On the Impact of Cluster Configuration on RoCE Application Design

Yanfang Le
Mojtaba Malekpourshahraki
Brent Stephens
Aditya Akella
Michael M. Swift

mmalek3@uic.edu
Designing and deploying high performance applications

• How to design and deploy applications to have
  • Low tail latency
  • High throughput
  • Low CPU load

Remote Direct Memory Access (RDMA)
RDMA overview

Zero copy

Kernel Bypass

Protocol Offload

RDMA cluster an applications need to be configured
Key question

How should I configure my RDMA cluster and application to achieve best performance?
How to transmit data?

How to handle incoming events?

Should applications be co-located?

Are NICs fair among all applications?

Can I use jumbo frames?
Design configurations – sending data options

• Design configurations
  • **How to transmit data?**
  • How to handle incoming events?

---

**Pros**
- One network round trip regardless to the structure

**Cons**
- Transfer just a single memory value

---

Data transfer that needs two memory access (e.g., hash table)
Design configurations – sending data options

• Design configurations
  • **How to transmit data?**
  • How to handle incoming events?

![Diagram showing READ-2 process](image)

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only local CPU is involved</td>
<td>Needs multiple round trip</td>
</tr>
</tbody>
</table>

What if network is congested?
Design configurations – sending data options

• Design configurations
  • **How to transmit data?**
  • How to handle incoming events?

Combine WRITE/SEND or SEND/SEND
Design configurations – sending data options

- Design configurations
  - **How to transmit data?**
  - How to handle incoming events?

![Diagram showing data transmission between CPU and RNIC](image)

**Pros**
- Only one network transfer is required

**Cons**
- Both local and remote CPU are involved

What if remote CPU is busy?

- PCIe PIO operations
- DMA operations
- RDMA data packets
- ACKs of data packets
Design configurations – Interrupt vs polling

• Design configurations
  • How to transmit data?
  • How to handle incoming events?

Pros
- No CPU overhead for detecting arrivals

Cons
- Slower reaction to the arrival events
Design configurations – Interrupt vs polling

• Design configurations
  • How to transmit data?
  • **How to handle incoming events?**

![Diagram showing CPU, Completion Queue, and Memory connections]

**Pros**
- Fast in reacting to a received data

**Cons**
- CPU is busy for polling memory or CQ
Configurations

Decisions

- How to transmit data?
- How to handle incoming events?

Deployment Configurations

- Should applications be co-located?
- Are NICs fair among all applications?
- Can I use jumbo frames?
Deployment configurations

- Deployment configurations
  - Should applications be co-located?
    - Applications competing on CPU
    - RDMA applications competing on RNIC
  - Can I use jumbo frames?
Choosing the best configuration

• Each configure option has its **pros and cons**
• There are a **lot of options to configure** a cluster
• It is **difficult to choose** the best configure

We need an experiment to compare all different possible configurations
We evaluate RDMA performance to answer to these questions

<table>
<thead>
<tr>
<th>Decisions</th>
<th>Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>How to transmit data?</td>
<td>What is the best RDMA verbs?</td>
</tr>
<tr>
<td>How to handle incoming events?</td>
<td>Should we use polling or interrupts?</td>
</tr>
<tr>
<td>Should applications be co-located?</td>
<td>What is the effect of applications competing for CPU on RDMA?</td>
</tr>
<tr>
<td>Are NICs fair among all applications?</td>
<td>What is the effect of applications competing for RNIC on RDMA?</td>
</tr>
<tr>
<td>Can I use jumbo frames?</td>
<td>What is the effects of frame size on performance?</td>
</tr>
</tbody>
</table>
Methodology

- **Cloud Lab Topology**
  - A cluster of 17 servers
    - 8-core ARM Cortex-A57 processor
    - 64 GB of memory
    - 10 GbE Mellanox ConnectX-3 NIC
  - Ethernet switch
    - HP 45XGc

- **Traffic patterns**
  - 16 connections for each server
  - No congestion in the switch

