Петербургский институт ядерной физики им. Б. П. Константинова
Национального исследовательского центра «Курчатовский институт»

Компьютерная инфраструктура в обработке экспериментальных данных (по материалам весенней конференции HEPIX 15-19 апреля 2024)

Семинар: ОФВЭ

2024-06-25

Докладчик: А.Е. Шевель

План

- Объёмы экспериментальных данных и сложность обработки
- Конференция HEPIX
- Программа конференции
- Статистика конференции HEPIX-2024
- Тренды в инструментах обработки данных
- Дистрибутивы Линукс
- Состояние и обновления в компьютерной инфраструктуре
- Передача данных
- Прогнозы развития компьютерных компонентов
- Рассуждения

Running large physics experiments

- 4+ experiments CERN
- 2+ experiments at BNL
- 1+ at ORNL
- 2+ JINR
- And dozen+ others
  - Many hundreds of PB/year

Future large experiments in preparation

- Large Synoptic Survey Telescope (lsst.org)
- The Square Kilometre Array (https://www.skao.int)
- CTAO - Cherenkov Telescope Array Observatory (ctao.org)
  - Another many hundreds of PB/year

HEPIX.org

- The HEPiX forum brings together worldwide Information Technology staff, including system administrators, system engineers, and managers from the High Energy Physics and Nuclear Physics laboratories and institutes, to foster a learning and sharing experience between sites facing scientific computing and data challenges.
- The HEPiX organization was formed in 1991, and its semi-annual meetings are an excellent source of information and sharing for IT experts in scientific computing.

The conference program

- Site reports
- Networking&Security
- Storage&Filesystems
- Operating systems, Clouds & Virtualization, Grids
- Computing & Batch Services
- IT Facilities, Business Continuity and Green IT
- Basic and End-User IT Services
- "Show me your toolbox"
- In total: 54 contributions in the standard tracks (free abstract submissions) duration (incl. Q&A): 20 hours 40 minutes
  - Total number of participants = 107

Data toolkits
trends in laboratories

CernVM-FS

CernVM File System

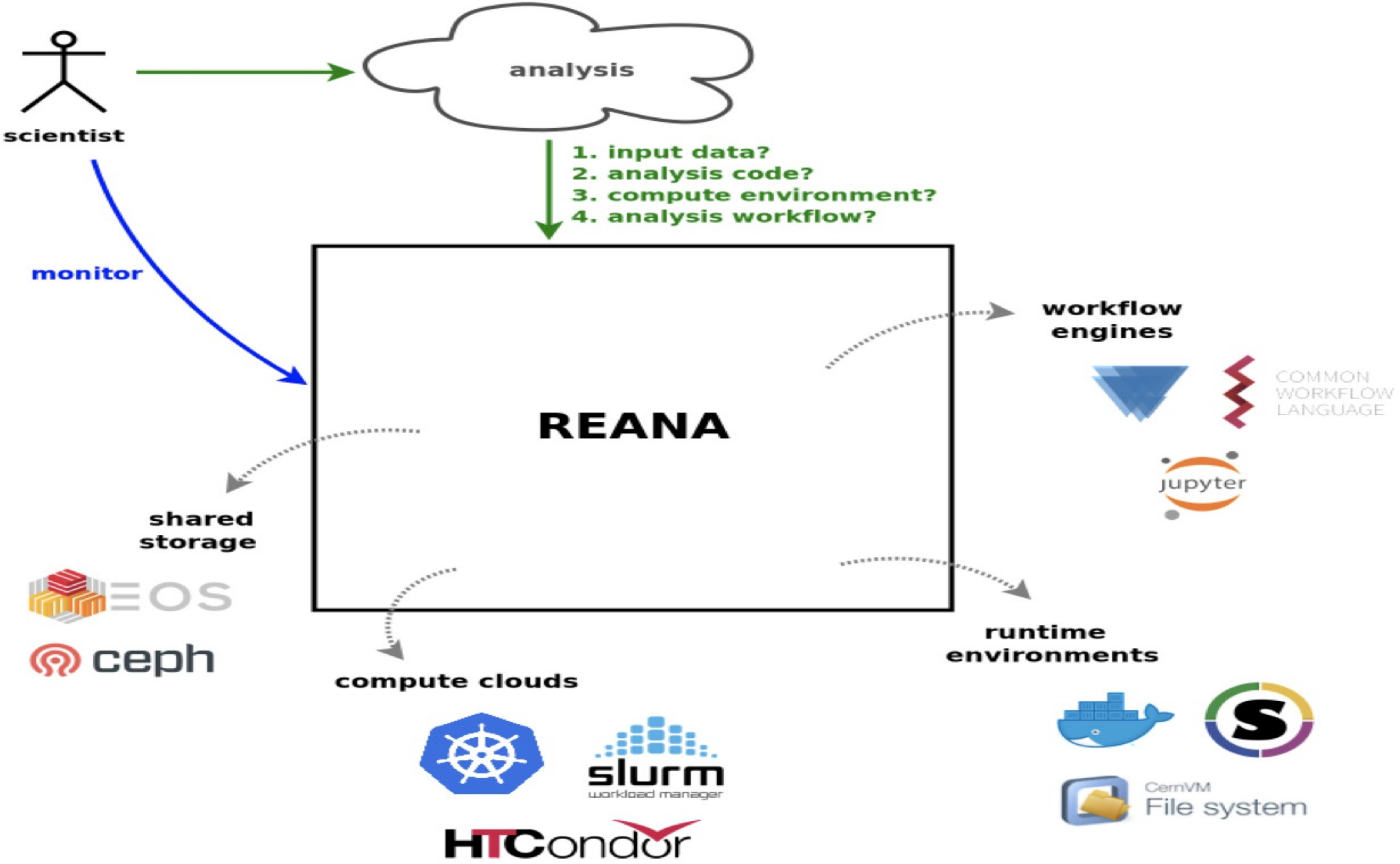
O

Data analysis toolkits

- **coffea** представляет собой прототип пакета для объединения всех типичных потребностей анализа экспериментов по физике высоких энергий (HEP) с использованием экосистемы Python.
  - https://coffeateam.github.io/coffea/

Data analysis workflow

- REANA (REusable ANAlysis) is a software platform developed at CERN (European Organization for Nuclear Research).
  - REANA allows users to store and execute data analysis workflows.
  - One of the key features of REANA is the removal of limits on the number of times a workflow can be restarted. This version also provides support for ARM.
  - REANA allows users to automatically mount any CVMFS repository.



scientist

analysis

1. input data?
2. analysis code?
3. compute environment?
4. analysis workflow?

