Ростех Объединенная приборостроительная корпорация

Вычислительные системы с высокоскоростной сетью Ангара

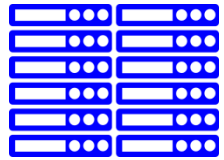
Симонов Алексей Сергеевич

НИЦЭВТ: История и настоящее

Ростех Объединенная приборостроительная корпорация

Содержание

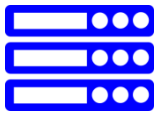
Назначение:



Коммуникационная сеть Ангара предназначена для осуществления передачи данных между узлами вычислительных систем с высокой скоростью и малой коммуникационной задержкой

Области применения:

1. Вычислительные кластеры для проведения научных расчетов, решения задач инженерного анализа, обработки
2. Системы хранения и обработки больших данных
3. В качестве интерконнекта вычислительного поля в центрах обработки данных

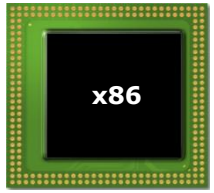


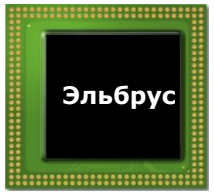
Совместимость:

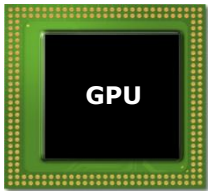
1. Системы на основе универсальных микропроцессоров x86, x86-64 (Intel, AMD), Эльбрус, GPU NVIDIA/AMD

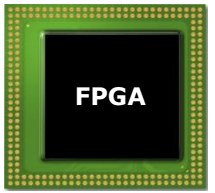
2. Специализированные системы на основе FPGA

3. Проводятся работы по стыковке с микропроцессорами с архитектурой ARM

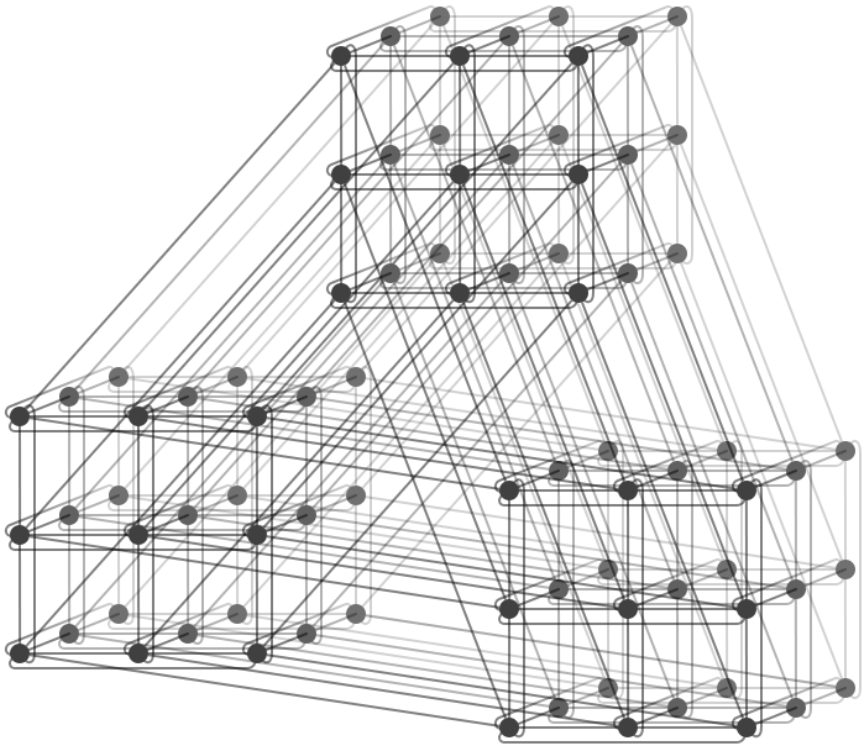








Коммуникационная сеть Ангара
Принципы построения



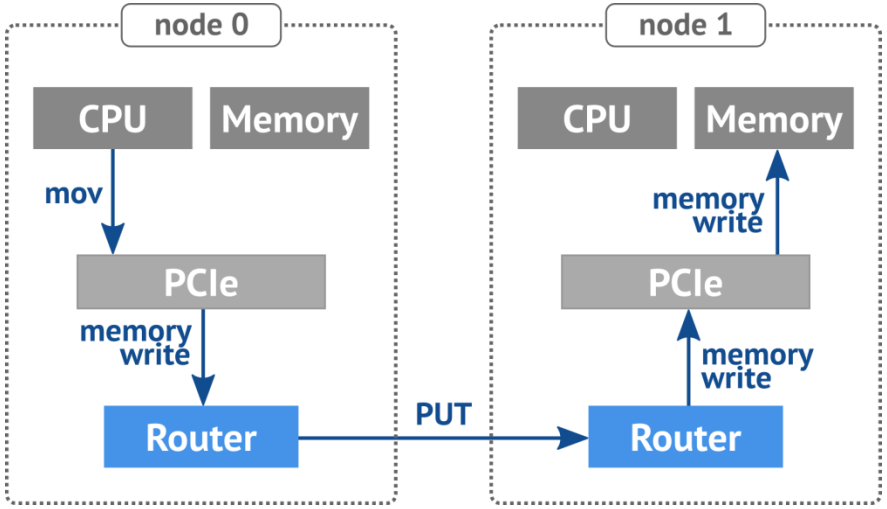
- Топология от «1D-mesh» до «4D-тор»
- Односторонние коммуникации:
  - put (запись в удалённую память)
  - get (чтение из удалённой памяти)
  - атомарные операции add и xor
- Прямой доступ в память удаленного узла (RDMA)
- Коллективные операции:
  - broadcast
  - reduce
- Поддержка многоядерности
- Адаптивная передача пакетов
- Механизмы синхронизации

Коммуникационная сеть Ангара
Основные низкоуровневые операции

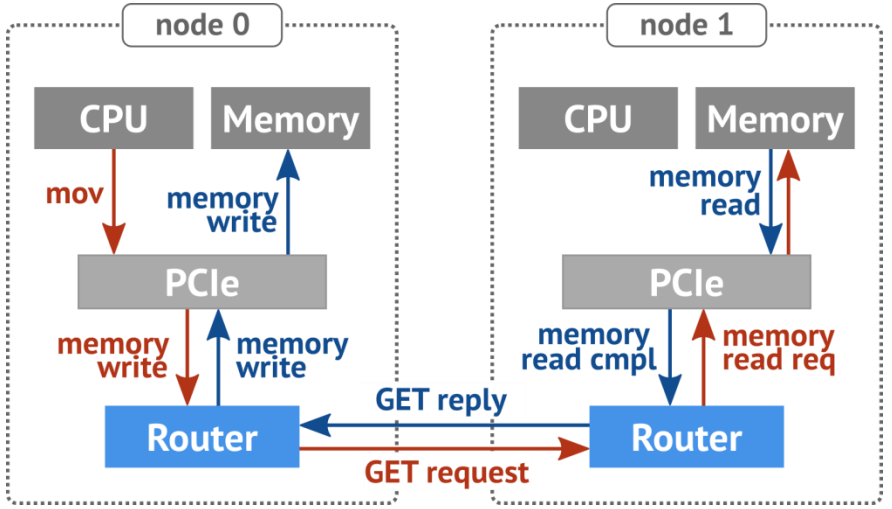
Операция	Кто инжектирует данные в сеть	Размер транзакции	Примечания
PUT vc0	Хост	1—256 байт	Эжекция мимо кроссбара (ответы)
PUT vc1	Хост	1—256 байт	Эжекция через кроссбар (запросы)
GET	Адаптер	1—256 байт	Гранулярность 8 байт
ADD/XOR	Хост	1—8 байт	Атомарность средствами адаптера
LWPUT	Адаптер	8—2^{30} байт	Гранулярность 8 байт
LGET	Адаптер	8—2^{30} байт	Гранулярность 8 байт

- *При инжекции коротких пакетов (PUT) на хосте выполняются операции MOV в адресное пространство ресурса PCI Express адаптера (инжекционный буфер).*
- *При инжекции длинных пакетов (LWPUT) адаптер выполняет memory read из памяти хоста, после чего отправляет полученные из памяти данные в сеть.*

