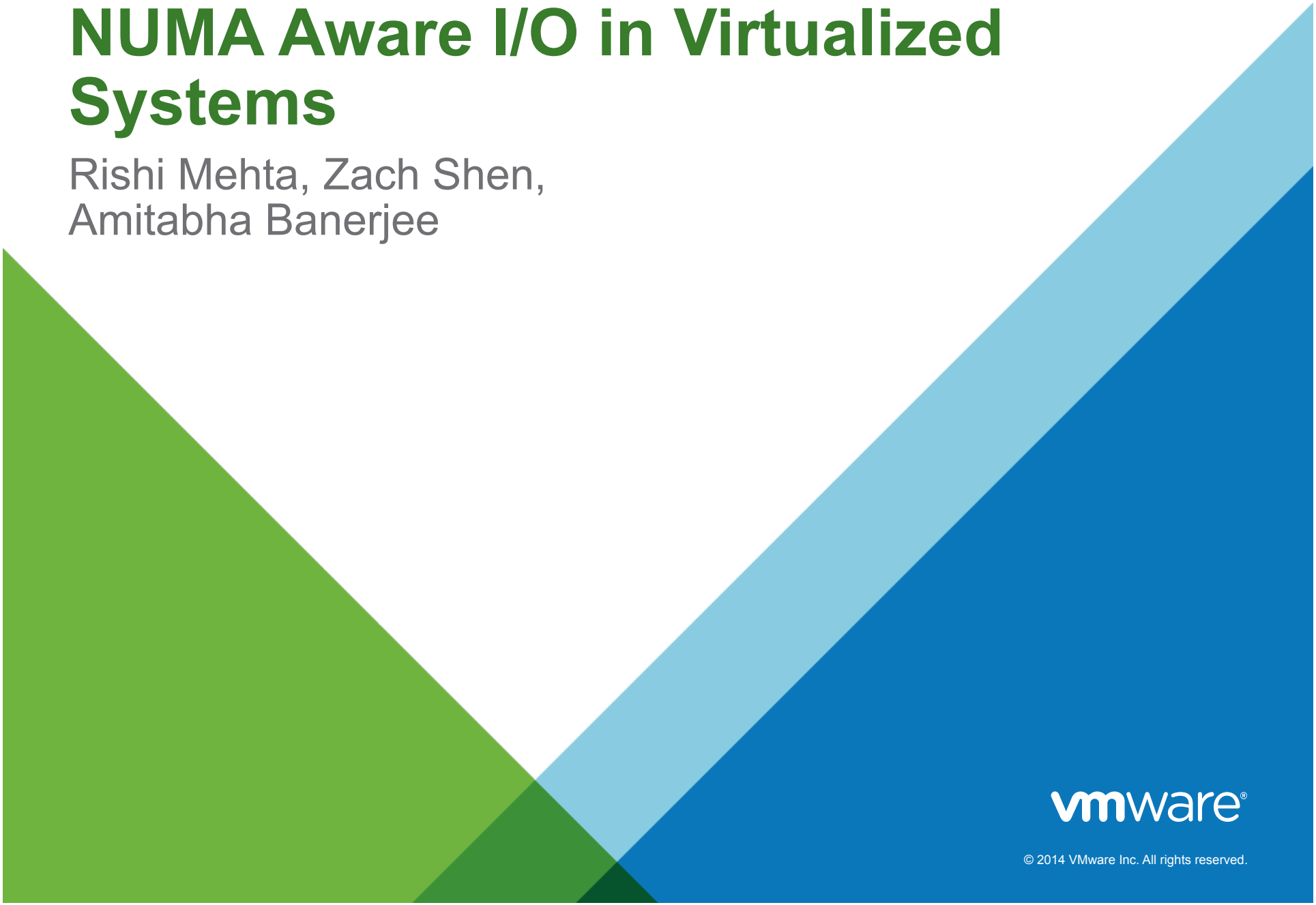
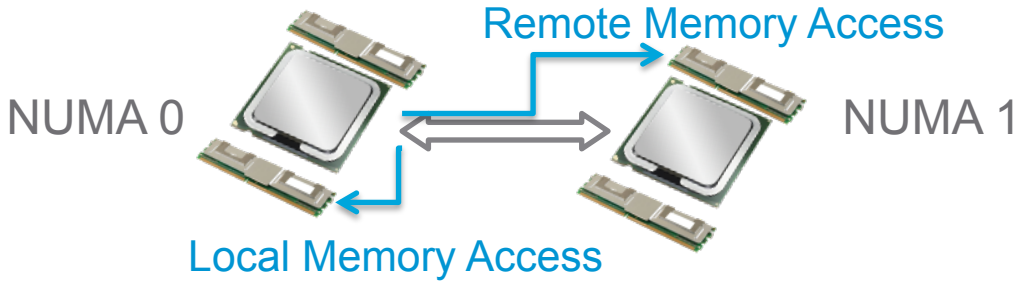
# NUMA Aware I/O in Virtualized Systems

