Федеральное государственное бюджетное учреждение науки
Институт вычислительной математики и математической геофизики
Сибирского отделения Российской академии наук

ИССЛЕДОВАНИЕ ПРОИЗВОДИТЕЛЬНОСТИ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ

Снытников Алексей Владимирович

Диссертация на соискание ученой степени доктора технических наук,
специальность 05.13.15, “Вычислительные машины, комплексы и компьютерные сети”

Научный консультант
доктор физико-математических наук,
профессор
Лаврентьев Михаил Михайлович

АКТУАЛЬНОСТЬ ТЕМЫ ИССЛЕДОВАНИЯ

- Рейтинги ВС в списках Top50 и Top500 выстроены в порядке убывания пиковой производительности и реальной производительности на тесте HPL
- Это дает определенную информацию о сравнительной скорости работы представленных там ВС.
- Но очень многие факторы остаются за пределами рассмотрения.
  - скорость работы и объем дисков,
  - Пропускная способность шины памяти и коммуникационной сети,
  - неоднородность оборудования и т.д.
- Это именно те проблемы, с которыми придется столкнуться при попытке посчитать на ВС большую задачу.
- По этой причине тестирование проводится с помощью измерения времени, затрачиваемого на различные этапы программы, решающей реальную физическую задачу.

АКТУАЛЬНОСТЬ ТЕМЫ ИССЛЕДОВАНИЯ

- Существует потребность в создании **комплексного** теста производительности вычислительных систем, охватывающего все аспекты, влияющие на производительность ВС:
  - процессорные ядра *(использование ускорителей вычислений и универсальных процессоров)*,
  - графические ускорители *(использование ускорителей вычислений)*,
  - контроллеры памяти, межпроцессорные коммуникационные интерфейсы, контроллеры коммуникационной сети, сетевые коммутаторы *(вопросы масштабируемости)*,
  - средства доступа к энергонезависимой подсистеме хранения данных (дисковой подсистеме) – *проблемы адаптации к архитектуре*
- таким образом выяснить реальное быстродействие системы на широком спектре задач математического моделирования.

АКТУАЛЬНОСТЬ ТЕМЫ ИССЛЕДОВАНИЯ

Создание специализированных тестов производительности

- С.А.Степаненко, 2016 (РФЯЦ-ВНИИЭФ): методика оценки эффективности и масштабируемости вычислительных систем до экзафлопса.
- J.Dongarra, 2013, (University of Tennesee): HPCG-комплексный тест производительности на основе метода сопряженных градиентов
- M.Heroux, 2009 (Sandia National Lab): Mantevo Project – набор многофункциональных мини-приложений для тестирования производительности
- Weeratunga, D., 1994, NAS Parallel Benchmark - набор тестов производительности на основе шаблонов вычислений, применямых в вычислительной гидродинамике.

АКТУАЛЬНОСТЬ ТЕМЫ ИССЛЕДОВАНИЯ

Сравнение с другими тестами

Тесты LinPack, HPL (High Performance LinPack), ***HPCG***, NAS Parallel Benchmarks, и другие тесты (кроме HPCG) построены на базе одного численного метода:

- Не охватывают вопросы масштабируемости расчетов на конкретной ВС
- Результат тестирования серьезно зависит от правильного запуска тестов

ЦЕЛЬ РАБОТЫ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Целью работы является:

- разработка методики и создание реализующей ее программы для комплексной оценки производительности ВС с учетом функционирования всех подсистем и выявления скорости решения реальных задач;
- создание альтернативы традиционной методике оценки производительности ВВС, основанной только лишь на количестве операций в секунду.

ЦЕЛЬ РАБОТЫ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Задачи исследования:

1. Создать программный комплекс, основанный на одном из наиболее часто применяемых в высокопроизводительных вычислениях численных методов, позволяющем определять производительность всех подсистем ВС и, в частности, выделить конкретную подсистему, наиболее заметно снижающую скорость счета реальных приложений.

2. Реализовать метод анализа производительности коммуникационной сети ВС для выработки рекомендаций по более эффективному распределению процессов приложения по узлам ВС.

3. Разработать и обосновать методику расчета абсолютной, не основанной на конкретном вычислительном методе или задаче, оценки пригодности ВС для решения реальных задач.

4. Разработать метод комплексного анализа производительности узлов мультиархитектурной ВС, позволяющий делать прогнозы эффективности данной мультиархитектурной ВС для решения конкретных задач, более достоверные по сравнению с синтетическими тестами.

СТРУКТУРА ДИССЕРТАЦИИ

Введение

Глава 1. Описание реализации метода частиц в ячейках для высокопроизводительных ВС

Глава 2. Физико-математические задачи и вычислительные методы в исследованиях, проводимых с использованием высокопроизводительных ВС

Глава 3. Комплексная оценка производительности ВС

Глава 4. Анализ масштабируемости, параллельной эффективности и ускорения параллельной ВС

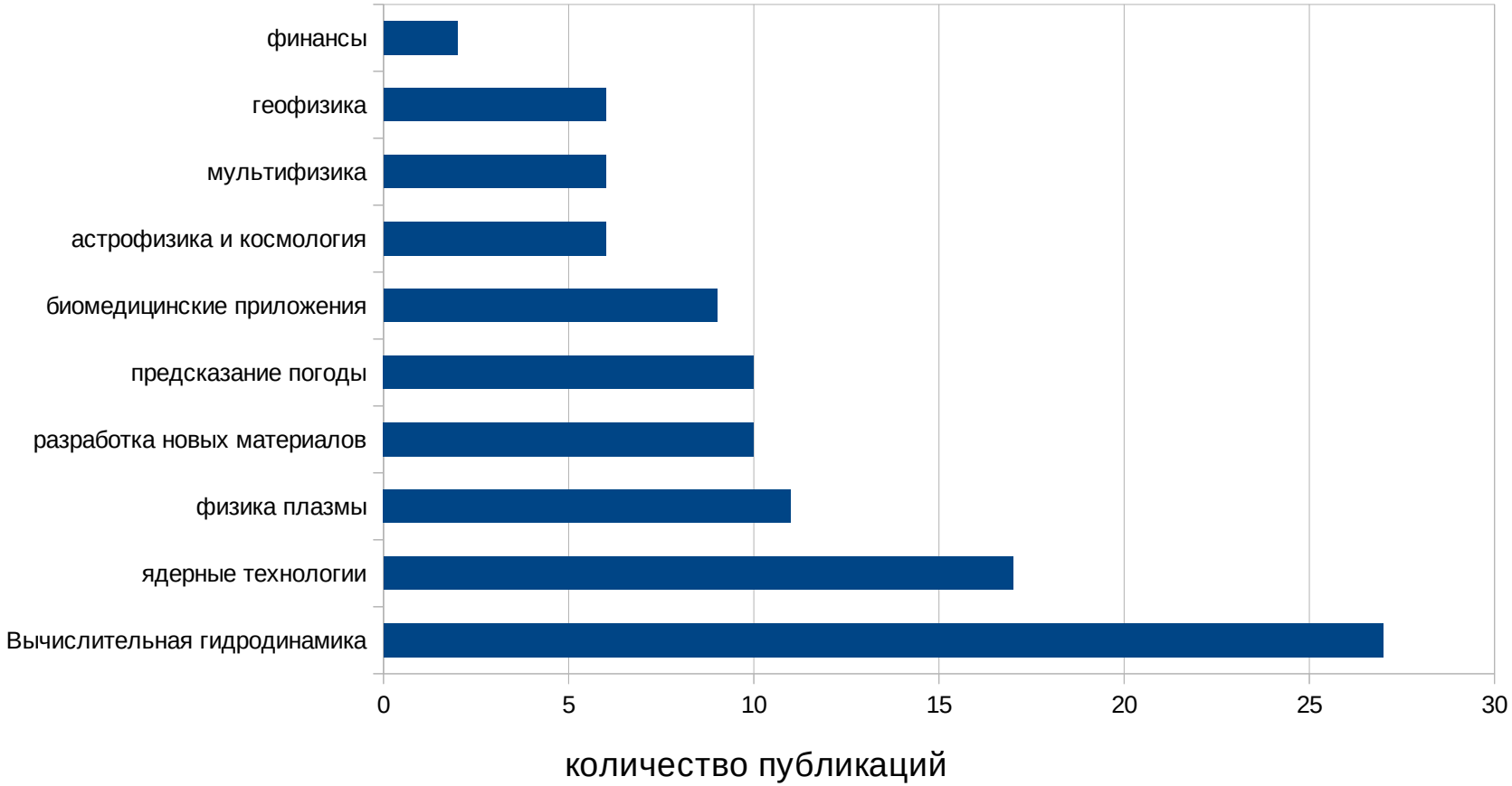
Глава 5. Анализ производительности узлов мультиархитектурной ВС.

Заключение
Список сокращений и условных обозначений
Список литературы
Список рисунков
Список таблиц

АКТУАЛЬНОСТЬ ТЕМЫ ИССЛЕДОВАНИЯ

Обзор публикаций о перспективных ВВС

Решаемые задачи



2010-2016 годы: Future Generation Computer Systems,
Procedia Computers Science,
Journal of Parallel and Distributed Computing, Parallel Computing,
Journal of Computational Physics, Computer Physics Communications и др.

