Optimalizált AI adatcenter switching

Laky István
Senior Systems Engineer

Juniper Networks | Driven by Experience™

AI DC infrastruktúra elemei és kihívásai

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

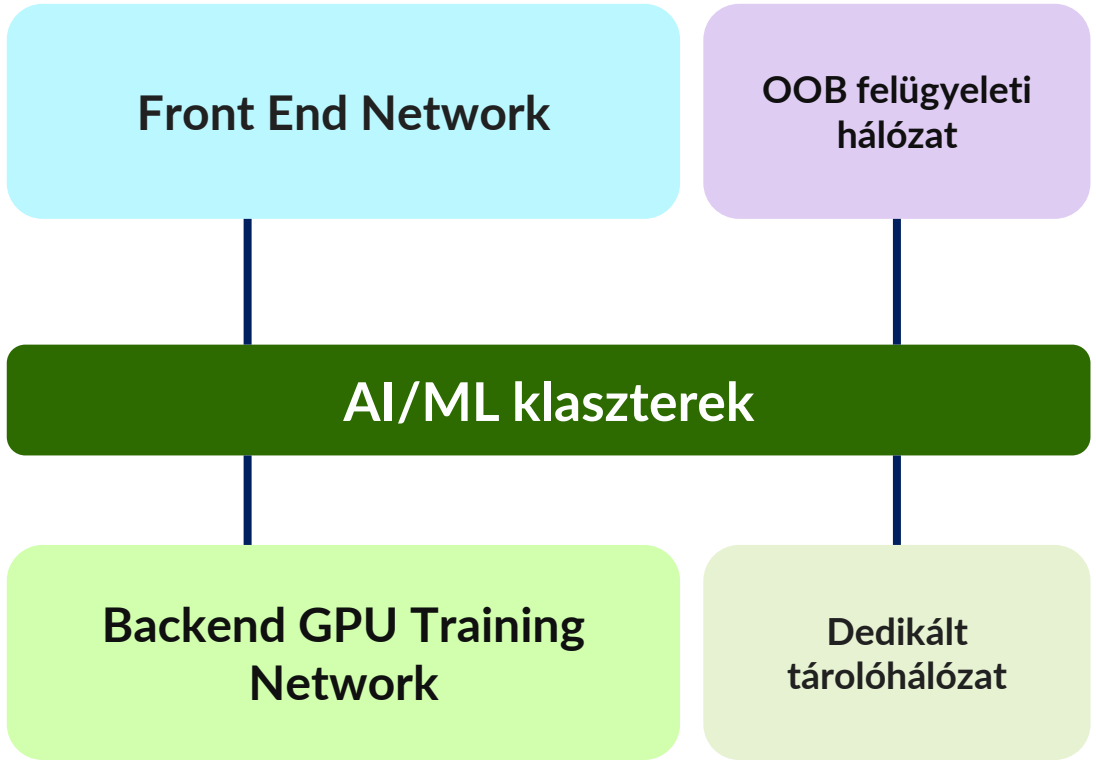
AI Cluster Networking kihívások

A frontend hálózatok

- 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 jellemzői

- 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 tanulási és tárolási forgalom elsősorban RDMA-t használnak



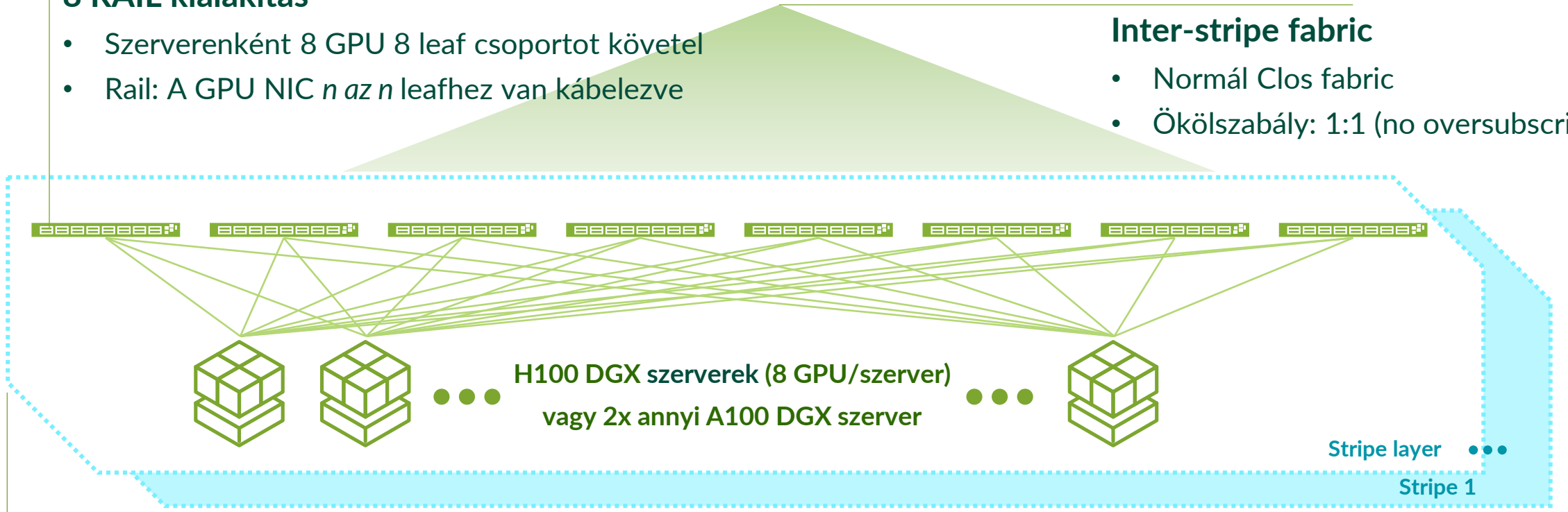
GPU Fabric Rail optimalizált kialakítás

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

NCCL (**N**VIDIA **C**ollective **C**ommunications **L**ibrary)

Fő funkciói és előnyei:

Több GPU közötti kommunikáció: Az NCCL lehetővé teszi a hatékony kommunikációt nemcsak egyetlen gépben található GPU-k között, hanem több gépen elhelyezkedő GPU-k között is (ezeket nevezzük **multi-node, multi-GPU** rendszereknek).

