Ростех

Возможности высокоскоростной сети Ангара при создании высокопроизводительных вычислительных систем, систем хранения и обработки Больших Данных

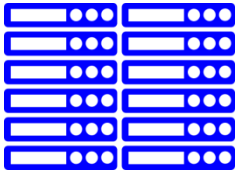
А.С. Симонов

План доклада

- Основные сведения о сети Ангара
- Инженерные пакеты и приложения
- Системы хранения и обработки Больших Данных
- Взаимодействие с научным сообществом

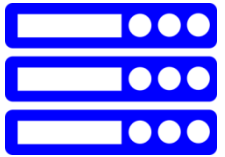
Назначение:

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)



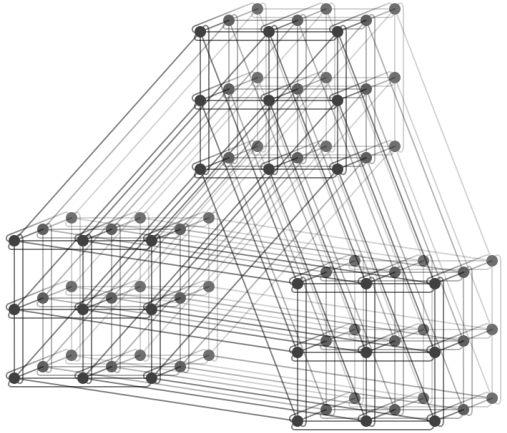
Коммуникационная сеть Ангара
Основные характеристики

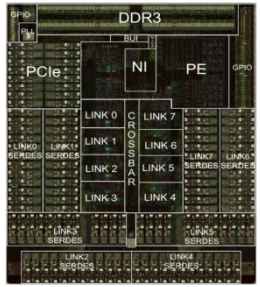
НИЦЭВТ

Ростех

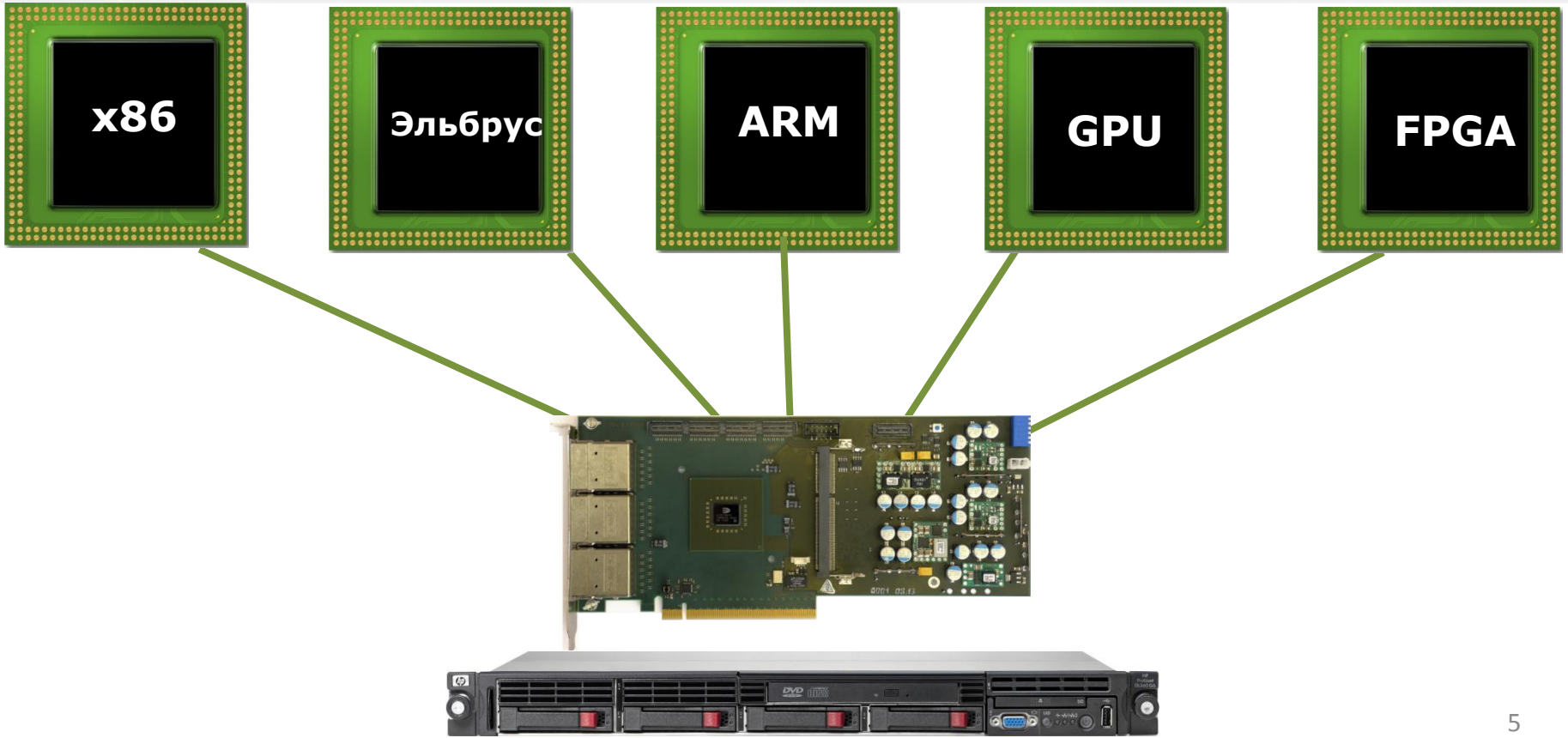
Ключевые особенности:

- Топология сети: 1D..4D-тор
- Адаптер на базе СБИС
- До 8 каналов связи с соседними узлами
- Прямой доступ в память удаленного узла (RDMA)
- Поддержка многоядерности
- Адаптивная передача пакетов
- Задержка на MPI ping-pong: 0,85 мкс
- Задержка на хоп: 130 нс
- Масштабирование: до 32K узлов
- Энергопотребление до 20 Вт
- Различные физические среды передачи данных

