Бенчмаркинг сетевых соединений

Обзор

Сафин Альберт

29 апреля 2015 г.

Benchmarking Methodology for Network Interconnect Devices

- RFC 2544
  Benchmarking Methodology for Network Interconnect Devices
- RFC 1242
  Benchmarking Terminology for Network Interconnection Devices
- Цель: тестирование устройств
- Опубликована в 1999

Организация теста

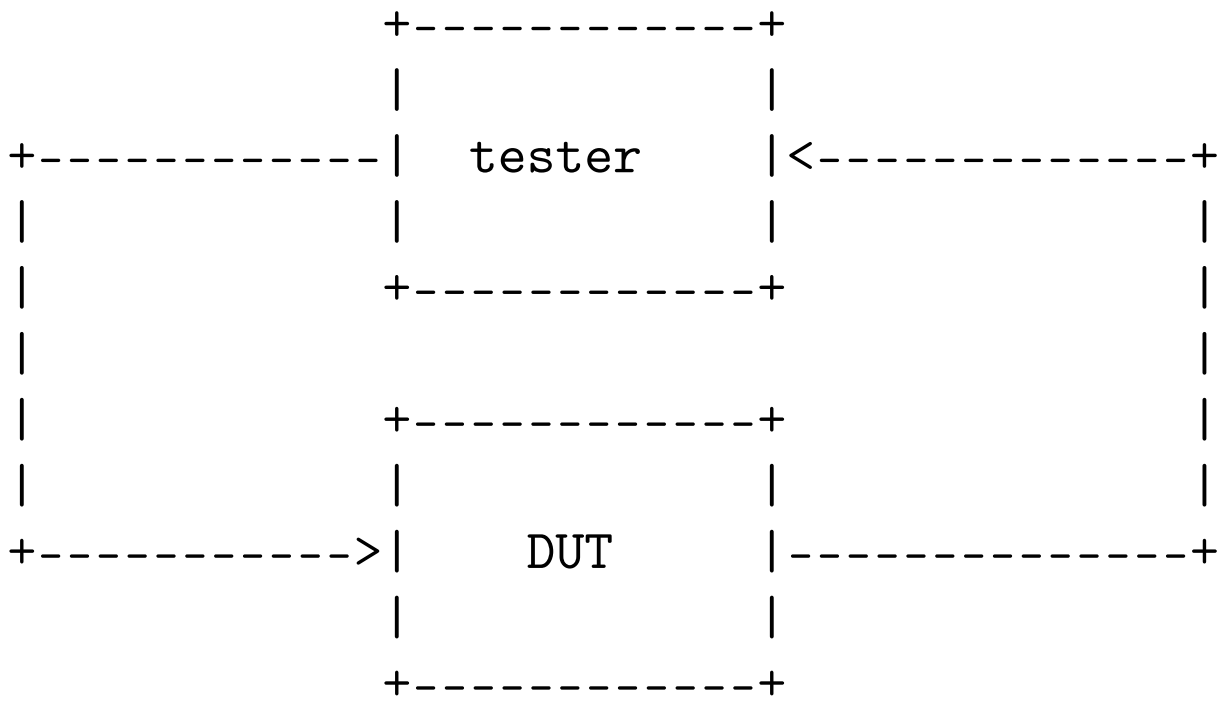


Figure 1

Тесты

- Throughput / Пропускная способность
- Latency / Латентность
- Frame loss rate / Частота потери кадров
- Back-to-back frames
- System recovery / Перегрузка
- Reset / Перезагрузка

Throughput / Пропускная способность

что Максимальное количество кадров в секунду, которое может передать устройство без ошибок.

тест Скорость варьируется методом бисекции.

Latency / Латентность

def The time interval starting when the end of the first bit of the input frame reaches the input port and ending when the start of the first bit of the output frame is seen on the output port.

тест
- поток с максимальной пропускной способностью, 120 секунд
- метка спустя 60 секунд
- результат — среднее время за 20 или более тестов

Frame loss rate / Частота потери кадров

что Процент кадров, полученных для пересылки, которые устройство при постоянной нагрузке не смогло переслать по причине нехватки ресурсов.

тест

- начинается с максимальной скорости
- скорость снижается с шагом 10% или менее
- пока два теста подряд не пройдут без потерь
- результат в форме графика:
  X — скорость
  Y — процент потерь

Back-to-back frames

что Количество кадров с минимальным межкадровым интервалом.

тест
- бисекция по количеству пакетов
- результат — среднее за 50 измерений

System recovery / Перегрузка

что Время восстановления после перегрузки.

