

Posodobitve Arnesove superračunalniške gruče v letu 2025 in načrti za leto 2026

Arnes je v letu 2025 nadaljeval z intenzivnim razvojem in nadgradnjo svoje superračunalniške infrastrukture ter s tem dodatno okrepil zmogljivosti nacionalnega okolja za visokozmogljivo računanje. Arnesova računska gruča, ki deluje že od leta 2009, predstavlja eno ključnih procesnih zmogljivosti za slovenski raziskovalni in izobraževalni prostor. S skupno računsko močjo približno 2,3 petaFLOPS je drugi najzmogljivejši superračunalnik v Sloveniji ter edini, ki je dostopen tako začetnikom kot tudi zahtevnejšim uporabnikom.

Gruča je vključena v konzorcij SLING, njene kapacitete pa so raziskovalcem in študentom na voljo prek Odprtega nacionalnega dostopa. Uporablja se v pedagoških procesih ter pri raziskavah na področjih umetne inteligence, tehnologij znanja, teoretske fizike in fizike delcev, genomike, računske kemije ter obdelave velikih podatkov.

Ključne nadgradnje v letu 2025

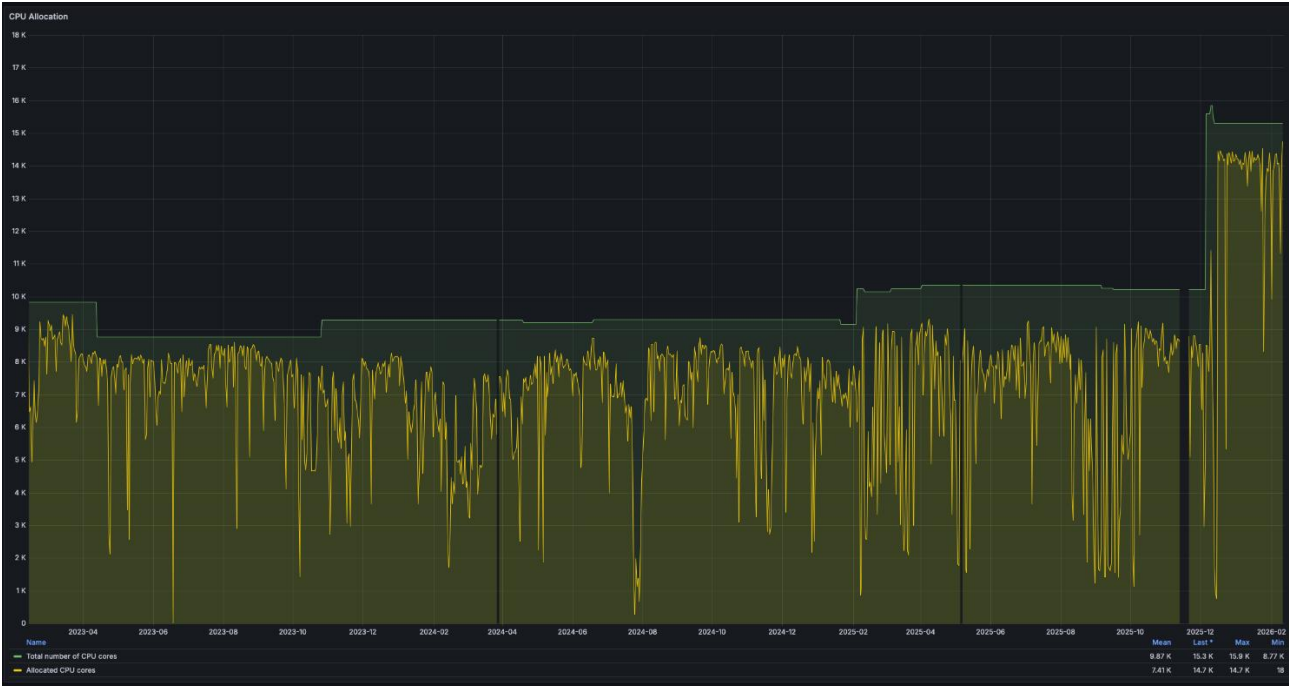
V letu 2025 je Arnes izvedel obsežno prenovno strojne opreme, pri čemer je zamenjal starejša CPU vozlišča z novimi zmogljivejšimi enotami, opremljenimi s procesorji AMD EPYC 9745 in do 1 TB delovnega pomnilnika. Skupno je bilo zamenjanih 41 starejših CPU vozlišč, ki so tudi bistveno zmogljivejša od predhodnikov. Ob koncu leta 2025 je teoretična zmogljivost CPE dela gruče znašala približno 0,5 PFLOPS.

Pomembna posodobitev je bila izvedena tudi na področju grafičnih pospeševalnikov. Dodana so bila tri nova GPU vozlišča s po štirimi pospeševalniki H100 SXM5, ki zagotavljajo bistveno višjo zmogljivost v primerjavi s prejšnjo generacijo. Iz uporabe je bilo umaknjenih sedem vozlišč z grafičnimi pospeševalniki V100S. Trenutno gruča vključuje 27 GPU vozlišč s skupno 60 grafičnimi pospeševalniki. Kljub manjšemu številu pospeševalnikov se je skupna zmogljivost GPU dela gruče povečala in je ob koncu leta dosegla približno 1,8 PFLOPS.

