Труды МАИ. Выпуск № 109 DOI: 10.34759/trd-2019-109-22

УДК 004.72

Имитационная модель высокоскоростной коммуникационной сети Ангара с топологией kD-тор

Симонов А.С.

Научно-исследовательский центр электронной вычислительной техники,

Варшавское шоссе, 125, Москва, 117587, Россия

e-mail: simonov@nicevt

Статья поступила 20.11.2019

Аннотация

В статье описана реализация и опыт применения имитационной модели маршрутизатора высокоскоростной коммуникационной сети Ангара, предназначенной для построения на её основе многопроцессорных вычислительных систем с высокой масштабируемостью производительности.

Проведённые с использованием имитационной модели исследования позволили отработать основные решения по микроархитектуре маршрутизатора, алгоритмы работы его составных частей, формат пакета, основные операции, алгоритмы маршрутизации и арбитража, провести оптимизацию архитектурных параметров маршрутизатора, в том числе базовых соотношений по пропускным способностям, связности сети, размерам буферов памяти, сбалансировать пропускные способности составных частей маршрутизатора и осуществить верификацию заказной СБИС. По результатам имитационного моделирования был сформирован окончательный технический облик маршрутизатора, реализованный в заказной СБИС и сетевом оборудовании Ангара на её основе.

Полученный опыт может быть использован при создании перспективных образцов сетевого оборудования.

Ключевые слова: имитационное моделирование, коммуникационная сеть.

Введение

Разработка современного высокоскоростного сетевого оборудования представляет собой сложный многоэтапный процесс, одним из важнейших элементов которого является разработка заказных СБИС, реализующих аппаратный функционал. Высокая сложность таких СБИС, существенная стоимость исправления возможных ошибок, значительные ограничения возможностей встроенных средств диагностики свидетельствуют о необходимости полноценной отработки с использованием методов компьютерного моделирования всех технических решений разрабатываемых СБИС до изготовления опытных образцов.

В АО «НИЦЭВТ» осуществляется разработка высокоскоростной коммуникационной сети (ВКС) Ангара, позволяющей обеспечить высокую масштабируемость производительности приложений. Одной из важнейших задач, решаемых в процессе проектирования сетевого оборудования Ангара [1-3], является отработка архитектуры его ключевого элемента – заказной СБИС маршрутизатора, в ходе которой необходимо выполнить следующее:

– проверить корректность работы коммуникационной сети в целом, взаимодействия структурных элементов между собой и с процессором вычислительного узла, корректность работы маршрутизатора узла в составе сети;

– выявление в реализации «узких мест», ограничивающих эксплуатационные характеристики маршрутизатора узла;

– осуществить отработку алгоритмов маршрутизации;

– оптимизировать параметры и функциональные возможности маршрутизатора узла, в том числе соотношения пропускных способностей составных частей, осуществить подбор оптимальных размеров буферов и т.д.;

– отработать и выбрать оптимальные алгоритмы арбитража, позволяющие обеспечить справедливое распределение ресурсов;

– осуществить оценку и сопоставление с техническим заданием достигаемых эксплуатационных характеристик заказной СБИС и сетевого оборудования на её основе с использованием различных простых и комплексных тестов;

– отработать алгоритмы обеспечения отказоустойчивости.

Для отработки архитектуры заказных СБИС в настоящее время, как правило, применяются инструментальные средства, основанные на различных методах поведенческого (имитационного) моделирования (simulation modeling, behavioral simulation), выполняемого на уровне структурных схем [4, 5]. Основным недостатком существующих инструментальных средств является относительно низкая скорость моделирования, далеко не всегда позволяющая за разумное время провести полноценную отработку архитектуры сложной заказной СБИС [6].

Возможно применение для этих целей различных ускорителей на основе ПЛИС [7, 8], однако при этом имеются существенные ограничения [9]. В связи с этим для архитектурного проектирования и высокой скорости моделирования разрабатываются программные симуляторы [10-15].

Для решения данной задачи в АО «НИЦЭВТ» была разработана параллельная многофункциональная имитационная модель маршрутизатора ВКС с топологией kD-тор [16], обеспечивающая реализацию алгоритмов работы всех его структурных элементов в транзакционном режиме и высокую скорость моделирования от нескольких сотен до тысяч тактов в секунду. Созданная имитационная модель явилась мощным инструментом для проведения исследований и позволила осуществить разработку и выбор такой комбинации технических решений по архитектуре и алгоритмам работы заказной СБИС маршрутизатора, при которой обеспечивается высокая масштабируемость производительности многопроцессорных вычислительных систем при решении прикладных задач.

Общие сведения

Имитационная модель реализована в виде моделирующего комплекса, обеспечивающего возможность последовательного моделирования различных вариантов конфигурации сети. Общие характеристики модели:

– число измерений тора – от 1 до 9;

– число узлов сети в каждом измерении – от 1 до 4 096;

– количество реализуемых подсетей (виртуальных каналов) на одном канале связи – от 1 до 16;

– глубина очередей виртуальных каналов – от 1 до 65 535 единиц передаваемых данных (флитов);

– задержка в канале связи – от 1 до 65 535 модельных тактов.

Общая структура имитационной модели представлена на рис. 1.

