Самая быстрая и энергоэффективная реализация BFS согласно рейтингу Graph500

Колганов А.С., МГУ ВМК

План доклада

- Алгоритм BFS
- Тенденции рейтингов Graph500 и GGraph500
- Реализация BFS на общей памяти (CPU / GPU)
- Прогнозируемая масштабируемость

План доклада

- Алгоритм BFS
- Тенденции рейтингов Graph500 и GGraph500
- Реализация BFS на общей памяти (CPU / GPU)
- Прогнозируемая масштабируемость

Поиск в ширину (BFS)

- Breadth first search – один из важных и фундаментальных алгоритмов в обработке графов;

- Алгоритмические трудности BFS:
  - Очень мало вычислений;
  - Нерегулярный доступ к памяти.

План доклада

- Алгоритм BFS
- Тенденции рейтингов Graph500 и GGraph500
- Реализация BFS на общей памяти (CPU / GPU)
- Прогнозируемая масштабируемость

Graph500 и GreenGraph500

- Использование алгоритма BFS для ранжирования суперкомпьютеров (**TEPS** – обработка количества ребер в секунду);
- Использование метрики MTEPS / WATT для ранжирования в рейтинге энергоэффективных машин;
- Оба списка до сих пор не заполнены:
  - Graph 500 (занято 201позиция);
  - GreenGraph 500 (занято 63 позиции).

Graph500 – поиск в графе

- Генерация ребер;
- Построение графа по полученным ребрам *(замеряется время, входит в таблицу);*
- Генерация 64 произвольных вершин из которых выполняется BFS;
- Для каждой вершины:
  - Выполнение алгоритма BFS *(замеряется время, входит в рейтинг);*
  - Выполнение процедуры проверки корректности;
- Вывод результирующей информации.

Graph500 – тенденции

		Узлов	Ядер	Масштаб	GTEPS
1	K computer (Fujitsu - Custom)	82944	663552	40	38 621
2	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	98304	1572864	41	23 751
3	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	49152	786432	40	14 982
4	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	16384	262144	38	5 848
5	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	8192	131072	37	2 567
6	Tianhe-2 (MilkyWay-2)	8192	196608	36	2 061
7	IBM - BlueGene/Q, Power BQC 16C 1.60GHz	4096	65536	36	1 427
7	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	4096	65536	36	1 427
7	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	4096	65536	36	1 427
7	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	4096	65536	36	1 427
7	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	4096	65536	36	1 427
12	HP - Cluster Platform SL390s G7	1024	12288	36	1 345
13	IBM - Power 775, POWER7 8C 3.836GHz	1960	7840	36	1 172
14	Fujitsu - Fujitsu - Fujitsu PRIMEHPC FX 10	4800	76800	36	1 003
15	Fujitsu - Fujitsu PRIMEHPC FX 10	4800	76800	38	993
16	Cray - XC30	4817	115600	36	865
17	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	2048	32768	35	769
17	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	2048	32768	35	769
17	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	2048	32768	35	769
17	IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	2048	32768	35	769

Graph500 – тенденции

		Узлов	Ядер	Масштаб	GTEPS
12,6 MW	1 K computer (Fujitsu - Custom)	82944	663552	40	38 621
7,8 MW	2 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	98304	1572864	41	23 751
3,9 MW	3 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	49152	786432	40	14 982
	4 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	16384	262144	38	5 848
	5 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	8192	131072	37	2 567
17,8 MW	6 Tianhe-2 (MilkyWay-2)	8192	196608	36	2 061
	7 IBM - BlueGene/Q, Power BQC 16C 1.60GHz	4096	65536	36	1 427
	7 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	4096	65536	36	1 427
	7 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	4096	65536	36	1 427
	7 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	4096	65536	36	1 427
	7 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	4096	65536	36	1 427
	12 HP - Cluster Platform SL390s G7	1024	12288	36	1 345
	13 IBM - Power 775, POWER7 8C 3.836GHz	1960	7840	36	1 172
	14 Fujitsu - Fujitsu - Fujitsu PRIMEHPC FX 10	4800	76800	36	1 003
	15 Fujitsu - Fujitsu PRIMEHPC FX 10	4800	76800	38	993
	16 Cray - XC30	4817	115600	36	865
	17 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	2048	32768	35	769
	17 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	2048	32768	35	769
	17 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	2048	32768	35	769
	17 IBM - BlueGene/Q, Power BQC 16C 1.60 GHz	2048	32768	35	769

x2

x2

GreenGraph500 – тенденции

Big DATA

Rank	MTEPS/W	Machine	Scale	GTEPS	Nodes	Cores
1	62,93	GraphCREST (CPU)	30	31,33	1	32
2	61,48	GraphCREST (CPU)	30	28,61	1	32
3	51,95	GraphCREST (CPU)	32	59,9	1	60
4	48,28	GraphCREST (CPU)	30	31,95	1	32
5	44,42	GraphCREST (CPU)	32	55,74	1	60

Small DATA

Rank	MTEPS/W	Machine	Scale	GTEPS	Nodes	Cores
1	815,68	TitanX (GPU)	26	132,14	1	28
2	540,94	Titan (GPU)	25	114,68	1	20
3	445,92	Colonial (GPU)	20	112,18	1	12
4	243,42	Monty Pi-thon	26	35,83	32	128
5	235,15	GraphCREST (ARM)	20	1,03	1	4
6	230,4	GraphCREST (ARM)	20	0,74	1	4
7	204,38	EBD	21	1,64	1	5

GreenGraph500 – тенденции

Big DATA: scale от 30 (256 ГБ для int64 и 128 ГБ для int32)