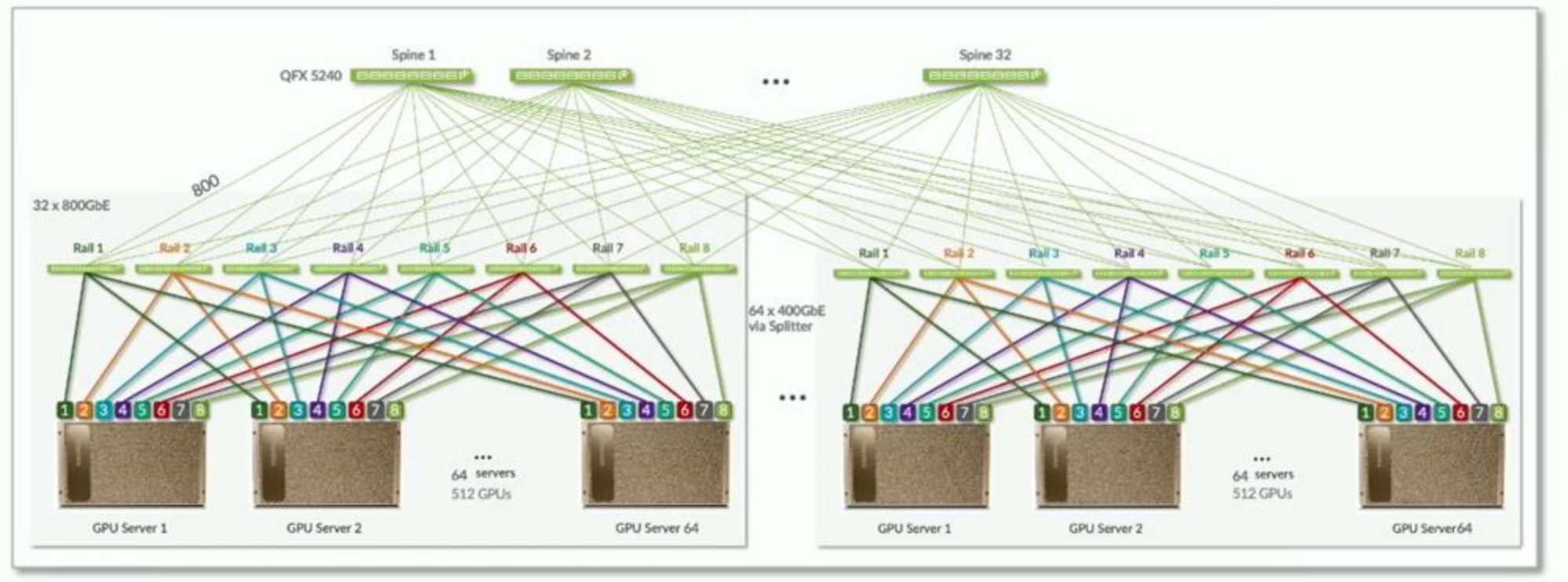
Kollektív műveletek: Az NCCL számos kollektív kommunikációs műveletet támogat, például:

- **All-Reduce**: Az összes GPU összeadja az adatait, majd megosztják az eredményt.
- **Broadcast**: Egy GPU elküldi az adatait az összes többi GPU-nak.
- **All-Gather**: Minden GPU megosztja a saját adatait az összes többi GPU-val.
- **Reduce-Scatter**: Az adatokat összegyűjtik, redukálják, majd szétszórják az eredményt a GPU-k között.

Skálázhatóság: Az NCCL képes zökkenőmentesen skálázni egyetlen GPU-tól egészen több ezer GPU-ig, beleértve a multi-node rendszereket, ahol az egyes GPU-k több fizikai gépen helyezkednek el.

Teljesítmény optimalizálás: Az NCCL optimalizálva van **NVLink PCIe** és **InfiniBand** hálózati technológiákra,

