OSPF for large-scale networks with regular topologies
draft-smirnov-ospf-dive-01

Abstract

Many popular topologies for large-scale networks have highly regular structure with distinctive design pattern. Examples of such topologies include hub-and-spoke (also known as "star") common in enterprise WAN networks, fat-tree and Clos topologies common in datacenters. For number of reasons in such large-scale networks distance-vector protocols perform better than OSPF. On the other hand network backbones have no highly regular topology pattern and there OSPF outperforms distance-vector protocols. As a result large-scale networks frequently employ different routing protocols in different regions of the network, complicating network operations.

This document proposes OSPF extensions to improve scalability of routing for large-scale networks.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on October 12, 2015.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.
This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction ............................................. 3
2. Requirements Language .................................. 3
3. Problem definition ...................................... 4
  3.1. Typical regular network topologies .................. 4
    3.1.1. Hub-and-spoke topology .......................... 4
    3.1.2. Fat-tree topology .................................. 5
    3.1.3. Clos topology .................................. 5
  3.2. Problems with OSPF routing in large-scale networks ... 5
4. Solution requirements .................................... 7
5. Functional summary ...................................... 8
6. Protocol Details ......................................... 10
  6.1. The DIVE area ...................................... 10
  6.2. Hello packets and the database exchange on DIVE area interfaces .................................. 11
  6.3. LSA generation into the DIVE area .................. 12
    6.3.1. Metric Sub-TLV .................................. 13
  6.4. SPF calculation in the DIVE area .................. 14
  6.5. Translation of LSAs and route propagation ............ 15
    6.5.1. Hub routers: Propagation of routes from the core network into the DIVE area .................. 16
    6.5.2. Hub routers: Propagation of routes from the DIVE area into the backbone area ............ 17
    6.5.3. Hub routers: Propagation of routes from the DIVE area into the non-backbone area ....... 18
    6.5.4. Route propagation on Spoke routers ............... 19
7. Other considerations for the DIVE area .................... 20
  7.1. Routing considerations .............................. 20
  7.2. LSDB size considerations ............................ 20
  7.3. Optimal DIVE area design ............................ 21
8. Backward Compatibility ................................... 21
9. Security Considerations ................................ 21
10. IANA Considerations .................................. 21
11. Acknowledgements ..................................... 22
12. References ........................................... 22
    12.1. Normative References ............................. 22
    12.2. Informative References ............................ 22
1. Introduction

OSPF is a link-state protocol which was designed to provide routing in networks of arbitrary topology. Big modern networks may have thousands of routers providing the same type of service. To simplify network design and operations as well as to unify hardware and software configurations of routers such networks are frequently built by replicating a basic design element hundreds and thousands of times. Resulting network has highly regular topology exhibiting a distinctive pattern. Such regular designs include hub-and-spoke topology common in enterprise networks, "fat-tree" and Clos topologies common in data center networks. Running routing protocols in such networks poses number of problems arising mostly from the very large number of routers in the network. On the other hand, regular pattern of the topology allows certain simplifications.

OSPF (and link state protocols in general) can be used to provide routing in networks with regular topologies but it does not make any use of the regularity. This makes OSPF especially vulnerable to the elements of the large scale.

Real-life networks combine regions of regular topologies with (smaller scale) regions of free topology where OSPF works the best. Continuing examples above, these are the headquarter (HQ) network of the enterprise or interconnections between datacenters. For operational simplicity it is desirable to have the same routing protocol running in both parts of the network. The present document specifies extensions to OSPF to improve its scalability in the very large scale networks with regular topologies.

Section 3 of this document details typical problems seen if OSPF is used as the routing protocol in the large scale networks with regular topologies. Section 4 lists requirements for the routing solution. Section 5 gives brief overview of the solution. Section 6 defines details of the new protocol behavior.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].
3. Problem definition

3.1. Typical regular network topologies

3.1.1. Hub-and-spoke topology

Hub-and-spoke topology (also frequently called a "star" topology) comes in many variants but they all possess set of common properties allowing to group them into a single category:

- A few (usually one or two) routers with central location which on one side connect to each remote site by a point-to-point Layer-2 connection and on another side to the network backbone. These routers are usually referred as 'hubs'.

- Remote site may have a single router possibly dual-homed to hub routers or multiple routers each connecting to a hub router. These routers are frequently referred as 'spokes'. Internally site’s network may consist of multiple devices but typically it is relatively small and simple. To the outside world site’s internal reachability is described by just a few prefixes.

