Effects of Packet Pacing for MPI Programs in a Grid Environment

Ryousei Takano[1,2], Motohiko Matsuda[1], Tomohiro Kudoh[1], Yuetsu Kodama[1], Fumihiro Okazaki[1], Yutaka Ishikawa[3,1]

[1)]Grid Technology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Japan
[2)]AXE, Inc. [3)]University of Tokyo

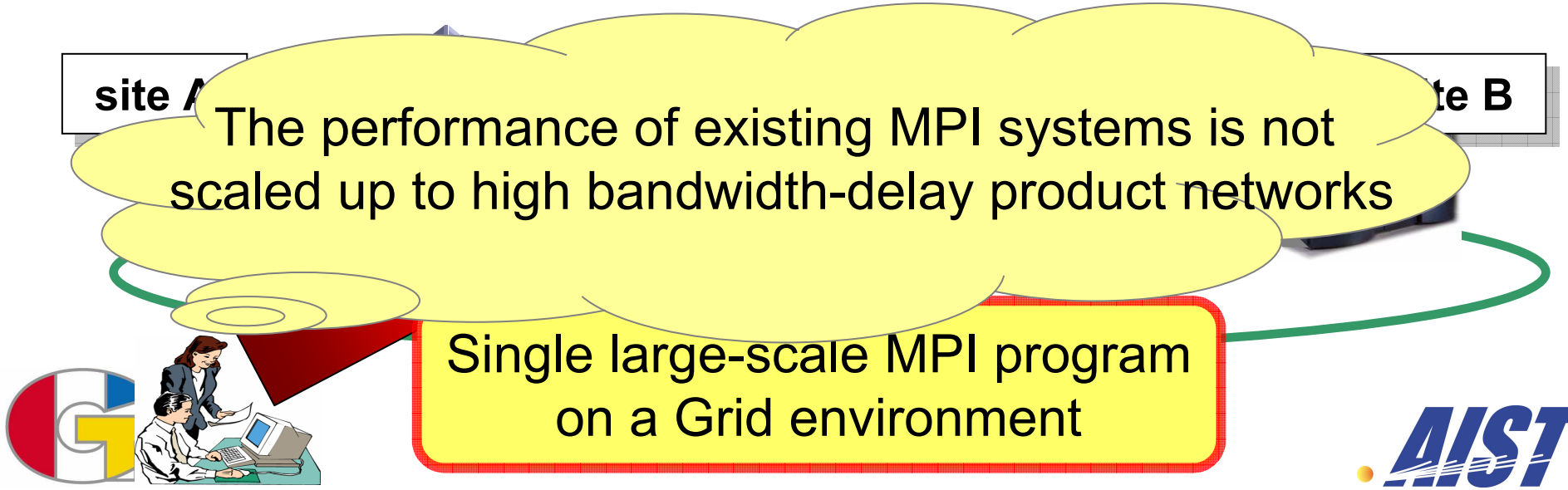
Cluster 2007, Sep. 19 2007, Austin Texas

Agenda

- Motivation
  - GridMPI
- Traffic control method for MPI programs
- Implementation
- Evaluation
- Conclusion

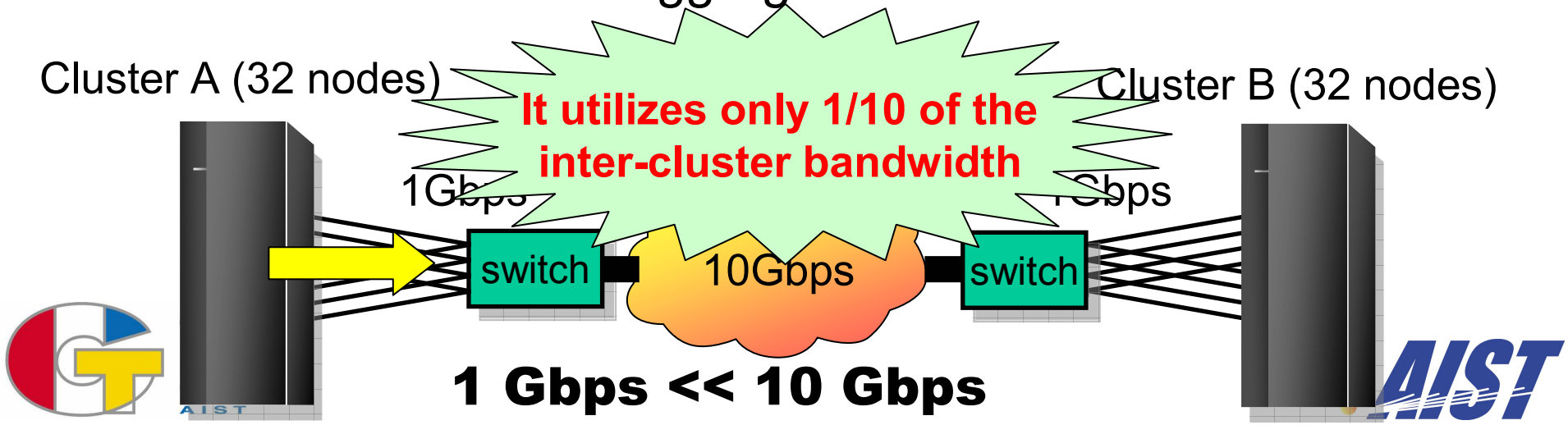
MPI on the Grid

- MPI is widely used for parallel applications
- Some MPI systems are designed for the Grid
  - MPICH-G2, PACX-MPI, MPICH-Madeleine, …
- GridMPI is focused on metropolitan-area networks:
  - ≥ 10Gbps, ≤ 10ms delay (roughly 1000km)



