TCP improvements in the Windows network stack

Praveen Balasubramanian

pravb@microsoft.com

Quick recap

- Anniversary update for Windows 10 on nearly all 400 million+ devices running Windows 10
- Server 2016 in market
- Transport improvements
  - Tail Loss Probe (TLP) enabled by default when RTT > 10 msec
  - Recent ACKnowledgement (RACK) enabled by default when RTT > 10 msec
  - IW10 enabled by default for all connections
  - TFO (TCP Fast Open) available as a experimental feature in the Edge browser
  - LEDBAT* being used for internal workloads like crash dump uploads

    * with some proprietary modifications
- Coming soon – Windows 10 Creators update, free update to all Windows 10 devices

TCP Fast Open updates

- TCP global setting was already enabled by default
- Ending the Mexican standoff
  - TFO is now on by default in Microsoft Edge browser in Windows Insider Preview builds 14986 and higher
  - HTTPS only, no proxy
  - Telemetry issues so no data to share – we will share data at a later time
- Fallback heuristics
  - Stop negotiating or using TFO on SYN retransmit
  - Per network, persisted
  - Exponential backoff and retry
- Fully functional server side support
- Request to community: Enable TFO on servers, report issues, report server success metrics, fix broken middleboxes

Experimental support for CUBIC

- Based on draft-ietf-tcpm-cubic
- Includes a fix for the “Quiescence bug”
- No HyStart – standard slow start
- On a system with Creators update (builds 15014+), run elevated:
  - **netsh int tcp set supplemental template=internet congestionprovider=cubic**
- Some observations from lab measurements:
  - CUBIC has better single flow performance than both CTCP and New Reno
  - CUBIC dominates when competing with CTCP or New Reno flows on a shared bottleneck link
  - CUBIC has better RTT fairness than both New Reno and CTCP
  - CUBIC builds up large buffers in absence of AQM

Delayed ACKs, TLP and WCDelAckT, ABC

- Switched the default delayed ACK timeout to 40 msec
- In Tail Loss Probe for the case where one packet is outstanding:

```
PTO = max(PTO, 1.5 * SRTT + WCDelAckT)
```

  WCDelAckT is set to 200 msec which makes TLP less effective, switching to lower values causes issues with ping-pong apps talking to older OS

- Suggested improvement: Negotiation / Receiver delayed ACK heuristic
- RFC recommends the ABC (appropriate byte counting) limit of SMSS bytes even in slow start:

  > We note that [RFC3465] allows for cwnd increases of more than SMSS bytes for incoming acknowledgments during slow start on an experimental basis; however, such behavior is not allowed as part of the standard.

- Windows used a value of 4 SMSS previously, now switched to 8 SMSS to better handle stretch ACKs, ACK coalescing, LRO etc.

TCP stats API

- Since Vista / Server 2008 – Estats API which is admin only
- In Creators update, a new per socket API called SIO_TCP_INFO
  - Modeled after the Linux TCP_INFO API
  - Versioned, so we can expand it to add more information over time

```
typedef struct _TCP_INFO_v0 {
    TCPSTATE State;
    ULONG Mss;
    ULONG64 ConnectionTimeMs;
    BOOLEAN TimestampsEnabled;
    ULONG RttUs;
    ULONG MinRttUs;
    ULONG BytesInFlight;
    ULONG Cwnd;
    ULONG SndWnd;
    ULONG RcvWnd;
    ULONG RcvBuf;
    ULONG64 BytesOut;
    ULONG64 BytesIn;
    ULONG BytesReordered;
    ULONG BytesRetrans;
    ULONG FastRetrans;
    ULONG DupAcksIn;
    ULONG TimeoutEpisodes;
    UCHAR SynRetrans;
} TCP_INFO_v0, *PTCP_INFO_v0;
```

```
TCP_INFO_v0 info;
DWORD version = 0;
DWORD bytes_returned;
int ret;

ret =
    WSAIoctl(
        s, // SOCKET
        SIO_TCP_INFO,
        &version, sizeof(version),
        &info, sizeof(info),
        &bytes_returned,
        0, 0);
if (ret == SOCKET_ERROR) {
    printf("ERROR: %d\n", WSAGetLastError());
    return;
}
```

Q&A