Rishi Mehta, Zach Shen,
Amitabha Banerjee

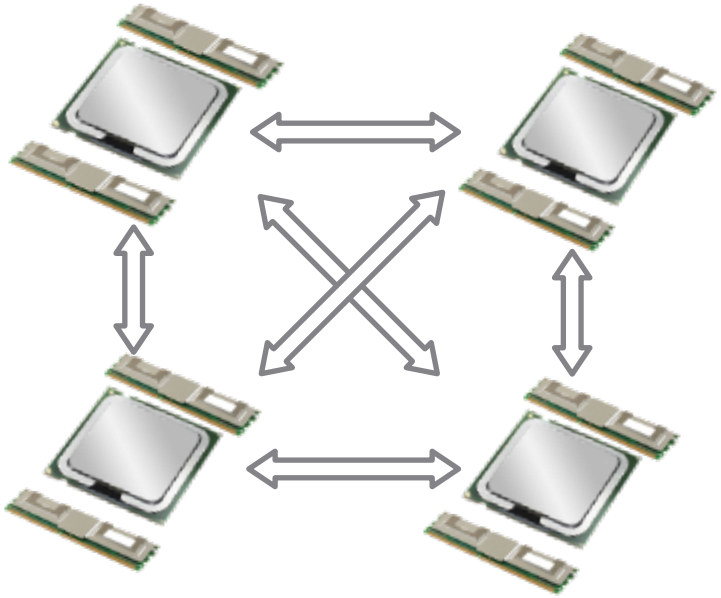


# Non-Uniform Memory Access

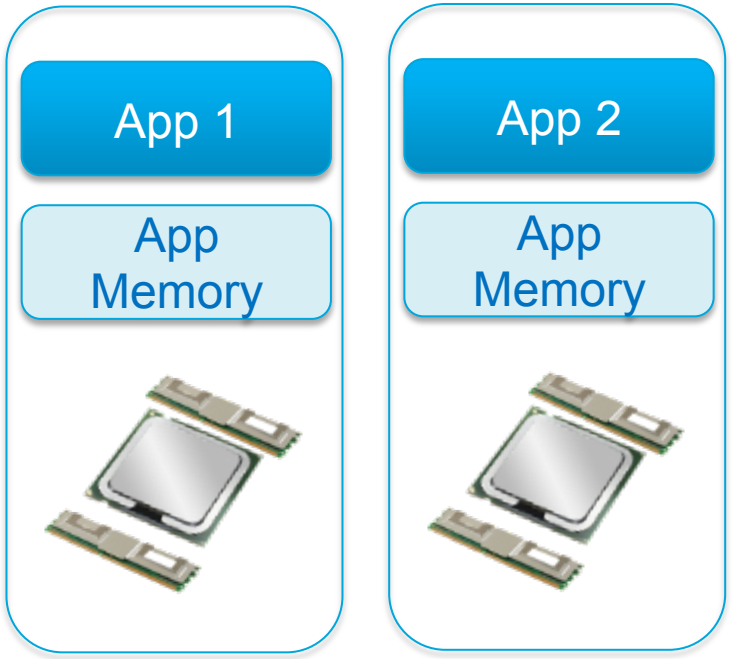




vmware

# Non-Uniform Memory Access

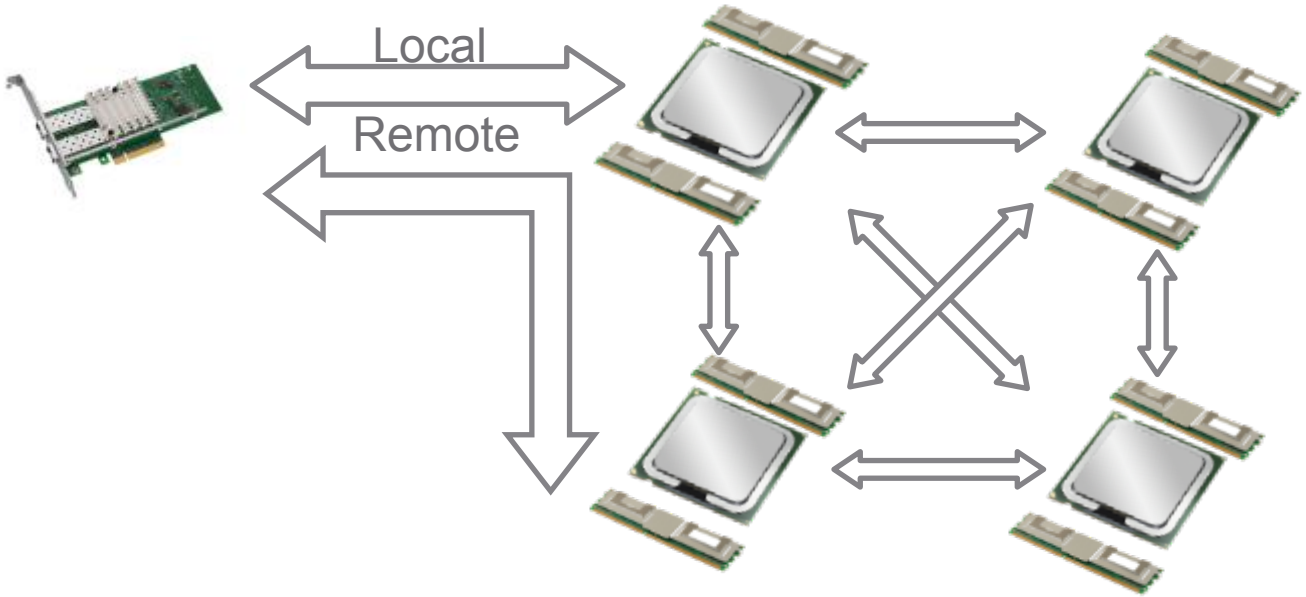


Modern Server Architecture



