QBiC

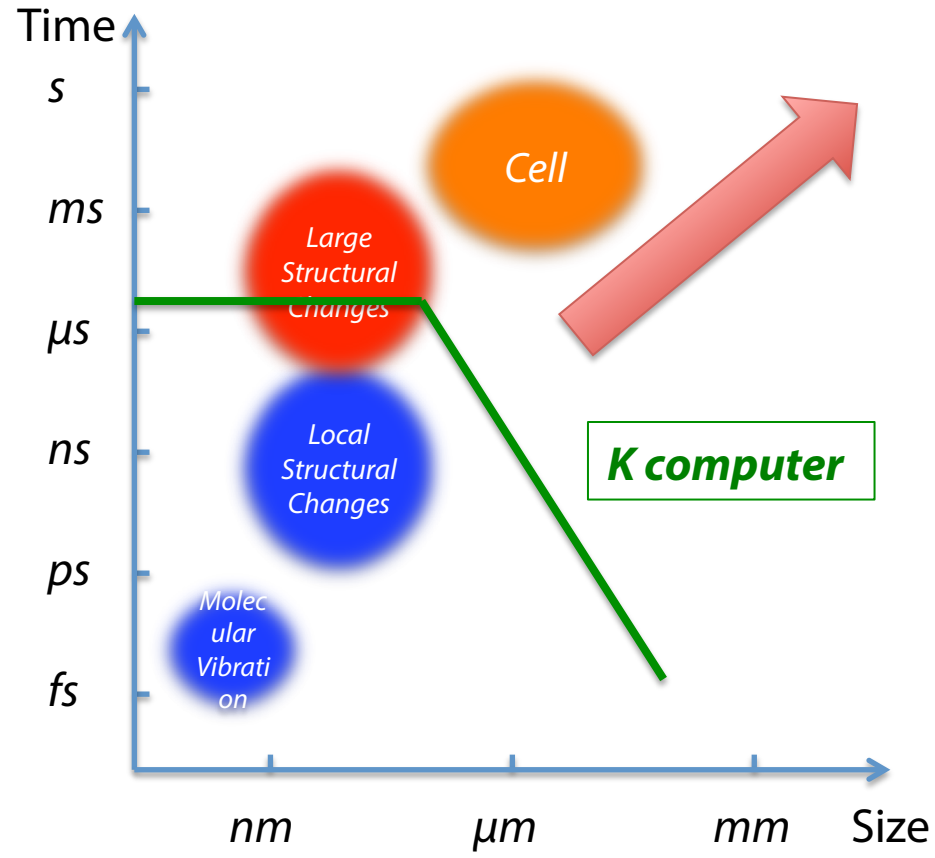
次世代MD専用計算機 MDGRAPE-4開発と今後の展開

泰地 真弘人

理化学研究所
生命システム研究センター 副センター長
主任研究員

taiji@riken.jp

Challenges in Molecular Dynamics simulations of biomolecules



- Timestep
  ～fsec
- Target timescale
  μsec～sec

To cover 10^9～10^{15} timescale difference

QBiC

Scaling challenges in MD

- ～50,000 FLOP/particle/step
- Typical system size : $N=10^5$
- 5 GFLOP/step
- 5TFLOPS effective performance

  1msec/step = 170nsec/day

  Rather Easy
- 5PFLOPS effective performance

  1μsec/step = 200μsec/day???

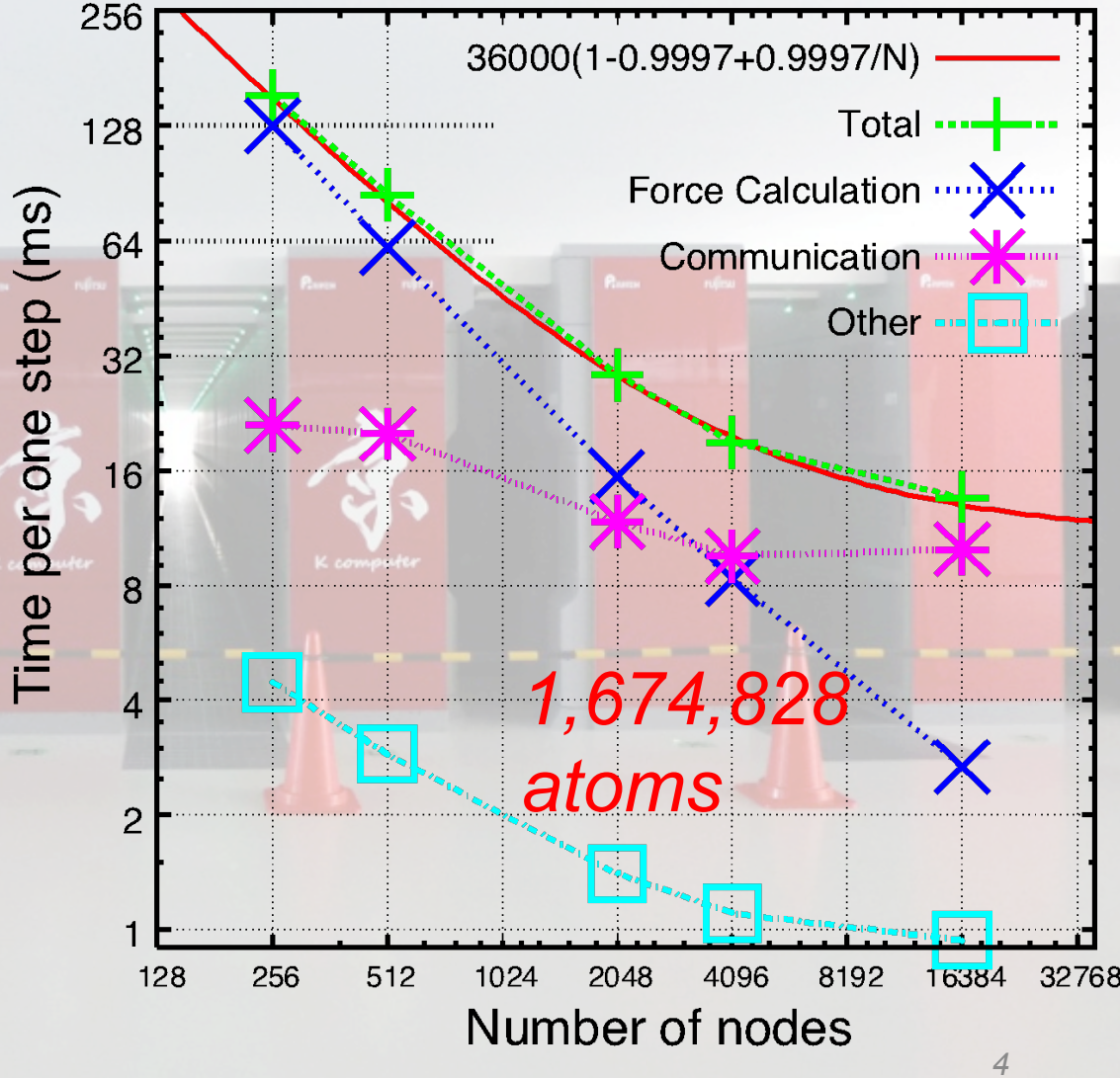
  Difficult, but important

QBiC

Scaling of MD on K Computer

Strong scaling
～50 atoms/core

~3M atoms/Pflops



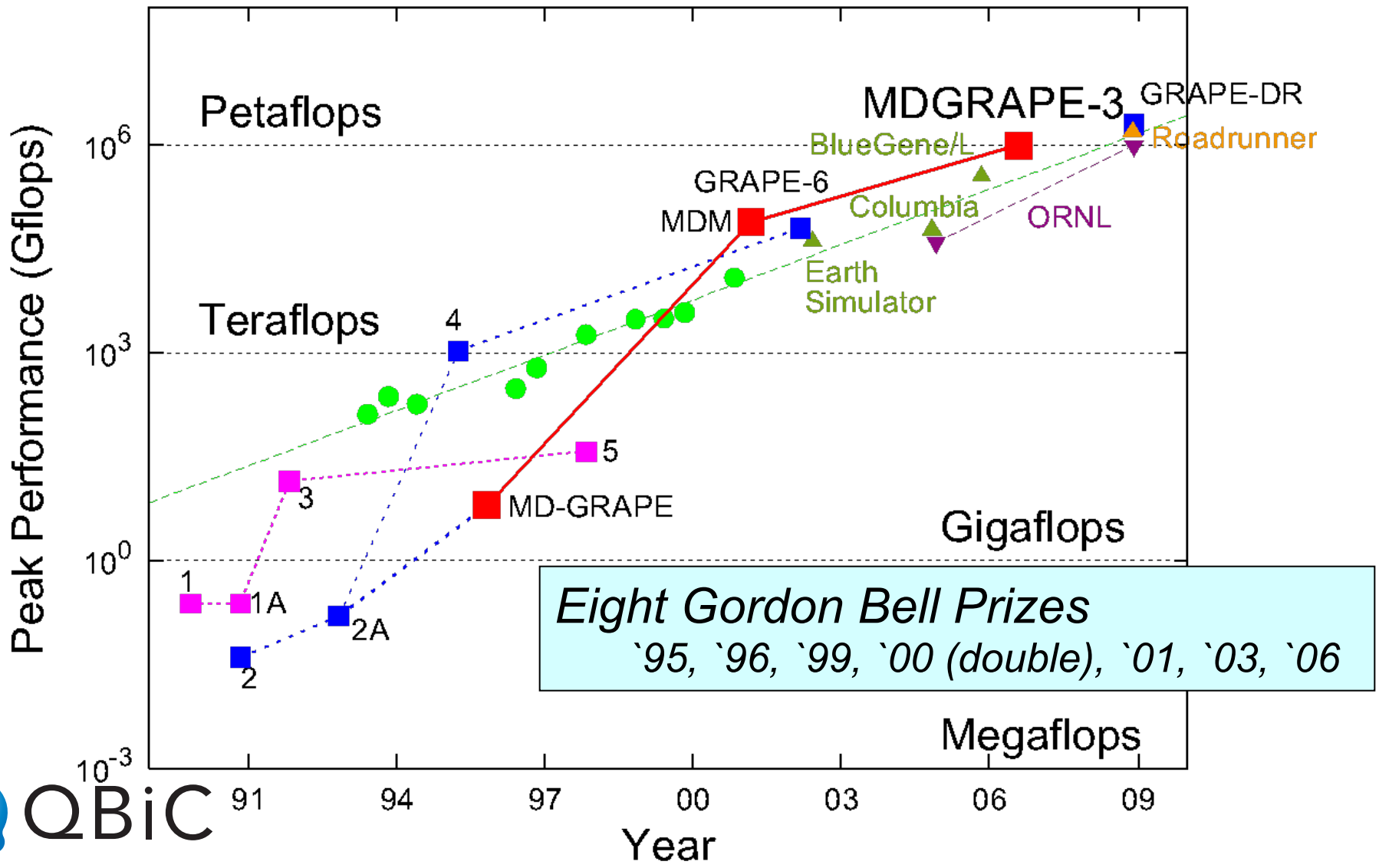
What is GRAPE?

