Применение технологий автоадаптации программ для решения CFD задач на структурированных сетках с использованием GPU

Авторы:

Кривов М.А.,
Притула М.Н,
Иванов П.С.

Ростов-На-Дону, 2014

Применение технологий автоадаптации программ для решения CFD задач на структурированных сетках с использованием GPU

Авторы:

Кривов М.А.,
Притула М.Н,
Иванов П.С.

Ростов-На-Дону, 2014

Зачем всё это нужно

- ***Проблема***
  - Иногда даже после проведения оптимизации скорость расчётов остаётся неудовлетворительной

Зачем всё это нужно

- ***Проблема***
  - Иногда даже после проведения оптимизации скорость расчётов остаётся неудовлетворительной
- ***Схема разработки***
  - Запрограммировали алгоритм

Зачем всё это нужно

- ***Проблема***
  - Иногда даже после проведения оптимизации скорость расчётов остаётся неудовлетворительной
- ***Схема разработки***
  - Запрограммировали алгоритм
  - Провели «общую» оптимизацию

Зачем всё это нужно

- ***Проблема***
  - Иногда даже после проведения оптимизации скорость расчётов остаётся неудовлетворительной
- ***Схема разработки***
  - Запрограммировали алгоритм
  - Провели «общую» оптимизацию
  - Перенесли на GPU

Зачем всё это нужно

- ***Проблема***
  - Иногда даже после проведения оптимизации скорость расчётов остаётся неудовлетворительной
- ***Схема разработки***
  - Запрограммировали алгоритм
  - Провели «общую» оптимизацию
  - Перенесли на GPU
  - Провели ещё раз «общую» оптимизацию

Зачем всё это нужно

- ***Проблема***
  - Иногда даже после проведения оптимизации скорость расчётов остаётся неудовлетворительной
- ***Схема разработки***
  - Запрограммировали алгоритм
  - Провели «общую» оптимизацию
  - Перенесли на GPU
  - Провели ещё раз «общую» оптимизацию
  - Провели оптимизацию под конкретную модель GPU

Зачем всё это нужно

- ***Проблема***
  - Иногда даже после проведения оптимизации скорость расчётов остаётся неудовлетворительной
- ***Схема разработки***
  - Запрограммировали алгоритм
  - Провели «общую» оптимизацию
  - Перенесли на GPU
  - Провели ещё раз «общую» оптимизацию
  - Провели оптимизацию под конкретную модель GPU
  - **?**

Зачем всё это нужно

- ***Проблема***
  - Иногда даже после проведения оптимизации скорость расчётов остаётся неудовлетворительной
- ***Схема разработки***
  - Запрограммировали алгоритм
  - Провели «общую» оптимизацию
  - Перенесли на GPU
  - Провели ещё раз «общую» оптимизацию
  - Провели оптимизацию под конкретную модель GPU
  - Встроили механизм автоадаптации

Суть технологий автоадаптации программ

- При компиляции «закладывается» возможность перестраивать программу во время работы
- Во время или до начала работы на целевой системе отдельный модуль производит донастройку программы

Суть технологий автоадаптации программ

- При компиляции «закладывается» возможность перестраивать программу во время работы
- Во время или до начала работы на целевой системе отдельный модуль производит донастройку программы
- ***Итог:*** производительность удаётся повысить за счёт учёта особенностей, не известных в момент компиляции

Суть технологий автоадаптации программ

- При компиляции «закладывается» возможность перестраивать программу во время работы
- Во время или до начала работы на целевой системе отдельный модуль производит донастройку программы
- ***Итог:*** производительность удаётся повысить за счёт учёта особенностей, не известных в момент компиляции

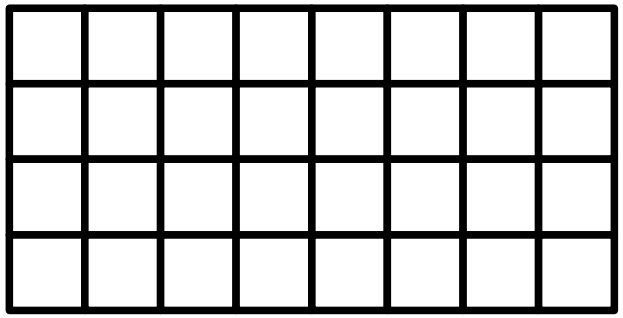
Во всех примерах далее используется автотюнер TTG Apptimizer!

Виды оптимизируемых параметров

- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

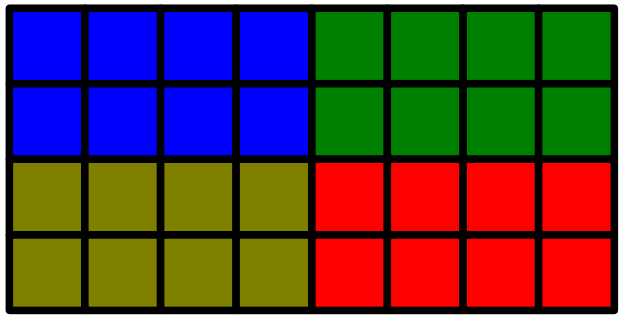
- ***Размеры GPU-блоков***



- Тяжеловесность нитей
- Схема адресации памяти
- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

- ***Размеры GPU-блоков***

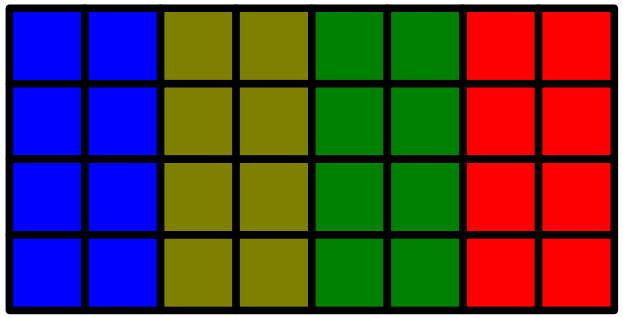


Размеры:
32x16

- Тяжеловесность нитей
- Схема адресации памяти
- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

- ***Размеры GPU-блоков***

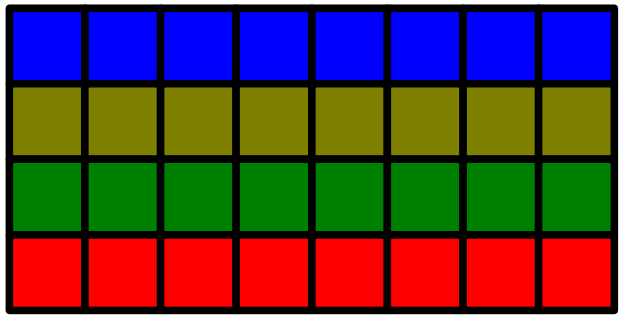


Размеры:
32x16
16x32

- Тяжеловесность нитей
- Схема адресации памяти
- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

- ***Размеры GPU-блоков***

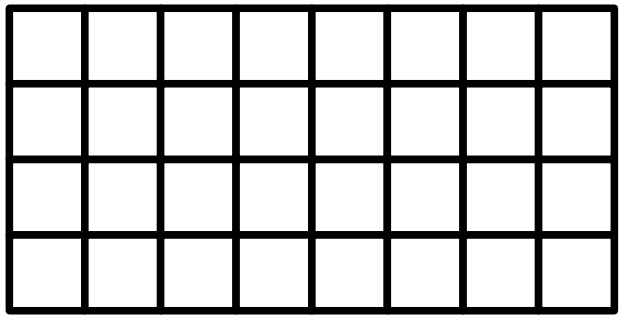


```
Размеры:
   32x16
   16x32
   64x8 ...
```

- Тяжеловесность нитей
- Схема адресации памяти
- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

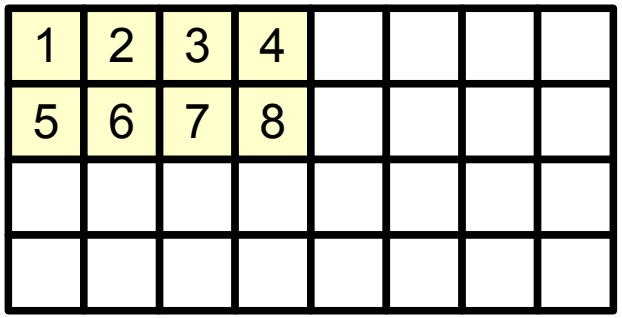
- Размеры GPU-блоков
- ***Тяжеловесность нитей***



- Схема адресации памяти
- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

- Размеры GPU-блоков
- ***Тяжеловесность нитей***

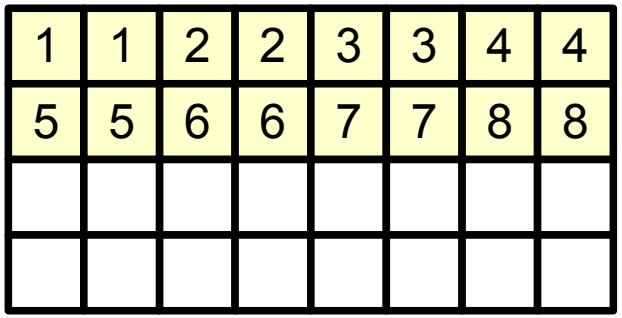


Размер «ячейки»:
1x1

- Схема адресации памяти
- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

- Размеры GPU-блоков
- ***Тяжеловесность нитей***

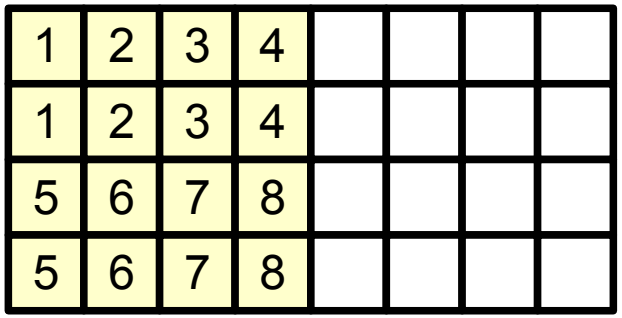


Размер «ячейки»:
1x1
1x2

- Схема адресации памяти
- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

- Размеры GPU-блоков
- ***Тяжеловесность нитей***

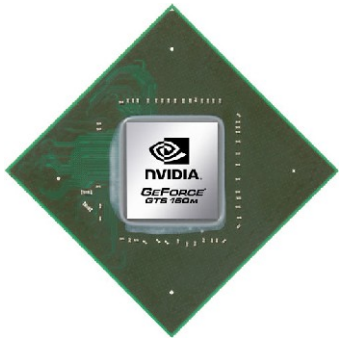



```
Размер «ячейки»:
   1x1
   1x2
   2x1 ...
```

