NEXT GEN CORE NETWORKS

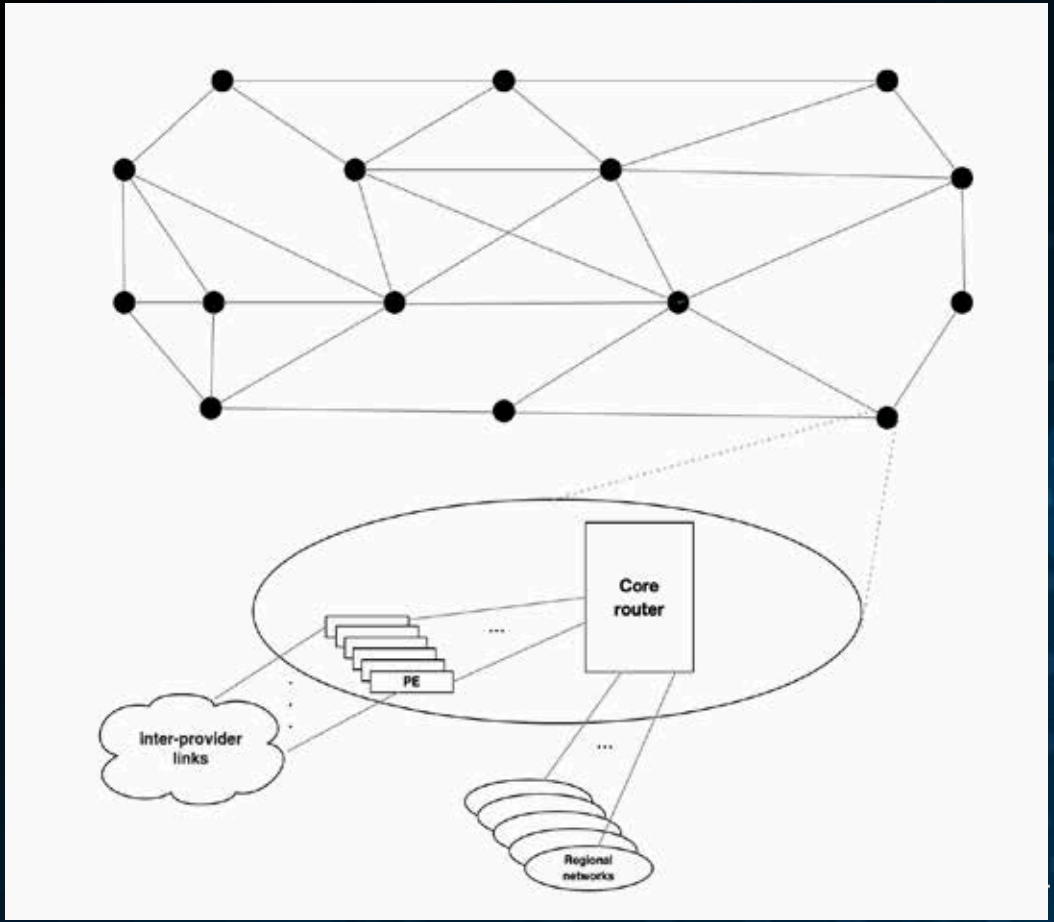
February, 2021

COMCAST

NATIONAL BACKBONE

A TYPICAL CORE SITE

- Core routers are typically mega routers, supporting thousands of ports, multi-Tbps of critical traffic
- Mainly scale up hardware, some scale out options
- Provides connectivity between external networks and internal regional networks
- Interconnects with all other core routers over long-haul fiber
- Carries Terabits of transit traffic



