

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ДАЛЬНЕВОСТОЧНОЕ ОТДЕЛЕНИЕ



ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

ЦКП “ЦЕНТР ДАННЫХ ДВО РАН”: ТЕКУЩЕЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ РАЗВИТИЯ

Сорокин А.А., Мальковский С.И.

Москва, 2017



ГИБРИДНЫЙ ВЫЧИСЛИТЕЛЬНЫЙ КЛАСТЕР*



Кластер состоит из 5 вычислительных узлов со следующими характеристиками (каждый узел):

- 2 десятиядерных процессора IBM POWER8 2.86 ГГц;
- память ECC, 256 ГБ;
- 2 x 1 ТБ 2.5" 7K RPM SATA HDD;
- 2 x NVIDIA Tesla P100 GPU, NVLink;
- контроллер EDR InfiniBand (100 Гбит/с).

Introducing IBM Power System S822LC for HPC First Custom-Built GPU Accelerator Server with NVLink



- Custom-built GPU Accelerator Server
- High-Speed NVLink Connections between CPUs & GPUs and among GPUs
- Features novel NVIDIA P100 Pascal GPU accelerator



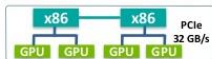
NVIDIA P100 Pascal GPU

2.5x Faster CPU-GPU Data
Communication via NVLink



POWER8 NVLink Server

No NVLink between CPU & GPU
for x86 Servers: PCIe Bottleneck



x86 Servers with PCIe

Сеть передачи данных: EDR InfiniBand (100 Гбит/с).

Управляющая сеть: Gigabit Ethernet.

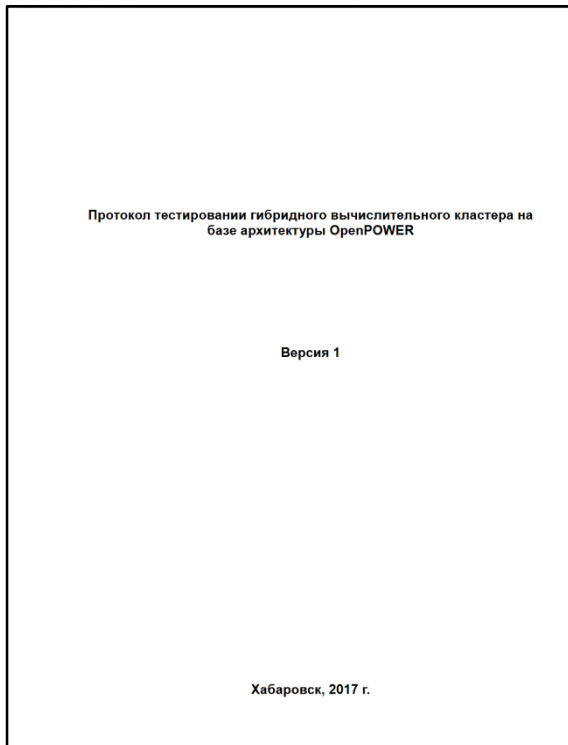
Хранилище: 50+ ТБ

*20 млн. рублей (6.6% от всех средств на модернизацию СКЦ ФАНО России в 2016 г.)



ПРОИЗВОДИТЕЛЬНОСТЬ ВЫЧИСЛИТЕЛЬНОГО КЛАСТЕРА

Подготовлен подробный протокол тестирования кластера



Оглавление	
Лист согласований и изменений	2
Местоположение документа	2
Перечень изменений	2
Описание кластера	4
Задачи тестирования и инструменты	5
Инструменты тестирования	6
Общие замечания	7
Тест на доступность нод кластера с mgmt узла	8
Тест работоспособности Infiniband соединения	9
Определение характеристик работы Infiniband сети	10
Определение характеристик работы кластера при помощи HPL	12
ВЫВОДЫ	16
Приложение 1. Характеристики работы Infiniband сети	17

Определение характеристик работы кластера при помощи HPL.

Для получения экспериментального показателя производительности кластера использовался скомпилированный тест HPL от NVIDIA. Он запускался с использованием IBM Spectrum MPI.

1. Редактируется файл HPL.dat с параметрами запуска LINPACK. Ниже представлен конфигурационный файл для запуска теста

```
HPL.dat
HPLinpack benchmark input file
Innovative Computing Laboratory, University of Tennessee
HPL.out output file name (if any)
6 device out (6=stdout,7=stderr,file)
1 # of problems sizes (N)
380160 Ns
1 # of NBS
788 NBS
1 PMAP process mapping (0=Row-,1=Column-major)
1 # of process grids (P x Q)
5 Ps
2 Qs
16.0 threshold
1 # of panel fact
2 PFACTs (0=left, 1=Crout, 2=Right)
1 # of recursive stopping criterion
2 8 NBMINs (>= 1)
1 # of panels in recursion
2 NDIVs
1 # of recursive panel fact.
2 RFACTs (0=left, 1=Crout, 2=Right)
1 # of broadcast
2 0 2 BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
1 # of lookahead depth
0 DEPTHs (>=0)
1 SWAP (0=bin-exch,1=long,2=mix)
192 swapping threshold
1 L1 in (0=transposed,1=no-transposed) form
0 U in (0=transposed,1=no-transposed) form
1 Equilibration (0=no,1=yes)
8 memory alignment in double (> 0)
```

2. В файле запуска run_linpack определяется переменная окружения «CPU_CORES_PER_RANK», равная кол-ву физических ядер, т.е. 10.
3. В файле запуска run_linpack определяется переменная окружения

http://lits.ccfеbras.ru/assets/files/protocol_v1.pdf

Rpeak = 55,83 Тфлопс

Linpack = 40,39 Тфлопс (~ 72%)

46 место в рейтинге ТОП-50 СНГ (27 редакция)



РАЗВИТИЕ КОМПЕТЕНЦИЙ

OpenPOWER™

Search

- Home
- About Us
- Membership
- Technical
- News/Events
- Get Involved
- Summit



Membership

Membership

OpenPOWER Membership

The goal of the OpenPOWER Foundation is to create an open ecosystem, using POWER architecture, to share expertise, investment and server-class Intellectual property to serve the evolving needs of industry and end users. To accomplish this goal, OpenPOWER has established a tiered membership base to attract and promote industry involvement from all members of its diverse technical ecosystem.

By becoming a member of the OpenPOWER Foundation, entities support the POWER architecture for custom open servers and components for Linux-based data centers as well as the development and proliferation of Open Source Software. OpenPOWER ecosystem partners can optimize the interactions of server building blocks — microprocessors, networking, I/O and other components — to fine tune for their needs to offer greater flexibility and solutions options to industry.