Rank	MTEPS/W	Machine	Scale	GTEPS	Nodes	Cores
1	62,93	GraphCREST (CPU)	30	31,33	1	32
2	61,48	GraphCREST (CPU)	30	28,61	1	32
3	51,95	GraphCREST (CPU)	32	59,9	1	60
4	48,28	GraphCREST (CPU)	30	31,95	1	32
5	44,42	GraphCREST (CPU)	32	55,74	1	60

Small DATA

Rank	MTEPS/W	Machine	Scale	GTEPS	Nodes	Cores
1	815,68	TitanX (GPU)	26	132,14	1	28
2	540,94	Titan (GPU)	25	114,68	1	20
3	445,92	Colonial (GPU)	20	112,18	1	12
4	243,42	Monty Pi-thon	26	35,83	32	128
5	235,15	GraphCREST (ARM)	20	1,03	1	4
6	230,4	GraphCREST (ARM)	20	0,74	1	4
7	204,38	EBD	21	1,64	1	5

GreenGraph500 – тенденции

Big DATA: scale от 30 (256 ГБ для int64 и 128 ГБ для int32)

- GTX Titan X – 12ГБ, Tesla K80 – 24ГБ;
- Для обработки scale 30 необходимо ~192ГБ;
- <GTX Titan X> x 16 = 192 ГБ 4 kW пик!
- <Tesla K80> x 8 = 192 ГБ 2.4 kW пик!

Small DATA

Rank	MTEPS/W	Machine	Scale	GTEPS	Nodes	Cores
1	815,68	TitanX (GPU)	26	132,14	1	28
2	540,94	Titan (GPU)	25	114,68	1	20
3	445,92	Colonial (GPU)	20	112,18	1	12
4	243,42	Monty Pi-thon	26	35,83	32	128
5	235,15	GraphCREST (ARM)	20	1,03	1	4
6	230,4	GraphCREST (ARM)	20	0,74	1	4
7	204,38	EBD	21	1,64	1	5

План доклада

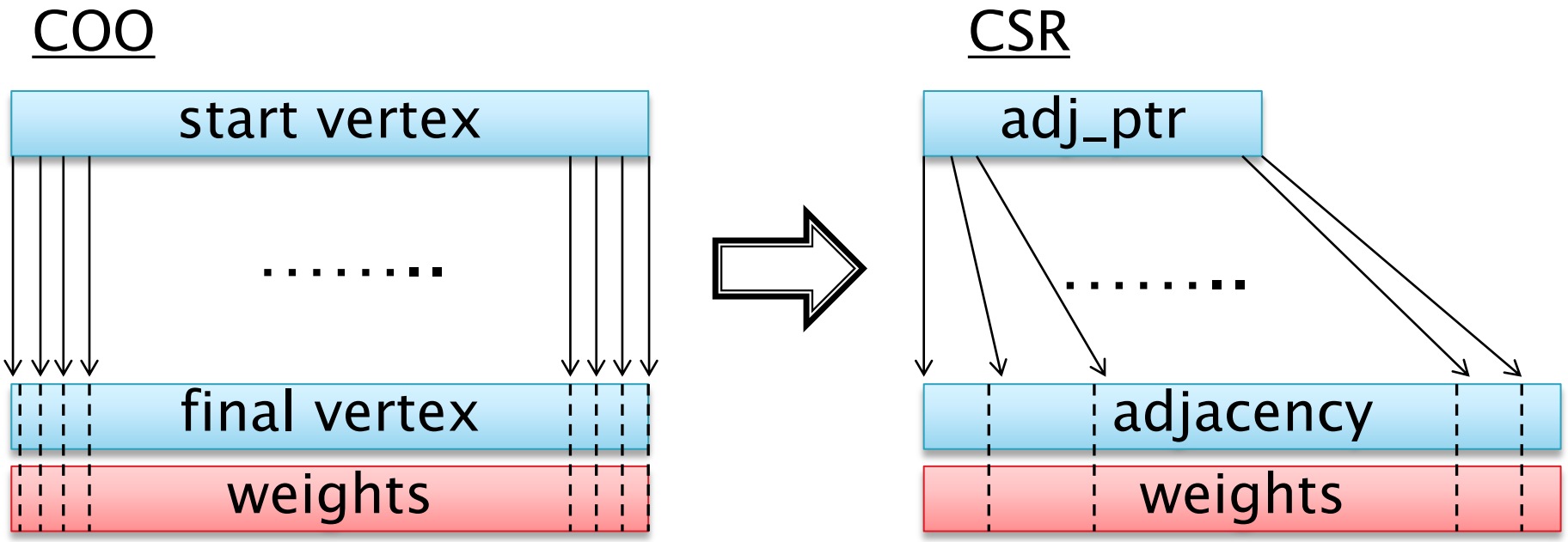
- Алгоритм BFS
- Тенденции рейтингов Graph500 и GGraph500
- Реализация BFS на общей памяти (CPU / GPU)
- Прогнозируемая масштабируемость

Реализация BFS

- Фаза 1:
  - преобразование полученного графа в CSR;
  - загрузка в память GPU;
- Фаза 2:
  - основной цикл алгоритма по 64 ключам;
  - гибридный алгоритм Top Down + Bottom Up.
- Реализация основана на идеях GraphCREST: *«Fast and Energy-efficient Breadth-First Search on a Single NUMA System, 2014»*

