分子動力学計算専用計算機 MDGRAPE-4の開発

泰地　真弘人　taiji@riken.jp

理化学研究所　主任研究員

生命システム研究センター　副センター長

Current MDGRAPE-4 System

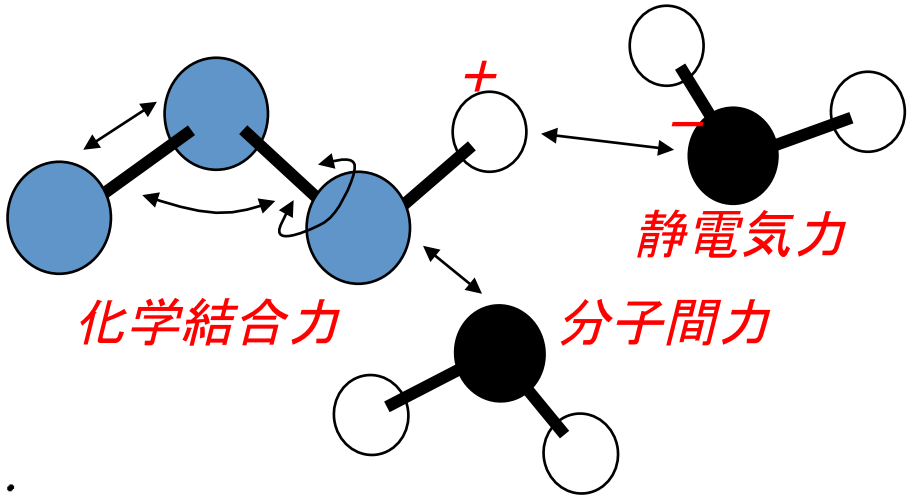
分子動力学シミュレーション

- 力の計算

$$\boldsymbol{F}_i = \sum_j \boldsymbol{f}_{ij}$$

- ニュートンの法則で原子の運動を計算

$$m_i \frac{d^2 \boldsymbol{x}_i}{dt^2} = \boldsymbol{F}_i$$

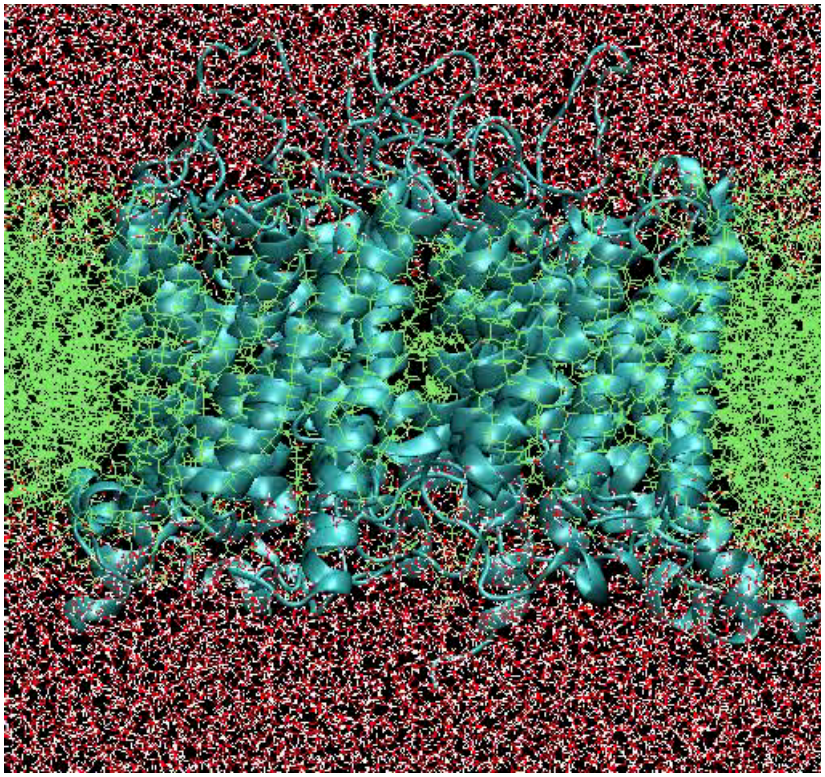


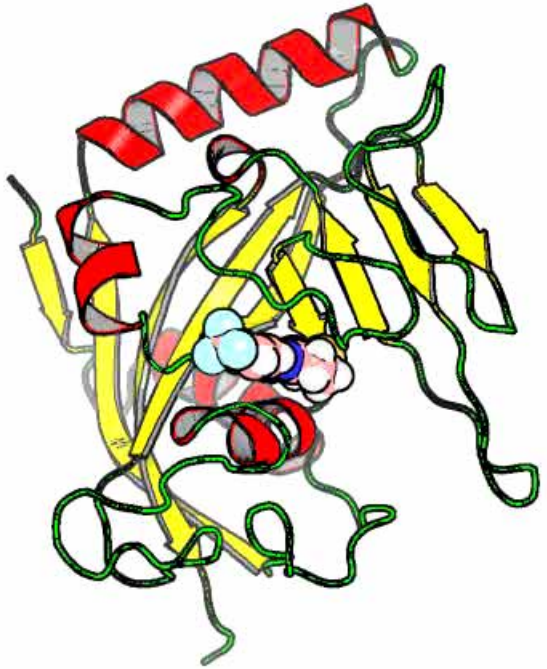
- 求めたい値を計算
- 時間刻み 1fsec

→ μsec～msecまで追いかける

QBiC

タンパク質の動的な性質



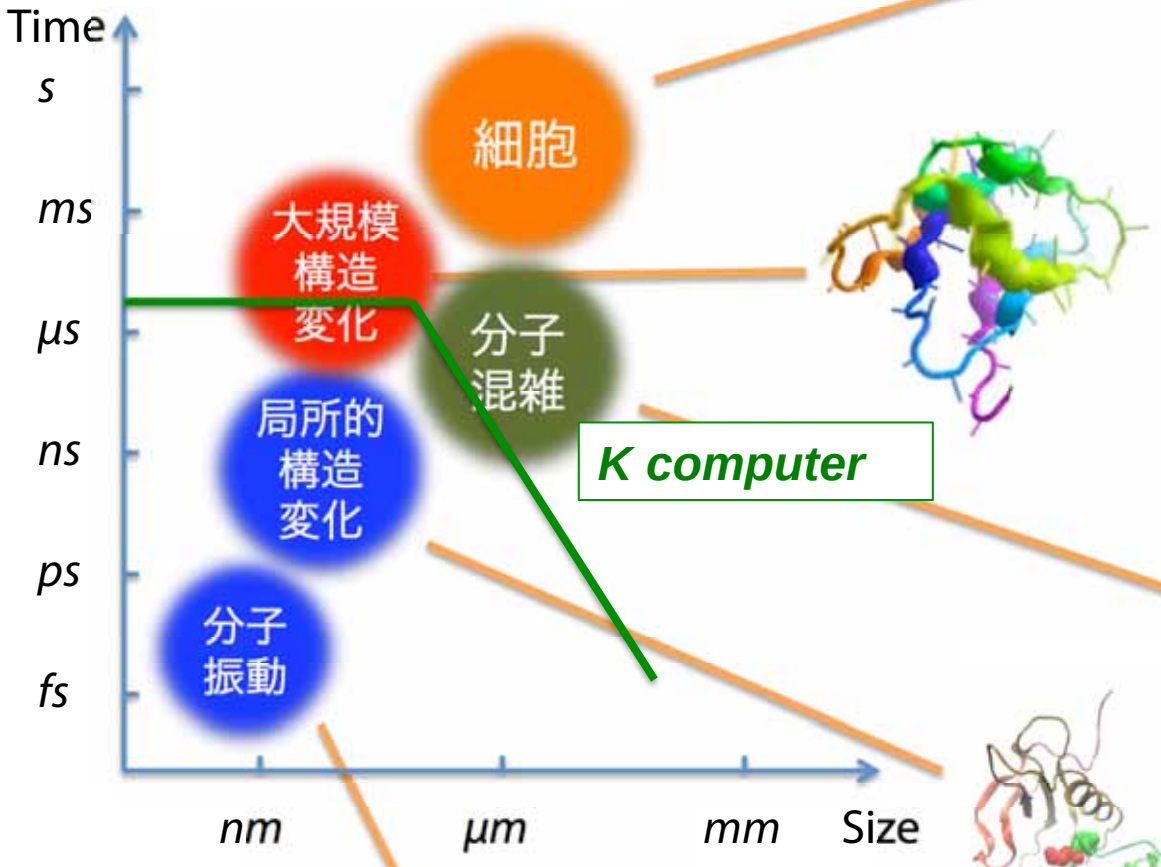


実験的に調べることが（比較的）難しい

→シミュレーションへの期待

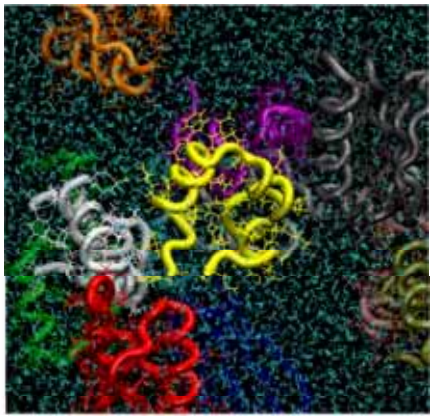
QBiC

原子から細胞へ

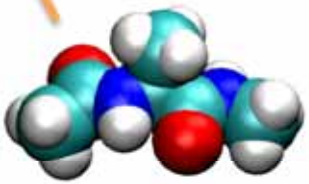


J. E. Steinberg et al.,
on youtube.com

K computer



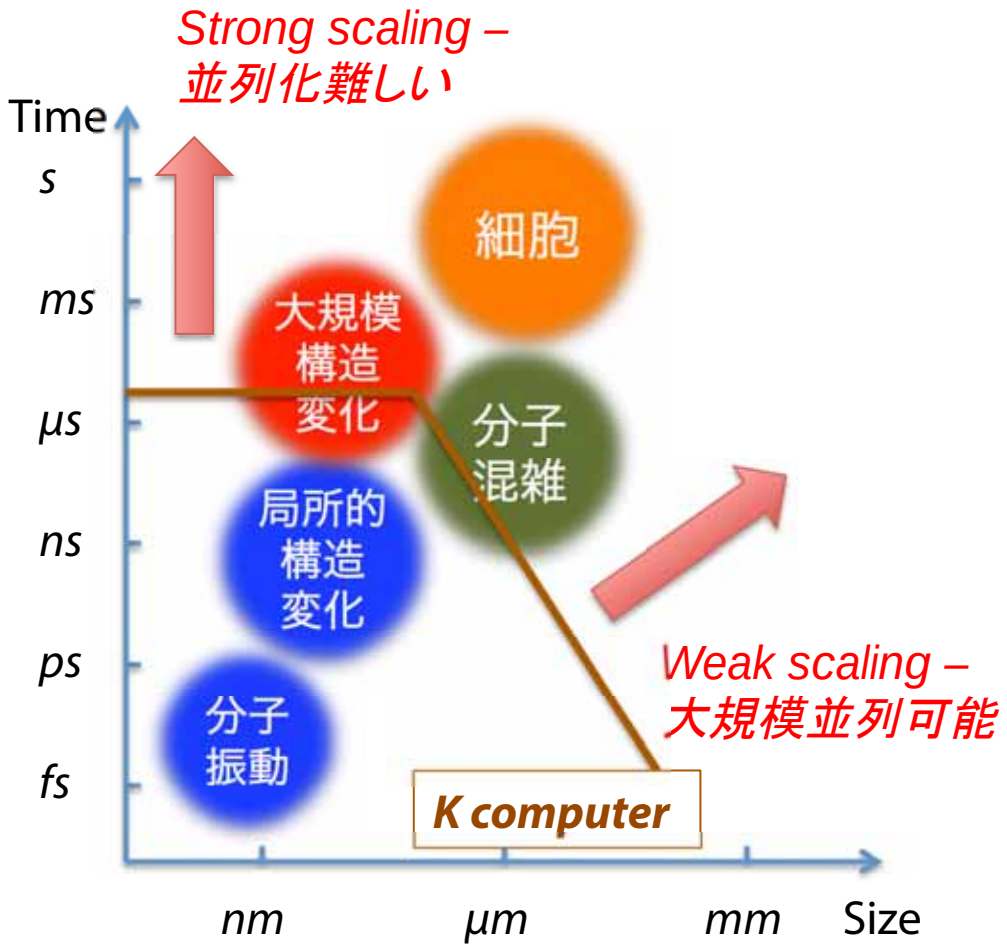