Глава 1. Описание реализации метода частиц в ячейках для высокопроизводительных ВС

Моделирование плазмы методом частиц в ячейках (осн.рез. 1)

Основные уравнения

$$\frac{\partial f_{i,e}}{\partial t}+\vec{v}\frac{\partial f_{i,e}}{\partial \vec{x}}+\vec{F}\frac{\partial f_{i,e}}{\partial \vec{v}}=0$$

$$\nabla\times\vec{B}=4\pi\vec{j}+\frac{1}{c}\frac{\partial \vec{E}}{\partial t}$$

$$\nabla\times\vec{E}=-\frac{1}{c}\frac{\partial \vec{B}}{\partial t}$$

$$\nabla\vec{E}=4\pi\rho$$

$$\nabla\vec{B}=0$$

$$\vec{p}=\gamma\vec{v},\ \gamma^{-1}=\sqrt{1-v^2}$$

$$\vec{F}=q_{i,e}\left(\vec{E}+\frac{1}{c}[\vec{v},\vec{B}]\right)$$

$$\vec{j}=\sum_{i,e}q_{i,e}\int f_{i,e}\vec{v}\,d\vec{v}$$

$$\rho=\sum_{i,e}q_{i,e}\int f_{i,e}\,d\vec{v}$$

Граничные условия: периодические

Начальные условия

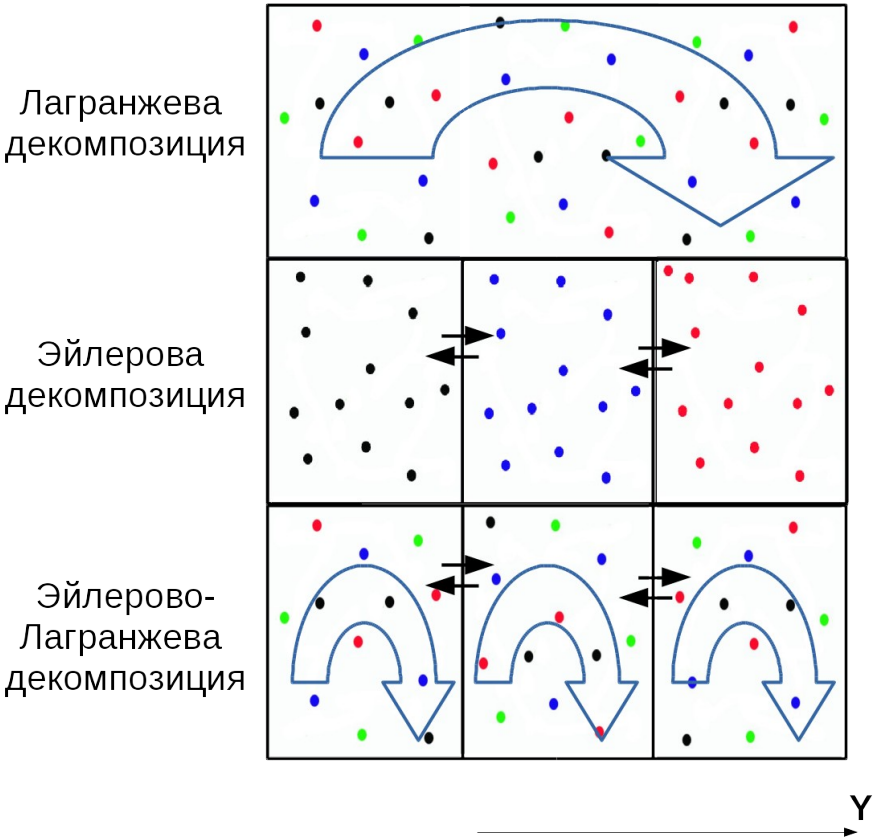
- $\rho_e = 1000,\ \rho_b = 1$

  $\rho = \rho_e+\rho_b$
- Импульсы электронов плазмы: $\boldsymbol{p_x, p_y, p_z}$ — максвелловское распределение, $\sigma = T_e = 1.0$

$$f=\exp\left(\frac{-p^2}{\sigma}\right)$$

- Импульсы ионов плазмы:0
- Импульс электронов пучка: $\boldsymbol{p_x = 50\ p_y = p_z = 0}$

Реализация метода частиц в ячейках для высокопроизводительных ВС



Операция MPI_Allreduce

Операция MPI_Sendrecv

Оригинальный вариант декомпозиции, созданный в данной работе

- В силу того, что метод частиц в ячейках создает “плохой” (очень нерегулярный) шаблон доступа к памяти, то полученная по данному методу производительность фактически является **оценкой снизу** для других вычислительных методов
- В рамках диссертации созданы несколько вариантов численного решения задачи моделирования динамики плазмы (созданы пакеты программ):
  - Реализовано три вида пространственной декомпозиции
  - Разработаны реализации для GPU и MIC (Intel Xeon Phi)
  - Выполнены реализации на Fortran и C/C++

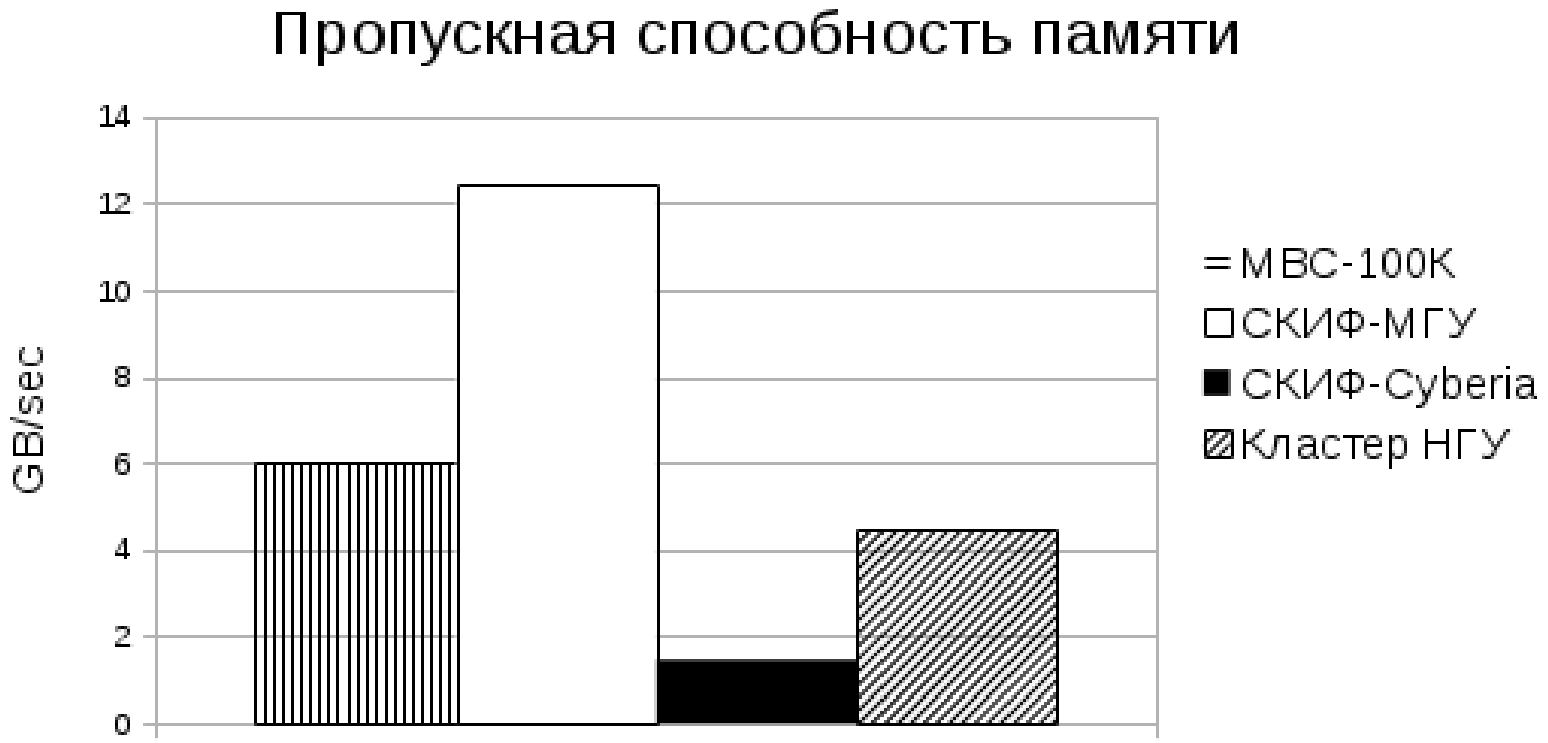
Глава 3. Комплексная оценка производительности ВС

- Общий обзор показателей производительности
- Измерение производительности системы памяти
- Измерение производительности процессорных ядер и ускорителей вычислений
- Расчет производительности процессорных элементов на основе движения модельных частиц
- Сравнение с известными тестами.

Измерение производительности системы памяти

$$W_{PIC,GB/sec} = \frac{W_P \times N_P P_{core}}{\Delta t}$$

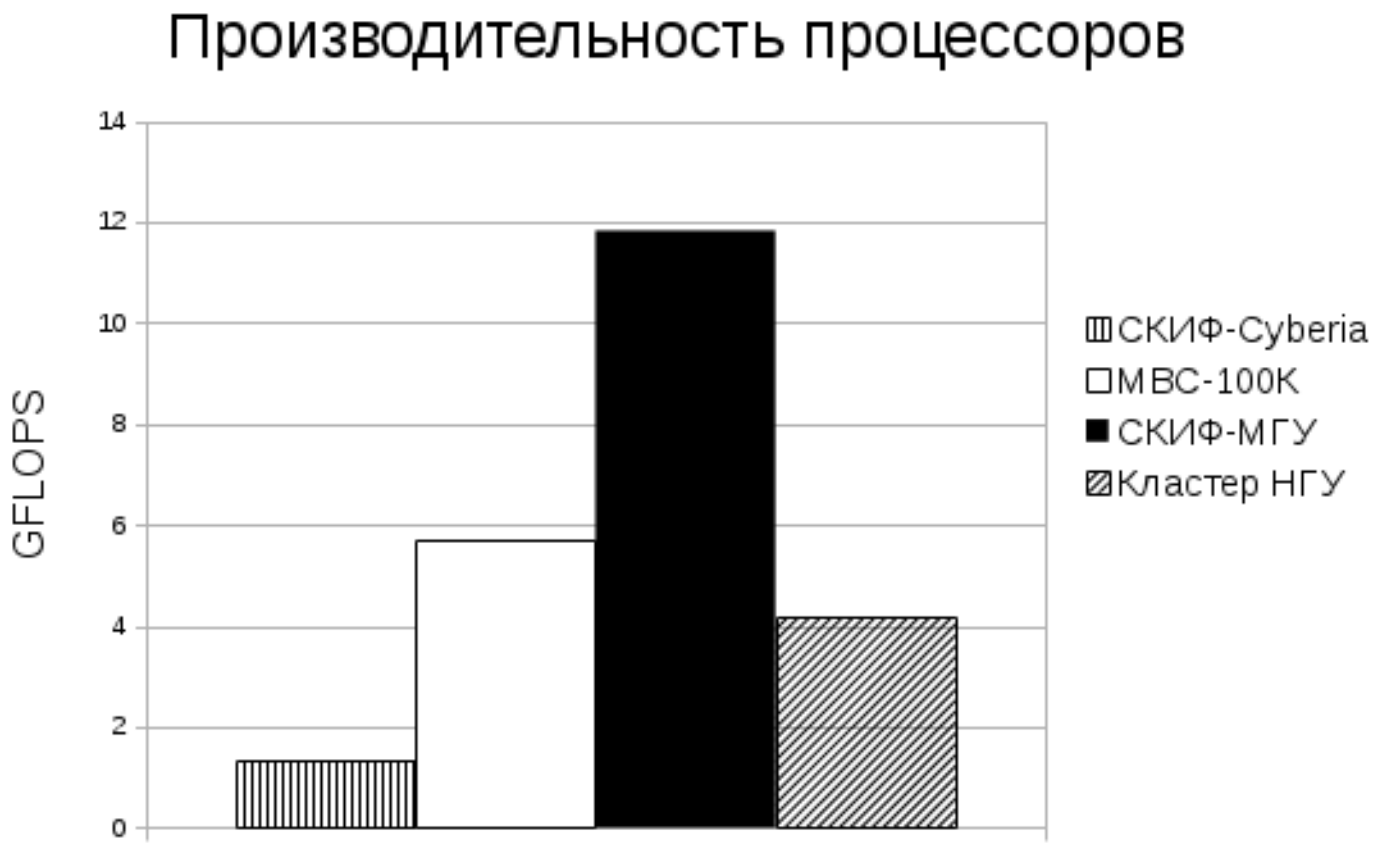
- W_P - количество байт на одну модельную частицу, W_P = 576;
- N_P - количество модельных частиц на одно процессорное ядро (2.5×10^6);
- P_{core} - количество процессорных ядер;
- Δt - длительность временного шага, сек.



