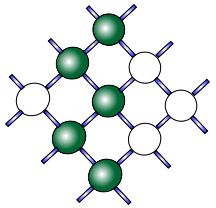
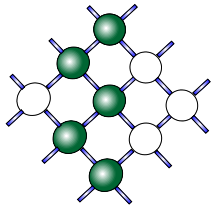


Федеральное государственное бюджетное учреждение науки
Институт физики полупроводников им. А.В. Ржанова СО РАН
Лаборатория вычислительных систем
пр. Ак.Лаврентьева, 13 , 630090, Новосибирск, Россия
т. +7 (383) 333-21-71

Анализ и организация функционирования распределенных вычислительных систем, разработка параллельных алгоритмов и программ

ПАВСКИЙ КИРИЛЛ ВАЛЕРЬЕВИЧ
заведующий Лабораторией ВС ИФП СО РАН
доктор технических наук, доцент

Новосибирск, 2024 г.



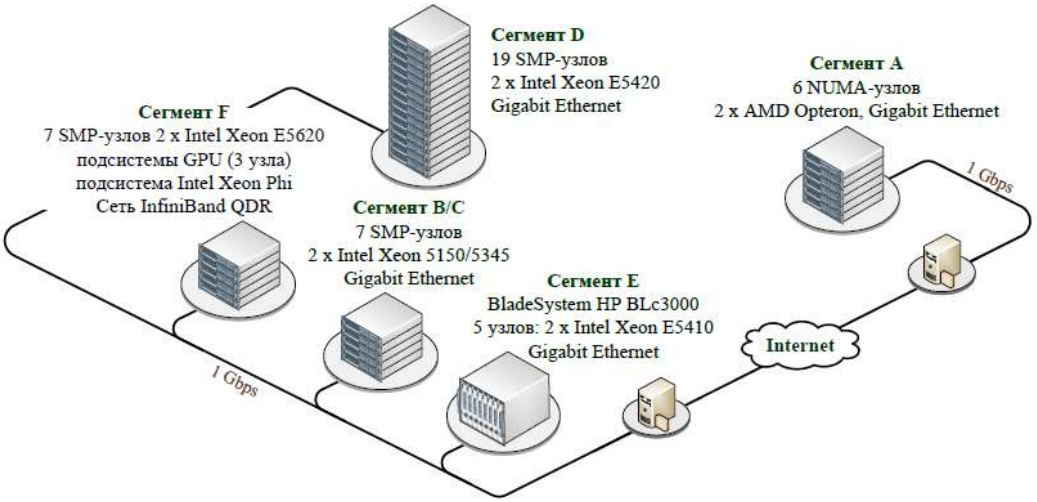
Темы работ

1. Исследование и разработка средств анализа функционирования масштабируемых вычислительных систем и расчет показателей надежности.
2. Разработка отказоустойчивых параллельных алгоритмов и программ для моделирования физических процессов при гетероэпитаксии германия на структурированных подложках кремния.
3. Разработка алгоритмических и программных средств исследования отказов и сбоев в высокопроизводительных вычислительных системах.
4. Разработка алгоритмов и программ моделирования объектов, построение проекций при томографии.
5. Разработка алгоритмов и программ обработки изображений для реконструкции объектов исследования методом птихографии.

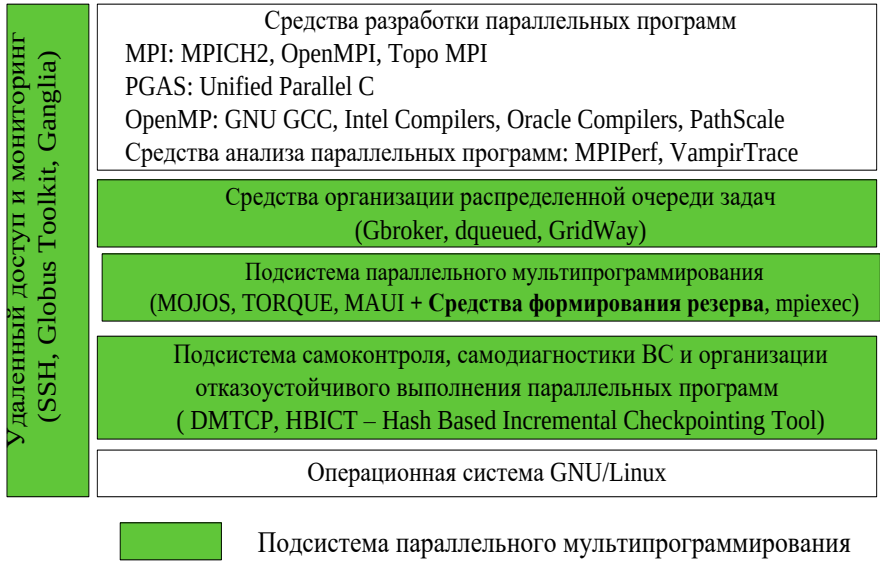
Федеральное государственное бюджетное учреждение науки Институт физики полупроводников им. А.В. Ржанова СО РАН
Лаборатория вычислительных систем
пр. Ак.Лаврентьева, 13 , 630090, Новосибирск, Россия
т. +7 (383) 333-21-71, 330-56-26

В Лаборатории вычислительных систем ИФП СО РАН ведутся научные исследования по следующим направлениям:

- Методы, алгоритмы и системное программное обеспечение организации функционирования распределенных вычислительных систем (ВС);
- Моделирование и анализ функционирования распределенных вычислительных систем;
- Разработка параллельных алгоритмов и программ.



Структура программного обеспечения мультикластерной вычислительной системы

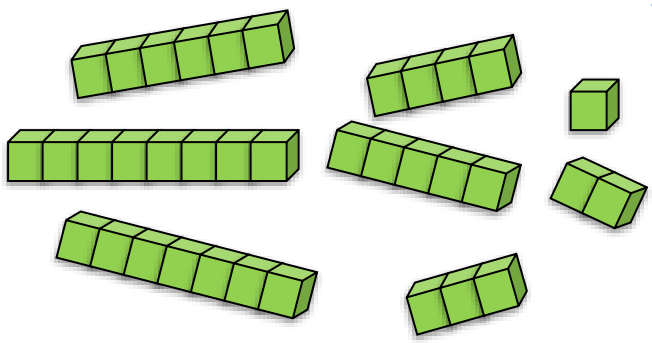


Мультикластерная ВС создана ЦПВТ СибГУТИ совместно с Лабораторией ВС ИФП СО РАН.
Мультипрограммные режимы функционирования организуются средствами, разработанными членами ведущей научной школы РФ (НШ-9505.2006.9, НШ-2121.2008.9, НШ-5176.2010.9, НШ-2175.2012.9, руководитель – чл.корр. РАН В.Г. Хорошевский).

Основное назначение ВС – исследование архитектуры и отработка инструментария параллельного мультипрограммирования, моделирование распределенных вычислительных технологий и подготовка научно-педагогических кадров.

Инструментарий параллельного мультипрограммирования

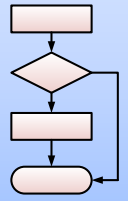
Модели, методы и программное обеспечение ,
предназначенное для оптимизации использования ресурсов распределенных ВС
при решении множества задач, представленных параллельными программами