CHALLENGES

CHALLENGES SUPPORTING CORE ROUTERS

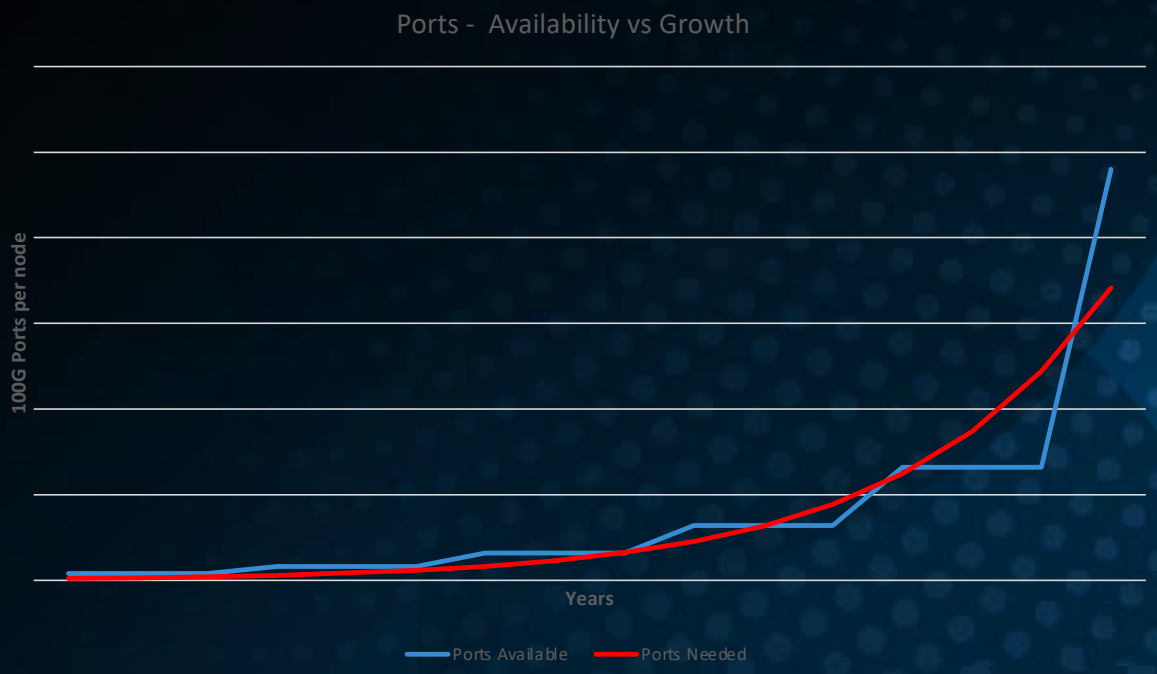
- Increase in ports supported by single node, driven by traffic growth
- ASIC bandwidth vs features
- Cost of SW (features) and HW (ports, ASICs)
- Failure zone (blast radius)
- Maintenance window
- Environmental constraints

CHALLENGES

CORE ROUTER PORTS

- New hardware certification and deployment has 2-3 year cycles
- Hardware refresh cycle takes 12 month period
- Always behind with demand at end of the product life cycle

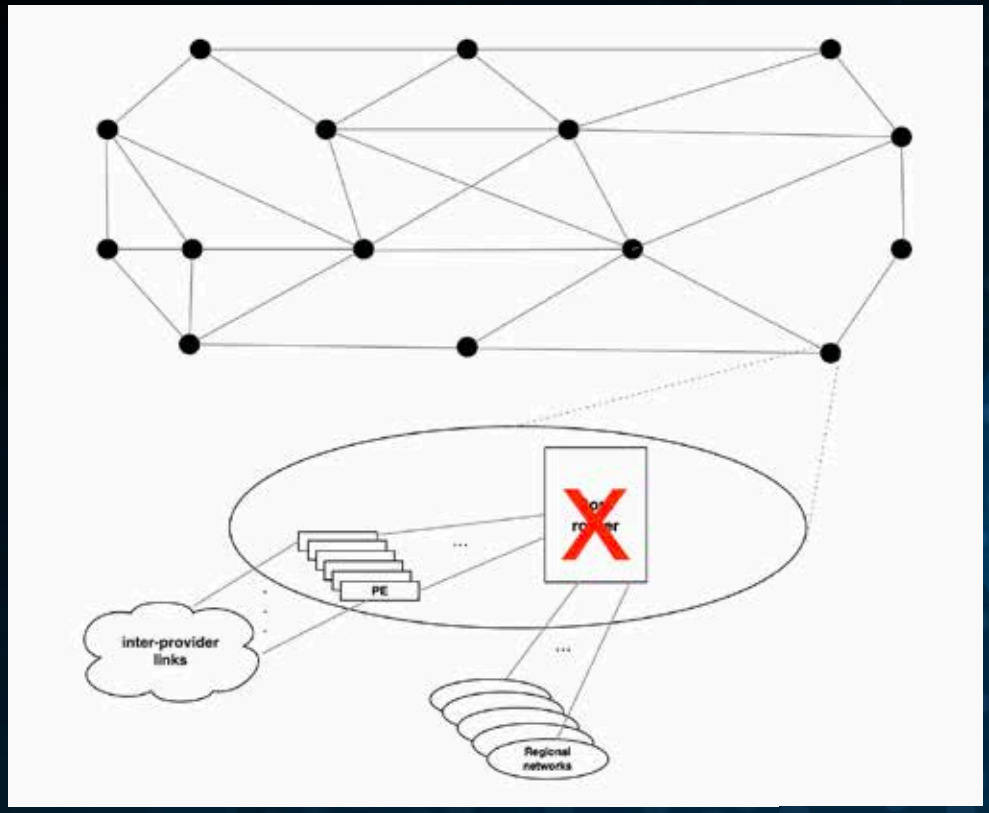
PORTS NEEDED TO SUPPORT GROWTH VS PORTS AVAILABLE ON SINGLE CARRIER GRADE ROUTER



