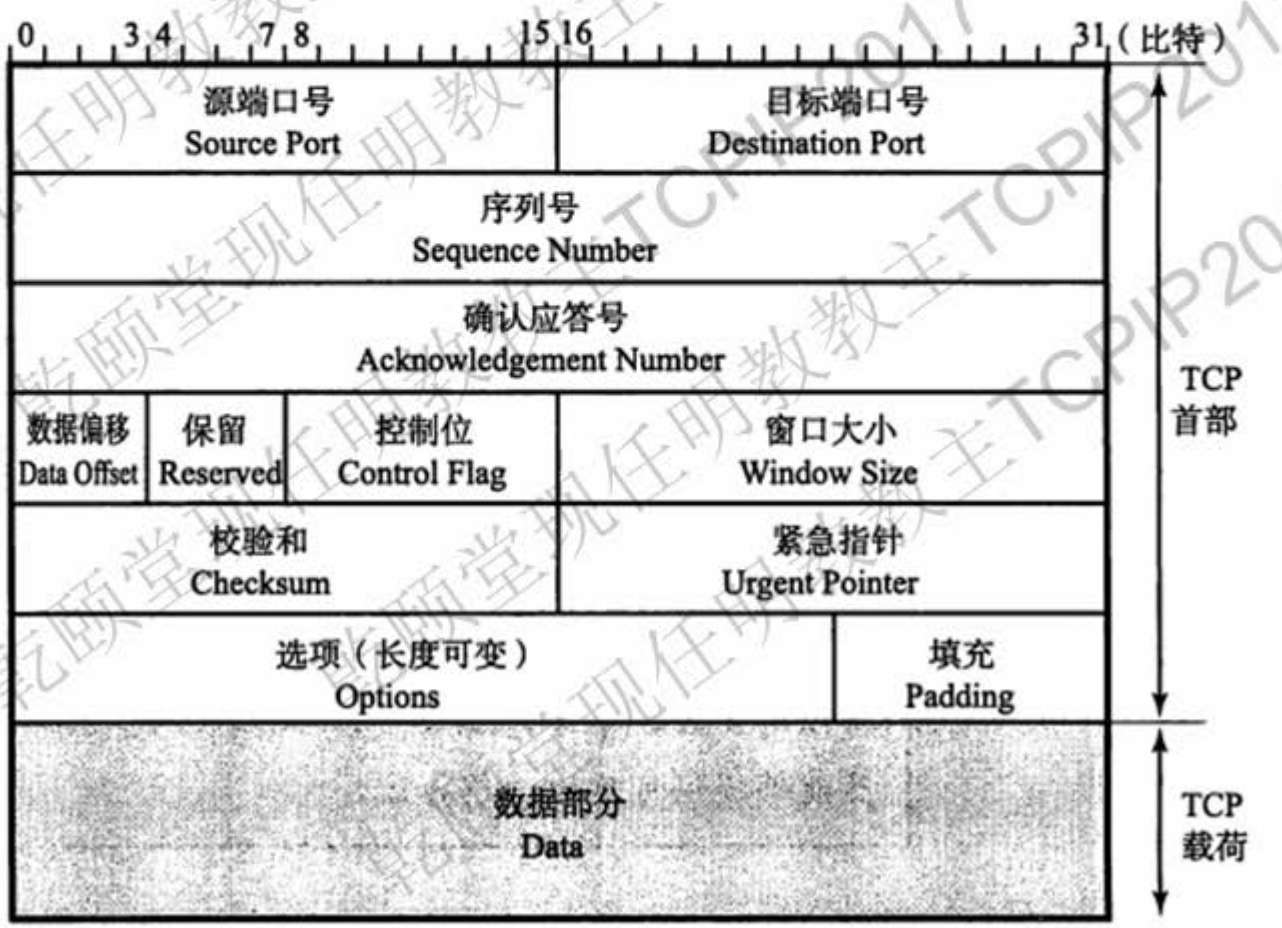
第五部分.2
TCP首部介绍

TCP的首部



TCP的首部（续）

- **源端口号（Source Port）**
  表示发送端端口号，字段长16位。

- **目标端口号（Destination Port）**
  表示接收端端口号，字段长度16位。

- **序列号（Sequence Number）**
  字段长32位。序列号（有时也叫序号）是指发送数据的位置。每发送一次数据，就累加一次该数据字节数的大小。
  序列号不会从0或1开始，而是在建立连接时由计算机生成的随机数作为其初始值，通过SYN包传给接收端主机。然后再将每转发过去的字节数累加到初始值上表示数据的位置。此外，在建立连接和断开连接时发送的SYN包和FIN包虽然并不携带数据，但是也会作为一个字节增加对应的序列号。

TCP的首部（续）

- **确认应答号（Acknowledgement Number）**

确认应答号字段长度32位。是指下一次应该收到的数据的序列号。实际上，**它是指已收到确认应答号减一为止的数据。**发送端收到这个确认应答以后可以认为在这个序号以前的数据都已经被正常接收。

- **数据偏移（Data Offset）[TCP首部长度]**

该字段表示TCP所传输的数据部分应该从TCP包的哪个位开始计算，当然也可以把它看作**TCP首部的长度**。该字段长4位，单位为4字节（即32位）。不包括选项字段的话，TCP的首部为20字节长，因此数据偏移字段可以设置为5。反之，如果该字段的值为5，那说明从TCP包的最一开始到20字节为止都是TCP首部，余下的部分为TCP数据。

