РСК

Группа компаний

Разработка и реализация российских суперкомпьютеров

А.А. Московский, генеральный директор

ИТ-Школа ЛИТ ОИЯИ, 17.10.2023

Высокопроизводительные системы с 2009 года

1. Опыт разработки и реализации платформ
2. Перспективы развития
3. Решения РСК

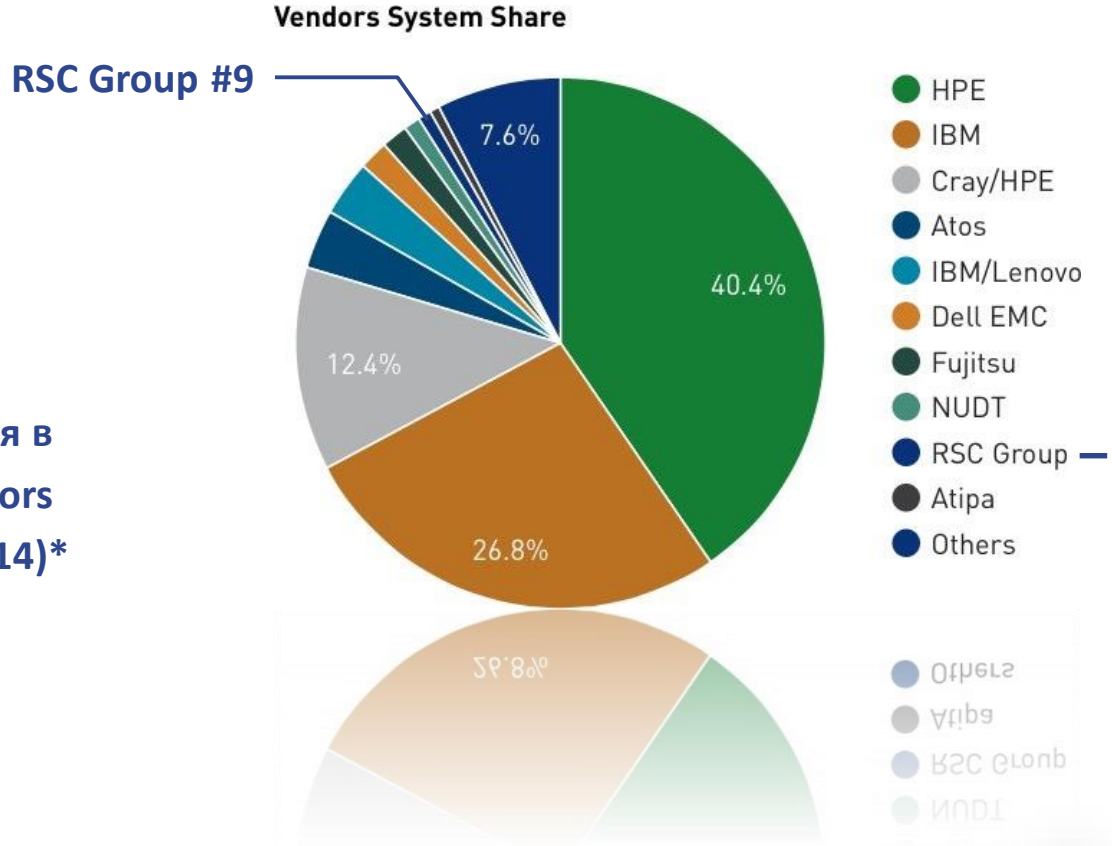
О группе компаний РСК

Ведущий российский разработчик и интегратор инновационных суперкомпьютерных решений

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

TOP500

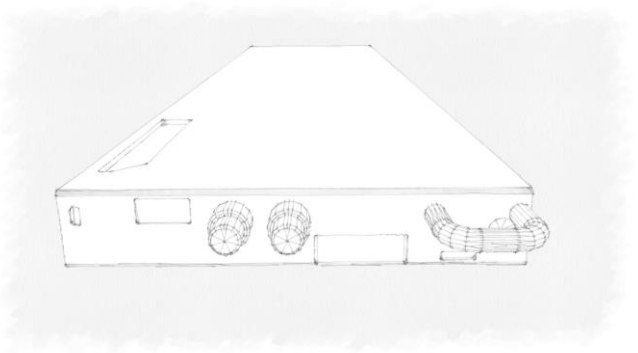
Единственная российская компания в мировом рейтинге Top10 HPC Vendors System Share by Top500 (Ноябрь 2014)*

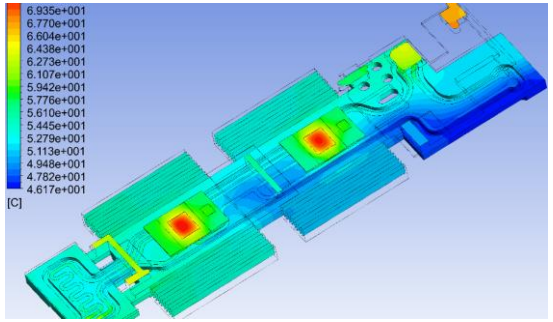


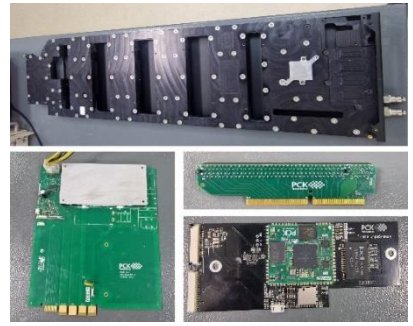
*** Топ 10 поставщиков по объему рынка**
https://www.top500.org/statistics/list/

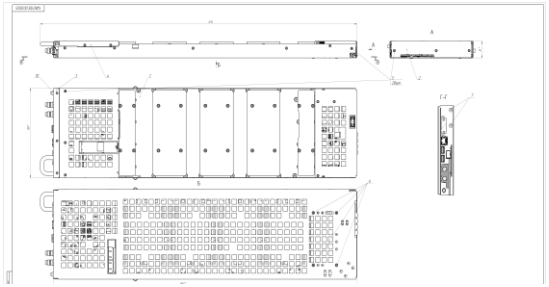
Полный цикл разработки РСК

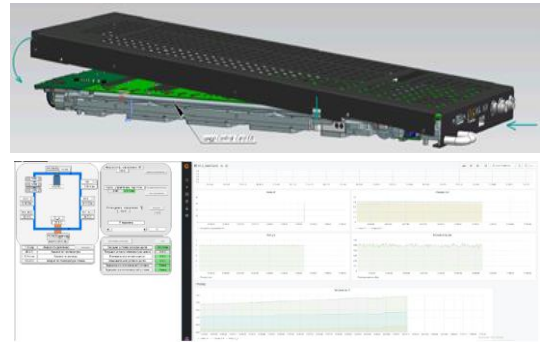
Дизайн-концепт → CFD симуляции → Производство тестового образца











Серийный продукт ← Предпроизводственные модификации ← Сборка и тестирование образца

Некоторые проекты 2010-2012 гг

«РСК Торнадо ЮУрГУ»

- **Создан в 2009, модернизирован в 2012,2013**
  Изначально узлы:
  - 2x Intel Xeon X5680 @ 3,33 ГГц (130 Вт TDP)
  - 24/48 ГБ DDR3R-1333 ОЗУ
  - QDR Infiniband, Fat-tree. Lustre FS
- В 2012 и 2013 **добавлено 284** узла:
  - 2x процессора Intel Xeon X5680 @ 3,33 ГГц
  - 1x сопроцессор Intel Xeon Phi SE10X (300 Вт)
  - СХД Panasas (возможность монтирования на MIC)
- **Рейтинги**
  - #128 в Top500 (11/13)
  - #50 в Green500 (06/13), #2 в России
- **Размеры**
  - 9 вычислит. стоек, 3 шкафа инфраструктуры
  - 50 кв. м./350 кВт
  - 473 TFLOPS Rpeak,
  - 288 TFLOPS Rmax (HPL)
  - 995 MFLOPS/Вт

Проект iScalare в Факультете Радиотехники и Кибернетики МФТИ

- Реализация проекта - 2011-2013 гг, лаборатория Интел-МФТИ под руководством В. Пентковского
- 224 узла 2xIntel Xeon E5-2690 83 ТФЛОПС Rpeak, 70 ТФЛОПС Rmax
- 10 место в рейтинге Top50, апрель 2013

- Специально для решения задач быстрого развертывания IT инфраструктуры
- Отсутствуют затраты на строительство ЦОД
- Не требует специальной подготовки помещения
- Размещение до 128 серверов и системы хранения на 1.6 м2
- Внешний модуль охлаждения минимального размера
- Низкие затраты на эксплуатацию комплекса

