Optimising MPI Applications for Heterogeneous Coupled Clusters with MetaMPICH

Carsten Clauss, Martin Pöppe, Thomas Bemmerl

carsten@lfbs.rwth-aachen.de
http://www.mp-mpich.de

RWTH Scalable Computing Aachen

Lehrstuhl für Betriebssysteme
RWTH Aachen

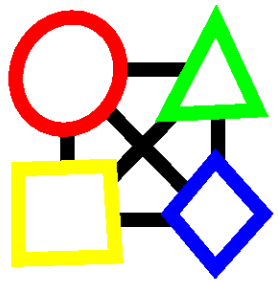
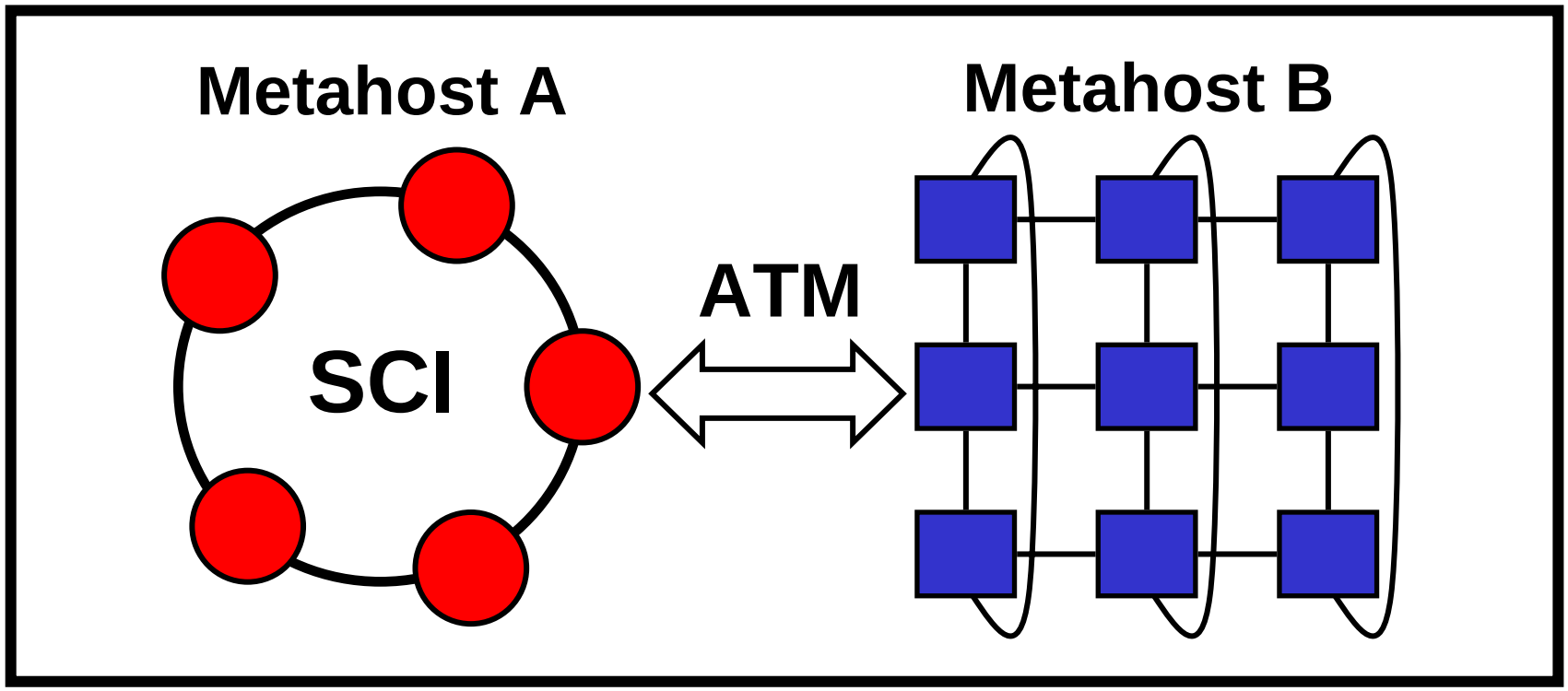
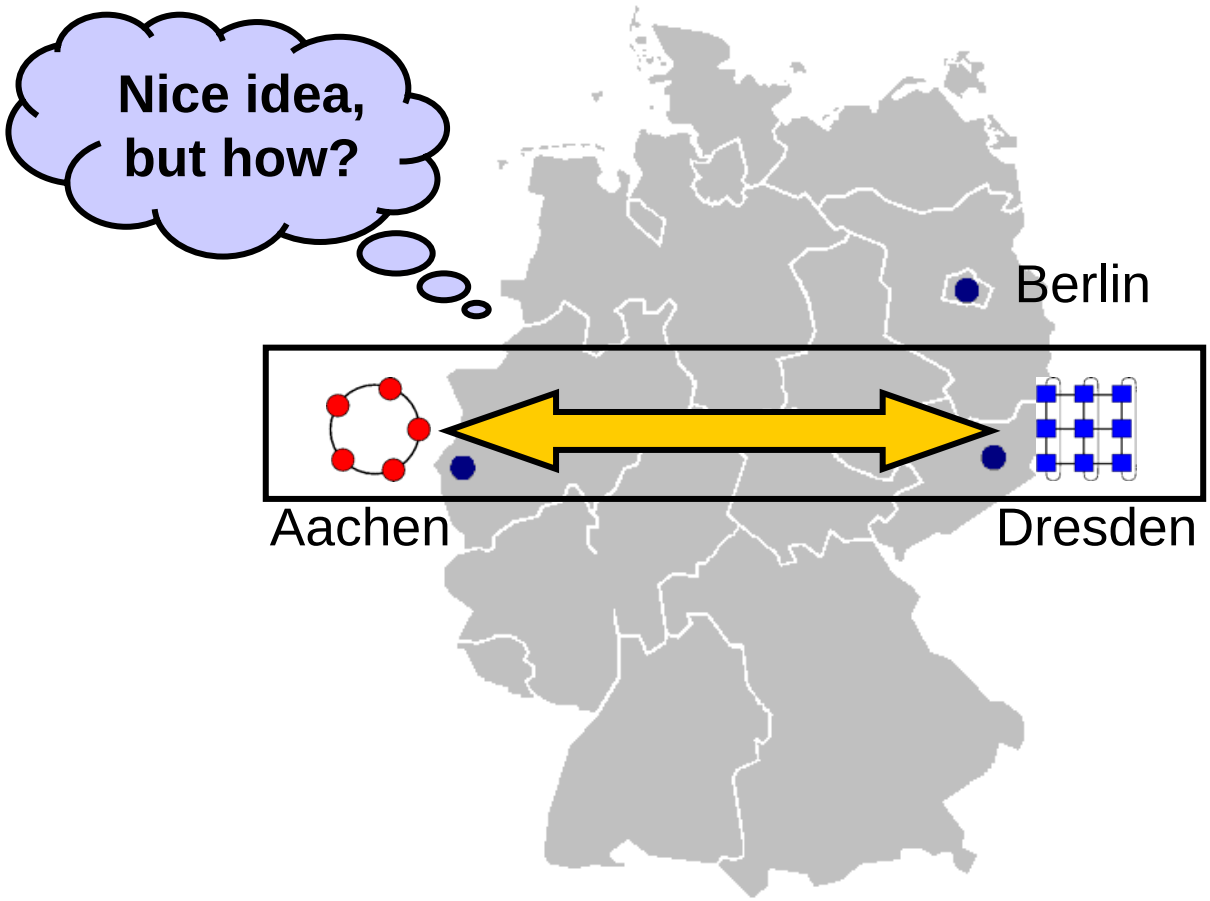


Table of Contents

Coupled Clusters

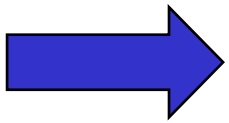


Example of a Possible Use



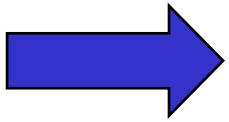
How to *build* a Metacomputer?