тест
- передаётся поток, 110% от максимальной ПС
- в момент времени A скорость снижается до 50%
- записывается момент B последней потери кадра
- результат — среднее значение $B - A$

Reset / Перезагрузка

что Время востановления после программного или аппаратного сброса.

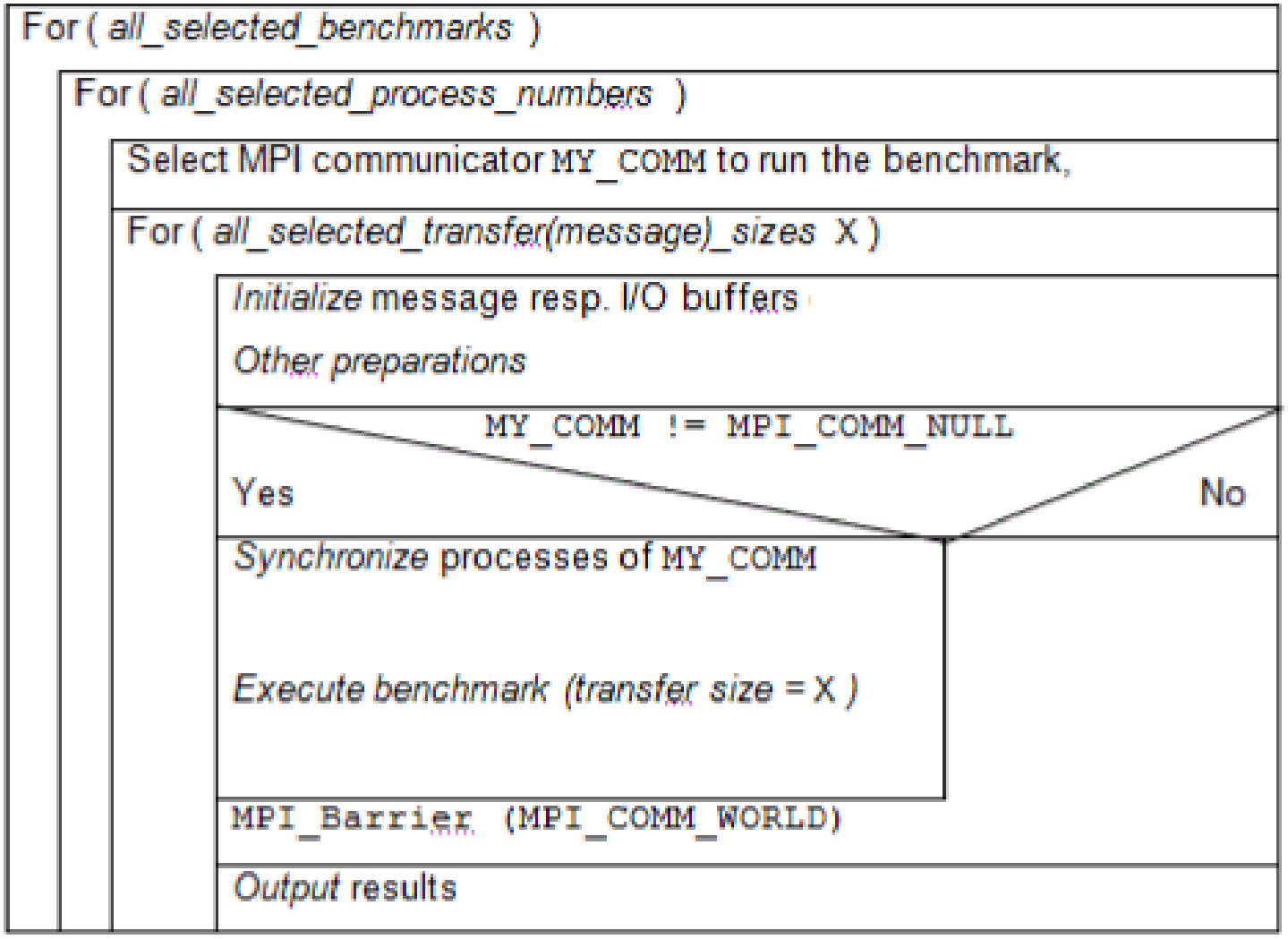
тест

Intel® MPI Benchmarks 4.0 Update 2

- URL: `https://software.intel.com/en-us/articles/intel-mpi-benchmarks`
- Intel® MPI Benchmarks User Guide and Methodology Description
- Цель: тестирование MP
- Компоненты:

IMB-MPI1	MPI-1
IMB-EXT	односторонние коммуникации (MPI-2)
IMB-IO	I/O (MPI-2)
IMB-NBC	non-blocking (MPI-3)
IMB-RMA	remote memory access (MPI-3)

