НИЦЭВТ

Ростех
Объединенная приборостроительная корпорация

Реализация задачи поиска минимального остовного дерева для суперкомпьютеров с сетью Ангара

Александр Семёнов, Артём Мазеев

DISLab/АО «НИЦЭВТ»

GraphHPC-2016

03.03.2016

План доклада

- Введение
  - Основные проблемы, возникающие при суперкомпьютерной обработке графов
- Алгоритм GHS решения задачи построения минимального остовного дерева (MST)
- Высокоскоростная сеть Ангара
- Оптимизированная реализация алгоритма GHS
- Сравнение результатов Ангара vs Infiniband
- Планы на будущее

Проблемы анализа больших графов

НИЦЭВТ

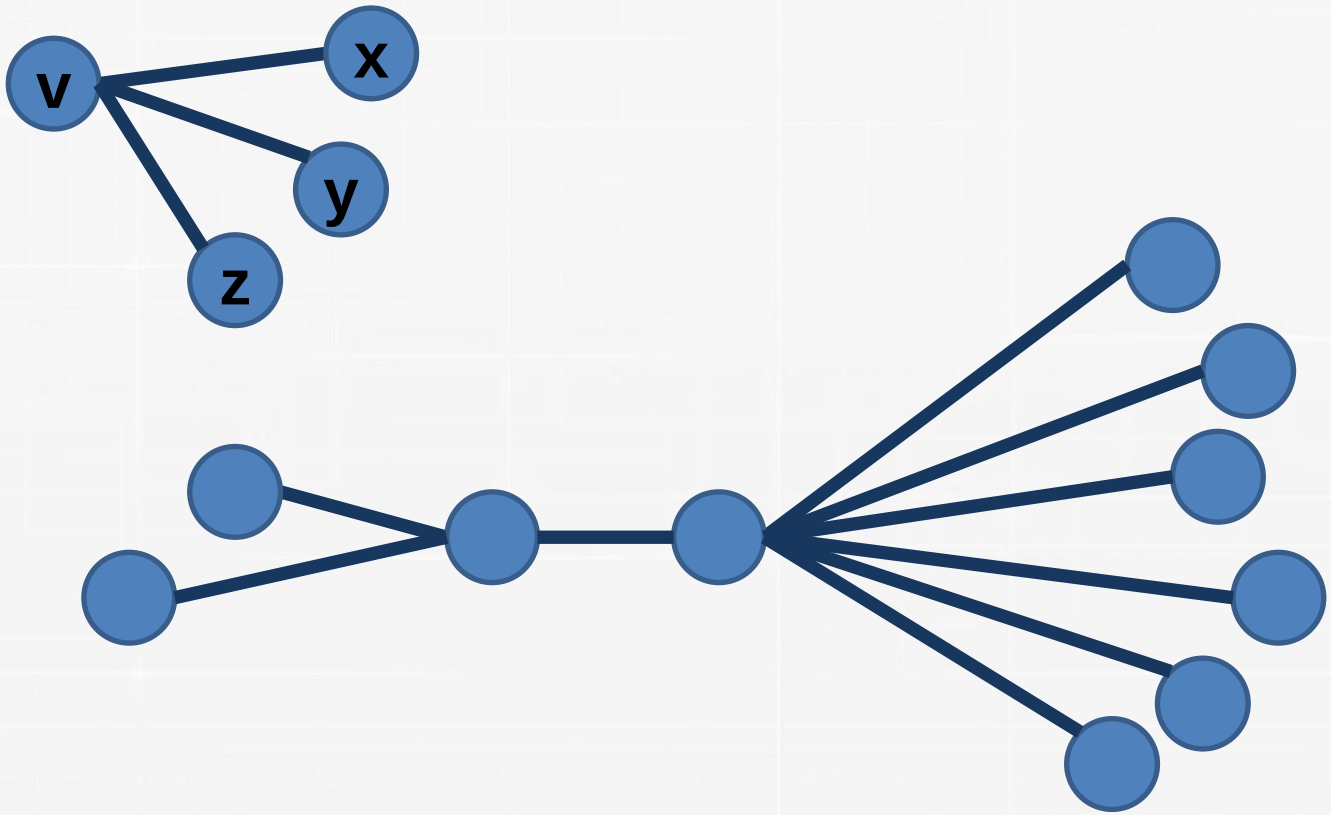
Ростех
Объединенная приборостроительная корпорация

- **Data-driven computations.** Зависимость вычислений от данных (топологии графа). Невозможность применения методов статического распараллеливания вычислений.
- **Unstructured problems.** Работа с нерегулярными, неструктурированными данными, трудность распараллеливания.
- **Poor locality.** Низкая пространственно-временная локализация обращений к памяти.
- **High data access to computation ratio.** Преобладание команд доступа к памяти над командами выполнения арифметических операций.

Проблемы анализа больших графов (1)

Объединенная приборостроительная корпорация
Ростех

- **Data-driven computations.** Зависимость вычислений от данных (топологии графа). Невозможность применения методов статического распараллеливания вычислений.

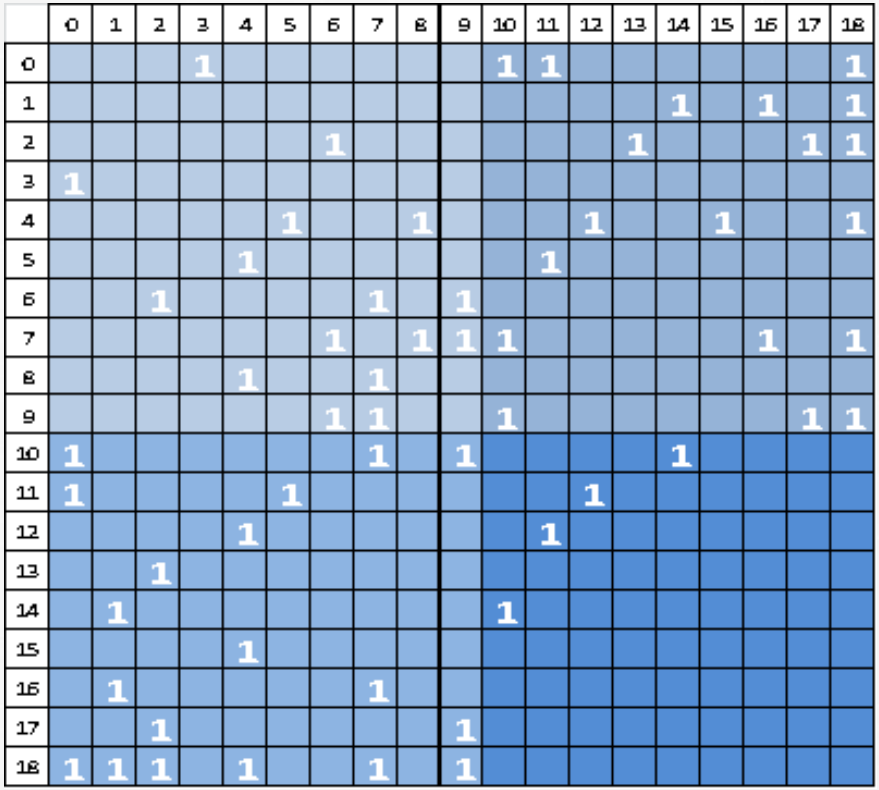


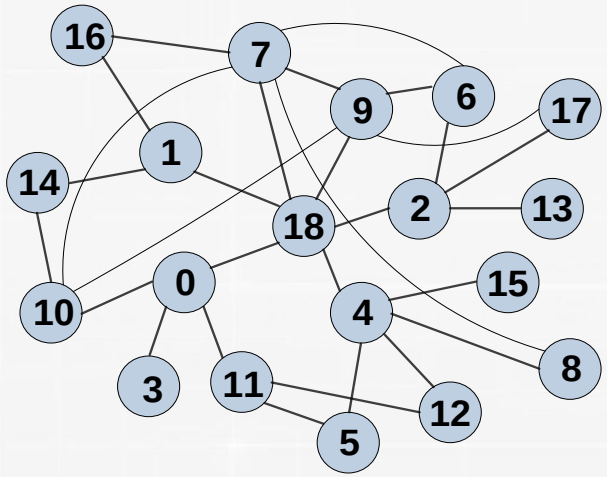
Проблемы анализа больших графов (2)

НИЦЭВТ

Объединенная приборостроительная корпорация
Ростех

- **Unstructured problems.** Работа с нерегулярными, неструктурированными данными, трудность распараллеливания.

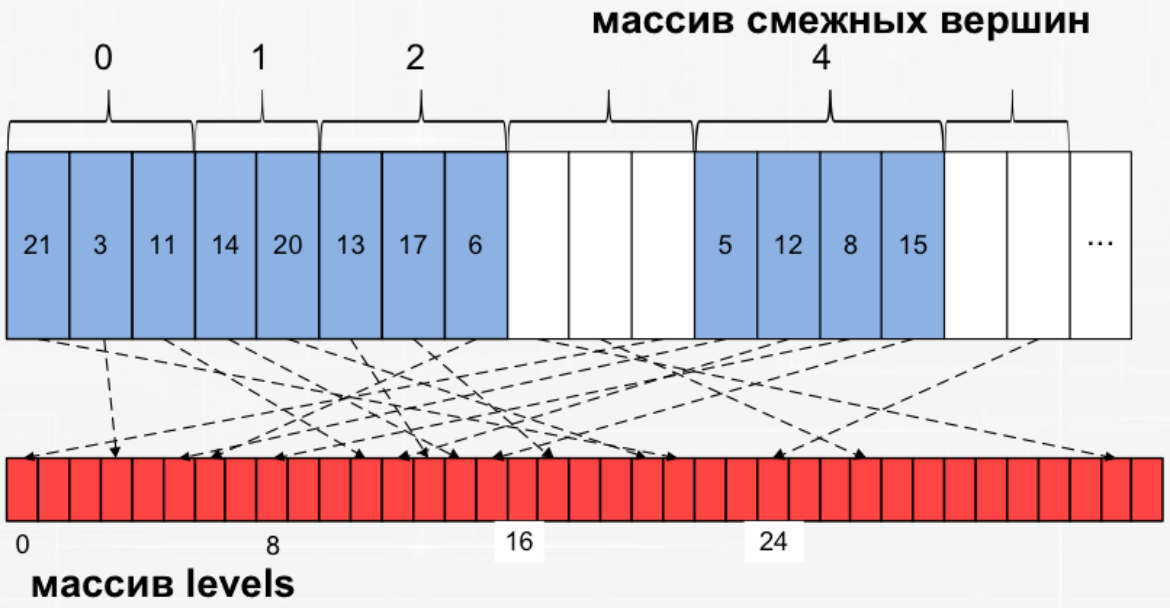


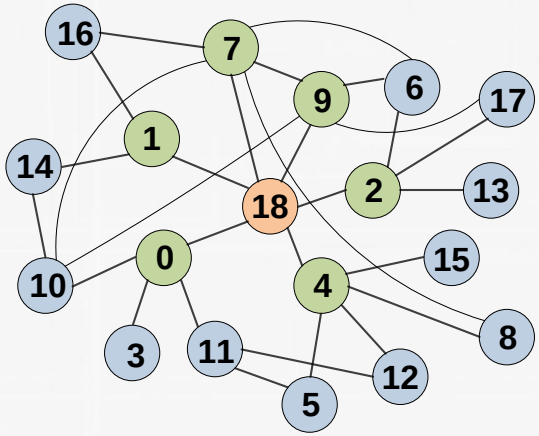


Проблемы анализа больших графов (3)

Объединенная приборостроительная корпорация
Ростех

- **Poor locality.** Низкая пространственно-временная локализация обращений к памяти.





Проблемы анализа больших графов (4)

НИЦЭВТ

Ростех — Объединенная приборостроительная корпорация

- **High data access to computation ratio.** Преобладание команд доступа к памяти над командами выполнения арифметических операций.