MetaMPICH



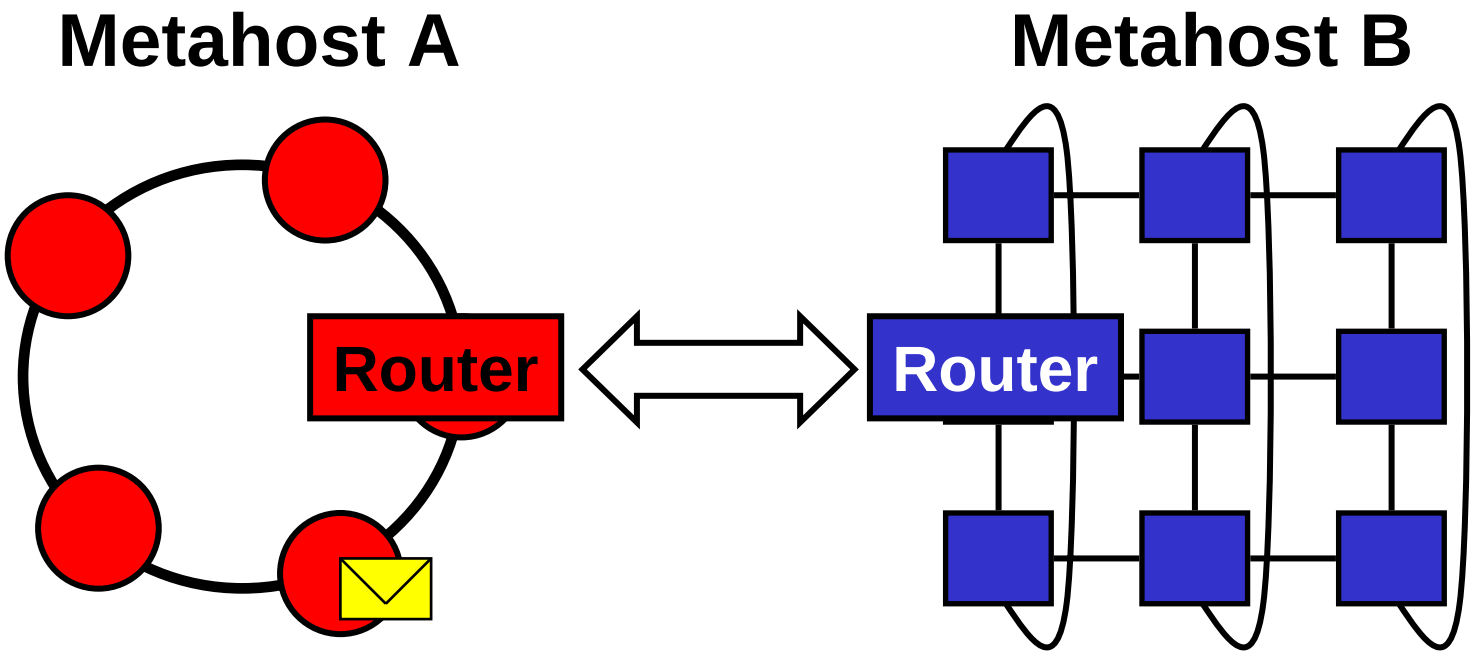
Transparent **MPI system on metacomputers**

