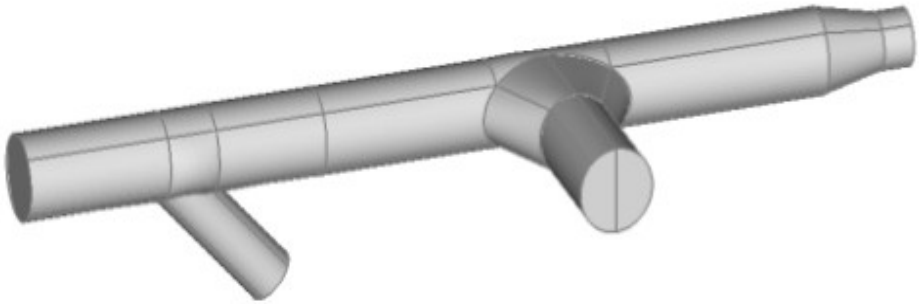
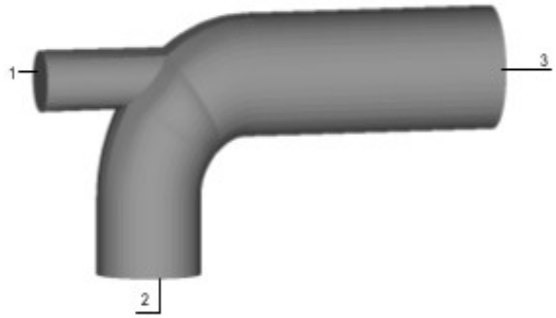
Оптимизация пересылок данных по MPI при решении системы уравнений Навье-Стокса на GPU-кластере

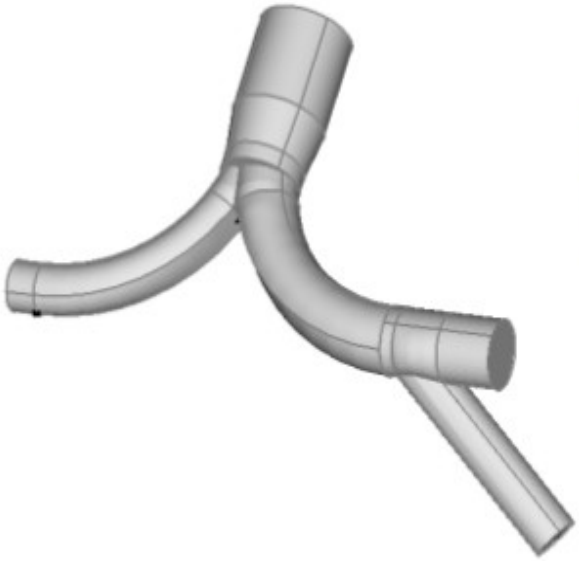
Авторы:
Кривов М.А.,
Притула М.Н.,
Гаврилов А.А.,
Дектерёв А.А.