Поток параллельных задач

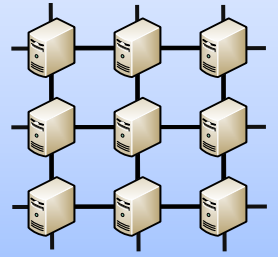
– единичный ранг

Распределенная операционная система



&

Вычислительная система

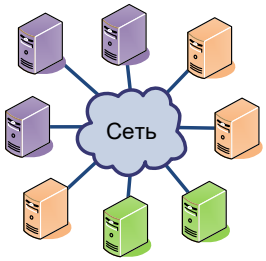


Монозадачный режим

Мультизадачные режимы

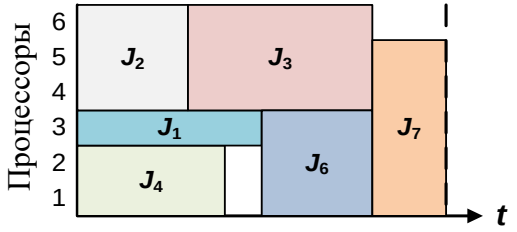
Обслуживание потоков задач

Генерация подсистем в пределах ВС



Обработка наборов задач

Формирование расписаний решения параллельных задач

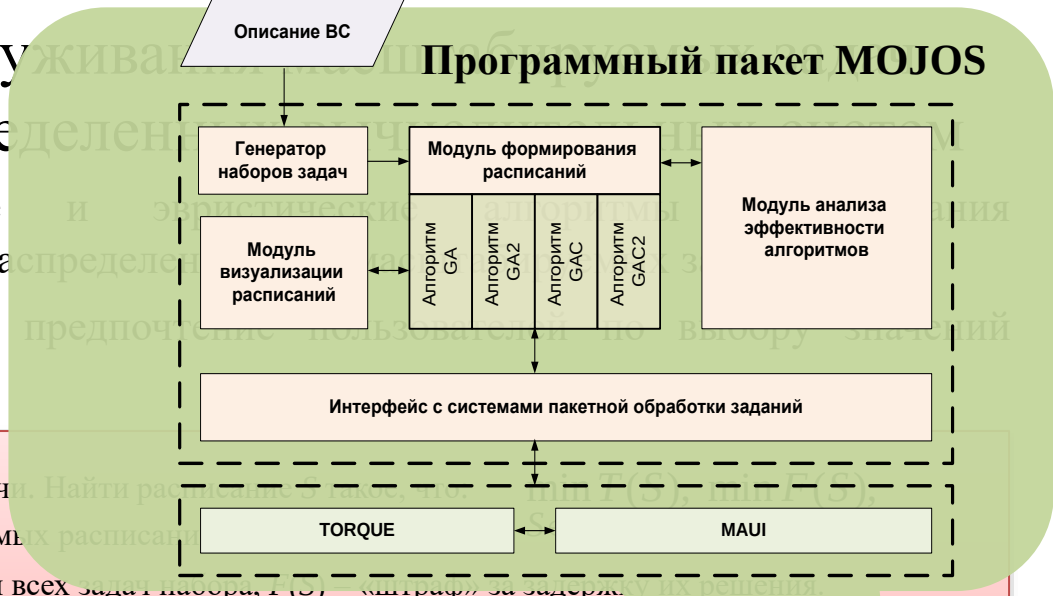


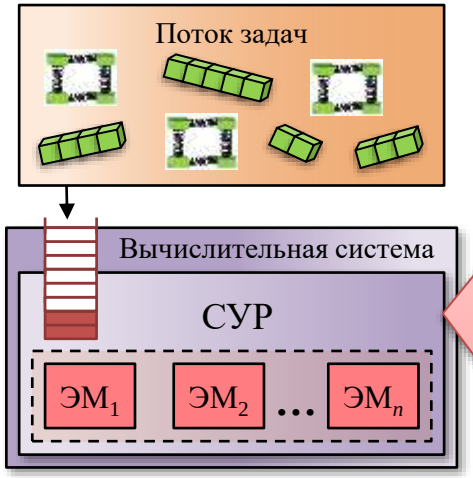
Точные, эвристические, стохастические
методы и алгоритмы

Алгоритмы обслу… на ресурсах распре…

Разработаны стохастические … субоптимальных расписаний решения на ра…

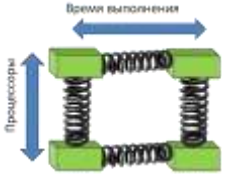
Разработанные **алгоритмы учитывают** … параметров задач (ранг и время решения).





Параллельная задача

Масштабируемая задача



– единичный ранг

Свойством масштабируемости обладают более 80% задач, решаемых на вычислительных системах.

Постановка задач… Найти расписание S такое, что… $\min T(S)$, $\min P(S)$,

Ω – область допустим… расписан… S.

$T(S)$ – время решения всех задач набора, $P(S)$ – «штраф» за задержку их решения.

Задача относится к целевому программированию, является NP-тр… ной.

Разработано семейство полиномиальных алгоритмов многокритериал… ной оптими…зации мультипрограммного функционирования распределённых вычислительных си… стем в р…ежиме обработки наборов масштабируемых задач.

Результаты экспериментального исследования показали, что отклонение суммарного времени решения задач от нижней границы целевой функции T(S) составляет 15 – 25 %.

Предложен способ описания масштабируемой задачи:

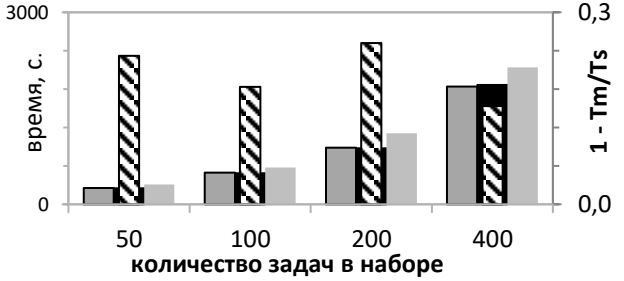
$$Request_1[,Request_2] \dots [,Request_i] \dots [Request_n]$$

nodes=value[@ppn=value][@walltime=value][@weight=value]

Разработано семейство эвристических алгоритмов обслуживания масштабируемых задач на распределенных ВС.

Модернизировано программное обеспечение СУР PBS Torque и планировщика MAUI.

Проведено экспериментальное исследование эффективности решения масштабируемых задач.

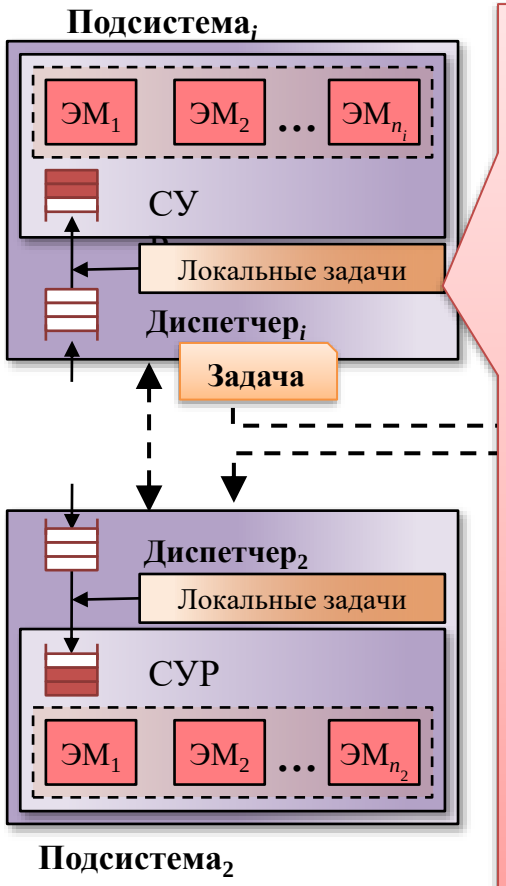


До 25 % снижение времени решения всех задач набора,
до 15 % снижение среднего времени ожидания задачи в очереди.

Децентрализованная диспетчеризация параллельных программ в пространственно-распределённых мультикластерных и GRID-системах

Разработаны алгоритмы децентрализованной диспетчеризации, учитывающие текущую загрузку подсистем распределённых ВС и производительность каналов связи.

При децентрализованной диспетчеризации коллектив диспетчеров совместно принимает решение о выборе ресурсов для задач. Это позволяет достичь живучести ВС и снизить сложность поиска ресурсов для задач.



Шаг 1. Выбирается подсистема j^* с минимальным значением $F(j), j \in S(j)$

$$F(j) = \begin{cases} \dfrac{t_j}{t_{\max}} + \dfrac{c_j^{-1}}{c_{\max}^{-1}} + \dfrac{w_j}{w_{\max}}, & \text{если } c_j < r \text{ или } q_j > 0, \\ \dfrac{t_j}{t_{\max}}, & \text{иначе.} \end{cases}$$

c_j – количество свободных ЭМ в подсистеме j;

q_j – количество задач в состоянии ожидания;

$w_j = q_j / n_j$ – количество задач в очереди, приходящееся на одну ЭМ;

$t_j = \sum_{l=1}^{k} t(h_l, j, z_l)$ – оценка времени доставки файлов задачи до подсистемы j

Шаг 2. Задача направляется в очередь локальной СУР подсистемы j*.

