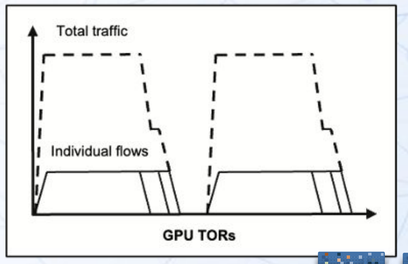
ARISTA

AIファブリックに関する3つの画期的なアイデア

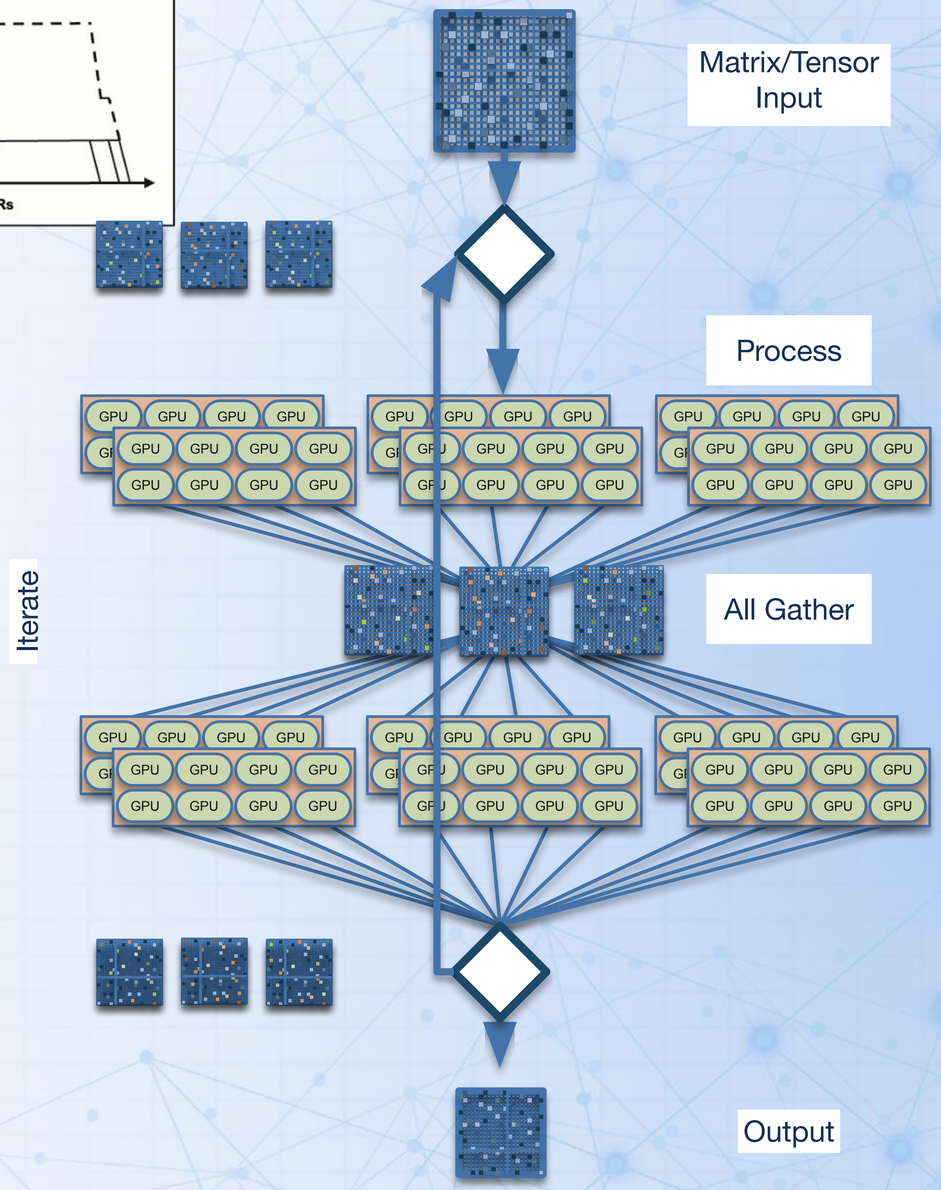
～Three Genius Ideas for AI Fabrics～

Shishio Tsuchiya
shtsuchi@arista.com

AI/MLネットワーク要件



- スケールアウト要件
  - 100 -> 1000アクセラレーター->100K アクセラレーター
- ノンブロック高帯域
  - **400G->800Gのコンピュートノードへの接続**
  - **ゼロオーバーサブスクライブ** と回復力のあるフェイルオーバー
- 難しいトラフィックプロファイルへの対応
  - 大きなメッセージサイズ
  - 少ないエントロピーの長時間フロー
  - 同期性の高いバースト
- ロスレスフォワーディング
  - RoCEとDCQCNをサポートするPFCとECN
  - **輻輳を起こさないロードバランシング**
  - 輻輳解消のためのバッファリング



AIトレーニングのワークロードはネットワークパフォーマンスの低下や遅延の影響を受けやすく極めて相関性が高い

Resilient AI Supercomputer Networking using MRC and SRv6

https://cdn.openai.com/pdf/resilient-ai-supercomputer-networking-using-mrc-and-srv6.pdf

