Nabava infrastrukture za projekt BrAIn - Podrška primjeni digitalnih tehnologija u obrazovanju

Obavijest gospodarskim subjektima s ciljem istraživanja tržišta

Sadržaj

1 UVOD

1.1 Općenito o projektu BrAIn

Projekt "BrAIn" koji provodi Hrvatska akademska i istraživačka mreža - CARNET, usmjeren je na integraciju digitalnih tehnologija u obrazovni proces s ciljem poboljšanja kvalitete i produljenja vremena koje učenici provode u odgojno-obrazovnom procesu. Projekt teži personaliziranom pristupu učenju i poučavanju, razvoju kurikuluma i digitalnog obrazovnog sadržaja iz područja digitalnih tehnologija. Specifični ciljevi uključuju razvoj digitalnih kompetencija učenika i nastavnika, razvoj sustava pametnih preporuka za bolji uvid u učenička postignuća, te automatizaciju nadzora i upravljanja za visoku dostupnost digitaliziranih obrazovnih usluga.

Projekt, koji traje od 1. rujna 2023. do 31. kolovoza 2029., obuhvaća edukaciju i istraživanje za razvoj digitalnih kompetencija, izradu kurikuluma i digitalnih obrazovnih sadržaja, te provedbu obrazovanja za učitelje i nastavnike. U sklopu istraživačke komponente, planirano je provođenje kontinuiranih istraživanja utjecaja digitalnih tehnologija na učenike, kao i razvoj kurikuluma za izvannastavne aktivnosti i fakultativne predmete usmjerene na razvoj digitalnih kompetencija.

1.2 Općenito o nabavi

Sukladno Zakonu o javnoj nabavi (NN 120/16, NN114/22) sa svrhom pripreme nabave i informiranja gospodarskih subjekata o svojim planovima i zahtjevima u vezi s nabavom, u nastavku CARNET objavljuje zahtjeve vezane za nabavu i isporuku infrastrukture namijenjene za potrebe umjetne inteligencije.

Radi daljnjeg planiranja i provedbe postupka nabave te izrade Dokumentacije o nabavi molimo sve zainteresirane gospodarske subjekte da dostave primjedbe i prijedloge prema traženim informacijama i troškovnikom **najkasnije do 04.06.2024. na adresu elektroničke pošte nabava@carnet.hr.**
Prilikom provođenja istraživanja tržišta CARNET će postupati na način da svojim postupcima ne narušava tržišno natjecanje niti krši načela zabrane diskriminacije i transparentnosti.

Rezultati provedenog istraživanja ne obvezuju CARNET niti se stvara bilo kakav pravni posao ili odnos s gospodarskim subjektima koji sudjeluju u istraživanju.

2 OPIS PREDMETA NABAVE

2.1 Infrastruktura

Predmet ove nabave je nabava infrastrukture namijenjene za potrebe umjetne inteligencije, uključujući, ali ne ograničavajući se na, hardver i pripadajući softver potreban za efikasno izvođenje aplikacija umjetne inteligencije. Infrastruktura će podržati napredne aplikacije kao što su inferencija velikih jezičnih modela, prepoznavanje prirodnog jezika, Omniverse aplikacije, te GPU-ubrzana analitika inteligentnog videa.

Izvedba aplikacija za inferenciju znatno je ubrzana upotrebom NVIDIA GPU-ova i uključuje radne opterećenja kao što su:

- Inferencija velikih jezičnih modela
- Prepoznavanje prirodnog jezika (NLR)
- Omniverse aplikacije
- DeepStream – GPU-ubrzana analitika inteligentnog videa (IVA)
- NVIDIA® TensorRT™ Triton – softver za inferenciju s ubrzanjem GPU-a GPU server dizajniran za izvođenje radnih opterećenja inferencije može biti raspoređen na rubu mreže ili u podatkovnom centru. Svaka lokacija servera ima svoj skup zahtjeva za okoliš i usklađenost. Na primjer, server na rubu mreže može zahtijevati NEBS usklađenost s strožim termičkim i mehaničkim zahtjevima.

Tablica 1 pruža zahtjeve konfiguracije sustava za server inferencije koji koristi NVIDIA GPU-ove. Veliki jezični modeli trebali bi ciljati na specifikacije višeg kraja. Upotreba Omniverse i aplikacija za vizualizaciju trebat će L40S/L40.

