Adatcenterek evolúciója AI szerverek kiszolgálására

Laky István – vezető rendszermérnök

JUNIPER NETWORKS | Driven by Experience™

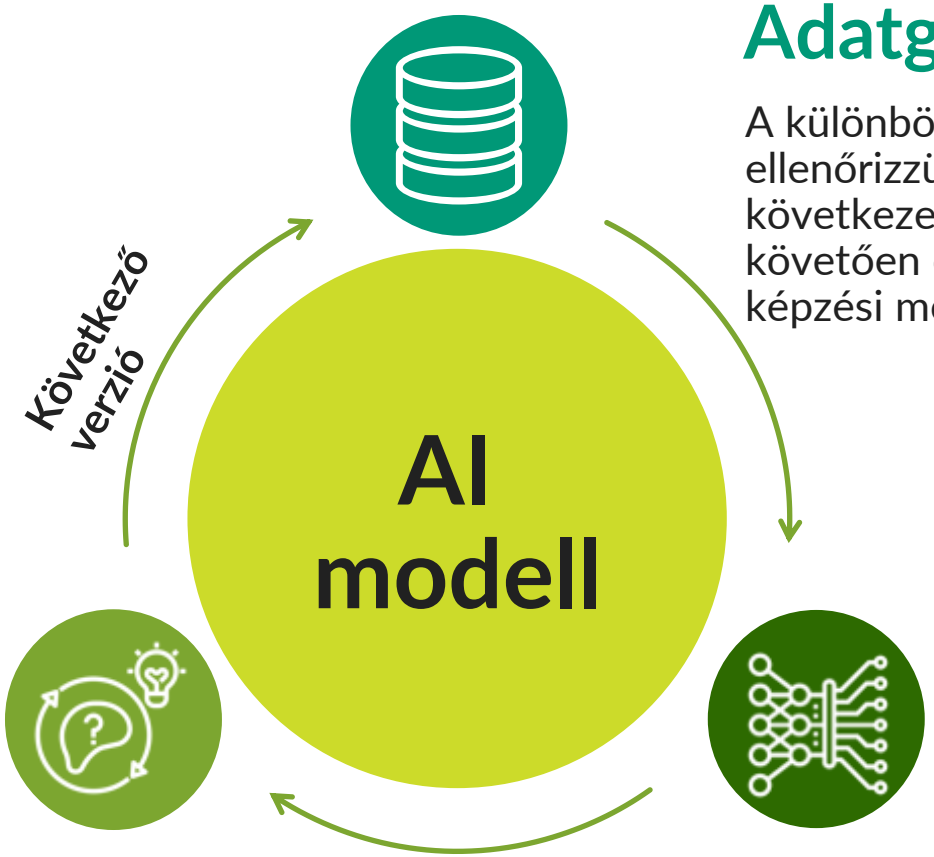
A DC A SZÁMÍTÁSTECHNIKA ÚJ EGYSÉGE

AI/ML INTRO

Mesterséges intelligencia modell

Adatgyüjtés

A különböző forrásokból gyűjtött adatokat ellenőrizzük a megbízhatóság és a következetesség szempontjából. Ezt követően előkészítik és összeállítják a képzési modell általi használatra.



Inferálás

A betanított modellt a következtetési összefüggések alkalmazzák, hogy a felhasználói bemenetek eredményeként használhatók legyenek.

Tanulás

Az AI-modellt a kurált adatkészlettel és a GPU csoportokkal mély tanulási keretrendszerével képezték ki.

Az AI/ML-alkalmazások életciklusa folyamatos, iteratív folyamat lehet a modellek tervezésében és fejlesztésében, a képzésben és a kurált adatokkal való validálásában, valamint a termelésben történő üzembe helyezésében, miközben folyamatosan figyelemmel kíséri teljesítményüket a folyamatos finomítás és fejlesztés érdekében.

AI rendszer fogalmak

- **GPU-k:** Grafikus feldolgozó egység(ek): AI szerver számítási egysége
- **RDMA:** Közvetlen távoli memóriaelérés (dedikált tároláshoz használatos)
- **RoCE(v2):** RDMA Converged Etherneten keresztül (v2 = L3 alapú)
- **JCT:** Munka befejezési ideje (= a legfontosabb KPI az AI ML DC-kben)
- **RAIL :** szerveren belül egy adott helyen/indexen található összes GPU-ból áll
- **Stripe:** olyan leaf-ek halmaza , amelyekhez egyetlen GPU-node csatlakozik

Kihívások az AI/ML érdekében

Üzleti

- Az adatokban rejlő lehetőségek azanosítása
- Képzés és szakértő IT személyzet felvétele
- Szállítási idők és telepítés
- Minimalizálni a tanulási modell befejezési idejét
- Költségek kezelése, különösen a GPU költségeinek kezelése
- Adatok bizalmas kezelése

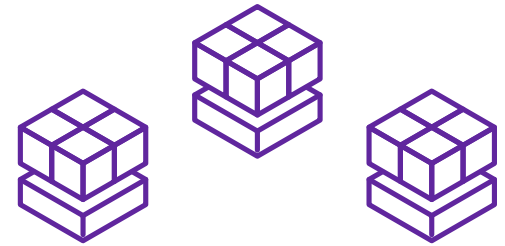
Technológia

- Hely, teljesítmény és hűtés
- Új mesterséges intelligencia szoftvercsomag integrálása
- Automatizálás, felügyelet, hibaelhárítás
- Az infrastruktúra megbízhatósága és rugalmassága
- GPU-k és tárolók nagy teljesítményű hálózatba kapcsolása, különösen nagy tanulási hálózatokhoz
- Az innováció sebessége: az infrastruktúra optimalizálása változik és függhet a modelltől és a párhuzamosítástól

