Edge Fabric:

Steering Oceans of Content to the world

Robel Kitaba
Network Engineer, Facebook

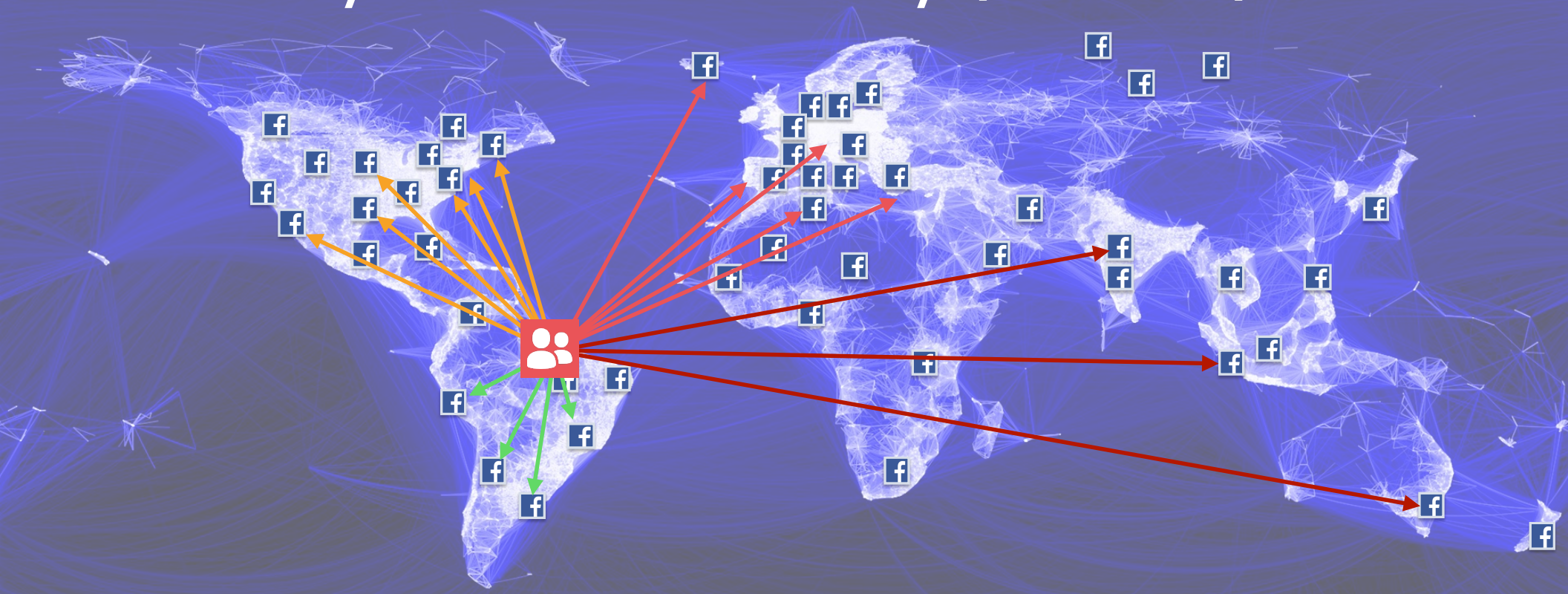
facebook INFRASTRUCTURE

Global Load Balancer

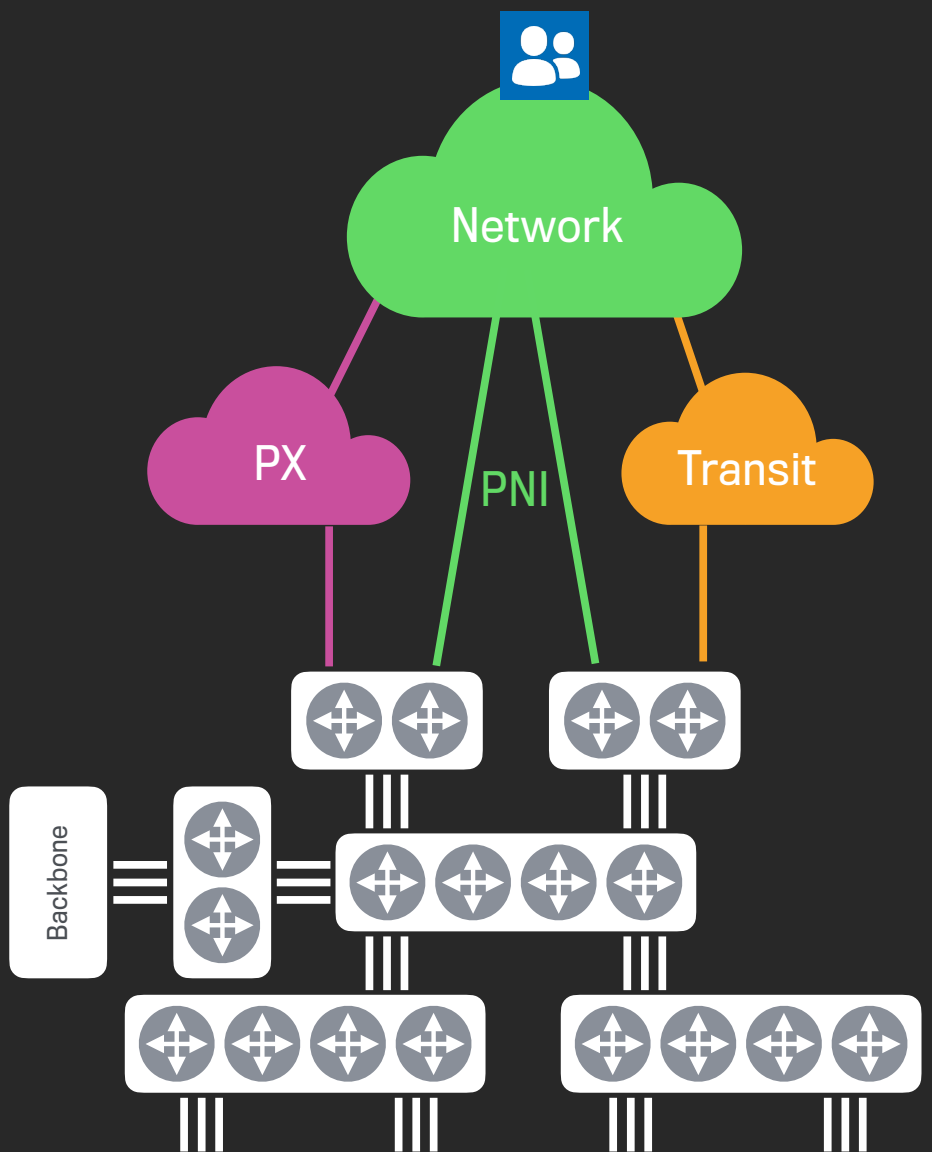
Manages ingress traffic

Locations just for visualization purposes, it does not reflect current configuration.

Latency based telemetry (SONAR)



Locations just for visualization purposes, it does not reflect current configuration.

