TA40

ESX Server Networking Performance

Bhavjit Walha
MTS
VMware

Shilpi Agarwal
Senior MTS
VMware

This session may contain product features that are currently under development.

This session/overview of the new technology represents no commitment from VMware to deliver these features in any generally available product.

Features are subject to change and must not be included in contracts, purchase orders, or sales agreements of any kind.

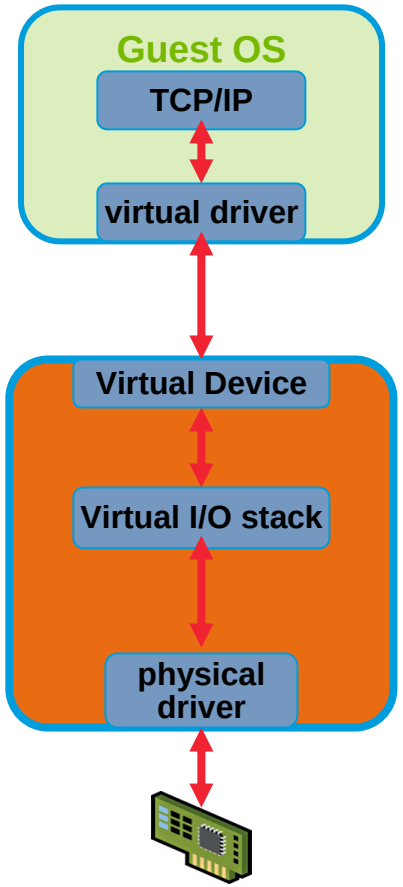
Technical feasibility and market demand will affect final delivery.

Pricing and packaging for any new technologies or features discussed or presented have not been determined.

Agenda

- **ESX Architecture - Network I/O**
- **Benchmarking methodology and results**
  - Performance trends – ESX 2.5 vs. ESX 3.x vs. Native
  - Comparison of virtual network devices
  - Performance scalability
- **Future Directions**
- **Benchmarking Guidelines**

ESX Architecture – Network I/O

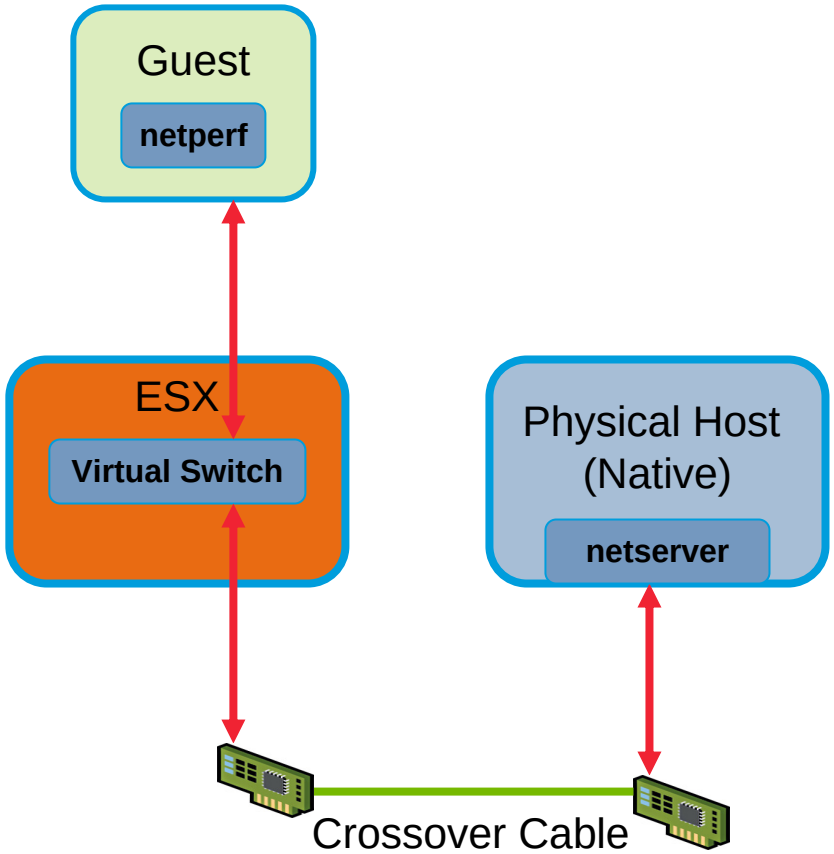


- **Virtual I/O stack**
  - Virtual Switch, NIC Teaming, Vlan Tagging, etc.
- **Sources of overheads**
  - Guest to vmkernel transitions
    - Address space switch
  - Virtual Interrupts
  - Virtual I/O stack
    - Packet copy
    - Packet routing

Performance Methodology

- **Benchmark**
  - Netperf 2.4.2
  - Parameters
    - Socket size: 64KB
    - Message size: 512B - 32KB
- **Performance Metrics**
  - Throughput
  - Request-Response per second
    - Single connection test
    - Inversely proportional to latency
- **Network I/O Paths**
  - Transmit(Tx), Receive(Rx), VM-VM

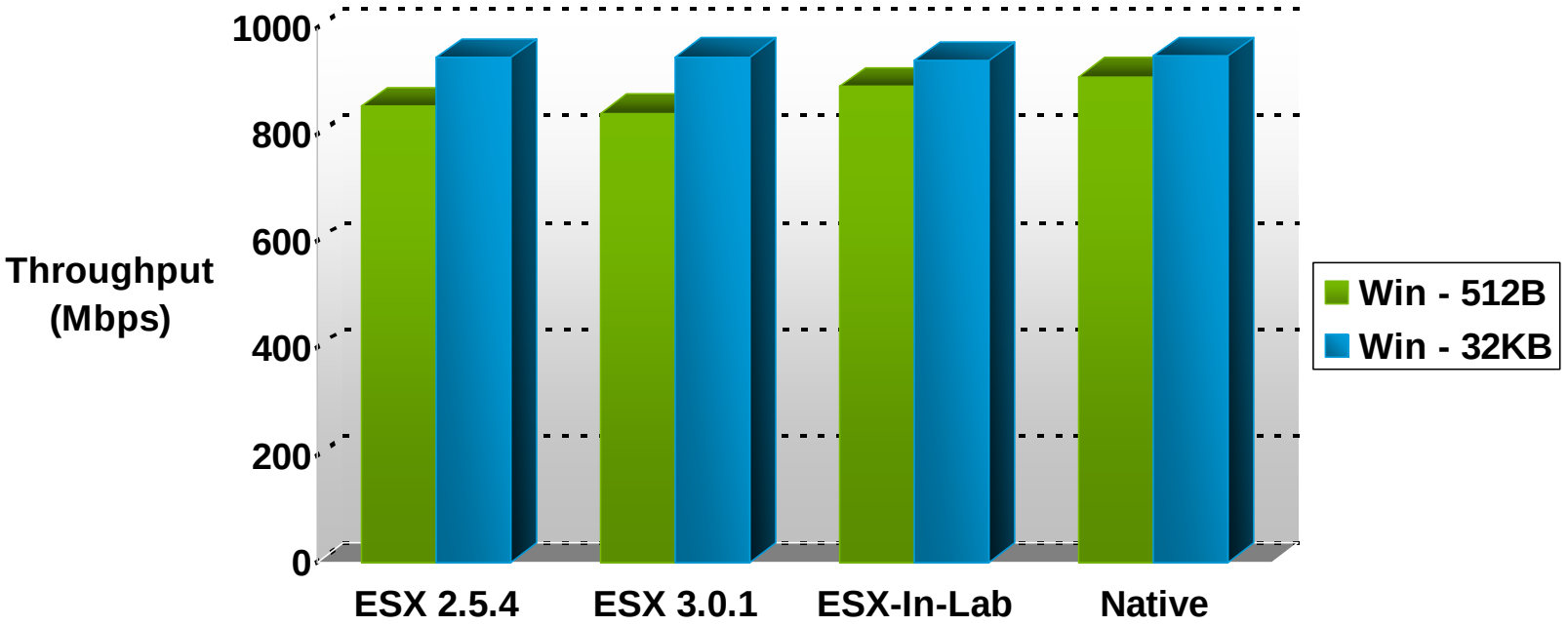
Tx Performance - Experimental Setup



- **Guest**
  - OS Version: Win2003/RHEL4 32-bit
  - 1 Virtual CPU, 512 MB memory
  - Virtual device - vmxnet
- **ESX**
  - ESX Version: 2.5.4, 3.0.1
  - H/W: Intel Xeon 5150 Quad Core, 2 socket, 2.66 GHz. 8GB RAM
  - NIC: Intel e1000, 1Gbps
- **Physical Host**
  - OS Version: Win2003 32-bit
  - H/W: Intel Xeon 5150 Quad Core, 2 socket, 2.66 GHz. 8GB RAM
  - NIC: Intel e1000, 1Gbps

