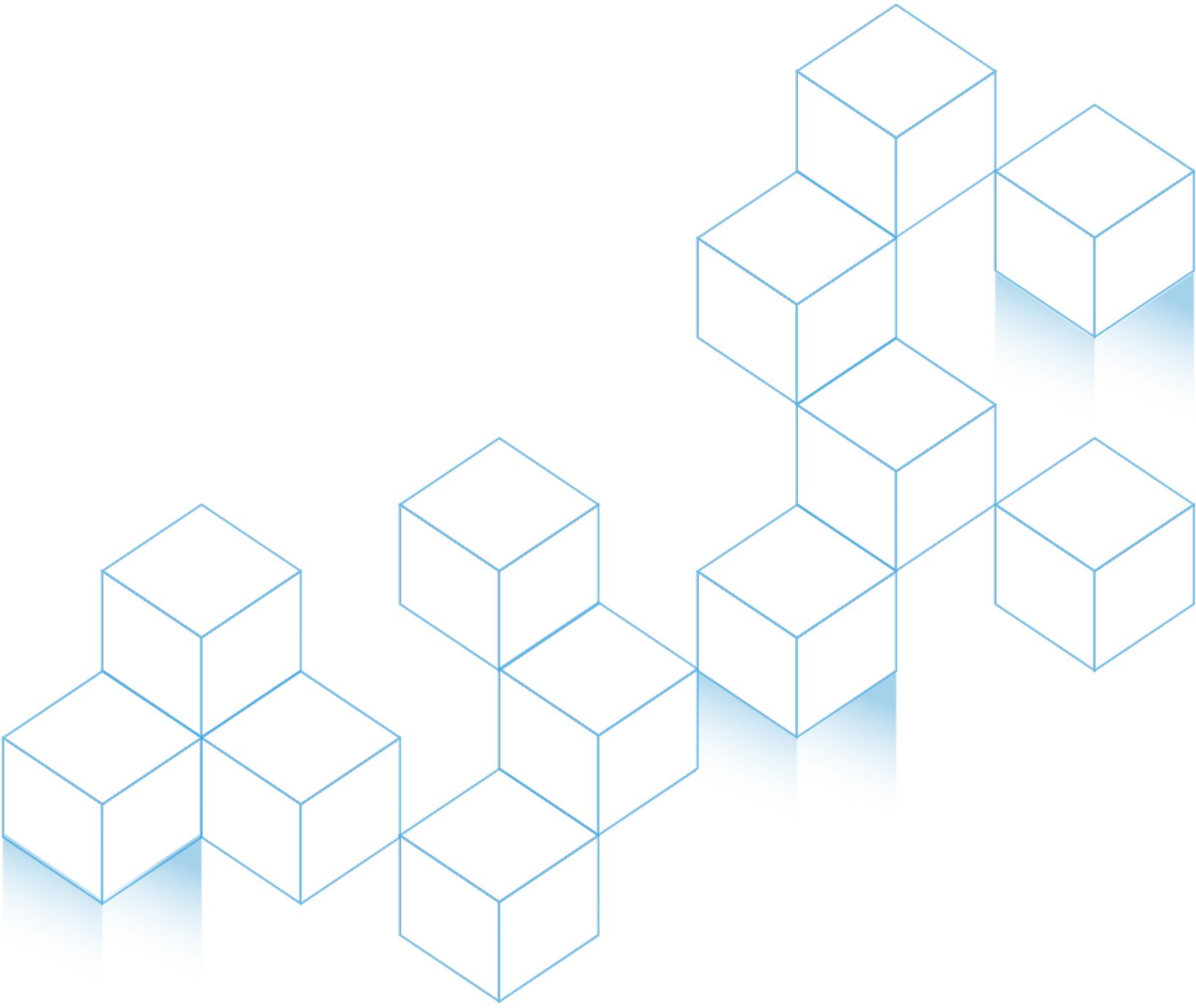
TCP/IP Performance Tuning for Skytap on Azure IBM Power LPARs



May 2023

Introduction

This article discusses common best practices for tuning IBM Power IBM i and AIX workloads for Skytap on Azure (SOA). One of the biggest challenges organizations have when migrating and running Power workloads in SOA is tuning the TCP/IP stack for both the local (within a SOA) environment and externally either between on-premises and SOA, or between SOA deployments in different SOA regions. This article makes several references to general Azure tuning techniques as well as those from IBM for AIX and IBM i.