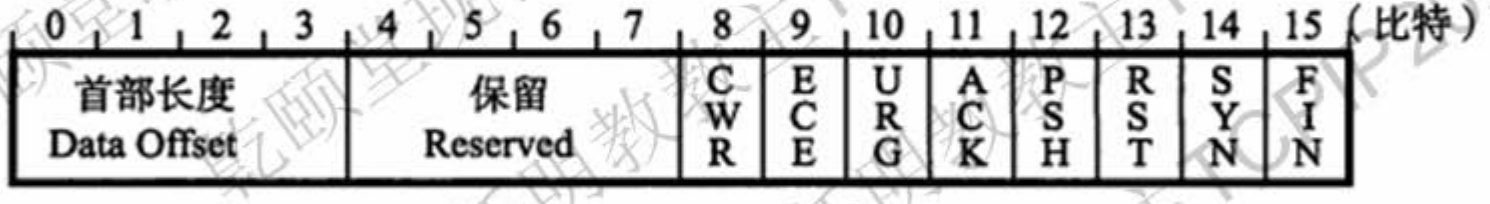
- **保留（Reserved）**

该字段主要是为了以后扩展时使用，其长度为4位。一般设置为0，**但即使收到的包在该字段不为0（安全问题），**此包也不会被丢弃。

TCP的首部（续）

- 控制位（Control Flag）

字段长为8位，每一位从左至右分别为URG、ACK、PSH、RST、SYN、FIN。这些控制标志也叫做控制位。当它们对应位上的值为1时，具体含义如下图所示。



1. URG（Urgent Flag）

该位为1时，表示包中有需要紧急处理的数据。对于需要紧急处理的数据，会在后面的紧急指针中再进行解释。

2. ACK（Acknowledgement Flag）

该位为1时，确认应答的字段变为有效。TCP规定除了最初建立连接时的SYN包之外该位必须设置为1。

TCP的首部（续）

3. PSH（Push Flag）

该位为1时，表示需要将受到的数据立刻传给上层应用协议。PSH为0时，则不需要立即传而是先进行缓存。

4. RST（Reset Flag）

该位为1时表示TCP连接中出现异常必须强制断开连接。例如，一个没有被使用的端口即使发来连接请求，也无法进行通信。此时就可以返回一个RST设置为1的包。此外，程序宕掉或切断电源等原因导致主机重启的情况下，由于所有的连接信息将全部被初始化，所以原有的TCP通信也将不能继续进行。这种情况下，如果通信对方发送一个设置为1的RST包，就会使通信强制断开连接。