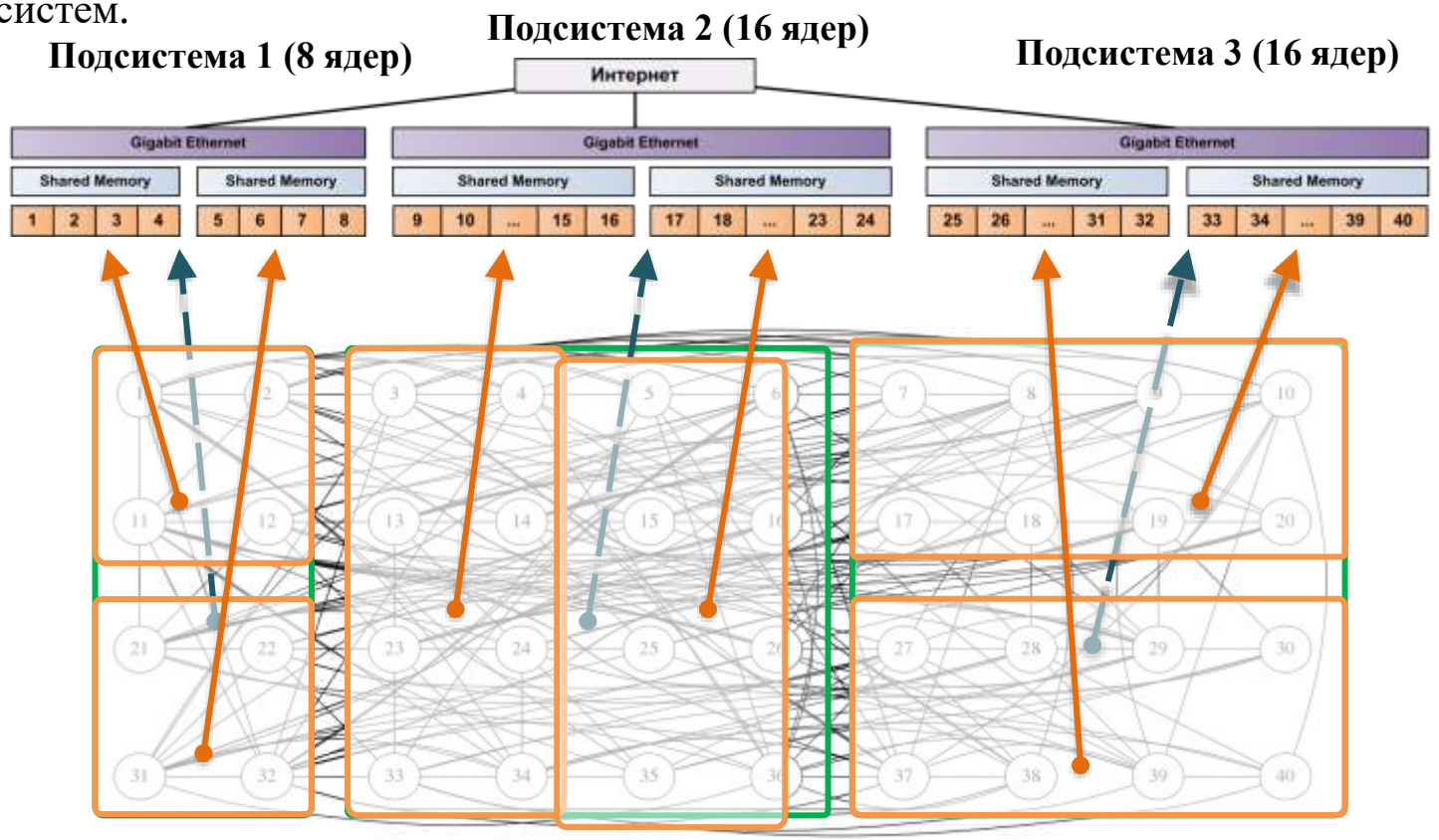
Разработанные алгоритмы (создан пакет GBroker) позволяют снизить в среднем в 1,2 раза время обслуживания задач по сравнению с централизованной диспетчеризацией (GridWay)

Вложение параллельных программ в пространственно-распределённые ВС

Разработаны эффективные алгоритмы вложения в иерархические ВС параллельных программ. Алгоритмы в сравнении с известными средствами учитывают все уровни иерархической организации коммуникационных сред ВС и обеспечивают суб.мин. время выполнения дифф.-ных обменов.

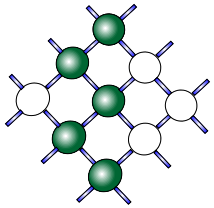
В частности, **предложен метод HierarchicTaskMap**, основанный на рекурсивном разбиении графа задачи и позволяющий учитывать все коммуникационные уровни пространственно-распределённых вычислительных систем.

Для современных распределенных ВС систем характерны иерархическая организация и различные пропускные способности каналов связи между их ресурсами.



Вложение параллельной программы The Parallel Ocean Program (POP), *N* = 40 в мультикластерную ВС из трёх подсистем

Время выполнения программы уменьшилось в 1,6 раз по сравнению со стандартным вложением



Анализ функционирования распределенных вычислительных систем

Разработка методов и построение математических моделей для расчета показателей эффективности функционирования большемасштабных распределенных вычислительных систем.

Анализ функционирования распределенных вычислительных систем

Объект исследования

Объектом исследования являются распределенные вычислительные системы

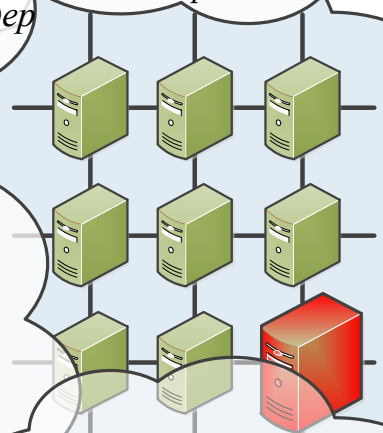
- Распределенная ВС представляется в виде композиции множества **элементарных машин** (узлов, процессоров, ядер и т.п.) соединенных коммуникационной сетью.
- Все основные ресурсы являются и логически технически **распределенными**.

Sunway TaihuLight

***40 960** выч.узлов,*
***10 649 600** ядер*
1 место в 50-ой ред. Top 500

Tianhe-2

***16 000** выч.узлов, **3012000** ядер*
2 место в 50-ой ред. Top 500



Cray XK7 Titan

***18 688** выч.узлов,*
***299008** ядер*
5 место в 50-ой ред. Top 500

IBM BlueGene/Q Sequoia

***98 304** выч.узлов,*
*1 **572 864** ядер*
6 место в 50-ой ред.

Проблема – *Надежность ВС*

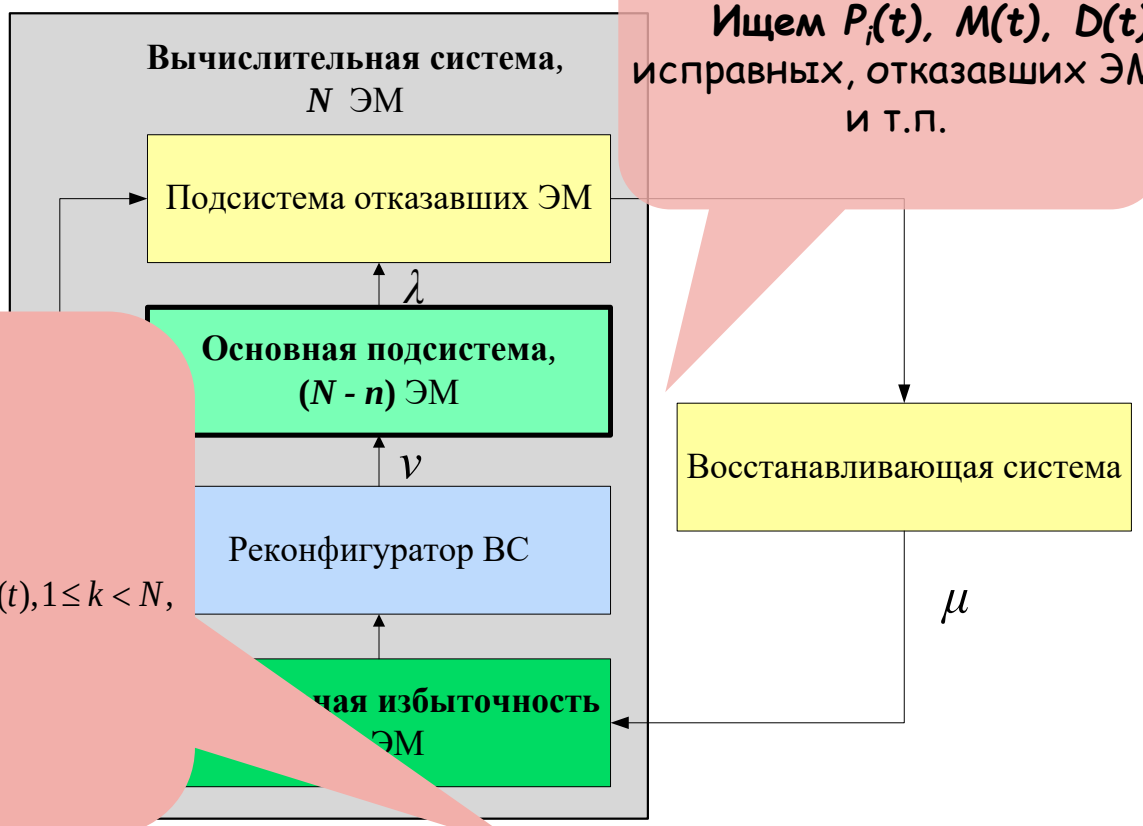
В силу своей **большемасштабности** такие ВС эффективно исследуются **стохастическими методами**.