- GRAvity PipE
- Special-purpose accelerator for classical particle simulations
  - Astrophysical *N*-body simulations
  - Molecular Dynamics Simulations

J. Makino & M. Taiji, Scientific Simulations with Special-Purpose Computers, John Wiley & Sons, 1997.

QBiC

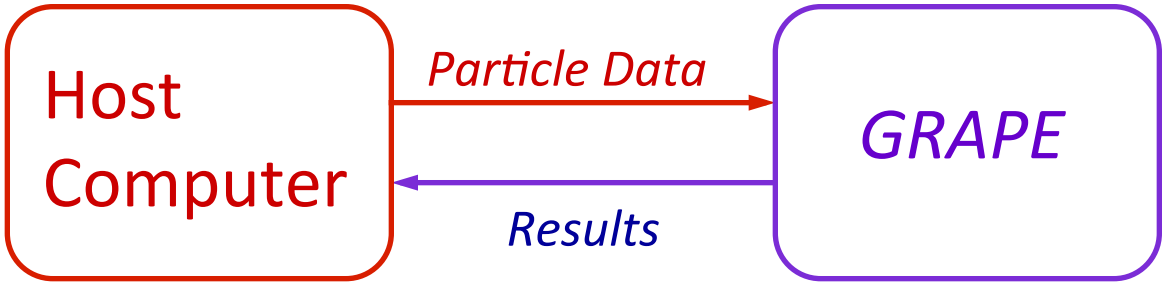
History of GRAPE computers



QBiC

GRAPE as Accelerator

- Accelerator to calculate forces by dedicated pipelines



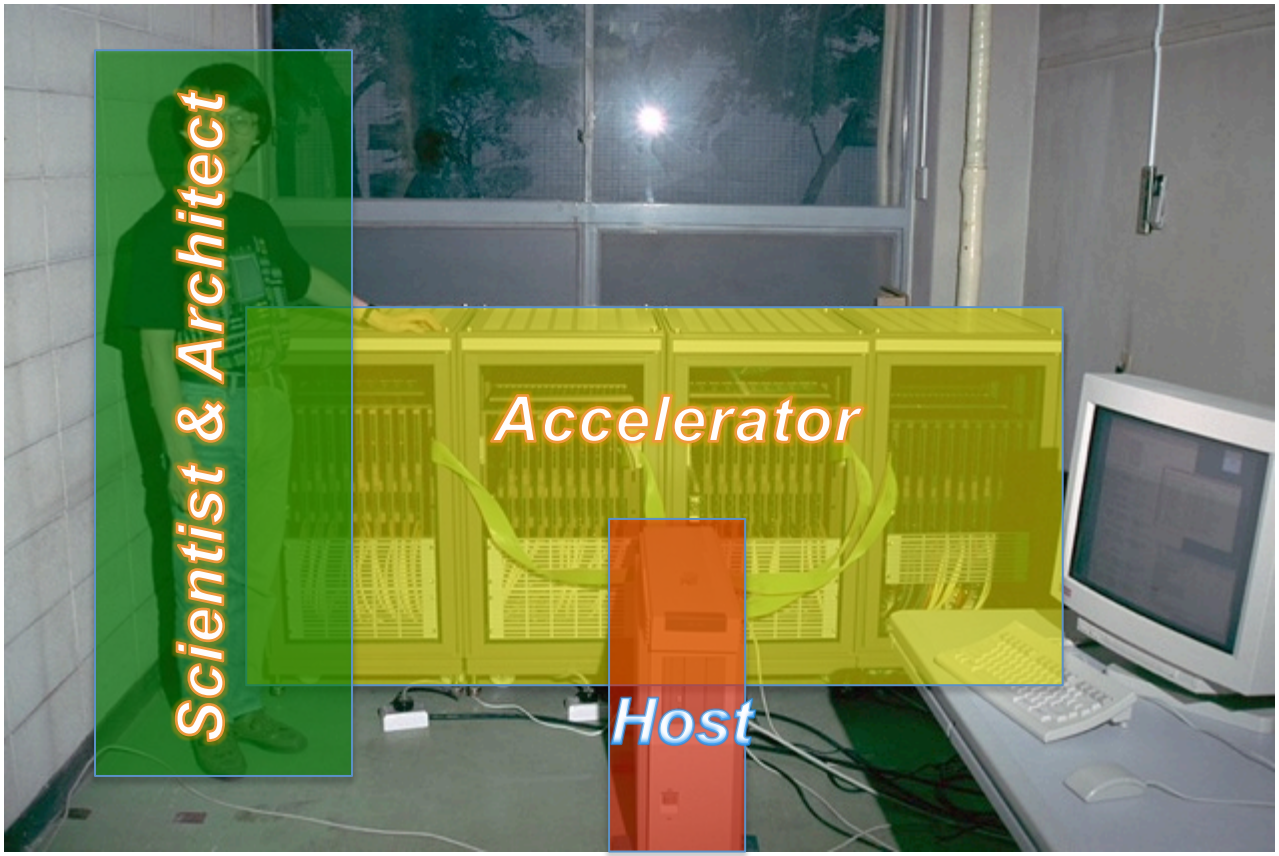
Most of Calculation → *GRAPE*
Others → *Host computer*

- *Communication = O (N) << Calculation = O (N^2)*
- *Easy to build, Easy to use*
- *Cost Effective*

QBiC

GRAPE in 1990s

- GRAPE-4(1995): The first Teraflops machine



Host CPU
~ 0.6 Gflops

Accelerator PU
~ 0.6 Gflops

Host:
Single or SMP

QBiC

GRAPE in 2000s

- MDGRAPE-3: The first petaflops machine

Host CPU
~ 20 Gflops

Accelerator PU
~ 200 Gflops

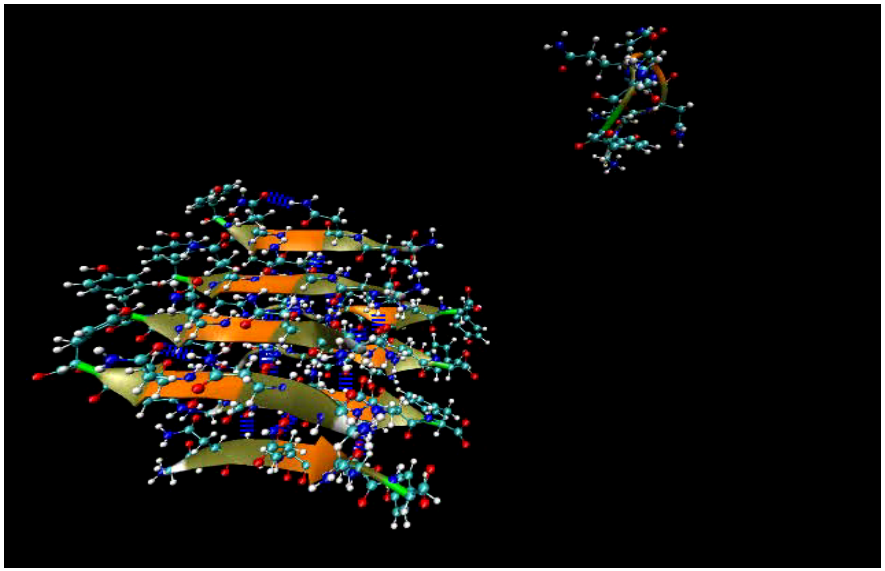
Host:
Cluster

QBiC

Scientists not shown but exist

Sustained Performance of Parallel System (MDGRAPE-3)