CHALLENGES

FAILURE BLAST RADIUS

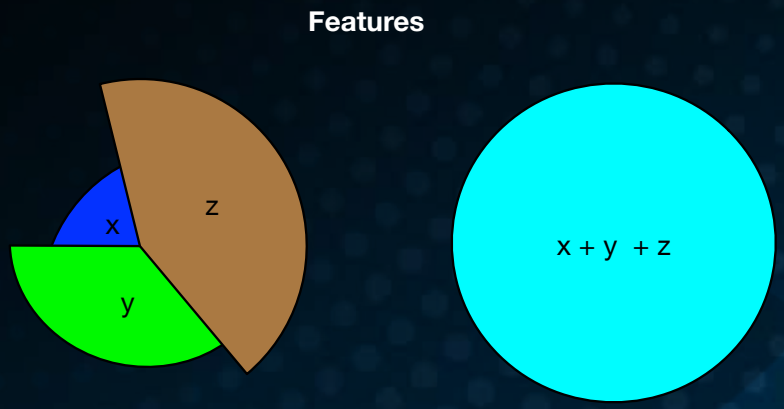
- Single node failure impacts millions of customers
- Creates a snowball effect, impacting regional networks, core links, and peers
- Another failure in connected network with a core router failure could be a disaster



CHALLENGES

ECONOMICS AND FEATURES

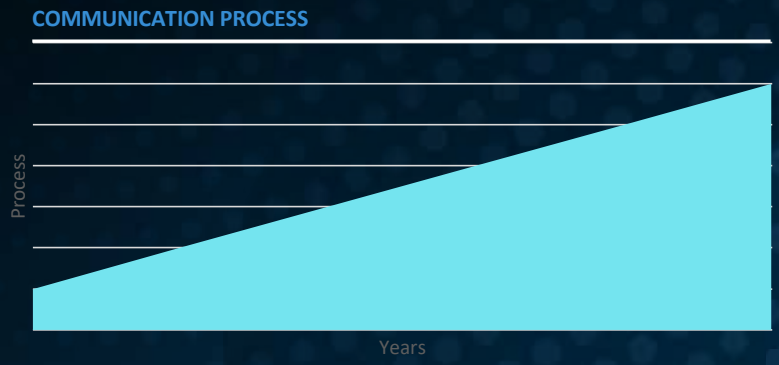
- Port cost to support complex features is more expensive compared to basic IP routing features
  - Difficult to mix ports on a single chassis
- Port cost of multi-chassis router is more expensive compare to fixed or modular chassis router
- ASIC bandwidth inversely proportional to features supported. More bandwidth added on ASIC chipsets, impacts the features supported.



CHALLENGES

OPERATIONAL CHALLENGES

- Limited number of maintenance windows available per year
- Require tight coordination between various Ops teams, with no to zero room for error
- Takes long time to push config updates, code upgrades, Business As Usual augments
  - Node swap out every ~5 years because of traffic growth, detailed in earlier slide



