

*Институт математического моделирования  
Российской академии наук*

# ***Балансировка загрузки процессоров***

***М.В.Якобовский***

*mail: [lira@imamod.ru](mailto:lira@imamod.ru)*

*web: <http://lira.imamod.ru>*

*Нижний Новгород*

*2009*

# Задачи большого вызова

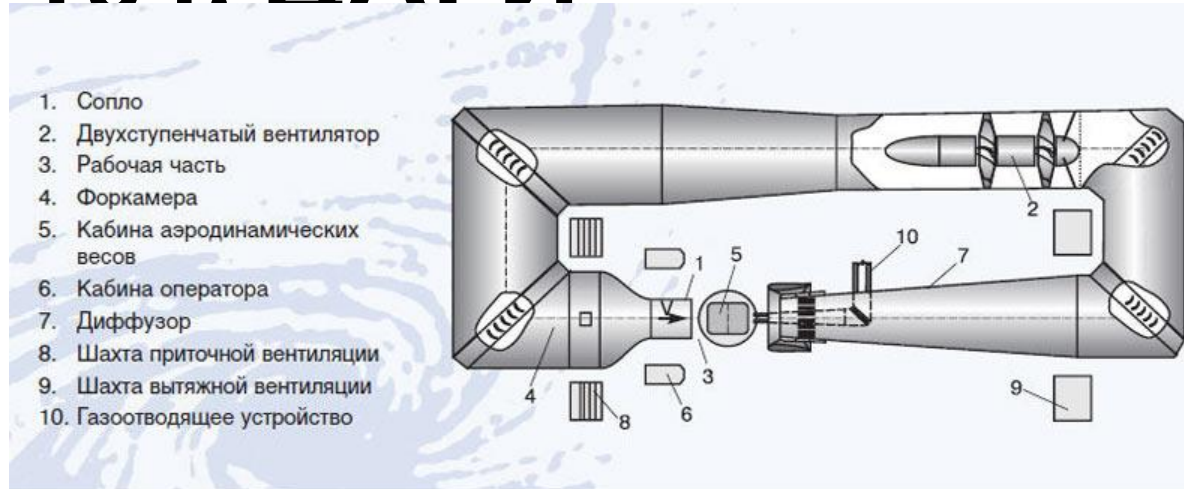
(Kenneth G. Wilson, Cornell University, 1987)

- Вычислительная газовая динамика:
  - Создание летательных аппаратов, эффективных автомобилей
  - Предсказание погоды, и глобальных климатических изменений
  - Оптимизация нефтедобычи, ...
- Молекулярная динамика:
  - Создание материалов с заданными свойствами
  - Разработка новых лекарственных соединений
  - Сверхпроводимость, Свойства веществ в экстремальных состояниях, ...
- Символьные вычисления
  - Распознавание речи
  - Компьютерное зрение
  - Изучение сложных систем
  - Автономные системы управления
- Квантовая хромодинамика и теория конденсированных сред
- Управляемый термоядерный синтез, Геном человека, ...

# Дозвуковая аэродинамическая труба Т-104. ЦАГИ

- Скорость потока **10–120 м/с**
- Диаметр сопла 7 м
- Длина рабочей части 13 м
- Мощность вентилятора **28.4 МВт**

<http://www.tsagi.ru/rus/base/t104>



## Суперкомпьютер СКИФ МГУ «ЧЕБЫШЁВ»

- Пиковая производительность 60 TFlop/s
- Мощность комплекса **0.72 МВт**

[http://parallel.ru/cluster/skif\\_msu.html](http://parallel.ru/cluster/skif_msu.html)



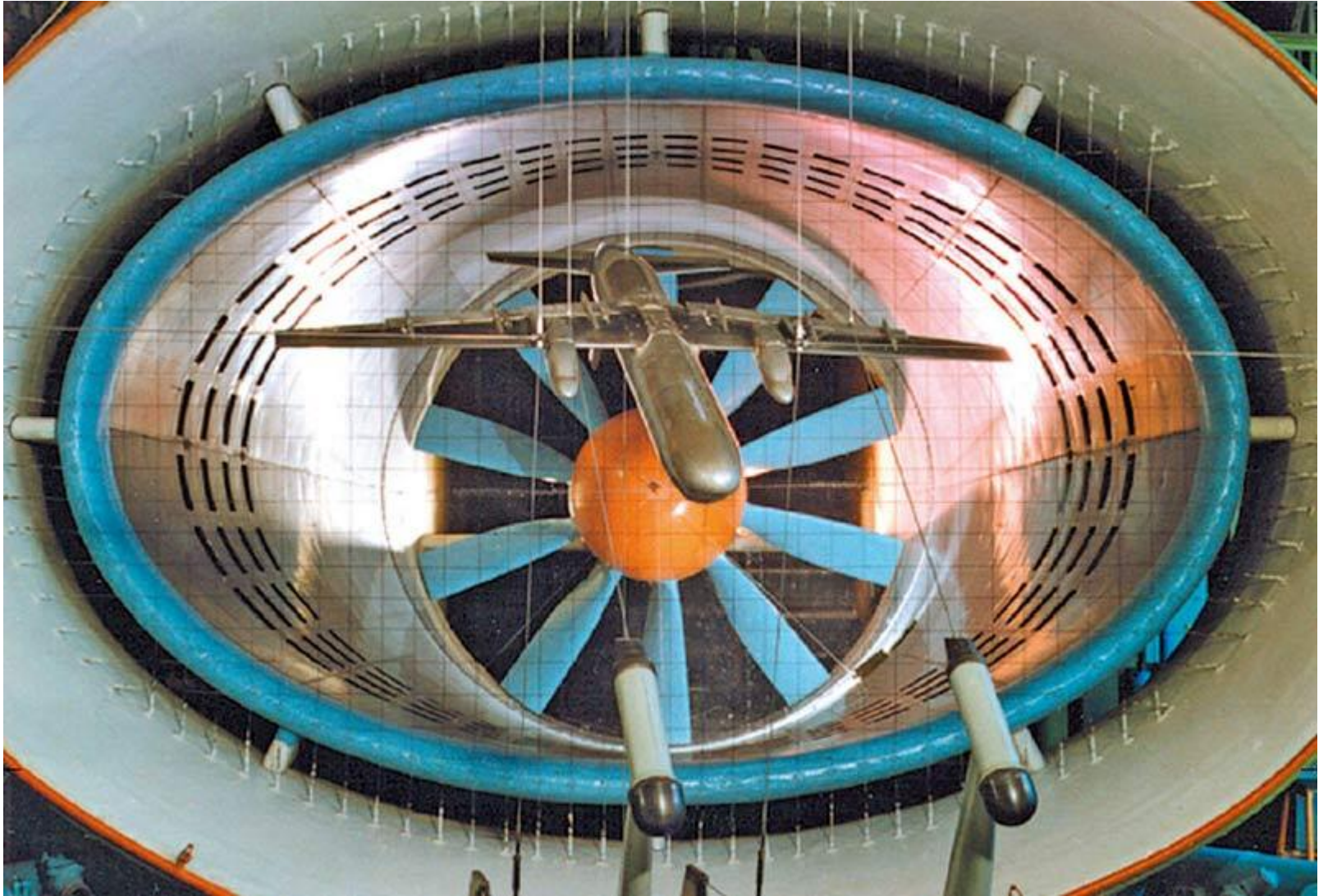












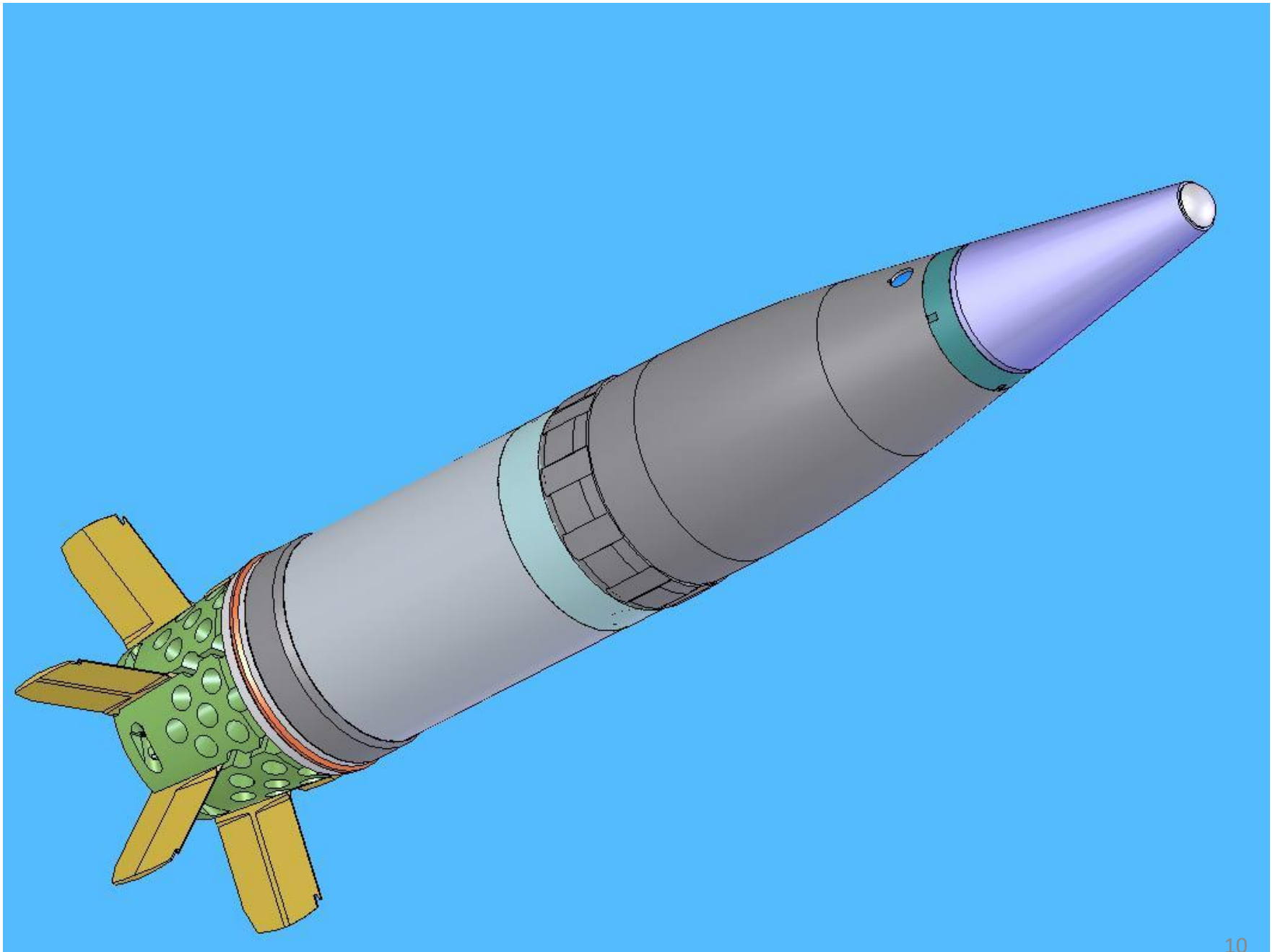
# Суперкомпьютеры

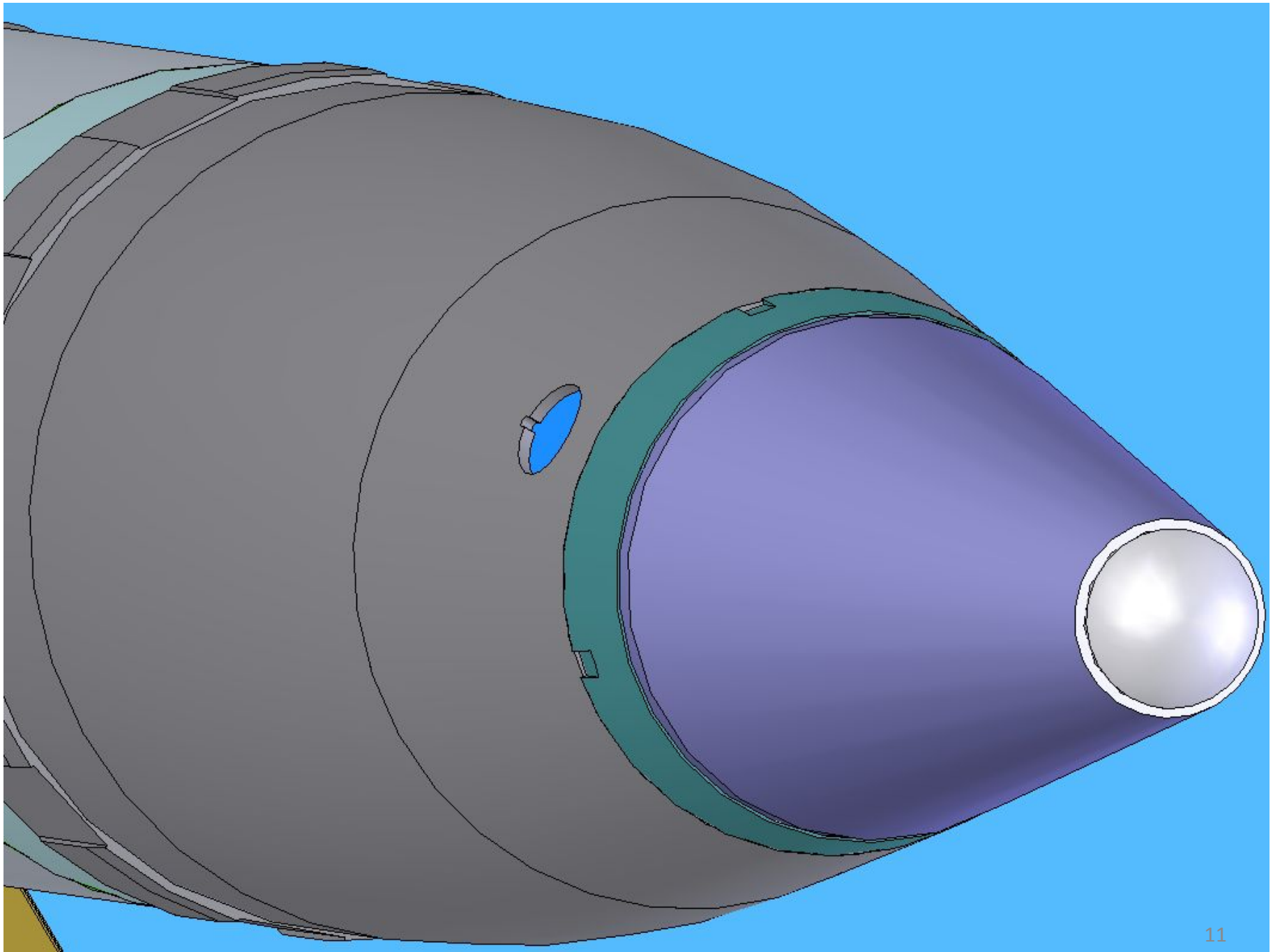
- Не просто составляют конкуренцию натурному эксперименту, но:
  - Необходимы для его проведения
  - Позволяют делать то, что натурный эксперимент делать не позволяет

# Суперкомпьютеры

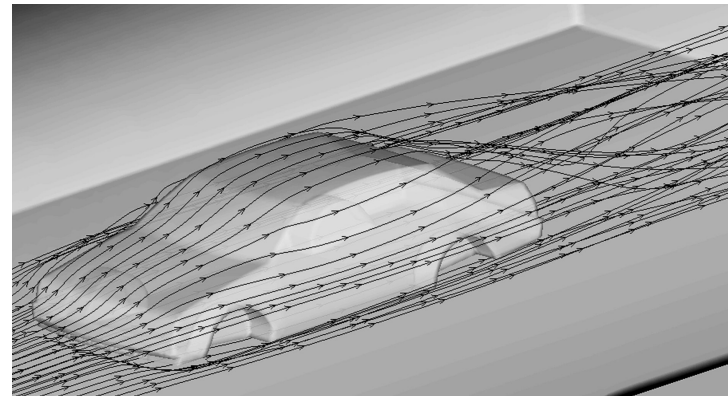
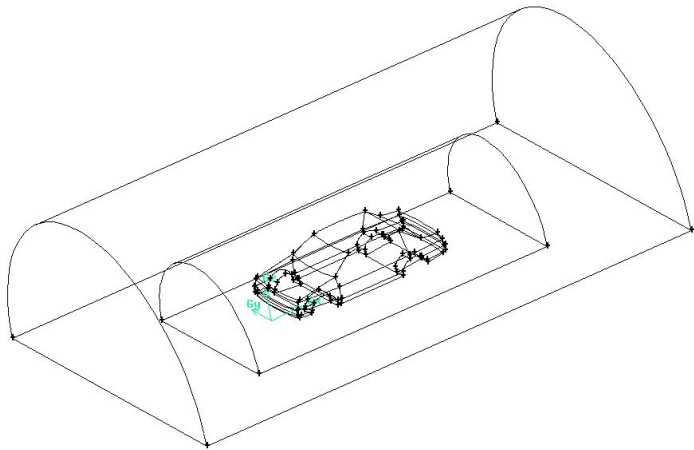
- Используются неэффективно и далеко не в полной мере
- Необходимы:
  - Вычислительное ядро: адаптация алгоритмов к архитектуре многопроцессорных систем с распределённой памятью
  - Специальное математическое обеспечение: визуализация, генерация сеток, рациональное разбиение на подобласти, динамическая балансировка загрузки процессоров, использование CAD-технологий, использование гетерогенных систем и GRID-технологий
  - Интеграция в единый программный комплекс







# НЕВЯЗКОЕ ОБТЕКАНИЕ КУЗОВА АВТОМОБИЛЯ ( $M = 0.12$ )



**СЕТКА: 430 949 УЗЛОВ, 2 430 306 ТЕТРАЭДРОВ**

# НЕВЯЗКОЕ ОБТЕКАНИЕ КУЗОВА АВТОМОБИЛЯ

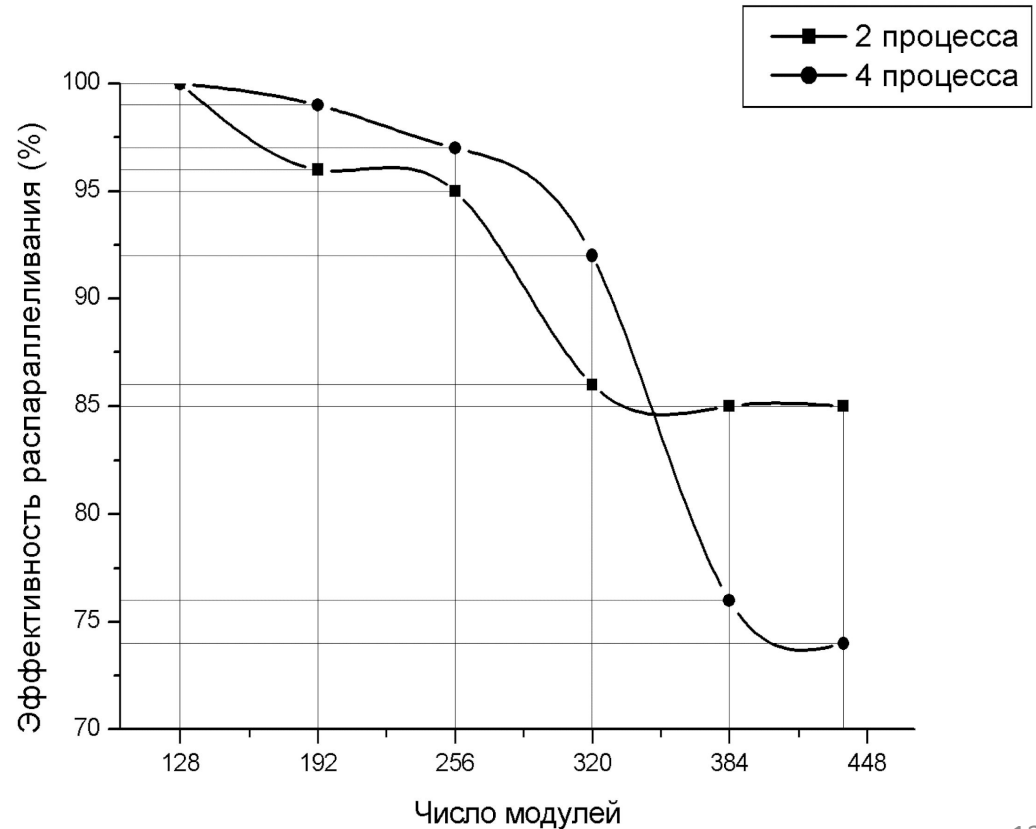
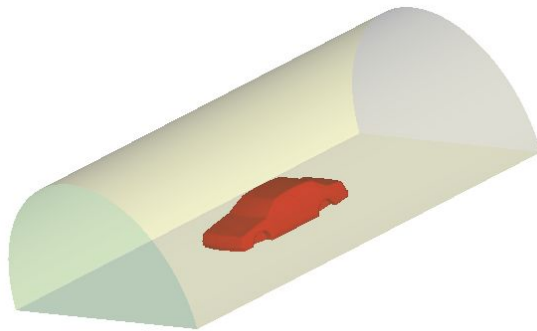


Сетка: 209 028 730 узлов, 1 244 316 672 тетраэдра (24 Гб)

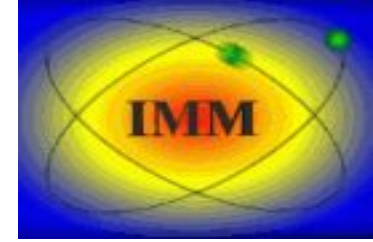
МВС: МВС-100К

1. Запуск задачи на 128, 192, 256, 320, 384 и 437 модулях с порождением 2 и 4 параллельных MPI процессов (до 1748 параллельных процессов).

