PowerVM Network Performance Update

IBM Power System User Group: Technical Update 2022

15. – 16. November 2022, Zürich

Alexander Paul
Power Systems Virtualization and Performance
paul.alex@ch.ibm.com

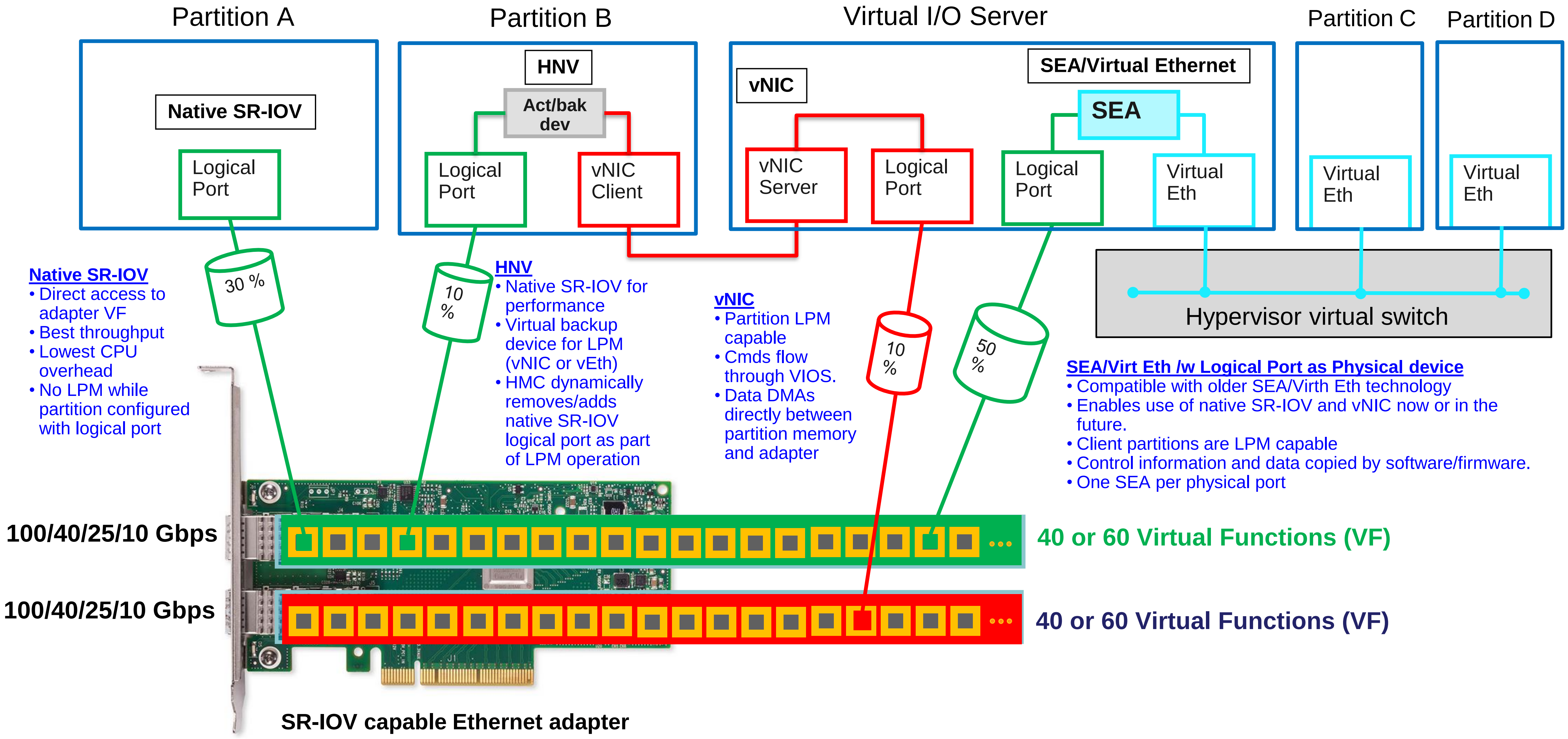
Agenda

- PowerVM network options - quick refresher
- E1080 network performance update and experiences
- Software virtualization vs. hardware assisted virtualization
- SR-IOV and vNIC performance update
- Power10 FW1030 sneak preview
- Case studies
- Linux performance best practice

PowerVM Network Options

- **PowerVM Virtual Ethernet and Shared Ethernet Adapter**
  - Well known and established technology (since POWER5)
  - Build-in failover capabilities (SEA failover on Virtual I/O Server)
  - Supports Live Partition Mobility and Simplified Remote Restart
- **PowerVM SR-IOV**
  - Hardware assisted network virtualization
  - Excellent performance
  - Client OS needs to care about redundancy as logical adapter ports are directly hardware related.
  - Does not allow the use of features like Live Partition Mobility (LPM).
- **PowerVM vNIC**
  - Based on PowerVM SR-IOV technology but using indirect assignment of adapter resources (via VIOS).
  - Allows that one vNIC can map to multiple SR-IOV backing devices.
  - Provides hypervisor-based failover and support for LPM and SRR.
- **Hybrid Network Virtualization**
  - Combination of native SR-IOV and either Virtual Ethernet or vNIC.
  - Performance characteristics during normal operations similar to native SR-IOV.
  - Build-in failover capabilities within the OS and LPM support, due to HMC controlled automation.

SR-IOV/vNIC/HNV Architectures



Native SR-IOV
- Direct access to adapter VF
- Best throughput
- Lowest CPU overhead
- No LPM while partition configured with logical port

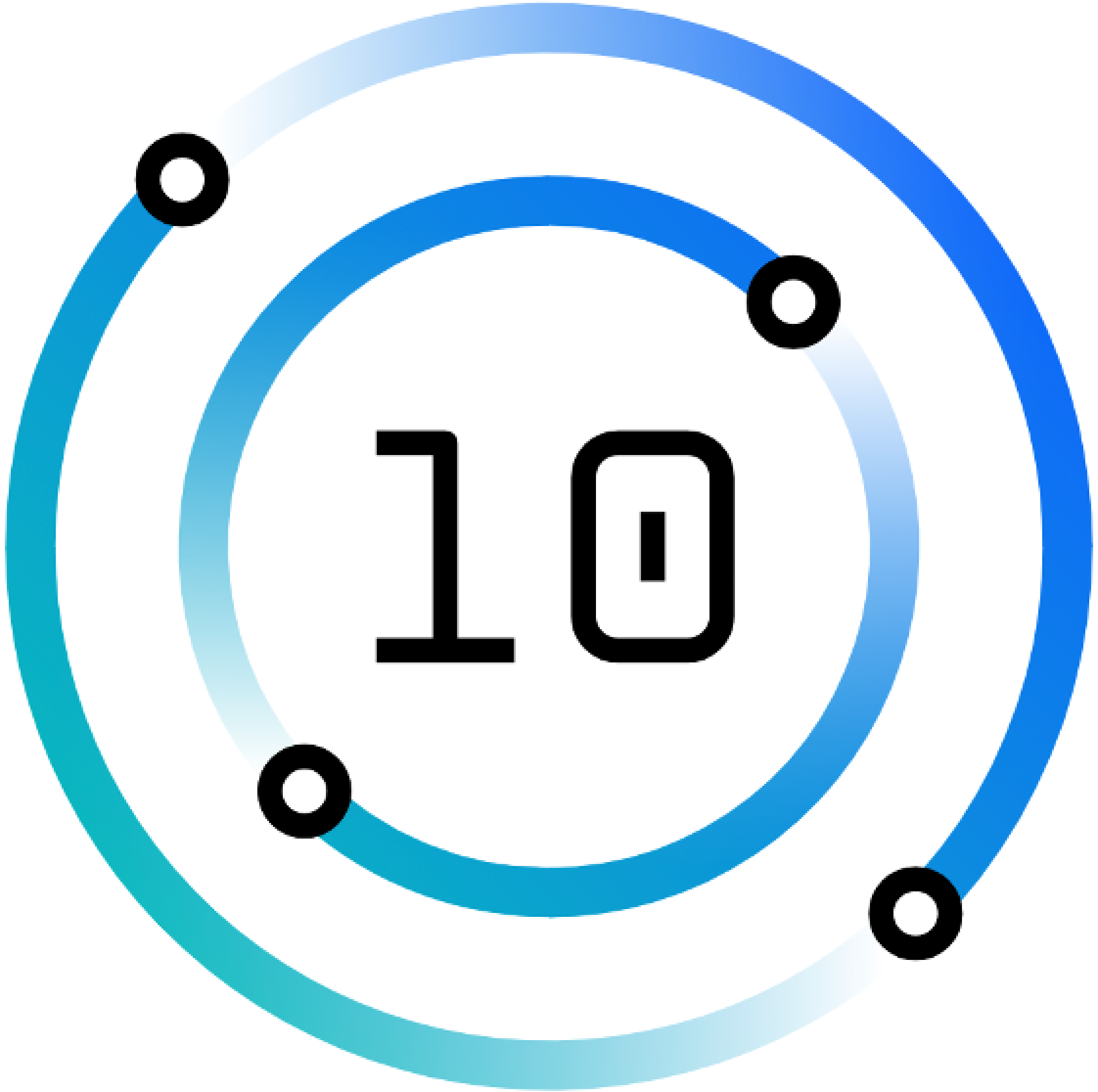
HNV
- Native SR-IOV for performance
- Virtual backup device for LPM (vNIC or vEth)
- HMC dynamically removes/adds native SR-IOV logical port as part of LPM operation

vNIC
- Partition LPM capable
- Cmds flow through VIOS.
- Data DMAs directly between partition memory and adapter

SEA/Virt Eth /w Logical Port as Physical device
- Compatible with older SEA/Virth Eth technology
- Enables use of native SR-IOV and vNIC now or in the future.
- Client partitions are LPM capable
- Control information and data copied by software/firmware.
- One SEA per physical port

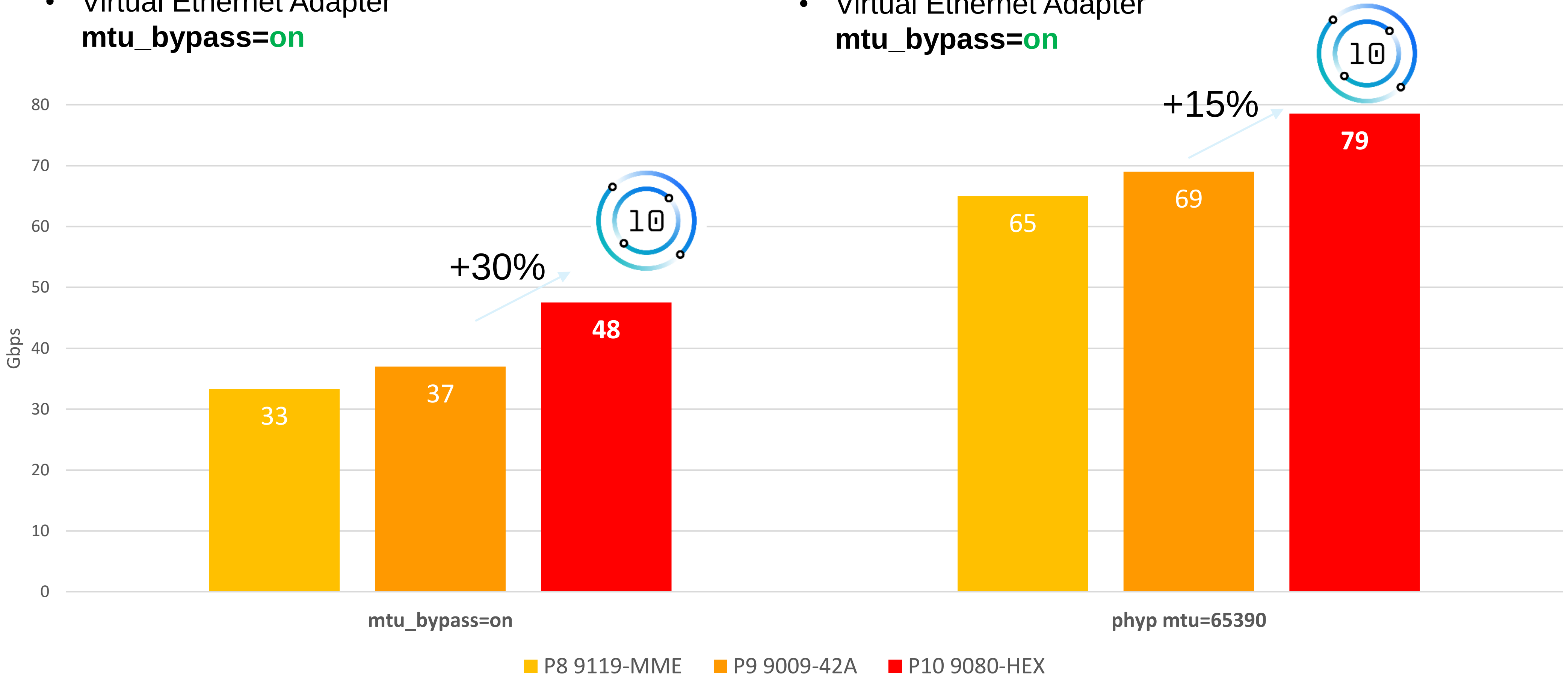
Power10

Network performance



Power10 Virtual Ethernet internal switching throughput

- **Benchmark with 8 parallel TCP sessions**
  - **Sender LPAR**:
    - **Power E1080 (9080-HEX)**
    - AIX 7.3
    - Entitled capacity: 4.00 / VCPUs: 4
    - Virtual Ethernet Adapter **mtu_bypass=on**
  - **Receiver LPAR**:
    - **Power E1080 (9080-HEX)** (same as sender)
    - AIX 7.3
    - Entitled capacity: 4.00 / VCPUs: 4
    - Virtual Ethernet Adapter **mtu_bypass=on**



Test with pre GA code. All subject to change.