CHALLENGES

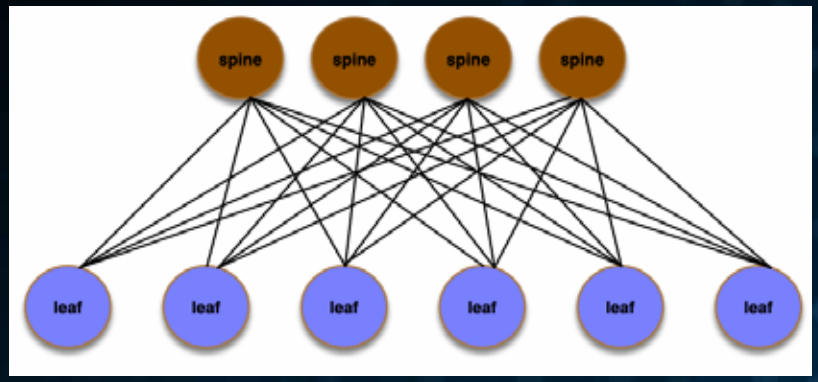
ENVIRONMENTAL DEMANDS

- Power per rack is limited in most of the colocation facilities
  - New NG (Next Gen) router with 200+ Tbs port capacity has high power demands (60kW+ per rack)
  - Throughput and port capacity is directly proportional to per rack power needed
  - Though ‘power per gig’ is going down with new ASICs, net bandwidth increase requires higher ‘power to rack’
  - Cooling is directly proportional to port count and power used
- Space
  - Colocation facilities have rack spacing requirements which forces us to evacuate surrounding racks to deploy high density NG routers

NEXT GEN CORE CLUSTERS

NEXT GEN CORE CLUSTERS

- Built using platforms with NG chipsets
- Built using Clos model; spine/leaf
- Flexible scaling, horizontal and vertical
- Seamless integration with the existing network
- Small failure domains (blast radius)
- Reduces the impact to services
- Single failure doesn’t impact flows in all directions
- Move the focus from platforms to Architecture, opening more options for deployments
- N+1 spine design allows hitless maintenances and updates

