Efficient MPI Collective Operations for Clusters in Long-and-Fast Networks

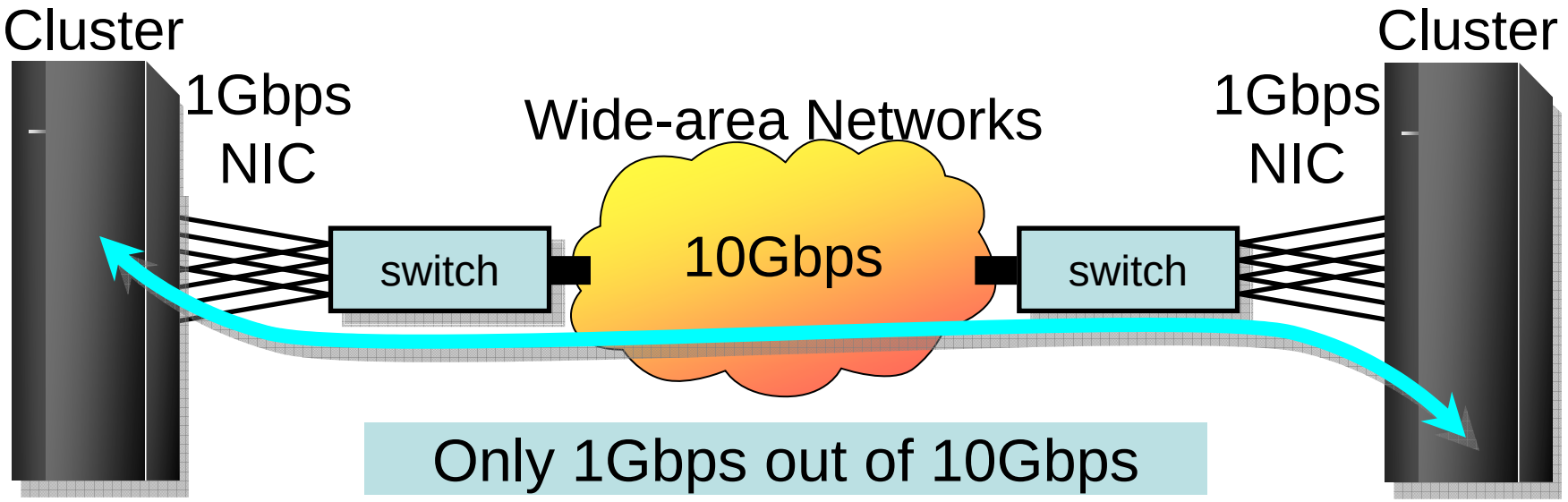
Motohiko Matsuda **Tomohiro Kudoh**
Yuetsu Kodama **Ryousei Takano**

Grid Technology Research Center
National Institute of Advanced Industrial Science and Technology

Yutaka Ishikawa
The University of Tokyo

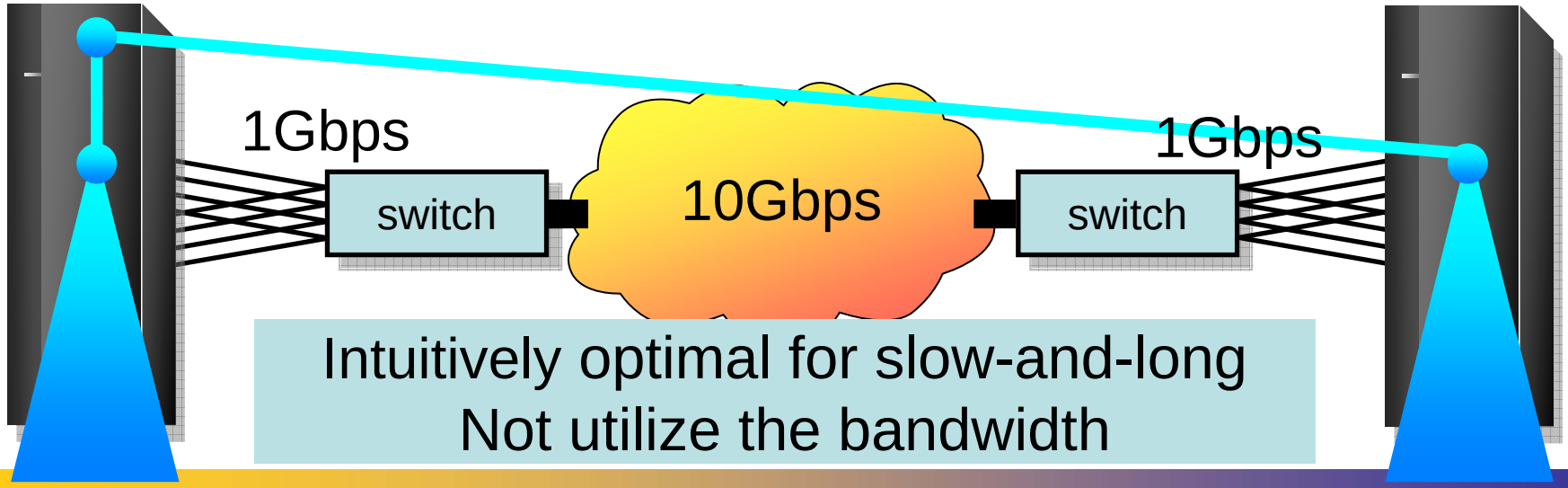
Motivation – Environment (Past/Now)

- Past assumption: long-and-narrow networks
  - Only uses a single wide-area connection at once
  - MPICH-G2, MagPIe, PACX-MPI, ...
- Current state: long-and-fast networks
  - 10 or 40Gbps Networks vs 1Gbps majority NIC
  - Need new collectives using multiple connections