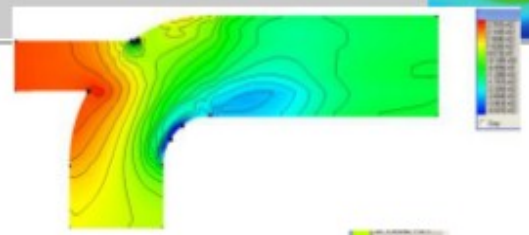
Нижний Новгород, 2013

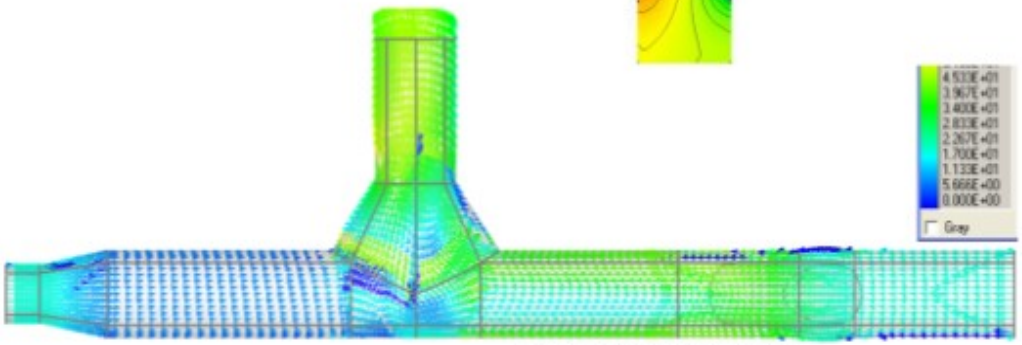
ГАЗОХОДЫ





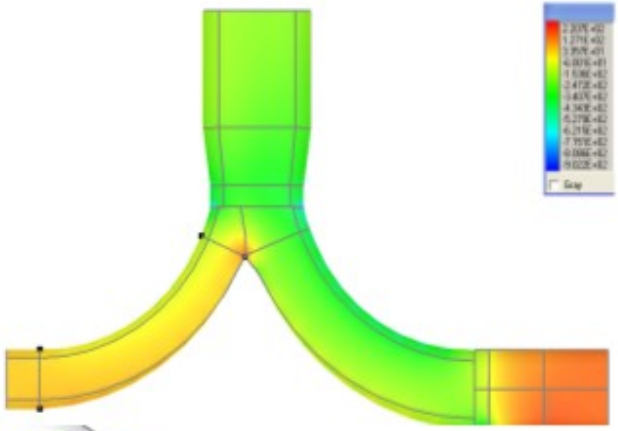


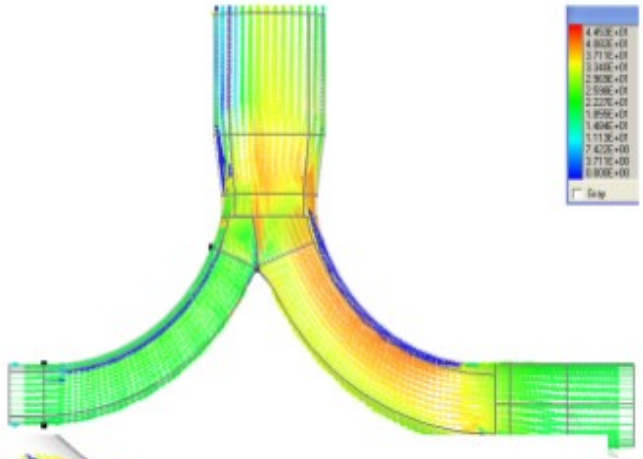


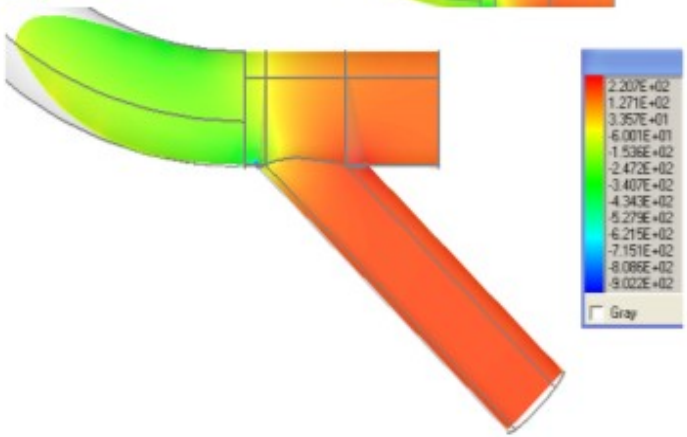


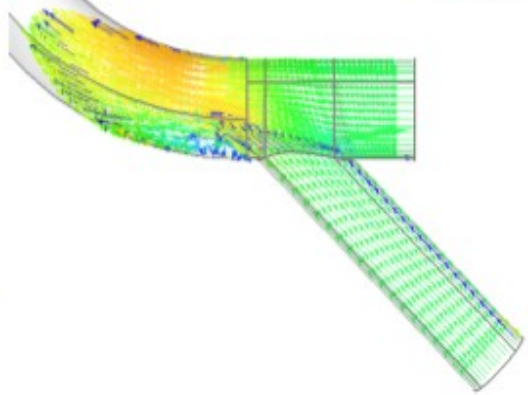
Поле давления

Поле скорости



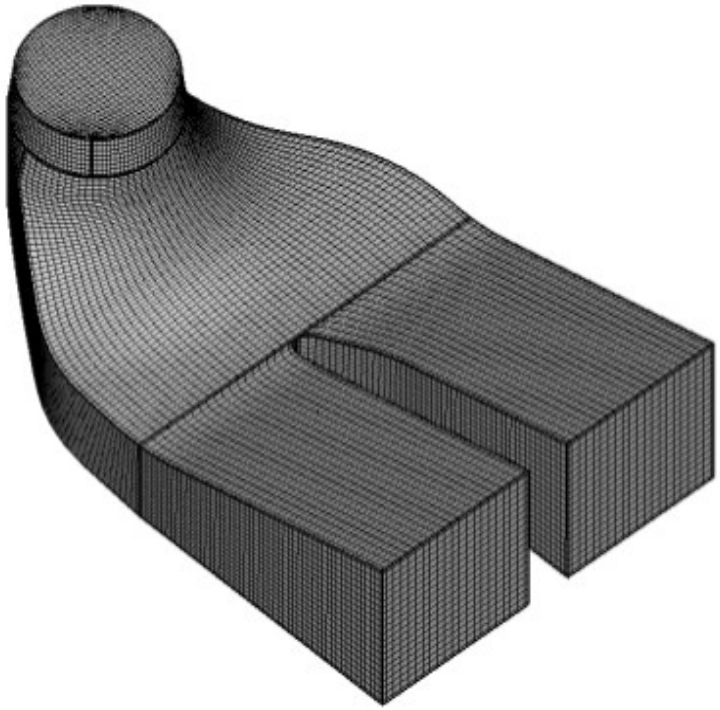




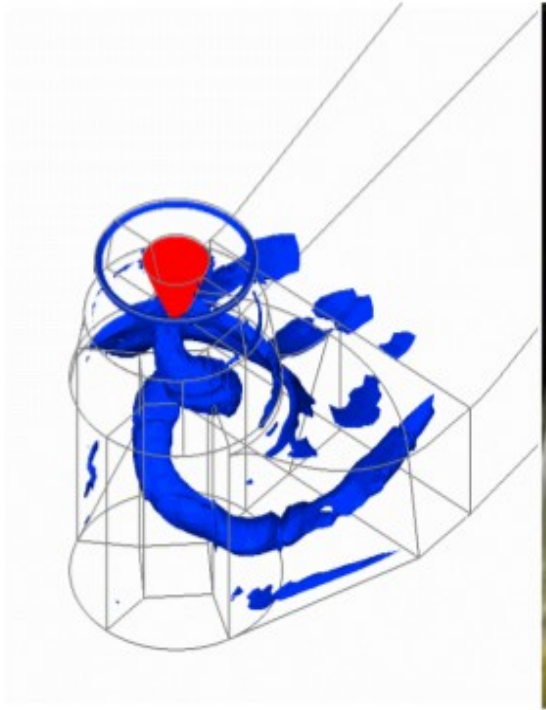


low

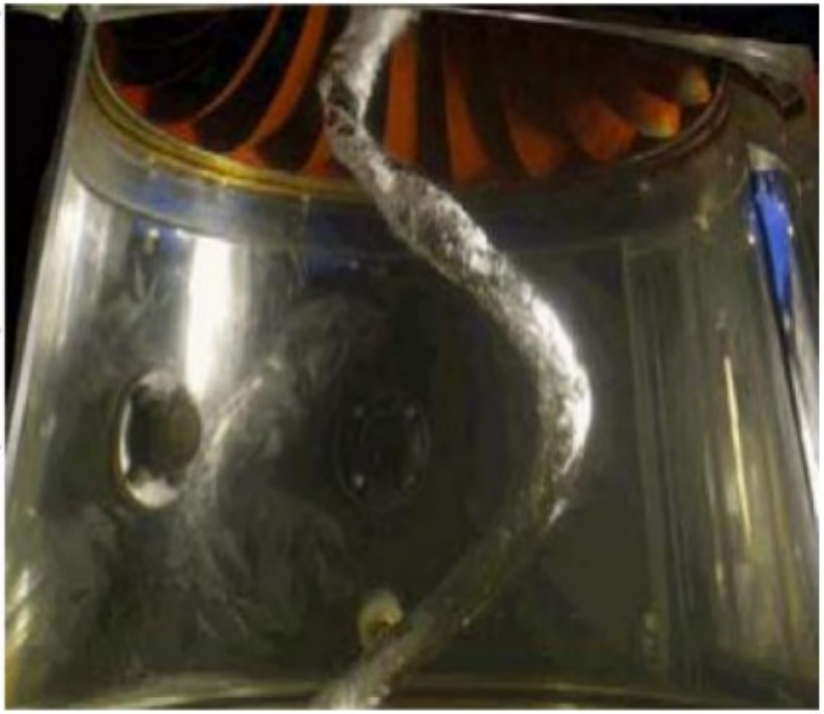
Расчет течения в отсасывающей трубе Усть-Ильимской ГЭС