monitor

REANA

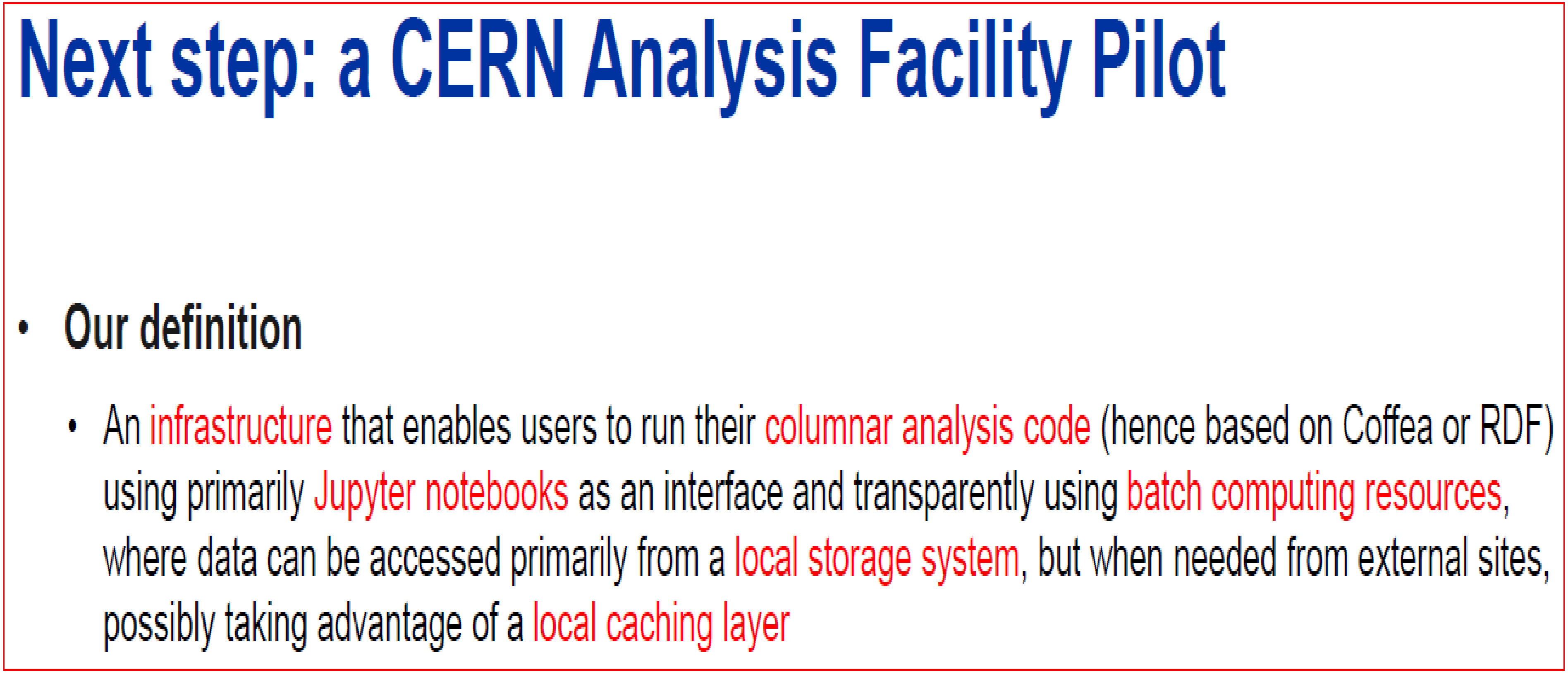
workflow engines

shared storage

compute clouds

runtime environments

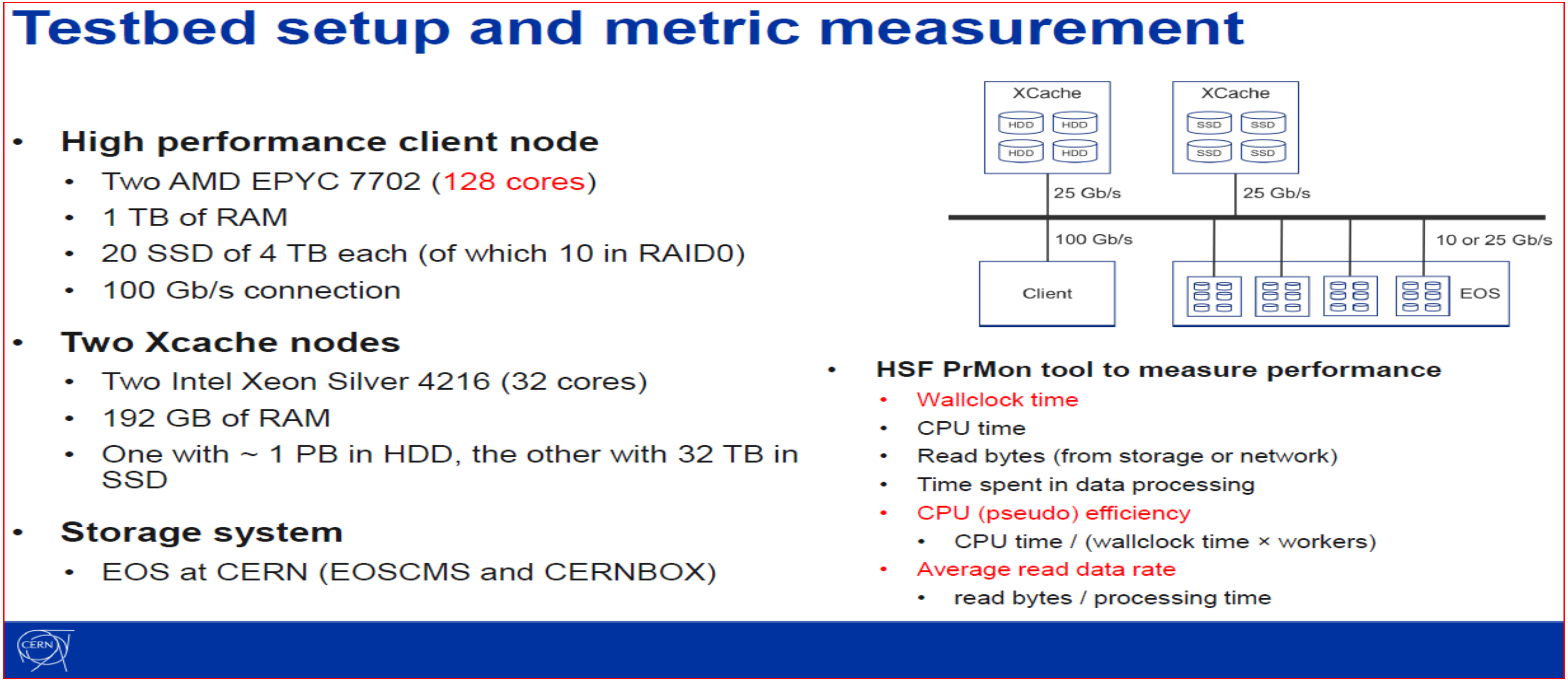
Testbeds for analysis facility



Next step: a CERN Analysis Facility Pilot

- Our definition
  - An infrastructure that enables users to run their columnar analysis code (hence based on Coffea or RDF) using primarily Jupyter notebooks as an interface and transparently using batch computing resources, where data can be accessed primarily from a local storage system, but when needed from external sites, possibly taking advantage of a local caching layer

Testbeds for analysis facility



Testbed setup and metric measurement

- **High performance client node**
  - Two AMD EPYC 7702 (128 cores)
  - 1 TB of RAM
  - 20 SSD of 4 TB each (of which 10 in RAID0)
  - 100 Gb/s connection
- **Two Xcache nodes**
  - Two Intel Xeon Silver 4216 (32 cores)
  - 192 GB of RAM
  - One with ~ 1 PB in HDD, the other with 32 TB in SSD
- **Storage system**
  - EOS at CERN (EOSCMS and CERNBOX)

XCache
HDD
HDD
HDD
HDD
XCache
SSD
SSD
SSD
SSD
25 Gb/s
25 Gb/s
100 Gb/s
10 or 25 Gb/s
Client
EOS