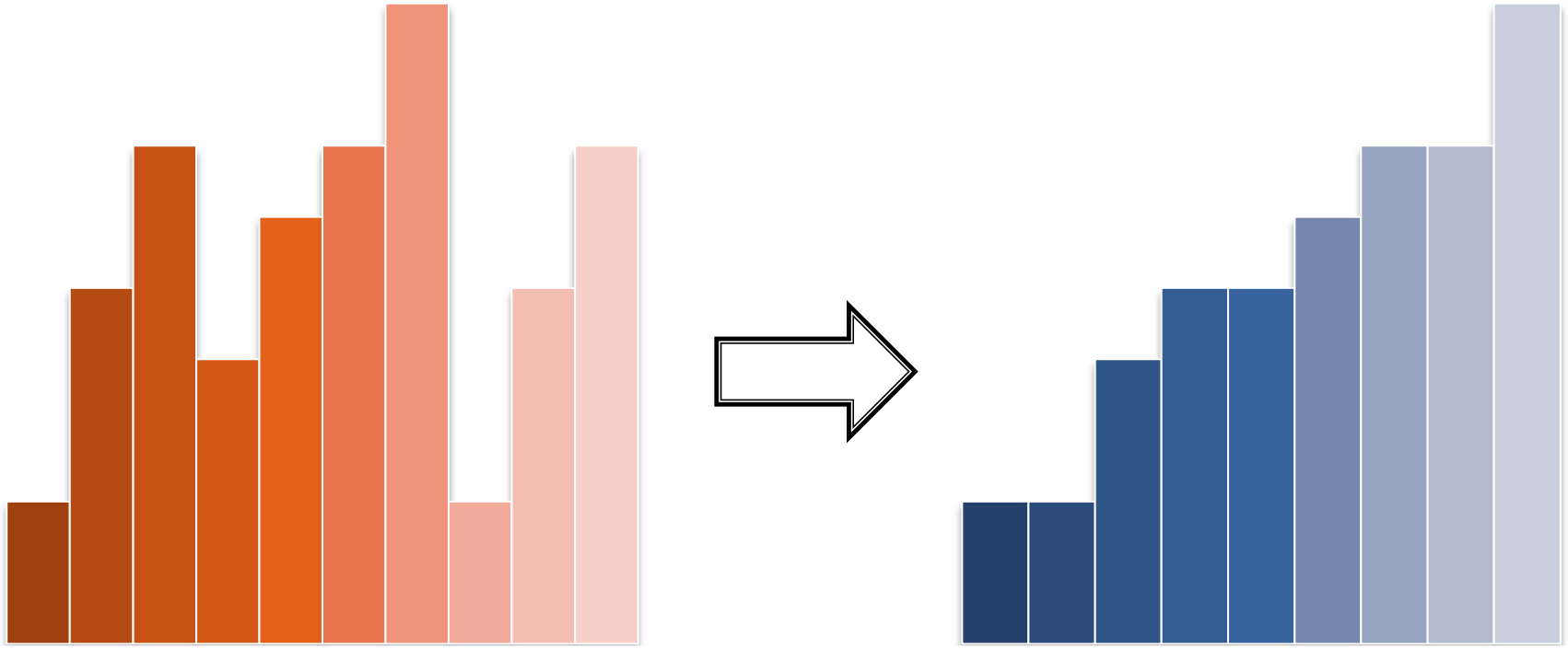
Реализация BFS: фаза1

- Преобразование в CSR (compressed sparse rows)



