Development of Next-Generation Massively Parallel Computers for Continuous Physical Systems

Akira Ukawa

Center for Computational Physics
University of Tsukuba

Computational Needs of Physical Sciences

- *Need more computing power to analyze:*
  - *larger system sizes/more degrees of freedom*
  - *longer time intervals/smaller time steps*
  - *more complex systems*

1 TFLOPS (1996) → 10TFLOPS (2000) → 100TFLOPS (2004?)

- *Need improved quality to tackle:*
  - *complex phenomena having*
    multiples of interactions types
    multiples of scales

Difficulties with conventional supercomputer development

- vector-parallel machines VPP,SR,SX,ES

  High efficiency, *but large development and running cost*

- scalar-parallel machines ASCI/Red,Blue,White,Q

  Less expensive,

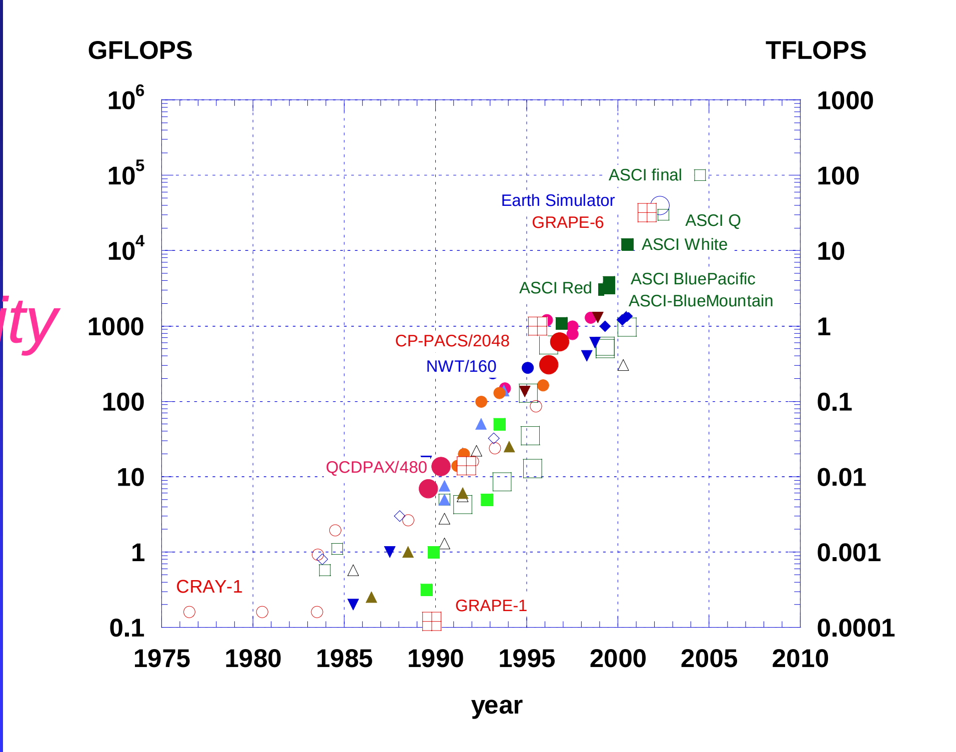
  but lower efficiency and

  large system size due

  to lower packaging density

⇓

Further enhancement along these architectures facing serious difficulties



Our Strategy: Classification and Synthesis

- two broad categories of physical systems
  - ***continuous systems: fluids,wave functions***,...
    - **local but complex force laws**
    - **only O(N) computations but need general-purpose CPU**
  - ***particle-based system: DNA/proteins,galaxies,…***
    - **long-ranged but universal and simple force laws**
    - **$O(N^2)$ computations but special-purpose CPU effective**

→ ***Pursue Hybrid approach:***

- *general/special processors for continuous/particle systems*
- *hybridization of the two systems for efficient processing of complex physical systems*

RFTF Project

Development of Next-Generation Massively Parallel Computers

Leader: Yoichi Iwasaki, University of Tsukuba

- *Development for Continuous Physical Systems*

  core member: Akira Ukawa, Taisuke Boku
  Center for Computational Physics,
  University of Tsukuba

- *Development for particle-based Systems*

  core member: Junichiro Makino
  Graduate School of Science,
  University of Tokyo

  → ***Development of GRAPE-6***

Target of our Project

- *R&D of key technologies* for next-generation MPP for continuous physical systems
  - **new processor architecture *SCIMA***
  - **interconnect and system design**
  - **parallel I/O and visualization environment *PAVEMENT***

- development of *hybrid multi-computer system (HMCS) for coupled continuous/particle-based simulations*

  in collaboration with the Makino subproject
  - **prototype system development**

    CP-PACS(continuous) and GRAPE-6(particles)
  - **novel astrophysics simulation with the prototype**

    galaxy formation under UV radiation

Project Organization

Leader : Yoichi Iwasaki (Univ. of Tsukuba)

Computer Science

- **Core member : Taisuke Boku (Univ. of Tsukuba)**
- Shigeru Chiba (Tokyo Inst. of Technology)
- Tsutomu Hoshino (Univ. of Tsukuba)
- Hiroshi Nakamura (Univ. of Tokyo)
- Ikuo Nakata (Housei Univ.)
- Kisaburo Nakazawa (Meisei Univ.)
- Shuichi Sakai (Univ. of Tokyo)
- Mitsuhisa Sato (Univ. of Tsukuba)
- Tomonori Shirakawa (Univ. of Tsukuba)
- Daisuke Takahashi (Univ. of Tsukuba)
- Moritoshi Yasunaga (Univ. of Tsukuba)
- Yoshiyuki Yamashita (Saga Univ.)
- Koichi Wada (Univ. of Tsukuba)
- Yoshiyuki Watase (KEK)

Computational Physics

- **Core member : Akira Ukawa (Univ.of Tsukuba)**
- Sinya Aoki (Univ. of Tsukuba)
- Kazuyuki Kanaya (Univ. of Tsukuba)
- Taishi Nakamoto (Univ of Tsukuba)
- Masanori Okawa (KEK)
- Hajime Susa (Univ. of Tsukuba)
- Masayuki Umemura (Univ. of Tsukuba)
- Tomoteru Yoshie (Univ. of Tsukuba)

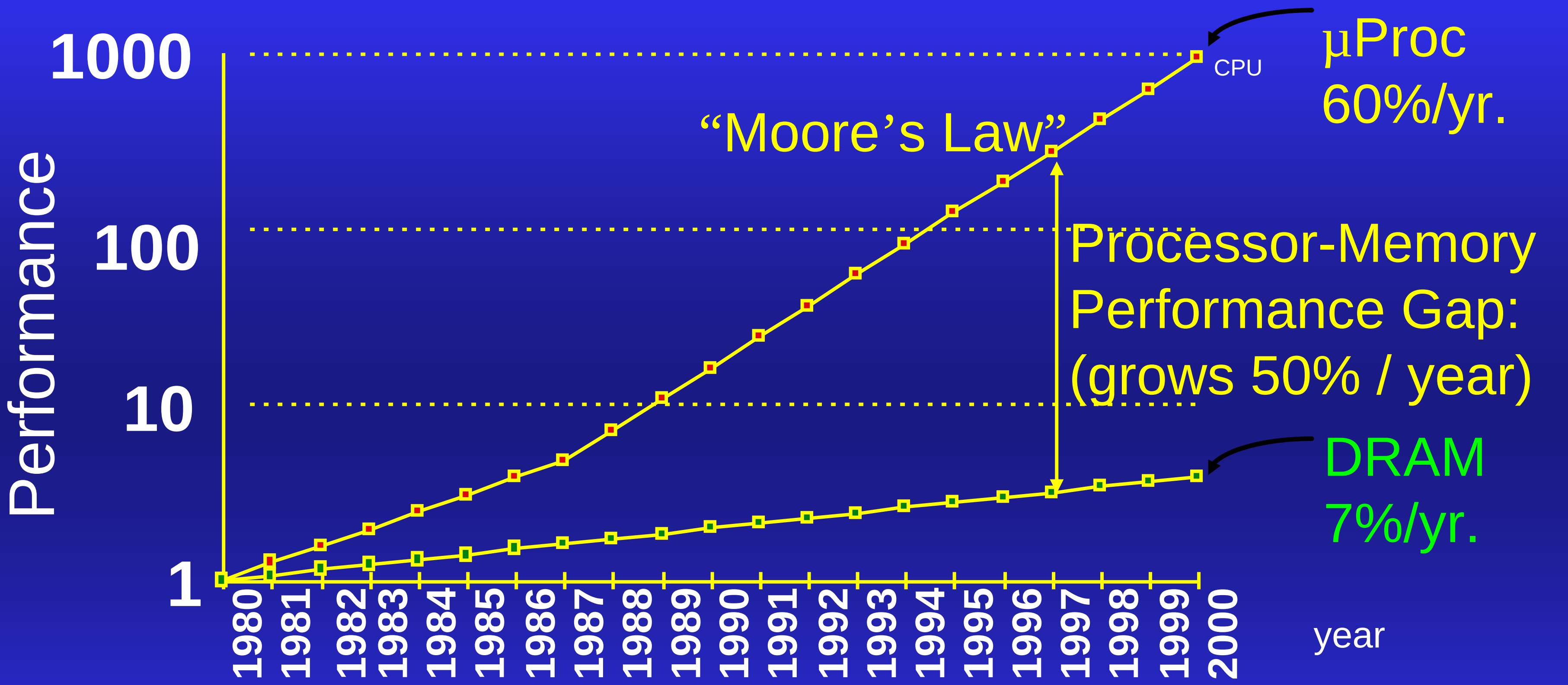
Project Organization II

- Processor architecture SCIMA
  - H. Nakamura,S. Chiba, M. Sato, D. Takahashi, S. Sakai
  - M. Kondo, M. Fujita, T. Ohneda, C. Takahashi, M. Nakamura
- Interconnect and system design
  - T. Boku, E. Oiwa
- Parallel I/O and visualization environment PAVEMENT
  - T. Boku, K. Itakura, M. Umemura, T. Nakamoto
  - M. Matsubara, H. Numa, S. Miyagi
  - Kubota Graphics Technologies
- HMCS and astrophysics simulation
  - T. Boku, M. Umemura, H. Susa
  - M. Matsubara
  - J. Makino, T. Fukushige

Novel Processor Architecture SCIMA for HPC applications

- Memory-CPU gap
- Design concept
- Performance evaluation
- RTL design
- Compiler

