Company Overview

InfiniComm is an imaginary Internet Service Provider (ISP) in the United States that owns its fiber transmission facilities as well as a Layer 2 switching infrastructure (ATM) across the country. InfiniComm has been offering Internet access for many years to other service providers (wholesale), large enterprises, and small/medium business customers. It currently has an installed base of more than 40,000 Internet ports. These Internet ports are supported on 500 Internet edge routers located at 100+ Points of Presence (POPs) that are scattered across the country. Internet connectivity is obtained via transit Internet Service Providers, private peering links, and connections in major cities to various Public Internet Peering points.

In addition to Internet access, for many years InfiniComm has been offering premium VPN-type services to large enterprise customers leveraging a purpose-built dedicated ATM network. Currently about 3000 ATM ports are installed across the country, and this number is growing considerably every year. The customer-managed customer edge (CE) routers are connected via 200 ATM switches hosted at various InfiniComm's POPs.

Network Overview

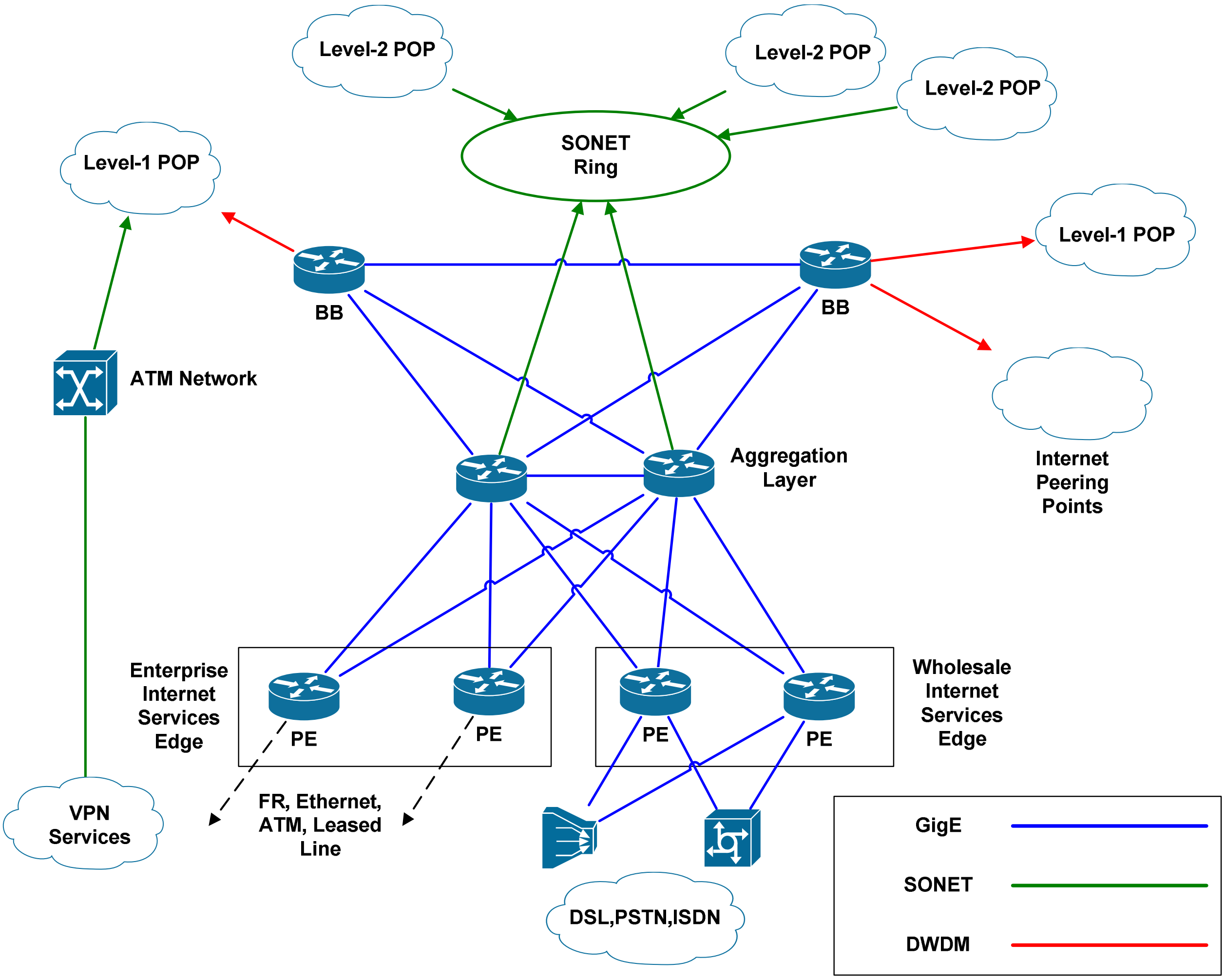
InfiniComm owns fiber across the country and has a long-distance optical core deployed, built on DWDM technology to optimize fiber use. Fiber paths normally follow the major railroad communication veins to save costs on fiber deployment. The use of DWDM provides quick availability and lower cost of additional optical paths if those are needed to scale network bandwidth. The high-speed core links are provided to routers as native wavelengths straight from the DWDM equipment using high-speed (e.g. OC192) SONET interfaces. Optical/TDM traffic grooming is used to optimize port utilization on DWDM and SONET ADM equipment. Notice that those SONET circuits do not benefit from any sort of protection at the optical level – SONET is simply the framing used for high-speed links. The network assumes reliance on IGP for network restoration upon a core link failure. Shorter distance links interconnecting various POPs could be SONET based and protected by means of optical features provided by Bidirectional Line Switch Rings (BLSRs). Intra-POP connectivity is achieved using point-to-point or switched Gigabit Ethernet circuits or Ethernet link bundles.

