2. odprti dan za gospodarstvo

Računski viri za podjetja danes in kaj bo prinesel novi superračunalnik

Dejan Valh
IZUM, Institut informacijskih znanosti

10. April 2026

O IZUM-u

- Javni infrastrukturni in raziskovalni zavod
- 129 zaposlenih (10 zaposlenih HPC/AI)
- Knjižnični informacijski sistem
- Informacijski sistem o raziskovalni dejavnosti
- UNESCO regionalni center
- Operacija HPC RIVR
- Konzorcij SLING
- Najnovejše:
  - Koordinator SLAIF
  - Upravljanje novega HPC/AI (Vega 2)

IZUM
Institut informacijskih znanosti, Maribor
Regionalni center za knjižnične informacijske sisteme in informacijske sisteme o raziskovalni dejavnosti pod okriljem UNESCO

unesco

COBISS.net

COBISS

SLING
Slovensko nacionalno superračunalniško omrežje

SICRIS

HPC RIVR

University of Maribor

FIS
Fakulteta za informacijske študije
Faculty of Information Studies

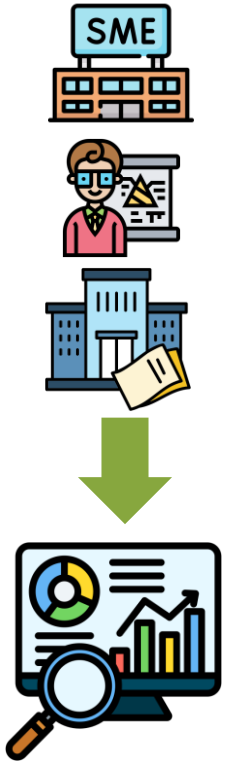
arnes

Jožef Stefan Institute

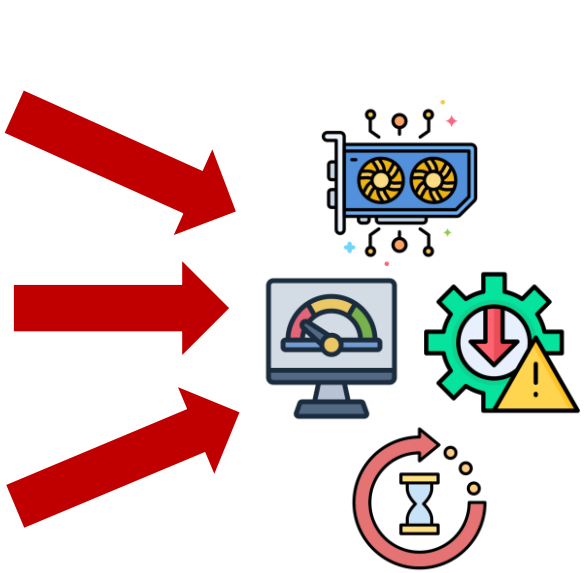
SLAIF
Slovenska tovarna UI

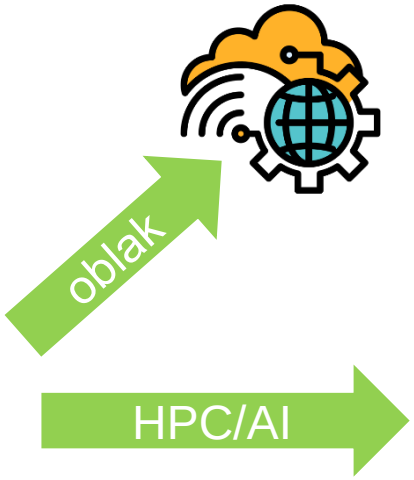
VEGA

Superračunalnik za umetno inteligenco?



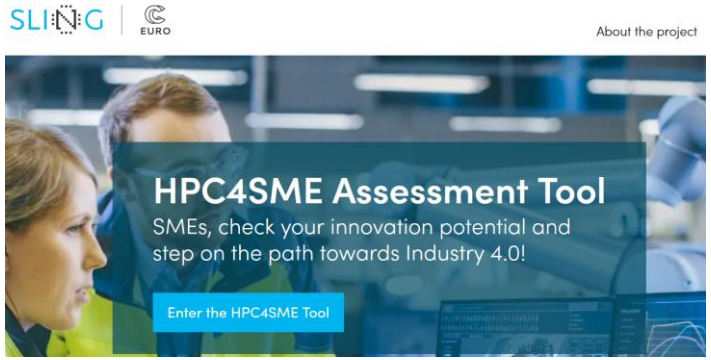
ANALIZA POTREB





SUPERRAČUNALNIK

Namenjen, orientiran, optimiziran za umetno inteligenco



https://www.hpc4sme.eu/

Programi financiranja in javni razpisi za industrijo

https://www.sling.si/programi-financiranja-in-javni-razpisi-za-industrijo/

Ikone: Flaticon.com

Kaj lahko naredijo SME-ji?