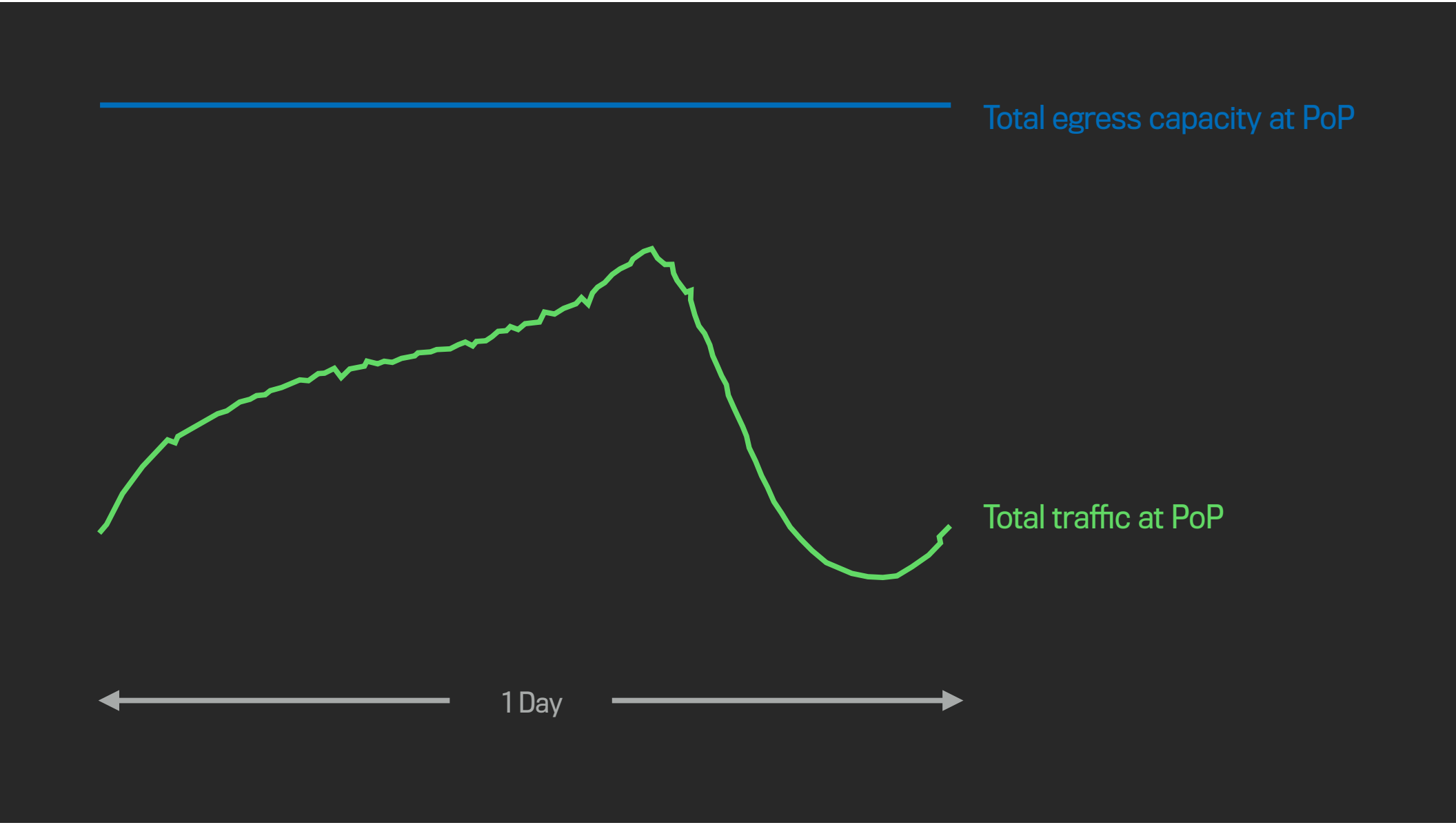


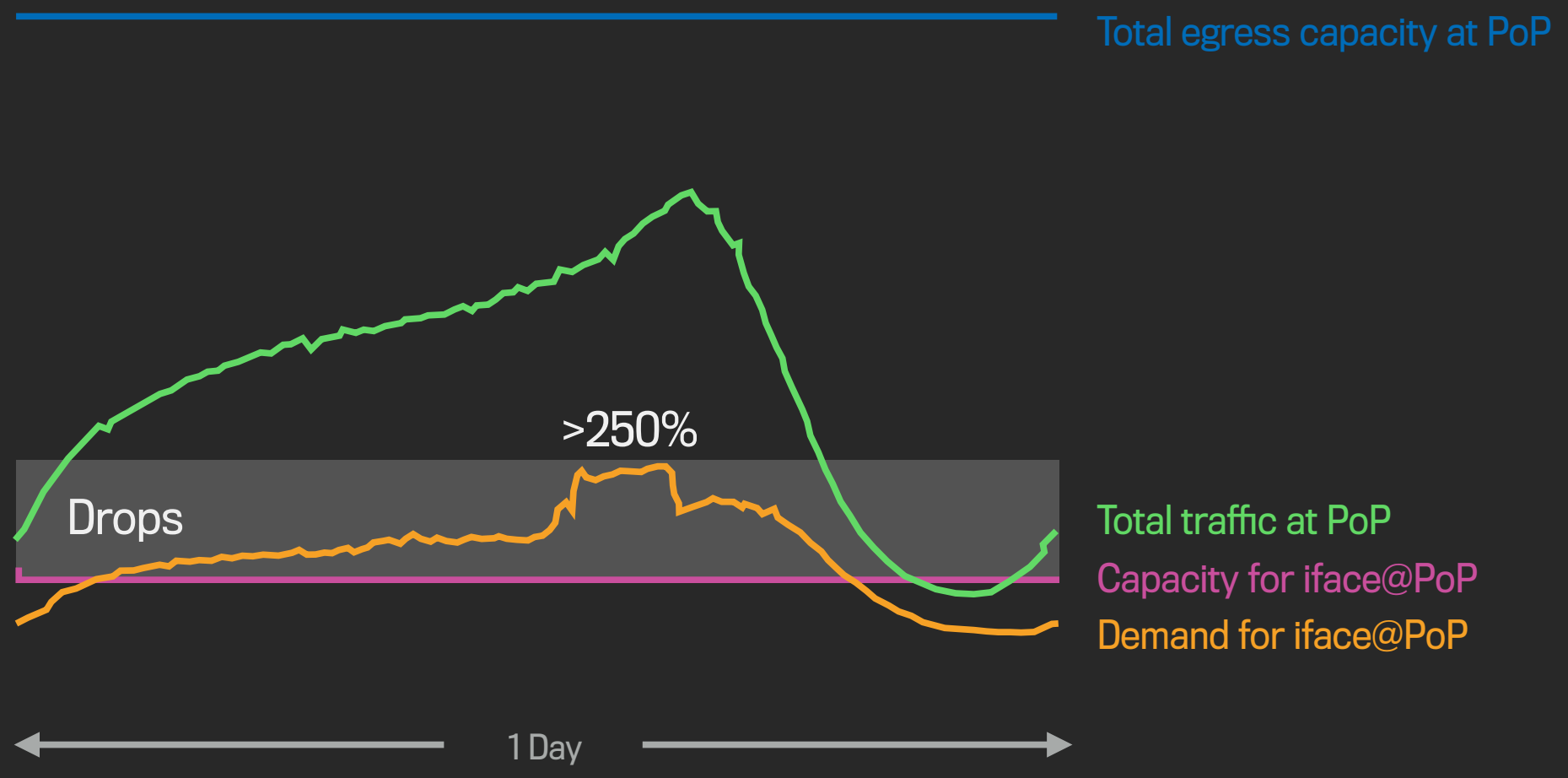
PoP: Point of Presence (colo facilities)

PNI Links: Direct peering with user networks

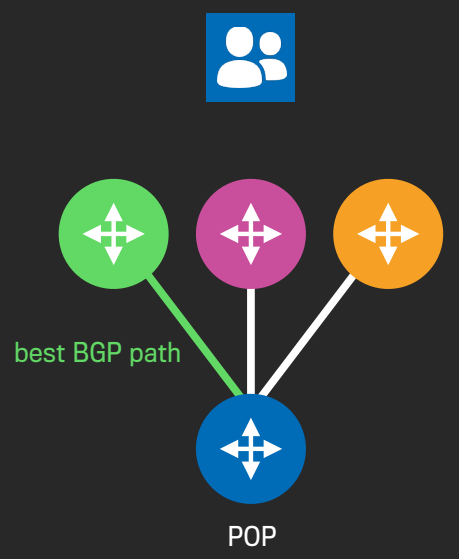
PX Links: Peering with networks over shared infrastructure

Transit Links: Peering with intermediate networks that provide global reachability


Total egress capacity at PoP
Total traffic at PoP
1 Day


Total egress capacity at PoP
>250%
Drops
Total traffic at PoP
Capacity for iface@PoP
Demand for iface@PoP
1 Day

Why demands exceeds capacity



Peering with other networks using BGP

BGP (STATIC)

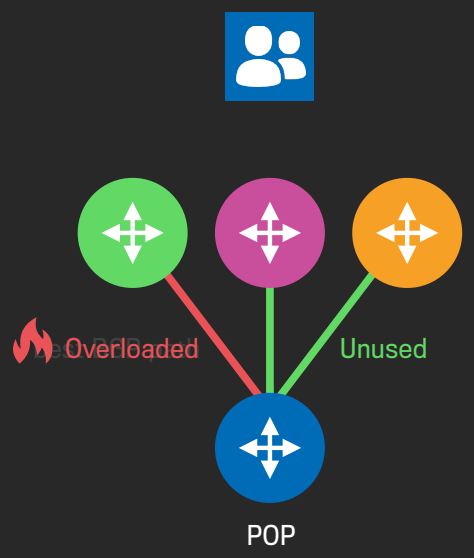
Local Preference

Med

AS Path length

Communities

Why demands exceeds capacity



Peering with other networks using BGP

BGP (STATIC)	REALITY (DYNAMIC)
Local Preference	Traffic demand changes
Med	Limited capacity
AS Path length	Performance variations
Communities	Transient failures

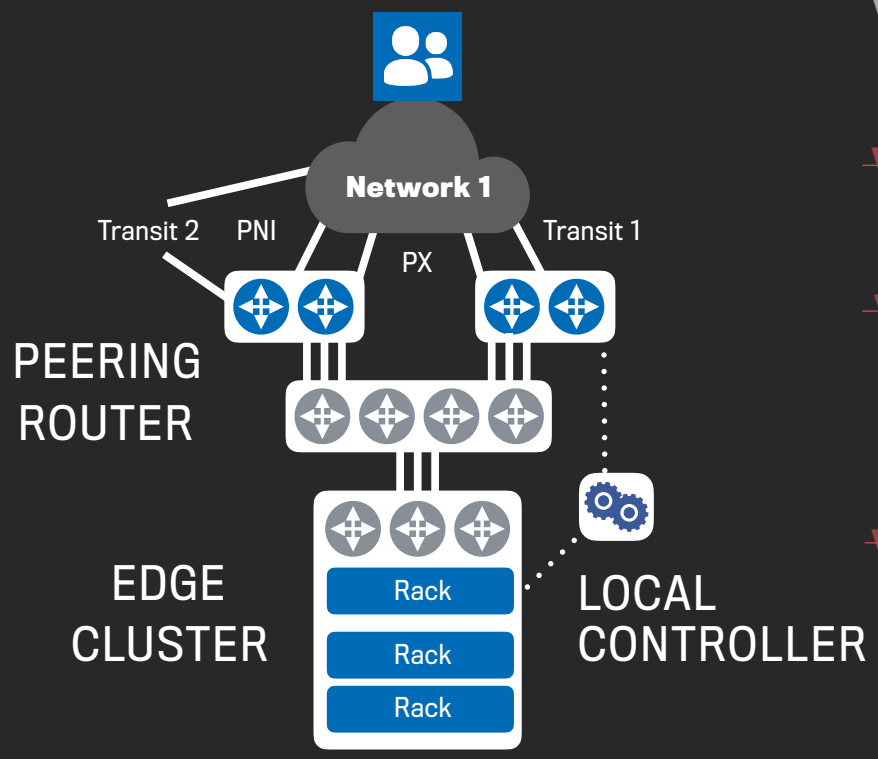
Local Edge Controller

Edge Fabric

"Engineering Egress with Edge Fabric: Steering Oceans of Content to the World",
Brandon Schlinker et al, SIGCOMM 2017

LOCAL CONTROLLER'S JOURNEY

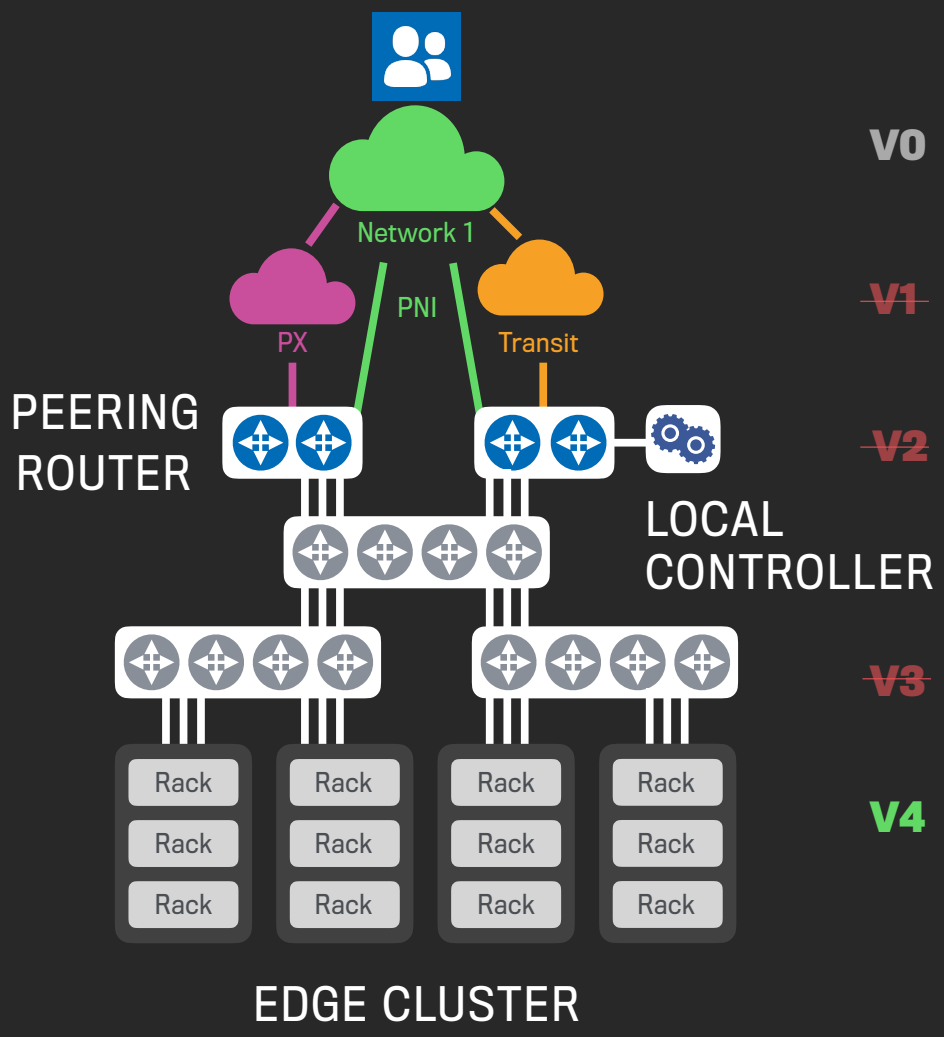
V0 ····· V1 ····· V2 ····· V3 ····· V4



V0	~~Manual interventions to change BGP policy when there were failures in PNIs~~	not scalable, too slow, error prone
V1	~~Setup MPLS paths from end hosts to PRs in order to choose egress links~~	Restrictions on hw
V2	~~Use DSCP marking at the end hosts to indicate egress link~~	Not scalable, coordination of config, rigid assumptions
V3	~~Use GRE tunnels from end hosts to PRs~~	Coordination of config, vendor bug

