Fast Barrier Synchronization for InfiniBand™

Torsten Hoefler

Chair of Computer Architecture
Technical University of Chemnitz

IPDPS'06 - CAC'06 Workshop
Rhodes Island, Greece
25th April 2006

CHEMNITZ UNIVERSITY OF TECHNOLOGY

CHEMNITZ UNIVERSITY OF TECHNOLOGY

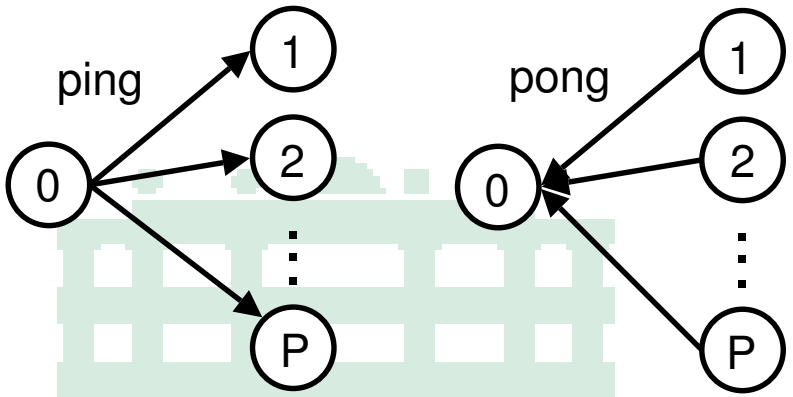
CHEMNITZ UNIVERSITY OF TECHNOLOGY

1:n n:1 Microbenchmark

- Developed to analyze the InfiniBand™ network
- Especially for collective communication
- Measures single message performance (RDTSC)
- MPI based
- Supports (nearly) all transport types

CHEMNITZ UNIVERSITY OF TECHNOLOGY

1:n n:1 Microbenchmark - principle



1. (0): Take time
2. (1..n-1): Send a single message to n-1 hosts
3. (1..n-1): Hosts respond immediately
4. (0): Wait for message receiption from all hosts
5. (0): Take time

CHEMNITZ UNIVERSITY OF TECHNOLOGY

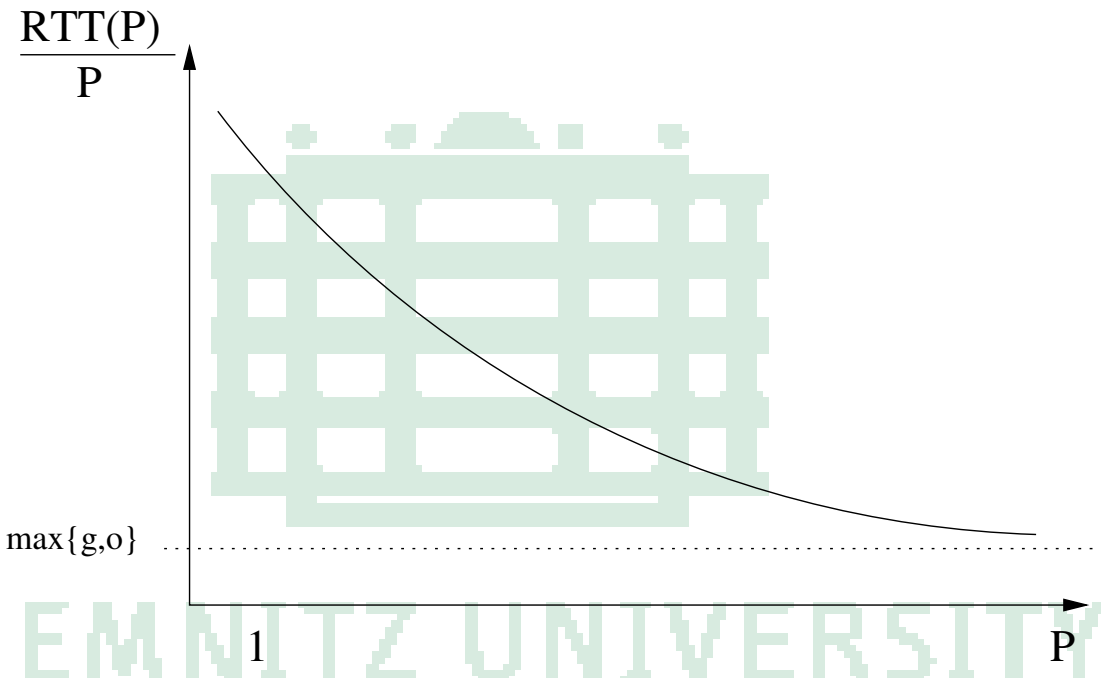
The LogP Model

LogP model by Culler et.al. 1993

LogP Parameters

- L - Latency
- g - Bandwidth-limiting Gap between consecutive messages ($g \approx 1/BW$)
- o - Send-/Receive Overhead
- P - Number of involved Processors