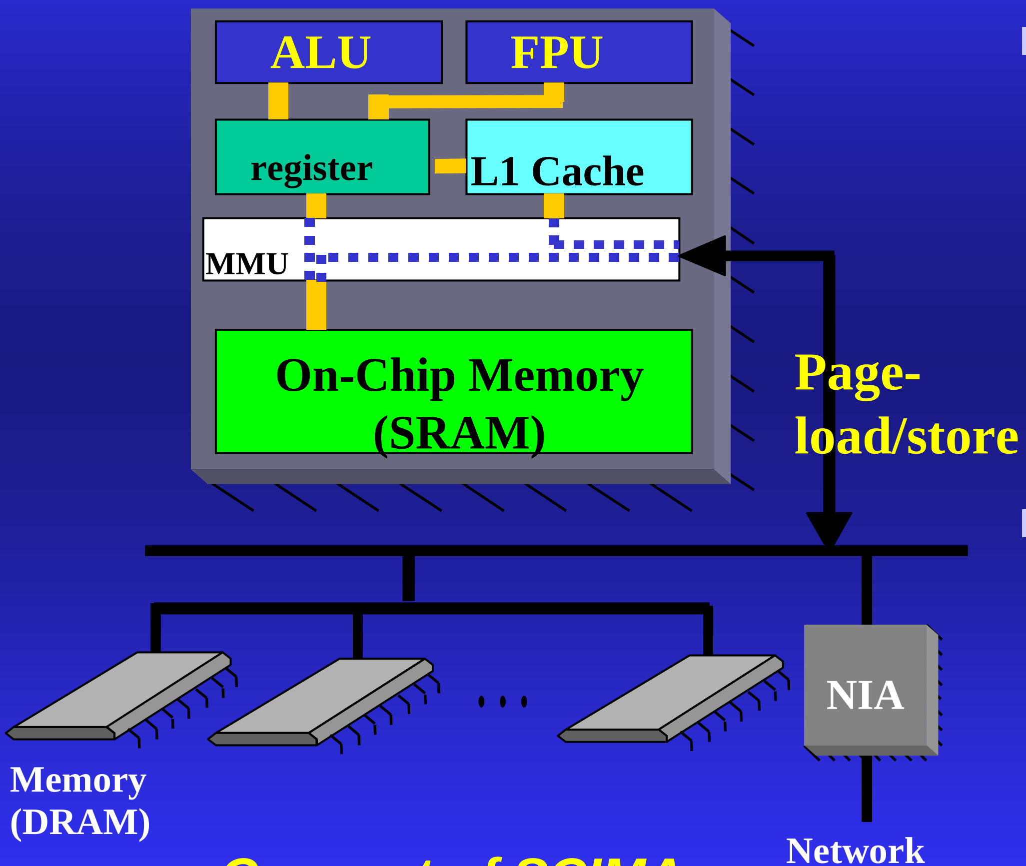
Growing gap of CPU and DRAM



cited from: Fig 5.1. "Computer Architecture A Quantitative Approach (2nd Editio
by J.Henessy and D.Patterson, Morgan Kaufmann (ISBN: 1-55860-329-8)

Effective performance of CPU limited by slow memory access

SCIMA: Software Controlled Integrated Memory Architecture



Concept of SCIMA

- *addressable On-Chip Memory* in addition to ordinary cache for *controlled allocation of data frequently used*
- *page-load/page-store* instruction between on-chip memory and off-chip memory for *large granularity* data transfer

Benefit of On-Chip Memory

- allocation / replacement is under software control
  - frequently used data is guaranteed to reside
  - interferences between/within arrays are avoided

  → can fully exploit data reusability
- data transfer from/to off-chip memory is under software control
  - large granularity transfer & block stride transfer

  → effective use of off-chip memory bandwidth

only regularly accessed data can enjoy this benefit

- data cache is still provided for irregular data

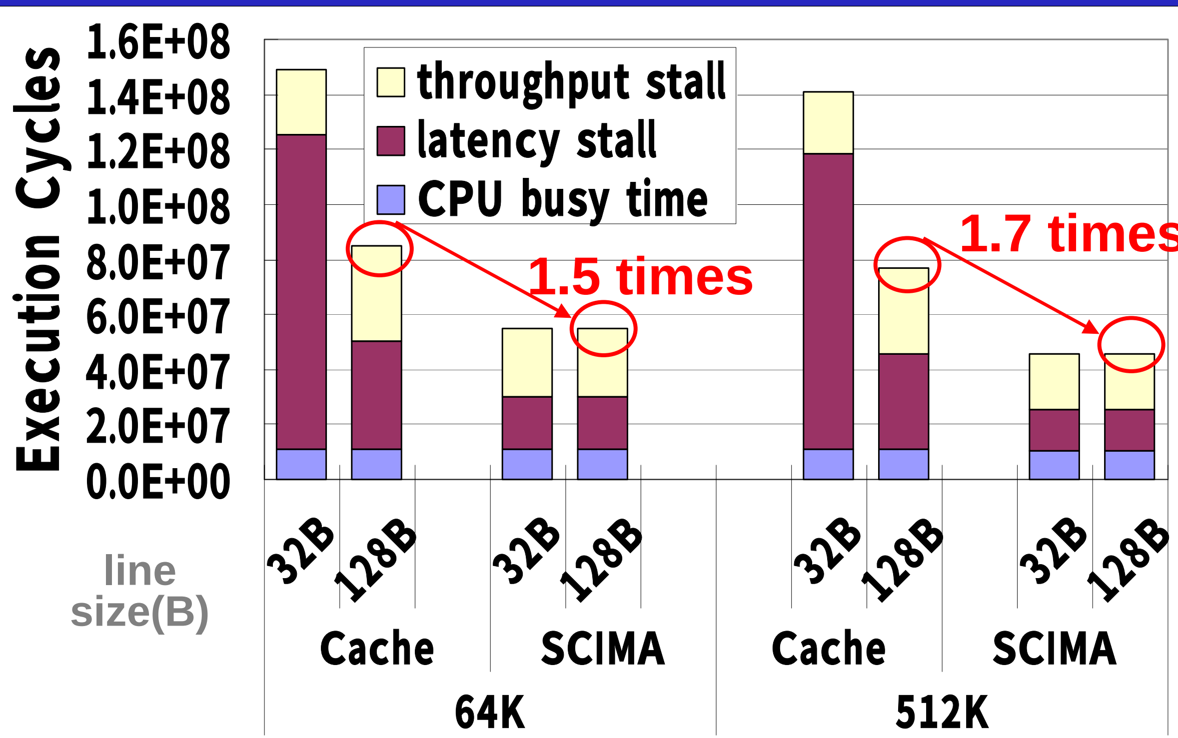
Performance Evaluation

- Benchmarks
  - NAS Parallel Benchmark: CG, ***FT***
  - ***QCD (quantum chromodynamics)***
    - real application: quantum mechanical system
- Optimization :
  by hand following the optimization strategy

- Data set :
  - NAS PB: class-W for saving simulation time
  - QCD: practical data size
    6 x 6 x 12 x 12 (4 dim. space-time)
    [48 x 48 x 48 x 96 on 2048 PU MPP]

Results of QCD

latency=160cycle, throughput=1B/cycle
longer latency, narrower throughput

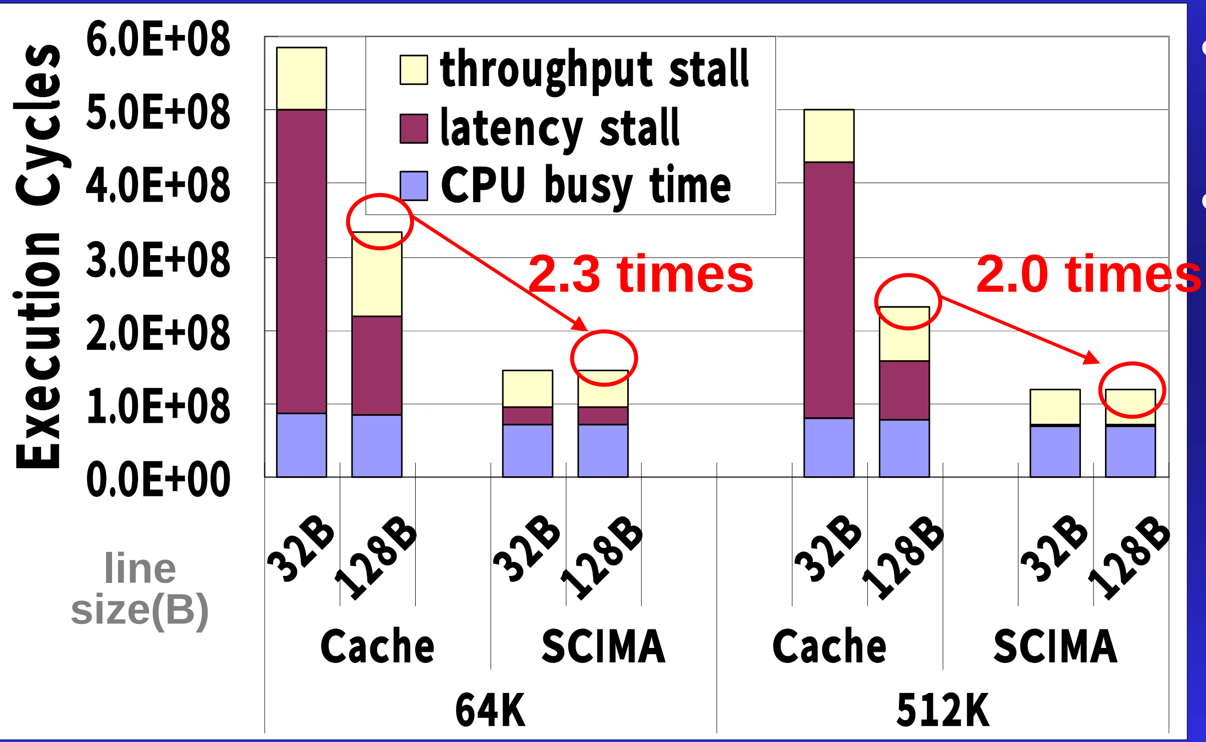


- SCIMA:
  1.5-3.0 times faster
- Cache
  larger line →
  ☺ latency stall
  ☹ throughput stall

- conflict on copied data
- unnecessary data transfe

Results of FT

latency=160cycle, throughput=1B/cycle
longer latency, narrower throughput

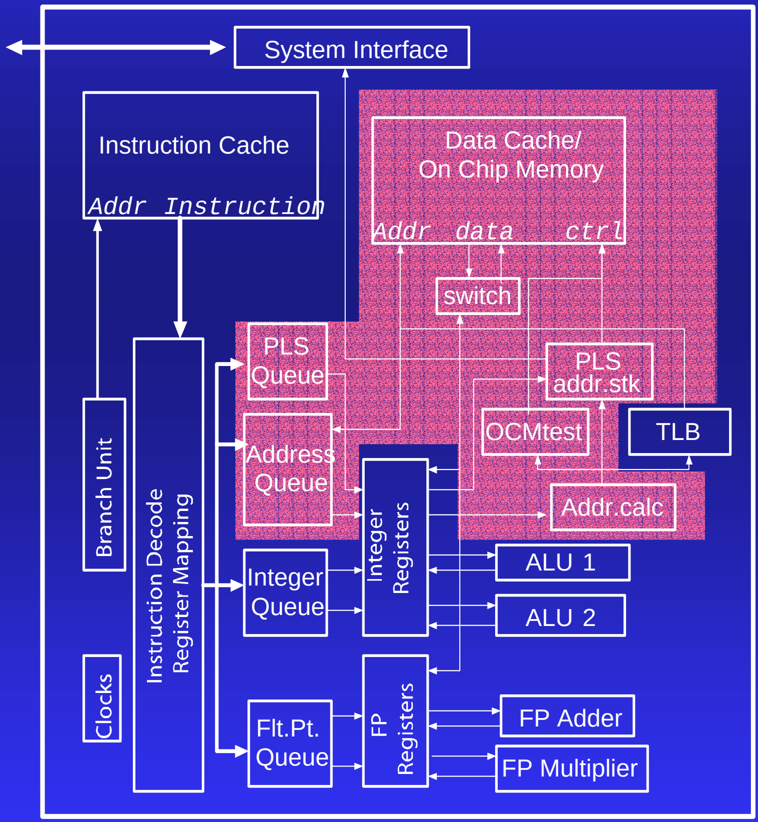


- SCIMA:
  2.0-4.2 times faster
- Cache
  - larger line →
  - ☺ latency stall
  - ☹ throughput stall

