Высокопроизводительные компьютерные системы в задачах машинного обучения

Александр Московский

DLCP 2023, Санкт-Петербург

22 июня 2023 г.

Высокопроизводительные системы с 2009 года

Разработка инновационных, энергоэффективных, высокопроизводительных и высокоплотных вычислительных систем для решения уникальных задач

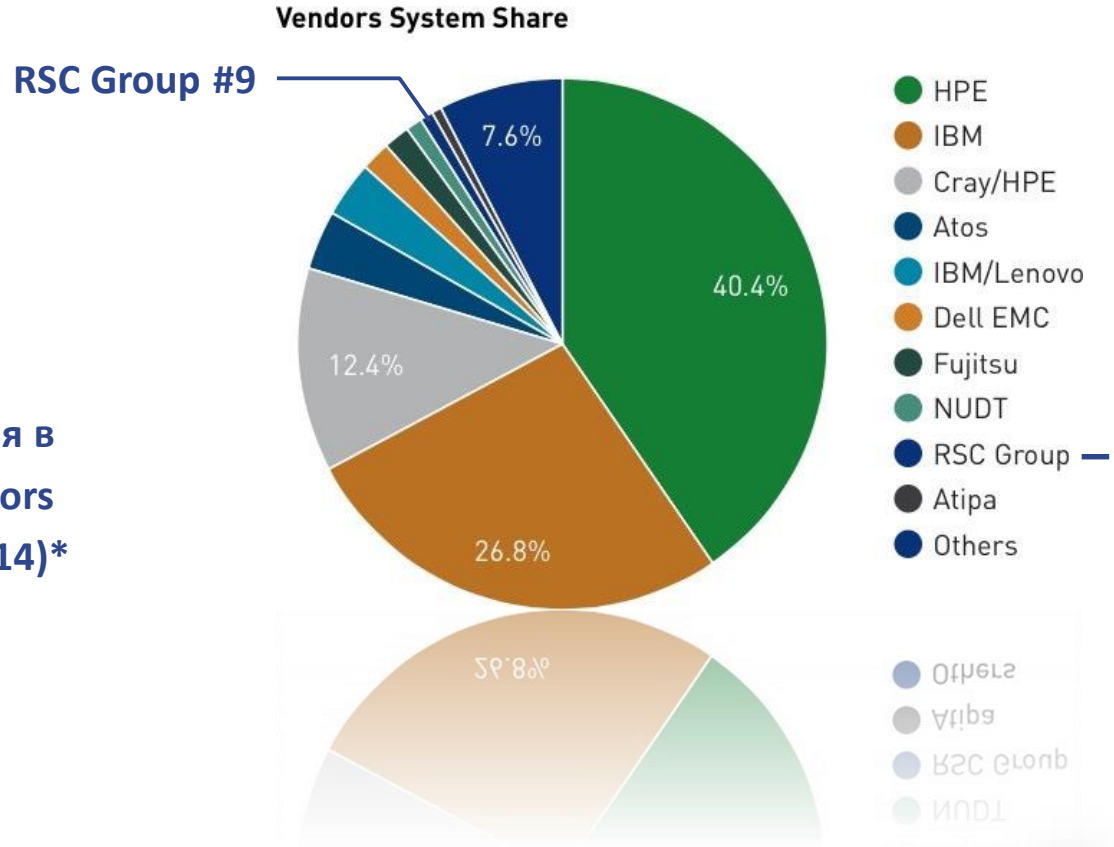
О группе компаний РСК

Ведущий российский разработчик и интегратор инновационных суперкомпьютерных решений

Russian DC Awards 2020 в номинации «Лучшее ИТ-решение для ЦОДа»

TOP500

Единственная российская компания в мировом рейтинге Top10 HPC Vendors System Share by Top500 (Ноябрь 2014)*



*** Топ 10 поставщиков по объему рынка**
https://www.top500.org/statistics/list/

Рыночные позиции

Ведущий российский разработчик и интегратор инновационных суперкомпьютерных решений

Joint Institute for Nuclear Research

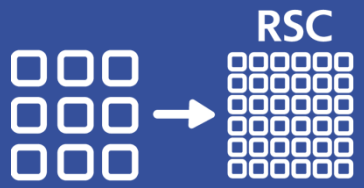
Наиболее **энергоэффективная система в России**

Больше **70%** всех российских систем в **мировом рейтинге HPCG**

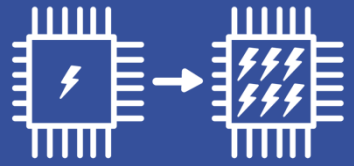
4 системы РСК – единственные представители России в **мировом рейтинге IO500**

