Measuring MPI Send and Receive Overhead and Application Availability in High Performance Network Interfaces

Douglas Doerfler

Ron Brightwell

Sandia National Laboratories

Euro PVM/MPI 2006

Sandia National Laboratories

Motivation

- Overhead and availability are important measurements
- Lack of open-source benchmarks that measure overhead and availability
- Continued evaluation of modern interconnects
  - SeaStar
  - InfiniPath

Methodology

- Use post-work-wait loop with non-blocking send and receive calls
- Overhead
  - Length of time that a processor is engaged in the transmission or reception of each message
- Availability
  - Fraction of total transfer time that the application is free to perform non-MPI work

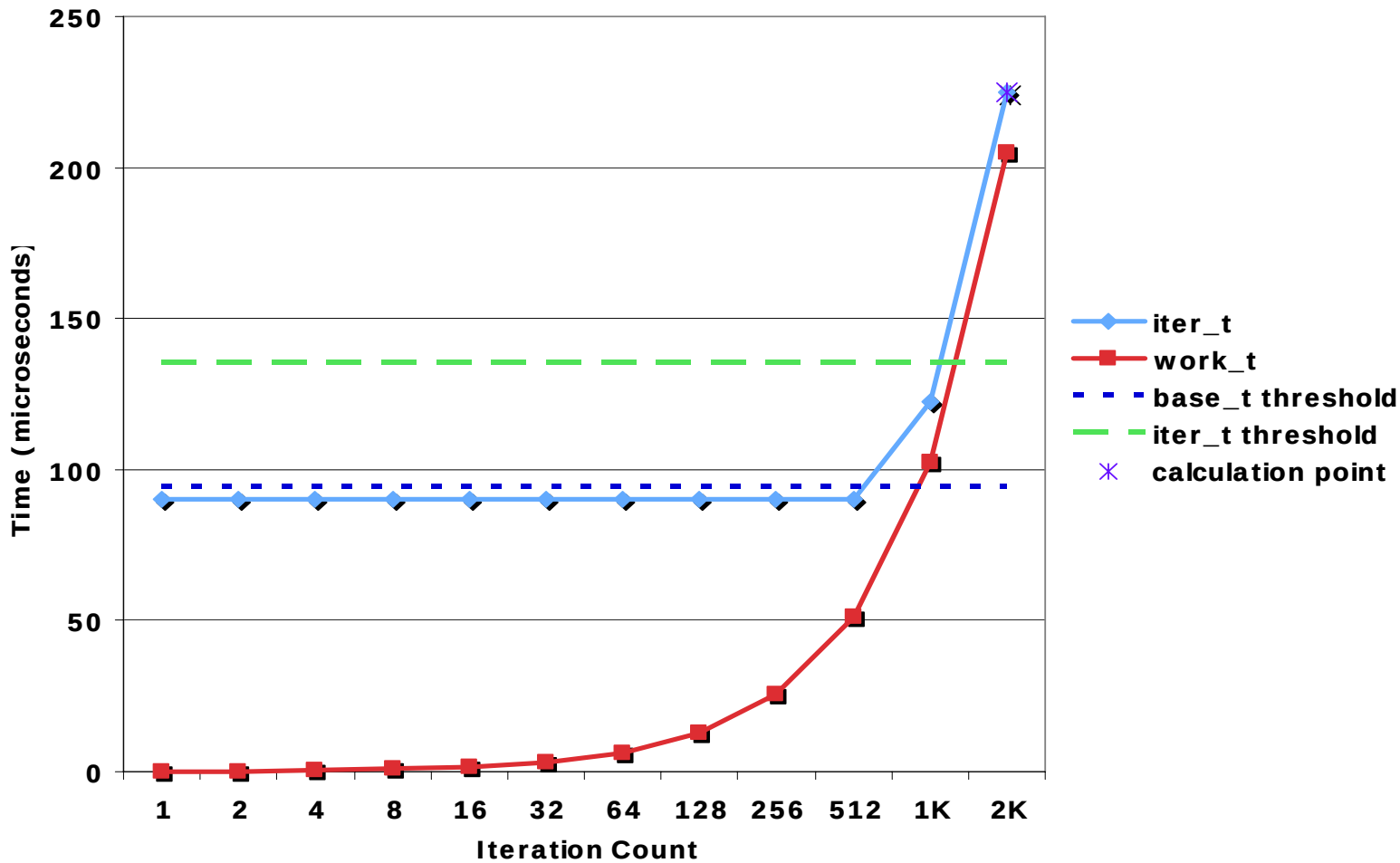
Determining Overhead

- For each iteration of the post-work-wait loop
  - Initiate communication
  - Perform work
  - Complete communication
  - Increase work
- Loop time
  - Time required to perform work plus time for communication
- Overhead
  - Subtract time to perform only work

