Virtualizing HPC on vSphere: Choosing a Best Method for RDMA HCA/NIC Deployment for MPI Workloads

vmware®

Introduction

The demand for HPC workloads is increasing as they play a critical role in scientific advancements across various fields. Researchers are satisfied as long as their application performance requirements are met, while staying within budget constraints. Simultaneously, HPC administrators and platform providers strive to modernize their infrastructure cost-effectively and expedite deployment, aiming to extend the use of their traditional HPC clusters for a broader user community and scenarios. This approach of improving utilization and making HPC easier to use has garnered widespread agreement among panels, committees, and audiences at the *SC'22* conference.

As part of this trend to modernize HPC infrastructure, virtualized HPC clusters have emerged as a natural solution. VMware, with its extensive experience in optimizing virtualization for CPU/GPU, memory, and network devices, has made significant strides in achieving near bare-metal performance. While compute resources are needed for software defined services like network management and hyper-converged storage (vSAN) services, virtualization offers administrators notable advantages in terms of flexibility and ease of use. It enables the transition to create from a virtualized HPC cluster to a private cloud (Software Defined Data Center, SDDC), and a multi-cloud platform for HPC [*Ref19*] and GenAI. VMware effectively tackles key challenges in bare metal HPC across five distinct categories.

Table 1: Benefits of virtualizing HPC compared to bare-metal HPC clusters.

	Bare metal	VMware virtualization
Performance	Optimal	Achieve near bare-metal performance [*Ref1*, *Ref2*, *Ref3*, *Ref4*]
Security	Require additional support	Intrinsic security tools (NSX & DFW) available for multi-tenancy [*Ref5*, *Ref6*]
Deployment	Long time and is not flexible	Aria automation (used in *HPC on Demand*)
Resource Utilization	Need reservations to prioritize workloads for a user Lack of load balancing No high availability for critical components	Resource isolation for multi-tenancy Resource pools vSphere DRS and HA for load balancing and critical components
Reproducibility	Checkpoint	Snapshot, Migration, Checkpoint

Tightly coupled HPC workloads, especially those using MPI, often face communication bottlenecks, posing challenges for HPC applications. Compared to other parallel and distributed programming systems, MPI offers the lowest minimum effective task granularity (METG), minimizing the overhead typically associated with such systems and enabling optimal performance [*Ref22*]. To address these communication challenges, hardware vendors have developed RDMA host channel adapters (HCA) and network interface cards (NIC) with improved bandwidth and reduced latency, aiming to enhance communication efficiency in these workloads.

VMware has been at the forefront of virtualizing these network devices to tackle the challenges in HPC effectively. VMware has published papers providing valuable insights into the performance considerations of RDMA transfer in vSphere, covering the setup and performance evaluation using SRIOV-IB [*Ref7*], SRIOV-RoCE[*Ref8*], Direct Path I/O with IB and ROCE concurrently [*Ref9*], and PVRDMA [*Ref10*]. This study addresses common customer inquiries regarding **performance differences among different VMware network adapter types** (VMXNET3, PVRDMA, SRIOV, DirectPath I/O) **for MPI workloads**, **performance disparities between IB and RoCE in vSphere**, **virtualization tax compared to bare metal**, and **recommended products** for running their HPC workloads.

To help clarify these common concerns, this study guides you through different methods for RDMA HCA/NIC usage in VMware stacks, highlighting the pros and cons of each approach. This empowers you to make informed decisions that best serve your HPC user community. We understand that vendor alignment may play a role in your choice, and based on our experiences with vSphere customer scenarios, we provide further advice on navigating this decision-making process.

Audience

This study is intended for traditional HPC administrators and decision-makers who are familiar with VMware vSphere, vSAN, and VMware vCenter Server. It assumes a certain level of understanding of HPC interconnect networking. Additionally, customers interested in distributed training using MPI in vSphere for GenAI will find relevant information within this paper.

Choosing adapter types for an RDMA HCA/NICs in vSphere

The choice of adapter type depends on the specific requirements of the workload and the desired balance between virtualization features and performance.

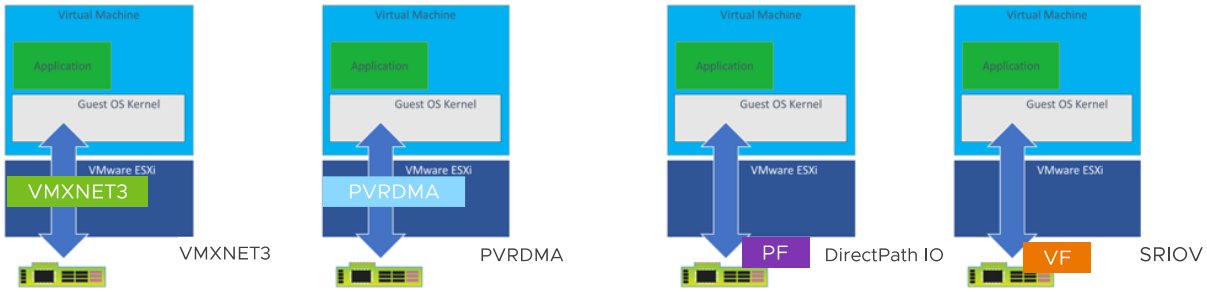


Figure 1. Adapter Type in VMs for an RDMA HCA/NIC

VMXNET3

VMXNET Generation 3 (VMXNET3) is a paravirtual (hypervisor aware) NIC type that offers broad compatibility and is widely supported across VMware products such as vSphere and Tanzu. It provides enhanced performance by eliminating the need for emulation of the corresponding physical network device (NIC card). Traditional network drivers rely on emulation, which can lead to performance issues, especially for VMs with high traffic loads like SQL or web servers. VMXNET3, on the other hand, operates directly with the hypervisor, resulting in a performance boost of approximately 20% and reduced CPU consumption [*Ref18*]. It's important to note that VMXNET3 is a virtual network device and is installed and enabled automatically together with VMware Tools.