Tx Performance - TCP Throughput

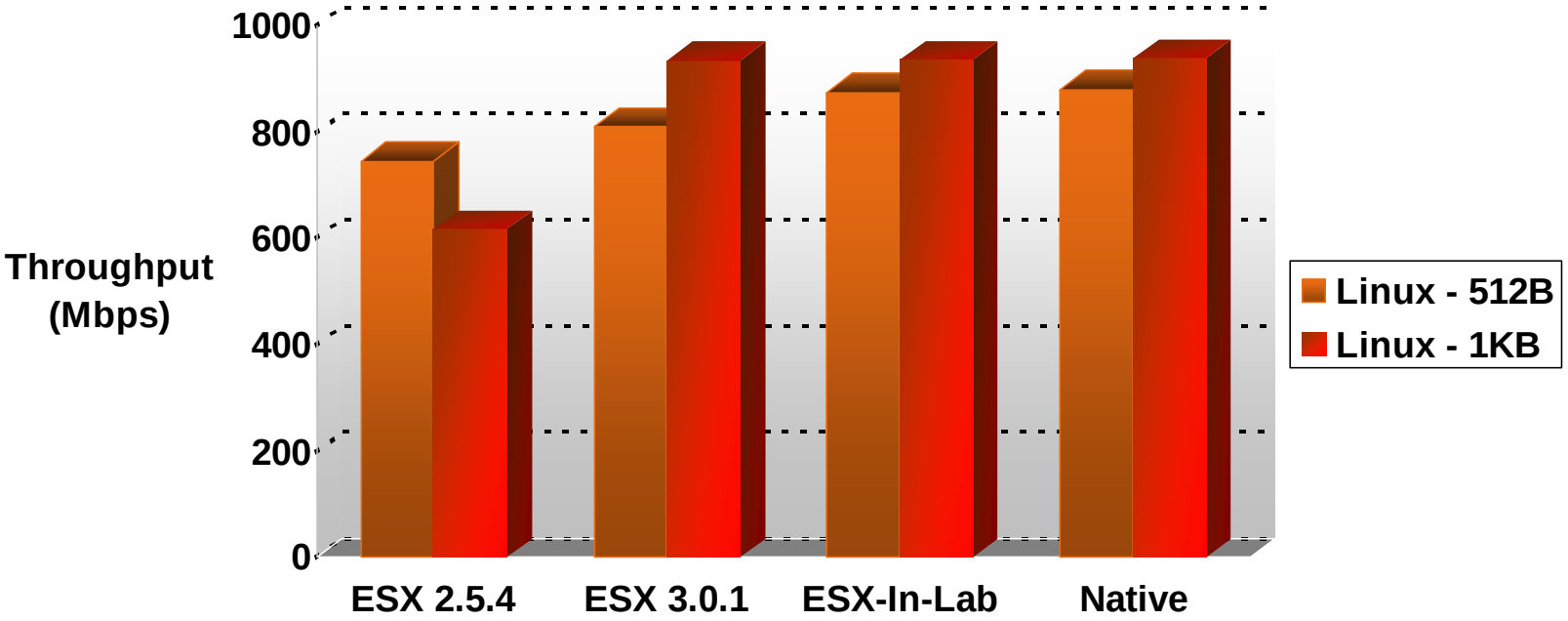
- Near native throughput
- Linux guest throughput is also close to native



H/W: Intel Xeon 5150 Quad Core @ 2.66 GHz, Intel e1000 NIC

Tx Performance - UDP Throughput

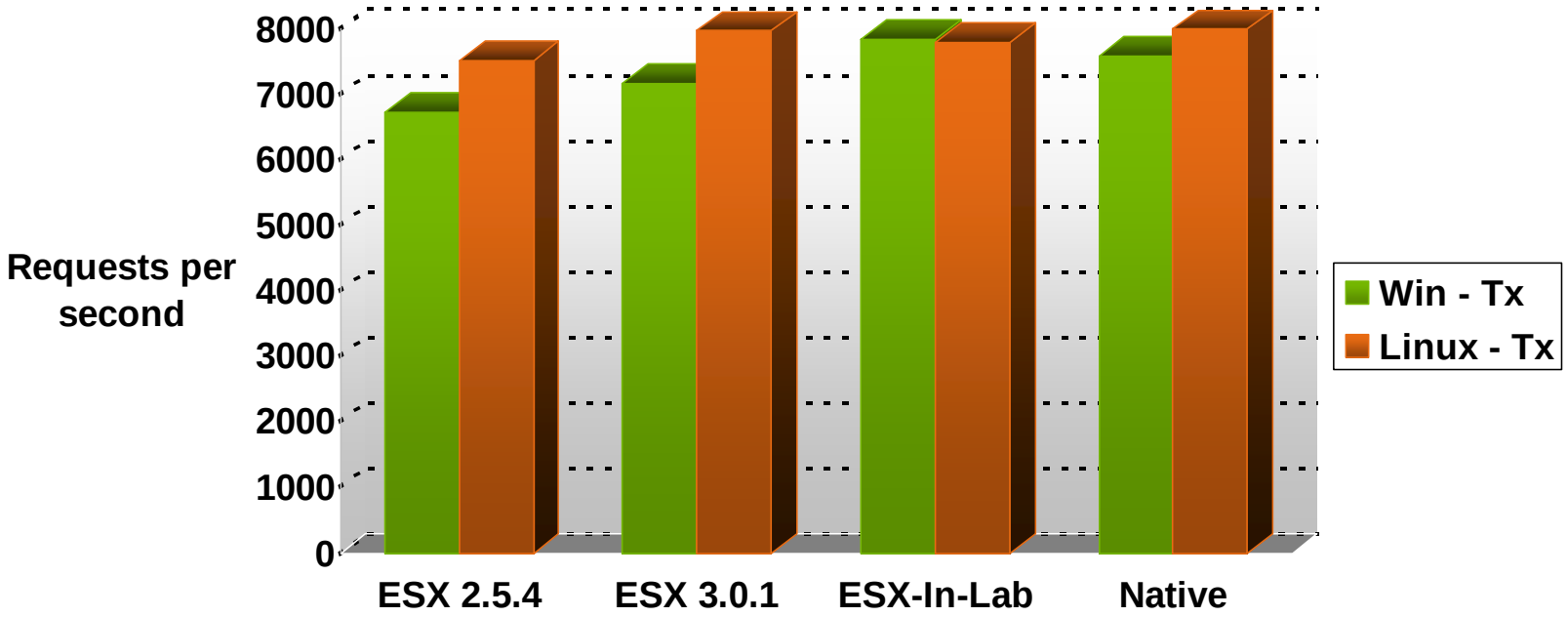
- Up to 40% improvement since ESX 2.5
- Throughput close to native for Windows as well



