A Non-Aligned Perspective - Battle of the Network Stars SC2003 Panel

Anthony Skjellum

University of Alabama at Birmingham,
Dept. of Computer and Information Sciences
tony@cis.uab.edu

and

MPI Software Technology, Inc

November 17, 2003

Preface

- I gave an oral presentation at the panel
- These slides are the retrospective
- I’ve covered all the topics from my oral presentation in the slides
- I’ve put some further thoughts in the “Epilog”

How important is RDMA for cluster networks?

- It is very important… the ability to hide communication behind computation is essential to improving net performance
- RDMA does not currently optimize…
  - small messages (copies needed)
  - very large messages (kernel intervention needed)
- RDMA is for medium-size messages (e.g., .1 – 10 Megabytes)
- RDMA, when successful, offers a way to hide an inherent overhead of parallel computing
- RDMA can be reduced in value by polling MPI implementations (like MPICH and LAM)

MPI Implementer's view, I

- First, we've implemented MPI-1.2 and MPI-2.1 for all the networks mentioned … e.g., ChaMPIon/Pro and MPI/Pro
- All of them work well with MPI-1, some support MPI-2 one-sided nicely also
- The question is, which make it easier for MPI to deliver net performance to applications
- The trend here is that offload engines are helpful (more on that to come)

MPI Implementer's view, II

- Commercial MPI users in clusters – from our viewpoint – are about 50% ethernet based
- Mostly these ethernet users are gigabit users now, some *still* use 100 Mbit/s
- Logically, these and other users will find it helpful to have continuing choice of Ethernet moving forward… and many will select 10 GigE
- In the "real world" there are Linux clusters, and Windows clusters, plus "the others" like MacOS

MPI Implementer's view, III

- Cost of purchase of the "fast networks" like Myrinet, IB, and Quadrics largely irrelevant
- Let's talk about cost of ownership
- What controls cost of ownership
  - Ease of use
  - Reliability
  - Driver quality ☺
- Giganet was a network, which, for its time, had good TCO … drivers still work years after the company stopped supporting them!

Facts about HPC Procurement and Rating

- The HPC world is controlled by two benchmarks
  - Ping-pong (measures latency)
  - HPL (measures a mixture of features)
- High speed SAN networks give support for RDMA and zero-copy semantics
- Neither benchmark reveals these powers directly, NOR the ability to overlap communication and computation

The “Microsecond” Marketing Lie

- One of the “ways to fool the masses”
- You can’t determine the value of an MPI on a network with X microseconds latency and Y Mbyte/s, unless you know the CPU overhead
- If there’s no CPU left for your application, the performance of the microbenchmark may have little to say about your real results
- MPI’s and network drivers designed to optimize ping-pong may really hurt application performance

Trends, I – SAN Refinements

- The “two level multicomputer” model of Seitz et al is represented in Myrinet and Quadrics
- This will continue to be good for many reasons if the MPI load can be moved, and attached MPI’s don’t poll in order to use the offload engines
- Quadrics and Myrinet are trending toward this model strongly
- CPU overhead reductions, not latency reductions will win users of these networks

Trends, II – 10GigE TOE

- 10GigE with TCP offload engines will be very cheap and probably made by companies that currently offer SANs
- We want to lower overhead, but we may be able to tolerate Ethernet one-way latencies in many applications especially as switches are optimized
- 10GigE plus load CPU overhead will be an important player, and possibly totally displace InfiniBand and others in the commercial space
- Extremely low overhead, low latency and high bandwidth interconnects will still persist, but probably at 10-50x the cost of 10GigE in its “prime”