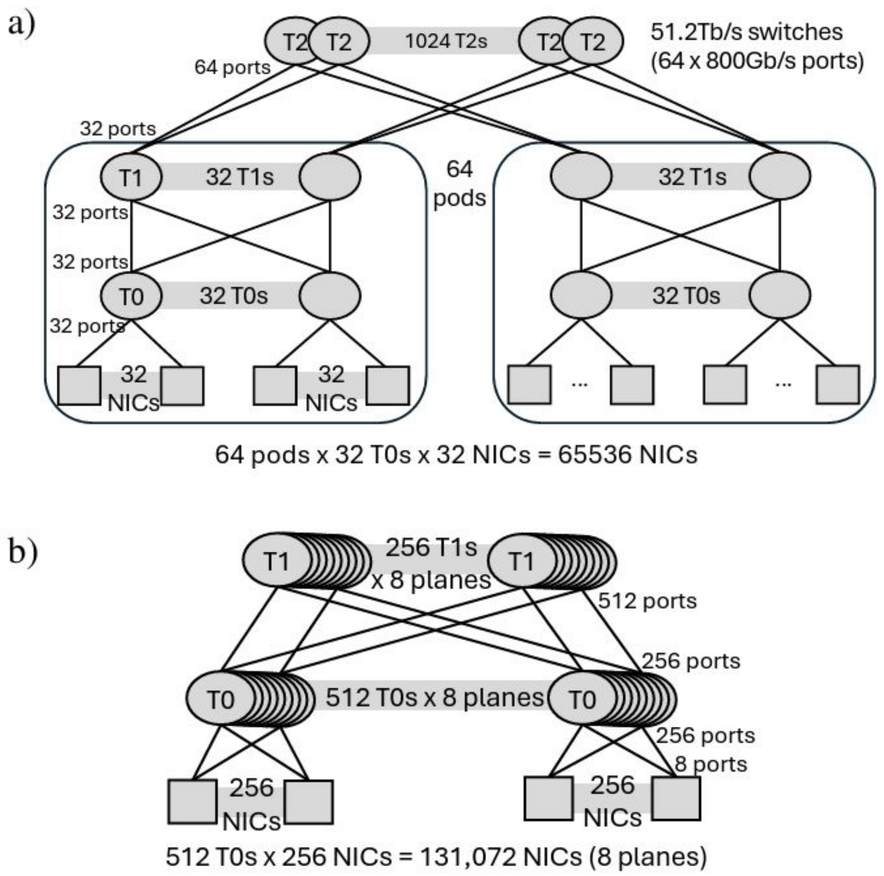


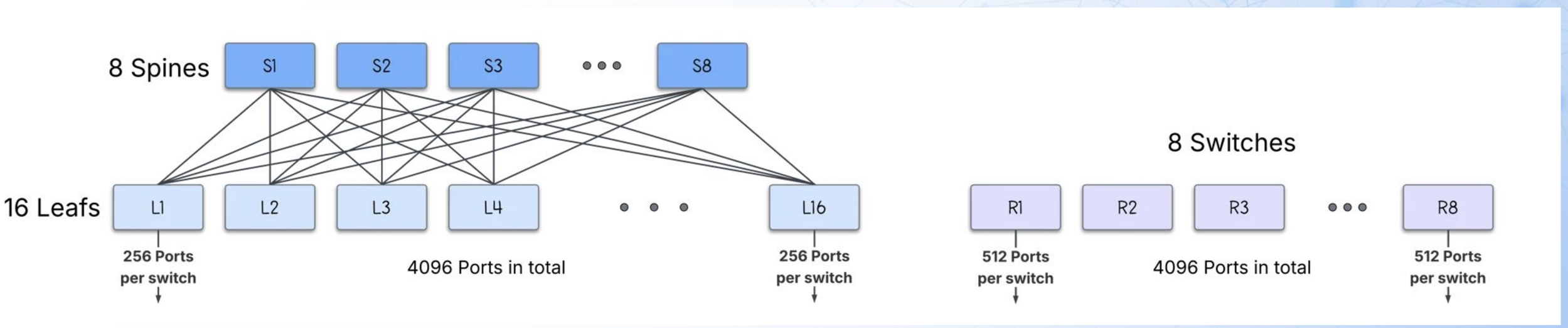
Figure 1: (a) 3-Tier 800 Gb/s single-plane topology vs (b) 2-Tier 8x100 Gb/s multi-plane topology

- Microsoft, OpenAI, AMD, Nvidia,Broadcomが論文を発表
- 従来のRoCE通信だと問題になっていたフロー衝突による輻輳,TCP incast問題,リンクやスイッチの障害によるjob中断を下記3つの手法で解決する為のアプローチとそのテスト結果を共有した
  - マルチパスRDMAプロトコル”MRC (Multipath Reliable Connection)”
  - Multi-plane Clos Topology
  - SRv6による静的ソースルーティング

 Copyright © Arista 2026. All rights reserved.

Public.

マルチプレーンリーフスパインネットワーク



- 合計4096ポートを提供する場合
  - 標準的なclosトポロジー
    - 24台の512ポートスイッチが必要
  - 8プレーンネットワーク
    - わずか8台のスイッチで済む
- 各800G NICが8つの独立した100Gポートに分岐し、各NICが各プレーンに100G接続。プレーンをまたぐことなく、8つの異なる方法で1つのNICから他のNICへアクセス可能。

 Copyright © Arista 2026. All rights reserved.

Public.

ARISTA

800Gブレークアウトオプション

https://www.arista.com/assets/data/pdf/Datasheets/Arista-800G_Optics_FAQ.pdf

QDD/OSFP 800G-2XDR4, 800G-2PLR4, 800G-XDR8, LPO-800G-2DR4, LPO-800G-DR8 supported modes								
Logical port configuration								
Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7	Lane 8	Notes
400G-4 (400G-XDR4 / PLR4)				400G-4 (400G-XDR4 / PLR4)				1
800G-8 (800G-XDR8 / PLR8)								2
200G-2		200G-2		200G-2		200G-2		3
100G-1 (100G-FR / LR)	100G-1 (100G-FR / LR)	100G-1 (100G-FR / LR)	100G-1 (100G-FR / LR)	100G-1 (100G-FR / LR)	100G-1 (100G-FR / LR)	100G-1 (100G-FR / LR)	100G-1 (100G-FR / LR)	4

- 通常2本の400G-XDR4/PLR4リンクは1つの800Gポートを集約する為に使用
- 800G-8は800Gのポイントーツーポイントのリンクで使用される
- 200G-2は4つの200Gリンクに分割する事が可能
- 8x100Gモードは8つの100G-FR/100G-LRに分割可能で1つの800Gポートに集約可能

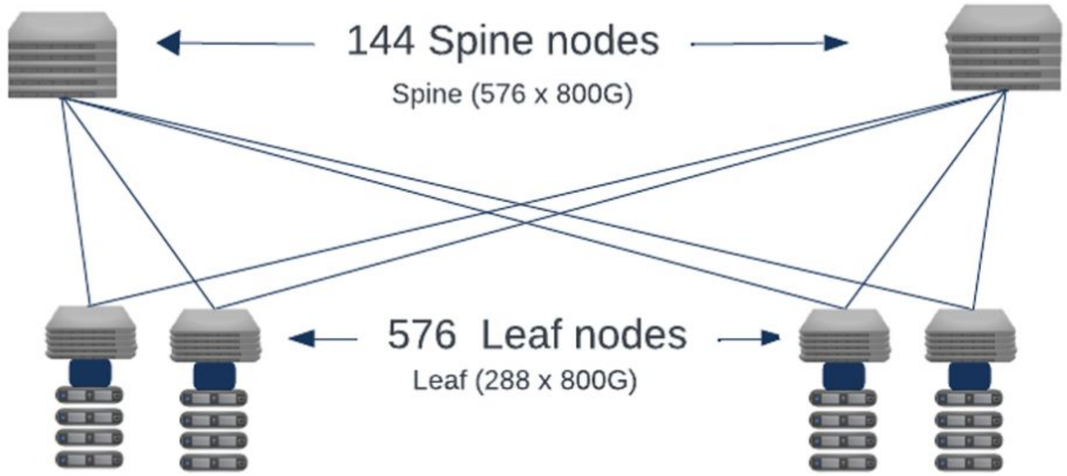
 Copyright © Arista 2026. All rights reserved.

Public.

ARISTA

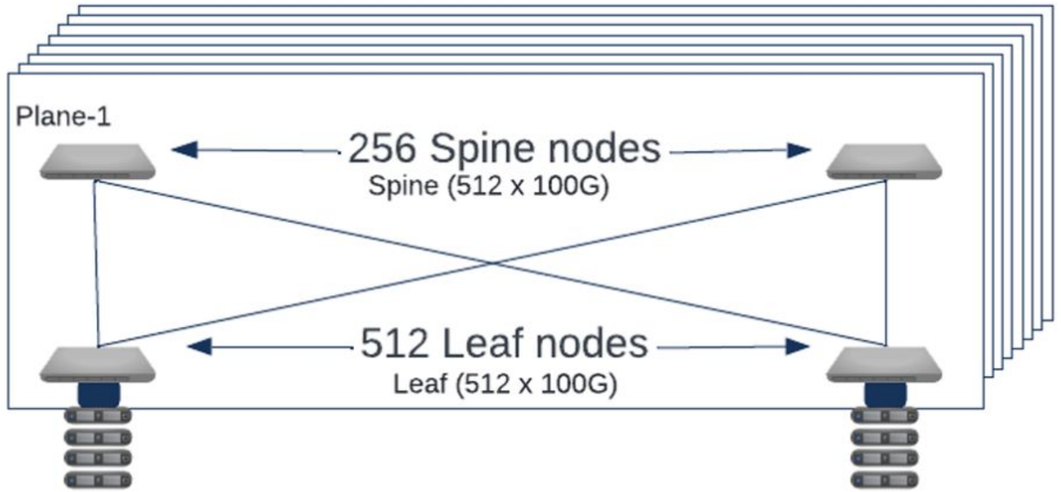
Single PlaneとMulti-Planeの比較

Single plane
Two-tier Scale Out fabric



Spineあたり接続数: 576 Leafs (@800G)
Leafあたり接続数: 144 XPUs (@800G)
総サポート規模: 最大82K × 800G XPU