Kerülni kell el az infrastruktúra szűk keresztmetszetét

Ha a hálózat szűk, akkor késlelteti a tanulási folyamat befejezését, akkor drága
A GPU idő kihasználása rossz

A hálózat kulcsfontosságú a ROI optimalizálása szempontjából



A GPU-k képviselik a legtöbbet
a CapEx befektetésből

A hálózat összekapcsolja
a GPU-kat elosztott képzésben

JUNIPER NETWORKS

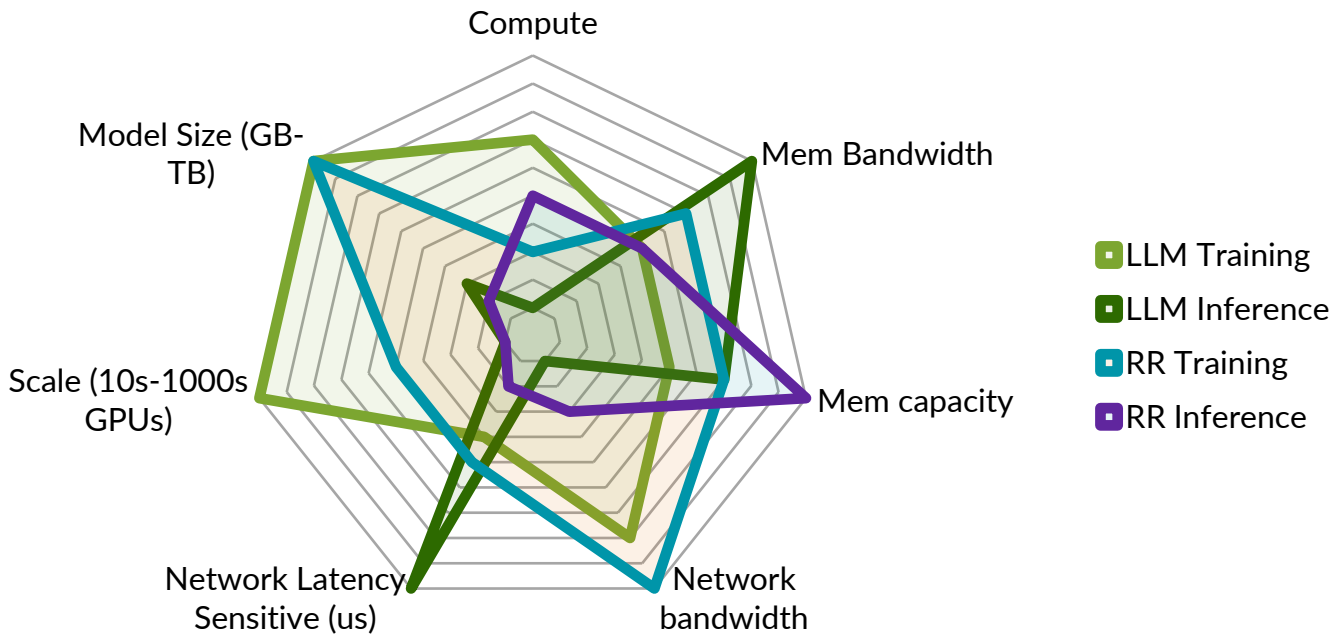
Az AI-követelmények eltérőek

A tanulás nagy klaszterekben történik

- A tanulas a GPU-k között osztják el
- Nagy hálózati sávszélesség
- A JCT latency fontos, de a hálózati késleltetés nem kritikusan érzékeny

Az inferálás kis klaszterekben elosztva

- Az inferálás többnyire 1 CPU/GPU-ra vonatkozik, kivéve az LLM-eket, ahol lehet több is



AI/ML infrastruktúra

JUNIPER NETWORKS

Az AI/ML DC terhelések egyedi jellemzői

- Nagy, elefant flow-k
- Kevesebb flow, alacsony entrópia
- A forgalom elsősorban a GPU-k között folyik
  - A GPU-k nagyobb konzisztens sávszélesség-sűrűséget igényelnek a szűk keresztmetszetek elkerülése érdekében
  - A GPU-forgalom RDMA-alapú (nem TCP)
- A csomópontok egyszerre kezdenek továbbítani, gyorsan telítve a linkeket
- Nagyon érzékeny a csomagvesztésre és a jitterre