Жидкостное охлаждение в Top 20

RSC

Системы с воздушным охлаждением в меньшинстве

Top500 Rank	System	Cooling technology
1	Frontier	Direct cold water cooling
2	Fugaku	Direct cold water cooling
3	LUMI	Direct cold water cooling
4	Leonardo	Direct warm water cooling
5	Summit	Direct cold water cooling
6	Sierra	Direct cold water cooling
7	Sunway TaihuLight	Airflow cooling
8	Perlmutter	Direct cold water cooling
9	Selene	Airflow cooling
10	Tianhe-2A	Airflow cooling

Top500 Rank	System	Cooling technology
11	Explorer-WUS3	Airflow cooling
12	Adastra	Direct cold water cooling
13	JUWELS Booster Module	Direct warm water cooling
14	Pre-Eos 128 Node DGX SuperPOD	Direct cold water cooling
15	HPC5	Airflow cooling
16	Voyager-EUS2	Airflow cooling
17	Setonix – GPU	Direct cold water cooling
18	Discovery 5	Direct cold water cooling
19	Polaris	Airflow cooling
20	SSC-21	Airflow cooling

Модернизация суперкомпьютерного комплекса ЛИТ ОИЯИ: 2022 год

Суперкомпьютер «Говорун» – создание

Гиперконвергентная архитектура

4 типа узлов для масштабирования всех типов нагрузок:

- С массивным параллелизмом
- Стандартные big-core
- С большой памятью
- Хранения

Энергоэффективность

На охлаждение решения расходуется менее 3% от общих энергозатрат системы (PUE = 1,0277)

Суперкомпьютер «Говорун» в (ОИЯИ) Дубна – 2018 год

- **536 ТФЛОП/с** пиковой производительности - **#18 и #45** в Top50
- Охлаждение «горячей» водой **всего оборудования**
- **Самый энергоэффективный** Суперкомпьютер в России (**PUE = 1,027**)
- Система управления РСК “BasIS”

2 сегмента системы

Вычислительные узлы сегмента Skylake

Пиковая производительность – **138** ТФлоп/с
Intel® Xeon® Gold 6154 processors (18 ядер)
Intel® Server Board S2600BP
RAM – 192 GB DDR4 2933 Ghz
Intel® Omni-Path 100 Gbit/s
48-port Intel® Omni-Path Edge Switch 100 Series w 100% liquid cooling

Узлы с процессорами Intel® Xeon Phi™ :

Пиковая производительность – **72,576** ТФЛОПС
Intel® Xeon Phi™ 7190 CPUs (72 cores)
Intel® Server Board S7200AP
Intel® SSD DC S3520 (SATA, M.2)
RAM – 96 GB DDR4 2400 Ghz
Intel® Omni-Path 100 Гбит/с
48-port Intel® Omni-Path Edge Switch 100 Series 100% liquid cooling

Суперкомпьютер «Говорун» в (ОИЯИ) Дубна – 2019 год

- **860 ТФЛОП/с** пиковой производительности - **#10** в Top50
- Программно-определяемая архитектура
- **Теоретическая производительность системы хранения >300 ГБ/с**
- Масштабируемая **Система Хранения-по-Требованию**
- Многоуровневая система хранения данных
- Охлаждение «горячей» водой **всего оборудования**
- **Самый энергоэффективный с**уперкомпьютер в России (**PUE = 1,027**)
- Система управления «РСК БазИС»

Гиперконвергентная система

Вычислительные узлы

Пиковая производительность – **463**TFLOPS
Intel® Xeon® Platinum 8268 processors (24 ядра)
Intel® Server Board S2600BP
Intel® SSD DC S4510 (SATA, M.2),
2 x Intel ® SSD DC P4511 (NVMe, M.2) 2TB
RAM – 192 GB DDR4 2933 Ghz
Intel® Omni-Path 100 Gbit/s
48-port Intel® Omni-Path Edge Switch 100 Series w 100% liquid cooling

Узлы хранения данных

18 узлов с 12 слотами под NVMe SSD
4 узлов с Optane™, 3,4TB IMDT
12 узлов OSS, NVMe SSD – **256TB**
2 узла MDS 12 **Optane™ 375GB**
Основная файловая система Lustre
Система Хранения-по-Требованию с ПО **РСК БазИС**

Узлы с процессорами Intel® Xeon Phi™ :

Пиковая производительность – **72,576** ТФЛОПС
Intel® Xeon Phi™ 7190 CPUs (72 cores)
Intel® Server Board S7200AP
Intel® SSD DC S3520 (SATA, M.2)
RAM – 96 GB DDR4 2400 Ghz
Intel® Omni-Path 100 Гбит/с
48-port Intel® Omni-Path Edge Switch 100 Series 100% liquid cooling

Суперкомпьютер «Говорун» – модернизация 3 этап

2021-2022 год

СХД RSC Tornado AFS
1 ПБ на узел 1U

Суммарно 8 ПБ

Суперкомпьютер «Говорун» – модернизация 4 этап

2022 год

32 новых вычислительных узла

Прирост производительности

202 Тфлопс (23,5%)