- conflict on copied data
- unnecessary data transfer

RTL design to check surface area and delay

- Compare MIPS R10000 processor (cache model) and SCIMA processor built on R10000 architecture
- RTL level design in Verilog-HDL
- Synthesis to gate level using
  - VDEC 0.35 μm process library
  - Synopsis design compiler
- evaluate chip surface area and delay



Designed part in red

Result of evaluation

- *Surface area*
  - address queue(load/store issuing mechanism) occupies 3% of R10000 chip
  - this area expands by 1.7 for SCIMA chip;

    negligible effect on chip surface area
- *Delay*
  - longest delay in cache model comes from instruction issue mechanism; possible critical path
  - 4.9% increase for SCIMA processor

    SCIMA model still much faster than the cache model

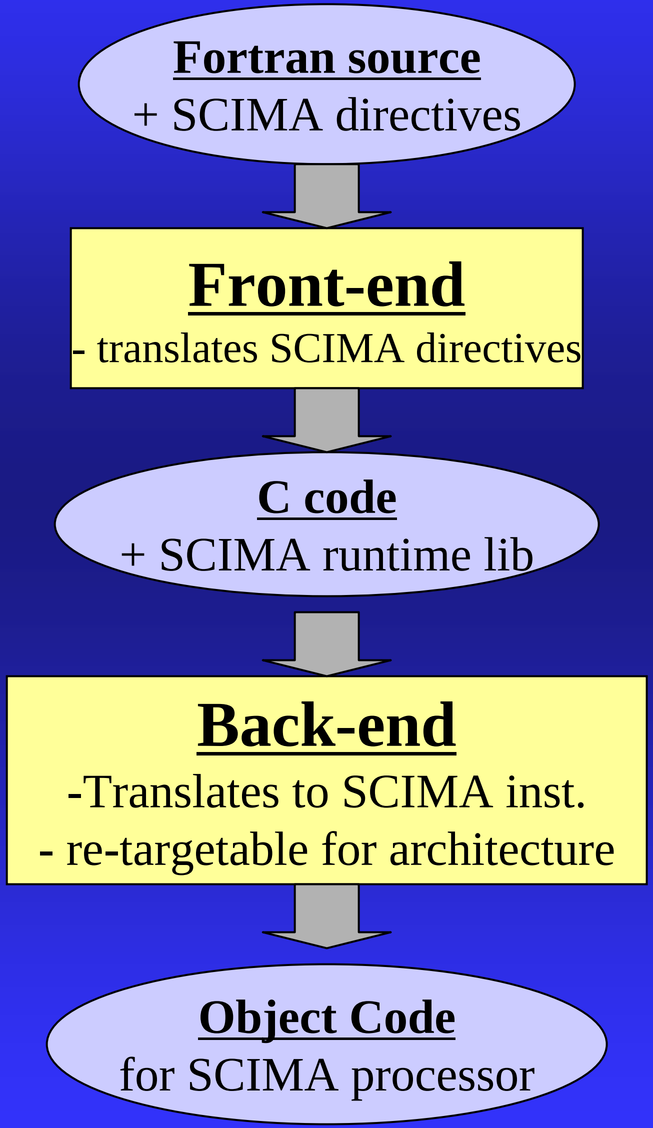
	Delay in instruction issue[ns]
cache model	6.11
SCIMA model	6.48

Compiler for SCIMA processor

- **SCIMA directives**
  - On-chip memory mapping
  - Control on-chip and off-chip memory transfer.
- backend: **re-targetable code generator** for various architectural parameters.
  - Number of registers
  - Instructions

Sample program

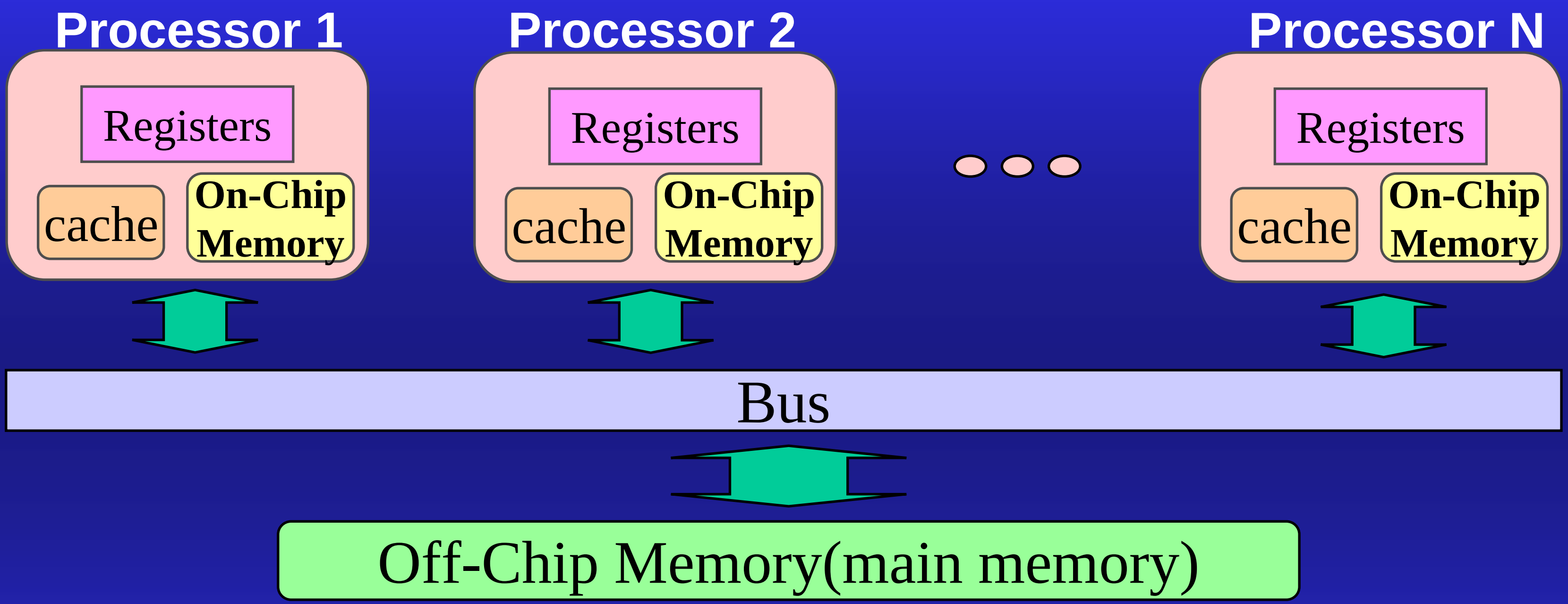
```
    double precision sum
    double precision a(N*2,N*2)
!$scm begin (a, N, N, 0, 0)
!$scm load (a, N + 1, N + 1, N, N)
    sum = 0.0
    do i = N + 1, N * 2
       sum = sum + a(i, i)
    enddo
!$scm end (a)
```



Interconnect and system design

- SMP configuration of SCIMA
- system image

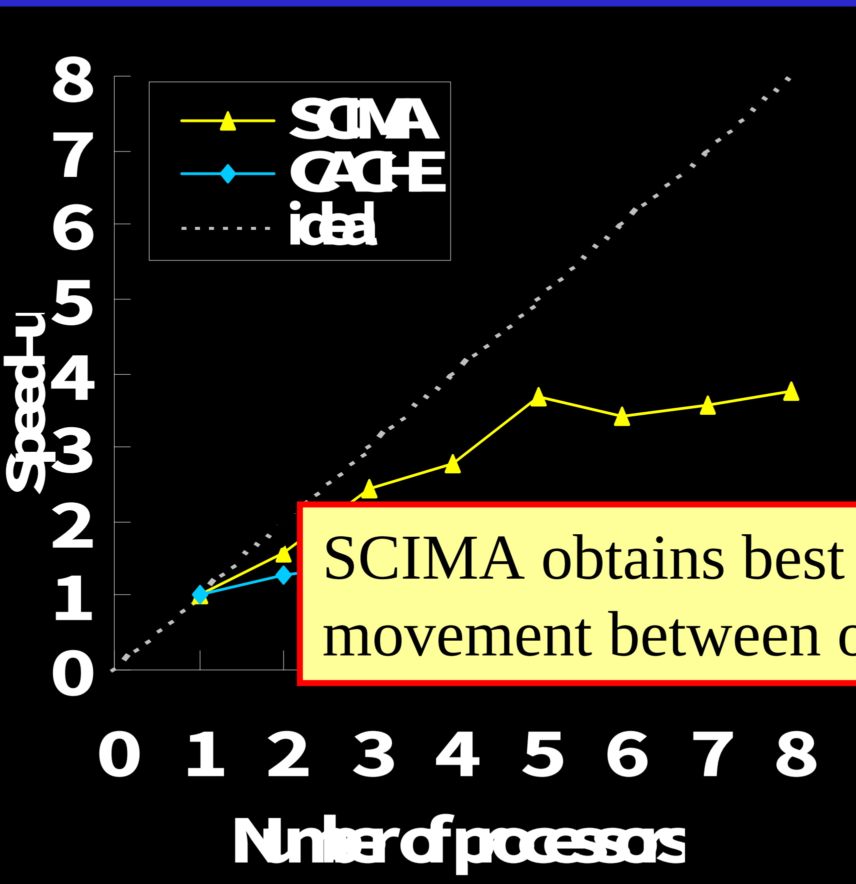
SMP configuration of SCIMA



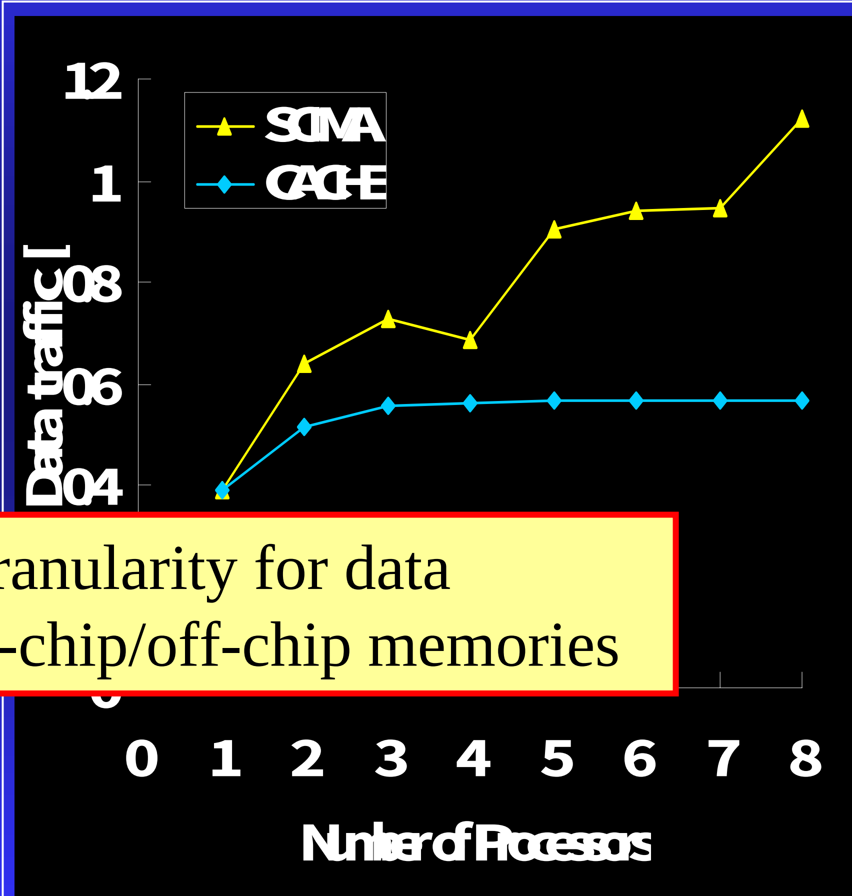
- Address space is partially mapped onto On-Chip Memory (exclusive among processors)
- Memory access management system is slightly modified for SMP

