РСК
Группа компаний

Hardware and software solutions by RSC: “Govorun” supercomputer refresh 2024-2025.

Alexander Moskovsky, CEO and co-founder

GRID 2025, JINR, Dubna

7 July 2025

15+ years of innovations for HPC, Data Centers and AI compute infrastructure

Development of innovative energy-efficient compute, storage, and software solutions delivering unique features and addressing specific end-user needs

About RSC Group

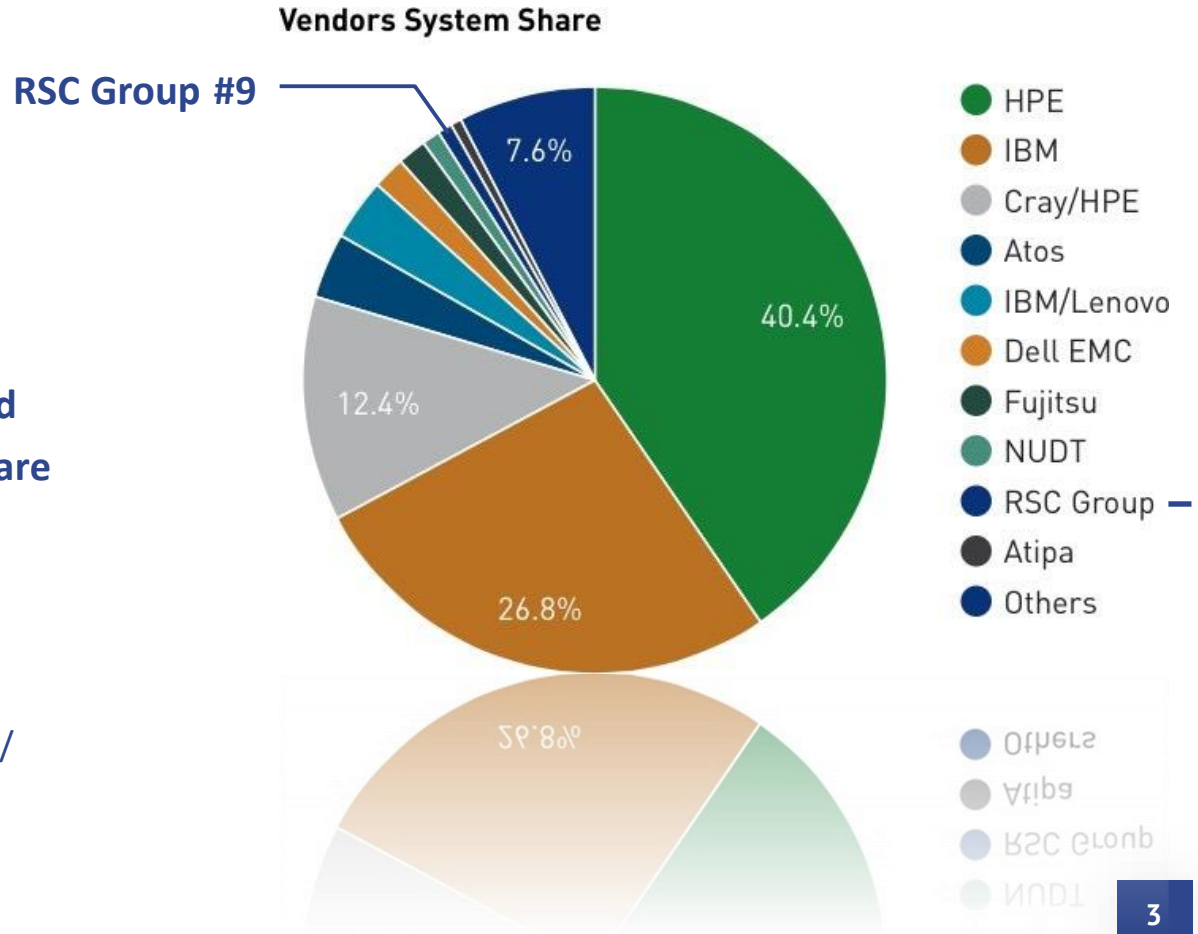
Leading innovative HPC solution provider in Russia/CIS and EMEA

«Best IT-solution for Data Center» nomination at Russian DC Awards 2020

TOP500

The only Russian company ranked in Top10 HPC Vendors System Share by Top500 (Nov 2014)*

*** Top10 suppliers by market volume**
https://www.top500.org/statistics/list/



Strong Market Position

Leading innovative HPC solution provider in Russia/CIS and EMEA

Joint Institute for Nuclear Research

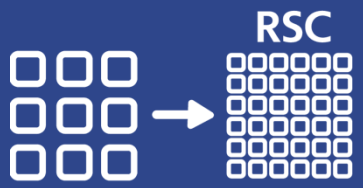
The most **energy efficient** system in Russia