H/W: Intel Xeon 5150 Quad Core @ 2.66 GHz, Intel e1000 NIC

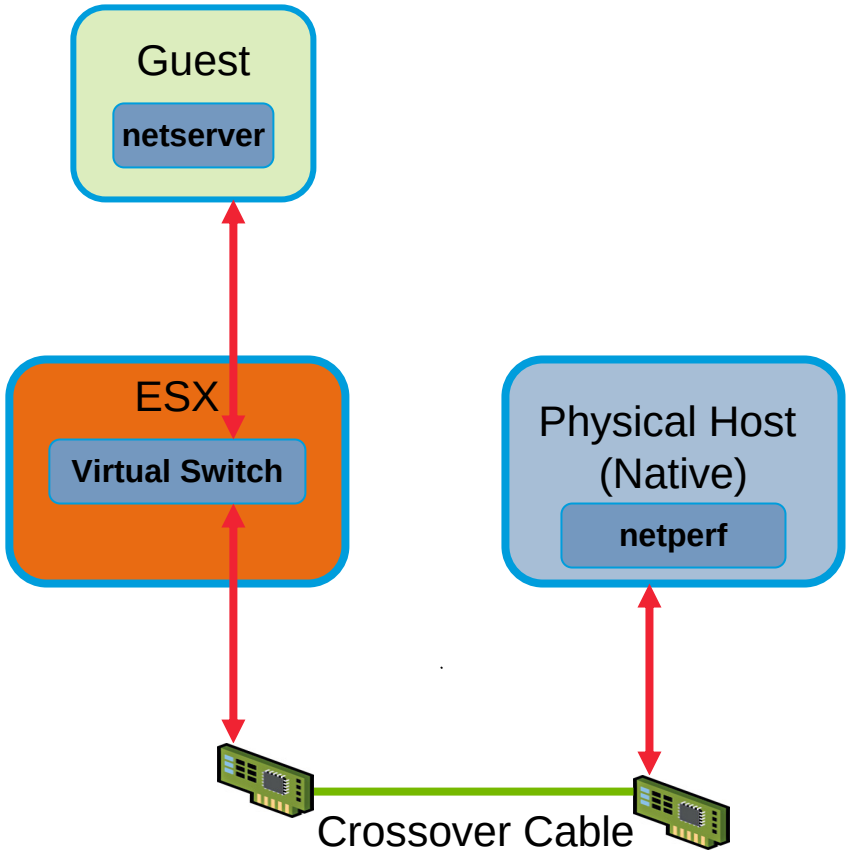
Tx Performance - Latency

- Upto 14% improvement since ESX 2.5
- Rx path: about the same, VM-VM path: 2x better



H/W: Intel Xeon 5150 Quad Core @ 2.66 GHz, Intel e1000 NIC

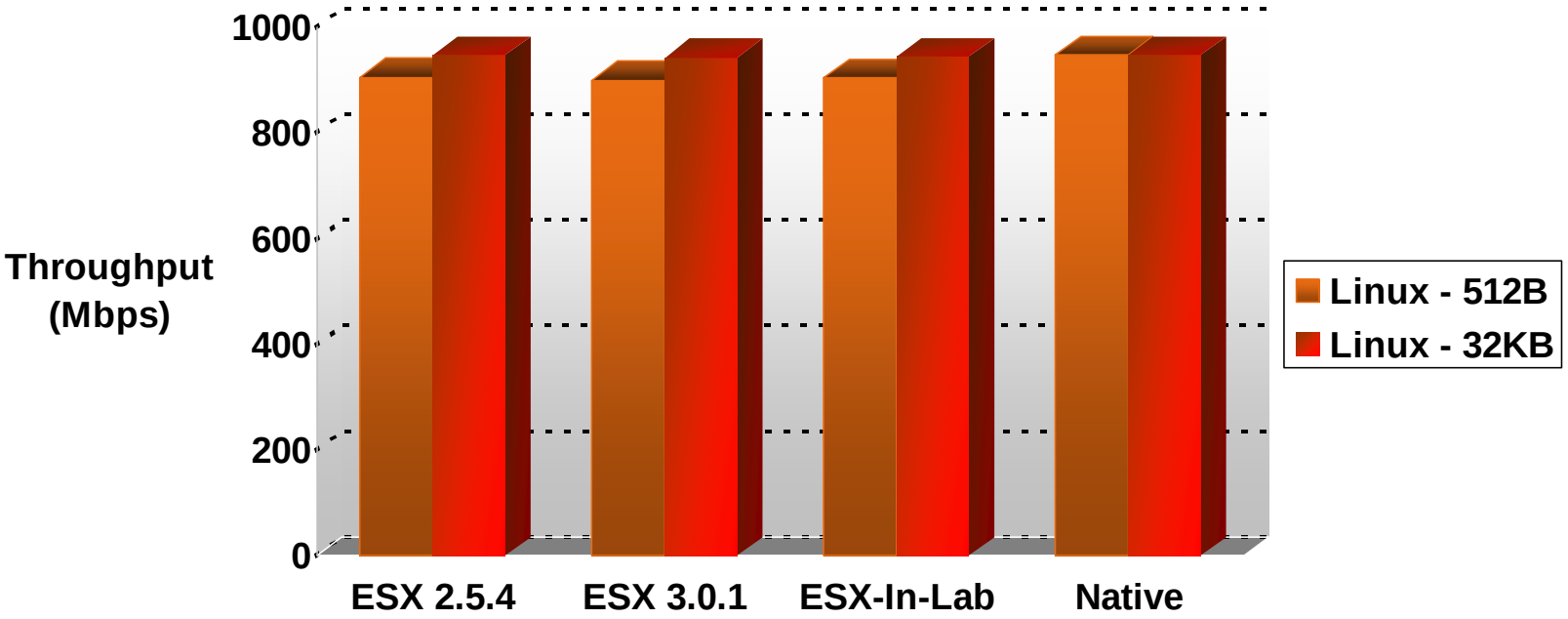
Rx Performance - Experimental Setup



- **Guest**
  - OS Version: Win2003/RHEL4 32-bit
  - 1 Virtual CPU, 512 MB memory
  - Virtual device - vmxnet
- **ESX**
  - ESX Version: 2.5.4, 3.0.1
  - H/W: Intel Xeon 5150 Quad Core, 2 socket, 2.66 GHz. 8GB RAM
  - NIC: Intel e1000, 1Gbps
- **Physical Host**
  - OS Version: Win2003 32-bit
  - H/W: Intel Xeon 5150 Quad Core, 2 socket, 2.66 GHz. 8GB RAM
  - NIC: Intel e1000, 1Gbps

