Present and Future of GPU HPC in Hungary

National HPC development

Zoltan Kiss, Head of HPC Dpt.
KIFÜ

HPC @hu
Kompetencia Központ

slido

Join at slido.com
#2522236

ⓘ Start presenting to display the joining instructions on this slide.

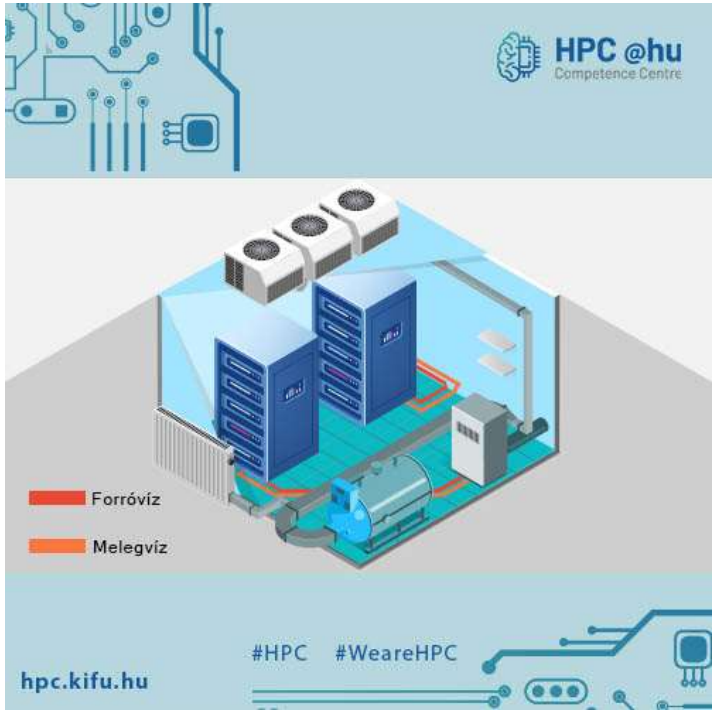