Tablica 1 Konfiguracija sustava servera

Tablica 1. Konfiguracija sustava servera

Parametar	Konfiguracija servera za inferenciju
GPU	L40S L40 L4 H100 H100 HGX
Konfiguracija GPU-a	2x / 4x / 8x GPU-a po serveru, preporučeno 4x kako bi se izbjegla potreba za PCIe prekidačem. GPU-ovi bi trebali biti balansirani preko CPU utora i root portova.
CPU	Preporučuju se x86 PCIe Gen5 sposobni CPU-ovi kao što su Intel Xeon skalabilni procesor (Sapphire Rapids) ili AMD Genoa.
CPU utori	Minimalno 2 CPU utora
Brzina CPU-a	Minimalna osnovna frekvencija od 2.1 GHz
CPU jezgre	6x fizičkih CPU jezgri po GPU-u
Sistemska memorija	Minimalno 1.5x ukupne memorije GPU-a / preporučuje se 2.0x. Ravnomjerno raspoređeno preko svih CPU utora i memorijskih kanala.

DPU	Jedan Bluefield®-3 DPU po serveru
PCI Express	Minimalno jedna Gen5 x16 veza po Gen5 GPU-u se preporučuje. Minimalno jedna Gen4 x16 veza po Gen4 GPU-u se preporučuje. Minimalno jedna Gen5 x16 veza po 2x GPU-ima za konfiguracije s PCIe prekidačem.
PCIe topologija	Za uravnoteženu PCIe arhitekturu, GPU-ovi bi trebali biti ravnomjerno raspoređeni preko CPU utora i PCIe root portova. NIC-ovi i NVMe drajvovi bi trebali biti postavljeni unutar istog PCIe prekidača ili root kompleksa kao i GPU-ovi. Važno je napomenuti da PCIe prekidač može biti opcionalan za ekonomične servere za inferenciju.
PCIe prekidači	Direktno povezivanje s CPU-om je preferirano. ConnectX®-7 Gen5 PCIE prekidači prema potrebi.
Mrežni adapter (NIC)	ConnectX®-7 (do 400 Gbps), BlueField®-3 DPU u NIC modu (do 400 Gbps). Vidi Odjeljak Mreža za detalje.
Brzina NIC-a	Do 400 Gbps po GPU-u, minimalno 50 Gbps po GPU-u za inferenciju na jednom čvoru, minimalno 200 Gbps za multi
Pohrana	Jedan NVMe po CPU utoru
Upravljanje udaljenim sustavima	Kompatibilno s Redfish 1.0 (ili većom verzijom)
Sigurnost	Modul TPM 2.0 (siguran pokretanje)

U sklopu projekta, nabava opreme za umjetnu inteligenciju od ključne je važnosti za ostvarivanje ciljeva projekta koji uključuje obradu velike količine podataka, treniranje i korištenje modela umjetne inteligencije, posebno za jezične modele velikih dimenzija (LLM). Za to su nam potrebni snažni računalni resursi koji omogućuju brzu i efikasnu obradu podataka, uključujući specijalizirane grafičke procesore (GPU) koji omogućuju paralelnu obradu podataka, što je ključno za treniranje složenih modela umjetne inteligencije. Osim tog, potrebni su nam i procesori visokih performansi, velike količine memorije za pohranu podataka te napredni mrežni adapteri za brz prijenos podataka. Ova oprema nam omogućava da efikasno treniramo modele umjetne inteligencije, razvijamo napredne algoritme i implementiramo rješenja koja mogu obraditi i analizirati.

2.2 Računalni uređaji

Dodatno, nabava obuhvaća i računalne uređaje visokih performansi, dizajnirane za mobilnu upotrebu, s integriranim grafičkim procesorima optimiziranim za zahtjevne zadatke strojnog učenja. Ti uređaji će

omogućiti fleksibilnost i mobilnost korisnicima, podržavajući istovremeno i složene računalne operacije potrebne za istraživanje i razvoj u području umjetne inteligencije.

U Prilogu 1. Tehnička specifikacija – računala se nalazi tehnička specifikacija.

3 JAMSTVO

Odabrani Ponuditelj jamči dostupnost i ispravan rad isporučene infrastrukture opisane pod točkom 2 tijekom cijelog trajanja ugovora. Nadalje, Ponuditelj se obvezuje osigurati produženo jamstvo za otklanjanje nedostataka koje traje ukupno 5 godina od dana isporuke.

Troškove otklanjanja nedostataka i/ili kvarova za vrijeme jamstvenog roka u cijelosti snosi odabrani Ponuditelj.