CRD

Performance Comparison of Cray X1E, XT3, NEC SX8, and AMD/IB

Hongzhang Shan

Erich Strohmaier

Lawrence Berkeley National Laboratory

Office of Science
U.S. DEPARTMENT OF ENERGY

Motivation

- XT3, X1E recently developed by Cray, need to understand their performance
  - Using synthetic benchmarks
  - Using scientific kernels or applications
- Relations between results of synthetic benchmarks and applications
  - Focus on communication

Outline

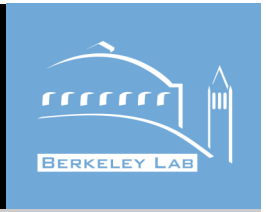
- Network Performance
  - Single Pair
    - Uni-directional
    - Bi-directional
  - Multi Pair
    - Bi-directional
- Application Performance
  - BeamBeam3D
- Modeling
  - Relations between benchmark results and application performance

CRD Platform Highlights

Platform	SMP	CPU		Mem	Network		
		Type	Peak	Peak	Type	Topology	Peak*
Cray X1E	4 (MSP)	X1E 1.13GHz	18GF/s	34GB/s	Custom	4D-Hyper cube	25.6 GB/s
NEC SX8	8	SX8 2GHz	16GF/s	64GB/s	IXS	Crossbar	16GB/s
Cray XT3	1	Opteron 2.4GHz	4.8GF/s	6.4GB/s	SeaStar	Torus	3.8GB/s
AMD/IB	2	Opteron 2.2GHz	4.4GF/s	6.4GB/s	Infini-Band	Fat-tree	1GB/s

Peak: Unidirectional, per network link

CRD Single Pair Unidirectional, Bi-directional Bandwidth Test



Unidirectional:

```
Clock(start)
For (I = 1; I < N; I++) {
  If (myid == 0) {
    MPI_Send();
    MPI_Recv();
  }
  Else {
     MPI_Recv()
    MPI_Send();
  }
}
Clock(end)
BW-Uni = N*size/(end - start)
```

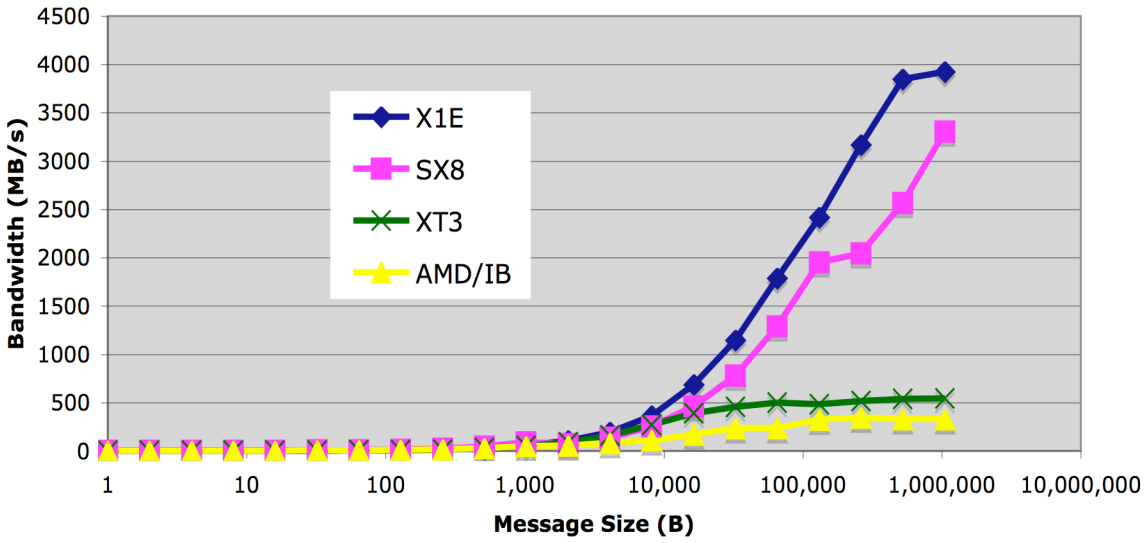
Bidirectional:

```
Clock(start)
For (I = 1; I < N; I++) {
   MPI_Irecv();
   MPI_Send();
   MPI_Wait();
}
Clock(end)
BW-Bi = N *size/(end - start)
```

