

ОАО «Научно-исследовательский центр электронной вычислительной техники»

Опыт разработки отечественной высокоскоростной коммуникационной сети для суперкомпьютеров



А.С. Симонов, А.И. Слуцкий, Д.В. Макагон, Е.Л. Сыромятников, А.Н. Щербак,
И.А. Жабин, А.С. Фролов

1 ноября 2012 года

- Мотивы выполнения проекта
- Что сделано в рамках проекта
- Что даст внедрение результатов

Мотивы выполнения проекта

Суперкомпьютерные технологии

Позиция Правительства РФ

1. В Перечень критических технологий Российской Федерации, утверждённый Указом Президента РФ от 7 июля 2011 года № 899, включены «Технологии и программное обеспечение распределенных и высокопроизводительных вычислительных систем»
2. Решением президиума Правительственной комиссии по высоким технологиям и инновациям, протокол от 2 августа 2010 г. № 3, создана Национальная Суперкомпьютерная Технологическая Платформа
3. Решением Комиссии при Президенте РФ по модернизации и технологическому развитию экономики России в 2010 г. запущен проект «Суперкомпьютерное образование»

Суперкомпьютеры за рубежом

1 Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom

2 K computer, SPARC64 VIIIx 2.0GHz, Tofu interconnect

3 Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom

4 SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR

5 Tianhe-1A - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050

6 Jaguar - Cray XK6, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA 2090

7 Fermi - BlueGene/Q, Power BQC 16C 1.60GHz, Custom

8 JuQUEEN - BlueGene/Q, Power BQC 16C 1.60GHz, Custom

9 Curie thin nodes - Bullx B510, Xeon E5-2680 8C 2.700GHz, Infiniband QDR

10 Nebulae - Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050

Рейтинги TOP-500, GREEN-500, GRAPH-500

В TOP-10 представлены:

Страны – США (6 систем), Япония (1 система),
ЕЭС (1 система), Китай (2 системы)

Компании – IBM (5 систем), CRAY (1 система),
Fujitsu (1 система), Bull SA (1 система),
NUDT(1 система), Dawning (1 система)

Характеристики:

- Производительность – 2,9 ÷ 20 Пфлопс
- Число процессорных ядер – 77 тыс. ÷ 1,5 млн.
- Потребляемая мощность – 650 кВт ÷ 12 МВт

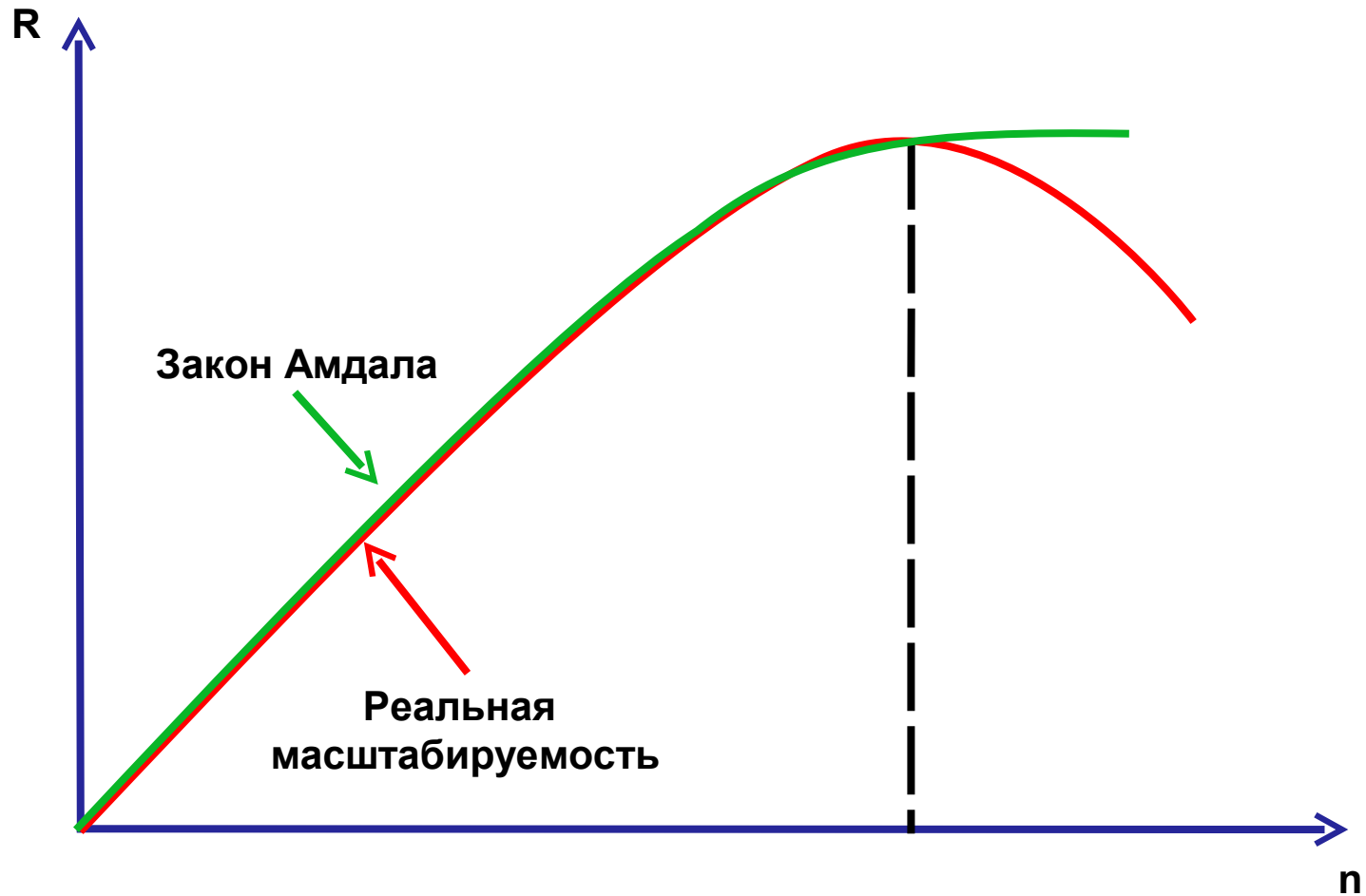
Суперкомпьютеры у нас в стране



Пять наиболее мощных суперкомпьютеров в России

№	Где установлен	Произв., Тфлопс	Поставщик	Год
1.	МГУ им. М.В. Ломоносова	1700 / 901	Т-Платформы	2012
2.	МСЦ РАН	227 / 119	Hewlett-Packard	2009
3.	РНЦ Курчатовский институт	123 / 101	Hewlett-Packard	2010
4.	ЮУрГУ	117 / 100	РСК	2010
5.	Институт математики и механики УрО РАН	160 / 75	Hewlett-Packard	2011





Мотивы разработки СБИС, почему не ПЛИС или Hardcopy FPGA?

