A Fair Benchmark for Evaluating the Latent Potential of Heterogeneous Coupled Clusters

Carsten Clauss
Chair for Operating Systems (LfBS)
RWTH Aachen University, Germany

ISPDC 2007, Hagenberg, Austria

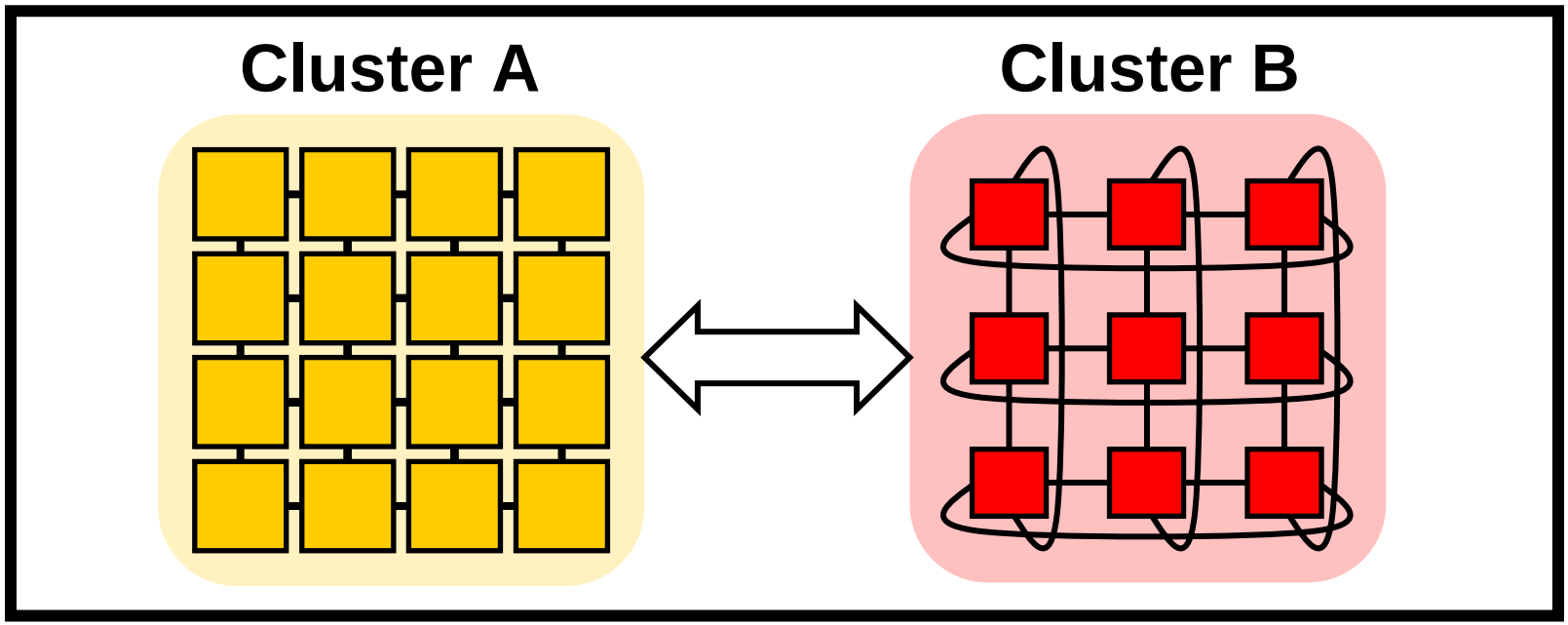
Overview

- Introduction
- Coupled Clusters: Issues and Benefits
- A Fair Benchmark for Coupled Clusters
- Selected Measurement Results
- Conclusions and Outlook

Cluster of Clusters

- **Dedicated Cluster Networks (e.g.):**
  InfiniBand, Myrinet, Quadrix, SCI

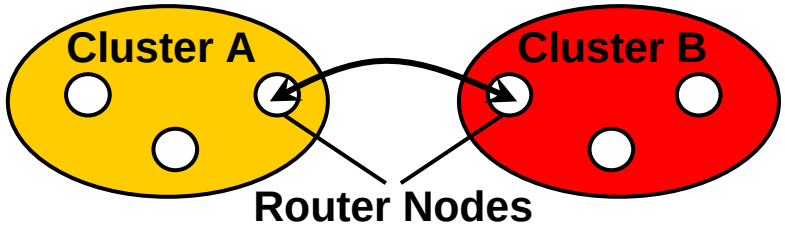
- **Cluster of Clusters / Meta-Computer:**



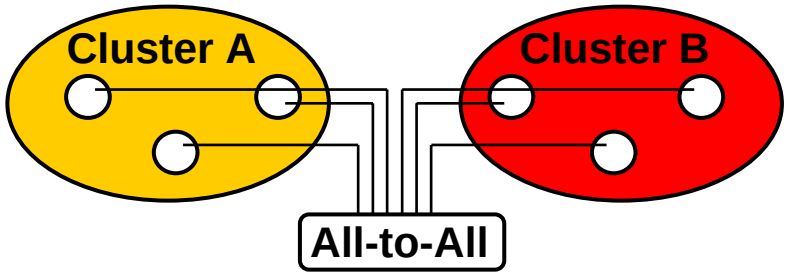
- **Interlinking Network (e.g.):**
  WAN, LAN, …

Cluster of Clusters

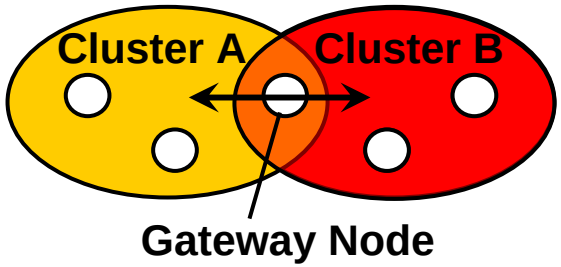
- **Clusters Coupled via Router Nodes:**



