НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

ОТДЕЛ СУПЕРКОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ НИУ ВШЭ:

О НАСТРОЙКЕ, ОПТИМИЗАЦИИ И РАЗВИТИИ СУПЕРКОМПЬЮТЕРНОГО КОМПЛЕКСА

Начальник отдела
Костенецкий Павел Сергеевич, к.ф.-м.н., доцент

Ведущий инженер
Чулкевич Роман Андреевич

Москва, 4 февраля 2020.

Сокращенный вариант.

ОТДЕЛ СУПЕРКОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ

ОСМ создан в структуре НИУ ВШЭ 14 октября 2019 г. (приказ ректора № 6.18.1-01/1410-08).

Основные задачи отдела

1. Методическая поддержка применения суперкомпьютерных вычислений подразделениями.
2. Администрирование суперкомпьютеров.
3. Администрирование пользователей.
4. Управление документацией и разработка инструкций.

С 21 октября 2019 г. все технические работы, связанные с суперкомпьютером, проводятся силами сотрудников ОСМ.

ХАРАКТЕРИСТИКИ СУПЕРКОМПЬЮТЕРА НИУ ВШЭ

- **6 место в ТОП 50 СНГ**
- Пиковая производительность: **912.4 Терафлопс**
- LINPACK-производительность: **568.5 Терафлопс**
- **26** вычислительных узлов
  - 10 узлов с **1,5 ТБ ОЗУ**
  - 16 узлов с **768 ГБ ОЗУ**
- **2** управляющих узла
- **104** графических процессора **NVIDIA Tesla V100 32 ГБ**
- **56** центральных процессоров **Intel Xeon Gold 6152**
- Оперативная память: **27.7 ТБ**
- Дисковая память: **885 ТБ**
  - параллельная СХД на базе Lustre: **840 ТБ**
  - SSD в узлах: **2**x240 ГБ в RAID1
  - SSD в управляющих узлах: **20x1,7 ТБ**
- Коммуникационная сеть: **InfiniBand** EDR (**2x100 Гбит/с**, топология **Fat Tree**)