Egy tipikus Spine-Leaf fabric

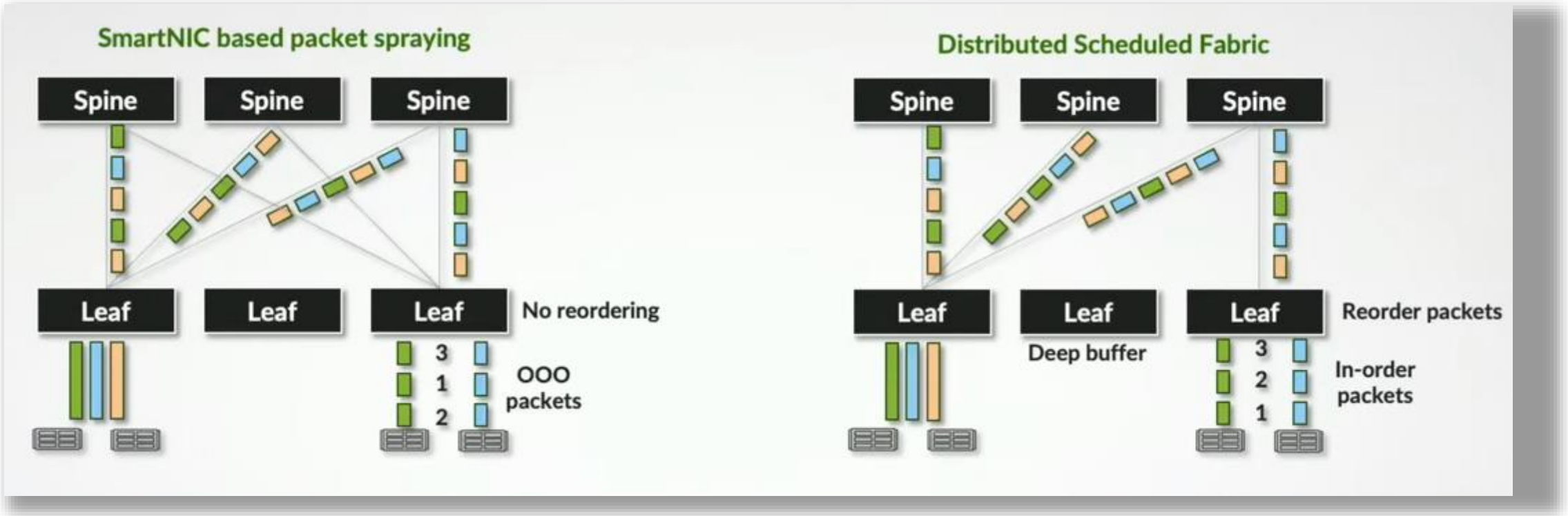


- Nagy sávszélességű flow-k nehézkessé teszik a forgalom megosztásást (load balancing)
- GPU-k a szinkonizáció végéig “állnak” (idle)

TCP/IP stack optimalizálási technikák

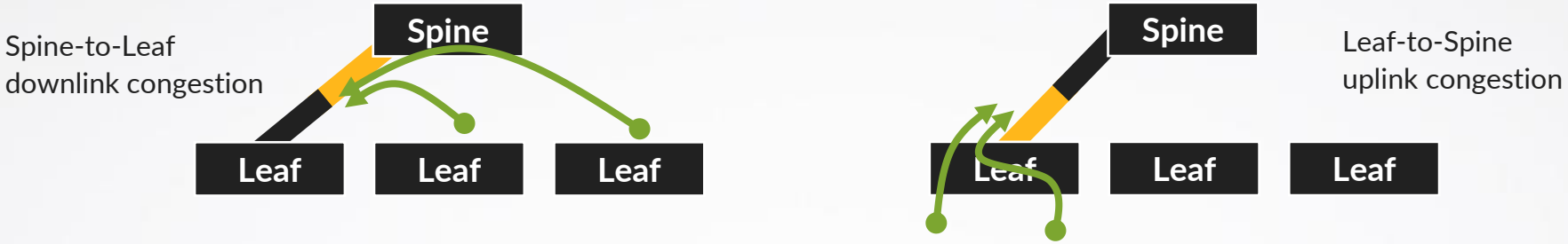
JUNIPER NETWORKS

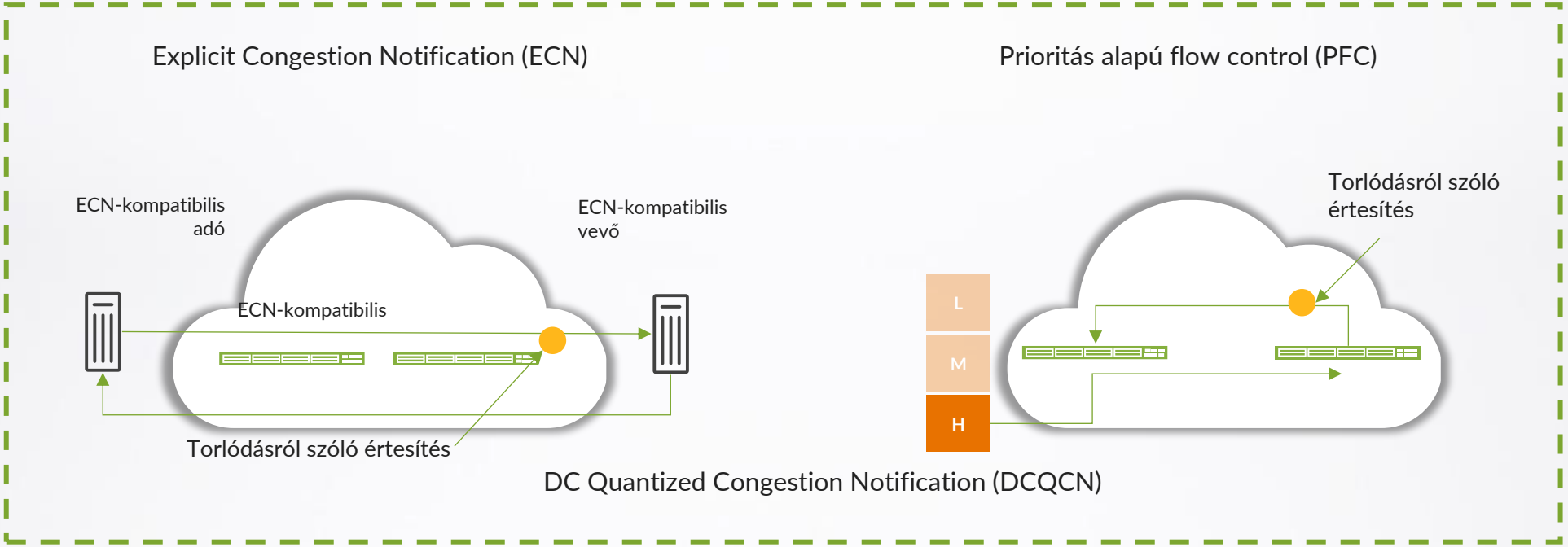
Forgalom elosztási technikák



	SuperNIC alapu Spraying	Elosztott ütemezett FABRIC
Forgalom elosztás módja	Spraying	Credit alapú elosztás
Költség	speciális NIC igény	Deep buffer switch igény
Gyártó függő	Igen	Igen