Executive Summary

An efficient network is key to rapid cloud migration and ongoing application performance for any cloud deployment. There are several factors that impact network performance, two of those being latency and fragmentation.

To establish a baseline for network performance based on several tuning parameters for both AIX and IBM i, Skytap tested the following scenarios:

- Two LPARs on the same Power Hosting Node (PHN)
- Two LPARs within the same Skytap environment and subnet
- Two LPARs, one in SOA's Singapore region to a second in SOA's Hong Kong region using VPN to connect both regions
- Two LPARs, one in SOA Singapore region to a second in SOA's Hong Kong region using Azure ExpressRoute provisioned at 1Gb/sec

Each of these scenarios were done for AIX and IBM i. For AIX, Skytap tested using SCP as a file transfer protocol in addition to IPERF3. For IBM i, Skytap used FTP only. Skytap testing found:

- LPARs of both Operating Systems performed better with TCP Send Offload enabled.
- Network performance between the two systems was best when source and target machines matched for both MTU and Send/Receive buffer settings.
- A Maximum Transmission Unit (MTU) of over 1500 for WAN connections was problematic. Skytap recommends using an MTU between 1250 and 1330 for ExpressRoute and VPN connections.
- When using a VPN, set an MSS Clamp no lower than 1200 and no higher than 1254.
- When transferring large amounts of data, look at using applications that can support multiple file transfer connections or transfer multiple files at once. This allows the client to maximize available bandwidth of a connection.

TABLE 1: AIX NETWORK PERFORMANCE RECOMMENDATIONS FOR LOCAL TRAFFIC AND BETWEEN SOA SINGAPORE AND HONG KONG REGIONS

Connection Type	Large Send Offload	MTU	RMTU	TCP Send / Receive Buffer
Host	Enabled	1500	576	2097152
LAN	Enabled	1500	576	2097152
VPN	Enabled	1500	576	2097152
ExpressRoute	Disabled	1500	576	2097152

TABLE 2: IBM I NETWORK PERFORMANCE RECOMMENDATIONS FOR LOCAL TRAFFIC AND BETWEEN SOA SINGAPORE AND HONG KONG REGIONS

Connection Type	Large Send Offload	MTU	RMTU	TCP Send / Receive Buffer
Host	Enabled	1330	1330	4194304
LAN	Enabled	576	576	65535
VPN	Enabled	1200	1200	4194304
ExpressRoute	Enabled	1200	1200	4194304

These settings were shown to have the best performance between LPARs of like Operating System in these scenarios but should only be considered a baseline. Performance will vary based on the configuration and settings within your Skytap environment.

The Role of MTU, Fragmentation, and Large Send Offload

The MTU is the largest size frame (packet) that can be sent over a network interface that is measured in bytes. The general default setting for most devices is 1500. However, for Power workloads this is commonly not the case as discussed later in this article.

Fragmentation

Fragmentation occurs when a packet is larger than the device, LPAR, and/or virtual machine (VM) is set to handle. When a larger packet is received, it is either fragmented into one or more packets and resent, or dropped altogether. If fragmented, the device on the receiving end then must reassemble the original packet so it can be processed by the end system.

Fragmentation generally has a negative impact on performance. When a packet must be fragmented, it requires more CPU on the device or receiving system. The device must hold all the fragmented packets, reassemble them, and then send it to the next device. If it has an MTU size that doesn't match, this process repeats until the traffic finally reaches the target system. The impact is increased latency due to devices having to reassemble the packets and the potential for packets to be received out of order by the device. In some cases, they can be dropped causing the packet to have to be resent.

The Pros and Cons of Modifying the MTU

Generally, the larger the MTU, the more efficient your network and the faster you can move data between two systems. This is the first consideration as to network tuning with Power workloads as they often don't default to the standard 1500 MTU. For AIX, it is common to see Jumbo Frames enabled which has an MTU of 9000. For IBM i, the default is at the opposite end of the spectrum at 576, which dates back to being optimized for dial up modems.

If the packet is set larger than devices in the path can handle, the packets will be fragmented, new header information will be added, and the packet is sent onto its destination. This all adds overhead. For low latency connections, it will prevent you from consuming maximum bandwidth. For WAN connections such as ExpressRoute and VPN over greater distances, it can have a much greater impact. This is especially apparent during the migration phase where you are working to move a lot of data in a short amount of time.