Измерение производительности процессорных ядер и ускорителей вычислений

$$N_{PIC,FLOPS} = \frac{F_P \times N_P \times P_{core}}{\Delta t}$$

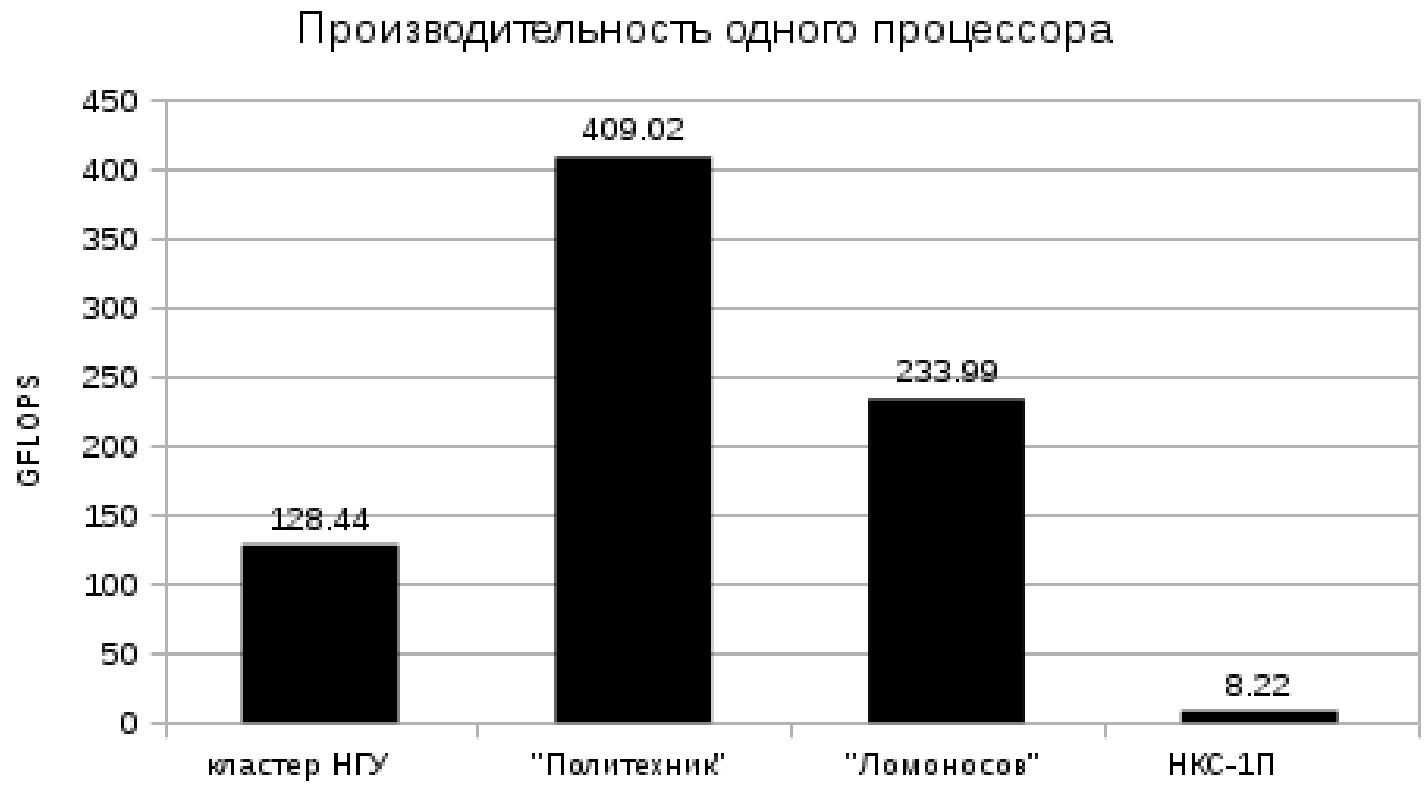
- F_P - количество операций на одну модельную частицу, F_P = 500;
- N_P - количество модельных частиц на одно процессорное ядро (2.5× 10^6);
- P_{core} - количество процессорных ядер;
- Δt - длительность временного шага, сек.



Расчет производительности процессорных элементов на основе движения модельных частиц

$$N_{PIC,FLOPS} = \frac{F_P \times N_P \times P_{core}}{\Delta t}$$

- F_P - количество операций на одну модельную частицу, F_P = 500;
- N_P - количество модельных частиц на одно процессорное ядро;
- P_{core} - количество процессорных ядер;
- Δt - длительность временного шага, сек.



Сравнение с известными тестами. Тест HPCG (“Политехник”, СПбПУ)

- Для операции DDOT (скалярное произведение в двойной точности): 1237.46 GFLOPS
- Для операции WAXPBY (сложение векторов с множителем): 10.766 GFLOPS
- Для операции SpMV (умножение разреженной матрицы на вектор): 13.0012 GFLOPS
- Для операции MG (трехуровневый многосеточный метод): 13.8033 GFLOPS.

Для созданного в данной работе теста для расчета движения модельных частиц (близко к операции DDOT): 409.02 GFLOPS

Глава 4. Анализ масштабируемости, параллельной эффективности и ускорения параллельной ВС

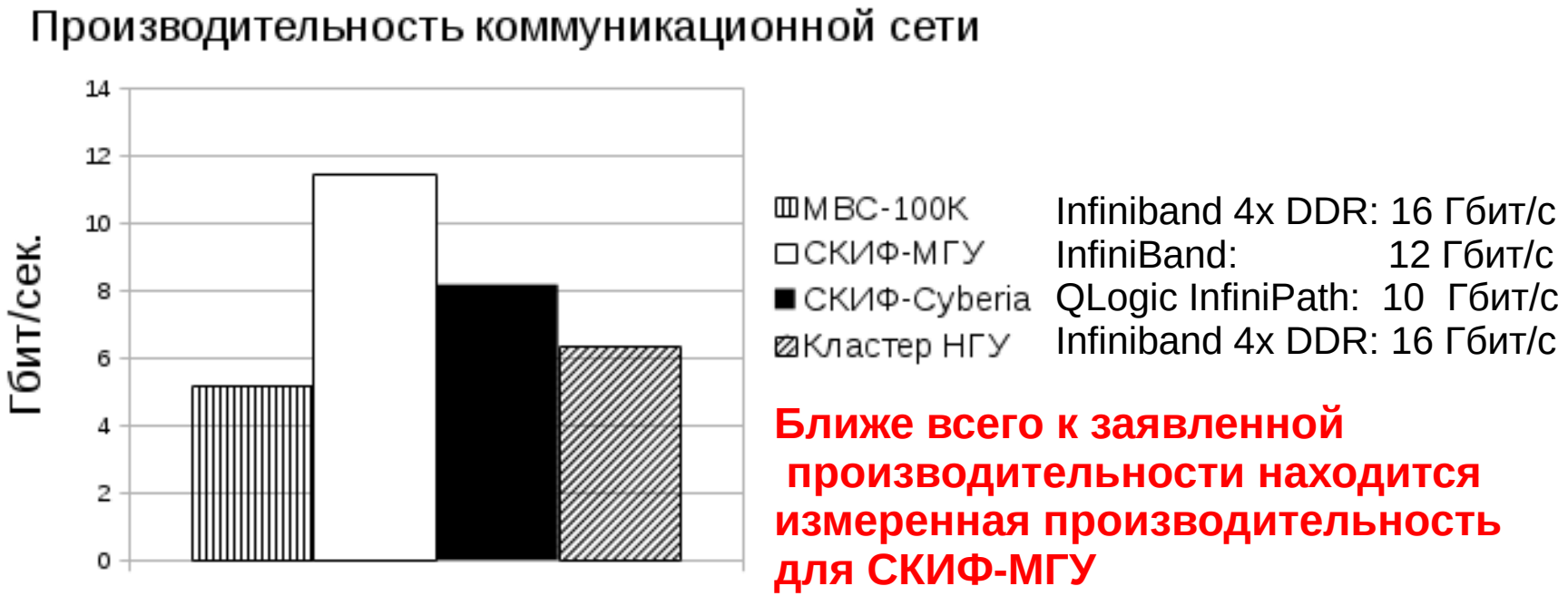
Определение

- Производительность коммуникационной сети – **фактический** объем пересылаемых данных в единицу времени

Производительность коммуникационной сети. Peer-to-peer

$$W_S = \frac{U_P \times \nu N_P \times P_{core}}{T_{S,PIC}}$$

- U_P - количество байт на одну модельную частицу (48)
- N_P - количество модельных частиц на одно процессорное ядро (2.5×10^6);
- P_{core} - количество процессорных ядер;
- ν - доля пересылаемых частиц ($\nu = 0.05$);
- $T_{S,PIC}$ - время пересылки частиц, сек.



