ОЦЕНКА ЗАДЕРЖКИ И ПРОПУСКНОЙ СПОСОБНОСТИ СЕТИ С ТОПОЛОГИЕЙ «МНОГОМЕРНЫЙ ТОР» ПРИ НАЛИЧИИ ОТКАЗАВШИХ КАНАЛОВ СВЯЗИ

И.А. Пожилов

ОАО «НИЦЭВТ»

Введение

В ОАО «НИЦЭВТ» ведется разработка высокоскоростной коммуникационной сети «Ангара» с топологией «многомерный тор» [1, 2]. В 2013 году был выпущен кристалл маршрутизатора этой сети, к 2015 году ожидается построение суперкомпьютера на его основе.

Современные суперкомпьютерные системы включают в себя большую совокупность элементов: процессоров, модулей памяти, сетевых устройств, устройств ввода-вывода, связанных в единый программно-аппаратный комплекс. И хотя каждый из этих элементов независимо обладает большой степенью надежности, с ростом количества подобных элементов система в целом обнаруживает тенденцию к возникновению непредвиденных неисправностей. Возникает необходимость контроля за развитием ситуации и поддержания работоспособности системы даже в условиях ограниченной функциональности отдельных ее элементов.

В работе [3] была решена задача связности, т. е. задача построения детерминированного маршрута, отвечающего правилу порядка направлений, в обход отказавших каналов связи. Но при построении подобных маршрутов страдает производительность сети: увеличивается диаметр, так как маршруты в обход отказов в большинстве случаев оказываются неминимальными, уменьшается общая пропускная способность сети в следствие уменьшения количества работающих каналов связи. Для оценки производительности сетей широко применяется понятие бисекционной пропускной способности [4]. В данной работе применяются результаты статьи [5] для обобщения понятния бисекционной пропускной способности для получения более точной оценки производительности.

Показатели производительности

В данной работе предлагается исследование двух классов показателей производительности, отвечающих таким параметрам, как задержка передачи и пропускная способность.

Задержка при передаче между двумя узлами сети определяется фиксированными свойствами аппаратуры (тактовой частотой процессора и маршрутизатора, шины связи между ними, архитектурными особенностями маршрутизатора), а так же топологией сети: она пропорциональна количеству промежуточных узлов на пути между двумя узлами, т. е. длине пути. Таким образом, предлагается исследовать диаметр сети.

Множество узлов сети обозначим N. Множество соединяющих их каналов связи E:

$$E=\{(u,v)|u,v\in N, существует\,канал\,u\to v\}.$$

Пусть $d(u,v)$ обозначает длину пути между двумя узлами сети u и v.

Определение 1. *Диаметром* сети называется максимальное расстояние между двумя узлами сети:

$$D=\max_{u,v\in N} d(u,v).$$

Определение 2. *Средней длиной пути* называется усредненная по всем парам узлов длина пути между двумя узлами:

$$\overline{D}=\frac{1}{N(N-1)}\sum_{u,v\in N} d(u,v).$$

Эти два показателя характеризуют коммуникационную задержку в худшем случае и в среднем.

Для того, чтобы характеризовать пропускную способность сети, введем несколько определений.

Рассмотрим разбиение множества узлов сети N на два непересекающихся подмножества U и W:

$$U\cup W=N,\ U\cap W=\varnothing.$$

Разрезом назовем множество каналов связи, соединяющих узлы из разных множеств:

$$\mathrm{cut}(U,W)=\{(u,w)\in E|u\in U, w\in W\}.$$

Если считать, что каждый канал связи обладает некоторой пропускной способностью BW_{link}, то пропускной способностью разреза назовем величину $cut(U,W)\cdot BW_{\mathrm{link}}$.

Классической метрикой для оценки пропускной способности является бисекционная пропускная способность.

