

Программная модель и оценка производительности для гетерогенных систем с графическими процессорами

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

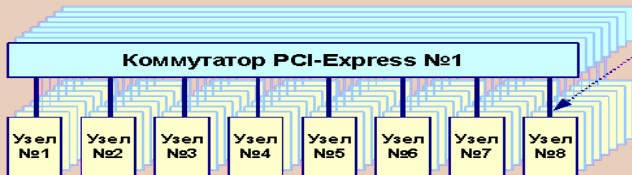
ИПМ им. М.В. Келдыша РАН

2012

Гибридный вычислительный кластер K-100

k100.kiam.ru

Скорость до 700 МБайт/с
Латентность ~ 12мкс
Время выдачи слова ~ 70нс
Время чтения слова ~ 2.5мкс



Вычислительный узел

2 x CPU
Intel Xeon X5670

DDR3
SDRAM

3 x GPU
Fermi C2050

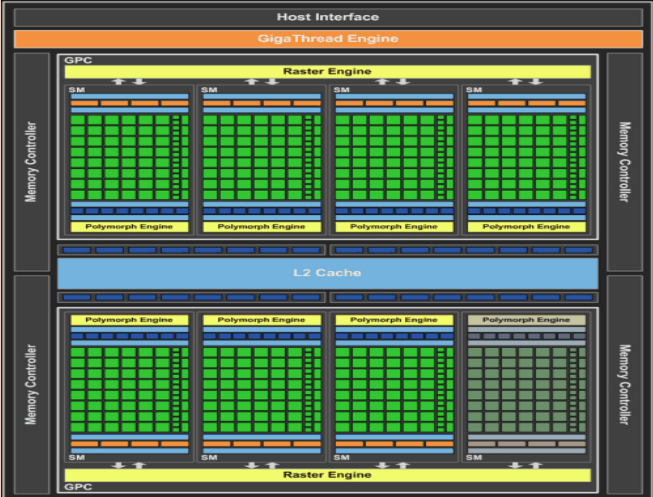
6 ядер на процессоре (12 на узле)
Кэш 12МБайт
Частота ядра 2.93 ГГц

DDR3 SDRAM 96 ГБайт

448 ядер CUDA (1344 на узле)
Частота ядра 1.15 ГГц
2,5 ГБ памяти GDDR5 (7,5 на узле)
Частота памяти 1.5 ГГц
Пропускная способность памяти 144 Гб/с

