Extended Routing Table Generation Algorithm for the Angara Interconnect

А.В. Мукосей, А.С. Семенов, А.С. Симонов

24.09.2019

План доклада

- Сеть Ангара
  - Общие сведения
  - Правила маршрутизации
- Описание проблемы
- Расширение алгоритма построения таблиц маршрутов
  - Граф зависимости каналов
  - Безопасное нарушение DOR
  - Алгоритм поиска дополнительных ребер
  - Алгоритм построения таблиц маршрутов
- Исследование разработанного алгоритма построения таблиц маршрутов
- Заключение

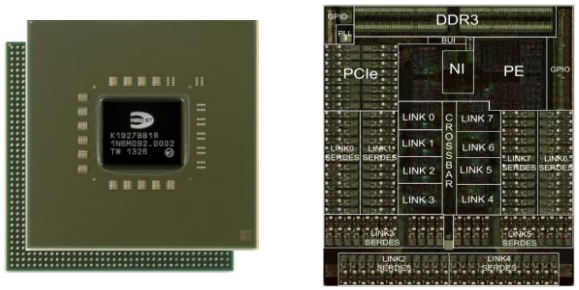
Сеть Ангара

- Топология сети:
  - 1D..4D-тор
- Коммутаторное исполнение
- До 8 каналов связи с соседними узлами
- Задержка на хоп: 130 нс
- Скорость канала связи: 75 Гбит/с
- Масштабирование: до 32К узлов
- Второе поколение Ангара-2: 200 Гбит/с (2020 год)

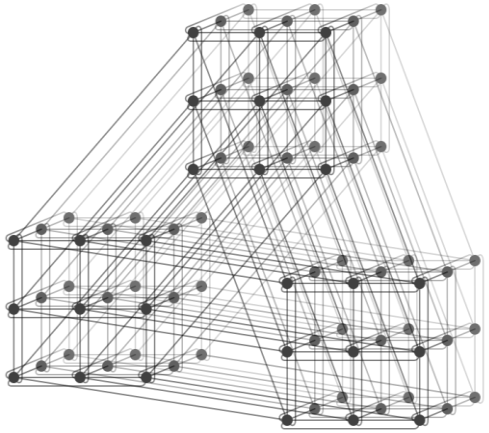
Существующие суперкомпьютеры на базе сети Ангара

Организация	Количество узлов	Конфигурация
АО «НИЦЭВТ»	36	3x3x2x2
ОИВТ РАН	32	4x2x2x2
	24	коммутатор
Центр компьютерного моделирования	96	4x4x3x2
Центр обработки данных	40	5x2x2x2
	8	2x2x2

СБИС маршрутизатора

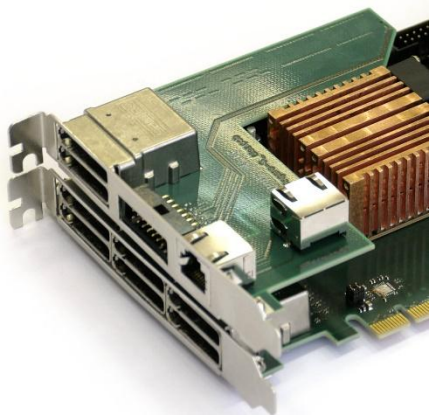


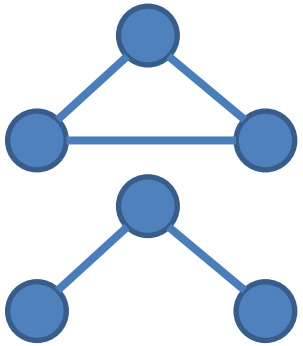
Топология – 3x3x3x3



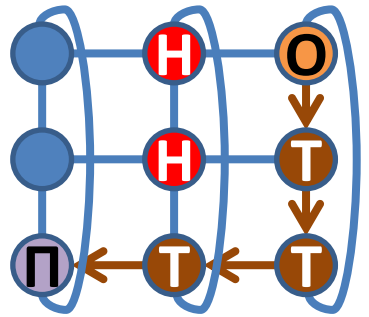
Сеть Ангара

- Сетевой адаптер:
  - 8 портов: $+X, -X, +Y, -Y, +Z, -Z, +K, -K$
  - Объединение в топологию «многомерный тор» (до 4х измерений)
- Ресурсы сети:
  - Каждый вычислительный узел с установленным адаптером сети Ангара представляет из себя **узел сети**
  - **Узлы сети** соединяются друг с другом по средством **каналов связи** (линков)
  - **Узлы сети** отправляют и принимают **сообщения** (сетевые пакеты)
  - Отправленные сообщения доставляются узлу назначения, проходя через **транзитные узлы** и **каналы связи** (линки)
  - Сообщения передаются по сети по реализованным и фиксированным в маршрутизаторе **правилам**