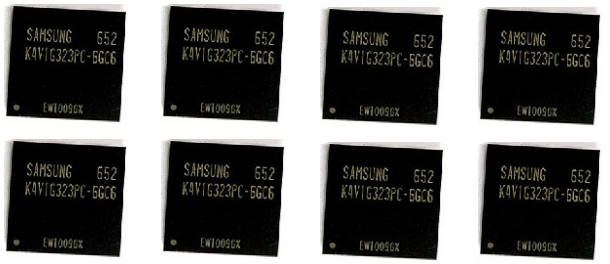
- Схема адресации памяти
- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

- Размеры GPU-блоков
- Тяжеловесность нитей
- ***Схема адресации памяти***



GPU

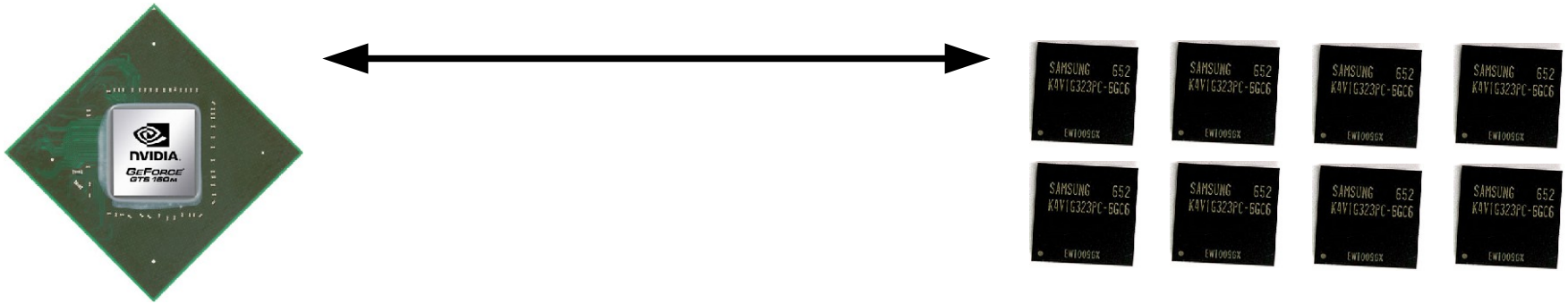


Глобальная память

- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

- Размеры GPU-блоков
- Тяжеловесность нитей
- ***Схема адресации памяти***

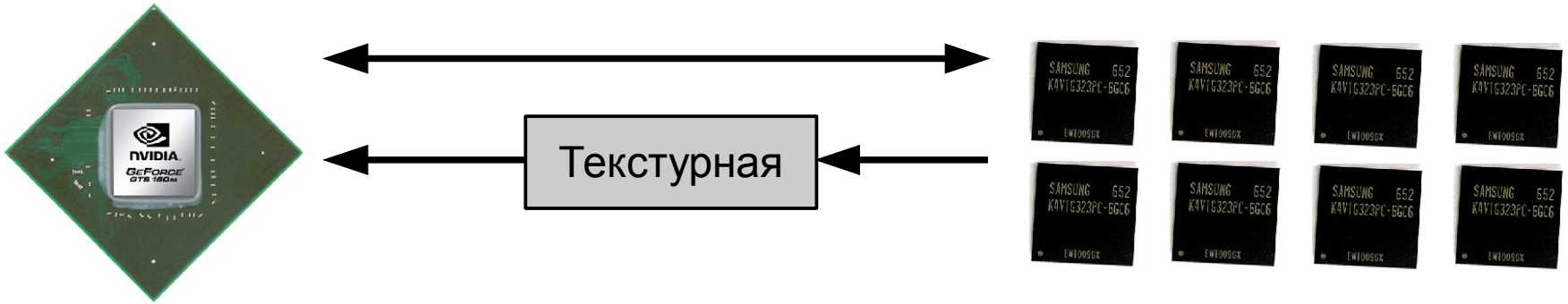


GPU Глобальная память

- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

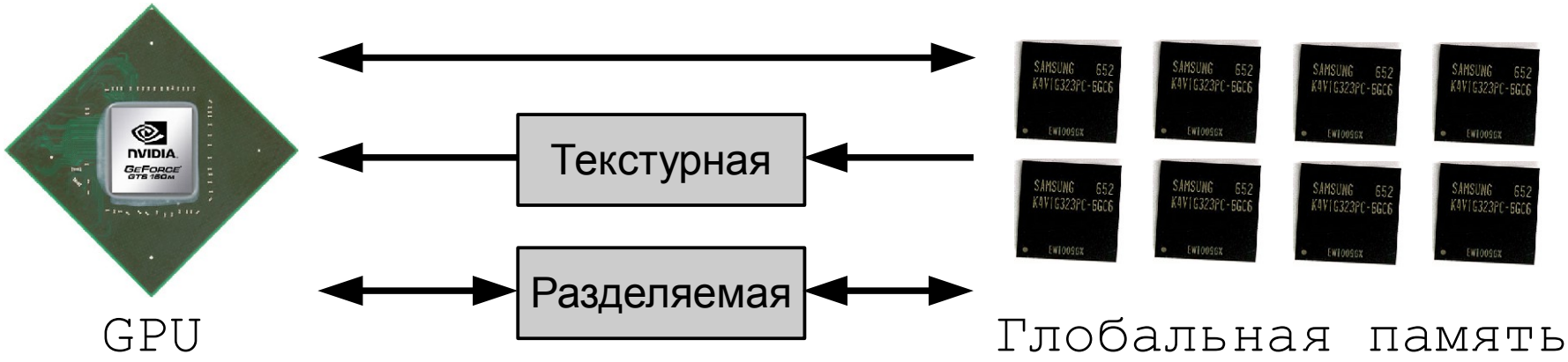
- Размеры GPU-блоков
- Тяжеловесность нитей
- ***Схема адресации памяти***



- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

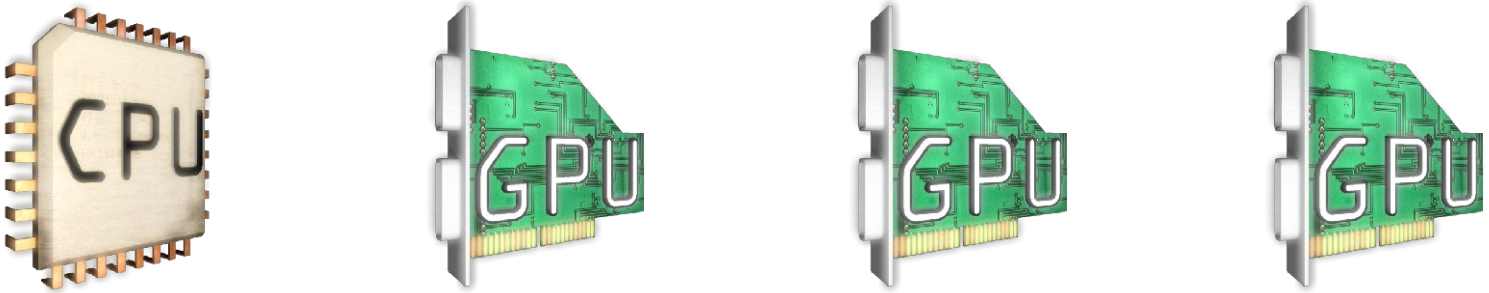
- Размеры GPU-блоков
- Тяжеловесность нитей
- ***Схема адресации памяти***



- Выбор устройств
- Размер теневых граней

Виды оптимизируемых параметров

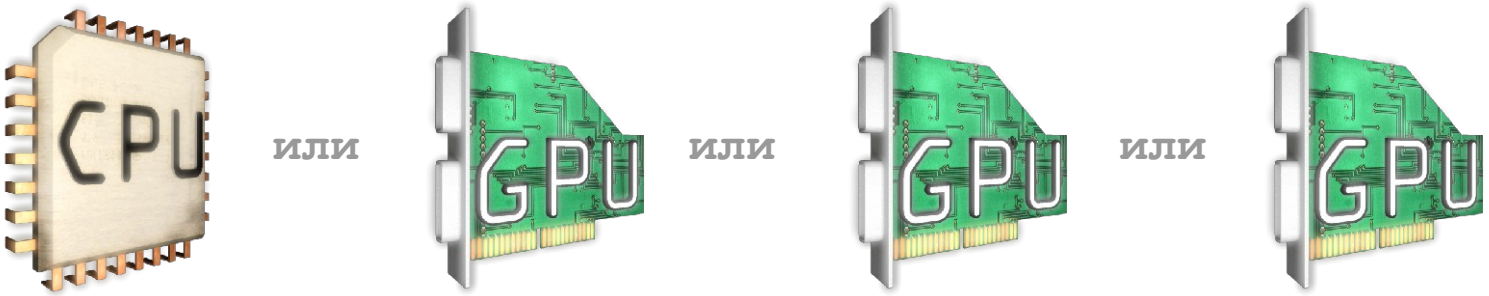
- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- ***Выбор устройств***