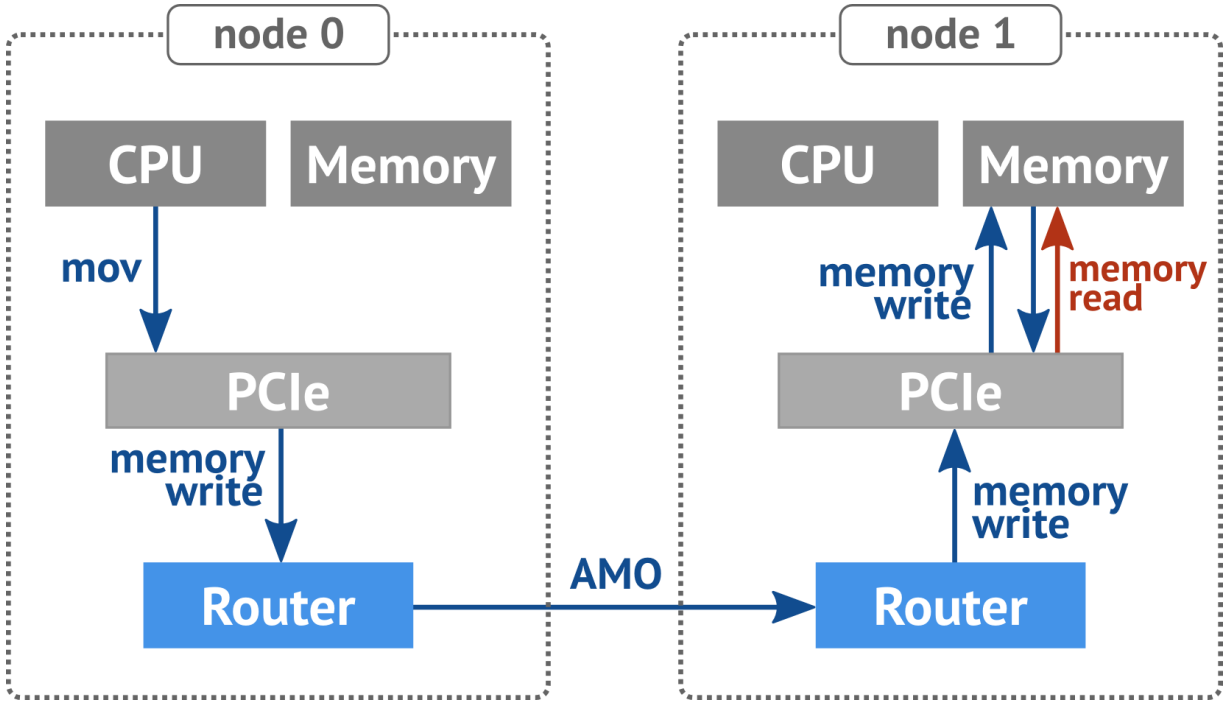
Запись



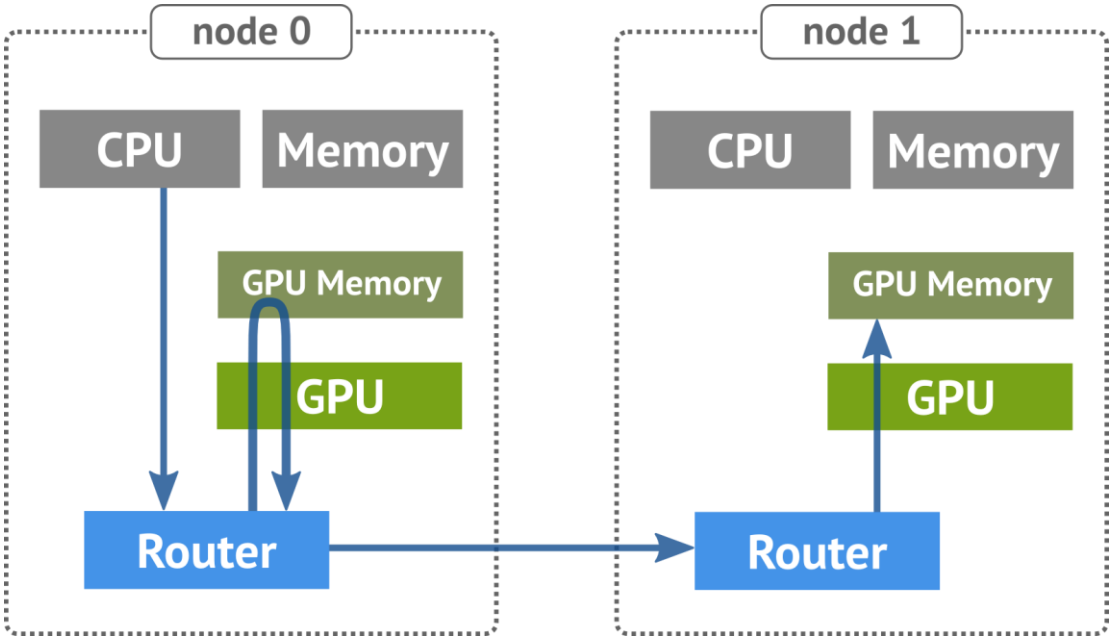
Чтение



Удалённая атомарная операция (сложение, XOR)

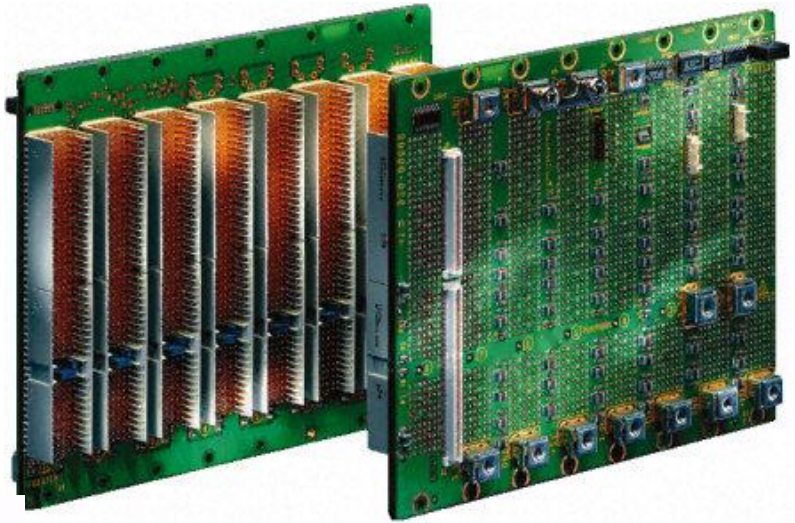


Реализация GPU Direct RDMA
Прямая запись из памяти локального GPU в память GPU на удаленном узле

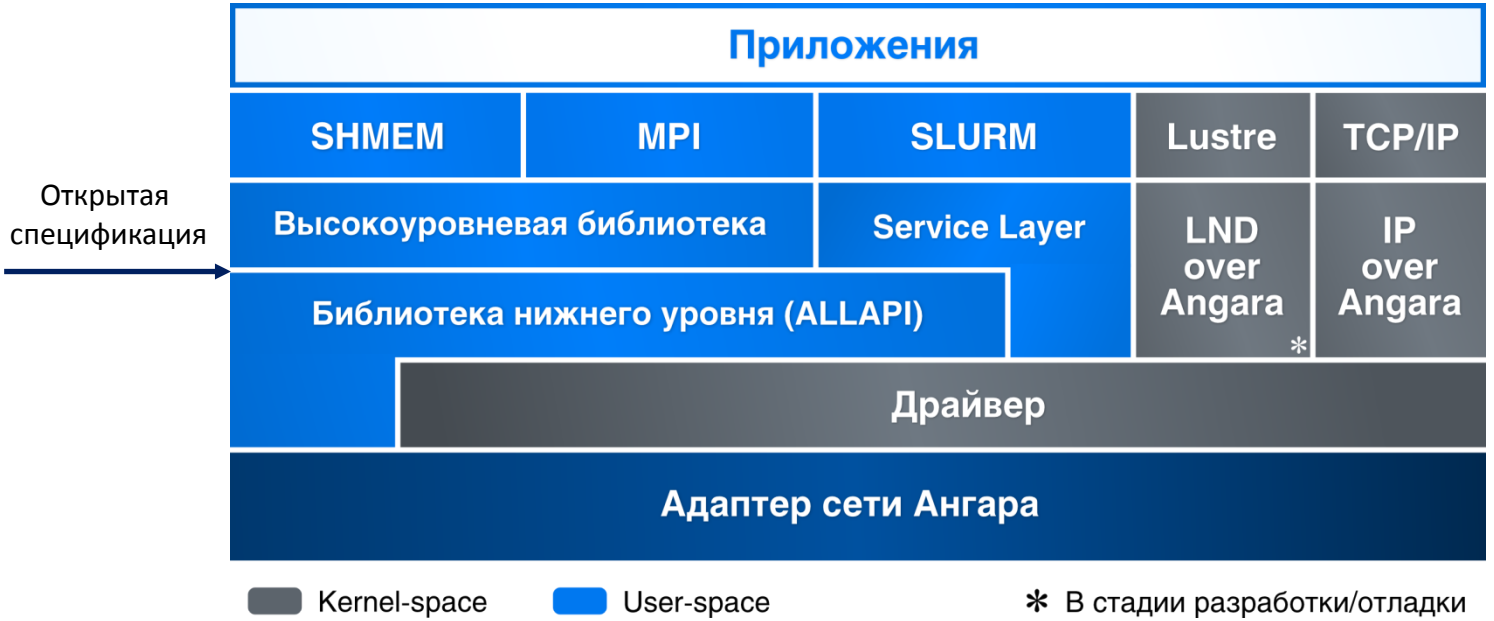


Передача данных между узлами

- Поддержка различных физических сред передачи данных (кабели медные и оптические, объединительные платы)
- Настройка и управление параметрами приемопередатчиков
- Коммуникационная задержка 670 нс

