The bridge to possible

Demystify NCS5500/5700 Resources

Effective Network Design and Operations

Deepak Balasubramanian, Technical Leader
Technical Marketing Engineering
@DeepakBalas5

Cisco Webex App

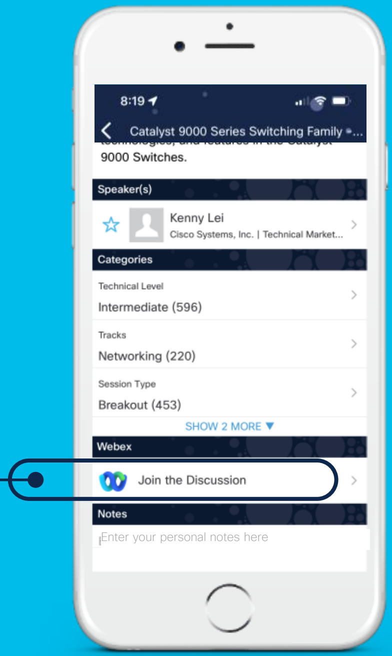
Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated
until February 24, 2023.



8:19

Catalyst 9000 Series Switching Family •...

9000 Switches.

Speaker(s)

Kenny Lei
Cisco Systems, Inc. | Technical Market...

Categories

Technical Level
Intermediate (596)

Tracks
Networking (220)

Session Type
Breakout (453)

SHOW 2 MORE ▼

Webex

Join the Discussion

Notes

Enter your personal notes here

Agenda

- NCS5500/NCS5700 NPUs Overview
- Double click into on-chip resources of NCS
- Breaking Forwarding Information Base(FIB)
- Resource consumption by applications
- Addressing Common Network design resource issues
- Key Take Aways

NCS5500/NCS5700 NPUs (Network Processing Units)

NCS 5500/5700 – Fixed Portfolio

High Scale Aggregation evolution

NCS5500 Products (Q-MX, J, J+)

1G | 10G | 25G

NCS 5501/SE ●●

40G |100G

NCS 5502/SE ●●

25G | 40G | 100G

NCS-55A1-48Q6H ●●●●

NCS-55A1-24Q6H-S/SS ●●●●

40G | 100G

NCS 55A1-36H-S/SE ●●●●

NCS 55A1-24H ●●●

10G | 25G | 100G

NCS 55A2-MOD-S/SE ●●●●

NCS5700 Products (J2/J2C/Q2C/J2C+)

NCS-57B1-6D24 ●●●●●

- 400G ZR/ZR+
- 1RU; 4.8 Tbps throughput
- 24x100G + 6x400G
- MACSEC, Timing

NCS-57B1-5DSE ●●●●●

- 400G ZR/ZR+
- 1RU; 4.4 Tbps throughput
- 24x100G + 5x400G
- MACSEC, Timing
- External TCAM

NCS-57D2-18DD-S ●●●●●

- 400G ZR/ZR+
- 2RU; 7.2 Tbps throughput
- Flexible 66 ports 2x400G + 16x400G/64x100G
- MACSEC*, IPSEC*, Timing

NCS-57C3-MOD-S ●●●●●

- 400G ZR/ZR+
- 3RU; 2.4T throughput
- Fixed: 48x1/10/25G + 8x100G QSFP28
- 3 x MPA: 2x800G + 1x 400G
- MACSEC, Timing

NCS-57C3-MOD-SE ●●●●●

- 400G ZR/ZR+
- 3RU; 2.4T throughput
- Fixed: 48x1/10/25G + 4x100G QSFP28
- 3 x MPA: 2x800G + 1x 400G
- MACSEC, Timing
- External TCAM

NCS-57C1-48Q6D-S ●●●●●

- 400G ZR/ZR+
- 1RU; 2.4T throughput
- 32x1/10/25G + 16x1/10/25/50G + 6x400G
- MACSEC, Timing

● Segment Routing ● EVPN ● MACSec ● Timing ● 400G ZR/ZRP

NCS 5500/5700 – Modular Portfolio

High Scale Aggregation evolution

NCS5500 Products (J, J+)

40G |100G

NC55-24H12F-SE 🔵🟡

NC55-18H18F 🔵🟡

100G

NC55-36X100G-S 🔵🟡

NC55-24X100G-SE 🔵🟡

NC55-6x200-DWDM-S 🔵🟡

40G | 100G

NC55-36X100G-A-SE 🔵🟡🔴🟣

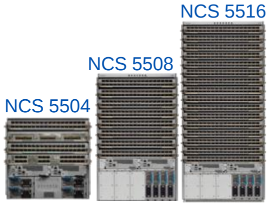
Modular

NC55-MOD-A-S/SE 🔵🟡🔴🟣

NC55-MOD-A-SE

10G | 25G | 100G

NC55-32T16Q4H-A 🔵🟡🔴🟣



- 🔵 Segment Routing
- 🟡 EVPN
- 🔴 MACSec
- 🟣 Timing
- 🟢 400G ZR/ZRP

NCS5700 Products (J2)

NC57-24DD 🔵🟡🟢

- 400G ZR/ZR+
- 24x400G,
- Through put 9.6 Tbps
- No eTCAM

NC57-18DD-SE 🔵🟡🟢

- 400G ZR/ZR+
- 18x400G, 30x200G/100G
- Through put 7.2 Tbps
- External TCAM

NC57-36H6D-S 🔵🟡🔴🟣🟢

- 400G ZR/ZR+
- 100G, 400G
- Throughput 4.8 Tbps
- Timing, MACSEC,

NC57-36H-SE 🔵🟡🔴

- 400G ZR/ZR+ (1x100G mode)
- 100G
- Throughput 3.6 Tbps
- External TCAM

NC57-MOD-S 🔵🟡🔴🟣🟢

- 400G ZR/ZR+
- 10G, 25G, 50G, 100G, 400G
- Throughput 4.8 Tbps
- Timing, MACSEC, 800G-MPA

NCS5500/5700 – NPU Evolution

	Jericho	Jericho +	Jericho2	Jericho2C	Jericho2C+
Bandwidth	720G	900G	4.8T	2.4T	7.2T
Power/100G	16.6W	16.6W	7.3W	5-6.7W	6.3W
Performance (pps)	720M	835M	2B	1B	2.83B
OCB	16MB	16MB	32MB	32MB	32MB
Buffer	4GB (GDDR)	4GB (GDDR)	8GB (HBM)	4GB (HBM)	8GB (HBM)
VOQ	96K	96K	64K per core	128K per core	256K per core
Counters	256K	256K	384K	192K	384K
Network IF	24x 25G+36x 12.5G	48x25G+24x12.5G	96x 50G	32x50G+96x25G	144x 50G
Fabric IF	36x 25G	48x 50G	112x 50G	48x 50G	192x 50G
MC Groups	-	128K	256K	256K	256K
Timing / Encryption	Class B / No	Class B / No	Class B / No	Class C / No	Class C / Yes

Double click on-chip resources of NCS

NCS On-chip Critical Databases

Database	Stores
LPM(Longest Prefix Match)	IPv4/IPv6, Multicast prefixes
LEM(Large Exact Match)	MPLS Labels, MAC, host routes*
ITCAM (Internal TCAM)	QOS, ACL, LPTS, LI , Stats
ETCAM (External TCAM) - Scale variants(SE) only	Prefixes (unicast, multicast) Service labels, ACLs, Flow-spec rules, stats, LPTS
FEC (Forwarding Equivalence class)	Next hop info - ENCAP pointers, VOQ ids
ECMP FEC	Multipath Next hop info – Pointer to FEC array
EEDB (Egress Encap Database)	MPLS out-labels, SRv6 remote SIDs, GRE, ARP/ND

* Host routes also go to LPM in J2 base

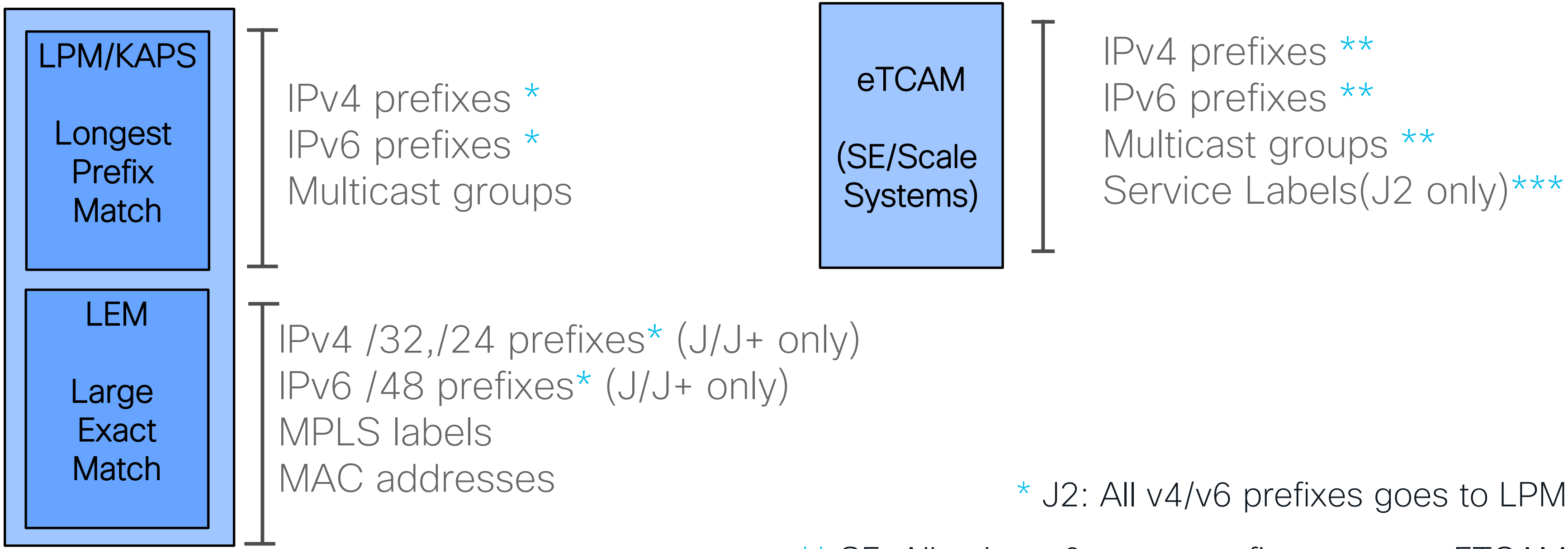
NCS On-chip Termination & Lookup DBs

Database	Stores
IOEM, EOEM(Ingress/Egress OAM Exact Match)	OAM classification (BFD/CFM)
VSI(Virtual Switch Interface)	L3 Interfaces, Bridge Domains
RIF (Routed Interface)/Routable VSI	L3 Interfaces
ISEM, ESEM (Ingress/Egress Small exact Match)	Ingress Tunnel Termination, Egress vlan translation, returns OUTLIF
In-LIF, Out-LIF (Ingress/Egress Logical interface table)	Logical Interface table (L2/L3 sub-interfaces)
GLEM(Global LIF Exact Match)	Global ID match

Hardware Resource monitoring CLIs

Current CLI options show controllers npu resources <-/->	New CLI options show controllers npu resources <-/->
LEM/LPM	encapAC
Externaltcam (v4/v6)	encapARP
FEC	encapPWE
ECMP_FEC	encaptunnels
Encap	LIF, RIF
Stats	VoQ, SEM

Database Overview – LEM/LPM



IPv4 prefixes *
IPv6 prefixes *
Multicast groups

IPv4 /32,/24 prefixes* (J/J+ only)
IPv6 /48 prefixes* (J/J+ only)
MPLS labels
MAC addresses

IPv4 prefixes **
IPv6 prefixes **
Multicast groups **
Service Labels(J2 only)***

* J2: All v4/v6 prefixes goes to LPM

** SE: All unicast & mcast prefixes goes to ETCAM

*** J2-SE: Label in ETCAM for specific scenarios

Database Overview – FEC & EEDB

FEC (Forwarding Equivalence Class)

- Ingress Object points to Egress Objects(Encap) + Egress Queue(Voq)
- FEC can be hierarchical pointing to another FEC
- ECMP FEC : Used for multipath pointing to regular FEC array
- Protected FEC : Comes in Primary/Backup pair for FRR

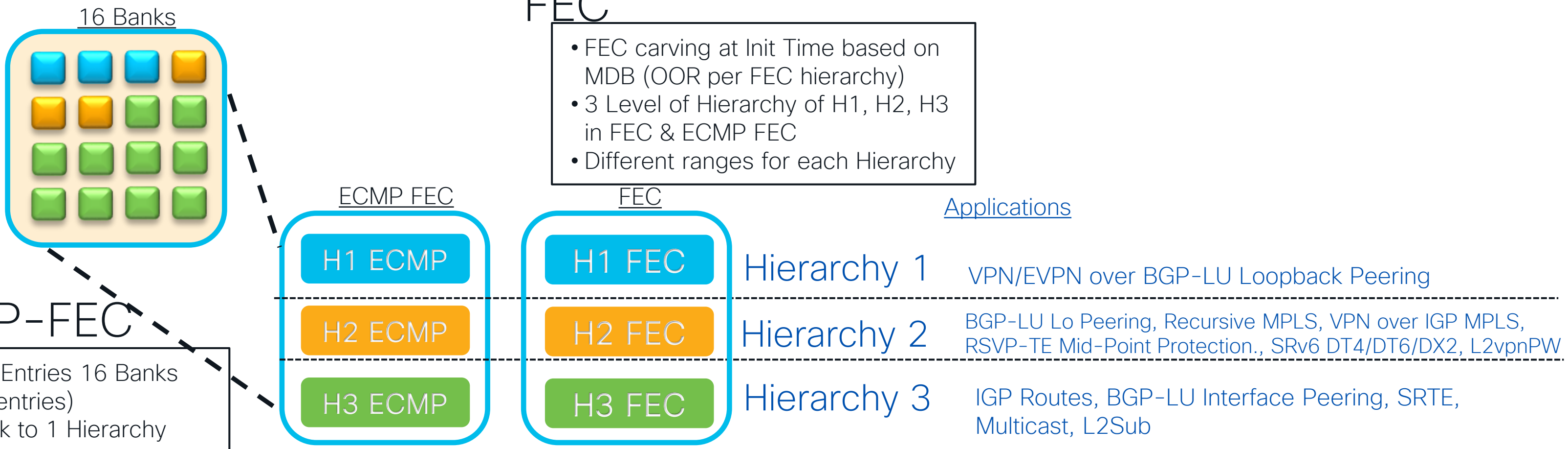
EEDB (Egress Encapsulation Database)

