Инженерно-техническая школа МИЭМ
Учебный центр НИУ ВШЭ "Вороново», 7 февраля 2016 г.

Архитектура суперкомпьютеров и прорыв в эксафлопсную эру

Магистерская программа
"Математические методы моделирования и компьютерные технологии"

В.В.Стегайлов

профессор департамента прикладной математики,
д.ф.-м.н., зав. отделом ОИВТ РАН

Развитие вычислительной техники

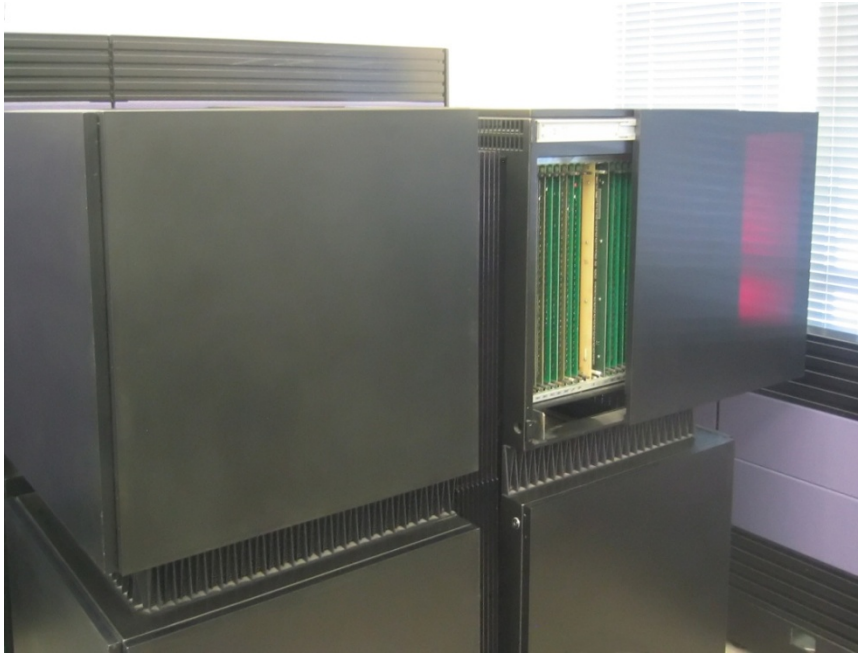
CDC 3600

peak: 0.0013 GFlops
1 CPU
clock fq.: 0.7 MHz
memory: 32 Kb

peak: 367000 GFlops
131072 CPUs
clock fq.: 700 MHz
memory: 250 Mb

Blue Gene/L

Суперкомпьютеры 1986-87 гг.



Thinking Machines CM-2:
16384 однобитовых процессора совместно с 512 арифметическими ускорителями Weitek

Meiko Computing Surface:
64 транспьютерных узлов с процессорами Intel i860

… и параллельные алгоритмы для первопринципных расчетов

VOLUME 68, NUMBER 9 PHYSICAL REVIEW LETTERS 2 MARCH 1992

Ab Initio Theory of the Si(111)-(7×7) Surface Reconstruction: A Challenge for Massively Parallel Computation

Karl D. Brommer,[1] M. Needels,[2] B. E. Larson,[3] and J. D. Joannopoulos[1]
[1]*Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139*
[2]*AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974*
[3]*Thinking Machines, Cambridge, Massachusetts 02139*
(Received 8 November 1991)

An *ab initio* investigation of the Si(111)-(7×7) surface reconstruction is undertaken using the state of the art in massively parallel computation. Calculations of the total energy of an ~700 effective-atom supercell are performed to determine (1) the fully relaxed atomic geometry, (2) the scanning tunneling microscope images as a function of bias voltage, and (3) the energy difference between the (7×7) and the (2×1) reconstructions. The (7×7) reconstruction is found to be energetically favorable to the (2×1) surface by 60 meV per (1×1) unit cell.

PACS numbers: 73.20.−r, 68.35.Bs, 68.35.Md

Thinking Machines CM-2

VOLUME 68, NUMBER 9 PHYSICAL REVIEW LETTERS 2 MARCH 1992

Ab Initio Total-Energy Calculations for Extremely Large Systems: Application to the Takayanagi Reconstruction of Si(111)

I. Štich, M. C. Payne, R. D. King-Smith, and J-S. Lin
Cavendish Laboratory (TCM), University of Cambridge, Madingley Road, Cambridge CB3 0HE, United Kingdom

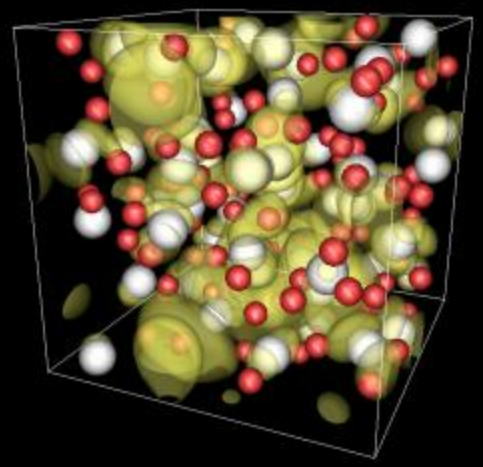
L. J. Clarke
Edinburgh Parallel Computer Centre, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom
(Received 8 November 1991)

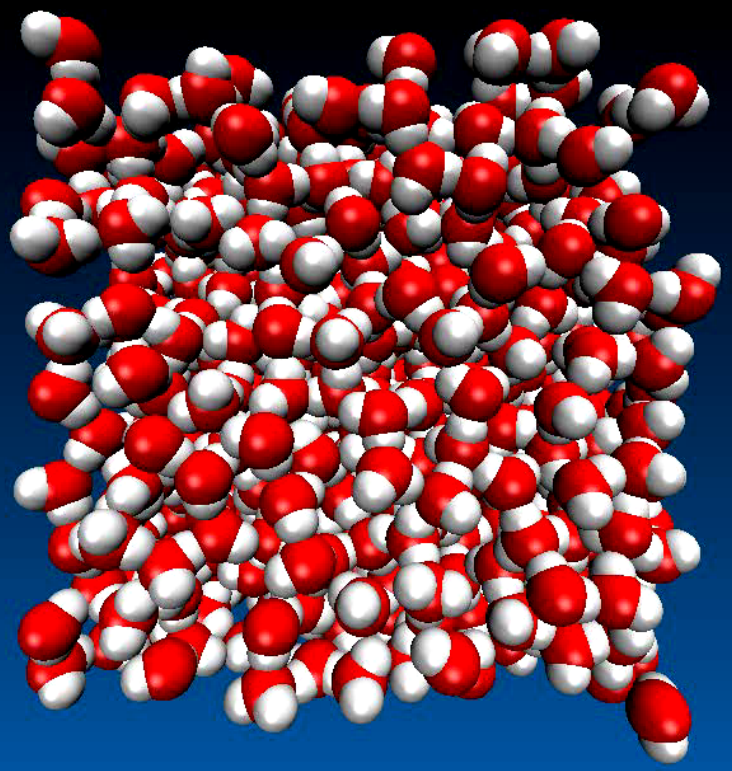
We have implemented a set of total-energy pseudopotential codes on a parallel computer which allows calculations to be performed for systems containing many hundreds of atoms in the unit cell. Using these codes we have calculated the total energies and structures of the 3×3, 5×5, and 7×7 Takayanagi reconstructions of the (111) surface of silicon. We find that the 7×7 structure minimizes the surface energy and observe structural trends across the series which can be correlated with the degree of charge transfer between the dangling bonds on the adatoms and rest atoms.

