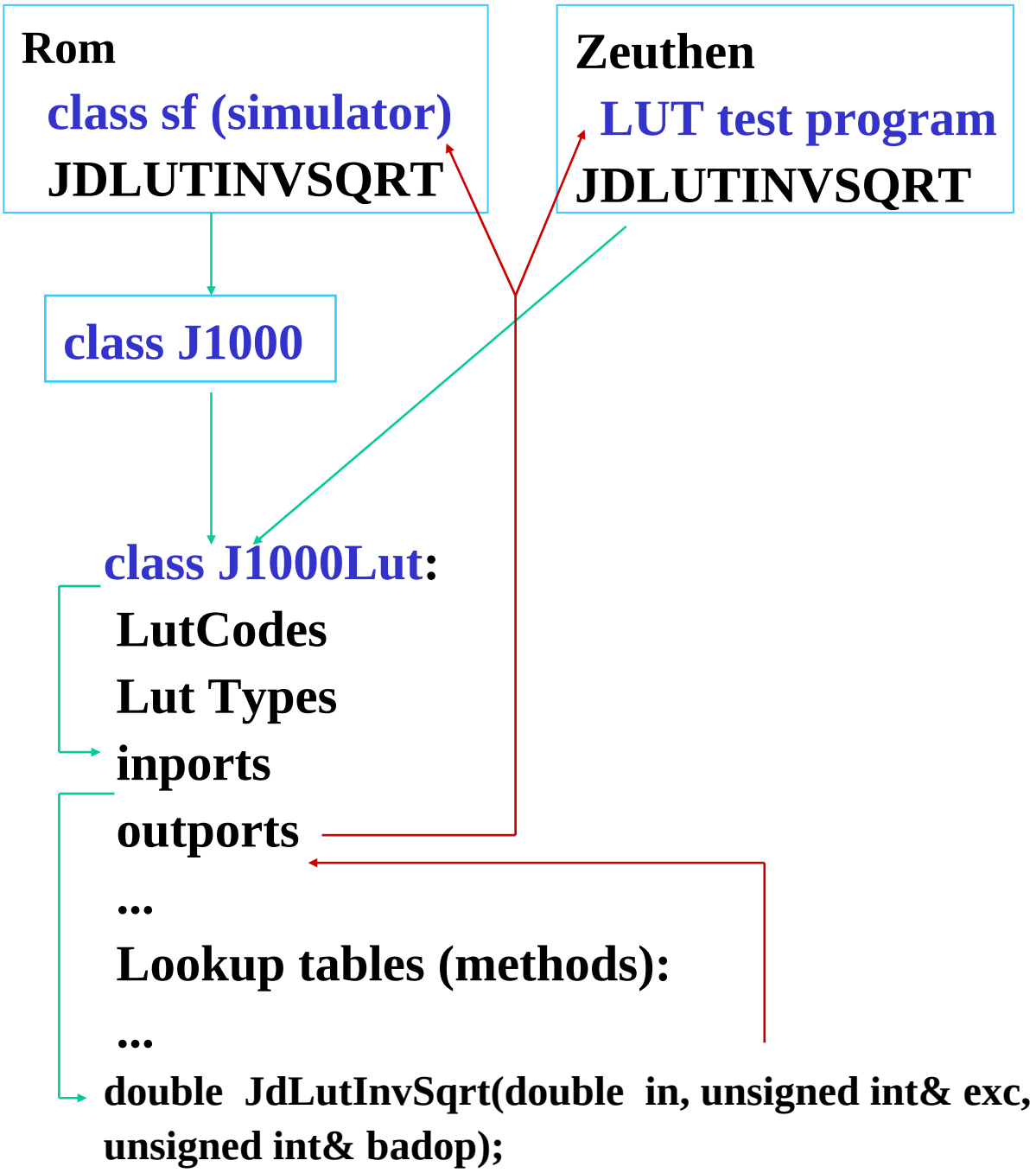
```
\include instr.exp
\include ksys.exp

$main
\equ R0 12
\equ R1 14
\equ R2 16
\equ R3 18
\equ R4 20
\equ R5 22
\equ R6 24
\equ R7 26
\equ R8 28
\equ R9 30
JCONSTD 100 {D0.0}
JCONSTD 102 {D2.0}
JCONSTD 104 {D0.5}
JCONSTD 106 {D1.5}

$puts("Input number:")
$accept(DEV_JANE_DATA,FMT_DOUBLE,1,102)
$print(DEV_JANE_DATA,FMTHEX,2,102)

! a --> R0
MEMTOJ1 R0 :2 102

! 0.0 --> R8
MEMTOJ1 R8 :2 100

! 0.5 --> R5
MEMTOJ1 R5 :2 104

! 1.5 --> R6
MEMTOJ1 R6 :2 106
```