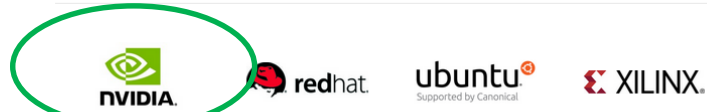
OpenPOWER Ecosystem welcomes Members at all levels



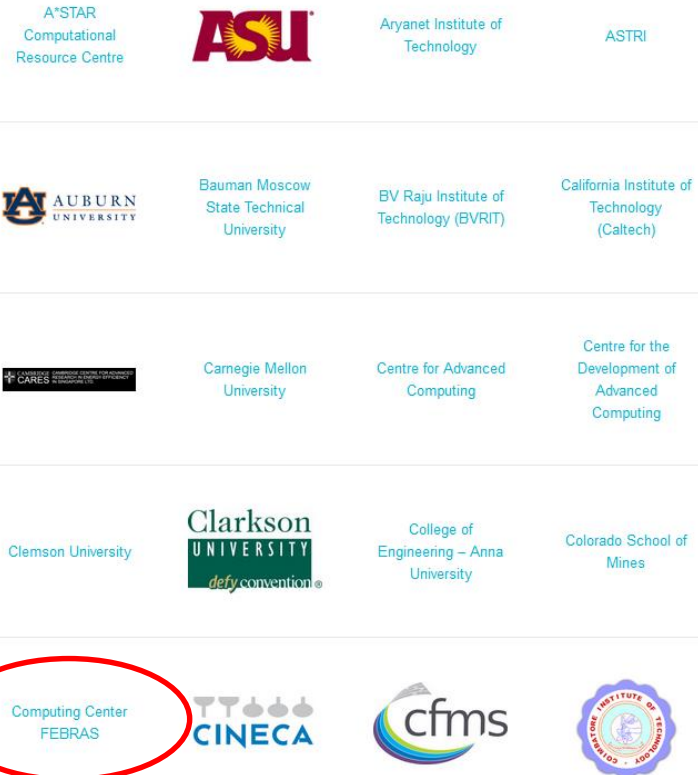
2 июня 2017 г., ПМЭФ,
подписание меморандума о
сотрудничестве с IBM

Current Members

Platinum Level



Associate & Academic Level



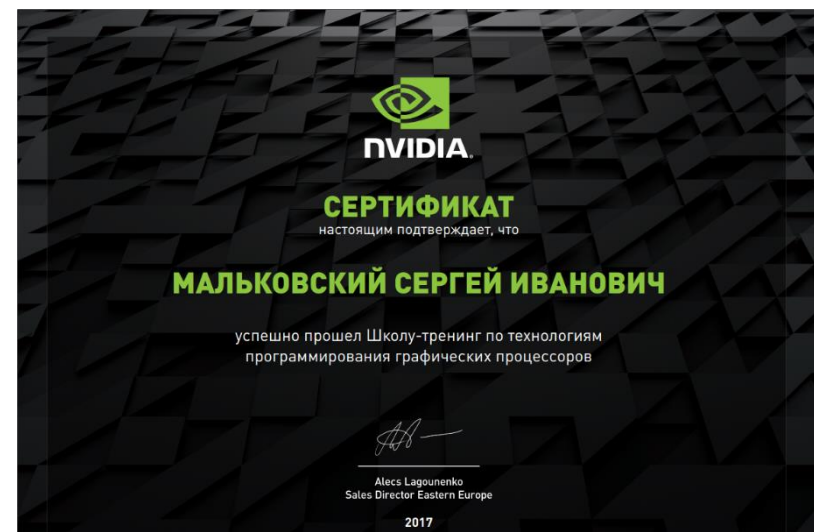
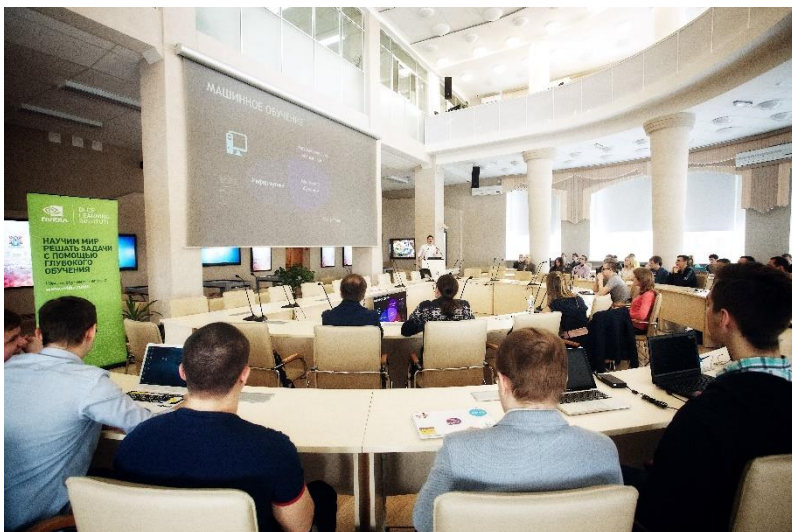


ОБУЧЕНИЕ И НАУЧНЫЕ СЕМИНАРЫ

Семинар по глубокому обучению и нейронным сетям, 20 марта 2017 г., г. Хабаровск



Школа по CUDA, 21-23 марта 2017 г., г. Хабаровск





ОБУЧЕНИЕ И НАУЧНЫЕ СЕМИНАРЫ

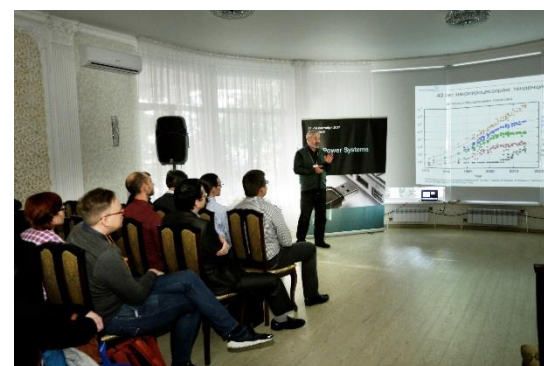
Семинар для пользователей ЦКП,
14 июня 2017 г., ВЦ ДВО РАН (г. Хабаровск)



Семинар в ИПУ РАН,
31 июля 2017 г. (г. Москва)



IV Всероссийская научно-практическая конференция
«Информационные технологии и
высокопроизводительные вычисления»
(сентябрь, 2017 г., г. Хабаровск)





ИССЛЕДОВАНИЯ ЭФФЕКТИВНОСТИ ГИБРИДНОЙ ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЫ

Для исследований были выделены 3 группы приложений:

1. Готовые пакеты прикладных программ или авторские алгоритмы, которые изначально были разработаны для вычислительных систем на архитектуре x86.
2. Компьютерные алгоритмы, использующие вычислительные ресурсы графических ускорителей (CUDA).
3. Библиотеки и сборки программного обеспечения для решения задач с использованием технологий машинного обучения (ML – machine learning), глубокого обучения (DL – deep learning) и систем искусственного интеллекта (AI – artificial Intelligence).