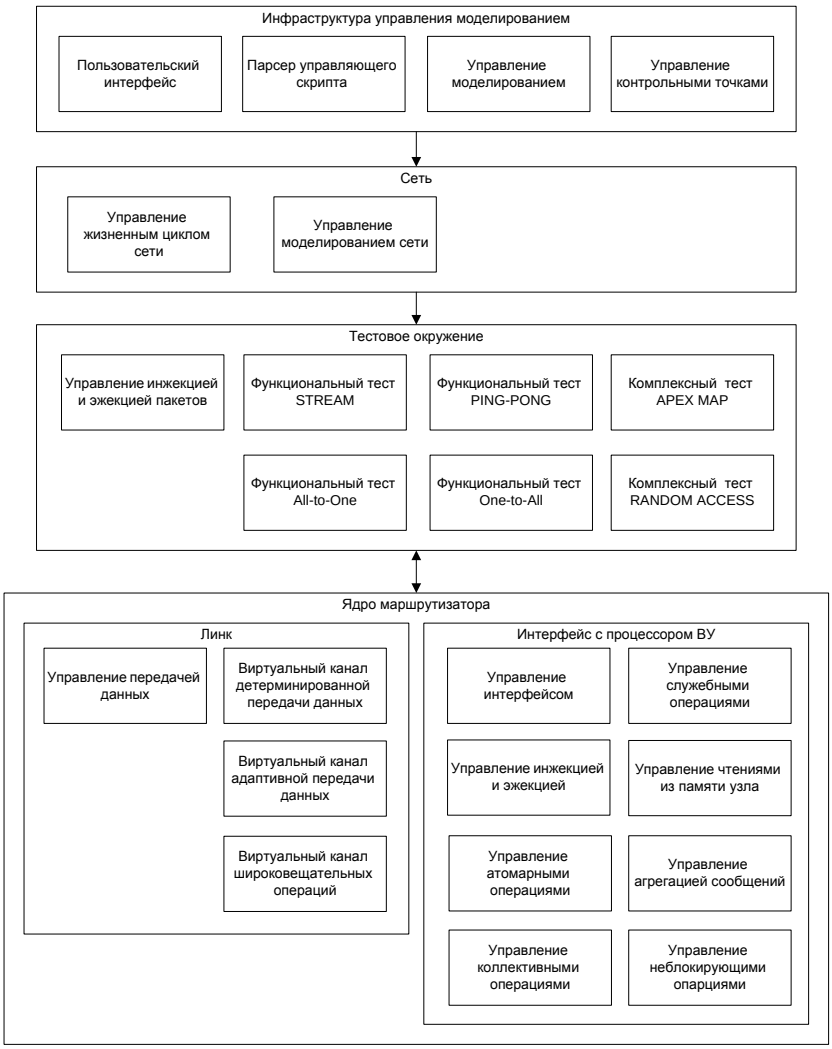


Рис. 1. Общая структура имитационной модели

Управление работой имитационной модели осуществляется инфраструктурой управления моделированием.

Объектом моделирования является сеть, создаваемая на каждом варианте конфигурации и включающая множество экземпляров связанных между собой в kD-тор сборок, имитирующих работу вычислительных узлов и включающих ядро маршрутизатора и тестовое окружение. В качестве примера на рис. 2 приведена структура сети с топологией 2D-тор размерностью 3x4.

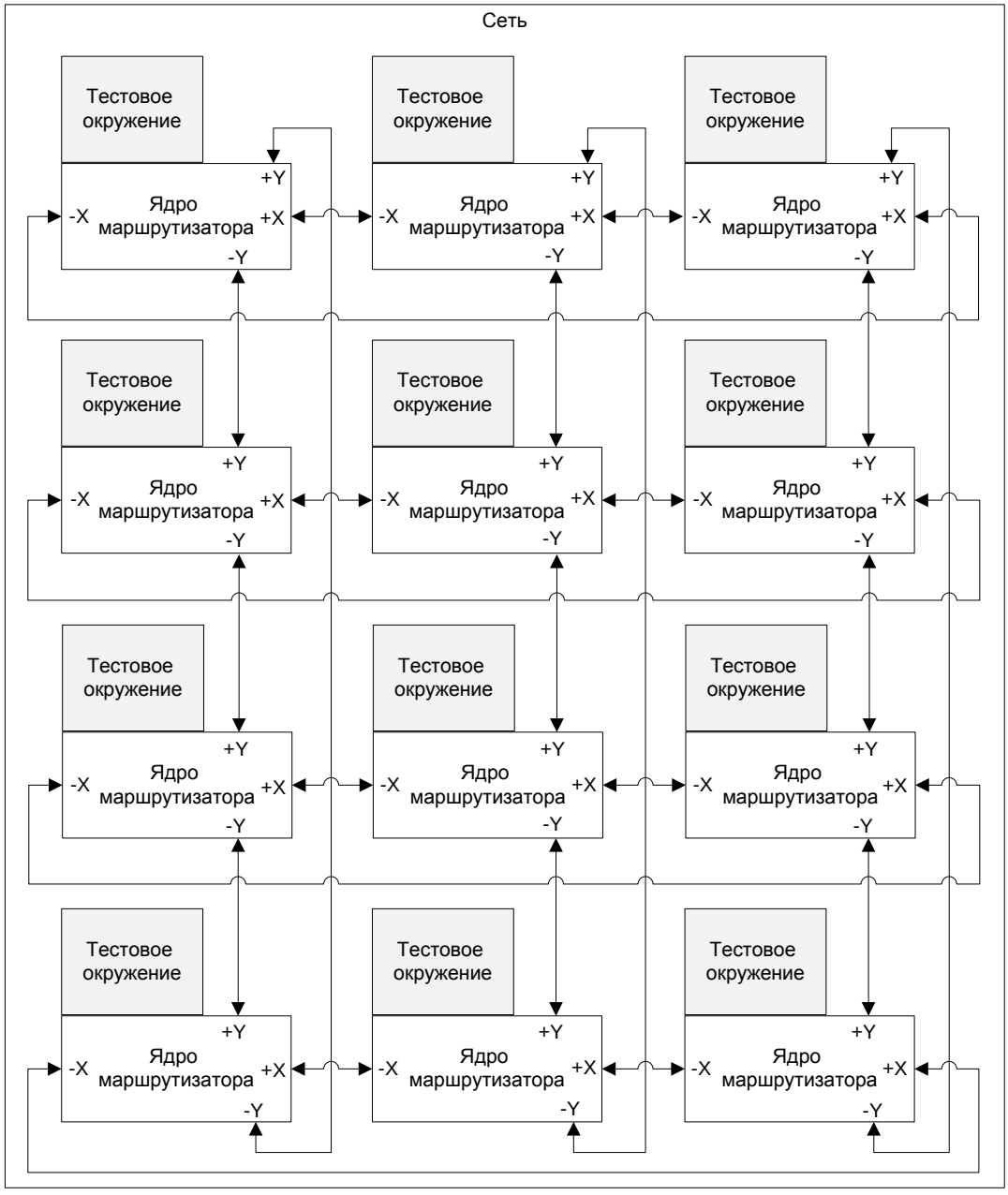


Рис. 2. Пример модельного варианта сети с топологией 2D-тор

В каждой сборке тестовое окружение имитирует функциональную нагрузку со стороны центрального процессора вычислительного узла, а ядро маршрутизатора осуществляет передачу пакетов между узлами. Обобщённый алгоритм работы имитационной модели приведен на рис. 3.

