Strateški Tehnički Izvještaj: Arhitektura AI Studija na Infrastrukturi HPE ProLiant DL380 Gen10

1. Izvršni Sažetak i Arhitektonska Vizija

1.1. Strateški Mandat i Opseg Projekta

Uloga Strateškog Arhitekta za AI Agent (AI Studio) zahtijeva holistički pristup dizajnu infrastrukture koji nadilazi puko nabrajanje hardverskih komponenti. Cilj ovog dokumenta je pružiti iscrpnu, dubinsku analizu implementacije računalnog ekosustava temeljenog na poslužiteljima **HPE ProLiant DL380 Gen10**, koji će služiti kao kralježnica novog AI Studija. S obzirom na složenost generativnih AI modela, potrebu za masivnim paralelnim procesiranjem i zahtjeve za niskom latencijom u pristupu podacima, ova infrastruktura ne smije biti tretirana kao generički IT resurs, već kao precizno ugođen instrument visokih performansi (HPC - High Performance Computing).

Analiza se temelji na dostavljenoj inventurnoj listi [1] i specifičnim tehničkim zahtjevima korisnika. Fokus je na dvama kritičnim čvorovima (nodovima):

1. **Server 1 (Operativan):** Postojeći HPE DL380 Gen10 konfiguriran s fokusom na heterogeno procesiranje (CPU + GPU) i hibridnu pohranu podataka. Ovaj server nosi teret trenutnih operacija, koristeći Intel Xeon Gold 6138 procesore i mješavinu NVMe i SAS diskova.
2. **Server 2 (Buduća implementacija - Veljača 2026.):** Planirani čvor optimiziran za propusnost podataka (Throughput-optimized), temeljen na Intel Xeon Gold 6230R procesorima i potpunoj NVMe arhitekturi (8x NVMe bay), dizajniran da eliminira I/O uska grla (bottlenecks) u treniranju modela.

Ovaj izvještaj služi kao definitivan tehnički priručnik (Blueprint), adresirajući kritične točke neuspjeha (SPOF), termalne izazove integracije pasivnih i aktivnih GPU akceleratora u 2U kućištu, te forenzičku analizu naponskih kabela nužnih za stabilan rad NVIDIA Quadro RTX 8000 kartica.

1.2. Profil Opterećenja AI Studija (Workload Profile)

Razumijevanje prirode AI opterećenja ključno je za arhitektonske odluke donijete u ovom izvještaju. Za razliku od transakcijskih baza podataka ili web servisa, AI Studio generira opterećenja karakterizirana:

- **Tenzorskim operacijama visokog intenziteta:** Kontinuirano korištenje FP32, FP16 i INT8 preciznosti na GPU jezgrama zahtijeva stabilno napajanje i hlađenje jer GPU često radi na

100% TDP-a satima ili danima.
- **Memorijskom propusnošću (Memory Bandwidth Bound):** Veliki jezični modeli (LLM) i generativni modeli ovise o brzini kojom se podaci mogu prenijeti iz sistemske memorije (RAM) u VRAM GPU-a. Ovdje uloga PCIe stazica (Lanes) i NUMA (Non-Uniform Memory Access) topologije postaje kritična.
- **Latencijom pohrane:** Proces treniranja ("Training") i finog podešavanja ("Fine-tuning") modela zahtijeva brzo učitavanje datasetova ("Data Loading"). Tradicionalni SAS diskovi, prisutni u inventaru [1], predstavljaju usko grlo, stoga je pravilna implementacija NVMe sloja (Tiering) imperativ.

---

2. Platformska Arhitektura: HPE ProLiant DL380 Gen10 – Dubinska Analiza

HPE ProLiant DL380 Gen10 nije statična komponenta; to je modularna platforma čije performanse drastično ovise o internoj konfiguraciji. Za potrebe AI Studija, moramo dekonstruirati server na razini mehaničkih, termalnih i električnih podsustava.

2.1. Mehanička i Termalna Fizika Kućišta (Chassis Dynamics)

Prema dostavljenom opisu [1], Server 1 je konfiguriran s kavezom za diskove "16 SFF + 2 SFF". Ova konfiguracija ima značajne implikacije na protok zraka (Airflow Impedance), što je kritično za hlađenje GPU-a.

