



**High performance on demand storage for inference support – the
benefit of composable disaggregated infrastructure"**

Alexander Moskovsky, CEO and co-founder

DLCP 2026, SINP, Dubna

6 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

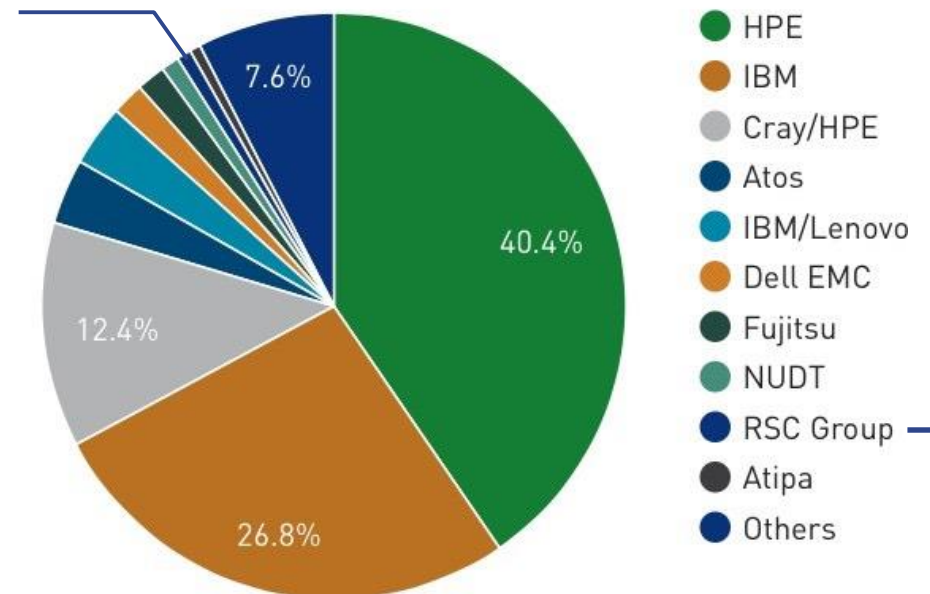


The only Russian company ranked
in Top10 HPC Vendors System Share
by Top500 (Nov 2014)*

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

Vendors System Share

RSC Group #9



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

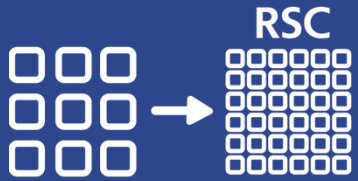


4 of RSC systems –
the only Russian
systems in **IO500**
rating

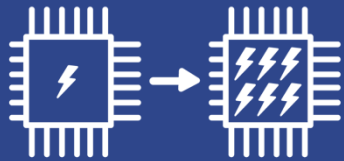


- **24%**
• share in
Russian **Top50**
rating

HPC. Points Of Excellence



Computing
Density



Power
Density



Energy
Efficiency



Ease to manage
and maintain

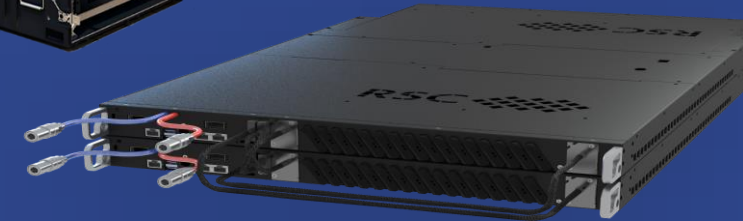
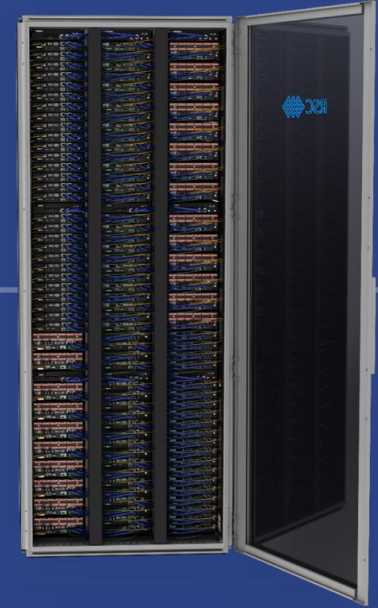
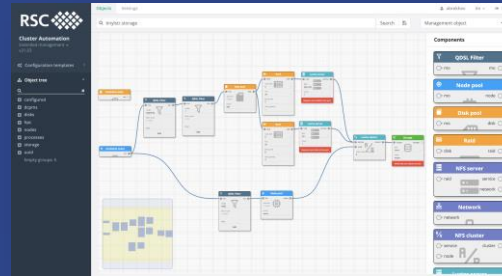