LogP Prediction



CHEMNITZ UNIVERSITY OF TECHNOLOGY

- $RTT(P)/P = (4o + 2L + (P-1) \cdot max\{g,o\})/P$

CHEMNITZ UNIVERSITY OF TECHNOLOGY

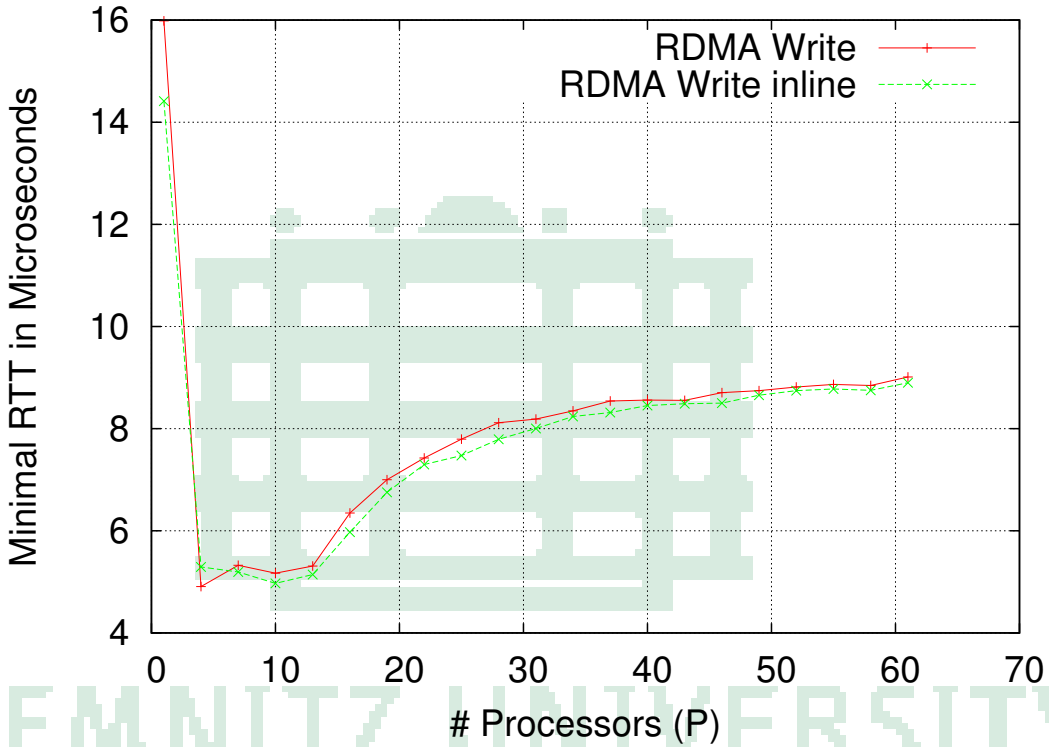
1:n n:1 Benchmark Results

Test Environment

- 8 Nodes
- Dual Xeon 2.066 GHz
- Red Hat Linux release 9 (Shrike)
- Kernel: 2.4.27 SMP
- HCA: Mellanox ”Cougar” (MTPB 23108)