- Размер теневых граней

Виды оптимизируемых параметров

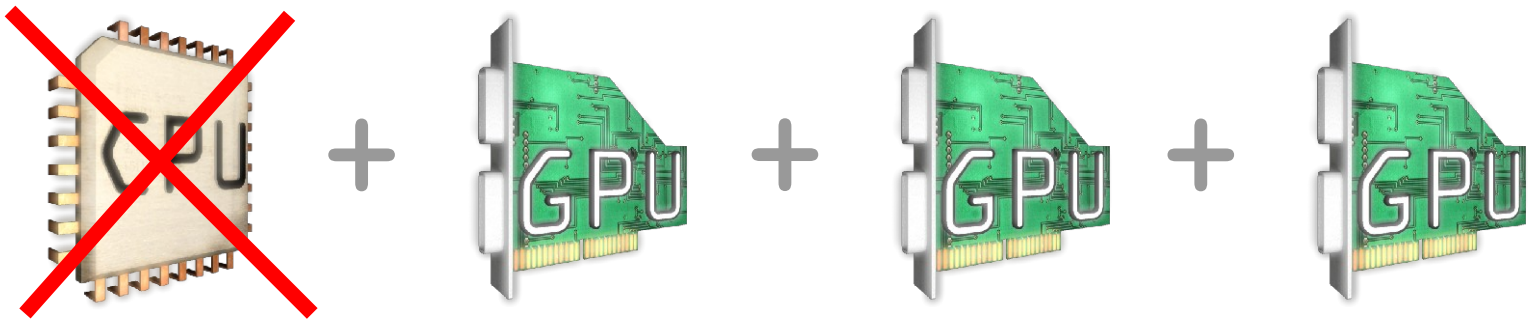
- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- ***Выбор устройств***



- Размер теневых граней

Виды оптимизируемых параметров

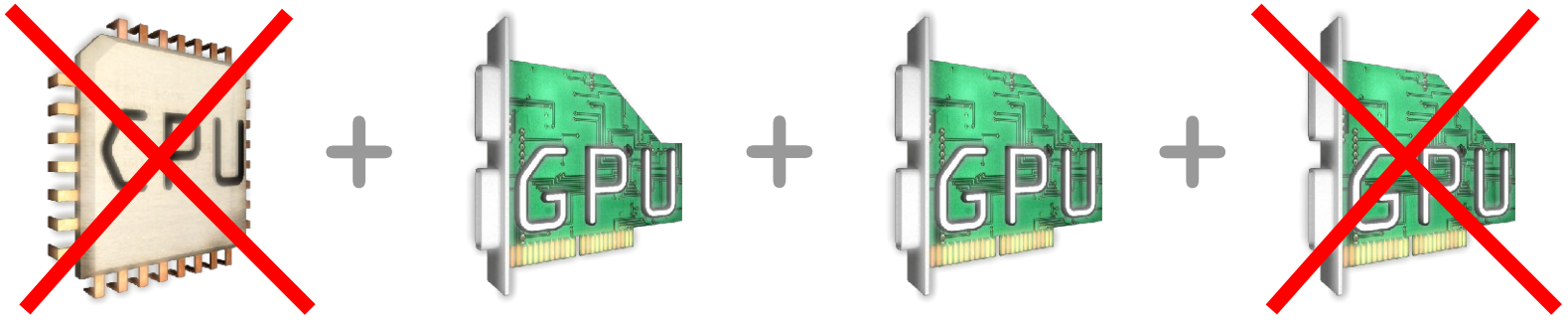
- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- ***Выбор устройств***



- Размер теневых граней

Виды оптимизируемых параметров

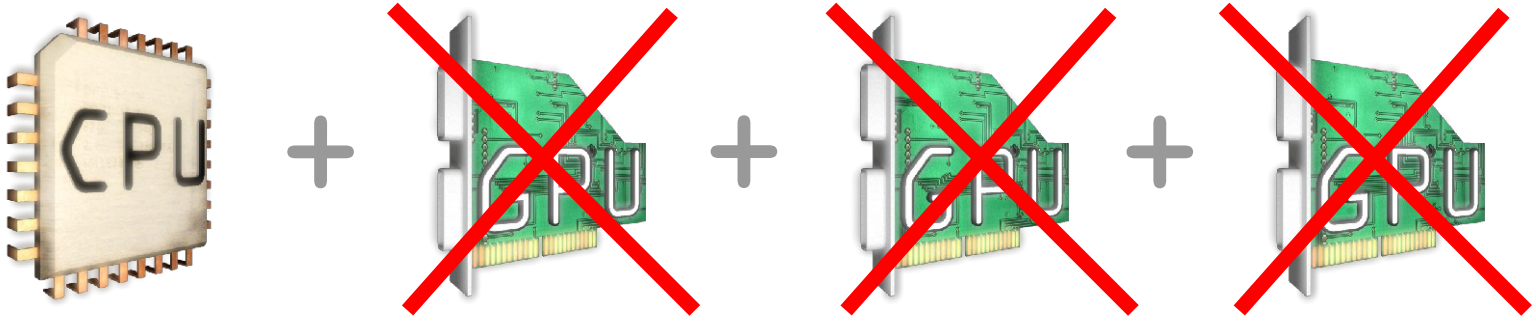
- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- ***Выбор устройств***



- Размер теневых граней

Виды оптимизируемых параметров

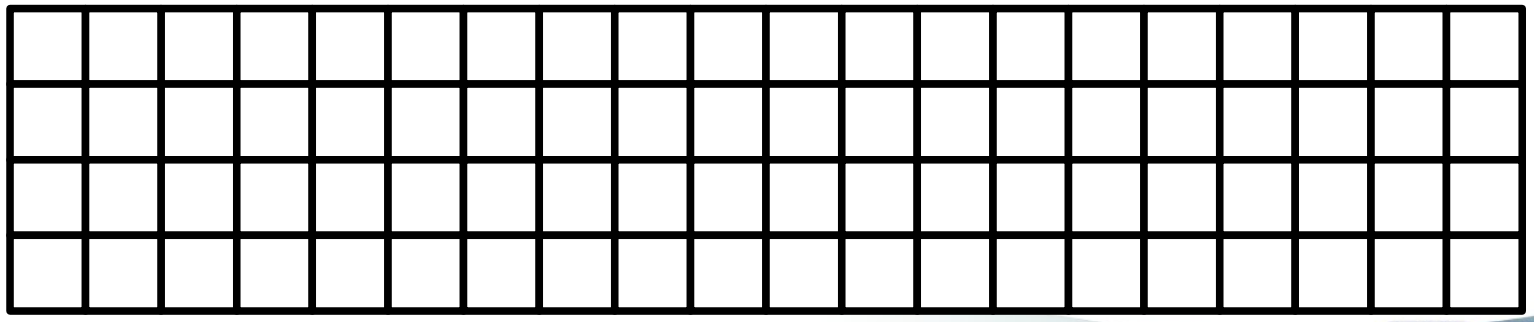
- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- ***Выбор устройств***



- Размер теневых граней

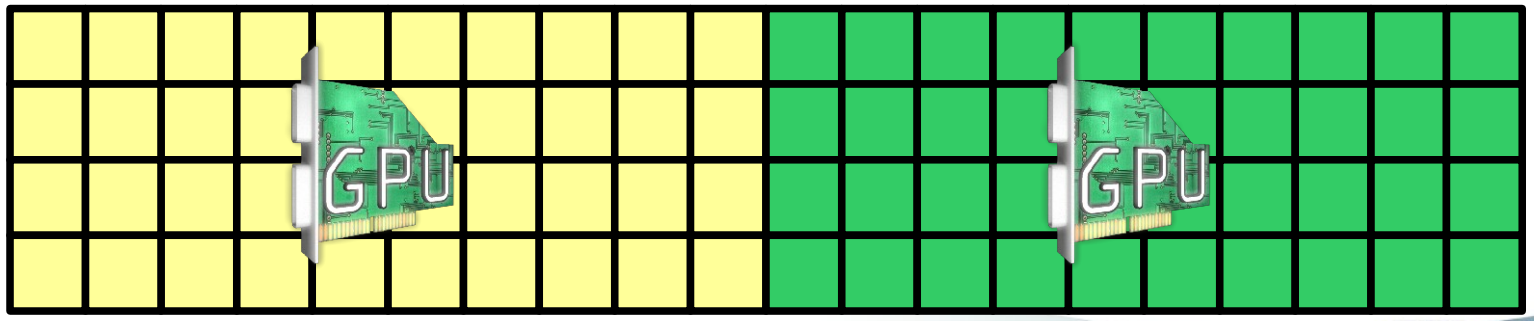
Виды оптимизируемых параметров

- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- Выбор устройств
- ***Размер теневых граней***