Access from Enterprise CE router to PE router for Internet connectivity is provided via Frame Relay, ATM, leased line, or SONET (PoS) connections. Each of these physical (or logical) links is dedicated to a single CE router. Access speeds may range from 64 kbps to OC-48. Wholesale Internet access is provided by terminating virtual dial-in connections for xDSL and various forms of dialup connections from other providers. Private peering links are used for the purpose of route exchange to facilitate VPDN tunneling.

The InfiniComm network is structured in two-level POP hierarchy. Each POP is classified as either a core (Level 1) or aggregation (Level 2). The POP level depends on the density of the customer access and combined traffic throughput requirements.

Level-1 POP

Level 1 POPs represent the high-capacity IP backbone dedicated to long-distance transit and interconnection of level 2 POPs to this long-distance transit backbone.

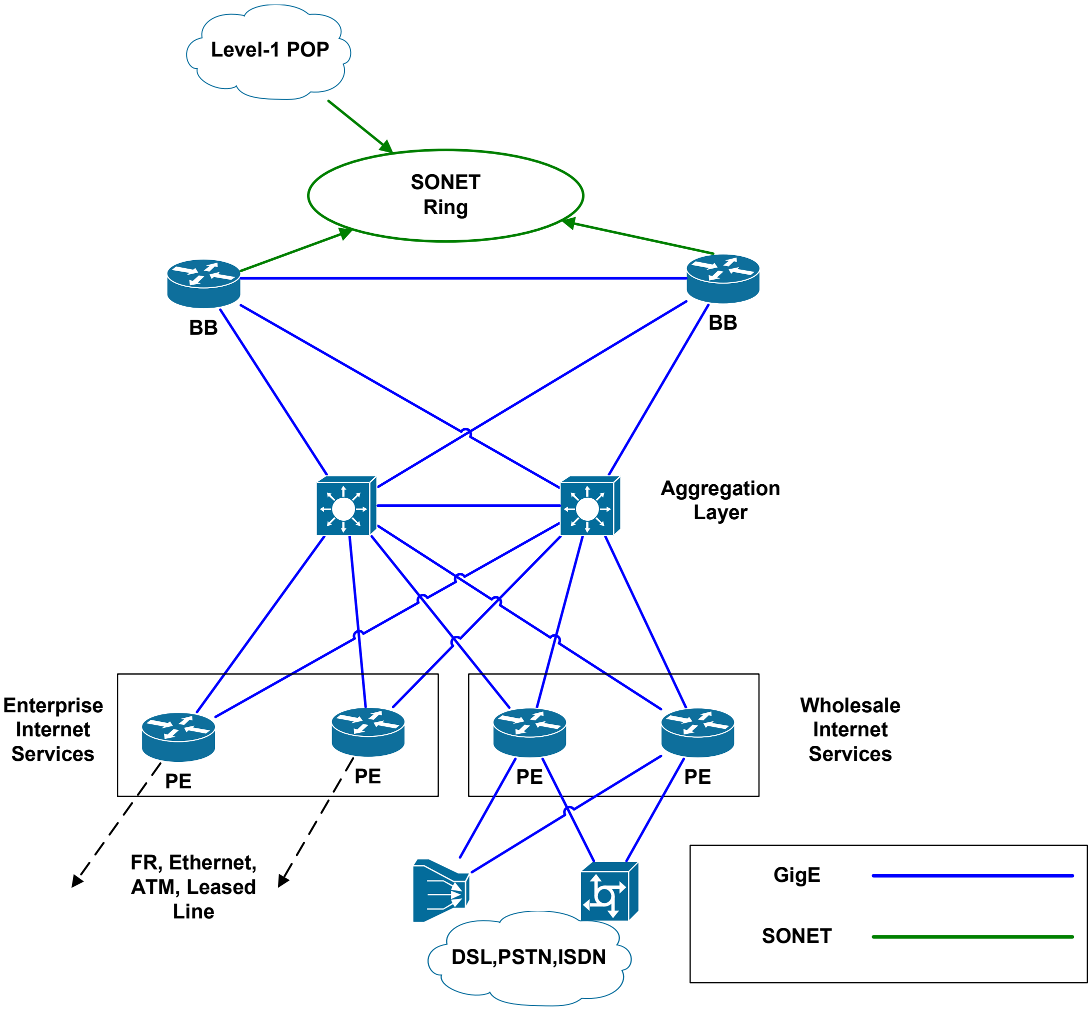


The BB (backbone) routers provide connectivity to other Level-1 POPs and communicate with the PE routers and Level-2 POPs using the aggregating P (provider) routers. As mentioned previously, connections between Level-1 POPs are provisioned using high-speed unprotected SONET circuits running over DWDM lambdas. In addition to providing network core services, some Level-1 backbone routers (BB) provide connections to large Internet peering points at various US locations exchanging full Internet routing tables with their peers.

Every Layer 1 POP edge is subdivided in three main sections – Enterprise Internet Services, Wholesale services and ATM network for Enterprise VPN services. Enterprise Internet services are delivered via dedicated set of PE routers terminating various circuit types using Frame-Relay, Ethernet, ATM encapsulations over leased lines. As a premium service, permanent ATM connections (PVC) could be provisioned for companies looking for private WAN clouds. Alongside with Enterprise services every POP terminates Virtual Private Dial-in sessions for wholesales connections. Every Level-1 POP has dedicated set of PE devices for this purpose, providing Internet access to DSL and PSTN customers from other ISPs.

Level 2 POP