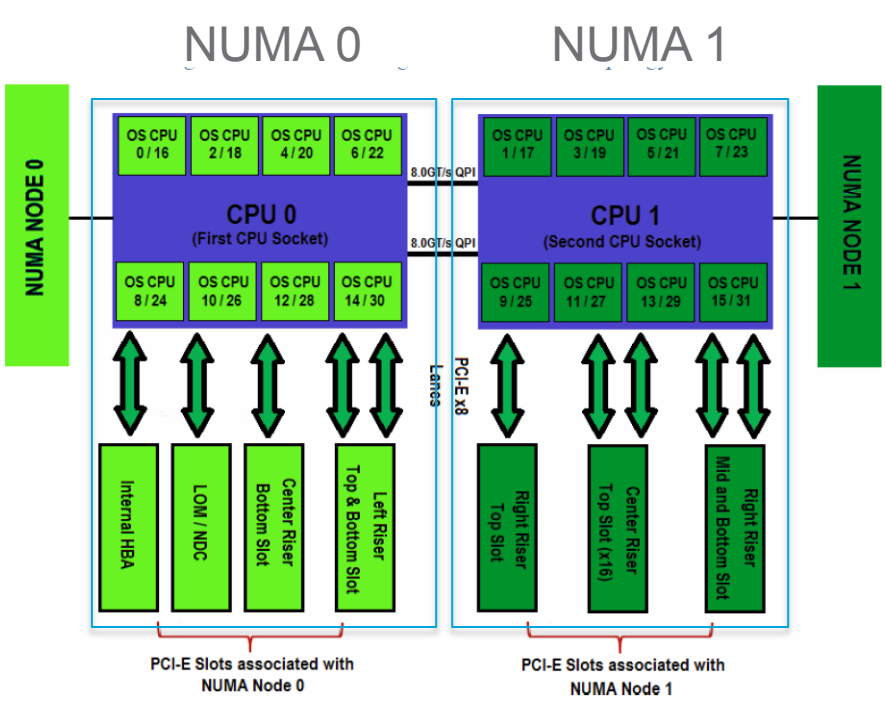

NUMA 1 NUMA 2

# Non-Uniform I/O Access

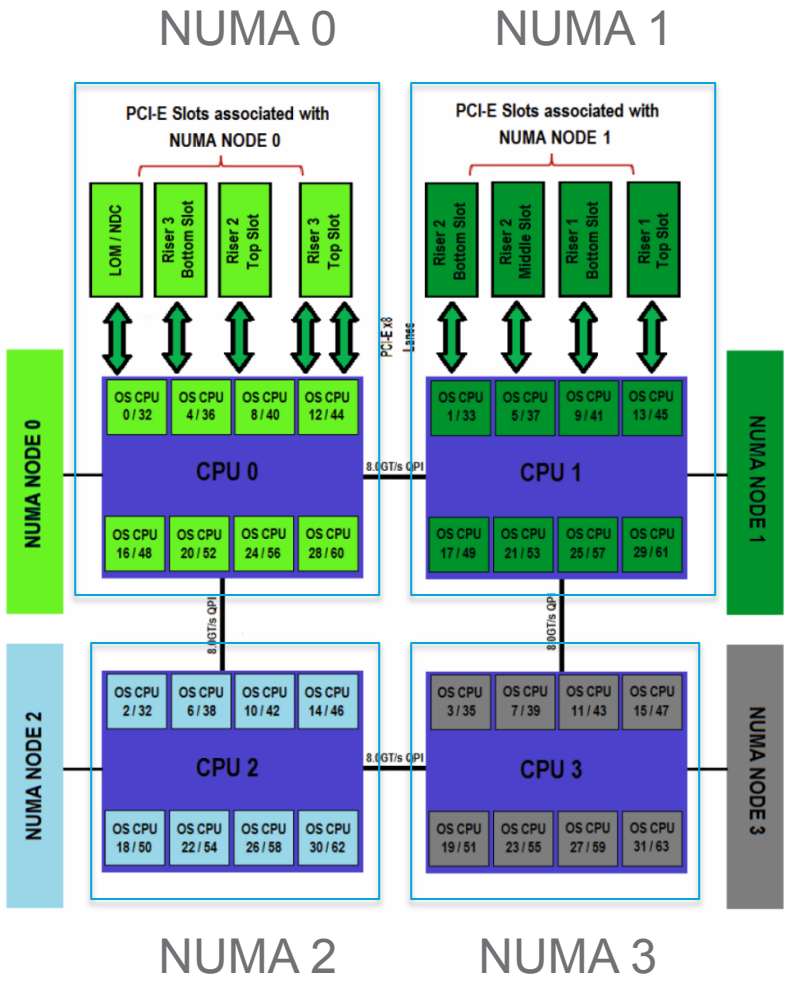



vmware

# Non Uniform I/O Access



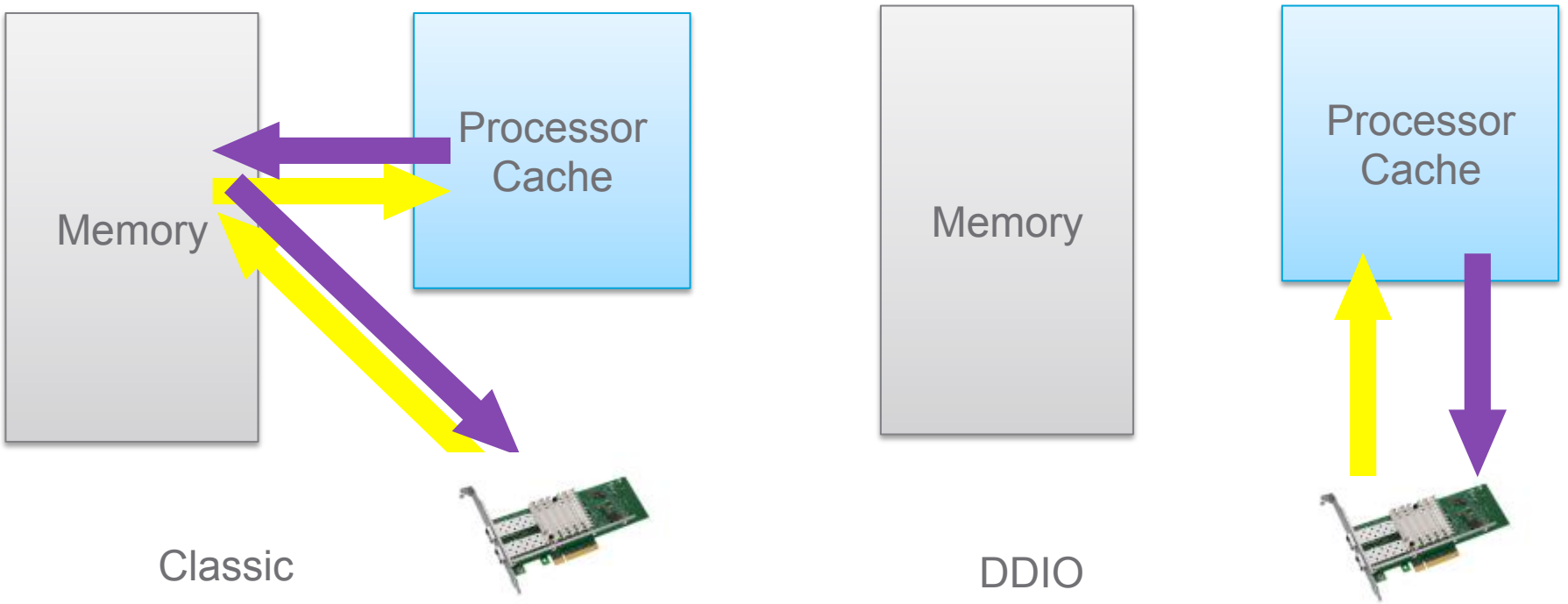

2 –socket Dell PE R720




4 –socket Dell PE R840

vmware

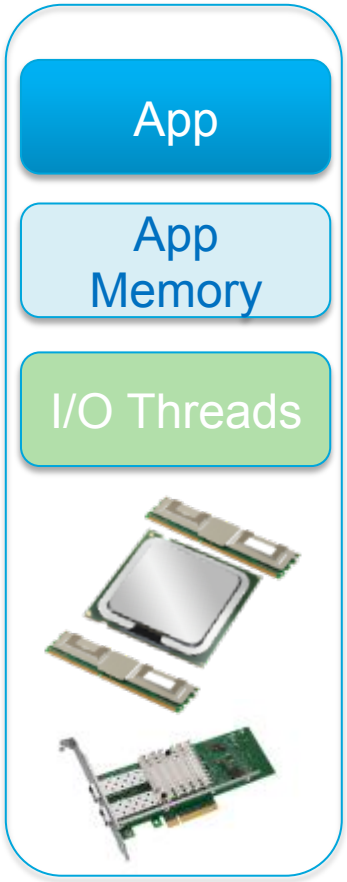
