Cisco IOS XR Deployment Best Practices for OSPF/IS-IS and BGP Routing

Contents

DISCLAIMER

This document provides a high-level summary of some established best practice recommendations for OSPF/IS-IS and BGP routing. These recommendations do not represent a Cisco validated design, and due care and attention are required for deployment in any specific operating environment. They should be read in conjunction with the configuration guides and technical documentation for the relevant products that describe in greater detail how these best practice recommendations can be implemented. References in this document to configuration guides and technical documentation for particular products are intended as examples only. Refer the configuration guides and technical documentation for your specific products.

Introduction

This document outlines some established best practices and recommendations for building simplified, efficient, and scalable networks powered by IOS XR routing platforms. This document focuses on specific implementation techniques and feature support options available in IOS XR to help customize OSPF/IS-IS and BGP deployments.

Implementing OSPF

The OSPF protocol, defined in RFC 2328, is an IGP used to distribute routing information within a single Autonomous System. OSPF offers several benefits over other protocols, but a proper design is required to create a scalable and fault-tolerant network.

For more information on OSPF, refer to:

- TechNote on OSPF: https://www.cisco.com/c/en/us/support/docs/ip/open-shortest-path-first-ospf/7039-1.html#anc13
- Configuration Guide for OSPF: https://www.cisco.com/c/en/us/td/docs/routers/asr9000/software/asr9k-r7-6/routing/configuration/guide/b-routing-cg-asr9000-76x/implementing-ospf.html
- Command Reference: https://www.cisco.com/c/en/us/td/docs/routers/asr9000/software/711x/routing/configuration/guide/b-routing-cg-asr9000-711x/implementing-ospf.html?dtid=osscdc000283

Key Concepts

- Hierarchy: A hierarchical network model is a useful high-level tool for designing reliable network infrastructure and helps break complex network design problems into smaller and more manageable areas.
- Modularity: By splitting various functions on a network into modules, the network is much easier to design. Cisco has identified several modules, including the enterprise campus, services block, data center, and Internet edge.
- Resiliency: The network is available in both normal and abnormal conditions. Normal conditions include expected traffic flows, patterns, and scheduled events such as maintenance windows. Abnormal conditions include hardware or software failures, extreme traffic loads, unusual traffic patterns, denial-of-service (DoS) events, and other planned or unplanned events.

- Flexibility: The ability to modify portions of the network, add new services, or increase capacity without going through a significant forklift upgrade (i.e., replacing major hardware devices).

As a general best practice, the network deployment should account for the “span” of the network to contain the routes within a specific boundary and routes that are relevant and required by the routers within a domain for forwarding. Effective use of OSPF areas helps reduce the number of link-state advertisements (LSAs) and other overhead traffic sent across the network. One of the advantages of creating a hierarchy is that this approach helps ensure that the size of the topology database that each router will need to maintain is manageable and conforms to the memory profile of the router.

OSPF Domain and BGP Redistribution

OSPF is designed to carry just a few thousand routes. At a high level, OSPF “areas” are sections of a network where any router knows about the routing capability of every other router in the area. This allows fast convergence when any device has a problem, but at the cost of reduced scalability. As such, OSPF is used in a Service Provider core to provide the base-level connectivity between all the core devices, and all the core devices are configured within the same OSPF area. This is a standard design of an “underlay” network.

By contrast, BGP is designed to carry significantly more routes than most IGPs, like OSPF. Risks associated with redistributing BGP routes into an IGP like OSPF. If a Service Provider requires BGP routes to be redistributed into the IGP domain for any use case, then this needs to be managed by the Service Provider with proper filtering at the Autonomous System Boundary Routers (ASBRs) and with the overload protection configured on the receiving router. If BGP redistribution is unfiltered into an OSPF, every OSPF device in the ASBR will begin receiving routes far beyond its capacity to handle at the same time. Cisco IOS XR routers, for example, will only allow 10,000 BGP routes to be redistributed into OSPF by default. When BGP routes are redistributed into the IGP, it is possible that all routers within the IGP domain may receive these routes, depending upon the IGP design. In accordance with OSPF protocol RFC, any external route being redistributed into OSPF must be distributed to all routers in the OSPF area.

Managing Redistribution into IGP

As a general best practice, redistribution should only be done in a careful and planned manner when there are no other options to learn the routes for reachability that a redistribution function will provide.

As a general practice, you should:

- Avoid redistribution
- Avoid carrying routes in an IGP domain
- Implement BGP for external reachability
- Use IGP to carry next-hop information only; for example, Loopback 0

OSPF Route Redistribution Limitations

The scale of prefixes redistributed from BGP into OSPF is managed with the overload protection (max-lsa) configuration. This is the only protection against leaking a large number of routes into the OSPF domain. In case of redistribution into a single OSPF area, you should implement multiple layers of protection against route redistribution.

Here are some of the options that are available for protection against route redistribution:

- Redistribution filtering using ACL
- Redistribution limit – global setting to prevent more than a specific number of routes from being redistributed. If the filter is removed, the global redistribution limit is the second line of defense and will protect the cores.
- Max-LSA configurations on all devices in the OSPF area – if protections mentioned in the above bullets fail, force the receiving routers to refuse the incoming excessive LSAs.

OSPF Link-State Database Overload Protection

The OSPF Link-State Database Overload Protection feature provides a mechanism at the OSPF level to limit the number of non-self-generated LSAs for a given OSPF process. If other routers in the network have been misconfigured, they may generate a high volume of LSAs, for instance, to redistribute large numbers of prefixes into OSPF. This protection mechanism helps in preventing routers from receiving many LSAs and therefore experiencing CPU and memory shortages.

