13th ANNUAL WORKSHOP 2017

VMWARE PARAVIRTUAL RDMA

DEVELOPER PERSPECTIVE

Adit Ranadive, Aditya Sarwade, Jorgen Hansen, Bryan Tan, Shelley Gong, George Zhang, Na Zhang, Josh Simons

VMware, Inc.

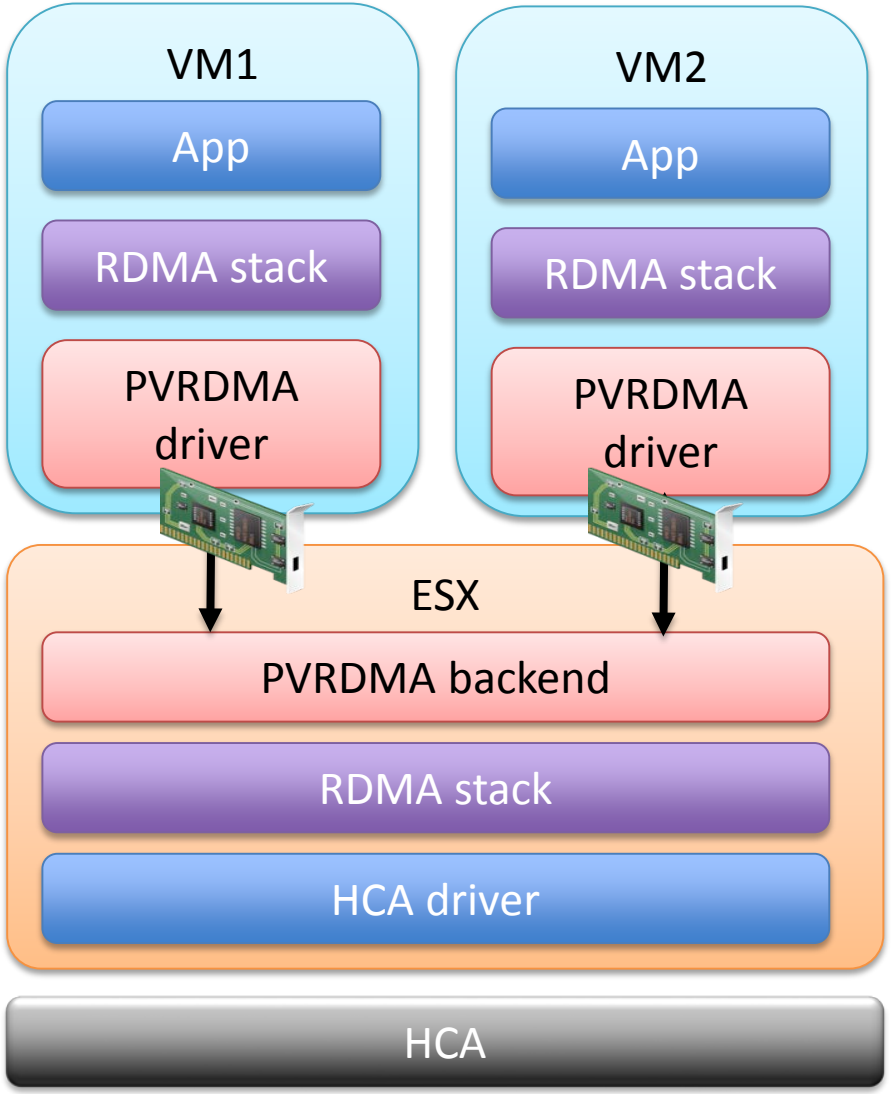
[28th March, 2017]

AGENDA

- **Overview of Paravirtual RDMA Device**
- **Device development process**
- **Challenges for device development and upstreaming**
- **Passthrough RDMA Updates**
- **Conclusion/Future Work**