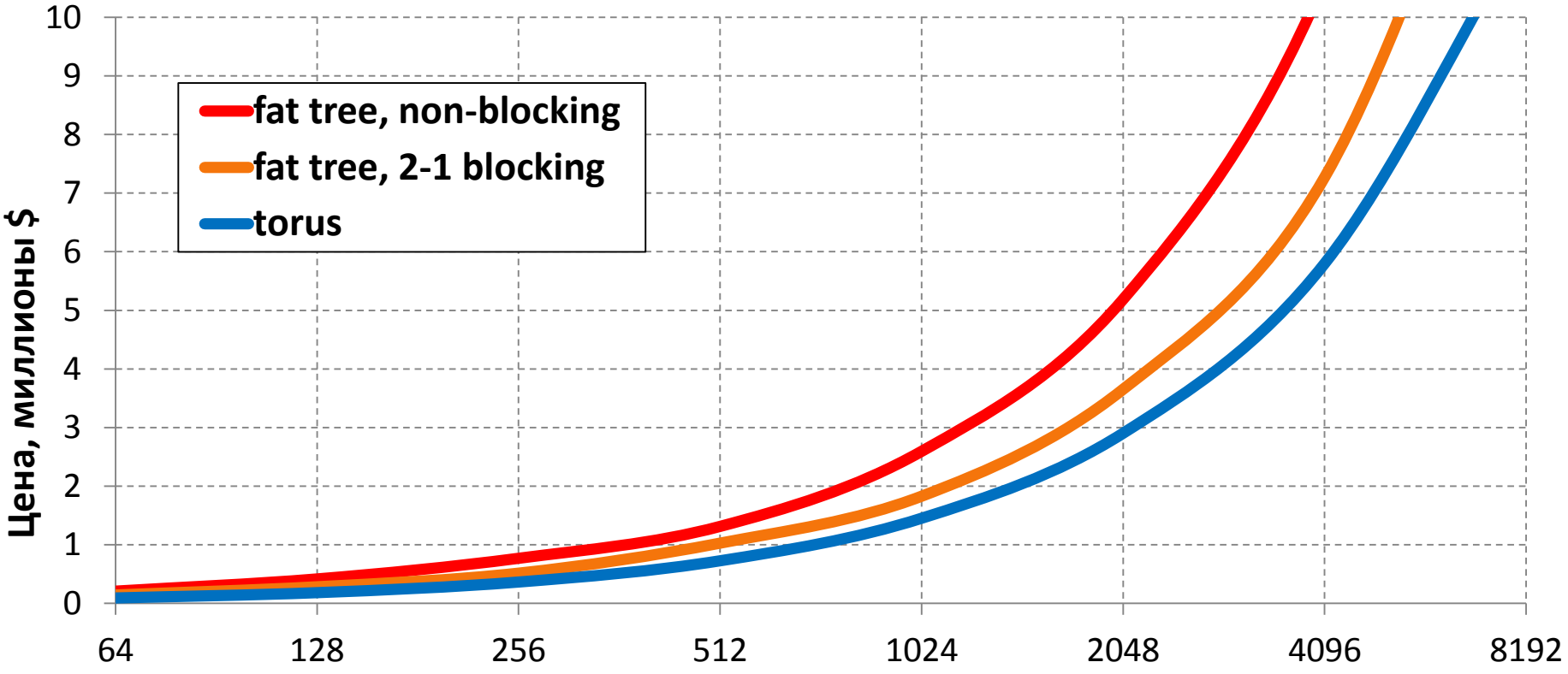


Стек программного обеспечения

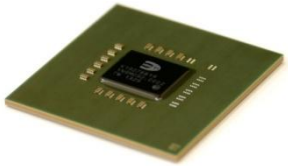


- Поддержка ОС : Astra Linux SE 1.3, ОС «Эльбрус» ,OpenSUSE/SLES 11SP3, CentOS 6.0-6.3, Версия ядра Linux от 2.6.21 до 3.16.0
- Поддержка компиляторов языков Fortran 77/90/95 (GNU, Intel), C/C++ (GNU, Intel)

Порядок затрат на сетевое оборудование для разных топологий



Сетевое оборудование Ангара и сервисы



Чип ЕС8430

FCBGA 1521
40×40 мм
35 Вт

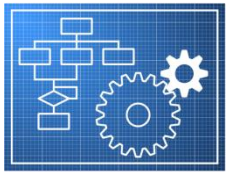
Development kit

PCIe gen 2 x16
Full-height,
full length

Варианты исполнения Адаптера PCIe

PCIe gen 2 x16
- Full-height, full length
- Full-height, half length

Заказная разработка платы адаптера



Доработка ПО адаптера и адаптация библиотек

Адаптация и профилирование прикладного ПО

Содержание

Тестовый вычислительный кластер «Ангара-К1»
36 узлов

НИЦ ВТ

Ростех Объединенная приборостроительная корпорация

24 вычислительных узла

- Supermicro SuperServer 5017GR-TF
- 2 процессора Intel Xeon E5-2630 (LGA2011, 6 ядер, 2.3 ГГц)
- 64 ГБ

12 вычислительных узлов

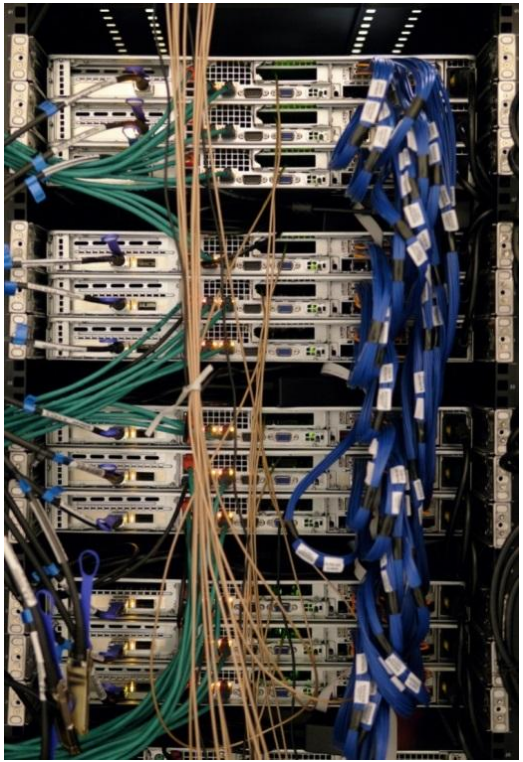
- Supermicro SuperServer 5017GR-TF
- процессор Intel Xeon E5-2660 (LGA2011, 8 ядер, 2.2 ГГц)
- 64 ГБ

Сеть «Ангара»

- Адаптер EC8430, топология 3D-тор 3x3x4
- Собственная реализация OpenSHMEM
- MPI: MPICH 3.0.4

Операционная система

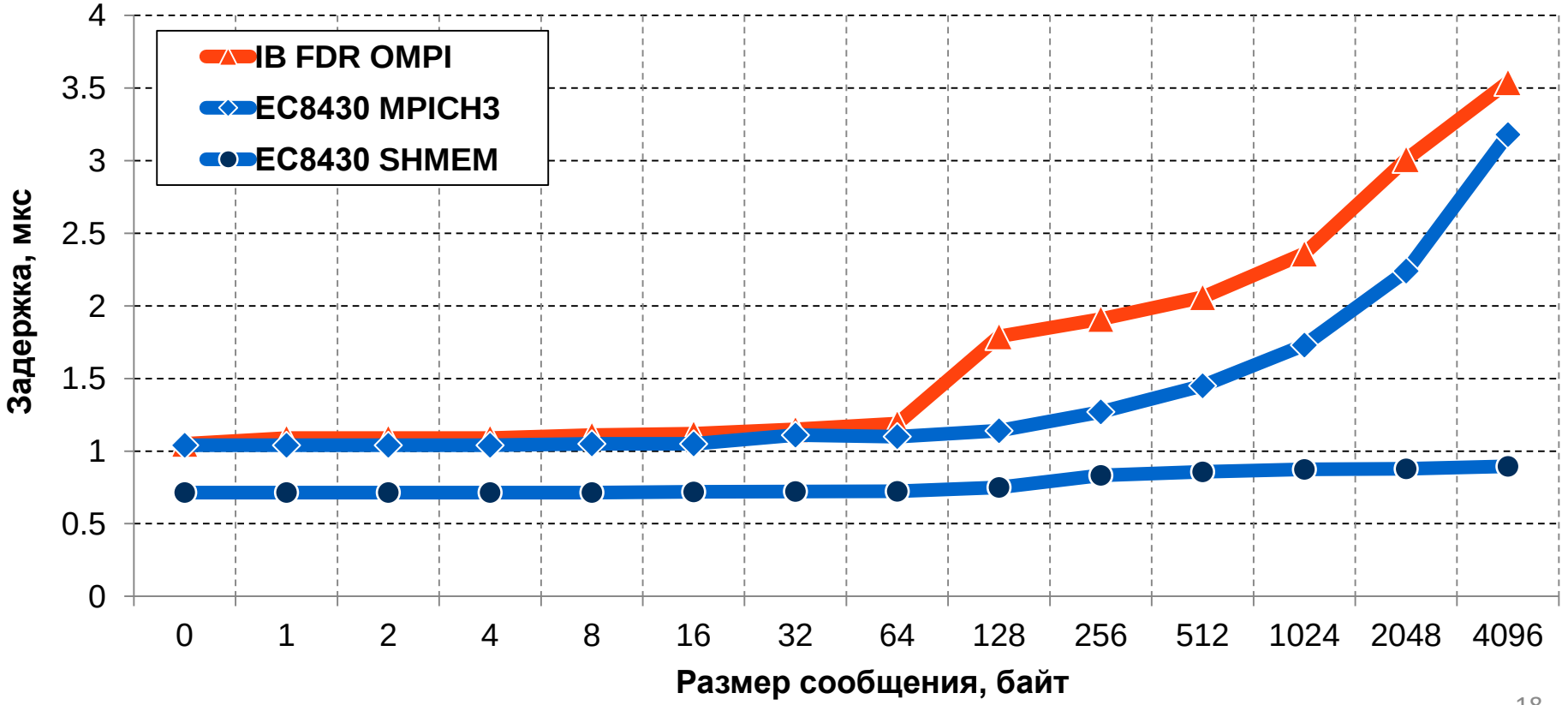
- SLES 11 SP2, Linux 3.0.13-0.27-default
- GCC 4.3.4 (revision 152973)