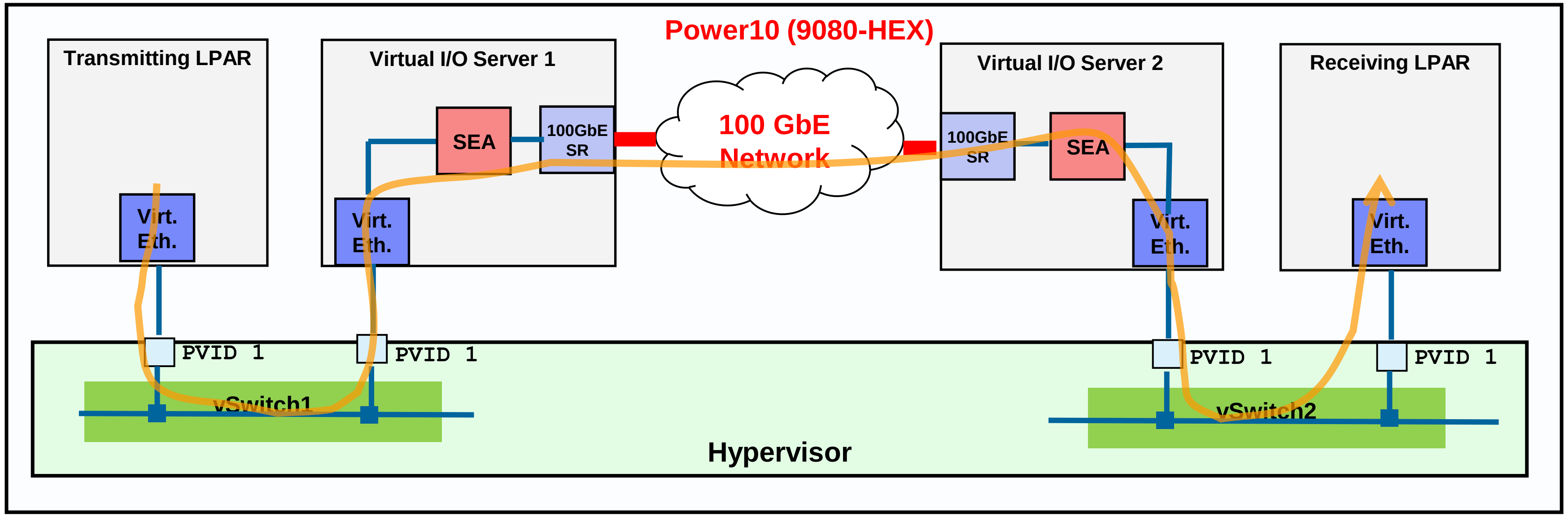
Power10 Shared Ethernet Adapter test setup

– Client LPARs

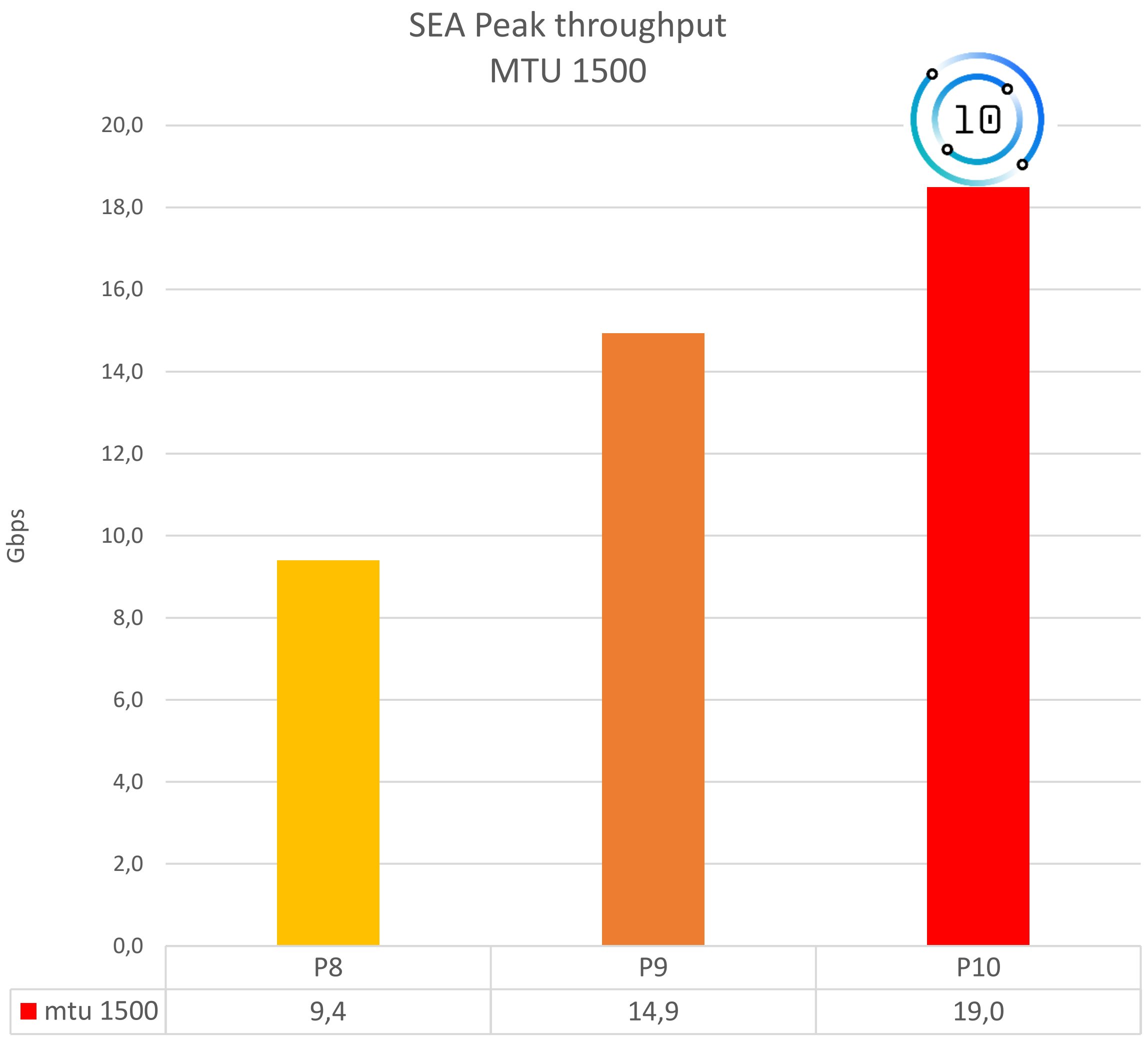
- Power E1080 (9080-HEX)
- AIX 7.3
- Entitled capacity: 8.00 / VCPUs: 8
- Virtual Ethernet Adapter
  mtu_bypass=on

–VIOS LPARs

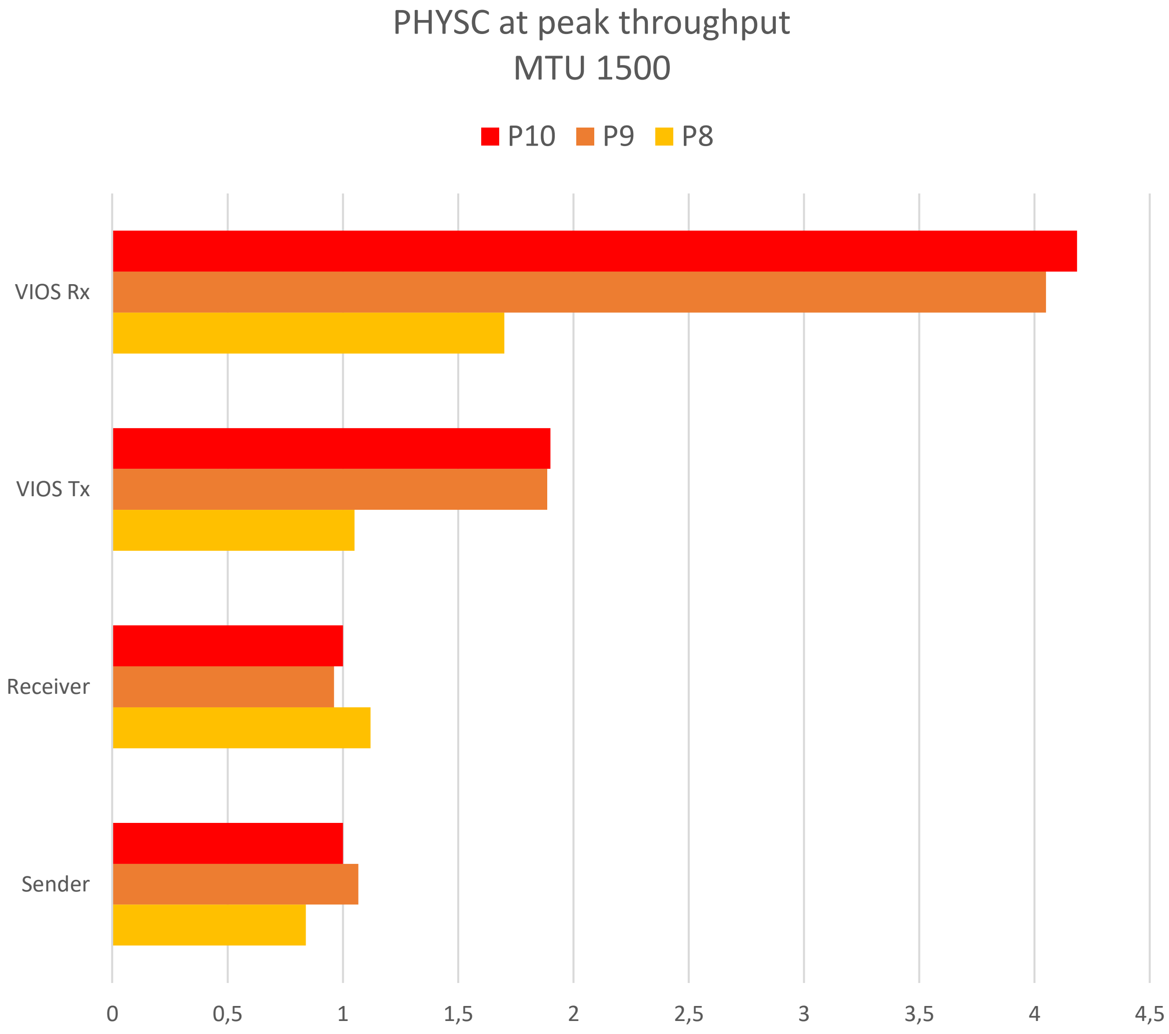
- Power E1080 (9080-HEX)
- VIOS 3.1.3.10
- Entitled capacity: 8.00 / VCPUs: 8
- Shared Ethernet Adapter
  largesend=1
  large_receive=yes



Power10 Shared Ethernet Adapter throughput



	P8	P9	P10
🟥 mtu 1500	9,4	14,9	19,0



Test with pre GA code. All subject to change.

