Scaling IGPs in ISP Networks

Philip Smith

SANOG 8, Karachi

3rd August 2006

Agenda

- Definition of Scaling
- IGP Design
- Tuning OSPF

What Does Scaling a Network Mean?

- Scaling is very important for an SP network
- Non-scalable:
  - A large network which doesn’t converge
  - Complexity which is impossible to support
  - A solution which does not afford a plan for growth
  - Addition of devices or links which no longer provided incremental benefit, and their existence may even be a detriment to performance
- Remember, Occam’s Razor

“One should not increase, beyond what is necessary, the number of entities required to explain anything”

William of Ockham, 1295-1349

What Does Scaling a Network Look Like?

- Size
  - Number of Devices
  - Number of Prefixes
  - Number of logical divisions
- Speed
  - Convergence time
  - Service restoration
- Stability
  - Can the network take a hit and still work?
  - Or will your business end up tangled in a knot???

What Does Scaling a Network Look Like?

- Services
  - Addition of capabilities to the existing framework of the network.
  - These can be switching capacity, performance, functionality, amongst others
- Simplicity
  - Networks can scale their complexity without growing in size
  - Networks can scale their size without increasing complexity
- Which is better?
  - Depends on the problem you are trying to solve

The $1 Million question: Can you fix the network at 3am on a weekend without your notes in front of you?

The Wise Network Geek

Routing Scaling Factors: Route processor

- Faster is Better, Right? Well… maybe.
  - **PRP2 = MPC7457 CPU at 1263Mhz, Rev 1.1, 512KB L2, 2048KB L3 Cache**
  - **NPE-G1 = SB-1 CPU at 700MHz, Implementation 1, Rev 0.2, 512KB L2**
- Two of the fastest CPU's however in some tests G1 outperforms PRP-2 - why?
- CPU architecture is important. Don't just compare clock speed!

Routing Scaling Factors: Architecture

- Platform Architecture
  - Each box may have different bottlenecks and limitations
- Software Limitations versus Hardware Limitations
  - Software Limitations are more far-reaching than they would be for the protocols within operating system revisions
  - Hardware Limitations would be based on platform specific hardware
- Software Limitations versus Reality
  - Anyone can “define” support for 32 Million interfaces but it doesn’t mean the router is capable of doing it.

Routing Scaling Factors: Memory

- More is better!
  - Memory for Software features allows larger tables, greater scaling
  - Hardware memory is often the gating factor, and more does allow for greater table sizes, but lookup performance may suffer based on architecture
- Generally speaking, with sufficient main memory a router can hold 1M+ prefixes, however:
  - Platform architecture and hardware memory may not be able to hold the resulting FIB tables.
  - Prefix distribution is important, as it directly influences table structure and FIB data structure size and shape.

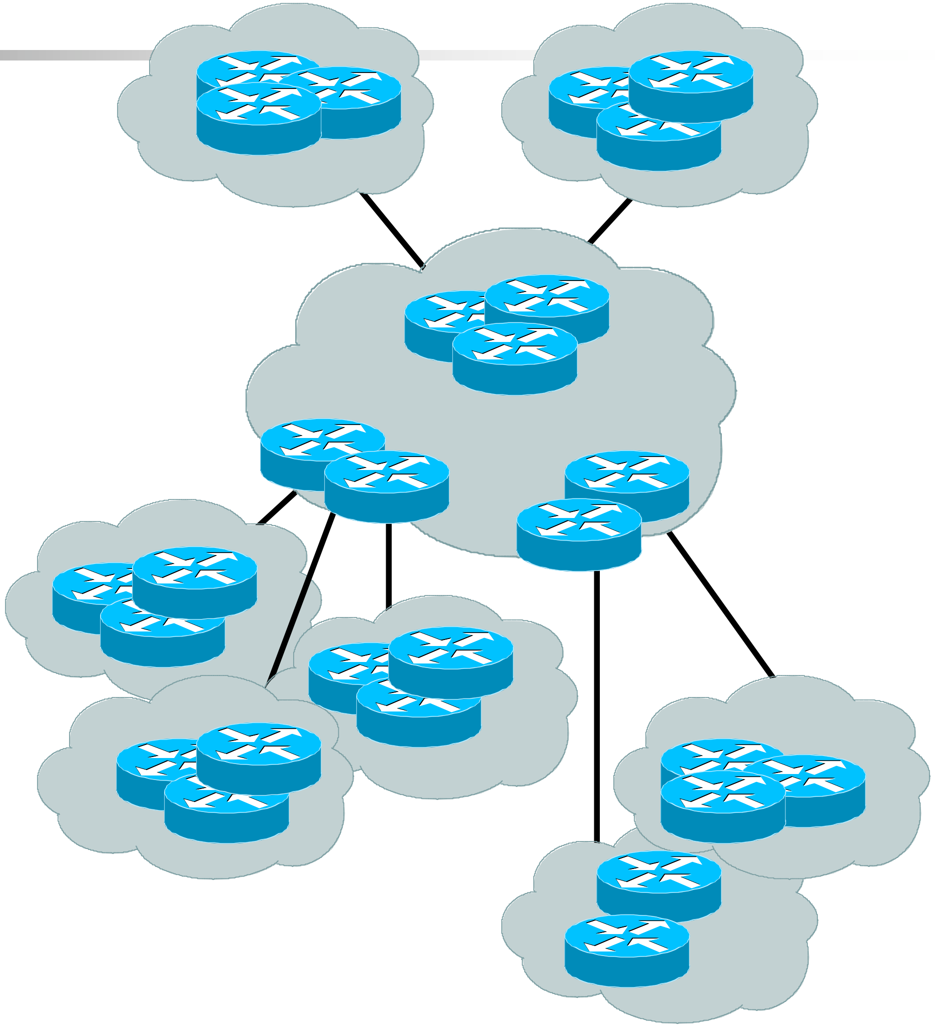
Routing Scaling Factors: Input/Output Speed