- **Clusters with an All-to-All Connectivity:**



- **Coupled Clusters with a Gateway Node:**



Cluster of Clusters

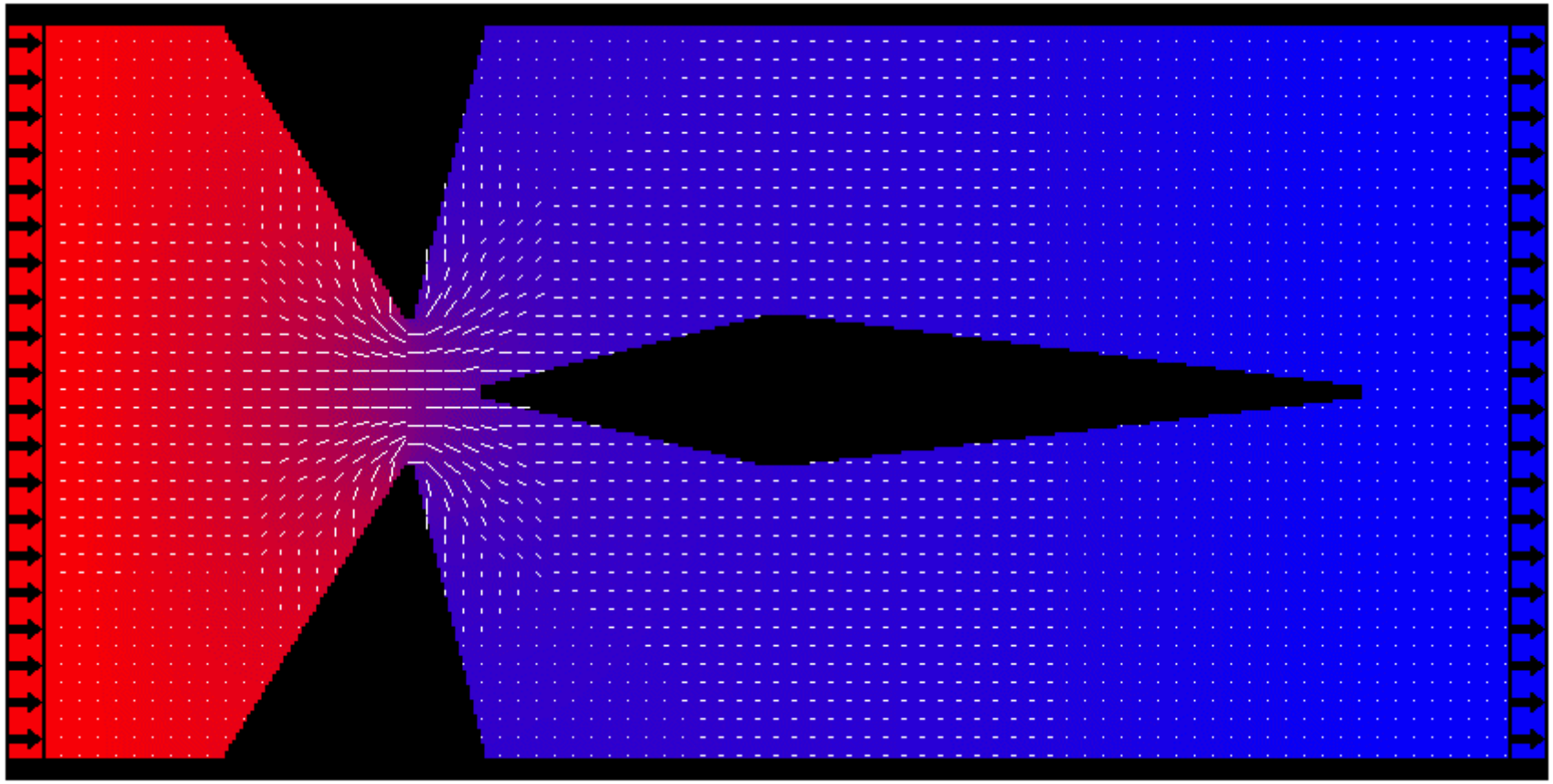
- **Heterogeneous Communication Structures!**
- **→ Need: Topology-aware Message-Passing!**
- **Meta-computing and Grid-enabled MPI-libraries:**
  - **GridMPI**
  - **MPICH-G2**
  - **PACX-MPI**
  - **MetaMPICH**
  - **MPICH/Madeleine**
  - **...**

Cluster of Clusters

- **Inter-Cluster Communication is the Bottleneck!**
- **→ Further Need: Topology-aware Applications!**
- **Adaptation Features Offered by the Libraries:**
  - **New Pre-defined Communicators (e.g. `MPI_COMM_LOCAL` / MetaMPICH)**
  - **Additional Communicator Attributes (e.g. topology *depths* and *colors* / MPICH-G2)**

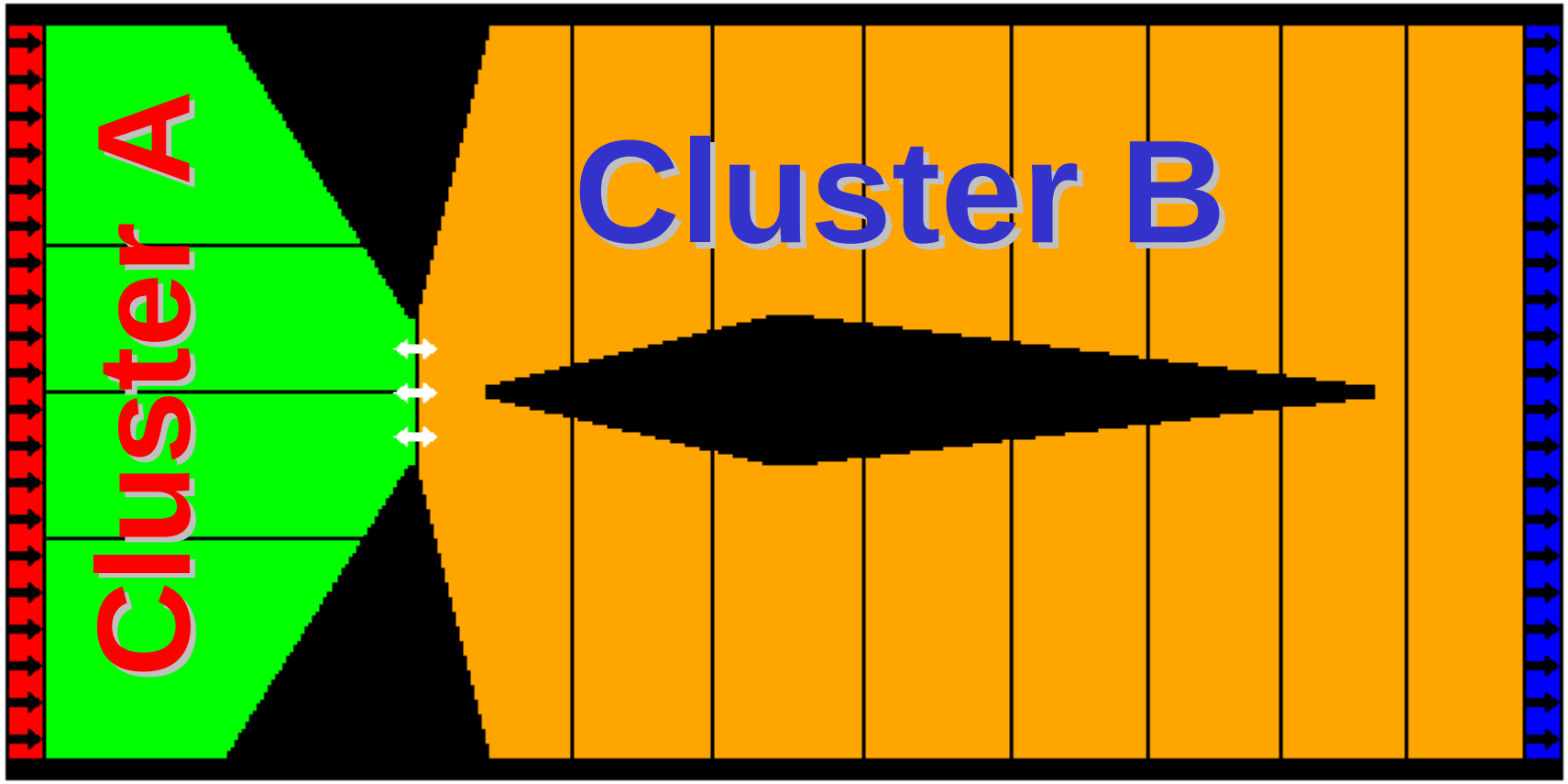
Cluster of Clusters

- **Example of an Adapted Application:**



Cluster of Clusters

- **Example of an Adapted Application:**



Cluster of Clusters

- **Advantages / Profits:**

  - **many more MPI processes by using the remote computational power**

  - **much more *total* cache memory available**

- **Disadvantages / Challenges:**

  - **inter-cluster communication is the bottleneck**

  - **non-adapted applications cannot benefit from the dedicated internal networks**