R. Harada, Y. Sugita, M. Feig,
JACS 134(10), 4842-4849 (2012).



Y. Shan et al., PNAS 106(1), 139–144 (2009).

QBiC

分子シミュレーションの問題



- 時間刻み
  ～fsec
- 目標到達時間
  μsec～sec

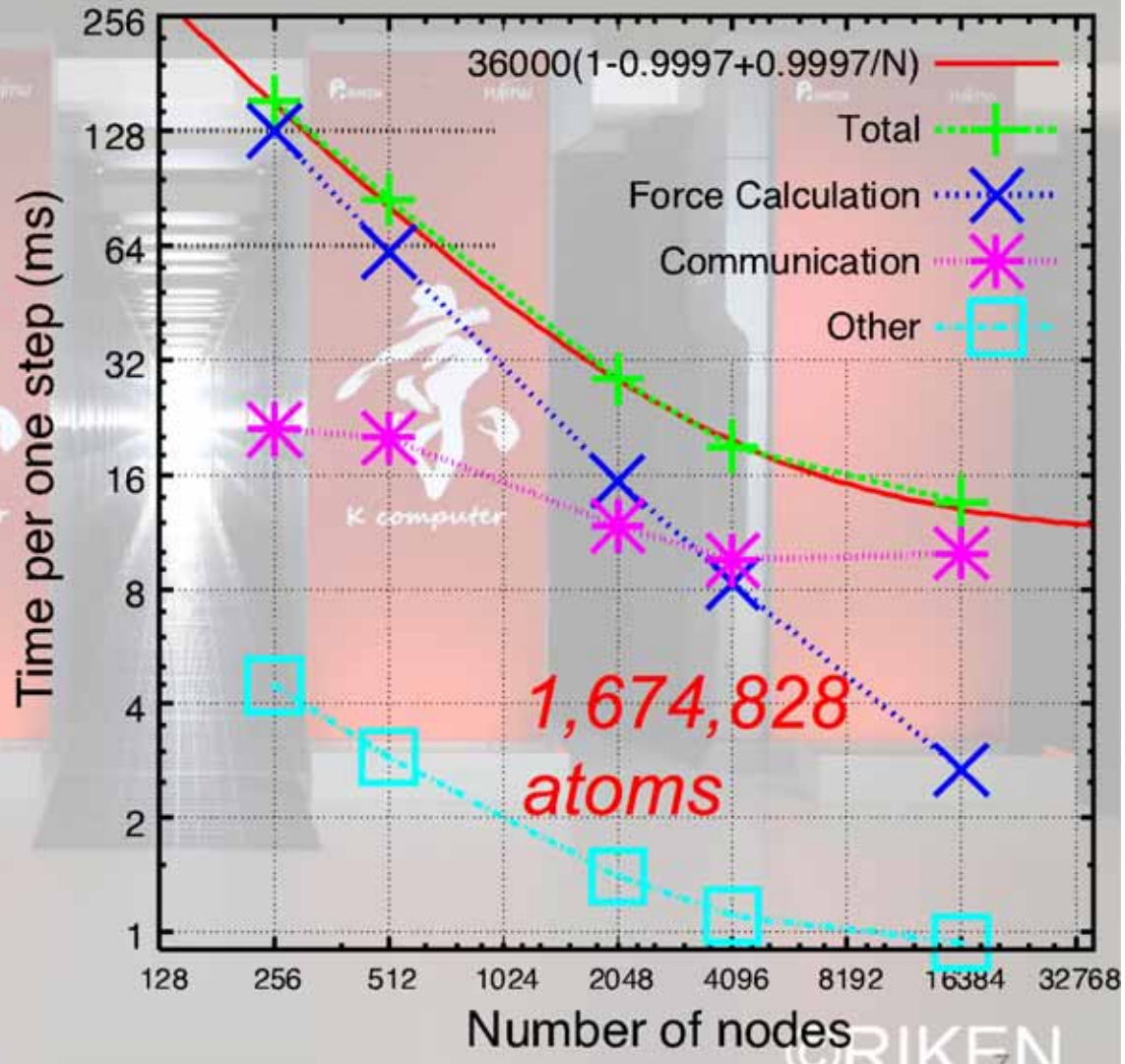
10^9～10^{15} のギャップを埋める必要性

QBiC

京コンピュータ上での分子動力学シミュレーションのスケーリング

強スケーリング
~50 原子/コア

~300万原子/Pflops



専用計算機の歴史

- 第一世代 : 初期の電子計算機
  - ▷Atanasoff-Berry Computer
- 第二世代 : 「丸ごと」専用計算機
  - ▷デルフト分子動力学計算機
    (Delft Molecular Dynamics Processor)など
  - ▷開発が大変で、時間かかった
- 第三世代 : m-TIS/GRAPE
  - ▷パソコンやワークステーションの加速装置
- 第四世代 : Anton, MDGRAPE-4
  - ▷System-on-Chip専用計算機・アクセラレータ

QBiC

m-TIS:スピン系のモンテカルロ計算専用計算機

- 日本における専用計算機のさきがけ

- 東京大学理学部
  現工学部伊藤伸泰助教授と共同
- 1987-88：1号機
  学部生のとき、趣味で製作
- 1990-92: 2号機
  書き換え可能ゲートアレイを使った
  たぶん最初の計算機

