11 декабря, 2017, Москва

Оптимизация топологии размещения MPI-процессов на кластерах с интерконнектом Ангара

М. Халилов, А. Тимофеев

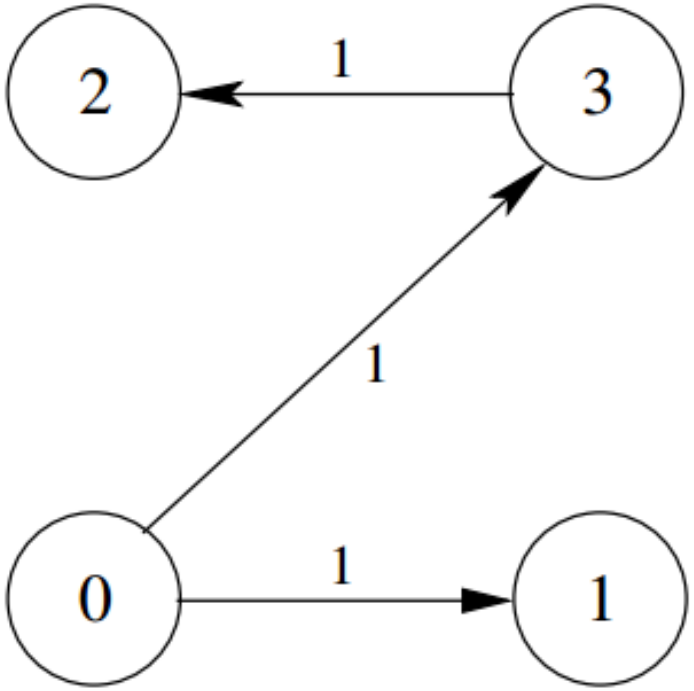
Московский институт электроники и математики им. А.Н. Тихонова ВШЭ

Московский физико-технический институт

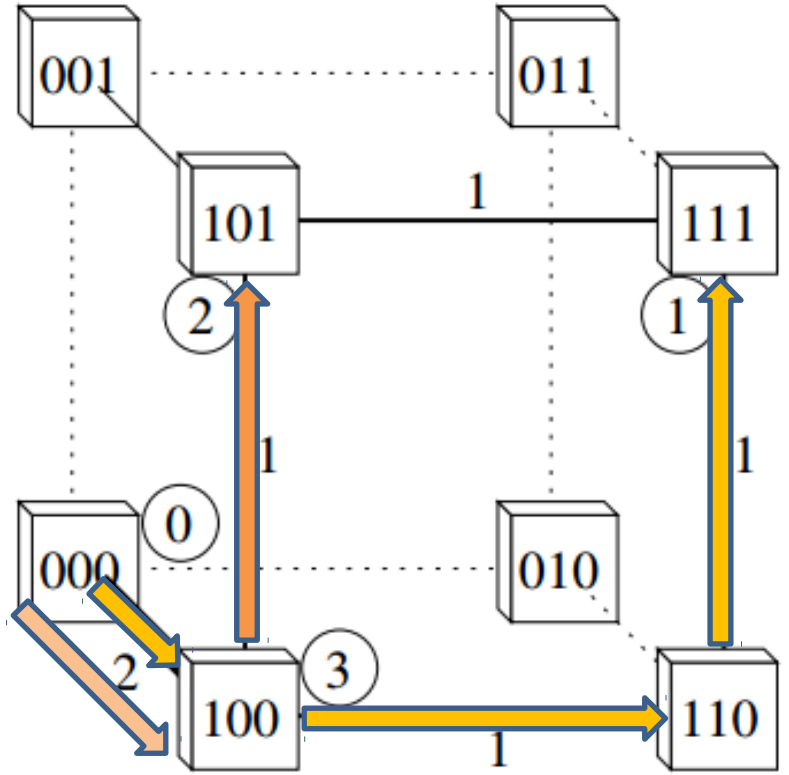
План доклада

- Задача оптимального отображения процессов
- Описание метода оптимизации
- Описание работы разработанной библиотеки оптимизации
- Тестирование и анализ полученных результатов
- Выводы