Ideal : BW-Bi = 2 * BW-Uni

CRD

Single-Pair Uni-directional Bandwidth

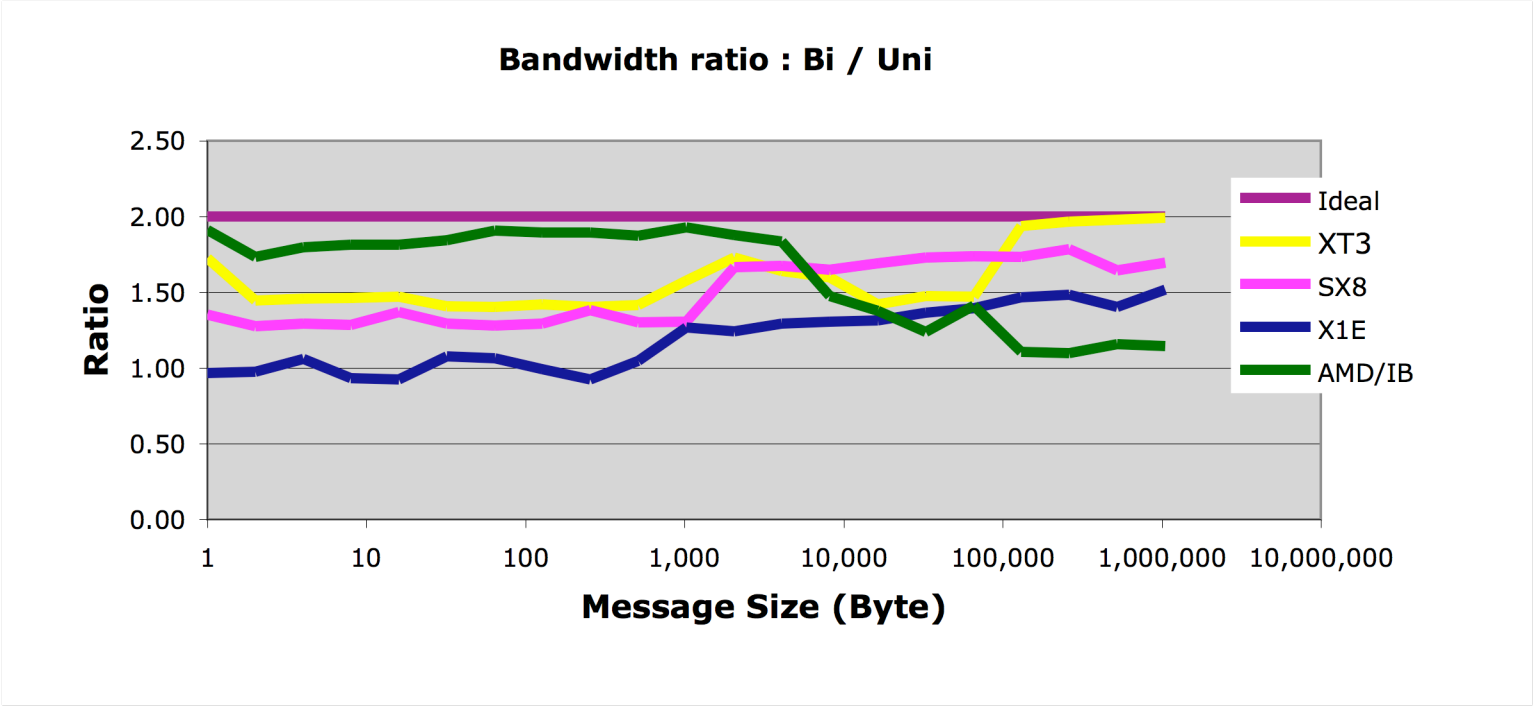


X1E	25.6 GB/s
SX8	16GB/s
XT3	3.8GB/s
AMD/IB	1GB/s

- The results are measured by selecting one processor from each of the two SMP nodes.
- The order correlates well with network link peak performance
- Vector platforms achieve significant higher bandwidth than superscalar platforms for large message sizes
- XT3 performs better than AMD/IB cluster

Office of Science
U.S. DEPARTMENT OF ENERGY

Single-pair Bidirectional Bandwidth



- For most cases, the ratio is well below ideal value of 2
- Different platforms show different pattern
- Performance on AMD/IB limited by PCI bus

Multi-Pair Bandwidth Test

Find pair:

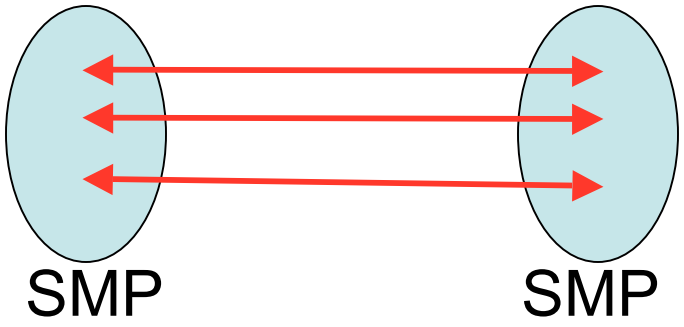
Pair.first = my_rank

Pair.second = my_rank .XOR. (nprocs -1)