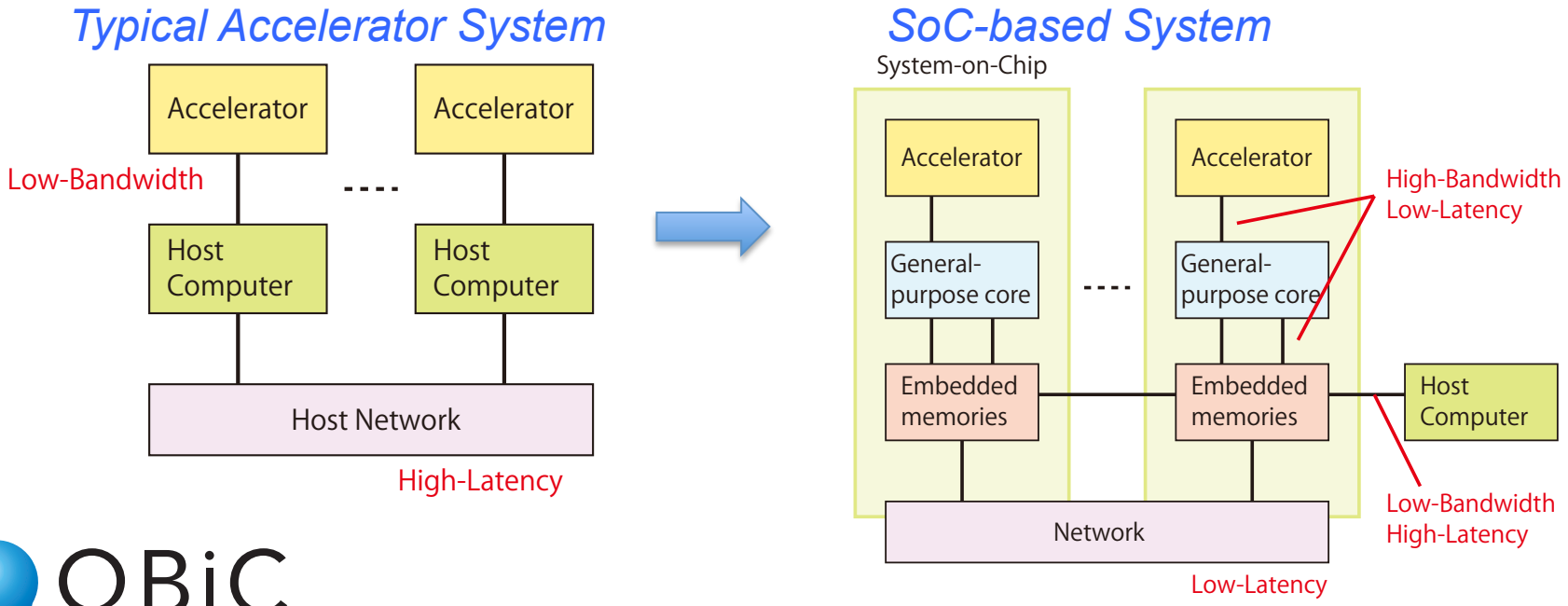
- Gordon Bell 2006 Honorable Mention, Peak Performance
- Amyloid forming process of Yeast Sup 35 peptides
- Systems with 17 million atoms
- Cutoff simulations (Rcut = 45 Å)
- 0.55sec/step
- Sustained performance:
  185 Tflops
- Efficiency ~ 45 %
- ~million atoms necessary
  for petascale



QBiC

Problem in Heterogeneous System - GRAPE/GPUs

- In small system
  - Good acceleration, High performance/cost
- In massively-parallel system
  - Scaling is often limited by host-host network, host-accelerator interface



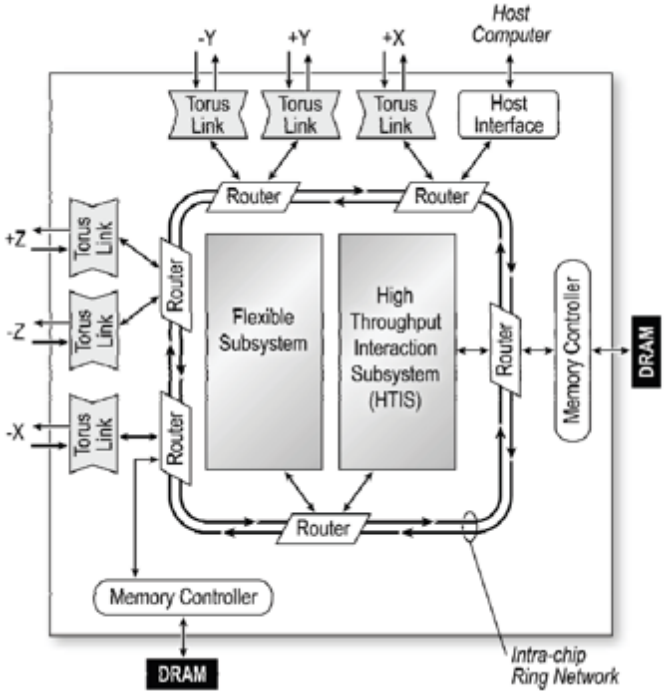
QBiC

Anton

- D. E. Shaw Research
- Special-purpose pipeline

  + General-purpose CPU core

  + Specialized network
- Anton showed the importance of the optimization in communication system



	GROMACS time		Anton time	
	small cutoff (9Å) large mesh (64^3)	large cutoff (13Å) small mesh (32^3)	small cutoff (9Å) large mesh (64^3)	large cutoff (13Å) small mesh (32^3)
Nonbonded forces				
Range-limited forces	111 ms (61%)	308 ms (88%)	1.8 μs (3%)	3 μs (13%)
FFT & inverse FFT	29 ms (16%)	3 ms (1%)	38 μs (66%)	12 μs (50%)
Mesh interpolation	19 ms (10%)	18 ms (5%)	10 μs (17%)	5.5 μs (23%)
Correction forces	7 ms (4%)	6 ms (2%)	2 μs (3%)	2.5 μs (10%)
Bonded forces	9 ms (5%)	9 ms (2%)	5 μs (9%)	5 μs (21%)
Integration	7 ms (4%)	7 ms (2%)	3 μs (5%)	2.5 μs (10%)
Total	181 ms (100%)	351 ms (100%)	58 μs (100%)	24 μs (100%)

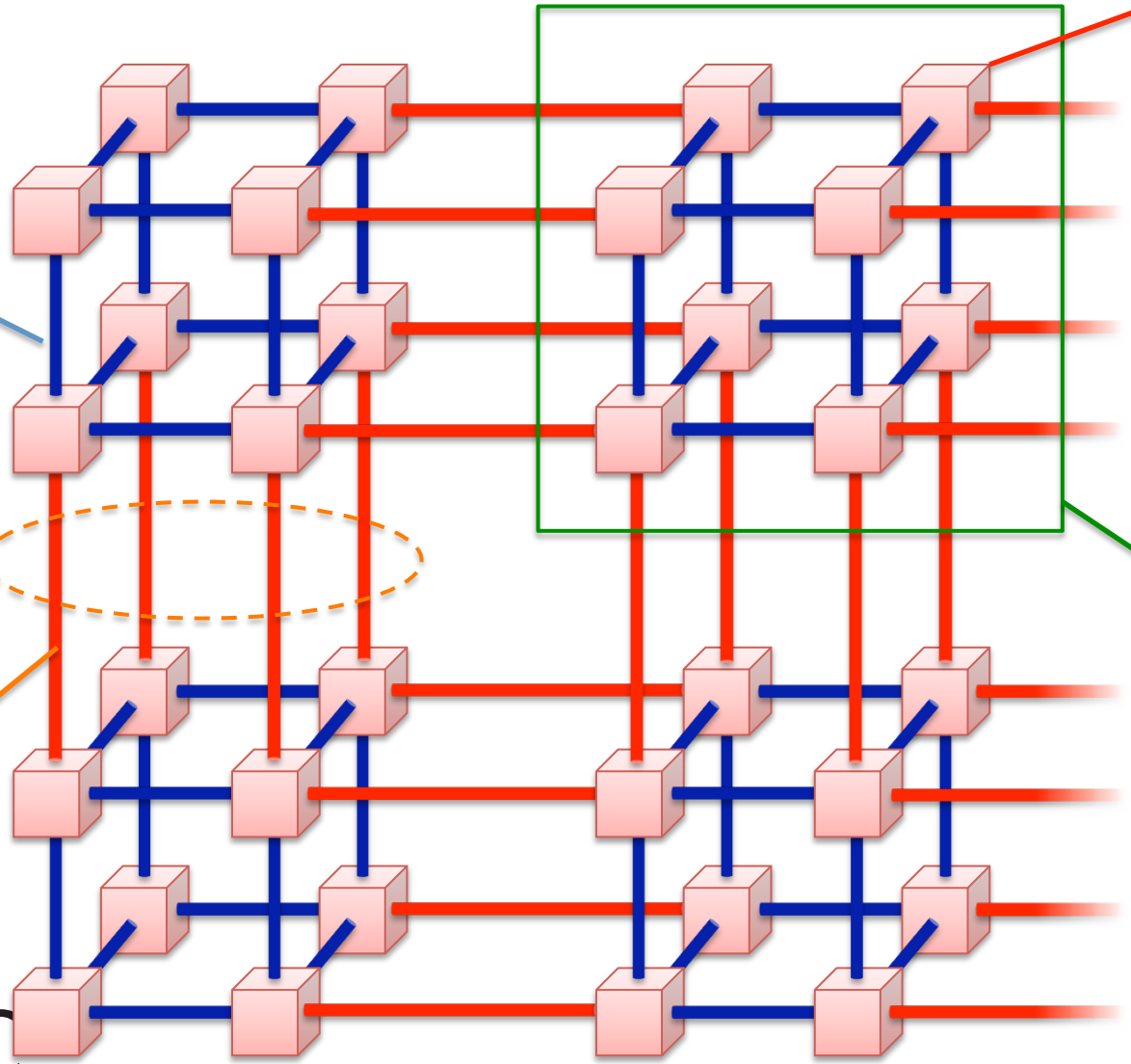
R. O. Dror et al., Proc. Supercomputing 2009, in USB memory.