1D кольцо и решетка



2D кольцо/решетка

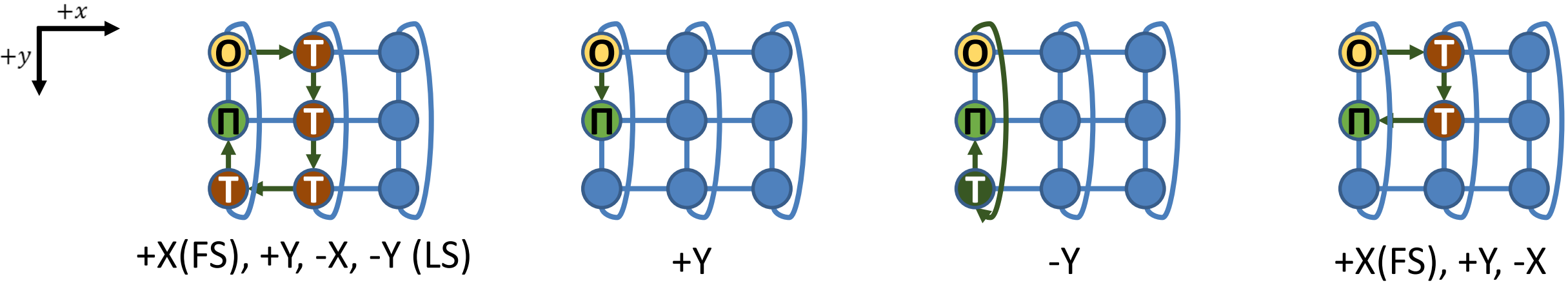
Правила маршрутизации

- Пакеты в сети передается согласно фиксированным в чипе маршрутизатора правилам детерминированной маршрутизации. Пакет по очереди совершает движения в различных направлениях, переходя от одного узла сети к другому.

- **Правила детерминированной маршрутизации:**
  1. Один первый шаг FS (First Step) в положительном направлении: $+X, +Y, +Z$ или $+K$ (если есть)
  2. Набор шагов в различных направлениях, таких что:
     - Движение осуществляется в заданном порядке направлений: $+X, +Y, +Z, +K, -X, -Y, -Z, -K$ (правило порядка направлений Direction Order или DOR-маршрутизация)
     - Движение возможно только в положительную или в отрицательную сторону по измерению (правило dirbit-маршрутизации). Движение по измерению может отсутствовать
  3. Один последний шаг LS (Last Step) в отрицательном направлении: –X, –Y, –Z, или –K (если есть)

Таблица маршрутов

- Правила маршрутизации позволяют задавать различные варианты маршрутов между двумя узлами
- *Таблица маршрутов* – множество путей для каждой пары узлов сети

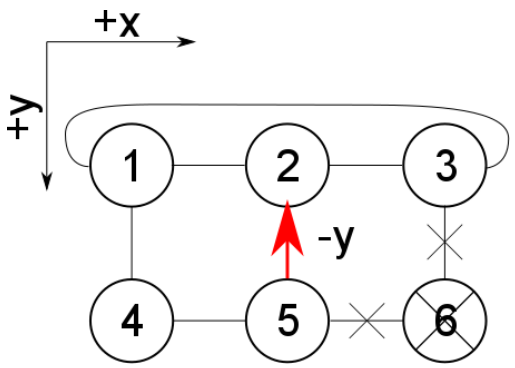


+X(FS), +Y, -X, -Y (LS) +Y -Y +X(FS), +Y, -X

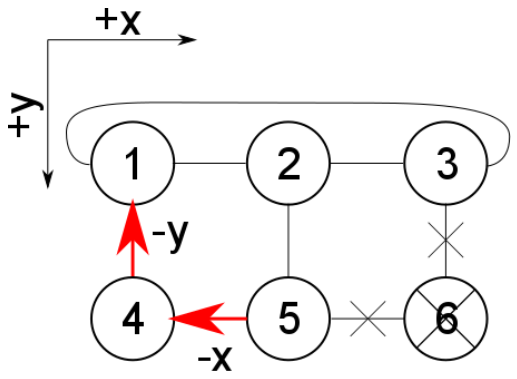
- Построение таблицы маршрутов – задание для каждой пары узлов маршрута, удовлетворяющего правилам маршрутизации
- В первой версии алгоритма построения таблиц маршрутов для гарантии отсутствия взаимных блокировок правило FS/LS подчиняется правилу DOR