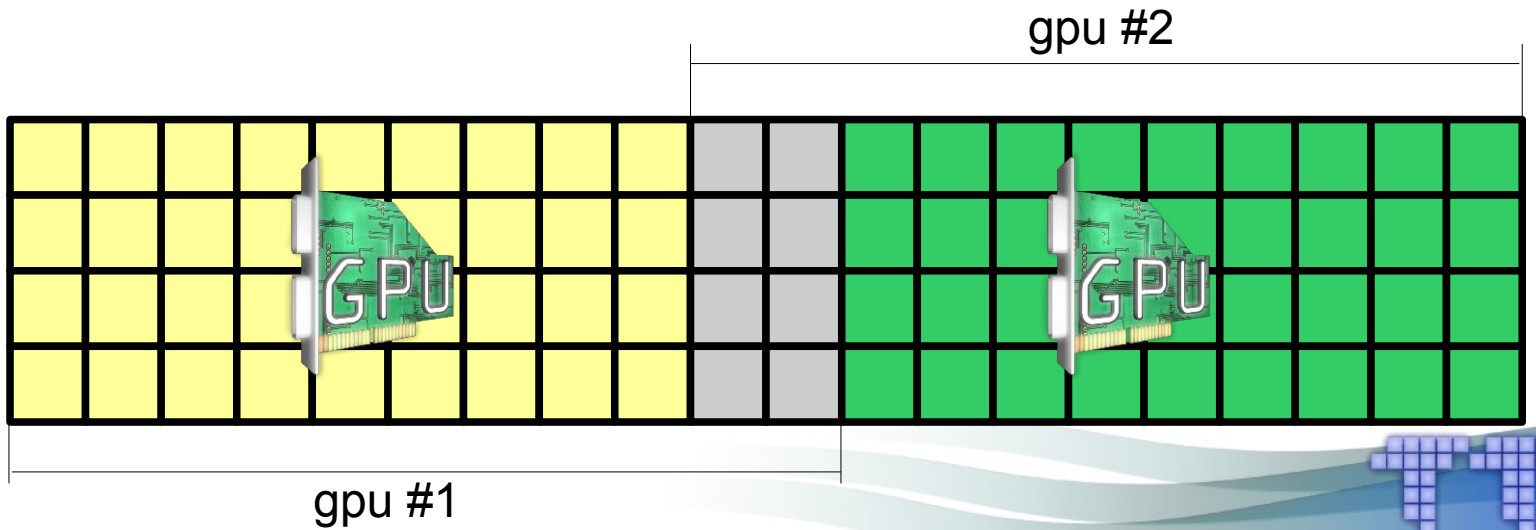
Виды оптимизируемых параметров

- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- Выбор устройств
- ***Размер теневых граней***



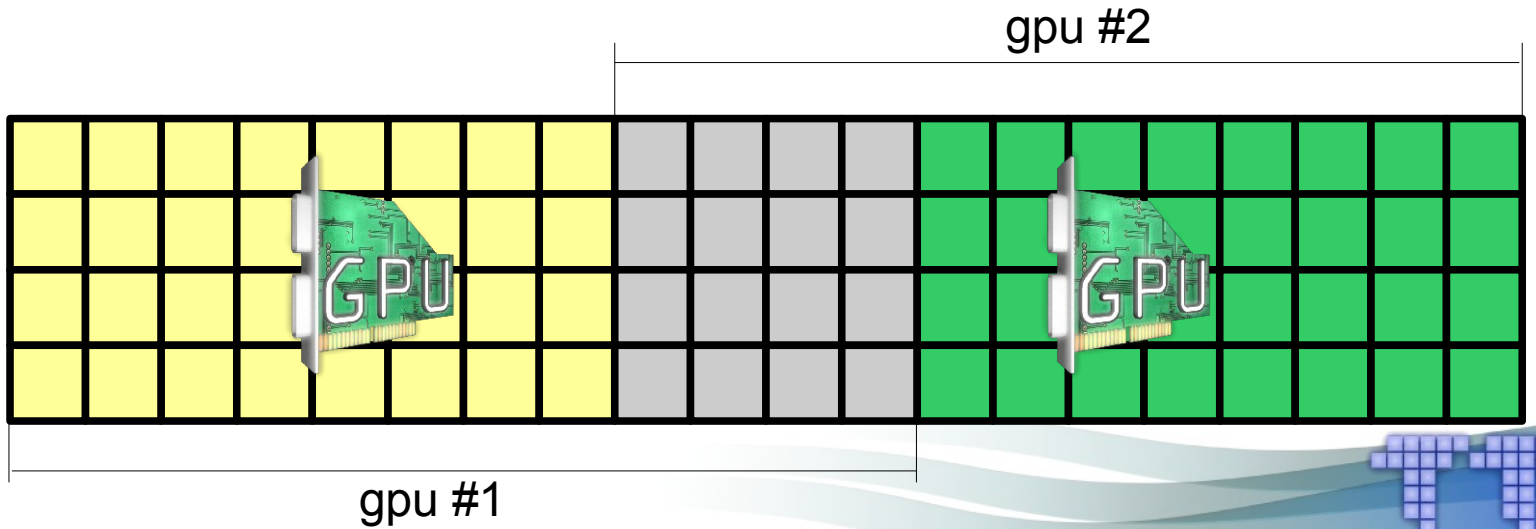
Виды оптимизируемых параметров

- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- Выбор устройств
- ***Размер теневых граней***



Виды оптимизируемых параметров

- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- Выбор устройств
- ***Размер теневых граней***



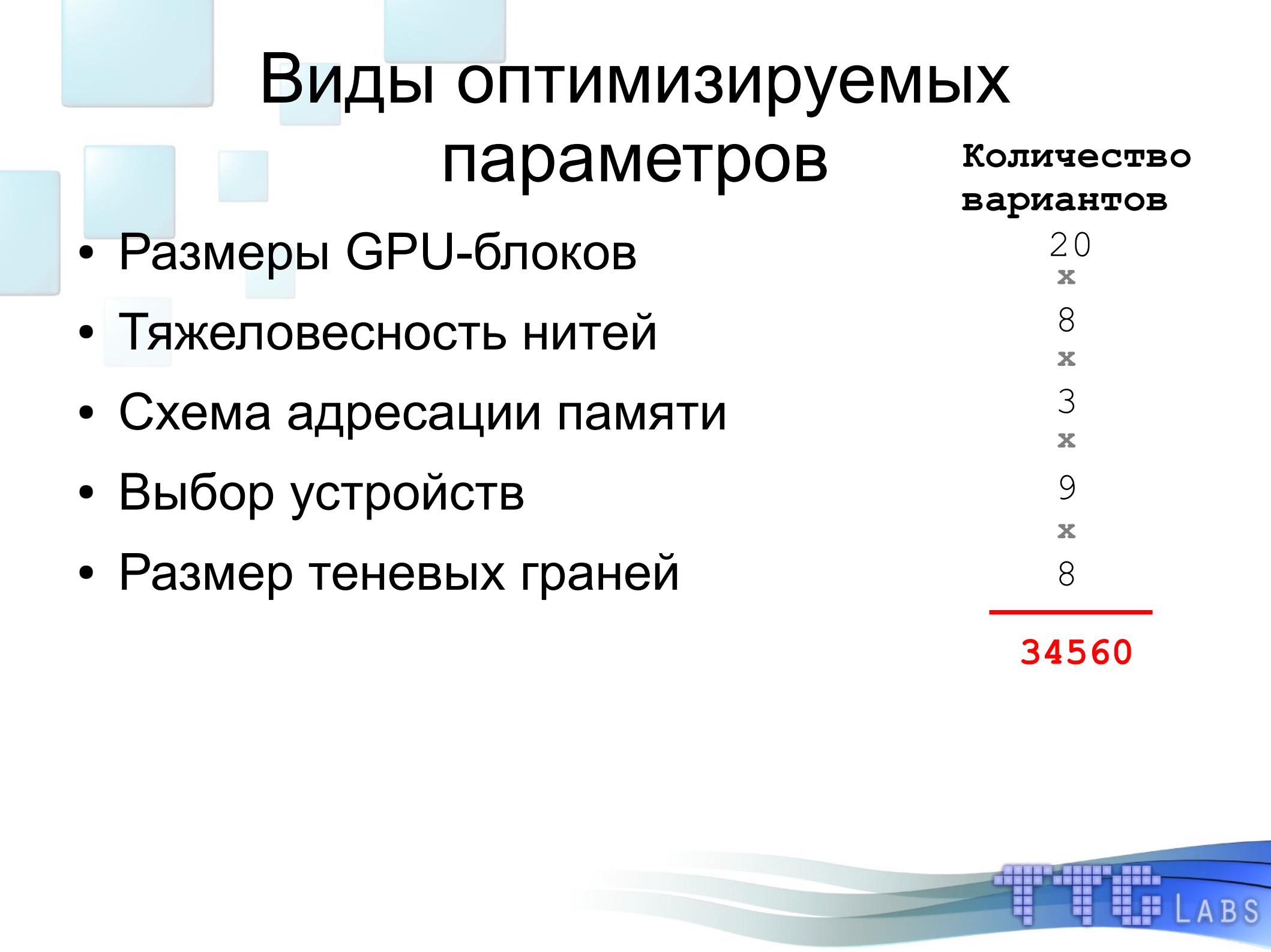
Виды оптимизируемых параметров

	Количество вариантов
• Размеры GPU-блоков	20
• Тяжеловесность нитей	8
• Схема адресации памяти	3
• Выбор устройств	9
• Размер теневых граней	8

Виды оптимизируемых параметров

- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- Выбор устройств
- Размер теневых граней

Количество вариантов
20
x
8
x
3
x
9
x
8
34560



Виды оптимизируемых параметров

- Размеры GPU-блоков
- Тяжеловесность нитей
- Схема адресации памяти
- Выбор устройств
- Размер теневых граней