# Intel DDIO




*Currently works with Local PCI Device

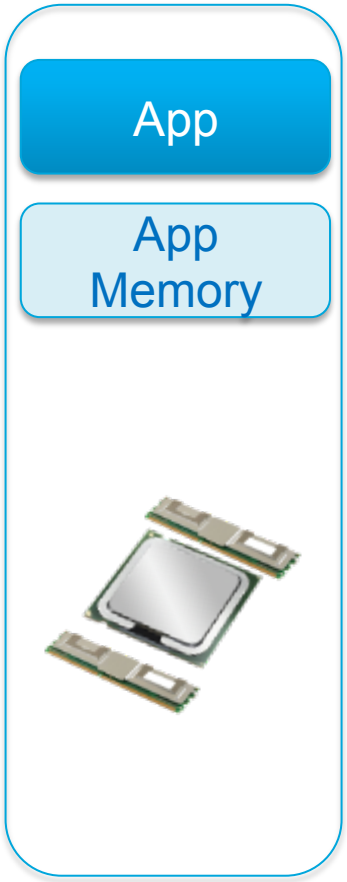
vmware

# Non-Uniform I/O Access




Ideal Situation

Intel DDIO Benefits

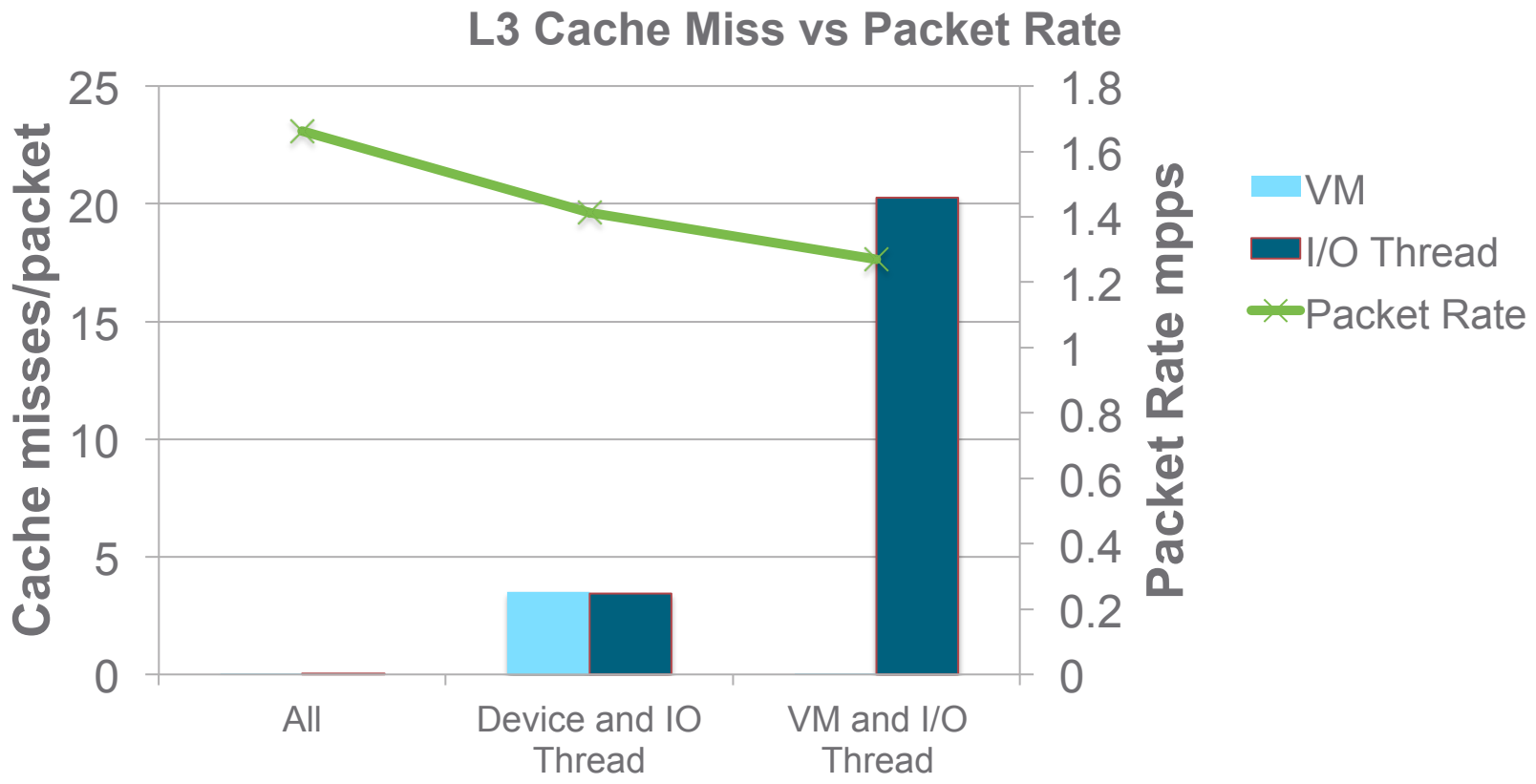




Local Memory Access

Real Situation

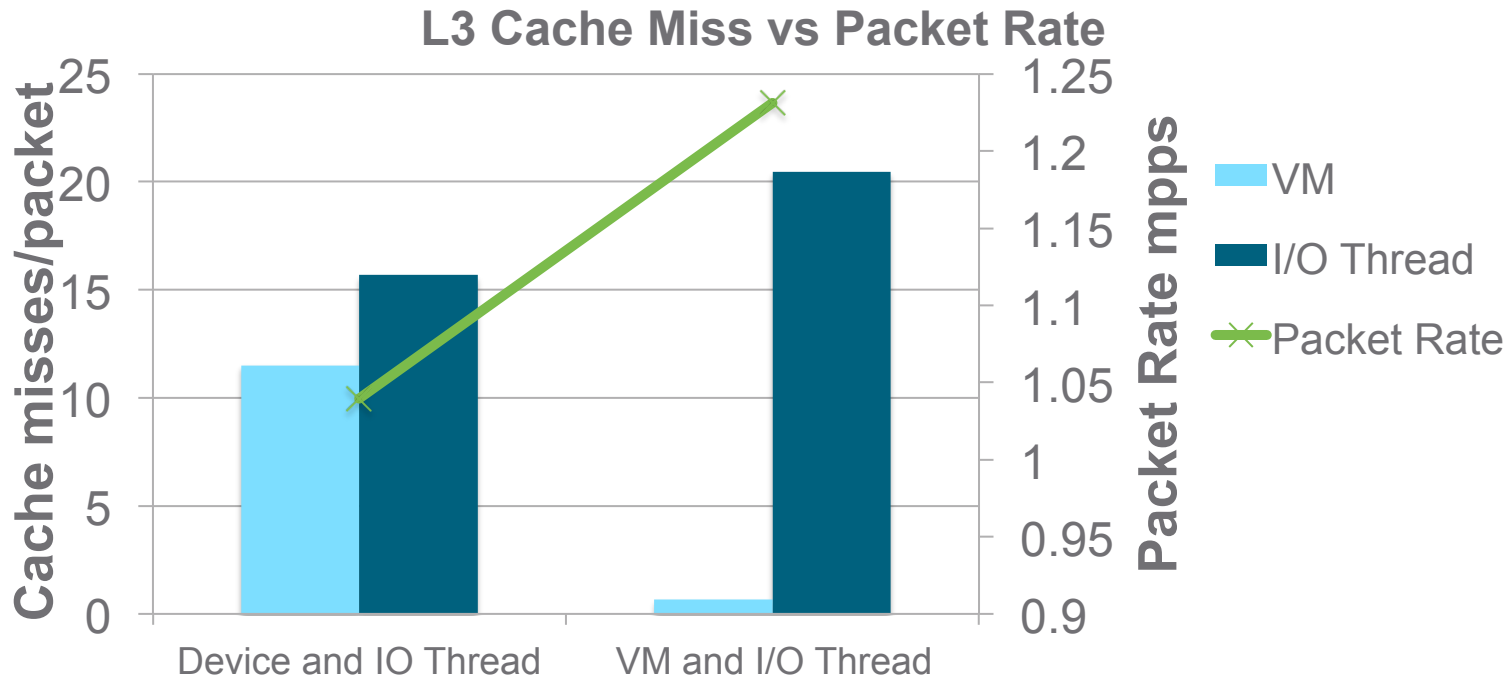