OF TECHNOLOGY

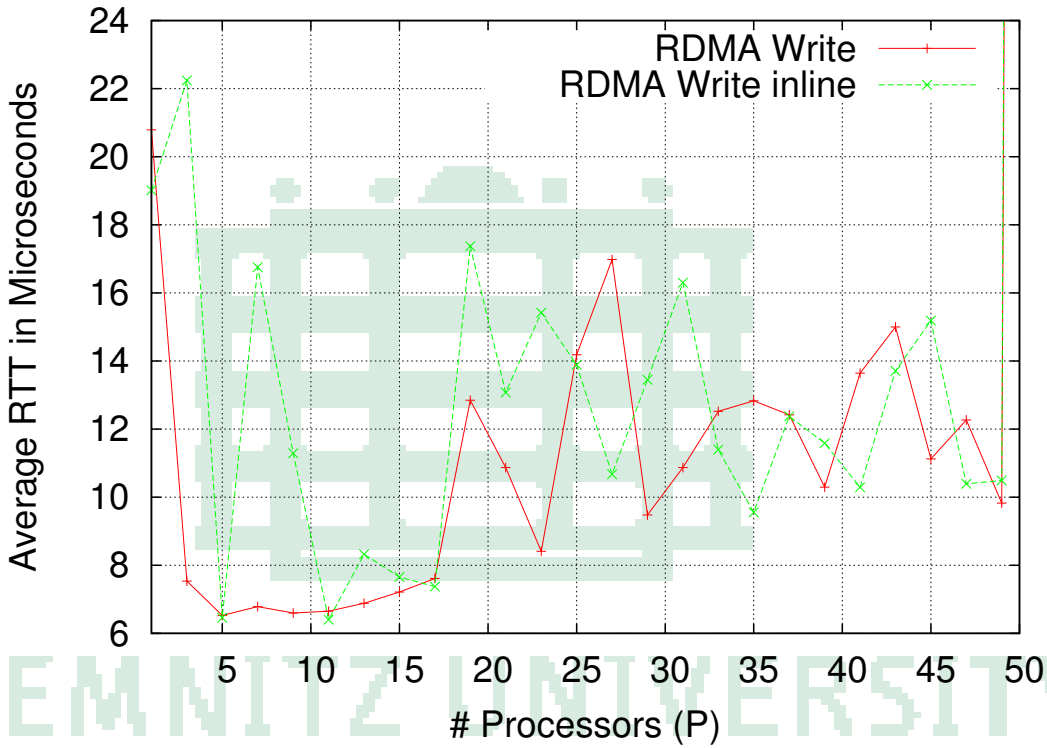
1:n n:1 Benchmark Results



CHEMNITZ UNIVERSITY OF TECHNOLOGY

- RDMA/W - fastest transport type in our tests
- Graph shows minimal values
- We benefit from sending to multiple hosts simultaneously
- Atomic was not available on our HCAs

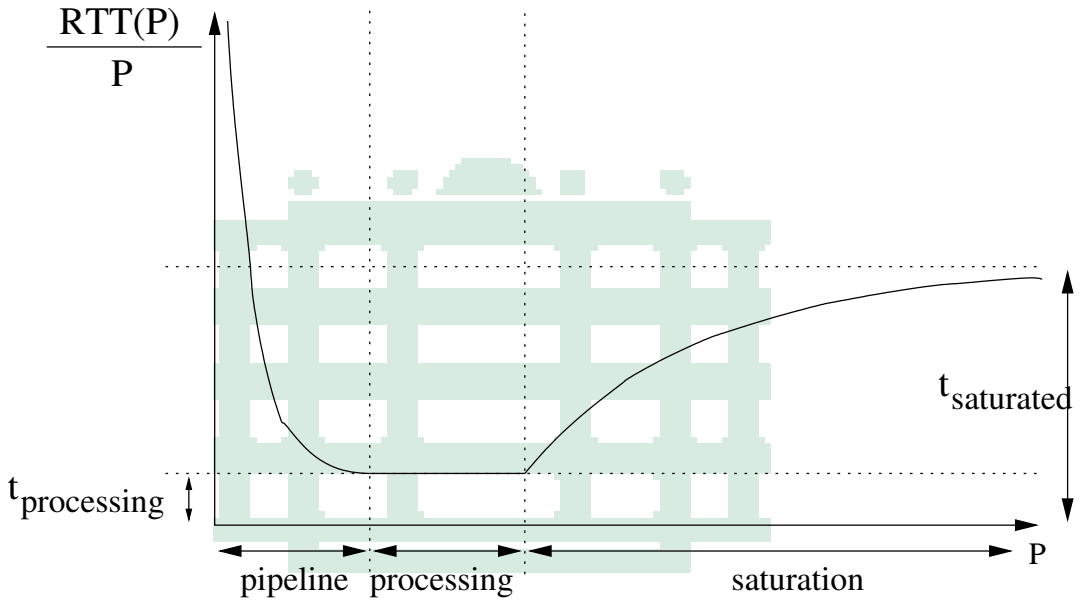
1:n n:1 Benchmark Results



CHEMNITZ UNIVERSITY OF TECHNOLOGY

- Average Graph has many outliers
- Still same ”shape”