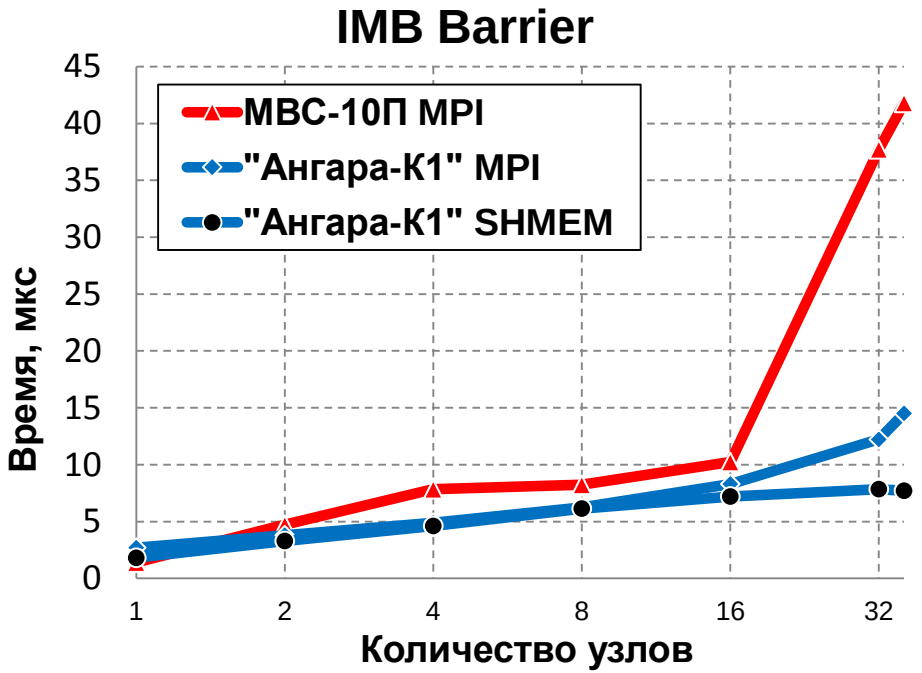


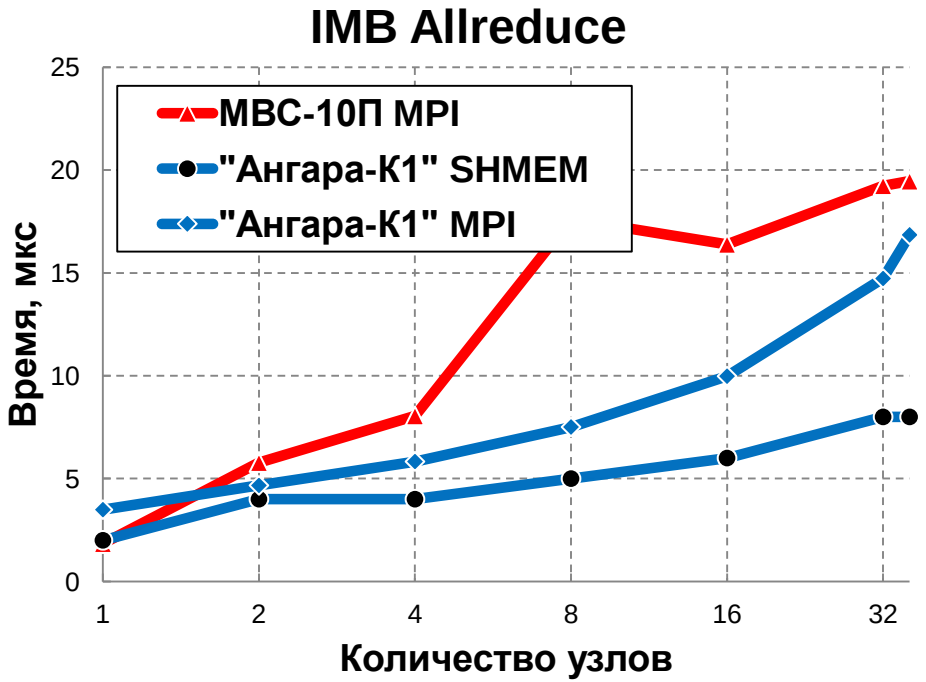
Ангара-К1 и МВС-10П (МСЦ РАН)

	Ангара-К1		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 IB FDR
2 узла – задержка на MPI (osu_latency)