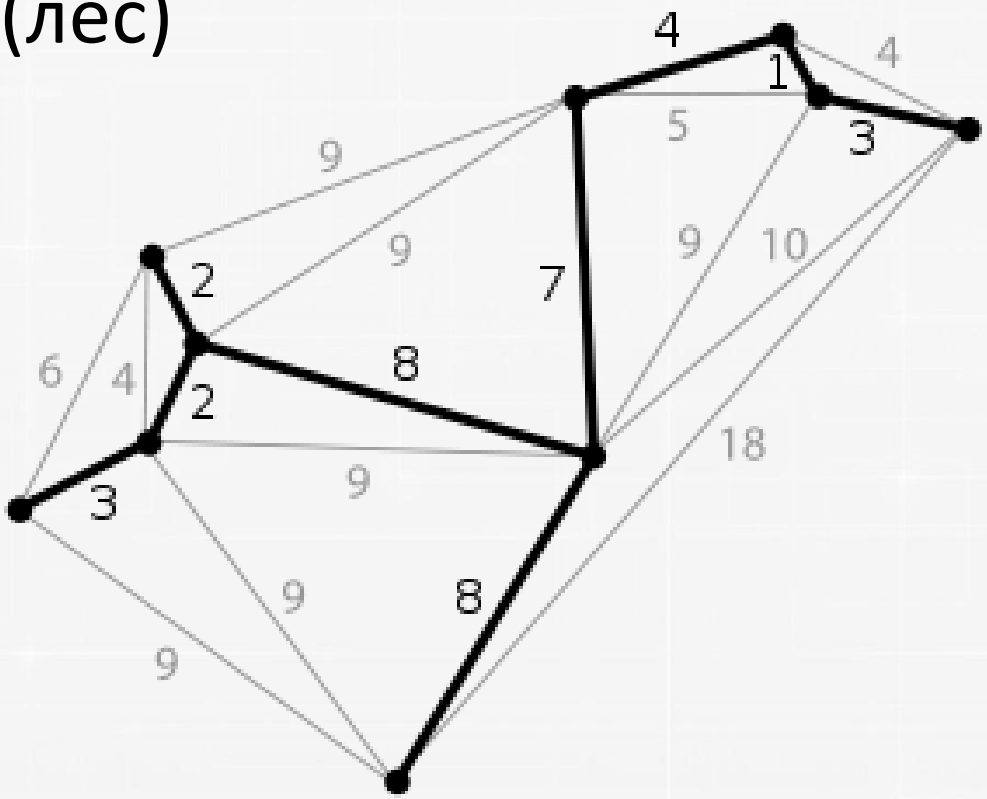
Intel Nehalem 5570, 2.933 МГц

Параметр	Задержка, нс (такты)	ПС, ГБ/с
Регистр	(1)	–
Кэш L1	1.3 (4)	46
Кэш L2	3.4 (10)	31
Кэш L3	13 (38)	24
Память своего сокета	65 (190)	14
Память чужого сокета	106 (310)	9
Сеть Ангара	MPI – 1000, SHMEM – 600	

Задача построения минимального остовного дерева (MST)

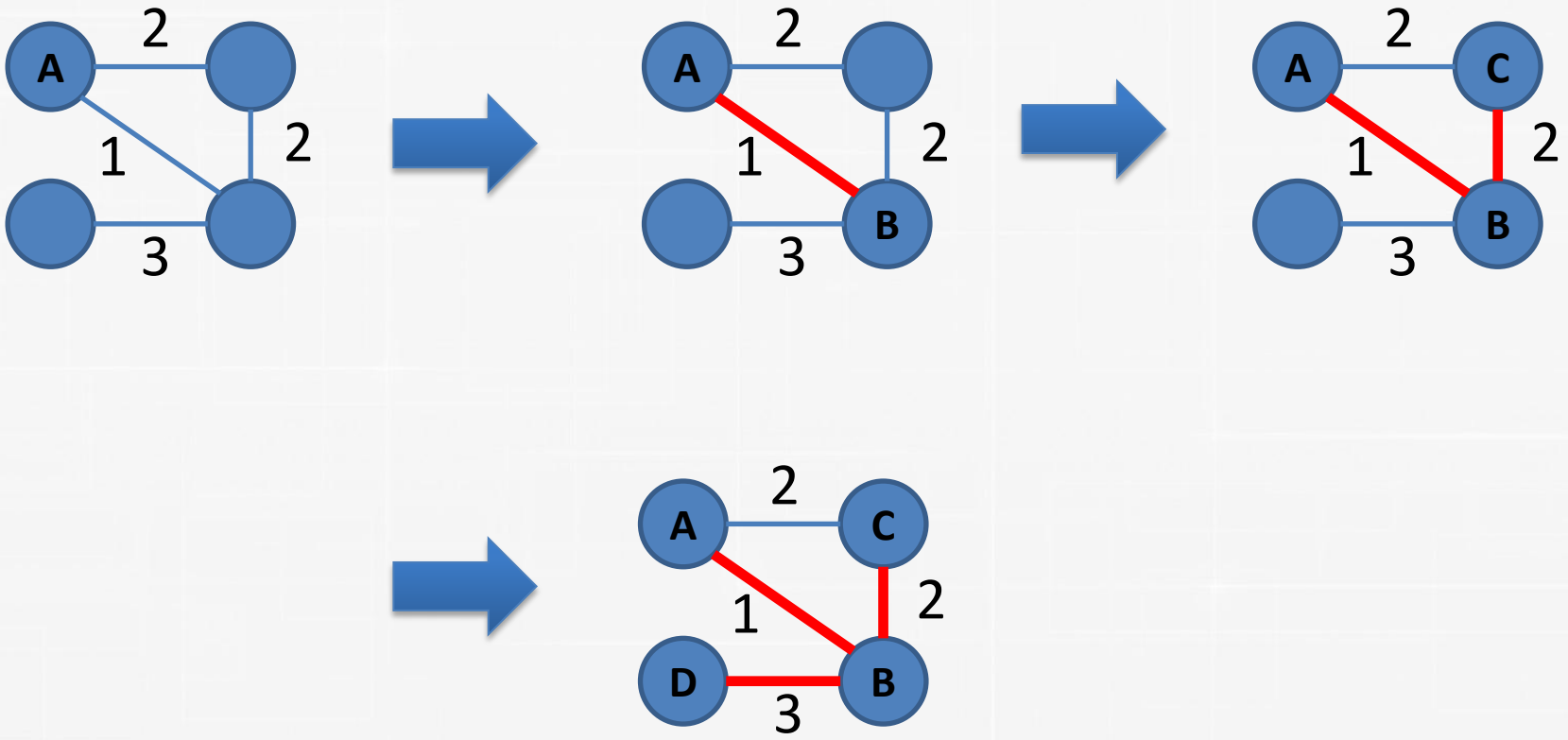
Постановка задачи

- G = (V, E) - неориентированный граф с весами. |V| = n, |E| = m
- Требуется найти минимальное остовное дерево (лес)



Алгоритмы решения задачи

- Жадный последовательный алгоритм Прима



Алгоритмы решения задачи

- Параллельный алгоритм Борувки – общая память
- Параллельный алгоритм GHS – распределенная память

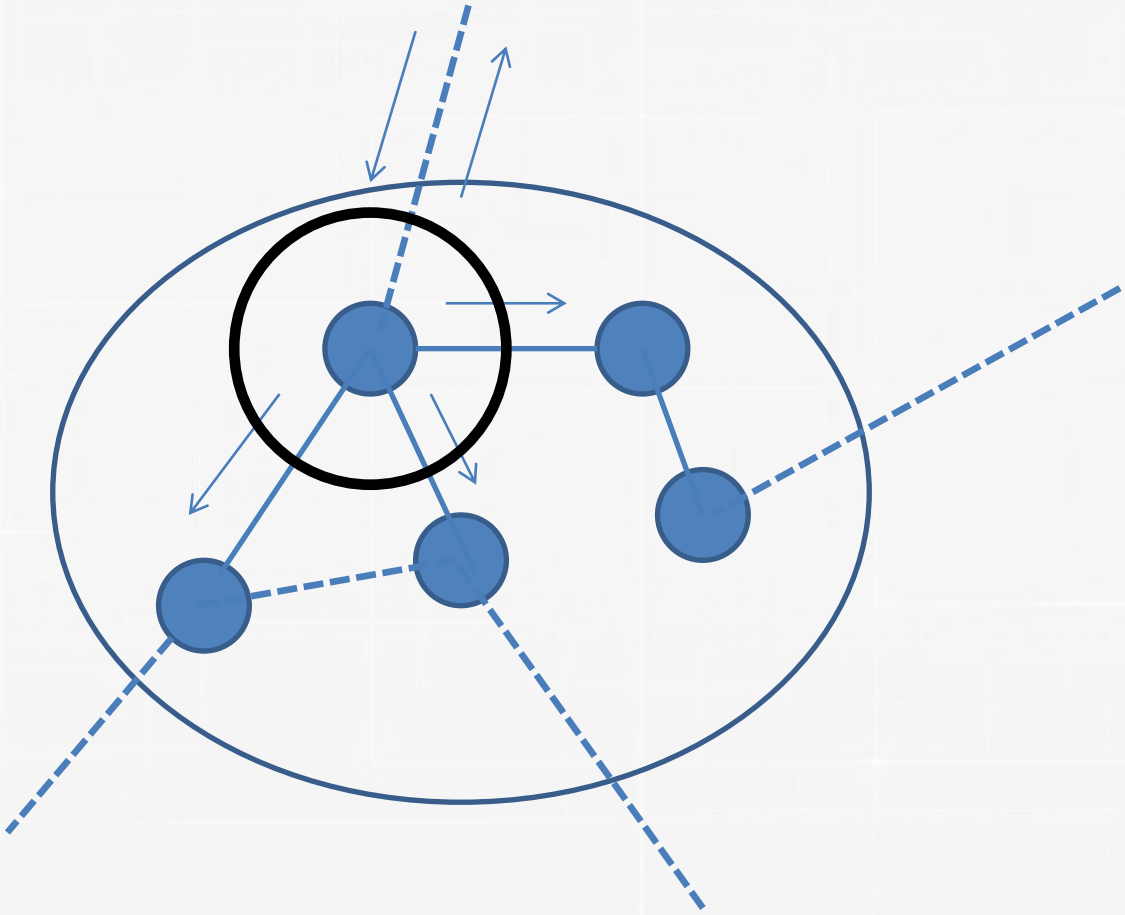
Асинхронный алгоритм GHS (R. Gallager, P. Humblet, P. Spira, 1983)

- Изначально каждая вершина – фрагмент
- Фрагмент ищет инцидентное ребро минимального веса
- Фрагменты постепенно объединяются
- Фрагмент представляет из себя дерево, наверху – ядровое ребро

Объединение фрагментов

- У каждого фрагмента есть переменная – уровень L
- Сначала уровень каждого фрагмента L = 0
- Два фрагмента с одинаковым уровнем L могут объединится во фрагмент с уровнем L + 1
- Фрагмент не может присоединиться к фрагменту с меньшим уровнем