5. SYN（Synchronize Flag）

用于建立连接。SYN为1表示希望建立连接，并在其序列号的字段进行序列号初始值的设定。

TCP的首部（续）

6. FIN（FIN Flag）

该位为1时，**表示今后不会再有数据发送，希望断开连接（单向传输断开）**。当通信结束希望断开连接时，通信双方的主机之间就可以相互交换FIN位置为1的TCP段。每个主机又对对方的FIN包进行确认应答以后就可以断开连接。不过，**主机收到FIN设置为1的TCP段以后不必马上回复一个FIN包，而是可以等到缓冲区中的所有数据都因已成功发送而被自动删除之后再发（半闭连接）**。

TCP的首部（续）

7. ECN-Echo（ECE）

8. Congestion Window Reduced（CWR）

TCP支持使用TCP头中的两个标记（flag）来支持ECN。分别为用于回传拥塞指示（即指示发送者应减少信息发送量）和确认接收到了拥塞指示回应。这即是ECN-Echo（ECE）和Congestion Window Reduced（CWR）位。

在TCP连接上使用ECN是可选的；当ECN被使用时，它必须在连接创建时通过SYN和SYN-ACK段中包含适当选项来协商。

当在一个TCP连接上协商ECN后，发送方指示连接上的TCP段携带IP分组传输流量，将支持ECN的传输用ECT码点标记。这使支持ECN的中间路由器可以标记具有CE码点的IP分组而不是丢弃它们，以指示即将发生的阻塞。

当接收到具有遇到阻塞码点时，TCP接收者使用TCP头中的ECE标记回传这个阻塞指示。当一个端点收到TCP带有ECE位的段时，它减少其拥塞窗口来代替丢包。然后，它设置段的CWR位来确认阻塞指示。

节点保持传输设置有ECE位的TCP段，直到它接收到设置有CWR的段。

TCP的首部（续）

- 窗口大小（Windows Size）

该字段长为16位。用于通知从相同TCP首部的确认应答号所指位置开始能够接收的数据大小（8位字节）。TCP不允许发送超过此处所示大小的数据。不过，如果窗口为0，则表示可以发送窗口探测，以了解最新的窗口大小。但这个数据必须是1个字节。

- 校验和（Checksum）

TCP校验和覆盖TCP首部和TCP数据，而IP首部中的校验和只覆盖IP的首部，不覆盖IP数据报中的任何数据。

TCP的校验和是必需的，而UDP的校验和是可选的。

TCP和UDP计算校验和时，都要加上上图所示的一个12字节的伪首部。

TCP的首部（续）

- **紧急指针（Urgent Pointer）**

该字段长为16位。只有在URG控制位为1时有效。该字段的数值表示本报文段中紧急数据的指针。正确来讲，从数据部分的首位到紧急指针所指示的位置为止为紧急数据。因此也可以说紧急指针指出了紧急数据的末尾在报文段中的位置。

如何处理紧急数据属于应用的问题。一般在暂时中断通信，或中断通信的情况下使用。例如在Web浏览器中点击停止按钮，或者使用TELNET输人Ctrl+C 时都会有URG为1的包。此外，紧急指针也用作表示数据流分段的标志。