For more information, please refer to the following *link*.

PVRDMA

PVRDMA is also a paravirtual NIC type that enables RDMA over Ethernet (RoCE) between virtual machines. It supports features like vSphere vMotion and snapshots and is compatible with VMs using hardware version 13 and Linux kernel 4.6+. With PVRDMA, multiple VMs can access the RDMA device using the industry standard OFED verbs API. It offers OS bypass, zero-copy, low latency, high bandwidth, reduced power consumption, and faster data access. To utilize PVRDMA, all VMs must attach a PVRDMA device and be connected to a distributed switch. For detailed instructions on setting up PVRDMA, please refer to the following [*Ref10*].

DirectPath IO and Dynamic DirectPath IO

DirectPath I/O and Dynamic DirectPath I/O are both passthrough options for RDMA HCA/NICs in vSphere 7 and 8, ideal for communication-intensive workloads like HPC simulations and scientific applications. They offer minimal overhead and enable dedicated utilization of RDMA HCA/NICs without the need for sharing on the same HCA/NIC device. They can also serve as an initial testing environment for teams exploring virtualized HPC.

DirectPath I/O assigns a specific physical device on an ESXi host using the Segment/Bus/Device/Function format, limiting the VM's access to that specific host. However, it has limitations, including exclusive device access for a single VM and the unavailability of features like high availability (HA), distributed resource scheduler (DRS), live migration (vMotion) , snapshot, suspension, and resume.

In contrast, Dynamic DirectPath I/O, first introduced in vSphere 7, leverages the *assignable hardware framework*, which allows for the flexible assignment of PCI devices to VMs using a key-value method with custom or vendor-generated labels. It decouples the static relationship between VMs and devices, enabling passthrough devices to work with DRS "*initial placement*" and subsequent vSphere HA[*Ref20*, *Ref21*]. But it still lacks support for vMotion, snapshot, suspension, and resume to support real-time load balancing in the current release.

The absence of vMotion, snapshot, suspension, and resume support for assigned RDMA HCA/NIC devices in both passthrough options is due to dependencies on the vmx/vmkernel device framework and associated driver software. Only Nvidia vGPUs, implemented as "Mediated" passthrough, currently support vMotion, while other passthrough devices do not. Future advancements, such as the *DVX* framework in vSphere 8, aim to address these limitations in collaboration with hardware vendors.

For detailed steps on setting up DirectPath I/O and the performance evaluation for IB and RoCE, refer to the article *InfiniBand and RoCE DirectPath IO Setup and Performance Study on vSphere 7.x*. Additional information can be found in the following links: *vSphere*

VMDirectPath I/O and Dynamic DirectPath I/O: Requirements for Platforms and Devices and *vSphere ML Accelerator Spectrum Deep Dive – Using Dynamic DirectPath IO (Passthrough) with VMs*.

SRIOV passthrough

SR-IOV passthrough is supported in ESXi 6.0 and later. Similar to (Dynamic) DirectPath I/O, SR-IOV provides performance benefits and tradeoffs. It is particularly advantageous in workloads with high packet rates or low latency requirements in the HPC area by exchanging data without using the VMkernel as an intermediary. However, like DirectPath I/O, SR-IOV is currently not compatible with certain core virtualization features such as vMotion/snapshot/resume. On the other hand, SR-IOV allows for the sharing of a single physical device among multiple guest VMs, thereby increasing utilization. It's important to note that to fully utilize the features of SR-IOV, it requires the use of port groups on either VMware Distributed Switch (*VDS*) or Standard Virtual Switch (*SVS*). These switches enable functionality such as VLAN tagging for proper operation with SR-IOV. Besides, SRIOV is now part of the assignable hardware framework in vSphere 8U2, supporting DRS initial placement like Dynamic DirectPath IO.

VMware published three papers in 2022 that discussed the setup of SRIOV-IB [*Ref7*] and SRIOV-RoCE[*Ref8*] and performance evaluation on vSphere.

For network adapter SRIOV compatibility considerations, see the *VMware Compatibility Guide*.

Summary

The VMXNET3 and PVRDMA adapters utilize paravirtual transfer methods, providing support for various virtualization features such as vMotion and snapshots, but with some performance overhead. On the other hand, (Dynamic) DirectPath IO and SR-IOV are passthrough methods that offer near bare-metal performance but lack certain virtualization benefits like vMotion.

Table 2: Summary of RDMA HCA/NIC Adapter Types in VMs.

Category	Paravirtual		Passthrough		
Network adapter type	VMXNET3	PVRDMA	DirectPath IO	Dynamic DirectPath IO	SRIOV
MPI performance	Normal latency & bandwidth	Better	Best	Best	Nearly the same as DirectPath IO
vMotion/ Snapshot/Suspend & Resume supported?	Yes	Yes	No	Currently No. (DVX can help in the future)	Currently No. (DVX can help in the future)
DRS/HA supported?	Yes	Yes	No	Yes* (DRS initial placement in vSphere 7)	Yes* (DRS initial placement in vSphere 8U2)
VDS/SVS required?	Yes	Yes	No	No	Yes
Multi-VMs sharing?	Yes	Yes	No. Exclusive to 1 VM	Yes. But exclusive to 1 VM once powered on.	Yes
RDMA capability	No. TCP only	RoCE only	IB or RoCE	IB or RoCE	IB or RoCE

Enhanced VMDirectPath I/O (*DVX*) introduced in vSphere 8 builds upon Dynamic DirectPath I/O, introducing a new framework and API for hardware-backed virtual devices. It offers enhanced support for virtualization features, including live migration with vMotion, suspending and resuming VMs, and disk and memory snapshots. DVX requires hardware compatibility with the framework.