Reviewing Existing Collectives

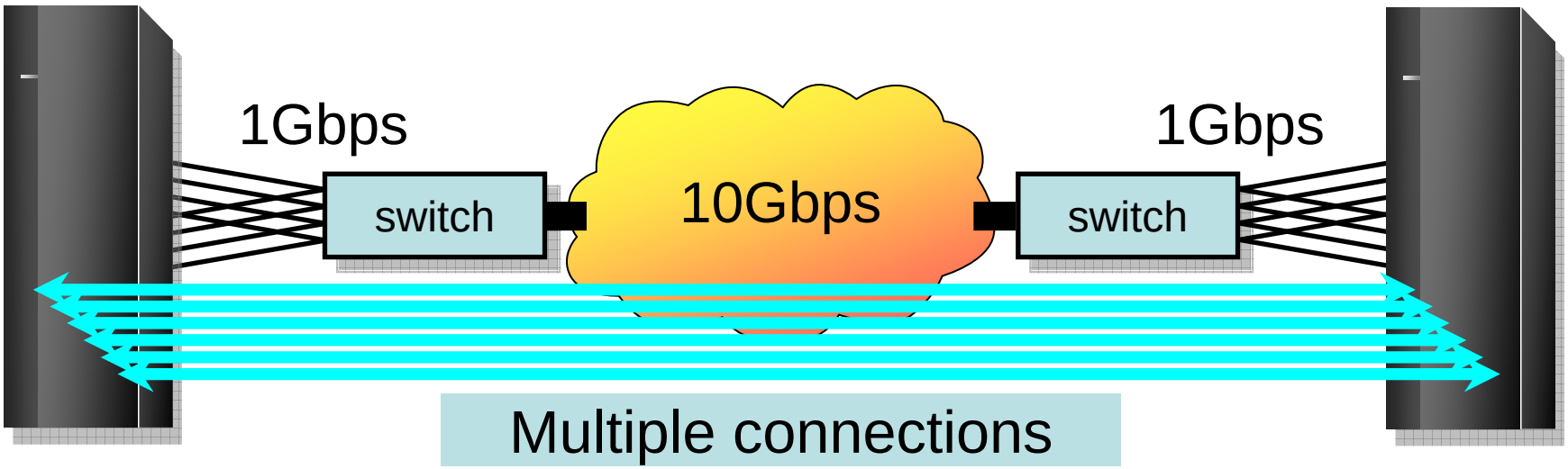
- Bcast (far-first):
  - (Step#1) Send data to the other cluster (by the root)
  - (Step#2) Bcast data in each cluster
- Allreduce (two-tier):
  - (Step#1) Reduce in both clusters
  - (Step#2) Exchange data and reduce (by the roots)
  - (Step#3) Bcast reduced data in each cluster



Intuitively optimal for slow-and-long
Not utilize the bandwidth

Algorithms Needed

- Utilize the bandwidth of inter-cluster network
  - Use multiple connections
- Avoid congestion
  - Control the #connections



Search Algorithms from ones for Clusters

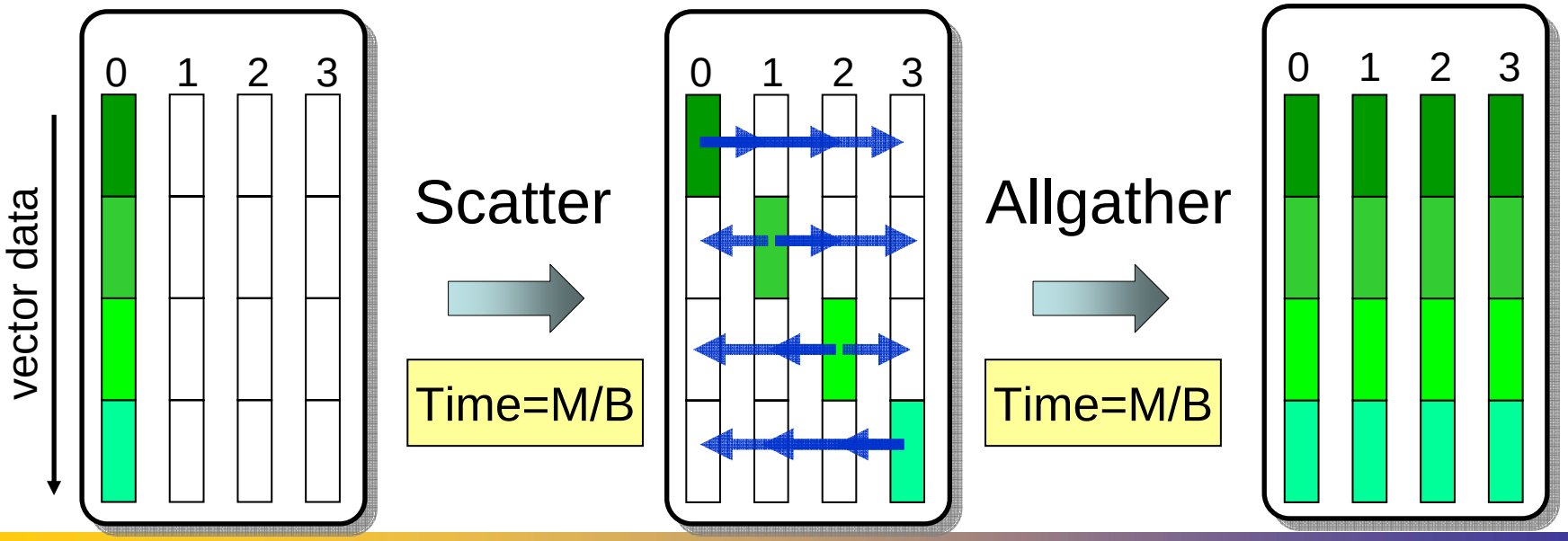
- Existing MPI algorithms cannot be extended, becuase
  - Only a root node has data
- Efficient algorithms for high bi-section bandwidth environment
  - Fast Bcast by *van de Geijn*, et al
  - Fast Allreduce by *Rabenseifner*

Simple Cost Model (in Time)