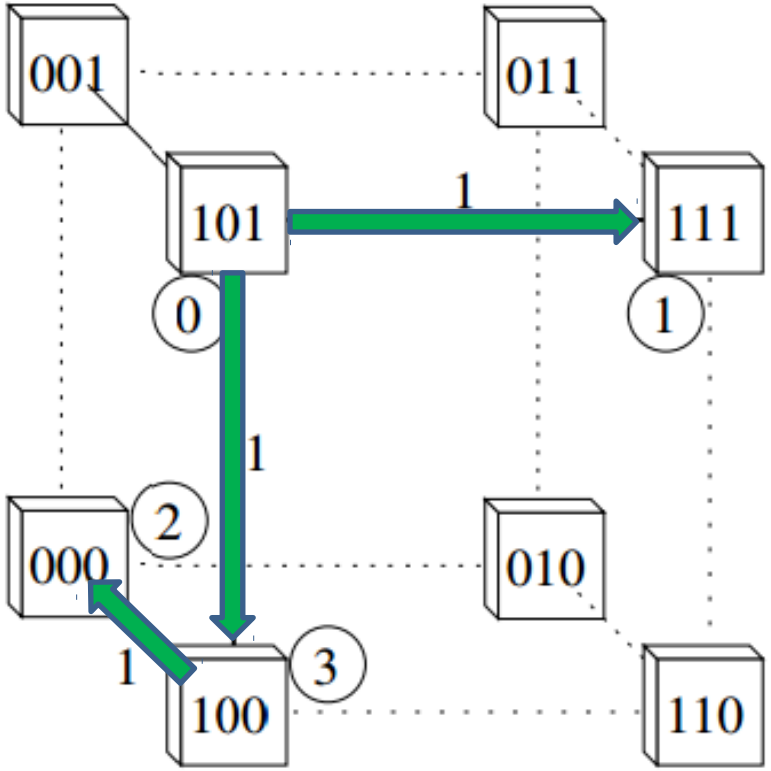
Задача поиска оптимального мэппинга MPI-программы



(b) Process Topology $\mathcal{G}$.



(c) Mapping Γ_1.



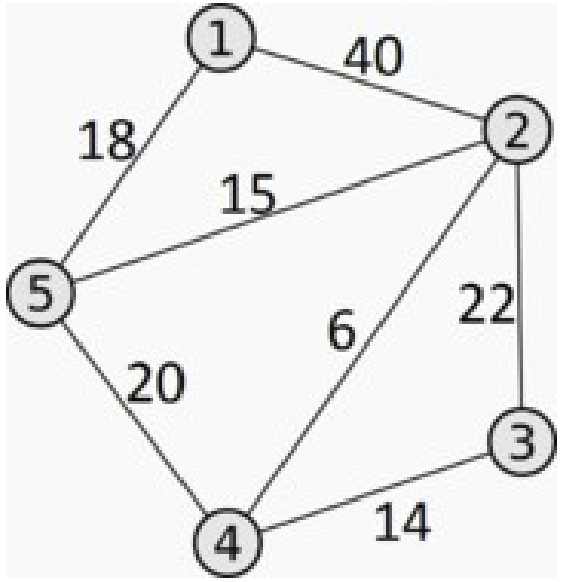
(d) Mapping Γ_2.

- Имеется **4 MPI-процесса** с рангами: 0, 1, 2, 3 (рис. b) и **8 узлов**;
- **Неоптимальное** отображение процессов (рис. c);
- **Оптимальное** отображение (рис. d);

Форматы представления информации о коммуникационных обменах

$$\begin{pmatrix} 0 & 40 & \infty & \infty & 18 \\ 40 & 0 & 22 & 6 & 15 \\ \infty & 22 & 0 & 14 & \infty \\ \infty & 6 & 14 & 0 & 20 \\ 18 & 15 & \infty & 20 & 0 \end{pmatrix}$$