The Level-2 (aggregation) POPs are similar in structure to Level-1 but lack the ATM "VPN service" component. Due to lower equipment density there is no intermediate routing aggregation layer and Layer-2 switches are used instead to aggregate physical cabling and connect PE routers to BB devices.


Level-1 POP
SONET Ring
BB
BB
Aggregation Layer
Enterprise Internet Services
PE
PE
PE
PE
Wholesale Internet Services
FR, Ethernet, ATM, Leased Line
DSL,PSTN,ISDN
GigE
SONET

Routing

InfiniComm has been allocated a single BGP autonomous system number. At the present moment, InfiniComm uses IS-IS as the IGP of choice inside its autonomous system. Single IS-IS L2 area is deployed encompassing both Level-1 and Level-2 POPs. IS-IS is used to carry infrastructure routing information only, such as transit core links, PE router and Loopback interfaces. The total size of the IGP routing table is slightly above a thousand prefixes. BGP is used to carry the full Internet routing table and peer with enterprise customers requiring dynamic routing protocols. Internal BGP connections are organized as full-mesh of iBGP peering sessions between all P and PE devices. Furthermore, eBGP connections are used to connect to major Internet peering points and large enterprise customers. Every router involved in BGP peering has plenty of DRAM to keep multiple copies of BGP routing table, though there are some considerations about growing forwarding tables on the router line-cards. InfiniComm supports both customers using the ISP’s PA address space or advertising their own individual PI address blocks.

Initial Analysis:

There is quite a lot of information presented, but you need to grasp the key points only to minimize time spent on reading the documents. Firstly, the network topology – it looks like we have a partially meshed core and aggregation layer. To some extent, this could be classified as “snowflake” topology with partially meshed core. The next important thing is type of the connections. Almost every link is point-to-point – even connections over SONET collector rings are in effect just P2P links. Most of the links are not protected at transport level, which means active IGP tuning is required to achieve better convergence. We also notice that intra-POP connections within Level-2 POPs are based on switched Ethernet and therefore are not truly P2P, which may create problems with fast failure detection.

