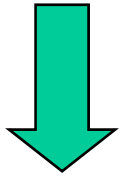
GAPE-DR Project: a combination of peta-scale computing and high-speed networking

Kei Hiraki

Department of Computer Science
The University of Tokyo

Computing System for real Scientists

- **Fast Computation at low cost**
  - High-speed CPU, huge memory, good graphics
- **Very fast data movements**
  - High-speed network, Global file system, Replication facilities
- **Transparency to local computation**
  - No complex middleware, or no modification to existing software



- **Real Scientists are not computer scientists**
- **Computer scientists aren't work forces for real scientists**

Our goal

- HPC system
  - as an infrastructure of scientific research
  - Simulation, data intensive computation, searching and data mining
- Tools for real scientists
  - High-speed, low-cost, and easy to use

    ⇒ Today's supercomputer cannot

    - High-speed/Low cost computation
    - High-speed global networking
    - OS/ ALL IP network technology

    ] GRAPE-DR project

GRAPE-DR

- GRAPE-DR:Very high-speed attached processor
  - 2Pflops(2008) Possibly fastest supercomputer
  - Successor of Grape-6 astronomical simulator
- 2PFLOPS on 512 node cluster system
  - 512 Gflops / chip
  - 2 M processor / system
- Semi-general-purpose processing
  - N-body simulation, gridless fluid dyinamics
  - Linear solver, molecular dynamics
  - High-order database searching

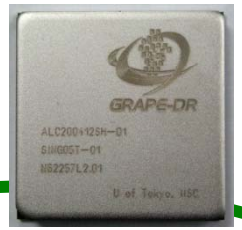
Data Reservoir

- Sharing Scientific Data between distant research institutes
  - Physics, astronomy, earth science, simulation data
- Very High-speed single file transfer on Long Fat pipe Network
  - > 10 Gbps, > 20,000 Km, > 400ms RTT
- High utilization of available bandwidth
  - Transferred file data rate > 90% of available bandwidth
    - Including header overheads, initial negotiation overheads
- OS and File system transparency
  - Storage level data sharing (high speed iSCSI protocol on stock TCP)
  - Fast single file transfer

Our Next generation Supercomputing system

- Very high-performance computer
- Supercomputing/networking system for non CS area

1 chip ---512Gflops,
1 system -- 2Pflops



GRAPE-DR processor
- very fast (512 Gflops/chip)
- low power (8.5Gflops/W)
- low cost

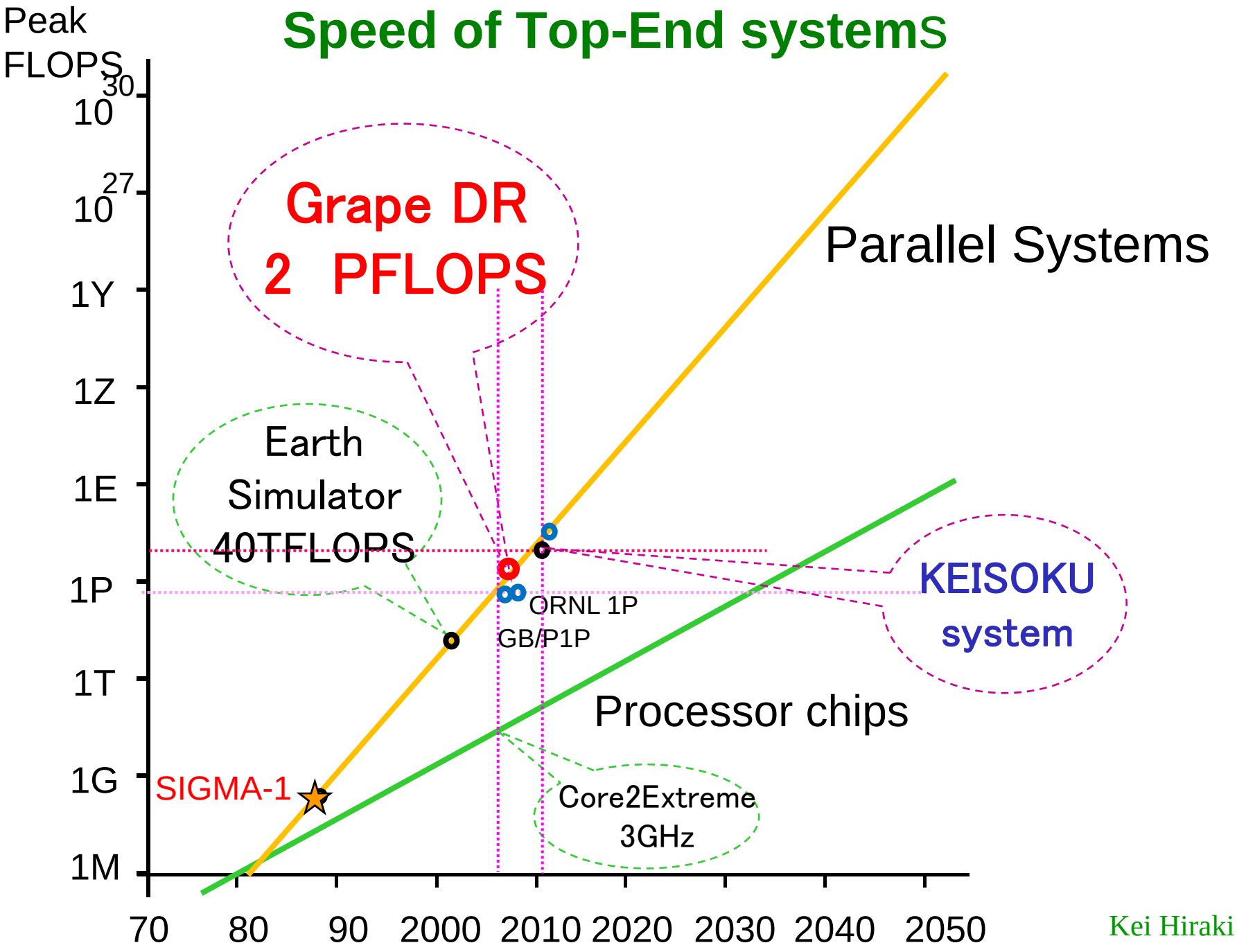
Next generation Supercomputing system

All IP computer architecture
- All components of computers Have own IP, and connect freely
- Processor, memory, Display, keyboard, mouse And disks etc.

(Keio Univ. Univ. of Tokyo)
University of Tokyo

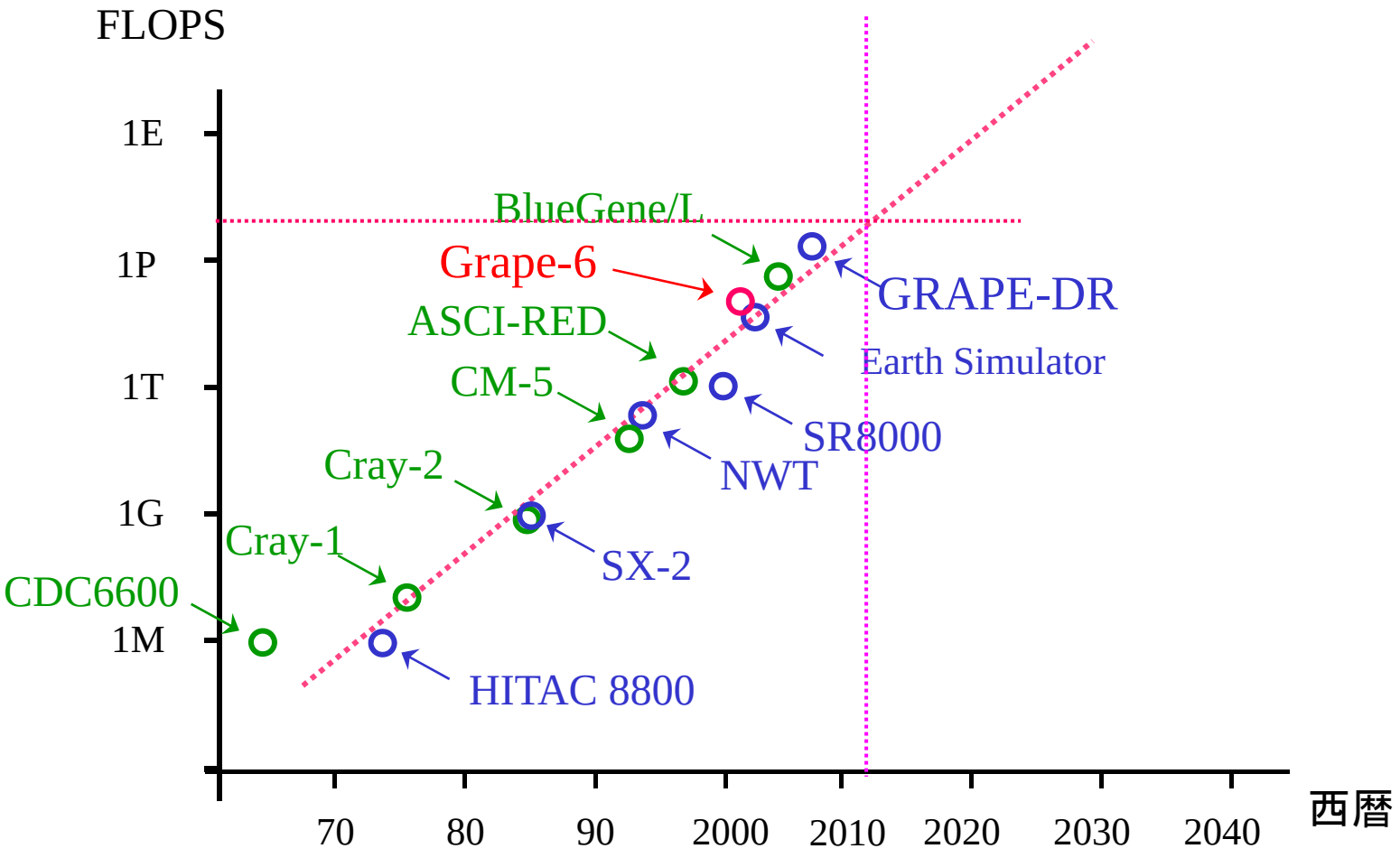
Long distance TCP technology
- same performance From local to global communication
- 30000Km,1.1GB/s