Модель функционирования вычислительной системы

- ***N*** – количество элементарных машин (ЭМ) ВС;
-

Пример:

$$\begin{cases} \frac{d}{dt}P_0(t) = -\lambda_0 \cdot P_0(t) + \mu_1 \cdot P_1(t), \\ \dots \\ \frac{d}{dt}P_k(t) = -(\lambda_k + \mu_k) \cdot P_k(t) + \lambda_{k-1} \cdot P_{k-1}(t) + \mu_{k+1} \cdot P_{k+1}(t), 1 \le k < N, \\ \dots \\ \frac{d}{dt}P_N(t) = -\mu_N \cdot P_N(t) + \lambda_{N-1} \cdot P_{N-1}(t). \end{cases}$$



Определяем процесс функционирования ВС. **Ищем $P_i(t)$, $M(t)$, $D(t)$** исправных, отказавших ЭМ и т.п.

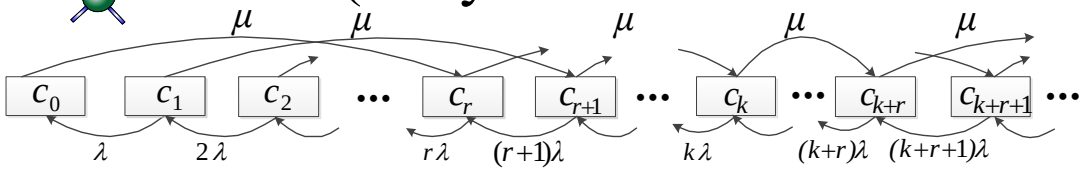
Распределение вероятностей состояния системы $\{P_0(i, t), P_1(i, t), \dots, P_N(i, t)\}$

$P_j(i, t)$ – вероятность того, что в системе, начавшей функционировать в состоянии $i \in \{0, 1, \dots, N\}$, в момент времени $t \ge 0$ будет j отказавших машин

1. Хорошевский В.Г. Модели анализа и организации функционирования большемасштабных распределенных вычислительных систем. // Электронное моделирование. - Киев, 2003. - Т. 25, № 6, С. 21-35.

2. Хорошевский В.Г. Павский В.А., Павский К.В. Вычисление показателей живучести распределенных вычислительных систем и осуществимости решения задач// Искусственный интеллект, №4, «Наука і освіта» ДонДІШІ, ,2006, С. 28 – 34.

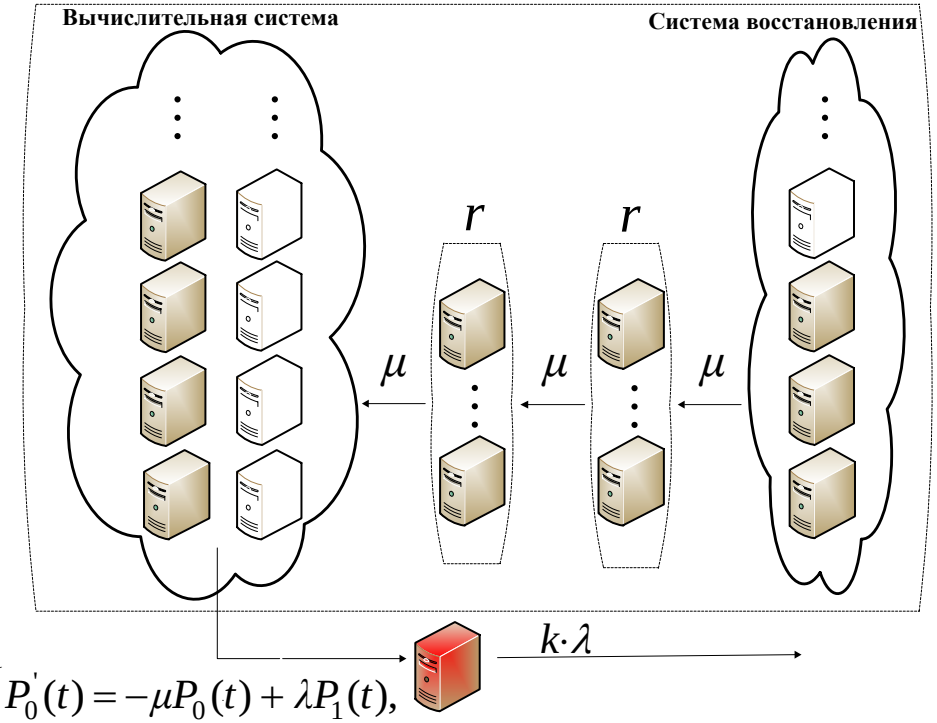
Модель функционирования распределенной вычислительной системы при групповом восстановлении (получена качественная и количественная оценка)



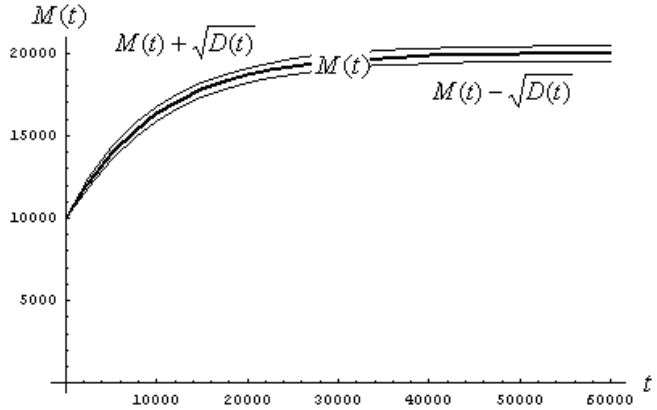
Построена математическая модель. **Найдены** решения для расчета мат. ожидания и дисперсии машин ВС.

$$\begin{cases} \dfrac{\partial}{\partial t}M(t)+\lambda M(t)=r\mu, \\ \dfrac{\partial}{\partial t}Q(t)+2\lambda Q(t)=2r\mu M(t)+r(r-1)\mu, \\ D(t)=Q(t)+M(t)-M^2(t), \end{cases}$$

$$M(0)=n, \qquad D(0)=0$$



$$\begin{cases} M(t)=\dfrac{r\mu}{\lambda}+\left(n-\dfrac{r\mu}{\lambda}\right)e^{-\lambda\cdot t}, \\ D(t)=Ce^{-2\lambda\cdot t}+\dfrac{r\mu}{\lambda}\left(\dfrac{r\mu}{\lambda}+2\left(n-\dfrac{r\mu}{\lambda}\right)e^{-\lambda\cdot t}\right)+\dfrac{r(r-1)\mu}{2\lambda}+M(t)-M^2(t), \\ C=-\dfrac{r\mu}{\lambda}\left(2n-\dfrac{r\mu}{\lambda}+\dfrac{(r-1)}{2}\right)-n+n^2. \end{cases}$$



Математическое ожидание исправных машин

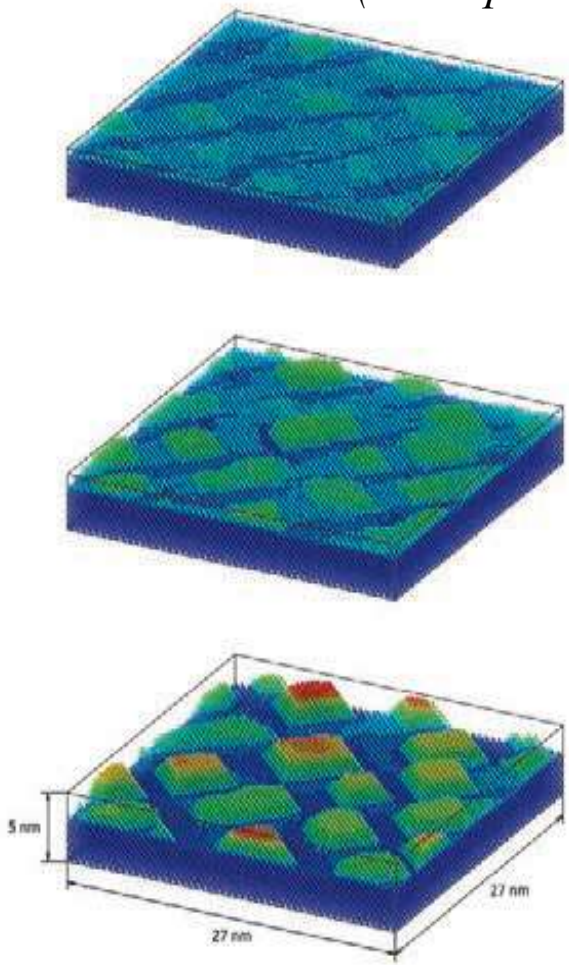
$$\begin{cases} P_0'(t)=-\mu P_0(t)+\lambda P_1(t), \\ P_k'(t)=-(\mu+k\lambda)P_k(t)+(k+1)\lambda P_{k+1}(t), \ 0<k<r, \\ P_k'(t)=-(\mu+k\lambda)P_k(t)+\mu P_{k-r}(t)+(k+1)\lambda P_{k+1}(t), \quad k\geq r. \end{cases}$$