Scalability on number of processors on SMP
(Matrix-Matrix Multiplication Benchmark)

Speed-up Ratio



Bus traffic per cycle

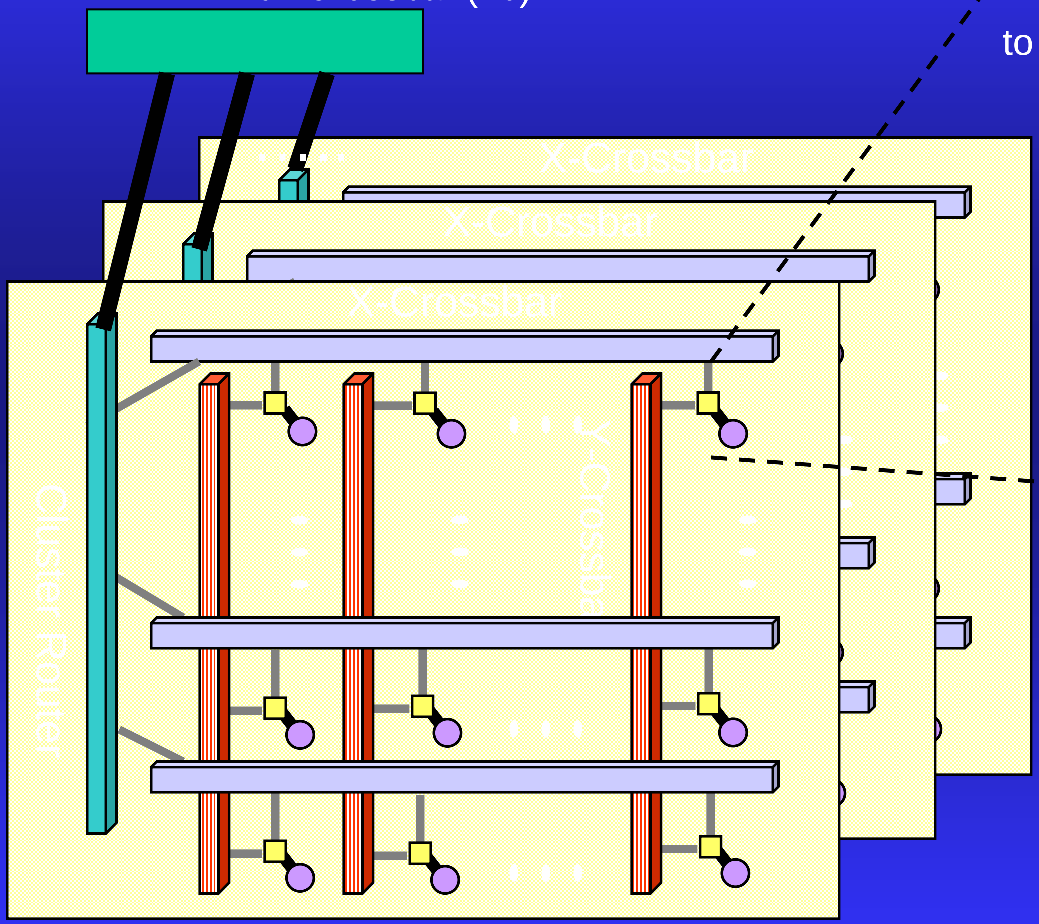


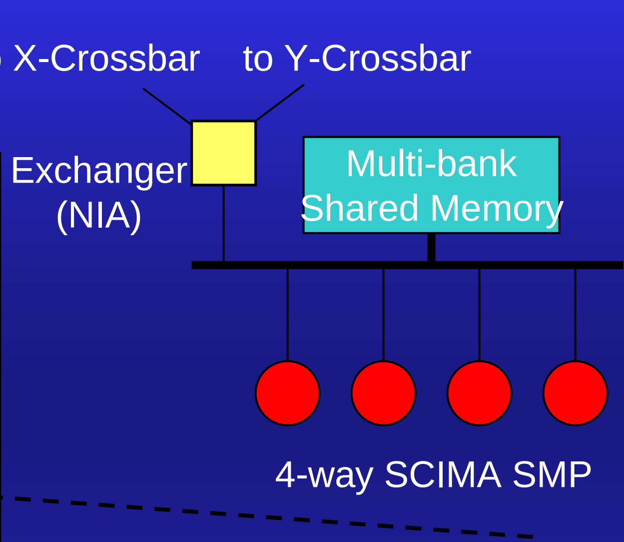
SCIMA obtains best granularity for data movement between on-chip/off-chip memories

N = 300, Bus band width: 4 [byte/cycle],
Off-Chip memory access latency: 40[cycle]

Network and Total System Image

High bandwidth Optical Link & Switch
(Cluster Link & Switch)
Full Crossbar (16)





Low Latency Electrical Link & Switch
(Intra-Cluster Link & Switch)
2-D Hyper Crossbar (16x16)

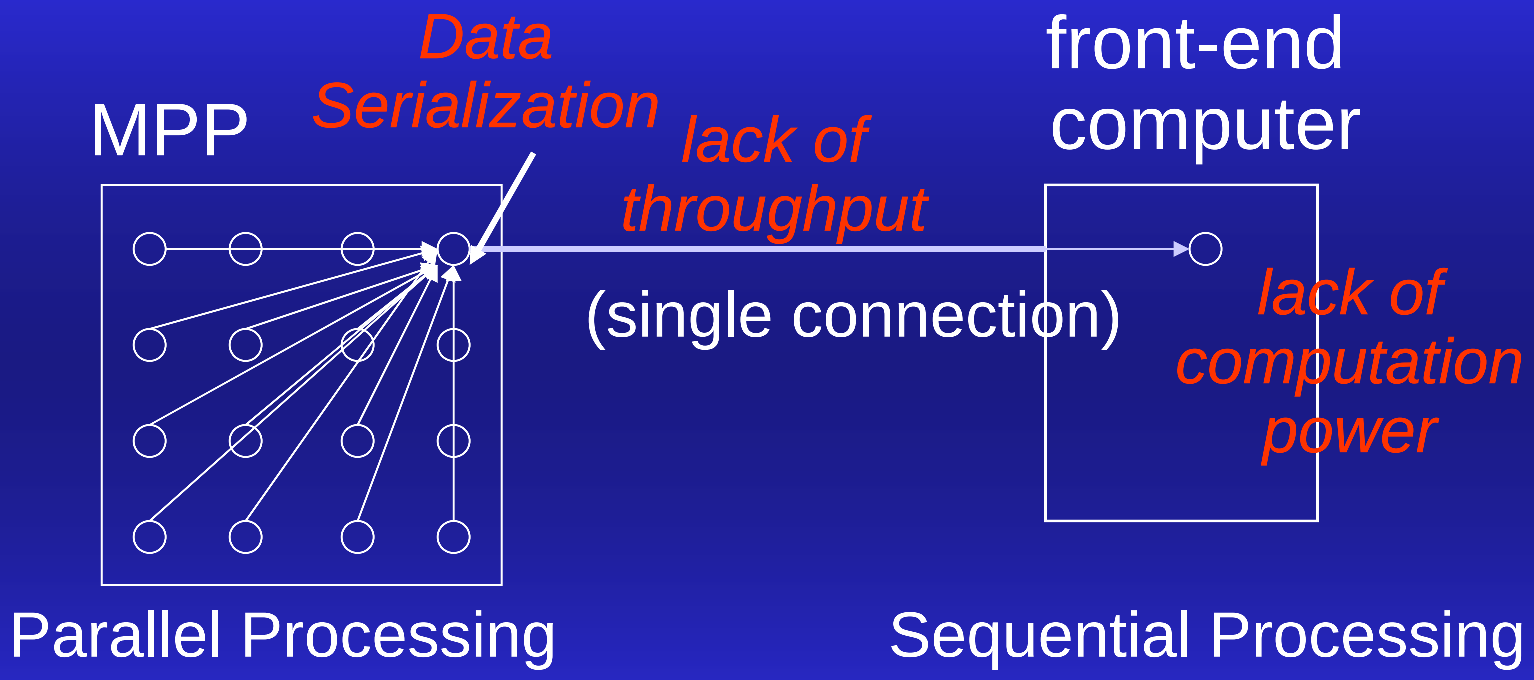
2 GHz SCIMA
(4 operations / clock)
× 4 = 32 GFLOPS / node

32 GFLOPS × (16 x 16) x 16
= 131 TFLOPS

Parallel I/O environement PAVEMENT

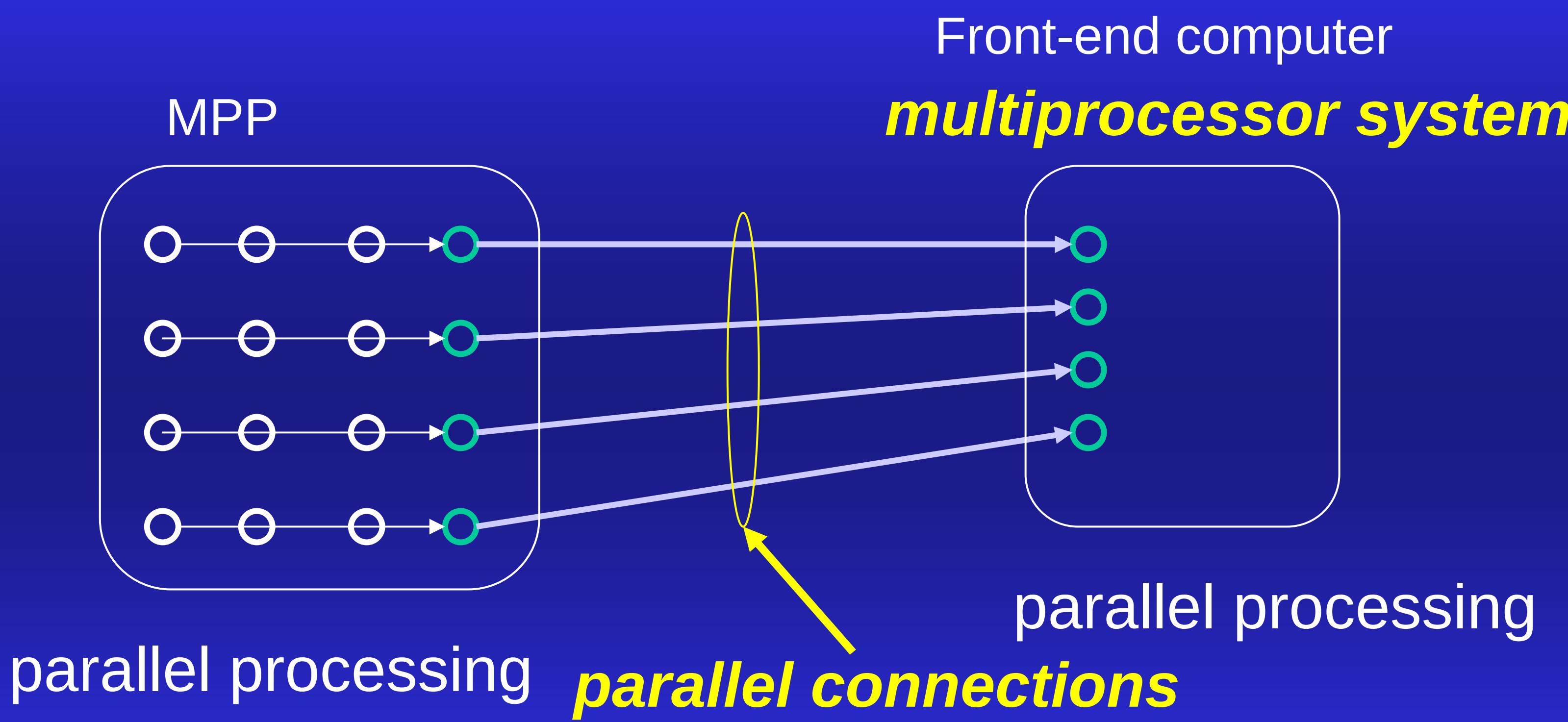
- Data I/O and visualization issues
- PIO (parallel I/O system)
- PFS (parallel file system)
- VIZ (parallel visualizer)