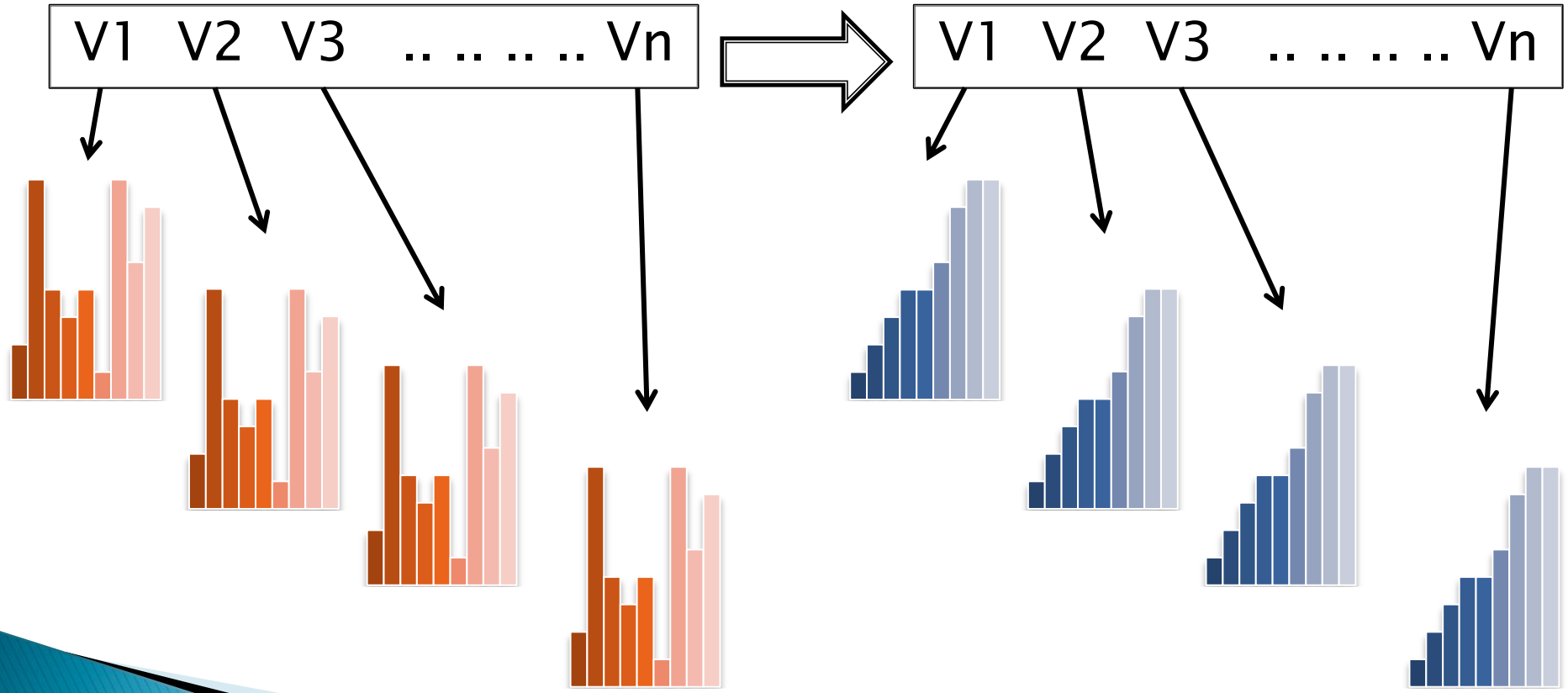
Реализация BFS: фаза1

- Глобальная сортировка вершин по степени связности



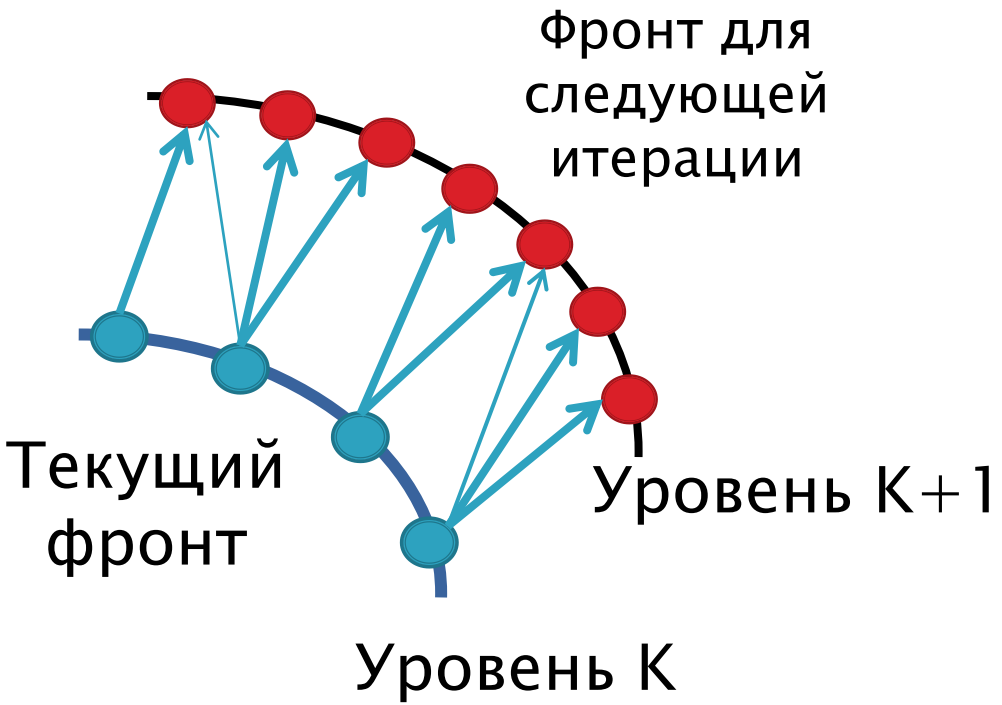
Реализация BFS: фаза1

- Локальная сортировка соседей по степени связности



Реализация BFS: фаза2

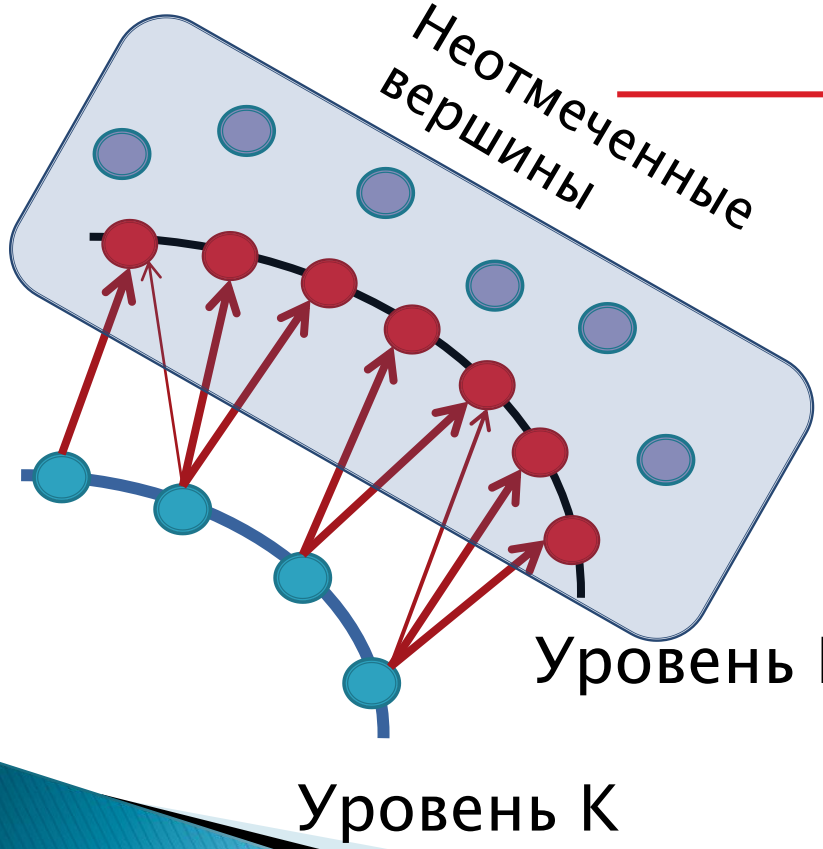
- Синхронный по уровням **Top-Down** обход




```
foreach (i = [0, N])
{
    foreach (k = =[rInd[i], rInd[i+1])
    {
        unsigned v = endV[k];
        if (levels[v] == 0)
        {
            levels[v] = lvl;
            parents[v] = i;
        }
    }
}
```