- **选项（Options）**

选项字段用于提高TCP的传输性能。因为根据数据偏移（首部长度）进行控制，所以其长度最大为字节。

另外，选项字段尽量调整其为**32位的整数倍**。具有代表性的选项如下图所示。

TCP的首部（续）

类型	长度	意 义	RFC
0	-	End of Option List	RFC793
1	-	No-Operation	RFC793
2	4	Maximum Segment Size	RFC793
3	3	WSOPT-Window Scale	RFC1323
4	2	SACK Permitted	RFC2018
5	N	SACK	RFC2018
8	10	TSOPT-Time Stamp Option	RFC1323
27	8	Quick-Start Response	RFC4782
28	4	User Timeout Option	RFC5482
29	-	TCP Authentication Option （TCP-AO）	RFC5925
253	N	RFC3692-style Experiment 1	RFC4727
254	N	RFC3692-style Experiment 2	RFC4727

TCP的首部（续）

类型2： 的MSS选项用于在建立连接时决定最大段长度的情况。这选项用于大部分操作系统。

类型3： 的窗口扩大，是一个用来改善TCP吞吐量的选项。TCP首部中窗口字段只有16位。因此在TCP包的往返时间(RTT)内，只能发送最大64K字节的数据。如果采用了该选项，窗口的最大值可以扩展到1G字节。由此，即使在一个RTT较长的网络环境中，也能达到较高的吞吐量。

窗口大小与吞吐量

TCP 通信的最大吞吐量由窗口大小和往返时间决定。假定最大吞吐量为 T_{max}，窗口大小为 W，往返时间是 RTT 的话，那么最大吞吐量的公式如下：

$$T_{max}=\frac{W}{RTT}$$

假设窗口为 65535 字节，RTT 为 0.1 秒，那么最大吞吐量 T_{max} 如下：

$$T_{max}=\frac{65535\ (字节)}{0.1\ (秒)}=\frac{65535\times 8\ (比特)}{0.1\ (秒)}$$

$$=5242800\ (bps)\fallingdotseq 5.2\ (Mbps)$$

以上公式表示 1 个 TCP 连接所能传输的最大吞吐量为 5.2Mbps。如果建立两个以上连接同时进行传输时，这个公式的计算结果则表示每个连接的最大吞吐量。也就是说，在 TCP 中，与其使用一个连接传输数据，使用多个连接传输数据会达到更高的网络吞吐量。在 Web 浏览器中一般会通过同时建立 4 个左右连接来提高吞吐量。

多连接能提升吞吐量

窗口扩大选项介绍

Kind:8 bits. Set to 3.
Length:8 bits. Set to 3.
Shift count:8 bits.

TCP窗口扩大因子是一个新的TCP选项，一些新的实现才会包含该选项，为了是新旧协议兼容，做了如下约定：

- 只有主动连接方的第一个SYN可以发送窗口扩大因子
- 被动连接方接收到带有窗口扩大因子的选项后，如果支持，则可以发送自己的窗口扩大因子，否则忽略该选项
- 如果双方支持该选项，那么后续的数据传输则使用该窗口扩大因子

The scale factor is used to shift the window field before the data segment is sent. The scale factor is limited to 14 to guarantee the window is less than the 2^{31} maximum. Thus, $2^{14} * (2^{16} - 1) < 2^{31}$ (or $(2^{16} - 1) << 14 < 2^{31}$, if you prefer).