Problem of conventional I/O and visualization



single connection and/or sequential processing limits the overall performance

Goal of parallel I/O and visualization

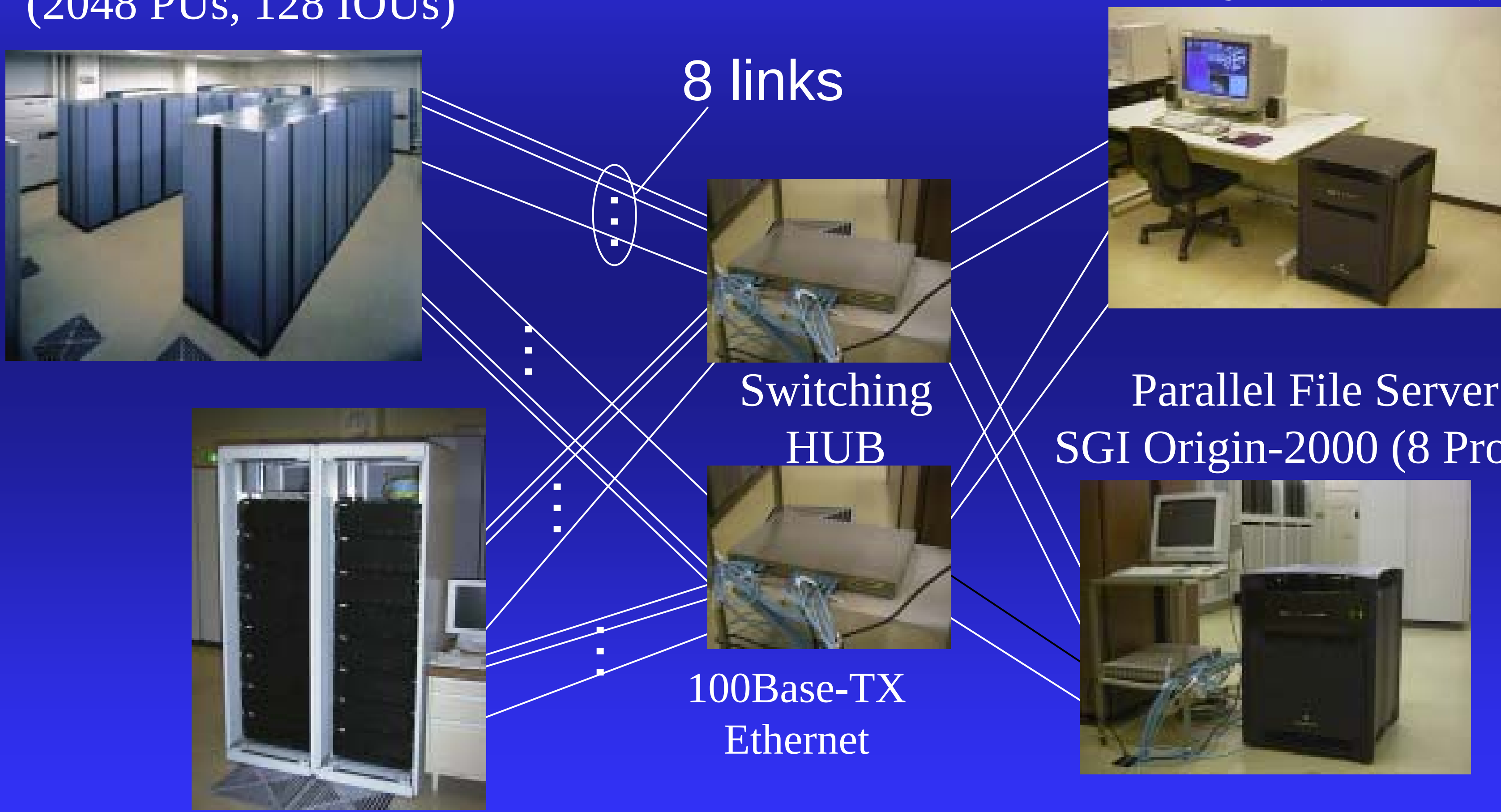


NO serialization during the whole computation

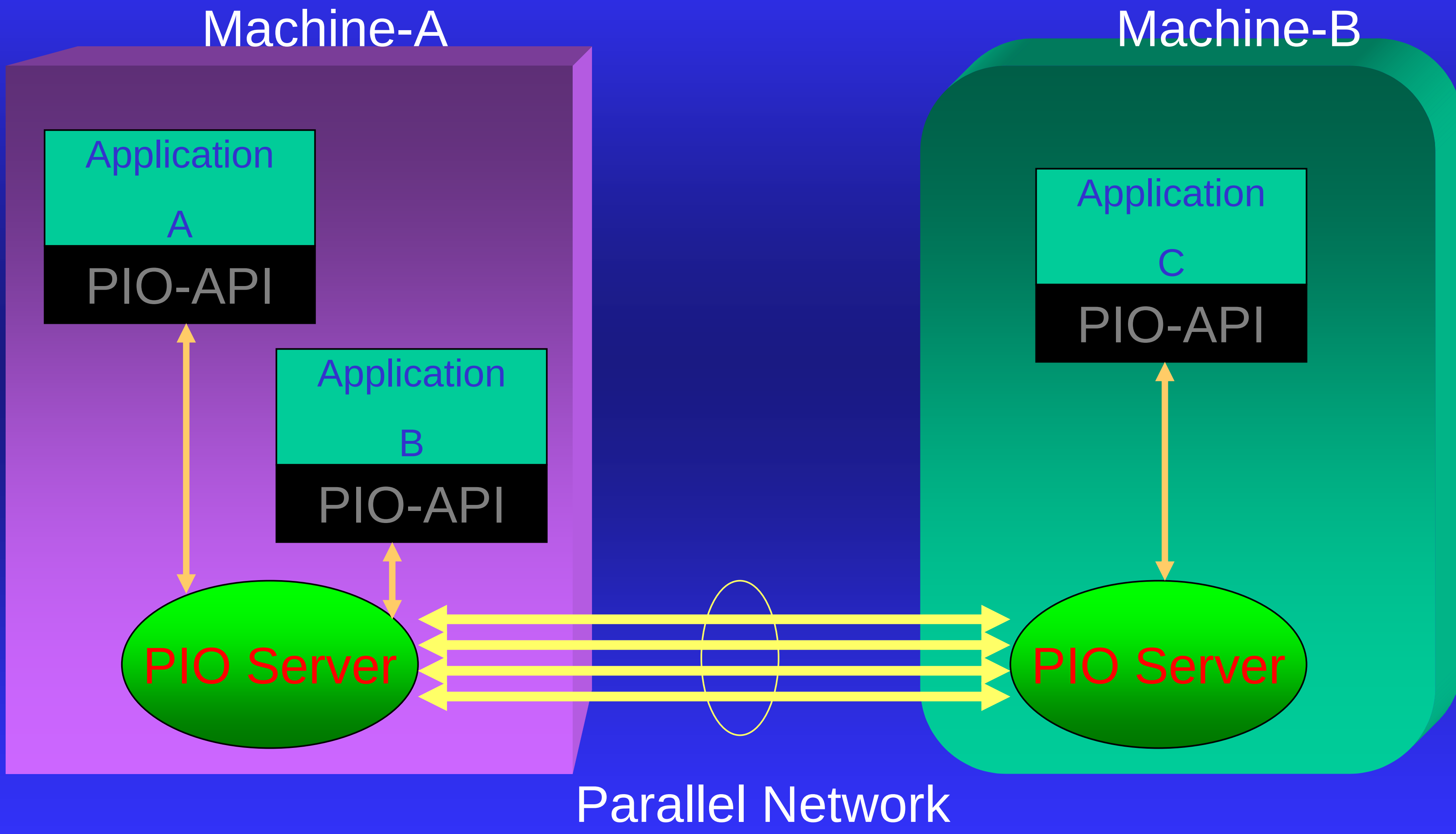
Design Strategy of Parallel I/O System

- Parallel Processing both in Network and Front-end System
  - making full use of parallel I/O processors in MPP
  - free from bottleneck caused by serialization
- Commodity-based Network Interface
  - 100base-TX, Gigabit Ethernet, etc.
  - widely used protocol (TCP/IP) → plenty of portability
  - high-performance and inexpensive I/O system
- dynamic load balancing

Experimental Environment



System Image



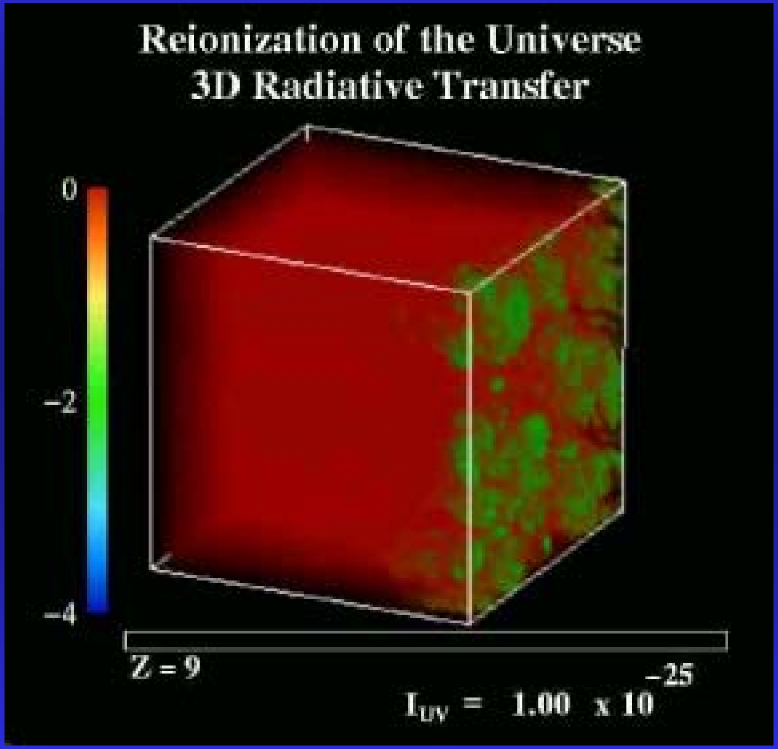
PAVEMENT/VIZ

- Parallelized 3-D volume rendering module
- Extension modules for AVS/Express (defacto standard visualization environment);

  compatible user interface with original module
- Use PAVEMENT/PIO for high throughput parallel data streaming