Матрица коммуникационного паттерна

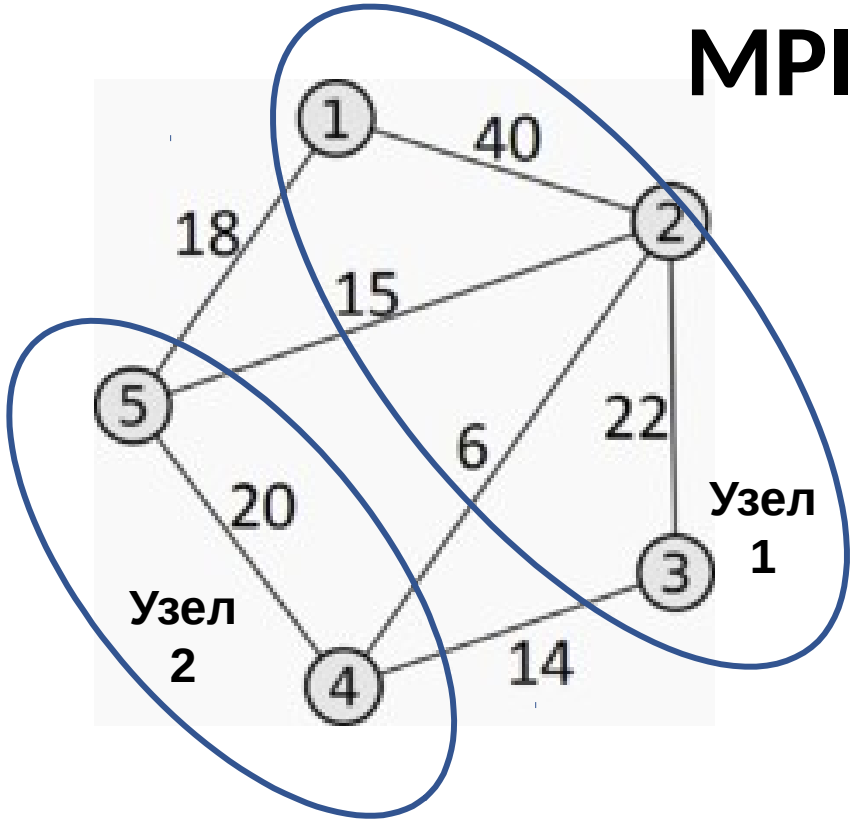


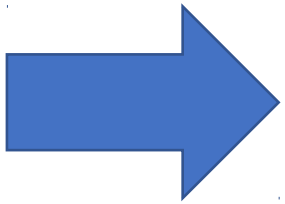
Информационный граф

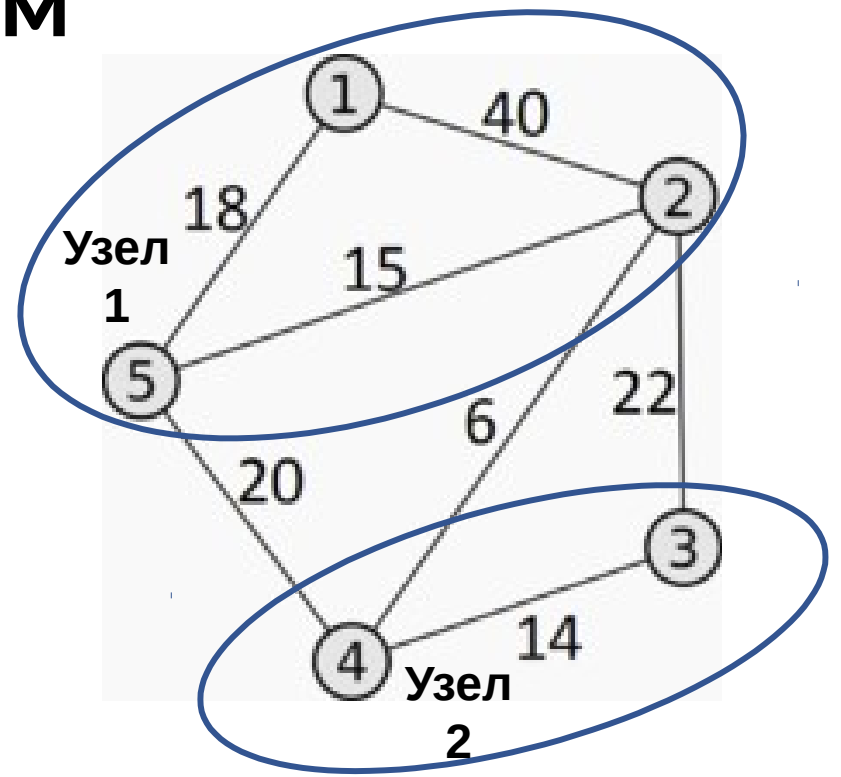
Строки — процессы-отправители
Столбцы — процессы-получатели
Элементы — количество переданных слов/интенсивность обменов