Ethernet alapú megoldások

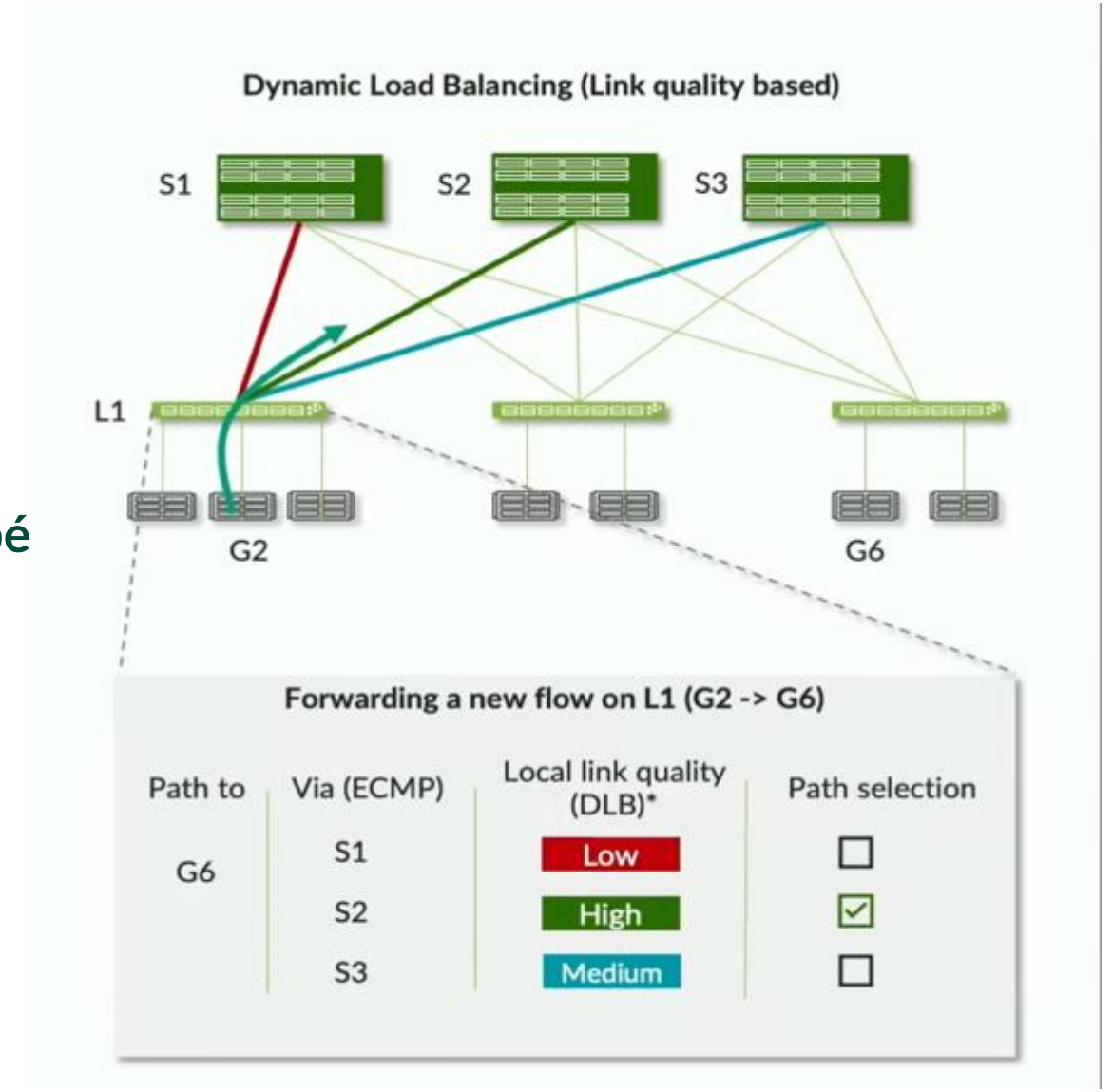




Load Balancing optimalizálási technológiák

Dinamikus Load Balancing (DLB)

- A DLB vizsgálja a link állapotát
  - ECMP path alapján
  - valós idejű link kihasználtság alapján
  - queue telítettség alapján
- Ezek alapján besorolja a link-eket és keresi a kevésbé használt utakat
- Két üzemmódban is működhet
  - Flowlet mód
  - Packet mód



Load Balancing mérési eredmények

	DLRM Training Time (min)			
	Background traffic == 0%	Background traffic == 17.5%	Background traffic == 32.5%	Background traffic == 37.5%
Static ECMP	2.8	3.045	3.225	4.693
DLB (Flowlet)	3.068	3.115	3.163	3.178
DLB (Per-Packet spraying)	3.026	3.085	3.127	3.155