2.1.1. Problem Impedancije Zraka u 2U Formatu

U standardnom 2U poslužitelju, zrak se kreće linearno od prednje strane prema stražnjoj (Front-to-Back).

- **Opstrukcija diskovima:** Konfiguracija s 16 SFF (Small Form Factor) diskova pokriva gotovo cijelu prednju usisnu površinu servera. Iako SFF kavezi imaju proreze za zrak, popunjavanje tih utora mehaničkim diskovima (kako sugerira inventar [1] s velikim brojem SAS HDD-ova) stvara zonu visokog statičkog tlaka na usisu.
- **Posljedica za GPU:** NVIDIA Quadro RTX 8000 (osobito pasivna verzija) ovisi o razlici tlaka (ΔP) između prednje i stražnje strane servera kako bi zrak prošao kroz gusto raspoređena rebra hladnjaka. Ako je usis blokiran diskovima, ventilatori moraju raditi na znatno višim okretajima (RPM) kako bi kompenzirali otpor, što povećava potrošnju energije i vibracije.

2.1.2. Kohabitacija Aktivnog i Pasivnog Hlađenja

Jedan od najsloženijih zahtjeva ovog projekta je istovremena instalacija jedne **aktivne** (s vlastitim ventilatorom) i jedne **pasivne** (bez ventilatora) NVIDIA Quadro RTX 8000 kartice.

- **Pasivna RTX 8000:** Ova kartica nema vlastiti mehanizam za pomicanje zraka. Ona se

oslanja isključivo na *sistemske ventilatore* servera. Da bi ovo funkcioniralo, HPE DL380 Gen10 mora biti opremljen **HPE High Performance Fan Kitom** (obično 6 ventilatora visokih performansi). Standardni ventilatori nisu sposobni generirati dovoljan CFM (Cubic Feet per Minute) pritisak kroz pasivni hladnjak GPU-a kada je usis blokiran diskovima. Također, nužna je ugradnja plastičnih usmjerivača zraka (Air Baffles) iz **GPU Enablement Kita**, koji fizički prisiljavaju zrak da prolazi kroz PCIe kaveze umjesto da ih zaobilazi.

- **Aktivna RTX 8000:** Ova kartica ima radijalni (blower) ventilator koji ispuhuje zrak kroz stražnji panel. Iako je termalno neovisnija, njeno fizičko tijelo u PCIe slotu može blokirati protok zraka prema susjednim komponentama.
- **Arhitektonska Preporuka:**
  - **Lokacija Pasivne Kartice:** Mora biti postavljena u PCIe Riser kavez koji ima najdirektniju zračnu liniju s ventilatorima, idealno u sekundarni riser (Secondary Riser) poravnat s CPU 2, uz uvjet da su instalirani odgovarajući usmjerivači zraka.
  - **Lokacija Aktivne Kartice:** Fleksibilnija je, ali treba izbjegavati postavljanje neposredno ispred pasivne kartice kako ne bi "krala" svježi zrak ili ispuhivala toplinu prema pasivnom hladnjaku.

2.2. PCIe Riser Arhitektura i Raspodjela Stazica (Lane Distribution)

HPE DL380 Gen10 koristi modularni sustav PCIe podizača (Risers).

- **Primary Riser:** Povezan je na CPU 1.
- **Secondary Riser:** Povezan je na CPU 2.
- **Tertiary Riser:** Povezan je na CPU 2.

Strateški Imperativ: Za AI Studio, ključno je izbalansirati opterećenje između dva procesora.

- Ako se obje RTX 8000 kartice stave na isti riser (npr. Primary), sav GPU promet prolazit će kroz CPU 1. Ako aplikacija trči na CPU 2, podaci moraju putovati preko UPI (Ultra Path Interconnect) veze između procesora, što uvodi latenciju i troši propusnost UPI linka potrebnog za koherenciju memorije cachea.
- **Konfiguracija:** Jedna RTX 8000 mora biti na Primary Riseru (CPU 1), a druga na Secondary Riseru (CPU 2). Ovo osigurava tzv. **NUMA balans**. Svaki CPU ima izravan pristup jednoj GPU kartici, maksimizirajući propusnost za paralelne procese treniranja.