- Direct Index table stores encapsulation info (MPLS labels, ARP/ND, GRE, SRv6)
- Lookup happens in egress pipeline to access the data and create packet headers

Database Overview – ECMP_FEC

- ECMP_FEC a sparse resource in NCS platforms (4K on J/J+, 32K on J2)
- ECMP_FECs allocated for
  - IPv4/IPv6 multipath – ecmp_fec reused b/w prefixes for same NH set
  - Labelled multipaths – Abuser ☺ as unique ecmp_fec for every label
  - Indirection ECMP FEC for BGP PIC Edge
  - Binding SID for SR Policies
- Optimizations with SR, BGP-SR, FEC sharing (will see more details)

NCS5700 FEC Hierarchy



16 Banks
ECMP FEC
FEC
H1 ECMP
H2 ECMP
H3 ECMP
H1 FEC
H2 FEC
H3 FEC
Hierarchy 1
Hierarchy 2
Hierarchy 3
Applications
VPN/EVPN over BGP-LU Loopback Peering
BGP-LU Lo Peering, Recursive MPLS, VPN over IGP MPLS, RSVP-TE Mid-Point Protection., SRv6 DT4/DT6/DX2, L2vpnPW
IGP Routes, BGP-LU Interface Peering, SRTE, Multicast, L2Sub

FEC

- FEC carving at Init Time based on MDB (OOR per FEC hierarchy)
- 3 Level of Hierarchy of H1, H2, H3 in FEC & ECMP FEC
- Different ranges for each Hierarchy

ECMP-FEC

- Total 32K Entries 16 Banks (each 2K entries)
- Every Bank to 1 Hierarchy only
- OOR NOT per hierarchy

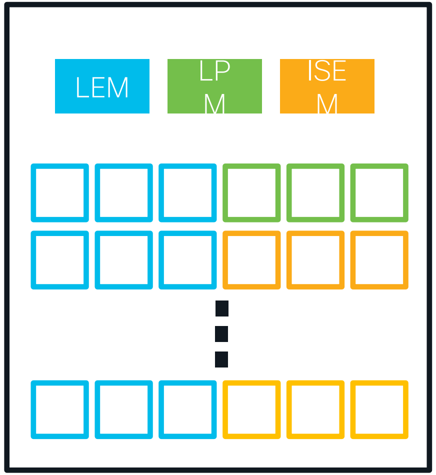
NCS5700(J2)	NCS5500(J/J+)
3 Level FEC hierarchy	2 Level (No dedicated carving)
~3x J/J+ scale (based on MDB)	~126K FEC
32K ECMP FEC	4K ECMP FEC

NCS5700 Encap phase: Application Map

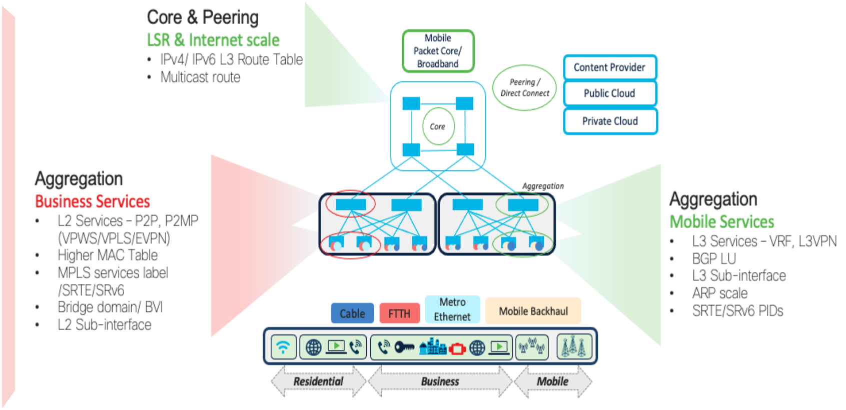
Cluster Bank Pair	Physical Phase (size)	Logical Phases (from L2MAX MDB)	Applications
EEDB_S2_XL	S2	Logical_phase : 1 (Encap_Rif)	RIF, SRv6 H.Encap
	XL	Logical_phase : 6 (Tunnel4)	IGP, BGP-LU interface peering, SRTE(1,2 label), SRv6 T.Insert
EEDB_L2_M3	L2	Logical_phase : 2 (Encap_NativeArp)	VPN/EVPN/SRTE(9,10 label), SRv6 T.Insert, PWE
	M3	Logical_phase : 8 (Encap_Ac)	AC Outlif, Non-VRF ND
EEDB_M1_M2	M1	Logical_phase : 3 (Encap_NativeAc or Tunnel1)	BGP-LU Loopback peering, SRTE(7,8 label), SRv6 T.Insert
	M2	Logical_phase : 5 (Tunnel3)	SRTE(3,4 label), SRv6 T.Insert
EEDB_S1_L1	S1	Logical_phase : 4 (Encap_Tunnel2)	SRTE(5,6 label), SRv6 T.Insert
	L1	Logical_phase : 7 (Encap_Arp)	ARP, ND, SRv6 H.Encap

NCS5700(J2): Flexible & Bank Pair to phase map based on MDB profile. ~3 x J/J+ scale
NCS5500(J/J+) : No phase map. 96K/112K full entries (28 banks). Any application can consume most banks (except a few reserved banks 0,1 for RIF & 2-5 for accounting encaps)

NCS5700 MDB Profile resource carving



L2 Profile



Core & Peering

LSR & Internet scale

- IPv4/ IPv6 L3 Route Table
- Multicast route

Aggregation

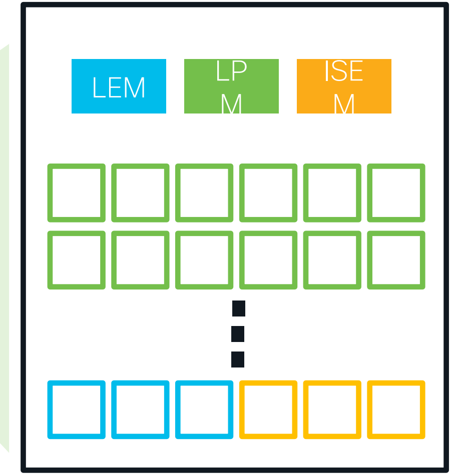
Business Services

- L2 Services – P2P, P2MP (VPWS/VPLS/EVPN)
- Higher MAC Table
- MPLS services label /SRTE/SRv6
- Bridge domain/ BVI
- L2 Sub-interface

Aggregation

Mobile Services

- L3 Services – VRF, L3VPN
- BGP LU
- L3 Sub-interface
- ARP scale
- SRTE/SRv6 PIDs



L3 Profile

4 MDB Profiles on NCS5700 (J2 Family)

- L3MAX-SE, L2MAX-SE : Scale variants
- L3MAX, L2MAX : Base variants

No MDB profiles on NCS5500(J/J+)

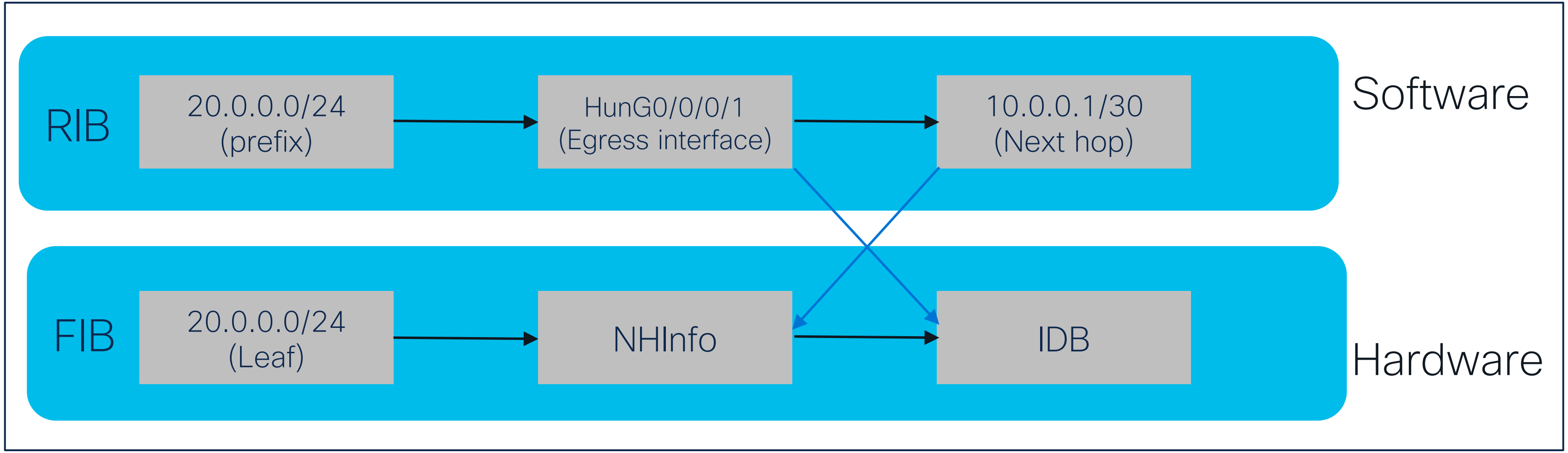
Breaking Forwarding Information Base(FIB)

CEF building blocks

- CEF built with 2 Key components
  1. Forwarding Information Base has Prefixes with corresponding next-hop information (Nhinfo) & egress interface (IDB), MPLS labels
  2. Adjacency table holds L2 adjacency (ARP/ND) for the next-hops
- FIB is the center-piece in forwarding infrastructure
- CEF process (fib_mgr) maintains the FIB databases in
  - Control plane (Route processors) which is PI (Platform independent)
  - Data plane (Line cards) which has both PI & PD (Platform dependent) representations

