



WP4: User-Centric AI Services

Vid Kranjec | Univerza v Mariboru | LHRS

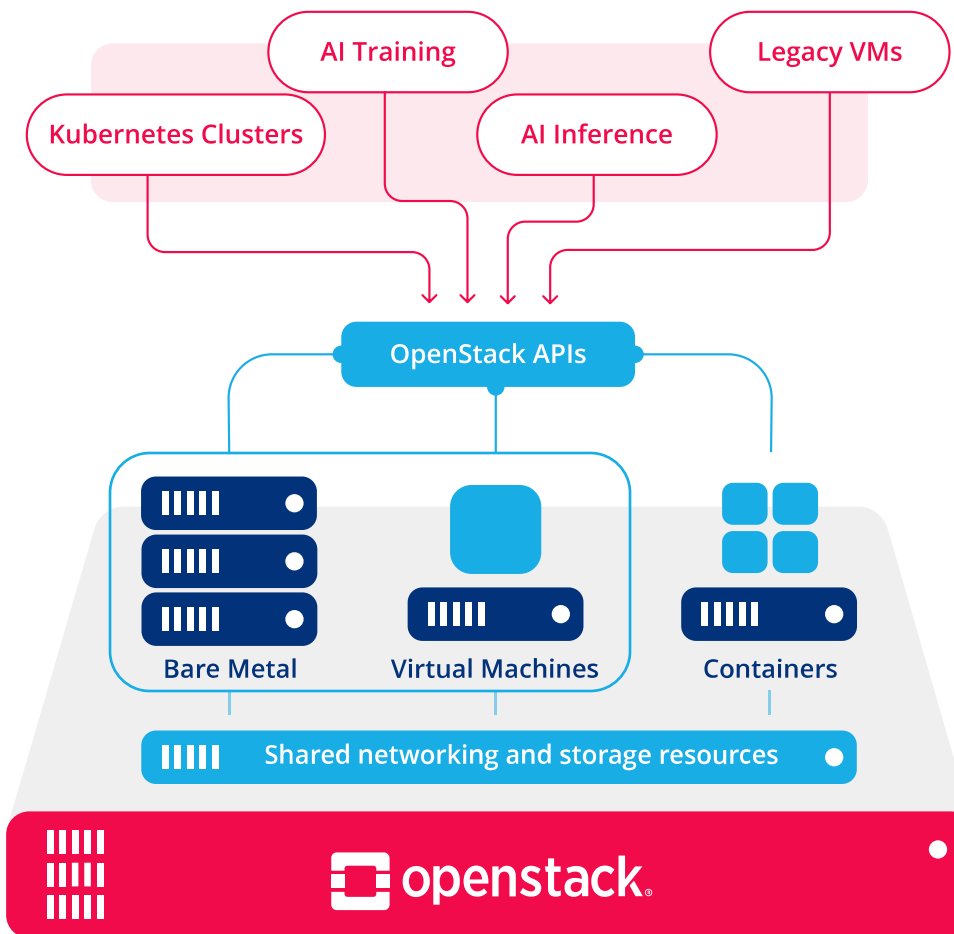
Vid Kranjec,

Laboratorij za heterogene računalniške sisteme, FERi UM

Platforma OpenStack®

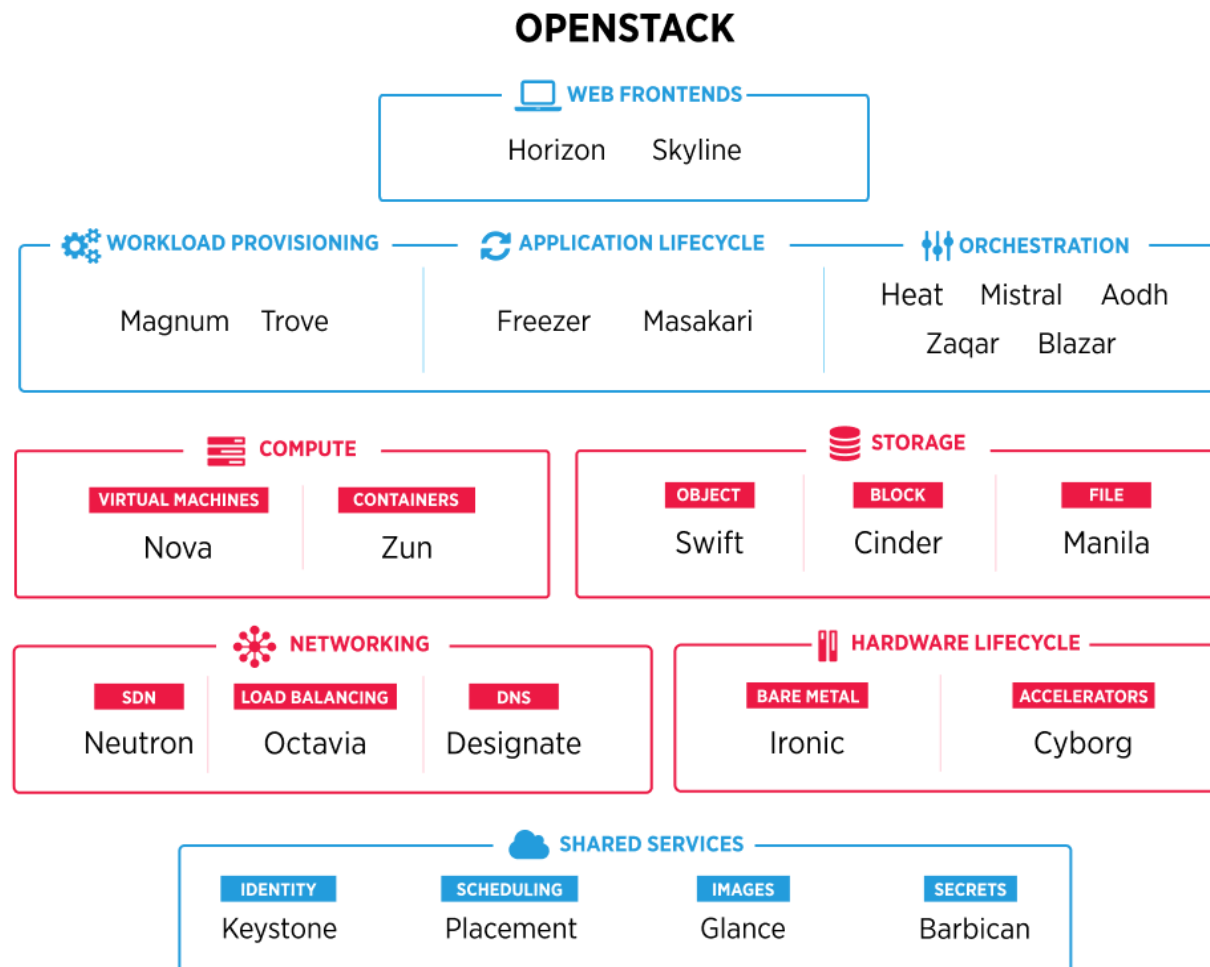
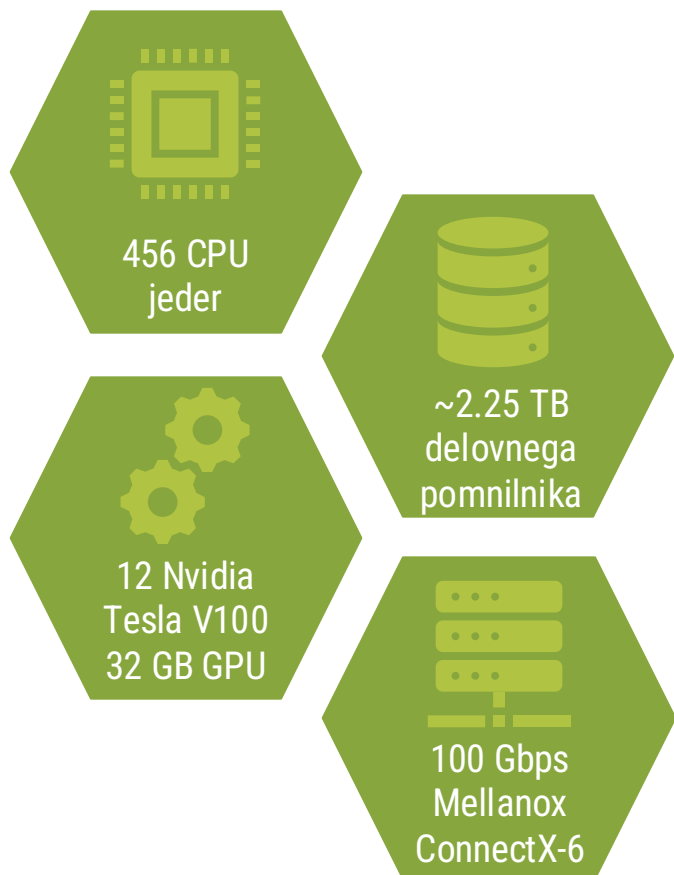