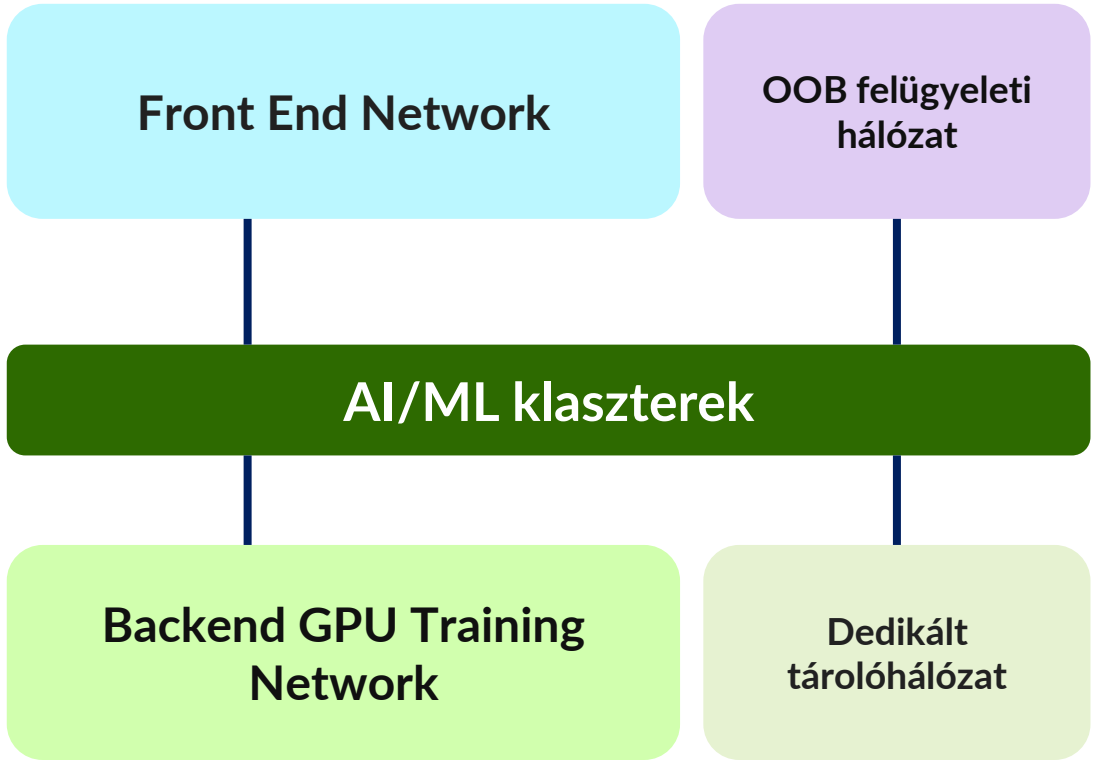
AI Cluster Networking kihívások

A frontend hálózatok mindig Ethernetek

- Itt fut a tanulás vezérlés
- Inferens ciklusok is futhak itt
- Az inferálás egyetlen szerveren fut, kivéve a ritka LLM eseteket, 16-64 GPU-kat használhat

A háttérhálózatok lehetnek InfiniBand vagy Ethernet

- Nincs oversubscription
- Offline, például a mesterséges intelligencia fejlesztői/teszt területei
- A tanulás gyakran 100-1000 GPU-t használ
- A betanítás néhány elefant flow-t eredményez, amelyek eltorlaszolják a fabric-ot, ha nem foglalkoznak velük (a kis flow entrópia kihívást jelent az ECMP-hashing számára)
- A képzési és tárolási forgalom elsősorban RDMA



Az AI/ML klaszter és hálózat anatómiája

Klaszter komponensek

- Tanulási klaszterek
- Inferálási klaszterek
- Megosztott tároló kapacitás
- Dedikált Cluster Storage

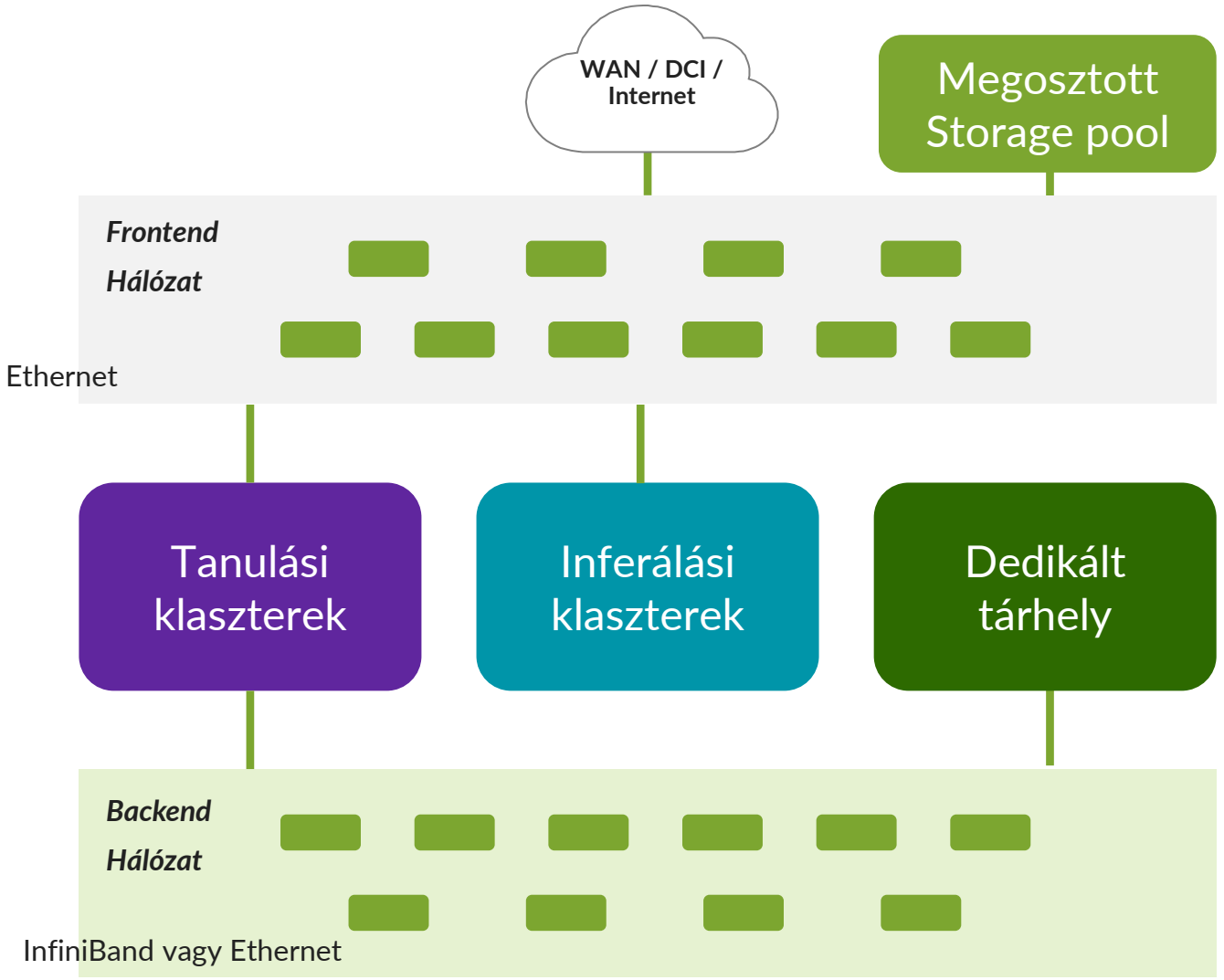
Klaszter hálózatok

„Frontend”