slido

What ICT Infra are you missing the most?

ⓘ Start presenting to display the poll results on this slide.

Contents

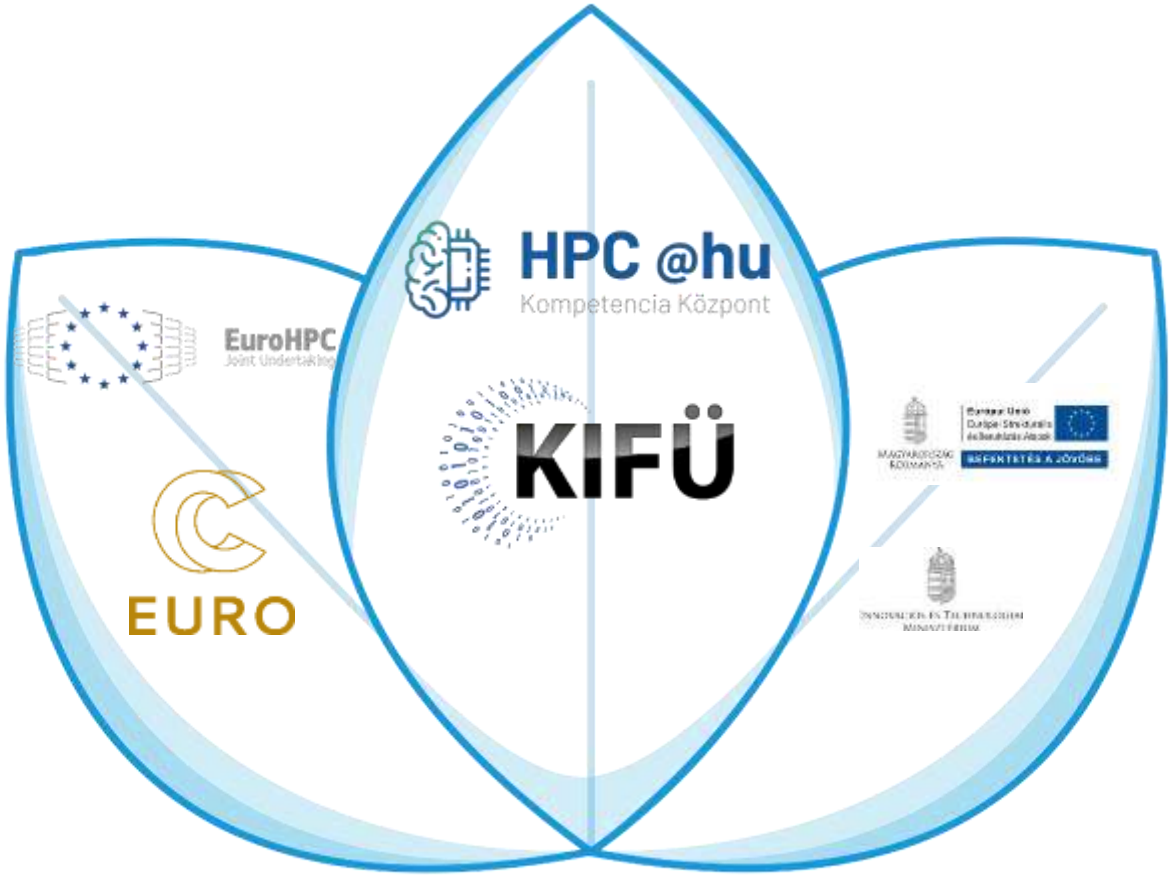
- HPC Competence Centre
- IT development
- HPC infrastructure
- Portals
- EuroHPC
- Trends and Input