Kei Hiraki

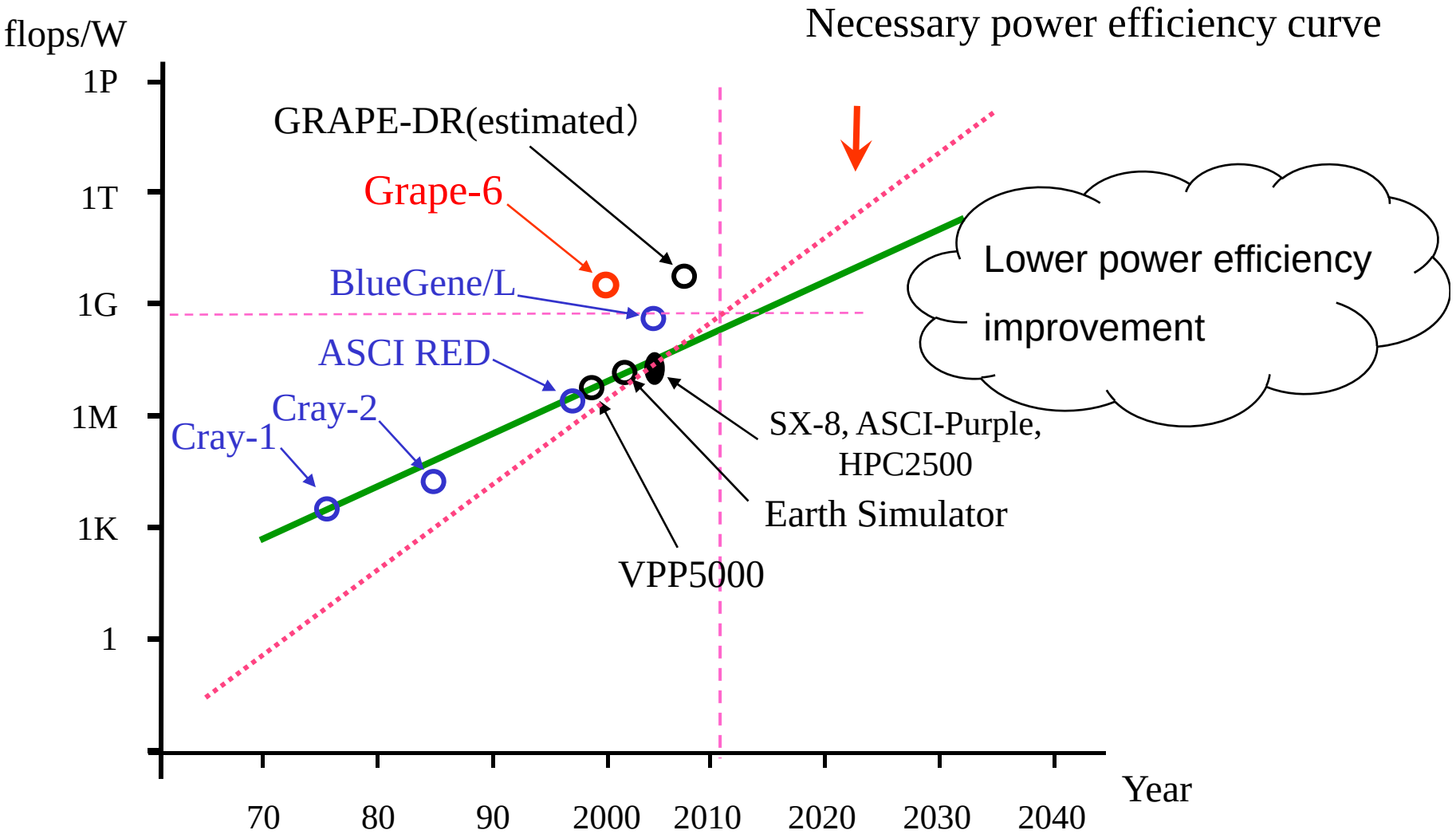

Speed of Top-End systemS
Peak FLOPS
10^30
10^27
1Y
1Z
1E
1P
1T
1G
1M
70
80
90
2000
2010
2020
2030
2040
2050
Grape DR
2 PFLOPS
Parallel Systems
Earth Simulator
40TFLOPS
KEISOKU system
ORNL 1P
GB/P1P
Processor chips
SIGMA-1
Core2Extreme
3GHz

Kei Hiraki

History of speedup（some important systems）



History of power efficiency



GRAPE-DR project (2004 to 2008)

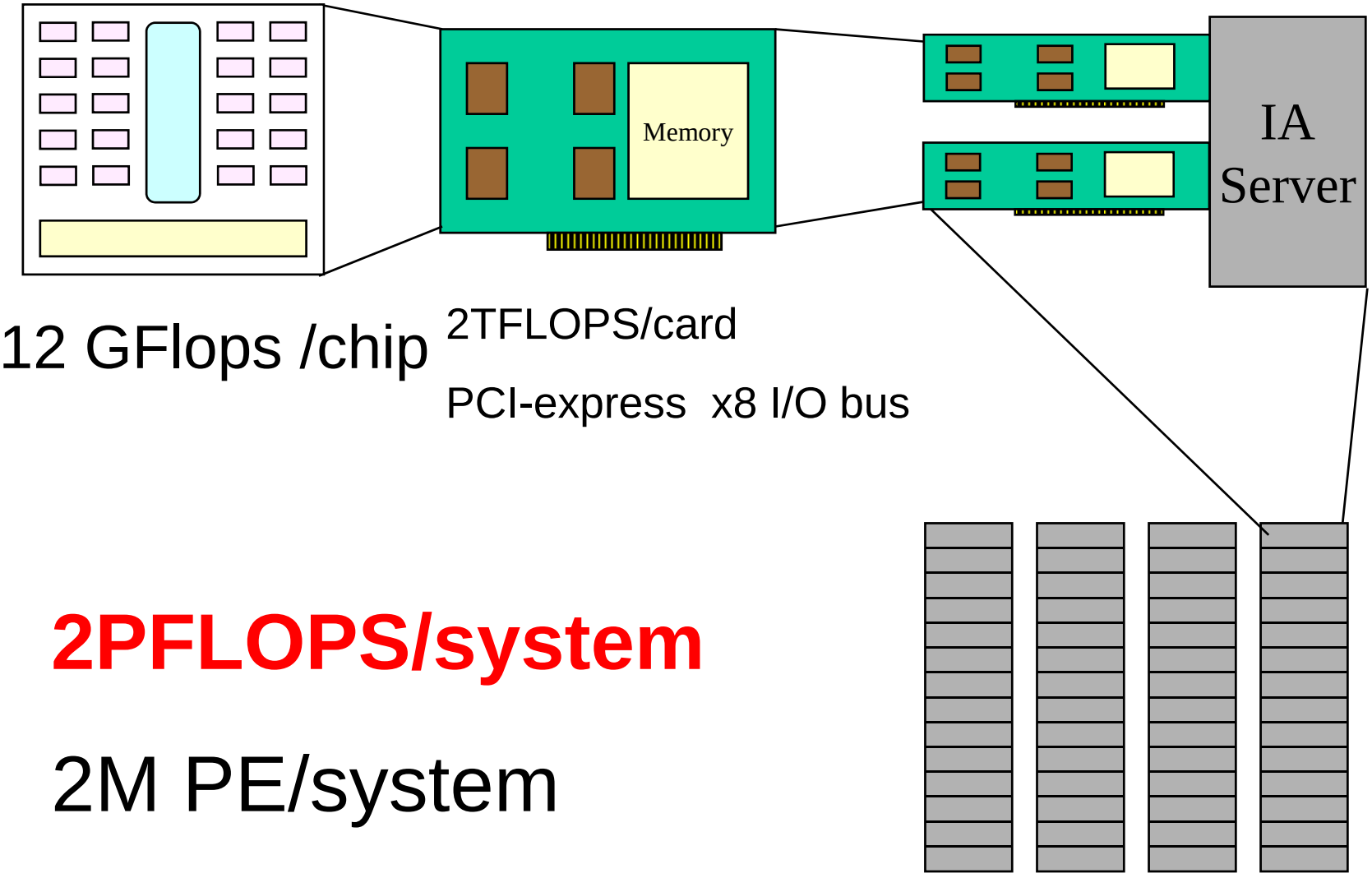
- Development of practical MPP system
  - Pflops systems for real scientists
- Sub projects
  - Processor system
  - System software
  - Application software
    - Simulation of stars and galaxies、CFD、MD, Linear system
- 2-3 Pflops peak、20 to 30 racks
- 0.7 – 1Pflops Linpack
- 4000 – 6000 GRAPE-DR chips
- 512 – 768 servers ＋ Interconnect
- 500Kw/Pflops

GRAPE-DR system

- 2 Pflops peak, 40 racks
- 512 servers ＋ Infiniband
- 4000 GRAPE-DR chips
- 0.5MW/Pflops(peak)
- World fastest computer?