Multi-plane
Two-tier Scale Out fabric



Spineあたり接続数: Planeごとに512 Leafs (@100G)
Leafあたり接続数: Planeごとに256 XPUs (@100G)
総サポート規模: 最大131K × 800G XPU

- ポート分割を活かしたマルチプレーン・アーキテクチャ
- 各プレーンへの接続速度を下げて(800G->100G)スイッチのRadix（収容数）を増やすことで、XPUあたりの総帯域幅を妥協せず、13万台超の2層ファブリックを実現。

 Copyright © Arista 2026. All rights reserved.
Public.
ARISTA

Multi Planeのメリットとチャレンジ

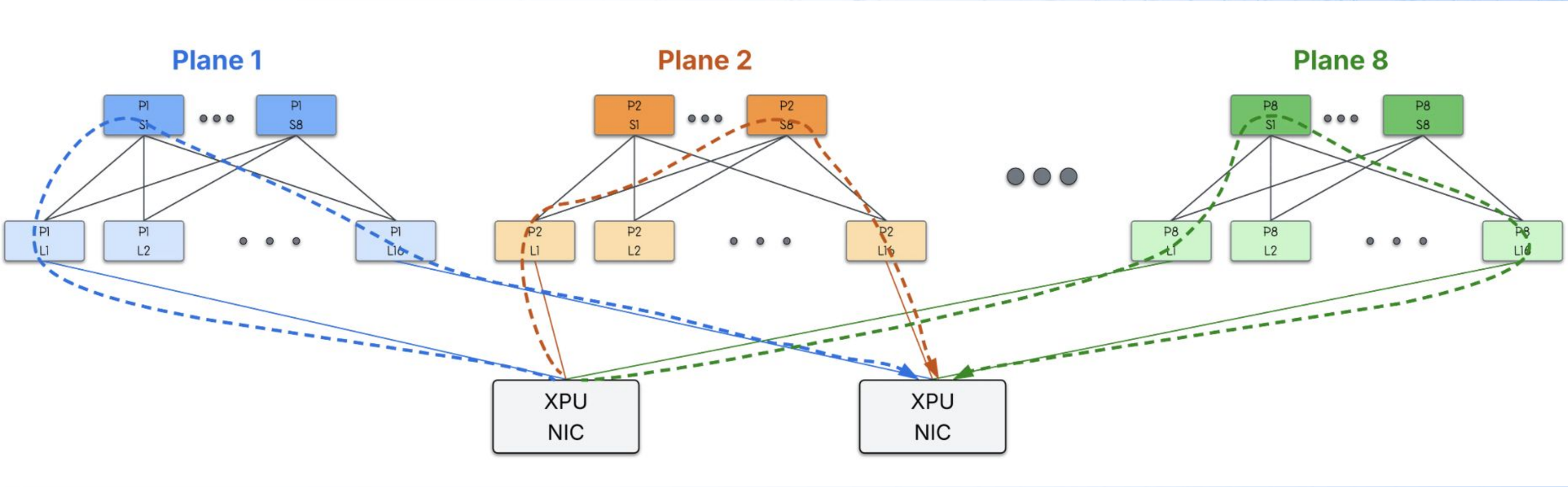
- メリット
  - ホップ数が少ない為、伝送遅延が最小に出来る
  - 1ホップで到達可能なノードが大幅に向上
  - コストと消費電力が削減される
  - Spine-Leaf間のリンク障害の影響が削減される。
  - ホスト-Leaf間のリンク障害が起こっても学習のためのJobを継続出来る
- チャレンジ
  - ワークロードはリンクやNICポートの障害が発生しても正常に動作し続けられる必要がある
  - ネットワークを最大限に活用するには、すべてのネットワークプレーンに均等に負荷を分散させ、1つのプレーン内の多数のパス間で負荷分散を行うと同時に、フローの競合によるパフォーマンスの低下を招かないようにする必要がある

 Copyright © Arista 2026. All rights reserved.

Public.

ARISTA

Multipath Reliable Connection (MRC)

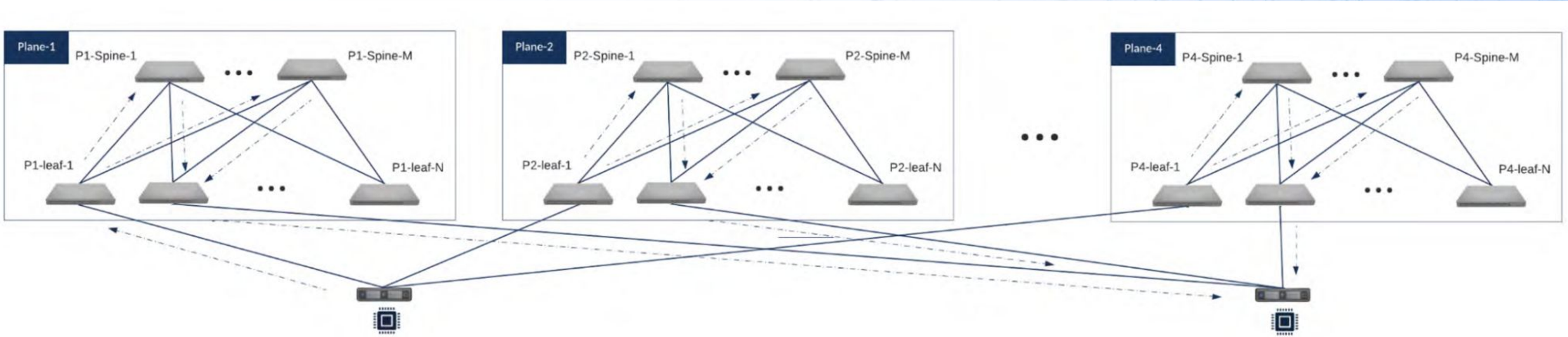


- エンドステーションのNICは受信側へのトラフィックを複数のリンクとパスへ分散。順不同なパケットは自動的に処理される
- MRCはECNとパケットトリミングに応答し、最もパフォーマンスの高いパスに負荷を分散させる

 Copyright © Arista 2026. All rights reserved.

Public.

Multipath Reliable Connection (MRC)



- MRCはRoCEv2トランスポートを拡張し既存の命令(Verb)は維持
  - サポート対象をメモリ書き込み操作（Write / Write-IMM）に限定
- Packet Spray/輻輳通知/パケットトリミングなどUEC機能の仕様を統合
- ファブリック内の特定の物理パスへの依存を排除し、トラフィックを複数経路に分散
  - 順序逆転(Out-of-order）のパケットは受信側で処理をする

 Copyright © Arista 2026. All rights reserved.

Public.