Вершины — MPI-процессы
Рёбра — наличие обменов
Веса рёбер — количество переданных слов/интенсивность обменов

Описание метода оптимизации отображения MPI-программ







Разбиение А

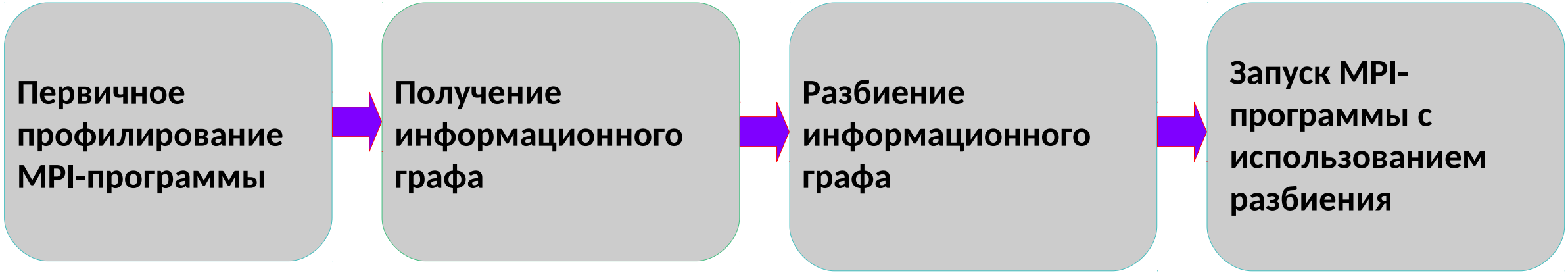
S_a = 18 + 15 + 6 + 14 = **53**

Разбиение Б

S_b = 20 + 6 + 22 = **48**

Разбиение выполняется для **минимизации суммы** рёбер, соединяющих разные подмножества разбиения

Функциональная схема работы библиотеки оптимизации запуска MPI-программ



Характеристики кластеров "Десмос" и "Ангара К-1"

Кластер "Десмос" ОИВТ РАН:

- 32 вычислительных узла
- Топология 4D-тор 4 x 2 x 2 x 2

Каждый из узлов оборудован:

- Intel Xeon E5-1650 (6 cores), 3.5 GHz
- 16 GB ОЗУ
- сетевой адаптер сети Ангара на базе СБИС EC8430

Кластер "Ангара К-1" НИЦЭВТ:

- 24 узла А-типа, 12 узлов Б-типа
- Топология 3D-тор 3 x 3 x 4

Узлы А-типа оборудованы:

- 2x Intel Xeon CPU E5-2630 (6 cores), 2.30 GHz
- 64 GB ОЗУ
- сетевой адаптер сети Ангара на базе СБИС EC8430

Узлы Б-типа оборудованы:

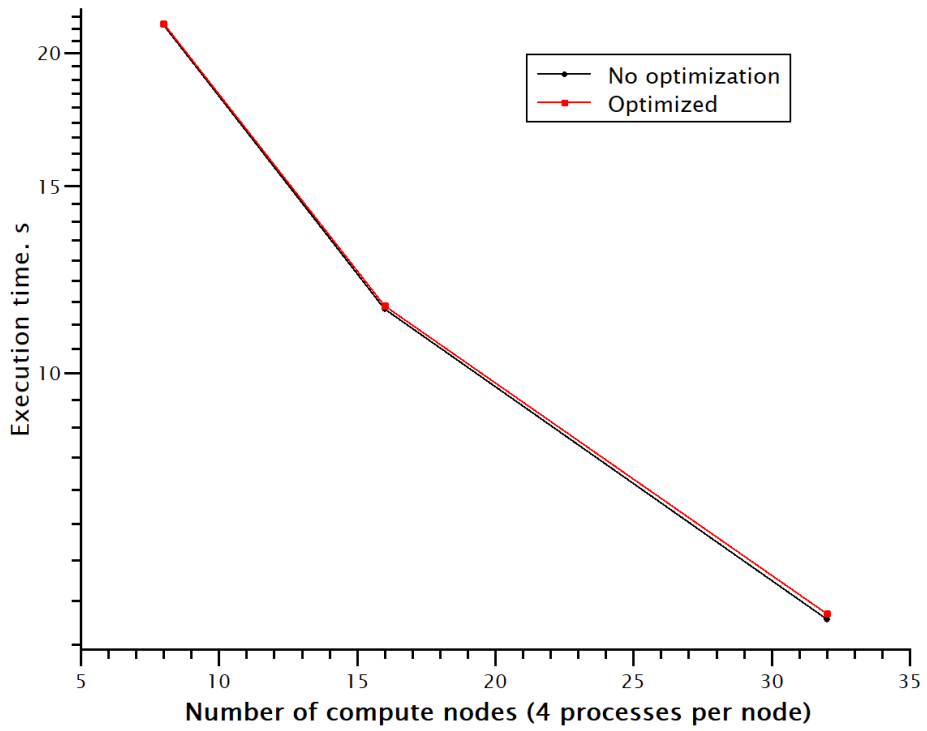
- Intel Xeon CPU E5-2660 (8 cores), 2.20 GHz
- 64 GB ОЗУ
- сетевой адаптер сети Ангара на базе СБИС EC8430

Режимы тестирования работы MPI-программ

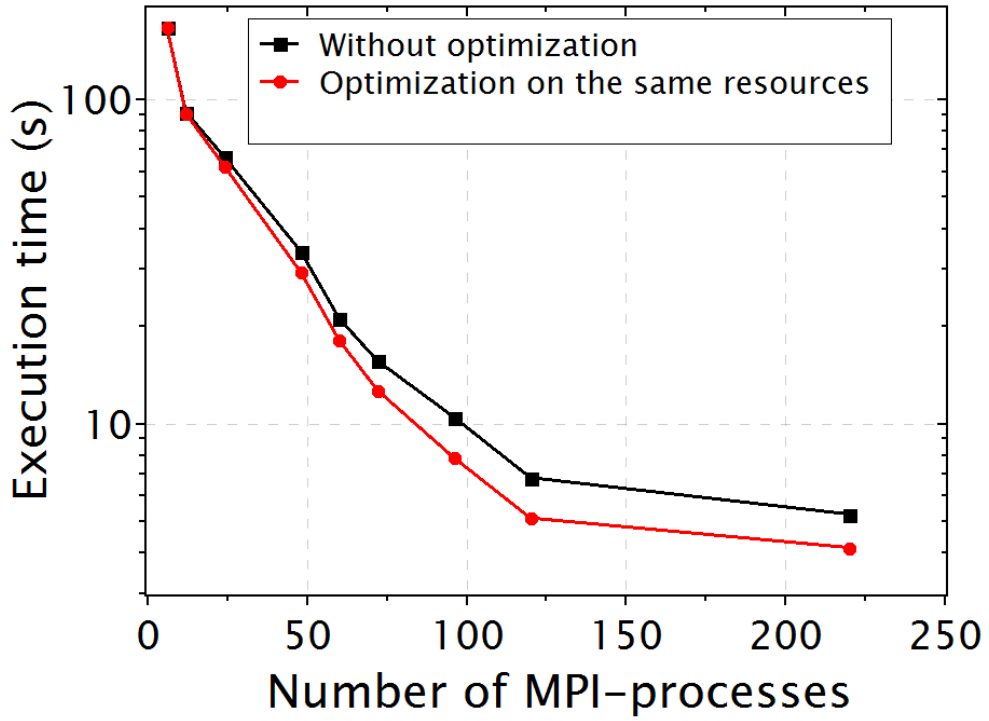
А) запуск MPI-программ осуществлялся без использования оптимизации отображения;

Б) запуск программ с использованием оптимизации отображения;

