Challenges of Layer 2 NID Based Architecture For vCPE/NFV Deployments

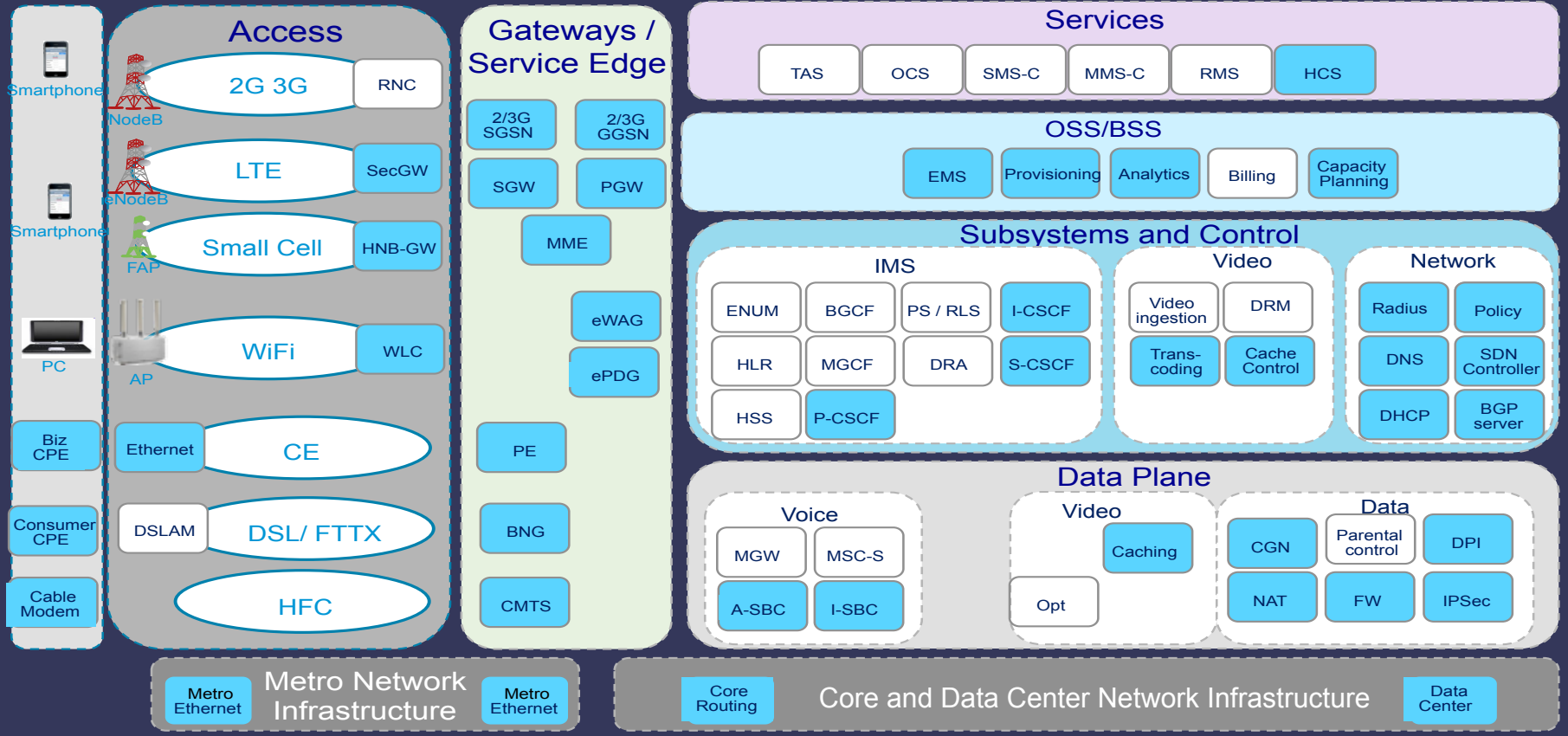
SANOG 26

Santanu Dasgupta
Sr. Consulting Engineer – Global Service Provider Network Architecture

August, 2015

Brief NFV Introduction

“Network Functions” in SP Network Architecture Landscape



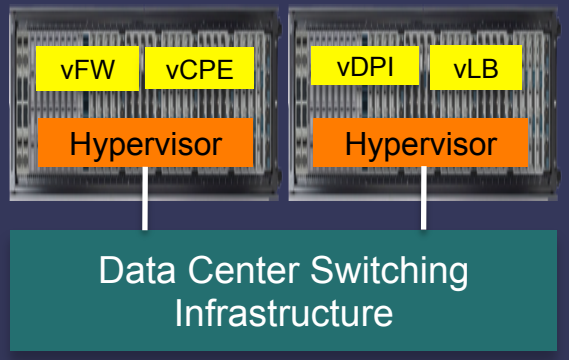
Virtualization of “Network Functions”

Existing Hardware / Appliance based Network Functions (NFs)

Step 1: Decouple software from underlying hardware

Step 2: Port it as a VM/ container on x86 Server platform running as a Network Function