We may also notice that core bandwidth is cheap thanks to the DWDM infrastructure and company-owned fiber. This most likely means there are no under-provisioning issues in the network core. Furthermore, we notice that the company provides Internet connection services over IP network and private WAN (VPN) services using separate ATM infrastructure, that is the network is not converged for multiple services.

Finally we look at the routing model, which is very simple – single area IGP and full-mesh BGP. This simplicity obviously creates scalability issues due to network being more vulnerable to various failure events and management complexities with scaling BGP full mesh. Therefore, we may already see a few challenges to this network: (1) scaling (2) service convergence and (3) resiliency issues at L2 POPs

Incoming Email
From: InfiniComm Network Manager

As you already know our network has been built around simple and conservative design model to facilitate predictability and easier management. However, it appears we are reaching our limits with the existing routing deployment model. We estimate adding one new Level-2 POP on average every 3-4 months in the next few years and the management burden of provisioning new devices becomes a serious challenge. Specifically, adding new BGP speakers becomes problematic due to the full-mesh of BGP peering sessions. However, we are still fine with our existing IGP model and do not want to change it. We are, therefore, looking for a way to simplify the BGP routing model and seeking your advice on this.

Question 1:

What additional information you need in order to make a BGP conversion decision?

- Traffic engineering requirements inside InfiniComm AS
- Network failure statistics
- Network Convergence Requirements
- BGP table size and average number of BGP paths per prefix
- No additional information is required to make a decision.

Answer:

Based on the baseline information a choice of BGP RR is obvious – there is single IGP and migrating to RRs is less disruptive than to BGP confederations. Furthermore, migrating to BGP confederations could be a reasonable choice if the network was about to be split into multiple IGP domains, e.g. to improve fault isolation. The question becomes: how much should we shrink the BGP full-mesh? Keep in mind that having full-mesh of BGP adjacencies allows for accurate traffic routing decisions at every device, since they all have information of every exit point from the local AS. Furthermore, storing multiple BGP paths may allow for efficient load-sharing via iBGP multipath and unequal-cost load-balancing toward external routes. Shrinking BGP full-mesh also has effect on BGP convergence – when backup paths are not present it is not possible to engage any sort of fast failure recover mechanism in BGP and rely on slower BGP convergence via withdrawals and announcements. Therefore, we should be interested the following: Traffic Engineering Requirements and Network Convergence Requirements.

Additional Information:

From: InfiniComm Engineer

Speaking of our convergence goals, upon a single core link failure the network should recover in no longer than one second. The same should apply in the event of a single P-router failure. This requirement applies to any non-edge router in any POP, as implementing edge node redundancy is more complicated. Additionally, we are very concerned with the major Internet peering link or router failures. We want to provide quick recovery in case of either an Internet peering link or peering router failure as re-converging on a full BGP table takes significant amount of time. It is critical for us to provide resilient Internet transit services to our major peers. Right now we have tuned router ingress queues and have BGP TCP transport parameters tuned to improve performance, but the convergence times are still significant. Also, we rely on BGP keepalive times for detection of peering node failures, and we can't tune those down too much as this will expose us to the risk of BGP session flapping.

At the present moment we are not concerned with fast restoration times upon a PE-CE link/router failure for dual-homed customers: they do not inject much routing information so BGP based re-convergence yields acceptable times, provided that we tune BGP MRAI timer.