ОТЛИЧИЯ СУПЕРКОМПЬЮТЕРА НИУ ВШЭ ОТ ДРУГИХ

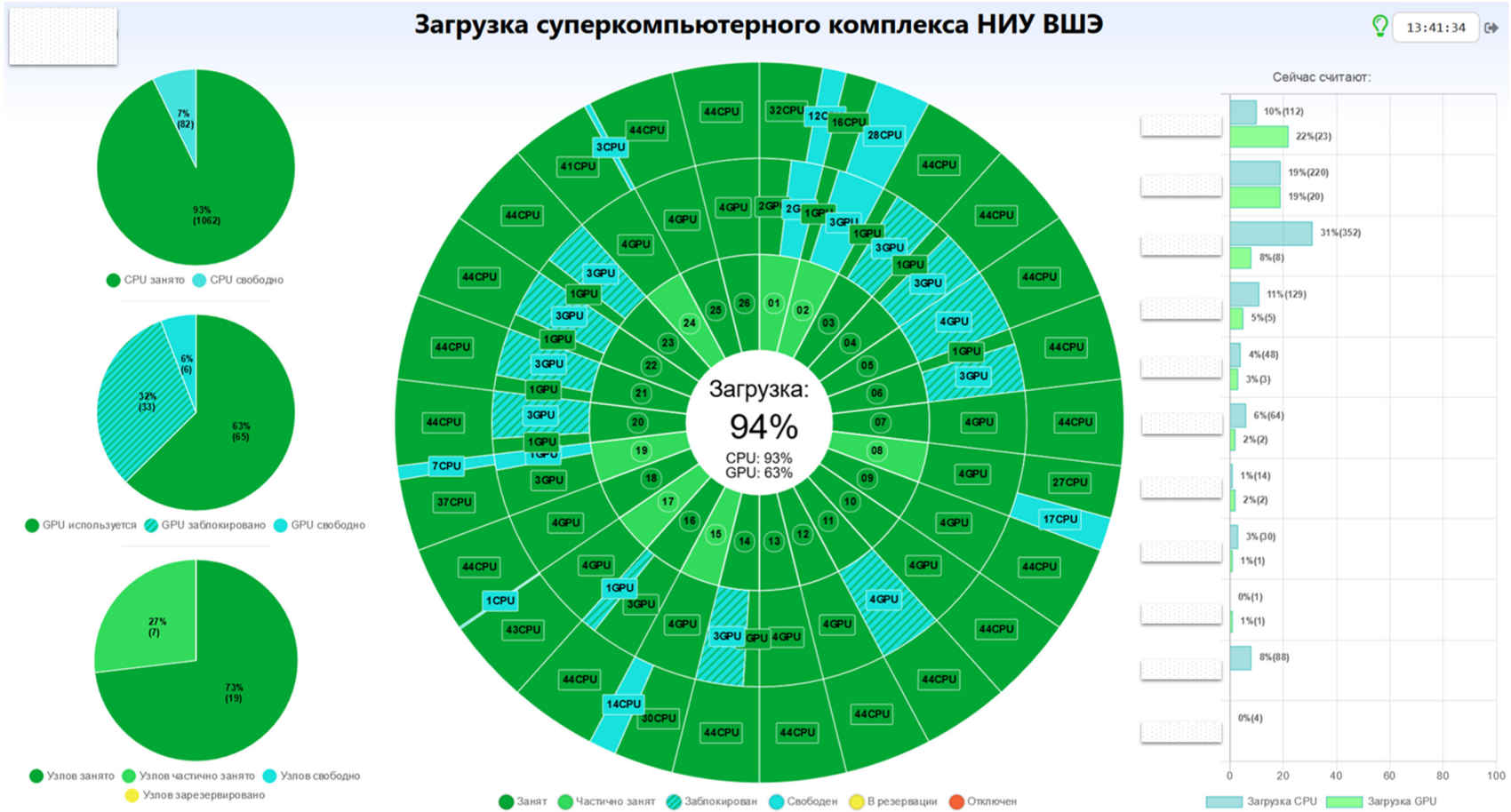
Аппаратные отличия

- Очень большая оперативная память в вычислительных узлах – до **1,5 ТБ.**
- По 4 самых современных **GPU Tesla V100 32GB**
- Сеть InfiniBand EDR
  - **100 Гбит/с**,
  - топология **толстое дерево**,