Определение 3. *Бисекционной пропускной способностью* называется *минимальная* пропускная способность разреза сети на две равные по величине части:

$$BW_{\text{bisect}} = \min_{|U|=|W|} \text{cut}(U, W) \cdot BW_{\text{link}}.$$

Для многомерного тора размером $N_1 \times N_2 \times \cdots \times N_k$, бисекционная пропускная способность вычисляется как $2N / N_{\max} \cdot BW_{\text{link}}$, где $N_{\max}$ – максимальная из размерностей тора N_j.

Покажем важную связь между бисекционной пропускной способностью и производительностью сети. Рассмотрим паттерн обмена сообщениями «все всем»: каждый узел сети u производит отправку некоторого сообщения размером M каждому из остальных $N-1$ узлов сети. Рассмотрим разбиение множества узлов на подмножества U и W. Каждый из узлов множества U должен отправить сообщение каждому из узлов множества W, причем сообщения обязательно проходят через разрез $\text{cut}(U, W)$, обладающий некоторой пропускной способностью. Таким образом, можно оценить время выполнения операции «все всем» снизу следующим образом:

$$T \geq \frac{M_{\text{total}}}{BW_{\text{total}}} = \frac{M|U||W|}{BW_{\text{link}} \cdot \text{cut}(U, W)} = \frac{M}{BW_{\text{link}}} \frac{1}{c}, \tag{1}$$

где

$$c = \frac{\text{cut}(U, W)}{|U||W|}.$$

Заметим, что оценка времени T справедлива для *произвольного* разбиения U, W, но наиболее интересной она становится при оптимальном значении c_0:

$$c_0 = \min_{U,W} \frac{\text{cut}(U, W)}{|U||W|},$$

т. е. когда правая часть оценки (1) достигает максимума.

Разбиение U, W, отвечающее минимуму этого функционала, называется *пропорциональным разбиением* (ratio cut partitioning, [5]), а величина c_0 – *ценой* этого разбиения.

В работе [5] предложен спектральный метод построения приближения пропорционального разбиения. Для этого строится *лапласиан* графа связей (в нашем случае – многомерного тора). *Лапласиан* – это матрица $L = A - D$, где A – матрица смежности, D – диагональная матрица степеней вершин. Лапласиан графа обладает рядом интересных свойств.

1. Она является самосопряженной (из неориентированности графа).
2. Положительно определенной (следует из свойства диагонального доминирования).
3. Имеет собственное значение $\lambda_1 = 0$ (т. к. сумма элементов в каждой строке равна нулю, по построению).
4. Из симметричности и положительной определенности имеет множество собственных значений $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N \in \mathbb{R}$.
5. Количество нулевых собственных значений равно числу компонент связности графа.
6. Для оптимальной величины c_0 выполнено неравенство $c_0 \geq \lambda_2 / N$, это доказано в [5].

Заметим, что любое разбиение U, W дает верхнюю оценку на c_0:

$$c_0 \leq \frac{\text{cut}(U, W)}{|U||W|},$$

т. к. c_0 – минимум этого отношения. Задача – найти хорошую верхнюю оценку.

В [5] обоснован эвристический способ нахождения верхней оценки: выбирается собственный вектор $v = v_2$, соответствующий собственному значению λ_2, и рассматриваются разбиения вида

$$U_j = \{ i \in \overline{1, N} \mid v_i < v_j \} \text{ для } j \in \overline{1, N},$$

и среди них находится оптимальное:

$$c^{\text{eigen}} = \min_{j \in \overline{1, N}} \frac{\text{cut}(U_j, N \setminus U_j)}{|U_j||N \setminus U_j|}.$$

Другую оценку сверху можно получить, перебрав все разбиения тора на две равные части аналогично построению бисекционной пропускной способности:

$$c^{\text{bisect}} = \min_{|U|=|W|} \frac{\text{cut}(U, W)}{|U||W|}.$$

Итоговая оценка пропорционального разбиения:

$$\bar{c} = \min\left\{c^{\text{eigen}}, c^{\text{bisect}}\right\}.$$

Пусть получено значение $\bar{c}$ для сети без отказов. Для сети с отказами обозначим соответствующую оценку $\bar{c}_F(N_f)$, где N_f — число отказавших каналов связи, и будем рассматривать отношение $\bar{c}_F/\bar{c}$ с целью исследования влияния отказов на пропускную способность. Отметим, что это лишь оценка сверху. Для более точного оценивания производительности сети при наличии отказов необходимо применять имитационное моделирование, что лежит за рамками данной статьи.