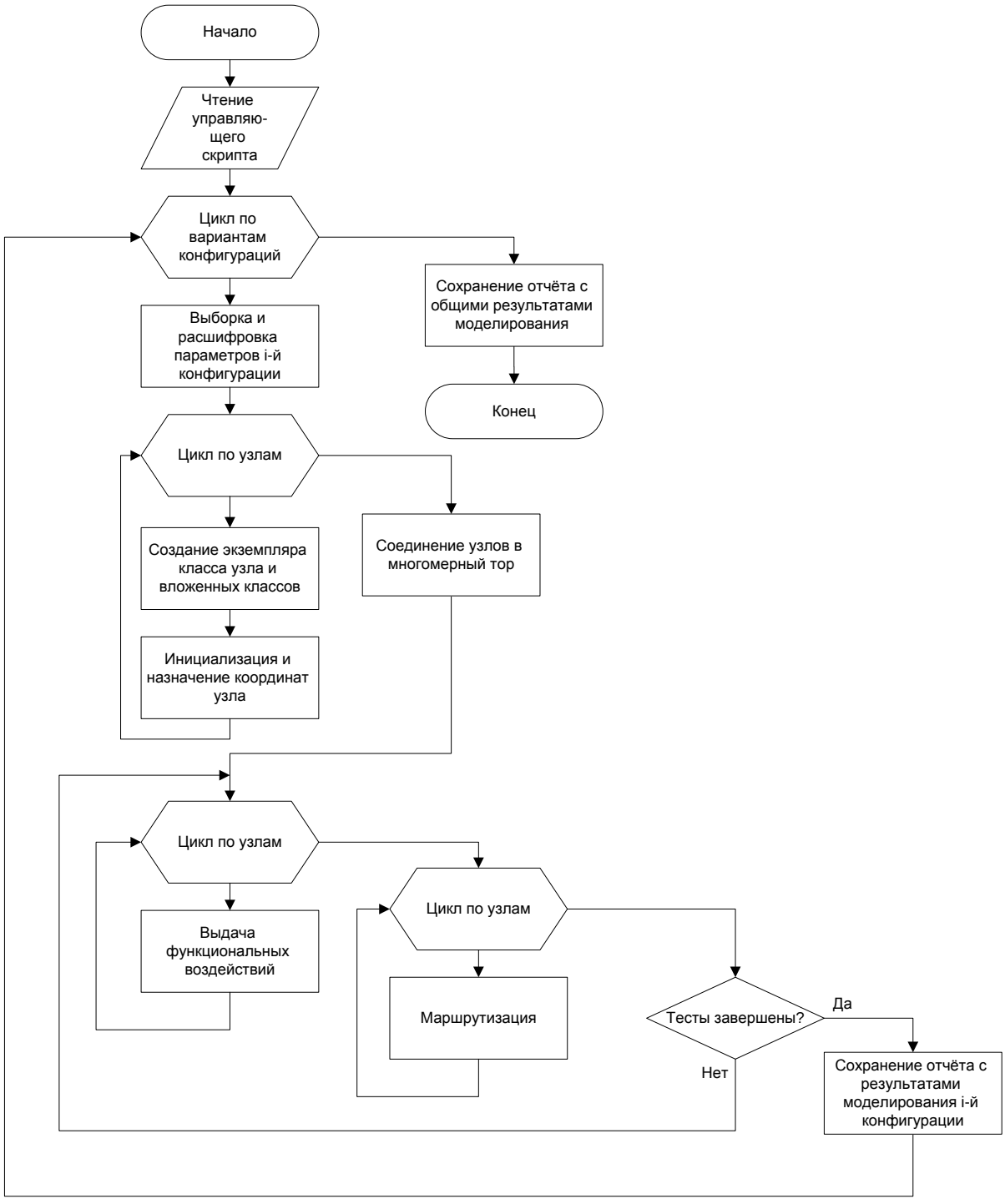


Рис. 3. Обобщённый алгоритм работы имитационной модели

После выборки и расшифровки параметров конфигурации осуществляется создание сети, в ходе которого для каждого узла создаётся свой экземпляр сборки из классов ядра маршрутизатора и тестового окружения, осуществляется их инициализация, включающая задание физических и логических адресов, адресов соседних узлов, формирование таблицы маршрутизации ядра маршрутизатора, а для тестового окружения задаются начальные условия для функциональных или комплексных тестов. После создания и инициализации всех узлов они соединяются между собой в структуру kD-тор.

Далее запускается цикл моделирования, включающий два вложенных цикла по узлам. В первом цикле для каждого узла вызывается метод, обслуживающий тестовое окружение. В нём осуществляется формирование пакетов, их приём и обработка. Во втором цикле для каждого узла вызывается метод, обслуживающий ядро маршрутизатора, в котором осуществляется имитация одного модельного такта работы для каждого функционального блока маршрутизатора.

После прохождения циклов по всем узлам осуществляется контроль завершения теста, при котором проверяется, что все отправленные пакеты достигли адресатов, а функциональные тесты завершили своё выполнение. Моделирование варианта завершается, когда все узлы перейдут в состояние завершения теста, после чего осуществляется сохранение отчёта с результатами моделирования конфигурации.