PACS numbers: 68.35.−p, 31.20.−d, 71.45.Nt

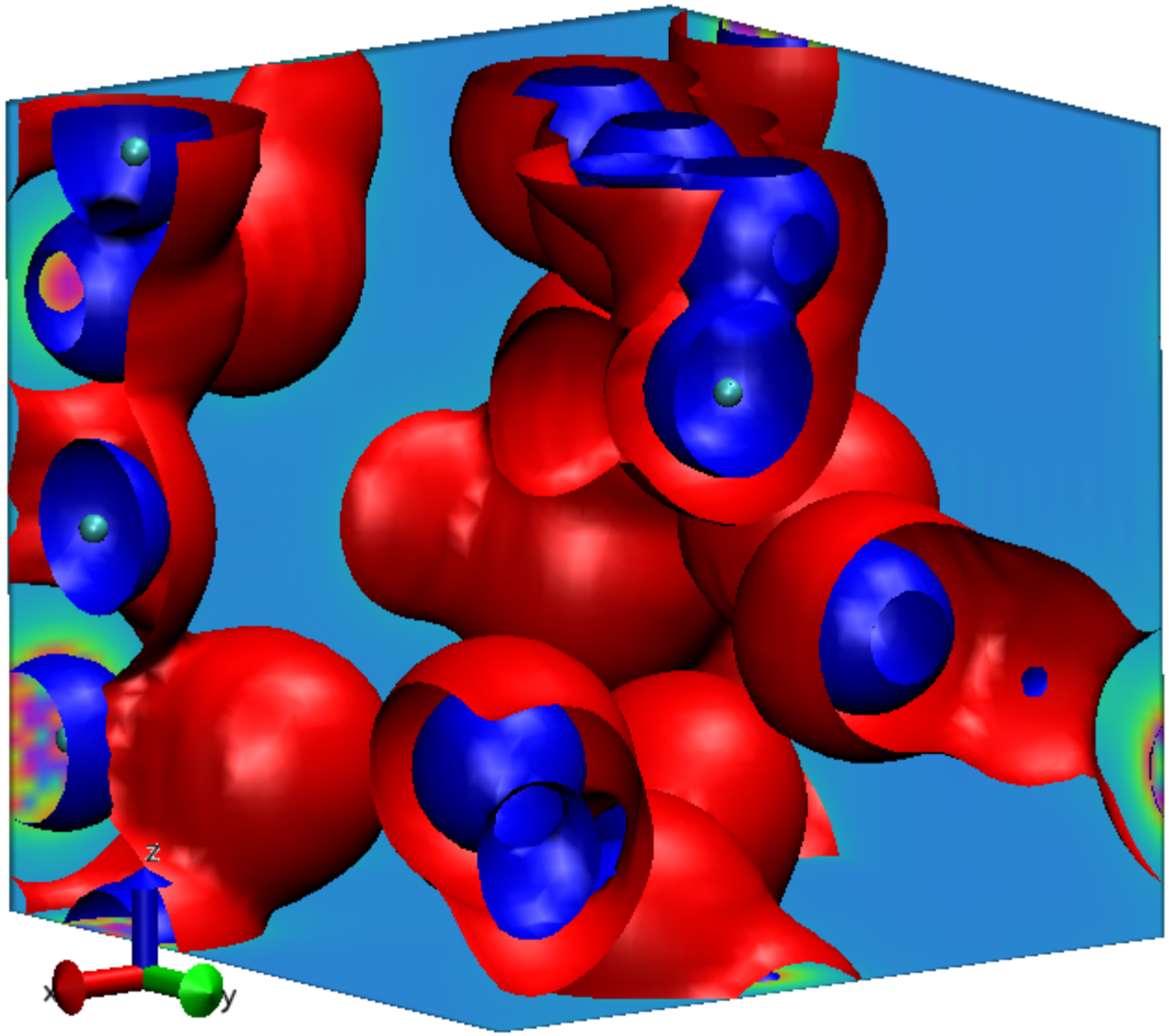
Meiko Computing Surface

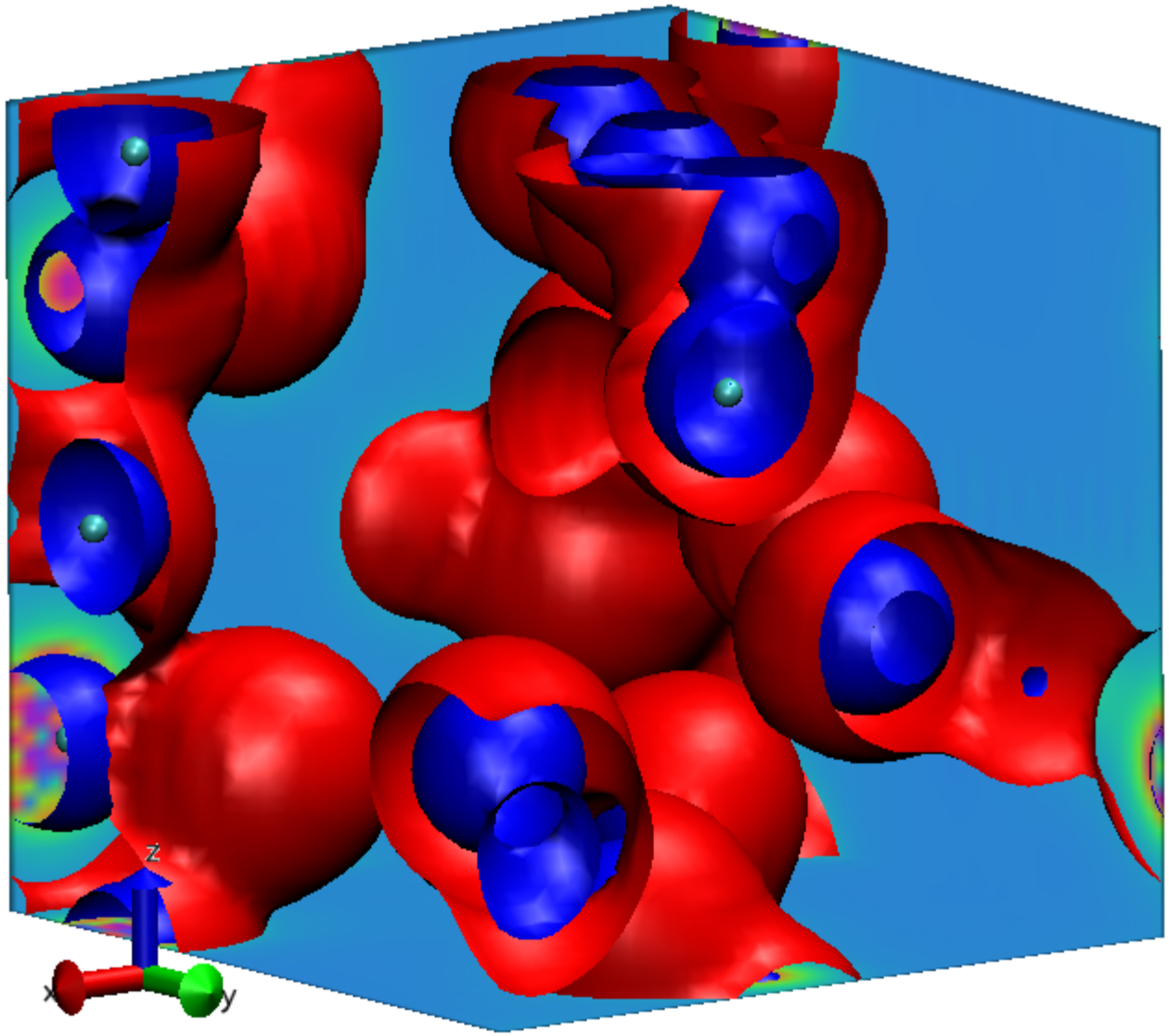
Атомистические модели воды



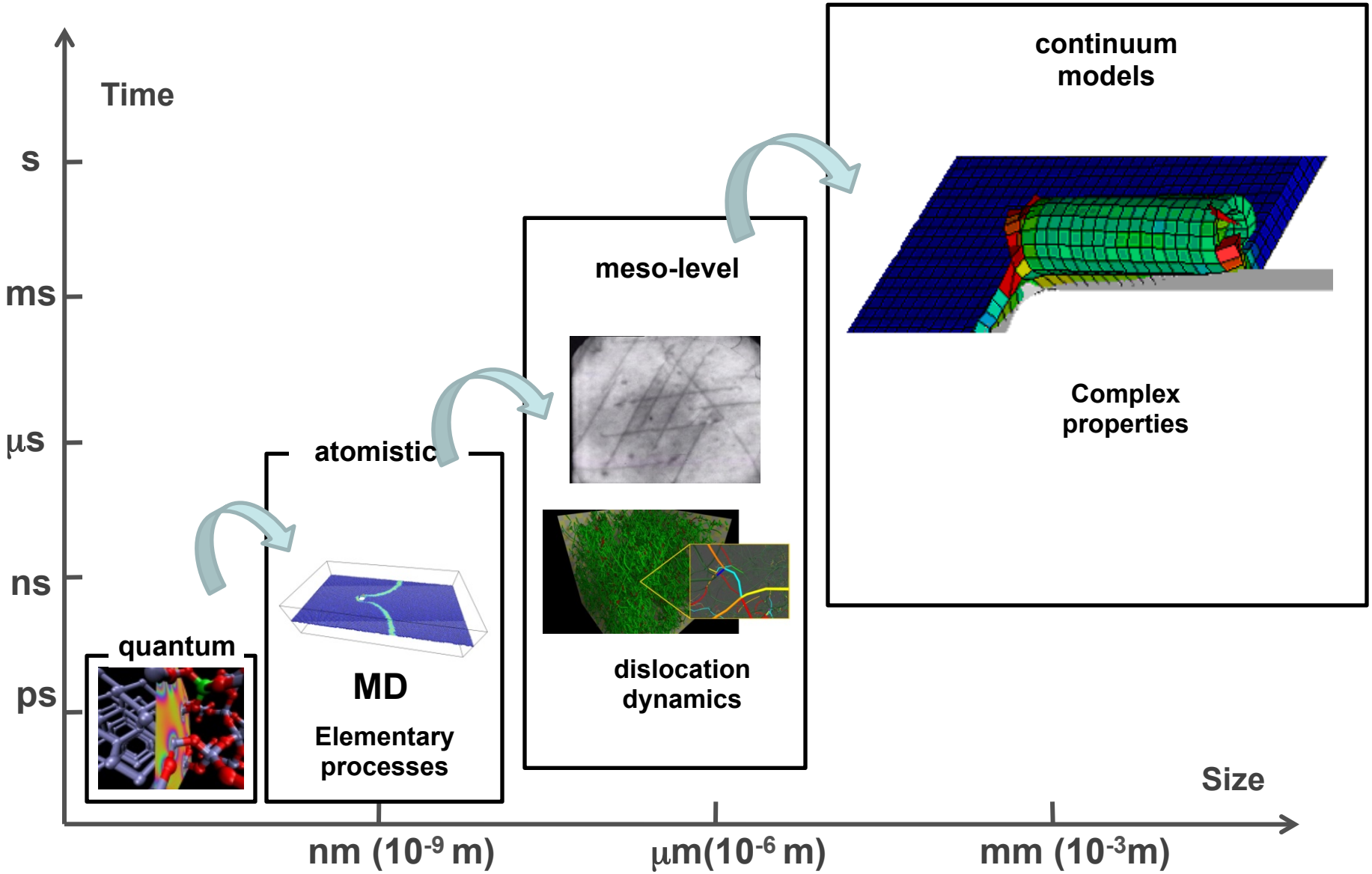


© 2011 by Yann von Hansen


z
x
y


z
x
y

Многомасштабный подход в теории и моделировании



Стратегия развития МД modeling&simulation

геометрия поверхности

Пластичность нанокристаллических материалов

Композиты с нанотрубками

пластичность и разрушение монокристаллов

одиночные дислокации

теоретическая физика,
computer science,
сравнение с экспериментом

зарождение полостей и дислокаций, кристаллизация

нанотрубки

УРС

оптимальное число вычислительных ядер M

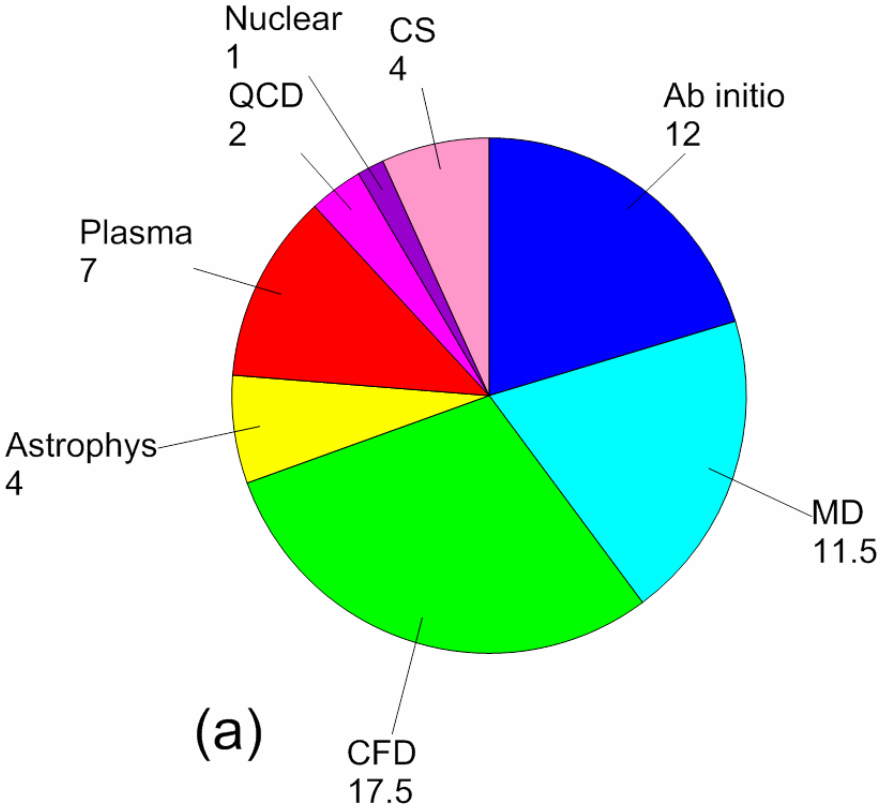
10^0 10^1 10^2 10^3 10^4 10^5 10^6

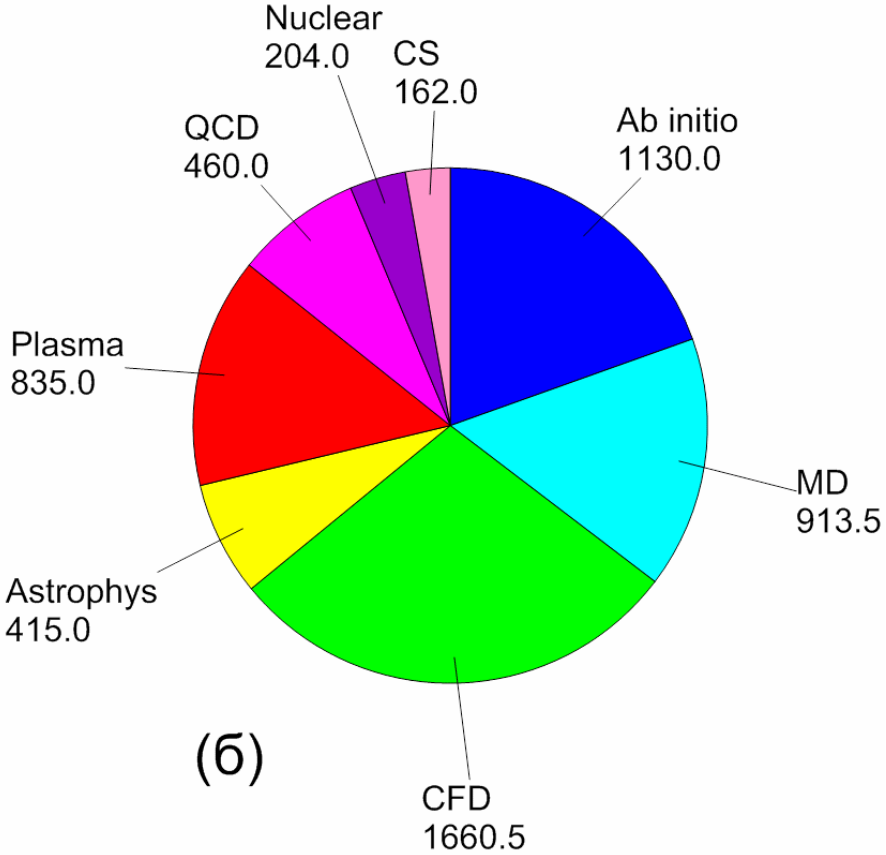
число атомов

10^2 10^3 10^4 10^5 10^6 10^7 10^8 10^9 10^{10} 10^{11} 10^{12}

Деление проектов INCITE 2014 г. по тематикам

Innovative and Novel Computational Impact on Theory and Experiment





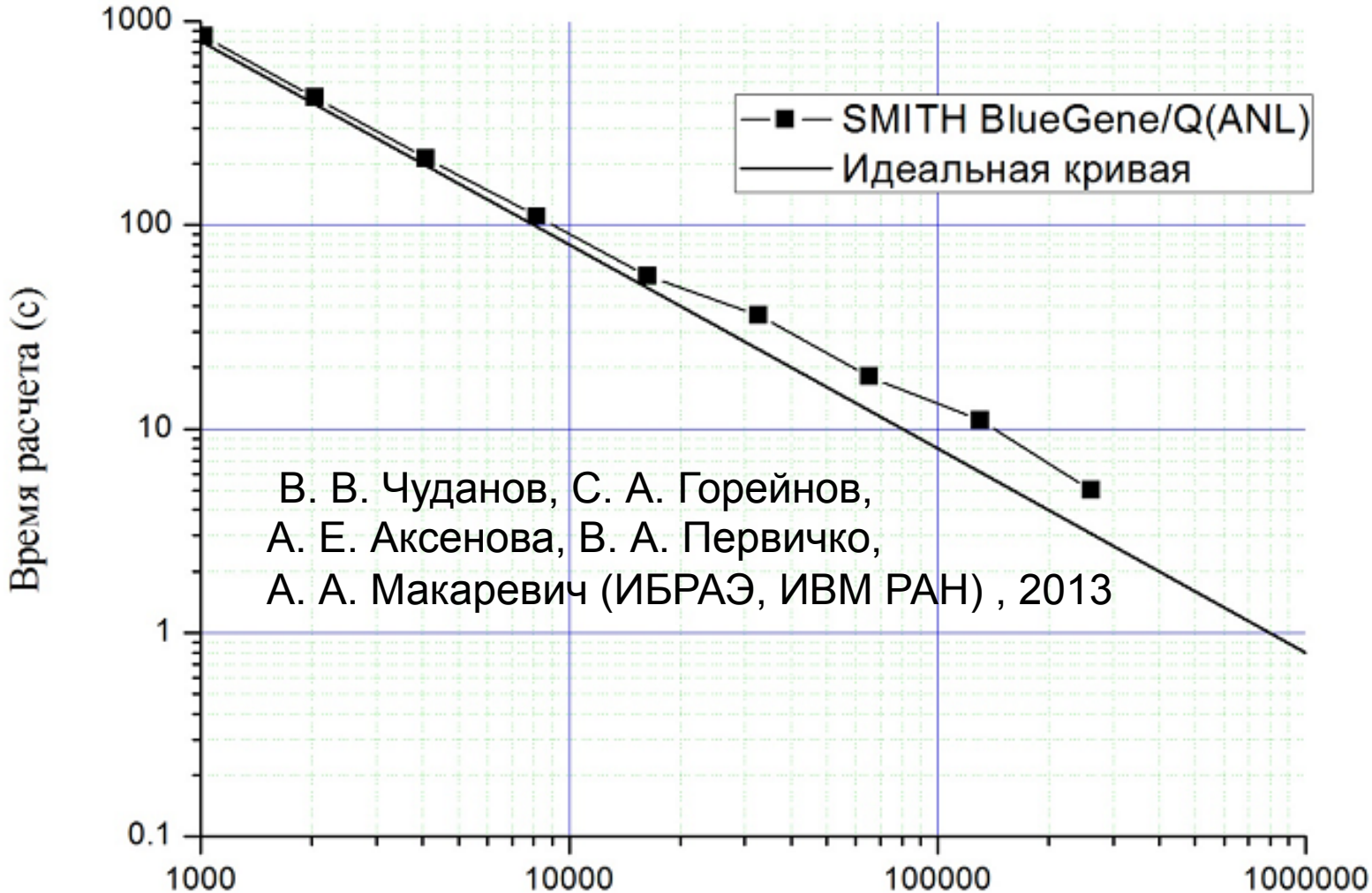
Число проектов по каждой тематике

Выделенное на каждую тематику вычислительное время в миллионах процессоро-часов

Computational fluid dynamics

SMITH: «сильная» масштабируемость, $\frac{p_0 T(N,p_0)}{pT(N,p)}$.