(potem ko ugotovijo, da bi potrebovali superračunalnik ali vsaj oblak za umetno inteligenco)

- Kadrovski potenciali:
  - Imate raziskovalce, sodelujete z akademskim svetom?
  - Imate kader IT (razvijalce, sistemske inženirje)?
  - Imate informatiko „outsourcano“?
- Programska oprema za posel/ind. procese/UI:
  - Razvijate lastno programsko opremo? Jo prodajate?
  - Uporabljate kupljene rešitve za LOB?
  - Uporabljate programske storitve v oblaku?
- Infrastruktura IT:
  - Imate lastno „on-premise“ infrastrukturo, svoje strežnike?
  - Imate hibridno infrastrukturo?
  - Uporabljate izključno infrastrukturo v oblaku?

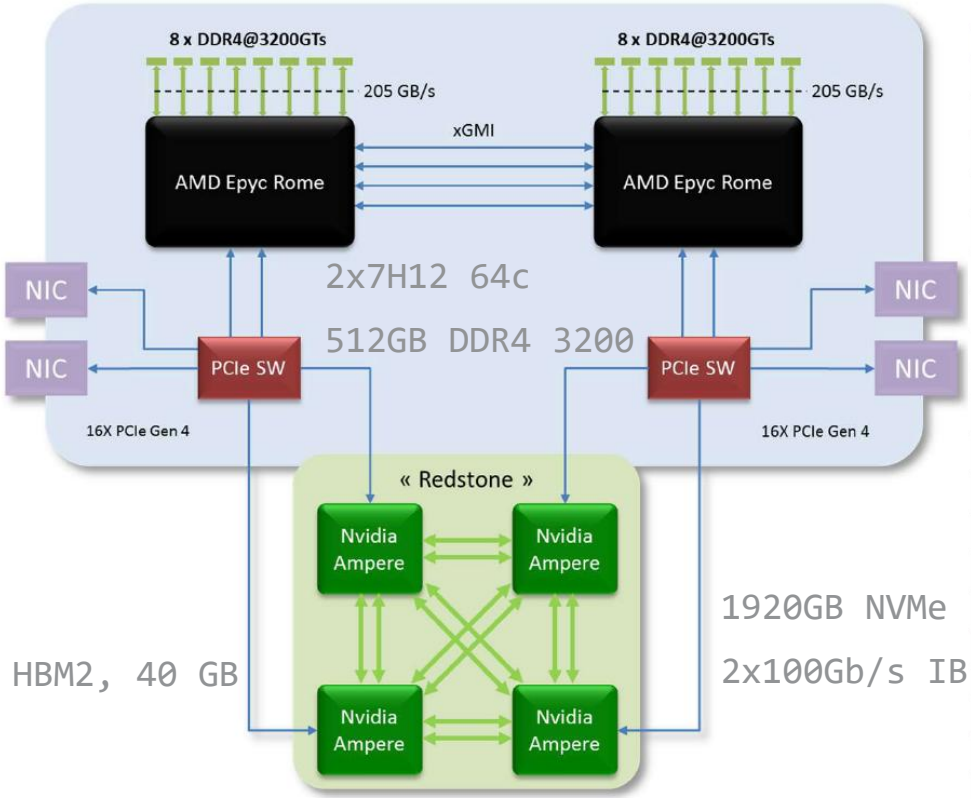
Pridobite kompetence in izkoristite testne, razvojne dostope do superračunalniških in oblačnih virov

Najdite partnerja (SLAIF!) ali pa preverite možnosti pri vašem ponudniku storitev IT

Investirajte v uporabo specifične programske opreme

Preučite možnosti uporabe superračunalniške infrastrukture za posamezne projekte UI

HPC Vega – particija GPU



Nvidia Datacenter GPU	Nvidia A100
GPU codename	GA100
GPU architecture	Ampere
Launch date	May 2020
GPU process	TSMC 7nm
Die size	826mm2
Transistor Count	54 billion
FP64 CUDA cores	3,456
FP32 CUDA cores	6,912
Tensor Cores	432
Streaming Multiprocessors	108
Peak FP64	9.7 teraflops
Peak FP64 Tensor Core	19.5 teraflops
Peak FP32	19.5 teraflops
Peak FP32 Tensor Core	156 teraflops/312 teraflops*
Peak BFLOAT16 Tensor Core	312 teraflops/624 teraflops*
Peak FP16 Tensor Core	312 teraflops/624 teraflops*
Peak INT8 Tensor Core	624 teraflops/1,248 TOPS*
Peak INT4 Tensor Core	1,248 TOPS/2,496 TOPS*
Mixed-precision Tensor Core	312 teraflops/624 teraflops*
Max TDP	400 watts

*Effective TOPS / TFLOPS using the new Sparsity feature

50 pospeševalnikov GPU za potencialno rabo podjetij

60 BullSequana XH2415 vozlišč = 240 NVIDIA A100

Načini dostopa do HPC Vega, Leonardo

Odprti dostop SLING:
- samostojno – nižji TRL-ji, izzivi pri pripravi dokumentacije za prijavo ALI
- sodelovanje z univerzami, inštituti, dokumentacijo priprav raziskovalci
- testiranje/razvoj/običajni/veliki
- osnovna podpora
- ni za komercialne storitve