Conversely, if you set the MTU size too small, then the destination system must process more packets and that too hinders performance. The key is to find the right MTU that is optimized for the speed and latency of the connection you are using.

Skytap and LPAR MTU

Skytap Power Hosting Nodes (PHNs) use VIOS. VIOS is the virtualization layer provided by PowerVM which controls all virtual input and output from a given LPAR. It virtualizes the physical hardware. This exists both for network traffic to local and WAN connections, in addition to traffic to Skytap's Software Defined Storage layer. Traffic that leaves the LPAR will be governed by VIOS and the Shared Ethernet Adapter (SEA) of the VIOS LPAR. Skytap sets

the MTU of the SEA to 1500. As a result, setting your LPAR's MTU higher than 1500 will result in fragmentation. However, this changes if you have multiple LPARs on the same network that are on the same subnet and PHN. In this situation, traffic never leaves the PHN and VIOS essentially "steps out of the way" allowing the LPAR virtual adapters to communicate directly with one another. The result is very fast network throughput and the ability to use Jumbo Frames.

Azure and Fragmentation

Azure's Virtual Network stack will drop out of order fragments or fragmented packets that do not arrive in the order that they were supposed to. When this happens, the sender must resend those packets. Microsoft does this to protect itself against a vulnerability called FragmentSmack that was announced in November of 2018. This can have a substantial impact when you have high latency connections or connections where there is a lot of fragmentations because of an MTU that is set too high.

Skytap and Azure ExpressRoute

Skytap and Azure ExpressRoute use an MTU size of 1438. Sending traffic between two SOA regions will result in an MSS of 1348. This includes 20 bytes of IP header and 20 bytes of TCP header, then Azure modifies this which adds another 50 bytes.

SOA and VPN

SOA's virtual networking stack includes native support for IPSec. This allows users to create VPN tunnels on the SOA side without having to add IPSec Gateways and/or virtual appliances. Like ExpressRoute, the MTU of IPSec connections should not be set higher than 1438 and an MSS of 1348.

Setting the correct MSS is critical in reducing fragmentation. To illustrate the point, here is an example of an IPERF test where Skytap set the MTU within the MSS of 1254 vs. setting it one byte above the MSS as 1255, resulting in fragmentation.

```
Ubuntu 18 <-> Skytap VPN <-> Ubuntu 18 | IPSec MSS 1255 | No guest MTU changes
$ iperf3 –client 10.0.0.1 –time 15
45 Mbps
62 Mbps
50 Mbps

Ubuntu 18 <-> Skytap VPN <-> Ubuntu 18 | IPSec MSS 1254 | No guest MTU changes
$ iperf3 –client 10.0.0.1 –time 15
500 Mbps
545 Mbps
388 Mbps
```

The above example was taken between the Hong Kong and Singapore regions. The differences are dramatic - approximately a 10x delta. The lesson learned is to avoid fragmentation at all costs as file transfer speed is greatly reduced as the latency between source and destination increases.

Large Send Offload

Large Send Offload (LSO) allows the LPAR in SOA to send large packets and frames through the network. When enabled, the Operating System creates a large TCP packet and sends it to the Ethernet Adapter for segmentation before sending it through the network. For LPARs that are on the same PHN with LSO enabled, the result is extremely high network throughput. SOA's VIOS SEA has LSO enabled, so that using this setting can improve communication between LPARs on the same subnet within a Skytap Environment. However, when the traffic is destined to another LPAR in the network, it must go through the SEA with an MTU of 1500, which will result in fragmentation.

For AIX with LSO enabled, you will see maximum throughput between two AIX LPARs on the same subnet at ~2Gb/sec due to the MTU size of 1500. With two LPARs on the same PHN, VIOS and the physical network layer is bypassed allowing far greater speeds at a maximum of about 2.6Gb/sec.[1]

For IBM i with LSO enabled, performance is going to be very protocol dependent. Using FTP as a test, maximum throughput between two IBM i systems is ~472Mb/sec with an MTU of 576

[1] Tests were done using IPERF3 set to use up to eight connections.

with two IBM i systems on the same subnet running on different PHNs. When running on the same PHN, performance increases slightly at ~482Mb/sec with an MTU size of 1330. It is important to note that you can expect higher throughput with multiple connections. Therefore, when transferring a lot of data from an IBM i system, you should use multiple FTP streams. For example, if you have three 5GB files to transfer, utilize three separate interactive sessions to increase your overall throughput.

