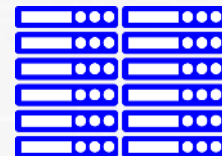


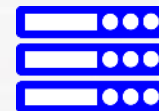
Niagara & Angara: Interconnect solution

Назначение:

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

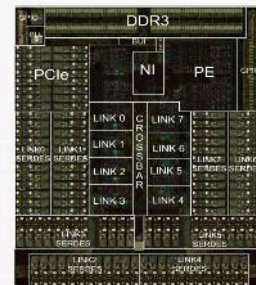
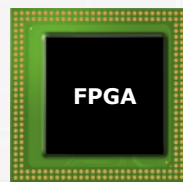
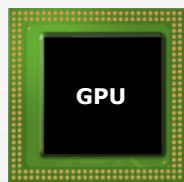
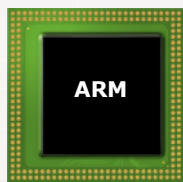
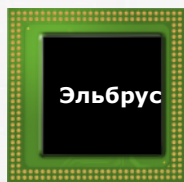
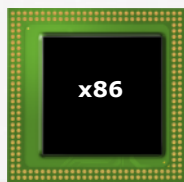
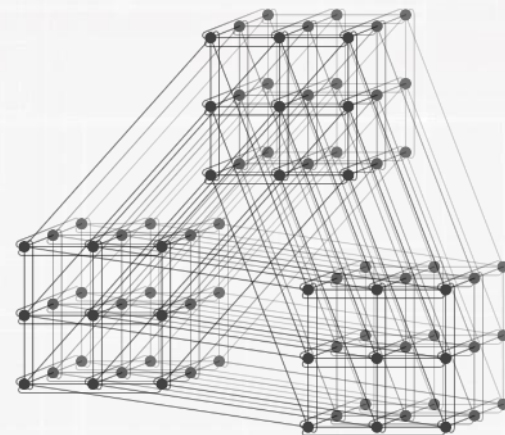
**Области применения:**

1. Вычислительные кластеры для расчетно-информационных задач, математического моделирования и виртуального прототипирования, решения задач инженерного анализа
2. Системы хранения и обработки Больших Данных
3. В качестве коммуникационной сети вычислительного поля в центрах обработки данных (ЦОД)



Ключевые особенности:

- Топология сети: 1D..4D-тор
- До 8 каналов связи с соседними узлами
- Прямой доступ в память удаленного узла (RDMA)
- Прямой доступ в память GPU (GPUDirect)
- Адаптивная передача пакетов
- Задержка на MPI ring-pong: 0,85/ 1,54 мкс (x86/Эльбрус-8С)
- Задержка на хоп: 130 нс
- Масштабирование: до 32К узлов
- Энергопотребление до 20 Вт
- Различные физические среды передачи данных



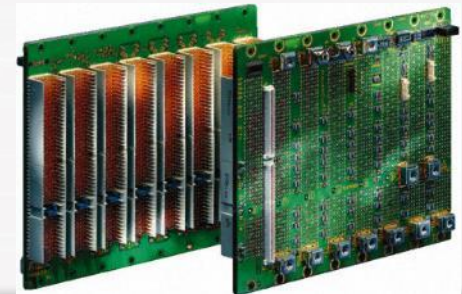
1. Высокопроизводительное решение на базе FHFL адаптера и Samtec кабеля





2. Универсальное решение на базе 24-портового коммутатора, low-profile адаптера и CXР кабеля



3. Заказное решение на базе объединительной платы и оптических кабелей




 Kernel-space

 User-space

* В стадии разработки/отладки

- Поддержка ОС : Astra Linux SE 1.3 – 1.5, ОС «Эльбрус», OpenSUSE/SLES 11 SP3/4, CentOS 6.0-7.3, ОС «Нейтрино» 6.5, Версия ядра Linux от 2.6.21 до 3.16.0
- Поддержка компиляторов языков Fortran 77/90/95 (GNU, Intel), C/C++ (GNU, Intel)

- **Ангара-К1: 36 вычислительных узлов (2014)**
 - 12 узлов с 1 процессором Intel Xeon E5-2660 (8 ядер, 2.2 ГГц)
 - 24 узла с 2 процессорами Xeon E5-2630 (6 ядер, 2.3 ГГц)
 - 64 ГБ
 - 3D-тор 4x3x3
 - Удаленный доступ (более 40 сторонних пользователей)
- **ОИВТ РАН: 32 вычислительных узла (4 кв. 2016)**
 - 1 процессор Intel Xeon E5-1650 v3 (6 ядер, 3.0 ГГц)
 - Nvidia GeForce GTX 1070
 - DDR4 16 ГБ
 - 4D-тор 4x2x2x2



