Программная модель и оценка производительности для гетерогенных систем с графическими процессорами

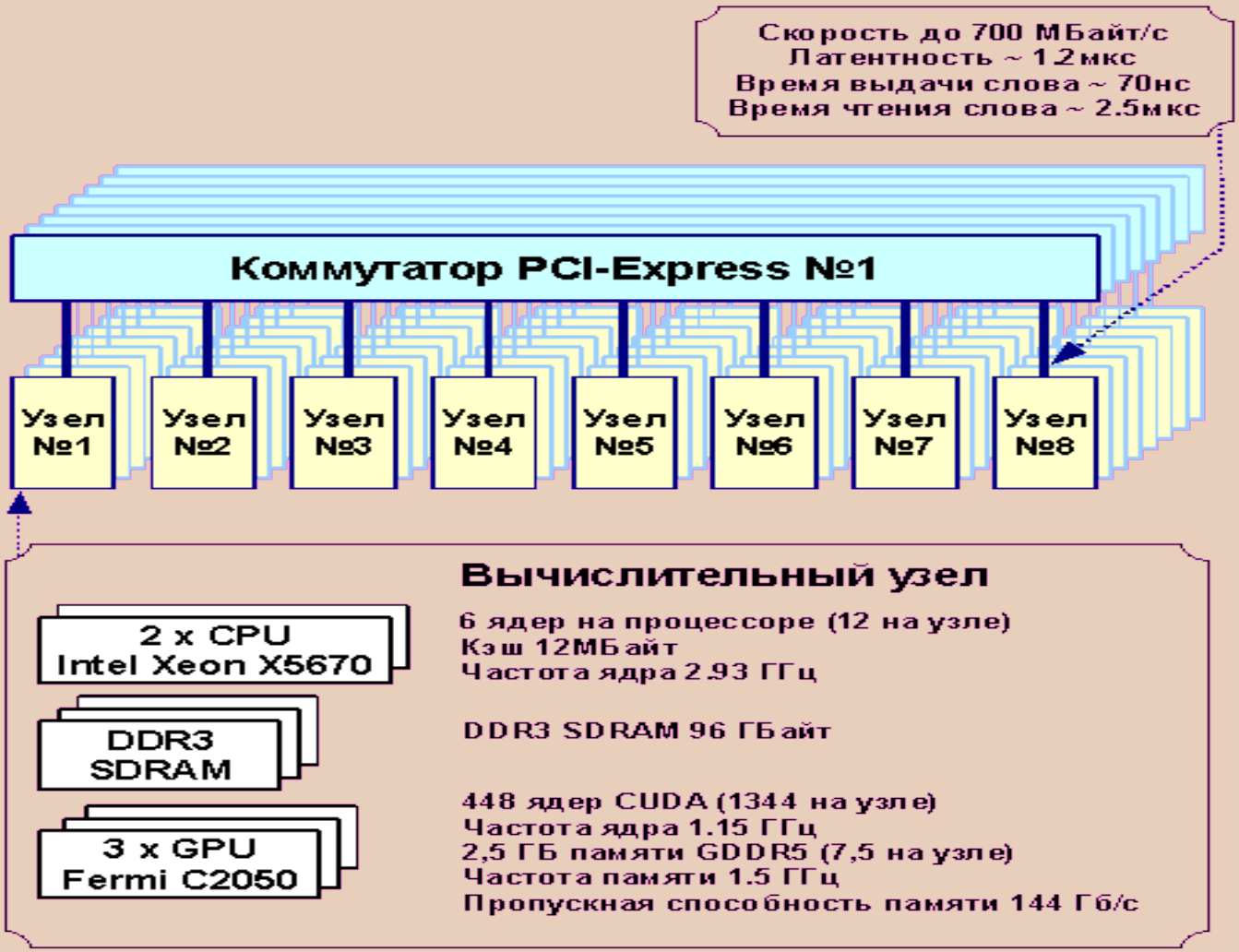
И.А. Горячев, В.Д. Левченко, А.Ю. Перепёлкина