Количество вариантов
20
x
8
x
3
x
9
x
8
34560

Проблема: нет «универсального» экстремума

Двумерная задача: уравнение теплопроводности

$$\frac{U}{\partial t} = \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2}$$

$$0 < x < 1$$

$$0 < y < 1$$

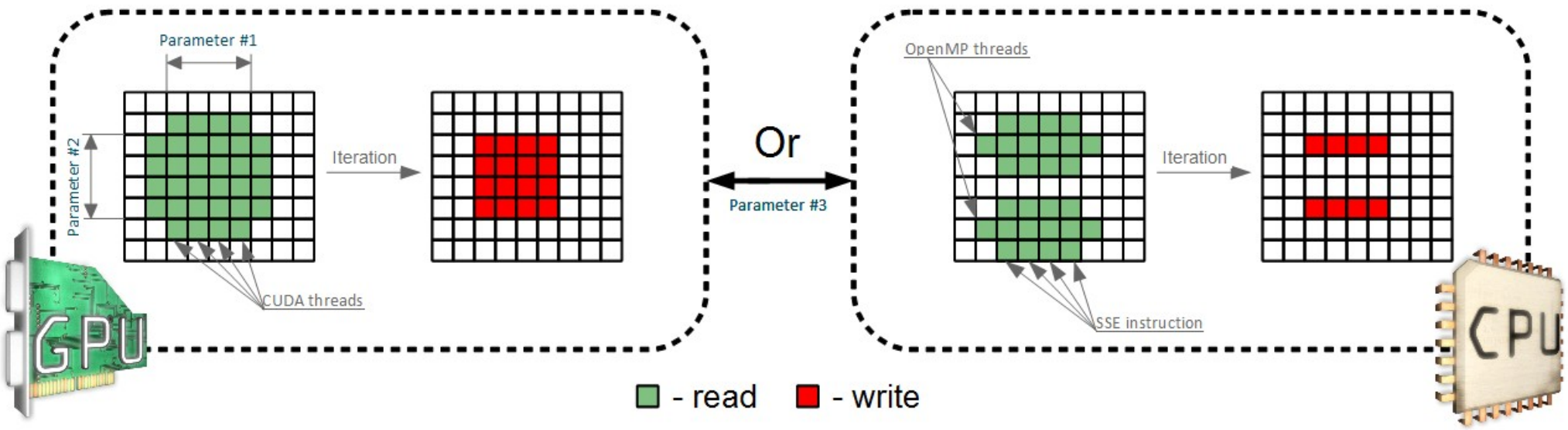
$$0 < t < T$$

$$U(x, y, t=0) = f(x, y)$$

$$U(x, y, t)|_{x=0,1} = \phi(y, t)$$

$$U(x, y, t)|_{y=0,1} = \theta(x, t)$$

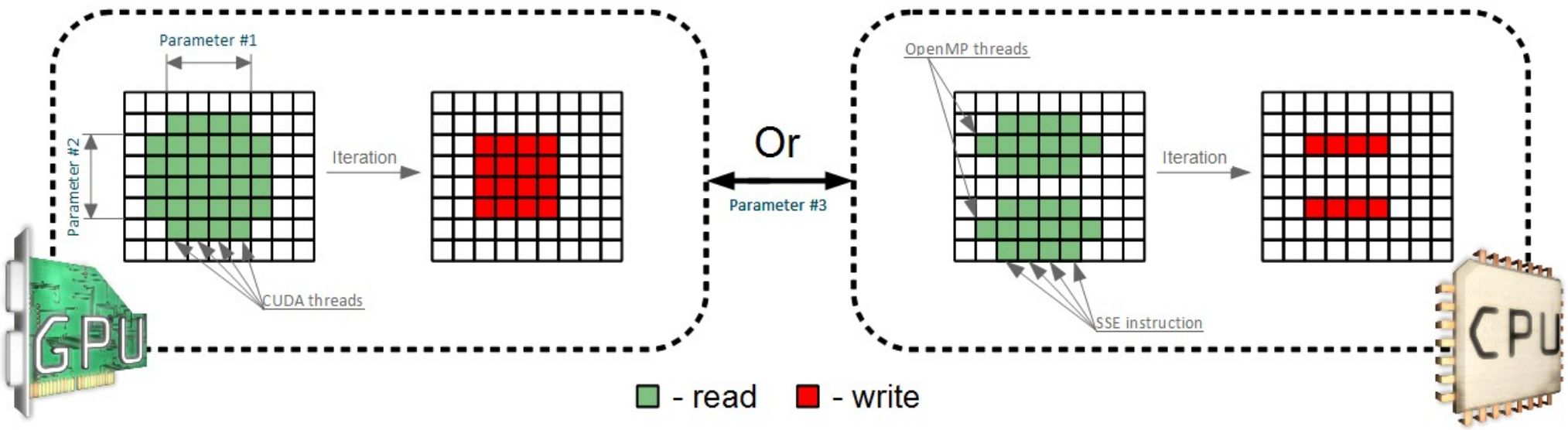
Двумерная задача: уравнение теплопроводности



Параметр 1. Высота CUDA-блока.
Параметр 2. Ширина CUDA-блока.
Параметр 3. Устройство (CPU или GPU).

Ограничение 1. Параметр 1 X Параметр 2 < N

Двумерная задача: уравнение теплопроводности



Параметр 1. Высота CUDA-блока. **=16**

Параметр 2. Ширина CUDA-блока. **=16**

Параметр 3. Устройство (CPU или GPU). **=GPU**

Ограничение 1. Параметр 1 X Параметр 2 < N

Двумерная задача: уравнение теплопроводности

Размер сетки	GF 8800	GF 9800	GF 285	GF 480	GF 580	GF 680
256x256	3,63+3,86%	3,32+0,60%	3,00+2,00%	8,12+7,02%	9,11+0,11%	7,84+21,56%
384x384	3,90+1,54%	3,76-0,27%	3,53-0,85%	9,22+12,26%	11,95+6,86%	10,24+12,79%
512x512	5,07+2,37%	4,83+0,83%	5,17+16,05%	9,49+26,98%	14,05+15,09%	12,49+18,33%
768x768	6,26+2,4%	6,11+0,65%	7,68+23,18%	10,66+40,24%	17,08+28,57%	15,00+39,47%
1024x1024	6,84+1,75%	6,66+0,30%	9,35+28,45%	11,28+44,15%	18,00+27,11%	16,29+36,22%
1536x1536	7,29+2,06%	6,68+3,14%	11,05+37,83%	11,8+42,20%	19,91+27,32%	17,57+33,92%
2048x2048	7,49+1,47%	6,88+2,62%	11,87+42,21%	11,89+42,39%	20,78+25,99%	18,31+31,95%
3072x3072	7,58+1,98%	6,82+2,79%	12,5+45,6%	11,99+43,70%	21,5+21,35%	18,66+32,10%
4096x4096	7,67+1,43%	6,75+3,41%	12,76+46,63%	11,95+46,44%	21,98+18,38%	18,79+30,81%
6144x6144	7,59+2,64%	6,72+3,72%	12,93+46,64%	11,92+46,06%	21,98+17,93%	18,63+32,37%
Среднее ускорение	+ 2,14%	+ 1,78%	+ 28,77%	+ 35,14%	+ 18,87%	+ 28,95%

Используется только GPU

Двумерная задача: уравнение теплопроводности

Размер сетки	GF 8800	GF 9800	GF 285	GF 480	GF 580	GF 680
256x256	3,63+3,86%	3,32+0,60%	3,00+2,00%	8,12+7,02%	9,11+0,11%	7,84+21,56%
384x384	3,90+1,54%	3,76-0,27%	3,53-0,85%	9,22+12,26%	11,95+6,86%	10,24+12,79%
512x512	5,07+2,37%	4,83+0,83%	5,17+16,05%	9,49+26,98%	14,05+15,09%	12,49+18,33%
768x768	6,26+2,4%	6,11+0,65%	7,68+23,18%	10,66+40,24%	17,08+28,57%	15,00+39,47%
1024x1024	6,84+1,75%	6,66+0,30%	9,35+28,45%	11,28+44,15%	18,00+27,11%	16,29+36,22%
1536x1536	7,29+2,06%	6,68+3,14%	11,05+37,83%	11,8+42,20%	19,91+27,32%	17,57+33,92%
2048x2048	7,49+1,47%	6,88+2,62%	11,87+42,21%	11,89+42,39%	20,78+25,99%	18,31+31,95%
3072x3072	7,58+1,98%	6,82+2,79%	12,5+45,6%	11,99+43,70%	21,5+21,35%	18,66+32,10%
4096x4096	7,67+1,43%	6,75+3,41%	12,76+46,63%	11,95+46,44%	21,98+18,38%	18,79+30,81%
6144x6144	7,59+2,64%	6,72+3,72%	12,93+46,64%	11,92+46,06%	21,98+17,93%	18,63+32,37%
Среднее ускорение	+ 2,14%	+ 1,78%	+ 28,77%	.+ 35,14%	+ 18,87%	+ 28,95%

Используется только GPU

Чёрный - ускорение до 10%
Синий - ускорение от **10%** до **20%**
Зелёный - ускорение от **20%** до **30%**
Жёлтый - ускорение от **30%** до **40%**
Красный - ускорение более **40%**

