



# AIネットワークのボトルネックを打破する： JuniperによるAI ロードバランシングの革新

ジュニパーネットワークス株式会社

技術本部 ソリューション技術部 - 塚本 広海

DC BU PLM - Mahesh Subramaniam

**JUNIPER**  
NETWORKS

## 塚本 広海 (Tsukamoto Hiromi)



ジュニパーネットワークス(約9年) - ソリューション技術推進本部  
2002-2016 某SIer : 製品担当、R&D、テクニカルプリセールス  
2016 - Juniper : APAC PLM/TME/COA,  
現在 - Software,DC,Sec 関連のテクニカルプリセールスチーム mgr  
MPLS Japan初登壇 - JANOG, Internet Week, CODT etc.登壇有

## Mahesh Subramaniam



Data Center Product BU – Sr. Director Product Management

Mainly Focus AI DC Platform Software

UEC Summit, OCP'25 & NANOG



HPE Networking

**HPE** JUNIPER  
networking

**HPE** aruba  
networking

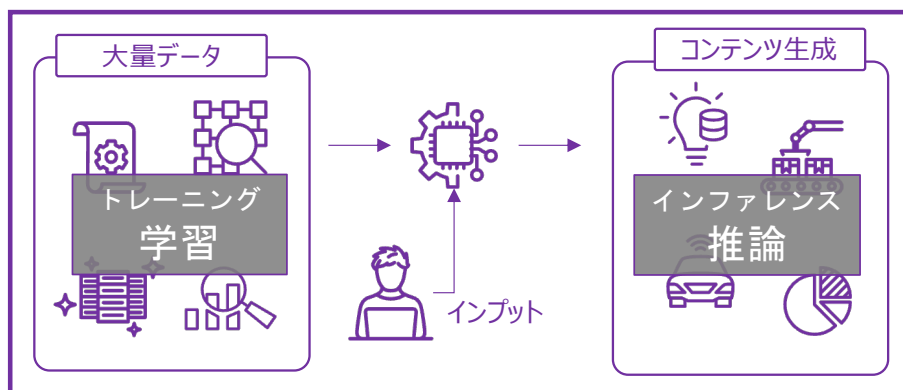
日本法人も2025.11.1よりHPEへ移籍

いままで大変お世話になりました。そして、これからもよろしくお願いします。

# 本セッションの内容

- AI DCの輻輳対策と輻輳管理の振り返り
  1. AIパフォーマンスを高めるための自動チューニング手法 (塚本)
  2. 輻輳回避・負荷分散の新手法  
– RLB with BGP-DPF (Mahesh) ✨
- まとめ

# AI/MLクラスターとNW構成・要件概要の振り返り

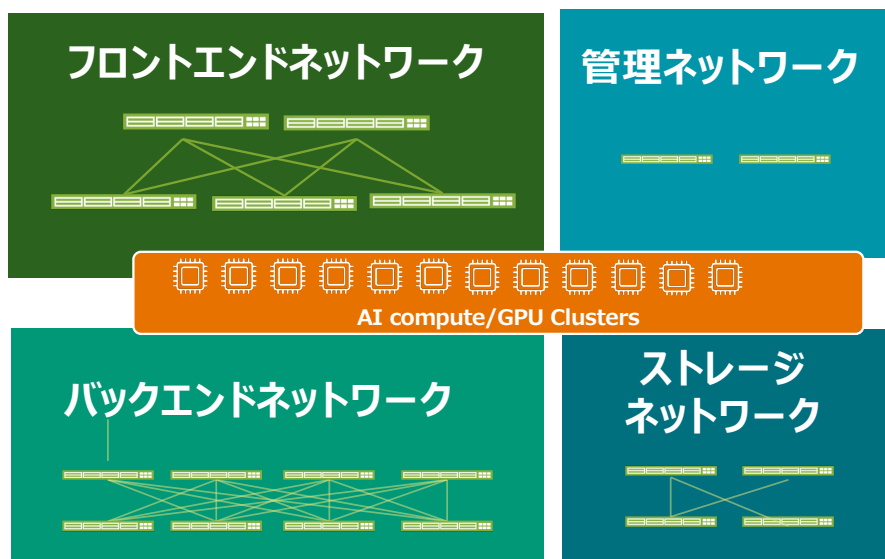


## フロントエンドネットワーク

- 主に推論や外部接続に利用
  - クライアントからのアクセス処理、マルチテナント対応
  - リアルタイムアクセスやRAGなどの低遅延要件も
  - マルチノード推論(大規模, MoE)の期待も

## バックエンドネットワーク

- 学習で利用するGPU間 RDMAネットワーク
  - 各GPU間の高帯域接続(400G/800G/1.6T )
  - Roce v2, ロスレスNW(輻輳対策、輻輳回避)
  - トレーニングタスクはJCT(ジョブ完了時間) 重要

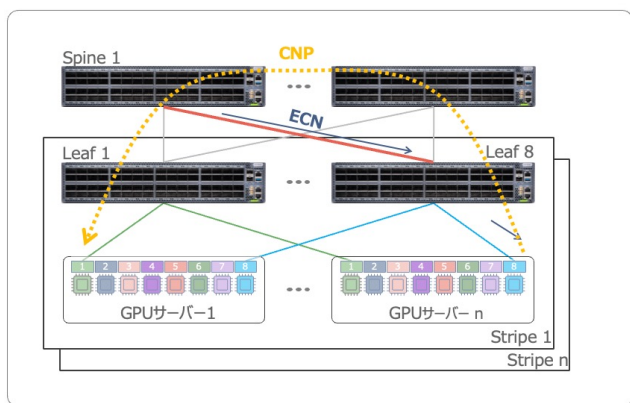




# バックエンド ネットワークの輻輳対策

## 輻輳制御：

輻輳状況をリアルタイムに監視・通知し、  
トラフィック流量調整 & バッファリング, (トリミング)

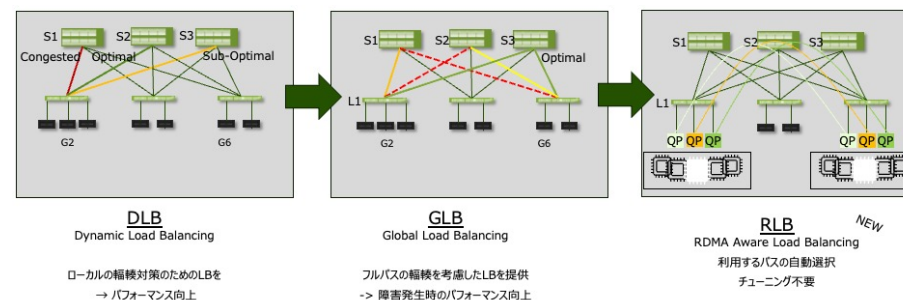


- DCQCN (PFC+ECN)
- PFC Watch-dog
- DCN(Drop Congestion Notification)



## 輻輳回避・負荷分散：

Leaf-Spine間のECMP環境の輻輳を最適化  
トラフィックの偏りを極小化して輻輳を回避



- DLB (Dynamic LB)
- GLB (Global LB)
- sDLB (Selective DLB)\*
- WPS (Weight Packet Spray)\*
- RLB (RDMA-aware LB)\*

\*Juniper Innovation

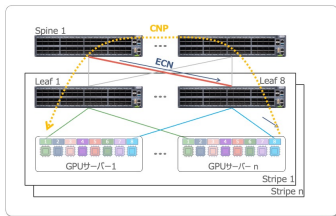
JUNIPER  
NETWORKS

# 本セッションの内容

- AI DCの輻輳対策と輻輳管理の振り返り
  1. AIパフォーマンスを高めるための自動チューニング手法（塚本）
  2. 輻輳回避・負荷分散の新手法  
– RLB with BGP-DPF (Mahesh) ✨
- まとめ

# AIパフォーマンス/JCT 改善のための様々なパラメーター

## 輻輳制御:



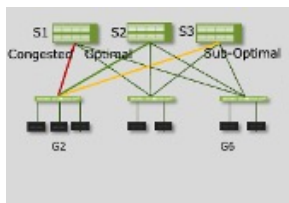
### – DCQCN

- ECN drop-profile - ECN マーク閾値と割合
- PFC X-ON/(X-OFF) - PFC送信/(停止) 閾値
- PFC watchdog interval - PFC pauseストームの緩和

### – Buffer Profile

- Dedicated Buffer/Shared - 共有バッファの割り当て
  - Lossless /Lossy/headroom buffer-partition - キューのバッファ割り当て
  - Alpha Value - 共有バッファからの各キューの動的バッファ割り当て
- etc.