- Масштабируемость – универсальные сети (Ethernet, Infiniband) отстают от заказных сетей современных суперкомпьютеров, а заказные сети недоступны
- Характеристики (Bandwidth, Latency)
- Цена (СБИС vs ПЛИС)
- Эффективная аппаратная поддержка современных парадигм программирования с использованием односторонних коммуникаций и PGAS

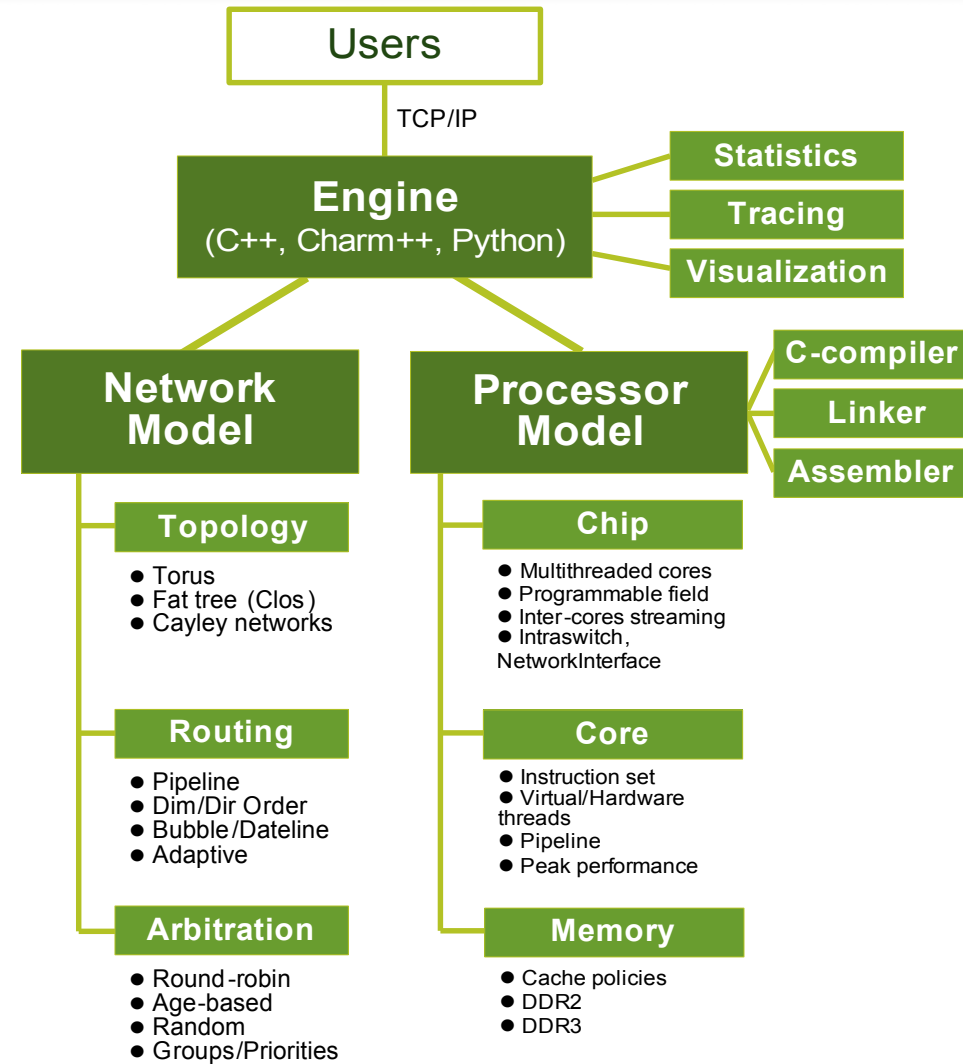
Что сделано в рамках проекта?

Этапы проекта

- Анализ зарубежного опыта (W. Dally, J. Duato, etc.)
- Разработка оценочной модели и проведение исследований сетей с различными топологиями на различных тест-паттернах
- Выбор и оценка эффективности алгоритмов маршрутизации
- Разработка спецификации на маршрутизатор, программную модель, стек ПО

