A Communication Model for Small Messages with InfiniBand

Torsten Höfler, Wolfgang Rehm
TU Chemnitz

23.06.2005
Introduction

Motivation

Previous Work

InfiniBand Specialities

A new Model

Architectural Considerations

The LoP Model

Measuring the Parameters

Results and Conclusion

Modeling Results

Conclusions
A Communication Model for Small Messages with InfiniBand

Outline

1 Introduction
   • Motivation
   • Previous Work
   • InfiniBand Specialities

2 A new Model
   • Architectural Considerations
   • The LoP Model
   • Measuring the Parameters

3 Results and Conclusion
   • Modeling Results
   • Conclusions
Motivation

- advantages of a model
  - proof a lower bound to a problem
  - understand architectural details

⇒ models have to be very accurate
  - why InfiniBand?
    - state of the art technology
    - offloading based network

⇒ special model for offloading based networks
  - Optimizing Barriers?
    - InfiniBand Barrier is well tuned (Panda et. al.)
    - others are optimal in abstract models (Finkel et. al.)
Motivation

- advantages of a model
  - proof a lower bound to a problem
  - understand architectural details

⇒ models have to be very accurate
  - why InfiniBand?
    - state of the art technology
    - offloading based network

⇒ special model for offloading based networks
  - Optimizing Barriers?
    - InfiniBand Barrier is well tuned (Panda et. al.)
    - others are optimal in abstract models (Finkel et. al.)
Motivation

- advantages of a model
  - proof a lower bound to a problem
  - understand architectural details

⇒ models have to be very accurate
- why InfiniBand?
  - state of the art technology
  - offloading based network

⇒ special model for offloading based networks
- Optimizing Barriers?
  - InfiniBand Barrier is well tuned (Panda et. al.)
  - others are optimal in abstract models (Finkel et. al.)
Motivation

- advantages of a model
  - proof a lower bound to a problem
  - understand architectural details
- models have to be very accurate
  - why InfiniBand?
    - state of the art technology
    - offloading based network
- special model for offloading based networks
  - Optimizing Barriers?
    - InfiniBand Barrier is well tuned (Panda et. al.)
    - others are optimal in abstract models (Finkel et. al.)
Motivation

- advantages of a model
  - proof a lower bound to a problem
  - understand architectural details

⇒ models have to be very accurate
- why InfiniBand?
  - state of the art technology
  - offloading based network

⇒ special model for offloading based networks
- Optimizing Barriers?
  - InfiniBand Barrier is well tuned (Panda et. al.)
  - others are optimal in abstract models (Finkel et. al.)
Outline

1. Introduction
   - Motivation
   - Previous Work
   - InfiniBand Specialities

2. A new Model
   - Architectural Considerations
   - The LoP Model
   - Measuring the Parameters

3. Results and Conclusion
   - Modeling Results
   - Conclusions
Known Models

- PRAM, $C^3$, BSP are too inaccurate (→ paper)
- LogP as base model
  - L - Hardware latency
  - o - Processor overhead
  - g - gap between consecutive messages
  - P - number of processors
1 Introduction
   • Motivation
   • Previous Work
   • InfiniBand Specialities

2 A new Model
   • Architectural Considerations
   • The LoP Model
   • Measuring the Parameters

3 Results and Conclusion
   • Modeling Results
   • Conclusions
InfiniBand Specialities

- user-level communication
- requests are queued in hardware
- HCA fetches a request from the top of the queue
- application is notified in Completion Queue (CQ)
- CQ can be shared between different connections
- different possibilities for sending Data (SEND, RDMA, Reliable, Unreliable ...)