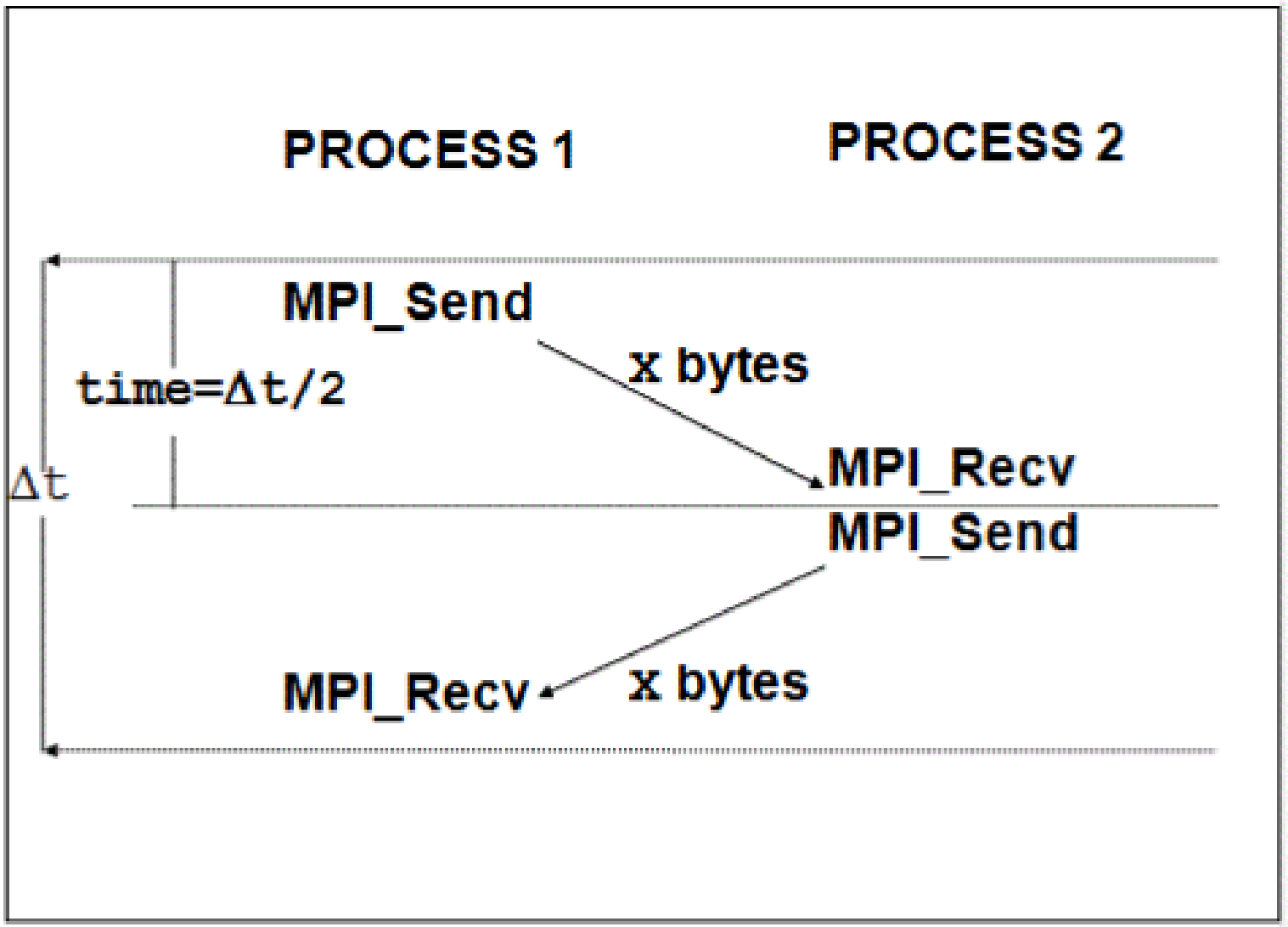
Control flow of Intel® MPI Benchmarks



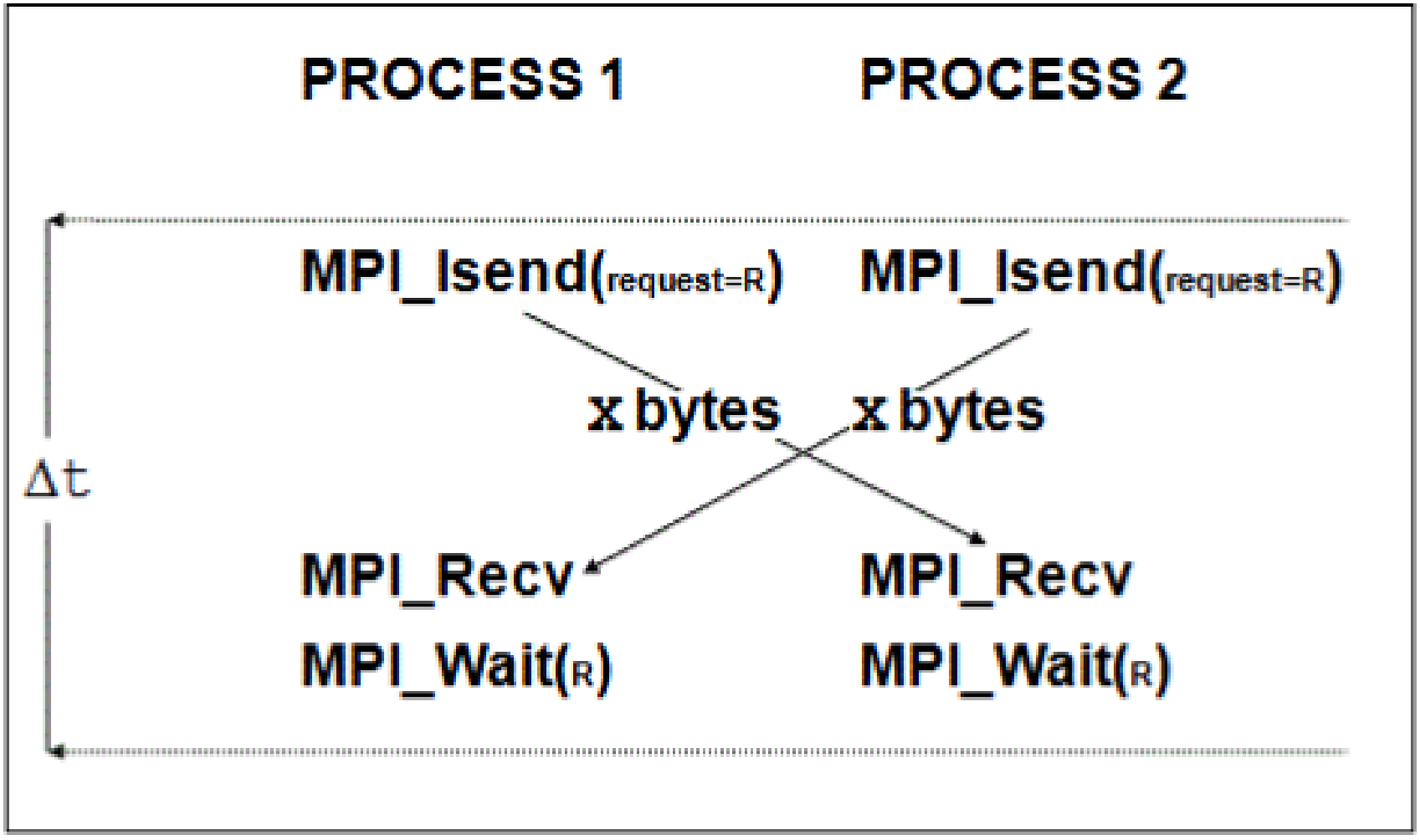
MPI-1

- Классификация:
  - Single Transfer
    - PingPong, PingPongSpecificSource
    - PingPing, PingPingSpecificSource
  - Parallel Transfer
    - Sendrecv
    - Exchange
    - Multi-PingPong
    - Multi-PingPing
    - Multi-Sendrecv
    - Multi-Exchange
  - Collective
- Типы данных: MPI_BYTE, MPI_FLOAT
- Варьируются размеры сообщений
- Измеряется пропускная способность
- Результат усредняется