Rx Performance - TCP Throughput

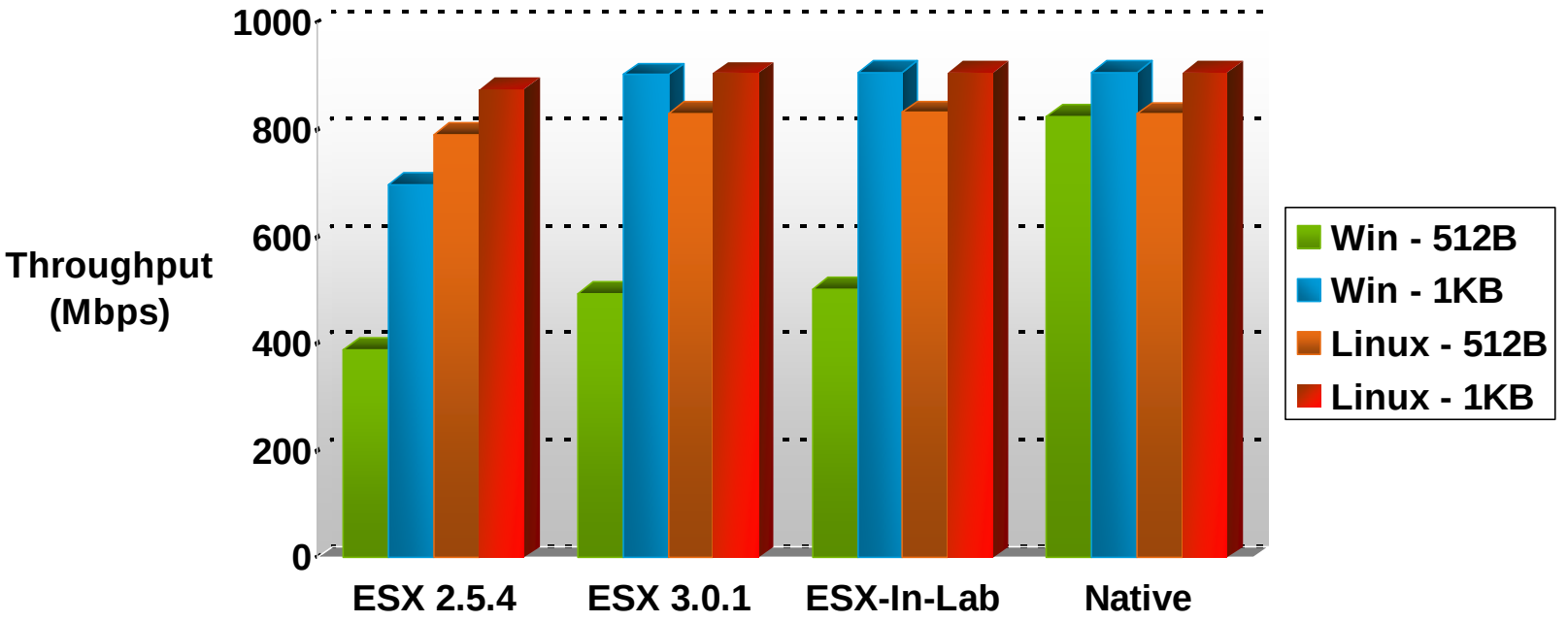
- Close to native throughput
- Throughput is about the same for Windows guest



H/W: Intel Xeon 5150 Quad Core @ 2.66 GHz, Intel e1000 NIC

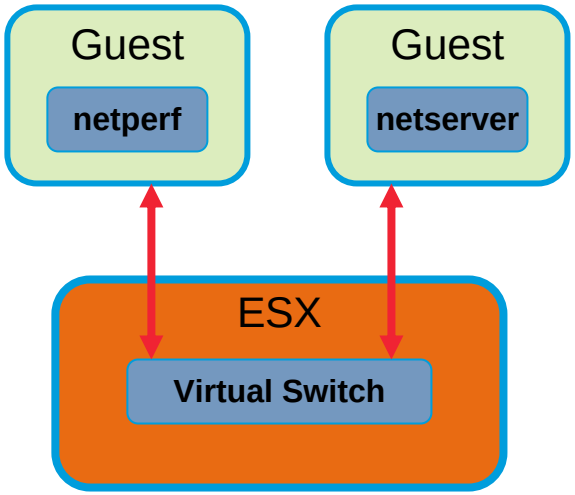
Rx Performance - UDP Throughput

- Up to 22% improvement since ESX 2.5
- Win - 512B is 50% of native due to non-connected UDP



H/W: Intel Xeon 5150 Quad Core @ 2.66 GHz, Intel e1000 NIC

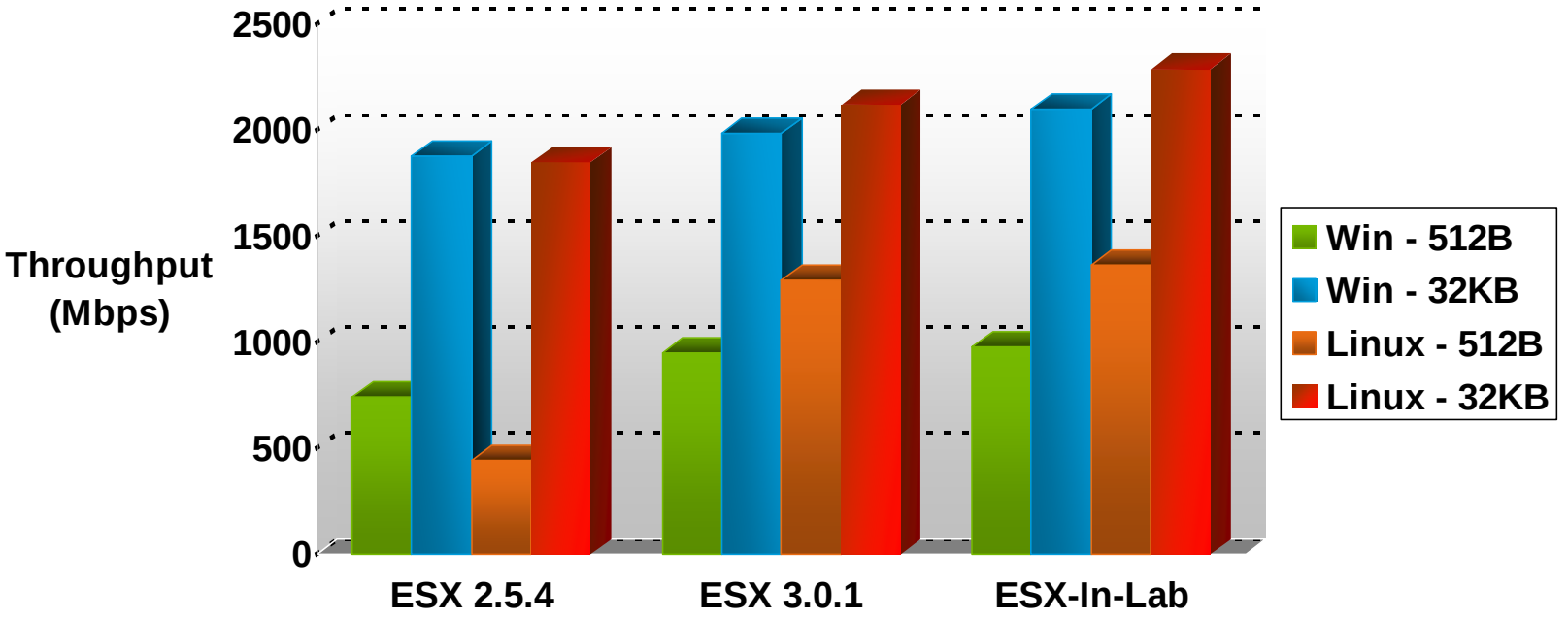
VM-VM Performance: Experimental setup



- **Guest**
  - OS Version: Win2003/RHEL4 32-bit
  - 1 Virtual CPU, 512 MB memory
  - Virtual device - vmxnet
- **ESX**
  - ESX Version: 2.5.4, 3.0.1
  - CPU: Intel Xeon 5150 Quad Core, 2 socket, 2.66 GHz
  - Memory - 8GB