---

3. Procesorska Logika: Analiza i Kompatibilnost CPU Platformi

Korisnik raspolaže dvjema generacijama Intel Xeon Scalable procesora. Iako dijele isti fizički socket (LGA 3647), njihove mikroarhitekture imaju različite karakteristike koje su presudne za AI opterećenja.

3.1. Server 1: Intel Xeon Gold 6138 (Skylake-SP)

- **Specifikacije:** 20 jezgri / 40 dretvi, 2.00 GHz Base, 125W TDP.[1]
- **Arhitektura:** Prva generacija Scalable (Skylake). Koristi Mesh interkonekciju jezgri umjesto prstenaste (Ring), što smanjuje varijaciju u latenciji, ali blago povećava prosječnu latenciju L3 predmemorije.
- **AVX-512 Penalty (Kritično za AI):** Skylake arhitektura poznata je po agresivnom smanjenju takta (Downclocking) kada se izvršavaju AVX-512 instrukcije, koje su česte u bibliotekama za linearnu algebru (NumPy, TensorFlow na CPU). Iako je bazni takt 2.0 GHz, pri punom AVX-512 opterećenju, takt može pasti na 1.6 GHz ili niže kako bi se održao TDP od 125W.
- **Zaključak:** Ovaj procesor je izvrstan za orkestraciju (Kubernetes master), prijenos podataka i opće zadatke, ali nije idealan za "CPU-only" inferenciju modela zbog pada frekvencije. Njegova uloga u Serveru 1 je primarno hranjenje GPU-ova podacima.

3.2. Server 2: Intel Xeon Gold 6230R (Cascade Lake Refresh)

- **Specifikacije:** 26 jezgri / 52 dretve, 2.10 GHz Base, 150W TDP.
- **Značenje sufiksa "R":** "R" označava "Refresh" seriju druge generacije. Ovi procesori nude performanse slične Platinum seriji po cijeni Gold serije.
- **DL Boost (VNNI):** Ključna prednost 6230R nad 6138 je podrška za **Intel Deep Learning Boost (VNNI - Vector Neural Network Instructions)**. Ove instrukcije omogućuju spajanje triju operacija (FMA) u jednu pri radu s INT8 preciznošću. To dramatično ubrzava inferenciju kvantiziranih AI modela na samom CPU-u.
- **Kompatibilnost Ploče:** Iako DL380 Gen10 podržava Gen2 procesore, serija "R" lansirana je kasnije. **Nužno je provjeriti reviziju matične ploče i verziju BIOS-a.** Starije revizije matičnih ploča (prije 2019.) možda neće moći isporučiti stabilan napon potreban za "Refresh" modele unatoč BIOS nadogradnji. Potrebno je provjeriti QuickSpecs za specifičnu podršku matične ploče (System Board Spare Part Number).

3.3. Interoperabilnost i Miješanje

Izričito je zabranjeno miješati Xeon 6138 i Xeon 6230R u istom serveru. UPI linkovi neće se sinkronizirati zbog različitih frekvencija i protokola upravljanja energijom. Plan odvajanja u Server 1 i Server 2, kako je navedeno u zadatku, je tehnički ispravan i jedini moguć.

---

4. Grafički Akceleratori: NVIDIA Quadro RTX 8000 Integracija

NVIDIA Quadro RTX 8000 sa 48GB GDDR6 memorije predstavlja srce ovog AI Studija. Njena velika količina VRAM-a omogućuje učitavanje masivnih modela (npr. Llama-3-70B s

kvantizacijom) bez potrebe za sporim prebacivanjem na sistemski RAM (Offloading).

4.1. Instalacija i Podrška

- **Dimenzije:** RTX 8000 je kartica pune visine i pune duljine (FHFL - Full Height, Full Length).
- **Zahtjev za Riserom:** Za instalaciju su potrebni **HPE DL380 Gen10 x16 Riseri**. Standardni riseri često dolaze s x8 slotovima. Instalacija x16 kartice u x8 slot fizički je moguća (ako je slot otvoren) ili električno ograničavajuća, što bi prepolovilo propusnost prema CPU-u.
- **Secure Boot:** Prilikom instalacije GPU-a, potrebno je provjeriti postavke BIOS-a za Secure Boot. Ponekad nepotpisani upravljački programi (NVIDIA proprietary drivers na Linuxu) mogu stvarati probleme ako Secure Boot nije ispravno konfiguriran ili ako ključevi nisu upisani (MOK - Machine Owner Key).