Reliability

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**



Liquid cooling in Top20@Top500 rating

Liquid cooling now prevails in Top20 supercomputers

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

Exastream AI

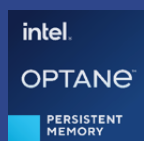
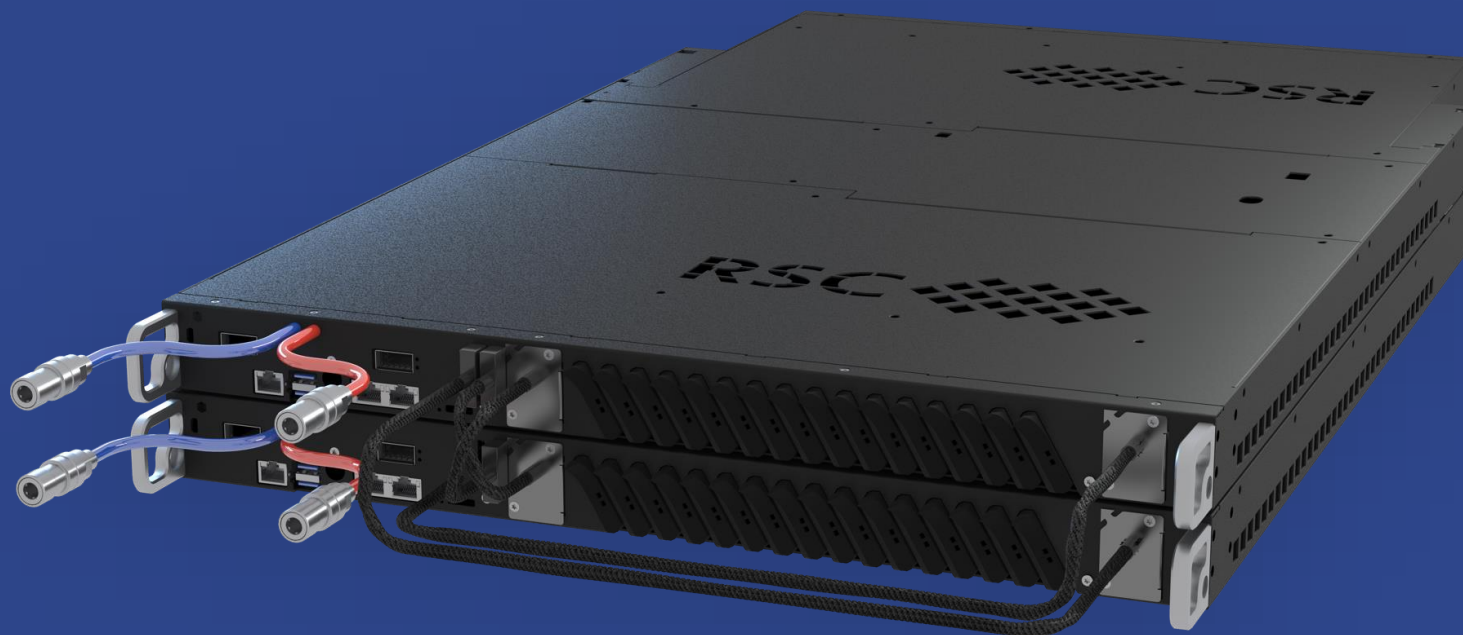
- **2U** form factor
- **Dual Intel Xeon Platinum 84XX/85XX CPU,**
- **8 GPUs NVidia H100** (PCIe, 80 GB), pairwise NVlink
- Up to **2 TB RAM ECC,**
- Up to **8 SSD** form factor **EDSFF 1.S,** up to **128 TB total**
- NVMe SSD interface
- Up to 4 high speed 200/400 Gbps ports
- 10 Gig E auxillary network
- **4 PSU** liquid cooled
- 100% liquid cooled, warm coolant
- Peak performance **208/408 TFLOPS (FP64/TF64) ,**



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)**

Scale Stream - C

Внешний массив PCIe-ускорителей оптимизирован для установки самых производительных GPU/TPU-карт и предназначен для расширения возможностей серверов с целью решения сложных задач в областях высокопроизводительных вычислений, машинного обучения и ИИ.