Изменение оценок при росте числа отказов

Рассмотрим зависимость введенных ранее оценок от количества отказавших каналов связи. Для фиксированного количества отказавших каналов будем рассматривать генеральную совокупность всех возможных расположений отказов в сети. Размер генеральной совокупности не позволяет перебрать все возможные варианты, поэтому будем применять метод Монте-Карло с числом итераций N= 10000.

Для каждого показателя из метода случайного поиска были найдены минимальное, среднее и максимальное значение.

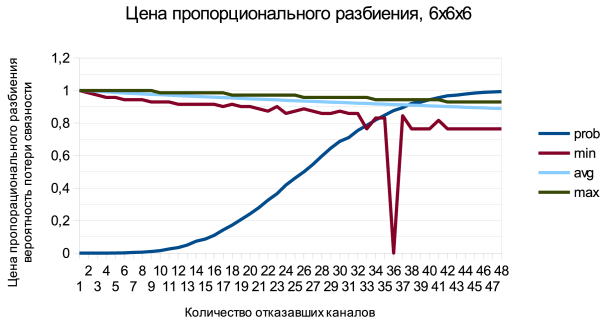

Рис. 1. Цена пропорционального разбиения и вероятность потери связности для тора 6x6x6

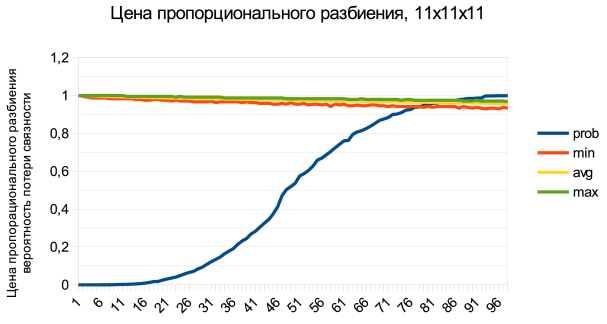

Рис. 2. Цена пропорционального разбиения и вероятность потери связности для тора 11x11x11

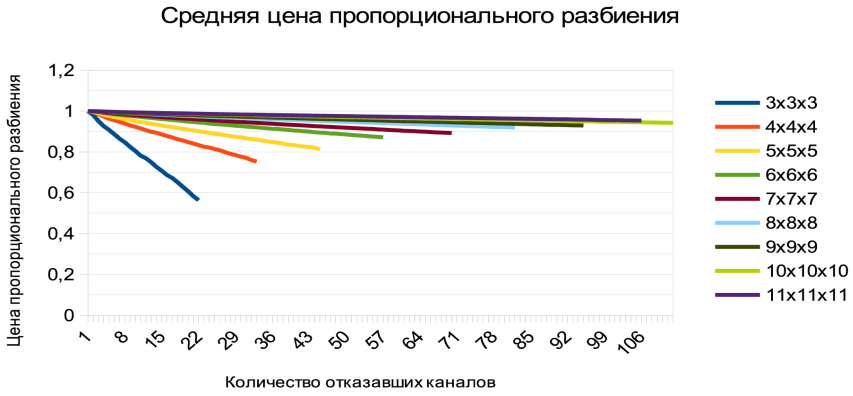

Рис. 3. Средняя цена пропорционального разбиения для 3D-торов разных размеров

На Рис.1 и Рис. 2 изображены графики $\bar{c}_F/\bar{c}$, т. е. относительного изменения цены пропорционального разбиения в зависимости от числа отказавших каналов связи. Изображены графики минимального, среднего и максимального значения этой величины, полученные из метода Монте-Карло. Для сравнения на этих графиках нанесен также график вероятности потери связности для соответствующих размеров торов (взятый из [3]). Из этих графиков видно, что даже при количестве отказов, когда с большой вероятностью будет невозможно построить детерминированный маршрут между некоторой парой узлов, оценка производительности, полученная из (1), ухудшится не сильно. Это позволяет судить о том, что пропускная способность сети снизится незначительно, и более серьезной проблемой является связность детерминированной сети.