4.2. Softverska Podrška i Virtualizacija

S obzirom na to da je riječ o "AI Agent" studiju, vjerojatno će se koristiti kontejnerizacija (Docker/Kubernetes) s NVIDIA Container Toolkitom.

- **vGPU (Virtual GPU):** Ako planirate dijeliti GPU resurse između više virtualnih mašina, Quadro RTX 8000 podržava NVIDIA vGPU softver (zahtijeva licencu). Za *bare-metal* Linux instalacije (npr. Ubuntu Server), licenca nije potrebna za Compute mode.

---

5. KRITIČNO: Forenzika Napajanja i Kabliranja (Dubinska Analiza)

Ovo poglavlje adresira najkritičniju točku upita: verifikaciju kabela **869820-001** i **869805-001** za napajanje RTX 8000. Pogreška ovdje može rezultirati trajnim oštećenjem matične ploče risera ili samog GPU-a.

5.1. Energetski Profil NVIDIA Quadro RTX 8000

- **TDP (Thermal Design Power):** 260 W.
- **Distribucija Snage:**
  - PCIe Slot (Matična ploča/Riser): Pruža do 75 W.
  - Potrebno vanjsko napajanje: 260 W - 75 W = 185 W.
- **Konektori na GPU:** Referentni dizajn RTX 8000 obično zahtijeva **1x 8-pin PCIe + 1x 6-pin PCIe** konektore za napajanje. (Neke OEM varijante mogu imati drugačiji raspored, npr. 1x 8-pin EPS stil, ali to je rijetko za Quadro seriju).

5.2. Analiza HPE Riser Konektora

HPE DL380 Gen10 riseri ne koriste standardne ATX/PCIe konektore na strani risera. Koriste

specifične **10-pinske** konektore (često označene kao "GPU Power").

5.3. Detaljna Verifikacija Kabela

Ovdje leži ključna opasnost. Korisnik posjeduje dva različita kabela.

Kabel 1: P/N 869820-001

- **Službeni Opis:** HPE Gen10 GPU Power Cable, 8-pin to 8-pin.
- **Fizička Konfiguracija:** 10-pin (Riser strana) na 8-pin (GPU strana).
- **Namjena:** Ovaj kabel je dizajniran da prenese do 150W (ili više, ovisno o debljini vodiča, AWG) na jedan 8-pinski konektor.
- **Kompatibilnost s RTX 8000:** Ovaj kabel će fizički pasati u 8-pinski utor na RTX 8000.
    - *Problem:* RTX 8000 treba i dodatni 6-pinski konektor. Sam ovaj kabel **NIJE DOVOLJAN** za pokretanje kartice.

Kabel 2: P/N 869805-001

- **Službeni Opis:** HPE Gen10 GPU Power Cable, 8-pin to 6+2 pin.
- **Fizička Konfiguracija:** 10-pin (Riser strana) na 6+2 pin (GPU strana).
- **Namjena:** Univerzalni kabel za kartice koje trebaju 6-pin ili 8-pin napajanje.
- **Kompatibilnost s RTX 8000:** Ovaj kabel može napajati ili 8-pinski ili 6-pinski utor na kartici.

5.4. Zaključak o Kabelima i Upozorenje (Rizik od Požara/Kvara)

KRITIČNI NALAZ: Korisnik vjerojatno ima jedan kabel po kartici (jedan za pasivnu, jedan za aktivnu). Ovo je NEDOVOLJNO.
NVIDIA Quadro RTX 8000 (260W) ne može raditi samo s jednim 8-pinskim kabelom. Ako spojite samo jedan kabel, kartica će detektirati nedostatak napona na drugom portu i odbiti se pokrenuti (crvena LED dioda) ili će raditi u "Safe Mode" s drastično smanjenim performansama, a vodiči u kabelu mogu se pregrijati jer pokušavaju povući 185W kroz konektor specificiran za 150W.
Potrebna Konfiguracija:
Svaka RTX 8000 kartica mora biti spojena s DVA izvora napajanja s risera (ili koristiti Y-splitter visoke kvalitete ako riser podržava tu amperažu po portu, što kod HPE-a često nije slučaj za 260W kartice).
Tablica 1: Matrica Ispravnog Kabliranja za RTX 8000 u DL380 Gen10