Compute, Storage and Networking infrastructure
for all your workloads



<https://www.openstack.org/>

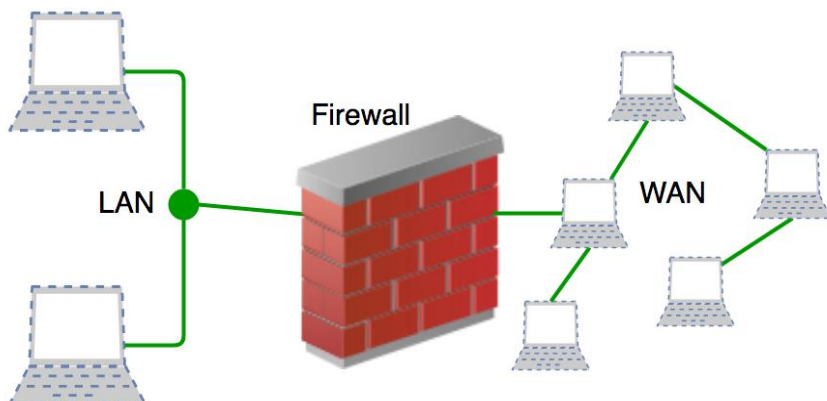
Tehnologije, storitve in računska moč



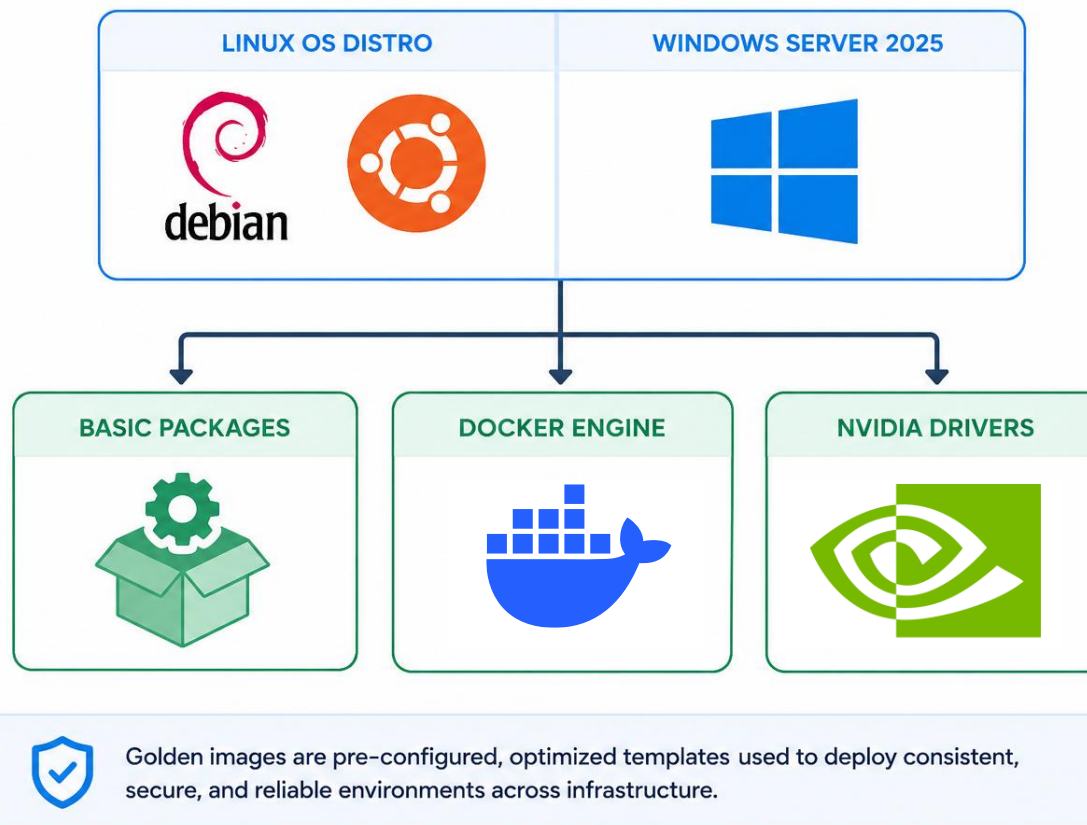
<https://www.openstack.org/>

Kaj je na voljo uporabnikom

- Zlate slike:
 - Debian 13
 - Ubuntu Cloud 24.04.3 LTS
 - Windows Server 2025
- Lastni telemetrični agent za zbiranje podatkov porabe po VM-ih
- Robustno omrežje in varnost