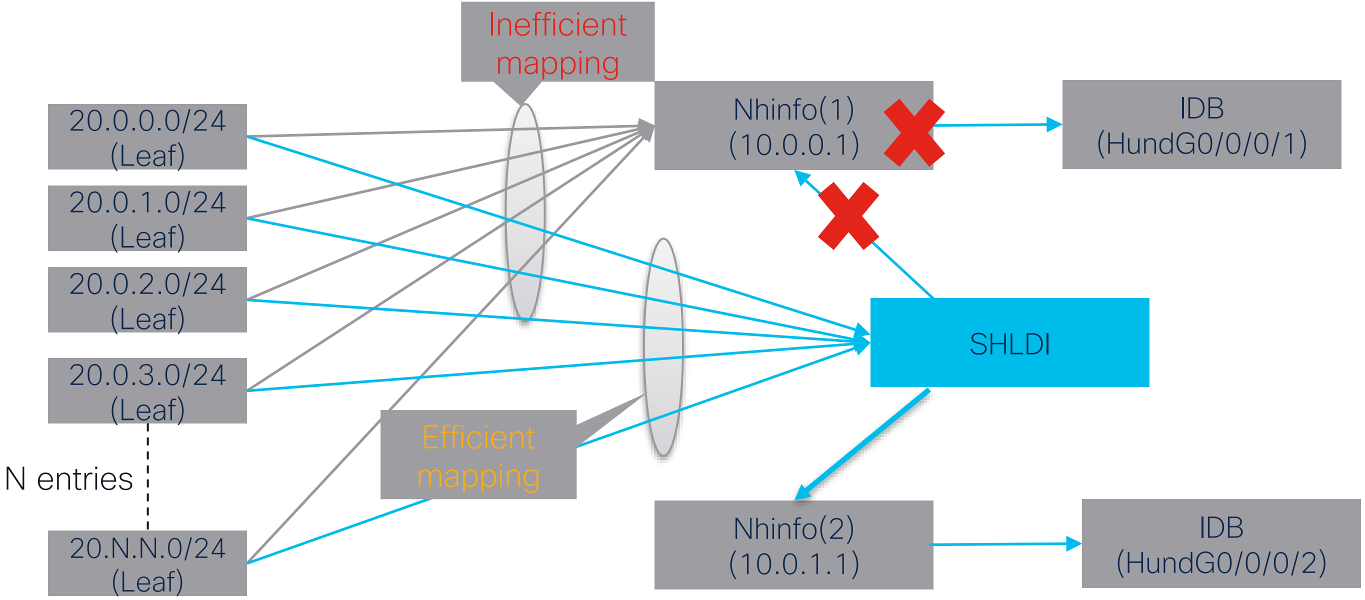
RIB to FIB mapping

- FIB derived from routing table (RIB) maintains the mirror image



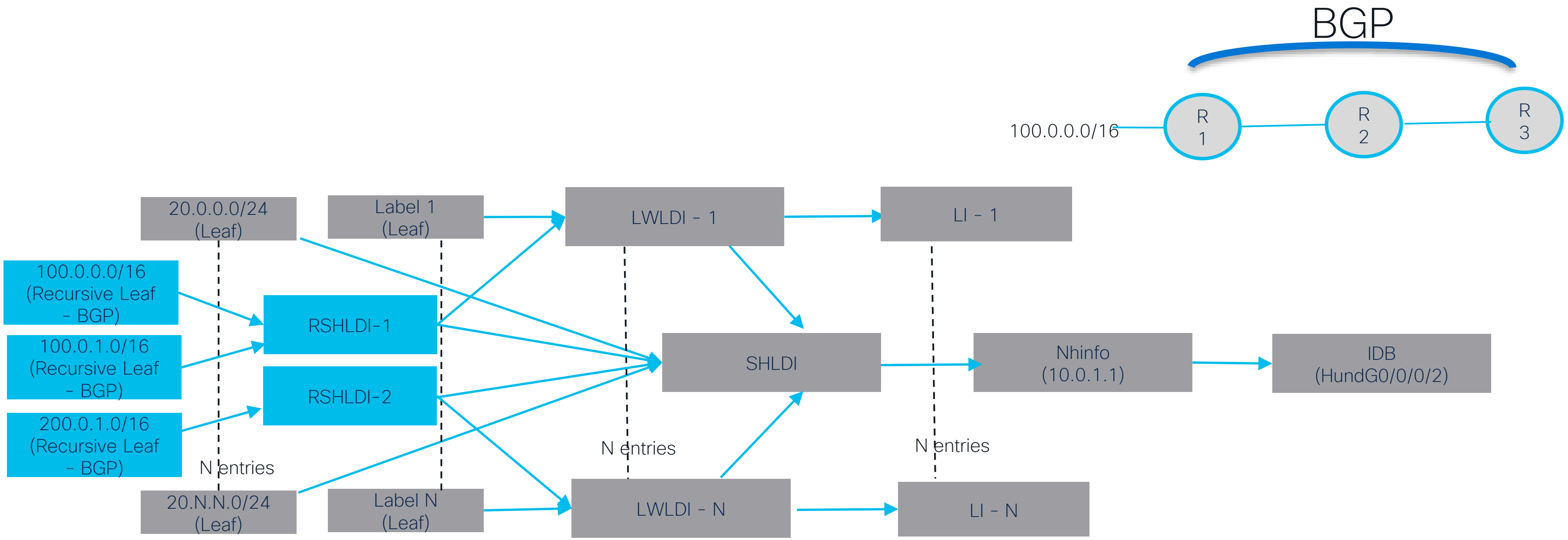
- Each family of platforms (NCS5500/5700/500 or ASR9k or CRS) have its own PD FIB implementations

FIB Objects – Shared Load Info(SHLDI)



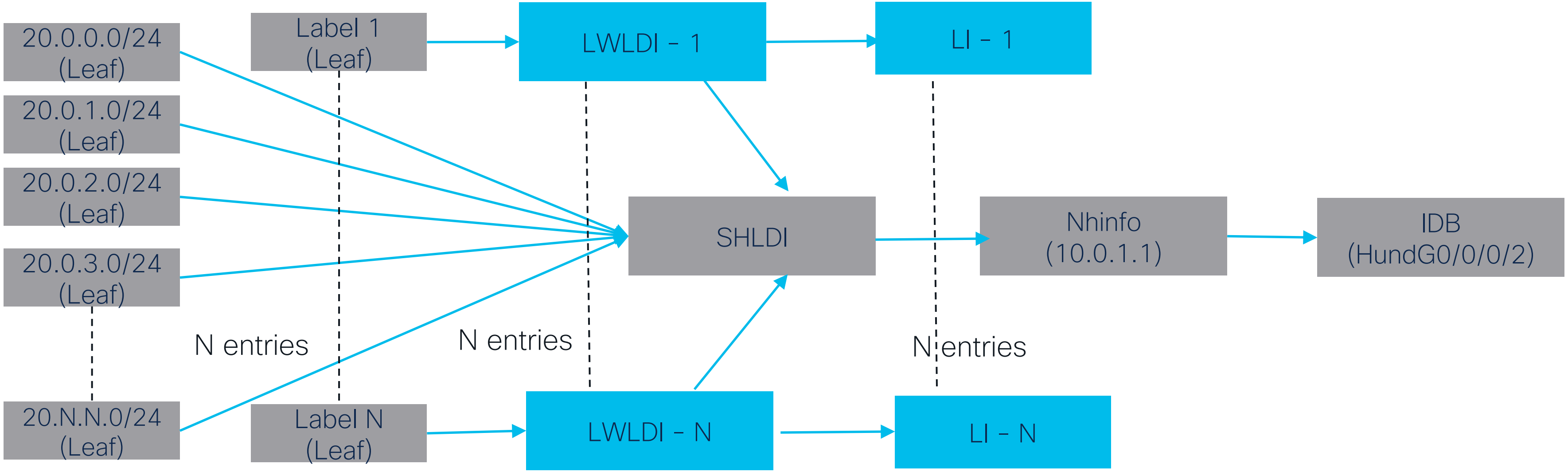
- SHLDI is an indirection object (no reprograming leafs for Nhinfo change)
- Single SHLDI update over N(1000’s) of Leaf updates, Better convergence

FIB Object – Recursive Shared Load Info(RSHLDI)



- RSHLDI used for service routes (BGP) points to SHLDI or LWLDI
- For example, R3 learns about 100.0.0.0/16 with next hop as R1, but R1 is not directly connected but learnt via IGP
- Many RSHLDIs point to same SHLDI if IGP next-hop same
- Many Recursive Leaves point to same RSHLDI if BGP next-hop same

FIB Objects – Light weight Load Info(LWLDI)

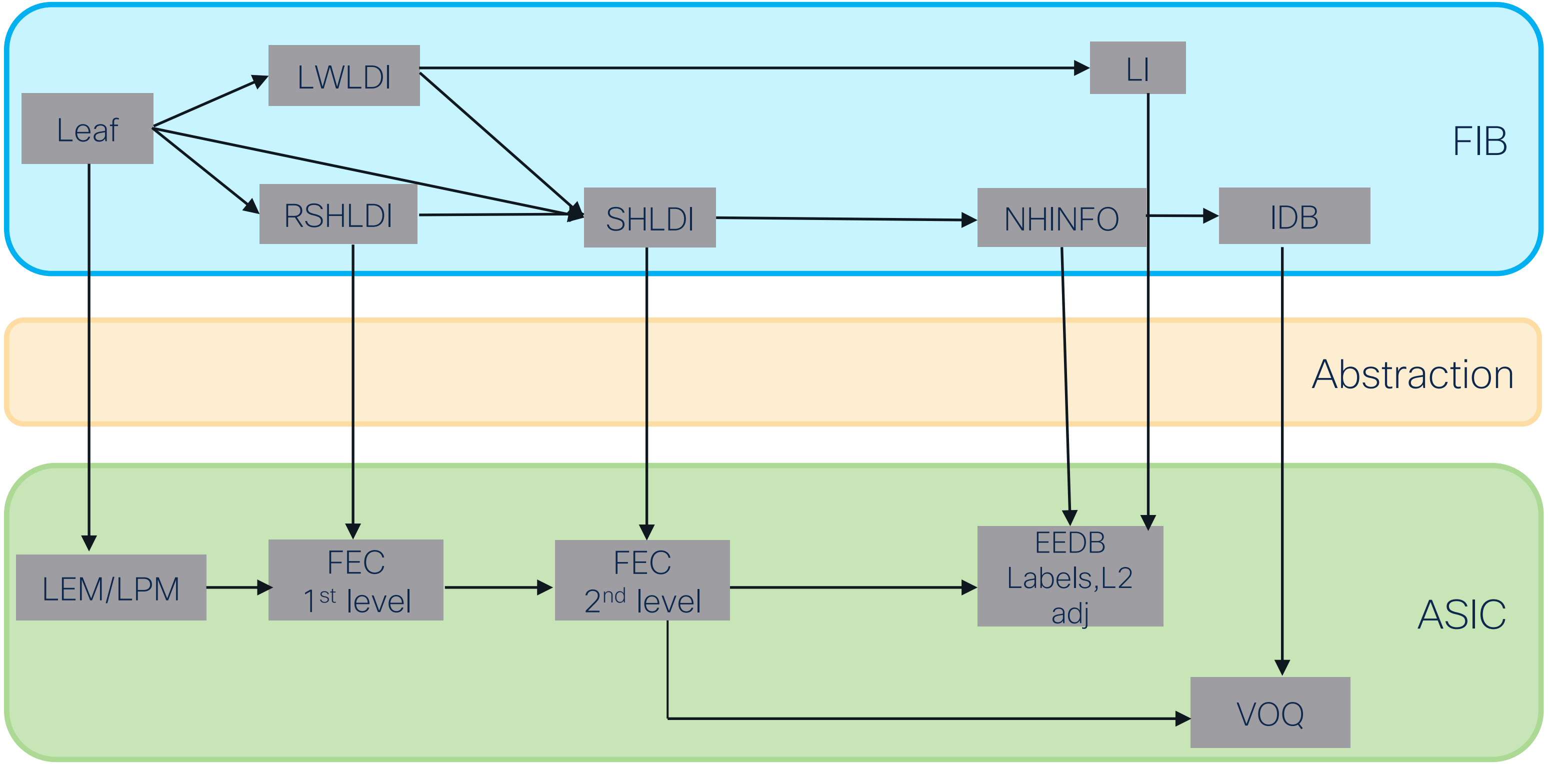


- LWLDI is Leaf extension (unique per leaf) which is a pointer towards LI (Label information)
- For “N” Leaves it will be “N” LWLDI as the out-labels are unique

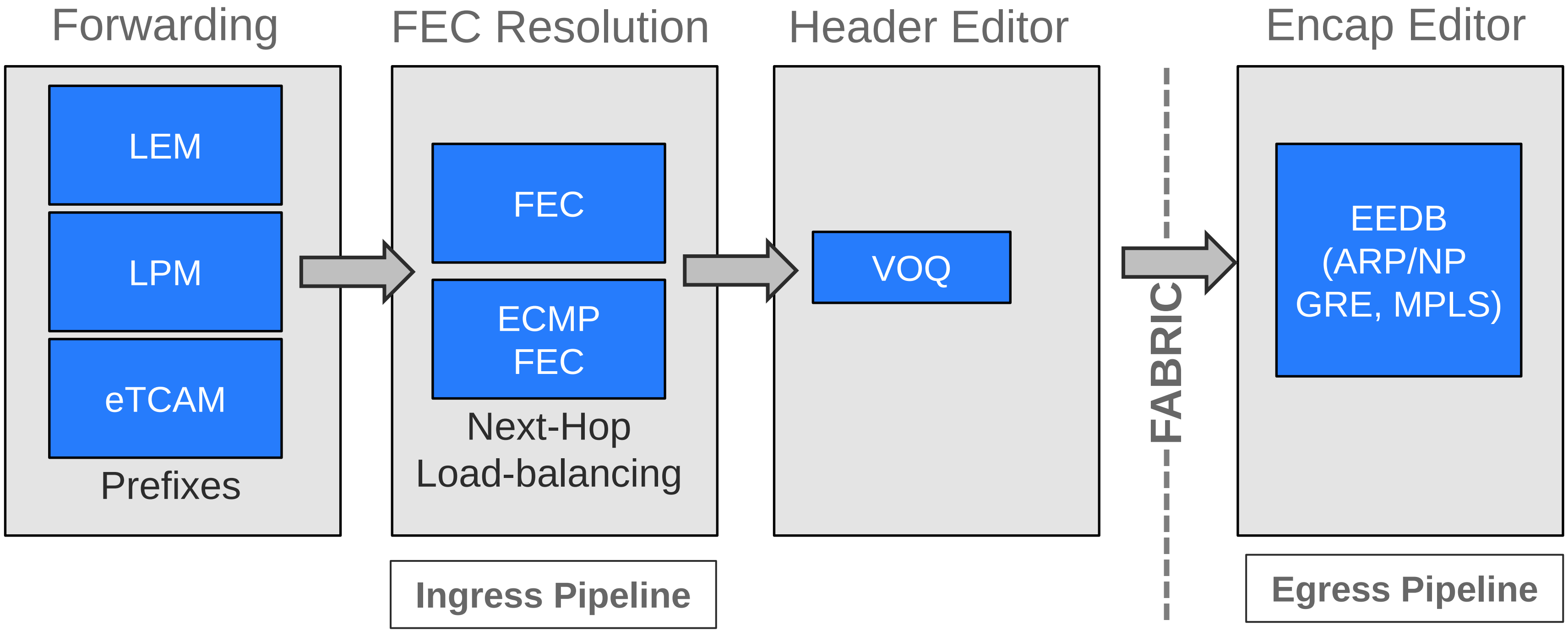
FIB object mapping to NCS PD (Hardware)

FIB Objects	Stores	HW Objects
Leaf	Destination IP Local Label/SID	LPM/LEM/ETCAM
LWLDI	Paths having MPLS Label binding	FEC
LDI (SHLDI, RSHLDI, COLL LDI)	Paths having no MPLS Label binding	FEC
LI	Out/Remote Label	EEDB (MPLS Encap)
SR6I	SRv6 Remote SID	EEDB (SRv6 Encap)
NHINFO (Type - TX)	Outgoing If & DMAC	EEDB (Link Layer Encap)
NHINFO (Type - TE)	TE Tunnel Path	FEC

NCS PD Object mapping



NCS CEF Implementation “Summary”



