

A mesterséges intelligencia adatközponti infrastruktúrájának rendszerszintű kihívásai és megoldási lehetőségei

System-Level Challenges and Potential Solutions for Artificial Intelligence Data Center Infrastructure

Dr. GÁL Zoltán

Debreceni Egyetem, Informatikai Kar, Debrecen, Magyarország
4028 Debrecen, Kassai út 26., Magyarország, gal.zoltan@inf.unideb.hu, <http://www.inf.unideb.hu>

Abstract

The rapid scaling of artificial intelligence (AI) has made data centres a critical physical infrastructure of the digital economy. Based on the supplied professional materials and verified scientific, standards-based and industrial sources, this paper systematises the key challenges of AI data centres, including electricity supply and grid-development lead times, cooling of high-density GPU systems, water and carbon efficiency, communication and storage bottlenecks, and the different capacity profiles of training and inference. It proposes an energy-to-solution approach that complements facility metrics such as PUE, WUE and CUE with service-level outcomes. International and Hungarian examples illustrate liquid cooling, waste-heat recovery, flexible energy portfolios and hybrid HPC-cloud infrastructures. The paper concludes that sustainable and sovereign AI capacity requires an integrated energy, network, data and computing masterplan.

Keywords: artificial intelligence; data centre; energy efficiency; liquid cooling; digital sovereignty

Kivonat

A mesterséges intelligencia (AI) gyors skálázódása az adatközpontokat a digitális gazdaság kritikus fizikai infrastruktúrájává tette. A tanulmány a feltöltött szakmai anyagokra és ellenőrzött tudományos, szabványügyi, illetve ipari forrásokra támaszkodva rendszerezi az AI-adatközpontok legfontosabb kihívásait. Vizsgálja a villamosenergia-ellátás és a hálózatfejlesztés eltérő időhorizontját, a nagy teljesítménysűrűségű GPU-rendszerek hűtését, a víz- és karbonhatékonyságot, a hálózati kommunikáció és adattárolás szűk keresztmetszeit, valamint az AI-oktatás és következtetés eltérő kapacitásigényeit. A cikk a PUE, WUE és CUE létesítményi mutatókat kiegészítő, szolgáltatási eredményhez kötött „energy-to-solution” szemléletet javasol. Nemzetközi és hazai példákon keresztül bemutatja a folyadékűtés, a hulladékhő-hasznosítás, a rugalmas energiaportfólió és a hibrid HPC-felhő infrastruktúra lehetőségeit. Következtetések szerint a fenntartható és szuverén AI-kapacitás integrált energia-, hálózati, adat- és számítási mestertervet igényel.

Kulcsszavak: mesterséges intelligencia; adatközpont; energiahatékonyság; folyadékűtés; digitális szuverenitás

1. BEVEZETÉS

A mesterséges intelligencia (AI) gyors fejlődése alapvetően átalakítja a digitális infrastruktúrával szembeni követelményeket. Az AI-szolgáltatás nem pusztán szoftveres jelenség: működése nagy teljesítményű gyorsítók, nagy sávszélességű hálózatot, többretegű tárolást, folyamatos villamosenergia-ellátást és egyre fejlettebb hűtést igényel. Modern megközelítésben az adatközpont öt összekapcsolt ökoszisztéma – adat, hálózat, számítás, szolgáltatás és energia – integrációs pontjaként létezik. E rendszerszemlélet indokoltságát az IEA (International Energy Agency) legfrissebb adatai is alátámasztják: 2025-ben az adatközpontok villamosenergia-fogyasztása mintegy 17%-kal nőtt, míg az AI-központú létesítmények fogyasztása 2030-ig a 2025-ös szint közel háromszorosára emelkedhet [1]. A fizikai infrastruktúra bővítése ugyanakkor eltérő időléptékeken zajlik: egy adatközpont tipikusan 1–3 év alatt megépíthető, miközben az új hálózati infrastruktúra tervezése és kiépítése 5–15 évet is igénybe vehet [2]. A tanulmány célja a feszültségek rendszerszintű feltárása, a releváns szakirodalom szintézise, valamint olyan műszaki és stratégiai megoldások azonosítása, amelyek egyszerre javítják a számítási hasznosulást, az energia- és vízhatékonyságot, az üzemfolytonosságot és a szuverenitást. Külön figyelmet kap az AI-oktatás (tréning) és a következtetés (inferencia) eltérő terhelési profilja:

előbbi hosszú, párhuzamosítható futásokkal, utóbbi folyamatos, késleltetésérzékeny kiszolgálással terheli az infrastruktúrát. A kapacitástervezésnek ezért az AI teljes életciklusát kell követnie, nem csupán a gyorsítók névleges teljesítményét.