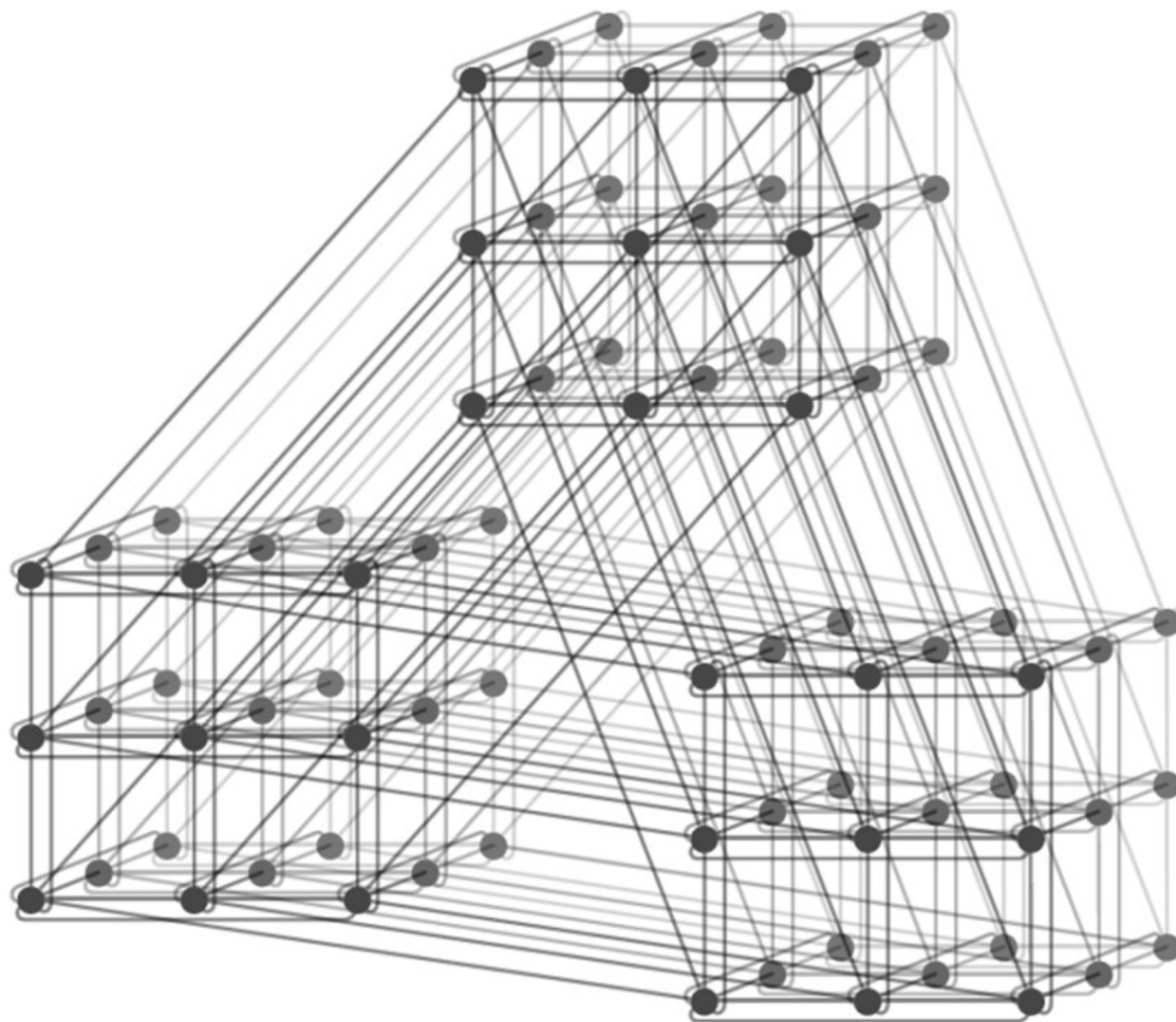
Параллельная имитационная модель

- Потактовая модель на языке Charm++
- Используется:
 - для исследования новых архитектур
 - для оценки производительности и верификации разрабатываемой коммуникационной сети
- Масштабирование производительности модели до 256 узлов суперкомпьютера “Ломоносов”

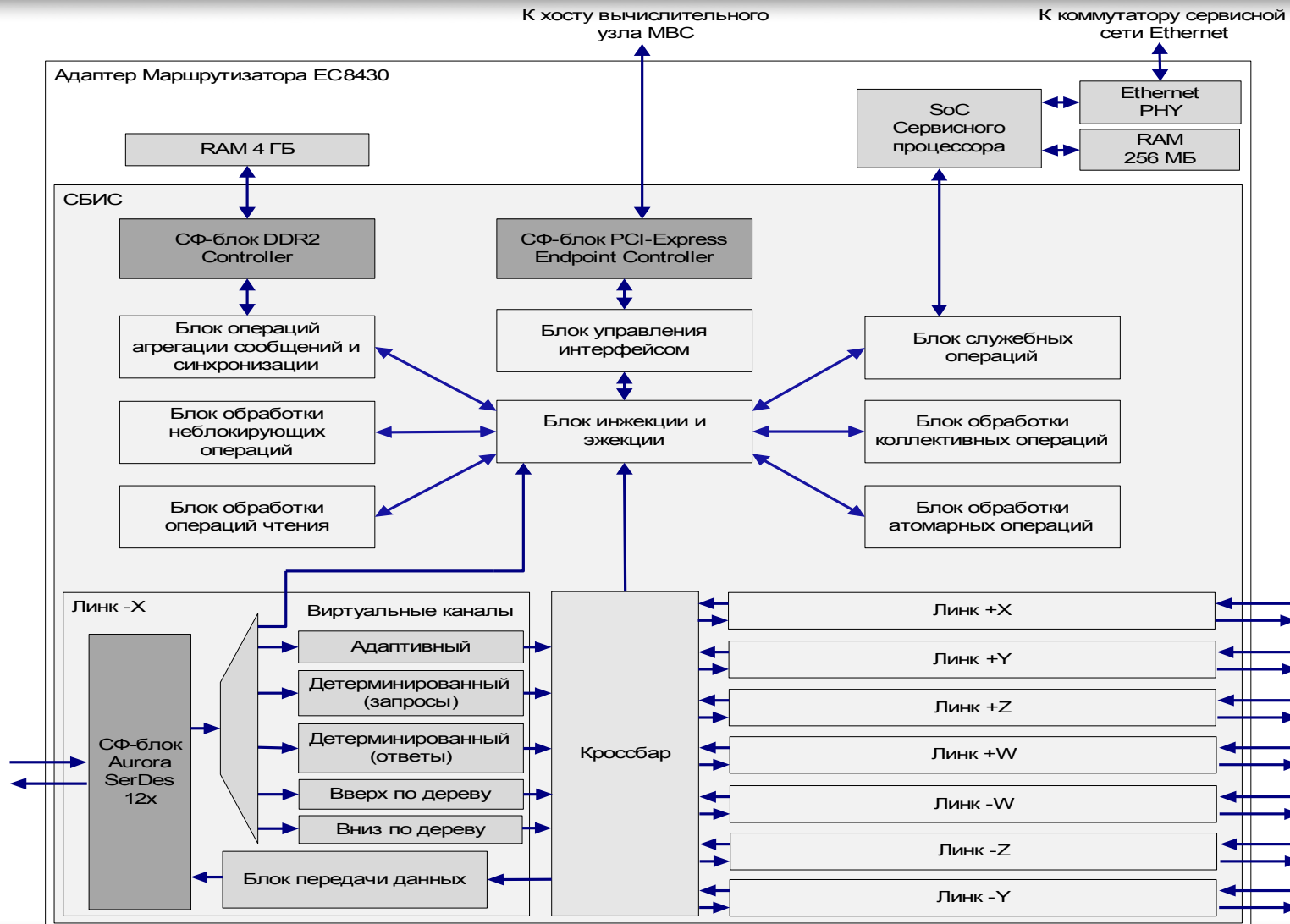


- Топология — многомерный тор
- Аппаратная поддержка глобально адресуемой памяти
- Детерминированная и адаптивная передача пакетов
- Односторонние коммуникации (RDMA)
- Эффективная поддержка MPI
- Аппаратная поддержка коллективных операций