24% доля в **российском рейтинге Top50**

Ключевые характеристики решений



Вычислительная плотность



Энергетическая плотность

Энергоэффективность

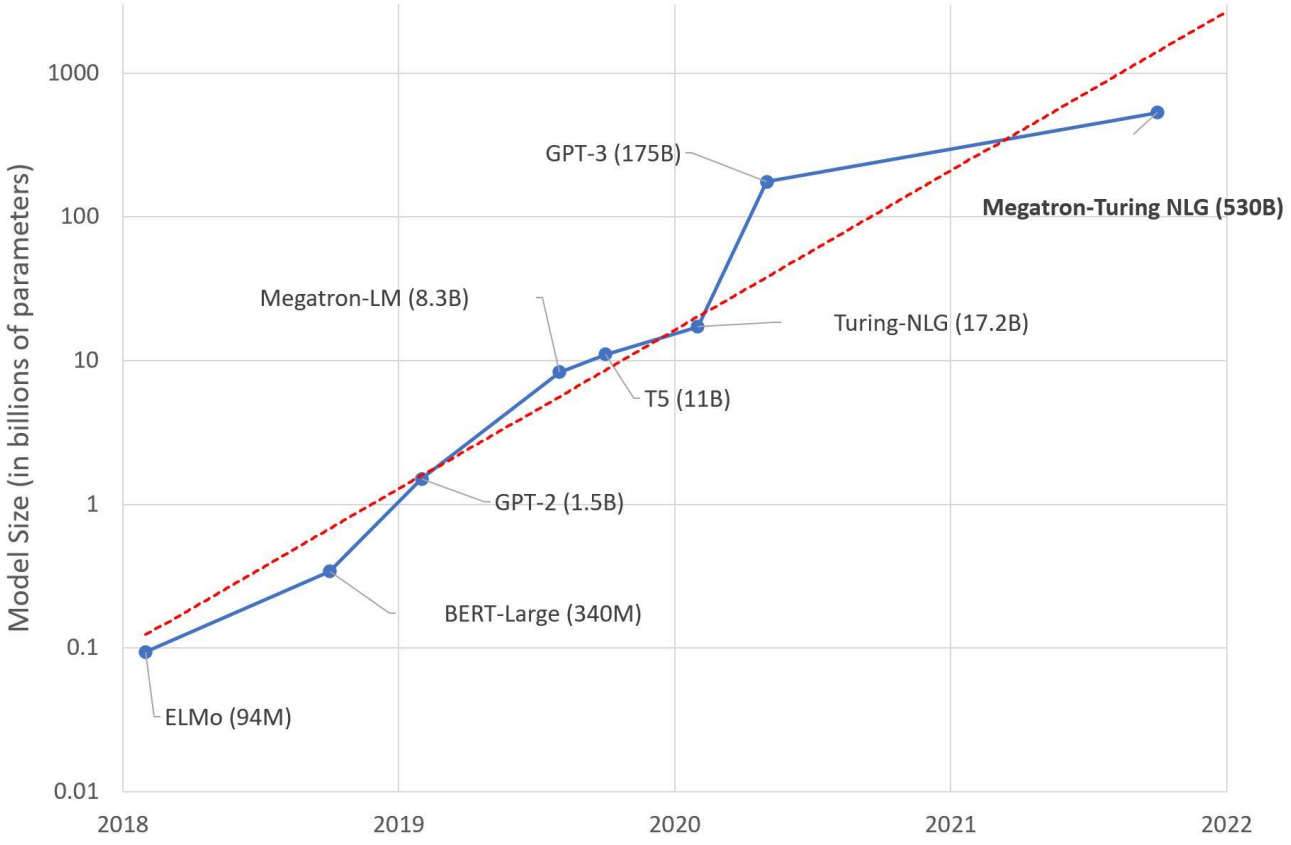
Легкость управления и обслуживания

Надежность

Машинное обучение с использованием высокопроизводительных систем

Большие модели и быстрый рост

- Сложность обучения модели-транфсформера [1]:
  - $C \approx 6ND$
  - *N – число параметров, D- число токенов*
- Для GPT-3 – $3{,}14*10^{23}$ операций
- Размер набора для обучения ~ терабайты
- Трансформеры – не только NLP (см., например, GenSLM [2])

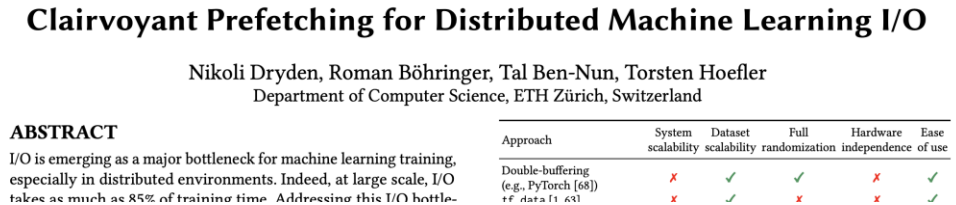


[1] – J. Caplan et. al “Scaling Laws for Neural Language Models” https://doi.org/10.48550/arXiv.2001.08361