Virtualized NFs running as VM on x86 Server Platform



NFV, SDN & Orchestration Together

Partial list, just a few main ones are mentioned here

VM / VNF Lifecycle Management in End-to-end manner

Network Plumbing to orchestrate dynamic topologies

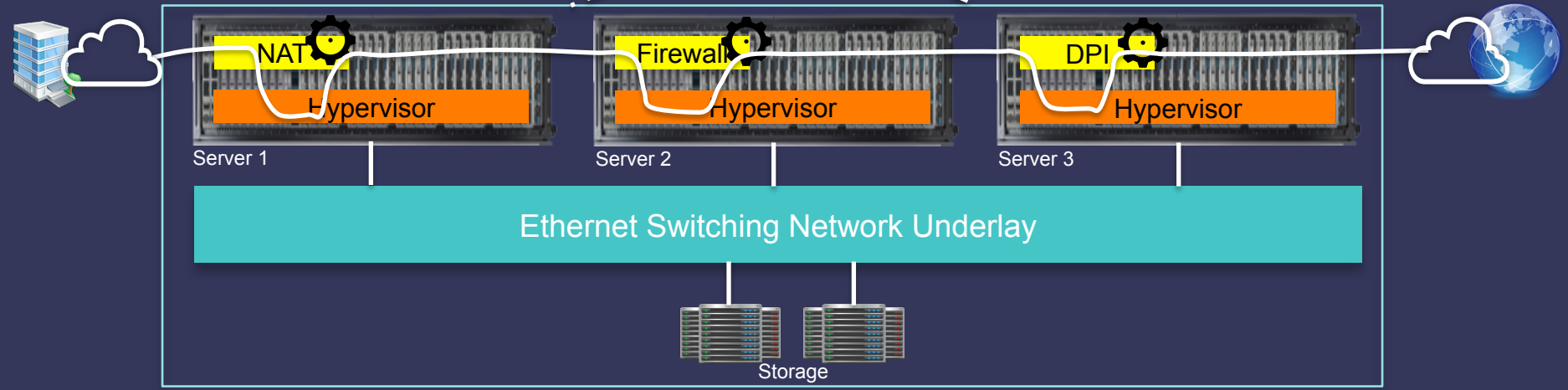
Configuration Management of the VNFs

Integration with Other DC/POD And the WAN

OAM, Assurance, Analytics

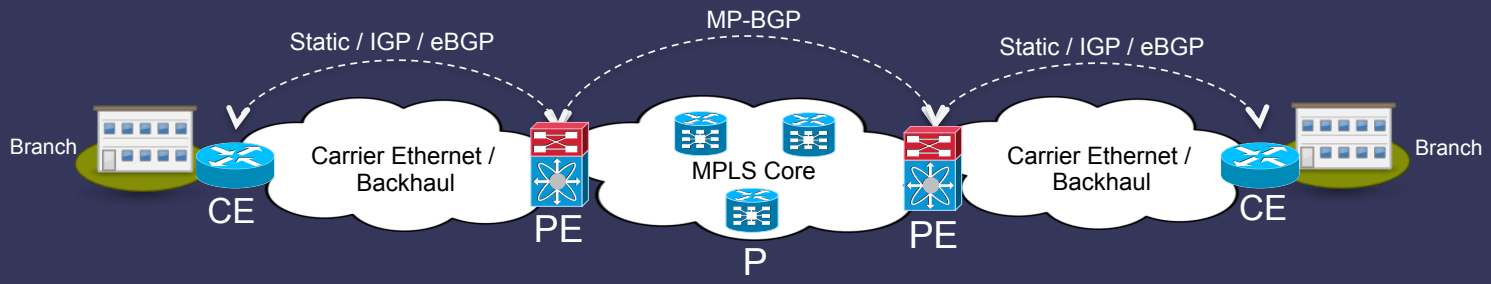
Orchestration and SDN Control Function

Standard APIs



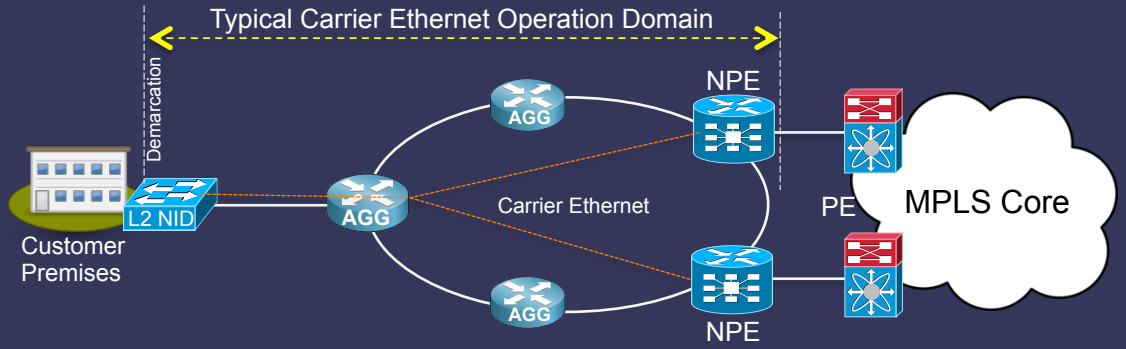
L2NID for vCPE/NFV Background & Context

Traditional Managed CPE with IP/MPLS L3VPN



- There are multiple genuine and perceived issues in the traditional service delivery model –
  - CPE provisioning and servicing often require truck roll (sending engineers) → high OPEX
  - The amount of feature sets enabled on the on-premise CPE makes the solution complex to operate
  - Service delivery is not agile, lacks automation, service turn-up / changes takes a lot of time
  - On site CPE's are often expensive and not an open platform
- The industry is expecting something that is more open, agile, fully automated, flexible to address different market segments and can help the operators to reduce their TCO
  - This is where L2 NID on-premises + vCPE architecture discussion for business VPN started almost 2 years back in the industry

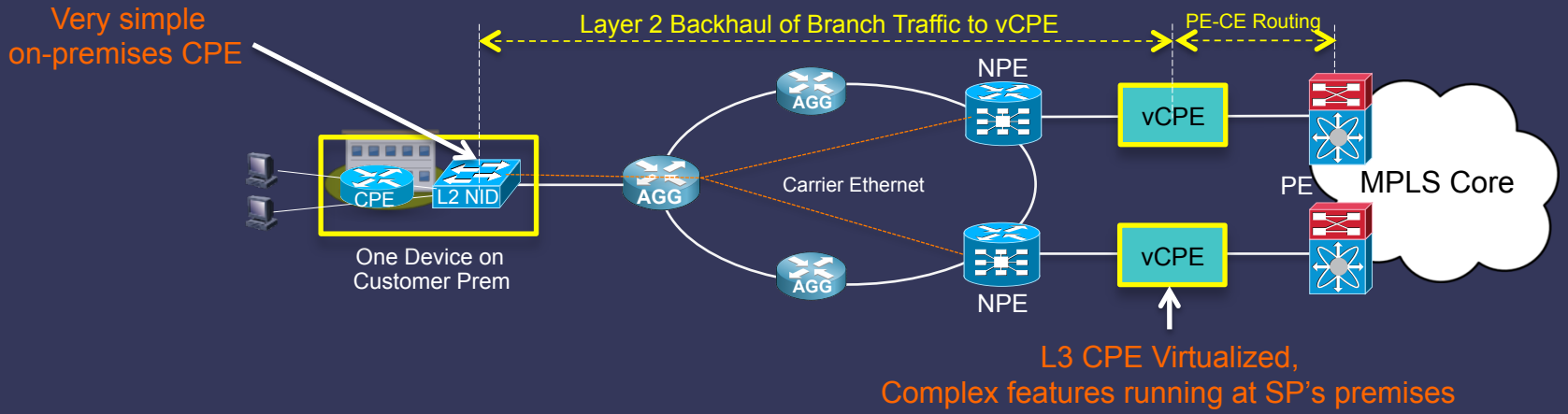
What is “L2 NID” ?