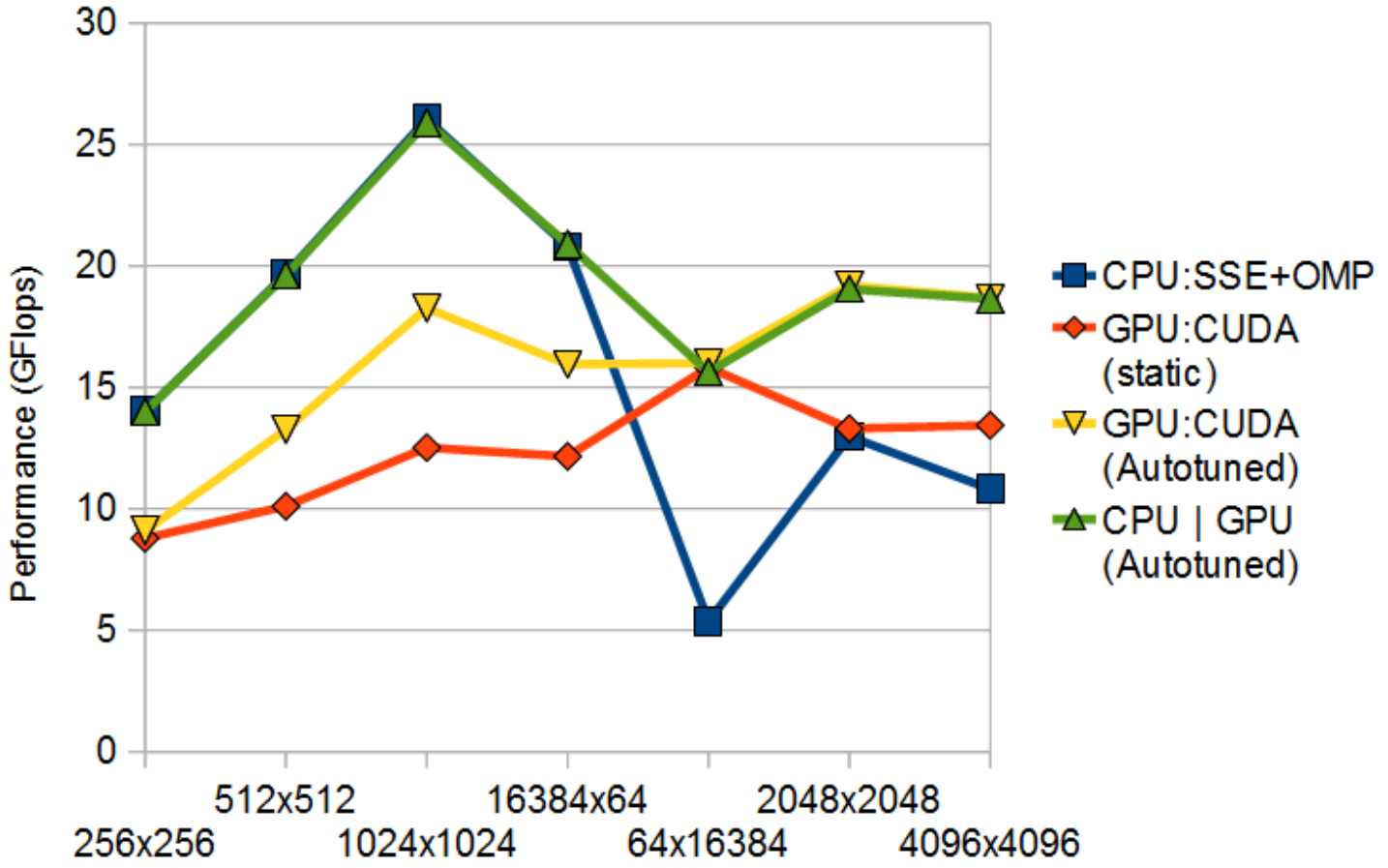
Двумерная задача: уравнение теплопроводности

Размер сетки	GF 8800	GF 9800	GF 285	GF 480	GF 580	GF 680
256x256	3,63+127,83%	3,32+156,10%	3,00+193,23%	8,12+7,02%	9,11+0,11%	7,84+21,56%
384x384	3,90+251,26%	3,76+244,85%	3,53+273,19%	9,22+46,75%	11,95+8,76%	10,24+12,79%
512x512	5,07+193,95%	4,83+212,99%	5,17+193,91%	9,49+60,38%	14,05+15,09%	12,49+18,33%
768x768	6,26+172,93%	6,11+177,40%	7,68+117,97%	10,66+61,73%	17,08+28,57%	15,00+39,47%
1024x1024	6,84+142,39%	6,66+162,30%	9,35+82,96%	11,28+49,20%	18,00+27,11%	16,29+36,22%
1536x1536	7,29+23,47%	6,68+42,714%	11,05+37,83%	11,8+42,20%	19,91+27,32%	17,57+33,92%
2048x2048	7,49+27,82%	6,88+35,16%	11,87+42,21%	11,89+42,39%	20,78+25,99%	18,31+31,95%
3072x3072	7,58+26,12%	6,82+38,70%	12,5+45,60%	11,99+43,70%	21,5+21,35%	18,66+32,10%
4096x4096	7,67+22,36%	6,75+44,06%	12,76+46,63%	11,95+46,44%	21,98+18,38%	18,79+30,81%
6144x6144	7,59+37,71%	6,72+48,68%	12,93+46,64%	11,92+46,06%	21,98+17,93%	18,63+32,37%
Среднее ускорение	.+ 102,59%	116,30%	.+ 108,02%	.+ 44,59%	+ 19,06%	.+ 28,95%

Используются GPU и CPU

Чёрный - ускорение до 10%
Синий - ускорение от **10%** до **20%**
Зелёный - ускорение от **20%** до **30%**
Жёлтый - ускорение от **30%** до **40%**
Красный - ускорение более **40%**

Двумерная задача: уравнение теплопроводности



Система: 2x ***Intel*** Xeon 5650 + 3x ***NVidia*** Tesla C2050

Двумерная задача: уравнение теплопроводности

Размер сетки	GF 8800	GF 9800	GF 285	GF 480	GF 580	GF 680
256x256	3,63+127,83%	3,32+156,10%	3,00+193,23%	8,12+7,02%	9,11+0,11%	7,84+21,56%
384x384	3,90+251,26%	3,76+244,85%	3,53+273,19%	9,22+46,75%	11,95+8,76%	10,24+12,79%
512x512	5,07+193,95%	4,83+212,99%	5,17+193,91%	9,49+60,38%	14,05+15,09%	12,49+18,33%
768x768	6,26+172,93%	6,11+177,40%	7,68+117,97%	10,66+61,73%	17,08+28,57%	15,00+39,47%
1024x1024	6,84+142,39%	6,66+162,30%	9,35+82,96%	11,28+49,20%	18,00+27,11%	16,29+36,22%
1536x1536	7,29+23,47%	6,68+42,714%	11,05+37,83%	11,8+42,20%	19,91+27,32%	17,57+33,92%
2048x2048	7,49+27,82%	6,88+35,16%	11,87+42,21%	11,89+42,39%	20,78+25,99%	18,31+31,95%
3072x3072	7,58+26,12%	6,82+38,70%	12,5+45,60%	11,99+43,70%	21,5+21,35%	18,66+32,10%
4096x4096	7,67+22,36%	6,75+44,06%	12,76+46,63%	11,95+46,44%	21,98+18,38%	18,79+30,81%
6144x6144	7,59+37,71%	6,72+48,68%	12,93+46,64%	11,92+46,06%	21,98+17,93%	18,63+32,37%
Среднее ускорение	.+ 102,59%	116,30%	.+ 108,02%	.+ 44,59%	+ 19,06%	.+ 28,95%

Используются GPU и CPU

Чёрный - ускорение до 10%
Синий - ускорение от **10%** до **20%**
Зелёный - ускорение от **20%** до **30%**
Жёлтый - ускорение от **30%** до **40%**
Красный - ускорение более **40%**

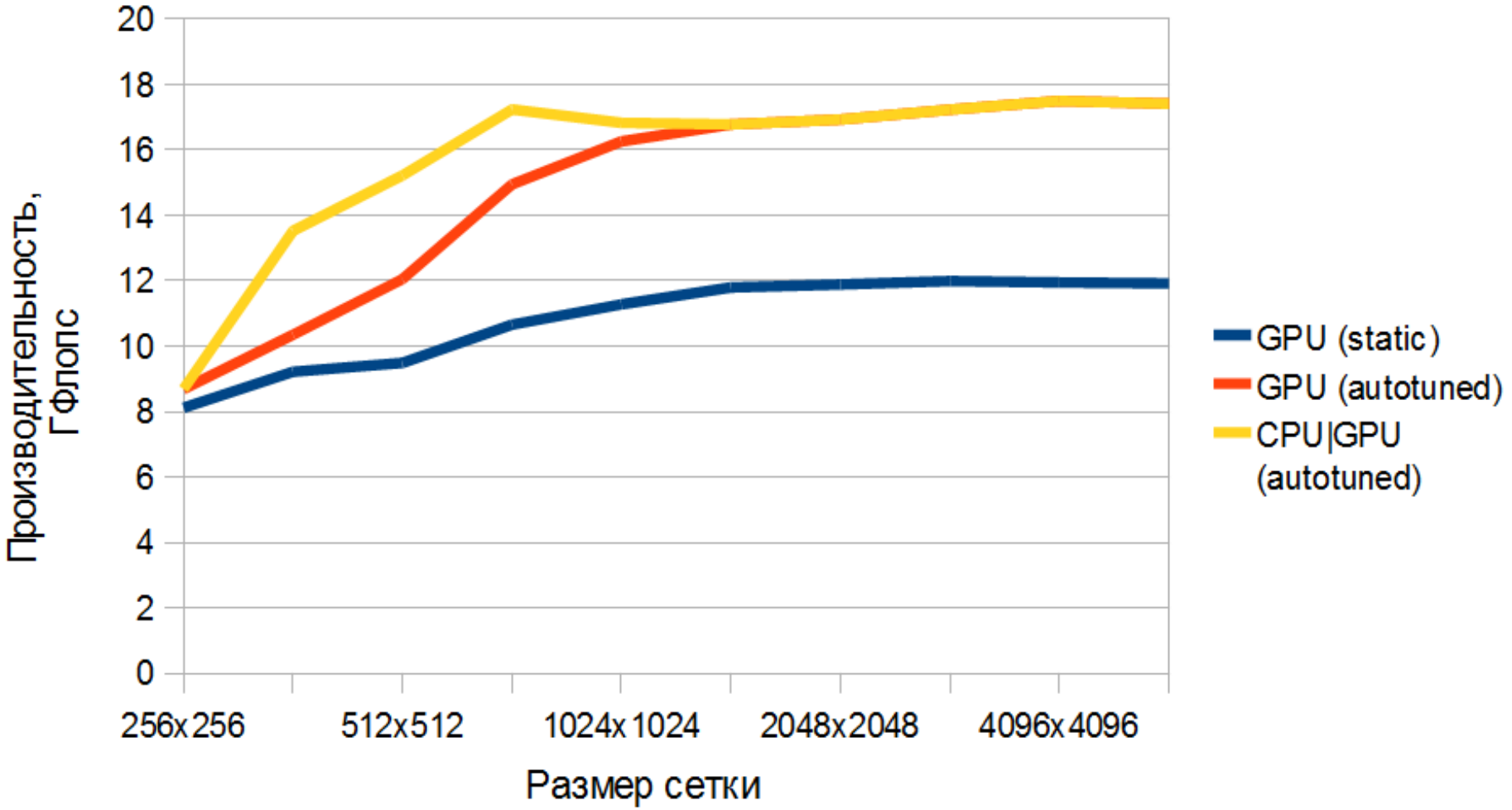
Двумерная задача: уравнение теплопроводности