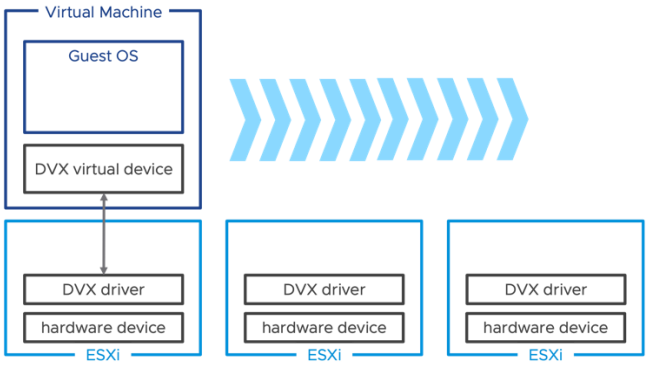


Figure 2. Enhanced VMDirectPath I/O (aka Device Virtualization Extensions or DVX)

Performance Evaluation

In virtualized infrastructure, HPC performance depends on hardware virtualization support and the capability of the virtual infrastructure. VMware vSphere and advancements in hardware virtualization have allowed applications running on a single VM to approach native speed. However, MPI applications, which require intensive communication between nodes, are more challenging and sensitive to interconnect performance. VMware has made efforts to optimize network stack performance, resulting in low overhead and near-bare metal performance for HPC workloads.

To evaluate MPI performance, the OSU microbenchmark is commonly used to measure p2p bandwidth and latency. Real-world HPC applications are also run to assess the performance efficiency of virtualized HPC clusters. These tests stress the CPU, networking, and storage resources of the cluster.

Table 3 presents the HPC applications and benchmark input datasets used in real-world application testing which are also used in [*Ref2, Ref7*, *Ref8*, *Ref9*]. Each experiment was performed five times, and the average throughput across the runs was measured. The performance on a single bare metal node serves as the baseline for relative performance comparisons.

Table 3: Workload and Benchmark Details

Application	Vertical Domain	Benchmark Dataset	Version
OSU Benchmark	Base RDMA performance testing	N/A	5.8
GROMACS	Life Sciences – Molecular Dynamics	HECBioSim 12M atoms	2020.5
Nanoscale Molecular Dynamics (NAMD)	Life Sciences – Molecular Dynamics	*STMV* – 8M atoms	2.14
LAMMPS	Molecular Dynamics – Fluid and materials	*EAM Metallic Solid Benchmark 1M atoms*	20210310
Weather Research and Forecasting (WRF)	Weather and Environment	*Conus 2.5KM resolution*	3.9.1.1
OpenFOAM	Manufacturing – Computational Fluid Dynamics (CFD)	*Motorbike 20M cell mesh*	9

The evaluation consists of two environments provisioned by *vhpc-toolkit* [*Ref12*, *Ref13*]. The first subsection evaluates the virtual performance of three different VMware network adapter types (VMXNET3, PVRDMA, SRIOV) using the **ConnectX-4 LX 25GbE** NIC in the *Google Cloud VMware Engine (GCVE)* environment. This environment consists of Dell R640 compute nodes running CentOS 7 on vSphere 7U2. Each node is equipped with two Intel Xeon Gold 6240 CPUs, providing a total of 32 cores and 768GB of memory.

In the second and third subsections, we evaluate the virtual performance of the **ConnectX-5 VPI** adapter (supporting EDR IB 100Gb/s and 100GbE). The testing environment aligns with the environment described in [*Ref7*, *Ref8*, *Ref9*], and we extract the relevant data presented in those papers for our analysis. The second subsection focuses on quantifying the virtualization overhead of using SRIOV compared to bare metal and DirectPath IO. The third subsection investigates the performance difference between InfiniBand (IB) and RDMA over Converged Ethernet (ROCE) in a vSphere environment.

For virtualization purposes, we reserve 4 cores per host in both environments. For detailed hardware information, including specifications and configurations, please refer to the "Functionality evaluation" section of the three respective papers.

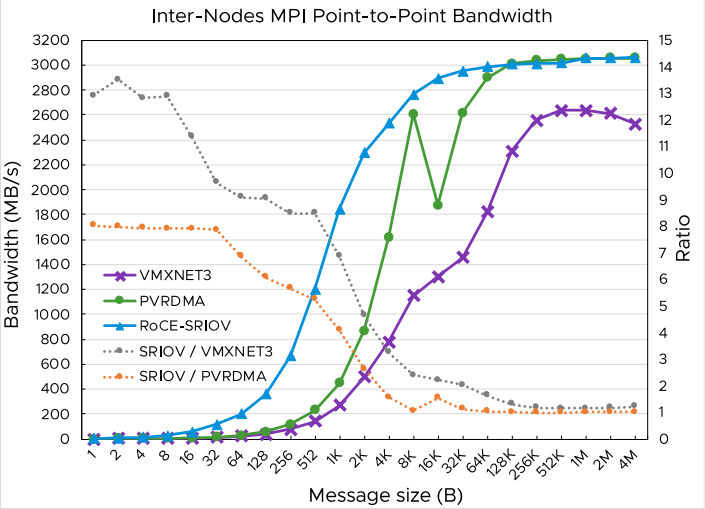
Performance Comparison between VMXNET3, PVRDMA, and SRIOV for MPI Workload

We initially performed tests using two OSU micro-benchmarks on a Mellanox ConnectX-4 25GbE Ethernet NIC in the GCVE environment. These tests focused on internode point-to-point bandwidth and latency measurements, specifically evaluating the performance between two processes running on separate VMs hosted on different hosts. In Figure 3, the solid lines represent the bandwidth and latency values for each message size, while the dotted lines depict the performance ratio of SRIOV vs VMXNET3 and SRIOV vs PVRDMA. The obtained results are as follows:

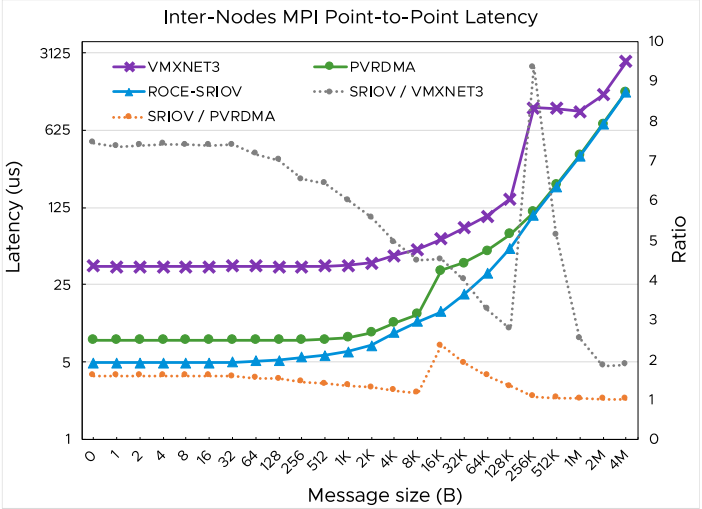
Figure 3a demonstrates that SRIOV achieves the highest bandwidth, followed by PVRDMA, and VMXNET3 with the lowest performance. For small message sizes ranging from 1B to 1KB, SRIOV outperforms VMXNET3 by a factor of 12.9X to 12.9X, and PVRDMA by 8.0X to 4.1X. After reaching message sizes of 64KB to 4MB, SRIOV and PVRDMA exhibit similar bandwidth performance, while VMXNET3 still experiences a 20% to 60% overhead.

Similarly, Figure 3b illustrates that SRIOV demonstrates the lowest latency, followed by PVRDMA, and VMXNET3. The latency of PVRDMA is on average 1.4 times that of SRIOV on all message sizes, while VMXNET3 experiences a latency 5.6 times higher than SRIOV.

Based on these results, it is evident that SRIOV outperforms PVRDMA and VMXNET3 in terms of bandwidth and latency for the pure MPI communication workload.



(a) osu_bw

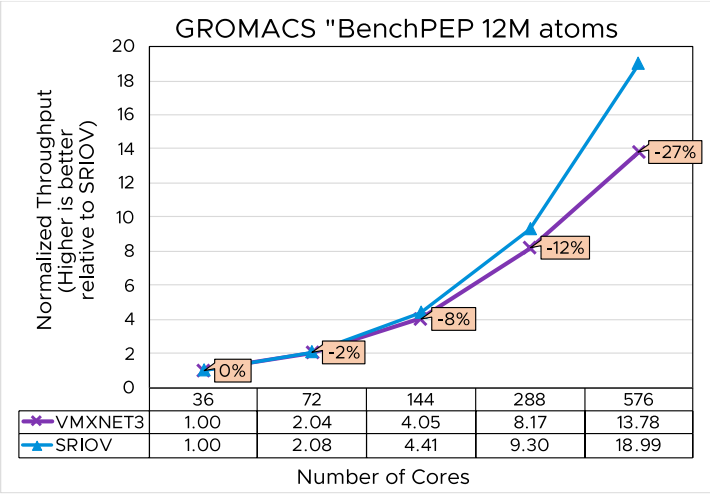


(b) osu_latency

Figure 3. Point-to-Point OSU micro-benchmark between 2 hosts using ConnectX-4 LX 25GbE NIC.

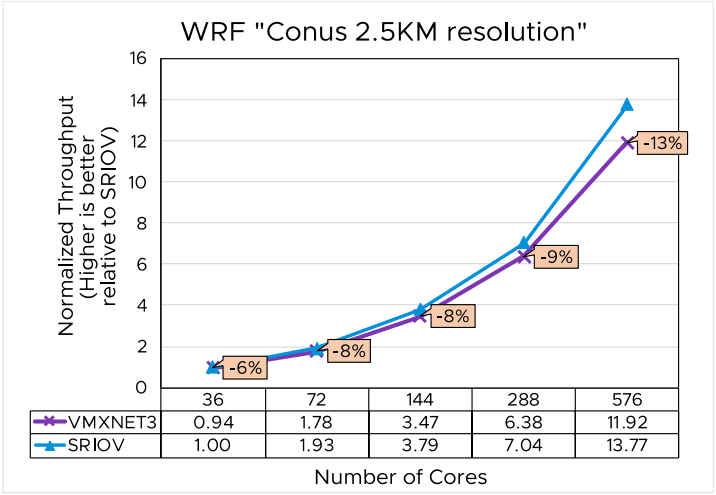
In our evaluation of real-world scenarios, we tested two HPC applications (GROMACS and WRF) using SRIOV and VMXNET3. On 576 cores (16 nodes), VMXNET3 showed up to 27% lower performance for GROMACS and 14% lower performance for WRF compared to SRIOV (Figure 4). However, it's important to consider that real-world applications involve computational work beyond communication, narrowing the time-to-solution gap between SRIOV and VMXNET3 observed in micro-benchmarks. Regarding PVRDMA, it may not be suitable for MPI applications highly sensitive to message rate. Therefore, we excluded PVRDMA from the results for those specific tests. However, we found favorable results with PVRDMA for HPC workloads when the number of processes per node is 20 or less [*Ref10, Ref23*]. We anticipate seeing improvements in future versions of vSphere to address these findings.

Based on these findings, we recommend customers prioritize SRIOV over VMXNET3 and PVRDMA for MPI workloads.



	36	72	144	288	576
VMXNET3	1.00	2.04	4.05	8.17	13.78
SRIOV	1.00	2.08	4.41	9.30	18.99

(a) GROMACS



	36	72	144	288	576
VMXNET3	0.94	1.78	3.47	6.38	11.92
SRIOV	1.00	1.93	3.79	7.04	13.77

(b) WRF

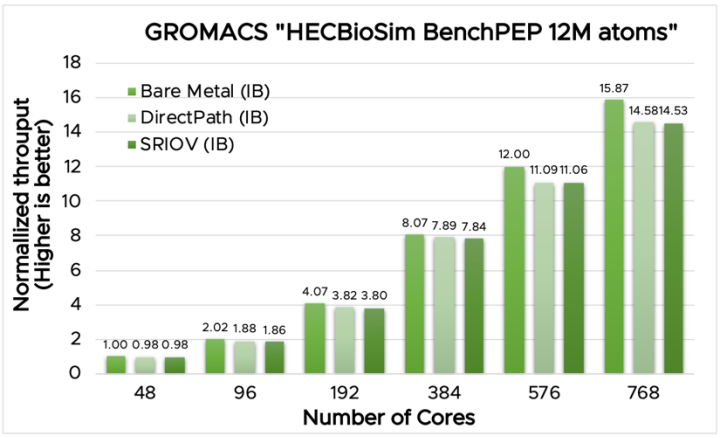
Figure 4. GROMACS and WRF Performance using SRIOV and VMXNET3 on ConnectX-4 LX 25GbE NIC.