  - **две сетевые карты IB** в каждом узле.
- В вычислительных узлах **по 2 SSD**.
- Всё вычислительное оборудование **DELL**.

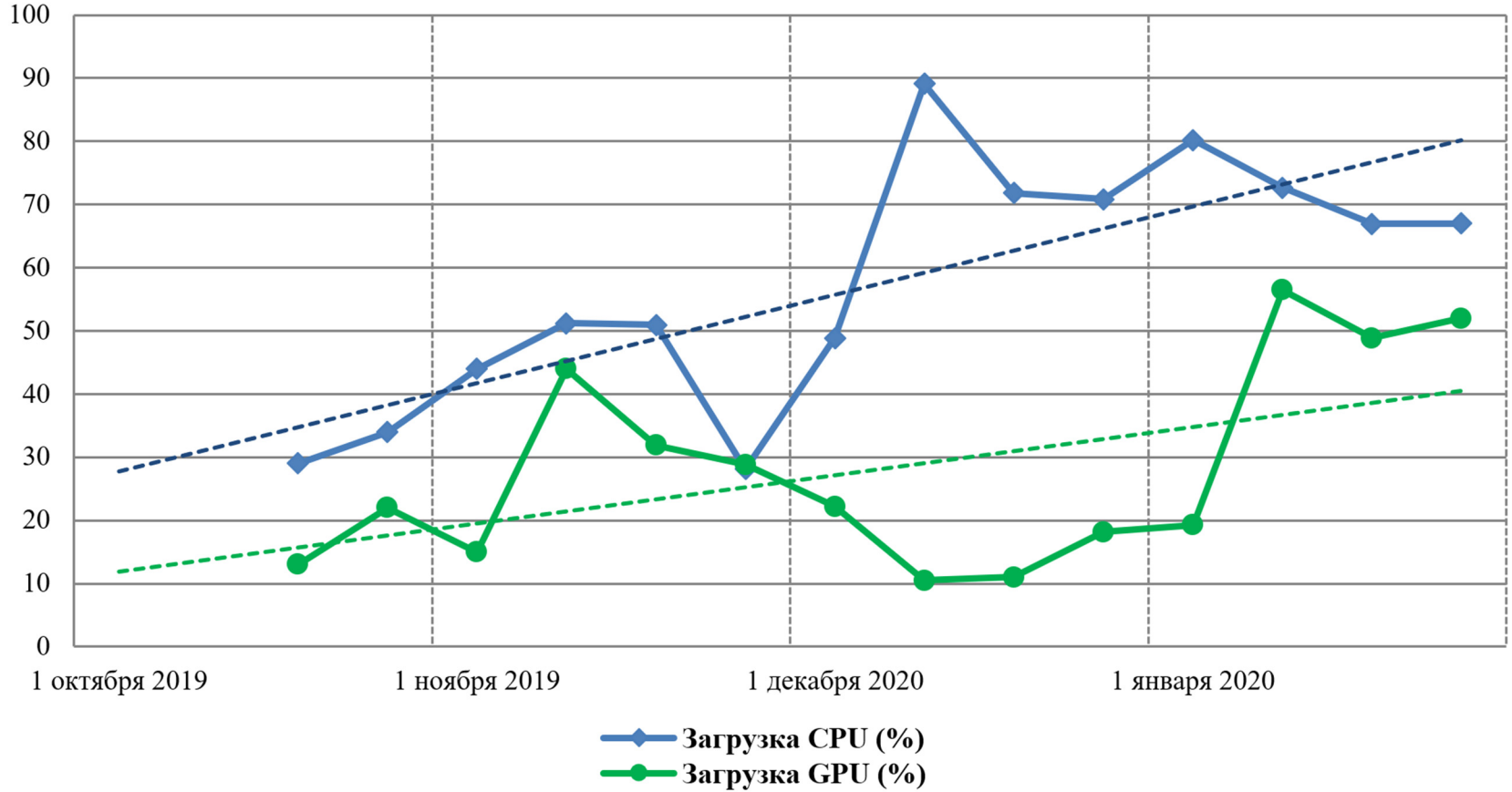
Программное отличие

- На каждом вычислительном узле могут работать **сразу несколько задач разных пользователей**.