ИПМ им. М.В. Келдыша РАН

2012

Гибридный вычислительный кластер K-100

k100.kiam.ru

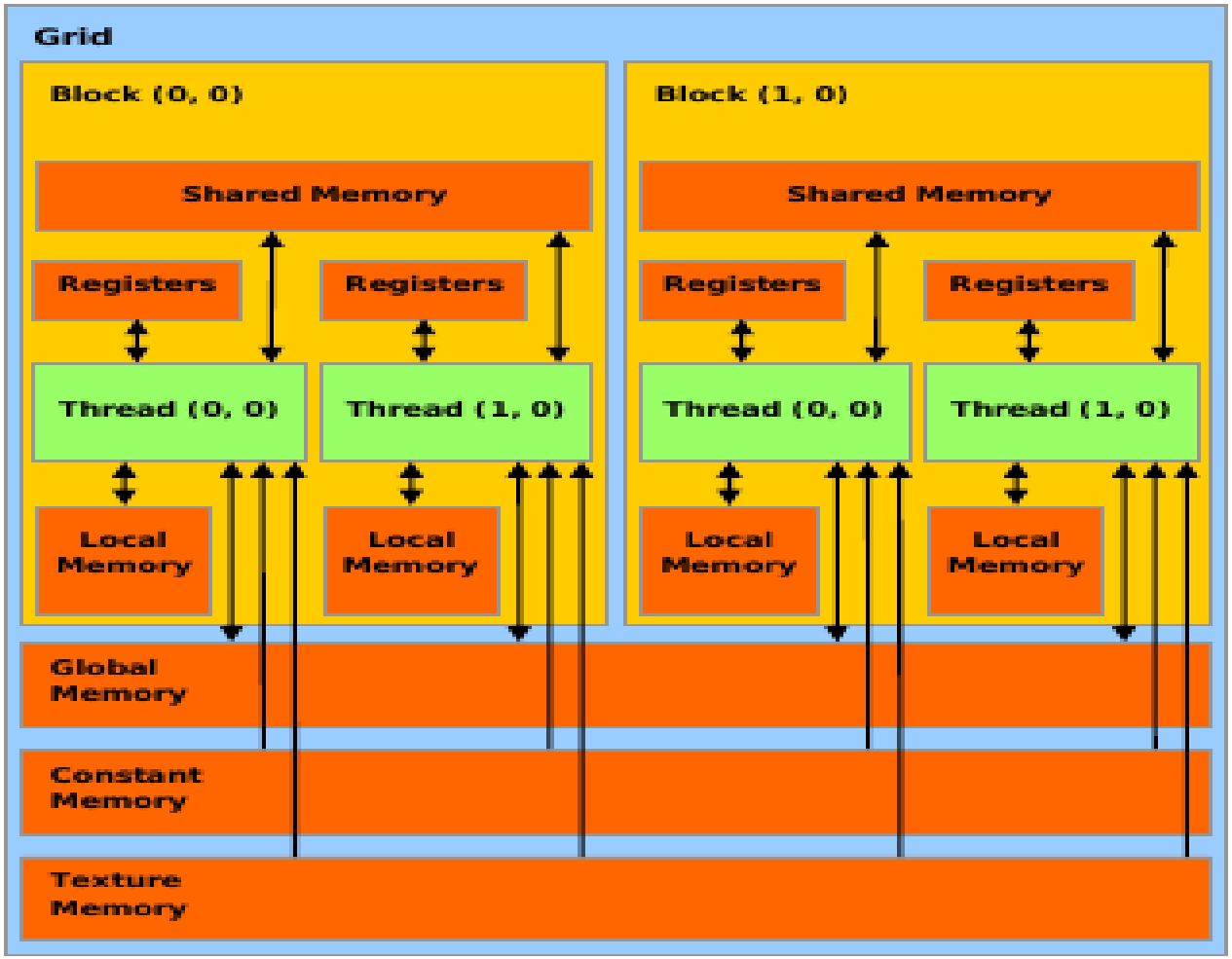


Вычислительные HW-уровни

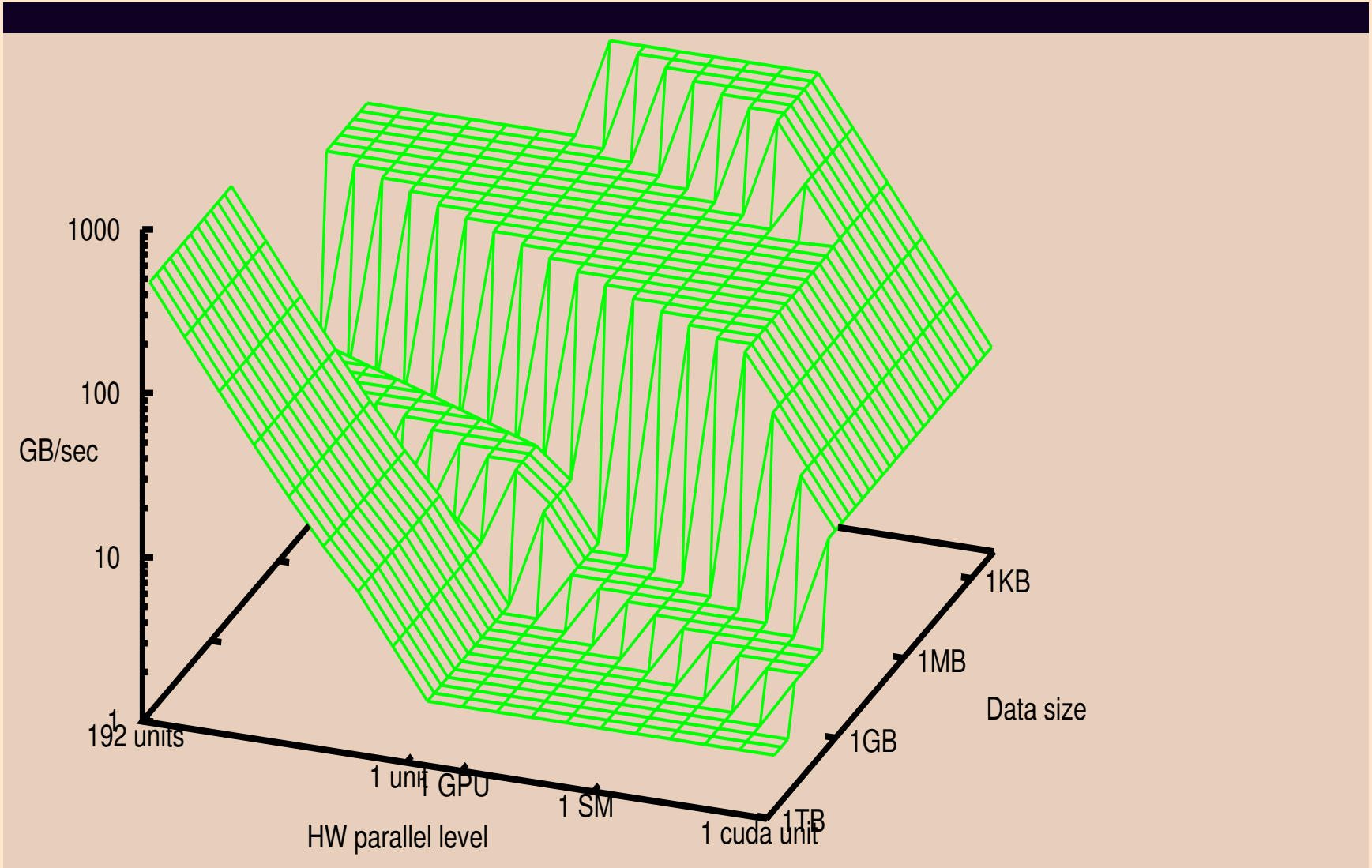
Архитектура Fermi

