TCP/IP Window Sizes in Data Center Migrations

by Tony Hunt, Russ Conner, and Clea Zolotow

Datacenters use networks to move their data and to migrate their datacenters. These networks utilize a large pipe, such as Optical Carrier lines (like an OC-12, 622.08 Mbp/s, or OC-48, 2405.376 Mbit/s). Lately, that methodology has been extended to migrate terabytes of data over a long distance, i.e., a fast and wide WAN link.

These links can show poor performance when the extant operating system has a small TCP Window size. The TCP Window size value travels in the TCP header as below:

TCP Header

Bit offset	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0	Source port																Destination port															
32	Sequence number																															
64	Acknowledgment number																															
96	Data offset				Reserved				CWR	ECE	URG	ACK	PSH	RST	SYN	FIN	Window Size															
128	Checksum																Urgent pointer															
160 …	Options (if Data Offset > 5) …																															

Figure 1: TCP segment structure showing the Window Size

"TCP has a sliding window that limits the transmission speed in order to reduce congestion and data loss. This is a variable-duration window that allows the transmitting end-host to send a specified number of data units before an acknowledgement is received. The TCP window contains the amount of outstanding data a transmitting end-host can send on a particular connection before it gets acknowledgment back from the receiving end-host." http://www.nren.nasa.gov/tcpwindows.html

During initial line testing, we saw TCP window sizes from 48000 to 135000. With a typical latency of 26 to 29ms, this translated to transmission rates from 4.8MB to 1.3MB per second. Given that the OC48 is generally capable of about 2547MB per second, this performance is unacceptable.

Before attempting fast and wide WAN migrations, the OS must be tuned properly. First, test with IPerf to ensure the TCP Window is set properly. Sun Solaris and Linux ship with default TCP Window sizes of 64K, limiting bandwidth to 1.3MB in many cases. Tuning can be done with the help of a document from the Pittsburg Supercomputing Center at http://www.psc.edu/networking/projects/hpn-ssh/ and must be applied to specific versions of SSH. This patch makes the TCP window size dynamic from the start based on the OS settings and not the default code's limit of 131K, or the size of a C++ u_int variable. RedHat Linux had an available RPM that was pre-patched and tested for a very recent 5.3 OpenSSH release. This OS was selected to test the fix.

The patched version yielded 26 to 30MB per second, nearly the same performance as on the local Network. Multiple RSYNC threads could be supported yielding 120MB in some cases. Limitations were now set by the capacity of the CPU to encrypt and un-encrypt data.

References:

http://www.speedguide.net/bdp.php