A második szekcióban a témával kapcsolatos fontosabb szakirodalom nézőpontját mutatjuk be. A harmadik fejezetben az AI és az adatközpontok köré csoportosuló ökoszisztéma problémáit és megoldásait tárgyaljuk. Az utolsó rész összefoglalja a tárgyalt témakört és középtávú következtetéseket fogalmaz meg.

2. HIVATKOZOTT IRODALMI ÁTTEKINTÉS

Az adatközpontok energiaigényének megítélése az elmúlt években jelentősen változott. Masanet és munkatársai 2020-ban arra hívták fel a figyelmet, hogy a korábbi évtizedben a szerver-, virtualizációs és létesítményi hatékonyság javulása mérsékelte a globális villamosenergia-igény növekedését, ezért a számítási volumenből közvetlenül nem vezethető le az energiafogyasztás [4]. Az új generatív AI-hullám azonban új skálázási pályát nyitott. Az IEA 2025-ös „Energy and AI” jelentése 2030-ra mintegy 945 TWh adatközponti fogyasztást vetített előre [3], a 2026-os frissítés pedig 2025 és 2030 között közel kétszereződő teljes adatközponti igényt és az AI-fókuszú központok fogyasztásának mintegy háromszoros növekedését jelzi [1]. Ezzel párhuzamosan a hálózati csatlakozási sorok világszerte rekordszintre emelkedtek, miközben az adatközponti beruházások időigénye jóval rövidebb, mint a nagyfeszültségű hálózat fejlesztése [2].

A hardversűrűség növekedése a kommunikáció és a hőelvezetés problémáját is kielezi. Az NVIDIA GB200 NVL72 egy számítástechnikai szekrényben (rack) 72 GPU-t kapcsol össze. Az NVLink tartomány összesített sávszélessége 130 TB/s [5], a referencia-rack hűtési kapacitásigénye pedig 120 kW nagyságrendű [6]. Ez a teljesítménysűrűség a közvetlen folyadékűtést a nagy sűrűségű AI-rendszerek egyik alaptermékéjévé teszi. Wang és munkatársainak áttekintése a DTP (Direct-To-Chip), immerziós és más folyadékűtési megoldások energiahatékonysági és hulladékhő-hasznosítási előnyeit összegzi [11]. A hulladékhő energetikai újrahasznosításának lehetőségeit Yuan és munkatársai részletesen rendszerezik, külön kiemelve a távhő- és épületfűtési alkalmazásokat [12].

A fenntarthatóság értékelésében önmagában a PUE (Power Usage Effectiveness) jellemző nem elegendő. A PUE, WUE (Water Usage Effectiveness) és CUE (Carbon Usage Effectiveness) mutatókat nemzetközi szabványok rögzítik [8–10], míg Lei és munkatársai szerint a vízfelhasználást a hűtési mód, az éghajlat, a villamosenergia-mix és a szerverhatékonyság együttesen határozza meg [13]. A számítási teljesítmény oldalán az MLCommons konzorcium energia-mérési protokolljai a teljes rendszer falnál mért AC-fogyasztását kapcsolják konkrét benchmark-terhelésekhez [7]. A szakirodalom közös tanulsága, hogy a fenntartható AI-infrastruktúra nem egyetlen mérőszámmal, hanem az „energy-to-solution” szemlélettel értékelhető: a feladatminőséget, késleltetést, átvitelt, vizet és energiát együtt kell optimalizálni. A gyakorlati demonstrációk ezt a rendszerszemléletet erősítik: a JUPITER melegvizes hűtését hulladékhő-visszanyerésre tervezték [14], a LUMI hidroenergiát és városi távhő-integrációt kapcsol össze [15], míg a Stargate Norway zárt körű direct-to-chip hűtéssel és megújuló villamos energiával készül [16].

3. AI ÖKOSZISZTÉMA PROBLÉMÁK FELVETÉSE ÉS MEGOLDÁSOK

Öt szempont szerint fogjuk megvizsgálni az adatközpontok tervezésével, összehangolt működtetésével kapcsolatos kihívásokat, amelyeket az AI speciális igényei aktuálisan támasztanak a szolgáltatók felé. Fontos szerepet kap a környezet fizikai adottsága, illetve az előfizetők felhasználási szokásai is.