- Layer 2 NID – Layer 2 Network Interface Device (example Cisco ME 1200)
- Some calls it Layer 2 Network Termination Device (NTD) → we will call it L2 NID for this presentation
- L2 NID is the device that Carrier Ethernet Operator drops at Customer Premises to terminate the Ethernet last mile
  - It is managed by the operator
  - It has user facing interfaces (UNI) and network facing interfaces (NNI) – typically all Ethernet
  - It marks the demarcation point between the Operator and Customer Network
- The L2 NID is typically a 4 to 6 port FE/GE/10GE L2 switch with some other capabilities such as –
  - Ethernet OAM (CFM, Y.1731) for fault & performance management, Service Activation (Y.1564), timing support (mobile b/h) etc.

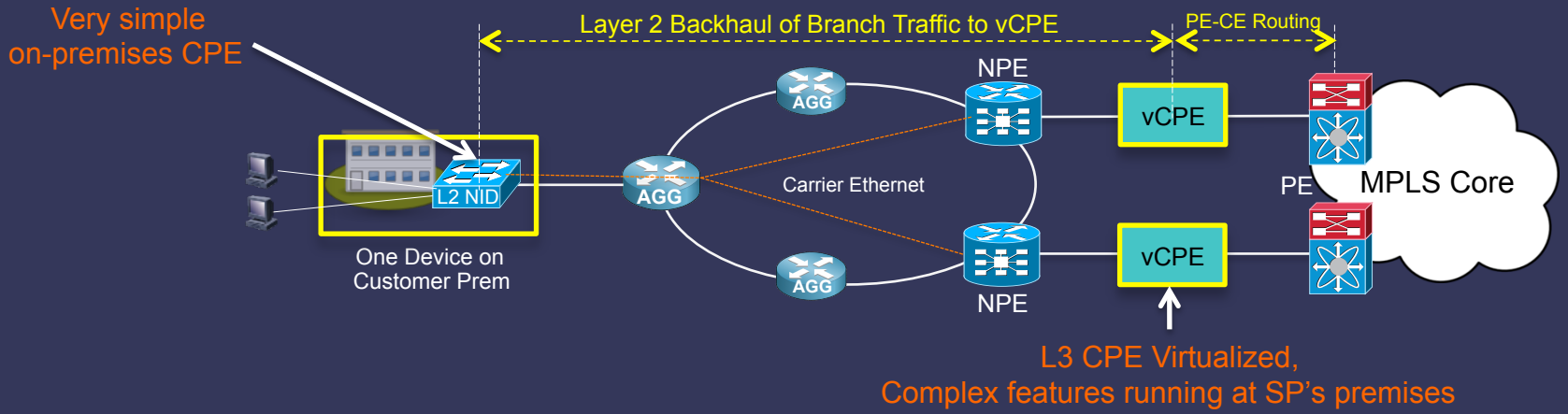
Animated

L2 NID + vCPE Architecture Proposal for Managed CPE/VPN



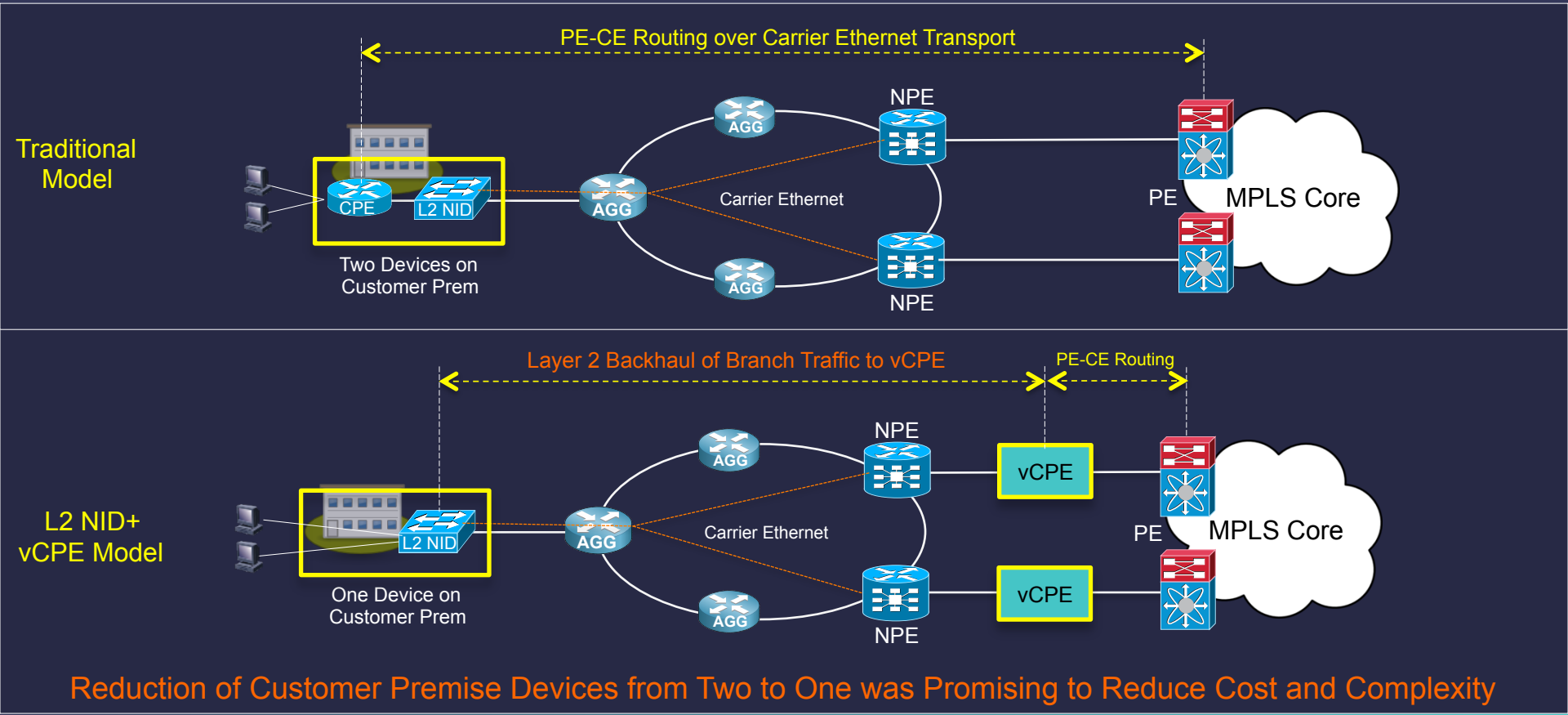
- No need to have a Layer 3 CPE at Customer premises anymore
- Virtualize the L3 CPE and Put that at SP's POP or Cloud / NFV Data Center
- Make the branch simplified with only one device, where complex features are running at SP's Cloud making it easier to operate → may also help to reduce cost