HPC @hu
Competence Centre

HPC ecosystem

Competence Centre and SME relations

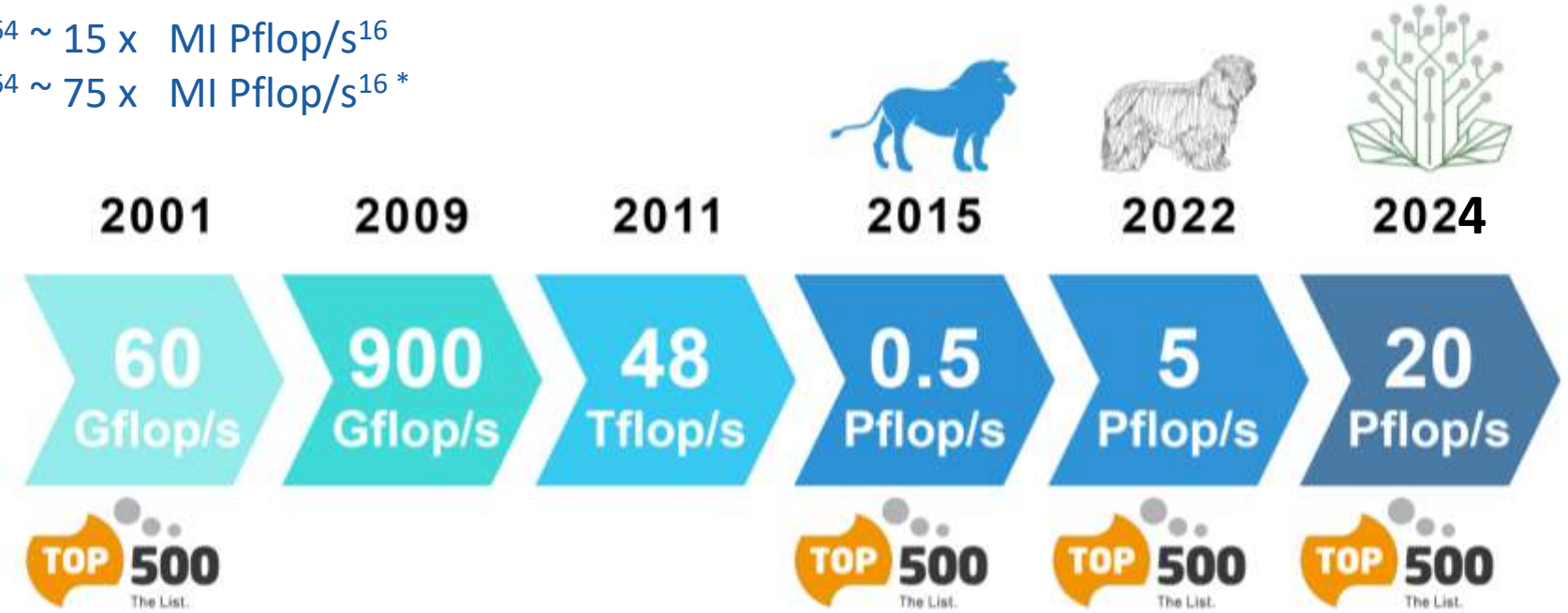


Infrastructure development

HPC @hu
Competence Centre

20 years of Hungarian HPC

1 HPC Pflop/s^{64} ~ 15 x MI Pflop/s^{16}
5 HPC Pflop/s^{64} ~ 75 x MI Pflop/s^{16} *

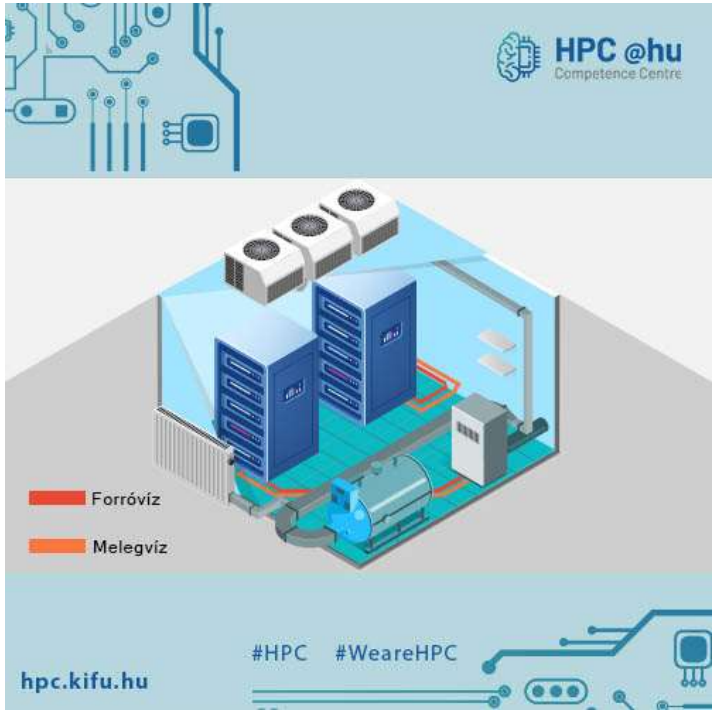


* A100 values

HPC @hu Competence Centre

Infrastructure development

- Hardware
- Software



Green DC

- Hot water used
- 90% air cooled

2022 Q3 – 5+ PF HPC

- Efficiency
- Extreme density
- Award winning building designer
- In Debrecen Campus

Komondor

HPC @hu
Competence Centre

CRAY EX EXASCALE INFARSTRUCTURE

CRAY
THE SUPERCOMPUTER COMPANY

OAK RIDGE National Laboratory

1.6 ExaFlops (US)

Lawrence Livermore National Laboratory

2 ExaFlops (US)

CSC LUMI

550 Pflops (EU)

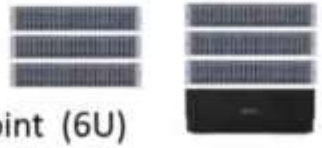
Argonne NATIONAL LABORATORY

180 Pflops (US)

COMPUTING

70 Pflops (US) ~ Komondor specs

Hewlett Packard Enterprise



All Flash entry point (6U)
Up to **80/50 GB/sec** read/write and **115 TB** capacity*