MDGRAPE-4

- Special-purpose computer for MD simulation
- Test platform for special-purpose machine
- Target performance
  - 20µsec/step for 100K atom system
  - 8.6µsec/day (2fsec/step)
- Target application : GROMACS
- Completion: 1Q 2014
- Enhancement from MDGRAPE-3
  - 130nm → 40nm process
  - **SoC Integration of Network / CPU**
  - Keep system as simple as possible

QBiC

MDGRAPE-4 System



MDGRAPE-4 SoC

Total 512 chips (8x8x8)

12 lane 6Gbps Electric = 7.2GB/s (after 8B10B encoding)

48 Optical Fibers

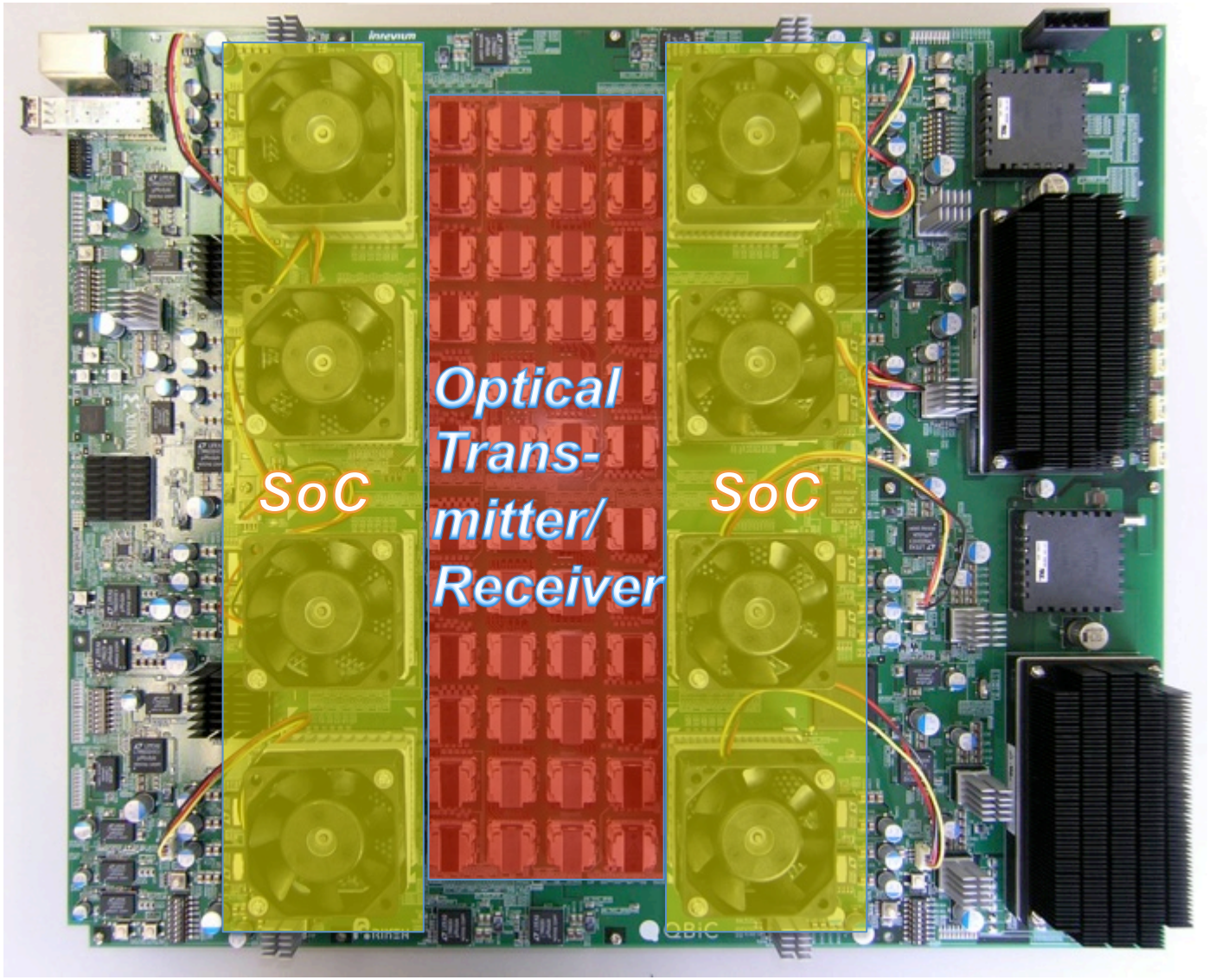
12 lane 6Gbps Optical

Node (2U Box)

Total 64 Nodes (4x4x4) =4 pedestals

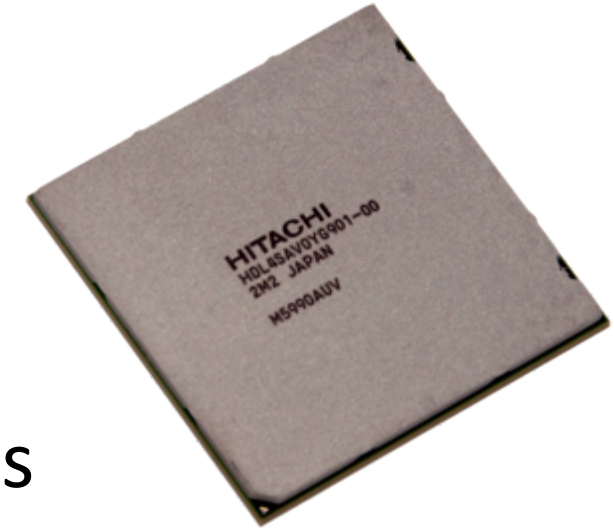
QBiC

MDGRAPE-4 Node



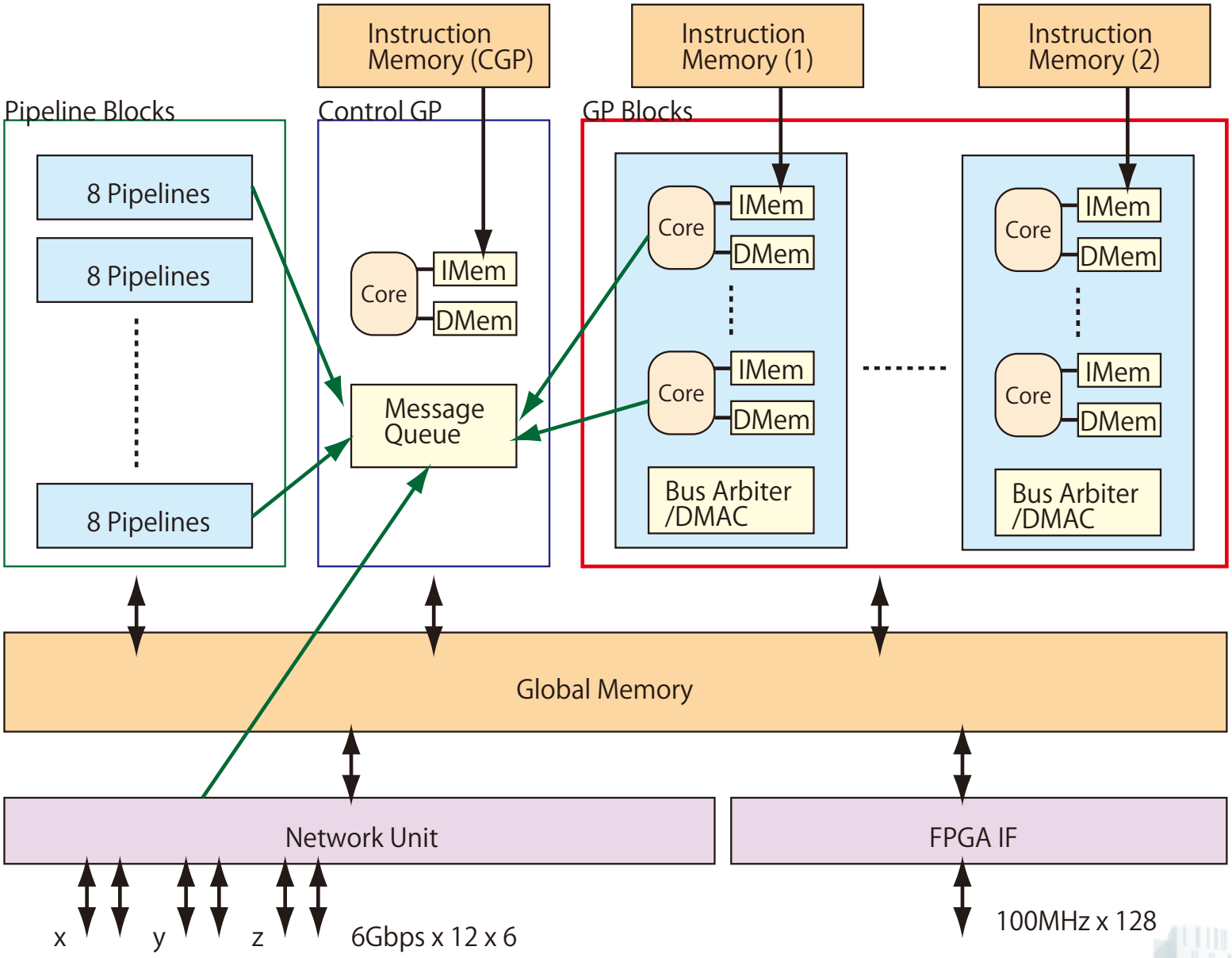
MDGRAPE-4 System-on-Chip

- 40 nm (Hitachi), ~ 230mm^2
- 64 force calculation pipelines
  @ 0.8GHz
- 64 general-purpose processors
  Tensilica Extensa LX4
  @0.6GHz
- 72 lane SERDES @6GHz