- Even if the CPU is fast, how quickly can we get the data to and from that CPU for processing?
- Inbound
  - How fast can we drain queues for inbound data and control plane traffic?
  - CPU dependencies - more power, the more processing that can be done, more work completed in less time
  - Software - there are software queues for the interfaces (sort of) and for the processes behind them. You may not overrun the interface, but may overrun the process level queue.
- Outbound
  - Framing packets is harder than you think
  - Routers are optimized for switching traffic fast
  - We can generally receive more than we can send

IGP Design

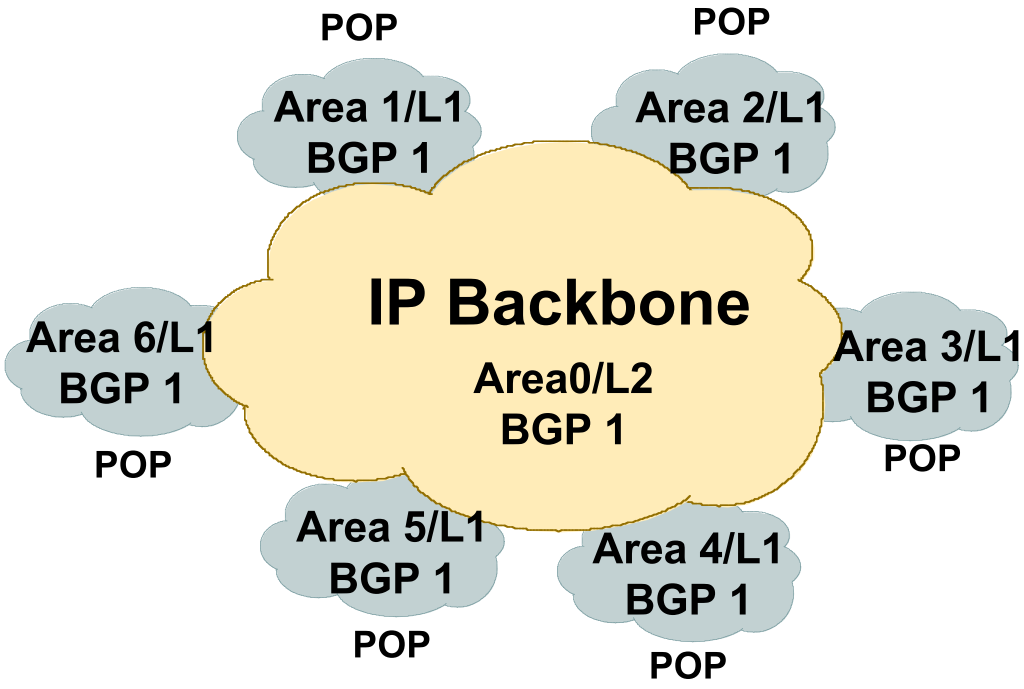
Service Providers

- SP networks are divided into PoPs
- Transit routing information is carried by BGP
- Customer address blocks are carried by BGP
- IGP is used to carry next hop within SP backbone only
- Optimal path to the next hop is critical