Performance Comparison between SRIOV and DirectPath I/O for MPI Workload

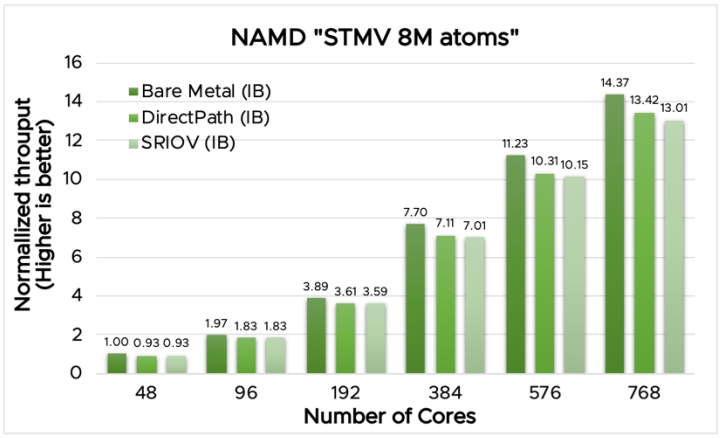
To assess the performance impact of SRIOV compared to bare metal and DirectPath IO, we conducted strong scaling tests on five real HPC workloads, varying the number of cores from 48 to 768 across 1 to 16 hosts using ConnectX-5 VPI 100Gbps HCA. We evaluated these workloads in both IB (Figure 5) and RoCE (Figure 6) configurations using SRIOV, DirectPath IO, and bare-metal setups.

We used the core count difference between bare metal (48 PPN) and the virtual environment (44 PPN) as the basis for an acceptable performance delta, representing an 8.3% decrease in cores for virtual. If the performance difference falls within this 8.3% threshold, it is deemed acceptable, considering the supplementary virtualization features provided by vSphere for cluster management, securities for multi-tenancy, and life cycle management.

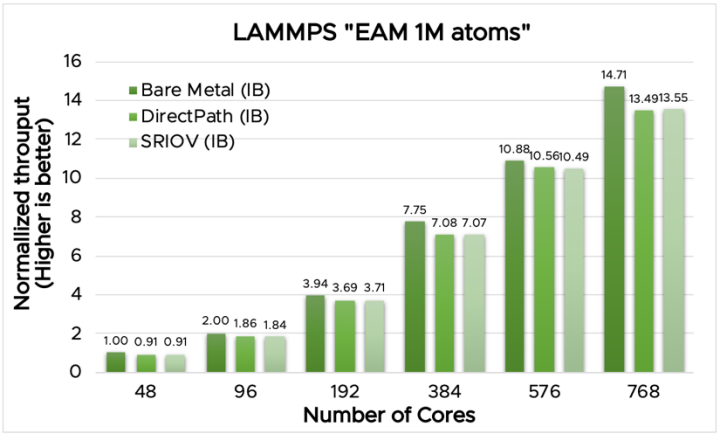
When utilizing InfiniBand, our results (Figure 5a~5e) show that at the largest scale of 768 cores (16 nodes), SRIOV exhibits a virtual tax of 8.4% on GROMACS, 9.4% on NAMD, 7.9% on LAMMPS, 4% on WRF, and a marginal improvement of 0.4% on OpenFOAM compared to bare metal. In contrast, DirectPath IO shows a performance impact of 8.1% on GROMACS, 6.6% on NAMD, 8.3% on LAMMPS, 4% on WRF, and 0.7% on OpenFOAM compared to bare metal. Except for NAMD in SRIOV-IB, which has an additional 1.1% overhead, all the other workloads fall within the acceptable performance delta of 8.3%.



