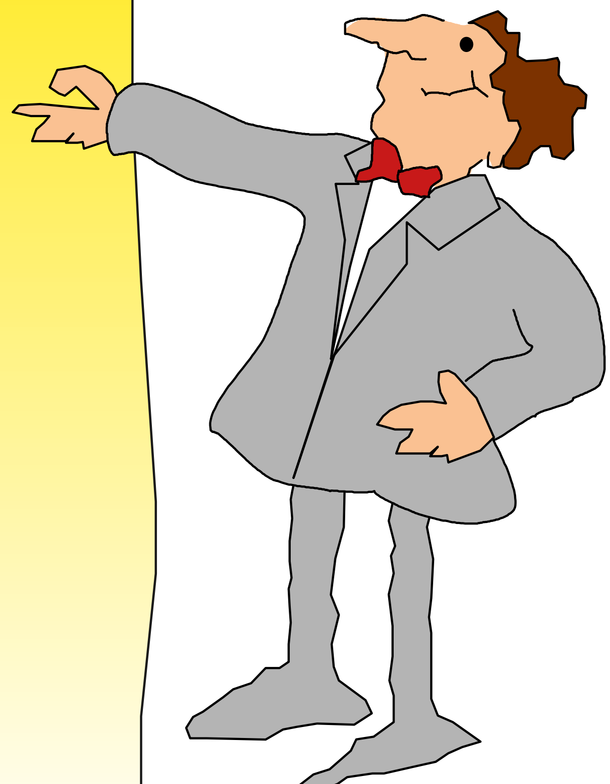
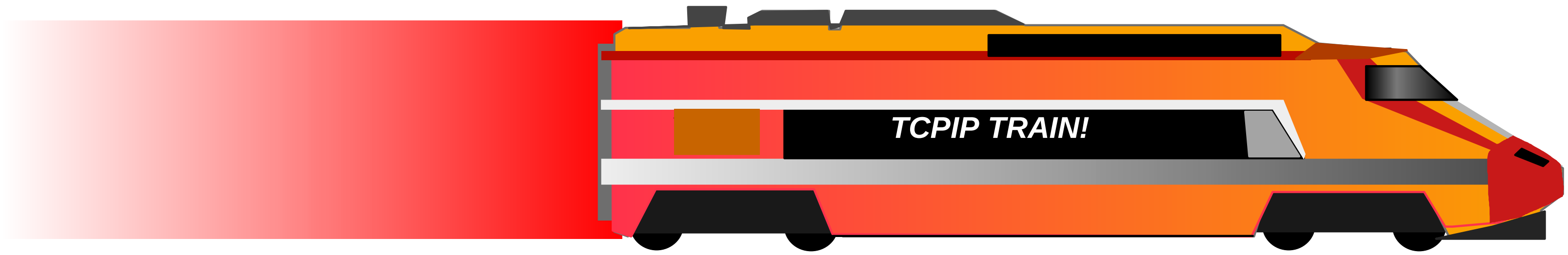
Using Traces for TCP/IP Throughput Performance Problems

- How do we determine if we have a throughput performance problem?
- Ways in z/OS CS to measure throughput performance
- Typical FTP data transfer
- Thruput issues:
  - The effect of high network latency
  - The effect of dropped packets
  - The effect of dropping window size
  - The effect of packet fragmentation
- What to look for in SYSTCPDA Component Trace





Debugging Throughput Performance Problems

What constitutes a throughput performance problem?

- We typically classify problems into two issues:
  - Slow data transfers
  - Terminated or aborted data transfers
- Note that--for the purposes of this presentation--there are really two types of terminated/aborted transfers:
  - Those caused by severe performance problems leading to timeout conditions (dropped packets, failure to recover from a zero sized window, etc.)
  - Those caused by non-performance related issues (connections reset by firewalls, application errors, etc.)
- Only performance related issues will be discussed in this presentation

What metrics do we can we use to calculate a transfer's performance?

- Network Latency
  - This is defined as the amount of time it takes for a packet to reach its destination
  - We have three ways of obtaining a rough estimate of this from our packet traces:
    - The delta time between syn packets in a transfer
    - The average RTT (round trip time) listed in a **session** formatted packet trace
    - Find the data delta between the last acknowledged packet and the last data packet
  - In both the delta time and the RTT, the latency will be roughly half the value

- Bandwidth
  - This is defined as the average amount of data that is transferred over a set period of time
  - We can determine this from the throughput listed in a session formatted packet trace

Gathering SYSTCPDA trace on z/OS

- SYSTCPDA (PKTTRACE) parameters
  - V TCPIP,tcp_proc,PKTTRACE,CLEAR
  - V TCPIP,tcp_proc,P,IP=xx.xx.xx.xx
  - TRACE CT,ON,COMP=SYSTCPDA,SUB=(tcp_proc)
    - R xx,WTR=PKTWRT,END

Step by step:

1) TRACE CT,WTRSTART=PKTWRT (starts the writer for packet tracing)

2) Clear the previous packet trace settings:
   V TCPIP,tcp_proc,PKTTRACE,CLEAR

3) Set TCPIP Packet Trace settings
   V TCPIP,tcp_proc,PKTTRACE,**FULL**,IP=ipaddress of the client

4) Start Packet trace/Connect the writer
   TRACE CT,ON,COMP=SYSTCPDA,SUB=(tcpipproc)
   R xx,WTR=PKTWRT,END

5) *** RECREATE SCENARIO ***

6) TRACE CT,OFF,COMP=SYSTCPDA,SUB=(tcpipproc)

7) TRACE CT,WTRSTOP=PKTWRT,FLUSH

Formatting SYSTCPDA trace on z/OS

- Use IPCS
  - IPCS Interactive

Option 2.7.1.d

```
System       ===>           (System name or blank)
Component    ===>           (Component name (required))
Subnames     ===>

GMT/LOCAL    ===> G                        (G or L, GMT is default)
Start time   ===>                          (mm/dd/yy,hh:mm:ss.dddddd or
Stop time    ===>                           mm/dd/yy,hh.mm.ss.dddddd)
Limit        ===> 0         Exception ===>
Report type ===> SHORT      (SHort, SUmmary, Full, Tally)
User exit    ===>           (Exit program name)
Override source ===>
Options         ===>

To enter/verify required values, type any character
Entry IDs ===>   Jobnames ===>   ASIDs ===>   OPTIONS ===>   SUBS ===>
```

ENTER = update CTRACE definition. END/PF3 = return to previous panel.

```
S = start CTRACE.  R = reset all fields.
```

  - IPCS in batch

IPCS
SETDEF NOCONFIRM PRINT NOTERM
DROPDUMP DDNAME(PKTTRACE)
CTRACE COMP(SYSTCPDA) LOCAL FULL OPTIONS((SESS))
SETDEF CONFIRM NOPRINT TERM

- Refer to:
  z/OS V1R6.0 MVS Interactive Problem Control System (IPCS) Commands - chapter 8

SYSTCPDA SESSion output

```
10 packets summarized

 Local Ip Address:                                   10.1.4.225
 Remote Ip Address:                                12.106.255.57


Host:                                     Local,          Remote
 Client or Server:                      Unknown,         Unknown
 Port:                                      397,            2765
 Application:                              mptn,
 Link speed (parm):                          10,              10 Megabits/s

 Connection:
 First timestamp:                 2005/02/15 11:00:07.494635
 Last timestamp:                  2005/02/15 11:11:34.991160
 Duration:                                   00:11:27.496525
 Average Round-Trip-Time:                              0.000 sec
 Final Round-Trip-Time:                                0.000 sec
 Final state:                                    ESTABLISHED
 Out-of-order timestamps:                                  0
```

SYSTCPDA SESSion output, continued

```
Data Quantity & Throughput:               Inbound,         Outbound
 Application data bytes:                        0,                0
 Sequence number delta:                         0,                0
 Total bytes Sent:                              0,                0
 Bytes retransmitted:                           0,                0
 Throughput:                                    0,                0 Kilobytes/s
 Bandwidth utilization:                     0.67%,            0.67%
 Delay ACK Threshold:                         200,              200 ms
 Minimum Ack Time:                      0.000000,          0.000000
 Average Ack Time:                      0.000000,          0.000000
 Maximum Ack Time:                      0.000000,          0.000000

Data Segment Stats:                       Inbound,         Outbound
 Number of data segments:                       0,                0
 Maximum segment size:                          0,                0
 Largest segment size:                          0,                0
 Average segment size:                          0,                0
 Smallest segment size:                         0,                0
 Segments/window:                             0.0,              0.0
 Average bytes/window:                          0,                0
 Most bytes/window:                             0,                0
```