Pilotni dostop SLING:
- **SAMO ZA GOSPODARSTVO**
- za hitro testiranje
- 2000 CPU ali 400 GPU vozliščnih ur
- hitra, enostavna prijava
- ni poročanja
- SLING + SLAIF podpora
- ni za komercialne storitve

SLING Slovensko nacionalno superračunalniško omrežje

VEGA

SLOVENSKI DELEŽ

EuroHPC JU DELEŽ

LEONARDO CINECA

Vir: https://leonardo-supercomputer.cineca.eu/hpc-system/

Odprti dostopi EuroHPC JU:
- za testiranje (Benchmark)
- za razvoj (Development)
- za raziskave (Regular):
  - obsežna prijava, poročilo
  - znanstvena recenzija
- podpora osnovna + ekspertna EPICURE
- ni za komercialne storitve

EPICURE Unlocking European-level HPC Support

EuroHPC JU Playground Access:
- dostop za tovarne, tudi SLAIF
- za testiranje, samo za podjetja
- ni znanstvene recenzije
- enostavna prijava
- dostop v dveh dneh!
- ni za komercialne storitve

AI Factories

EuroHPC Joint Undertaking

SLING pilotni dostop za podjetja

Dostop te vrste je namenjen izključno podjetjem za pilotna testiranja, z računskimi kapacitetami v obsegu do 2000 zliščnih ur.

https://www.sling.si/zaposleni-v-gospodarstvu-podjetniki/

SLING | EURO

SLING Podpora Uporabniki Izobraževanja in dogodki Repozitorij

Kratki cilji in spletna stran projekta *

Na kratko predstavite projekt (spletna stran je dobrodošla). Navedite razloge za uporabo gruče, na primer: preizkušanje programske opreme in algoritmov, optimizacija obdelav, izvedba analiz, vizualizacija, promocija, izobraževanje. Opišite cilje, ki jih boste zasledovali v okviru dela na gruči.

Limit is 100 words. Words remaining: 100.

Raziskovalna metoda, algoritmi ter programska oprema *

Opišite tipično obdelavo, ki jo boste izvajali v okviru raziskovalne metode, navedite ključne algoritme in programsko opremo, zaradi katerih potrebujete računsko gručo. Navedite dosedanje izkušnje z računanjem na gručah.

Limit is 100 words. Words remaining: 100.

Izbira particije *

- [x] Vega – CPU particija
- [x] Vega – GPU particija

You have exceeded the number of allowed selections: 1.

PILOTNI DOSTOP ZA GOSPODARSTVO

https://www.sling.si/prijava-pilotni/

Vega – CPU particija – želena količina virov v vozliščnih urah (do 2000 vozliščnih ur)

Vega – GPU particija – želena količina virov v vozliščnih urah (do 400 vozliščnih ur)

Računski viri in infrastrukturne potrebe *

Opišite vaše potrebe po računskih virih, podatkovnih shrambah in programski opremi, vključno z uporabljanimi knjižnicami. Jasno navedite posebne zahteve, na primer za vizualizacijo podatkov, interaktivno uporabo, dostop do zunanjih virov.

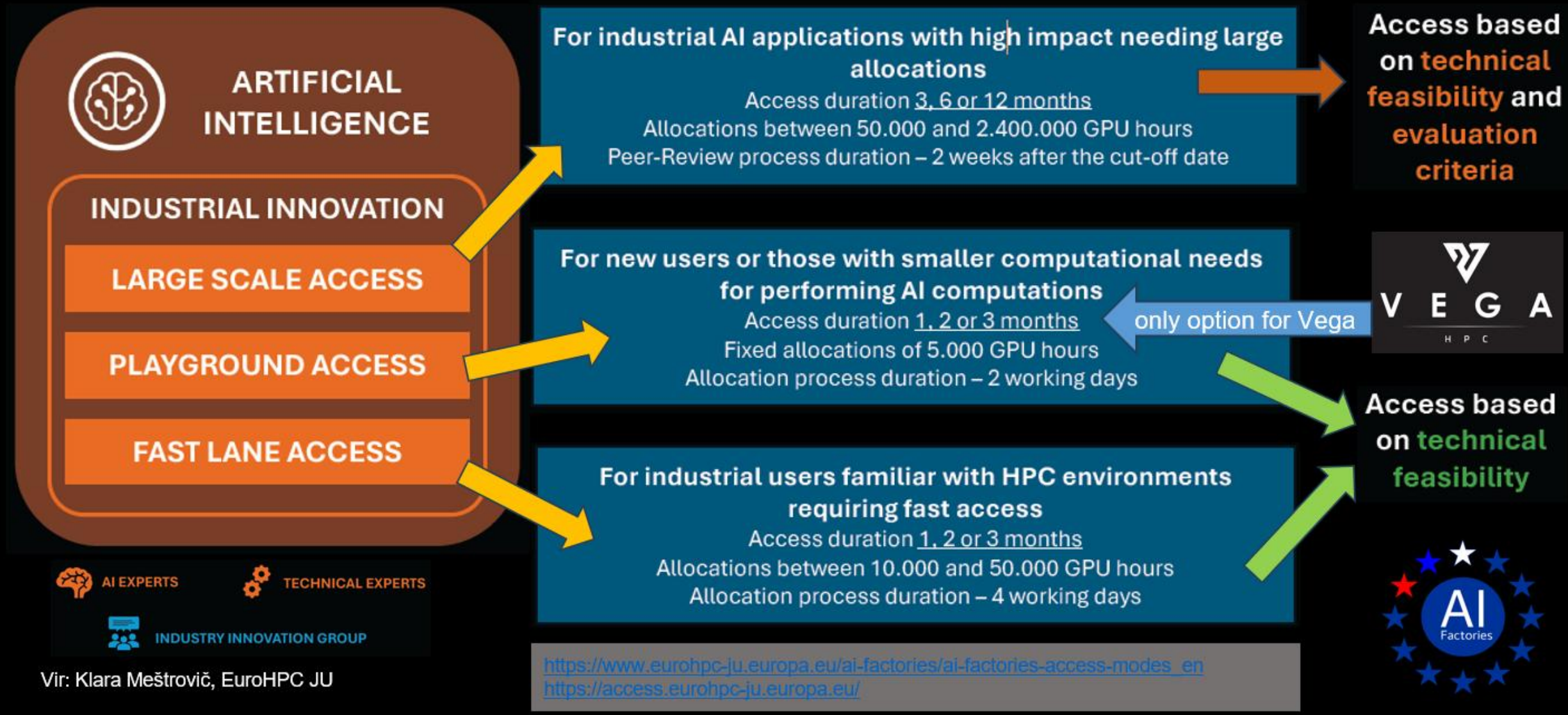
Limit is 100 words. Words remaining: 100.