MP-MPICH

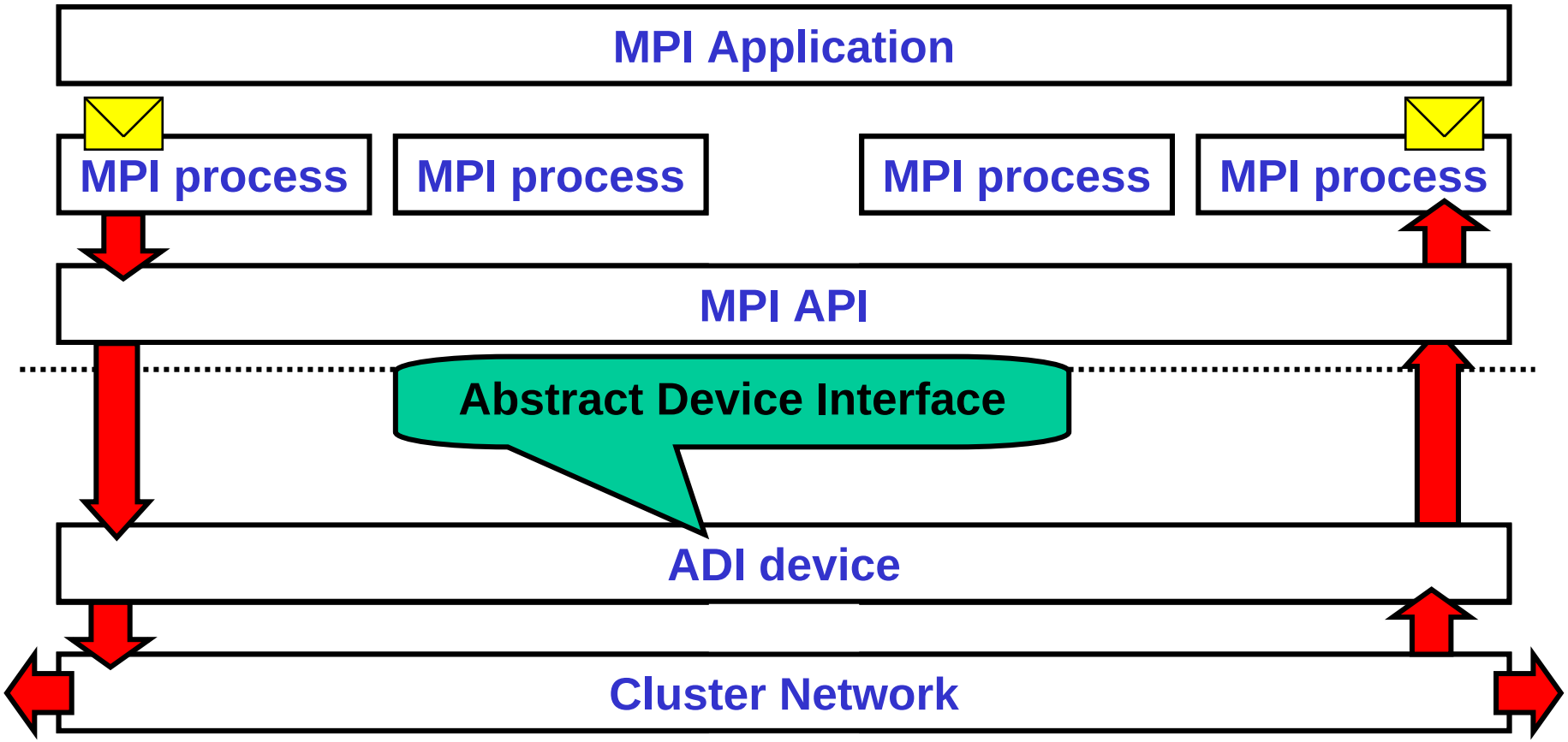


- **Windows 2000/XP, Linux, Solaris**
- **SCI, TCP, SHMEM** **internal**
- **TCP and AAL5** **external**

