OSPF、MTU、および LSA パッキングのテクニカル ノート

内容

概要

このドキュメントでは、Cisco Bug ID CSCse01519 における Open Shortest Path First（OSPF）パケット、最大伝送ユニット（MTU）、リンクステート アドバタイズメント（LSA）、リンク ステート（LS）更新パケットの相互作用について説明します。

OSPF パケットのサイズ

ルータのリンクには MTU があります。OSPF パケットなどの発信パケットをインターフェイス MTU よりも大きくすることはできません。

[Request for Comments（RFC）2328 に OSPF プロトコルのバージョン 2 が記載されています。](#) RFC 2328 の付録 A.1 には、OSPF パケットのカプセル化について次のように記されています。

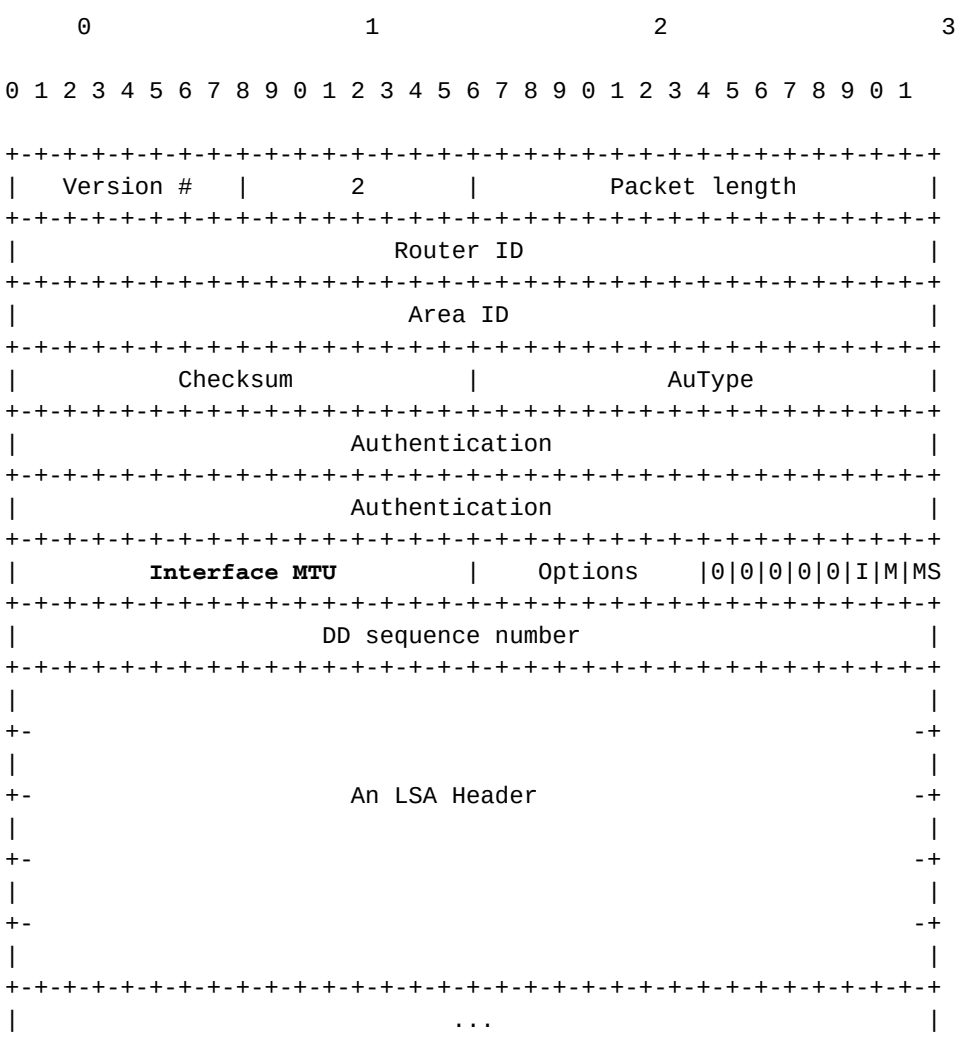
OSPF はインターネット プロトコルのネットワーク層で直接動作します。したがって、OSPF パケットは IP およびローカル データリンクのヘッダーのみによってカプセル化されます。