Potrebe po tehnični pomoči in izobraževanju

Ste seznanjeni z delovanjem superračunalnika, poznate glavne elemente in njihove funkcije, poznate vmesno programsko opremo Slurm in izvajati posle, poznate okoljske module in vsebnike, znate pripravljati izvajalne skripte itd...

Skupen potreben prostor za hrambo datotek – Vega (GB)

Načini dostopa za tovarne UI – EuroHPC JU



https://www.eurohpc-ju.europa.eu/ai-factories/ai-factories-access-modes_en
https://access.eurohpc-ju.europa.eu/

Vir: Klara Meštrovič, EuroHPC JU

SYSTEM*	SITE (COUNTRY)	PARTITION	PROCESSOR	ACCELERATOR	FIXED ALLOCATION** Playground Access	MAX. (MIN.) REQUEST** Fast Lane Access	MAX. PER MONTH** Large Scale Access	MAX. (MIN.) REQUEST*** Large Scale Access
MN5 MARENOSTRUM	BSC (ES)	**MN5 ACC**	Intel Sapphire Rapids	Nvidia Hopper	5 000	50 000 (Min. 10 000)	339 000	4 066 000 (Min. 50 000)
LEONARDO CINECA	CINECA (IT)	**Leonardo Booster**	Intel Xeon	NVIDIA A100	5 000	50 000 (Min. 10 000)	1 368 000	16 420 000 (Min. 50 000)
LUMI	CSC (FI)	**LUMI-G**	AMD Epyc	AMD Instinct	5 000	50 000 (Min. 10 000)	2 473 000	29 681 000 (Min. 50 000)
MELUXINA HIGH PERFORMANCE COMPUTING IN LUXEMBOURG	LuxProvide (LU)	**MeluXina GPU**	AMD Epyc	NVIDIA A100	5 000	50 000 (Min. 10 000)		
VEGA HPC	IZUM Maribor (SI)	**Vega GPU**	AMD Epyc	NVIDIA A100	5 000			
DISCOVERER Sofia Petascale Supercomputer	Sofia Tech Park (BG)	**Discoverer GPU**	AMD Epyc	NVIDIA H200	5 000			

Vir: https://www.eurohpc-ju.europa.eu/ai-factories/ai-factories-access-modes_en

EuroHPC JU Playground Access to AI factories

The Project

Project details

Project title: EU Legal LLM Fine-tuning Validation

Project summary (abstract):

We aim to fine-tune an open-source LLM on EU legal data. The goal is to test the hypotheses and gain more experience in building a specialized AI assistant for EU regulatory compliance that helps European SMEs understand and comply with EU legislation, including GDPR, the AI Act, and NIS2.

Keywords:

LLM, fine-tuning EU law, legal AI, compliance, Mistral, , GDPR, AI Act, NIS2, multilingual

Instructions: Not provided

Proposal for civilian purposes: true

Is any part of the project confidential?: No

Artificial Intelligence (AI) technology #1

AI set of technologies selection: Natural Language Processing

Share (%): 100

Application Domain #1

Application domain title: SH2 Institutions, Governance and Legal Systems

Organization details

Instructions: Not provided

Organization name: [redacted]

Organization type: Small and medium enterprise (SME)

Company VAT number: [redacted]

Organization with research activity: Yes

Organization head office is located in Europe: Yes

Organization department: R&D / IT deparment

AI Factory Selection

AI Factory Selection:

Vega GPU

Code(s) used: PyTorch, LoRA, PEFT, Hugging Face Transformers

Requested amount of resources (GPU hours): 5000

Maximum number of GPUs: 10

Total storage required (GB): 300

Total amount of data to transfer to/from (GB): 600

Komercialni dostopi (plačilo po uporabi)