Message Transfer Time

- Loop time before work interval becomes dominant factor
- To get accurate estimate of transfer time
  - Loop values are accumulated and averaged
  - Only values measured before work interval starts contribute to loop time
  - Values used in average calculation are determined by comparing iteration time to a given threshold
  - Threshold must be sufficiently long to avoid premature stop in accumulation of values used for average calculation
  - Our method does not automatically determine threshold
- Transfer is defined by MPI semantics

MPI Overhead Method



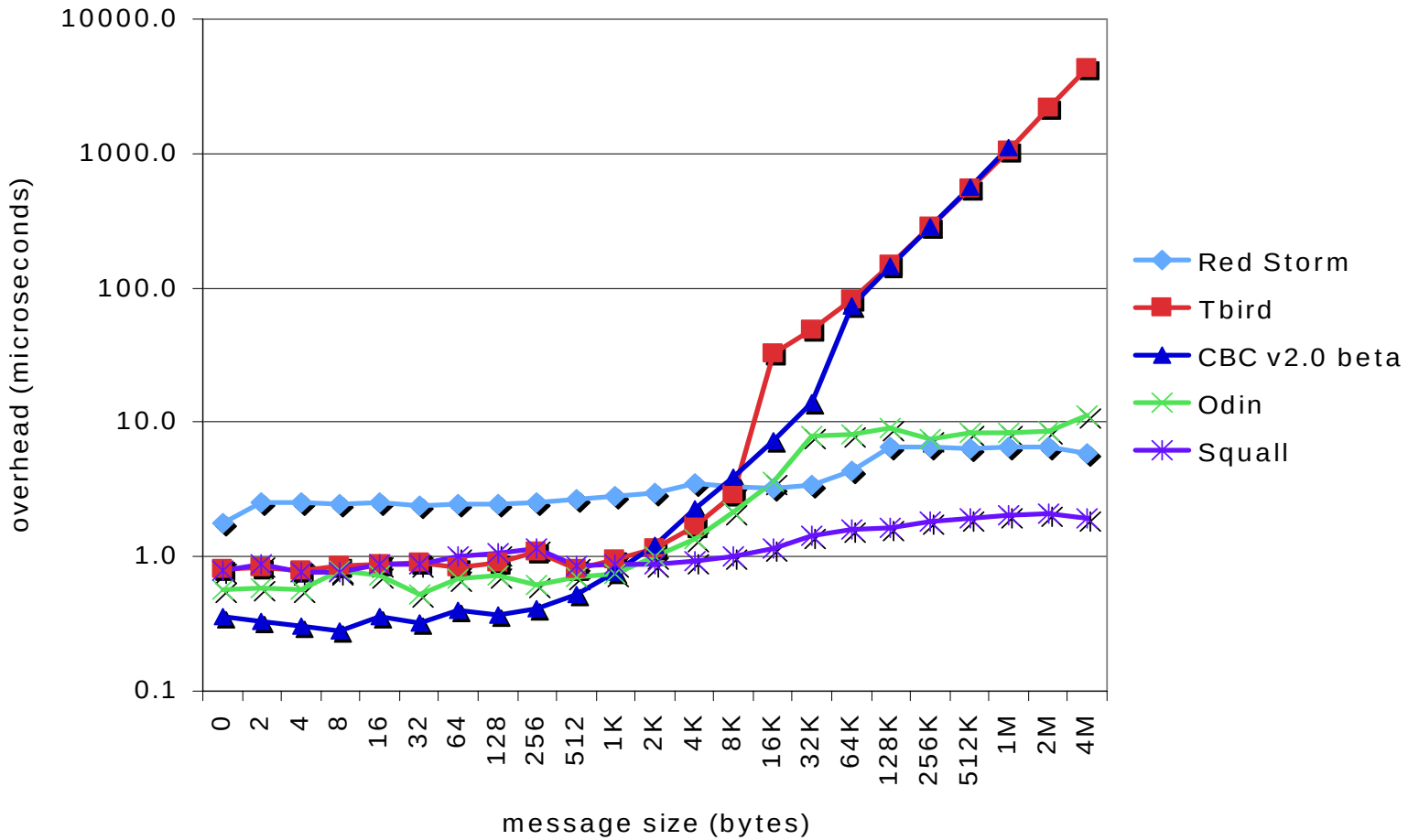
Sandia National Laboratories

Platforms

	Red Storm	*Thunderbird*	*CBC-B*	*Odin*	*Red Squall*
Interconnect	**Seastar 1.2**	**InfiniBand**	**InfiniBand**	**Myrinet 10G**	**QsNetII**
Manufacturer	**Cray**	**Cisco/Topspin**	**PathScale**	**Myricom**	**Quadrics**
Adaptor	**Custom**	**PCI-Express HCA**	**InfiniPath**	**Myri-10G**	**Elan4**
Host Interface	**HT 1.0**	**PCI-Express**	**HT 1.0**	**PCI-Express**	**PCI-X**
Programmable coprocessor	**Yes**	**No**	**No**	**Yes**	**Yes**
MPI	**MPICH-1**	**MVAPICH**	**InfiniPath**	**MPICH-MX**	**MPICH QsNet**