3.1. Energiaellátás és hálózati csatlakozás

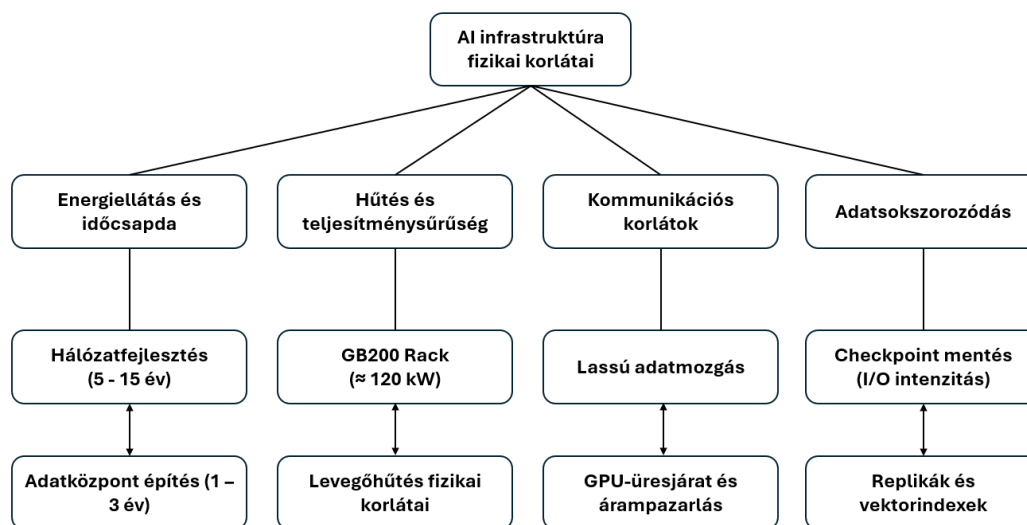
Az AI-ökoszisztéma fenntartható skálázásának első problémája az energiaellátás és az energiahálózati csatlakozás időbeli aszimmetriája. A feltöltött anyagokban szereplő „időcsapda” lényege, hogy a számítási infrastruktúra lényegesen gyorsabban telepíthető, mint ahogyan az energiahálózat megerősíthető. A megoldás ezért nem kizárólag új termelőkapacitás építése, hanem a telephely és az energiarendszer közös mesterteve. A fázisos adatközpont-építés, a nem merev hálózati csatlakozások, a keresletoldali válasz, a tréningfeladatok időbeli áthelyezése, valamint az akkumulátoros tárolás együttesen csökkenthetik a csúcsterhelést és gyorsíthatják a ténylegesen használható kapacitás rendelkezésre állását [2]. Az energiaportfólióban a gyorsan telepíthető nap- és szélenergia mellett szükség van kiszámítható, alacsony szénintenzitású forrásokra is. Ezt a tendenciát jól jelzi a Microsoft és a Constellation Energy cégek hosszú távú nukleáris megállapodása [20]. Az ilyen portfólióelemek nem helyettesítik a hálózatfejlesztést, hanem a 24/7 ellátás kockázatát mérséklék.

Az energiahálózati integráció technikai oldalán az AI-terhelés két időskáláját külön kell választani. A GPU-k és a tápegységek milliszekundumos–másodperces teljesítményváltozásai lokálisan, UPS-sel, akkumulátoros energiatárolóval és rack- vagy klasterszintű teljesítménysapka (*power cap*) alkalmazásával simíthatók, míg a perces–órás rugalmasságot az ütemező képes kihasználni. Mivel a tréningfeladatok egy része megszakítható, köztes állapotmentésből (checkpoint) újraindítható vagy alacsonyabb karbonintenzitású időablakba tolható, ezért célszerű a stabil és a rugalmas számítási igényt külön kapacitáosztályként kezelni. A karbonintenzitás-alapú ütemezésnél a határidőt, a checkpoint-költséget, az adatmozgatás energiaigényét és a hálózati korlátokat együtt szükséges figyelembe venni. Ellenkező esetben a lokálisan kedvező időzítés a rendszer más részén többletterhelést vagy késleltetést okozhat.

3.2. Hűtés, vízhasználat és hulladékhő

A nagy teljesítménysűrűségű rackeknél a kizárólagos léghűtés műszaki és energetikai korlátokba ütközik; a közvetlen folyadékhűtés magasabb hőmérsékletű hőhordozót tesz elérhetővé, ami javítja a hővisszanyerés lehetőségét [11–12]. A cél nem egyszerűen a PUE minimalizálása. A vízstresszes térségekben a WUE, karbonintenzív villamosenergia-mix esetén pedig a CUE válhat domináns korláttá [8–10,13]. A hulladékhő akkor válik valódi másodlagos terméké, ha a telephely közelében stabil hőigény és megfelelő hőmérsékletszint áll rendelkezésre. A LUMI 100%-ban vízenergiára támaszkodó adatközpontja hulladékhővel a Kajaani, finn város távhőigényének akár ötödét képes fedezni [15]. A JUPITER melegvízes hűtése a németországi Jülich campusz fűtési hálózatába integrálható volt [14]. Magyar példaként a Komondor közvetlen folyadékhűtése lehetővé teszi több száz kilowatt hulladékhő hasznosítását a Debreceni Sportuszoda épületének és medencéinek fűtésében [17].