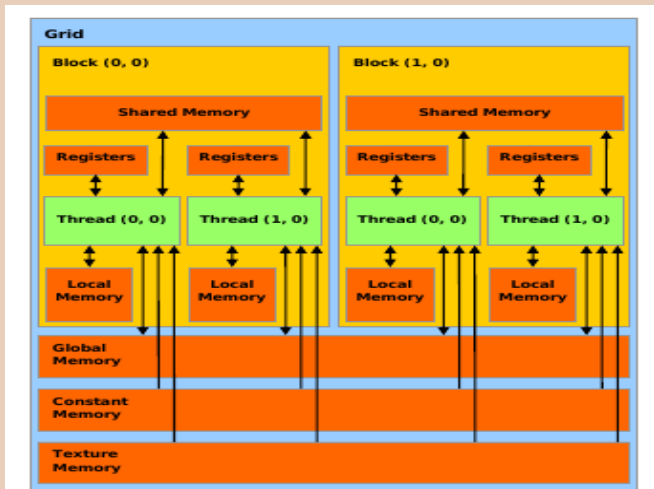
Вычислительные HW-уровни

Архитектура Fermi

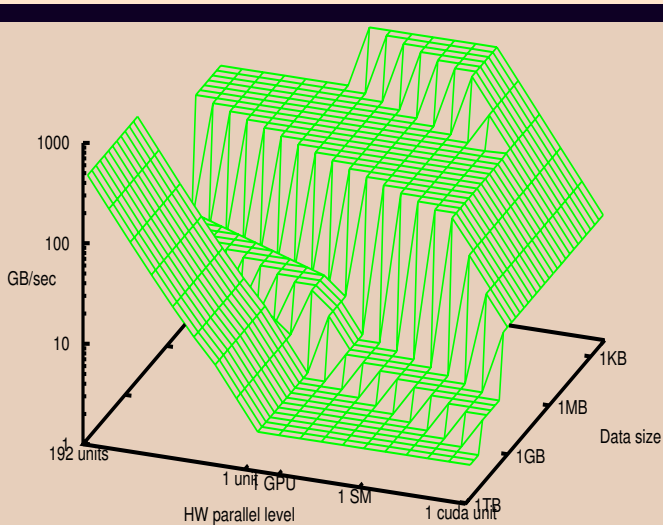


Иерархия памяти в устройствах с графическими процессорами

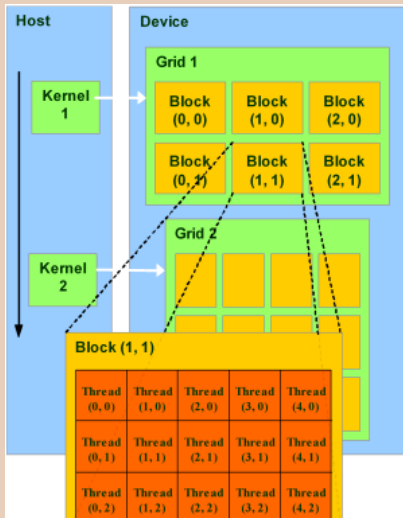
Комплекс Tesla 2050



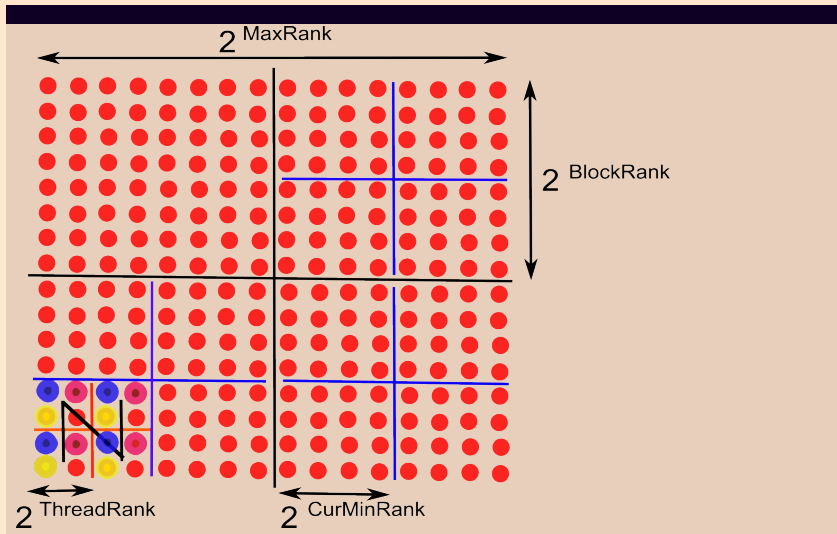
Пирамидальная модель вычислительной системы, включающей GPGPU



Программная модель. Выбор оптимальных параметров

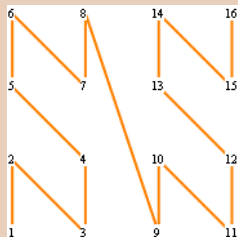
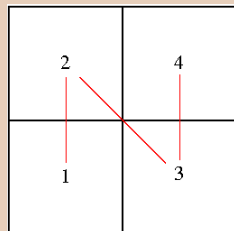