- Az inferálási klaszterek ezt a hálózatot használják
- Megosztott tároló terület
- Menedzsment hálózat a képzéshez

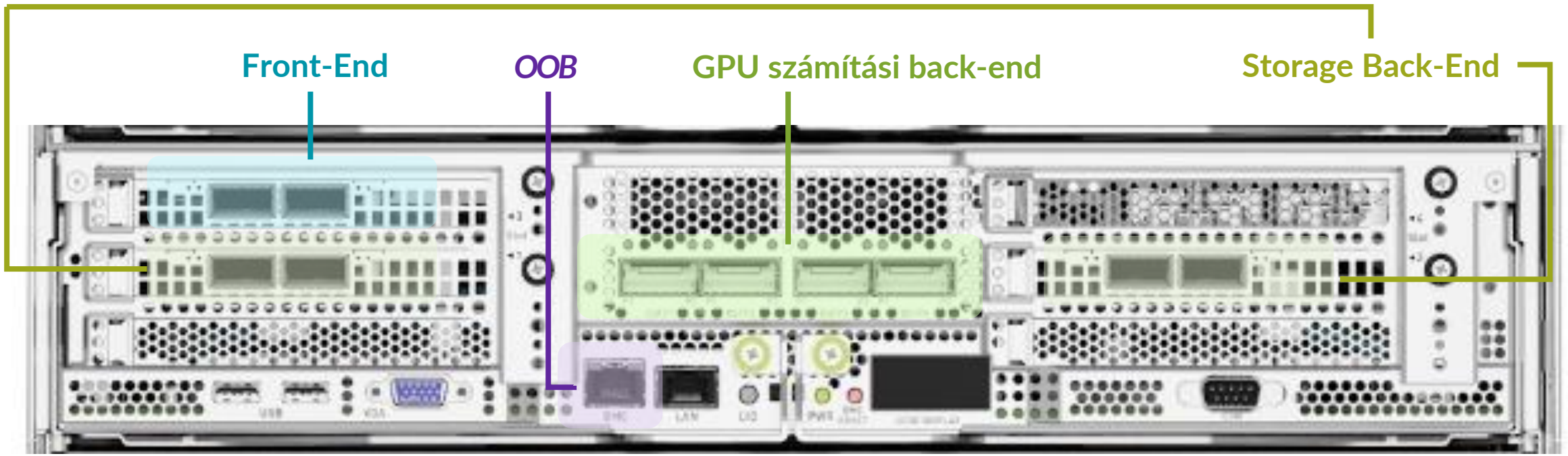
”Backend"