A hő- és vízgazdálkodásnál kerülni kell a lokális optimumokat. A PUE javítása például növelheti a helyszíni vízhasználatot, ha intenzívebb párologtató (evaporatív) hűtéssel jár, miközben a zárt folyadékkör mellett a villamosenergia-termelés vízlábnyoma továbbra is jelentős [13]. Ebből kifolyólag telephelyenként többdimenziós célfüggvény szükséges az energia, víz, karbon, hőértékesítés és szolgáltatási minőség kompromisszumaira vonatkozóan.



1. ábra. Az AI-feldolgozás fizikai és infrastruktúra jellegű kihívásai

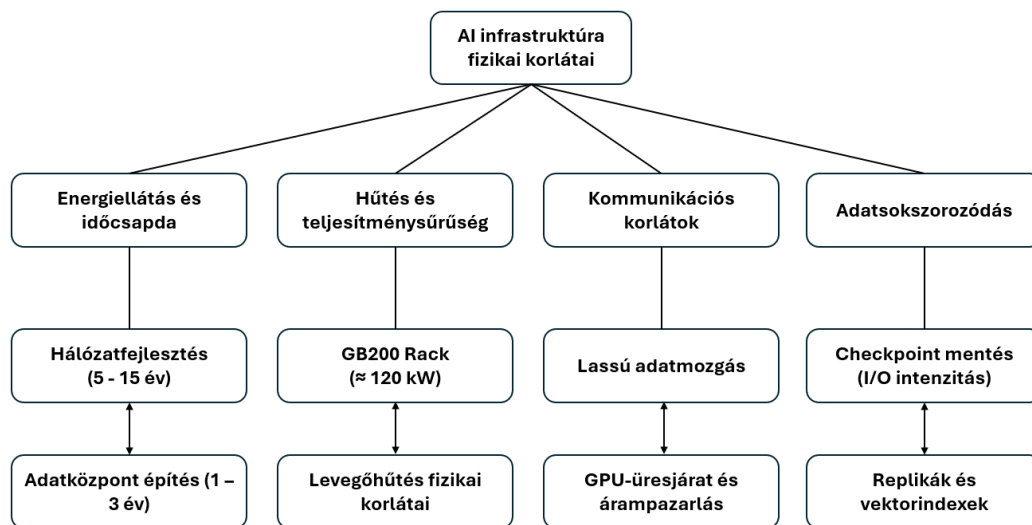
A folyadékhűtés hatékonyságát a teljes hőút határozza meg: GPU-lapka → hőközvetítő réteg → hűtőlap (cold plate) → szekunder folyadékkör → CDU (Coolant Distribution Unit) → primer létesítményi kör → szárashűtő, hűtőgép vagy hőhasznosító rendszer. Az egyes szakaszok hőellenállása, a tömegáram, a folyadék fajhője és az előremenő–visszatérő hőmérséklet-különbség együtt szabják meg a szükséges szivattyúteltjesítményt és a kinyerhető hő hőmérsékletszintjét. A magasabb visszatérő vízhőmérséklet javítja a hulladékhő hasznosíthatóságát, ugyanakkor szűkítheti a komponensek termikus tartalékát. Emiatt a hűtés optimalizálása többváltozós feladat, nem egyszerűen a ventilátorenergia kiváltása.

A korszerű rendszerekhez szivárgásérzékelés, redundáns pumpák, vízkémiai kontroll, korrózió- és galvanikus kompatibilitás, valamint rack-szintű hőmérséklet- és nyomástelemetria szükséges [11].

3.3. Kommunikáció és adattárolás

Egy AI-klaszterben a drága gyorsító akkor is fogyaszt energiát, amikor kommunikációra vagy I/O-ra vár. A racken belüli 130 TB/s-os NVLink adattovábbító (fabric) [5] önmagában nem oldja meg a több rackes skálázást: a hálózati topológiának, a modellpárhuzamosításnak és az adat-particionálásnak összehangoltnak kell lennie. A megoldási elv az adatlokalitás, vagyis az aktív adatkészlet kerüljön a számítás közelébe, a modellfelosztás kövesse a fizikai adattovábbítót, a kommunikált byte-ok száma pedig kvantálással, ritkítással és gradiens-tömörítéssel csökkenjen. Az adattárolás ugyanezt a hierarchikus logikát igényli. Az AI-életciklus a nyers adatok mellett tisztított és címkézett változatokat, keverés/funkció (shuffle/feature) réteget, sűrű checkpoint-okat, log-okat, replikákat és vektorindexeket is létrehoz. Emiatt a kapacitás tervezése nem egyszerű tárhelyszorzó, hanem az I/O-intenzitás, a másolatok száma, a megőrzési idő és a visszaállítási célidő közös optimalizálása. A forró memória rétegben HBM/DRAM/NVMe szolgálja az aktív batch-eket és KV-cache (Key-Value) gyorsítót. A meleg réteg SSD- és objektumtárat használ az adat-kontingensekhez és checkpoint-okhoz; a hideg réteg az archív, ritkán elért verziókat kezeli. A cél az, hogy a legdrágább tároló csak az aktív munkakészletet szolgálja ki.