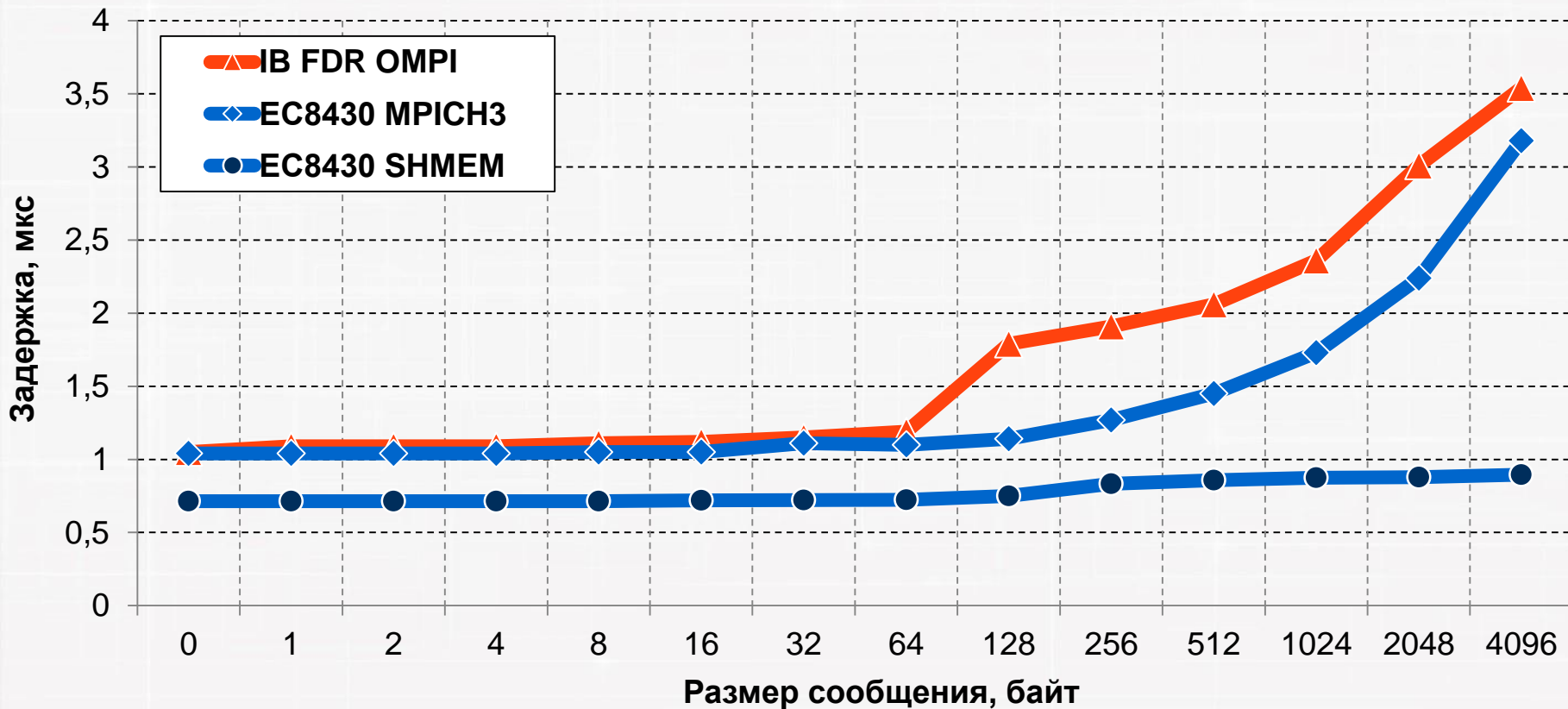


ИВМ РАН

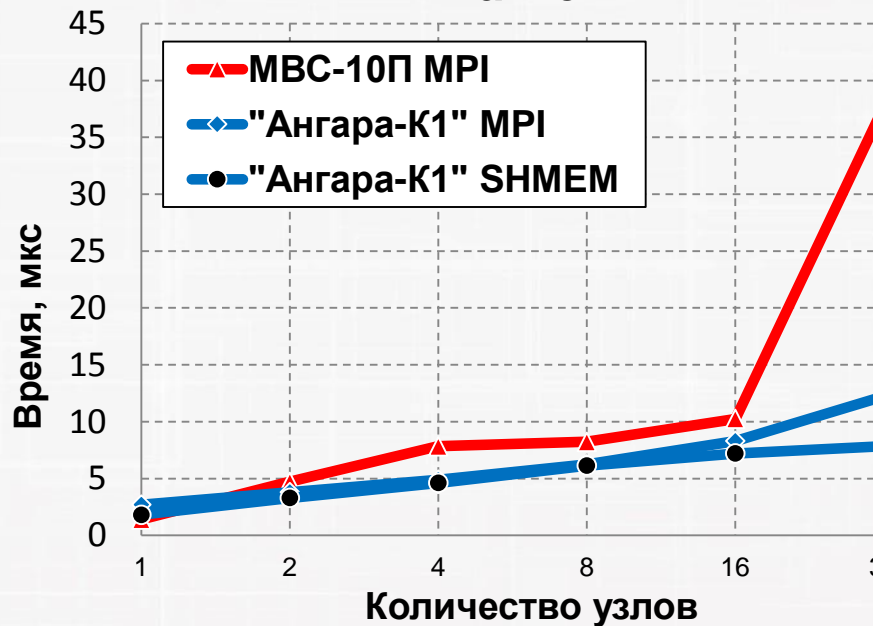
ИДСТУ СО РАН

Оценочное тестирование

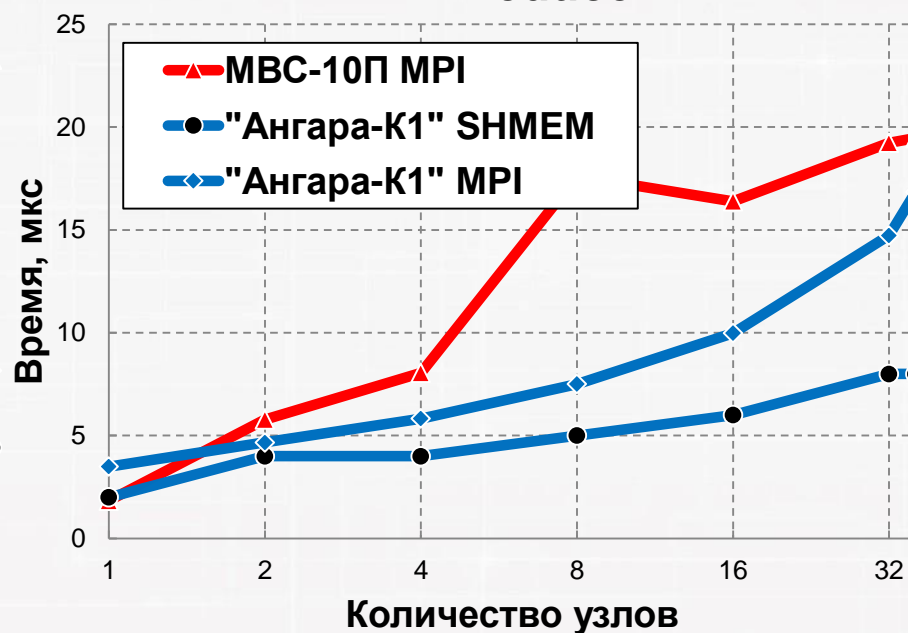
	Ангара-К1		МВС-10П
Узлы	А	2x Xeon E5-2630 по 6 ядер, 2.3 ГГц	2x Xeon E5-2690 по 8 ядер, 2.9 ГГц
	В	Xeon E5-2660 по 8 ядер, 2.2 ГГц	
Количество узлов	24*А+12*В = 36		207 (36)
Память узла	64 ГБ		64 ГБ
Сеть	Ангара 3D-топ 3x3x4		Infiniband 4xFDR Fat Tree



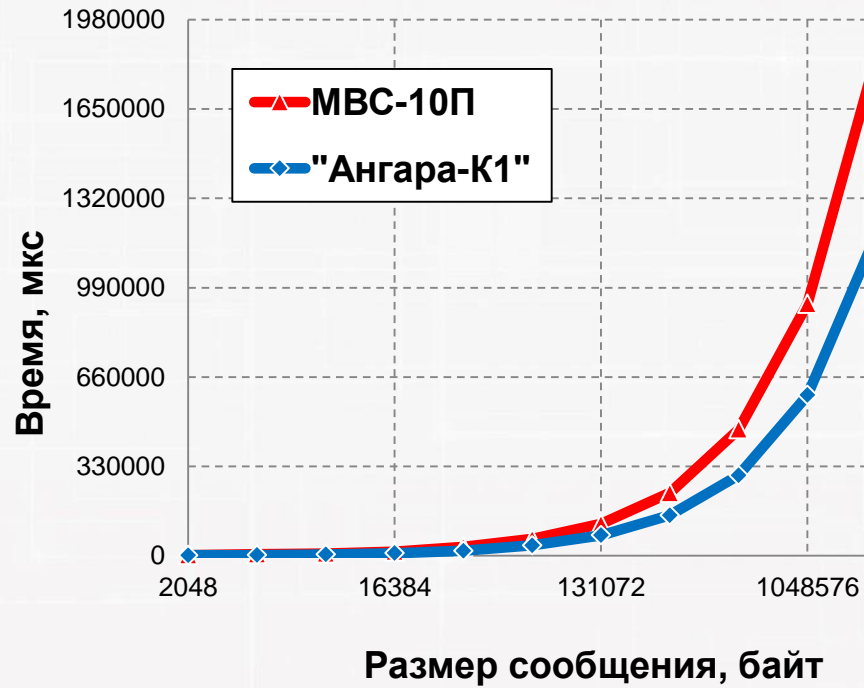
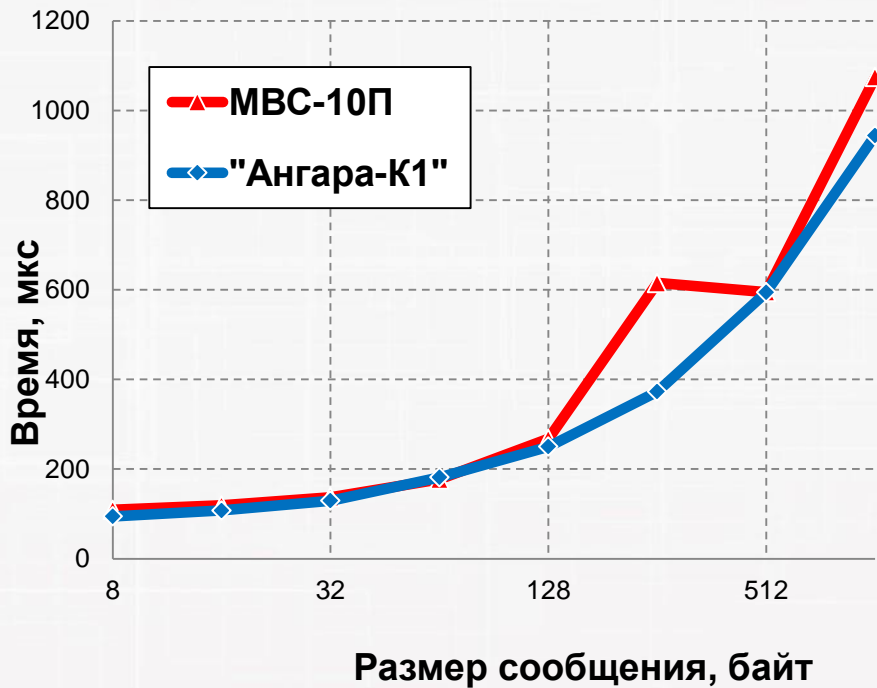
IMB Barrier



IMB Allreduce

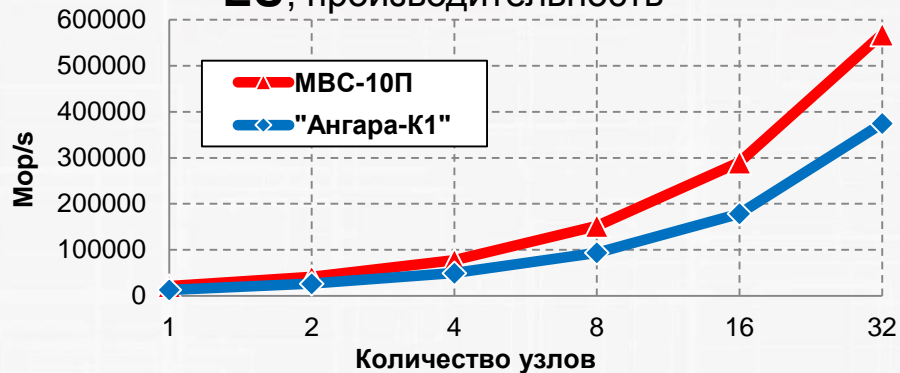


размер сообщения 8 байт

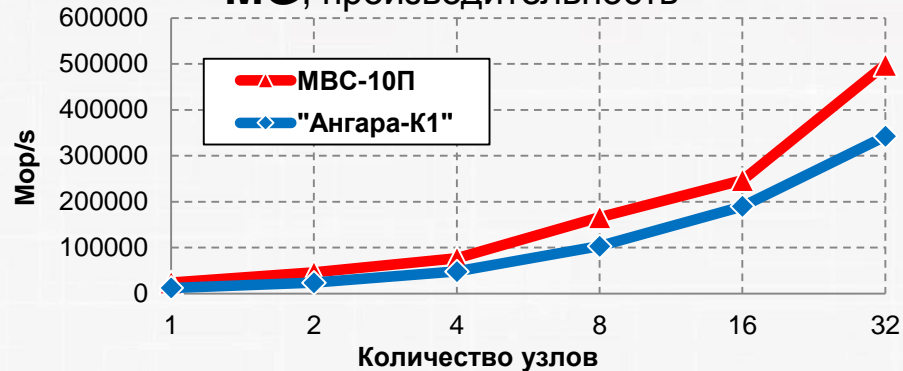


		Ангара	МВС-10П
HPL	Тфлопс	4.44	—
	% пиковой	85 %	—
HPCG MPI	Гфлопс	279	363
	% пиковой	5.3 %	5.4 %
HPCG SHMEM	Гфлопс	342	—
	% пиковой	6.5 %	—