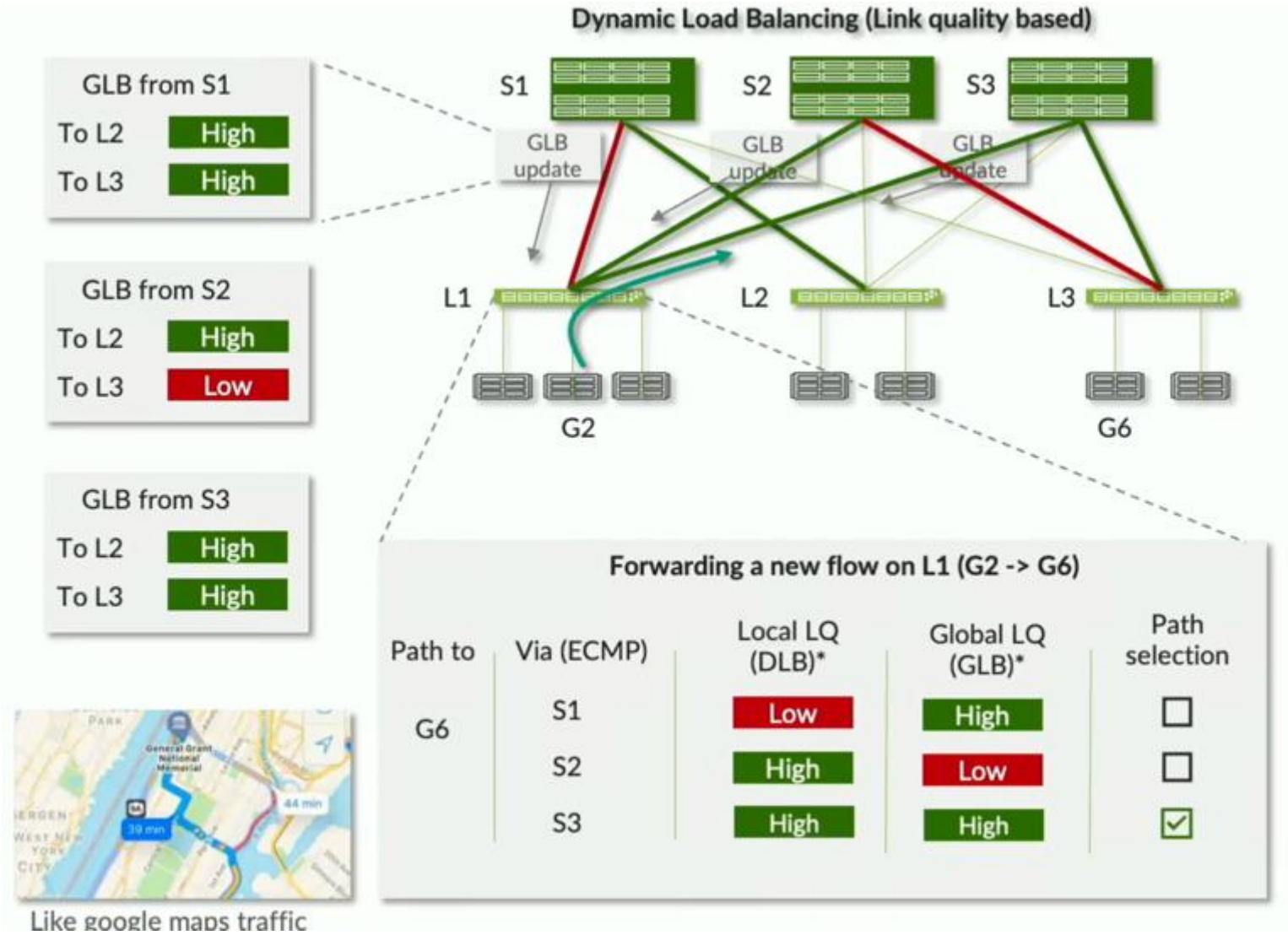
Static hashing performance deteriorates quickly with background traffic

Highly Deterministic

Packet spraying only marginally better than DLB-Flowlet (0.08%)

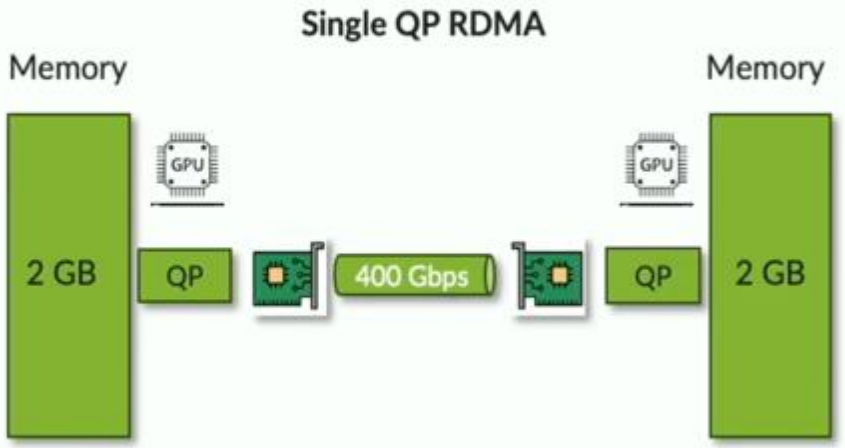
Global Load Balancing (GLB)