GOLDEN IMAGES

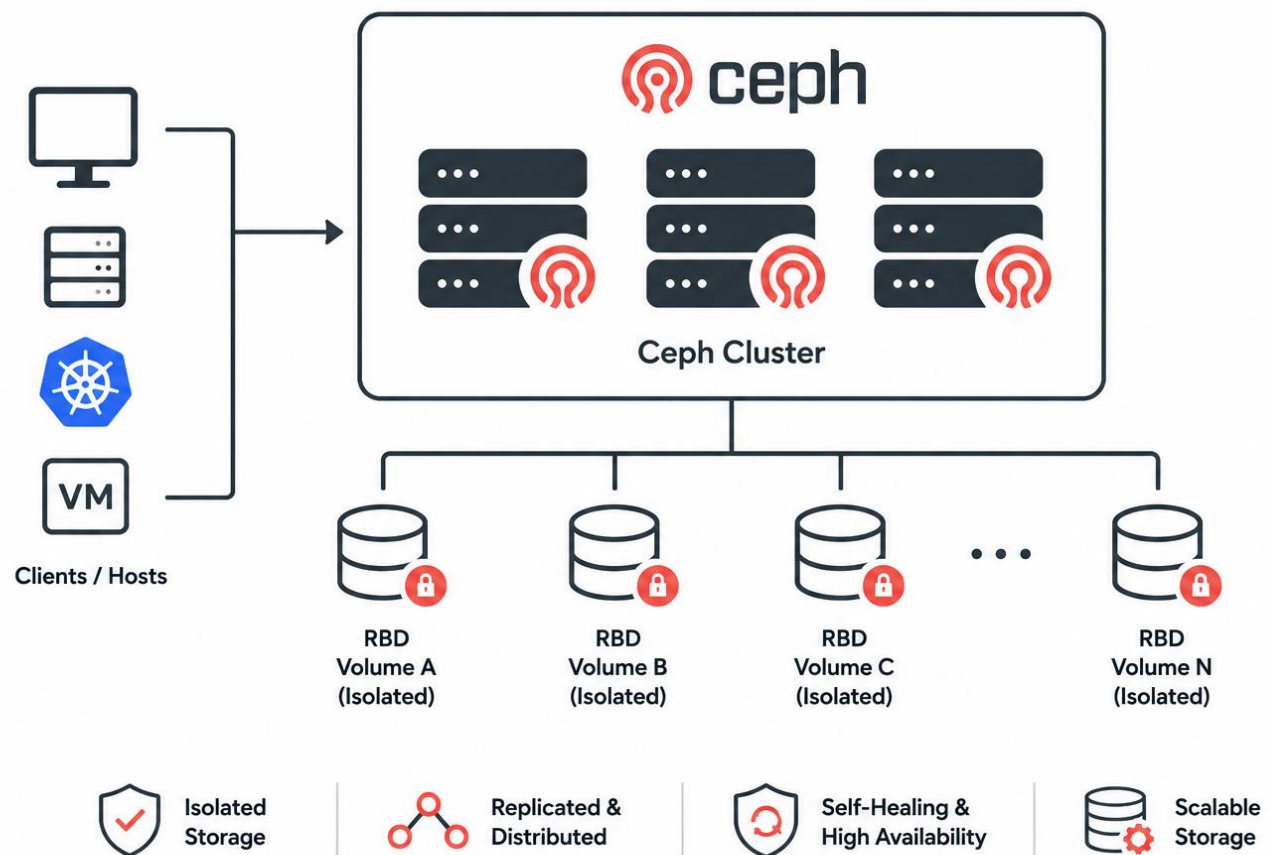


Future plans and upgrades 1/3

SSO for login option

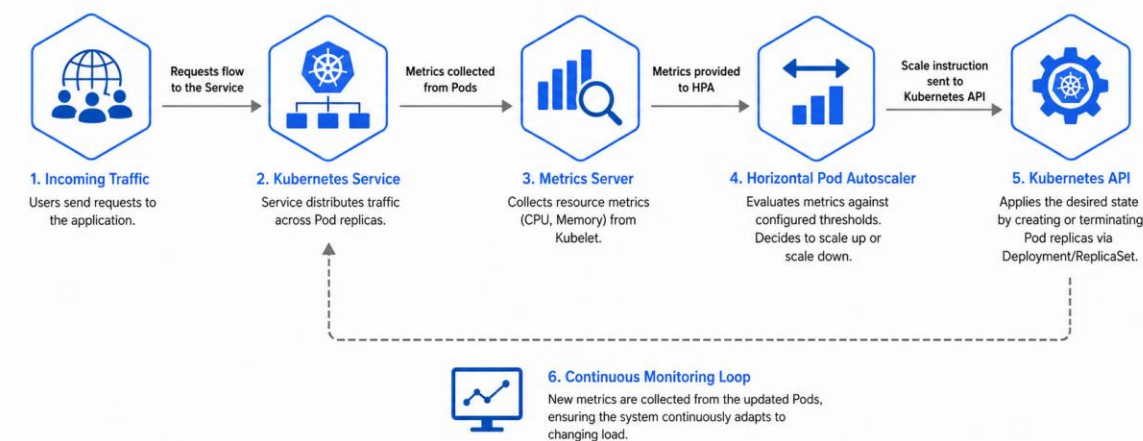


Isolated Ceph volumes (storage)