ИССЛЕДОВАНИЯ ЭФФЕКТИВНОСТИ ГИБРИДНОЙ ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЫ

Регламент работ:

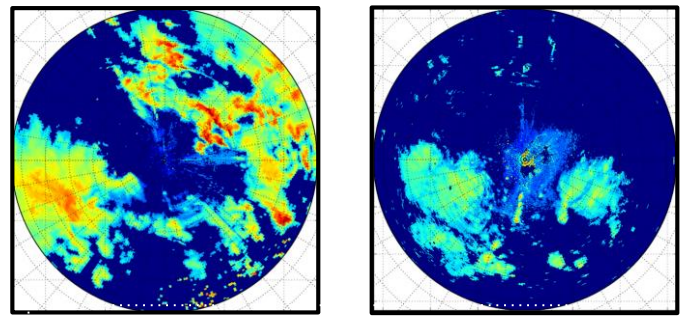
1. Тестирование проводилось на реальных научных задачах (проведен отбор по 1-2 задачи для каждой группы исследуемых приложений).
2. Заявки ученых прошли регистрацию и отбор в информационной системе “Центр коллективного пользования” и документально подтверждены заявлениями от организаций.
3. По каждой задаче сформирована группа, состоящая из сотрудников ЦКП и специалистов из предметной области.
4. При установке программ преимущественно используются компиляторы, библиотеки, сборки приложений, оптимизированные под гибридную архитектуру.



РЕШЕНИЕ ЗАДАЧ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ И СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Сверхкраткосрочный прогноз явлений погоды (в пределах 0 – 6 ч от срока наблюдения)

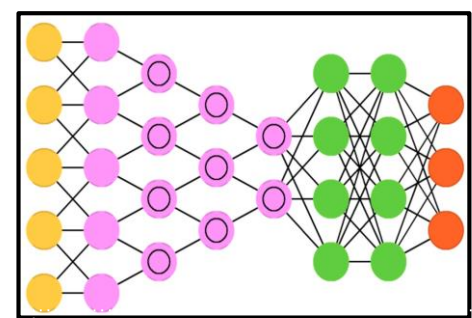
Данные: сеть наземных метеорадаров



Пространственное разрешение: 900*900 км;
Временное разрешение: 5 минут;
Продолжительность наблюдений: 20 лет;
Объем данных для обучения: > 2 Тб

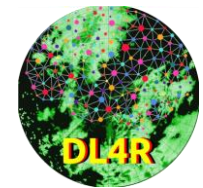


Модель: сверточная нейронная сеть

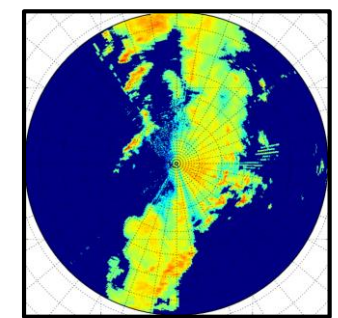


Количество слоев: > 20;
Регуляризация: dropout, pooling;
Количество параметров: > 1 млн.;
Время обучения: > 3 дней (100 Гб данных, неглубокая сеть, урезанное пространственное разрешение).

asimovinstitute.org/neural-network-zoo/



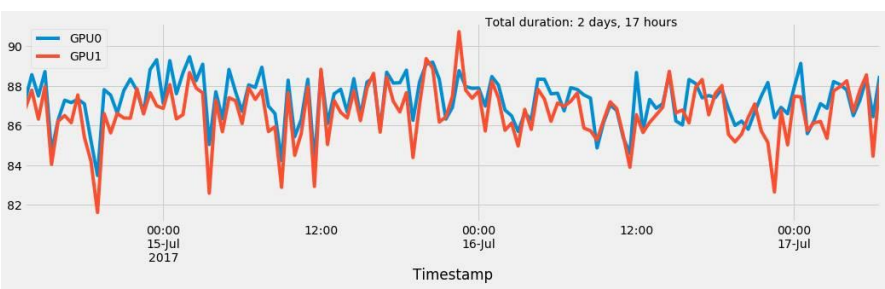
Прогноз: 1 час вперед



wradlib.org/wradlib-docs/



Ускорение: ~ 100 раз



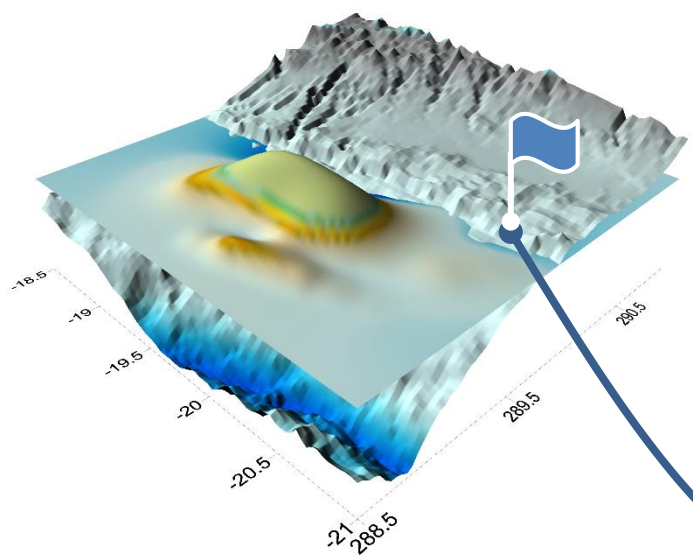


РЕШЕНИЕ ЗАДАЧ С ИСПОЛЬЗОВАНИЕМ CUDA



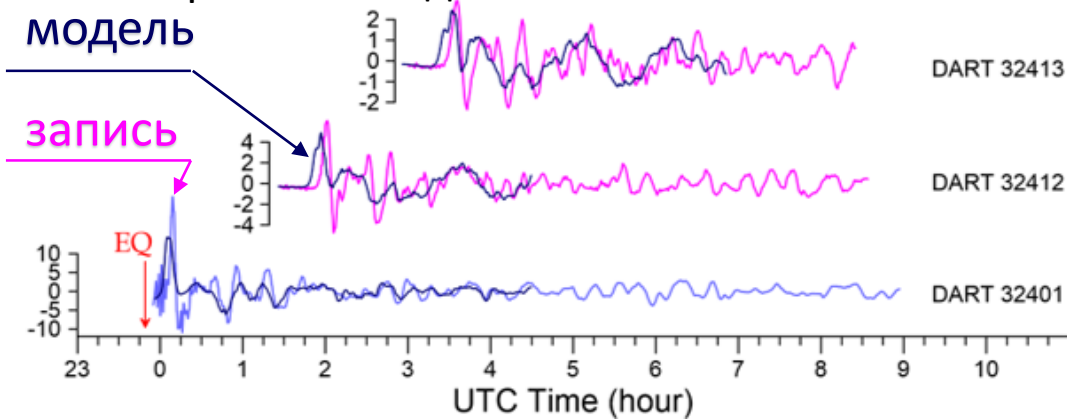
Институт морской геологии и геофизики
Дальневосточного отделения
Российской академии наук

Модель источника цунами
(начальное возвышение поверхности океана)
1 апреля 2014 г. в Чили