Проблема

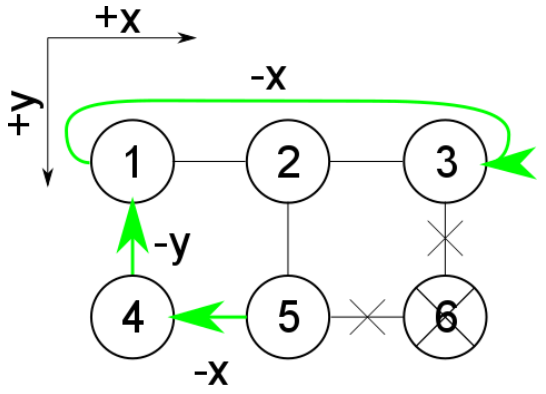
- Рассмотрим топологию 3x2 с одним недоступным узлом.
  - Узел 6 недоступен
- Отсутствует путь из узла 5 в узел 3. Из узла 5 можно пойти:
  - а) по направлению $-Y$, переходя при этом в узел 2. По правилам маршрутизации дальнейшее движение возможно только в $-Y$, но дальше нет путей
  - б) по направлению $-X$, переходя при этом в узел 4. Из узла 4 есть только путь в направлении $-Y$ к узлу 1 аналогично предыдущему варианту
- Нарушение правила порядка направлений



а)



б)



в)

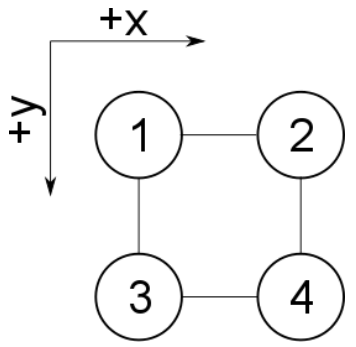
Постановка задачи

- Модификация (расширение) алгоритма построения таблиц маршрутов:
  - FS/LS не обязаны удовлетворять правилу порядка направлений

- Решение задачи
  - Необходимо определить, когда FS/LS может нарушить правило порядка направлений

Граф зависимости каналов

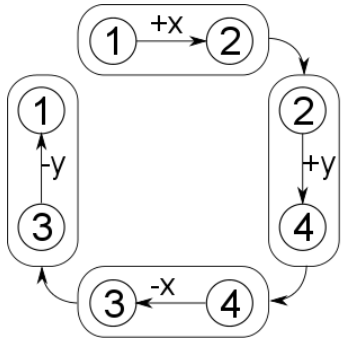
- Множество всех узлов в сети обозначим N
- Если узел u соединен каналом связи с узлом v в направлении s, то мы будем говорить что $v = u + s$
- Канал связи обозначим $\alpha_{u,s}$, где $u \in N, s \in \mathcal{D}$, где $\mathcal{D}$ – множество направлений $\{+X, +Y, +Z, +K, -X, -Y, -Z, -K\}$
- Множество $\mathcal{D}$ упорядочено, будем говорить, что $s_i > s_j$, если s_i идет после s_j.
- Граф зависимости каналов (channel dependency graph, CDG) $\mathrm{G(V, E)}$:
  - Множество V вершин графа G – множество каналов связи
  - Множество E ребер графа G – пара каналов связи $(\alpha_{v,s_v}, \alpha_{u,s_u})$, где $\alpha_{v,s_v}, \alpha_{u,s_u} \in V$, $s_v, s_u \in \mathcal{D}$, $v, u \in N, u = v + s_v$, а также переход из канала связи α_{v,s_v} в канал связи α_{u,s_u} разрешен по

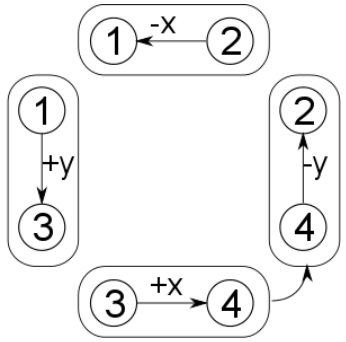


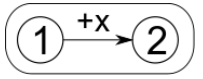
2D решетка

Узел сети

Линк







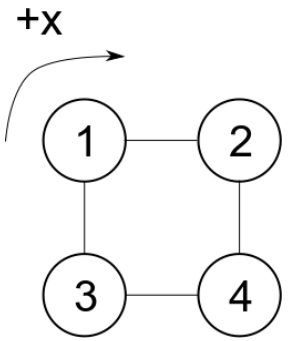
Вершина CDG $\alpha_{1,+x}$

Ребро CDG

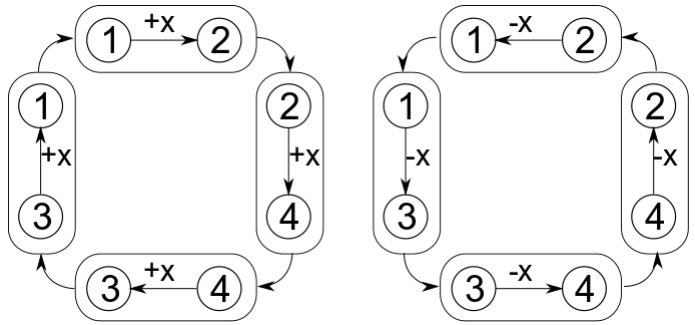
Соответствующий CDG для 2D решетки с DOR

