What if MPI Collectives Were Instantaneous?

Cray User Group (CUG) Meeting
Lugano, Switzerland

Rolf Riesen
Courtenay Vaughan
Sandia National Laboratories

Torsten Hoefler
Technical University Chemnitz

May 11, 2006

Talk Overview

Section Outline

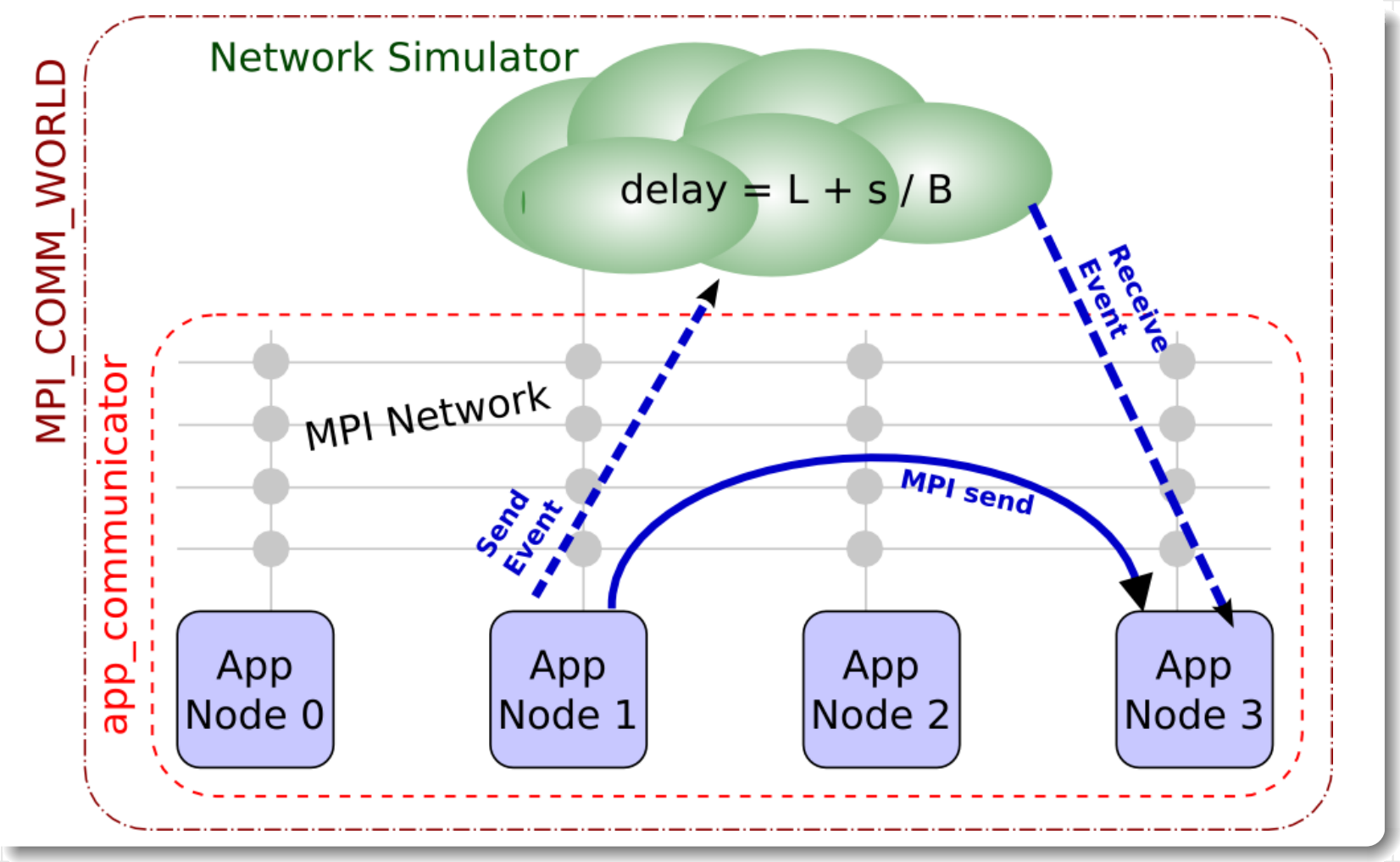
Introduction

- Collective operations are a fertile field for research
- Optimize collectives for specific topology
- Move collectives into NIC
- etc.