- GPU számítási Fabric
- Dedikált Storage Fabric



Az AI DC tipikus kliense: DGX H100 Server

Csatlakozás: 3,2 Tbps back-end, 800 Gbps Storage, 100 Gbps Front-end



4,1 Tbps szerverenként a három fő szegmensre

AI-adatközpontok

JUNIPER NETWORKS

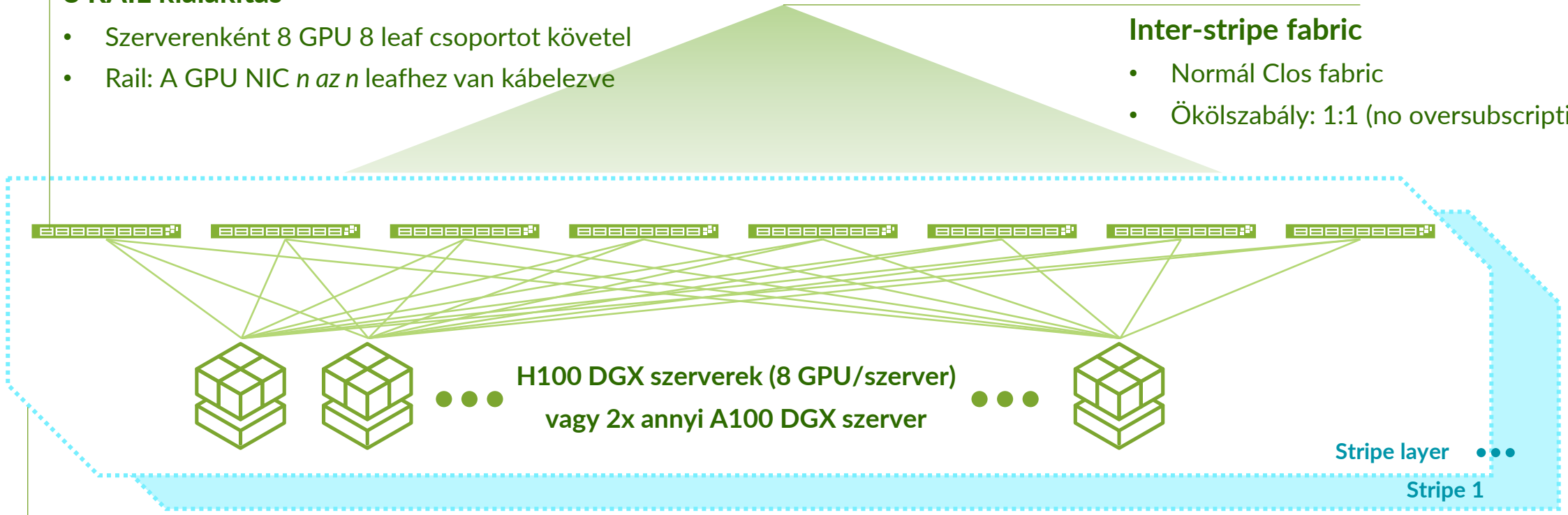
GPU Fabric Rail optimalizált kialakítás

8 RAIL kialakítás

- Szerverenként 8 GPU 8 leaf csoportot követel
- Rail: A GPU NIC *n az n* leafhez van kábelezve

Inter-stripe fabric

- Normál Clos fabric
- Ökölszabály: 1:1 (no oversubscription)



Stripe

- A „Stripe”-ban vagy a 8 leaf-ből álló csoportban az NCCL optimalizálhatja a kapcsolatot, hogy a csoportban tartsa a forgalmat, és minimalizálja a spine-okon áthaladó forgalmat

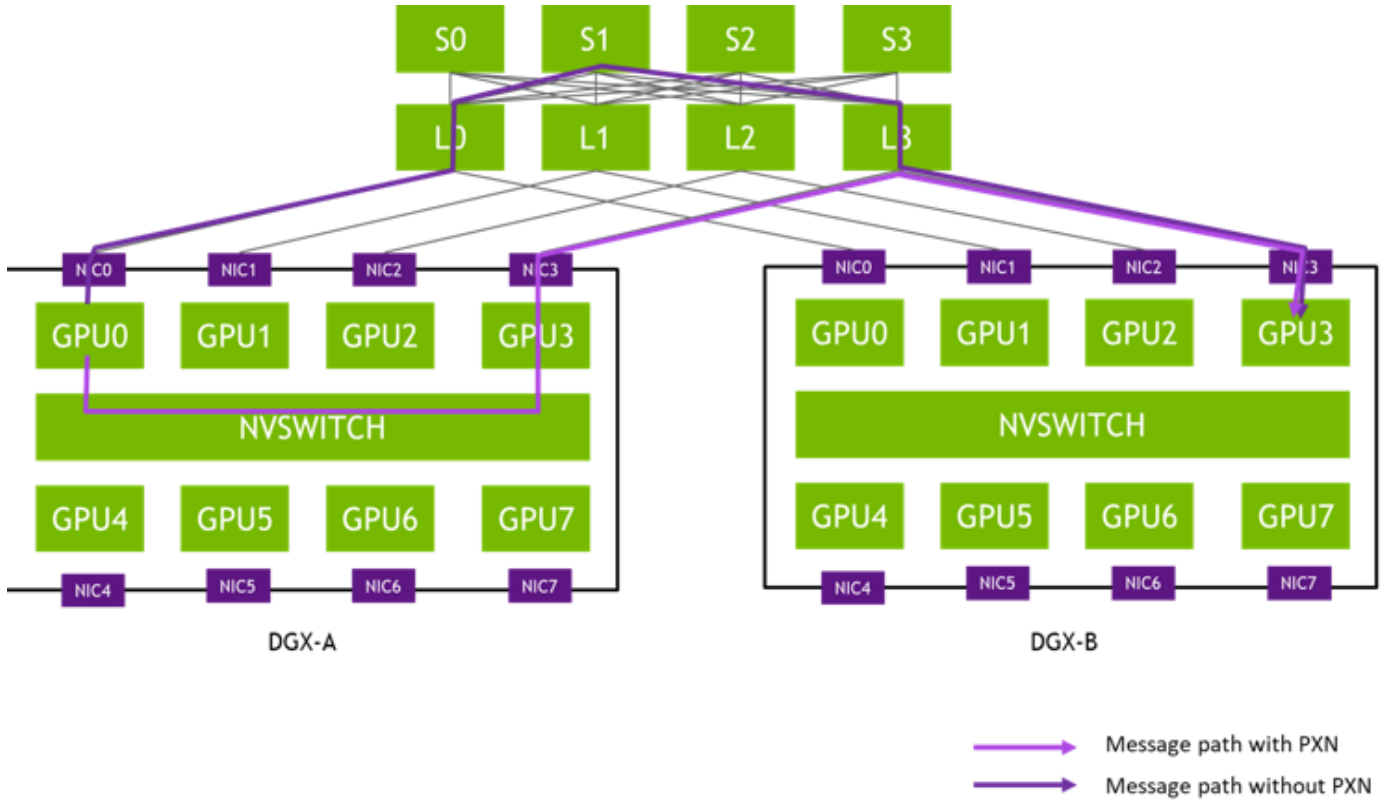
RAIL optimalizált hálózati konfiguráció

Az NVIDA NCCL + PXN optimalizálja a GPU-GPU kommunikációt a GPU-k közötti belső NVLink kapcsolat kihasználásával a kommunikációs útvonal optimalizálása érdekében.

Ebben a példában a DGX-A GPU0-nak üzenetet kell küldenie a DGX-B GPU3-nak.

Az NCCL + PXN nélkül az üzenet áthaladna a hálózat leaf és spine elemén.

Az NCCL + PXN esetén az üzenet a DGX-A GPU3-ra kerül, amely azután a hálózati kártyán keresztül továbbítja a DGX-B GPU3-nak.