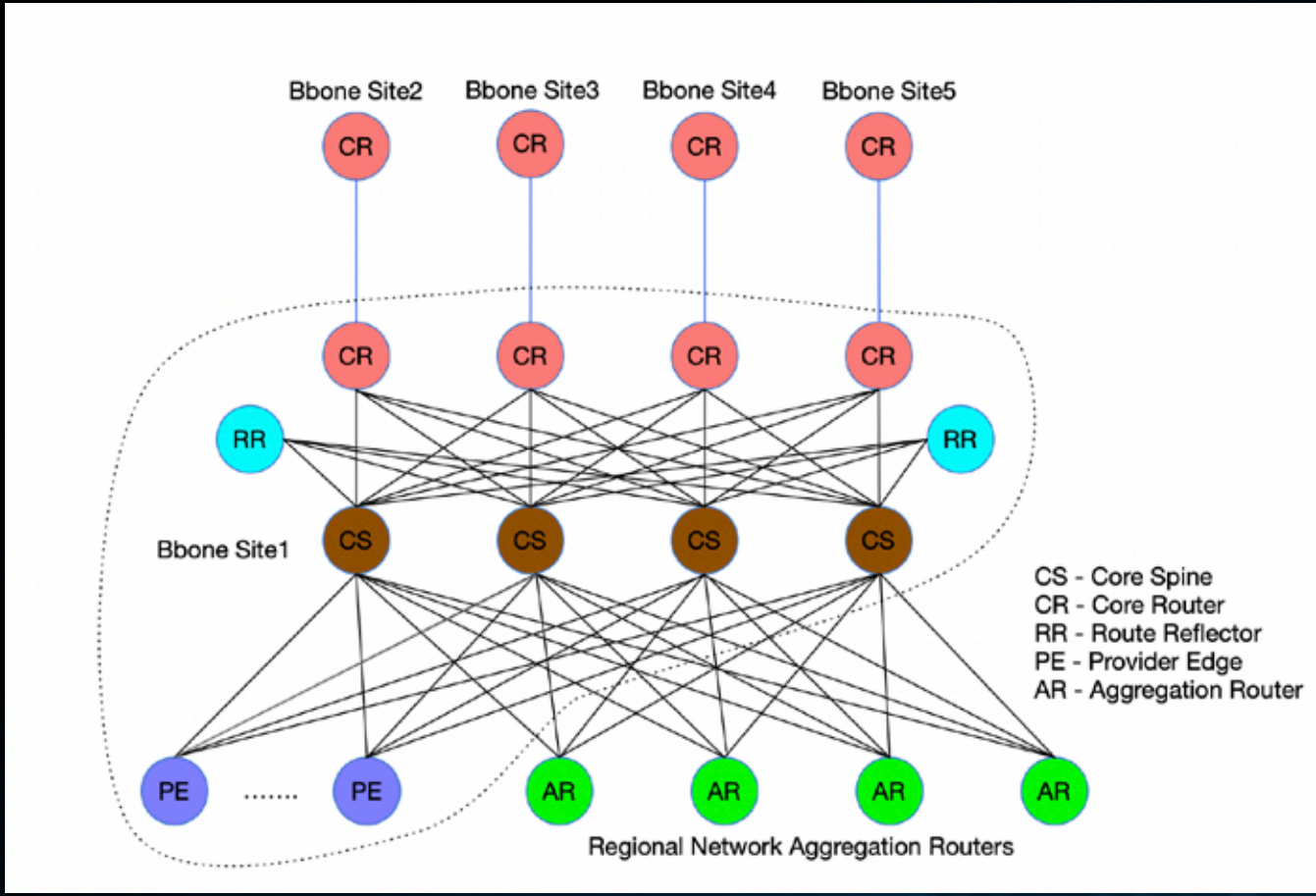


CLOS FOR CORE ROUTERS

IMPACTS OF DEPLOYING CLOS IN CORE LOCATIONS

- Deploying Clos in core locations had its own challenges
  - Requirements in Core locations are different compare to Datacenters
  - Clos deployments increase the number of nodes in the IGP domain
  - High number of ECMP paths
  - Need tools to monitor load sharing over ECMP paths
  - Need to rework BGP design
  - With multi-level BGP design, need enhancements and features to improve recovery time during convergence
  - Need tools and dashboards to monitor Clos
  - Some of the links are over long-haul fibers.