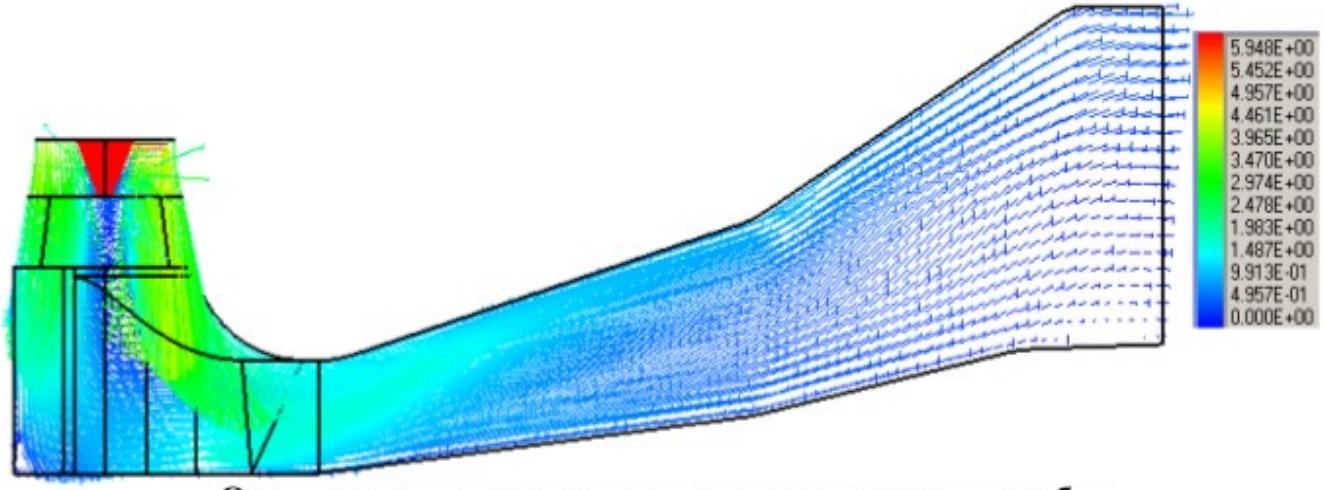


Расчетная сетка

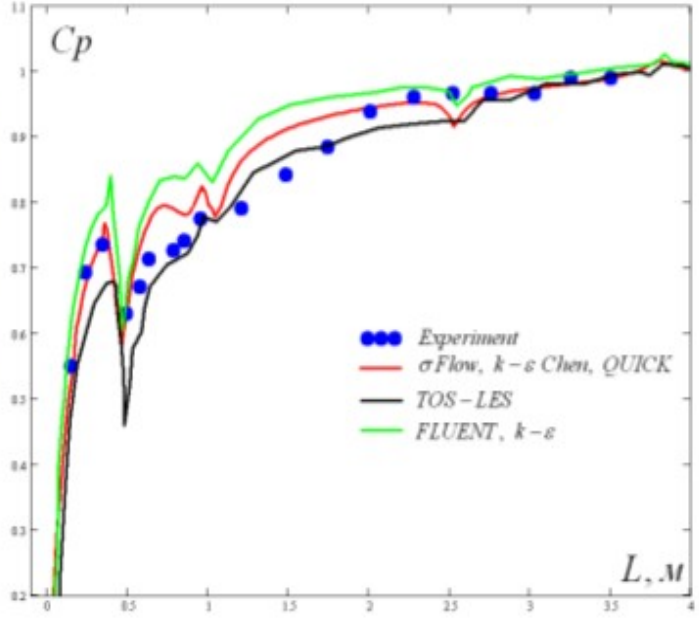


Прецессия вихревого ядра

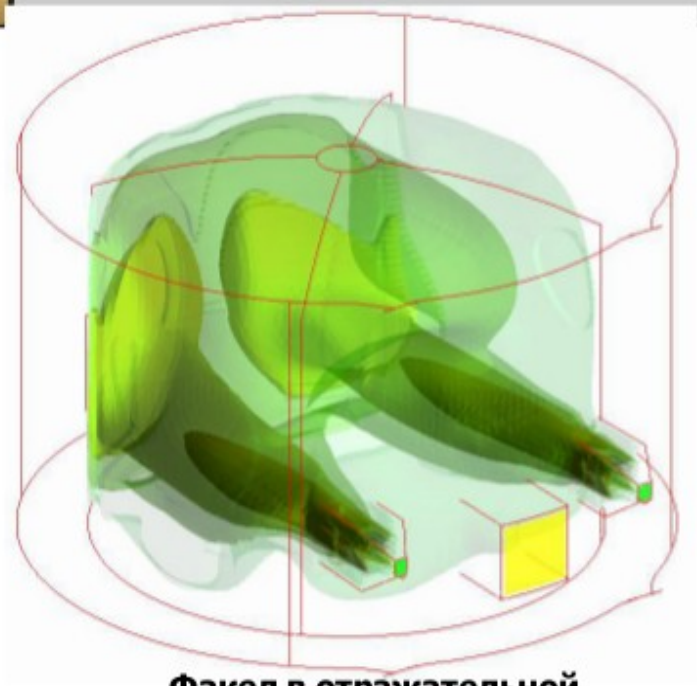




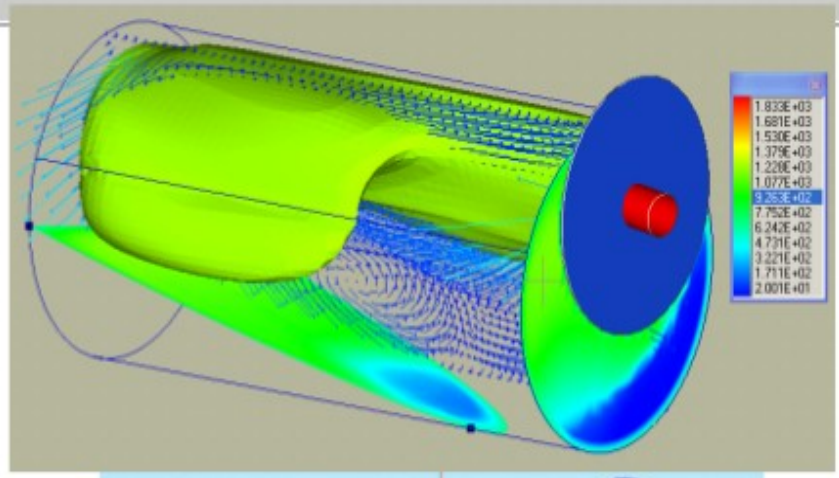
Осредненное поле вектора скорости сечении трубы