VM-VM Performance - TCP Throughput

- More than 100% improvement in some cases
- 2Gbps+ throughput



H/W: Intel Xeon 5150 Quad Core @ 2.66 GHz, Intel e1000 NIC

Performance Optimizations in ESX 3.0

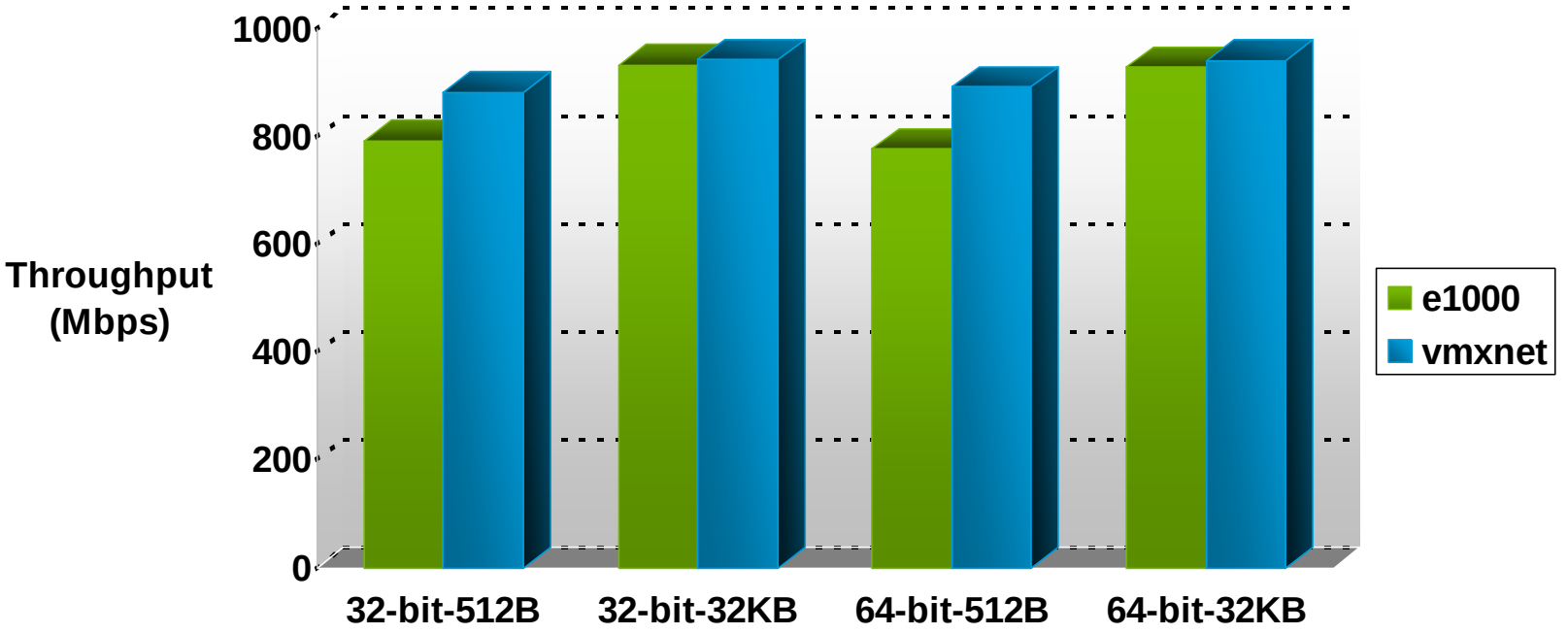
- **Virtual Interrupt coalescing**
  - Reduces interrupt processing overhead in guest
- **Tx Coalescing**
  - Reduces guest to vmkernel transitions
- **Tx zerocopy**
  - Do not copy the packet from guest memory to vmkernel memory
  - Translate the physical page numbers into machine page numbers
  - Cache the translated addresses
- **Other performance enhancements too …**

Virtual Network Devices

- **Supported virtual network devices : vlance, vmxnet**
  - Vlance provides good out of box experience
  - Vmxnet is the high performance virtual device
- **In ESX 3.0.1, e1000 virtual device was introduced**
  - Default and only supported device in 64bit guests
  - Provides good out of box experience and performance
  - Why do we need vmxnet ?
    - In some cases, vmxnet is better than e1000
    - Vmxnet is virtualization aware device

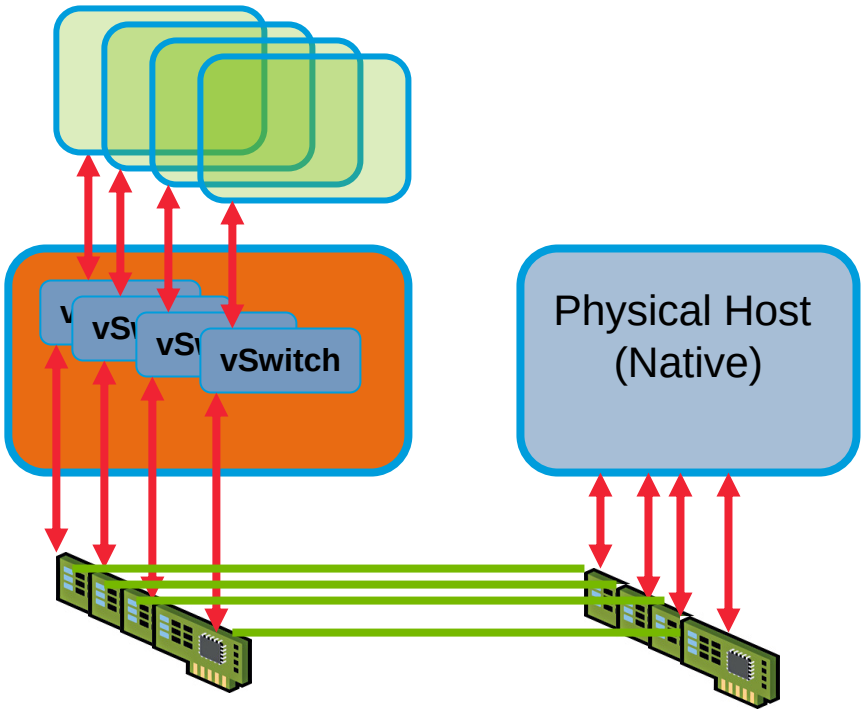
Virtual Network Devices (Win2003 TCP Tx)

- E1000 and vmxnet both can achieve line rate
- Vmxnet is better in Hw-assist environments



H/W: Intel Xeon 5150 Quad Core @ 2.66 GHz, Intel e1000 NIC, VT enabled for 64-bit guest, **S/W:** ESX in lab

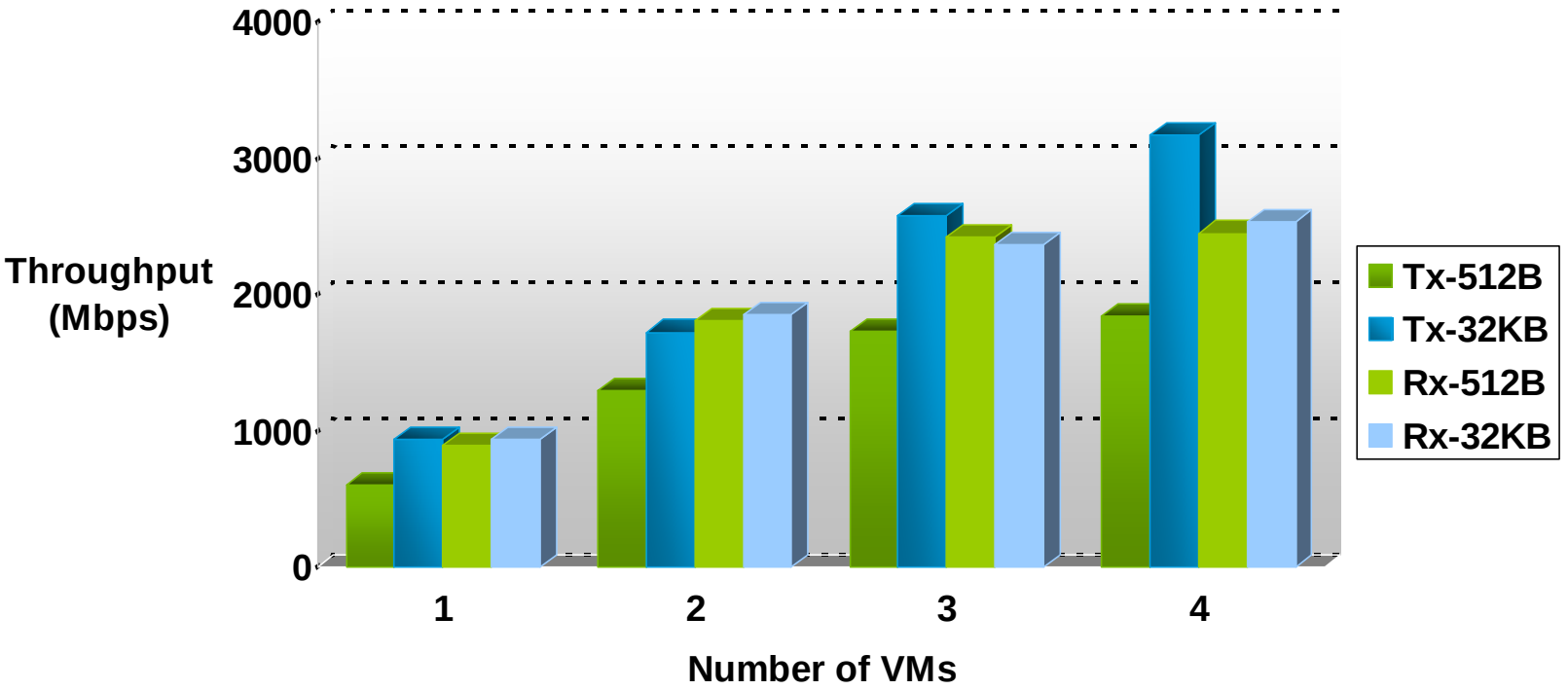
Performance Scalability – Experimental Setup