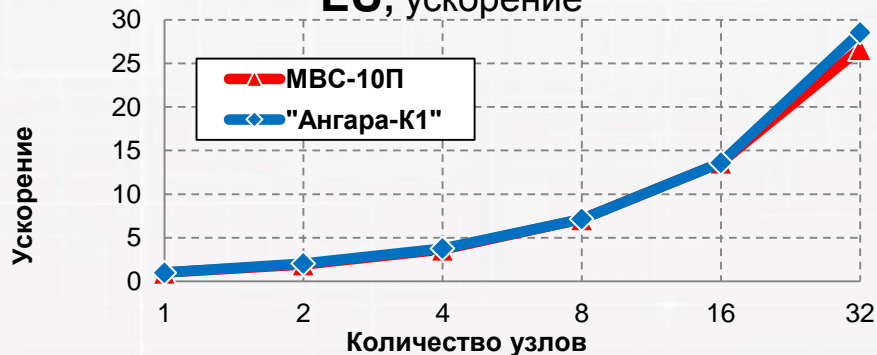
LU, производительность



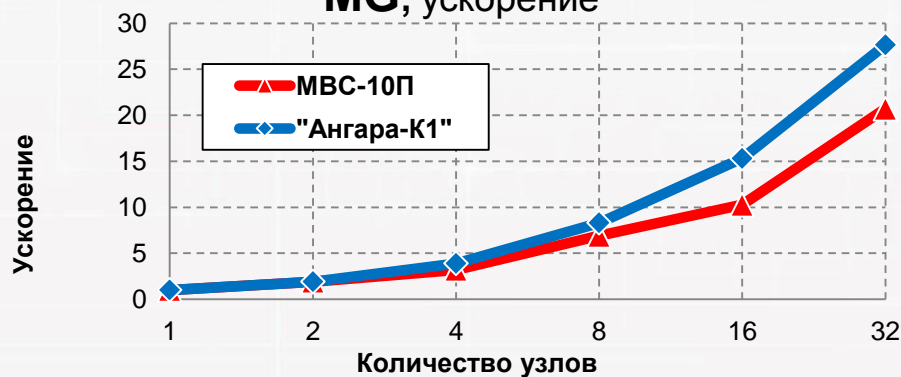
MG, производительность

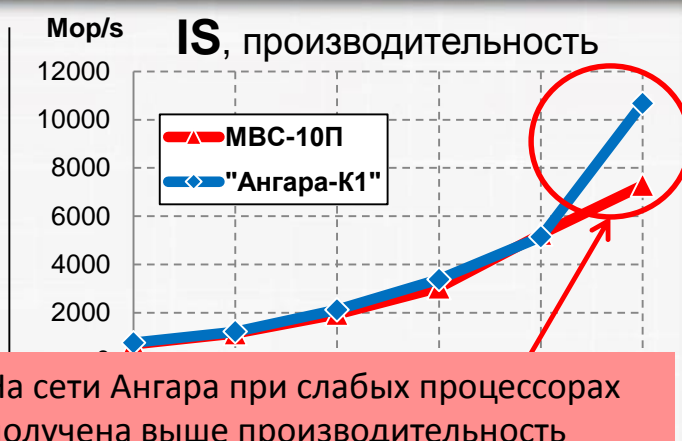
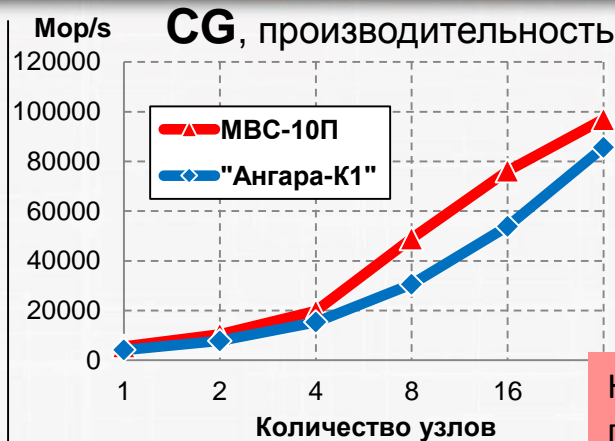
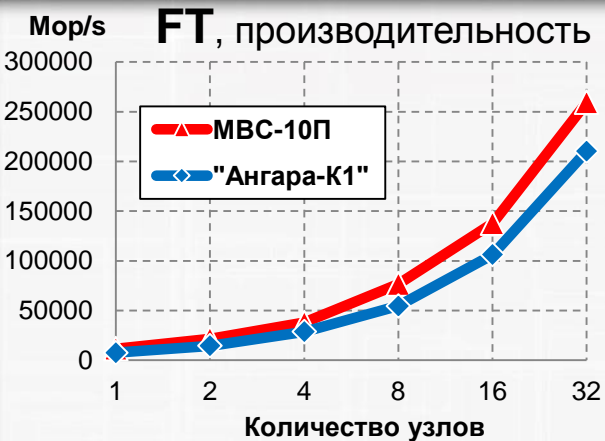


LU, ускорение

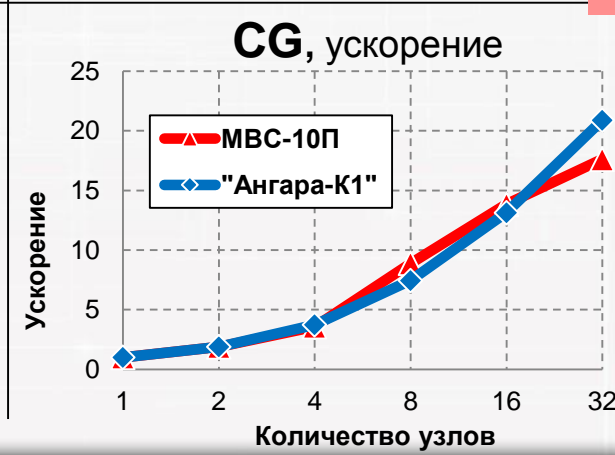
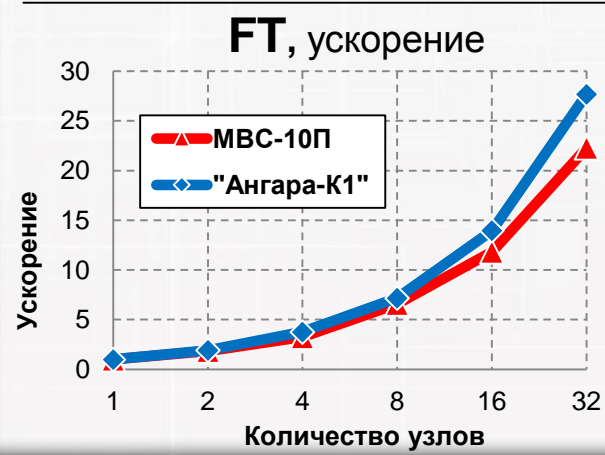


MG, ускорение





На сети Ангара при слабых процессорах получена выше производительность





	Desmos	Политехник
Узлы	1x Xeon E5-1650v3 6 ядер, 3.0 ГГц	2x Xeon E5-2697 v3 по 14 ядер, 2.6 ГГц
Количество узлов	32	207 (36)
Память узла	8 ГБ	64 ГБ
Ускоритель	Nvidia GeForce GTX 1070	
Сеть	Ангара 4D-топ 4x2x2x2	Infiniband 4xFDR Fat Tree 2:1
Компилятор	Intel Parallel Studio XE 2017	Intel Parallel Studio XE 2016