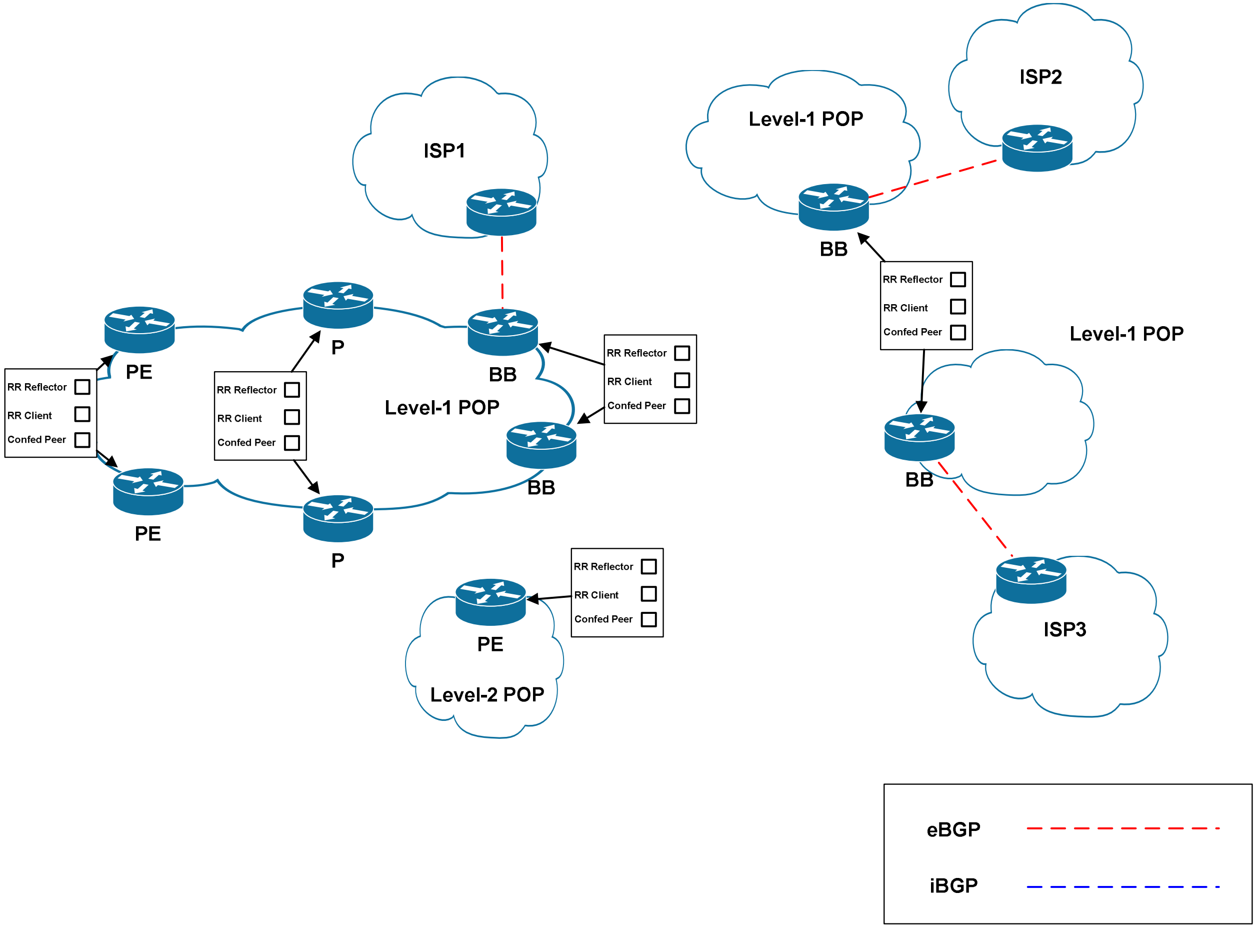
We don't implement any special form of traffic engineering, our main goal is to minimize amount of transit bandwidth we provide to our customers and peers. Therefore, we reset MED attribute for all routes we receive from our peers. Having the BGP full-mesh provides all routers with full view of all external connections and using IS-IS metrics based on link bandwidth results fairly good link load distribution and minimizes the amount of transit traffic through our AS.

Analysis:

Notice how text intertwines together IGP and BGP convergence requirements. The link failures should be handled by IGP and we should not pay attention to this information right now. As far as BGP is concerned, we are required to provide fast recovery upon a peering link failure. What we also need to notice is that the system uses hot-potato routing, or in other words relies on internal IGP cost as BGP tie-breaker to minimize the amount of transit traffic.