Architektúra opciók AI Back-end hálózathoz

InfiniBand

Megfelelő Mellanox hálózati kártyák és kapcsolók

Előnyök

- Többnyire veszteségmentes és nagy teljesítményű
- Beépített torlódás elkerülés
- Alacsony késleltetés

Hátrányok

- Egyedülálló gyártótól származó hálózati kártyák kapcsolói és szoftverei, valamint L1-L4
- Szállítási idők és az NVIDIA, mint szűk keresztmetszet az értékesítésben

Ethernet Fabric

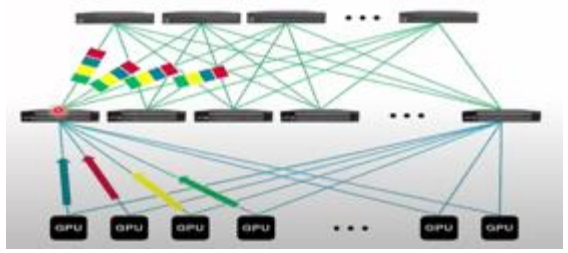
Sekély puffer, mély puffer, hibrid a levélen és a gerincen

Előnyök

- Chip- és berendezésgyártók gazdag ökoszisztémája, erős fejlesztési háttér. pl. jól érthető biztonsági lehetőségek
- Szabvány alapú, átjárható, az IT szakértők számára jól ismert
- Torlódások elkerülése és szabályozása DCQCN (ECN+PFC) használatával

Hátrányok

- Alacsony késleltetés (1-3 us, de az TH chipek segítenek)
- Előfordulhat, hogy az ALB/DLB nem tökéletes a torlódások elkerülésére, és a torlódásszabályozás átmenetileg lelassíthatja a dolgokat



Scheduled Fabric

Cell or Ethernet-based

disaggregated chassis

Előnyök

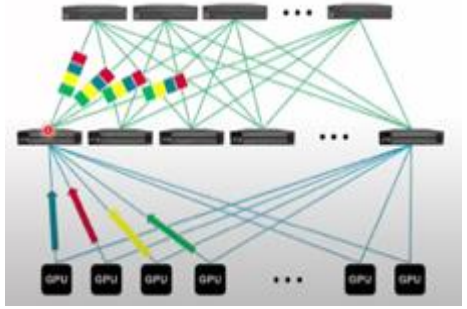
- Felbontja a vonalkártyáit és a fabric kártyáit spine és leaf switch-ekre.
- VoQ-t alkalmaz és cella alapú továbbítást a jobb terheléselosztás érdekében a spine/leaf fabric-ban

Hátrányok

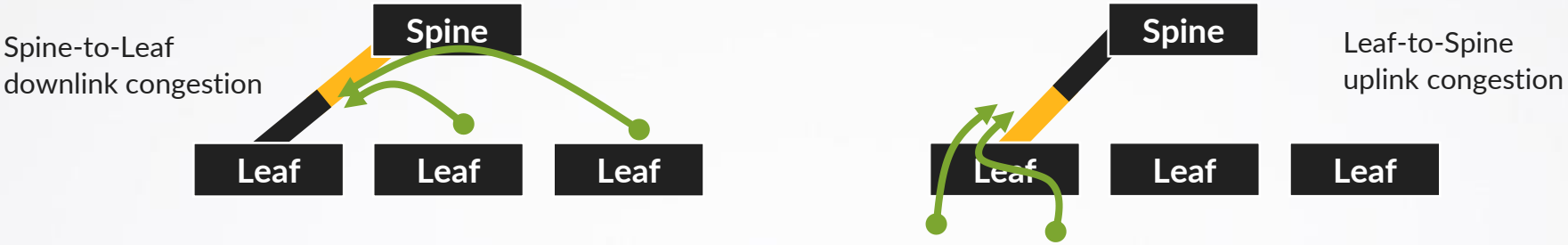
- Nem skálázható nagy telepítésekre, mégha azt is állítják
- A valódi teljesítménynövekedést ellenőrizni kell
- Chip korlátozások
- Gyártó szabvány miatti kötött upgrade-ek

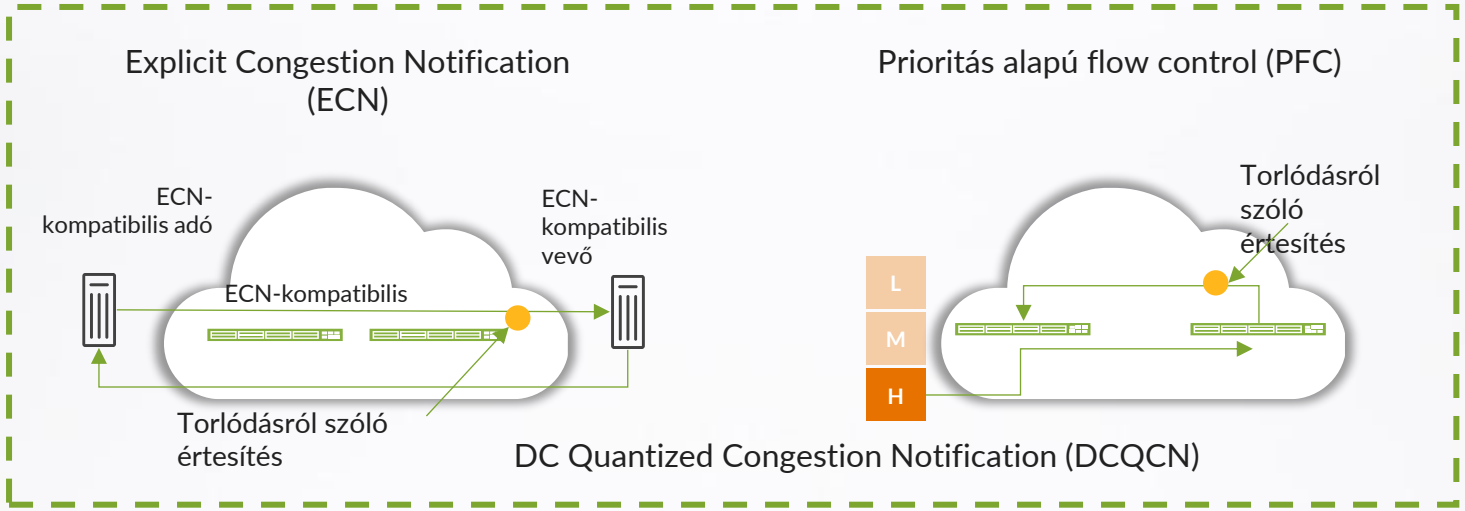
NVIDIA JUNIPER NETWORKS NVIDIA ARISTA CISCO Q Fabric BROADCOM CISCO