L2 NID + vCPE Architecture Proposal for Managed CPE/VPN



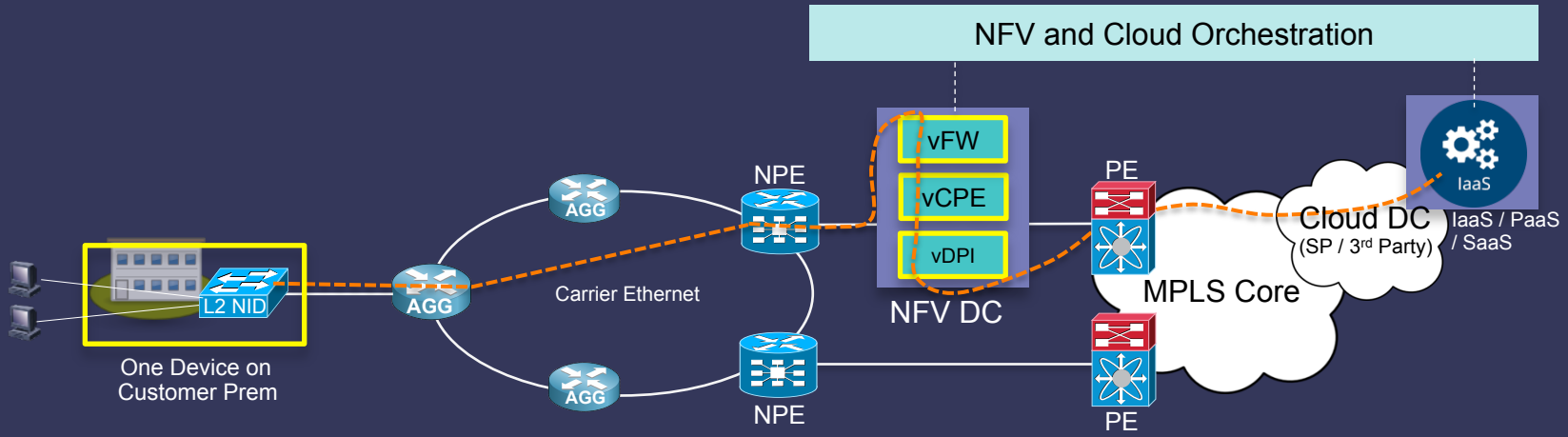
- No need to have a Layer 3 CPE at Customer premises anymore
- Virtualize the L3 CPE and Put that at SP's POP or Cloud / NFV Data Center
- Make the branch simplified with only one device, where complex features are running at SP's Cloud making it easier to operate → may also help to reduce cost

Why The L2 NID Based Alternate Looked Promising ?



Reduction of Customer Premise Devices from Two to One was Promising to Reduce Cost and Complexity

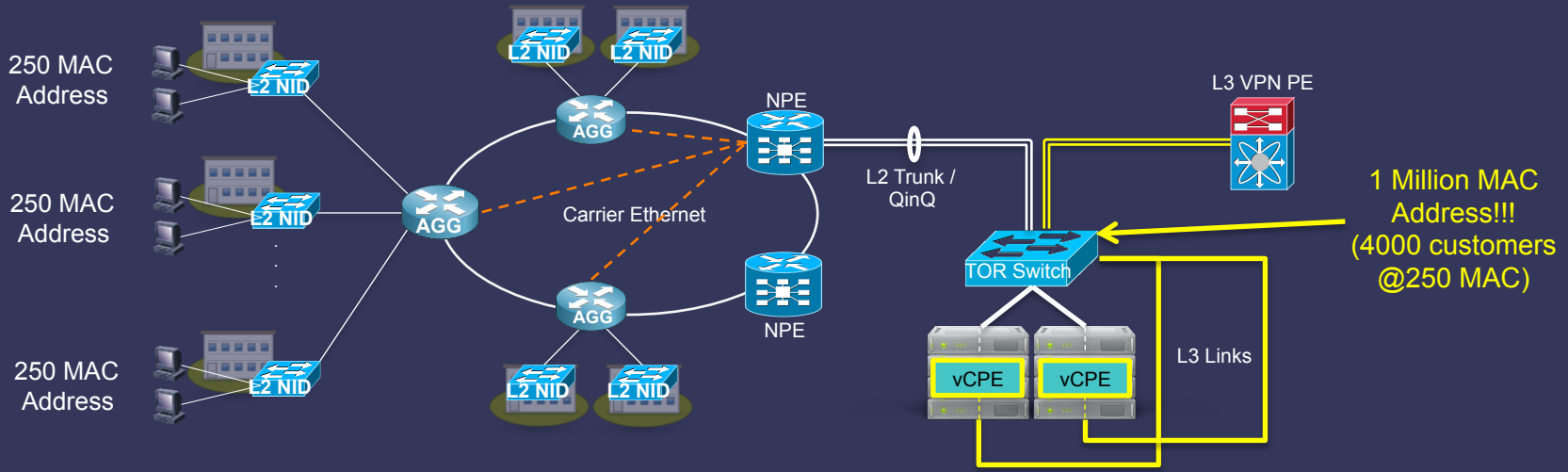
This Would Enable Agile Service Creation too