Сеть с топологией 4D-тор



Микроархитектура маршрутизатора



Функциональные возможности

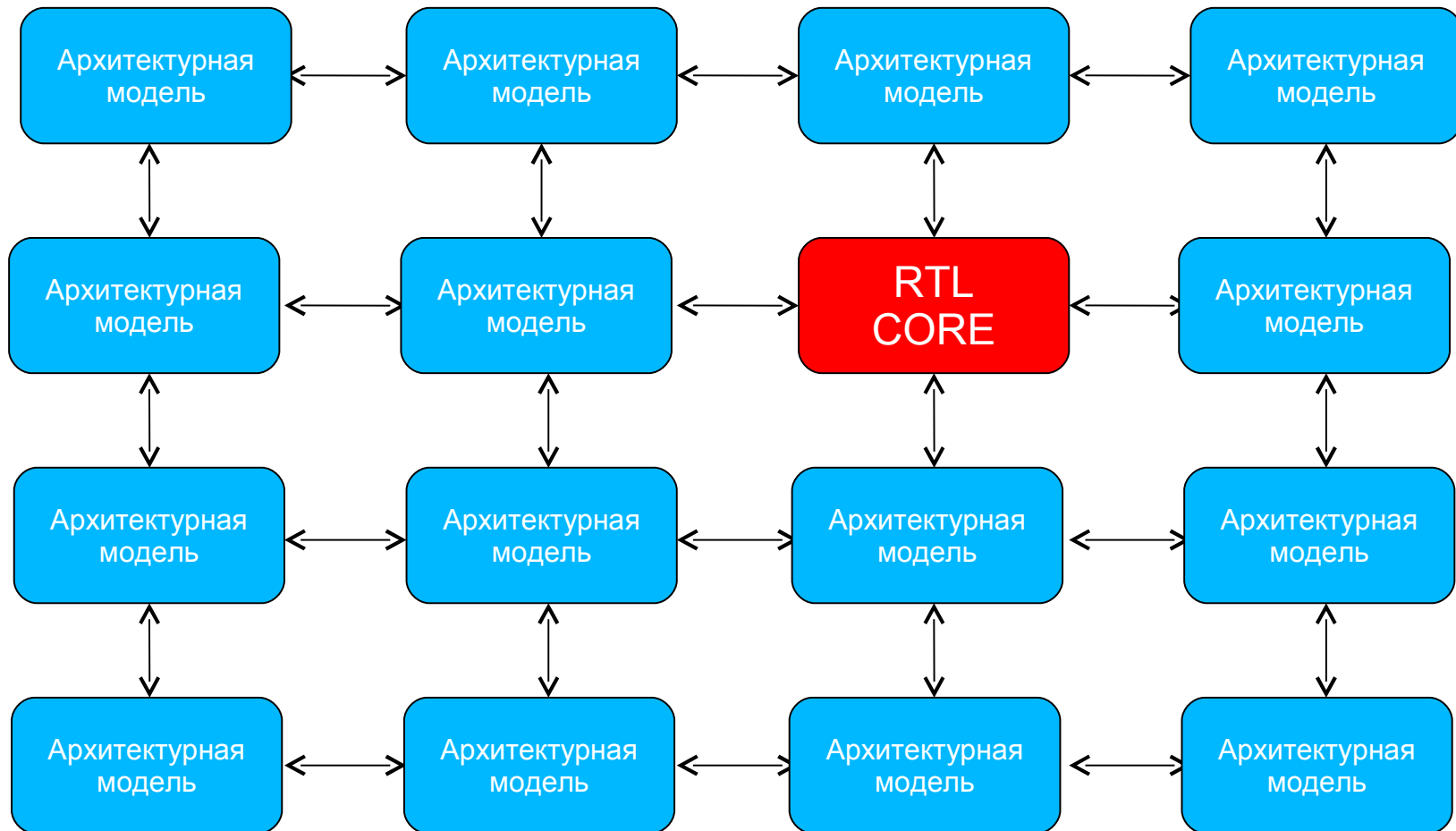
- Протокол надёжной доставки пакетов
- Аппаратная поддержка многопоточности
- Реализация сборки массивов в памяти маршрутизатора с последующим копированием в память узла
- Система обеспечения отказоустойчивости
- Аппаратная реализация операций:
 - Запись в память удалённого узла
 - Запись в память удалённого узла с сохранением порядка
 - Атомарные операции в памяти удалённого узла
 - Чтение из памяти удалённого узла
 - Неблокирующая запись больших массивов в память удалённого узла
 - Чтение больших массивов из памяти удалённого узла

- Реализация backend для OpenSHMEM, MPI, GASNET (UPC, Co-Array Fortran)
- Адаптированная версия параллельной файловой системы Lustre
- Библиотеки поддержки вычислений ScaLAPACK, FFTW, BLAS.....
- Инфраструктура запуска, сервисная подсистема

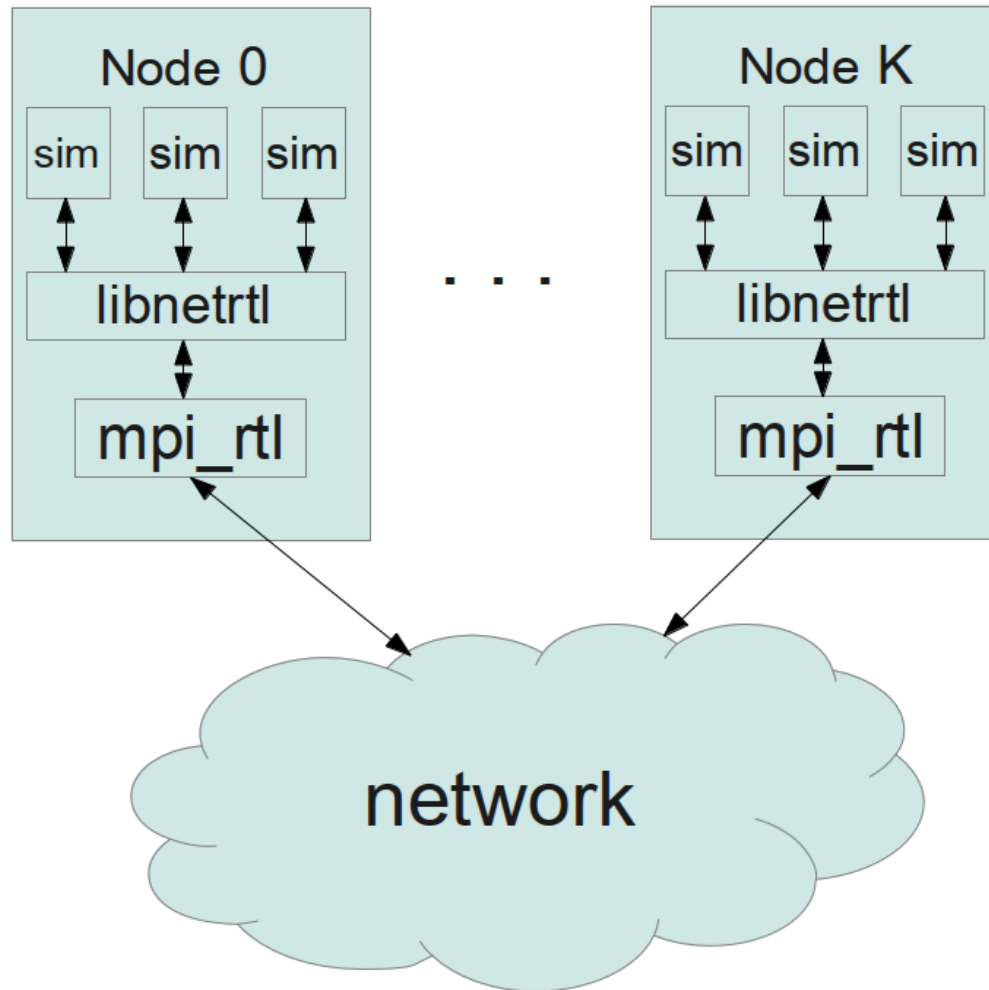
Верификация СБИС

- Проблемы верификации:
 - Сетевые (дедлоки, ливлоки, голодание, число узлов)
 - Скорость верификации
- Уровни верификации:
 - Вычислительная система
 - Маршрутизатор в целом (со всеми интерфейсами)
 - Функциональное ядро маршрутизатора
 - Крупные блоки функционального ядра