На каждом варианте моделирования создаётся сеть, параметры которой определяются значениями соответствующих полей управляющего скрипта.

Создаваемая на каждом варианте моделирования сеть состоит из связанных между собой узлов, фрагмент такой сети представлен на рис. 4.

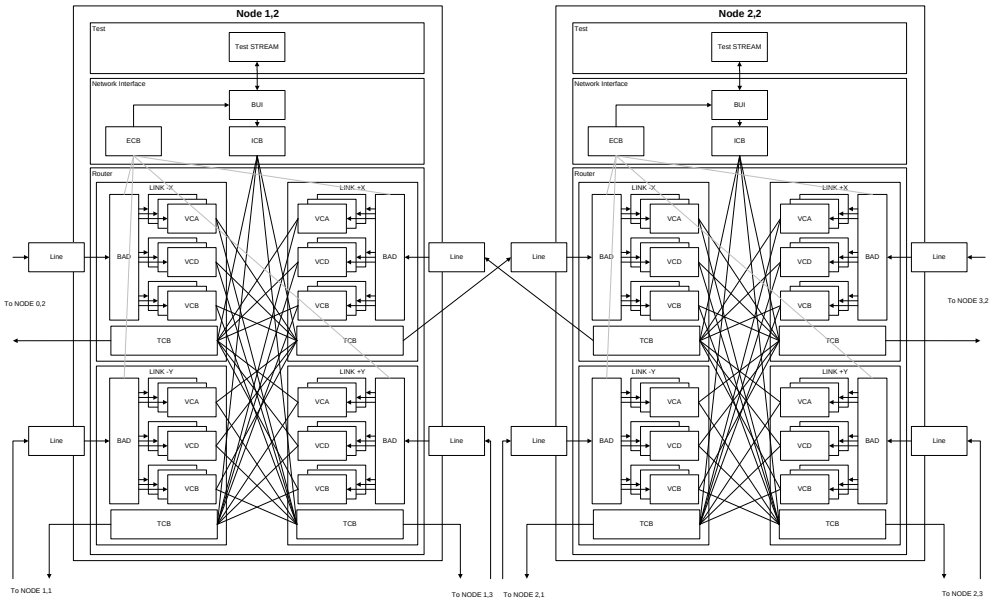


Рис. 4. Фрагмент сети

На приведённом рисунке представлены два узла с физическими координатами 1,2 и 2,2 в топологии 2D-тор, линиями и стрелками обозначены указатели, используемые для связи между собой экземпляров классов, соответствующих различным блокам маршрутизатора. Часть функциональных блоков и связей на рисунке условно не показана.

Для передачи пакетов между узлами был выбран метод виртуальной сквозной передачи VCT (Virtual cut-through) [17], позволяющий обеспечить минимальную латентность. При этом каждый передаваемый в имитационной модели пакет данных представляет собой отдельный экземпляр соответствующего класса, что позволяет не только имитировать передачу, но и включить в его состав методы, обеспечивающие необходимый уровень информативности при проведении диагностики и оценочного тестирования.

В составе модели реализованы три вида тестов:

– тесты оценки работоспособности;

– нагрузочные тесты;

– комплексные тесты.

Тесты оценки работоспособности предназначены для проверки корректности работы модели. Нагрузочные тесты предназначены для контроля основных характеристик ВКС – пропускной способности и коммуникационной задержки, и включают представительную выборку тестов All2One, One2All, Random Uniform, Ping Pong, Stream. Комплексные тесты предназначены для получения предварительной информации о производительности вычислительной системы при выполнении приложений. В состав данной группы входят APEX MAP [18] и имитаторы трафика тестов Linpack [19] и HPCG [20].

По результатам моделирования создаётся отчёт, включающий полученные в ходе моделирования характеристики для каждого варианта. Пример отчёта приведен на рис. 5.

Datafile for complete results (Final time 15:50)
Duration of computation - 722:29: 2

Test All2One. Results: msg size (bytes), average latency (ns), min latency (ns), max latency (ns), BW on recv node (MBits/s), message rate (MT/s)

Var #:	0	1	2	3	4	5	6	7
Length, bytes	32,00	128,00	512,00	4096,00	16384,00	65536,00	262144,00	1048576,00
Average Latency, ns	15537,56	44592,15	183593,03	948253,31	989114,13	1127951,38	1247732,25	1294388,88
Min latency, ns	1066,00	1090,00	1090,00	1090,00	1090,00	1090,00	1090,00	1090,00
Max Latency, ns	31324,00	96402,00	431896,00	3269518,00	13363460,00	28681798,00	62292784,00	133642320,00
Average BW, MBit/s	30785,30	30182,45	26394,83	27184,60	26602,69	24834,44	22968,79	21623,05
Message Rate, MT/s	106666,67	29629,63	7054,67	75,14	2,97	0,65	0,16	0,05

Рис. 5. Пример отчёта по результатам моделирования

Практическое применение имитационной модели

Одним из важнейших применений разработанной имитационной модели является проверка корректности разработанных принципов работы маршрутизатора, в ходе которой решается следующий набор задач:

– проверка взаимодействия структурных элементов маршрутизатора между собой;

– проверка взаимодействия структурных элементов маршрутизатора с процессором вычислительного узла;

– проверка корректности работы маршрутизатора в целом в составе коммуникационной сети.

Перечисленные проверки осуществляются путём встраивания в имитационную модель набора ASSERT'ов и специальных тестовых программных модулей-чекеров, расположенных в интерфейсах между составными частями маршрутизатора и между экземплярами маршрутизаторов в составе модели коммуникационной сети. Перечисленные средства осуществляют постоянный контроль выполнения общих правил работы коммуникационной сети:

– все отправленные пакеты должны быть доставлены по адресу получателя;

– пакеты не должны теряться;

– пакеты не должны дублироваться;

– пакеты не должны искажаться.