Реализация BFS: фаза2

- Синхронный по уровням **Bottom-Up** обход




```
foreach (i = [0, N])
{
    if (levels[i] == 0)
    {
        foreach (k=[rInd[i], rInd[i+1] ])
        {
            unsigned endk = endV[k];
            if (levels[endk] == lvl - 1)
            {
                parents[i] = endk;
                levels[i] = lvl;
                break;
            }
        }
    }
}
```

Реализация BFS: фаза2

- Гибридный алгоритм **Top-Down** + **Bottom-Up** (оптимизация по направлениям)

Граф SCALE 26 |V| = 2^26 (67,108,864) |E| = 2^30 (1,073,741,824)

Уровень	Top-Down	Bottom-Up	Hybrid
0	2	2,103,840,895	2
1	66,206	1,766,587,029	66,206
2	346,918,235	52,677,691	52,677,691
3	1,727,195,615	12,820,854	12,820,854
4	29,557,400	103,184	103,184
5	82,357	21,467	21,467
6	221	21,240	221
Всего:	2,103,820,036	3,936,072,360	65,689,625
	100%	187%	3.12%

= 2x|E|

Существенное уменьшение количества просмотренных ребер

Реализация BFS: фаза2

- Гибридный алгоритм **Top-Down** + **Bottom-Up** (оптимизация по направлениям)

Граф SCALE 26 |V| = 2^26 (67,108,864) |E| = 2^30 (1,073,741,824)

Уровень	Top-Down	Bottom-Up	Hybrid	Выбор обхода
0	2	2,103,840,895	2	Фаза роста
1	66,206	1,766,587,029	66,206	
2	346,918,235	52,677,691	52,677,691	
3	1,727,195,615	12,820,854	12,820,854	Фаза убывания
4	29,557,400	103,184	103,184	
5	82,357	21,467	21,467	
6	221	21,240	221	
Всего:	2,103,820,036	3,936,072,360	65,689,625	
	100%	187%	3.12%	

= 2x|E|

Реализация BFS: оптимизации для GPU

- Использование CUDA Dynamic Parallelism для балансировки нагрузки в Top-Down;
- Использование векторизации внутри каждого потока (чтение 4х значений levels за 1 операцию);
- Использование частичного выравнивания части ребер для Bottom-Up;
- Использование Bottom-Up с очередью на более поздних итерациях.