Иерархия памяти в устройствах с графическими процессорами

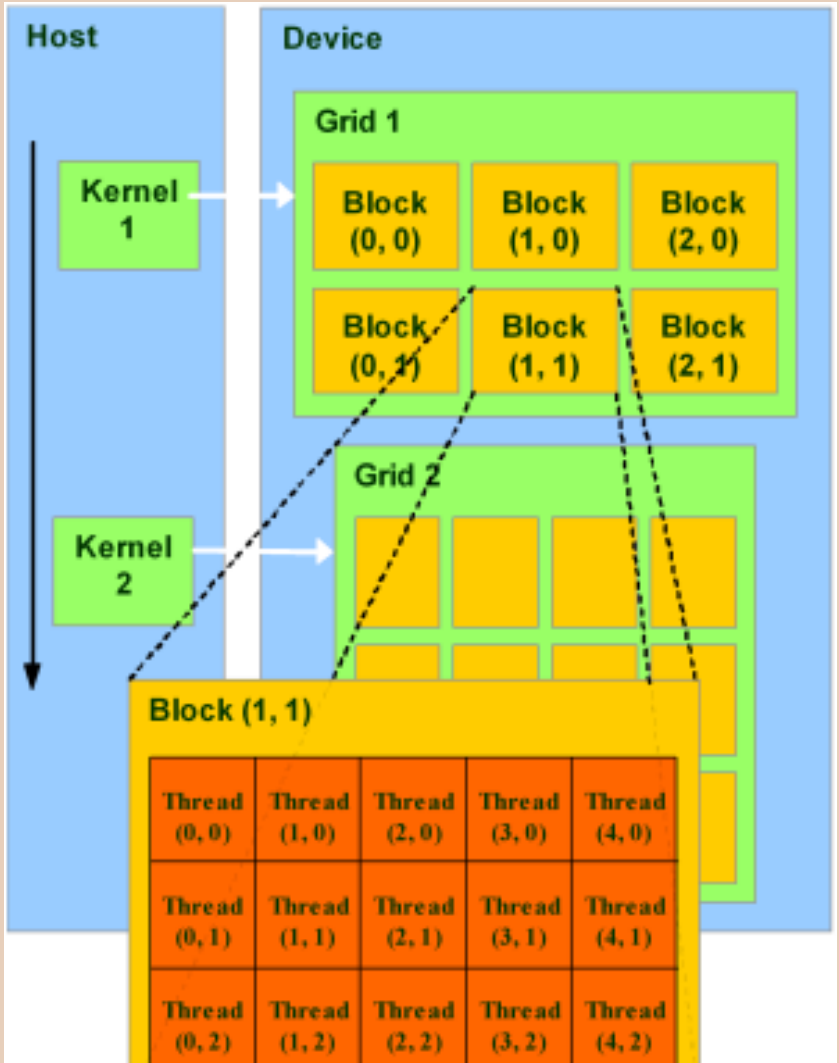
Комплекс Tesla 2050



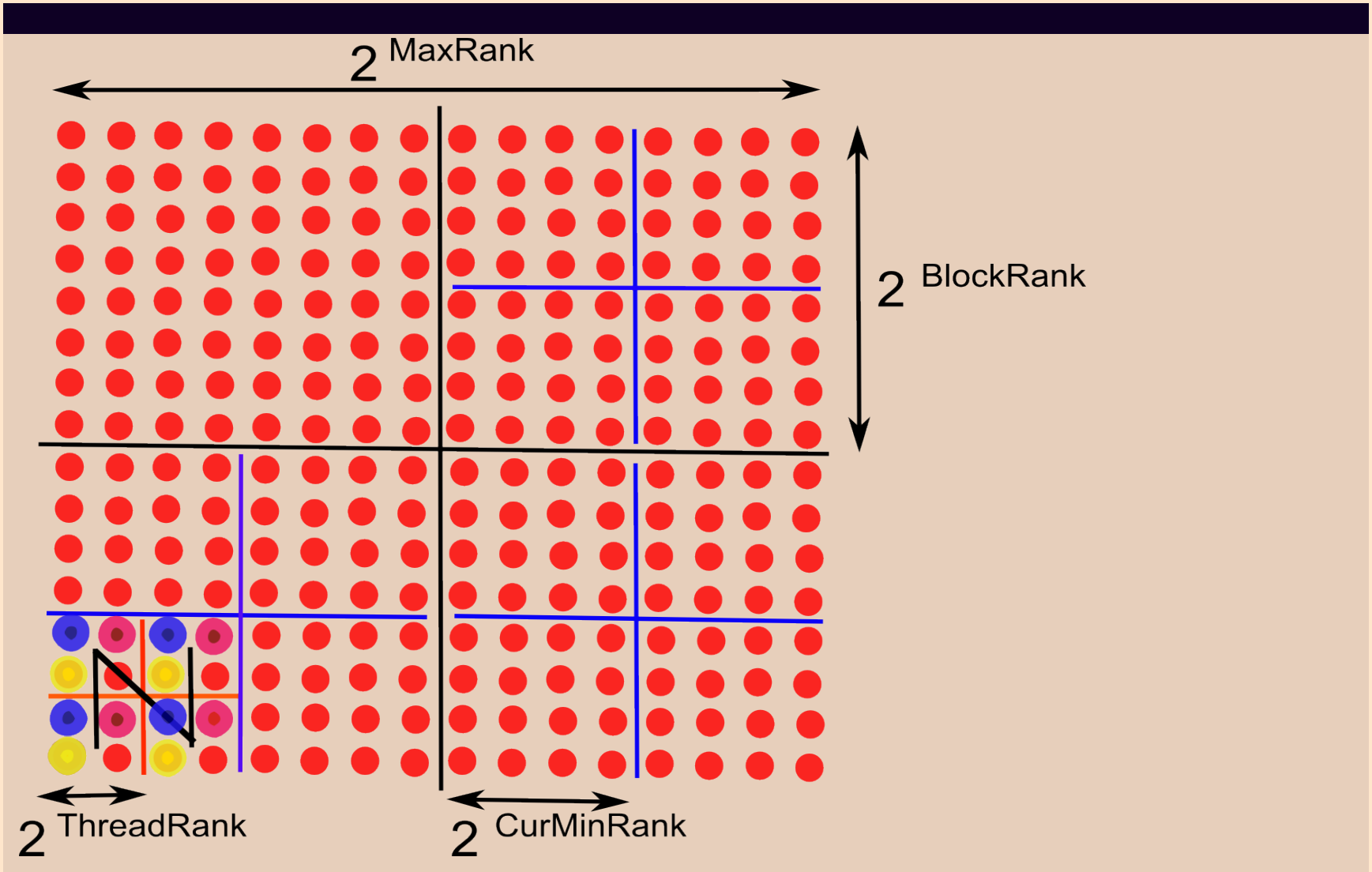
Пирамидальная модель вычислительной системы, включающей GPGPU



