Communication Models for Resource Constrained Hierarchical Ethernet Networks

Speaker: Konstantinos Katrinis#

Jun Zhu+, Alexey Lastovetsky*, Shoukat Ali#, Rolf Riesen#

+ Technical University of Eindhoven, Netherlands
* University College Dublin, Ireland
# Dublin Research Laboratory, IBM, Ireland
Outline

• Introduction
• Related work
• Network properties
• Communication model
• Experiments
• Conclusion
Introduction

• **Cost effective yet powerful computer cluster**
  – COTS computers: multi-core to many-core
  – Ethernet vs. custom interconnects
  – **Shared** resources: network and memory
  – Open-source software stack: Linux and OpenMPI

• **Concerns in cluster-based parallel computing**
  – Computers are tightly coupled
  – **Communication models** are non-trivial
• Two star-configured racks connected via backbone
• Communication contention happens on different levels
  – Network interface cards (NICs)
  – Backbone cable
• Communication **times prediction** is hard yet important
Goals and Contributions

- To derive network properties on parameterized network topology from simultaneous point-to-point MPI operations
- Our work is the first effort to discover the asymmetric network property on TCP layer for concurrent bidirectional communications
- To propose communication models for concurrent communications in resource-constrained Ethernet clusters
- We show that the communication time predictions become significantly less accurate, if the asymmetric network property is excluded from the model
Related Work

No network contention

- Hockney model [PMPC 94]- point-to-point communication time for a message with size $m$ is: $a + m*b$, where $a$ is latency and $b$ inversed bandwidth

- Similar models: LogP [Culler 93] for small messages and LogGP [Hoefler 06]

Network contention-aware

- A recent communication model [Martinasso 11] considers NIC level contention for InfiniBand clusters

Our proposed model for Ethernet clusters, with

- NIC and backbone levels contention-aware
- Asymmetric communication property - from benchmarking
MPI Micro-benchmark

**sender process**

```c
for i := 0 to (maxIter-1)  
  // Message 'msg' initialization
  ...

  // Synchronization
  MPI_Barrier()

  // Sending a message
  MPI_Send(&msg, msgSize, rankRecv, id0);
```

**receiver process**

```c
for i := 0 to (maxIter-1)  
  // Pre-post receive
  MPI_Irecv(&msg, msgSize, rankSend, id0, request);

  // Timing the communication operation
  MPI_Barrier()
  t0 := MPI_Wtime();
  MPI_Wait(request, status);
  tArray[i] := MPI_Wtime() - t0;

  /* If the estimation in the first 'i+1' elements of tArray
     indicate enough statistical reliability, exit the loop */
  if (isStatisticallyReliable(tArray, i+1) )
    i := maxIter;

  // Synchronize the value of 'i'
  MPI_Recv(&i, 1, rankRecv, id1);
```

- Point-to-point MPI benchmarking
- A 95% confidence level of averaged timings
- Setup for any given number of simultaneous communications
• Up to 15 nodes (RHEL 5.5 x86-64) in each rack
  o Dual-socket six-core (Intel Xeon X5670 6C@2.93GHz)
  o 1Gb NIC tuned, ToR IBM BNT Rack Switch G8264 1-10Gb
• OpenMPI 1.5.4 as the MPI Implementation
• Large message sizes (10MB) in benchmarking
Network Property - Fairness

To set **unidirectional communication** for $|E|$ number of point-to-point MPI operations in testbed

A. Intra-rack communication: sender on the same node
B. Inter-rack communication: sender on different nodes

We expect

- Bandwidth is **fairly distributed** over all links
- In experiment B, when $|E|$ is bigger enough, the bandwidth of the backbone may **saturate**
Network Property – Fairness (contd.)

Formal model:

\[ \beta_{a,b} = \begin{cases} 
\beta \cdot |E|, & \text{if } \beta = \beta_O \text{ and } |E| > 10 \text{ or } \beta = \beta_E \\
\beta_E, & \text{if } \beta = \beta_O \text{ and } |E| \leq 10 
\end{cases} \]

Verified properties for unidirectional communication

- **Fairness**
- **Network saturation**

Fig. Average bandwidth of unidirectional logical links on an optical backbone
Network Property - Asymmetric

- To study bidirectional communication, we swap the mapping policy for some of the sender and receiver processes in the previous experiments.

- We expect the previous properties hold, i.e. fairness and network saturation.

- However, an asymmetric property appears, which has not yet been reported in the literature.

- Iperf has been used to verify the property, and we double-check in a different Ethernet cluster in HCL laboratory in UCD.
Network Property – Asymmetric (contd.)

For instance, when $\delta + (\cdot) = 2$ and $\delta - (\cdot) = 1$, i.e. two incoming and one outgoing links

- The outgoing link should get 940Mbps bandwidth, according to a fair dynamic bandwidth allocation in full
- However, it gets 470Mbps, the same as incoming links

Fig. Average bandwidth for bidirectional logical links on a NIC

Formal model:

$$\beta_{a,b} = \begin{cases} 
\beta \cdot \delta_{max}(\cdot), & \text{if } \beta = \beta_O \text{ and } \delta_{max}(\cdot) > 10 \text{ or } \beta = \beta_E \\
\beta_E, & \text{if } \beta = \beta_O \text{ and } \delta_{max}(\cdot) \leq 10
\end{cases}$$
Communication Model

To find the most congested logical link (Eq. A-5, Appendix) to be analyzed (Line 5)

Is congestion caused by with-flow traffic (Line 7)?

YES  NO

To find the with-flow bottleneck network resource (Line 9)  To find the contra-flow bottleneck network resource (Line 13)

Is contra-flow physical link is saturated (Line 16)?

YES  NO

To derive logical link bandwidth based on fairness property (Line 10, 26-32)  To derive logical link bandwidth based on asymmetric property (Line 16-17)

Are all logical links analyzed? (Line 3)

NO  YES

END

Insert the logical link into queue to be analyzed, and reset its contra-flow congestion factor to NULL (Line 19-20)
The communication times depend on message sizes and the derived communication bandwidth of logical links, as in [Martinasso 11].

- the bandwidth of logical links may be redistributed dynamically.

- The predicted communication time $T_{a,b}$ for each communication operation is calculated until all logical links are analyzed.
Experiments

- Cluster has been configured with 1 GbE for intra-rack and 10 GbE for inter-rack communication
- Each time the same number of nodes are configured in both racks, with a total nodes $|N|$ up to 30

![Diagram](image)

Figure 6. The communication patterns of two test instances, when $|N| = 10$ and $d = 3$. 
Experimental Results

- Fig. Histogram of times prediction errors.
- 9 experiments with a set of values for parameters $|N|$ and $d$
- A total of 354 randomly generated communication patterns are tested
- The prediction error with pure fairness property: can be as worse as $-80\%$, i.e. predicted times are 5 times lower than the measured ones
- Our model is quite accurate: worst averaged 9.5\%, and much better worse case ($-50\%$, no more than 2 times difference)
Conclusion & Future Work

Conclusion:

• We derive an ‘asymmetric network property’ on TCP layer for concurrent bidirectional communications on Ethernet clusters.

• We develop a communication model to characterize the communication times on resource constrained networks accordingly.

• We conduct statistically rigorous experiments to show that our model can be used to predict the communication times for simultaneous MPI operations effectively, only when asymmetric network property is considered.

Conclusion:

• As the future work, we plan to generalize our model for more complex network topologies.

• On the other hand, we would also like to investigate how the asymmetric network property can be tuned below TCP layer in Ethernet networks.
Thank you!

Questions?