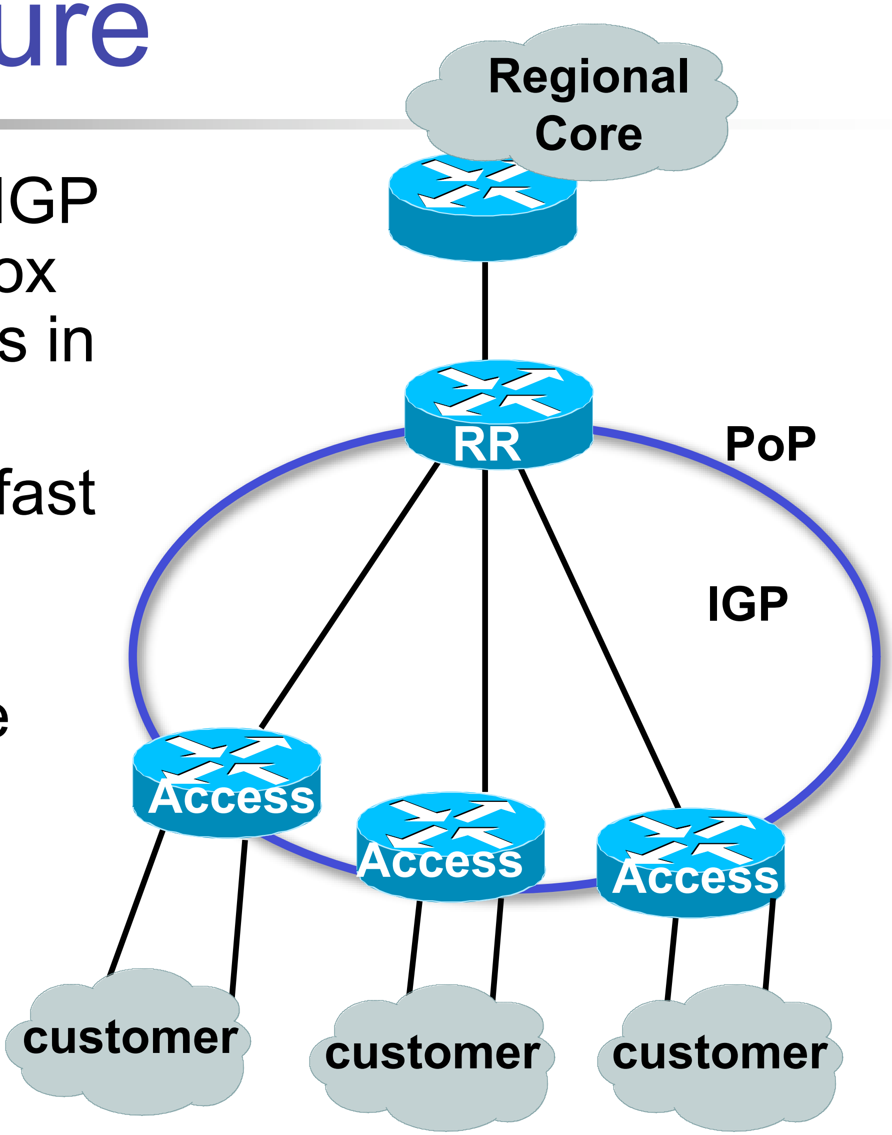
SP Architecture

- Major routing information is ~190K prefixes via BGP
- Largest known ISP IGP routing table is ~6–7K
- Total of 196K
- 6K/196K ~ 4% of IGP routes in an ISP network
- A very small factor but has a huge impact on network convergence!