- Some optimizations are difficult to implement and debug
- Would be nice to know in advance whether effort is worth it

Section Outline

Hybrid Simulation



Section Outline

NAS Parallel Benchmarks

Number of collectives used by NAS FT

	Class A			Class B			Class C	
nodes	4	16	64	4	16	64	16	64
Reduce	6	6	6	20	20	20	20	20
Allreduce	2	2	2	2	2	2	2	2
Alltoall	8	8	8	22	22	22	22	22
Alltoallv	0	0	0	0	0	0	0	0
Barrier	1	1	1	1	1	1	1	1
Bcast	6	30	126	6	30	126	30	126

- Don’t expect any gain from optimized collectives.

NAS Parallel Benchmarks

Number of collectives used by NAS FT

	Class A			Class B			Class C	
nodes	4	16	64	4	16	64	16	64
Reduce	6	6	6	20	20	20	20	20
Allreduce	2	2	2	2	2	2	2	2
Alltoall	8	8	8	22	22	22	22	22
Alltoallv	0	0	0	0	0	0	0	0
Barrier	1	1	1	1	1	1	1	1
Bcast	6	30	126	6	30	126	30	126

- Don’t expect any gain from optimized collectives.

NAS Parallel Benchmarks

Number of collectives used by NAS MG

	Class A			Class B			Class C		
nodes	4	16	64	4	16	64	4	16	64
Reduce	1	1	1	1	1	1	1	1	1
Allreduce	88	88	88	88	88	88	88	88	88
Alltoall	0	0	0	0	0	0	0	0	0
Alltoallv	0	0	0	0	0	0	0	0	0
Barrier	6	6	6	6	6	6	6	6	6
Bcast	18	90	378	18	90	378	18	90	378

- NAS parallel benchmarks are not good tests for collectives performance.

NAS Parallel Benchmarks

Number of collectives used by NAS MG

	Class A			Class B			Class C		
nodes	4	16	64	4	16	64	4	16	64
Reduce	1	1	1	1	1	1	1	1	1
Allreduce	88	88	88	88	88	88	88	88	88
Alltoall	0	0	0	0	0	0	0	0	0
Alltoallv	0	0	0	0	0	0	0	0	0
Barrier	6	6	6	6	6	6	6	6	6
Bcast	18	90	378	18	90	378	18	90	378

- NAS parallel benchmarks are not good tests for collectives performance.

All-to-All Benchmark

```
for (size= 1;  size  < max_cnt; size++)   {
      snd_cnt= rcv_cnt= size;
      t1= MPI_Wtime();
      for (rep= 0;  rep < NUM_REP; rep++) {
            MPI_Alltoall(snd_buf, snd_cnt, MPI_INT, rcv_buf,
                        rcv_cnt,  MPI_INT, MPI_COMM_WORLD);
      }
      t2= MPI_Wtime();
      if  (my_rank == 0)   {
            printf  ("%8d %12.9f\n", snd_cnt, (t2 – t1)  /  NUM_REP);
      }
}
```

- Expect large performance gain from improved collectives

Abinit

- Package to calculate total energy, charge density and electronic structure of systems composed of electrons and nuclei
- Kernel of code consumes about 98% of the running time
- Kernel uses **MPI_Alltoall()** and **MPI_Allreduce()** exclusively
- Previous work showed that Abinit is mainly limited by the **MPI_Alltoall()** collective operation
- Expect that improved collectives will help Abinit to scale better

Abinit

- Package to calculate total energy, charge density and electronic structure of systems composed of electrons and nuclei
- Kernel of code consumes about 98% of the running time
- Kernel uses **MPI_Alltoall()** and **MPI_Allreduce()** exclusively
- Previous work showed that Abinit is mainly limited by the **MPI_Alltoall()** collective operation

- Expect that improved collectives will help Abinit to scale better

Section Outline

Methodology

- Run each application and benchmark in native, stand-alone mode
- Run each with simulator
- Run each with simulator, collectives cost set to 0

- Current collectives model is too optimistic for messages $\leq$ 128 kB
- Limits accuracy of simulator runs
- Does not matter when we set cost artificially to zero

Methodology

- Run each application and benchmark in native, stand-alone mode
- Run each with simulator
- Run each with simulator, collectives cost set to 0

- Current collectives model is too optimistic for messages $\leq$ 128 kB
- Limits accuracy of simulator runs
- Does not matter when we set cost artificially to zero

Section Outline

NAS Parallel Benchmarks

Runtime for NAS FT

	16 nodes		
	min	median	max
normal	59.54	59.74	59.83
sim	59.53	59.69	59.82
zero	59.54	59.63	59.78