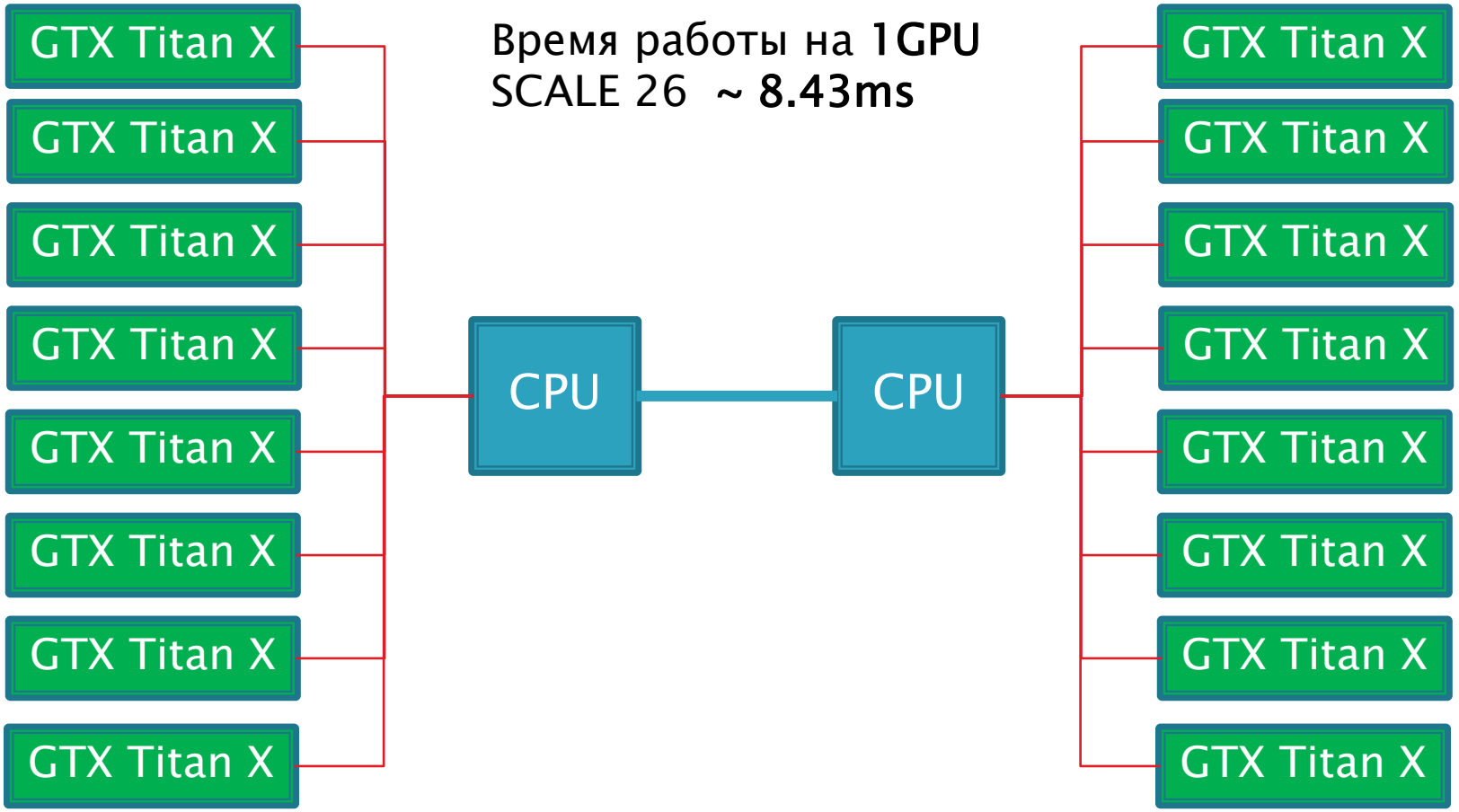
Полученные результаты

- Первая позиция в GGraph500: Small DATA:
  GTX Titan X – 132 GTEPS, 815 MTEPS/W, SCALE:**26**;
- Вторая позиция в GGraph500: Small DATA:
  GTX Titan – 114 GTEPS, 540 MTEPS/W, SCALE:**25**;
- 15я позиция в GGraph500: Small DATA:
  Intel Xeon E5 – 10.6 GTEPS, 81 MTEPS/W, SCALE:**27**;
- Достигнута пропускная способность памяти GPU примерно 140–150 ГБ/с (50–60% от пиковой);
- Энергопотребление GPU составило 50% от максимума.

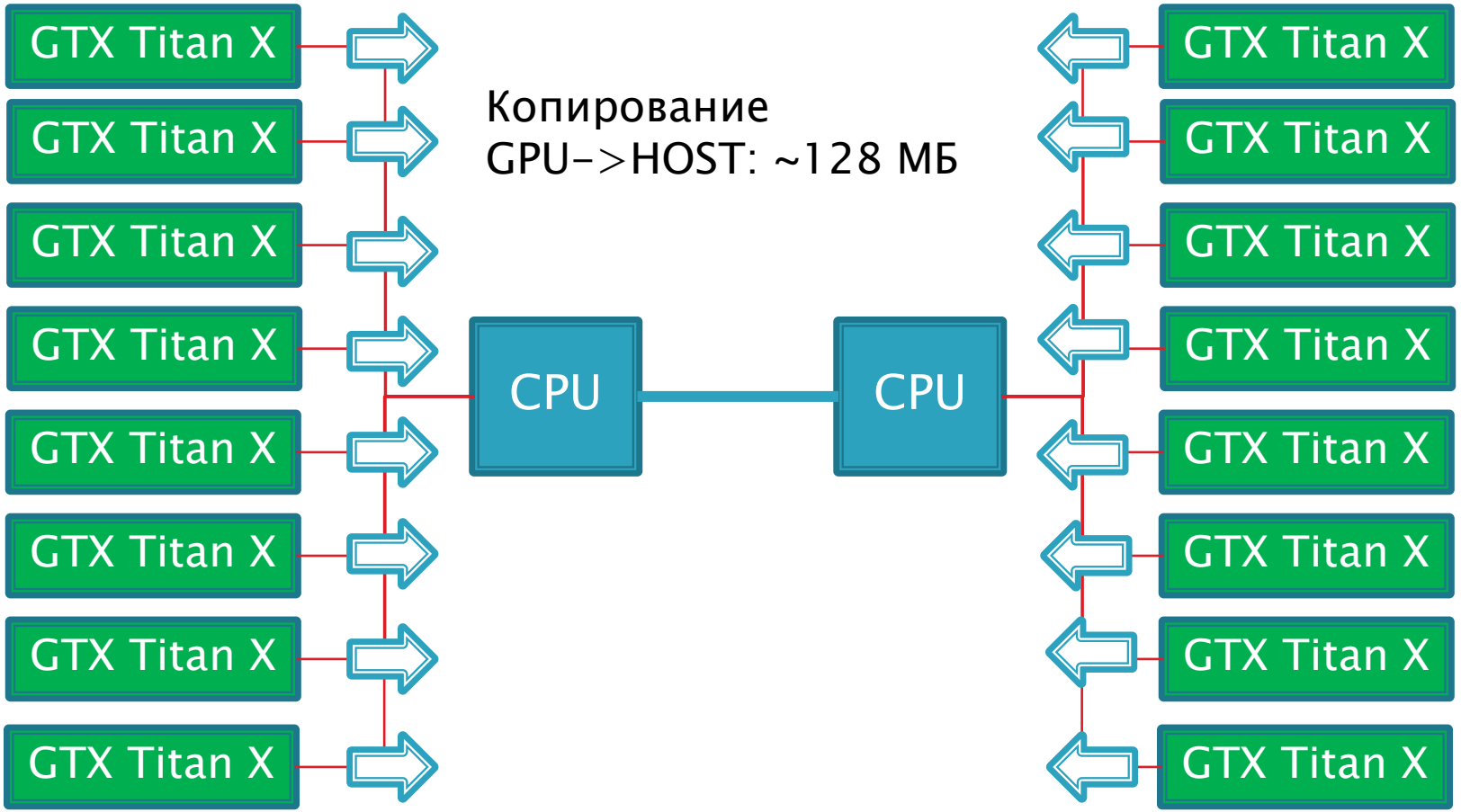
План доклада

- Алгоритм BFS
- Тенденции рейтингов Graph500 и GGraph500
- Реализация BFS на общей памяти (CPU / GPU)
- Прогнозируемая масштабируемость