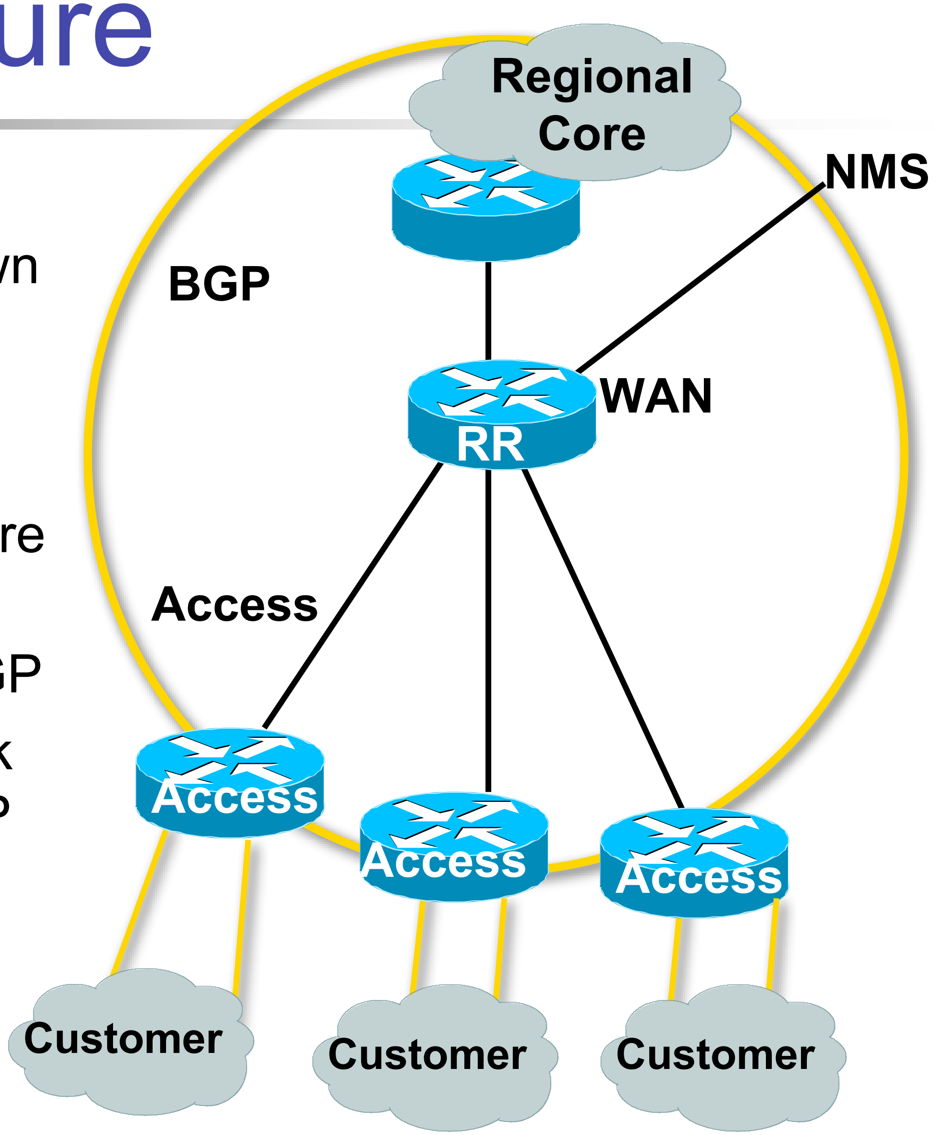
SP Architecture

- You can reduce the IGP size from 6K to approx the number of routers in your network
- This will bring really fast convergence
- Optimise where you must and summarise where you can
- Stops unnecessary flapping