Development partly with Kubota Graphics Technologies
To be included in their free software package

Example of PAVEMENT/PIO and VIZ at work



Reionization of the Universe

CP-PACS 2048PU 128IOU
(614GFLOPS for calculation)
+
16 channels
(For I/O)
+
Origin2000 8CPU
(For volume rendering)

⇓

Real-Time Visualization

Heterogeneous Multi-Computer System (HMCS)

with J. Makino and T. Fukushige

- Motivations
- Concept
- Prototype with CP-PACS/GRAPE-6
- Application to galaxy formation

Multi-Scale Physics Simulation

Simulation of complex phenomena involving

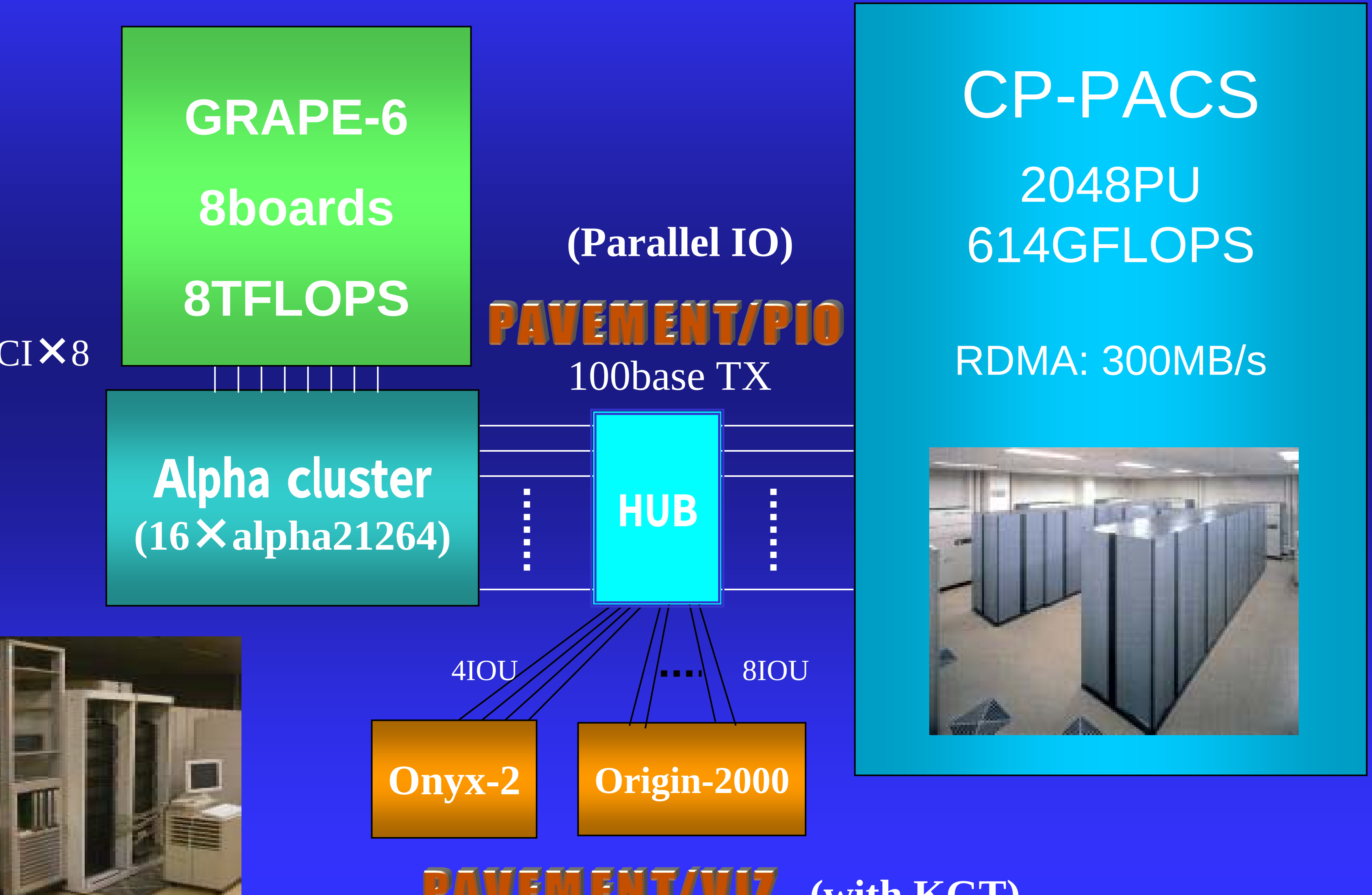
- Multiple types of interactions
  - Classical (gravity, electromagnetizm, …)
  - Quantum Dynamics, …
- Short- and long-ranged Interactions
- Difference in Computation Order
  - $O(N^2)$ ex. gravitational force
  - $O(N \log N)$ ex. FFT
  - $O(N)$ ex. straight-CFD

Concept of HMCS

– Heterogeneous Multi-Computer System –

- *Combining Particle Simulation (ex: Gravity interaction) and Continuum Simulation (ex: SPH) in a Platform*
- *Combining General Purpose Processor (flexibility) and Special Purpose Processor (high-speed)*
- *Connecting General Purpose MPP and Special Purpose MPP through High-throughput parallel Network*

Prototype HMCS SYSTEM



g6cpplib

API to access GRAPE-6 from CP-PACS

- `g6cpp_start(myid, nio, mode, error)`
- `g6cpp_unit(n, t_unit, x_unit, eps2, error)`
- `g6cpp_calc(mass, r, f_old, phi_old, error)`
- `g6cpp_wait(acc, pot, error)`
- `g6cpp_end(error)`

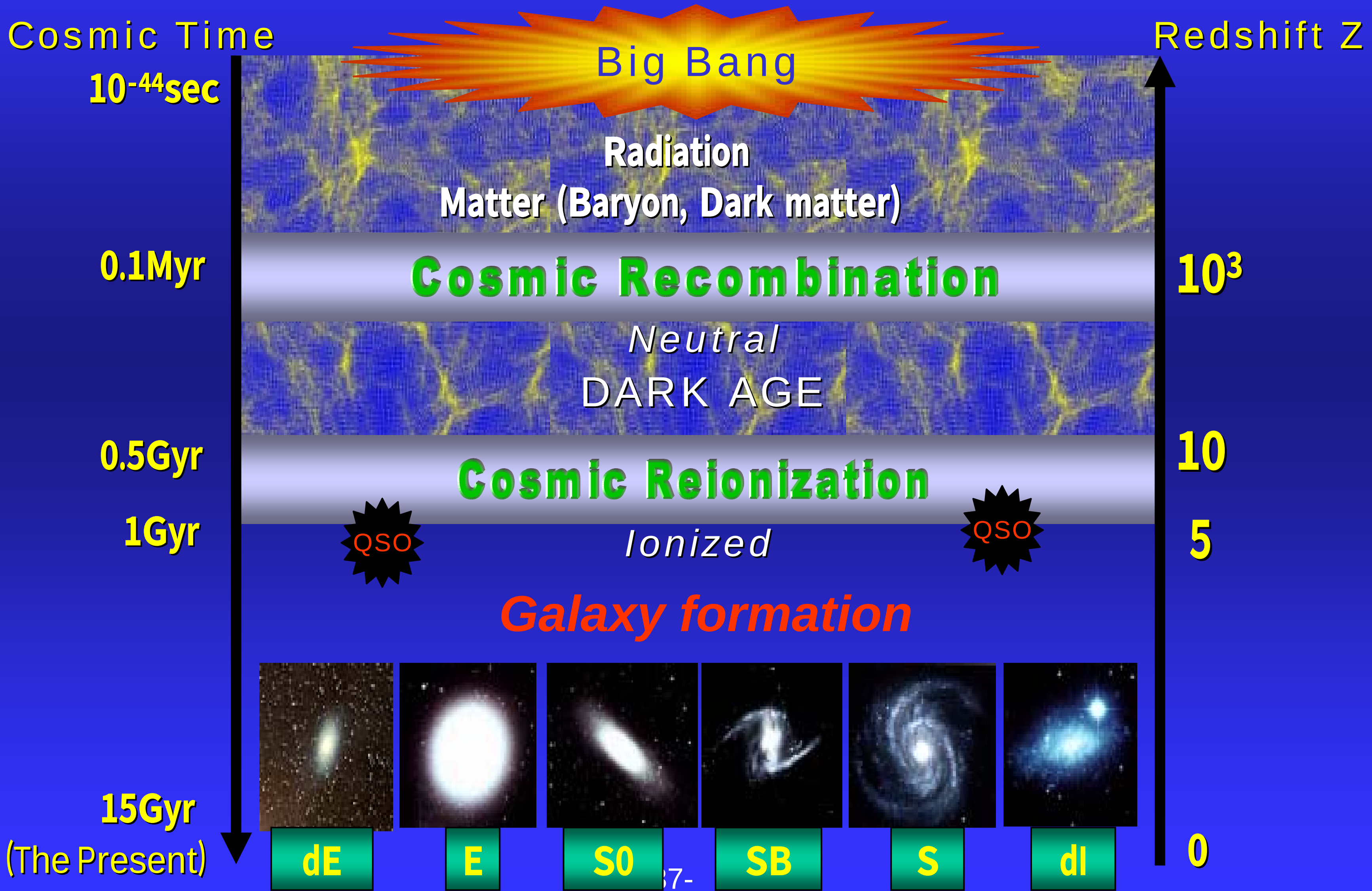
Breakdown after sent to GRAPE-6 cluster

(sec)

# of particles	8K	16K	32K	64K	128K
communication	0.695	1.647	2.139	2.478	4.952
all-to-all	1.012	1.436	1.786	2.249	2.357
set j-particle	0.139	0.165	0.229	0.357	0.610
calculation	0.012	0.029	0.063	0.158	0.440

GRAPE-6 takes just 1 sec for 128K particles

Galaxy formation on HMCS



Galaxy formation under UV radiation

Hydrodynamic motion of matter
+ Gravity acting on matter
+ Radiative transfer

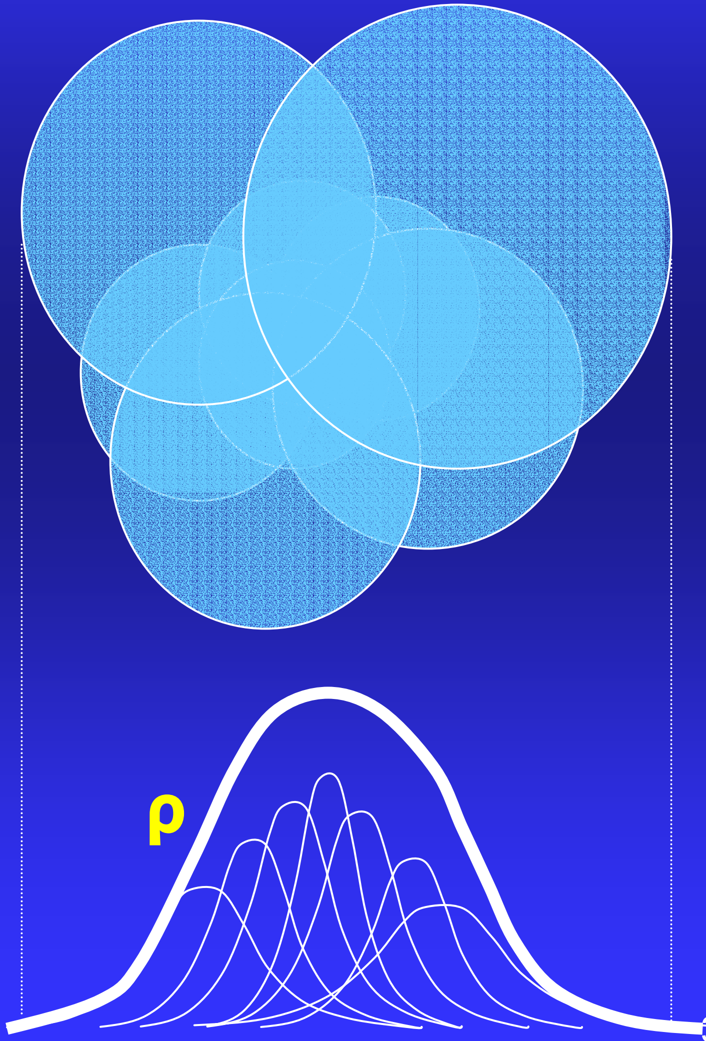
1. **Hydrodynamics:**
   Smoothed Particle Hydrodynamics (SPH) method is employed.