Встроенные в имитационную модель ASSERT'ы позволяют выявить явные нарушения перечисленных правил. Работа программных модулей-чекеров основана

на подсчёте числа и контроле SEQ-номеров пакетов, проверке контрольных сумм, хранении дополнительной адресной информации в поле данных пакета и т.д.

Взаимодействие структурных элементов маршрутизатора между собой в составе имитационной модели осуществляется с использованием транзакций. На каждом модельном такте между структурными элементами, например, между виртуальным каналом одного канала связи и блоком передачи данных другого канала связи, может передаваться одна транзакция, информационно соответствующая передаче одного флита данных.

Проверка корректности взаимодействия структурных элементов маршрутизатора между собой, в том числе информационной достаточности транзакций, осуществляется на этапе отладки имитационной модели.

Отработка взаимодействия ядра маршрутизатора с процессором вычислительного узла также осуществляется с использованием транзакций, размер и информационное наполнение которых соответствуют транзакциям внутреннего TRN-интерфейса СФ-блока контроллера интерфейса PCI Express.

Отработка взаимодействия ядер маршрутизаторов, соединенных между собой в сеть посредством соединения по каналам связи в пространственную структуру kD-тор, осуществляется с использованием транзакций, размер и информационное наполнение которых соответствуют транзакциям внутреннего интерфейса СФ-блока Aurora, предназначенного для логического объединения нескольких блоков SerDes в единый канал передачи данных.

Для выполнения проверки корректности работы маршрутизатора в целом в составе коммуникационной сети в состав имитационной модели также встроено тестовое окружение, представляющее собой набор тестов проверки работоспособности. В его состав входят как тривиальные тесты, позволяющие проверить корректность пересылок типа «точка-точка», так и комплексные тесты, направленные на проверку корректности работы маршрутизатора в условиях предельных нагрузок на сеть.

Выявление в реализации (dataflow) «узких мест», ограничивающих эксплуатационные характеристики маршрутизатора, а также выбор оптимальных значений микроархитектурных параметров, осуществляются в процессе имитационного моделирования при выполнении различных тестов путём наблюдения интенсивности трафика в контрольных точках ядра маршрутизатора.

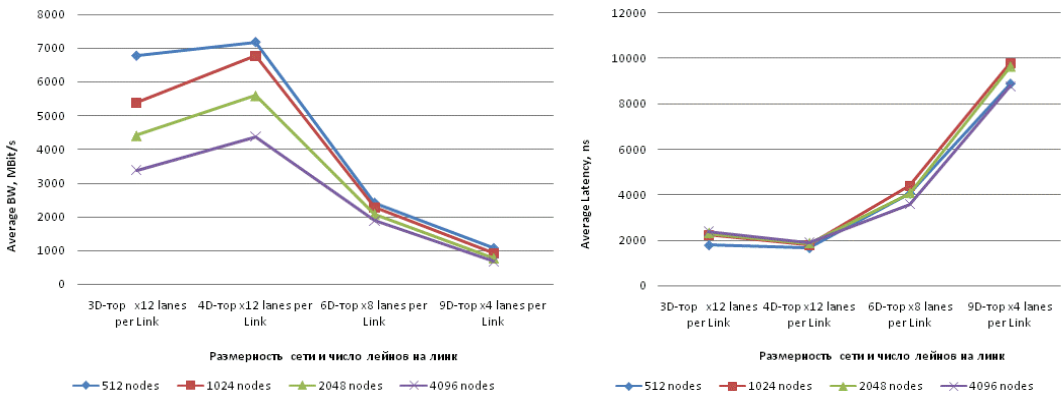
Одной из первых задач оптимизации параметров маршрутизатора был выбор оптимального соотношения между агрегатной пропускной способностью каналов связи маршрутизатора и пропускной способностью интерфейса с процессором (B_L/B_P).

В ходе моделирования варьировалась агрегатная пропускная способность всех каналов связи маршрутизатора BL при передаче пакетов различной длины. Для тестов взяты две вычислительные системы размером 512 и 4 096 узлов. Результаты имитационного моделирования позволили определить, что для вычислительных систем небольшого размера (512 узлов), начиная с $B_L/B_P>4.8$, наблюдается прекращение масштабирования производительности. Аналогичный эффект для

вычислительных систем среднего размера (4096 узлов) наблюдается, начиная с $B_L/B_P>6.4$. Таким образом, для выбранного интерфейса PCI Express Gen2 16x с пиковой пропускной способностью 80 ГБит/с агрегатная пропускная способность каналов связи должна быть не менее 512 ГБит/с.

Другой важной задачей является выбор оптимального соотношения между размерностью сети и шириной каналов связи. Важно ещё на этапе архитектурного проектирования правильно ответить на вопрос, что лучше: увеличить размерность сети, но при этом получить более «тонкие» каналы связи, или снизить её размерность, но при этом получить более «толстые» каналы связи.