P9 and P10 Shared Ethernet Adapter best practice - 1 / 3

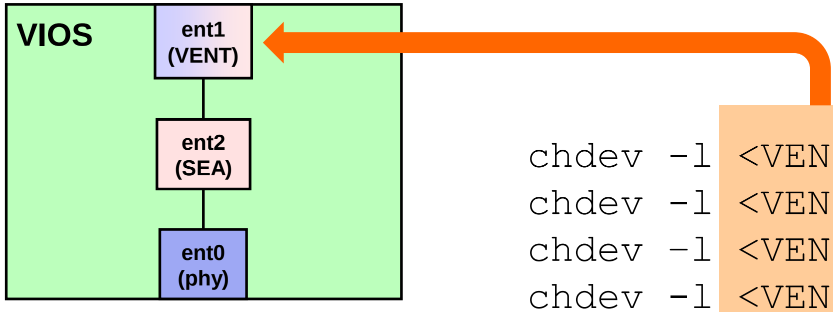
- **Rules of thumb for throughput**:
- SEA typically reaches a throughput of **~15 Gbps (mtu 1500)** or **~35 Gbps (mtu 9000)** per virtual trunk adapter.
- Multiple trunk adapters can be used to further increase the overall throughput with SEA.
- SEA Load Sharing already follows this approach by design.
- Note: All clients in the same VLAN / same vswitch are bound to one active trunk adapter and have to share the above maximum bandwidths.

- **Large Segment Offload** is required for gaining the above throughputs and to keep CPU utilization affordable.
- **Client: mtu_bypass=on**
- **VIOS: largesend=1 and large_receive=yes**

- The use of **Jumbo Frames** (mtu=9000) is recommended, especially when using 25, 40 or 100 Gigabit adapters.

P9 and P10 Shared Ethernet Adapter best practice - 2 / 3

- **Note: For high demanding workloads, the `max_min` setting might make no resource errors to not fully disappear.**
- In such case, manual tuning of minimum and maximum VETH buffers is required.
- Reboot required, so use `chdev -P` if SEA is in use.



```
chdev -l <VENT> -a max_buf_huge=128 -P
chdev -l <VENT> -a min_buf_huge=128 -P
chdev -l <VENT> -a max_buf_large=256 -P
chdev -l <VENT> -a min_buf_large=256 -P
chdev -l <VENT> -a max_buf_medium=2048 -P
chdev -l <VENT> -a min_buf_medium=2048 -P
chdev -l <VENT> -a max_buf_small=4096 -P
chdev -l <VENT> -a min_buf_small=4096 -P
chdev -l <VENT> -a max_buf_tiny=4096 -P
chdev -l <VENT> -a min_buf_tiny=4096 -P
```

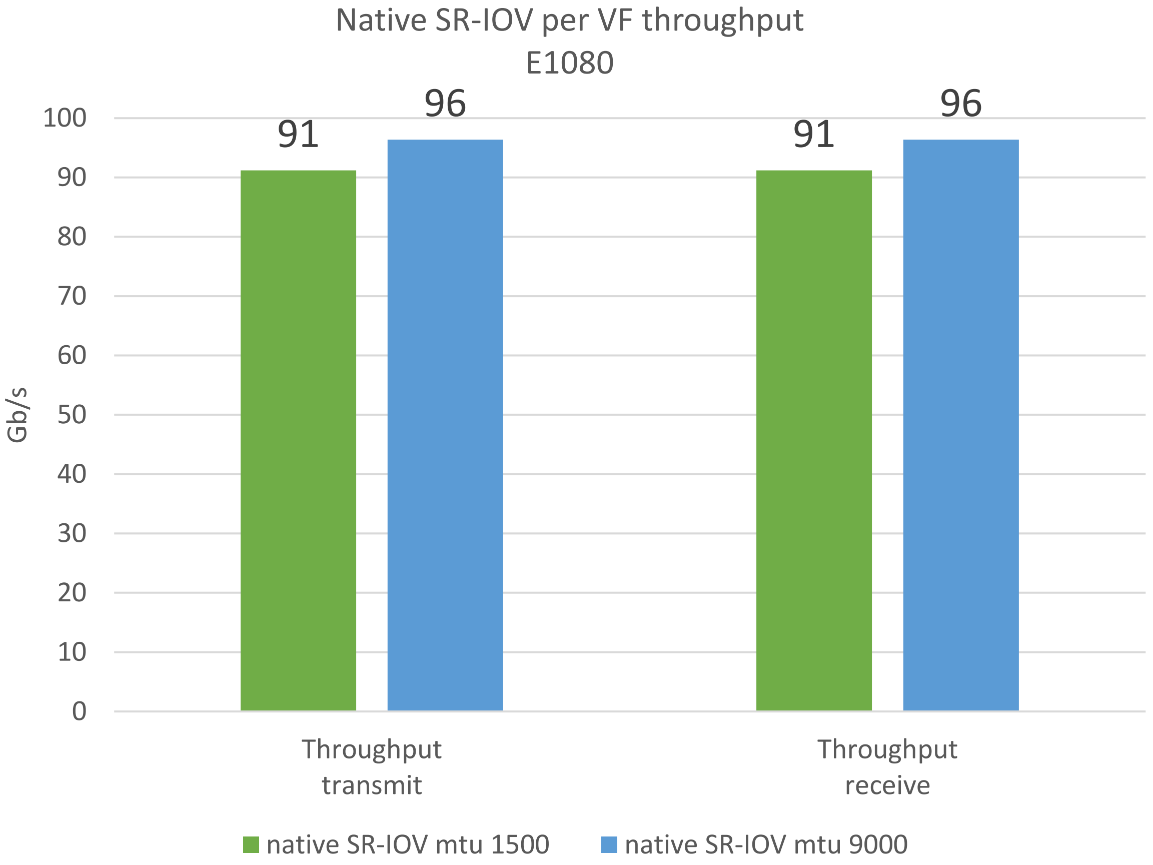
P9 and P10 Shared Ethernet Adapter best practice - 3 / 3

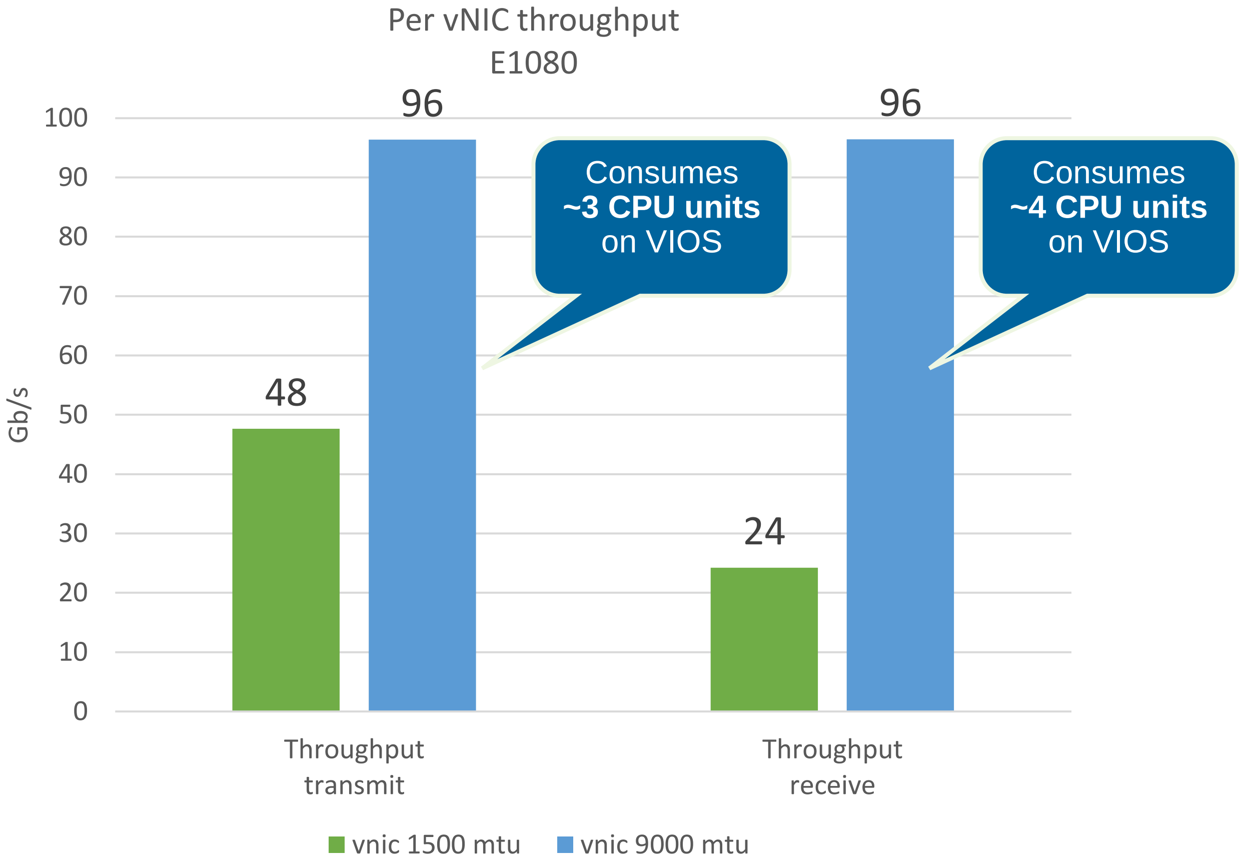
- Per SEA, ensure VIOSes have sufficient **CPU exclusively for network** available.
- SEA with **10 Gigabit: 2 CPU units** (dedicated or shared).
- SEA with **>10 Gigabit: 3-4 CPU units** (dedicated or shared).
- When using VIOS with shared processors, always set weight to 255.

- Depending on latency between sender and receiver, TCP send- and receive buffers need to be increased.
- AIX: **tcp_sendspace** and **tcp_recvspace** isno parameter
- Linux: **net.ipv4.tcp_rmem** and **net.ipv4.tcp_wmem**
- IBMi: CHGTCPA -> TCP receive and send buffer sizes

Power10 AIX native SR-IOV and vNIC throughput

- Power E1080 (9080-HEX) Systems Firmware MH1010
- PCIe4 2-port 100 Gigabit Ethernet Adapter (#EC66/#EC67)
- **AIX 7.3** and VIOS 3.1.3.10 (for vNIC)