OSPF では、プロトコル パケットをフラグメント化する方法が定義されず、ネットワーク MTU よりも大きいパケットの送信時には IP フラグメンテーションに依存します。必要に応じて、OSPF パケット長を最大で 65,535 バイト（IP ヘッダーを含む）まで指定できます。 サイズが大きくなる可能性がある OSPF パケット タイプ（Database Description パケット、Link State Request、Link State Update、および Link State Acknowledgment パケット）は、機能が損なわれないように、通常は複数の異なるプロトコル パケットに分割できます。これは次の理由により推奨されています。IP フラグメンテーションは可能な限り回避する必要があります。

LS Update パケットには 1 つ以上の LSA が含まれます。1 つの LS Update パケット内の複数の LSA は、LS Update パケットへの LSA のパッキングとされます。

DBD パケットの MTU

RFC 2328 にも記述されている Database Description（DBD）パケットは、OSPF のリンクステート データベースの内容を示しています。



RFC 2328 の付録 A.3.3 には、インターフェイス MTU について次のように記されています。

フラグメンテーションせずに、関連付けられているインターフェイスに送信できる最大 IP データグラムのバイト単位のサイズ。

OSPF 隣接関係が初期化されると、リンクに接続されたルータは、DBD パケットに含まれるそれぞれのインターフェイス MTU 値を交換します。

RFC 2328 のセクション 10.6 に次の記述があります。

Database Description パケットのインターフェイス MTU フィールドに示されている IP データグラム サイズが、ルータがフラグメンテーションせずに受信インターフェイスで受け取ることができるサイズより大きい場合、Database Description パケットは拒否されます。
debug ip ospf adj コマンドを使用すると、これらの DBD パケットの受信を確認できます。

この例では、2 つの OSPF ネイバー間で MTU 値が一致していません。このルータの MTU は 1600 です。

```
OSPF: Rcv DBD from 10.100.1.2 on GigabitEthernet0/1 seq 0x2124 opt 0x52 flag 0x2
len 1452 mtu 2000 state EXSTART
OSPF: Nbr 10.100.1.2 has larger interface MTU
```

もう 1 つの OSPF ルータのインターフェイス MTU は 2000 です。

```
OSPF: Rcv DBD from 10.100.100.1 on GigabitEthernet0/1 seq 0x89E opt 0x52 flag 0x7
len 32 mtu 1600 state EXCHANGE
OSPF: Nbr 10.100.100.1 has smaller interface MTU
```

OSPF 隣接関係が完全に解除されるまで、DBD パケットが継続的に再送信されます。

```
OSPF: Send DBD to 10.100.1.2 on GigabitEthernet0/1 seq 0x9E6 opt 0x52 flag 0x7
len 32
OSPF: Retransmitting DBD to 10.100.1.2 on GigabitEthernet0/1 [10]
OSPF: Send DBD to 10.100.1.2 on GigabitEthernet0/1 seq 0x9E6 opt 0x52 flag 0x7
len 32
OSPF: Retransmitting DBD to 10.100.1.2 on GigabitEthernet0/1 [11]
%OSPF-5-ADJCHG: Process 1, Nbr 10.100.1.2 on GigabitEthernet0/1 from EXSTART to
DOWN, Neighbor Down: Too many retransmissions
```

OSPF の動作および LS Update パケットへの LSA のパッキング

Cisco Bug ID CSCse01519 以前

Cisco Bug ID CSCse01519 以前は、インターフェイス MTU にかかわらず、Cisco IOS® ソフトウェアの OSPF は 1500 バイトを超える OSPF パケットを作成できませんでした。したがって、インターフェイス MTU が 1500 バイトより大きい場合でも、OSPF は OSPF パケットに 1500 バイトまでしかパッキングしませんでした。OSPF はリンク上でより大きいパケットを送信して、高いスループットを実現できるわけですから、これは非効率的です。