Feature Behavior

Here is how the feature behaves:

- When this feature is enabled, the router keeps a count of the number of all received (non-self-generated) LSAs.
- When the configured threshold value is reached, an error message is logged.
- When the configured max number of received LSAs is exceeded, the router stops accepting new LSAs.

```
max-lsa <max-lsa-count> <%-threshold-to-log-warning> ignore-count <ignore-count-value> ignore-time <ignore-time-in-minutes> reset-time <time-to-reset-ignore-count-in-minutes>
```

OSPF States

If the received LSAs count is higher than the configured max number after a minute, the OSPF process brings down all adjacencies and clears the OSPF database. This state is called the ignore state. In this state, all OSPF packets received on all interfaces belonging to the OSPF instance are ignored, and no OSPF packets are generated on the interfaces. The OSPF process remains in the ignore state for the duration of the configured ignore-time (default is 5 minutes). When the ignore-time expires, the OSPF process returns to normal operation and builds adjacencies on all its interfaces.

If the LSA count exceeds the max number as soon as the OSPF instance returns from the ignore state, the OSPF instance can oscillate endlessly between its normal state and the ignore state. To prevent this infinite oscillation, the OSPF instance counts how many times it has been in the ignore state. This counter is called the ignore-count. If the ignore-count (default ignore-count is 5) exceeds its configured value, the OSPF instance permanently remains in the ignore state.

You must issue the clear ospf command to return the OSPF instance to its normal state.

The ignore-count is reset to zero if the LSA count does not exceed the maximum number again during the time configured by the reset-time keyword.

If you use the warning-only keyword, the OSPF instance never enters the ignore state. When the LSA count exceeds the maximum number, the OSPF process logs an error message, and the OSPF instance continues in its normal state operation.
There is no default value for max-lsa. The limit is checked only if it is specifically configured.

Once max-lsa is configured, other parameters can have default values:

- default %-threshold-to-log-warning - 75%
- default ignore-count-value – 5
- default ignore-time-in-minutes - 5 minutes
- default time-to-reset-ignore-count - 10 minutes

Here is an example of the implementation which shows how to configure the OSPF instance to accept 12000 non-self-generated LSAs and 1000 non-self-generated LSAs in VRF V1.

```
RP/0/RSP0/CPU0:router# configure

RP/0/RSP0/CPU0:router(config)# router ospf 0

RP/0/RSP0/CPU0:router(config-ospf)# max-lsa 12000

RP/0/RSP0/CPU0:router(config-ospf)# vrf V1

RP/0/RSP0/CPU0:router(config-ospf)# max-lsa 1000
```

The following example shows how to display the current status of the OSPF instance.

```
RP/0/RSP0/CPU0:router# show ospf 0

    Routing Process "ospf 0" with ID 10.0.0.2

    NSR (Non-stop routing) is Disabled

    Supports only single TOS(TOS0) routes

    Supports opaque LSA

    It is an area border router

    Maximum number of non-self-generated LSA allowed 12000

       Current number of non self-generated LSA 1

       Threshold for warning message 75%

       Ignore-time 5 minutes, reset-time 10 minutes

       Ignore-count allowed 5, current ignore-count 0
```

Implementing BGP

BGP address families make the BGP a “multiprotocol” routing protocol. It is highly recommended that you understand how the address families are used to create scalable topologies that are easy to implement and manage. Using address families, the operator can create different topologies for different technologies, for example, EVPN, Multicast, and so on.

For more information on BGP, see the BGP configuration guide: https://www.cisco.com/c/en/us/td/docs/iosxr/ncs5500/bgp/76x/b-bgp-cg-ncs5500-76x/implementing-bgp.html

BGP and BFD

BGP convergence in a Service Provider network is important to meet customer expectations for building resilient and fault-tolerant networks. By default, BGP has a Keepalive timer of 60 seconds and a Hold timer of 180 seconds. All this means that BGP will be very slow to converge unless there is help available from supporting protocols. BFD Bi-directional Forwarding (BFD) is one such protocol that is designed to help the client protocols converge faster. With BFD, protocols can converge within seconds.

Additional Information

- This guide provides conceptual and configuration information for BFD: https://www.cisco.com/c/en/us/td/docs/iosxr/ncs5500/routing/76x/b-routing-cg-ncs5500-76x/implementing-bfd.html
- This whitepaper presents a Service Provider-centric view on fast convergence using BFD on the Cisco NCS 5500 and Cisco Network Convergence System 500 Series routers: https://xrdocs.io/ncs5500/tutorials/bfd-architecture-on-ncs5500-and-ncs500/
- For a deeper dive into using BFD on Bundle interfaces and implementing Multipath and MultiHop BFD, refer to https://xrdocs.io/ repository.

BGP Slow peer detection

A slow peer is a peer that cannot keep up with the rate at which the router is generating update messages over a prolonged period (in the order of minutes) in an update group. When a slow peer is present in an update group, the number of formatted updates pending transmission builds up. When the cache limit is reached, the group does not have any more quotas to format new messages. For a new message to be formatted, some existing messages must be transmitted using the slow peer and then removed from the cache. The rest of the members of the group that are faster than the slow peer and have completed transmission of the formatted messages will not have anything new to send, even though there may be newly modified BGP networks waiting to be advertised or withdrawn. This effect of blocking the formatting of all the peers in a group when one of the peers is slow in consuming updates is the "slow peer" problem.

Events that cause a significant churn in the BGP table (such as connection resets) can cause a brief spike in the rate of update generation. A peer that temporarily falls behind during such events but quickly recovers after the event is not considered a slow peer. For a peer to be marked as slow, it must be incapable of keeping up with the average rate of generated updates over a more extended period (in the order of a few minutes).

BGP Slow peer may be caused due to:

- Packet Loss or high traffic on the link to the peer.
- A BGP peer could be heavily loaded in terms of CPU and hence cannot service the TCP connection at the required speed.
- In this case, platform hardware capability and the offered load must be checked.
- Throughput issues with the BGP connection
- For more information on BGP Slow peer detection, see: https://www.cisco.com/c/en/us/td/docs/iosxr/ncs5500/bgp/76x/b-bgp-cg-ncs5500-76x/implementing-bgp.html#concept_ir5_j4w_p4b

Here are some mitigations and best practices for managing slow peers:

■ End-to-end QoS, which reserves bandwidth for BGP control plane traffic during congestion.

■ Use of correct and appropriate MSS / MTU values using BGP PMTUD and/or TCP MSS settings.