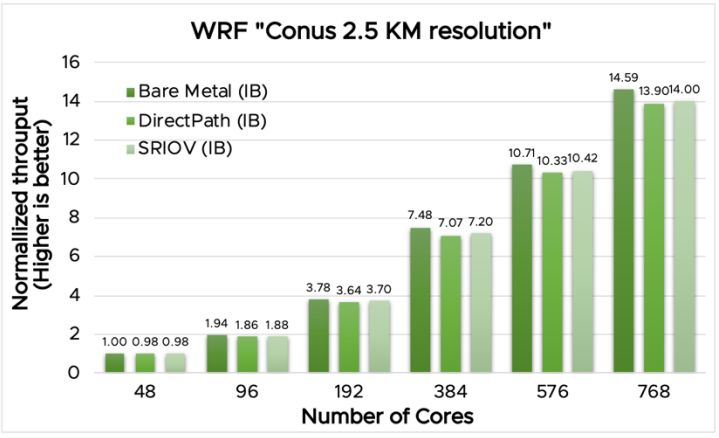
(a) GROMACS



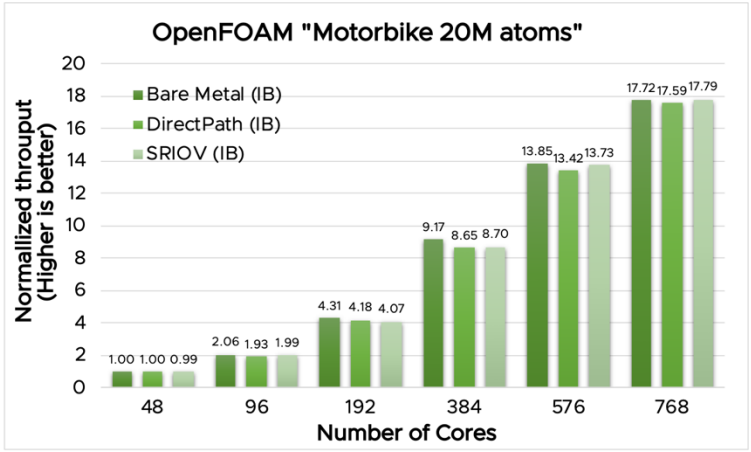
(b) NAMD



(c) LAMMPS



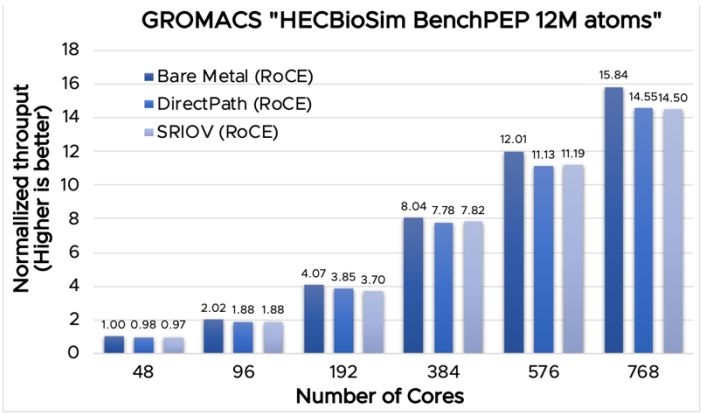
(d) WRF



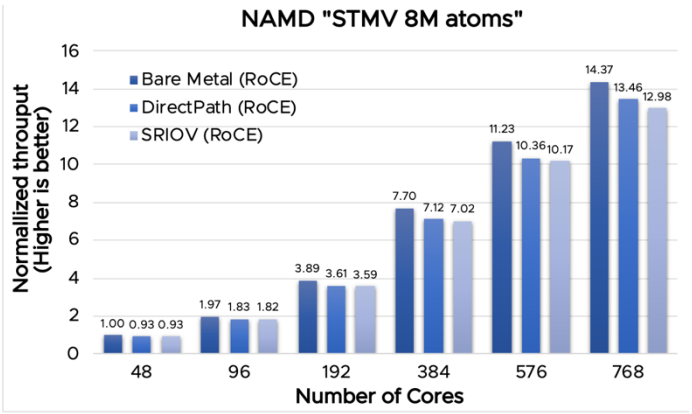
(e) OpenFOAM

Figure 5. GROMACS, NAMD, LAMMPS, WRF & OpenFOAM Performance on Bare Metal, DirectPath IO, and SRIOV using InfiniBand.

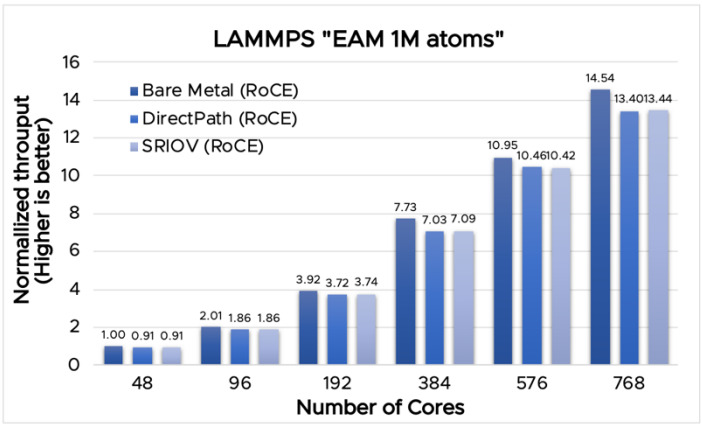
When using RoCE, Figures 6a~6e demonstrate that at the largest scale of 768 cores (16 nodes), SRIOV introduces an overhead of 8.5% on GROMACS, 9.7% on NAMD, 7.6% on LAMMPS, 4% on WRF, and interestingly, a performance improvement of 1.1% on OpenFOAM compared to bare metal. On the other hand, DirectPath IO exhibits an overhead of 8.1% on GROMACS, 6.3% on NAMD, 7.9% on LAMMPS, 3.5% on WRF, and 1.1% on OpenFOAM compared to bare metal. Similarly, except for NAMD in SRIOV-IB with an additional 1.1% overhead, all the other workloads fall within the acceptable performance range indicated by our 8.3% performance ruler.