На Рис.1 видно падение оценки до нуля для 36 отказов, это следствие недостаточного размера выборки: оценка 0 говорит о том, что нашлась конфигурация отказов, которая приводит к появлению двух компонент связности в графе, при этом цена пропорционального разбиения получается равной нулю (из определения). Для 3D-тора минимальное число отказавших каналов связи, приводящее к подобной ситуации, равно 6 (отказ всех

каналов из одной вершины), но подобная конфигурация маловероятна, поэтому на графике подобная ситуация встретилась только при числе отказов 36.

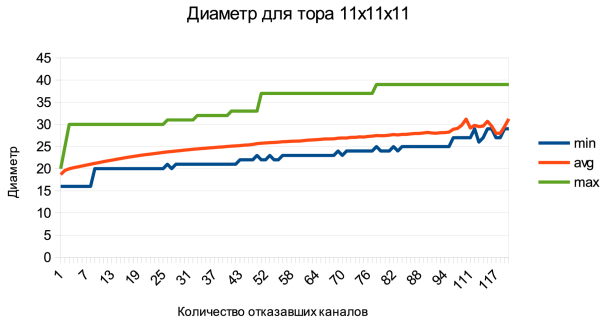

Рис. 4. Диаметр для тора размером 11x11x11, минимальное, среднее и максимальное (худшее значения)

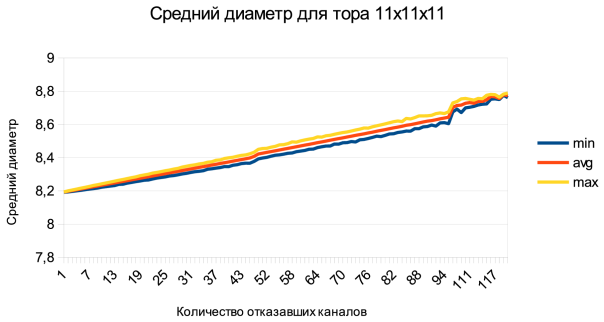

Рис. 5. Средняя длина пути для тора 11x11x11, минимальное, среднее и худшее значения

На Рис. 3 изображены графики усредненной по итерациям метода Монте-Карло цены пропорционального разбиения для 3D-торов разных размеров. Из этого графика видно, что с увеличением размера тора относительное падение пропускной способности уменьшается. На этом изображении не показаны графики вероятности потери связности, чтобы не усложнять иллюстрацию. Для тора 3x3x3 вероятность, близкая к единице достигается для числа узлов 20 (из [3]). Таким образом, для тора 3x3x3 падение производительности будет около 40% даже при наличии связности.

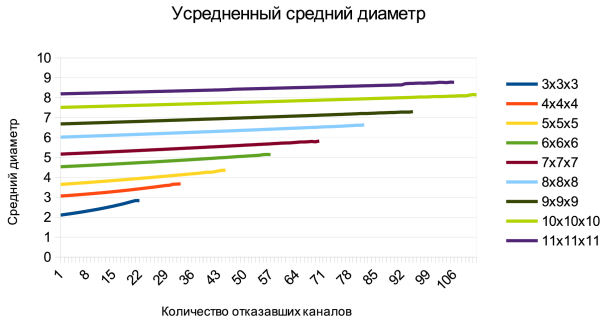

Рис. 6. Усредненное (по итерациям метода Монте-Карло) значение средней длины пути для торов разных размеров