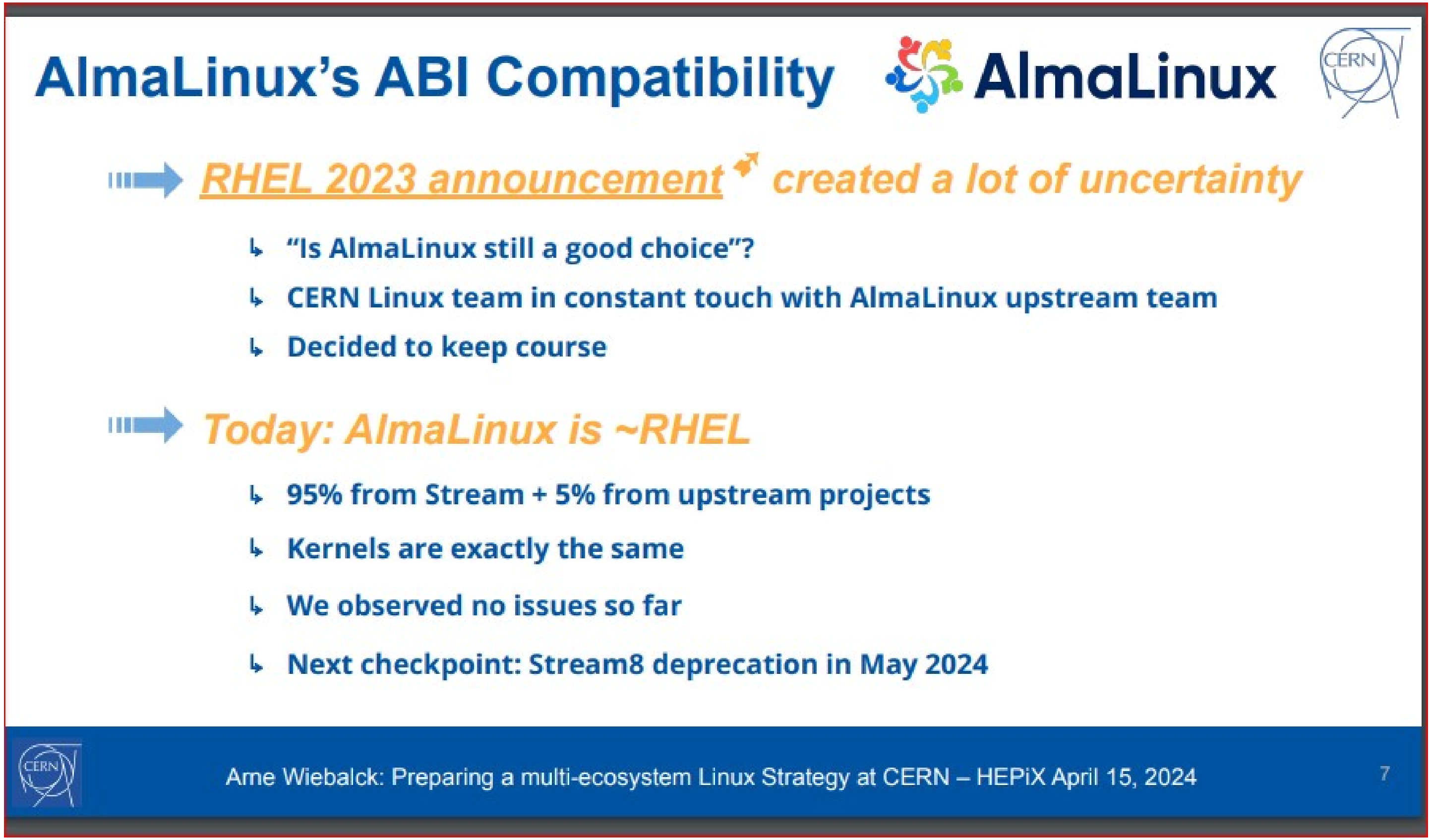
- **HSF PrMon tool to measure performance**
  - Wallclock time
  - CPU time
  - Read bytes (from storage or network)
  - Time spent in data processing
  - CPU (pseudo) efficiency
    - CPU time / (wallclock time × workers)
  - Average read data rate
    - read bytes / processing time

Linux distributions

Linux future discussion

- CentOS 7 is in EOL
- What will be after … ?
  - AlmaLinux (Rocky) 9.x
  - Debian 10
- Now most laboratories plan (or use already) AlmaLinux (Rocky) 9.3

AlmaLinux



AlmaLinux’s ABI Compatibility

AlmaLinux

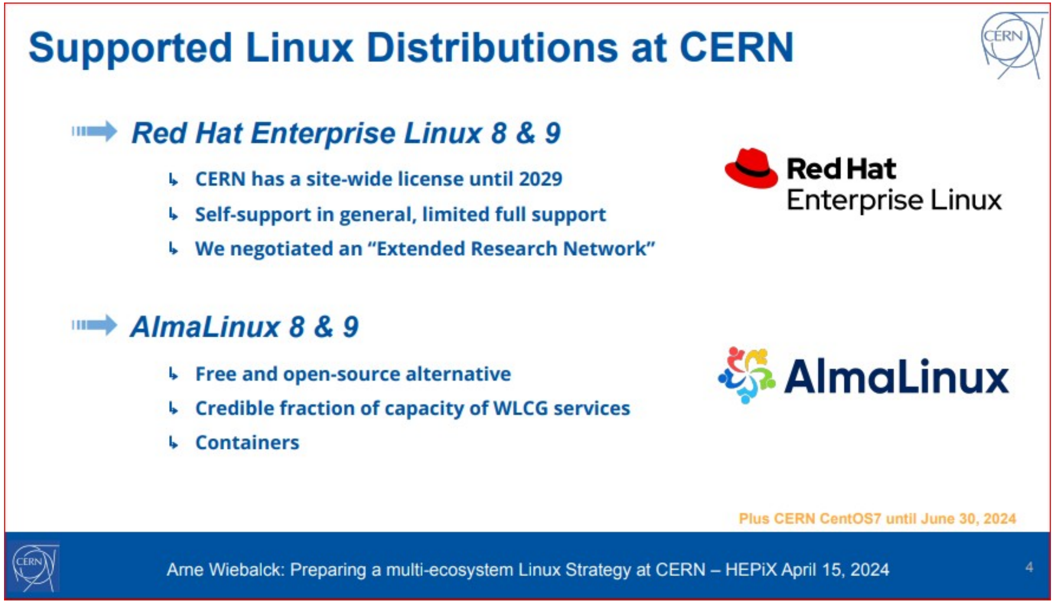
CERN

- *RHEL 2023 announcement created a lot of uncertainty*
  - “Is AlmaLinux still a good choice”?
  - CERN Linux team in constant touch with AlmaLinux upstream team
  - Decided to keep course
- *Today: AlmaLinux is ~RHEL*
  - 95% from Stream + 5% from upstream projects
  - Kernels are exactly the same
  - We observed no issues so far
  - Next checkpoint: Stream8 deprecation in May 2024

Arne Wiebalck: Preparing a multi-ecosystem Linux Strategy at CERN – HEPiX April 15, 2024 7