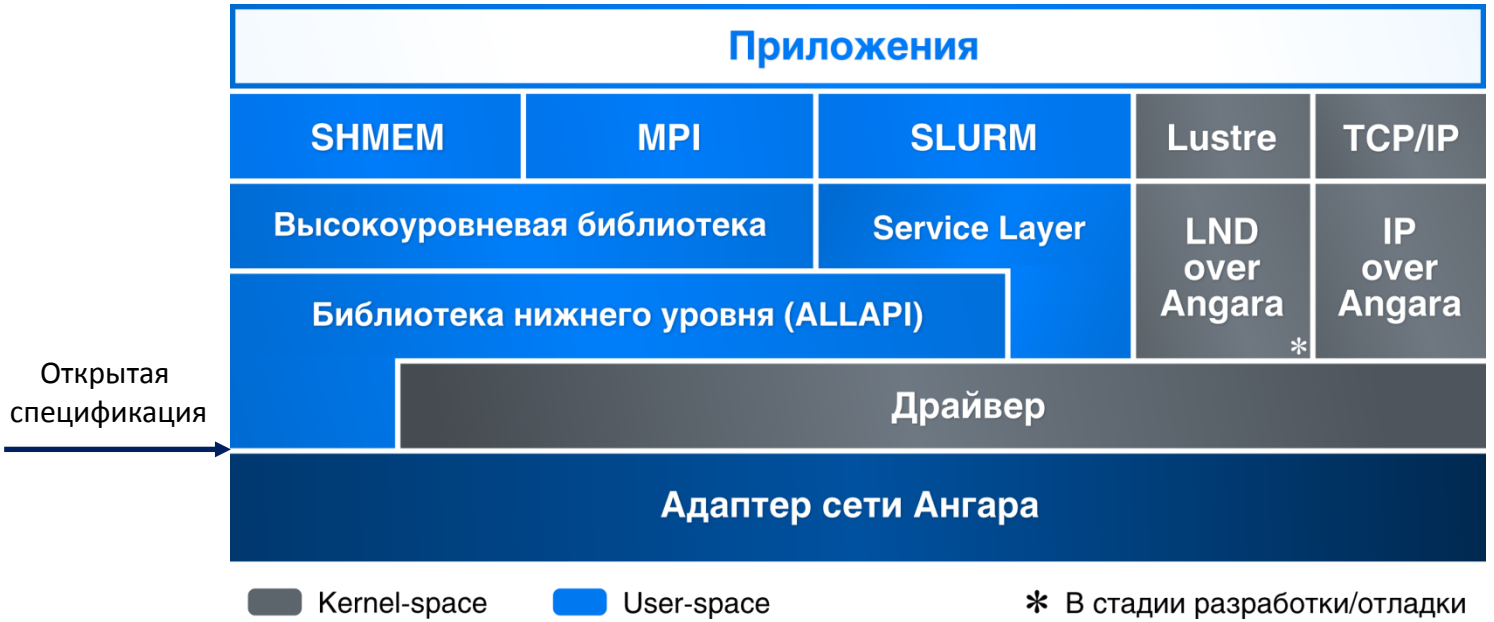




Коммуникационная сеть Ангара
Поддержка вычислительных элементов



Стек программного обеспечения

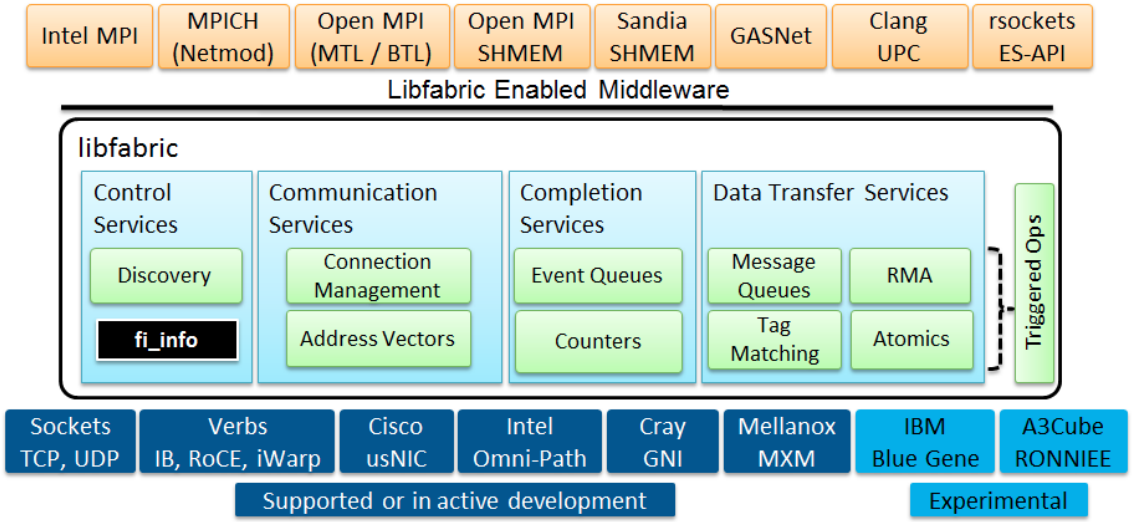


- Поддержка ОС : Astra Linux SE 1.3, ОС «Эльбрус», OpenSUSE/SLES 11 SP3/4, CentOS 6.0-7.3, Версия ядра Linux от 2.6.21 до 3.16.0
- Поддержка компиляторов языков Fortran 77/90/95 (GNU, Intel), C/C++ (GNU, Intel)

Варианты библиотеки MPI

Варианты реализации:

- MPICH 3.0.4
  - Оптимизированная версия
- MPICH 3.2
- OpenMPI 1.10.2
- Разработки по libfabric -> Intel MPI



Настройка стека ПО и MPI для каждого конкретного кластера заказчика

Варианты сетевого оборудования Ангара

1. Высокопроизводительное решение на базе FHFL адаптера и Samtec кабеля (**доступно**)

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

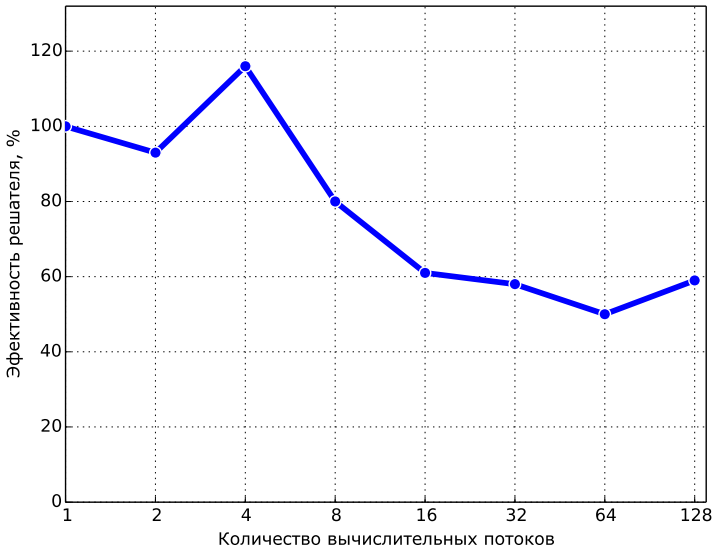
- **ОИВТ РАН: 32 вычислительных узла**
  - 1 процессор Intel Xeon E5-1650 v3 (6 ядер, 3.0 ГГц)
  - Nvidia GeForce GTX 1070
  - DDR4 16 ГБ
  - 4D-тор 4x2x2x2
- **Ангара-К1: 36 вычислительных узлов**
  - 12 узлов с 1 процессором Intel Xeon E5-2660 (8 ядер, 2.2 ГГц)
  - 24 узла с 2 процессорами Xeon E5-2630 (6 ядер, 2.3 ГГц)