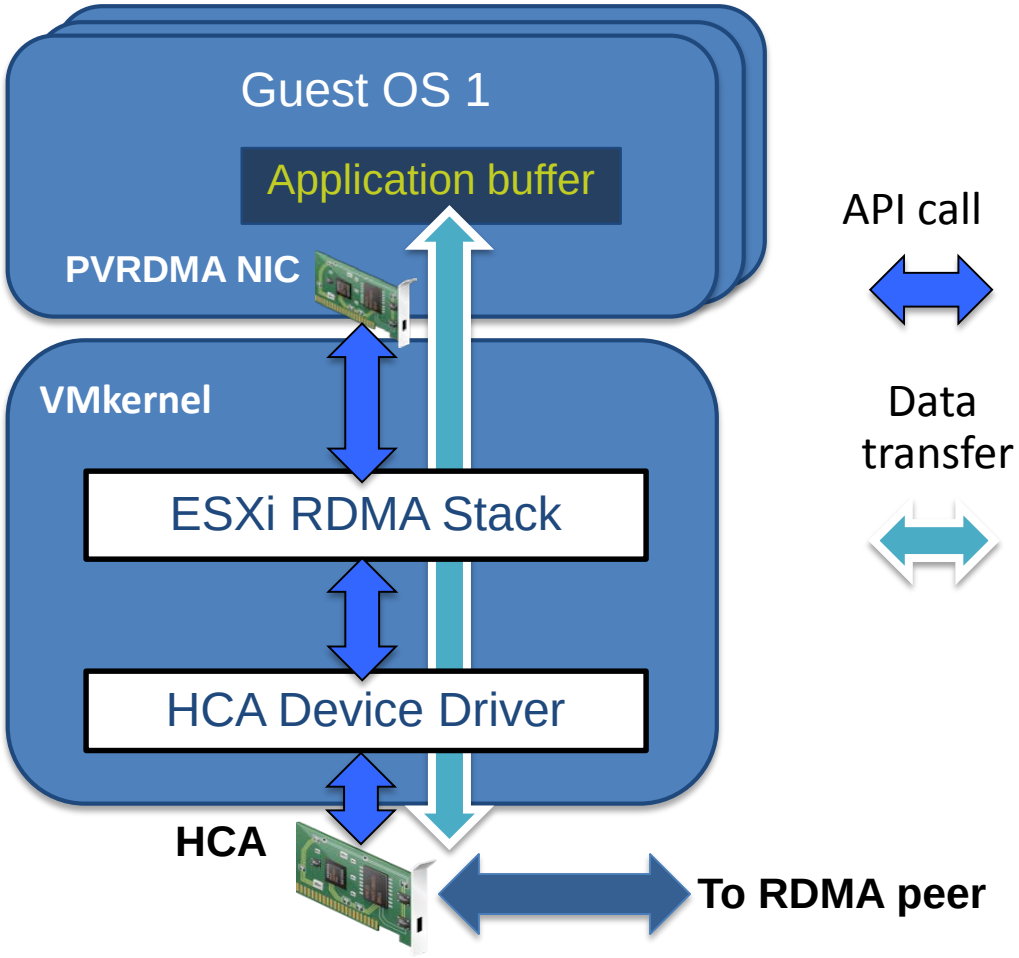
PVRDMA DEVICE

- **Paravirtual RDMA (PVRDMA) is a new PCIe virtual NIC**
  - A network interface (VMXNet3)
  - An RDMA provider (RoCE)
  - RDMA provider plugs in to the OFED stack
    - Verbs-level emulation
    - In kernel and user-space
- **ESX**
  - Leverage native RDMA stack
  - Physical HCA services all VMs
- **Uses HCA for performance, but works without it**
- **Virtual devices can share an HCA without SR-IOV**
- **Supports vMotion (live migration)!**



PVRDMA ARCHITECTURE

- **PVRDMA exposes virtual resources to VM**
  - E.g., PD, CQ, QP, MR
- **ESX creates corresponding HW resources**
- **Guest memory regions are registered as host physical addresses in HW**
- **Work Requests (WRs) are queued on vQPs**
- **PVRDMA backend posts these WRs on to corresponding QPs opened in the ESXi RDMA stack**
- **HCA performs DMAs to/from application memory without any SW involvement**
  - Enables direct zero-copy data transfers in HW!