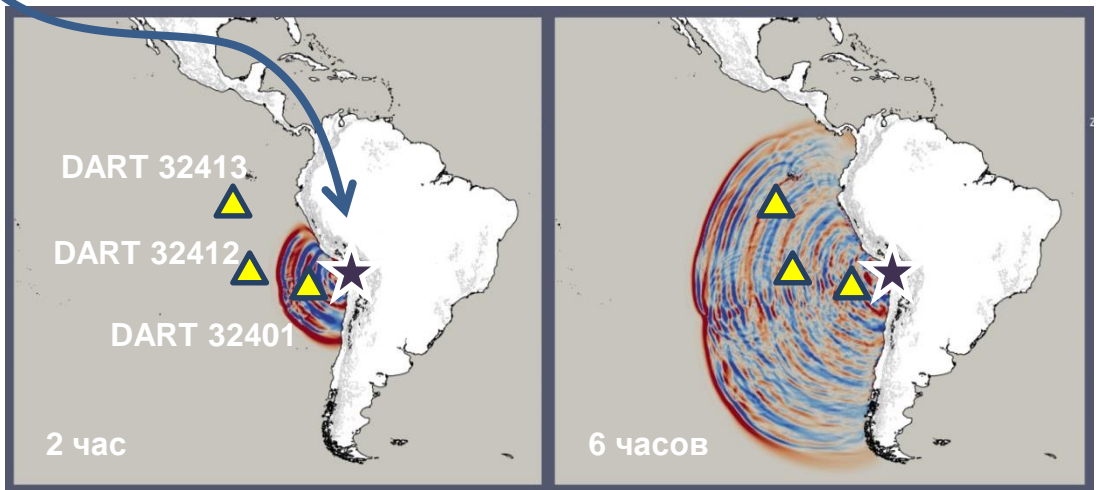


Результаты расчета:

Сравнение с данными NOAA DART:



Модель распространения волны цунами



Характеристики расчета:

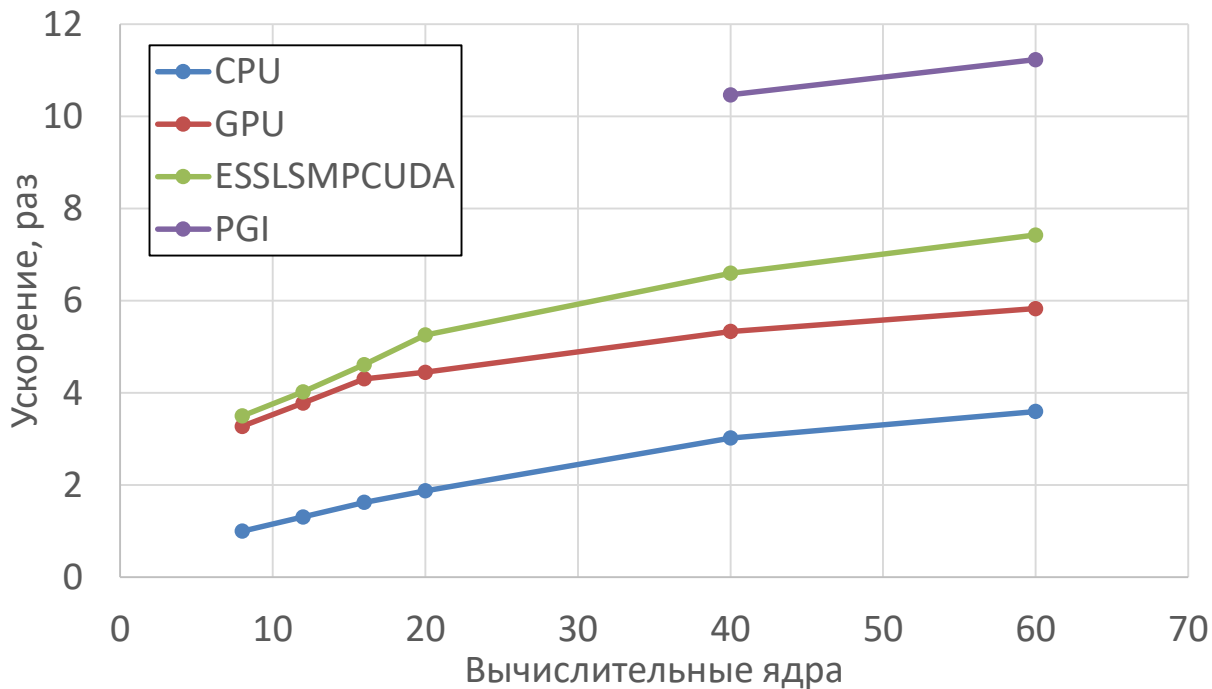
Размерность сетки	10801 x 6701 (весь океан)
Шаг сетки	2 угл. минуты
Расчетное время	48 часа
Компьютерное время	~30 мин



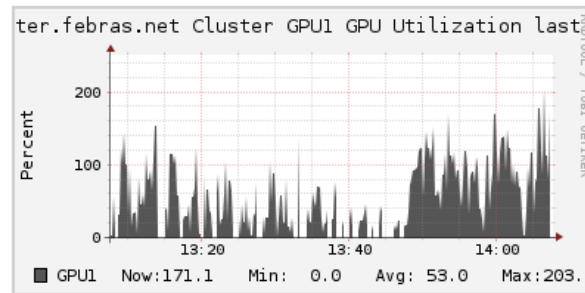
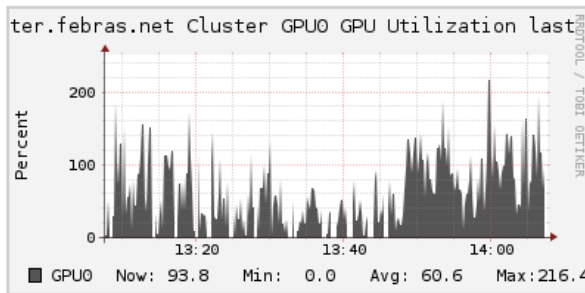
ПАКЕТЫ ПРИКЛАДНЫХ ПРОГРАММ, КОТОРЫЕ ИСПОЛЬЗУЮТСЯ ПРЕИМУЩЕСТВЕННО НА КЛАСТЕРАХ С АРХИТЕКТУРОЙ x86

Quantum ESPRESSO, бенчмарк PSIWAT

(взаимодействие поверхности золота, покрытой тиолом, с водой)

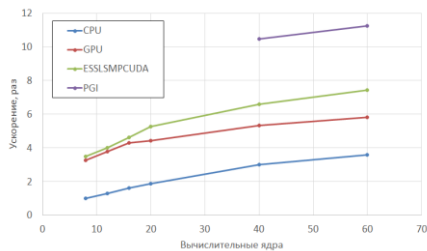


Использование GPU позволяет получить ускорение в 3,1 раза (по сравнению с версией, работающей только на CPU)