- **Guest**
  - OS Version: Win2003/RHEL4 32-bit
  - 1 VCPU, 512 MB memory
  - Virtual device: vmxnet
- **ESX**
  - ESX Version: 2.5.4, 3.0.1
  - H/W: Opteron 270 dual core, 2 socket @ 1.99 GHz. 4GB RAM
  - NICs: 2 dual port Intel e1000, 1Gbps
- **Physical Host**
  - OS Version: Win2003 32-bit
  - H/W: Opteron 270 dual core, 2 socket @ 1.99 GHz. 4GB RAM.
  - NICs: 2 dual port Intel e1000, 1Gbps

Performance Scalability – WS2003 TCP Tx & Rx

- Linear scaling until CPU is saturated
- **With lower network usage per VM, can scale up to many more VMs**



H/W:AMD Opteron 270, dual core, 2 socket, 1.99 GHz, Intel e1000 NICs **S/W:** ESX 3.0.1

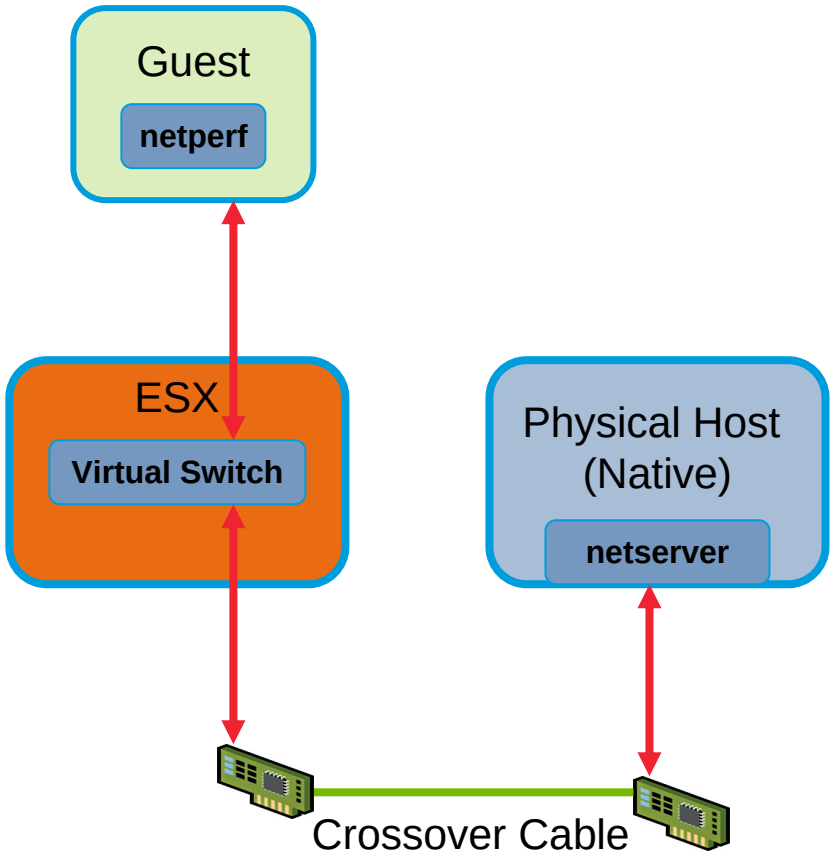
Agenda

- **ESX Architecture - Network I/O**
- **Benchmarking methodology and results**
  - Performance trends – ESX 2.5 vs. ESX 3.x vs. Native
  - Comparison of virtual network devices
  - Performance scalability
- **Future Directions**
- **Benchmarking Guidelines**

Future Directions

- **10 Gbps networks**
- **TCP Segmentation Offload (TSO)**
  - Offload segmentation of large TCP messages to the physical NIC
  - For TCP send traffic only
- **Jumbo Frames (JF)**
  - Use a large MTU for communication
    - Standard Ethernet MTU is 1500 bytes
    - Typical Jumbo Frame MTU = 9000 bytes
  - Benefits all IP protocols, on both send and receive paths
  - Limited to local networks, great for IP storage
- **Other performance enhancements too … but not discussed in detail**

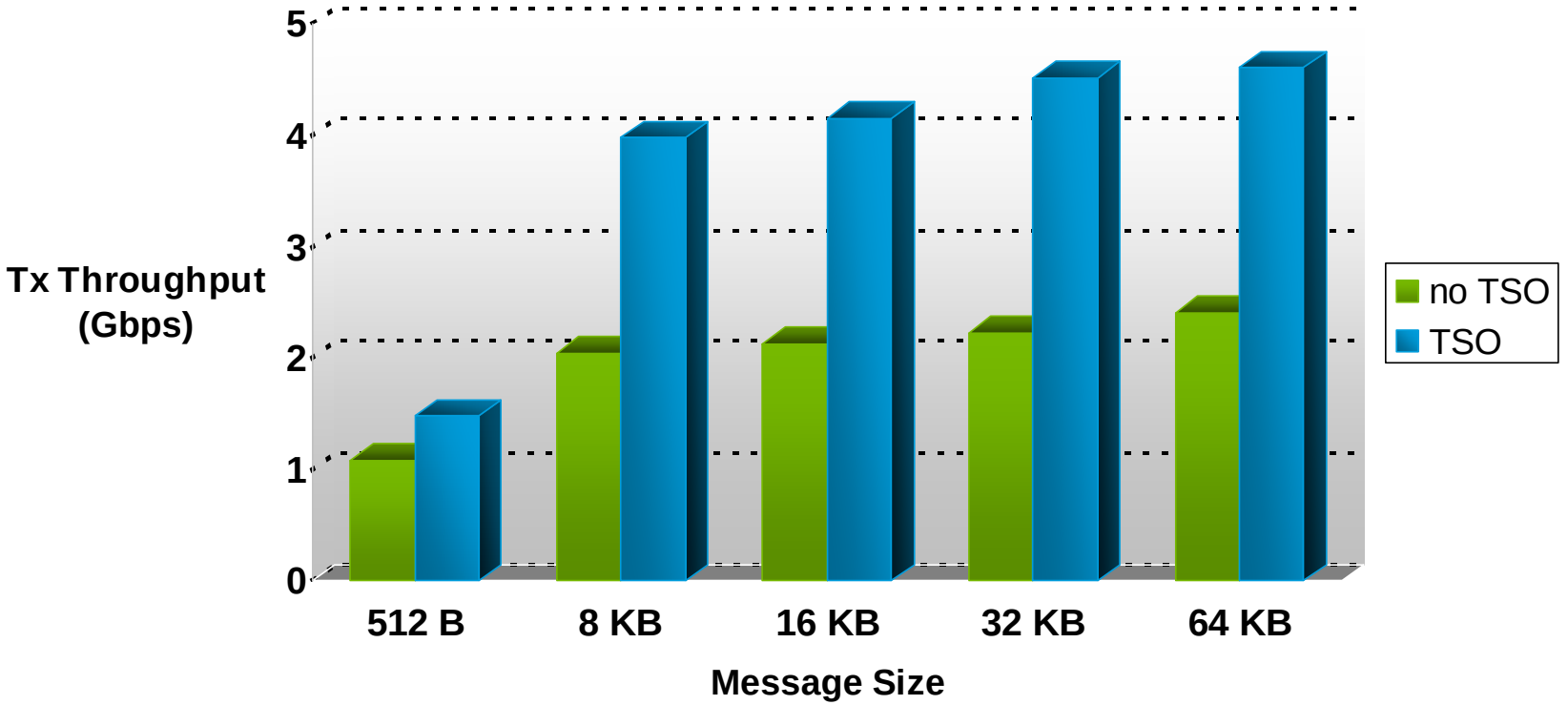
10 GigE Performance - Experimental Setup