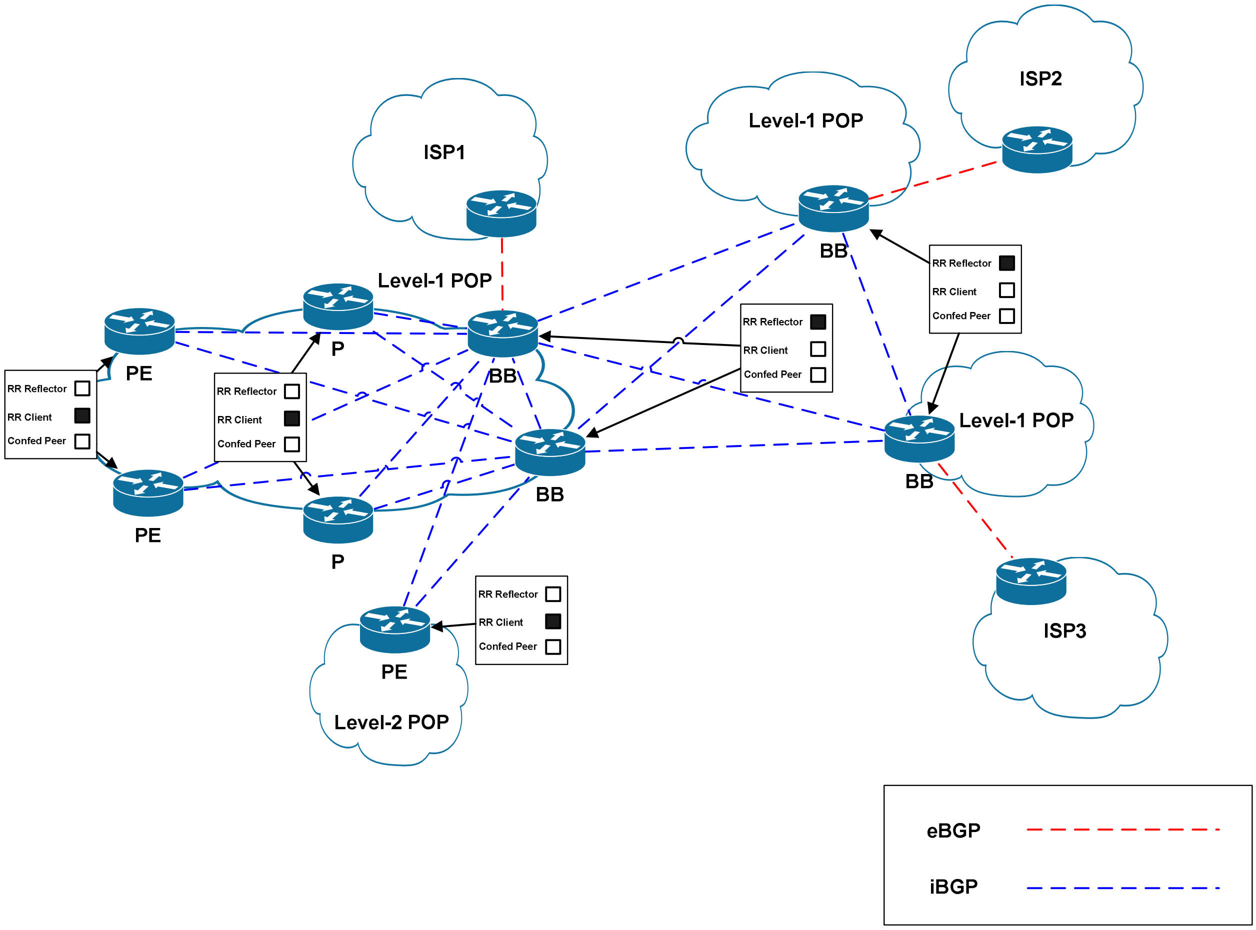
Question 2:

Diagram your solution using the figure below as template. You can add BGP peering sessions and designate routers as either BGP reflectors, clients or BGP confederation peers. You can add new routers for special purposes if needed.



Answer:

Here is how a viable solution may look like.



The main issue is proper placing of route reflectors. Since we are using classical IP routing, RR topology has to be congruent with physical topology. In other words, the RR's must in line with traffic paths to prevent routing loops. Another question is how deep into the L1 POPs we should push the BGP full-mesh boundary. Obviously, we want to keep the full mesh as small as possible but we have to satisfy the requirement of topology congruency. It is clear that all POP's must be interconnected via a full-mesh of sessions, but should that mesh go deeper in Level-1 POP's and include the aggregating P routers – thus, following the physical topology between L2 and L1 POPs? Or should we use two-level hierarchy of router-reflectors and make the aggregating P routers BGP reflectors as well – this also reflects the topology properly. Firstly, it is not recommended to add layers of reflection unless it's really necessary – information hiding may have serious impact on optimum exit point selection and resulting BGP convergence. Therefore, we should probably stick with just one level of route-reflectors, since our network is not that big (though "big" is relative).

Next, for the expansion of BGP full-mesh, the answer depends on whether we need the aggregating routers to have the option of quickly recovering from BGP failures. So far, the fast recovery is only required for the ISP peering sessions, which are terminated at the BB routers. Therefore, there is no need to push the full mesh down to the aggregating routers at every Level 1 and we could keep it just to the BB routers.