A Fair Benchmark

- **Why a *Fair* Benchmark?**

  - non-adapted applications do not scale

  - regular benchmarks will report poor performance

- **Discover and Deploy the Latent Potential:**

  → evaluate the system with an adapted benchmark

  → forecast the scalability of adapted applications

A Fair Benchmark

- **Required Benchmark Characteristics:**

  - comprehensible and common core algorithms

  - simple parallelization and communication patterns

  - potentials to adapt the algorithm to heterogeneity

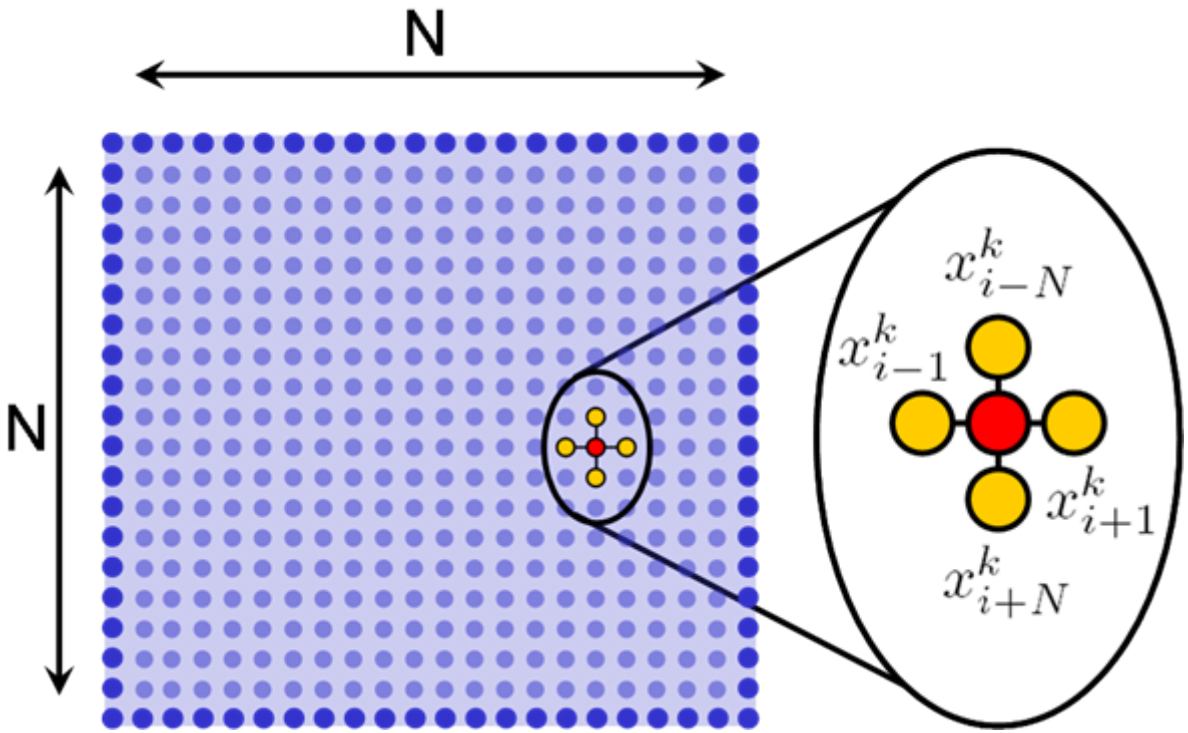
- **Our Choice:**

  → a simple JOR kernel solving the well-known Laplace problem on a rectangular domain

A Fair Benchmark

- **The Simple Benchmark Kernel:**

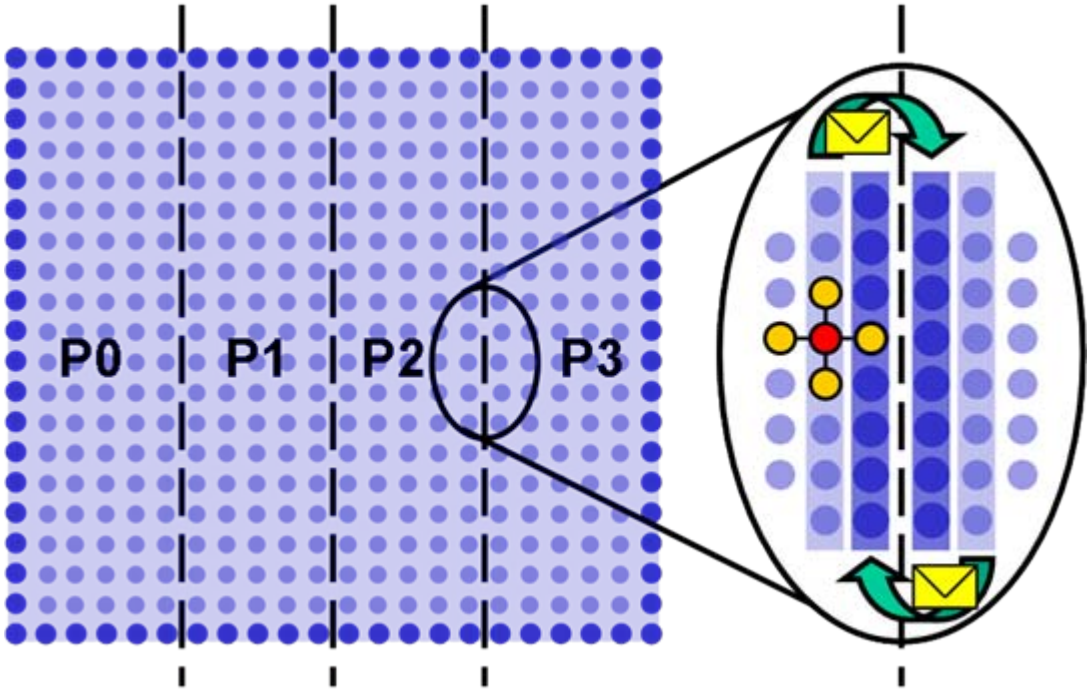
$$x_i^{k+1} = \frac{x_{i-N}^k + x_{i-1}^k + x_{i+1}^k + x_{i+N}^k}{4}$$