VPN and MTU

If you are using a VPN between Skytap Environments in different SOA regions or between SOA and an external environment, you need to account for the increased overhead for the VPN encapsulation.

For AIX, testing was done between two SOA regions with ~34ms of latency using SCP between both systems. Skytap recommends setting the MTU to 1200 with an MSS of 1250 with LSO enabled. In this configuration, Skytap has seen transfer rates of ~366Mb/sec. When the physical network layer is bypassed, Skytap has seen far greater transfer speeds at a maximum of about 2.6Gb/sec.

For IBM i, testing was done between two SOA regions with ~34ms of latency using FTP between both systems. Skytap recommends setting the MTU to 1200 with an MSS of 1250 with LSO enabled. In this configuration, Skytap has seen transfer rates of ~190Mb/sec.

Azure ExpressRoute and MTU

If you are using an ExpressRoute between Skytap Environments in different SOA regions or between SOA and an external environment, you need to account for two important factors:

1. Azure's virtual networking stack will drop out of order packets. This often increases as latency increases. As a result, fragmentation needs to be managed carefully for optimal network performance.
2. ExpressRoute adds some IP overhead when used with SOA. The amount added is 50 bytes per frame/segment.

For AIX, testing was done between two SOA regions with ~34ms of latency using SCP between both systems. Skytap recommends setting the MTU to 1200 with an MSS of 1250 with LSO enabled. In this configuration, Skytap has seen transfer rates of ~366Mb/sec, and when the physical network layer is bypassed, it allows far greater speeds at a maximum of about 2.6Gb/sec.

For IBM i, testing was done between two SOA regions with ~34ms of latency using FTP between both systems. Skytap recommends setting the MTU to 1200 with an MSS of 1250 with LSO enabled. In this configuration, Skytap has seen transfer rates of ~190Mb/sec per any given connection.

Latency and Round-Trip Time

Network latency is governed by the speed of light. As a result, the maximum network throughput for TCP/IP is going to be dependent on the round-trip time (RTT) between source and target devices. This table shows the distance between SOA regions.

TABLE 3: EXAMPLE LATENCY AND ROUTE-TRIP TIME

Route	One-way time	RTT
East Asia (Hong Kong) to Southeast Asia (Singapore)	38 ms	76 ms
North Europe (Ireland) to West Europe (Amsterdam)	21 ms	42 ms
East US (Virginia) to South Central Texas)	37 ms	54 ms

Power OS Tunning Summary

Every deployment of AIX and IBM i that is on-premises and/or in SOA will have different variables from what was tested and described in this article. Skytap's tests were in a relatively controlled environment between two SOA regions with a known network path and devices. When communicating with systems in SOA regions to those located on-premises and other cloud providers, different devices and their configurations will have a significant impact on network performance. This needs to be accounted for early in your network design and solutions. It is always best to tune and optimize from a position of knowing vs. not knowing. As a result, Skytap has provided a summary of the testing done and the tuning parameters that were used for each. A list of the parameters is provided below with a link to the appropriate documentation for you to learn more, and/or make the change to your system.

AIX Network Tuning[2]

For AIX network tuning, IBM's guide for AIX Network Adapter Performance is a great reference. Key tuning parameters that were used to tune Skytap performance include:

[2] One of the best "unofficial" resources on AIX Tuning, Network Tuning in AIX by Jaqui Lynch it is a must read for anyone working on AIX Networking

- **LSO** AIX/IBM i: The TCP LSO option allows the AIX TCP layer to build a TCP message up to 64 KB long. The adapter sends the message in one call down the stack through IP and the Ethernet device driver. LSO is enabled by default on SOA's PHN SEAs (Shared Ethernet Adapter of VIOS).
- **Maximum Transmission Unit** (MTU) AIX/IBM i: The maximum network packet size that can be transmitted or received by a network device.
- **Remote Maximum Transmission Unit** (RMTU): The MTU size that is used for remote networks. AIX attempts to calculate which networks are remote by looking at the route table. IBM i allows you to set the MTU for each route.
- **Default TCP/IP Send Receive Buffer Size** AIX/IBM i: The tcp_recvspace option specifies how many bytes of data the receiving system can buffer in the kernel for receiving traffic, likewise the tcp_sendspace tunable specifies how much data the sending application can buffer when sending data. For these tests, Skytap found that the default AIX Send/Receive Buffer size of 2097152 bytes yielded the best performance. For IBM i, there was some variation of values that performed well.
- **Maximum Segment Size** (MSS): The largest amount of data, measured in bytes, that a computer or communications device can handle in a single, unfragmented piece. For VPN testing between regions, an MSS of 1200 was used on both sides of the VPN tunnel.
- **Latency**: Inside of a Skytap Environment yields near zero network latency. All remote tests were done between Hong Kong (CN-HongKong-M-1) and Singapore (SG-Singapore-M-1). Latency between these regions was between 33ms and 38ms.