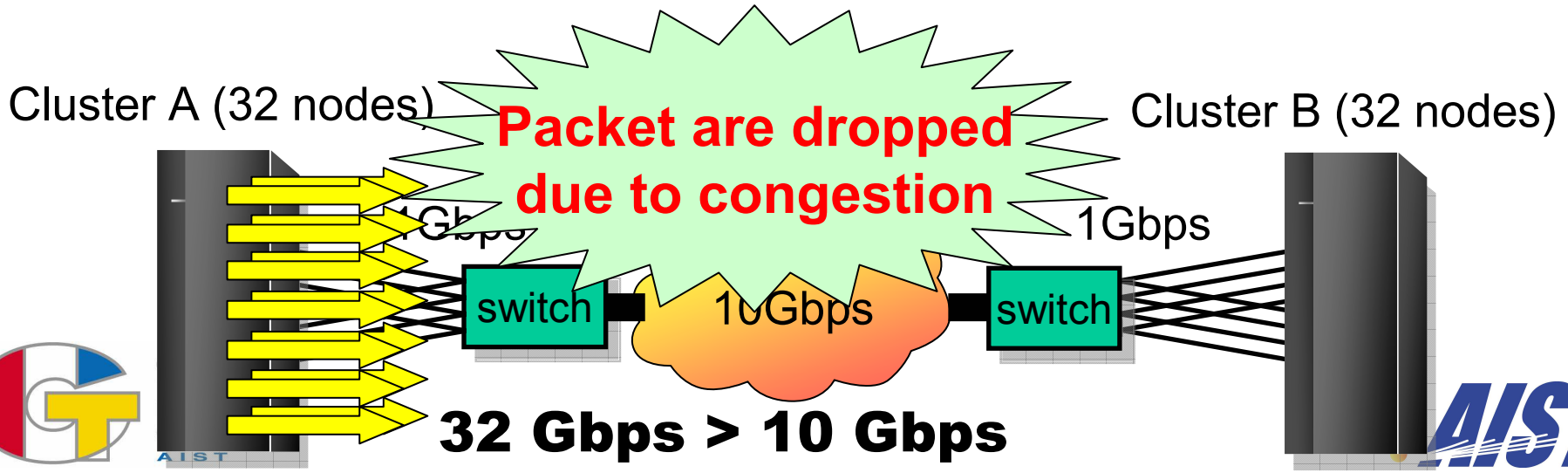
Motivation (1)

- TCP is used for the inter-cluster communication
- Optimizing the TCP performance is the key to successful deployment of MPI programs to the Grid
- Assumption:
  - inter-cluster BW > interconnect BW in cluster
  - inter-cluster BW < aggregate interconnect BW in cluster



Motivation (2)

- How do we maximize use of the network?
  - We should use multiple connections without congestion
  - TCP performance can be degraded due to excessive contention (Especially, worse as the BDP increases)
- Traffic control is needed to fully utilize the inter-cluster network



Agenda

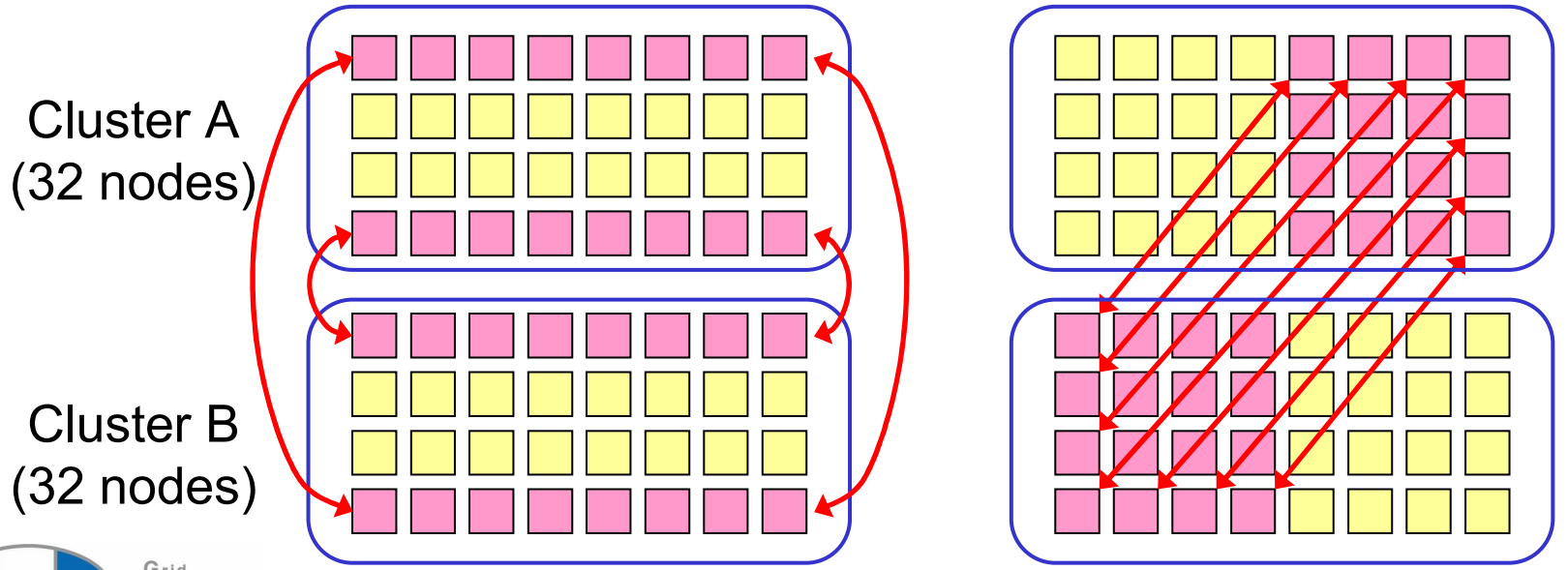
- Motivation
- Traffic control method for MPI programs
  - MATB: Maximum Allowable Transmission Bandwidth
- Implementation
- Evaluation
- Conclusion

MATB: Maximum Allowable Transmission Bandwidth

- How do we decide the transmission bandwidth of each node in cluster to avoid congestion?
  - "Inter-cluster BW / #nodes" is not fully utilize network
- MATB: Maximum allowable transmission bandwidth
  - "Inter-cluster BW / # nodes participated in the inter-cluster communications"
  - Depends on the communication pattern of applications

Examples of inter-cluster communication

- NAS Parallel Benchmarks (BT, SP, and CG)
- Only half nodes of each cluster participate in the inter-cluster communication
  - MATB: 10 Gbps / 16 nodes