UEC Traffic Forwarding

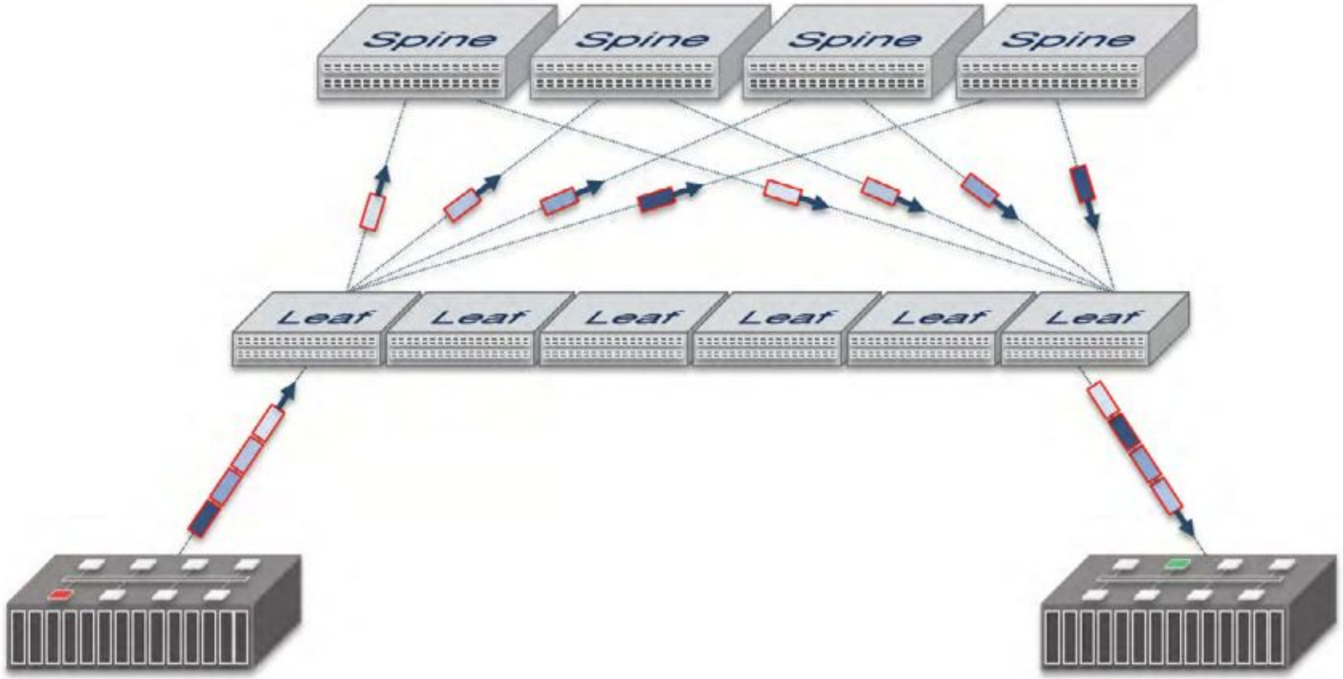


Figure 3: Packet Spraying Overview

- フローベースの負荷分散（ECMP）から、NICによるパケット単位の分散（Packet Spraying）へシフト
- NICは順序逆転（Out-of-order）で到着したパケットを処理し、直接システムメモリへ展開
- 欠落したパケットのみの選択的再送（Selective Retransmit）が可能に

 Copyright © Arista 2026. All rights reserved.

Public.

UEC Congestion Management

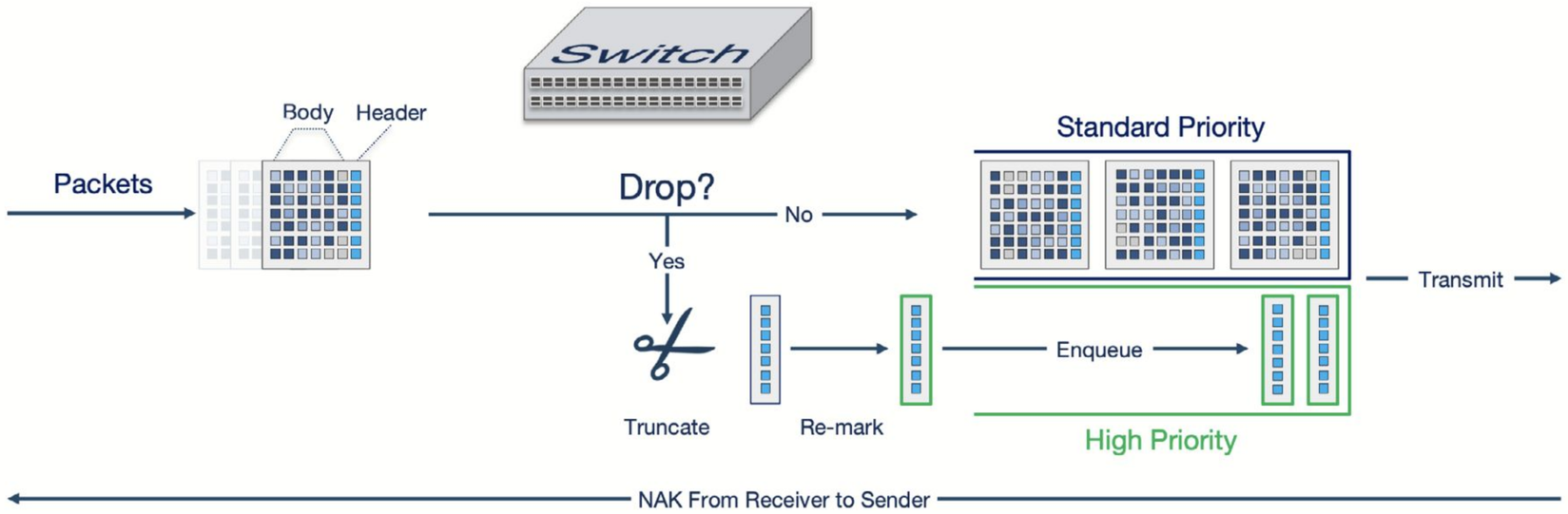


Figure 4: Packet Trimming Mechanism

- 従来はシーケンス番号でパケットの取りこぼしを検出 go-back N
  - 1つのパケット損失の為に多くのパケットが再送され非効率
- UECでは、輻輳スイッチに到達した際、パケットを完全廃棄せずヘッダーのみにトランケーション（切り詰め）」する
- Truncationされたパケットは優先キューに入れられ、宛先はこれを見てNAKの形で送信元に戻す
- これは2つの大きな役割を果たす
  - ドロップした事を明示的に通知
  - 輻輳通知となり、転送レートを落とし、輻輳パスを通らない様に迂回する

 Copyright © Arista 2026. All rights reserved.

Public.

ARISTA

UEC Advanced Connection Setup and Host-based Flow Control

- UECでは輻輳を未然に防ぐ仕組みとして“Ephemeral Connections” (儚い接続)と2つの輻輳制御スキームを導入
- エフェメラル・コネクションは、データが流れ始めるまでのラウンドトリップ・ハンドシェイクの遅延をなくすことで、高速な接続開始を可能にする。明示的な終了も必要無い

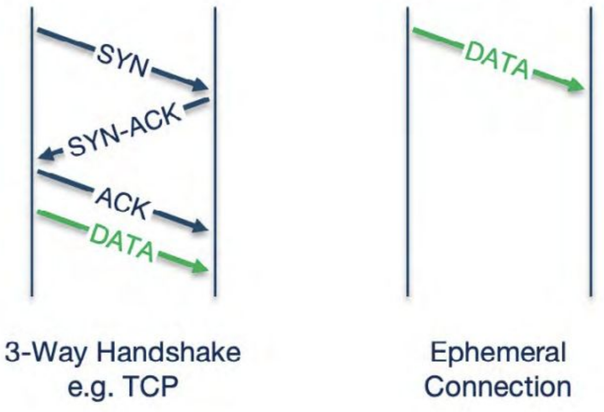


Figure 5: Comparing 3-Way Handshake to Ephemeral Connections

 Copyright © Arista 2026. All rights reserved.

Public.