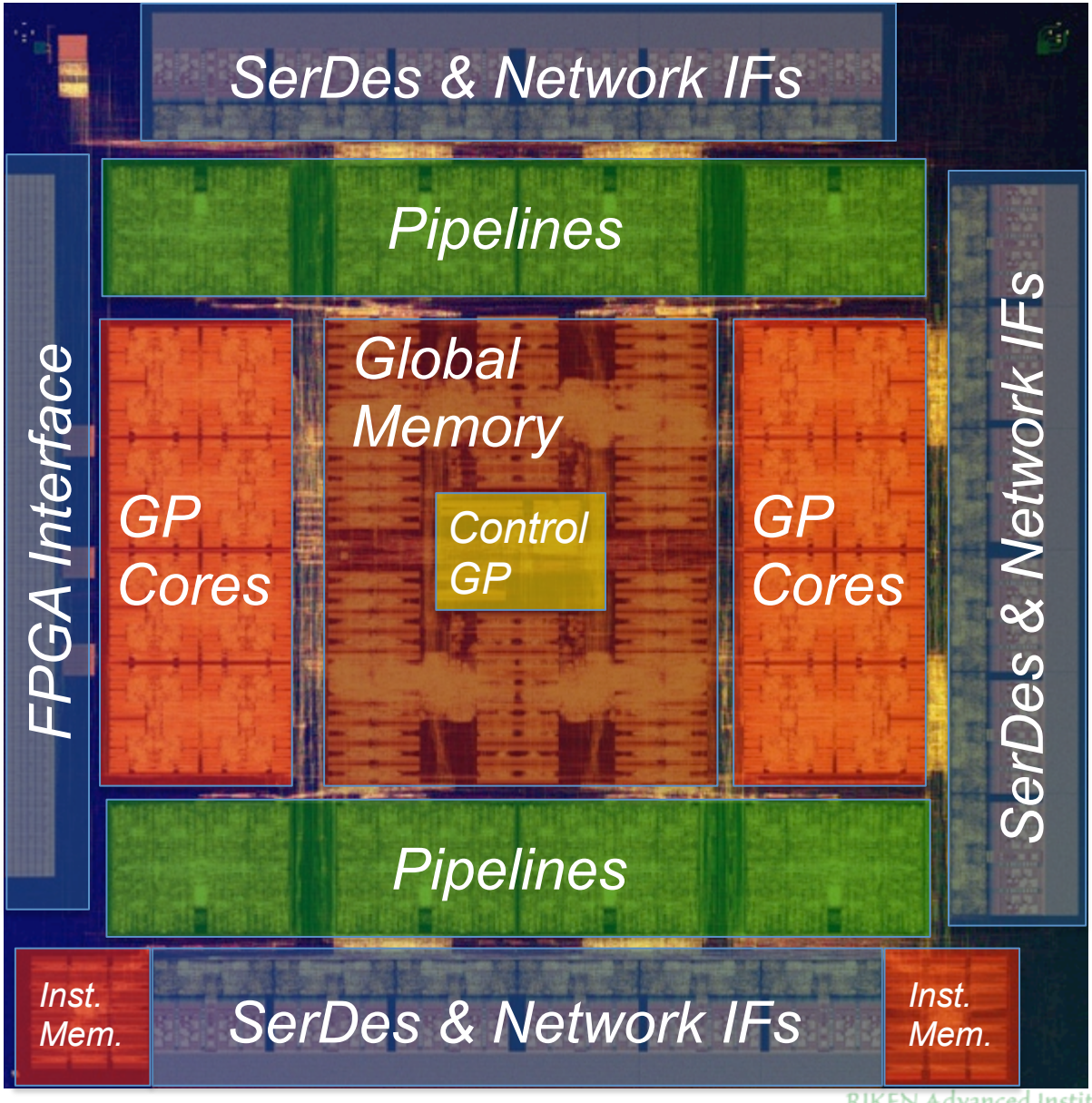


QBiC

SoC Block Diagram

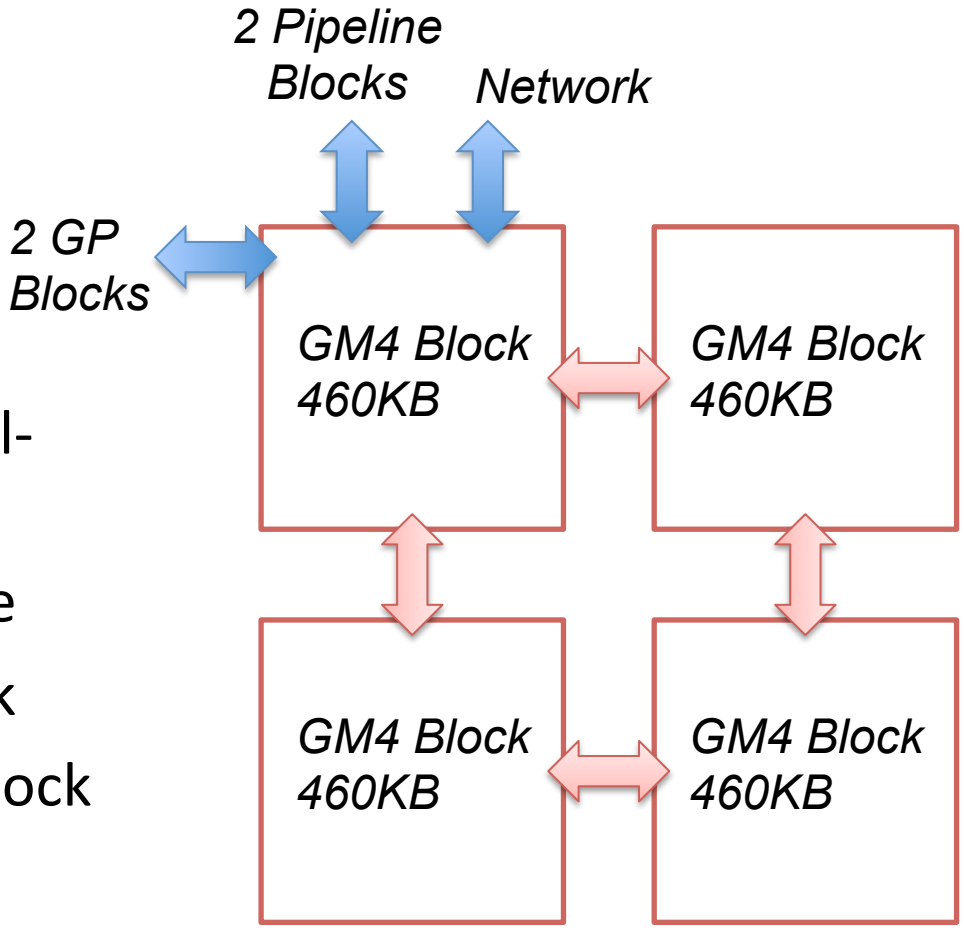


SoC Physical Image



Embedded Global Memories in SoC

- ~1.8MB
- 4 Block
- For Each Block
  - 128bit X 2 for General-purpose core
  - 192bit X 2 for Pipeline
  - 64 bit X 6 for Network
  - 256bit X 2 for Inter-block



QBiC

Parallelization of Force Calculation

- in small number of particles
- We have to explore many ways of parallelization

$$F_i = \sum_j f(r_i, r_j)$$

1. Index i GRAPE-3
2. Index j MDGRAPE-3
3. Calculations of f GRAPE-1

QBiC

i-Parallelization

$$F_i = \sum_j f(r_i, r_j)$$

- Parallel Calculation of F_i
- $O(N)$ Reads → $O(NN_c)$ calc → $O(N)$ writes
- Use of Action-Reaction Symmetry:

  $O(NN_c)$ writes with possible conflicts.

  Most of GPU implementation avoid this.
- $N_{proc} \gtrsim N \sim 10^{5\text{-}6}$

  Not enough!