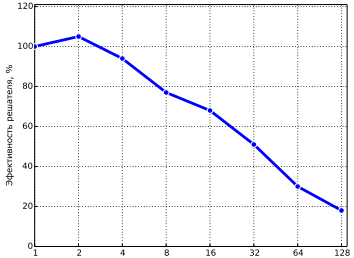
25.09.2017 В. Стегайлов (ОИВТ РАН)

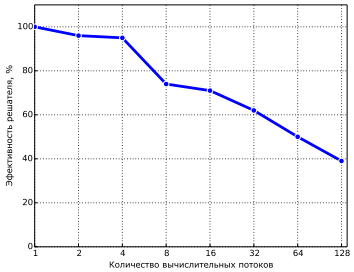
Гибридный суперкомпьютер на базе сети Ангара для задач вычислительного материаловедения

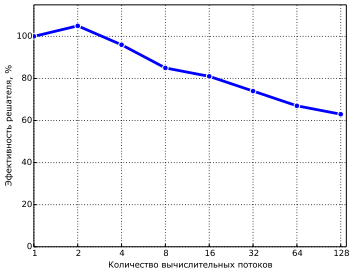
Инженерные пакеты

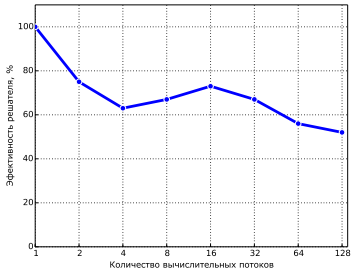
(1) ANSYS Mechanical 18.2
Совместно КАДФЕМ Си-Ай-Эс

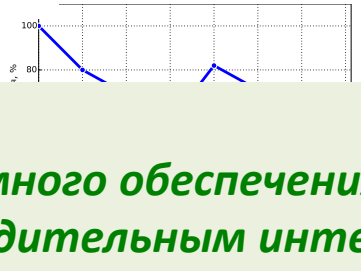


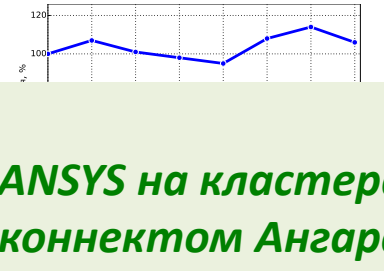


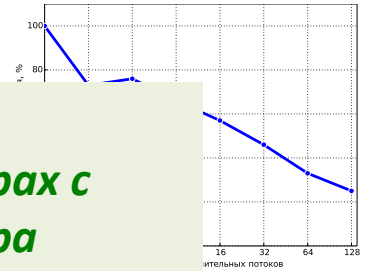












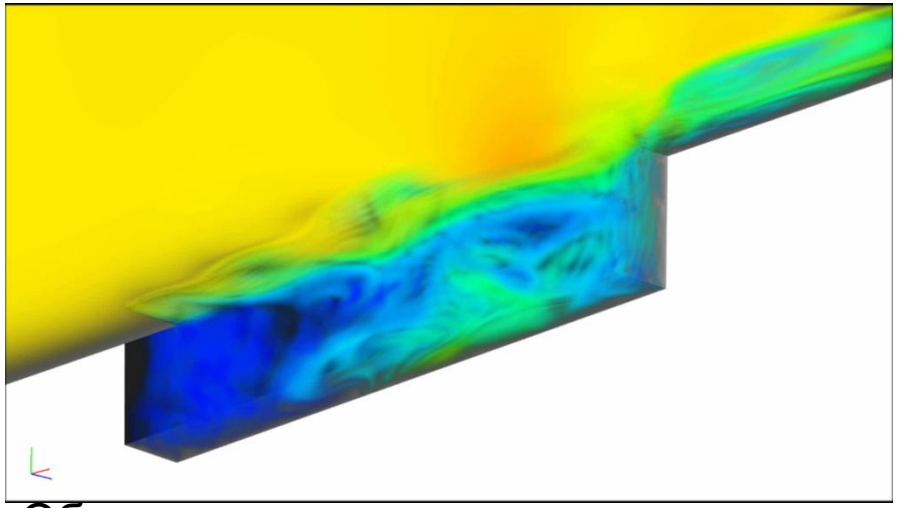
25.09.2017 Ю. Новожилов.

Тестирование работы программного обеспечения ANSYS на кластерах с отечественным высокопроизводительным интерконнектом Ангара