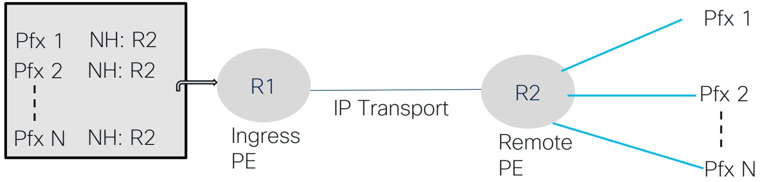
Resource consumption by applications

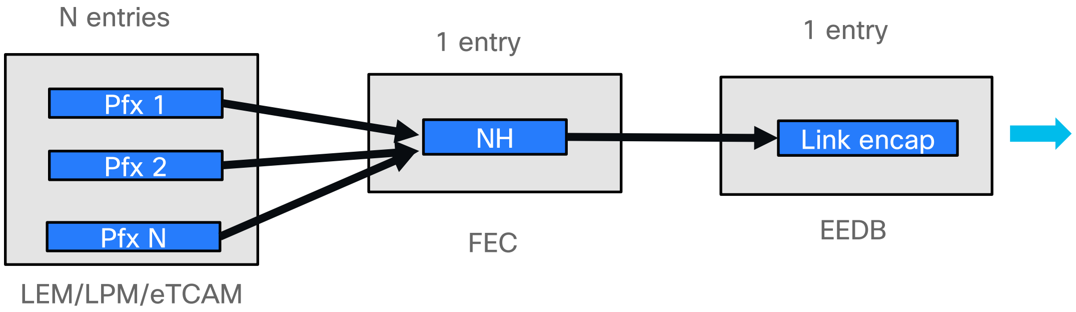
Numbers that matters

Database	NCS5500 Max entries	NCS5700 Max entries (*Depends on MDB)
LEM	768K	1.2M*
FEC	124K	340K+*
ECMP FEC	4K	32K
EEDB	96K (J)/112K(J+)	224K+*
LPM	350K 1.5M (Large LPM)	2.6M (Base)*
eTCAM	4M (SE version)	5M+ (SE version)*

Resource utilization – IP Forwarding (Uni-path)

“N” IP prefixes Uni-path via single Next-hop

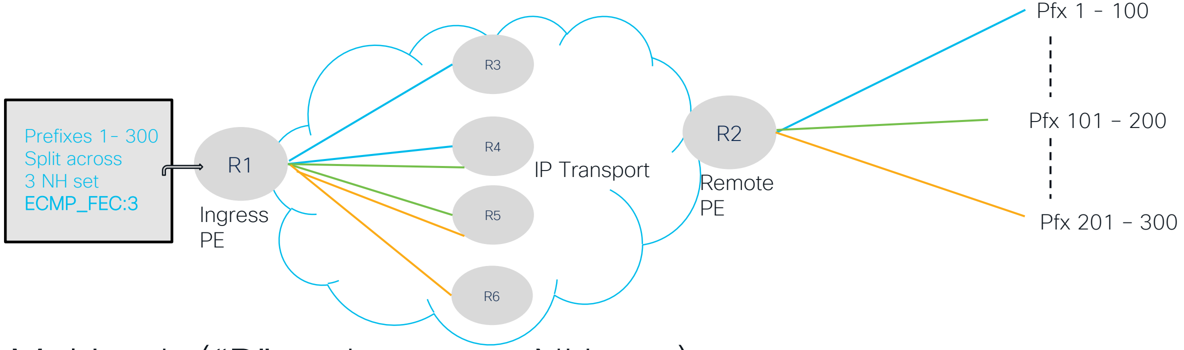




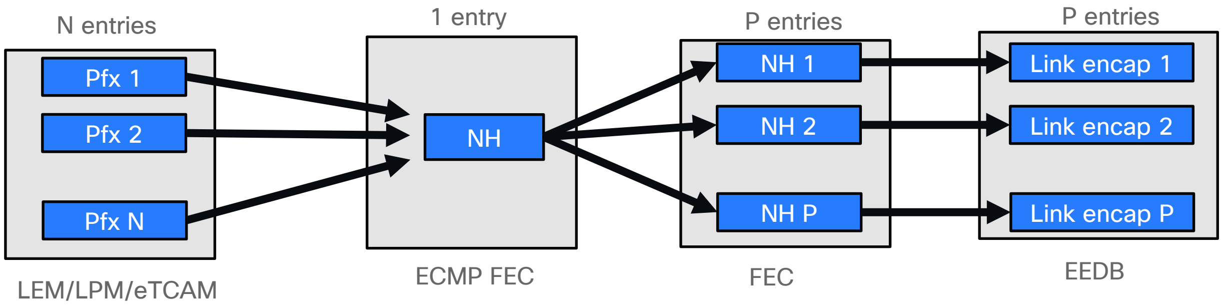
LPM	ECMP FEC	FEC	EEDB
N	0	1	1

Resource utilization – IP Forwarding (Multipath)

“300” IP prefixes Multipath via “3” different NH set



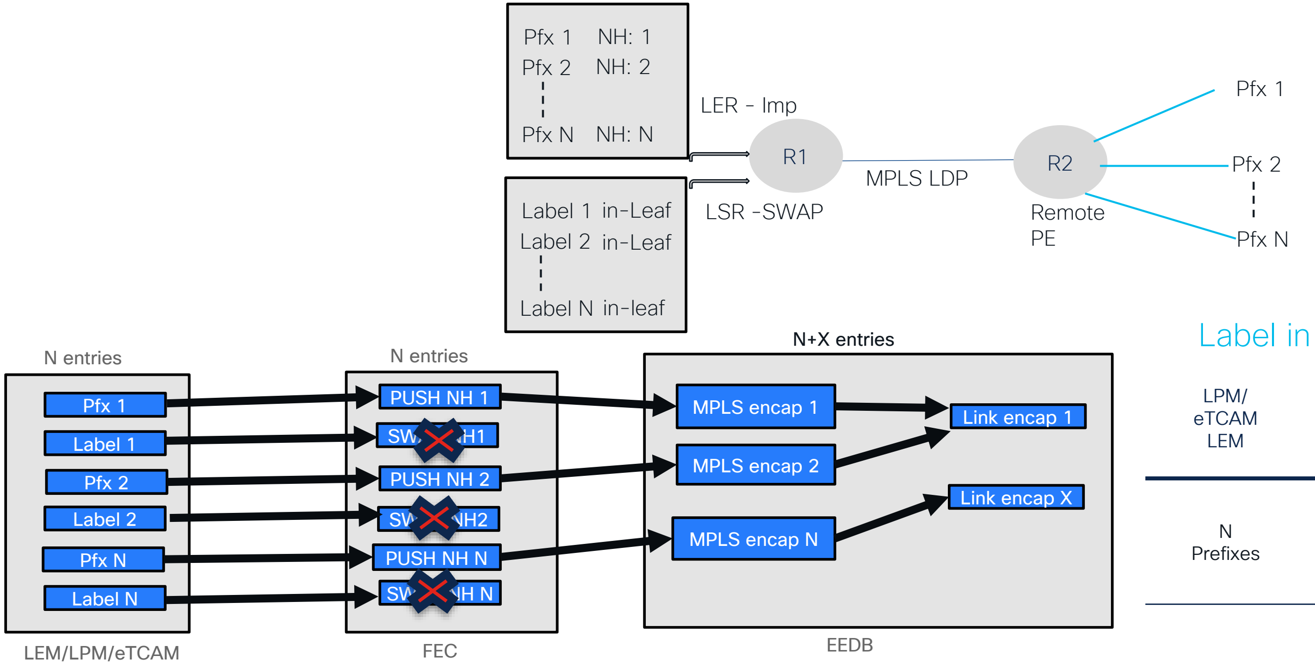
“N” IP prefixes Multipath (“P” paths, same NH set)



LPM	ECMP FEC	FEC	EEDB
N	1	P	P

Resource utilization – MPLS Forwarding

“N” MPLS (LDP) prefixes Uni-path via “X” Next-hops

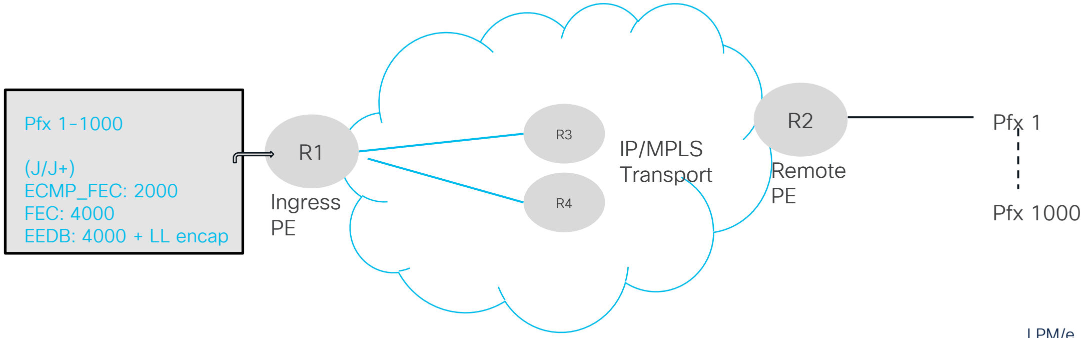


Label in Leaf (LEM) for SWAP

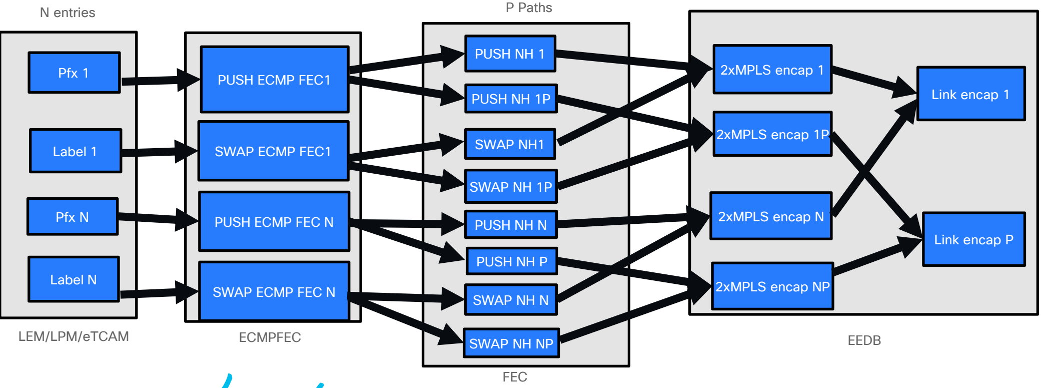
LPM/ eTCAM LEM	LEM	ECMP FEC	FEC	EEDB
N Prefixes	N Labels	0	N	(N) + X

Resource utilization – MPLS Forwarding (ECMP)

“N=1000” MPLS LDP prefixes Multipath (“P=2” paths, same IGP NH set)



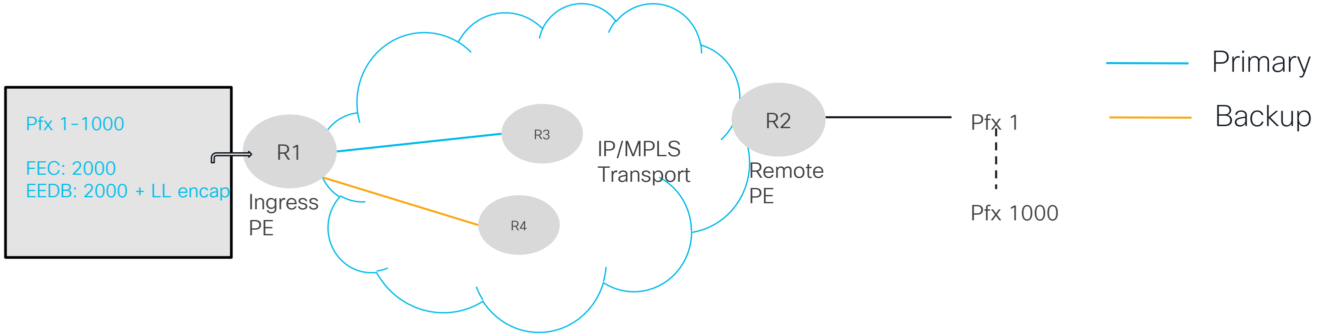
J2 Optimized: Encap label action

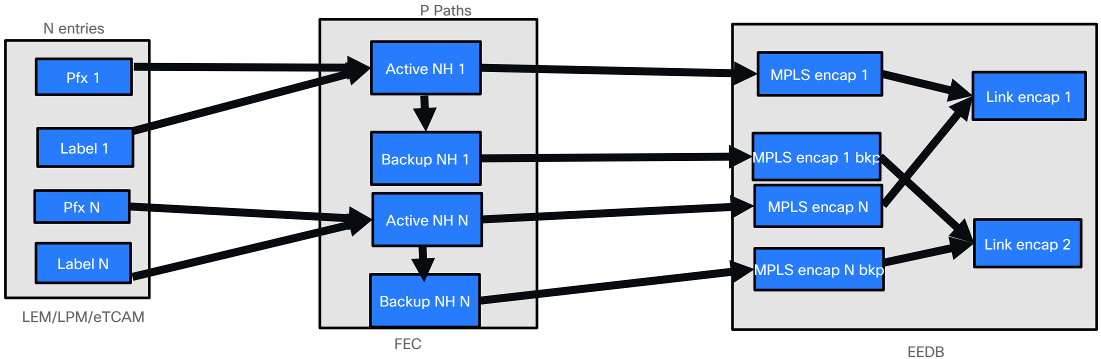


Databases	LPM/eTCAM LEM	LEM	ECMP FEC	FEC	EEDB
Calculation	N Prefixes	N Labels	2 x N N(J2)	ECMP_ FEC x P	P x (ECMP_ FEC) + P
Illustration 1K , 2 paths	1000	1000	2000 1000 (J2)	4000 2000(J2)	4002 2002(J2)