Disk entry point (10U):
15 GB/sec and
315 TB capacity*

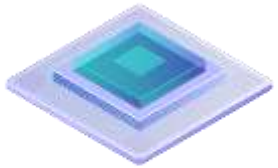
All Flash base rack: > 1TB/sec and up to 4.5 PB capacity*
- Expansion rack: > 2 TB/sec and 4.6 PB*

Disk base rack: 90 GB/sec and 7.5 PB capacity*
- Expansion rack: 120 GB/sec and 10 PB*

*usable capacity

komondor

20 000+
CPU mag



CPU 0,7+
petaflops

CPU

200+
GPU

GPU 4+
petaflops

GPU

Big Data 9+ TB
memory

BIG DATA

8GPU/node

MI

10 PB

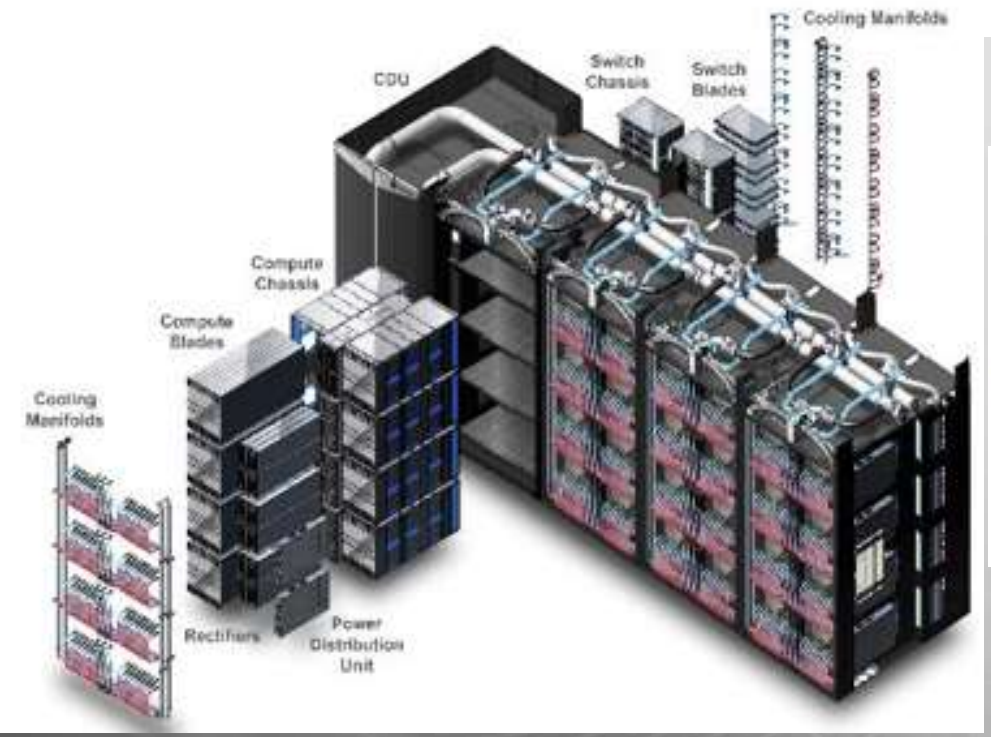
Tape archive