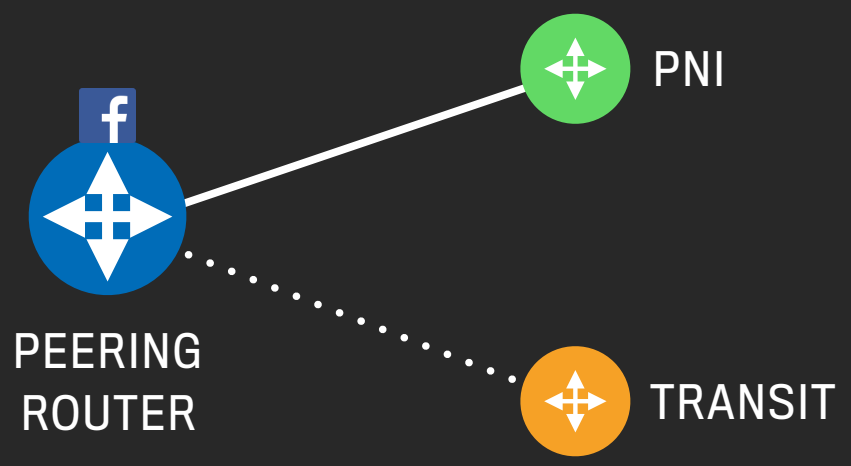
LOCAL CONTROLLER'S JOURNEY

V0 ····· V1 ····· V2 ····· V3 ····· V4



V0	~~Manual interventions to change BGP policy when there were failures in PNIs~~	not scalable, too slow, error prone
V1	~~Setup MPLS paths from end hosts to PRs in order to choose egress links~~	Restrictions on hw
V2	~~Use DSCP marking at the end hosts to indicate egress link~~	Not scalable, coordination of config, rigid assumptions
V3	~~Use GRE tunnels from end hosts to PRs~~	Coordination of config, vendor bug
V4	Use BGP injections at PRs	Flexible, dynamic, decouples decisions from PoP architecture

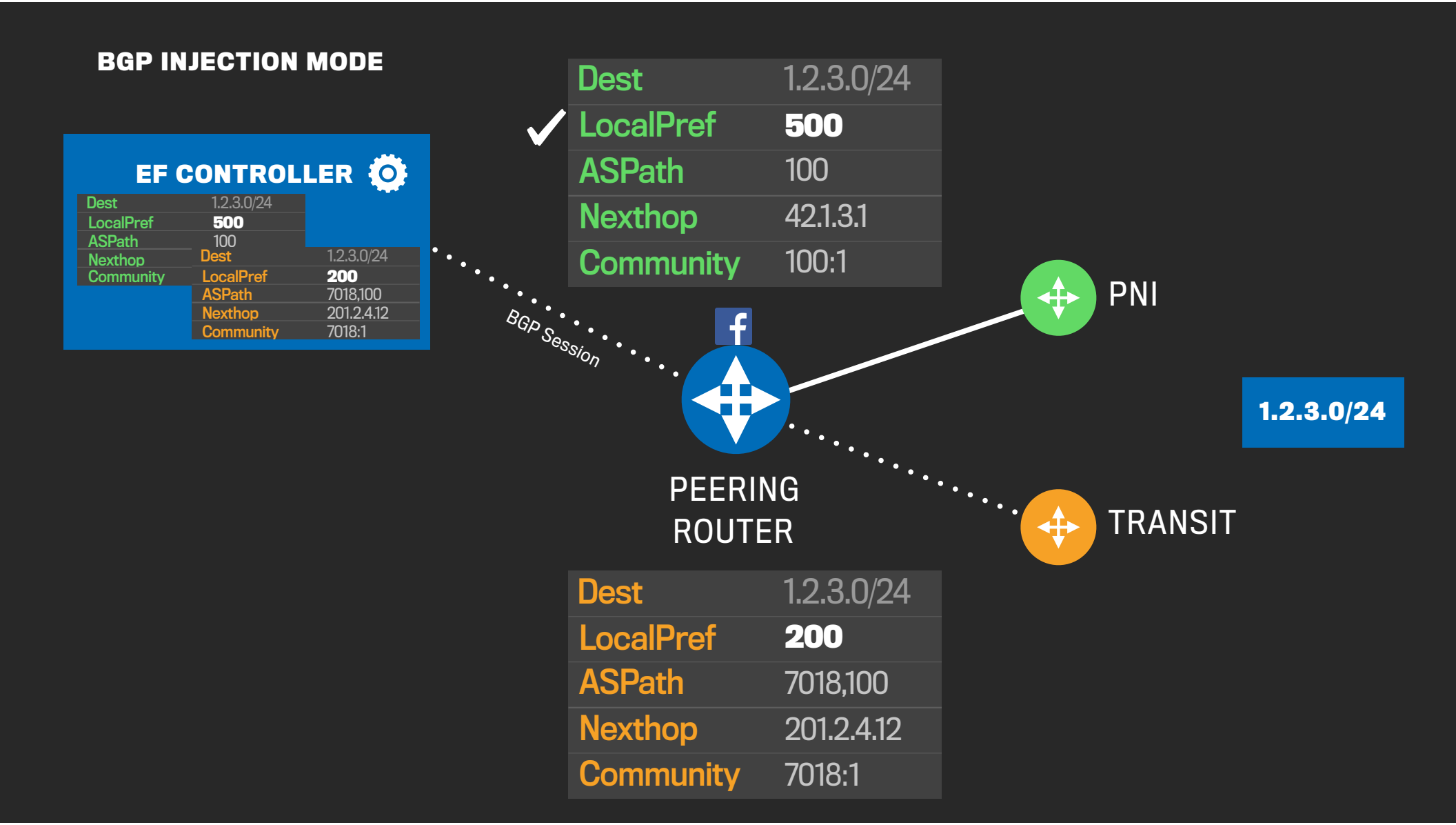
BGP INJECTION MODE

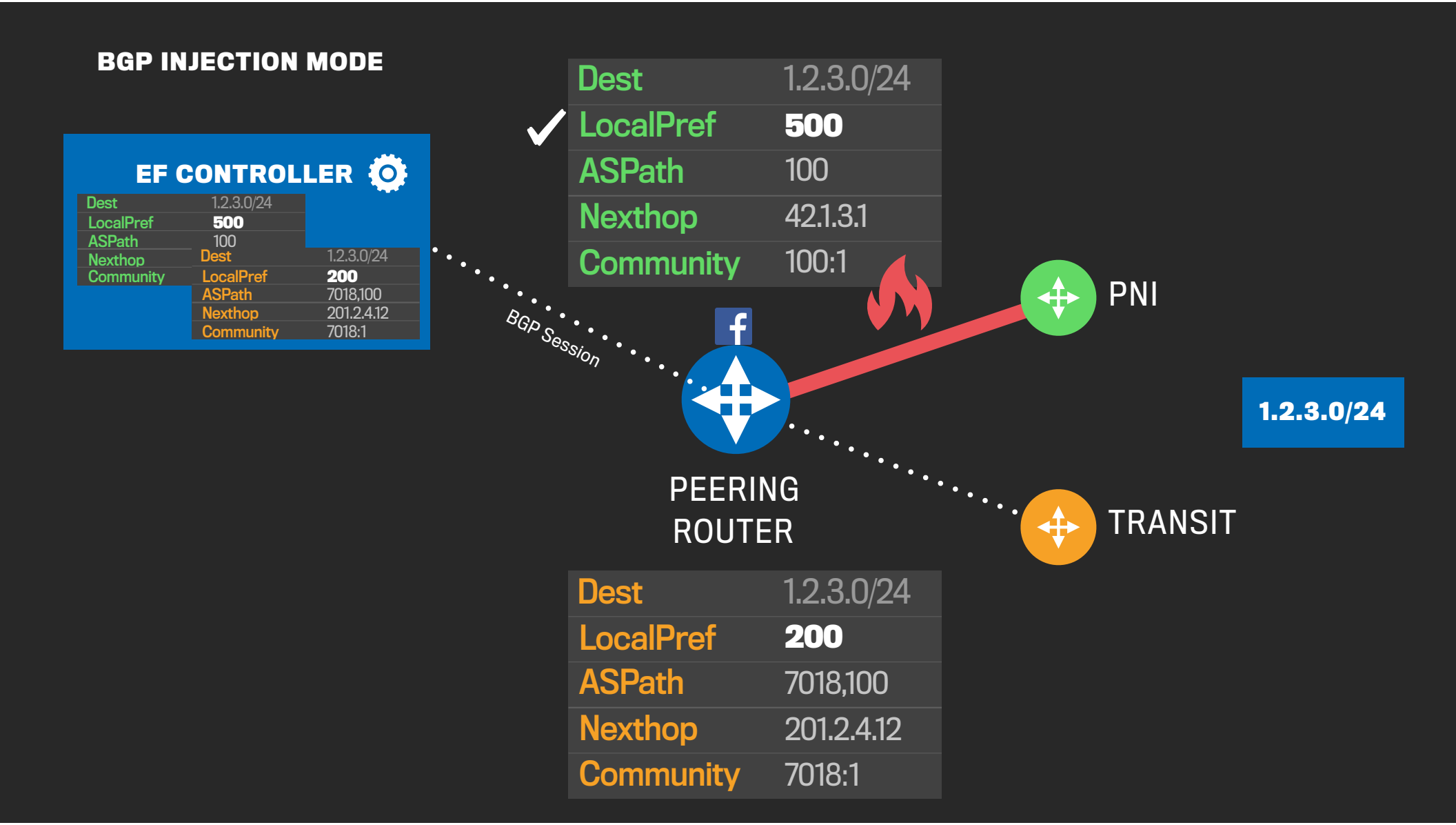


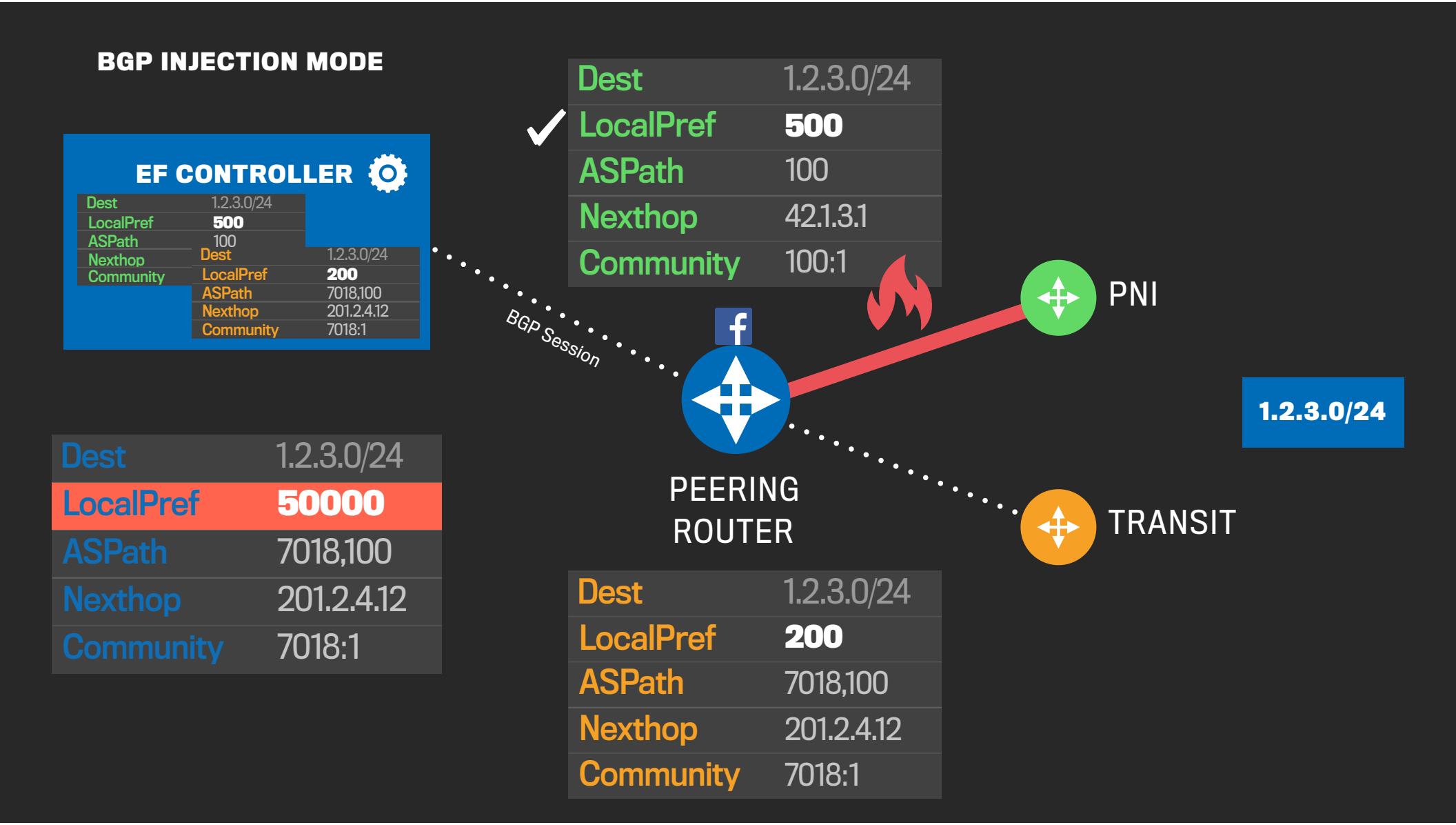
Dest	1.2.3.0/24
LocalPref	**500**
ASPath	100
Nexthop	42.1.3.1
Community	100:1