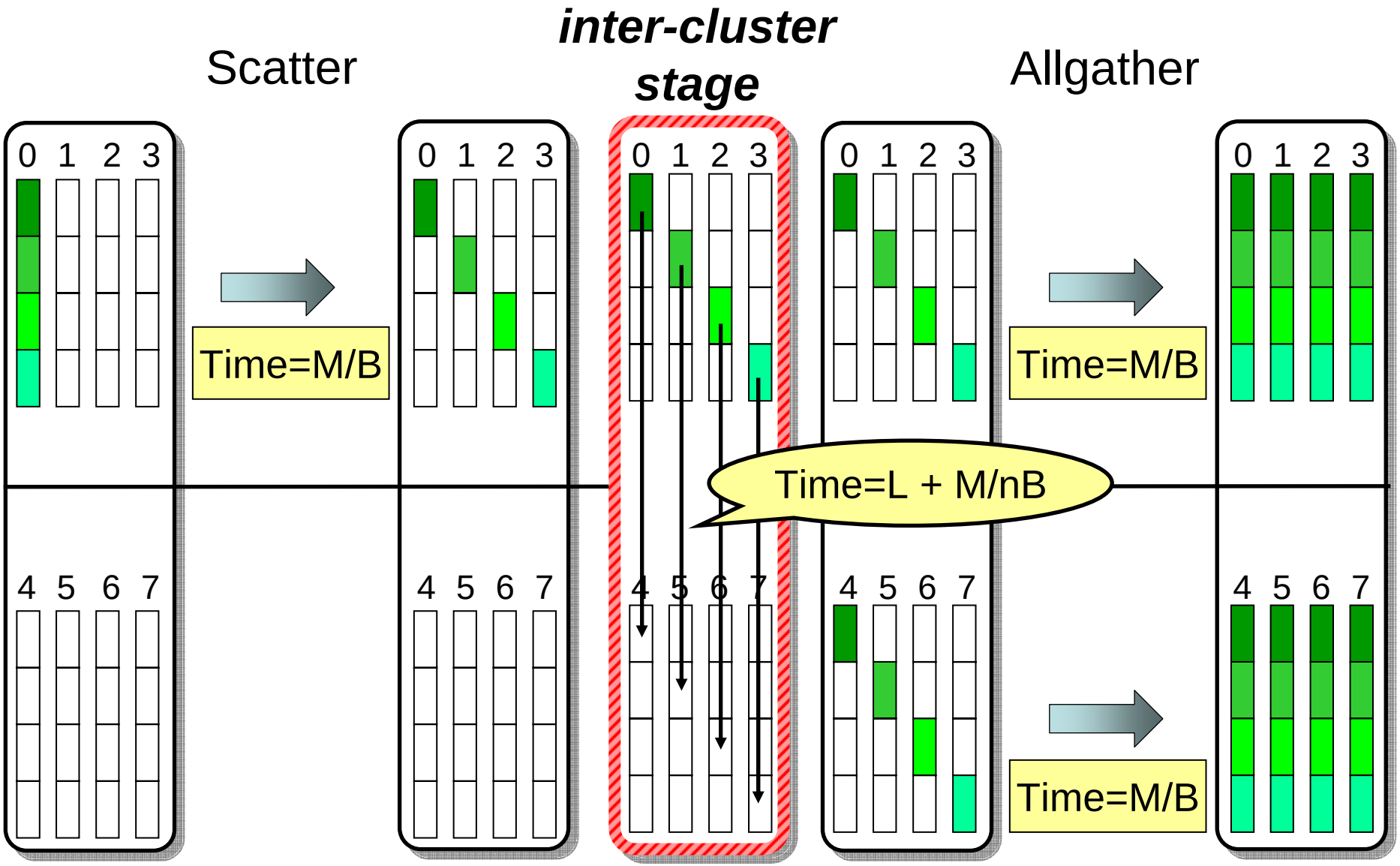
- Very simplified communication cost model:
  - M: Message size
  - B: Bandwidth of a node (NIC)
  - L: Inter-cluster latency
  - n: #connections between clusters
- Time of message transfer:
  - Intra-cluster Time= M/B
  - Inter-cluster Time= (L + M/nB)
- Assumption:
  - Ignore intra-cluster latency
  - Ignore communication overhead

van de Geijn Bcast (original version)

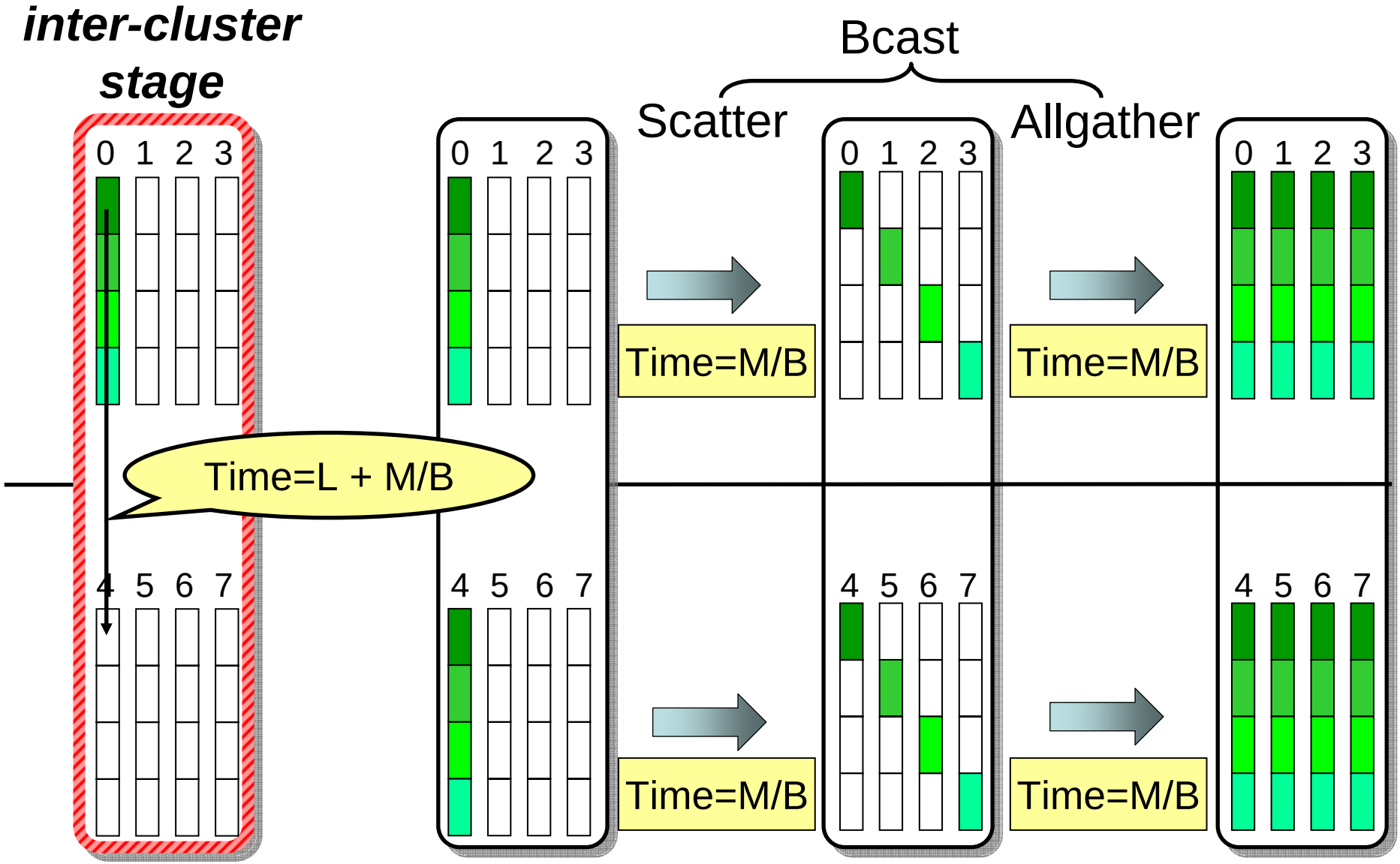
- Bcast by van de Geijn, et al
  - Fast in high bi-section bandwidth environment
  - Very efficient for long messages
- Algorithm:
  - Scatter + Allgather
  - Start Bcast from multiple roots after Scatter