- **Guest**
  - OS: Win2003 / RHEL5 64-bit
  - 1 Virtual CPU, 512 MB RAM
  - Virtual device - vmxnet
- **ESX**
  - ESX Version: experimental
  - H/W: Intel Xeon 5150 Dual Socket, Dual core @ 2.66 GHz. 8GB RAM
  - NIC: Neterion Xframe II 10 Gbps adapter
- **Physical Host**
  - OS Version: RHEL4 64-bit
  - H/W: Intel Xeon 5150 Dual Socket, Dual core @ 2.66 GHz. 4GB RAM.
  - NIC: Neterion Xframe II 10 Gbps adapter

TSO impact: 10GigE testbed

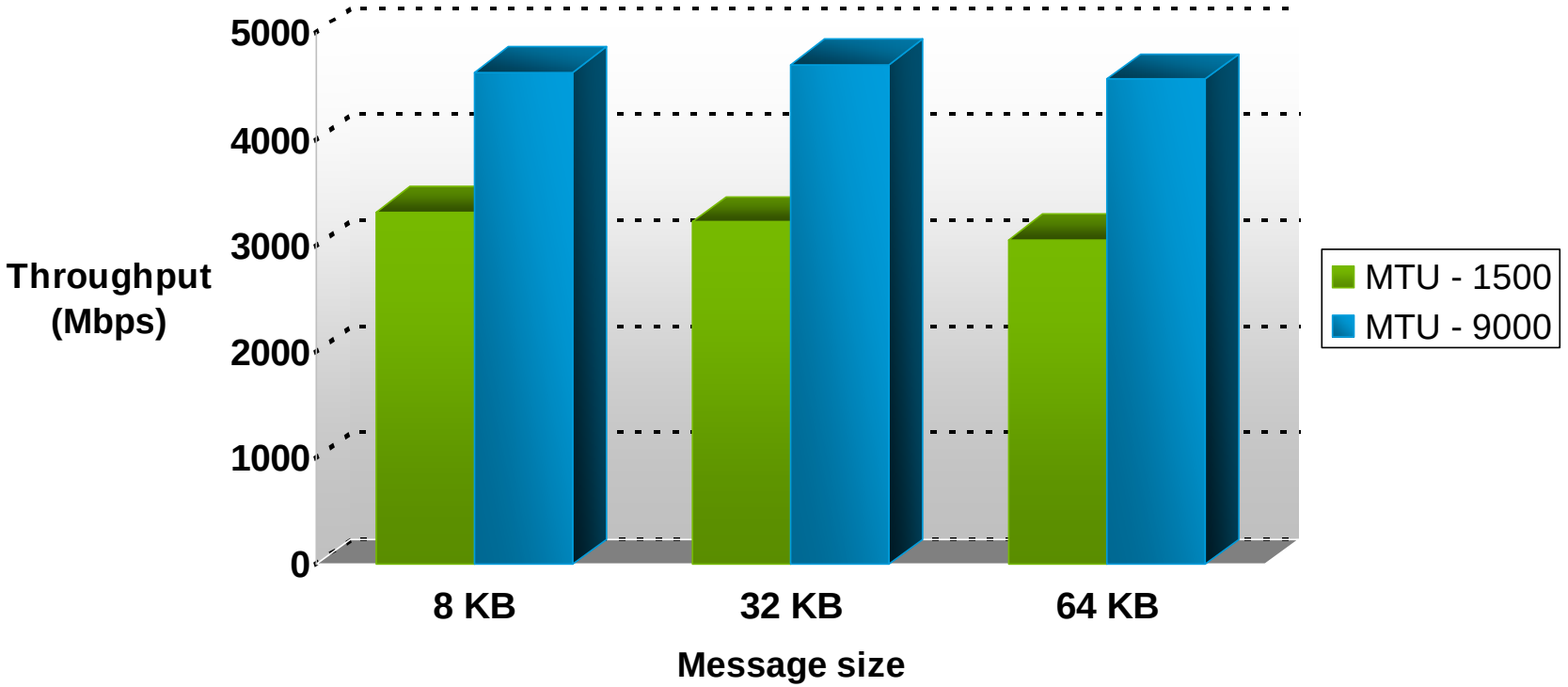
- Up to 2x improvement in throughput across message sizes
- Similar gains in Linux – up to 1.7x



H/W: 4-way Xeon @ 3.0 GHz, Neterion PCI-X 10GigE NIC. **S/W**: Win 2003 64-bit. Netperf socket size = 64 KB.

Jumbo Frames: Linux TCP receive throughput

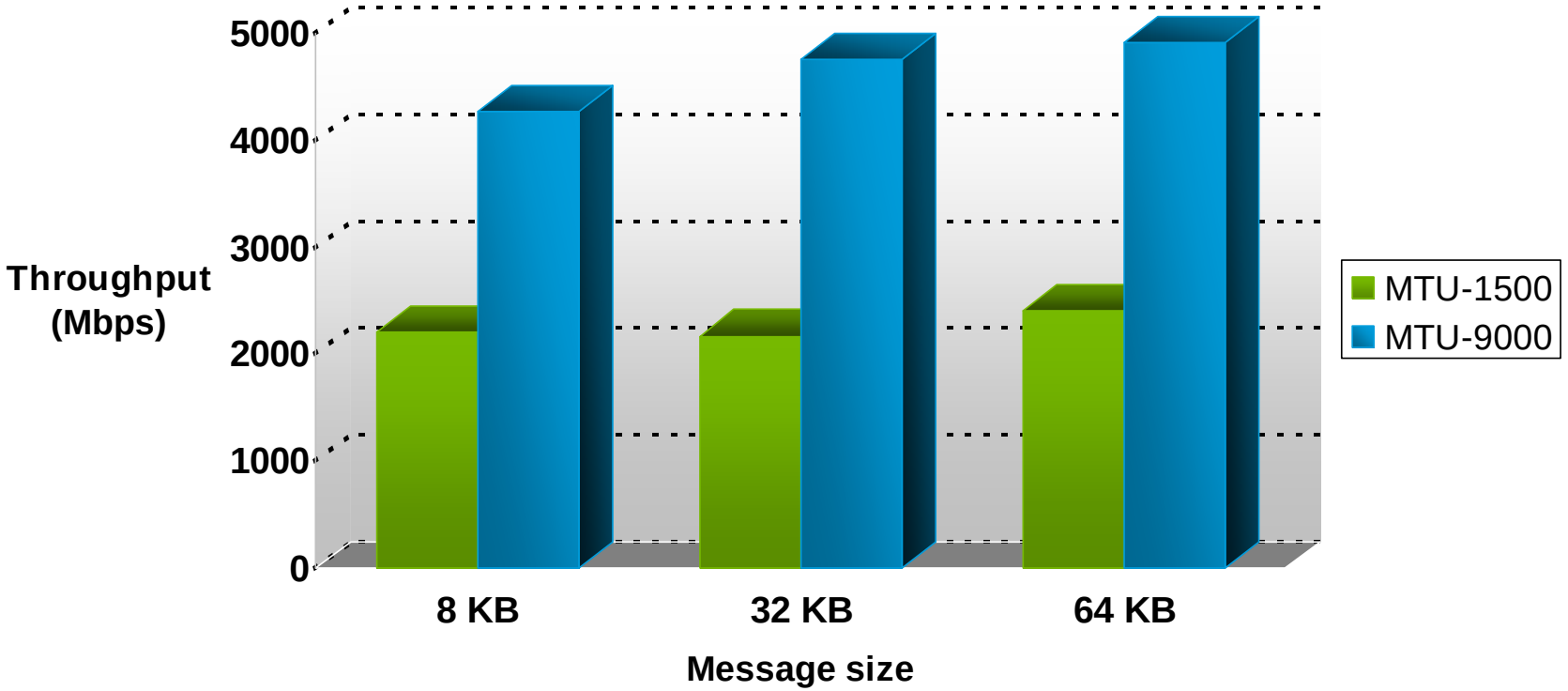
- Up to 50% improvement in Receive throughput