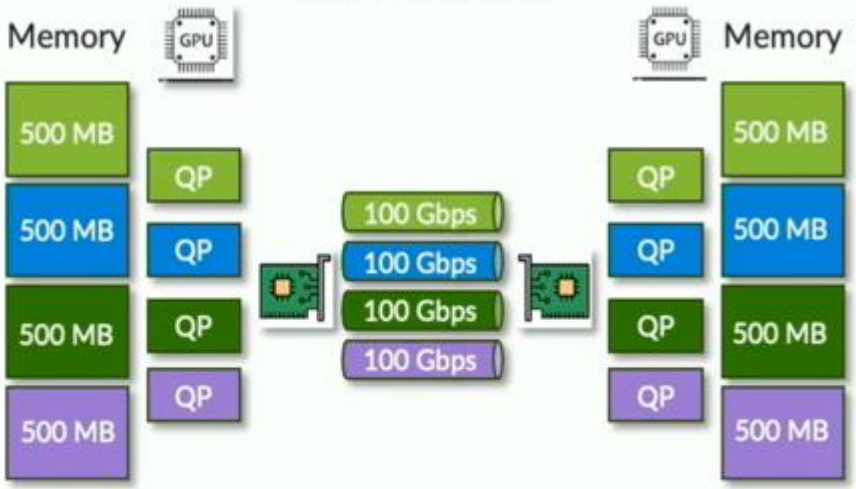
- A Spine switch moitorozza a leaf kapcsolatait és hirdeti vissza a DLB döntésekhez
- A Leaf switch ezeket feldolgozza és alkalmazza a DLB-hez
- Az „upstream” switch elkerüli a túlterhelt linkeket