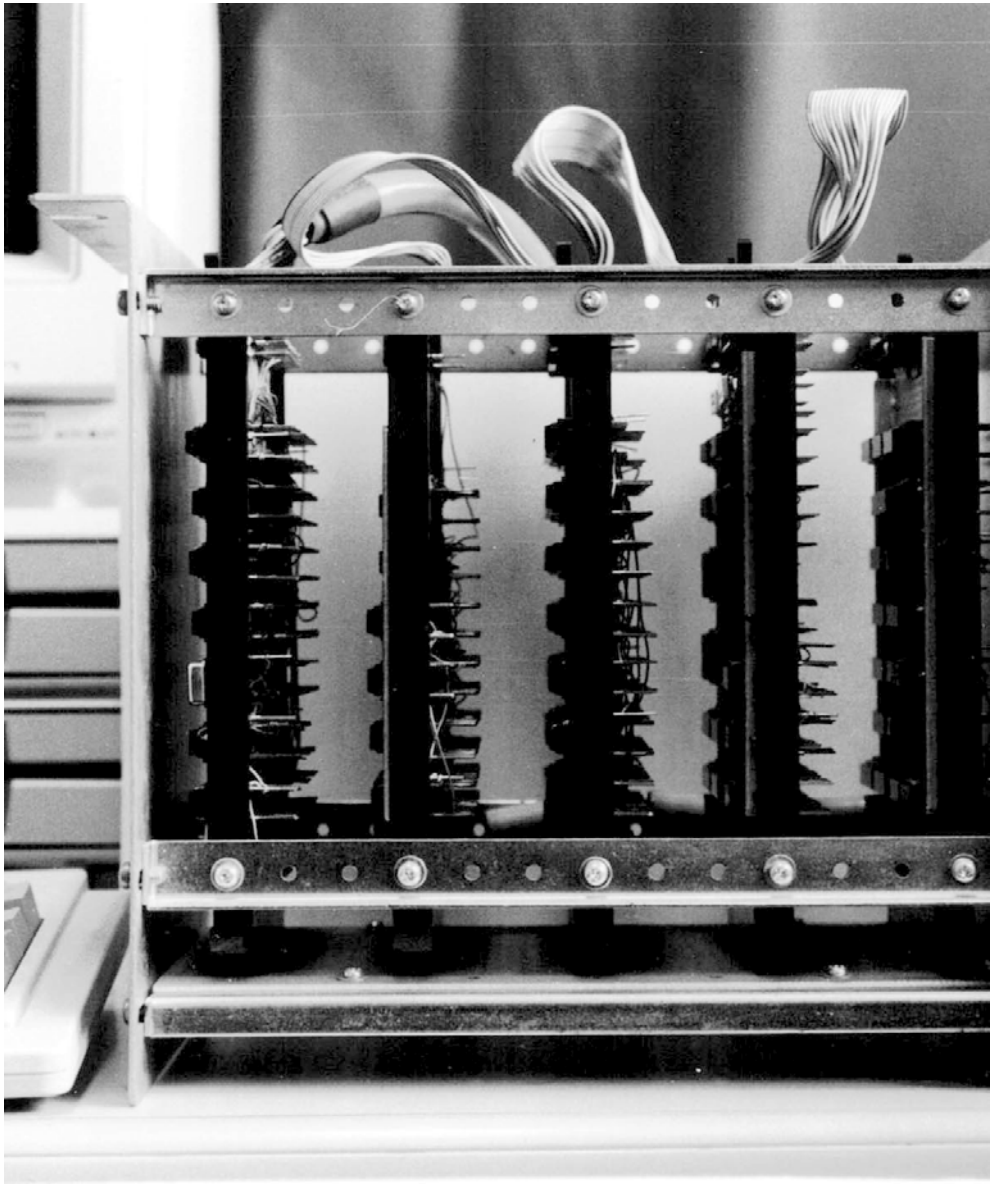
QBiC

m-TIS（1987)

秋葉原で買ってきた部品で
完全手作り

最初は手弁当

半年ぐらいで完成



QBiC

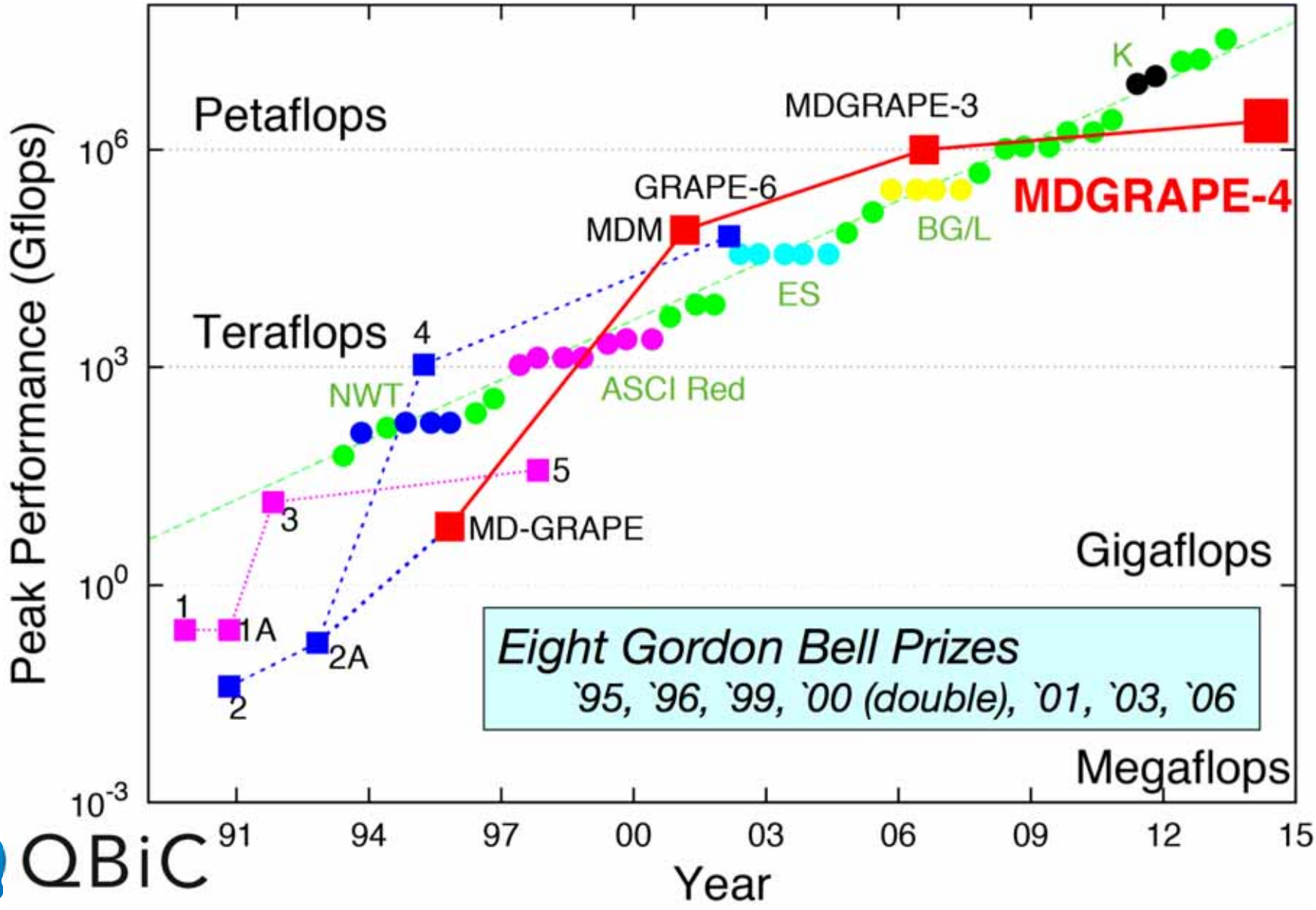
GRAPE計画

- 1989～
- GRAPE (GRAvity PipE）計画
- 銀河・銀河団、球状星団、惑星形成、宇宙論など
- 重力多体問題のための専用計算機
- 原子を星と思えば、分子にも使える