Программная модель. Выбор оптимальных параметров

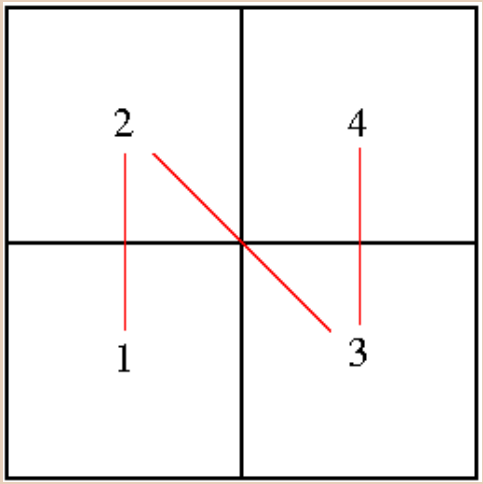


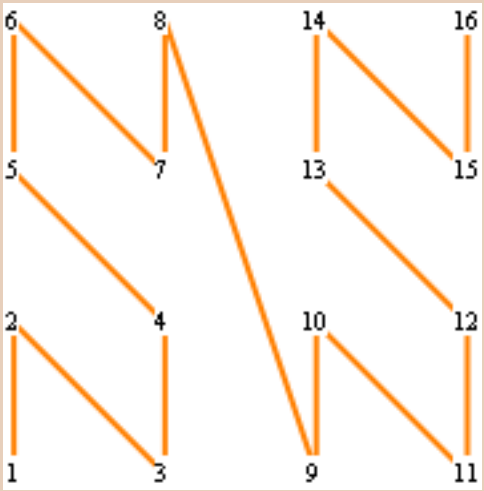
Трехуровневый алгоритм разбиения матриц

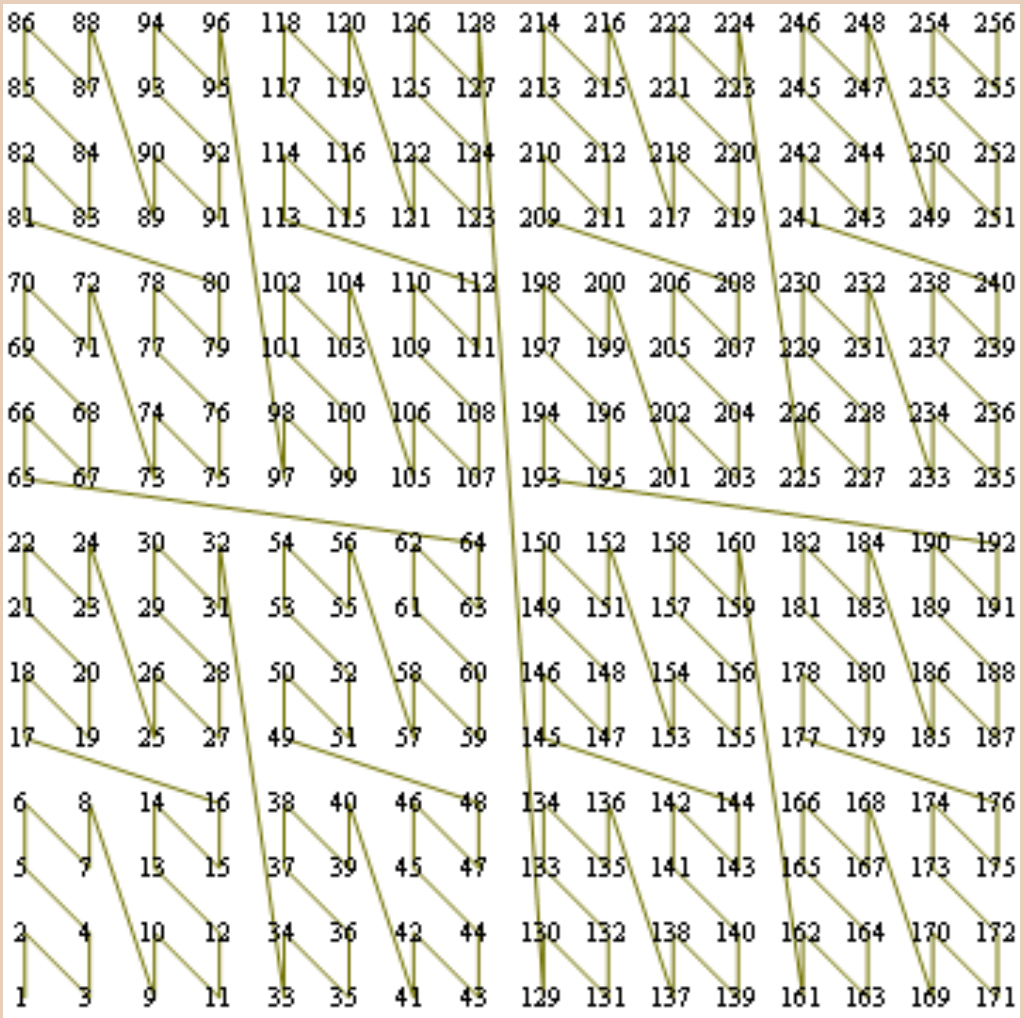


Вопрос о способе хранения данных

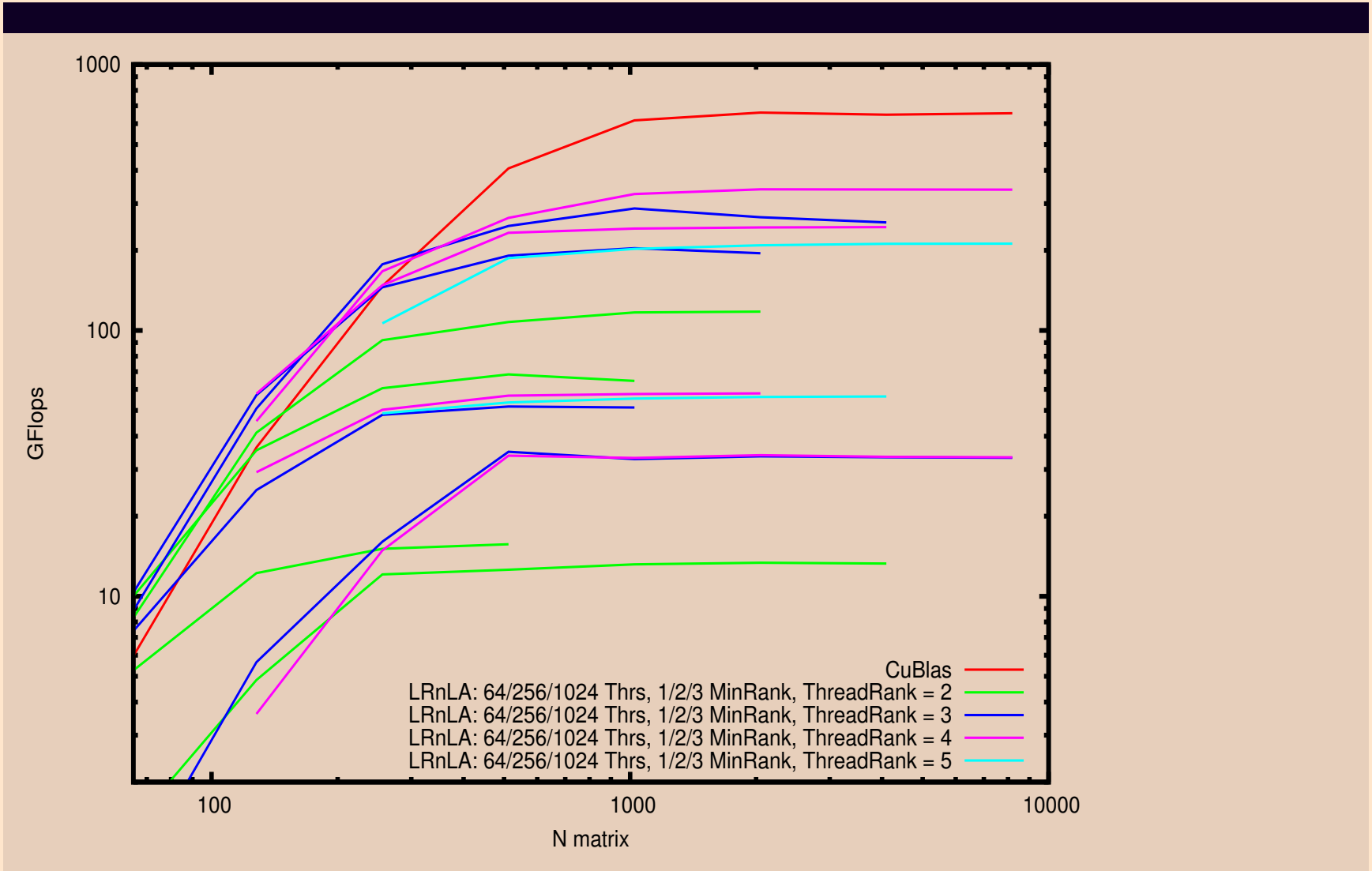