A több ezres GPU-klaszterek kommunikációja szorosan összefügg a párhuzamosítási stratégiával. Adatpárhuzamos tanításhoz a gradiens-összegzés tipikusan „erőteljes csökkentés” műveletet igényel; tensor- és pipeline-parallelizmus esetén pont-pont adatfolyamok jelennek meg, míg Mixture-of-Experts modelleknél a mindegyik mindegyikkel (*all-to-all*) kommunikáció válhat erőteljessé. A szoftveres modellfelosztás ebből adódóan közvetlenül meghatározza a hálózat terhelési mintázatát. A tenzorok párhuzamos feldolgozása kommunikációs műveleteket épít a számítás kritikus útjába. Ugyanakkor az optimalizáló állapotok, gradiens és paraméterek particionálása csökkenti az egy GPU-ra jutó memóriaigényt, de növeli a kollektív adatcserék szerepét. Következésképpen a topológia-tudatos ütemezőnek nemcsak szabad GPU-kat, hanem megfelelően összekötött GPU-csoportokat kell kiosztania, minimalizálva a rackek és fabric-szintek közötti átlépéseket.



2. ábra. Infrastrukturális és szoftveres megoldások az AI számára

Az adatmozgás energiaigénye a GPU-n belül is meghatározó. A modern gyorsítók számítási teljesítménye gyorsabban nő, mint ahogy a memóriamozgatás költsége csökken, ezért sok AI-kernel memória- és I/O-korlátos. Az I/O-tudatos implementációk csempézéssel és lokális újrafelhasználással csökkentik a HBM és az áramköri SRAM közötti olvasási/írási műveletek számát. Ez adatközponti szinten azért fontos, mert a rövidebb GPU-foglaltsági idő és a kisebb memóriatranzakciók száma közvetlenül javíthatja a hatékonyságot. Hasonló szemlélet szükséges a tárolási pipeline-ban. Lokális NVMe memória búferszempufferrel, inkrementális checkpointtal, tömörítéssel, deduplikációval és megfelelő metaadat-szolgáltatással csökkenthető a központi objektumtárra jutó I/O-csúcs és a helyreállítási idő.

3.4. Orkesztráció és hasznos AI-kapacitás

Kiemelt problémakör a névleges kapacitás és a hasznos AI-teljesítmény különbségéből származó kihívás. Az AI-tanítás hosszú, nagy, részben áthelyezhető futásokat jelent, míg az inferencia 24/7, ingadozó és szigorú p95/p99 késleltetéshez kötött szolgáltatás. Emiatt a közös infrastruktúrán belül eltérő ütemezési politikára van szükség. A batch-ütemezés, topológia-tudatos erőforrás-kiosztás, dinamikus teljesítménysapka,

gyors checkpoint-helyreállítás, inferenciánál pedig batching, cache, modellrouting és kvantálás növelik a hasznos munkára jutó energia arányát. Az MLCommons konzorcium mérési szemlélete [7] alapján a teljes rendszerfogyasztást konkrét szolgáltatási eredményhez kell kötni. A feltöltött anyagokban javasolt a minőségkorrigált token/MWh metrika szolgáltatásszintű továbbfejlesztése.

A hatékonyság a teljes stack tulajdonsága. Hardverszinten az alacsonyabb numerikus pontosság, szoftverszinten a kvantálás, ritkítás, MoE-aktiválás, tudásdesztilláció és hatékonyabb optimalizálók csökkenthetik az egy célminőség eléréséhez szükséges műveletek számát. A mérési rendszernek ugyanakkor meg kell akadályoznia, hogy az energia-megtakarítás a pontosság, a válaszidő vagy a rendelkezésre állás romlásával „látószolagos” hatékonysággá váljon.

Az inferencia optimalizálásában a számítási teljesítmény mellett egyre gyakrabban a memória-sávszélesség és a KV-cache kapacitása a szűk keresztmetszet. Autoregresszív dekódoláskor minden aktív kérés dinamikusan növekvő kulcs-érték állapotot tart fenn. Ha ennek kezelése töredezett memóriát és nagy másolási költséget okoz, csökken az egyszerre kiszolgálható kérések száma. A lapozáselvű KV-cache-kezelés a virtuális memóriához hasonlóan kisebb blokkokban osztja ki a gyorsítómemóriát, így mérsékli a töredezettséget és javítja a dinamikus batching lehetőségét. A gyakorlati következmény az, hogy az inferencia-kapacitás méretezésénél nem elegendő a GPU FLOP/s vagy HBM-kapacitás. A kontextushossz, a batch összetétele, a cachetalálási arány és a memóriafragmentáció is magas prioritású tervezési paraméterek.