- NFV and Orchestration also enables agile service creation and turn-up
- vCPE can be chained with rich set of NFV, Cloud IaaS, PaaS and SaaS services (SP hosted or 3rd party)
- This is true irrespective of the on-premises CPE type – be in L2 NID or L3 CPE or whatever else

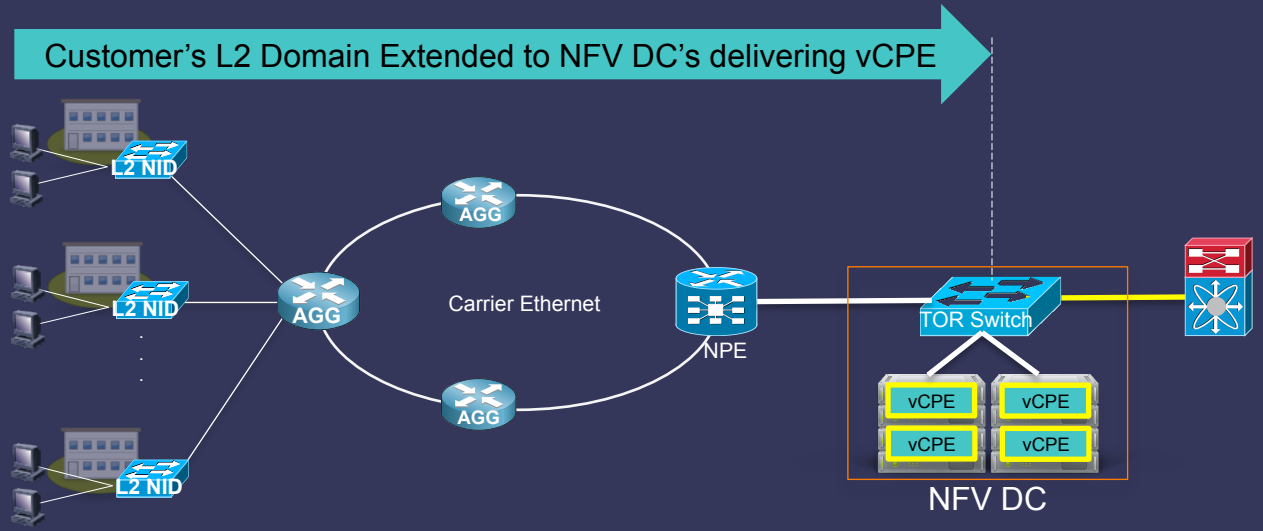
Challenges with the L2 NID Based Architecture, where there is no L3 CPE at the Branch

MAC Address Scale Issues for the NFV DC



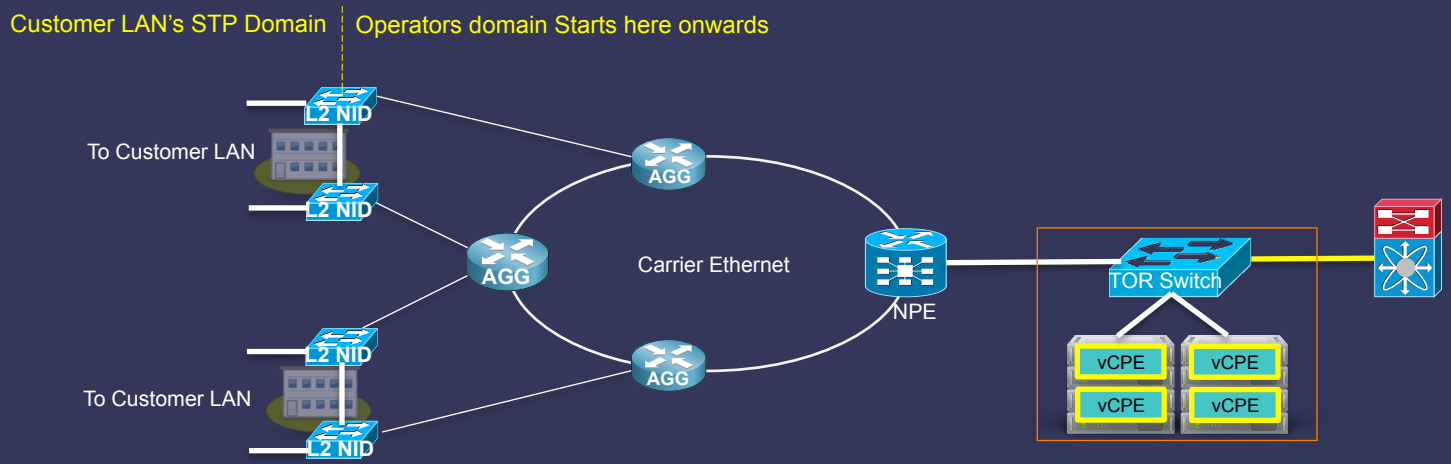
- NFV DC's are built with a network switching underlay → servers aren't directly connected to the NPE
- With layer 2 backhaul of traffic from customer branches, the NFV DC switching layer will learn all customer MAC addresses
- An example site with 4000 customer sites and 250 MAC address per site means 1 Million MACs
  - The switching underlay / TOR switches will now need to support and learn 1Million MAC addresses
  - Impacting cost of the network, service scale, convergence time upon failure due to large table size
- This can be technically solved with end-to-end overlay (like GRE or MPLS PW) from branch to vCPE or NPE to vCPE → defeating the original simplicity of the proposed architecture to a major extent