■ Use the correct hardware and minimize the number of routes with respect to the hardware.

Slow-peer detection is enabled by default in Cisco IOS XR starting from Release 7.1.2. Slow peers are peers which are slow to receive and process the inbound BGP updates and acknowledge the updates to the sender. If the slow peer is participating in the same update group as other peers, this can slow down the update process for all peers. In this release, when IOS XR detects a slow peer, it will create a syslog that has the details about the specific peer.

Fast Convergence using BGP Prefix Independent Convergence

For BGP prefixes, fast convergence is achieved using BGP Prefix Independent Convergence (PIC), in which BGP calculates an alternate best path and primary best path and installs both paths in the routing table as primary and backup paths.

If the BGP next-hop remote becomes unreachable, BGP immediately switches to the alternate path using BGP PIC instead of recalculating the path after the failure.

If the BGP next-hop remote PE is alive, but there is a path failure, IGP TI-LFA FRR handles fast re-convergence to the alternate path, and BGP updates the IGP next-hop for the remote PE.

BGP PIC is configured under VRF address-family for fast convergence of VPN Prefixes if a remote PE becomes unreachable.

For more information on BGP Prefix Independent Convergence, see:
https://www.cisco.com/c/en/us/td/docs/iosxr/ncs5500/bgp/76x/b-bgp-cg-ncs5500-76x/bgp-pic.html

BGP Security with BGP Flowspec

BGP Flowspec, in a nutshell, is a feature that allows you to receive IPv4/IPv6 traffic flow specifications (source X, destination Y, protocol UDP, source port A, and so on) and actions that need to be taken on that traffic (such as drop, police, or redirect) via BGP update.
Inside the BGP update, the Flowspec matching criteria are represented by BGP NLRI, and BGP extended communities represent the actions.

This feature is based on RFC 5575 and can be used to help mitigate DDoS attacks. When a certain host inside of a network is being attacked, we can send a Flowspec update to edge routers so that attack traffic can be policed or dropped, or even redirected elsewhere, maybe to an appliance that can clean the traffic (filter out the ‘bad’ traffic and forward only the ‘good’ traffic toward the affected host).

Once Flowspecs are received by a router and programmed in applicable line cards, any active L3 ports on those line cards will start processing ingress traffic according to Flowspec rules.

For more information on implementing BGP FlowSpec, see:

- BGP FlowSpec whitepaper: https://xrdocs.io/ncs5500/tutorials/bgp-flowspec-on-ncs5500/
- BGP Configuration Guide: https://www.cisco.com/c/en/us/td/docs/iosxr/ncs5500/bgp/76x/b-bgp-cg-ncs5500-76x/implementing-bgp.html#concept_uqv_bxq_h2b

BGP Maximum Prefix Feature

The Maximum-Prefix feature is useful when, at a change of outbound policy at the remote peering site, a router starts to receive more prefixes than the resources of the peering router can handle but also to protect resources or the internal BGP peers where these external prefixes will be forwarded. Such resource overhead could be disruptive.

BGP maximum-prefix feature imposes a maximum limit on the number of prefixes that are received from a neighbor for a given address family. By default, whenever the number of prefixes received exceeds, the maximum number configured, the BGP session sends a cease notification to the neighbor and the session gets terminated. One address-family crossing the maximum-prefix will bring the whole BGP session down, impacting all other address-families enabled in that BGP session.

This feature is commonly used for external BGP peers to protect the internal infrastructure of a Service Provider. It serves as a guardrail to prevent router resource depletion that could be caused by a misconfiguration, either locally or on the remote neighbor. Configuring maximum-prefix is highly recommended to protect against local or remote misconfigurations that could trigger route table flooding. This also protects against prefix de-aggregation attacks.

BGP maximum-prefix configuration should be explicitly enabled on all eBGP routers to limit the number of prefixes that it should receive from a particular neighbor, whether customer or peering AS. It is recommended that the operator configures an acceptable margin of extra prefixes that the system may be able to sustain after careful evaluation of the available system memory. It should be noted that there is no one size fits all configuration that may be applied to all the routers and the threshold should be carefully adjusted based on the role of the device in the network. For instance, if BGP maximum-prefix is to be configured on IBGP neighbors, then the maximum-prefix value must be lower on the neighbors configured on the route-reflector vs. that for the neighbors configured on the route-reflector-clients. The route-reflector aggregates prefixes received from multiple peering routers and then re-advertises the full table to the route-reflector-clients. Therefore, the route-reflector will advertise more prefixes to its clients than what it receives from each individual peer. Similarly, a peering router may also re-advertise more prefixes towards the route-reflector than what it receives from each individual external peer.

In summary, it is recommended to carefully review and configure the appropriate action to take when the maximum-prefix threshold is reached on a production device. Some attributes of the maximum-prefix command options are described as follows:

- When a BGP session is explicitly configured with the maximum-prefix feature without any additional keywords (such as warning-only or potentially restart), the session will be torn down as default behavior. The default action of peer session being brought down with no auto-recovery could lead to a prolonged outage within the core.

```
%ROUTING-BGP-5-ADJCHANGE_DETAIL : neighbor 10.10.10.10 Down - BGP
Notification received, maximum number of prefixes reached (VRF:
default; AFI/SAFI: 1/1, 1/128, 2/4, 2/128, 1/133, 2/133) (AS: 65000) "
%ROUTING-BGP-5-NBR_NSR_DISABLED_STANDBY : NSR disabled on neighbor
10.10.10.10 on standby RP due to Peer exceeding maximum prefix limit
(VRF: default)
```

- Configuring the discard extra paths option drops all excess prefixes received from the neighbor above the configured maximum value threshold. This drop does not result in session flap. The benefits of this option include limiting the BGP process memory utilization and stopping the flapping of the peers within the core network. However, this may result in forwarding loops for the prefixes being discarded as the forwarding entries may become inconsistent between routers in the network.
- When using add-path, the configured maximum-prefix value applies to paths instead of prefixes as the NLRI is made of the prefix and the path attributes. Refer to the following command reference for more information:

https://www.cisco.com/c/en/us/td/docs/iosxr/ncs5500/bgp/b-ncs5500-bgp-cli-reference/b-ncs5500-bgp-cli-reference_chapter_01.html