Torsten Höfler, Wolfgang Rehm TU Chemnitz
A Communication Model for Small Messages with InfiniBand
Outline

1. Introduction
   - Motivation
   - Previous Work
   - InfiniBand Specialities

2. A new Model
   - Architectural Considerations
   - The LoP Model
   - Measuring the Parameters

3. Results and Conclusion
   - Modeling Results
   - Conclusions
RTT Model

- three sections → NIC warmup, maximum, saturation
- warmup → $t_{pipeline} = \frac{\lambda_1}{\lambda_2 + p}$
- maximum → $t_{processing} = \lambda_3$
- saturation → $t_{saturation} = \lambda_4 \cdot (1 - e^{\lambda_5 \cdot (p - \lambda_6)})$
Overhead Model

- cache and pipelining on the host-cpu
- pipeline startup: \( t_{ov}(\lambda_1...3) = \lambda_1 + \frac{\lambda_2}{\lambda_3 + p} \)

\[ t \]

\[ t_{\text{min}} \]

\[ p \]
Introduction
- Motivation
- Previous Work
- InfiniBand Specialities

A new Model
- Architectural Considerations
- The LoP Model
- Measuring the Parameters

Results and Conclusion
- Modeling Results
- Conclusions
The LoP Model

- model every possible Transport Type separately
- HCA offers additional level of parallelism
- new possibilities for overlapping
- implicit parallelism on the HCA proposed by IBA standard
LoP Problems

- $h$ parameter cannot be measured directly
- Linear model for $g$ is not appropriate
- $h$ is modeled as part of the $L \rightarrow L(p)$
- Architectural assumptions are used to model RTT
1 Introduction
   • Motivation
   • Previous Work
   • InfiniBand Specialities

2 A new Model
   • Architectural Considerations
   • The LoP Model
   • Measuring the Parameters

3 Results and Conclusion
   • Modeling Results
   • Conclusions
Parametrization

- $o_s(p)$ - time to complete VAPI_post_sr()
- $o_r(p)$ - time to complete VAPI_post_rr()
- $L(p) = \frac{RTT(p)}{2} - (p \cdot o_s(p) + o_s(1))$

Sender

IBA

Receiver

```
take_time(t0);
p * VAPI_post_rr();
take_time(t1);
p * VAPI_post_sr();
take_time(t2);
p * VAPI_poll_cq();
take_time(t3);
p * VAPI_poll_cq();
take_time(t4);
```
### Outline

1. Introduction
   - Motivation
   - Previous Work
   - InfiniBand Specialities

2. A new Model
   - Architectural Considerations
   - The LoP Model
   - Measuring the Parameters

3. Results and Conclusion
   - Modeling Results
   - Conclusions
RDMA $o_s(p)$ Results

\[ t_{srov}^{rdmaw,n}(p) = 0.6 + \frac{0.2}{-0.8+p} \]
RDMA $RTT(p)$ Results

$$t_{\text{rtt, min}}^{\text{rdmaw, n}}(p) = 4.5 + \frac{16.8}{0.01+p} + 4.5 \cdot (1 - e^{-0.06 \cdot (p-12.9)})$$
Deriving the Hardware Latency

\[
L_{\text{send},n}^{\text{min}}(p) = \frac{t_{\text{send},n}^{\text{rtt},\text{min}}(p)}{2} - \left( t_{\text{sr},\text{ov}}^{\text{send},n}(1) \right) - \left( t_{\text{sr},\text{ov}}^{\text{send},n}(p) \right)
\]
Introduction

Motivation

Previous Work

InfiniBand Specialities

A new Model

Architectural Considerations

The LoP Model

Measuring the Parameters

Results and Conclusion

Modeling Results

Conclusions
Conclusions

- analysis of small messages performance for IBA
- development of a new very accurate model
- LogP is quite accurate for saturated networks
- LoP offers different optimization chances
- e.g. sending more than one message together
- \( \Rightarrow \) optimized barrier \( \rightarrow \) 40% speedup
Future Work

- analyze different algorithms in the LoP context
- simplification of the LoP model
- expansion to arbitrary message sizes
- evaluation for different offloading based networks
Questions/Comments?