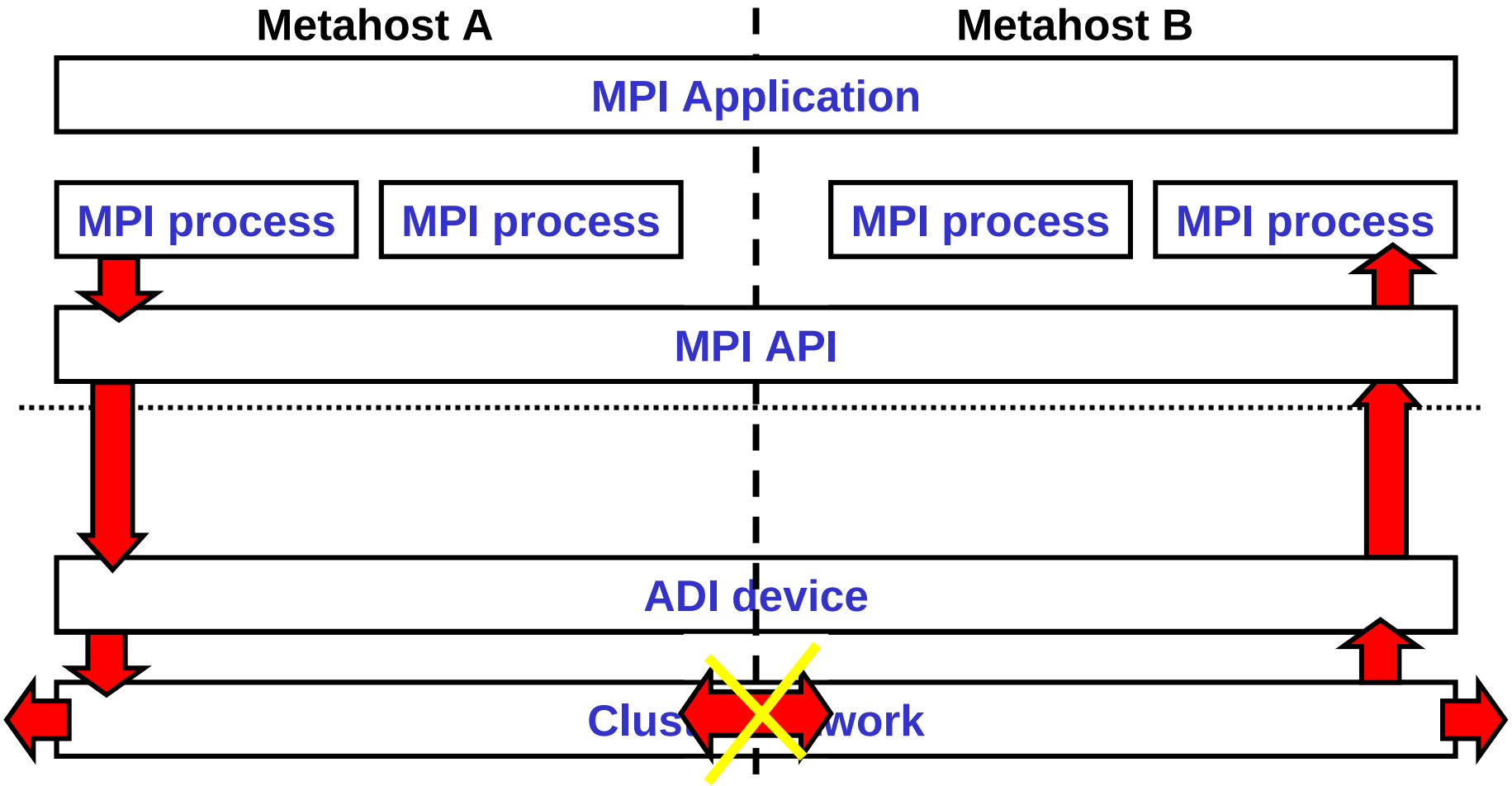
Architecture of MetaMPICH