- 東京大学教養学部宇宙地球科学教室で始まり、
- その後東京大学理学部/理化学研究所/国立天文台

QBiC

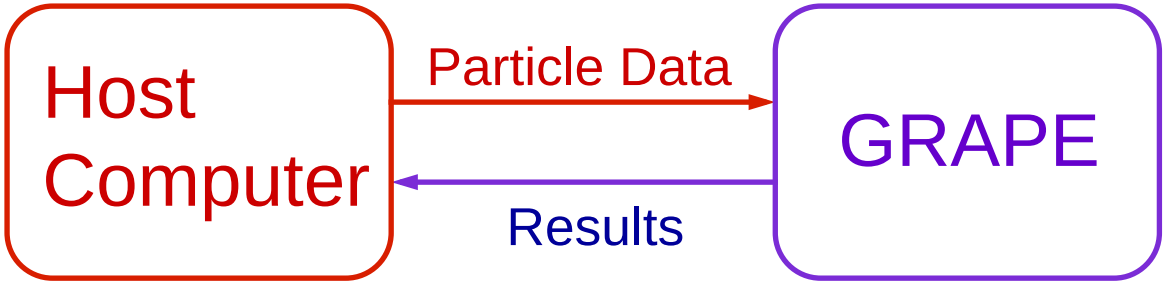
GRAPEの歴史



QBiC

GRAPE as Accelerator

- Accelerator to calculate forces
  by dedicated pipelines



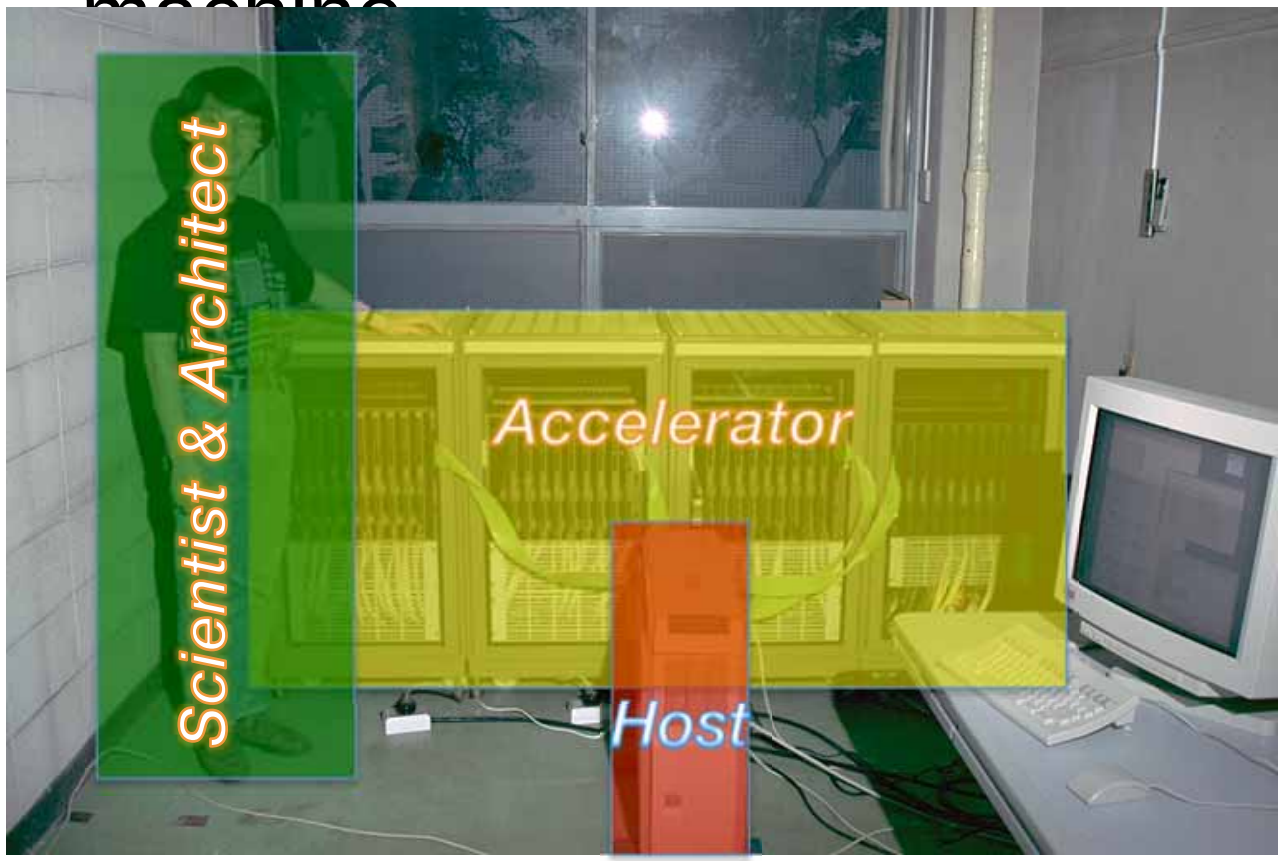
Most of Calculation → GRAPE
Others → Host computer