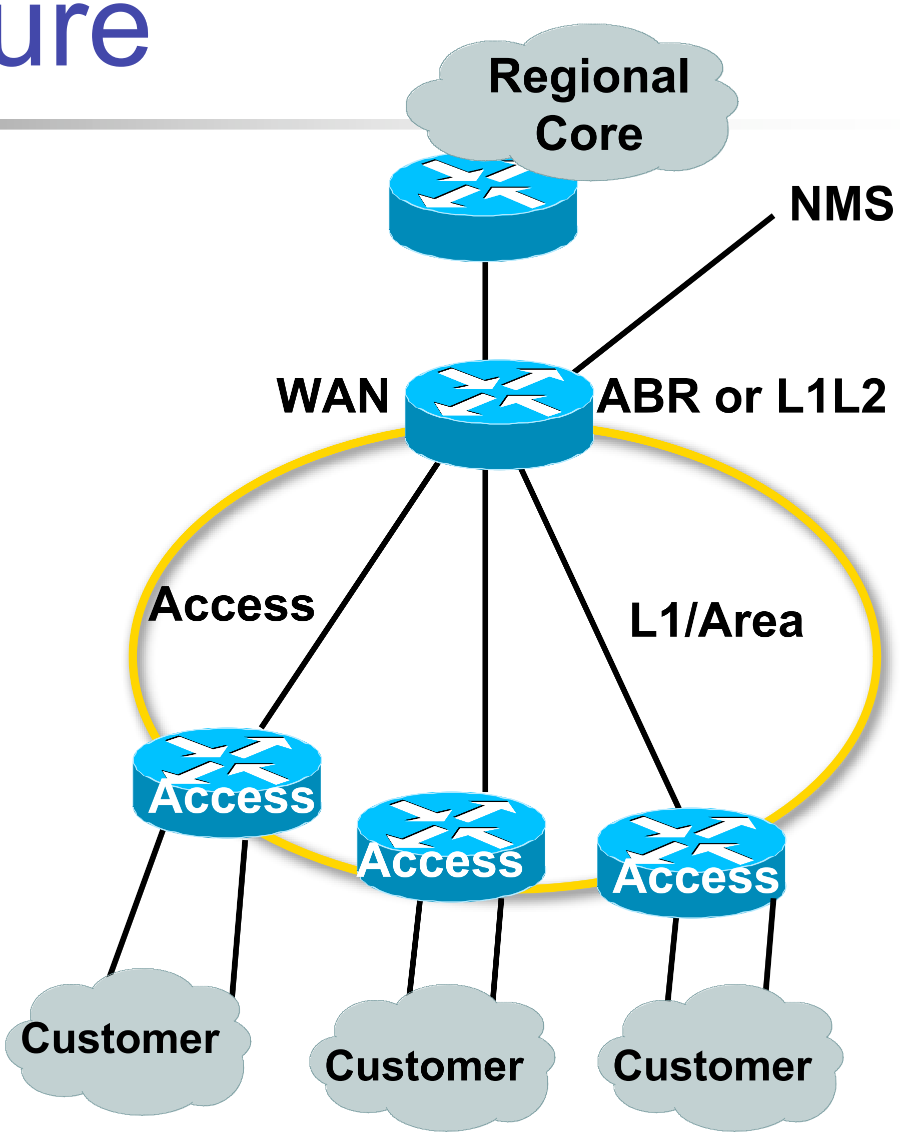
SP Architecture

- Link between ISP Access and Customer Router needs to be known for management purpose
- BGP next-hop-self should be used on all access routers — unless Customer and SP Access Router are on shared media (rare case)
- This will cut down the size of the IGP
- For Customer to Access Router link use redistributed connected in BGP
- These connected subnets should ONLY be sent through RR to NMS for management purpose; this can be done through BGP communities