  - GRAPE-DR (2008)
  - IBM　BlueGene/P(2008?）
  - Cray Baker（ORNL）（2008?）

GRAPE-DR chip



512 GFlops /chip

2TFLOPS/card

PCI-express x8 I/O bus

2PFLOPS/system

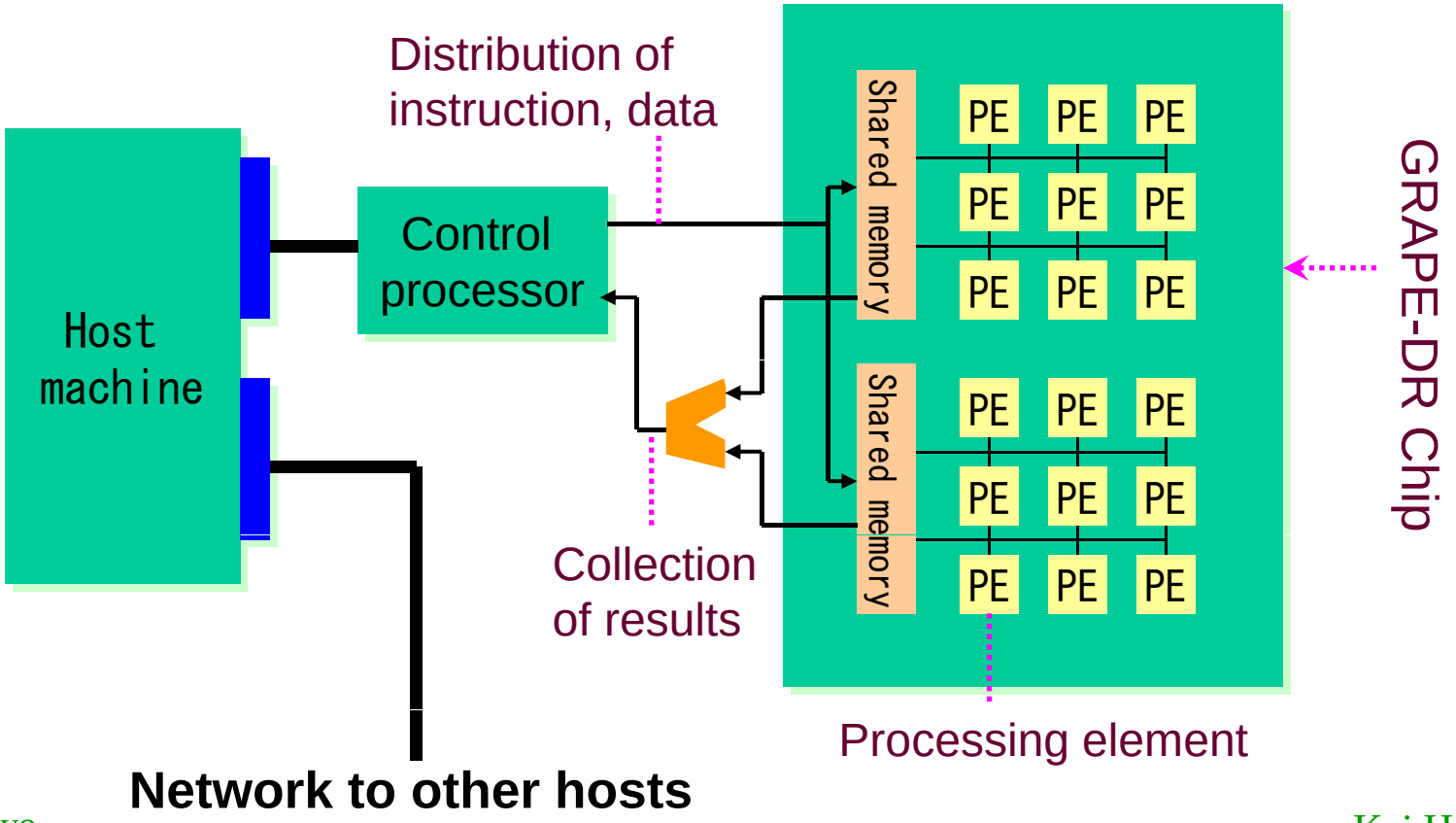
2M PE/system

20Gbps Infiniband

Kei Hiraki

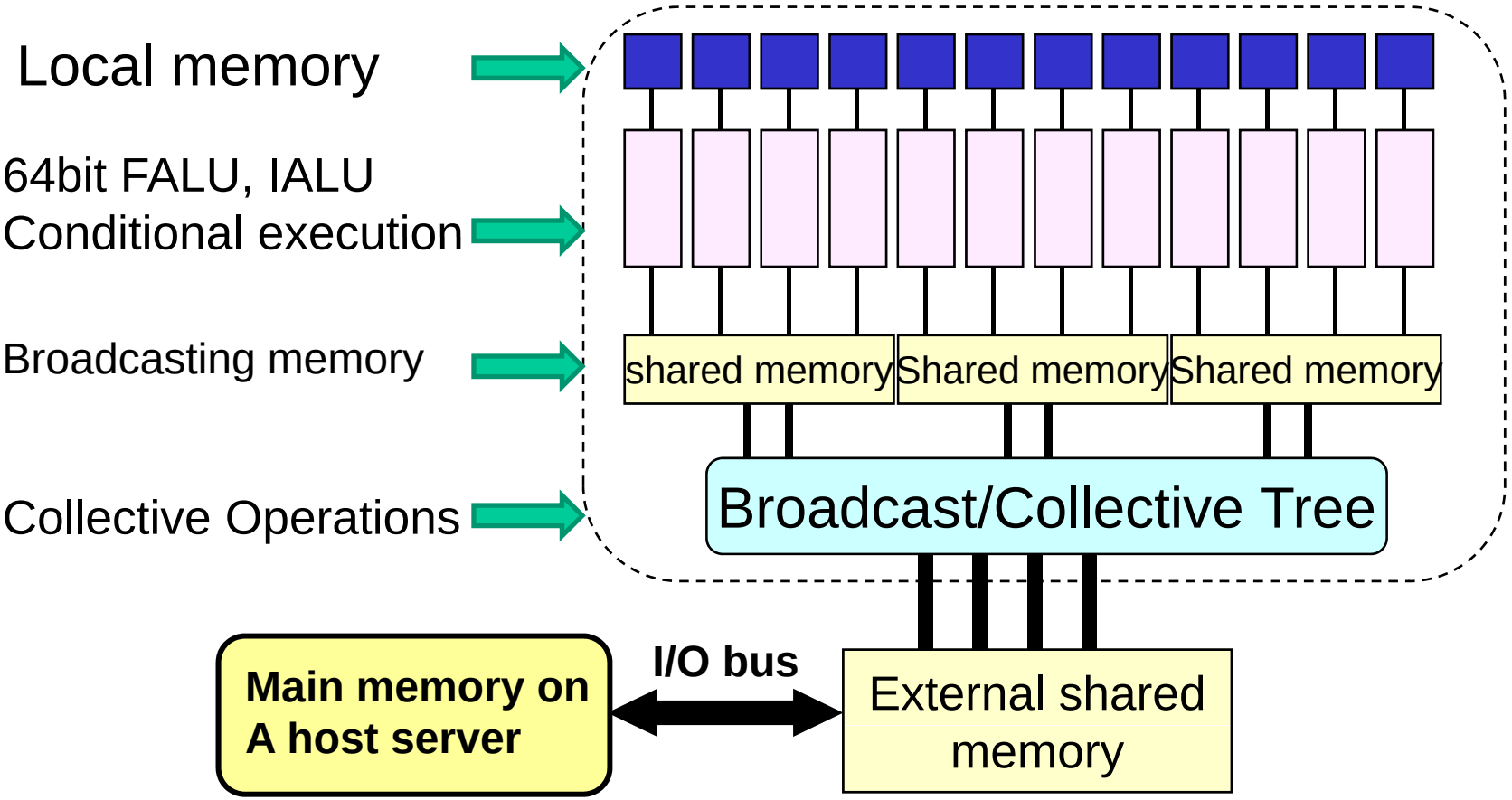
Node Architecture

- Accelerator attached to a host machine via I/O bus
- Instruction execution between shared memory and local data
- All the PE in a chip execute the same instruction (SIMD)



GRAPE-DR Processor chip

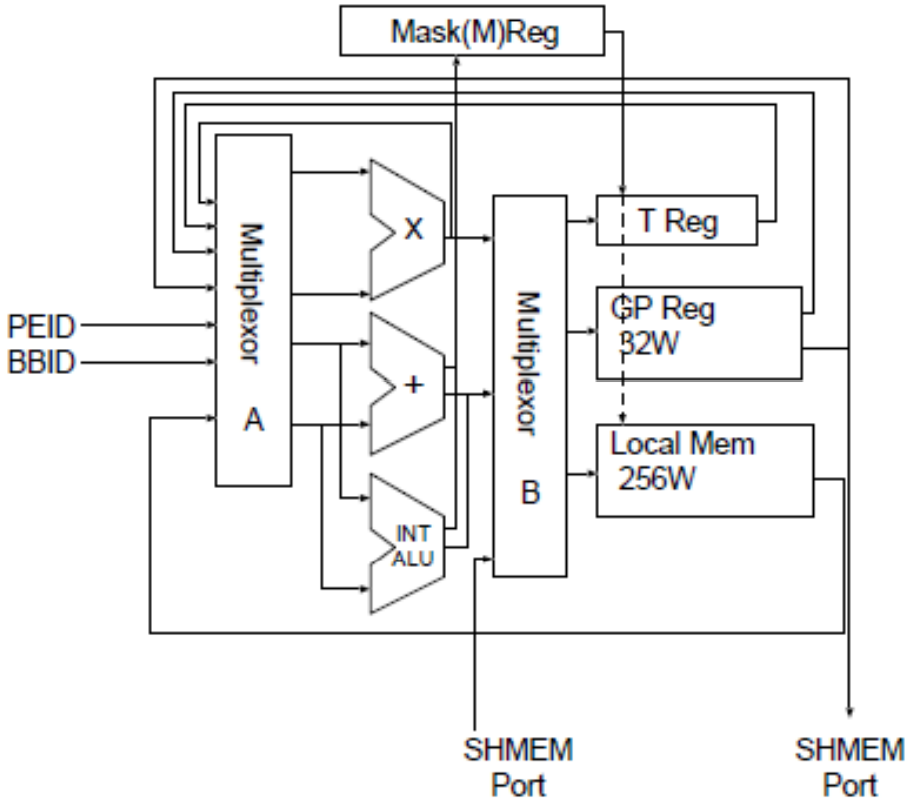
- SIMD architecture
- 512PEs
- No interconnections between PEs



Kei Hiraki

Processing Element

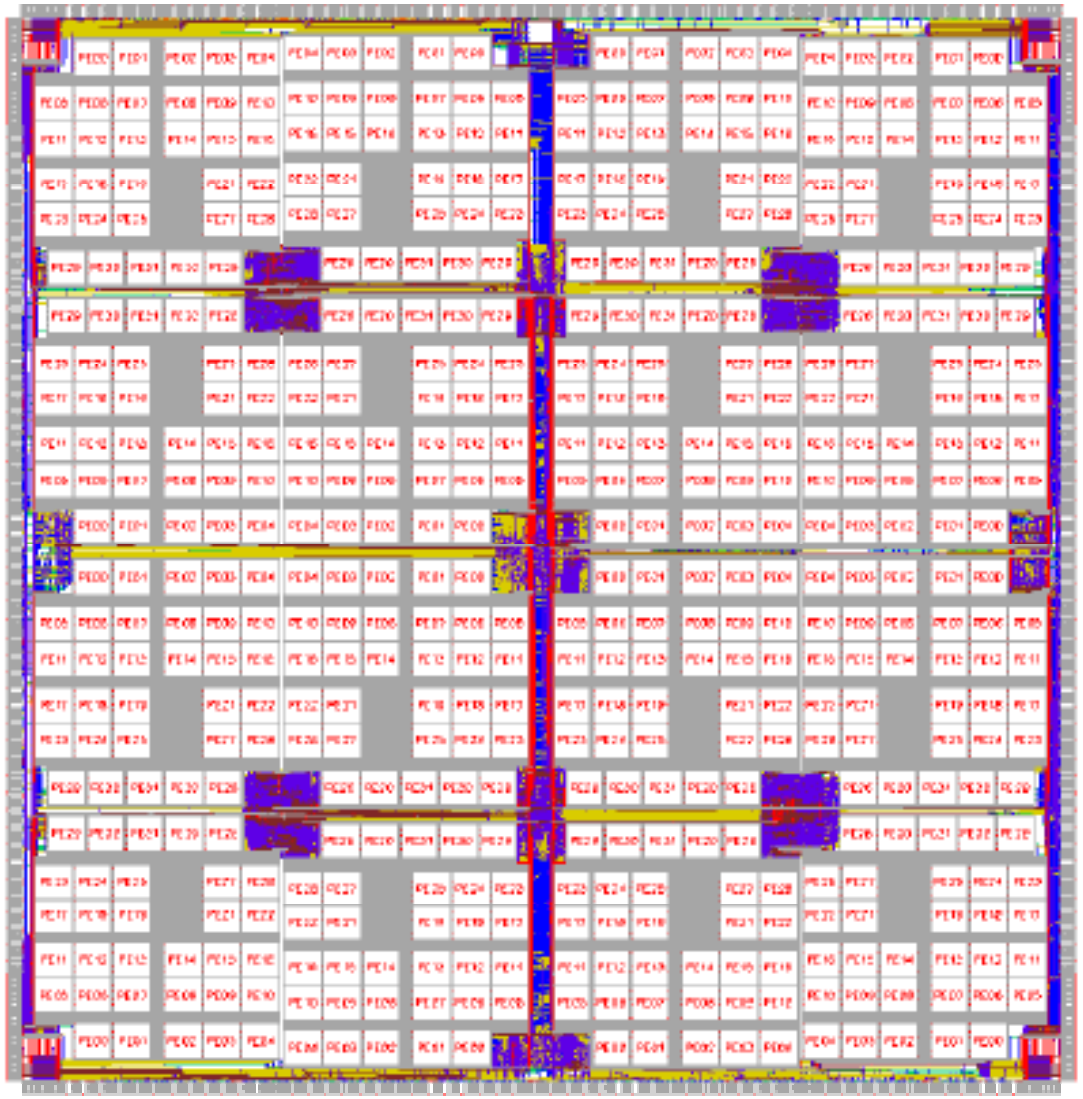
- 512 PE in a chip



- Float Mult
- Float add/sub
- Integer ALU
- 32-word registers
- 256-word memory
- communication port