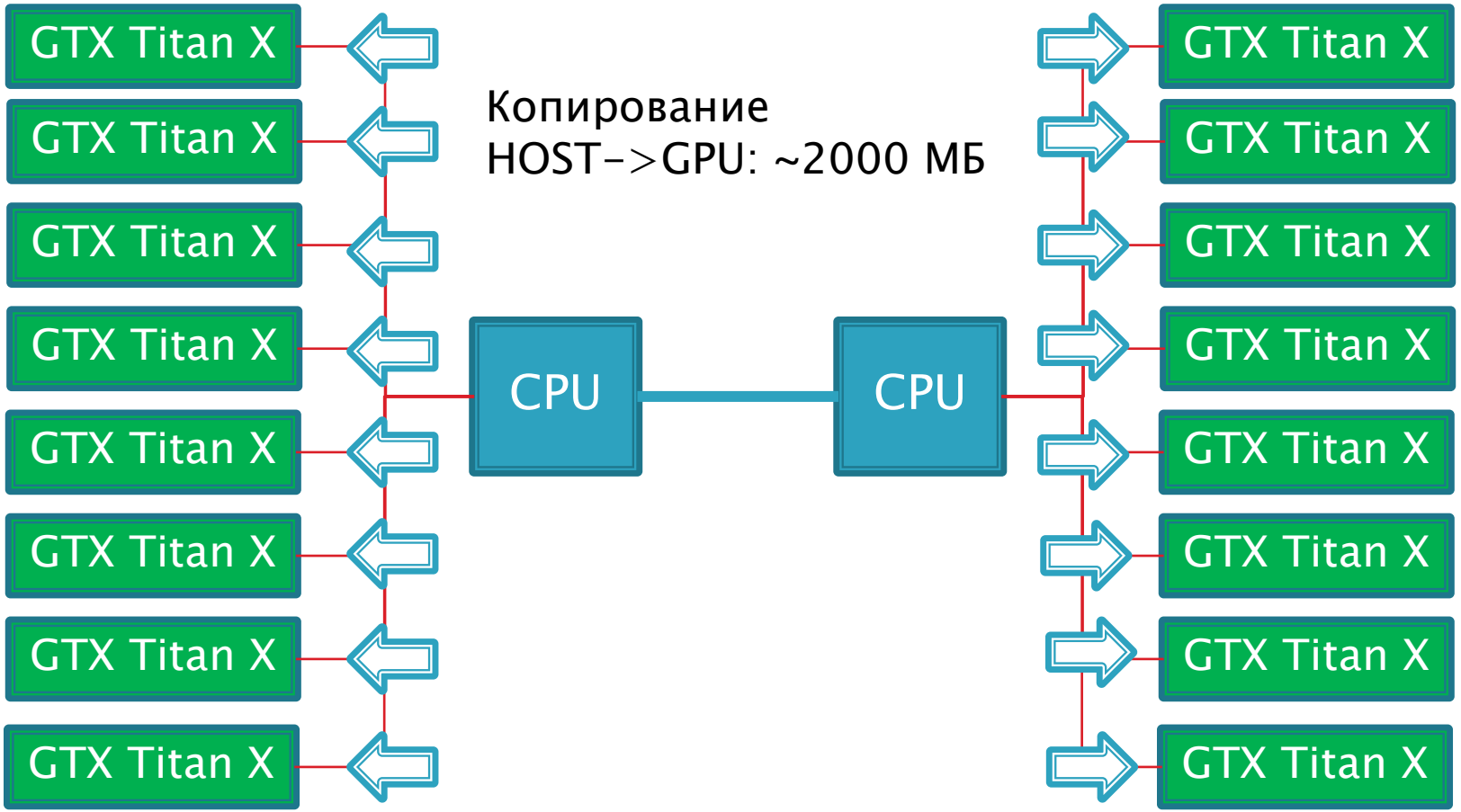
Прогнозируемое масштабирование на несколько GPU: PCIe 3.0



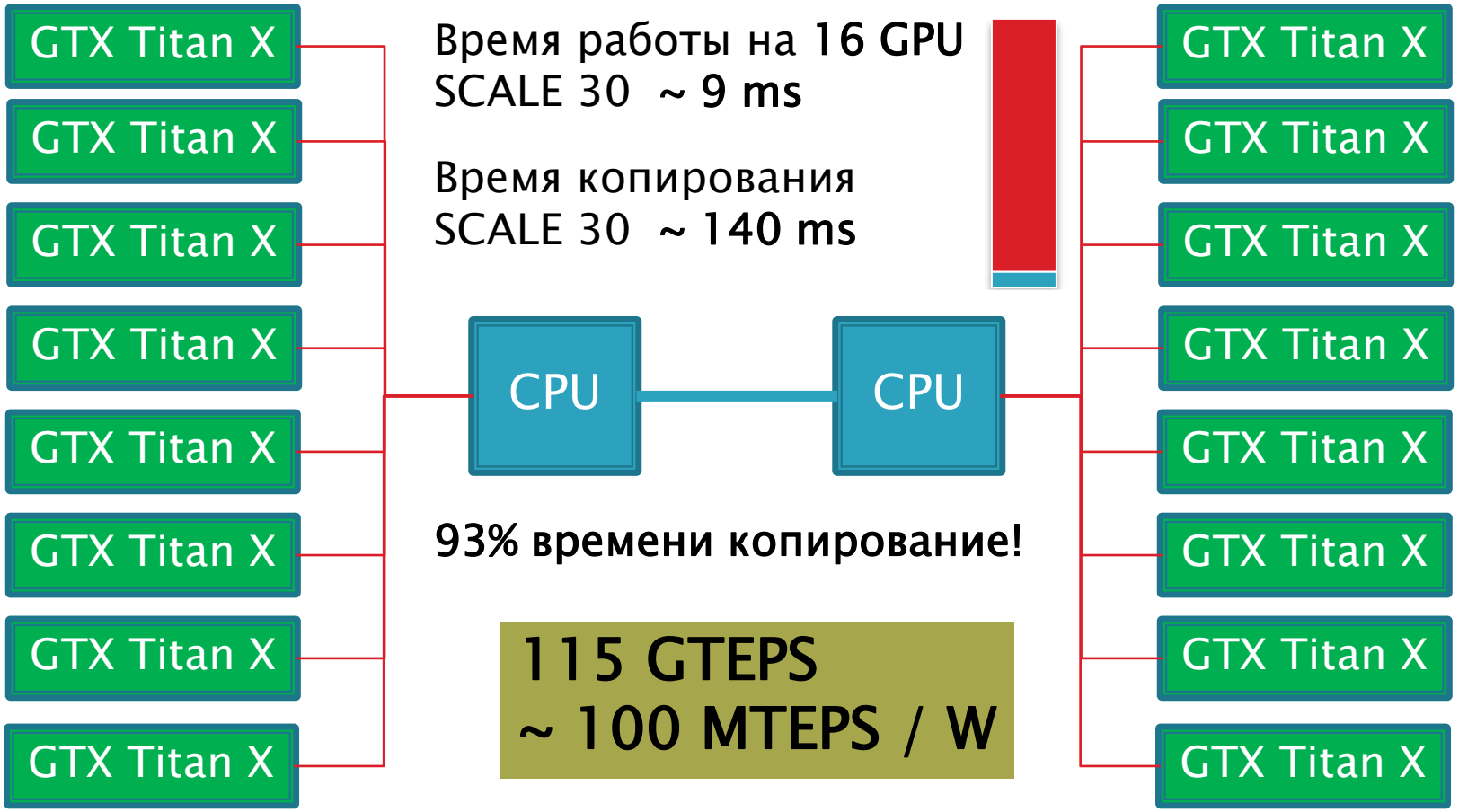
Прогнозируемое масштабирование на несколько GPU: PCIe 3.0