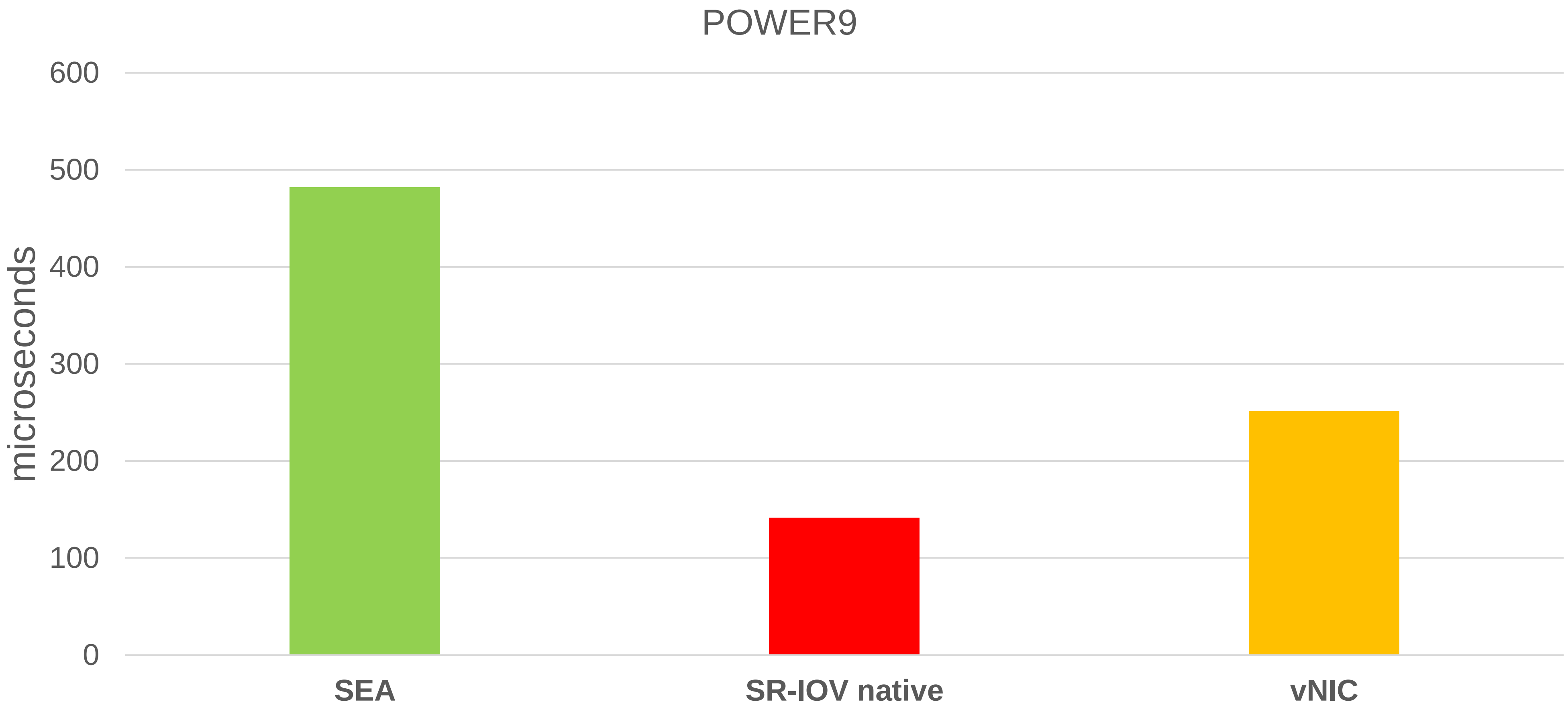
Consumes **~3 CPU units** on VIOS

Consumes **~4 CPU units** on VIOS

Test with pre GA code.

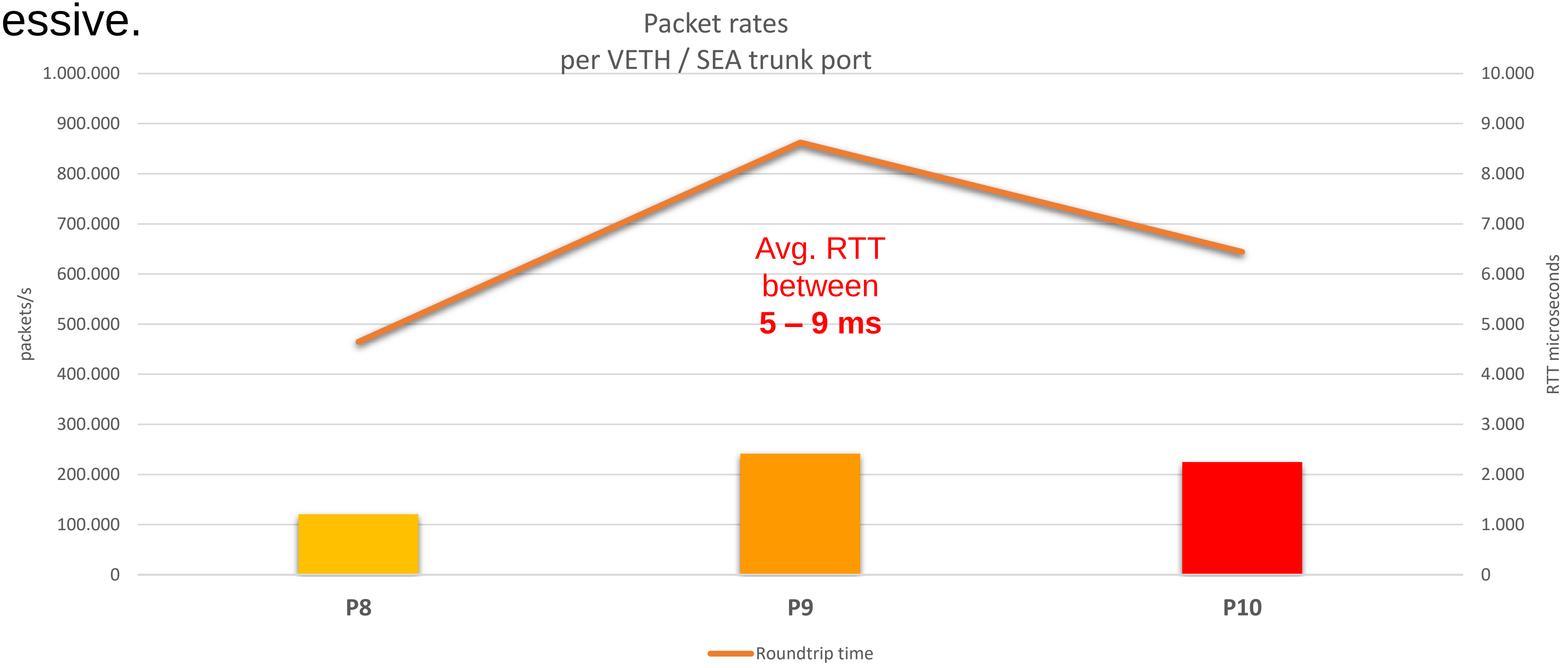
The impact of high transaction rates on latency

- Average Roundtrip Time (avg_rtt)
- Transactional workload with **100 parallel TCP sessions**



Virtual Ethernet / SEA packet rate and latency

- A typical maximum packet rate for a Virtual Ethernet (VETH) adapter is ~200K – 250K packets/s.
- This packet rate is per VETH client adapter or per trunk adapter for an SEA.
- Avoid running a VETH or trunk adapter into saturation, because RTT can become excessive.


Packet rates
per VETH / SEA trunk port

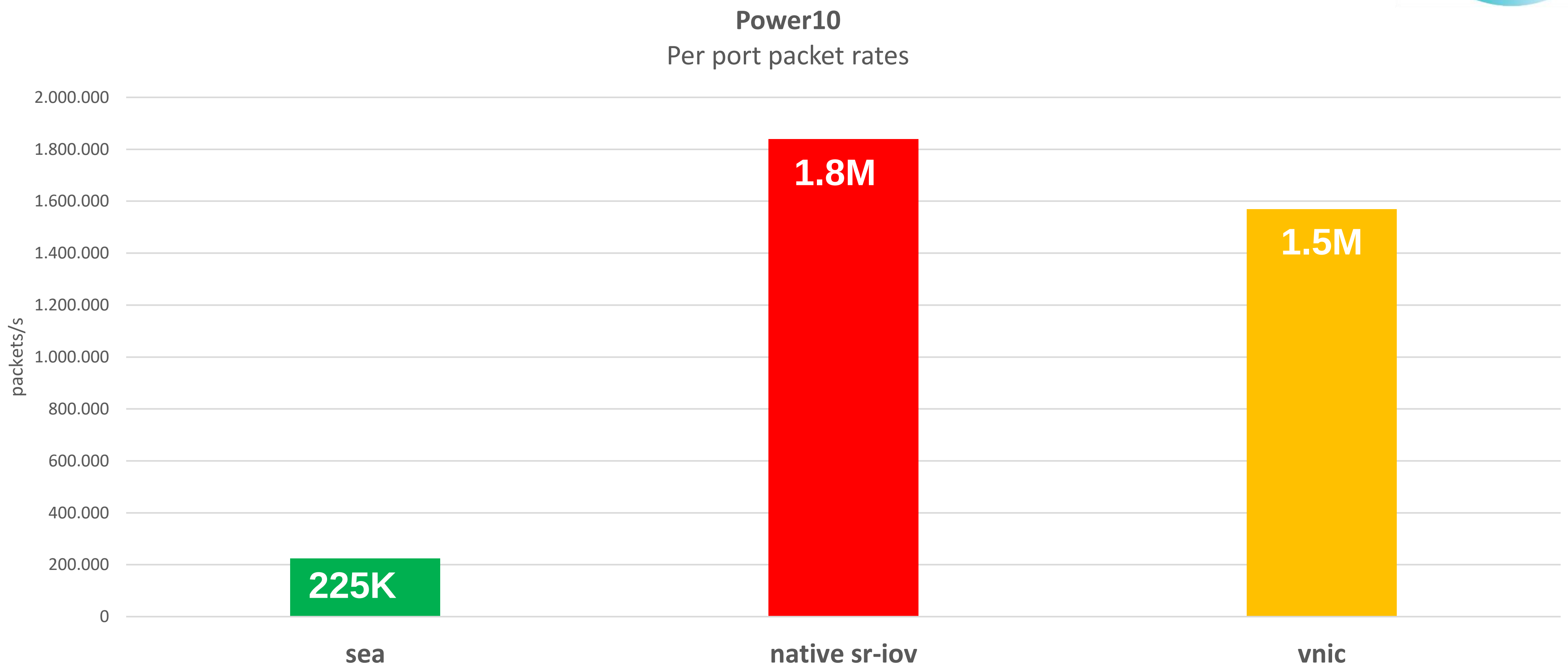
Avg. RTT between **5 – 9 ms**

Test with pre GA code. All subject to change.

2022 APaul

Power10 small packet transaction rates

- Power E1080 (9080-HEX) Systems Firmware MH1010
- PCIe4 2-port 100 Gigabit Ethernet Adapter (#EC66/#EC67)
- AIX 7.3 and VIOS 3.1.3.10 (for vNIC)

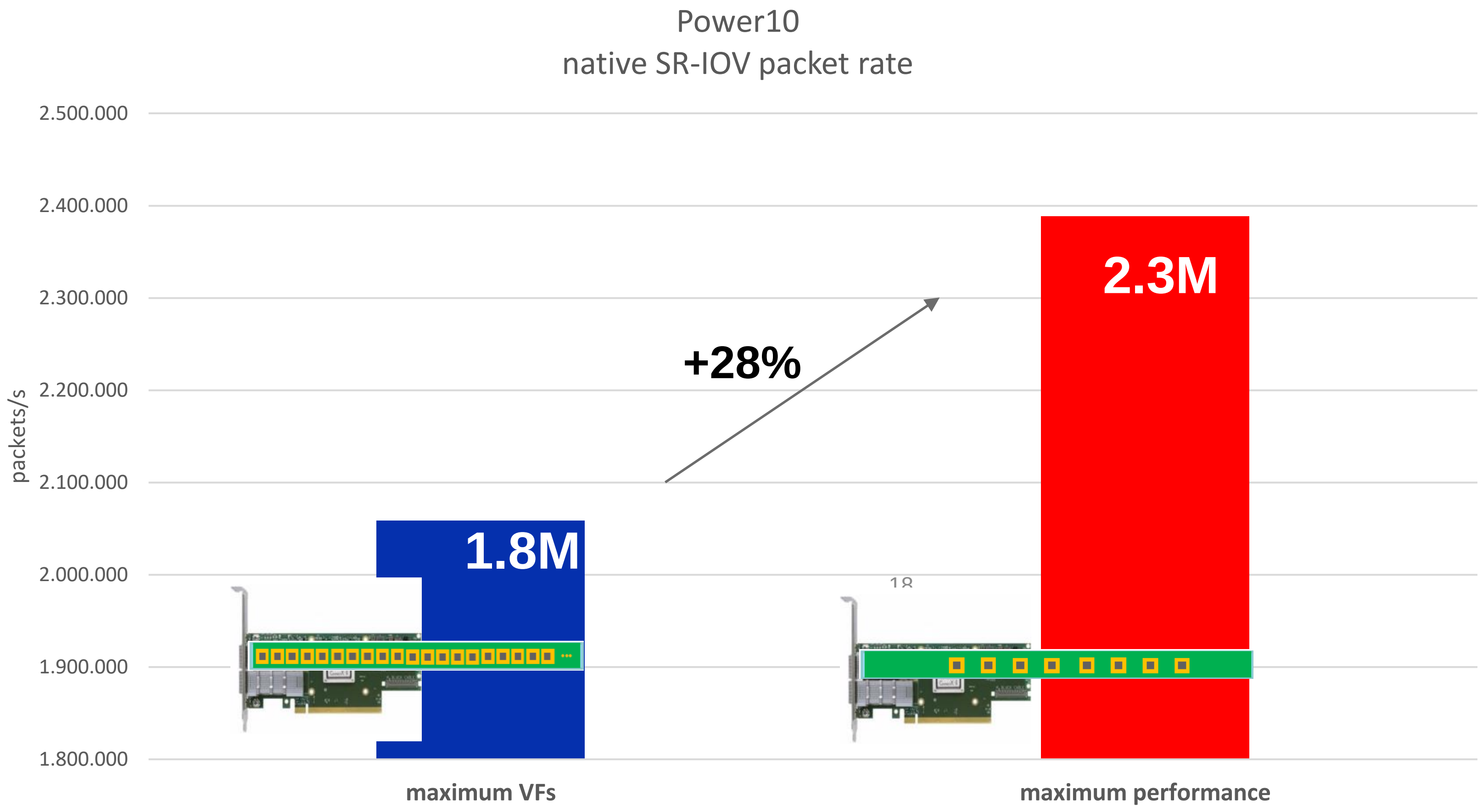


Test with pre GA code.