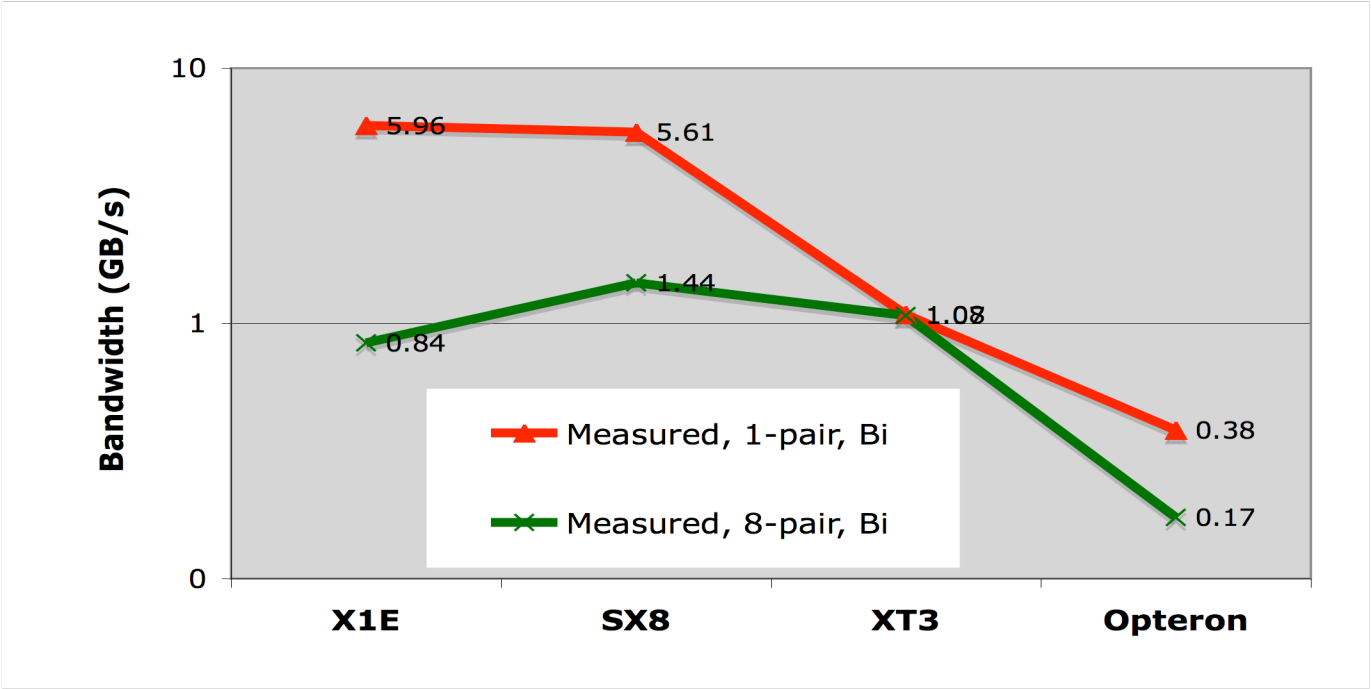
Measure:

```
Clock(start)
For (I = 1; I < N; I++) {
        Uni-directional bandwidth test()  or
        Bi-directional bandwidth test()
}
Clock(end)
```

Bandwidth = N*message size/(end - start)