Вершины в алгоритме GHS

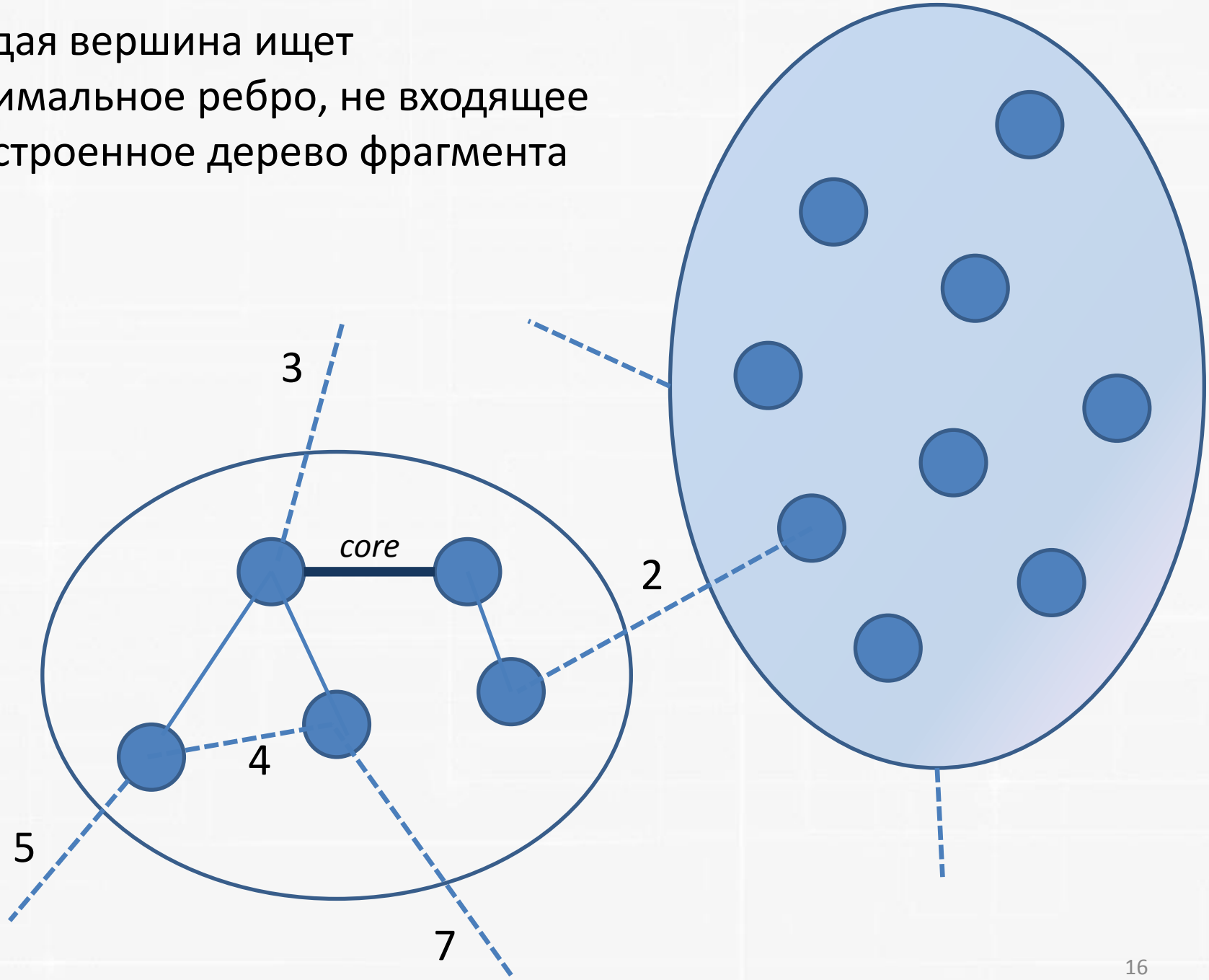


Сообщения между вершинами

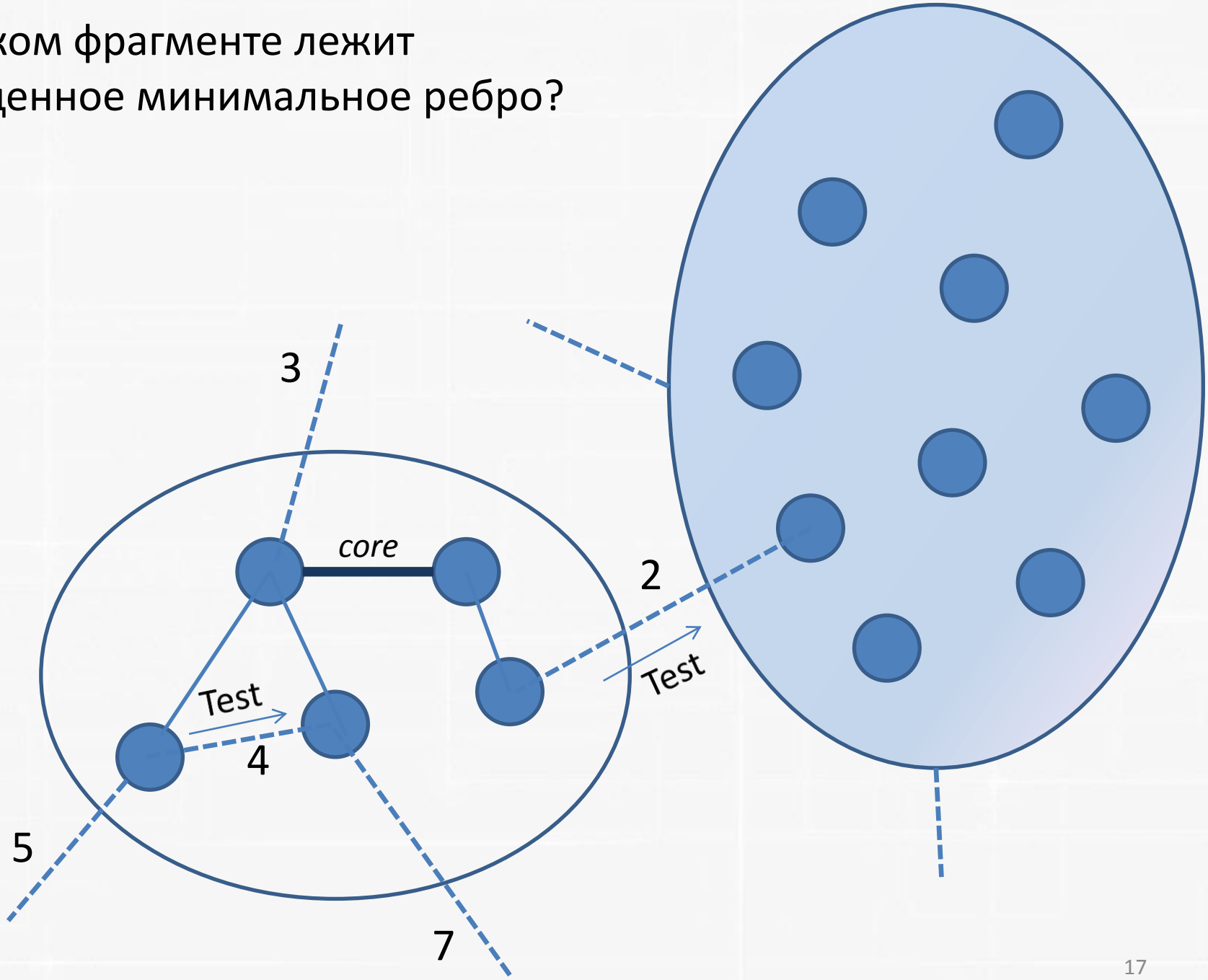
Всего 7 типов сообщений:

- **Test (различные фрагменты?)**
  - **Accept (различные)**
  - **Reject (один и тот же фрагмент)**
- **Report (отчет о наименьшем ребре)**
- **Change core (переход от корня к наилучшему ребру внутри фрагмента)**
- **Connect (запрос об объединении)**
- **Initiate (обновление данных и начало нового поиска наилучших ребер)**

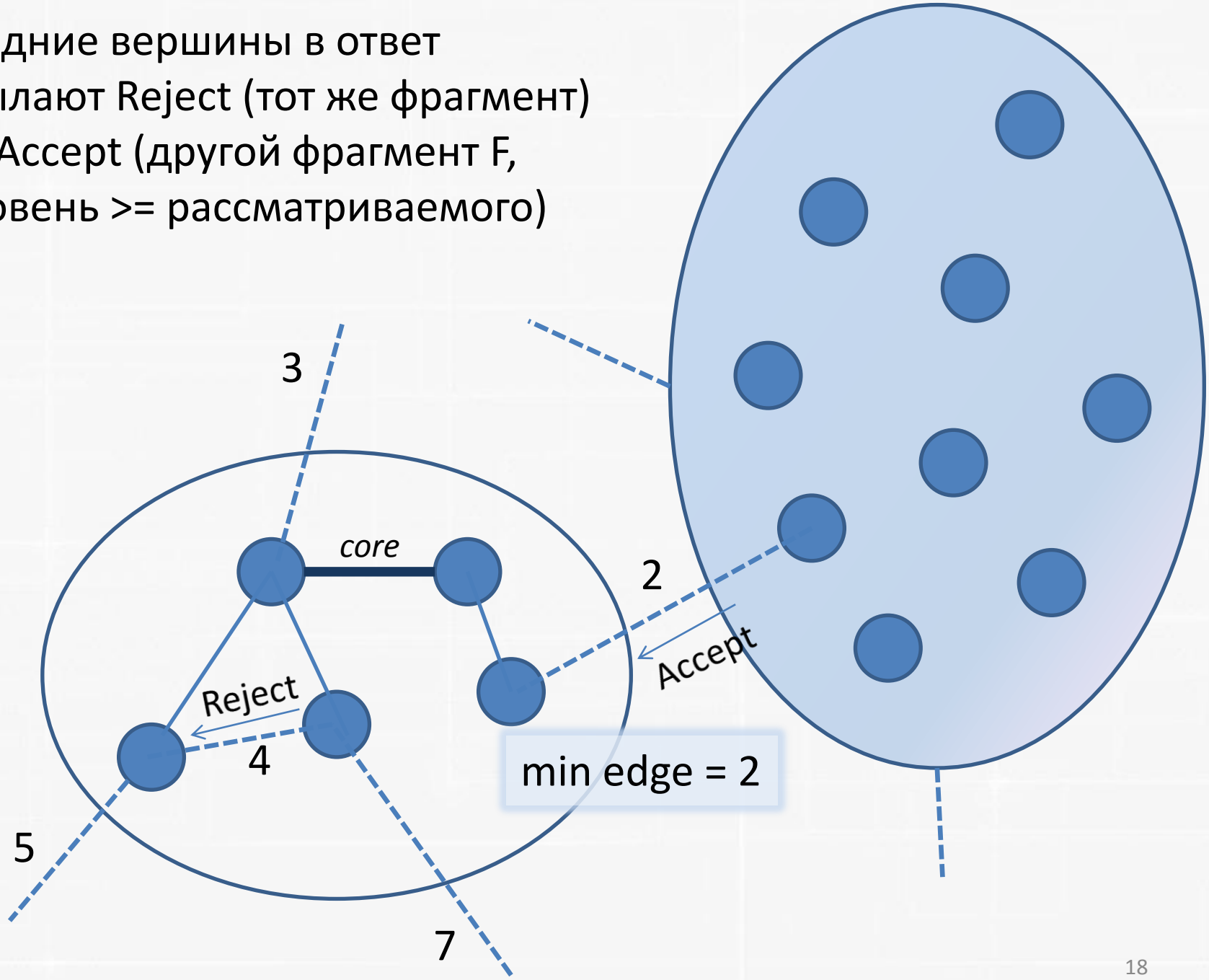
Каждая вершина ищет минимальное ребро, не входящее в построенное дерево фрагмента



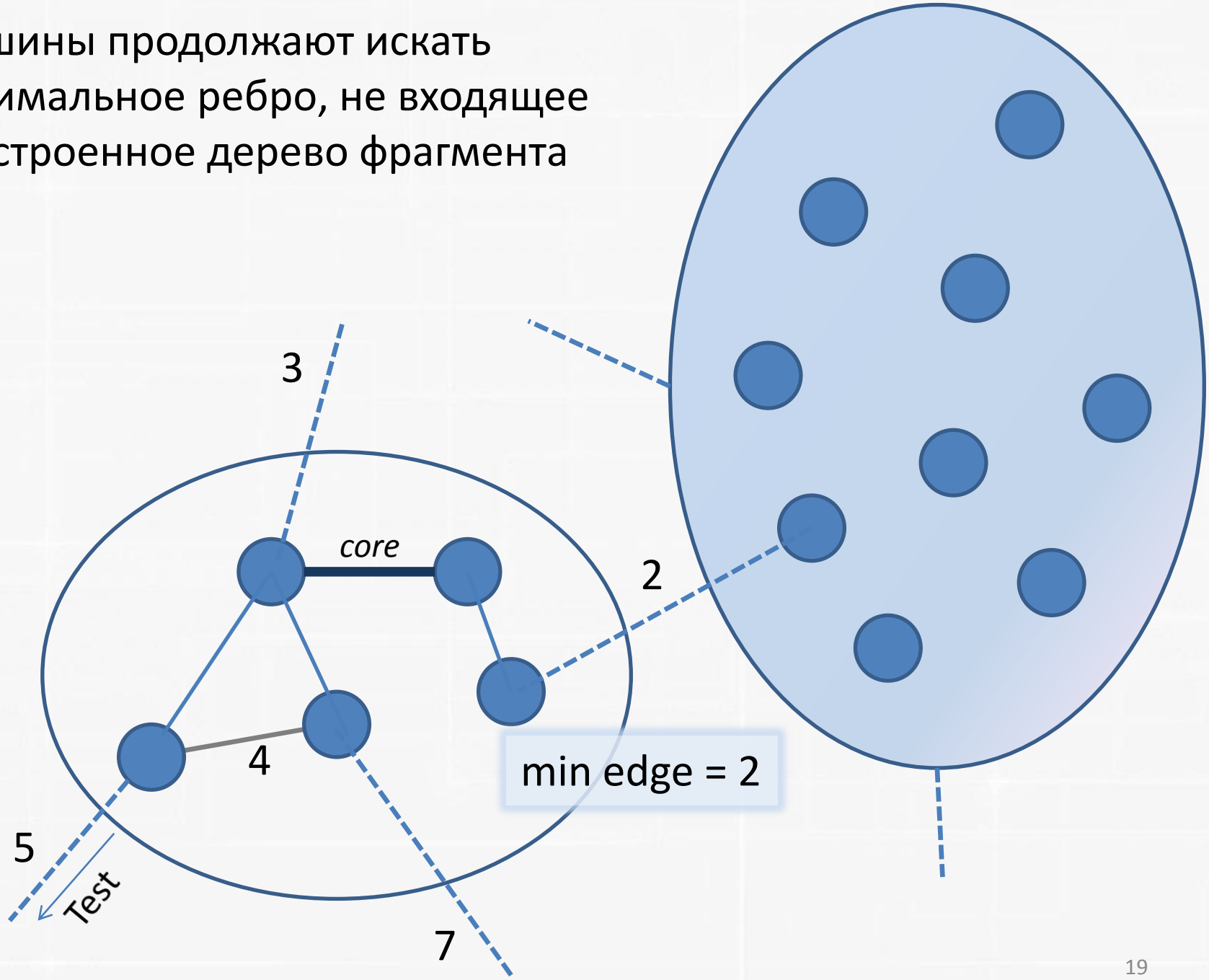
В каком фрагменте лежит найденное минимальное ребро?