Komercialni dostopi:
- plačilo po ceniku
- osnovna + ekspertna podpora
- enostavna prijava
- ni poročanja, NDA in pogodba
- komercialne storitve
- prodaja rešitev tretjim strankam

ARCTUR

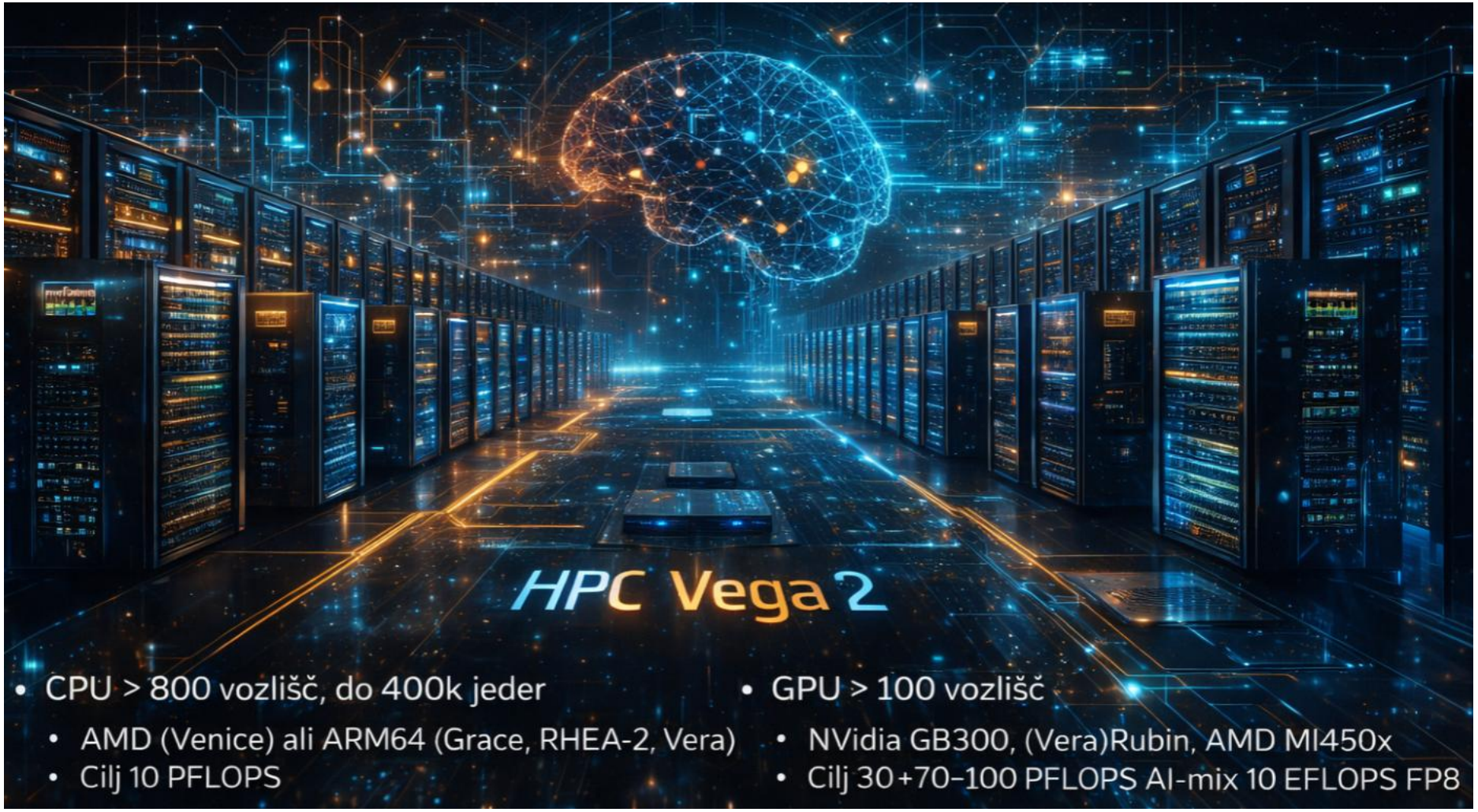
FORTISSIMO PLUS

VEGA

https://si-doc.vega.izum.si/cenik/

- CPU vozlišča (960):
  - 2 procesorja AMD (Zen 2), vsak 64 jeder
  - 256 GB (768xCPU), 1 TB (192xLM), DDR4
  - lokalni disk 2 TB, NVMe, priklop na IB 100Gb/s
- 60 vozlišč GPU s 4x NVidia A100 40 GB, 512 GB DDR4
- Diskovna sistema zmogljivosti 400GB/s in 200GB/s
- Letni pavšal (1000 EUR) vključuje:
  - tri uporabnike na enem projektu,
  - vsak uporabnik dobi 100 GB
  - uporaba osmih prijavnih vozlišč in razvojne particije
  - administracijo (izdaja računov) in priprava delovnega okolja
  - uro razvojnega inženirja