- 90nm CMOS
- 18mm x18mm
- 400M Tr
- BGA 725

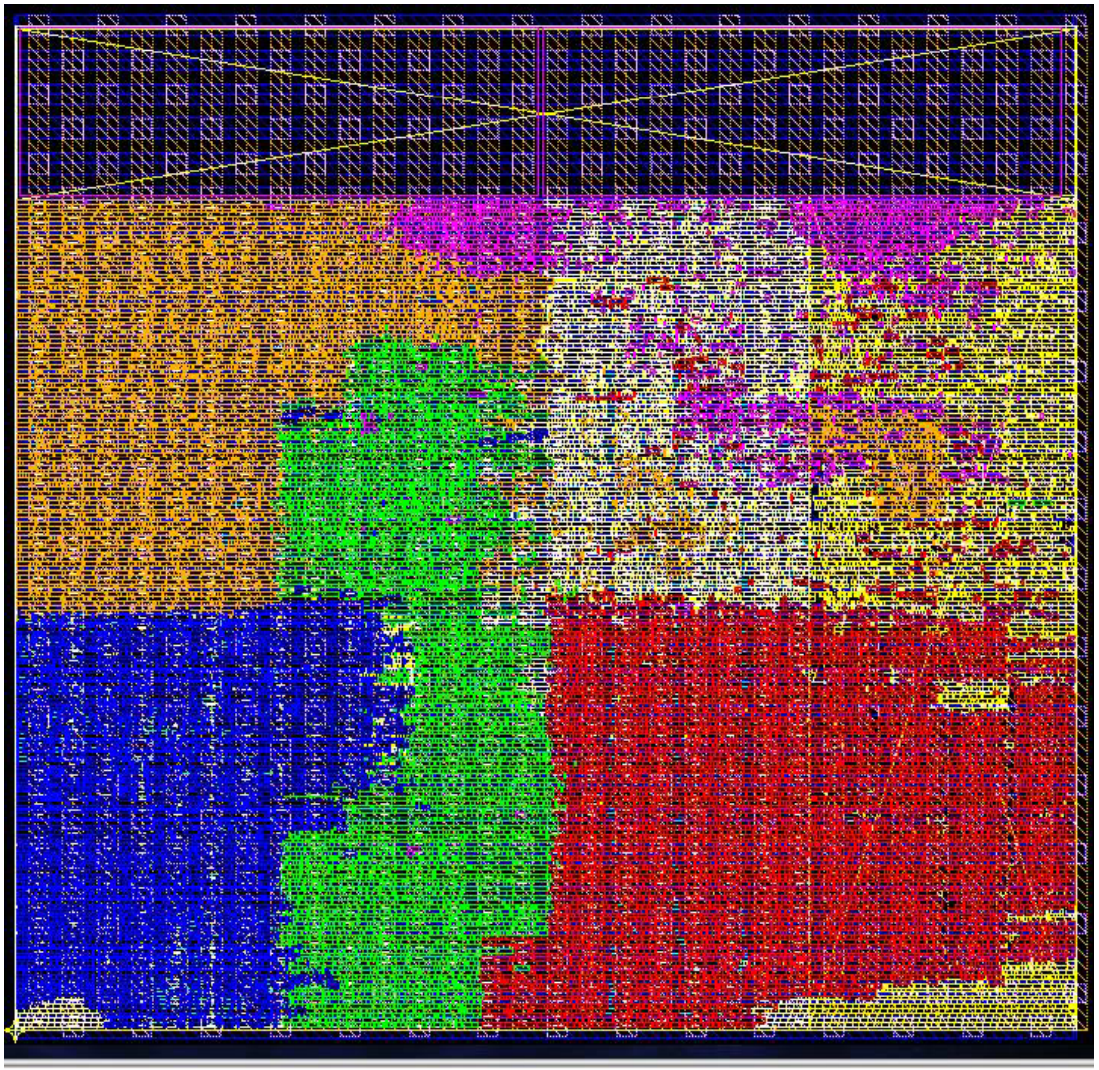
- 512Gflops(32bit)
- 256Gflops(64bit)
- 60W(max)



Kei Hiraki

Chip layout in a processing element

Module Name	Color
grf0	red
fmul0	orange
fadd0	green
alu0	blue
lm0	magenta
others	yellow



GRAPE-DR Processor chip

Engineering Sample (2006)

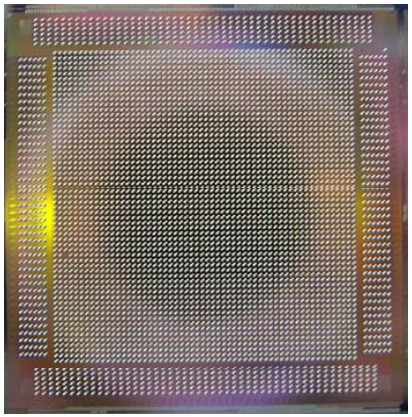
512 Processing Elements/Chip（Largest number in a chip）

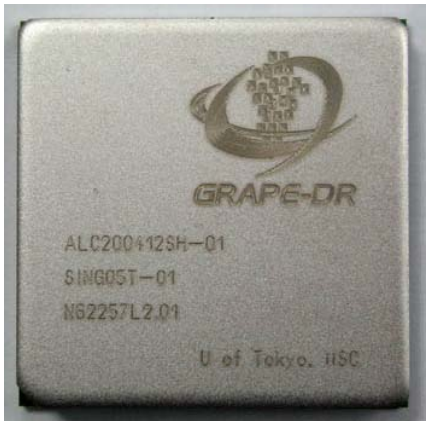
Working on a prototype board (at designed speed)

500MHz、512Gflops（Fastest chip for production）

Power Consumption max. 60W　　Idle 30W

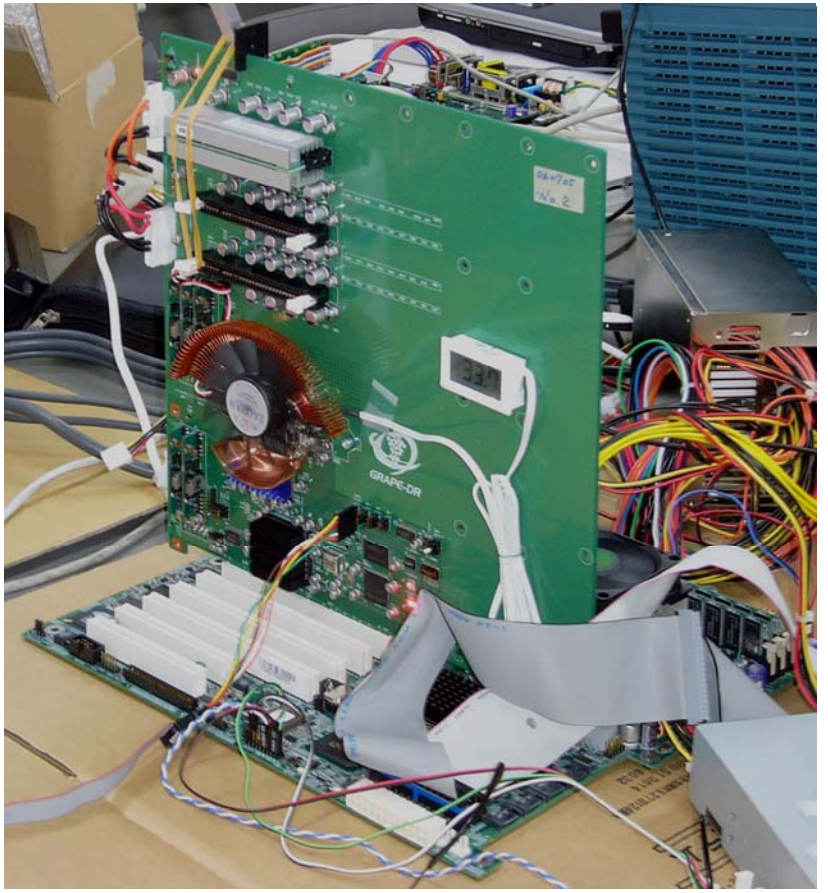
（Lowest power per computation）





GRAPE-DR Prototype boards

- For evaluation（1 GRAPE-DR chip、PCI-X bus）
- Next step is 4 chip board

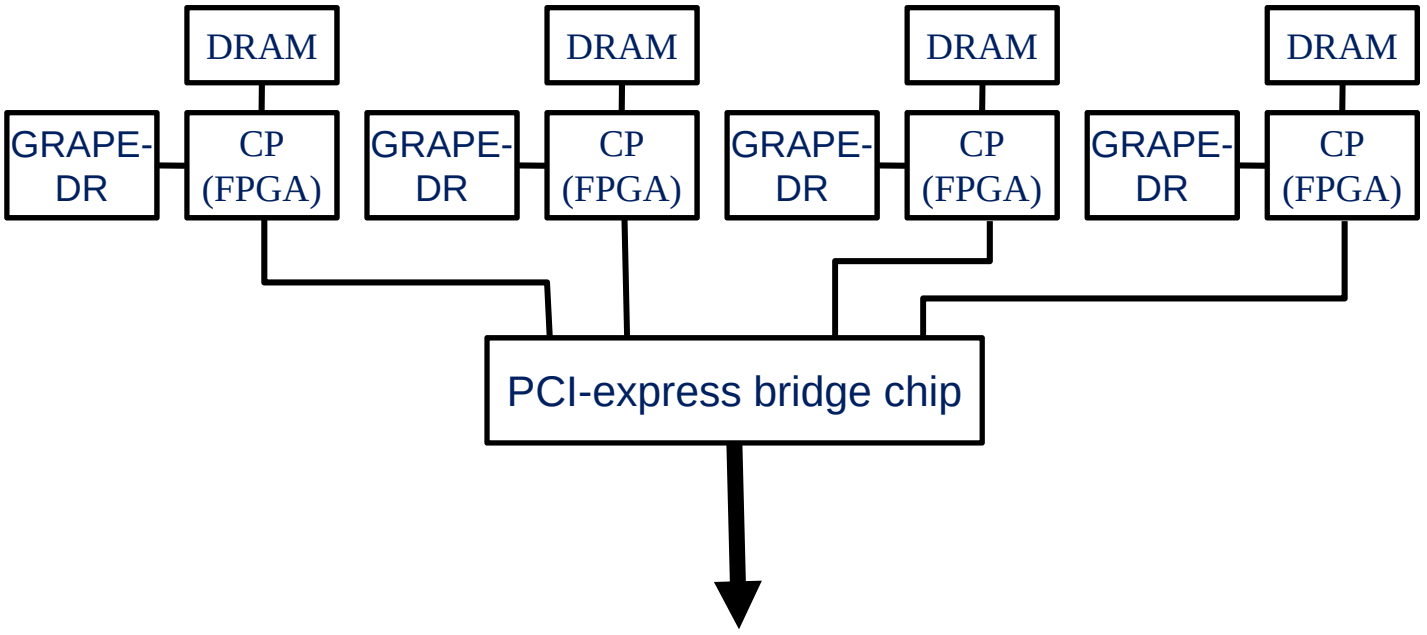


GRAPE-DR Prototype boards(2)

- 2nd prototype
- Single-chip board
- PCI-Express x8 interface
- On-board DRAM

Block diagram of a board

- 4 GRAPE-DR chip ⇒ 2 Tflops(single), 1Tflops(double)
- 4 FPGA for control processors
- On-board DDR2-DRAM 4GB
- Interconnection by a PCI-express bridge chip