$$P_n(0)=1, \quad P_k(0)=0, \quad k\neq n, \quad k=0,1,2,\ldots;$$

$\lambda=10^{-4}\ ч^{-1} \quad \mu=0.1\,ч^{-1} \quad r=20 \quad n=10^4\ ЭМ$

Разработка параллельных алгоритмов моделирования гетероэпитаксиального роста

Моделирование процесса зарождения и роста трехмерных островков Ge на Si(100)
(Лаборатория неравновесных полупроводниковых систем)



- Приоритетной задачей в области материаловедения является создание пространственно упорядоченных массивов полупроводниковых квантовых точек (КТ).
- Одним из методов получения пространственно упорядоченных массивов КТ является молекулярно-лучевая эпитаксия (МЛЭ) на структурированных подложках
- Механизм роста на поверхности со сложным рельефом изучен не достаточно. Поэтому актуальной проблемой является исследование зарождения и роста КТ в ямках с помощью алгоритмов моделирования данного процесса.

(Методы МК и МД)

Рис. Эволюция поверхности кристалла в процессе моделирования роста Ge на Si(100) в следующих условиях: температура 400° С, скорость осаждения Ge 0.1 МС/с. Количество осажденного германия: 3 МС (а), 4 МС (б), 5 МС (в). Цвет соответствует высоте рельефа поверхности

Актуальность

Эпитаксия - это закономерное нарастание одного кристаллического материала на другом, при более низких температурах, то есть ориентированный рост одного кристалла на поверхности другого (подложки).

Одним из видов эпитаксии является **гетероэпитаксия** - когда растущий слой отличается по химическому составу от вещества подложки.

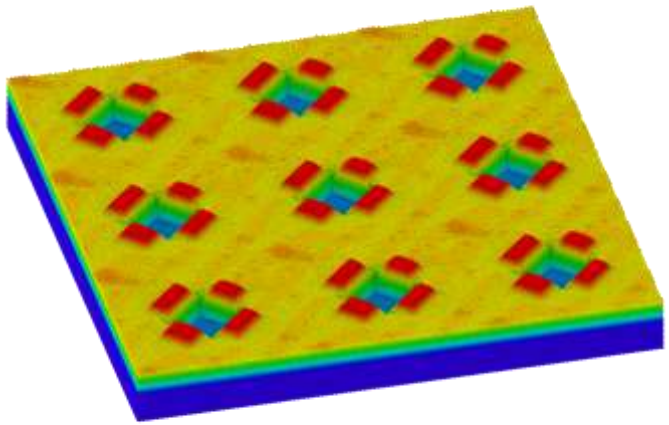


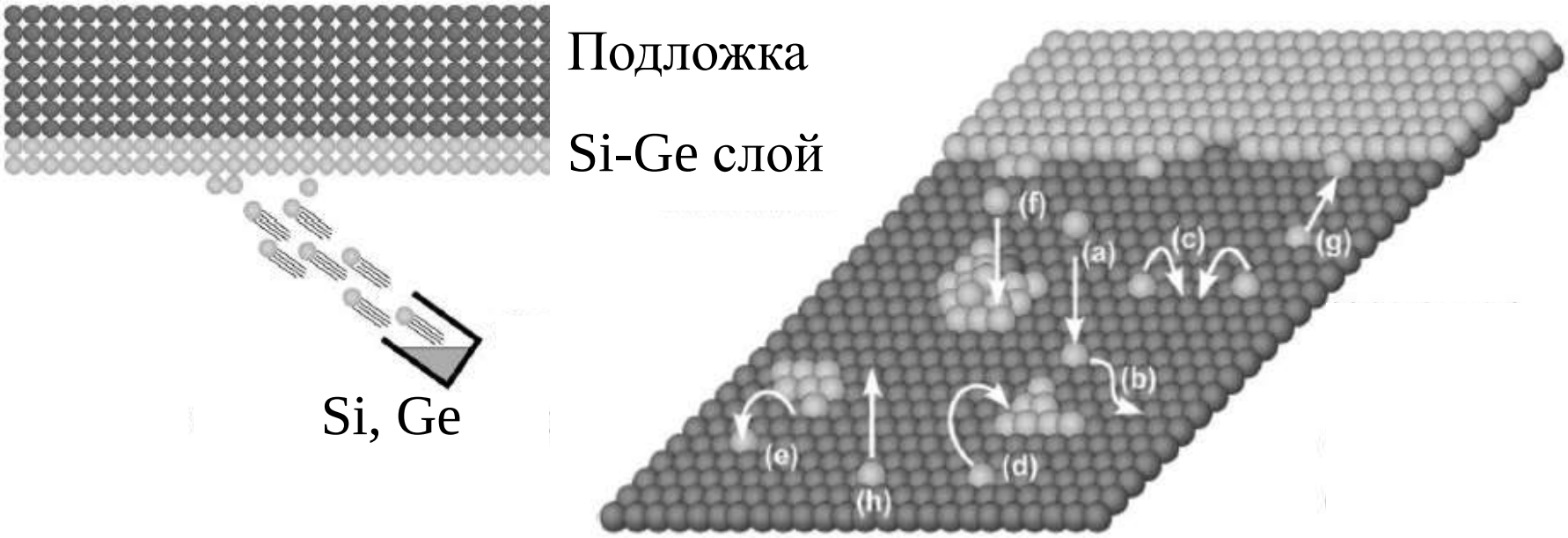
Рис. 1 - Ямки и зарождение Квантовых точек

Например, для моделирования процесса гетероэпитаксии [1] (**ИФП СО РАН**) на подложке размером **38×38×8 нм^3** (~586 тыс. атомов) требуется **2616 ч** (**~109 суток**) на узле кластера **ИВЦ НГУ**.

Следовательно, **актуальной** является **разработка параллельных алгоритмов и программ для моделирования гетероэпитаксии**.

[1] *Rudin S.A., Zinovyev V.A., Smagina Zh.V., Novikov P.L., Nenashev A.V., Pavsky K.V. Groups of Ge nanoislands grown outside pits on pit-patterned Si substrates //* Journal of Crystal Growth*, Available online 12 June 2022, 126763 In Press, Journal Pre-proof* https://doi.org/10.1016/j.jcrysgro.2022.126763 https://www.sciencedirect.com/science/article/pii/S0022024822002512?via%3Dihub