Массив поддерживает до **10-ти PCIe x16 ускорителей** с возможностью использования карт разной ширины и объединения их мостами, а также подключения до **4-х серверов** через внешние кабели на базе стандарта **PCIe x16 Gen4**.

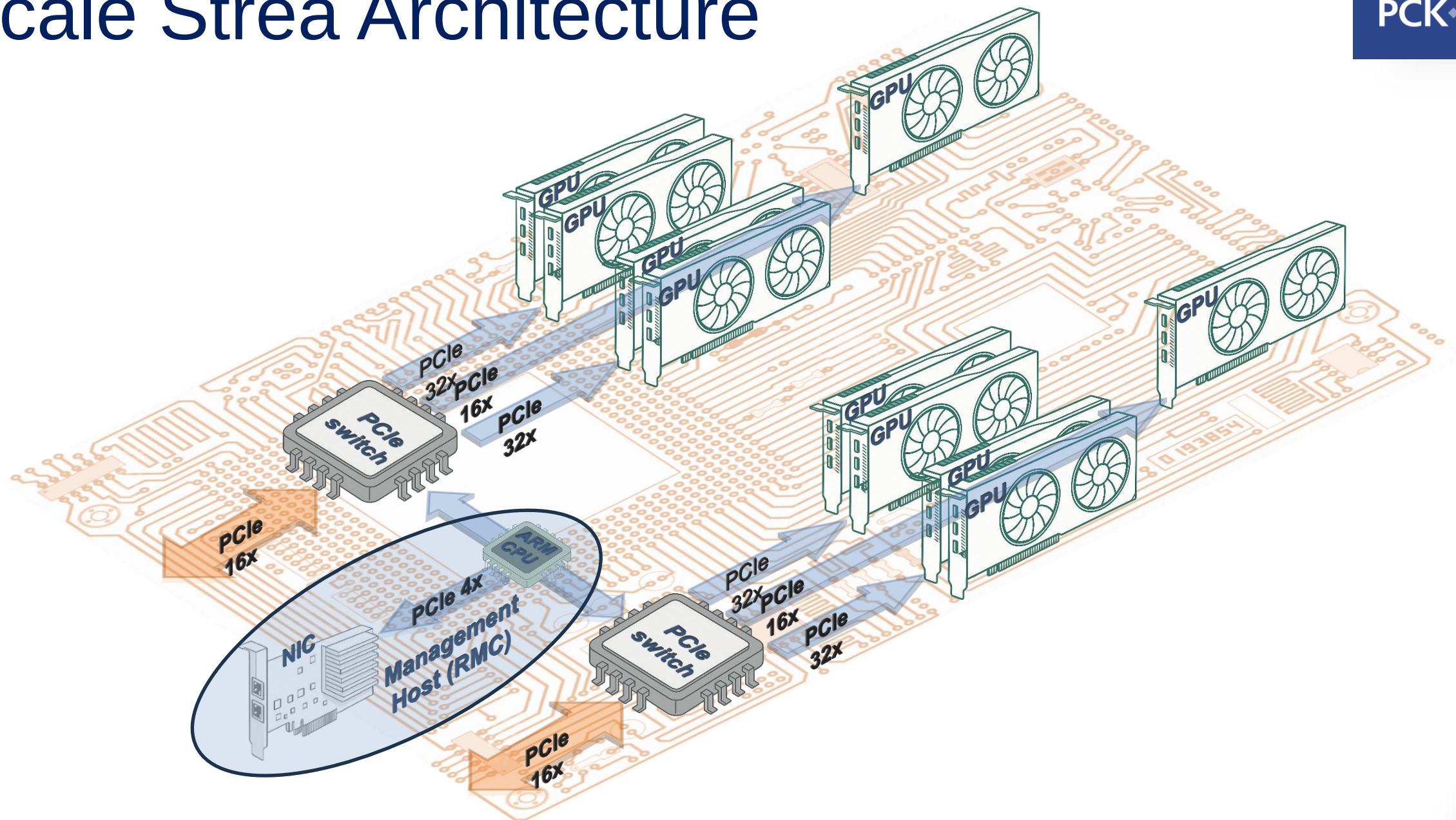
Достигаемая производительность:

До **300 ТФлопс (FP64)** на массив из 10-ти карт NVIDIA H200.

До **960 TOPS (int8)** на массив с 10-ю картами LinQ HPQ.