## 輻輳回避:



### – 負荷分散

- ECMP/Flowlet/Pinning/Spraying/Rebalancing etc. - 手法の選択
- DLB/GLB update interval - DLBやGLBのモニタリングや送信間隔
- DLB Inactivity Interval - Flowlet識別のための間隔
- DLB Flowset Table Size - ECMP 数に応じて変更

etc.



# AI DCの選択肢と実証ラボ(Juniper AI Lab)

マルチベンダー環境の事前POC/評価済みデザイン/ベンチマーク/相互接続試験/新技術評価 etc.

最適な負荷分散戦略は何？

最適なDCQCNプロファイルは？

1:1 No oversubが必須？  
2:1でも許容される？

VxLANオーバーレイは  
どのような影響を与える？

レール設計と非レール設計  
どちらが最適？

400Gと800GのJCTの違いは？



SpineのBufferはDeepとShallow  
どちらが適している？

遅延はどのような影響？

Leaf-SpineでPFCを有効化は必須？

障害シナリオにおけるジョブへの影響は？

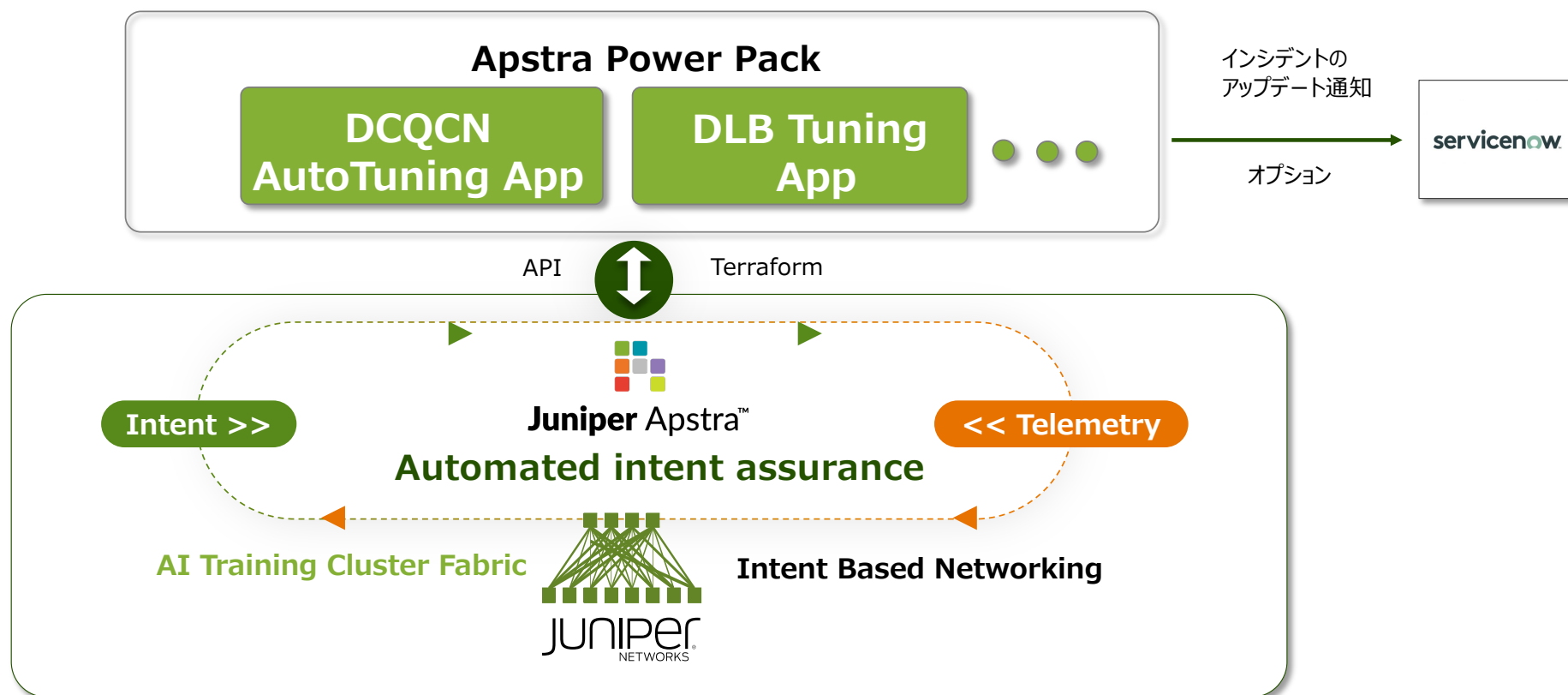
PXNを有効化すべきか無効化すべきか？

NCCLアルゴリズムと  
パフォーマンスは、どのように関連する？

AI DC JVD(Juniper Validated Design)/Tech Field Day/juniper.netなどで公開

# Apstra Auto Tune App

AIパフォーマンス チューニングの運用効率を改善するための自動化



<https://github.com/Juniper/ApstraPowerPacks>

Juniper Public

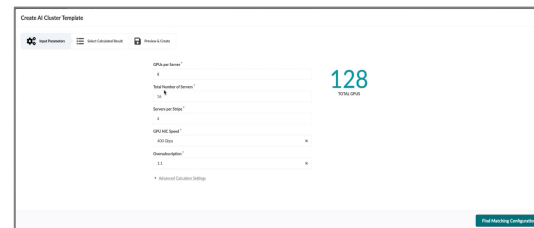
# Apstra - AI Cluster Template

フロントネットワークだけでなく、GPUのバックエンドネットワークの自動化・可視化にも対応

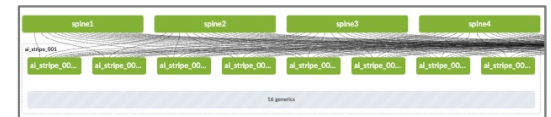


AI DC JVD

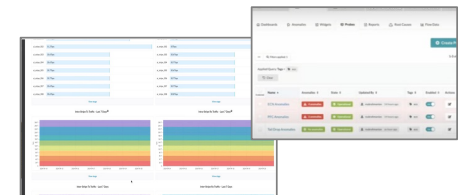
AI クラスターテンプレート  
Simplified Template designer with minimal inputs