Az autoregresszív generálás soros természetét a spekulatív dekódolás részben feloldhatja. A módszer egy kisebb, gyorsabb *draft* modellel több tokenjelöltet állít elő, amelyeket a nagy modell egyetlen vagy kevés párhuzamos ellenőrzési lépésben validál. A szolgáltató folyamatos batching, prefix/KV-cache megosztással, operátorfüzióval, kvantálással és dinamikus modellroutinggal kombinálva a terhelés minőségi követelményeihez illesztheti az egy kérésre jutó erőforrást. A modellrouting lényege, hogy egyszerű kéréshez kisebb modell, nehéz kéréshez nagyobb modell rendelhető, abban az esetben, ha a választási logika mérhető minőségi küszöbkhöz kötött.

3.5. Szuverenitás és hibrid infrastruktúra

Az AI-adatközpont szuverenitási és üzemfolytonossági infrastruktúra is ugyanakkor. A redundáns energia- és hűtési utak, a többzónás replikáció, a fizikai és kiberbiztonság, valamint az adatjoghatóság mind többletkapacitást igényelnek, ezért költségüknek és energiahatásuknak mérhetőnek kell lennie. Az európai válasz a közös HPC- és AI Factory infrastruktúra: 2026-ban 19 AI Factory működtetése/kiépítése zajlik, és az EU legfeljebb hét AI Gigafactory létrehozására indított felhívást [18]. Magyarország számára a HunAIFA antenna a JUPITER AI Factoryhoz kapcsolódva teremthet hozzáférést európai exaszintű erőforrásokhoz, miközben a HUN-REN Cloud és a hazai szakértelem helyi kontrollt biztosít [19]. A fenntartható modell ezáltal hibrid: szuverén hazai HPC, európai kapacitás, felhő, szerverelhelyezés (colocation) és peremhálózati (edge) erőforrások összehangolt használata.

A szuverenitás műszaki értelmezése nem az izolációt, hanem az ellenőrizhető függőségeket jelenti. A hibrid infrastruktúrában el kell különíteni az adat-, vezérlési és kulcskezelési síkot: érzékeny adatoknál meghatározható, hogy mely joghatóságon belül történhet tárolás és feldolgozás, miközben kevésbé érzékeny feldolgozási munkák külső kapacitásra terelhetők. Ehhez egységes identitás- és jogosultságkezelés, titkosított adatút, auditálható naplózás, reprodukálható konténerképek és szoftverösszetevő-jegyzék (SBOM: Software Bill of Materials) szükségesegek.

Az üzemfolytonosságot az AI-adatközpontokban nem elegendő kizárólag létesítményi redundanciával biztosítani; a hibatartományokat a teljes szolgáltatási láncon végig kell követni. A villamos betáplálás, az UPS és a hűtési körök N+1 vagy 2N kialakítása csak akkor jelent tényleges rendelkezésre állást, ha a hálózati fabric, a tárolási réteg, a vezérlősík és a klaszterütemező is képesek a hardverösszetevők hibájának izolálására. A tervezés alapegysége ezért nem az egyedi eszköz, hanem a kiesés által érintett hibatartomány: rack, sor, terem, villamos ág, hűtési ág vagy akár teljes telephely. A kapacitástartalékok ezekhez a tartományokhoz kell méretezni úgy, hogy egy tervezett karbantartás és egy további, független meghibásodás egyidejű fennállása se okozzon szolgáltatási szint alatti teljesítményt.

Az AI-klasztareknél különösen fontos az amortizált üzem kezelése. Részleges GPU- vagy fabric-kiesés esetén az ütemezőnek újra kell csomagolnia a feladatokat, a checkpoint-ból folytatható tréninget más erőforrásokra kell helyeznie, az inferencia-forgalmat pedig az SLA megtartása mellett kell újraosztania. Ehhez a klasszikus infrastruktúra-monitoringnál részletesebb, rétegeket összekapcsoló telemetria szükséges. A villamos teljesítmény, a hőmérséklet, a folyadékáram, a GPU-k kihasználtsága, a HBM-terhelés, a hálózati torlódás, az I/O-várakozás és az alkalmazási késleltetés idősorait közös időalapon célszerű korrelálni. Így elkülöníthető, hogy egy teljesítményromlás számítási, kommunikációs, tárolási, hűtési vagy energiaoldali eredetű-e. A kapacitásmenedzsmentben ezért hasznos a digitális iker és a „mi van akkor ha” szimuláció alkalmazása. Egy

új GPU-generáció, magasabb rack-sűrűség vagy eltérő tréningprofil hatása előre becsülhető a transzformátorokra, a CDU-kra, az optikai fabricra és a hőhasznosításra. Az így kialakított, mérhető üzemeltetési modell a szuverenitás gyakorlati feltétele is, mert a kritikus függőségeket, a szükséges tartalékokat és a helyreállítási képességet nem deklarációként, hanem ellenőrizhető műszaki paraméterekként teszi kezelhetővé.