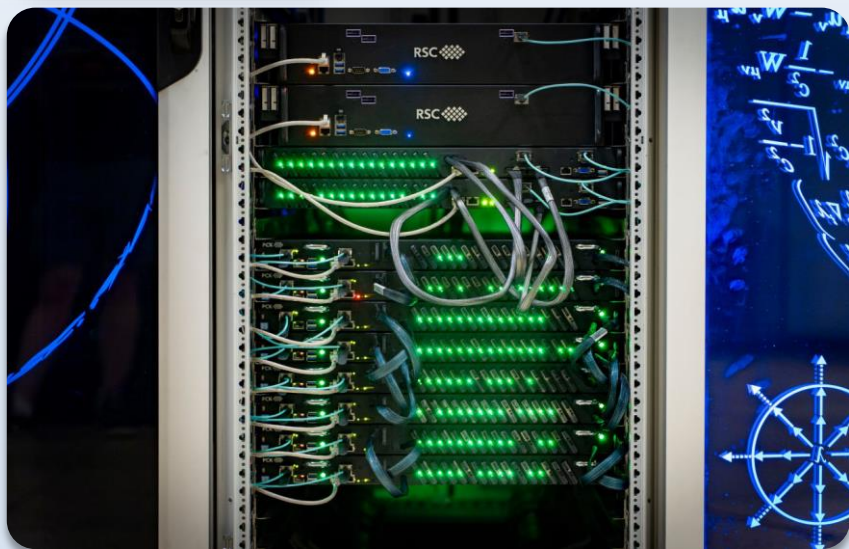


Scale Strea Architecture



Govorun Supercomputer (JINR) – Modernization, Stage 5

2024-2025



- 2 new RSC ExaStream AI nodes (8x Nvidia H100 cards each)
- 2 new RSC Tornado AFS storage nodes (2x1 PB)
- Performance gains **416 TFlops (37%)**
- The total system performance **2.2 PFlops**

Science

- Joint Institute for Nuclear Research (JINR)
- Russian Academy of Sciences (JSCC RAS)
- Institute of Mathematics SB RAS
- Siberian Supercomputer Center (SSCC SB RAS)
- Ioffe Physical-Technical Institute of the Russian Academy of Sciences (Ioffe PTI RAS)
- Institute of Oceanology of the Russian Academy of Sciences (IO RAS)
- Institute of Atmospheric Physics of the Russian Academy of Sciences (IAP RAS)
- Russian Weather Forecast Agency (Roshydromet)

Education

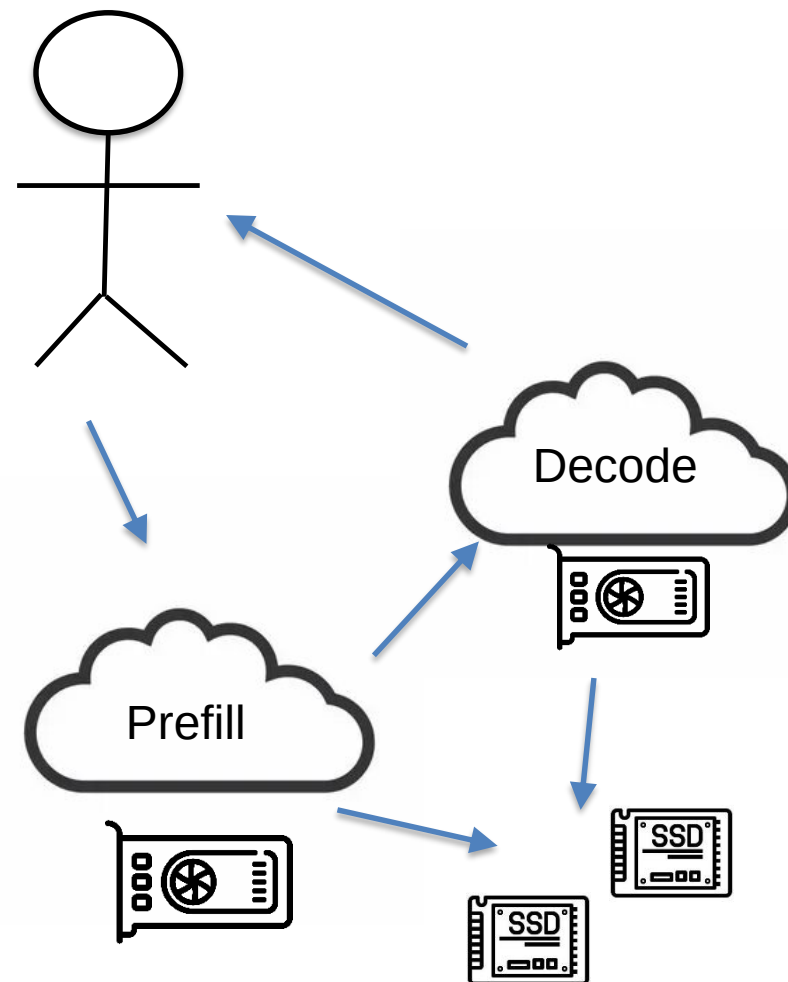
- Saint Petersburg Polytechnic University (SPbPU)
- Moscow State University (MSU)
- South Ural State University (SUSU)
- Moscow Institute of Physics and Technology (MIPT)
- State University of Nizhny Novgorod (UNN)