介绍选项与各种系统激活窗口扩大的方法
https://en.wikipedia.org/wiki/TCP_window_scale_option

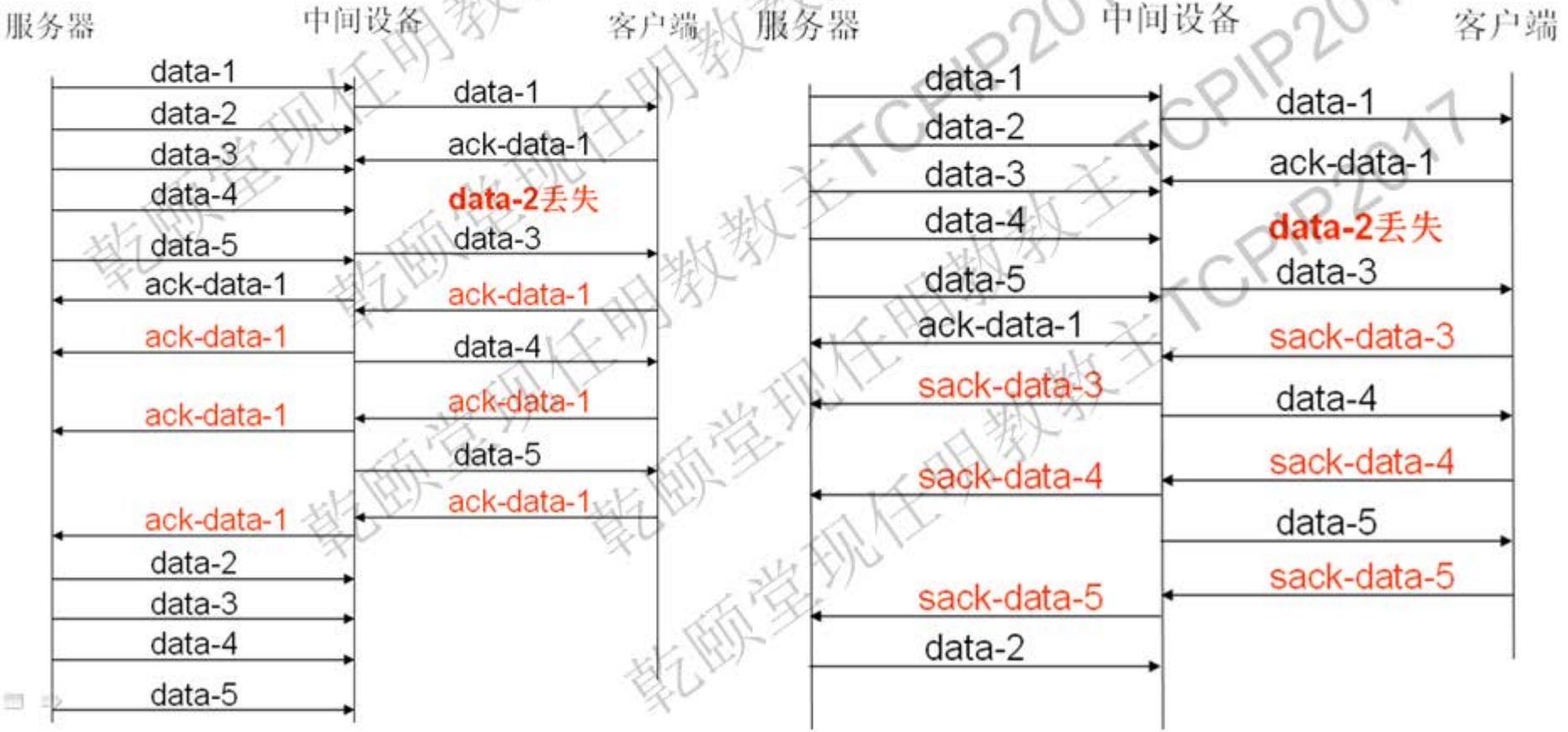
窗口扩大选项抓包

百度网盘抓包

TCP_Option_窗口扩大_选择ACK

```
TCP - Transport Control Protocol
    Source Port:            57137 [34-35]
    Destination Port:       443  https [36-37]
    Sequence Number:        519537490 [38-41]
    Ack Number:             0 [42-45]
    TCP Offset:             8  (32 bytes) [46 Mask 0xE0]
    Reserved:               %0000 [46 Mask 0x0F]
    TCP Flags:              %00000010  ......S. [47]
                              0... .... (No Congestion Window Reduction)
                              .0.. .... (No ECN-Echo)
                              ..0. .... (No Urgent pointer)
                              ...0 .... (No Ack)
                              .... 0... (No Push)
                              .... .0.. (No Reset)
                              .... ..1. SYN
                              .... ...0 (No FIN)
    Window:                 8192 [48-49]
    TCP Checksum:           0xD510  Checksum invalid. Should be:  0xF6EF [50-51]
    Urgent Pointer:         0 [52-53]
    TCP Options:
      Option Type:            2  Maximum Segment Size [54]
        Length:               4 [55]
        MSS:                  1460 [56-57]
      Option Type:            1  No Operation [58]
      Option Type:            3  Window Scale Factor [59]
        Length:               3 [60]
        Shift Count:          8 [61]
      Option Type:            1  No Operation [62]
      Option Type:            1  No Operation [63]
      Option Type:            4  Sack Supported [64]
        Length:               2 [65]
FCS - Frame Check Sequence
    FCS:                    0xFA8415D8  Calculated
```