Security Exposure Due to Extension of the Broadcast Domain



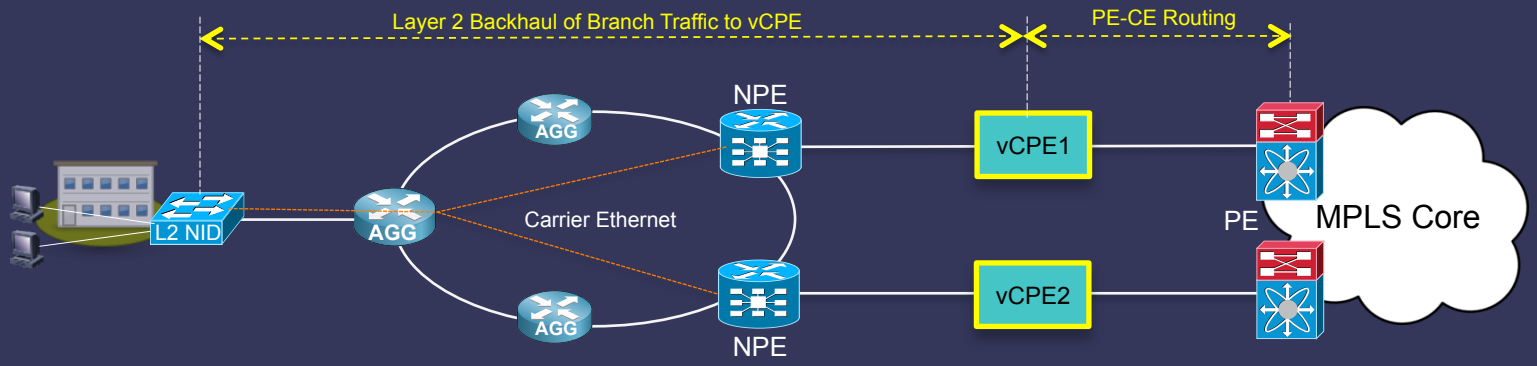
- By not having a L3 CPE at branch, and vCPE at NFV DC, it extends customer's Layer 2 domain all the way to the NFV DC
- For a POP with 4000 customers, it means extension of 4000 layer 2 domains hitting the NFV DC → SP typically has no control what assets are there at these branches and how secure they are
- This poses a significant security risk to SP's infrastructure for various DDoS/other attacks

Potential Risks Due to Layer 2 Loops



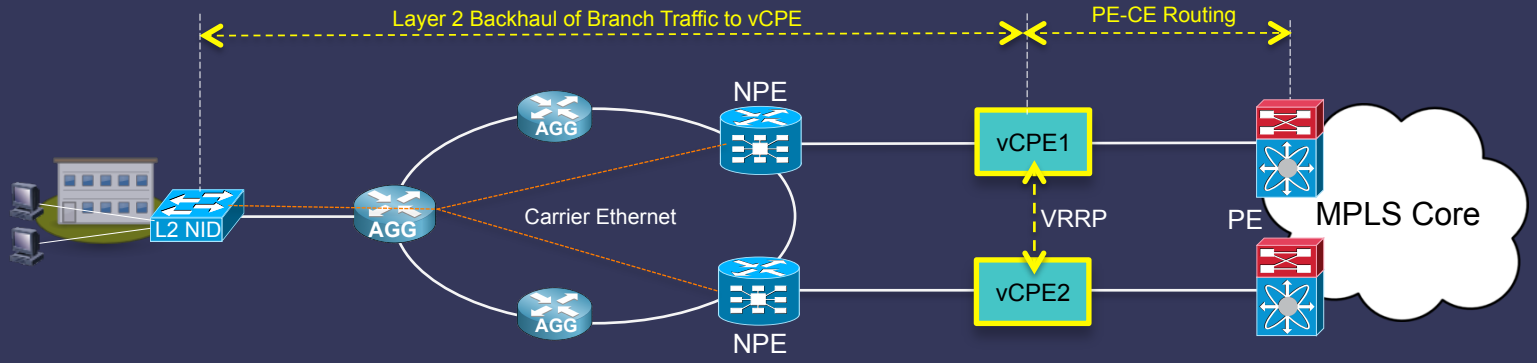
- In a L2 NID only based architecture, it is critical to demarcate customer's STP domains at the L2 NID
- There could be dual homing situations, where L2 NID may have to participate in Customer's Spanning Tree domain, also may require some form of loop prevention mechanism on the NNI side too
- Such dual homed connectivity requirement pose risks. Operational errors may cause the SP infrastructure to get impacted due to layer 2 loops originated from a customer branch

High Availability Design Challenges



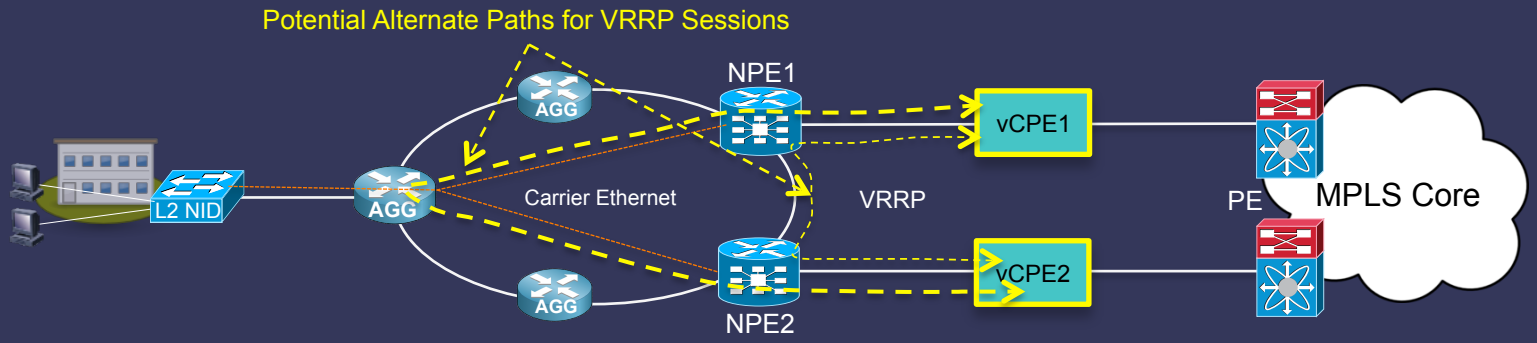
- For situations with two vCPE's for HA, the two vCPE's need to run HSRP / VRRP (lets consider VRRP)
- There are multiple ways to run the VRRP traffic between the two vCPE's that comes with different levels of complexity and different degree of reliability
- The "L2 NID ←→ vCPE" connectivity tracking becomes key for reliable bi-directional packet forwarding