2. Запуск задачи на 437 модулях в рамках гибридной модели параллелизма MPI + OpenMP (3496 параллельных процессов)



# Суперкомпьютеры



МСЦ РАН: процессор: Intel(R) Xeon(R) CPU X5365 @ 3.00GHz  
ядер на узел: 8  
память узла: 4/8 Гб  
число узлов: 782 (6256 ядер)  
коммуникации: InfiniBand DDR  
производительность: 75 TFLOPS

СКИФ МГУ: процессор: Intel(R) Xeon(R) CPU E5472 @ 3.00GHz  
ядер на узел: 8  
память узла: 8 Гб  
число узлов: 630 (5040 ядер)  
коммуникации: InfiniBand DDR  
производительность: 60 TFLOPS

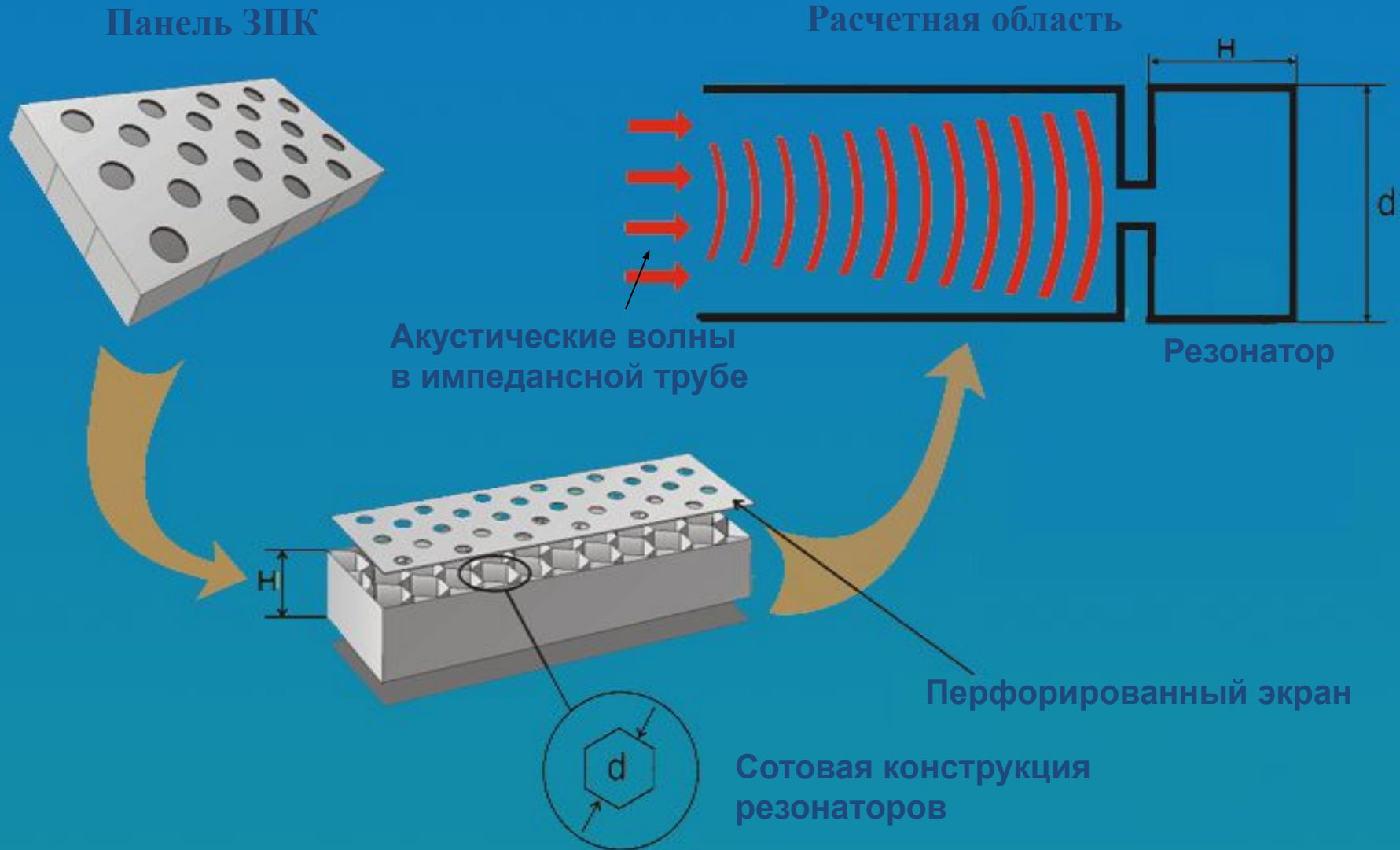


# Акустика

## Вычислительные эксперименты по ЗПК

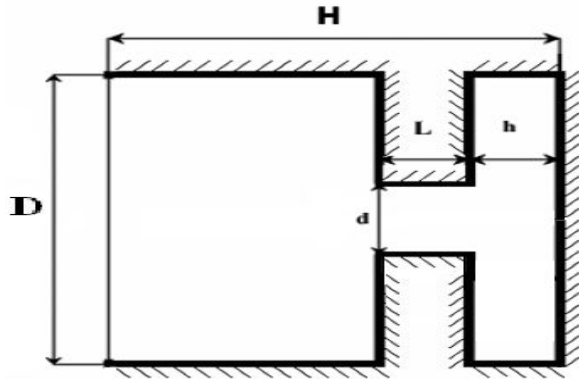


# Звукопоглощающие конструкции

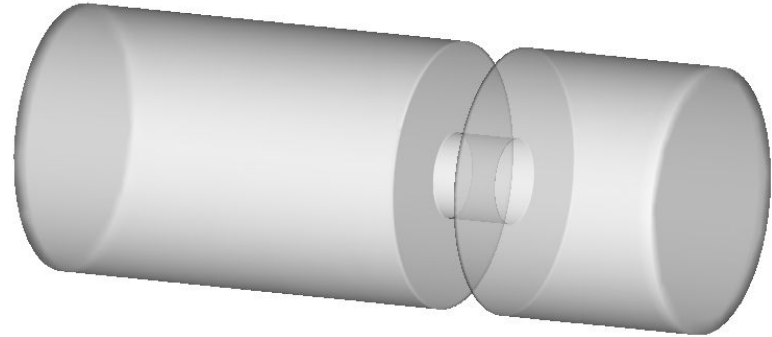


# Эксперимент 1: Модель 2D и 3D импедансной трубы

## 2D задача

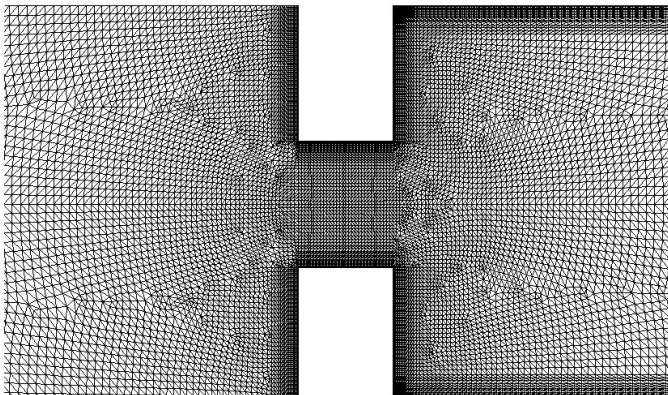


## 3D задача

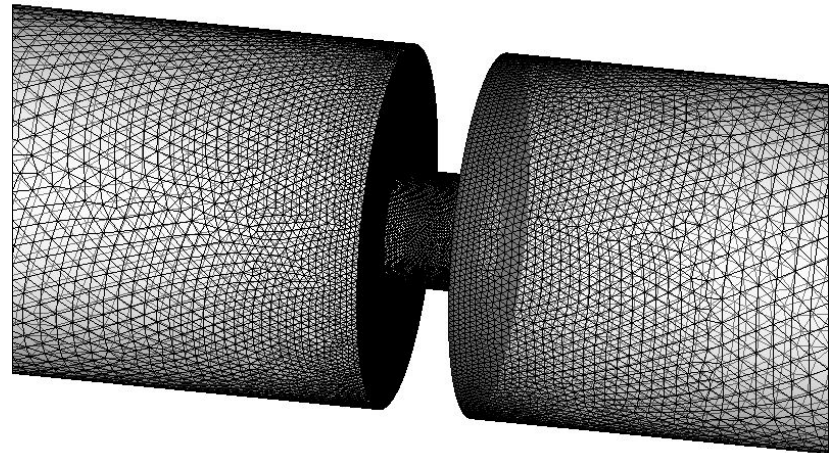


## Концентрация сетки около горла резонатора

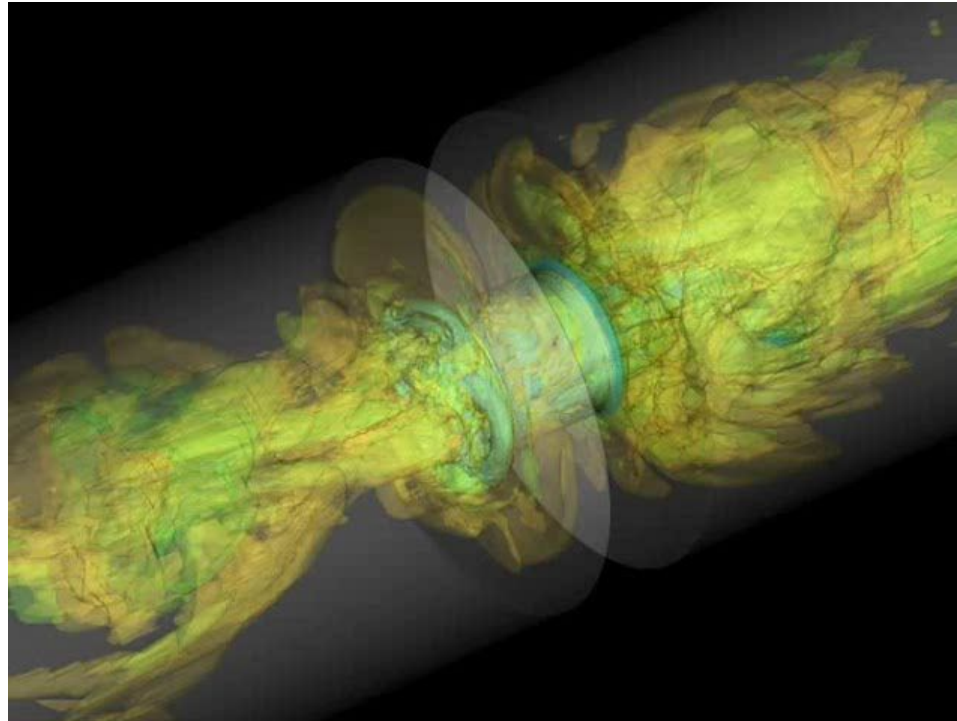
Размер сетки до 90К узлов



Размер сетки до 1М узлов



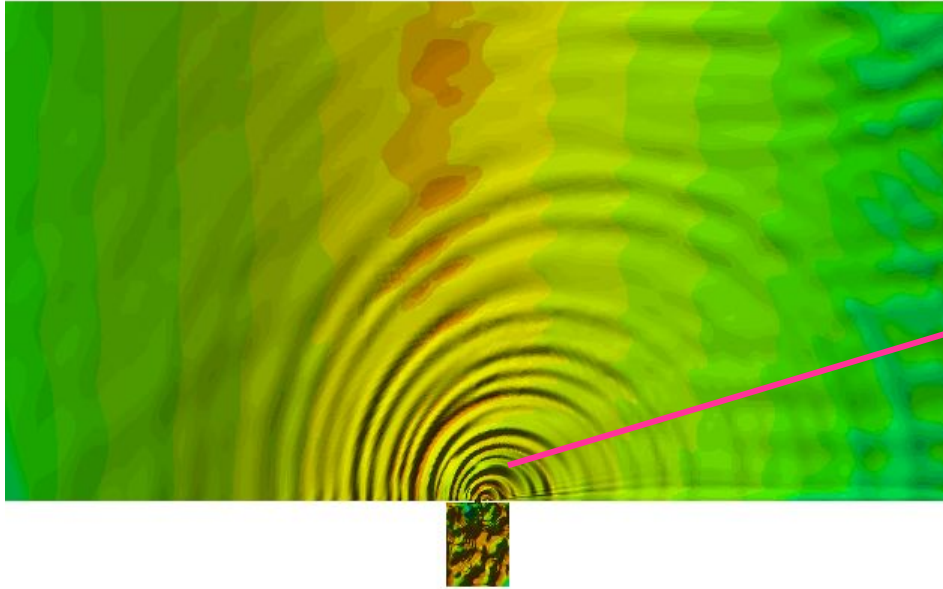
## Течение в отверстии резонаторной камеры



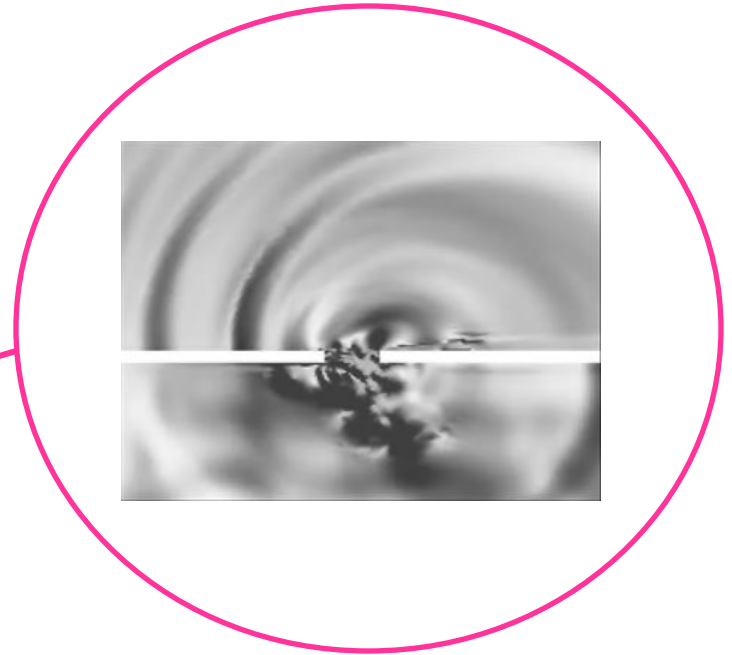


## Эксперимент 2: 2D канал с резонаторами (2/2)

Эффект свиста



Слой смешения



Возмущения плотности



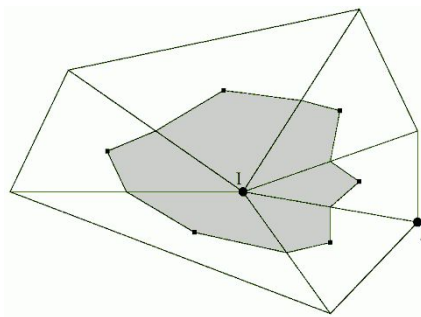
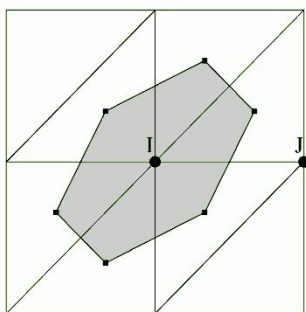
# Базовая численная схема (1/2)

## 2D контрольные объемы

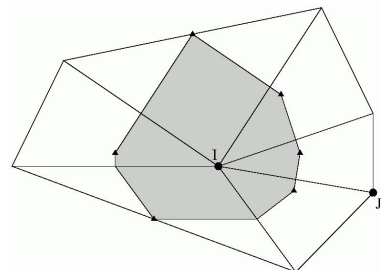
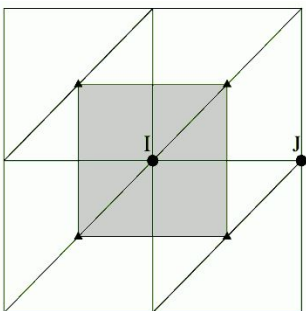
Декартова сетка

Неструктурированная  
треугольная сетка

Медианные ячейки



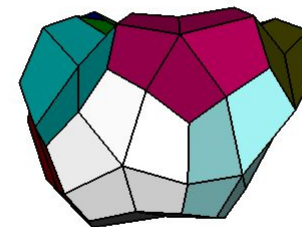
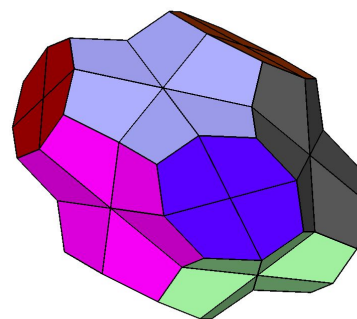
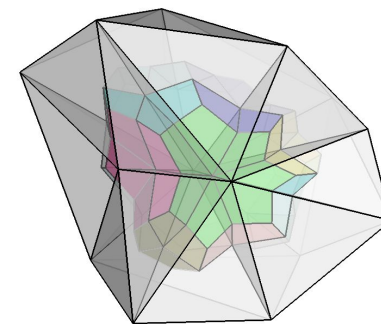
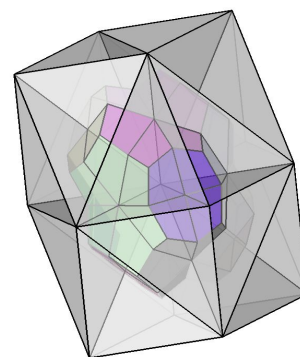
Ячейки на центрах описанных окружностей



## 3D контрольные объемы

Декартова сетка

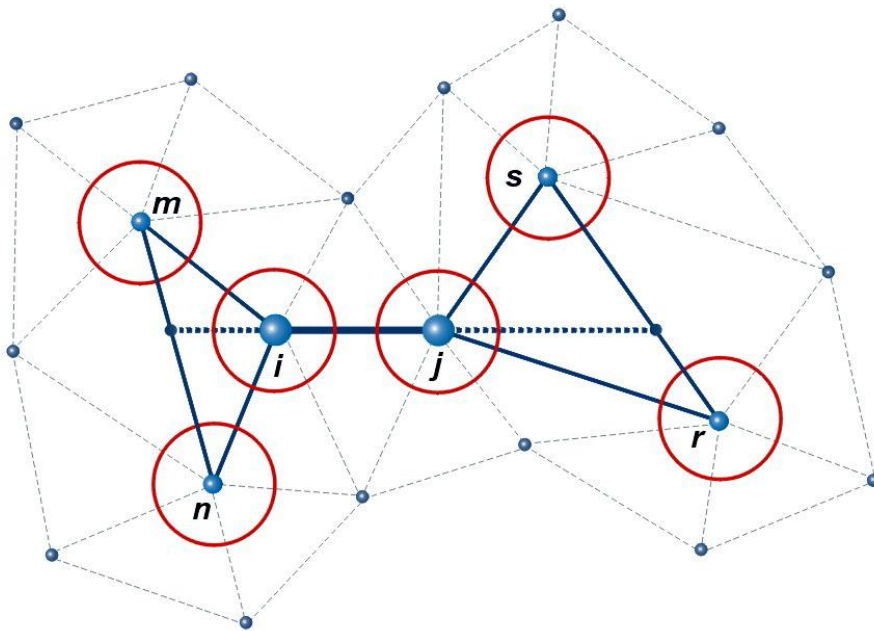
Неструктурированная  
тетраэдральная сетка



# Базовая численная схема

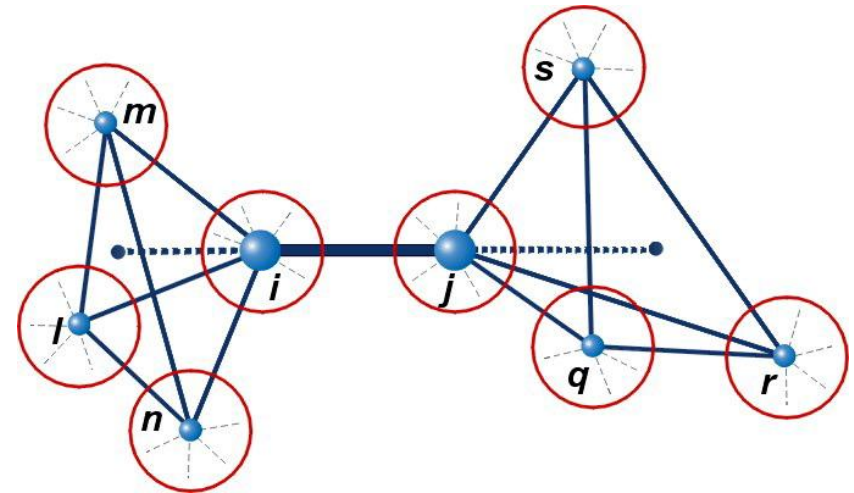
Пространственный шаблон для определения потока между узлами I и J  
(сложность для распараллеливания)

2D треугольная сетка



2D шаблон высокого порядка:  
Противопоточные треугольники + соседи

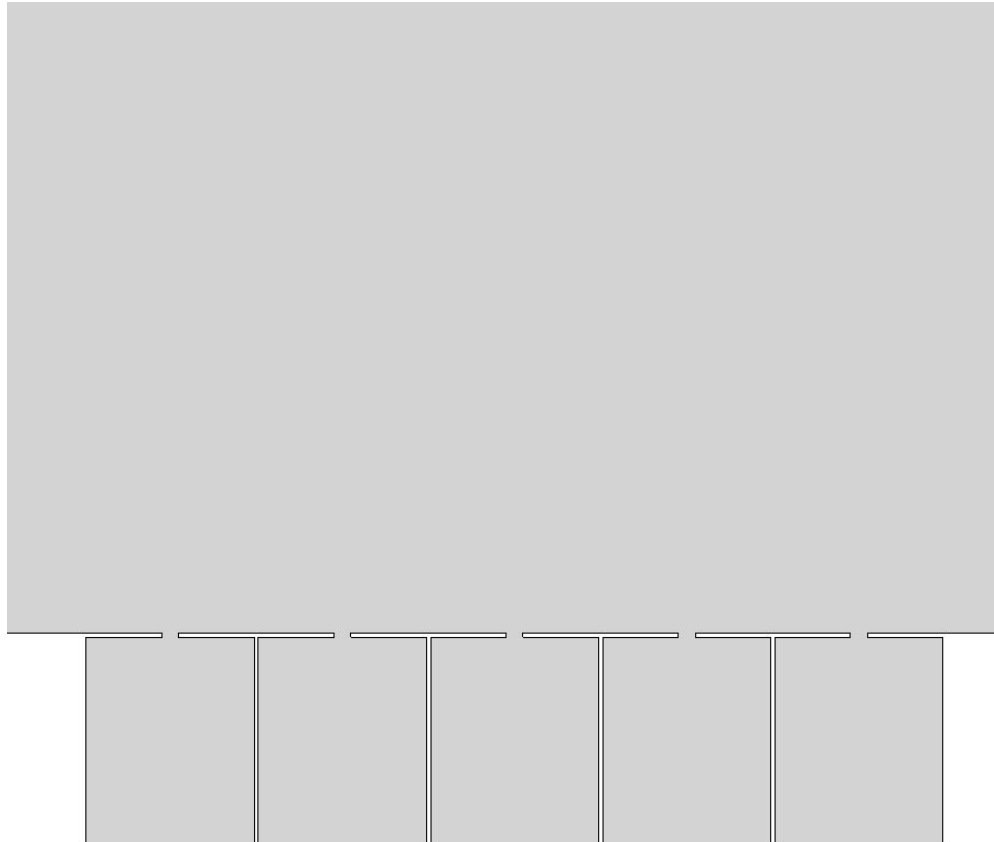
3D тетраэдральная сетка



3D шаблон высокого порядка:  
Противопоточные тетраэдры + соседи

# Канал с 5 резонаторами

Применимость не только суперкомпьютеров,  
но и Grid технологий



Возмущения плотности

Уравнения Эйлера, нет погранслоя,  $M=0.4$

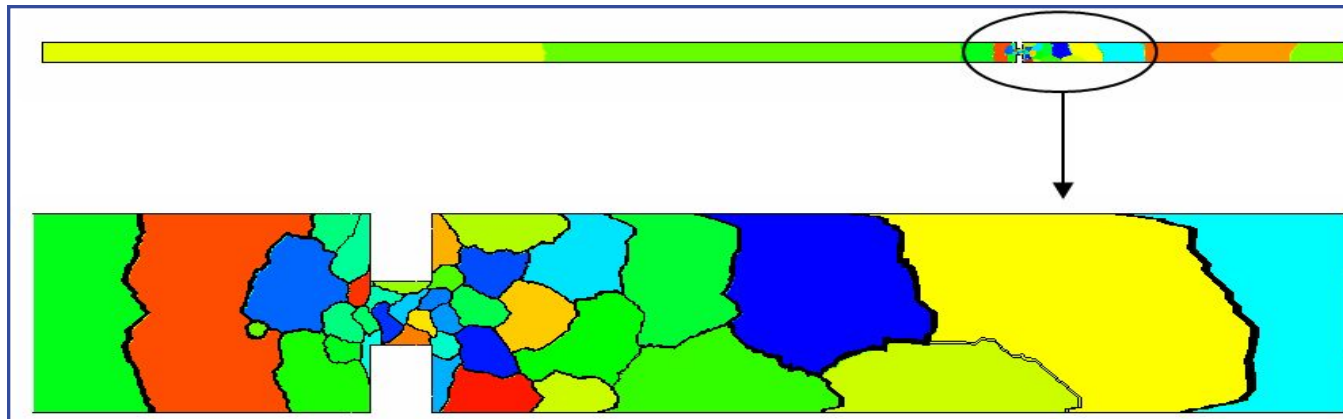
# Производительность вычислений

Две различные параллельные системы использовались для тестов

- 1) Типичный малобюджетный кластер с обычной сетью Ethernet  
Узел: 2CPU Intel Xeon 3GHz  
Сеть: Ethernet 1Gbit
- 2) “Продвинутый” кластер с высокопроизводительной сетью низкой латентности  
Узел: 2CPU AMD Opteron 2.4Hz  
Сеть: Myrinet