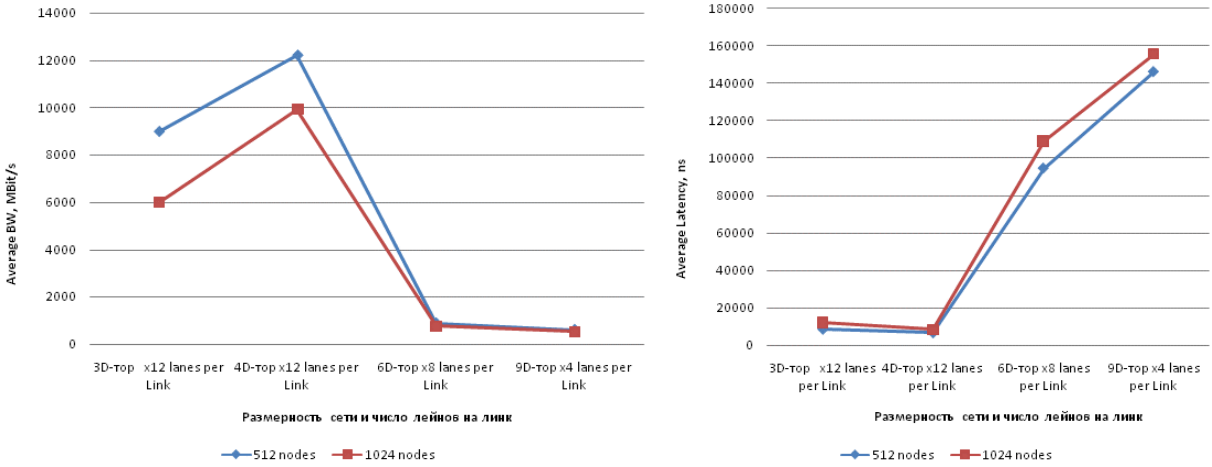
С учётом технических ограничений для имитационного моделирования были подготовлены различные варианты размерностей топологии сети и ширины каналов связи. Для определения оптимального соотношения размерности топологии сети и ширины каналов связи были использованы тесты RU, One2All и All2One. Результаты моделирования приведены на рис. 6-8.



«а» «б»

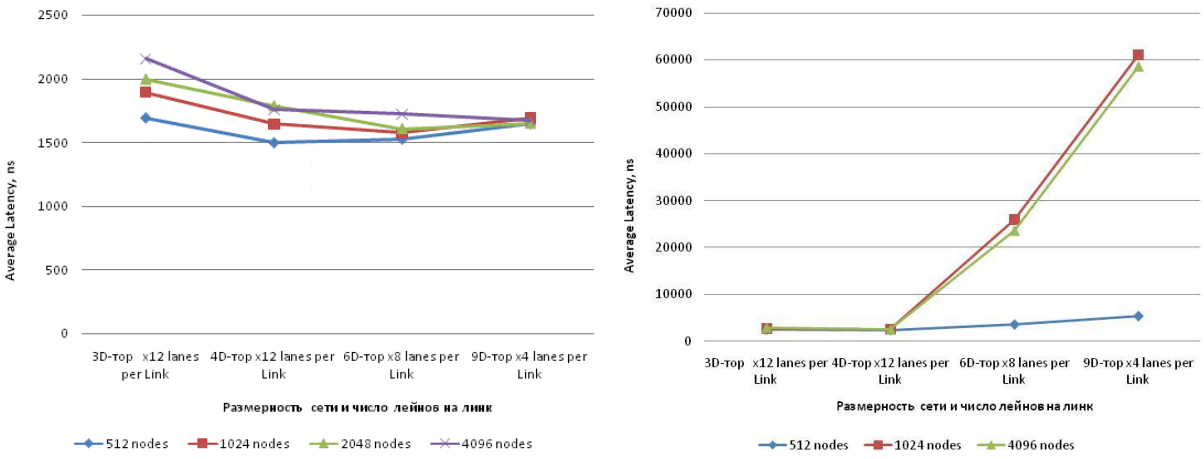
Рис. 6. Зависимость пропускной способности («а») и коммуникационной задержки («б») от размерности топологии сети (сообщения 32 байта, тест RU)

 DOI: 10.34759/trd-2019-109-22



«а» «б»

Рис. 7. Зависимость пропускной способности («а») и коммуникационной задержки («б») от размерности топологии сети (сообщения 4 Кбайта, тест RU)



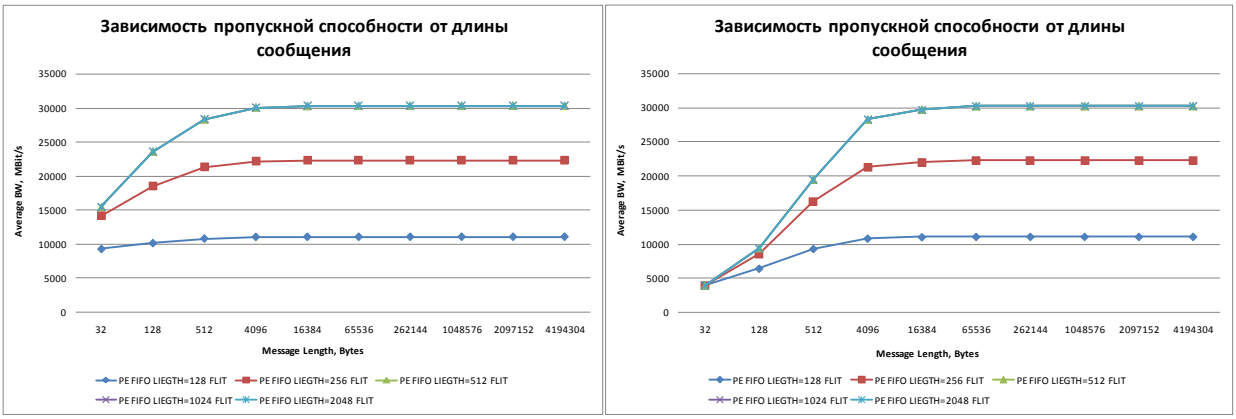
«а» «б»

Рис. 8. Зависимость коммуникационной задержки от размерности топологии сети (сообщения 32 байта («а») и 4 Кбайт («б»), тест One2All)

По результатам имитационного моделирования был сделан вывод о том, что при существующих технических ограничениях оптимальной является топология 4D-тор с каналами связи шириной 12x.

Одним из важных ресурсов при проектировании заказной СБИС является встроенная статическая оперативная память, используемая для организации различных FIFO-буферов, в том числе буферов виртуальных каналов и интерфейса с процессором, буферов отображения BAR интерфейса PCI Express и т.д. Для определения оптимального размера входных FIFO-буферов виртуальных каналов на имитационной модели был проведен запуск представительной выборки тестов All2One, One2All, RU, Ping Pong, Stream для нескольких наиболее интересных с практической точки зрений конфигураций многопроцессорных вычислительных систем.