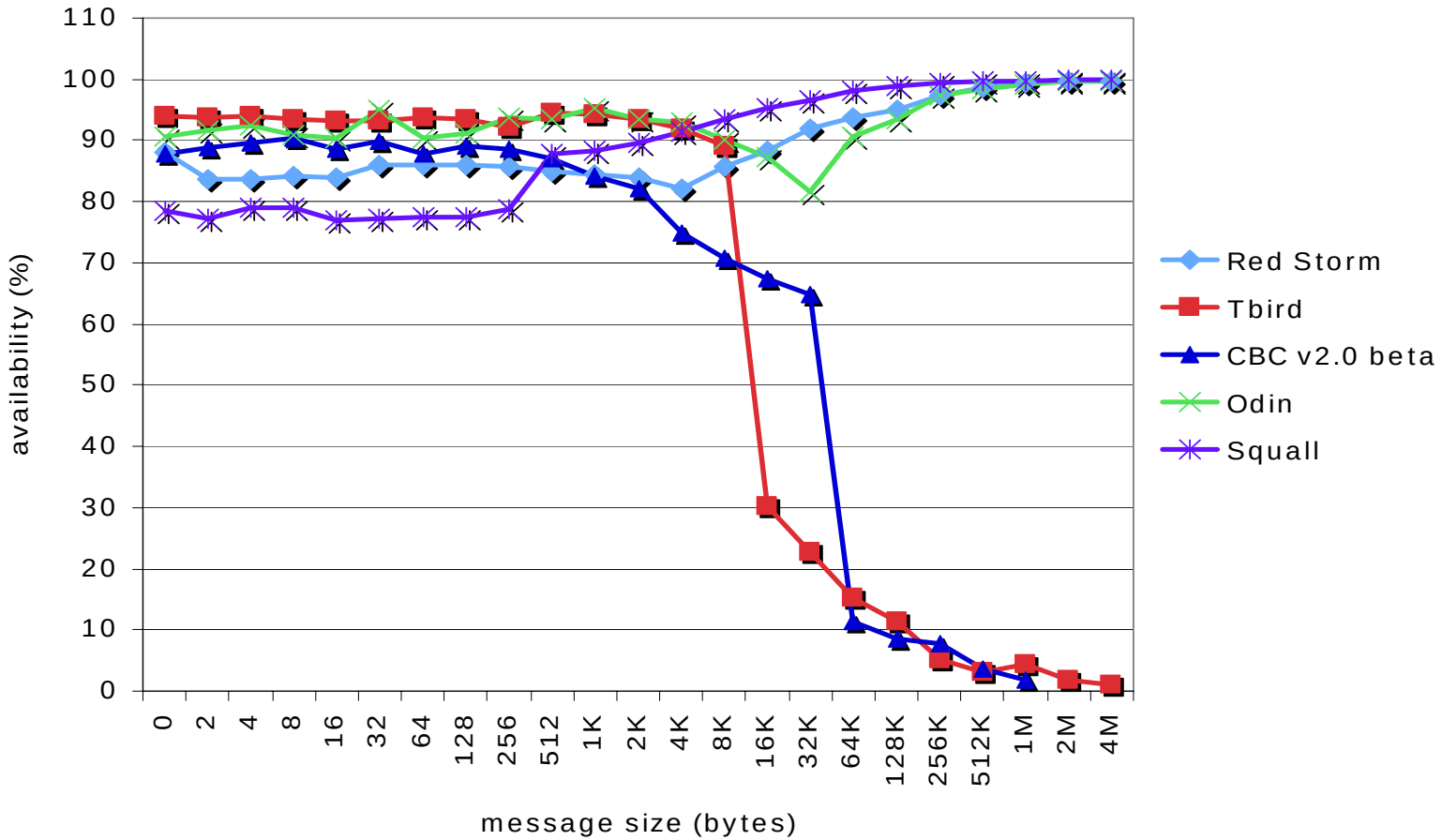
Sandia National Laboratories

Send Overhead

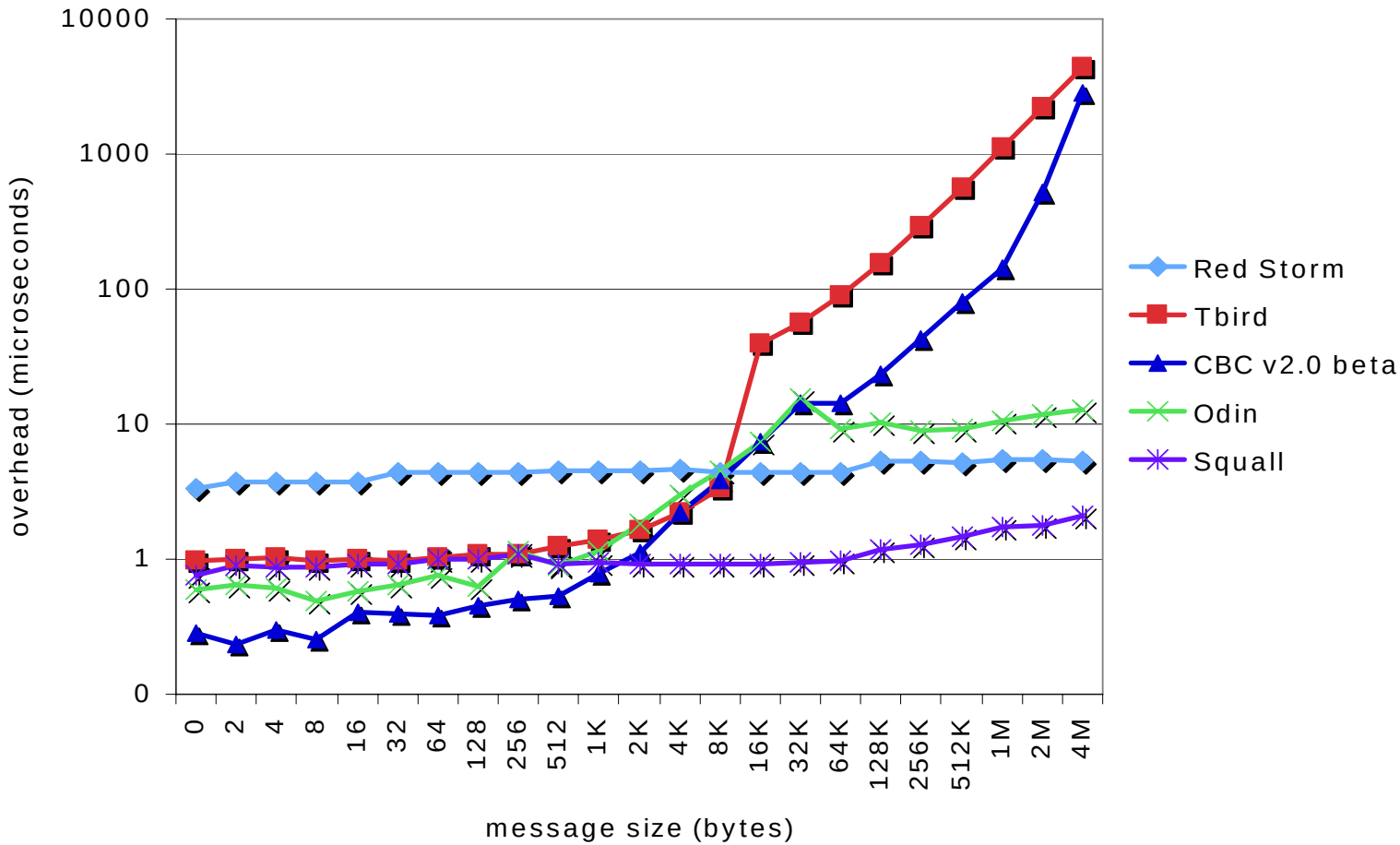


Sandia National Laboratories

Send Availability

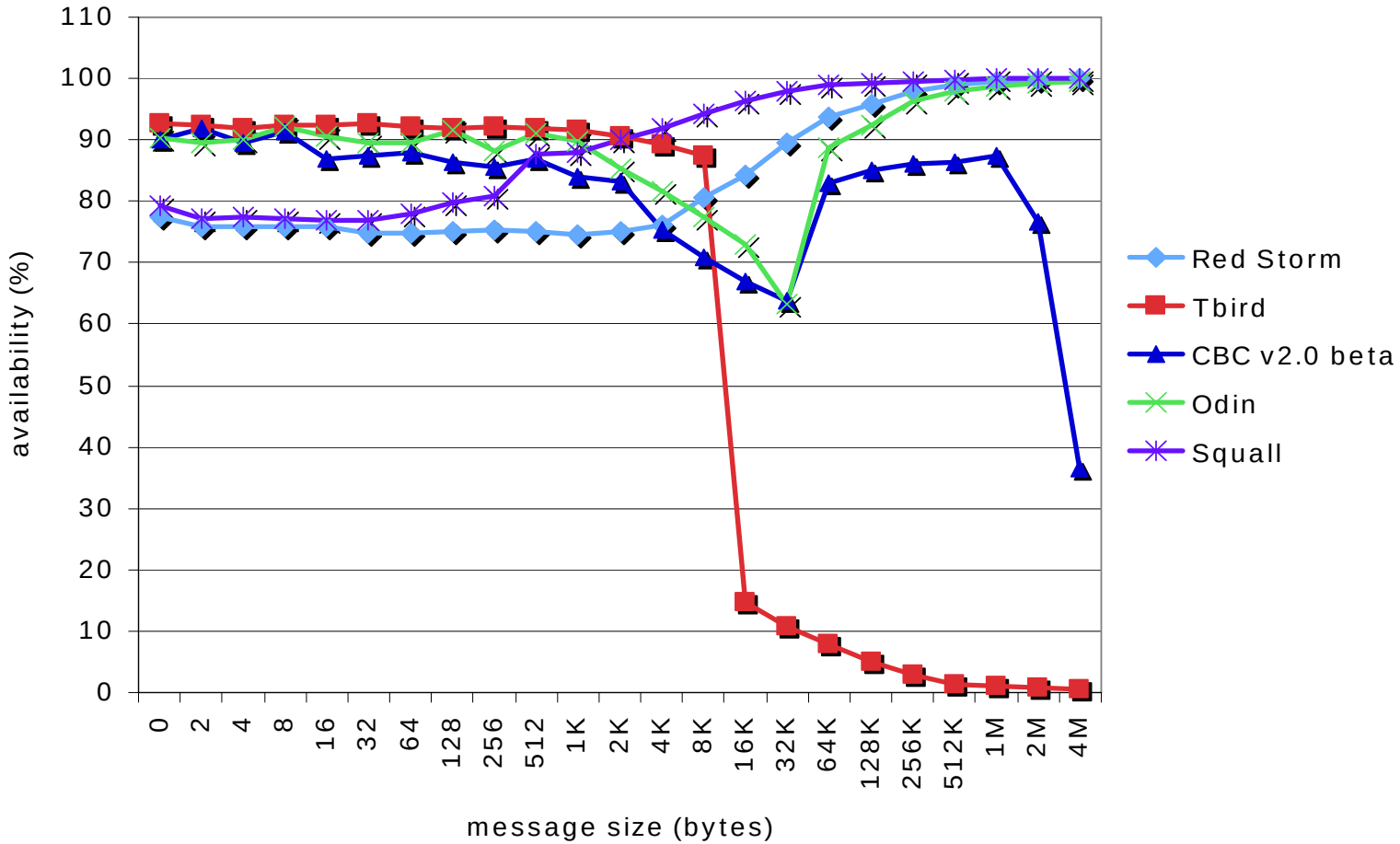


Receive Overhead



Sandia National Laboratories

Receive Availability



Observations

- Significant difference in overhead and availability between networks that have progress and those that do not
- Availability probably not important for small messages
  - Helps to understand interface characteristics
    - Effect of protocol crossover points
- Small message availability may also help determine gap and/or message rate

Related Work

- Initial work by UC-Berkeley showing the importance of overhead relative to latency and bandwidth (ISCA'97)
- Work by LBL on evaluating high-performance networks (IPDPS'03)
- COMB benchmark suite
  - New benchmark is cleaner approach (and code)
- OSU overlap benchmark (x86 specific)
- Several Sandia papers analyzing the impact of overlap, offload, and progress

Conclusions

- Overhead and availability are important measurements
- Need effective ways of measuring overlap of computation/communication and communication/communication
- Transports that do not provide independent progress have significantly worse overhead and availability performance

Future Work

- Address multi-processor nodes
- Make benchmark more realistic
  - Address issues that Keith raises ☺
- Continue to develop methods to measure
  - Ability to overlap communication/communication
  - Impact of resource allocation strategies like memory pinning

http://www.cs.sandia.gov/smb

Sandia National Laboratories