PCI-express x16

Compiler for GRAPE-DR

- GRAPE-DR optimizing compiler
  - Parallelization with global analysis
  - Generation of multi-threaded code for complex operation

○ Sakura-C compiler

- Basic optimization

  （PDG,flow analysys、pointer analysis etc.）
- Souce program in C ⇒ intermeddiate language

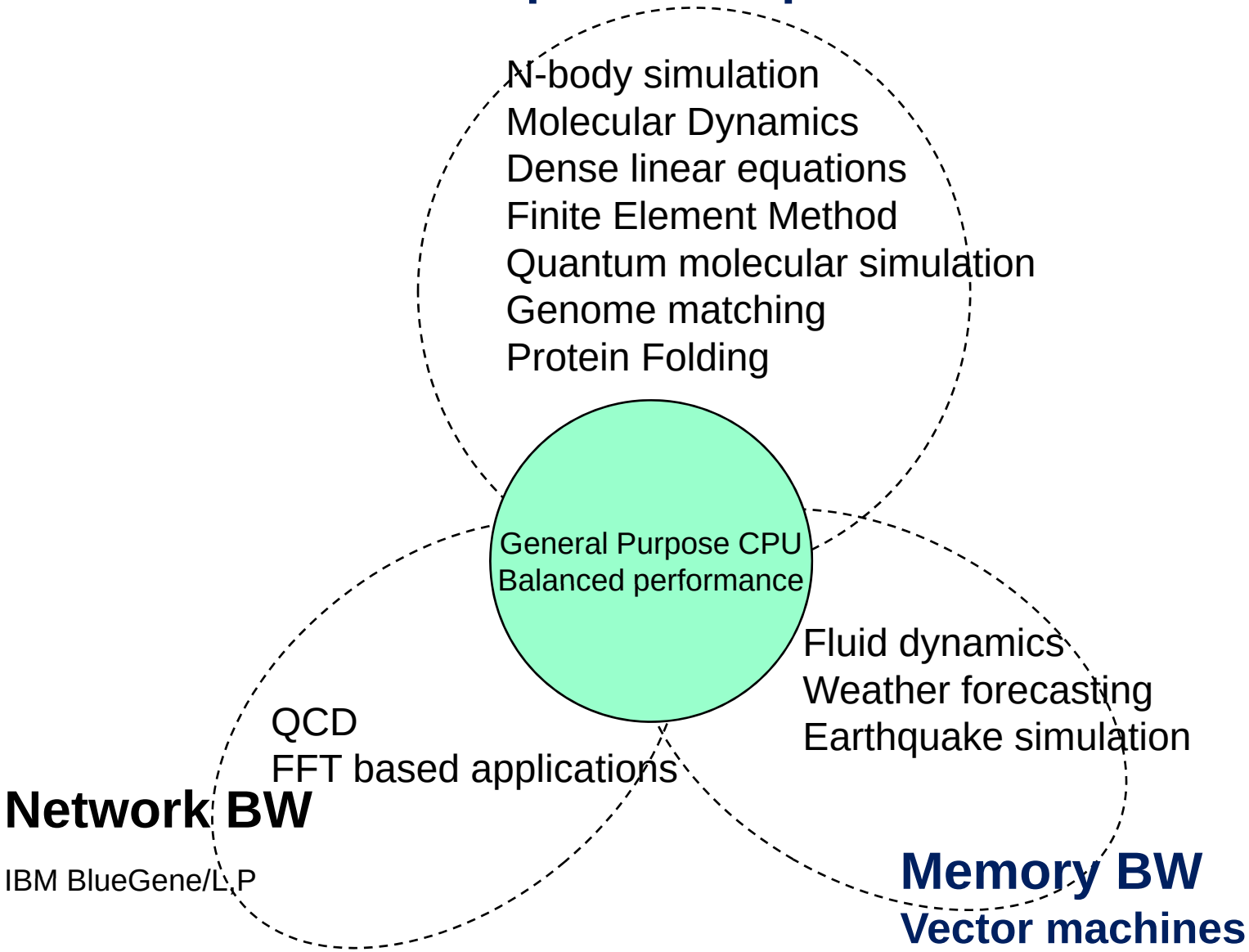
  ⇒ GRAPE-DR machine code

○ Currently, 2x to 3x slower than hand optimized codes

GRAPE-DR Application software

- Astronomical Simulation
- SPH(Smoothed Particle Hydrodynamics)
- Molecular dynamics
- Quantum molecular simulation
- Genome sequence matching
- Linear equations on dense matrices
  - Linpack

Application fields for GRAPE-DR

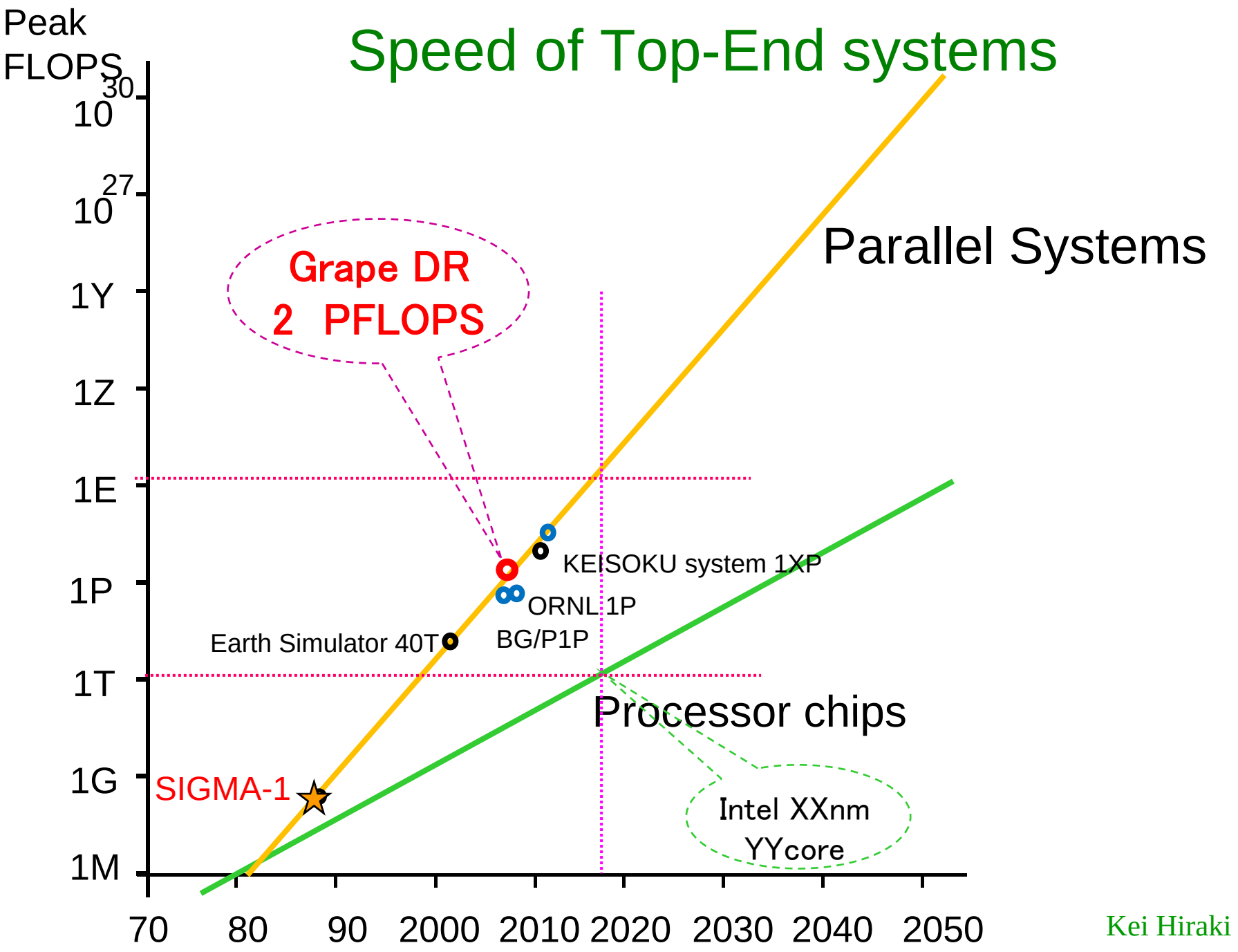


Eflops system

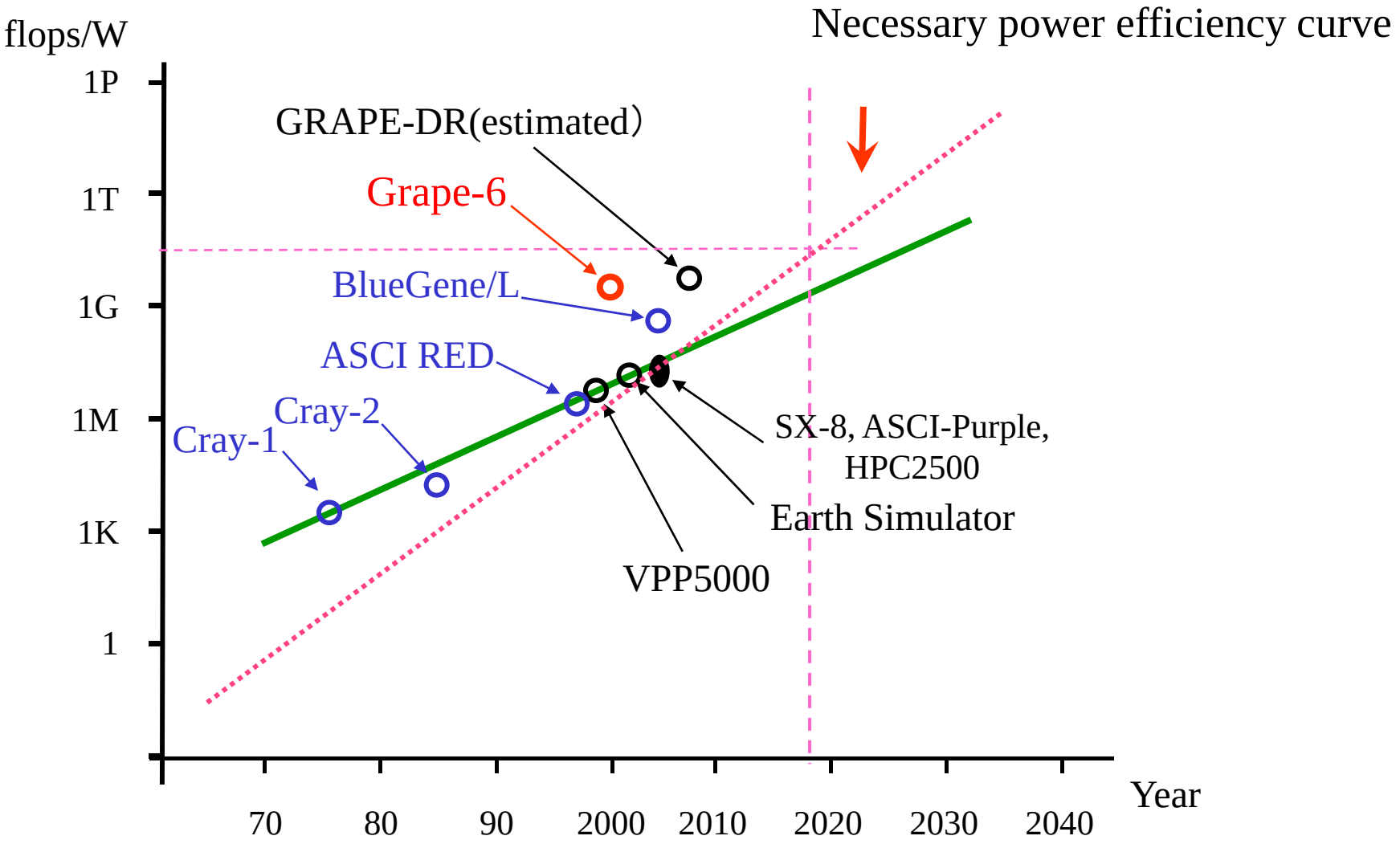
- 2011 10Pflops systems will be appeared
  - Next Generation Supercomputer by Japanese Government
  - IBM BlueGene/Q system
  - Cray Cascade system (ORNL)
  - IBM HPCS