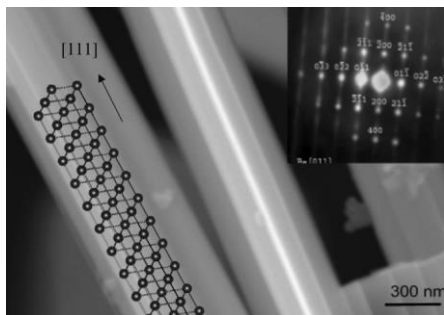
ПАКЕТЫ ПРИКЛАДНЫХ ПРОГРАММ, КОТОРЫЕ ИСПОЛЬЗУЮТСЯ ПРЕИМУЩЕСТВЕННО НА КЛАСТЕРАХ С АРХИТЕКТУРОЙ X86



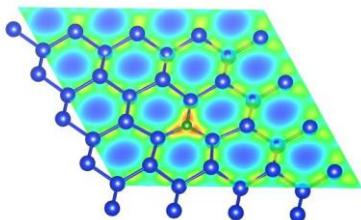
Версии пакета

Версия	Используемая библиотека	Поддержка GPU	Модуль	Сайт проекта
5.4	IBM ESSL	-	espresso/cpu/5.4	http://www.quantum-espresso.org/
5.4	IBM ESSL	+*	espresso/gpu/5.4	https://github.com/fspiga/qe-gpu-plugin
5.4	IBM ESSL (SMPCUDA)	++	espresso/gpu/5.4-esslcuda	http://www.quantum-espresso.org/
6.0	IBM ESSL	+++	espresso/gpu/6.0	https://github.com/RSE-Cambridge/qe-gpu

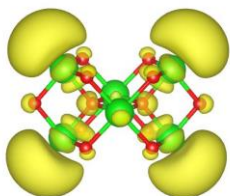
* – число символов "+" соотносится с производительностью пакета (больше – лучше)



Нанопроволоки вольфрама – как интеллектуальные покрытия, литий-ионные батареи и наноструктурные датчики.
Исследованы атомная структура и упругие свойства.



Предсказание свойств новых 2D материалов.
Силицен – как возможный материал для квантовых компьютеров.
Исследованы атомная и электронная структура, упругие свойства.



Наночастицы ZrO_2 , TiO_2 и SiO_2 – как диэлектрический затвор для нанотранзисторов.
Исследована атомная и электронная структура.

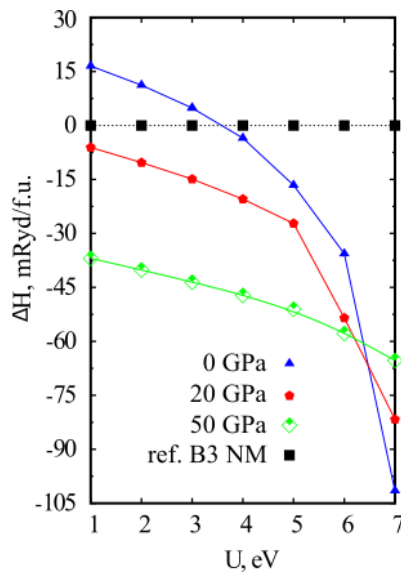
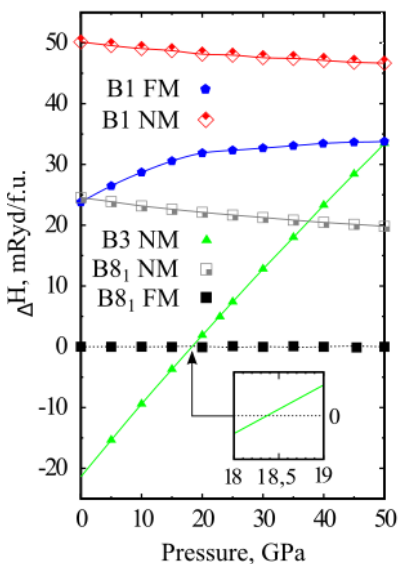
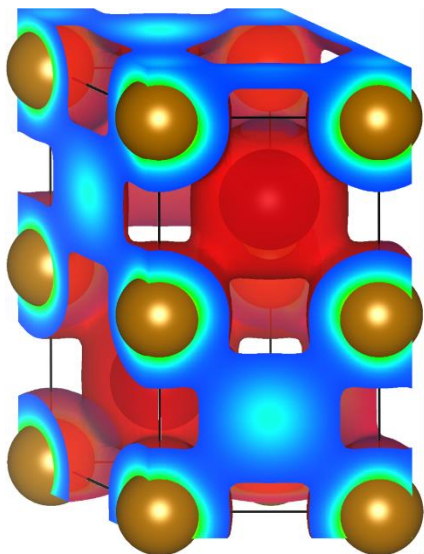
ускорение в 5-6 раз

Лаборатория численных методов математической физики ВЦ ДВО РАН (рук. группы А.Н. Чибисов)



ПЕРВОПРИНЦИПНЫЙ РАСЧЕТ СВОЙСТВ НИТРИДОВ D МЕТАЛЛОВ

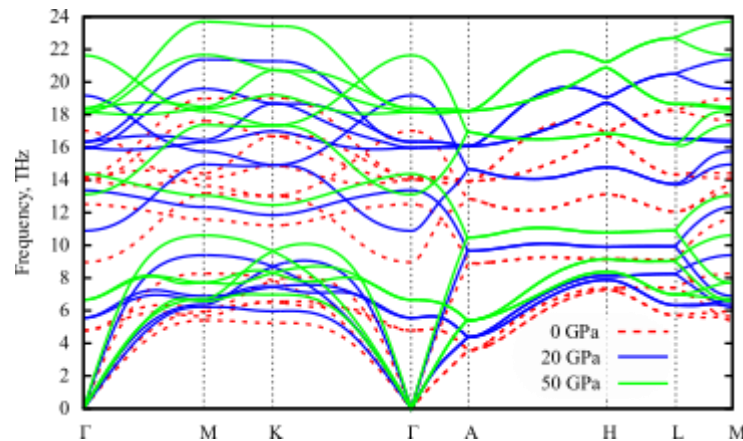
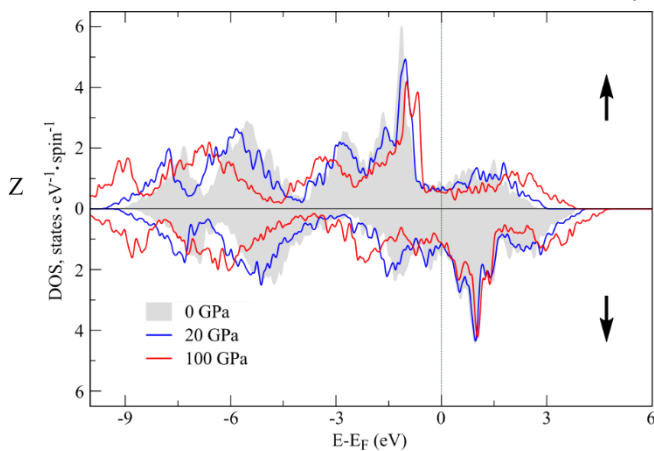
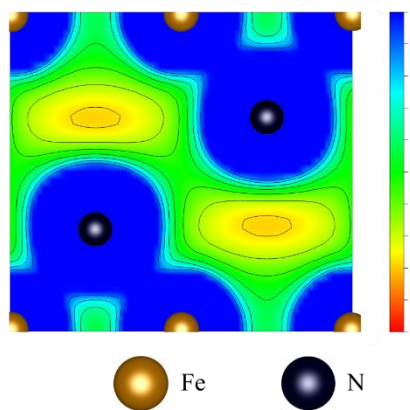
- Ab initio* расчеты в рамках метода DFT
- Изучение электронных, магнитных и динамических свойств
- Фазовые переходы под давлением
- Предсказание стабильности соединений
- Учет влияния сильных электронных корреляций