Хранение данных CubeLR







Производительность. Сравнение с библиотекой CUBLAS



Вопрос о компиляторе

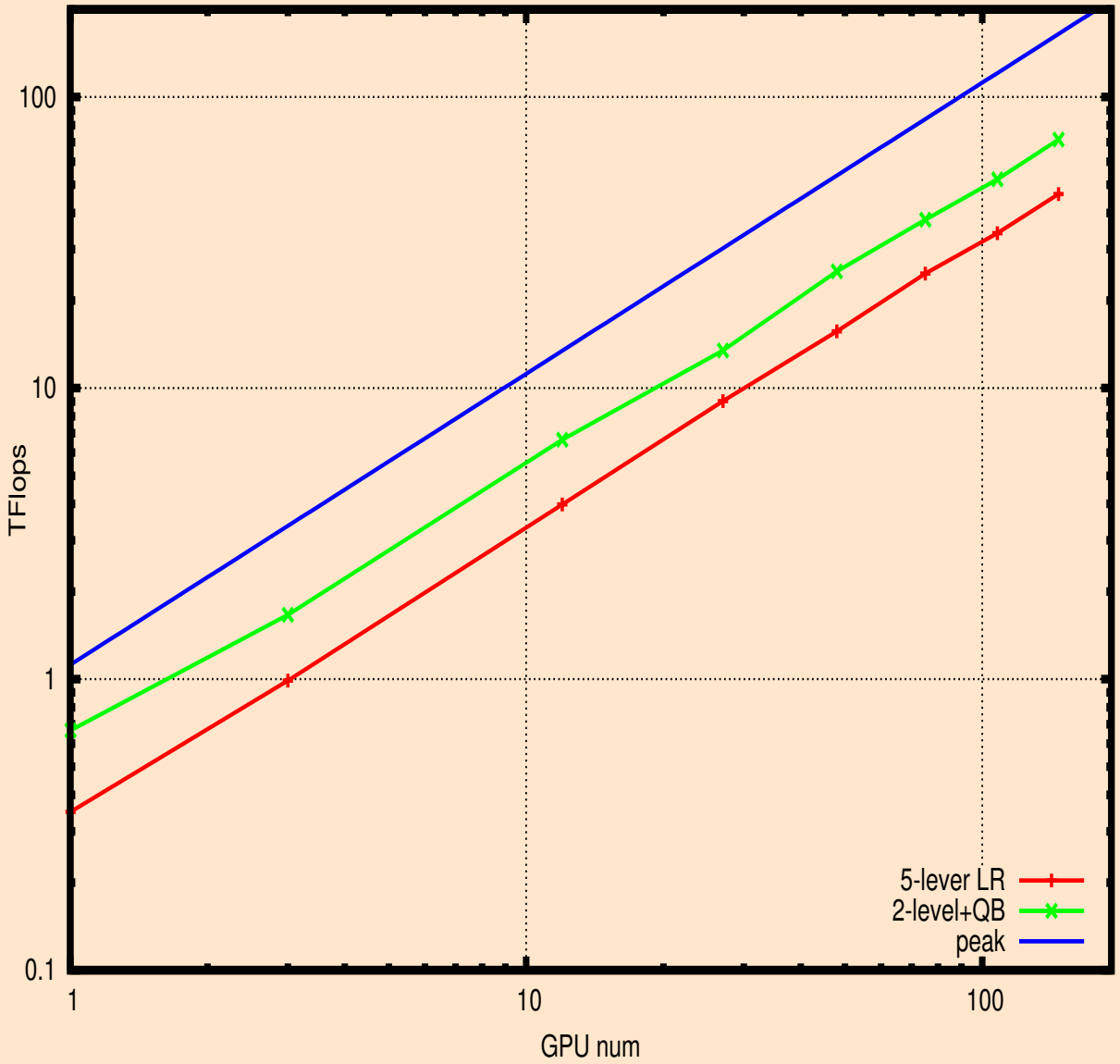
CPU реализация

1. Пиковая производительность:
   (2.93ГГц)*(11 ядер)*(2 unit per clock:
   умножение+сложение)*(4 float вектора) = 253.33 Gflops