Ближе всего к заявленной производительности находится измеренная производительность для СКИФ-МГУ

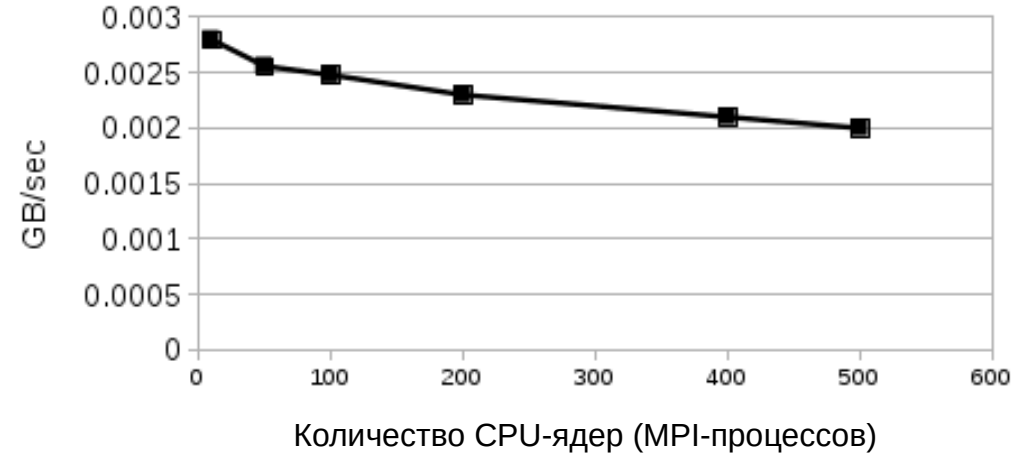
Производительность коммуникационной сети. Collective.

$$W_A = \frac{N_X \times N_Y / P_{SUB} \times N_Z \times 24}{T_A}$$

N_X, N_Y, N_Z - количество узлов сетки по X,Y и Z соответственно;
P_{SUB} - количество подобластей (если используется эйлерова декомпозиция)
T_A - длительность операции MPI_Allreduce (суммирование токов по всей области), сек.

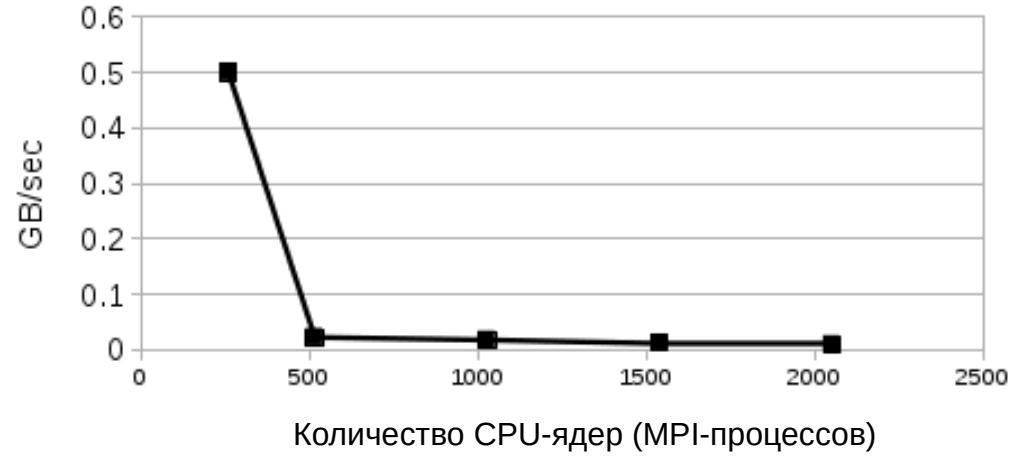
Сеть: **QDR Infiniband 4x: 40 Гбит/с**
MPI: **Intel MPI 4**

“Ломоносов”



Сеть: **Infiniband 4x DDR: 16 Гбит/с**
MPI: **MVAPICH-1.2rc1**

МВС-100К



Формула для комплексной оценки ВС

$$W_S = \frac{U_P \times \nu N_P \times P_{core}}{T_{S,PIC}}$$

$$W_A = \frac{N_X \times N_Y / P_{SUB} \times N_Z \times 24}{T_A}$$

$$\Big\} \approx W_{PIC,GB/sec} = \frac{W_P \times N_P P_{core}}{\Delta t}$$

Для того, чтобы параллельная ВС могла быть признана адаптированной к задачам математического моделирования, она должна соответствовать следующим требованиям:

- Очень высокая производительность коммуникационной сети (W_S, и W_A), позволяющая пересылать все необходимые для расчета данные, не задерживая вычислений;
- Относительно высокая производительность оперативной памяти ($W_{PIC,GB/sec}$), позволяющая эффективно использовать ресурсы процессоров, т.е. **фактически совпадающая** с производительностью процессора ($N_{PIC,\ FLOPS}$).

Методика расчета абсолютной оценки(осн.рез.3)

- **“Абсолютная”** – в данном случае означает отсутствие привязки к каким-либо алгоритмам или задачам
- Последовательность действий:

  1)Измерение производительности системы памяти

  2)Измерение производительности коммуникационной системы

  3)Измерение проиводительности процессорных элементов

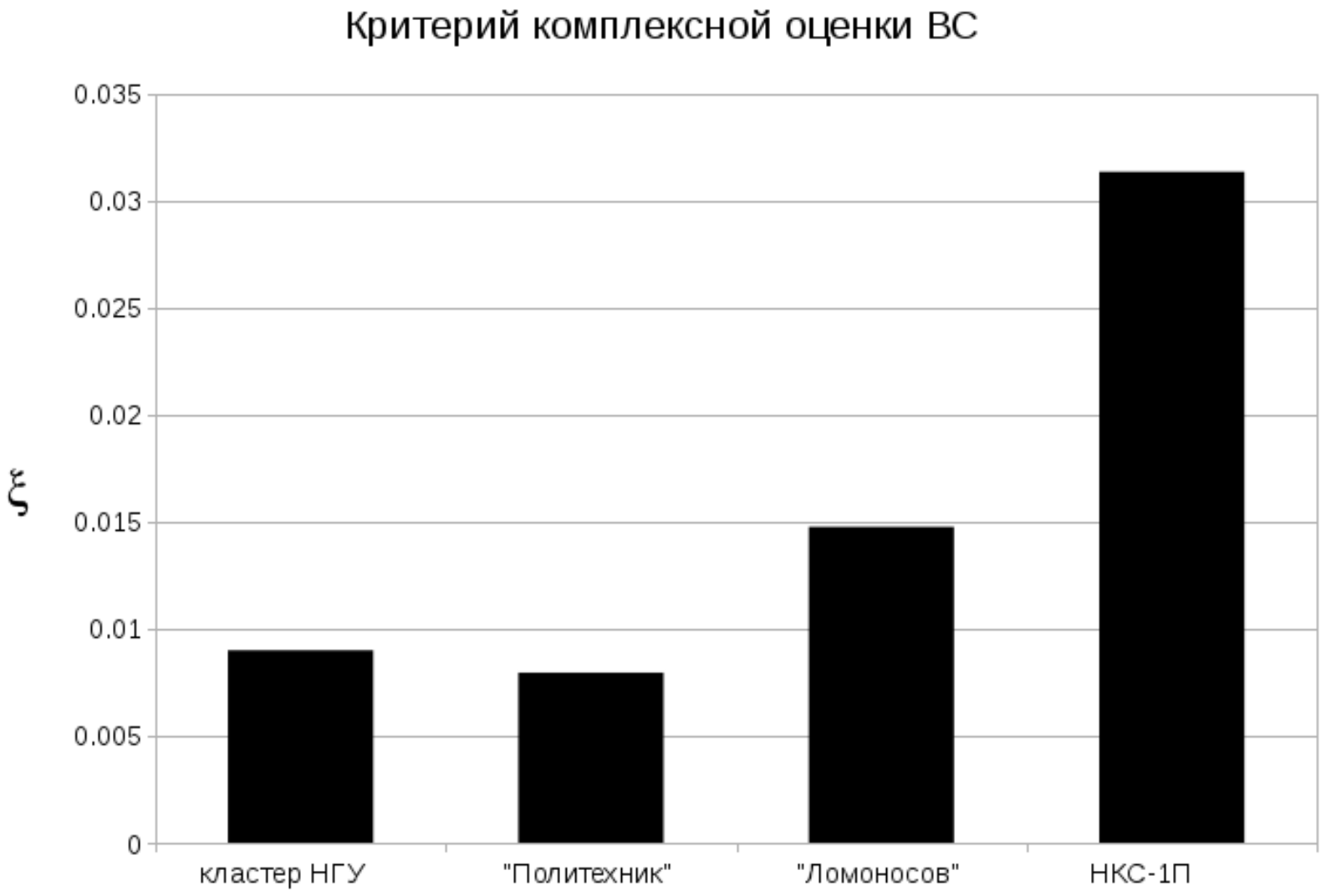
  4)Сравнение производительности памяти и коммуникационной системы

  5)Оценка будет представлять собой отношение производительности коммуникационной системы к производительности системы памяти

методика расчета абсолютной оценки(осн.рез.3)