Modified van de Geijn Bcast



cf. Far-First Bcast (existing algorithm)

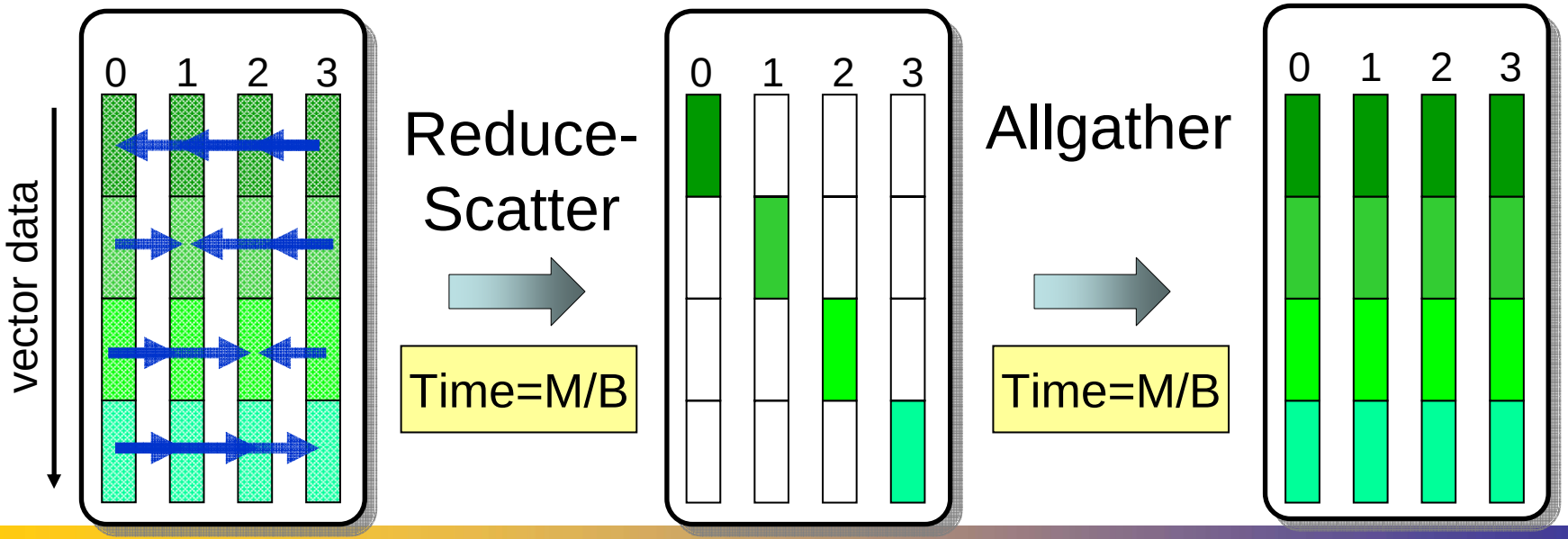


Bcast Time Cost Summary

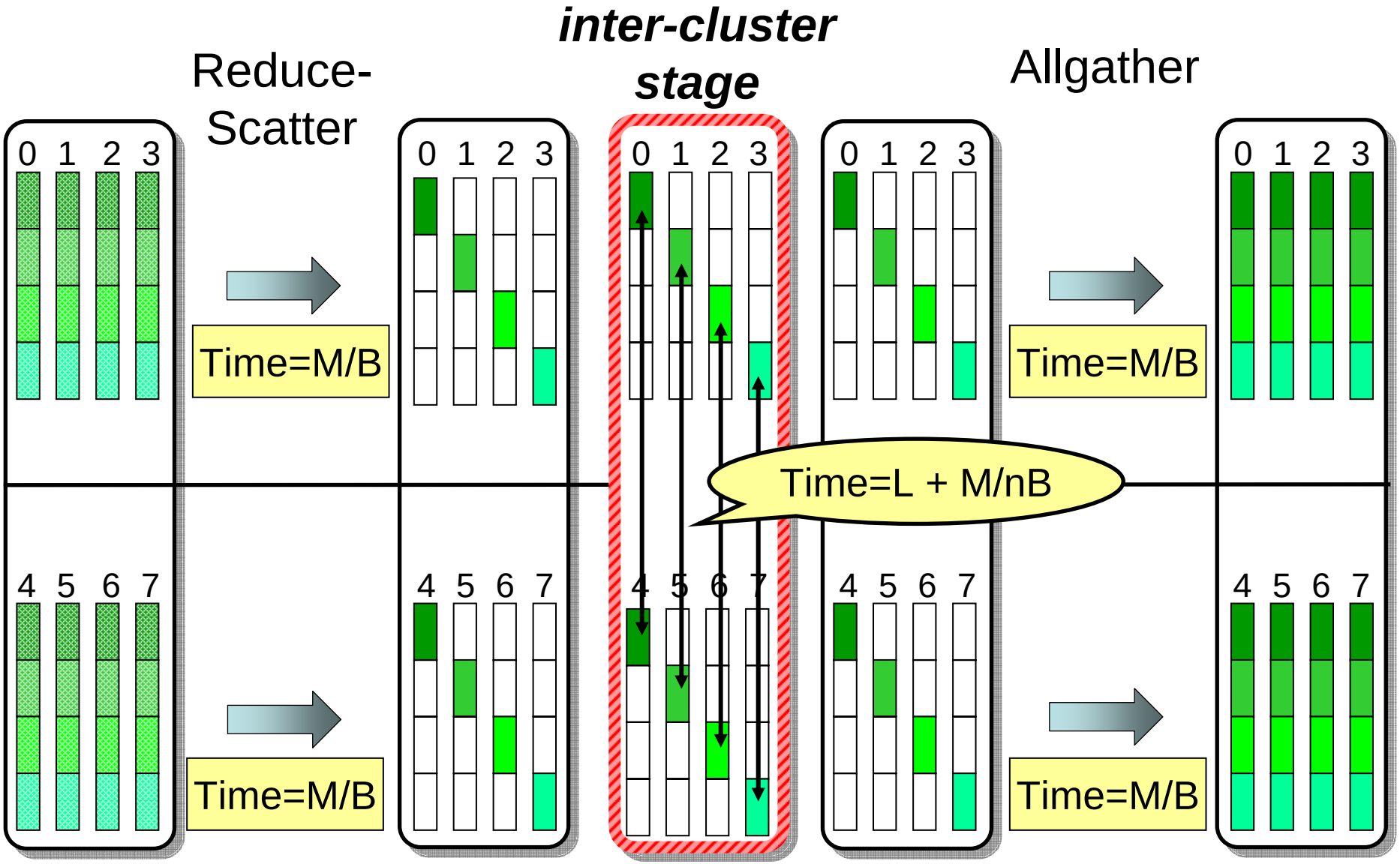
- Modified van de Geijn Bcast
  - *Time*= L + M/nB + M/B + M/B
- cf. Far-First Bcast (existing algorithm):
  - *Time*= L + M/B + (M/B + M/B)