UEC Advanced Connection Setup and Host-based Flow Control

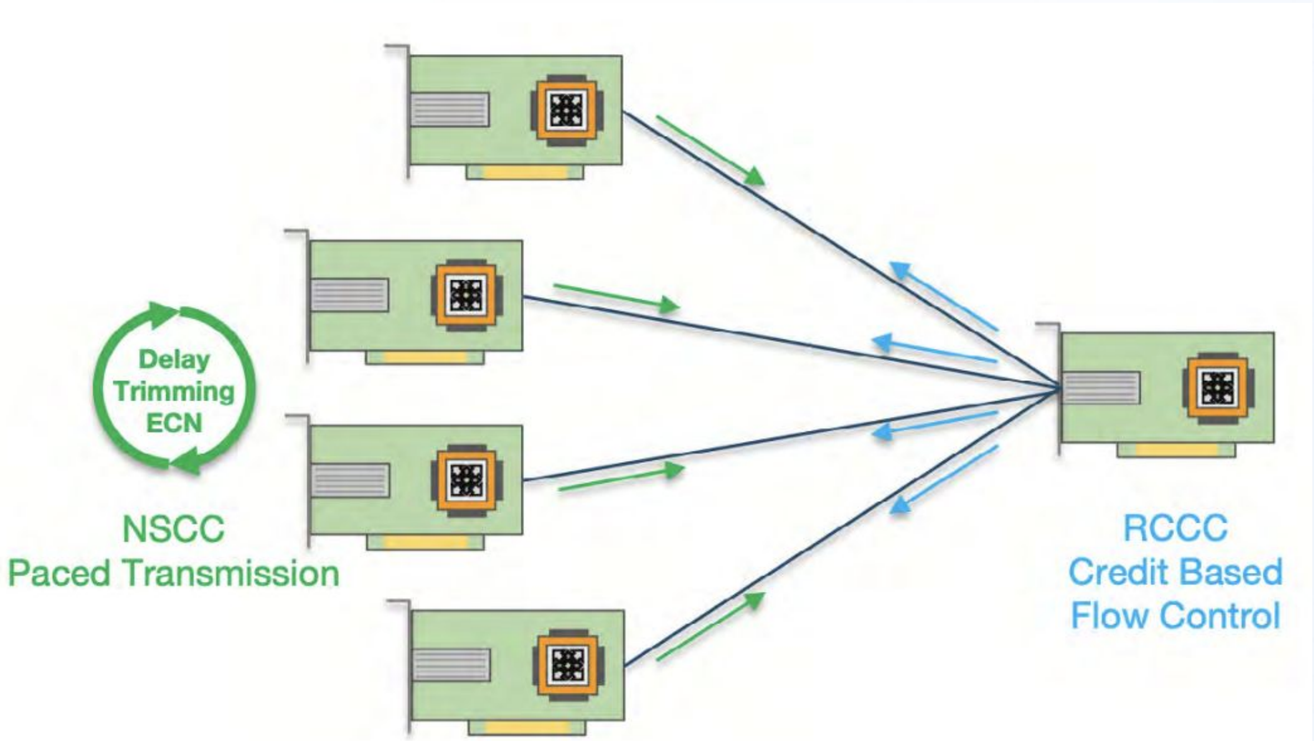


Figure 6: NSCC and RCCC Operation

- Network Signal Congestion Control (NSCC)は送信者ベースの輻輳回避手法で下記を利用し送信ペースを配分する
  - ネットワーク遅延
  - トリミングパケット
  - ECN通知
- Receiver Credit Congestion Control: RCCC)は受信者ベースの輻輳回避。各受信者がクレジットを作成し、全ての送信者に公平に割り当てる。これによりTCPインキャストの様な事を排除出来る。
- NSCCとRCCCは独立して動作する事が出来る

 Copyright © Arista 2026. All rights reserved.

Public.

MRCの特徴

- Entropy Value (EV)
  - QPフロー内のパケットの分散はEntropy Valueによって制御
    - IPv6アンダーレイの場合、UDP送信ポートとフローラベルにまたがった32ビット値として設定
- Out-of-order packet delivery
  - 各パケットにはRDMA仮想アドレスとリモートキーが格納
  - NICは到着順位に関係なく、システムメモリに書き込む可能
  - 受信側はSelective ACKを使用して、どのパケットが到着したかを通知
    - 従来のgo-back-Nモデルでは無く欠落してるパケットのみを再送信可能
- Path selection とmonitoring
  - NACKもしくはSACKを受信しパケット損失を検知した場合。そのパスは “inactive” に移行し除外
  - “inactive” パスに対してプローブを送信して、到達可能性をテスト
  - 確認応答を返した時は再びactiveになりPacket Sprayの候補になる

 Copyright © Arista 2026. All rights reserved.

Public.

ARISTA

MRCの特徴

- Packet trimming と congestion feedback
  - UECで定められているECNとパケットトリミングを利用
  - ECNは単なる輻輳通知では無く負荷分散の為に使用
  - 送信側はECNを受け取るとその特定のEVパスを転送経路として回避し非効率を解消
  - 受信側のラストホップではECNは無効。代わりに輻輳時はパケットトリミングがされ、ヘッダーのみを送信し、受信側はNACKメッセージで再送信を要求
    - これにより送信者は経路自体は機能してる事が確認出来る
- Path failure
  - パケットがトリミングされず、破棄されて受信されなかった場合はSACKとして送信元に通知
  - 送信元はパス障害と解釈し、badとマーク
  - その後プローブが送信され、パス全体で所定の閾値で正常に受信した場合はEVはActiveに戻る事が出来る

 Copyright © Arista 2026. All rights reserved.

Public.

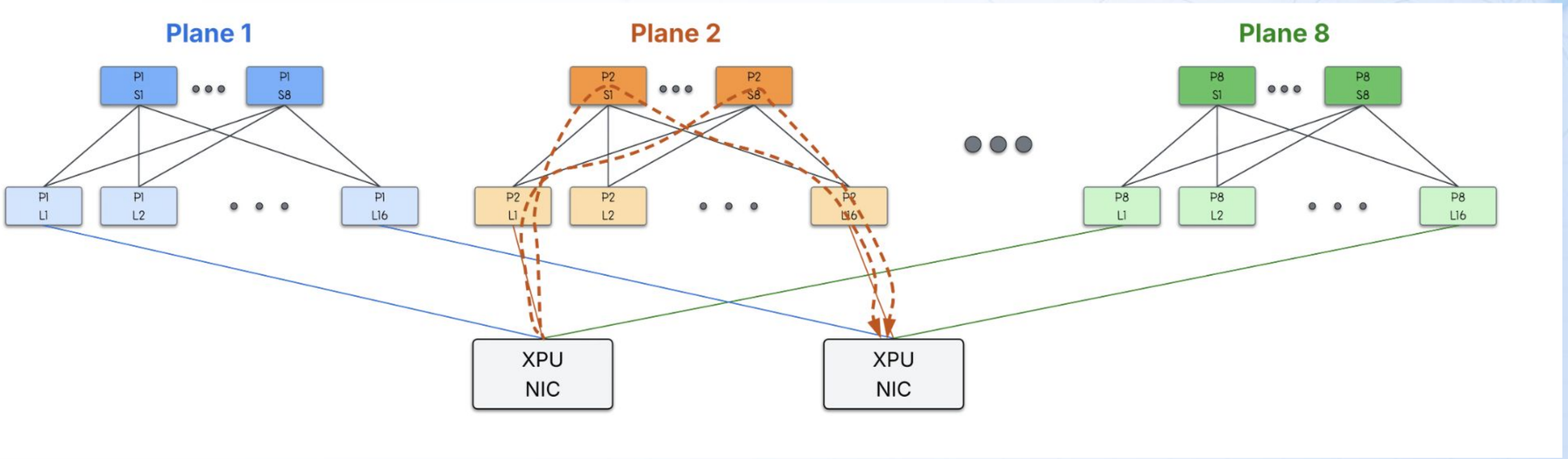
MRCの動作まとめ

シグナル	原因	送信側の反応 （Sender reaction）	EVの状態（EV state）
ECNマーク	特定パスの輻輳	再バランス処理：フローを一時的に他のEV値へ退避	アクティブのまま（ただし一回スキップ）
パケットトリミング ＆ NACK	インキャスト（一斉集中）による輻輳	トリミングされたパケットの再送	アクティブのまま（ただし一回スキップ）
パケット損失 （トリミングなし）	リンクまたはスイッチの障害	EVセットから対象のEV値を一時的に除外	非アクティブ（障害）：プローブによる導通確認成功で復旧
SACKに付与されたリンク状態	対向側のNICポートダウン	宛先の別NICポート／プレーンへEVセットを再マップ	プレーンレベルの障害

 Copyright © Arista 2026. All rights reserved.