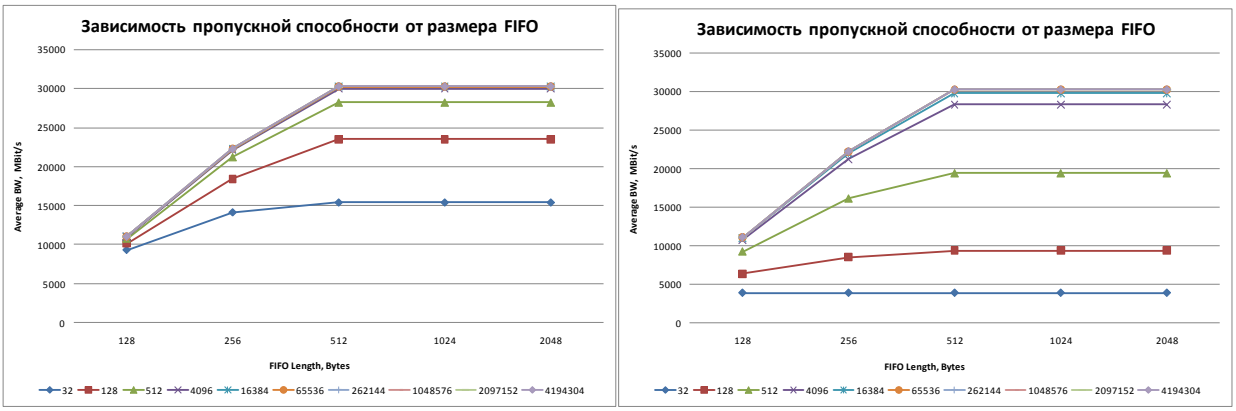
Результаты моделирования на тестах Stream для соседних узлов и для расстояния между узлами 4 шага приведены на рис. 9-12.



«а» «б»

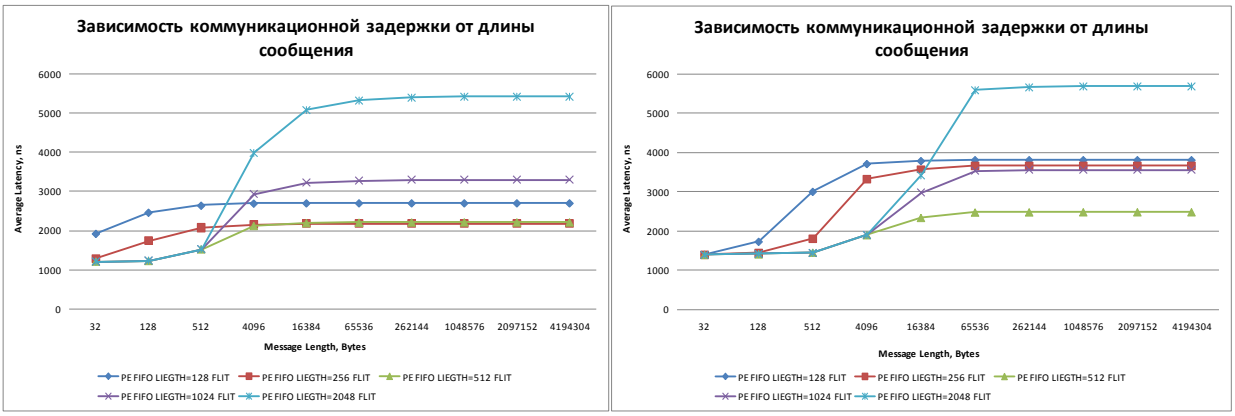
Рис. 9. Зависимость пропускной способности от размера сообщения при выполнении теста Stream, Мбит/с («а» – 1 шаг, «б» – 4 шага)

 DOI: 10.34759/trd-2019-109-22



«а» «б»

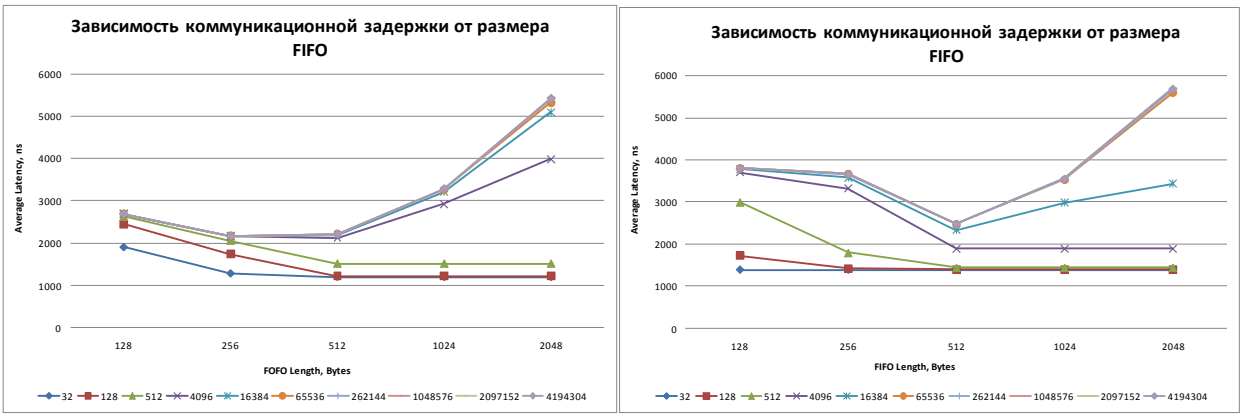
Рис. 10. Зависимость пропускной способности от размера FIFO при выполнении теста Stream, Мбит/с («а» – 1 шаг, «б» – 4 шага)



«а» «б»

Рис. 11. Зависимость коммуникационной задержки от размера сообщения при выполнении теста Stream, нс («а» – 1 шаг, «б» – 4 шага)

 DOI: 10.34759/trd-2019-109-22



«а» «б»

Рис. 12. Зависимость коммуникационной задержки от размера FIFO при выполнении теста Stream, Мбит/с («а» – 1 шаг, «б» – 4 шага)

Как видно из приведенных графиков, наблюдается существенная зависимость пропускной способности при выполнении теста Stream от числа шагов. Чем больше расстояние между узлами, тем более существенное влияние на пропускную способность оказывает размер FIFO. При этом на графиках отчётливо наблюдается точка оптимума при размере FIFO 512 флит. Именно с этой точки рост пропускной способности практически прекращается.