Экспериментальные результаты моделирования работы оптимизации на тесте NPB LU

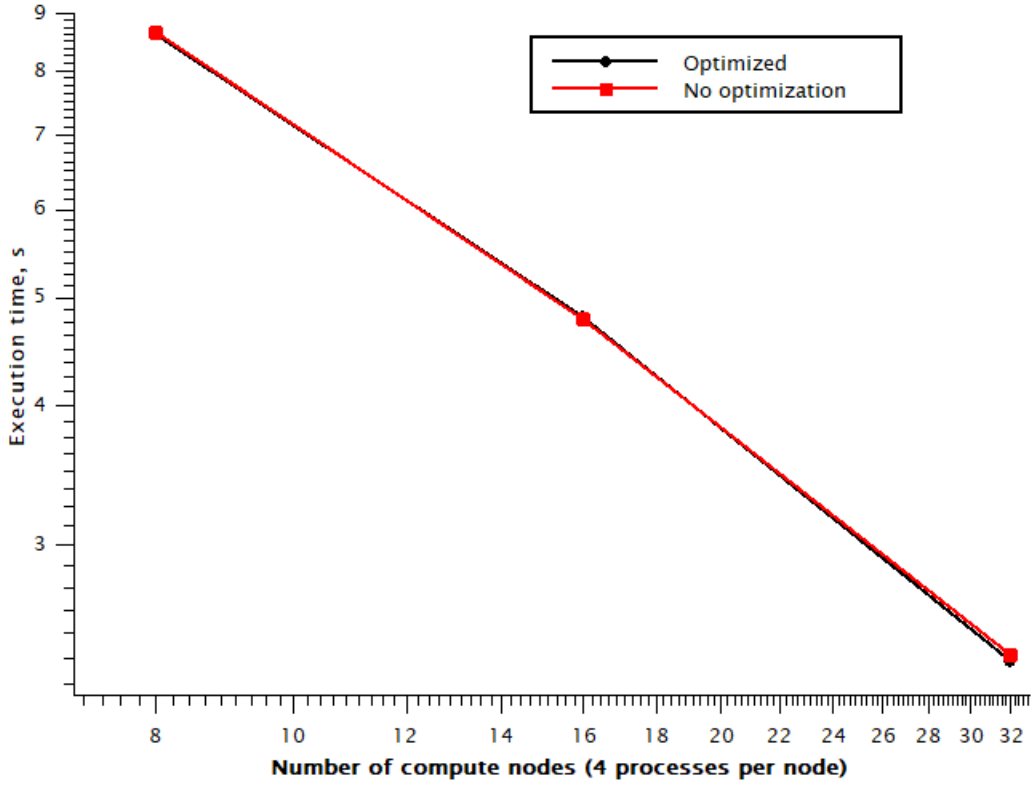


Десмос (1 сокет)