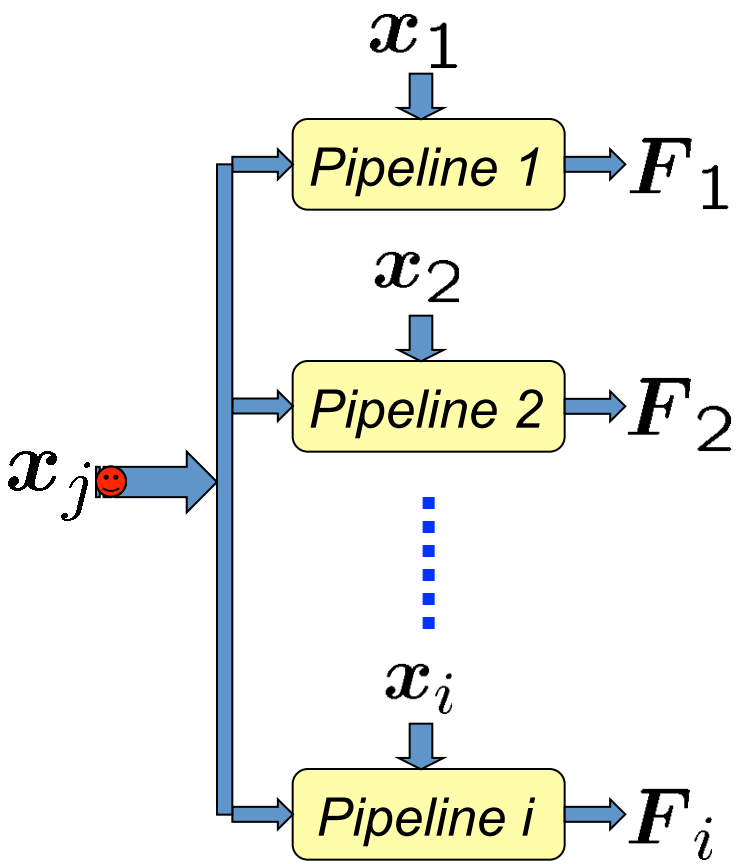
QBiC

Broadcast Parallelization

- In particle simulations
- Two-body forces

$$\boldsymbol{F}_i = \sum_j \boldsymbol{f}(x_i, x_j)$$

- For parallel calculation of $\boldsymbol{F}_i$, we can use the same $\boldsymbol{x}_j$



- Broadcast Parallelization
  - relax Bandwidth Problem

QBiC

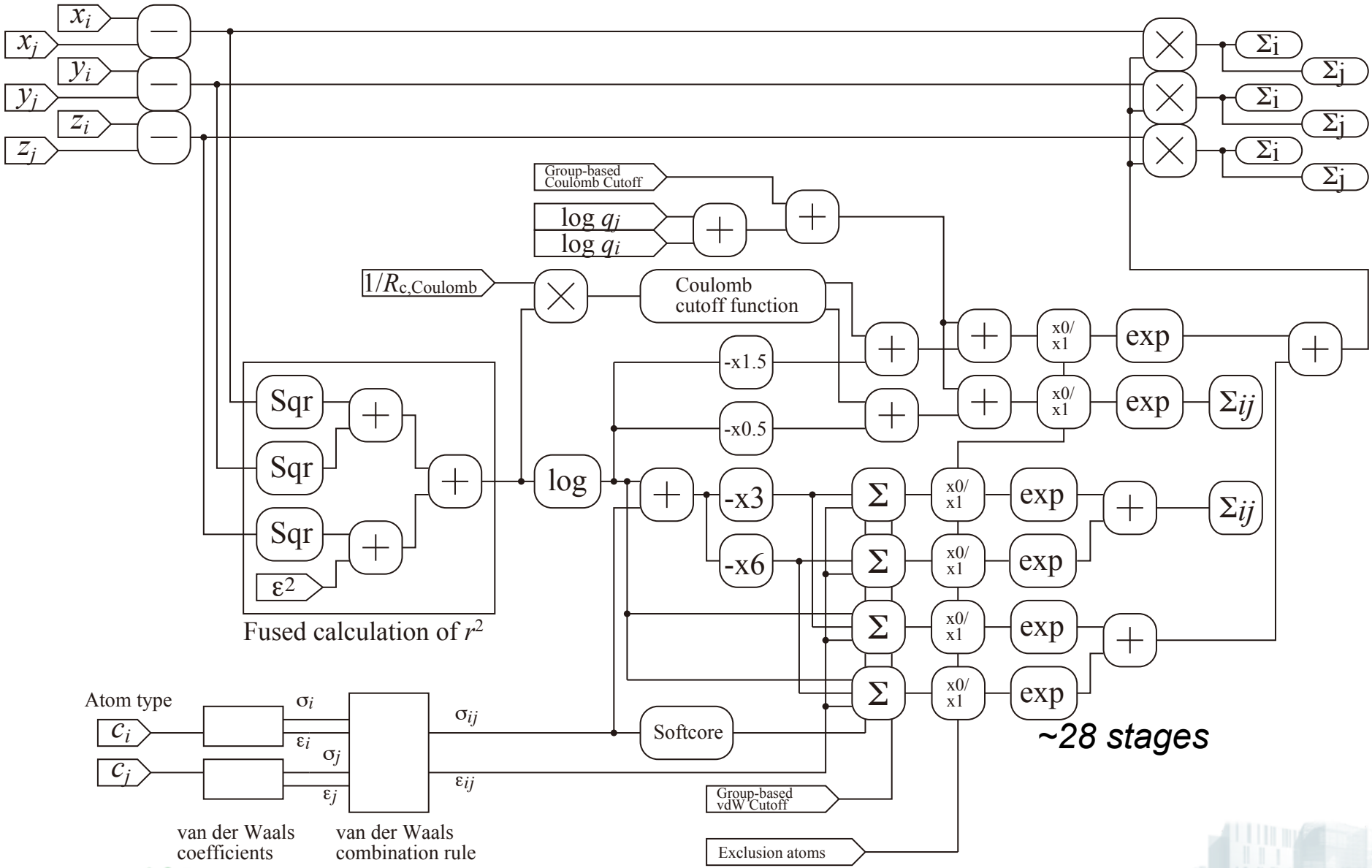
j-Parallelization

$$F_i = \sum_k \sum_{j=j0(k)}^{j1(k)} f_{ij}$$

- Parallel Calculation of $F_i^{(k)} = \sum_{j=j0(k)}^{j1(k)} f_{ij}$
- Reduction of $F_i^{(k)}$ is necessary
  - Fast reduction / synchronization is required
  - This kind of parallelization will be important

QBiC

Pipeline Block Diagram



Pipeline Functions

- Nonbond forces

$$\vec{f}_{ij} = \vec{r}_{ij} \cdot \left[\frac{q_i q_j}{r_{ij}^3} g_c(\alpha r_{ij}^2) + \frac{\epsilon_{ij}}{r_{ij}^2} \left\{ 12(r_{ij}/\sigma_{ij})^{-12} - 6(r_{ij}/\sigma_{ij})^{-6} \right\} \right]$$

$$\vec{F}_i = \sum_j \vec{f}_{ij}, \qquad \vec{F}_j = -\sum_i \vec{f}_{ij}$$

- and potentials

$$\phi_c = \sum_j \frac{q_i q_j}{r_{ij}} g_{c,\phi}(\alpha r_{ij}^2)$$

$$\phi_v = \sum_j \epsilon_{ij} \left\{ (r_{ij}/\sigma_{ij})^{-12} - (r_{ij}/\sigma_{ij})^{-6} \right\}$$

- Gaussian charge assignment & back interpolation
- Soft-core
- ~50G interactions/sec/chip

QBiC

Pipeline speed

- Tentative performance
  - 8x8 pipelines @0.8GHz(worst case)

    64x0.8=51.2 G interactions/sec

    512 chips = 26 T interactions/sec

    Lcell=Rc=12A, Half-shell ..2400 atoms

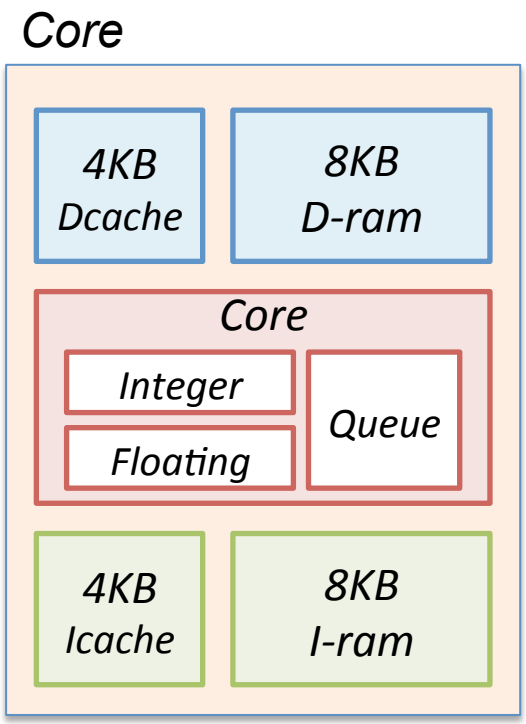
    10^5 X 2400/ 26T ~ 9.2 μsec
  - Flops count ~50 operations / pipeline