CURRENT STATUS

- **PVRDMA Device released as part of vSphere 6.5**
  - ESXi hypervisor + Virtual Center Management Platform + Virtual Networking Management
  - Supports RoCEv1
  - RC, UD QPs
- **Linux Driver**
  - Included in 4.10
  - OFED 4.8 RC1 (as tech preview)
- **User-space Library**
  - rdma-core-13
  - OFED 4.8 RC1

DEVELOPING PVRDMA DEVICE

- **RDMA APIs (though numerous) are well-defined!**
  - Both user-level and kernel are similar
- **VM Compatibility is a big priority at VMware!**
  - Virtual Devices expected to work when the VM is moved to a different host without an HCA
  - Has to work with virtualization features - vMotion, Snapshots
- **3 Devices/Transports**
  - Memory copy - VMs on same host
  - HW RDMA – VMs have access to HCAs
  - TCP Emulation – Either VM peer has no HCA
- **Enforce consistent RDMA behavior between transports**
  - Completely hidden from guest software
  - Interoperate QPs between transports
  - Testing scenarios increase

DEVICE DEVELOPMENT CHALLENGES

- **Putting 3 devices together**
  - Started with the Memcpy/TCP modes
    - Limited to APIs in virtualization environment
  - Physical HCA support was added later to ESX
    - Special ULP to VMkernel RDMA

- **What is consistent RDMA behavior?**
  - Differences between IB specification and existing RDMA NIC behavior
  - How do you test IB spec compliance or OFED compliance?
    - Standardized compliance tests from OFIWG?
  - Important from a testing perspective – what are your expected results/failures?

- **Memory Regions**
  - Emulated user MR support (when there was no physical support)
  - Physical MRs added to the ESX device driver to support remote read/write (user MRs)
  - No support for DMA MRs with remote read/write

DEVICE DEVELOPMENT CHALLENGES

- **Unreliable Datagram**
  - Receives from multiple transports
  - No way to un-enqueue a WQE
  - Post “bounce buffers” on physical HCA – copy to guest buffers
- **vMotion support**
  - Partial - VM has to be stopped at the end of a posted WR
  - Cannot communicate with native host
  - Need hardware support to perform this gracefully
    - Create specific resource identifiers
    - Suspend/Resume of Queue Pairs
- **Stuck at RoCEv1**
  - RoCEv2 wasn't finalized till closer to our release date
  - Not enough support in distros as yet for RoCEv2
- **Harder to release device updates**
  - Ensure VM compatibility through virtual hardware versioning
  - Cannot just release device firmware updates