Ниже $p_0 = 1024$, максимальное $p = 262144$, $N = 4097^3$. Сплошная линия соответствует зависимости $T = p_0 T(N, p_0)/p$.



В. В. Чуданов, С. А. Горейнов,
А. Е. Аксенова, В. А. Первичко,
А. А. Макаревич (ИБРАЭ, ИВМ РАН) , 2013

Estimates from the ISC-2015 Conference

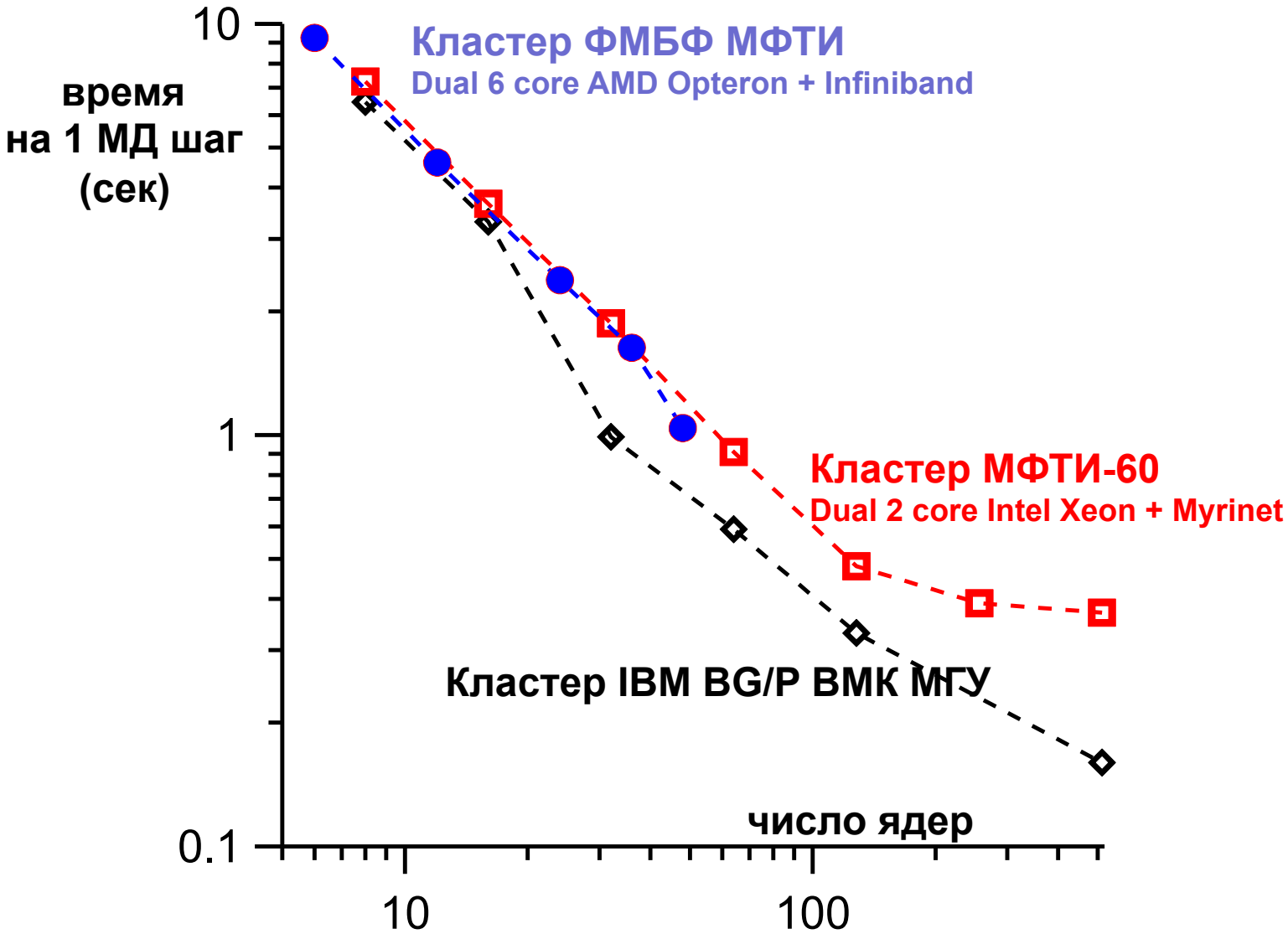
Rough estimate of DNS flow over an aircraft

An airbus 310 cruising at 250 m/s at 10000 m

For 1 second of simulated flight:
1 teraflops machine – $8*10^5$ years
50 exaflops machine – 1 weak

(based on John Kim, TSFP-9, 2015)

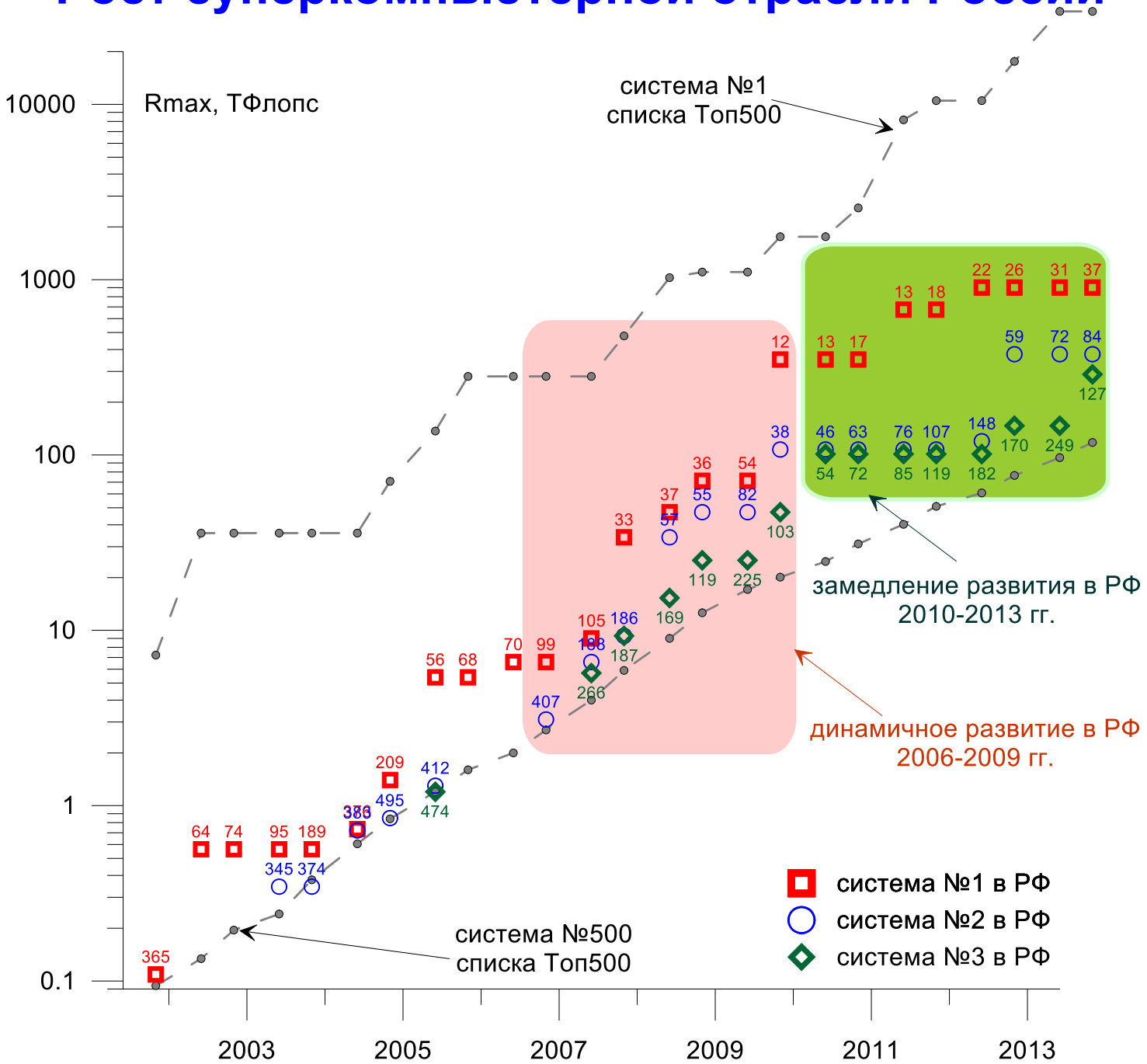
Электронно-структурная модель 32 молекул воды



Жиляев П.А., Стегайлов В.В. **Выч. методы и прогр.** 2012. Т.13. С.37-45.

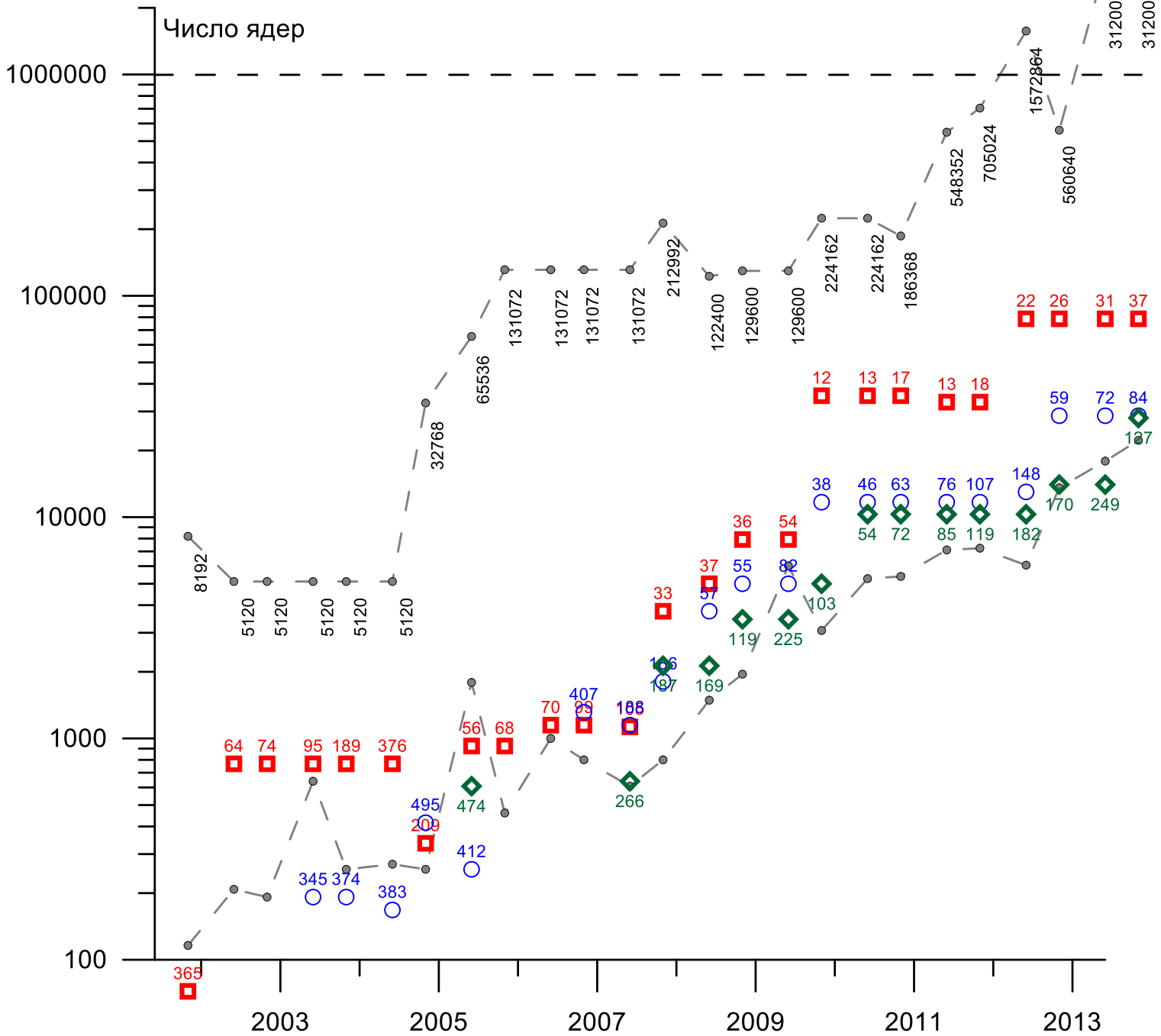
Сложный путь развития суперкомпьютерной отрасли России

Рост суперкомпьютерной отрасли России

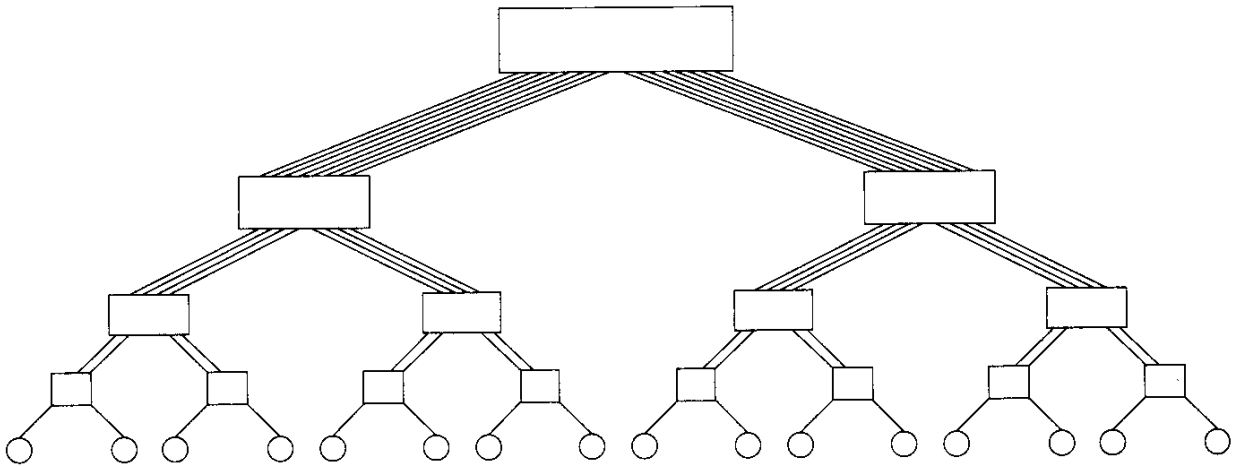


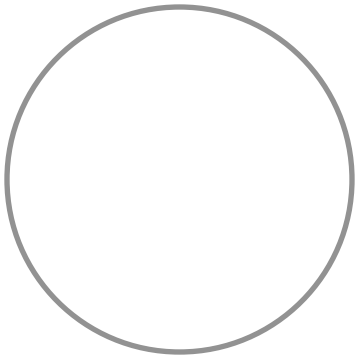
Движение в «эру экзафлопса»

Суперкомпьютеры

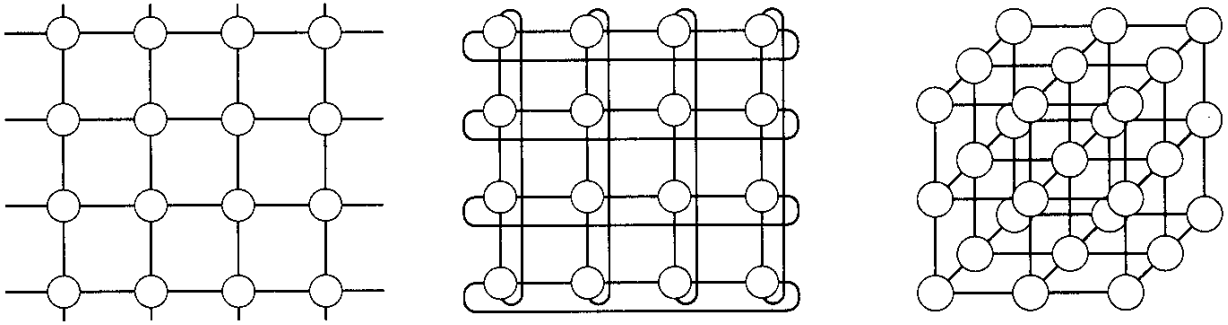


Топология интерконнекта и архитектура





Message Passing Interface



(a) (b) (c)

OpenMP
CUDA
OpenCL
ПЛИС
….