SYSTCPDA SESSion output, continued

```
Window Stats:                         Inbound,        Outbound
 Number of windows:                         0,               0
 Maximum window size:                       0,           32697
 Largest window advertised:                 0,           32697
 Average window advertised:                 0,           32697
 Smallest window advertised:                0,           32697
 Window scale factor:                       0,               0
 Window frequency:                          0,               0 Windows/s
 Time Stamp updates:                        0,               0
 Total Round Trip Time:              0.000000,        0.000000 (  0%), (  0%)
 Average Round Trip Time:            0.000000,        0.000000
 Maximum Data in Pipe:                      0,               0
 Maximum retransmisson:                     0,               0
```

SYSTCPDA SESSion output, continued

```
Number of:                               Inbound,        Outbound
 Packets:                                     0,              10
  (x) Untraced Packets:                       0,               0
  (.) In-order data:                          0,               0 ( 0.00%), ( 0.00%)
  (a) Acknowledgments:                        0,               0 ( 0.00%), ( 0.00%)
  (+) Data and ACK:                           0,               0 ( 0.00%), ( 0.00%)
  (u) Duplicate ACKs:                         0,               9 ( 0.00%), (90.00%)
  (w) Window size updates:                    0,               1 ( 0.00%), (10.00%)
  (z) Zero window sizes:                      0,               0 ( 0.00%), ( 0.00%)
  (p) Window probes:                          0,               0 ( 0.00%), ( 0.00%)
  (k) Keepalive segments:                     0,               0 ( 0.00%), ( 0.00%)
  (r) Retransmissions:                        0,               0 ( 0.00%), ( 0.00%)
  (o) Out-of-order:                           0,               0 ( 0.00%), ( 0.00%)
  (d) Delayed ACKs:                           0,               0 ( 0.00%), ( 0.00%)
  (f) Fragments:                              0,               0 ( 0.00%), ( 0.00%)

Time Spent on:                           Inbound,        Outbound
  (.) In-order data:            00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (a) Acknowledgments:          00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (+) Data and ACK:             00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (u) Duplicate ACKs:           00:00:00.000000, 00:11:27.496525 ( 0.00%), (78.12%)
  (w) Window size updates:      00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (z) Zero window sizes:        00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (p) Window probes:            00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (k) Keepalive segments:       00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (r) Retransmissions:          00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (o) Out-of-order:             00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (d) Delayed ACKs:             00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (f) Fragments:                00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
```

SYSTCPDA SESSion output, continued

```
Number of:                                   Inbound,        Outbound
  (   S ) SYN:                                     0,                0 ( 0.00%), ( 0.00%)
  ( A S ) ACK SYN:                                 0,                0 ( 0.00%), ( 0.00%)
  (    F) FIN:                                     0,                0 ( 0.00%), ( 0.00%)
  ( A  F) ACK FIN:                                 0,                0 ( 0.00%), ( 0.00%)
  (  R  ) RST:                                     0,                0 ( 0.00%), ( 0.00%)
  (U    ) URG:                                     0,                0 ( 0.00%), ( 0.00%)

Time Spent on:                               Inbound,        Outbound
  (   S ) SYN:                     00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  ( A S ) ACK SYN:                 00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (    F) FIN:                     00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  ( A  F) ACK FIN:                 00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (  R  ) RST:                     00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (U    ) URG:                     00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
```

SYSTCPDA SESSion output, continued

TcpHdr	IO	F	Seq	Ack	RcvWnd	Data	Delta Time	TimeStamp	RcdNr	State	Inf	Ip_id	Rtt
AP	O	w	2691883512	1129000329	32697	0	0.000000	11:00:07.494635	76	ESTABLISHED	1	A828	0.00
AP	O	u	2691883512	1129000329	32697	0	76.474337	11:01:23.968972	87	ESTABLISHED	1	E477	0.00
AP	O	u	2691883512	1129000329	32697	0	76.421369	11:02:40.390341	123	ESTABLISHED	1	41B3	0.00
AP	O	u	2691883512	1129000329	32697	0	76.445911	11:03:56.836252	136	ESTABLISHED	1	6BD1	0.00
AP	O	u	2691883512	1129000329	32697	0	76.416353	11:05:13.252605	169	ESTABLISHED	1	96C4	0.00
AP	O	u	2691883512	1129000329	32697	0	76.427137	11:06:29.679742	180	ESTABLISHED	1	DA2D	0.00
AP	O	u	2691883512	1129000329	32697	0	76.497060	11:07:46.176802	212	ESTABLISHED	1	085B	0.00
AP	O	u	2691883512	1129000329	32697	0	76.567341	11:09:02.744143	225	ESTABLISHED	1	46BB	0.00
AP	O	u	2691883512	1129000329	32697	0	75.626664	11:10:18.370807	239	ESTABLISHED	1	6FC6	0.00
AP	O	u	2691883512	1129000329	32697	0	76.620353	11:11:34.991160	269	ESTABLISHED	1	9CD9	0.00

Estimating network latency from a packet trace

Using the RTT:

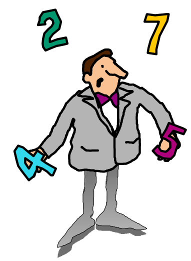
```
Connection:
 First timestamp:                   2004/03/08 18:28:34.828020
 Last timestamp:                    2004/03/08 18:28:53.995294
 Duration:                                     00:00:19.167274
 Average Round-Trip-Time:                                0.194 sec
 Final Round-Trip-Time:                                  1.545 sec
 Final state:                         TIME_WAIT (ACTIVE CLOSE)
 Out-of-order timestamps:                                    1
```

Using the delta time between packets:

TcpHdr	IO	F	Seq	Ack	RcvWnd	Data	Delta Time
S	I		2334250646	0	262144	0	0.000000
A S	O		3133347901	2334250647	262144	0	0.000113
A	I	u	2334250647	3133347902	262144	0	0.201817

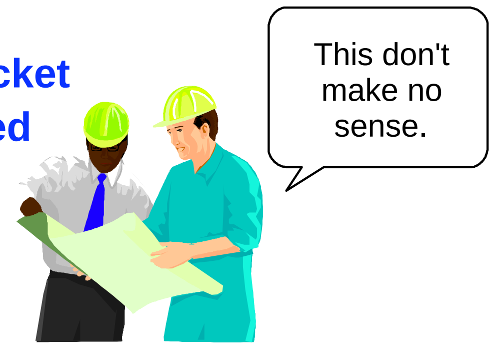
Using the data delta between the last acknowledged packet and the last data packet sent.

TcpHdr	IO	F	Seq	Ack	RcvWnd	Data	Delta	Time
A	O	.	3134061590	2334250647	262144	1448	0.000001	18:28:39.503446
A	O	.	3134063038	2334250647	262144	1448	0.000001	18:28:39.503447
A	O	.	3134064486	2334250647	262144	1448	0.000021	18:28:39.503468
AP	O	.	3134065934	2334250647	262144	1448	0.000001	18:28:39.503469
A	I	a	2334250647	3133974710	262144	0	0.000064	18:28:39.503533



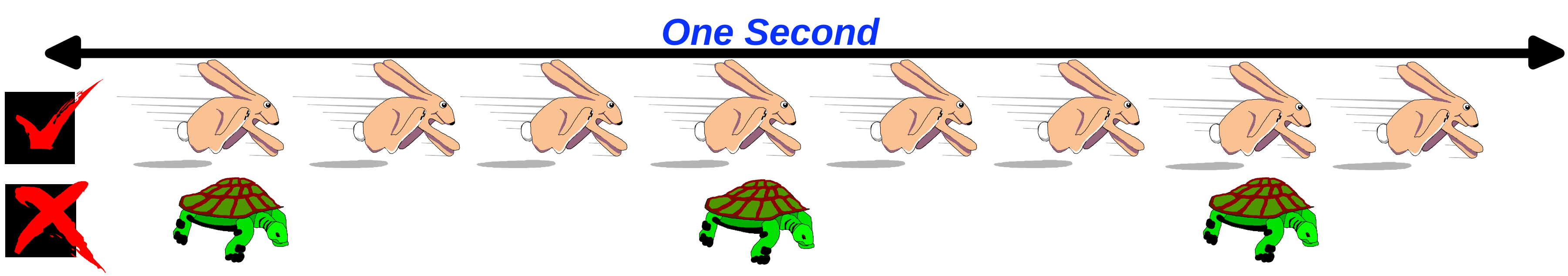
1. **Take the SEQ number of the last data packet sent.**
   3134065934
2. **Add in the data sent on that packet**
   3134065934 + 1448 = 3134067382
3. **Subtract the ACK number of the last acknowledgment received**
   3134067382 - 3133974710 = 92672
4. **Thus we know that 92672 bytes of data still remain in the pipe.**
   Another way to think of this is that there are 92672 bytes worth of packets that are not acknowledged. This is 64 packets that have been sent by one host but not acknowledged by the receiving host.