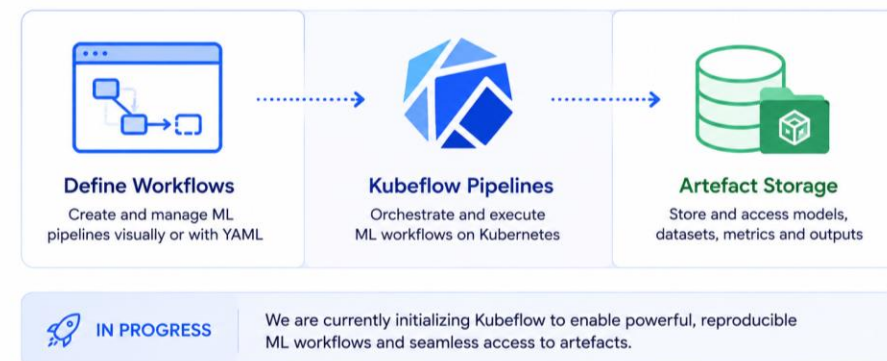
Future plans and upgrades 2/3

Kubernetes Autoscaling



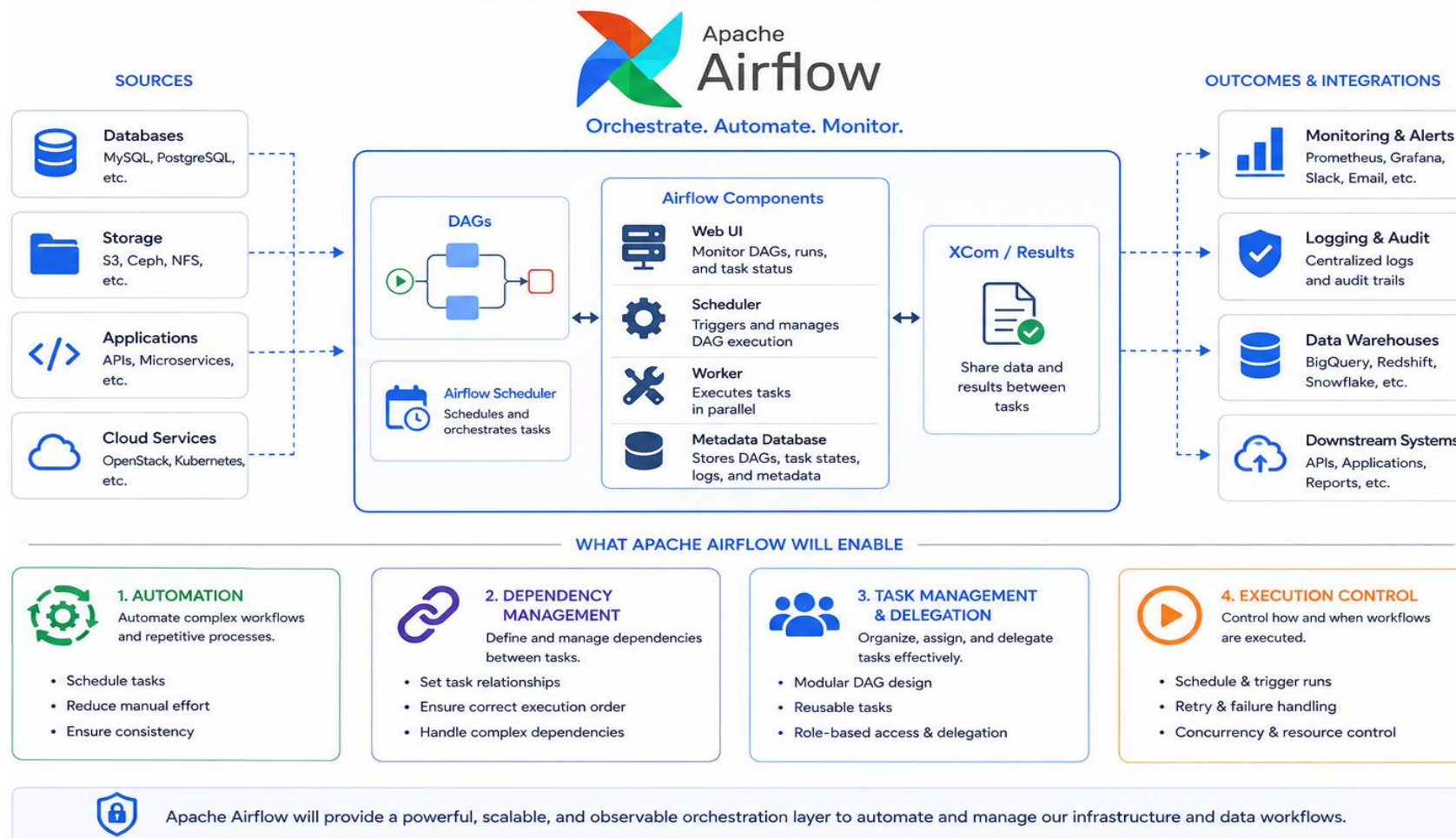
Kubeflow