SR-IOV Maximum Performance Configuration Mode

- Power E1080 (9080-HEX) Systems Firmware MH1010
- PCIe4 2-port 100 Gigabit Ethernet Adapter (#EC66/#EC67)
- AIX 7.3 (7300-00-00)

- **Ethernet Logical Port Limits Maximum**
  - **8 VFs (high performance configuration mode)**
    - **`queues_rx=16`**
    - **`queues_tx=16`**
  - **60 VFs (high fanout configuration mode)**
    - **`queues_rx=8`**
    - **`queues_tx=8`**

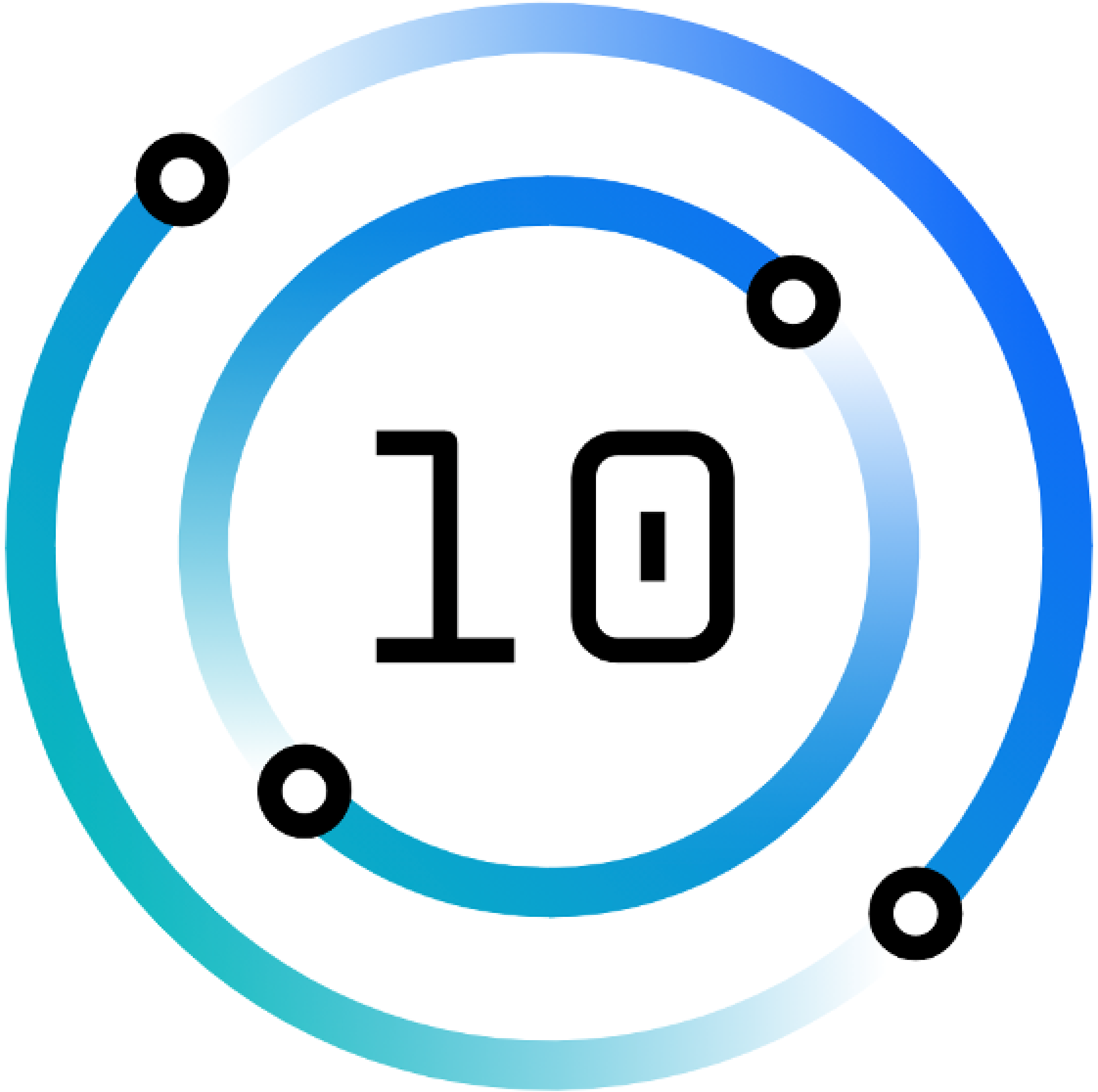


Test with pre GA code. All subject to change.

Power10

Sneak preview

FW1030



All subject to change.

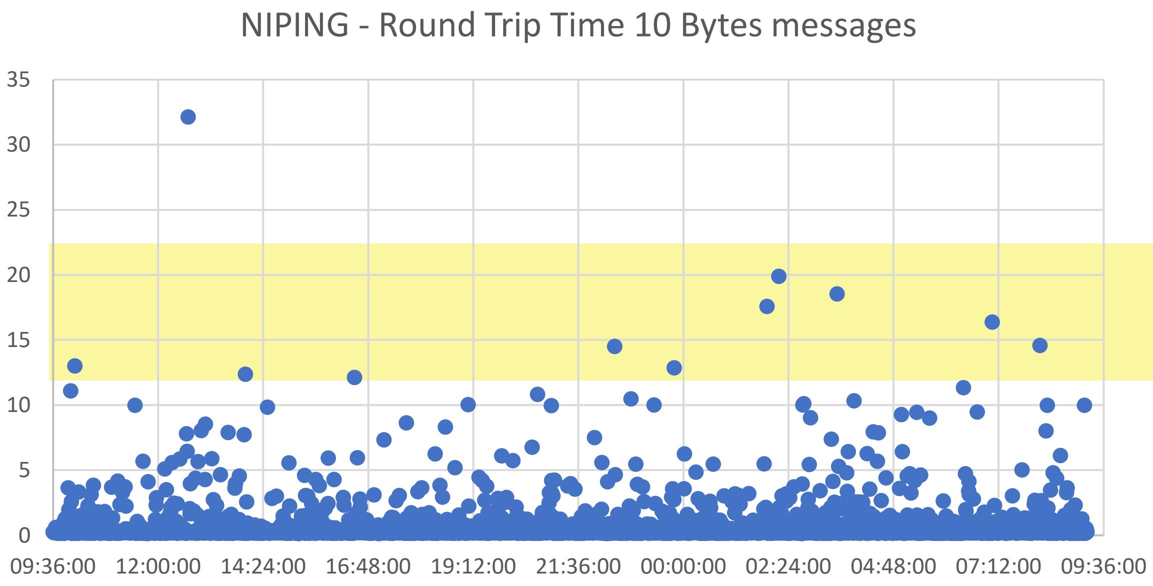
Case studies

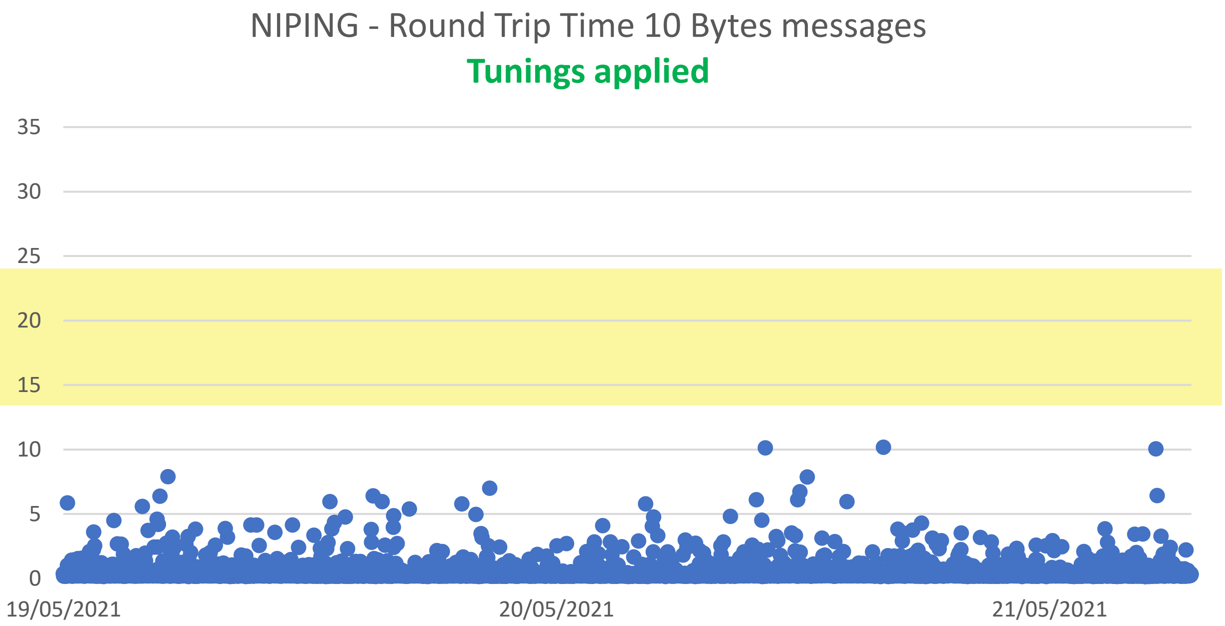
Case study #1 - SAP latency case study - make sure you grab low hanging fruits

Test #	note	#samples	RTT avg.	RTT > 1ms (% of total)		RTT > 5ms (% of > 1 ms)		RTT > 10ms (% of > 1 ms)		RTT > 20ms (% of > 1 ms)	
1	baseline	2907	0,71	391	13%	73	19%	18	5%	1	0%

- Two tunings were performed for test #2

```
schedo -p -o vpm_xvcpus=2
no -p -o ifstats32=0
```