Over **70%** of all Russian systems in HPCG rating

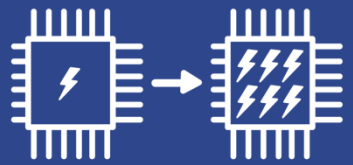
4 of RSC systems – the only Russian systems in **IO500 rating**

- **24%**
- share in Russian **Top50 rating**

HPC. Points Of Excellence



Computing Density



Power Density

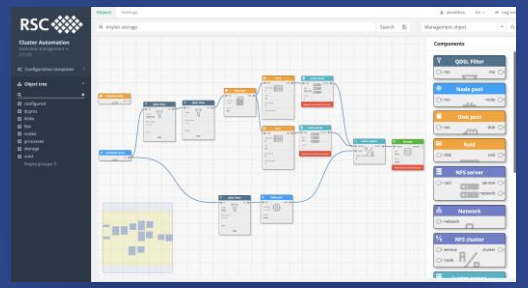
Energy Efficiency

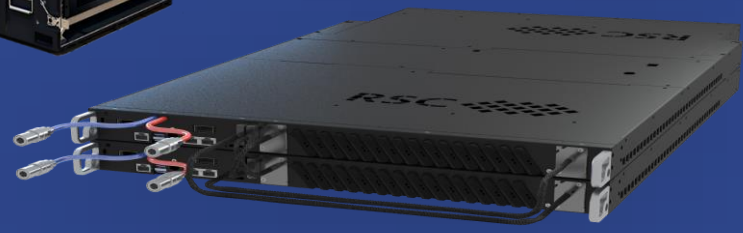
Ease to manage and maintain

Reliability

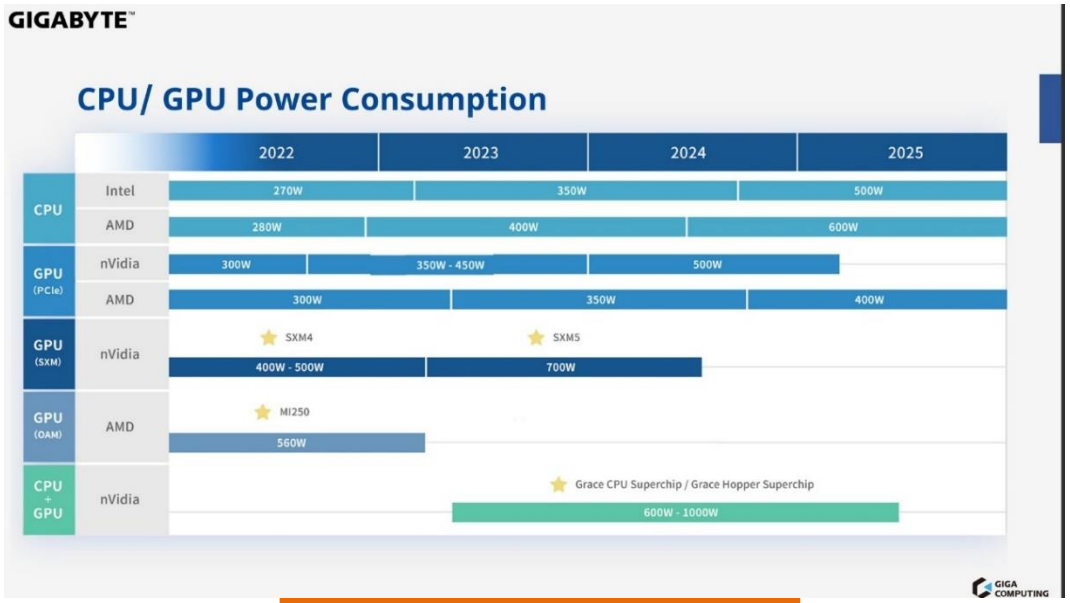
State-of-the-Art & Turn-Key Solutions:

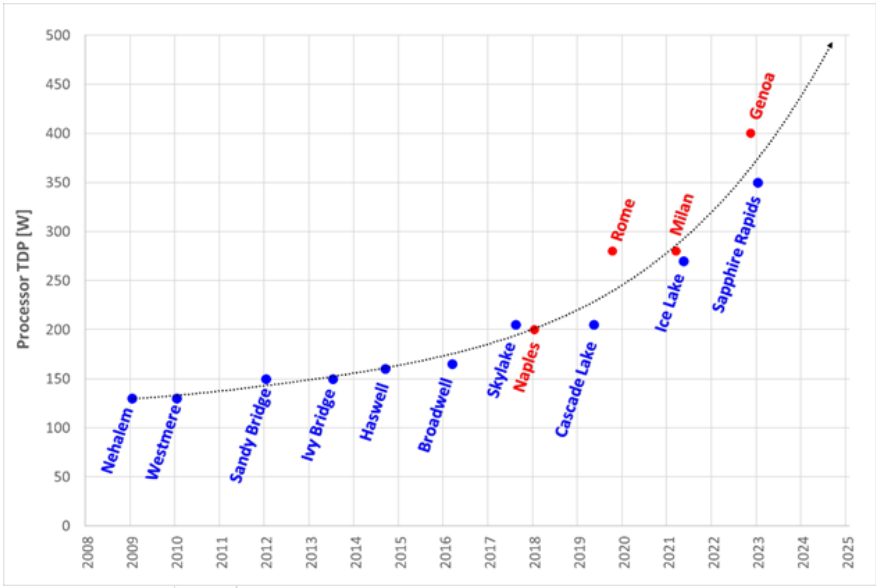
- Energy-Efficient Supercomputers
- High Performance Computing
- Data Centers
- Edge Computing
- AI Compute Infrastructure
- Intellectual Data Storage Systems
- Integrated Data Center Management Software



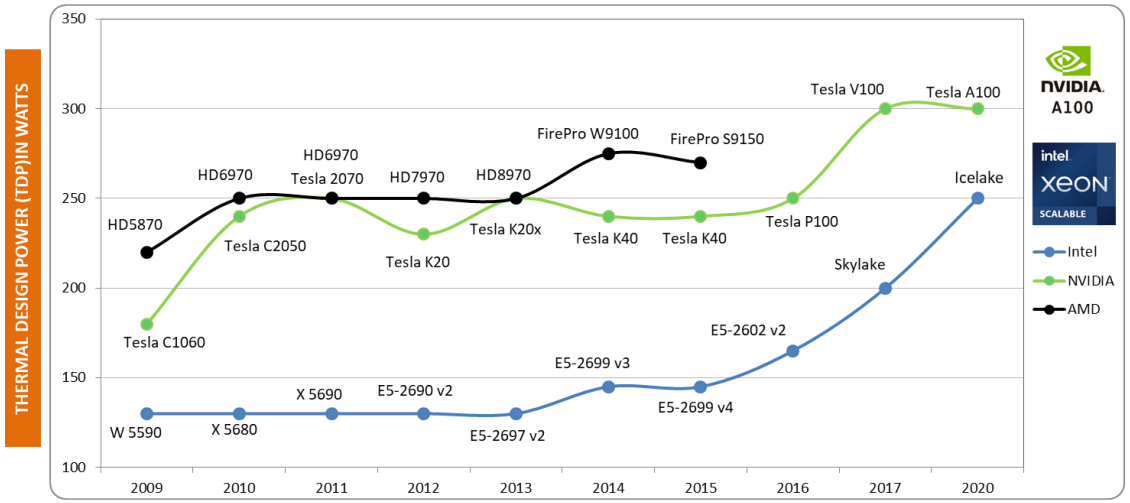


CPU/GPU power consumption is growing fast





GPU and CPU TDP TREND



Liquid cooling in Top20@Top500 rating

15 РСК лет ИННОВАЦИИ ДЛЯ ЛИДЕРСТВА

Большинство самых мощных систем из списка Top500 используют жидкостное охлаждение

Top500 Rank	System	Cooling technology
1	El Capitan	Direct cold water cooling
2	Fugaku	Direct cold water cooling
3	Aurora	Direct cold water cooling
4	JUPITER Booster	Direct warm water cooling
5	Eagle	Airflow cooling
6	HPC6	Direct cold water cooling
7	Fugaku	Direct cold water cooling
8	Alps	Direct cold water cooling
9	LUMI	Direct cold water cooling
10	Leonardo	Direct warm water cooling

Top500 Rank	System	Cooling technology
11	Isambard-AI phase 2	Direct cold water cooling
12	Tuolumne	Direct cold water cooling
13	ISEG2	Direct cold water cooling(?)
14	MareNostrum 5 ACC	Direct warm water cooling
15	ACBI 3.0	Direct warm water cooling
16	Eos NVIDIA DGX SuperPOD	Direct cold water cooling
17	Discovery 6	Direct cold water cooling
18	SCC-24	Direct warm water cooling
19	Venado	Direct cold water cooling
20	Sierra	Direct cold water cooling

Возможно ли охладить воздухом стойку 100 кВт?

15 РСК лет ИННОВАЦИИ ДЛЯ ЛИДЕРСТВА

При увеличении плотности набивки ИТ-стойки увеличивается внутреннее сопротивление потоку воздуха. Электрическая мощность вентилятора пропорционально кубу его производительности.