Верификация на уровне вычислительной системы

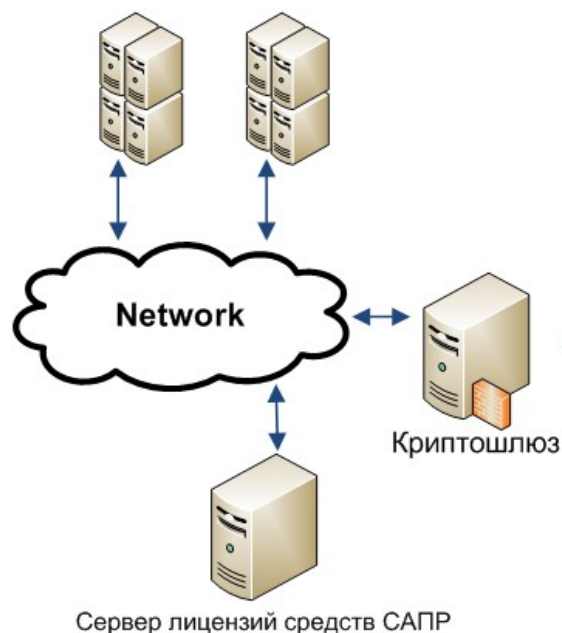


Параллельная система верификации



Вычислительные ресурсы, задействованные для верификации

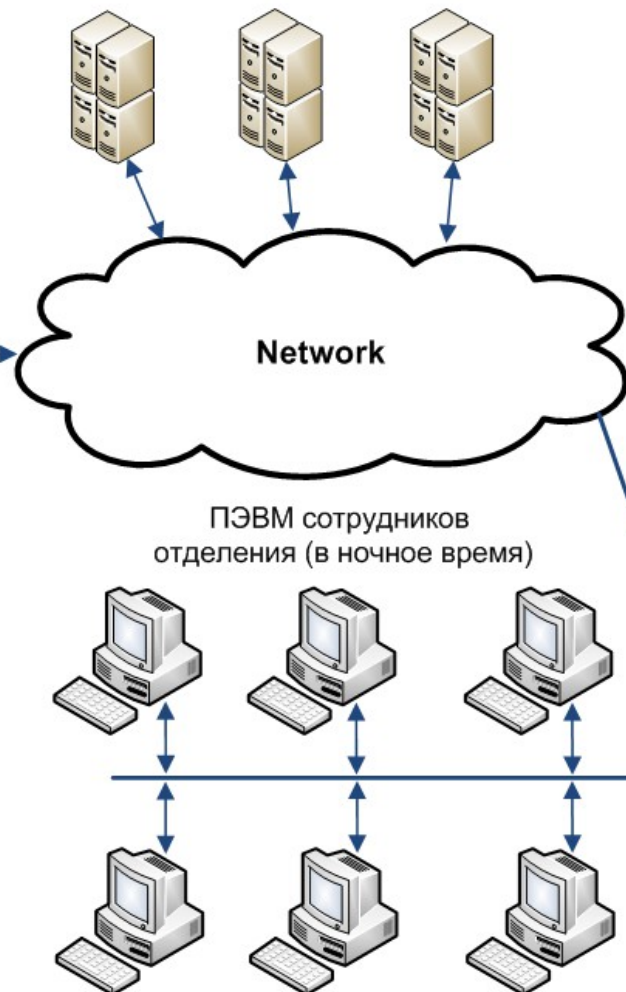
Кластеры ЦОД
ОАО «Концерн радиостроения «Вега»



Интернет

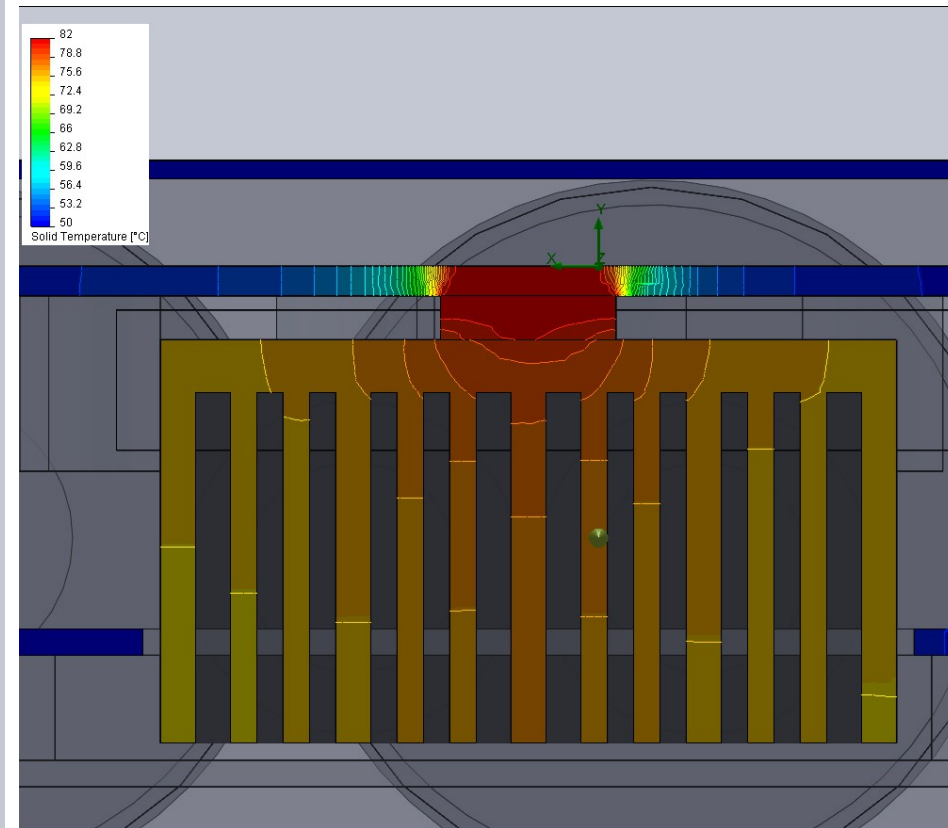
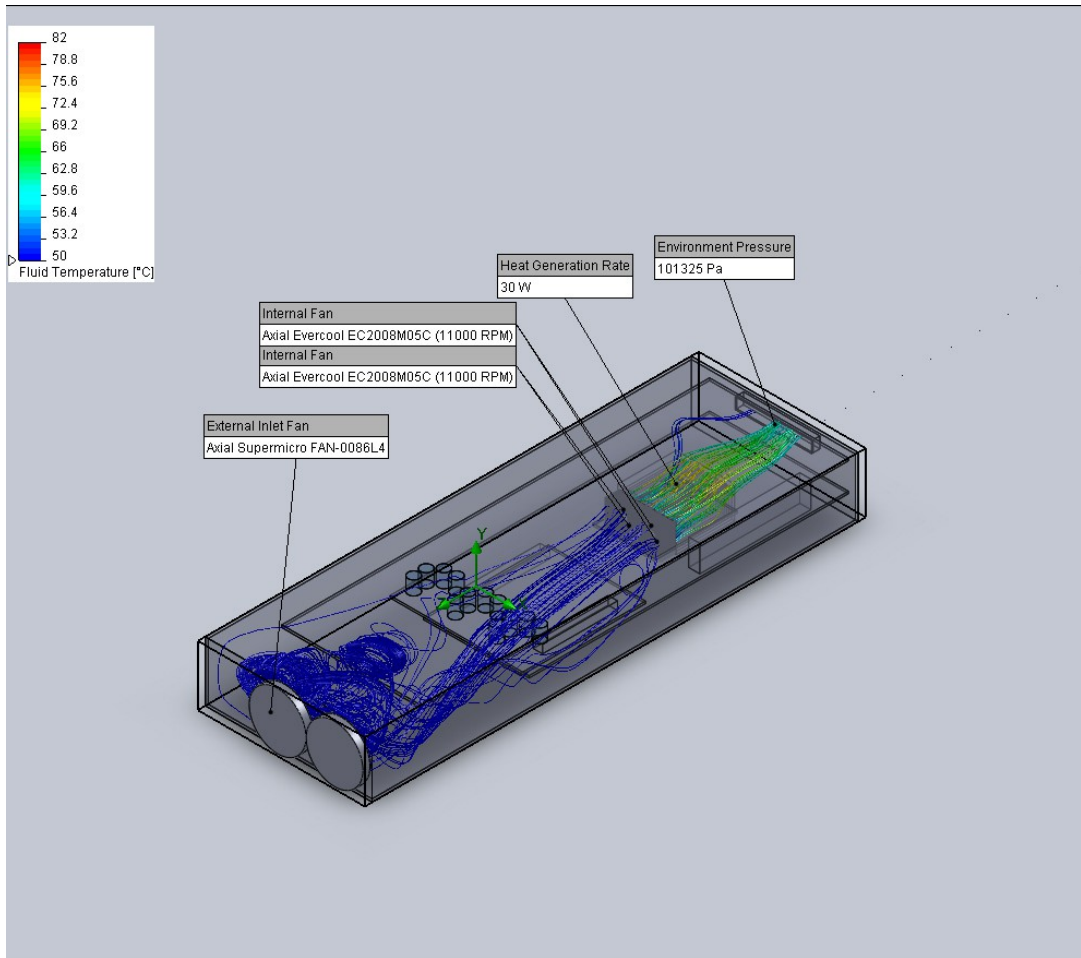