(a) BT, SP (b) CG

MATB for the NPB

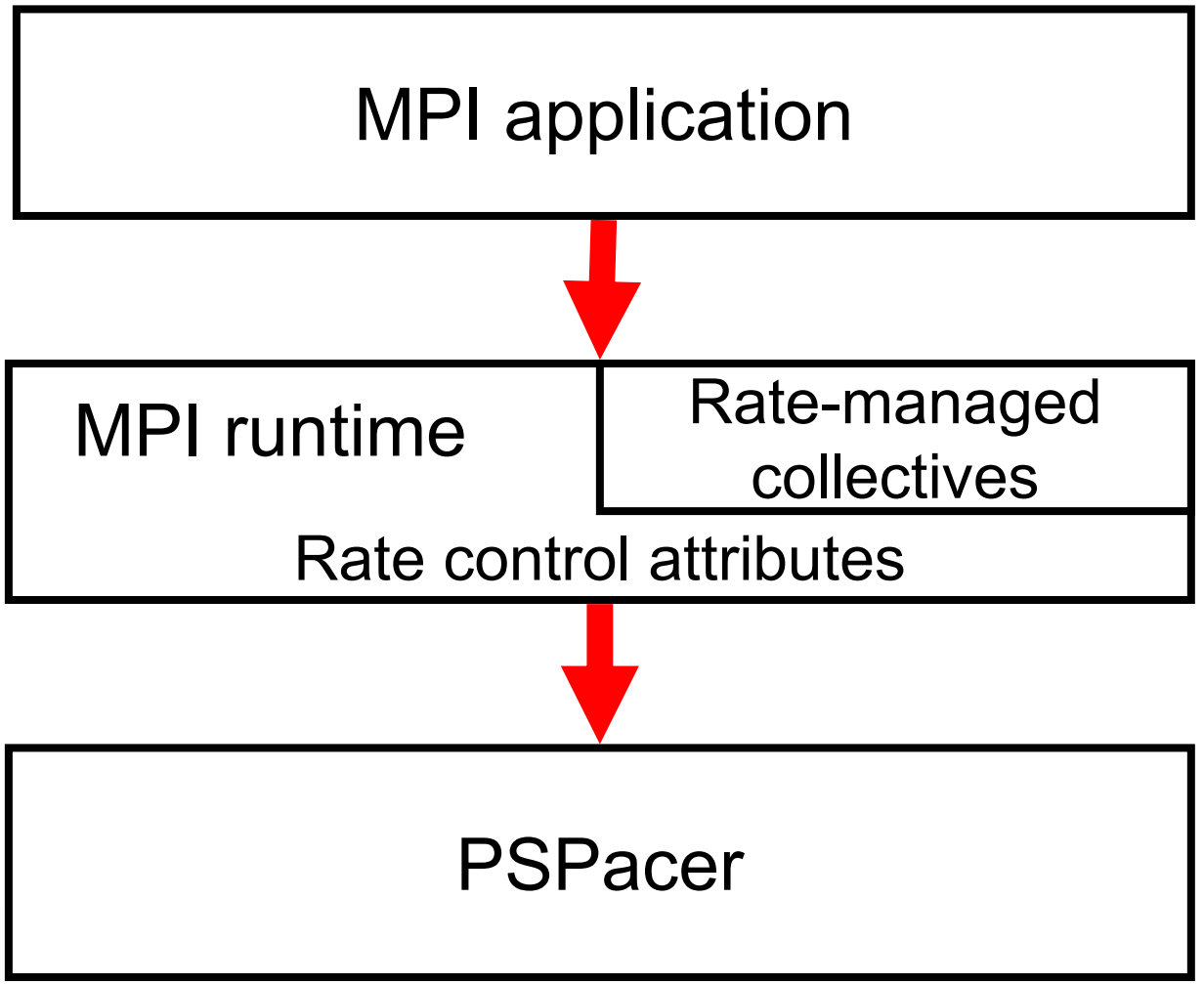
Benchmarks	MATB	(B=10 G, N=32)
BT	$B/(2\sqrt{2N})$	625 Mbps
CG	$B/(N/2)$	625 Mbps
LU, MG	B/N	312.5 Mbps
IS, FT (all-to-all)	B/N	312.5 Mbps

- 2 clusters with the same number of nodes
  - *B*: Inter-cluster bandwidth
  - *N*: The number of nodes at each cluster

Agenda

- Motivation
- Traffic control method for MPI programs
- Implementation
  - Rate control attributes
  - PSPacer: packet pacing software
- Evaluation
- Conclusion

Implementation



MPI-level API

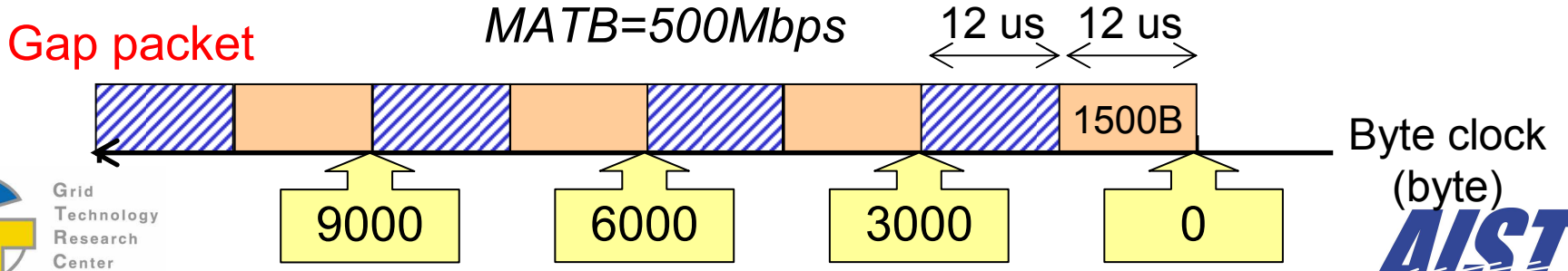
- Rate control attributes
  - MPI attributes (MPI-1.2/2.0 standard)
  - Predefined attribute keys
    - YAMPI_PSP_MAXRATE (inter-cluster bandwidth)
    - YAMPI_PSP_MATB (MATB)
  - MPI program can explicitly set MATB

```
int *rate, *matb, flag;
        :
MPI_Attr_get(comm, YAMPI_PSP_MAXRATE, &rate, &flag);

*matb = *rate / n;
MPI_Attr_put(comm, YAMPI_PSP_MATB, (void *)matb);
```