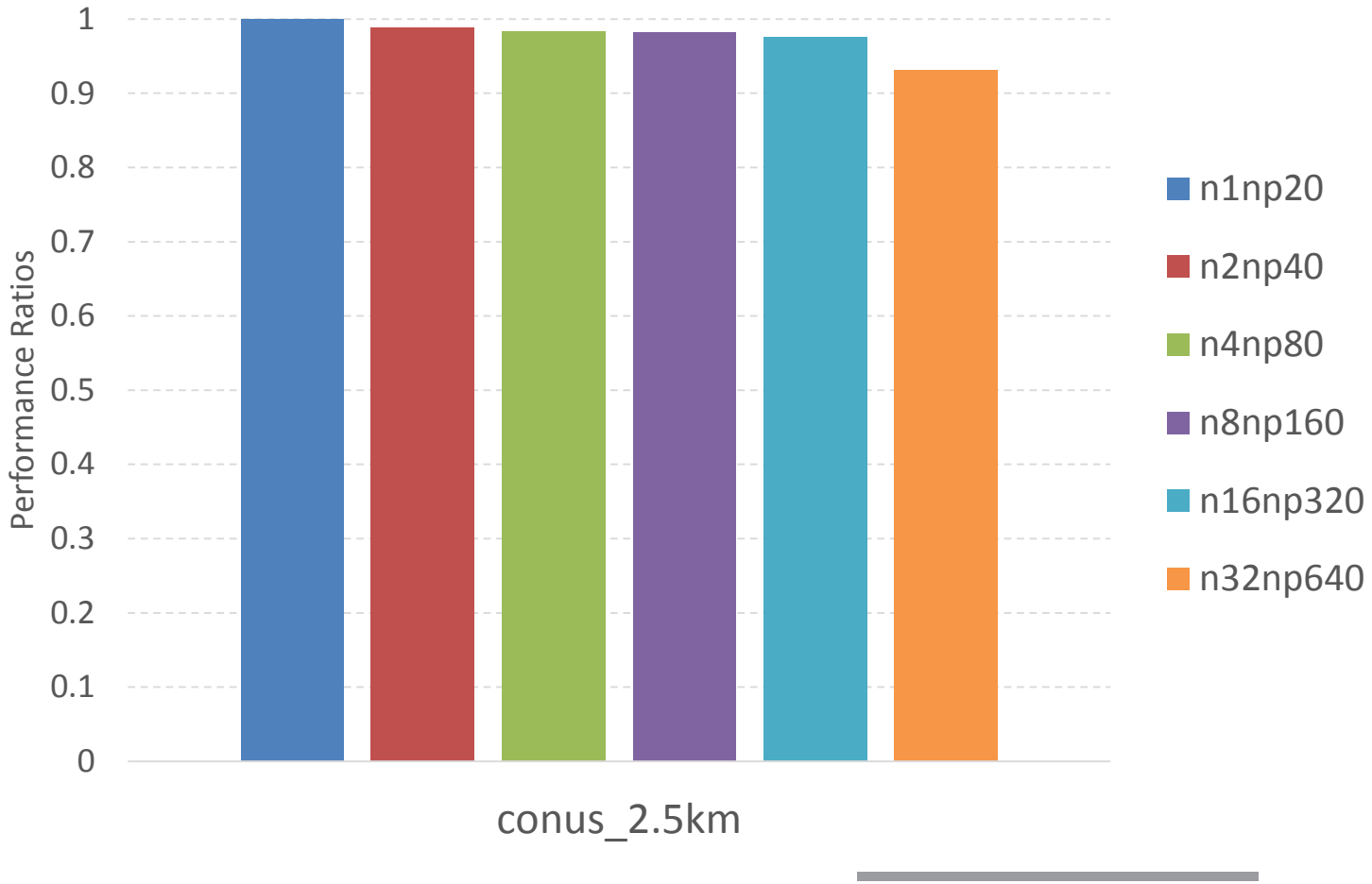
GUEST SOFTWARE DEVELOPMENT CHALLENGES

- **OFED in guests**
  - Separate driver/library repositories
  - Compatible with distro/OFED – Didn't want to deal with changing upstream code
  - Dev and Test environment was same to test more easily
- **First experiences with upstreaming**
  - Not just driver but user-space as well
  - Learning experience!
  - Awesome to see the "Applied" message from Doug!
- **OFED/Kernel changes**
  - ABI changes
  - Keep updated with other API changes
  - Addition of rdma-core
- **Integration with OFED 4.8**
  - Slightly different process than upstreaming – Tech preview
- **Driver versioning**
  - Keep track of fixes and changes to driver/library

PASSTHROUGH RDMA UPDATE

Office of CTO
Na Zhang, Josh Simons

WEATHER RESEARCH & FORECASTING (WRF)