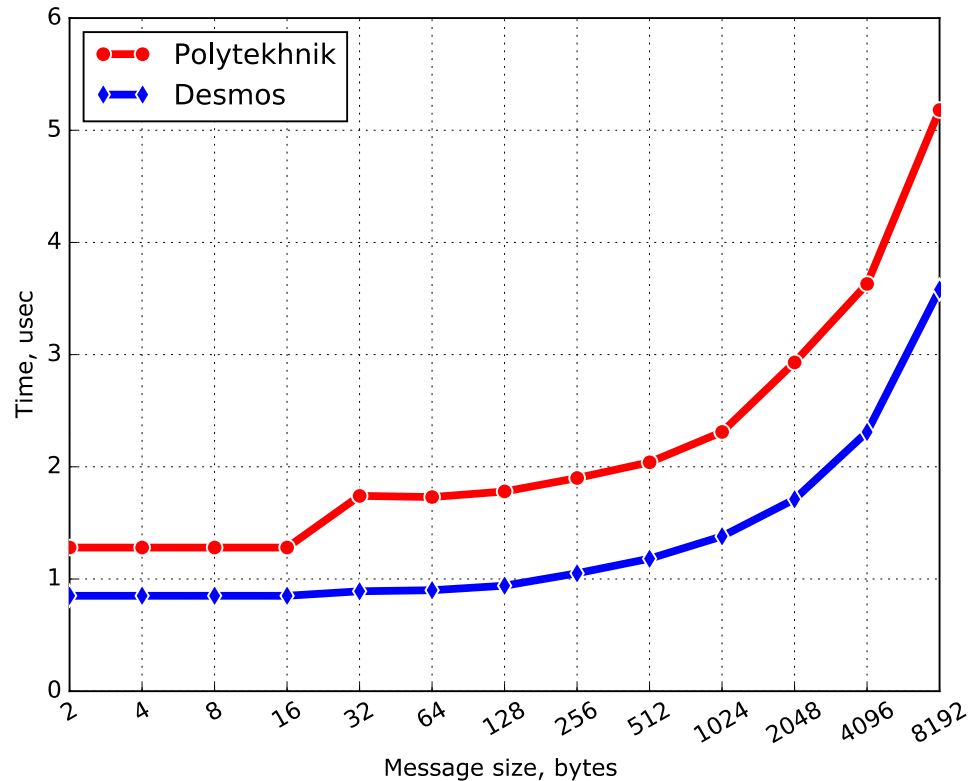
Desmos vs Политехник

- задержка на MPI между двумя узлами (osu_latency)

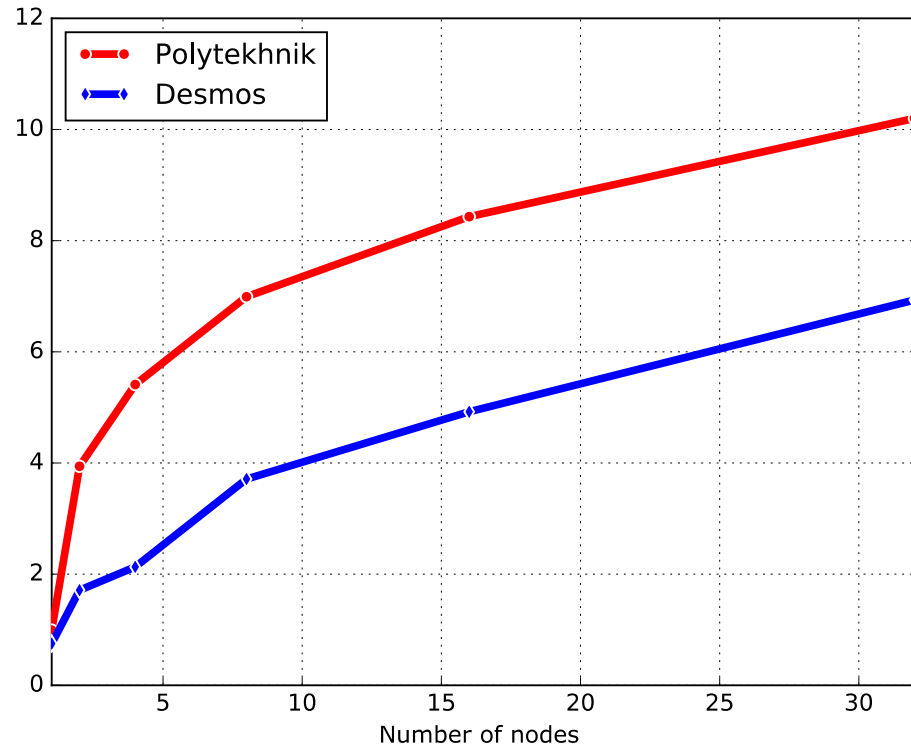
- время выполнения MPI_Barrier



osu_latency



barrier, ppn=4

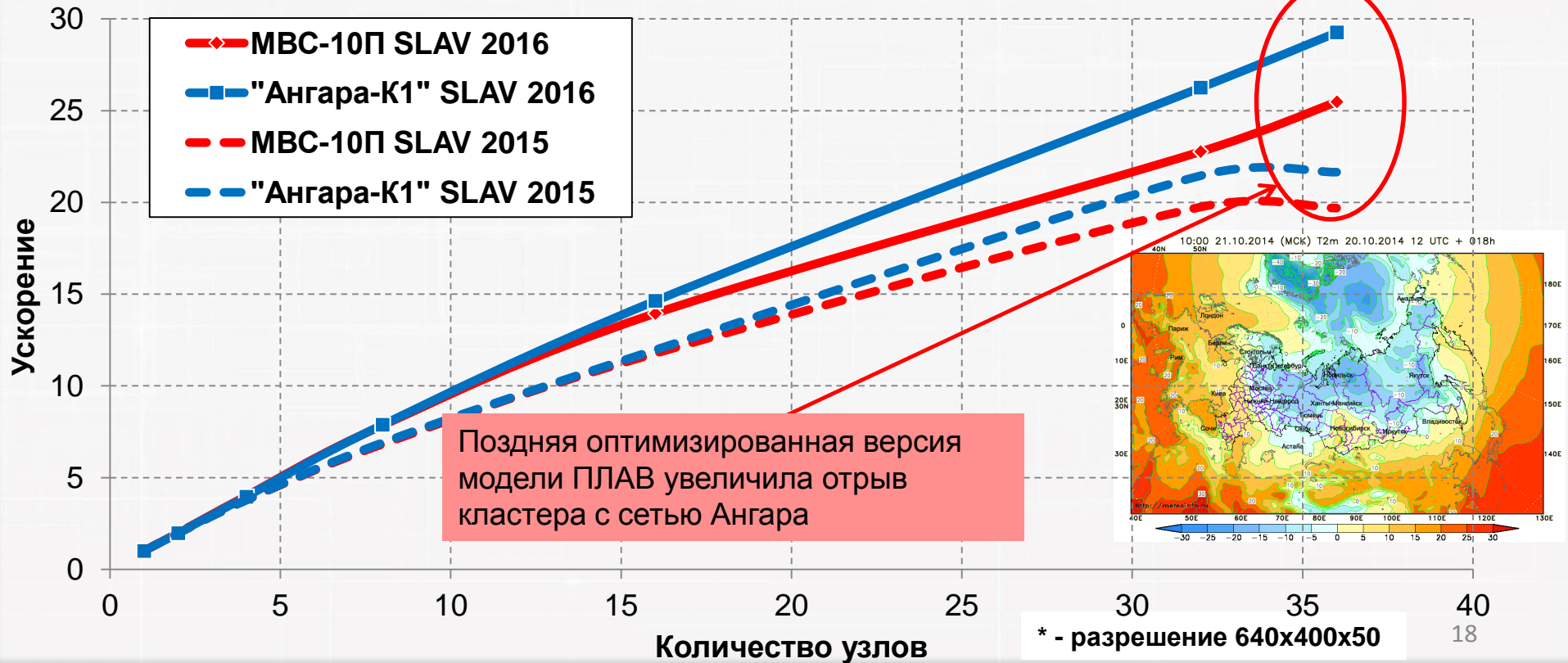




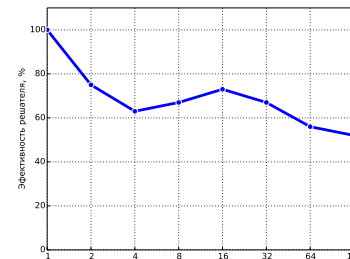
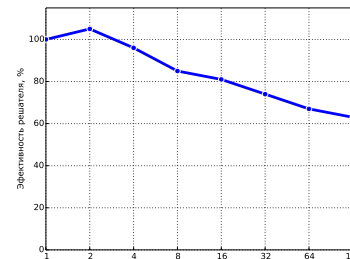
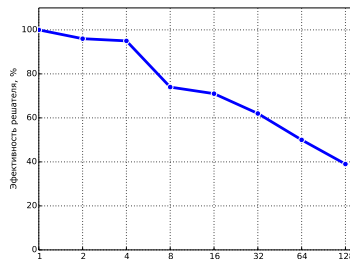
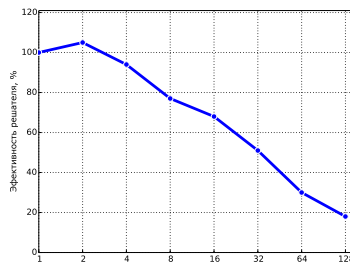
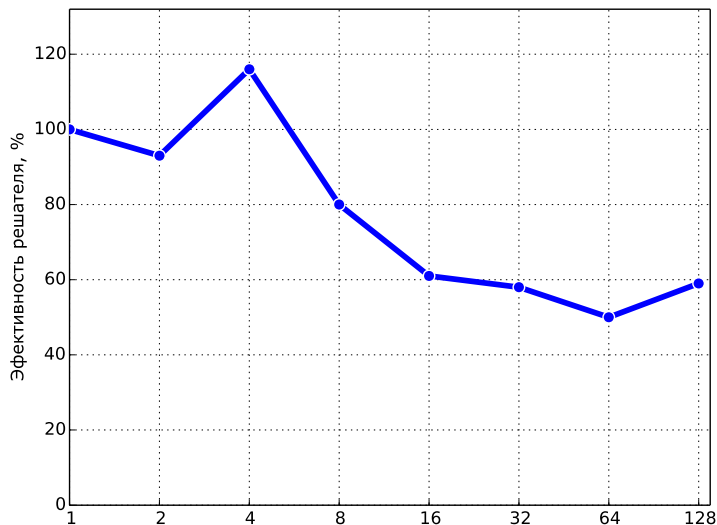
Коммуникационная сеть Ангара

Модель прогноза погоды ПЛАВ*.