Эти системы имеют существенно различное отношение производительности процессора и сети

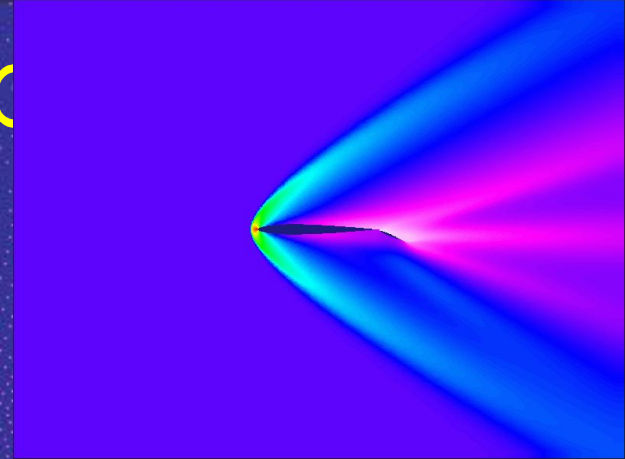
**Тестовая задача:** Модельная 2D задача – импедансная труба.  
Размер сетки 80 000 узлов, схема 5-го порядка



Пример разбиения сетки



# Статическая балансировка загрузки



$$G = (V, E), \quad V = \{v_i\}, \quad |V| = n$$

$$R(V) = (V_1, \dots, V_p)$$

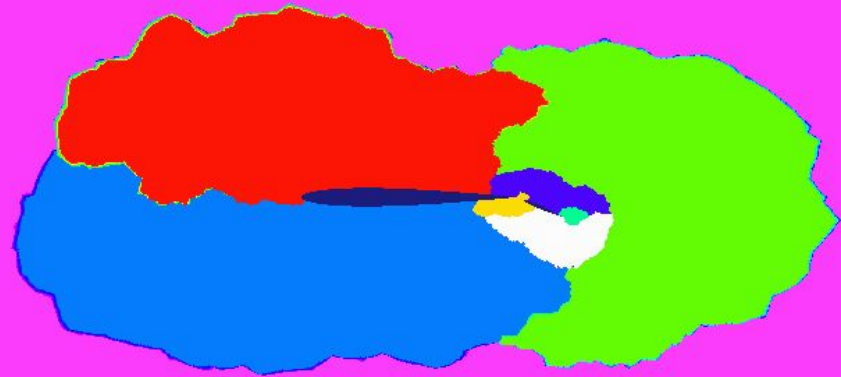
$$V = \bigcup_{k=1}^p V_k, \quad V_i \cap V_j = \emptyset, \quad i \neq j$$

$$\min_{R(V)} \left\{ J = \max_{k=1, \dots, p} \sum_{v_i \in V_k} \left( w(v_i) + \alpha \sum_{v_j \notin V_k} w(v_i, v_j) \right) \right\}$$



# Критерии декомпозиции графов

- Равномерное распределение суммарного веса узлов/рёбер
- Минимизация максимального веса исходящих из домена ребер
- Минимизация суммарного веса разрезанных ребер
- Минимизация максимальной степени доменов
- Обеспечение связности доменов
- Обеспечение связности множества внутренних узлов доменов



А.Н. Андрианов, А.В. Жохова, Б.Н. Четверушкин

Процессоров	11	31	47	63
New_sort	13.59	5.59	4.38	4.16
METIS	13.61	11.00	11.10	10.56

Чему равно  $25/4$  ?

6.25

$$25/4=$$

~~6.25~~

$$25/4=$$

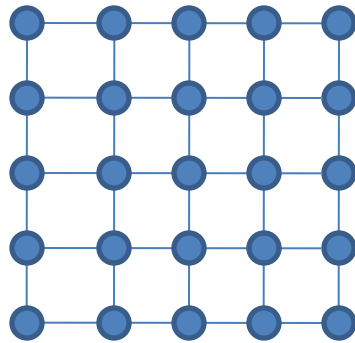
6 ~~6.25~~ 9

4

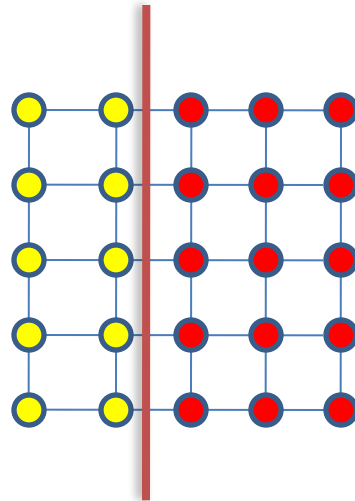


$$25/4 = 4 ? 6 ? 9$$

- Разрезать решетку 5 x 5 на 4 части

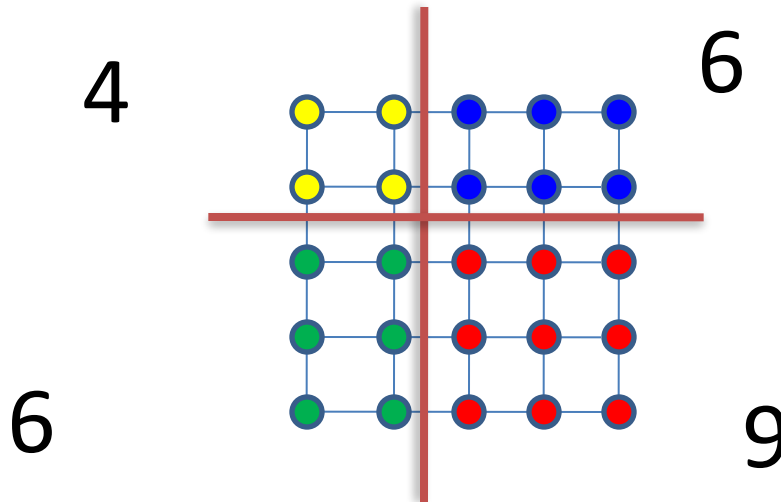


# Декомпозиция сетки из 25 узлов на 4 части



$$25/4 = 4 ? 6 ? 9$$

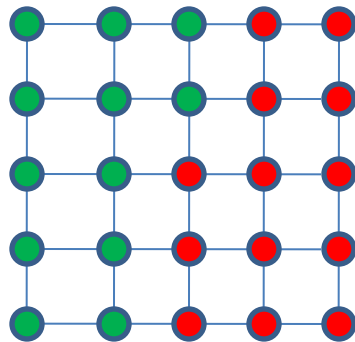
- Декомпозиция решетки 5 x 5 на 4 домена



- Дисбаланс  $9/4=2.25$

$$25/4 = 4 ? 6 ? 9$$

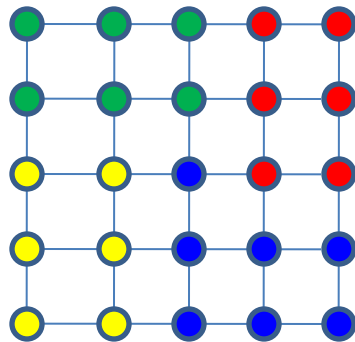
- Декомпозиция решетки 5 x 5 на 2 домена
- Дисбаланс 13/12 : 8%





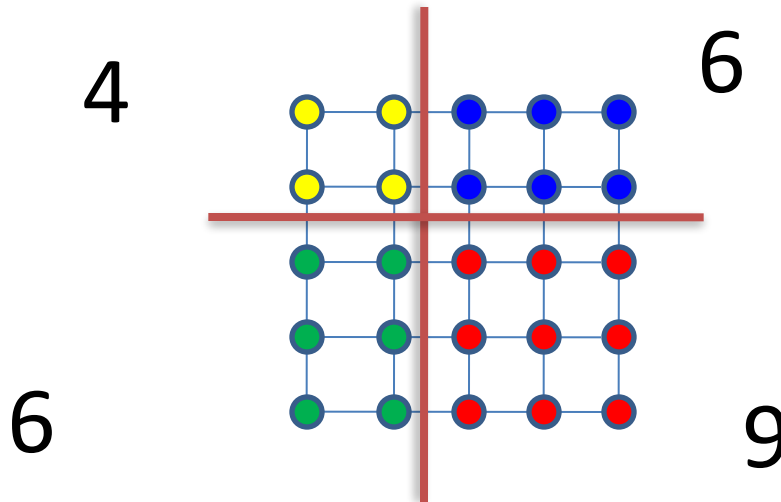
$$25/4 = 4 ? 6 ? 9$$

- Декомпозиция решетки 5 x 5 на 4 домена
- Дисбаланс 7/6 : 17%



$$25/4 = 4 ? 6 ? 9$$

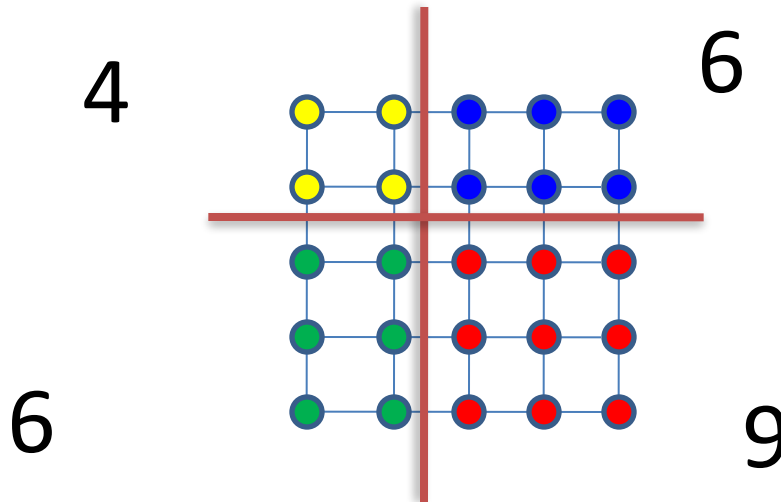
- Декомпозиция решетки 5 x 5 на 4 домена



- Дисбаланс  $9/4=2.25$

$$25/4 = 4 ? 6 ? 9$$

- Декомпозиция решетки 5 x 5 на 4 домена

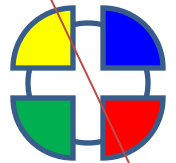
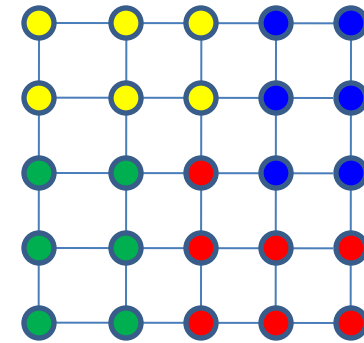
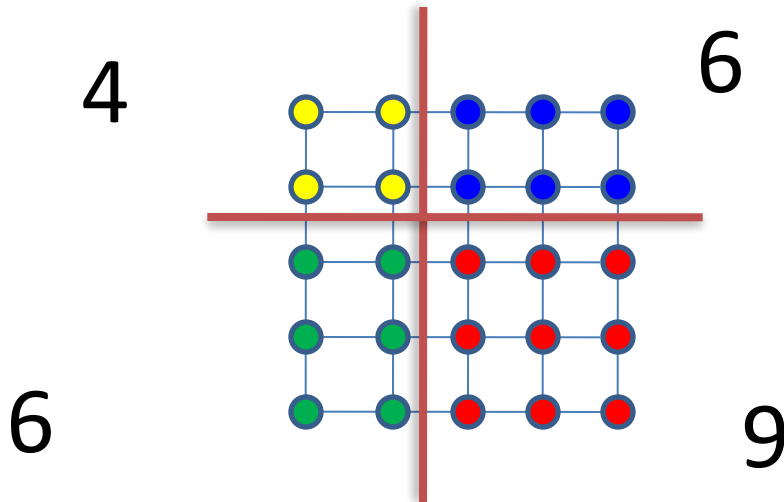


- ~~• Дисбаланс  $9/4=2.25$~~

$$25/4 = 4 ? 6 ? 9$$

- Декомпозиция решетки 5 x 5 на 4 домена

Потери  $9/6.25=1.44$

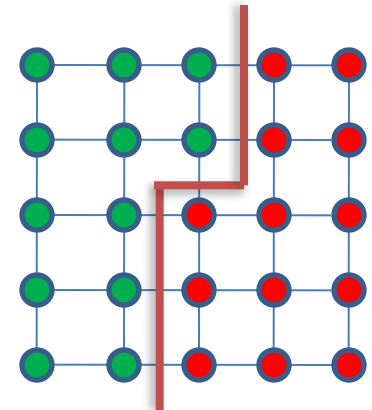
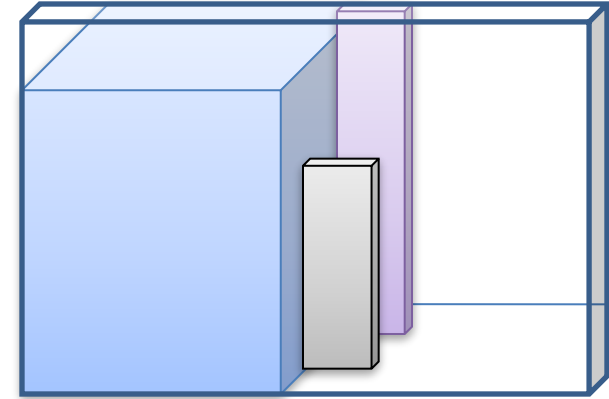
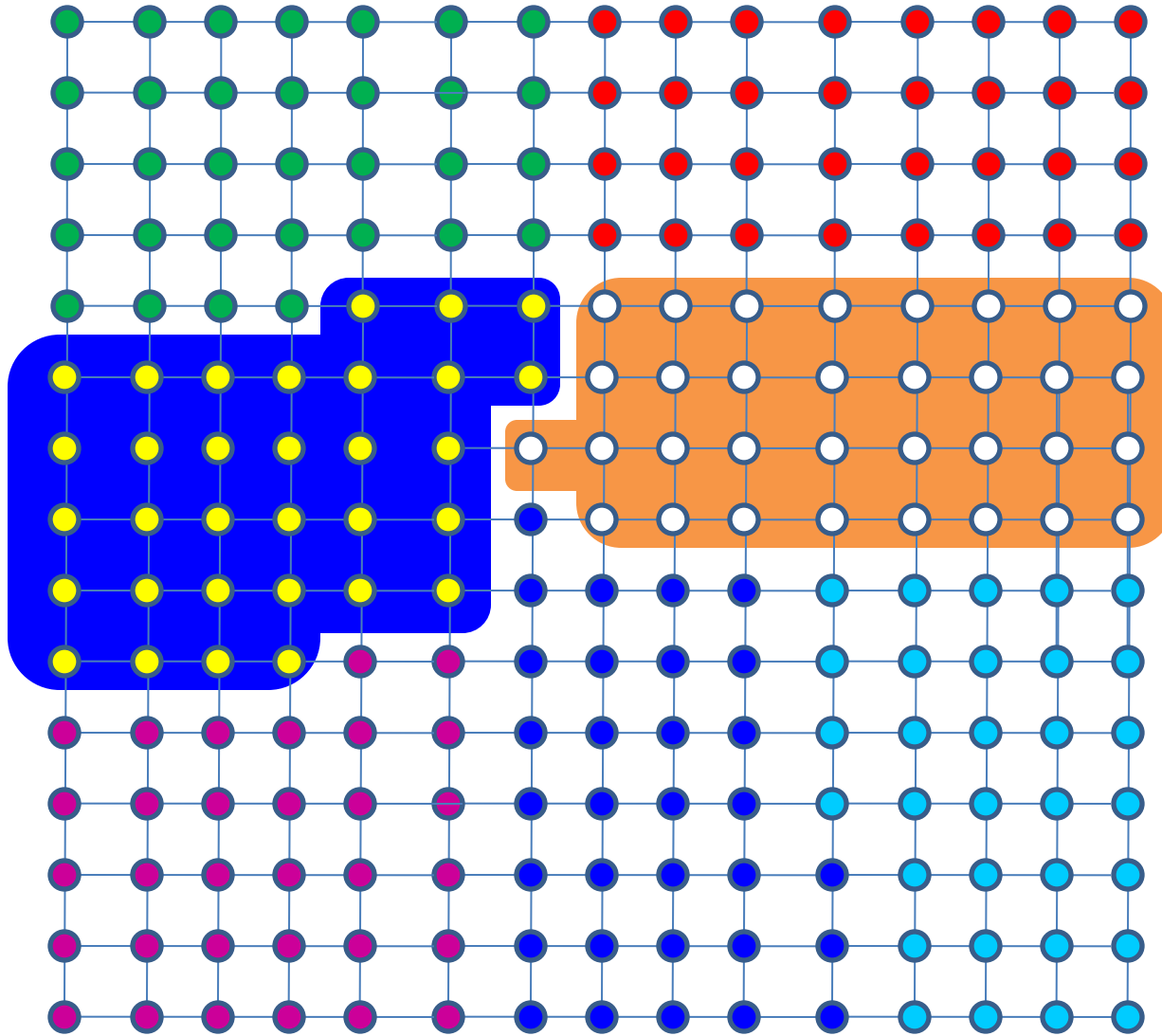


- ~~• Дисбаланс  $9/4=2.25$~~

**Потери**  
 $9/7=1.29$



# Декомпозиция сетки 25x25 на 7 частей



# Разбиение тетраэдральной сетки, содержащей $2 \cdot 10^8$ узлов, на 125 процессорах

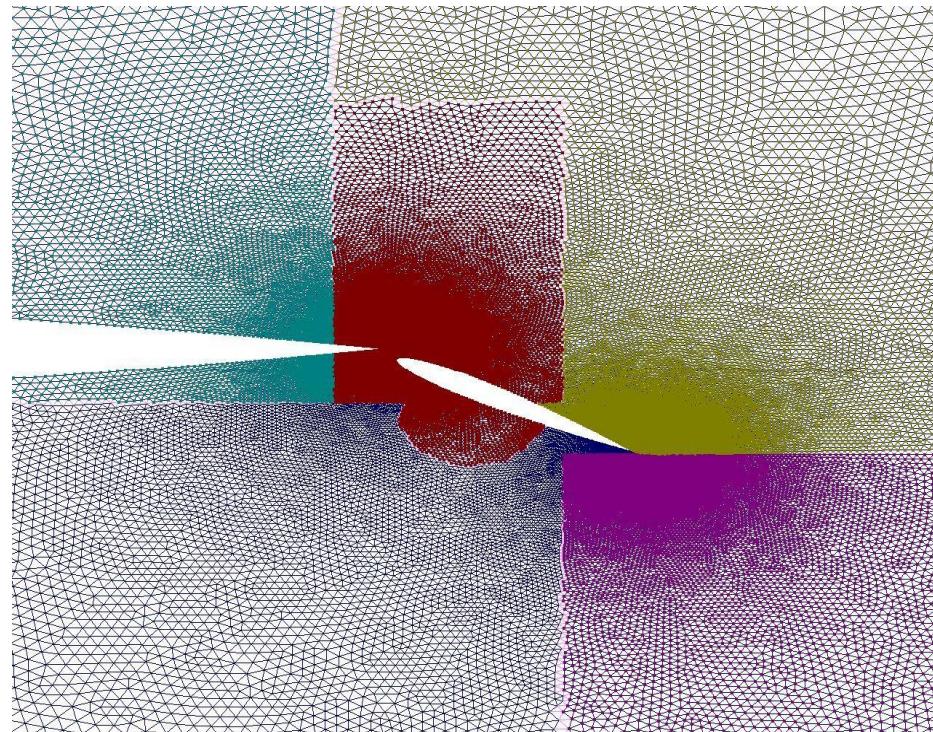
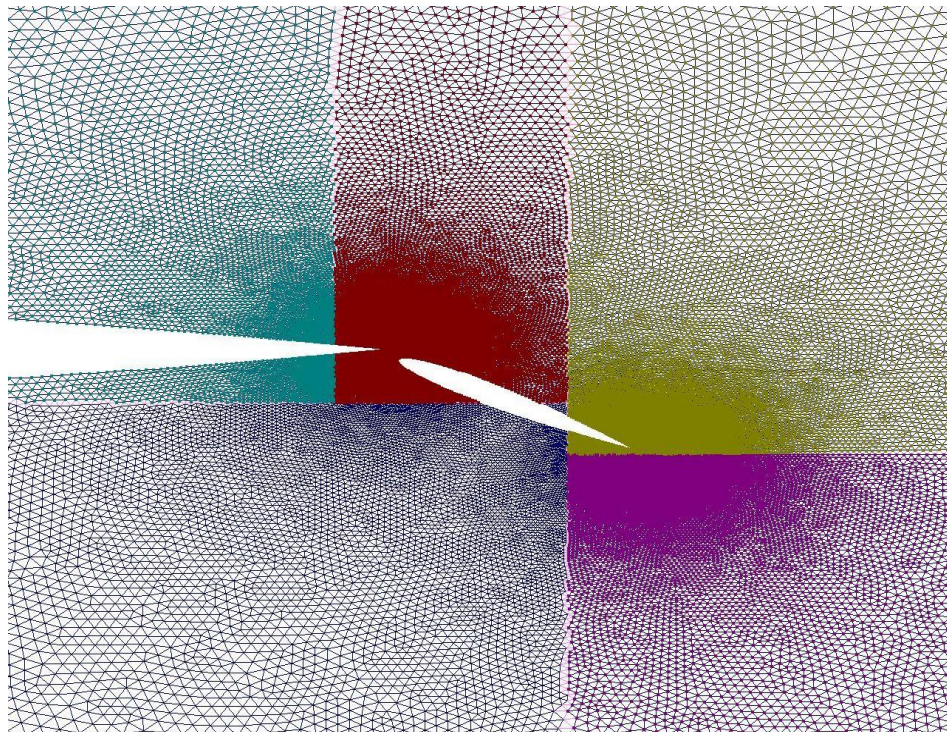
- вычисления производились на кластере СКИФ МГУ (1250 4-ядерных процессоров, 60 TFlop/s)