Resource utilization – MPLS Forwarding (FRR)

“N=1000” MPLS LDP prefixes FRR/LFA (Backup using different label)



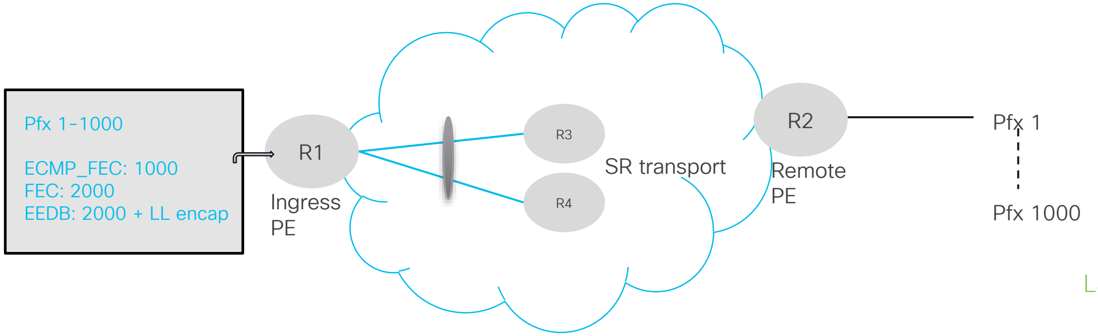


IMP_LSP Shared encap case

Databases	LPM/e TCAM LEM	LEM	FEC	EEDB
Calculation	N Prefix es	Label s	(Nx2) + (2 x unique NH set)	Nx2
Illustration 1K with FRR (1 NH set)	1000	1000	2002	2002

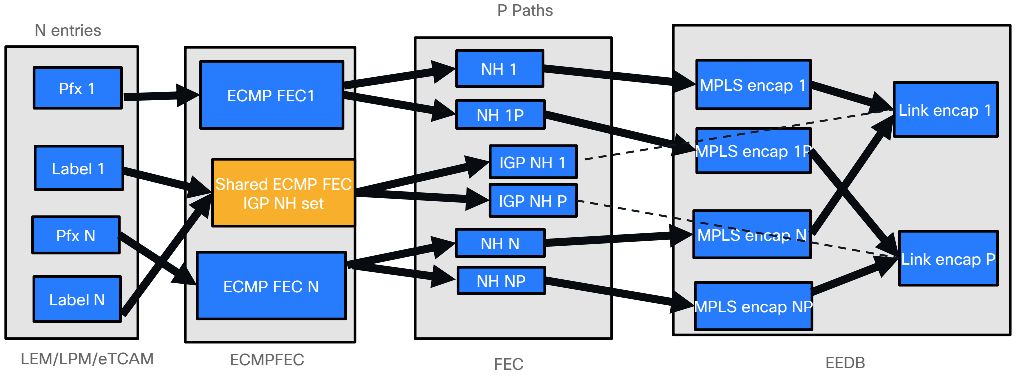
Resource utilization – SR Forwarding (ECMP)

“N=1000” prefixes Multipath (P=2 paths, same NH set)



Pfx 1-1000

ECMP_FEC: 1000
FEC: 2000
EEDB: 2000 + LL encap

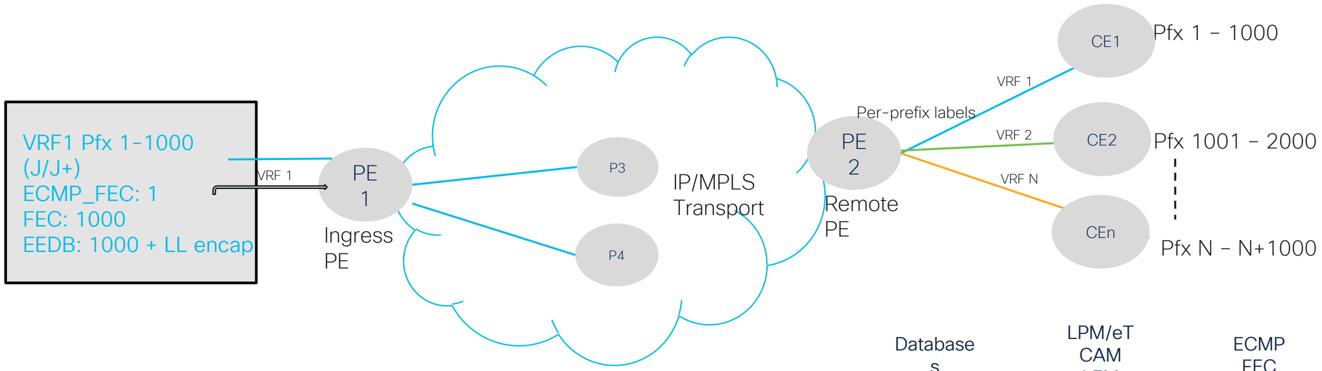


LSP – Label in Leaf/LEM , Shared ECMP FEC

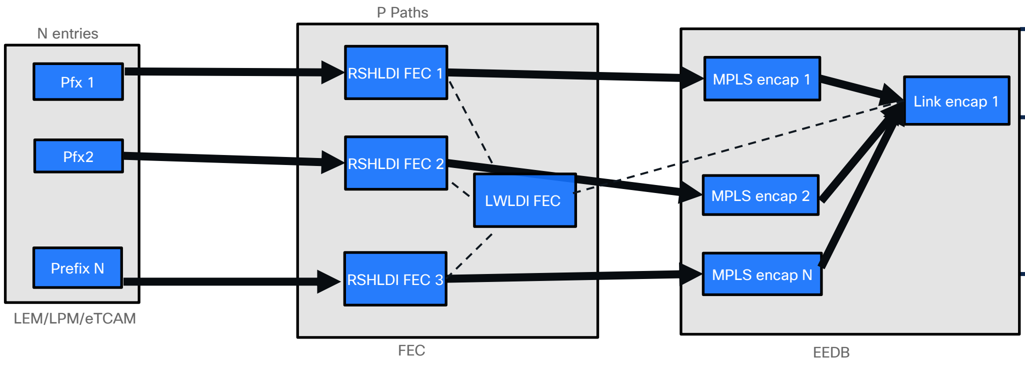
Databases	LPM/eTCAM LEM	LEM	ECMP FEC	FEC	EEDB
Calculation	N Prefixes	N Labels	N	ECMP_FEC x P	P x (ECMP_FEC) + P
Illustration 1K , 2 paths	1000	1000	1000 + 1	2002	2002

Resource utilization – L3VPN over IGP

“N=1000” VRF prefixes (per-prefix mode) – Single level recursion. Same IGP NH


VRF1 Pfx 1-1000 (J/J+)
ECMP_FEC: 1
FEC: 1000
EEDB: 1000 + LL encap

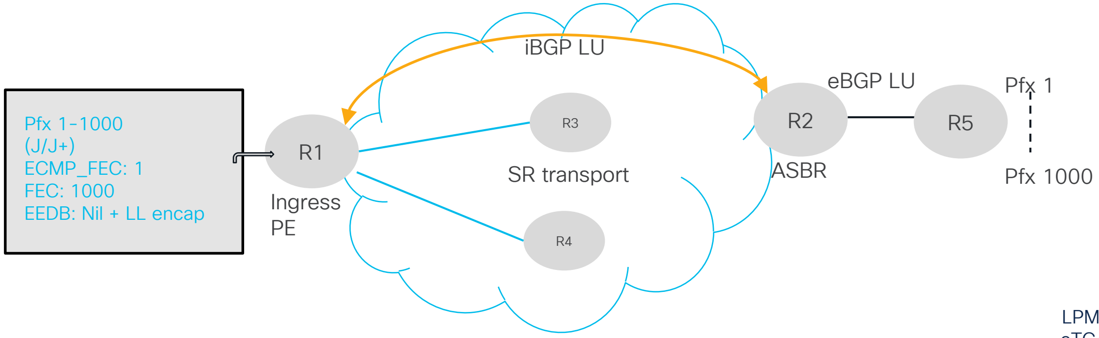
EEI Push of service labels
Label in FEC or Label in eTCAM

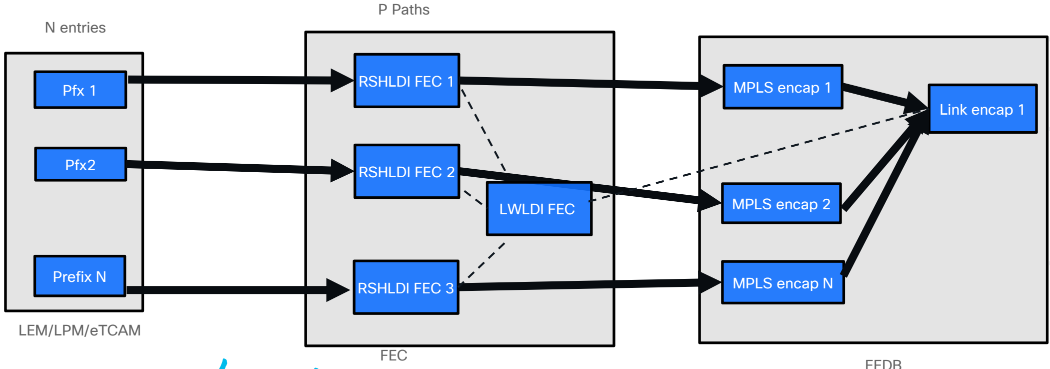


Databases	LPM/eTCAM LEM	ECMP FEC	FEC	EEDB
Calculation	N Prefixes	-	N EEI Push (New)	N (Old) Nil (New)
J/J+, J2 base 1K prefixes, same IGP-NH	1000	1 Indirection-FEC per NH set	~1000	~1000(Old) ~Nil (New – EEI Push Label in FEC)
J2 SE (OP2)	1000	1 Indirection-FEC per NH set	~Nil (EEI push: Label in eTCAM)	~Nil

Resource utilization – BGP LU Loopback Peering

“N=1000” BGP LU prefixes– Single level recursion, Same NH


Pfx 1-1000
(J/J+)
ECMP_FEC: 1
FEC: 1000
EEDB: Nil + LL encap

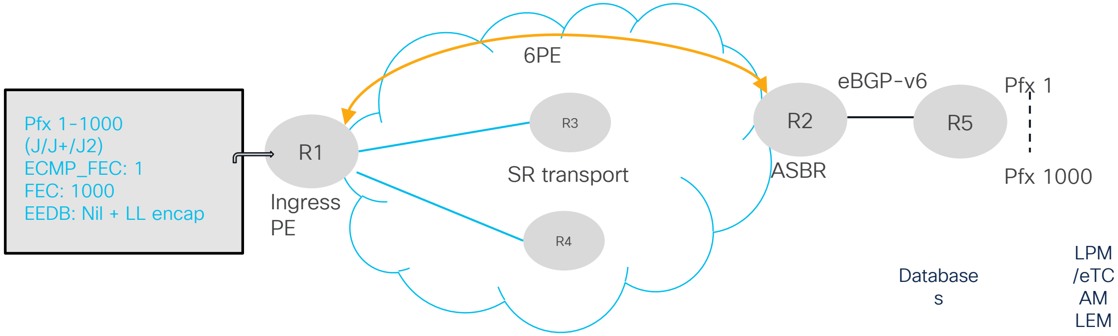


EEI Push of LU labels
Label in FEC

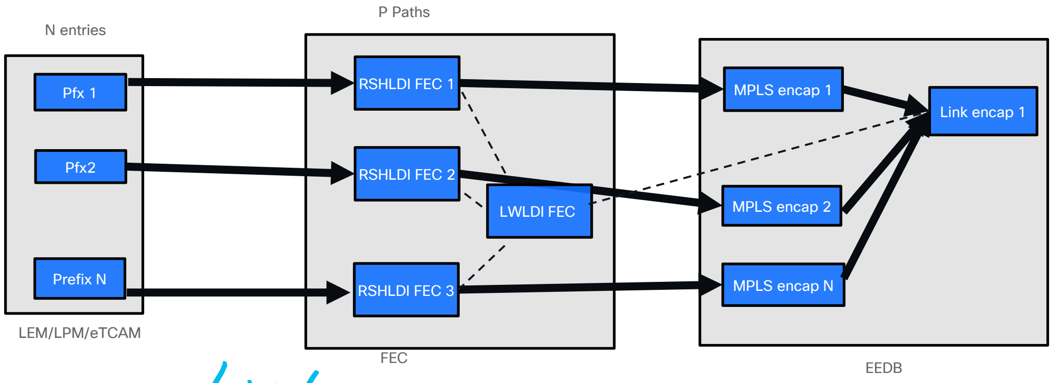
Databases	LPM/ eTCAM LEM	ECMP FEC	FEC	EEDB
Calculation	N Prefixes	-	N	N
J/J+, 1K prefixes, same IGP-NH	1000	1 Indirection-FEC per NH set	~1000	~1000(Old) ~Nil (New – EEI Push Label in FEC)
J2	1000	1 Indirection-FEC per NH set	~1000	~1000