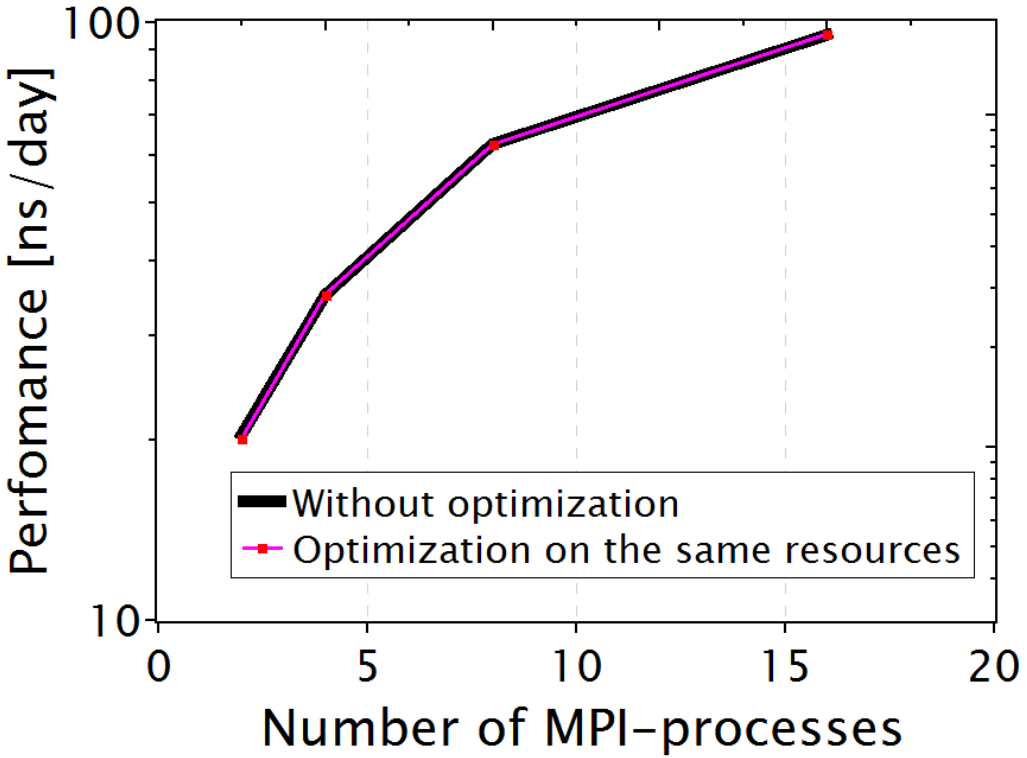


Ангара К-1 (2 сокета)

Экспериментальные результаты моделирования работы оптимизации на тестах NPB FT и GROMACS



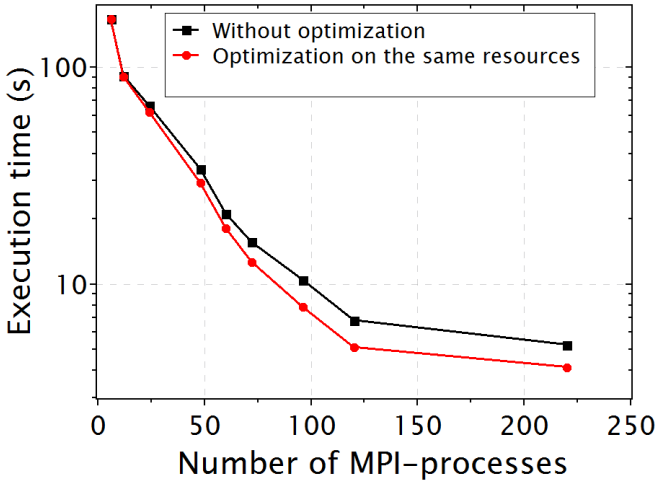
NPB Fourier Transform

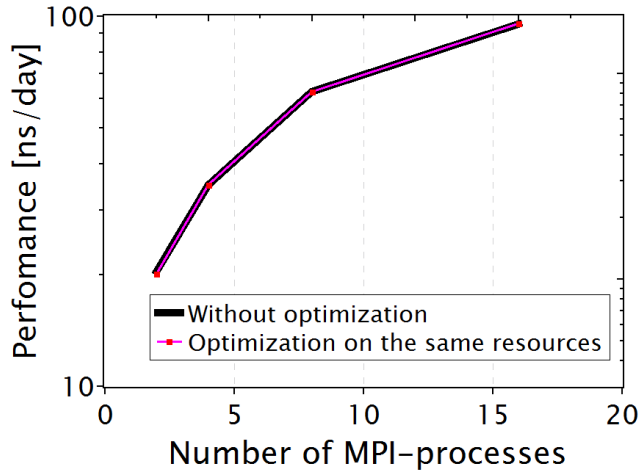


GROMACS: membrane channel protein embedded in a lipid bilayer surrounded by water (MEM, ~100k atoms)