Building the foundation for ML workflows

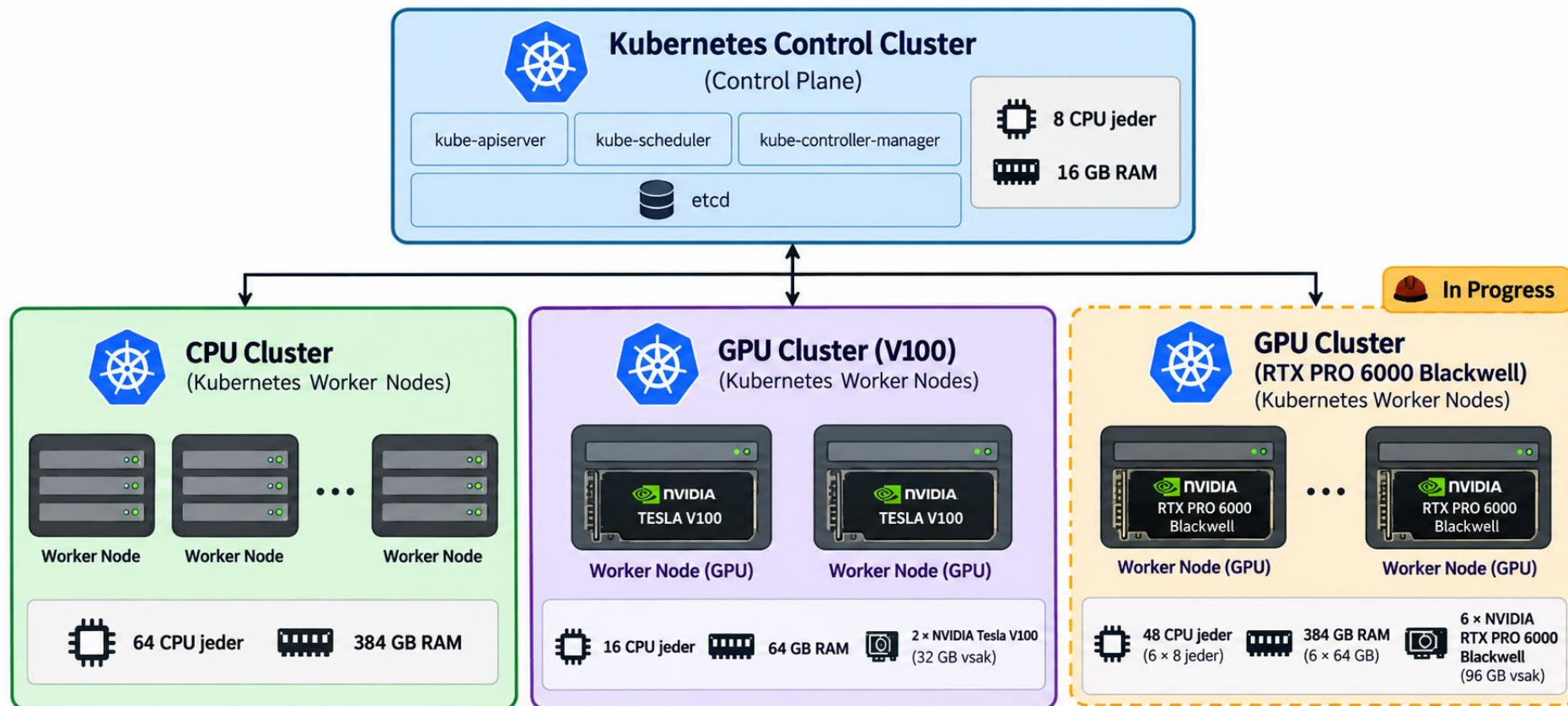


Future plans and upgrades 3/3

Integrating Apache Airflow for Intelligent Orchestration