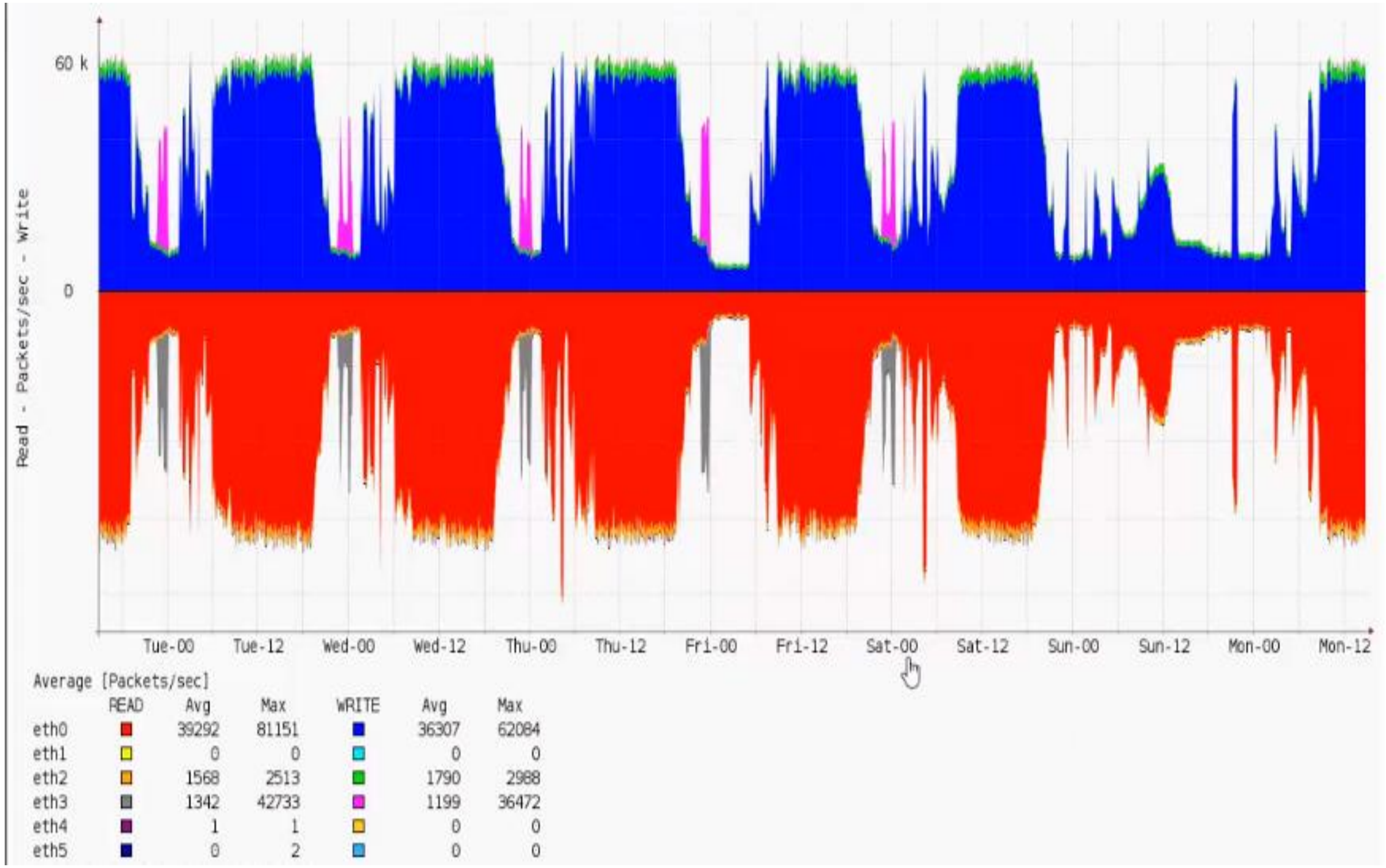

NIPING - Round Trip Time 10 Bytes messages


NIPING - Round Trip Time 10 Bytes messages

Tunings applied

Case study #2 - Network issue or not?

- Financial (transactional) workload on a E980 with dual VIOS Shared Ethernet Adapter setup w. 25 Gigabit physical adapters in sharing mode.
- Flat average utilization of ~60K packets/s in sending and ~80K packets/s receiving direction can be observed every business day.
- Application showed frequent high latency events as well as application timeouts.

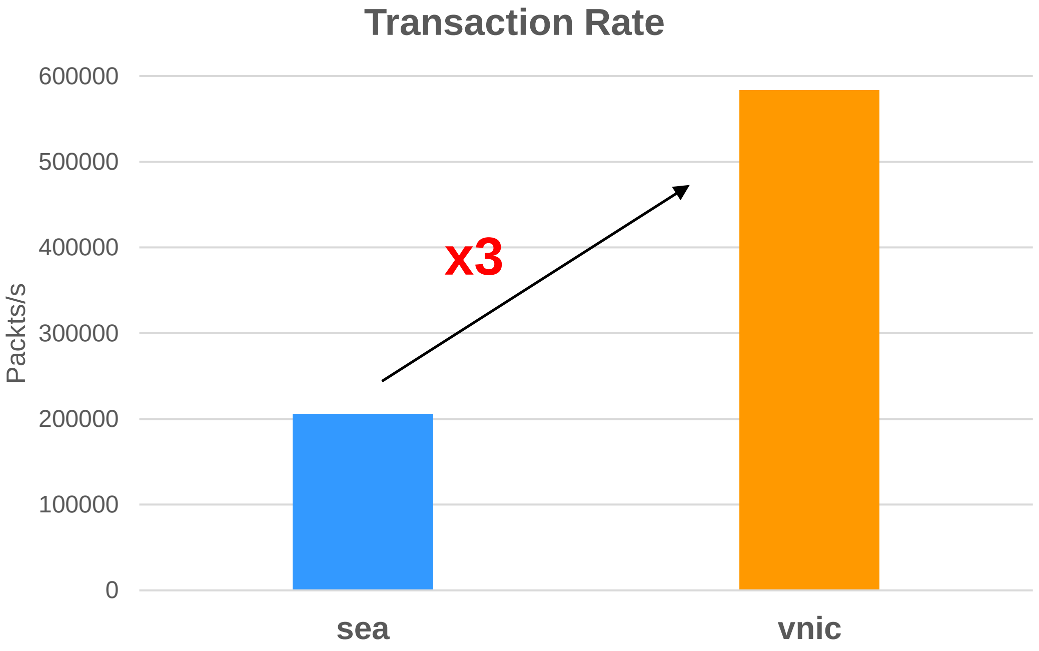


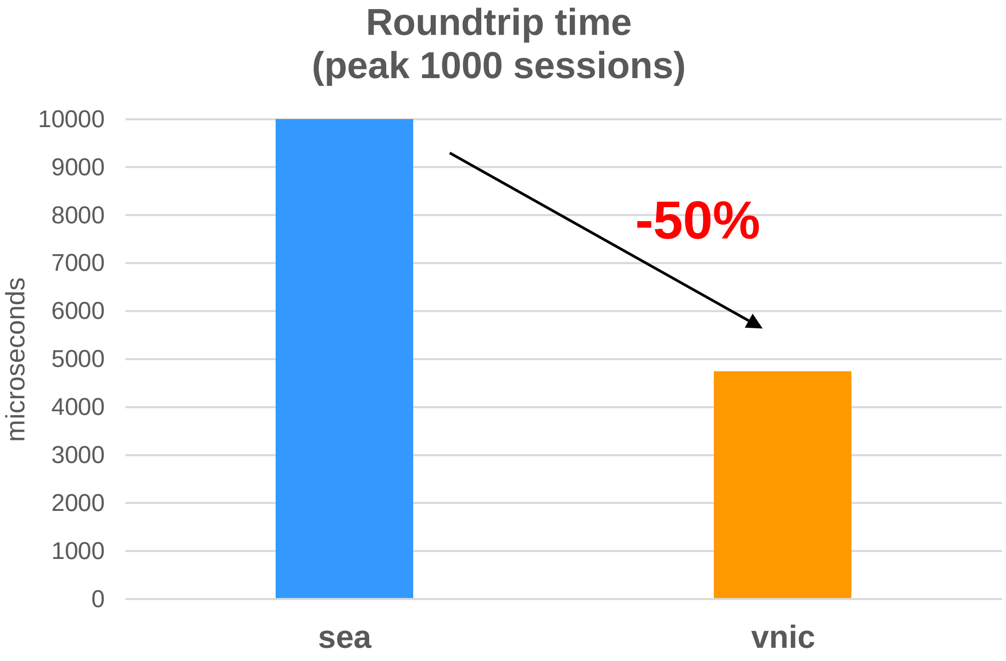
Average [Packets/sec]

	READ	Avg	Max	WRITE	Avg	Max
eth0	🟥	39292	81151	🟦	36307	62084
eth1	🟨	0	0	🟦	0	0
eth2	🟧	1568	2513	🟩	1790	2988
eth3	■	1342	42733	🟪	1199	36472
eth4	🟪	1	1	🟨	0	0
eth5	🟦	0	2	🟦	0	0

Case study #2 - Network issue or not?

- Evaluated vNIC for user (frontend) connection with high packet rates.
- KPI stress tests confirmed that vNIC with default settings provided 3x transaction rates but half the roundtrip time than SEA.
- After going live with vNIC in production, application response times normalized, and timeouts did not occur again.

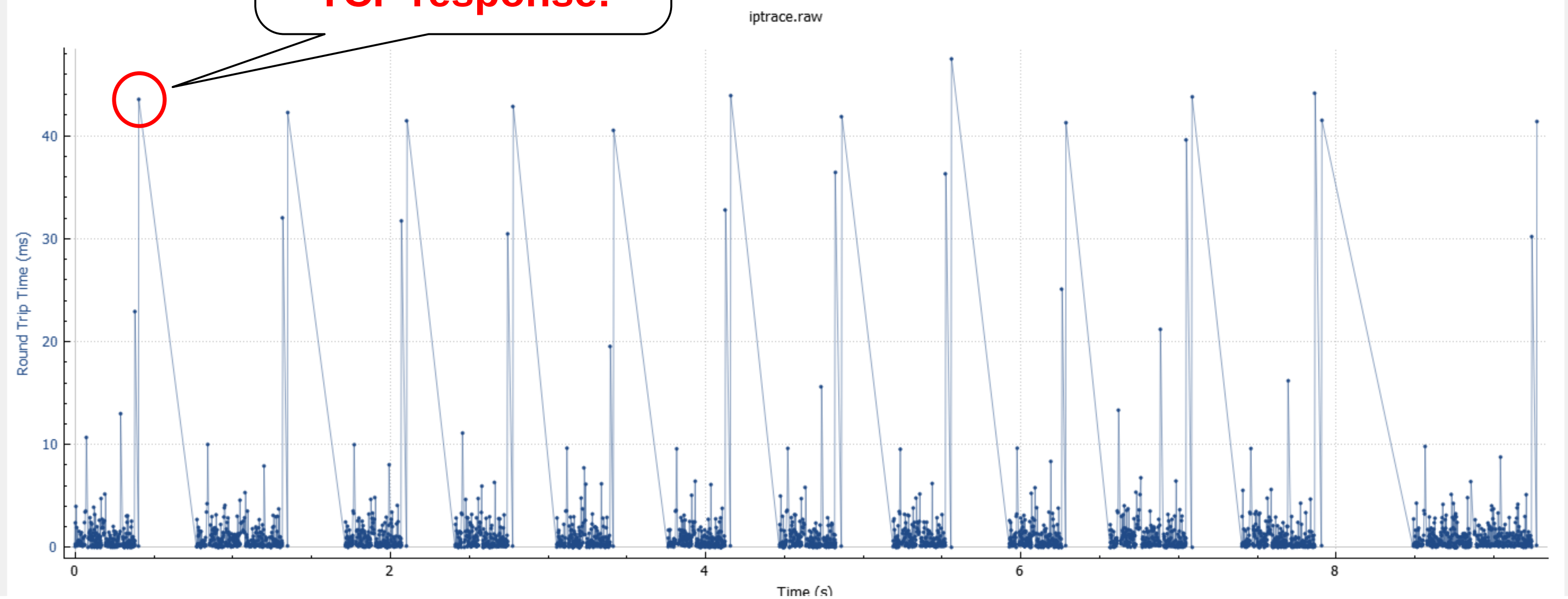




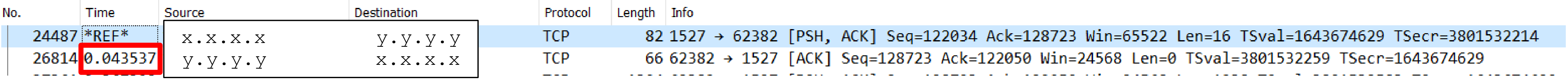
Case study #3 – Yet another network issue?

- Multiple application servers were experiencing significant increase in response time after migration from Power9 to Power10.

App Server takes 40 ms to send TCP response!