[2] – M. Zvyagin et. al. “GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics” https://doi.org/10.1101/2022.10.10.511571

Проблема ввода-вывода

- Большие объемы наборов данных для обучения
- Необходим параллельный доступ к данным при обучении
- Доступ может быть нерегулярным
- Некоторые задачи ограничены доступом к данным



Clairvoyant Prefetching for Distributed Machine Learning I/O

Nikoli Dryden, Roman Böhringer, Tal Ben-Nun, Torsten Hoefler
Department of Computer Science, ETH Zürich, Switzerland

ABSTRACT

I/O is emerging as a major bottleneck for machine learning training, especially in distributed environments. Indeed, at large scale, I/O takes as much as 85% of training time. Addressing this I/O bottle-

Approach	System scalability	Dataset scalability	Full randomization	Hardware independence	Ease of use
Double-buffering (e.g., PyTorch [68])	✗	✓	✓	✗	✓
tf.data [1, 63]	✗	✓	✗	✗	✓

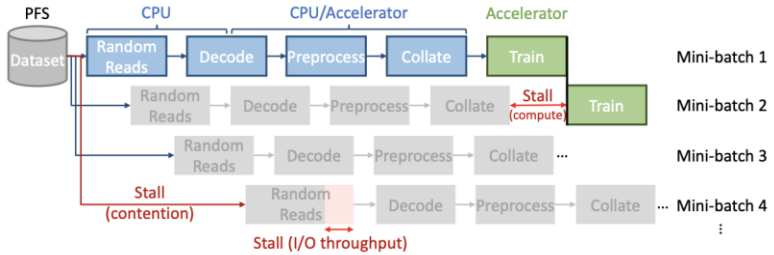
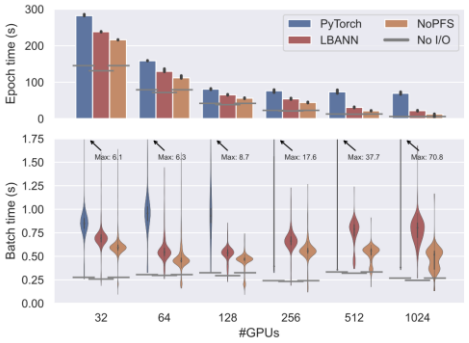
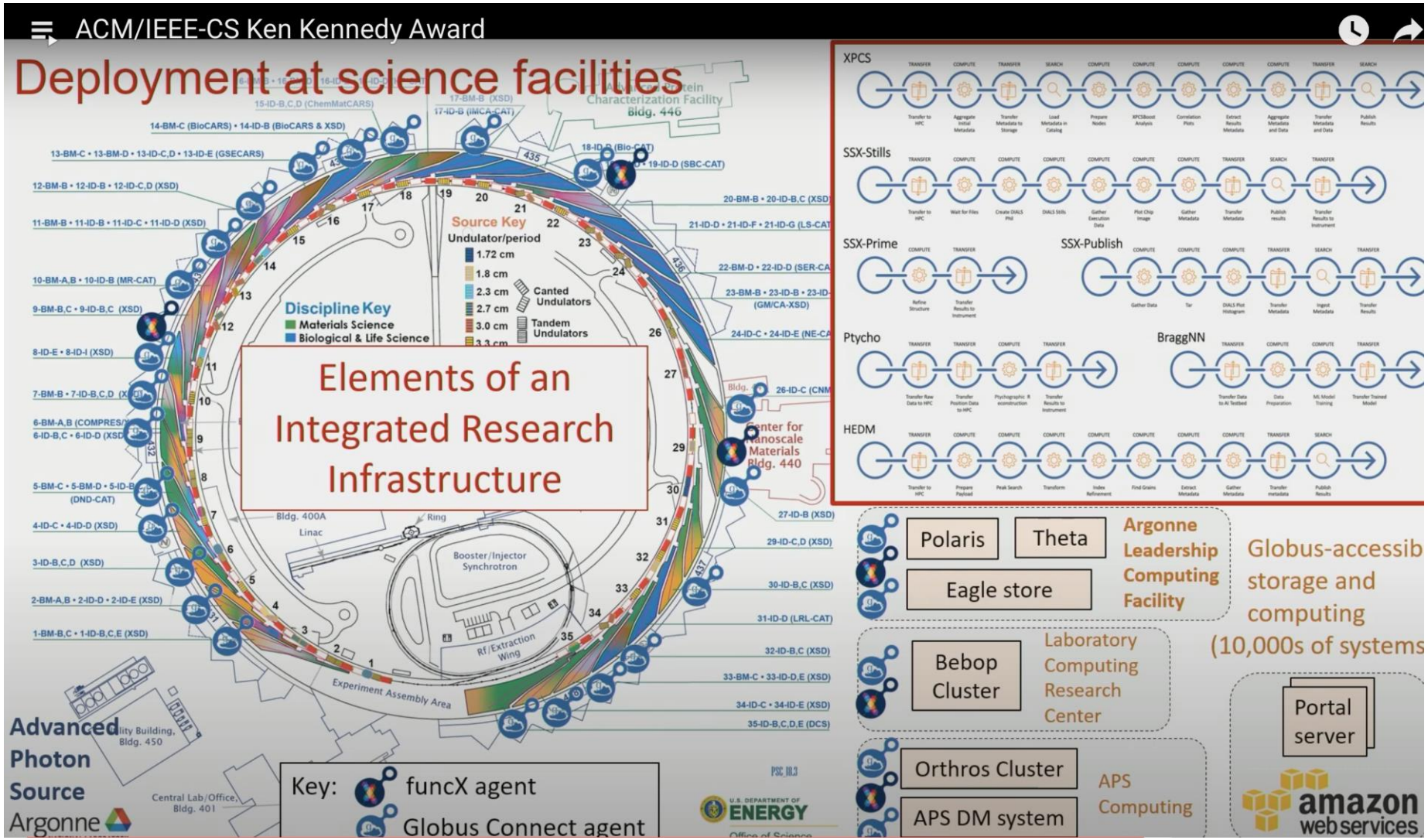