2. **Self-gravity:**
   Barnes-Hut Tree is effective for serial calculation, but difficult to parallelize.
   → Direct summation is made by GRAPE-6.

3. **Chemical reaction & radiative cooling are included.**

4. **Radiative transfer (RT):**
   RT is solved with a method by Kessel-Deynet &Burkert.

SPH
(Smoothed Particle Hydrodynamics)



Matter density represented as a collection of particles

$$\rho(\mathbf{r}_i) = \sum_j \rho_{j0} W(|\mathbf{r}_i - \mathbf{r}_j|)$$

W : kernel function

Radiative Transfer Equation

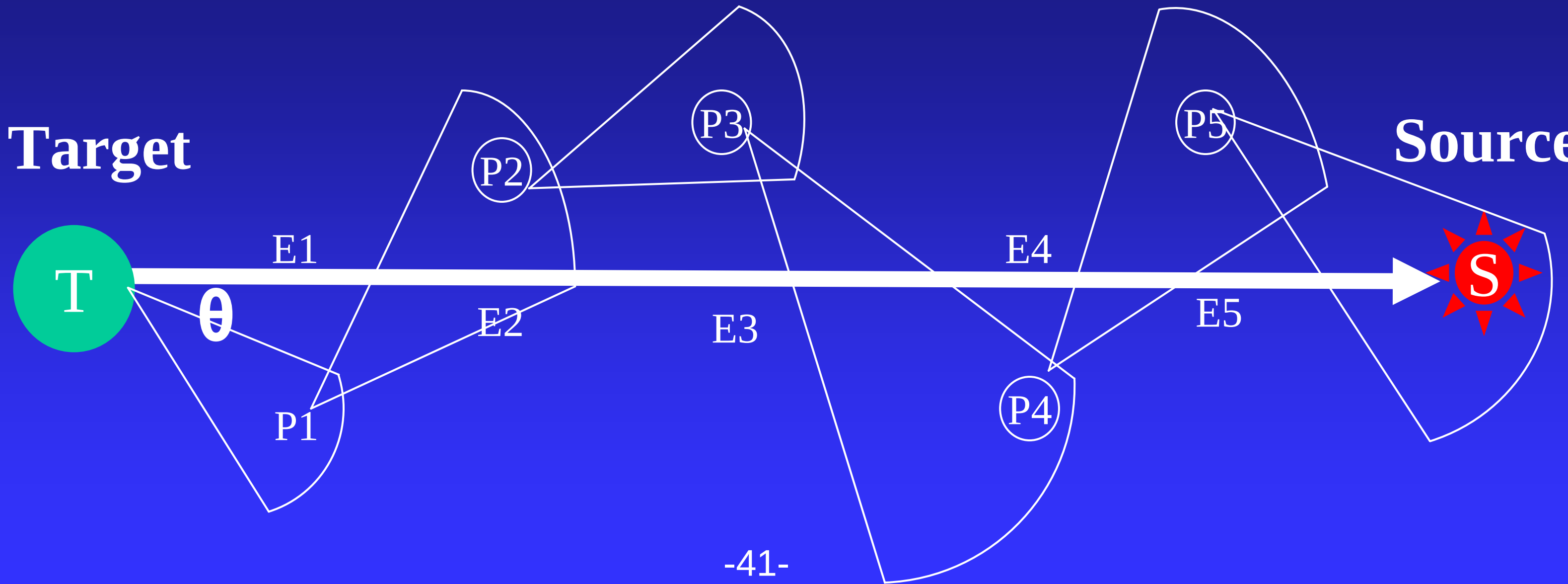
$$\frac{1}{c}\frac{\partial I_\nu}{\partial t} + \mathbf{n}\cdot\nabla I_\nu = \chi_\nu (S_\nu - I_\nu)$$

- Boltzmann equation for photon distribution function
- 6D calculation, hence computationally heavy
  (3D for space + 2D for ray direction + 1D for frequency)

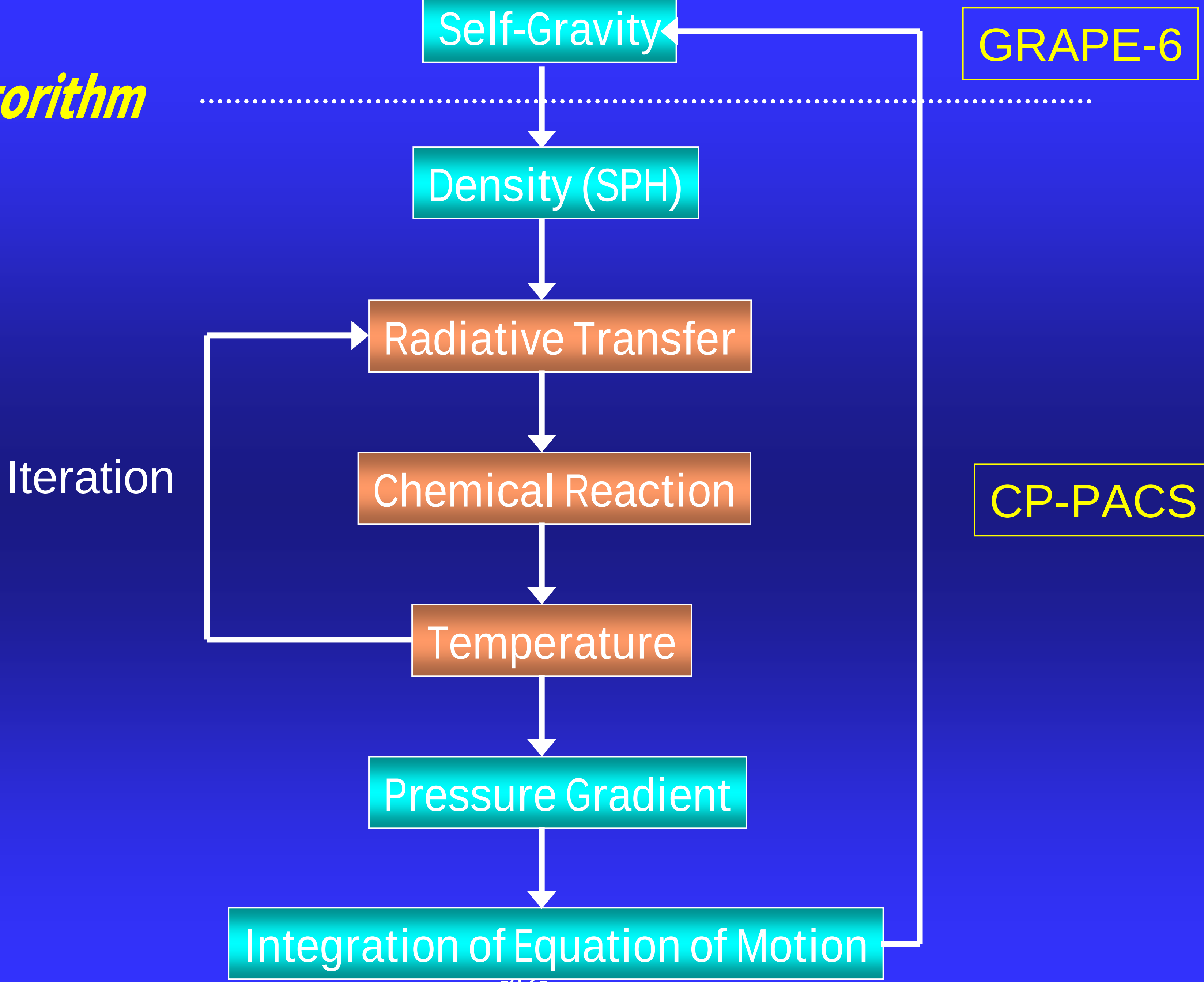
Radiative Transfer for SPH

- Accurate calculation of optical depth along light paths required.
- Use the method by Kessel-Deynet & Burkert (2000) .

$$\tau_{TS} = \sum_i \frac{\sigma}{2}\left(n_{E_i} + n_{E_{i+1}}\right)\left(s_{E_{i+1}} - s_{E_i}\right)$$



Algorithm


Self-Gravity
GRAPE-6
Density (SPH)
Radiative Transfer
Iteration
Chemical Reaction
CP-PACS
Temperature
Pressure Gradient
Integration of Equation of Motion