Linux at CERN

Now



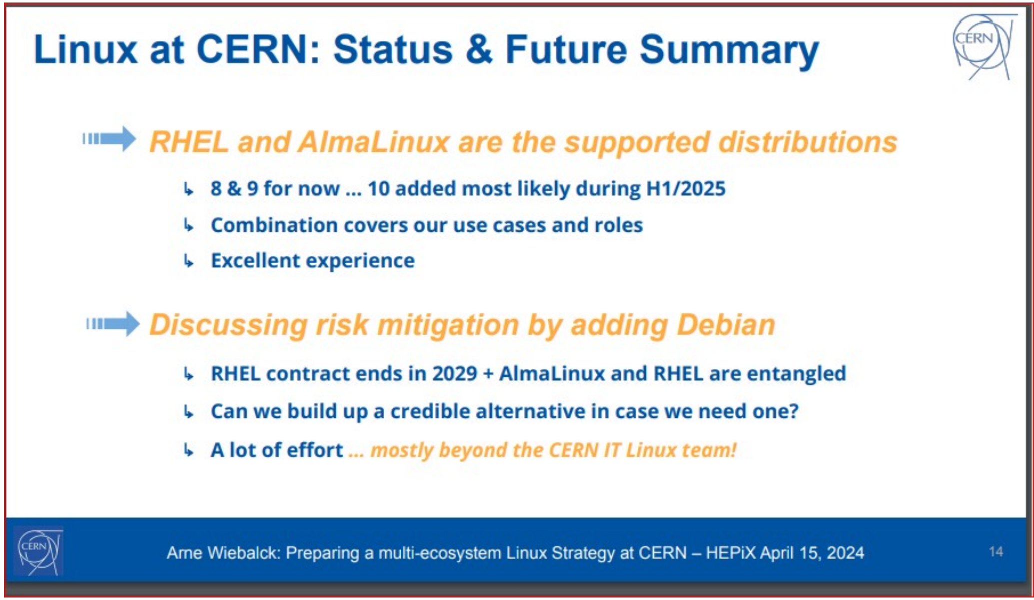
Supported Linux Distributions at CERN

- *Red Hat Enterprise Linux 8 & 9*
  - ↳ CERN has a site-wide license until 2029
  - ↳ Self-support in general, limited full support
  - ↳ We negotiated an "Extended Research Network"
- *AlmaLinux 8 & 9*
  - ↳ Free and open-source alternative
  - ↳ Credible fraction of capacity of WLCG services
  - ↳ Containers

Plus CERN CentOS7 until June 30, 2024

Arne Wiebalck: Preparing a multi-ecosystem Linux Strategy at CERN – HEPiX April 15, 2024 4

Nearest future



Linux at CERN: Status & Future Summary

- *RHEL and AlmaLinux are the supported distributions*
  - ↳ 8 & 9 for now … 10 added most likely during H1/2025
  - ↳ Combination covers our use cases and roles
  - ↳ Excellent experience
- *Discussing risk mitigation by adding Debian*
  - ↳ RHEL contract ends in 2029 + AlmaLinux and RHEL are entangled
  - ↳ Can we build up a credible alternative in case we need one?
  - ↳ A lot of effort … *mostly beyond the CERN IT Linux team!*

Arne Wiebalck: Preparing a multi-ecosystem Linux Strategy at CERN – HEPiX April 15, 2024 14

Linux distribution for the Future

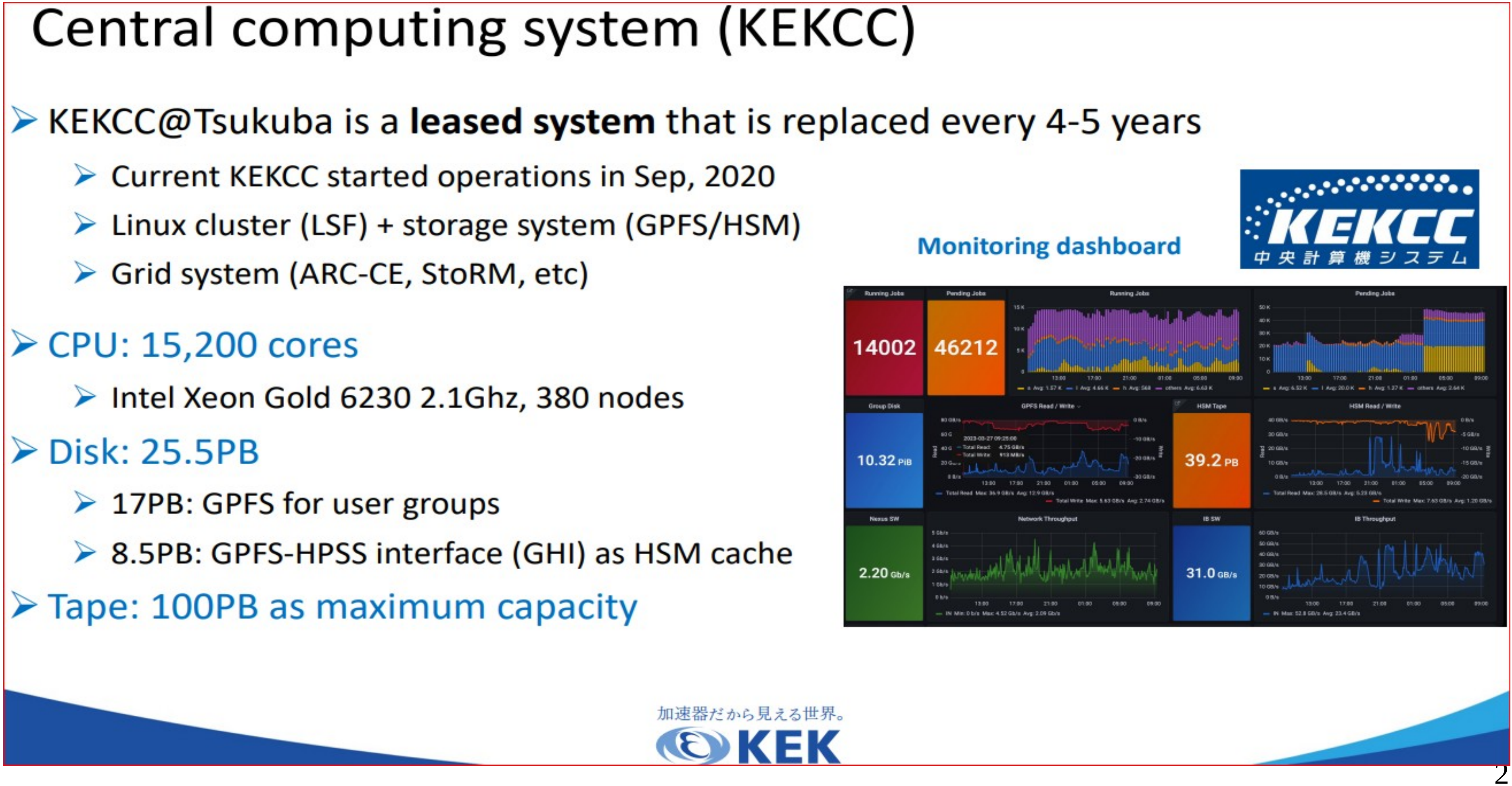
- **RHEL** (Red Hat Enterprise Linux) and clones are the primary Linux distribution at numerous sites.
- **AlmaLinux** is currently the most popular **RHEL** clone. While Rocky Linux is still in use, it is less common than it was previously.
- Some sites continue to use Ubuntu as personal workstations and **Debian** for server side**.** There is interest in positioning Debian as a potential alternative to RHEL and its clones.