H/W: 4-way Xeon @ 3.0 GHz, Neterion PCI-X 10GigE NIC. **S/W**: RHEL 5 64-bit. Netperf socket size =128 KB

Jumbo Frames: Windows TCP send throughput

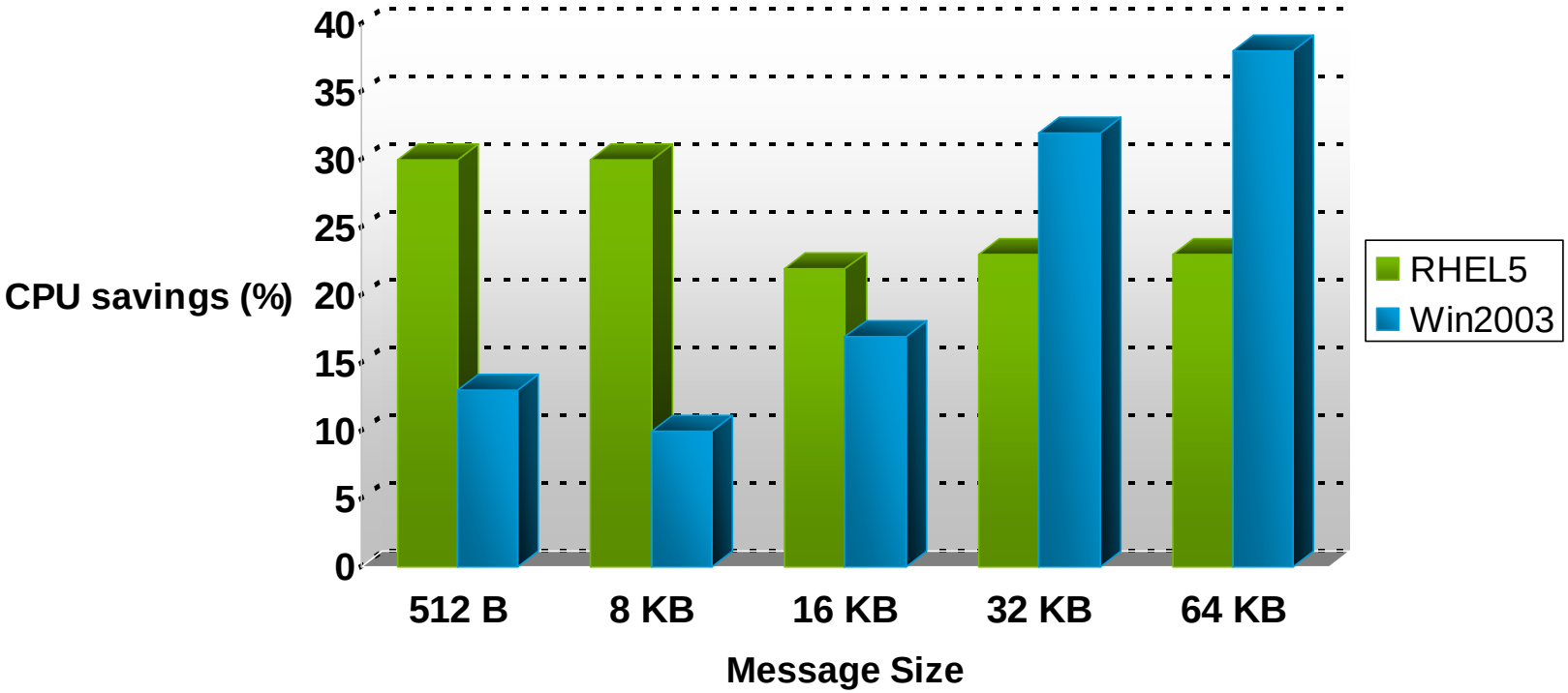
- 2x improvement in send throughput
- TSO was disabled in the guest OS



H/W: 4-way Xeon @ 3.0 GHz, Neterion PCI-X 10GigE NIC. **S/W**: Win 2003 64-bit. Netperf socket size = 300KB.

TSO impact: 1 Gbps testbed

- Line rates achieved even w/o TSO.
- 20% - 30% CPU savings by enabling TSO



H/W: 4-way Xeon @ 3.0 GHz, Intel 82571 NIC. **S/W**: Virtual device: vmxnet. Netperf socket size = 64 KB.

ESX-In-Lab Peak Performance

- **Near Line Rates on 10 Gbps testbed**

	Send	Receive
Peak Throughput	8.7 Gbps	9.2 Gbps

  - Peak Rx throughput: 9.2 Gbps
    - 3 VMs, Intel Oplin 10GigE NIC, Dual Socket Dual-Core Xeon @ 2.66 GHz
    - VM config: RedHat FC5 - 32-bit.
  - Peak Tx throughput: 8.7 Gbps
    - 4 VMs, Intel Oplin 10GigE NIC, Dual Socket Quad-Core Xeon @ 2.66 GHz.
    - VM Config: RedHat FC5 - 32-bit.

Benchmarking – Best Practices

- **Network Configuration**
  - Isolate testbed
  - Use dedicated link like a crossover cable for maximum throughput
  - Make sure that there are no throughput bottlenecks between the client and the VM
  - For VM-VM experiments, attach VMs to same vswitch
  - Ensure client machine can drive the desired throughput

Benchmarking – Best Practices

- **VM Configuration**
  - vmxnet performance superior in most cases
  - Use uniprocessor VMs for benchmarking single-threaded apps
  - Disable extra services and background jobs.
- **Others**
  - Multiple threads may be needed
    - Single netperf / iperf thread cannot saturate 10 Gbps link
  - Measure CPU usage using esxtop
    - Report aggregate CPU usage of the system - not that of the VMM world

Benchmarking – Pitfalls

- **Interrupt sharing (KB article # 1290)**
  - Console OS and VMkernel may share interrupt lines

cat /proc/vmware/interrupts

Vector PCPU0 PCPU1 PCPU2 PCPU3

0x79: 559 4707 6121 3428 <COS irq 16 (PCI level)>,VMK aic79xx

0x89: 39806 0 0 0 COS irq 19 (PCI level),VMK vmnic1

  - May cause performance variation
  - Remove modules from Console OS, disable devices, shuffle cards

cat /proc/interrupts

17: 2406 vmnix-level ehci-hcd, usb-uhci

18: 11760 vmnix-level usb-uhci

19: 995557 vmnix-level usb-uhci

Benchmarking – Pitfalls

- **Hardware Limitations**
  - PCI bus or Physical NIC may be the bottleneck
  - Run native experiments to get a baseline
- **Netperf behavior**
  - Ensure there are no version mismatches
  - Intractable errors when used across different OSes
  - Avoid cross-compilation

Conclusion

- **Network performance has improved over releases**
  - In most cases, we achieve near-native throughput.
- **Both vmxnet and e1000 can easily saturate 1 Gbps link**
  - In some cases, vmxnet is superior to e1000
- **Enabling TSO or JF cause up to 2X throughput increase**
  - Substantial CPU savings on 1 Gbps networks
- **In Lab, close to line speeds on 10 Gbps networks**

Questions?

TA40

ESX Networking Performance

Bhavjit Walha

VMware

Shilpi Agarwal

VMware

VMWORLD 2007

EMBRACING YOUR VIRTUAL WORLD

BREAKOUT SESSION