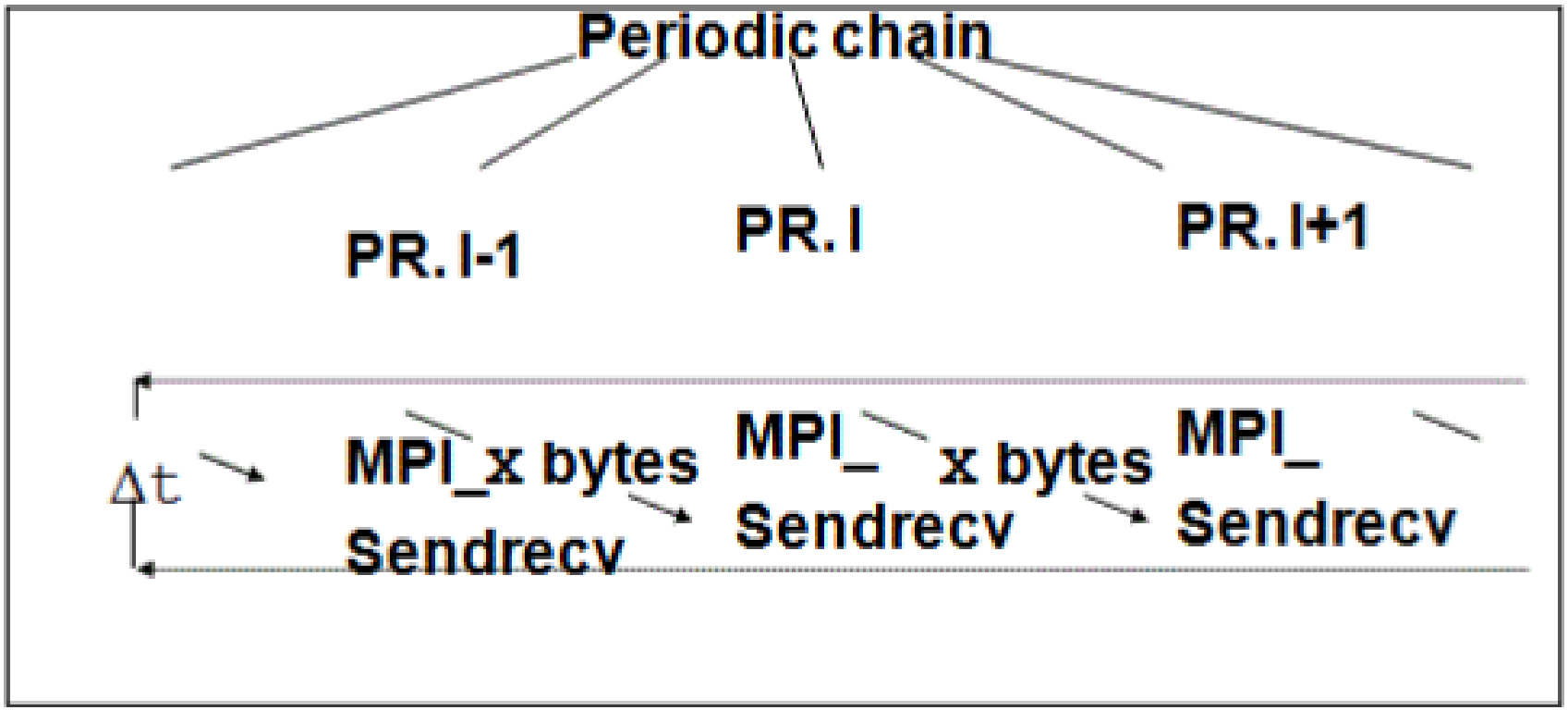
PingPong, PingPongSpecificSource



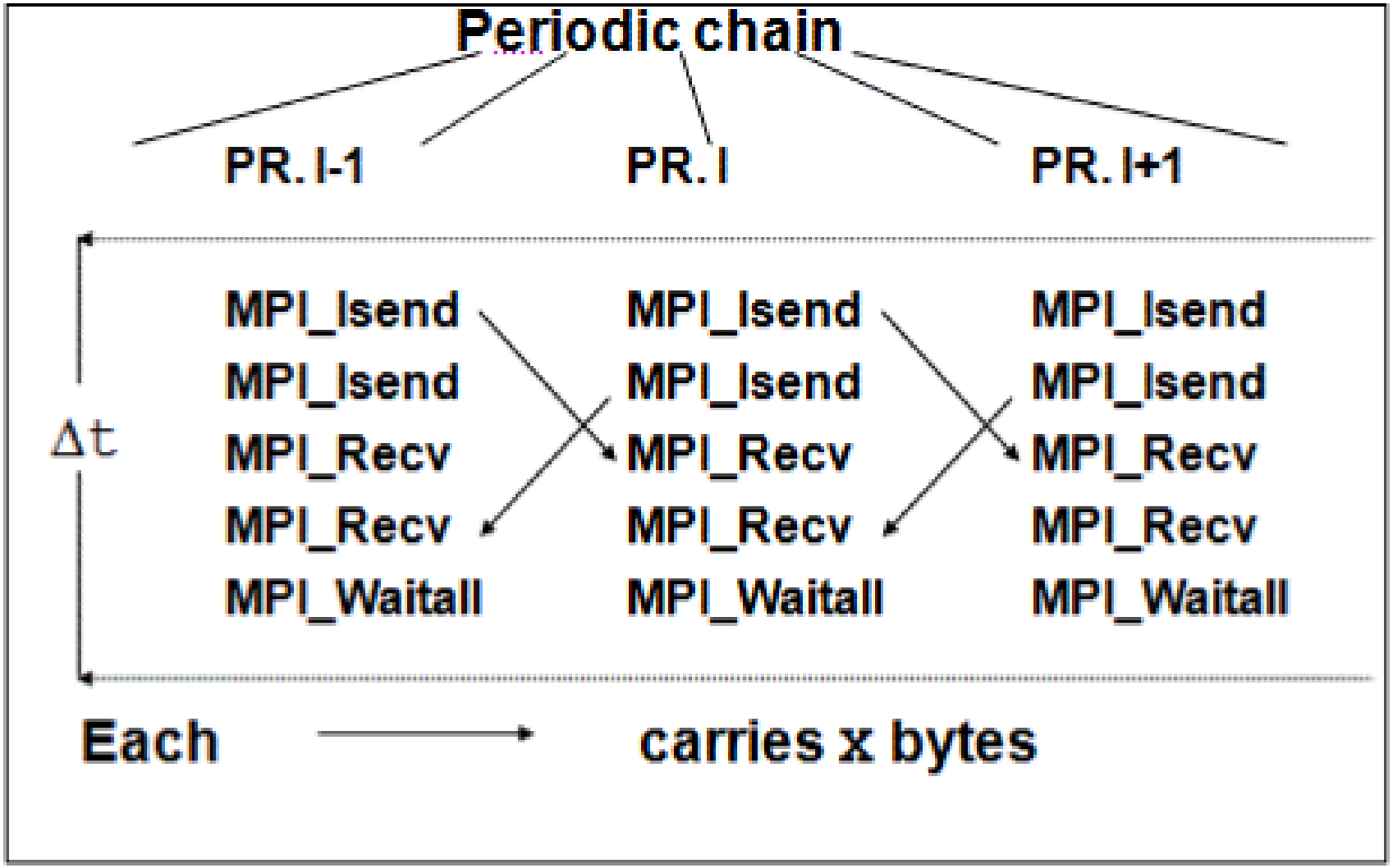
PingPing, PingPingSpecificSource



Sendrecv


Periodic chain
PR. I-1
PR. I
PR. I+1
Δt
MPI_x bytes
Sendrecv
MPI_ x bytes
Sendrecv
MPI_
Sendrecv

Exchange


Periodic chain
PR. I-1
PR. I
PR. I+1
MPI_Isend
MPI_Isend
MPI_Recv
MPI_Recv
MPI_Waitall
Δt
Each
carries x bytes

Multiple mode benchmarks

- Multi-PingPong
- Multi-PingPing
- Multi-Sendrecv
- Multi-Exchange

«The definitions of the multiple mode benchmarks are analogous to their standard mode counterparts in the single transfer class.»