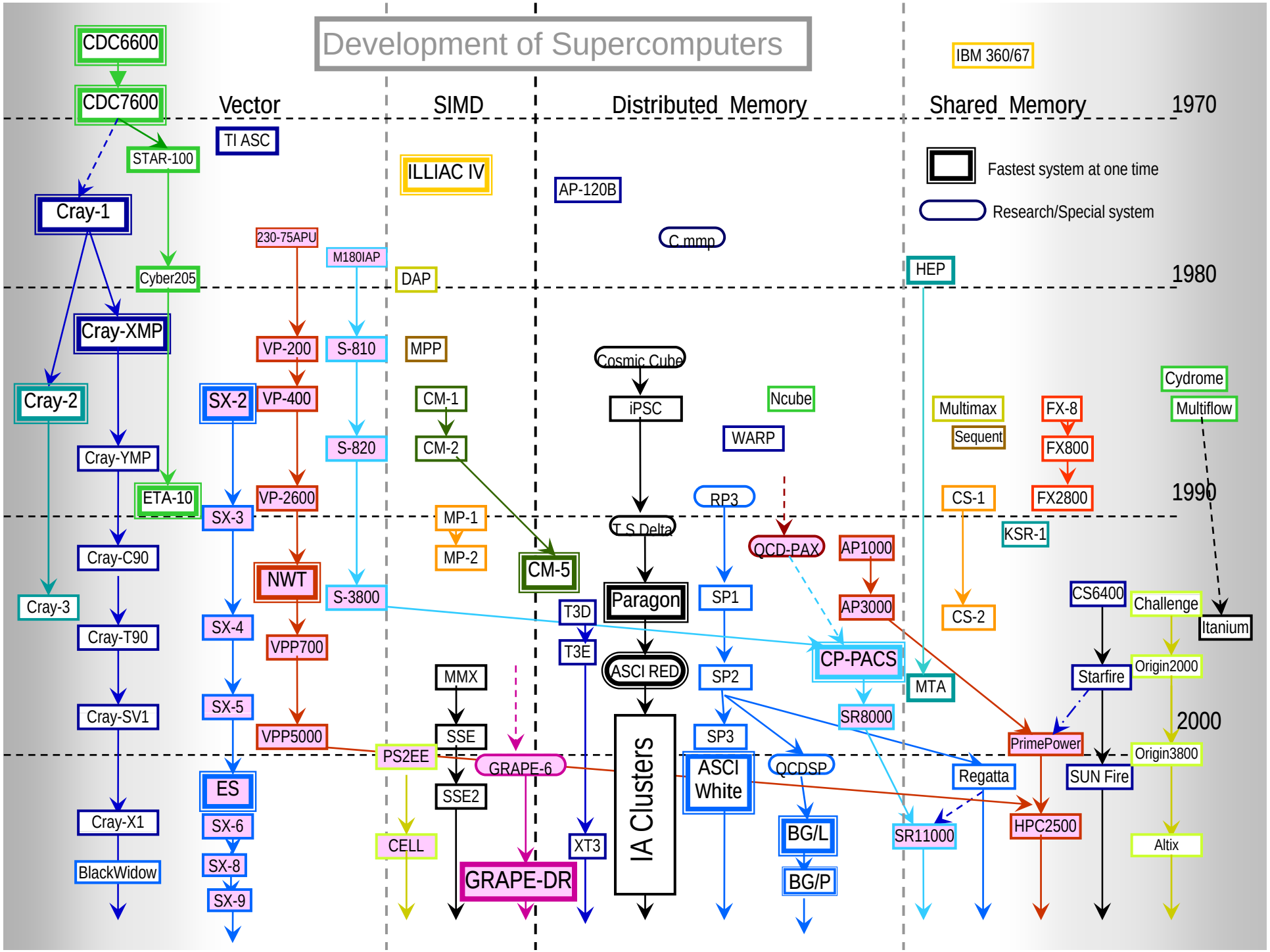
Architecture

- Cluster of Comodity MPU (Sparc/Intel/AMD/IBM Power)
- Vector processor
- Densely integrated low-power processors
- 1 ～ 1.5 ～2.5 MW/Pflops power efficiency


Speed of Top-End systems
Peak FLOPS
10^30
10^27
1Y
1Z
1E
1P
1T
1G
1M
70
80
90
2000
2010
2020
2030
2040
2050
Grape DR
2 PFLOPS
Parallel Systems
KEISOKU system 1XP
ORNL 1P
BG/P1P
Earth Simulator 40T
SIGMA-1
Processor chips
Intel XXnm
YYcore
Kei Hiraki

History of power efficiency




Development of Supercomputers
Vector
SIMD
Distributed Memory
Shared Memory
1970
1980
1990
2000
Fastest system at one time
Research/Special system
CDC6600
CDC7600
STAR-100
Cray-1
Cyber205
Cray-XMP
Cray-2
Cray-YMP
ETA-10
Cray-C90
Cray-3
Cray-T90
Cray-SV1
Cray-X1
BlackWidow
TI ASC
SX-2
SX-3
SX-4
SX-5
ES
SX-6
SX-8
SX-9
230-75APU
VP-200
VP-400
VP-2600
NWT
VPP700
VPP5000
M180IAP
S-810
S-820
S-3800
ILLIAC IV
DAP
MPP
CM-1
CM-2
CM-5
MP-1
MP-2
MMX
SSE
SSE2
PS2EE
GRAPE-6
GRAPE-DR
CELL
AP-120B
C.mmp
Cosmic Cube
iPSC
T.S.Delta
Paragon
ASCI RED
T3D
T3E
XT3
IA Clusters
Ncube
WARP
RP3
SP1
SP2
SP3
ASCI White
QCD-PAX
QCDSP
BG/L
BG/P
IBM 360/67
HEP
MTA
Multimax
Sequent
CS-1
CS-2
AP1000
AP3000
CP-PACS
SR8000
SR11000
Regatta
HPC2500
FX-8
FX800
FX2800
KSR-1
CS6400
Starfire
PrimePower
SUN Fire
Cydrome
Multiflow
Challenge
Itanium
Origin2000
Origin3800
Altix

Eflops system

- Requirements for building Eflops systems
  - Power efficiency
    - MAX Power ～ 30 MW
    - Efficiency must be smaller than 30 Gflops/W
  - Number of processor chips
    - 300,000 chips ? (3 Tflops/chip)
  - Memory system
    - Smaller memory / processor chip ⇒ Power consumption
    - Some kind of shared memory mechanism is a must
  - Limitation of cost
    - 1 billion $/ Eflops ⇒ 1000 $/Tflops

Effective approaches

- General purpose MPUs (clusters)
- Vectors
- GPGPU
- SIMD (GRAPE-DR)
- FPGA based simulating engine

Unsuitable architecture

- Vectors
  - Too much power for numerical computation
  - Too much power for wide memory interface
  - Less effective to wide range of numerical algorithms
  - Too expensive for computing speed

- General purpose MPUs (clusters)
  - Too low FPU speed per die size
  - Too much power for numerical computation
  - Too many processor chips for stable operation

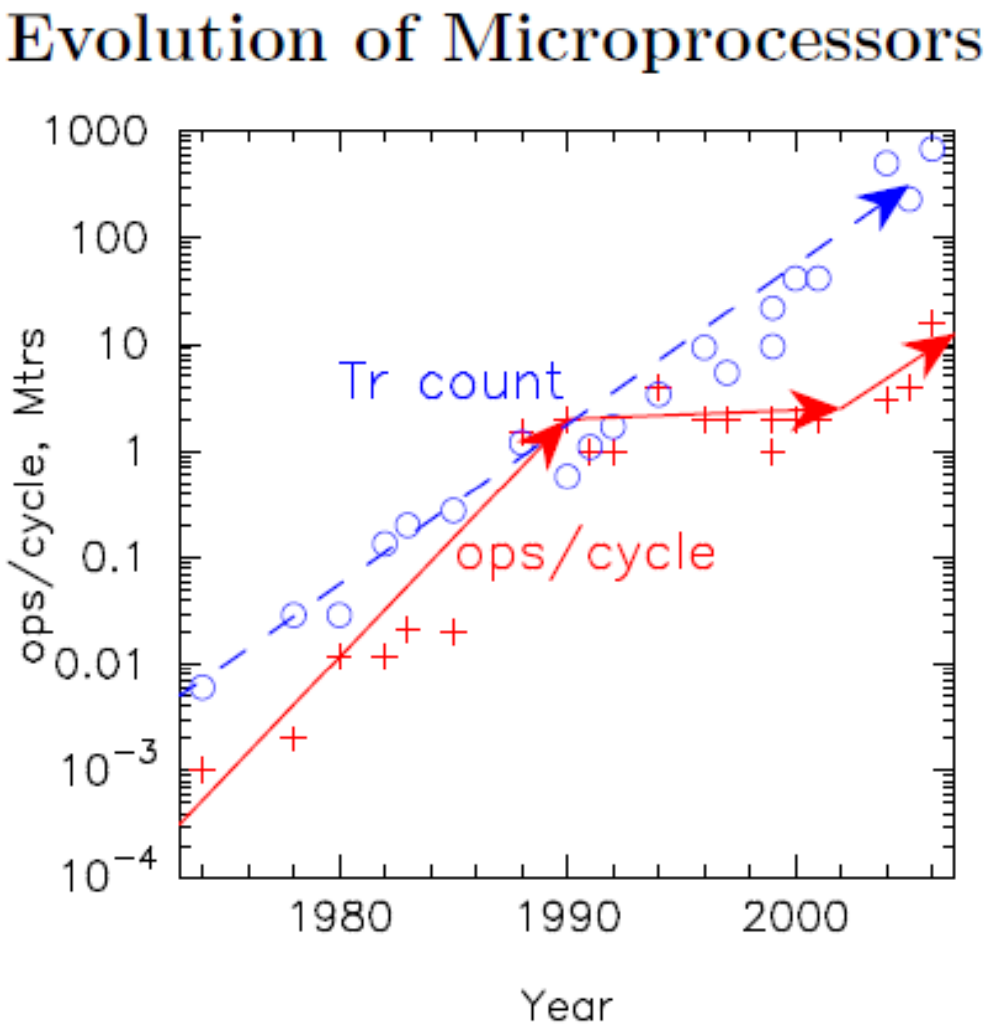
Unsuitable architecture

- Vectors
  - Current version 65nm Technology
    - 0.06 Gflops/W peak
    - 2,000,000$/Tflops