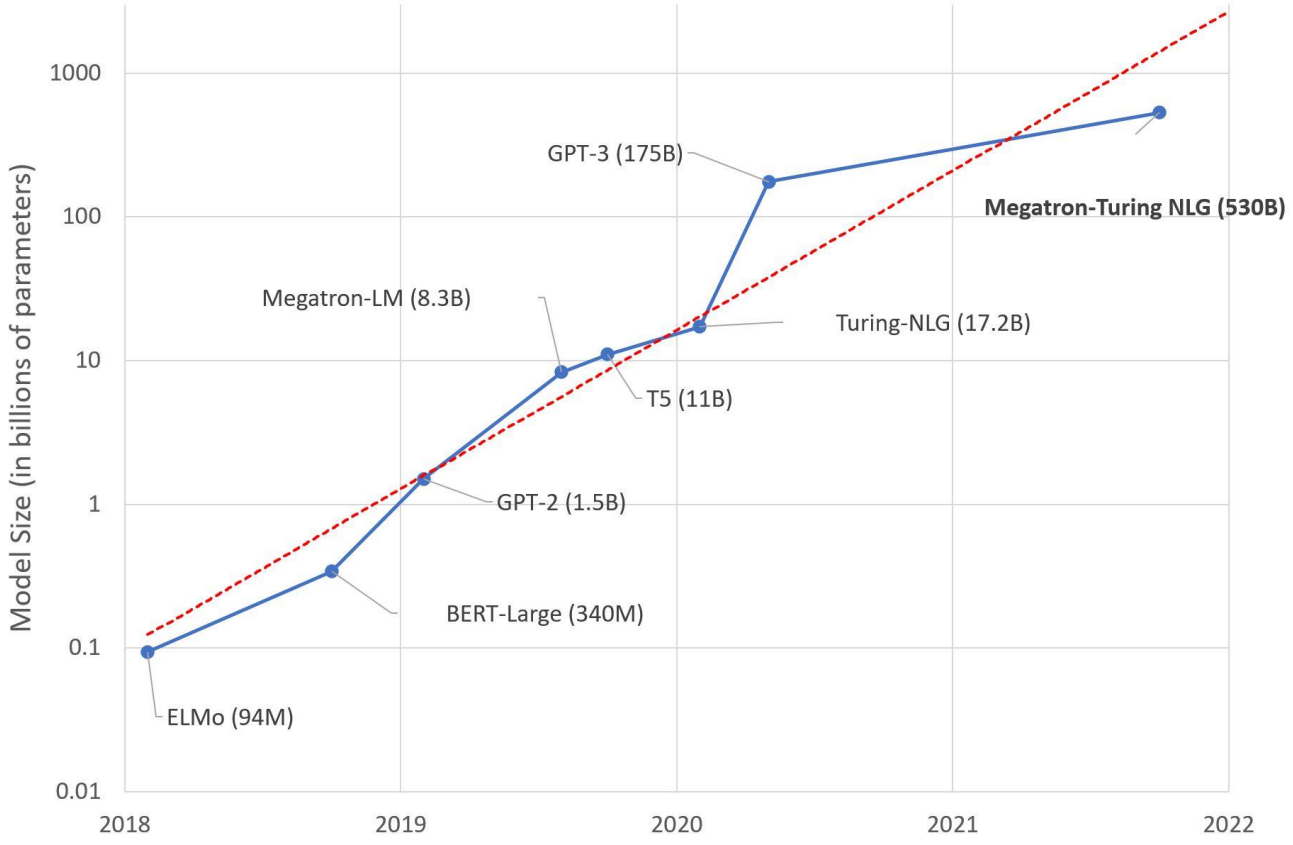
Суммарная производительность системы

1,1 ПФлопс

Тренды микроэлектроники и высокопроизводительных систем

Большие модели ИИ и их быстрый рост

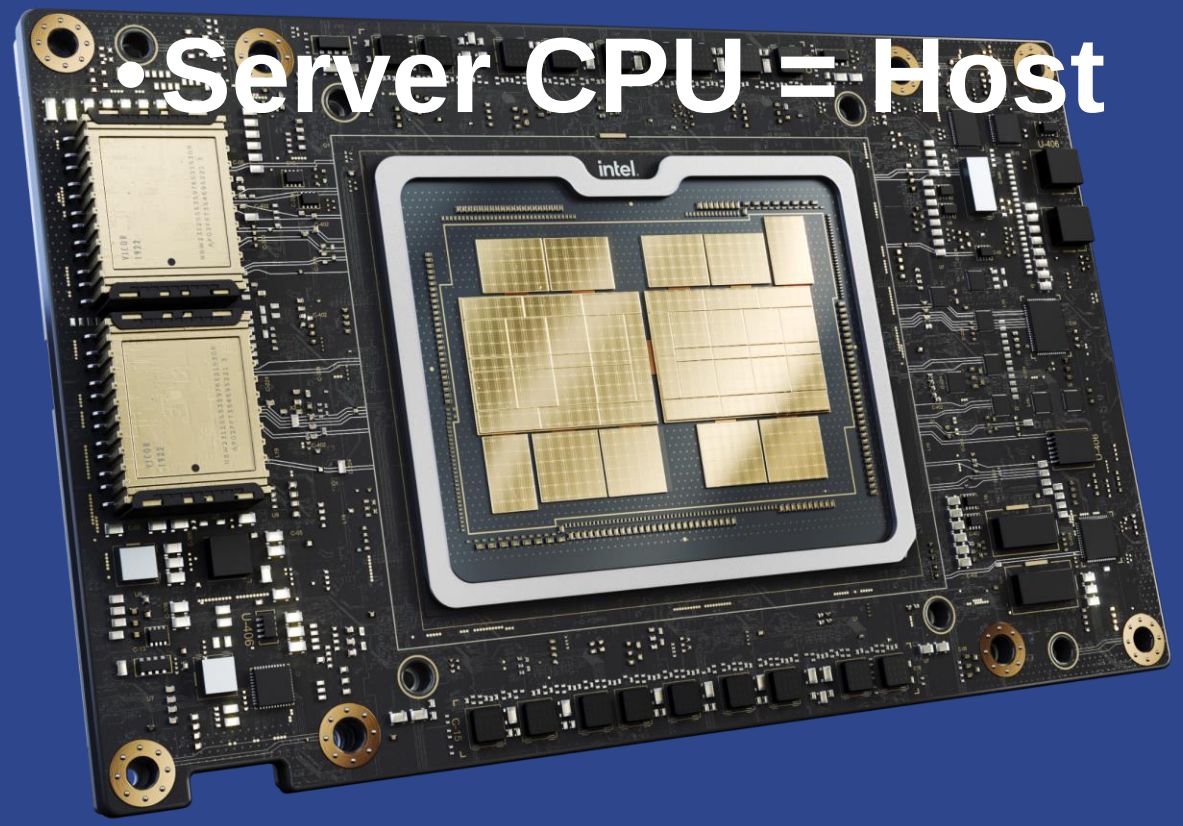
- Сложность обучения модели-транфсормера [1]:
  - $C \approx 6ND$
  - *N – число параметров, D- число токенов*
- Для GPT-3 – $3{,}14*10^{23}$ операций
- Размер набора для обучения ~ терабайты
- Трансформеры – не только NLP (см например GenSLM [2])



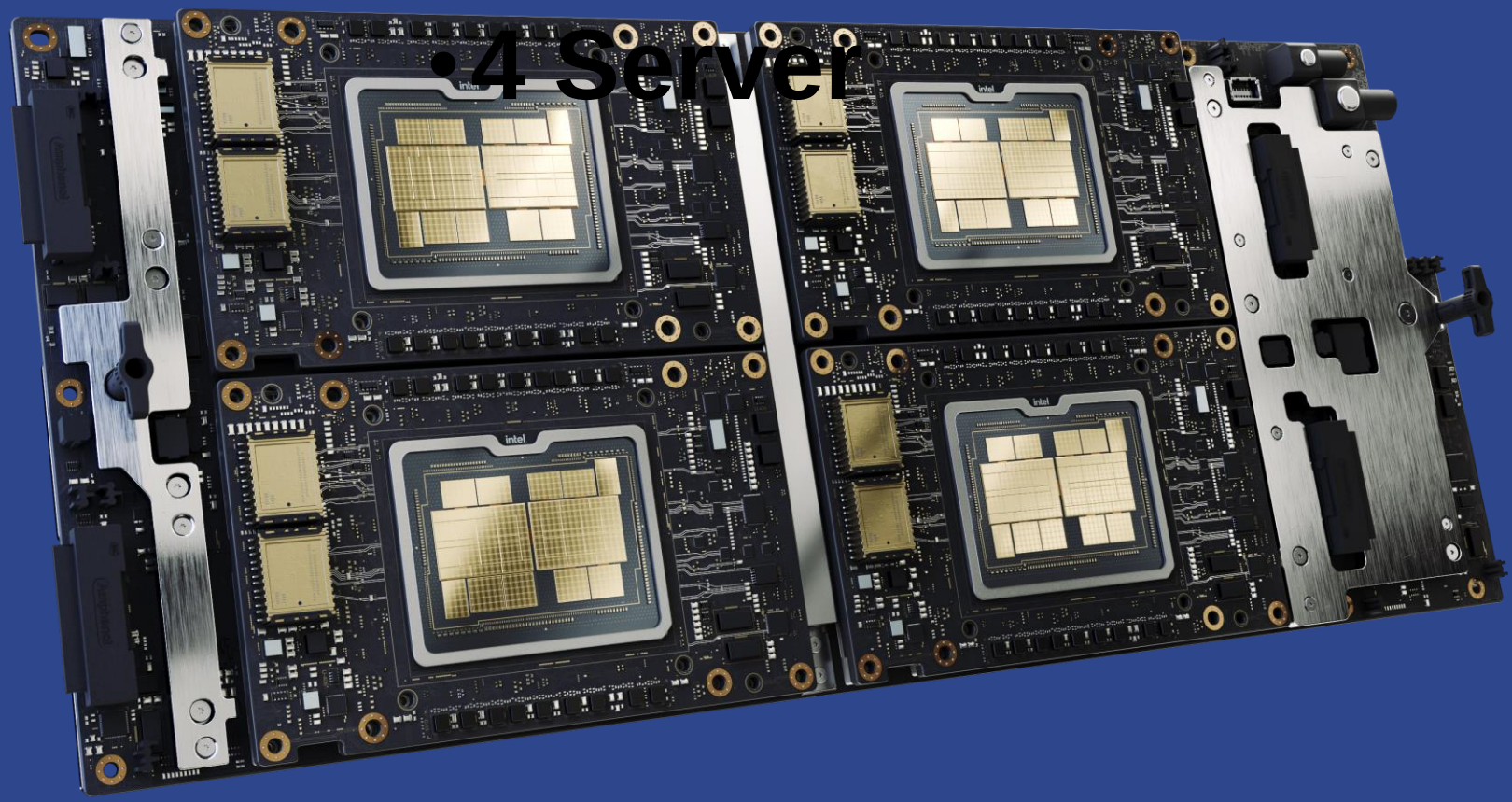
[1] – J. Caplan et. al “Scaling Laws for Neural Language Models” https://doi.org/10.48550/arXiv.2001.08361