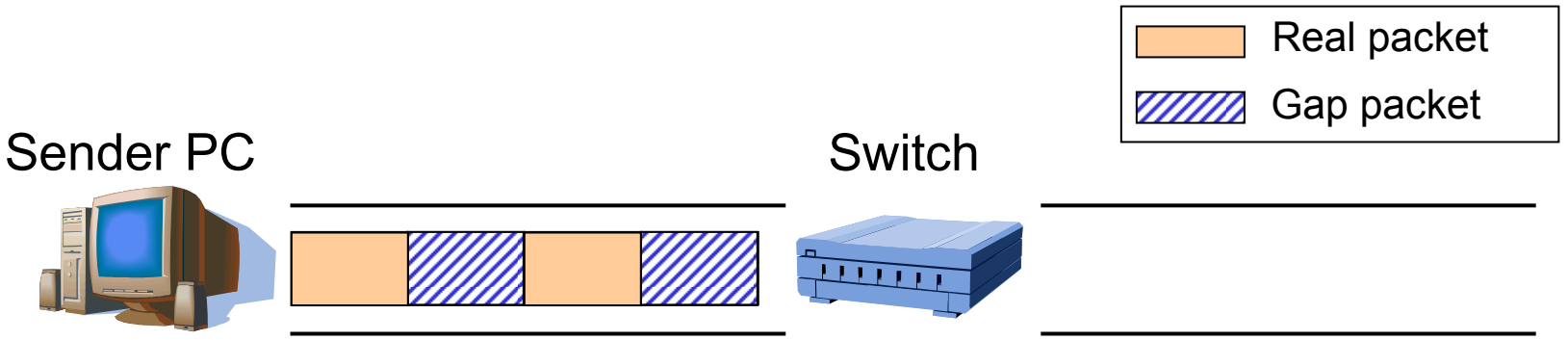
PSPacer: packet pacing software

- Existing method: timer interrupt driven
  - Precise control is difficult for high speed network
  - Token bucket cannot prevent microscopic bursty traffic
- PSPacer: byte clock
  - Transferred bytes (byte clock) are used as a timer
    - For GbE, 1 byte＝8 nsec
  - If packets are sent back-to-back, transmission timing can be precisely controlled
  - For the purpose of padding between packets, dummy packets (gap packets) are inserted.



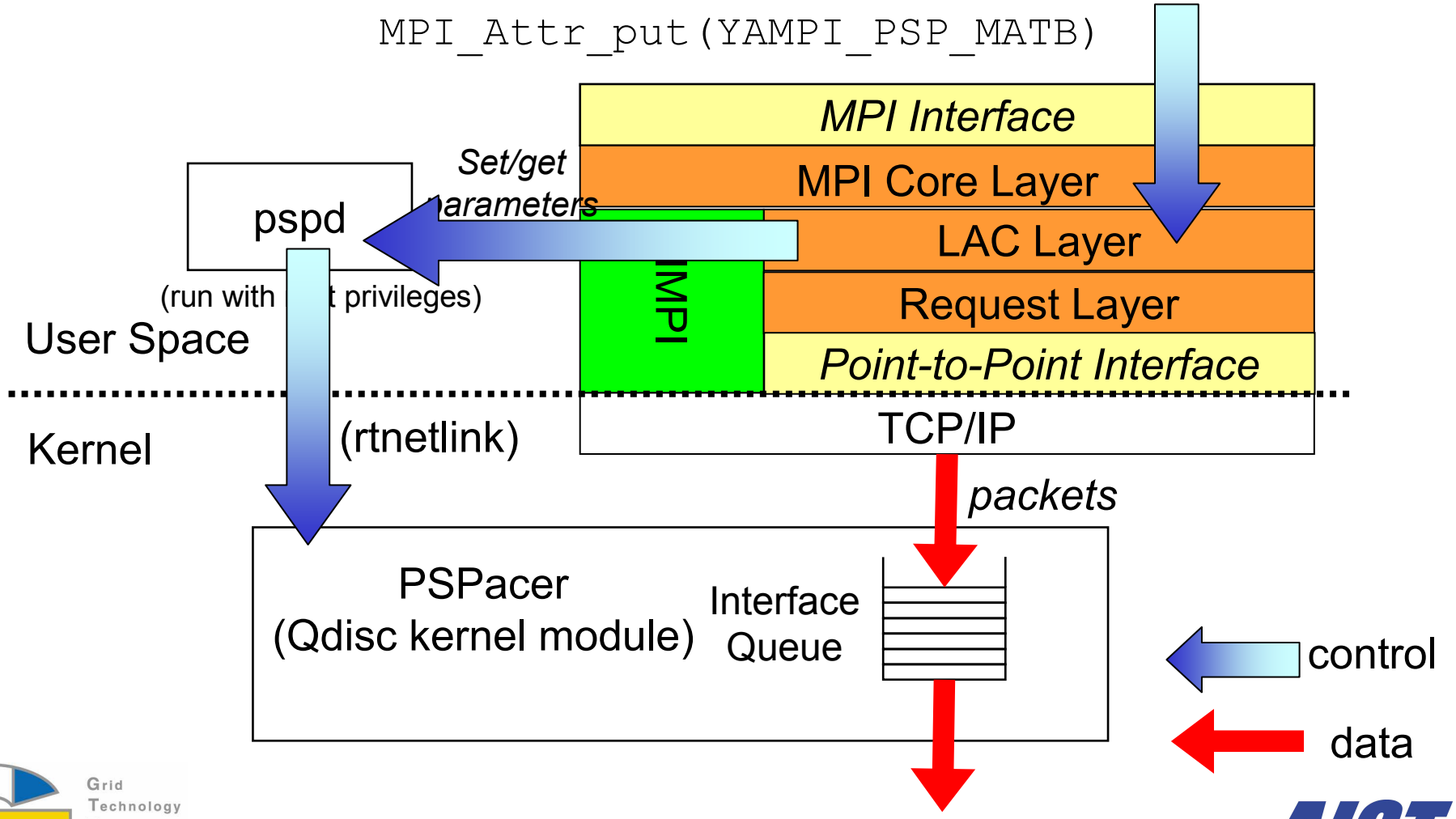
Implementation of a gap packet on Ethernet