Figure 2: Overview of I/O pipelines and potential issues.



http://htor.inf.ethz.ch/publications/img/dryden-nopfs.pdf

Сложные конвейеры обработки

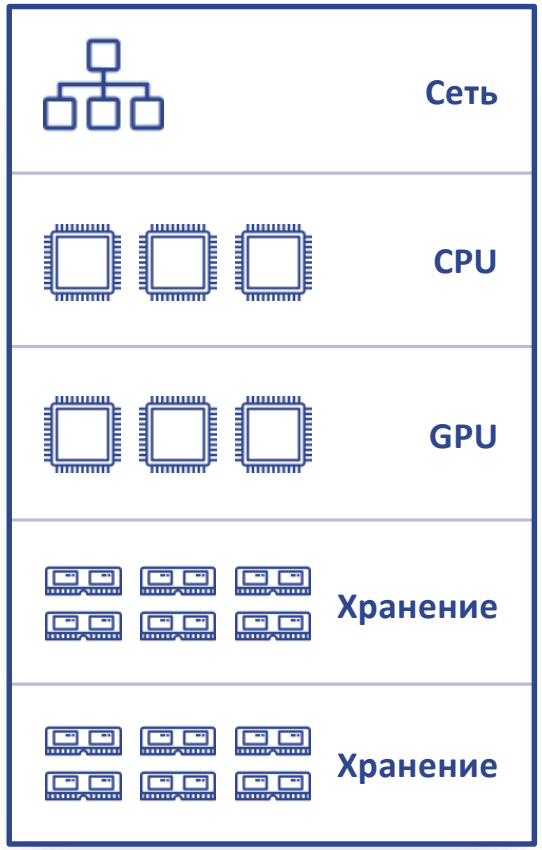


Ian Foster Ken Kennedy award talk, SC22 Обработка данных на Advanced Photon Source

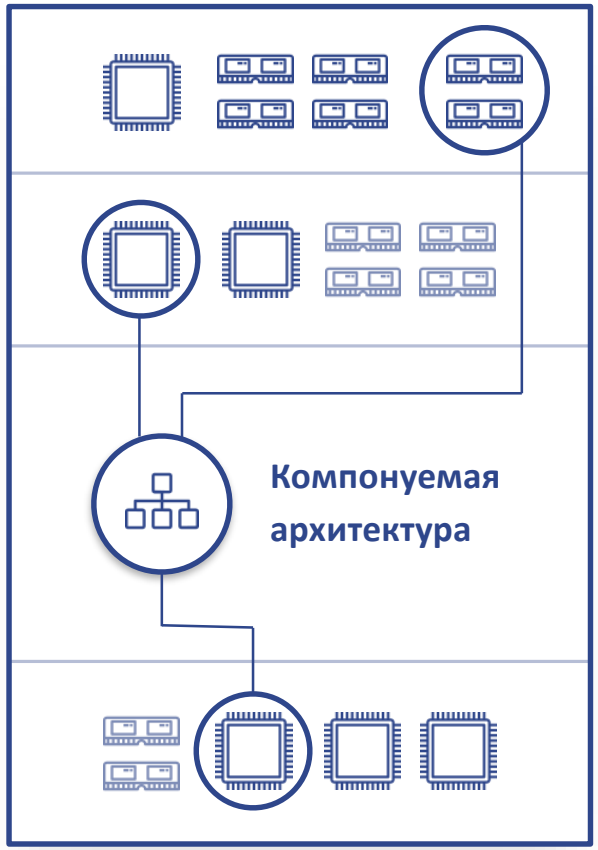
Суперкомпьютерные решения

Переход к архитектурам компонуемых дезагрегируемых сред

Архитектура уровня стойки (Rack Scale Architecture) → **Компонуемая Дезагрегированная Инфраструктура (CDI)**