Размер сетки	GF 8800	GF 9800	GF 285	GF 480	GF 580	GF 680
256x256	3,63+127,83%	3,32+156,10%	3,00+193,23%	8,12+7,02%	9,11+0,11%	7,84+21,56%
384x384	3,90+251,26%	3,76+244,85%	3,53+273,19%	9,22+46,75%	11,95+8,76%	10,24+12,79%
512x512	5,07+193,95%	4,83+212,99%	5,17+193,91%	9,49+60,38%	14,05+15,09%	12,49+18,33%
768x768	6,26+172,93%	6,11+177,40%	7,68+117,97%	10,66+61,73%	17,08+28,57%	15,00+39,47%
1024x1024	6,84+142,39%	6,66+162,30%	9,35+82,96%	11,28+49,20%	18,00+27,11%	16,29+36,22%
1536x1536	7,29+23,47%	6,68+42,714%	11,05+37,83%	11,8+42,20%	19,91+27,32%	17,57+33,92%
2048x2048	7,49+27,82%	6,88+35,16%	11,87+42,21%	11,89+42,39%	20,78+25,99%	18,31+31,95%
3072x3072	7,58+26,12%	6,82+38,70%	12,5+45,60%	11,99+43,70%	21,5+21,35%	18,66+32,10%
4096x4096	7,67+22,36%	6,75+44,06%	12,76+46,63%	11,95+46,44%	21,98+18,38%	18,79+30,81%
6144x6144	7,59+37,71%	6,72+48,68%	12,93+46,64%	11,92+46,06%	21,98+17,93%	18,63+32,37%
Среднее ускорение	.+ 102,59%	116,30%	.+ 108,02%	.+ 44,59%	+ 19,06%	.+ 28,95%

Используются GPU и CPU

Чёрный - ускорение до 10%
Синий - ускорение от **10%** до **20%**
Зелёный - ускорение от **20%** до **30%**
Жёлтый - ускорение от **30%** до **40%**
Красный - ускорение более **40%**

Двумерная задача: уравнение теплопроводности



Система: ***Intel*** Xeon E3-1230 + ***NVidia*** GeForce 480 GTX

Трёхмерная задача: уравнение Пуассона

$$\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}+\frac{\partial^2 U}{\partial z^2}=f(x,y,z)$$