SACK示意图


服务器
中间设备
客户端
data-1
data-2
data-3
data-4
data-5
ack-data-1
ack-data-1
ack-data-1
ack-data-1
data-2
data-3
data-4
data-5
data-1
ack-data-1
data-2丢失
data-3
ack-data-1
data-4
ack-data-1
data-5
ack-data-1
未激活SACK
服务器
中间设备
客户端
data-1
data-2
data-3
data-4
data-5
ack-data-1
sack-data-3
sack-data-4
sack-data-5
data-2
data-1
ack-data-1
data-2丢失
data-3
sack-data-3
data-4
sack-data-4
data-5
sack-data-5
激活SACK

现任明教教主
乾颐堂TCP/IP 2017

SACK抓包

SACK允许选项:其类型值为4，该选项只允许在有SYN标志的TCP包中，也即TCP握手的前两个包中，分别表示各自是否支持SACK。

```
TCP - Transport Control Protocol
    Source Port:            57137 [34-35]
    Destination Port:       443 https [36-37]
    Sequence Number:        519537490 [38-41]
    Ack Number:             0 [42-45]
    TCP Offset:             8 (32 bytes) [46 Mask 0xF0]
    Reserved:               %0000 [46 Mask 0x0F]
  TCP Flags:                %00000010 .....S. [47]
      0... .... (No Congestion Window Reduction)
      .0.. .... (No ECN-Echo)
      ..0. .... (No Urgent pointer)
      ...0 .... (No Ack)
      .... 0... (No Push)
      .... .0.. (No Reset)
      .... ..1. SYN
      .... ...0 (No FIN)
    Window:                 8192 [48-49]
    TCP Checksum:           0xD510 Checksum invalid. Should be: 0xF6EF [50-51]
    Urgent Pointer:         0 [52-53]
  TCP Options:
    Option Type:            2 Maximum Segment Size [54]
      Length:               4 [55]
      MSS:                  1460 [56-57]
    Option Type:            1 No Operation [58]
    Option Type:            3 Window Scale Factor [59]
      Length:               3 [60]
      Shift Count:          8 [61]
    Option Type:            1 No Operation [62]
    Option Type:            1 No Operation [63]
    Option Type:            4 Sack Supported [64]
      Length:               2 [65]
FCS - Frame Check Sequence
    FCS:                    0xFA8415D8 Calculated
```

SACK选项:其选项类型值为5，选项长度可变，用来选择性的确认数据的序列号范围。

```
TCP - Transport Control Protocol
    Source Port:            443 https [34-35]
    Destination Port:       57305 [36-37]
    Sequence Number:        933743551 [38-41]
    Ack Number:             2731764466 [42-45]
    TCP Offset:             8 (32 bytes) [46 Mask 0xF0]
    Reserved:               %0000 [46 Mask 0x0F]
  TCP Flags:                %00010000 ...A.... [47]
      0... .... (No Congestion Window Reduction)
      .0.. .... (No ECN-Echo)
      ..0. .... (No Urgent pointer)
      ...1 .... Ack
      .... 0... (No Push)
      .... .0.. (No Reset)
      .... ..0. (No SYN)
      .... ...0 (No FIN)
    Window:                 632 [48-49]
    TCP Checksum:           0xF481 [50-51]
    Urgent Pointer:         0 [52-53]
  TCP Options:
    Option Type:            1 No Operation [54]
    Option Type:            1 No Operation [55]
    Option Type:            5 Selective Acknowledgement [56]
      Length:               10 [57]
      SACK From:            2731763334 [58-61]
      SACK To:              2731764466 [62-65]
FCS - Frame Check Sequence
    FCS:                    0xD5F6078E Calculated
```