д.ф,-м.н. М.А.Толстых, ИВМ РАН, Гидрометцентр



Инженерные пакеты

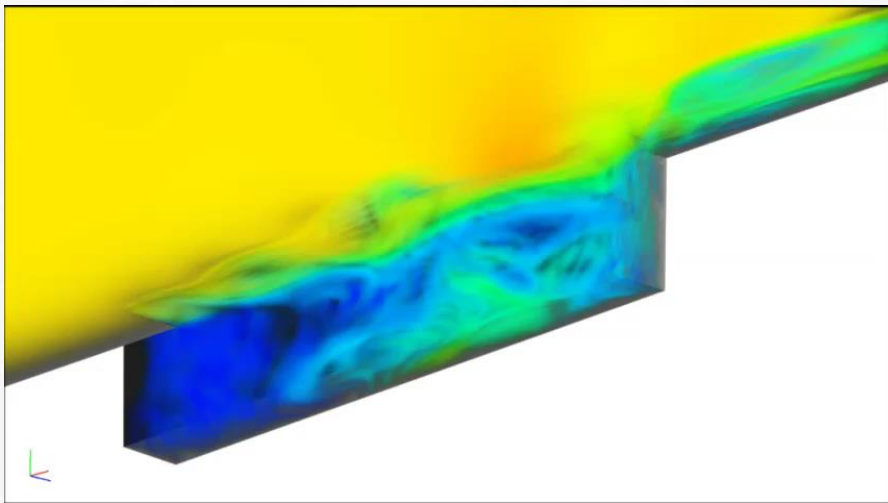


Ю. Новожилов. Тестирование работы программного обеспечения ANSYS на кластерах с отечественным высокопроизводительным интерконнектом Ангара. Международная конференция Суперкомпьютерные дни в России, 2017.

Ю. Новожилов. Работа решателей ANSYS на российском интерконнекте Ангара. XIV конференция пользователей CADFEM/ANSYS, 31 октября, 2017.

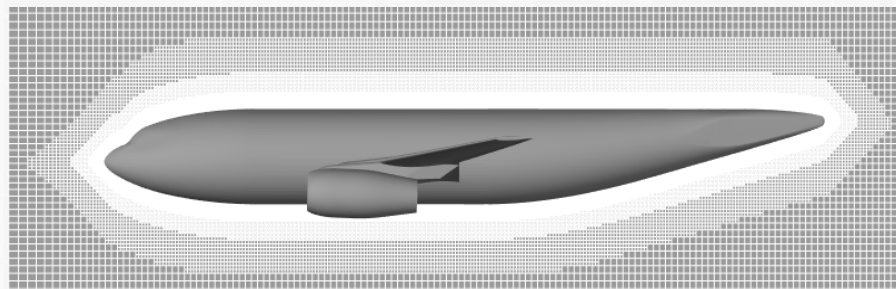
M219 Cavity case

Обтекание каверны воздухом, 5.5 млн ячеек



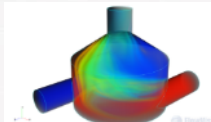
Объемная визуализация скорости

Неоднородная сетка
Основная – 17.5 млн. ячеек,
Приповерхностная – 9.3 млн. ячеек
(всего – 26.8 млн. ячеек)



Задача Смеситель, 260 тыс. ячеек

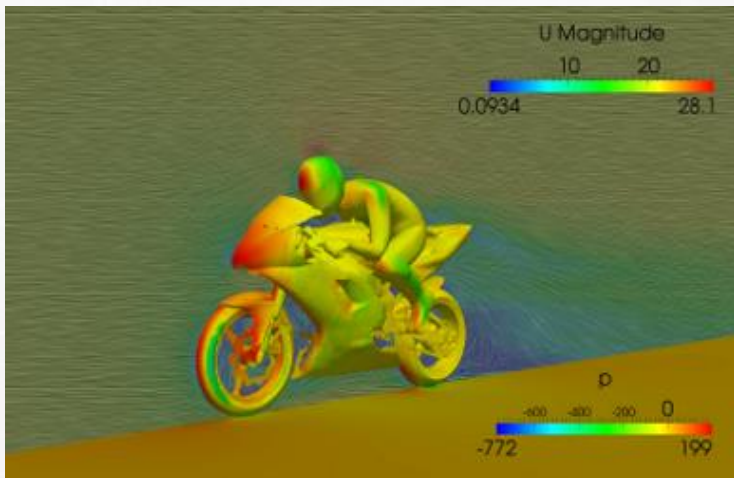
Распределение температуры



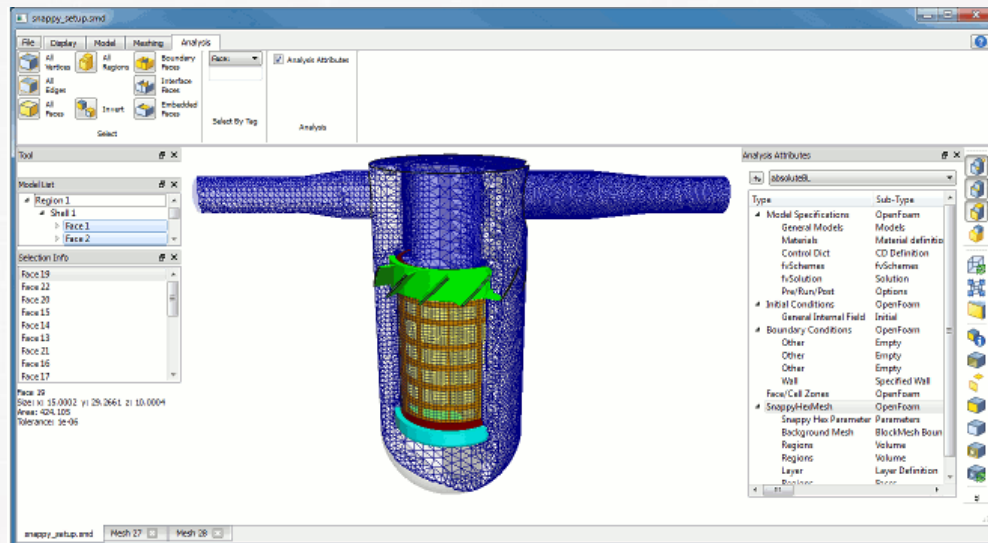
В. Акимов. Исследование масштабируемости FlowVision на кластере с сетью Ангара. Ряд докладов на международных и российских конференциях.