RDMA optimalizálás

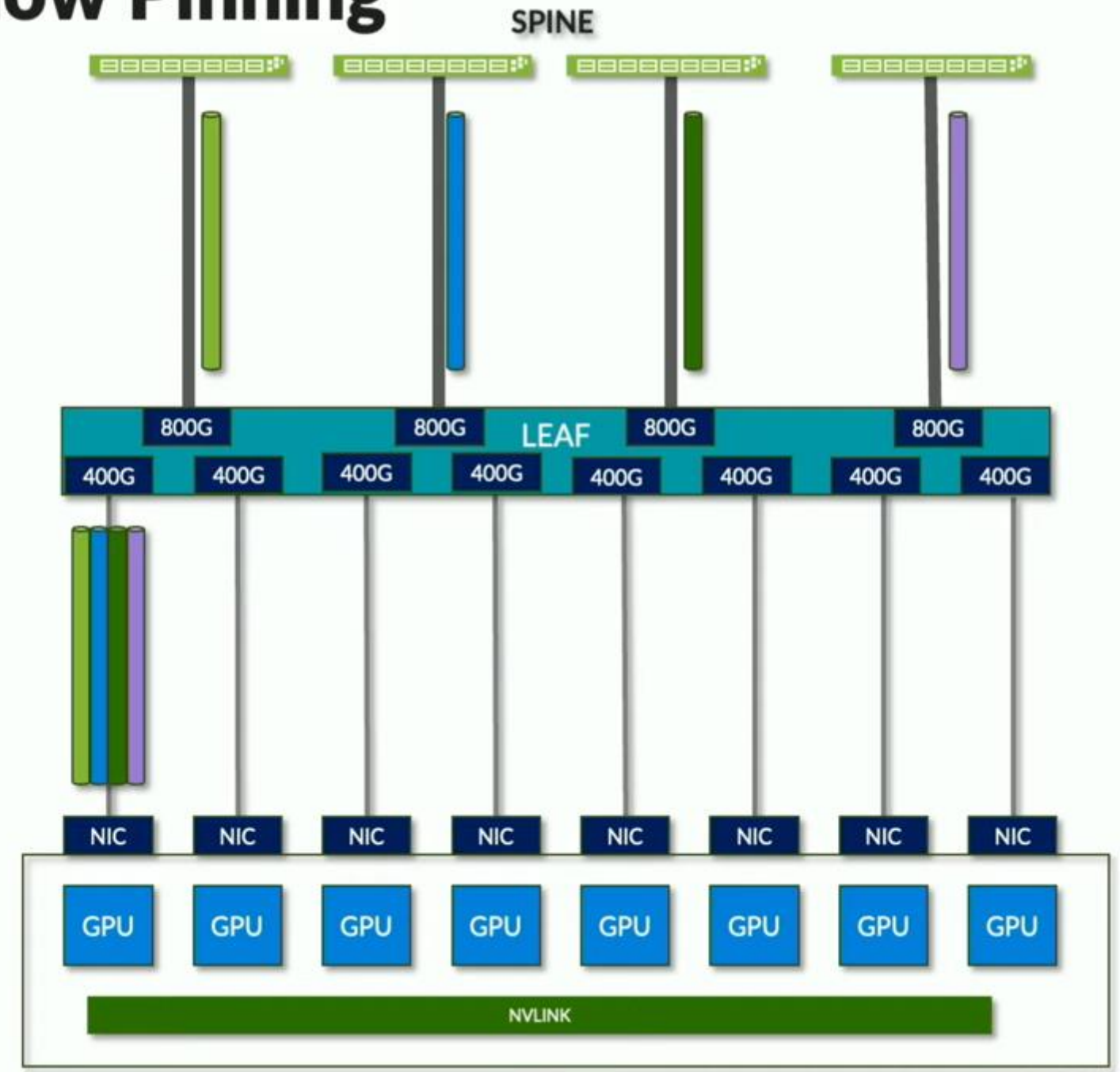
Deterministic RDMA Flow Pinning



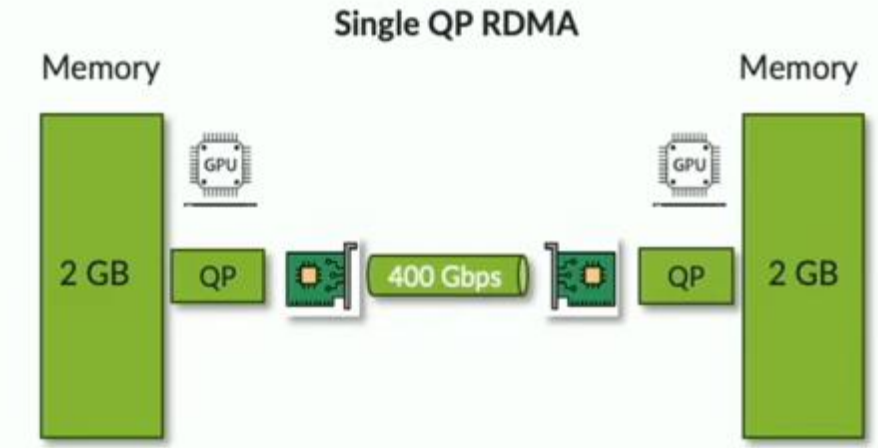


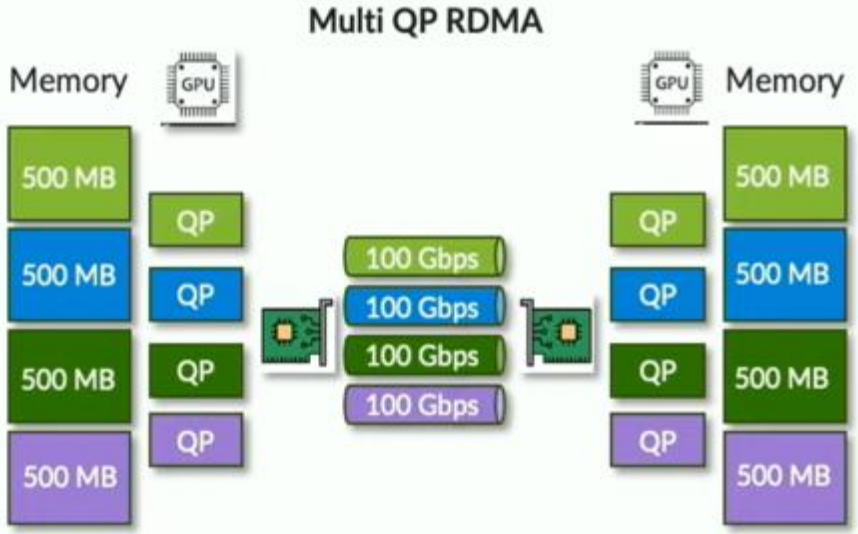
Num of QPs = Num of leaf-spine links = N = 4