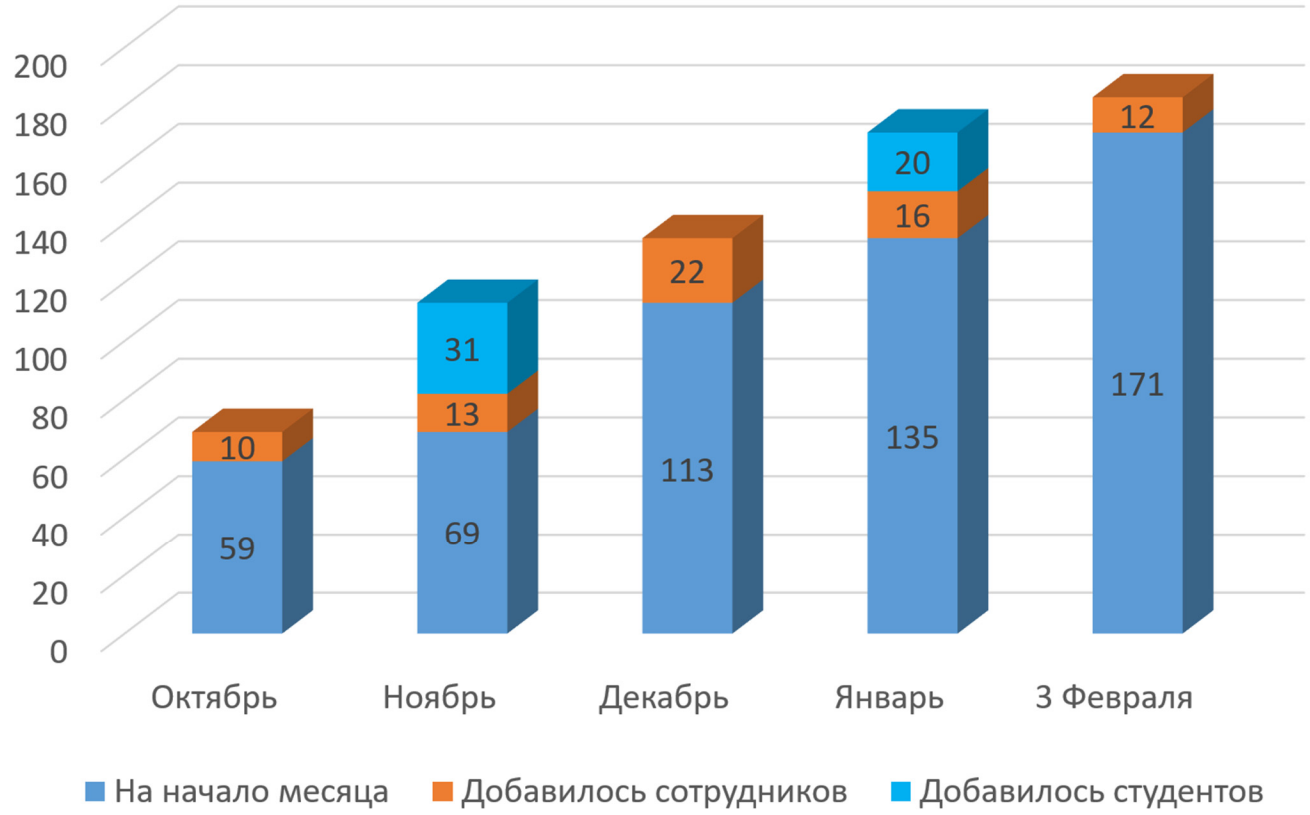
СОБСТВЕННАЯ СИСТЕМА МОНИТОРИНГА



РОСТ ЗАГРУЗКИ СУПЕРКОМПЬЮТЕРА



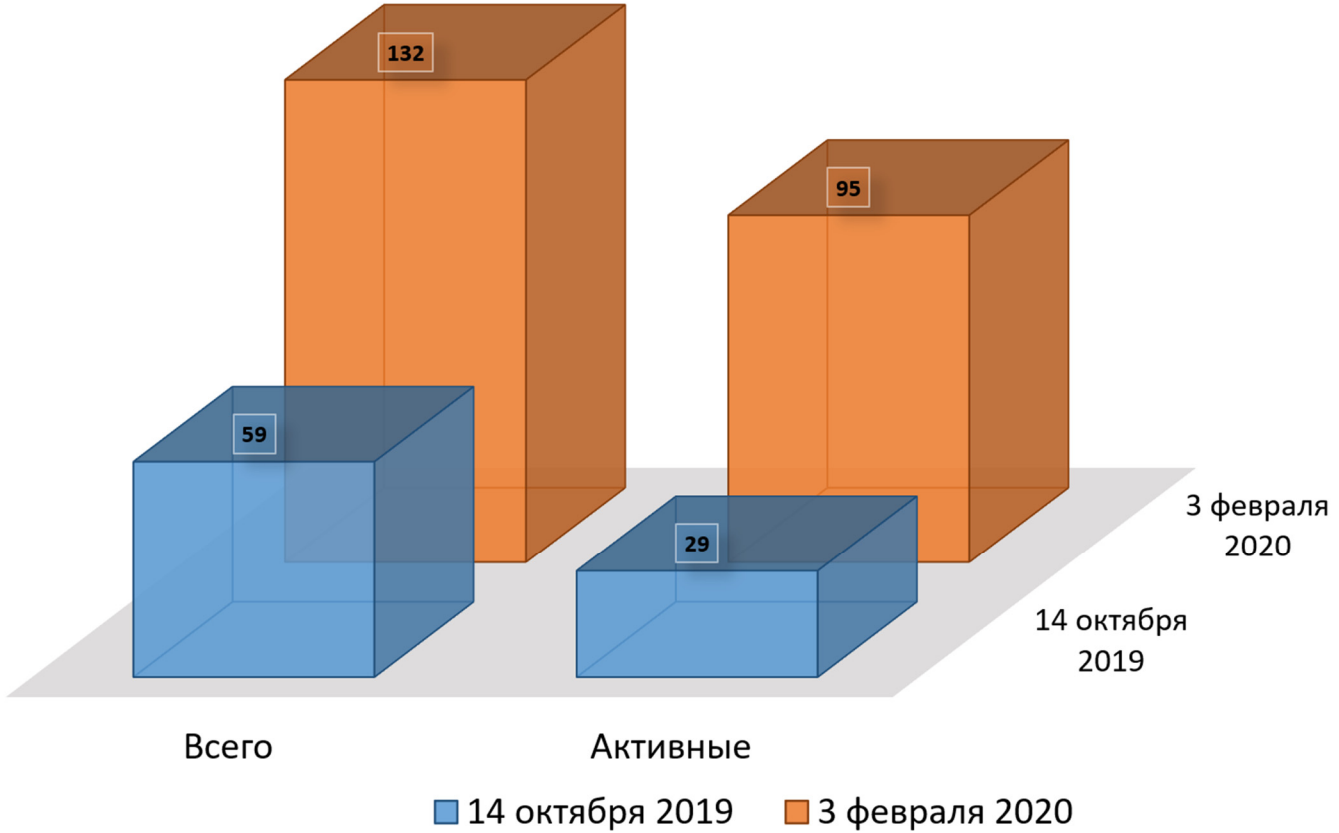
ПРИРОСТ КОЛИЧЕСТВА ПОЛЬЗОВАТЕЛЕЙ



На 3 февраля 2020 г.
171 пользователь:

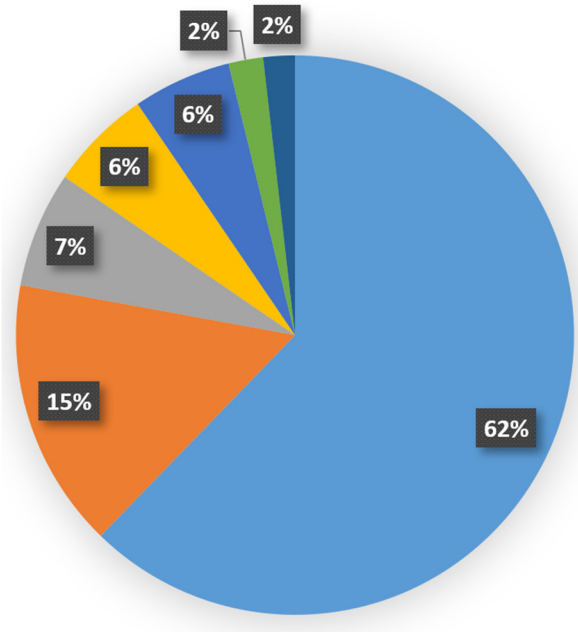
- 132 сотрудника
- 51 студент

ПРИРОСТ КОЛИЧЕСТВА АКТИВНЫХ ПОЛЬЗОВАТЕЛЕЙ-СОТРУДНИКОВ

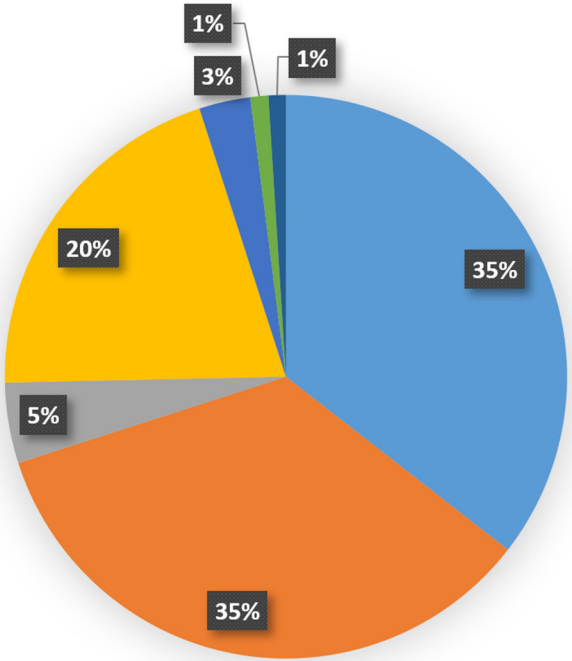


ЗАГРУЗКА ПО ПОДРАЗДЕЛЕНИЯМ

Загрузка CPU

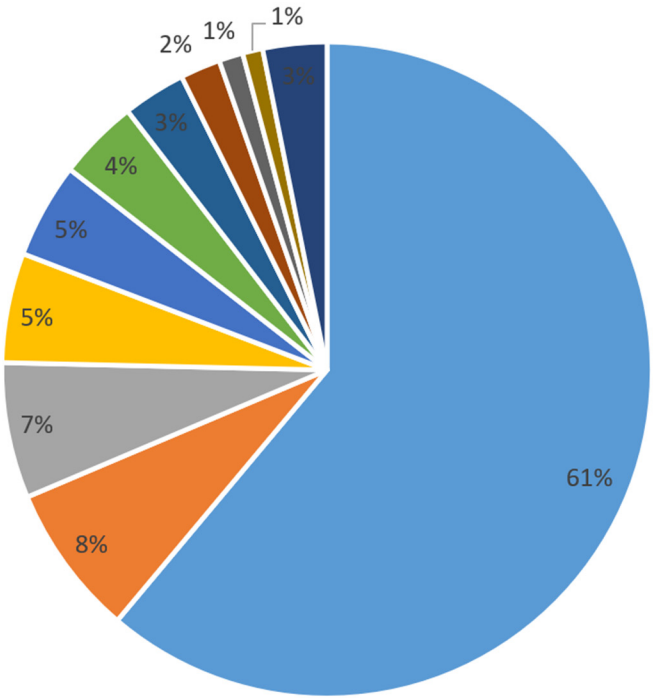


Загрузка GPU



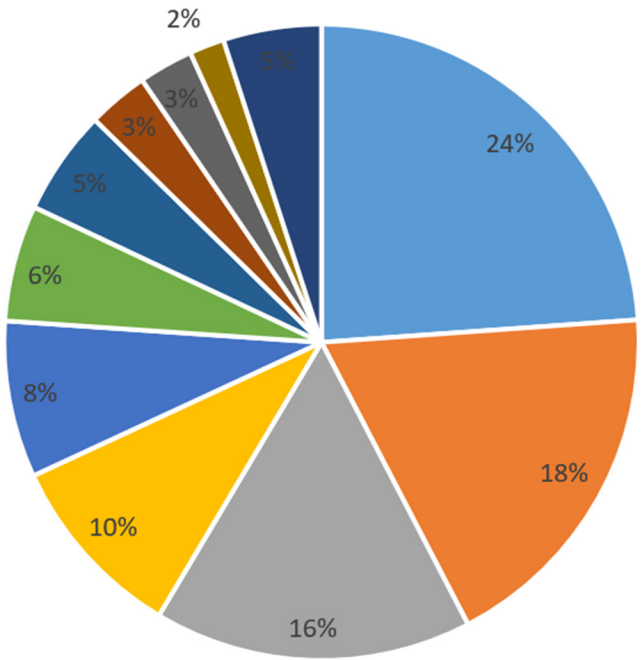
- Международная лаборатория суперкомпьютерного атомистического моделирования и многомасштабного анализа
- ФКН, Центр глубинного обучения и байесовских методов
- Международная лаборатория теории игр и принятия решений
- ФКН, Лаборатория компании Самсунг
- СПб, Центр анализа данных и машинного обучения
- Отдел информационно-аналитических систем (Институт статистических исследований и экономики знаний, Центр стратегической аналитики и больших данных)
- Другие

ЗАГРУЗКА ПО ПРОЕКТАМ – CPU



- Атомистическое многомасштабное моделирование материалов
- Large-scale DNN ensembling study
- ИИ-законодатель
- Исследование гипотезы о лотерейных билетах для обучения нейронных сетей
- Learn to move
- Методы обнаружения объектов новых классов
- Обусловленные Нормализационные Потоки для Задачи Ч
- Обучение pretrained модели BERT на данных более 300 млн научных материалов
- Атомистическое и континуальное моделирование в задачах физики
- Embeddings for code diffs
- Другие

ЗАГРУЗКА ПО ПРОЕКТАМ – GPU



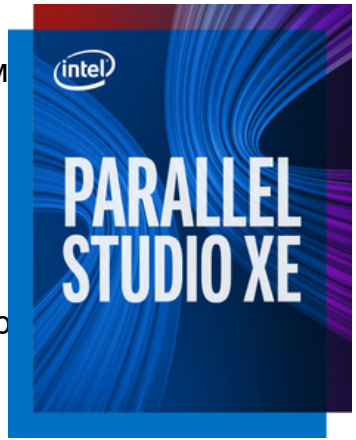
- Исследование гипотезы о лотерейных билетах для обучения нейронных сетей
- Методы обнаружения объектов новых классов
- Атомистическое многомасштабное моделирование материалов
- Large-scale DNN ensembling study
- Обусловленные нормализационные потоки для задачи Ч
- Проект Шпильмана
- Эффективное использование данных по взаимодействиям со средой в обучении с подкреплением
- Learn to move
- Методы глубинного обучения для семантического парсинга текста и синтеза программ
- Атомистическое и континуальное моделирование в задачах физики
- Другие

INTEL PARALLEL STUDIO XE 2020 CLUSTER EDITION

В январе 2020 г. приобретен комплект Intel Parallel Studio XE 2020 Cluster Edition (for Linux – Floating Academic 2 seats).