A Fair Benchmark

- **… and its Parallelization:**

$$x_i^{k+1} = \frac{x_{i-N}^k + x_{i-1}^k + x_{i+1}^k + x_{i+N}^k}{4}$$

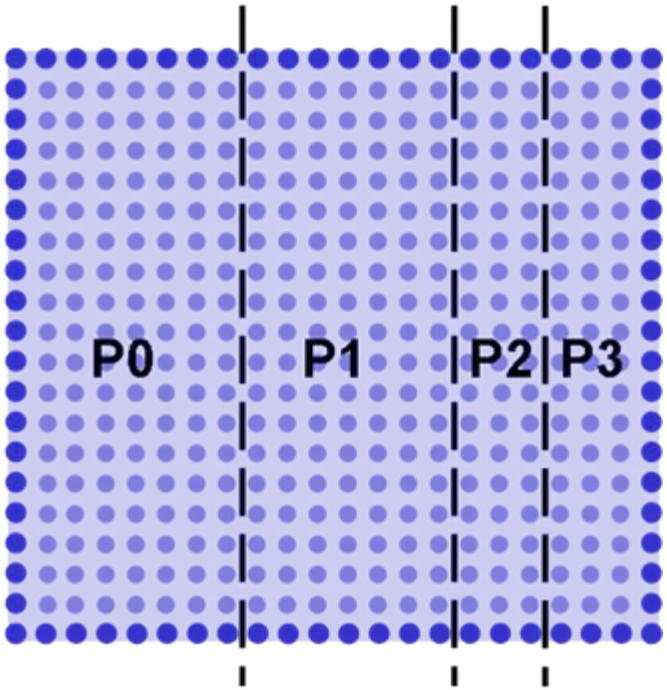


A Fair Benchmark

- **Adaptation to Heterogeneity:**

(1) Heterogeneous Computational Power:

→ apply load balancing via an adapted domain decomposition!



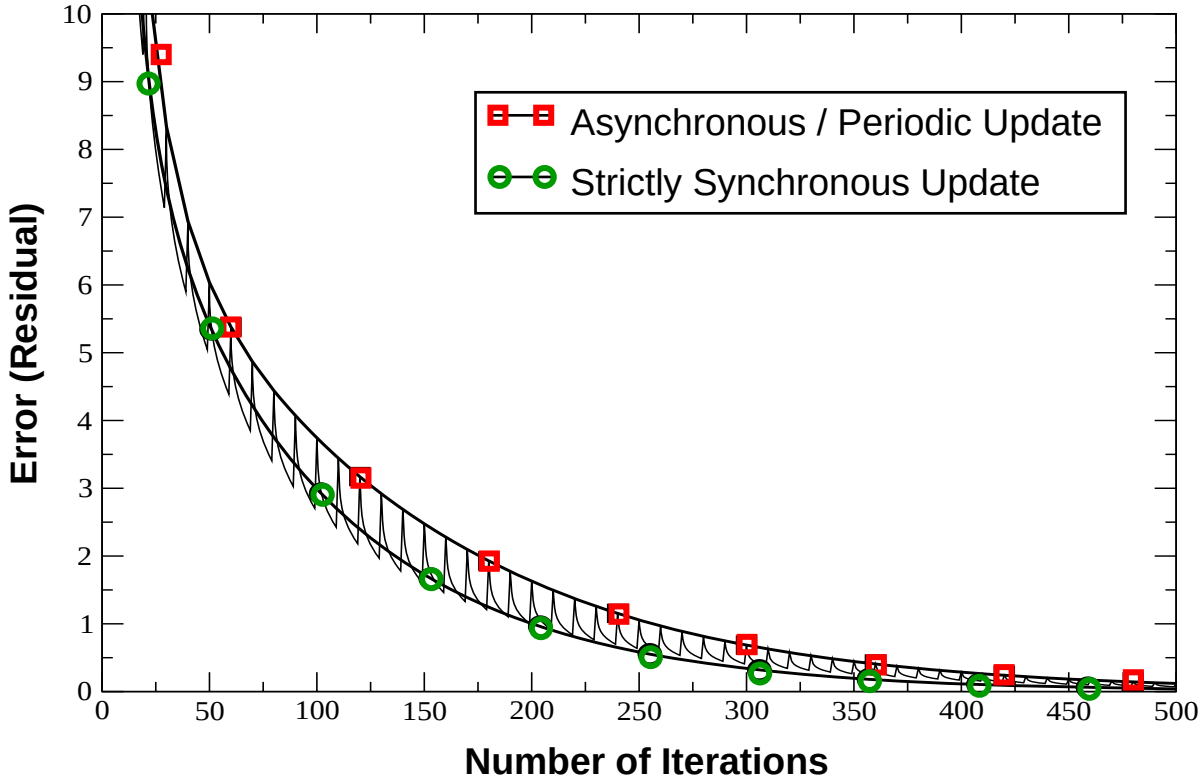
f (P0,P1) > f (P2,P3)

A Fair Benchmark

- **Adaptation to Heterogeneity:**

(2) Bottlenecks in Communication

→ compensate weak communication resources by asynchronous / periodic relaxation!