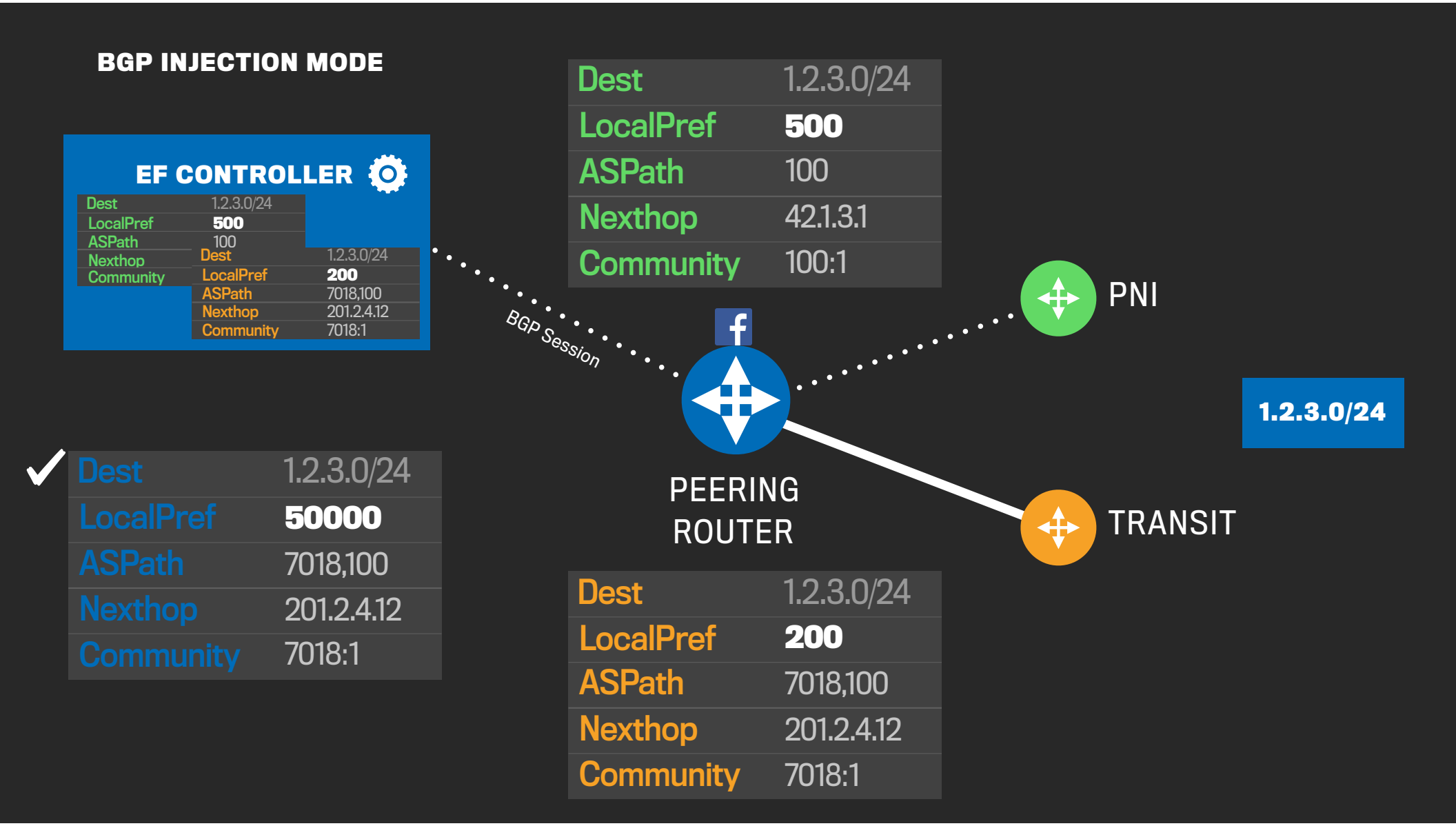
1.2.3.0/24

Dest	1.2.3.0/24
LocalPref	**200**
ASPath	7018,100
Nexthop	201.2.4.12
Community	7018:1

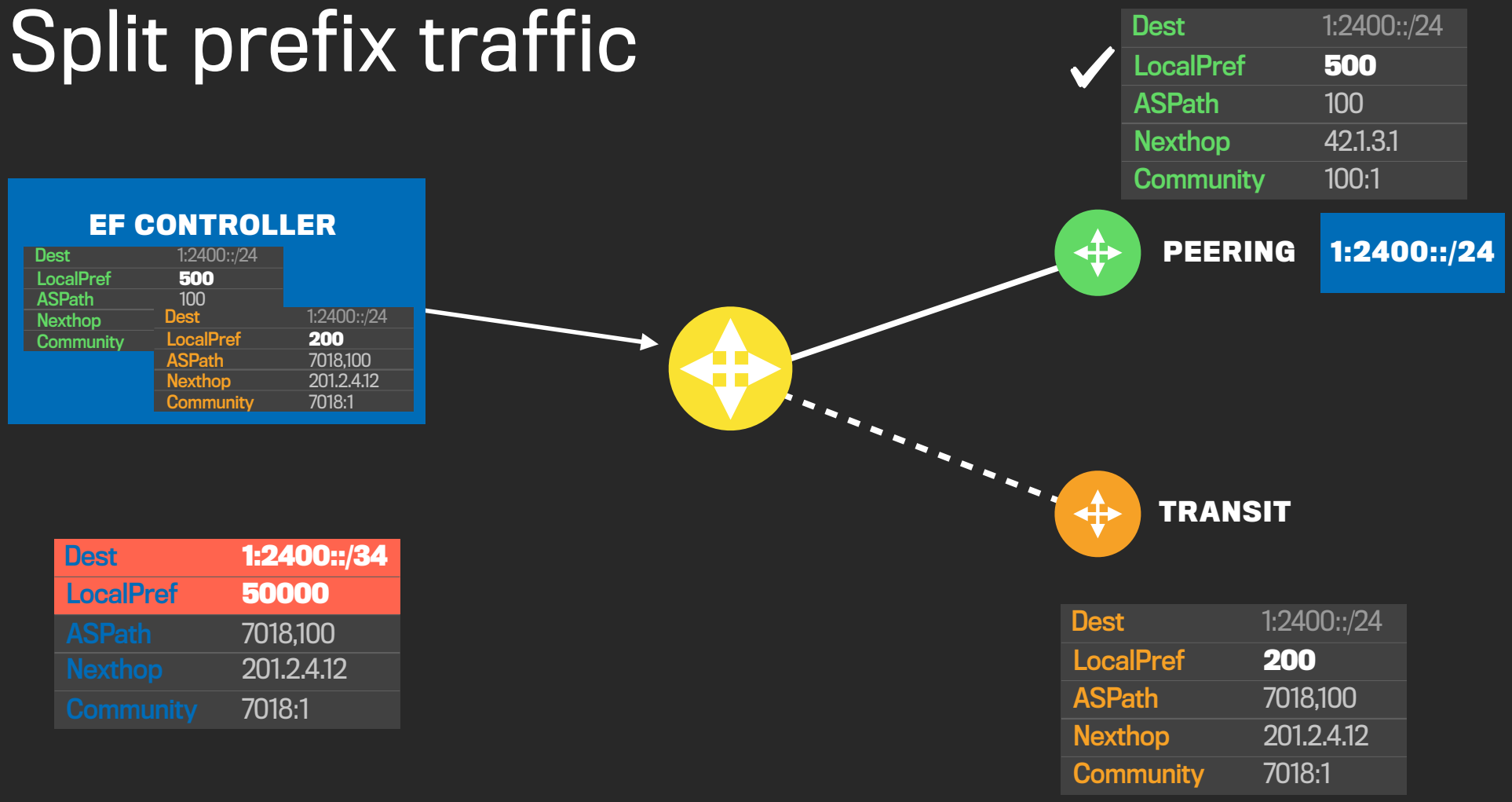

BGP INJECTION MODE
EF CONTROLLER
Dest 1.2.3.0/24
LocalPref 500
ASPath 100
Nexthop
Community
Dest 1.2.3.0/24
LocalPref 200
ASPath 7018,100
Nexthop 201.2.4.12
Community 7018:1
BGP Session
Dest 1.2.3.0/24
LocalPref 500
ASPath 100
Nexthop 42.1.3.1
Community 100:1
PNI
1.2.3.0/24
PEERING ROUTER
TRANSIT
Dest 1.2.3.0/24
LocalPref 200
ASPath 7018,100
Nexthop 201.2.4.12
Community 7018:1


BGP INJECTION MODE
EF CONTROLLER
Dest 1.2.3.0/24
LocalPref 500
ASPath 100
Nexthop
Community
Dest 1.2.3.0/24
LocalPref 200
ASPath 7018,100
Nexthop 201.2.4.12
Community 7018:1
BGP Session
Dest 1.2.3.0/24
LocalPref 500
ASPath 100
Nexthop 42.1.3.1
Community 100:1
PNI
1.2.3.0/24
PEERING ROUTER
TRANSIT
Dest 1.2.3.0/24
LocalPref 200
ASPath 7018,100
Nexthop 201.2.4.12
Community 7018:1