		геометрическая декомпозиция		ParMETIS	
число доменов		80 000		20 000	
время		21 сек.		10 сек.	
число вершин в домене		2612	2613	2 328	10 932
мин.	макс.				
среднее число связей с соседними доменами		14		14	
число некомпактных доменов		229		364	

# Фрагмент треугольной сетки из 75790 вершин

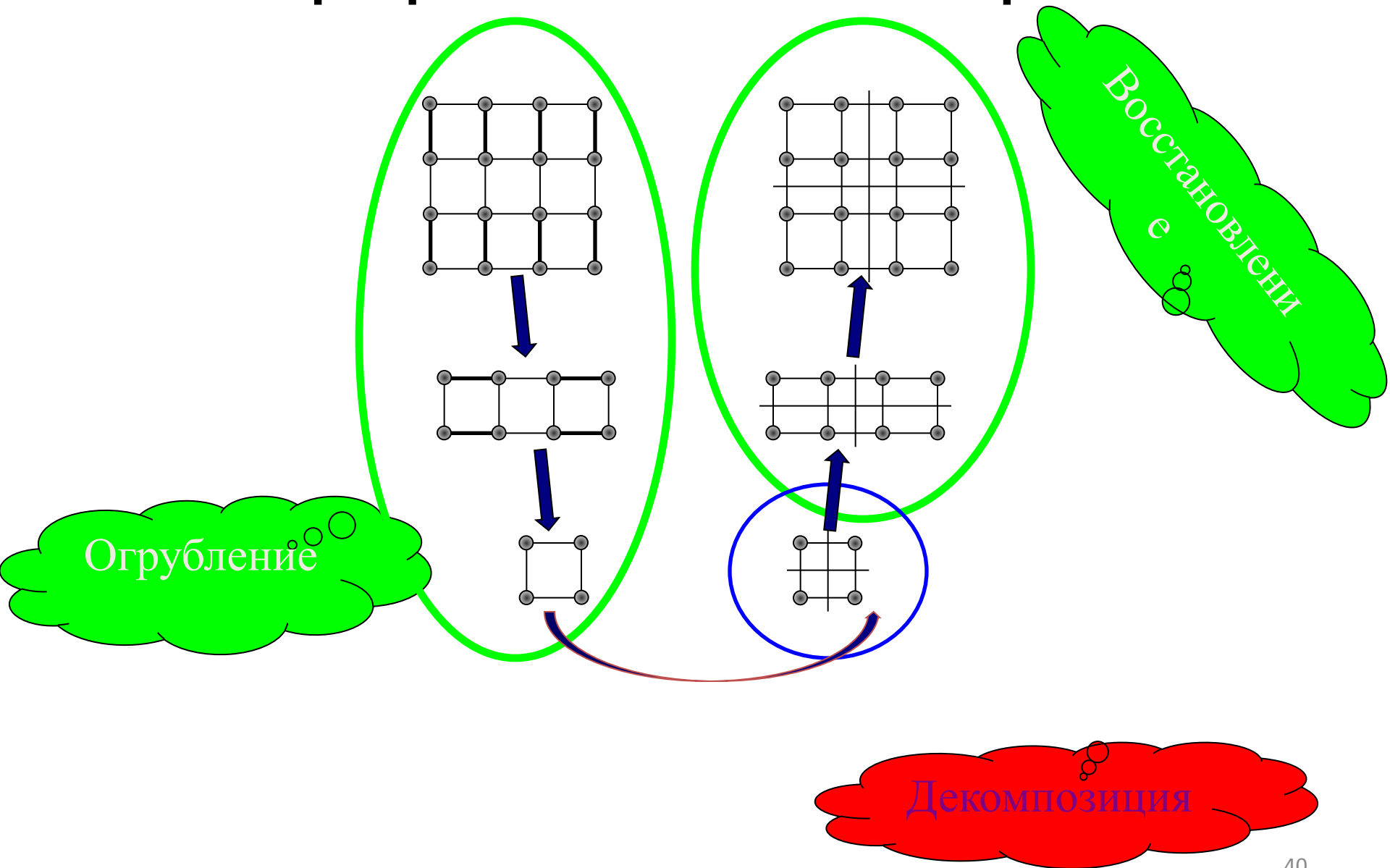
результат геометрической  
декомпозиции

результат перераспределения  
малых блоков вершин

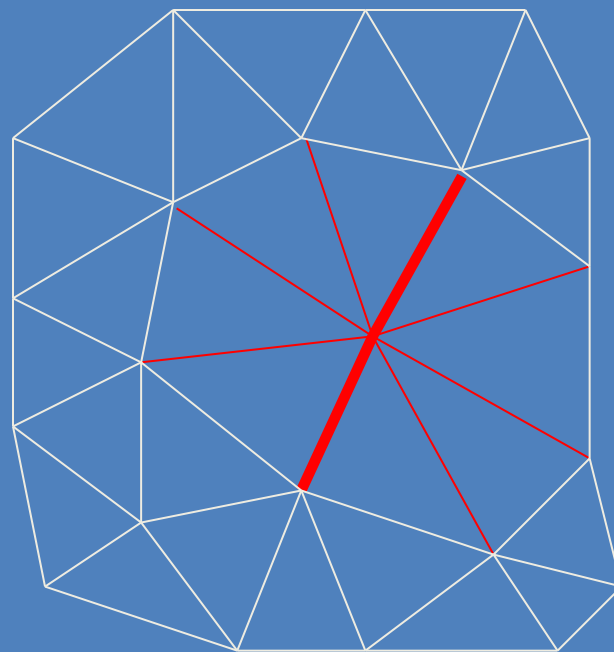
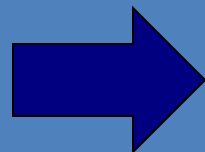
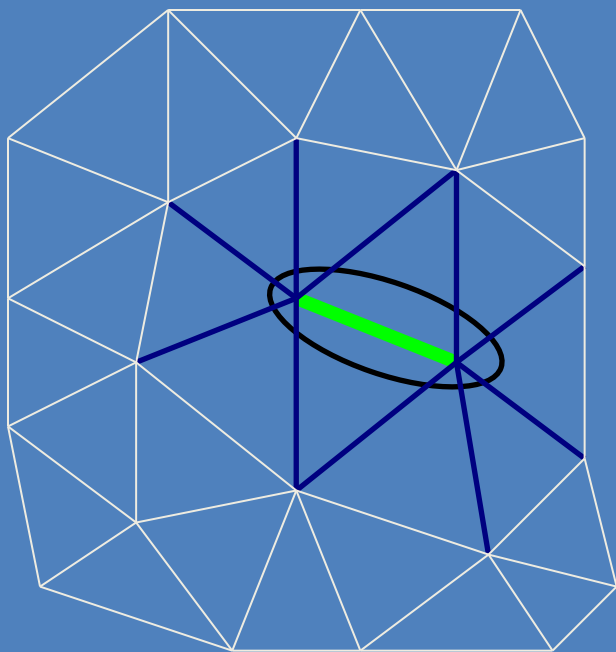




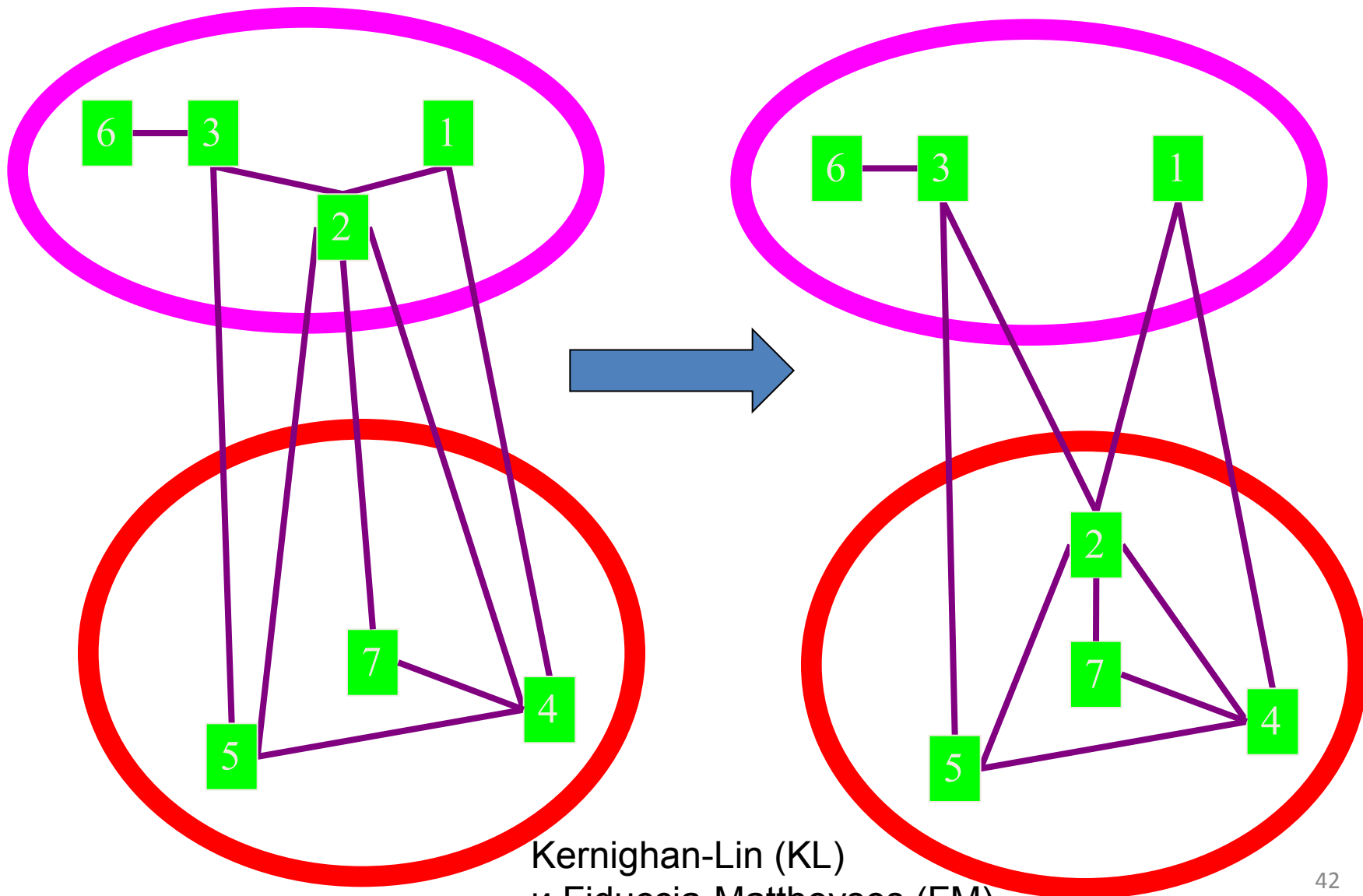
# Иерархический алгоритм



# Огрубление графа



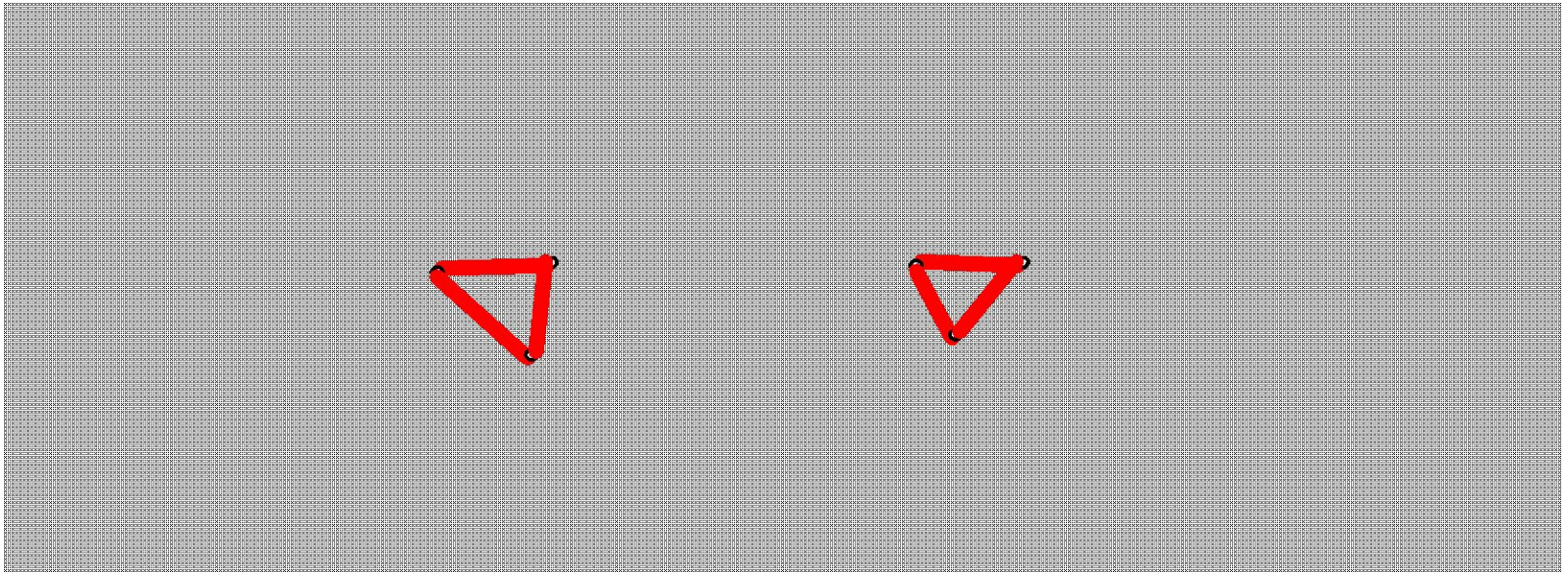
# Локальное уточнение



Kernighan-Lin (KL)  
и Fiduccia-Mattheyses (FM)



# Связность ядер доменов



$$d(i) = \min_k : \mathbf{A}^k v_i \cap B \neq \emptyset$$

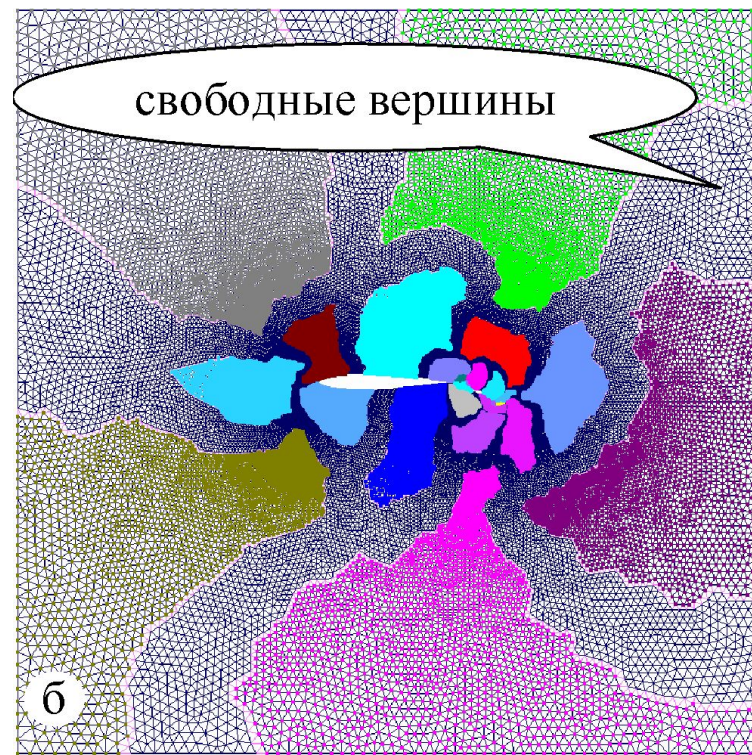
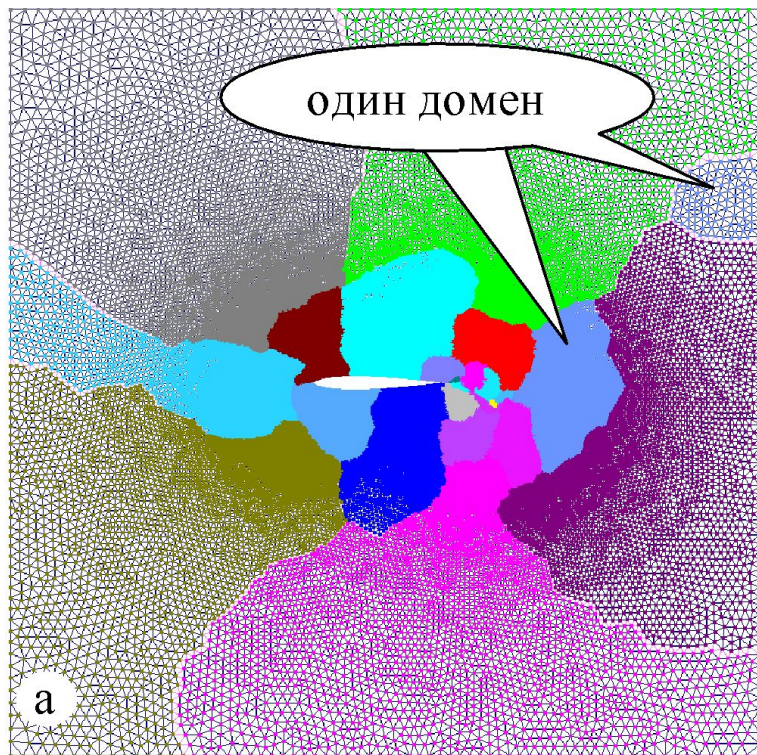
$$T_{k+1} = \mathbf{A}T_k \setminus T_k \setminus T_{k-1}, \quad T_{-1} = \emptyset$$