Граф зависимости каналов

- **Теорема (1988)**
  - Если в графе зависимости каналов CDG отсутствуют циклы, то в сети отсутствуют взаимные блокировки
- **Замечания**
  - Кольца в одном измерении разрешаются с помощью правила пузырька
  - Граф зависимости каналов CDG не позволяет описать полностью маршрутизацию в сети Ангара (из за dirbit-маршрутизации)
  - Правило dirbit-маршрутизации в сети Ангара вносит дополнительные ограничения к правилу порядка направлений
  - В дальнейшем мы будем строить CDG для DOR маршрутизации



1D кольцо

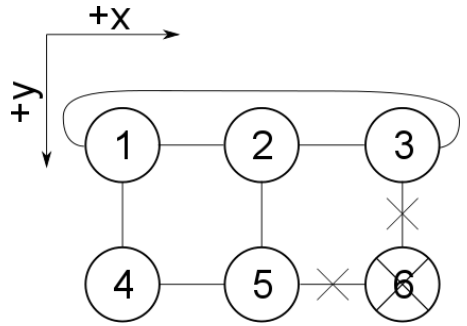


Соответствующий CDG для 1D кольца с DOR

Безопасное нарушение DOR

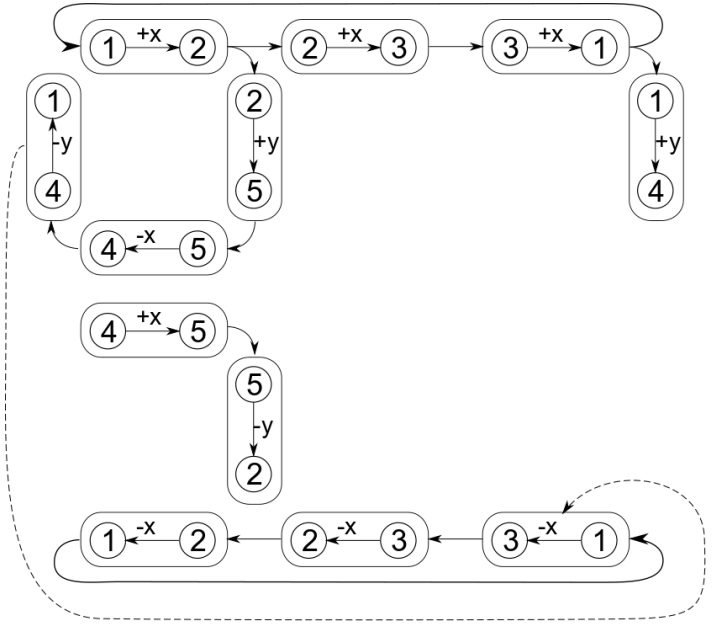
- Рассмотрим топологию 3x2 с одним не доступным узлом
  - Узел 6 недоступен

- Рассмотрим соответствующий граф CDG. Если добавить ребро от вершины $\alpha_{4,-y}$ к вершине $\alpha_{1,-x}$ (пунктирная линия), то:
  - Нарушается правило порядка направлений: $-Y - X$
  - Не возникает цикл
  - Возникает путь от узла сети 5 к узлу 3
    - $5 + -X = 4$
    - $4 + -Y = 1$
    - $1 + -X = 3$