Automated Rail-optimized topology

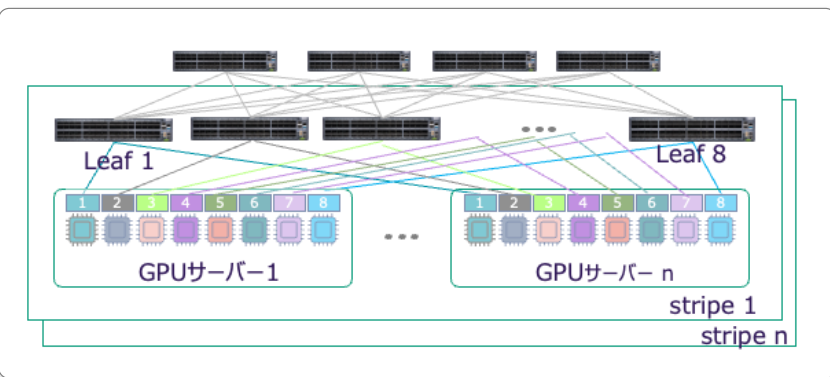
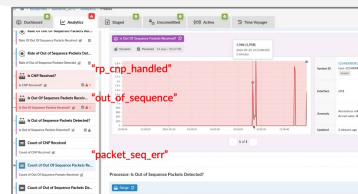


AI DC & GPUモニタリング  
Dashboards Stripe & Rail Traffic  
Queue-Level traffic management visibility

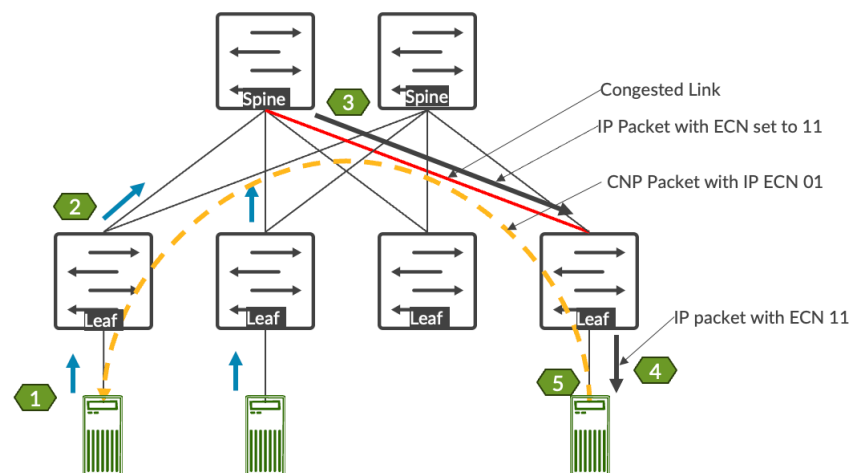


AI Auto Tune App performance tuning

Apstra GPU エージェント  
GPU Hardware Traffic Monitoring probe

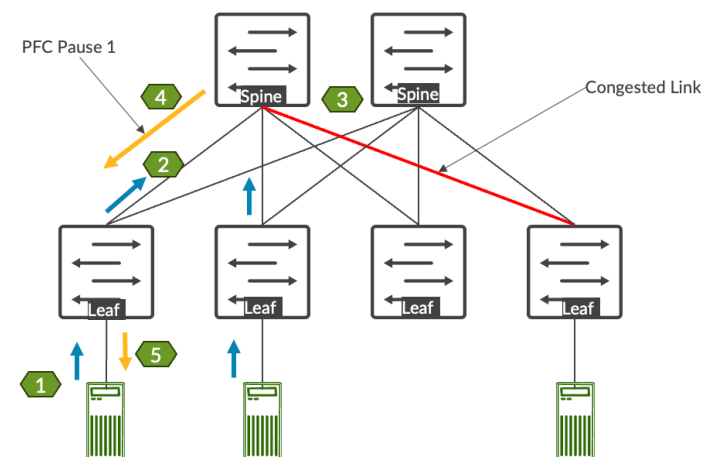


# 輻輳制御 - DCQCN (ECN + PFC)



## ECN(Explicit Congestion Notification)

- Egressキューのバッファ使用率が一定レベルに達したとき (WRED)にパケットにECNマーキング開始
- ECN・CNP受信により送信元が送信レートを調整
- バッファ資料率に応じてマーキングの割合を変化
- ECN - 「早期輻輳通知」- ソフトブレーキ



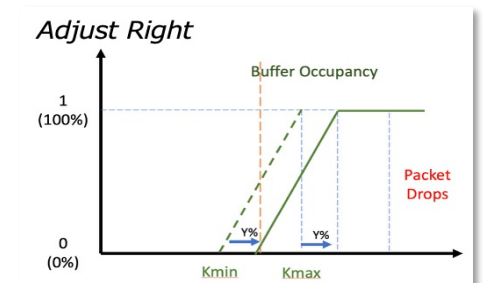
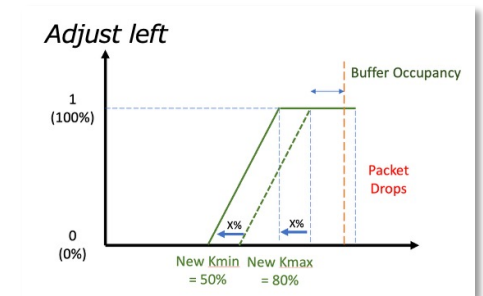
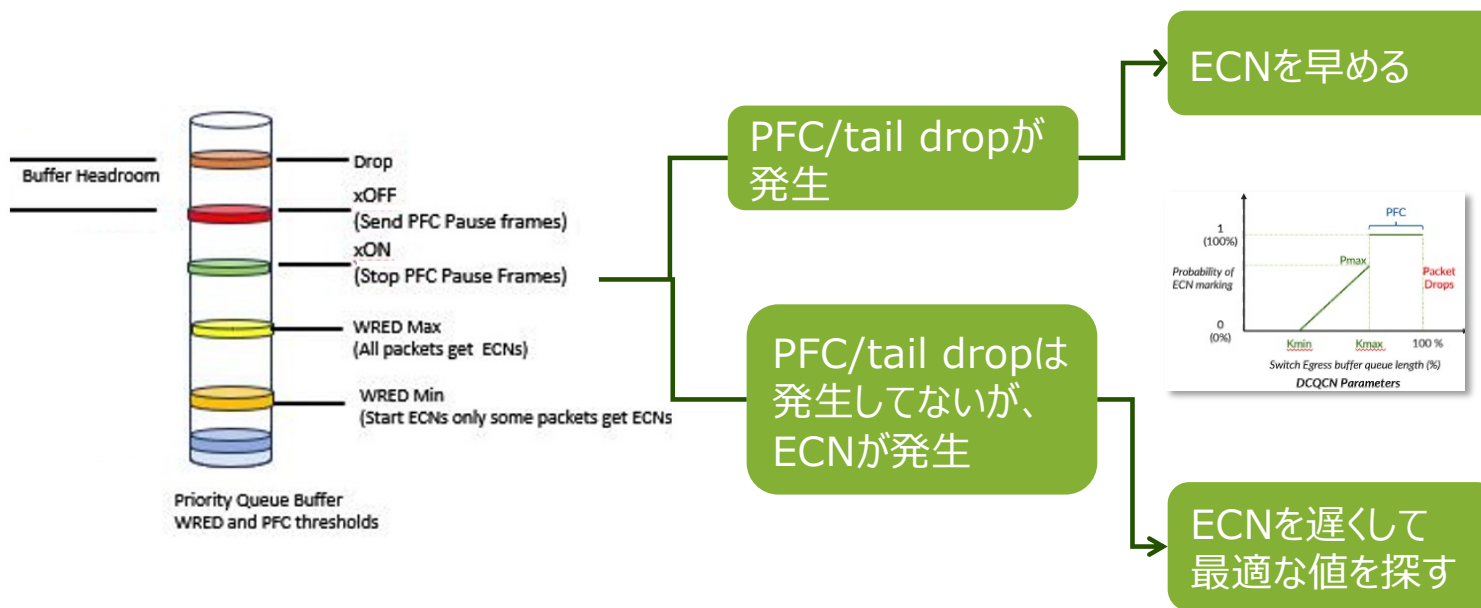
## PFC(Priority Flow Control)