CRD Network Injection Contention

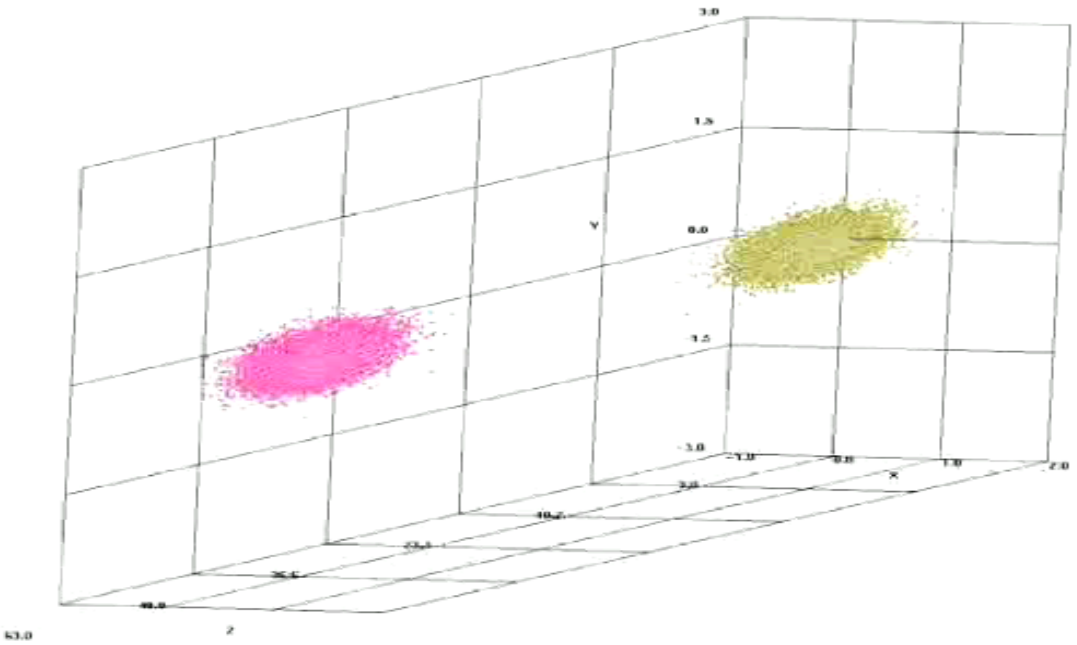


- **Contention is not an issue at the measured scale on XT3**

Outline

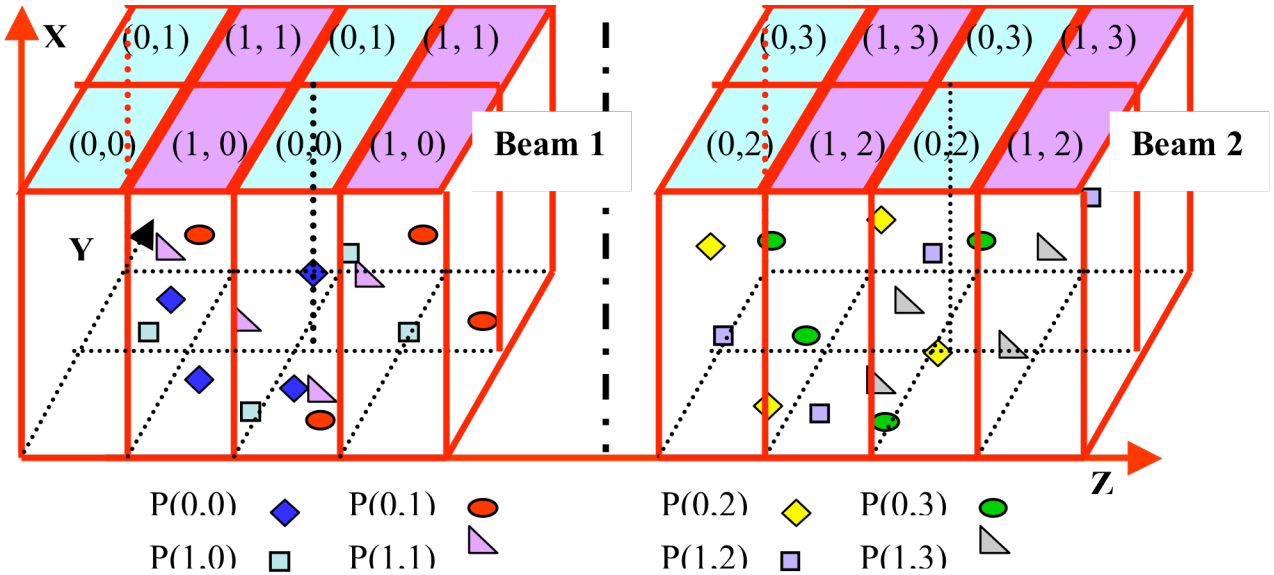
- Network Performance
  - Single Pair
    - Uni-directional
    - Bi-directional
  - Multi Pair
    - Bi-directional
- **Application Performance**
  - **BeamBeam3D**
- **Modeling**
  - **Relations between benchmark results and application performance**