Simulation and Tests (sqrt assembler,cont.)

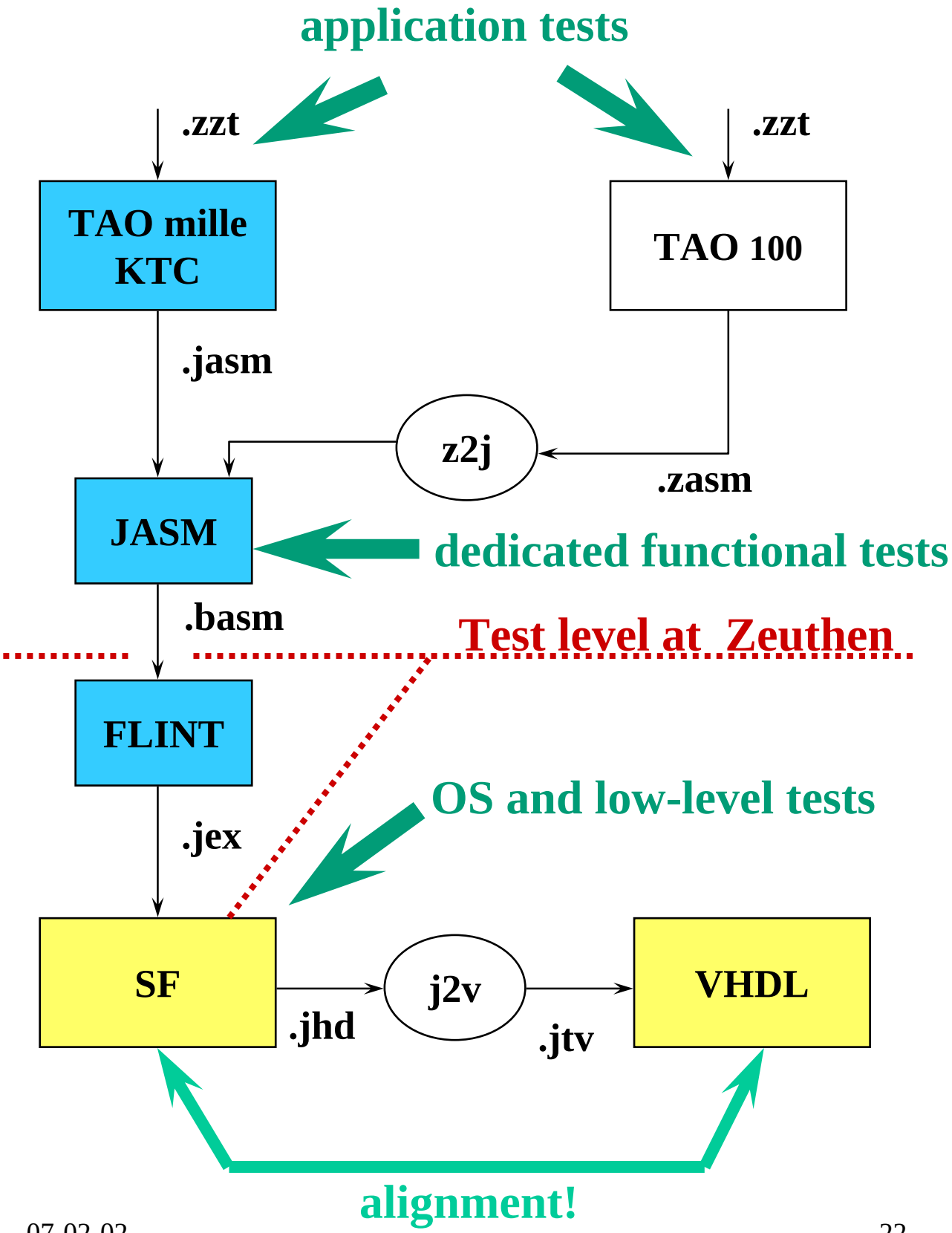
```
MEMTOJ1 R0 :2 102

! 0.0 --> R8
MEMTOJ1 R8 :2 100

! 0.5 --> R5
MEMTOJ1 R5 :2 104

! 1.5 --> R6
MEMTOJ1 R6 :2 106

! a/2 --> R1  ----------
JDNORM_PP R1 R0 R5 R8

! seed --> R2
JDTOS R9 R0
JSLUTINVSQRT R2 R9
JSTOD R2 R2

JTOMEM1 400 R2 :2
$puts("seed:")
$print(DEV_JANE_DATA,FMT_DOUBLE,1,400)

! seed*seed --> R3

\dup i = 1 to 5
  JDNORM_PP R3 R2 R2 R8
  JDNORM_MM R4 R3 R1 R6
  JDNORM_PP R2 R2 R4 R8
  JDNORM_PP R9 R2 R0 R8
  JTOMEM1 400 R9 :2
  $print(DEV_JANE_DATA,FMT_DOUBLE,1,400)
\enddup
$puts("End of 5th iteration...")

$stop(HS_SUCCESS)
```