- バッファがさらにある閾値がを超えるとPFC Pause/xOFFを下位スイッチへ送信
- PFC pauseを受信し、xONまでトラフィックxON送信を停止
- PFC - 「緊急停止」- ハードブレーキ

# DCQCN Auto Congestion Algorithm

課題: ECNマーク閾値が低いと送信レートが上がらない  
高すぎるとPFCによる送信停止やパケットドロップ発生

➡ PFC・テールドロップ有無を監視、ECNのタイミングを調整し、最適な値を決定

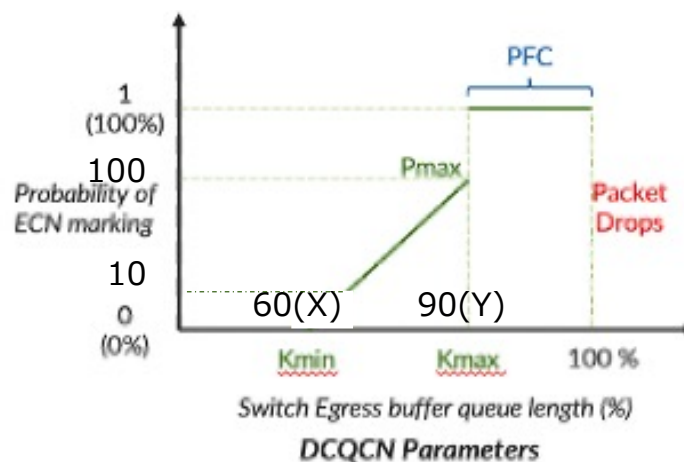




# Apstra DCQCN Auto Tuning App

ECN,PFC,Tail-Dropsを監視 ➡ ECN通知タイミングのDrop Profile(Kmin,Kmax)を調整

ECN(drop-profile)  
fill-level [60 90];  
drop-probability [ 10 100];



Demo: Automated Congestion Management in the AI DC  
<https://www.juniper.net/us/en/the-feed/topics/data-center/automated-congestion-management-in-the-ai-data-center-with-juniper-networks.html>



調整サイクル例 :

ECN & PFC検出 ➡ 下げる - 60/90 → 40/70 (%)



ECN & PFC検出 ➡ 下げる - 40/70 → 30/60 (%)

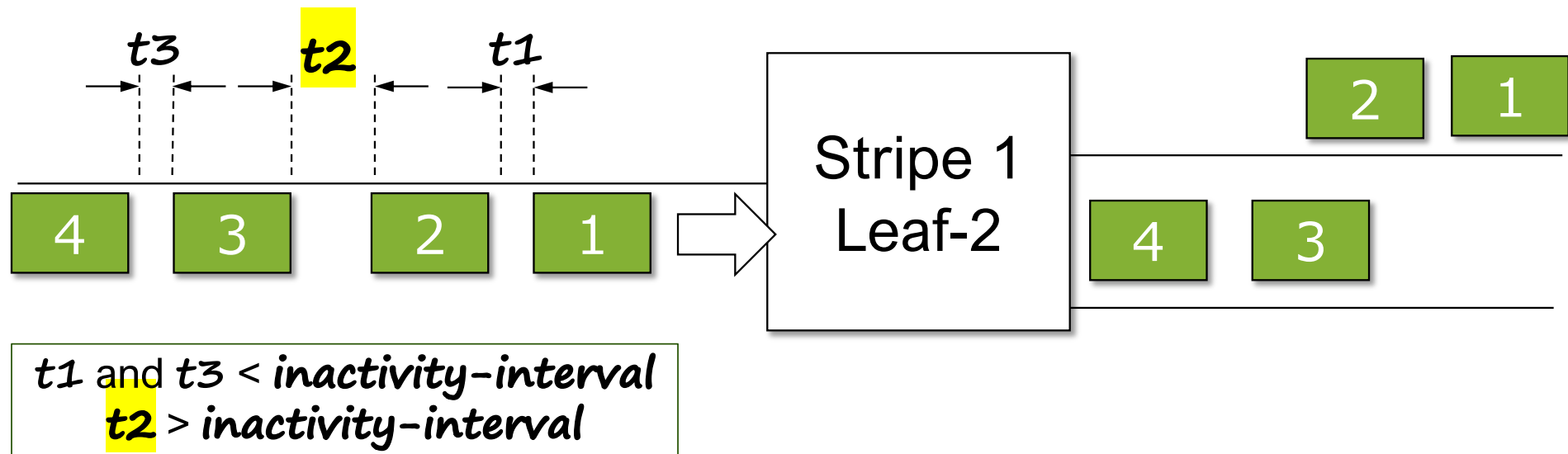


ECN発生 & PFC検出せず ➡ 上げる - 30/60 → 35/65

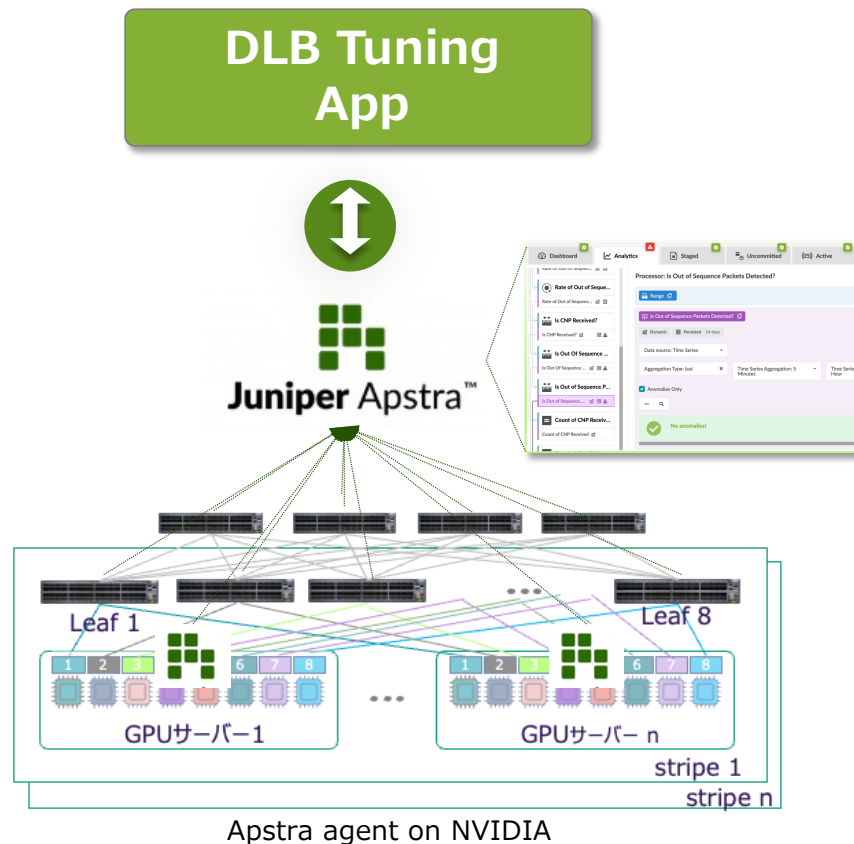
\*ECNの効果を最大化するために、ECN通知が早すぎず、PFCやDropが発生しない、最適な値を決定