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inlined message size</td>
<td>256 B</td>
</tr>
<tr>
<td>Largest RDMA packet</td>
<td>4 KB</td>
</tr>
<tr>
<td>Transport (WRITEs, SENDs)</td>
<td>Unconnected</td>
</tr>
<tr>
<td>Total experiment time</td>
<td>10 seconds</td>
</tr>
</tbody>
</table>
Methodology

• Test cases
  • READ-Intr
  • READ-Poll
  • WRITE/SEND-Poll
  • SEND/SEND-Intr
  • READ-Intr-$n$
What to measure?

• Three scenarios
  • Dedicated server for applications
  • Co-located applications competing on CPU
  • Co-located applications competing on RNIC

• RDMA performance in the terms of
  • Latency
  • Throughput

What is the effect of cluster configuration on RDMA performance
Which is the best verb?

- Scenario I: No contention

Best verb changes when the payload size changes
One side vs two side

- Scenario I: No contention

WRITE/SEND-poll $\equiv$ READ-Intr

WRITE/SEND-poll $>$ READ-Intr

One side verbs are better
Effect of RDMA verbs on throughput

- Scenario I: No contention

Correct verb size depends on the size of the payload
Effect of RDMA READs on throughput

- Scenario I: No contention

Interrupts or polling has little impact on throughput (No CPU contention)
Application needs two memory access

- Scenario I: No contention

Best verb changes when the payload size changes

WRITE/SEND-poll < READ-Intr-2

WRITE/SEND-poll > READ-Intr-2
Should I use jumbo frames?

- Scenario I: No contention

Stable latency for small messages

Jumbo frames should not be used because they increase the latency
Effect of CPU contention on performance

Scenario II: CPU contention

CPU contention changes the best verb choice

READ-poll $\cong$ READ-intr

READ-poll $>$ READ-intr
RNIC fairness among short and long messages

Scenario III: RNIC contention

- Two traffic patterns (Facebook):
  - Storage (4 MB)
  - memcached

RNIC contention hurts the throughput of small messages for all verbs
### Difficult to decide what is best configuration

<table>
<thead>
<tr>
<th>Decisions</th>
<th>Measurements</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>How to transmit data?</td>
<td>What is the best RDMA verbs?</td>
<td>Choosing the best performing verb is difficult</td>
</tr>
<tr>
<td>How to handle incoming events?</td>
<td>Should we use polling or interrupts?</td>
<td>Correct choice depends on the CPU load</td>
</tr>
<tr>
<td>Should applications be co-located?</td>
<td>What is the effect of applications competing for CPU on RDMA?</td>
<td>CPU contention changes the best verbs</td>
</tr>
<tr>
<td>Are NICs fair among all applications?</td>
<td>What is the effect of applications competing for RNIC on RDMA?</td>
<td>RDMA applications do not fairly share the RNIC</td>
</tr>
<tr>
<td>Can I use jumbo frames?</td>
<td>What is the effects of frame size on performance?</td>
<td>Never use Jumbo frames because it increases latency</td>
</tr>
</tbody>
</table>

Design Configurations

Deployment Configurations

Observations

Measurements

Results
Observations

• Correct verb choice depends on:
  • Object size
  • CPU utilization
  • RNIC utilization
  • The number of dependent memory accesses

Impossible for application to monitor and measure and adjust verbs
• Needs of a higher-level library
  • Hide complexities from applications
  • Monitor important system parameters
  • Configure the verbs automatically
Conclusion

• We studied
  • The impact of co-located applications contending for RNIC and CPU
  • The impact of large frame sizes hurt tail latency

• We observed
  • Correct verb choice is dependent on many variables
  • It's impossible for application to monitor and measure all variables

• Future work
  • Implement the library

A higher-level library that hides complexities from applications is needed
Thanks for the attention

Mojtaba Malekpourshahraki

Email: mmalek3@uic.edu
Website: cs.uic.edu/~mmalekpo