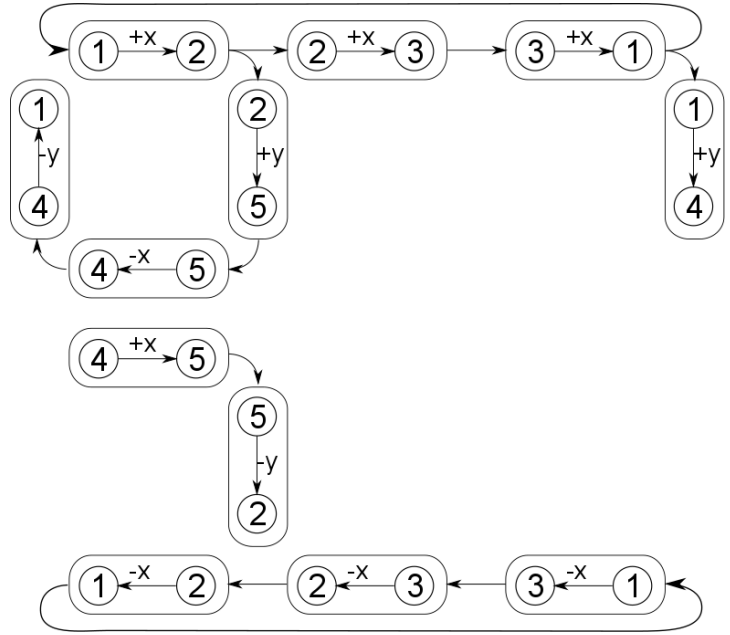
Сеть с топологией 3x2

Узел 6 сломан

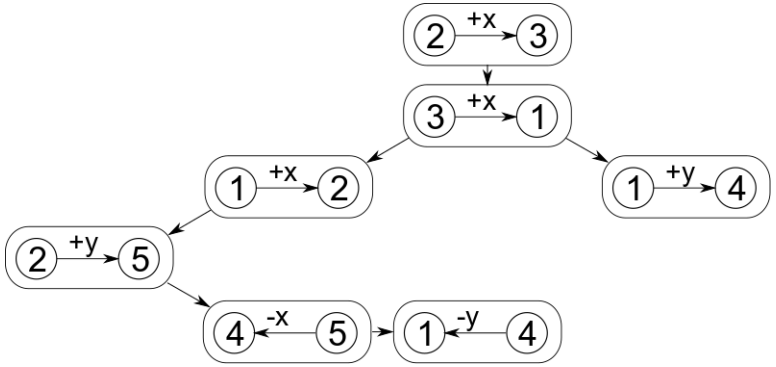


Соответствующий CDG для топологии 3x2

Безопасное нарушение DOR



- Множество используемых направлений $B(\alpha_{u,s_u})$ – это такой набор всевозможных направлений, которые будут проходить все возможные пути в графе G, начинающиеся из вершины α_{u,s_u}.
- На рисунке: $B(\alpha_{2,+x}) = \{+X, +Y, -X, -Y\}$.



Безопасное нарушение DOR

Теорема

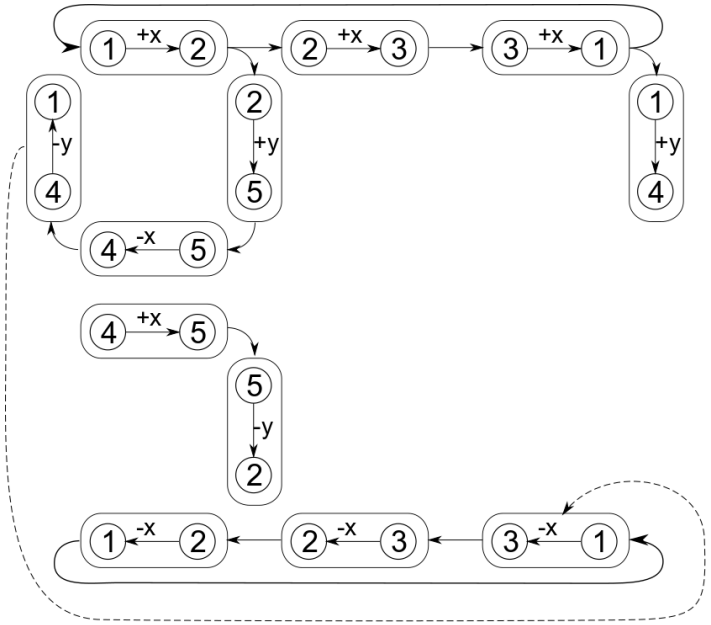
Дополнительное ребро $(\alpha_{u_j,s_j}, \alpha_{u_k,s_k})$, такое что $s_j > s_k$ и $u_k = u_j + s_j$ не создаст цикла в графе зависимости каналов G, если $s_j \notin B(\alpha_{u_k,s_k})$

Доказательство

Допустим, дополнительное ребро $(\alpha_{u_j,s_j}, \alpha_{u_k,s_k})$ создает цикл в графе G . Это означает, что в графе G существовал путь из вершины α_{u_k,s_k} в вершину α_{u_j,s_j}. Это в свою очередь означает, что $s_j \in B(\alpha_{u_k,s_k})$, но это противоречит условиям теоремы. ■

Безопасное нарушение DOR

- Дополнительное ребро $(\alpha_{4,-y}, \alpha_{1,-x})$ не создаст цикла, так как:
  $-y \notin B(\alpha_{1,-x})$, где $B(\alpha_{1,-x}) = \{-X\}$



- Заметим, что множество используемых направлений изменится для вершины $\alpha_{4,-y}$ и для всех вершин графа G, из которых есть путь в вершину $\alpha_{4,-y}$

Алгоритм поиска дополнительных ребер

- 1 этап
  - Методом поиска вширь (BFS) для каждой вершины α_{u,s_u} графа G находятся множество используемых направлений $B(\alpha_{u,s_u})$.
- 2 этап
  - Поиск всех пар $(\alpha_{u_j,s_j}, \alpha_{u_k,s_k})$, где $s_j > s_k$, $u_k = u_j + s_j$ и $s_j \notin B(\alpha_{u_k,s_k})$. Если такая пара существует, то в G добавляется ребро
  - После каждого добавленного ребра необходимо обновить множества используемых направлений
  - Для каждой вершины, посещенной методом BFS по графу G с инвертированными ребрами $\bar{E}$, запущенным из вершины α_{u_j,s_j}, необходимо добавить к множеству используемых направлений множество $B(\alpha_{u_k,s_k})$
- Замечание
  - Второй этап выполнялся сначала для $s_j \in \{-X, -Y, -Z, -K\}$, затем для $s_j \in \{+X, +Y, +Z, +K\}$