# DLB – Dynamic Load Balancing : Flowlet mode

- パケット間のInactivity Interval(アイドル間隔)によりフローを識別し、最適リンクを選択。
  - Intervalが長すぎると、フローが別フローとして認識されづらく偏る可能性あり
  - Intervalが短すぎると、別フローとして認識されるが、リオーダーが発生する可能性あり



# Apstra DLB Tuning App



- GPU側のOOS(Out-of-Sequence)を監視
- OOSが増えていたら
  - inactivity\_interval を増加
    - JUNOS default 256  $\mu$ sec
    - apstra default 128  $\mu$ sec
    - default\_deltaは 24  $\mu$ secずつ増加
- OOSが変わらず
  - 変更なし・最適な値として判断
    - 将来的に減少する実装を予定

# アジェンダ

- AI DCの輻輳対策と輻輳管理の振り返り
- 1. AIパフォーマンスを高めるための自動チューニング手法（塚本）
- 2. 輻輳回避・負荷分散の新手法  
– RLB with BGP-DPF (Mahesh)
- まとめ

# AI DC Software

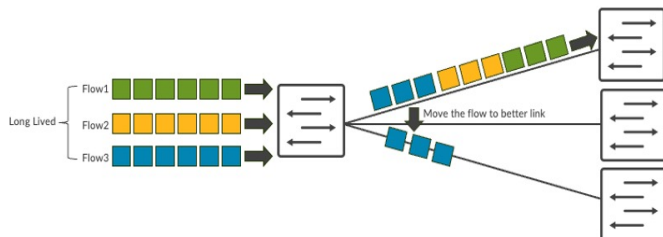


## Congestion Avoidance

(Effective utilization of fabric capacity)

### Efficient Load Balancing Features

- DLB (Dynamic LB)
- GLB (Global LB)
- sDLB (Selective DLB)\*\*
- WPS (Weight Packet Spray)\*\*
- RLB : GPU/QP pinning & BGP-DPF\*\*



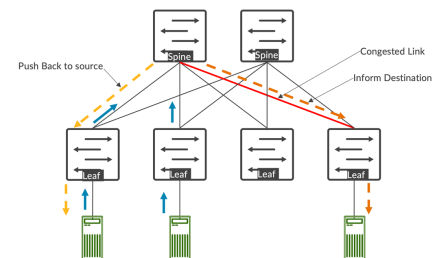
\*\* - Juniper Innovations

## Congestion Control

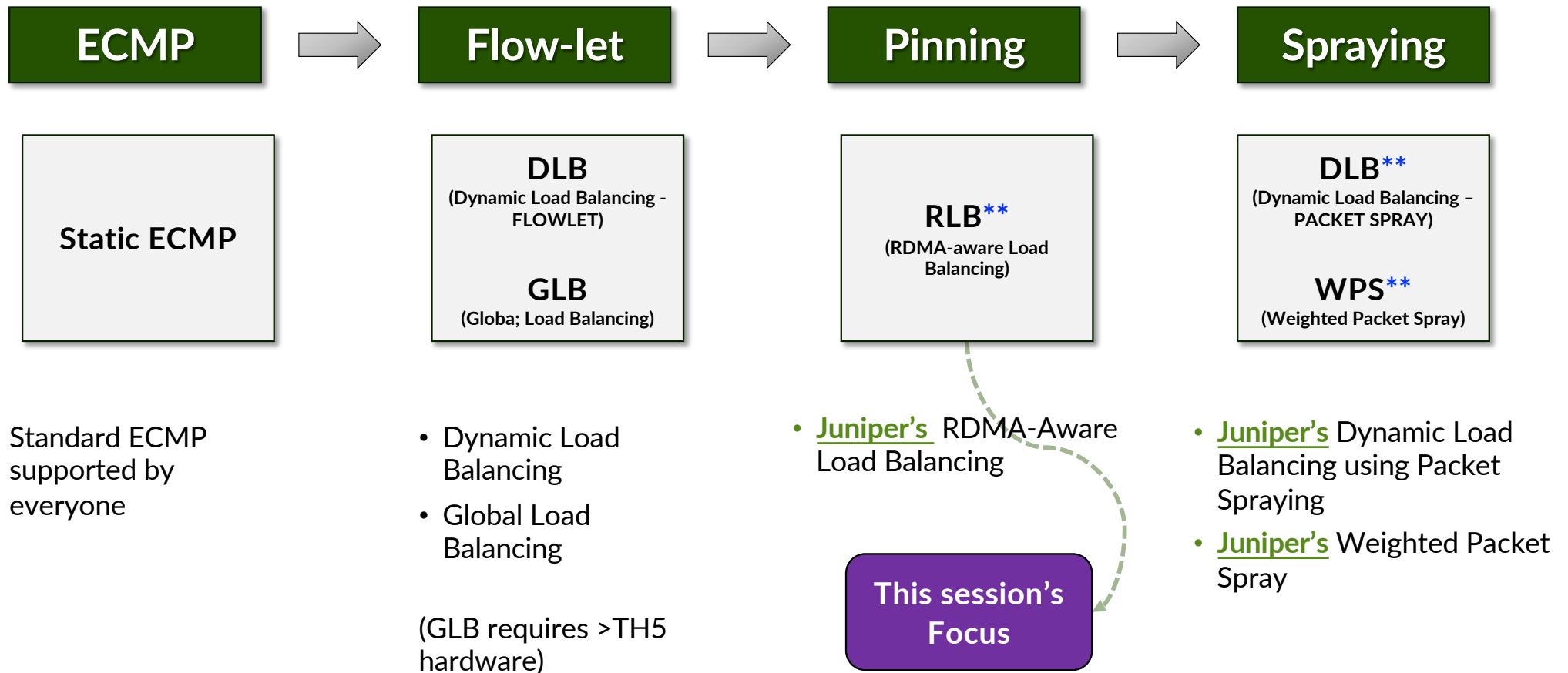
Preventing packet drops (lossless)

### Congestion Management Features

- DCQCN (PFC+ECN)
- Telemetry:
  - Interface level
  - Queue level
- PFC Watch-dog
- Per-Q Alpha Threshold value support



# Load Balancing Evolution





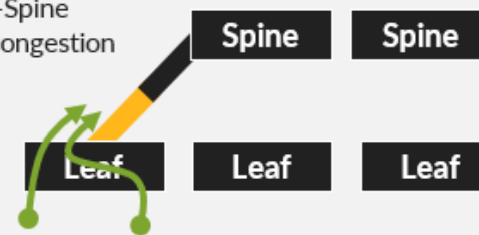
# Launching AI Load Balancing

The Problem

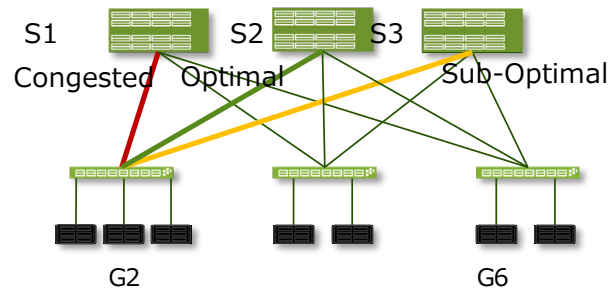
Spine-to-Leaf  
downlink congestion



Leaf-to-Spine  
uplink congestion

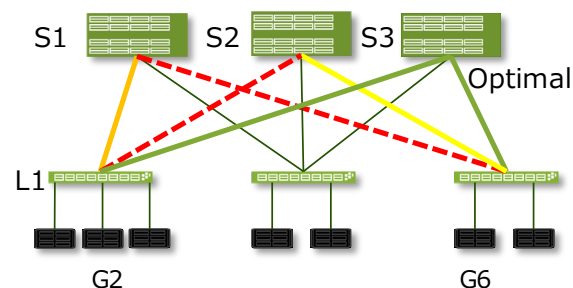
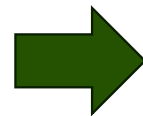