Industry

- VK Social Media Corp.
- Aviation
- Automotive
- Energy
- Computer Graphics
- Oil&Gas
- And other

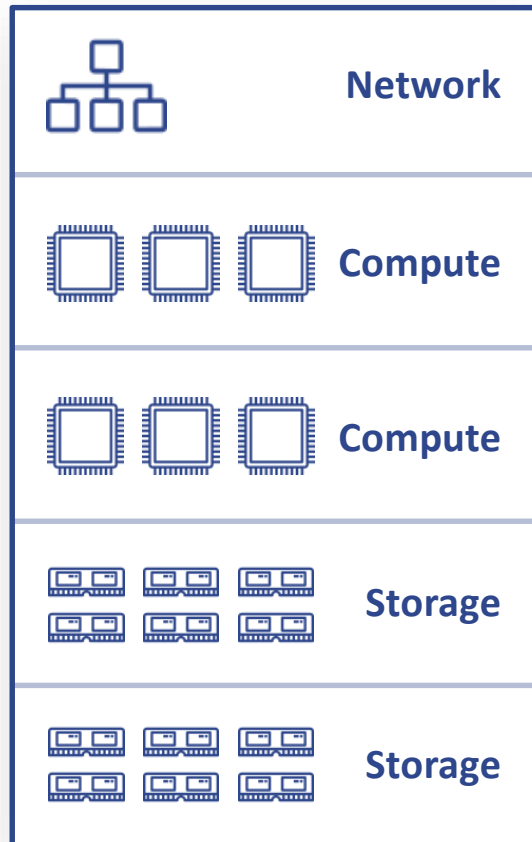
KV-cache problem with LLM

- LLM inference today – two phases
- **Prefill phase** – memory and compute intensive
- **Decode phase** – less numeric intensive, more memory intensive
- **KV-cache is a result of prefill phase**
- One could take advantage of two phases
- Less expensive GPUs for decode phase
- Cache survives disruptions