High Availability Design Challenges
VRRP on Directly Connected Links



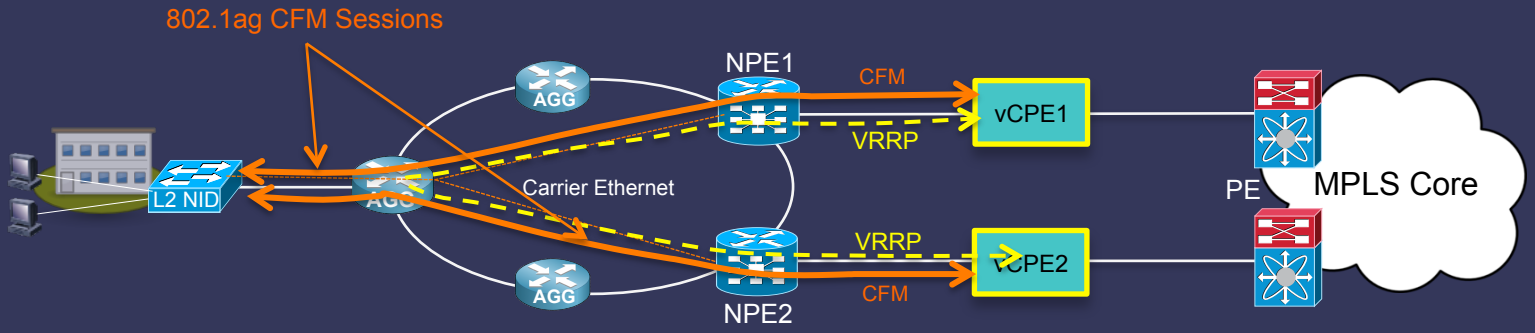
- Simplest way to run VRRP → but requires a L2 segment between two NFV DCs (typically two sites)
- Less reliable, since VRRP operation is blind to the connectivity failures from vCPE to L2 NID
- If vCPE1 is active on VRRP segment, and if vCPE1 to the L2 NID connectivity fails, vCPE1 may continue to remain active → will cause service outage

High Availability Design Challenges
VRRP via the Carrier Ethernet Network



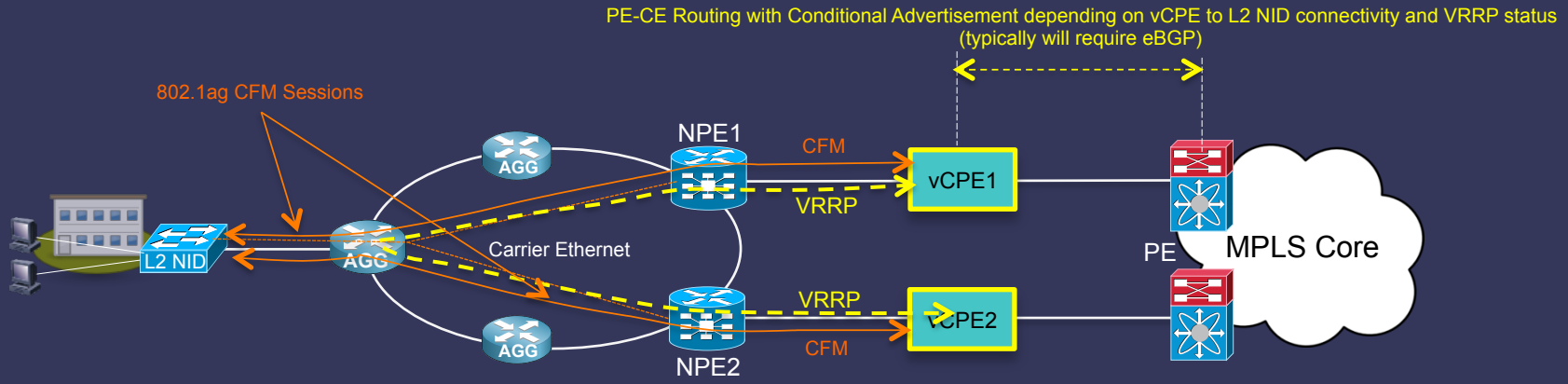
- To address the reliability issue, VRRP may be carried across the Carrier Ethernet network
  - vCPE1 – NPE1 – NPE2 – vCPE2 → need a L2 path, FRR enabled explicit path between NPE's to force the path
  - vCPE1 – NPE1 – AGG … – AGG – NPE2 – vCPE2
- More reliable now, since VRRP traffic will be dropped if the L2NID to vCPE connectivity fails, but –
  - Carrier Ethernet network to ensure VRRP packets aren't dropped during congestion – that will trigger false failover
  - VRRP delay timers may not be very aggressive, Carrier Ethernet network to ensure minimum delay
  - This is more complex to provision and operate

High Availability Design Challenges
VRRP via the Carrier Ethernet Network + IEEE 802.1ag (CFM)



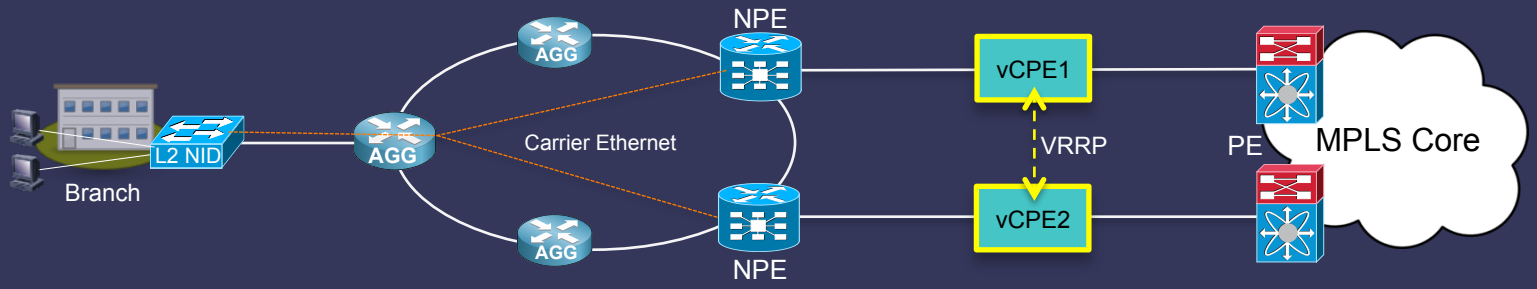
- The previous solution is still not end-to-end, not covering the AGG to the L2 NID connectivity
  - For an end-to-end reliable operations, that is a key requirement
- A way to address this challenge is to use VRRP and CFM (802.1ag) together
  - CFM runs end-to-end from L2NID to vCPE. When due to any failure on the path, CFM session expires → interface of vCPE goes to line protocol "down" state → VRRP traffic cannot go out any more out of the interface → VRRP switchover takes place to the standby
- Solves this HA issue, but brings back a lot of complexities in the network
  - We're trying to remove complexities by removing L3 CPE from branch → but we introduced different complexities now