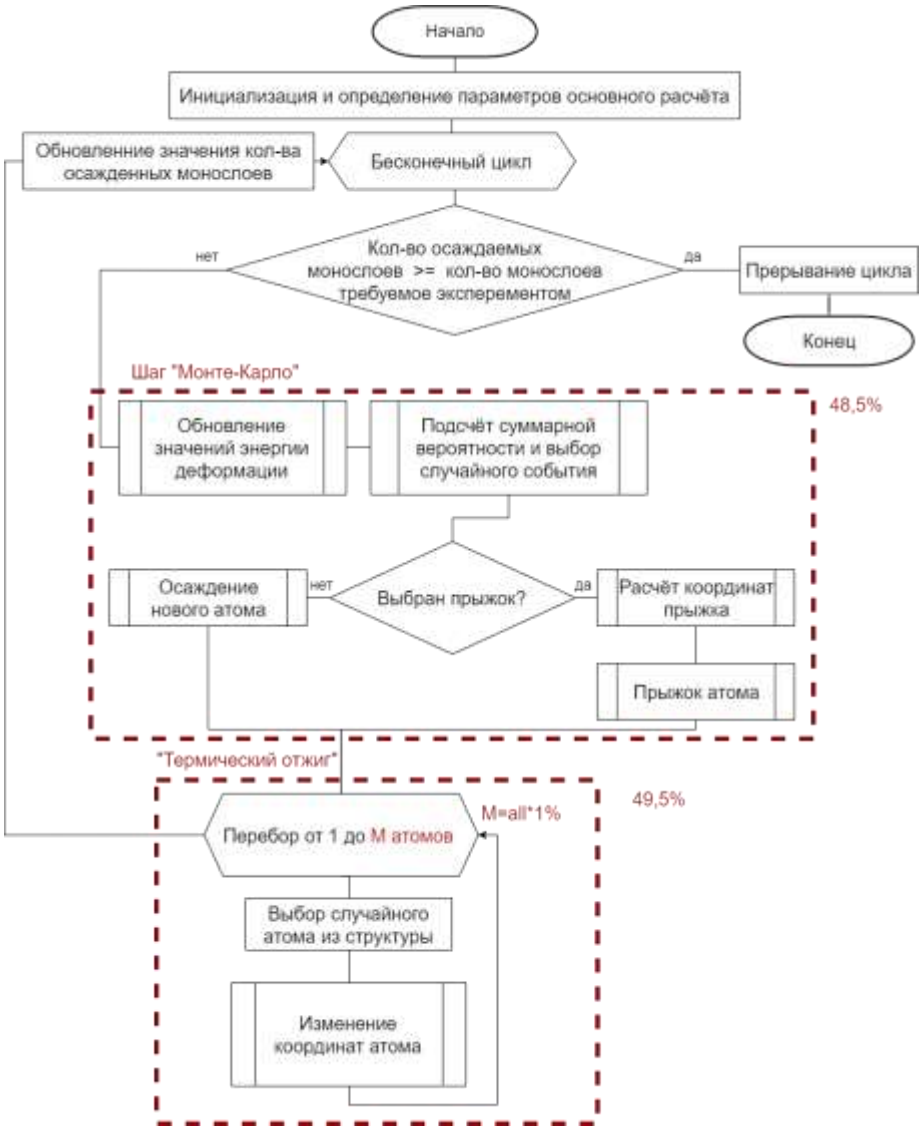
Молекулярно-лучевая эпитаксия
Метод Монте -Карло



(a) осаждение на поверхность
(b) диффузия по поверхности
(c) образование островка
(d) присоединение атомов к островку
(e) отсоединение атомов от островка
(f) осаждение на островок
(g) присоединение к ступени
(h) десорбция с поверхности

Bert Voigtländer. Fundame`tal processes in Si/Si and Ge/Si epitaxy studied by scanning tunneling microscopy during growth // Surface Science Reports 2001. V. 43, Issues 5-8, P. 127-254

Монте-Карло моделирование гетероэпитаксии
Основной цикл



Обновление значения энергии деформации {

```
…
#pragma omp parallel for private(dir)
reduction(+:Bx, By, Bz)
…
#pragma omp parallel for private(dir)
…
}
```

Вычисление суммарной вероятности прыжков атомов {

```
…
#pragma omp parallel for private(x,y,z)
reduction(+:real_sum)
…
#pragma omp parallel for private(dir)
reduction(+:E_a)
…
}
```

Прыжок атома {

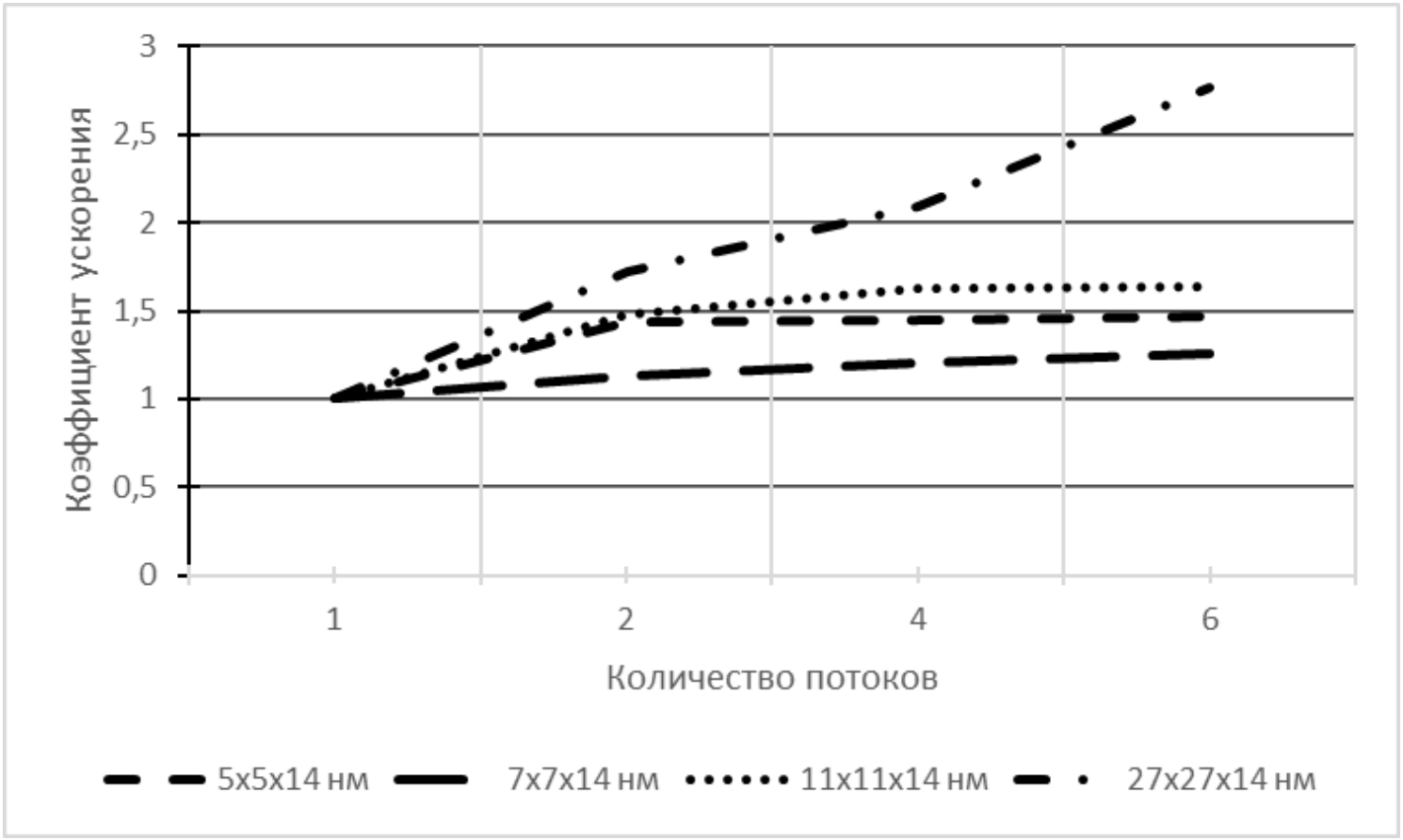
```
…
#pragma omp parallel for private(n)
shared(bad_jump)
…
#pragma omp parallel for private(dir)
reduction(+:Bx, By, Bz)
…
#pragma omp critical v_ochered_Edef
…
}
```

Осаждение атома {

```
…
#pragma omp parallel private(dir)
…
#pragma omp parallel for private(dir)
reduction(+:Bx, By, Bz)
…
#pragma omp critical v_ochered_Edef
…
#pragma omp parallel for private(dir)
reduction(+:n_jumps)
…}
```

Эффективность распараллеливания

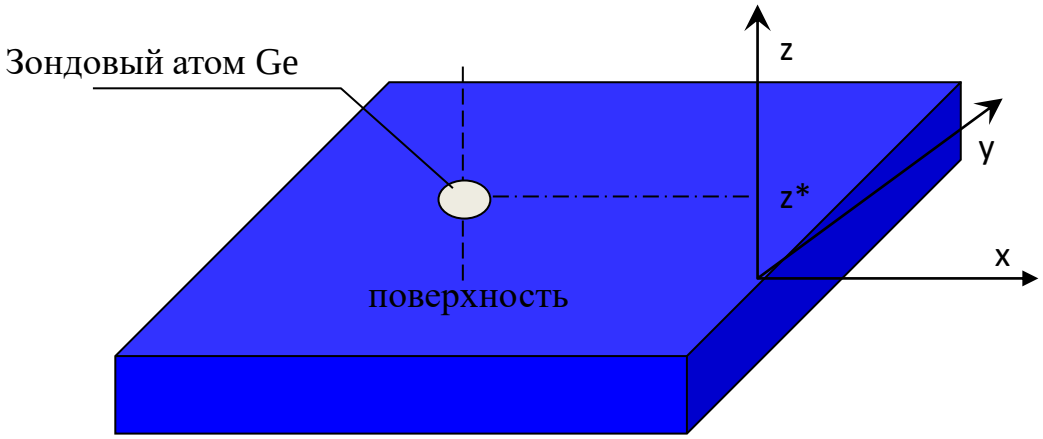
Моделирование проводилось на вычислительном узле с параметрами: **8-ядерный процессор Intel Core i7-9700K с тактовой частотой 3.60 ГГц, 32 ГБ ОЗУ**