Алгоритм построения таблиц маршрутов

- Таблица маршрутов – множество путей для каждой пары узлов сети.
- В прошлых работах был разработан граф путей, описывающий всевозможные пути в сети Ангара. Для каждого узла сети $i \in N$ в графе путей строились вершины:
  - U^i_{begin} – вершина соответствующая началу маршрута
  - U^i_{FS} – вершина в которую можно попасть совершив первый положительный шаг $FS \in \mathcal{D}$
  - $U^i_{dirbit_k}$ – вершина в которую можно попасть совершив шаги в направлениях $dirbit_k$. Направления $dirbit_k$ соответствуют правилу порядка направлений и dirbit-маршрутизации
  - U^i_{LS} – вершина в которую можно попасть совершив последний отрицательный шаг $LS \in \mathcal{D}$
  - U^i_{end} – вершина соответствующая концу маршрута
- Вершины графа связаны таким образом, что переход от одной вершины к другой соответствует пути пакета в сети и удовлетворяет правилам маршрутизации

Свойства

- Путь в сети между двумя узлами $u, v \in N$ существует тогда и только тогда, когда существует путь в графе путей из вершины U^u_{begin} в вершину U^v_{end}
- Пути в сети могут быть построены с помощью алгоритма поиска вширь

Алгоритм построения таблиц маршрутов

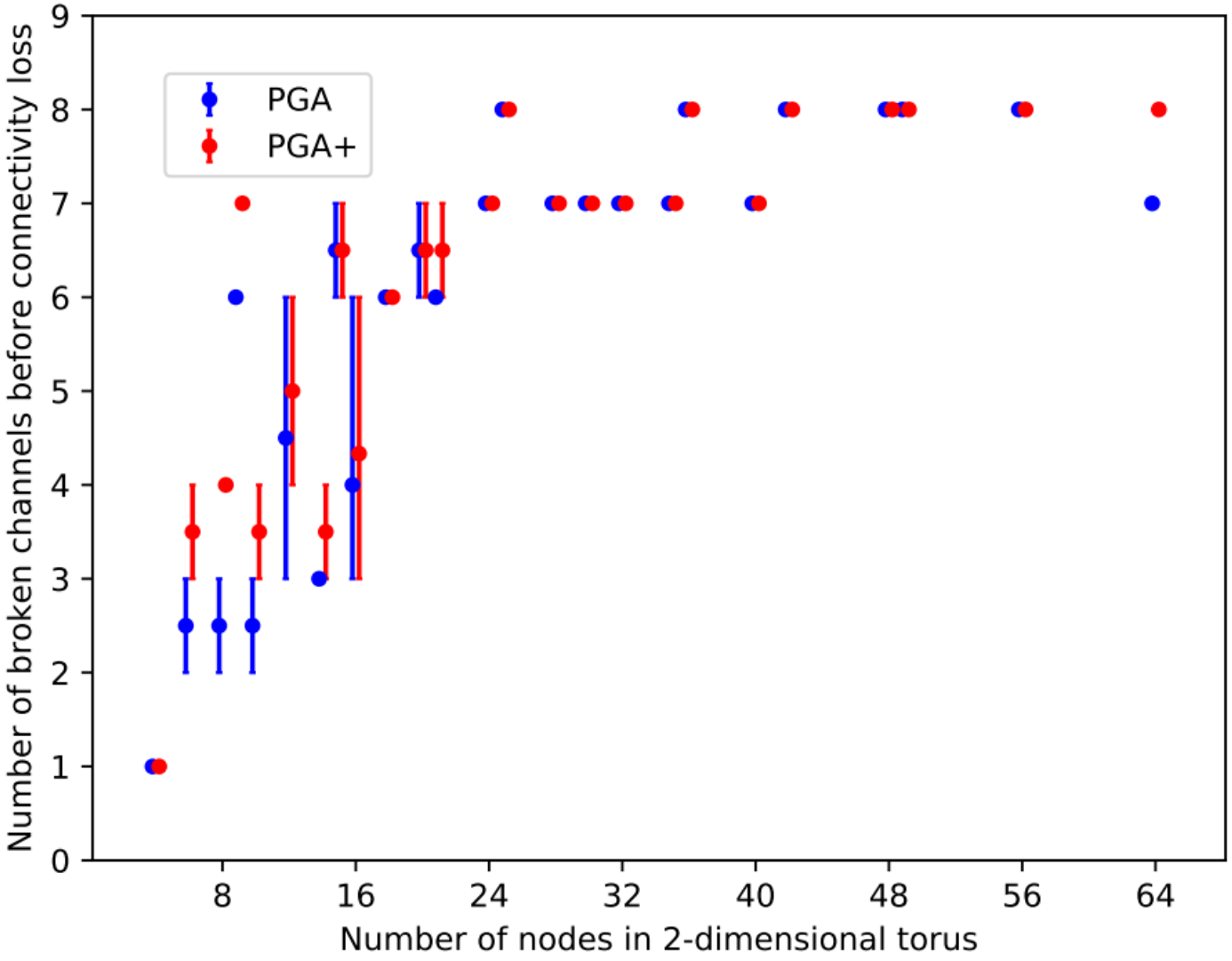
- В первой версии графа путей все ребра соответствовали правилу порядка направлений
  - Алгоритм построения путей по первой версии графа обозначим как PGA