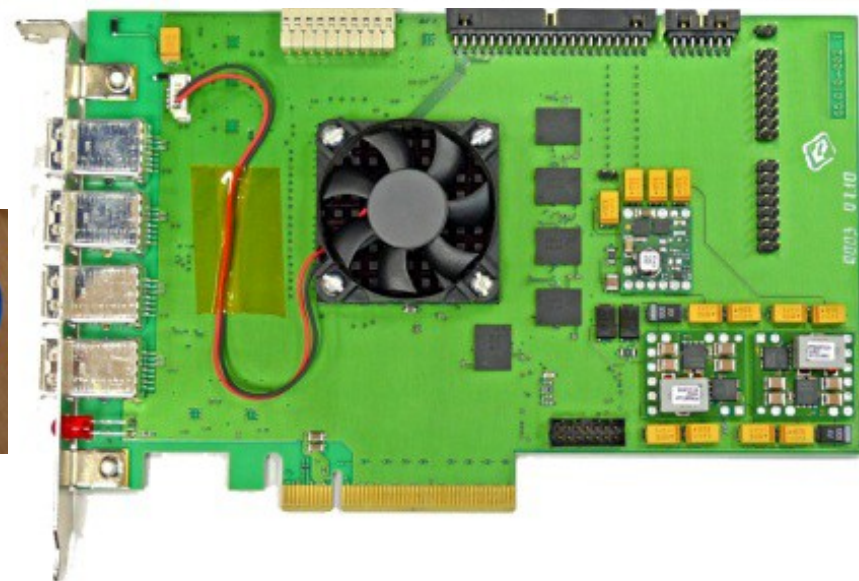
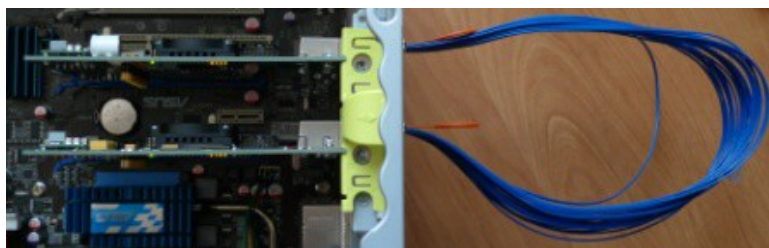
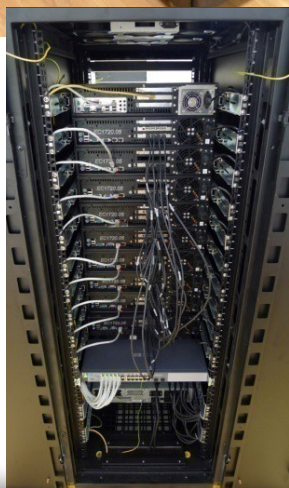
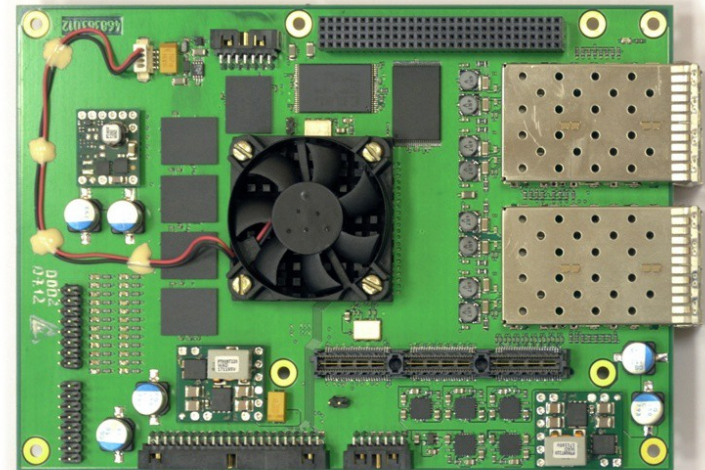
Кластеры ВЦ
ОАО «НИЦЭВТ»