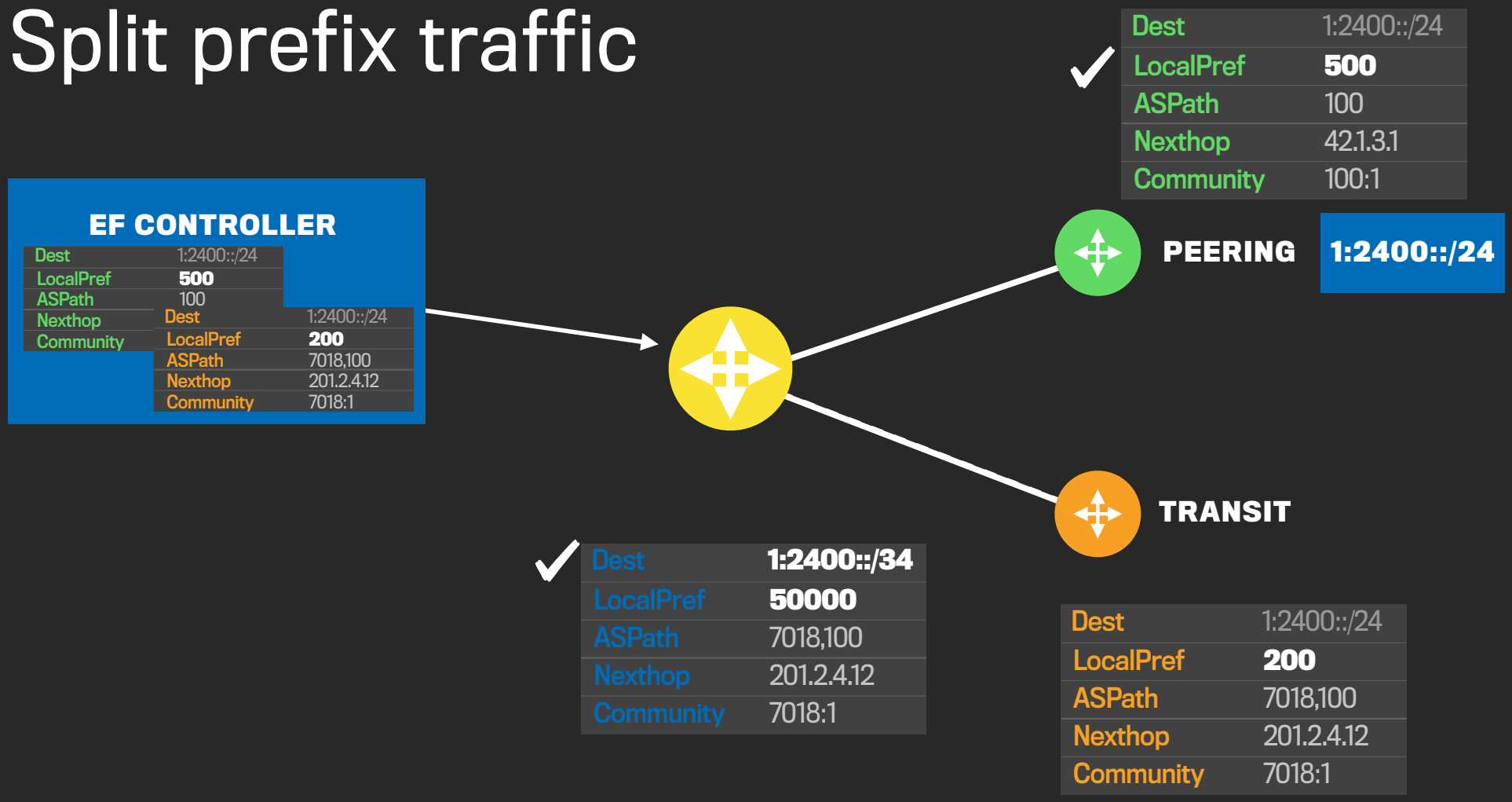

BGP INJECTION MODE
EF CONTROLLER
Dest 1.2.3.0/24
LocalPref 500
ASPath 100
Nexthop
Community
Dest 1.2.3.0/24
LocalPref 200
ASPath 7018,100
Nexthop 201.2.4.12
Community 7018:1
BGP Session
Dest 1.2.3.0/24
LocalPref 500
ASPath 100
Nexthop 42.1.3.1
Community 100:1
PNI
1.2.3.0/24
PEERING ROUTER
TRANSIT
Dest 1.2.3.0/24
LocalPref 50000
ASPath 7018,100
Nexthop 201.2.4.12
Community 7018:1
Dest 1.2.3.0/24
LocalPref 200
ASPath 7018,100
Nexthop 201.2.4.12
Community 7018:1

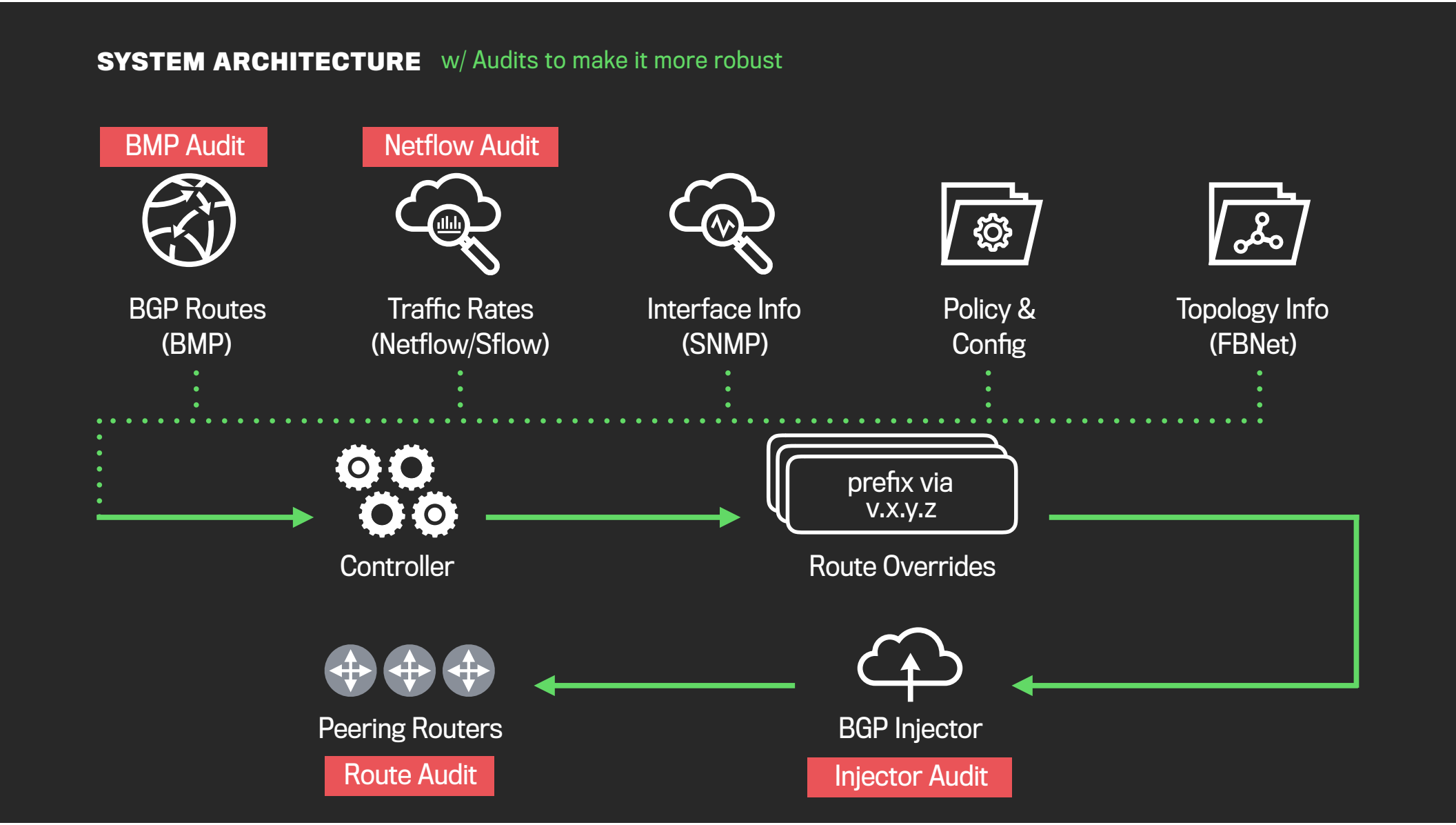

BGP INJECTION MODE
EF CONTROLLER
Dest 1.2.3.0/24
LocalPref 500
ASPath 100
Nexthop
Community
Dest 1.2.3.0/24
LocalPref 200
ASPath 7018,100
Nexthop 201.2.4.12
Community 7018:1
BGP Session
Dest 1.2.3.0/24
LocalPref 500
ASPath 100
Nexthop 42.1.3.1
Community 100:1
PNI
1.2.3.0/24
PEERING ROUTER
TRANSIT
Dest 1.2.3.0/24
LocalPref 50000
ASPath 7018,100
Nexthop 201.2.4.12
Community 7018:1
Dest 1.2.3.0/24
LocalPref 200
ASPath 7018,100
Nexthop 201.2.4.12
Community 7018:1