Architecture Options for AI Back-end Network

	 **InfiniBand** *Szabadalmaztatott Mellanox NIC-ek és kapcsolók*	 **Ethernet fabric** *sekély puffer, mély puffer, Hibrid a leveleken és a tüskén át*	 **Scheduled Fabric** *Cella vagy Ethernet alapú*
Vendor lock	Igen	Nem	Igen
Működési következetesség	Nem	Igen	Nem
Költség	Magasabb	Alacsonyabb	Magasabb
Skála	Centrali/korlátozott	Elosytott. /Magasabb	Central/korlátozott
Teljesítmény AI munkaterhelésekhez	Igen	Igen	Igen

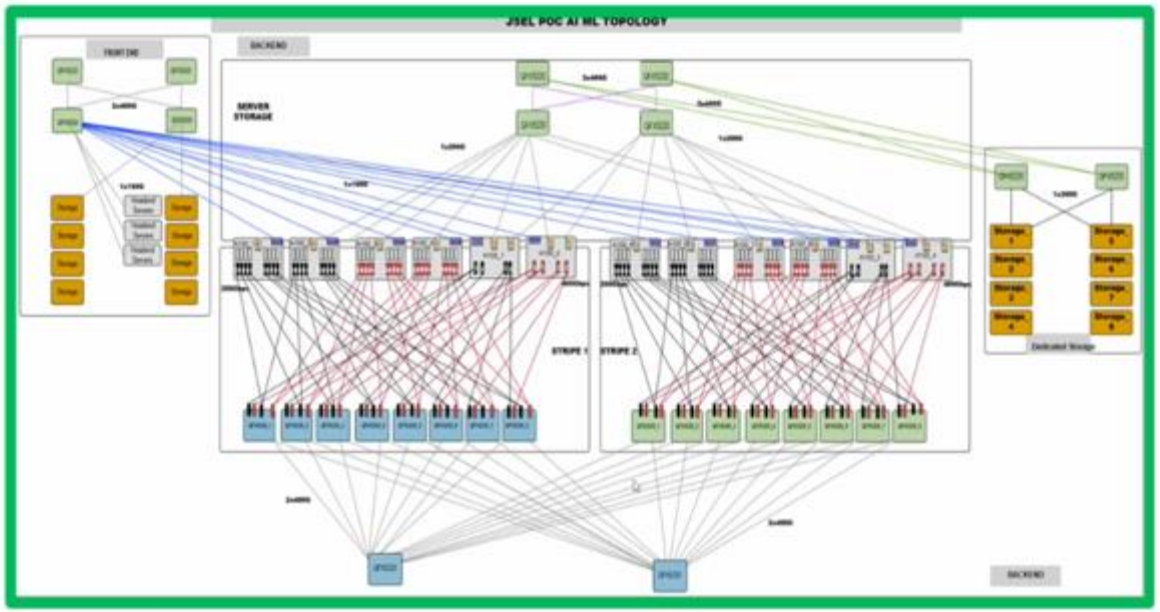
Ethernet: InfiniBand teljesítmény magas költségek nélkül

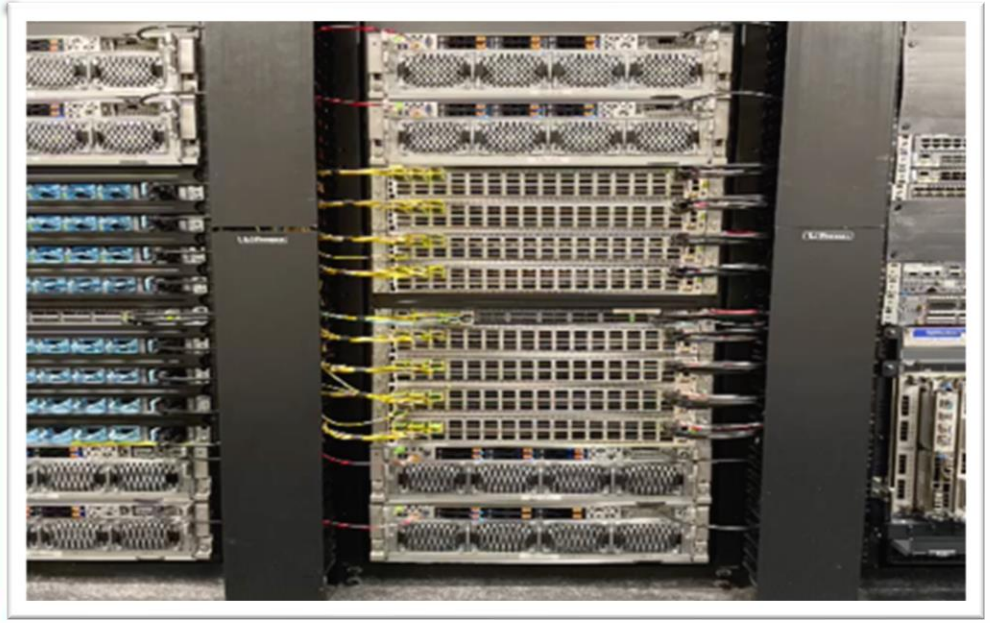




A Juniper Lab Benchmark tesztjei az „AI Optimized Ethernet” teljesítményét az IB-hez hasonló teljesítményt mutatják.