Трёхуровневый алгоритм разбиения матриц



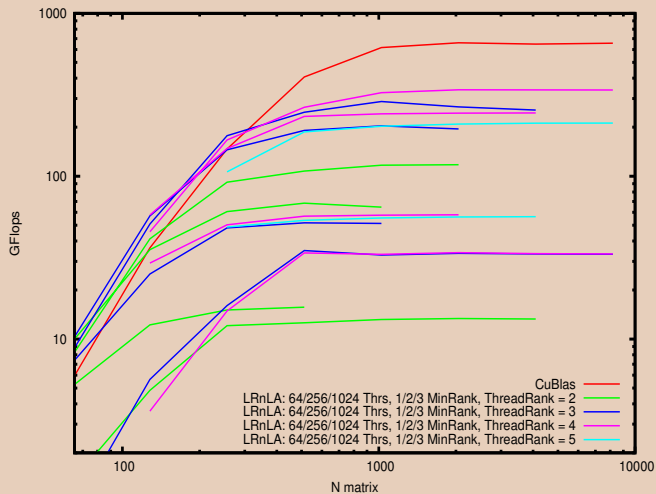
Вопрос о способе хранения данных

Хранение данных CubeLR



86	88	94	96	118	120	126	128	214	216	222	224	246	248	254	256
85	87	93	95	117	119	125	127	213	215	221	223	245	247	253	255
82	84	90	92	114	116	122	124	210	212	218	220	242	244	250	252
81	83	89	91	113	115	121	123	209	211	217	219	241	243	249	251
70	72	78	80	102	104	110	112	198	200	206	208	230	232	238	240
69	71	77	79	101	103	109	111	197	199	205	207	229	231	237	239
66	68	74	76	98	100	106	108	194	196	202	204	226	228	234	236
65	67	73	75	97	99	105	107	193	195	201	203	225	227	233	235
22	24	30	32	54	56	62	64	150	152	158	160	182	184	190	192
21	23	29	31	53	55	61	63	149	151	157	159	181	183	189	191
18	20	26	28	50	52	58	60	146	148	154	156	178	180	186	188
17	19	25	27	49	51	57	59	145	147	153	155	177	179	185	187
6	8	14	16	38	40	46	48	134	136	142	144	166	168	174	176
5	7	13	15	37	39	45	47	133	135	141	143	165	167	173	175
2	4	10	12	34	36	42	44	130	132	138	140	162	164	170	172
1	3	9	11	33	35	41	43	129	131	137	139	161	163	169	171

Производительность. Сравнение с библиотекой CUBLAS



Вопрос о компиляторе

CPU реализация

1. Пиковая производительность:

$(2.93\text{ГГц}) * (11 \text{ ядер}) * (2 \text{ unit per clock:}$

$\text{умножение} + \text{сложение}) * (4 \text{ float вектора}) = 253.33 \text{ Gflops}$