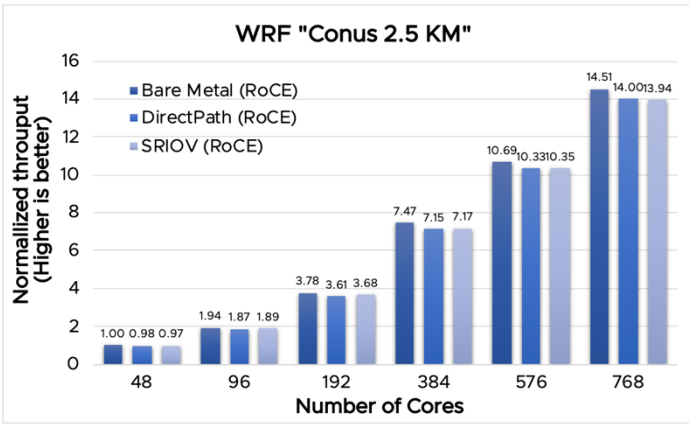
(a) GROMACS



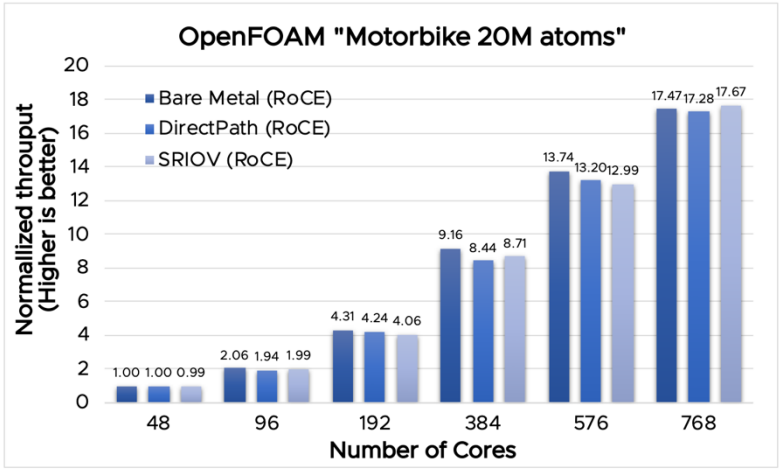
(b) NAMD



(c) LAMMPS



(d) WRF



(e) OpenFOAM

Figure 6. GROMACS, NAMD, LAMMPS, WRF & OpenFOAM Performance on Bare Metal, DirectPath IO, and SRIOV using RoCE.

Based on our findings, SRIOV demonstrates performance comparable to DirectPath IO and even approaches the performance of bare metal. While there is a cost of losing 4 cores compared to bare metal, the advantages for users' production, cluster management, and maintenance are significant. SRIOV allows for improved resource utilization and the ability to deploy multiple VMs sharing a single RDMA HCA/NIC. Therefore, we recommend choosing SRIOV over DirectPath IO for MPI workloads.

Performance Comparison between RoCE and IB for MPI Workload

Over the years, InfiniBand (IB) has been the dominant technology in HPC networks, with a peak usage of 44% in 2012 (Figure 7). However, since 2017, Remote Direct Memory Access over Converged Ethernet (RoCE) has gained momentum, becoming the preferred choice for high-speed Ethernet-based HPC networks and cloud ecosystems. As of 2022, approximately 40% of network designs in the Top500 leverage RoCE, while IB usage has gradually declined to 23% (from 44% in 2012 to 30% in 2017).

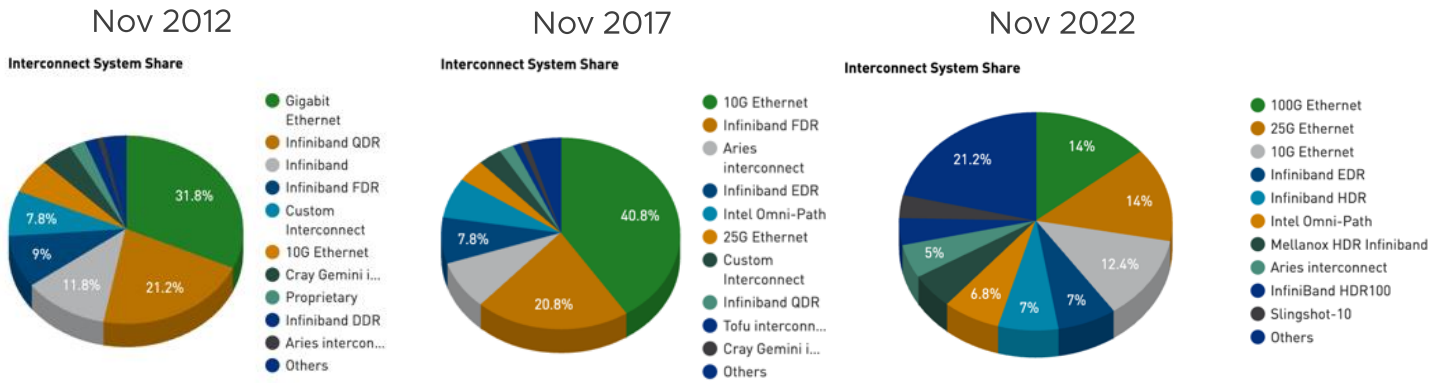


Figure 7. Interconnect system share from 2012 to 2022. (Source: https://www.top500.org/statistics/list/)

To adapt to the evolving landscape of HPC network technologies, VMware acknowledges the significance of supporting both IB and RoCE. While ESXi traditionally only supported Ethernet, our collaboration with hardware vendors has enabled the utilization of the IB protocol for data transfer through DirectPath IO since 2009. In this subsection, we will explore the performance differences between IB and RoCE for MPI workloads in recent vSphere versions.

Table 4: Strong Scaling Performance Difference between IB and RoCE using SRIOV on five Real-world HPC Workloads.

+: IB is faster -: RoCE is faster					
# cores	GROMACS	NAMD	LAMMPS	OpenFOAM	WRF
48	0.4%	-0.2%	0.0%	-0.1%	-1.0%
96	-1.1%	0.2%	-0.7%	-0.1%	0.2%
192	2.6%	0.0%	-0.4%	-0.3%	-0.5%
384	0.1%	-0.2%	0.2%	0.0%	-0.4%
576	-1.2%	-0.1%	1.1%	-5.5%	-0.7%
768	0.2%	0.3%	1.1%	-0.7%	-0.4%

Table 5: Strong Scaling Performance Difference between IB and RoCE using DirectPath IO on five Real-world HPC Workloads.

+: IB is faster -: RoCE is faster					
# cores	GROMACS	NAMD	LAMMPS	OpenFOAM	WRF
48	-0.2%	0.2%	0.5%	0.0%	0.2%
96	0.2%	-0.1%	0.4%	0.2%	0.2%
192	-0.7%	0.2%	-0.5%	1.3%	-0.8%
384	1.4%	-0.1%	1.0%	-2.5%	1.2%
576	-0.4%	-0.4%	1.4%	-1.7%	0.1%
768	0.1%	-0.2%	1.0%	-1.8%	0.8%

In our laboratory environment, we possess a dual-port CX-5 VPI card that can be configured as either IB or Ethernet. Specifically, Port 1 is connected to an IB switch (Mellanox SB7800 100 Gb), while Port 2 is linked to an Ethernet switch (Dell PowerSwitch S5232F 100GbE). Leveraging the results already presented in [*Ref7*, *Ref8*, *Ref9*] on five real-world HPC workloads, we present the performance disparity between RoCE and IB in SRIOV (Table 4) and DirectPath IO (Table 5) scenarios. Generally, the performance gap between RoCE and IB is negligible, typically around 1%. We also observe an interesting finding worth noting. When utilizing SRIOV on 12 nodes, RoCE showcases a 5.5% advantage over IB. This performance improvement has piqued our interest, and we plan to conduct further investigation to better understand the underlying factors contributing to this phenomenon.