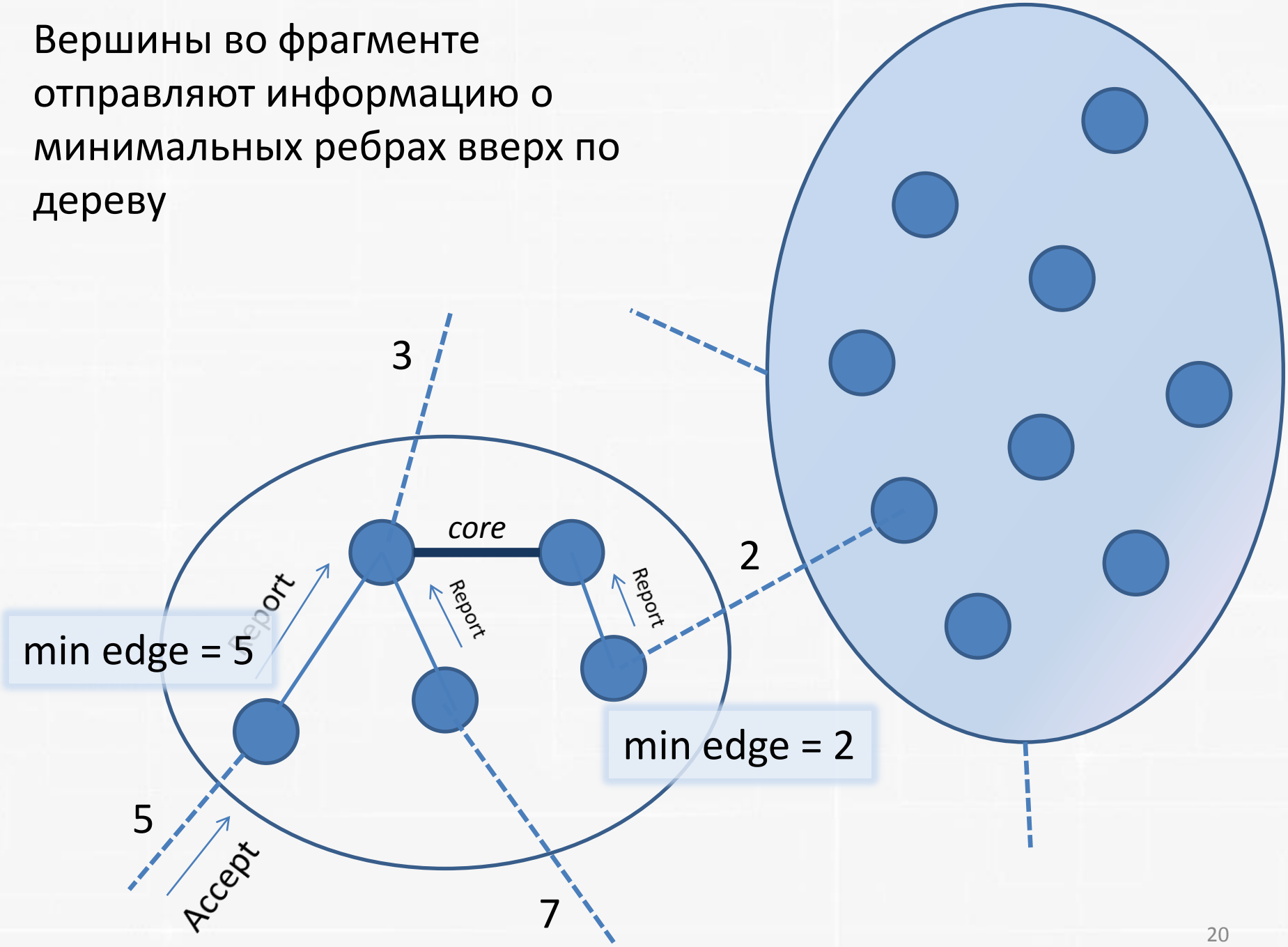
Соседние вершины в ответ посылают Reject (тот же фрагмент) или Accept (другой фрагмент F, F.уровень >= рассматриваемого)



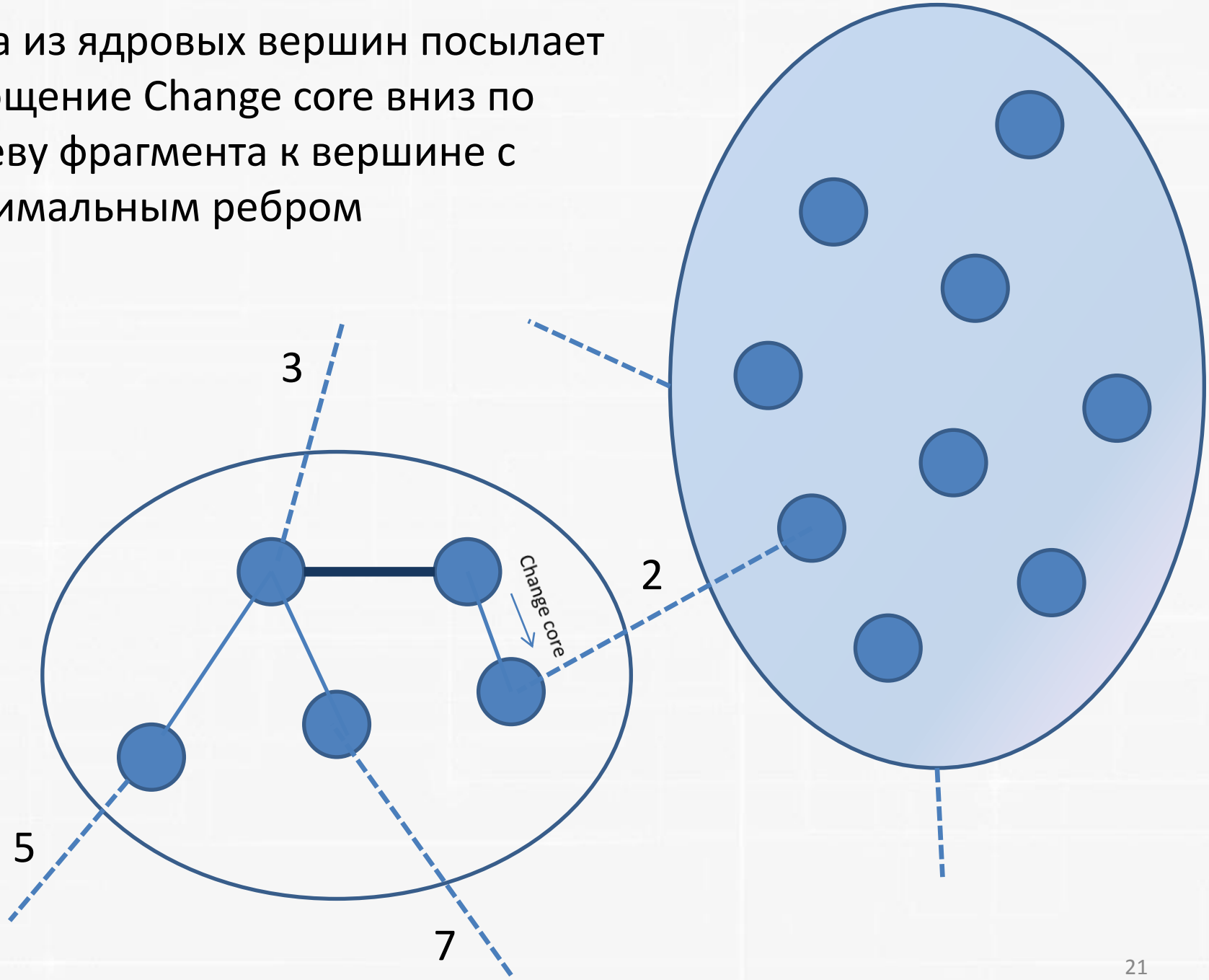
Вершины продолжают искать минимальное ребро, не входящее в построенное дерево фрагмента



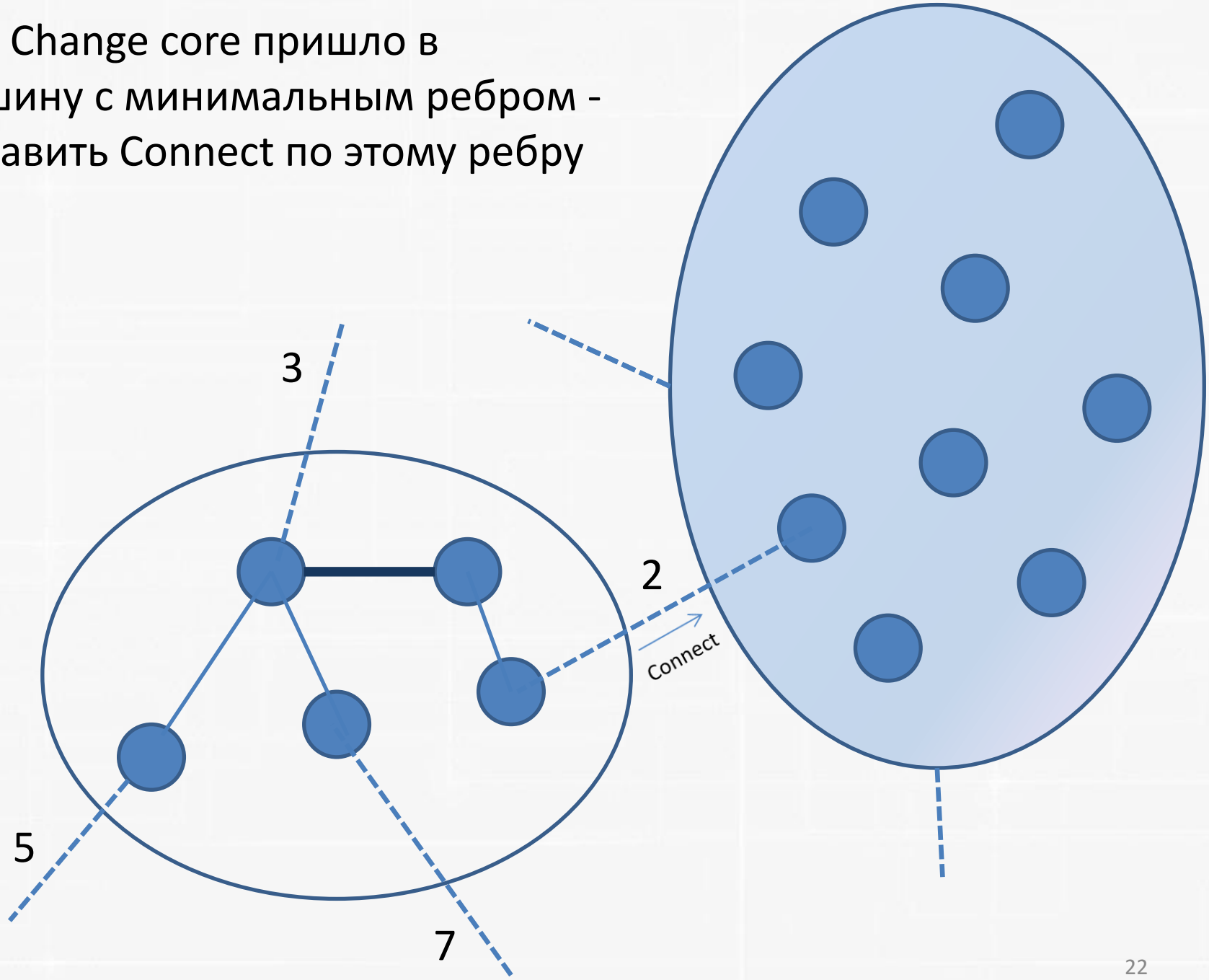
Вершины во фрагменте отправляют информацию о минимальных ребрах вверх по дереву