# Инкрементный алгоритм декомпозиции графа



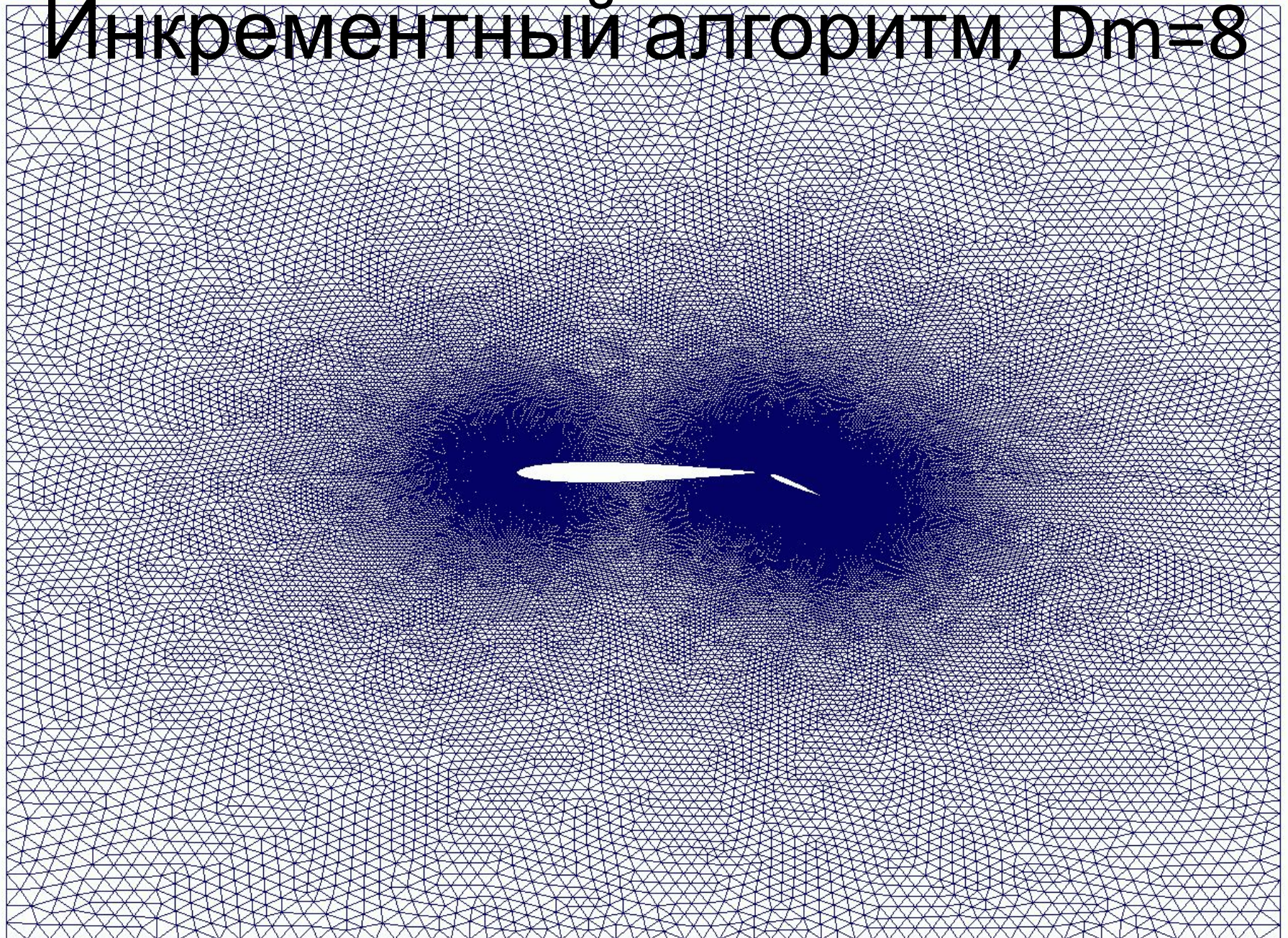


# Редуцирование доменов



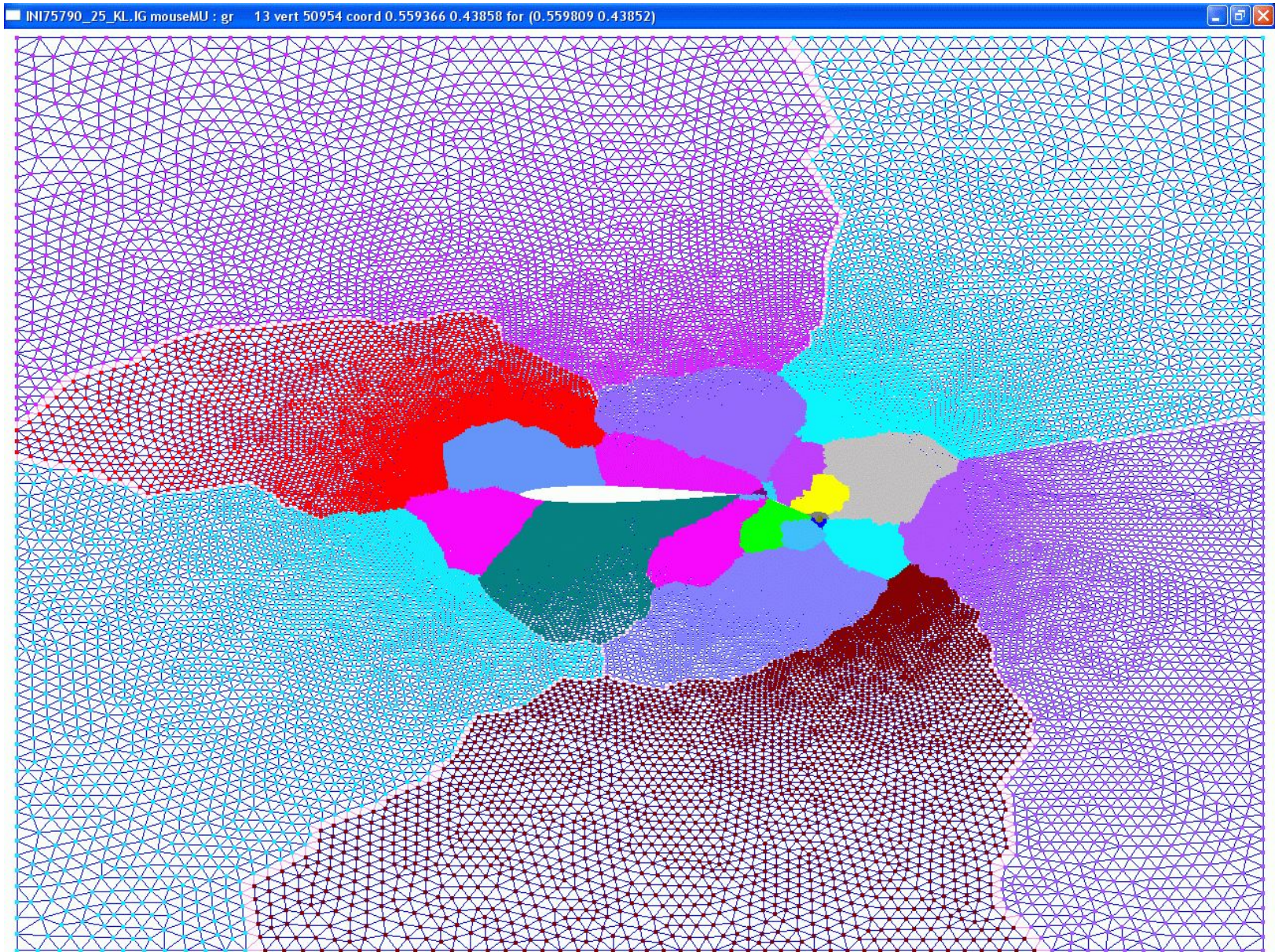


# Инкрементный алгоритм, $Dm=8$



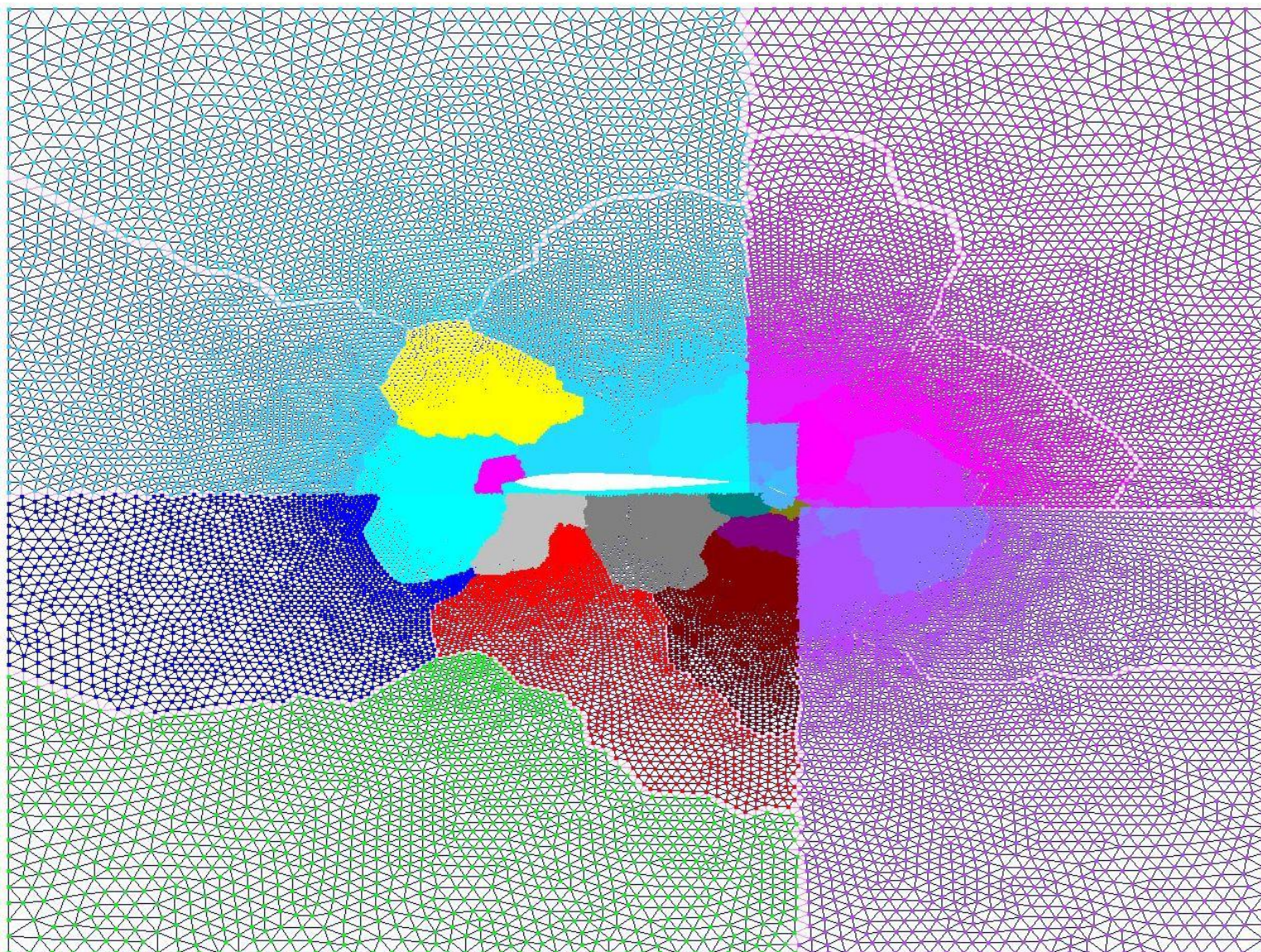


# Инкрементный алгоритм, $Dm=25$



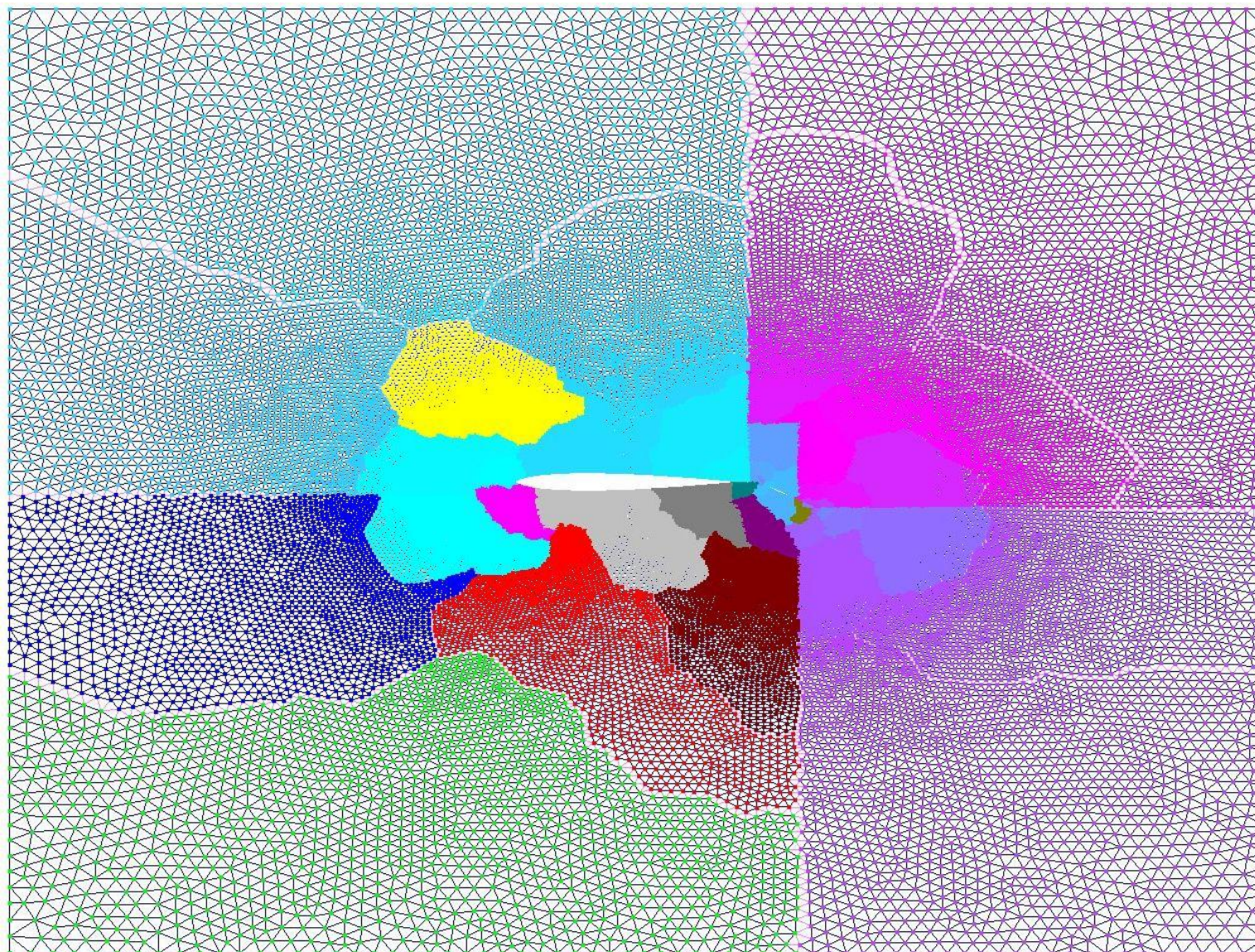


# Результат локального разбиения сетки из 75790 вершин на 50 доменов на 5 процессорах





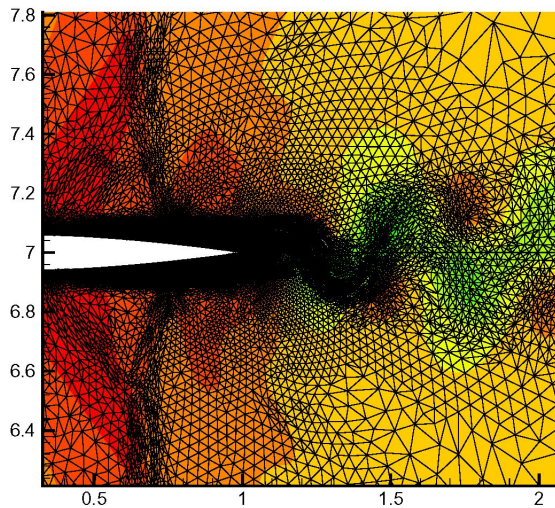
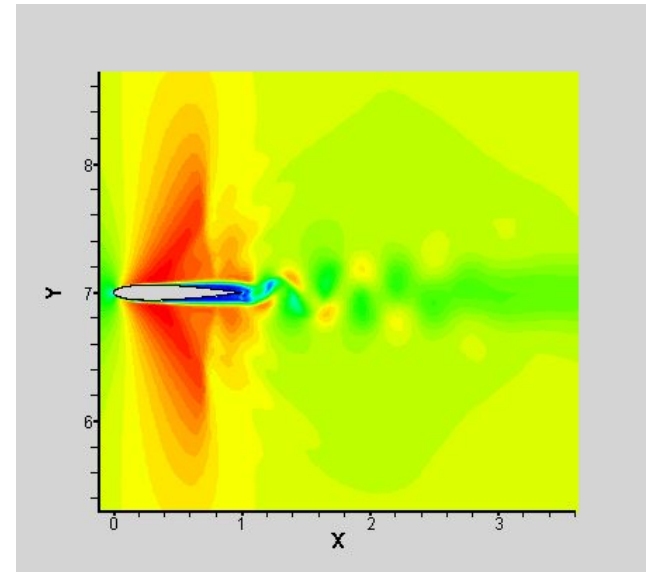
## Результат сбора плохих групп доменов и их повторного разбиения



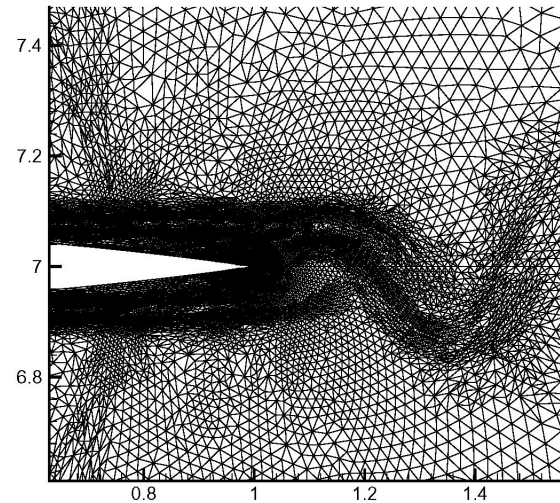


# Адаптивные сетки

Обтекание профиля NASA0012  
( $M=0.85$ ,  $Re=10^4$ )  
под нулевым углом атаки:



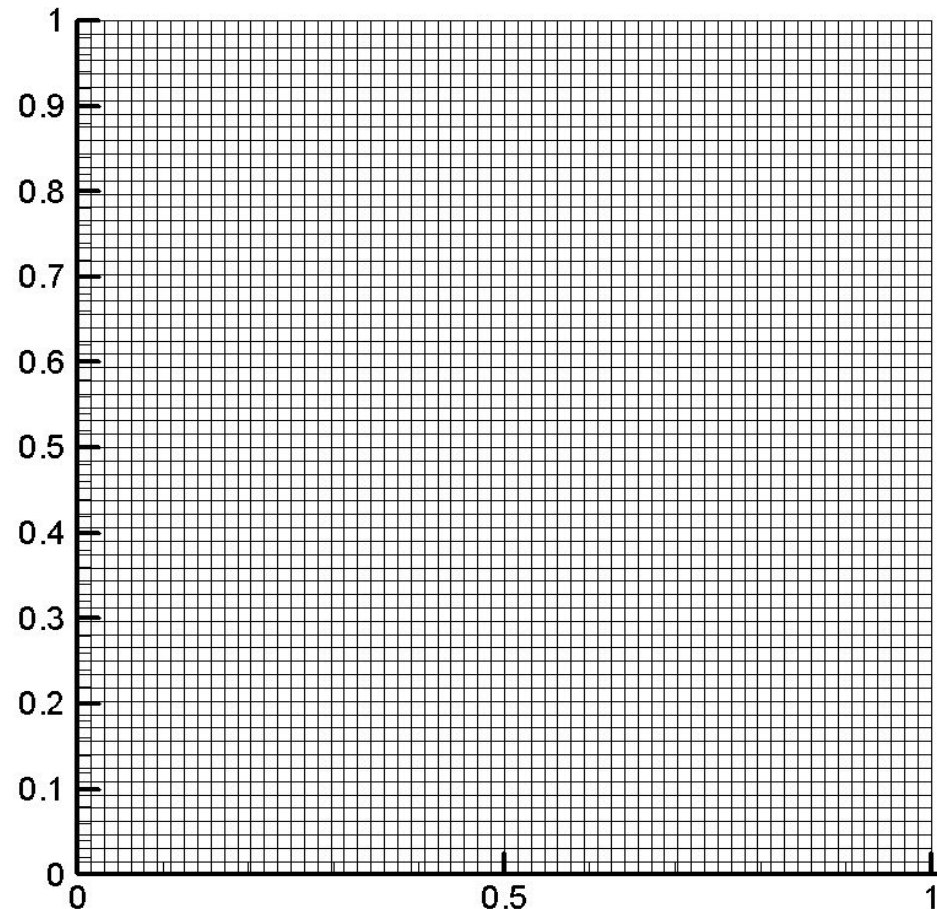
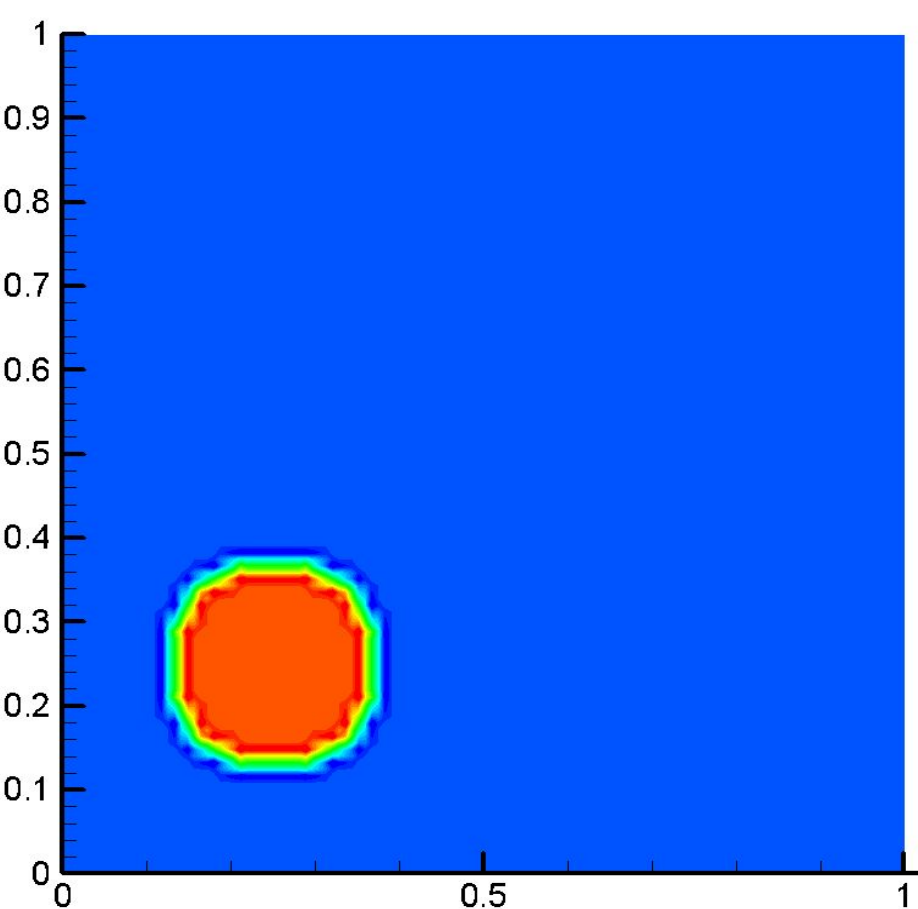
Поле продольной скорости