Computing infrastructure status and upgrades

Computing cluster Upgrades

- Most laboratories shown upgrade in computing performance (from 10% to 30% and more year-2-year).
- All laboratories have at least 100Gbit external connectivity.
  - KEK upgraded to 400Gbit
  - CERN tested 800Gbit

Cluster organization



Central computing system (KEKCC)

- KEKCC@Tsukuba is a **leased system** that is replaced every 4-5 years
  - Current KEKCC started operations in Sep, 2020
  - Linux cluster (LSF) + storage system (GPFS/HSM)
  - Grid system (ARC-CE, StoRM, etc)
- CPU: 15,200 cores
  - Intel Xeon Gold 6230 2.1Ghz, 380 nodes
- Disk: 25.5PB
  - 17PB: GPFS for user groups
  - 8.5PB: GPFS-HPSS interface (GHI) as HSM cache
- Tape: 100PB as maximum capacity

Monitoring dashboard

Storage systems

- EOS, dCache;
- Lustre;
- Ceph;
- HPSS (Hierarchical High Performance Storage System from IBM);
- Tape storage.

Docs/Video/Zoom

- Streaming Science Globally: CERN’s Live Streaming Service
  - https://indico.cern.ch/event/1377701/contributions/5875772/attachments/2837852/4959552/HEPIX_2024_cern_live_streaming_service.pdf

–

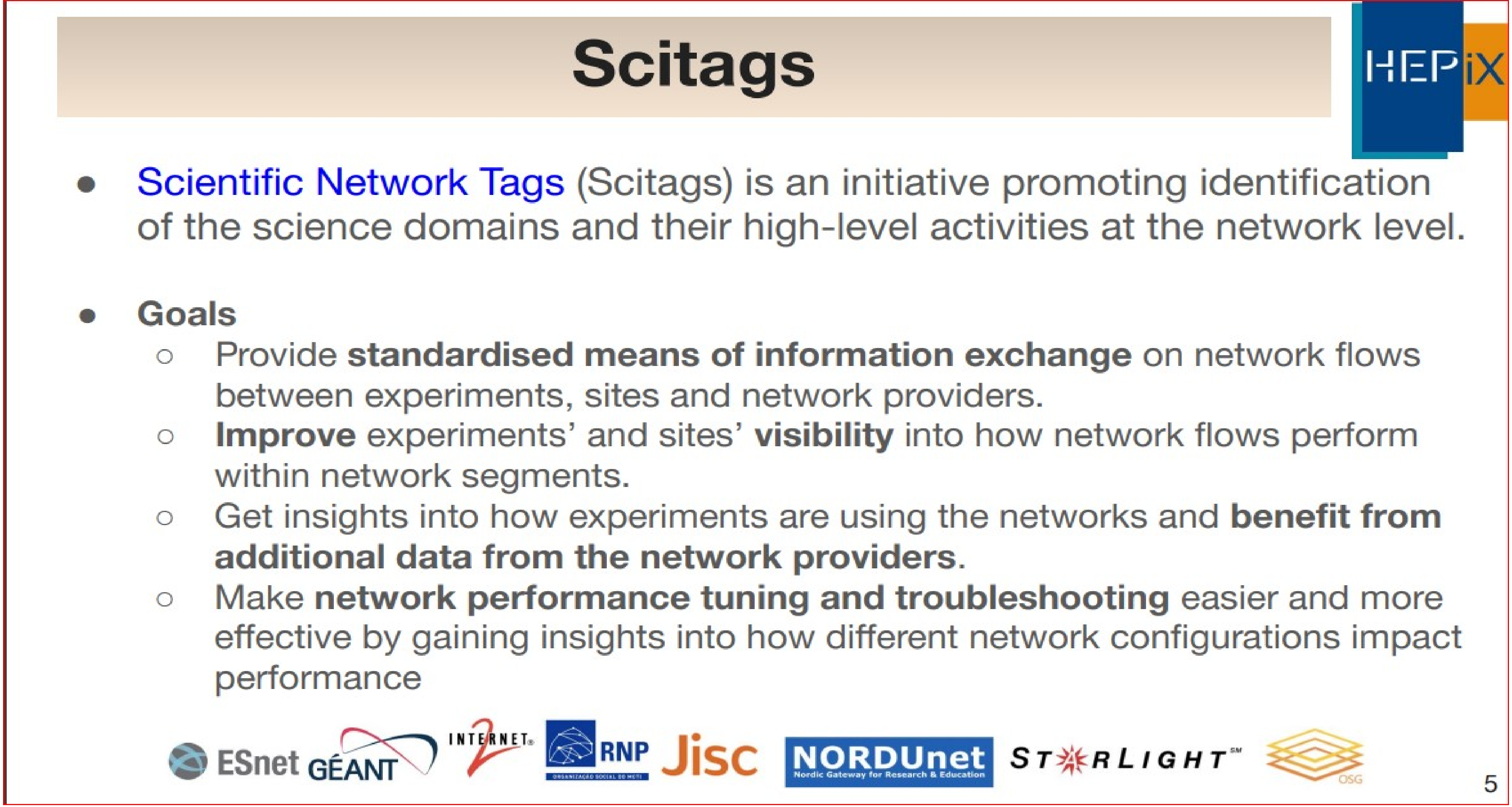
Very much attention

- Cluster engineering infrastructure (cooling)
- Cluster engineering infrastructure monitoring
- Saving electricity & water

Data transfer operations

WLCG Data Challenge 24 (DC24)

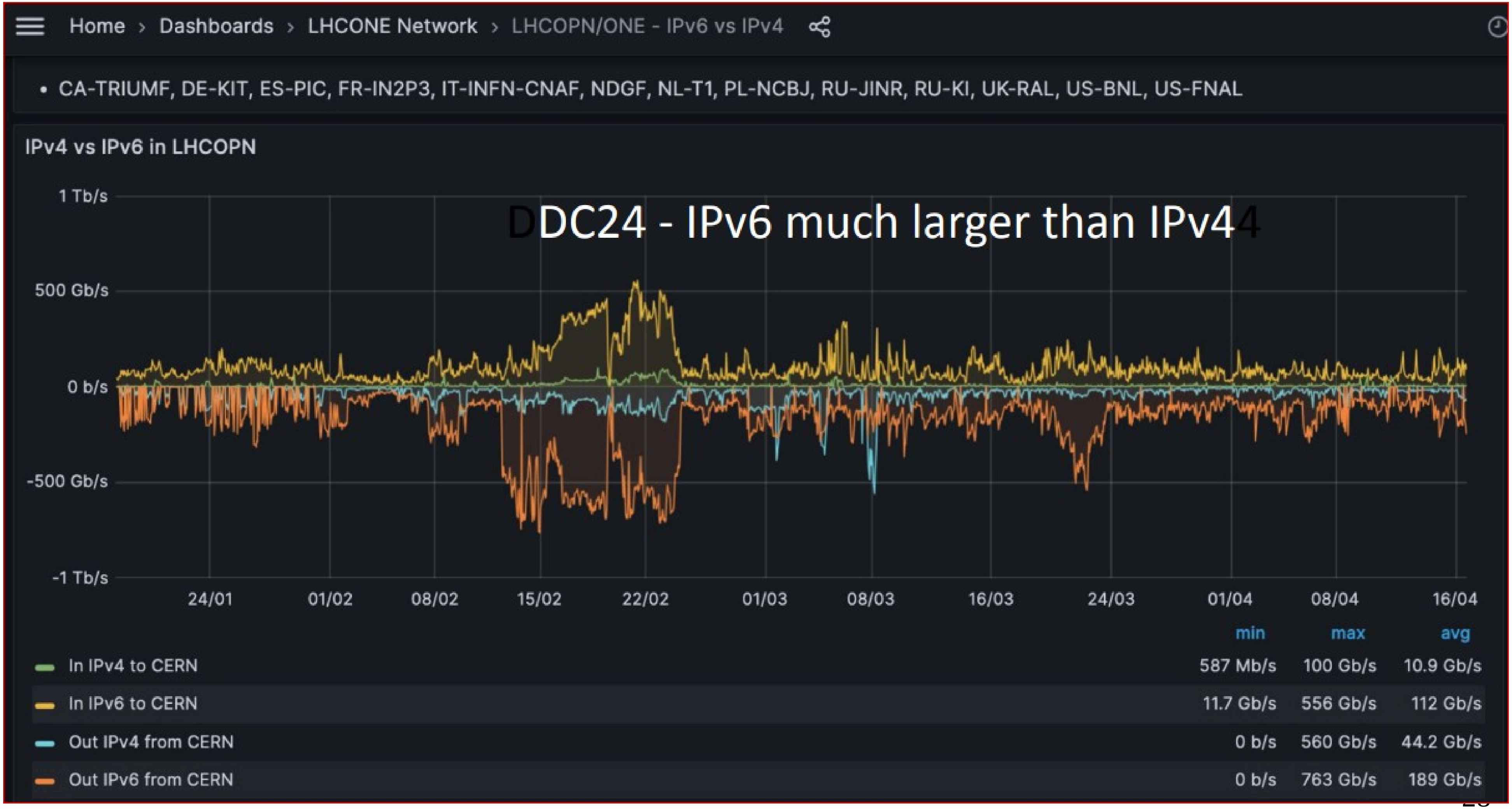
- Five experiments (ALICE, ATLAS, Belle II, CMS, and LHCb) participated in DC24. In general the center's infrastructure, including dCache and the network, successfully handled the load without bottlenecks.
- The WLCG Data Challenge 24 (DC24) was a significant event, with multiple sources mentioning it in the context of data transfer, networking, and infrastructure testing. For example during DC24, BNL reached its peak transfer target goals, exceeding 200 Gb/s in both inbound and outbound traffic.