Resource utilization – 6PE

“N=1000” ipv6 6PE prefixes -Same NH. Per-prefix label mode


Pfx 1-1000
(J/J+/J2)
ECMP_FEC: 1
FEC: 1000
EEDB: Nil + LL encap

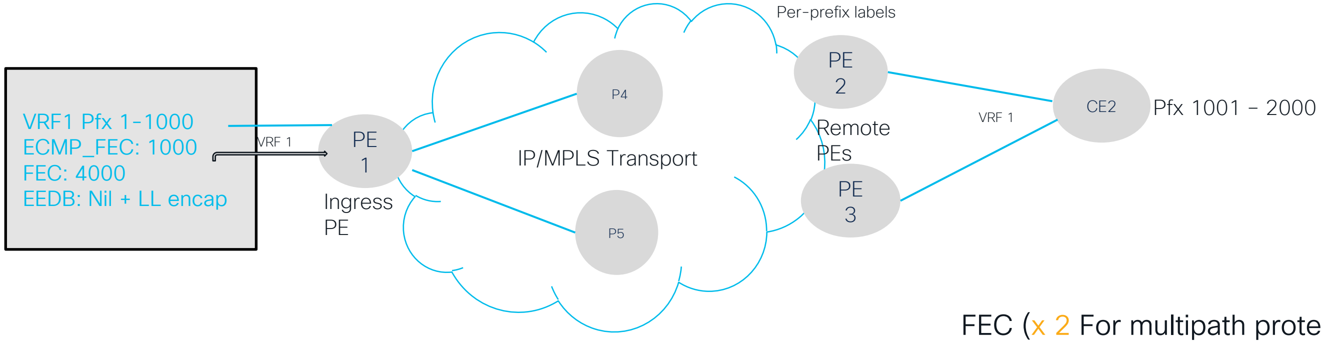
EEI Push of 6PE labels
Label in FEC/TCAM

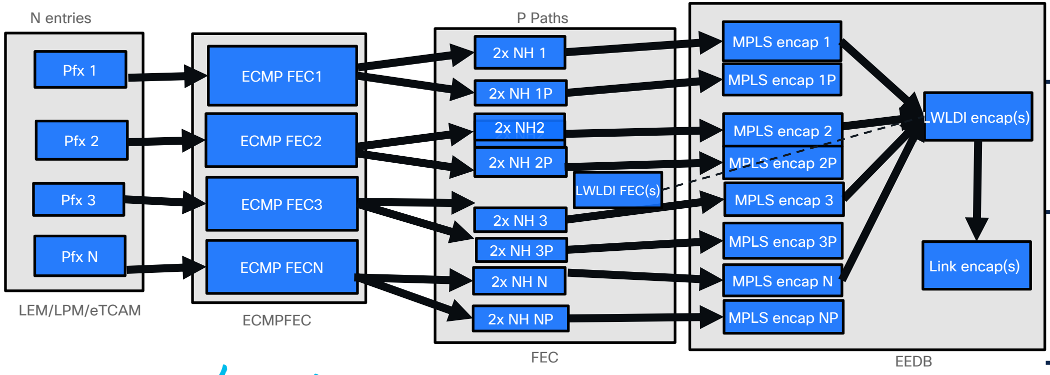


Databases	LPM /eTCAM LEM	ECMP FEC	FEC	EEDB
Calculation	N Prefixes	-	N	N
J/J+,/J2 1K prefixes, same IGP-NH	1000	1 Indirection-FEC per NH set	~1000	~1000(Old) ~Nil (New – EEI Push Label in FEC)
J2-SE	1000	1 Indirection-FEC per NH set	~Nil (Label in TCAM)	~Nil + Link encap

Resource utilization – L3VPN (Multipath) over IGP

“N=1000” vpnv4/v6 prefixes (per-prefix label allocation) via Multipath (P=2 BGP paths), same BGP NH-set (PE2,PE3)



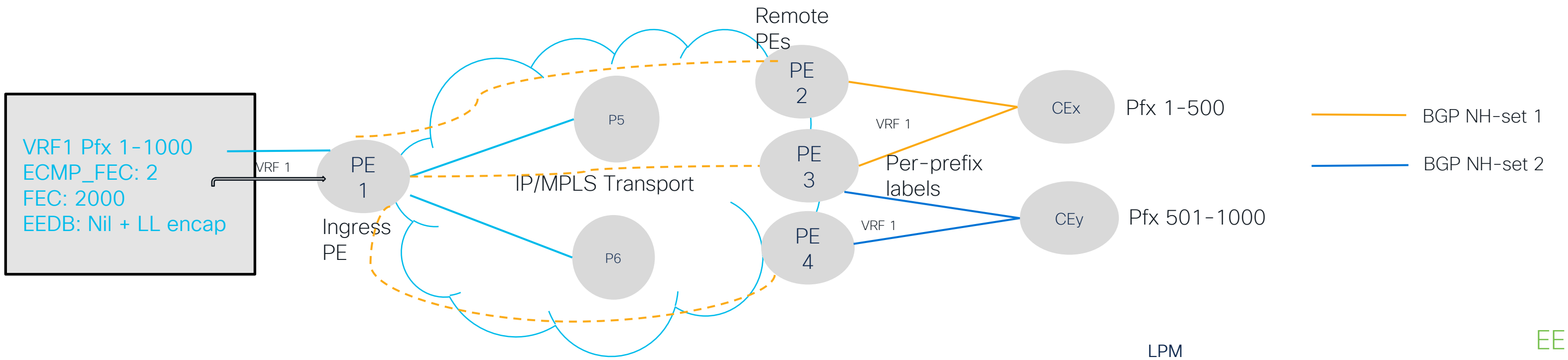


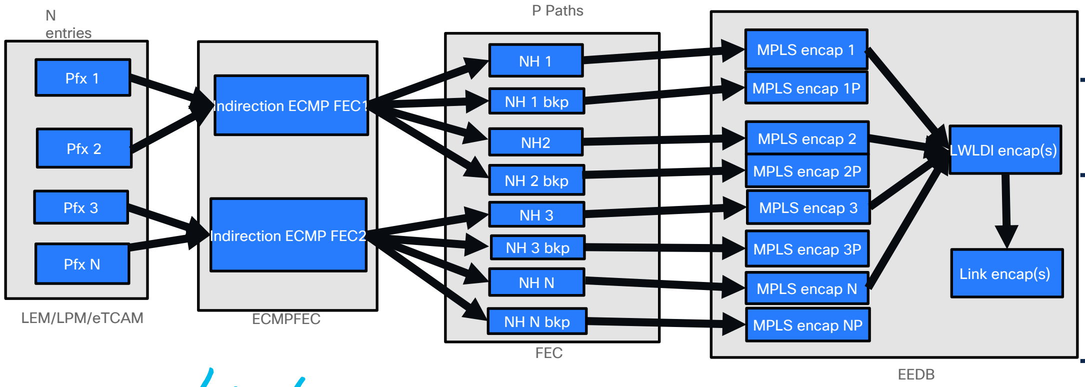
FEC (x 2 For multipath protection) EEI PUSH

Databases	LPM/ eTCAM LEM	ECMP FEC	FEC	EEDB
Calculation	N Prefixes	N	N x P x 2	N x P
Illustration 1K , 2 BGP paths (same BGP NH-set)	1000	~1000	4000	~2000 (Old) ~Nil (New -> EEI Push Label in FEC)

Resource utilization - L3VPN (PIC) over IGP

“N=1000” vpnv4/v6 prefixes (per-prefix label allocation) with PIC
“P=2” BGP paths, 2 different BGP NH-sets (PE2,PE3) (PE3,PE4)

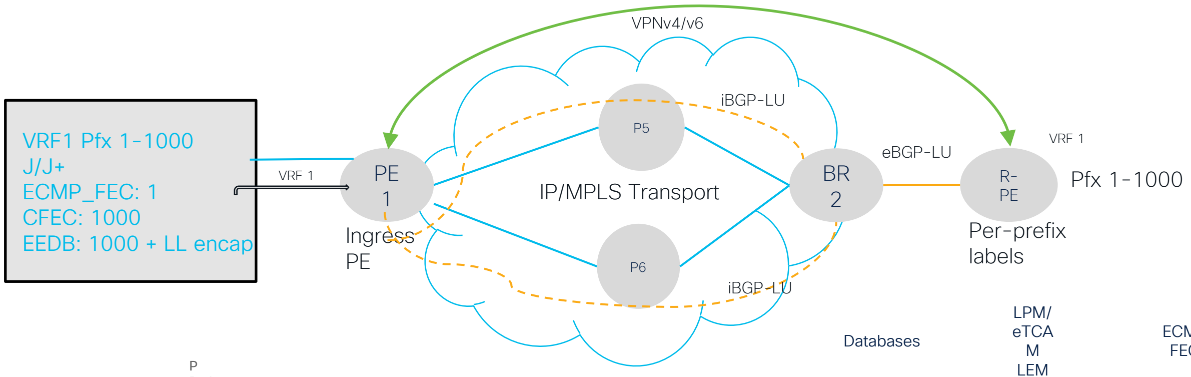

VRF1 Pfx 1-1000
ECMP_FEC: 2
FEC: 2000
EEDB: Nil + LL encap



Databases	LPM/eTCAM LEM	ECMP FEC	FEC	EEI PUSH EEDB
Calculation	N Prefixes	~	N x P(2)	NxP(2)
Illustration 1K , 2 BGP paths (2 BGP NH set)	1000	2 indirection ecmp-fec (1 per BGP-NH set)	2000	2000 (Old) ~Nil (New – EEI Push Label in FEC)

Resource utilization - L3VPN over BGP-LU (3 levels)

“N=1000” VRF prefixes (per-prefix mode) – Two level recursion (VPN over BGP-LU over IGP-SR/LDP)



VRF1 Pfx 1-1000
J/J+
ECMP_FEC: 1
CFEC: 1000
EEDB: 1000 + LL encap

VRF 1

PE 1

Ingress PE

VPNv4/v6

iBGP-LU

P5

IP/MPLS Transport

P6

iBGP-LU

BR 2

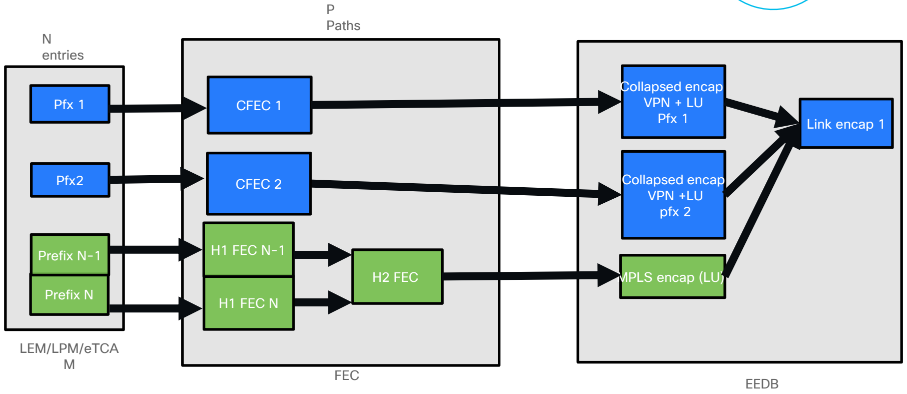
eBGP-LU

VRF 1

R-PE

Pfx 1-1000

Per-prefix labels



N entries

Pfx 1

Pfx2

Prefix N-1

Prefix N

LEM/LPM/eTCAM

P Paths

CFEC 1

CFEC 2

H1 FEC N-1

H1 FEC N

H2 FEC

FEC

Collapsed encap VPN + LU Pfx 1

Collapsed encap VPN +LU pfx 2

MPLS encap (LU)

Link encap 1

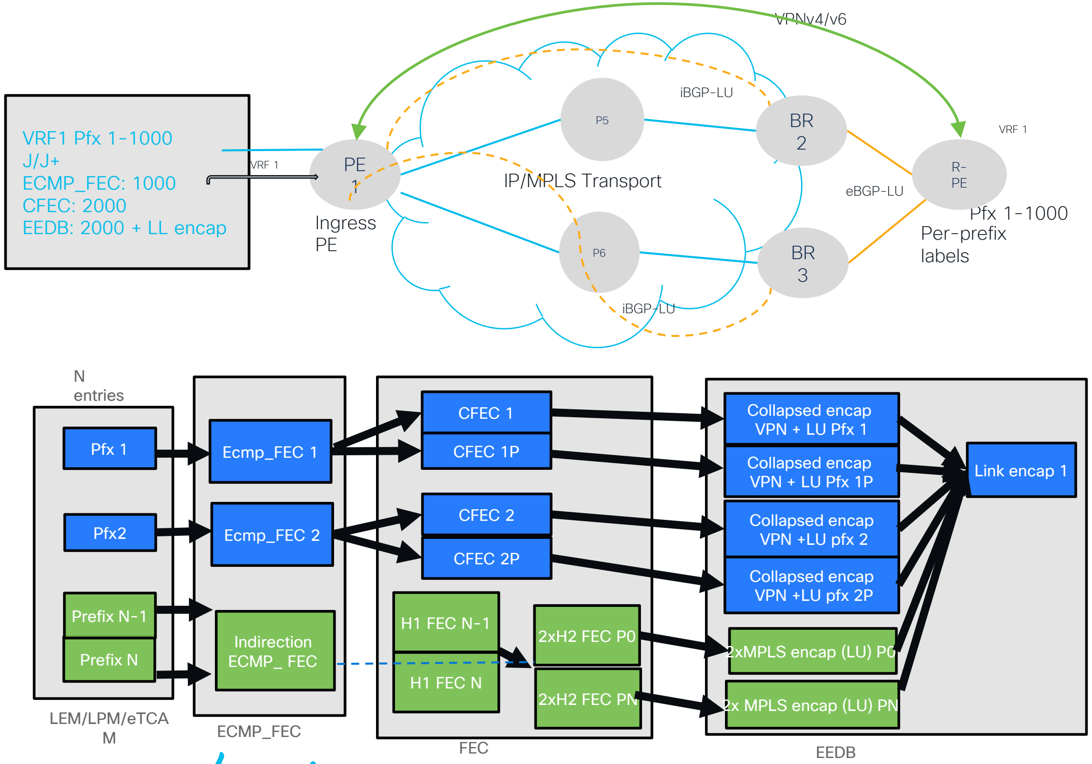
EEDB

Collapsed LDI – J/J+ ; 3 level FEC – J2

Databases	LPM/ eTCAM LEM	ECMP FEC	FEC	EEDB
Calculation	N Prefixes	-	N	N
J/J+ 1K prefixes	1000	1 Indirection -FEC per NH set	~1000 (CFEC)	~1000(CFEC)
J2-nonSE	1000	1 Indirection -FEC per NH set	~1000 (H1)	~Nil (EEI push from H1 FEC)
J2 SE (eXR)	1000	1 Indirection -FEC per NH set	~Nil (Label in TCAM) + limited FECs for LU, IGP	~Nil (EEI push Label in TCAM) + limited encaps for LU, IGP