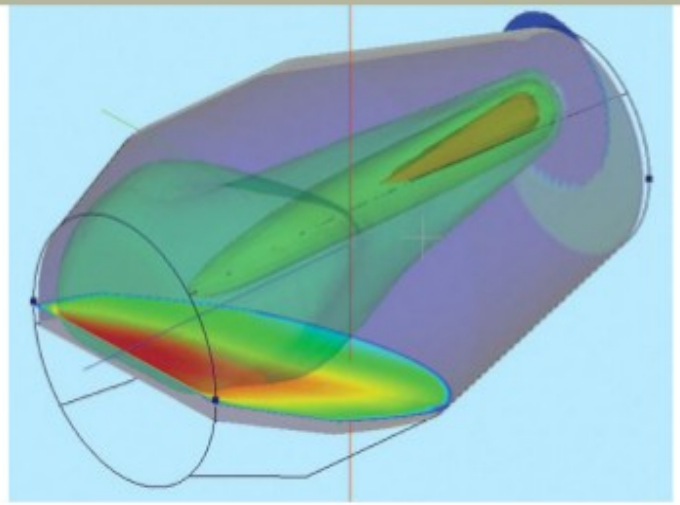


МЕТАЛЛУРГИЧЕСКИЕ ПЕЧИ

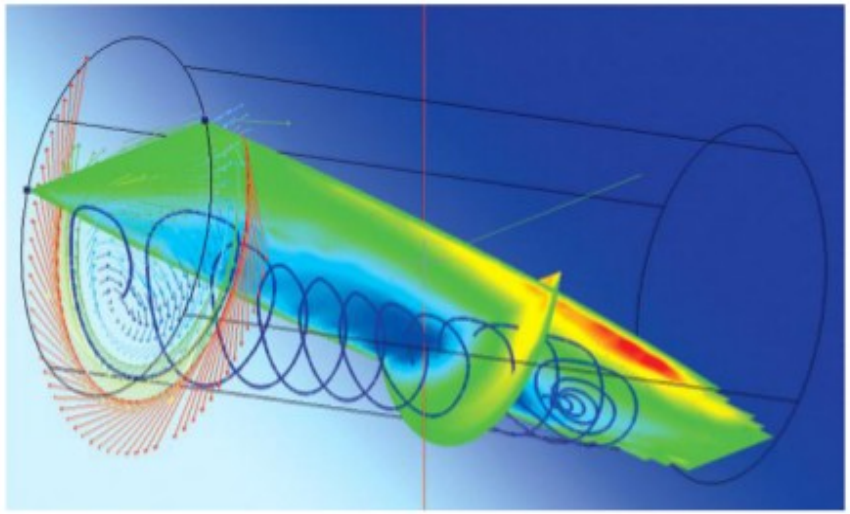


Факел в отражательной печи

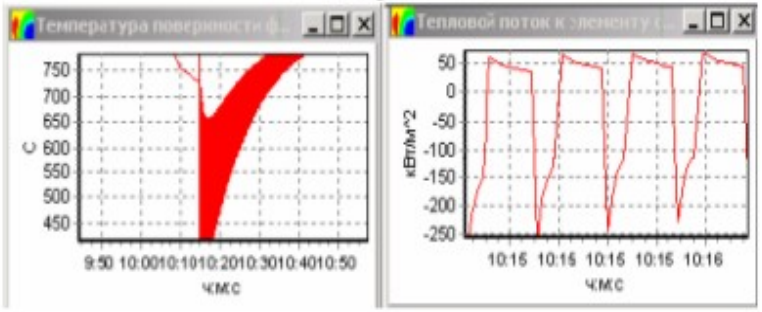




Температура факела и поверхности шихты в роторной печи



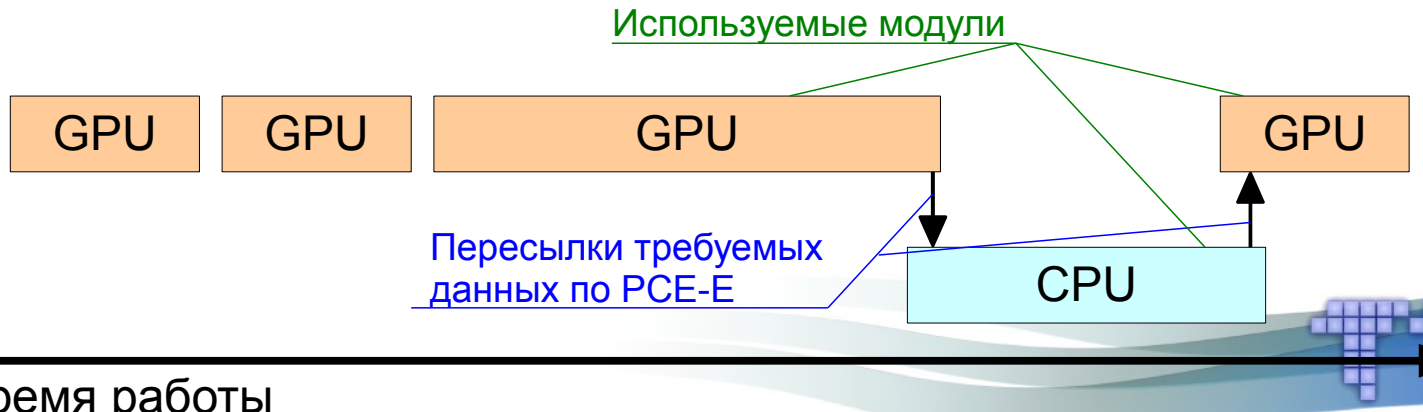
Движение металла в роторной печи



σFlow

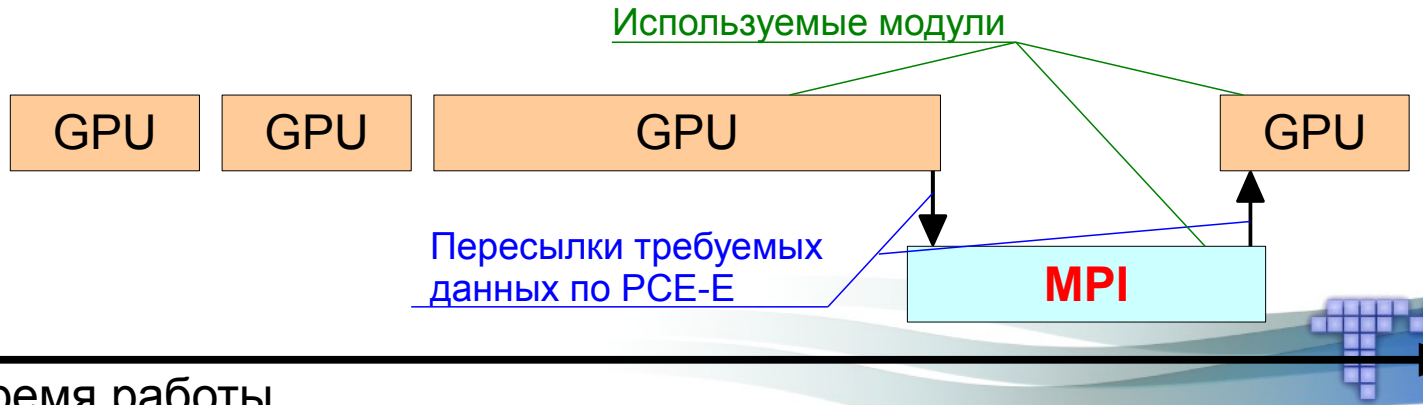
Решаемая проблема

- **Что есть:** огромный MPI-пакет, в котором часть решателей уже портированы на CUDA для работы в режиме Single-GPU
- **Что требуется:** добавить поддержку выполнения расчётов в режиме Multi-GPU
- **В чём проблема:** требуется существенно уменьшить объём пересылаемых данных по шине PCI-E.



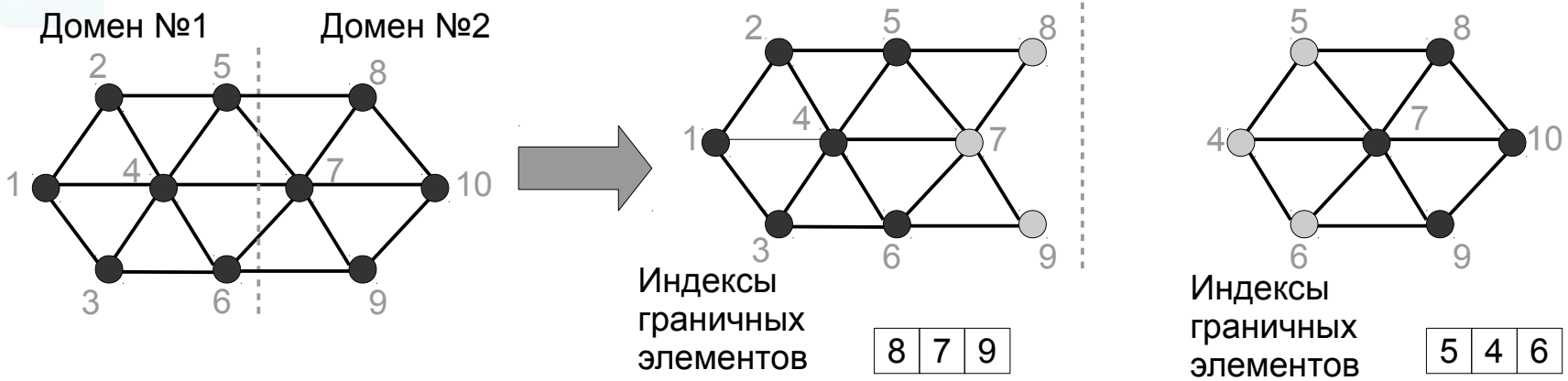
Решаемая проблема

- **Что есть:** огромный MPI-пакет, в котором часть решателей уже портированы на CUDA для работы в режиме Single-GPU
- **Что требуется:** добавить поддержку выполнения расчётов в режиме Multi-GPU
- **В чём проблема:** требуется существенно уменьшить объём пересылаемых данных по шине PCI-E.

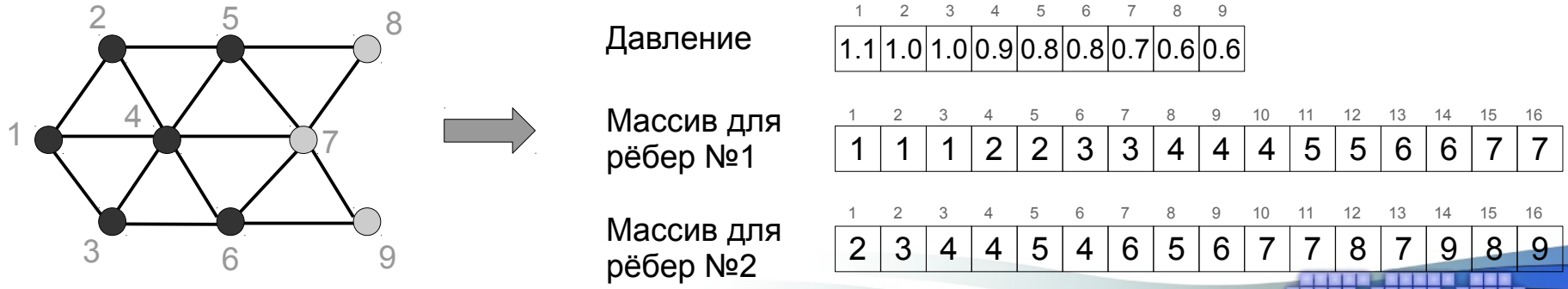


Представление данных

- Разбиение сеток на домены (делается с помощью METIS)



- Хранение сеток в памяти GPU

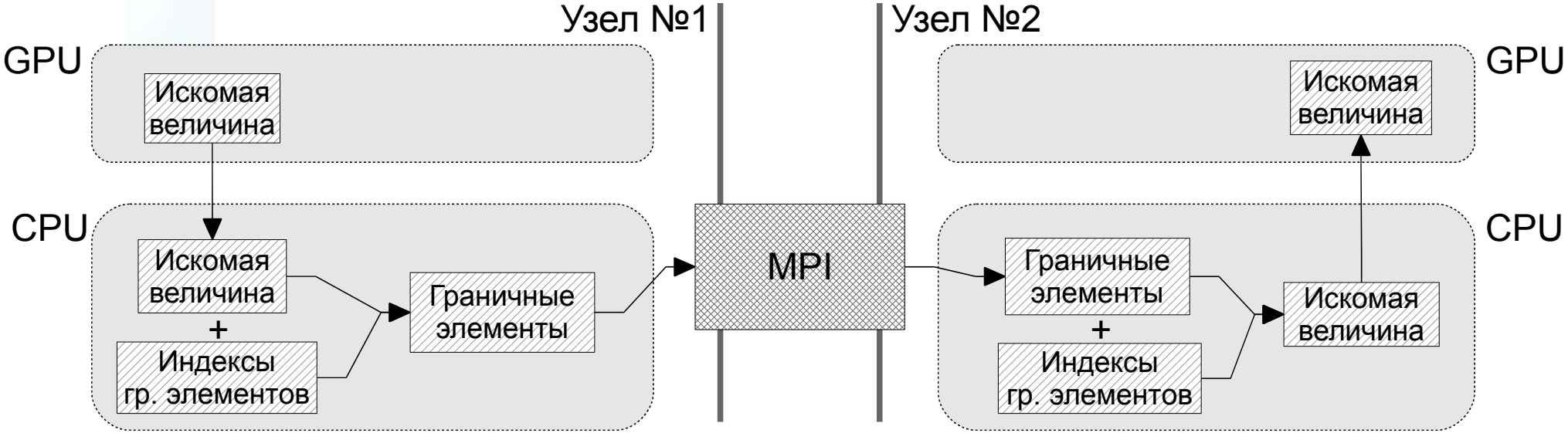


Варианты пересылок

- Вариант 1. «Original»
- Вариант 2. «Bandwidth-optimized»
- Вариант 3. «Latency-optimized»