Rabenseifner Allreduce (original version)

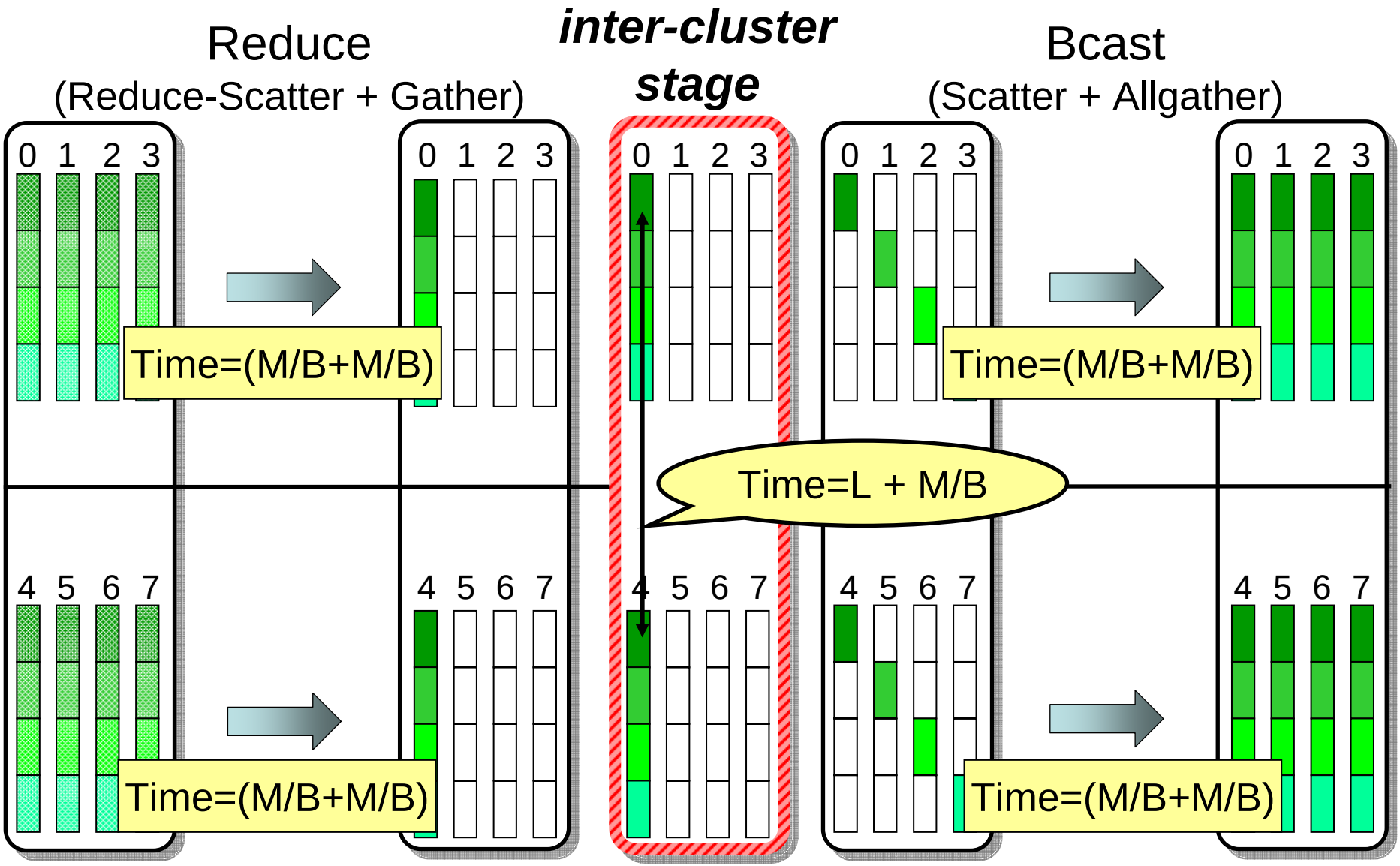
- Allreduce designed by Rabenseifner
  - Fast in high bi-section bandwidth environment
  - Very efficient for large messages
- Algorithm:
  - Reduce-Scatter + Allgather
  - Based on an similar idea of van de Geijn Bcast



Modified Rabenseifner Allreduce



cf. Two-Tier Allreduce (existing algorithm)