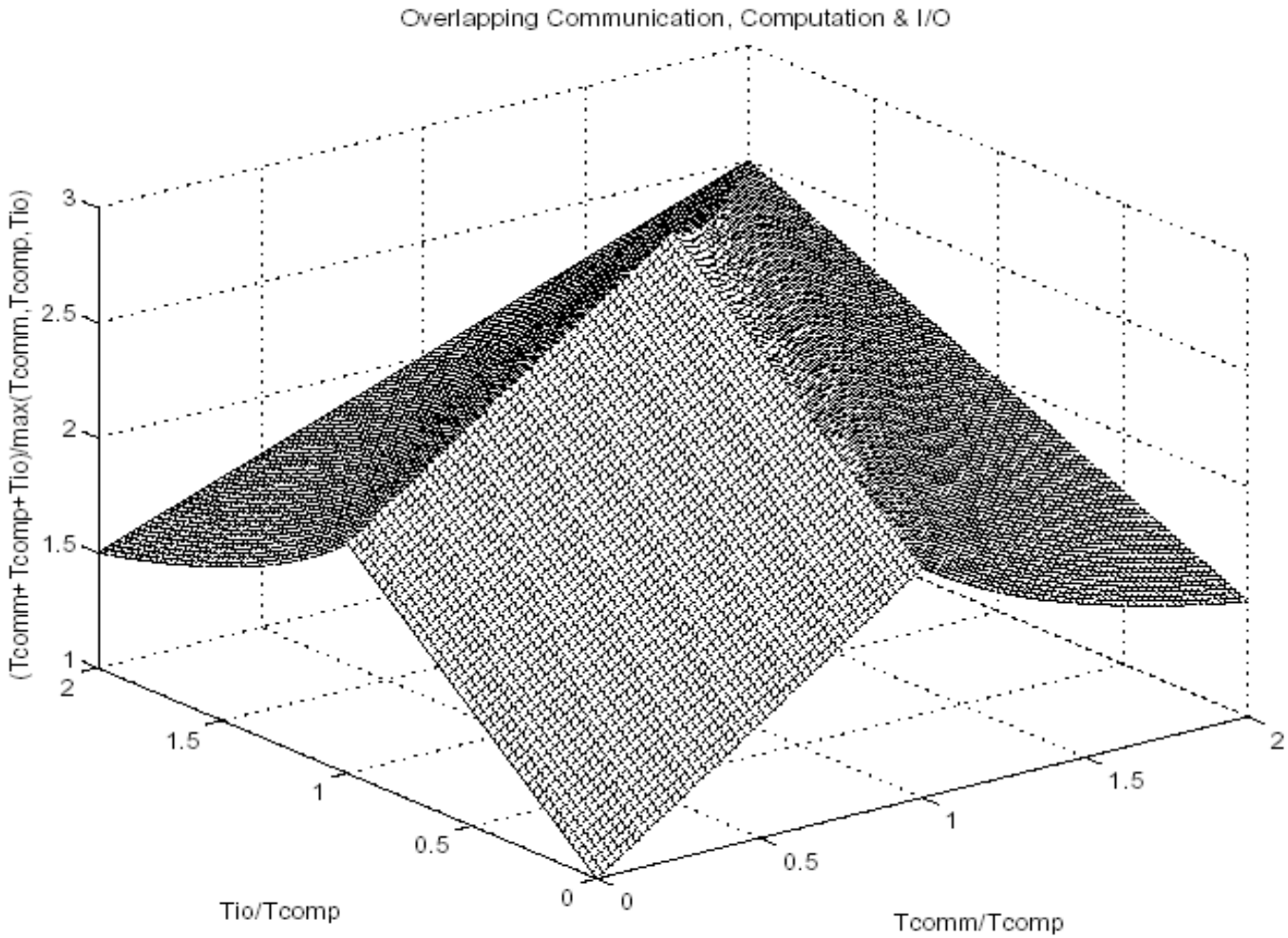
Fair way to rate the cost of a SAN vs. Commodity 1Gig Ethernet

- What is the incremental reduction in time to solution of your application (remember to use an MPI that actually rewards low overhead networks)
- The cost vs. benefit ratio can then be computed versus the GigE baseline
- This incremental benefit can be measured against the capital cost of more cluster equipment or faster cluster equipment, depending on how your application scales (or doesn't scale)
- No absolute performance is interesting, relative to baseline performance is fair
- You can fairly compare two networks for your applications by looking at the absolute time to solution and also the cost/benefit computation
- If you have many applications, or expect application mix to change during your cluster's lifetime, account for that in weights
- Any reference to ping-pong and HPL dilutes the key argument

EPILOG

There is significant work on studying overlap of communication, computation, and I/O… at an MPI level and with SANs, not just low latency and high bandwidth

Up to 3x speedup…



Where to learn more

- The papers documenting work on overlap in MPI/Pro and ChaMPIon/Pro are at
  - http://www.mpi-softtech.com/company/publications
  - And under HPC Lab at UAB
  - http://ardra.hpcl.cis.uab.edu/publications
- There are also unpublished PhD dissertations addressing this further… such as
  - Boris Protopopov's PhD dissertation… http://library.msstate.edu/etd/show.asp?etd=etd-06272003-120226

Disclaimer

- These opinions are my own, not those of MPI Software Technology, Inc nor the University of Alabama…