(2) Flowvision
Совместно с ТЕСИС

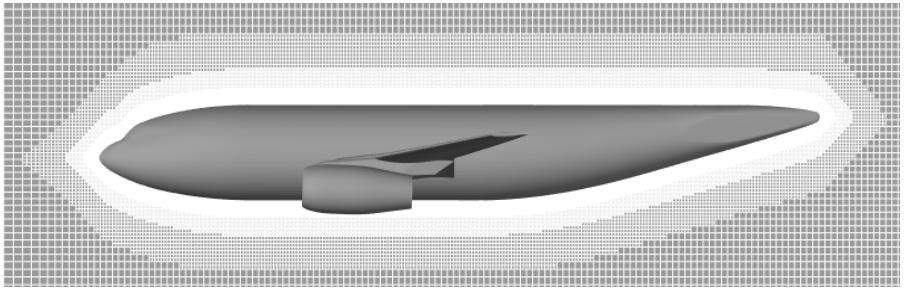
M219 Cavity case

Обтекание каверны воздухом, 5.5 млн ячеек



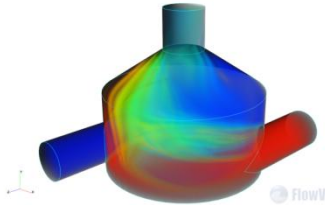
Объемная визуализация скорости

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

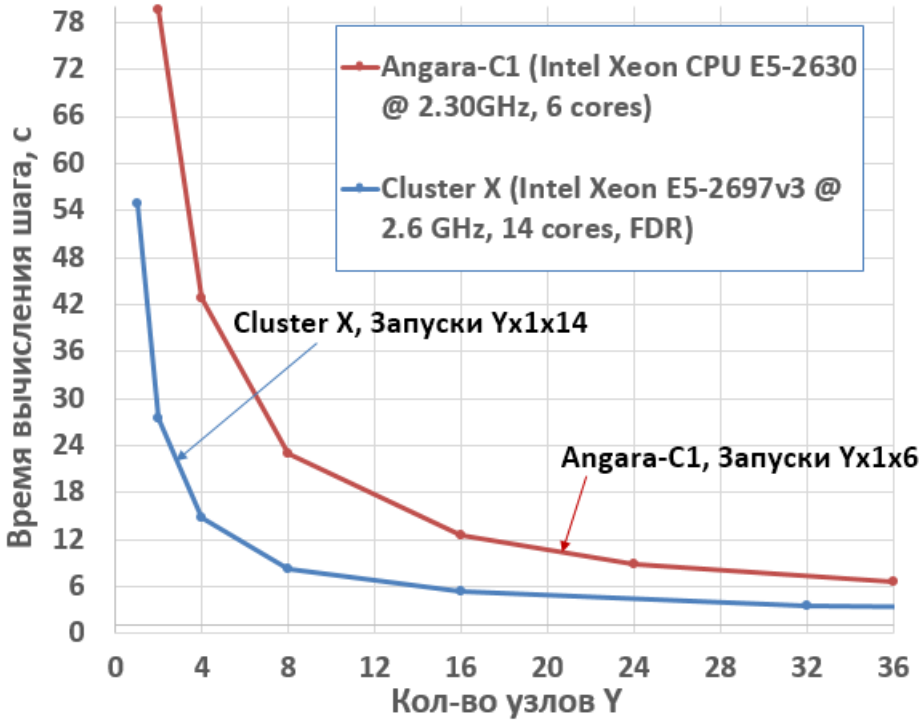


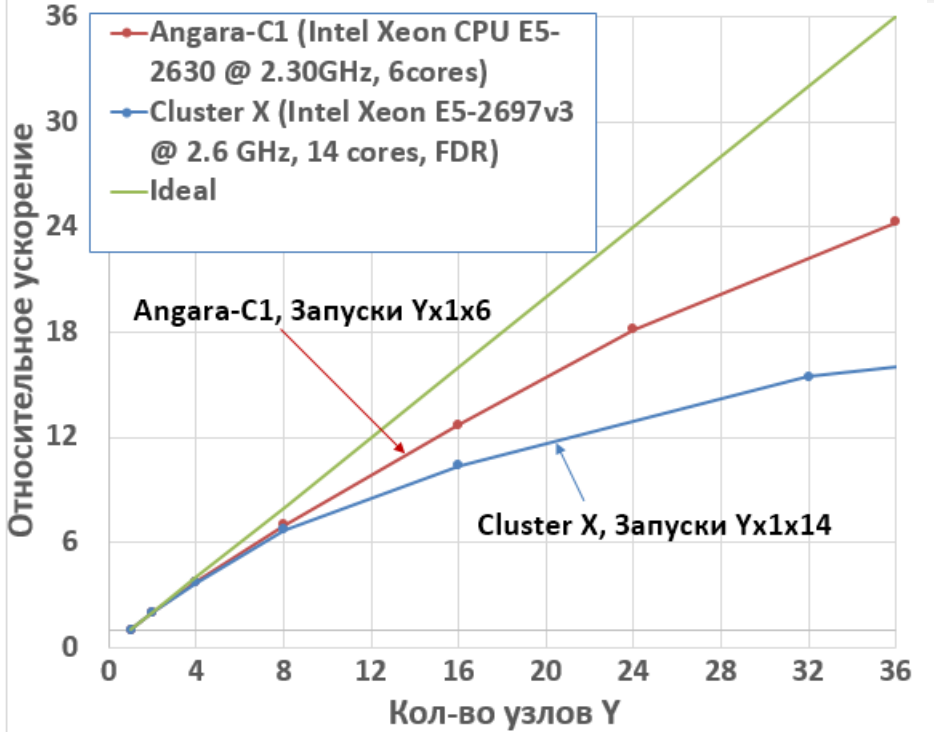
Задача Смеситель, 260 тыс. ячеек

Распределение температуры



(2) Flowvision. M219 Cavity case
Совместно с ТЕСИС





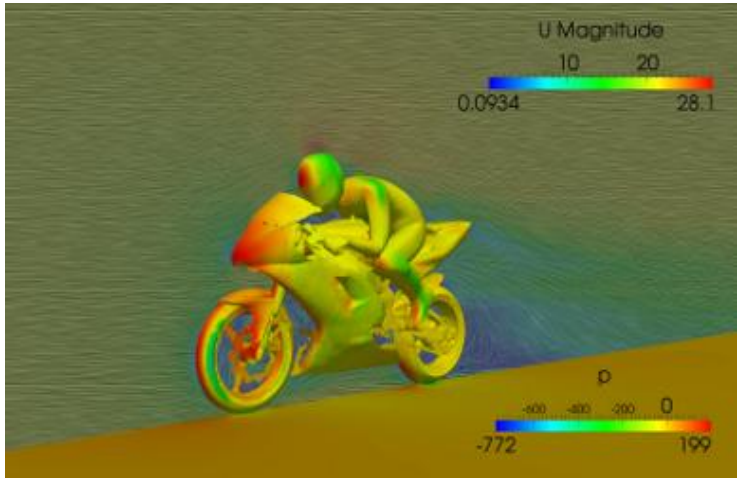