Scitags

- Scientific Network Tags (Scitags) is an initiative promoting identification of the science domains and their high-level activities at the network level.

- Goals
  - Provide **standardised means of information exchange** on network flows between experiments, sites and network providers.
  - **Improve** experiments' and sites' **visibility** into how network flows perform within network segments.
  - Get insights into how experiments are using the networks and **benefit from additional data from the network providers**.
  - Make **network performance tuning and troubleshooting** easier and more effective by gaining insights into how different network configurations impact performance



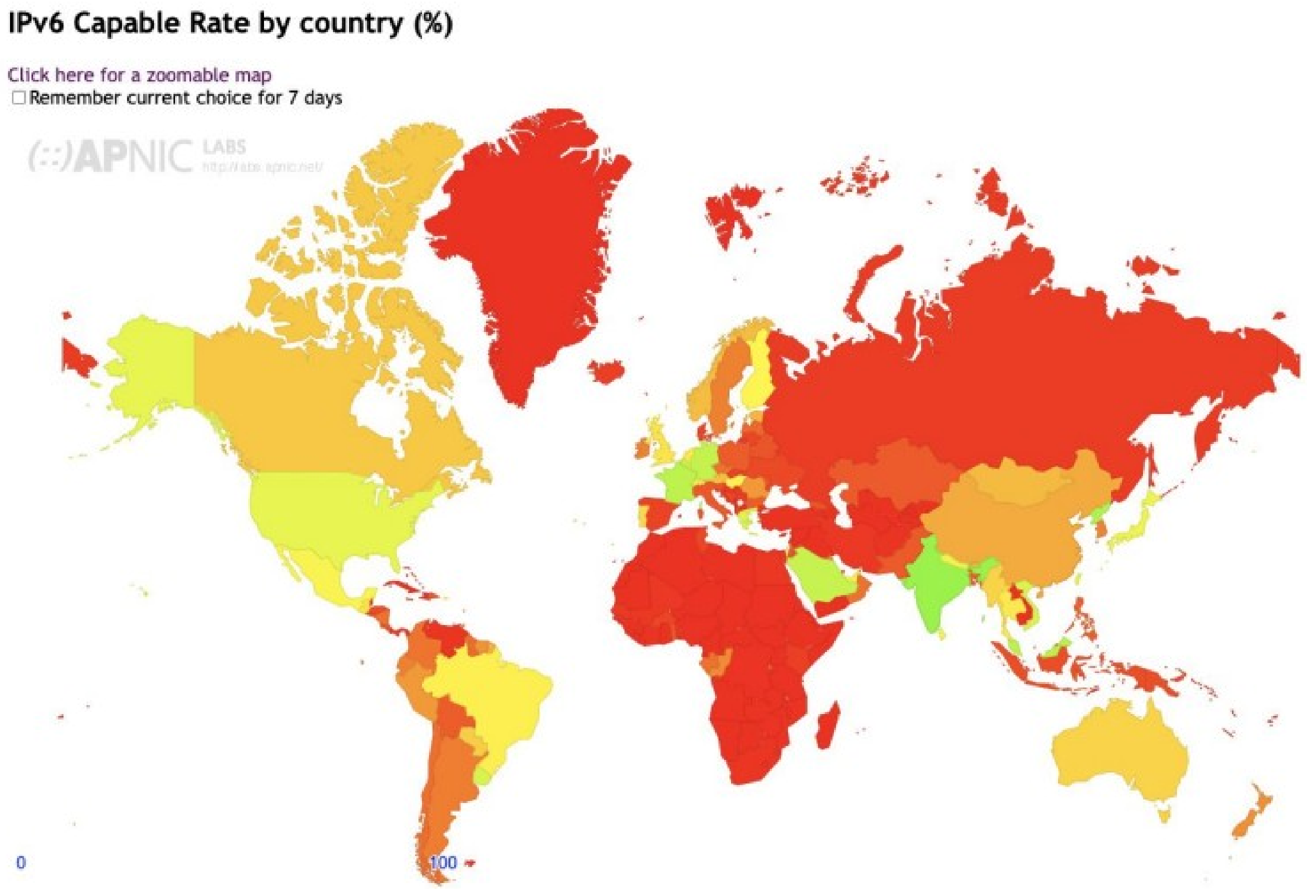
- CA-TRIUMF, DE-KIT, ES-PIC, FR-IN2P3, IT-INFN-CNAF, NDGF, NL-T1, PL-NCBJ, RU-JINR, RU-KI, UK-RAL, US-BNL, US-FNAL

IPv4 vs IPv6 in LHCOPN

DC24 - IPv6 much larger than IPv4

	min	max	avg
In IPv4 to CERN	587 Mb/s	100 Gb/s	10.9 Gb/s
In IPv6 to CERN	11.7 Gb/s	556 Gb/s	112 Gb/s
Out IPv4 from CERN	0 b/s	560 Gb/s	44.2 Gb/s
Out IPv6 from CERN	0 b/s	763 Gb/s	189 Gb/s

https://stats.labs.apnic.net/ipv6



IPv6 Capable Rate by country (%)