Public.

SRv6



- MRCはECMP環境でも動作するが、SRv6によりトラフィックを複数のプレーンに分散させるだけでなく、各プレーン内のさまざまなパスを活用するために、トラフィックの直接ソースルーティングも可能に

 Copyright © Arista 2026. All rights reserved.

Public.

MRC deployment with SRv6

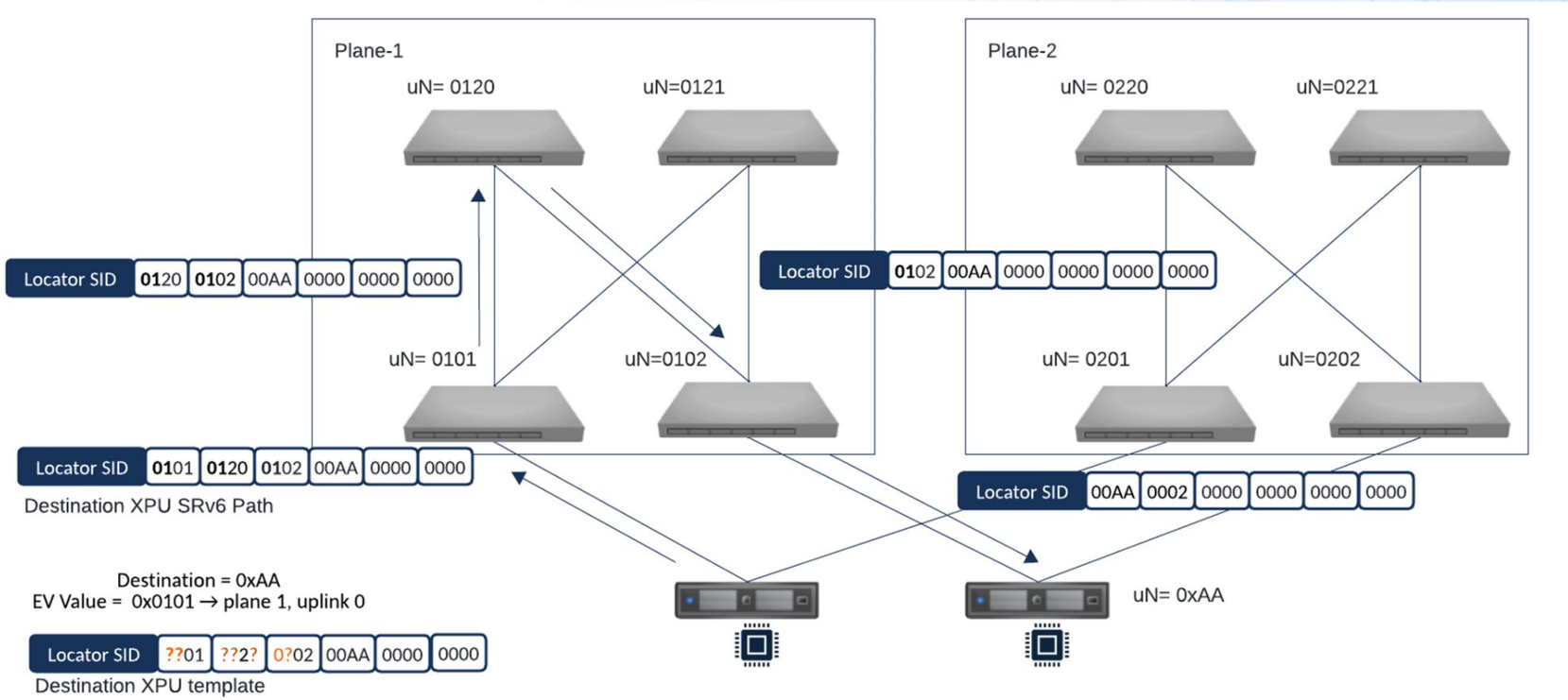


Figure 4: Diagram of the MRC multi-planar transport with SRV6

- ファブリック内のノードはuN SIDをサポートするSRv6トランスポートノードとして動作
- Open AIでは動的なルーティングを使わずに静的にエントリーを設定

 Copyright © Arista 2026. All rights reserved.

Public.

SRv6の転送処理

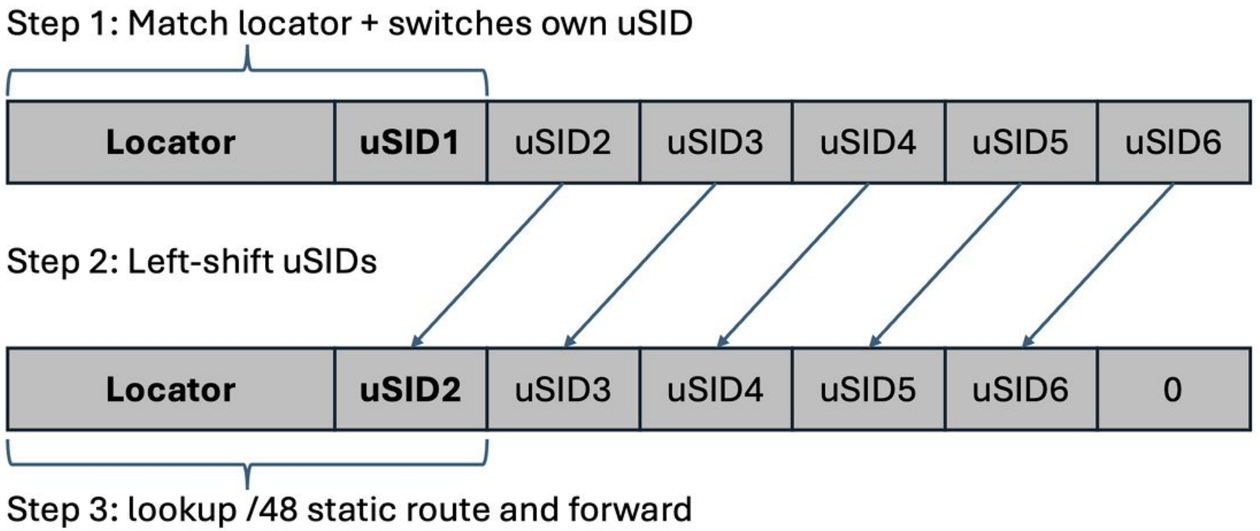


Figure 2: SRv6 forwarding using uN uSIDs

- 宛先IPv6アドレスの中に、通過する各スイッチを指定する16ビットの識別子（uSID）を順番に並べる
- Match: スイッチはパケットのアドレスを見て”自分宛てのSRv6パケットか”を確認。
- Left-Shift: アドレス内の経路情報を16ビット左にズラし、次に通過すべきスイッチのアドレスを先頭に繰り上げる

 Copyright © Arista 2026. All rights reserved.