Варианты пересылок

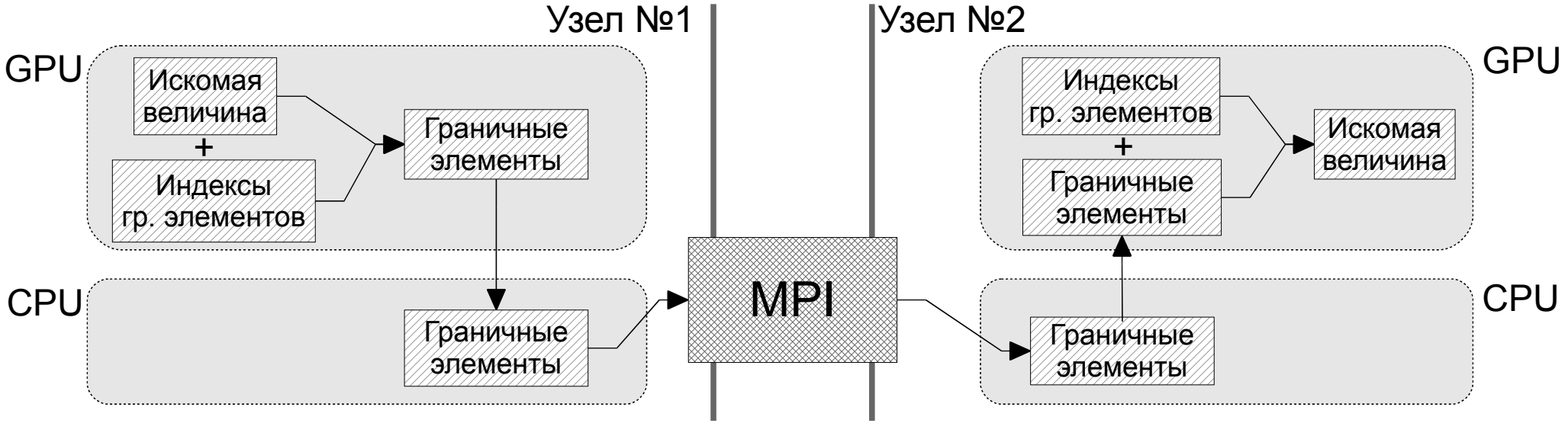
- **Вариант 1. «Original»**



- Вариант 2. «Bandwidth-optimized»
- Вариант 3. «Latency-optimized»

Варианты пересылок

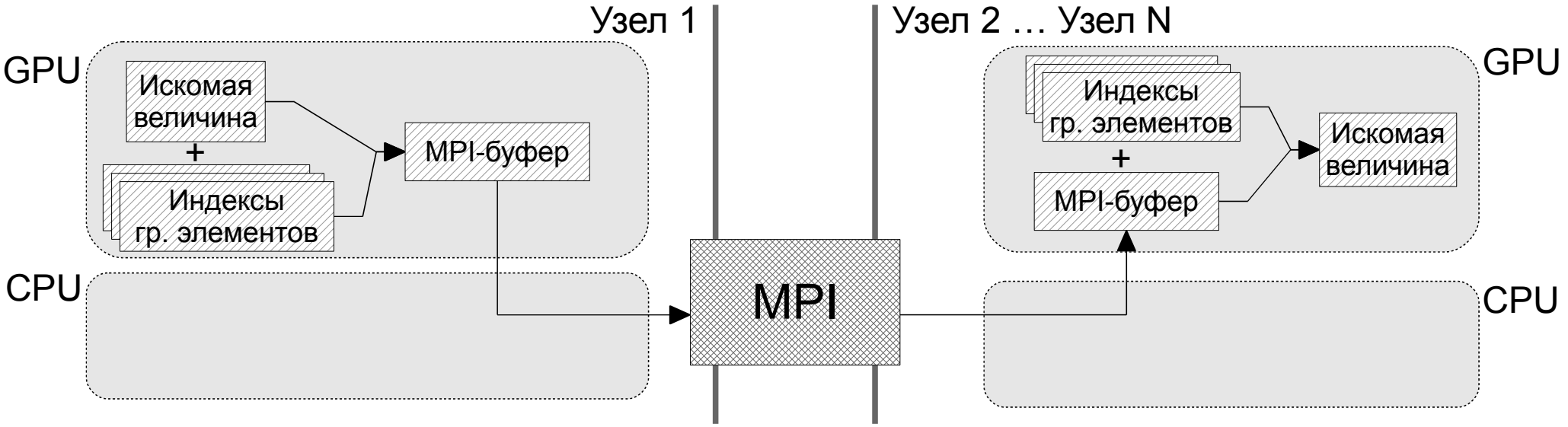
- Вариант 1. «Original»
- **Вариант 2. «Bandwidth-optimized»**



- Вариант 3. «Latency-optimized»

Варианты пересылок

- Вариант 1. «Original»
- Вариант 2. «Bandwidth-optimized»
- **Вариант 3. «Latency-optimized»**



Сравнение всех трёх реализаций

Реализация	Объём пересылаемых данных по шине PCI-E	Количество операций копирования по шине PCI-E	Асинхронность обмена по MPI
Original	Большой ($M*2*S_1$)	M * 2	Да
Bandwidth-optimized	Небольшой (порядка $M*2*S_2$)	M * (N-1) * 2	Да
Latency-optimized	Небольшой (порядка $M*2*S_2$)	2	Нет

Где

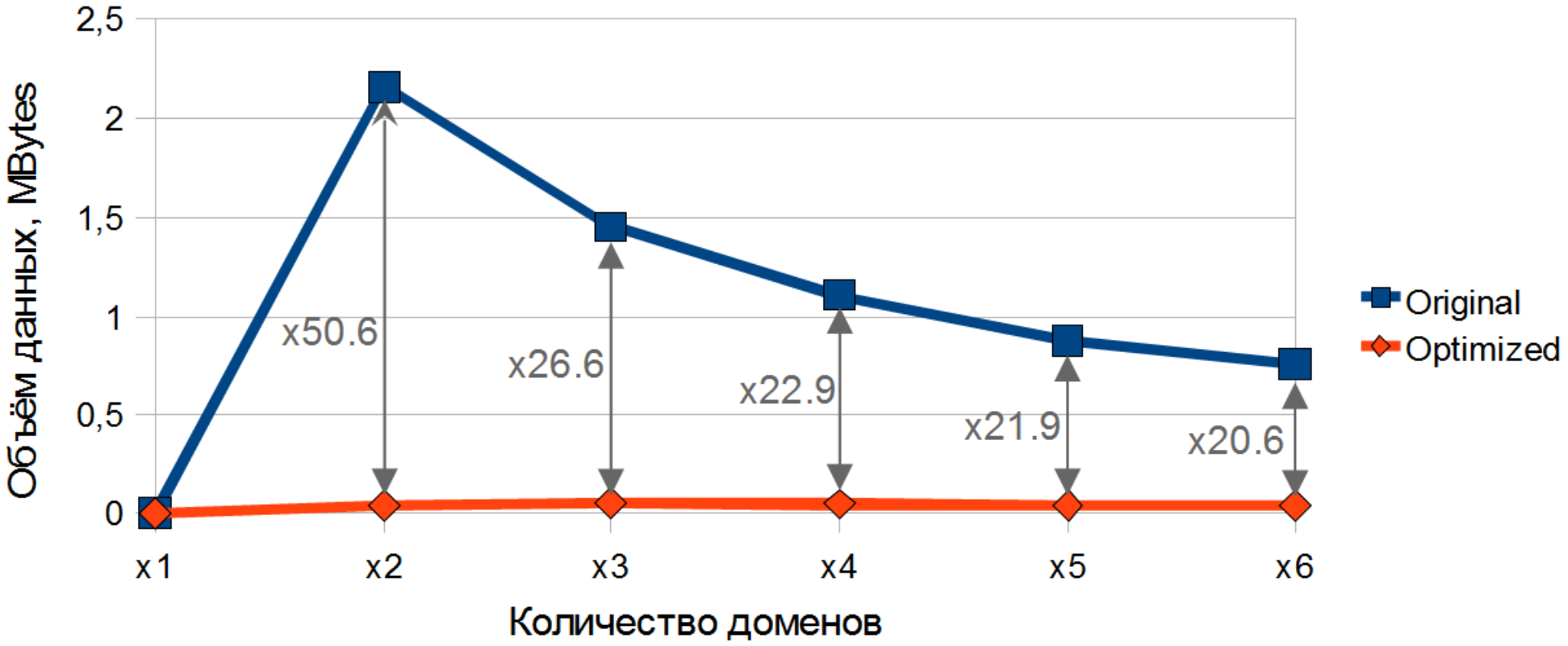
S_1 – размер массива с синхронизируемой величиной;

S_2 – размер массива со всеми граничными элементами (для всех смежных доменов);

N – количество узлов кластера;

M – число синхронизируемых скалярных величин;