- ⊕ **Гиперконвергерция**
- ⊕ **Современные технологии хранения и передачи**
- ⊕ **Программная оркестрация**
- ⊕ **Системы хранения «по запросу»**
- ⊖ Программная виртуализация



Дезагрегированная компонуемая инфраструктура

Вычислительные узлы

с поддержкой процессоров Intel, AMD, NVidia и «Эльбрус»

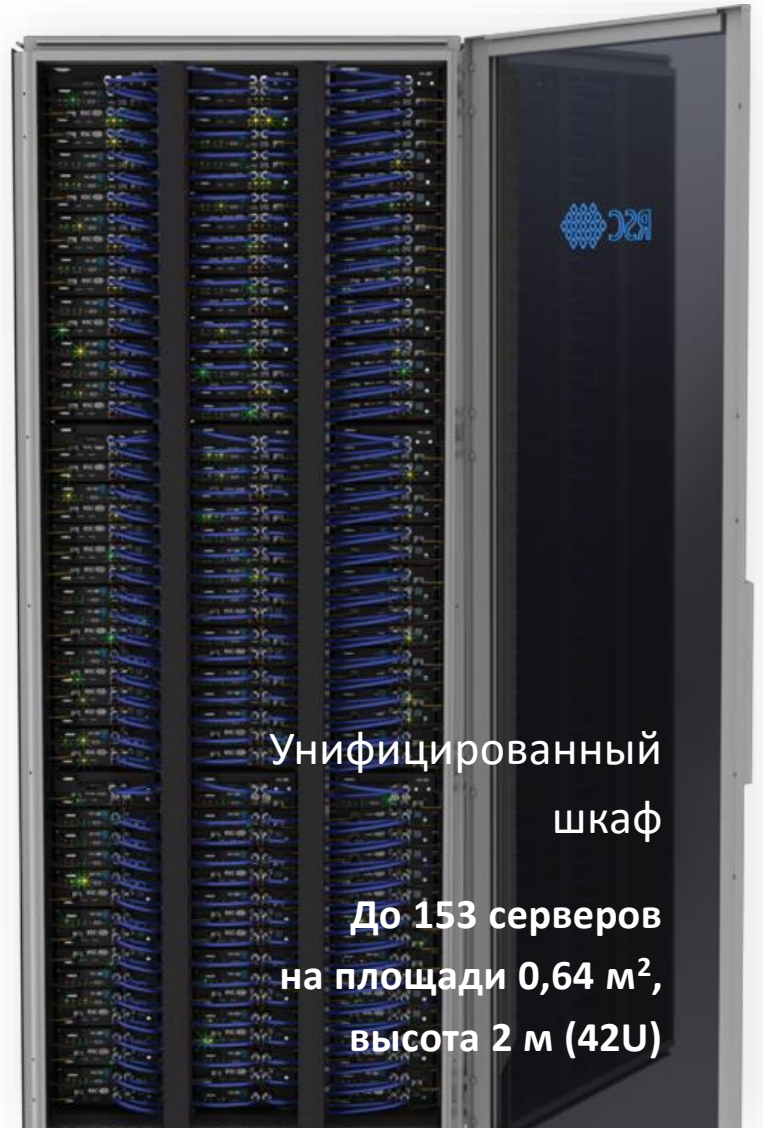
Гиперконвергентные узлы

с устройствами хранения NVMe

Модули избыточного питания

Програмный стек управления

RSC BasIS Software Platform



Унифицированный шкаф

До 153 серверов на площади 0,64 м², высота 2 м (42U)