Что это дает

1) Компиляторы Intel – прирост производительности сборок по сравнению с GNU.
2) Intel MKL – выигрывающий по производительности у BLAS+LAPACK/ATLAS/OpenBLAS.
3) Intel MPI – выигрывающий у других реализаий MPI.
4) Набор ПО для профилирования, анализа и оптимизации разрабатываемых програм
5) Оптимизированный Python.



Особенности использования

Загрузка модуля: *module load INTEL/parallel_studio_xe_2020_ce*

Уже доступен консольный запуск:

- *Intel Trace Analyzer and Collector* – профилировщик производительности для MPI пр
- *Advisor XE* – средство для моделирования параллельной работы;
- *Inspector XE* – отладчик (динамический анализ кода);
- *Vtune Amplifier XE* – сбор и анализа данны о производительности кода;

В планах

1) Разработка базовой инструкции по использованию ПО;
2) Доступ к GUI профилировщика.

PYTHON: JUPYTER NOTEBOOK

Реализована возможность упрощенного запуска JUPYTER NOTEBOOK на вычислительных узлах комплекса

Использование: только для запуска уже подготовленных расчетов на суперкомпьютере. Запрещается использовать установленный на суперкомпьютере JUPYTER NOTEBOOK для учебной работы и для процесса разработки!

Что это дает:

1) удобный интерфейс(web) для интерактивного исполнения python кода;
2) использование графических возможностей (например, построение графиков);
3) JUPYTER видит все GPU на узле (выделенные пользователю через очередь).

```
In [1]: from tensorflow.python.client import device_lib

        def get_available_gpus():
            local_device_protos = device_lib.list_local_devices()
            return [x.name for x in local_device_protos if x.device_type == 'GPU']

In [2]: get_available_gpus()
Out[2]: ['/device:GPU:0', '/device:GPU:1', '/device:GPU:2', '/device:GPU:3']
```

Особенности запуска:

1) module load Python/Anaconda_v10.2019
2) sbatch <привычные флаги ресурсов> run_notebook <номер порта: 1400-5000>
3) «forwarding» порта на локальную машину
4) доступ из браузера локальной машины;

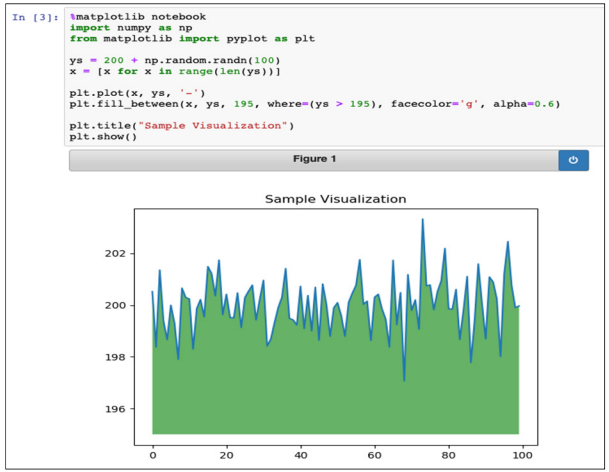
```
localhost:8080/notebooks/Untitled5.ipynb?kernel_name=python3
```

```
In [3]: %matplotlib notebook
        import numpy as np
        from matplotlib import pyplot as plt

        ys = 200 + np.random.randn(100)
        x = [x for x in range(len(ys))]

        plt.plot(x, ys, '-')
        plt.fill_between(x, ys, 195, where=(ys > 195), facecolor='g', alpha=0.6)

        plt.title("Sample Visualization")
        plt.show()
```



НОВАЯ СИСТЕМА ОЧЕРЕДЕЙ ЗАДАЧ

С 27 декабря 2019 г. на суперкомпьютере установлена новейшая система очередей задач SLURM 19.05.05 (ранее была SLURM 17)

Что это дает:

1) Возможность планирования GPU-ресурсов подобно ядрам CPU;
2) Возможность учитывать архитектурные особенности вычислительных узлов;
3) Возможность движения в сторону выделения процессов на GPU (а не целиком);
4) Возможность корректного запуска нескольких задач из разных разделов очереди на 1 узле;

Особенности использования

1) Больше нет необходимости в обязательном порядке указывать флаг --gres;
2) Новые флаги для постановки задач:
   - *--gpus* (или -G) – количество GPU для задачи;
   - *--gpus-per-node* – количество GPU на каждый выделяемый узел;
   - *--gpus-per-task* – количество GPU на каждый процесс;
   - *--cpus-per-gpu* – количество CPU на каждый выделенный GPU;
   - *--gpu-bind* – не подходит для нашей архитектуры (не влияет на результат);
3) Использованию флага «-s» или «--oversubscribe» больше не рекомендуется.

slurm
workload manager
v.19.05.05

Планы

Удалить все очереди кроме *normal*. Уже настроена отдельная система, более успешно управляющая ограничениями.