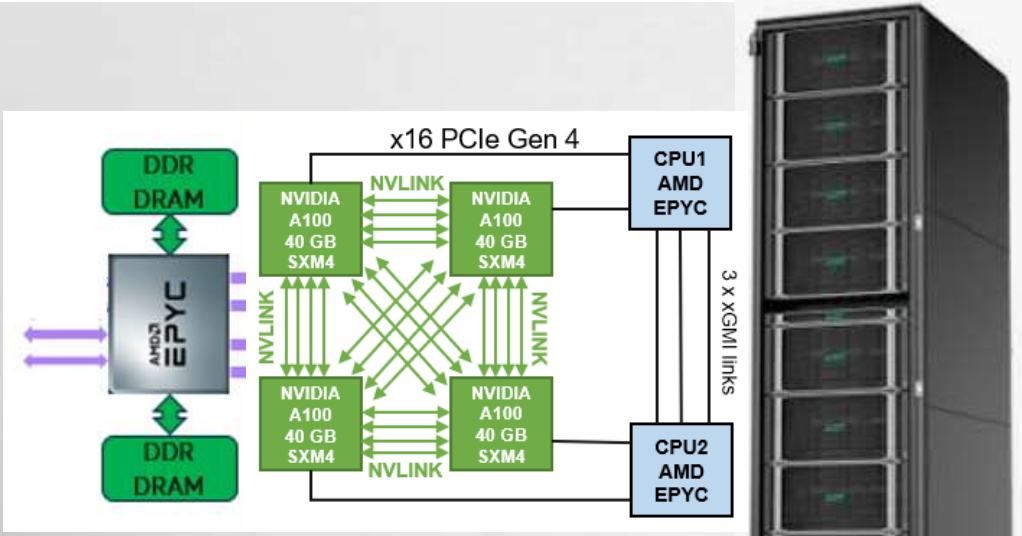
2 PB

Ultrafast

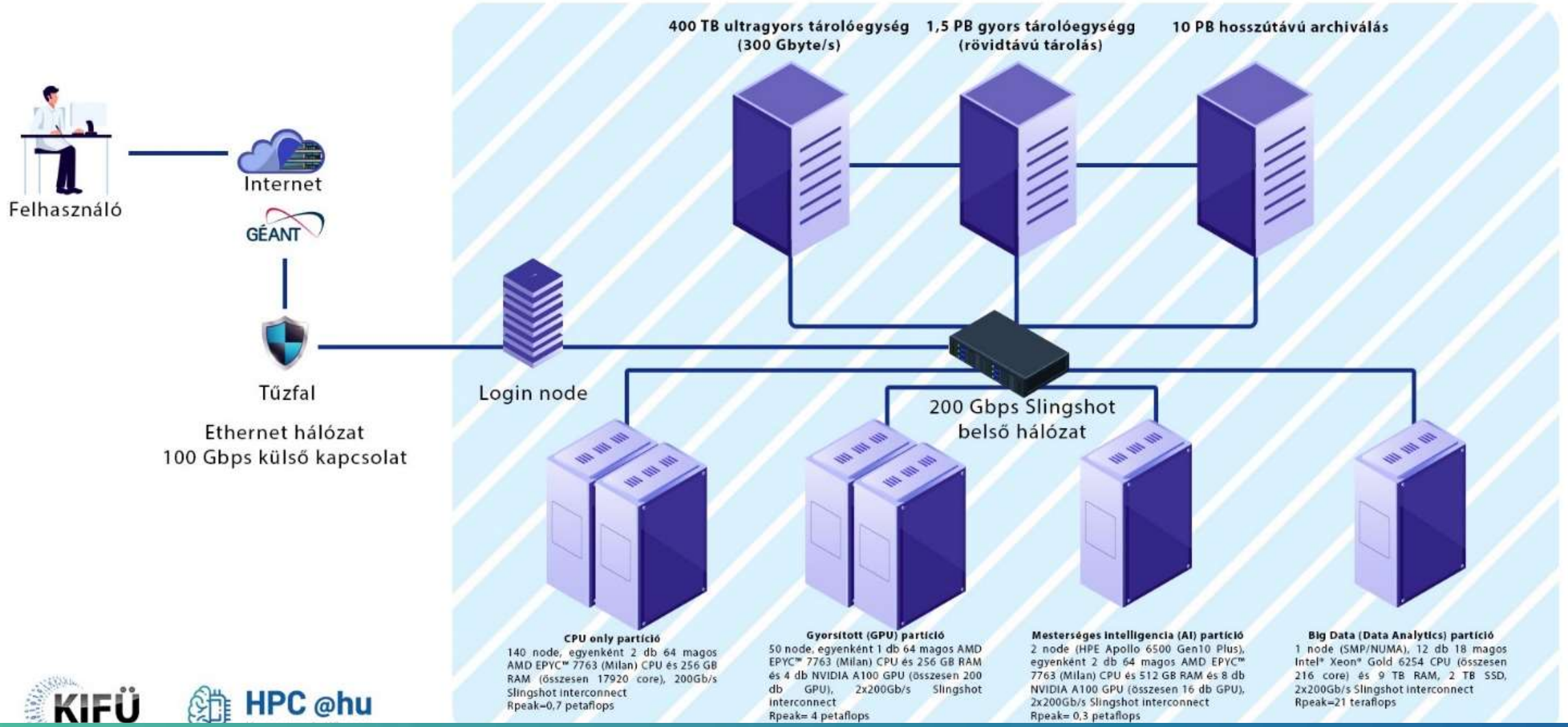
STORAGE


Hewlett Packard Enterprise
komondor
HPC @hu
KIFÜ


Cooling Manifolds
CDU
Switch Chassis
Switch Blades
Compute Chassis
Compute Blades
Cooling Manifolds
Rectifiers
Power Distribution Unit


x16 PCIe Gen 4
DDR DRAM
AMD EPYC
DDR DRAM
NVIDIA A100 40 GB SXM4
NVIDIA A100 40 GB SXM4
NVIDIA A100 40 GB SXM4
NVIDIA A100 40 GB SXM4
NVLINK
NVLINK
NVLINK
NVLINK
CPU1 AMD EPYC
CPU2 AMD EPYC
3 x xGMI links