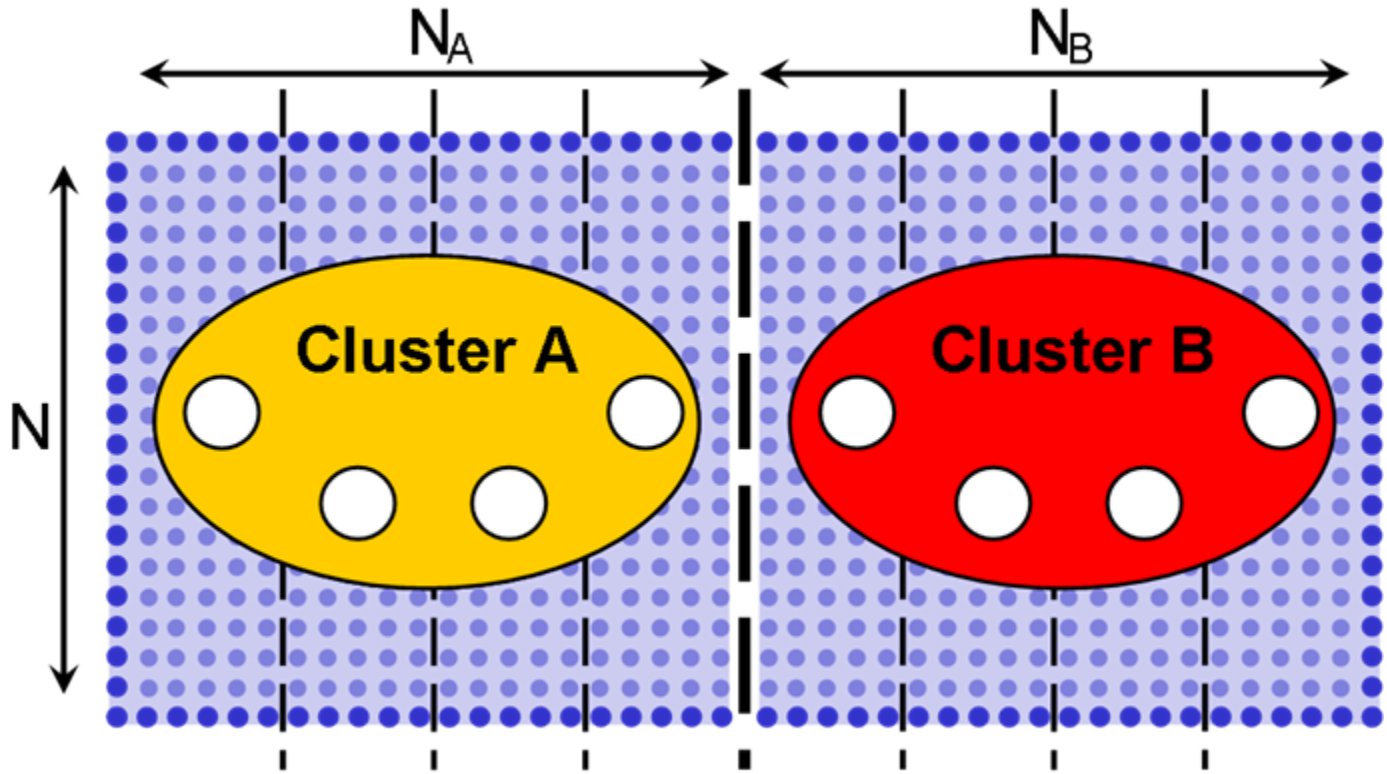


A Fair Benchmark

- **Benchmark Description:**

(1) Domain Decomposition

→ **Load Balance:** $N_A + N_B = 2\ N$

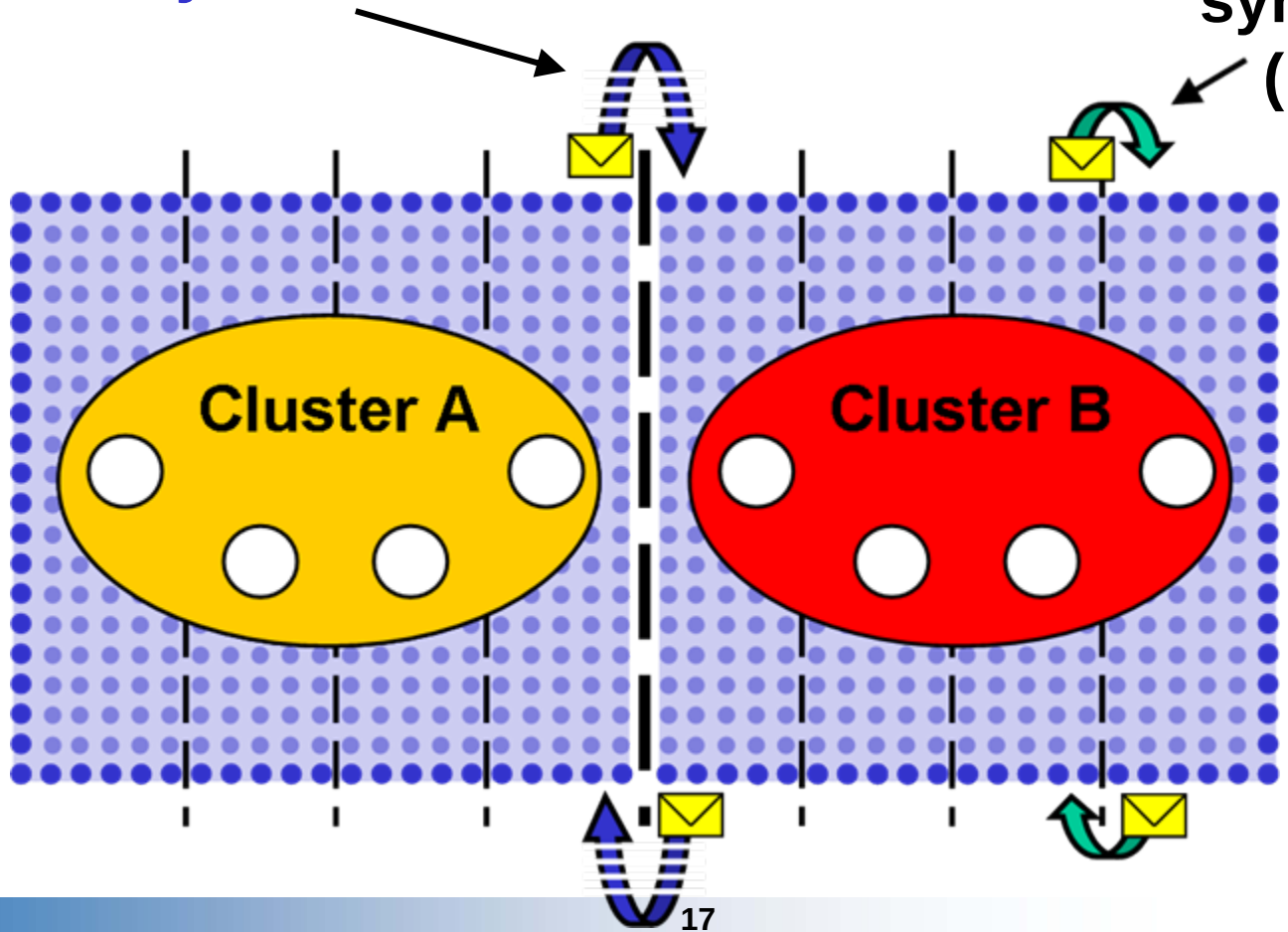


A Fair Benchmark

- **Benchmark Description:**

(2) The Inter-Cluster Bottleneck

→ Asynchronous Relaxation

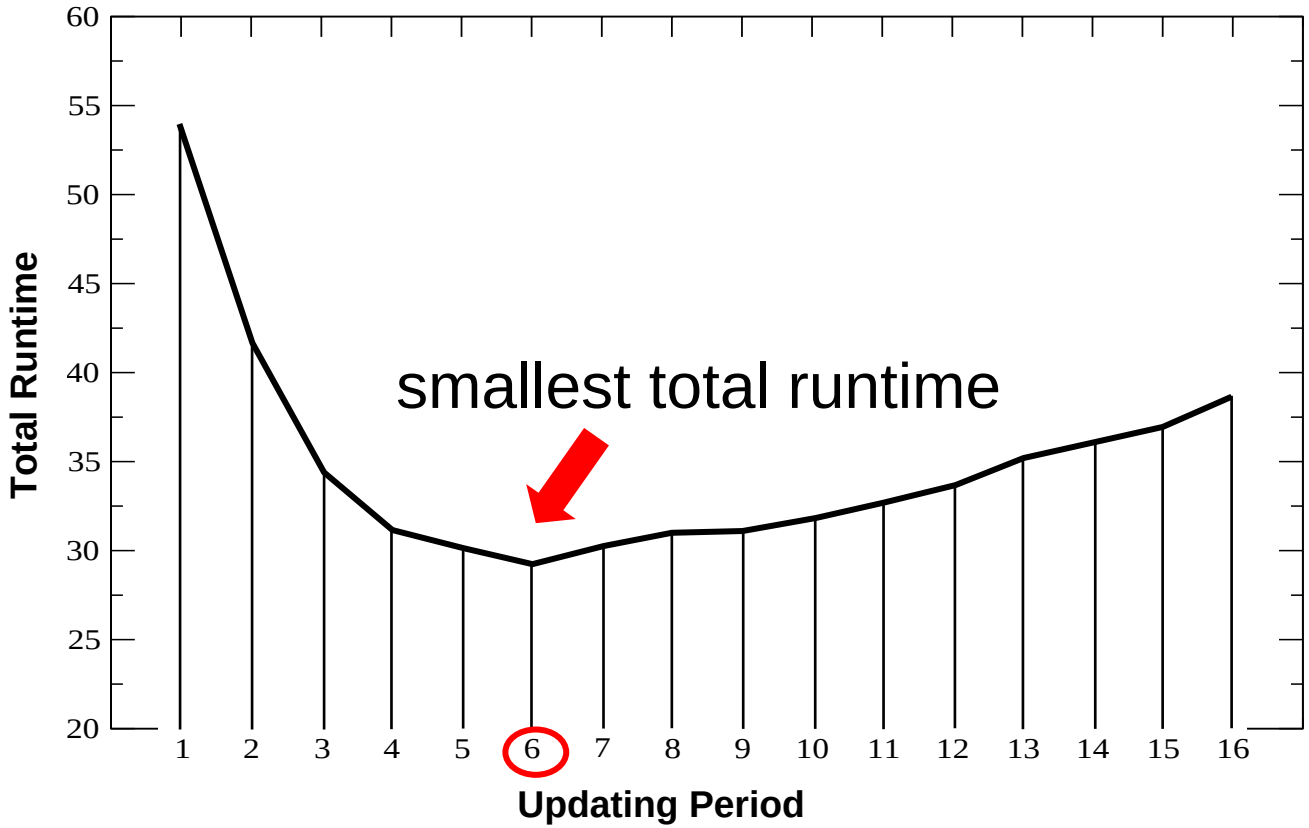


A Fair Benchmark

- **Benchmark Description:**

(2) The Inter-Cluster Bottleneck

→ Find the best periodic updating scheme:



A Fair Benchmark

- **Benchmark Procedure:**

1. **local run (no inter-cluster communication)**
   → **ratio of computational power**