> **注：**このシナリオには例外が 1 つありました。1 つの LSA が 1500 バイトを超えている場合、OSPF は単一の LSA をフラグメント化できないため、サイズに関係なくそのパケットが作成されました。その後ルータの IP スタックが、発信インターフェイスの MTU に収まるようにパケットをフラグメント化します。これは通常、OSPF ルータに複数のリンクがあり、ルータ LSA がリンク MTU よりも大きくなった場合に発生しました。

同様に、発信インターフェイスの MTU が 1500 バイトより小さい場合も、OSPF プロセスによって 1500 バイトを超える OSPF パケットが作成またはパッキングされました。そしてルータの IP スタックが、そのパケットを発信リンクの MTU に収まるようにより小さい IP パケットにフラグメント化します。これは通常、OSPF を実行している 2 台のルータ間の IPSec トンネルで発生しました。トンネルのカプセル化で追加されるオーバーヘッドのバイト数により、MTU は 1500 バイト未満になります。OSPF が最大 1500 バイトの OSPF パケットを作成し、パケットはルータによって送信される前にフラグメント化されます。これがさらに効率を低下させていました。

Cisco Bug ID CSCse01519 以降

Cisco Bug ID CSCse01519 以降、IOS の OSPF では OSPF パケットを 1500 バイトを超えてパッキングすることができます。これは、発信インターフェイスの MTU が 1500 バイトを超える場

合に発生します。多くの情報を 1 つの大きなパケットにパッキングできるので、送信がより効率的になります。つまり、1 台の OSPF ルータが複数の外部 LSA を OSPF ネイバーに送信する必要がある場合、そのルータが Cisco Bug ID CSCse01519 が実装された IOS を実行していれば、より多くの外部 LSA を 1 つの LS Update パケットにパッキングすることができます。

また、Cisco Bug ID CSCse01519 によって、OSPF は 1500 バイトまでのパケットを作成できます。一部のシナリオでは、2 つの OSPF ネイバー間の MTU は 1500 バイト未満です。前述の例の IPSec トンネルでは、OSPF は 1500 バイト未満の OSPF パケットを送信して IP フラグメンテーションを回避します。ここでも、インターフェイス MTU よりも大きい LSA の場合は例外です。

Cisco Bug ID CSCse01519

OSPF ルータのアップグレード時に、Cisco Bug ID CSCse01519 が原因で生じる OSPF MTU 問題が検出される場合があります。

概要

多くのネットワークの OSPF ネイバーは、レイヤ 2（L2）スイッチド ネットワークまたは転送ネットワークを介して接続され、L2 VPN サービスまたは同期デジタル階層/Synchronous Optical NETwork（SDH/SONET）のネットワークで構成されます。これらの転送ネットワークでは、OSPF を実行しているルータと異なる MTU が設定されている場合があります。

MTU はすべてのルータで適切に設定され、実際の MTU を反映している必要がありますが、間違いが見過ごされることがよくあります。

次の例は 2 台のルータが OSPF を実行しているネットワークです。ルータ 1（R1）とルータ 2（R2）は L2 スイッチを介して接続されています。

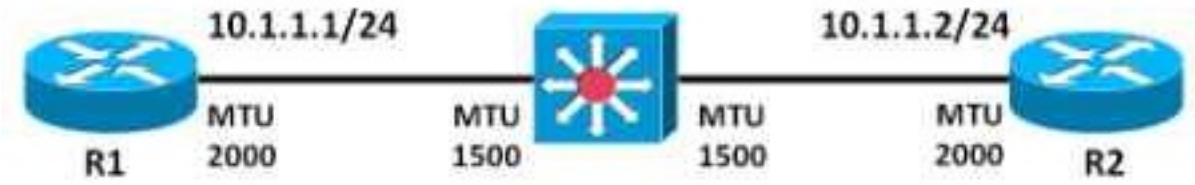


Figure 1： Example network

この例では、ルータの GigabitEthernet インターフェイスで MTU が 2000 に設定されています。L2 スイッチの MTU は 1500 バイトのみです。