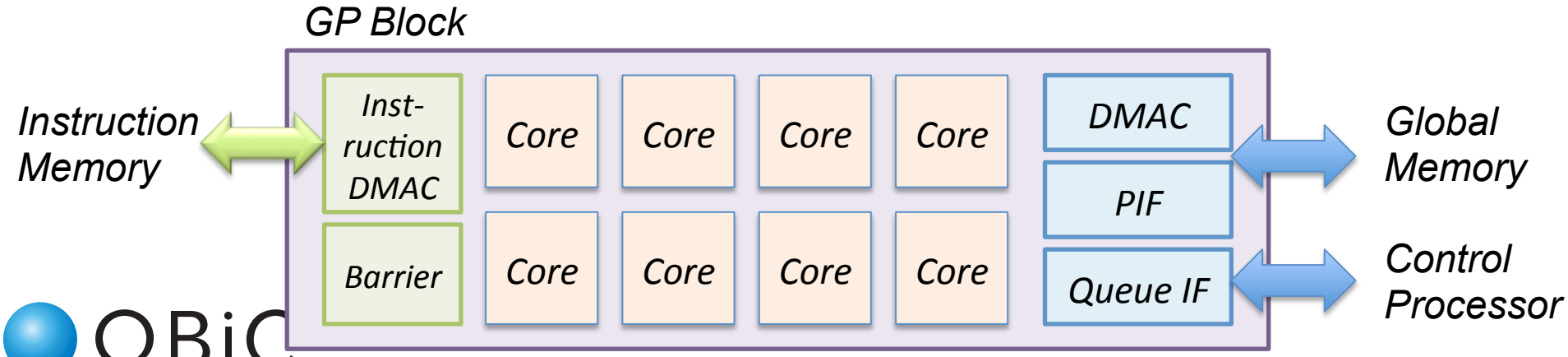
    2.56 Tflops/chip

QBiC

General-Purpose Core

- Tensilica LX @ 0.6 GHz
- 32bit integer / 32bit Floating
- 4KB I-cache / 4KB D-cache
- 8KB Local Memory
  - DMA or PIF access
- 8KB Local Instruction Memory
  - DMA read from 512KB Instruction memory





QBiC

Software / Programming Environment

- Porting GROMACS
- C (C++)Language for GP core
- Minimal library support
  - No external memory
  - No VM
- Manual Load to local instruction memory
  - 8kB (4kB+4kB double buffer)
  - for performance

QBiC

Synchronization

- 8-core synchronization unit
- Tensilica Queue-based synchronization

  send messages
  - Pipeline → Control GP
  - Network IF → Control GP
  - GP Block → Control GP
- Synchronization at memory

  – accumulation at memory

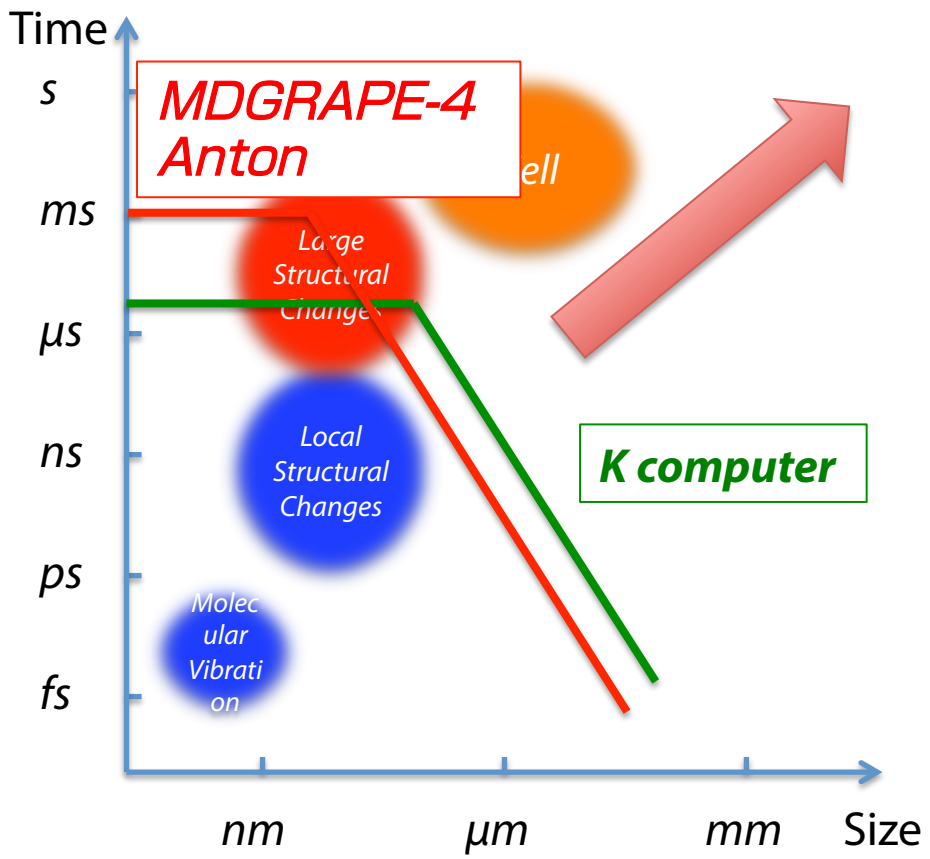
QBiC

Power Dissipation (Tentative)

- Dynamic Power (Worst) < 40W
  - Pipeline ~ 50%
  - General-purpose core ~ 17%
  - Others ~ 33%
- Static (Leakage) Power
  - Typical ~5W, Worst ~30W
- ~ 50 Gflops/W
  - Low precision
  - Highly-parallel operation at modest speed
  - Energy for data movement is small in pipeline

QBiC

MD Simulations enabled by MDGRAPE-4



- Protein structure prediction ~ 100 residues
- Structure optimization
  - ▷ fluctuations
  - ▷ mutations

Enhances biological applications of MD simulations

QBiC

Current status of MDGRAPE-4

- Board evaluation
  - Force calculation pipelines … mostly confirmed
  - General-purpose cores … mostly confirmed
  - Network … under evaluation
- System completion
  - Within FY2013

QBiC

Reflection

Though the design is not finished yet…

- Latency in Memory Subsystem
  - More distribution inside SoC
- Latency in Network
  - More intelligent Network controller
- Pipeline / General-purpose balance
  - Shift for general-purpose?
  - # of Control GP

QBiC

Shared Use of MDGRAPE-4

- 2014 1Q : Hardware completion
- 2014 3Q : Software ready
- 2014 3Q – 2016 1Q :

  Use under research collaboration agreement
- 2016 2Q –

  Shared use
- ~6 research subject / year
  - 3 QBiC / Research collaboration

QBiC

Future Perspectives

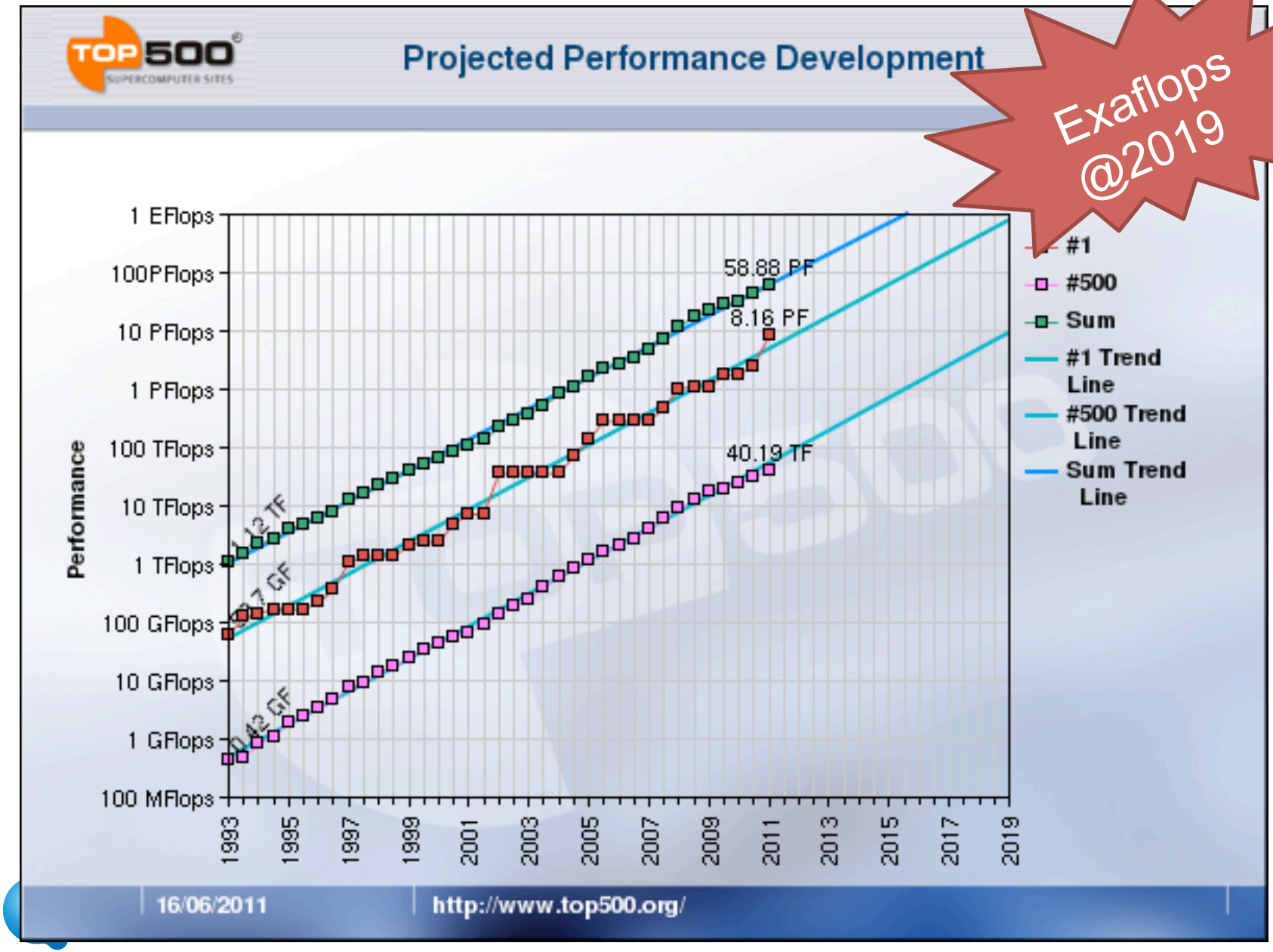
For Molecular Dynamics

- Single-chip system
  - >1/10 of the MDGRAPE-4 system can be embedded with 11nm process
  - For typical simulation system it will be the most convenient
  - Still network is necessary inside SoC
- For further performance improvement
  - # of operations / step / 20Katom ~ 10^9
  - # of arithmetic units in system ~ 10^6 /Pflops