МОЩНОСТЬ ИТ-СТОЙКИ, кВт	ПЕРЕПАД ТЕМПЕРАТУРЫ, ^{0}C	ТЕМПЕРАТУРА ВХОД / ВЫХОД, ^{0}C	РАСХОД ВОЗДУХА, М^3/ЧАС	ПОТЕРИ ДАВЛЕНИЯ ВНУТРИ СЕРВЕРА, Па	МОЩНОСТЬ ВЕНТИЛЯТОРОВ , кВт	МОЩНОСТЬ ИТ, кВт	СООТНОШЕНИЕ МОЩНОСТИ, %
15	15	25 / 40	2 994	50	0,10	15	0,7
30	15	25 / 40	5 988	200	0,83	29	2,8
60	15	25 / 40	11 976	800	6,65	53	11,1
100	15	25 / 40	19 960	2222	30,80	69	30,8
100	12	40 / 52	24 950	3472	60,16	40	60,2

* Выполнен оценочный расчет

Интегрированный подход: адаптивность на разных уровнях

Уровень	Средства адаптации
Прикладное ПО	Модели программирования, средства разработки
Связующее ПО (Промежуточное ПО)	Конфигурации под задачу Оркестратор (РСК БазИС)
Аппаратные платформы	Программно-определяемые конфигурации
Инженерная инфраструктура	Агенты мониторинга, управления

Ключевые проекты

Наука

- Объединенный институт ядерных исследований (ОИЯИ)
- Российская академия наук (МСЦ РАН)
- Физико-технический институт имени Иоффе РАН
- Институт математики имени Соболева СО РАН
- Сибирский суперкомпьютерный центр (ИВМиМГ 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 Social Media Corp.
- Авиастроение
- Автомобилестроение
- Энергетика
- Компьютерная графика
- Нефте- и газодобыча

и другие

Решение «РСК Экзастрим ИИ» для развития вычислительной инфраструктуры ИИ

5th phase Govorun modernization

2024-2025 гг.

2 новых узла
«РСК Экзастрим ИИ»
(8 карт Nvidia H100 в каждом)

2 новых узла хранения
RSC Tornado AFS (2x1 ПБ)

Прирост производительности
416 Тфлопс (37%)

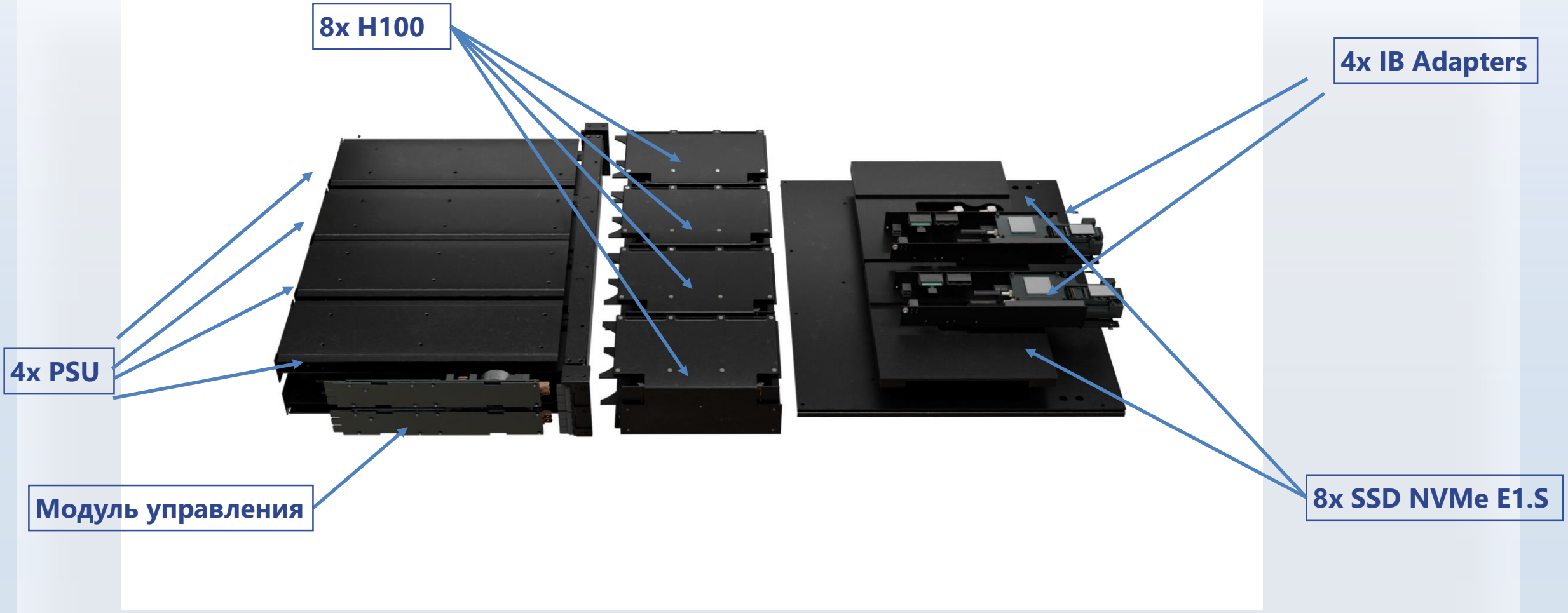
Суммарная производительность системы
2,2 ПФлопс