НОВАЯ СИСТЕМА ОЧЕРЕДЕЙ ЗАДАЧ

Информацию по использованию флага «-s» для разделения ресурсов (он же «--oversubscribe=yes») :

1) ранее, когда в старой системе очередей v.17 задачи плохо распределялись по узлам, пользователи начинали использовать данный флаг для ускорения постановки задачи к выполнению;
2) флаг означает возможность использования одного и того же ядра CPU сразу несколькими процессами, что вызывает замедление обоих расчетов.

Вопрос: Почему же не запускается задача, даже при наличии ресурсов и указании флага «-s»?

Ответ: Задачи, запущенные из разных разделов очереди (например, normal и gpu-1) следуют правилу:

	Первая задача с разделением ресурсов	**Первая задача обычная**
Первая задача с разделением ресурсов	Обе задачи могут выполняться на 1 узле и «делиться» ресурсами	Задачи не могут выполняться на 1 узле
Вторая задача обычная	Задачи не могут выполняться на 1 узле	Обе задачи могут выполняться на 1 узле, но без использования одних и тех же ресурсов

slurm workload manager
v.19.05.05

Вывод: использовать режим oversubscribe в версии 19 на суперкомпьютере НИУ ВШЭ нецелесообразно.

ПРЕДЛОЖЕНИЕ ПО УПЛОТНЕНИЮ ПОТОКА ЗАДАЧ

ПРИМЕР

Есть задача, требующая 220 ядер CPU и не требующая GPU.

Обычный запуск:

sbatch –n 220 my_task.sh

Выделится 220 ÷ 44 = 5 вычислительных узлов, на которых другие пользователи уже не смогут взять ни одной GPU (т.к. нет ни одного свободного CPU, которой смог бы обслуживать GPU).

Компромиссный запуск:

sbatch –n 220 --ntasks-per-node=40 my_task.sh

Выделится $\lceil 220 \div 40 \rceil = 6$ вычислительных узлов, на которых другие пользователи смогут выделить от 1 до 4 GPU.

(реально задача запустится на $6 \times 40 = 240$ ядрах).

Наиболее гуманный запуск:

sbatch –n 220 --ntasks-per-node=28 my_task.sh

Выделится $\lceil 220 \div 28 \rceil = 8$ вычислительных узлов, на которых параллельно с вашим расчетом можно полноценно взять все 4 GPU.

(реально задача запустится на $8 \times 28 = 224$ ядрах).

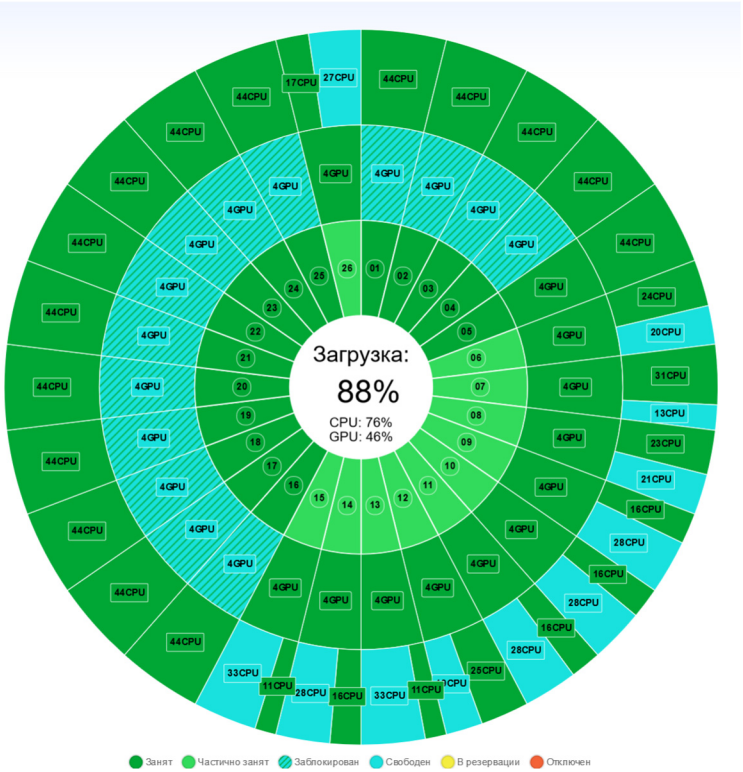


Рис.1 Пример столкновения интересов разных типов пользователей (CPU vs GPU)

КАК СТАТЬ ПОЛЬЗОВАТЕЛЕМ

- Бесплатно для штатных работников, аспирантов и студентов НИУ ВШЭ.
- Для сотрудников и студентов-исследователей обязателен научный проект.
  - ➢ Руководителем проекта может быть сотрудник,
  - ➢ Возможны индивидуальные проекты.
- Студентов, проходящих учебную дисциплину, регистрирует преподаватель. Для студентов создаются учетные записи вида student-1…student-N на один семестр. Преподаватель сам распределяет учетные записи по студентам.
- Обрабатываются заявки пришедшие только с доменов @hse.ru и @edu.hse.ru
- Формы регистрации доступны на странице отдела суперкомпьютерного моделирования: https://www.hse.ru/org/hse/se/hpc

В планах: разработка online-системы регистрации пользователей.