- В новой версии для вершин U_{FS}^{i} и U_{LS}^{i} были добавлены дополнительные ребра, нарушающие правило порядка направлений, но не приводящие к взаимным блокировкам согласно анализу соответствующего графа зависимости каналов
  - Алгоритм построения путей по новой версии графа обозначим как PGA+

Исследования

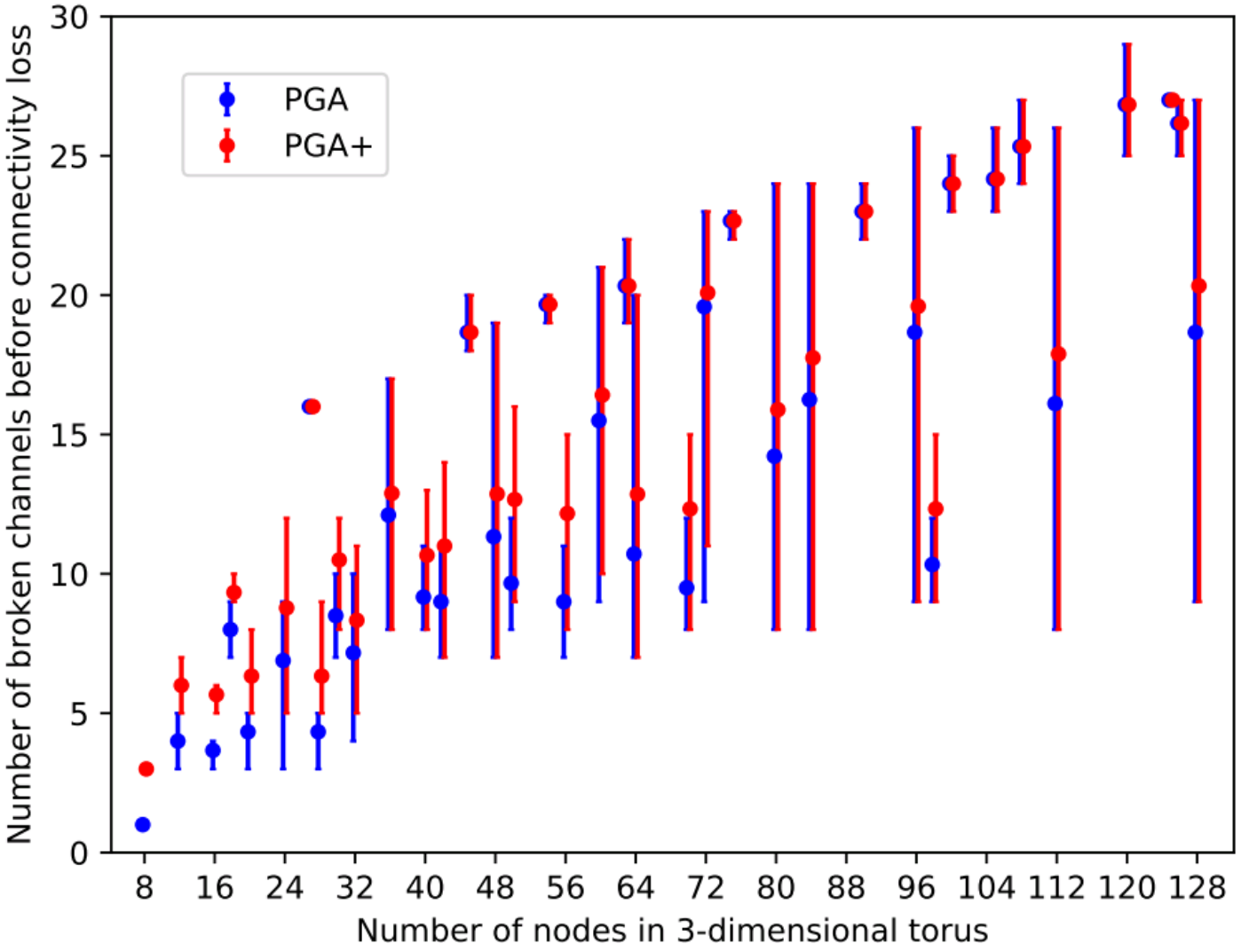
- Исследования проводились на различных топологиях
  - Число узлов до 128
  - 2D, 3D и 4D торы
  - Размерности тора ограничены $2 \leq d_i \leq 8$, где $d_i -$ размерность тора в i-ом измерении
- Для каждой системы случайно ломалось некоторое количество каналов связи
- Для каждого числа сломанных каналов связи происходило как минимум 100 попыток генерации сломанных каналов связи и построения таблиц маршрутов алгоритмами PGA и PGA+
- Для каждой системы и каждого метода фиксировалось количество сломанных каналов связи, после которого система становилась несвязной (отсутствовал путь между некоторой парой узлов) для всех 100 попыток