BeamBeam3D



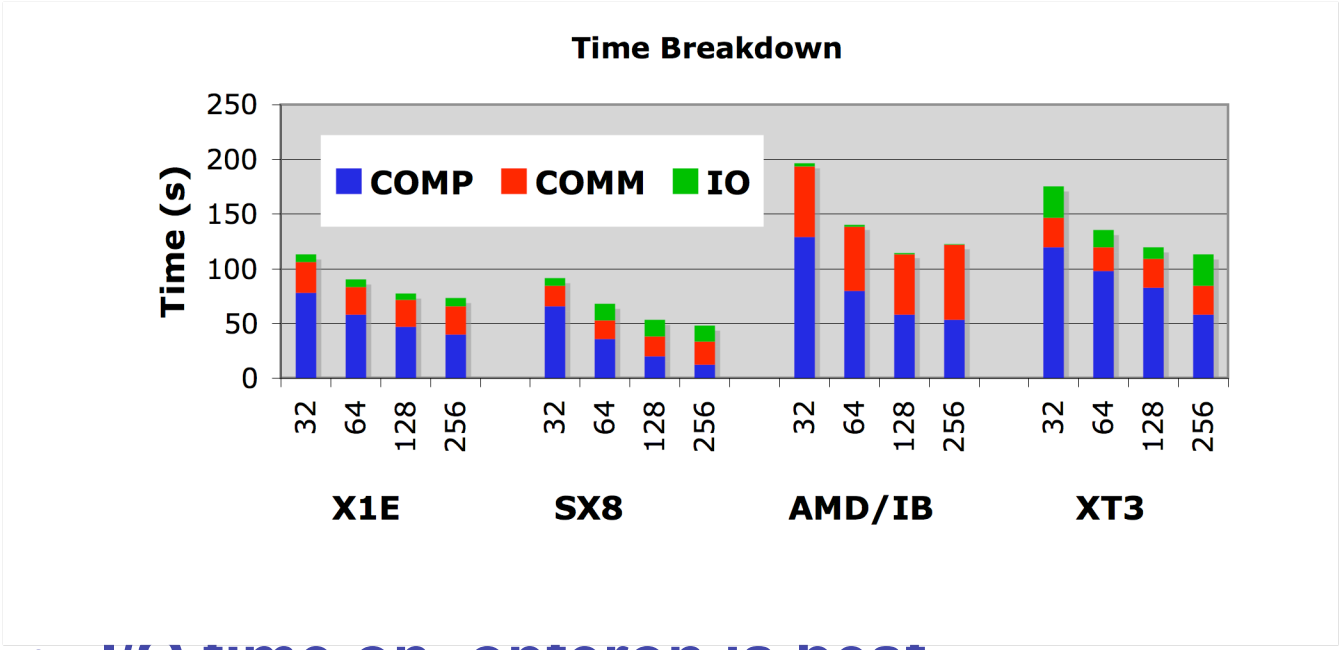
- Simulate Beam-Beam Colliding Process in Ring Colliders
- Important SciDAC application

Computational Method



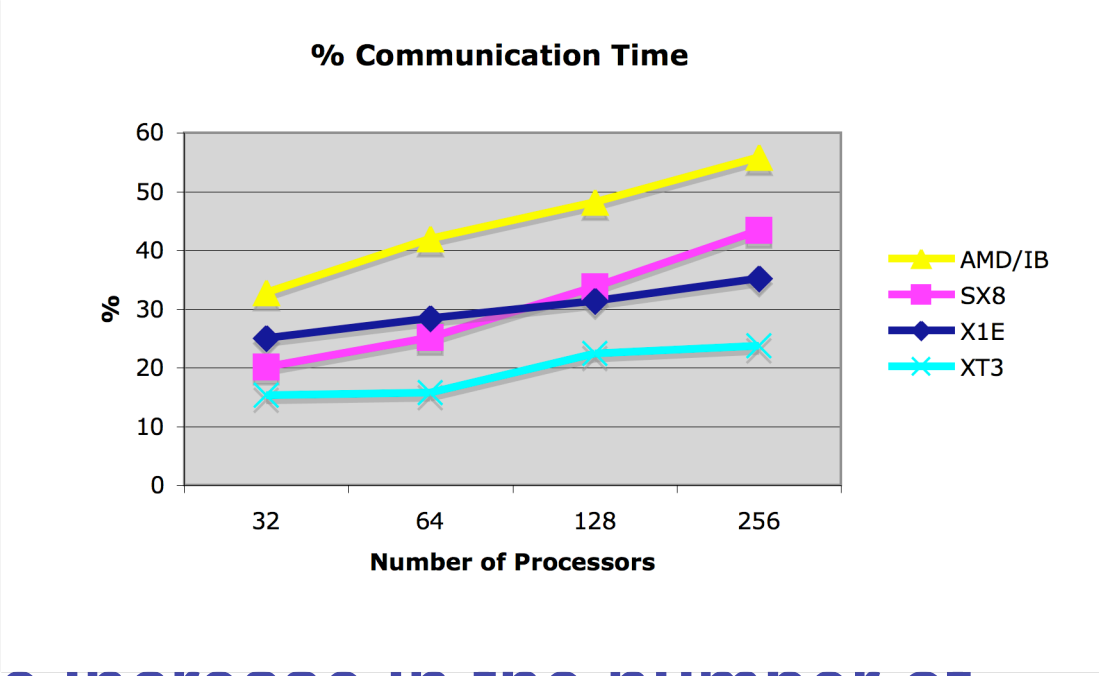
- Particle-in-cell method with two main data structures, particles and field domain
- Using Particle-field decomposition, field grids are partitioned in 2D: Pz * Py