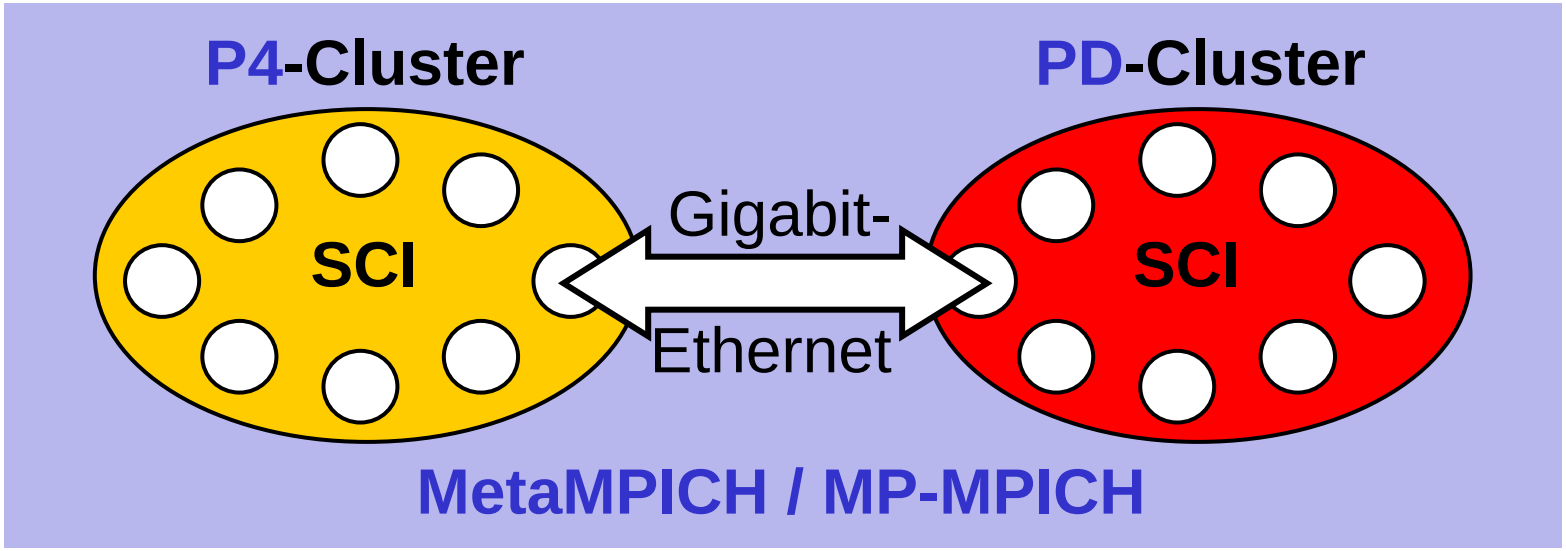
2. **transparent run with no optimization applied**
   → ***Artless* Speedup**

3. **determination of the best updating period**

4. **optimized run with load-balance and the best periodic updating scheme**
   → ***Artful* Speedup**

Selected Results

- **The Scenario:**



- **P4-Cluster:**
  - 8 Xeon Nodes, 16 CPUs
  - 2.4 GHz, 512 kByte Cache
  - 8 GByte Total Memory
  - SCI-Network, 2*4 2D-Torus
  - Gigabit-Ethernet

- **PD-Cluster:**
  - 16 PentiumD DualCore CPUs
  - 2.8 GHz, 1 MByte Cache
  - 8 GByte Total Memory
  - SCI-Network, 3*4 2D-Torus
  - Gigabit-Ethernet