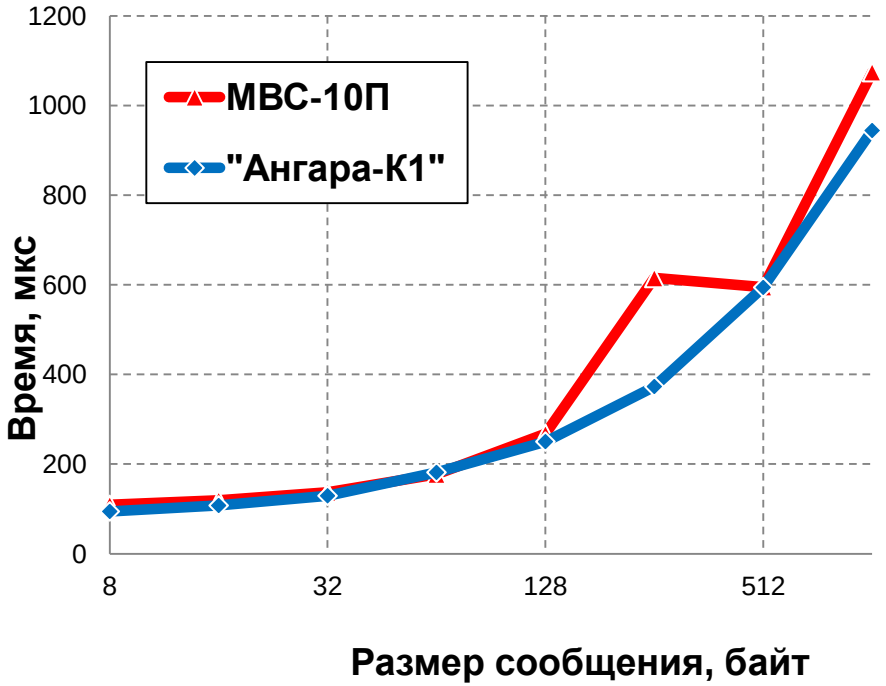


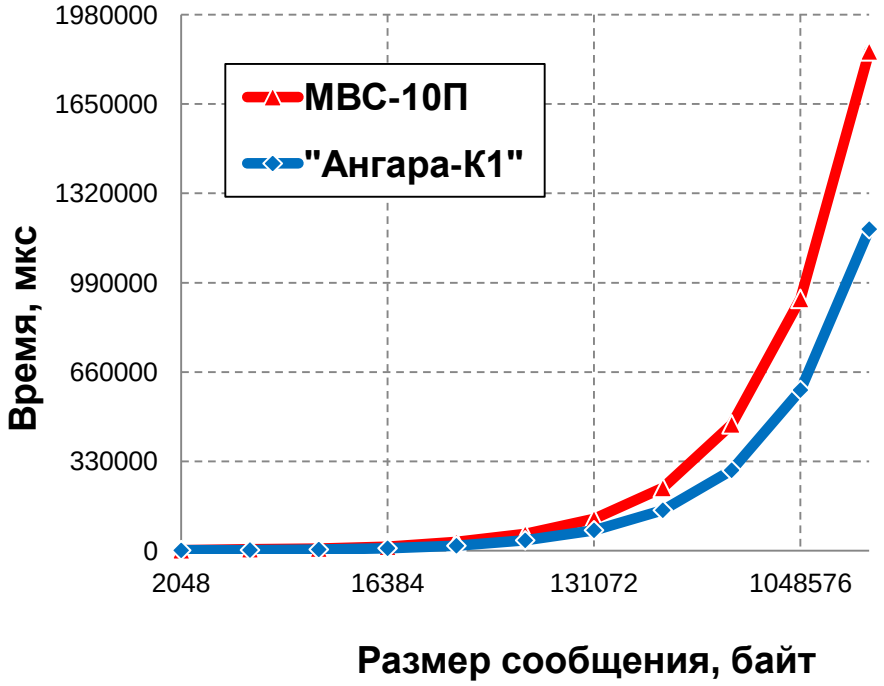




размер сообщения 8 байт

Коллективные операции
IMB AlltoAll, 32 узла

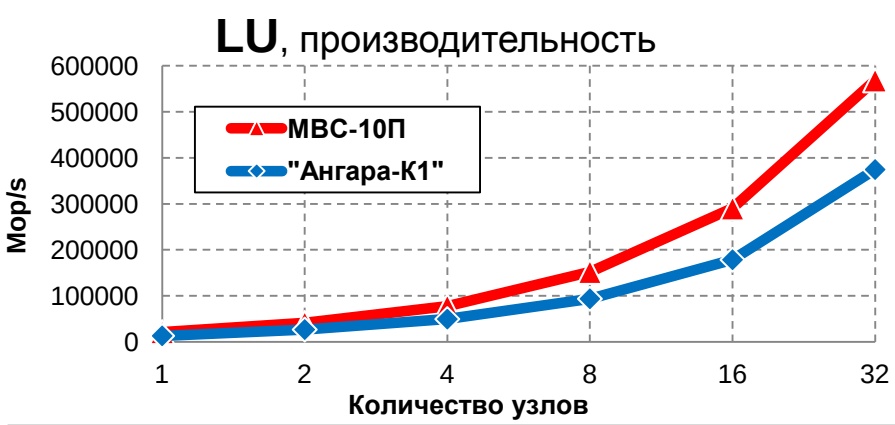


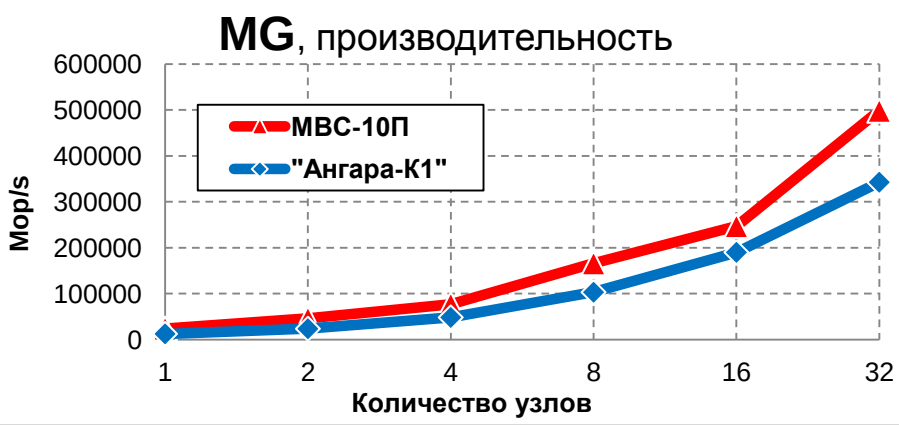


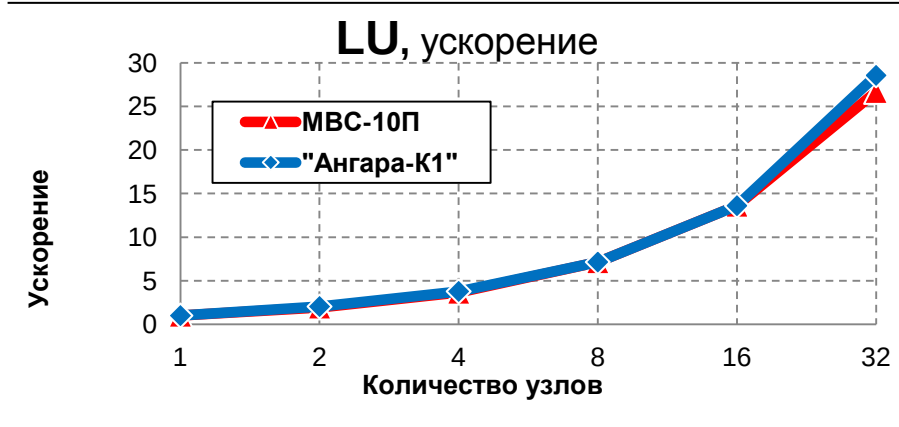
Результаты тестов HPL и HPCG

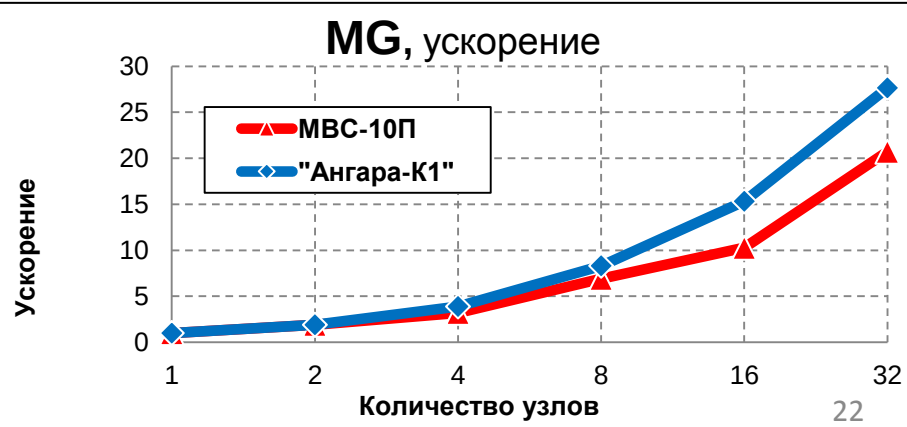
		Ангара	МВС-10П
HPL	Тфлопс	4.44	–
	% пиковой	85 %	–
HPCG MPI	Гфлопс	279	363
	% пиковой	5.3 %	5.4 %
HPCG SHMEM	Гфлопс	342	–
	% пиковой	6.5 %	–

Результаты тестов NPB LU, MG (класс C)