Time Breakdown



- I/O time on Opteron is best
- I/O time on other systems could be reduced by aggregation
- Computation time scales best on the SX8
- Communication time on Infiniband is worst

Communication Ratio



- With the increase in the number of processors, the communication volume keeps constant, leading to higher % of communication time

Outline

- Network Performance
  - Single Pair
    - Uni-directional
    - Bi-directional
  - Multi Pair
    - Bi-directional
- Application Performance
  - BeamBeam3D
- **Modeling**
  - **Relations between benchmark results and application performance**

CRD Communication Characteristics

Phase	Name	Pattern	Direction	Beam	Size [Byte]	# messages per turn
1:	Greenf2D	FFT Transpose	Column	Same	(Nx/Pcol+1)* (Ny/Pcol)*16*2	(Pcol-1)*(Nslice*2-1)
2a:	Guardsum2D	All-to-All Reduce	Column	Same	Nx*Ny/Pcol*8	(Pcol-1)*Nslice*Nslice
2b:	Guardsum2Drow	All-to-All Reduce	Row	Same	Nx*Ny/Pcol*8*I I = 1, Nslice/Prow	(Prow-1)*MIN(2*Prow, CEILING(Nslice/I, 1)*2-1)
3:	Fieldsolver2D	FFT Transpose	Column	Same	(Nx/Pcol+1)* (Ny/Pcol)*16	(Pcol-1)*Nslice* (Nslice+Prow-1)/Prow*2
4a:	Guardexch2Drow	All-to-All Broadcast	Row	Same	Nx*Ny/Pcol*8*I I = 1, Nslice/Prow	(Prow-1)*MIN(2*Prow, CEILING(Nslice/I, 1)*2-1)
4b:	Guardexch2D	All-to-All Broadcast	Column	Other	Nx*Ny/Pcol*8	Pcol*Nslice*Nslice

Nx*Ny is the field grid size, Nslice is the number of slices per beam
Pcol*Prow is the processor grid