データ トラフィックのサイズが 1500 バイト以下であれば、OSPF パケットが 1500 バイトを超えることはないので、Cisco Bug ID CSCse01519 を実装していない IOS を使用できます。ただ

し、たとえば 1800 バイトの LSA が存在する場合、R1 または R2 の OSPF プロセスによって 1500 バイトを超える LS Update パケットが作成および送信されますが、パケットはルータ間の L2 スイッチによって廃棄されます。

R2 の OSPF データベースに十分なネットワークがあると、非常に大きい LSA がローカルから発信されるため、LS Update パケットがインターフェイス MTU を超える可能性があります。

- これらのネットワークがcovering networkコマンドによって発信された場合、ネットワークはR2のルータLSAに表示されます。R2は2000バイトを超えるルータLSAを構築して送信しますが、IPはインターフェイスMTUにフラグメント化します。ただし、L2 スイッチはこれらのパケットを廃棄します。その後 OSPF がこのパケットを無限に再送信し、OSPF 隣接関係が完全な状態になることはありません。したがって、Cisco Bug ID CSCse01519 を実装していない IOS を実行していても、問題はすぐに検出されます。
- **redistribute connected コマンドによって発信されるネットワークは、外部 LSA に表示されます。**OSPF は、サイズが最大 1500 バイトの 1 つの LS Update パケットに外部 LSA をパッキングしようとします。この場合、インターフェイス MTU が 2000 バイトであるため、OSPF 隣接関係は「FULL」状態に達します。基本 MTU が不適切である問題はすぐに検出されません。この問題は、1 台のルータが Cisco Bug ID CSCse01519 を実装した IOS にアップグレードされたときに検出されます。

シナリオ

両方のルータが Cisco Bug ID CSCse01519 を実装していない IOS バージョンを実行すると仮定します。

OSFP 隣接関係が構築されると、インターフェイスの MTU が 2000 であっても、R1 は 1500 バイトを超える OSPF パケットを受信しないことに注意してください。

debug ip ospf packets コマンドを有効にします。

```
OSPF: rcv. v:2 t:1 l:48 rid:10.100.1.2
     aid:0.0.0.0 chk:72CF aut:0 auk: from GigabitEthernet0/1
...
OSPF: rcv. v:2 t:4 l:1468 rid:10.100.1.2
     aid:0.0.0.0 chk:8389 aut:0 auk: from GigabitEthernet0/1
OSPF: rcv. v:2 t:4 l:136 rid:10.100.1.2
...
```

このデバッグ出力にある「l: 1468」は OSPF パケットの長さなので、最大 OSPF パケットが 1468 バイトであることが確認できます。「t:4」は、OSPF パケットがタイプ 4 の Link State Update パケットであることを示しています。RFC 2328 のセクション 4.3 から抜粋した次の表には、各 OSPF パケット タイプが定義されています。

Type	パケット名	プロトコル機能
1	Hello	ネイバーの検出および保持
0	Database Description	データベース コンテンツの要約
3	リンク ステート要求	データベース ダウンロード
4	リンク ステート更新	データベース更新
5	Link State Ack	フラッディング確認応答

OSPF 隣接関係は「FULL」状態に達します。

```
R1#show ip ospf neighbor gigabitEthernet 0/1

Neighbor ID     Pri   State           Dead Time   Address         Interface
10.100.1.2        0   FULL/  -        00:00:34    10.1.1.2        GigabitEthernet0/1

R2#show ip ospf neighbor gigabitEthernet 0/1

Neighbor ID     Pri   State           Dead Time   Address         Interface
10.100.100.1      0   FULL/  -        00:00:34    10.1.1.1        GigabitEthernet0/1
```