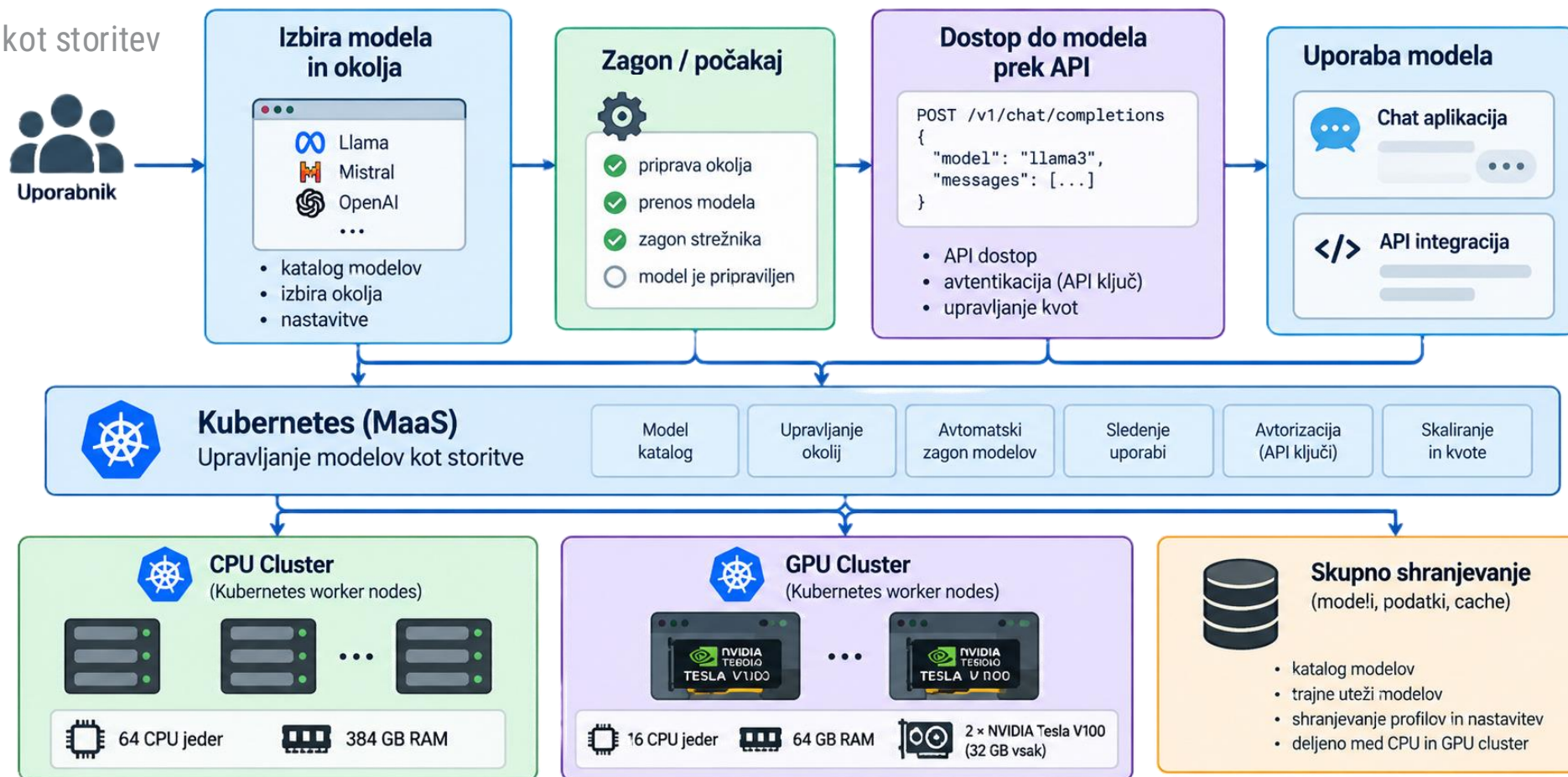


Kubernetes gruče



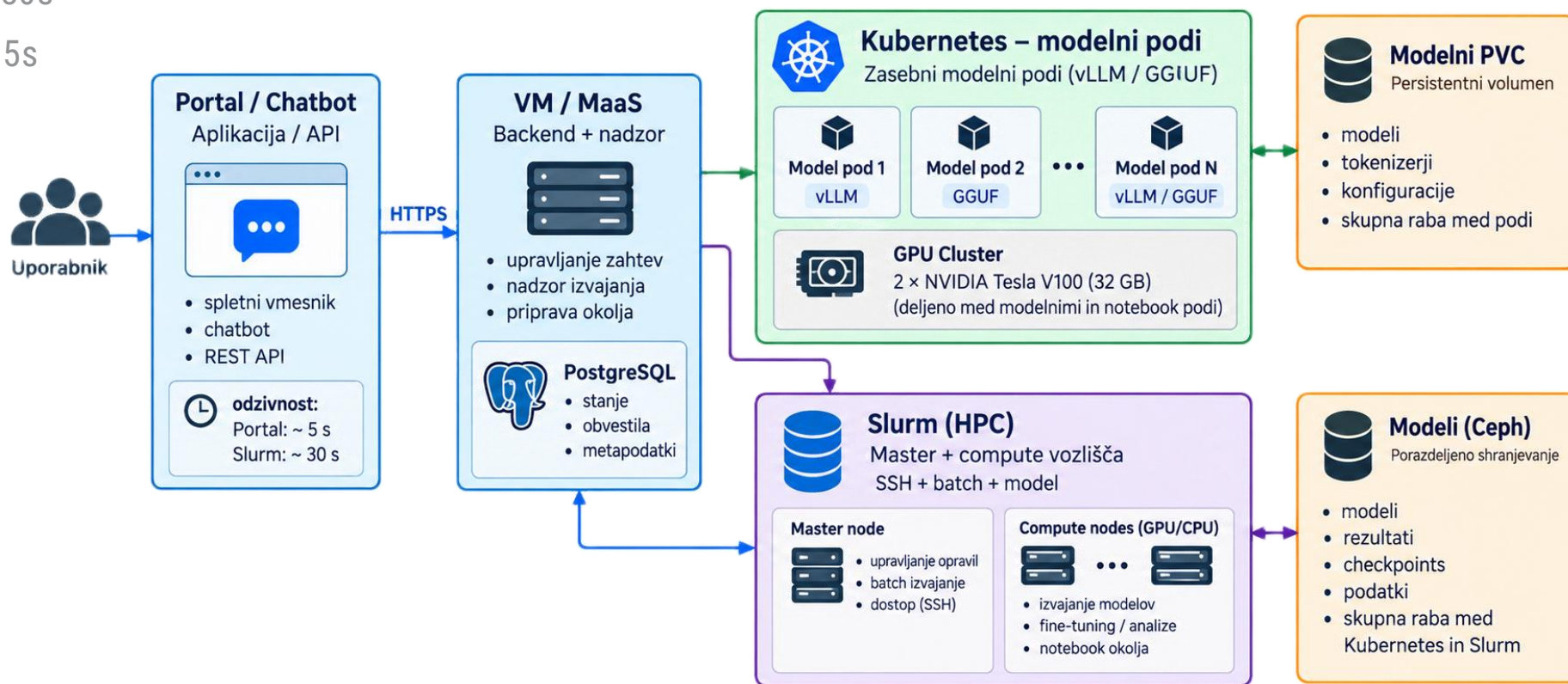
Model-as-a-Service (1)