There are fewer large-sized elephant flows in AI Training networks making the problem worse



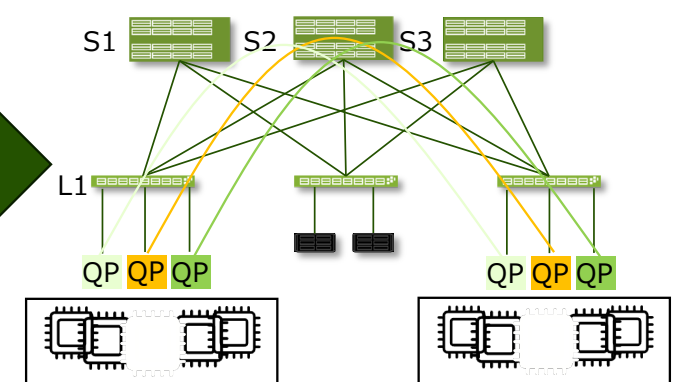
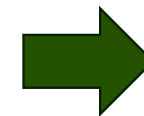
DLB-  
Dynamic Load Balancing

Delivered local congestion aware LB ->  
Better performance



GLB-  
Global Load Balancing

Delivered full path congestion aware LB  
-> Better performance during failures



RLB-  
RDMA Aware Load Balancing

Highest Performance  
Lowest variability

No tuning

JUNIPER  
NETWORKS

# RLB:RDMA-aware Load Balancing

## Highest Consistent Performance

Eliminates Congestion in most scenarios

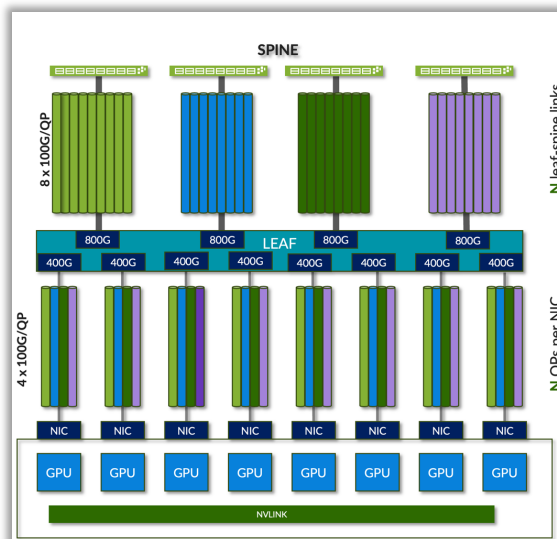


## In-order Delivery

Delivers packets in order improving JCT



## RDMA-aware Load Balancing



## Expensive HW not required

Doesn't require expensive HW  
(SuperNICs or deep-buffer switches)



## Delivers non-rail performance

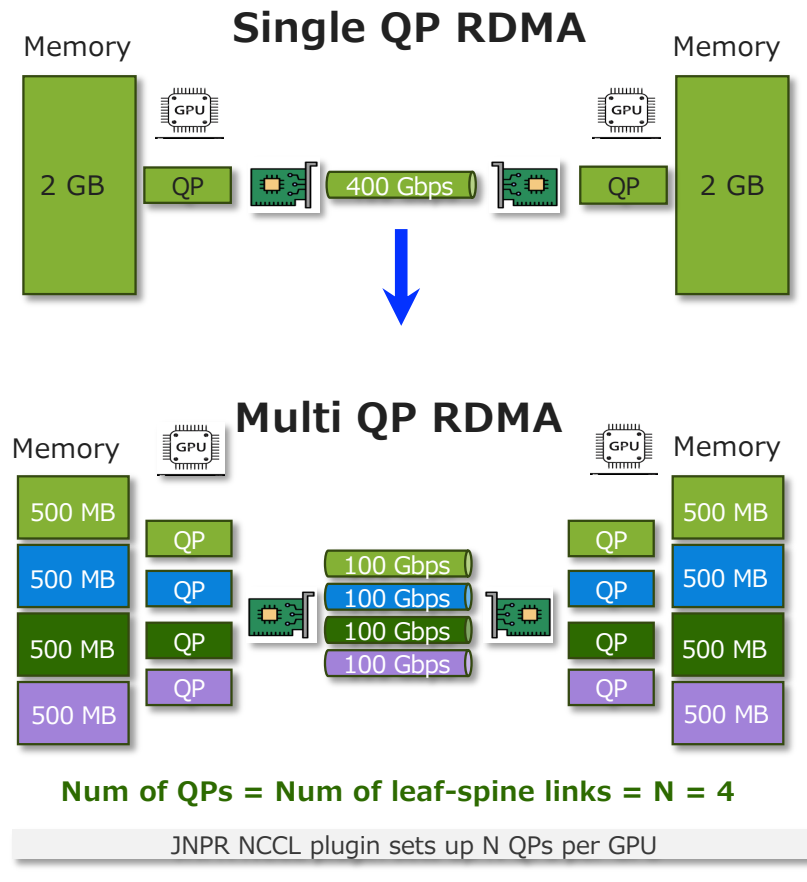
Delivers highest performance even for non-rail, lowering TCO.



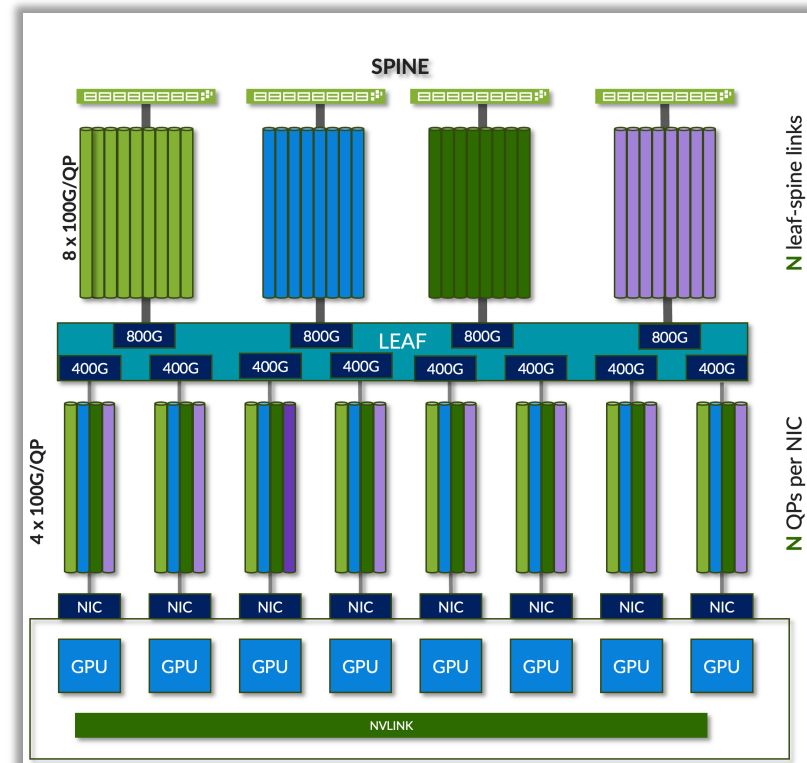
Ditch the proprietary tax: scale smarter with Juniper AI load balancing

<https://blogs.juniper.net/en-us/industry-solutions-and-trends/ditch-the-proprietary-tax-scale-smarter-with-juniper-ai-load-balancing>

# Deterministic RDMA Flow Pinning

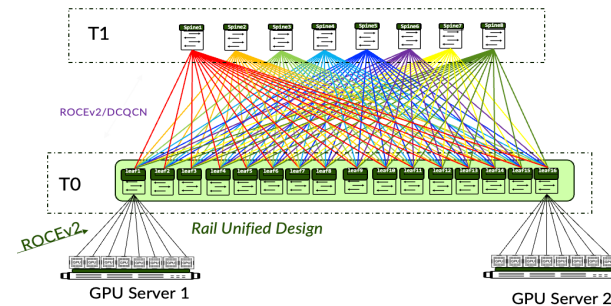


ROCEv2 BTH Header **QPAIRs**  
are pinned to different paths.