Allreduce Time Cost Summary

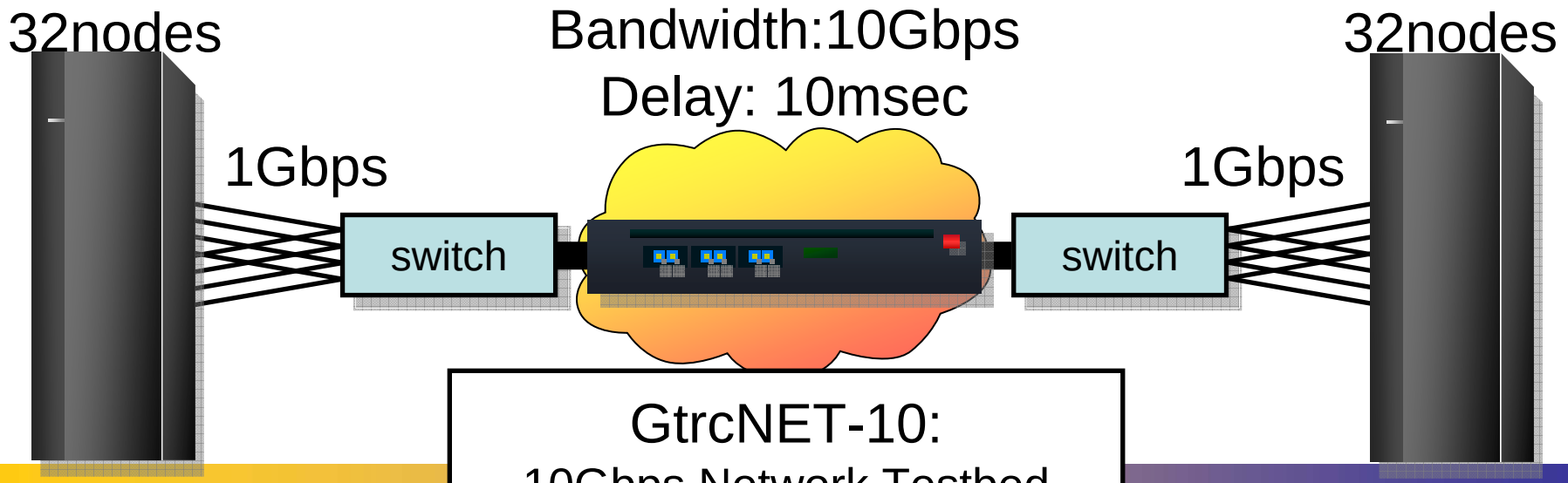
- Modified Rabenseifner Allreduce
  - *Time*= L + M/nB + M/B + M/B
- cf. Two-Tier Allreduce (existing algorithm):
  - *Time*= L + M/B + (M/B + M/B) + (M/B + M/B)

Avoiding Congestion

- Restrict the #connections
  - Selected nodes can communicate
  - Other nodes forward messages to the selected nodes
- 10Gbps network with 1Gbps NIC
  - Up to 10 connections
  - Totally avoids congestion

Experimental Setting

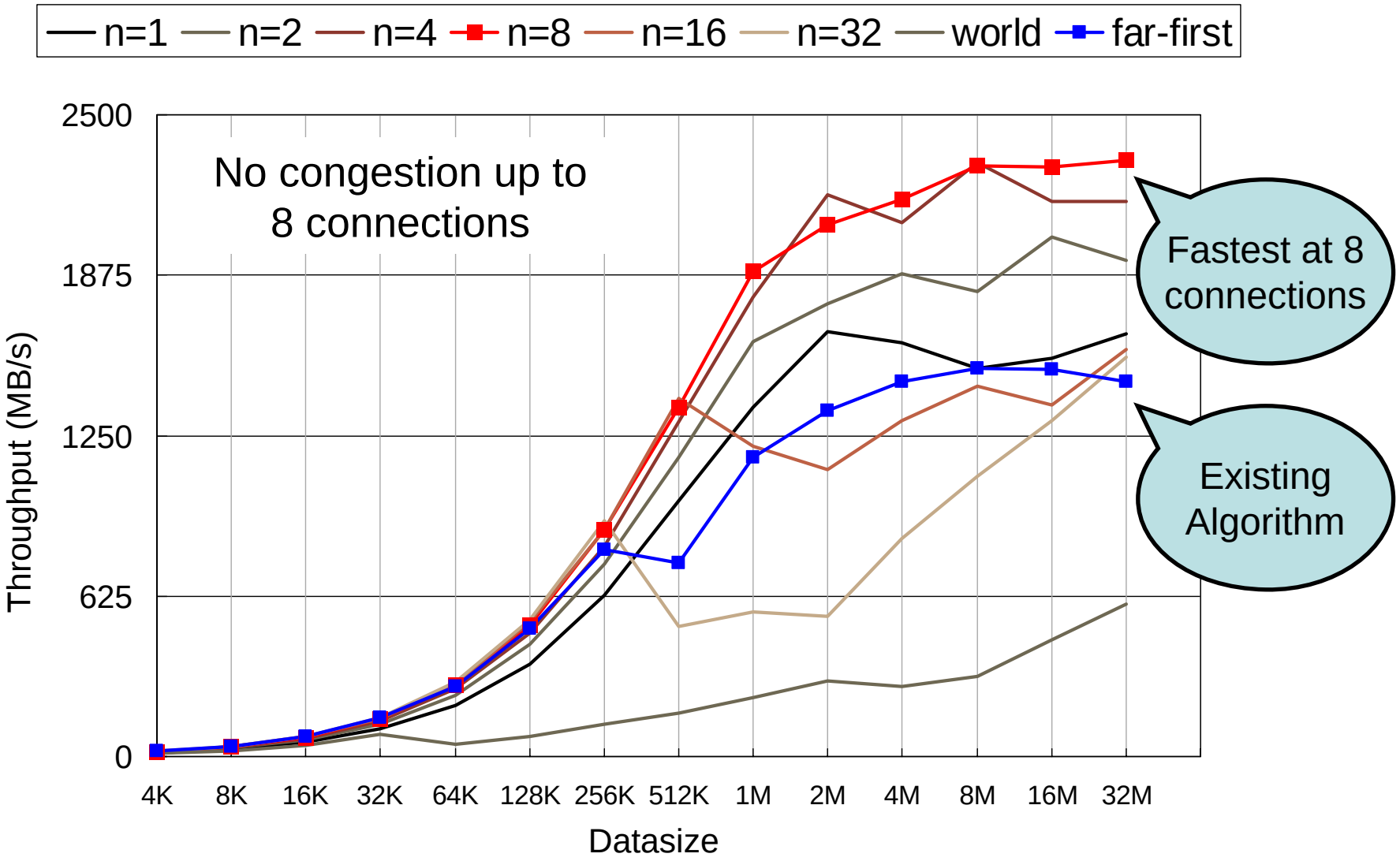
CPU	Opteron (2.0GHz) x 2
Memory	6GB DDR333
NIC	Broadcom BCM5704
OS	Fedore Core 5
Switch	Huawei-3Com Quidway S5648