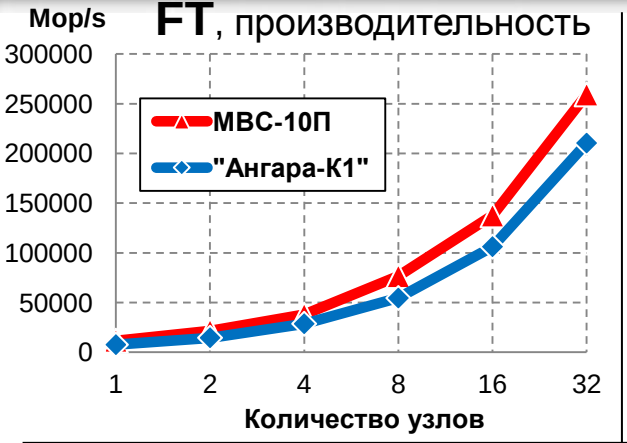


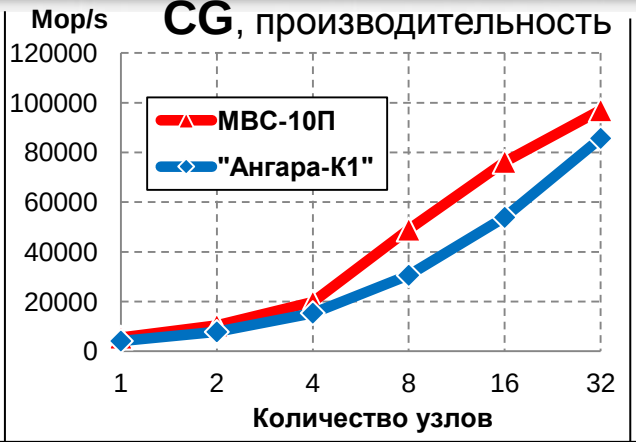


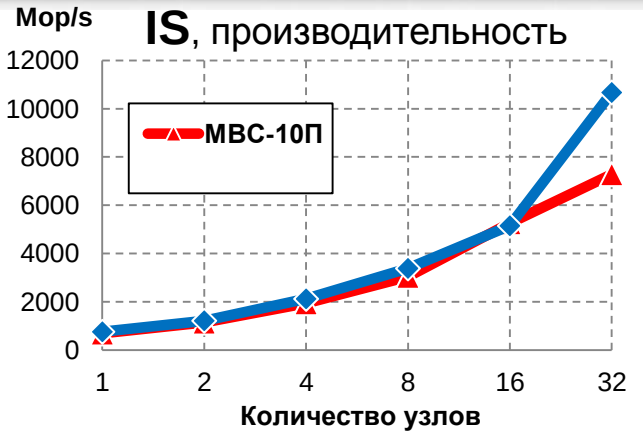


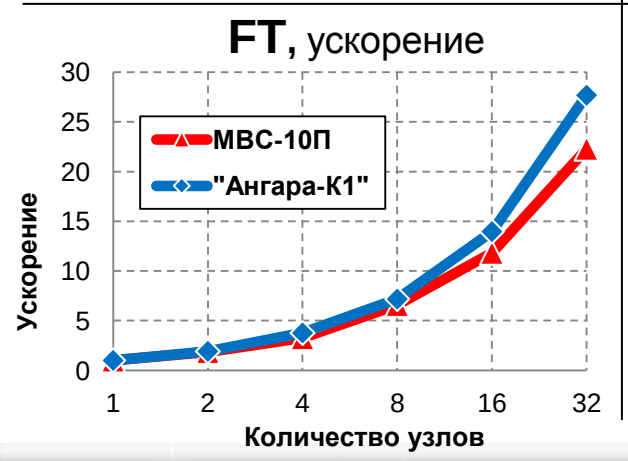


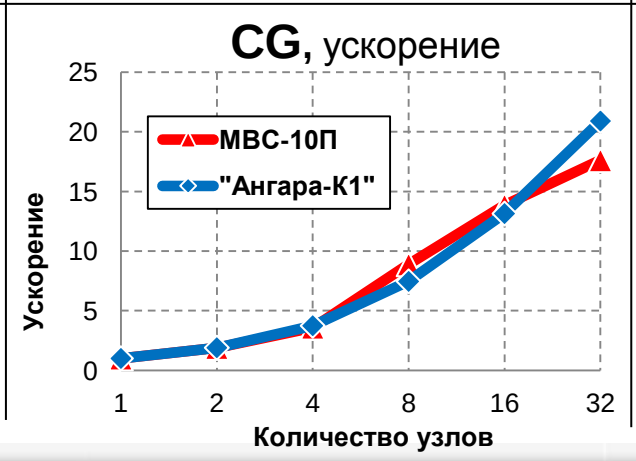
Результаты тестов NPB FT, CG, IS (класс C)