# Traffic Engineer in the DC IP Fabrics

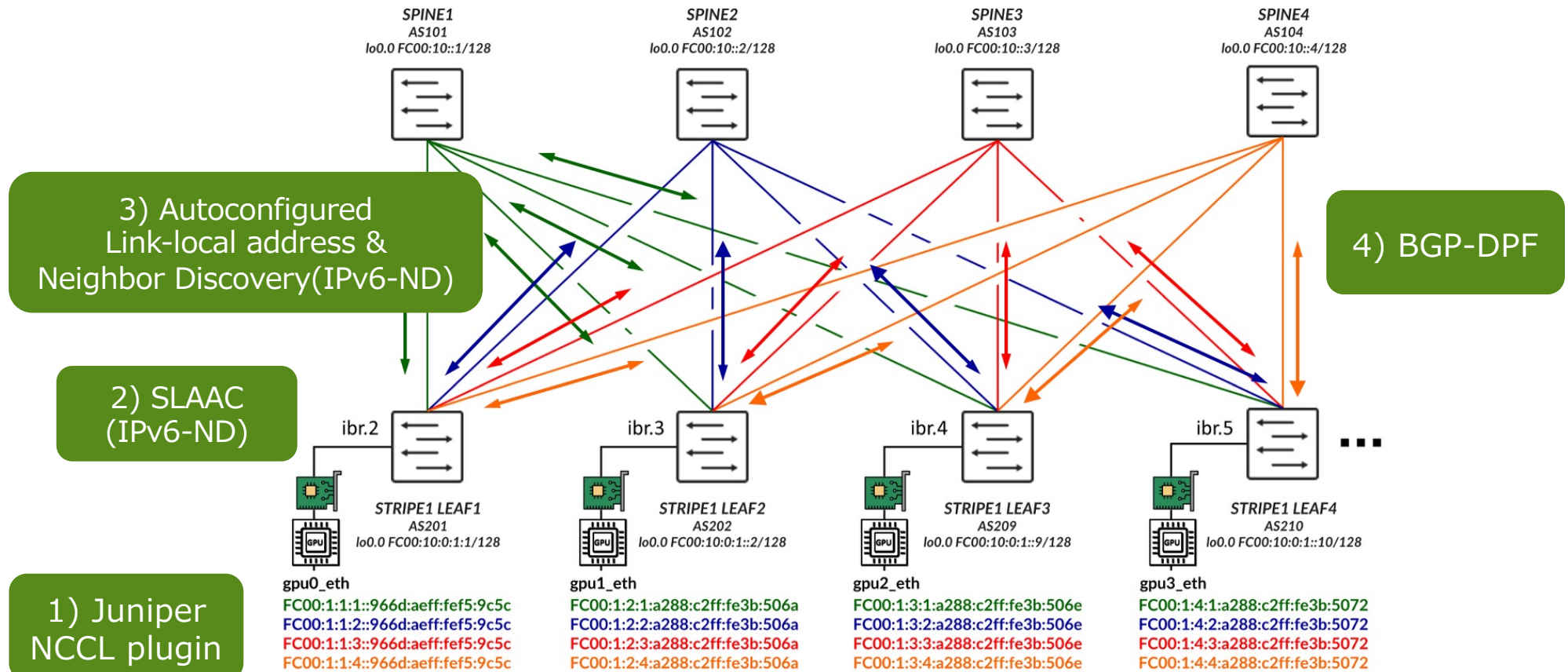
## - DPF



Deterministic Path Forwarding (DPF)

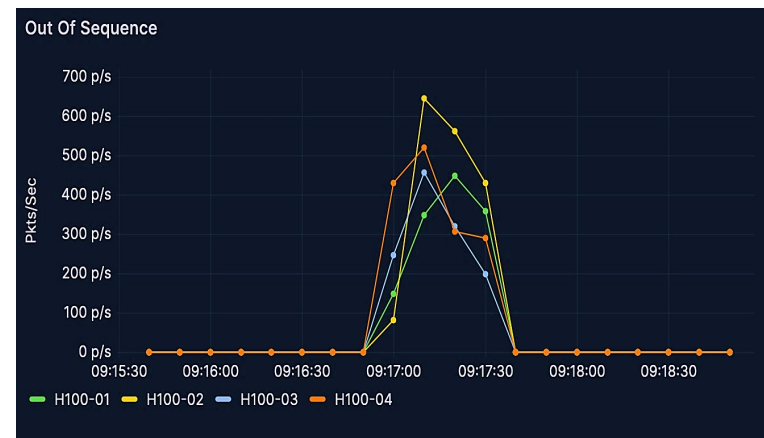
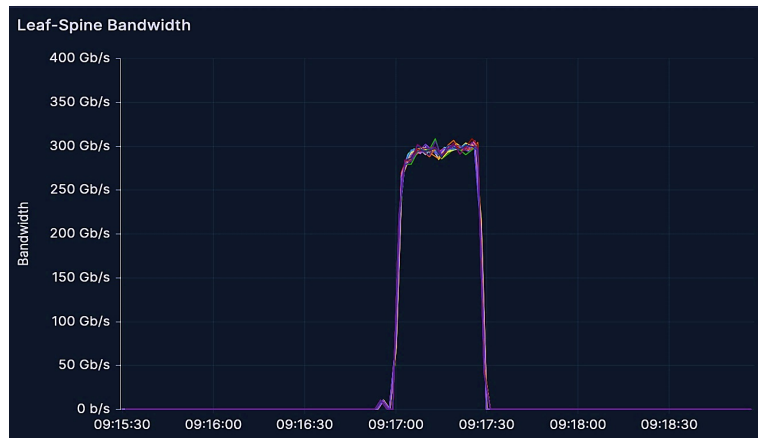
- RSVP-TE was traditionally used for traffic-engineering
- Many data center networks uses pure IP without MPLS
- BGP is usually used as the only routing protocol in the DC IP Fabrics
- BGP DPF is an alternative to RSVP-TE for the DC IP Fabrics
- Separate a physical IP fabric into multiple logical fabrics
- Flows can be mapped to different logical fabrics for:
  - Different SLAs
  - Load balancing purpose
  - No fate sharing                      etc...
- A lightweight traffic engineering for IP fabrics