TABLE 4: AIX NETWORK PERFORMANCE WITH LSO ENABLED USING SCP

Test Type	MTU Size	RMTU Size	Result (MB/sec)	Result (Mb/sec)2	Latency (ms)
Same Host	1500	576	130	1040	0
Same Skytap Subnet	1500	576	130	1040	0
Internet	1500	576	49.8	398	33
ExpressRoute	1500	576	53.3	426	33
VPN	1200	1200	45.8	366	34

TABLE 5: AIX NETWORK PERFORMANCE WITH LSO DISABLED USING SCP

Test Type	MTU Size	RMTU Size	Result (MB/sec)	Result (Mb/sec)2	Latency (ms)
Same Host	1500	576	89.2	713	0
Same Skytap Subnet	1500	576	91.4	731	0
Internet	1500	576	48.3	452	33
ExpressRoute	1500	576	55	440	33
VPN	1200	576	35.1	280	34

TABLE 6: AIX NETWORK PERFORMANCE WITH LSO ENABLED USING IPERF8 WITH EIGHT CONNECTIONS

Test Type	MTU Size	RMTU Size	Connections	Result (MB/sec)	Result (Mb/sec)2	Latency (ms)
Same Host	1500	576	8	3312.5	26500	0
Same Skytap Subnet	1500	576	8	2412.5	19300	0
Internet	1500	576	8 (512K packet)	88	705	33
ExpressRoute	1500	576	8	35.6	285	33
VPN	1200	1200	8 (512K packet)	35.6	285	34

TABLE 7: AIX NETWORK PERFORMANCE WITH LSO DISABLED USING IPERF8 WITH EIGHT CONNECTIONS

Test Type	MTU Size	RMTU Size	Connections	Result (MB/sec)	Result (Mb/sec)	Latency (ms)
Same Host	1500	576	8	356	2855	0
Same Skytap Subnet	1500	576	8	351	2810	0
Internet	1500	576	8	56.5	452	34
ExpressRoute	1500	576	8	4.6	37	33
VPN	1200	576	8 (512K packet)	5	40.2	34

TABLE 8: IBM I NETWORK PERFORMANCE WITH LSO ENABLED USING FTP

Test Type	MTU Size	RMTU Size	Send/Receive Buffer	Result (MB/sec)	Result (Mb/sec)	Latency (ms)
Same Host	1330	1330	4194304	60.26	482.15	0
Same Skytap Subnet	576	576	65535	59.05	472.41	0
VPN	1200	1200	4194304	23.85	190.8	36
ExpressRoute	1200	1200	4194304	22.35	178.85	36

TABLE 9: IBM I NETWORK PERFORMANCE WITH LSO DISABLED USING FTP

Test Type	MTU Size	RMTU Size	Send/Receive Buffer	Result (MB/sec)	Result (Mb/sec)	Latency (ms)
Same Host	1200	1200	4194304	48.94	391.53	0
Same Skytap Subnet	1440	1440	4194304	58.75	470.04	0
VPN	576	576	4194304	1.58	12.64	38
ExpressRoute	1330	1330	4194304	15.23	12.18	34

Summary

To maximize the network performance between systems in SOA, one must consider latency and fragmentation and work to reduce both variables. This is not only true when looking at source and destination systems within SOA regions, but also when moving data in and out of SOA externally. In Skytap's testing, it found these points critical to ensuring maximum network performance:

- TCP Send Offload should be enabled for both AIX and IBM i workloads.
- Make sure that source and target machines match both MTU and Send/Receive buffer settings.
- Do not use an MTU larger than 1500. Skytap recommends using an MTU between 1250 and 1330 for ExpressRoute and VPN connections.
- When using a VPN, look at setting an MSS Clamp no lower than 1200 and no higher than 1254.
- When transferring large amounts of data, look at using applications that can support multiple file transfer connections or transfer multiple files at once to maximize available bandwidth.

It is important to remember each system and network can be different, so this needs to be accounted for in your tuning efforts.