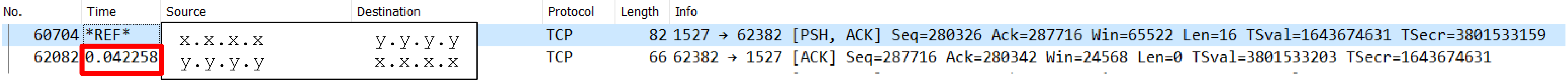


DB Server



No.	Time	Source	Destination	Protocol	Length	Info
24487	*REF*	x.x.x.x	y.y.y.y	TCP	82	1527 → 62382 [PSH, ACK] Seq=122034 Ack=128723 Win=65522 Len=16 TSval=1643674629 TSecr=3801532214
26814	0.043537	y.y.y.y	x.x.x.x	TCP	66	62382 → 1527 [ACK] Seq=128723 Ack=122050 Win=24568 Len=0 TSval=3801532259 TSecr=1643674629

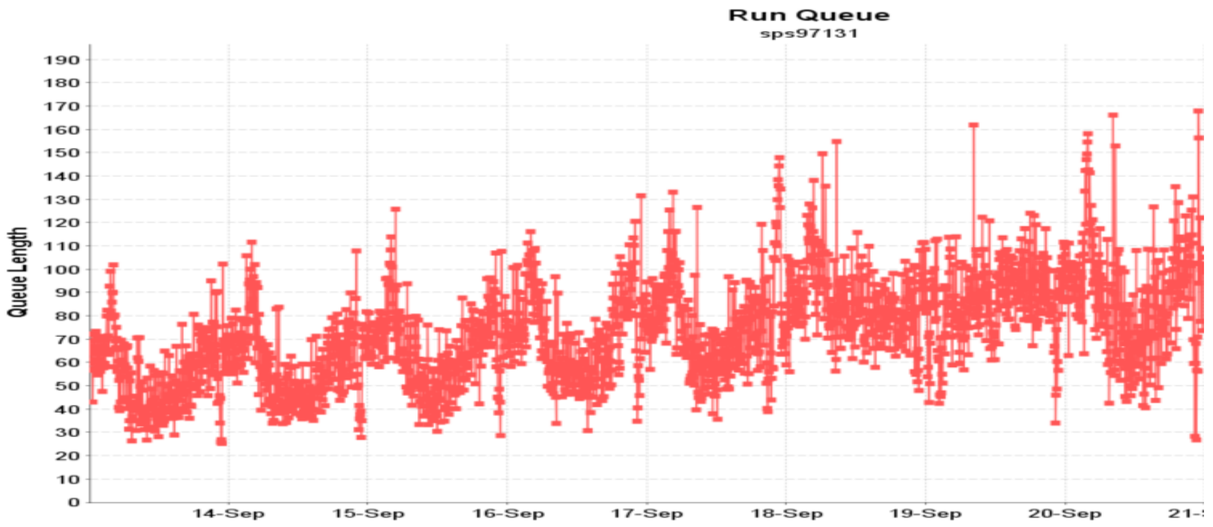
App Server



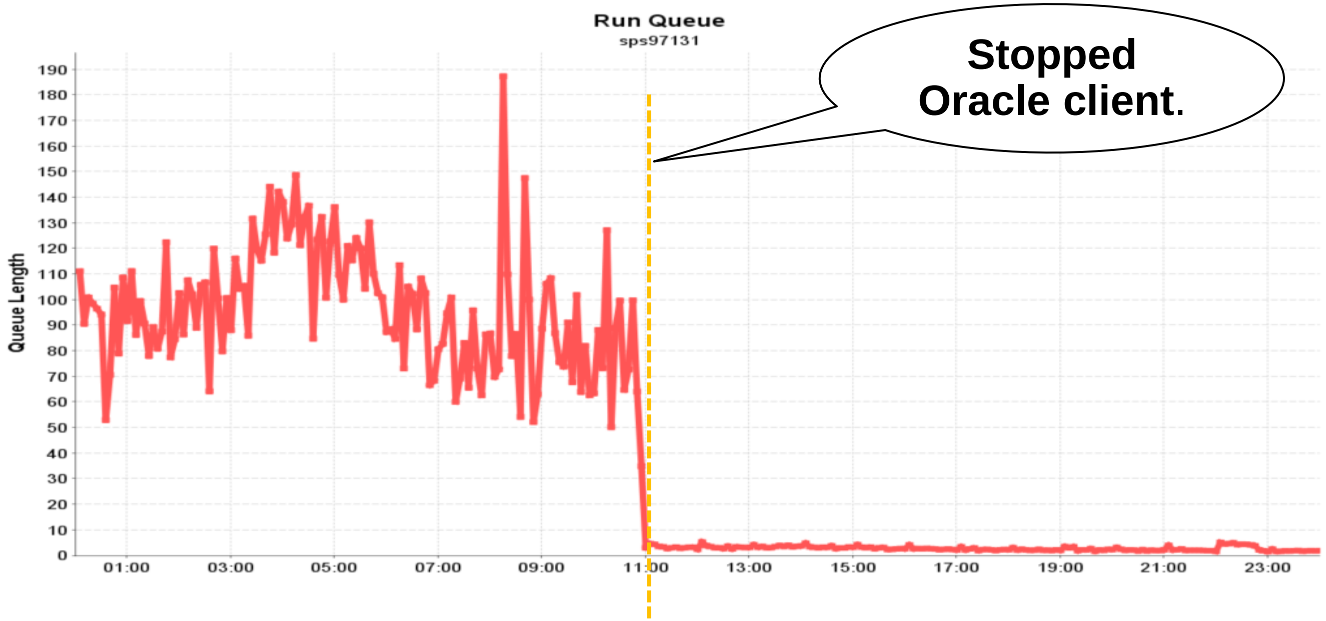
No.	Time	Source	Destination	Protocol	Length	Info
60704	*REF*	x.x.x.x	y.y.y.y	TCP	82	1527 → 62382 [PSH, ACK] Seq=280326 Ack=287716 Win=65522 Len=16 TSval=1643674631 TSecr=3801533159
62082	0.042258	y.y.y.y	x.x.x.x	TCP	66	62382 → 1527 [ACK] Seq=287716 Ack=280342 Win=24568 Len=0 TSval=3801533203 TSecr=1643674631

Case study #3 – Yet another network issue?

- Application server hit a bug with a recently installed update of an Oracle agent.
- The agent software spawned accumulating, CPU intensive threads.
- Consequence: Over time, LPAR became CPU bound.



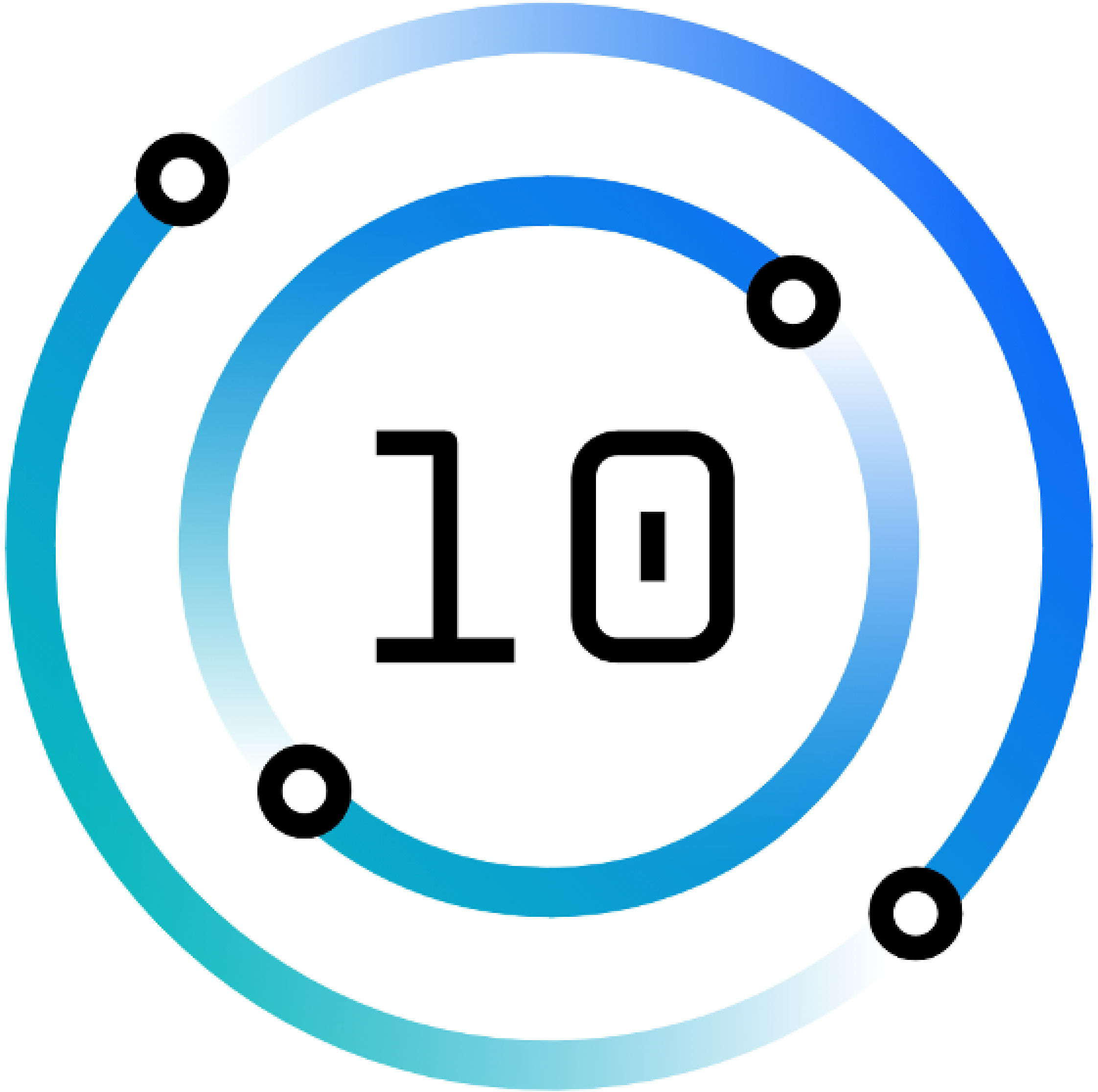
Not a network issue!



Stopped Oracle client.

Linux

Performance best practice



Manage Adapter Queues for vNIC

The ibmvnic kernel module offers multiple queues dependent on the physical adapter.

vNIC by default utilized 2x send and 8x receive queues.

For better throughput and/or latency:

```
# ethtool -L eth# rx 8 tx 8
```

If the physical SR-IOV adapter port is configured for **maximum performance**, up to 16 queues are available per vNIC.

```
# ethtool -L eth# rx 16 tx 16
```

vNIC Adapter

```
# ethtool -l eth0
Channel parameters for eth0:
Pre-set maximums:
RX:             16
TX:             16
Other:          0
Combined:       0
Current hardware settings:
RX:             8
TX:             2
Other:          0
Combined:       0
```

IRQ binding for ibmvnic

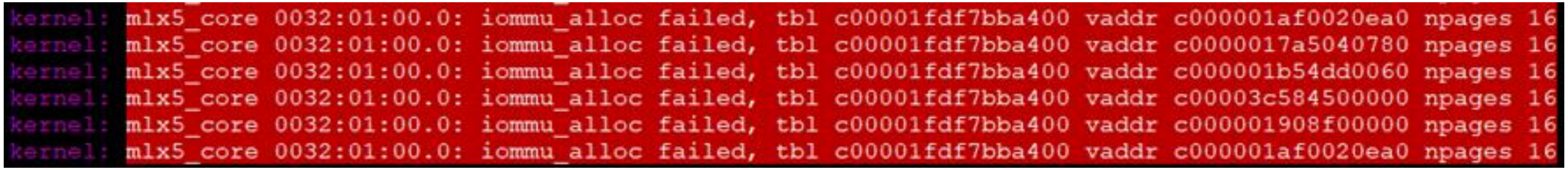
- Manual IRQ binding can further increase vNIC performance.
- Irqbalancer nedds to be stopped first.
  `#systemctl stop irqbalance`
- Example for a LPAR with 8 Virtual CPUs:

```
ethtool -L $eth rx 8 tx 8
TXQs=$(cat /proc/interrupts | grep ibmvnic | grep tx | cut -d: -f2 | tr -d ' ')
RXQs=$(cat /proc/interrupts | grep ibmvnic | grep rx | cut -d: -f2 | tr -d ' ')
i=0 ; for rxq in $RXQs
do
    echo $i > /proc/irq/${rxq}/smp_affinity_list
    let i=i+4
done

i=2 ; for txq in $TXQs
do
    echo $i > /proc/irq/${txq}/smp_affinity_list
    let i=i+4
done
```