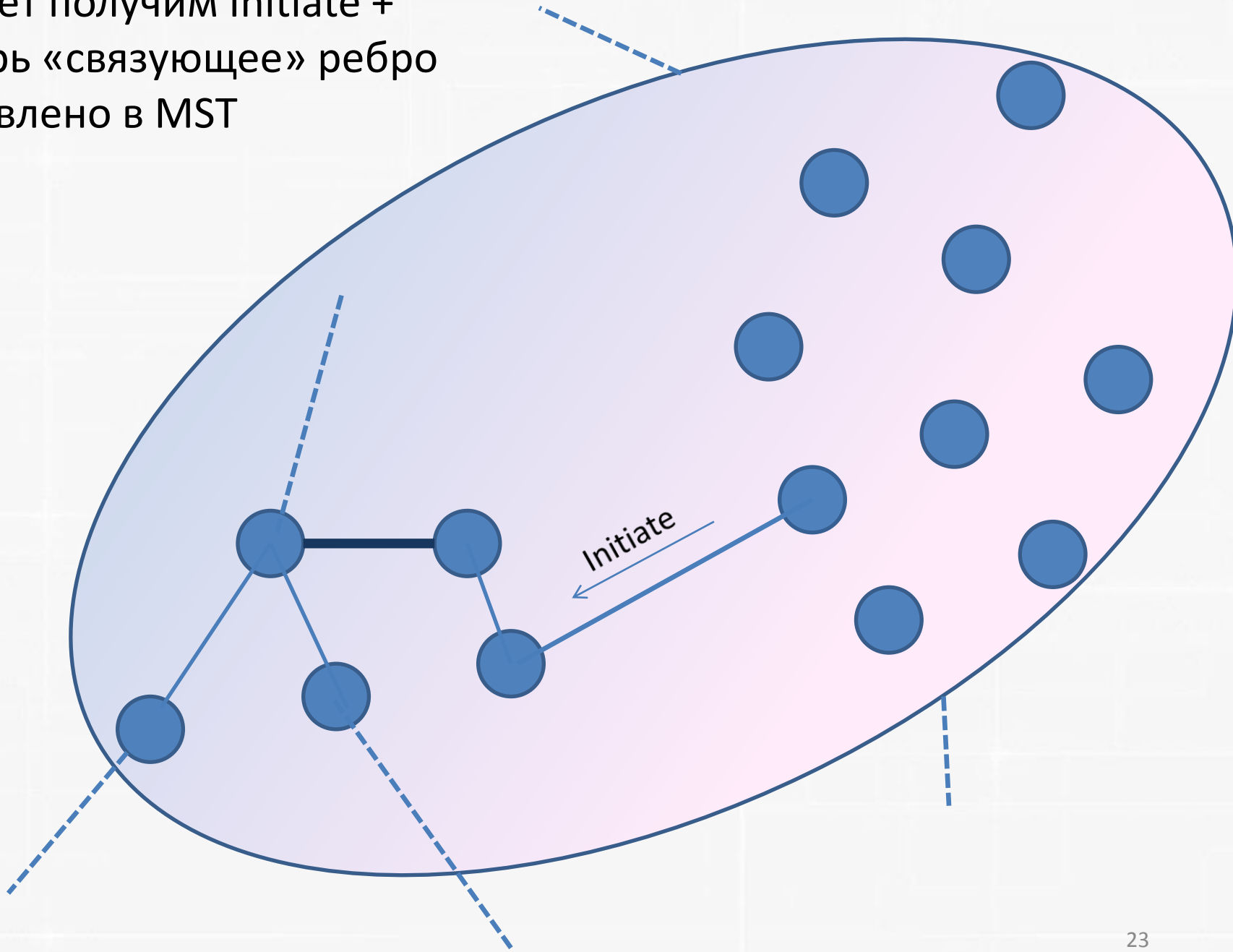
Одна из ядровых вершин посылает сообщение Change core вниз по дереву фрагмента к вершине с минимальным ребром



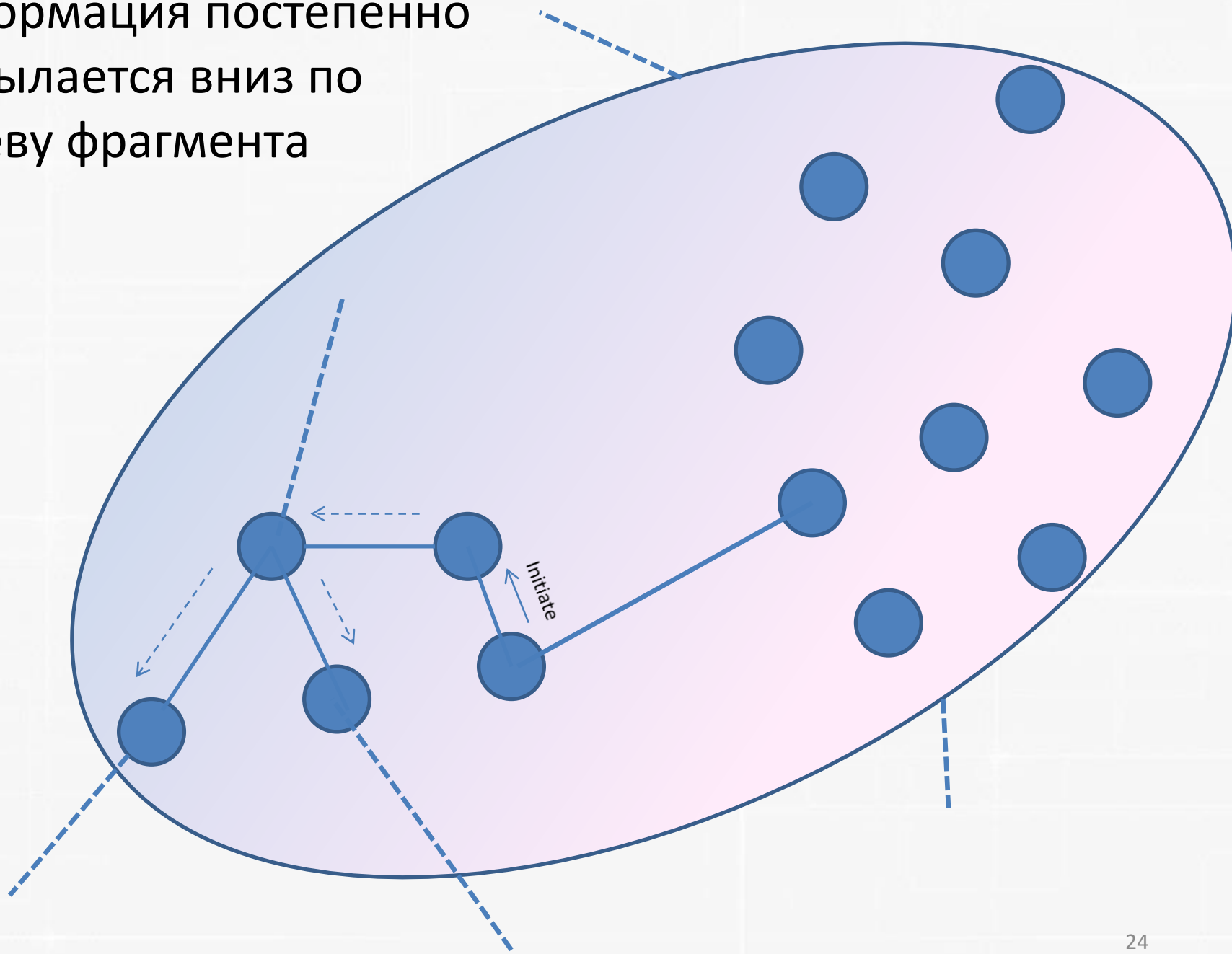
Если Change core пришло в вершину с минимальным ребром - отправить Connect по этому ребру



В ответ получим Initiate + теперь «связующее» ребро добавлено в MST



Информация постепенно рассылается вниз по дереву фрагмента



Обработка сообщений

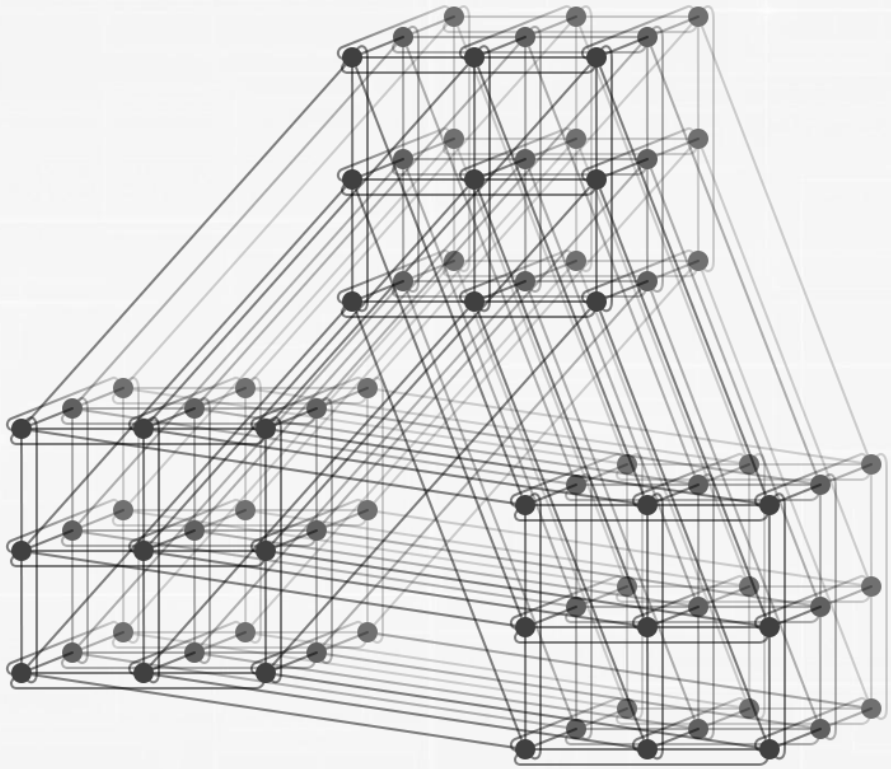
- Каждая вершина: принимает сообщения, кладет в очередь, обрабатывает сообщения из очереди
- При обработке сообщения вершина может посылать сообщения соседним вершинам
- Некоторые сообщения (Test, Connect, Report) нужно откладывать, они будут обработаны позднее
- Сообщения по одному ребру не должны обгонять друг друга

Сложность алгоритма GHS

- Временная сложность – O (n logn)
- Количество сообщений – O (m + n logn)