Resource utilization L3VPN over BGP-LU Multipath

"N=1000" VRF prefixes (per-prefix mode) – Two level recursion (VPN over BGP-LU Multipath(P=2) over IGP-SR/LDP)

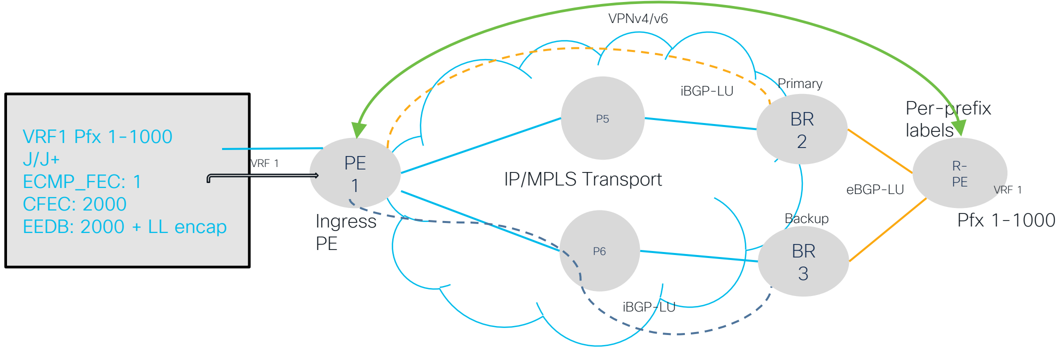


Collapsed LDI – J/J+ 3 level FEC – J2

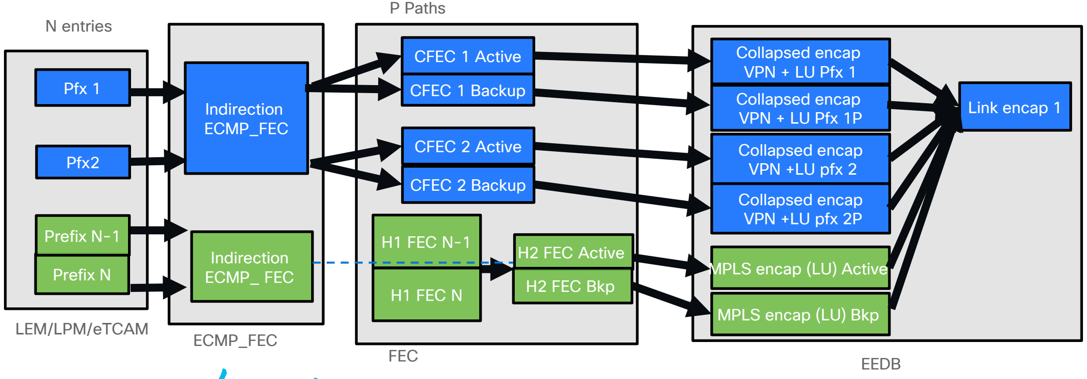
Databases	LPM/eTCAM LEM	ECMP FEC	FEC	EEDB
Calculation	N Prefixes	N	(NxP) + (Px2) x 2	N
J/J+ 1K prefixes 1 BGP(LU)-NH set(2 paths)	1000	1000 + 2 Indirection-FEC per NH set for LU	~2000 (CFEC) + 8	~2000(CFEC)
J2-nonSE	1000	1 Indirection-FEC per BGP-LU NH set	~1000 (H1) + 4	~Nil + 4 (EEI push from H1 FEC)
J2-SE (eXR)	1000	1 Indirection-FEC per BGP-LU NH set	~Nil + 4	~Nil + 4 (EEI push from LEAF)

Resource utilization L3VPN over BGP-LU PIC

“N=1k” VRF prefixes (per-prefix mode) – Two level recursion (VPN over BGP-LU PIC over IGP-SR/LDP)



VRF1 Pfx 1-1000
J/J+
ECMP_FEC: 1
CFEC: 2000
EEDB: 2000 + LL encap



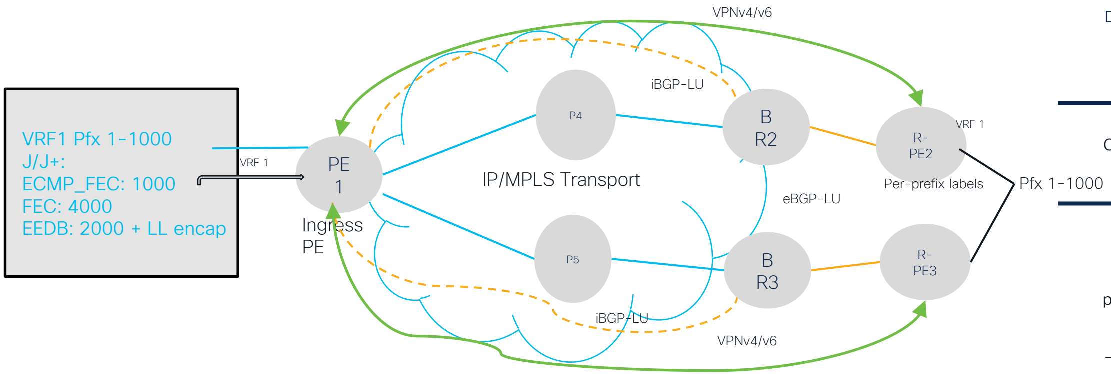
Collapsed LDI – J/J+   3 level FEC – J2

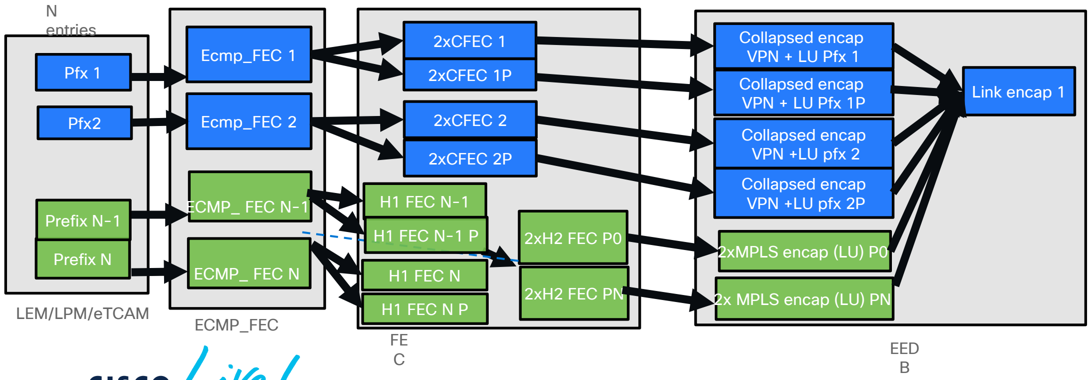
Databases	LPM/eTCAM LEM	ECMP FEC	FEC	EEDB
Calculation	N Prefixes	N	(N x P) + 4	N
J/J+ 1K prefixes 1 BGP(LU)-NH set(2 paths)	1000	1Indirection-FEC per NH set for LU	~2000 (CFEC) + 4 4 – LU Imp/swap active/bkp	~2000(CFEC)
J2-nonSE	1000	1 Indirection-FEC per BGP-LU NH set	~1000 (H1) + 2	~Nil + 4 (EEI push from H1 FEC)
J2-SE (eXR)	1000	1 Indirection-FEC per BGP-LU NH set	~Nil + 2	~Nil + 4 (EEI push from LEAF)

Resource utilization L3VPN Multipath over BGP-LU

“N=1K” VRF prefixes (per-prefix mode) – Two level recursion (VPN MP over BGP-LU over IGP-SR/LDP)

Collapsed LDI – J/J+   3 level FEC – J2

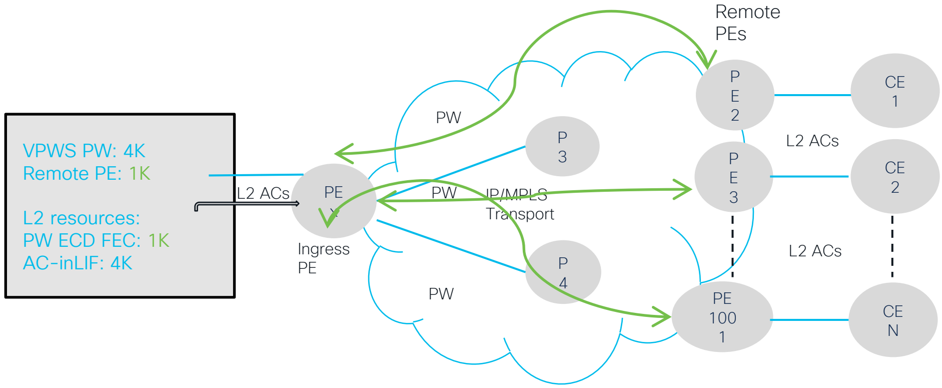

VRF1 Pfx 1-1000
J/J+:
ECMP_FEC: 1000
FEC: 4000
EEDB: 2000 + LL encap



Databases	LPM/eTCAM LEM	ECMP FEC	FEC	EEDB
Calculation	N Prefixes	N	(NxPx2) + (Px2)	N
J/J+ 1K vpn prefixes – 2 paths BGP LU – 2 peers	1000	1000 + 2 Indirection-FEC per NH set for LU	~4000 (CFEC) + 4	~2000
J2-nonSE	1000	1000 + 1 Indirection-FEC per BGP NH set	N x (P) ~2000 (H1) + 4	~Nil + 4 (EEI push from H1 FEC)
J2-SE (eXR)	1000	1000+ 1 Indirection-FEC per BGP NH set	~2000(H1) + 4	~Nil + 4 (EEI push from H1 FEC)

Resource utilization – L2VPN over IGP

“4K” L2VPN PW over IGP-SR/LDP across 1K remote-PEs



Database usage	4K VPWS PW, 1K Remote PEs
LPM/eTCAM,or LEM	1K Prefixes (RPE Loopbacks) 1K Transport labels
LIF, SEM	4K (AC-Lif), 4K (PWE-Lif) 4K VC local labels(iSEM)
FEC	1K ECD FEC 1K IGP FEC
EEDB	4K PWE encap (Remote VC labels), 4K AC-encap 1K encap (Transport labels)

- Imposition : AC to MPLS
  - AC-Lif lookup points to indirection-FEC (ECD) which points to IGP FEC. (ECD FEC unique per Remote PE)
  - AC-Lif also points to encap (PWE-lif) which stores remote VC label
- Disposition: MPLS to AC
  - Local PW VC label (iSEM lookup) will point to PWE-lif which resolves to AC-lif (l2 rewrite) & destination port
- AC-Lif & PWE-Lif entries are symmetric for ingress and egress pipeline stages

ARP/ND scale in NCS5500/5700

- ARP/ND scale depends on the encap allocation(Controlled by GRID)
- NCS5700 has better encap scale
  - Dedicated encap phase for ARP/ND (phase 7)
  - Applicable for physical/BVI
  - Sharing the cluster bank pair with other applications
- L3MAX-SE illustration
  - ARP uses 2 x 60b encap entries (for all MDB)
  - GRID restricts ARP/ND for physical to 32K
  - BVI can use the full encap cluster bank
  - 64K ARP (BVI or Phy+BVI combo) possible as 1D scale (Comes at the cost of tunnel1 resources)
  - Shares the cluster bank resources with tunnel1 (reserved for BGP LU, SRTE, SRV6 T.Insert)
- New CLI on NCS5700 “sh controller npu resources encapARP location <> “
- NCS5500 Max encap entries 8K
- New enhancements to optimize the encap allocation for ND based on MAC instead of # of ipv6 addresses

Common Network Design problems with NCS5500

Designing with multipaths/ECMP

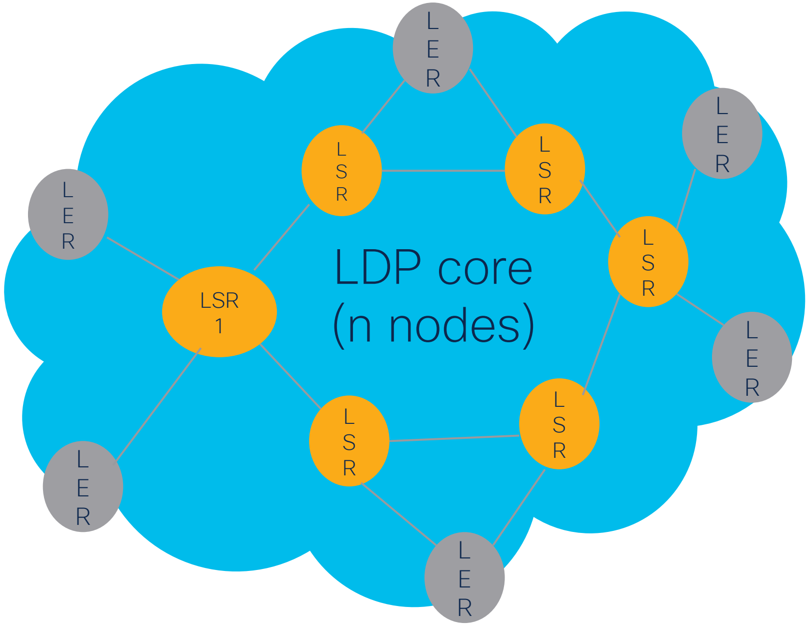
- ECMP_FEC is a sparse resource in NCS5500 (only 4K) & allocated when programming chains with multipaths
- ECMP_FECs are used for
  - IP unlabeled ECMP
  - IP labeled / MPLS LSP ECMP
  - BGP LU, L3 VPN multipath
  - EVPN multi-homing
  - SRTE (first SID ECMP, multi-SID-List candidate paths)
- For every labelled prefix with ecmp, we burn an ECMP_FEC
- Limits the use of labelled ECMP deployments (typically multipath for transport labels with BGP_LU advertising the remote PE loopbacks)