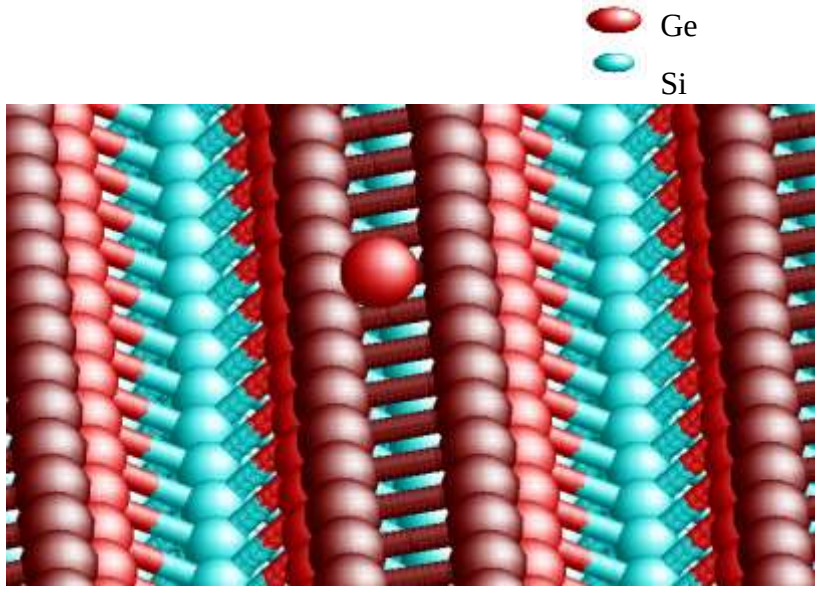


Моделирование осаждения 4 **МС (~0,54 нм)** Ge
на подложки Si различных размеров при **500ºС**

$K_n = T_1/T_n$, где T_n – время выполнения программы при n потоках.

Расчет потенциального рельефа
Метод молекулярной динамики





- Зондовый атом Ge помещается в точку (x,y) над поверхностью
- Зондовый атом имеет только одну степень свободы вдоль z
- Имитируется релаксация системы, зондовый атом находит равновесное положение z*
- Вычисляется энергия зондового атома U(z*) в точке (x,y,z*)
- Потенциальный рельеф строится путем сканирования по области x-y

Расчет методом молекулярной динамики потенциального рельефа структурированных подложек (расчеты методом молекулярной динамики)

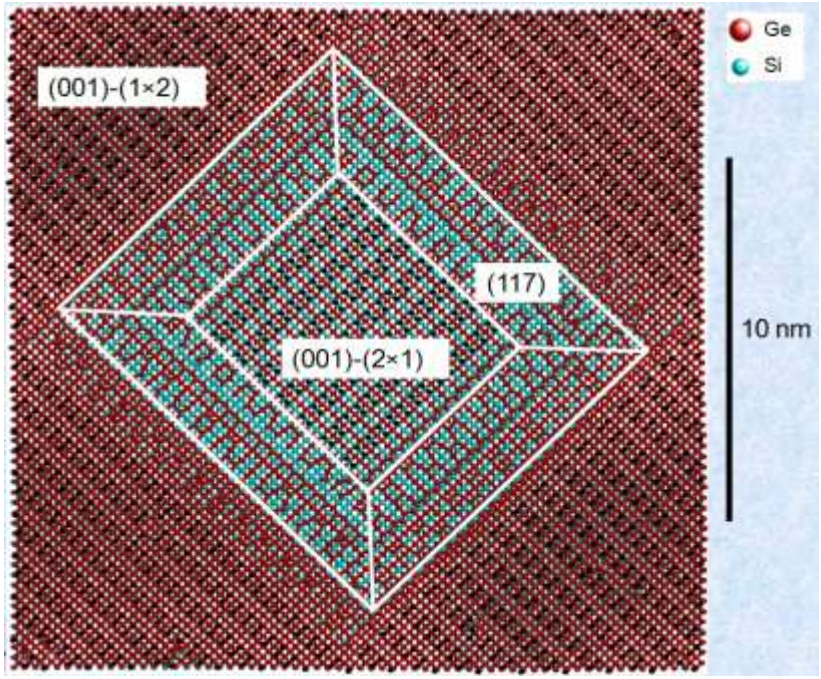


Рис. 1. Моделируемая структура с ямкой в форме перевернутой усеченной пирамиды (вид сверху). Атомы Si показаны голубым цветом, Ge - красным, Ge в димерах - черным. Белыми контурами отмечены грани усеченной пирамиды.