# RLB and BGP-DPF:Control Plane



# DLB vs. RLB performance

DLB



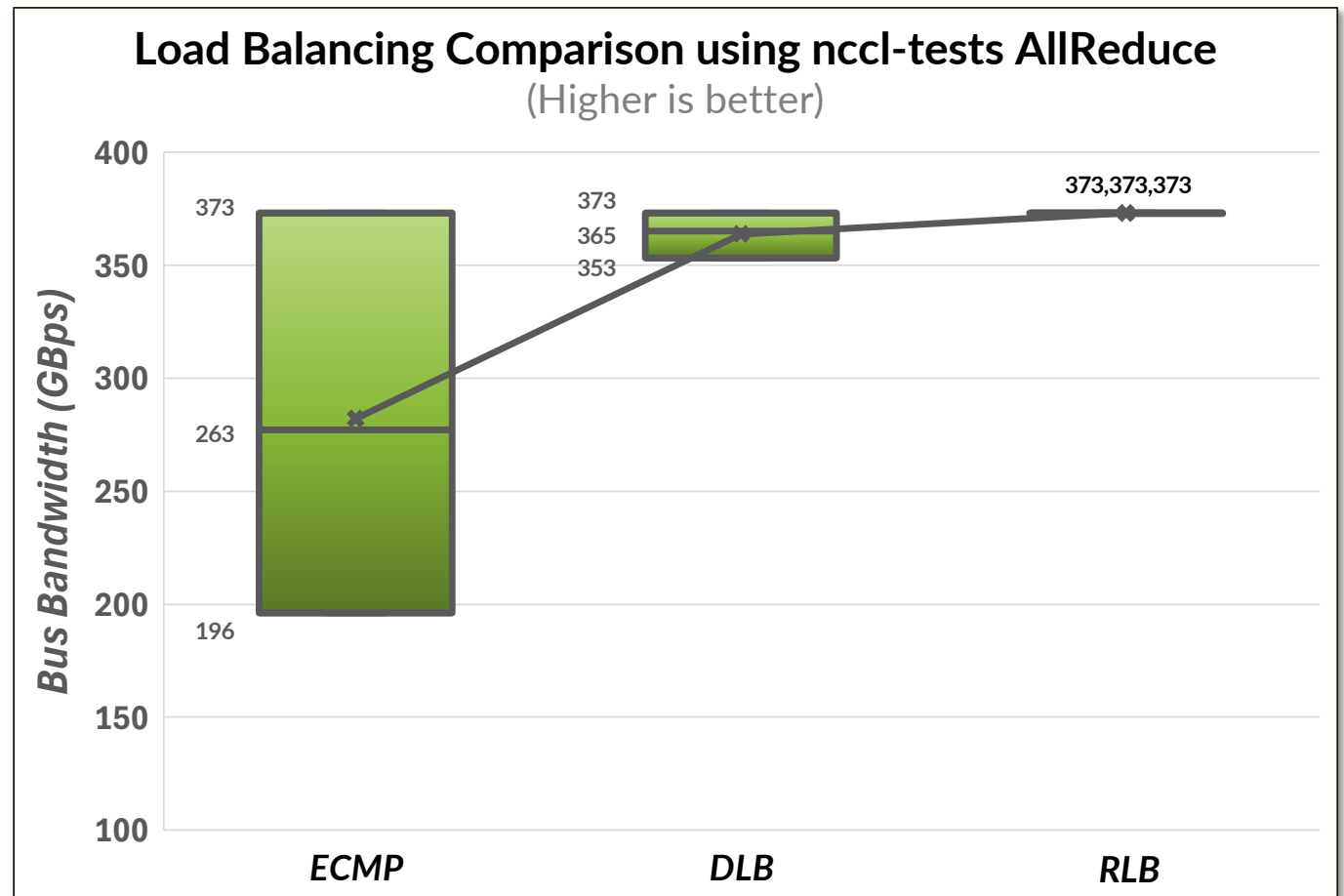
RLB



# Performance Variability Comparison

## Performance Variability with 3 consecutive tests

- ECMP has the highest **variability**.
  - 196, 263, 373 GBps
- DLB has **good** performance variability
  - 353, 365, 373 GBps
- RLB delivers **highest** consistent performance
  - No performance variability (373,373,373) GBps





# アジェンダ

- AI DCの輻輳対策と輻輳管理の課題の振り返り
- 1. AIパフォーマンスを高めるための自動チューニング手法
- 2. 輻輳回避・負荷分散の新手法  
– RLB with BGP-DPF
- まとめ



# PUBLISHED AI JVDs

## Juniper Validated Design

JUNIPER NETWORKS | Engineering  
Simplicity

AI Data Center GPU Backend Fabric with  
Juniper RDMA-aware LB and BGP-DPF—  
Juniper Validated Design (JVD)

Published  
2025-09-29



AI Data Center Network with  
Juniper Apstra, NVIDIA GPUs,  
and WEKA Storage

Published Q4 2024

AI Data Center Network with Juniper  
Apstra, AMD GPUs, and Vast Storage



Published Q1 2025, later enhanced with  
Pollara



AI Data Center Multitenancy  
with EVPN/VXLAN

Published Q2 2025

AI Data Center GPU Backend Fabric  
with Juniper RDMA-aware LB and  
BGP-DPF—Juniper Validated Design  
(JVD). Published Q3 2025



# Juniper's AI DC Networking Solution

AIのポテンシャルを開放し比類のないAIパフォーマンスと容易なオペレーションを実現



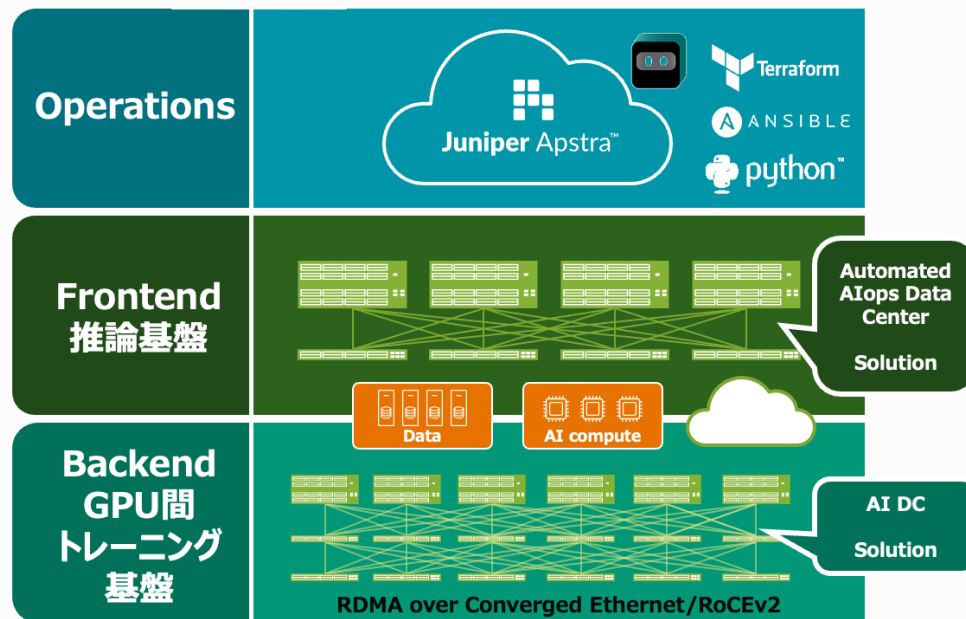
Openness



Operations



E2E Solutions



800G



2024 800GE  
No.1 OEM  
Market Share

Public

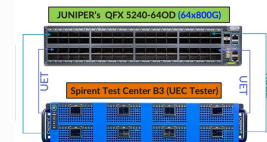


kakao



Ultra Ethernet  
Consortium

Interop Tokyo  
世界初UEC相互検証





# Thank you

**JUNIPER**  
NETWORKS | Driven by  
Experience™