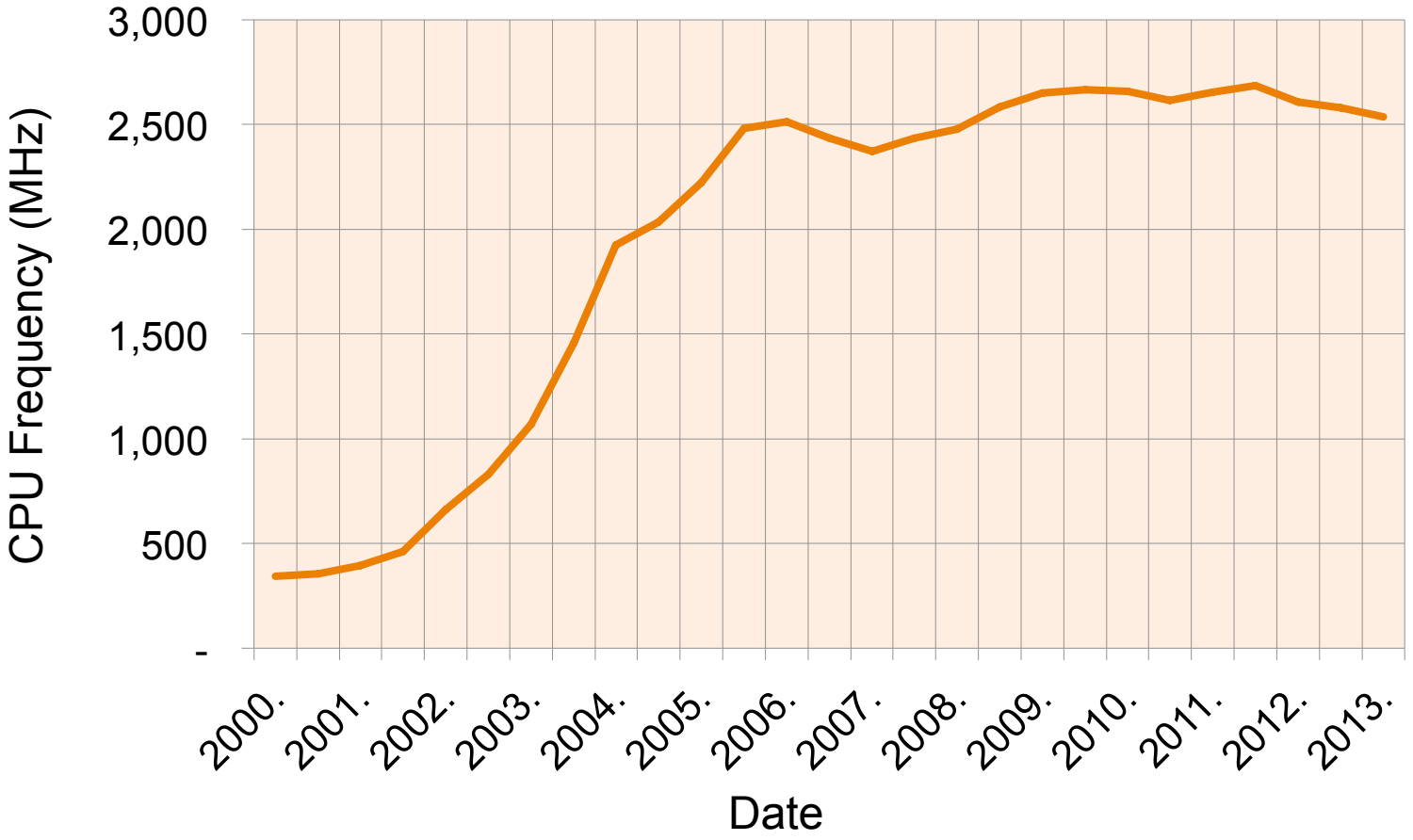
    Exascale means "Flash" (one-path) calculation
  - More specialization is required

QBiC

What happens in future?



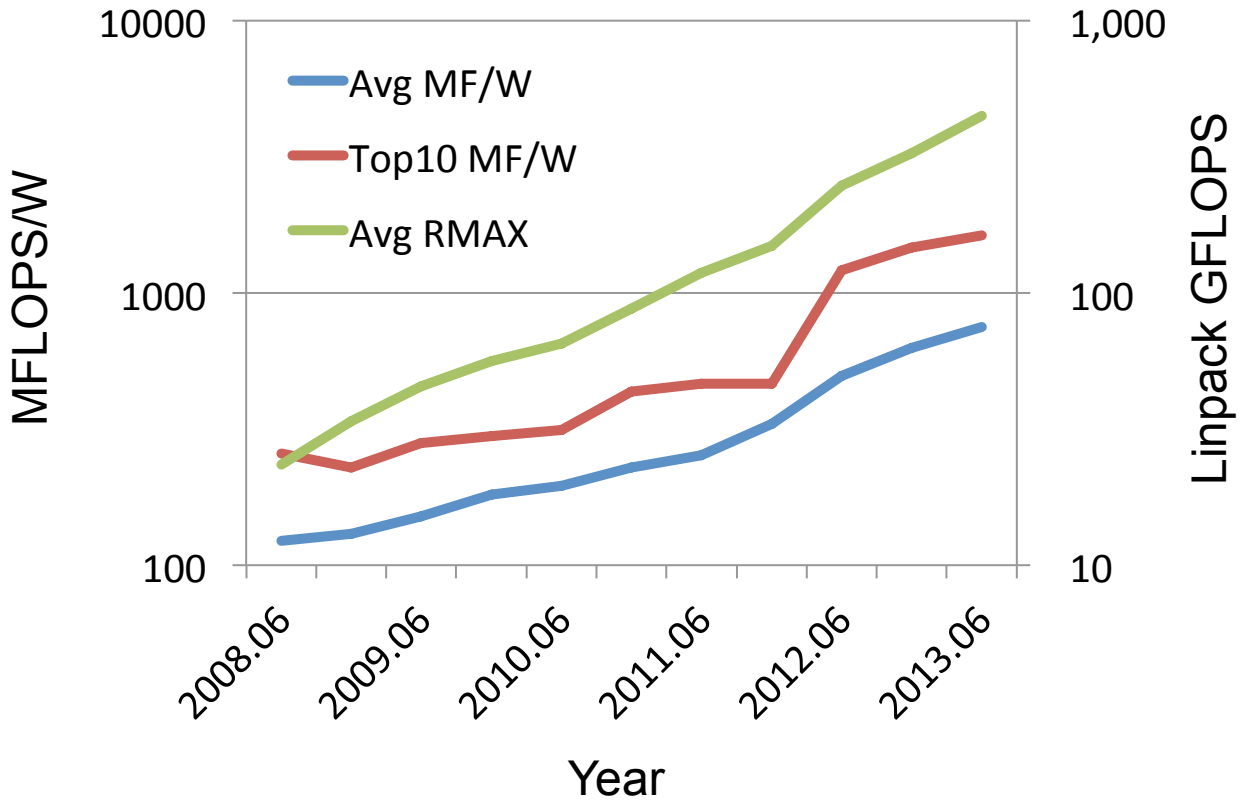
TOP500 Average Frequency



- CPU frequency became constant after 2005
- Performance increase relies on parallelization

QBiC

TOP500 Average Power Efficiency (MFLOPS/W)



- Pro : Exponential Growth
- Con: Growth speed slower than performance growth

QBiC

Exascale Systems

- 10^{18} FLOPS @ 2GHz
- 2.5 X 10^{8} FP ALUs
- ~10^{5} Processors → 10^{3} ALUs / Processors
- Possible directions
  - SoC Integration
  - BlueGene/L type machines
  - Heterogeneous

QBiC

What Exaflops means

- 10^8 arithmetic units in parallel
- Different from TFLOPS -> PFLOPS
- Which application can use this???
  - It may be better to have a center with exascale performance
  - However, the real target performance for applications will be ~10PFLOPS
  - Design should consider this point

QBiC

Future of high performance computing

- End of Si cycle ~ 2020
- Production cost of semiconductor will continue to decrease for a while
- Dedicated approach / semi-dedicated approach will be an answer to improve performance after the end of Si cycle

QBiC

Acknowledgements

- RIKEN
  - Mr. Itta Ohmura
  - Dr. Gentaro Morimoto
  - Dr. Yousuke Ohno
  - Mr. Aki Hasegawa
- Japan IBM Service
  - Mr. Ken Namura
  - Mr. Mitsuru Sugimoto
  - Mr. Masaya Mori
  - Mr. Tsubasa Saitoh
- Hitachi Co. Ltd.
  - Mr. Iwao Yamazaki
  - Mr. Tetsuya Fukuoka
  - Mr. Makio Uchida
  - Mr. Toru Kobayashi
  - and many other staffs
- Hitachi JTE Co. Ltd.
  - Mr. Satoru Inazawa
  - Mr. Takeshi Ohminato