- Connection between hubs and spokes is the only connection between a site and rest of the network. All traffic going in/out of the site flows through the hubs. Inter-spoke traffic is minimal and may even be blocked, for example for security reasons.

- Hub-and-spoke network usually provides some sort of redundancy. Redundant hub routers and redundant connections to sites are common; redundant spoke routers are not unusual.

- Number of sites/spokes may be very large requiring each hub router to handle thousands of spoke neighbors.

- All sites share the same routing administrative policy, i.e. the same route filtering and summarization rules.

On hub routers hub-and-spoke layer-2 connections may be presented to OSPF as either single point-to-multipoint interface with large number of neighbors or as multiple point-to-point Layer-3 interfaces.

Examples of layer-2 technologies commonly used to create hub-and-spoke networks are (from historical to modern technologies): SMDS, ISDN, Frame-Relay, IPsec VPN tunnels, VPLS.

Hub-and-spoke topology is the most common topology for building enterprise WAN networks. It naturally maps into HQ/branch enterprise
structure and provides optimal link capacity for traffic flowing between branches and centralized services located in the HQ.

3.1.2. Fat-tree topology

Fat-tree topology is hierarchical structure resembling tree turned upside down with link capacity growing from the "leafs" up toward the "roots". This topology is common in datacenter designs. Since parts of the tree may run Layer-2 switching, IP routing protocols usually 'see' 2 or 3 levels in the tree hierarchy. Typically lower level nodes have redundant connections to the upper level.

This topology may be seen as hub-and-spoke topology where each spoke site in turn has the tree-like organization.

3.1.3. Clos topology

Clos topology is multistage switching network where each node in a stage connects to multiple nodes in previous and subsequent stages. This is another topology frequently used in datacenters. The key advantage of this topology is that number of elements in each stage and number of links connecting an element with previous and subsequent stages may be chosen in such way that all upstream and downstream links have the same capacity. Clos network always provides multiple same-cost paths between each pair of leaf nodes. On a node number of ECMP paths for each destination would be equal to number of connections to elements in the next stage of the Clos topology. In real-life deployments there may be as many as several dozens of ECMP paths for each routing destination prefix.

3.2. Problems with OSPF routing in large-scale networks

Hub-and-spoke networks pose challenges to routing scalability:

- All spoke sites placed into single OSPF area will, by the virtue of link-state protocol, receive full topology information describing each spoke in the area and its connection to hubs. This information is at best excessive as all links from the site go to hubs and knowledge of many other spoke links in the area cannot reveal alternative paths to destinations outside of the site. Worse yet, spoke routers are usually much smaller devices than hub routers and they are intended to serve tiny site network with a few routes and light traffic; spoke routers do not have sufficient resources (either memory or CPU) to hold and process the same link-state database as hub routers. Distance-vector protocols in the same topology propagate only prefix reachability information and do not tax spoke routers with topology view.
- Inter-site visibility may be undesirable to decrease size of the routing table on spoke devices or because of security reasons. Link-state protocols do not allow routing information to be filtered within area flooding scope. To compare, distance-vector protocols allow route filtering and summarization on per-neighbor basis.

- Having full topology visibility within an area may also lead hub routers to compute suboptimal paths. Consider example hub-and-spoke network with two hubs A and B and two spoke sites S1 and S2. Each spoke site is dual-connected to both hubs. Hubs A and B are Area Border Routers (ABRs) between hub-and-spoke WAN and the backbone. If link between hub A and site S1 goes down then A will choose (or at least consider) intra-area WAN route A -> S2 -> B -> S1. But typically a spoke site (S2 in this example) must not be considered for any transit traffic.

- Size and stability of Router LSA of hub routers is also problematic due to very large number of point-to-point links in it describing connections to all spokes in the area. Both the size and probability of need to rebuild the Router LSA grow directly proportional to N, number of spokes in the area. Thus overall CPU resources consumed by flooding and processing hub’s Router LSA grow as \( O(N^2) \).

- Alternative network design is to separate each spoke into area of its own. This solves problems of spoke routers but transfers cumulative burden of supporting multiple areas to hub routers. This design requires hub router to be able to support thousands of NSSA areas, originate as many router LSAs, translate multiple LSAs from/into each area. Managing common route filtering and summarization policy is also difficult.