   This is slightly different than the time it takes a packet to travel from one host to the other, but can be used to gauge the relative amount of time it takes for a packet of data to be received at the remote host.



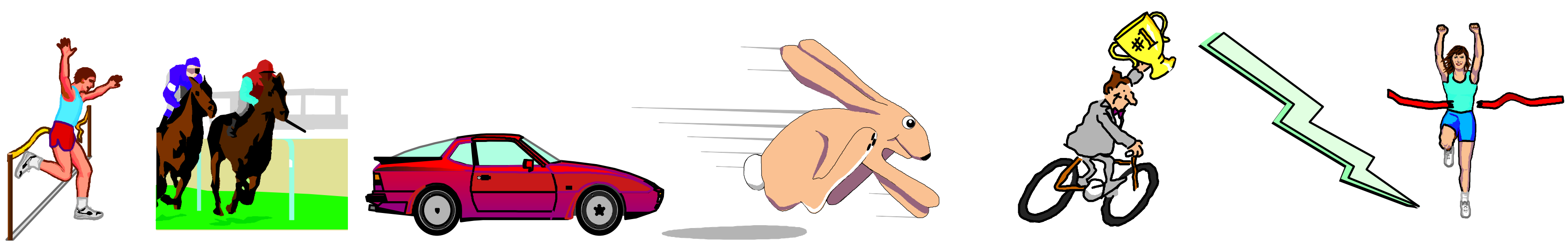
Estimating throughput from a packet trace

Data Quantity & Throughput:	Inbound,	Outbound
Application data bytes:	0,	9788417
Sequence number delta:	1,	9788418
Total bytes Sent:	0,	9788416
Bytes retransmitted:	0,	5792
Throughput:	0,	638.356 KBytes/s
Bandwidth utilization:	0.00%,	52.29%
Delay ACK Threshold:	200,	200 ms
Minimum Ack Time:	0.000002,	0.000091
Average Ack Time:	0.019908,	0.000091
Maximum Ack Time:	0.199157,	0.000091



What attributes do we find in a typical (well performing) z/OS trace?

- Use of the slow start algorithm after the three-way handshake
  - This allows the stack to gauge how fast to send data
- Data is acknowledged in a timely manner
- Little or no duplicate acknowledgments or retransmitted packets
- Constant flow of data
  - Any periods in which a host is not sending data is wasted time



Example of a typical trace

TcpHdr	IO	F	Seq	Ack	RcvWnd	Data	Delta Time	TimeStamp	RcdNr	State	Inf
S	I		2334250646	0	262144	0	0.000000	18:28:34.828020	32	SYN_RCVD	1
A S	O		3133347901	2334250647	262144	0	0.000113	18:28:34.828133	33	SYN_RCVD	1
A	I	u	2334250647	3133347902	262144	0	0.201817	18:28:35.029950	35	ESTABLISHED	1
A	O	.	3133347902	2334250647	262144	1448	[illegible]	[illegible]	[illegible]	[illegible]	1
AP	O	.	3133349350	2334250647	262144	1448	[illegible]	[illegible]	[illegible]	[illegible]	1
A	I	d	2334250647	3133350798	[illegible]	0	[illegible]	[illegible]	39	ESTABLISHED	1
A	O	.	3133350798	[illegible]	[illegible]	1448	0.000092	18:28:35.872116	40	ESTABLISHED	1
A	O	.	3133352246	[illegible]	[illegible]	1448	0.000001	18:28:35.872117	41	ESTABLISHED	1
AP	O	.	3133353694	[illegible]	262144	1448	0.000001	18:28:35.872118	42	ESTABLISHED	1
A	I	a	[illegible]	[illegible]	262144	0	0.197702	18:28:36.069820	44	ESTABLISHED	1
A	O	.	[illegible]	[illegible]	262144	1448	0.000097	18:28:36.069917	45	ESTABLISHED	1
A	O	.	[illegible]	[illegible]	262144	1448	0.000002	18:28:36.069919	46	ESTABLISHED	1
AP	O	.	[illegible]	2334250647	262144	1448	0.000001	18:28:36.069920	47	ESTABLISHED	1
A	I	a	[illegible]	3133358038	262144	0	0.197876	18:28:36.267796	48	ESTABLISHED	1
A	O	.	3133359486	2334250647	262144	1448	0.000105	18:28:36.267901	49	ESTABLISHED	1
A	O	.	3133360934	2334250647	262144	1448	0.000026	18:28:36.267927	50	ESTABLISHED	1
A	O	.	3133362382	2334250647	262144	1448	0.000001	[illegible]	51	ESTABLISHED	1
AP	O	.	3133363830	2334250647	262144	1448	0.000001	[illegible]	52	ESTABLISHED	1
A	I	a	2334250647	3133363830	262144	0	0.197944	[illegible]	53	ESTABLISHED	1
A	O	.	3133365278	2334250647	262144	1448	0.000160	[illegible]	54	ESTABLISHED	1
A	O	.	3133366726	2334250647	262144	1448	0.000002	[illegible]	55	ESTABLISHED	1
A	O	.	3133368174	2334250647	262144	1448	0.000001	18:28:36.466036	[illegible]	ESTABLISHED	1
A	O	.	3133369622	2334250647	262144	1448	0.000001	18:28:36.466037	[illegible]	[illegible]	1
AP	O	.	3133371070	[illegible]	[illegible]	1448	0.000001	18:28:36.466038	58	[illegible]	1
A	I	a	2334250647	[illegible]	[illegible]	0	0.197832	18:28:36.663870	59	[illegible]	1
A	O	.	3133372518	[illegible]	262144	1448	0.000106	18:28:36.663976	60	ESTABLISHED	1
A	O	.	[illegible]	[illegible]	262144	1448	0.000002	18:28:36.663978	61	ESTABLISHED	1
A	O	.	[illegible]	2334250647	262144	1448	0.000001	18:28:36.663979	62	ESTABLISHED	1
A	O	.	[illegible]	2334250647	262144	1448	0.000001	18:28:36.663980	63	ESTABLISHED	1
A	O	.	3133378310	2334250647	262144	1448	0.000001	18:28:36.663981	64	ESTABLISHED	1
AP	O	.	3133379758	2334250647	262144	1448	0.000001	18:28:36.663982	65	ESTABLISHED	1
A	I	d	2334250647	3133379758	262144	0	0.200345	18:28:36.864327	66	ESTABLISHED	1