次に、R2 の IOS を Cisco Bug ID CSCse01519 を実装した IOS バージョンにアップグレードします。

```
R2#show ip ospf neighbor gigabitEthernet 0/1

Neighbor ID     Pri   State           Dead Time   Address         Interface
10.100.100.1      0   LOADING/  -     00:00:33    10.1.1.1        GigabitEthernet0/1

R2#show ip ospf neighbor gigabitEthernet 0/1 detail
Neighbor 10.100.100.1, interface address 10.1.1.1
   In the area 0 via interface GigabitEthernet0/1
   Neighbor priority is 0, State is LOADING, 5 state changes
   DR is 0.0.0.0 BDR is 0.0.0.0
   Options is 0x12 in Hello (E-bit L-bit )
   Options is 0x52 in DBD (E-bit L-bit O-bit)
   LLS Options is 0x1 (LR)
   Dead timer due in 00:00:39
   Neighbor is up for 00:00:49
   Index 1/1, retransmission queue length 0, number of retransmission 0
   First 0x0(0)/0x0(0) Next 0x0(0)/0x0(0)
   Last retransmission scan length is 0, maximum is 0
   Last retransmission scan time is 0 msec, maximum is 0 msec
   Number of retransmissions for last link state request packet 9
   Poll due in 00:00:00

R2#show ip ospf neighbor gigabitEthernet 0/1 detail
Neighbor 10.100.100.1, interface address 10.1.1.1
   In the area 0 via interface GigabitEthernet0/1
   Neighbor priority is 0, State is LOADING, 5 state changes
   DR is 0.0.0.0 BDR is 0.0.0.0
   Options is 0x12 in Hello (E-bit L-bit )
   Options is 0x52 in DBD (E-bit L-bit O-bit)
   LLS Options is 0x1 (LR)
   Dead timer due in 00:00:33
   Neighbor is up for 00:02:06
   Index 1/1, retransmission queue length 0, number of retransmission 0
   First 0x0(0)/0x0(0) Next 0x0(0)/0x0(0)
   Last retransmission scan length is 0, maximum is 0
   Last retransmission scan time is 0 msec, maximum is 0 msec
   Number of retransmissions for last link state request packet 25
   Poll due in 00:00:03

%OSPF-5-ADJCHG: Process 1, Nbr 10.100.100.1 on GigabitEthernet0/1 from LOADING
to DOWN, Neighbor Down: Too many retransmissions
```

OSPF 隣接関係は「LOADING」状態のままになり、「FULL」状態に達しません。OSPF が再送信の 25 回の制限に達するまで、再送信が行われます。OSPF は隣接関係の確立を再試行し、同じ問題が再度発生してループが無限に続きます。

つまり、R2 でのアップグレードによってこれまで隠れていた問題が発見されます。それは基本 MTU が OSPF ルータで使用する MTU より小さいということです。

スイッチが MTU を 2000 に変更すると、1500 バイトよりも大きい OSPF パケット（「l: 1980」）が問題なく送信されます。

```
R1#
OSPF: rcv. v:2 t:3 l:1980 rid:10.100.1.2
     aid:0.0.0.0 chk:AC5B aut:0 auk: from GigabitEthernet0/1
```

基本 MTU の問題を確認するため、常に MTU と DF（フラグメントなし）ビットの設定と同じサイズで OSPF ネイバー IP アドレスの ping を実行します。

基本 MTU の値を検出するには、ping を実行してサイズをスイープします。適切な MTU を特定するために、出力内の感嘆符（!）の数をカウントします。この例では、**ping コマンドからの最後のエコー応答のサイズは 1500 バイトです。**

```
R2#ping
Protocol [ip]:
Target IP address: 10.1.1.1
Repeat count [5]: 1
Datagram size [100]:
Timeout in seconds [2]:
Extended commands [n]: yes
Source address or interface:
Type of service [0]:
Set DF bit in IP header? [no]: yes
Validate reply data? [no]:
Data pattern [0xABCD]:
Loose, Strict, Record, Timestamp, Verbose[none]:
Sweep range of sizes [n]: yes
Sweep min size [36]: 1460
Sweep max size [18024]: 1540
Sweep interval [1]:
Type escape sequence to abort.
Sending 81, [1460..1540]-byte ICMP Echos to 10.1.1.1, timeout is 2 seconds:
Packet sent with the DF bit set
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!.............................
...........
Success rate is 49 percent (40/81), round-trip min/avg/max = 1/1/4 ms
```