25.09.2017 В. Акимов

Исследование масштабируемости FlowVision на кластере с сетью Ангара

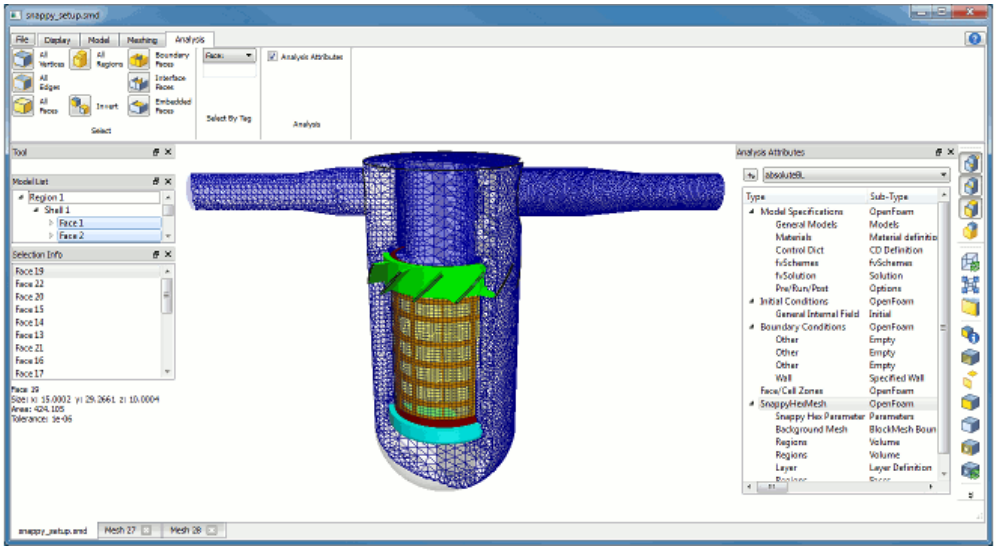
(3) Open FOAM

OpenFOAM

The Open Source CFD Toolbox

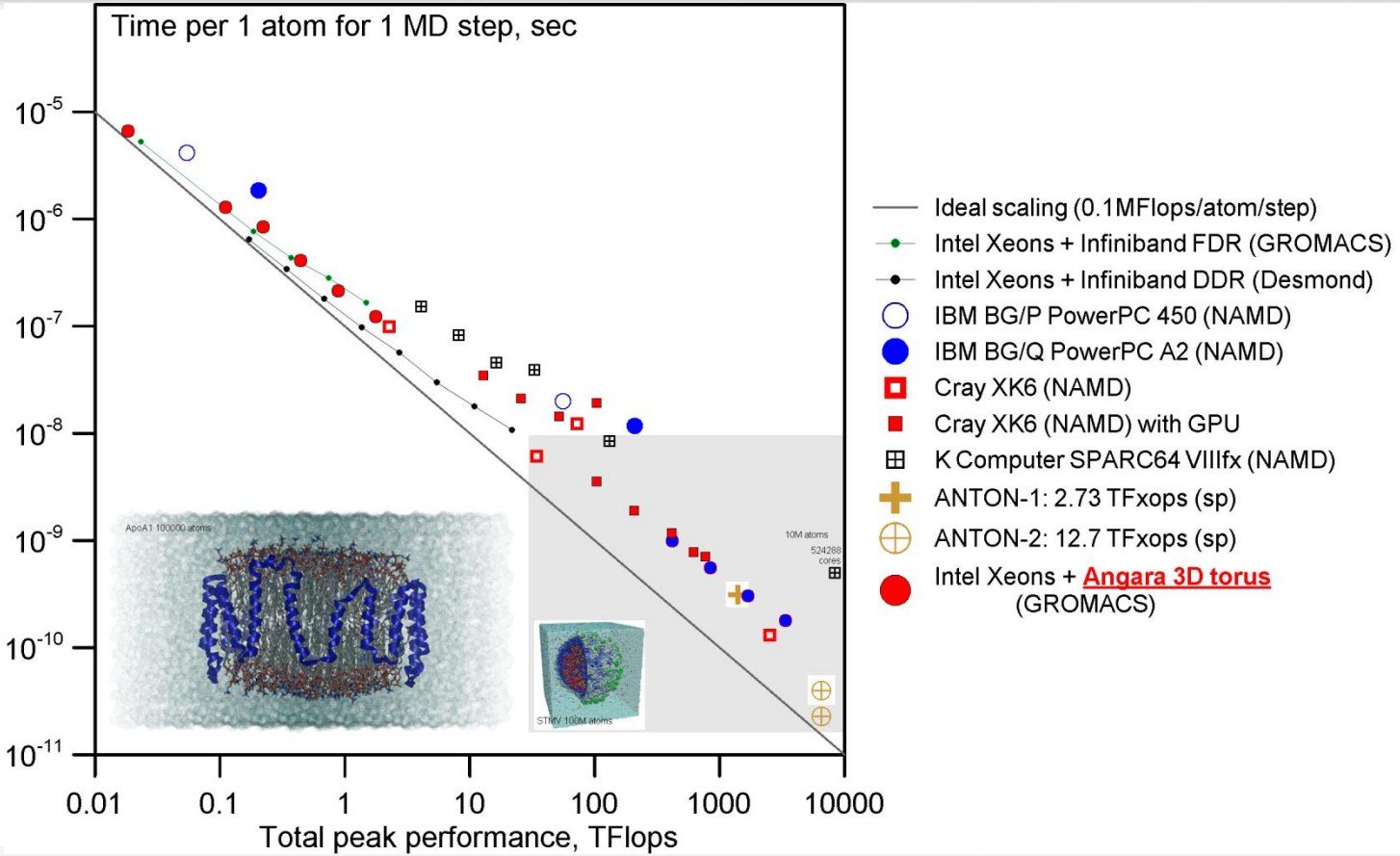


Версия 3.0.0

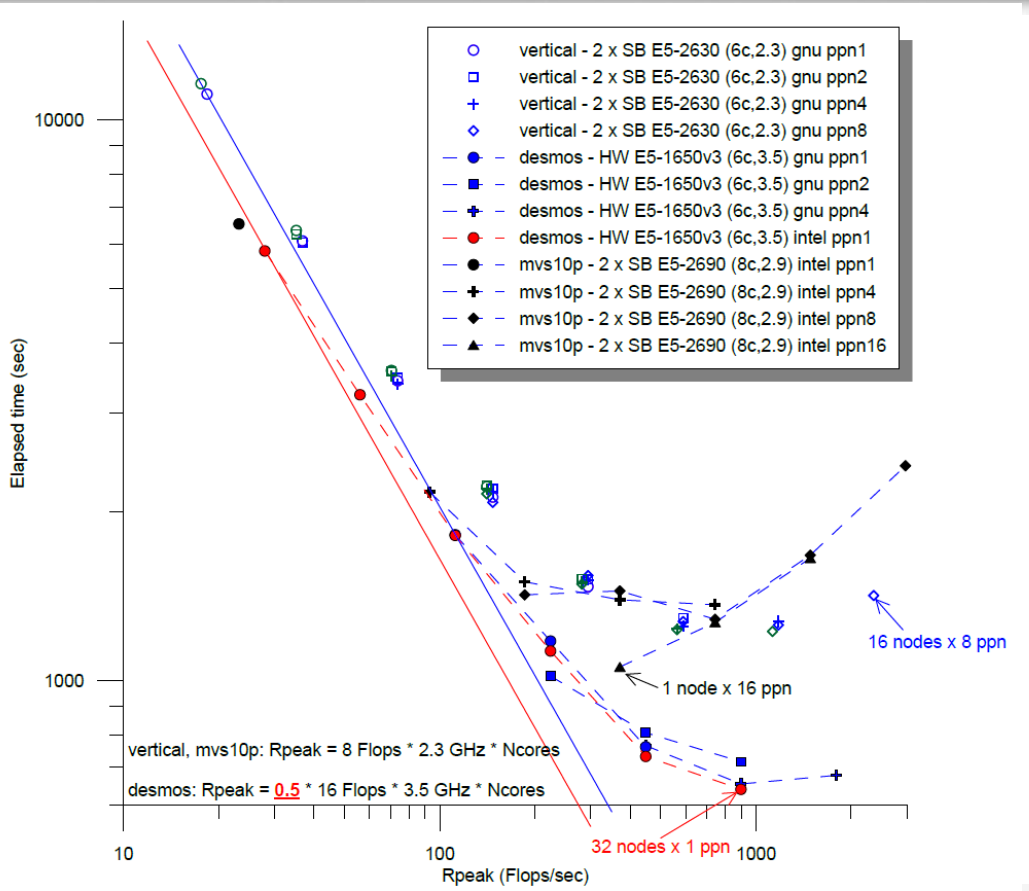


Приложения

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

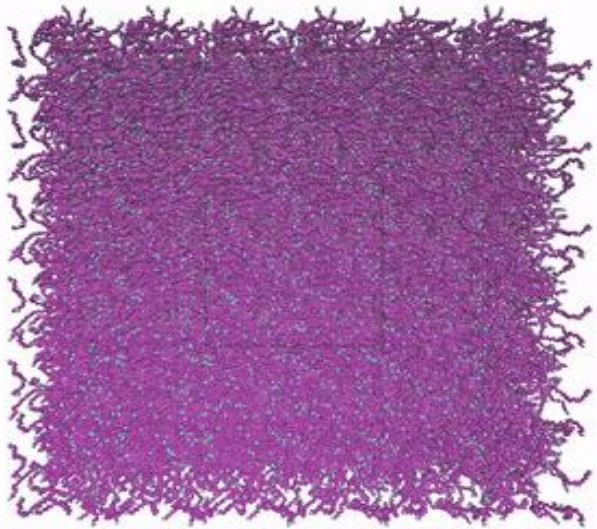


(2) Молекулярная динамика. Кристалл золота (4 атома в расчетной ячейке, k-сетка 21x21x21), VASP. д.ф.-м.н. В.В. Стегайлов, ОИВТ РАН