Click here for a zoomable map

□Remember current choice for 7 days

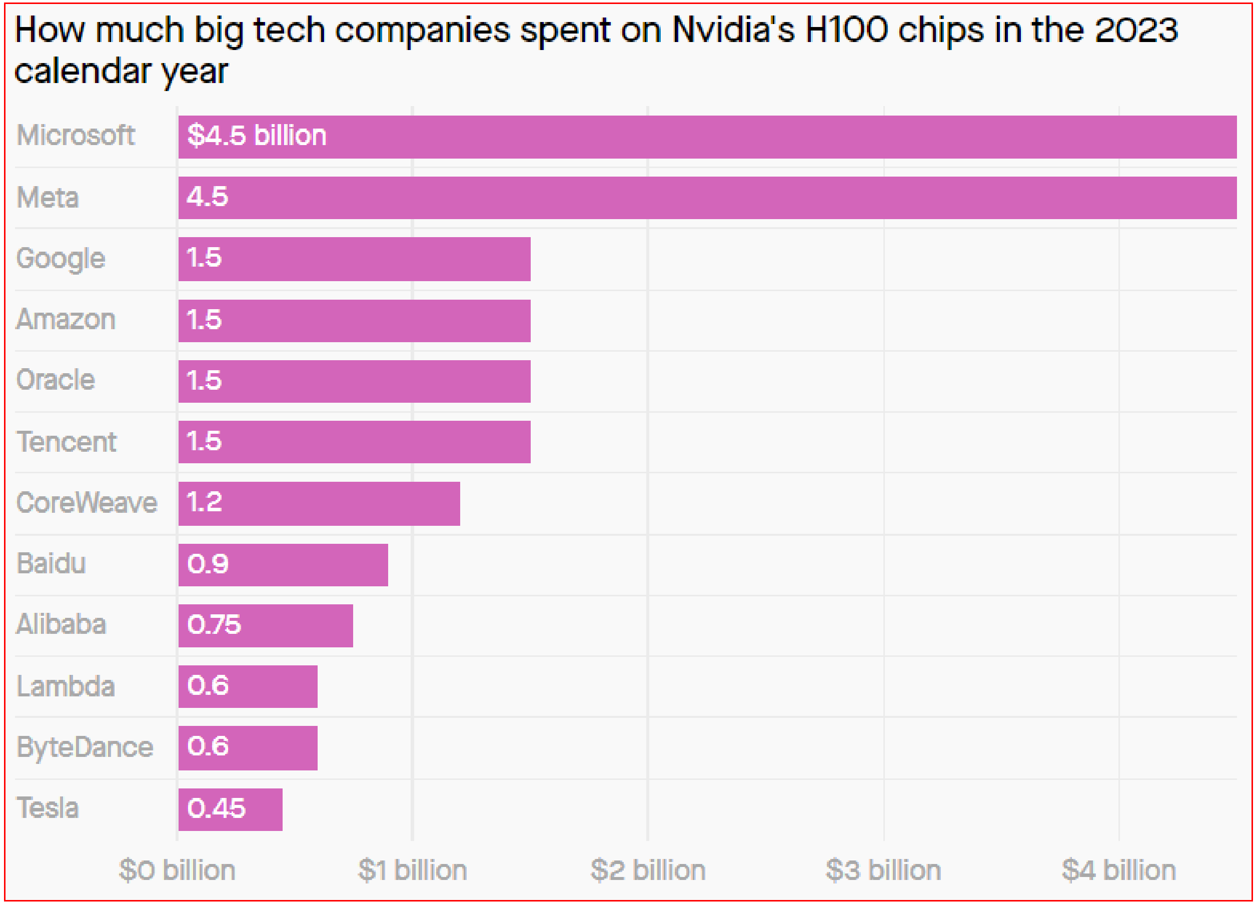
Science and Technology Facilities Council

Computer components future expectations

Trends in CPU

- Up to 128 cores today, 200+ announced
- Termal Design Power (TDP) to reach 1 kW/socket (liquid cooling)
  - Current last generation 1U systems need heatpipes with big radiators to cool
  - Expecting for next generation CPU’s that 1U systems will become rare and 2U or even bigger will become the standard
- GPU (Nvidia, AMD, Intel)
- AI Accelerators
  - Captive processors from Amazon (Trainium2),
  - Google (TPU v5) and Microsoft (Maia 100)
  - Intel Gaudi2, SambaNova SN40L, and others

AI accelerators



How much big tech companies spent on Nvidia's H100 chips in the 2023 calendar year

From https://qz.com/nvidia-generative-ai-google-microsoft-meta-1851206854

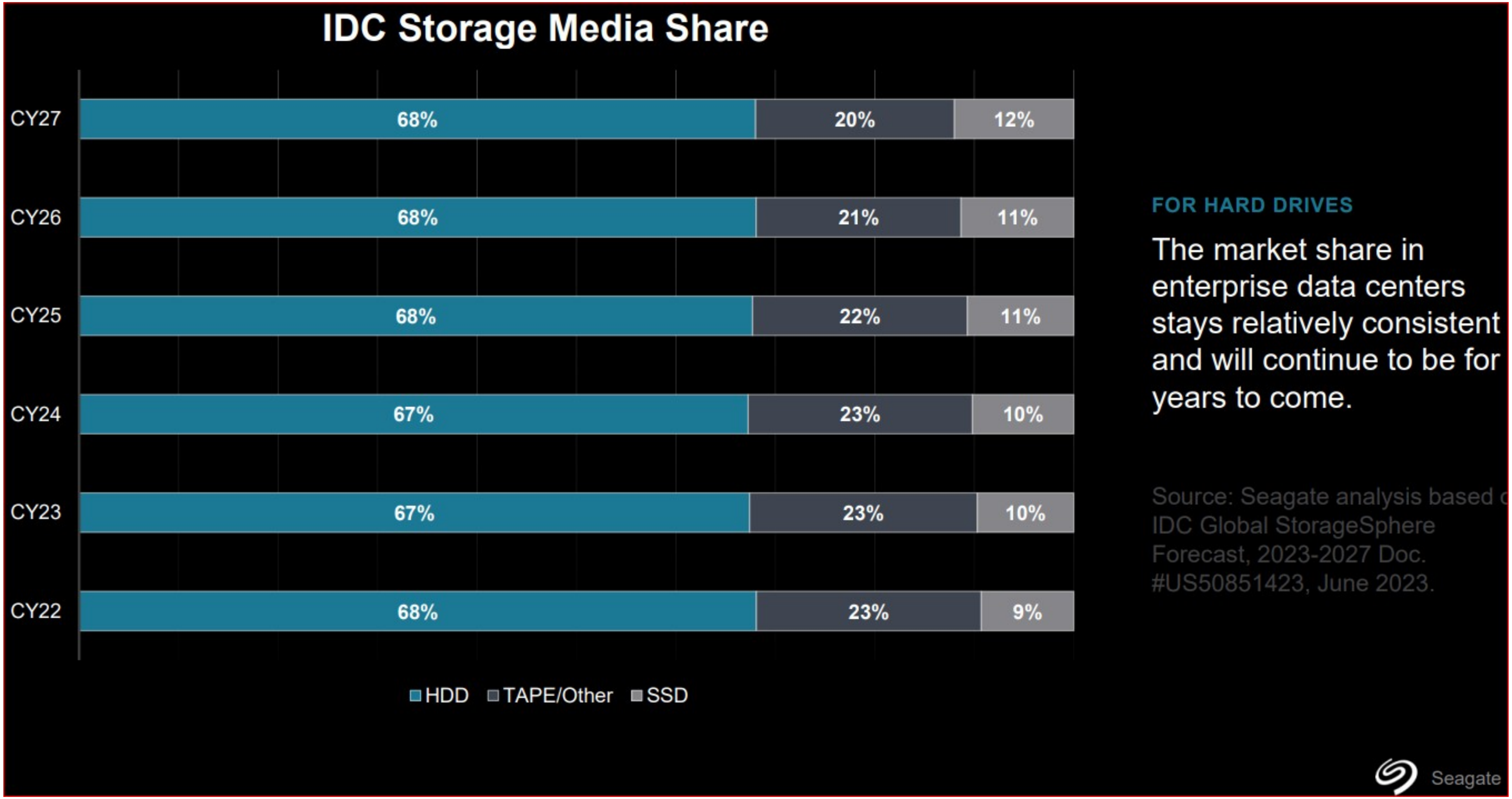
Memory

- DRAM memory
- CPU’s transitioning from DDR4 to DDR5 memory (up to 8400?)
  - DDR6 in 2026