vmware

# DDIO benefits vs Local I/O : 1VM




vmware

# DDIO benefits vs Local I/O : 4VM



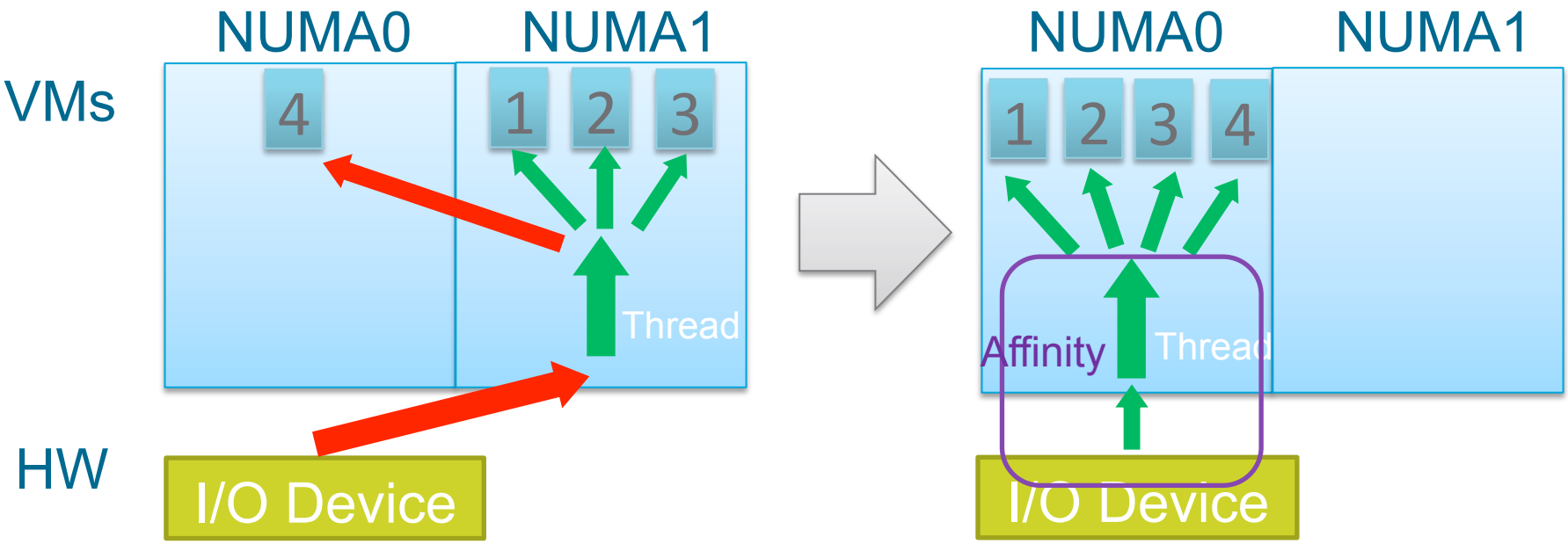

vmware

# NUMA I/O Scheduler

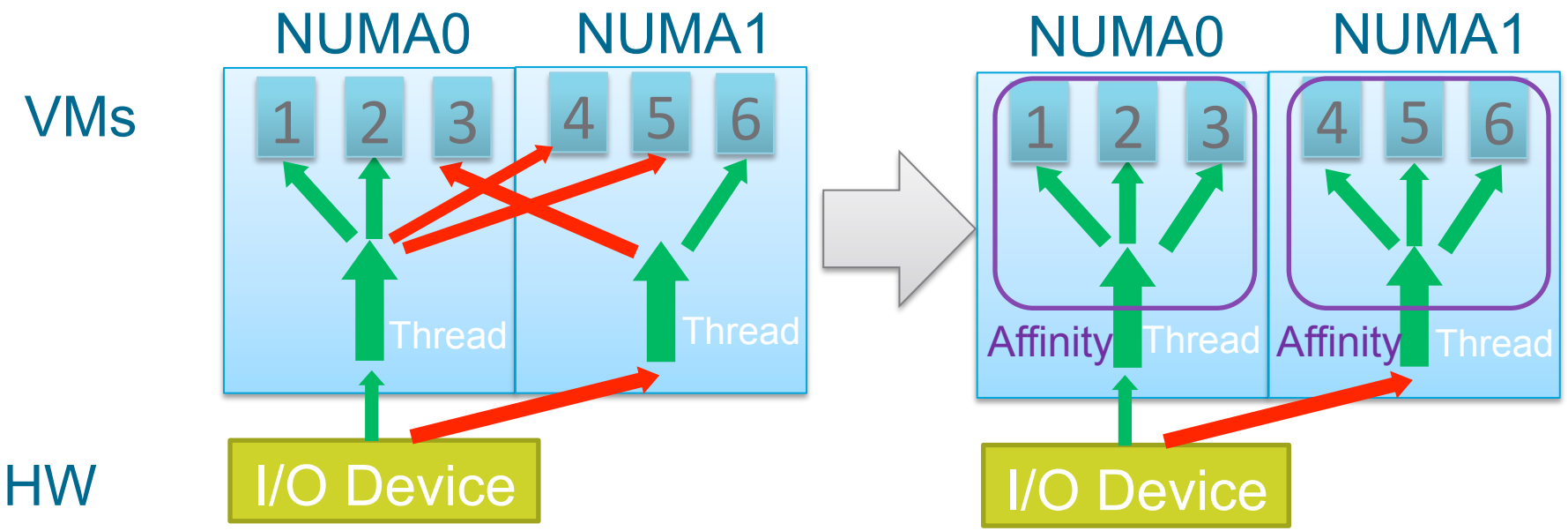
- Hybrid Mode
  - Low Load :
    - One I/O thread is sufficient for networking traffic
    - Pin I/O Thread to device NUMA Node
    - Let the scheduler migrate I/O intensive VM to device NUMA Node
  - High Load:
    - Sufficient load for multiple I/O Threads.
    - Create I/O Thread per NUMA Node.
    - Use I/O Thread based on Virtual Machines home node.