[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

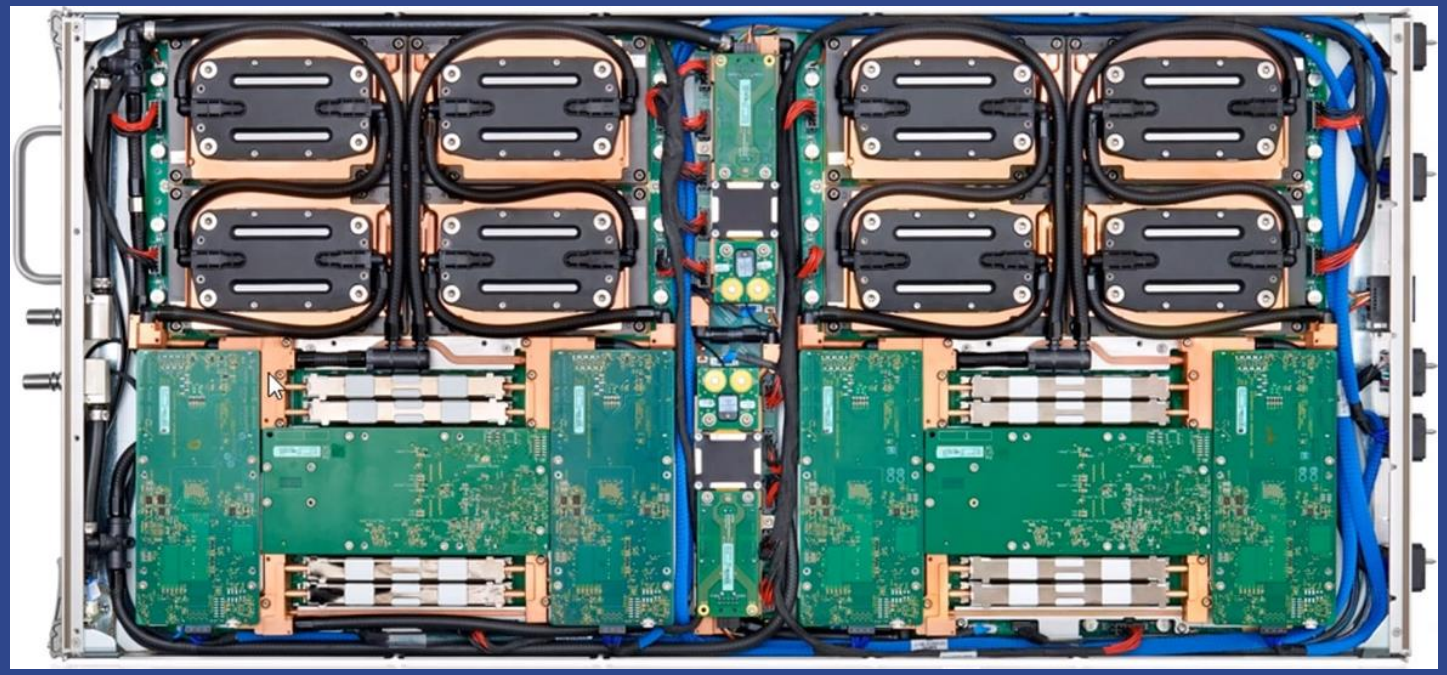
- Развитие специализированных процессоров.
- Развитие быстрого интерконнекта (CXL) возможность создания сервера по запросу (подключение ускорителей, вычислителей, памяти).



- Развитие специализированных процессоров.
- Развитие быстрого интерконнекта (CXL) возможность создания сервера по запросу (подключение ускорителей, вычислителей, памяти).



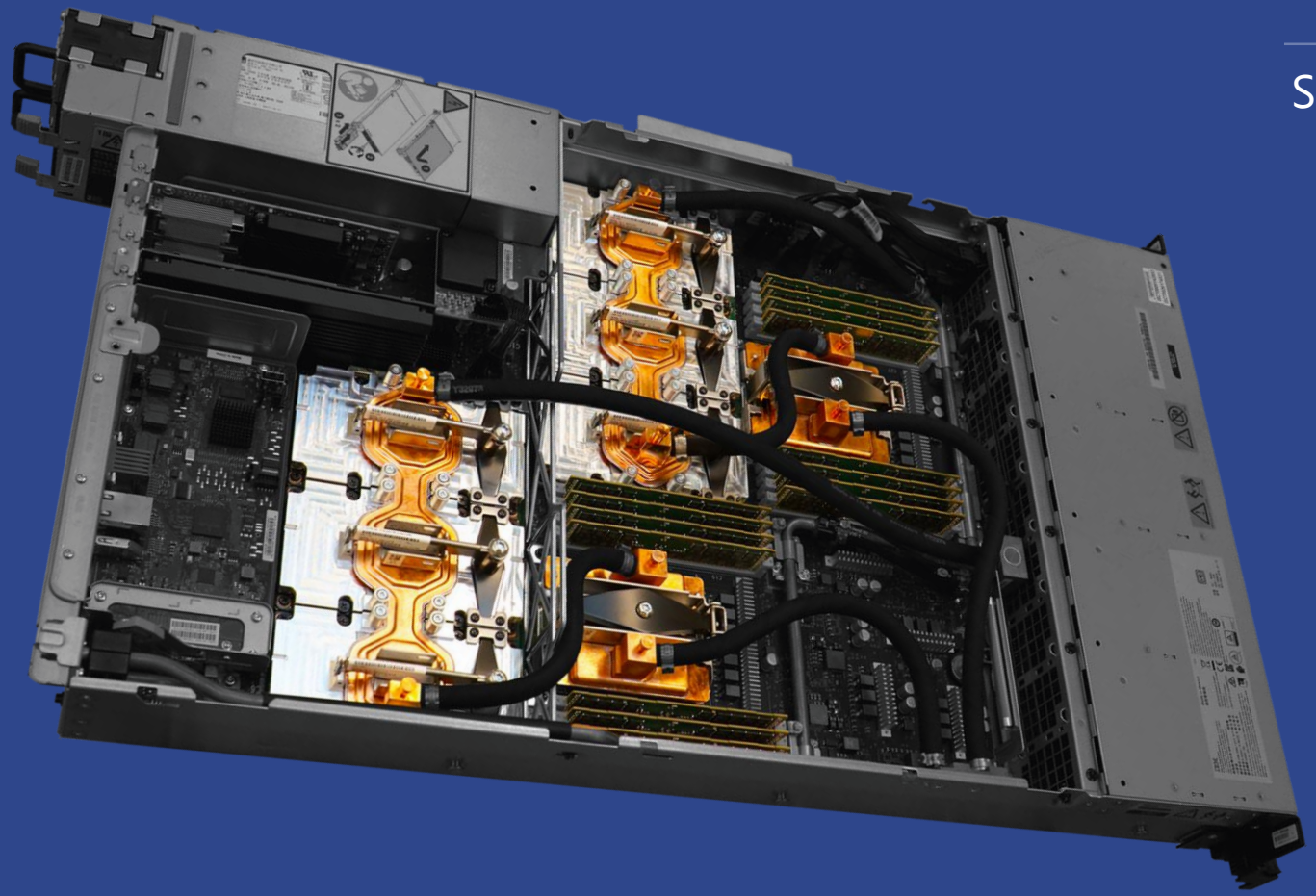
- Продолжающаяся интеграция электроники – повышение плотности энергии (системы больше потребляют и больше выделяют). Существующие привычные форматы размещения оборудования перестают работать.
- 19 дюймов шкаф – 20 кВт. Воздушное охлаждение (20 серверов в шкаф).
- Различные OCP платформы - 15кВт, можно увеличить до 30кВт (4 AI системы в шкаф).
- Существующие подходы исчерпали себя.
- Крупные компании быстрее отрасли формируют свои платформы (Google TPU, Huawei, Yandex).