A possible Explanation - The LoP Model

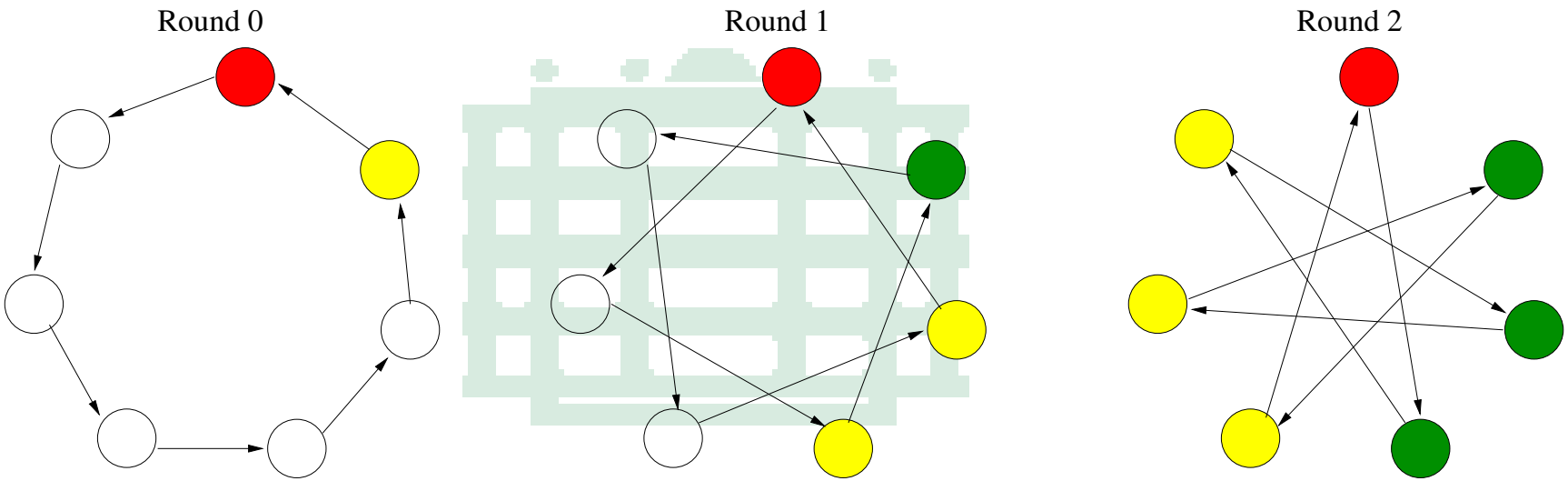


CHEMNITZ UNIVERSITY OF TECHNOLOGY

- Pipeline startup function - hardware pipe, caches
- Minimal processing time - hardware
- Network saturation - network hardware / transceiver