Juniper Networks PoC labor Kaliforniában





BERT-Large

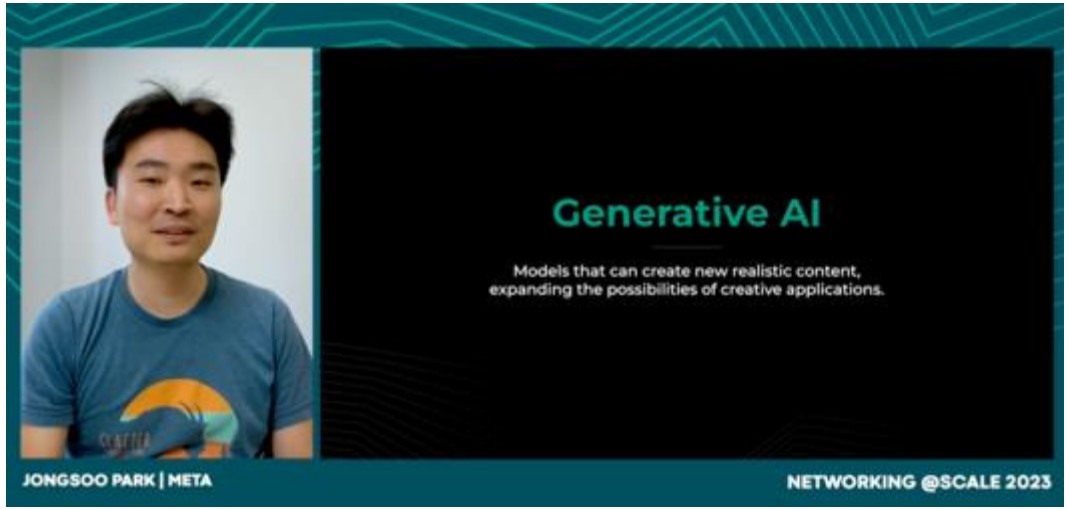
2.52 min
Juniper Ethernet Training Time

2.5-3.3 min
Other benchmarks posted (including IB)

Konklúziók

RDMA hálózati kapcsolat Ethernet alapon

- Több szállító lehetősége, vonali sebesség, optikai opció, valamint gyorsabb innováció, mint az InfiniBand
- Az RDMA over Converged Ethernet (RoCE) a GPU számítási és Storage Fabric-ban egyaránt használható
- Sokkal jobb biztonsági lehetőségek, például VXLAN több helyszínű környezetekben
- A Junos támogatja a RoCEv2 torlódások elkerülését és kezelését (DLB, DCQCN / ECN / PFC)



"Hasonló eredményeket értünk el Etherneten, mint az InfiniBand, és reméljük, hogy ez szabad piacot ad az LLM hardver szükségletében."

- Jungsoon Park, Meta

A GPU-k drágák és szűkösek

Egyetlen GPU

33k USD

8 x GPU szerver

350k USD

Kis AI klaszterek

AI front-end hálózat

AI hálózat training

5-10 millió USD

Nagy AI-klaszterek

100M USD

Adatközpont CapEx összehasonlítás

	Hagyományos DC	AI Training DC
Számítsa ki	55%	80%
Tárolás	35%	14%
Hálózat	10%	6%

A Meta tavaly azt állította, hogy az AI/ML-ben eltelt idő 33%-át a hálózatra várással töltik.

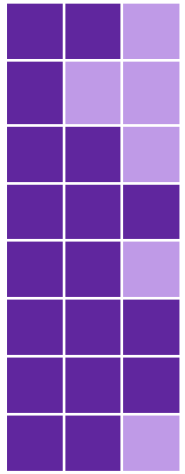
Forrás: Dell'Oro

...a hálózat kritikus

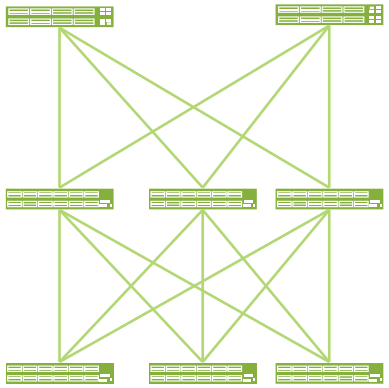
JUNIPER NETWORKS

Az AI-modell teljesítménye és gazdaságossága a Job Completion Time (JCT) minimalizálásán alapul

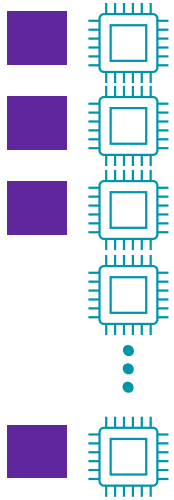
Az adatok darabolva…



FED to GPU
DC fabric-on keresztül…



A GPU-k végzik a **SZÁMÍTÁSOKAT**



…számítások **összefűzve hálózaton keesztül**

Az IDŐ PÉNZ

Az összes csomópont előrehaladását bármilyen késleltetett flow visszatarthatja

Thank you

JUNIPER NETWORKS | Driven by Experience™