- General purpose MPUs (clusters)
  - Current version 65nm Technology
    - 0.3 Gflops/W peak (BlueGene/P, TACC etc.)
    - 200,000$/Tflops

Goals are
30Gflops/W, 1000$/Tflops, 3Tflops/chip

General purpose CPU



- Transistors: 1000 times in last 15 years
- FPUs: 8 times more in the same period
- a factor of 100 "lost"

(J.Makino 2007)

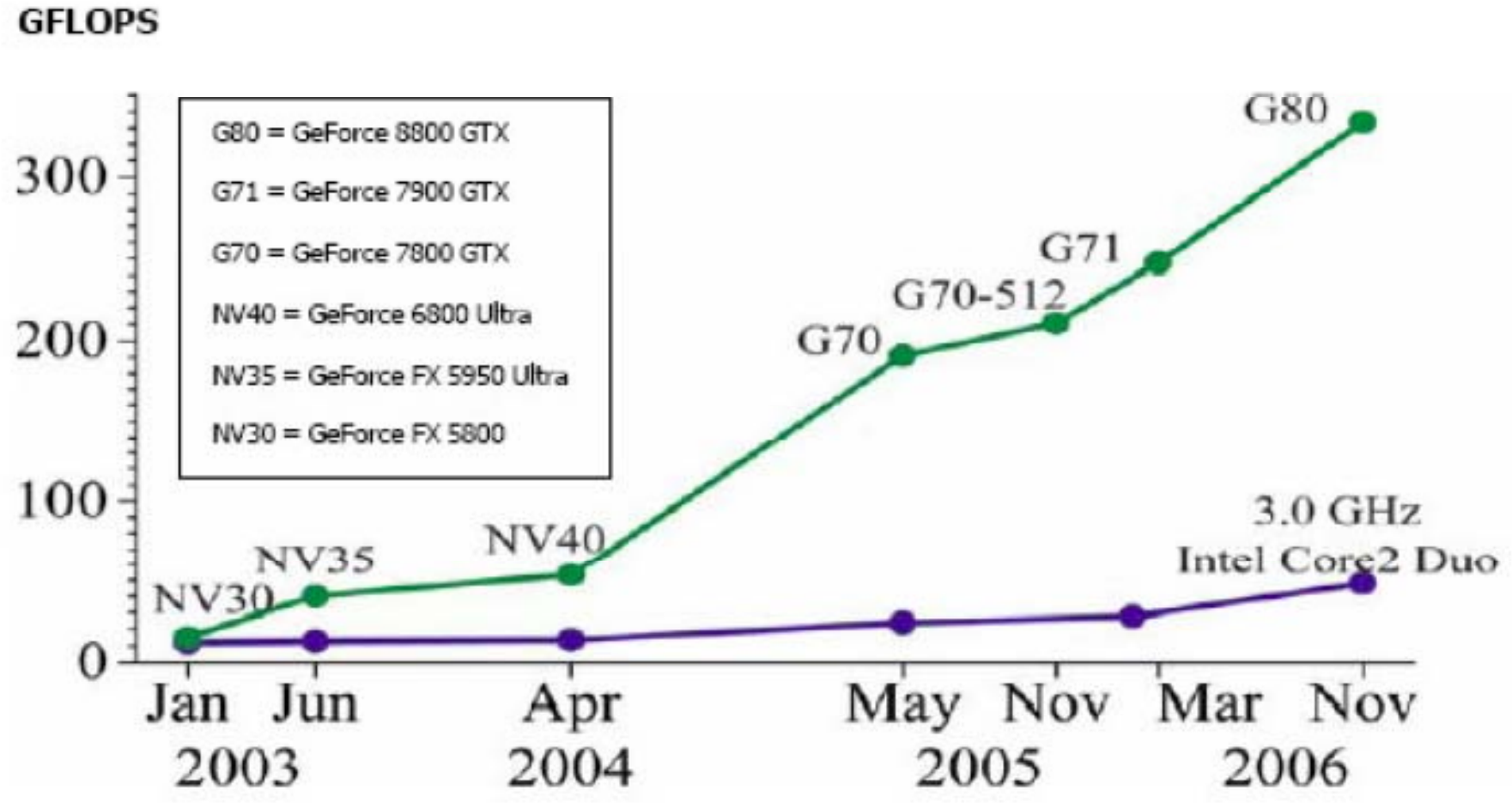
Architectural Candidates

- SIMD (GRAPE-DR)
  - 30% FALU
  - 20% IALU
  - 20% Register file
  - 20% Memory
  - 10% Interconnect
- GPGPU
  - Less expensive due to the number of production
  - Less efficient than properly designed SIMD processor
    - Larger die size
    - Larger power consumption

About 50% is used for computation

Very efficient

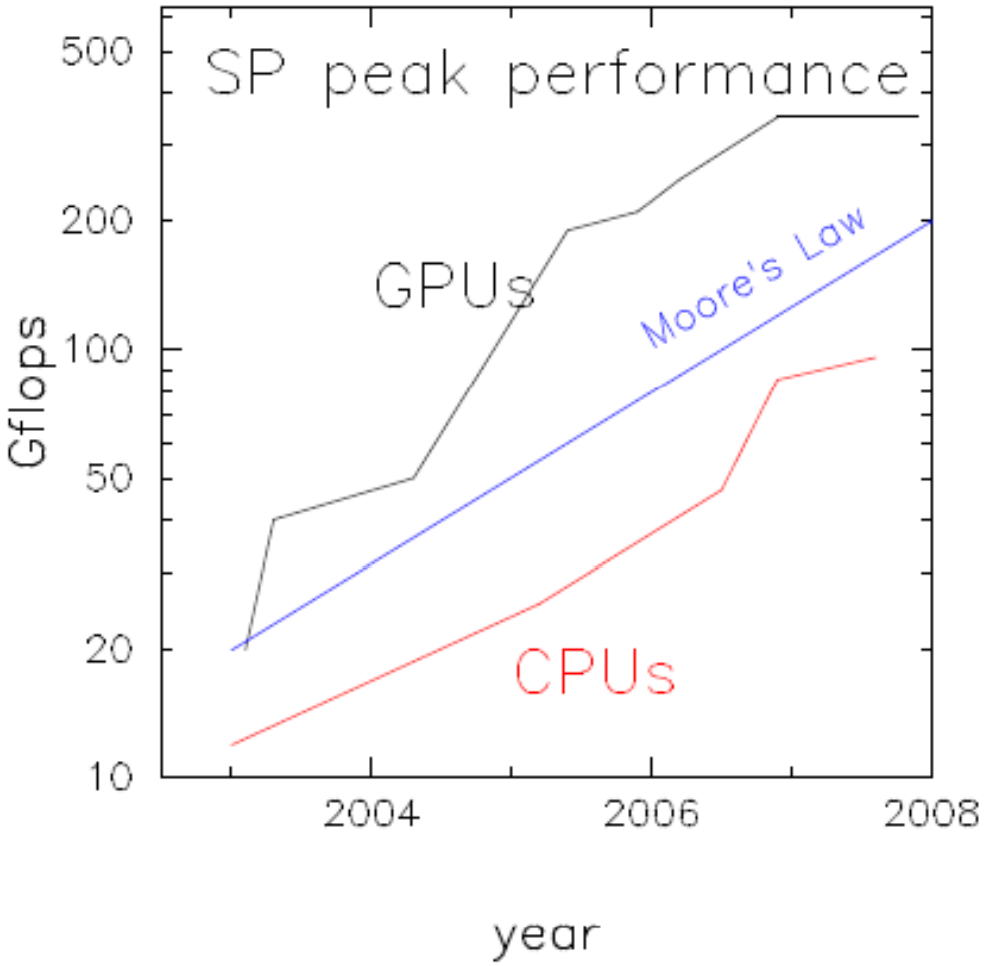
GPGPU(1)



"GPUs beat Moore's Law!"

(J.Makino 2007)

GPGPU(2)



- Faster-than-Moore period ended in 2005
- Microprocessors are catching up
- DP performance?
- Design limit with memory bandwidth

(J.Makino 2007)

Comparison with GPGPU

Pros:

- Significantly better silicon usage
  (512PEs with 90nm)
- Designed for scientific applications
  reduction, small communication overhead, etc

Cons:

- Higher cost per silicon area...
  (small production quantity)
- Longer product cycle... 5 years vs 1 year

Good implementations of N-body code on GPGPU are coming (Hamada, Nitadori, Portegies Zwart, Harris, ...)

Comparison with GPGPU(2)

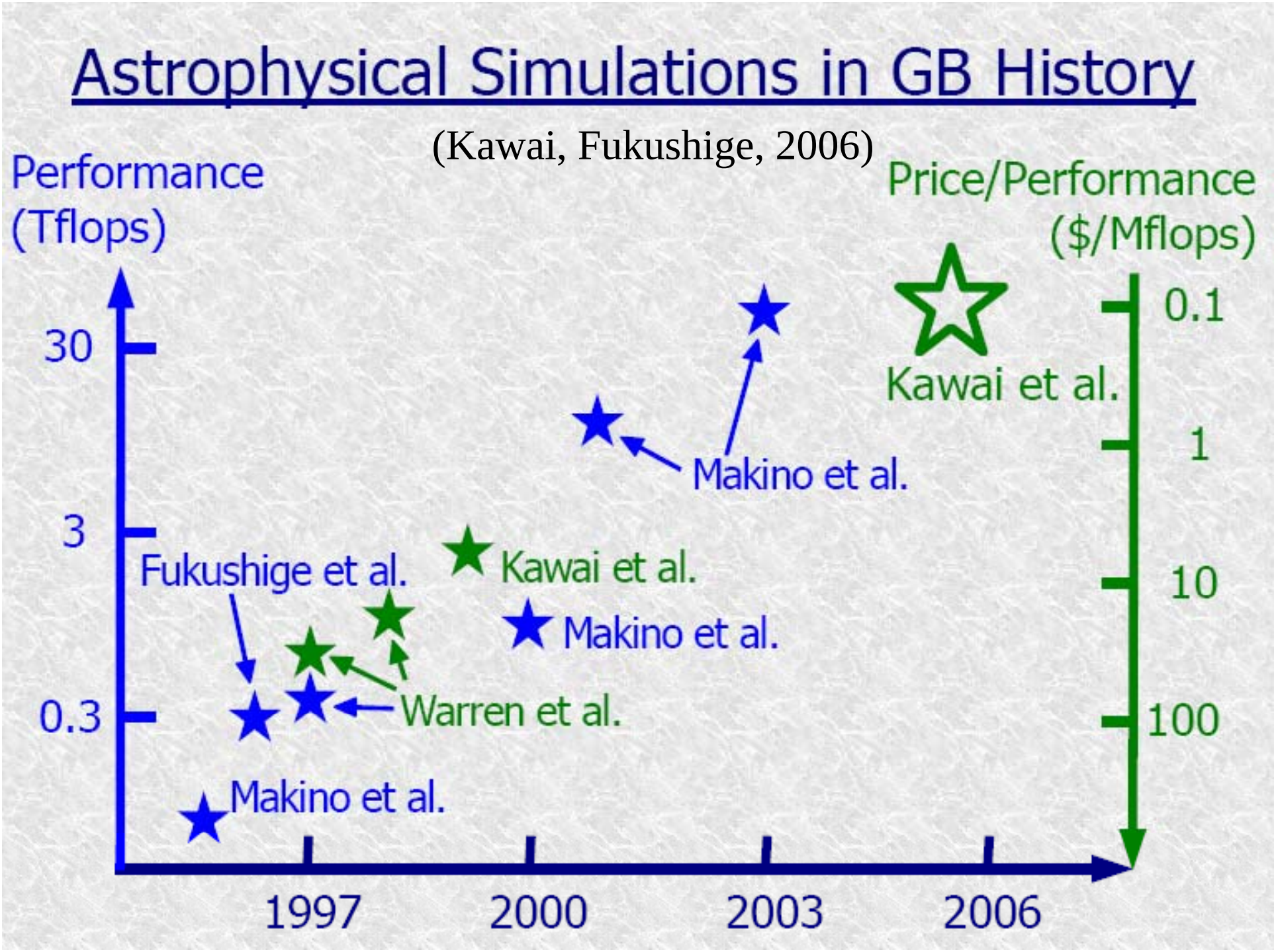
	GRAPE-DR	nV G92	AMD FS9170
Design rule	90	65	55
Clock(GHz)	0.5	1.5	0.8
# FPUs	512	112	320
SP peak(GF)	512	336	512
DP peak(GF)	256	—	?
Power(W)	65	70?	150?