- 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

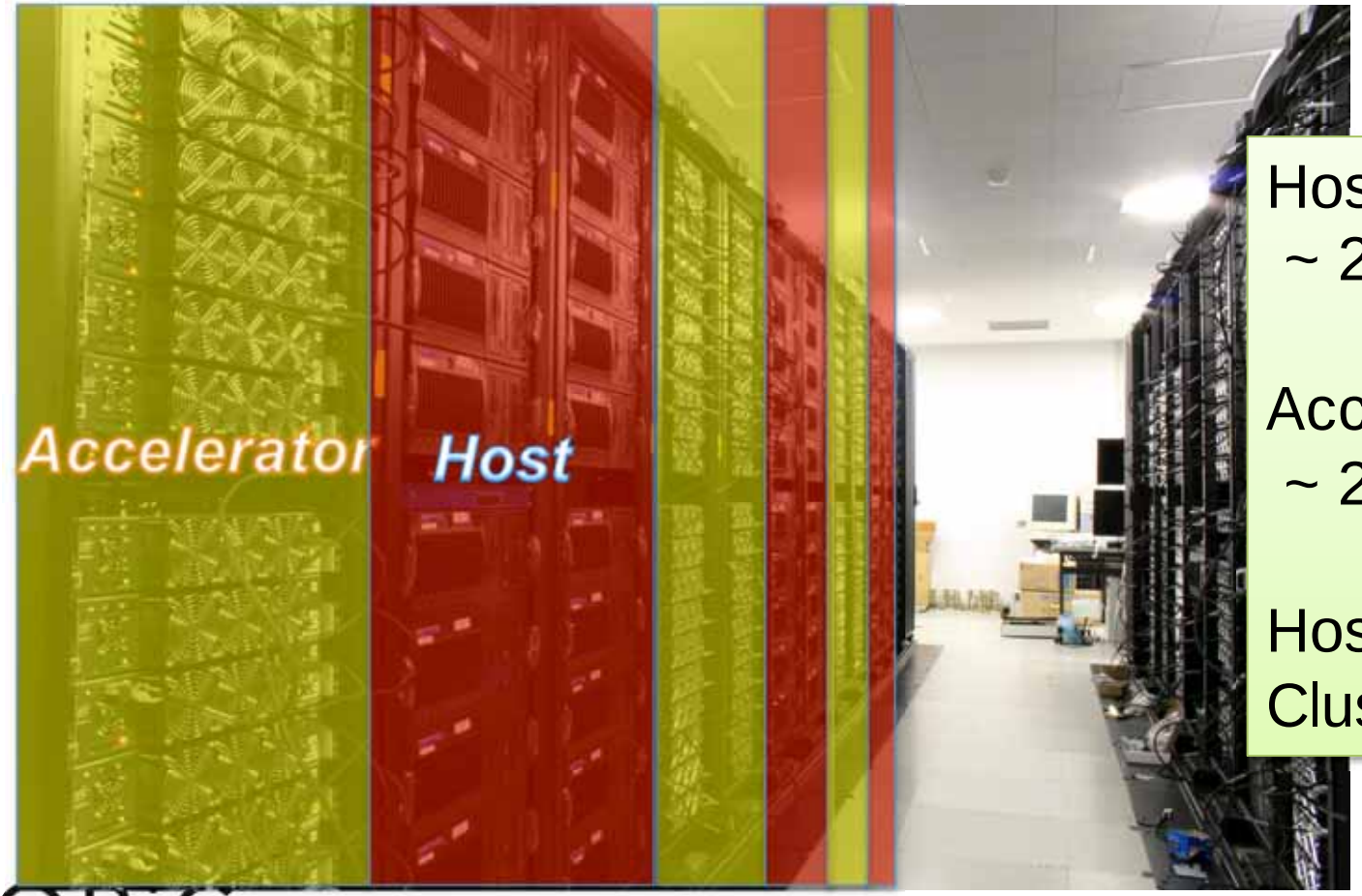
Accelerator PU
~ 0.6 Gflops

Host:
Single or SMP

QBiC

GRAPE in 2000s

- MDGRAPE-3: The first petaflops machine



Host CPU
~ 20 Gflops

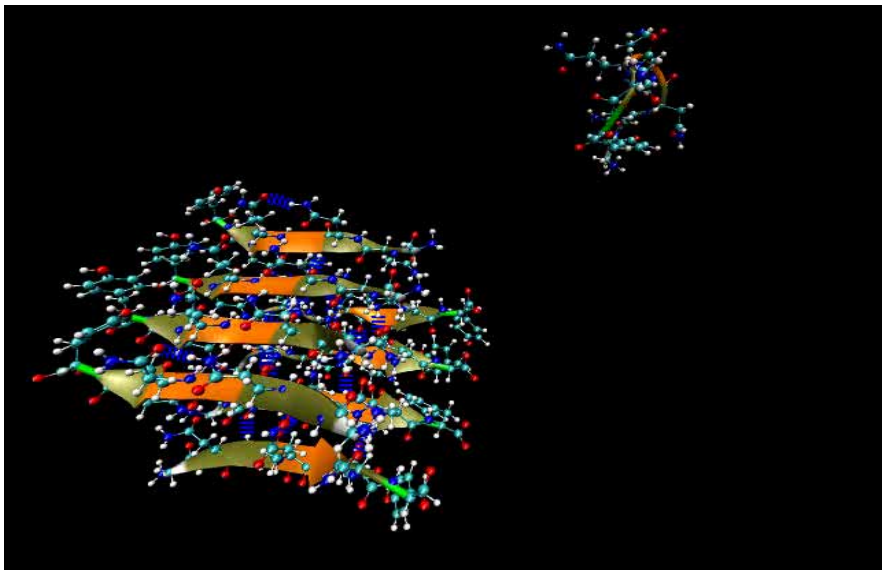
Accelerator PU
~ 200 Gflops

Host:
Cluster

QBiC

MDGRAPE-3: 並列性能の問題

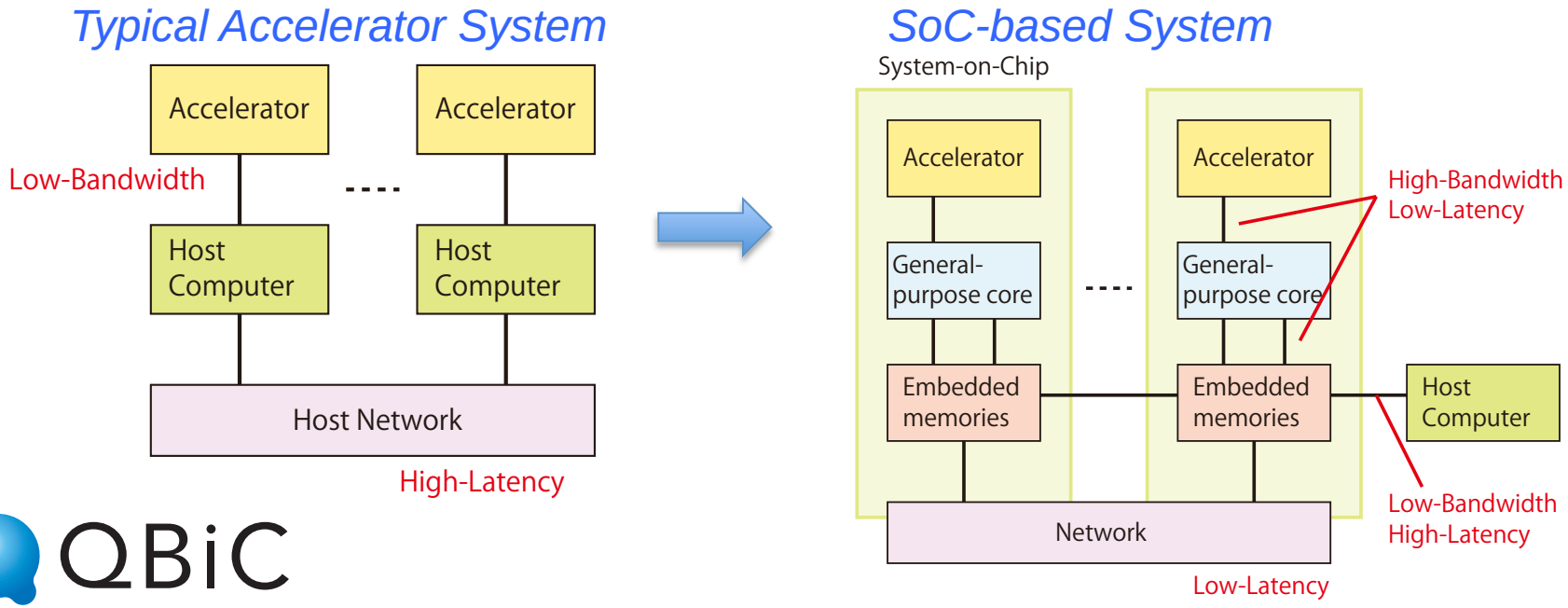
- Gordon Bell 2006 Honorable Mention, Peak Performance
- Amyloid forming process of Yeast Sup 35 peptides
- 1700万原子
- 0.55sec/step
- 実効性能
  185 Tflops
- 効率 ~ 45 %
- ペタスケールの性能には
  100万原子以上が必要
- ホスト計算機＋ネットワークの制約



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研究所)

Forbes 400 Richest Americans | World's Billionaires | World's Most Powerful People | America's B Small Comp

#164 A. Jerrold Perenchio — Browse list

David Shaw

Net Worth **$2.2 B**

As of March 2011

+ Follow David Shaw 6

Like 4 likes. Sign Up to see what your friends like.

At a Glance

Age: **60**

Source: **hedge funds , self-made**

Residence: **New York, NY**

Country of citizenship: **United States**

Education: **PHD, Stanford University; BA/BS, University of California, San Diego**

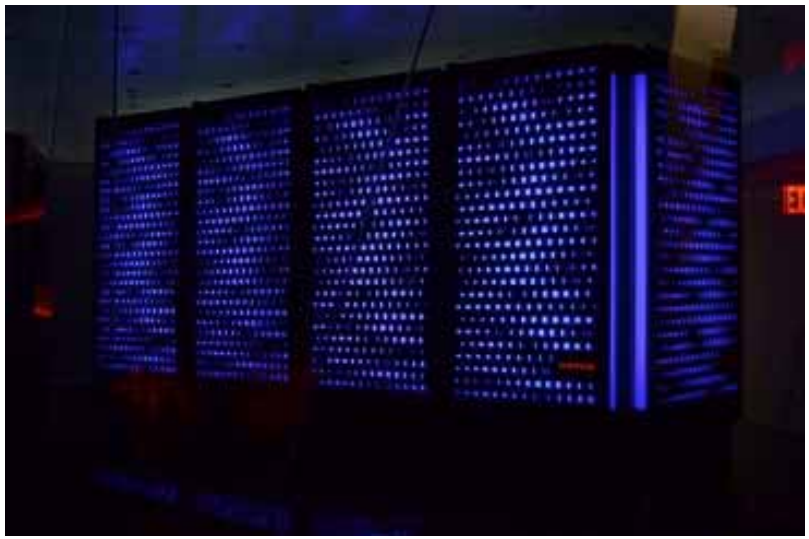
Marital Status: **Married**

Forbes Lists

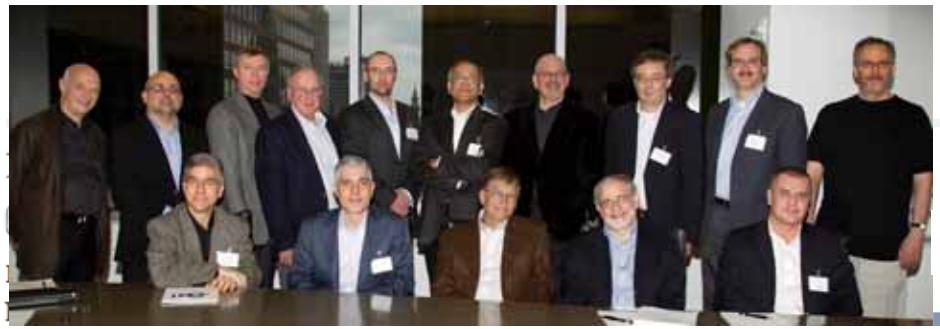
#540 Forbes Billionaires

#179 in United States

#164 Forbes 400



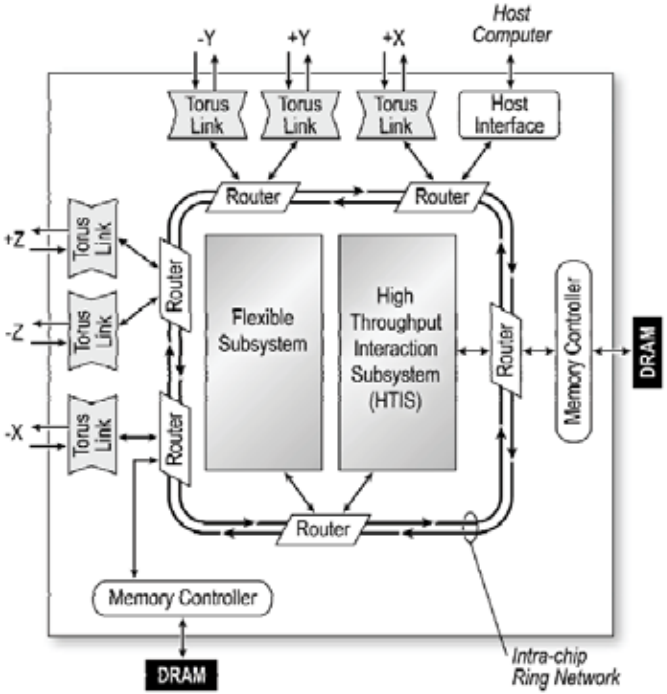
分子シミュレーション
専用計算機



Bill Gates visit at Schroedinger Inc.

Anton

- D. E. Shaw Research
- 専用パイプライン（GRAPE類似）
  + 汎用CPUコア
  + 専用ネットワーク
- **~20μsec/step, ~10μsec/day**
- **Anton-2 : ~2μsec/step, ~85μsec/day**
- 汎用部・ネットワークとの統合の重要性を示した



	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%)