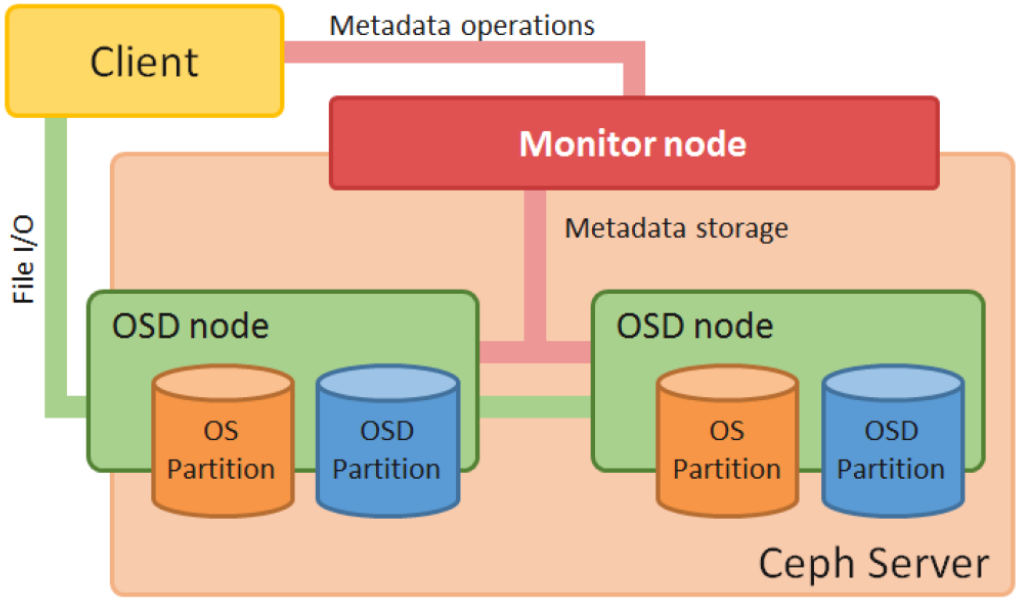
SLAIF AI-HPC (Vega 2) - splošen HPC + HPC za UI



Vir: MS 365 Copilot

SLAIF AI-HPC (Vega 2) - splošen HPC + HPC za UI

- Diskovni sistemi:
  - Visoka zmogljivost: Weka, VAST, DDN (cilj 10 PB)
  - Velika kapaciteta: CEPH (FS, S3,...) (cilj 100 PB)



- Omrežje:
  - Interconnect - Infiniband 800 Gb/s ali 1,6Gb/s
  - Povezljivost v internet 1,2 Tbit/s

- CLOUD particija (lokalni oblak):
  - Preko 100 strežnikov z vsaj eno grafično kartico, okoli 50 običajnih strežnikov
  - Sistemska infrastruktura, upravljanje, podporni servisi, monitoring …
  - Lokalni servisi, storitve za uporabnike (lokalni oblak), multitenency (izolacija)
  - Sistemi za sklepanje (inferenčni strežniki)