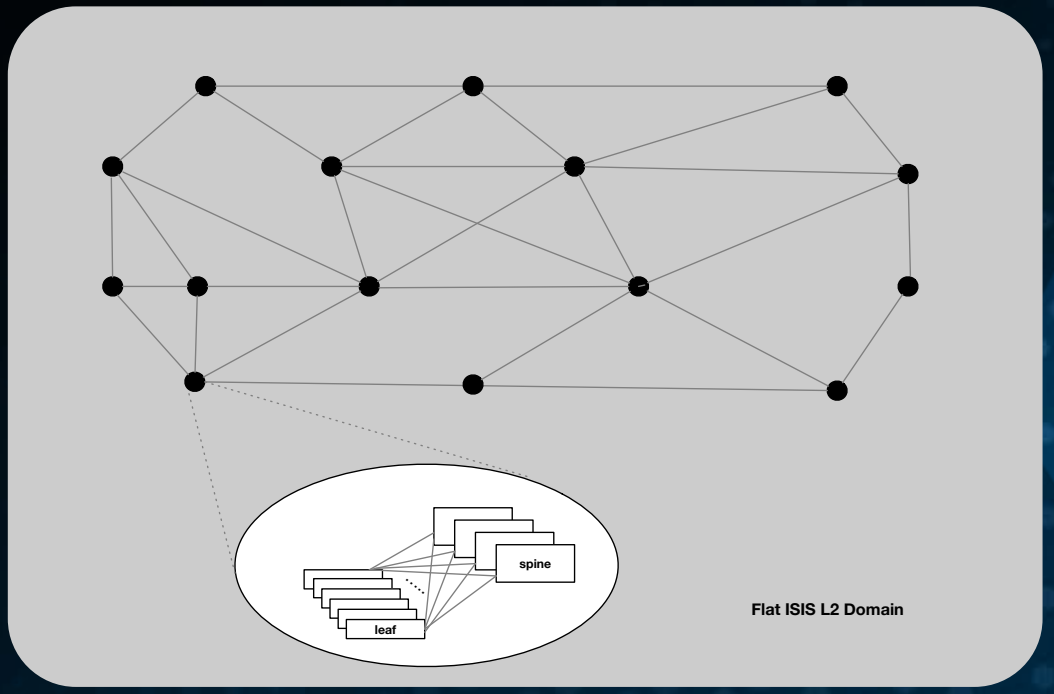
PHYSICAL VIEW



IGP DESIGN

IGP

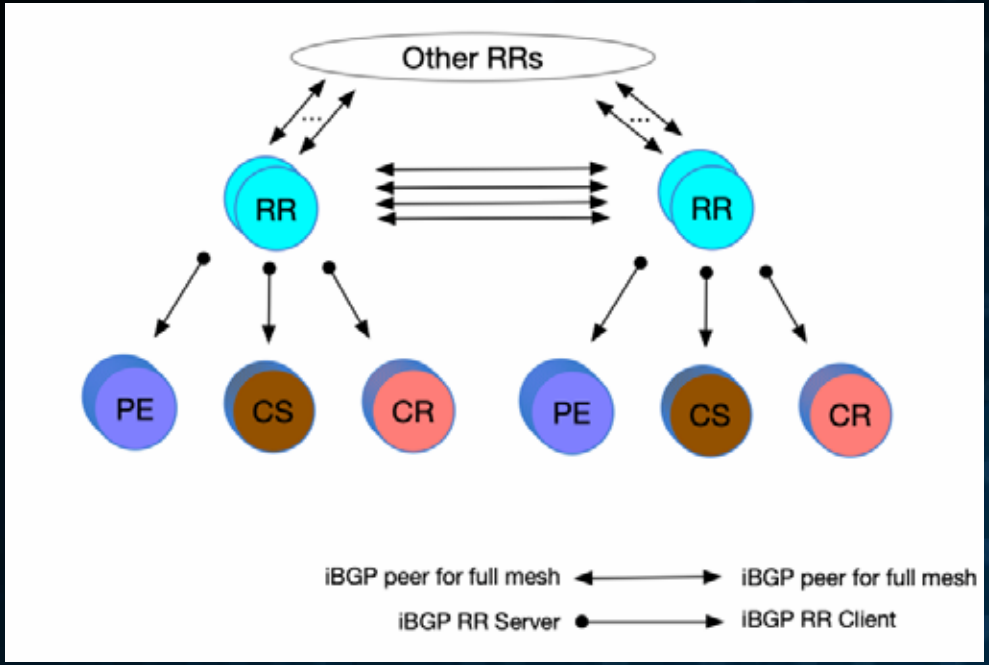
- All nodes and links are configured in one ISIS L2 domain
  - Closely watching the IGP dbase growth and thresholds, impact to convergence
  - Our Core network is in its own IGP domain, separated from regional networks, which helps managing IGP dbase.
  - Couple of options on the table to reduce IGP dbase in the future, if needed
  - Couple of IETF drafts out there that we are watching, could help us if needed in future



BGP DESIGN

BGP

- Two-layer BGP design
- Dedicated x86 servers for Route Reflectors
- All RRs in iBGP full mesh
- RRs are acting as servers for all local nodes



BGP FEATURES AND ENHANCEMENTS

BGP ADD-PATH

- Needed for ECMP load sharing

BGP - PREFIX INDEPENDENT CONVERGENCE (PIC)

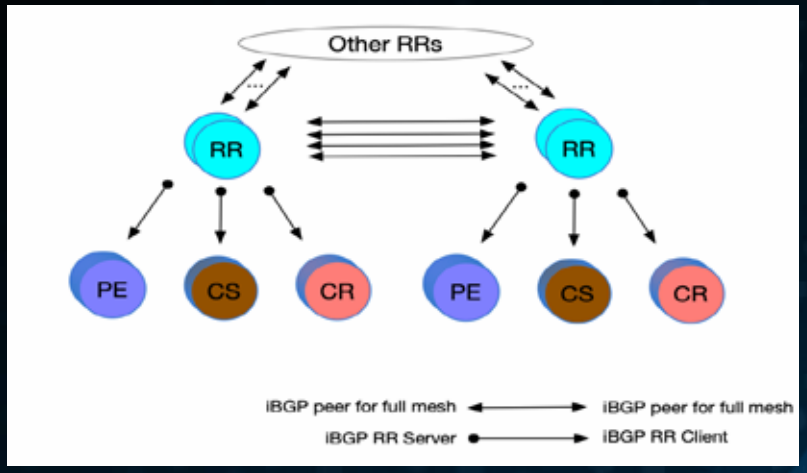
- Needed for better convergence
- Used BGP PIC multipath feature
- Working with vendors on new enhancements for better convergence