Мировые рекорды

Компонуемая архитектура «РСК Торнадо»
на базе Intel Xeon Scalable 3-го поколения

967,45 ТФлопс**
Вычислительная плотность
на шкаф*

814,56 ТФлопс/м³
Производительность на объем

3,67 ТБ/с
Пропускная способность
распределенной системы
хранения

130 кВт
Энергетическая
плотность на шкаф

* шкаф 42U 80x80 см
** для решений на базе Intel® Xeon®

Мировые рекорды

РСК Торнадо AFS с Intel SSD P5316

41,3 ПБ

Рекордная емкость теплой системы хранения на шкаф

1 ТБ/с

Пропускная способность распределенной системы хранения

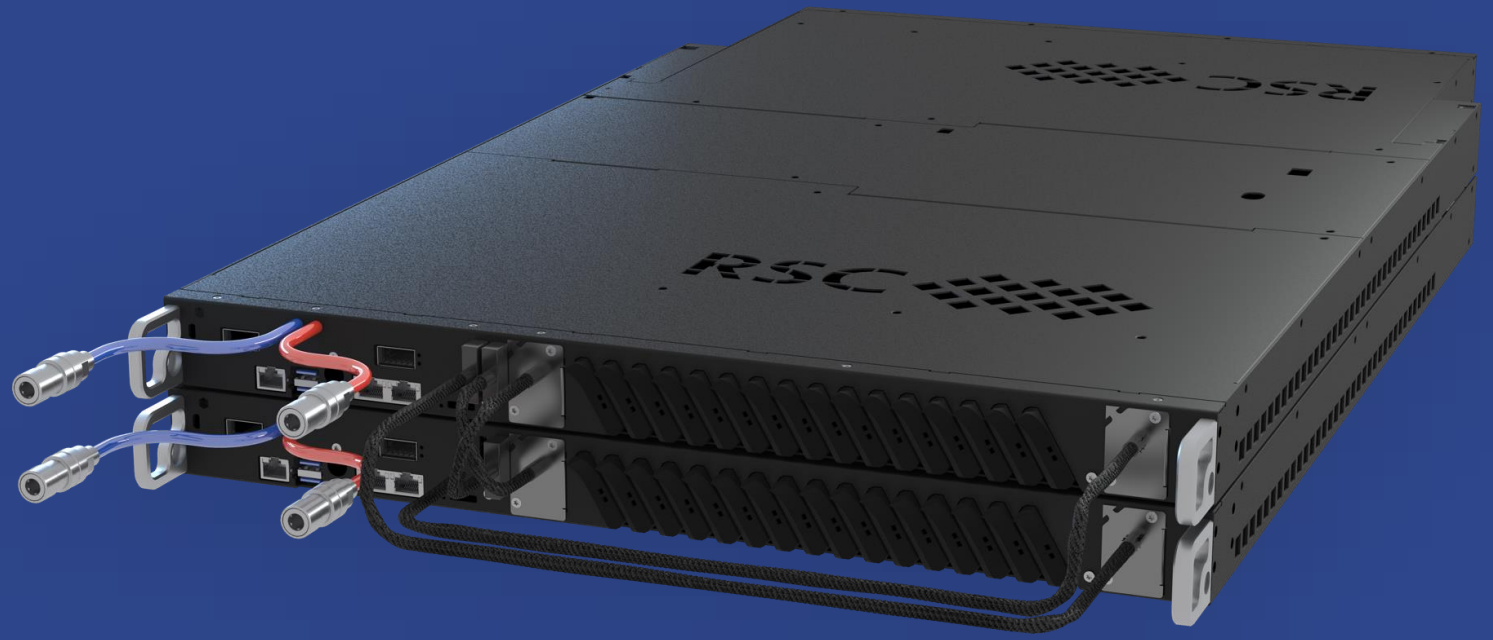
54 кВт

Энергетическая плотность на шкаф

шкаф 42U 60x100 см

СХД РСК Tornado AFS рекордно большого объема

E1.L Intel® Data Center SSDs
в форм-факторе EDSFF



Сверхвысокая емкость – 1 ПБ
в одном сервере (1U)
Надежное объединение 2 серверов
в СХД емкостью **2 ПБ (2U)**

Первое решение на 100% жидкостном охлаждении с высочайшей плотностью на базе 32x Intel EDSFF SSD, двух процессоров Intel® Xeon® Scalable и памяти Intel® Optane™ DC Persistent Memory

100% охлаждение «горячей водой» позволяет достичь рекордной энергоэффективности (PUE < 1,04)

Мировые рекорды

РСК Торнадо ИИ Решение с nVidia A100

1,895 ТФлопс (FP64)

Высочайшая вычислительная плотность на шкаф

105 кВт

Высочайшая энергетическая плотность на шкаф

104,83/209,66 ПОПс (INT8/INT4)

Лидирующие показатели по производительной плотности для ИИ

шкаф 42U 60x100 см

AI/ML/DL с РСК Торнадо ИИ

- 38,8 ТФлопс (FP64) в одном сервере
- 2,49/4,99 ПОПс (INT8/INT4) в одном сервере
- 100% охлаждение **«горячей водой» (PUE < 1,04)**