Формула для комплексной оценки ВС

$$\xi = \frac{W_{MPI}}{W_{PIC,GB/sec}},$$



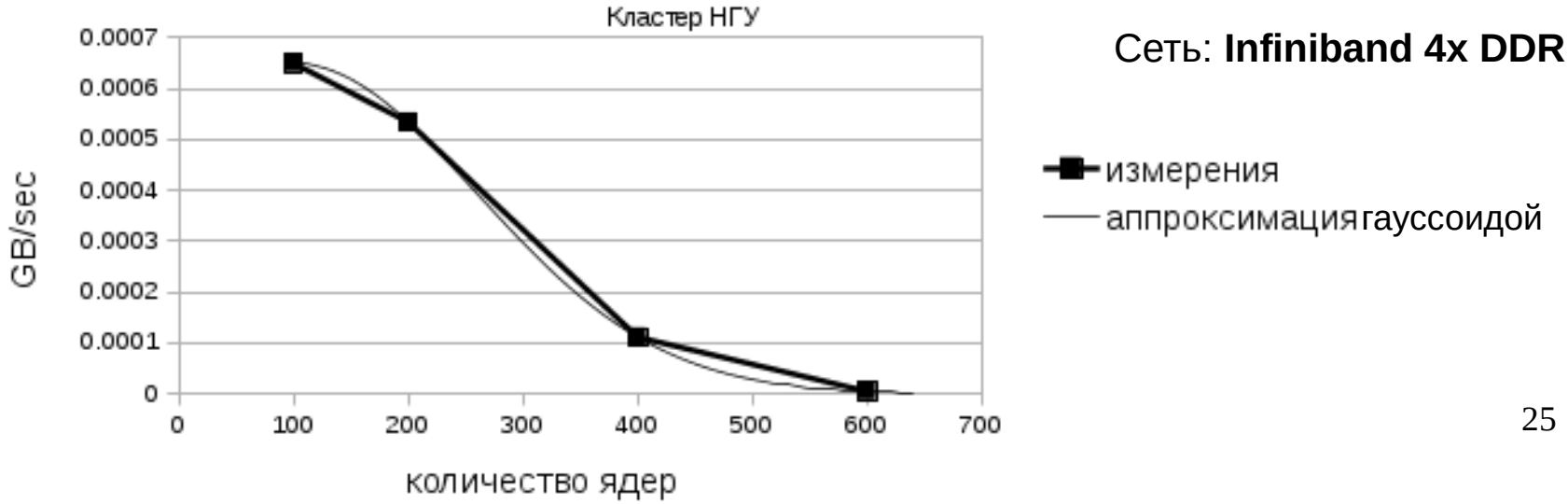
Анализ возможности функционирования ВС как целого для решения одной большой задачи

$$W_A = \frac{N_X \times N_Y / P_{SUB} \times N_Z \times 24}{T_A}$$

Эффективный коммуникационный размер ВС - максимальное количество процесоров, которое может быть в рамках данной ВС эффективно использовано для решения одной задачи.

Количество процессоров, для которого производительность коммуникационной сети падает не более чем в ***e*** раз по сравнению с максимальной: **226**

Производительность коммуникационной сети



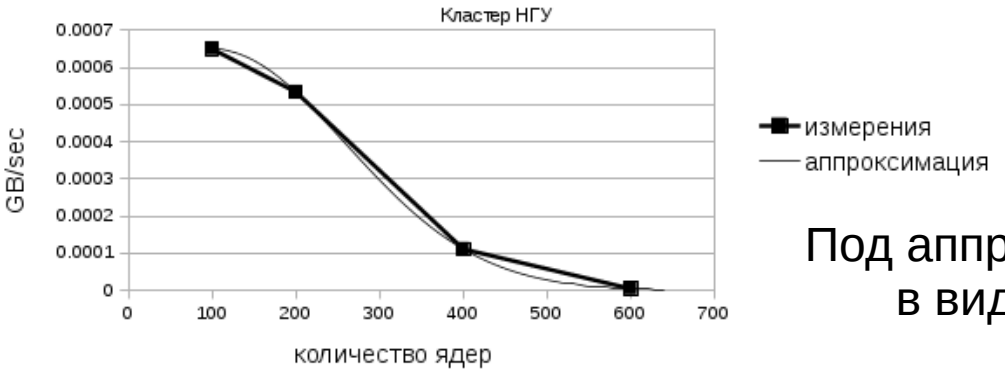
Сеть: **Infiniband 4x DDR**

Эффективный коммуникационный размер

Производительность коммуникационной сети

Кластер НГУ

Сеть: **Infiniband 4x DDR**
MPI: **Intel MPI 5**

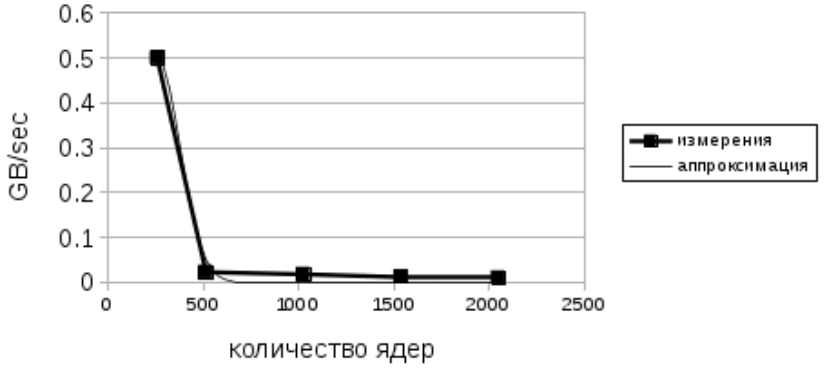


226

Под аппроксимацией имеется в виду аппроксимация гауссоидой

МВС-100К

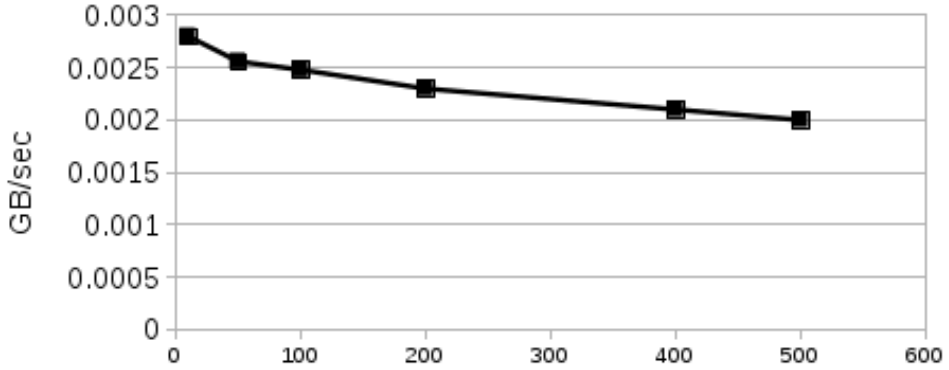
Сеть: **Infiniband 4x DDR: 16 Гбит/с**
MPI: **MVAPICH-1.2rc1**



425

“Ломоносов”

Сеть: **QDR Infiniband 4x: 40 Гбит/с**
MPI: **Intel MPI 4**



239

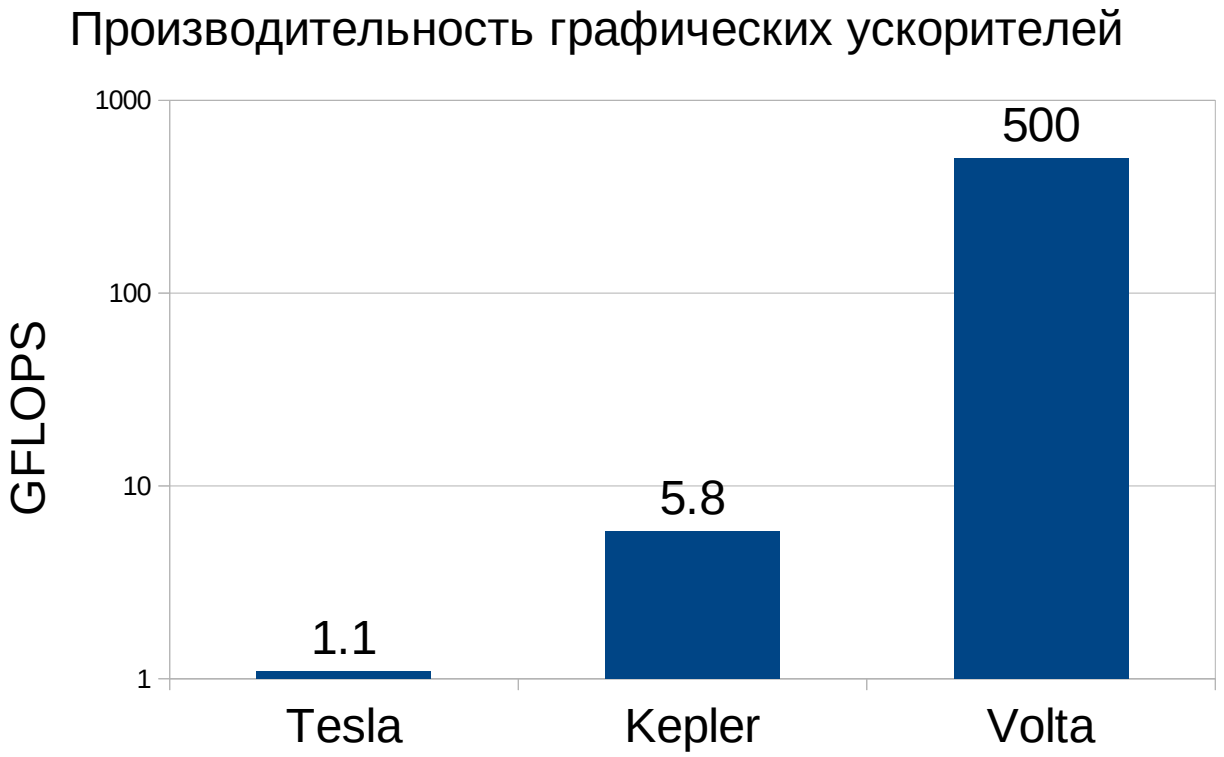
Количество CPU-ядер (MPI-процессов)

Глава 5. Анализ производительности узлов мультиархитектурной ВС

измерение производительности процессорных ядер и ускорителей вычислений

$$N_{PIC,FLOPS} = \frac{F_P \times N_P \times P_{core}}{\Delta t}$$

- F_P - количество операций на одну модельную частицу, F_P = 500;
- N_P - количество модельных частиц на одно процессорное ядро (2.5×10^6);
- P_{core} - количество процессорных ядер;
- Δt - длительность временного шага, сек.