Recommendation: Carefully evaluate the following options when configuring the maximum-prefix command:

- No explicit action defined: Router will bring down the session and keep the BGP neighbor relationship down until the operator manually clears the BGP session. [clear bgp command]
- Restart [time-interval]: Bring down the session and attempt an automatic restart of the BGP session periodically after a configured timer. This will succeed if the remote peer stops advertising the excess prefixes otherwise the BGP session will go down again (hence causing periodic instability).
- Discard-extra-paths: With the discard-extra-paths option, the BGP session stays up but prefixes over the maximum-prefix limit are discarded. This option does not impact other address-families where the maximum-prefix has not been reached and ensures that local resources are not exhausted but this may lead to forwarding loops for the prefixes being discarded. Note that the discard extra paths option cannot co-exist with the soft reconfig knob.
- Warning-only: Log a warning only when the threshold is reached so that the operator can take manual actions to clear the condition.

For more information, please refer to the Routing Configuration Guide as follows:

https://www.cisco.com/c/en/us/td/docs/routers/asr9000/software/asr9k-r7-3/routing/configuration/guide/b-routing-cg-asr9000-73x/implementing-BGP.html#concept_5AF38064B1D044B7B5F439C10BCF9808

Best Practices and Recommendations

The following list provides an overview of general best practices and recommendations, listed in no specific order:

- Network audit for the general health of the system. Start with a configuration audit and move sequentially from interface configurations up to routing and services.
- Have a monitoring strategy in place. While SNMP is standard practice, consider deploying more robust and descriptive techniques using Streaming Telemetry. Refer to the following whitepaper for best practice recommendations on implementing Telemetry on an IOS XR Router: https://xrdocs.io/telemetry/

OSPF

Here are general best practices and recommendations for OSPF:

- Implement route summarization for intra-area routes for OSPF.
- Configure the router-ID explicitly inside OSPF as one of the OSPF-enabled loopback addresses.
- Design a hierarchical network to limit the LSAs within an area for OSPF. Keep the number of ABRs for an area within a reasonable range (~3 to 4).
- Implement OSPF “max-lsa” configuration for OSPF, or equivalent, to limit the LSAs in the database to effectively use the system’s memory.
- Limit the maximum number of routes that can be distributed from BGP to OSPF. In IOS-XR, the default limit is 10K.
- Use Route Policy (RPL) to redistribute the routes into OSPF.
- Summarize the inter-area route and external type 5 routes wherever applicable.
- Use of authentication when necessary.
- Always use NSF and NSR.
- Configure redistribution filtering at the source instead of the destination.
- Use passive interface where applicable.
- OSPF should only carry Loopback and Router-Interface routes – remove any other BGP-to-OSPF redistribution.
- Consider moving each primary hub into its own area (NSSA).
- Use BFD for fast failure detection vs. the aggressive routing protocol timers.
- Do not use the mtu-ignore command as much as possible.
- Consider using IGP-LDP sync in an MPLS environment to avoid sending traffic on an unlabeled path.
- Consider scale within supported platform limits (number of prefixes, number of labels, ECMP, Number of Areas, and so on).
- Avoid mutual redistribution at multiple points.
- Configure administrative distance so that each prefix native to each protocol or process is reached via the corresponding domain’s protocol or process.
- Control the prefixes (using distance or prefix-list combination) so that the same prefix is not advertised back to the originating domain.
- Although OSPF process ID has local significance to the router, it is recommended to have the same process ID for all the routers in the same OSPF domain. This improves configuration consistency and eases automatic configuration tasks.

■ When configuring OSPF for hub-and-spoke environments, design the OSPF areas with a smaller number of routers.
■ Configure the OSPF auto-cost reference bandwidth throughout the OSPF domain to the highest bandwidth link in the network.
■ From a design perspective, we recommend that you implement IGP peering with domains under the same administrative or operational controls to help avoid an unplanned or rogue IGP update propagating across the network. This should allow for better serviceability and ease of troubleshooting in case errors occur. In case a large IGP domain is a business necessity, plan on using BGP in those cases to limit the number of routes in the IGP network domain.
■ If you need end-to-end MPLS connectivity, continue using hierarchy/segmentation and use options such as RFC3107 BGP-LU or inter-domain path computation via PCE, or select redistribution/leaking with policy as a last resort.
■ OSPF Shortest Path First Throttling feature may be used to configure SPF scheduling in millisecond intervals and to potentially delay the SPF calculations during network instability.
■ OSPF SPF Prefix Prioritization feature enables an administrator to converge important prefixes faster during route installation.

IS-IS

Here are general best practices and recommendations for IS-IS:

■ If you run a flat single-level network, think about the scale. Configure all routers as L2 only. By default router is L1-L2, and leaking routing information from L1 to L2 is enabled by default. This could lead to all routers leaking all L1 routes to L2, bloating the link-state database.
■ If you run a multi-level (multiple areas) network, ensure the Layer-3 topology follows the ISIS hierarchy. Do not create backdoor links between L1 areas.
■ If you run a multi-level (multiple areas) network, ensure L1 and L2 routers are connected via both L1 and L2 areas. This does not require multiple physical or virtual connections between them; run the link between L1 and L2 routers as an L1/L2 circuit.
■ If you run a multi-level (multiple areas) network, summarize what can be summarized – for example, in the case of MPLS, the loopback of PE routers need to be propagated between areas, but infrastructure link addresses do not.
■ Create and follow the proper addressing plan if possible. That allows summarization and helps scale.
■ Set the LSP lifetime to a maximum of 18 hours.
■ Avoid redistribution by any means. Redistribution is complex and needs to be managed manually to avoid routing loops. Use multiarea/level design if possible.
■ If you must use redistribution, use route tagging during redistribution and "distribute-list in" filtering based on tags to manage it. Summarize during redistribution if possible.
■ Configure interfaces as "point-to-point" whenever possible. This improves the performance and scalability of the protocol.
■ Do not use ISIS in highly meshed topology. Link-state protocols behave poorly in highly meshed environments.
■ Configure a high default metric in the ISIS address-family submode. This prevents newly added links from attracting traffic if they are inadvertently configured without a metric.
■ Configure "log adjacency changes" to help with connection troubleshooting.