Фрагмент сетки

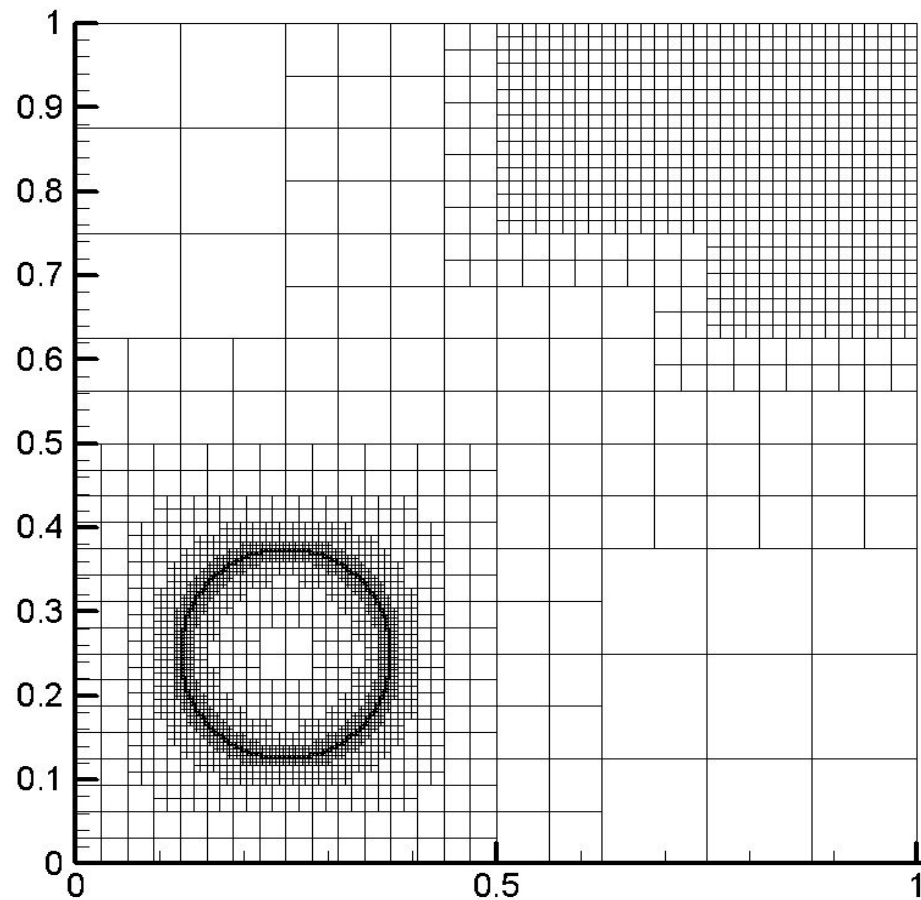
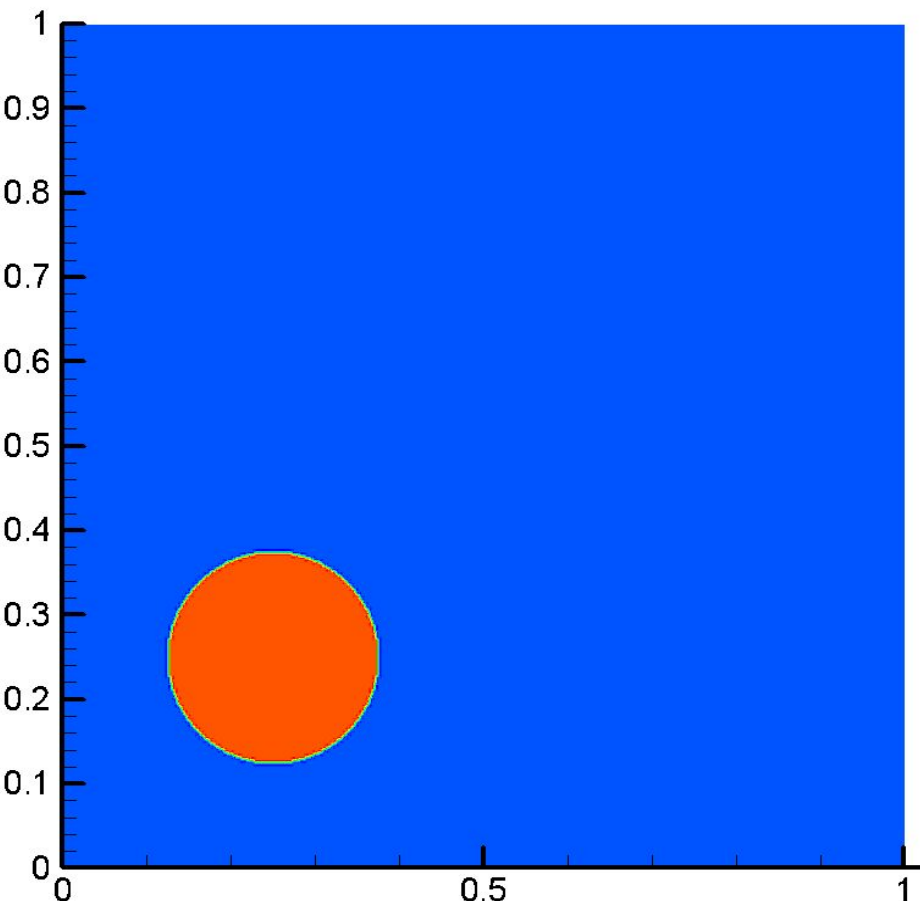
# Равномерная сетка

Слева – ??*круглое*?? пятно примеси



# Адаптивная сетка

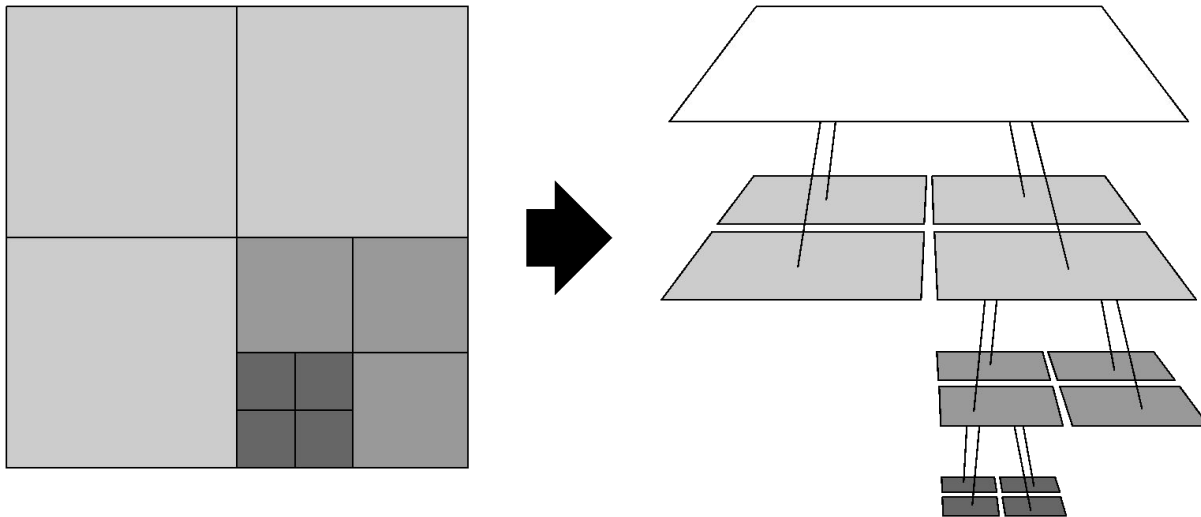
Слева – *круглое* пятно примеси



# Адаптивные декартовы сетки

- Вначале сетка состоит из одной прямоугольной ячейки
- Каждая ячейка может быть **разделена** на четыре ячейки одинакового размера
- Если ячейки когда-то составляли одну ячейку, то они могут быть **объединены** обратно
- Каждая ячейка хранит **величину**, описывающую среднее значение неизвестной функции в пределах ячейки (метод конечных объёмов)

При данных предположениях сетку удобно хранить в виде **четверичного дерева**:



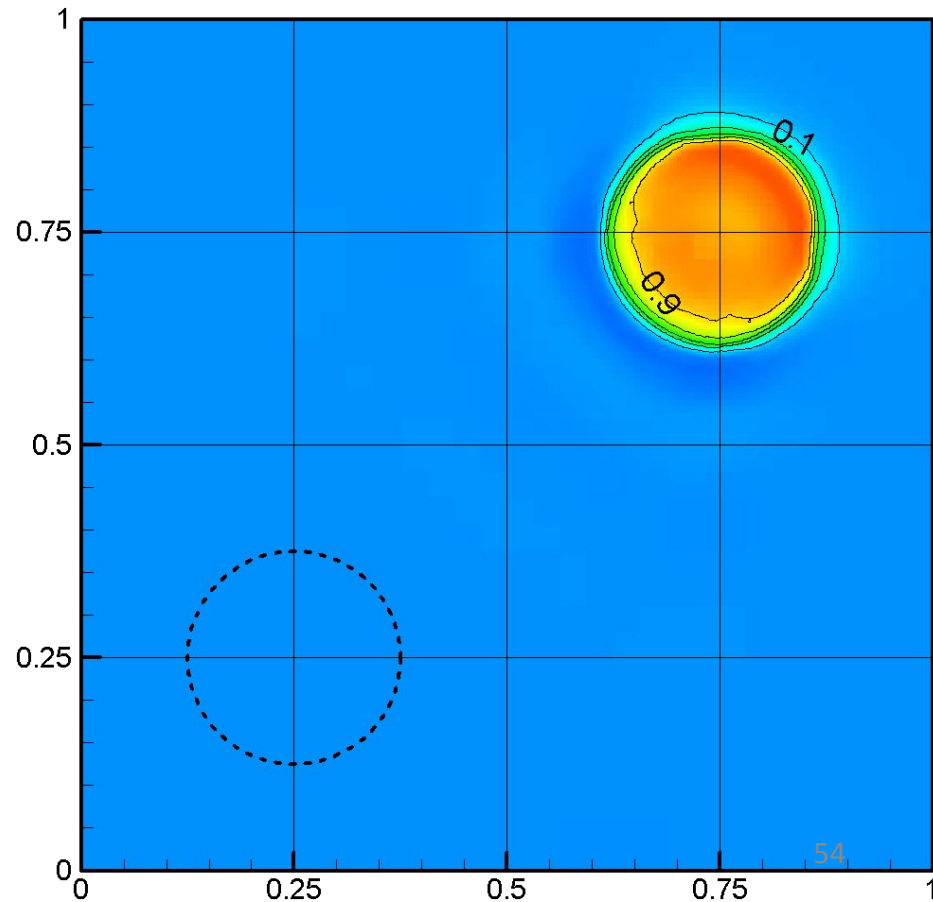
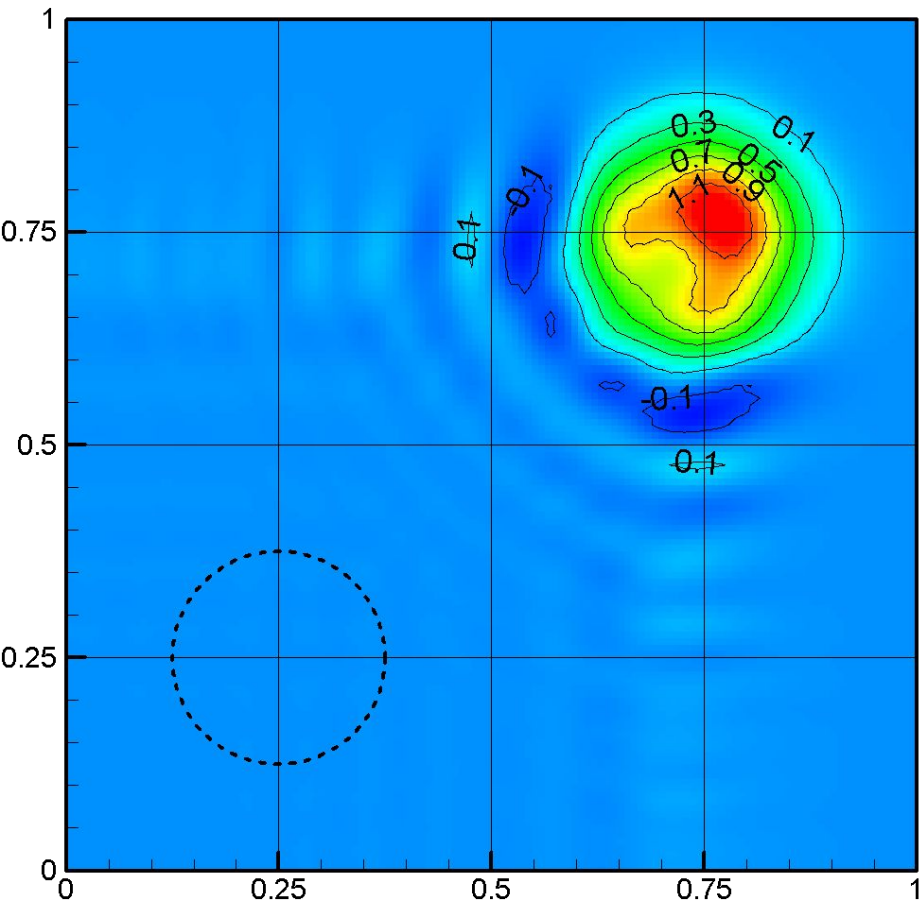
Дополнительные ограничения на размеры ячеек:

- Задан **максимально допустимый** размер ячеек
- Задан **минимально допустимый** размер ячеек
- Размеры соседних ячеек должны различаться **не более, чем в 2 раза**

# Сравнение с равномерной сеткой

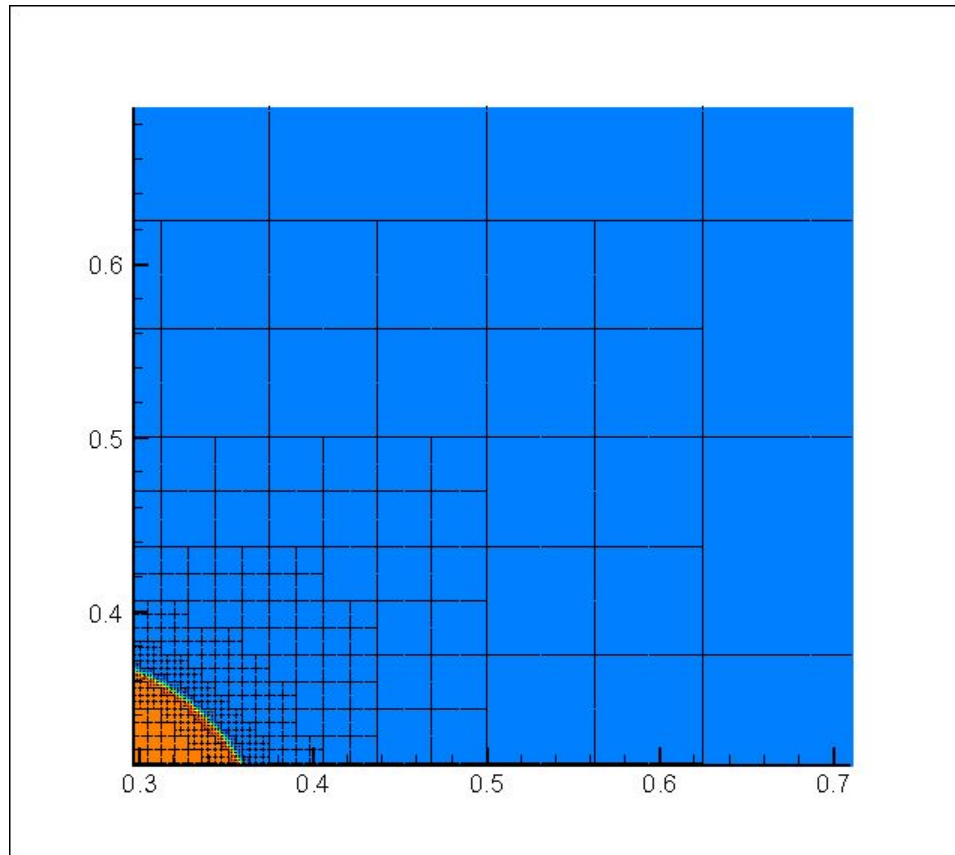
## сеткой

На рисунках показаны результаты решения простейшей задачи переноса на равномерной (слева) и адаптивной (справа) сетках с одинаковым числом ячеек (4096 штук). Скорость переноса направлена под углом  $45^\circ$  к линиям сетки; начальное условие показано пунктиром





# Адаптивная сетка

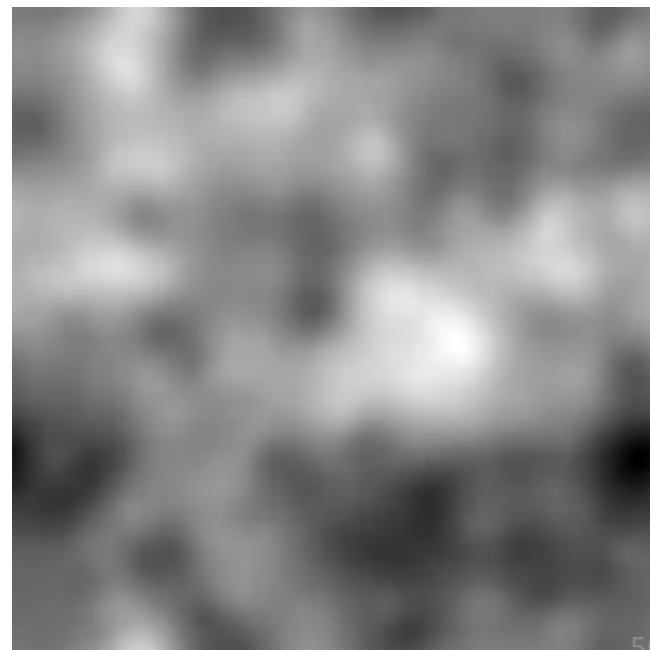
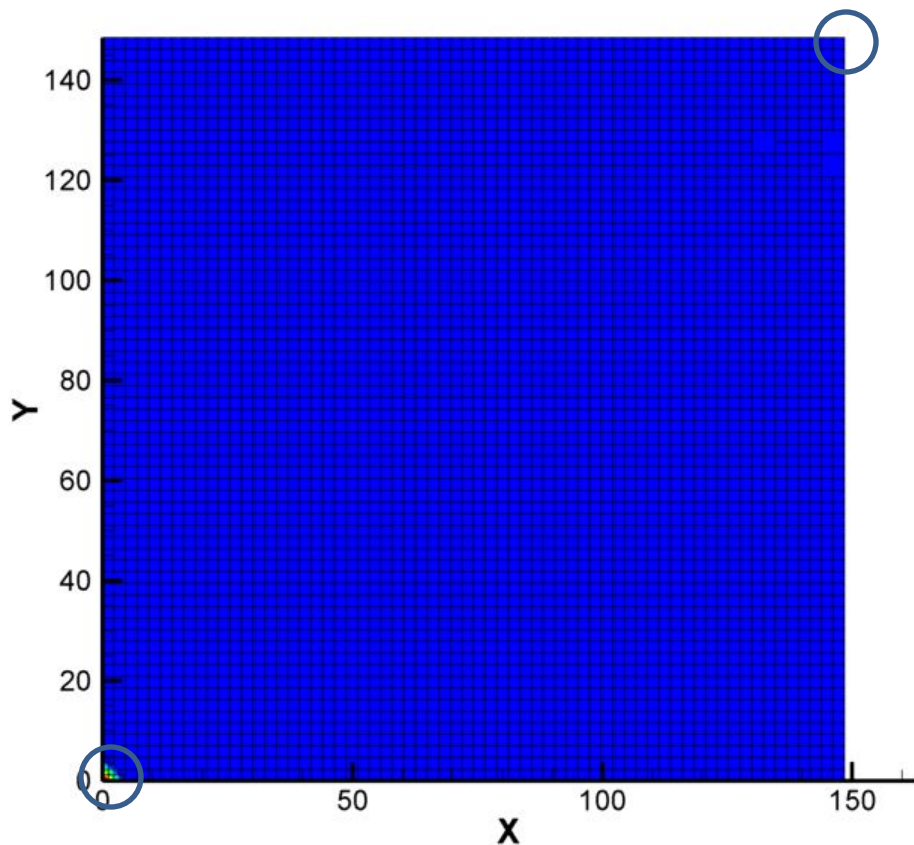


# Решение двумерной задачи фильтрации нефтеводяной смеси в области с неоднородной проницаемостью

В юго-западном углу находится скважина, нагнетающая воду, в северо-восточном углу — добывающая скважина.

5-ти точечная схема

Поле проницаемости с разбросом значений на 4 порядка).

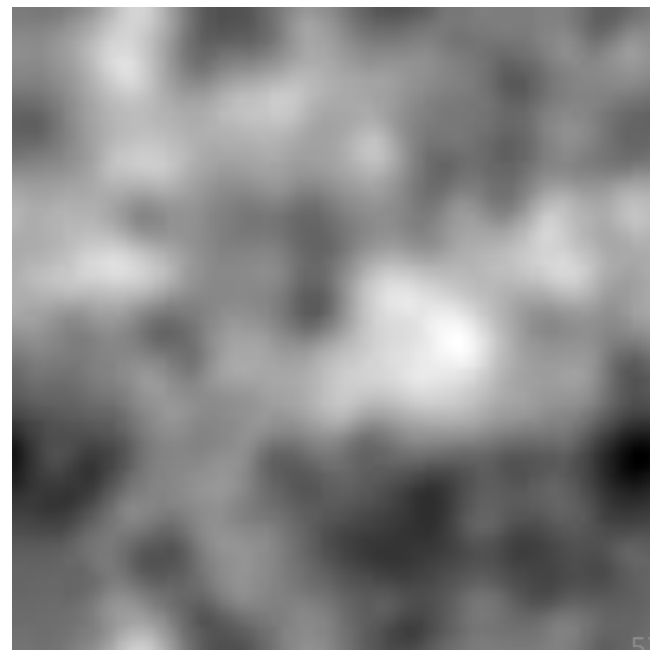
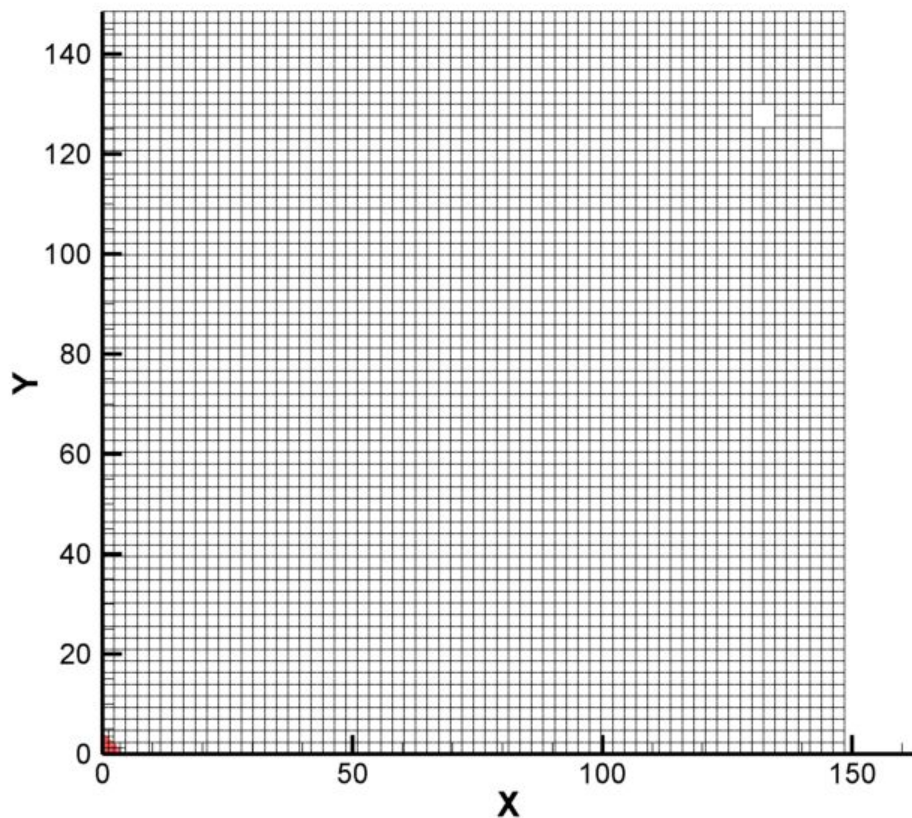


# Решение двумерной задачи фильтрации нефтеводяной смеси в области с неоднородной проницаемостью

В юго-западном углу находится скважина, нагнетающая воду, в северо-восточном углу — добывающая скважина.