Komponenta	Potreban Ulaz	HPE Kabel (Opcija A - Dual Cable)	HPE Kabel (Opcija B - Y-Cable Kit)
RTX 8000 Port 1	8-pin PCIe	Kabel P/N 869820-001	Dio kompleta 871830-B21

		(spojen na Riser Port 1)	
RTX 8000 Port 2	6-pin PCIe	Kabel P/N 869805-001 (spojen na Riser Port 2)	Dio kompleta 871830-B21
Ukupno kabela po GPU		**2 komada**	**1 komad (bifurkiran)**
Napomena		Troši 2 porta na riseru.	Zahtijeva provjeru max struje po portu risera.

Akcijski Plan:

1. Provjerite stražnju stranu RTX 8000 kartica. Imaju li 8-pin + 6-pin ulaze? (99% vjerojatnost: DA).
2. Provjerite Riser kartice u serveru. Svaki Riser obično ima 2x 10-pin power izlaza.
3. Za **jednu** karticu trebate iskoristiti **oba** porta na riseru (jedan kabel u 8-pin, drugi u 6-pin na GPU).
4. Ako imate samo navedena dva kabela (ukupno), **ne možete spojiti obje kartice**. Nedostaju vam još dva kabela.

---

6. Podsustav Pohrane Podataka: SSD i NVMe Strategija

Pohrana podataka u AI studiju dijeli se na tri razine (Tiering): "Hot" (podaci koji se trenutno obrađuju), "Warm" (često korišteni datasetovi) i "Cold" (arhiv).

6.1. Server 1: Hibridna SSD/SAS Implementacija

Prema inventaru [1], Server 1 ima "16SFF+2SFF" i kontroler "P816i-a". Također su navedeni "NVMe SSD PM1725 1.6TB (3x)".

- **Izazov NVMe u SFF kavezu:** Standardni SFF kavezi na DL380 Gen10 su spojeni na SAS kontroler (P816i-a). SAS kontroler **ne može** upravljati NVMe diskovima. NVMe diskovi zahtijevaju izravnu PCIe vezu.
- **Rješenje:** Da bi PM1725 diskovi (ako su u U.2 2.5" formatu) radili u Serveru 1, morate imati instaliran **HPE DL380 Gen10 NVMe Enablement Kit**. To podrazumijeva specijalne kabele koji idu s matične ploče (ili NVMe risera) izravno na stražnju stranu kaveza diskova

(Backplane). Kavez mora podržavati NVMe (tzv. U.3 ili Combo backplane).

- **Preporuka za Diskove (Server 1):**
  - **OS (Boot):** 2x 400GB ili 600GB SAS SSD u RAID 1 (Mirror) na P816i-a kontroleru. Ovo osigurava redundanciju operativnog sustava.
  - **Cache/Scratch:** 3x PM1725 NVMe konfigurirani kao LVM Stripe (RAID 0 ekvivalent) ili ZFS Cache. Ovdje se drže podaci koji se trenutno "hrane" u GPU. Brzina ovih diskova (6GB/s read) je ključna da GPU ne čeka podatke.
  - **Data Lake (Warm):** Ostatak SAS slotova popuniti s 1.2TB 10K SAS HDD-ovima u RAID 6 polju za pohranu modela i checkpointa.

6.2. Server 2: NVMe Only (Budućnost)

Server 2 dolazi s "8x NVMe bays". Ovo je superiorna konfiguracija za AI.

- **VROC (Virtual RAID on CPU):** Budući da nema hardverskog RAID kontrolera za NVMe, koristi se Intel VROC.
- **Zahtjev:** Za kreiranje RAID polja (npr. RAID 5 ili 10) na NVMe diskovima, potrebna je **Intel VROC Premium Licenca (Hardware Key)** koja se uštekava na matičnu ploču. Bez ovoga, diskovi su vidljivi samo kao pojedinačni "Passthrough" uređaji ("JBOD").