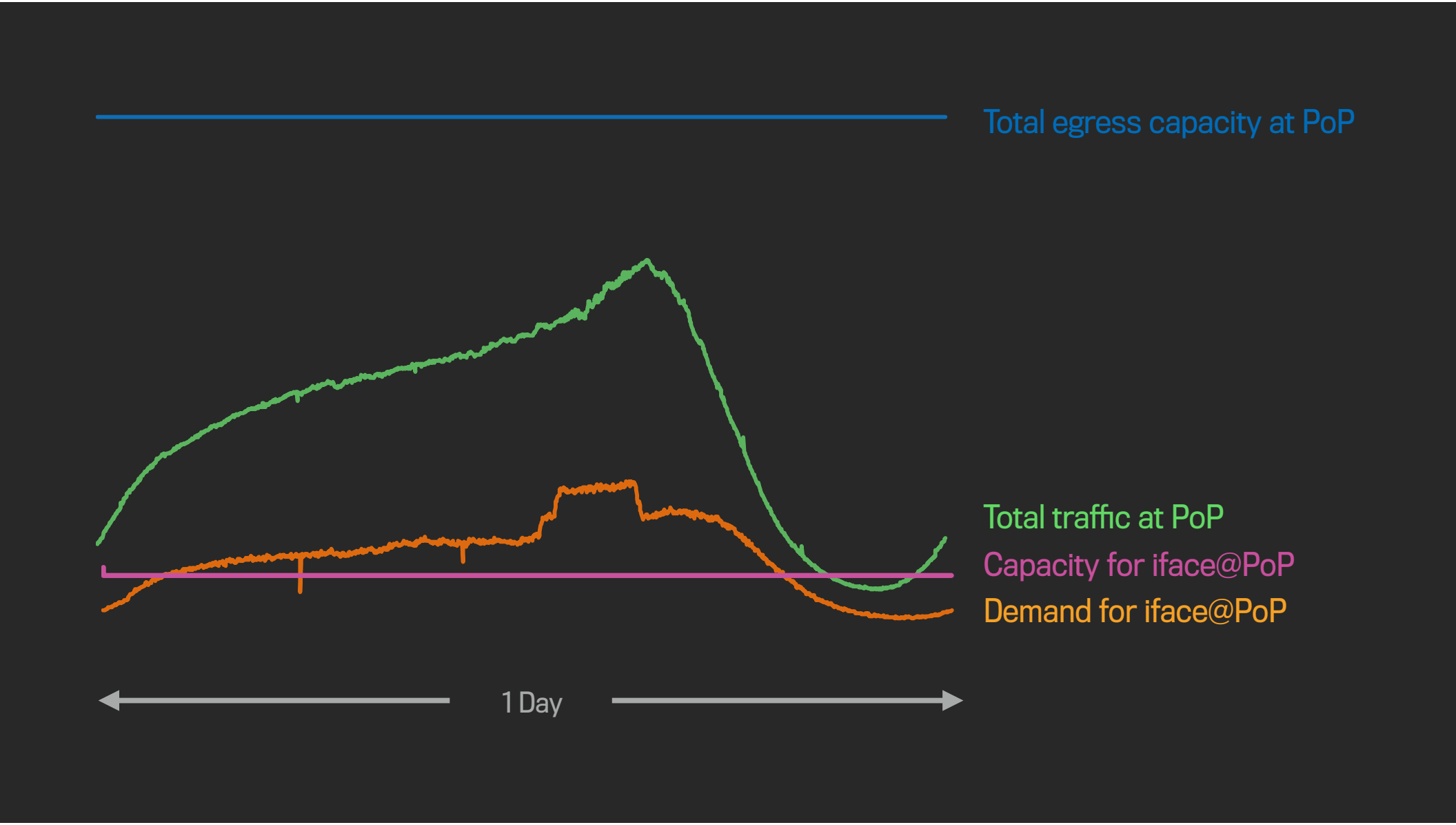
Split prefix traffic

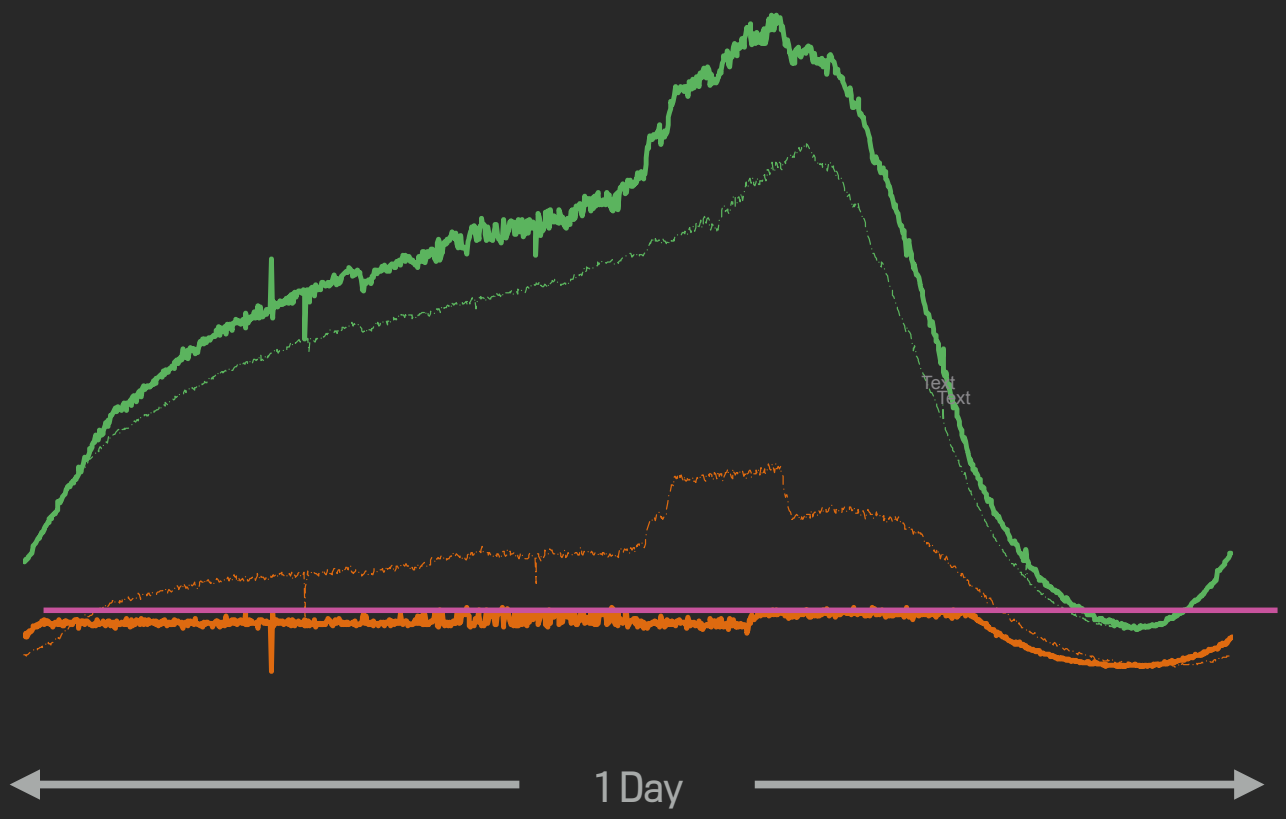


Split prefix traffic

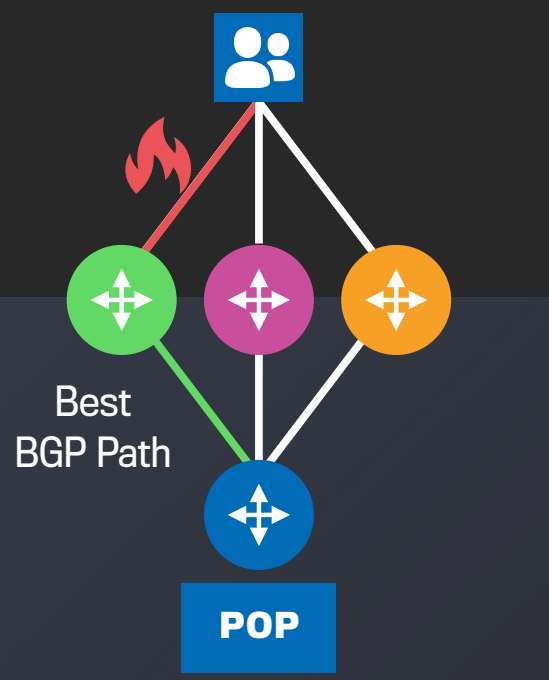



SYSTEM ARCHITECTURE
w/ Audits to make it more robust
BMP Audit
Netflow Audit
BGP Routes (BMP)
Traffic Rates (Netflow/Sflow)
Interface Info (SNMP)
Policy & Config
Topology Info (FBNet)
Controller
prefix via v.x.y.z
Route Overrides
Peering Routers
Route Audit
BGP Injector
Injector Audit


Total egress capacity at PoP
Total traffic at PoP
Capacity for iface@PoP
Demand for iface@PoP
1 Day


Total egress capacity at PoP
Avoid packet drops while maintaining high link utilization
Total traffic at PoP
Capacity for iface@PoP
Demand for iface@PoP
Traffic on iface@PoP w/Edge Fabric
1 Day

Looking beyond Facebook's network



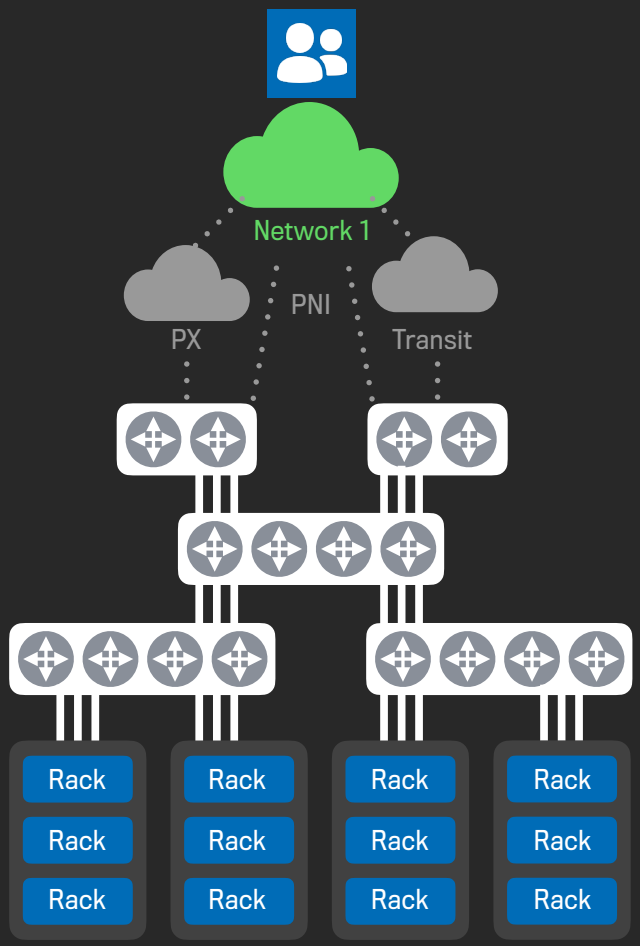
Facebook's Network

?

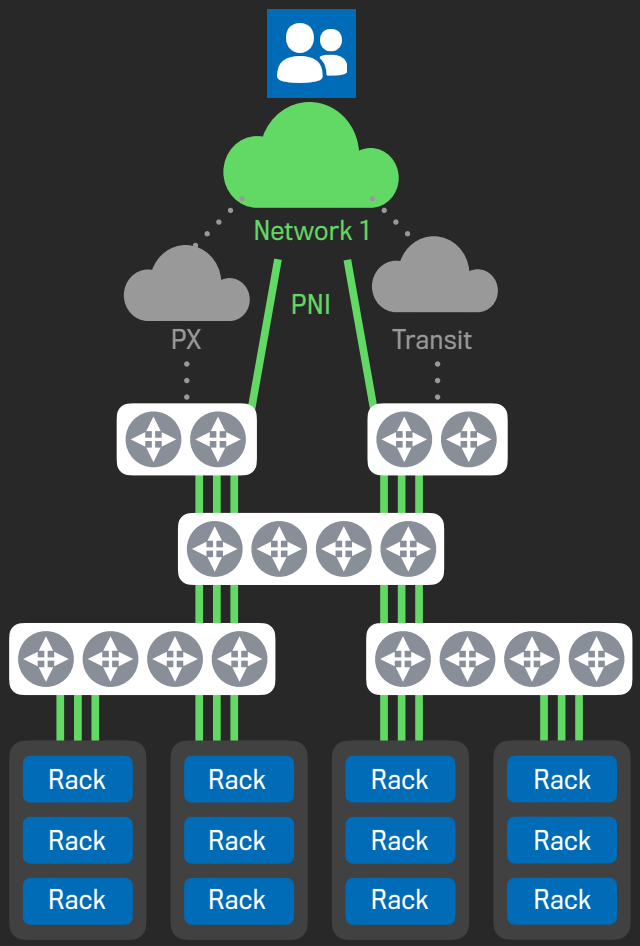
BGP (STATIC)	REALITY (DYNAMIC)
Local Preference	Traffic demand changes
Med	Limited capacity
AS Path length	Performance variations
Communities	Transient failures

Performance Routing

Alternative Path Measurements

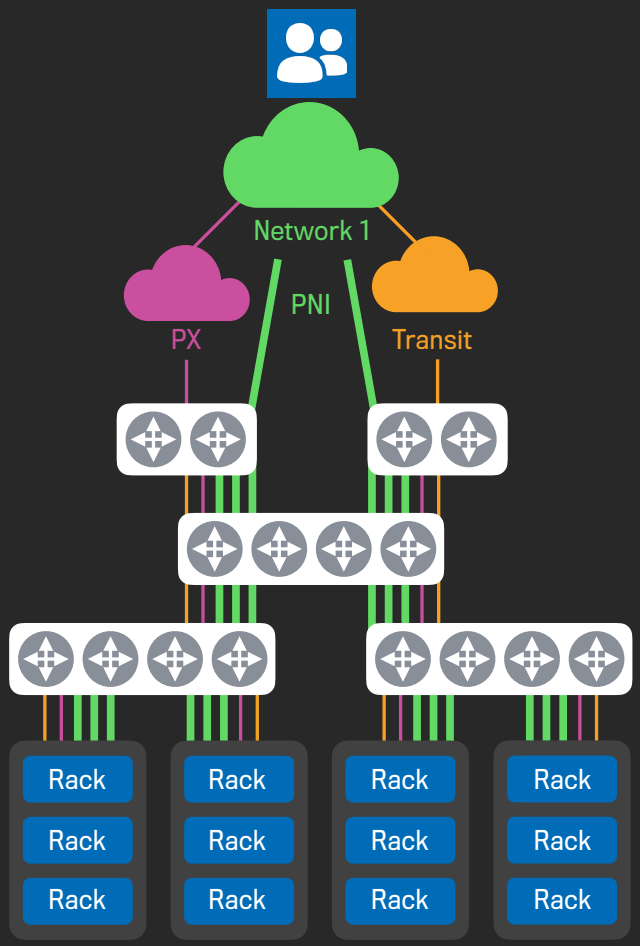


Collect TCP stats for transactions
(RTT, packet loss, throughput)