5-ти точечная схема

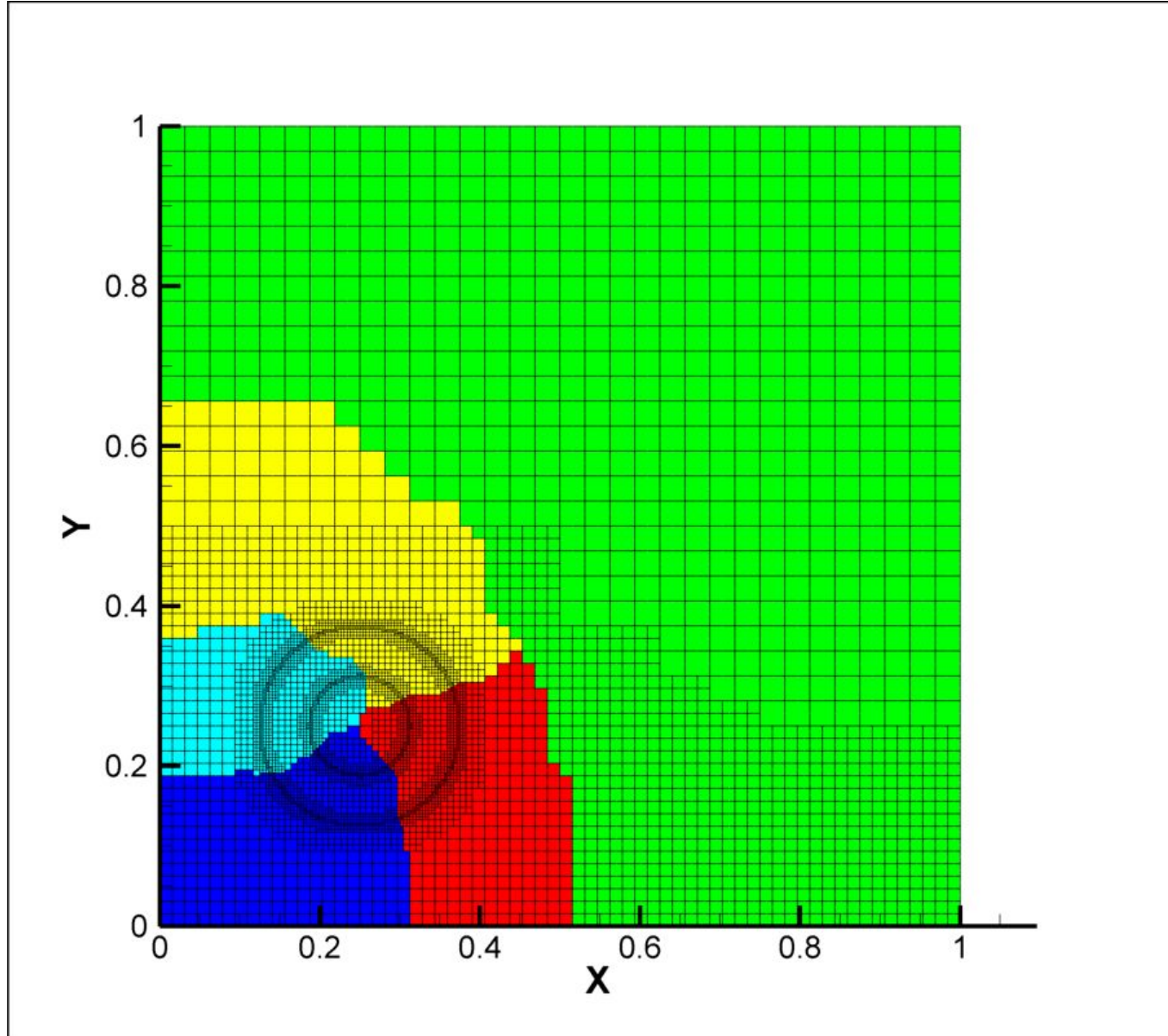
Поле проницаемости с разбросом значений на 4 порядка).

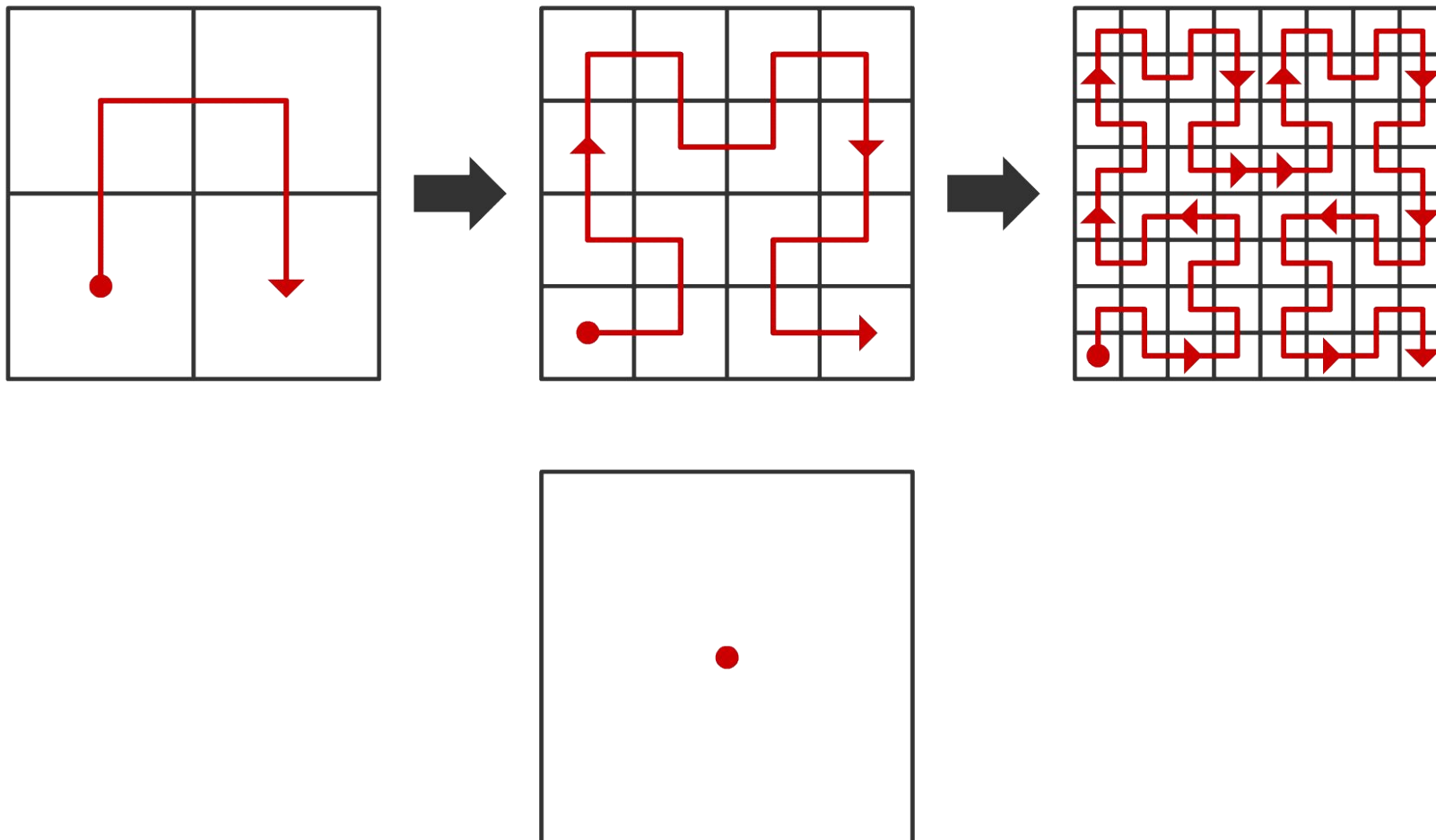


# Динамическая балансировка загрузки

- Перераспределение вычислительных узлов между процессорами необходимо:
  - При изменении конфигурации сетки
  - При изменении вычислительной сложности обработки узлов
  - При изменении эффективной производительности процессоров

# Декомпозиция пакетом Metis





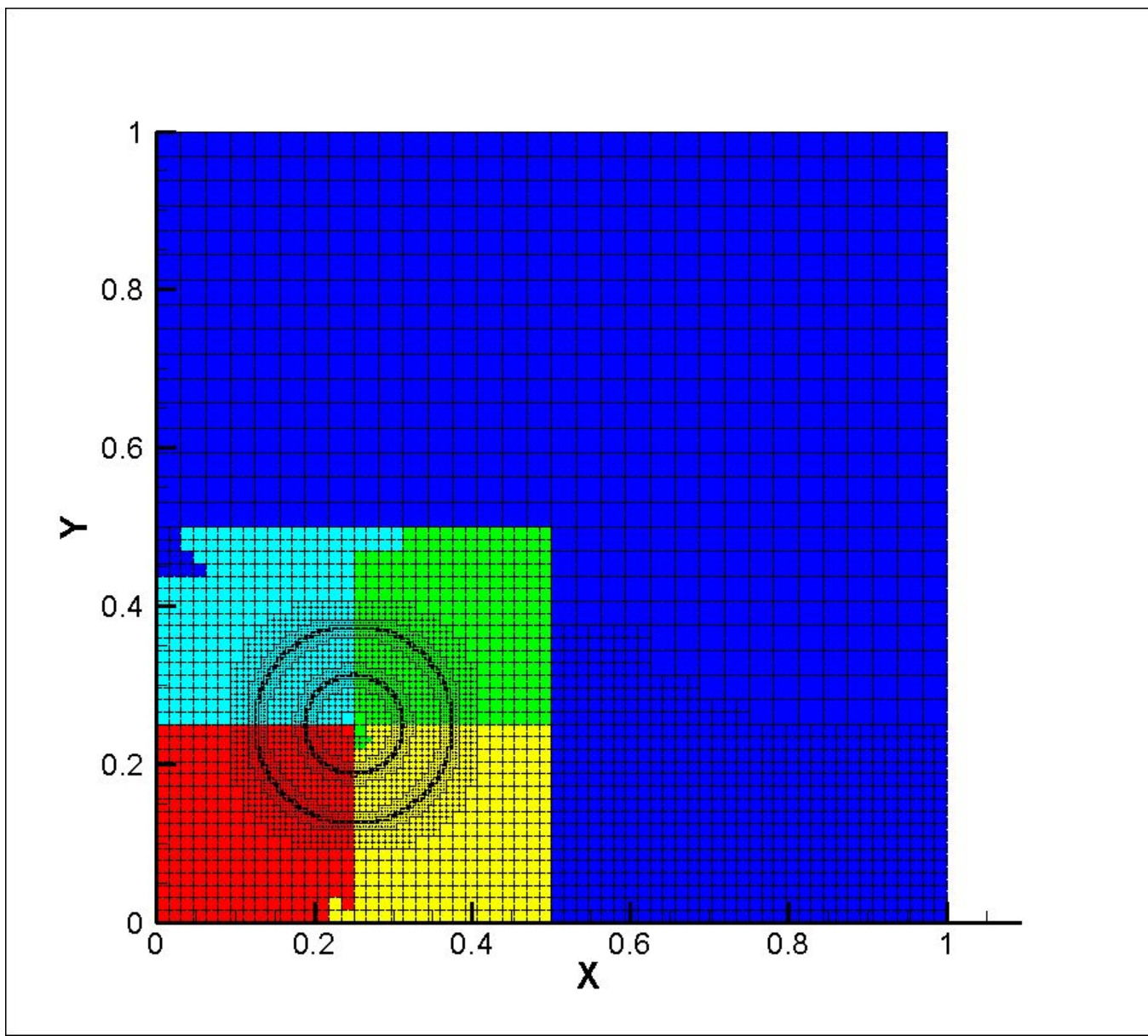
## Нумерация с помощью кривой Гильберта

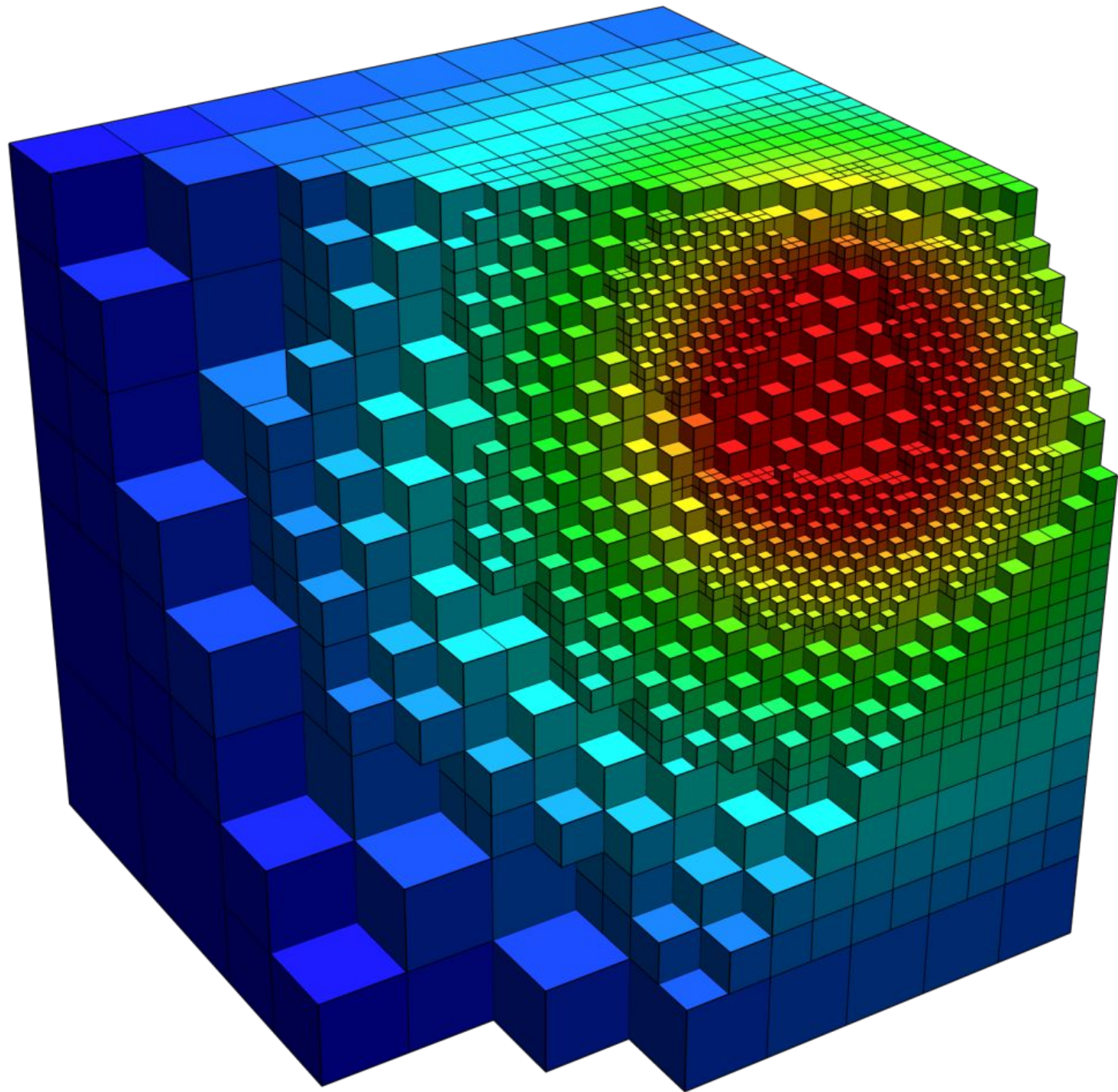
Формируется простой рекурсивной процедурой

Локальное изменение сетки приводит к локальному изменению кривой



# Декомпозиция с помощью кривой Гильберта





# Стратегии балансировки загрузки

$W_i^j$  - вычислительная нагрузка,  
ассоциированная с узлом сетки  $i$  на

Статическая

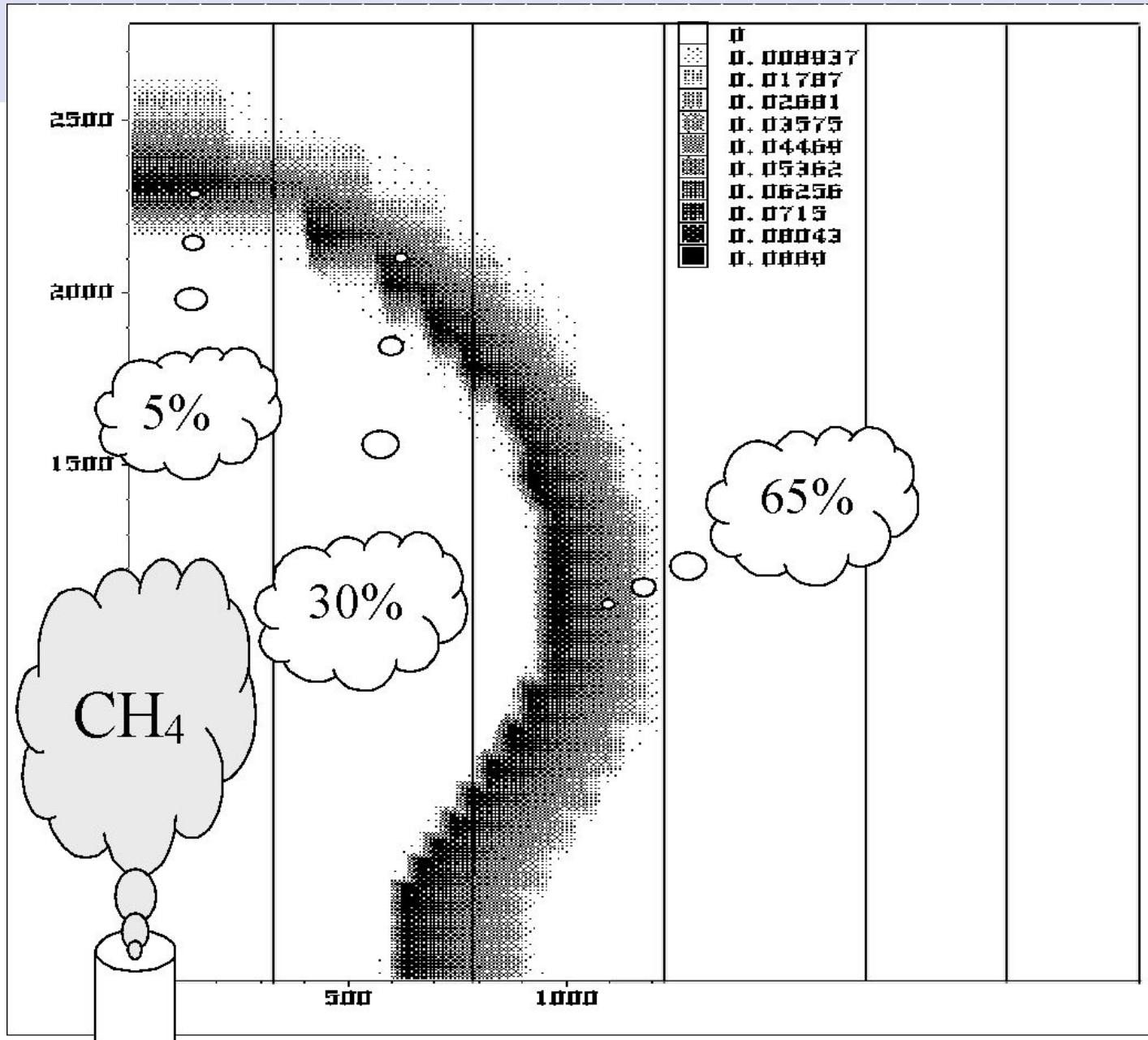
Динамическая  
диффузная

- $W_i^j = W_i^j$  – не зависит от времени
- $W_i^j \approx W_i^{j-1}$  – меняется медленно
- $W_i^j \neq W_i^{j-1}$  – меняется значительно и прогнозируемо

Динамическая

?

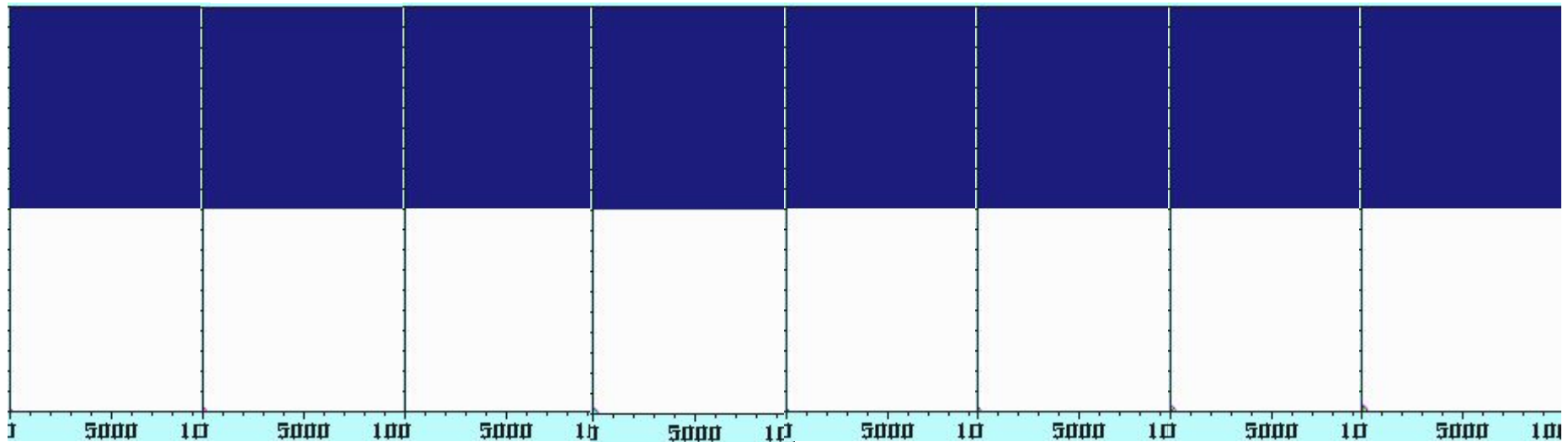
# Моделирование задач горения на многопроцессорных системах





# Methane combustion

$\text{CH}_4$     $\text{NO}$     $\text{NO}_2$     $\text{NO}_2$     $\text{CH}_3$     $\text{CO}$     $\text{CO}_2$     $\text{T}$





# Моделирование задач горения

$$\frac{\partial \mathbf{U}}{\partial t} + A\mathbf{U} = f, \quad \mathbf{U} = (\rho, \rho y^{(i)}, \rho u, \rho v, E)^T \quad f = (0, \omega_i, 0, 0, 0)^T$$

Здесь  $A$  оператор,  $\rho$  - плотность,  
 $y^{(i)}$  – массовые доли  $i$ -х компонент,  
 $u, v$  - скорости,  
 $p$  - давление,  $E$  – полная энергия,  
 $\omega_i$  – скорости образования компонент.