High Availability Design Challenges
Upstream Routing from vCPE to the L3VPN PE



- The vCPE's need to run PE-CE routing with eBGP / IGP or Static routing
- The vCPE's now need to perform conditional route advertisement to the L3 VPN PE's depending on the reachability of vCPE to L2 NID and VRRP status
- If vCPE1 is the preferred path for the downstream traffic, but vCPE1 has lost connectivity to L2 NID, the vCPE1 needs to make L3VPN PE aware by advertising routes with less preferred attribute than vCPE2
- Typically restrict the PE-CE routing protocol to eBGP and may add more complexity in the design

Lack of L3 Capability @Branch Will Limit Available Services

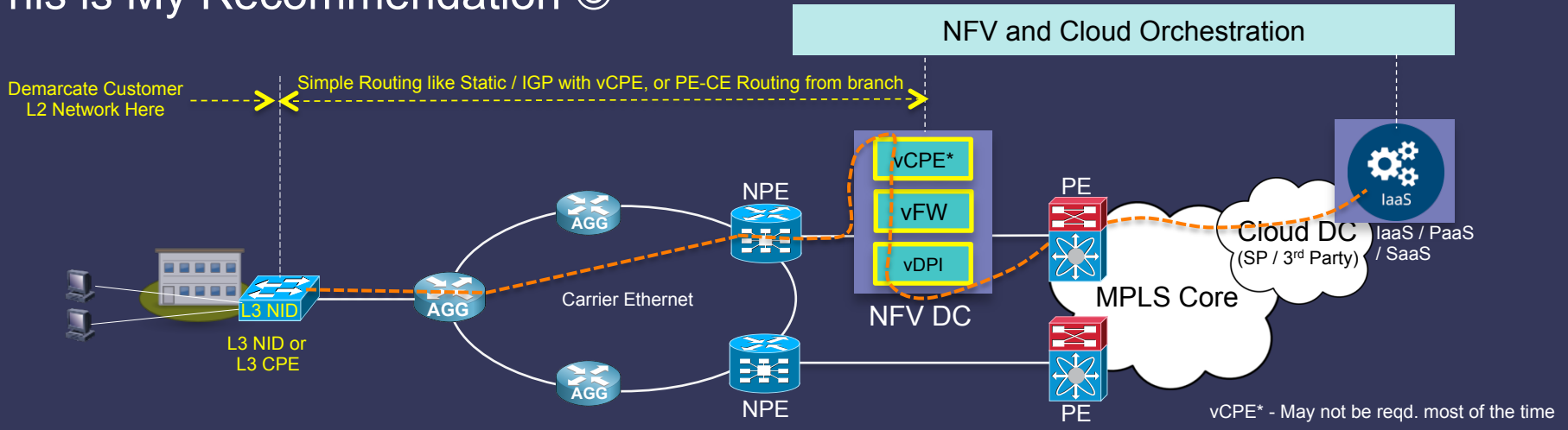


- Many customer may require capabilities at branch that requires Layer 3 devices
  - Such as IPSec VPN or WAN Acceleration
- Many Service Providers are looking forward to use 3G or 4G LTE as backup connectivity
  - Typical L2 NIDs do not have those interfaces → forcing another CPE for the backup
- If Hierarchical & granular QOS is a requirement at the branch, this could be challenging with L2 NID too

Conclusion of L2 NID + vCPE Architecture

- We were attempting to simplify the architecture by removing L3 CPE from the branch in the managed CPE / VPN architecture
- But in that process, complexities of other types got injected back into the network
  - MAC address scaling issue impacting service scale, convergence, cost
  - Security exposure of the vCPE / NFV DC and the SP infrastructure due to extension of L2 domains
  - Possible chances of Layer 2 loops due to operational errors
  - Complex design requirements to satisfy high availability → more difficult to operate
- It may create further limitations when it comes to service availability at the branch
  - Layer 3 services such as IPSec, WAN Acceleration etc. are not possible from the branch anymore
  - 3G / 4G LTE on the same device
  - Hierarchical and Granular QOS from the branch

So What We May Do To Approach the Problem ?
This is My Recommendation ☺



- Keep a L3 CPE at branch, may be physical or virtual → call it a L3 NID or L3 CPE or whatever you like
  - We need to demarcate customer's L2 network at the branch itself, that's how the networks scaled
  - This helps avoid MAC address scale, security & L2 Loop issues. Also avoids additional issues with VRRP design
  - Make the L3 CPE at branch Zero Touch Provisioning (ZTP) capable – to achieve automation and agility
- If required, try and make the L3 CPE at branch simplified by reducing the footprint of "enabled features"
  - Provision complex CPE features on NFV DC (may include advanced routing on a vCPE)
  - Have the ability to service chain vCPE with other rich set of functions using NFV orchestration system → make it agile!

Thank you.

NFV – How to build / Augment Operations skillsets

- Most existing technologies, protocols and associated skills are equally required
- On top of that, there are needs for acquisition of New Skills
  - x86 Server Virtualization
  - Virtualization on Linux (and KVM/QEMU) Environment
  - Cloud Orchestration Systems – such as OpenStack
  - Virtual Switches – OVS, Netmap/VALE, Snabbswitch, Vendor Specific etc
  - SDN Controllers – OpenDayLight, Vendor Specific
  - Device Programmability and APIs – NETCONF, Yang, RESTCONF, REST APIs, OF….
  - Service Function Chaining – specially NSH (Network Service Header)
  - Network based Virtual Overlay transport – VXLAN, MPLSoGRE/UDP, LISP, L2TPv3…..
  - Automation Tools – puppet / chef etc.
  - Management, Orchestration, OSS Fundamentals,
  - …..