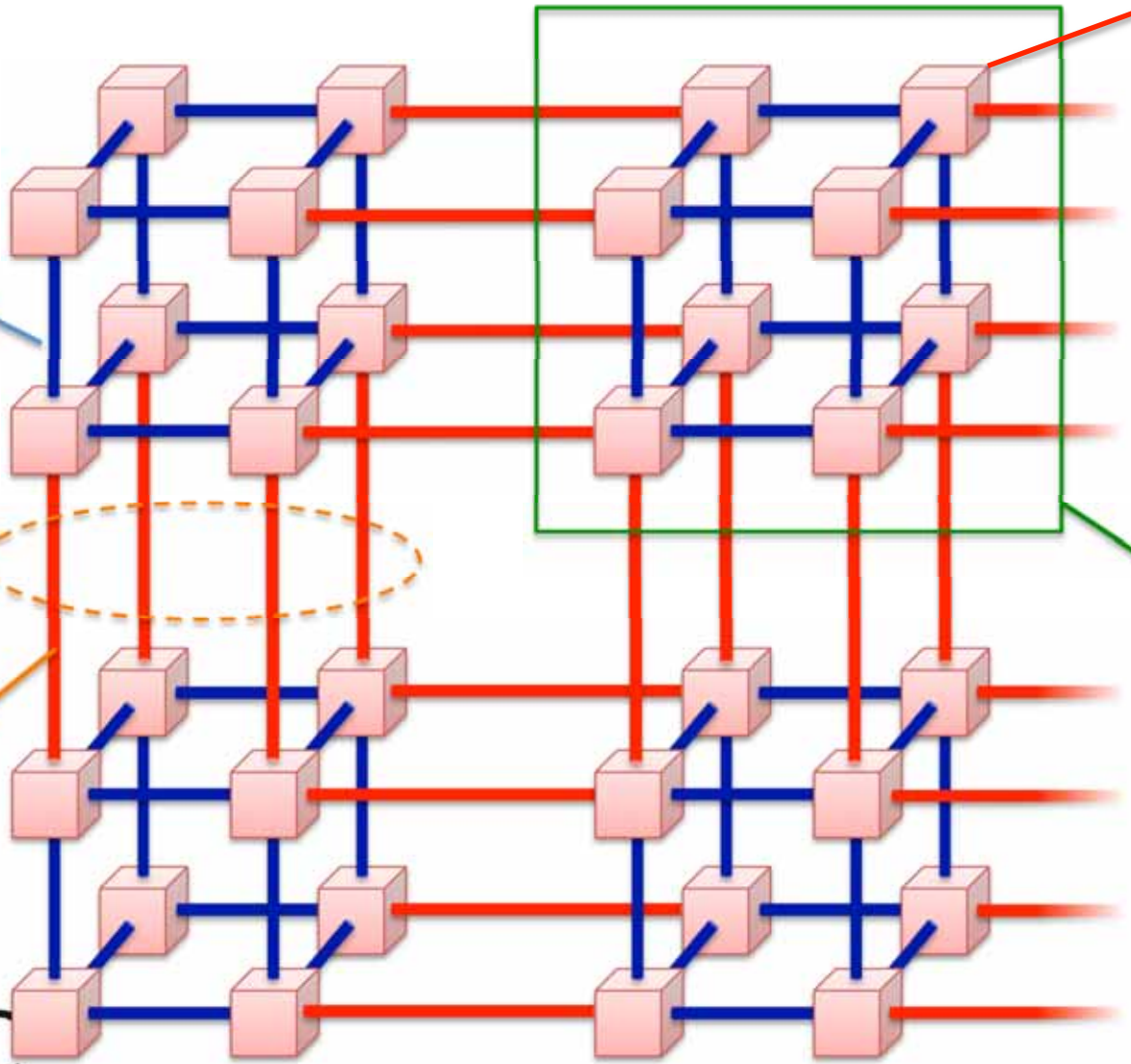
R. O. Dror et al., Proc. Supercomputing 2009, in USB memory.

MDGRAPE-4

- Antonに追いつき追い越す…
- 目標性能
  - ▷5万原子の系を一日に数μsec
- 世界的に使われているソフトウェア GROMACSを移植
- 完成: 2014年
- MDGRAPE-3からの拡張
  - ▷130nm → 40nm process
  - ▷**ネットワークと汎用CPUをシステムオンチップに集積**

QBiC

MDGRAPE-4 System



MDGRAPE-4 SoC

Total 512 chips (8x8x8)

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

48 Optical Fibers

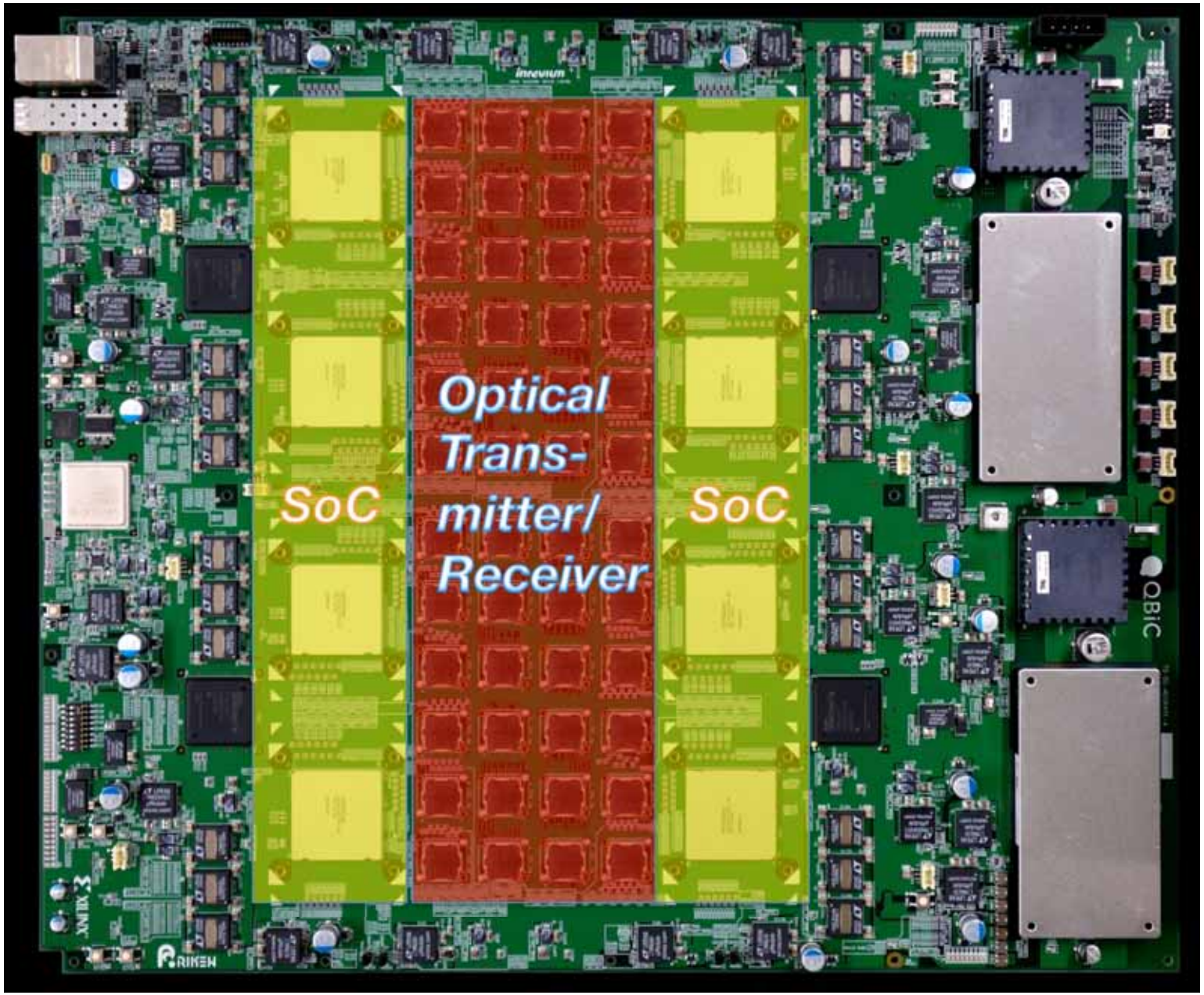
12 lane 6Gbps Optical

Node (2U Box)