	64 nodes		
	min	median	max
normal	15.36	15.42	15.62
sim	15.37	15.43	15.70
zero	15.35	15.42	15.74

NAS Parallel Benchmarks

Runtime for NAS IS

	16 nodes		
	min	median	max
normal	2.73	2.81	4.29
sim	2.77	2.81	4.44
zero	2.77	2.81	4.44

	64 nodes		
	min	median	max
normal	1.45	1.46	2.14
sim	1.45	1.46	2.18
zero	1.41	1.46	2.15

NAS Parallel Benchmarks

Runtime for NAS MG

	16 nodes		
	min	median	max
normal	14.53	14.69	14.70
sim	14.54	14.68	14.70
zero	14.53	14.68	14.71

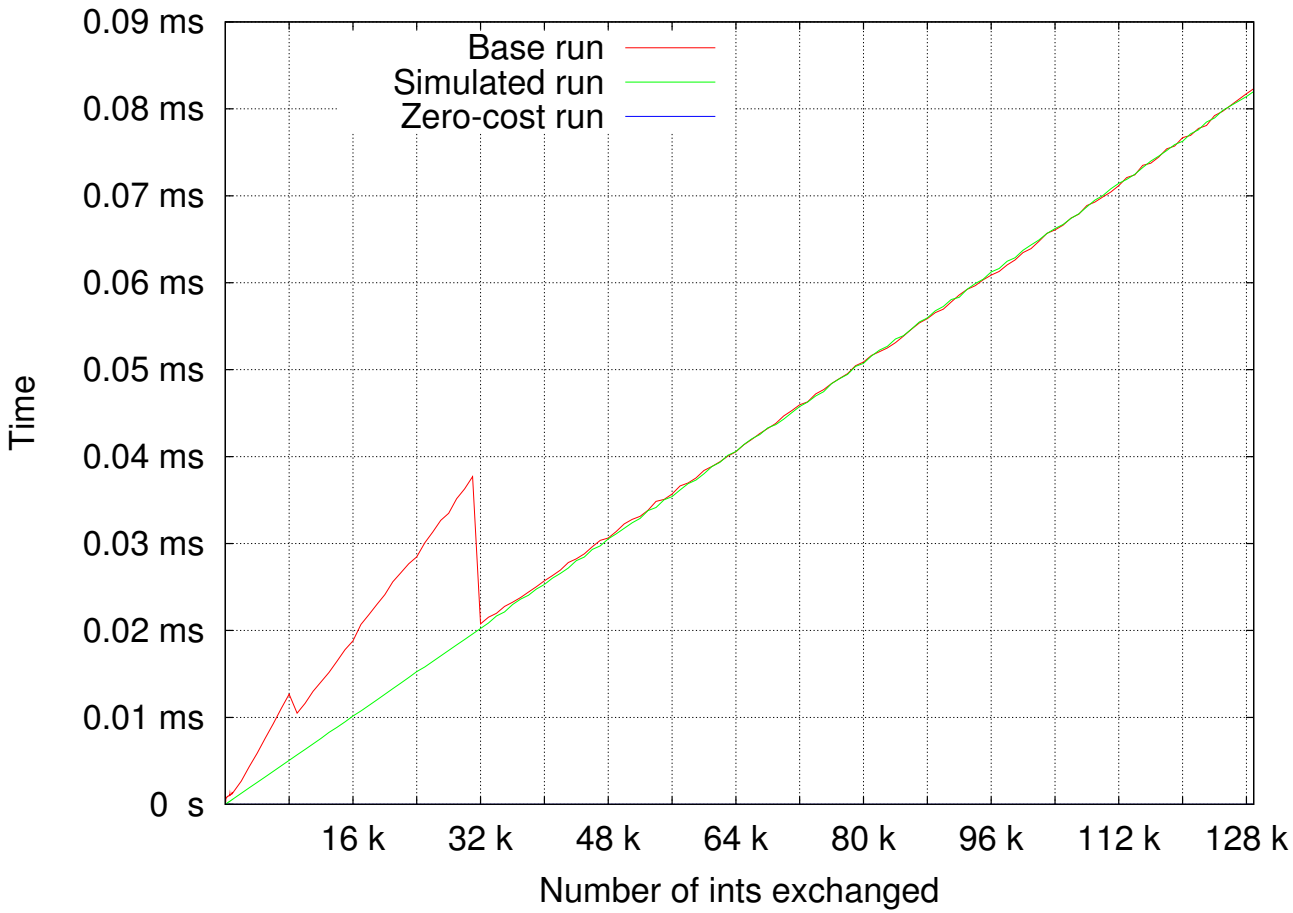
	64 nodes		
	min	median	max
normal	3.40	3.56	3.56
sim	3.40	3.55	3.56
zero	3.40	3.55	3.57