---

7. Mrežna Arhitektura i Huawei Integracija

Mreža povezuje računalne čvorove međusobno i s vanjskim svijetom. Huawei S5735-L24T4X-A1 je gigabitni switch s 4x 10G uplink portovima.

7.1. Fizička Povezivost (Layer 1)

- **Switch Portovi:** 4x 10GE SFP+.
- **Server Kartice:** DL380 Gen10 obično koristi FlexibleLOM (FLR) adaptere, npr. HPE 562FLR-SFP+ (Intel X710 čipset).
- **Medij:** Inventar navodi "SFP+ 10GBASE-SR (10x)". Ovo su optički primopredajnici (Transceivers).
  - *Kompatibilnost:* Huawei switchevi su generalno tolerantni, ali HPE mrežne kartice mogu biti izbirljive. Idealno je koristiti HPE brandirane optike u serveru i Huawei brandirane u switchu.
  - *Kabel:* Potreban je **OM3 (Aqua)** ili **OM4 (Violet)** višemodni optički kabel (LC-LC konektori).
- **Alternativa:** Ako je udaljenost manja od 5-7 metara, preporučuje se korištenje **DAC (Direct Attach Copper)** kabela. Oni su jeftiniji, troše manje struje i imaju manju latenciju od optike jer nema konverzije signala iz električnog u svjetlosni i natrag.

7.2. Logička Konfiguracija (Layer 2)

Za AI Studio, propusnost od 10Gbps po serveru je minimum. Preporučuje se agregacija

linkova.

- **LACP (Link Aggregation Control Protocol - 802.3ad):**
  - Spojite dva 10G porta sa Servera 1 na dva 10G porta na Huawei switchu.
  - Konfigurirajte **Eth-Trunk** u LACP modu na Huawei strani.
  - Na strani OS-a (Linux Bond) koristite mode=4 (802.3ad).
  - *Rezultat:* 20Gbps ukupne propusnosti i redundancija u slučaju kvara jednog kabela.

7.3. Buffer Limitacije (Flow Control)

Huawei S5735 je "Campus" switch, ne "Data Center" switch. Ima relativno plitke buffere za pakete.

- **Rizik:** Prilikom distribuiranog treniranja (ako se koriste oba servera), dolazi do "Micro-burst" prometa. Plitki bufferi mogu dovesti do odbacivanja paketa (TCP Drops), što drastično ruši performanse AI treninga.
- **Mitigacija:** Omogućite **Global Flow Control (802.3x)** na switchu i na mrežnim karticama servera. To omogućuje slanje "Pause" okvira kako bi se spriječilo prelijevanje buffera.

---

8. Zaključak i Operativni Hodogram

Arhitektura AI Studija na HPE DL380 Gen10 platformi je robustna, ali zahtijeva preciznu egzekuciju kako bi se izbjegle zamke nekompatibilnosti.

Sažetak rizika i preporuka:

1. **Napajanje GPU-a:** Trenutni broj kabela je nedostatan za siguran rad dviju RTX 8000 kartica. Potrebna je hitna nabavka dodatnih HPE GPU naponskih kabela ili kompleta **871830-B21**.
2. **Hlađenje:** Miješanje pasivnih i aktivnih GPU-a u kućištu punom diskova zahtijeva **High Performance Fan Kit** i postavljanje BIOS profila na "Increased Cooling".
3. **Procesor:** Server 1 (Skylake) koristite za pripremu podataka, a Server 2 (Cascade Lake R) rezervirajte za tešku inferenciju i vektorske operacije zbog VNNI podrške.
4. **Mreža:** Iskoristite LACP na 10G portovima Huawei switcha kako biste osigurali propusnost za NVMe-over-Fabric promet u budućnosti.

Implementacijom ovih smjernica, HPE ProLiant DL380 Gen10 transformirat će se iz standardnog poslužitelja u visokoučinkovitu platformu sposobnu podržati najzahtjevnije izazove moderne umjetne inteligencije.

Works cited

1. Inventar_Opreme - Copy.xlsx