Compute node “RSC Exastream AI”

Конфигурация серверов «РСК Экзастрим ИИ», установленных в ОИЯИ:

- высота узла **2U**,
- **два процессора Intel Xeon Platinum 8468** (4-го поколения, 48 ядер, тактовая частота 2,1-3,8 ГГц, объем кэш-памяти 105 МБ),
- **8 графических ускорителей NVidia H100** (PCIe, 80 ГБ),
- **1 ТБ** оперативной памяти,
- **16 ТБ** емкости хранения данных на базе SSD-дисков с интерфейсом NVMe,
- **4 блока питания** производства РСК,
- **система прямого жидкостного охлаждения** разработки РСК.

Inside «RSC Exastream AI»



Сервер «РСК Экзастрим ИИ» 2U: 208/408 TFLOPS (FP64/TF64) до 2 ТБ ОЗУ, до 128 ТБ SSD

21 сервер в шкафу «РСК Экзастрим» 42U: 4,368/8,568 PFLOPS (FP64/TF64), 115 кВт на шкаф

Вычислительный узел «РСК Экзастрим ИИ»

- Графические карты с прямым жидкостным охлаждением (до 8 ГПУ) попарно объединены мостами на базе технологии высокоскоростных соединений NVLink для обеспечения быстрой передачи данных между графическими процессорами.

- Имеет локальную подсистему хранения «теплых данных», сетевую подсистему с доступом на основе технологии GPUDirect.

- Реализована возможность расширения ресурсов путем подключения дополнительных пар GPU или системы внешнего хранения данных на базе пула твердотельных дисков (JBOF), подключаемой напрямую к серверу.

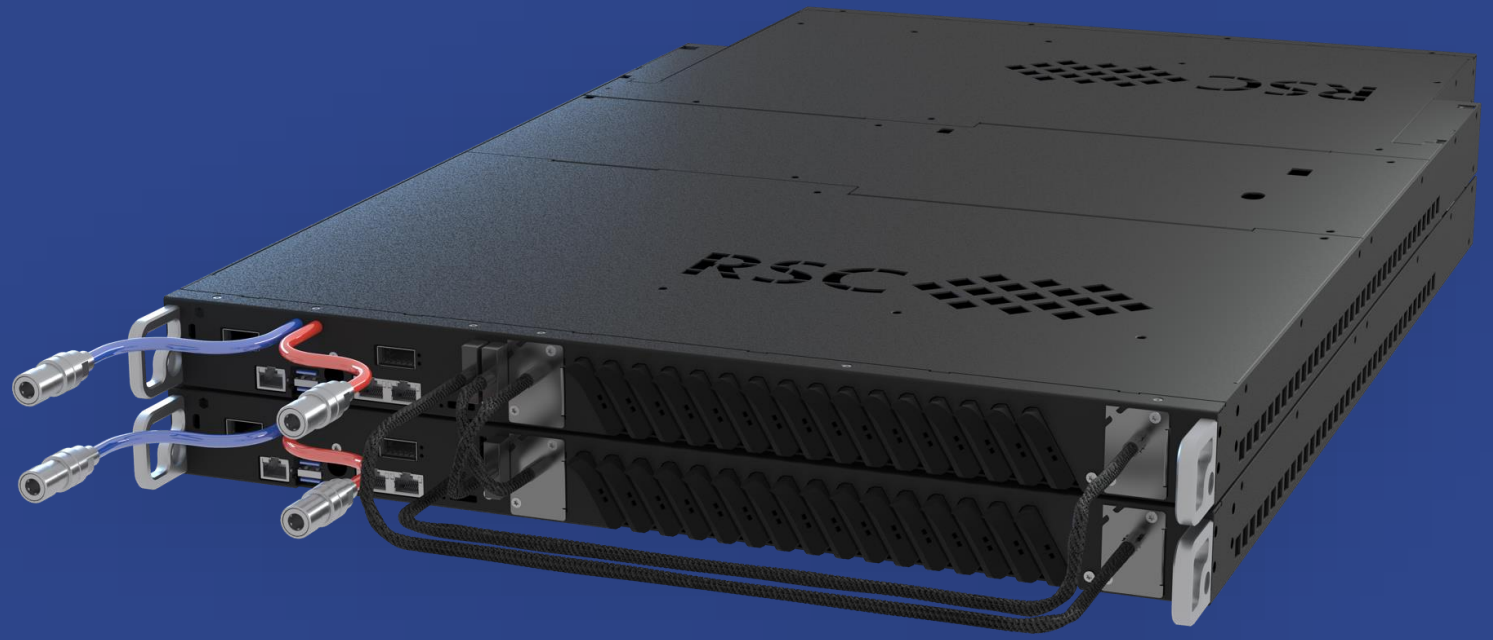
Узел «РСК Экзастрим ИИ»