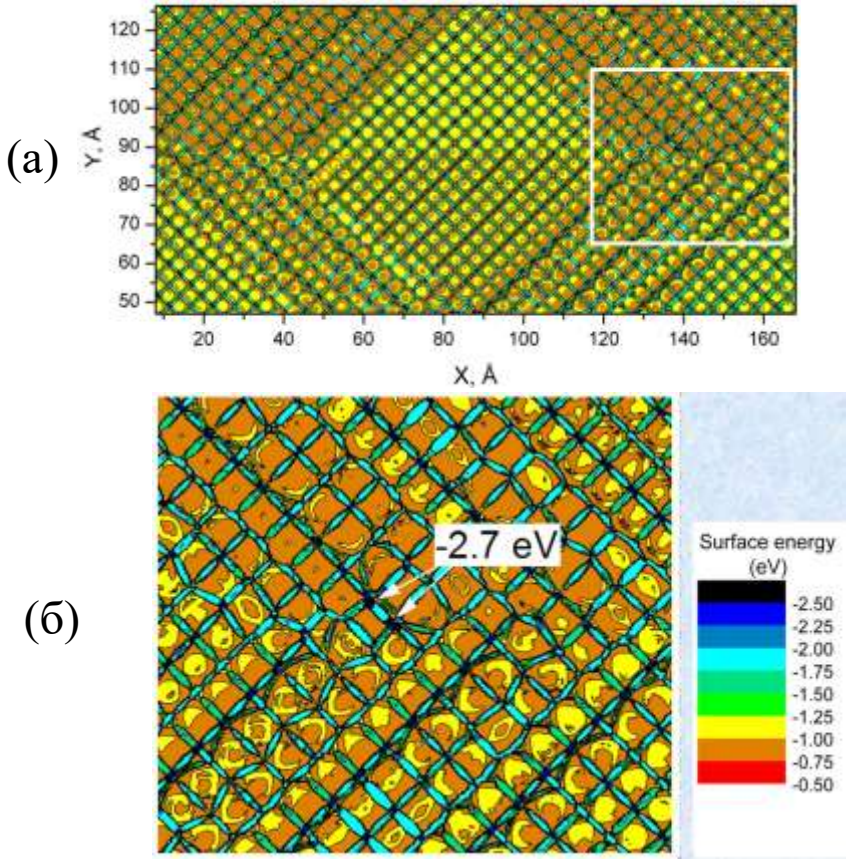


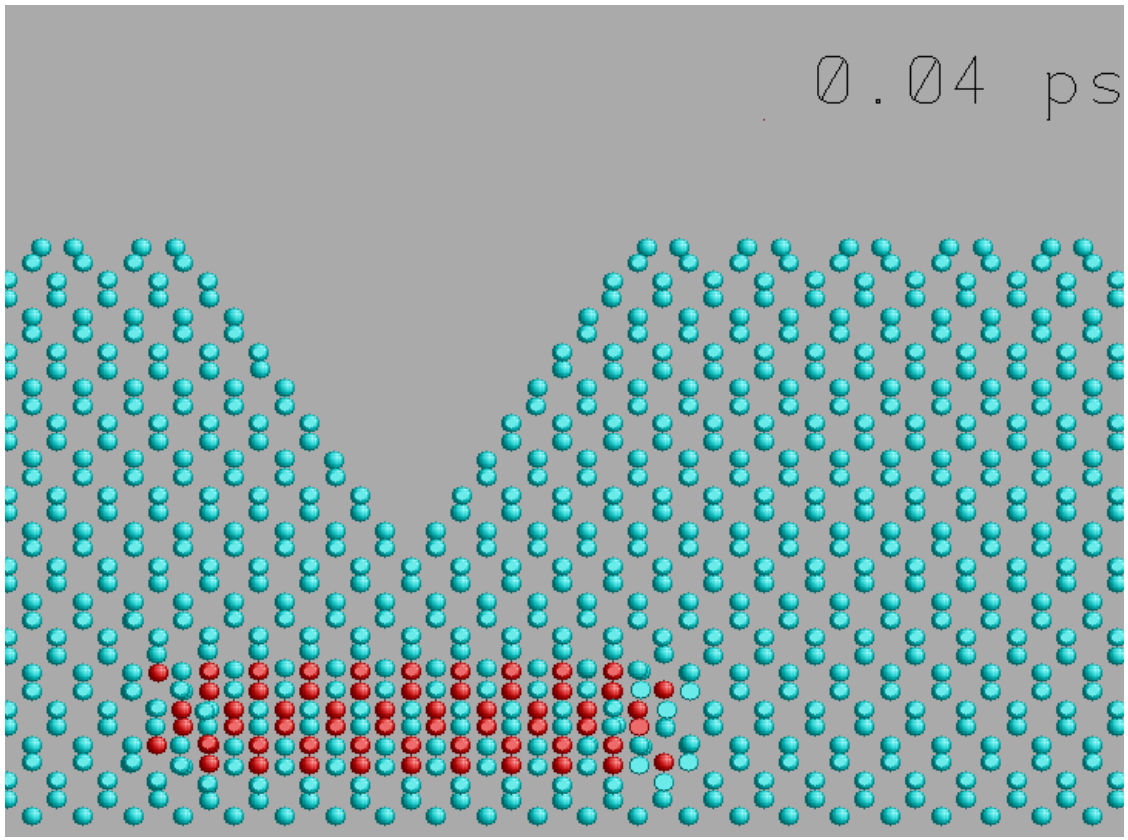
Рис.2. (а) Энергетическая поверхность моделируемой структуры. (б) увеличенный фрагмент энергетической поверхности в области, отмеченной белым контуром на рис. 2 (а).

Зондовый атом приближается к подложке на расстояние характерного межатомного взаимодействия и для полной системы решается уравнение движения на интервале времени $2 \cdot 10^{-12}$ с. Поверхностная энергия определяется как энергия зондового атома в конечном (равновесном) положении.

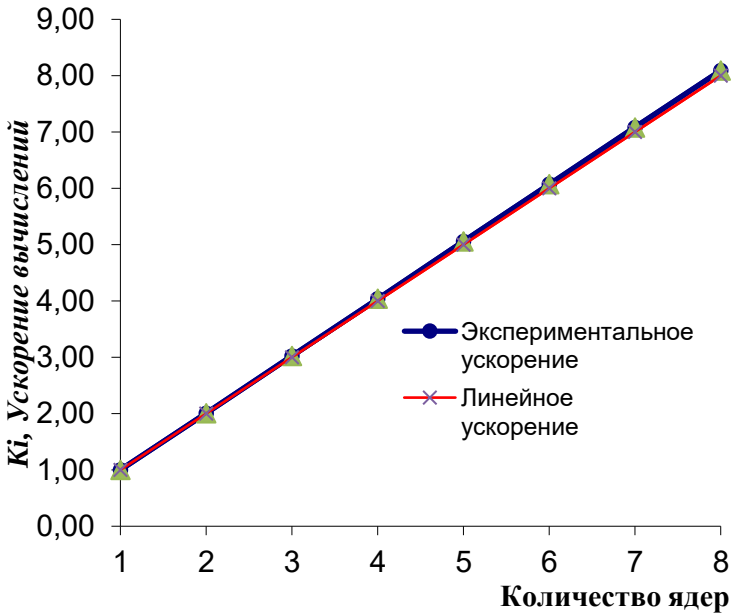
Энергетическая поверхность строится с помощью зондового атома сканированием поверхности структуры в пределах исследуемой области.

Изменение деформаций в структурированной подложке Si, обусловленные введенными междоузельными атомами Ge (расчеты методом молекулярной динамики)

А.В. Двуреченский , П.Л. Новиков, К.В. Павский, Ж.В. Смагина



Поперечное сечение канавки в моделируемой структуре.



Ускорение параллельной программы.

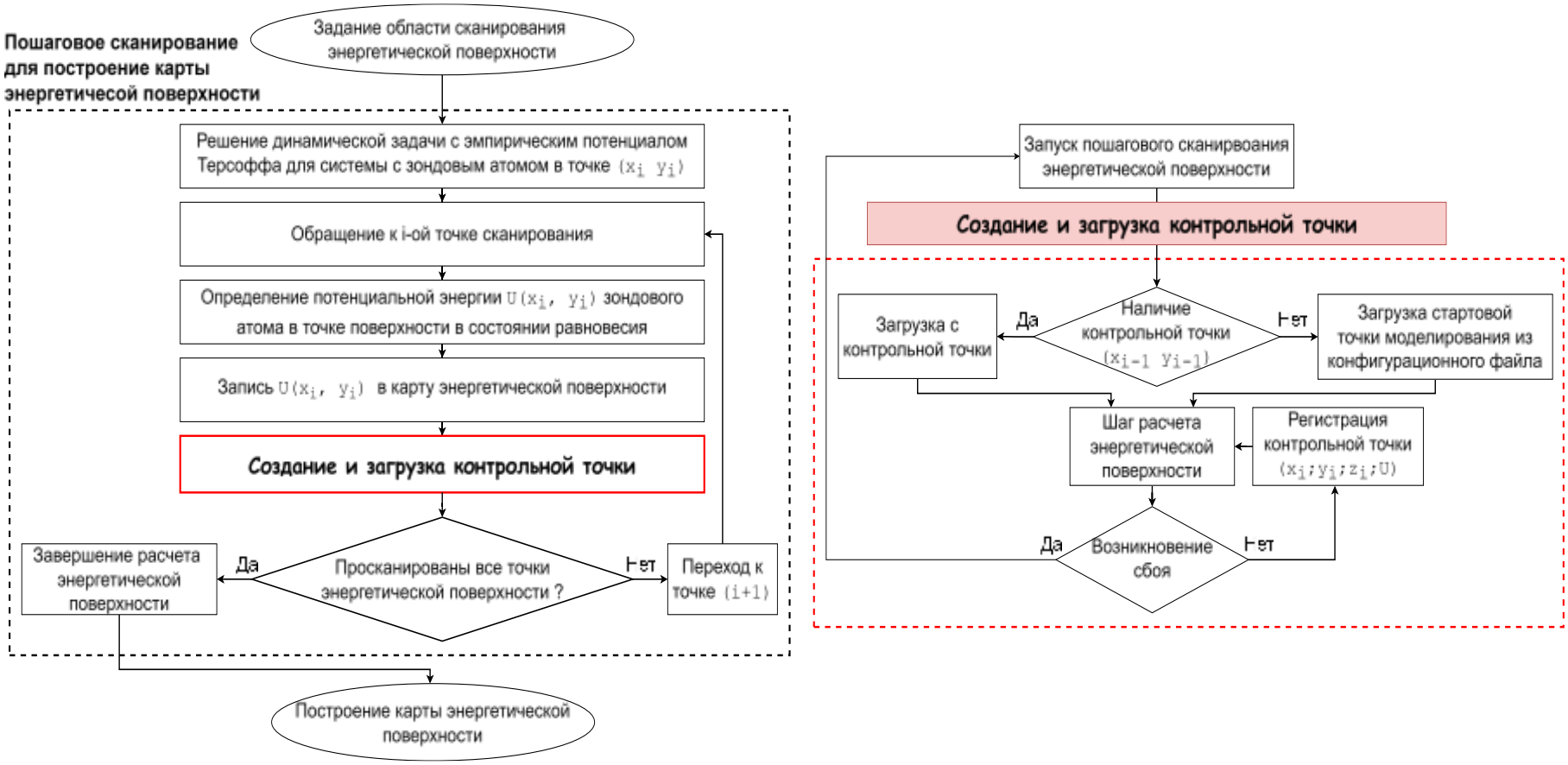
$K_i = T_1/T_i$

T_i – время исполнения программы на i ядрах.

Ориентация стенок канавки	Деформация (%) Концентрация межузельных атомов		Вклад межузельных атомов (%)
	10^{20} см$^{-3}$	0	
(111)	-5.35	-0.74	-4.61

Novikov P.L., Atovullaev T., SmaginaZh.V., Dvurechenskii A.V., Pavskii K.V. Effect of Interstitials Embedded in Pre-Patterned Si Substrate on Location of GeNanoislands // Physica Status Solidi C: Current Topics in Solid State Physics, 2017,14, 1700200(1-4).

Программная реализация отказоустойчивого расчета потенциального рельефа



СПАСИБО ЗА ВНИМАНИЕ