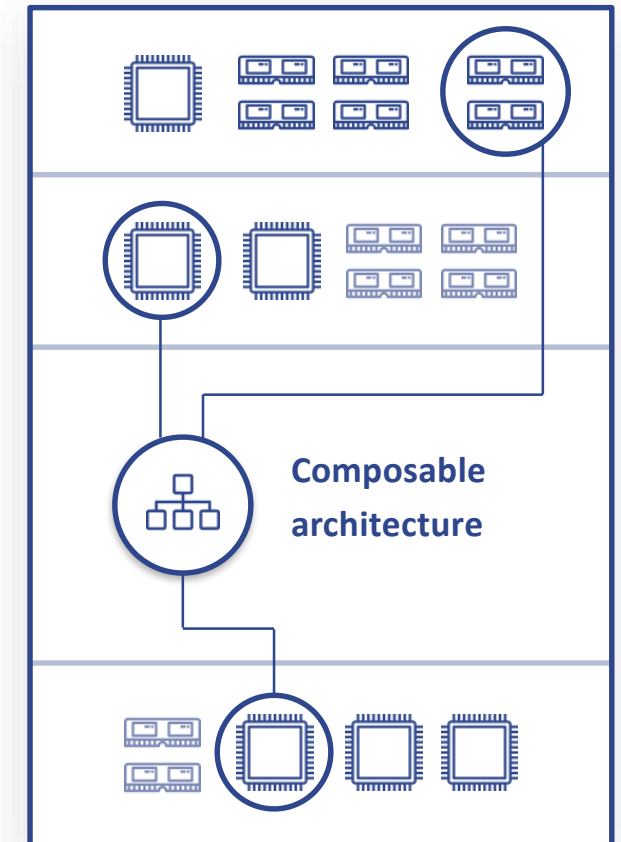
Composable Disaggregated Infrastructure

Rack Scale Architecture



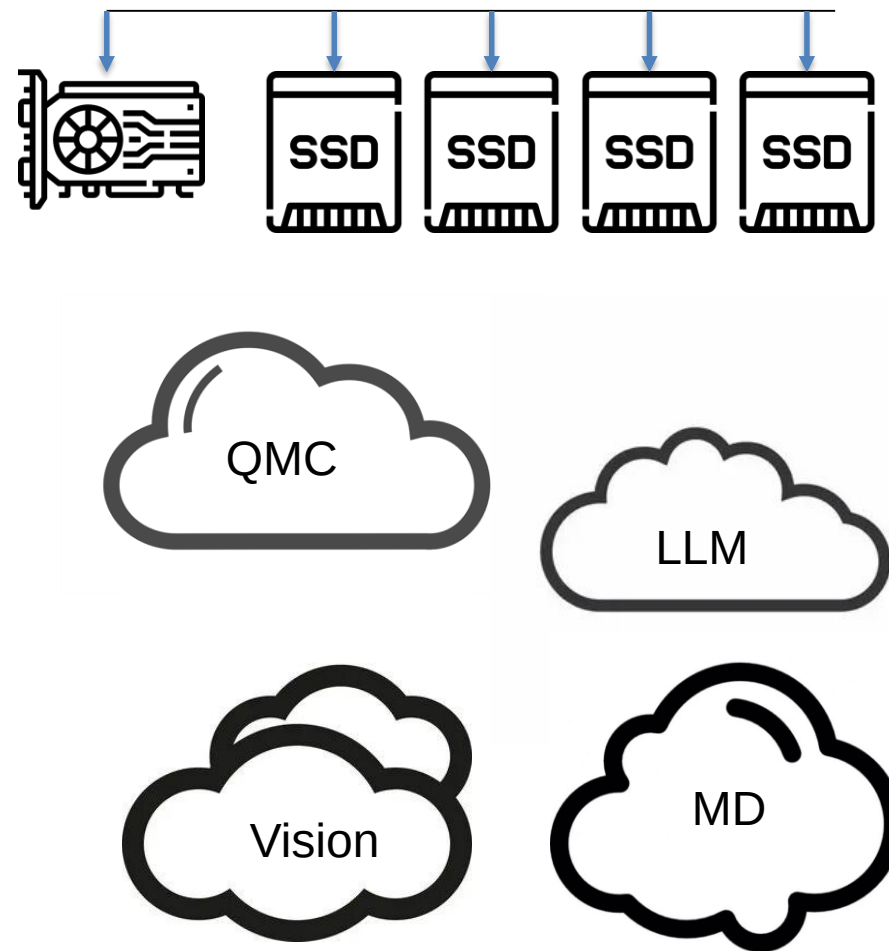
- ⊕ Hyperconvergence
- ⊕ Modern storage and network building blocks
- ⊕ Software orchestration
- ⊕ Storage on-demand
- ⊖ Software virtualization

Composable Disaggregated Infrastructure (CDI)