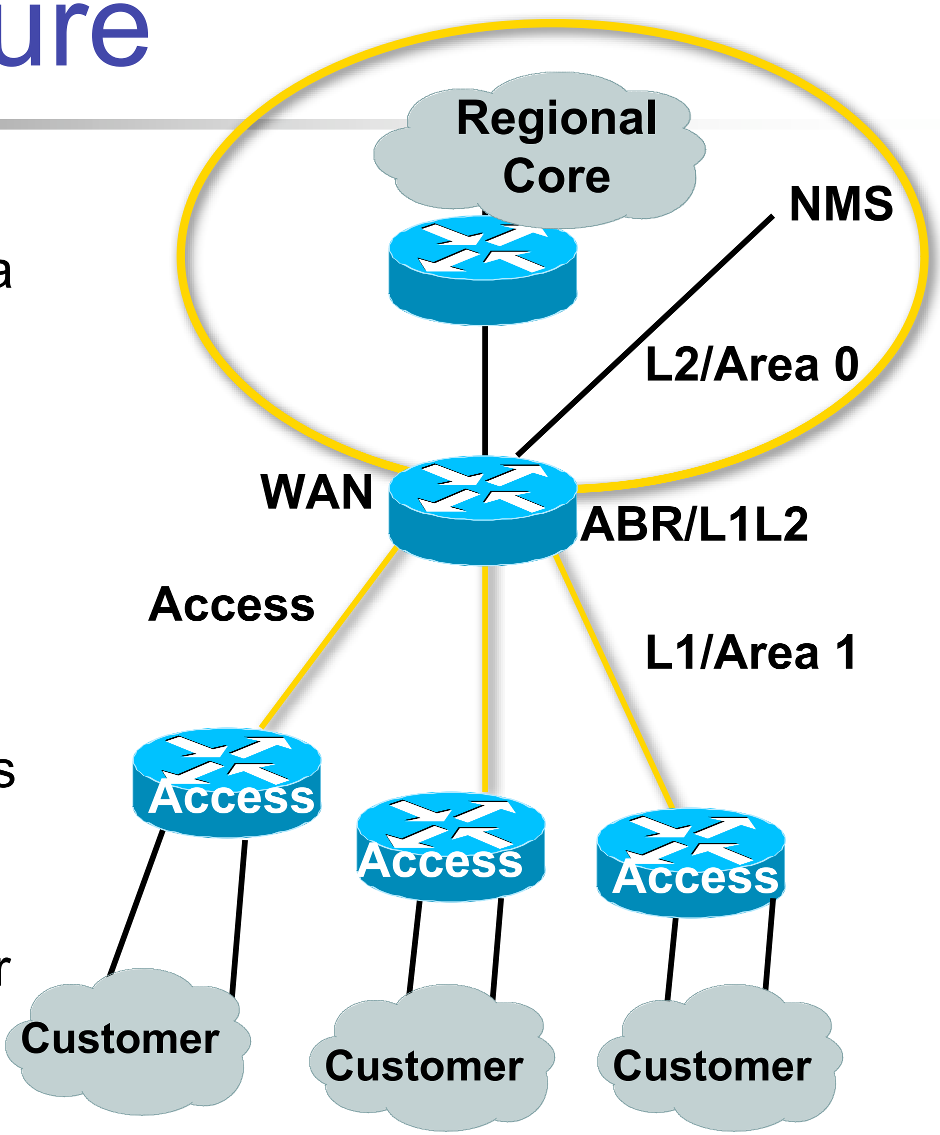
SP Architecture

- Where do we define area boundaries?
  - WAN routers can be L1L2 in ISIS or ABR in case of OSPF
- Hide the pop infrastructure from your core



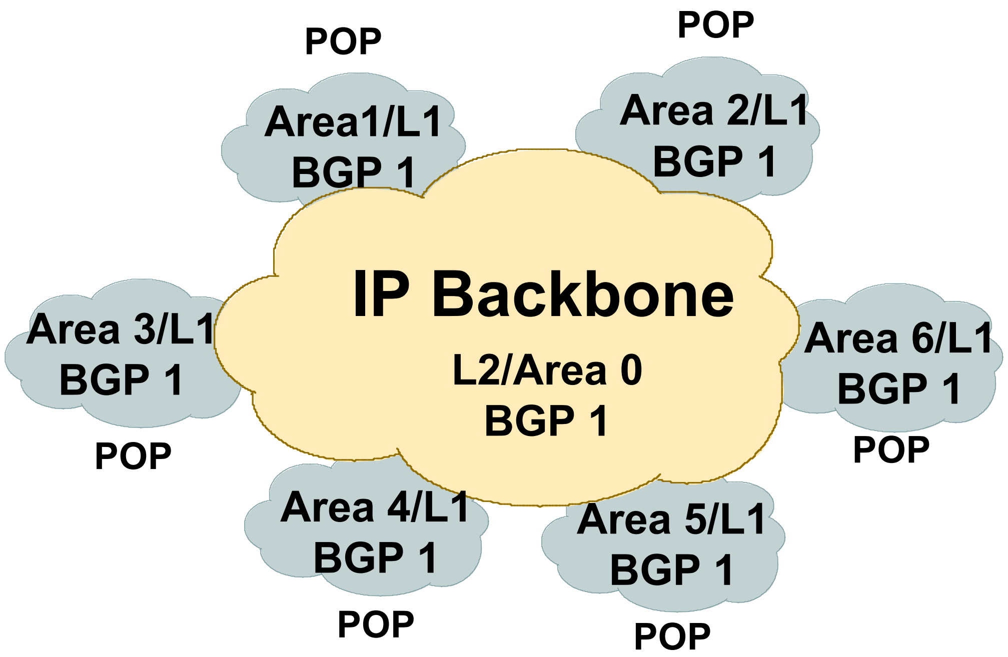
SP Architecture

- Physical address between ABR and Access Router should be in a contiguous blocks
- These physical links should be filtered via Type 3 filtering from area 0 into other areas
- Why? To reduce the size of the routing table within each pop
- Every area will carry only loopback addresses for all routers
- Only NMS station will keep track of those physical links
- Access Router will not carry other PoPs Access Routers physical addresses in the routing table