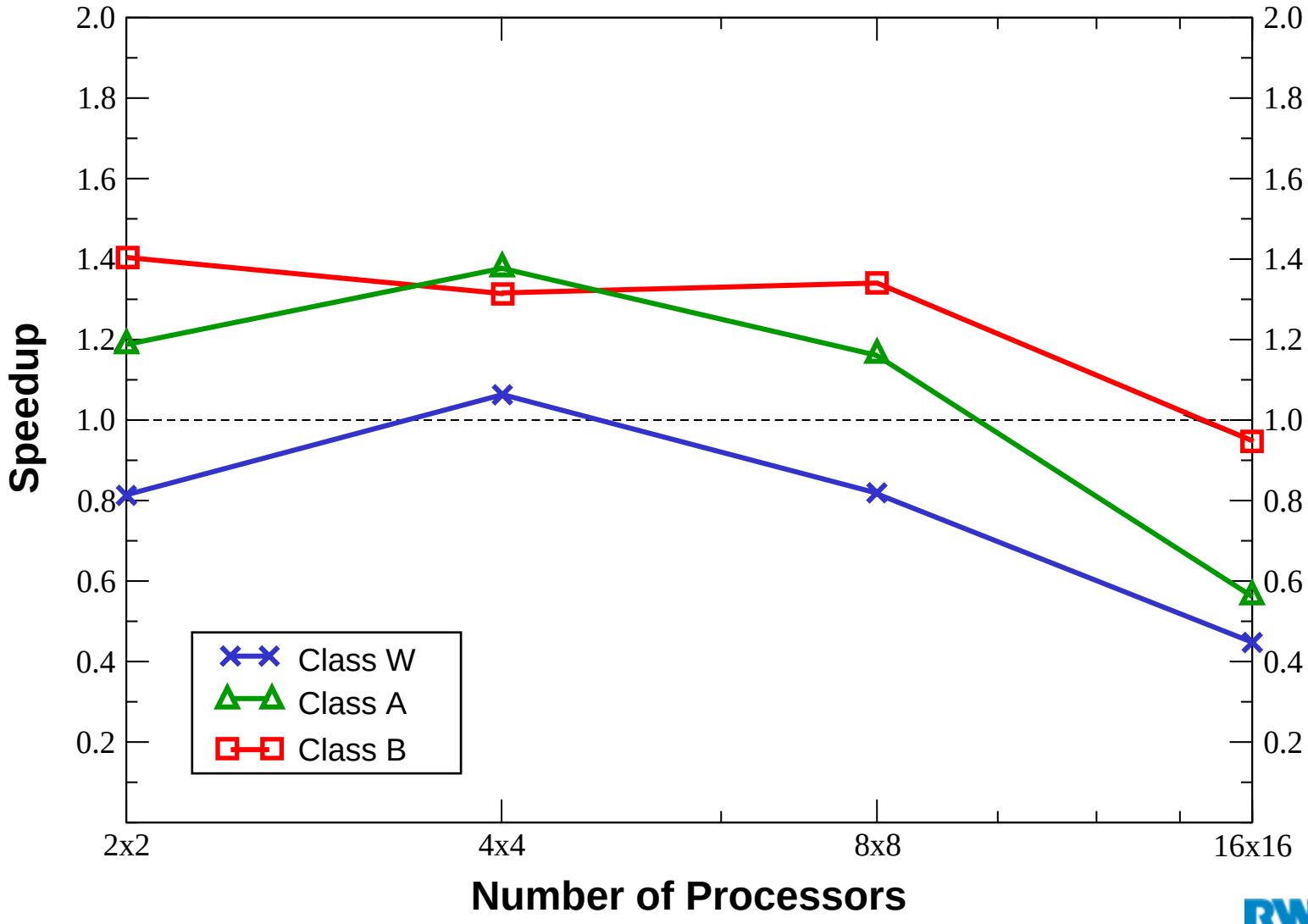
Selected Results

- **The Problem Classes:**

Class	Size of N	Purpose
S	16	Very Small Systems / Test-Runs
W	128	Small Clusters of Workstations
A	256	Midsize Productivity Clusters
B	512	High Performance Clusters
C	1024	Coupled Top-500 Systems

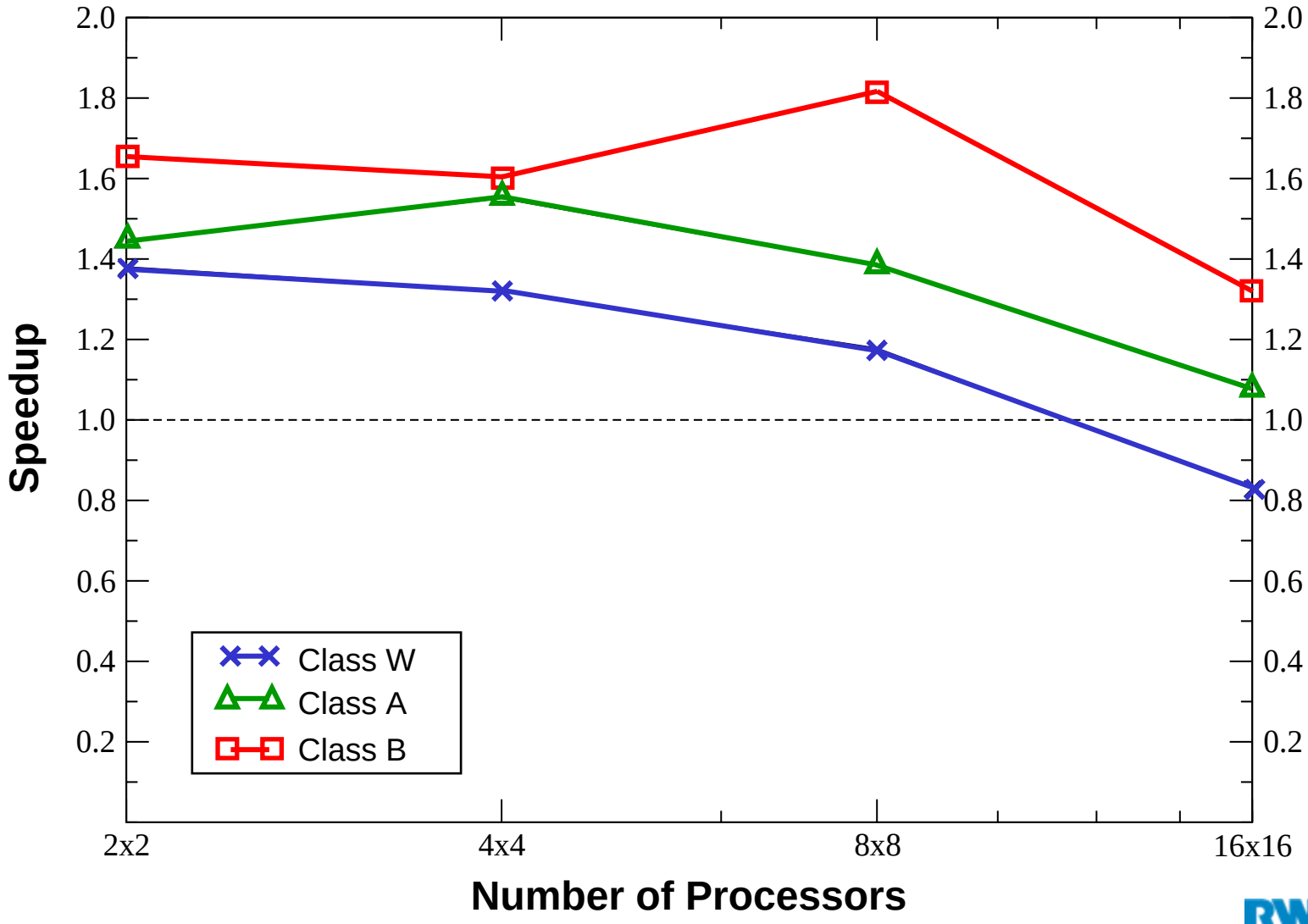
Selected Results

- Measured *Artless* Speedups:



