Accelerating Para-Virtual I/O with CBDMA

JIAYU HU, INTEL
Para-Virtual I/O

- Para-virtual I/O is a virtualization technique to enhance VM I/O performance.

- VirtIO is a standard of para-virtual I/O, which consists of VirtIO front-end in VM and backend in hypervisor.

- Backend exchanges data with front-end via **copying packet buffers** between host and VM memory.

The overhead of **copying large bulk of data** makes the **backend** become the I/O **bottleneck**.
Crystal Beach DMA

- Crystal Beach DMA (CBDMA) is a **DMA engine** in the processor, which is extremely efficient in performing **memory copy operations**.
- No CPU intervention during data transfer.

**Challenges of using CBDMA to accelerate the backend:**
- NUMA
- Copy buffer length
- CPU-CBDMA cooperation pipeline
NUMA

• Influence from CBDMA and memory NUMA nodes

**NUMA – Local:**
CBDMA and Memory in same node.

**NUMA – Min Remote:**
CBDMA and SRC Memory in same node, DST memory in another node.

**NUMA – Max Remote:**
CBDMA and Memory in different nodes.

Higher is better

CPU and CBDMA in the same NUMA node.
• Influence from CBDMA and memory NUMA nodes

• CBDMA and memory in same NUMA node improves throughput 4% ~ 13%.

CPU and CBDMA in the same NUMA node.
Influence from CBDMA and CPU NUMA nodes

NUMA – Local: **CBDMA and CPU in same node.**

NUMA – Remote: **CBDMA and CPU in different nodes.**
NUMA

- Influence from CBDMA and CPU NUMA nodes

- CPU and CBDMA in the same NUMA node improves throughput 16% ~ 66%.

Memory and CBDMA in the same NUMA node.
Copy Length

CBDMA NUMA–Local: CBDMA copy & CBDMA and CPU in same node.

CBDMA NUMA–Remote: CBDMA copy & CBDMA and CPU in different nodes.
Copy Length

• CBDMA NUMA-local vs. CPU

• When lengths exceed 1024 B and 2048 B, CBDMA NUMA-local outperforms CPU.
Copy Length

- **CBDMA NUMA-remote vs. CPU**
  
  When lengths exceed **2048 and 3072 B**, CBDMA NUMA-remote outperforms CPU.
Copy Length

- **CBDMA vs. CPU**

  - When lengths are smaller than 1024 B, CPU outperforms CBDMA.

  - CBDMA achieves higher performance with larger copy lengths.
Solutions to Address Challenges

- **NUMA-aware resource assignment scheme**
  - Dynamically assign CPU, memory and CBDMA devices, according to resource status.
  - Working in progress.

- **Increase packet lengths** via enabling **TCP Segmentation Offload (TSO)** and **UDP Fragmentation Offload (UFO)**.
  - E.g. 1.5 KB → 64 KB TCP packets

- **Adaptive CPU-CBDMA Pipeline**
Adaptive CPU-CBDMA Pipeline

One enqueue/dequeue operation
Adaptive CPU-CBDMA Pipeline

- Copy pipeline

**CPU Pipeline**
- Read Descriptors → CPU copy buffers

**CBDMA Pipeline**
- Read Descriptor → Translate VA to PA
- Select CBDMA & Enqueue Requests
- Doorbell CBDMA to transfer data

Repeat for batching

exist unprocessed descriptors

After doorbell, **CPU and CBDMA operations execute concurrently.**
Adaptive CPU-CBDMA Pipeline

• Copy pipeline

**Decision Maker** decides which pipeline to use, according to packet information.

1. **Read Descriptors**
2. **Translate VA to PA**
3. **Select CBDMA & Enqueue Requests**
4. **Doorbell CBDMA to transfer data**
5. **Wait CBDMA copy Completion**

**Repeat for batching**

exist unprocessed descriptors

---

Copy pipeline

CPU Pipeline

CBDMA Pipeline
Adaptive CPU-CBDMA Pipeline

- Decision maker

- **Min_Len_NL** is the minimal length that CBDMA outperforms CPU, when in the same NUMA node.

- **Min_Len_NR** is the minimal length that CBDMA outperforms CPU, when in different NUMA nodes.
Adaptive CPU-CBDMA Pipeline

- Set `Min.Len.NL` and `Min.Len.NR` to 2048 and 3072 bytes.

---

CBDMA and CPU core in the same NUMA node?

- Yes
  - Buffer Length > **2048 B**
    - Yes
      - CBDMA Pipeline
    - No
      - No

- No
  - Buffer Length > **3072 B**
    - Yes
      - CPU Pipeline
    - No
      - No
Experiment

<table>
<thead>
<tr>
<th>CPU</th>
<th>Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testpmd Information</td>
<td>1 core</td>
</tr>
<tr>
<td>VM Information</td>
<td>4 cores pre VM 1GB Huge-page Enable TSO 1 queue</td>
</tr>
<tr>
<td>Iperf</td>
<td>TCP packet size is 64 KB</td>
</tr>
<tr>
<td>CPU cores, CBDMA and memory locate in NUMA node 0.</td>
<td></td>
</tr>
</tbody>
</table>

- Use 1 GB huge-page to mitigate the address translation overhead, i.e. GPA to HPA.
Throughput (Gbps)

CBDMA Device Number

Results

CBDMA – 4 KB MBUF: Testpmd mbuf size is 4 KB.

CBDMA – 8 KB MBUF: Testpmd mbuf size is 8 KB.
Results

• Using 4 KB mbuf, CBDMA improves throughput up to 30%.
• Using 8 KB mbuf, CBDMA improves throughput up to 37%.
Thanks

jiayu.hu@intel.com