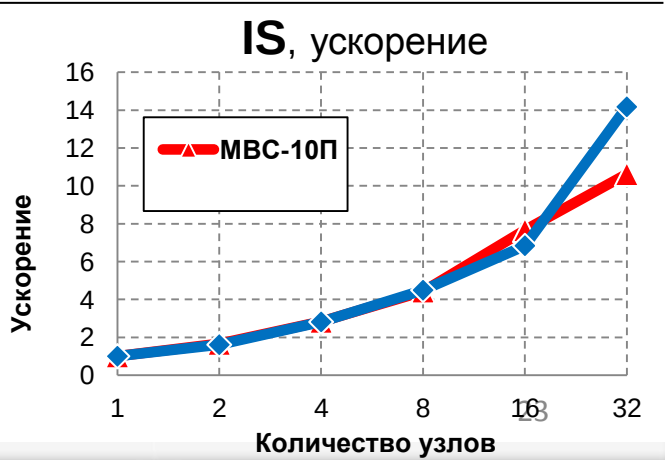








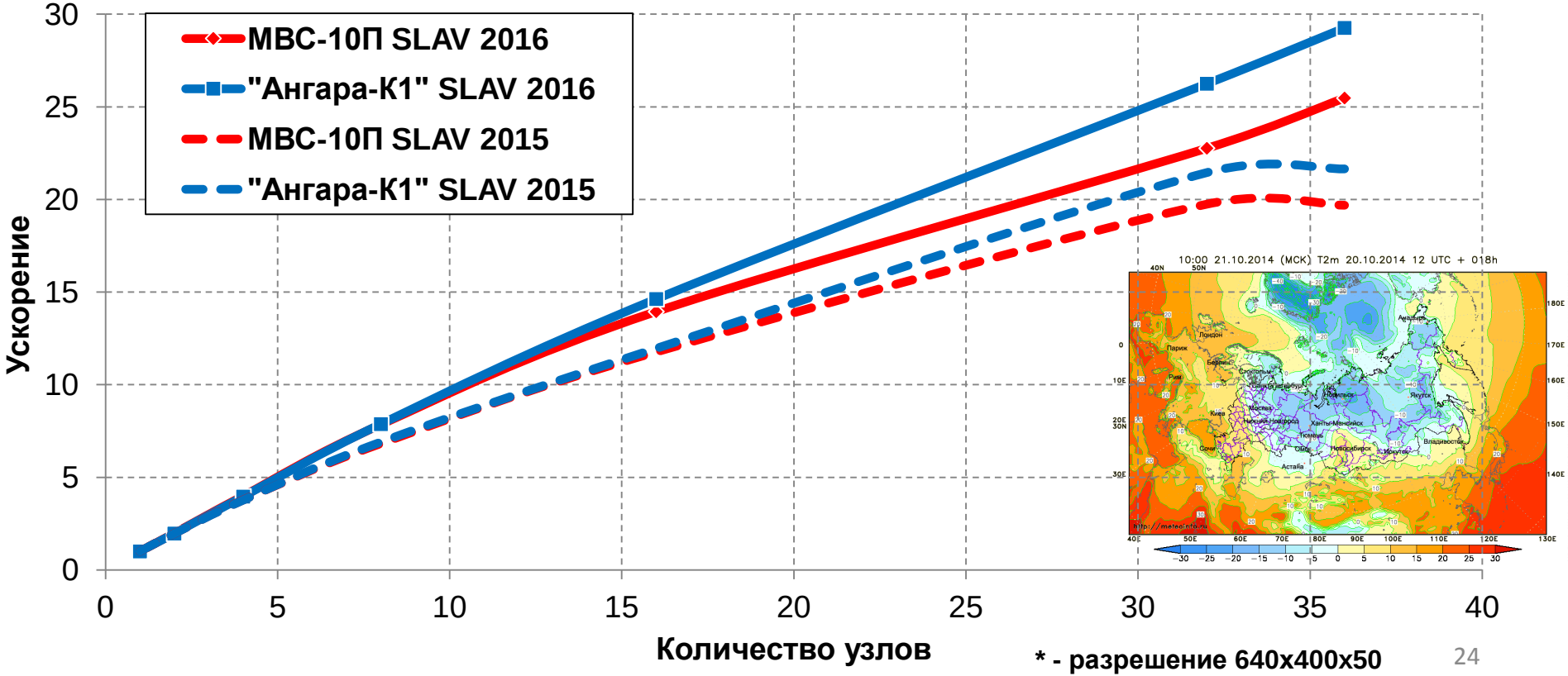




(1) Модель прогноза погоды ПЛАВ*.д.ф,-м.н. М.А.Толстых, ИВМ РАН, Гидрометцентр

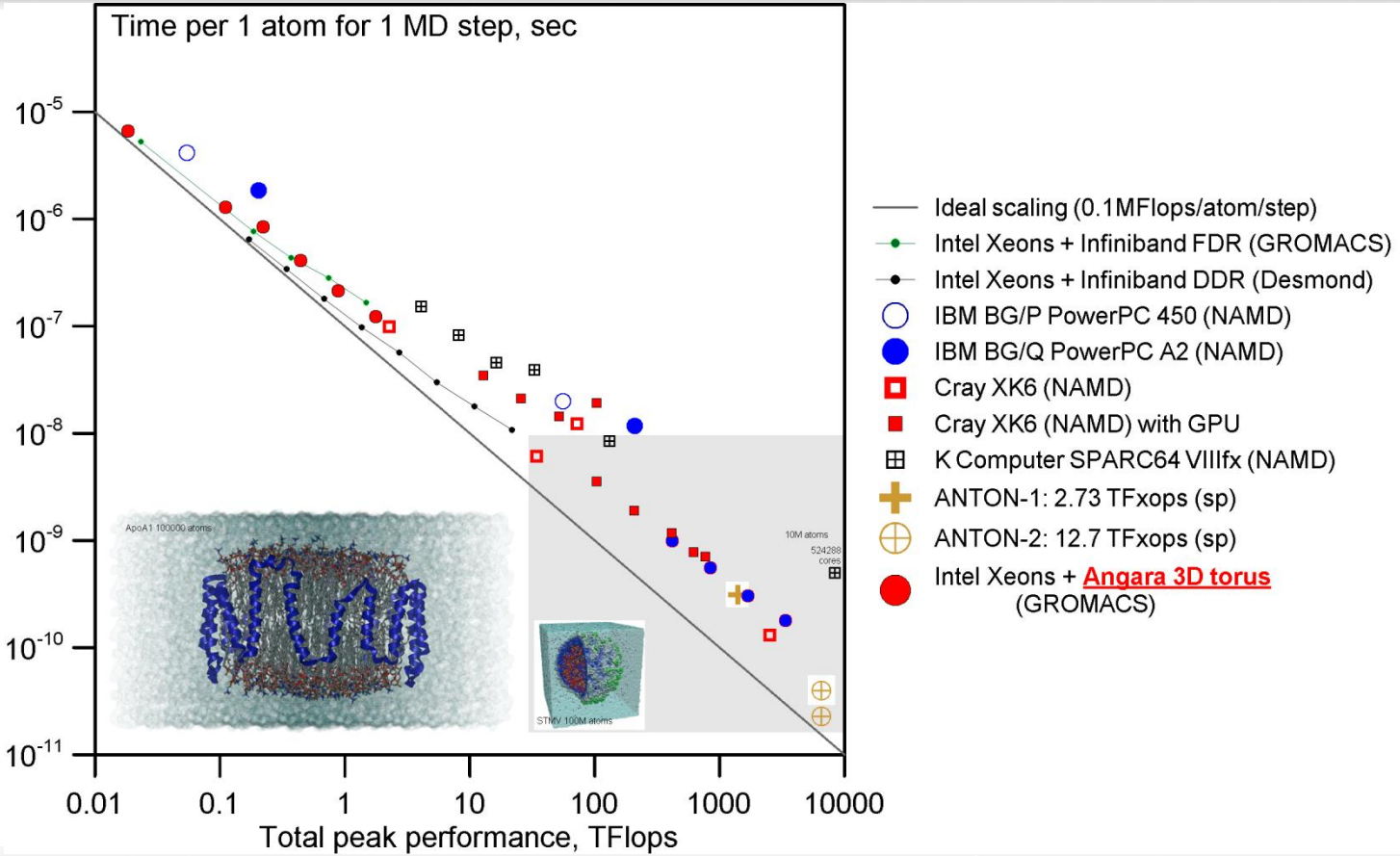
НИЦ ВТ

Ростех Объединенная приборостроительная корпорация



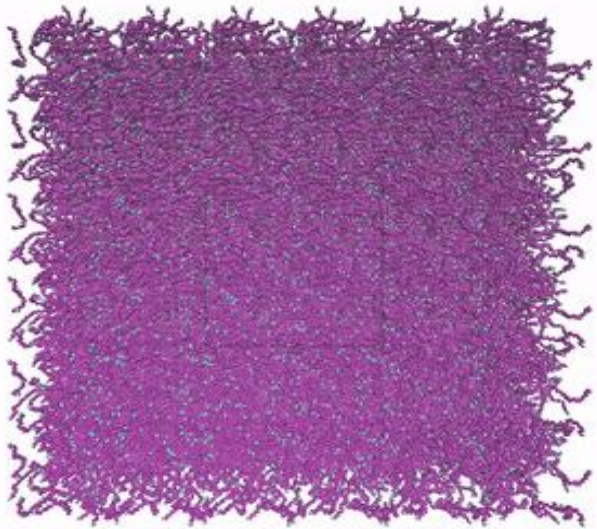
* - разрешение 640x400x50

(2) Молекулярная динамика. Модель белка, GROMACS. д.ф.-м.н. В.В. Стегайлов, ОИВТ РАН

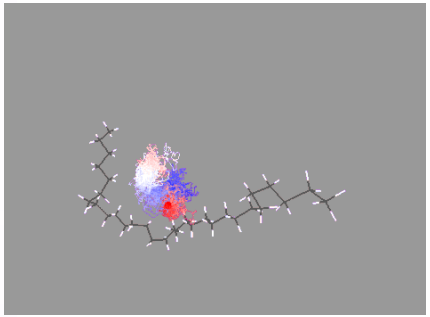


(3) Молекулярная динамика. Исследование свойств жидких углеводородов, LAMMPS.

д.ф.-м.н. В.В. Стегайлов, ОИВТ РАН



н-триаконтановая жидкость
T = 350 ÷ 490 K ; P = 1 атм
Количество молекул ~ 4 000



Траектория 1-й молекулы в исследуемой жидкости

Диффузия, вязкость жидких углеводородов, т.к. они входят в состав

трансформаторных масел,

топлив и смазочных материалов

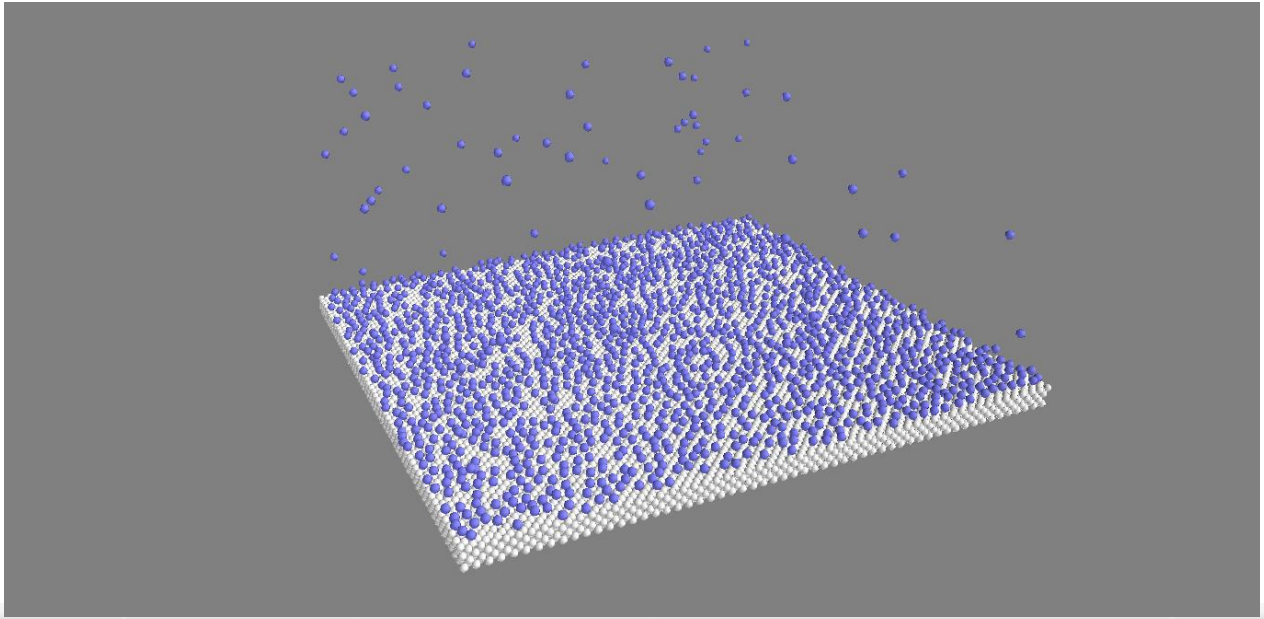
Молекулярная динамика-> макроскопические свойства

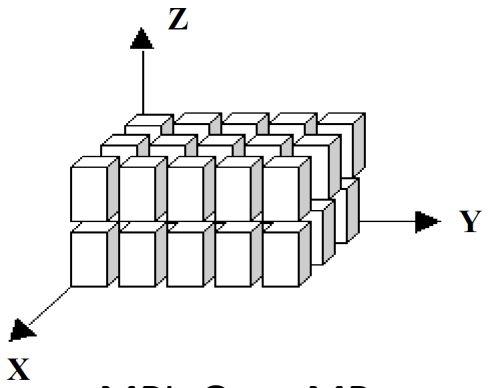
LAMMPS, 30 July 2016

(4) Моделирование термодинамического равновесия в системах газ-металл методами молекулярной динамики. д.ф.-м.н. С.В. Поляков, ИПМ РАН

Расчет по взаимодействию азота со стенками никелевого микроканала

Число частиц: 8 128 512 + 423 840 = 8 552 352,
Температура термостатов: T_{Ni} = 273.15 K, T_{N2} = 273.15 K
Число шагов по времени: 2 000 000 шагов, 1 шаг = 2 фс
Размер системы: 102x102x1534 нм^3



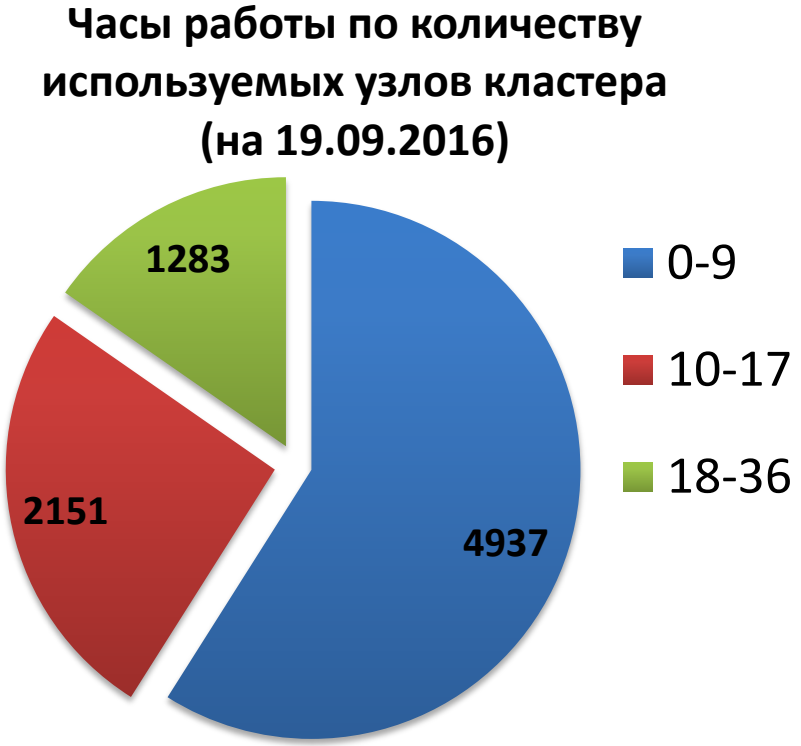


MPI+OpenMP

Фрагмент распределения молекул азота (область 20x20 нм) на поверхности никелевой пластины, в момент времени 2.3 нс