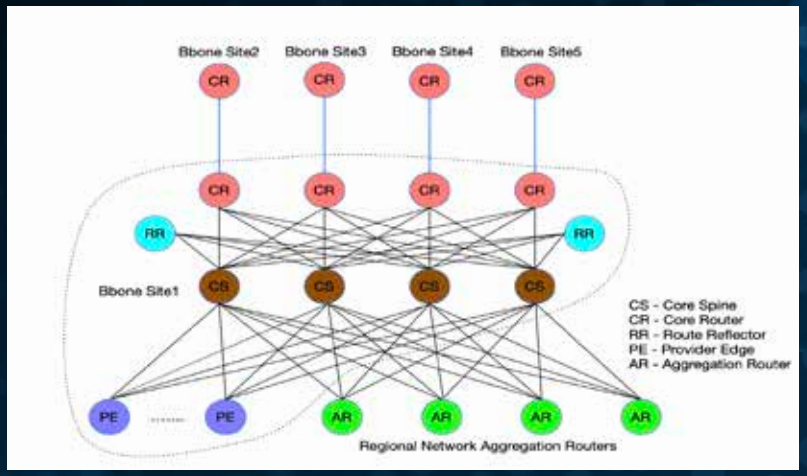
ANYCAST BGP NH

- Reduces the number of BGP paths
- Anycast advertisements tied to BGP state and other triggers

BGP PATHS EXPLODES WITH ADD-PATH AND FULL MESH

- Options used to reduce the BGP paths
  - Local full view for local clients only
  - Best + backup path for remote peers
  - Exploring other options to reduce BGP path count