It is also interesting to point out that when needed BGP full-mesh could be pushed down the particular POP if the layer or routers requires more detailed view of the network topology. This is very similar to the principle of area partitioning in IGP protocols.

Question 3:

How would you assign clusters IDs to route-reflectors?

- One cluster ID for all BB routers at every Level-1 POP
- Separate cluster ID across Level-1 POPs but the same cluster-ID for both BB routers within the same POP
- Separate cluster ID for every BB router at every Level-1 POP

Answer:

The first answer is obviously wrong, as it will require every client to peer with every BB routers. Therefore, we are left for deciding which of the remaining two is better in our case. Normally, sharing the same cluster-ID between RRs prevents information duplication, but opens a possibility of inconsistent routing, in case if a client loses connection to one of the RRs. This possibility, however, is very insignificant in networks with resilient underlying infrastructure, as IGP routing takes care of link failures and iBGP session almost never notices the faults. However, as we remember, memory on the devices is not a problem, so we should probably go with the last option and allocate a separate cluster ID per every BB router.

Question 4:

Will the modified BGP mesh automatically provide effective support for hot-potato routing? **[Yes/No]**

Answer:

The answer is **No**, because hot-potato routing requires special considerations in RR environment.

Question 5:

What conditions must be observed to allow for hot-potato routing in this scenario?

- BGP Multipath should be enabled
- Full-mesh iBGP within every single Level-1 and connected Level-2 sites.
- IGP metric should be tuned so that inter-cluster paths are less preferred than intra-cluster paths
- RR's should reset next-hop to self
- BGP Cost community should be used in this scenario

We know that MED values are reset as routes are accepted into the system, but this is not enough to provide hot-potato routing if BGP is not using full-mesh of iBGP peering sessions. The problem is that best-path selection is now performed by RR devices that may have different IGP metrics to exit points, as compared to PE devices. In order to keep the selection consistent, IGP costs on the inter-cluster links should be higher than those of any intra-cluster path (BB to PE). This will effectively enforce RR to perform best-path selection in the same manner as the client PE would perform it. However, this is not enough for intra-cluster path. For example, if a PE has two paths to a remote system across the local Level-1/Level-2 conglomeration, then it may select one based on the IGP cost that is different from what the RR would use. To resolve this issue, full-mesh of iBGP sessions should be configured within every cluster. This results in a hierarchy of full-mesh iBGP session – a core full-mesh and a full-mesh within every L2/L1 POP group. Still this design is much more manageable than a full-mesh covering all routers.

As for the remaining two answers, there are obvious logic flaws. First, enabling BGP multipath and changing BGP next-hop does have some relation to BGP but bear in sense in this scenario. As for BGP Cost community, using it will creates behavior that is more like cold-potato routing in most scenarios as opposed to the hot-potato routing, that uses IGP metrics.

Question 6:

What is your suggestion for handling transit link failures in the ISP network, specifically the links interconnecting Level-1 POPs?

- To accommodate the target convergence time, local link protection is required, e.g. by using IP Fast-Reroute or MPLS Fast-Reroute
- Ensure the link failure detection time is minimized by tuning carrier loss delay to a minimum and optimizing ISIS for fast LSP generation delay and SPF delay timers. This will allow for sub-second IGP convergence.
- Use link bundling for every inter-POP link and rely on link bundle control protocol to detect failure.
- Implement SONET optical-level protection mechanisms

Answer:

This question is no longer related to BGP convergence, now we are talking about the IGP process. From the information we got previously, we know that target convergence time should be under one second. There is a lot of ways to achieve that, but the simplest one is to configure IGP for fast convergence, by tuning LSP generation timers and SPF delay. In the network we have, almost every link is point-to-point with the exception of the Level-2 POPs that use Layer-2 switches for link aggregation. Since the question mentions Level-1 to Level-1 links, there should be no problems with fast link failure detection.

Using fast-reroute (local protection) is also possible, but requires either MPLS or other tunneling techniques to allow universal re-routing. This would require additional configuration and possible software updates. As long as the same goal could be achieved using simpler approach it should be preferred.

As for link bundling and SONET APS, those two are somewhat similar, though link bundling may provide bandwidth aggregation. The problem is excessive cost of this solution, as this means a 1+1 protection scheme.