Some problems described above call for designing as small areas as possible, while others vice versa are resolved by designing big areas. In a relatively small network it is possible to find a sensible design compromise but as number of spokes grows to thousands finding working compromise becomes more and more challenging and the balance becomes more and more fragile.

As noted in Section 3.1.2, fat-tree topology can be viewed as a particular case of the hub-and-spoke topology. For this reason many problems described above for hub-and-spoke networks are equally affecting fat-tree networks.

Clos networks add one more problem specific to this topology:
Section 3.1.3 underlined that Clos networks provide massive equal-cost multipath for most destinations. When a link goes down this rarely means that the node lost connectivity to any destination. OSPF in this situation rebuilds its Router LSA and floods it to all routers in the area. Ensuing SPF on all nodes in the area will result in no change in the routing on all routers not connected to the problematic link. For comparison, distance-vector routing protocols detect that there were no change in reachability of prefixes or their metrics and hence no update is sent to neighbors.

As can be seen from these examples full knowledge of topology within an area, what is a key property of link state protocols and works so well in networks with arbitrary topology, becomes the biggest factor limiting routing scalability in networks with regular topologies. For this reason distance-vector protocols are the tool of choice for network designer working with large hub-and-spoke networks. Factors specific to networks with regular topologies, such as link between hub and spoke being the only connectivity to rest of the network and low number of prefixes advertised in each direction, negate convergence slowness which affects distance-vector protocols in more complex topologies.

4. Solution requirements

For OSPF solution to be as scalable as distance-vector protocol these design goals were taken into consideration:

- Spoke routers MUST be protected from routing information sent from/to other spoke routers unless explicitly required by network’s policy.

- Solution MUST require no modification to OSPF routers other than those connected to hub-and-spoke network itself.

- Number of LSAs which hub router originates MUST NOT grow faster than $O(N)$ where $N$ is number of spoke sites and preferably should not depend on $N$. Size of LSAs MUST NOT depend on $N$.

- Solution MUST protect against routing loops should a spoke site becomes connected to another site and/or to rest of the network. Supporting such configuration is not a goal and such configuration looses important property defining spoke site of hub-and-spoke networks. But network must be protected from routing meltdown in case of accidental misconfiguration.

- Solution SHOULD provide easy and scalable way to apply common administrative routing policy via centralized configuration.
5. Functional summary

Area Border Routers in OSPF propagate inter-area routing information by announcing reachability and routing metrics. Thus inter-area and external routes are announced in OSPF as in distance-vector routing protocols.

The solution satisfying requirements laid in the Section 4 is to create new type of OSPF area which is devoid of LSAs carrying topology information (i.e. Router and Network LSAs). The routing information in this area is propagated only by prefix LSAs flooded with link-local scope. In many aspects such area behaves similarly to distance-vector routing protocols. Due to this the area type is called DIstance-VEctor or DIVE area.

In hub-and-spoke network DIVE area would cover only links between hubs and spoke routers. Other interfaces on hub routers will be placed into a regular OSPF area carrying link-state routing information. This may or may not be the backbone area. Further in this document this area and network behind it is called "the core network".

Spoke routers are connected to area covering the site network.

Example hub-and-spoke network using DIVE area is depicted on Figure 1.
Note that network using DIVE area design resembles [RFC4364] BGP/MPLS VPN network where OSPF is used as PE-CE protocol [RFC4577]. The DIVE area is analogue of BGP ‘superbackbone’ in MPLS VPN network and hub and spoke routers are analogues of MPLS VPN PE routers.

DIVE area does not have LSAs with area scope flooding. This solves problems related to excessive visibility of routing information where it is undesirable. There is also no visibility or reachability of routers other than immediately connected neighbors. Each router in the DIVE area is ABR and there is no concept of a router internal to the DIVE area. In this sense DIVE area is always a one-hop area.

Since DIVE area deviates from traditional one level hierarchy of OSPF areas (backbone area/all other areas connected to it) it must employ strict rules of accepting and propagating routing information to prevent routing information looping. These rules are further discussed in Section 6.5. ABRs connected to DIVE area perform translation of LSAs from and into the DIVE area similar to translation of NSSA LSA into External LSA on ABR between NSSA and backbone areas.