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

2. Версия компилятора: gcc 4.7.1

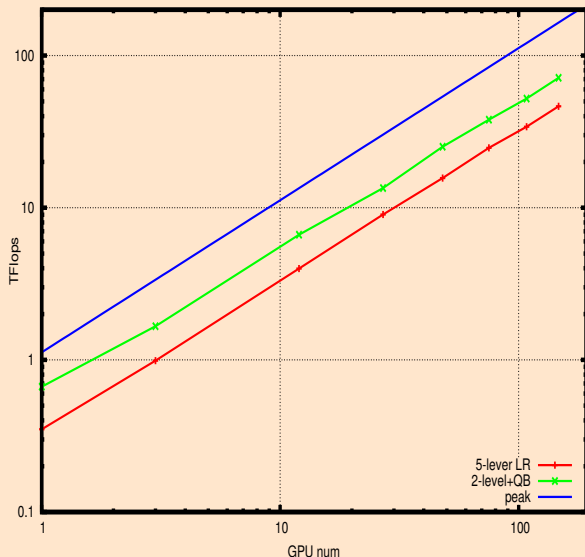
`g++ -march=corei7 -Ofast -O2 -lgomp`

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

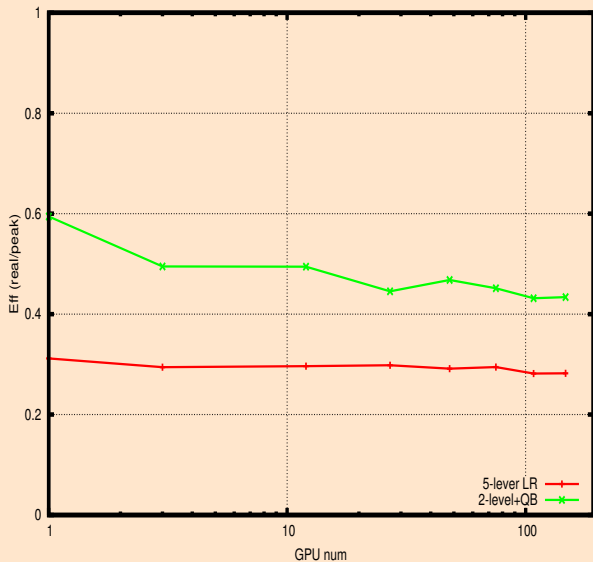
Производительность
пиковая на 64 узлах:
215TFlops

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:

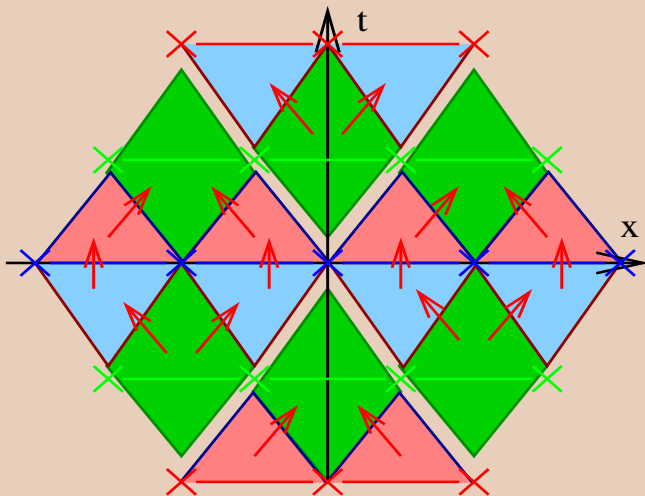
90%(LR), 72%(cublas)

полная:

28%(LR), 43%(cublas)

Локально-Рекурсивные неЛокально-Ассинхронные алгоритмы

LRnLA алгоритмы



Планируемое применение GPGPU для задач физики плазмы

Трёхмерная полностью кинетическая модель

$$\frac{1}{c} \frac{\partial \vec{B}}{\partial t} = -\nabla \times \vec{E}, \quad \frac{1}{c} \frac{\partial \vec{E}}{\partial t} = \nabla \times \vec{B} - \frac{4\pi \vec{j}}{c},$$
$$\nabla \cdot \vec{B} = 0, \quad \nabla \cdot \vec{E} = 4\pi \rho,$$

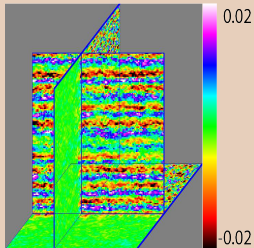
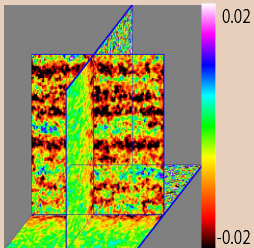
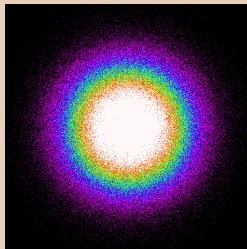
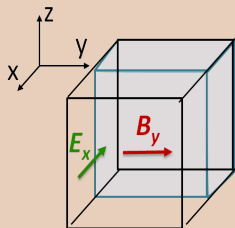
$$\rho = \sum_{\alpha} \int f_{\alpha} e_{\alpha} d\vec{v}, \quad \vec{j} = \sum_{\alpha} \int \vec{v}_{\alpha} f_{\alpha} e_{\alpha} d\vec{v}.$$

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

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

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

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 использование 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).