CHEMNITZ UNIVERSITY OF TECHNOLOGY

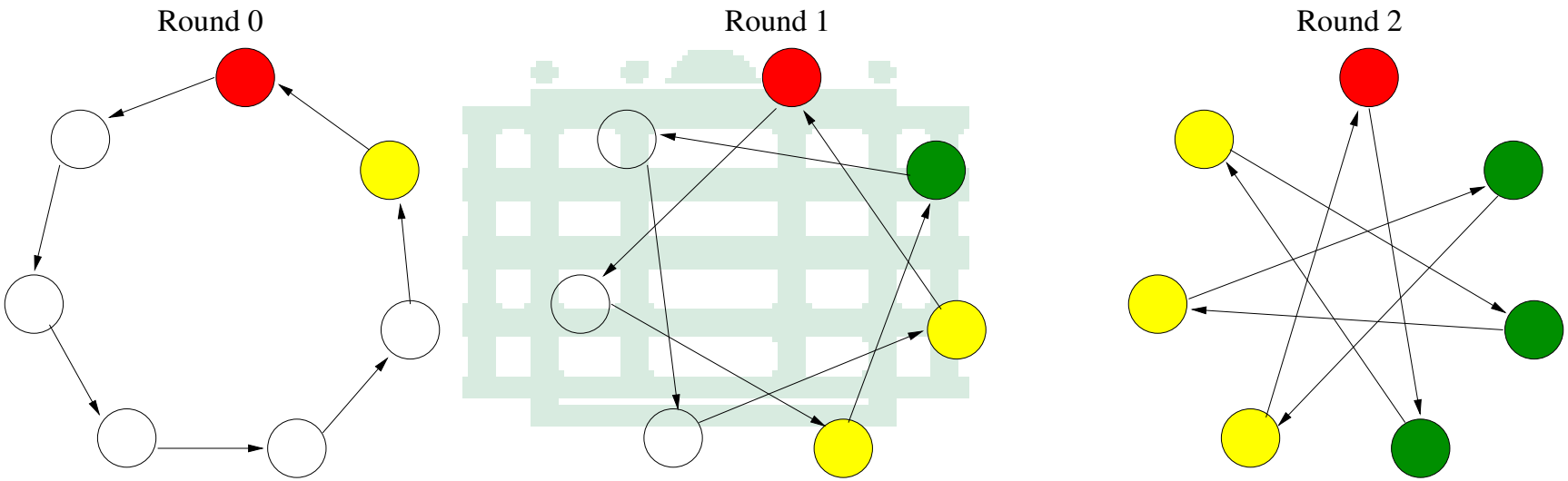
The Dissemination Algorithm



CHEMNITZ UNIVERSITY OF TECHNOLOGY

- Logarithmic running time ($O(log_2 P)$)
- Works with non-power of two P

Dissemination - Peer selection

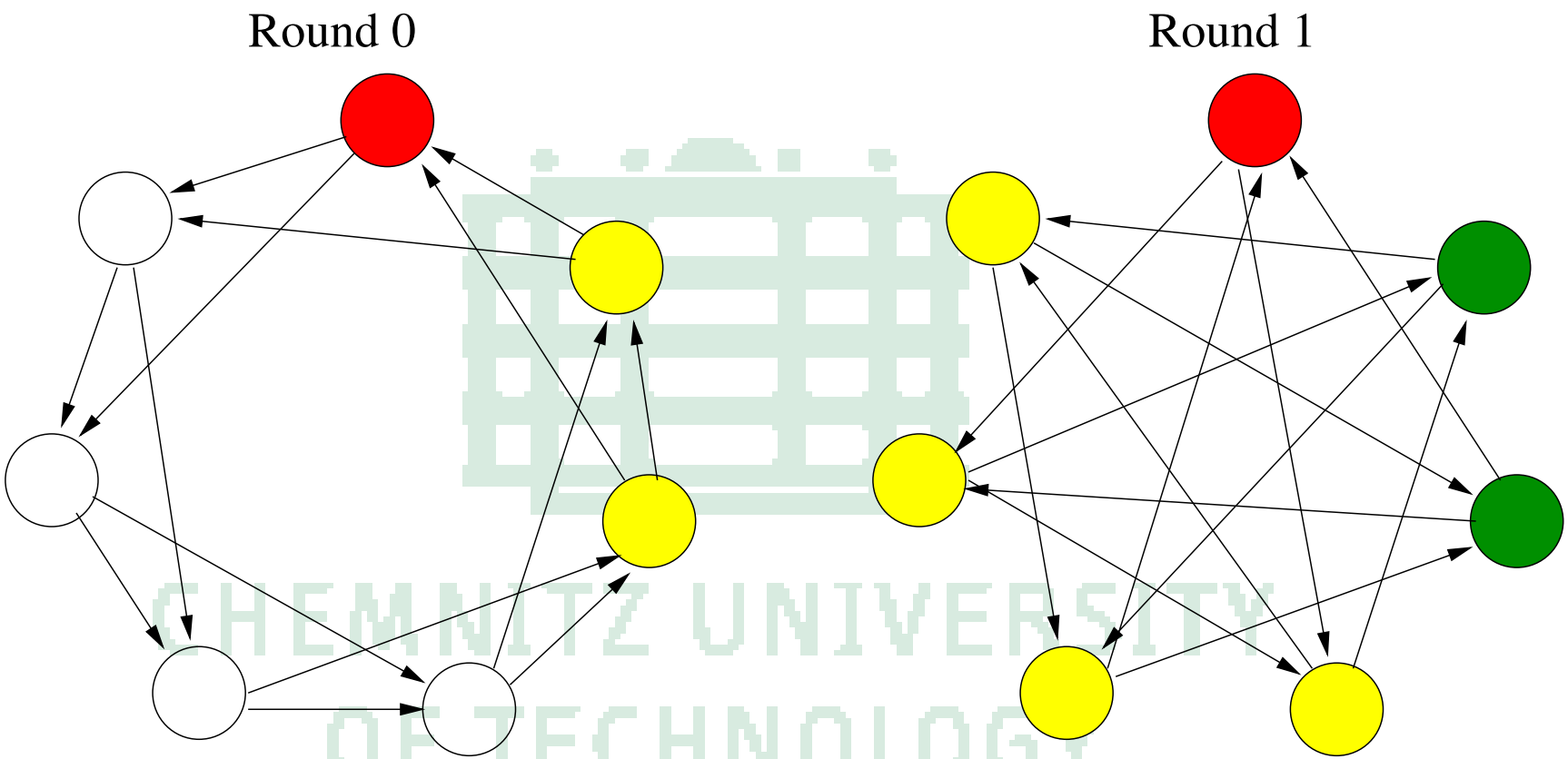


CHEMNITZ UNIVERSITY OF TECHNOLOGY

- $speer = (p + 2^r) \ mod \ P$
- $rpeer = (p - 2^r) \ mod \ P$