KIFÜ



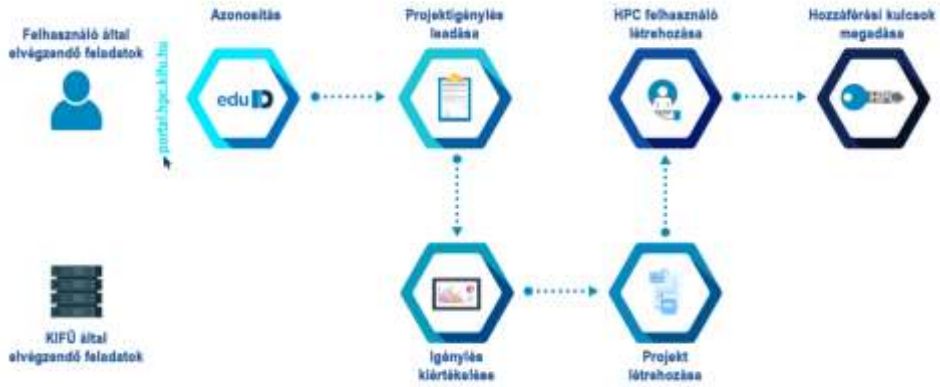
KIFÜ HPC @hu

HPC @hu
Competence Centre

Software

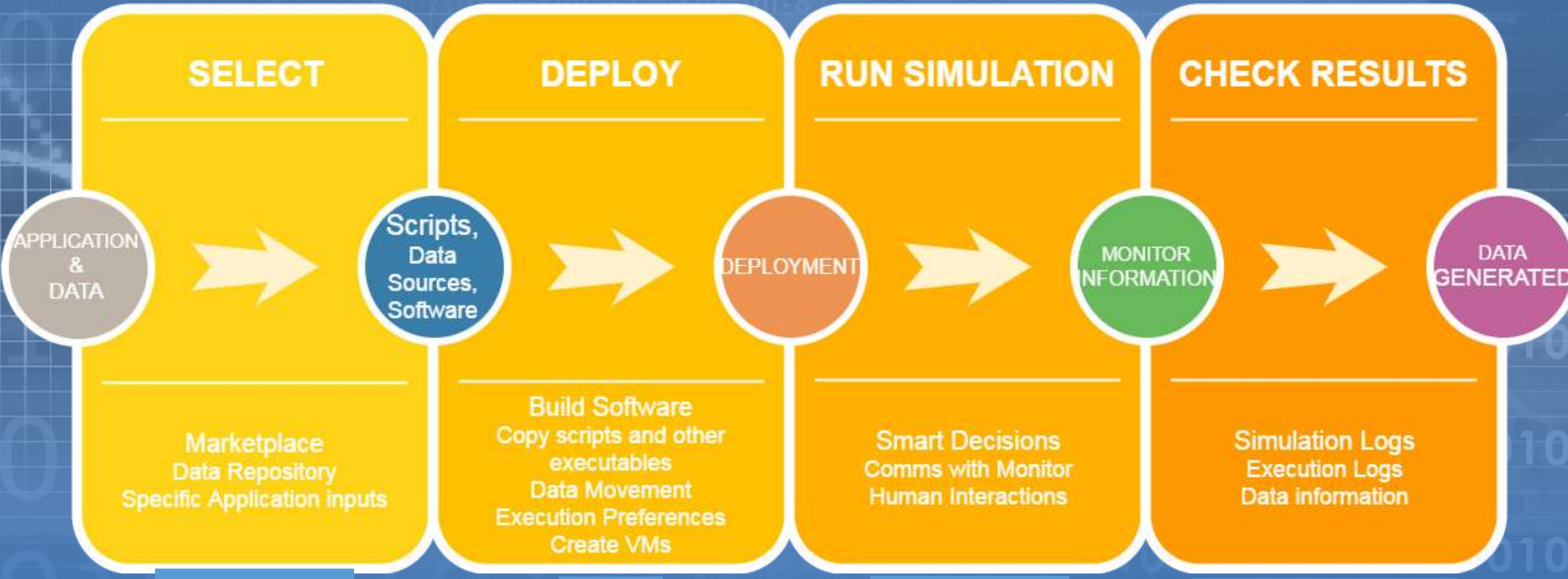
- Ansys
- Amber20
- Amsterdam Density Functional (ADF)
- CharMM
- Gaussian 16
- Terachem
- Schrödinger
- TotalView