Проектирование корпуса СБИС

Вопросы отвода тепла





- В ОАО «НИЦЭВТ» разработана и передана на изготовление СБИС маршрутизатора межузловой коммуникационной сети для вычислительных кластеров и суперкомпьютеров
- Характеристики СБИС:
 - Технология – 65 нм
 - Размер кристалла – 13.0 x 10.5 мм
 - Число транзисторов – более 180 млн.
 - Корпус СБИС – FCBGA-1521 40 x 40 мм
 - Потребляемая мощность – 30 Вт
 - Интерфейсы - PCI-Express 2.0, DDR3, SerDes 12x

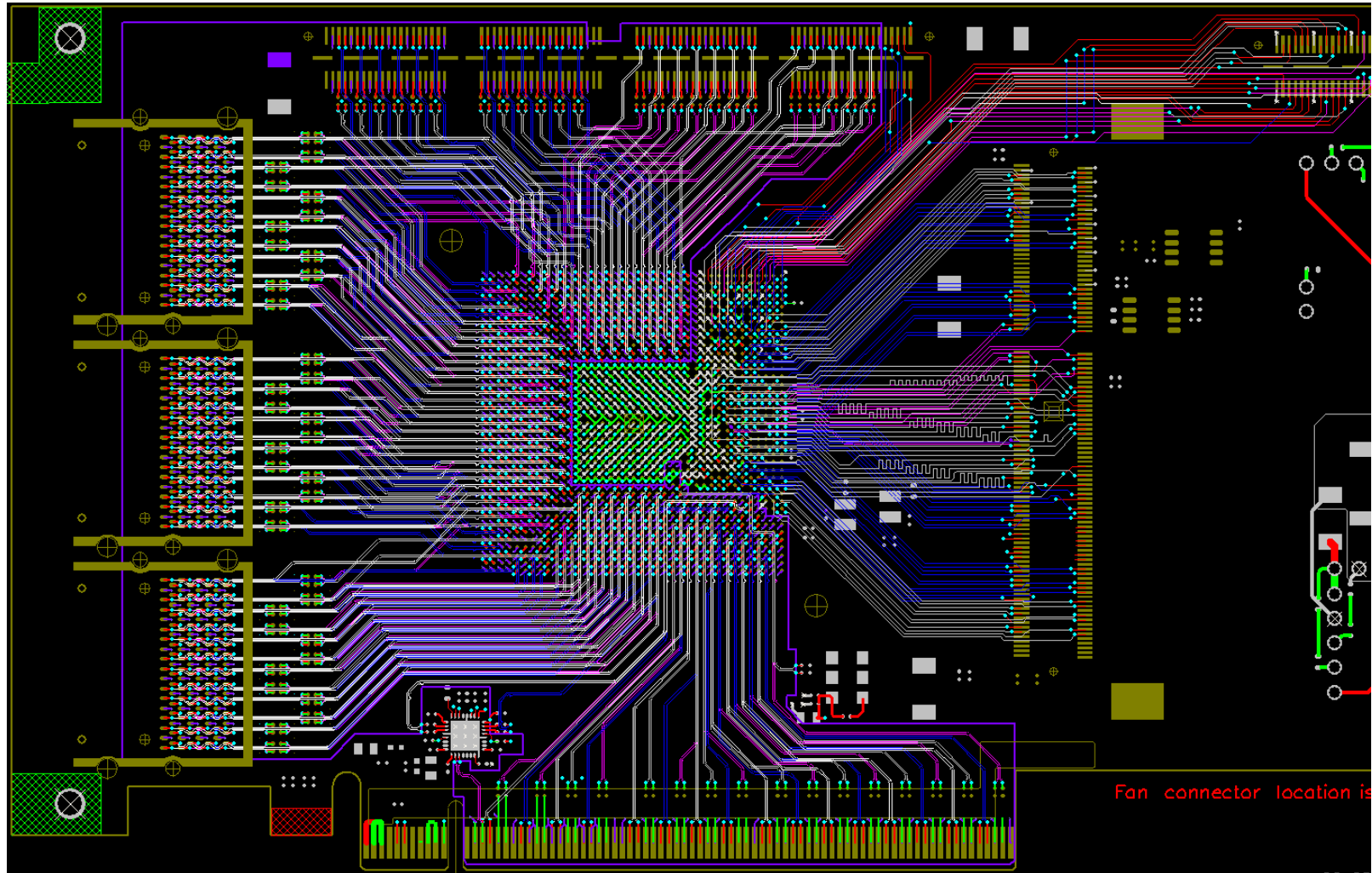
Варианты поставки

№	Наименование продукта	Прототипы	Начало серийных поставок
1.	СБИС K1927BB1Я	I кв. 2013 г.	III кв. 2013 г.
2.	Сетевой адаптер в форм-факторе PCI Express	II кв. 2013 г.	III кв. 2013 г.
3.	Сетевой модуль в форм-факторе EPIC-Express	III кв. 2013 г.	IV кв. 2013 г.
4.	В составе платформы	III кв. 2013 г.	IV кв. 2013 г.

Поддержка пользователей (Адаптер / СБИС К1927ВВ1Я)