- Model kot storitev



Model-as-a-Service (2)

- GPU kapaciteta Kubernetes: 2 × 32 GB, deljena med modelnimi in notebook podi
- Slurm: 30s
- Portal: 5s



Model-as-a-Service (3)

- Model in izvajalno okolje

The screenshot shows the 'Izberite svoj model' (Choose your model) page in the SLAIF console. It displays a grid of four AI models:

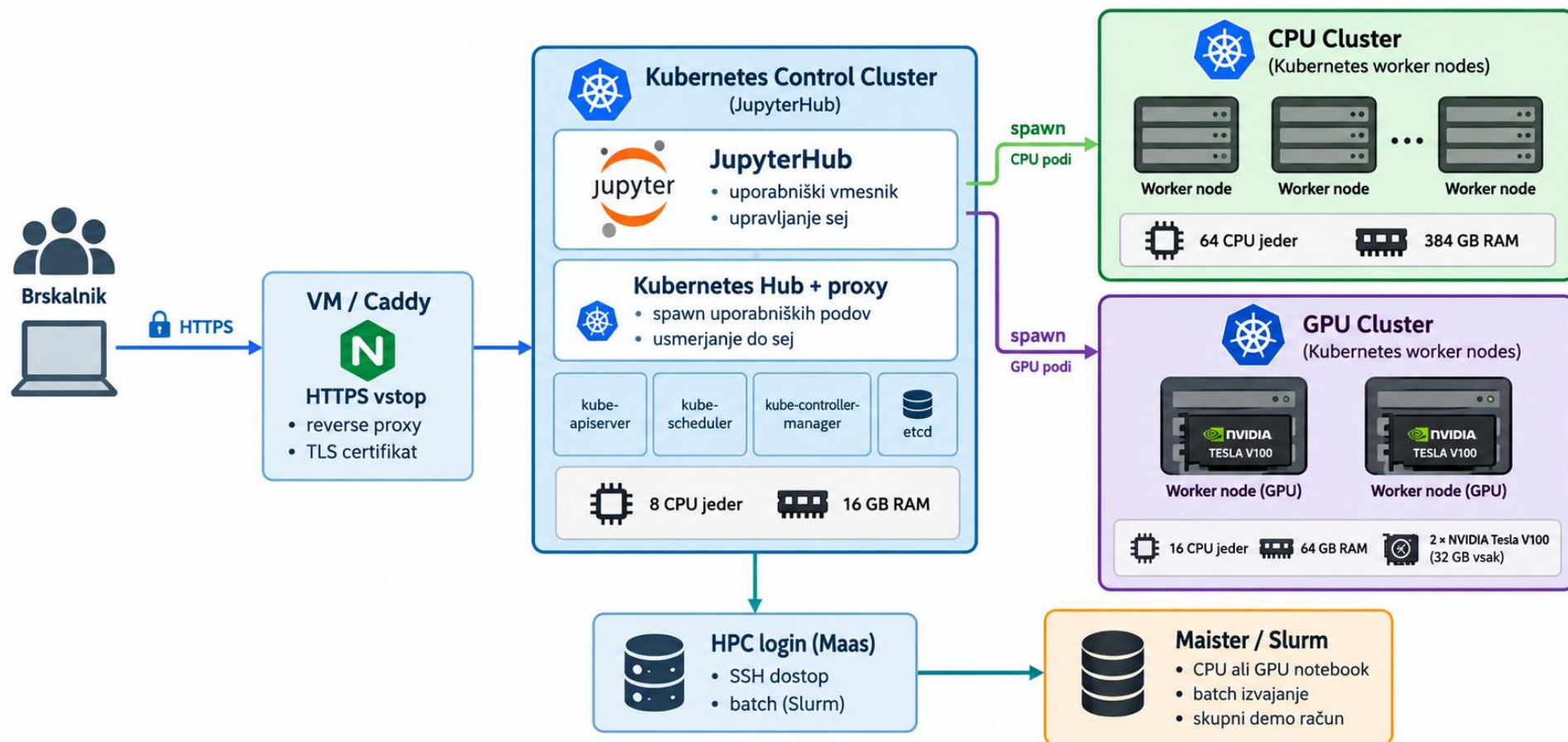
- Qwen 2.5 1.5B Instruct**: 1.5B parameters, 1.5B tokens, 1.5B tokens.
- Qwen3.5 27B - Q8**: 27B parameters, 27B tokens, 27B tokens.
- Qwen3.5 122B-A10B - Q3_A_M**: 122B parameters, 122B tokens, 122B tokens.
- Gemma 4 26B-A4B IT - Q8 - dolgi pogovori**: 26B parameters, 26B tokens, 26B tokens.