Selected Results

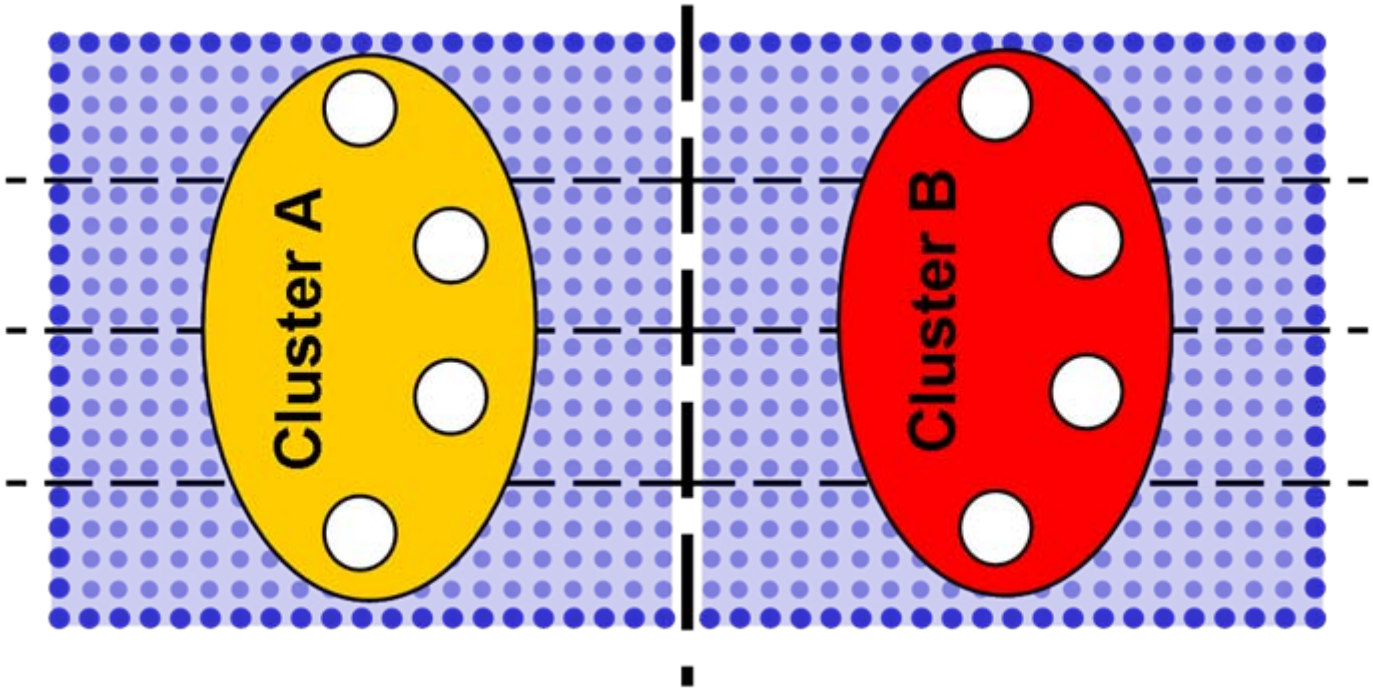
- **Measured *Artful* Speedups:**



Conclusions and Outlook

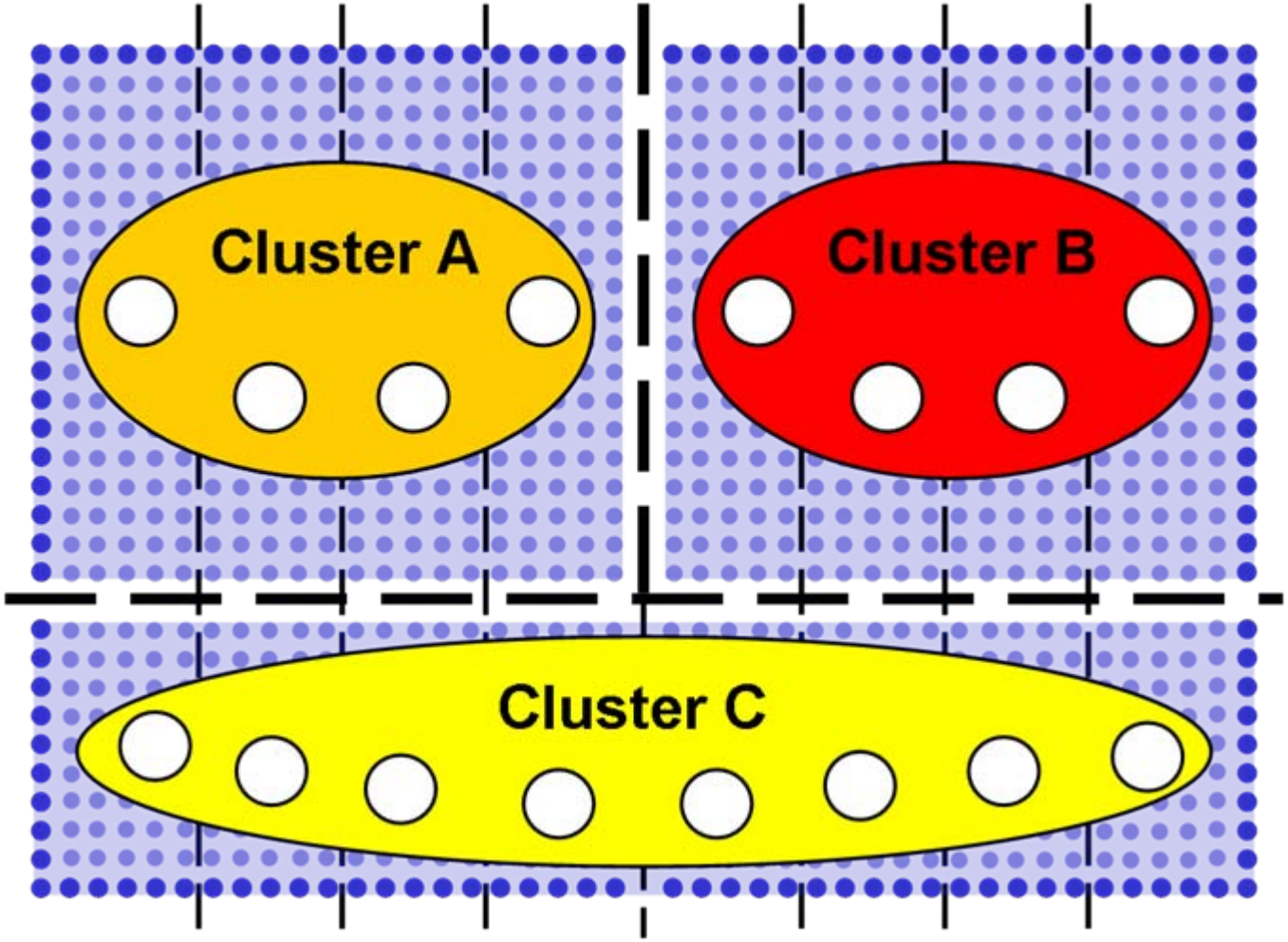
- **Alternative Domain Decomposition:**

  → **especially in case of an All-to-All connectivity?**



Conclusions and Outlook

- **More Than Two Sites?**



Thank you for your attention!

Carsten Clauss
Chair for Operating Systems (LfBS)
RWTH Aachen University, Germany

www.MP-MPICH.de