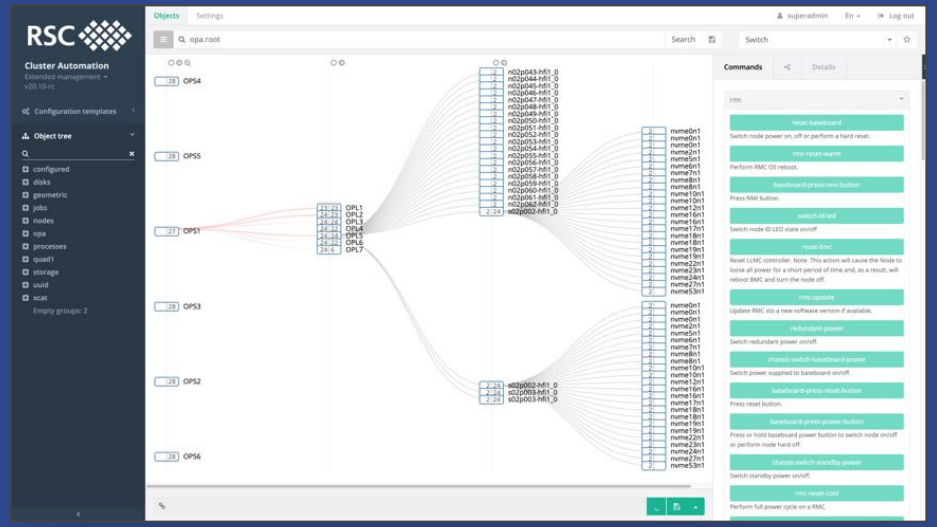
- 2x x86 CPUs
- 4x nVidia A100
- До 4 NVMe SSD PCIe Gen 4 E1.S (16 ТБ)
- До 4x 100-200Gb/s Omni-Path, Infiniband, Ethernet
- 2x резервных источника питания

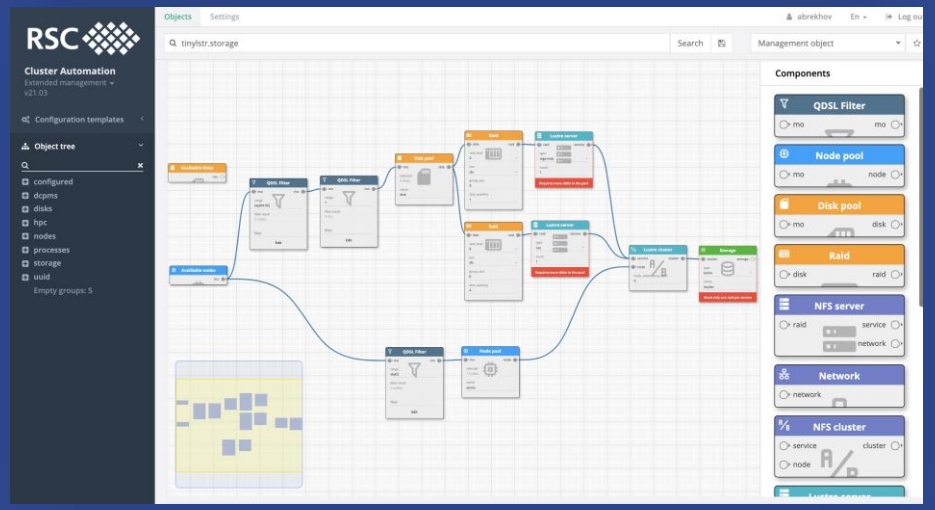
Программное обеспечение

Программный стек управления ЦОД

РСК Базис

«РСК БазИС» – это платформа управления для построения дезагрегированной компонуемой инфраструктуры как в крупных, так и в мобильных ЦОД





Реконфигурируемая программная платформа разделяет аппаратные компоненты и динамически компонует их гибким образом:

- построение проблемно-ориентированных конфигураций из объединенных вычислительных ресурсов, хранилищ и сетевых элементов
- поддержка быстрой переконфигурации
- адаптация к изменяющимся потребностям пользователей в ресурсах

РСК БазИС

Программная платформа управления вычислительными центрами

1

Суперкомпьютерный стек

Для конфигурации и управления высокопроизводительными кластерами

2

Система хранения

Для создания систем хранения из дезагрегированных ресурсов ЦОД под задачу пользователя

3

Управление ресурсами

Для учета использования любых ресурсов кластера и их аналитики

4

Гибкое облако

Для расширения вычислительных ресурсов в моменты пиковых нагрузок

На базе платформы оркестрации «РСК БазИС Автоматизация»

Система хранения «по запросу»

«РСК БазИС» позволяет создавать системы хранения данных «по запросу»:

Кластерная файловая система Lustre

Стандарт «де-факто» в мире суперкомпьютеров

Новая высокопроизводительная объектная система хранения DAOS

Разработана «с чистого листа» для поддержки высокоскоростных фабрик, устройств NVMe и Storage Class Memory

Предоставляет современные высокопроизводительные методы работы с данными:

HDF5	Apache Spark	MPI-IO	TensorFlow	NoSQL	S3	POSIX

Система хранения «по запросу» РСК БазИС

Группа компаний РСК получила престижную награду Russian DC Awards 2020 в номинации «Лучшее ИТ-решение для ЦОДа», победив с проектом «Высокопроизводительная система хранения для суперкомпьютера», реализованном в 2020 году в Объединенном институте ядерных исследований (ОИЯИ) в Дубне.