Furthermore, we have collaborated with Dell's HPC lab to certify the *HPC on Demand system* as an *Intel® Select Solution* for Simulation and Modeling. This certification was achieved on a 32-node cluster featuring Intel® Xeon® Scalable third-generation processors (IceLake) and ConnectX-6 100GbE **RoCE** network. Notably, the system achieved exceptional performance, with MPI-based HPC workloads showcasing near bare-metal performance with less than 10%. [*Ref3*].

Considering the maturity of RoCE, the broad development of safety techniques for Ethernet and near bare-metal performance, we recommend prioritizing RoCE over IB for your future virtualized HPC deployments.

Summary

We begin by guiding you through different methods for RDMA HCA/NIC usage in VMware stacks, discussing the advantages and disadvantages of each approach. This empowers you to make informed decisions that align with the specific needs of your HPC user community.

Next, we address inquiries about performance differences among VMware network adapter types (VMXNET3, PVRDMA, SRIOV, DirectPath I/O) for MPI workloads. Through strong scaling tests on five real-world HPC workloads, we evaluate these adapters using various core configurations and compare the performance disparities between InfiniBand (IB) and Remote Direct Memory Access over Converged Ethernet (RoCE) in vSphere.

Our findings highlight that SRIOV outperforms VMXNET3 and PVRDMA in terms of bandwidth and latency, not only for pure MPI communication but also for real-world workloads. We also observe performance differences between IB and RoCE, with RoCE being preferred due to its maturity and the availability of well-established safety techniques for Ethernet.

While virtualization incurs an acceptable impact compared to bare metal, the benefits it offers, such as enhanced resource utilization and security tools, and the flexibility to deploy multiple VMs sharing a single RDMA HCA/NIC, outweigh this virtualization tax. Based on our results, we recommend utilizing SRIOV for MPI workloads, especially when compared to DirectPath IO and bare metal deployments.

In summary, this paper addresses inquiries about VMware network adapter performance, explores IB and RoCE disparities in vSphere, discusses virtualization tax versus bare metal, and provides recommendations for running HPC workloads. By considering our insights, recommendations, and future technology innovation (e.g., DVX), you can optimize your VMware environment for HPC applications and achieve superior performance.

References

[1] *Achieving Near Bare-metal Performance for HPC Workloads on VMware vSphere 7*
[2] *Virtualizing High Performance Computing (HPC) on VMware vSphere® 7 Using Intel® Select Solution for HPC*
[3] *Performance Evaluation of HPC Applications on a Dell PowerEdge R650-based VMware Virtualized Cluster*
[4] *Performance Study of HPC Scale-Out Workloads on VMware vSphere 7*
[5] *Bringing Secure Multi-tenancy to High Performance Computing and Machine Learning*
[6] *Secure Networking for Multi-Tenant High-Performance Computing and Machine Learning*
[7] *InfiniBand SR-IOV Setup and Performance Study on vSphere 7.x*
[8] *RoCE SR-IOV Setup and Performance Study on vSphere 7.x*
[9] *InfiniBand and RoCE DirectPath IO Setup and Performance Study on vSphere 7.x*
[10] *VMware Paravirtual RDMA for High Performance Computing*
[11] Running HPC and Machine Learning Workloads on VMware vSphere
[12] *vHPC-Toolkit Full Documentation*
[13] *vHPC-Toolkit Tutorial Video in VMware User Group (VMUG) Meeting*
[14] *VMware vSphere*
[15] *VMware vSAN*
[16] *VMware vCenter Server*
[17] VMworld Barcelona 2018: VIN2677BE – Extreme Performance Series: Performance Best Practices
[18] *Virtual Networking Made Easy with VMware VMXNET3 Driver*
[19] *Get HPC on Demand*
[20] *vSphere ML Accelerator Spectrum Deep Dive – Using Dynamic DirectPath IO (Passthrough) with VMs*
[21] *How vSphere HA Works*
[22] *Task Bench: A Parameterized Benchmark for Evaluating Parallel Runtime Performance*
[23] *Interconnect Acceleration for Machine Learning, Big Data, and HPC*
[24] *HPC on Demand*
[25] *Intel® Select Solution*

About the Author

Yuankun Fu, Senior Member of Technical Staff, wrote the original content of this paper.

Yuankun is a Senior Member of Technical Staff in the VMware OCTO team. He holds a Ph.D. degree in Computer Science with a specialization in HPC from Purdue University. Since 2011, he has worked on a wide variety of HPC projects at different levels from hardware, middleware, to application. He currently focuses on the HPC/ML application performance on the VMware multi-cloud platform, from creating technical guides and best practices to root-causing performance challenges when running highly technical workloads on customer platforms. In addition, he serves on the Program Committee of international academic conferences, such as *eScience'23* and *BigData'23*. Yuankun also enjoys musicals, museums, basketball, soccer, and photography in his spare time.

Acknowledgments

The author would like to express his sincere gratitude to Ramesh Radhakrishnan, Brayn Tan, Janakiram Vantipalli, Michael DeMoney and Frank Denneman from VMware for their invaluable advice, expert insights, and thorough peer review of this study. Their guidance and feedback have been instrumental in shaping the findings and conclusions presented in this paper. The author also extends thanks Catherine Xu for her editing and publishing this document.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com.