

Slovensko superračunalništvo za umetno inteligenco

Avtorja: Jan Jona Javoršek, Florian Klauser, Institut »Jožef Stefan«

Tehnološke inovacije z zvenečimi termini često lepo odmevajo v medijih, a pogosto je to le brenčanje modne muhe. Zanimanje za umetno inteligenco ne bo tako minljivo, saj jo je pohod generativnih modelov – od trenutno najodmevnejših velikih jezikovnih modelov kot sta Claude in ChatGPT, preko še vedno popularnih slikovnih modelov kot na primer DeepAI in DALL-E, do manj znanih ali celo poslovno tajnih specializiranih industrijsko-razvojnih modelov – trdno postavil v srčiko razvojnih, tehnoloških, upravnih, uporabniško-vmesniških in umetniško-ustvarjalnih temeljev sodobne družbe.

Da bodo naši raziskovalci ter razvijalci novih tehnologij imeli dovolj dobra orodja za vse bolj zahtevne tehnološke in kompetenčne izzive, ki jih prinaša vzpon umetne inteligence, v skupnem evropskem raziskovalnem prostoru postavljamo številne tovarne umetne inteligence. Eno takšnih tovarn vzpostavljamo tudi v Sloveniji. Projekt bo vključeval številne podporne mehanizme za svetovanje in pomoč uporabnikom ter izobraževalne aktivnosti, njegovo jedro pa predstavlja vzpostavitev novega superračunalnika, zasnovanega posebej za naloge umetne inteligence.



Novi slovenski superračunalnik, ki bo poganjal tovarno umetne inteligence, bo nameščen v Arnesovem podatkovnem centru prihodnosti na območju Dravskih elektrarn v Mariboru.

Razvoj velikih jezikovnih modelov predstavlja nekaj dodatnih izzivov za današnje superračunalnike. Med temi izstopa predvsem količina podatkov, ki jih je treba zbrati, urediti, odstraniti podvajanje, normalizirati, razčleniti na manjše enote, preoblikovati v numerično obliko vektorjev in organizirati v enote za učenje nevronskega modela. Ta potem skozi več plasti kodiranja in dekodiranja prilagaja uteži svojih sinaps in tako lahko za poljubne vnose jezikovnih elementov izračuna verjetnost naslednjega jezikovnega elementa oziroma zaporedja besed.

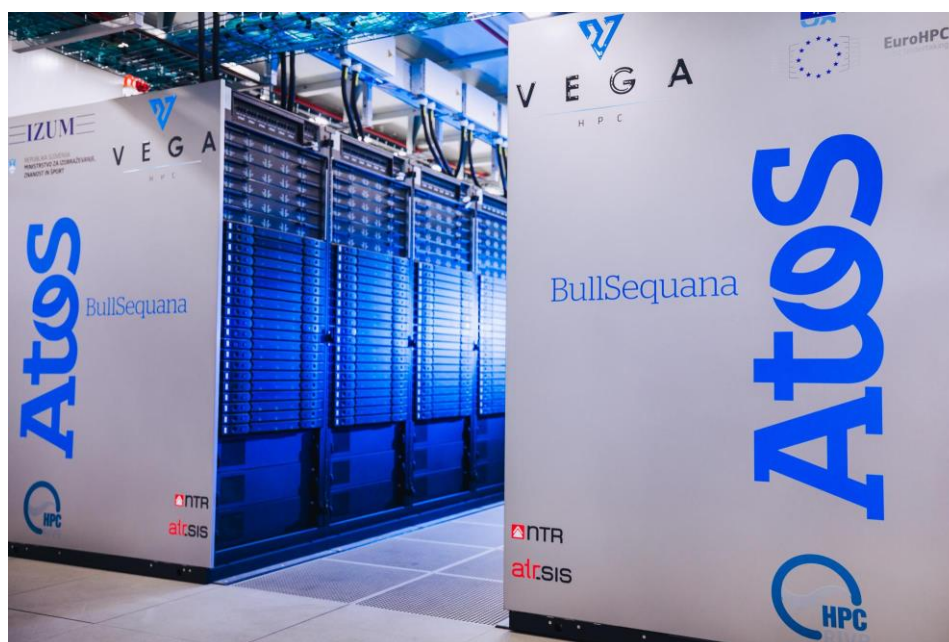
Že manjši jezikovni modeli so računsko zelo zahtevni: učenje slovenskega modela GAMS 1B, ki je doučen na podlagi modela OPT 1B, je zahtevalo 10 različnih poskusov, vsak poskus pa je trajal 16 ur na 6 vozliščih slovenskega superračunalnika Vega (vsak model je bil razporejen čez eno vozlišče, učenje je potekalo vzporedno).