Open FOAM

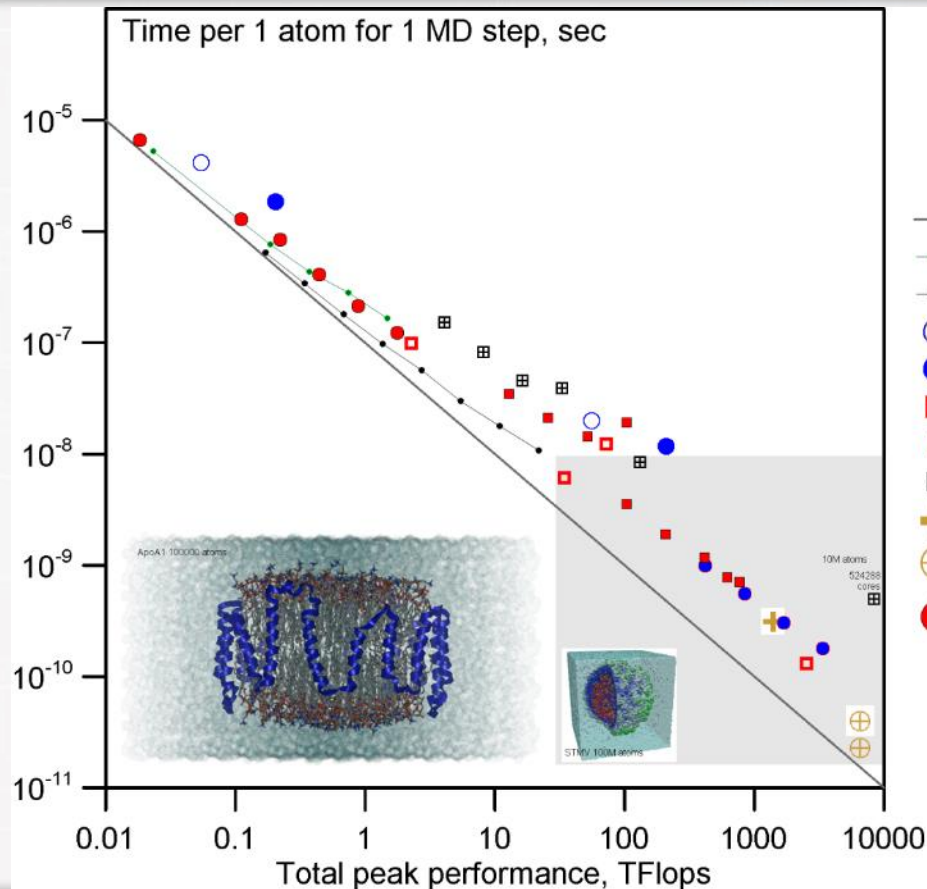
The Open Source CFD Toolbox



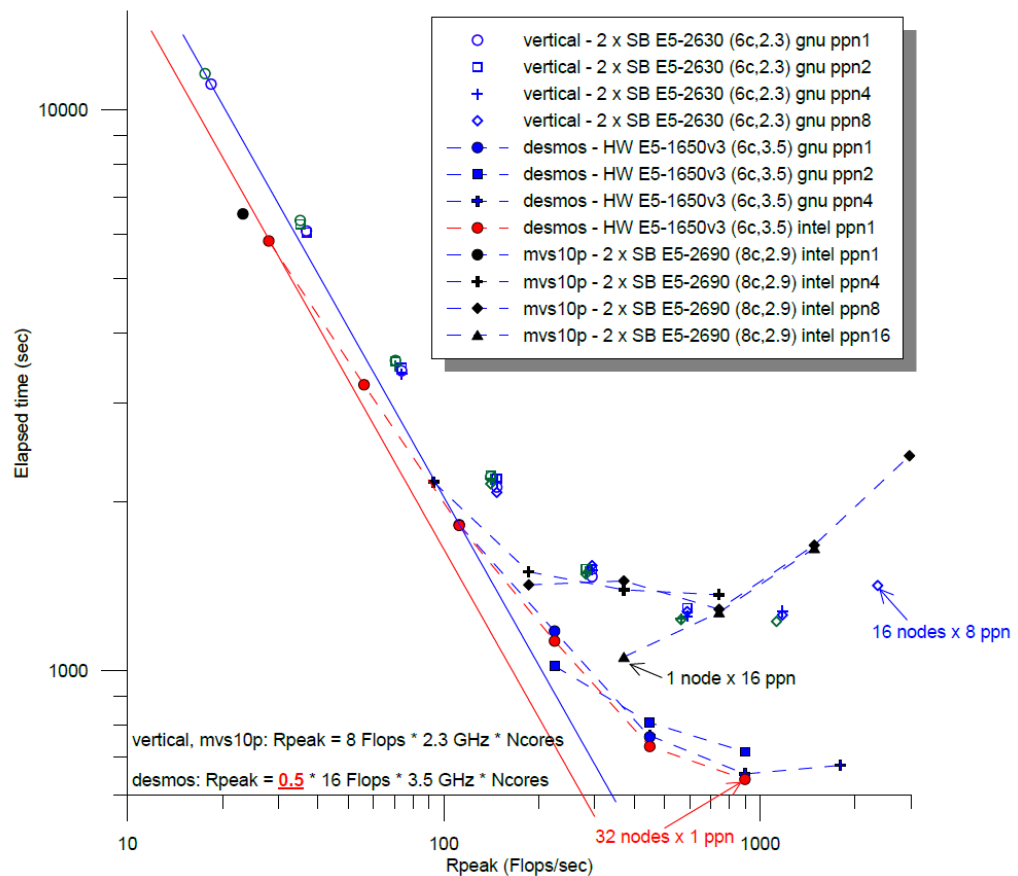
Версия 3.0.0

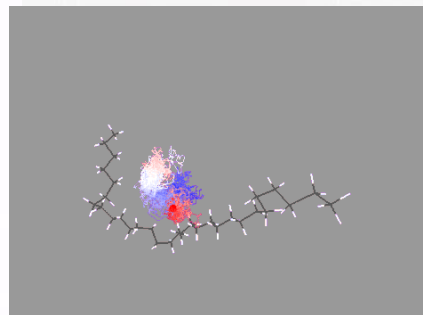
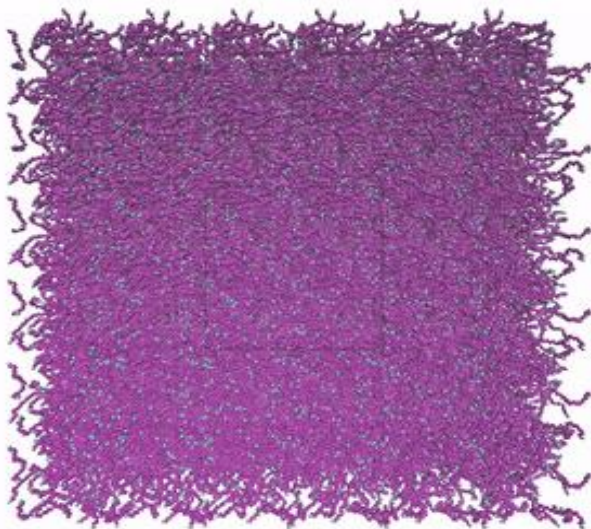


Приложения



- Ideal scaling (0.1 MFlops/atom/step)
- Intel Xeons + Infiniband FDR (GROMACS)
- Intel Xeons + Infiniband DDR (Desmond)
- IBM BG/P PowerPC 450 (NAMD)
- IBM BG/Q PowerPC A2 (NAMD)
- Cray XK6 (NAMD)
- Cray XK6 (NAMD) with GPU
- ▣ K Computer SPARC64 VIIIfx (NAMD)
- ⊕ ANTON-1: 2.73 TFxops (sp)
- ⊕ ANTON-2: 12.7 TFxops (sp)
- Intel Xeons + **Angara 3D torus** (GROMACS)





Траектория 1-й молекулы
в исследуемой жидкости

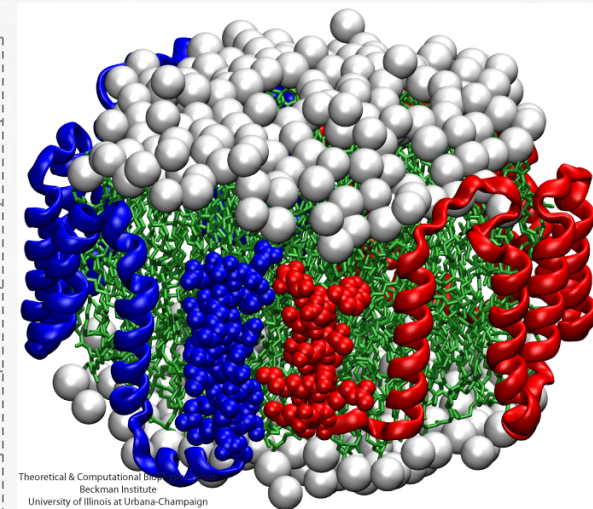
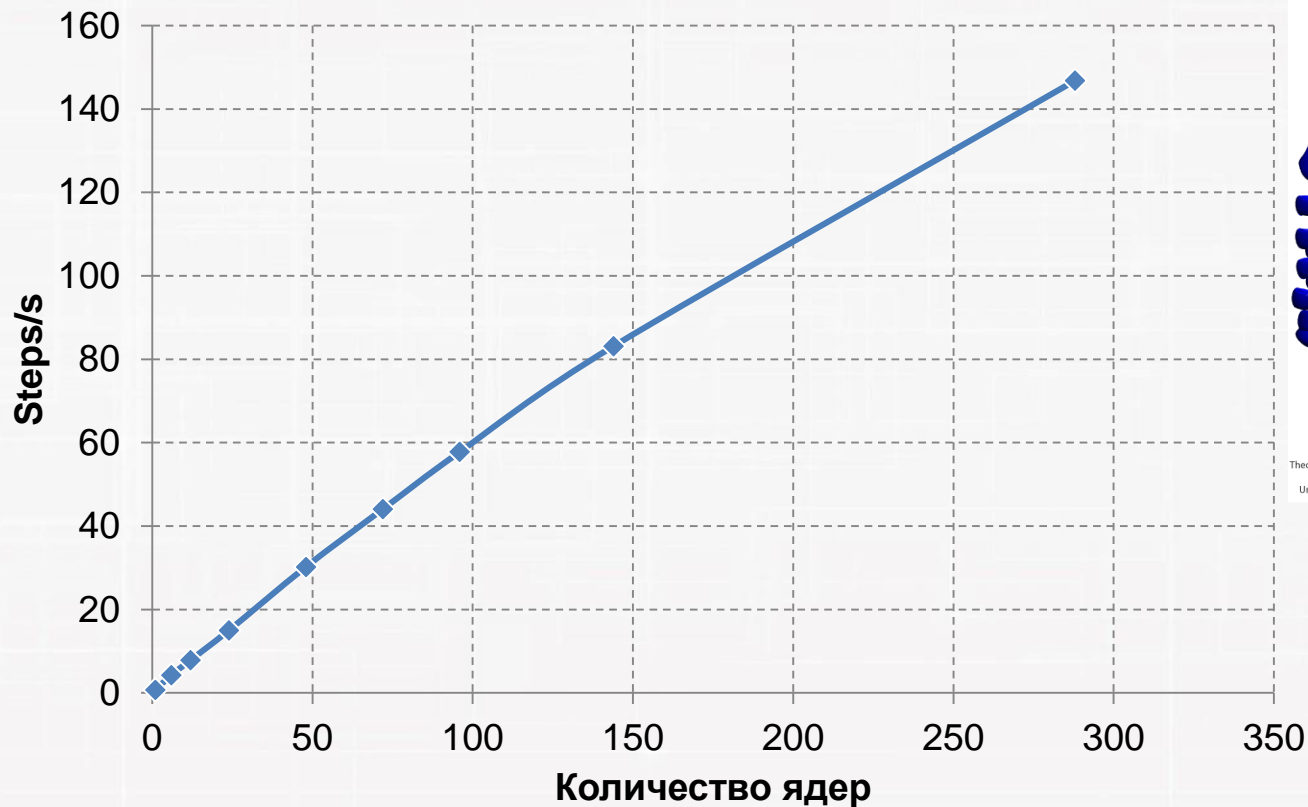
н-триаконтановая жидкость
 $T = 350 \div 490 \text{ K}$; $P = 1 \text{ атм}$
Количество молекул $\sim 4\ 000$