Slow Start

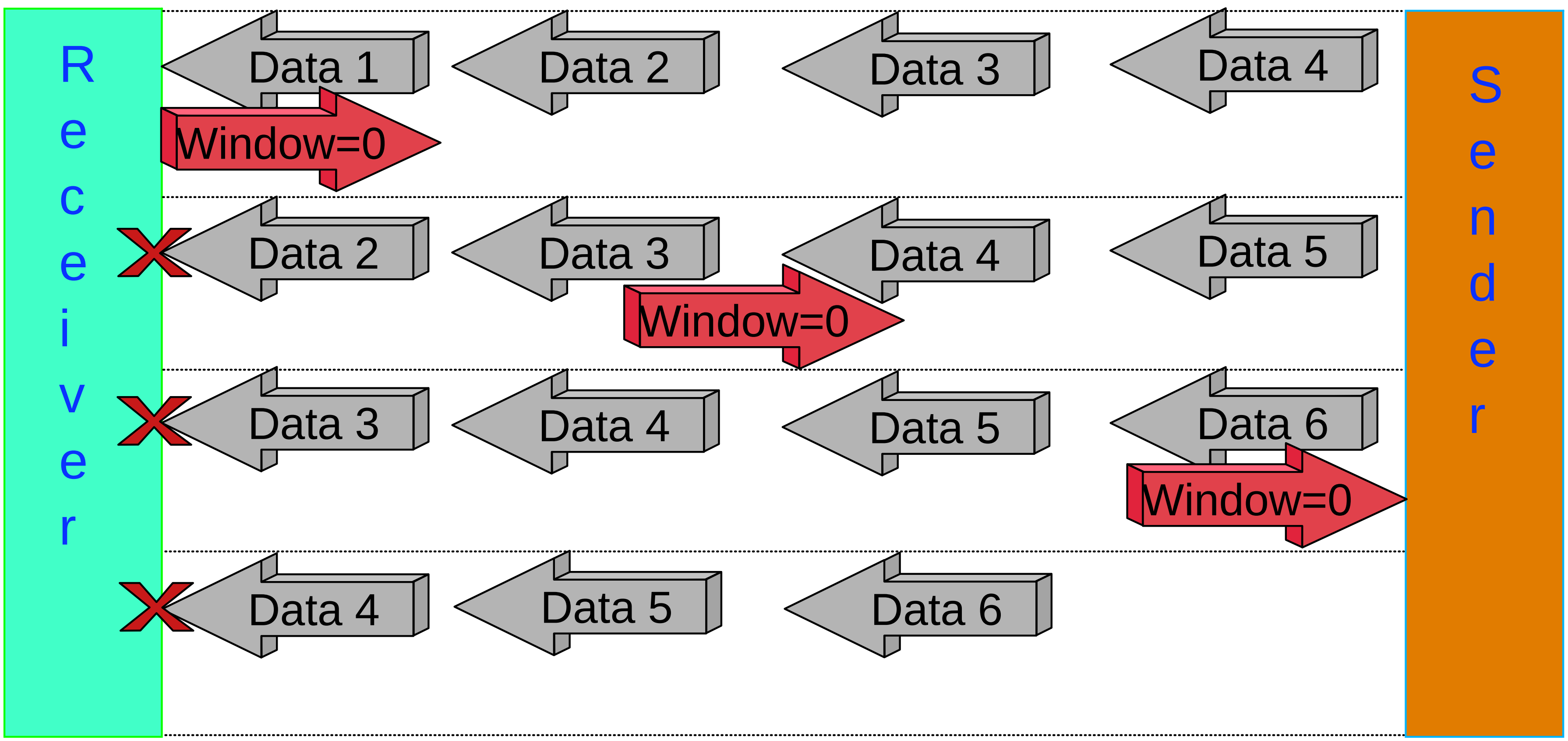
No duplicate acks or retransmissions

Decent response time

Constant data flow

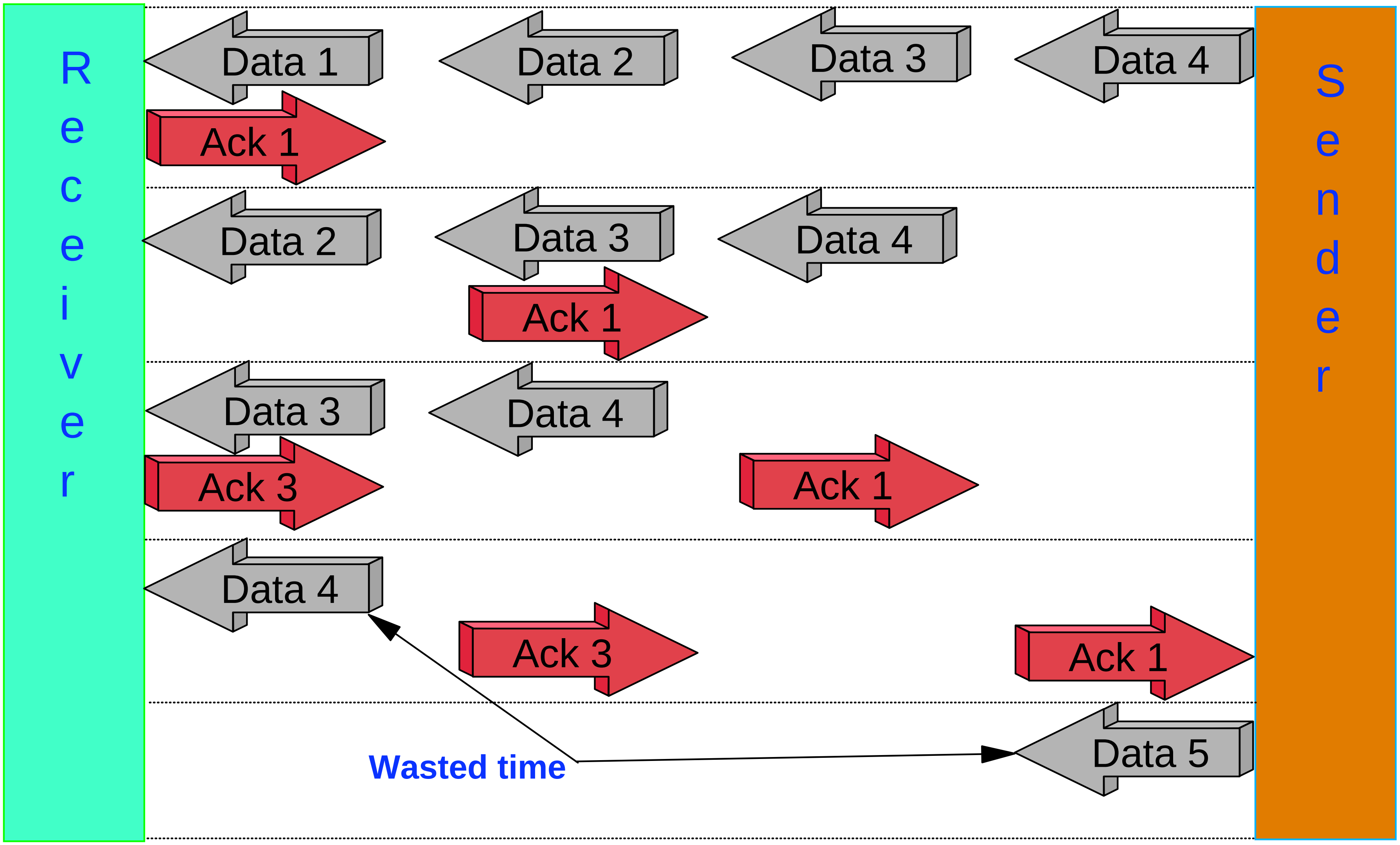
How does increased network latency effect a transfer?

- The longer the round trip time between two hosts, the longer it takes for adjustments to the transfer to occur
  - For example, window size updates may not occur in a timely manner (a similar situation can occur with retransmitted packets and duplicate ACKs)



By the time the window update arrives at the sender, 5 additional data packets have been sent

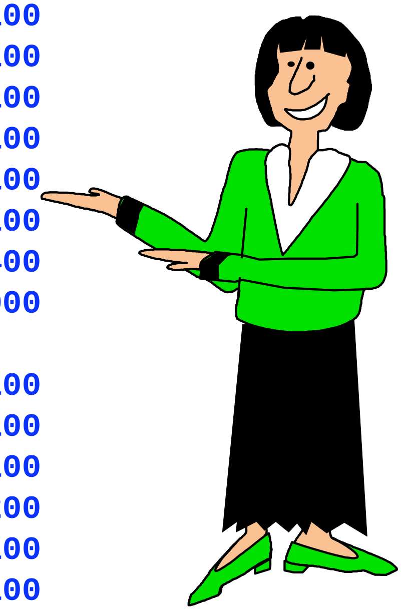
- Another example is filling up the advertised window size.
  - If the sending side fills up the window, it has to wait for an acknowledgment to arrive--opening the window back up--before additional data can be sent



Packet 4 fills the window (the unacknowledged data in the pipe). As such, no additional data can be sent until a previous packet is acknowledged.

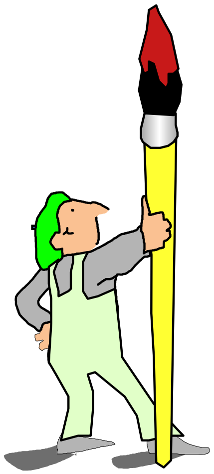
- Generally, high network latency by itself does not create a problem
  - It does, however, amplify other problems when they do occur
  - It can also prolong the amount of time required to recover from an error.
- Example of a transfer with a heightened network latency in which no error occurs:

	TcpHdr	IO	F	Seq	Ack	RcvWnd	Data	Delta Time
1<----	A	O	.	5462055798	2268250647	262144	1448	0.000200
2	A	O	.	5462057246	2268250647	262144	1448	0.000100
3	A	O	.	5462058694	2268250647	262144	1448	0.000100
4	A	O	.	5462060142	2268250647	262144	1448	0.000100
5	A	O	.	5462061590	2268250647	262144	1448	0.000100
6	A	O	.	5462063038	2268250647	262144	1448	0.000100
7	A	O	.	5462064486	2268250647	262144	1448	0.002100
8	AP	O	.	5462065934	2268250647	262144	1448	0.000100
9	A	I	a	2268250647	5461974710	262144	0	0.006400
10	A	O	.	5462067382	2268250647	262144	1448	0.010000
<...>								
75	A	O	.	5462151790	2268250647	262144	1448	0.000100
76	A	O	.	5462153238	2268250647	262144	1448	0.000100
77	A	O	.	5462154686	2268250647	262144	1448	0.000100
78	A	O	.	5462156134	2268250647	262144	1448	0.000200
79	A	O	.	5462157582	2268250647	262144	1448	0.000100
80	AP	O	.	5462159030	2268250647	262144	1448	0.000100
81--->	A	I	a	2268250647	5462057246	218400	0	0.006100



In this example, 80 packets (both inbound and outbound) are transferred before an acknowledgment arrives indicating that the first packet has been received.