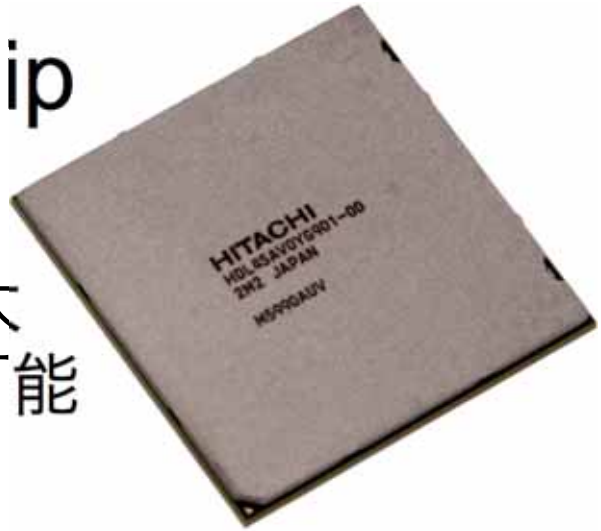
Total 64 Nodes (4x4x4) =4 pedestals

QBiC

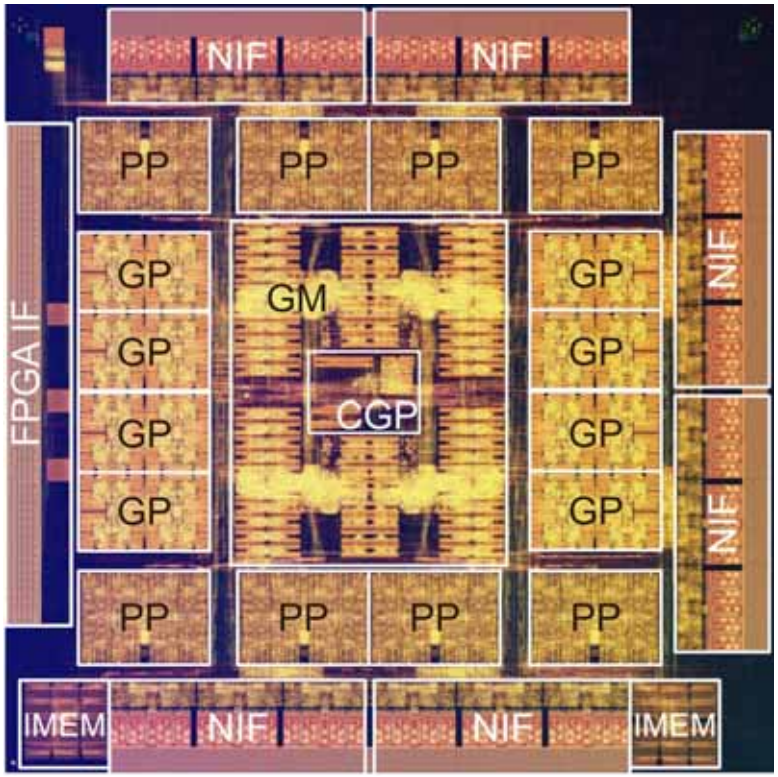
MDGRAPE-4 Board



MDGRAPE-4 System-on-Chip

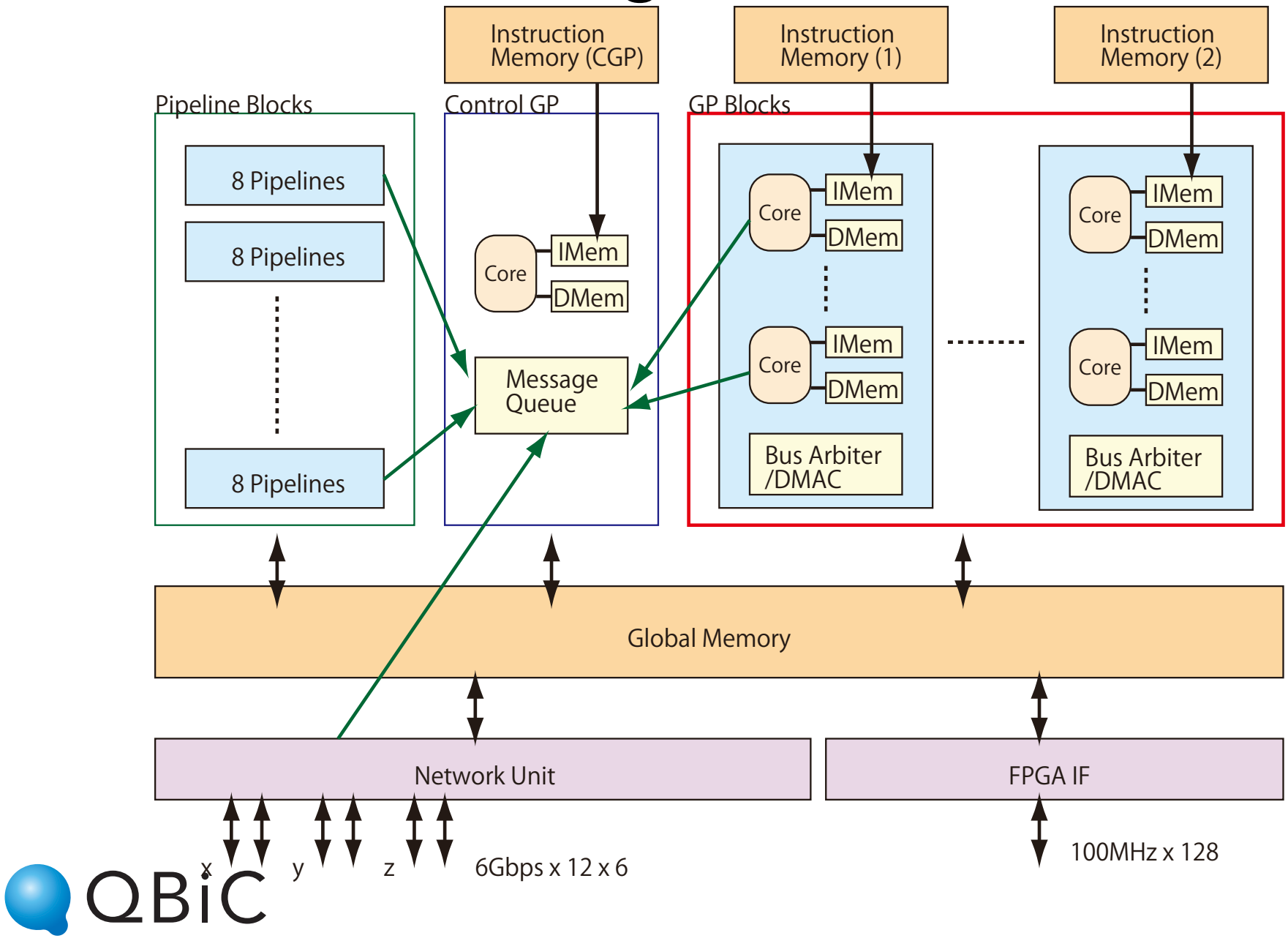


- **理研QBiCで設計**
  - ▷アカデミアで設計した*LSI*として最大
  - ▷理研と*D.E.Shaw Research* だけが可能
- 40 nm (日立製作所), ~ 230mm^2
- 専用パイプライン64本
  @ 0.8GHz 2.5TFLOPS
- 汎用コア65個
  Tensilica Extensa LX4
  @0.6GHz
- 3次元トーラスネットワーク
  @ 7.2GB/sec/一方向



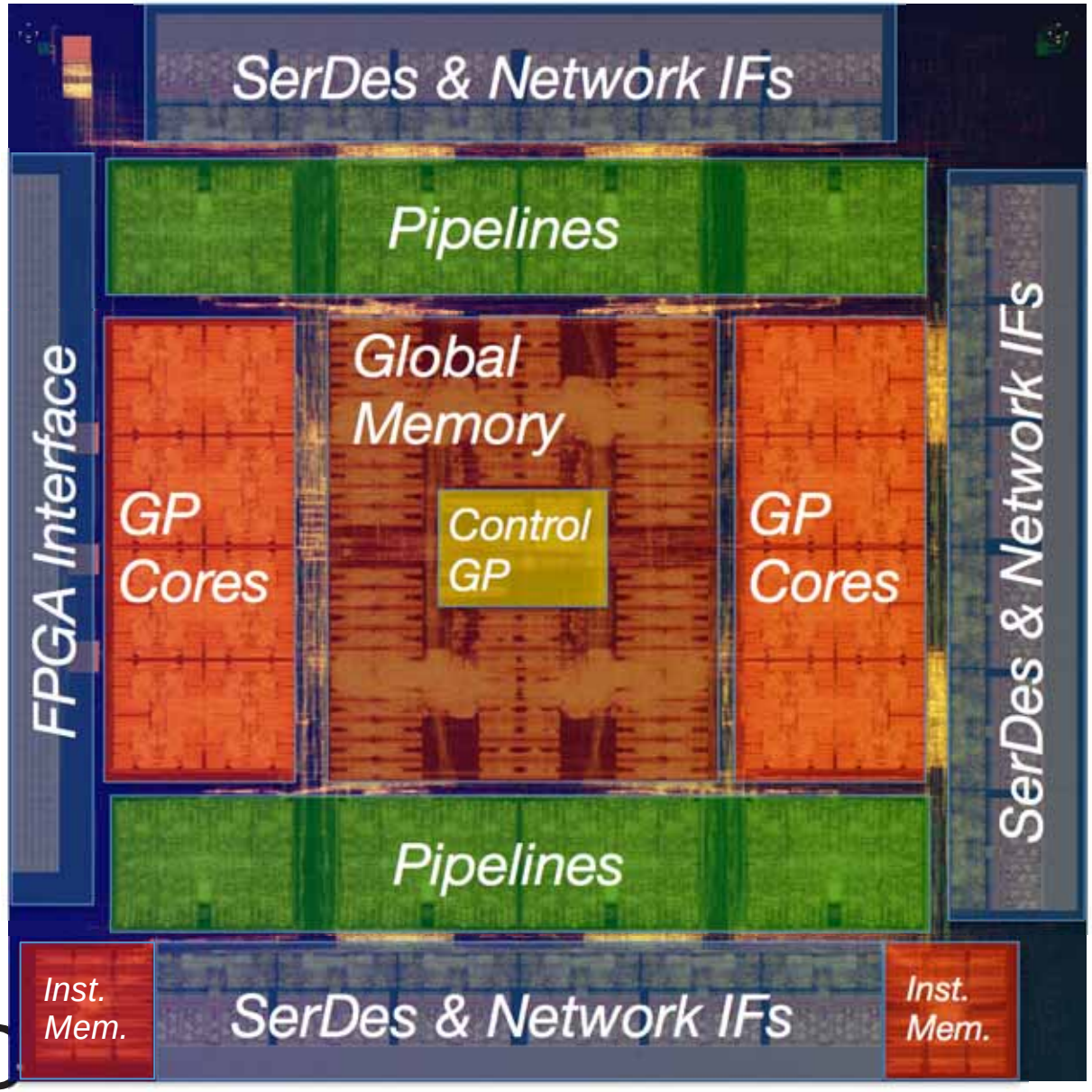
QBiC

SoC Block Diagram



QBiC

SoC Physical Image



QBiC

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 = 2.5 TFLOPS equiv.