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

трансформаторных масел,
топлив и смазочных материалов

Молекулярная динамика-> **макроскопические свойства**

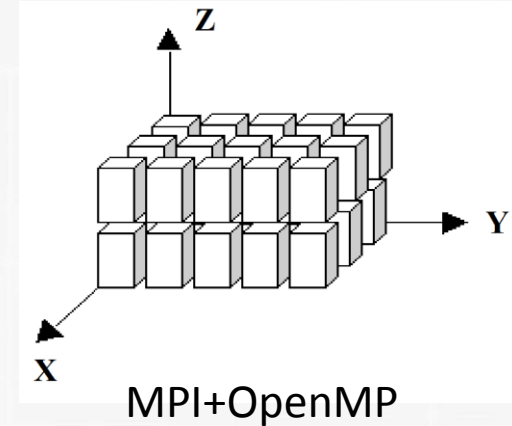
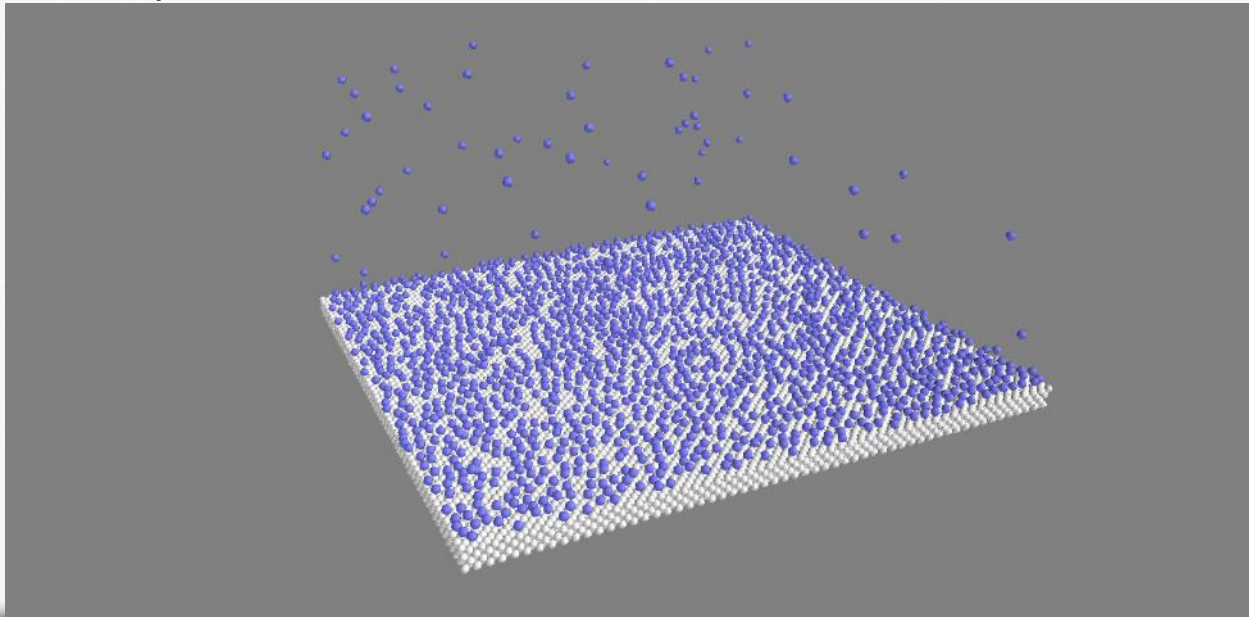
LAMMPS, 30 July 2016



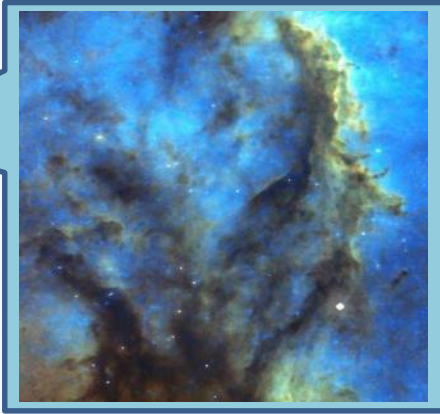
Расчет по взаимодействию азота со стенками никелевого микроканала

Число частиц: $8\,128\,512 + 423\,840 = 8\,552\,352$,Температура термостатов: $T_{Ni} = 273.15\text{ K}$, $T_{N_2} = 273.15\text{ K}$

Число шагов по времени: 2 000 000 шагов, 1 шаг = 2 фс

Размер системы: $102 \times 102 \times 1534\text{ нм}^3$ 

Фрагмент распределения
молекул азота (область
 $20 \times 20\text{ нм}$) на поверхности
никелевой пластины, в
момент времени 2.3 нс



The self-gravity magneto hydrodynamics equations

$$\frac{\partial}{\partial t} \begin{pmatrix} \rho \\ \rho_i \\ \rho \vec{v} \\ \rho E \\ \rho \varepsilon \end{pmatrix} + \nabla \cdot \begin{pmatrix} \rho \vec{v} \\ \rho_i \vec{v} \\ \rho \vec{v} \vec{v} \\ \rho E \vec{v} \\ \rho \varepsilon \vec{v} \end{pmatrix} = \begin{pmatrix} 0 \\ s_i \\ \nabla \cdot (\vec{B} \vec{B}) - \nabla p^* - \rho \nabla \Phi \\ -\nabla \cdot (p^* \vec{v} - \vec{B} (\vec{B}, \vec{v})) - (\rho \vec{v}, \nabla \Phi) - \Lambda + \Gamma \\ -(\gamma - 1) \rho \varepsilon \nabla \cdot \vec{v} - \Lambda + \Gamma \end{pmatrix}$$

$$\frac{\partial \vec{B}}{\partial t} = \nabla \times (\vec{v} \times \vec{B})$$

$$\nabla \cdot \vec{B} = 0$$

$$\Delta \Phi = 4\pi G \rho$$

$$\rho E = \rho \varepsilon + \frac{\rho v^2}{2} + \frac{B^2}{2}$$

$$p = (\gamma - 1) \rho \varepsilon$$

$$p^* = p + \frac{B^2}{2}$$

Технологии на базе сети Ангара

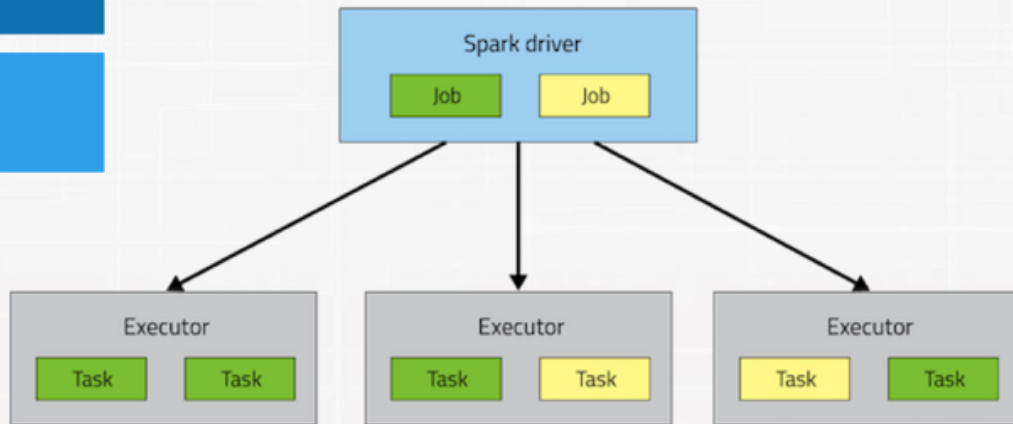
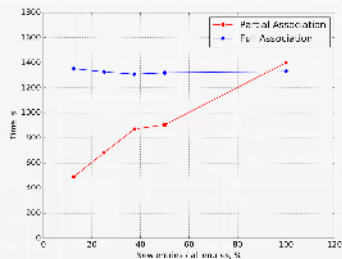
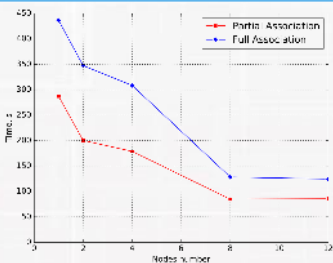
Spark SQL