- Example in which a window is filled and no additional data can be sent



	TcpHdr	IO	F	Seq	Ack	RcvWnd	Data	Delta Time
1	AP	O	.	3391804923	4198137376	65535	1380	0.000001
2	A	I	a	4198137376	3391800783	8280	0	0.974430
3	A	O	.	3391806303	4198137376	65535	1380	0.000074
4	AP	O	.	3391807683	4198137376	65535	1380	0.000002
5	A	I	a	4198137376	3391803543	8280	0	0.935102
6	A	O	.	3391809063	4198137376	65535	1380	0.000025
7	AP	O	.	3391810443	4198137376	65535	1380	0.000001
8	A	I	a	4198137376	3391806303	8280	0	0.973124
9	A	O	.	3391811823	4198137376	65535	1380	0.000031
10	AP	O	.	3391813203	4198137376	65535	1380	0.000001

Notice the ACK number on packet 2 (3391800783) and the RcvWnd size (8280)
--this tells the sender that more data can be sent as long as it does not exceed ACK number 3391800783 + 8280 = 3391809063

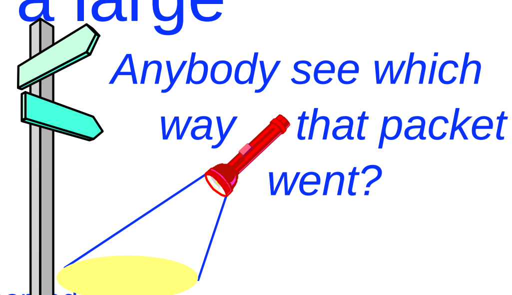
The sending side sends two more packets (packet 4 being the second) and stops

Note the SEQ number on packet 4 (3391807683) and the amount of data sent (1380)
-- 3391807683 + 1380 = 3391809063

Thus, we know that the sending side is filling the window and having to wait for it to be opened back up before sending more data.

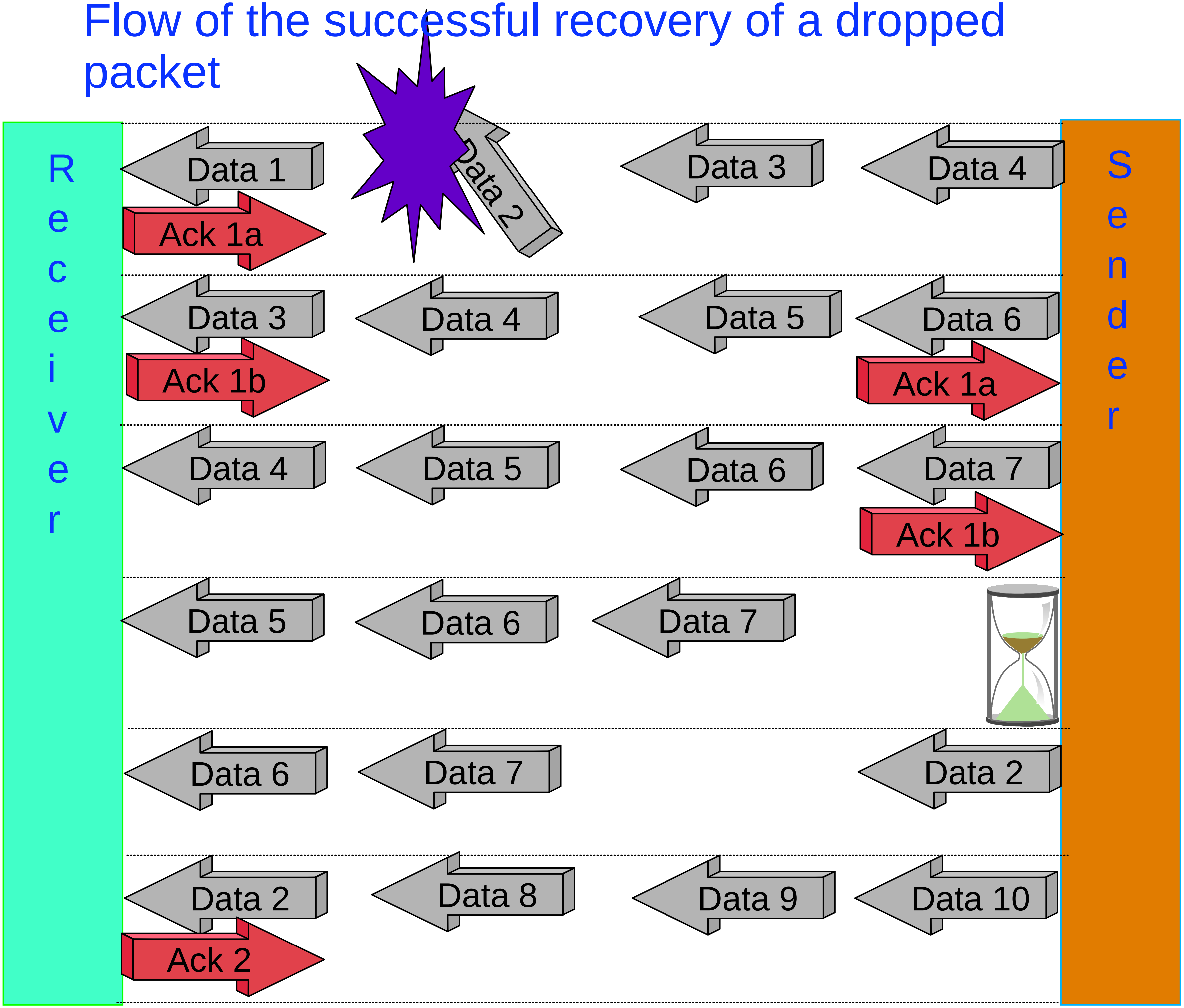
How do dropped or missing packets affect a data transfer?

- Dropped or missing packets are never a good sign as they may indicate other troubles within the network.
- However, in the context of a single data transfer, one or two dropped packets have little effect on the overall performance between the two hosts
  - ► TCP is architected ability to recover from lost data typically (depending on the network latency) allows the two TCPIP stacks to quickly recover from a lost packet
- Performance problems become noticeable when there occurs a consistent packet loss throughout the course of the transfer
  - ► All of the small recovery delays add up to a large cumulative delay