# NUMA I/O: Light load

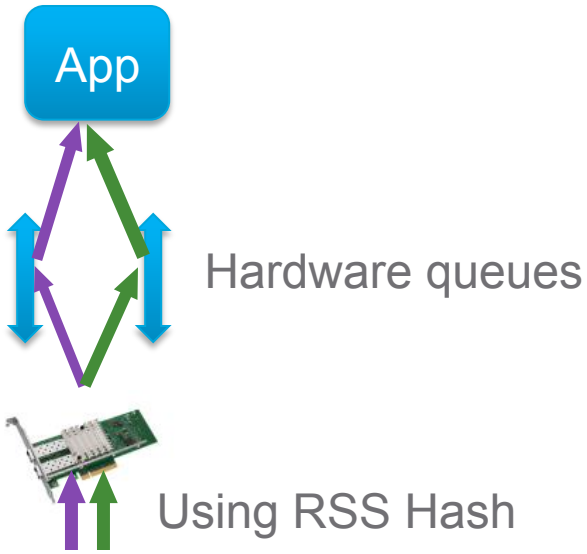


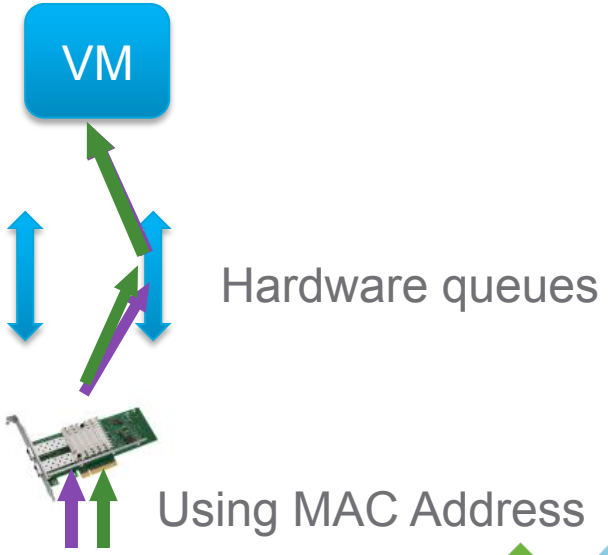
# NUMA I/O: Heavy load



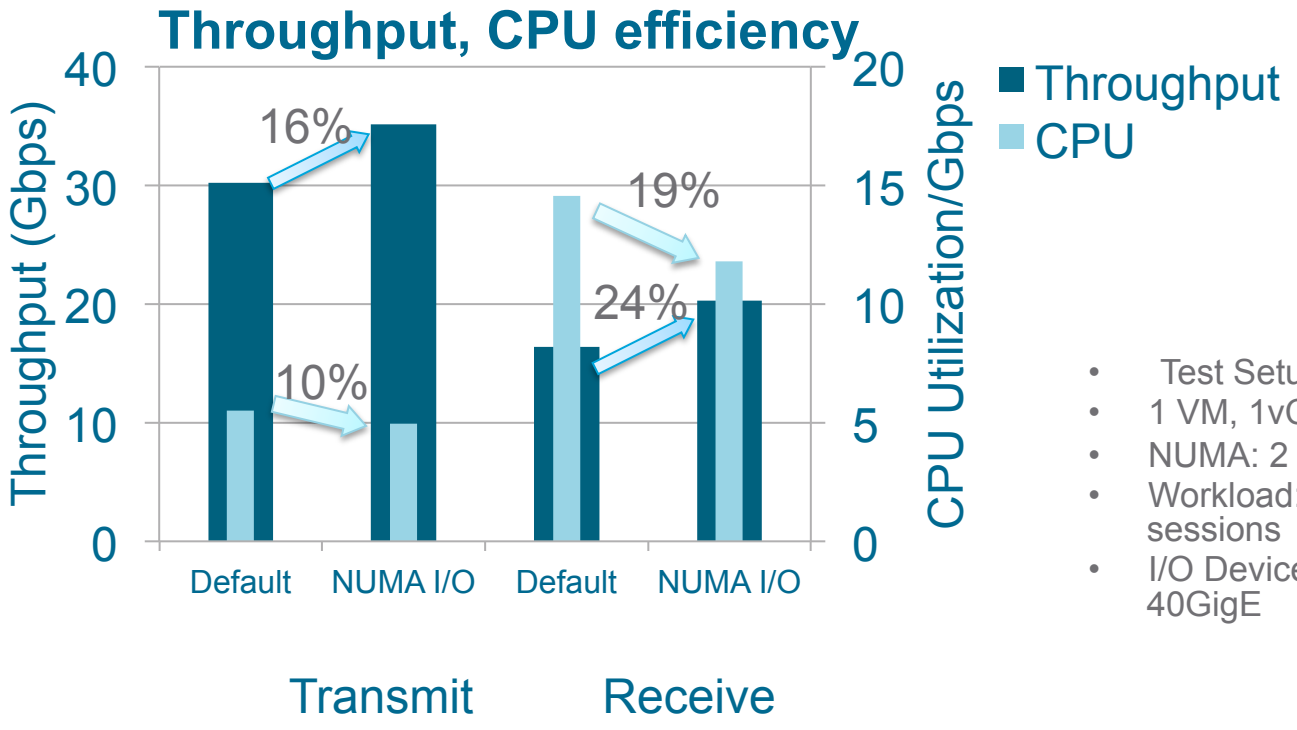
# Hardware Support

- Most NICs support multiple queues.
  - Multiple contexts to process I/O requests.
  - More scheduling choices
- Virtual machine provides abstraction at an application layer
  - Easier to Isolate Apps and I/O traffic using MAC Address as traffic classification instead of 5-tuple hash.