Статистика использования кластера «Ангара-К1»

- В режиме внешнего доступа работает с 1.11.2015 г.
- Число пользователей: 28
- Всего использовано 183778 процессоро-часов



Содержание

1. Внедрение в коммерческий сегмент (вычислительные кластеры для науки, образовательных учреждений, промышленных предприятий, создаваемые на основе коммерчески доступных комплектующих зарубежного производства) осуществляется в партнерстве с компаниями - интеграторами путем поставки им сетевого оборудования Ангара в стандартных форм-факторах и формирования для них гибкой ценовой политики

2. Внедрение в сегмент заказных разработок осуществляется путем работы напрямую с Заказчиками и создания под них специализированных версий сетевого оборудования Ангара или вычислительных систем на его основе

Партнеры по созданию вычислительных систем

РСК

NIAGARA
Российские Суперкомпьютеры

Ф-ПЛАТФОРМЫ

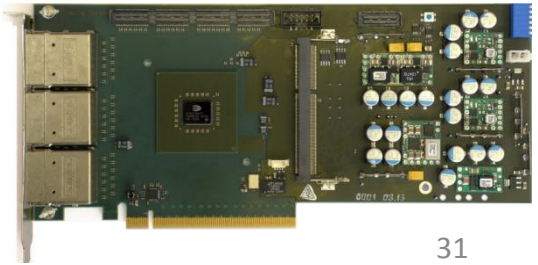
inline group

СТОРУС

ОТКРЫТЫЕ ТЕХНОЛОГИИ

ИНФОРМ Ростех

МЦСТ ЭЛЬБРУС



Типовые конфигурации вычислительных систем, оптимизированных по стоимости на узел

НИЦЭВТ

Ростех Объединенная приборостроительная корпорация

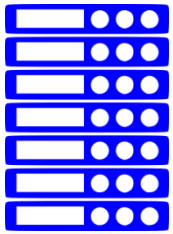
Кластер на базе ПК

До 16 узлов

CPU	Xeon 1600 v3/4
HDD	500 ГБ
ОЗУ	8-64 ГБ
GPU*	NVidia GTX™/ AMD Radeon™
Сеть	Ангара 2D-тор

Кластер: 1U сервера

1 стойка

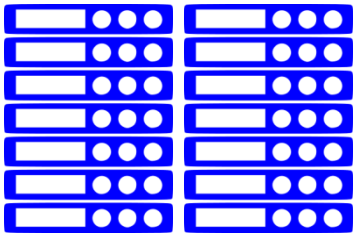


до 32 узлов

CPU	Xeon 1600 v3/4
HDD/SSD*	1 ТБ
ОЗУ	16-256 ГБ
GPU*	NVidia Tesla™/ AMD FirePro™
Сеть	Ангара 3D-тор

Кластер: 1U сервера

До 8 стоек



32-256 узлов

CPU	Xeon 1600 v3/4
HDD/SSD*	1 ТБ
ОЗУ	64-256 ГБ
GPU*	NVidia Tesla™/ AMD FirePro™
Сеть	Ангара 4D-тор

*Опционально

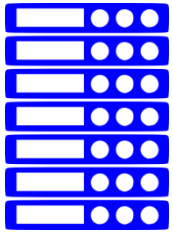
Типовые конфигурации вычислительных систем, оптимизированных по стоимости (руб./Тфлопс)

НИЦЭВТ

Ростех Объединенная приборостроительная корпорация

Кластер: 1U сервера

1 стойка

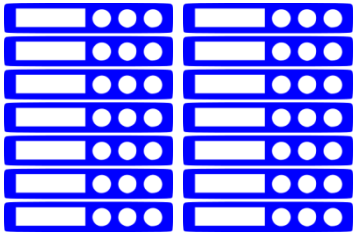


до 32 узлов

CPU	2x Xeon 2600 v3/4
HDD/SSD*	1 ТБ
ОЗУ	16-256 ГБ
GPU*	NVidia Tesla™/ AMD FirePro™
Сеть	Ангара 3D/4D-тор

Кластер: 1U сервера

До 8 стоек



32-256 узлов

CPU	2x Xeon 2600 v3/4
HDD/SSD*	1 ТБ
ОЗУ	64-256 ГБ
GPU*	NVidia Tesla™/ AMD FirePro™
Сеть	Ангара 4D-тор

*Опционально

Вычислительный кластер с на основе серверов x86-64 высотой 1U

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»

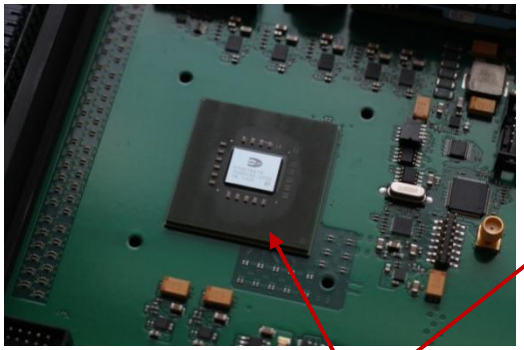
- Адаптер EC8430, топология 3D-тор 3x3x4
- Собственная реализация OpenSHMEM
- MPI: MPICH 3.0.4

Операционная система

- SLES 11 SP2, Linux 3.0.13-0.27-default
- GCC 4.3.4 (revision 152973)

Характеристики сервера на основе Вычислительной платформы ЕС1740.0003

Характеристика	ЕС1740.0003 (EATX)
Сокеты x86/ядра x86	2/32
Тип ОЗУ	4-х канальная DDR3 1866/1600/1333/1066 МГц
Объем ОЗУ	128 Гбайт DDR3
Модули ОЗУ	8 шт.
Интерфейсы	Ангара 2x10G Ethernet 4xSATA RAID
Накопителей	4 шт.
Потребляемая мощность	400 Вт
Габаритные размеры	321 x 367 x 30 мм



Совместные работы АО «НИЦЭВТ» и АО «МЦСТ»

	Испытательный стенд
Узлы	На основе процессоров Эльбрус-4С, частота 750 МГц, и КПИ1
Количество узлов	4
Память узла	24
Сеть	Ангара 1D-тор
HPL*	75.09 Гфлопс (78.2%)

* - предварительный результат

GRAPH HPC 2017

Научно-практическая конференция

Технологии параллельной обработки больших графов

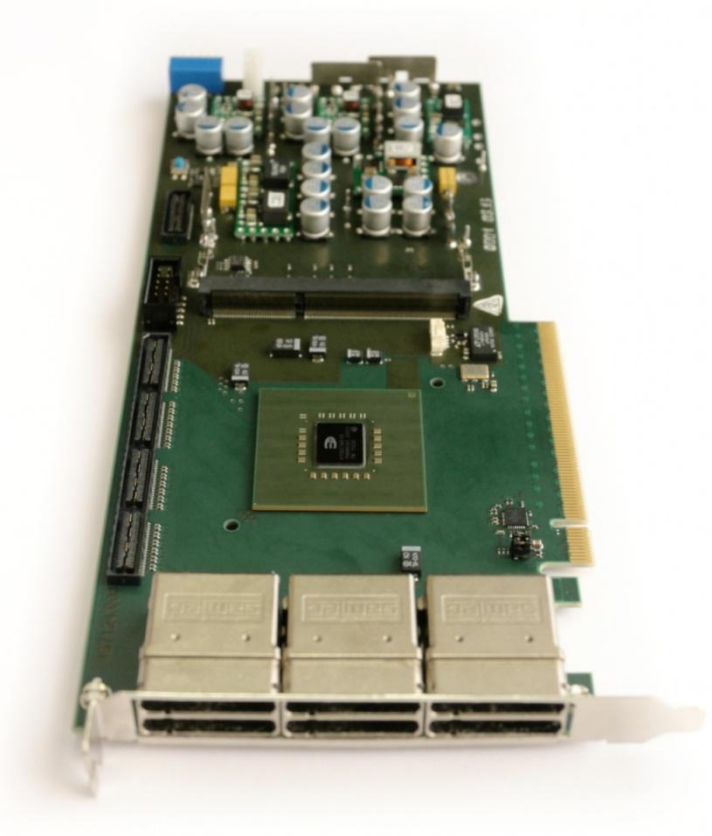
2 марта 2017, Москва

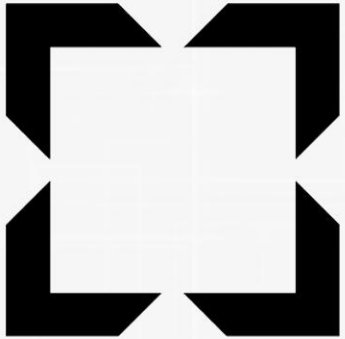
- Конференция GraphHPC: 2014, 2015, 2016, 2017 годы
- Конкурс с автоматической системой проведения
- Доступ на кластер с сетью Ангара
- Сайт конференции graphhpc.dislab.org
- Открытое ПО

Контакты:

117587, Москва, Варшавское ш, 125

angara@nicevt.ru





Ростех

Объединенная
приборостроительная
корпорация