Non-Uniform Memory Access

Memory + Socket = NUMA node

Remote Memory Access

Local Memory Access

NUMA 0

NUMA 1
Non-Uniform Memory Access

Modern Server Architecture

NUMA 1

NUMA 2

App 1

App Memory

App 2

App Memory
Non-Uniform I/O Access
Non Uniform I/O Access

2 -socket Dell PE R720

4 -socket Dell PE R840
Intel DDIO

Classic

*Currently works with Local PCI Device
Non-Uniform I/O Access

Ideal Situation

Real Situation

Intel DDIO Benefits

Local Memory Access
DDIO benefits vs Local I/O : 1VM

L3 Cache Miss vs Packet Rate

Cache misses/packet

Packet Rate mpps

VM
I/O Thread
Packet Rate

0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8

0
5
10
15
20
25

All
Device and IO Thread
VM and I/O Thread
DDIO benefits vs Local I/O : 4VM

L3 Cache Miss vs Packet Rate

Packet Rate mpps

Cache misses/packet

Device and IO Thread

VM and I/O Thread

VM

I/O Thread

Packet Rate
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

VMs

HW

I/O Device

NUMA0

1

2

3

NUMA1

Thread

NUMA0

1 2 3 4

NUMA1

Affinity

Thread

I/O Device
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
Improvement: Light Load

Test Setup:
1VM, 1vCPU
NUMA: 2 sockets
Workload: UDP packet forwarding
I/O Device: Intel 10G Device

DPDK L2 Forwarding

<table>
<thead>
<tr>
<th>Packet Size</th>
<th>rxpps</th>
<th>fwdpps</th>
<th>Total CPU%/1M fwd pkts</th>
</tr>
</thead>
<tbody>
<tr>
<td>64Bytes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>256Bytes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>512Bytes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024Bytes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

%age improvement

Test Setup:
1VM, 1vCPU
NUMA: 2 sockets
Workload: UDP packet forwarding
I/O Device: Intel 10G Device
Improvement: Heavy load

HTTP Request Response Time, CPU utilization

- Default CPU
- NUMA I/O CPU
- Default RRT
- NUMA I/O RRT

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