Routing information propagated throughout DIVE area is encoded into Prefix Attribute LSAs for OSPFv2 ([I-D.ietf-ospf-prefix-link-attr])
and Extended Prefix LSAs for OSPFv3 ([I-D.ietf-ospf-prefix-link-attr]) and Extended LSAs for OSPFv3 ([I-D.ietf-ospf-ospfv3-lsa-extend]). In both cases ‘old’ style LSAs carrying either topological or routing information are not originated or flooded into the DIVE area.

6. Protocol Details

6.1. The DIVE area

Current specification defines new type of OSPF area called DIVE area. There is no concept of router internal to the DIVE area, all DIVE area routers are ABRs. At the time of configuring DIVE area, role of the router in the area must be provided as configuration parameter. Currently supported DIVE area router roles are Hub and Spoke. Router role has important implications during translation of routes between the DIVE area and other connected area(s). Details of LSA translation are covered in Section 6.5. In the subsequent text terms ‘Hub’ and ‘Spoke’ (with capital letter) are used to denote routers’ configured role in the DIVE area while terms ‘hub’ and ‘spoke’ (with small letter) describe position of the router in the hub-and-spoke topology.

Router of any role may be connected to multiple DIVE areas. Hierarchical DIVE areas are not defined by the current specification. In other words, router’s role in all connected DIVE areas SHOULD be the same.

Hub router must be connected to either backbone or non-backbone regular area. Hub router cannot be connected to either stub or NSSA area.

Site area (or areas) connected to the Stub router must be a non-backbone regular area, NSSA or stub area. Note that route filtering and summarization are best to be applied on the hub routers. This will both protect the high-scale DIVE area from flooding unnecessary information and provide centralized location to manage the route filtering/summarization policy on a few hub routers rather than on many spokes. So stub and NSSA areas on spoke sites would provide limited benefit comparing to regular non-backbone area and SHOULD be used only if there exist direct spoke-to-spoke neighborships between some sites.

Virtual links through DIVE area are not supported.

Routers connected to the DIVE area MUST support Prefix Attribute LSAs for OSPFv2 ([I-D.ietf-ospf-prefix-link-attr]) and Extended LSAs for OSPFv3 ([I-D.ietf-ospf-ospfv3-lsa-extend]).
6.2. Hello packets and the database exchange on DIVE area interfaces

All routers connected to the DIVE area must agree on the area’s configuration and learn roles of neighbors. Roles of the local router and neighbor determine LSA translation and route propagation. Section 6.5 details these rules.

Router’s role is advertised in Hello packets sent on interfaces in the DIVE area. Two new bits, called DV-bits, are used to encode router’s role in the DIVE area. DV-bits are allocated in the Extended Options and Flags LLS TLV for OSPFv2 [RFC5613] and in the Options field of OSPFv3:

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| | | | | | | | | | | | | | | | | | | | | | | | | | |DV |F|I|RS|LR|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Bits in Extended Options and Flags TLV

```
0                   1                   2
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| | | | | | | | | | | | | DV|L|AF|*|*|DC|R|N|MC|E|V6|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

OSPFv3 Options Field

Meaning of DV bits is defined as:

- 0 0 - Interface of sending router does not belong to DIVE area
- 0 1 - Sending router has role Hub
- 1 0 - Sending router has role Spoke
- 1 1 - Reserved

When Hello packet is received from previously unknown neighbor DV bits are checked to see if neighbor’s interface belongs to a DIVE area. If neighbor advertises any of DV bits set and receiving interface does not belong to a DIVE area OR if both DV bits advertised by the neighbor are clear and receiving interface belongs to a DIVE area then received packet MUST be silently discarded.

DV bits advertised by the neighbor must be stored in neighbor’s data structure and compared when receiving subsequent Hello packets from
the neighbor. Change in advertised DV bits MUST generate BadLSReq Neighbor FSM event. Processing this event will cause adjacency with the neighbor to be reset and LSDB exchange to [re]start.

LS database of DIVE area may contain only Opaque LSAs for OSPFv2 and Extended LSAs for OSPFv3. LSA types defined in [RFC2328] and [RFC5340] are not flooded into the DIVE area, including AS External LSAs with the domain flooding scope. OSPFv2 opaque LSAs with domain flooding scope are not flooded into DIVE areas. OSPFv3 flooding of unknown LSA types is performed as described by [RFC5340].

Choosing neighbors to establish the full adjacency or to stop neighbor formation at the 2-Way Neighbor FSM state does NOT depend on DIVE area roles of the local router and of the neighbor and works as described in [RFC2328]. On broadcast and NBMA interfaces of Spoke routers in the DIVE area implementation SHOULD have Router Priority by default set to 0.

If during the LS database exchange with neighbor in DIVE area router receives Database Description packet describing LSA of a type not allowed in the DIVE area then SeqNumberMismatch Neighbor FSM event MUST be generated and LSDB exchange must restart.

If OSPF interface type is broadcast then implementation SHOULD support Incremental Hellos as described by [RFC5820]. If Incremental Hellos are supported then they MUST be enabled by default on broadcast interfaces in DIVE area. On point-to-multipoint interfaces Hub routers SHOULD default to sending unicast Hellos to discovered neighbors rather than sending multicast Hello packets listing all known neighbors.

6.3. LSA generation into the DIVE area

Following types of LSAs containing listed TLV types may be originated into the DIVE area:

For OSPFv2 (see [I-D.ietf-ospf-prefix-link-attr]):

- OSPFv2 Extended Prefix Opaque LSA
- * OSPFv2 Extended Prefix TLV

Extended Prefix Opaque LSA MUST have LSA Type 9.

Extended Prefix TLV as defined by [I-D.ietf-ospf-prefix-link-attr] may advertise attributes for several route types. Only following route types may be present in Extended Prefix TLV in LSAs originated into the DIVE area:
1 - Intra-Area
3 - Inter-Area
5 - AS External

This specification defines one new sub-TLV of OSPFv2 Extended Prefix TLV - Metric Sub-TLV, see Section 6.3.1.

For OSPFv3 (see [I-D.ietf-ospf-ospfv3-lsa-extend]):

- E-Intra-Area-Prefix-LSA
- E-Inter-Area-Prefix-LSA
- E-AS-External-LSA
- E-Link-LSA

Extended Prefix LSAs may contain following TLV types:

- 6 - Intra-Area Prefix TLV
- 3 - Inter-Area Prefix TLV
- 5 - External Prefix TLV

All Extended prefix LSAs originated into the DIVE area MUST have link-local flooding scope. Thus their LSA types will be:

<table>
<thead>
<tr>
<th>LSA function code</th>
<th>LS Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>0x8023</td>
<td>E-Inter-Area-Prefix-LSA</td>
</tr>
<tr>
<td>37</td>
<td>0x8025</td>
<td>E-AS-External-LSA</td>
</tr>
<tr>
<td>40</td>
<td>0x8028</td>
<td>E-Link-LSA</td>
</tr>
<tr>
<td>41</td>
<td>0x8029</td>
<td>E-Intra-Area-Prefix-LSA</td>
</tr>
</tbody>
</table>

6.3.1. Metric Sub-TLV

One new sub-TLV is defined for OSPFv2 to carry metric of the route. This is required because in the DIVE area Extended Prefix Opaque LSAs do not accompany [RFC2328] LSAs and must carry all route information.

The Metric Sub-TLV is a Sub-TLV of the OSPF Extended Prefix TLV defined in [I-D.ietf-ospf-prefix-link-attr]. It MAY appear more than once in the top level TLV and has the following format:
where:

Type: TBD, suggested value is 3.

Length: 4.

E: one bit field. For AS External routes this defines the type of external metric. Its function and meaning are fully analogous to the E-bit of Type-5 LSA [RFC2328]. This bit is always 0 for intra- and inter-area route metrics.

MT-ID: Multi-Topology ID (as defined in [RFC4915]).

Metric: The cost of this route. For inter-area and external routes all 24 bits of the field may be used to encode route metric. For intra-area routes upper 8 bits must be 0, thus valid metric for intra-area route is in the range 1 to 2^{16}-1.

If more than one instance of the Metric Sub-TLV is present in the Extended Prefix TLV then each instance MUST describe metric with different Multi-Topology ID.

6.4. SPF calculation in the DIVE area