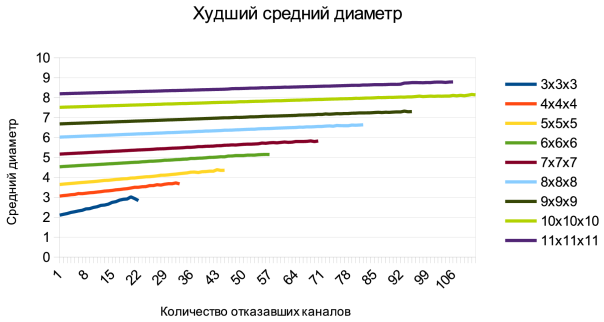

Рис. 7. Худшее значение средней длины пути среди проведенных итераций метода Монте-Карло

На следующих рисунках изображены графики диаметра и средней длины пути в сети. На Рис. 4-5 изображены графики зависимости диаметра и средней длины пути от количества отказавших каналов связи. Показаны минимальное, среднее и худшее значение диаметра, полученные из метода Монте-Карло. Видим, что средняя длина пути увеличивается незначительно: с 8.2 до 8.8 (около 7%), причем разница между минимальным и худшим показателем остается небольшой, т. е. средняя длина пути слабо зависит от конкретного набора отказавших каналов, в основном зависимость от их количества. С диаметром ситуация обстоит иначе: со значения 20 в худшем случае диаметр может увеличиться до 40 (в два раза), в среднем наблюдается увеличение до 30 (на 50%). Это говорит о том, что следует ожидать заметное увеличение максимальной задержки передачи при увеличении числа отказавших каналов связи.

На Рис. 6-7 показана зависимость средней длины пути от количества отказавших каналов связи для торов разных размеров. На Рис. 6 показано среднее значение средней длины пути по итерациям метода Монте-Карло, на Рис. 7 – худшее значение. Для всех торов наблюдается тот же рост на 7-8% на 100 отказов средней длины пути при увеличении числа отказов каналов связи.