CHEMNITZ UNIVERSITY OF TECHNOLOGY

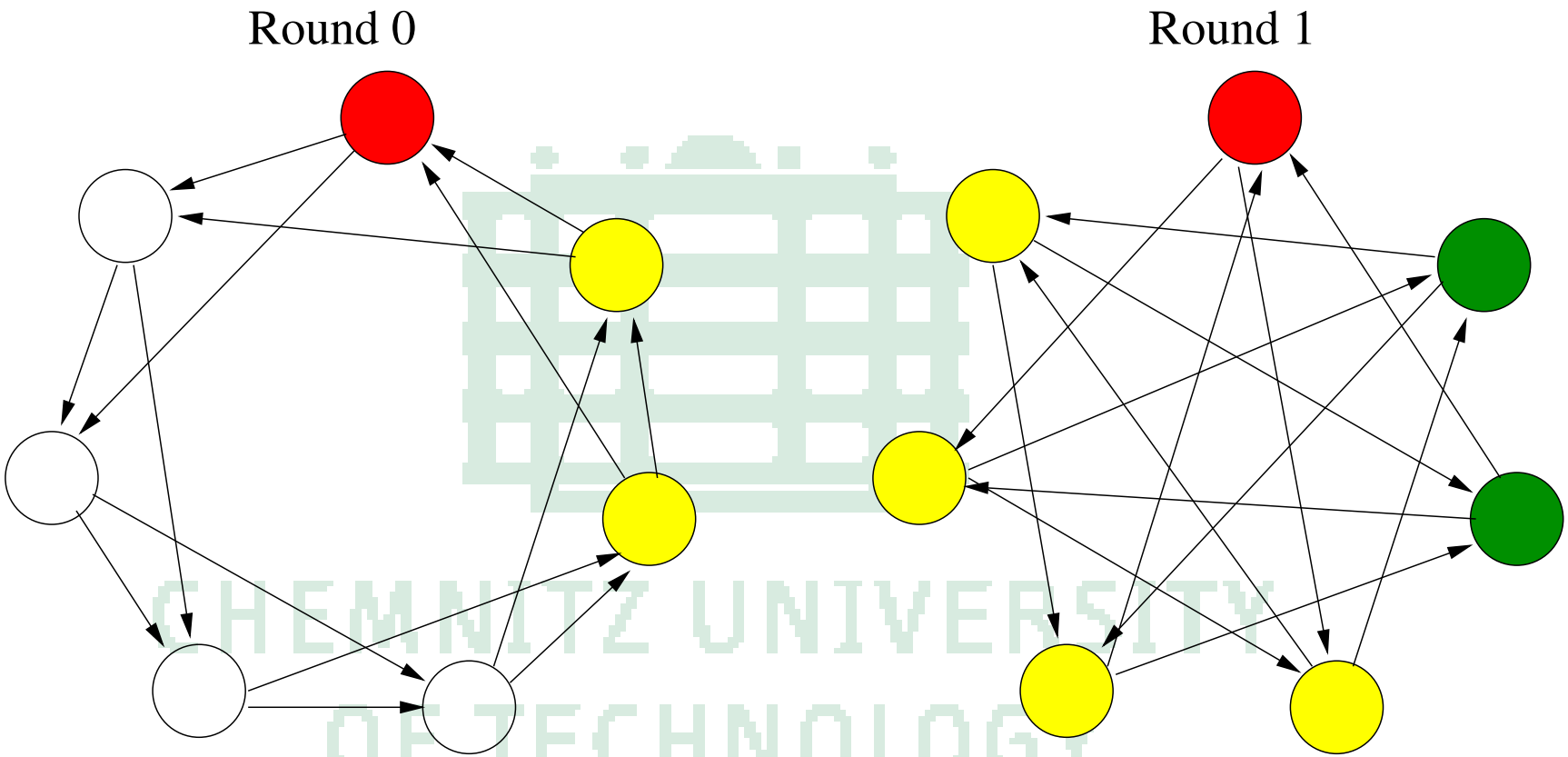
The n-way Dissemination Algorithm



CHEMNITZ UNIVERSITY OF TECHNOLOGY

- Logarithmic running time ($O(log_2 P) - O(log_n P)$?)
- Works with non-power of n P

n-way Dissemination - Peer selection



CHEMNITZ UNIVERSITY OF TECHNOLOGY

- $speer_i = (p + i \cdot (n+1)^r) \bmod P$
- $rpeer_i = (p - i \cdot (n+1)^r) \bmod P$