■ Use "metric-style wide" under the ISIS address-family ipv4 sub-mode. Narrow metrics aren't very useful and don't support features like segment-routing or flex-algo.
■ If you are using SR-MPLS TI-LFA remember to add "ipv4 unnumbered mpls traffic-eng Loopback0" to the configuration to allow ISIS to allocate TE tunnels when required.
■ Let the "lsp-gen-interval" and "spf-interval" configurations default unless you are sure that faster native convergence is required. With TI-LFA native convergence isn't as critical, since fast-reroute will handle single topology changes in 50 ms or less.
■ If you modify "lsp-gen-interval" or "spf-interval" do not use an initial delay shorter time than 50 ms.
■ In most cases, "set-overload-bit" is a better choice than "max-metric" as it is an atomic change that is supported by fast-reroute.
■ Use cryptographic authentication for Hellos (hello-password) and LSPs (lsp-password). Keychains provide the most flexibility and can
accommodate hitless key rollovers.
■ Configure "nsf cisco" for hitless authentication of ISIS process restarts and SMU installation. Despite the name, this provides better interoperability with other vendors than "nsf ietf".
■ On a platform with dual RPs, ALSO configure "nsr" to handle RP switchovers.
■ Use "group" and "apply-group" templates to configure repeated configuration sections. This is less error prone and easier to change if needed.
■ In a multi-level network, carefully consider whether you need to use "propagate" to leak prefixes down from Level 2 to Level 1. This can limit scalability, and often the level-1 default route provided by the Attached bit is sufficient.
■ If you are using multiple ISIS instances in the same VRF, consider configuring unique "distance" values for them. This will make route installation in the RIB more deterministic if each has a route to the same prefix.
■ Use BFD for quick link-down detection. With BFD providing this function, the ISIS hello-interval may be safely increased to improve scalability.

BGP

Here are general best practices and recommendations for BGP:

■ Use NSR and NSF / graceful restart with carefully tuned timers depending on the expected scale.
■ Configure BGP using the ‘always UP’ loopback interface, not the physical interface for IBGP peering.
■ Do not redistribute BGP (high-volume) routes into IGP (comparatively low volume) and vice-versa without proper RPL, restricting the number of redistributed routes from BGP to an IGP (OSPF/ISIS).
■ Doing BGP to IGP redistribution without a proper, well-tested policy (ACL) may cause resource (memory) exhaustion on the router.
■ Use of summary routes in BGP to decrease the routing table size and use of memory. Aggregate routes with summary-only wherever it makes sense
■ Use route filtering for advertising and receiving routes efficiently, especially in BGP.
■ We recommend using Route-Reflector (RR) and confederation to scale up the network.
■ Some of the Route Reflector design considerations are:

- Path Scale increases based on the number of clients/non-clients.
- In hierarchical RRs, use the same cluster-id at the same level (redundant RR) for loop prevention and scale.
- Control MTU within the BGP path or use PMTUD protocol to adjust BGP MSS automatically.
- Use BFD or tune BGP timers for faster fault detections.
- BGP scale is as per configuration and use-case, and no one size fits all. You need to have a good idea about:
- route scale
  - path scale (with soft reconfiguration, it will increase)
  - attribute scale
- If the add-path is configured, it consumes more memory.
- Careful understanding of the BGP neighbor policies:
  - pass-all (especially at a boundary router) may cause havoc as the memory scale will shoot up.
  - Use policy constructs that will avoid regular expression matches in RPL.
- With NSR, standby RP will use about 30% more virtual memory than active. Keep this in consideration if there is a standby.
- Look out for continuous churn in a significant number of routes (version bumps). This may keep the update generation memory in high watermark.
- Protect peers with max-prefix knob.
- Use next-hop-trigger delay parameters according to scale and convergence goals.
- In the network design, try to avoid new attributes. Unique attributes lead to inefficient packing and result in more BGP updates.
- Configuring multi-path across the network can lead to forwarding loops. Use with care.
- Use table-policy to avoid route install to rib if RR is not inline-RR (no next-hop-self)

Monitor System Memory for Routing Processes

No device has infinite resources – if we send an infinite number of routes to a device, the device must choose how it fails. The routers will attempt to service all the routes until the memory limits are exhausted, and this may cause all routing protocols and processes to fail.

Each process in the core router does have an “RLIMIT” defined. The “RLIMIT” is the amount of system memory each process is allowed to consume.

This section describes some standard techniques to monitor and check your system memory used by the BGP process.

Process Memory

Shows the amount of memory consumed by a process.

```
RP/0/RP0/CPU0:NCS-5501#show proc memory

JID      Text(KB)   Data(KB)  Stack(KB) Dynamic(KB) Process

------ ---------- ---------- ---------- ----------- -----------------------------
-

1150          896     368300        136       33462 lspv_server

380           316    1877872        136       32775 parser_server
```

```
1084         2092    2425220         136       31703 bgp

1260         1056    1566272         160       31691 ipv4_rib

1262         1304    1161960         152       28962 ipv6_rib

1277         4276    1479984         136       21555 pim6

1301           80     227388         136       21372 schema_server

1276         4272    1677244         136       20743 pim

250           124     692436         136       20647 invmgr_proxy

1294         4540    2072976         136       20133 l2vpn_mgr

211           212     692476         136       19408 sdr_invmgr

1257            4     679752         136       17454 statsd_manager_g
```