FRONTIER - HPE CRAY EX235A

6.1 КВТ

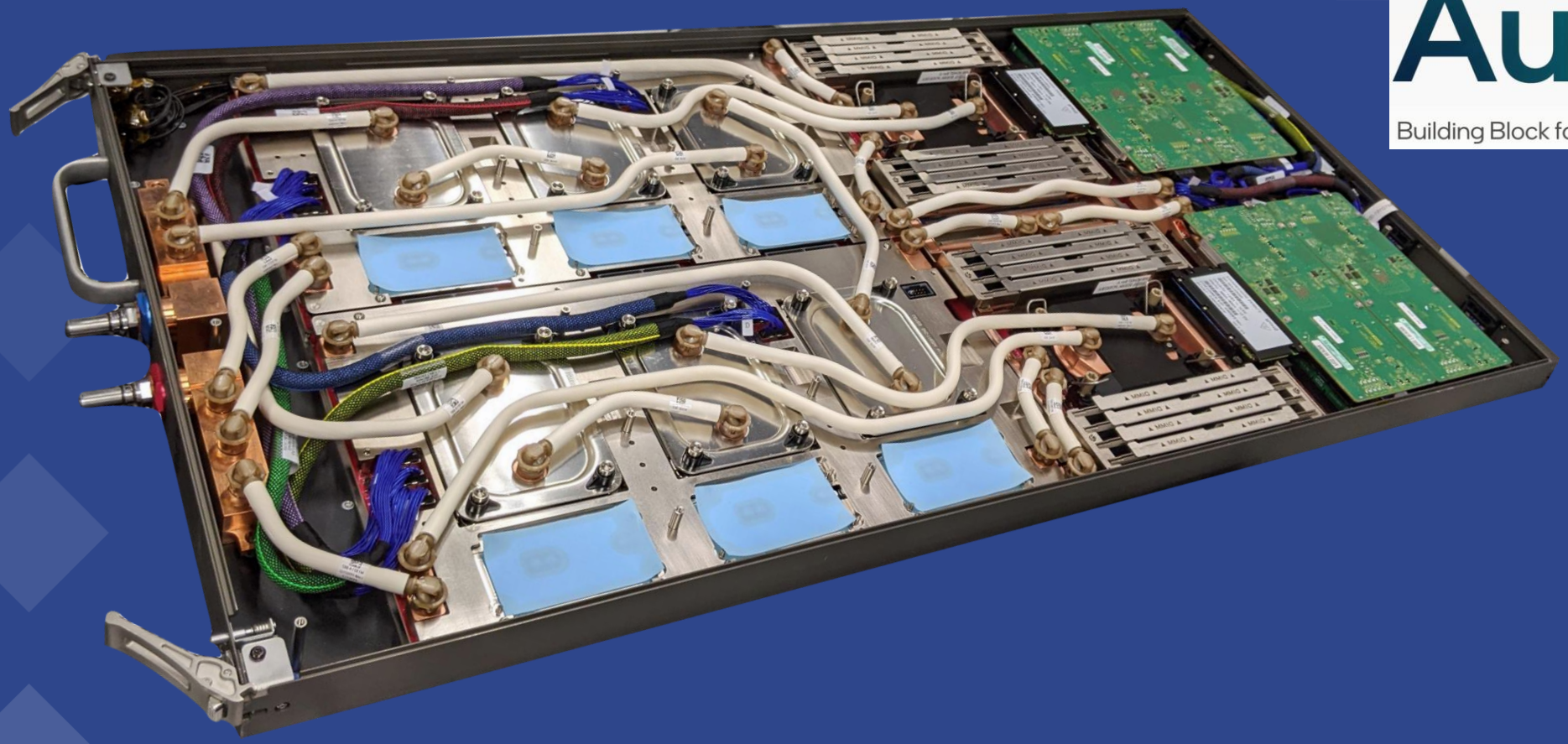
- Продолжающаяся интеграция электроники – повышение плотности энергии (системы больше потребляют и больше выделяют). Существующие привычные форматы размещения оборудования перестают работать.



SUMMIT - IBM POWER SYSTEM AC922

4.8 КВТ

PCK

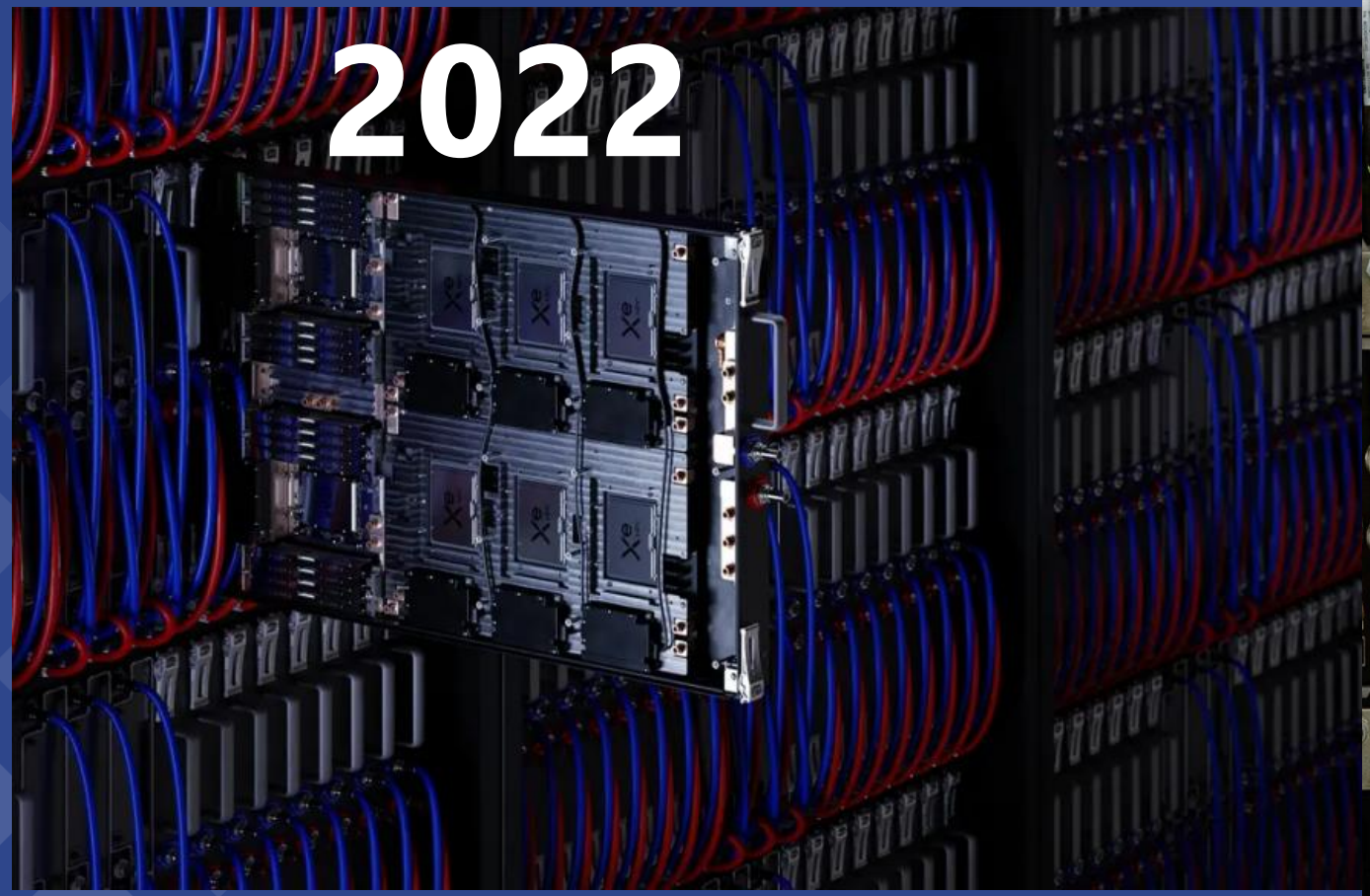


Aurora Blade

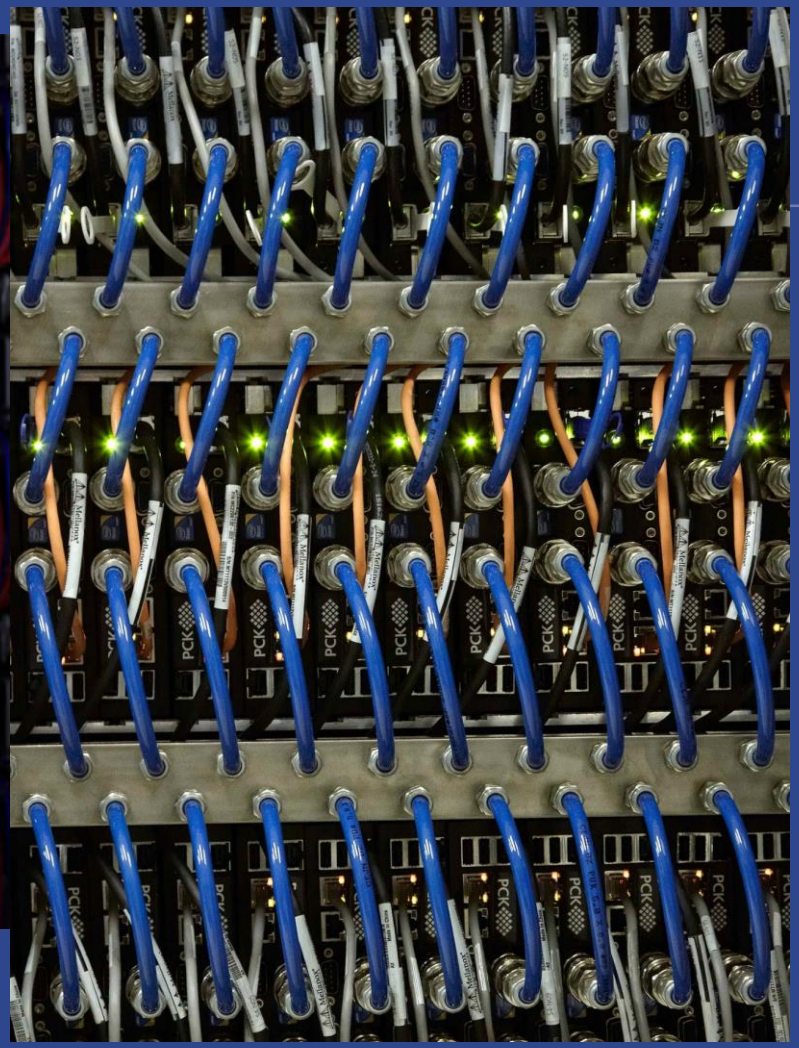
Building Block for the ExaScale Supercomputer