Arnes je posodobil tudi programsko okolje. Računska vozlišča in Ceph infrastruktura so bila nadgrajena na AlmaLinux 9, podatkovna gruča Ceph pa je bila postopoma različico 18 (Reef) in nato še na različico 19 (Squid). Zamenjanih je bilo tudi več starejših diskovnih enot, kar je povečalo razpoložljiv prostor za uporabnike na več kot 4 PB uporabnega prostora.

Rekordna uporaba in rast zahtevnosti nalog

Leto 2025 je zaznamovala tudi rekordna zasedenost superračunalnika. V decembru so uporabniki opravili kar 7.761.633 CPU ur, kar predstavlja nov mesečni rekord. Največ kapacitet trenutno porablja projekt ATLAS, povezan z raziskavami na CERN-u in odkritjem Higgsovega bozona.



Graf prikazuje izkoriščenost CPU-jev Arnesove gruč.

Arnes na svoji superračunalniški gruč mesečno izvede skoraj 665.000 nalog. Opazen je premik v naravi uporabe, saj se vse več zahtevnih obremenitev seli na grafične procesne enote, kar je predvsem posledica razmaha umetne inteligence. Naloge postajajo časovno daljše in kompleksnejše, Arnes pa temu trendu prilagaja razvoj infrastrukture.

Cluster Utilization 2025-12-01T00:00:00 - 2025-12-31T23:59:59

Usage reported in TRES Hours

Cluster	TRES Name	Allocated	Down	PLND	Dow	Idle	Planned	Reported
arnes	cpu	7761637	441522	19656		1476534	1181122	10880470
arnes	gres/gpu	21318	3226	2056		30848	0	57448

Cluster Utilization 2026-01-01T00:00:00 - 2026-01-31T23:59:59

Usage reported in TRES Hours

Cluster	TRES Name	Allocated	Down	PLND	Dow	Idle	Planned	Reported
arnes	cpu	10163148	22331	0		21661	1179036	11386176
arnes	gres/gpu	24019	419	0		29130	0	53568

Decembra 2025 so uporabniki na Arnesovi gruč opravili 7.761.637 CPU ur, januarja pa 10.163.148 CPU ur, kar je razvidno iz zgornjih izpisov.

Poenostavitev dostopa: JupyterHub in AAI prijava

Pomembna novost leta 2025 je bil tudi uspešno izveden preizkus nove storitve, ki temelji na JupyterHub in uporabnikom omogoča dostop do okolij JupyterLab in Jupyter Notebook neposredno prek brskalnika. Prijava poteka preko enotne AAI identitete, storitev pa je povezana z Arnesovo superračunalniško gručo in uporablja tako CPU kot GPU vire. Producerska različica bo uporabnikom na voljo v letu 2026.

Načrti za leto 2026: stabilnost, evropska integracija in prihodnje zmogljivosti

V letu 2026 Arnes načrtuje nadaljnje aktivnosti v treh ključnih sklopih: stabilnost in razvoj obstoječe infrastrukture, integracija v evropske raziskovalne infrastrukture ter načrtovanje prihodnjih zmogljivosti.

Med strateškimi cilji je vzpostavitev enotne nacionalne platforme s t. i. poolingom strojne opreme GPU, ki bi omogočila stroškovno učinkovito uporabo naprednih AI rešitev, dostop do sodobne opreme tudi manjšim raziskovalnim skupinam.

Arnes bo nadaljeval tudi s poenostavitvijo dostopa do virov v smeri »računske zmogljivosti kot storitve« ter z izboljšavami zalednih sistemov: posodobitvijo centralnega upravljanja konfiguracij, razširitvijo nadzornega sistema in uvedbo sodobnejšega obveščanja o napakah.

Med pomembnejšimi tehničnimi načrti za leto 2026 so še:

- vzpostavitev produkcijskega okolja JupyterHub,
- priprava načrta za integracijo v vozlišče EOSC,
- prenova centralnega sistema CVMFS in integracija EESSI,
- implementacija tehnologije RoCE za medvozliško komunikacijo z nizko latenco,
- nadgradnja razporejevalnika Slurm,
- optimizacija shranjevanja s Ceph in vključitev NVMe vozlišč,
- priprava novega sistema za upravljanje uporabnikov
- in preučitev možnosti virtualizacije z OpenStack.

Arnes bo ob tem še naprej krepil podporo uporabnikom, izboljševal procese dostopa in vzdrževal visoko raven kibernetike varnosti. Posebno pomembno ostaja tudi sodelovanje v strateškem projektu Tovarna umetne inteligence – SLAIF, ki predstavlja pomemben korak pri razvoju zmogljivosti umetne inteligence v Sloveniji.