CHEMNITZ UNIVERSITY OF TECHNOLOGY

The n-way Dissemination Algorithm

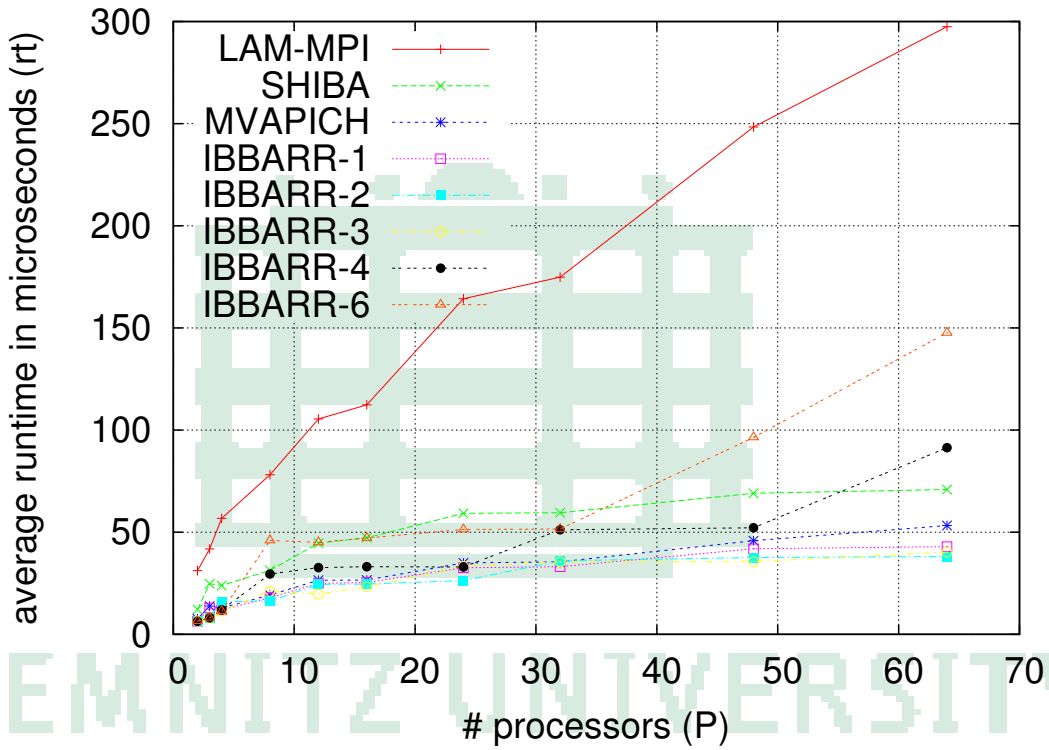
Implementation Details

- Implementation as collv1 component in Open MPI
- Communication peers are precomputed