5.3 KBT

Intel launches Data Center GPU: Ponte Vecchio to be equipped on Argonne Labs

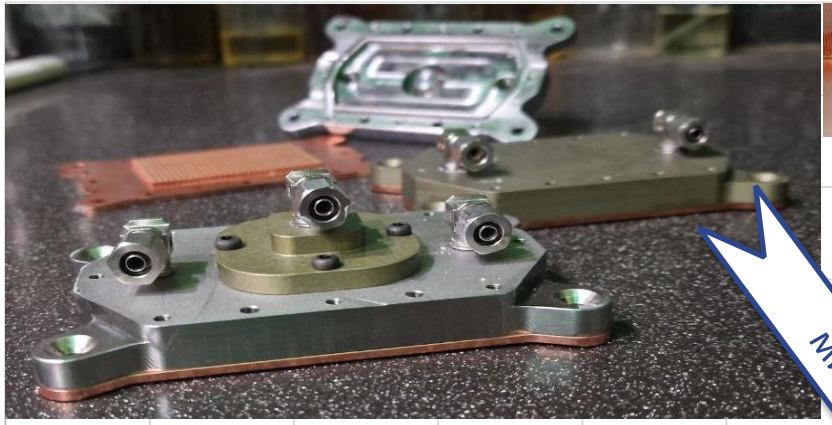


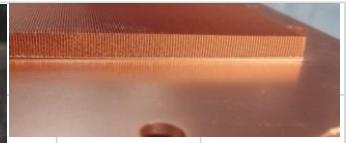
2013 г.

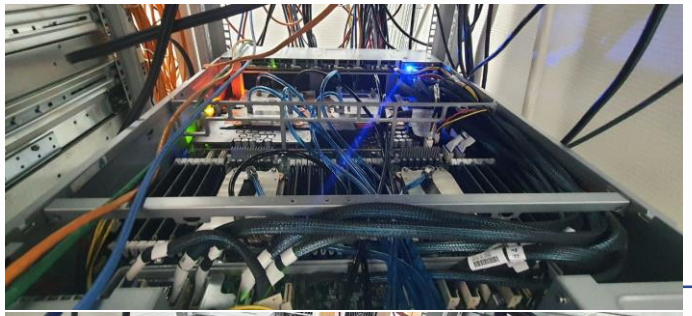


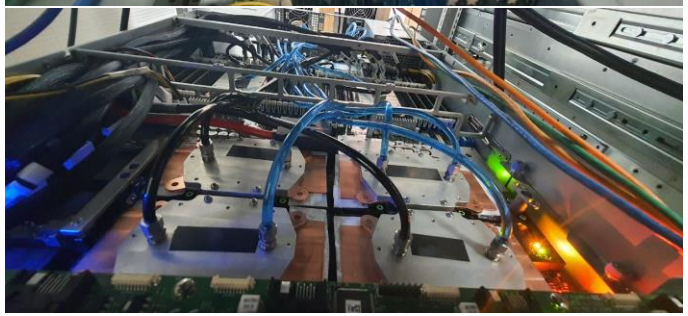
РСК Торнадо
МСЦ РАН
г. Москва

Системы охлаждения с фазовым переходом









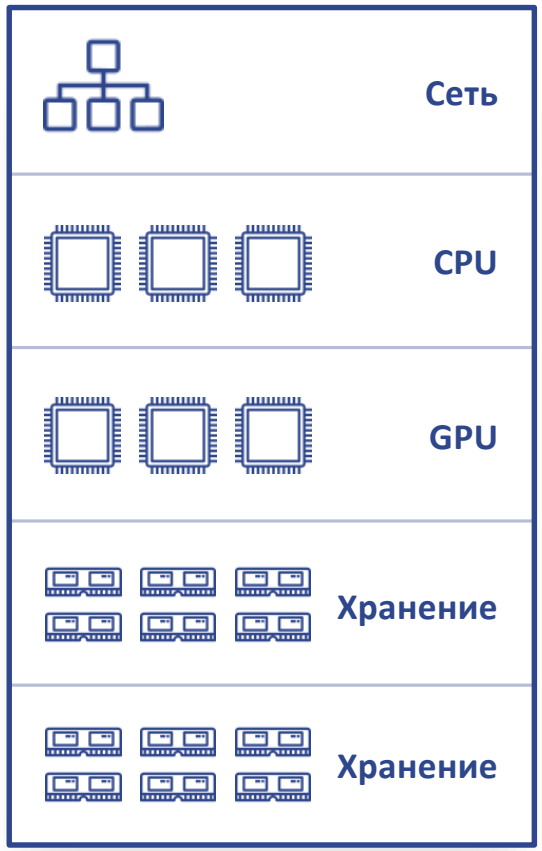
Микроканальная система охлаждения с фазовым переходом позволяет обеспечить максимальную производительность всех существующих продуктов на рынке.

Технология охлаждения спреем с фазовым переходом является новой разработкой для потенциальных чипов, которые появятся в ближайшем будущем. Она позволяет отводить тепловые потоки в 2МВт/м² при температуре корпуса в 85°C, что удовлетворяет требованиям современных чипов.

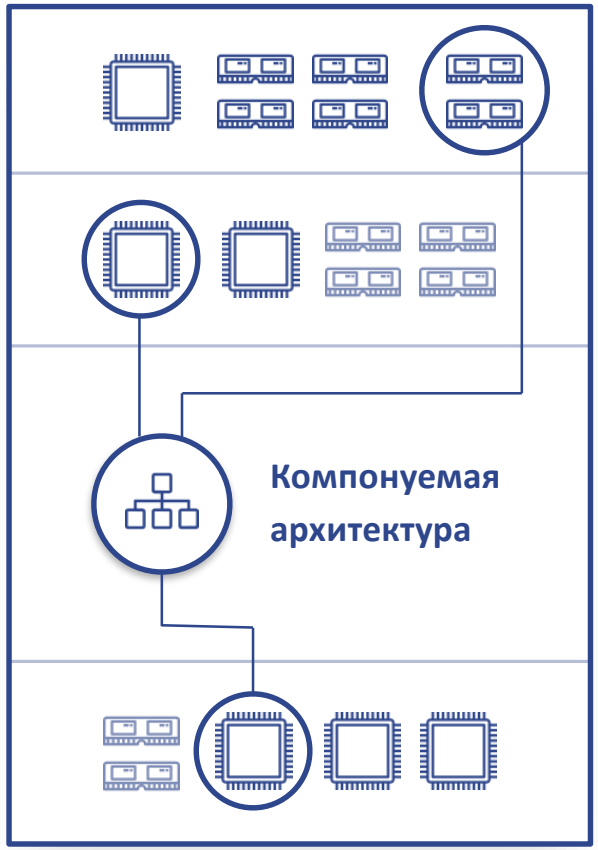
Суперкомпьютерные решения

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 и «Эльбрус»