Уменьшение объёма пересылаемых данных по шине PCI-E



Размер тестовой сетки — 500 тыс. узлов

Сравнение всех трёх реализаций

Реализация	Объём пересылаемых данных по шине PCI-E	Количество операций копирования по шине PCI-E	Асинхронность обмена по MPI
~~**Original**~~	Большой ($M*2*S_1$)	M * 2	Да
Bandwidth-optimized	Небольшой (порядка $M*2*S_2$)	M * (N-1) * 2	Да
Latency-optimized	Небольшой (порядка $M*2*S_2$)	2	Нет

Где

S_1 – размер массива с синхронизируемой величиной;

S_2 – размер массива со всеми граничными элементами (для всех смежных доменов);

N – количество узлов кластера;

M – число синхронизируемых скалярных величин;

Выигрыш от укрупнения блоков при копировании по PCI-E*

Размер в KiB \ Число блоков	4	6	8	10	12	14	16	20	24	28	32	40	48	64	72	80
4	-45,2	-39,502	-32,719	-27,929	-22,525	-19,946	-15,135	-6,717	-2,302	3,408	10,45	19,944	26,134	38,38	41,208	45,061
6	-46,033	-40,206	-33,656	-29,171	-25,79	-20,944	-19,089	-12,183	-6,452	-3,223	3,223	9,69	14,06	22,524	29,339	29,291
8	-44,408	-38,966	-33,192	-31,016	-27,235	-23,164	-19,774	-15,958	-9,804	-5,625	-1,825	4,713	9,547	15,9	19,734	20,859
10	-44,441	-38,966	-35,064	-31,571	-27,992	-24,667	-20,823	-16,035	-11,114	-8,764	-5,374	-2,338	2,127	5,207	11,82	10,903
12	-45,982	-40,506	-36,467	-33,304	-27,248	-26,313	-24,883	-20,119	-16,468	-13,684	-9,918	-7,332	1,991	-0,134	2,427	7,238
14	-44,349	-40,215	-35,33	-33,852	-29,936	-26,895	-23,279	-20,293	-16,8	-13,891	-11,255	-8,432	-8,456	-3,662	-3,103	-0,088
16	-46,17	-40,65	-35,585	-32,861	-30,963	-27,742	-27,269	-21,598	-19,445	-16,61	-15,3	-14,049	-12,063	-8,991	-3,906	-3,959
24	-45,334	-41,734	-38,158	-33,998	-31,381	-30,489	-27,447	-25,751	-24,038	-23,03	-22,393	-20,525	-17,686	-15,099	-16,035	-14,736
32	-44,645	-40,998	-38,628	-33,931	-34,345	-30,21	-29,812	-24,673	-29,064	-26,867	-26,178	-26,224	-25,275	-22,487	-21,464	-20,059
48	-45,97	-40,995	-36,827	-34,871	-32,304	-36,351	-31,756	-31,193	-30,407	-30,345	-26,017	-27,423	-26,927	-28,357	-24,714	-26,613
64	-46,055	-39,392	-39,067	-35,624	-36,758	-37,318	-34,102	-35,477	-31,316	-32,889	-33,375	-31,639	-30,939	-31,96	-32,311	-31,502
128	-42,544	-43,402	-39,617	-38,86	-37,985	-37,271	-39,746	-34,455	-36,233	-36,235	-40,641	-38,416	-40,627	-45,558	-46,201	-46,369
256	-41,578	-42,958	-40,463	-41,497	-40,056	-41,394	-45,082	-39,742	-43,521	-45,195	-47,605	-50,474	-50,674	-51,026	-49,785	-49,2
512	-44,464	-41,648	-39,921	-44,595	-46,056	-49,19	-49,506	-51,255	-50,719	-51,458	-52,261	-50,591	-52,826	-51,673	-51,98	-52,79
768	-45,47	-48,525	-45,657	-49,459	-50,774	-52,863	-53,733	-54,575	-53,047	-55,379	-53,661	-51,767	-52,824	-54,503	-51,885	-54,651
1024	-45,58	-45,929	-48,905	-51,855	-52,933	-53,994	-54,46	-54,971	-55,577	-53,833	-53,357	-55,3	-53,819	-54,386	-55,967	-54,27

Изменение скорости копирования в процентах

*Результаты, представленные на данном слайде, были получены на кафедре системного программирования ВМиК МГУ, и к докладчику никого отношения не имеют. Он их просто утащил.

Сравнение всех трёх реализаций

Реализация	Объём пересылаемых данных по шине PCI-E	Количество операций копирования по шине PCI-E	Асинхронность обмена по MPI
Original	Большой ($M*2*S_1$)	M * 2	Да
Bandwidth-optimized	Небольшой (порядка $M*2*S_2$)	M * (N-1) * 2	Да
Latency-optimized	Небольшой (порядка $M*2*S_2$)	2	Нет

Где

S_1 – размер массива с синхронизируемой величиной;

S_2 – размер массива со всеми граничными элементами (для всех смежных доменов);

N – количество узлов кластера;

M – число синхронизируемых скалярных величин;

Сравнение всех трёх реализаций

Реализация	Объём пересылаемых данных по шине PCI-E	Количество операций копирования по шине PCI-E	Асинхронность обмена по MPI
Original	Большой ($M*2*S_1$)	M * 2	Да
Bandwidth-optimized	Небольшой (порядка $M*2*S_2$)	M * (N-1) * 2	Да
Latency-optimized	Небольшой (порядка $M*2*S_2$)	2	Нет

Где

S_1 – размер массива с синхронизируемой величиной;

S_2 – размер массива со всеми граничными элементами (для всех смежных доменов);

N – количество узлов кластера;

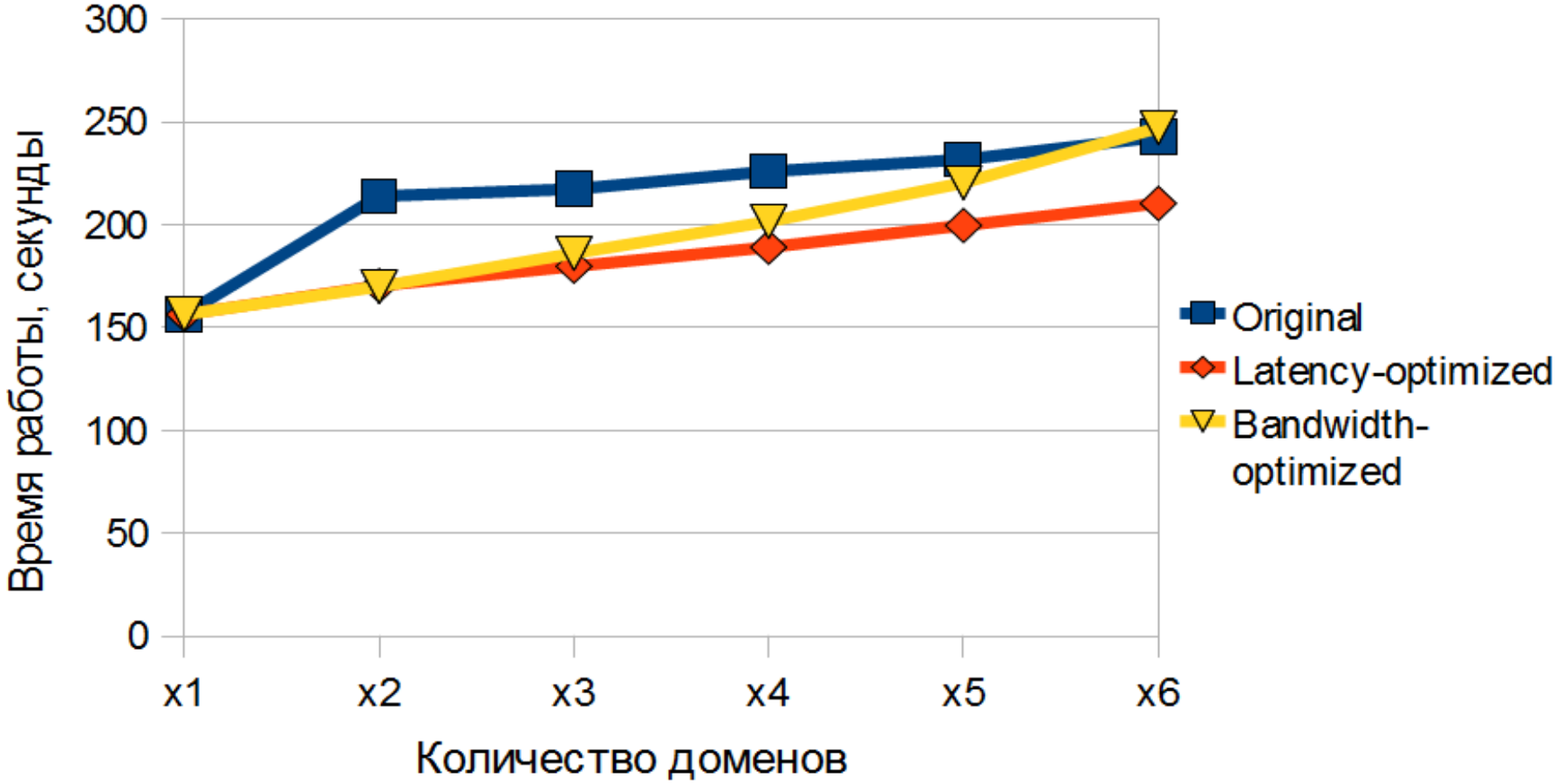
M – число синхронизируемых скалярных величин;

Верны ли оценки?