JNPR NCCL plugin sets up N QPs per GPU



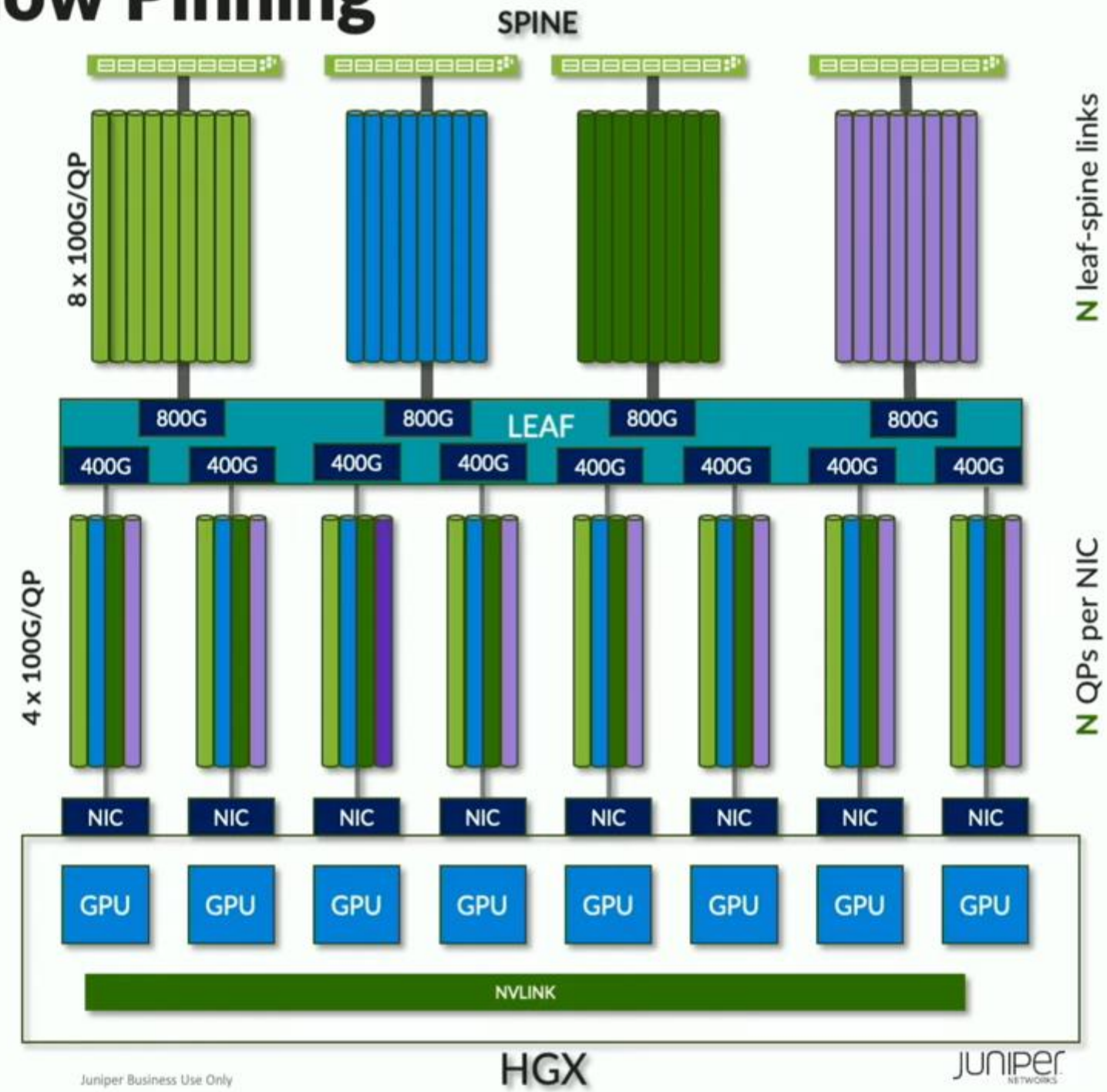
Deterministic RDMA Flow Pinning





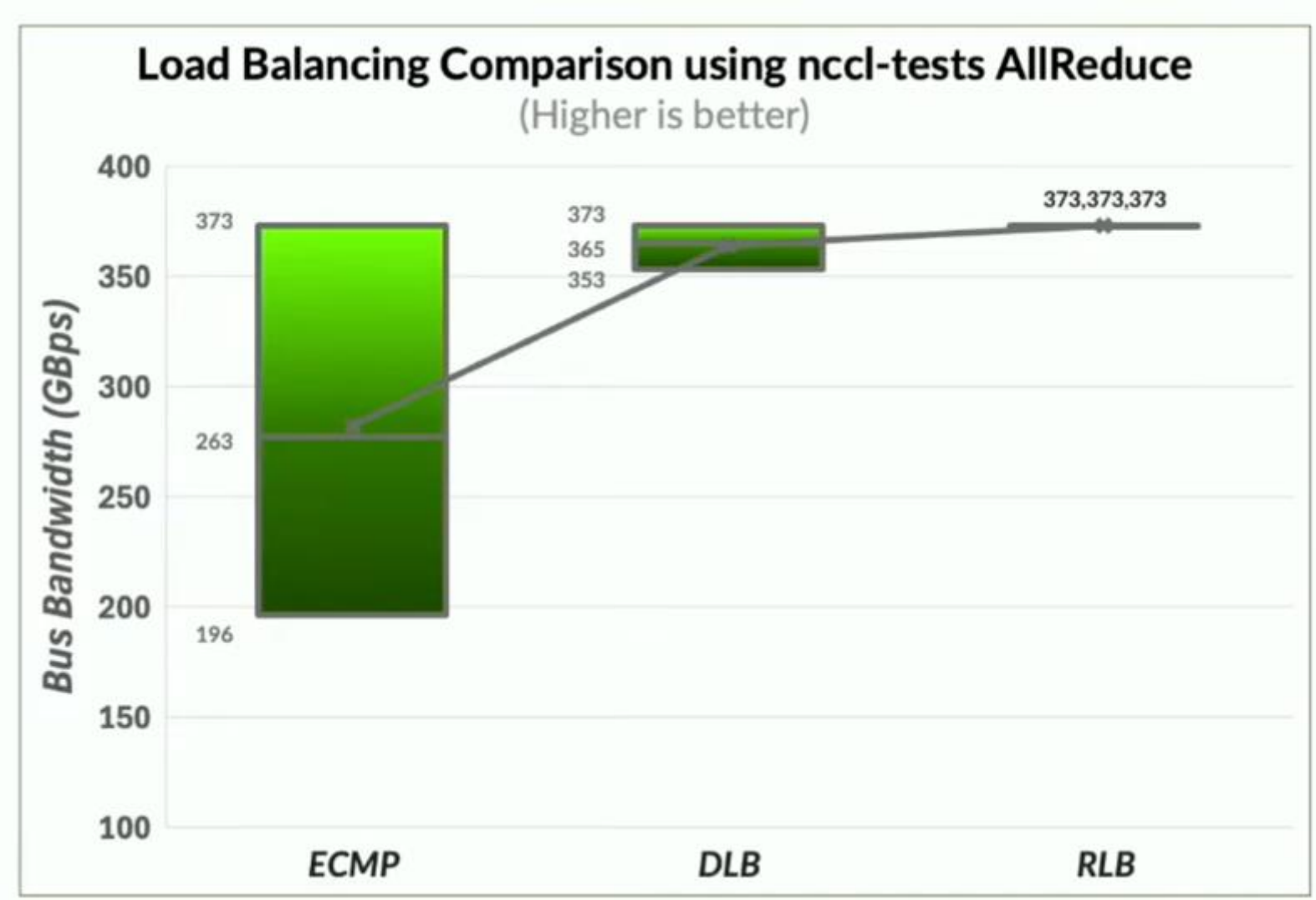
Num of QPs = Num of leaf-spine links = N = 4

JNPR NCCL plugin sets up N QPs per GPU



Konklúzió

Tesztelés eredmények szórása



Spine – leaf linkek kihasználtsága

Unequal Flow distribution (leaf-spine links)

Static ECMP

Equal Flow distribution (leaf-spine links)

DLB (Flowlet)

Thank you

JUNIPER NETWORKS | Driven by Experience™