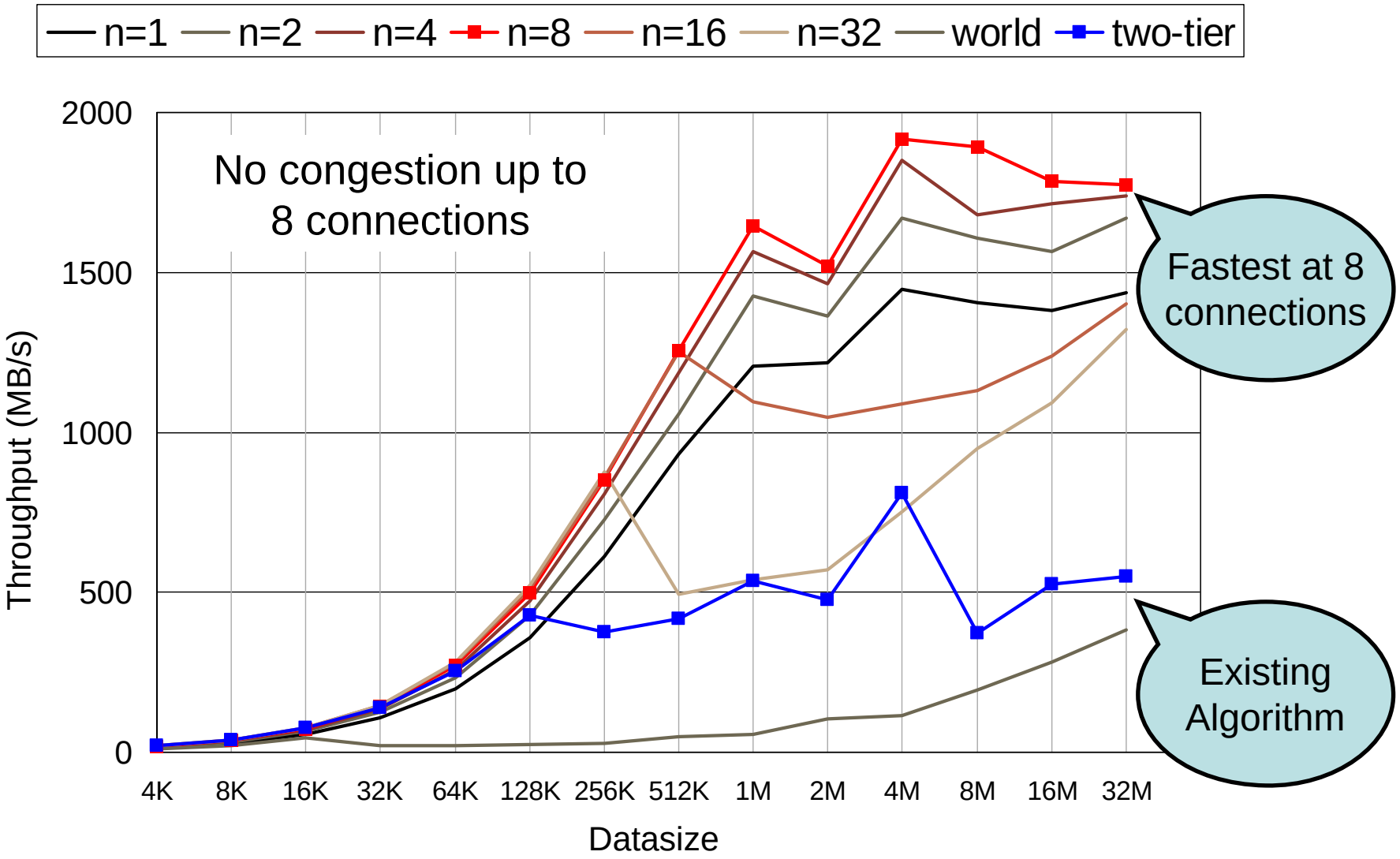
Comparisons by Throughput

- Normalized throughput for comparison between algorithms
  - Throughput value is *inverse of time*
- Bcast
  - Throughput (MB/s)= $\frac{\text{Message-Size x \#Nodes}}{\text{Bcast-Time}}$
- Allreduce
  - Throughput (MB/s)= $\frac{\text{Message-Size x \#Nodes}^2}{\text{Allreduce-Time}}$

Bcast (delay=10ms)



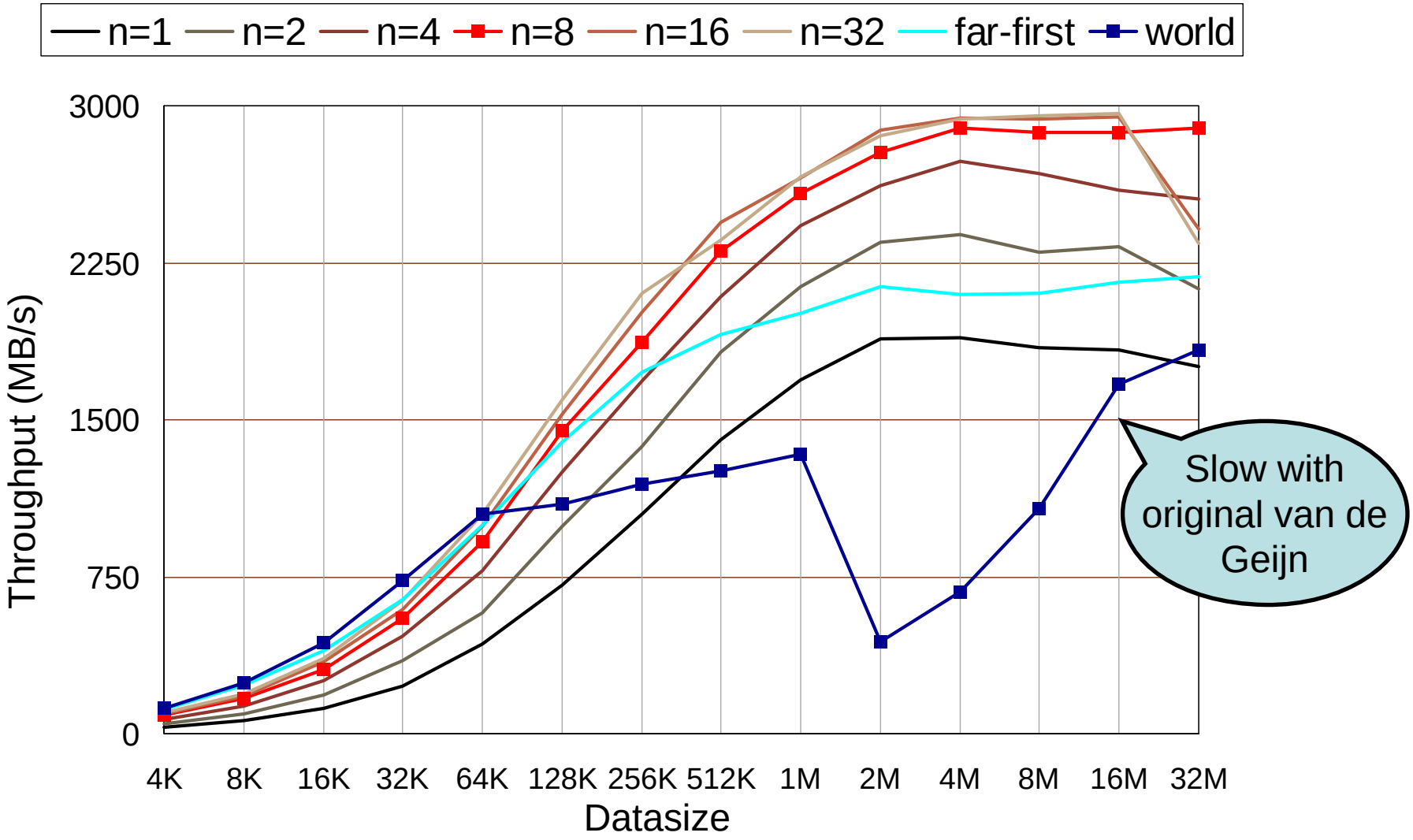
Allreduce (delay=10ms)