All-to-All Benchmark

All-to-all on 16 nodes

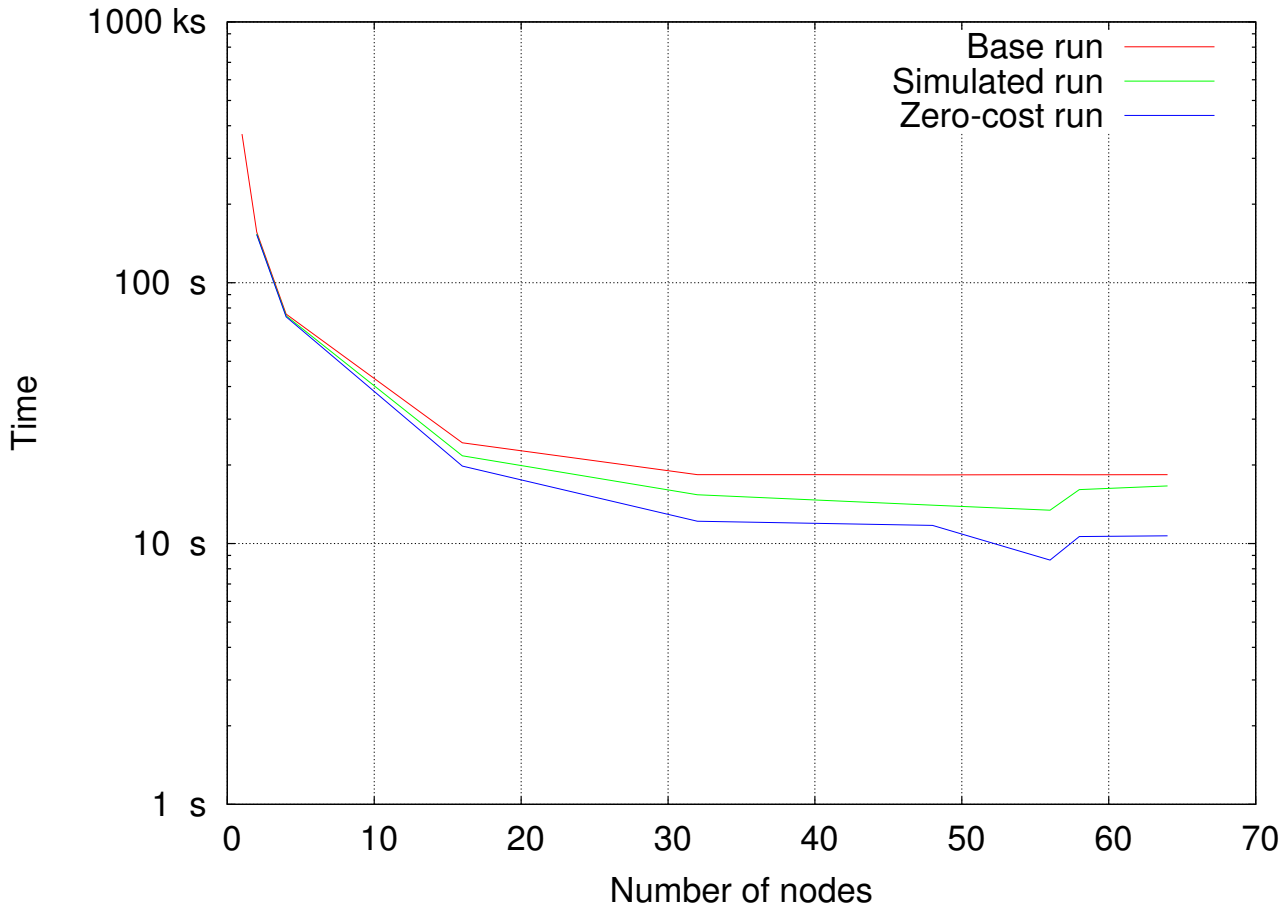
All-to-All Benchmark

All-to-all on 64 nodes



Abinit

Abinit with 43 atoms



Section Outline

Summary

- NAS benchmarks are not at all sensitive to the performance of collective operations
- All-to-all benchmark shows huge performance differences
- There is clearly an impact of collective performance on the runtime of Abinit
  - runtime of the inner loop for our particular atomic simulation is cut in half
  - Abinit shows such poor scaling; collective improvement does not help much