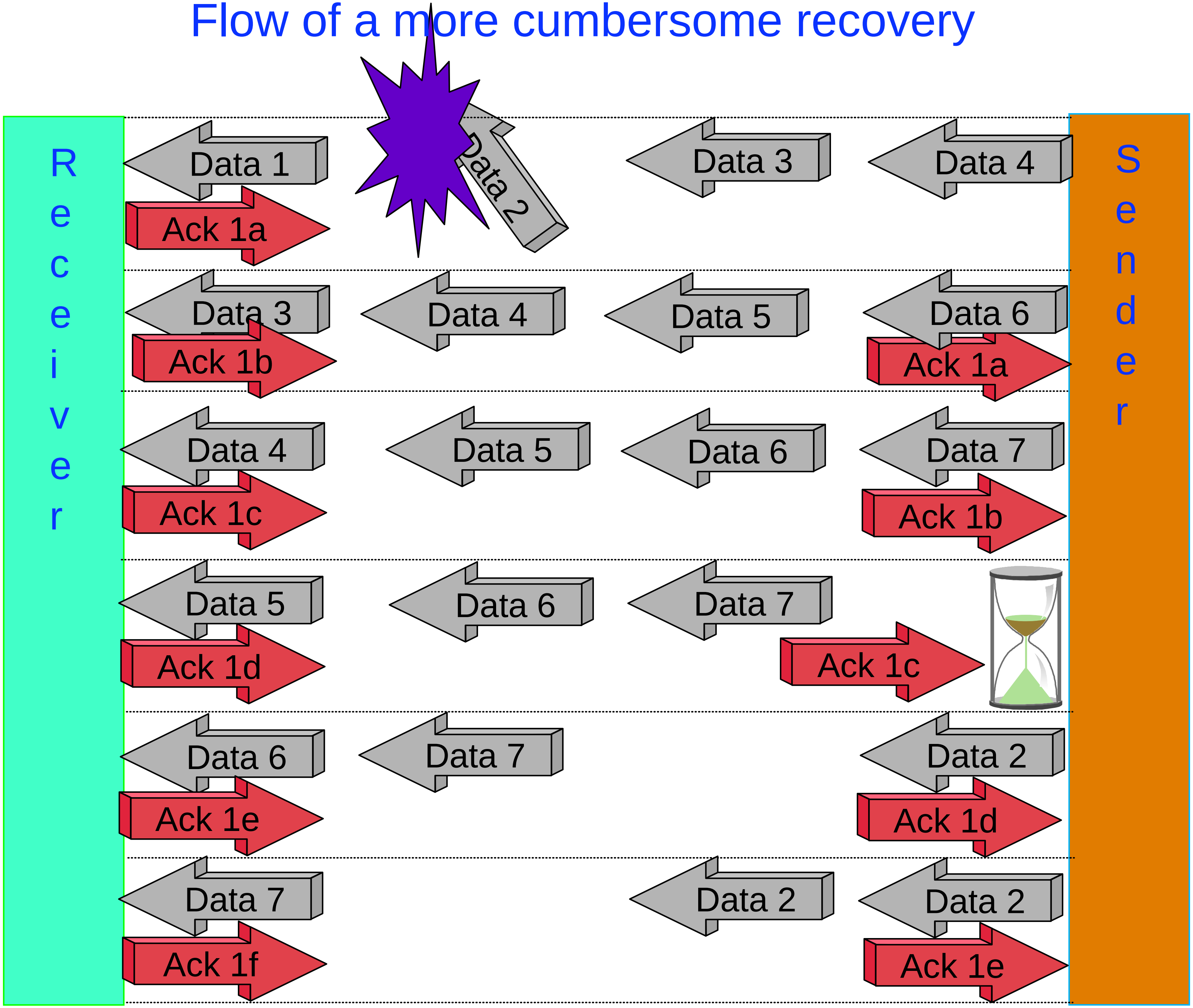


- Dropped packets result in the following delays:
  - The amount of time for a duplicate ACK (signifying a missing packet) to travel from the receiving stack to the sending stack
  - The amount of time the receiving stack may wait before retransmitting the missing packet (the stack will wait to retransmit in case the missing packet is still enroute)
  - The amount of time for the retransmitted packet to travel from the receiving stack to the sending stack.
    - After this point, the sending stack will typically pick up where it left off
    - However, if duplicate ACKs continue to arrive, the sending stack will begin retransmitting the packet in response
- All of this can be compounded by a high network latency

Flow of the successful recovery of a dropped packet



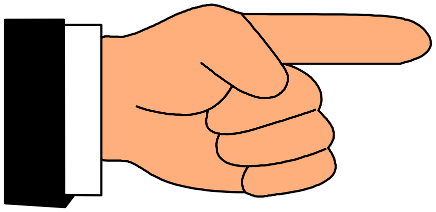
Flow of a more cumbersome recovery



Trace example of the previous flow

TcpHdr		IO	F	Seq	Ack	RcvWnd	Data	Delta Time	TimeStamp
1	A	I	a	1490971693	2739132871	32768	0	0.085756	10:24:58.135879
2	A	O	.	2739159151	1490971693	262144	1460	0.000054	10:24:58.135933
3	A	O	.	2739160611	1490971693	262144	1460	0.000005	10:24:58.135938
4	AP	O	.	2739162071	1490971693	262144	1460	0.000062	10:24:58.136000
5	A	I	u	1490971693	2739132871	32768	0	0.094051	10:24:58.230051
6	A	I	u	1490971693	2739132871	32768	0	0.205426	10:24:58.435477
7	A	I	u	1490971693	2739132871	32768	0	0.095687	10:24:58.531164
8	AP	O	r	2739132871	1490971693	262144	1460	0.000071	10:24:58.531235
9	A	I	u	1490971693	2739132871	32768	0	0.105971	10:24:58.637206
10	A	I	u	1490971693	2739132871	32768	0	0.100483	10:24:58.737689
11	A	I	u	1490971693	2739132871	32768	0	0.092407	10:24:58.830096
12	A	I	u	1490971693	2739132871	32768	0	0.115150	10:24:58.945246
13	A	I	u	1490971693	2739132871	32768	0	0.084884	10:24:59.030130
14	AP	O	r	2739132871	1490971693	262144	1460	0.085665	10:24:59.115795
15	A	I	u	1490971693	2739132871	32768	0	0.015070	10:24:59.130865
16	A	I	u	1490971693	2739132871	32768	0	0.099009	10:24:59.229874
17	A	I	u	1490971693	2739132871	32768	0	0.100359	10:24:59.330233
18	AP	O	r	2739132871	1490971693	262144	1460	0.000042	10:24:59.330275
19	A	I	u	1490971693	2739132871	32768	0	0.196344	10:24:59.526619
20	A	O	r	2739134331	1490971693	262144	1460	0.000090	10:24:59.526709
21	A	O	r	2739135791	1490971693	262144	1460	0.000004	10:24:59.526713

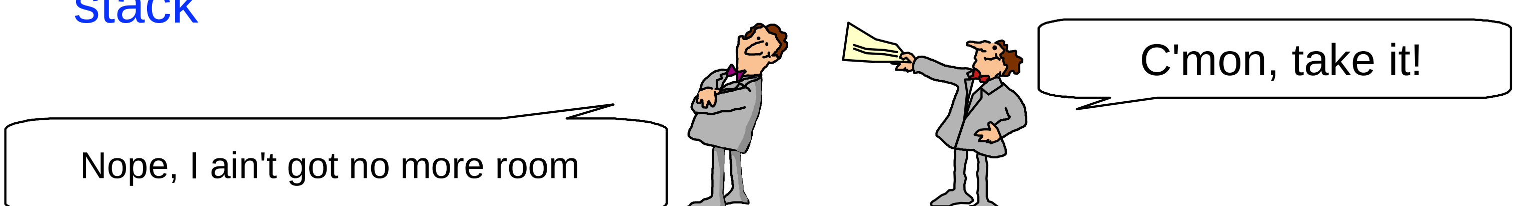
That's right....keep on going to the next page.