Matrix	NxN	DataSz	Thr
<f4,12>	4Kx4K	192MB	11
<f4,13>	8Kx8K	768MB	11

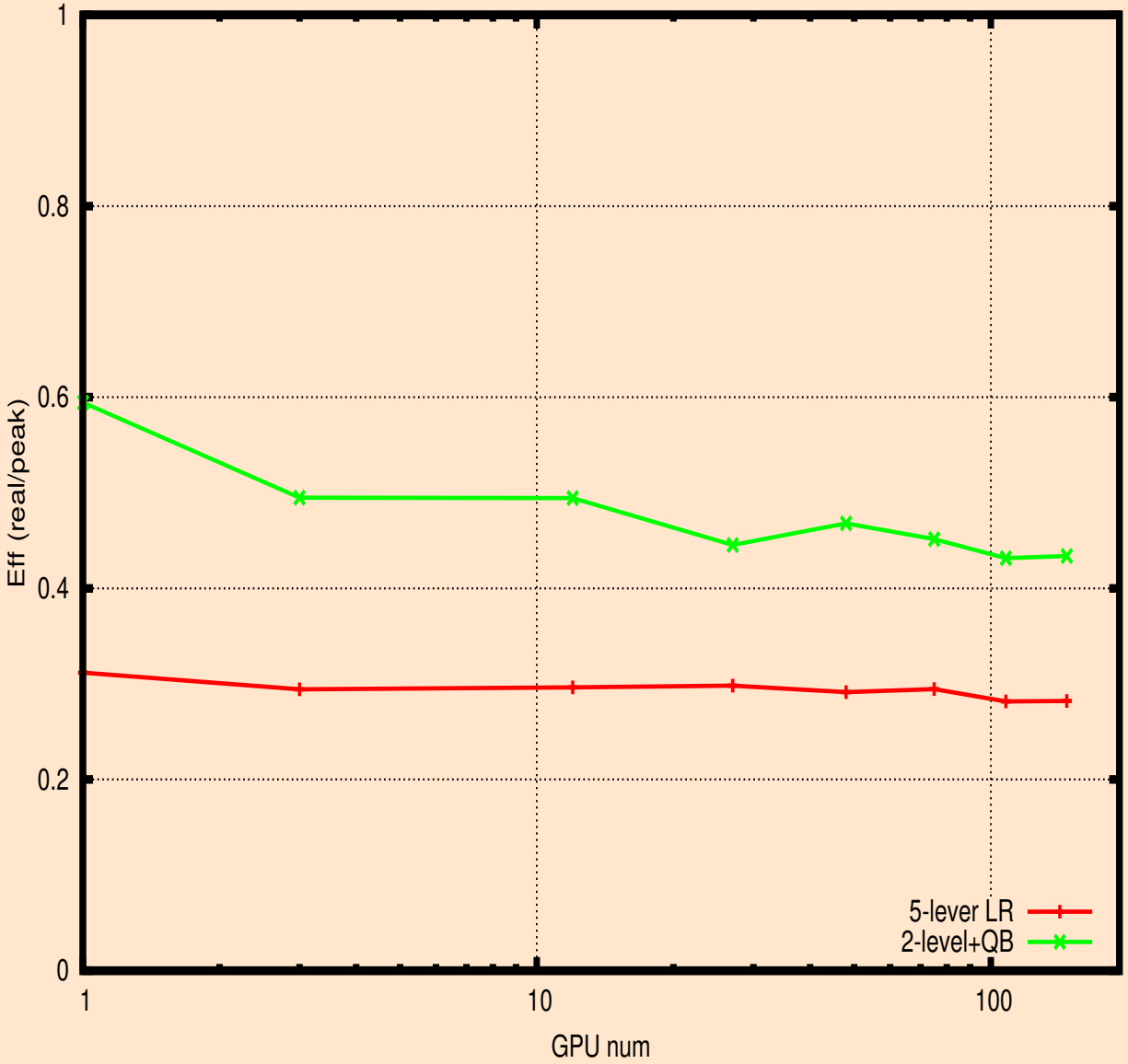
2. Версия компилятора: gcc 4.7.1
   `g++ -march=corei7 -Ofast -O2 -lgomp`
   Производительность: 174/253.33 = 70%
3. Версия компилятора: gcc-4.3
   Производительность: 90/253.33 = 35%