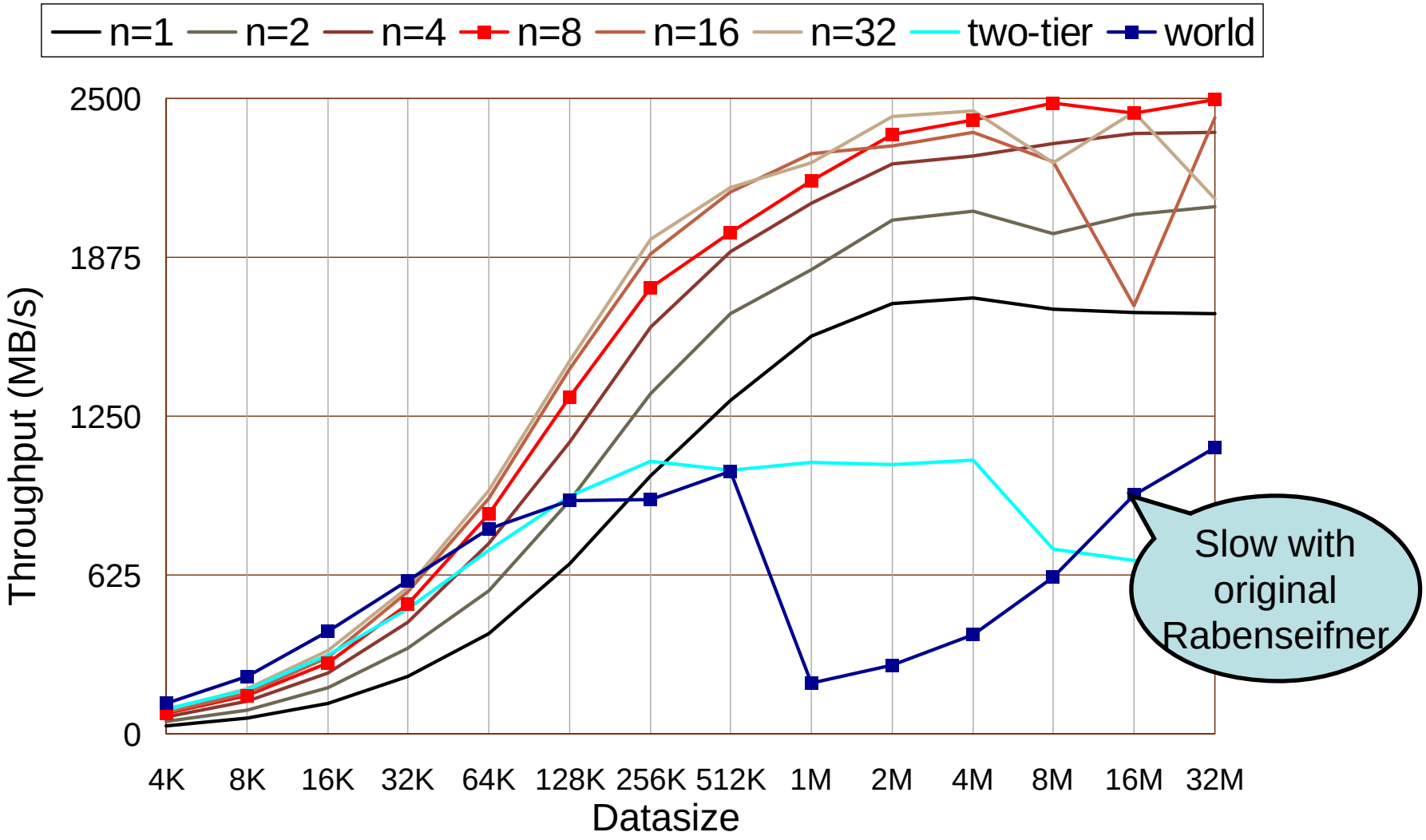
In Cluster Environment

- Reduced bi-section bandwidth environment, such as
  - Ethernet with multiple switches
  - Fat-Tree with reduced upper-level links
- Experimental setting:
  - 10Gbps up-link (same)
  - Delay time set to 0msec

Bcast (no delay)



Allreduce (no delay)



About Other MPI Collectives

- Reduce-scatter/Allgather/Reduce:
  - Subpart of Allreduce
- Scatter/Gather:
  - Limited by the bandwidth of a node
  - Smart algorhtims are unlikely
- Barrier:
  - Zero message size
- Alltoall:
  - Highly congesting
  - Much of TCP/IP issues

Summary

- Base algorithms for high bi-section bandwidth networks
  - van de Geijn Bcast
  - Rabenseifner Allreduce
- Collectives for long-and-fast networks
  - Perform inter-cluster communication in the middle stage
  - Utilize multiple connections
  - Avoid congestion by limiting the #connections

GridMPI™

http://www.gridmpi.org

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.

END