Intra-area SPF calculation within the DIVE area is reduced to walking the list of neighbors in the area and adding neighbors which have reached the FULL adjacency state to the table of routers’ reachability. Each reachable router is marked to be both ABR and ASBR. Cost of the routing table entry is equal to the cost of local interface associated with the neighbor.

Unlike of other area types in the DIVE area routers which do not have fully established adjacency between them do not have valid intra-area path to reach each other.

Calculation of inter-area and AS External routes follows algorithms described in [RFC2328] for OSPFv2 and [RFC5340] for OSPFv3 with following caveat. OSPFv2 Extended Prefix LSA does not provide ordering of prefixes by prefix types. Hence there are no separate phases of computing inter-area and then AS external routes. Instead,
all Extended Prefix LSAs and all Extended Prefix TLVs in them are examined in turn and type of calculated route is determined by the Route Type field of the Extended Prefix TLV being examined.

6.5. Translation of LSAs and route propagation

Each router connected to a DIVE area is an Area Border Router and will originate LSAs into connected non-DIVE areas to describe reachability of prefixes received via the DIVE area. And vice versa, it will originate LSAs into the DIVE area to describe reachability of routes learned via other connected areas.

Moreover, in DIVE area all LSAs propagating routing information have link-local scope. In those cases where routing information should propagate between routers which do not have direct adjacency, intermediate routers will originate their own LSAs carrying routing information one hop further. Accordingly to distance-vector routing principles metric of such routes will be increased to reflect cost of the path to reach destination from the router originating LSA. There are two cases when routing information has to be re-advertised within the DIVE area:

- If inter-spoke site traffic is not prohibited then hub routers must advertise to spokes inter-spoke routing information. This may be either in the form of summarized routing information covering multiple spoke sites (including advertisement of default route) or in the form of non-summarized routing information hub received from spoke routers. In the latter case hub would re-advertise in the DIVE area routing information received from spoke neighbors in the area.

- DIVE area is attached to the core network via redundant hub routers and hubs advertise into the network summarized routing information covering multiple site prefixes. If link between one of hubs and a spoke site is lost then the hub must know alternative paths to the spoke network via other hubs. Direct neighborhood between hub routers in the DIVE area would provide such alternate path. Thus in this scenario hub routers advertise summarized routing information into the core network and exchange non-summarized spoke prefix reachability via DIVE area adjacency.

Note that in both scenarios above re-advertisement of routing information within DIVE area is done by Hub routers and information being re-advertized was received by the Hub from Spoke routers. Routers whose role in DIVE area was configured as Spoke MUST NOT re-advertise into the DIVE area routing information received via a DIVE area. Routers whose role in DIVE area was configured as Hub MUST NOT
re-advertize routing information received from other Hub routers in this DIVE area.

Route filtering and/or summarization is frequently configured on Hub routers. Summarization reduces number of LSAs to originate, maintain and flood. Managing LSDB size is an important aspect of scalability in a large-scale network. Summarization may be performed in both directions - to summarize reachability of core networks advertised toward the spoke sites (in the ultimate summarization case Hubs may advertise toward spokes only one - default - route) and to summarize into the core reachability of remote sites connected by the hub-and-spoke network. To improve stability of LSA advertising summarized routing information an implementation MUST allow cost of the summary route to be statically provided via configuration and SHOULD have static assignment of summary cost (as opposed to dynamically computing cost of the summary route from costs of component routes falling into the summary range) as default cost selection mechanism.

A spoke site for redundancy reasons may be connected to the hub-and-spoke network by more than one spoke router. To prevent looping of routing information, routes propagated from the DIVE area into the spoke site network must not be re-advertised back into the DIVE area by another spoke router. This is achieved by setting by Spoke routers the Down bit in LSAs advertised into the spoke site network. Unlike of looping prevention for MPLS VPN PE routers [RFC4576], Spoke routers are allowed to install into their own routing table routes derived from LSAs with the Down bit set but they MUST NOT re-advertising them into the DIVE area.

Following chapters describe route propagation and re-advertisement rules. For Hub routers LSA translation rules for routes learned from the DIVE area depend on if the Hub is connected to the backbone area or non-backbone area.

6.5.1. Hub routers: Propagation of routes from the core network into the DIVE area