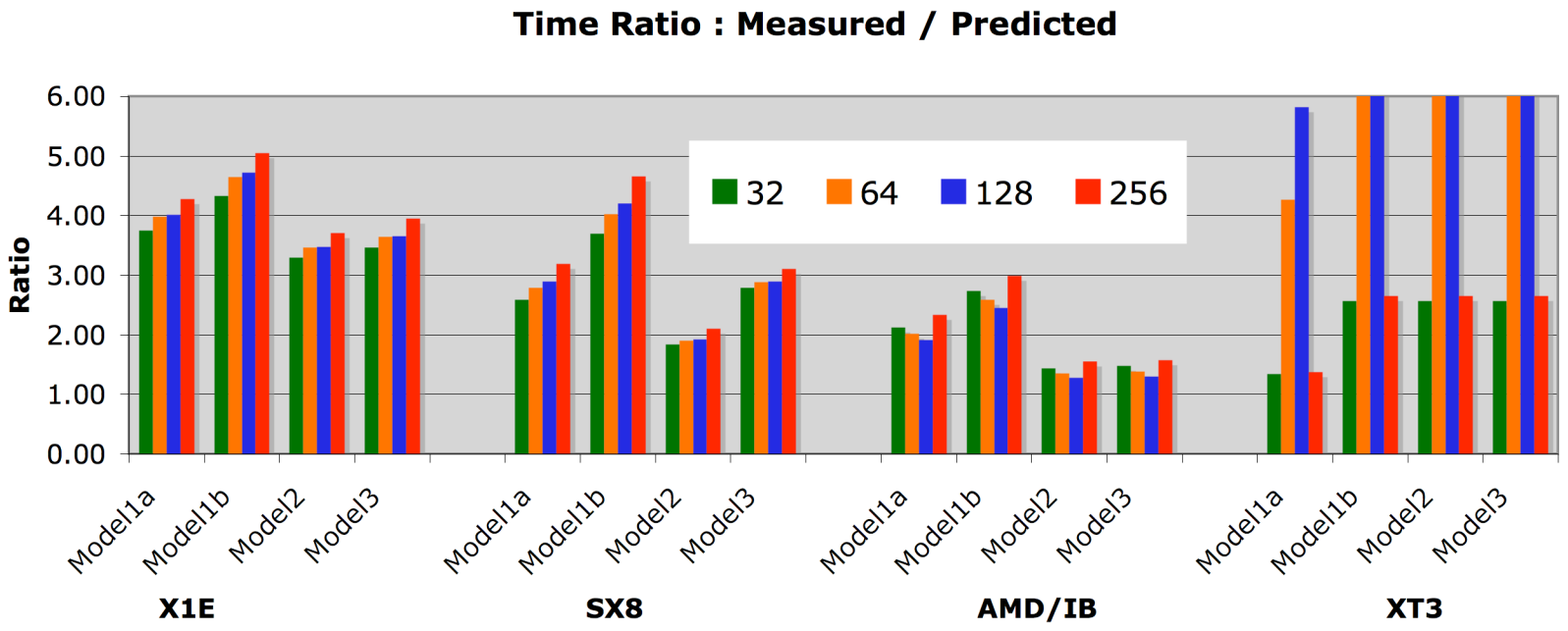
Models

$$T = L + S/B$$

T: time, L: latency, S: Message Size, B: bandwidth

- **Single layer:**
  - **Model 1a: Single pair, Uni, between SMP nodes**
  - **Model 1b: Single pair, Bi, between SMP nodes**
  - **Model 2 : Multi pair (# processors in a SMP), Bi, between SMP nodes**
- **Multi Layer:**
  - **Model 3 : Multi pair, Bi, inside SMP and between SMP**

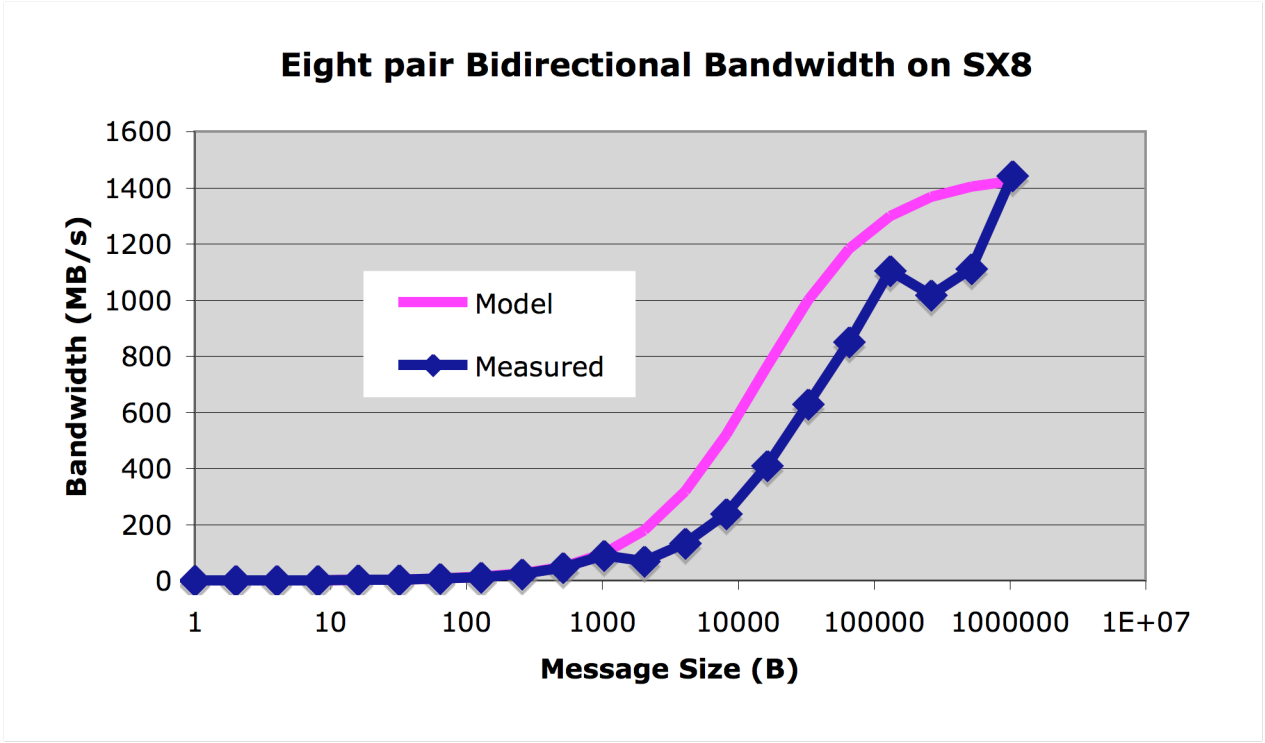
CRD Performance Prediction



- Model 1a, 1b do not correlate well with application performance
- Model 2, Multi-pair Bidirectional results is better than single-pair bi-directional results
- Model 3, multi-layer does not work well