15 РСК лет ИННОВАЦИИ ДЛЯ ЛИДЕРСТВА

Параметр	Значение
Поддержка ЦПУ	Intel Xeon Scalable 4-го или 5-го поколения
Оперативная память	DDR5, 32 слота, до 2 ТБ
Поддержка ГПУ	До 8-ми NVidia H100/H200 Pairwise NVLink
Постоянная память	До 8-ми SSD рулеров в форм-факторе EDSFF 1.S, суммарно 128 ТБ
Коммутация	До 4-х портов Infiniband HDR/NDR
Сеть	10 ГБ Ethernet
Система охлаждения	100% жидкостное охлаждение, температура 40-50 С
Блок питания	Разработка РСК, прямое жидкостное охлаждение

RSC Tornado AFS storage with record density

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



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

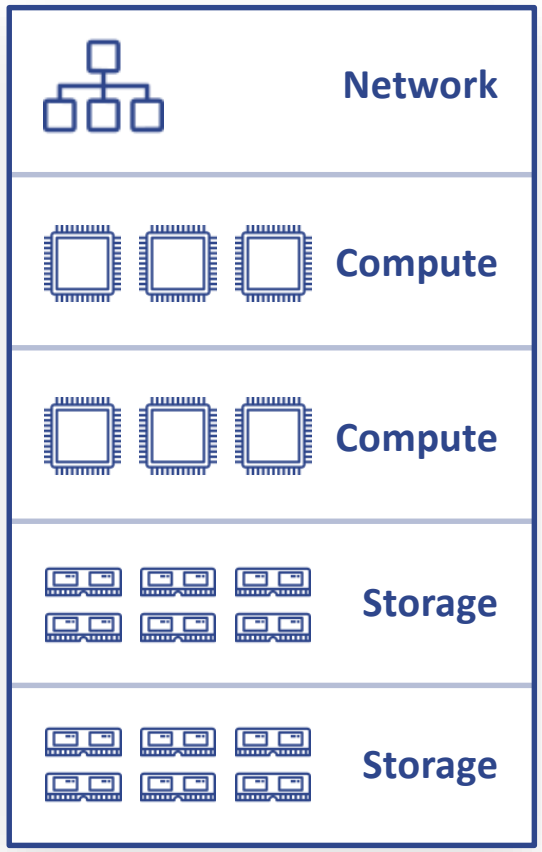
Первое решение на 100% жидкостном охлаждении с высочайшей плотностью на базе 32x Intel EDSFF SSD, двух процессоров Intel Xeon Scalable и памяти Intel Optane DC Persistent Memory

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

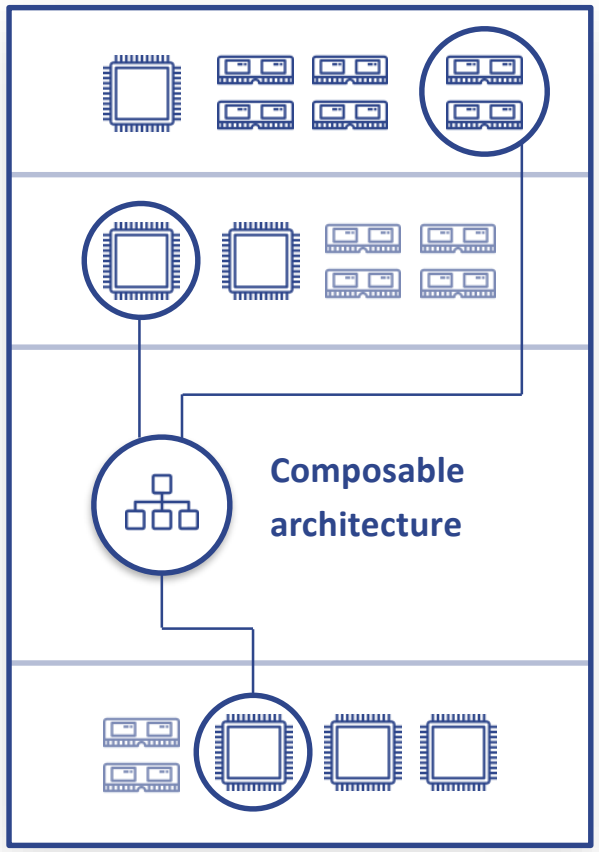
Composable Disaggregated Infrastructure

PCK

Rack Scale Architecture → Composable Disaggregated Infrastructure (CDI)