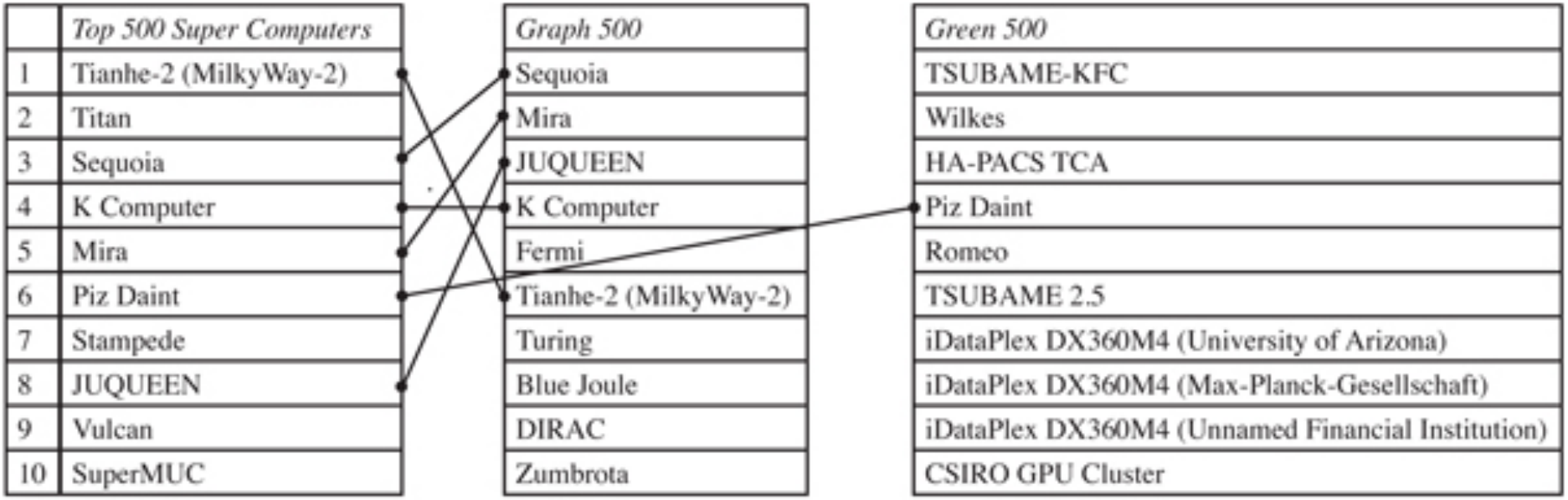
«MPP» vs «SMP»

Особенности устройства суперкомпьютеров №№1-3 текущего списка Топ500 2013 г.

	Кол-во узлов	Кол-во вычисл. ядер ЦПУ	Узел		
			ЦПУ (кол-во ядер)	Ускоритель (кол-во ядер)	Суммарная теор. произв. узла, ТФлопс
Tianhe-2	16 000	384 000	2 x Intel Xeon (2 x 12)	3 x Intel Phi (3 x 57)	3.43
Cray Titan	18 688	299 008	AMD Opteron (16)	Nvidia Kepler (14 потоковых мультипроцессора)	1.45
IBM Sequoia	98 304	1 572 864	IBM PowerPC (16 + 2)		0.205

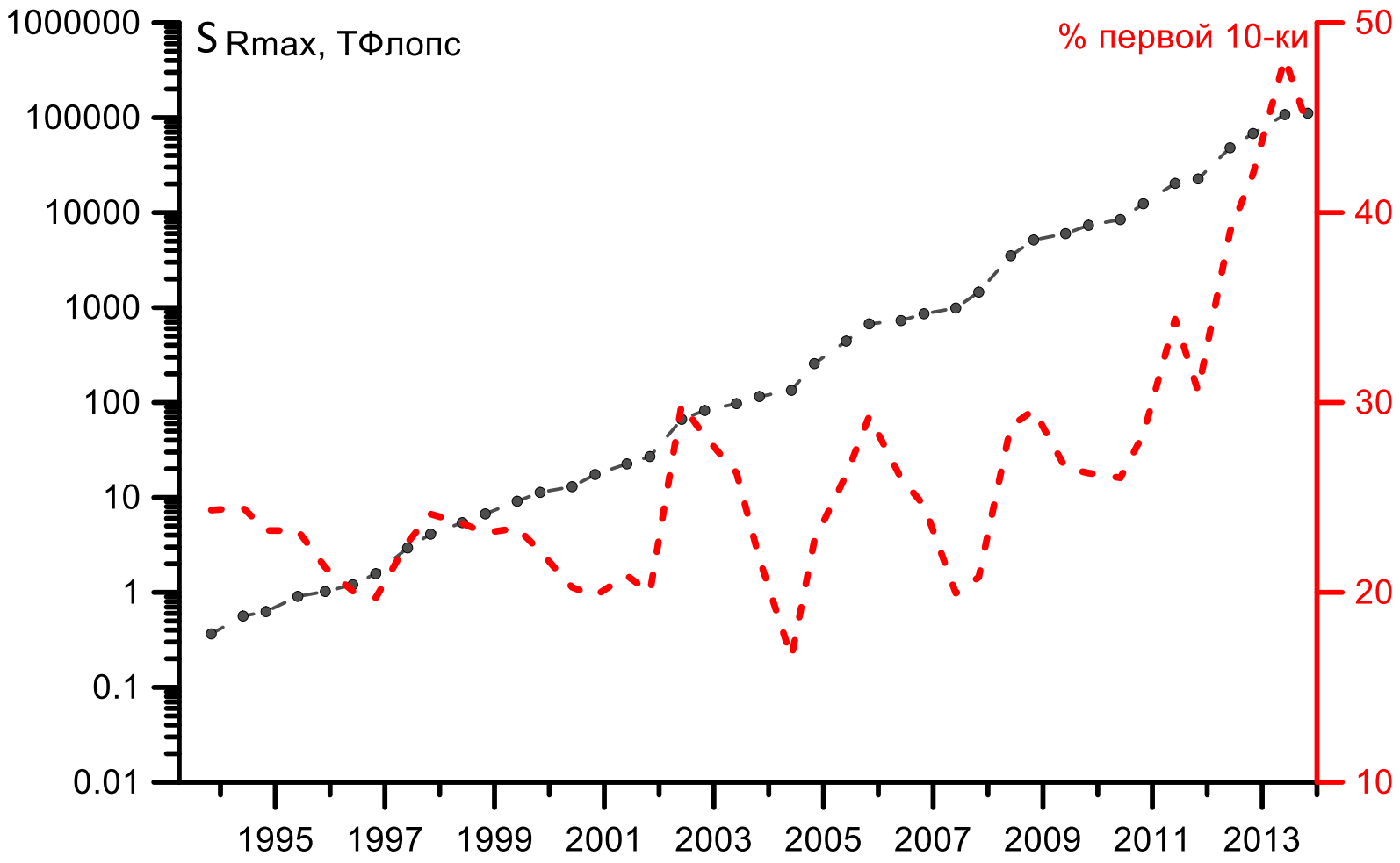
Эффективность Rmax/Rpeak, ГФлопс/Вт

	Rmax, ПФлопс	Rpeak, ПФлопс	Rmax/Rpeak, %	МВт	ГФлопс /Вт
Tianhe-2	33.7	54.9	61%	17.8	0.733
Cray XK7 «Titan»	17.6	27.1	65%	8.21	2.143
IBM BlueGene/Q «Sequoia»	17.2	20.1	81% → 85%	7.89	2.177
Кластер «Ломоносов»	0.9	1.7	53%	2.8	0.322

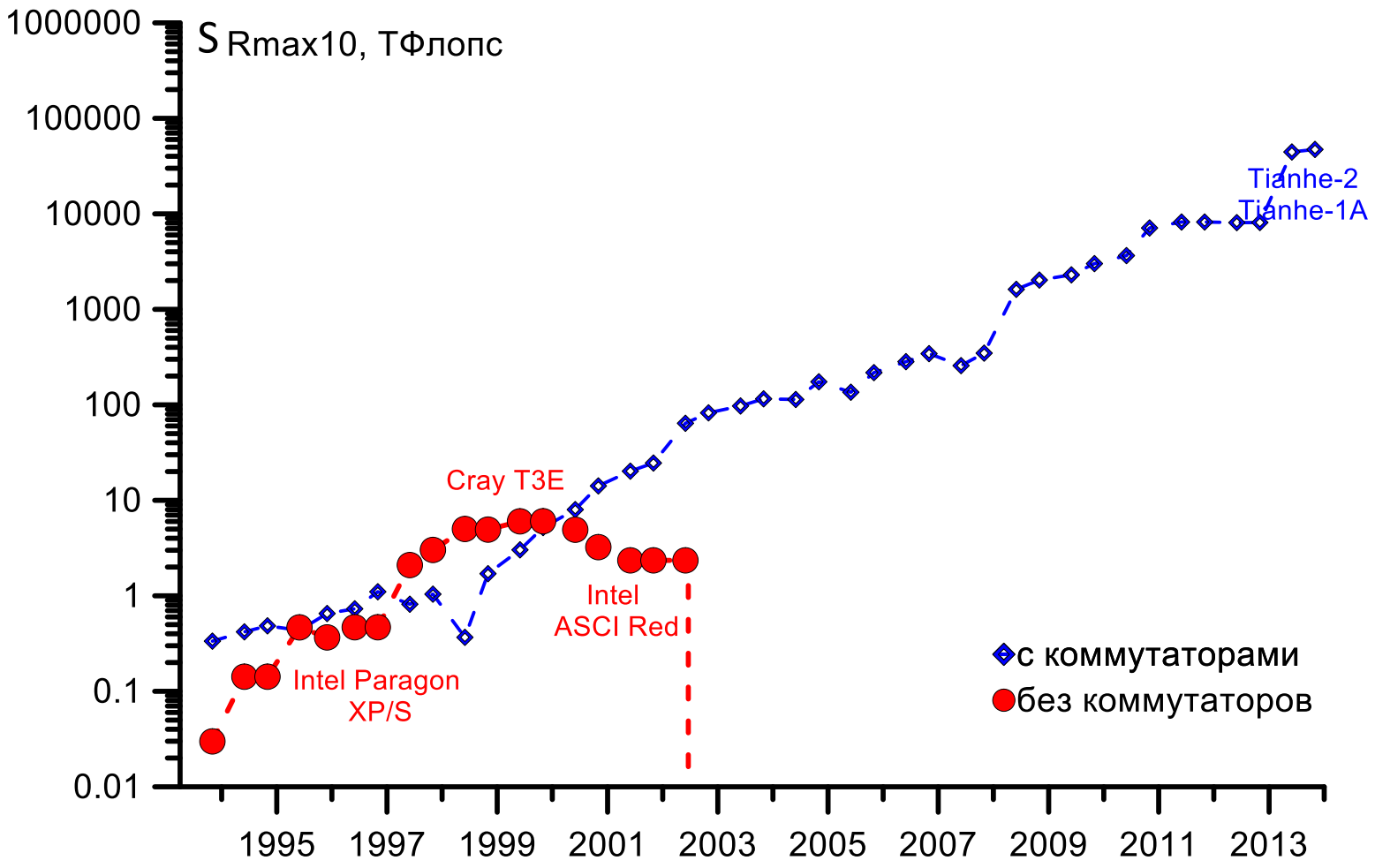


Топология интерконнекта (что скрывает MPI?)