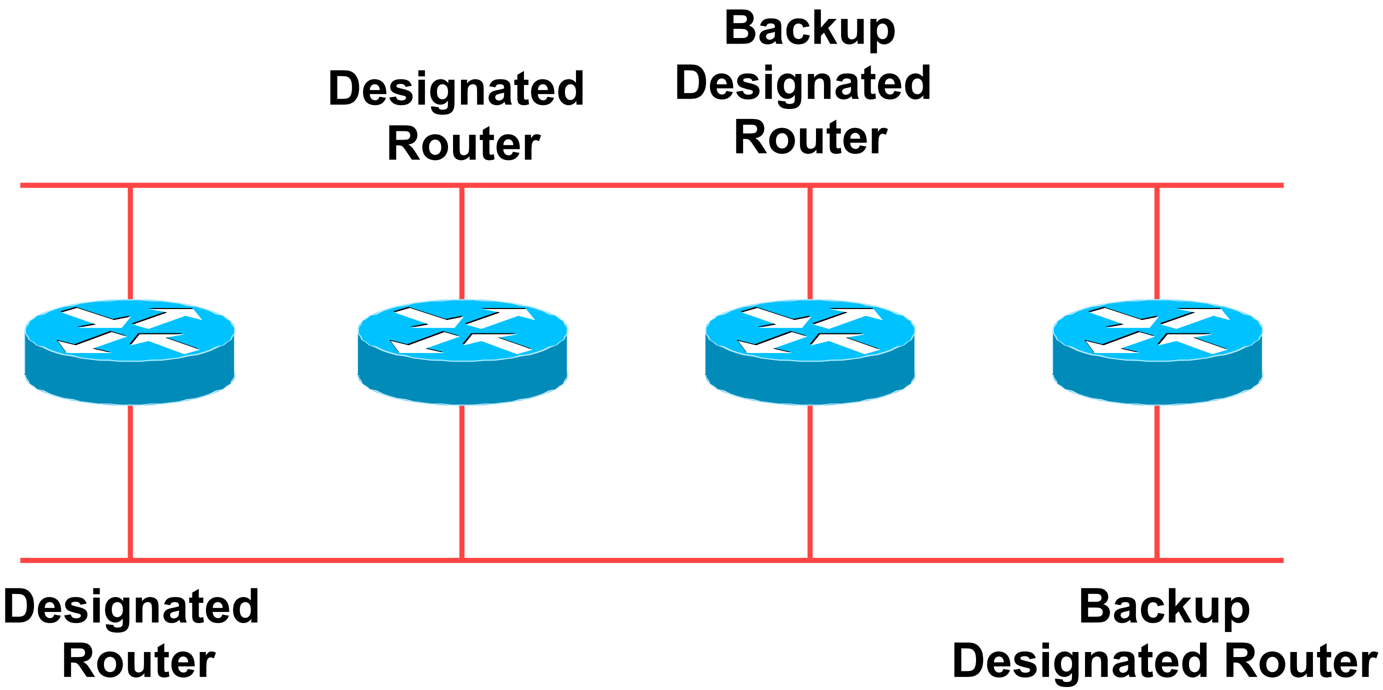
SP Architecture

- Backbone Area (0/L2) will contain all the routes
- This is the most intelligent form of routing and also there will not be too many routes in IGP
- If there are 500 PoPs and every pop contains 4 routers; then instead of having 6K routes you will only have 2K
- This is scalable and hack proof network!



OSPF Design: Designated Router

- There is ONE designated router per multi-access network
  - Generates network link advertisements
  - Assists in database synchronization