- A PAUSE frame（IEEE 802.3x flow control）is used as a gap packet
  - No side effects
    - PAUSE time = 0
    - Discarded at the switch/router's input port
  - No special hardware



R.Takano, et al, "Design and Evaluation of Precise Software Pacing Mechanisms for Fast Long-Distance Networks," PFLDnet2005

PSPacer + GridMPI



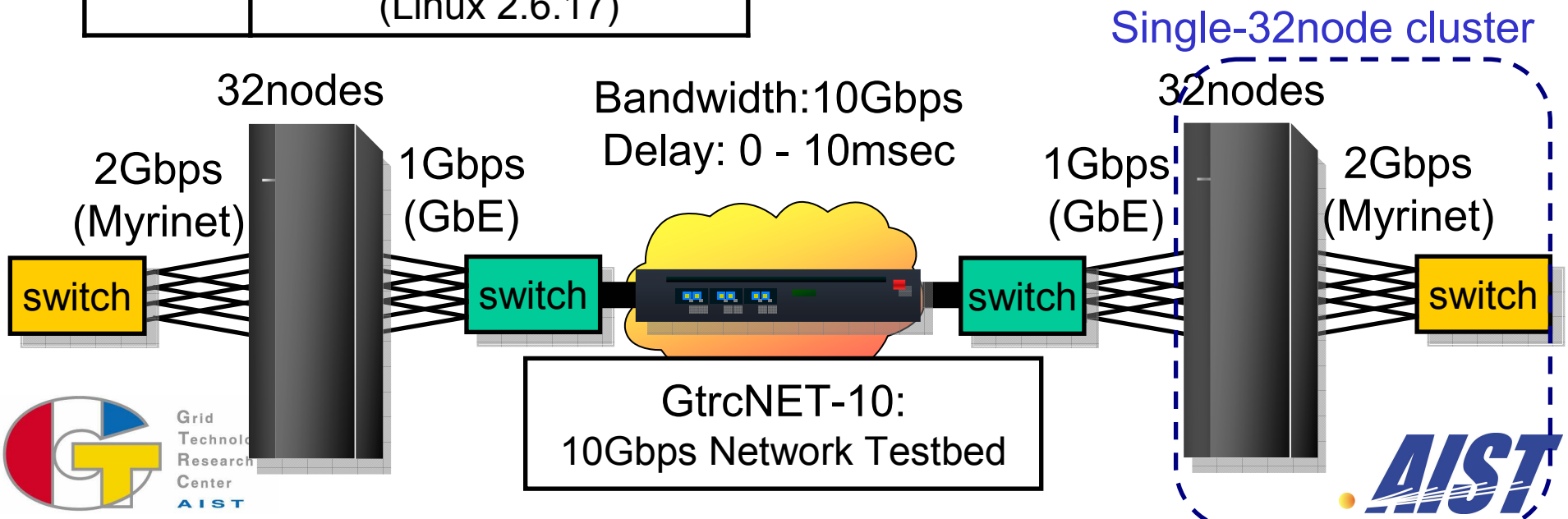
Agenda

- Motivation
- Traffic control method for MPI programs
- Implementation
- Evaluation
  - NPB 3.2 in an emulated WAN environment
  - Analysis of effects of packet pacing
- Conclusion

Experimental Setting

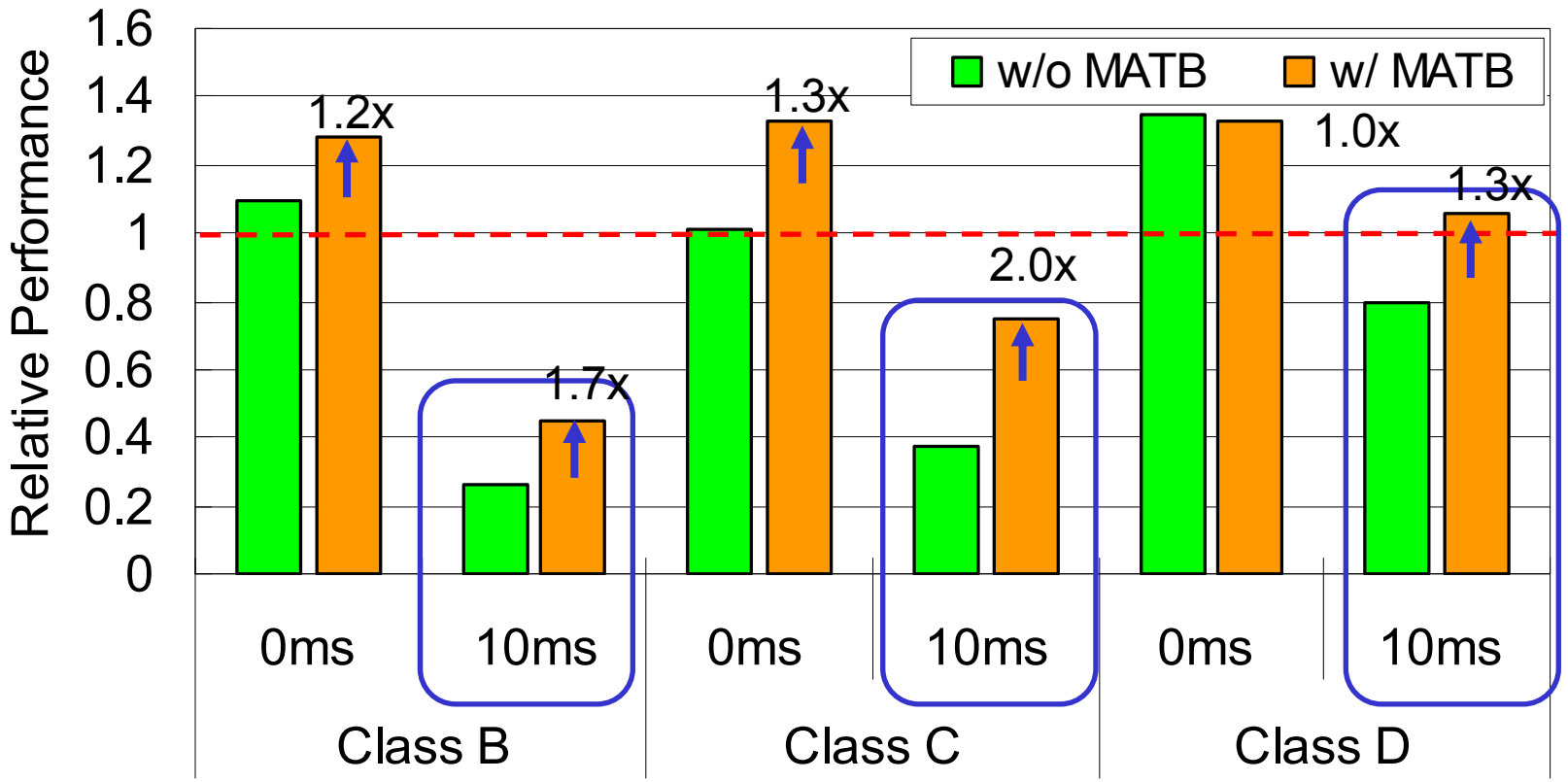
Node PC	
CPU	Opteron/2.0GHz dual
Memory	6GB DDR333
Ethernet	Broadcom BCM5704
Myrinet	Myricom M3F-PCIXD-2
OS	SuSE Enterprise Server 9 (Linux 2.6.17)