Each process is allocated a maximum amount of memory that it is allowed to consume. This is defined as the rlimit.

```
RP/0/RP0/CPU0:NCS-5501#show proc memory detail

JID         Text       Data       Stack      Dynamic    Dyn-Limit  Shm-Tot
Phy-Tot    Process

================================================================================
===========================

1150        896K       359M       136K        32M      1024M        18M
24M     lspv_server

1084          2M      2368M       136K        30M      7447M        43M
69M     bgp

1260          1M      1529M       160K        30M      8192M        38M
52M     ipv4_rib

380         316K      1833M       136K        29M      2048M        25M
94M     parser_server

1262          1M      1134M       152K        28M      8192M        22M
31M     ipv6_rib

1277          4M      1445M       136K        21M      1024M        18M
41M     pim6

1301         80K       222M       136K        20M       300M         5M
33M     schema_server

1276          4M      1637M       136K        20M      1024M        19M
41M     pim

250         124K       676M       136K        20M      1024M         9M
31M     invmgr_proxy

1294          4M      2024M       136K        19M      1861M        48M
66M     l2vpn_mgr

211         212K       676M       136K        18M       300M         9M
29M     sdr_invmgr
```

```
1257          4K       663M       136K         17M      2048M        20M
39M     statsd_manager_g

288           4K       534M       136K         16M      2048M        15M
33M     statsd_manager_l

...
```

Top Memory Consumers

```
RP/0/RP0/CPU0:NCS-5501#show memory-top-consumers

###################################################################

 Top memory consumers on 0/0/CPU0 (at 2022/Apr/13/15:54:12)

###################################################################

    PID             Process   Total(MB)    Heap(MB)      Shared(MB)

   3469          fia_driver         826      492.82             321

   4091             fib_mgr         175     1094.43             155

   3456                 spp         130        9.68             124

   4063     dpa_port_mapper         108        1.12             105

   3457              packet         104        1.36             101

   5097           l2fib_mgr          86       52.01              71

   4147           bfd_agent          78        6.66              66

   4958         eth_intf_ea          66        4.76              61

   4131       optics_driver          62      141.23              22

   4090             ipv6_nd          55        4.13              49

###################################################################

 Top memory consumers on 0/RP0/CPU0 (at 20xx/MMM/HH:MM:SS)

###################################################################

    PID             Process   Total(MB)    Heap(MB)      Shared(MB)

   3581                 spp         119        9.62             114

   4352     dpa_port_mapper         106        2.75             102

   4494             fib_mgr          99        7.71              90

   3582              packet          96        1.48              94

   3684       parser_server          95       64.27              25

   8144          te_control          71       15.06              55

   8980                 bgp          70       27.61              44

   7674           l2vpn_mgr          67       23.64              48

   8376      mibd_interface          65       35.28              28

   3608                 gsp          65       15.75              48
```

Total Memory – Used and Available

System components have a fixed amount of memory available.

```
RP/0/RP0/CPU0:NCS-5501#show memory summary location all

node:      node0_0_CPU0

------------------------------------------------------------------

 Physical Memory: 8192M total (6172M available)

 Application Memory : 8192M (6172M available)

 Image: 4M (bootram: 0M)

 Reserved: 0M, IOMem: 0M, flashfsys: 0M

 Total shared window: 226M

node:      node0_RP0_CPU0

------------------------------------------------------------------

 Physical Memory: 18432M total (15344M available)

 Application Memory : 18432M (15344M available)

 Image: 4M (bootram: 0M)

 Reserved: 0M, IOMem: 0M, flashfsys: 0M

 Total shared window: 181M
```

The shared memory window provides information on the shared memory allocations on the system.

```
RP/0/RP0/CPU0:NCS-5501#show memory summary detail location 0/RP0/CPU0

node:      node0_RP0_CPU0

------------------------------------------------------------------

 Physical Memory: 18432M total (15344M available)

 Application Memory : 18432M (15344M available)

 Image: 4M (bootram: 0M)

 Reserved: 0M, IOMem: 0M, flashfsys: 0M

 Shared window soasync-app-1: 243.328K

 Shared window soasync-12: 3.328K

 ...

 Shared window rewrite-db: 272.164K

 Shared window l2fib_brg_shm: 139.758K

 Shared window im_rules: 384.211K

 Shared window grid_svr_shm: 44.272M

 Shared window spp: 86.387M

 Shared window im_db: 1.306M
 Total shared window: 180.969M
 Allocated Memory: 2.337G
 Program Text: 127.993T
 Program Data: 64.479G
 Program Stack: 2.034G
System RAM:    18432M (   19327352832)
Total Used:     3088M (    3238002688)
 Used Private:     0M (             0)
 Used Shared:   3088M (    3238002688)
```

You can check the participant processes with a shared memory window.

```
RP/0/RP0/CPU0:NCS-5501#sh shmwin spp participants list
Data for Window "spp":
-----------------------------
List of current participants:-
NAME                                     PID             JID             INDEX
spp                                      3581            113             0
packet                                   3582            345             1
ncd                                      4362            432             2
netio                                    4354            234             3
nsr_ping_reply                           4371            291             4
aib                                      4423            296             5
ipv6_io                                  4497            430             6
ipv4_io                                  4484            438             7
fib_mgr                                  4494            293             8
...
snmpd                                    8171            1002            44
ospf                                     8417            1030            45
mpls_ldp                                 7678            1292            46
bgp                                      8980            1084            47
cdp                                      9295            337             48
RP/0/RP0/CPU0:NCS-5501#sh shmwin soasync-1 participants list
Data for Window "soasync-1":
-----------------------------
List of current participants:-
```

```
NAME                                        PID             JID
INDEX

tcp                                         5584            168
0

bgp                                         8980            1084
```

Resource Monitoring and Watchdogs

Memory utilization is monitored through a system watchdog in cXR and with Resmon in eXR.

```
RP/0/RP0/CPU0:NCS-5501#show watchdog memory-state

---- node0_RP0_CPU0 ----

Memory information:

    Physical Memory     : 18432.0   MB

    Free Memory         : 15348.0   MB

    Memory State        :   Normal

RP/0/RP0/CPU0:NCS-5501#

RP/0/RP0/CPU0:NCS-5501#show watchdog threshold memory defaults location
0/RP0/CPU0

---- node0_RP0_CPU0 ----

 Default memory thresholds:

 Minor:    1843      MB ß--10%

 Severe:   1474      MB ß--8%

 Critical:  921.599 MB  ß--5%

Memory information:

    Physical Memory     : 18432.0   MB

    Free Memory         : 15340.0   MB

    Memory State        :   Normal

RP/0/RP0/CPU0:NCS-5501#

RP/0/RP0/CPU0:NCS-5501(config)#watchdog threshold memory minor ?

  <5-40>  memory consumption in percentage
```