Прогноз эффективности мультиархитектурной ВС (осн.рез. 4)

- Вопрос, который наиболее часто задают применительно к мультиархитектурной ВС, оснащенной графическими ускорителями - это возможность **эффективной** реализации конкретного вычислительного алгоритма на даннной мультиархитектурной ВС.
- Для ответа на данный вопрос предлагается интерполяционная формула:

$$v = v_{PIC}\, k + (1-k)\, v_{B,E}$$

- v_{PIC} – производительность на стадии расчета движения частиц
- $v_{B,E}$ – производительность на стадии расчета электрического и магнитного полей

Прогноз эффективности мультиархитектурной ВС (осн.рез. 4)

большинство численных методов используемых в математическом моделировании находятся в промежуточном положении по отношению к используемым в методе частиц в ячейках алгоритму вычисления поля и алгоритму расчета движения частиц по следующим показателям:

1) вычислительной интенсивности (равномерное распределение вычислительно сложных фрагментов по тексту или отдельные высоконагруженные участки);
2) характеру доступа к оперативной памяти (регулярный или нерегулярный);
3) объему используемых данных (большой или маленький).

$$v_{pre} = v_{PIC}k + (1-k)v_{B,E}$$

Вычислительный алгоритм	Интенсивность	Доступ к оперативной памяти	Объем данных	k
Расчет движения модельных частиц	низкая	нерегулярный	большой	1.0
Метод Монте-Карло	низкая	нерегулярный	средний	0.9
Метод SPH	низкая	нерегулярный	небольшой	0.6
Метод конечных элементов	высокая	нерегулярный	большой	0.5
Конечно-разностные схемы (явные)	высокая	регулярный	большой	0.2
Конечно-разностные схемы (явные)-2	высокая	регулярный	большой	0.1
Вычисление электромагнитного поля	высокая	регулярный	большой	0.0

ЗАКЛЮЧЕНИЕ

В диссертации предложено решение важной научной проблемы – предложена и реализована оригинальная методика комплексного тестирования мультиархитектурных параллельных вычислительных систем.

1. Создан программный комплекс, основанный на **моделировании динамики плазмы методом частиц в ячейках** на высокопроизводительных ВС для **всестороннего исследования производительности ВС**, позволяющий в рамках одного запуска программы определить конкретную подсистему, наиболее заметно снижающую скорость счета реальных приложений.

2. Реализован и протестирован метод исследования коммуникационной структуры высокопроизводительных ВС, позволяющий давать рекомендации по более оптимальному распределению процессов в приложении на узлах высокопроизводительной ВС, а также вычислять **экстраполяцию** реально полученной **производительности на** аналогичные **системы с большим количеством узлов** и процессоров.

3. Разработана и обоснована **методика расчета абсолютной оценки** пригодности данной ВС для решения реальных задач, основанной на сбалансированности производительности различных подсистем конкретной ВС, в частности оперативной памяти, коммуникационной сети, дисковой подсистемы, процессоров и ускорителей вычислений и позволяющей сравнивать ВС безотносительно используемых программ и решаемых задач.

4. Предложен и протестирован метод комплексного анализа производительности узлов мультиархитектурной ВС, оснащенной многоядерными процессорами и графическими ускорителями или ускорителями Intel Xeon Phi, основанный на программе для моделирования динамики плазмы методом частиц в ячейках и позволяющий делать **прогнозы эффективности** данной **мультиархитектурной ВС** для решения конкретных задач, более достоверные по сравнению с синтетическими тестами.

Основные публикации

всего статей: **30**, тезисов конференций: 19, книги: 1

1. Snytnikov A.V. Large-Scale and Fine-Grain Parallelism in Plasma Simulation. – Chapter 3. Parallel Programming: Practical Aspects, Models and Current Limitations. – pp. 59-70. – 2014. ISBN 978-1633219571.

Статьи в журналах, рекомендованных ВАК РФ

2. Снытников А.В. Сравнительный анализ производительности кластерных суперЭВМ на примере задачи о релаксации электонного пучка в высокотемпературной плазме // Вестник Нижегородского Университета им. Н.И. Лобачевского. — 2011. — Т. 1, No 3. — С. 285-292.
3. Снытников А. В. Об одном методе распараллеливания решения уравнения Пуассона // Автометрия. — 2006. — Т. 43, No 3. — С. 62-68.
4. Вшивков В. А., Лазарева Г. Г., Снытников А. В. Эффективный параллельный алгоритм численного моделирования в моносилановой плазме тлеющего разряда // Автометрия. — 2008. — Т. 44, No 5. — С. 112-122.
5. Месяц Е. А., Снытников А. В., Лотов К. В. О выборе числа частиц в методе частиц-в-ячейках для моделирования задач физики плазмы //Вычислительные технологии. — 2013. — Т. 18, No 6. — С. 83-96.
6. Вшивков В. А., Лазарева Г. Г., Снытников А. В. Адаптивное изменение массы модельных частиц при моделировании тлеющего ВЧ-разряда в силановой плазме // Вычислительные технологии. — 2008. — Т. 13, No 1. — С. 22-30.
7. Романенко А. А., Снытников А. В., Тимофеев И. В. Трехмерный гибридный код для моделирования генерации высокочастотного электромагнитного излучения турбулентной плазмы // Вестник НГУ, Серия: Информационные технологии. — 2016. — Т. 14, No 3. —С. 81-90.
8. Многоуровневый подход к разработке алгоритмического и программного обеспечения экзафлопсных суперЭВМ / Б. М. Глинский [и др.] // Выч. мет. программирование. — 2015. — Т. 16, № 4. — С. 543-556.
9. Лотов К. В., Месяц Е. А., Снытников А. В. Моделирование кинетической неустойчивости электронного пучка в плазме // Математическое моделирование. — 2014. — Т. 26, No 11. — С. 45-50.
10. Романенко А. А., Снытников А. В., Черных И. Г. Адаптация параллельного вычислительного алгоритма к архитектуре суперЭВМ на примере моделирования динамики плазмы методом частиц в ячейках // Вестник НГУ Серия: Информационные технологии. — 2017. —Т. 15, No 4. — С. 74-86.
11. Вшивков В. А., Снытников А. В. Особенности проведения экзафлопс-расчетов в физике плазмы // Выч. мет. программирование. — 2012. —Т. 13, No 1. — С. 44-48.

Публикации в изданиях, индексируемых Scopus и Web of Science

9. B.M. Glinskiy, I.M.Kulikov,I.G.Chernykh, A.V.Snytnikov, A.A.Romanenko, V.A.Vshivkov. Co-design of Parallel Numerical Methods for Plasma Physics and Astrophysics / B. M. Glinskiy [et al.] // Supercomputing Frontiers and Innovations. — 2014. — Vol. 1, no. 3. — P. 88-98.
10. Goedheer W.J., Snytnikov A.V., Vshivkov V.A. Adaptive mass alteration to model ion-ion recombination in a Particle-in-Cell simulation of silane radio-frequency discharges. // Computer Physics Communications. — 2010. —V. 181, No 10. — P. 1743-1749.
11. Lotov, K.V., Timofeev, I.V., Mesyats, E.A., Snytnikov, A.V., Vshivkov, V.A. Note on quantitatively correct simulations of the kinetic beam-plasma instability // Physics of Plasmas — 2015. —V. 181, No 2.
12. Glinskiy B.M., Kulikov I.M., Chernykh I.G., Snytnikov A.V., Nenashev V., Andreev A., Egunov V., Kharkov E. The Co-design of Astrophysical Code for Massively Parallel Supercomputers // Lecture Notes in Computer Science, 2016. – Vol. 10049. – P. 342-353.
13. Glinskiy B.M., Kulikov I.M., Chernykh I.G., Snytnikov A.V., Sapetina A.F., Weins D.V. The Integrated Approach to Solving Large-Size Physical Problems on Supercomputers // Supercomputing. RuSCDays 2017. Communications in Computer and Information Science (CCIS), Springer. 2017. – P. 278-289.

Сравнение с известными тестами. Тест HPL

- кластер СПбПУ “Политехник” 4.3 (499)
- кластер ССКЦ СО РАН “НКС-1П” 3.14 (8.22)

Как видно, производительность получается значительно меньше, чем при расчете движения модельных частиц в тесте PIC-MANAS. Объяснение этому заключается в том, что подбор правильной конфигурации для теста Linpack, адекватно показывающий производительность рассматриваемой ВС представляет собой отдельную сложную задачу. Также важно отметить, что результаты теста PIC-MANAS в значительно меньшей степени зависят от конфигурации запуска в силу того, что в ходе выполнения вычилений по методу частиц в ячейках происходит многократное усреднение всех показателей.

ПУБЛИКАЦИИ И АПРОБАЦИЯ РЕЗУЛЬТАТОВ

- **По теме диссертации автором опубликовано более 15 работ:**

- **в журналах из перечня ВАК РФ**: **10**
- **в изданиях, индексируемых Scopus и Web of Science**: **5**
- **свидетельства о регистрации программ для ЭВМ**: **2**
- **Акт о внедрении: 2 (ССКЦ СО РАН, ИЯФ СО РАН)**

Апробация результатов. Основные результаты диссертационной работы докладывались и обсуждались на Международных научных конференциях в России и за рубежом: на международной научных конференциях серии Parallel Computing Technologies (Нижний Новгород, 2003, Красноярск, 2005, Новосибирск, 2009), международной конференции International Conference on Computational Science (Амстердам, 2009), международной конференции Open Magnetic Systems for Plasma Confinement (Новосибирск, 2010), международной конференция <<Параллельные вычисления и задачи управления>> (Москва, 2010), международной суперкомпьютерной конференции «Научный сервис в сети Интернет» (Новороссийск, 2009, 2011, 2014), международной научной конференции Russian Supercomputing Days (Москва, 2015, 2016), международной конференции «Супервычисления и математическое моделирование» (Саров, 2016), обсуждались на семинарах в Институте Вычислительной Математики и Математической Геофизики СО РАН, Институте Ядерной Физики СО РАН, Институте Вычислительных Технологий СО РАН, Институте Теоретической и Прикладной Механики СО РАН, Институте Прикладной Математики РАН, Научно-Исследовательском Вычислительном Центре МГУ.