Storage

- SSD account for ~12% of enterprise storage capacity
- PCIe Gen 5 SSDs now available
- Tapes: TS1170 - 50TB / cartridge

Storage future from Seagate



IDC Storage Media Share

FOR HARD DRIVES

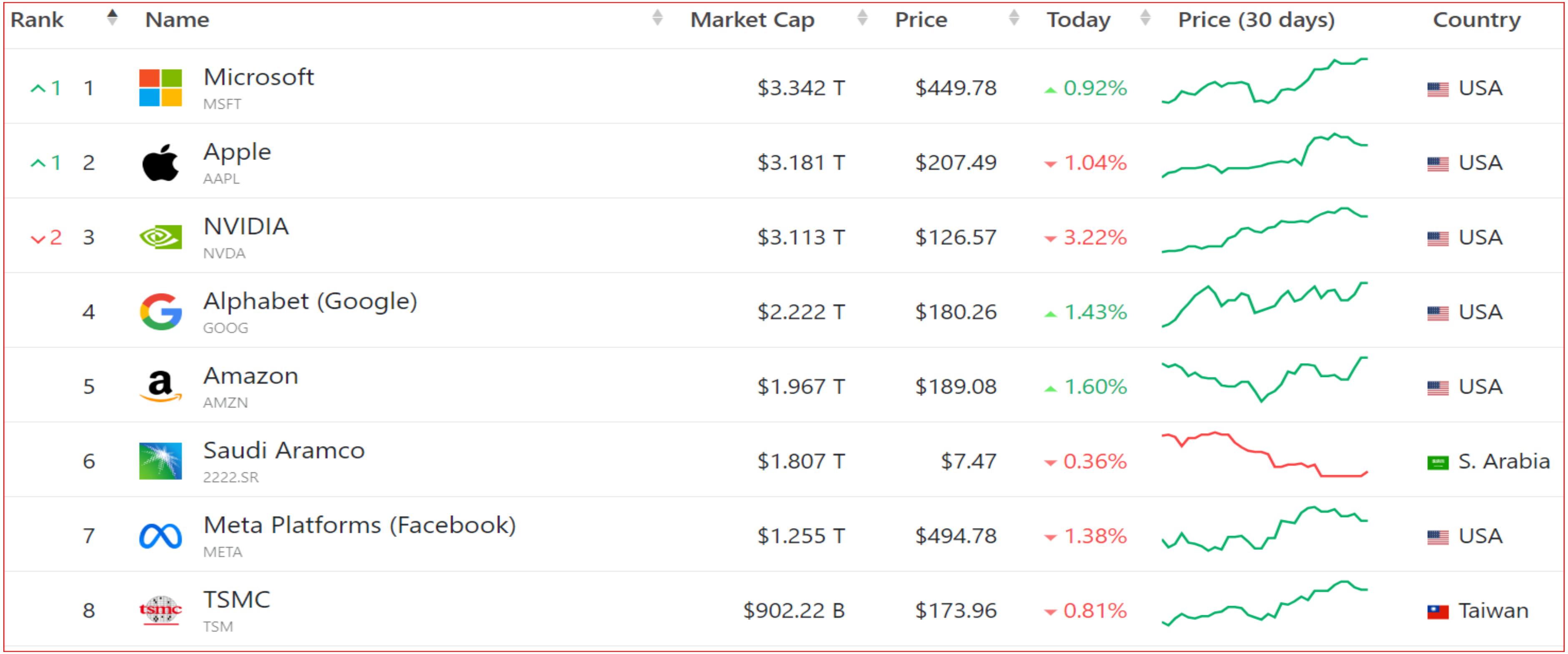
The market share in enterprise data centers stays relatively consistent and will continue to be for years to come.

Source: Seagate analysis based o
IDC Global StorageSphere
Forecast, 2023-2027 Doc.
#US50851423, June 2023.

Network

- Transition to 400GbE (4x100Gbs) in progress
- 800GbE (8x100Gbs) specification released in 2020
- Trends on WAN connectivity
  - LHC network traffic exponentially increasing, will need Tb/s links on major routes by 2029; Aggregate network traffic from ATLAS + CMS will be O(10 Tb/s)
  - Govs push moving to IPv6 soon (BNL, FNAL, CERN)
  - Developments
    - Better models and better automation
    - ML for system optimization

Крупнейшие компьютерные компании



Rank		Name	Market Cap	Price	Today	Price (30 days)	Country
⌃1	1	Microsoft MSFT	$3.342 T	$449.78	▲ 0.92%		USA
⌃1	2	Apple AAPL	$3.181 T	$207.49	▼ 1.04%		USA
⌄2	3	NVIDIA NVDA	$3.113 T	$126.57	▼ 3.22%		USA
	4	Alphabet (Google) GOOG	$2.222 T	$180.26	▲ 1.43%		USA
	5	Amazon AMZN	$1.967 T	$189.08	▲ 1.60%		USA
	6	Saudi Aramco 2222.SR	$1.807 T	$7.47	▼ 0.36%		S. Arabia
	7	Meta Platforms (Facebook) META	$1.255 T	$494.78	▼ 1.38%		USA
	8	TSMC TSM	$902.22 B	$173.96	▼ 0.81%		Taiwan

From https://companiesmarketcap.com/

Рассуждения

На что обратил внимание

- Наравне с большими кластерами продолжена линия микро-кластеров для интерактивной работы.
- Мало внимания уделено технологии машинного обучения, хотя на CHEP-2023 (май 2023) было около 10% выступлений целиком или косвенно упоминали эту тему.

Некоторые соображения

- Следует регулярно посещать конференцию типа HEPIX, чтобы быть в курсе событий.
- Для поддержания формы следует регулярно обновлять компьютерную инфраструктуру Отделения.
  - Старые работающие серверы мы готовы отдавать любым сотрудникам ОФВЭ для дальнейшего использования.

Рекомендации

- Серверную инфраструктуру ОФВЭ следует обновлять регулярно. Микро-кластеры (несколько хостов) для отладок и освоения имеющихся инструментов, например Jupyter Notebook, REANA, CVMFS, машинное обучение.

Заключение

- Сайт с презентациями весенней сессии HEPIX-2024
  - https://indico.cern.ch/event/1377701/

Spare references on networking

- NOTED https://indico.cern.ch/event/898285/contributions/4039620/attachments/2121618/3571075/NOTED___Hepix_Joanna_Waczynska.pdf
- Global Network Advancement Group https://www.gna-g.net/
- Shawn McKee / University of Michigan // Research Networking Technical Working Group Status and Plans
  - Scientific Network Tags (Scitags) is an initiative promoting identification
  - of the science domains and their high-level activities at the network level.
  - https://www.scitags.org/assets/img/chep_paper23.pdf
- REANA - https://www.epj-conferences.org/articles/epjconf/pdf/2019/19/epjconf_chep2018_06034.pdf