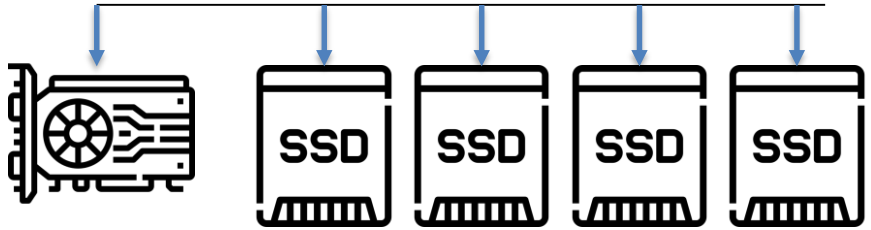


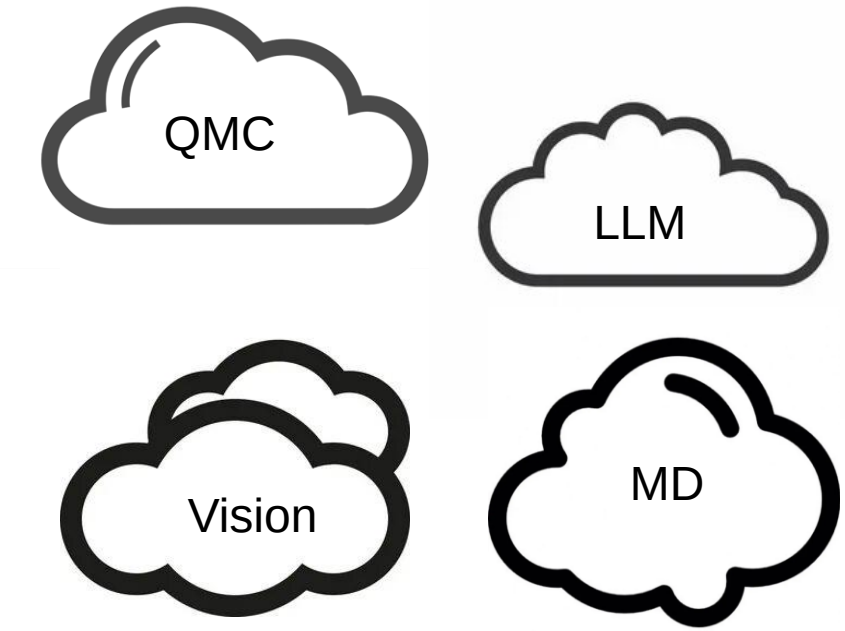
- ⊕ Hyperconvergence
- ⊕ Modern storage and network building blocks
- ⊕ Software orchestration
- ⊕ Storage on-demand
- ⊖ Software virtualization



RSC BasIS: GPU Direct and SSD management

- Оптимальное соотношение GPU и SSD **под задачу**
- GPU Direct – быстрый доступ к данным для приложений на GPU
- NVM-over-Fabric – управление устройствами хранения
- Патент РФ «**Метод построения высокопроизводительных отказоустойчивых систем хранения данных на основе распределенных файловых систем и технологии NVMe over Fabrics**»
- Реестр российского ПО (РСК БазИС, РСК Базис СХД)





Сэм Альтман, OpenAI

The rate of new wonders being achieved will be immense. It’s hard to even imagine today what we will have discovered by 2035; maybe we will go from solving high-energy physics one year to beginning space colonization the next year;

By TechCrunch - TechCrunch Disrupt San Francisco 2019 - Day 2, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=92008259

https://blog.samaltman.com/the-gentle-singularity

MProt-DPO: Breaking the ExaFLOPS Barrier for Multimodal Protein Design Workflows with Direct Preference Optimization