Switch	
Ethernet	Huawei-3Com S5648 + optional 10 Gbps port
Myrinet	Myricom M3-SW16-8F + M3-SPINE-8F



CG Benchmark: problem size

(Relative performance normalized to the single 32-node cluster)



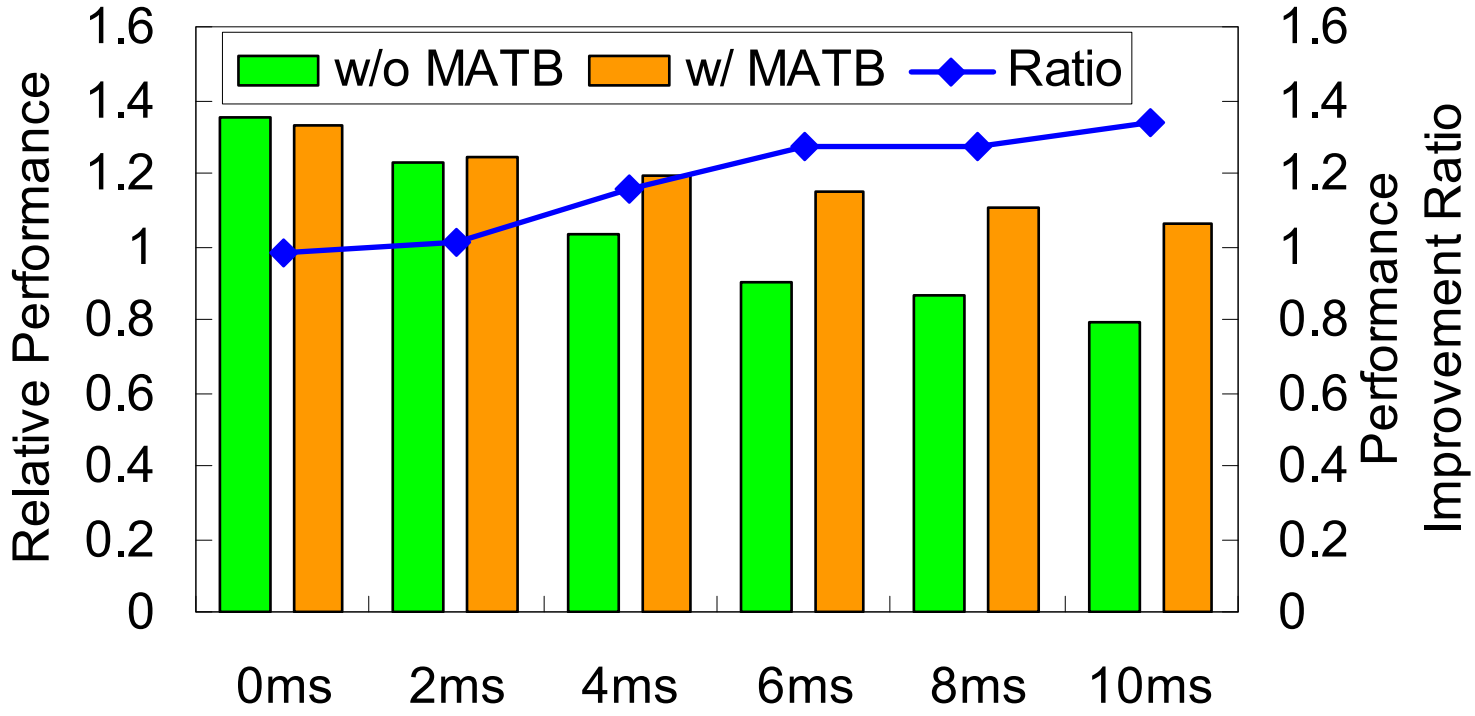
- In class D, the results with MATB are better than the single cluster case even though the delay is 10 ms.