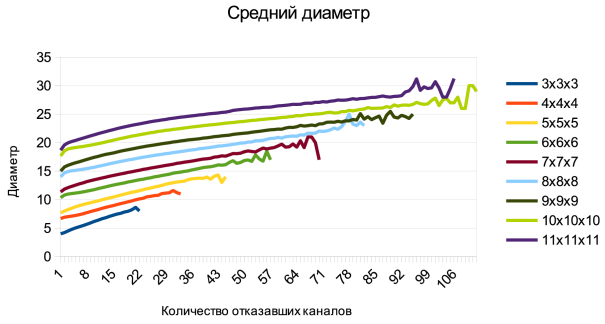

Рис. 8. Усредненное значение диаметра по итерациям метода Монте-Карло

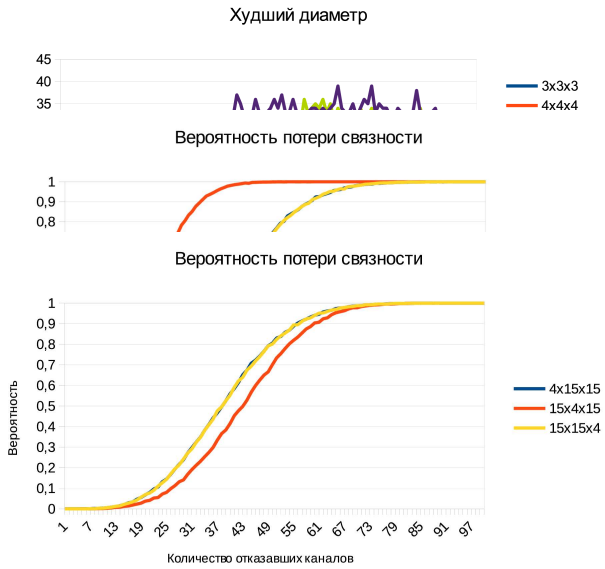

Рис. 11. Вероятность потери связности для «длинных» торов

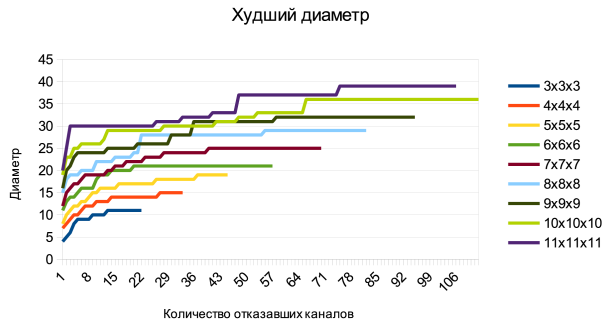

Рис. 10. Монотонная версия графика худшего значения диаметра

Рис. 12. Средний диаметр для «длинных» торов

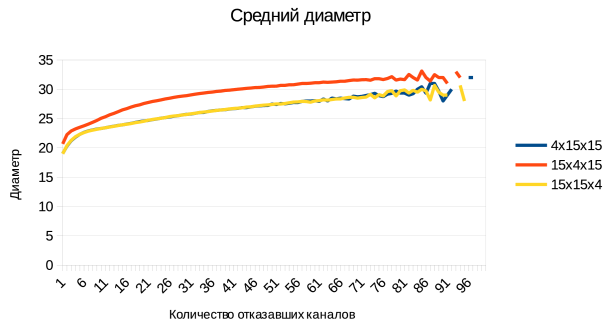

Рис. 13 Вероятность потери связности для «плоских» торов

Рис. 14. Средний диаметр для «плоских» торов

На Рис. 8-10 показана соответствующая картина для диаметра: усредненное значение и худшее. Артефакты на правых концах графиков объясняются размером выборки: при таком большом количестве отказов каналов связи появляется большая вероятность потери связности детерминированной сети, поэтому усреднение здесь происходит по очень малой выборке (10-20), откуда следует неустойчивость получаемых значений. График на Рис. 9 и вовсе полон артефактов, это объясняется тем, что на нем изображен худший полученный диаметр среди итераций метода Монте-Карло, и в силу случайности при большем количестве отказавших каналов связи иногда не находилось их комбинации, дающей больший диаметр. Заметим, что зависимость худшего диаметра от количества отказов должна быть монотонной: при добавлении еще одного отказавшего канала связи диаметр сети может только увеличиться или остаться неизменным. Этот факт использован при получении графика на Рис. 10 на нем все графики монотонны, но построен он из тех же исходных данных.

На Рис. 11-14 представлено сравнение исследуемых показателей (вероятность потери связности и средний диаметр) для несимметричных торов. Так называемых «длинных» торов, где одни из размерностей больше двух других и «плоских» торов, где одна из размерностей меньше. В качестве «длинных» торов рассмотрены торы 60x5x5, 5x60x5 и 5x5x60. Из рисунка видно, что показатели для торов 60x5x5 и 5x5x60 совпадают, это следует из антисимметричности существования пути: если из вершины u в вершину v существует путь, удовлетворяющий правилу порядка направлений, то из v в u существует путь, удовлетворяющий обратному правилу порядка направлений, таким образом при транспонировании тора мы получаем аналогичный результат. Аналогичное рассуждение верно и для плоских торов 4x15x15 и 15x15x4.

При сравнении вероятности потери связности и диаметра для торов 60x5x5 и 5x60x5 выявляется следующая закономерность: диаметр для тора 5x60x5 меньше при любом количестве отказов, причем начиная от 8 отказавших каналов в среднем на 30, далее с увеличением количества отказов разница уменьшается до 10. Заметим, что при отсутствии отказов диаметр тора 60x5x5 равен 34. Это различие объясняется следующим образом. Пусть задан порядок направлений +X, +Y, +Z, -X, -Y, -Z. Пусть при этом канал связи в направлении X вышел из строя. Рассмотрим узел сети, из которого выходит соответствующий отказавших канал связи. Пусть его координаты (x, y, z) . Пути из этого узла можно начинать только в направлениях +Y, +Z, а движение вдоль размерности X осуществлять только в направлении -X. Отсюда следует, что диаметр сети увеличивается до 60, т. к. путь в узлы вида $(x+1, \cdot, \cdot)$ будет включать движение в направлении -X. На графике Рис. 5 такого резкого увеличения не наблюдается, т. к. на нем представлено усредненное значение диаметра, а вероятность появления отказа именно в направлении X равна 1/3 для одного отказа или $1-(2/3)^n$ для n отказов.