VMDirectPath I/O Technology

Test Cluster Configuration:

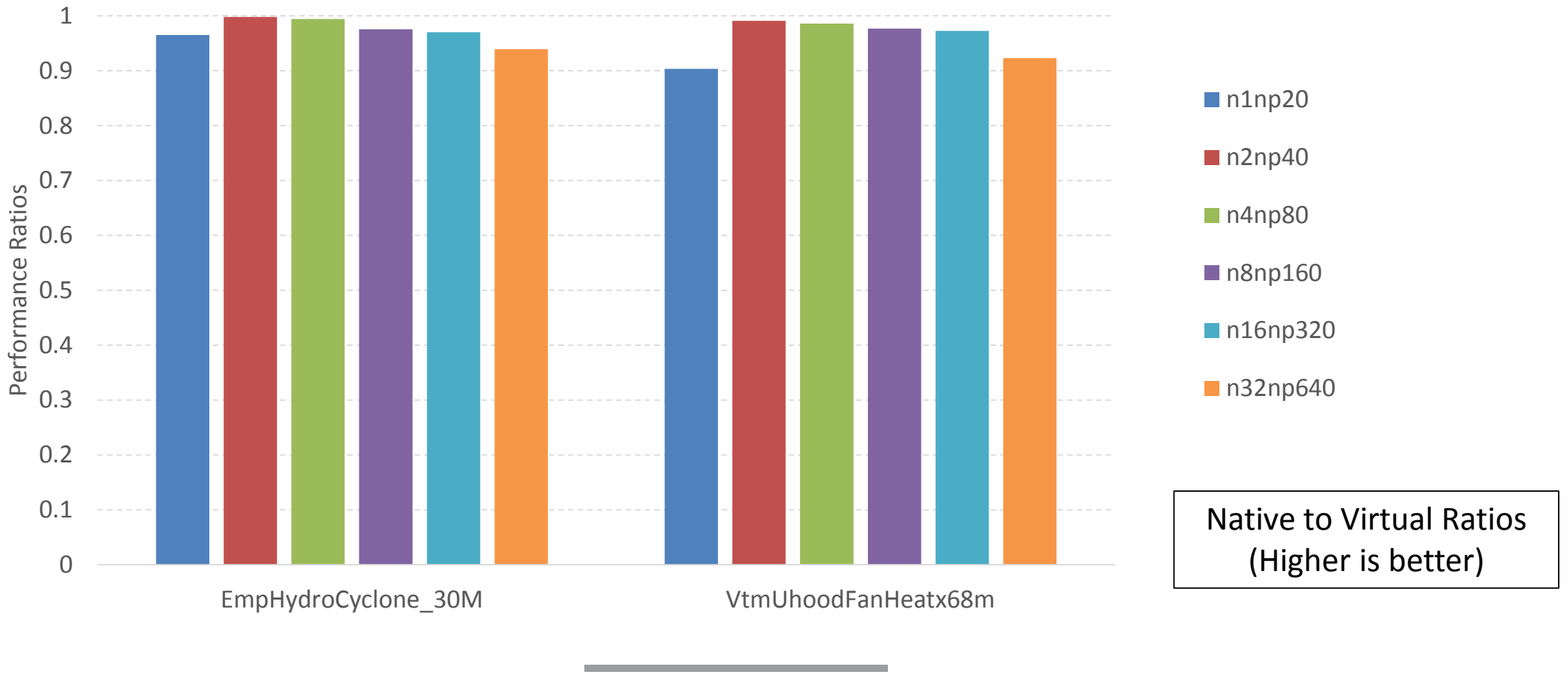
- 32-node Cluster
- Dell PowerEdge C6320
- Dual 10-core Haswell
- 128GB RAM
- 100Gb/s EDR InfiniBand
- ESX 6.5

VM Configuration:

- 1 VM per host
- 20 vCPUs
- 100GB RAM

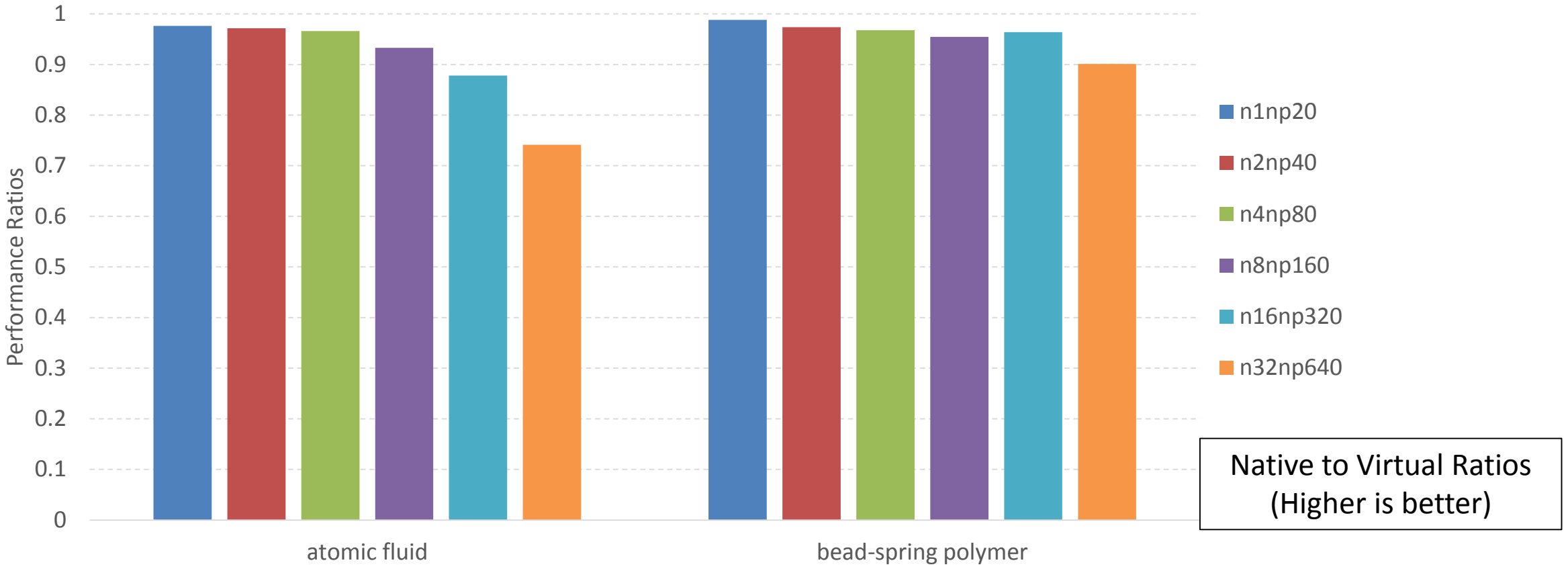
Native to Virtual Ratios
(Higher is better)