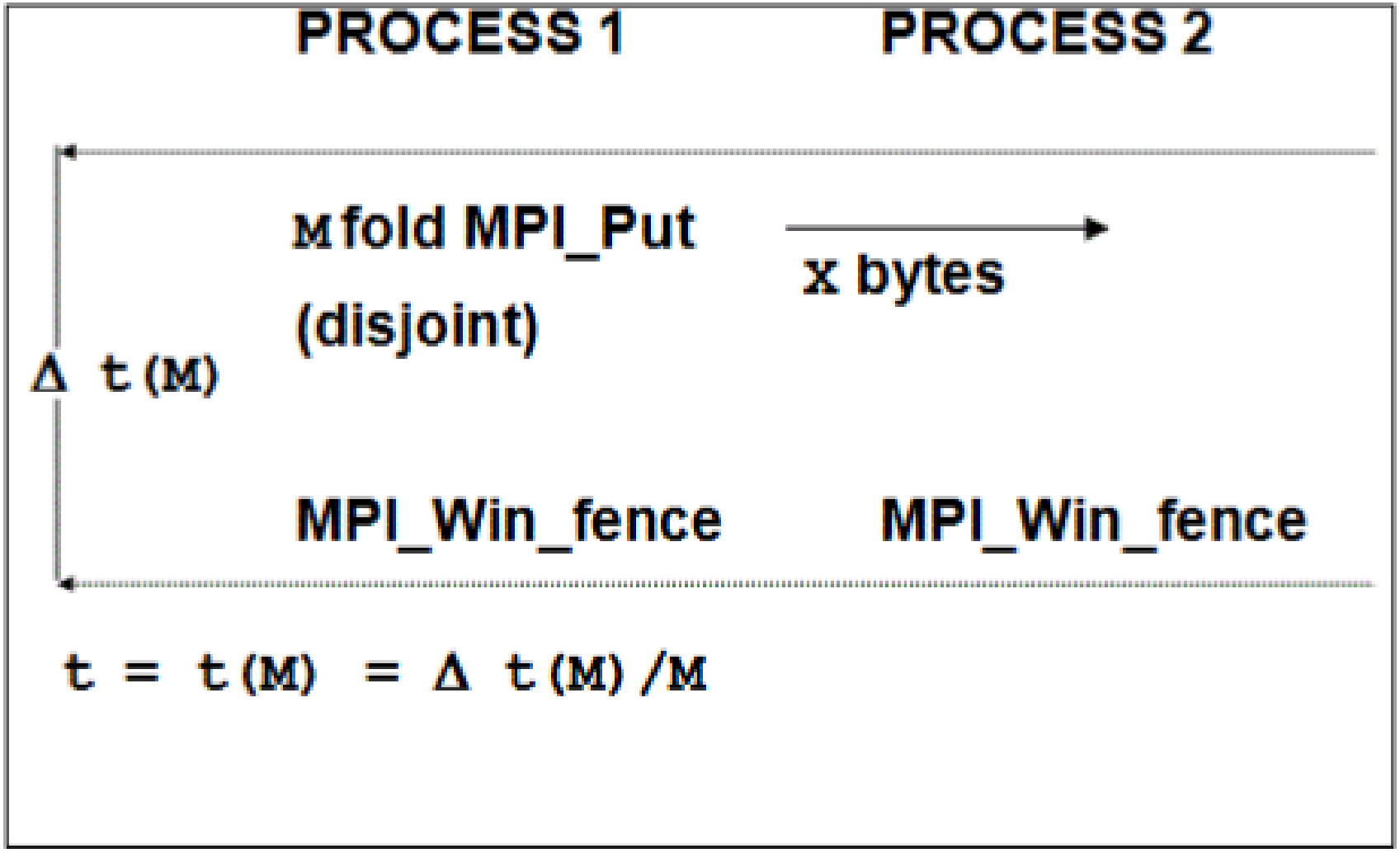
Collective Benchmarks

- Bcast/multi-Bcast
- Allgather/multi-Allgather
- Allgatherv/multi-Allgatherv
- Alltoall/multi-Alltoall
- Alltoallv/multi-Alltoallv
- Scatter/multi-Scatter
- Scatterv/multi-Scatterv
- Gather/multi-Gather
- Gatherv/multi-Gatherv
- Reduce/multi-Reduce
- Reduce_scatter/multi-Reduce_scatter
- Allreduce/multi-Allreduce
- Barrier/multi-Barrier

IMB-EXT(MPI-2)

- Unidir_Put
- Unidir_Get
- Bidir_Put
- Bidir_Get
- Accumulate
- Window

Unidir_Put Pattern



IMB-NBC

Бенчмарки для измерения ...

- пересечения передач и вычислений
  - Ibcast
  - Iallgather
  - Iscatter
  - ...
- чистого времени коммуникаций
  - Ibcast_pure
  - Iallgather_pure
  - Iscatter_pure
  - ...

Measuring Pure Communication Time

Алгоритм:

1. Замерить чистое время коммуникаций.
2. Начать неблокирующую коллективную операцию.
3. Начать вычисления (IMB_cpu_exploit).
4. MPI_Wait.

HPC challenge benchmark

- URL: `http://icl.cs.utk.edu/hpcc/faq/`
- HPCC состоит из 7 тестов.

HPC challenge benchmark

- URL: `http://icl.cs.utk.edu/hpcc/faq/`
- HPCC состоит из 7 тестов.

HPL	floating point
STREAM	memory bandwidth
RandomAccess	rate of random updates of memory
PTRANS	rate of transfer for larges arrays of data from multiprocessor's memory
FFTE	Fast Fourier Transform
DGEMM	$C := \alpha AB + \beta C$
b_{eff}	Latency/Bandwidth

HPC challenge benchmark

- URL: `http://icl.cs.utk.edu/hpcc/faq/`
- HPCC состоит из 7 тестов.