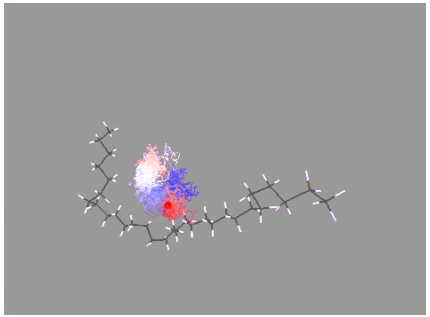


(3) Молекулярная динамика. Исследование свойств жидких углеводородов, LAMMPS.

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

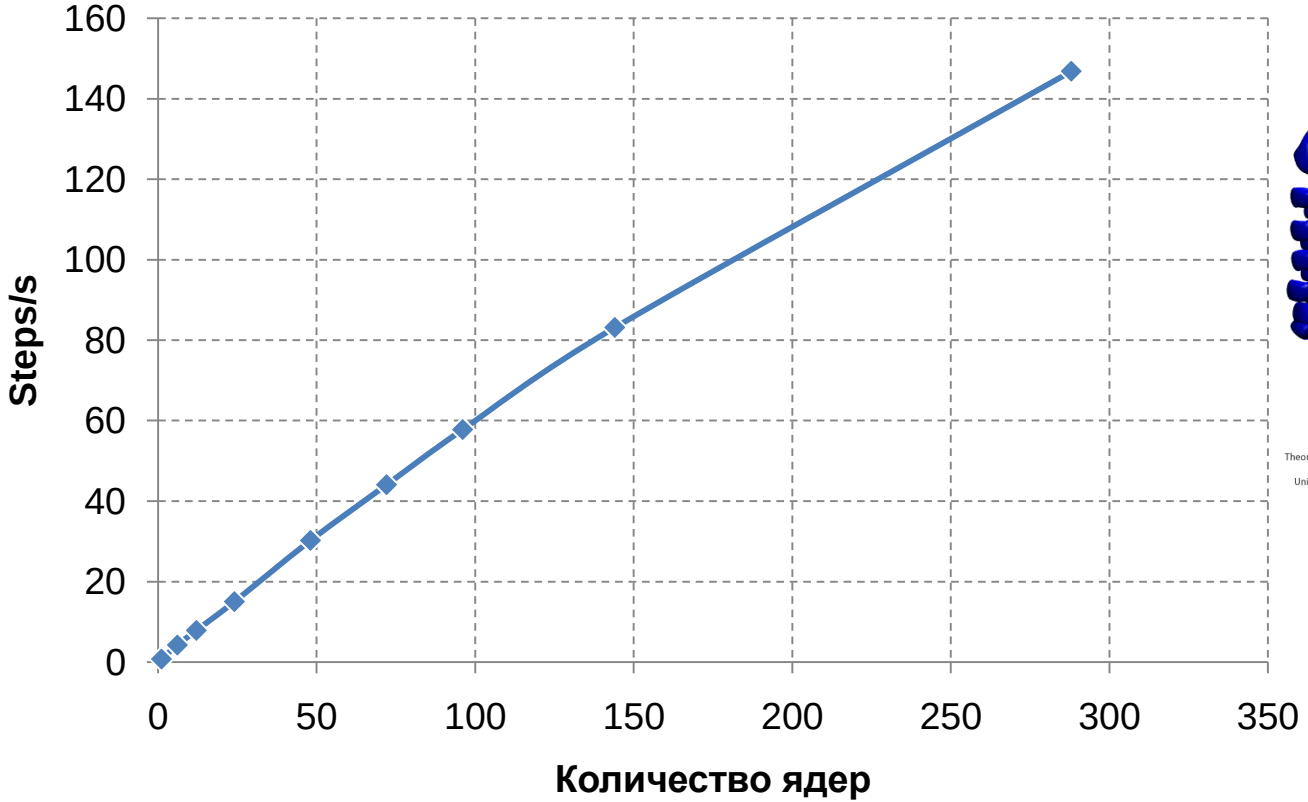
Диффузия, вязкость жидких углеводородов, т.к. они входят в состав