Вероятность потери связности для тора 5x60x5, напротив, оказывается больше при любом количестве отказов, чем для тора 60x5x5 (и 5x5x60), т. е. по этому показателю тор 60x5x5 предпочтителен.

Аналогично, для «плоских» торов вида 4x15x15 наблюдается больший диаметр в случае 15x4x15, но меньшая вероятность потери связности.

Заключение

В данной работе построены теоретические оценки задержки и пропускной способности сети с топологией «многомерный тор» при наличии отказавших каналов связи. Данные оценки могут использоваться при определении эффективности коммуникационного алгоритма на сети с отказами, например, обмен данными по схеме «все всем» ограничен сверху введенной в этой статье оценкой, таким образом, эффективность этого алгоритма можно определять как отношение времени, полученного из оценки к реальному времени выполнения алгоритма.

Из проведенного исследования поведения этих оценок с ростом числа отказавших каналов связи можно сделать следующие выводы:

1. Средняя длина пути при наличии неисправных каналов увеличивается на 7-10% за сто отказов, таким образом в среднем задержка передачи между парой узлов будет увеличена на 7-10%.
2. Того же нельзя сказать о диаметре, т. е. максимальном расстоянии между парой вершин: с большой вероятностью будет находиться пара вершин, задержка на передачу между которыми будет увеличена практически в два раза по сравнению с сетью без отказов, т. е. следует ожидать соответствующего падения производительности при выполнении синхронизационных примитивов вроде барьеров.
3. Цена пропорционального разбиения при приемлемых вероятностях потери связности уменьшается незначительно: уже для тора размером 5x5x5 получаем падение не ниже 80% исходного на границе потери связности детерминированной маршрутизации, таким образом, по оценкам данной статьи пропускная способность сети падает не более, чем на 20%, прежде чем сеть потеряет связность.
4. При выборе порядка направлений в торе выигрыш в диаметре сети с отказами дают торы, где наибольшая размерность тора является средней, но при этом вероятность потери связности будет больше.
5. Так как оценка через цену пропорционального разбиения является оценкой сверху, для исследования производительности самой сети необходимо проводить имитационное моделирование при наличии отказавших каналов связи. Работа в данном направлении ведется.

В дальнейшем планируется продолжать исследования отказоустойчивости сети «Ангара». В частности, по результатам данного исследования появляется возможность постановки задачи оптимальной отказоустойчивости, так же планируется исследование отказоустойчивости адаптивной маршрутизации и подсети коллективных операций. Планируется провести имитационное моделирование поведения сети при наличии отказов каналов связи и привести более точную оценку падения производительности.

ЛИТЕРАТУРА:

1. Жабин Иван, Макагон Дмитрий и Симонов Алексей: Кристалл для Ангары [Журнал] // Суперкомпьютеры. - 2013 г.. - Т. зима-2013. - стр. 46-49.
2. Корж А. А. [и др.] Отечественная коммуникационная сеть 3D-тор с поддержкой глобально адресуемой памяти для суперкомпьютеров транспетафлопсного уровня производительности [Конференция] // Параллельные вычислительныетехнологии (ПаВТ’2010): Труды международной конференции (Уфа, 29 марта - 2 апреля 2010 г.).. - Челябинск : Издательский центр ЮУрГУ, 2010. - стр. 227-237.
3. Пожилов Илья, Алгоритм решения задачи связности в условиях наличия неисправных каналов связи для детерминированной маршрутизации, основанной на правиле порядка направлений, в сети с топологией «многомерный тор» // Труды конференции ПАВТ-2014 — 2014 г.
4. C. D. Thompson (2003). *A complexity theory for VLSI*. PhD Thesis, Technical Report CMU-CS-80-140, Carnegie-Mellon University.
5. Hagen, L.; Kahng, A.B., New spectral methods for ratio cut partitioning and clustering. // Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on , vol.11, no.9, pp.1074,1085, Sep 1992