Allow us to monitor performance only to the primary path

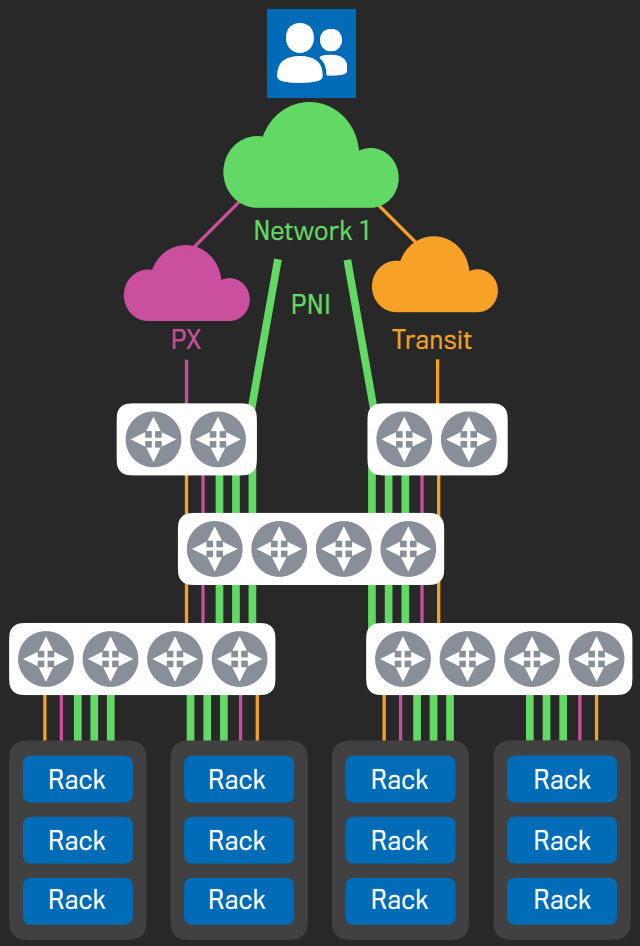
Collect TCP stats for transactions (RTT, packet loss, throughput)



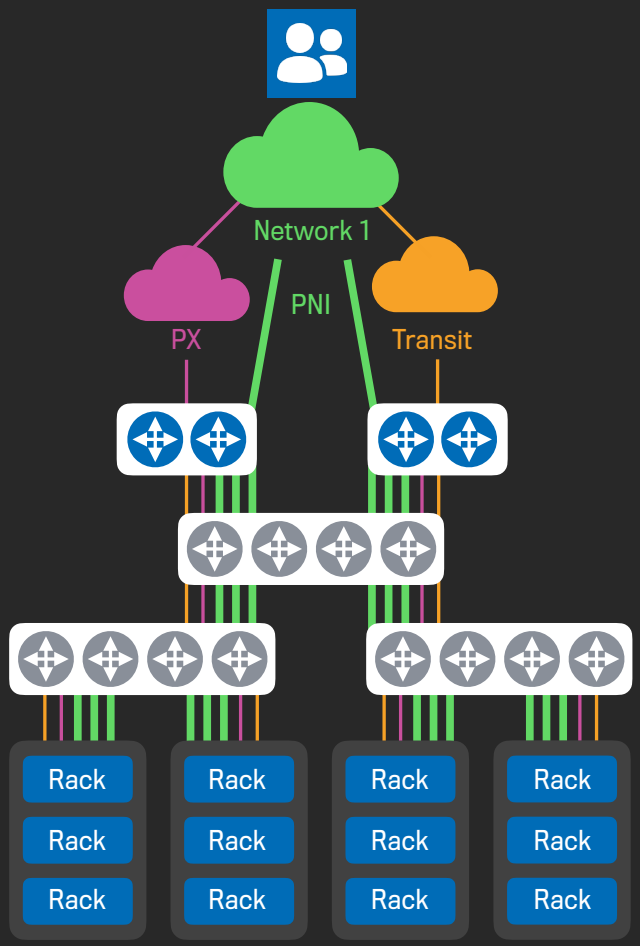
Allow us to monitor performance only to the primary path

Send a very small portion of traffic over alternate paths

Collect TCP stats for transactions (RTT, packet loss, throughput)

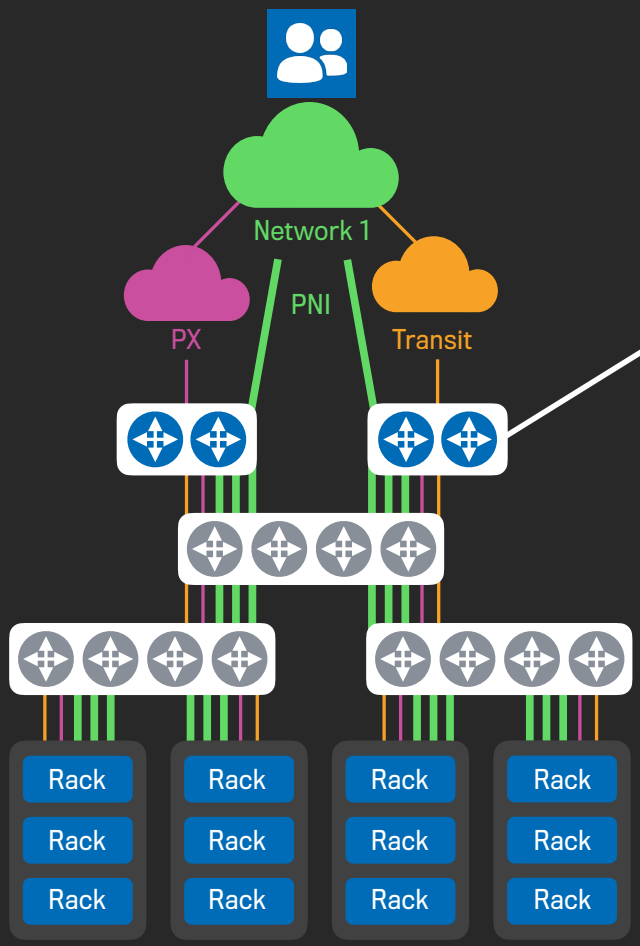


Mark random flows with special DSCP values



Configure alternate routing tables per DSCP value

Mark random flows with special DSCP values



APM CONTROLLER

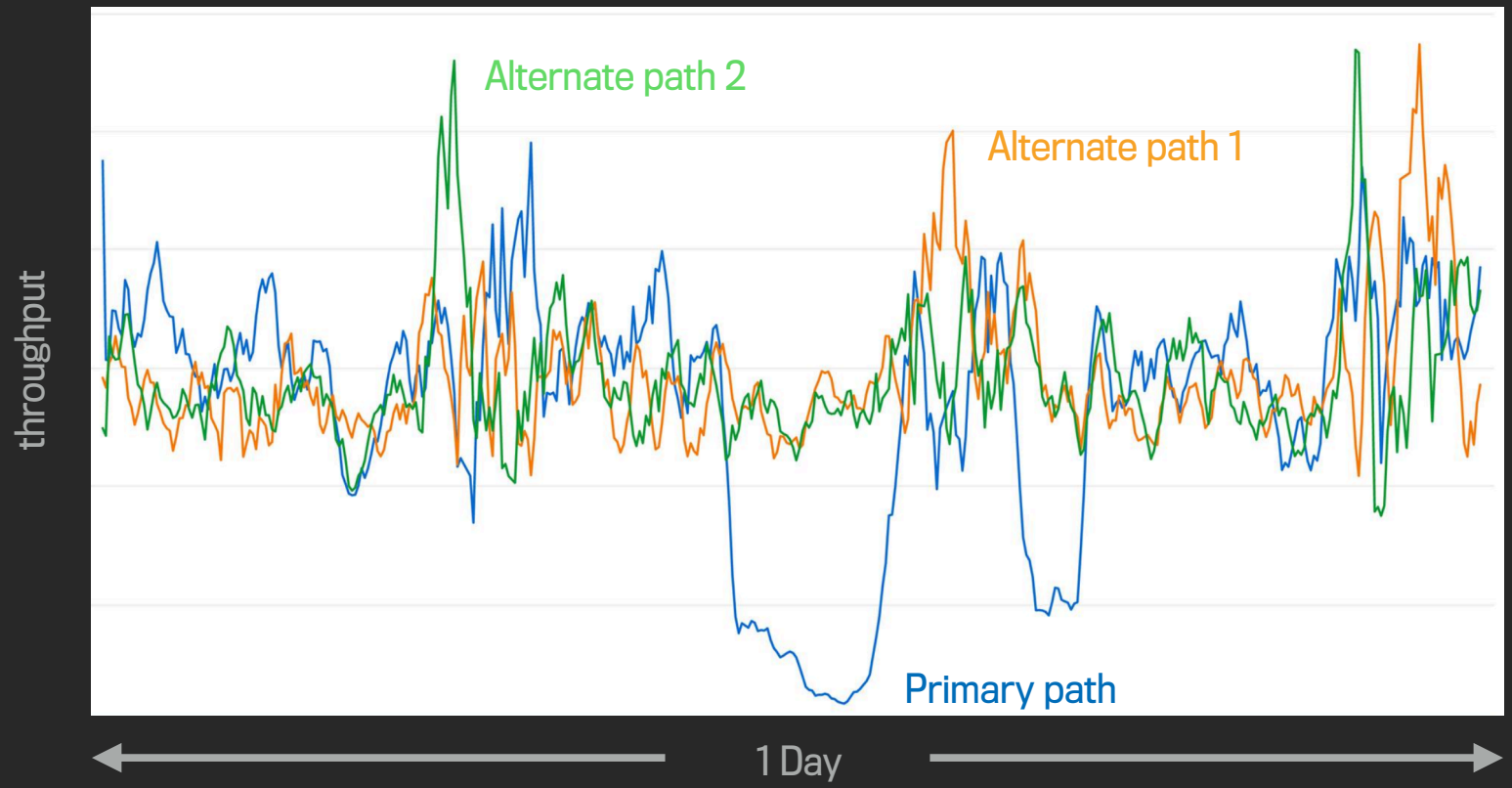
Insert routes into the alternate routing tables

Configure alternate routing tables per DSCP value

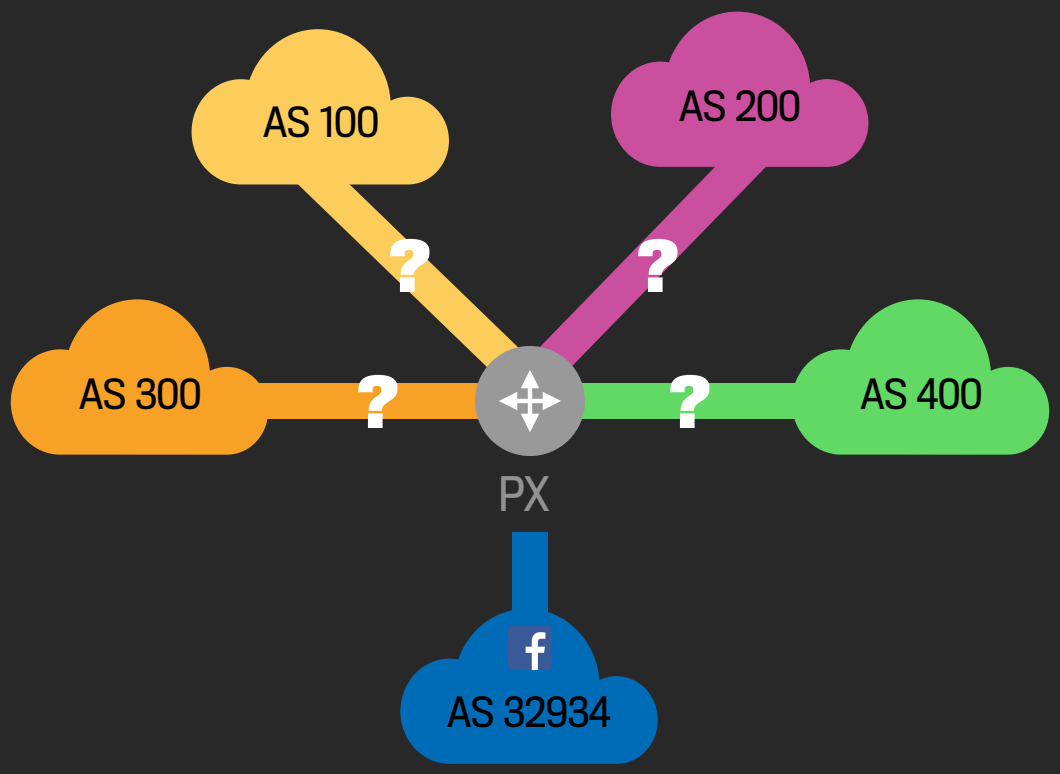
Mark random flows with special DSCP values