Public.

SRv6アドレス生成の 3ステップ

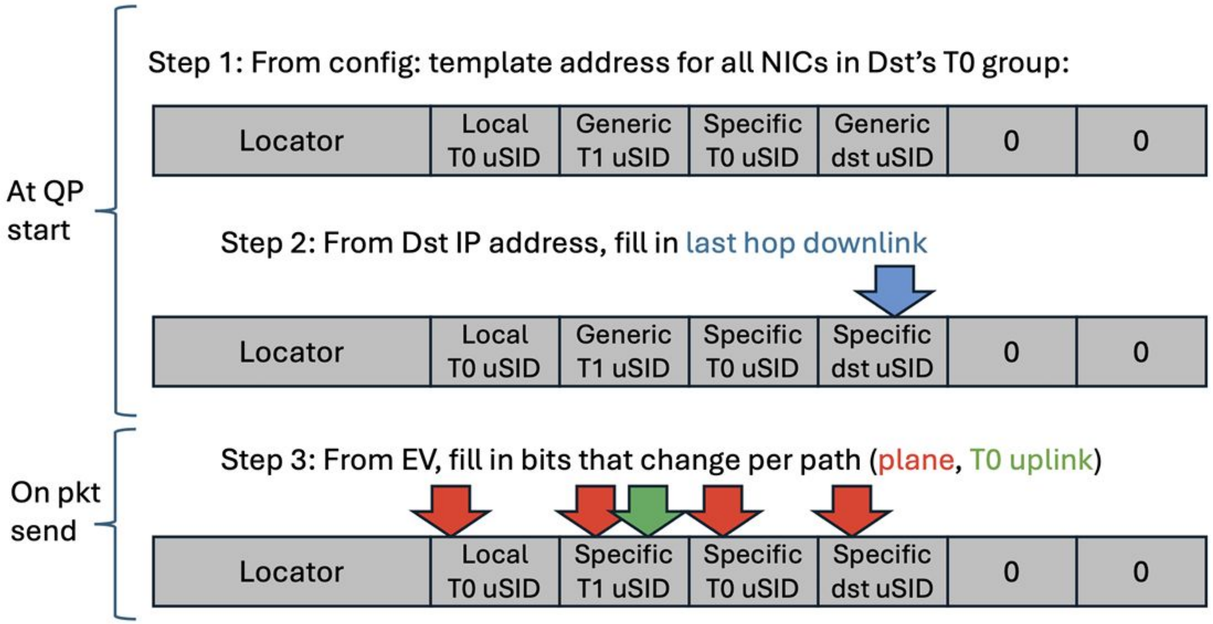


Figure 3: Creating the SRv6 address from an EV and template

- 宛先（Dst）のT0グループに属するすべてのNICで共通の”ベースとなるSRv6アドレスのテンプレート”を取得
- 宛先IPアドレスから、最終ホップとなるダウンリンク（port番号）の情報を埋め込み、宛先固有のテンプレートにする
- EVから動的なビットを流し込む

Plane番号

T0アップリンク番号

 Copyright © Arista 2026. All rights reserved.

Public.

まとめ

- 新しいRDMAベースのトランスポートプロトコルであるMRCは、多数のパスにトラフィックを分散させ、それらの間で負荷分散を行うことで、**フローの衝突という問題を解消**
- **マルチプレーン Closトポロジー**を採用
  - 高いスイッチ密度(Radix)と共にファブリック全体の冗長性を確保
  - 10万台をはるかに超えるGPUを備えたトレーニングクラスタを2層トポロジーで構築可能
  - 物理的な冗長性を高めることが可能
- SRv6を用いた静的ソースルーティングを採用
  - MRCが自ら障害箇所をバイパスする自由度を確保

 Copyright © Arista 2026. All rights reserved.

Public.

参考

- Three Genius Ideas for AI Fabrics
  - https://blogs.arista.com/blog/three-genius-ideas-for-ai-fabrics
- Multiplanar Scale Out fabrics with MRC Arista’s support of MRC for building an open multi-plane Scale Out Ethernet AI fabric
  - https://www.arista.com/assets/data/pdf/Whitepapers/MRC-Whitepaper.pdf
- Resilient AI Supercomputer Networking using MRC and SRv6
  - https://cdn.openai.com/pdf/resilient-ai-supercomputer-networking-using-mrc-and-srv6.pdf
- Multipath Reliable Connection (MRC) Specification
  - https://www.opencompute.org/documents/ocp-mrc-1-0-pdf

 Copyright © Arista 2026. All rights reserved.

Public.

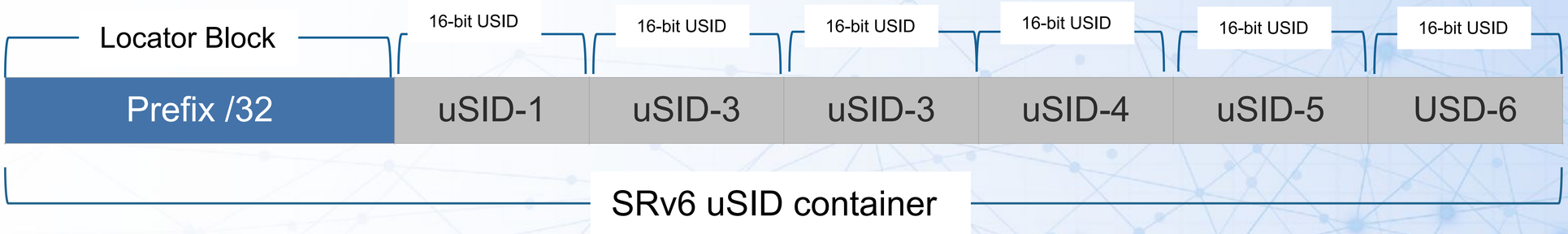
ARISTA

Thank You

www.arista.com

SRv6 - MicroSID (NEXT-C-SID Flavor)

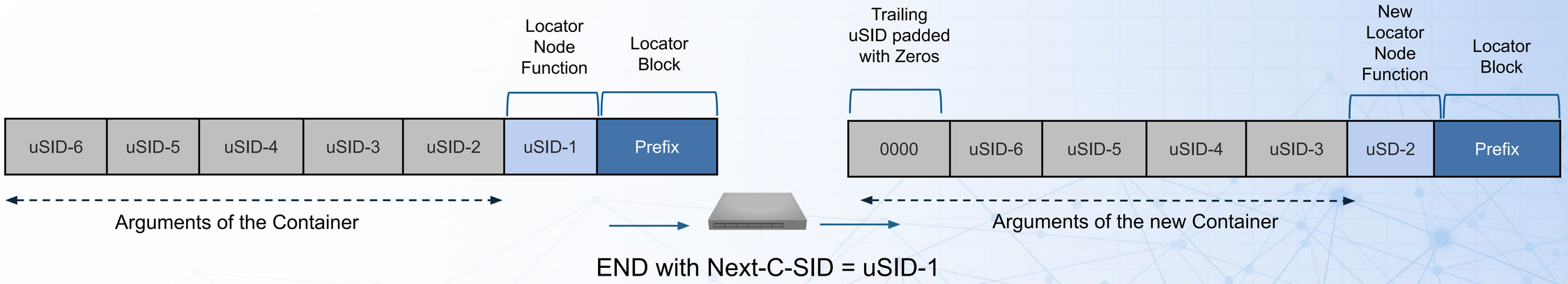
- uSIDでは1つのIPv6アドレスに複数のSIDを配置する事が可能
  - 1つのIPv6アドレスで最大6つのマイクロインストラクションが可能
  - 多くの標準的なSRv6のユースケースでSRHの挿入が不要に
  - 既存のハードウェアでSRv6を展開出来る可能性がある
- ドメイン内のSRv6 SIDはLocator Blockと呼ばれるIPv6プレフィックスから割り当てる
  - そのためドメイン内のSIDは全て同じLocator Block(上位ビット)から始まる
  - 共通のLocator Blockを持つ一連のSIDは単一のIPv6アドレス内に圧縮して収める事が出来る
  - 各SID毎のLocator Blockの重複を省略する事でこの圧縮を実現してる



 Copyright © Arista 2026. All rights reserved.
Public.