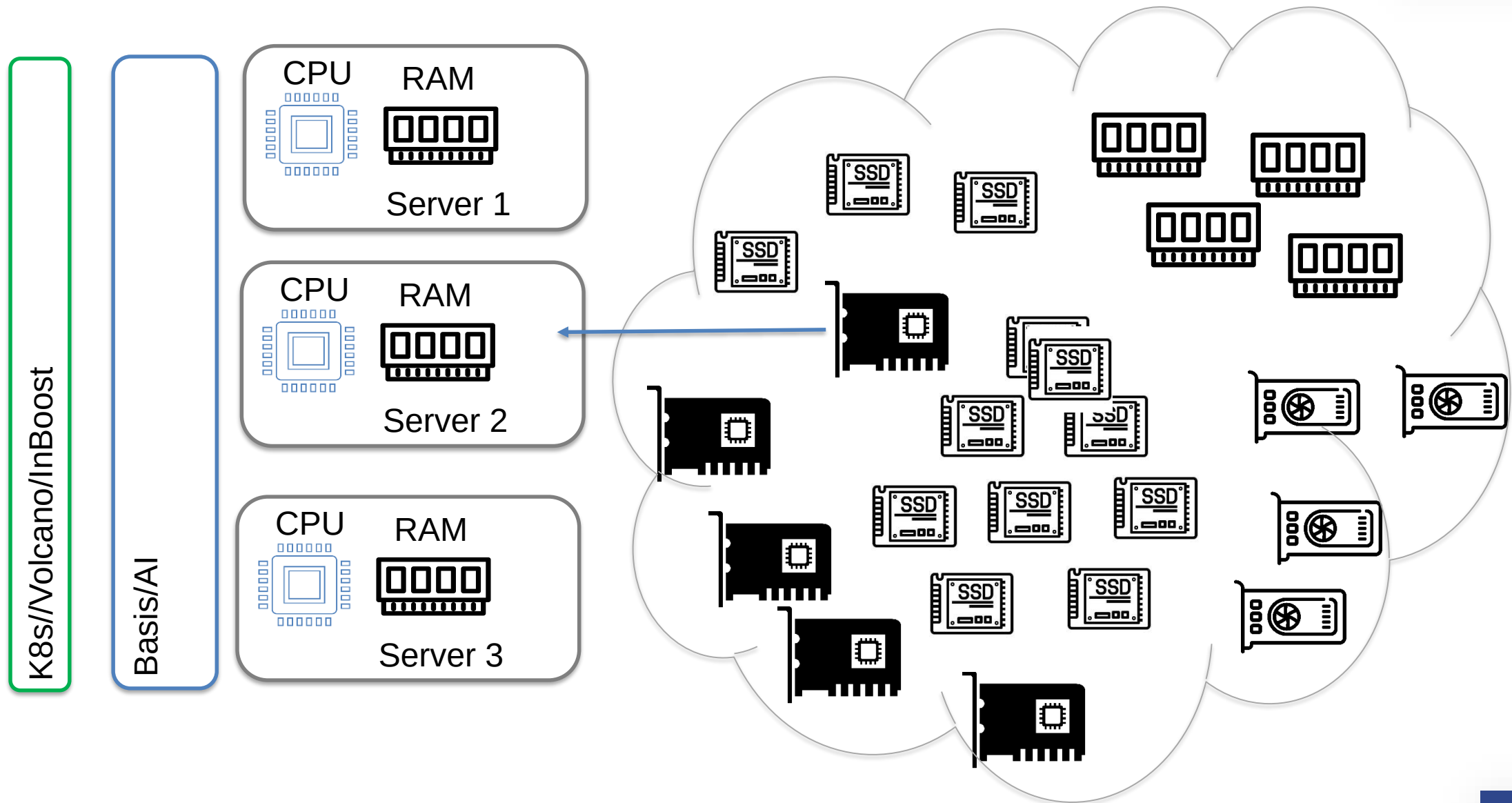


RSC BasIS: GPU Direct and SSD management

- Optimal balance for GPU and SSD for a task
- GPU Direct – fast access from GPU to data
- NVM-over-Fabric – storage device management
- RU patent РФ «Метод построения высокопроизводительных отказоустойчивых систем хранения данных на основе распределенных файловых систем и технологии NVMe over Fabrics»
- Ministry of Digital RU software registry (PCK БазИС, PCK Базис СХД)