Interesting Examples

Temporary congestion of the primary path

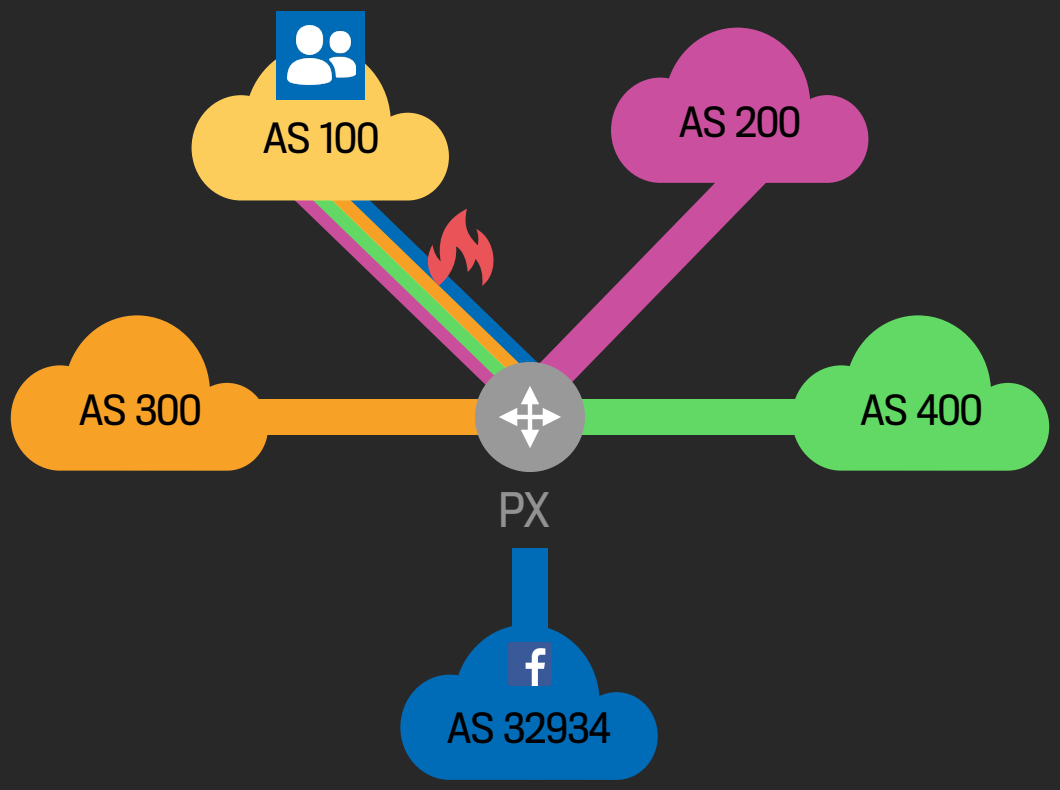


Public Exchange Performance problem



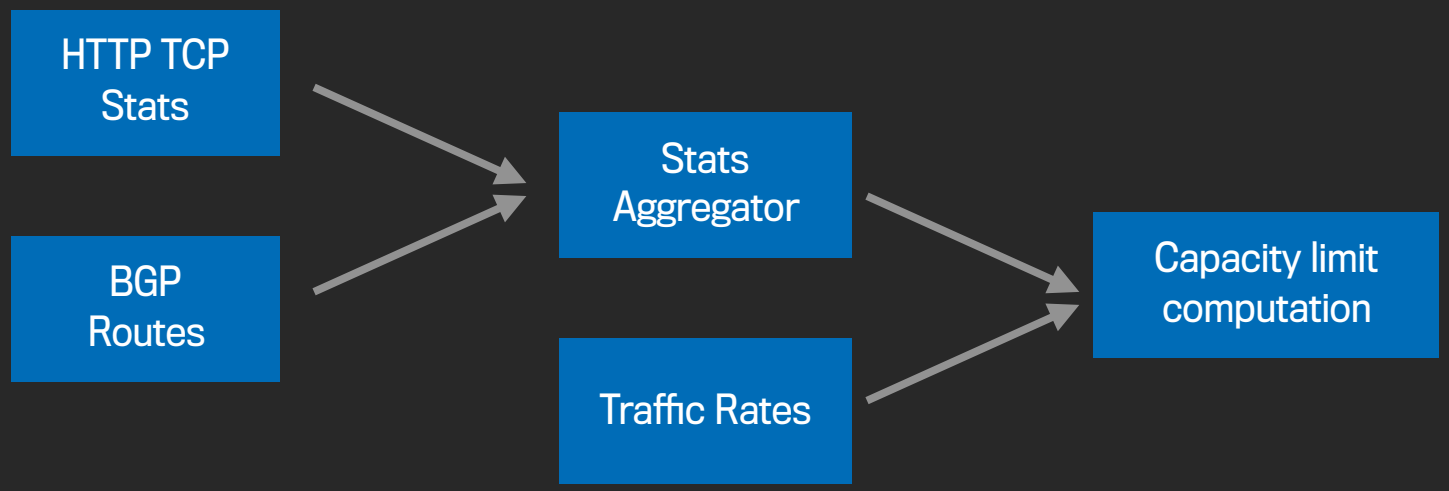
Peer's capacity is unknown

Public Exchange Performance problem



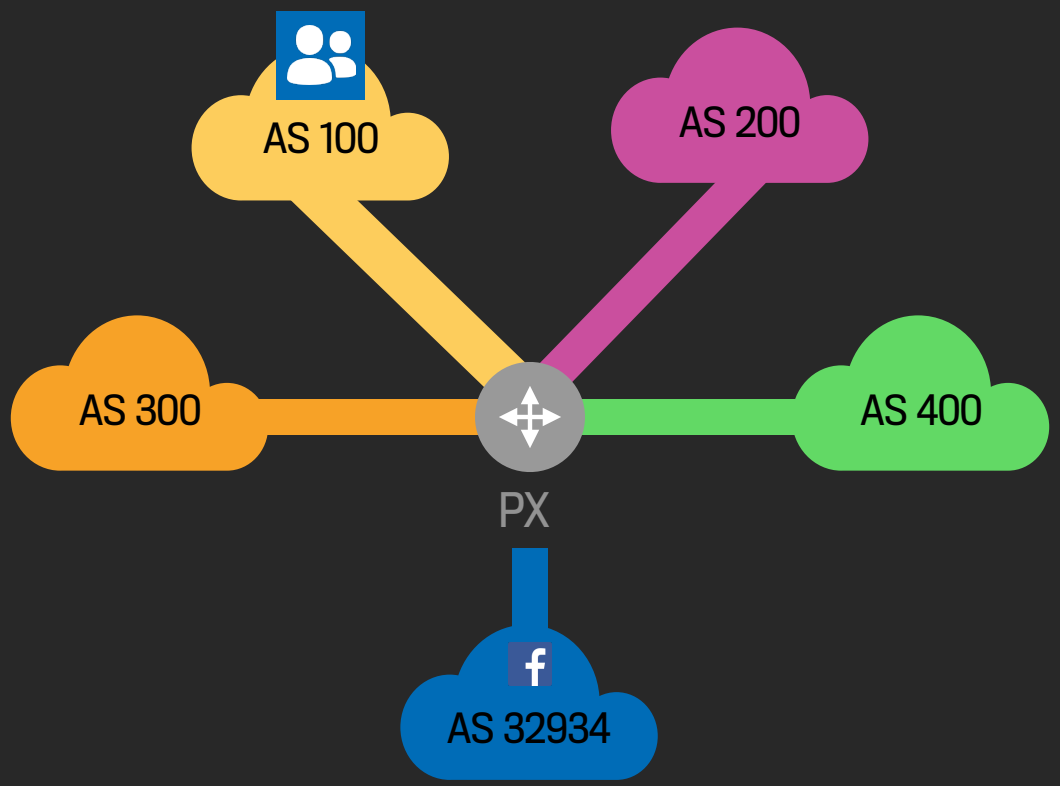
Peer's capacity is unknown

Path Performance Monitoring Service



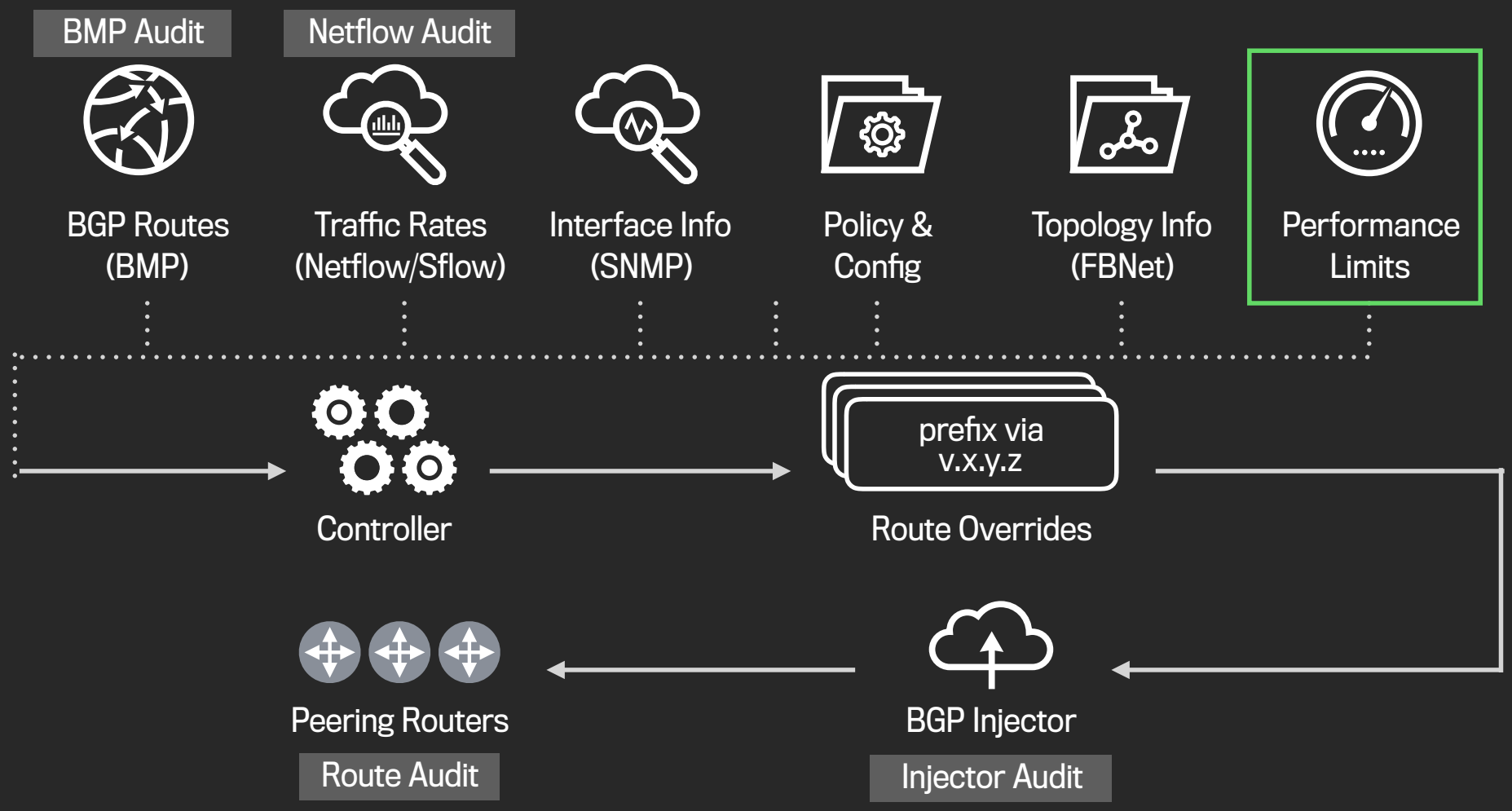
Computes effective Peer's capacity on PX

Public Exchange Performance problem



Infer how much traffic to send without overwhelming the peer

ENHANCE EDGE FABRIC W/ PERFORMANCE



Thanks