Верны ли оценки?

Как оказалось, нет...

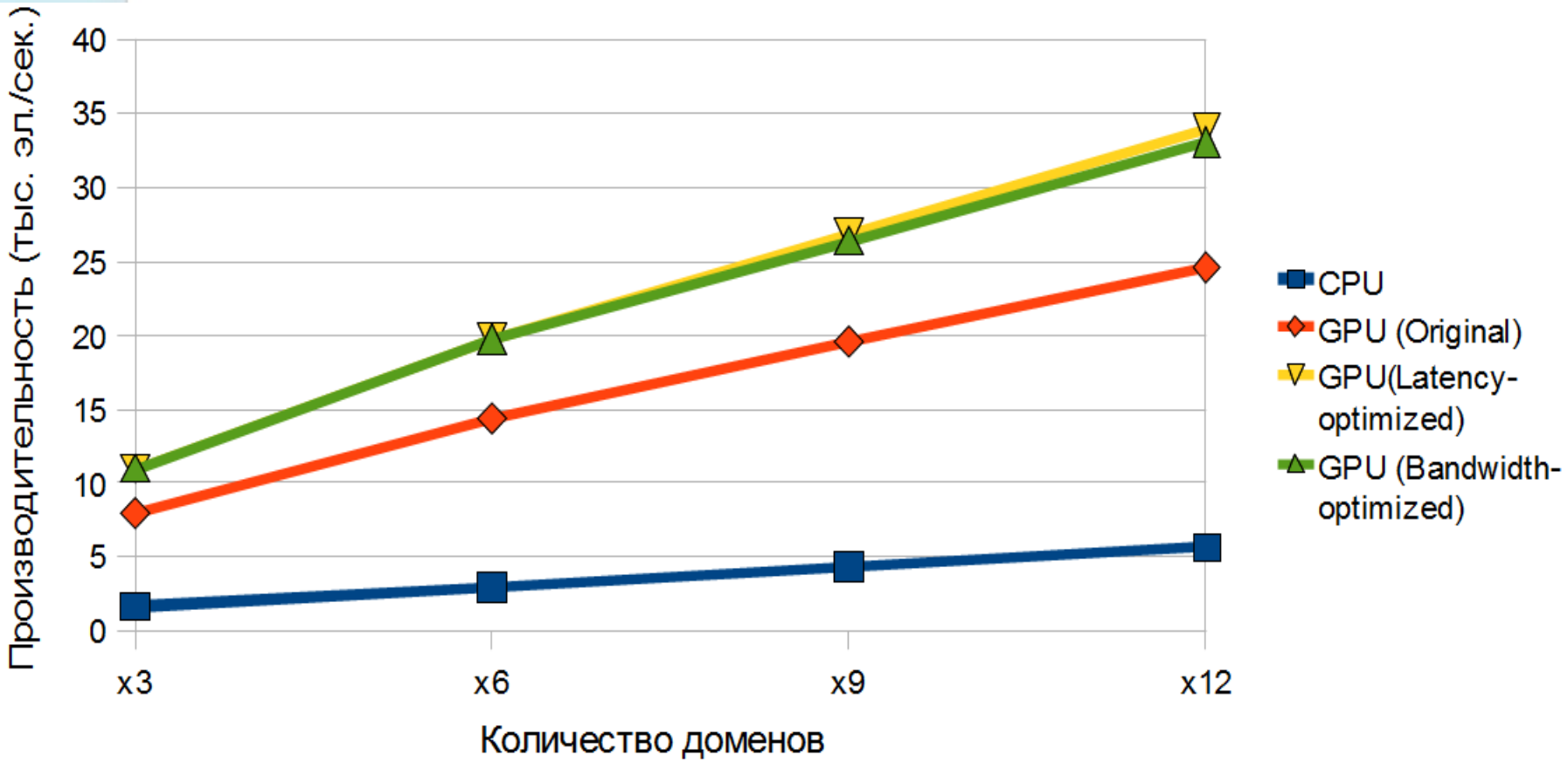
Тестирование: несколько процессов на одном GPU



Размер сетки — 2 млн. узлов

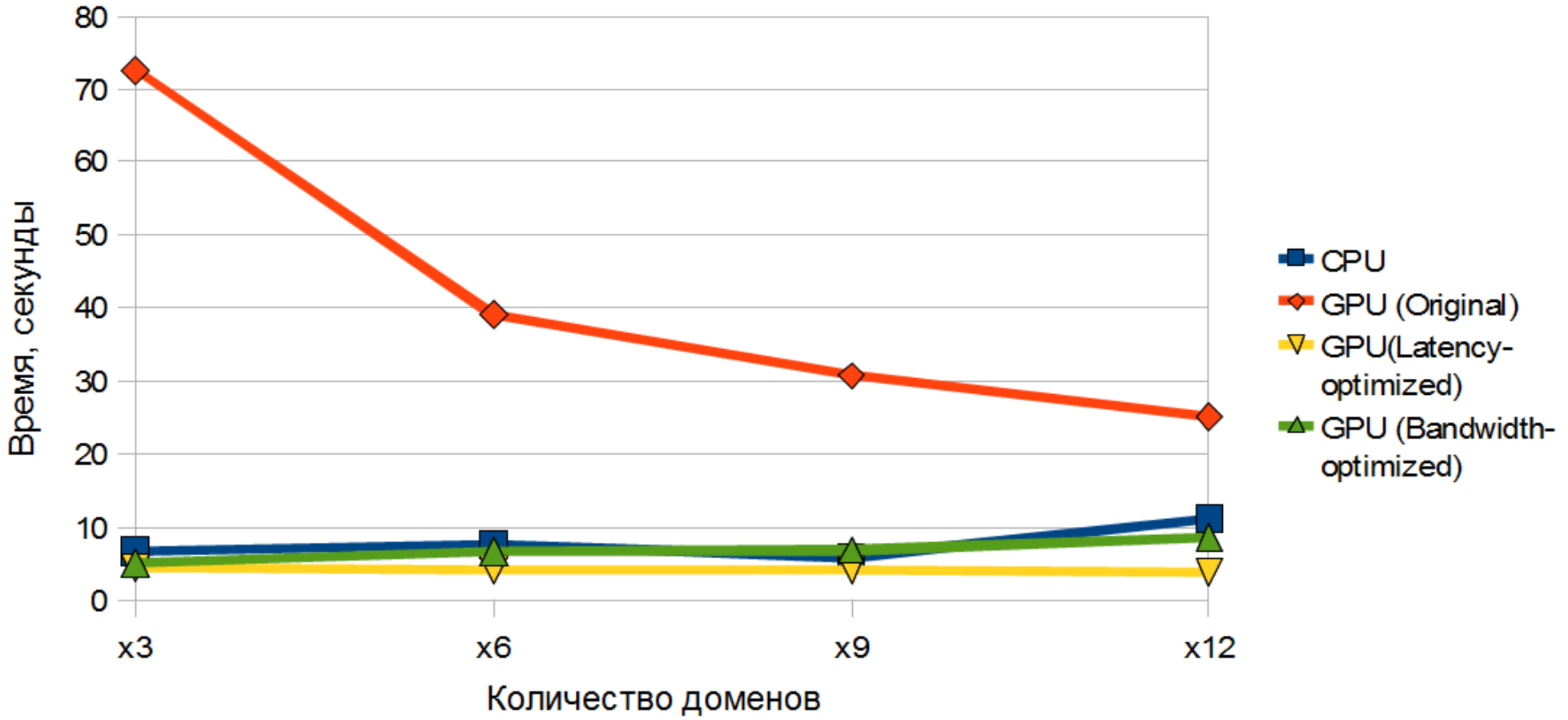
Аномалия: накладные расходы в версии «Bandwidth-optimized» увеличиваются намного быстрее, чем в «Original»