Развитие списка Топ500



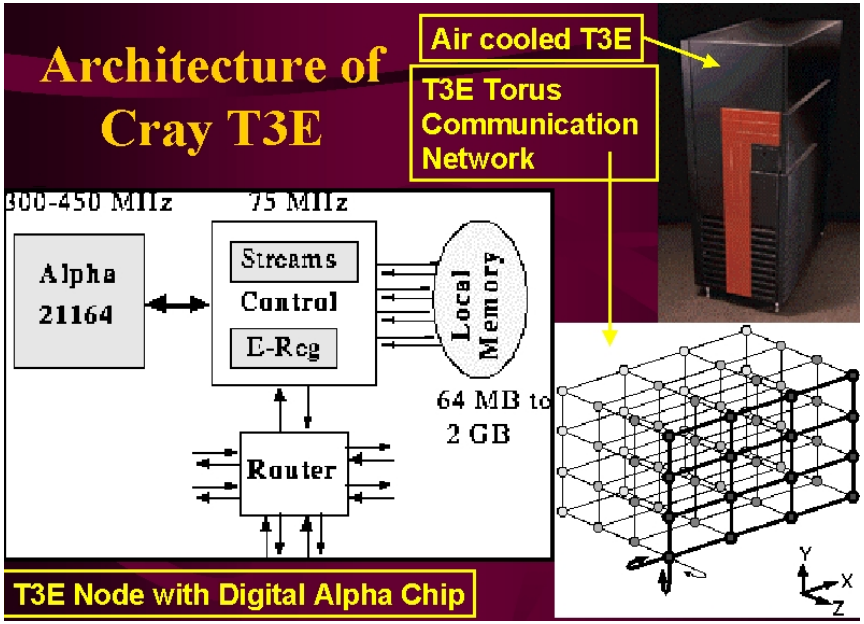
Развитие списка Топ500



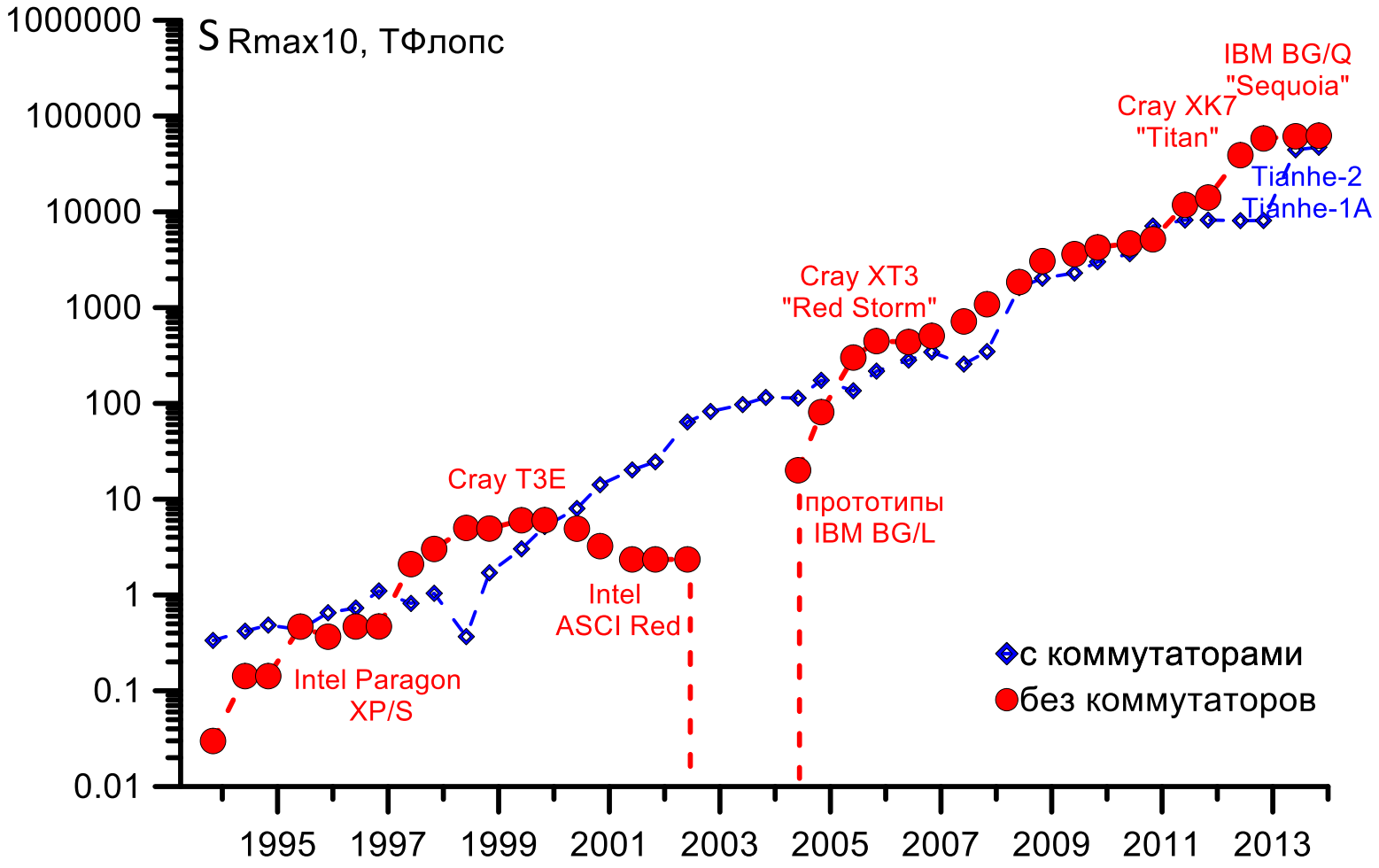
Cray T3E

Число суперкомпьютеров Cray T3E
в верхней 10-ке списков Топ500
в 1996-1999 гг.

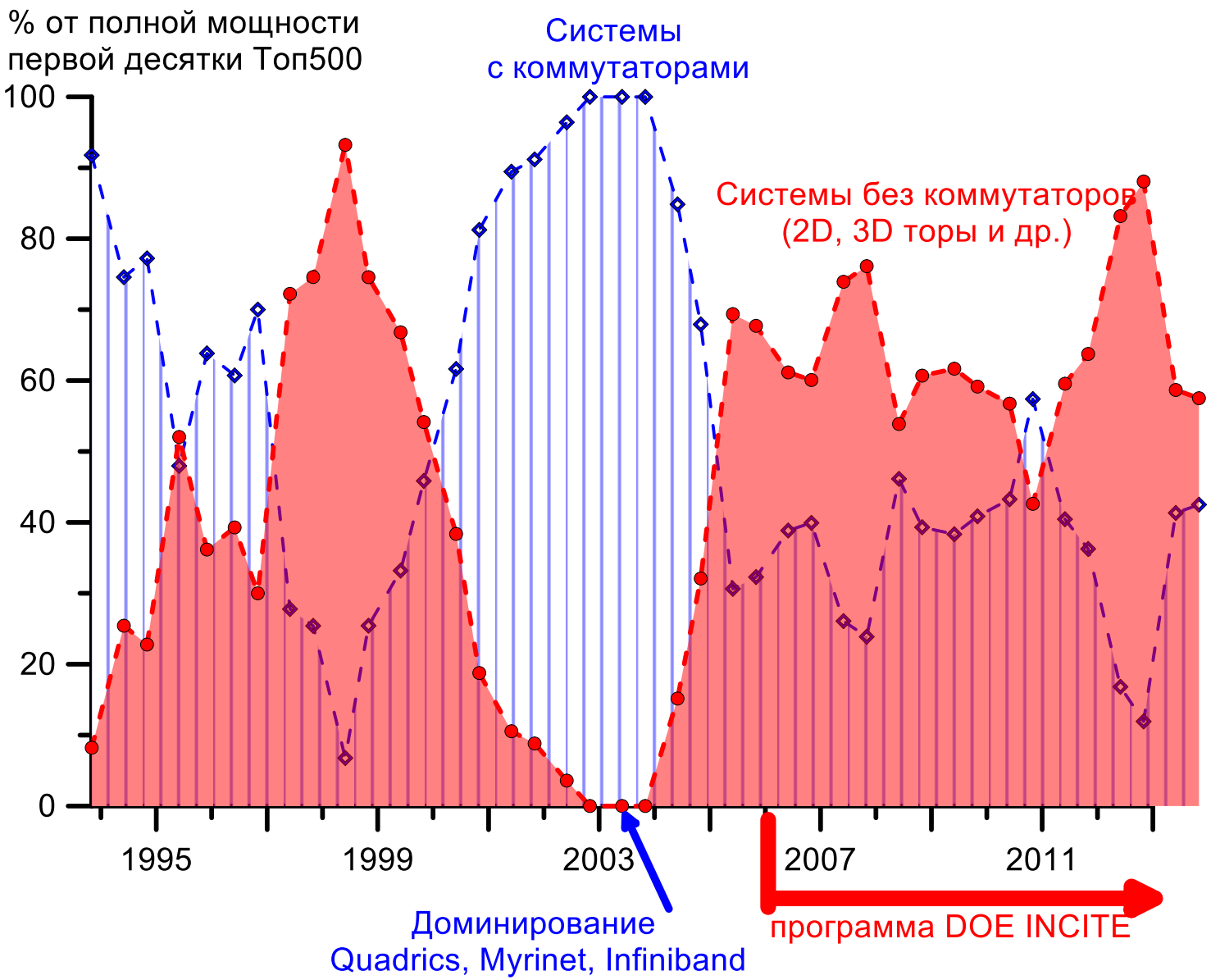
11/1996	06/1997	11/1997	06/1998	11/1998	06/1999	11/1999
1	**6**	**5**	**8**	**6**	**5**	**4**



Развитие списка Топ500



Развитие списка Топ500



Тороидальные системы в мире

Регион/страна	Кол-во	Число ядер	ΣRmax, ТФлопс	% от ΣRmax	Название
Россия	1	8192	23.9	0.01	IBM – 1.
Европа	17	1 406 240 (+45 712)	14 371.6	6.4	IBM – 8, Cray – 9.
Австралия, Бразилия, Канада, Сауд.Аравия	4	194 944 (- 8 192)	1 465.8	0.65	IBM – 3, Cray – 1.
США	35	4 810 064 (+367 416)	59 812.9	26.7	IBM – 18, Cray – 16, Sun – 1.
Япония и Южная Корея	10	1 042 688 (+23 808)	14 202.9	6.3	Fujitsu -3, IBM – 2, Cray – 5.
Итого:	67	7 462 128 (+428 744)	89 877.1	40.1	Fujitsu -3, IBM – 32, Cray – 31, Sun – 1.

Число и суммарные параметры систем с тороидальной топологией по странам мира в списке Топ500 июня 2013 г.: количество подобных систем, суммарное число ядер (в скобках – прирост по отношению к предыдущему списку), суммарная производительность, вклад в суммарную производительность списка июня 2013 г. (224 ПФлопс), распределение систем по производителям. (Китай пока не представлен в Топ500 такими суперкомпьютерами)

MPI: утрата универсальности?

Faraj A., Kumar S., Smith B., Mamidala A., Gunnels J., **MPI collective communications on the Blue Gene/P supercomputer: algorithms and optimizations** // 17th IEEE Symposium on High Performance Interconnects. 2009. P. 63-72.

за период немногим больше 3 лет была дана **31 ссылка** в научной периодике (авторы цитирующих статей **из Австрии, Германии, Индии, Италии, Испании, Китая, США и Турции**)

Chen D., Eisley N. A., Heidelberger P., Senger R. M., Sugawara Y., Kumar S., Salapura V., Satterfield D. L., Steinmacher-Burow B., Parker J. J. **The IBM BlueGene/Q interconnection network and message unit** // Proc. Int. Conf. for High Perf. Computing, Networking, Storage and Analysis (SC), 2011.

за период немногим больше 1 года было дана **21 ссылка** в научной периодике (авторы цитирующих статей из **Бельгии, Германии, Индии, Испании, Китая и США**).

MPI: утрата универсальности?

Адамович И. А., Климов А. В., Климов Ю. А., Орлов А. Ю., Шворин А. Б. **Опыт разработки коммуникационной сети суперкомпьютера «СКИФ-Аврора»** // Программные системы: теория и приложения. 2010. Т. 1. Вып. 3. С. 107–123.

Корж А. А., Макагон Д. В., Бородин А. А., Жабин И. А., Куштанов Е. Р., Сыромятников Е. Л., Черемушкина Е. В. **Отечественная коммуникационная сеть 3D-тор с поддержкой глобально адресуемой памяти** // Вест. ЮУрГУ - серия «Мат. модел. и программирование». 2010. Т. 211. № 35. С. 41-53.

Интерконнект

«Сердце» современного суперкомпьютера - коммутационная сеть (интерконнект) проявляется на 3 уровнях:

1. Оборудование и топология сети, т.е. принцип физического объединения узлов каналами обмена данных.

2. Системное программное обеспечение, реализующее стандартные процедуры обмена данными (один-одному, один-всем, все-всем и т.п.)

3. Алгоритмы параллельного решения математической задачи, основанные на указанном системном программном обеспечении.

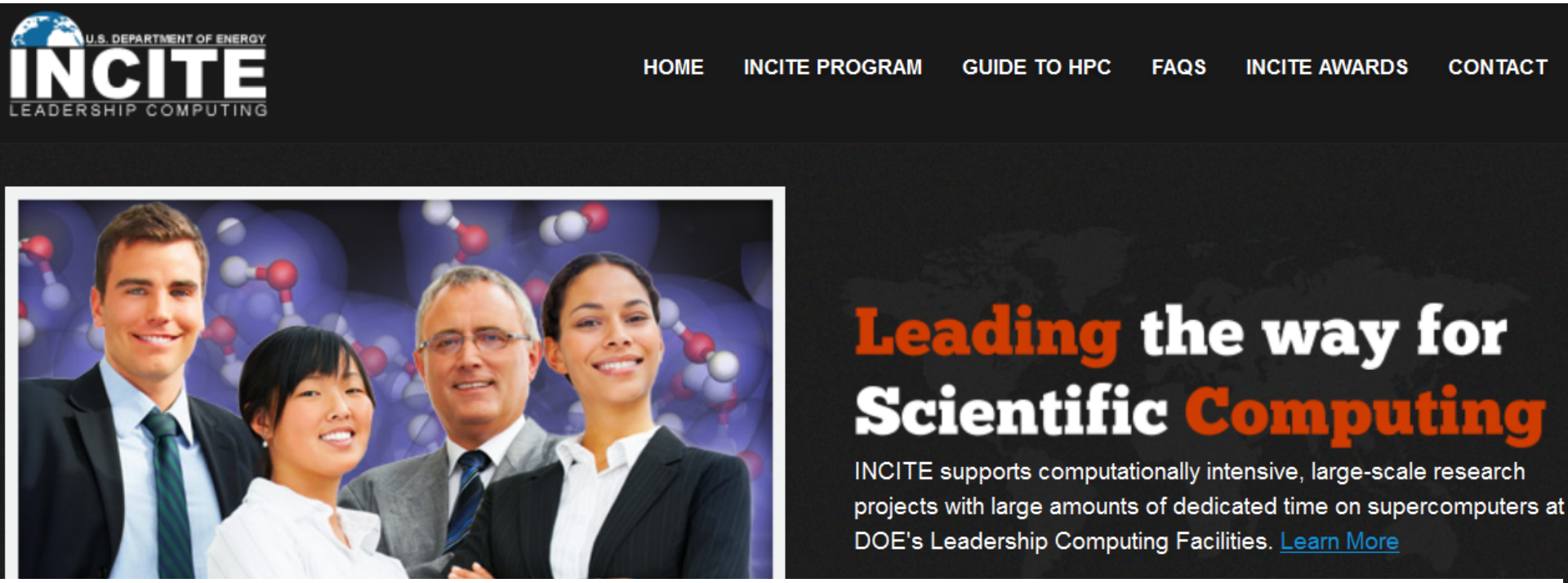
Инерционность развития!