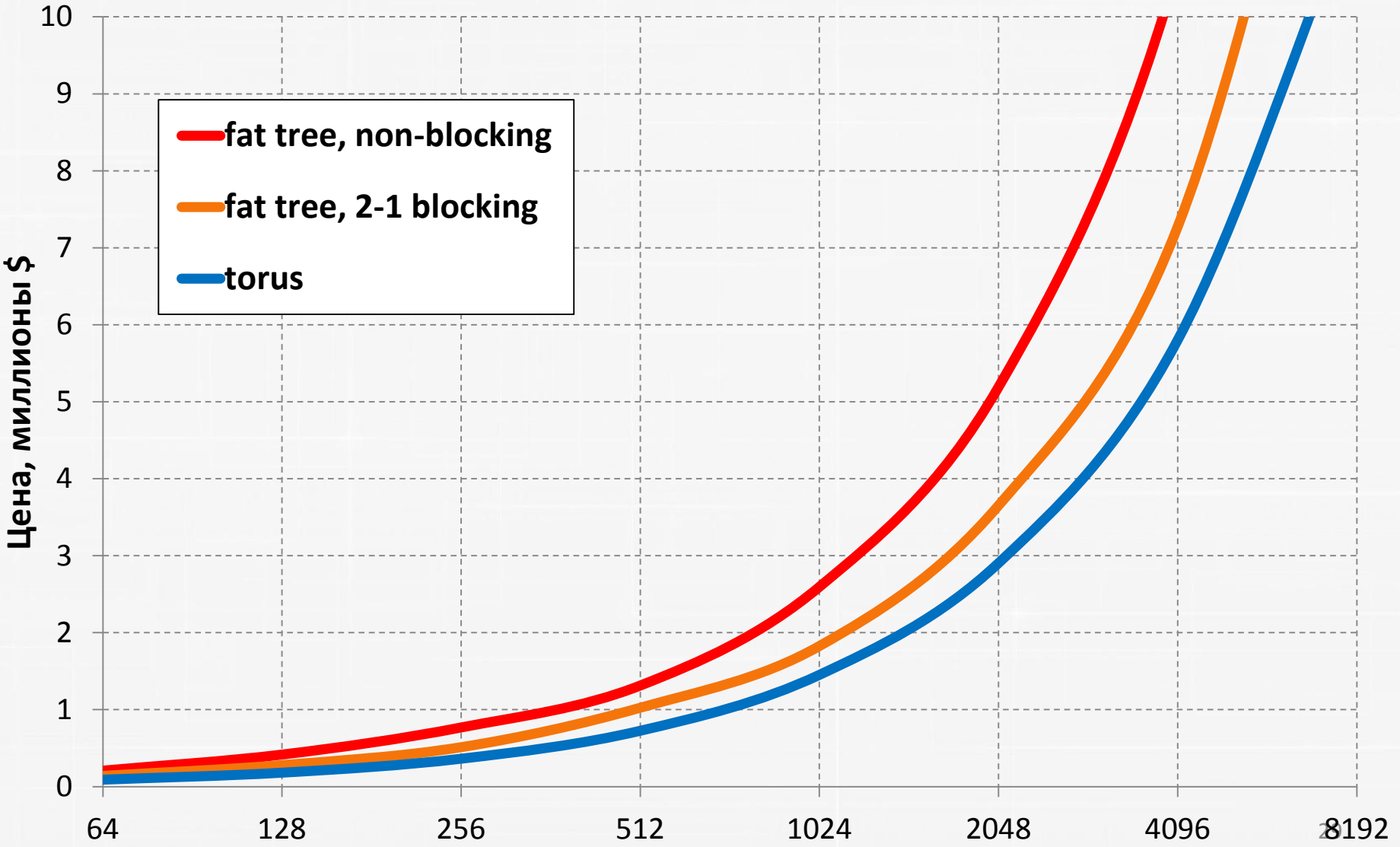
Высокоскоростная сеть Ангара

Высокоскоростная сеть Ангара



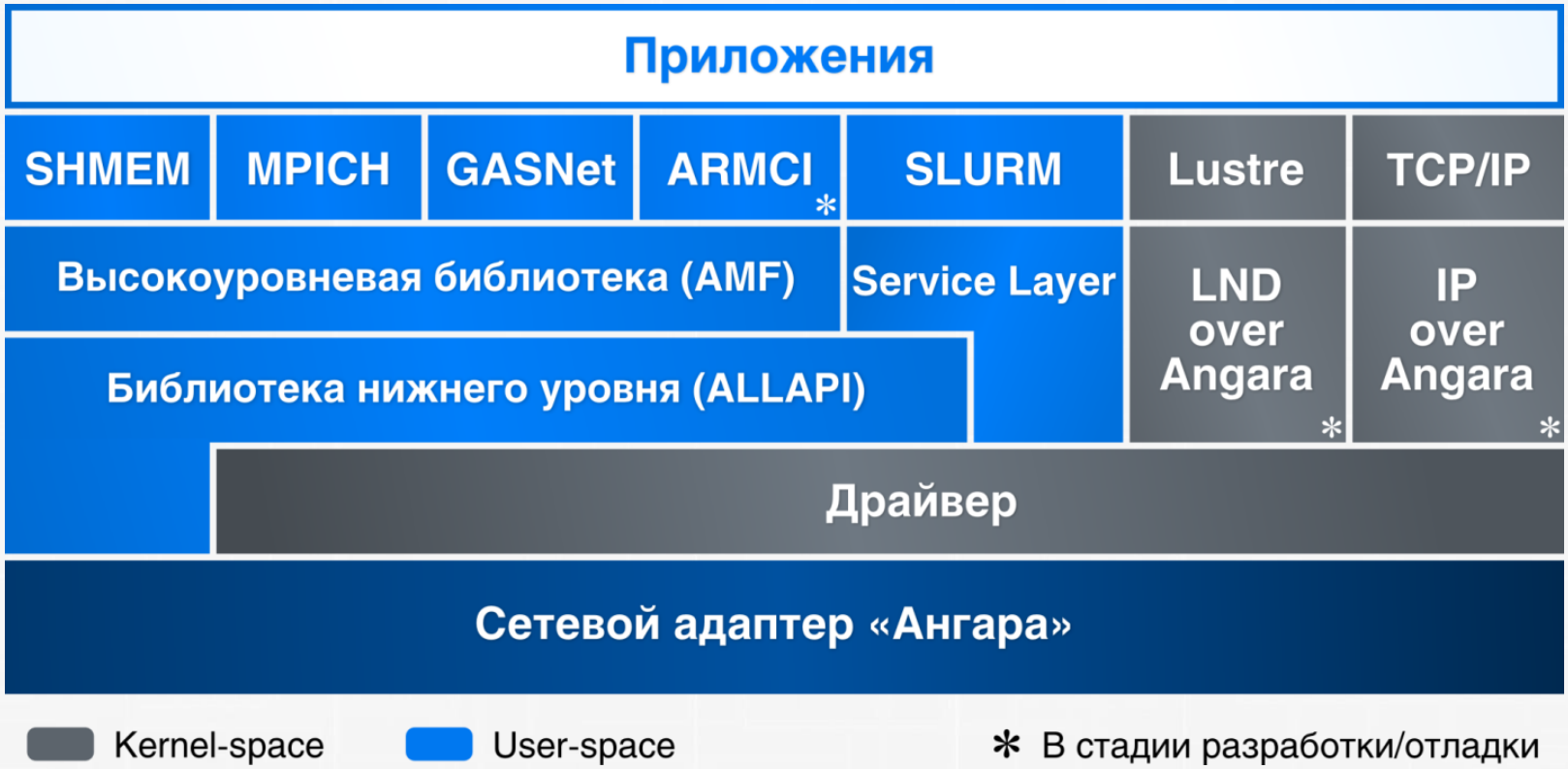
- **Топология «4D-тор»**
- **Односторонние коммуникации:**
  - put (запись в удалённую память)
  - get (чтение из удалённой памяти)
  - атомарные операции add и xor
- **Поддержка многоядерности**
- **Адаптивная передача пакетов**
- **Механизмы синхронизации**
- **Кристалл маршрутизатора выпущен по технологии 65 нм**

Порядок цен на сетевое оборудование для разных топологий



Высокоскоростная сеть Ангара
Стек ПО

Объединенная приборостроительная корпорация
Ростех

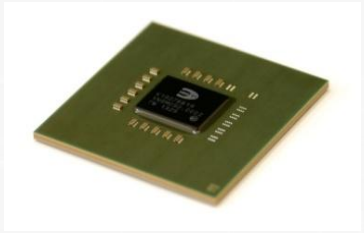


- Поддержка ОС : Astra Linux SE 1.3, ОС «Эльбрус» ,OpenSUSE/SLES 11SP3, CentOS 6.0-6.3, Версия ядра Linux от 2.6.21 до 3.16.0
- Поддержка компиляторов языков Fortran 77/90/95 (GNU, Intel), C/C++ (GNU, Intel), UPC, Co-Array Fortran.

Сетевое оборудование Ангара и сервисы

НИЦЭВТ

Ростех — Объединенная приборостроительная корпорация



Чип 1892ВМ8Я

FCBGA 1521
40 × 40 мм
35 Вт

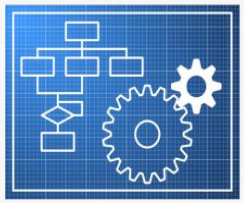
Development kit

Full-height,
full length

Адаптер

Full-height,
full length

Заказная разработка платы адаптера



Доработка ПО адаптера и адаптация библиотек

Адаптация и профилирование прикладного ПО

Коммуникационная сеть Ангара Gen2

НИЦЭВТ

Ростех Объединенная приборостроительная корпорация

Развитие архитектуры и функциональных возможностей:

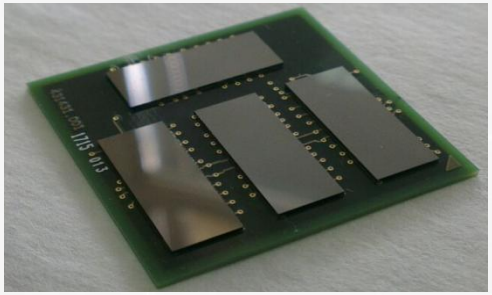
- Поддержка различных топологий (butterfly, dragonfly)
- Улучшение аппаратной поддержки MPI и парадигмы PGAS
- Улучшение аппаратной поддержки коллективных и синхронизационных операций

Интеграция и партнёрство:

- Интеграция с процессором
- Интеграция с ускорителем

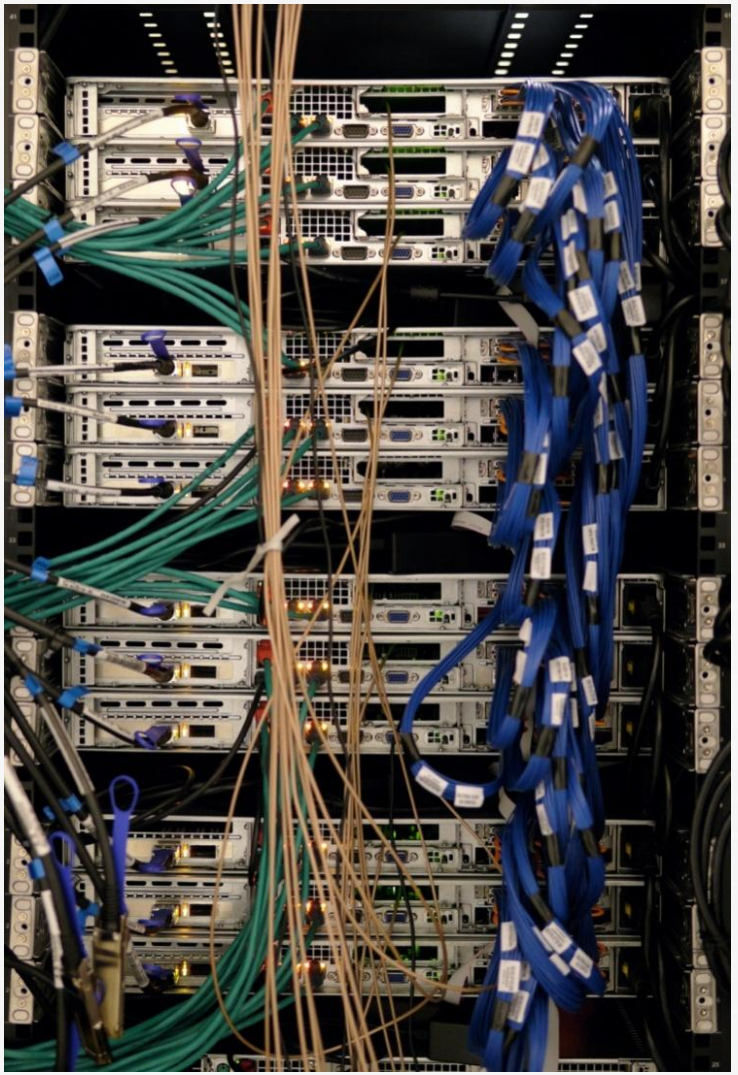
Усовершенствование технологий:

- Линк ~200 Гбит/с
- 4 PCI Express 3.0 x16
- Оптические трансиверы



Кластер «Ангара-К1» с сетью Ангара: 36 узлов

- **24 вычислительных узла**
  - Supermicro SuperServer 5017GR-TF
  - 2 процессора Intel Xeon E5-2630 (LGA2011, 6 ядер, 2.3 ГГц)
  - 64 ГБ
- **12 вычислительных узлов**
  - Supermicro SuperServer 5017GR-TF
  - процессор Intel Xeon E5-2660 (LGA2011, 8 ядер, 2.2 ГГц)
  - 64 ГБ
- **Сеть Ангара**
  - Адаптер EC8430, топология 3D-тор 3x3x4
  - Собственная реализация SHMEM
  - MPI: MPICH 3.0.4, MVAPICH2 1.9
- **Операционная система**
  - SLES 11 SP2, Linux 3.0.13-0.27-default
  - GCC 4.8.3



Оптимизированная реализация алгоритма GHS

Реализация алгоритма GHS

- Вершин графа больше, чем узлов кластера, на каждом узле хранится множество вершин, и связанная с ними информация
- Завершается, когда все «затихнет»
- Примечание: далее local_V – локальное множество вершин

Базовая версия (0)


```
for (v ∈ local_V) {
    wake_up (v); // отправка Connect по ребру min веса
}
while (1) {
    read_msgs (); // прием и запись сообщений в очередь

    /* обработка всей очереди, немедленная отправка
    ответных сообщений */
    process_queue ();

    /* check_finish () – проверка на завершение */
    if (time_to_check_finish) {
        if ( allreduce (queue_size) == 0 &&
             allreduce (send_cnt) == allreduce (recv_cnt) ) exit();
}
```