Sodobne generacije grafičnih pospeševalnikov vodilnih podjetij, kot sta NVIDIA in AMD so močno optimizirane za računske operacije, ki jih zahteva programska oprema za strojno učenje. Ker matrične operacije za to uporabo ne zahtevajo velike natančnosti, novi procesorji večino svojih kapacitet namenjajo krajšim reprezentacijam števil s plavajočo vejico: namesto 64 bitov se jih uporabi le 8 ali manj. Tako lahko izvedejo mnogo več operacij v danem času.

Za umetno inteligenco prilagojeni superračunalniki so opremljeni tudi s tenzorskimi procesnimi enotami in nevronskimi procesnimi enotami. Te enote so posebej prilagojene za matrične operacije z večdimenzionalnimi vektorji (tenzoriji) z majhno energetske porabo, ki jih potrebujemo za dejansko uporabo danega modela, a se pogosto uporabljajo tudi za šifriranje, obdelavo slik in zvoka, multimedijske algoritme in podobno.

Za spopadanje z vse večjimi količinami podatkov pa je bilo treba omogočiti tudi neposredno komunikacijo med računskimi procesorji. Podjetje NVIDIA je, na primer, uvedlo vodilo NVLink, ki z zunanjim nizkolatenčnim stikalom omogoča povezavo 72 grafičnih procesnih enot (in 36 novih splošnih procesorjev ARM) in posebna stikala s 36 porti s kapaciteto 800 Gbit/s, ki omogoča pasovno širino 900 GB/s. Druga podjetja razvijajo podobno učinkovite rešitve.

Zaradi velike količine podatkov je treba zagotoviti tudi visoko zmogljiv datotečni sistem, ki dandanes navadno temelji na kombinacij pred kratkim zelo napredovalih tehnologij računanja v oblaku in vse večjimi kapacitetami hitrih naprav flash.



Superračunalnik Vega

Novi slovenski superračunalnik, prilagojen nalogam umetne inteligence, bodo v okviru projekta Slovenska tovarna umetne inteligence (SLAIF) vzpostavili IZUM, Arnes in Institut »Jožef Stefan«. Nahajal se bo v novem podatkovnem centru Arnesa pri Dravskih elektrarnah, kar bo omogočilo energetsko učinkovito napajanje sistema. V projektu neposredno sodeluje deset partnerjev: Institut »Jožef Stefan«, Univerza v Ljubljani, Univerza v Mariboru, Univerza v Novi Gorici, IZUM, Arnes, Univerza na Primorskem, Fakulteta za informacijske študije Novo mesto, Gospodarska zbornica Slovenije in Tehnološki park Ljubljana.

Za razvijanje in uporabo umetne inteligence za raziskovalne, administrativne ali poslovne namene ni treba čakati na vzpostavitev novega superračunalnika. Sistem Vega že prilagajamo za uporabo v kontekstu tovarne umetne inteligence, ki bo omogočila hitrejši dostop, lažje poganjanje delotokov za gradnjo, analizo, dograjevanje in uporabo modelov ter predvsem učinkovit in bogat ekosistem ekspertov, ki bodo omogočali raziskovalcem in podjetnikom dostop do novih tehnologij. Slovensko nacionalno superračunalniško omrežje (SLING) kot del projekta EuroCC2 s superračunalniki v Mariboru (IZUM), Ljubljani (Arnes in IJS) in Novi Gorici (Arctur) nudi dostop do superračunalniških storitev, podpore in izobraževanja, uporabnikom pa tudi pomaga pridobiti dostop do velikih sistemov evropske iniciative EuroHPC. Uporabnikom so na voljo brezplačna izobraževanja in svetovanja. V okviru projekta raziskovalci in podjetja lahko izvedejo tudi »proof-of-concept« in se tako podajo na pot inovacij s pomočjo umetne inteligence in superračunalništva. Več informacij na: www.sling.si

Projekt EuroCC2 financira Evropsko Skupno podjetje za evropsko visokozmogljivo računalništvo (JU) v okviru sporazuma o dodelitvi sredstev št. 101101903. JU podpirajo EU program Digitalna Evropa, Nemčija, Bolgarija, Avstrija, Hrvaška, Ciper, Češka republika, Danska, Estonija, Finska, Grčija, Madžarska, Irska, Italija, Litva, Latvija, Poljska, Portugalska, Romunija, Slovenija, Španija, Švedska, Francija, Nizozemska, Belgija, Luksemburg, Slovaška, Norveška, Turčija, Republika Severna Makedonija, Islandija, Črna gora in Srbija. Delovanje Nacionalnega kompetenčnega centra SLING sofinancira RS, Ministrstvo za visoko šolstvo, znanost in inovacije.