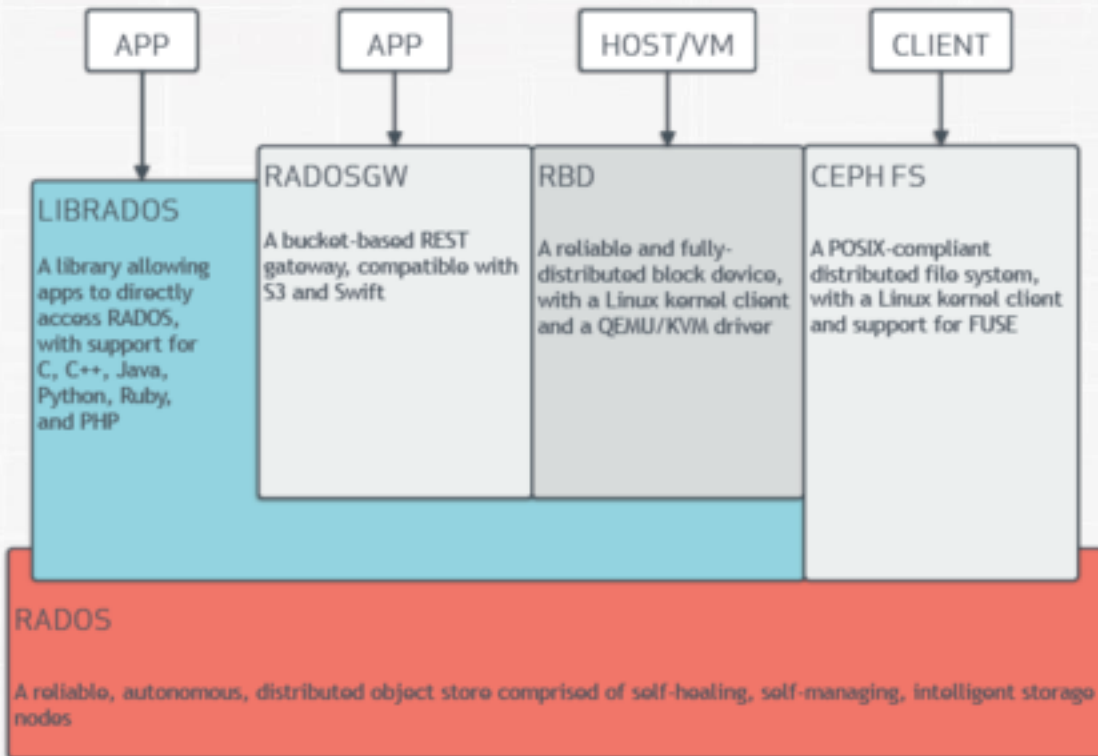
Spark Streaming

MLlib (machine learning)

GraphX (graph)

Apache Spark





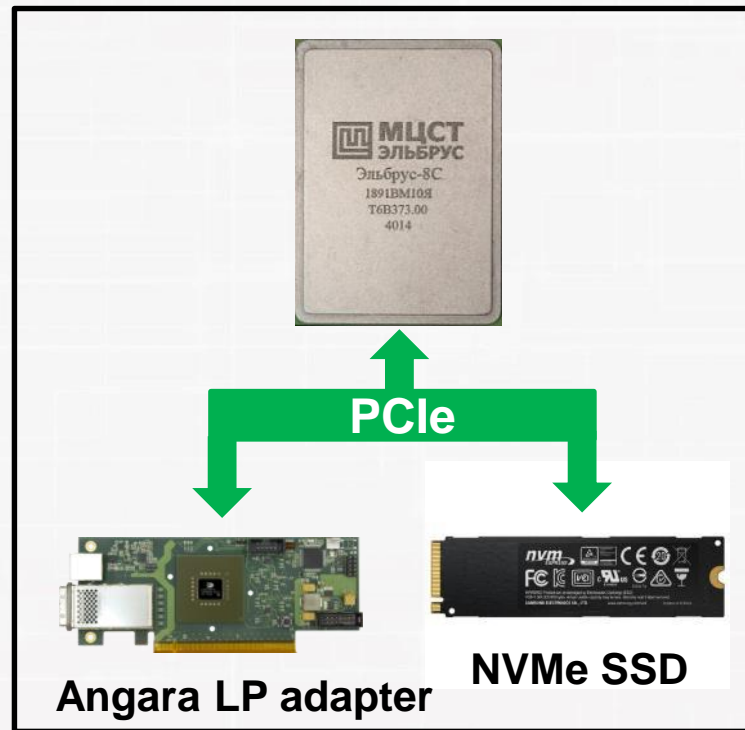
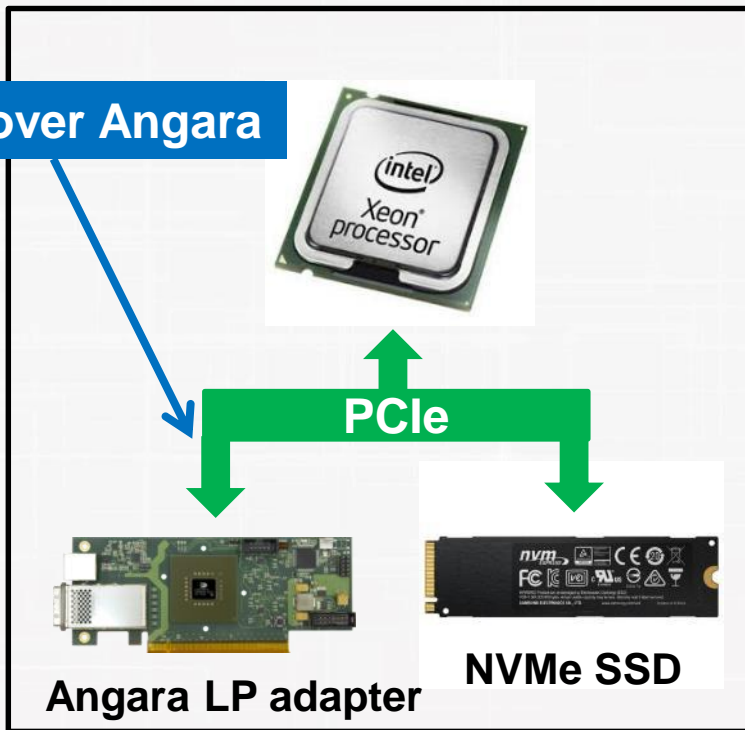
```

$# rados bench -p scbench 10 rand
sec Cur ops started finished avg MB/s cur MB/s last lat(s) avg lat(s)
0 0 0 0 0 0 0 - 0
1 16 258 242 967.695 968 0.142073 0.0610817
2 16 487 471 941.789 916 0.0234243 0.0647762
3 15 739 724 965.153 1012 0.145909 0.0643161
4 15 1049 1034 1033.83 1240 0.0233676 0.0603486
5 16 1361 1345 1075.84 1244 0.0055336 0.0579456
6 16 1714 1698 1131.84 1412 0.0299221 0.0556169
7 16 2065 2049 1170.7 1404 0.012719 0.0536391
8 16 2419 2403 1201.34 1416 0.0165833 0.0523875
9 16 2754 2738 1216.73 1340 0.0138274 0.0517339
10 15 3103 3088 1235.04 1400 0.0764744 0.0510114
  
```

```

Total time run: 10.090779
Total reads made: 3104
Read size: 4194304
Object size: 4194304
Bandwidth (MB/sec): 1230.43
Average IOPS: 307
Stddev IOPS: 49
Max IOPS: 354
Min IOPS: 229
Average Latency(s): 0.0512704
Max latency(s): 0.22856
Min latency(s): 0.00462222
  
```


NVMe over Angara



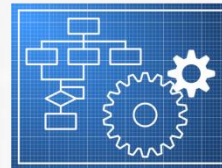
Взаимодействие с научным сообществом

- Исследование производительности программных систем и библиотек на системах с сетью Ангара
- Отображение процессов на топологию с учетом маршрутизации сети Ангара
- Оптимизация коллективных операций для MPI
- Разработка (или портирование) эффективной коммуникационной библиотеки, например, SHMEM, GASNet
- Разработка системы поддержки контрольных точек задачи

- **4 опубликованных статьи** от **3х научных коллективов**, в названии которых присутствует сеть Ангара, одна – в трудах конференции Parallel Processing and Applied Mathematics, Польша
- **10 научных публикаций** за 2016-2017 год
- **6 докладов** на конференции Суперкомпьютерные дни в России 2017

- Практикум 2017 года для студентов
- Научная группа ВМК МГУ
- С.В. Поляков, ИПМ РАН
- ТЕСИС, Flowvision
- А.В. Созыкин, ИММ УрО РАН

- Настройка программного обеспечения на вычислительных системах, в том числе MPI
- Оперативная поддержка пользователей
angara.nicevt.ru
support@angara.nicevt.ru
- Профилирование и адаптация прикладного ПО



Контакты:

117587, Москва, Варшавское ш, 125

angara@nicevt.ru