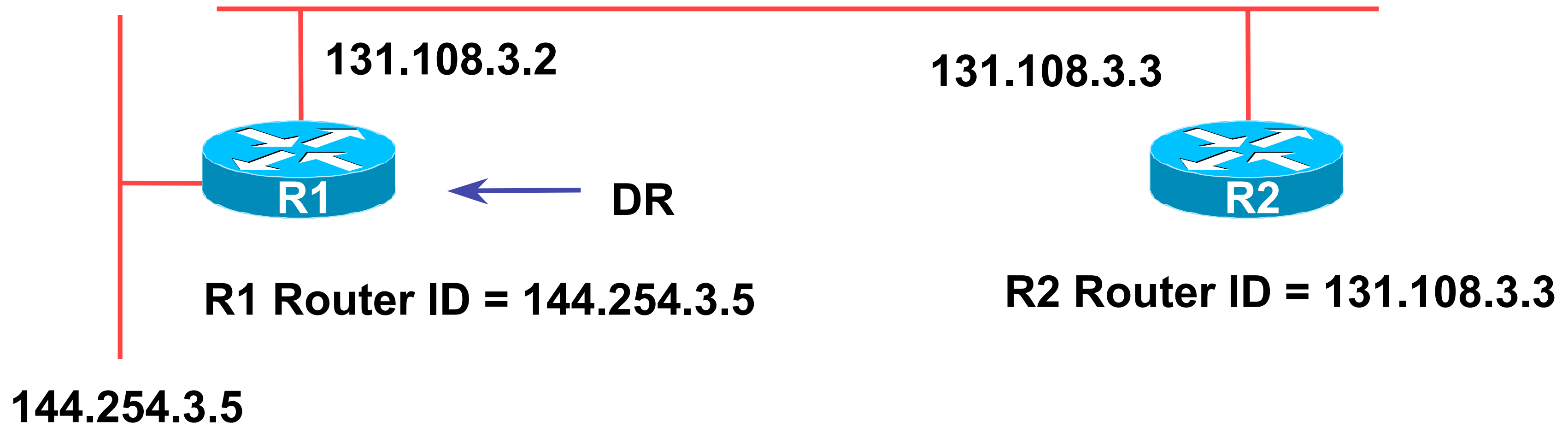
OSPF Design: Designated Router

- Configured priority (per interface)

  ISPs choose router with fastest route processor as DR/BDR

- Else determined by highest router ID

  Router ID is the loopback interface address, if configured, otherwise the highest IP address



Common Problems/Mistakes

- Redistribution from other routing protocols into IGP
  - Don’t do it!
  - If you’re an SP you should not be carrying external information in your IGP
  - Let BGP take care of external reachability
    - Use “next-hop-self” in iBGP
- Poor addressing plans
  - Infrastructure should be addressed per PoP/area
    - Easier to summarise on boundaries
  - Single address block for loopbacks
  - Customer addressing is completely separate!

Tuning OSPF

Tuning OSPF (1)

- Hello/Dead Timers
  - *ip ospf hello-interval 3 (default 10)*
  - *ip ospf dead-interval 15 (default is 4x hello)*
  - This allows for faster network awareness of a failure, and can result in faster reconvergence, but requires more router CPU and generates more overhead
- LSA Pacing
  - *timers lsa-group-pacing 300 (default 240)*
  - This is a great feature; allows grouping and pacing of LSA updates at configured interval; reduces overall network and router impact

Tuning OSPF (2)

- DR/BDR Selection
  - *ip ospf priority 100 (default 1)*
  - This feature should be in use in your OSPF network; forcibly set your DR and BDR per segment so that they are known; choose your most powerful, or most idle routers; try to keep the DR/BDR limited to one segment each
- OSPF Internal Timers
  - *timers spf 2 8 (default is 5 and 10)*
  - Allows you to adjust SPF characteristics; first number sets wait time from topology change to SPF run; second is hold-down between SPF runs; BE CAREFUL WITH THIS COMMAND; if you’re not sure when to use it, it means you don’t need it; default is 95% effective

Tuning OSPF (3)

- OSPF startup
  - *max-metric router-lsa on-startup wait-for-bgp*
  - Avoids black holing traffic on router restart
  - Causes OSPF to announce its prefixes with highest possible metric until iBGP is up and running
  - When iBGP is running, OSPF metrics return to normal, make the path valid
  - (Equivalent to ISIS setting over-load-bit)

Tuning OSPF (4)

- LSA filtering/interface blocking
  - *Per interface:*
  - *ip ospf database-filter all out (no options)*
  - *Per neighbor:*
  - *neighbor 1.1.1.1 database-filter all out (no options)*
  - OSPFs router will flood an LSA out all interfaces except the receiving one; LSA filtering can be useful in cases where such flooding unnecessary (i.e., NBMA networks), where the DR/BDR can handle flooding chores
  - *area <area-id> filter-list <acl>*
  - Filters out specific Type 3 LSAs at ABRs
- Improper use can result in routing loops and black-holes that can be very difficult to troubleshoot

Using OSPF Authentication

- Use authentication; too many people overlook this basic feature
- When using authentication, use the MD5 feature
  - area <area-id> authentication message-digest
  - (whole area)
  - ip ospf message-digest-key 1 md5 <key>
  - (activate per interface)
- Authentication can selectively be disabled per interface with:
  - ip ospf authentication null

Scaling IGPs in ISP Networks

Philip Smith

SANOG 8, Karachi

3rd August 2006