A warning is printed if the thresholds are crossed.

```
RP/0/RP0/CPU0:Feb 17 23:30:21.663 UTC: resmon[425]: %HA-HA_WD-4-MEMORY_ALARM :
Memory threshold crossed: Minor with 1840.000MB free. Previous state: Normal

RP/0/RP0/CPU0:Feb 17 23:30:21.664 UTC: resmon[425]: %HA-HA_WD-6-
TOP_MEMORY_USERS_INFO : Top 5 consumers of system memory (1884160 Kbytes free):

RP/0/RP0/CPU0:Feb 17 23:30:21.664 UTC: resmon[425]: %HA-HA_WD-6-
TOP_MEMORY_USER_INFO : 0: Process Name: bgp[0], pid: 7861, Heap usage: 12207392
kbytes.
```

```
RP/0/RP0/CPU0:Feb 17 23:30:21.664 UTC: resmon[425]: %HA-HA_WD-6-
TOP_MEMORY_USER_INFO : 1: Process Name: ipv4_rib[0], pid: 4726, Heap usage:
708784 kbytes.

RP/0/RP0/CPU0:Feb 17 23:30:21.664 UTC: resmon[425]: %HA-HA_WD-6-
TOP_MEMORY_USER_INFO : 2: Process Name: fib_mgr[0], pid: 3870, Heap usage: 584072
kbytes.

RP/0/RP0/CPU0:Feb 17 23:30:21.664 UTC: resmon[425]: %HA-HA_WD-6-
TOP_MEMORY_USER_INFO : 3: Process Name: netconf[0], pid: 9260, Heap usage: 553352
kbytes.

RP/0/RP0/CPU0:Feb 17 23:30:21.664 UTC: resmon[425]: %HA-HA_WD-6-
TOP_MEMORY_USER_INFO : 4: Process Name: netio[0], pid: 3655, Heap usage: 253556
kbytes.

LC/0/3/CPU0:Mar  8 05:48:58.414 PST: resmon[172]: %HA-HA_WD-4-MEMORY_ALARM :
Memory threshold crossed: Severe with 600.182MB free. Previous state: Normal

LC/0/3/CPU0:Mar  8 05:48:58.435 PST: resmon[172]: %HA-HA_WD-4-
TOP_MEMORY_USERS_WARNING : Top 5 consumers of system memory (624654 Kbytes
free):

LC/0/3/CPU0:Mar  8 05:48:58.435 PST: resmon[172]: %HA-HA_WD-4-
TOP_MEMORY_USER_WARNING : 0: Process Name: fib_mgr[0], pid: 5375, Heap usage
1014064 Kbytes.

LC/0/3/CPU0:Mar  8 05:48:58.435 PST: resmon[172]: %HA-HA_WD-4-
TOP_MEMORY_USER_WARNING : 1: Process Name: ipv4_mfwd_partner[0], pid: 5324, Heap
usage 185596 Kbytes.

LC/0/3/CPU0:Mar  8 05:48:58.435 PST: resmon[172]: %HA-HA_WD-4-
TOP_MEMORY_USER_WARNING : 2: Process Name: nfsvr[0], pid: 8357, Heap usage 183692
Kbytes.

LC/0/3/CPU0:Mar  8 05:48:58.435 PST: resmon[172]: %HA-HA_WD-4-
TOP_MEMORY_USER_WARNING : 3: Process Name: fia_driver[0], pid: 3542, Heap usage
177552 Kbytes.

LC/0/3/CPU0:Mar  8 05:48:58.435 PST: resmon[172]: %HA-HA_WD-4-
TOP_MEMORY_USER_WARNING : 4: Process Name: npu_driver[0], pid: 3525, Heap usage
177156 Kbytes.
```

Some processes might take specific actions based on the watchdog memory state. For example, BGP does the following:

- in the minor state, BGP stops bringing up new peers
- in the severe state, BGP gradually brings down some peers.
- in a critical state, BGP process shuts down.

Processes can be configured to register for memory state notifications.

```
Show watchdog oor-aware-process
```

Users can disable automatic process shutdown due to watchdog timeout.

```
watchdog restart memory-hog disable
```

Where to find more information?

- Cisco IOS XR Blogs and Whitepapers repository (xrdocs.io)
    - Core Fabric Design: https://xrdocs.io/design/blogs/latest-core-fabric-hld : This whitepaper discusses the recent trends and evolution in core backbone networks.
    - Peering Fabric Design: https://xrdocs.io/design/blogs/latest-peering-fabric-hld : This whitepaper provides a comprehensive overview of the challenges and best practice recommendations for peering design with a focus on network simplification.
- Configuration Guide Reference: Implementing BGP

https://www.cisco.com/c/en/us/td/docs/iosxr/ncs5500/bgp/710x/b-bgp-cg-ncs5500-710x/implementing-bgp.html

Feature Enhancements

Autonomous System Boundary Router Isolation and Adjacency Control for LSA Overflows	Introduced in 7.10.1 on NCS 5500 fixed port routers: NCS 5700 fixed port routers In a network employing an Autonomous System Boundary Router (ASBR) and other routers, you are now assured of uninterrupted traffic flow even if the ASBR generates LSAs that exceed the limit you configured. This is made possible as you can now isolate ASBRs and also control the duration of adjacency in the EXCHANGE or LOADING phase. By isolating the ASBR from its immediate neighbors, the remaining network topology can continue to function without disruption, effectively preventing any adverse impact on traffic flow. This approach also simplifies the recovery process, as manual intervention is only necessary for the immediate neighbors of the ASBR routers. This feature introduces these changes: CLI: • max-external-lsa • exchange-timer YANG Data Model: • Cisco-IOS-XR-ipv4-ospf-cfg.yang • Cisco-IOS-XR-ipv4-ospf-oper.yang • Cisco-IOS-XR-um-router-ospf-cfg.yang (see GitHub, YANG Data Models Navigator)
Automatically Reestablish a BGP Neighbor Session	Introduced in this release on: NCS 5500 fixed port routers; NCS 5700 fixed port routers; NCS 5500 modular routers (NCS 5500 line cards; NCS 5700 line cards [Mode: Compatibility; Native]) You can now configure the router to automatically re-establish a BGP neighbor session that has been disabled because the maximum-prefix limit has been exceeded. The feature introduces these changes: CLI • maximum-prefix-restart-time