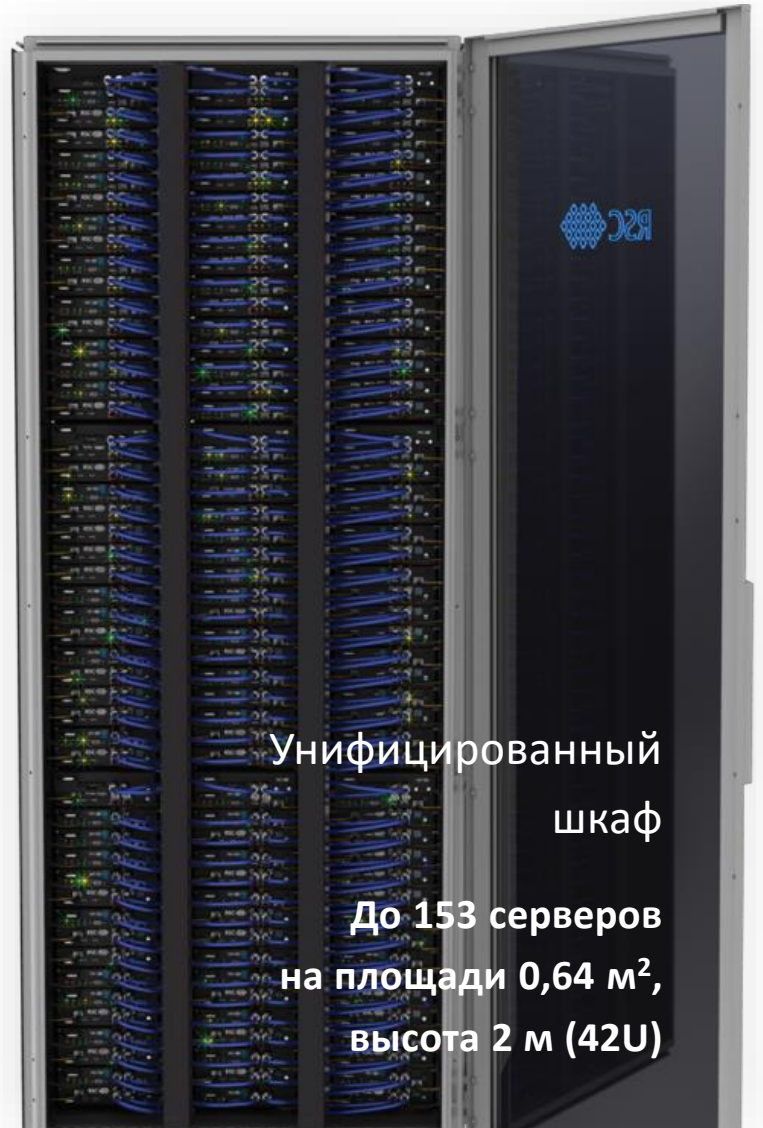
Гиперконвергентные узлы

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)

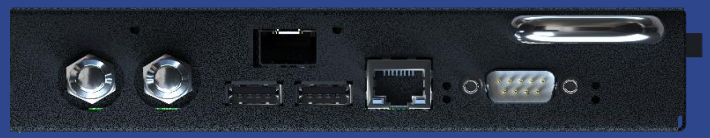
Вычислительный узел «РСК Торнадо» на базе «Эльбрус-8СВ»

Разработан в рамках **архитектуры уровня вычислительной стойки** – подходу к проектированию не единичного сервера, а всего ЦОД для удаленного управления без обслуживающего персонала

Содержит интегрированный модуль управления – для интеграции с программной платформой удаленного управления ЦОД "РСК БазИС"

Работает в рамках интероперабельной платформы, поддерживающей как сервера на базе Intel, так и Эльбрус

Высочайшая энергоэффективность благодаря 100% охлаждению ‘горячей водой’



- 4 процессора «Эльбрус-8СВ» (8 ядер, 1500 МГц),
- 2,3 TFLOPs SP / 1,1 Tflops DP
- Оперативная память: до 256 Гбайт DDR4
- До 3 дисков mSATA
- Установка высокоскоростных сетевых адаптеров

Модуль управления выполняет расширенный мониторинг, удаленное управление, следит за предотвращением возможных аварийных ситуаций

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)**

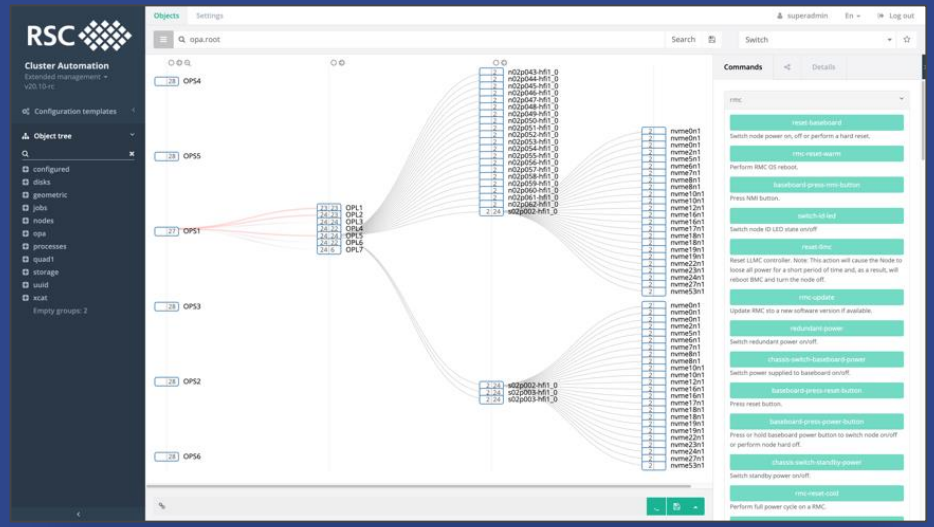
- 2x x86 CPUs
- 4x nVidia A100
- До 4 NVMe SSDs PCIe Gen 4 E1.S (16TB)
- До 4x 100-200Gb/s Omni-Path, Infiniband, Ethernet
- 2x резерных источника питания

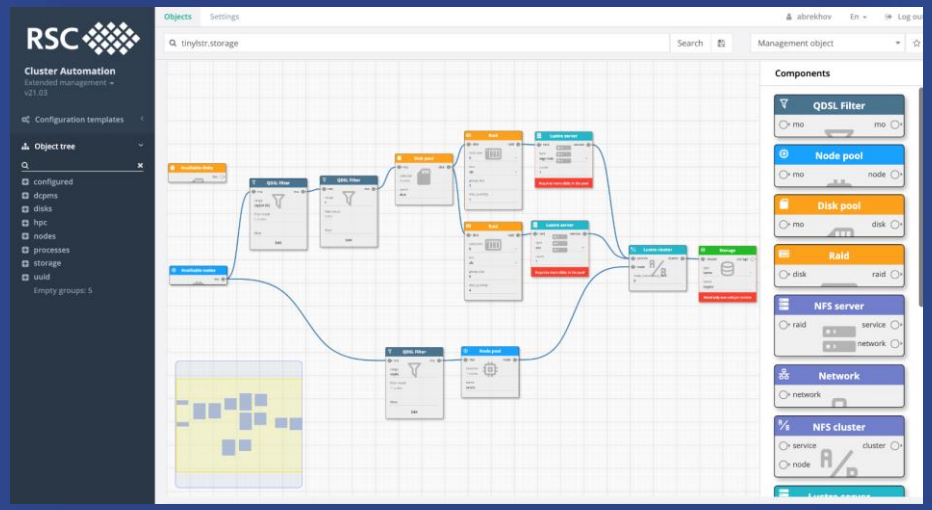
Программное обеспечение

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

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:

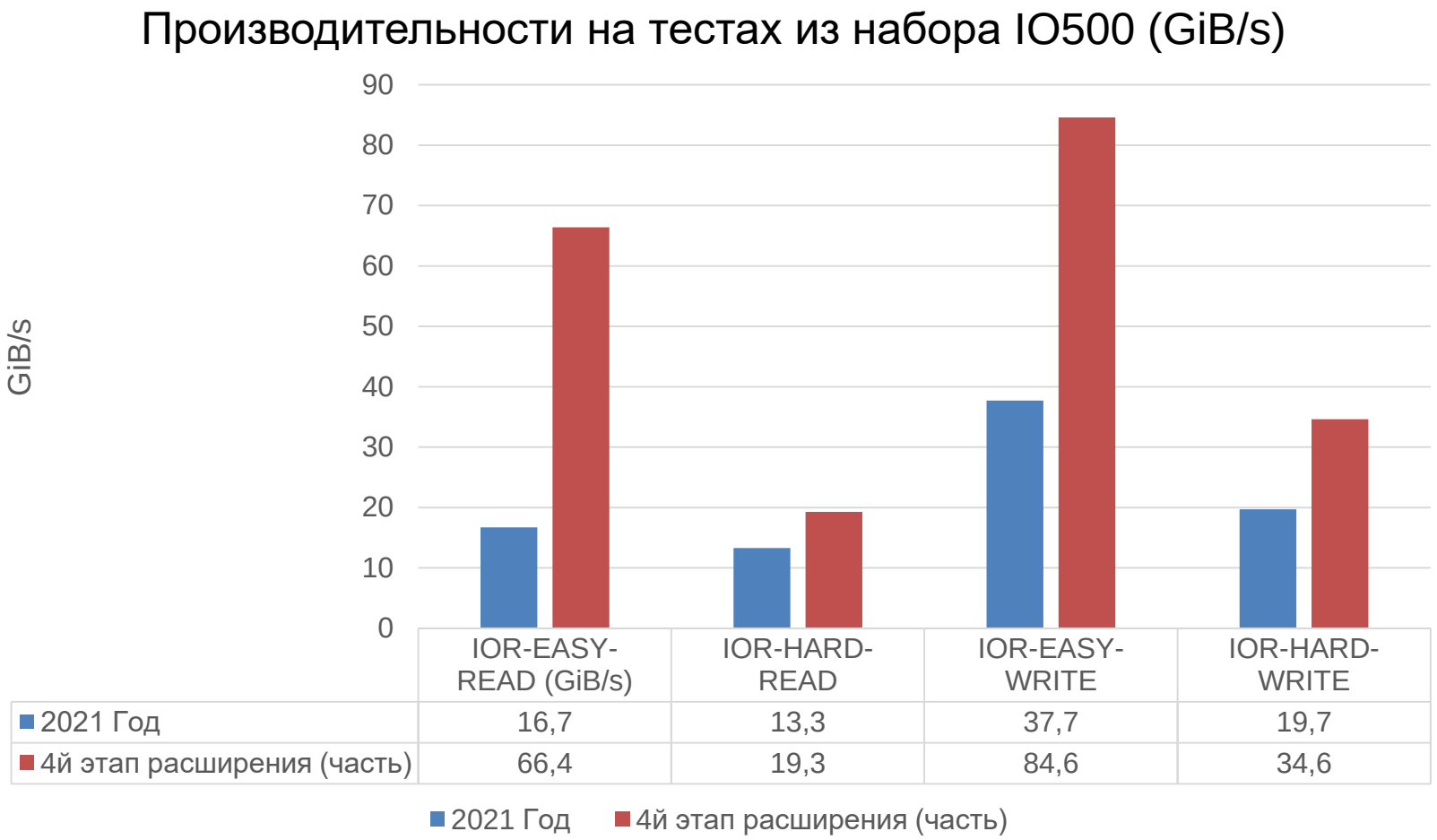
HDF5	Apache Spark	MPI-IO	TensorFlow	NoSQL	S3	POSIX

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.

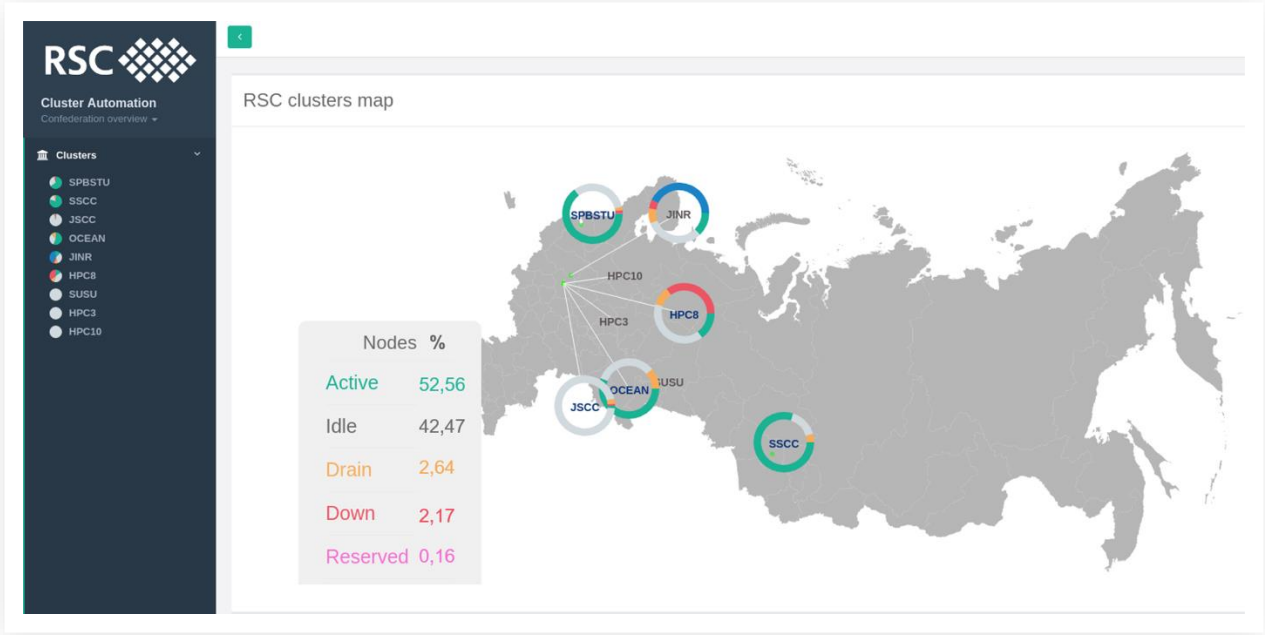
Архитектура РСК БазИС

Рост производительности подсистемы данных суперкомпьютера «Говорун» в ЛИТ ОИЯИ после модернизации 2022 года:



*Приведены данные производительности, полученные только на части добавляемого сегмента

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

PRESENTATION

(RP25) Development of Performance Assessment Method Based on Aspen DSL and Micro-Kernels Benchmarking

Session: Research Posters Session

Poster Authors: Ekaterina Tyutlyaeva, Alexander Moskovsky, Igor Odintsov, Sergey

Event Type: Research Poster

Step 1: Profiling

Starting point of any application performance analysis is a profiling. The profiling allows us to understand an application dynamical structure and identify sections of program code consuming the most parts of execution time.
The Intel® VTune™ Amplifier XE [2] results have been used to visualize call graph of the application under the study.

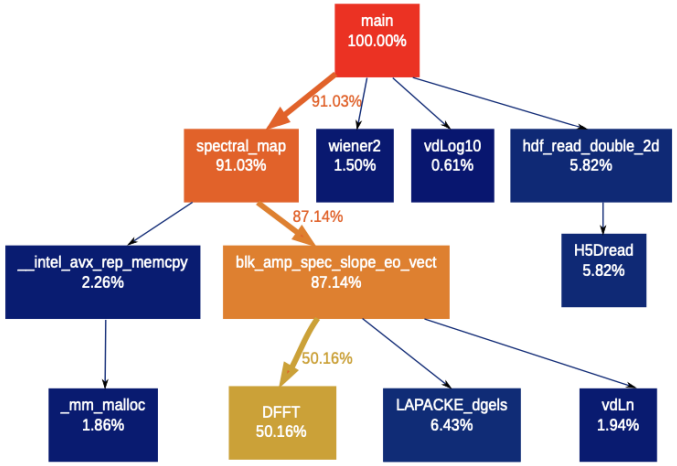


Figure 1: Visualised Call graph of the Studied Application for Skylake testbed

Step 2: Aspen

Step 3: Basic Blocks

Step 4: FLOPs Estimations

Step 5: Memory Usage

- Analytical performance model is created

Step 6: Modeling

Evaluate for exec time for data size/clock freq./cache size etc.

Table 4: Modeling Results for Broadwell Testbed

NxN	24x24	32x32	40x40	48x48	56x56	64x64
T_{exp}	40.268	20.920	28.078	26.095	25.039	21.603
T_{basic_model} Basic Frequency	22.444	15.988	22.263	22.593	23.042	18.755
T_{basic_model} Turbo Mode	16.993	12.333	16.869	16.959	17.438	14.345
$T_{memory\_extended\_model}$ Basic Frequency	25.276	18.030	23.929	24.051	24.369	19.996
$T_{memory\_extended\_model}$ Turbo Frequency	18.255	13.022	17.282	17.37	17.599	14.442

Table 5: Modeling Results for Skylake Testbed

NxN	24x24	32x32	40x40	48x48	56x56	64x64
T_{exp}	32.275	16.737	22.128	19.435	19.469	16.343
T_{basic_model} Basic Frequency	22.856	15.210	23.249	23.702	24.325	19.016
T_{basic_model} Turbo Mode	16.635	11.07	16.921	17.251	17.705	13.84
$T_{memory\_extended\_model}$ Basic Frequency	23.014	15.533	23.782	24.496	25.428	20.480
$T_{memory\_extended\_model}$ Turbo Frequency	16.75	11.305	17.309	17.829	18.507	14.906
$T_{basic\_model\_Broadwell}$ Basic Frequency	21.766	15.534	21.592	21.911	22.345	18.207

Спасибо!

РСК

rscgroup.ru
hq@rsc-tech.ru