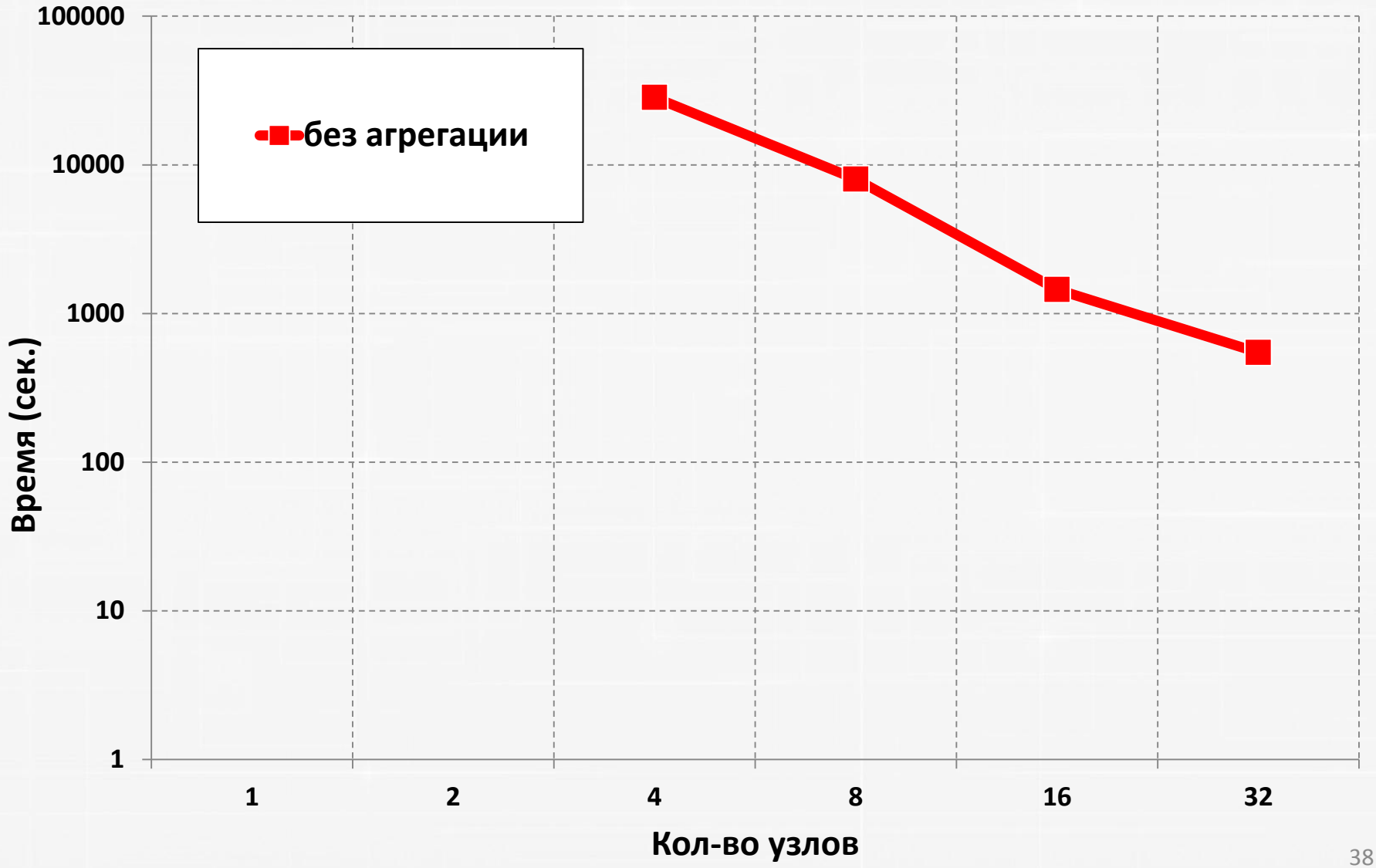
Агрегация сообщений (1)


```
for (v ∈ local_V) {
    wake_up (v); // отправка Connect (запись в буфер)
}

while (1) {

    read_msgs (); // прием и запись сообщений в очередь

    /* обработка очереди. Отправка сообщений при
    переполнении буфера */
    process_queue ();

    // отправка всех сообщений
    if (time_to_send) send_all_bufs ();

    check_finish (); // проверка на завершение

}
```

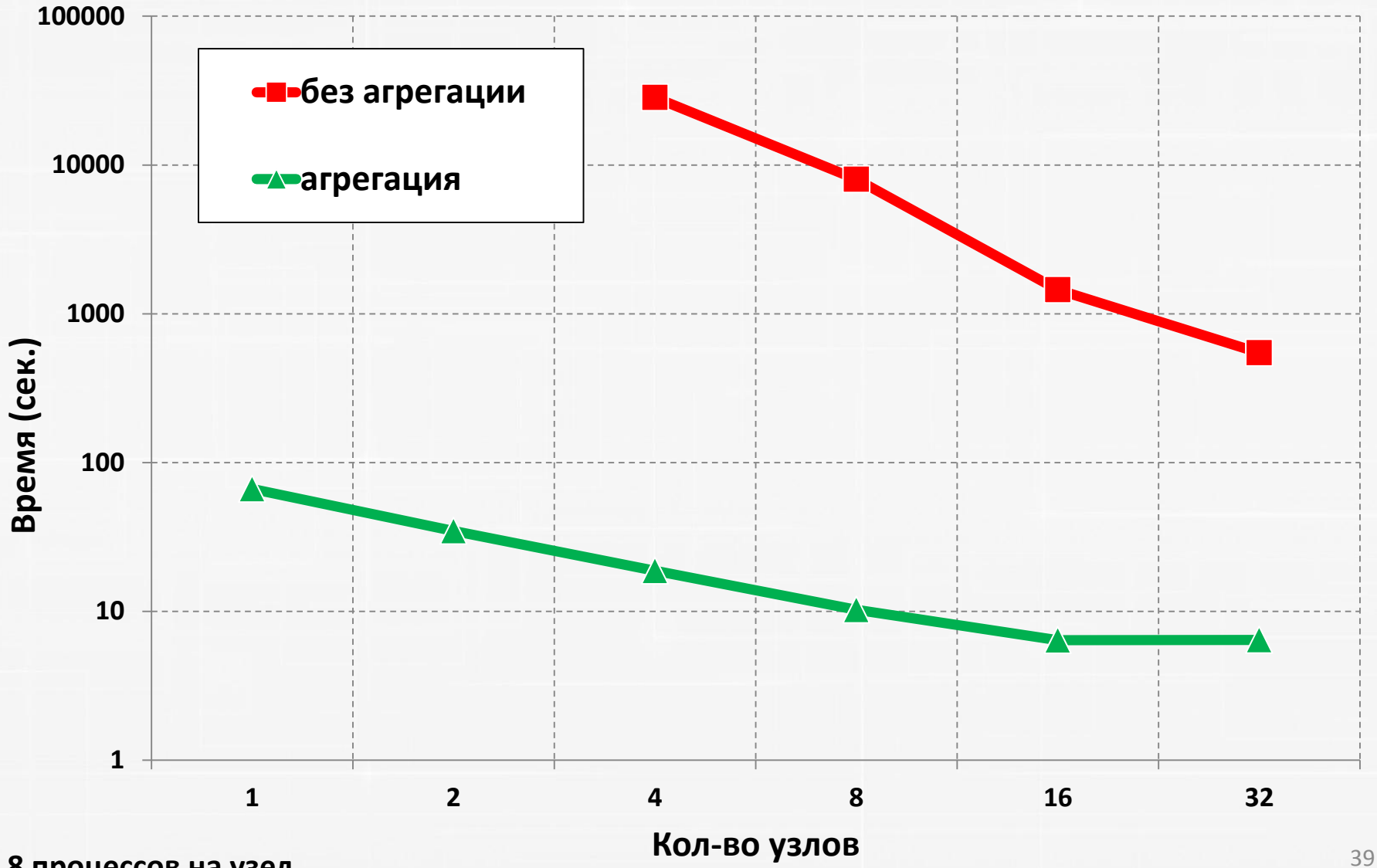
Время работы реализаций на кластере «Ангара-К1». Граф RMAT-23, 2^{23} вершин

НИЦЭВТ

Ростех — Объединенная приборостроительная корпорация



Время работы реализаций на кластере «Ангара-К1». Граф RMAT, 2^{23} вершин



8 процессов на узел

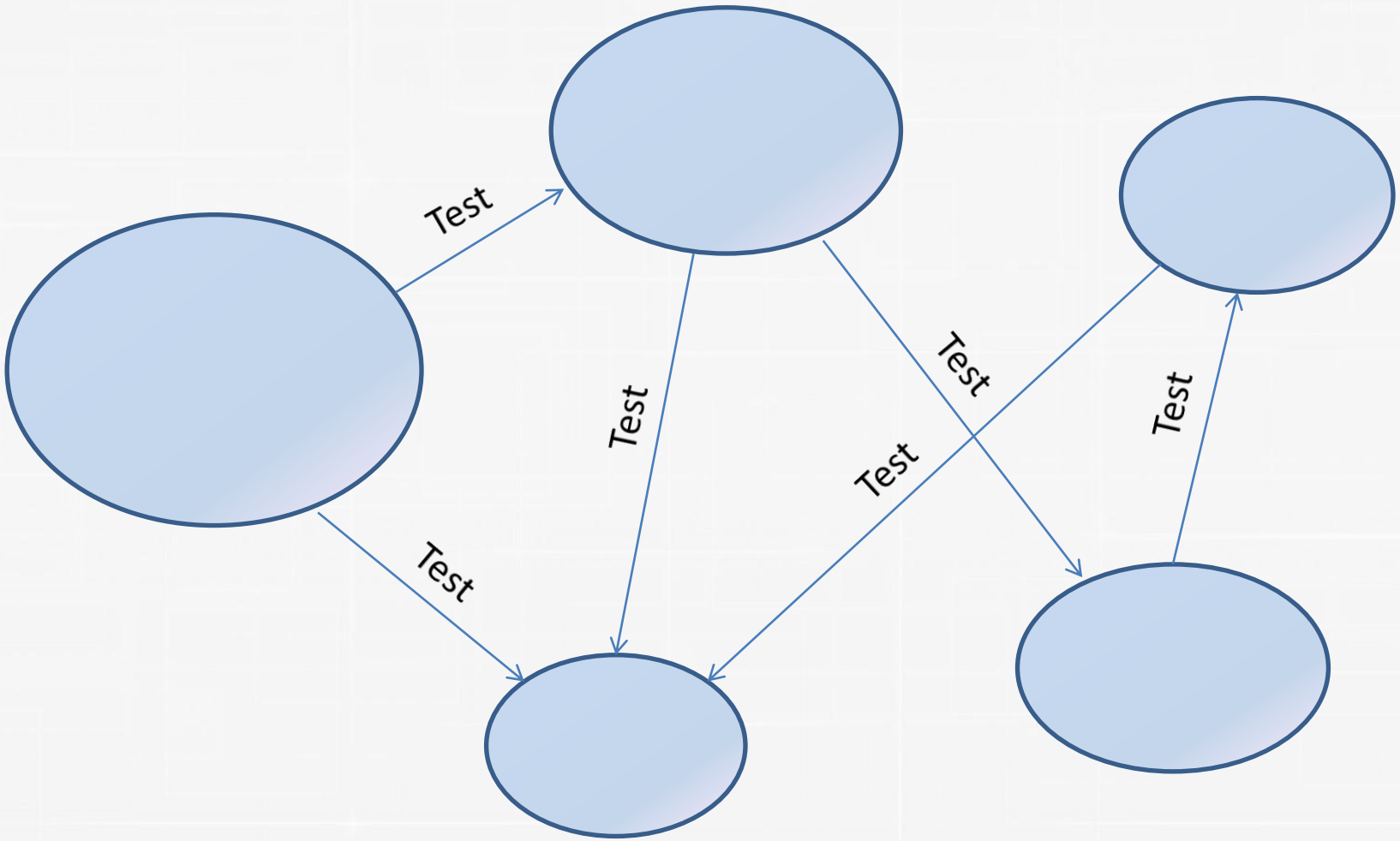
Функция find_corresponding_edge (2)

- Когда приходит сообщение (u, v) нужно узнать локальный номер ребра, по которому оно пришло

Варианты реализации:

- Линейный поиск
- Бинарный поиск
- Хеширование (линейное исследование и вставка)

Много отложенных сообщений типа Test



Обработка сообщений типа Test в отдельной очереди (3)

```
for (v ∈ local_V) {
    wake_up (v); // отправка Connect (запись в буфер)
}
while (1) {
    /* прием сообщений, сообщения типа Test
    записываются в отдельную очередь */
    read_msgs ();

    process_queue (); // обработка основной очереди

    /* обработка очереди с сообщениями типа Test */
    if (time_to_process_test_queue) process_test_queue ();

    if (time_to_send) send_all_bufs ();

    check_finish (); // проверка на завершение
}
```

Оптимизация памяти, типы сообщений, сжатие сообщений (4)

- Оптимизация объема используемой локальной памяти
- Оптимизация длины сообщений
  - special_id: (u, v) -> (proc_id, v)
- Передача сообщений разного типа
  - длина сообщений 12 и 24 байт

Сравнение вариантов реализаций. «Ангара-К1». Граф RMAT, 2^{23} вершин

НИЦЭВТ

Ростех Объединенная приборостроительная корпорация

N	Реализация	1 узел	32 узла	
		Время, сек.	Время, сек.	Ускорение
1	Агрегация	66,51	6,42	10,36
2	Хеширование при поиске	54,51	5,05	10,80
3	Сообщения Test в отдельной очереди	47,77	2,27	21,08
4	Оптимизация памяти, типы и сжатие сообщений	36,05	1,74	20,74