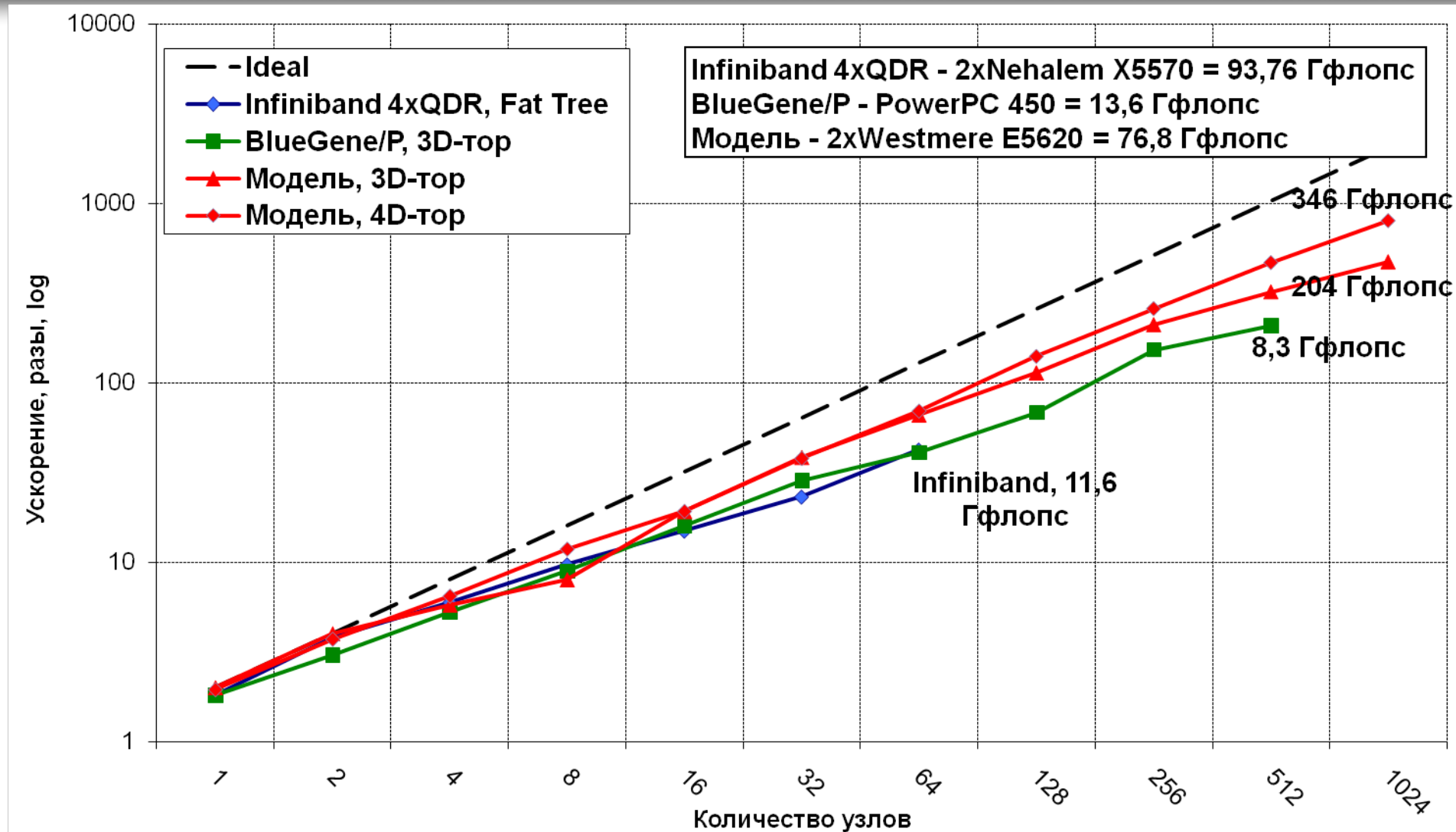
- Технические условия
- Руководство пользователя
- Руководство программиста (Program Manual)
- Техническое описание СБИС К1927ВВ1Я (Datasheet)
- Руководство по разработке устройств на основе СБИС К1927ВВ1Я (Design Guide)
- Reference Design

Reference Design

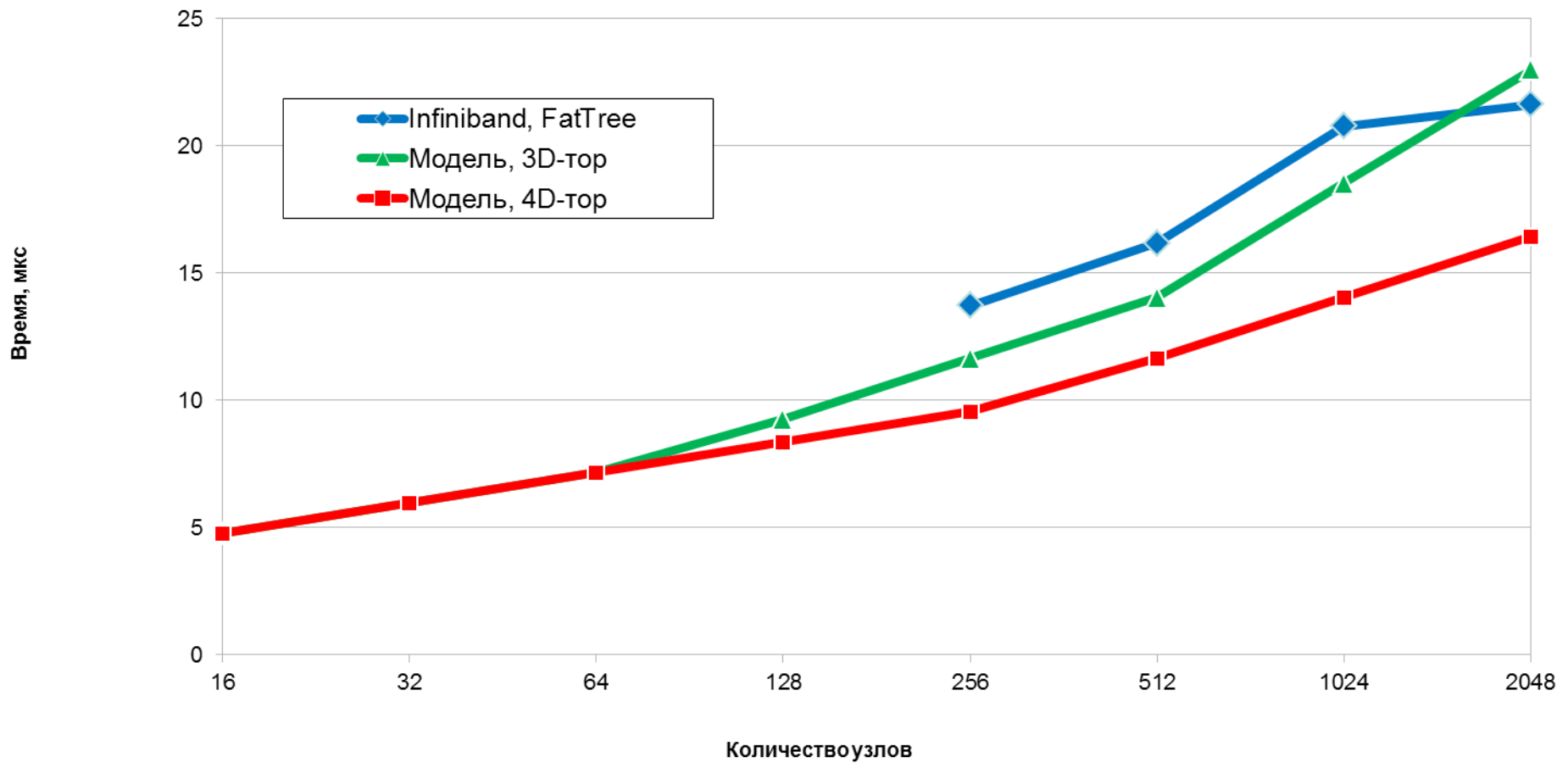


Что даст внедрение результатов?

Сравнение производительности на задаче умножения разреженной матрицы на вектор



Барьерная синхронизация



- Коллектив ОАО «НИЦЭВТ» справился с проблемами, возникшими в ходе выполнения проекта, ждём результат
- В ОАО «НИЦЭВТ» создана инфраструктура и сформирована команда, способная решать задачи по разработке современных сложных СБИС для НРС