STAR-CCM+

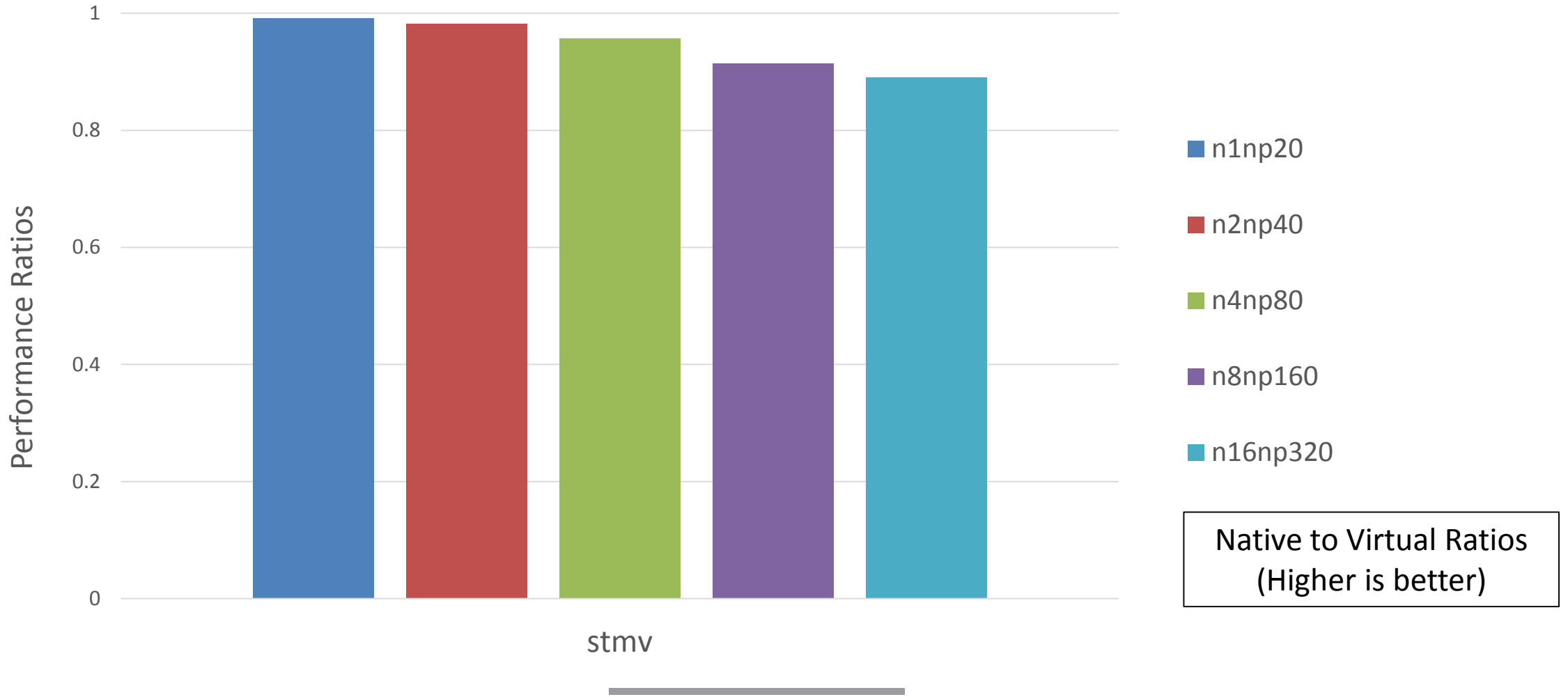


Native to Virtual Ratios
(Higher is better)

LAMMPS


Performance Ratios
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
atomic fluid
bead-spring polymer
n1np20
n2np40
n4np80
n8np160
n16np320
n32np640
Native to Virtual Ratios
(Higher is better)

NAMD


Performance Ratios
1
0.8
0.6
0.4
0.2
0
stmv
n1np20
n2np40
n4np80
n8np160
n16np320
Native to Virtual Ratios
(Higher is better)

CONCLUSIONS/FUTURE WORK

- **Driver/Library for VMware's PVRDMA device added to 4.10 kernel and OFED 4.8 RC1**
- **Some unique challenges for a paravirtual RDMA provider**
- **Overall great experience to work with open source and OFED community**
- **Looking at adding RoCEv2, Shared Receive Queues to PVRDMA**
- **RDMA is gaining more importance in virtualization settings**
  - Paravirtualization is one aspect
  - HCA vMotion support to talk to native hosts
- **Passthrough RDMA performance is pushing closer to native**
  - Hardware virtualization support keeps getting better

13th ANNUAL WORKSHOP 2017

THANK YOU

Adit Ranadive [aditr@vmware.com]

VMware, Inc.