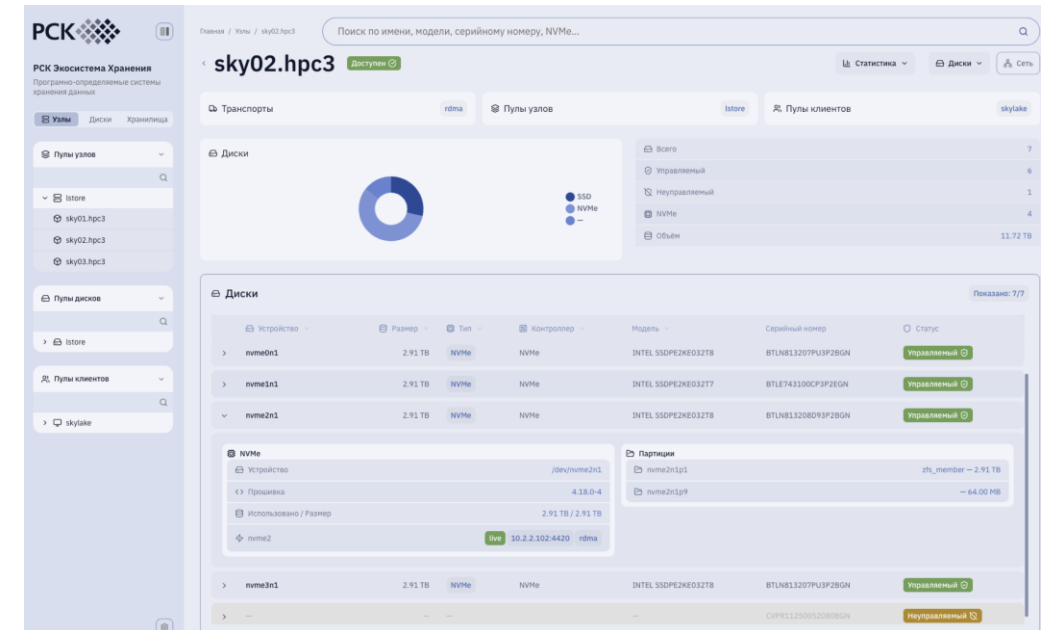


Software-defined ML/Inference factory



High performance file system tuning for KV cache

- GPU direct access
- Sharding
- KV cache pieces are kb-mb size
- Write once, read many (replication)
- Hierarchy and placement management (future works)



Genesis mission (USA)

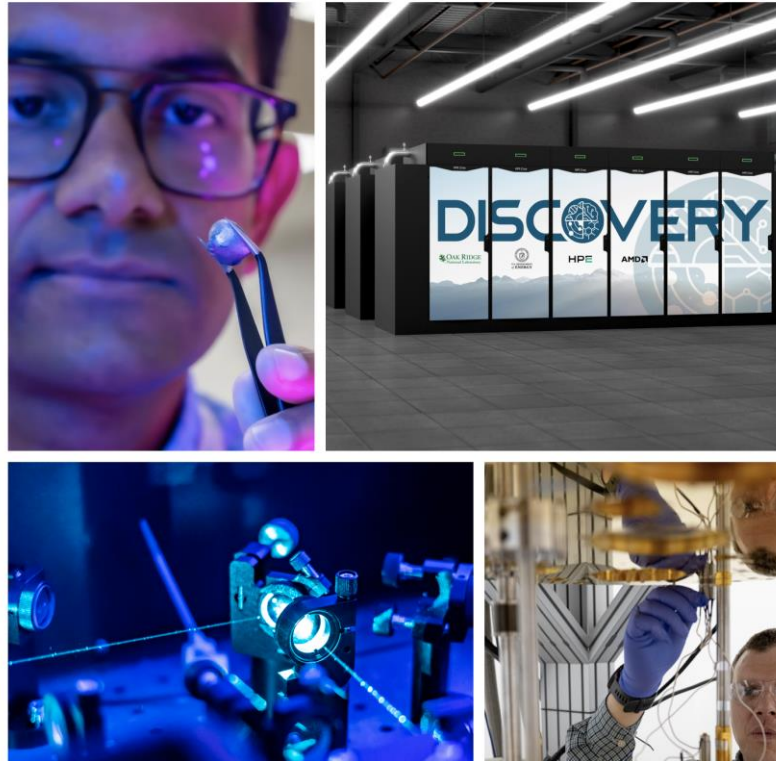
- AI for Science, data and compute integration

ACCELERATING DISCOVERY

A new platform for AI-driven scientific discovery

The Genesis Mission is a national initiative led by the Department of Energy and its 17 national laboratories to build the world's most powerful scientific platform to accelerate discovery science, strengthen national security, and drive energy innovation. It does so by enabling AI-driven, exascale-powered advances that enhance America's energy innovation, global competitiveness, and security.

By connecting computing, data, and experimental facilities into a unified system, Genesis accelerates discovery and shortens the path from research to real-world impact, strengthening U.S. leadership in energy, security, and advanced technology.



Discovery

Discovery is an HPE system powered by AMD processors and accelerators that will arrive in 2028 featuring impressive speed and enhanced bandwidth across the system.

[Learn More >](#)



Frontier

ORNL's exascale supercomputer is delivering world-leading performance.

[Learn More >](#)

Autonomous Laboratories at Oak Ridge National Laboratory

Autonomous laboratories integrate artificial intelligence with automated experimentation and advanced instrumentation to dramatically accelerate scientific discovery. At Oak Ridge National Laboratory, researchers are developing interconnected research platforms that combine AI, leadership-class computing, and cutting-edge instrumentation into an emerging ecosystem of self-driving labs. These prototype "Labs of the Future" are paving the way for a new generation of AI-driven scientific discovery.



Thanks



hq@rscgroup.ru
www.rscgroup.ru
+7(495)6403107