TCP_Option_选择ACK.p

百度网盘抓包

密钥认证BGP穿越ASA问题

默认密钥认证的BGP无法穿越ASA

```
*Mar  1 04:16:27.566: %TCP-6-BADAUTH: No MD5 digest from 10.1.1.1(19778
) to 202.100.1.1(179)
```

EBGP配置

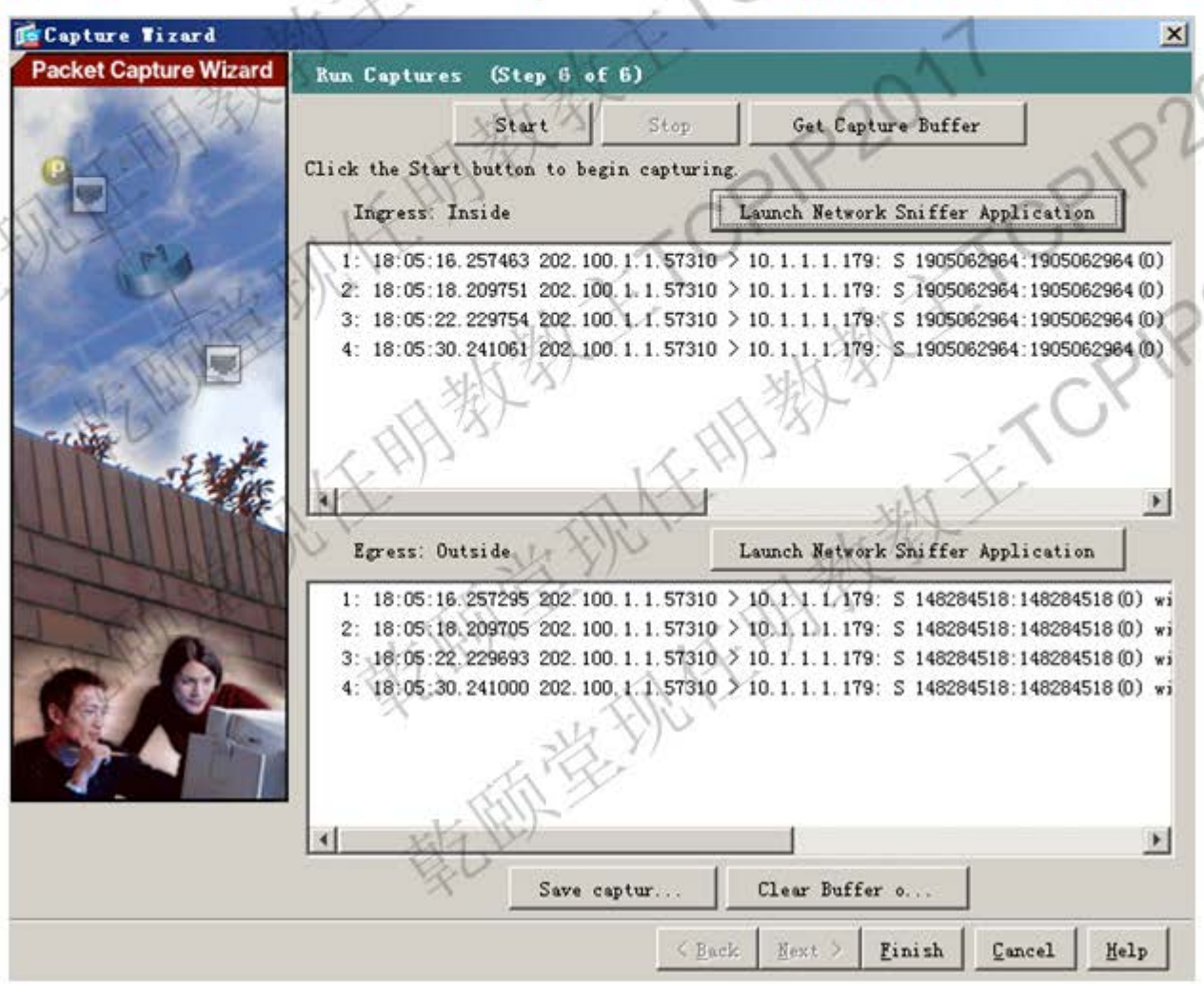
Outside BGP配置:

```
router bgp 101
 bgp log-neighbor-changes
 neighbor 10.1.1.1 remote-as 102
 neighbor 10.1.1.1 password cisco
 neighbor 10.1.1.1 ebgp-multihop 255
 neighbor 10.1.1.1 update-source GigabitEthernet1
ip route 10.1.1.0 255.255.255.0 202.100.1.10
```