Основные этапы исследования выполнены при поддержке грантов Российского фонда фундаментальных исследований №№ 14-07-00241 (руководитель), 18-07-00364 (руководитель), 16-07-00434, 16-07-00916, а также при финансовой поддержке Министерства образования и науки РФ (уникальный идентификатор работ (проекта) RFMEFI57417X0145, соглашение № 14.574.21.0145 от 26.09.2017 г.) в рамках федеральной целевой программы “Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014—2020 годы”.

Спасибо за внимание!!!

Замечания официального оппонента М.Э.Рояка

1. Наиболее существенным недостатком работы считаю то, что одним из защищаемых положений диссертационной работы является положение 3 («Разработана и обоснована методика расчета *абсолютной* оценки пригодности данной ВС для решения реальных задач, основанной на сбалансированности производительности различных подсистем конкретной ВС, в частности оперативной памяти, коммуникационной сети, дисковой подсистемы, процессоров и ускорителей вычислений и позволяющей сравнивать ВС безотносительно используемых программ и решаемых задач»). Это положение слабо обосновано, более того, в самой диссертации можно найти подтверждение того, что разработка такого теста малореальна. Например, в п.1.4.1 из таблицы 1 видно, что даже в самом методе частиц в ячейках в зависимости от используемой декомпозиции расчетной области меняется время выполнения программы, откуда, очевидно, следует что оценка вычислительной системы, сделанная тестом диссертанта, изменится даже при относительно небольшой смене алгоритма метода частиц в ячейках, не говоря уже о переходе к решению задач каким-либо принципиально другим методом (типа метода конечных элементов, коллокаций, граничных элементов и др.). Кроме того, результаты теста диссертанта, как следует из приведенных в работе сравнений с другими тестами (например, с тестом Intel MPI Bebchmarks на стр.113), не полностью совпадают с результатами других тестов, что также говорит об отсутствии универсальности (поскольку другие тесты также пытаются эмулировать решение реальных задач, но другими способами). Не отрицая полезность проведенных автором исследований, хочу заметить, что разработка абсолютного теста все же пока является неосуществимой мечтой. Необходимо

С замечанием согласен. Использование термина “абсолютный” является в данном случае неудачным… Вместе с тем речь идет именно о сбалансированности ВС, т.е. о том, например, что подсистема памяти генерирует данные не с большей скоростью, чем подсистема памяти способна их пересылать, поэтому здесь имеет значение лишь объем данных и тип декомпозиции области, но не используемый вычислительный метод

Замечания официального оппонента М.Э.Рояка

2. На стр.14 в перечислении конференций в апробации работы иногда путаются числа и падежи
3. Обозначения в формулах иногда вводятся без пояснений, что затрудняет их понимание, например, в формулах 1.2 и 1.3 без пояснений вводятся *u* и *x* с индексами, причем в 1.2 *u* жирная, а в 1.3 – нет.
4. Приведенные листинги в п. 1.6.1 и 1.6.2 показывают, что автор, учитывая вещественные операции, не заботится в своих программах о целочисленных, например, о копировании при передаче параметра – структуры Field в листинге 1.4, а также о значительных дополнительных затратах при вызове функций. Несколько странен и неединообразный стиль программирования автора – параметры типа double иногда передаются копированием, а иногда по ссылке. Кроме того, подсчет операций выполняется автором тоже не всегда корректно, например, на стр. 36 в фрагменте

```
i=abs(i2.x-i1.x);
l=abs(i2.y-i1.y);
k=abs(i2.z-i1.z);
m =4*i +2* l+k
```

автор в комментариях указывает 3 операции, причем, судя по описаниям типов, все переменные в этом фрагменте целые.

С замечаниями 2,3 полностью согласен.
По замечанию 4: параметры передаются по ссылке в основном там, где в этом есть прямая необходимость, в отдельных случаях это неточность, вызванная автоматической переработкой исходного текста программы, первоначально написанной на языке Фортран
Учет количества целочисленных операций не проводился ввиду того, что интерес представляют исключительно операции с плавающей точкой.

Замечания официального оппонента М.Э.Рояка

```
i=abs(i2.x-i1.x);
l=abs(i2.y-i1.y);
k=abs(i2.z-i1.z);
m =4*i +2* l+k
```

автор в комментариях указывает 3 операции, причем, судя по описаниям типов, все переменные в этом фрагменте целые.

5. На стр.130 при обсуждении таблицы 24 сказано, что зависимость в ней близка к линейной, хотя в ней всего 3 точки и при первом увеличении числа потоков вдвое время изменилось более чем в 3 раза, в при следующем удвоении – менее чем в два раза, а других данных в таблице нет.

5. С замечанием согласен.

Замечания официального оппонента В.Т.Калайды

В работе, при обосновании использования метод частиц в ячейках для построения теста, автор аргументирует выбор высокими требованиями этого метода к быстродействию процессоров, оперативной памяти, и системы хранения данных, перечисляя его преимущество перед другими вариантами. При этом никаких альтернативных подходов не приводится.

С замечанием согласен

В разделе 1диссертации декларируется, что число операций вычислительного алгоритма метода частиц в результате автоматической оптимизация программы «не даёт почти ничего». Оценок, подтверждающий это утверждение в работе нет.

Замечание справедливо, вместе с тем таблица 25, стр. 131,показывающая влияние оптимизирующих опций компилятора Intel C/Fortran, а также таблица 7, стр.63, в которой показана сравнительная эффективность различных подходов к организации вычислений с частицами, отчасти отвечают на этот вопрос в том смысле, что ни один из перечисленных в обоих таблицах подходов не имеет явных преимуществ над другими

Замечания официального оппонента В.Т.Калайды

Значения оценок производительности систем (и подсистем), приведённых в работе, из-за ограниченного объёма расчётов, можно рассматривать как предварительные.

С замечанием согласен. Вместе с тем можно сказать что расчеты проведены на всем доступном оборудовании, и более того, эти расчеты проводились именно как моделирование (получить доступ к ресурсам ВВС для разработки тестов вряд ли возможно...), а не как тесты производительности, с чем отчасти связан их ограниченный объем

Положения, выносимые на защиту, содержат не чёткие, а зачастую и некорректные определения типа: «наиболее заметно снижающую…» (положение № 1), «по более оптимальному…» (положение № 2), «более достоверные…» (положение № 4).

С замечанием согласен.

Замечания официального оппонента Вл.В.Воеводина

1. В разделе 3 проведено сравнение с известными тестами производительности вычислительных систем, такими как HPL, HPCG. Сравнение показывает, что существует важное отличие разработанного теста от уже существующих, а именно, значительно меньшая зависимость от правильного подбора конфигурации запуска. Не понятно с чем это связано. Не приводится обоснование отсутствия этой меньшей степени зависимости: обычно тесты могут как раз и настраиваться на учет зависимости от архитектуры, чтобы выжать из последней все особенности и специфику.

С замечанием согласен. Меньшая зависимость связана с тем, что при расчете движения модельных частиц фактически происходит усреднение времен выполнения различных операций по очень большому (равному числу частиц) количеству измерений, котрые выполняются при разной загруженности ВС. Таким образом возможность запуска теста в “неудачный” момент фактически отсутствует.

Замечания официального оппонента Вл.В.Воеводина

2. Описывается методика измерений, которая проверяется на четырех доступных системах. Отсутствуют обоснования выбора именно этих систем.

С замечанием согласен. Вместе с тем можно сказать что расчеты проведены на всех ВВС, доступных на момент выполнения работы

Замечания официального оппонента Вл.В.Воеводина

3. Рубрикация работы не во всех разделах удобна и обоснована. Например, в главе 1 параграф 1.4 «Параллельная реализация» имеет только один подпараграф. То же самое и в параграфе 3.4. главы 3. Согласно требованиям к рубрикации научной работы, если параграф делится на подпараграфы, то их должно быть не менее двух.
4. По Главе 1 нет выводов, а представлен только практический результат, который предшествует научным результатам.

С замечанием согласен.

Замечания ведущей организации

1)Вторую главу, посвященную обзору литературы, логично было бы сделать первой, как обычно принято.

С замечанием согласен

2)На стр. 11 имеется неточность. Автор диссертации указал себя как руководитель ряда грантов РФФИ. Однако из 5 грантов он является руководителем только в двух: 14-07-00241 и 18-07-00364

С замечанием согласен

3) стр. 18 используется схема с перешагиванием. А почему бы не использовать схему Верле, в которой еще есть скорость на целом шаге по времени, и которая хорошо себя зарекомендовала в задачах молекулярной динамики?

С замечанием частично согласен. С точки зрения корректности расчетов по методу частиц в ячейках основным является большое количество модельных частиц, а не качество расчета движения отдельных частиц, а для построения теста выбор расчетной схемы малосущественен

Замечания ведущей организации

4) В уравнении (1.16) индексы написаны с ошибками. Написан так, что порядок аппроксимации – первый, а не второй, как указано.

С замечанием согласен

5) Стр. 58. Компьютерная техника развивается очень быстро. При расчете энергопотребления надо ориентироваться не на суперкомпьютеры Sunway TaihuLight и Tianhe-2, которые сейчас находятся на 3 и 4 местах в Top-500, и которые построены на обычных многоядерных процессорах, а на суперкомпьютеры, построенные на новых GPU Nvidia Volta GV100. С июня 2018 г. Два первых места в рейтинге Top-500 занимают суперкомпьютеры Summit (200 петафлопс, 10 мегаватт) и Sierrra (126 петафлопс, 7 мегаватт). При пересчете на экзафлопс получится мощность только 50 мегаватт, что вполне реально.

С замечанием частично согласен. Возможно, при оценке энергопотребления экзафлопсных машин было бы полезно ориентрироваться не только на самые лучщие образцы, но также преполагать и большие показателя энергозатрат.

Замечания ведущей организации

5)В уравнении (1.16) индексы написаны с ошибками. Написан так, что порядок аппроксимации – первый, а не второй, как указано.