Архитектура РСК БазИС

РСК

POSIX

HDF5

Apache Spark

MPI-IO

TensorFlow

NoSQL

Другие файловые системы

lustre

Блочное хранилище

Система хранения «РСК БазИС»

Платформа «РСК БазИС Автоматизация»

Оборудование

Ключевые проекты

Наука

- Объединенный институт ядерных исследований (ОИЯИ)
- Российская академия наук (МСЦ РАН)
- Физико-технический институт имени Иоффе РАН
- Сибирский суперкомпьютерный центр СО РАН
- Институт океанологии имени Ширшова
- Институт физики атмосферы им. А. М. Обухова РАН
- Гидрометцентр России

Образование

- Санкт-Петербургский политехнический университет Петра Великого (СПбПУ)
- Московский государственный университет имени Ломоносова (МГУ)
- Нижегородский государственный университет (ННГУ)
- Южно-Уральский государственный университет (ЮУрГУ)
- Московский физико-технический институт (МФТИ)

Отрасли экономики

- VK
- Авиастроение
- Автомобилестроение
- Энергетика
- Компьютерная графика
- Нефте- и газодобыча

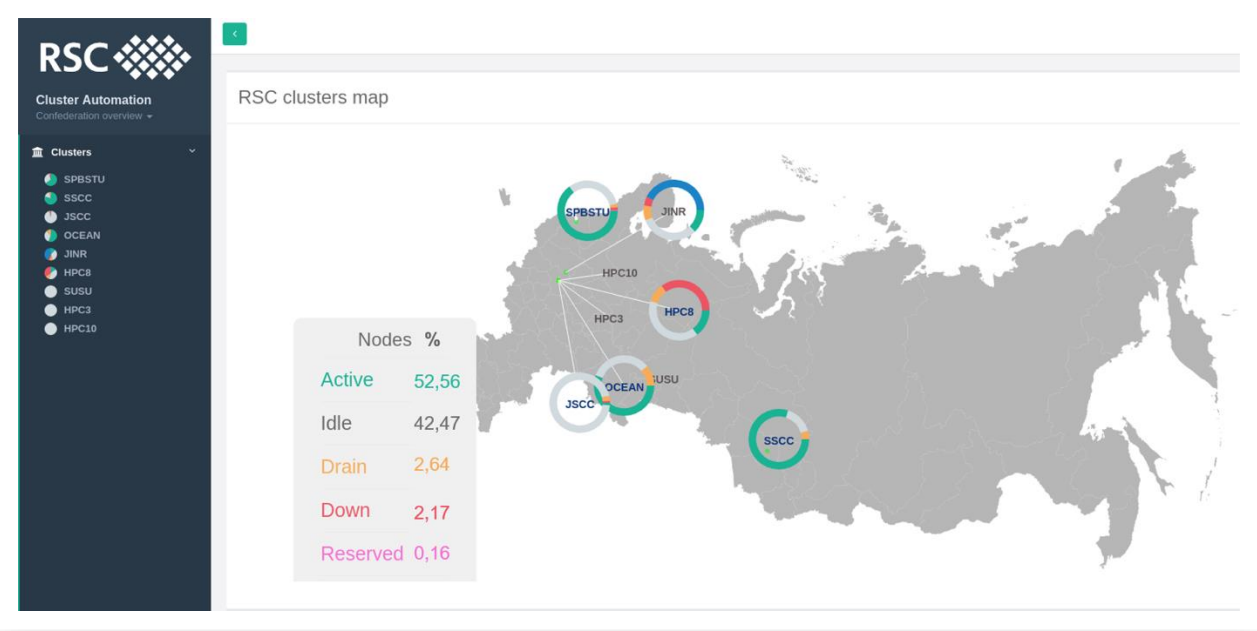
и другие

Примеры проектов

Объединенный институт ядерных исследований (ОИЯИ), Дубна

Санкт-Петербургский политехнический университет Петра Великого (СПбПУ)

Система управления и мониторинга распределенными ЦКП



- Управление и мониторинг центрами коллективного пользования (ЦКП), распределенными по территории России
- Единая платформа управления системой для HPC и облака, дополненная средствами развертывания, управления и поддержки, включая поддержку территориально распределенных систем
- Основные задачи ЦКП – предоставление вычислительных ресурсов и временного хранения данных
- Использована система управления жизненным циклом центра обработки данных

Спасибо!

РСК

rscgroup.ru
hq@rsc-tech.ru