A megbízhatóság ellenőrzését rendszeres, kontrollált hibainjektálással és terheléspróbákkal is célszerű kiegészíteni. A próba során mérhető a feladatok újraütemezési ideje, a checkpoint-ból történő visszaállítás, a hálózati kerülőutak felépülése, valamint a hűtési és villamos tartalékok tényleges reakcióideje. A helyreállítási céldíót és az elfogadható adatvesztést külön paraméterezni kell a tréning-, inferencia- és adattárolási szolgáltatásokra, mert ezek hibaköltsége és időérzékenysége eltérő. Így a redundancia nem statikus tartalék, hanem folyamatosan tesztelt, mérhető szolgáltatási képesség lesz.

4. ÖSSZEFOGLALÓ ÉS VÍZIO

Az elemzés alapján az AI-adatközpont nem önálló műszaki létesítmény, hanem egymástól függő adat-, hálózati, számítási, szolgáltatási és energiarétegek közös rendszere. A következő évek fő kockázata nem pusztán a villamosenergia-fogyasztás növekedése, hanem az eltérő fejlesztési időhorizontok, a nagy rack-sűrűség, a hűtési és vízkorlátok, az adatmozgatási veszteségek, valamint a névleges és ténylegesen hasznos számítási kapacitás közötti rés. A megoldások ezért csak rendszerszinten értelmezhetők.

A jövőkép öt pillérre építhető: i) A kapacitást MW helyett szolgáltatási eredményhez kötött mutatókkal – például kWh/feladat, illetve minőség- és SLA-korrigált token/MWh – kell mérni; ii) A telephelyválasztásban a villamos hálózat, az optikai kapcsolat, a vízigény és a hulladékhő helyi piaca egyetlen döntési modellben kell kezelni; iii) A rugalmas energiaportfóliót tárolással, keresletoldali válasszal és időben áthelyezhető tréningterheléssel kell összekapcsolni; iv) A hibrid számítási ökoszisztémának a hazai HPC-t, a felhőt, a colocationt, az edge-et és az EuroHPC infrastruktúrát egységesen kell összehangolni; v) A szuverenitást az adatjoghatóság, a helyi szakértelem, a redundancia és az ellátáslánc-biztonság együttesen alapozza meg.

Magyarország számára a Komondor hulladékhő-hasznosítása és a HunAIFA európai kapcsolódása már két olyan építőelem, amelyre integrált nemzeti AI-infrastruktúra fejleszthető [17,19]. A hosszú távú cél olyan adatközponti rendszer, amely nem passzív fogyasztó, hanem a villamos- és hőrendszer rugalmas, mérhetően hatékony résztvevője, és közben megbízható, biztonságos, szuverén számítási kapacitást biztosít. E vízióban a beruházási döntés alapegysége nem a telek vagy a rack, hanem a garantált, minőségkorrigált és energiaoldalról is kiszolgálható AI-kapacitás, amely közösen optimalizálja a modell-, hálózati, ütemezési és hőhasznosítási fejlesztéseket.

KÖSZÖNETNYILVÁNÍTÁS

Ezt a munkát a Debreceni Egyetem QoS-HPC-IoT Laboratórium és a TKP2021-NKTA-34 projekt támogatta. A TKP2021-NKTA-34 számú projekt a Kulturális és Innovációs Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, a TKP2021-NKTA pályázati program finanszírozásában valósult meg. A munkát az OMAA 121öu2 projekt is támogatta.

IRODALMI HIVATKOZÁSOK

- [1] IEA (International Energy Agency) (2026): Key Questions on Energy and AI. IEA, Paris. URL: <https://www.iea.org/reports/key-questions-on-energy-and-ai> (Utolsó letöltés: 2026. 09. 07.)
- [2] IEA (International Energy Agency) (2026): Electricity 2026. IEA, Paris. URL: <https://www.iea.org/reports/electricity-2026> (Utolsó letöltés: 2026. 09. 07.)
- [3] IEA (International Energy Agency) (2025): Energy and AI. IEA, Paris. URL: <https://www.iea.org/reports/energy-and-ai> (Utolsó letöltés: 2026. 09. 07.)
- [4] Masanet, E.; Shehabi, A.; Lei, N.; Smith, S.; Koomey, J. (2020): Recalibrating global data center energy-use estimates. *Science*, 367(6481), 984–986. DOI: 10.1126/science.aba3758. URL: <https://doi.org/10.1126/science.aba3758> (Utolsó letöltés: 2026. 09. 07.)
- [5] NVIDIA (é. n.): NVIDIA GB200 NVL72 – specifications and architecture. URL: <https://www.nvidia.com/en-us/data-center/gb200-nvl72/> (Utolsó letöltés: 2026. 09. 07.)
- [6] Elmeleegy, A. / NVIDIA (2024): NVIDIA Contributes NVIDIA GB200 NVL72 Designs to Open Compute Project. NVIDIA Technical Blog, 15 October 2024. URL: <https://developer.nvidia.com/blog/?p=90182> (Utolsó letöltés: 2026. 09. 07.)
- [7] MLCommons (2026): MLPerf Inference – Power Measurement documentation and benchmark methodology. URL: <https://docs.mlcommons.org/inference/power/> (Utolsó letöltés: 2026. 09. 07.)