5-уровневое разбиение. Производительность на K-100



размер матриц:
$N = (1 \div 7) \times 2^{15}$
(28GB на узел)
уровни разбиения 1-4:
$2 + 4 + 7 + 2 = 15$
уровень 4:
3xGPU+2xCPU,OpenMP
уровень 5:
1-49 nodes, MPI
Производительность
пиковая на 64 узлах:
215TFlops
достигнуто (49 узлов):
LR: 46.4TFlops
cublas: 71.4TFlops

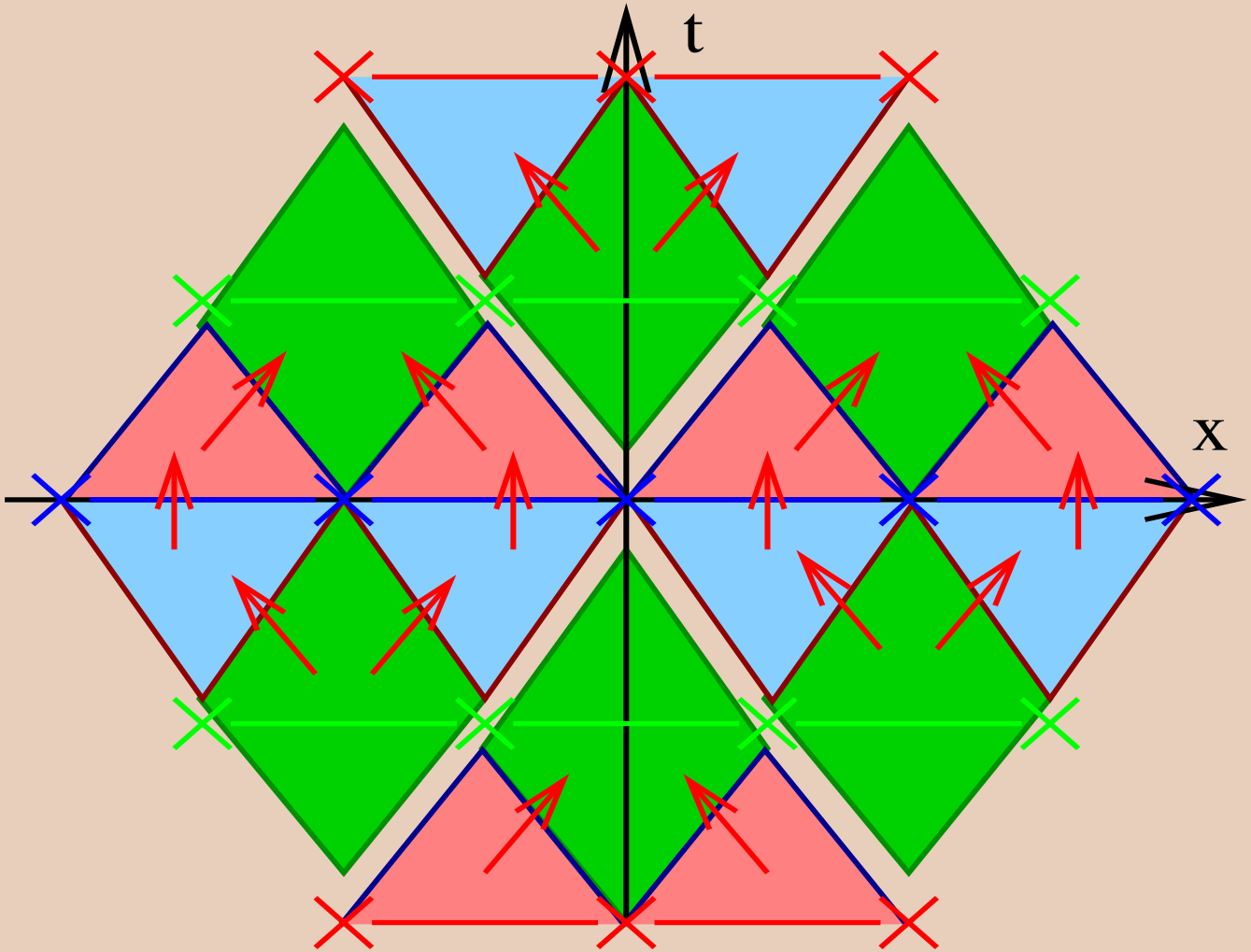
5-уровневое разбиение. Эффективность на K-100



размер матриц:
$N = (1 \div 7) \times 2^{15}$
(28GB на узел)
уровни разбиения 1-4:
$2 + 4 + 7 + 2 = 15$
уровень 4:
3xGPU+2xCPU,OpenMP
уровень 5:
1-49 nodes, MPI
Эффективность
достигнуто (49 узлов):
параллельная:
90%(LR),72%(cublas)
полная:
28%(LR),43%(cublas)