Calculation parameters

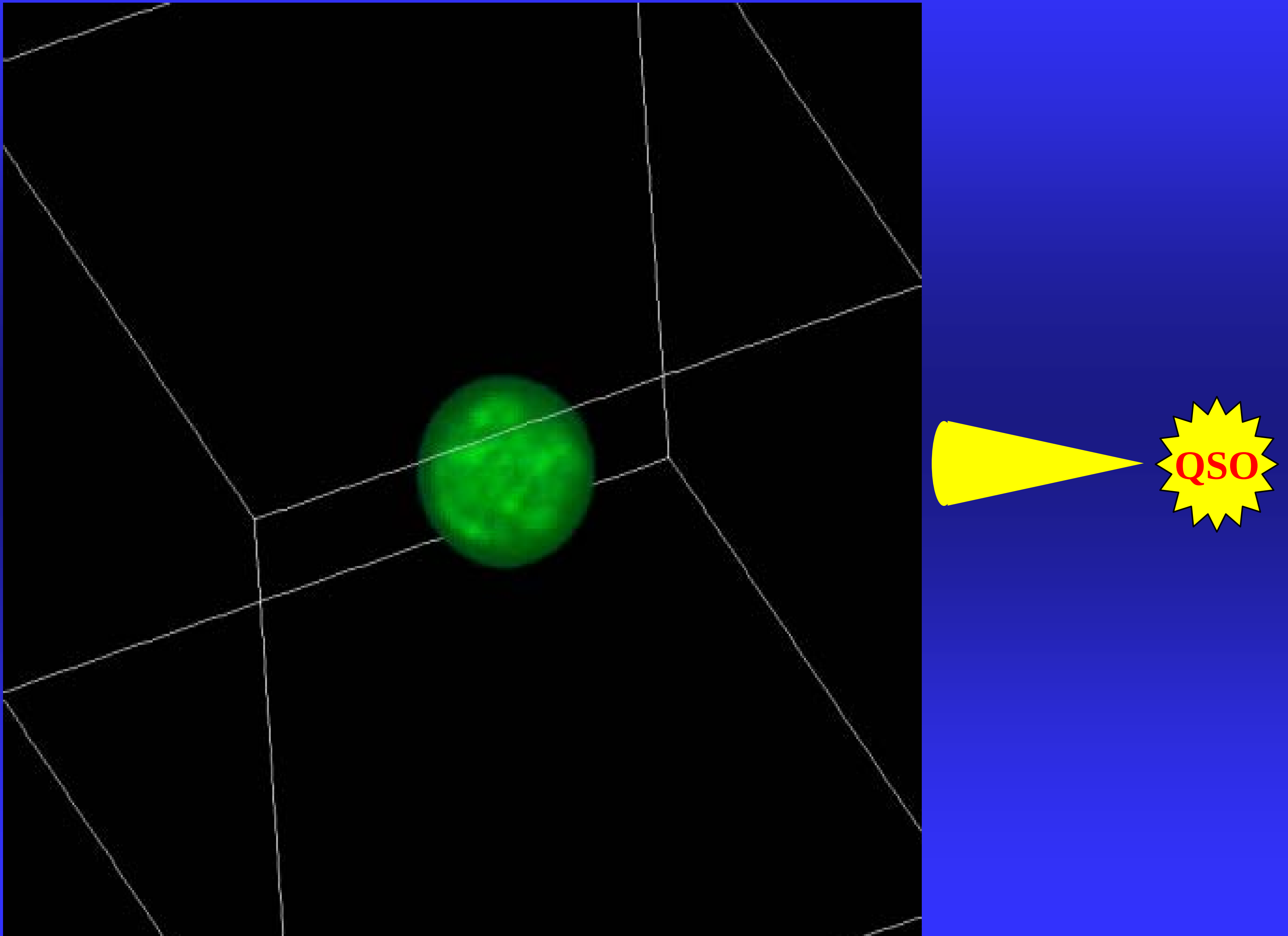
- *1024PU of CP-PACS 307GFLOPS*
  4 boards of GRAPE-6 4TFLOPS

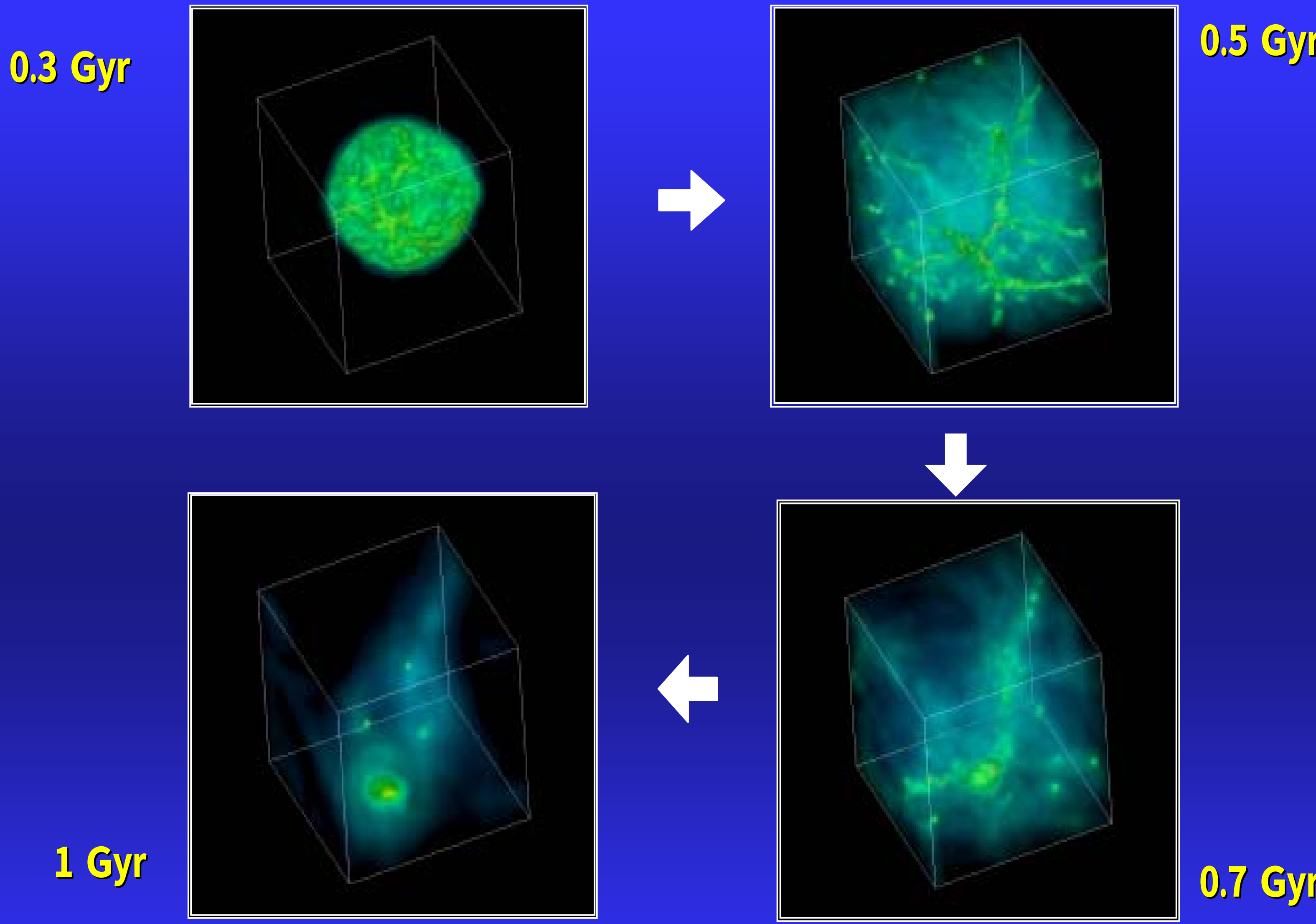
- *64K baryonic matter particles*
  + 64K dark matter particles

- *4 sec for calculation on CP-PACS*
  3 sec for communication to/from GRAPE-6
  0.1 sec for calculation on GRAPE-6

total 7.1 sec/time step

Galaxy Formation under UV Background





At the initial stage, a density fluctuation is generated to match a cold dark matter spectrum. This fluctuation expands with the cosmic expansion, and simultaneously smaller-scale density fluctuations develop inside to form filamentary structures. Tiny filaments evaporate due to the heating by background UV radiation, whereas larger filaments shrink to coalesce into a condensed rotating cloud. This rotating cloud would evolve into a galaxy. This simulation has revealed that the background UV radiation plays an important role for the final

Summary and Prospect

- *Carried out R&D of key element technologies for HPC of continuum physical systems :*
  - *processor architecture SCIMA*
  - *interconnect and total system design*
  - *parallel I/O and visualization environment*

- *Floating point power alone will be insufficient to process complex multi-scale simulations:*
  - *both continuum and particle degrees of freedom*
  - *both short- and long-ranged interactions*
  - *multiple of scales*

Summary and Prospect II

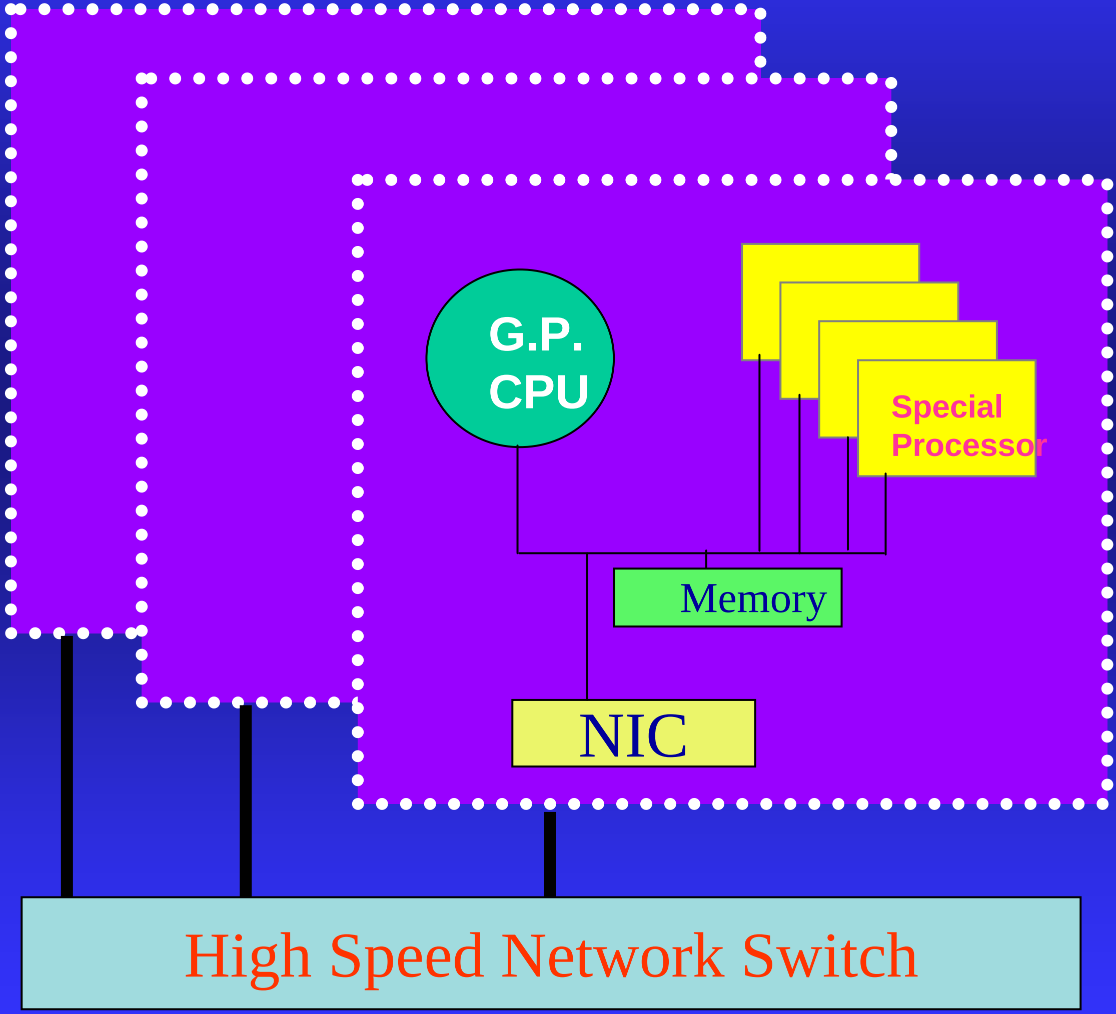
- *Proposed HMCS as a paradigm to treat such systems :*
  - *combines high flexibility of general-purpose systems and high performance of special-purpose systems, distributing the computation load in a best way possible to each sub-system*
  - *built a prototype HMCS and demonstrated the effectiveness of the concept with a novel galaxy formation simulation in astrophysics*

Summary and Prospect III

- *further development of the HMCS concept:*
  - *HMCS-R: remote general/special systems connected through high-speed network (e.g., superSINET)*
  - *HMCS-E: embed special-purpose processors in the node of general purpose systems*

→ *Will provide an ideal platform for next generation of large-scale scientific simulations of complex phenomena*

HMCS-E (Embedded)



- general and special purpose processor unified in each node
- Ideal combination of flexibility and high-performance