Нужен ли в действительности массовый паралелизм?

EU PRACE initiative

US DOE INCITE initiative

Innovative & Novel Computational Impact on Theory and Experiment



	DOE	University	Government	Industry	International
DOE	**12**	14	1	0	1
University		**17**	0	3	2
Government			**0**	0	1
Industry				**1**	0
International					**3**

University-DOE-Industry	2
University-DOE-International	3

US DOE INCITE initiative

Год	Всего проектов	Проекты по молекулярному и атомистическому моделированию	% от общего числа
2006	15	6	40%
2007	45	11	24%
2008	58	16	28%
2009	66	20	30%
2010	70	23	33%
2011	57	18	32%
2012	60	23	38%
2013	61	24	39%

2012 год:

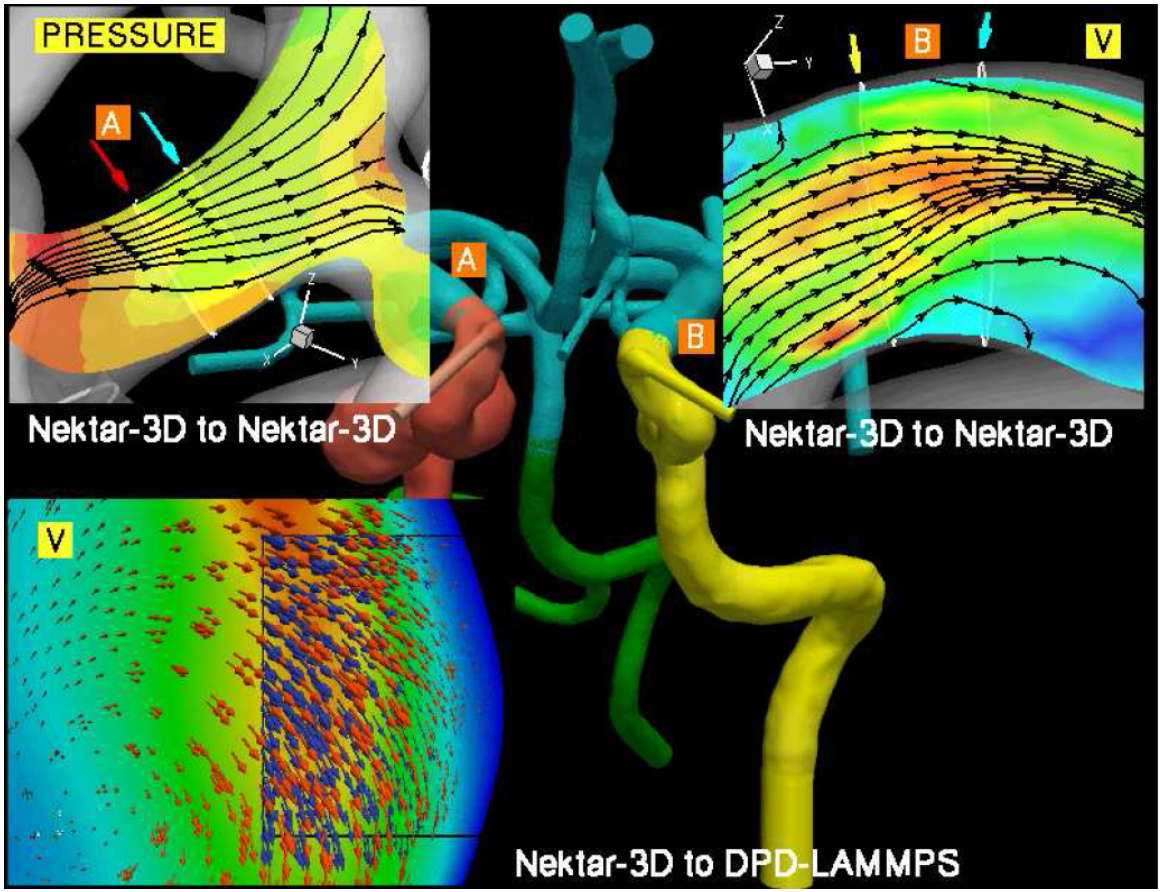
60 проектов = 1672 миллиона процессор-часов на IBM BlueGene/P и Cray XT5

23 проекта по молекулярному и атомистическому моделированию = 518 млн. час.

31 % общего времени

US DOE INCITE initiative

Пример результатов комбинированного расчета кровотока с эффектами отложения холестериновых бляшек в рамках многомасштабной модели, сочетающей атомистическое моделирование с моделью жидкости Навье-Стокса



Grinberg L., et al. **A new computational paradigm in multiscale simulations: Application to brain blood flow** // Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2011.

US DOE INCITE initiative

Демонстрация сохранения эффективности распараллеливания решения одной задачи на системах BlueGene/P и Cray XT5

Число ядер	Время загрузки ЦПУ, с	Эффективность
BlueGene/P (4 ядра на узел)		
28 672	3205.58	
61 440	1399.12	1.07
126 976	665.79	1.02
Cray XT5 (12 ядер на узел)		
17 280 (Kraken)	2194	
25 920 (Kraken)	1177	1.24
34 560 (Jaguar)	806	1.10
93 312 (Jaguar)	280	1.07
186 624 (Jaguar)	206	0.68

Grinberg L., et al. **A new computational paradigm in multiscale simulations: Application to brain blood flow** // Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2011.

US DOE INCITE initiative

Исследование, выполненное в приведенной работе, состоит из следующих частей:

- **связь континуального и дискретного алгоритмов с разными временными шагами,**
- **обмен данными в процессе расчета между частями системы и между континуальным и дискретным уровнями,**
- **оптимизация производительности с учетом специфики систем IBM BlueGene/P и Cray XT5,**
- **анализ параллельной эффективности на обеих системах.**

Grinberg L., et al. **A new computational paradigm in multiscale simulations: Application to brain blood flow** // Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2011.

Эти цифры демонстрируют переход к новой эре использования вычислительных методов в науке и технике

Пример из программы SC’13

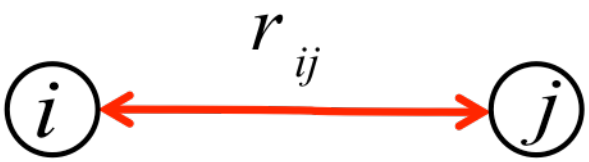
TITLE: **Petascale Direct Numerical Simulation of Turbulent Channel Flow on up to 786K Cores**
SESSION: Extreme-Scale Applications
EVENT TYPE: Papers, Awards, Best Student Paper Finalists
TIME: 1:30PM - 2:00PM
SESSION CHAIR: Michael A. Heroux
AUTHOR(S): Myoungkyu Lee, Nicholas Malaya, Robert D. Moser
ROOM:201/203
ABSTRACT:
We present results of performance optimization for direct numerical simulation (DNS) of **wall bounded turbulent flow (channel flow)**. DNS is a technique in which the fluid flow equations are solved without subgrid modeling. **Of particular interest are high Reynolds number (Re) turbulent flows over walls,** because of their importance in technological applications. Simulating high Re turbulence is a challenging computational problem, due to the high spatial and temporal resolution requirements. An optimized code was developed using **spectral methods**, the method of choice for turbulent flows. Optimization was performed to address three major issues: efficiency of banded matrix linear algebra, cache reuse and memory access, and communication for the global data transposes. Results show that performance is highly dependent on characteristics of the communication network, rather than single-core performance. In our tests, it exhibits approximately **80% strong scaling parallel efficiency at 786K cores relative to performance on 65K cores**.

Органические жидкости (трансформаторные масла)

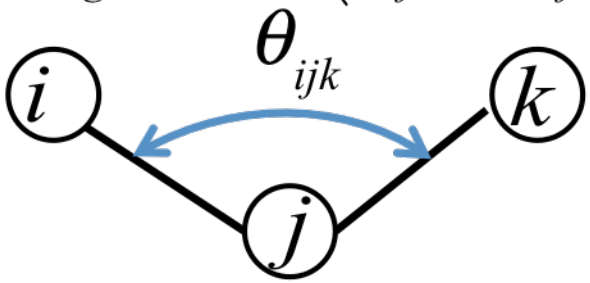
Межатомные и межмолекулярные взаимодействия

$$E = E_{bond} + E_{angle} + E_{dehedral} + E_{LJ} + E_{Coul}$$

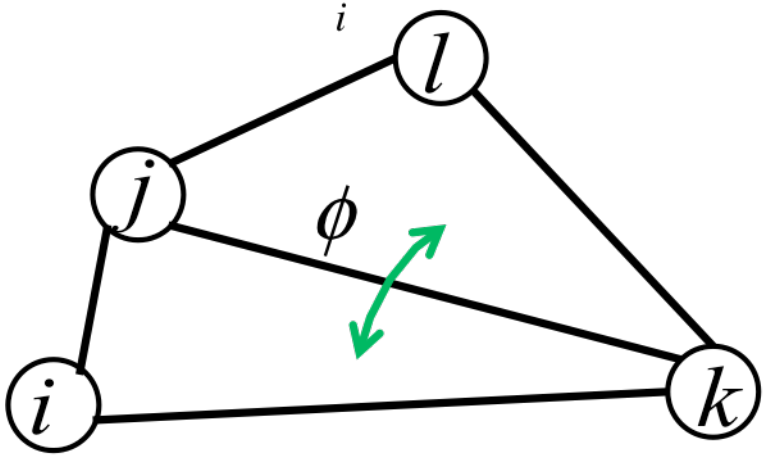
$$E_{bond} = K_b \left(r_{ij} - r_{ij}^{\ 0} \right)^2$$



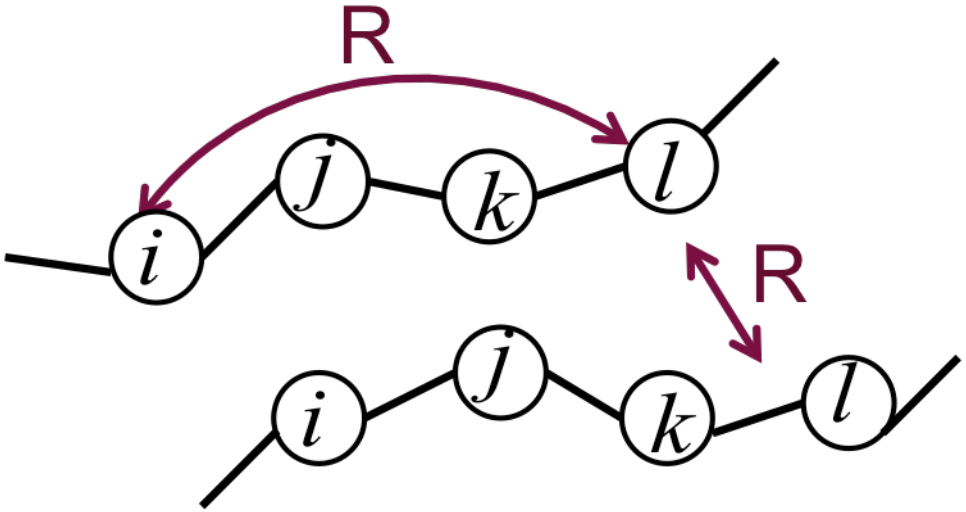
$$E_{angle} = K_a \left(\theta_{ijk} - \theta_{ijk}^{\ 0} \right)^2$$



$$E_{dehedral} = \sum_i K_i [1 - \cos(i \cdot \phi)]$$



$$E_{LJ} = 4\varepsilon[(\sigma / R)^{12} - (\sigma / R)^6]$$



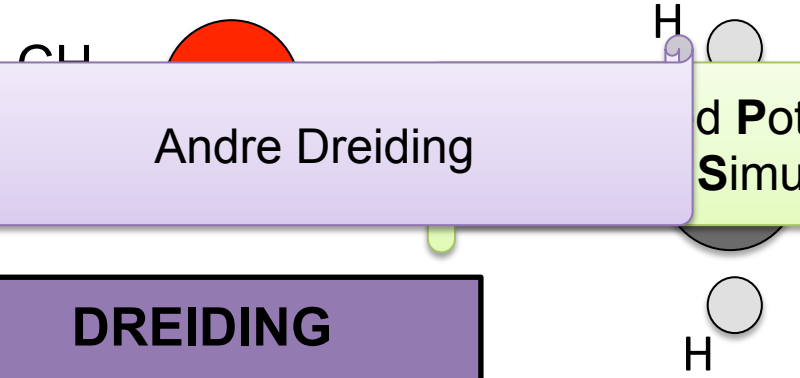
Различные приближения

количество молекул

Tranferable **P**otentials for **P**hase **E**quilibria

Модель объединенного атома

Полноатомное рассмотрение



Andre Dreiding

d **P**otential for **L**iquid **S**imulations

TraPPE

- объединенный атом
- простая параметризация
- нет зарядов
- Маленькое время на МД шаг

(~0.004 ГФлопс/шаг/мол)

DREIDING

- полноатомная
- средняя параметризация
- нет зарядов
- среднее время на МД шаг

(~0.04 ГФлопс/шаг/мол)

OPLS

- полноатомная
- сложная параметризация
- заряды на атомах
- большое время на МД шаг

(~0.3 ГФлопс/шаг/мол)

вычислительные затраты

Система

125-8000 молекул – TraPPE-UA (240000 атомов)