Национальный
исследовательский
**Томский
государственный
университет**



**QUEEN'S
UNIVERSITY
BELFAST**



СЕГМЕНТИРОВАНИЕ ТРЕХМЕРНЫХ МЕДИЦИНСКИХ ИЗОБРАЖЕНИЙ

Реализованы:

- сверточная нейросеть для сегментации изображений;
- предобработка изображений для обучения;
- сериализация и десериализация обученной нейросети на диск;
- отладочный веб-сервер для визуализации результатов обучения.

Произведен подбор гиперпараметров сети на основании экспериментов с обучением (сеть должна обладать достаточной сложностью для выбора существенной информации из данных, но не переобучаться и не деградировать).

Произведено обучение нейросети на наборе данных LITS (Liver Tumor Segmentation Challenge) - сегментирование печени и новообразований печени.

Результаты:

- достигнута мера Жаккара 0.54 – уровень, сравнимый с мировыми результатами;
- обучение нейросети до точности 0.45-0.5 занимает примерно 12 часов на одном графическом сопроцессоре.

МОЛЕКУЛЯРНО-ДИНАМИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПРОЦЕССА ВЫСОКОСКОРОСТНОГО ВЗАИМОДЕЙСТВИЯ ЧАСТИЦ (АТОМОВ)

Проведен:

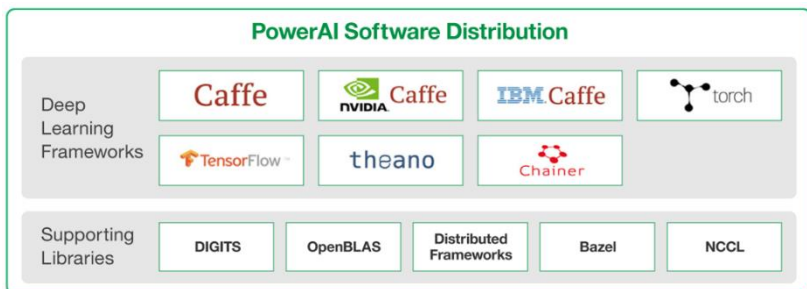
- расчет устойчивой конфигурации состоящей из 0.5 млн.- 50 млн. частиц, связанных потенциалом Леннарда-Джонса (термостатирование);
- МД-расчет устойчивой конфигурации атомов для структур с ковалентным типом химической связи (Si и C) (термостатирование), связанных потенциалом Терсофф;
- МД-моделирование процессов формирования нанокластеров, состоящих из атомов металлов А на поверхности (001) металлов В (Ni, Cu, Au), связанных потенциалом Rosato-Guillope-Legrand.

Результаты:

Достигнутое качество и показатели скорости расчетов позволяют говорить о существенном расширении возможностей практического применения предложенной модели высокоскоростного взаимодействия. В частности, стало возможным увеличить число расчетных частиц до 50 миллионов, что позволило существенно повысить качество вычислительных экспериментов.



ПРОГРАММНОЕ ОКРУЖЕНИЕ



.....



контейнер Singularity

1. Средства параллельных вычислений MPI:
библиотека IBM Spectrum MPI, OpenMPI
2. Компиляторы языков программирования:
IBM XL C/C++, IBM XL Fortran, GNU C/C++,
GNU Fortran, PGI C/C++, PGI Fortran
3. NVIDIA CUDA Toolkit 8.0
4. Математические библиотеки:
IBM ESSL и PESSL
5. Система диспетчеризации заданий:
PBS Professional
6. Мониторинг: Ganglia



GROMACS

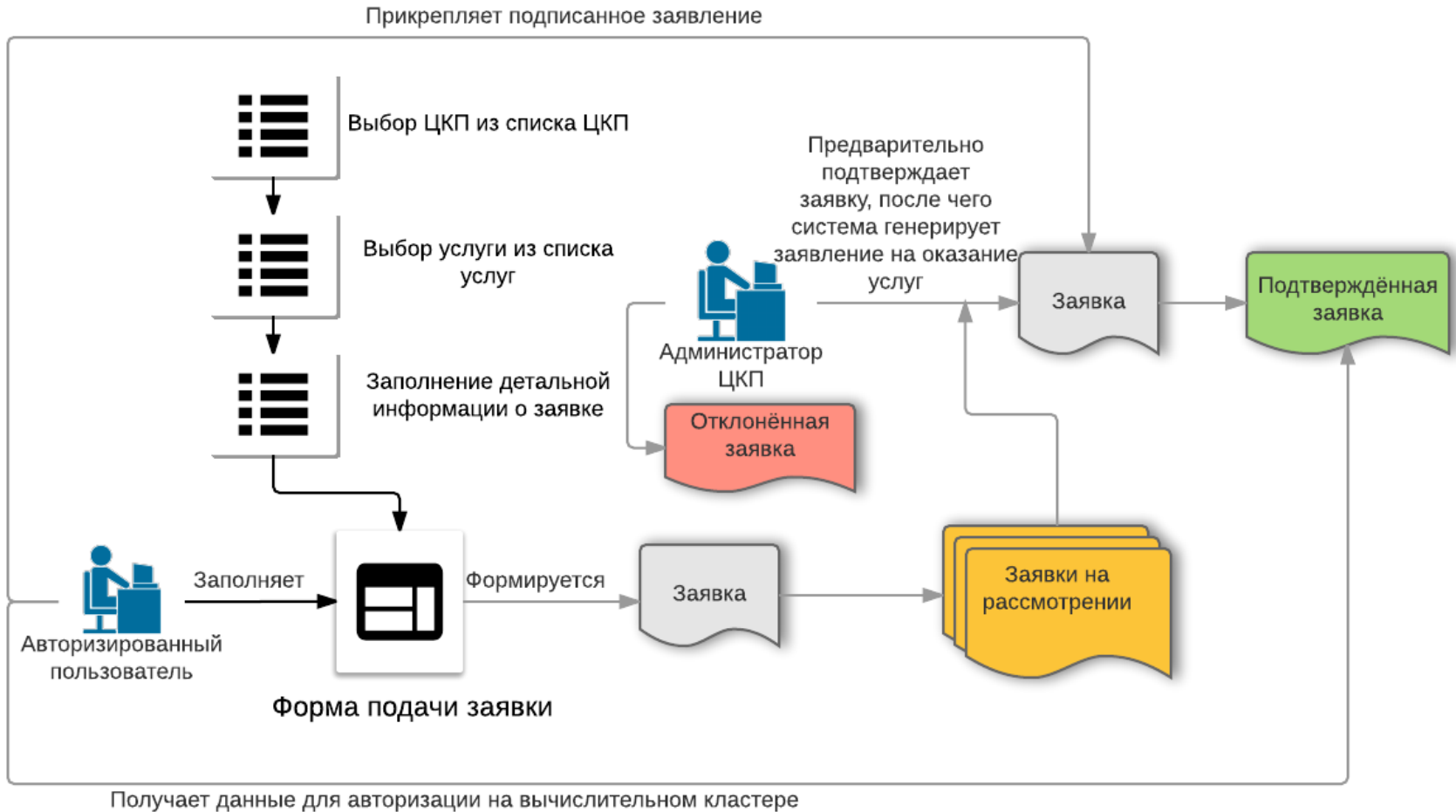
.....