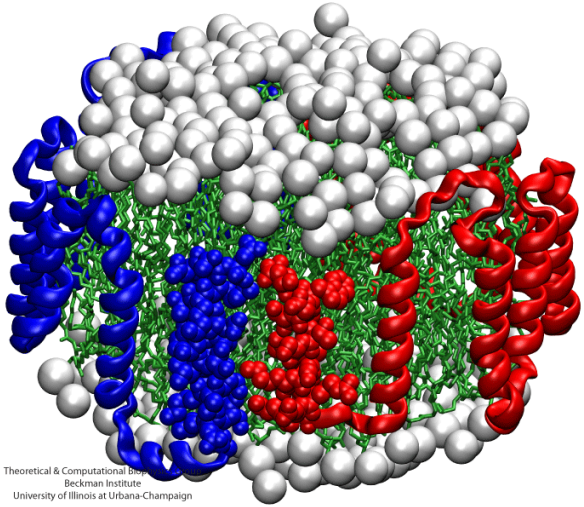
трансформаторных масел,

топлив и смазочных материалов

Молекулярная динамика-> макроскопические свойства

(4) Молекулярная динамика. ApoA1 benchmark, NAMD. РСК

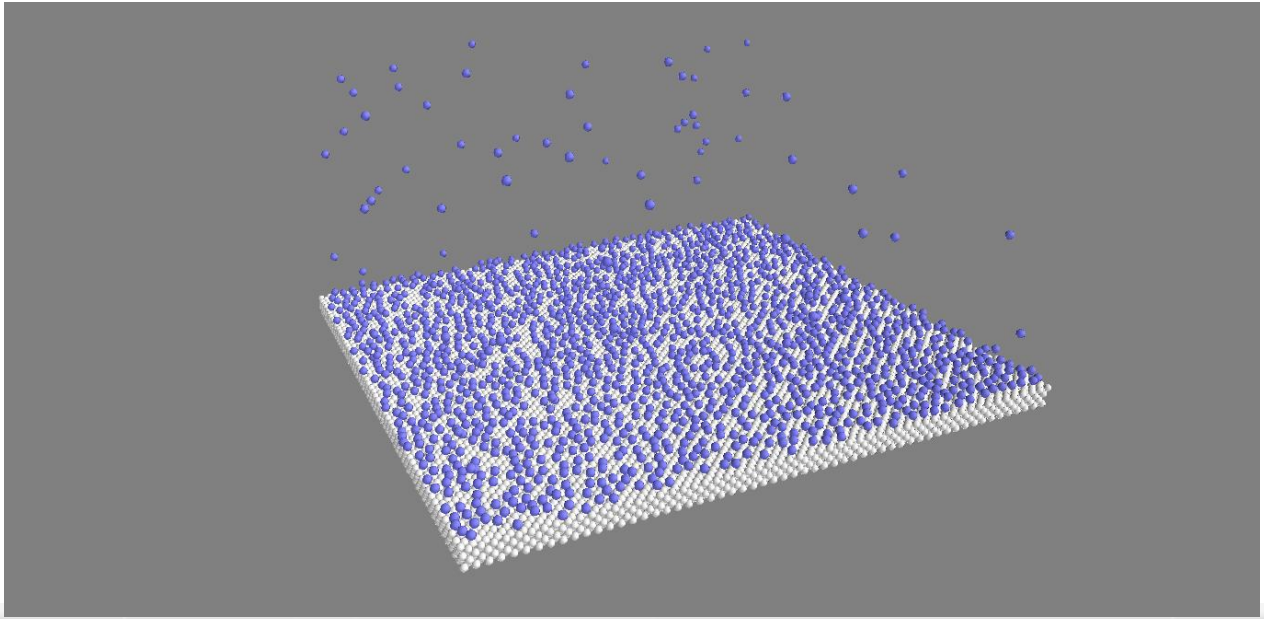


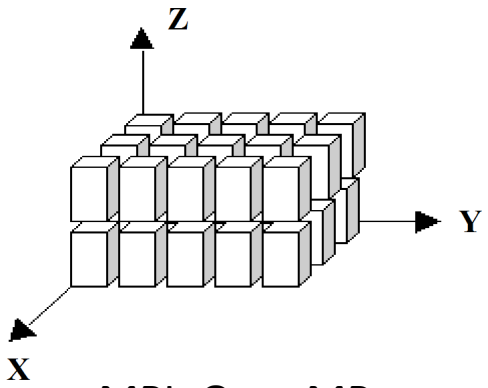


(5) Моделирование термодинамического равновесия в системах газ-металл методами молекулярной динамики. д.ф.-м.н. С.В. Поляков, ИПМ РАН

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



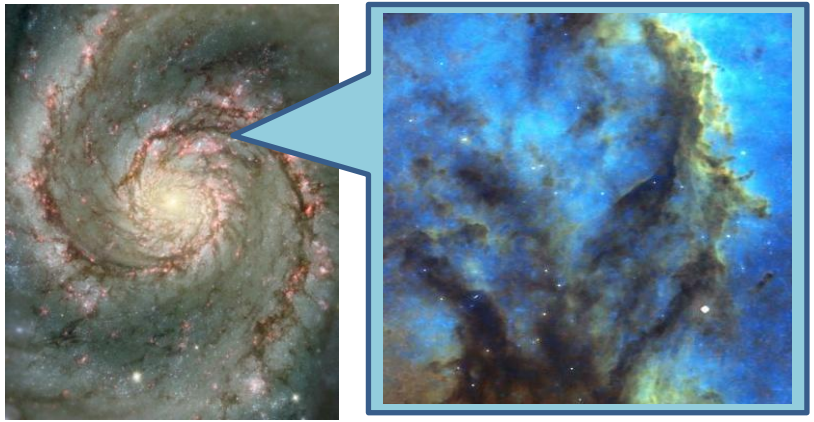


MPI+OpenMP

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

(6) Моделирование МГД турбулентности астрофизических тел

д.ф.-м.н. И.М. Куликов, ИВМиМГ СО РАН



The self-gravity magneto hydrodynamics equations

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

$$\frac{\partial\vec{B}}{\partial t}=\nabla\times\left(\vec{v}\times\vec{B}\right)\qquad\nabla\cdot\vec{B}=0\qquad\Delta\Phi=4\pi G\rho$$

$$\rho E=\rho\varepsilon+\frac{\rho v^2}{2}+\frac{B^2}{2}\qquad p=\left(\gamma-1\right)\rho\varepsilon\qquad p^*=p+\frac{B^2}{2}$$

НИЦЭВТ

Тестирование организациями

Ростех

NIAGARA
Российские Суперкомпьютеры

РСК

Т-ПЛАТФОРМЫ

ТеСИС

inline group

МЦСТ ЭЛЬБРУС

РФЯЦ ВНИИЭФ

ОРДЕНА ЛЕНИНА ИПМ им.М.В.КЕЛДЫША РАН

ОБЪЕДИНЕННЫЙ ИЗТАН

ИВМ РАН

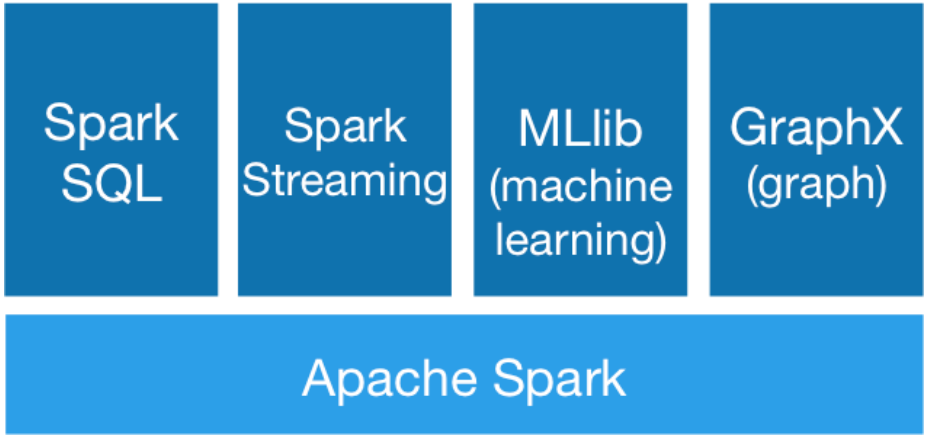
ИДСТУ СО РАН

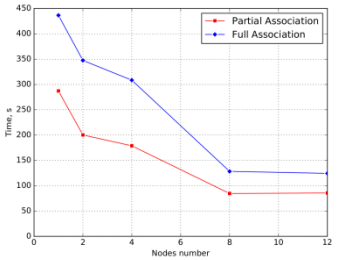
Статистика использования кластера «Ангара-К1»

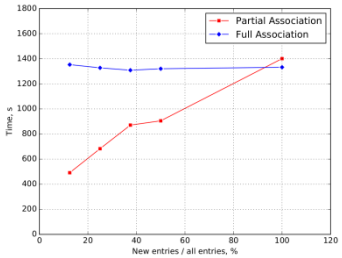
- В режиме внешнего доступа работает с 1.11.2015 г.
- Число пользователей: 40
- Всего использовано 269782 процессоро-часов