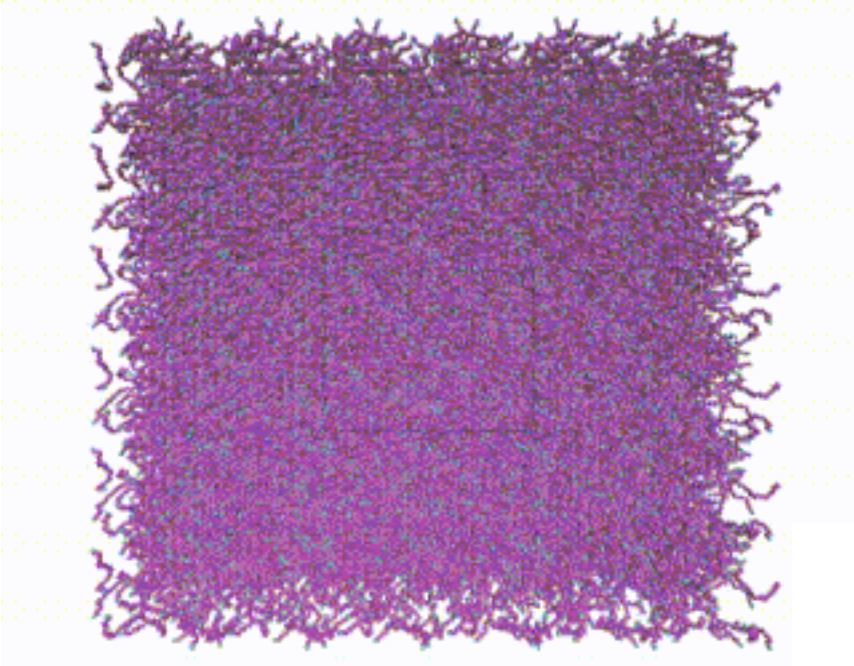
125-3375 молекул – DREIDING, OPLS (330000 атомов)

CH_3

CH_2

C. M.

$$\langle \Delta r^2 \rangle \qquad \langle v(0)v(t) \rangle$$



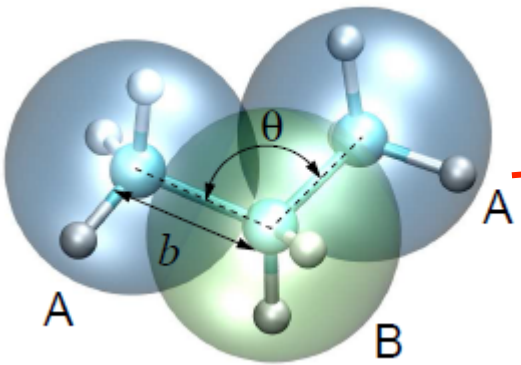
$\rho = 0.77$ g/cm^3

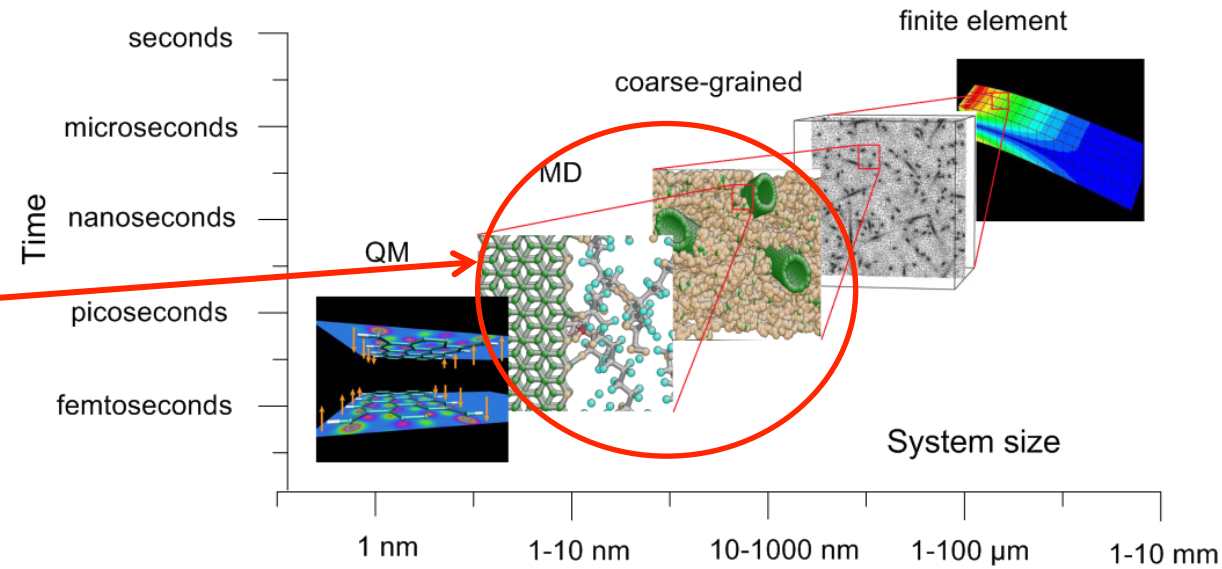
T = 353 K P ~ 1 atm

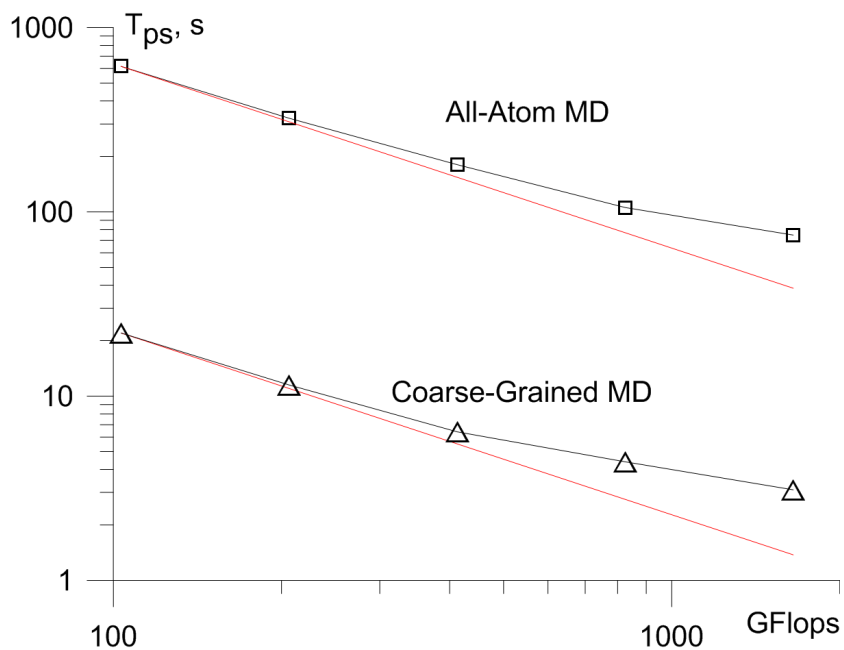
L = 4.8 ÷ 19.2 nm

Полимерные композиты

Coarse-grained Molecular Dynamics

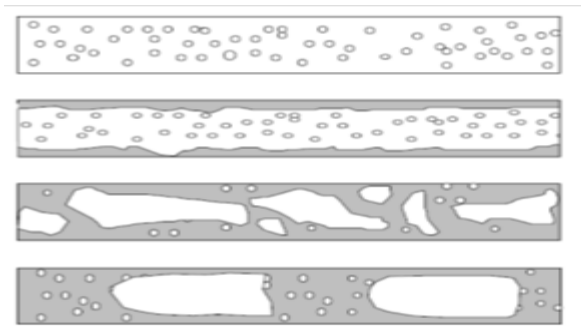




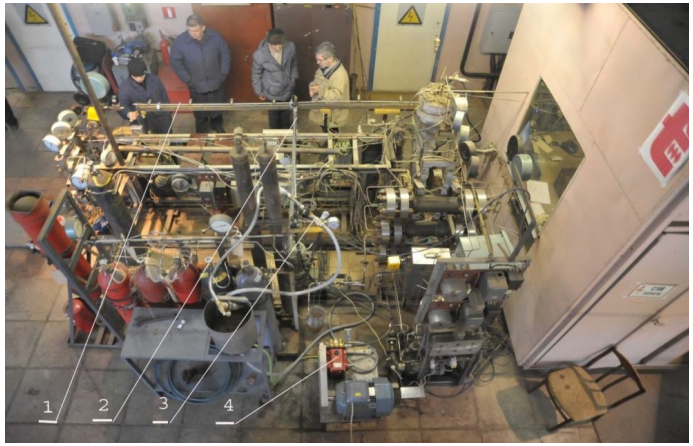


Газоконденсат
(фильтрация в пористых средах)

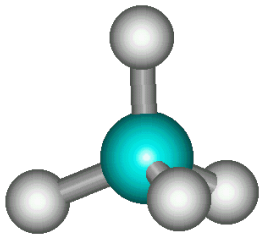
Мотивация работы

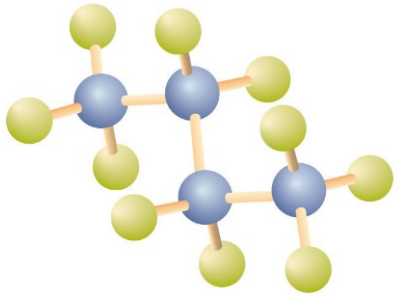


К гидродинамике двухфазных течений



С экспериментальной проверкой (стенд «Пласт», ОИВТ РАН)

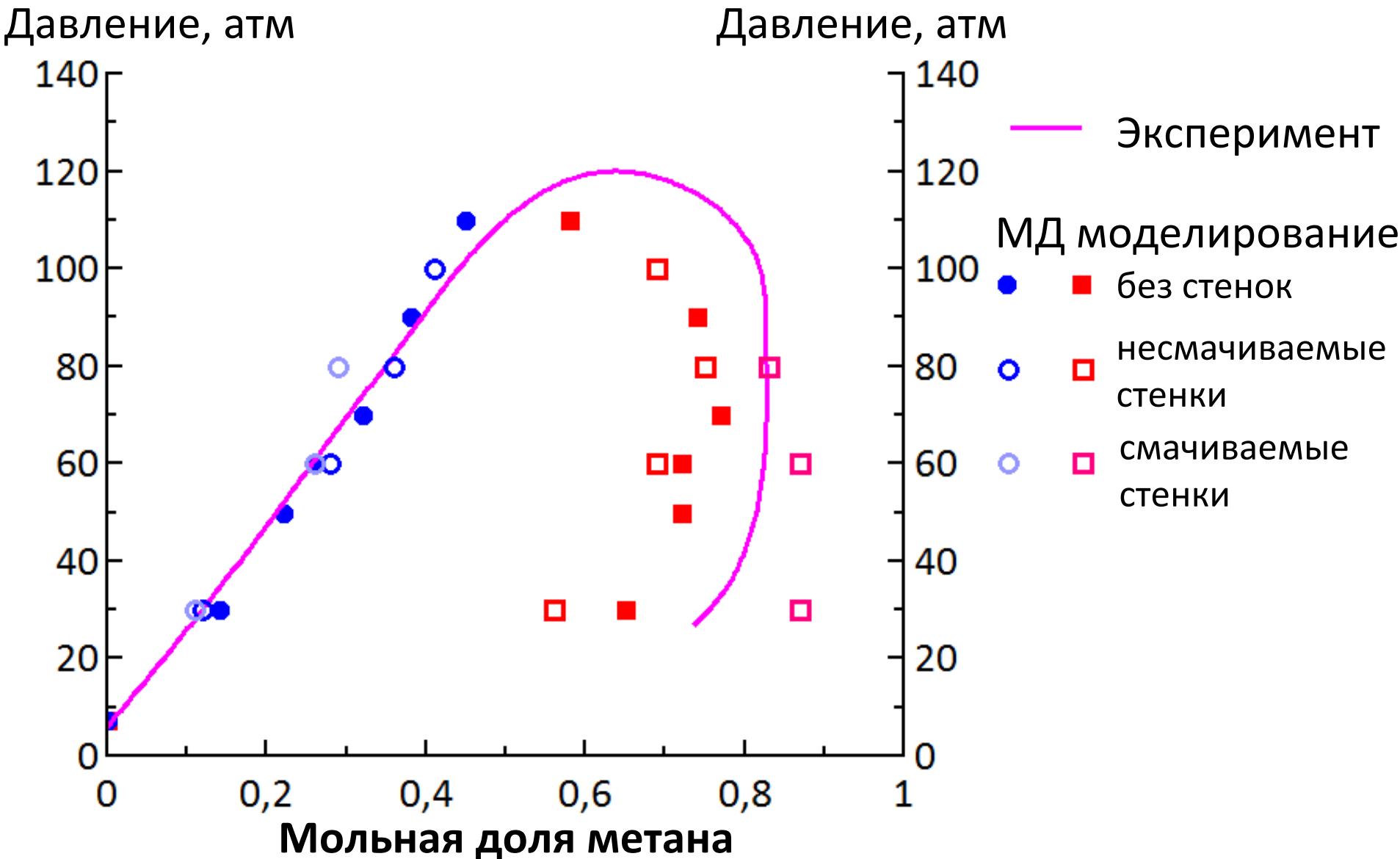




От атомистического моделирования

К решению одной из проблем газодобычи

Сверхкритическая фазовая диаграмма (метан+бутан, 330 К, TraPPE-UA)



Газовые гидраты

фазовые диаграммы
кинетика образования и разложения

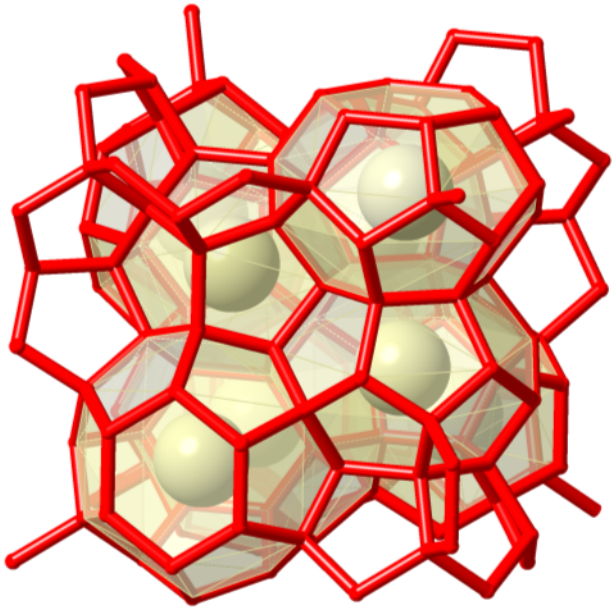
Methane Hydrates - Energy Source of the Future?