HPC portal



HPC @hu
Competence Centre

Deploy workflow - example



CKAN

CI

Containers

Visualization

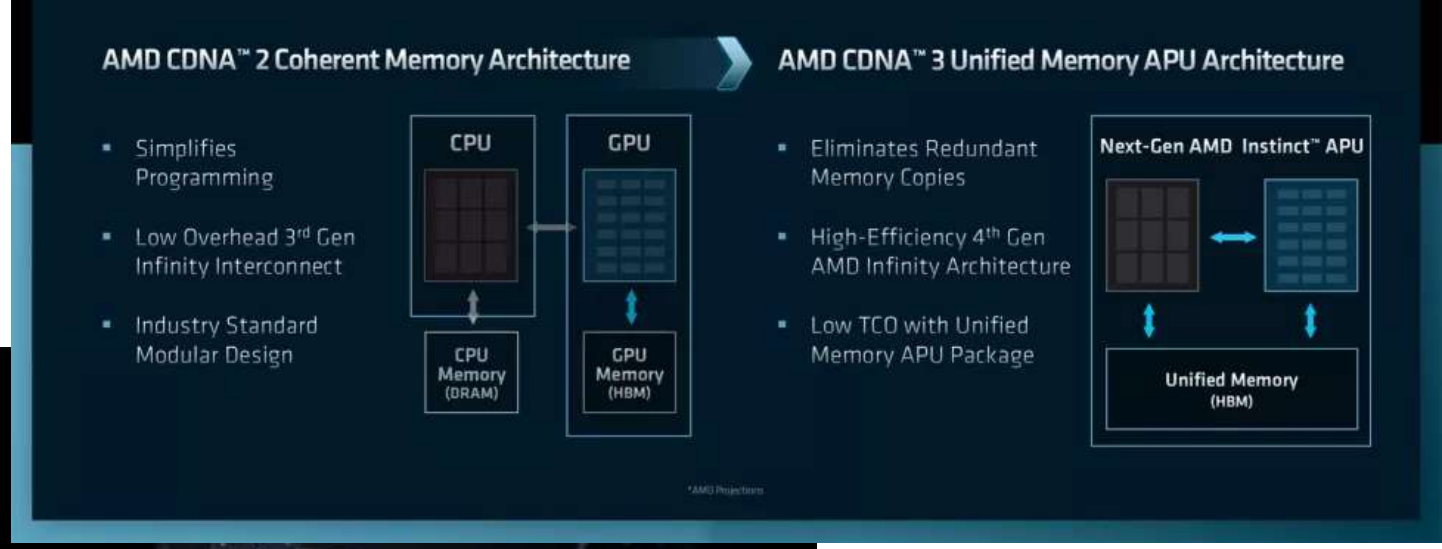
KIFÜ

EuroHPC

- 20 PF
- CPU + **GPU** + Big Data + MI
- Multilevel storage system
- Visualization subsystem
- Modular architecture
- Quantum HPC integration
- HU-DE (Juelich – TIM – KIFÜ) cooperation
- Portal and software developments

HPC @hu
Competence Centre

What's in the oven?



AMD INSTINCT™ MI300

THE WORLD'S FIRST DATA CENTER APU

- 4th Gen AMD Infinity Architecture: AMD CDNA™ 3 and EPYC™ CPU "Zen 4" Together
  CPU and GPU cores share a unified on-package pool of memory
- Groundbreaking 3D Packaging
  CPU | GPU | Cache | HBM
- Designed for Leadership Memory Bandwidth and Application Latency
- APU Architecture Designed for Power Savings Compared to Discrete Implementation

Available **2023**

>8X

Expected AI Training Performance vs. MI250X

See Endnote MI300-03. Preliminary data and projections, subject to change

HPC @hu
Competence Centre

What's in the oven?

Nvidia Datacenter GPUs		
Data by Nvidia, H100 specifications not finalized		
GPU Performance	**NVIDIA A100**	**NVIDIA H100 SXM5**
Peak FP8 Tensor TFLOPS with FP16 Accumulate	NA	2000/4000
Peak FP8 Tensor TFLOPS with FP32 Accumulate	NA	2000/4000
Peak FP16 Tensor TFLOPS with FP16 Accumulate	312/624	1000/2000
Peak FP16 Tensor TFLOPS with FP32 Accumulate	312/624	1000/2000
Peak BF16 Tensor TFLOPS with FP32 Accumulate	312/624	1000/2000
Peak TF32 Tensor TFLOPS	156/312	500/1000
Peak FP64 Tensor TFLOPS	19.5	60
Peak INT8 Tensor TOPS	624/1248	2000/4000
Peak FP16 TFLOPS (non-Tensor)	78	120
Peak BF16 TFLOPS (non-Tensor)	39	120
Peak FP32 TFLOPS (non-Tensor)	19.5	60
Peak FP64 TFLOPS (non-Tensor)	9.7	30
Peak INT32 TOPS	19.5	30

slido

What chip are you using the most now?

ⓘ Start presenting to display the poll results on this slide.

slido

What chip would you see the most potential for your use case in the future? (multiple choice)

ⓘ Start presenting to display the poll results on this slide.

Thank you for your feedback!

hpc.kifu.hu
kiss.zoltan@kifu.gov.hu
@HPC.CC.hu
@HPC_hu

HPC @hu
Kompetencia Központ

slido

Audience Q&A Session

ⓘ Start presenting to display the audience questions on this slide.