Исследование 2D топологий



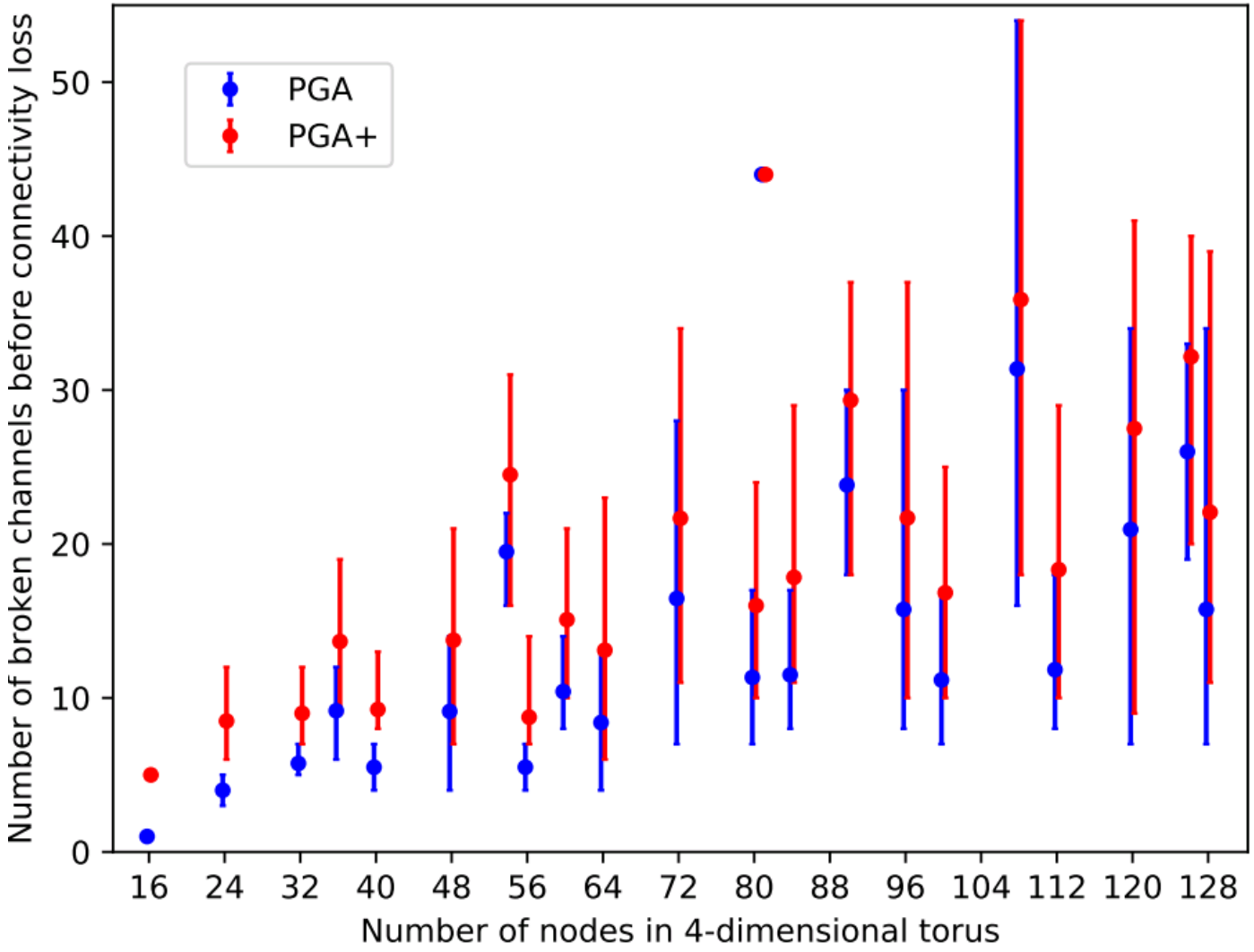
- Для систем с числом узлов, для которого возможно построить различные топологии на рисунке представлено максимальное, минимальное и среднее значение числа линков
- В среднем количество сломанных линков до потери связности увеличилось на 4,91%

Исследование 3D топологий



- В среднем количество сломанных линков до потери связности увеличилось на 8,17%

Исследование 4D топологий



- В среднем количество сломанных линков до потери связности увеличилось на 34,05%

Заключение

- Предложен алгоритм поиска и построения дополнительных поворотов, нарушающих правило порядка направлений и не допускающий возникновения взаимных блокировок
- Разработанный алгоритм интегрирован в алгоритм построения таблиц маршрутов на основе графа путей для сети Ангара
- Новый алгоритм увеличивает число отказов до потери связности:
  - На 4,91% на 2D топологиях
  - На 8,17% на 3D топологиях
  - На 34,05% на 4D топологиях

Спасибо за внимание!