HPL	floating point
STREAM	memory bandwidth
RandomAccess	rate of random updates of memory
PTRANS	rate of transfer for larges arrays of data from multiprocessor's memory
FFTE	Fast Fourier Transform
DGEMM	$C := \alpha AB + \beta C$
b_{eff}	**Latency/Bandwidth**

b_{eff}

$$b_{eff} = logavg\Bigg(\frac{logavg_{ringpatterns}(sum_L(max_{mthd}(max_{rep}(b(ringpat.,L,mthd,rep)))))}{21}, \frac{logavg_{randompatterns}(sum_L(max_{mthd}(max_{rep}(b(randompat.,L,mthd,rep)))))}{21} \Bigg)$$

- Размеры сообщений

$$1B, 2B, 4B, ..., 2kB, 4kB,$$

$$4kB \cdot a^1, 4kB \cdot a^2, ..., 4kB \cdot a^8.$$

$$4kB \cdot a^8 = L_{max} = \frac{memory\ per\ processor}{128}$$

- Методы
  - MPI_Sendrecv
  - MPI_Alltoallv
  - MPI_Irecv, MPI_Isend, MPI_Waitall
- ring patterns
- random patterns

Результат

Latency and Bandwidth Results - Optimized Runs Only - 40 Systems - Generated on Wed Apr 29 04:40:18 2015

System Information System - Processor - Speed - Count - Threads - Processes					Latency					Bandwidth				
					Random Ring	Natural Ring	Ping-Pong Latency			Ping-Pong Bandwidth			Random Ring	Natural Ring
							Maximum	Mininum	Average	Maximum	Minimum	Average		
MA/PT/PS/PC/TH/PR/CM/CS/IC/IA/SD					usec	usec	usec	usec	usec	GB/s	GB/s	GB/s	GB/s	GB/s
Cray XT5 AMD Opteron	2.6GHz	196608	3	65536	15.99	16.09	9.00	0.59	7.45	1.67	1.53	1.61	0.04	0.04
Cray XT5 AMD Opteron	2.6GHz	223112	2	111556	31.09	11.30	10.55	6.87	8.88	1.65	1.52	1.59	0.03	0.35
Cray Inc. mfeg8 Cray X1E	1.13GHz	248	1	248	14.58	16.54	11.59	6.42	8.07	11.03	8.19	8.90	0.30	3.15
Cray Inc. Red Storm/XT3 AMD Opteron	2.4GHz	12960	1	25920	15.76	14.58	9.31	5.68	7.17	2.10	1.98	2.02	0.06	0.15
Cray Inc. Red Storm/XT3 AMD Opteron	2.4GHz	12800	1	25600	19.25	19.19	10.25	5.25	7.45	2.10	1.98	2.02	0.04	0.06
Cray Inc. Red Storm/XT3 AMD Opteron	2.4GHz	12960	1	25920	19.58	19.41	10.37	5.69	7.61	2.10	1.98	2.03	0.04	0.10
Cray Inc. X1 Cray MSP	0.8GHz	252	1	252	22.64	18.82	10.28	8.04	9.15	9.22	4.89	8.41	0.44	2.60
Cray Inc. X1 Cray MSP	0.8GHz	60	1	60	21.16	20.97	9.27	7.99	8.67	9.46	4.41	8.43	1.01	3.43
Cray Inc. X1 Cray MSP	0.8GHz	124	1	124	20.85	18.09	9.69	8.11	8.98	9.10	5.00	8.50	0.80	4.12
Cray Inc. X1 Cray MSP	0.8GHz	124	1	124	20.85	18.09	9.69	8.11	8.98	9.10	5.00	8.50	0.80	4.12
Cray Inc. X1 Cray E	1.13GHz	1008	1	1008	16.30	14.66	9.58	6.40	8.44	10.83	5.96	8.32	0.15	3.03
Cray Inc. XT3 AMD Opteron	2.4GHz	5208	1	5208	9.33	8.51	8.51	6.38	7.37	1.15	1.15	1.15	0.20	0.58
Cray Inc. XT3 AMD Opteron	2.4GHz	5208	1	5208	9.33	8.51	8.51	6.38	7.37	1.15	1.15	1.15	0.20	0.58
Cray Inc. XT3 AMD Opteron	2.4GHz	5208	1	5208	9.18	8.51	8.18	6.12	7.07	1.15	1.15	1.15	0.20	0.65
Cray Inc. XT3 Dual-Core AMD Opteron	2.6GHz	10404	1	10404	17.04	16.14	8.69	5.36	7.01	1.15	1.15	1.15	0.08	0.20
Cray, Inc. XT5 AMD Opteron	2.6GHz	98304	3	32768	15.45	15.40	8.51	0.47	7.25	1.66	1.54	1.61	0.06	0.06

Спасибо за внимание.