After completing calculation of routes during SPF DIVE area’s Hub router will perform Area Border Router’s functions. This section lists rules to propagate routing information from the core network into the DIVE area. Each prefix being propagated will be described by one Prefix TLV, see Section 6.3. Strategy of packing Prefix TLVs into LSAs (one or multiple Prefix TLVs per LSA; LS ID selection etc) is outside of the scope of this document.

To advertise routing information into DIVE area a Hub router MUST:
o Examine each reachable prefix in its routing table. If the best path for the prefix lies through the DIVE area then proceed to the next prefix.

o Check if the route falls into any configured summary range. If it is, mark the summary as reachable. Compare type of the route with type of the summary using usual OSPF route preference rules (intra-area route is preferred over inter-area; external Type-1 is preferred over external Type-2 etc.). If route’s type is more preferable store it as new type of the summary route and proceed to the next route.

o Otherwise add into DIVE area’s LSAs route TLV of the appropriate type. The LSA MUST be added to LSDB of all interfaces where exist Spoke neighbor(s) in a state above Down:

* If route is intra-area or inter-area then originate inter-area route TLV. Use cost of the route as cost advertised in the TLV

* If route is Type-1 AS external route then originate Type-1 AS External route TLV. Use cost of the route as cost advertised in the TLV

* If route is Type-2 AS external route then originate Type-2 AS External route TLV. Use Type-2 metric of the LSA which contributed to the route plus one as cost advertised in the TLV

o For routes which became unreachable advertise the LSA without TLV corresponding to the route or flush the LSA if applicable

Calculation of the summary route reachability and type, as well as flushing TLVs of unreachable routes is the same for all router roles and route propagation scenarios, so for brevity they are omitted in the following sections.

6.5.2. Hub routers: Propagation of routes from the DIVE area into the backbone area

If the Hub router is connected to (i.e. has interfaces in) the backbone area then route advertisement rules are:

o For routes whose LSA was originated by a Spoke router originate into the backbone area LSA of the corresponding type:

* For intra- and inter-area routes originate Type-3 Summary LSA (OSPFv2) or Inter-area Prefix LSA (OSPFv3) using cost of the route.
* For Type-1 AS external routes originate Type-5 External LSA (OSPFv2) or AS-External LSA (OSPFv3) advertising Type-1 external route and using cost of the route

* For Type-2 AS external routes originate Type-5 External LSA (OSPFv2) or AS-External LSA (OSPFv3) advertising Type-2 external route and using metric received in the Prefix TLV plus one

* When advertising AS external routes the Hub router MUST also announce itself as ASBR

  o If LSA was not originated into the backbone because the route is subsumed by summarization then instead add TLV to the LSA in LSDB of all interfaces where exist Hub neighbor(s) in a state above Down. Otherwise to provide inter-spoke connectivity TLV MAY be added to the LSA in LSDB of all interfaces where exist Spoke neighbor(s) in a state above Down. In either case the TLV MUST have the same route type as route being advertised. For intra-inter-area and Type-1 external routes advertised cost is taken as cost of the route. For Type-2 external routes the cost is equal to metric received in the Prefix TLV plus one

  o Note that a Hub router MUST NOT advertise into either the backbone or to other Hubs routes received from Hubs

6.5.3. Hub routers: Propagation of routes from the DIVE area into the non-backbone area

If the Hub router is not connected to the backbone area then it cannot advertise inter-area routing information. To provide compromise between network design flexibility and compatibility with [RFC2328]/[RFC5340] implementations the Hub router will advertise routing information as AS external routes.

For routes whose LSA was originated by a Spoke the Hub router MAY originate into the non-backbone area LSA of the following type:

  o For intra- and inter-area routes originate Type-5 External LSA (OSPFv2) or AS-External LSA (OSPFv3) advertising Type-1 AS external route with metric equal to cost of the route

  o For Type-1 AS external routes originate Type-5 External LSA (OSPFv2) or AS-External LSA (OSPFv3) advertising Type-2 external route with metric of one
o For Type-2 AS external routes originate Type-5 External LSA (OSPFv2) or AS-External LSA (OSPFv3) advertising Type-2 external route and using metric received in the Prefix TLV plus one.

Propagation of the route to other Hub or Spoke routers in the same DIVE area is the same as described in the previous section.

6.5.4. Route propagation on Spoke routers

To advertise routing information received from the DIVE area into areas of the site network Spoke router MUST:

o For intra- and inter-area routes originate into the site area Type-3 Summary LSA (OSPFv2) or Inter-Area-Prefix LSA (OSPFv3). Metric advertised in the LSA is set equal to cost of the route.

o For Type-1 AS external routes originate into the site area Type-5 External LSA (OSPFv2) or AS-External LSA (OSPFv3) advertising Type-1 external route with metric equal to cost of the route.

o For Type-2 AS external routes originate Type-5 External LSA (OSPFv2) or AS-External LSA (OSPFv3) advertising Type-2 external route with metric equal to the metric in TLV contributing to the route.