	YANG Data Model: - New XPaths for openconfig-bgp-neighbor.yang(see GitHub, YANG Data Models Navigator)
BGP Flowspec on Bridge-Group Virtual Interfaces	Introduced in 7.10.1 release on: NCS 5500 modular routers (NCS 5700 line cards [Mode: Native]) You can now effectively employ BGP Flowspec on Bridge-Group Virtual Interface (BVI) to connect to broadcast domains that house host devices, as in the case of enterprise networks. This support means that your customers can safeguard their networks from network threats such as Distributed Denial of Service (DDoS) attacks incoming through the BVI.
Discard Incoming BGP Update Message	Introduced in 7.10.1 release on: NCS 5500 fixed port routers; NCS 5700 fixed port routers; NCS 5500 modular routers (NCS 5500 line cards; NCS 5700 line cards [Mode: Compatibility; Native]) You can now avoid the session reset when a BGP session encounters errors while parsing the received update message. This is made possible because the feature enables discarding the incoming update message as a withdraw message. CLI: - update in error-handling treat-as-withdraw YANG Data Model: - New XPaths for openconfig-bgp-neighbor.yang (see GitHub, YANG Data Models Navigator)
Exclusion of Label Allocation for Non-Advertised Routes	Introduced in 7.10.1 release on: NCS 5500 fixed port routers; NCS 5700 fixed port routers; NCS 5500 modular routers (NCS 5500 line cards; NCS 5700 line cards [Mode: Compatibility; Native]) We have enabled better label space management and hardware resource utilization by making MPLS label allocation more flexible. This flexibility means you can now assign these labels to only those routes that are advertised to their peer routes, ensuring better label space management and hardware resource utilization. Prior to this release, label allocation was done regardless of whether the routes being advertised. This resulted in inefficient use of label space.
Protection of Directly Connected EBGP Neighbors through Interface-Based LPTS Identifier	Introduced in 7.10.1 release on: NCS 5500 fixed port routers We have enhanced the network security for directly connected eBGP neighbors by ensuring that only packets originating from designated eBGP neighbors can traverse through a single interface, thus preventing IP spoofing. This is made possible because we've now added an interface identifier for Local Packet Transport Services (LPTS). LPTS filters and polices the packets based on the type of flow rate you configure.

	The feature introduces the following: CLI: • bgp lpts-secure-binding YANG Data Model: • Cisco-IOS-XR-um-router-bgp-cfg (see GitHub, YANG Data Models Navigator)
Reduce Recursions for eBGP Peering on Loopback Address on Bridge-Group Virtual Interface	Introduced in 7.10.1 release on: NCS 5500 modular routers (NCS 5700 line cards [Mode: Native]) You can now achieve eBGP peering on Loopback interfaces on Bridge-Group Virtual Interface (BVI) and reduce the recursion level from three to two. This reduction in the recursion level, achieved by removing the need to use the BVI name in the configuration of static routes, allows faster packet forwarding and better utilization of network resources.
BGP Policy Accounting	Introduced in release 7.9.1: Border Gateway Protocol (BGP) policy accounting measures and classifies IP traffic that is received from different peers. You can identify and account for all traffic by customer and bill accordingly. Policy accounting is enabled on an individual input interface basis. Using BGP policy accounting, you can now account for traffic according to the route it traverses. This feature is now supported on routers that have the Cisco NC57 based line cards with external TCAM (eTCAM) and operate in native mode. This feature introduces these changes: • CLI: The feature introduces the hw-module fib bgppa stats-mode command. • YANG Data Model: New XPaths for Cisco-IOS-XR-um-hw-module-profile-cfg.yang (see GitHub, YANG Data Models Navigator)
Detect Slow Peer in a BGP Group	Introduced in release 7.9.1: BGP peers process the incoming BGP update messages at different rates. A slow peer is a peer that is processing incoming BGP update messages very slowly over a long period of time compared to other peers in the update sub-group. Slow peer handling is important when routes are constantly changing over a long period of time. It is important to clean up stale information in the queue and send only latest state. It is helpful to know if there is a slow peer, which indicates there is a network issue, such as sustained network congestion or a receiver not processing updates on time, that the network administrator can address.
Limiting LSA numbers in a OSPF Link-State	Introduced in release 7.9.1: The nonself-generated link-state advertisements (LSAs) for a given Open Shortest Path First (OSPF)

Database	process is limited to 500000. This protection mechanism prevents routers from receiving many LSAs, preventing CPU failure and memory shortages, and is enabled by default from this release onwards. If you have over 500000 LSAs in your network, configure the max-lsa command with the expected LSA scale before upgrading to this release or later. This feature modifies the following commands: • show ospf to display the maximum number of redistributed prefixes. • show ospf database database-summary detail to display the number of LSA counts per router. • show ospf database database-summary adv-routerrouter ID to display the router information and the LSAs received from a particular router.
Limiting the Maximum Redistributed Type-3 LSA Prefixes in OSPF	Introduced in release 7.9.1: By default, the maximum redistributed Type-3 LSA prefixes for a given OSPF process is now limited to 100000. This mechanism prevents OSPF from redistributing a large number of prefixes as Type-3 LSAs and therefore preventing high CPU utilization and memory shortages. Once the number of redistributed prefixes is reached or exceeds the threshold value, the system log message is generated, and no more prefixes are redistributed.