(J.Makino 2007)

FPGA based simulator

- Commercial FPGAs cannot do FP operations faster than microprocessors.
  - Has been so in last decade
  - Will remain so for foreseeable future
- Custom reconfigurable processors cannot compete with commercial FPGAs in price-performance ratio.
  - Longer development cycle
  - Far smaller quantity

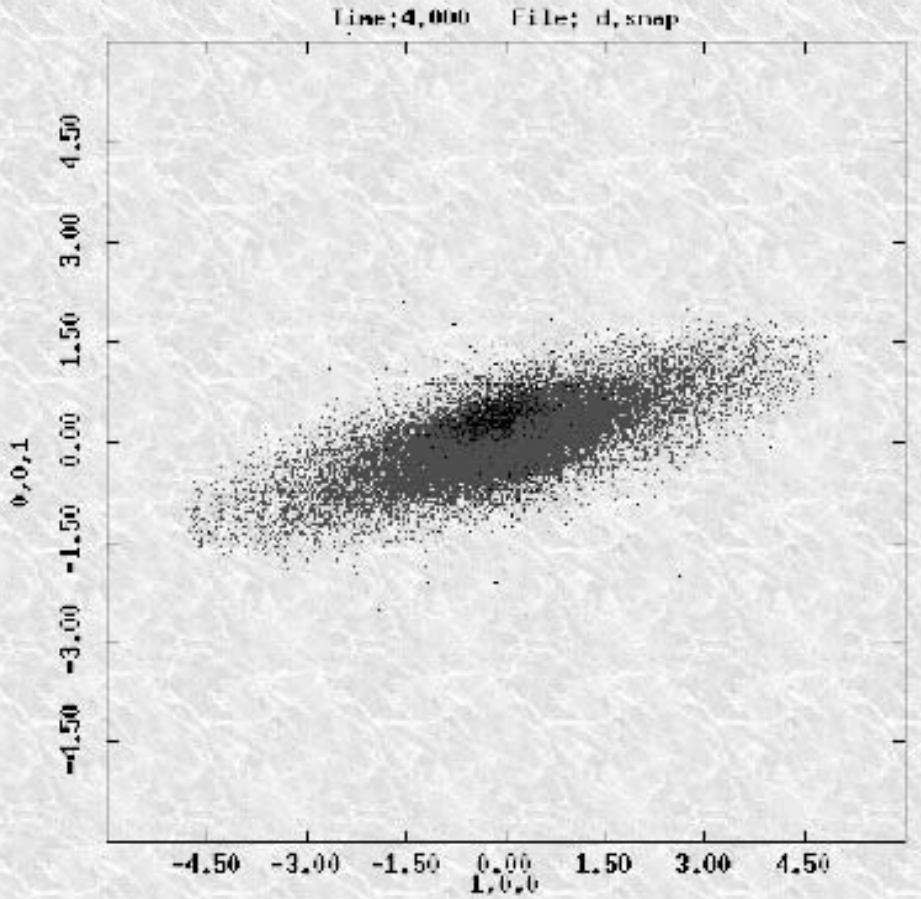
Good for applications with short-wordlength


Astrophysical Simulations in GB History
(Kawai, Fukushige, 2006)
Performance
(Tflops)
Price/Performance
($/Mflops)
30
3
0.3
0.1
1
10
100
Kawai et al.
Makino et al.
Fukushige et al.
Kawai et al.
Makino et al.
Warren et al.
Makino et al.
1997
2000
2003
2006

Astronomical Simulation

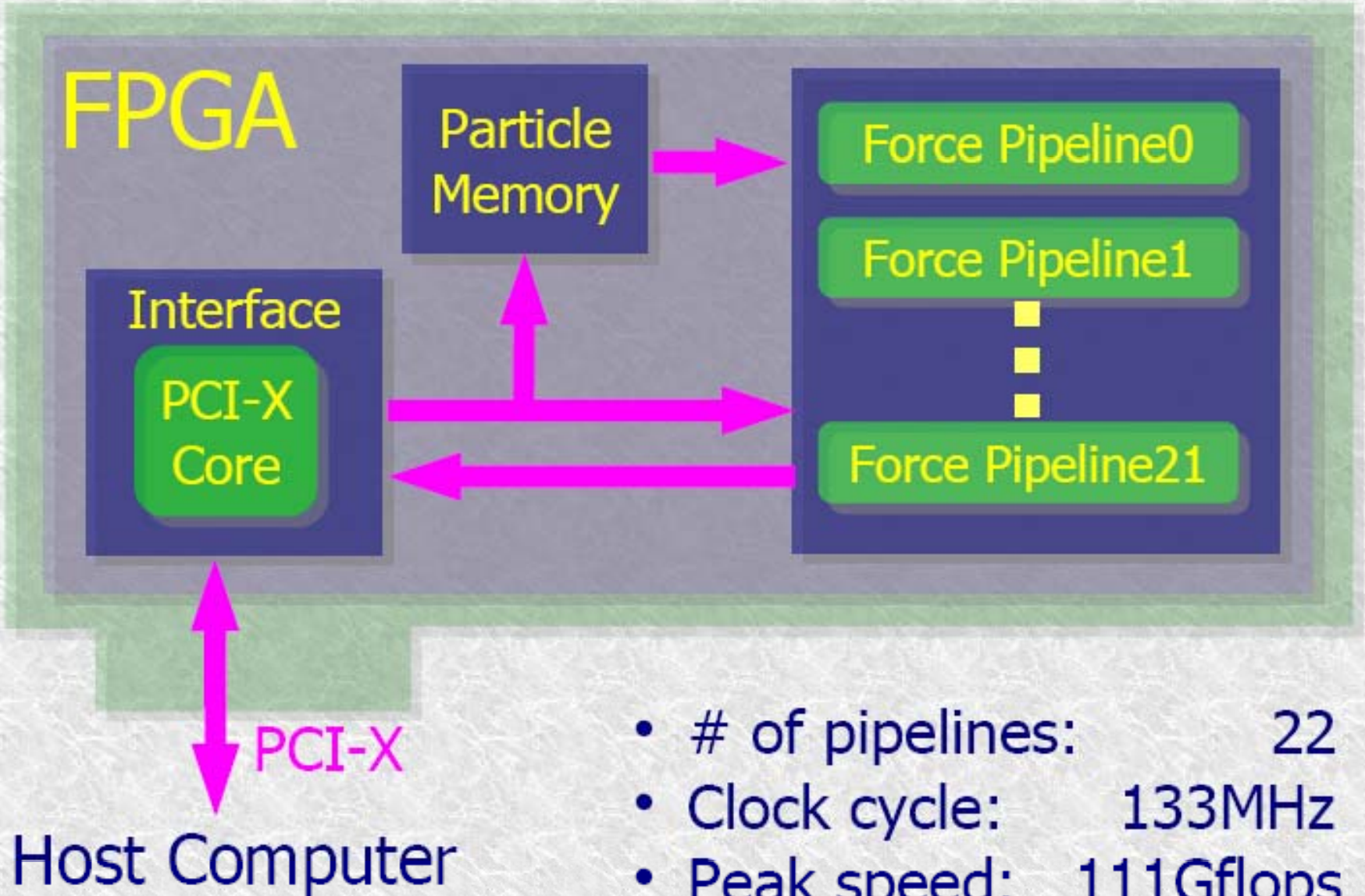
(Kawai, Fukushige, 2006)

An astronomical object



A representation by particles

Add-in Card Configuration



- # of pipelines: 22
- Clock cycle: 133MHz
- Peak speed: 111Gflops
- Power consumption: 3W

(Kawai, Fukushige, 2006)

Grape-7 board (Fukushige et al.)

(Kawai, Fukushige, 2006)

Price/Performance

Price (JPY) (Kawai, Fukushige, 2006)

Add-in Card:	187,619
Host PC:	91,238
Total:	278,857

→ $2,363

$1=118JPY

Corrected Performance: 22.59 Gflops

Price/Performance: $105/Gflops

(improved from $158/Gflops marked in July)

⇕

Previous record: $246/Gflops (Kim et al. 2001)

SIMD v.s. FPGA

- High-efficiency of FPGA based simulation
  - Small precision calculation
  - Bit manipulation
  - Use of general purpose FPGA is essential to reduce cost.

  - Current version 65nm Technology
    - 8 Gflops/W peak
    - 105,000$/Tflops

  - 30 Gflops/W peak will be feasible at 2017
  - 1000 $/ Tflops will be a tough target

SIMD v.s. FPGA

- GRAPE-DRx based simulation
  - Double precision calculation
  - Programming is much easier than FPGA based approach
    - HDL based programming v.s. C programming
  - Use of latest semiconductor technology is the key
  - Current version 90nm Technology
    - 4 Gflops/W peak
    - 10,000$/Tflops
  - 30 Gflops/W peak will be feasible at 2017
  - 1000 $/ Tflops will be feasible at 2017

Systems in 2017

- Eflops system
  - Necessary condition
    - Cost ～1000$/Tflops
    - Power consumption (30Gflops/W -> 30MW)
  - GRAPE-DR technology
    - 45nm CMOS -> 4Tflops/chip, 40Gflops/W
    - 22nm CMOS -> 16Tflops/chip, 160Gflops/W
  - GPGPU Low double precision performance
    - Power consumption, Die size
  - ClearSpeed
    - Low performance (25Gfops/chip)

To FPGA community

- Key issues to achieve next generation HPC
  - Low cost. Special purpose FPGA will be useless to survive in HPC
  - Low power. Low power is the most important feature of FPGA based simulation
  - High-speed. 500 Gflops/chip? (short precision)

- Programming environment
  - More important to HPC users (not skilled in hardware design)
  - Programming using conventional programming languages
  - Good compiler, run-time and debugger
  - New algorithm that utilize short-precision arithmetic is necessary