Выводы

Проведённые с использованием имитационной модели исследования позволили отработать основные архитектурные и алгоритмические решения, провести оптимизацию архитектурных параметров и осуществить верификацию заказной СБИС маршрутизатора высокоскоростной коммуникационной сети Ангара с топологией kD-тор. По результатам имитационного моделирования был

сформирован окончательный технический облик маршрутизатора, реализованный в заказной СБИС и сетевом оборудовании Ангара на её основе.

Библиографический список

1. Слуцкин А.И., Симонов А.С., Жабин И.А., Макагон Д.В., Сыромятников Е.Л. Разработка межузловой коммуникационной сети ЕС8430 «Ангара» для перспективных российских суперкомпьютеров // Успехи современной радиоэлектроники. 2012. № 1. С. 6 - 10.

2. Жабин И.А., Макагон Д.В., Поляков Д.А., Симонов А.С., Сыромятников Е.Л., Щербак А.Н. Первое поколение высокоскоростной коммуникационной сети «Ангара» // Наукоёмкие технологии. 2014. Т. 5. № 1. С. 21 - 27.

3. Симонов А.С., Жабин И.А., Куштанов Е.Р., Макагон Д.В., Семенов А.С., Щербак А.Н. Высокоскоростная сеть Ангара: архитектура и результаты применения // Вопросы кибербезопасности. 2019. № 4 (32). С. 46 - 53.

4. Ravi S., Raghunathan A., Chakradhar S.T. U.S. Patent No. 7,134,100. Washington, DC: U.S. Patent and Trademark Office. 2006, available at: https://patentimages.storage.googleapis.com/2b/1d/3c/7bdd5e86bdac9b/US7134100.pdf

5. Mathur A., Wang Q. Power reduction techniques and flows at RTL and system level // Processing of 2009 22nd International Conference on VLSI Design, IEEE, 2009, pp. 28 – 29, doi: 10.1109/VLSI.Design.2009.113

6. Байда Ю.В., Бутузов А.В., Ефимов А.Г., Цветков М.С. Методология перехода от программной потактовой модели микропроцессора к аппаратному симулятору на базе программируемой логики // Труды МФТИ. 2012. Т. 4. № 3. С. 114 - 121.

7. Матафонов Д.Е. Создание и отработка маршрутизатора в стандарте SpaceWire на отечественной программируемой логической интегральной схеме // Труды МАИ. 2018. № 103. URL: http://trudymai.ru/published.php?ID=100780

8. Kim D., Celio C., Biancolin D., Bachrach J., Asanovic K. Evaluation of RISC-V RTL with FPGA-accelerated simulation // First Workshop on Computer Architecture Research with RISC-V, 2017, available at: https://people.eecs.berkeley.edu/~biancolin/papers/carrv17.pdf

9. Pellauer M., Vijayaraghavan M., Adler M. et al. A-Port networks: preserving the timed behavior of synchronous systems for modeling on FPGAs // ACM Transactions on reconfigurable technology and systems, 2009, vol. 2, no. 3, pp. 1 - 26.

10.Zheng G., Kakulapati G., Kalé L.V. Bigsim: A parallel simulator for performance prediction of extremely large parallel machines // 18th International Parallel and Distributed Processing Symposium, 2004, Santa Fe, NM, pp. 78.

11.Paul D., Nakhla N.M., Achar R., Nakhla M.S. Parallel simulation of massively coupled interconnect networks // IEEE Transactions on Advanced Packaging, 2009, vol. 33, no. 1, pp. 115 - 127.

12.Mubarak M., Carothers C.D., Ross R.B., Carns P. Enabling parallel simulation of large-scale HPC network systems // IEEE Transactions on Parallel and Distributed Systems, 2016, vol. 28, no. 1, pp. 87 - 100.

13.Cerutti I., Corvera J.A., Dumlao S.M., Reyes R., Castoldi P., Andriolli N. Simulation and FPGA-based implementation of iterative parallel schedulers for optical interconnection networks // IEEE/OSA Journal of Optical Communications and Networking, 2017, vol. 9, no. 4, pp. 76 - 87.

14.Jain N., Bhatele A., White S., Gamblin T., Kale L.V. Evaluating HPC networks via simulation of parallel workloads // Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2016, pp. 154 - 165.

15.Fujimoto R. Parallel and distributed simulation // Winter Simulation Conference (WSC), IEEE, 2015, pp. 45 - 59.

16.Симонов А.С. Программа имитационного моделирования работы коммуникационной сети «Ангара» // Свидетельство о государственной регистрации программы для ЭВМ, № 2014662998, 12.12.2014.

17.Kermani P., Kleinrock L. Virtual cut-through: A new computer communication switching technique // Computer networks, 1979, vol. 3, pp. 267 - 286.

18.Weinberg V. The Apex-MAP Benchmark. PRACE Workshop «New Languages & Future Technology Prototypes», LRZ, 1-2, March 2010, available at: http://www.prace-ri.eu/IMG/pdf/17_apexmap_vw.pdf

19.Dongarra J., Luszczek P., Petitet A. The LINPACK benchmark: past, present and future // Concurrency and Computation: practice and experience, 2003, vol. 15, no. 9, pp. 803 - 820.

20.Dongarra J., Heroux A., Luszczek P. High-performance conjugate-gradient benchmark: A new metric for ranking high-performance computing systems // The International Journal of High Performance Computing Applications, 2016, vol. 30, no. 1, pp. 3 - 10.