- [8] ISO/IEC (2026): ISO/IEC 30134-2:2026 – Information technology — Data centres key performance indicators — Part 2: Power usage effectiveness (PUE). URL: <https://www.iso.org/standard/30134-2> (Utolsó letöltés: 2026. 09. 07.)
- [9] ISO/IEC (2022): ISO/IEC 30134-9:2022 – Information technology — Data centres key performance indicators — Part 9: Water usage effectiveness (WUE). URL: <https://www.iso.org/standard/77692.html> (Utolsó letöltés: 2026. 09. 07.)
- [10] ISO/IEC (2022): ISO/IEC 30134-8:2022 – Information technology — Data centres key performance indicators — Part 8: Carbon usage effectiveness (CUE). URL: <https://www.iso.org/standard/77691.html> (Utolsó letöltés: 2026. 09. 07.)
- [11] Wang, N.; Guo, Y.; Huang, C.; Shao, S. (2024): Advances in direct liquid cooling technology and waste heat recovery for data center: A state-of-the-art review. *Journal of Cleaner Production*, 477, 143872. DOI: 10.1016/j.jclepro.2024.143872. URL: <https://doi.org/10.1016/j.jclepro.2024.143872> (Utolsó letöltés: 2026. 09. 07.)
- [12] Yuan, X.; Liang, Y.; Hu, X.; Xu, Y.; Chen, Y.; Kosonen, R. (2023): Waste heat recoveries in data centers: A review. *Renewable and Sustainable Energy Reviews*, 188, 113777. DOI: 10.1016/j.rser.2023.113777. URL: <https://doi.org/10.1016/j.rser.2023.113777> (Utolsó letöltés: 2026. 09. 07.)
- [13] Lei, N.; Lu, J.; Shehabi, A.; Masanet, E. R. (2025): The water use of data center workloads: A review and assessment of key determinants. *Resources, Conservation and Recycling*, 219, 108310. DOI: 10.1016/j.resconrec.2025.108310. URL: <https://doi.org/10.1016/j.resconrec.2025.108310> (Utolsó letöltés: 2026. 09. 07.)
- [14] Forschungszentrum Jülich (2025): JUPITER Supercomputer Propels European Computing Power. 10 June 2025. URL: <https://www.fz-juelich.de/en/news/archive/press-release/2025/jupiter-supercomputer-propels-european-computing-power> (Utolsó letöltés: 2026. 09. 07.)
- [15] LUMI (2023): LUMI data center receives the Green Data Centre of the Year award. 9 March 2023. URL: <https://lumi-supercomputer.eu/lumi-data-center-receives-the-green-data-centre-of-the-year-award/> (Utolsó letöltés: 2026. 09. 07.)
- [16] OpenAI (2025): Introducing Stargate Norway. 31 July 2025. URL: <https://openai.com/index/introducing-stargate-norway/> (Utolsó letöltés: 2026. 09. 07.)
- [17] KIFÜ (2024): A hazai HPC infrastruktúra fejlesztésének eredménykommunikációja – Komondor szuperszámítógép és hulladékhő-hasznosítás. URL: https://kifu.gov.hu/wp-content/uploads/2024/08/GINOP315_eredmenykomm_A4_online.pdf (Utolsó letöltés: 2026. 09. 07.)
- [18] EuroHPC Joint Undertaking (2026): The EuroHPC Joint Undertaking launches the AI Gigafactories Call. 30 July 2026. URL: https://www.eurohpc-ju.europa.eu/eurohpc-joint-undertaking-launches-ai-gigafactories-call-2026-07-30_en (Utolsó letöltés: 2026. 09. 07.)
- [19] HUN-REN (2025): HUN-REN success: Hungary among the first to connect to Europe’s largest supercomputer under the AI Factory Antenna programme. 13 October 2025. URL: https://hun-ren.hu/hunren_news/ai-factory-antenna-programme-109669 (Utolsó letöltés: 2026. 09. 07.)
- [20] Constellation Energy (2024): Constellation to Launch Crane Clean Energy Center, Restoring Jobs and Carbon-free Power to the Grid. 20 September 2024. URL: <https://investors.constellationenergy.com/news-releases/news-release-details/constellation-launch-crane-clean-energy-center-restoring-jobs> (Utolsó letöltés: 2026. 09. 07.)