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

Technical Advantages of Special-Purpose Computers for MD

- High performance
  - > 2Tflops/chip
- **Low communication latencies**
- **Low power consumption**
  - 65W/chip (Worst), 50Gflops/W
  - 65KW system (MAX, currently ~40KW)
  - **~10 times better than GPUs** using the same technology
- Communication latencies and power consumption will become **more serious problems in future** high-performance MD simulations

QBiC

Current status of MDGRAPE-4

- Hardware
  - Mostly completed
- Software
  - Porting GROMACS 5
  - Collaboration with Prof. Lindahl, Stockholm University

QBiC

Reflection

Though the system is not finished yet…

- Latency in Memory Subsystem
  - More distribution inside SoC
- Latency in Network
  - More intelligent Network controller
  - Better SERDES…
- Pipeline / General-purpose balance
  - More general-purpose computing power
  - Specialized functions in GP
  - increase # of Control GP or faster messaging system
- Pipeline
  - Automatic balancing within a block
  - FMM support

QBiC

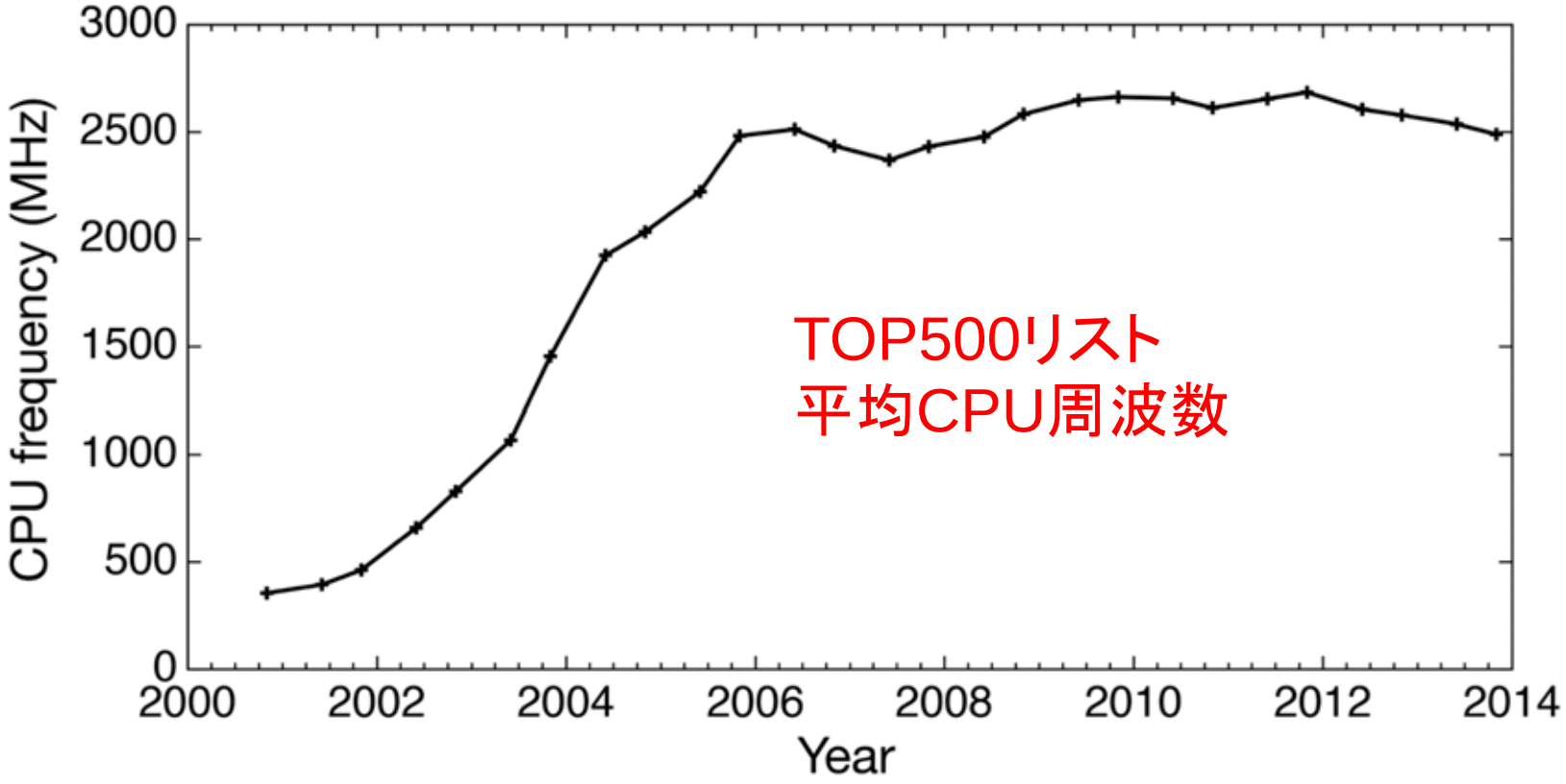
Future Perspectives of MD machine

- Single-chip system
  - Multichip Modules using 2.5D technology
  - >1/16 of the MDGRAPE-4 system can be embedded into a single module 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

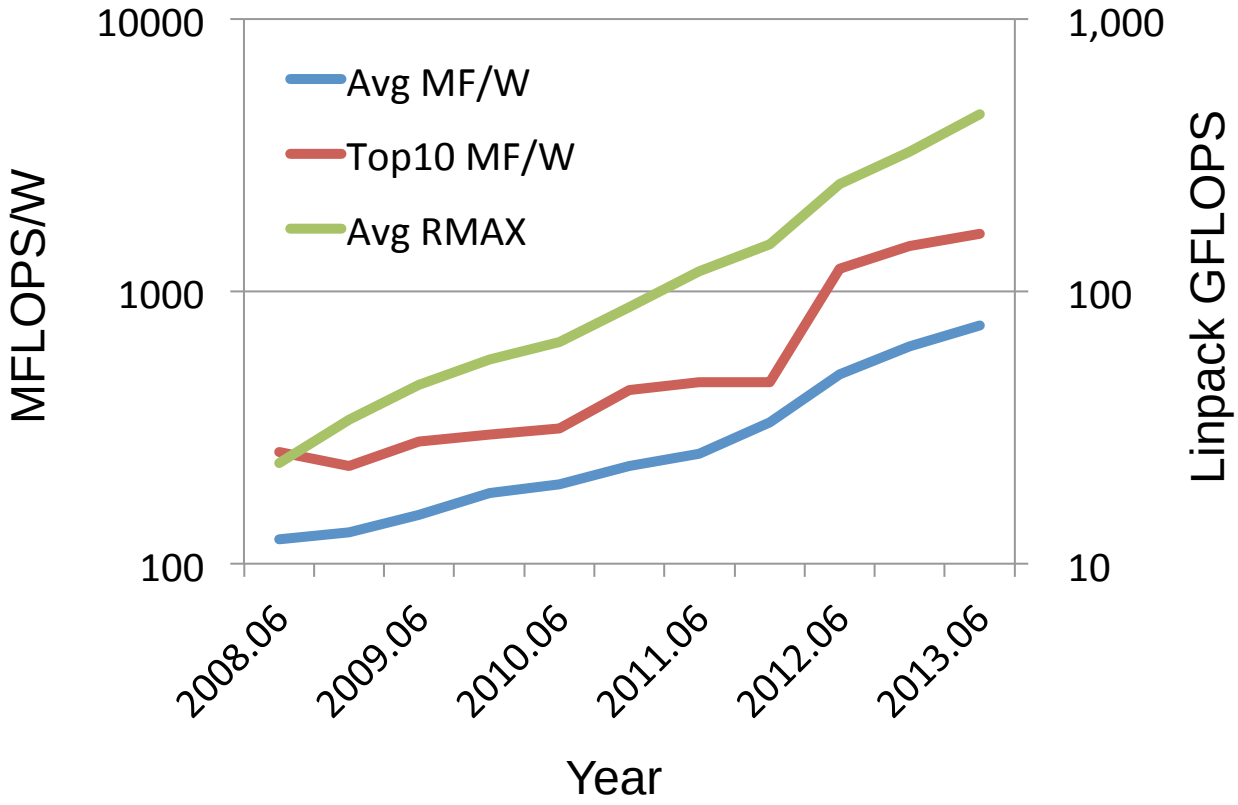
今後の高性能計算



- さらなる並列化の必要性
- ますます性能向上が困難に

QBiC

TOP500 Average Power Efficiency (MFLOPS/W)



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

QBiC

Dilemma in power performance

- Larger systems require better power performance
- For better power performance
  - Lower operational frequency
  - More arithmetic units
- These make “real efficiencies” of massively-parallel systems lower

Importance of strong scaling

Not only for MD

- Fine-grain parallelization
  - Hardware & software support
  - May decrease power efficiency & nominal peak performance
- Data flow computing
- Network-driven computing

- Large-scale Artificial Neural Network

QBiC

Programming models for fine-grain parallel operations

- Erlang
  - message送信/receive
  - 独立アドレス空間
- Future / Promise
  - http://en.wikipedia.org/wiki/Futures_and_promises
  - C++11, Java/Scala etc.
  - Futureで関数の「予約」・完了待ち
  - Promiseで関数の結果・完了通知
  - Promise pipelining による通信の削減

QBiC

京コンピュータから京コンピューティングへ

- 京の次＝エクサ(10^{18}) スケール?
- ほとんどの生命科学系計算は（単一計算では）Exaflopsに到達しない
  - ▷累積計算ではExaflopsを利用可能
  - ▷分子シミュレーションなど、京クラスの計算の高速化が重要なものが多い
- 京クラスの計算をより高い実効性能で多量に実行する

  「京コンピューティング」

  の実現が重要

QBiC

Acknowledgements

- RIKEN
  - Mr. Itta Ohmura
  - Dr. Gentaro Morimoto
  - Dr. Yousuke Ohno
  - Mr. Aki Hasegawa
  - Dr. Noriaki Okimoto
  - Dr. Yoshinori Hirano
- 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

QBiC