o In all above cases the LSA MUST have the Down bit set.

To advertise into DIVE area routing information learned from attached site network area Spoke router:

o MUST skip routes which were produced from LSAs with the Down bit set.

o Since site area is a non-backbone area Spoke router MUST NOT have inter-area routes learned via the site network.

o For other route types add into DIVE area’s LSAs route TLV of the type as listed below. The LSA MUST be added to LSDB of all interfaces where exist Hub neighbor(s) in a state above Down:

  * If route is intra-area then originate inter-area route TLV. Use cost of the route as cost advertised in the TLV.

  * If route is Type-1 AS external or translatable NSSA route then originate Type-1 AS External route TLV. Use cost of the route as cost advertised in the TLV.
* If route is Type-2 AS external or translatable NSSA route then originate Type-2 AS External route TLV. Use Type-2 metric of the LSA which contributed to the route as cost advertised in the TLV.

7. Other considerations for the DIVE area

7.1. Routing considerations

Route propagation rules in DIVE area make sure that information is advertised between Hubs and Spokes and into respective connected areas. These rules prohibit multiple re-advertisement of the routing information within the DIVE area. Thus the DIVE area may only serve as a shim layer between traditional OSPF areas and it is not possible to build full OSPF network functioning on principles of distance-vector protocol.

Routing information traveling through DIVE area loses track of its true originator. To prevent routing loops, routes delivered via the DIVE area are made worse. For routes carrying metric comparable with cost of intra-domain path this is done by adding cost of links to reach route’s origin. For routes carrying cost external to the OSPF domain this is done by incrementing the external cost.

This increment in the metric also solves problem of originator receiving back its own routing information. For example, if spokes are connected to a Hub by a point-to-multipoint interface and the Hub wants to advertise to spokes prefix received from a Spoke router then the Spoke router which originated the prefix will receive its own information back even though the LSA has link-local flooding scope.

Fast-poisoning of routes which became unreachable is ensured by rules which prevent a Spoke router from re-advertising back to Hubs (directly or indirectly via other Spoke routers connected to the same spoke site) any routing information received on the DIVE area interface.

7.2. LSDB size considerations

LSAs in DIVE area have link-local flooding scope. This solves scalability problems of spoke routers because they don’t have to deal with information originated for or from the other spokes (unless it is desired). This also solves input-output constraints on hub routers by limiting volume of information which has to be exchanged with each spoke. On the other hand this may have adverse effect on the size of the link-state database a hub router has to maintain.

This is the case when spoke routers are connected by point-to-point OSPF interfaces. In this case the database size of hub router is multiplied by number of interfaces to spoke sites. This problem can
be addressed by grouping spoke connections into smaller number of point-to-multipoint interfaces.

7.3. Optimal DIVE area design

Given these considerations, the recommended DIVE area design for Hub routers is:

- Spoke routers are connected via small number of point-to-multipoint interfaces
- Hub routers, if necessary, are interconnected within the DIVE area via interfaces separate from connections to Spokes
- Hub routers do route summarization of routing information they advertise both into the core network and into the DIVE area toward Spoke routers.

8. Backward Compatibility

Devices attached to the DIVE area MUST conform to this specification. Awareness of devices is checked via new options bits in Hello packets before the start of adjacency formation, thus devices not supporting this specification cannot join the DIVE area.

This specification is fully backward compatible with devices not immediately connected to DIVE area. New information defined by this specification is not propagated to such devices. Current specification includes measures to protect a network in case of basic misconfiguration or design problem.

9. Security Considerations

This document does not introduce any new security implications. General security considerations described in
[I-D.ietf-ospf-prefix-link-attr] and
[I-D.ietf-ospf-ospfv3-lsa-extend] apply to LSAs in DIVE area.

10. IANA Considerations

This specification updates several IANA OSPF registries:

- New bits (DV-bits) are reserved in the "LLS Type 1 Extended Options and Flags" registry of the Extended Options and Flags Link Local Signaling TLV
- New bits (DV-bits) are registered in the "OSPFv3 Options" registry
One new value is being added to the registry of OSPFv2 Extended Prefix TLV Sub-TLVs (Metric sub-TLV)

11. Acknowledgements

The author would like to thank Paul Wells and Alvaro Retana for early discussions.

12. References

12.1. Normative References

[I-D.ietf-ospf-ospfv3-lsa-extend]

[I-D.ietf-ospf-prefix-link-attr]


12.2. Informative References


Author’s Address

Anton Smirnov
Cisco Systems, Inc.
De Kleetlaan 6a
Diegem 1831
Belgium

Email: as@cisco.com