С замечанием согласен

6) Стр. 58. Компьютерная техника развивается очень быстро. При расчете энергопотребления надо ориентироваться не на суперкомпьютеры Sunway TaihuLight и Tianhe-2, которые сейчас находятся на 3 и 4 местах в Top-500, и которые построены на обычных многоядерных процессорах, а на суперкомпьютеры, построенные на новых GPU Nvidia Volta GV100. С июня 2018 г. Два первых места в рейтинге Top-500 занимают суперкомпьютеры Summit (200 петафлопс, 10 мегаватт) и Sierrra (126 петафлопс, 7 мегаватт). При пересчете на экзафлопс получится мощность только 50 мегаватт, что вполне реально.

С замечанием согласен. Возможно, при оценке энергопотребления экзафлопсных машин было бы полезно ориентрироваться не только на самые лучщие образцы, но также преполагать и большие показателя энергозатрат.

Замечания ведущей организации

6) Рис. 4.7 и 4.8. На оси подписано “количество ядер”, что противоречит таблице 16, где указано количество GPU, что совсем не одно и то же (каждый используемый GPU (Nvidia Tesla X2070 содержит 512 ядер).

С замечанием согласен.В обоих случаях имелось в виду количество MPI-процессов, т.е. количество CPU-ядер, задействованных в операциях MPI

7) Рис. 4.8. График аппроксимации даже близко не соответствует расчетным точкам. Тот же рисунок и в автореферате (рис. 4).

С замечанием согласен

8) Надо было все три аппроксимации (рис. 4.13;4.14 и 4.15 привести на одном графике, чтобы было легче их сравнивать. Более того , все три аппроксимации совсем не соответствуют полученным данным. Как это так вышло? То же самое относится к рисункам 4.16; 4.17 и 4.18.

С замечанием согласен

Замечания ведущей организации

9) Глава 5. Вопрос о переносе программ с GPU на ускорители Intel Xeon Phi в настоящее время становится не актуальным. Во-первых, потому, что линия Intel Xeon Phi полностью снимается с производства уже в этом году! Во-вторых, как уже говорилось, два самых мощных компьютера в мире “Summit” (200 петафлопс) и “Sierra” (126 петафлопс) построены на новых графических ускорителях Nvidia “Volta GV100” (5376 ядер CUDA каждый).

С замечанием согласен, но наработки Intel Xeon Phi будут использованы при разработке следующих поколений процессоров Intel Xeon, кроме того, речь идет в большей степени о переносе на OpenMP, чем конкретно на Phi. Упоминание Intel Xeon Phi связано с тем, что на тот момент эта архитектура рассматривалась как перспективное направление развития ССКЦ СО РАН и МСЦ РАН.

10) Автореферат, список литературы. Пункты [21] и [22] уже приведены под номерами [13] и [16].

С замечанием согласен

Сведение к минимуму архитектурно-зависимых участков кода

- В коде имеется 15-20 вызовов небольших вычислительных фрагментов,
  - выполняющих обработку узлов сетки, модельных частиц, границ расчетной области
  - оформленных в виде ядер CUDA
  - таким образом, не подлежащих компиляции в помощью компилятора Intel и пр.
- Идея: сделать такой “непроходной” участок кода по крайней мере единственным

Моделирование плазмы методом частиц в ячейках (осн.рез. 1)

Основные уравнения

$$\frac{\partial f_{i,e}}{\partial t}+\vec{v}\frac{\partial f_{i,e}}{\partial \vec{x}}+\vec{F}\frac{\partial f_{i,e}}{\partial \vec{v}}=0$$

$$\nabla\times\vec{B}=4\pi\vec{j}+\frac{1}{c}\frac{\partial \vec{E}}{\partial t}$$

$$\nabla\times\vec{E}=-\frac{1}{c}\frac{\partial \vec{B}}{\partial t}$$

$$\nabla\vec{E}=4\pi\rho$$

$$\nabla\vec{B}=0$$

$$\vec{p}=\gamma\vec{v},\ \gamma^{-1}=\sqrt{1-v^2}$$

$$\vec{F}=q_{i,e}\left(\vec{E}+\frac{1}{c}[\vec{v},\vec{B}]\right)$$

$$\vec{j}=\sum_{i,e}q_{i,e}\int f_{i,e}\vec{v}\,d\vec{v}$$

$$\rho=\sum_{i,e}q_{i,e}\int f_{i,e}\,d\vec{v}$$

- **Граничные условия:** периодические

Начальные условия

- $\rho_e = 1000,\ \rho_b = 1$

  $\rho = \rho_e+\rho_b$
- Импульсы электронов плазмы:

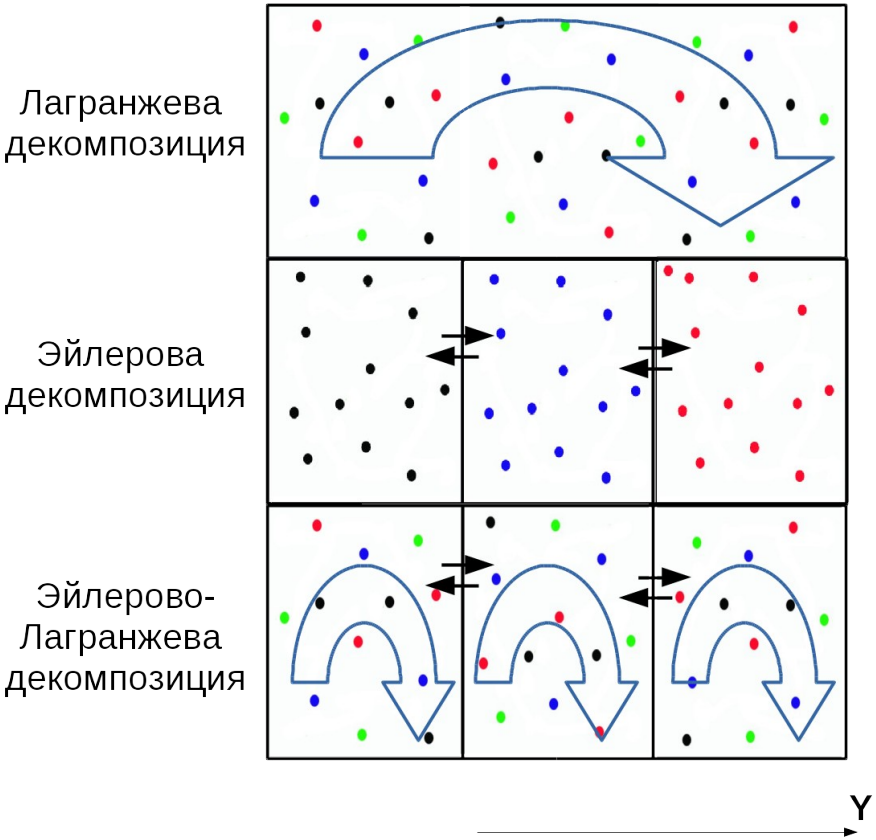
  $\boldsymbol{p_x, p_y, p_z}$ — максвелловское распределение, $\sigma = T_e = 1.0$

$$f=\exp\left(\frac{-p^2}{\sigma}\right)$$

- Импульсы ионов плазмы:0
- Импульс электронов пучка:

  $\boldsymbol{p_x = 50\ p_y = p_z = 0}$

Реализация метода частиц в ячейках для высокопроизводительных ВС



Операция MPI_Allreduce

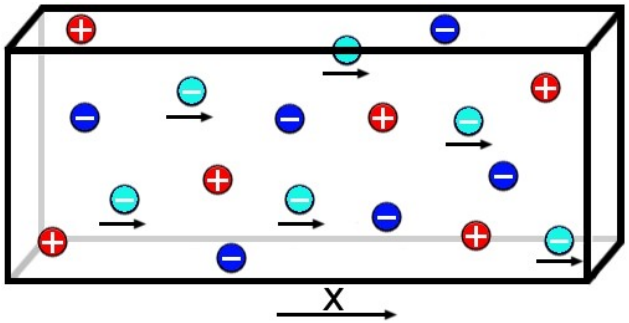
Операция MPI_Sendrecv

Оригинальный вариант декомпозиции, созданный в данной работе

- В силу того, что метод частиц в ячейках создает “плохой” (очень нерегулярный) шаблон доступа к памяти, то полученная по данному методу производительность фактически является **оценкой снизу** для других вычислительных методов
- В рамках диссертации созданы несколько вариантов численного решения задачи моделирования динамики плазмы (созданы пакеты программ):
  - Реализовано три вида пространственной декомпозиции
  - Разработаны реализации для GPU и MIC (Intel Xeon Phi)
  - Выполнены реализации на Fortran и C/C++

Метод частиц в ячейках в моделировании плазмы

В методе частиц в ячейках среда разбивается на модельные частицы, траекториями движения которых являются характеристики кинетического уравнения Власова



$$\frac{\partial p_{i,e}}{\partial t} = \kappa\left(E + [v, B]\right),$$

$$\frac{\partial r_{i,e}}{\partial t} = v_{i,e}.$$

$$p_{i,e} = \frac{v_{i,e}}{\sqrt{1 - v_{i,e}^2}}, \quad \kappa_e = -1, \quad \kappa_i = m_e / m_i.$$

- Уравнения движения модельных частиц решаются с помощью схемы с перешагиванием
- Уравнения Максвелла решаются методом FDTD (метод Yee)
- Вычислительная сложность: 500 операций с плавающей точкой на частицу

Реализация численных методов выполнена на основе работы: **Вшивков В.А.** и др.,[52] Вычислительные технологии, Том 6, № 2, 2001.

Реализация метода частиц в ячейках для GPU

Оригинальный результат, созданный в данной работе

- Движение модельных частиц является наиболее вычислительно-сложной частью расчета (до 90 %)
- В то же время именно эта часть алгоритма наиболее заметно ускоряется при переходе на GPU
- Поэтому:
  1. Частицы хранятся в виде массивов по ячейкам
  2. Расчет движения частиц выполняется на разделяемой памяти
  3. При пересылке частиц между ячейками исключена синхронизация

Кластер НКС-30Т (ИВМиМГ СО РАН):
Сервера SL390s G7 (85 Тфлопс):
CPU (2 x) 6-ядерный Xeon X5670 (2.93 Ггц),
RAM 96 Гбайт,
GPU (3 x) NVIDIA Tesla M 2090
на архитектуре Fermi (compute capability 2.0)