Тестирование: GPU-кластер



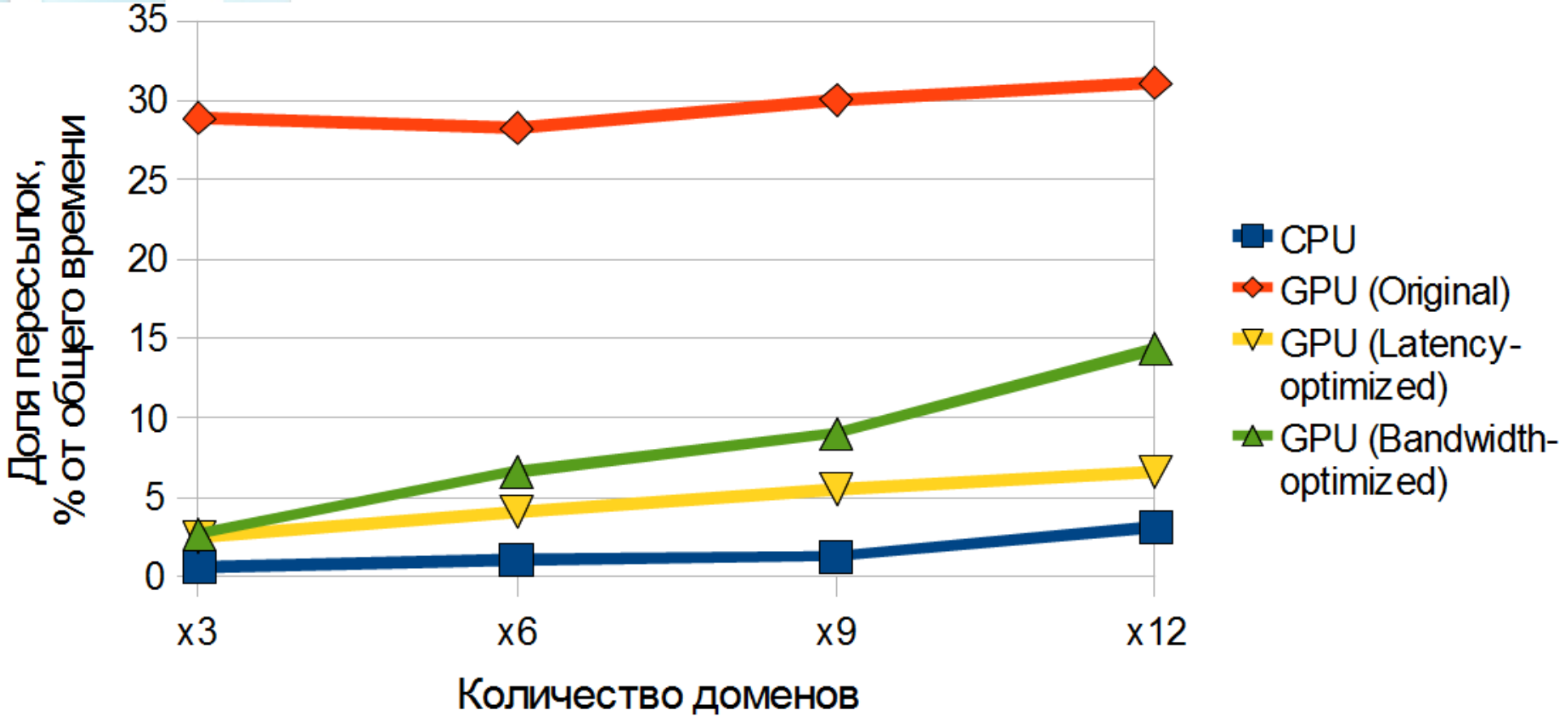
Производительность всех реализаций
Размер сетки — 2 млн. узлов

Тестирование: GPU-кластер



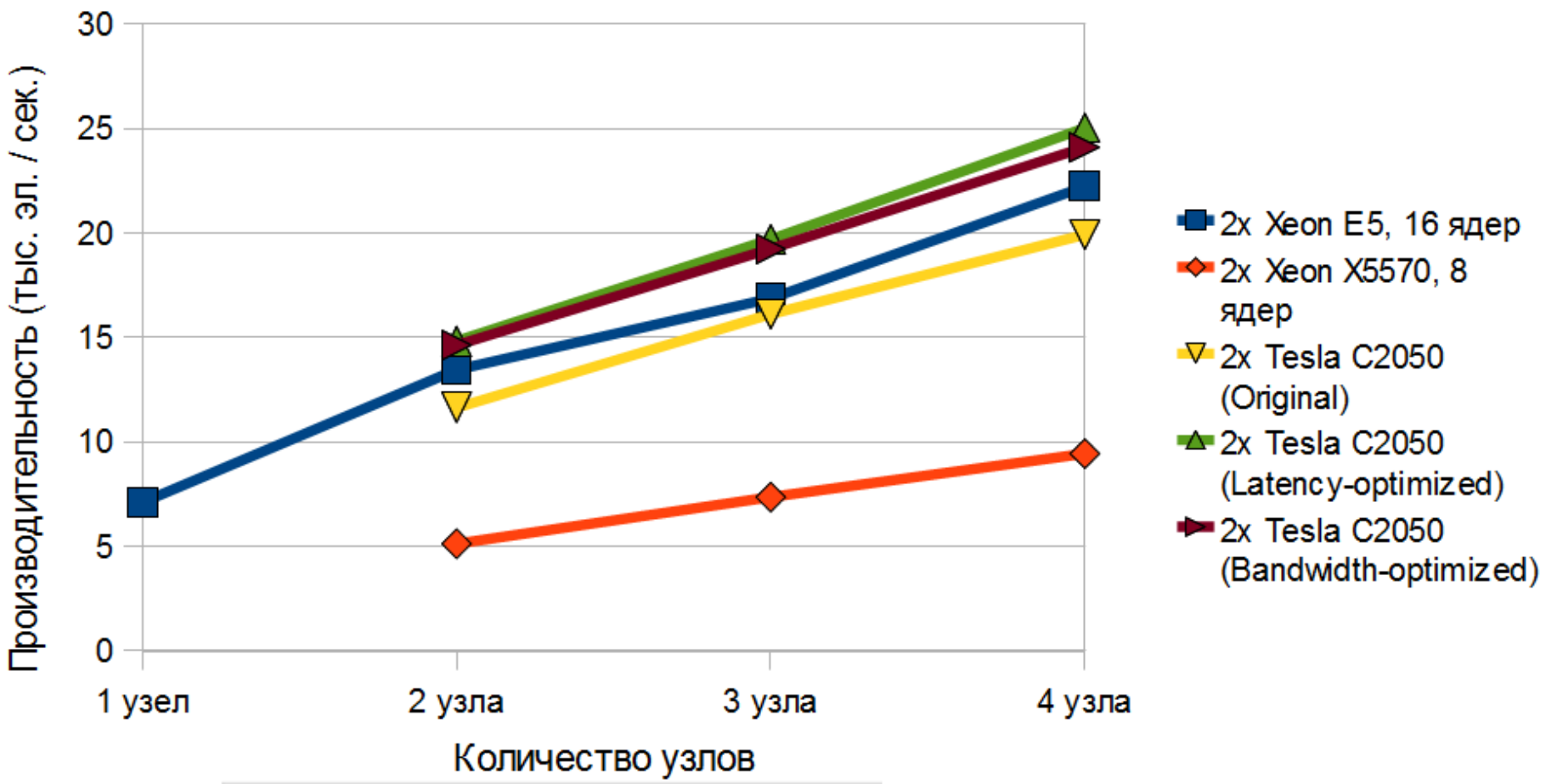
Время выполнения операций обмена
Размер сетки — 2 млн. узлов

Тестирование: GPU-кластер



Вклад операций обмена в общее время работы
Размер сетки — 2 млн. узлов

Тестирование: GPU-облако Amazon EC2



Производительность всех реализаций
Размер сетки — 8 млн. узлов

Мораль

- Для реальной и упрощённой задач могут требоваться разные подходы к оптимизации
- В зависимости от сценария использования, исходная версия программы может оказаться даже быстрее, чем оптимизированная
- Благодаря проведённым изменениям удалось сократить время пересылок данных в 7.5 раза, или с 30% до 6% от общего времени работы

Вопросы?

(m_krivov@ttgLabs.com)