SRv6 - MicroSID

- Next-C-SID Flavorは完全な形式を持つ128ビットのSID
  - C-SIDコンテナ（uSIDキャリヤとも呼ばれる）の転送処理に使われる
  - このIPv6アドレスはLocator-Blockの後に一連のC-SID(または命令処理)が続く構造を持つ
  - Locator-NodeとFunctionがコンテナ内の最初のC-SIDとなりそのArgument/引数部分に以降の一連のC-SIDが配置される
    - 2番目のC-SIDはArgumentの最上位ビットに配置され、3番目のC-SIDは2番目のC-SIDの直後のビットに続く（以降も同様）
  - NEXT-C-SID Flavor が指定され、自身がアドバタイズしているLocator Node（END）と一致するパケットを受信したノードは以下を行う:
    - 自身のLocator NodeのC-SIDを削除し、後続のC-SID群を左へシフトして空いた末尾のuSID領域をpadding(0埋め)する
    - ノードは、先頭に繰り上がった新しいC-SIDに基づいてパケットを転送する

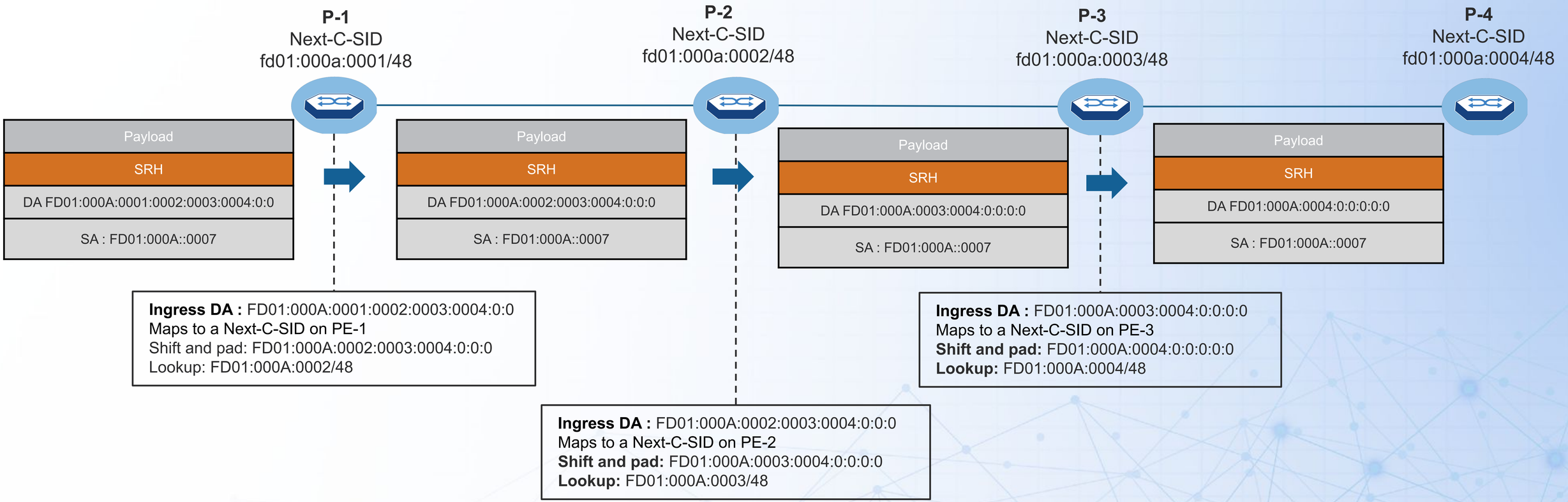


 Copyright © Arista 2026. All rights reserved.

Public.

SRv6 – MircoSID (NEXT-C-SID Flavor)

- NEXT-C-SID operation
  - Locator Block fd01:000A/32

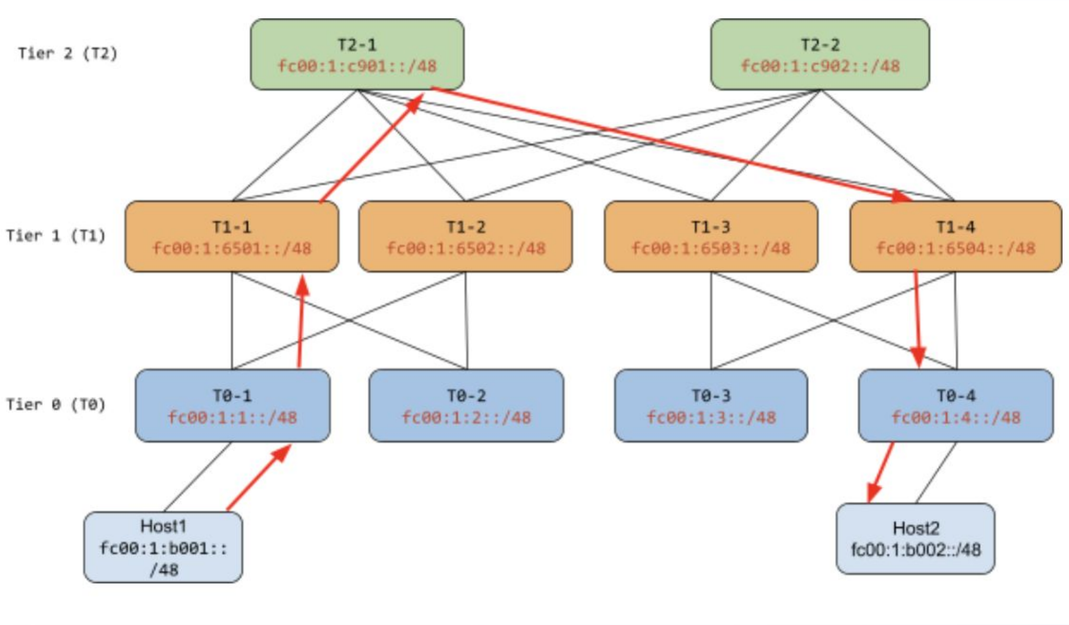


 Copyright © Arista 2026. All rights reserved.

Public.

SRv6 uN Support

https://www.arista.com/en/support/toi/eos-4-34-2f/22062-srv6-un-support




```
router srv6
   vrf default
      local address 2001:db8:100::1
      !
      micro-segment domain DOM1
         block fc00:42::/32
      !
      locator LOC1
         prefix micro-segment domain DOM1 end usid 3

interface Loopback0
   ipv6 address fc00:42:3::/48
```

- F3216(uSID block length of 32 uSID length of 16)
- **サポートプラットフォームは**
  7060X6/7280R3/7500R3/7800R3/7800R4/7700R4/7280R4/7282R4

 Copyright © Arista 2026. All rights reserved.

Public.

ARISTA

Thank You

www.arista.com