Grid Technology Research Center AIST

AIST

CG Benchmark: delay

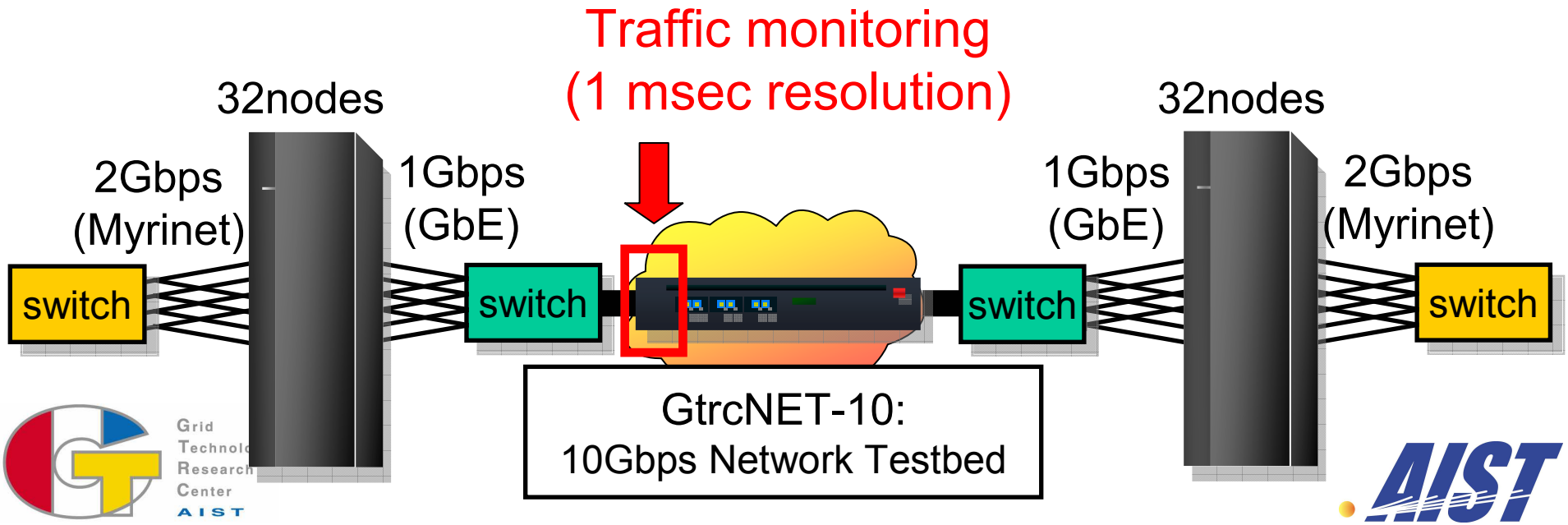
Performance improvement ratio compared with w/o MATB (Class D)



- The proposed method is effective on a Grid environment (In other benchmarks, we observed the same trend)

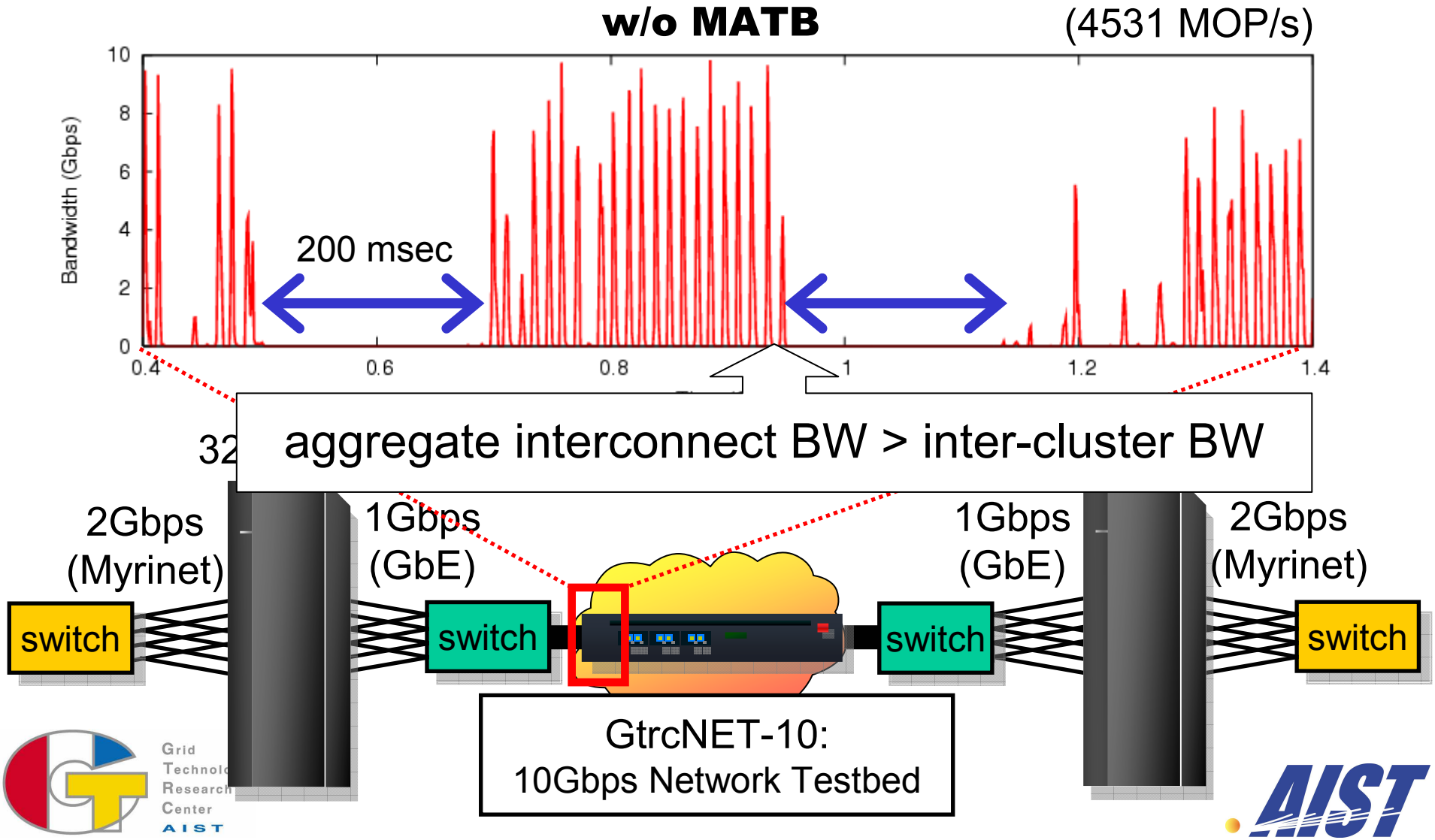
Effects of packet pacing

- Observe aggregate output traffic between clusters in 1 msec resolution by GtrcNET-10
- Target: CG (Class C, 0 msec delay)