Прогнозируемое масштабирование на несколько GPU: PCIe 3.0

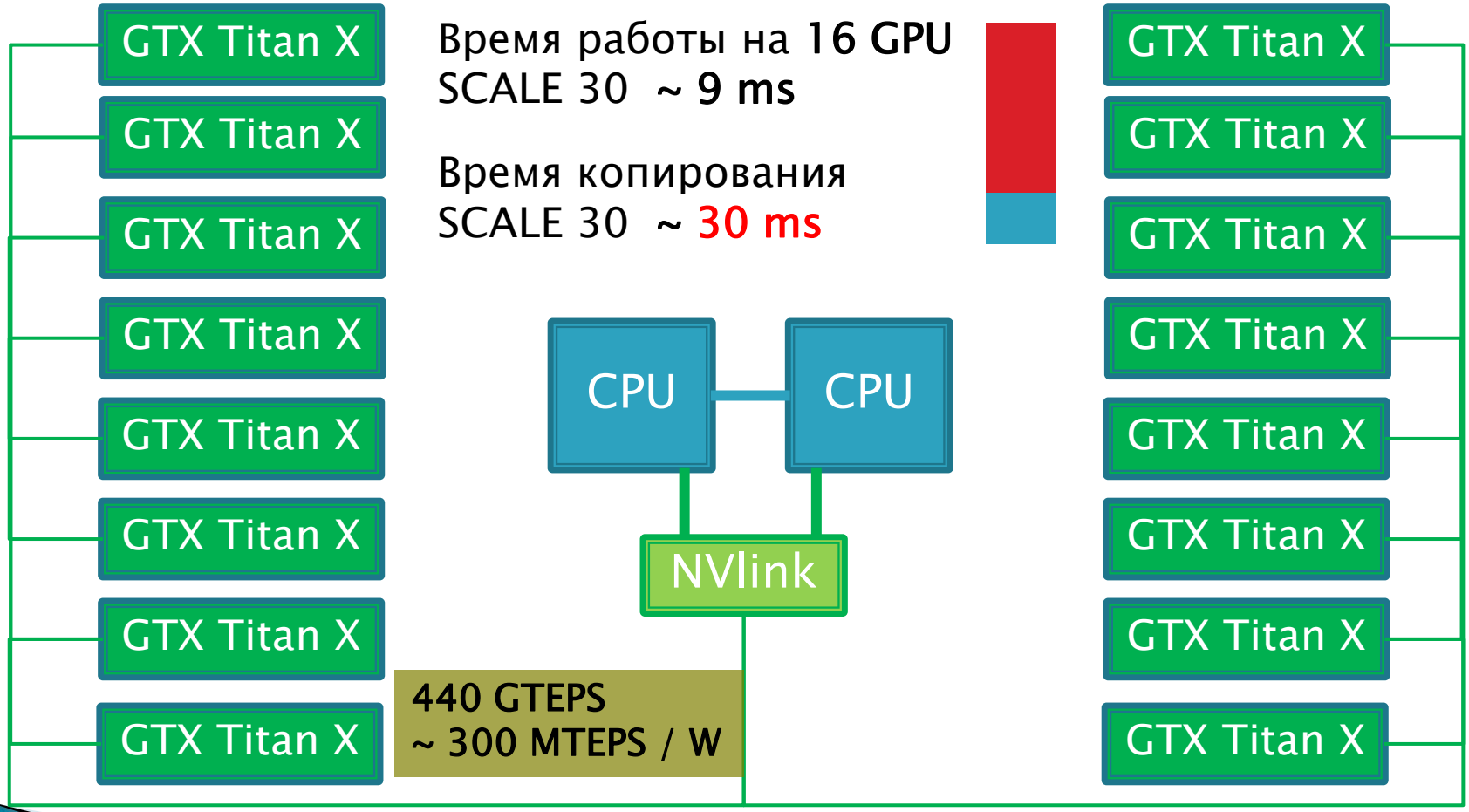


Прогнозируемое масштабирование на несколько GPU: PCIe 3.0



(1я позиция – 62,93 MTEPS / W)

Прогнозируемое масштабирование на несколько GPU: NVLINK



Спасибо за внимание!

Александр Колганов, МГУ ВМК,
alexander.k.s@mail.ru