CentOS



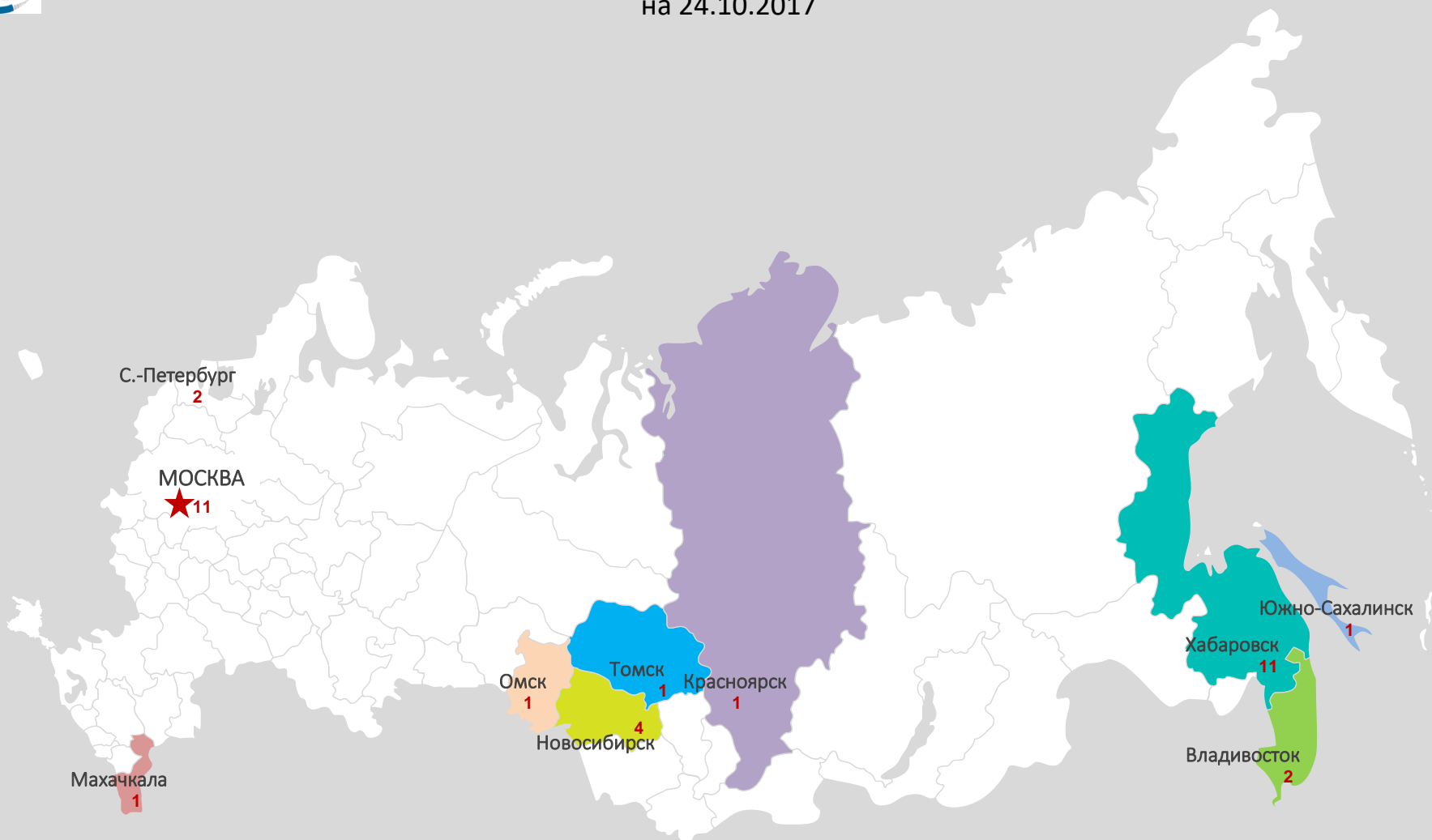
АЛГОРИТМ ПОДАЧИ И УТВЕРЖДЕНИЯ ЗАЯВКИ





ГЕОГРАФИЧЕСКИЙ ОХВАТ ЗАЯВОК

на 24.10.2017



37 заявок



СТРУКТУРА ЗАЯВОК

(приняты или на рассмотрении)

№	Название ВУЗа	шт.
1	Московский государственный университет	2
2	Московский физико-технический институт	1
3	Санкт-Петербургский государственный университет	1
4	Новосибирский национальный исследовательский государственный университет	1
5	Национальный исследовательский Томский государственный университет	1
6	Дальневосточный федеральный университет	1
7	Сибирский федеральный университет	1
8	Тихоокеанский государственный университет	1
9	Омский государственный университет	1
10	Дагестанский государственный университет	1
11	Казанский федеральный университет	1
	Всего заявок	12



СТРУКТУРА ЗАЯВОК

(приняты или на рассмотрении)

№	Название организации	ШТ.
1	Вычислительный центр ДВО РАН (г. Хабаровск)	8
2	ФИЦ Информатика и управление РАН (г. Москва)	3
3	Институт океанологии им. П.П.Ширшова РАН (г. Москва)	2
4	Институт водных проблем РАН (г. Москва)	1
5	Институт морской геологии и геофизики ДВО РАН (г. Южно-Сахалинск)	1
6	Институт теплофизики им. С.С. Кутателадзе СО РАН (г. Новосибирск)	1
7	Институт математики им. С. Л. Соболева СО РАН (г. Новосибирск)	1
8	Объединенный Институт высоких температур РАН (г. Москва)	2
9	Институт прикладной астрономии РАН (г. Санкт-Петербург)	1
10	Институт прикладной математики ДВО РАН (г. Владивосток)	1
11	ФИЦ Институт цитологии и генетики СО РАН (г. Новосибирск)	1
	Всего заявок	22



СТРУКТУРА ЗАЯВОК (приняты или на рассмотрении)

№	Название организации	шт.
1	Дальневосточный филиал ФГУП “Всероссийский научно-исследовательский институт физико-технических измерений”	1
2	Федеральное государственное бюджетное учреждение “Дальневосточное управление по гидрометеорологии и мониторингу окружающей среды”	1
3	ФГУП НИИ “Квант”	1
	Всего	3



ИСПОЛЬЗОВАНИЕ ГРАФИЧЕСКИХ СОПРОЦЕССОРОВ

Гибридная модель вычислений:

- последовательная часть кода выполняется на CPU;
- массивно-параллельные вычисления выгружаются на GPU.