Gautham Dharuman[1†], Kyle Hippe[1,2†], Alexander Brace[1,2†], Sam Foreman[1†], Väinö Hatanpää[1], Varuni K. Sastry[1], Huihuo Zheng[1], Logan Ward[1], Servesh Muralidharan[1], Archit Vasan[1], Bharat Kale[1], Carla M. Mann[1,2], Heng Ma[1], Yun-Hsuan Cheng[3], Yuliana Zamora[3], Shengchao Liu[5], [illegible], Murali Emani[1], Tom Gibbs[3], Mahidhar Tatineni[7], Deepak Canchi[8], Jerome Mitchell[8], K[illegible], Maria Garzaran[8], Michael E. Papka[1,9], Ian Foster[1,2], Rick Stevens[1,2], [illegible], Anima Anandkumar[10*], Venkatram Vishwanath[1,9*], Arvind Ramanathan[1,2]
[1]Argonne National Laboratory, [2]University of Chicago, [3]NVIDIA Inc., [4]Swiss National Superc
[5]University of California, Berkeley, [6]University of Wisconsin-Madison, Madison, [7]San Diego Supe
[8]Intel Corporation, [9]University of Illinois Chicago, [10]California Institute of Techno
[†]Joint first authors, *Contact authors: venkat@anl.gov, anima@caltech.edu, ramanathana

***Abstract*—We present a scalable, end-to-end workflow for protein design. By augmenting protein sequences with natural language descriptions of their biochemical properties, we train generative models that can be preferentially aligned with protein fitness landscapes. Through complex experimental- and simulation-based observations, we integrate these measures as preferred parameters for generating new protein variants and demonstrate**

sustained performance of 4.11 ExaFl
peak performance of 5.57 ExaFLOPS.
a breakthrough for protein design work
imental and simulation observables acr
(such as natural language and protein
an interface for end users to develop p

GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics

Maxim Zvyagin[1†], Alexander Brace[1,2†], Kyle Hippe[1†], Yuntian Deng[3,4†], Bin Zhang[5], Cindy Orozco Bohorquez[5], Austin Clyde[1,2], Bharat Kale[6], Danilo Perez-Rivera[1,7], Heng Ma[1], Carla M. Mann[1,2], Michael Irvin[1], J. Gregory Pauloski[2], Logan Ward[1], Valerie Hayot-Sasson[1,2], Murali Emani[1], Sam Foreman[1], Zhen Xie[1], Diangen Lin[1,2], Maulik Shukla[1,2], Weili Nie[3], Josh Romero[3], Christian Dallago[3,9], Arash Vahdat[3], Chaowei Xiao[8,3], Thomas Gibbs[3], Ian Foster[1,2], James J. Davis[1,2], Michael Brettin[1], Rick Stevens[1,2], Anima Anandkumar[3,11*], Venkatram Vishwanath[1*], Arvind Ramanathan[1*]
[illegible]atory, [2]University of Chicago, [3]NVIDIA Inc., [4]Harvard University, [5]Cerebras Inc., [6]Northern [illegible]York University, [8]Arizona State University, [9]Technical University of Munich, [10]University of Illinois Chicago, [11]California Institute of Technology
[illegible]rs, *Contact authors: venkat@anl.gov, anima@caltech.edu, ramanathana@anl.gov

Learning-at-Criticality in Large Language Models for Quantum Field Theory and Beyond

Xiansheng Cai,[1,*] Sihan Hu,[2,3,*] Tao Wang,[4,5,†] Yuan Huang,[6,‡] Pan Zhang,[1,7,§] Youjin Deng,[2,3,¶] and Kun Chen[1,**]

[1]*Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China*
[2]*Department of Modern Physics, University of Science and Technology of China, Hefei, Anhui 230026, China*
[3]*Hefei National Laboratory, University of Science and Technology of China, Hefei 230088, China*
[4]*Department of Physics, University of Massachusetts, Amherst, MA 01003, USA*
[5]*Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China*
[6]*DP Technology, Beijing 100080, China*
[7]*School of Fundamental Physics and Mathematical Sciences, Hangzhou Institute for Advanced Study, UCAS, Hangzhou 310024, China*

(Dated: June 11, 2025)

Fundamental physics often confronts complex symbolic problems with few guiding exemplars or established principles. While artificial intelligence (AI) offers promise, its typical need for vast datasets to learn from hinders its use in these information-scarce frontiers. We introduce learning at criticality (LaC), a reinforcement learning (RL) scheme that tunes Large Language Models (LLMs) to a sharp learning transition, addressing this information scarcity. At this transition, LLMs achieve peak generalization from minimal data, exemplified by 7-digit base-7 addition—a test of nontrivial arithmetic reasoning. To elucidate this peak, we analyze a minimal concept-network model (CoNet) designed to capture the essence of how LLMs might link tokens. Trained on a single exemplar,

Jun 2025

Спасибо!

РСК

ИННОВАЦИИ ДЛЯ ЛИДЕРСТВА

rscgroup.ru
hq@rsc-tech.ru
oleg.gorbachov@rscgroup.ru