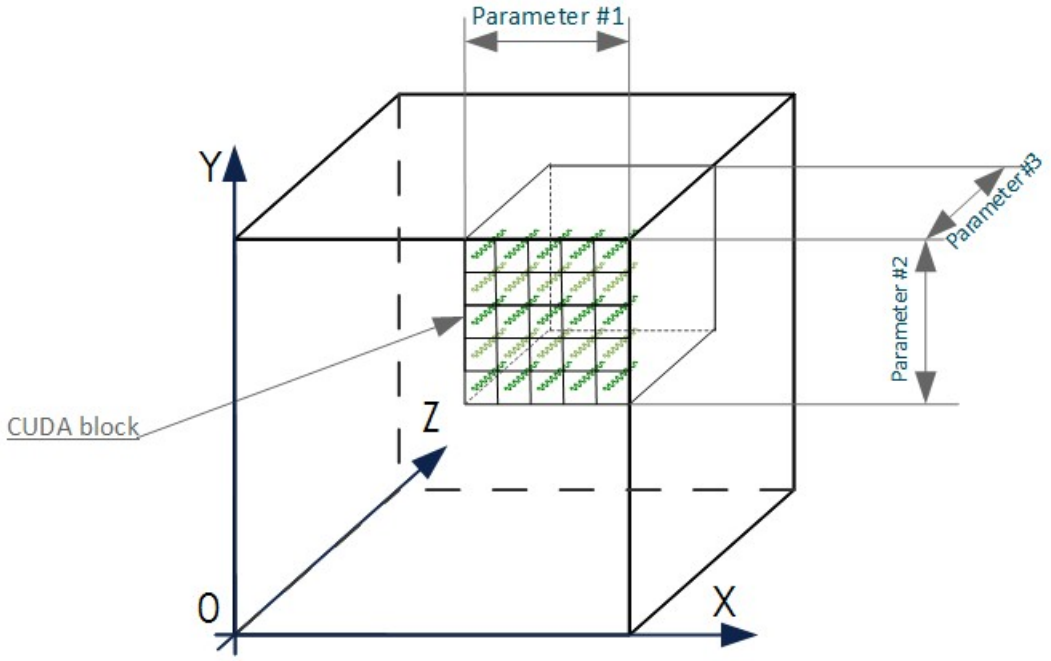
$$0<x<1$$

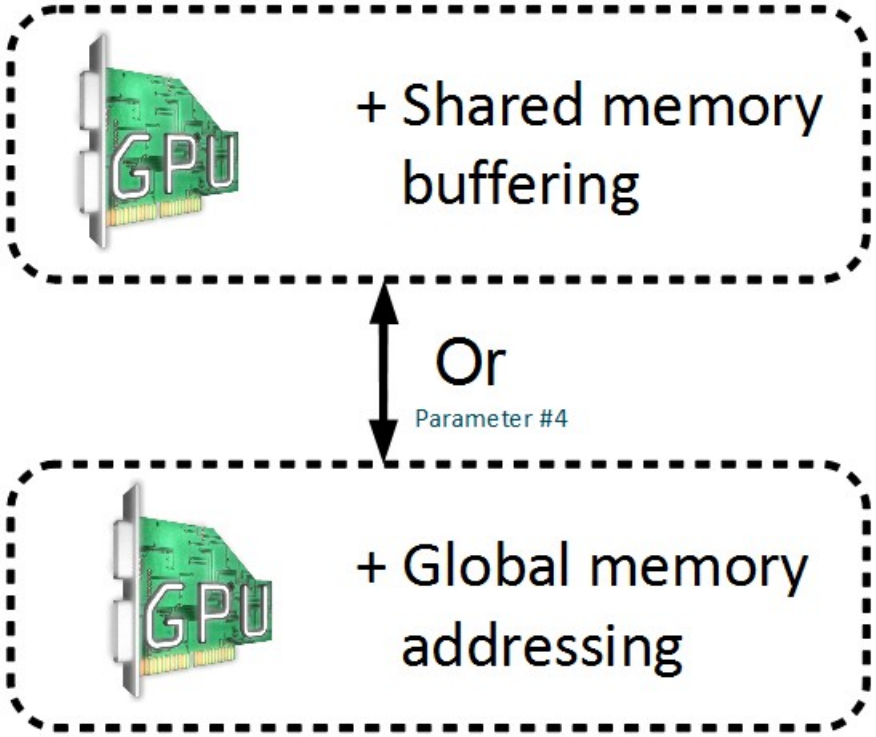
$$0<y<1$$

$$0<z<1$$

$$U(x,y,z)|_{(x,y,z)\in G}=g(x,y,z)$$

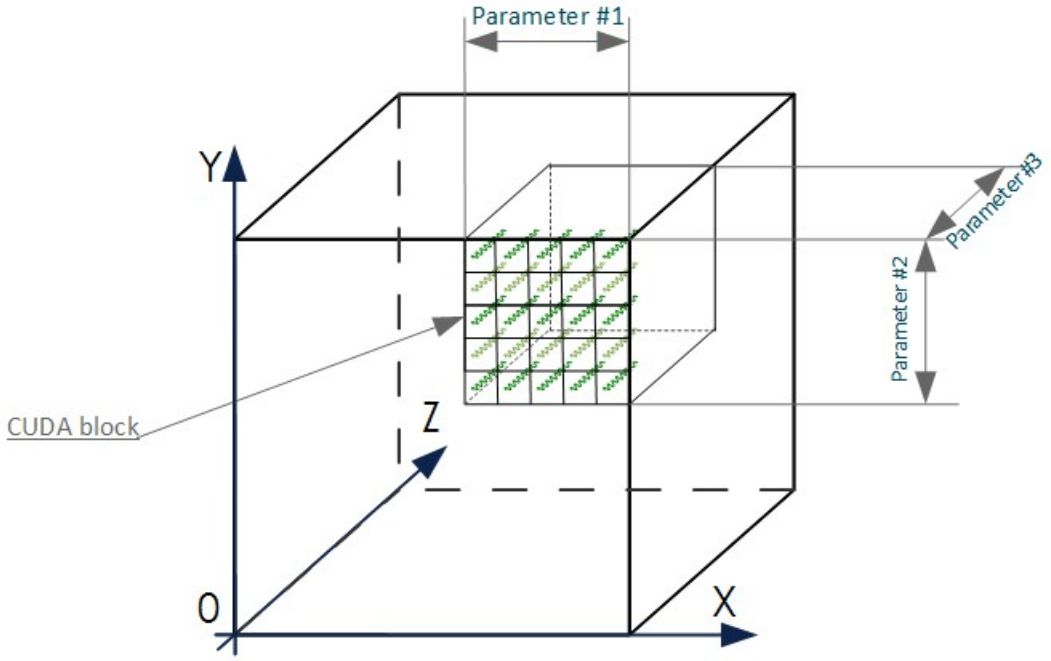
Трёхмерная задача: уравнение Пуассона

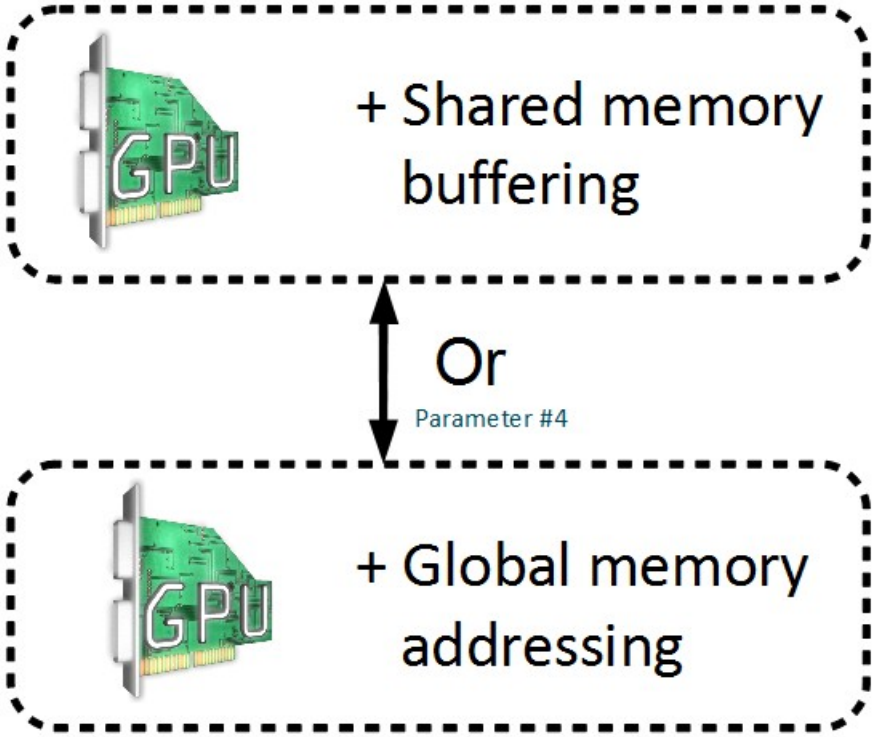




Параметр 1. Высота CUDA-блока
Параметр 2. Ширина CUDA-блока
Параметр 3. «Тяжеловесность» нитей
Параметр 4. Тип памяти (shared или global)

Трёхмерная задача: уравнение Пуассона





Параметр 1. Высота CUDA-блока **=16**

Параметр 2. Ширина CUDA-блока **=16**

Параметр 3. **«**Тяжеловесность» нитей **=4**

Параметр 4. Тип памяти (shared или global) **=shared**

Двумерная задача: уравнение теплопроводности

Размер сетки	GF 8800	GF 9800	GF 285	GF 480	GF 580	GF 680
64x64x64	9.7 + 10.1%	11.7 + 0.4%	23.1 + 4.1%	54.2 + 3.0%	61.1 + 17.9%	55.9 + 9.9%
96x96x96	10.1 + 18.5%	12.3 + 9.2%	29.9 + 1.9%	59.0 + 22.9%	69.3 + 25.0%	60.3 + 15.2%
128x128x128	10.2 + 16.1%	9.4 + 30.4%	23.1 + 4.2%	55.2 + 25.4%	63.5 + 28.6%	47.7 + 49.2%
160x160x160	10.3 + 21.2%	11.7 + 9.6%	25.5 + 16.8%	61.5 + 22.7%	71.4 + 33.1%	54.9 + 33.2%
192x192x192	10.2 + 21.2%	10.8 + 12.0%	25.8 + 6.1%	60.5 + 22.7%	70.4 + 33.5%	49.1 + 48.9%
Среднее ускорение	+17.4%	+12.3%	+6.6%	+19.4%	+27.6%	+31.3%

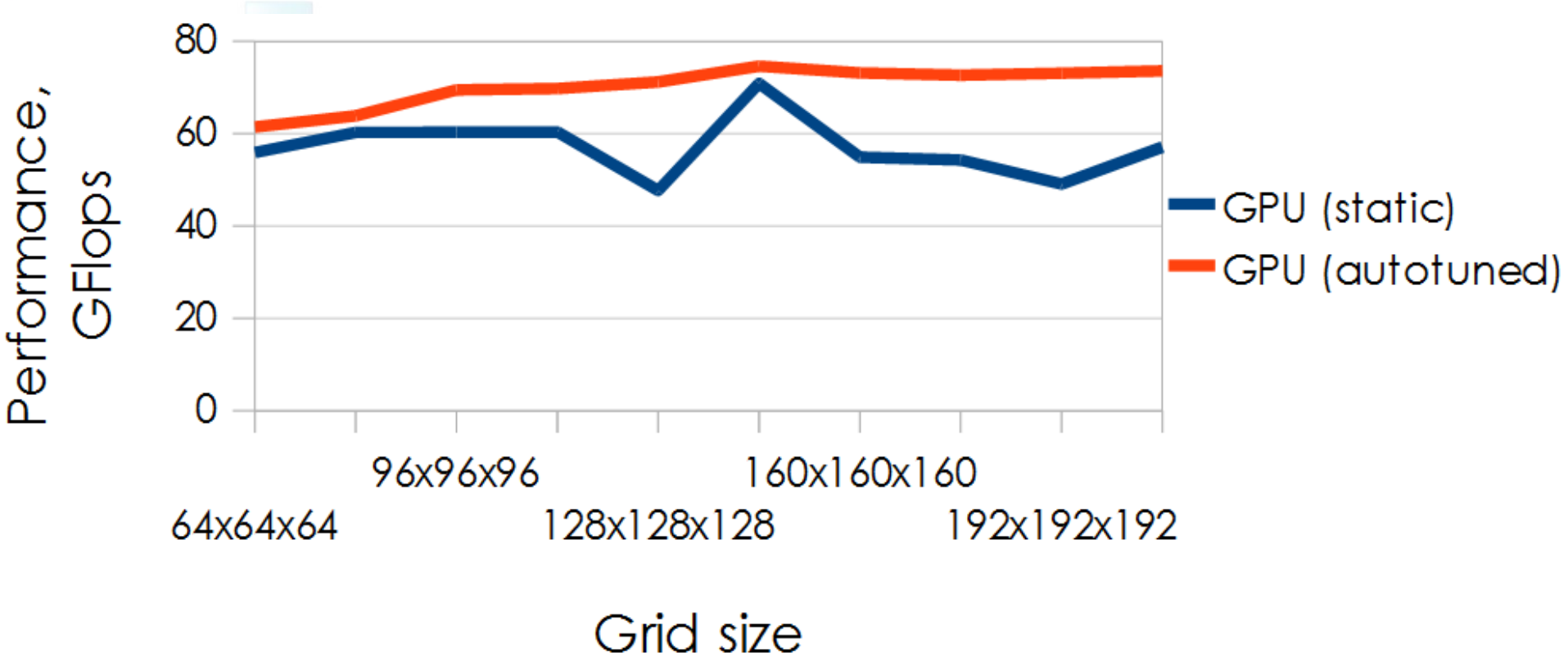
Чёрный - ускорение до 10%
Синий - ускорение от **10%** до **20%**
Зелёный - ускорение от **20%** до **30%**
Жёлтый - ускорение от **30%** до **40%**
Красный - ускорение более **40%**

Двумерная задача: уравнение теплопроводности

Размер сетки	GF 8800	GF 9800	GF 285	GF 480	GF 580	GF 680
64x64x64	9.7 + 10.1%	11.7 + 0.4%	23.1 + 4.1%	54.2 + 3.0%	61.1 + 17.9%	55.9 + 9.9%
96x96x96	10.1 + 18.5%	12.3 + 9.2%	29.9 + 1.9%	59.0 + 22.9%	69.3 + 25.0%	60.3 + 15.2%
128x128x128	10.2 + 16.1%	9.4 + 30.4%	23.1 + 4.2%	55.2 + 25.4%	63.5 + 28.6%	47.7 + 49.2%
160x160x160	10.3 + 21.2%	11.7 + 9.6%	25.5 + 16.8%	61.5 + 22.7%	71.4 + 33.1%	54.9 + 33.2%
192x192x192	10.2 + 21.2%	10.8 + 12.0%	25.8 + 6.1%	60.5 + 22.7%	70.4 + 33.5%	49.1 + 48.9%
Среднее ускорение	+17.4%	+12.3%	+6.6%	+19.4%	+27.6%	+31.3%

Чёрный - ускорение до 10%
Синий - ускорение от **10%** до **20%**
Зелёный - ускорение от **20%** до **30%**
Жёлтый - ускорение от **30%** до **40%**
Красный - ускорение более **40%**

Двумерная задача: уравнение теплопроводности



Система: ***Intel*** Xeon E3-1230 + ***NVidia*** GeForce 680 GTX

Мораль

- Даже после «ручной» оптимизации в программах есть потенциал для ускорения
- Концепция автоадаптации программ применима к некоторым задачам из CFD
- На рассмотренных модельных задачах удалось получить ускорение от 20% до 50%

Вопросы?

(m_krivov@ttgLabs.com)