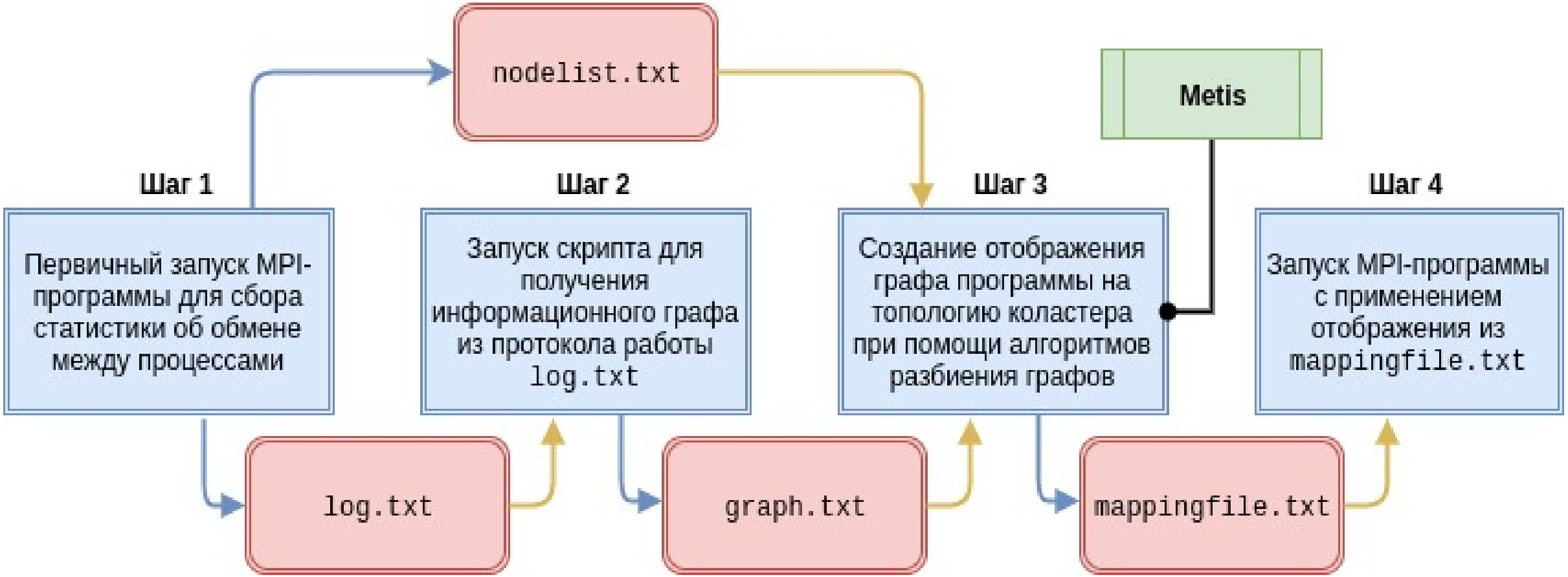
Результаты и выводы

- Выработан общий подход для оптимизации запуска MPI-программ на кластерах с сетью Ангара
- При малом количестве ядер и вычислительных узлов эффект от оптимизации незначителен
- Использование первого сокета на многосокетных узлах даёт прирост до 25%
- Использование ремаппинга теряет смысл для задач с симметричным коммуникационным паттерном





Функциональная схема работы библиотеки оптимизации запуска MPI-программ



Статьи, посвященные оптимизации запуска MPI-программ

- 1) Subramoni H. et al. Design of a scalable InfiniBand topology service to enable network-topology-aware placement of processes. // High Performance Computing, Networking, Storage and Analysis (SC), 2012 International Conference for. – IEEE, 2012. – С. 1-12.
- 2) Hoefler T., Snir M. Generic topology mapping strategies for large-scale parallel architectures. // Proceedings of the international conference on Supercomputing. – ACM, 2011. – С. 75-84.
- 3) Yu H. et al. Topology mapping for Blue Gene/L supercomputer. // Proceedings of the 2006 ACM/IEEE conference on Supercomputing. – ACM, 2006. – С. 116.
- 4) Пазников А. А., Курносов М. Г., Куприянов М.С. Многоуровневые алгоритмы отображения параллельных MPI-программ на вычислительные кластеры. // Проблемы информатики. 2015. Т1, С. 4-17

Разбиение графа с помощью METIS

- Задача разбиения графа **NP-полная**.
- В METIS реализован алгоритм* рекурсивной бисекции, его сложность: **T = O(|E|log_2(z))**, E - кол-во рёбер, z - число подмножеств.
- Требуется **N+1** разбиений, N - кол-во узлов с ненулевым количеством сокетов.

Karypis, G. and Kumar, V. A // SIAM Journal on Scientific Computing, 1999