8 процессов на узел

Сравнение результатов на «Ангара-К1» с суперкомпьютером с сетью Infiniband

Кластер «Ангара-К1» vs МВС-10П

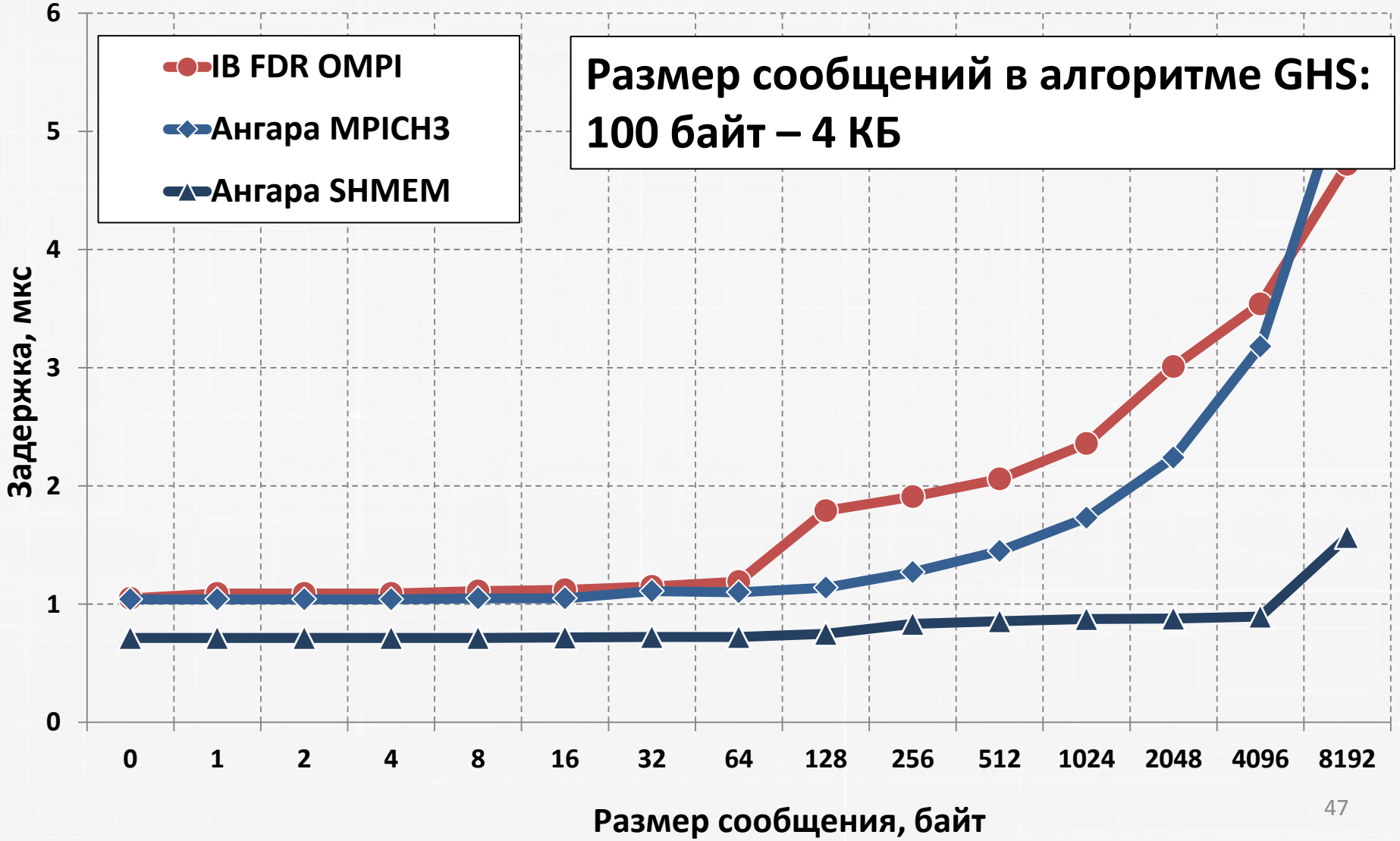
Объединенная приборостроительная корпорация
Ростех

	Ангара		МВС-10П
Узлы	A	2x Xeon E5-2630 по 6 ядер, 2.3 ГГц	2x Xeon E5-2690 по 8 ядер, 2.9 ГГц
	B	Xeon E5-2660 по 8 ядер, 2.2 ГГц	
Количество узлов	24*A+12*B = 36		207 (36)
Память узла	64 ГБ		64 ГБ
Сеть	Ангара 3D-тор 3x3x4		Infiniband 4xFDR Fat Tree

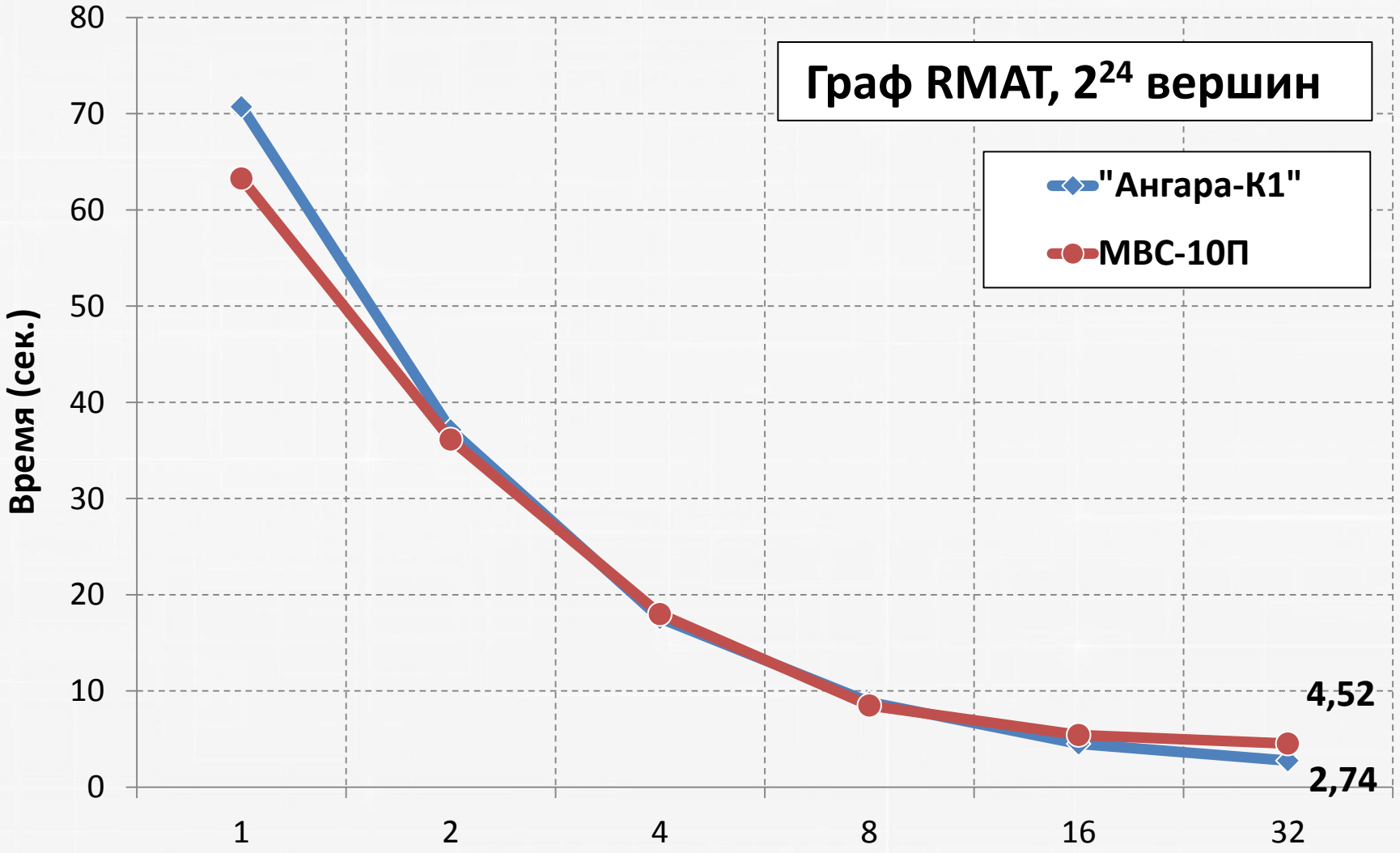
Ангара vs Infiniband. Задержка (тест osu_latency)

НИЦЭВТ

Ростех — Объединенная приборостроительная корпорация

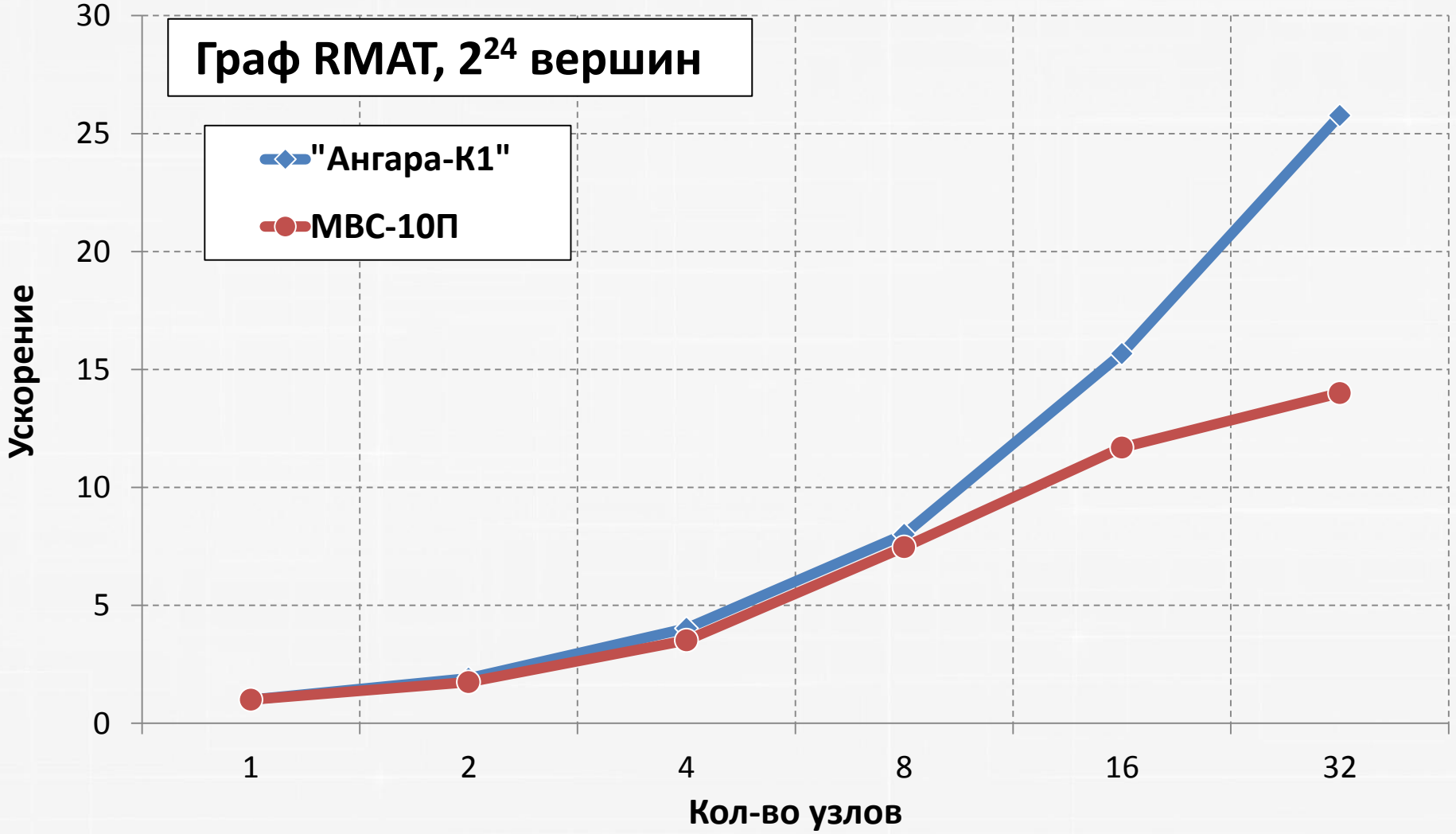


«Ангара-К1» vs МВС-10П. Алгоритм GHS: Время решения



8 процессов на узел

«Ангара-К1» vs МВС-10П. Алгоритм GHS: Ускорение



8 процессов на узел

Выводы и планы на будущее

- На 32 узлах оптимизации позволили ускорить реализацию в 3.7 раза на графе RMAT-23 (по сравнению с версией с агрегацией)
- Решение для графа RMAT-29 с 0.5 млрд вершин и 17 млрд ребер (~256 ГБ) получено за 84 секунды на 32 узлах «Ангара-К1»
- Дальнейшая оптимизация
- Масштабирование на большом числе узлов кластера (Ломоносов, Blue Gene/P)

- Разработка коммуникационной библиотеки для решения сложных задач, оптимизированной для сети Ангара, Infiniband

Спасибо за внимание!
Вопросы?