Direct Memory Access (DMA) Window and I/O Adapter Enlarged Capacity

- The **DMA window** is a specific memory address range that a PCIe adapter is allowed to access.
- Typically, the DMA window is ~2 GB in size.
- DMA allocation errors occur, when a device driver requests additional DMA mappings beyond the DMA window size.
- Example #1:
  - iommu_alloc failure messages indicate the driver cannot allocate any more DMA resources:



```
kernel: mlx5_core 0032:01:00.0: iommu_alloc failed, tbl c00001fdf7bba400 vaddr c000001af0020ea0 npages 16
kernel: mlx5_core 0032:01:00.0: iommu_alloc failed, tbl c00001fdf7bba400 vaddr c0000017a5040780 npages 16
kernel: mlx5_core 0032:01:00.0: iommu_alloc failed, tbl c00001fdf7bba400 vaddr c000001b54dd0060 npages 16
kernel: mlx5_core 0032:01:00.0: iommu_alloc failed, tbl c00001fdf7bba400 vaddr c00003c584500000 npages 16
kernel: mlx5_core 0032:01:00.0: iommu_alloc failed, tbl c00001fdf7bba400 vaddr c000001908f00000 npages 16
kernel: mlx5_core 0032:01:00.0: iommu_alloc failed, tbl c00001fdf7bba400 vaddr c000001af0020ea0 npages 16
```

- Example #2
  - Increasing the ring buffers for a PCIe adapter results in memory allocation error:

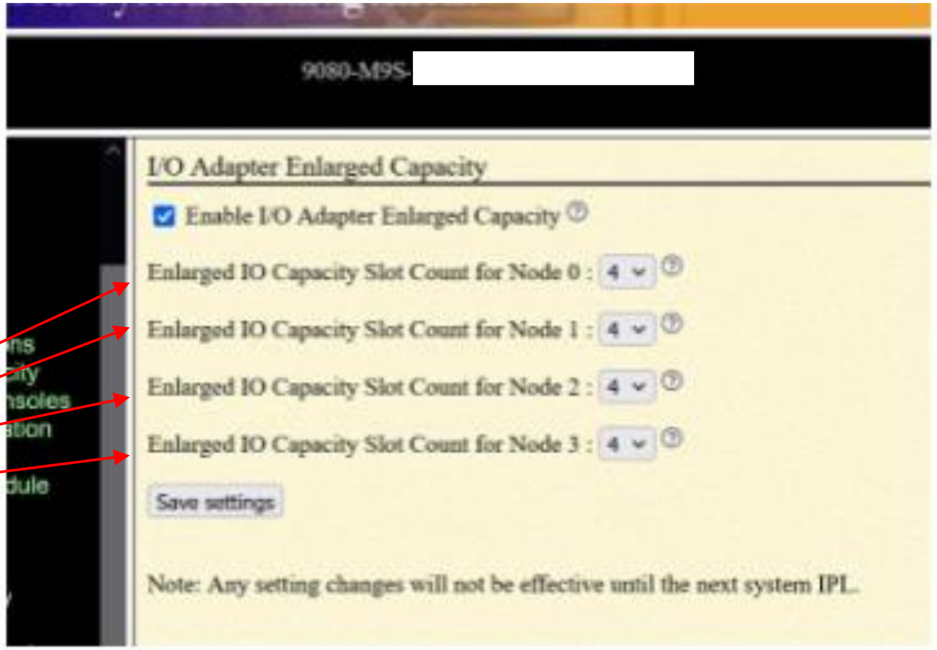
```
# ethtool -G eth22 rx 8192 tx 8192

Cannot set device ring parameters: Cannot allocate memory
```

Direct Memory Access (DMA) Window and I/O Adapter Enlarged Capacity

- **Solution**:
  - Enable I/O Enlarged Capacity for PCIe adapters provides a wider (64-bit Huge) DMA window.
  - With 64-bit DMA window, possibly all the partition’s memory could be mapped for DMA.
  - Avoids DMA allocation errors, improves latency but requires more system memory.
- To enable **I/O Adapter Enlarged Capacity**:

1. From the ASMI menu, select **System Configuration** > **I/O Adapter Enlarged Capacity**.
2. Select **Enable I/O Adapter Enlarged Capacity.**
3. Click **Save** to save settings.
4. Restart the system.



9080-M9S-[redacted]

I/O Adapter Enlarged Capacity

☑ Enable I/O Adapter Enlarged Capacity

Enlarged IO Capacity Slot Count for Node 0 : 4

Enlarged IO Capacity Slot Count for Node 1 : 4

Enlarged IO Capacity Slot Count for Node 2 : 4

Enlarged IO Capacity Slot Count for Node 3 : 4

Save settings

Note: Any setting changes will not be effective until the next system IPL.

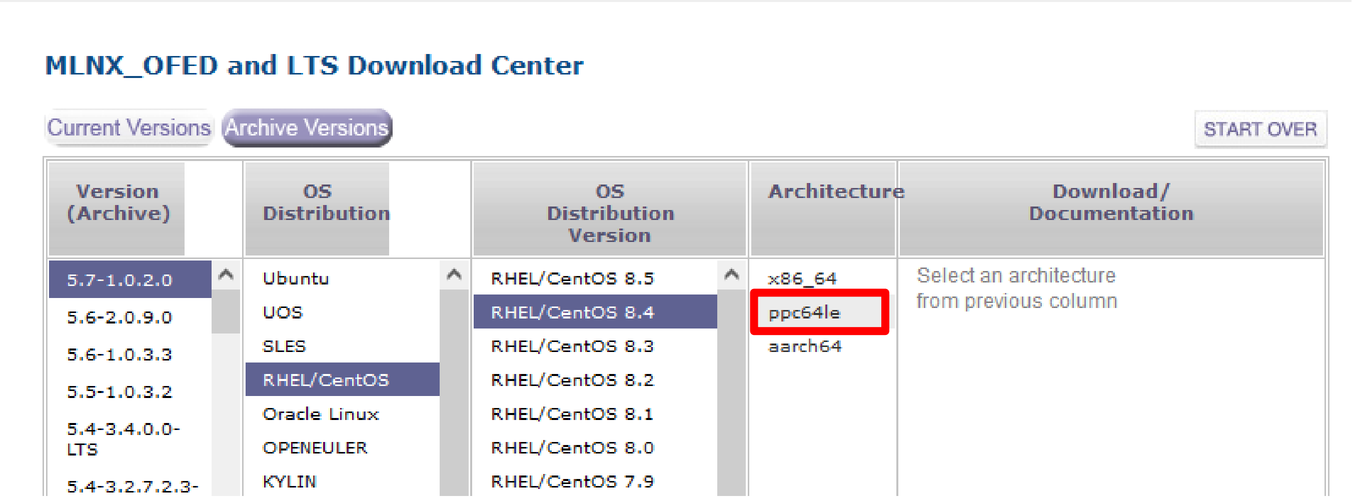
Example:
4x 2-port 100 Gigabit Ethernet adapters
(#EC66)

Mellanox OpenFabrics Enterprise Distribution for Linux

- RHEL and SLES distributions contain default mlx5_core kernel module.

```
# cat /etc/redhat-release
Red Hat Enterprise Linux release 8.3 (Ootpa)
# modinfo mlx5_core
filename:       /lib/modules/4.18.0-
250.el8.dt3.ppc64le/kernel/drivers/net/ethernet/mellanox/mlx5/core/mlx5_core.ko.xz
version:        5.0-0
license:        Dual BSD/GPL
description:    Mellanox 5th generation network adapters (ConnectX series) core driver
```

- Mellanox OpenFabrics Enterprise Distribution for Linux (MLNX_OFED)
  https://www.mellanox.com/products/infiniband-drivers/linux/mlnx_ofed



MLNX_OFED and LTS Download Center

Version (Archive)	OS Distribution	OS Distribution Version	Architecture	Download/ Documentation
5.7-1.0.2.0	Ubuntu	RHEL/CentOS 8.5	x86_64	Select an architecture from previous column
5.6-2.0.9.0	UOS	RHEL/CentOS 8.4	ppc64le	
5.6-1.0.3.3	SLES	RHEL/CentOS 8.3	aarch64	
5.5-1.0.3.2	RHEL/CentOS	RHEL/CentOS 8.2		
5.4-3.4.0.0-LTS	Oracle Linux	RHEL/CentOS 8.1		
5.4-3.2.7.2.3-	OPENEULER	RHEL/CentOS 8.0		
	KYLIN	RHEL/CentOS 7.9		

MLNX_OFED installation on RHEL

Example RHEL 8.4

```
# yum -y install kernel-modules-extra gcc-gfortran createrepo
# wget https://content.mellanox.com/ofed/MLNX_OFED-5.4-1.0.3.0/MLNX_OFED_LINUX-5.4-1.0.3.0-rhel8.4-ppc64le.tgz
# tar -xzvf MLNX_OFED_LINUX-5.4-1.0.3.0-rhel8.4-ppc64le.tgz
# cd MLNX_OFED_LINUX-5.4-1.0.3.0-rhel8.4-ppc64le

# ./mlnxofedinstall --add-kernel-support
# dracut -f
# reboot now

# To load the new driver, run:
/etc/init.d/openibd restart



# ibdev2netdev
mlx5_0 port 1 ==> eth0 (Down)
mlx5_1 port 1 ==> eth4 (Up)

# ibdev2netdev -v
000d:01:00.0 mlx5_0 (MT4122 - 01FT742SN  ) 2-Port PCIe4 100GbE RoCE Adapter x16 fw 16.30.1004 port 1 (ACTIVE) ==> eth0 (Up)
0012:01:00.0 mlx5_1 (MT4122 - 01FT742SN  ) 2-Port PCIe4 100GbE RoCE Adapter x16 fw 16.30.1004 port 1 (ACTIVE) ==> eth4 (Up)
```

MLNX_OFED installation on SLES

Example SLES 15.3

```
# zypper install createrepo_c kernel-source python2 python3-devel kernel-syms insserv-compat
```

Required when running a different kernel version

```
# zypper install --oldpackage kernel-source-5.3.18-59.19.1.noarch kernel-syms-5.3.18-59.19.1.ppc64le

# wget https://content.mellanox.com/ofed/MLNX_OFED-5.4-1.0.3.0/MLNX_OFED_LINUX-5.4-1.0.3.0-sles15sp3-ppc64le.tgz
# tar -xzvf MLNX_OFED_LINUX-5.4-1.0.3.0-sles15sp3-ppc64le.tgz
# cd MLNX_OFED_LINUX-5.4-1.0.3.0-sles15sp3-ppc64le

# ./mlnxofedinstall --add-kernel-support
# dracut -f
# reboot now

# To load the new driver, run:
/etc/init.d/openibd restart



# ibdev2netdev
mlx5_0 port 1 ==> eth0 (Down)
mlx5_1 port 1 ==> eth4 (Up)

# ibdev2netdev -v
000d:01:00.0 mlx5_0 (MT4122 - 01FT742SN   ) 2-Port PCIe4 100GbE RoCE Adapter x16 fw 16.30.1004 port 1 (ACTIVE) ==> eth0 (Up)
0012:01:00.0 mlx5_1 (MT4122 - 01FT742SN   ) 2-Port PCIe4 100GbE RoCE Adapter x16 fw 16.30.1004 port 1 (ACTIVE) ==> eth4 (Up)
```

Thank You!

Alexander Paul
Senior Systems and Network Engineer
paul.alex@ch.ibm.com

Trademarks and Disclaimers

-
- References in this document to IBM products or services do not imply that IBM intends to make them available in every country.
- Trademarks of International Business Machines Corporation in the United States, other countries, or both can be found on the World Wide Web at https://www.ibm.com/legal/us/en/copytrade.shtml
  - Trademarks owned by IBM, for details see the URL above.
  - Special attributions, the listed trademarks of the following companies require attribution:
    - Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
    - IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.
    - ITIL is a Registered Trade Mark of AXELOS Limited.
    - Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.
    - Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
    - Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
    - Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
    - Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
    - Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
    - UNIX is a registered trademark of The Open Group in the United States and other countries.
    - VMware, the VMware logo, VMware Cloud Foundation, VMware Cloud Foundation Service, VMware vCenter Server, and VMware vSphere are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.
- Information in this presentation is provided "AS IS" without warranty of any kind.
  - The customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.
  - Information concerning non-IBM products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by IBM. Sources for non-IBM list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. IBM has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-IBM products. Questions on the capability of non-IBM products should be addressed to the supplier of those products.
  - All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
  - Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in IBM product announcements. The information is presented here to communicate IBM's current investment and development activities as a good faith effort to help with our customers' future planning.
  - Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.
  - Prices are suggested U.S. list prices and are subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.
- The Client logo used with their permission within this presentation.