I. Блок Газовой динамики (GD):

$$\frac{\partial \mathbf{U}}{\partial t} + A\mathbf{U} = 0$$

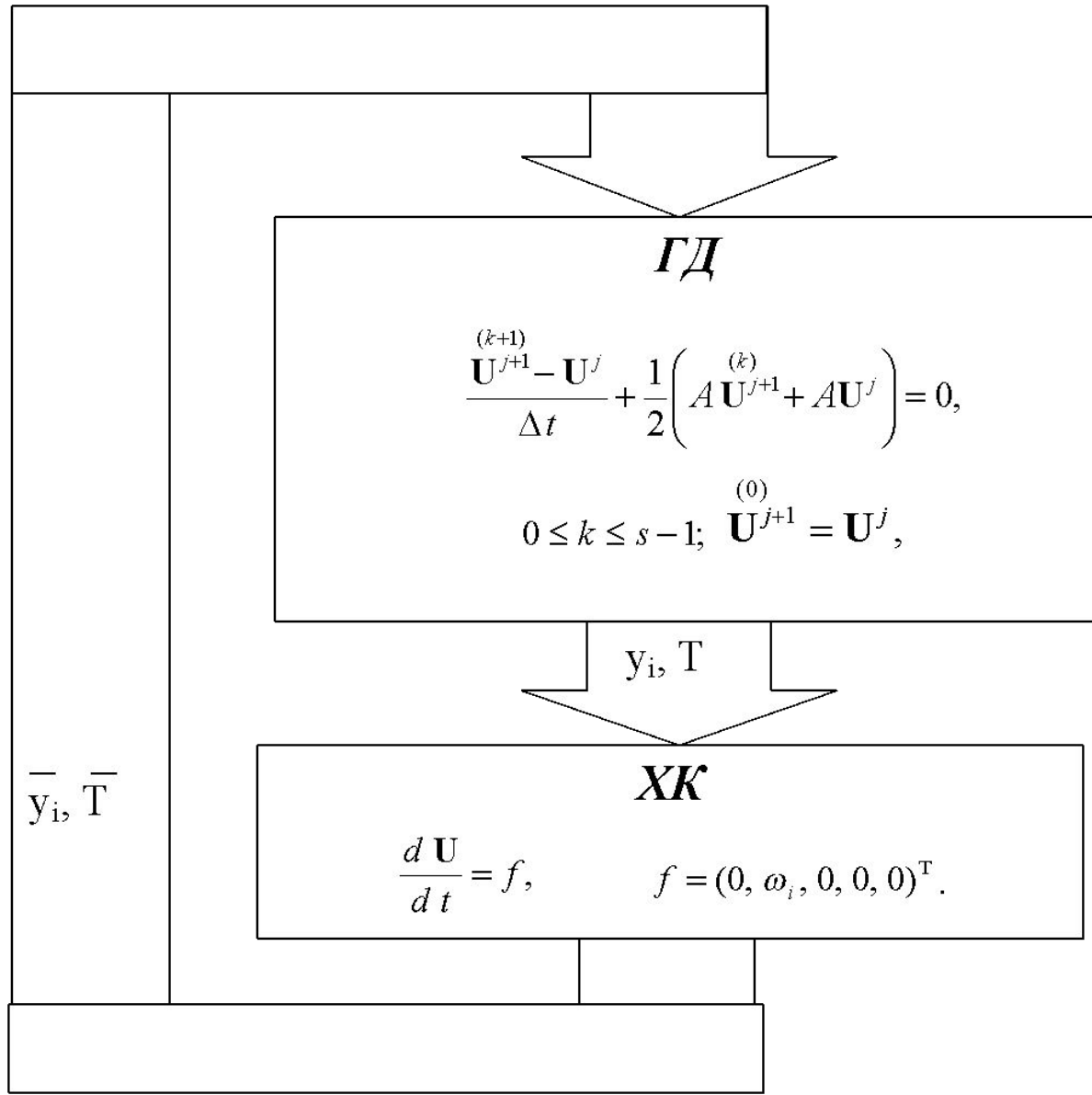
$$\frac{\mathbf{U}^{j+1} - \mathbf{U}^j}{\Delta t} + \frac{1}{2}(A\mathbf{U}^{j+1} + A\mathbf{U}^j) = 0.$$

II. Блок химической кинетики (CHEM):

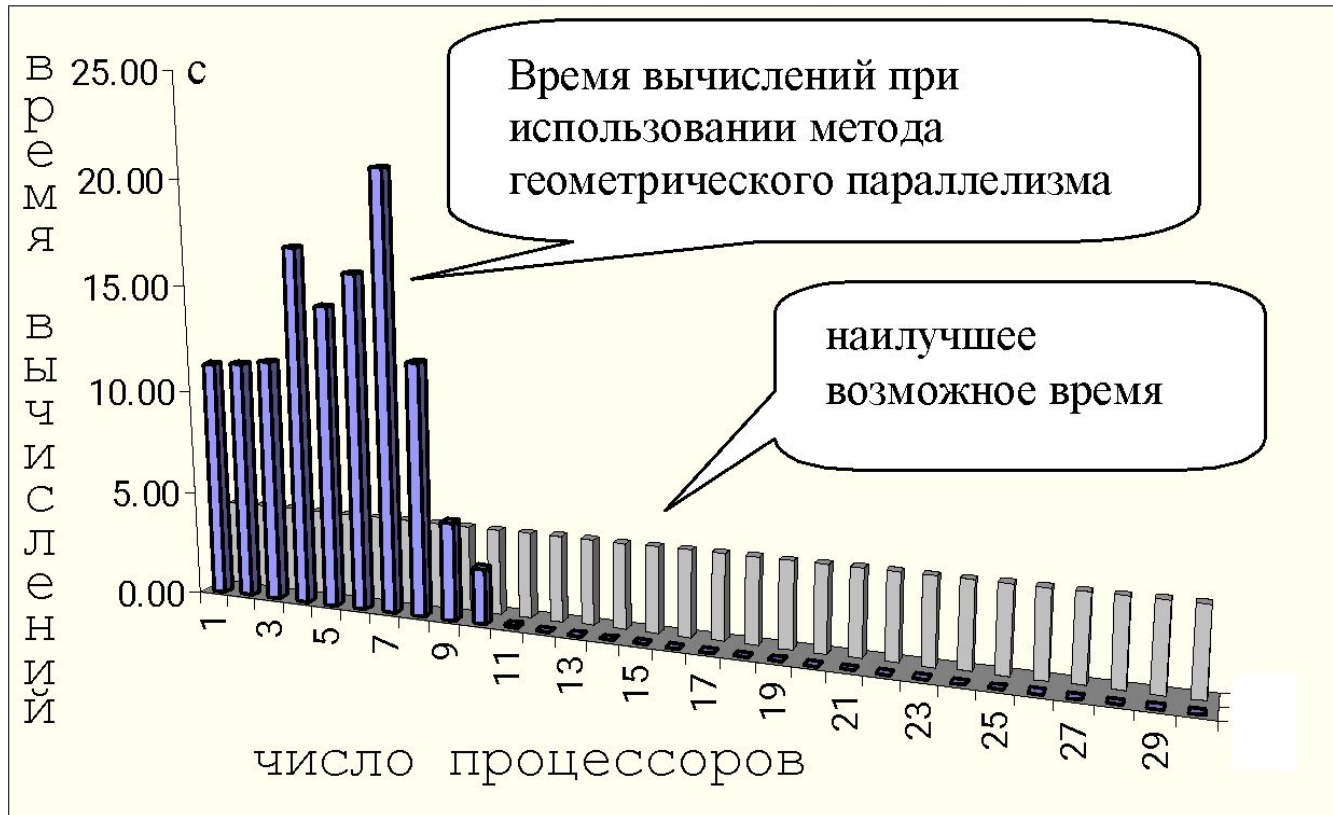
$$\frac{d\mathbf{U}}{dt} = f, \quad f = (0, \omega_i, 0, 0, 0)^T$$



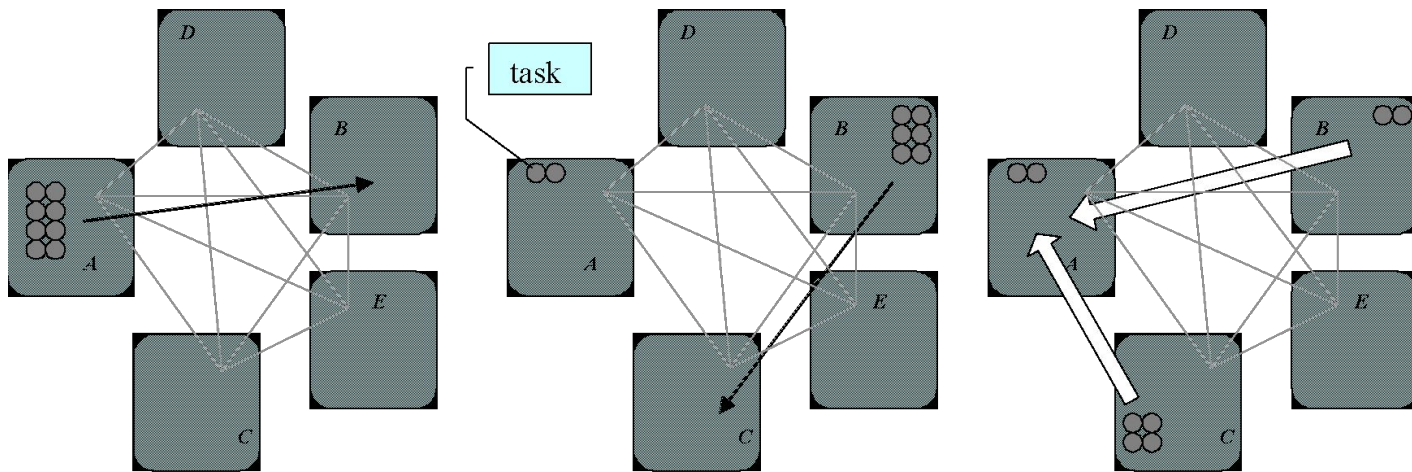
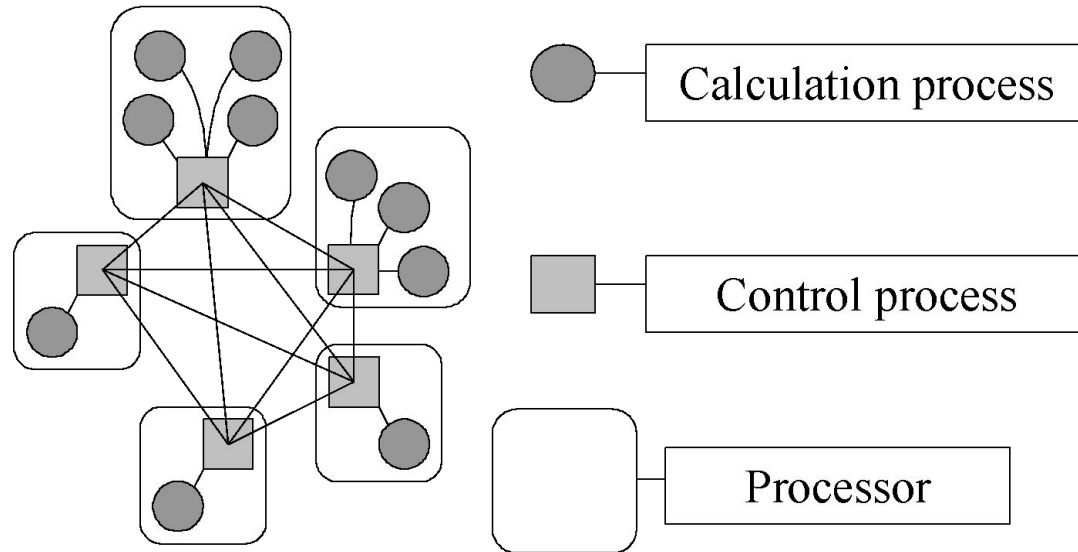
# Блок схема алгоритма



# Распределение времени счета



# Структура и возможности алгоритма



# Состояния обрабатывающего процесса

- **занят** - если установлен соответствующий флаг. Этот флаг устанавливается перед передачей обрабатывающему процессу необработанной точки (неважно локальной или внешней) и сбрасывается после того, как точка уже обработана и управляющий процесс получил от обрабатывающего процесса результат;
- **свободен** - если не занят, т.е. готов к получению очередной свободной точки.



# Управляющий процесс

- **1. если**
  - есть необработанные точки (неважно локальные или внешние) **и**
  - обрабатывающий процесс свободен,
- **то**  
установить **флаг обрабатываемой точки**, одна из необработанных точек передается на обработку обрабатывающему процессу.

# Управляющий процесс

- **2. если**
  - нет локальных необработанных точек ***и***
  - нет внешних точек ***и***
  - нет обрабатываемых точек ***и***
  - ***флаг запроса на получение необработанных точек*** не установлен ***и***
  - есть процессоры, которые еще не ответили, что не могут предоставить точки для обработки (соответствующий флаг ***флаг запрета обменов*** не установлен),
- ***то***
- послать запрос на получение необработанных точек одному из таких процессоров.
- установить ***флаг запроса на получение необработанных точек***

# Управляющий процесс

- **Иначе (если не 2)**
- **3. если**
  - все переданные точки получены обратно обработанными **и**
  - от всех процессоров получено сообщение о том, что точки для обработки предоставлены быть не могут **и**
  - всем процессорам послано сообщение о том, что точки для обработки предоставлены быть не могут,  
**то**  
завершение работы

# Управляющий процесс

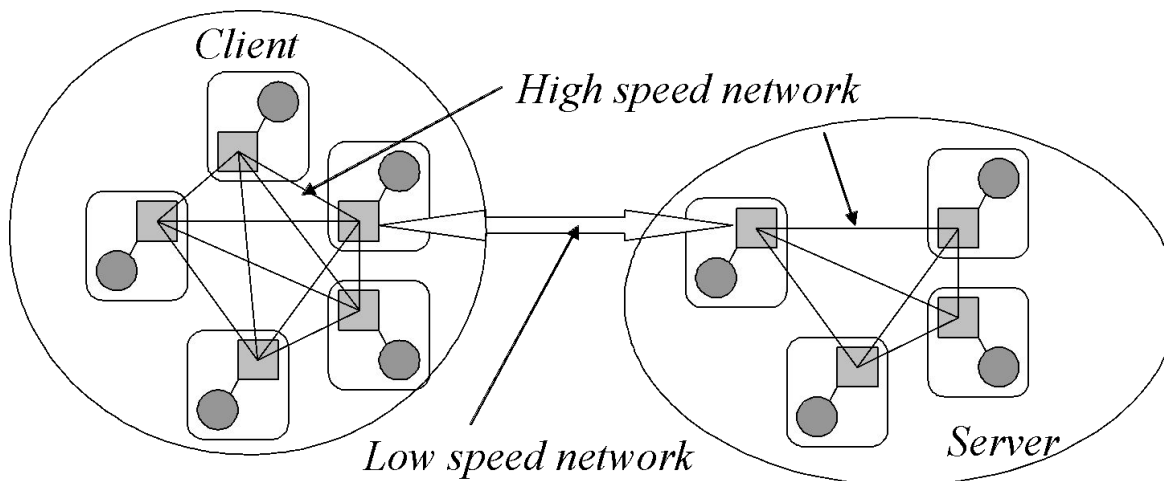
- 4. получить очередное сообщение от любого процессора или от своего обрабатывающего процесса.
- 5. обработать полученное сообщение
- 6. перейти к началу цикла (п. 1)



# Окончание при выполнении всех условий:

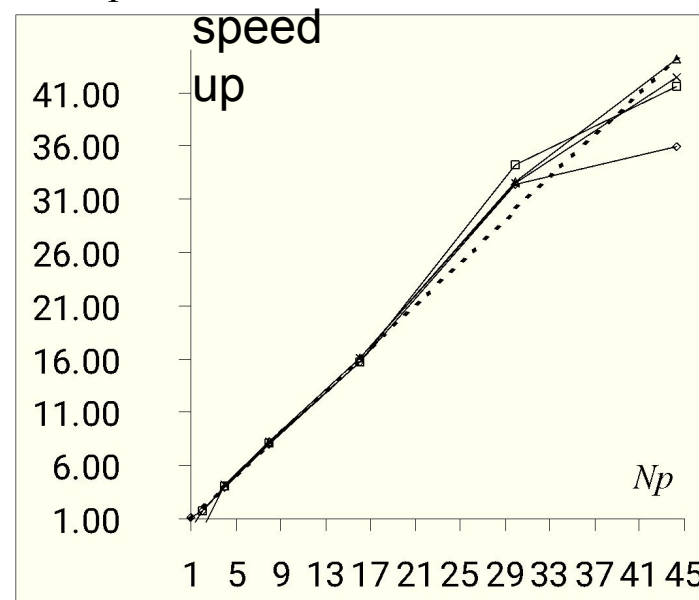
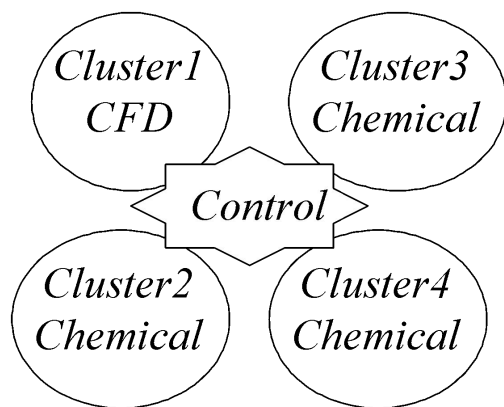
- нет локальных необработанных точек
- нет внешних точек
- нет обрабатываемых точек
- всем процессорам был послан запрос на получение необработанных точек
- всем процессорам было послано сообщение о том, что необработанные точки предоставлены быть не могут
- от всех процессоров получено сообщение о том, что необработанные точки предоставлены быть не могут
- все локальные точки обработаны и получены

# Кластеры и эффективность

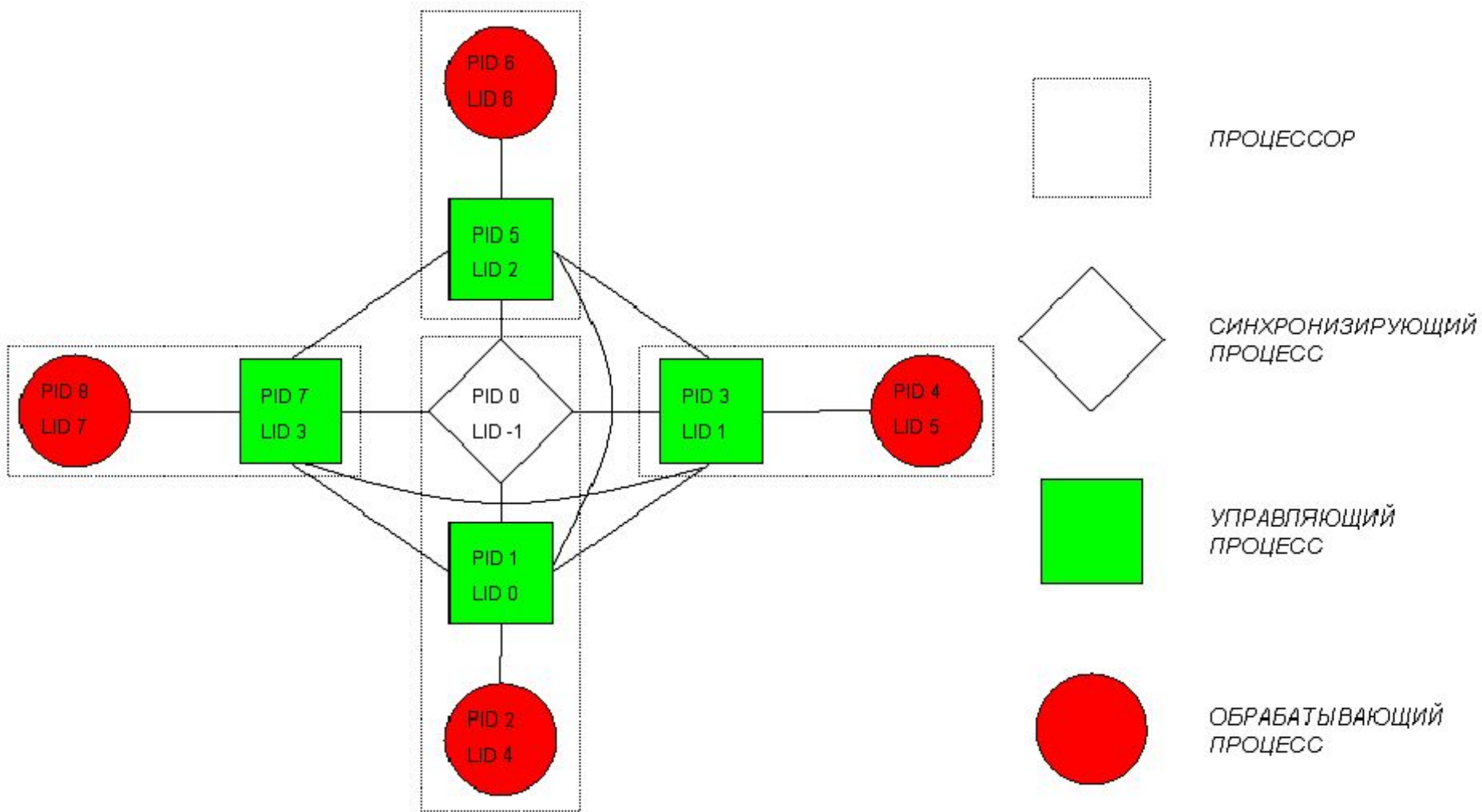


32 processors

12 processors



# Схема взаимодействия процессов



# Выводы

- Балансировка загрузки процессоров – ключевой этап обеспечения высокой эффективности использования многопроцессорной системы.
- С ростом числа процессоров возрастает актуальность использования динамической балансировки загрузки