Estimated resource 1-5 million cubic kilometers of gas (500-2500 gigatones of carbon)

Lattice parameters of structure I

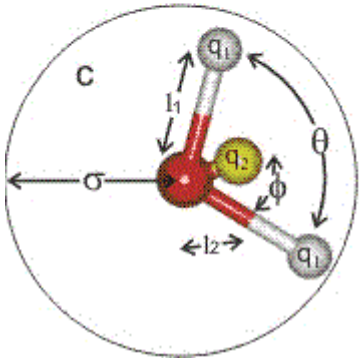
- Type: bcc
- Size ≈12x12x12 Å
- 8 methane molecules (100% cage occupancy)
- Elementary unit cell formula

$$(5^{12})_2 \cdot (5^{12}6^2)_6 \cdot 46\ H_2O$$

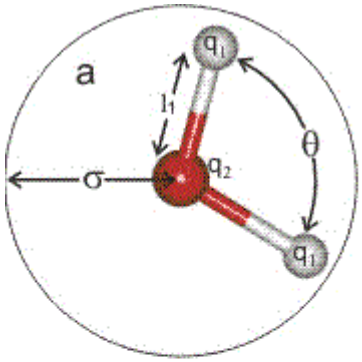


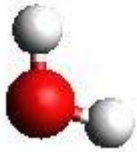
Interaction potentials for water

TIP4P - models



SPC/E



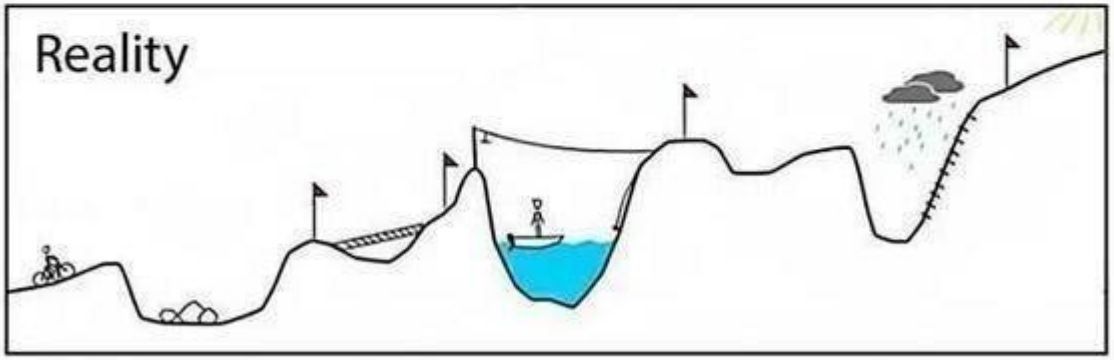


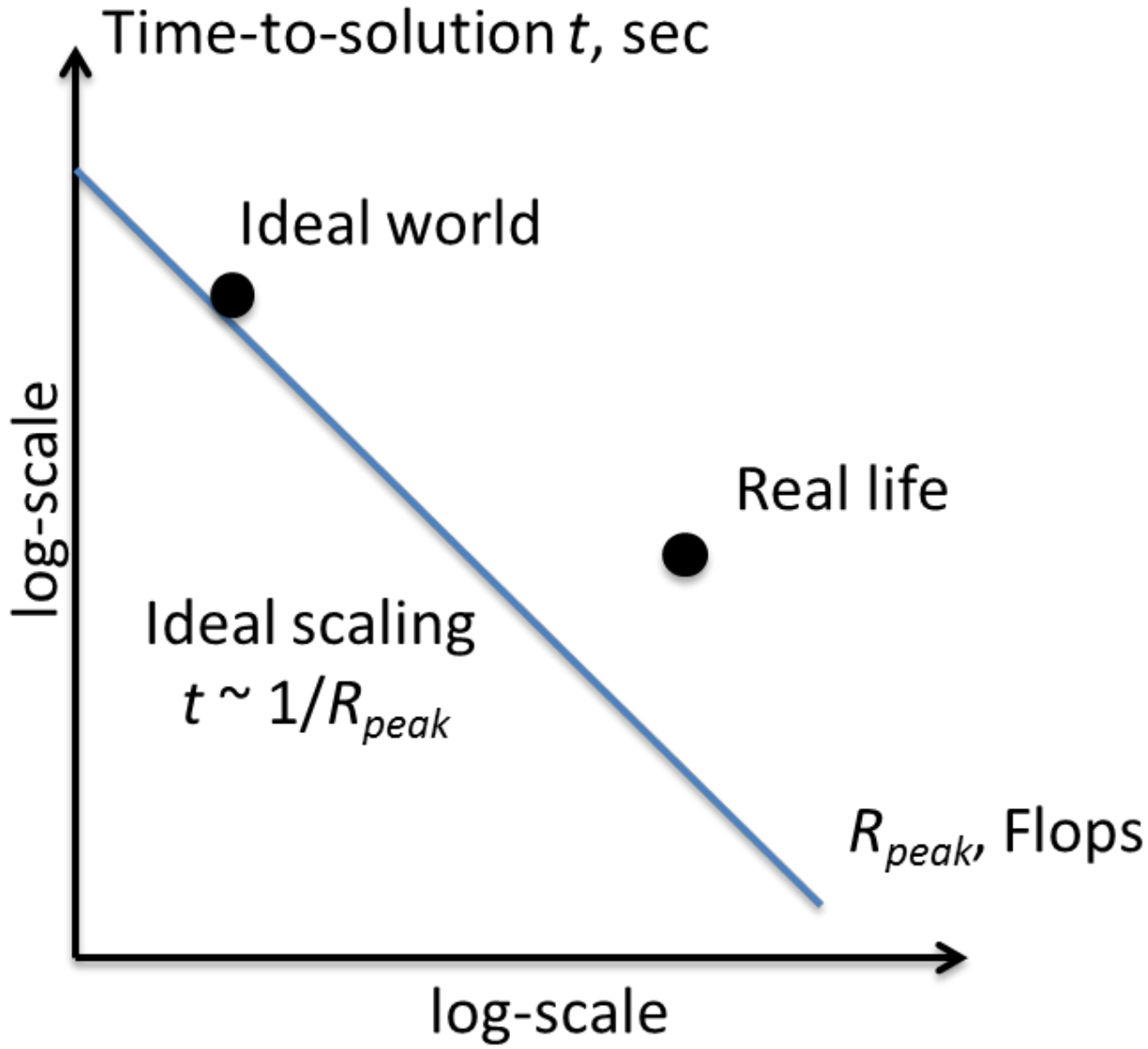
$$U_{ij} = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] + \frac{q_i q_j}{r}$$

Путь к экзафлопсу

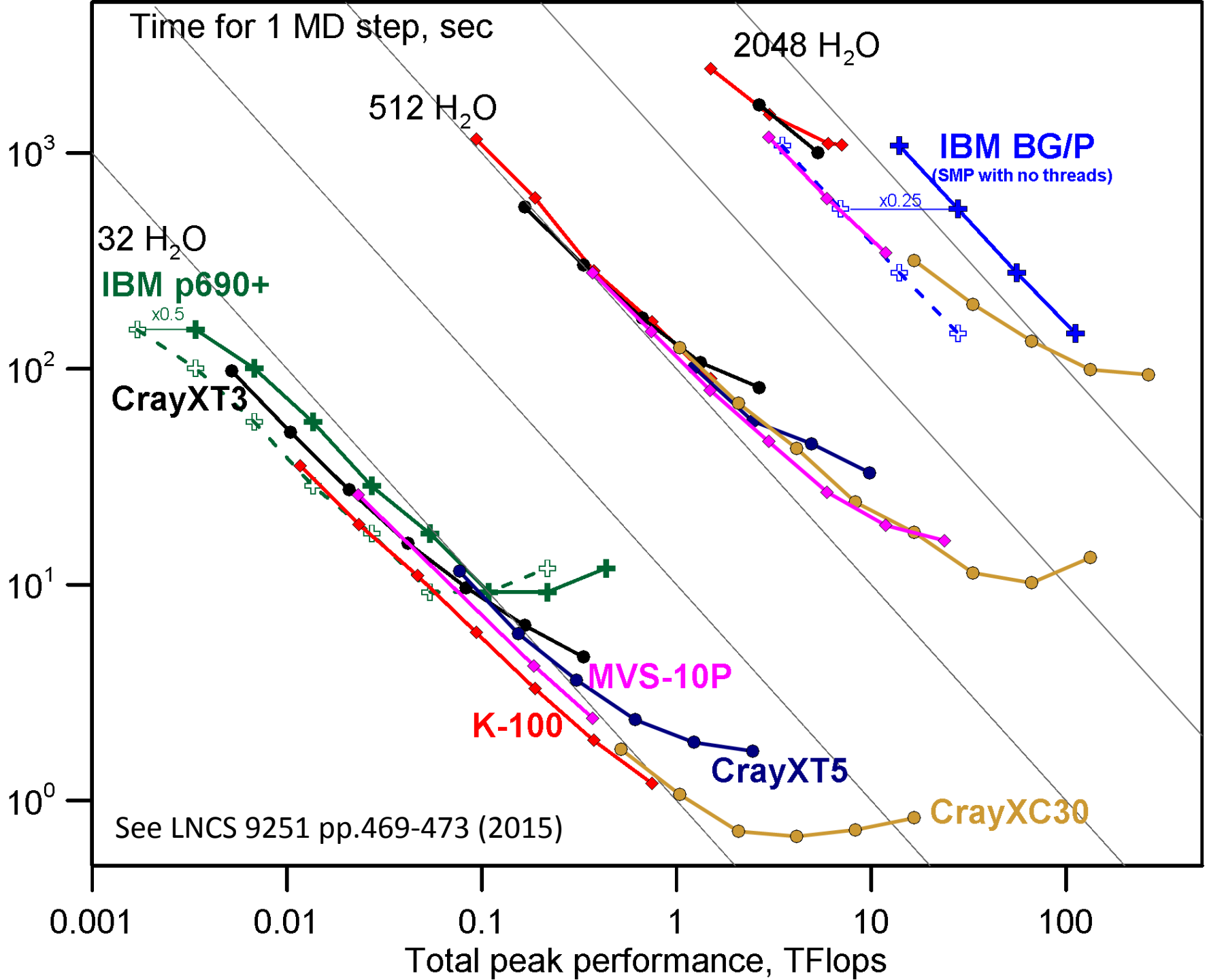
масштабируемость алгоритмов
и их адаптация к новому аппаратному обеспечению

$$\text{Time-to-solution} \sim \frac{1}{R_{peak}}$$





Vladimir V. Stegailov, Nikita D. Orekhov, and Grigory S. Smirnov,
HPC Hardware Efficiency for Quantum and Classical Molecular Dynamics
// V. Malyshkin (Ed.): PaCT 2015, LNCS 9251, pp. 469–473, 2015.


Time for 1 MD step, sec
2048 H_2O
512 H_2O
32 H_2O
IBM BG/P
(SMP with no threads)
x0.25
IBM p690+
x0.5
CrayXT3
MVS-10P
K-100
CrayXT5
CrayXC30
See LNCS 9251 pp.469-473 (2015)
10^3
10^2
10^1
10^0
0.001
0.01
0.1
1
10
100
Total peak performance, TFlops

ATTENDEES DEVELOPMENT

ISC
High Perform

ATTENDEES
GEOGRAPHICAL BREAKDOWN

Attendees from 54 Countries

Germany	875
USA	487
UK	235
China	150
France	132
Japan	136
...	

ISC High Performance

Car to Car Crashsimulation: smart vs. S-Class

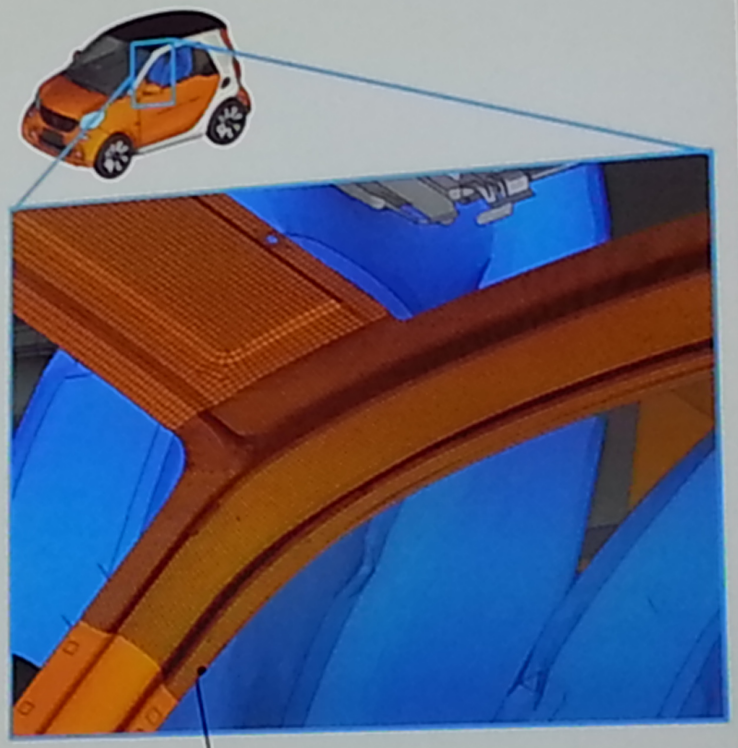
Mercedes-Benz Safety

Crashconfiguration:

- 50 km/h each, 50%-overlap for smart
- Weight S-class m = 2308 kg, Smart m = 1124 kg → **mass ratio 2:1**

Simulation:

- 192 CPUs, Solver Precision: single
- Elapsed Time 107994 seconds (29h 59min 54sec) for 222223 cycles, Total CPU Time 107323 seconds (29h 48min 43sec)
- nodes: 7,746,019, elements: 8,819,605
- Simulation time: 100ms
- Inital Timestep: 5E-4ms
- Size of shell elements for body in white: 3-10mm



Summary

- **Challenges on several front**

  But solutions are in sight

- **Future HPC node design**
  - Simple, efficient, cores
  - DRAM & PCM—3D, on package, hierarchy
  - Hierarchical, heterogeneous interconnect
- **The biggest challenge:**

  Improve System efficiency