Основные техники:

- использование оптимизированных библиотек (ESSL, NVBLAS)
- использование библиотек из состава CUDA Toolkit
- директивы (OpenACC)
- CUDA/OpenCL расширения C/C++/FORTRAN

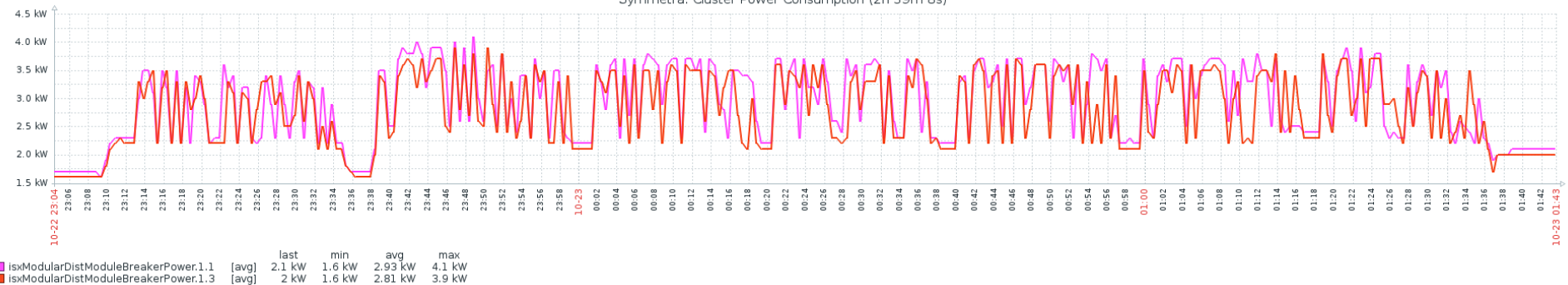
С
Л
О
Ж
Н
О
С
Т
Ь

С
К
О
Р
О
С
Т
Ь



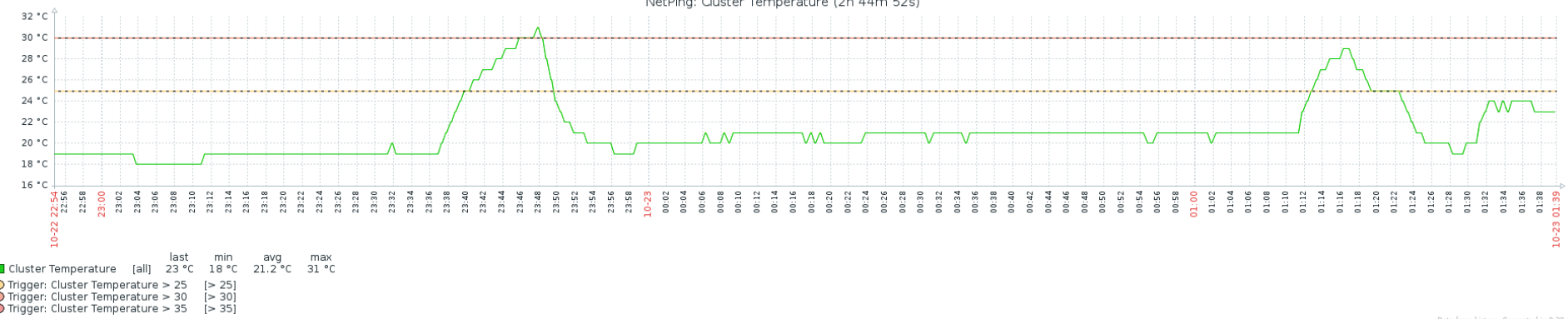
СТАТИСТИКА (ИНЖЕНЕРНАЯ ИНФРАСТРУКТУРА)

Symmetra: Cluster Power Consumption (2h 39m 8s)



Data from history. Generated in 0.59 sec.

NetPing: Cluster Temperature (2h 44m 52s)



Data from history. Generated in 0.28 sec.



КЛЮЧЕВЫЕ РЕЗУЛЬТАТЫ РАБОТ В 2016 г.

1. На базе гибридного кластера удалось сформировать эффективный “универсальный” вычислитель общего пользования. Количество заявок и их динамика говорят о востребованности ресурсов кластера.
2. Новая вычислительная система интегрирована в информационно-телекоммуникационную инфраструктуру, объединяющую крупнейшие национальные образовательные и научно-исследовательские центры страны (МСЦ РАН – ИВТ СО РАН – ВЦ ДВО РАН: 1-10 Гбит). Это дает возможность работать с её ресурсами не только учёным Дальнего Востока, но и специалистам из других регионов России.
3. На реальных научных задачах с использованием авторского и уже разработанного прикладного программного обеспечения, проведены исследования гибридной архитектуры и оценка её эффективности. Определены группы приложений, для которых новый кластер показывает наилучшие показатели производительности. Научные задачи, связанные с использованием технологий и систем ML/DL/AI и CUDA обеспечивают до 100% утилизации ресурсов кластера.



ПЕРСПЕКТИВЫ, ВОПРОСЫ, ПРОБЛЕМЫ

1. Рост числа заявок, востребованность в новых ресурсах на базе гибридных вычислительных систем, отсутствие подобных машин коллективного пользования в топ 50, ставят вопрос о необходимости дальнейшего системного развития проекта. Следующий шаг – рост производительности кластера до 440 Тфлопс (40 узлов).
2. Совместно с IBM и NVIDIA планируется дальнейшее развитие компетенций ЦКП для поддержки и обучения пользователей. Это в том числе влияет на эффективность использования вычислительных ресурсов.
3. В рамках действующего соглашения с МСЦ РАН и ИВТ СО РАН планируется дальнейшее развитие федеральной научной телекоммуникационной инфраструктуры для обмена данными между СКЦ и поддержки специализированных научных информационных систем.
4. Совместно с ФИЦ ИУ РАН будут продолжены работы по формированию распределенной вычислительной инфраструктуры на базе систем OpenPOWER.



ГЛАВНЫЕ ПРОБЛЕМЫ

1. Отсутствуют механизмы финансовой поддержки текущей деятельности ЦКП (электроэнергия, инженерная инфраструктура, продление гарантии на оборудование, персонал, программное обеспечение и т.п.).
2. Отсутствуют четкие правила и условия предоставления доступа к ресурсам (разные ведомства, разные источники финансирования исследований и т.п.).



СПАСИБО ЗА ВНИМАНИЕ!