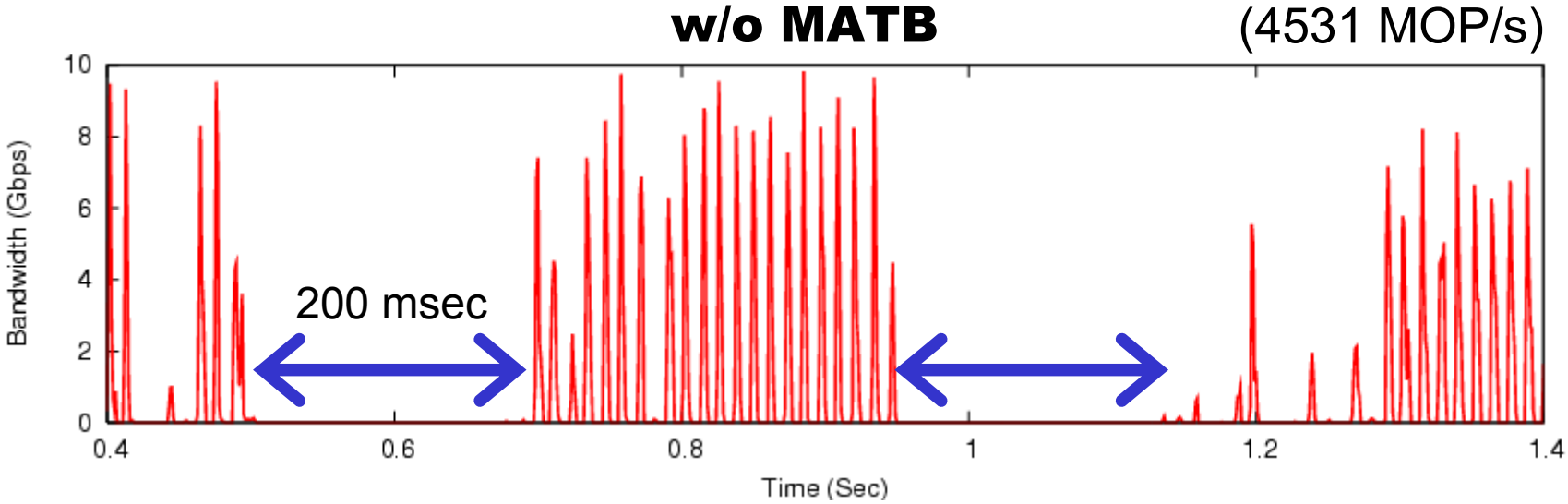
Inter-cluster traffic of CG

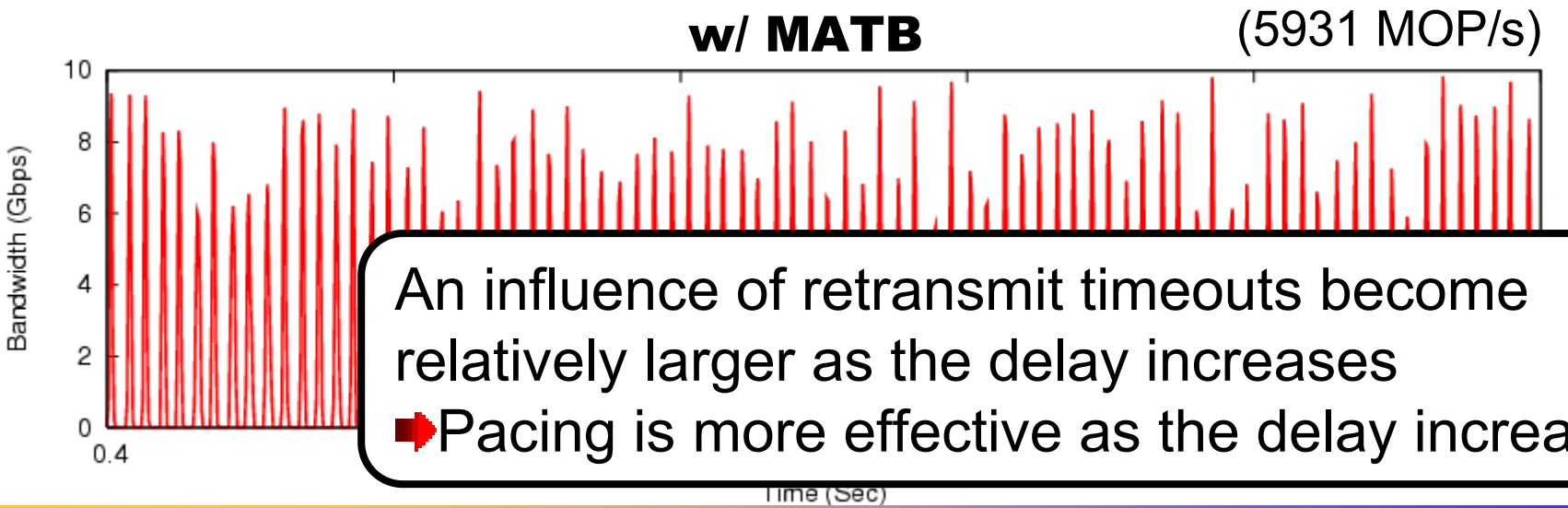
(Class C, 0 msec delay)



Inter-cluster traffic of CG

(Class C, 0 msec delay)





An influence of retransmit timeouts become relatively larger as the delay increases

- Pacing is more effective as the delay increases

Agenda

- Motivation
- Traffic control method for MPI programs
- Implementation
- Evaluation
- Conclusion

Conclusion

- Improving the TCP performance is the key to the successful deployment of MPI programs in a Grid environment
- We have proposed a traffic control method based on the communication pattern of applications
- The experimental results show that it is feasible to connect multiple clusters and run large-scale applications over distances up to 1000km

- GridMPI: http://www.gridmpi.org/
- PSPacer: http://www.gridmpi.org/gridtcp.jsp
- GtrcNET: http://projects.gtrc.aist.go.jp/gnet/

Part of this research was supported by a grant from the Ministry of Education, Sports, Culture, Science and Technology (MEXT) of Japan through the NAREGI (National Research Grid Initiative) Project.

Grid Technology Research Center AIST

AIST