PROVISIONING AND SUPPORT

AUTOMATION

- Device launches
- Cable validations
- Device configurations
- Code upgrades

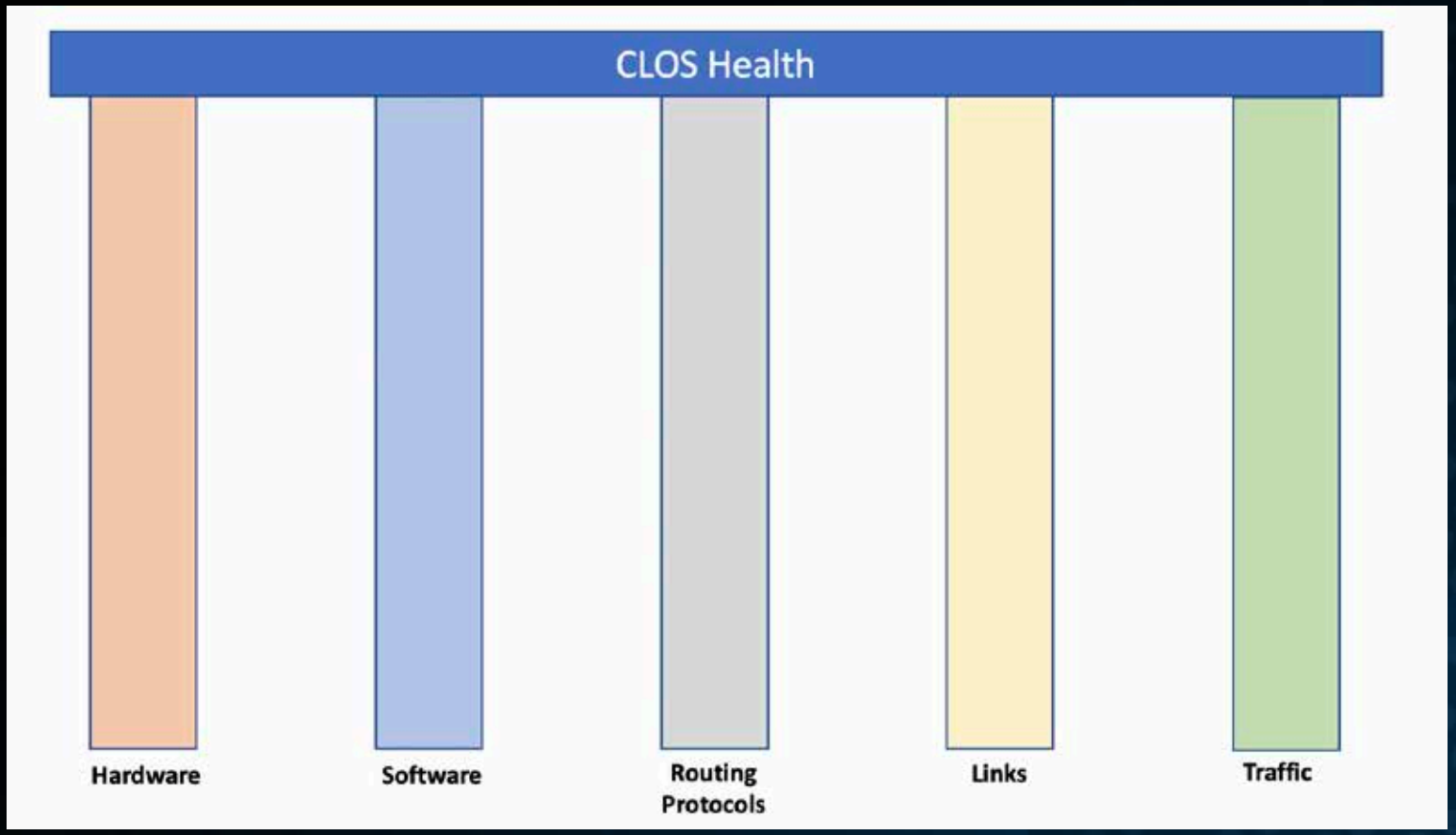
OPERATIONS

- Need more than SNMP
- Model driven telemetry data streaming
- New OPS dashboards and tools using telemetry data streams

SUPPORT - CLOS HEALTH

- Collection of routers make a Clos cluster
- All routers should be grouped and monitored as single unit for monitoring the health of the Clos clusters
- We are now managing the fabric components that are typically managed by vendor in single node design
- Dashboards and tools should monitor the entire group of routers in the clusters as single SRLG

CLOS HEALTH PILLARS



CLOS HEALTH - HARDWARE

Clos Health

- Monitor utilization against threshold for all hardware components and resources.
  - NPU resources
  - Memory tables
    - TCAM (Ternary Content-Addressable Memory)
    - LEM (Large Exact Match)
    - LPM (Longest Prefix Match)
    - FEC (Forward Error Correction)
    - ECMP-FEC (Equal-Cost Multi-Path FEC)
  - Port buffers
    - Ingress and Egress
  - Hardware errors

CLOS HEALTH - SOFTWARE

Clos Health

- Monitor Software processes
  - Bugs
  - SMUs
  - SW Signature checks

CLOS HEALTH – ROUTING PROTOCOLS

Clos Health

- Monitor all routing protocols against designed threshold.
- LLDP
  - Fabric links – Network learned vs Design
- ISIS
  - Fabric links – Network learned vs Design
- PIM
  - Fabric links – Network learned vs Design
- BGP
  - Network learned vs Design
    - Clos BGP sessions
    - Clos BGP prefixes received
  - Trend of prefixes received per neighbor
- Trend of route churn
  - Network stability
- Routing policies, route redistribution
- SRLG grouping across multiple routers
  - Edge router and external networks
    - ISIS
    - BGP

CLOS HEALTH – LINKS

Clos Health

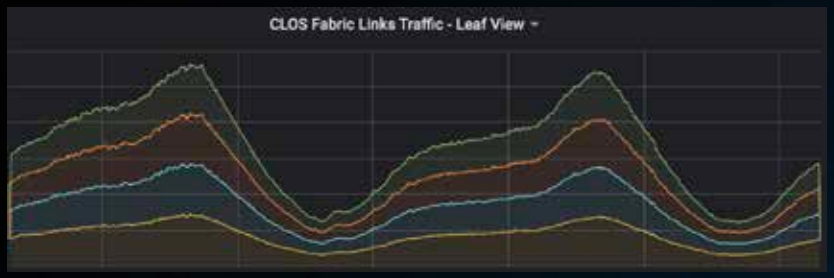
- Fabric Links
  - Fabric links count provisioned vs operational
  - Fabric bandwidth provisioned vs available
- Fabric Redundancy check
- Optic statistics
  - TX and RCV power

CLOS HEALTH - TRAFFIC

Clos Health

- Link Utilization
- ECMP Load sharing
  - Fabric links per leaf node
- Cluster/Node level packets/bits inbound vs packets/bits outbound
- Traffic grouping for external networks
- Traffic grouping for downstream edge routers
- Drops
  - Interface drops
  - Null0 drops
  - Ttl drops
  - No route drops
  - QoS drops
  - Hardware drops

SOME GRAPHS USING TELEMETRY STREAM DATAPOINTS



NEXT GEN CORE CLUSTERS

PROGRESS

- NG Core clusters broadly deployed in Comcast Backbone
- Able to do code upgrade and configuration changes during daytime
- Better traffic utilization and management
  - Single router failure doesn't impact multiple paths

LESSONS LEARNED

- Automation is the key
- Need better Operations process and tools
  - To manage fabric bandwidth, if over subscription used
- A lot of cabling work
  - Used telemetry data to validate the cabling
- Design to last longer, get all physical work done on day 1


COMCAST