ОТЧЕТЫ И ЕЖЕГОДНАЯ ПЕРЕРЕГИСТРАЦИЯ

- Научный или научно-практический проект регистрируется на срок до одного года
- Проект продлевается на следующий период после представления краткого отчета:
  1) один слайд в PowerPoint с научно-популярным описанием задачи и полученных результатов (в примечании к слайду указываются контакты исполнителей и подробности о задаче на полстраницы);
  2) такой же слайд, переведенный на английский язык;
  3) список статей со ссылкой на суперкомпьютер НИУ ВШЭ.

Во всех статьях пользователей обязательно должна быть ссылка на суперкомпьютер в разделе благодарностей:

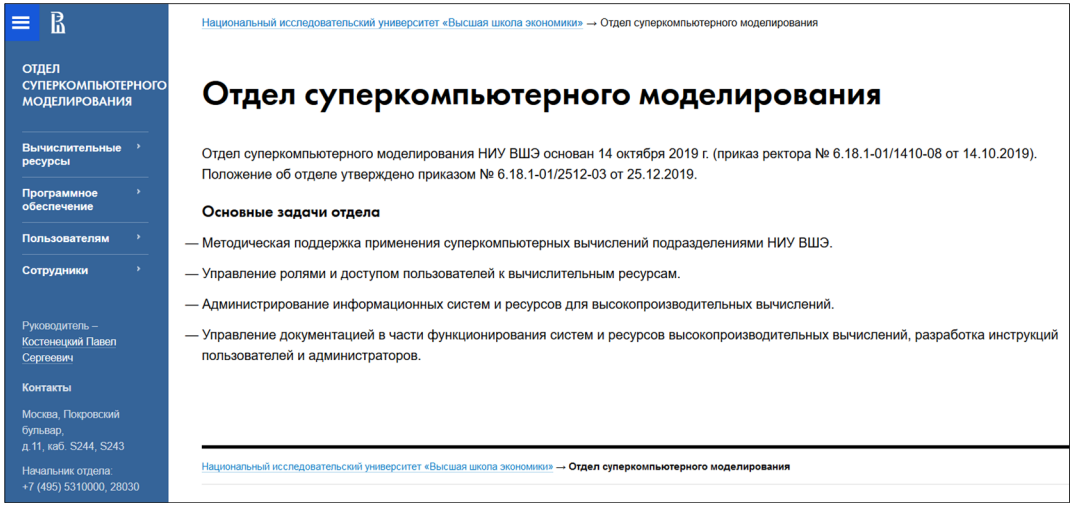
- «Работа выполнена с использованием суперкомпьютерного комплекса НИУ ВШЭ» или
- «This research was supported in part through computational resources of HPC facilities provided by NRU HSE».

РАСШИРЕНИЕ СУПЕРКОМПЬЮТЕРА

- Расширение запланировано на 2021 г.
- До конца 2020 г. будет собираться статистика о загрузке задачами пользователей ресурсов суперкомпьютера (RAM, GPU, CPU, IB).
- На базе статистики будет определена оптимальная конфигурация комплекта расширении.

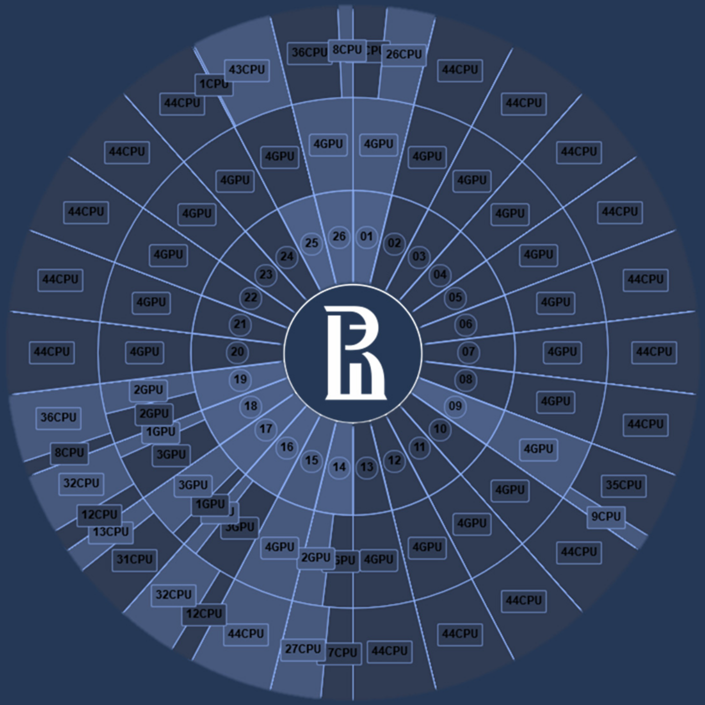
САЙТ ОТДЕЛА СУПЕРКОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ

https://hse.ru/org/hse/se/hpc



- Документы для регистрации
- Характеристики оборудования
- Инструкции
- История изменения конфигурации

О проблемах и пожеланиях сообщайте: HPC@hse.ru



https://hse.ru/org/hse/se/hpc
pkostenetskiy@hse.ru

Телефон: +7(495)7729590 (28030)

Адрес: АУК «Покровский бульвар», S244, S243