Inside BGP配置:

```
router bgp 102
 bgp log-neighbor-changes
 neighbor 202.100.1.1 remote-as 101
 neighbor 202.100.1.1 password cisco
 neighbor 202.100.1.1 ebgp-multihop 255
 neighbor 202.100.1.1 update-source GigabitEthernet1
ip route 202.100.1.0 255.255.255.0 10.1.1.10
```

使用Packet Capture捕获数据包



现任明教教主
乾颐堂TCP/IP 2017

抓包比较

外部原始数据包

```
TCP - Transport Control Protocol
    Source Port:            57310
    Destination Port:       179  bgp
    Sequence Number:        148284518
    Ack Number:             0
    TCP Offset:             11  (44 bytes)
    Reserved:               %0000
    TCP Flags:              %00000010  .....S.
        0... .... (No Congestion Window Reduction)
        .0.. .... (No ECN-Echo)
        ..0. .... (No Urgent pointer)
        ...0 .... (No Ack)
        .... 0... (No Push)
        .... .0.. (No Reset)
        .... ..1. SYN
        .... ...0 (No FIN)
    Window:                 16384
    TCP Checksum:           0x3FF4
    Urgent Pointer:         0
    TCP Options:
        Option Type:        2  Maximum Segment Size
            Length:         4
            MSS:            516
        Option Type:        19  MD5 Signature
            Length:         18
            Opt Value:      0x9E3BC3A2D53475DE7BACC18E63E80671
        Option Type:        0  End Of Option List
        Option Type:        0  End Of Option List
```

现在新版本默认放过Option19

穿越ASA后数据包

```
TCP - Transport Control Protocol
    Source Port:            57310
    Destination Port:       179  bgp
    Sequence Number:        1905062964
    Ack Number:             0
    TCP Offset:             11  (44 bytes)
    Reserved:               %0000
    F=%00000010 . . . . . . S .
    Window:                 16384
    TCP Checksum:           0xE5FE
    Urgent Pointer:         0
    TCP Options:
        Option Type:        2  Maximum Segment Size
            Length:         4
            MSS:            516
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
        Option Type:        1  No Operation
```

解决密钥认证BGP穿越ASA问题（CLI）

严重注意：两台路由器需要有抵达对方的明细路由

```
access-list out extended permit tcp host 202.100.1.1 host 10.1.1.1 eq bgp

class-map BGP
 match port tcp eq bgp

tcp-map Allow-TCP-Option-19
  tcp-options range 19 19 allow

policy-map global_policy
 class BGP
  set connection random-sequence-number disable
  set connection advanced-options Allow-TCP-Option-19
```