Software Tests in Zeuthen

Lookup table test (C++ environment)



Test Chains


application tests
.zzt
.zzt
TAO mille KTC
TAO 100
.jasm
z2j
.zasm
JASM
dedicated functional tests
.basm
Test level at Zeuthen
FLINT
OS and low-level tests
.jex
SF
j2v
VHDL
.jhd
.jtv
alignment!

Status, Milestones

- **1999**

January : **T1000 + COMM delivered**
First T1000 tests in system + run mode successful, Random Bit Generator (ranbit) works
March : **T1000 chip tested**
April : **Flink board ready**
June : **J1000, COMM tested**
June : **Processor Board ready**

October : **APE unit ready**

- **2000**

January : **APE crate ready = 64 Gflops installation**
... September : **Mass production**

- **2001**

256 Gflops installation, (2 towers = 4 crates)

DESY Zeuthen Activities

• **Host communication**

PCI flat link interconnection board (130 Mbyte/sec) with hardware implemented token ring protocol (K.-H. Sulanke)

• **Hardware tests and developments**

Contribution to board- and backplane tests (S. Menschikow, H. Leich, U. Schwendicke)

• **Software**

Test with physics programs, assembler tests, simulator tests, general test strategies (H. Simma)

Generation of test vectors, LUT tests, Compiler tests (H. Simma, P. Wegner, W. Friebel)

Tao programs for compiler tests (R. Sommer, A. Hoferichter, F. Neugebauer, J. Heitger)

Future QCD machine

- **Goal**

10 - 20 integrated TFlops in 2003

- **Architectur options**

APE-like full custom solution with architectural improvements

Integration of DSP's on custom boards and communication network

Commercial PC's with fast communication network

- **Activities**

Benchmarking and performance evaluation (a Dirac operator benchmark for LQCD on a single PC)

Technology studies for remote communications

Mechanisms for "hiding" remote communications

Topology optimization

Study of the impact of program cache (in case of APE-like solutions)

Proposal in summer 1999