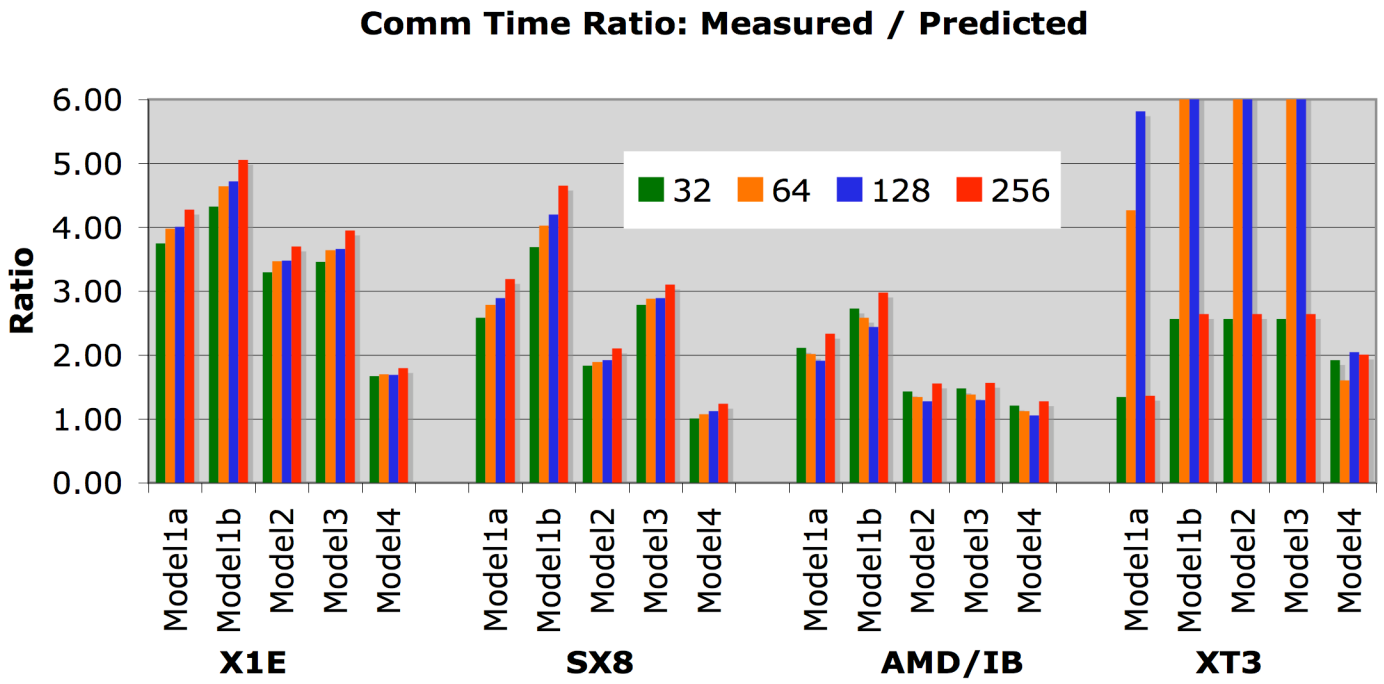
Office of Science
U.S. DEPARTMENT OF ENERGY

Model 4



- Linear model does not fit well
- Using measurement number for each message size directly (Model 4)

Performance Prediction



- Using measurement number directly (Model 4) works much better
- X1E, XT3 need more complex benchmarks due to network topology (HPCC ?)

Summary

- **Network:**
  - Multi pair benchmark captures contention from node adapter much better than single pair measurement
- **Application**
  - Vector platforms perform much better than superscalars
- **Modeling:**
  - Big gap between effective bandwidth on applications and the peak measured by single pair benchmarks
  - Multi pair results capture contention from node adapter much better than single pair
  - Using microbenchmark timings directly is more accurate than using a linear timing model
  - On X1E, XT3, synthetic benchmarks sensitive to network link contention are needed

Acknowledgement

- Thank Oak Ridge National Laboratory, NERSC, the High Performance Computing Center Stuttgart (HLRS) to provide the platforms.