Sistemska (kontrolna) infrastruktura

slurm workload manager

klasično HPC/AI upravljanje batch
in delovnih obremenitev MPI

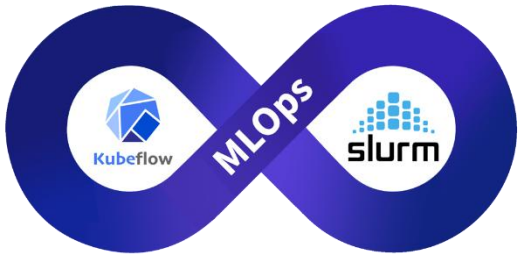
kubernetes

cloud-native orkestracija,
lifecycle, networking

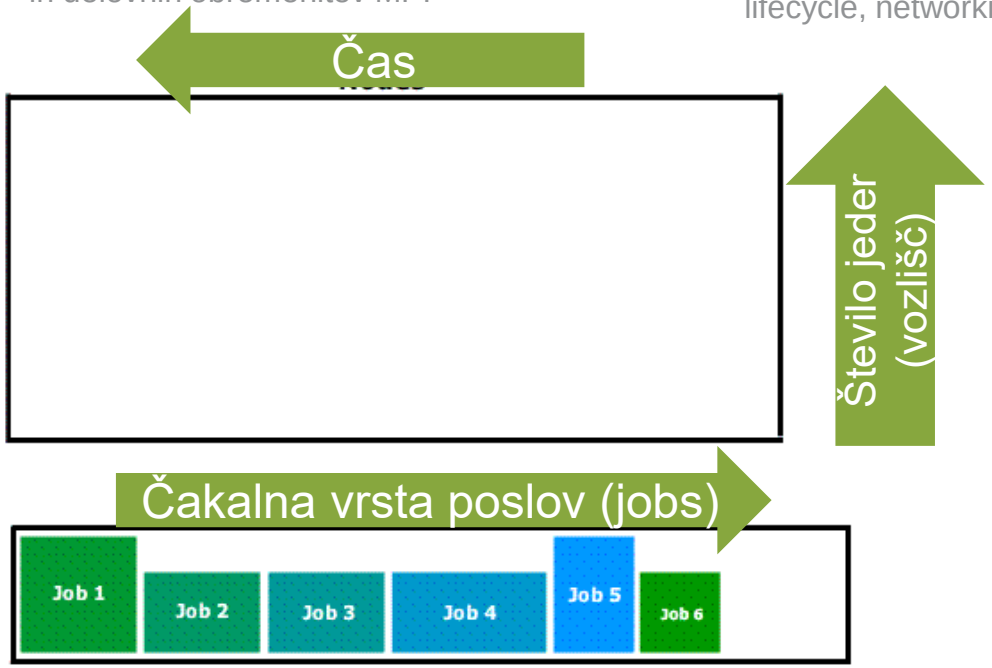
slinky

Kubernetes Operator for Slurm Clusters

Slurm zažene znotraj
Kubernetesa brez spremembe
uporabniške izkušnje



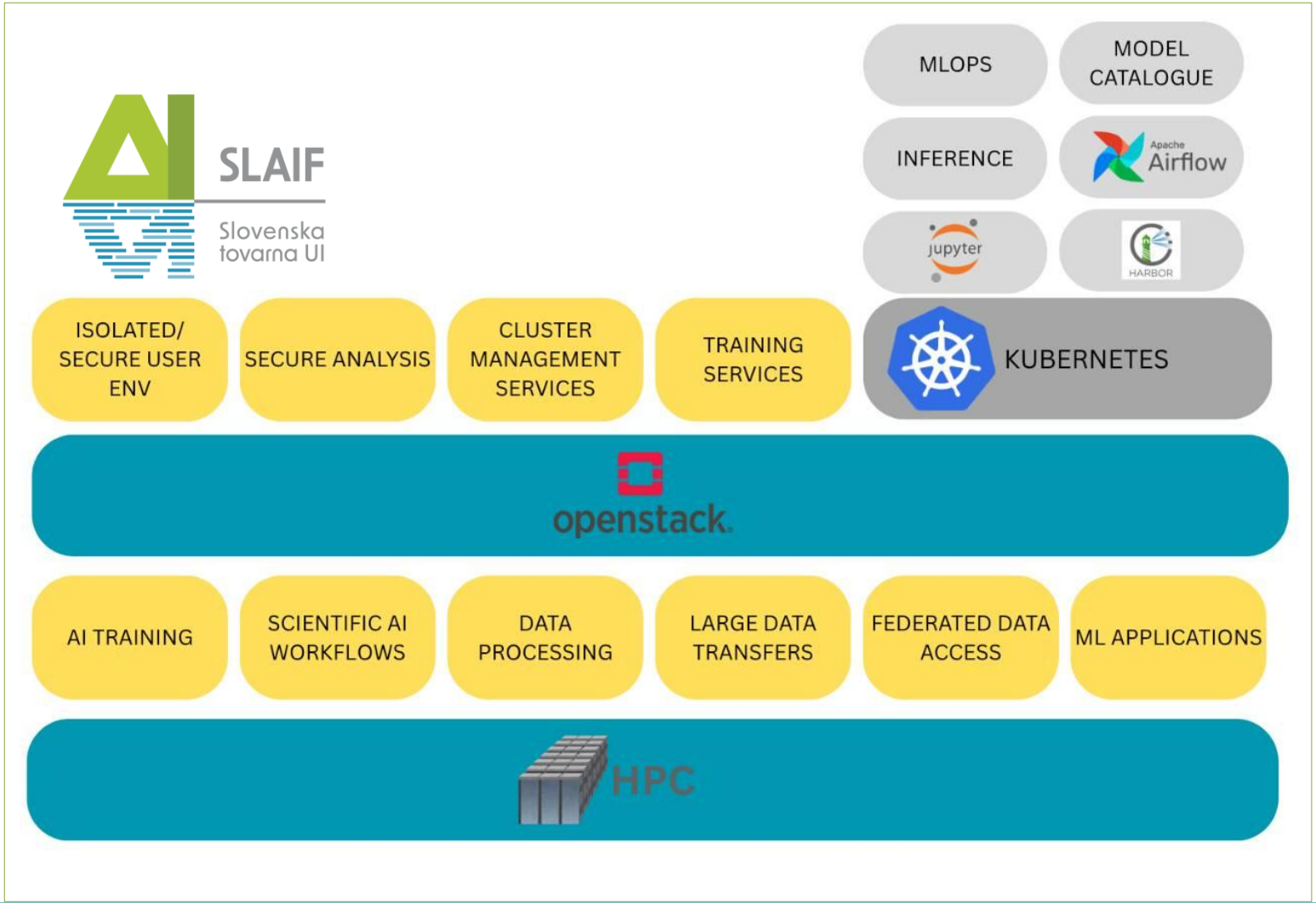
Vir: https://www.dkube.io/post/mlops-on-hpc-slurm-with-kubeflow



Plast	*Orodje*	*Vloga*
Orkestracija MLOps	**Kubeflow**	*kaj* se zgodi
Urniki, vrste HPC	**Slurm**	*kdaj in kje*
Infrastruktura	**Kubernetes**	*kako* teče