Локально-Рекурсивные неЛокально-Ассинхронные алгоритмы

LRnLA алгоритмы



Планируемое применение GPGPU для задач физики плазмы

Трёхмерная полностью кинетическая модель

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

$$\nabla \vec{B} = 0, \qquad \nabla \vec{E} = 4\pi\rho,$$

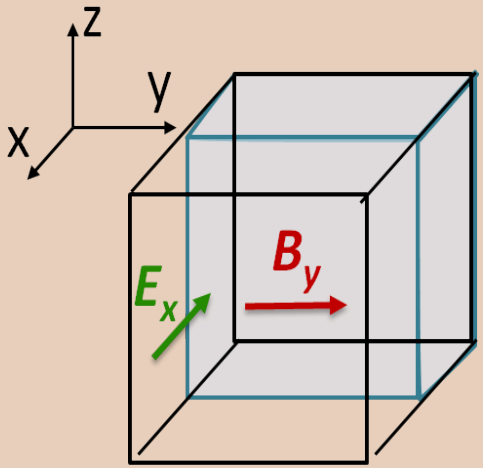
$$\rho = \sum_{\alpha} \int f_\alpha e_\alpha d\vec{v}, \qquad \vec{j} = \sum_{\alpha} \int \vec{v}_\alpha f_\alpha e_\alpha d\vec{v}.$$

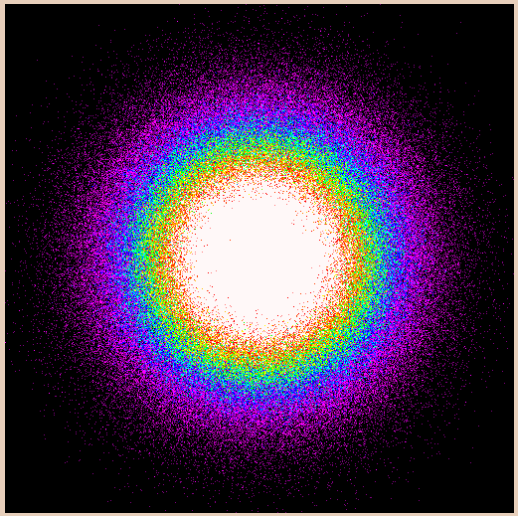
$$\frac{\partial f_\alpha}{\partial t} + \vec{v}_\alpha \frac{\partial f_\alpha}{\partial \vec{r}_\alpha} + \frac{e_\alpha}{m_\alpha \gamma}\left(\frac{1}{c}\vec{v}_\alpha \times \vec{B} + \vec{E}\right)\frac{\partial f_\alpha}{\partial \vec{v}_\alpha} = 0.$$

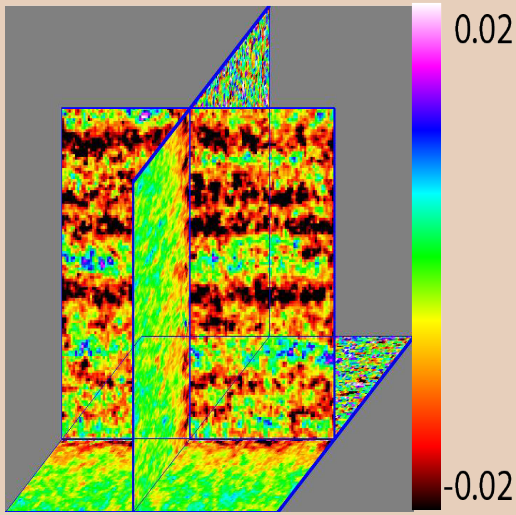
$$\frac{d\vec{v}_j}{dt} = \frac{e_\alpha}{m_\alpha}\left(\frac{\vec{v}_j}{c} \times \vec{B}_i(\vec{r}_j) + \vec{E}_i(\vec{r}_j)\right),$$

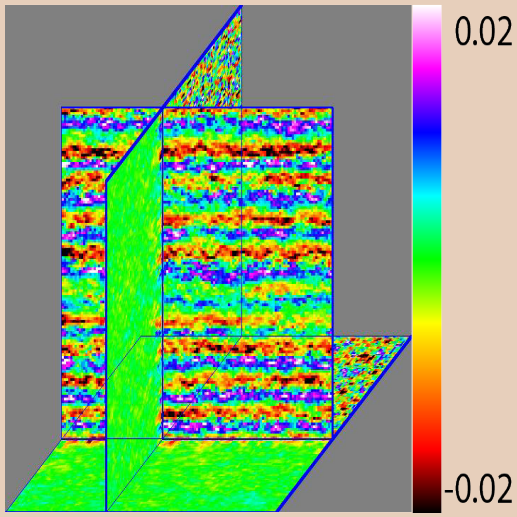
$$\frac{d\vec{r}_j}{dt} = \frac{\vec{v}_j}{\sqrt{1 + (\vec{v}_i/c)^2}},$$

Результаты моделирования холловского двигателя. CPU









Выводы

1. Реализован алгоритм перемножения матриц, хранящихся в локально-рекурсивном виде, использующий пятиуровневое разбиение, учитывающее организацию подсистемы памяти.
2. На Tesla 2050 без проведения каких-либо специфических оптимизаций достигнута производительность на уровне 350GFlops (1/3 от пика или 1/2 от результата cublas для матриц, хранящихся в виде двумерных массивов).
3. LRnLA алгоритмы применимы для GPGPU, оптимальное использования регистровой памяти требует выбора промежуточного количества тредов в блоке (<=256).
4. Параллельная эффективность на верхних уровнях для суперкомпьютера K100 близка к предельной (90% для LR и 72% для cublas для 147 GPGPU на 49 2xCPU узлах).