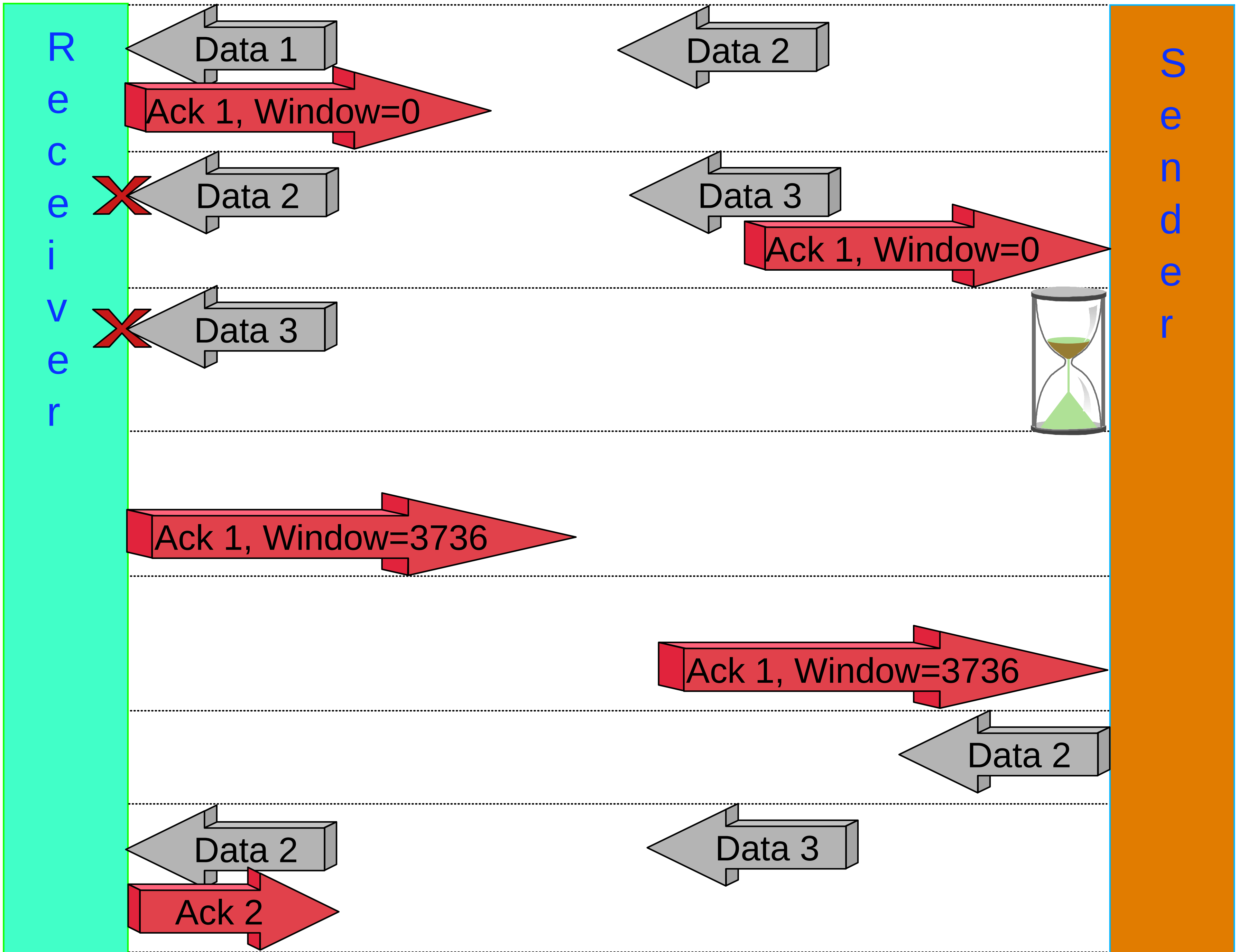
```
22 A       O r 2739137251 1490971693 262144  1460   0.000003 10:24:59.526716
23 AP      O r 2739138711 1490971693 262144  1460   0.000033 10:24:59.526749
24 A       I u 1490971693 2739132871  32768     0   0.099873 10:24:59.626622
25 A       I u 1490971693 2739132871  32768     0   0.099096 10:24:59.725718
26 A       I u 1490971693 2739132871  32768     0   0.100617 10:24:59.826335
27 A       O r 2739140171 1490971693 262144  1460   0.000099 10:24:59.826434
28 A       O r 2739141631 1490971693 262144  1460   0.000035 10:24:59.826469
29 AP      O r 2739143091 1490971693 262144  1460   0.000003 10:24:59.826472
30 A       I u 1490971693 2739132871  32768     0   0.094342 10:24:59.920814
31 A       O r 2739132871 1490971693 262144  1460   0.008104 10:24:59.928918
32 A       O r 2739134331 1490971693 262144  1460   0.000104 10:24:59.929022
33 A       O r 2739135791 1490971693 262144  1460   0.000003 10:24:59.929025
34 A       O r 2739137251 1490971693 262144  1460   0.000013 10:24:59.929038
35 A       O r 2739138711 1490971693 262144  1460   0.000003 10:24:59.929041
36 A       O r 2739140171 1490971693 262144  1460   0.000003 10:24:59.929044
37 A       O r 2739141631 1490971693 262144  1460   0.000013 10:24:59.929057
38 A       O r 2739143091 1490971693 262144  1460   0.000003 10:24:59.929060
39 AP      O r 2739144551 1490971693 262144  1460   0.000013 10:24:59.929073
40 A       I u 1490971693 2739132871  32768     0   0.091217 10:25:00.020290
41 A       I u 1490971693 2739132871  32768     0   0.239321 10:25:00.259611
42 A       I a 1490971693 2739150391  32768     0   0.062420 10:25:00.322031
```

What happens when the window size drops to zero?

- The window size is a TCP implementation that allows a receiving stack to communicate to the sending stack how much additional data it can receive
  - As the stack receives data, it writes it into a buffer (reducing the amount of space available in the buffer)
  - The application reads the data in from this buffer (increasing the amount of space available in the buffer)
  - The window size advertised by the receiving stack tells the sending stack how much space is available in the buffer
  - If the window size drops to zero, (meaning the application is no longer reading in data from the buffer) the sending stack is unable to send any additional data until the window size is increased.
- Things that would cause the window size to drop to zero:
  - The receiving application is hung
  - The receiving application isn't getting enough CPU cycles
  - The receiving stack has been overwhelmed with data from the sending stack



Flow of a window size going to zero



Trace example of a window size dropping to zero

TcpHdr		IO	F	Seq	Ack	RcvWnd	Data	Delta Time	TimeStamp
1	AP	O	.	3104443730	2351016260	65535	16	0.000001	15:19:23.997699
2	A	I	w	2351016260	3104408354	35392	0	0.000500	15:19:23.998199
3	A	I	a	2351016260	3104409426	34320	0	0.000003	15:19:23.998202
4	A	I	a	2351016260	3104410498	33248	0	0.000003	15:19:23.998205
<...>									
31	A	I	a	2351016260	3104439442	4304	0	0.000004	15:19:24.000324
32	A	I	a	2351016260	3104440514	3232	0	0.000003	15:19:24.000327
33	A	I	a	2351016260	3104441586	2160	0	0.000003	15:19:24.000330
34	A	I	a	2351016260	3104442658	1088	0	0.000003	15:19:24.000333
35	A	I	a	2351016260	3104443746	0	0	0.000227	15:19:24.000560
36	A	I	w	2351016260	3104443746	3736	0	0.593867	15:19:24.594427
37	A	O	.	3104443746	2351016260	65535	536	0.000082	15:19:24.594509

Interesting Notes:

In packet 1 the sending stack sends only 16 bytes of data. This is because it knows it has filled the window:

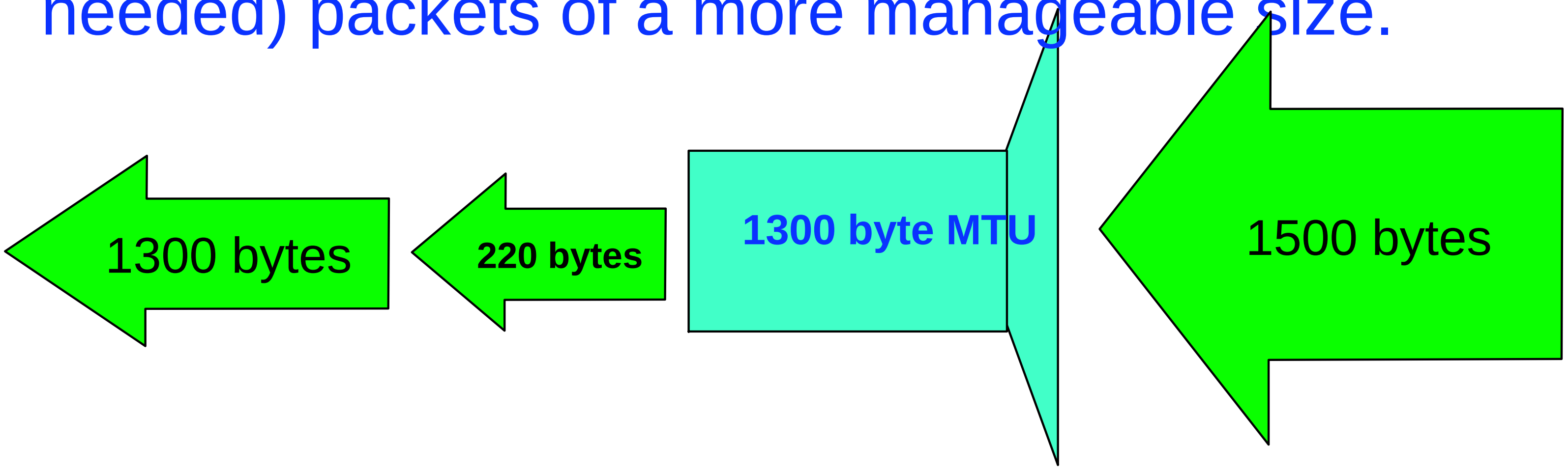
3104443730 + 16 - 35392 = 3104408354

The continued series of ACKs (and dropping window sizes) mirrors the arrival of the data packets at the receiving stack.

The zero window occurs in packet 35, and recovers half a second later after 3736 bytes are read in by the receiving application.

Finally, what about packet fragmentation?

- Packet fragmentation occurs when the sending stack uses an MTU greater than can be supported by a router on the path to the receiving stack.
  - Assuming the "Don't Fragment" bit is not turned on in the packet, the router will break the packet into two (or more if needed) packets of a more manageable size.



Note that the fragment has an extra 20 bytes. This accounts for the
additional IP header that must be added to the leftover data.