Benchmark Details

- LAM/MPI 7.1.1
- TUC SHIBA 1.0
- MVAPICH 0.9.4

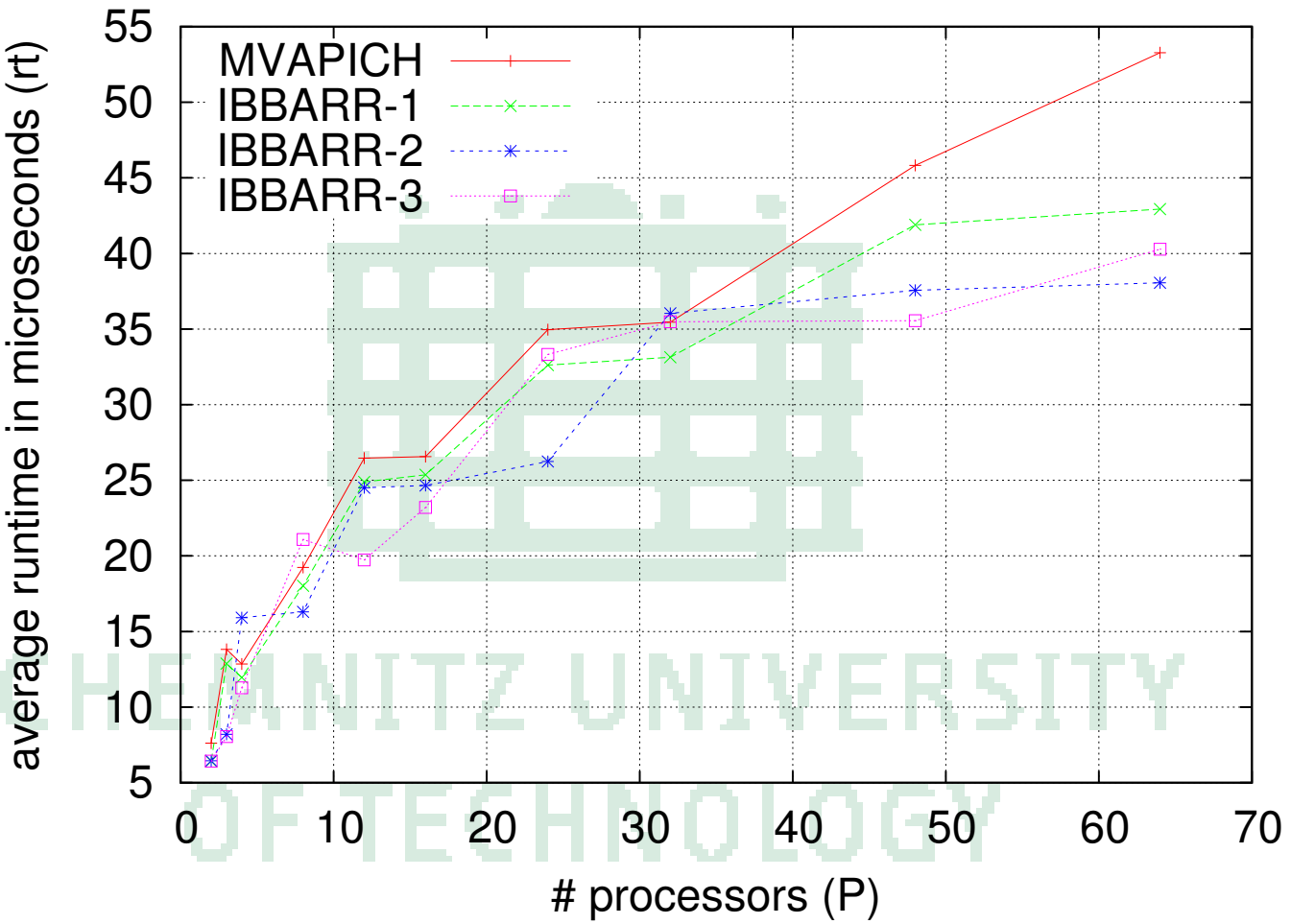
Benchmark Results



CHEMNITZ UNIVERSITY OF TECHNOLOGY

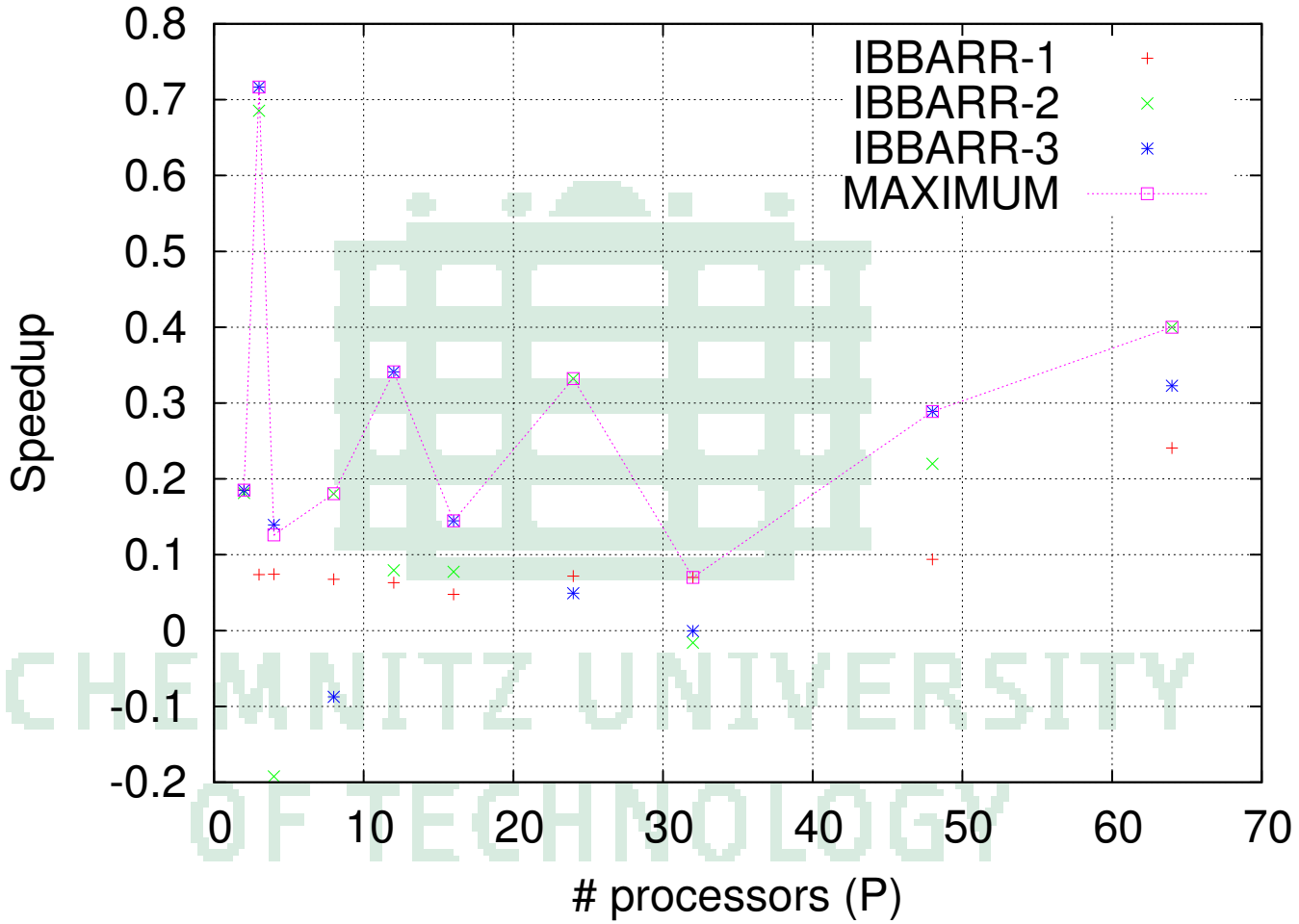
- LAM/MPI dominates
- Zoom in ...

Benchmark Results



- Fastest Barrier in test

Benchmark Results



CHEMNITZ UNIVERSITY OF TECHNOLOGY

- Speedup with regards to MVAPICH

CHEMNITZ UNIVERSITY OF TECHNOLOGY

Conclusions

InfiniBand™

- InfiniBand™ hardware shows parallelism
- n-way dissemination principle can lower barrier latency

MPI Layer

- Open MPI coll v1 provides a low overhead framework
- n-selection non trivial $\rightarrow$ coll v2

OF TECHNOLOGY

Future Work/Ongoing Efforts

InfiniBand™

- Implementation of InfiniBand™ specialized collectives

MPI Layer

- Open MPI collective framework version 2

CHEMNITZ UNIVERSITY OF TECHNOLOGY