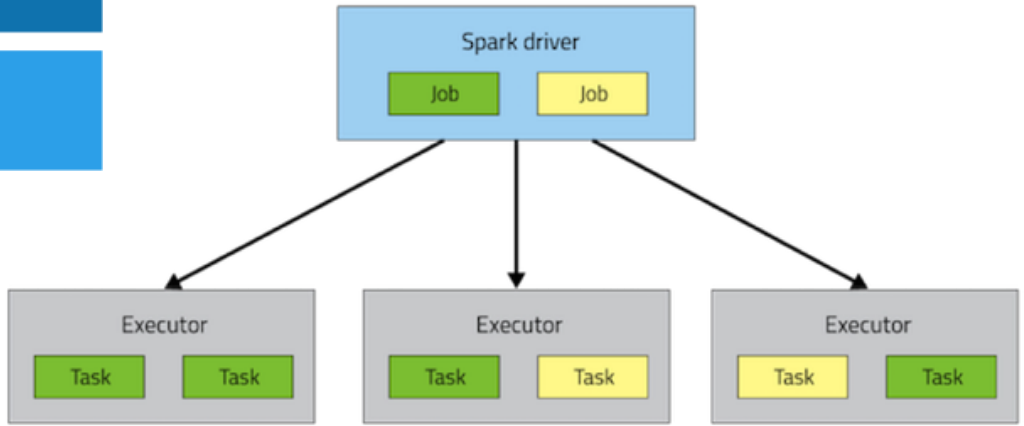
Системы хранения и обработки Больших Данных

Apache Spark





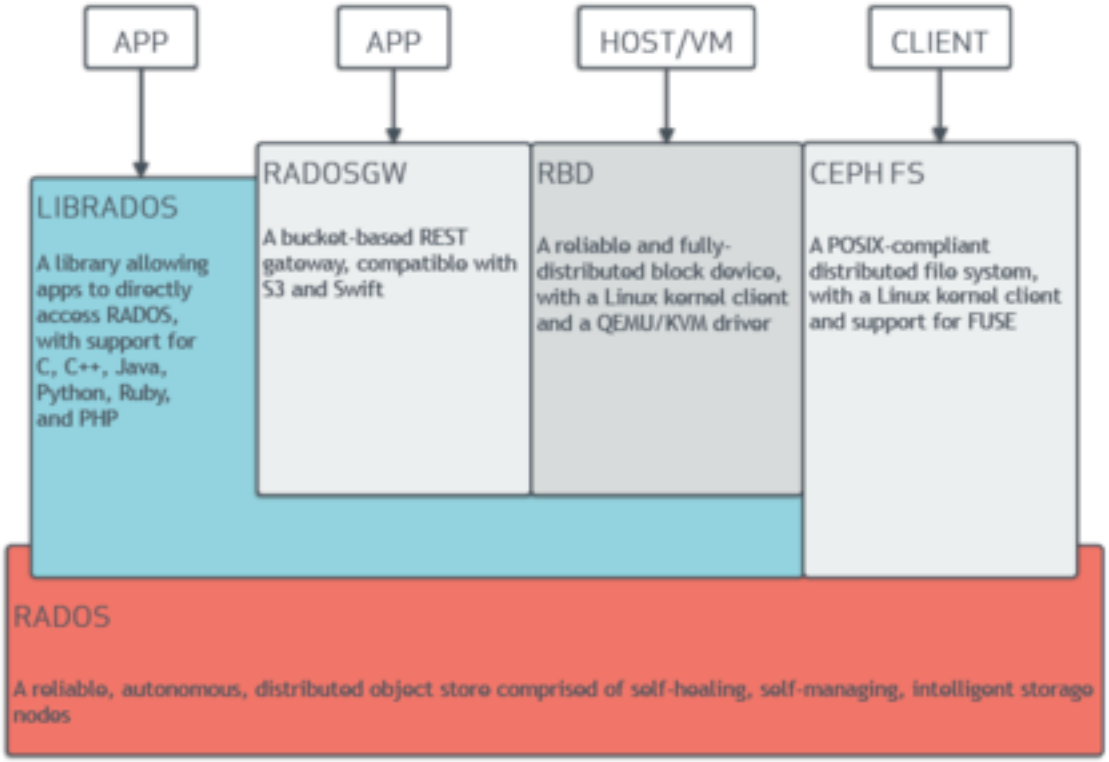




25.09.2017 А. Агарков

Оценочное тестирование Apache Spark на кластере с сетью Ангара

Программная система хранения данных Ceph



ceph

```
$# 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
    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
```

Взаимодействие с научным сообществом

Направления исследований для научных организаций

- 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
- другие …

26.09.2017 М.Р. Халилов, А.В. Тимофеев (НИУ ВШЭ)

Оптимизация работы MPI-программ с учётом особенностей топологии кластеров, использующих коммуникационную сеть Ангара

Партнеры по созданию вычислительных систем

NIAGARA
Российские Суперкомпьютеры

РСК

ПЛАТФОРМЫ

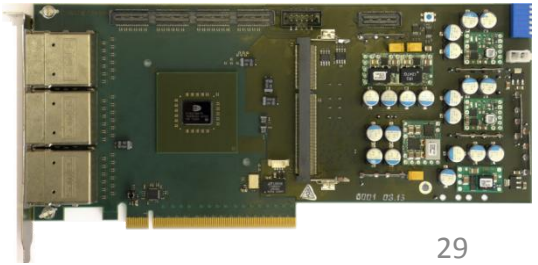
inline group

СТОРУС

ОТКРЫТЫЕ ТЕХНОЛОГИИ

Ростех ИНФОРМ

МЦСТ ЭЛЬБРУС



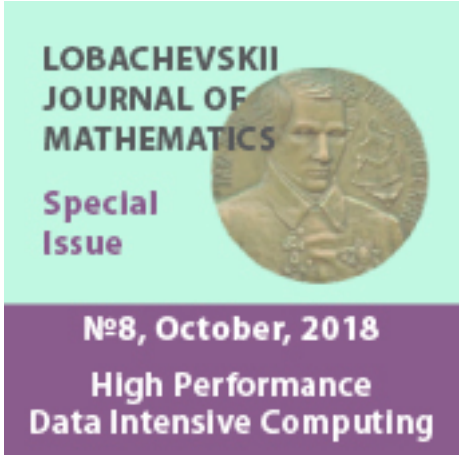
GRAPH HPC 2018

Научно-практическая конференция

Технологии параллельной обработки больших графов

1 марта 2018, Москва

- Конференция GraphHPC 2014-2018 годы
- Тенденции сближения технологий HPC и BigData
- Сайт конференции graphhpc.dislab.org



Контакты:

117587, Москва, Варшавское ш, 125

angara@nicevt.ru