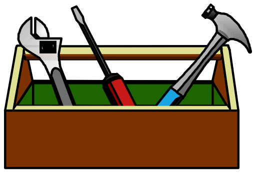
- Fragmentation slows down a data transfer on numerous levels:
  - There is overhead involved with the process of actually fragmenting the packet
    - Many routers put packets-to-be-fragmented on a separate queue, whereas smaller packets pass through more quickly. This can potentially lead to packets arriving out of order.
  - With two packets now traversing the network, there is a greater likelihood of packets getting dropped, and fragments can arrive out of order
    - Also note that most packets in a bulk data transfer are the same size. So if one packet is fragmented, then it's likely that most of the packets will be fragmented.
    - This doubles the number of packets involved in the transfer, adding to network congestion, and increasing network latency.
  - There's also additional overhead associated with reassembling the fragments once they reach the receiving stack.

So it's agreed...instead of buying bigger shoes, we'll chop our feet in half and buy more small pairs of shoes.

So now that we know what causes performance slow downs, what's the quickest way to diagnose a problem?

- The best method of diagnosing a data transfer performance issue is via a packet trace, using the session format.
- This format will provide you with:
  - Data transfer averages: RTT and throughput
  - Data transfer statistics: number of bytes retransmitted, duration of transfer, maximum amount of unacknowledged data, and window sizes
  - Packet counters: number of duplicate ACKs, retransmitted packets, out-of-order packets, fragments, and zero window sizes
- The session format also provides an easily readable output of the interaction between the two stacks
  - All inbound and outbound packets
  - Organized by connection
  - Output in the order in which the stack received them
  - Does packet analysis and highlights which packets are duplicate ACKs or retransmissions

The session formatter is a pretty handy dandy tool!



Sample session formatted output

```
 7726 packets summarized
  Local Ip Address:                                10.135.150.141
  Remote Ip Address:                               10.136.100.141
 Host:                                  Local,             Remote
  Client or Server:                    SERVER,             CLIENT
  Port:                                  1039,                 20
  Application:                               ,           ftp-data
  Link speed (parm):                       10,                 10 Megabits/s
 Connection:
  First timestamp:                 2004/03/08 18:28:34.828020
  Last timestamp:                  2004/03/08 18:28:53.995294
  Duration:                                   00:00:19.167274
  Average Round-Trip-Time:                              0.194 sec
  Final Round-Trip-Time:                                1.545 sec
  Final state:                      TIME_WAIT (ACTIVE CLOSE)
  Out-of-order timestamps:                                  1
 Data Quantity & Throughput:          Inbound,           Outbound
  Application data bytes:                   0,            9788417
  Sequence number delta:                    1,            9788418
  Total bytes Sent:                         0,            9788416
  Bytes retransmitted:                      0,               5792
  Throughput:                               0,            638.356
Kilobytes/s
  Bandwidth utilization:                0.00%,             52.29%
  Delay ACK Threshold:                    200,                200 ms
  Minimum Ack Time:                  0.000002,           0.000091
  Average Ack Time:                  0.019908,           0.000091
  Maximum Ack Time:                  0.199157,           0.000091
 Data Segment Stats:                  Inbound,           Outbound
  Number of data segments:                  0,               6792
  Maximum segment size:                  1460,               1460
  Largest segment size:                     0,               1448
  Average segment size:                     0,               1440
  Smallest segment size:                    0,                248
  Segments/window:                        0.0,               12.0
  Average bytes/window:                     0,              17294
  Most bytes/window:                        0,              81088
```

Length of transfer

# of bytes sent in the transfer

# of bytes retransmitted

Average throughput

Average time to ack data

```
Window Stats:                          Inbound,         Outbound
 Number of windows:                          0,              566
 Maximum window size:                   131072,           131072
 Largest window advertised:             262144,           262144
 Average window advertised:             240077,           262144
 Smallest window advertised:            141232,           262144
 Window scale factor:                        3,                3
 Window frequency:                           0,          37.2867 Windows/s
 Time Stamp updates:                       879,              450
 Total Round Trip Time:              19.096576,        19.167232 (99.63%), (  100%)
 Average Round Trip Time:             0.000000,         0.000000
 Maximum Data in Pipe:                       0,           156384
 Maximum retransmisson:                      0,                0
Number of:                             Inbound,         Outbound
 Packets:                                  929,             6797
  (x) Untraced Packets:                      0,                0
  (.) In-order data:                         0,             6782 ( 0.00%), (99.77%)
  (a) Acknowledgments:                     877,                1 (94.40%), ( 0.01%)
  (+) Data and ACK:                          0,                0 ( 0.00%), ( 0.00%)
  (u) Duplicate ACKs:                        1,                2 ( 0.10%), ( 0.02%)
  (w) Window size updates:                  46,                0 ( 4.95%), ( 0.00%)
  (z) Zero window sizes:                     0,                0 ( 0.00%), ( 0.00%)
  (p) Window probes:                         0,                0 ( 0.00%), ( 0.00%)
  (k) Keepalive segments:                    0,                0 ( 0.00%), ( 0.00%)
  (r) Retransmissions:                       0,                4 ( 0.00%), ( 0.05%)
  (o) Out-of-order:                          0,                6 ( 0.00%), ( 0.08%)
  (d) Delayed ACKs:                          3,                0 ( 0.32%), ( 0.00%)
  (f) Fragments:                             0,                0 ( 0.00%), ( 0.00%)
```

Window size data

The maximum amount of unacknowledged data at any given time

Packet counters and percentages

Use the symbols to locate and identify trouble spots within the session formatted trace itself.
Specifically:

- u - duplicate ACK
- z - zero window size
- r - retransmission
- f - fragment
- o - out of order packets

```
Time Spent on:                       Inbound,        Outbound
  (.) In-order data:          00:00:00.000000, 00:00:00.741193 ( 0.00%), ( 3.86%)
  (a) Acknowledgments:        00:00:17.459578, 00:00:00.000091 (91.09%), ( 0.00%)
  (+) Data and ACK:           00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (u) Duplicate ACKs:         00:00:00.201817, 00:00:00.000089 ( 1.05%), ( 0.00%)
  (w) Window size updates:    00:00:00.066940, 00:00:00.000000 ( 0.34%), ( 0.00%)
  (z) Zero window sizes:      00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (p) Window probes:          00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (k) Keepalive segments:     00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (r) Retransmissions:        00:00:00.000000, 00:00:00.000368 ( 0.00%), ( 0.00%)
  (o) Out-of-order:           00:00:00.000000, 00:00:00.000888 ( 0.00%), ( 0.00%)
  (d) Delayed ACKs:           00:00:00.608337, 00:00:00.000000 ( 3.17%), ( 0.00%)
  (f) Fragments:              00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
Number of:                           Inbound,        Outbound
  (   S ) SYN:                             1,               0 ( 0.10%), ( 0.00%)
  ( A S ) ACK SYN:                         0,               1 ( 0.00%), ( 0.01%)
  (    F) FIN:                             1,               1 ( 0.10%), ( 0.01%)
  ( A  F) ACK FIN:                         1,               1 ( 0.10%), ( 0.01%)
  (  R  ) RST:                             0,               0 ( 0.00%), ( 0.00%)
  (U    ) URG:                             0,               0 ( 0.00%), ( 0.00%)
Time Spent on:                       Inbound,        Outbound
  (   S ) SYN:                00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  ( A S ) ACK SYN:            00:00:00.000000, 00:00:00.000113 ( 0.00%), ( 0.00%)
  (    F) FIN:                00:00:00.079828, 00:00:00.008061 ( 0.41%), ( 0.04%)
  ( A  F) ACK FIN:            00:00:00.000177, 00:00:00.000091 ( 0.00%), ( 0.00%)
  (  R  ) RST:                00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
  (U    ) URG:                00:00:00.000000, 00:00:00.000000 ( 0.00%), ( 0.00%)
```

For More Information....

URL	Content
http://www.ibm.com/servers/eserver/zseries	IBM eServer zSeries Mainframe Servers
http://www.ibm.com/servers/eserver/zseries/networking	Networking: IBM zSeries Servers
http://www.ibm.com/servers/eserver/zseries/networking/technology.html	IBM Enterprise Servers: Networking Technologies
http://www.ibm.com/software/network/commserver	Communications Server product overview
http://www.ibm.com/software/network/commserver/zos/	z/OS Communications Server
http://www.ibm.com/software/network/commserver/z_lin/	Communications Server for Linux on zSeries
http://www.ibm.com/software/network/ccl	Communication Controller for Linux on zSeries
http://www.ibm.com/software/network/commserver/library	Communications Server products - white papers, product documentation, etc.
http://www.redbooks.ibm.com	ITSO redbooks
http://www.ibm.com/software/network/commserver/support	Communications Server technical Support
http://www.ibm.com/support/techdocs/	Technical support documentation (techdocs, flashes, presentations, white papers, etc.)
http://www.rfc-editor.org/rfcsearch.html	Request For Comments (RFC)