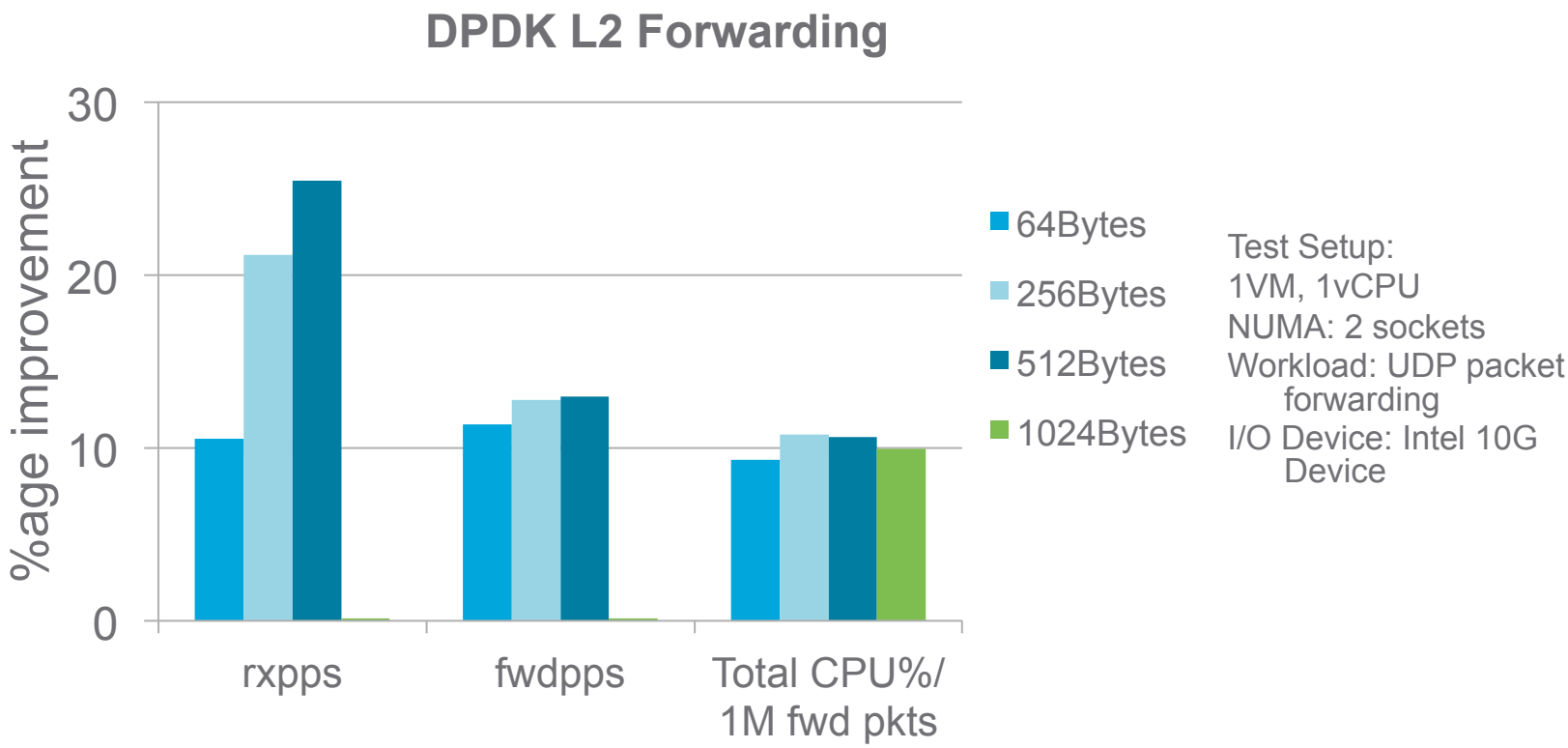
# Improvement: Light load




- Test Setup:
- 1 VM, 1vCPU
- NUMA: 2 sockets
- Workload: TCP bulk, 4 sessions
- I/O Device: Mellanox 40GigE

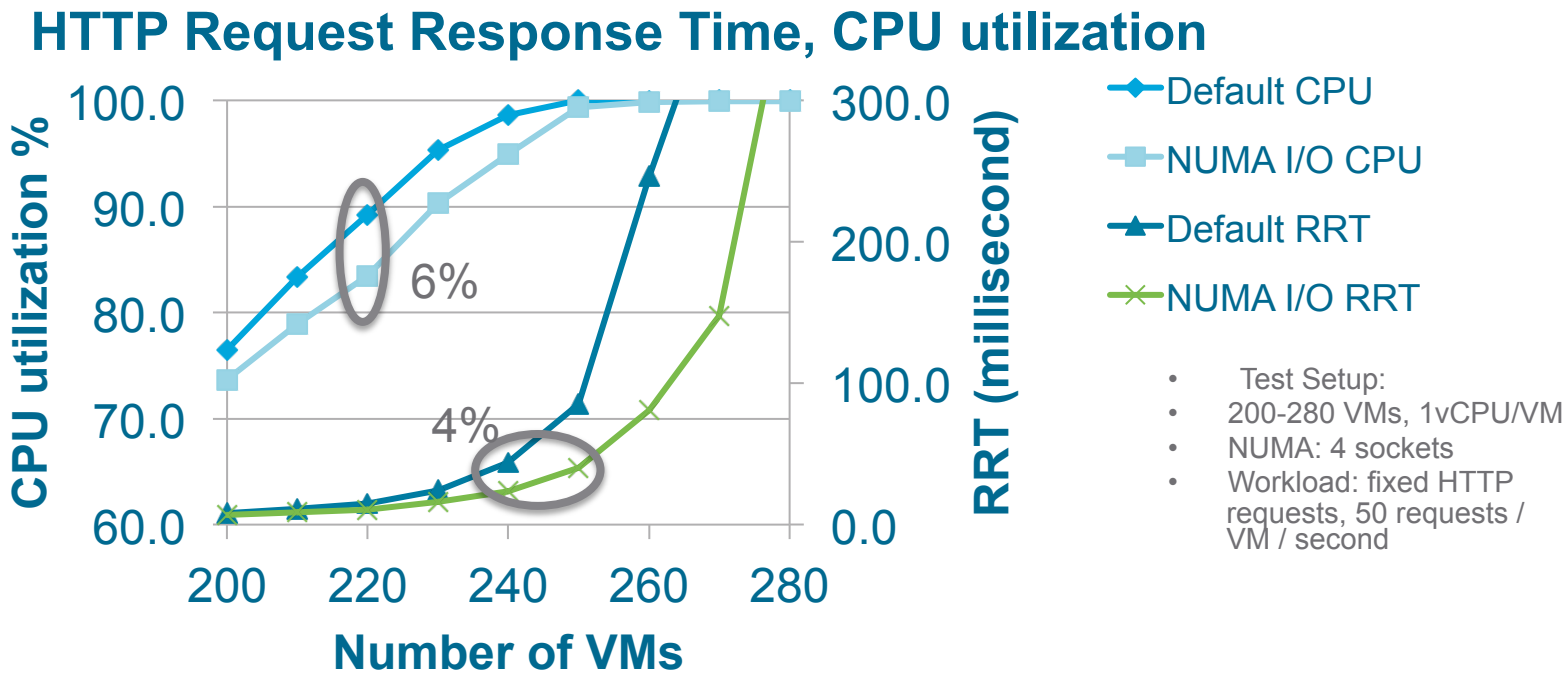
vmware

# Improvement: Light Load




vmware

# Improvement: Heavy load




- Test Setup:
- 200-280 VMs, 1vCPU/VM
- NUMA: 4 sockets
- Workload: fixed HTTP requests, 50 requests / VM / second

vmware

# Recap: NUMA aware I/O

- Hybrid design: low and high load
- Improvement:
  - higher throughput (Increased by 25%)
  - lower CPU (Reduced by 20%)
  - better VM consolidation (Up to 5% more VMs per host)
- Future work:  GPU, storage, RDMA

vmware

# Questions ?

vmware®