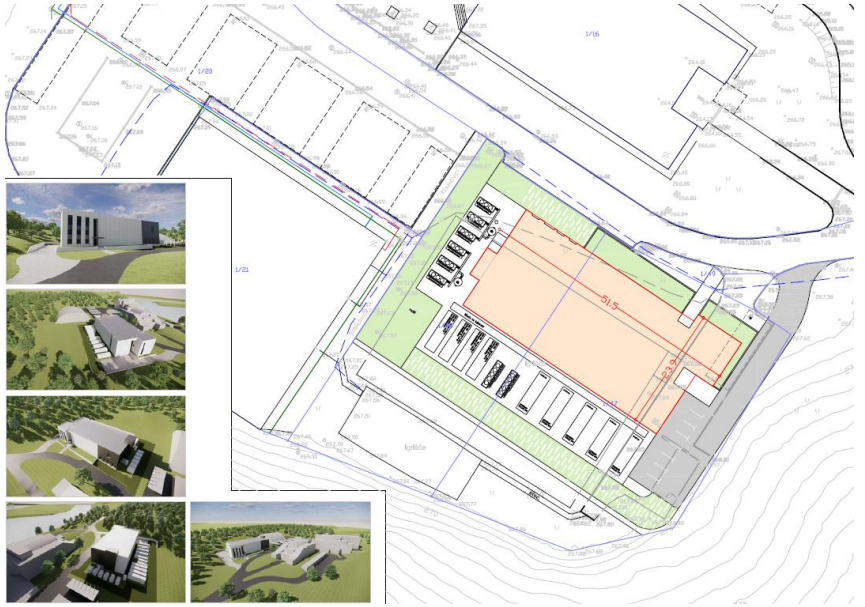
MLOps spravi ML iz Jupyter notebooka v produkcijo
Kubeflow vodi procese ML, Slurm pa vodi računske vire
Slinky omogoči sodelovanje Slurm+Kubernetes (Kubeflow)

Primer celovite sistemske arhitekture

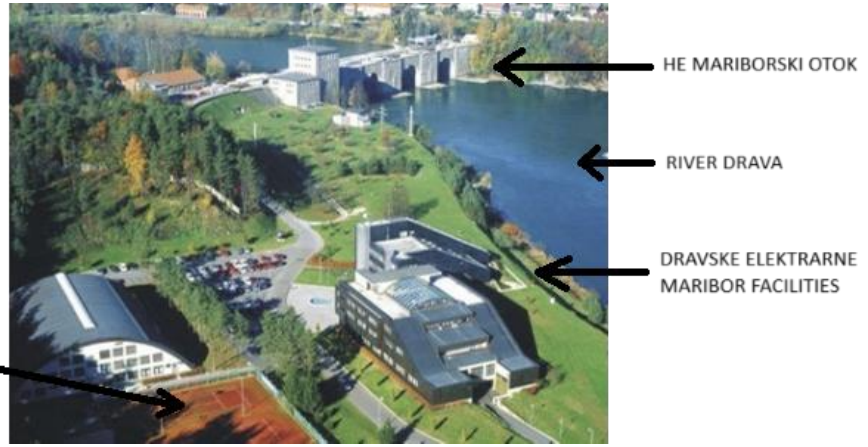


- Sistem vrst: HPC + OpenOnDemand
- Oblačni vmesnik (K18s)
- Testni inferenčni strežniki
- Delotoki in upravljanje podatkov
- Podatkovni viri in shrambe (S3, CVMFS, dCache)

Nov podatkovni center Arnes (HE MB otok)



arnes



Podpora SLING/SLAIF

Osebni pristop za zahteve, prošenje in želje podjetij

ENOTNA VSTOPNA TOČKA: support@slaif.si

podpora@sling.si / support@sling.si

info@sling.si

Spletna stran www.sling.si.

Naročite se na **Novičnik SLING**!

https://informator.arnes.si/?p=subscribe&id=7

Hvala za udeležbo! Vprašanja?

Thank you for attending! *Questions?*

IZUM
Institute of Information Science, Maribor, Slovenia

arnes

University of Maribor

Univerza v Novi Gorici

Gospodarska zbornica Slovenije
Chamber of Commerce and Industry of Slovenia

Jožef Stefan Institute
Ljubljana, Slovenia

UNIVERZA V LJUBLJANI
University of Ljubljana

FIS Fakulteta za informacijske študije
Faculty of information studies

Univerza na Primorskem

TP LJ01

Financerja / Financed by:

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA VISOKO ŠOLSTVO, ZNANOST IN INOVACIJE

EuroHPC Joint Undertaking

Projekt SLAIF: Slovenska tovarna umetne inteligence je finančno podprlo Ministrstvo za visoko šolstvo, znanost in inovacije. Projekt je bil na razpisu skupnega podjetja EuroHPC izbran za financiranje v okviru programov Obzorje Evropa ter Digitalna Evropa.

SLAIF: Slovenian AI Factory has been funded by the Ministry of Higher Education, Science and Innovation of Republic of Slovenia. At a call by EuroHPC JU, the project has received a positive funding decision under Horizon Europe and Digital Europe Programmes.