Each model card includes details about its architecture, supported languages, and a 'Poglej v zvezi' (View related) button.

- Pogovor, nastavitve in meritve

The screenshot shows the 'Chatbot' interface in the SLAIF console. It displays a conversation with the 'Gemma 4 26B-A4B IT - Q8 - dolgi pogovori' model. The chat history shows a user asking a question about the model's capabilities, and the chatbot responding with a detailed answer. The interface includes a 'Nov pogovor' (New chat) button and a 'Poglej v zvezi' (View related) button.

Jupyter Hub



Jupyter Hub (2)

- **CPE primer:** šest korakov

WP4 · JUPYTERHUB · KUBERNETES CPU

Učenje klasifikatorja na CPU

Sintetični podatki · 3 razredi · softmax klasifikator · ločena testna množica

97.8%

točnost na testnih podatkih

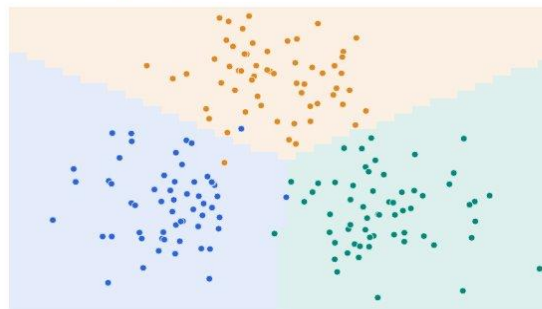
720 / 180

učni / testni primeri

1.46 s

učenje in vrednotenje · CPU

Naučene meje in testni primeri



Barva točke: pravi razred · ozadje: napoved modela

Ponovljiv primer (seed 42). Brez GPU, prenosa modela ali zunanjih podatkov.

Matrika napak

vrstice: pravi razred · stolpci: napoved

	0	1	2
0	55	1	1
1	1	59	0
2	1	0	62

Napaka med učenjem



200 korakov · izguba 1.099 → 0.124

- 97,8% točnost
- 180 testnih primerov

- **GPE primer:** štirje okoraki

```

3. Učenje majhne nevronske mreže na GPU

[ ]: model = torch.nn.Sequential(torch.nn.Linear(8, 16), torch.nn.Tanh(), torch.nn.Linear(16, 2)).cuda()
optimizer = torch.optim.Adam(model.parameters(), lr=0.04)
criterion = torch.nn.CrossEntropyLoss()
torch.cuda.synchronize()
start = time.perf_counter()
losses = []
for epoch in range(80):
    optimizer.zero_grad()
    loss = criterion(model(x_train), y_train)
    loss.backward()
    optimizer.step()
    losses.append(float(loss.detach()))
torch.cuda.synchronize()
training_seconds = time.perf_counter() - start
print(json.dumps({"korak": 3, "epoch": 80, "zacetna_izguba": losses[0], "koncna_izguba": losses[-1], "sekunde_ucenja": training_seconds}))

{"korak": 3, "epoch": 80, "zacetna_izguba": 0.6088966727256775, "koncna_izguba": 0.01072141993790865, "sekunde_ucenja": 0.2957458770251833}

4. Vrednotenje na nevidenih podatkih

[ ]: model.eval()
with torch.no_grad():
    prediction = model(x_test).argmax(dim=1)
    accuracy = float((prediction == y_test).float().mean())
    assert accuracy > 0.9
    assert losses[-1] < losses[0]
    result = {"korak": 4, "testna_točnost": accuracy, "GPU": torch.cuda.get_device_name(0),
              "vidne_GPU": torch.cuda.device_count(), "najvec_MB_tensorjev": torch.cuda.max_memory_allocated() / (1024 ** 2),
              "sekunde_ucenja": training_seconds, "status": "PASS"}
    print(json.dumps(result))

{"korak": 4, "testna_točnost": 0.996069375, "GPU": "Tesla V100-PCI-E-32GB", "vidne_GPU": 1, "najvec_MB_tensorjev": 18.43994140625, "sekunde_ucenja": 0.2957458770251833, "status": "PASS"}

```

- 99,6% točnost
- 512 testnih primerov

V prihodnje

- **V kratkem:**
 - **GPE platforma:** Vključitev 6 RTX Pro 6000 Blackwell GPE s kompatibilno programsko opremo
 - **Uporabniški dostop:** Dokončati Keycloak/SSO integracijo
 - **Pripravljena okolja:** Vzdrževanje in razširitev zlatih slik (*VM golden images*) ter razvojnih knjižnic
- **Pozneje:**
 - **Pripravljenost in fazna migracija:** Uskladitev z IZUM-om in migracija na VEGO
 - **Bodoči delovni tokovi:** Dokončanje implementacije Kubeflow in AirFlow

Kontakt

vid.kranjec1@um.si



marko.ferme@um.si



Hvala za udeležbo!

Thank you for attending!



Financerja / Financed by:



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.