"Abusers"

Optimize ECMP_FEC consumption

- Eliminate ECMP_FEC usage (Best option ☺)
    - IGP (ISIS) level, we can restrict it with knob "maximum-paths 1"
    - ISIS enhanced to randomly pick different path(next-hop) for different prefixes
    - Links are still not underutilized as we don't end up programing the same link for all prefixes
- Restrict ECMP_FEC usage
    - Use LDP filters to reduces the label allocation only to the remote PE loopbacks
- Move to Segment-Routing for optimized ECMP_FEC
    - SR FIB optimizations in-place for better ECMP_FEC handling
- Other FIB optimizations in-place to reduce usage of ENCAP, FEC

Flat IGP domain: Scaled labelled(LDP) ECMP routes

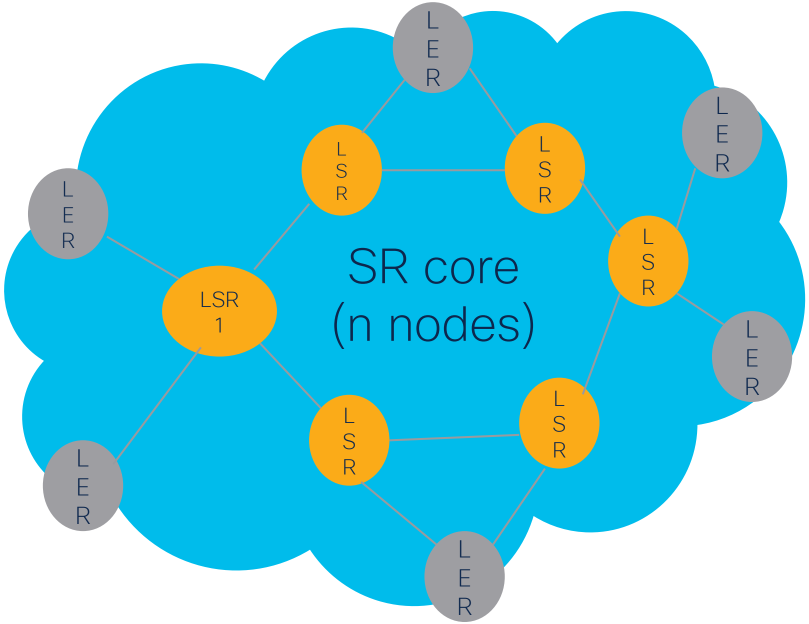


- With multipath enabled in IGP (OSPF/ISIS) LER/LSR nodes will have to program ecmp path for all IGP prefixes
- IP ECMP paths will use the same ECMP_FEC for many prefixes resolved via the same NH
- MPLS ECMP paths will use unique ECMP_FEC for the LDP prefixes
- In MPLS for multipath we assign one ECMP_FEC for push and one for SWAP

LSR1: (Illustration)

- Assume 5000 prefixes learnt via ISIS with LDP (all of them have multipath)
- IP ECMP_FEC will be less (= NH set) mostly < 10
- 2 x MPLS ECMP_FEC (PUSH, SWAP) will be allocated for 5K prefixes = 10K ECMP FEC required > 4K limits on J/J+

Flat IGP domain: Scaled labelled(SR) ECMP routes



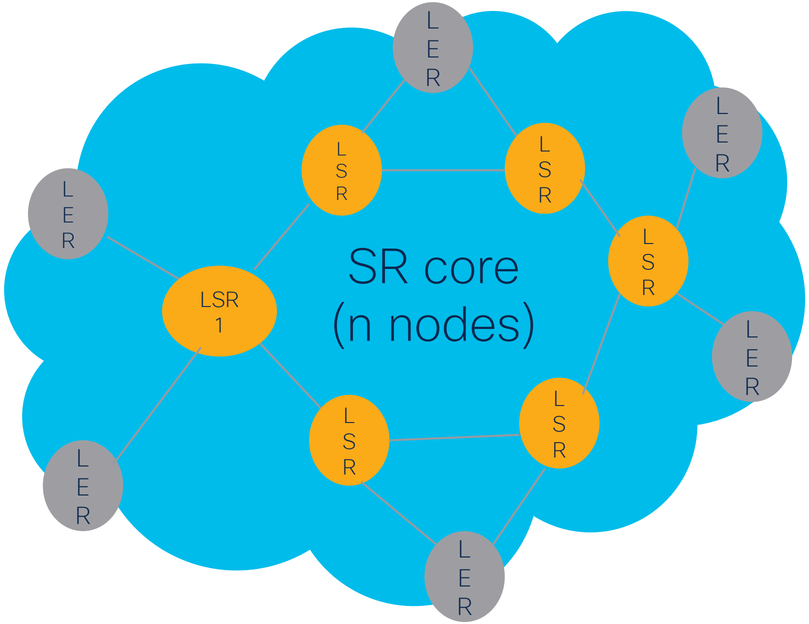
Innovation: LEM Optimization

- LEM stores OUT labels along with the corresponding Local labels
- Only possible if all OUT labels are same for all ECMP NH (Applicable only for SR prefixes)
- For SWAP shared ECMP FEC used (=NH Set)
- For PUSH scenario, use unique ECMP_FEC for each SR prefixes

LSR1: (Illustration with SR)

- Assume 5000 prefixes learnt via ISIS with SR (all of them have multipath)
- IP ECMP_FEC will be less (= NH set) mostly < 10
- SWAP ECMP_FEC = NH Set mostly <=10
- 1 x ECMP_FEC for PUSH per prefix = 5K ECMP FEC required > 4K limits on J/J+

Flat IGP domain: Scaled labelled(SR) ECMP routes With SR ECMP_FEC optimization



Innovation: SR ECMP_FEC optimization (needs a hw-mod profile)

- LEM stores OUT labels along with the host prefixes (/32)
- Only possible if all OUT labels are same for all ECMP NH (Applicable only for SR prefix-sid)
- For both PUSH & SWAP shared ECMP FEC used (=NH Set)

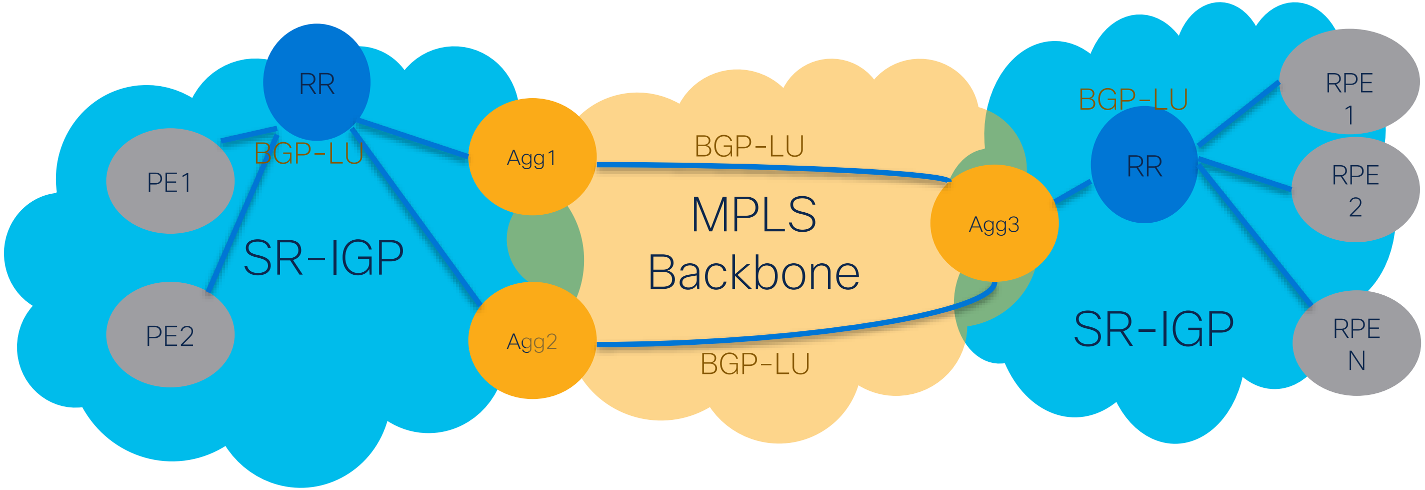
LSR1: (Illustration with SR)

- Assume 5000 prefixes learnt via ISIS with SR (all of them have multipath)
- IP ECMP_FEC will be less (= NH set) mostly < 10
- PUSH & SWAP ECMP_FEC = NH Set mostly <=10

Interdomain routing with BGP-LU Multipath

Agg1, Agg2 routers advertising remote PE(RPE) loopbacks over BGP LU to PE1,PE2

(BGP Multipath enabled on PE1, PE2)



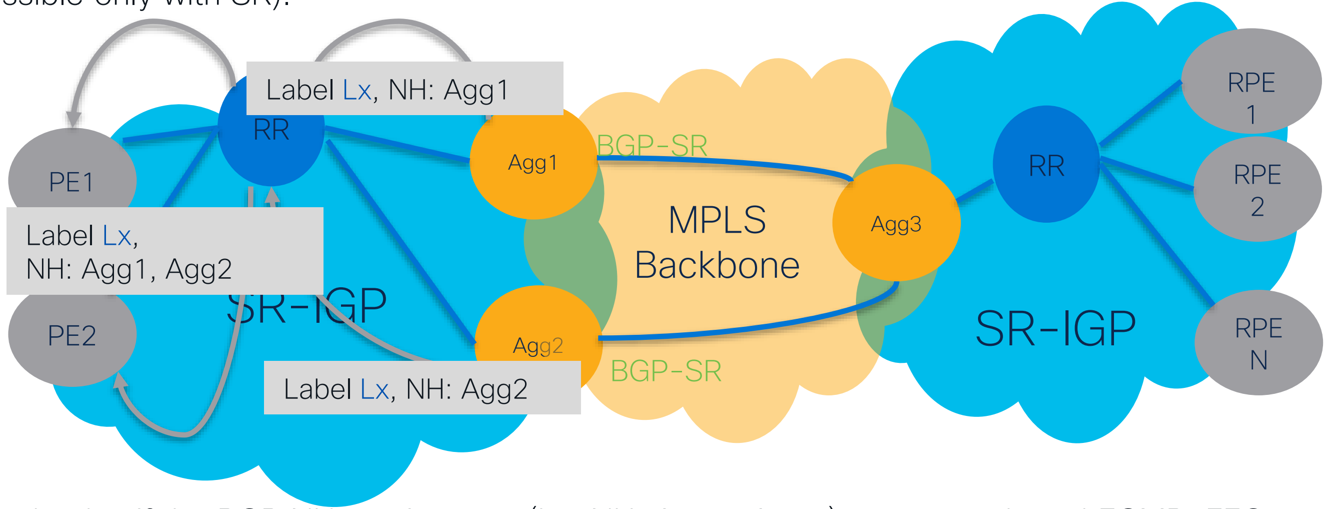
For N Remote PE loopbacks, ECMP FEC requirement on PE1/PE2 will be 2xN.

Illustration: for 5K remote PEs , 10K ECMP FEC required for BGP LU Multipath (> 4K on J/J+)

Interdomain routing BGP-SR Multipath optimization

Agg1,Agg2 routers advertising same labels for every remote PE(RPE) loopbacks over BGP SR(proxy SR) to PE1,PE2. (BGP Multipath enabled on PE1, PE2)

Used Shared ECMP_FEC optimization for the pattern “Label in leaf” with same outgoing labels for a given RPE loopback (possible only with SR).



For N RPE loopbacks, if the BGP NH-set is same (Lx, NH: Agg1, Agg2) we use a shared ECMP_FEC

Illustration with BGP-SR: For any # remote PEs with 10 different BGP NH-SETs, only 10 ECMP FEC required

Key Take Aways

NCS5500/NCS5700 an optimized transport platform

Super capable of various roles in Network !

- Scalable Network design made possible with hardware and software innovations
- Constraints can be addressed with better understanding of resources mapping
- Segment Routing (SR/SRv6) key piece solving many scalability and design issues

Complete your Session Survey

- Please complete your session survey after each session. Your feedback is very important.
- Complete a minimum of 4 session surveys and the Overall Conference survey (open from Thursday) to receive your Cisco Live t-shirt.
- All surveys can be taken in the Cisco Events Mobile App or by logging in to the Session Catalog and clicking the "Attendee Dashboard” at https://www.ciscolive.com/emea/learn/sessions/session-catalog.html

Continue Your Education

- Visit the Cisco Showcase for related demos.
- Book your one-on-one Meet the Engineer meeting.
- Attend any of the related sessions at the DevNet, Capture the Flag, and Walk-in Labs zones.
- Visit the On-Demand Library for more sessions at ciscolive.com/on-demand.

cisco The bridge to possible

Thank you

cisco Live!

cisco Live!