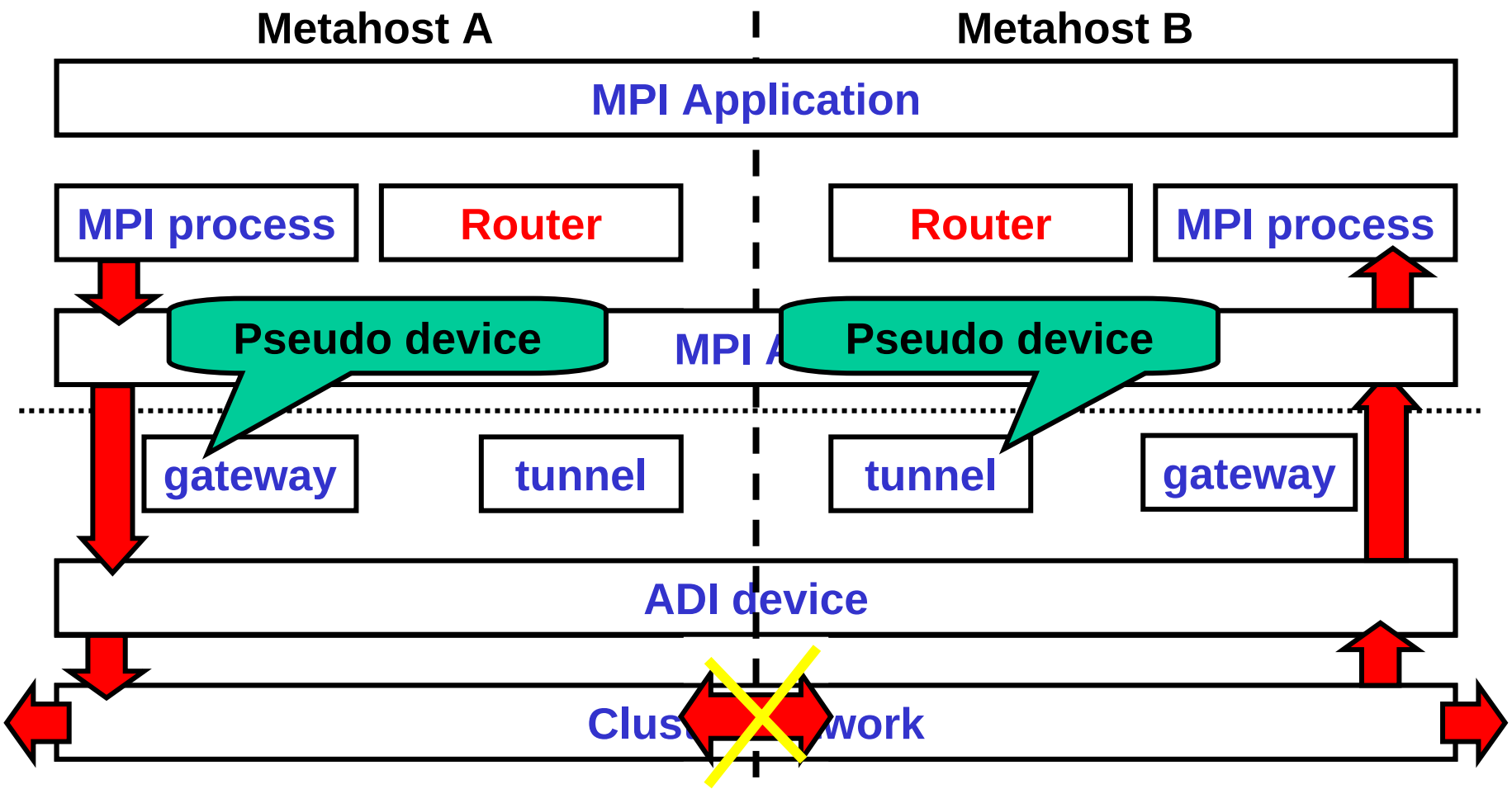
Architecture of MetaMPICH



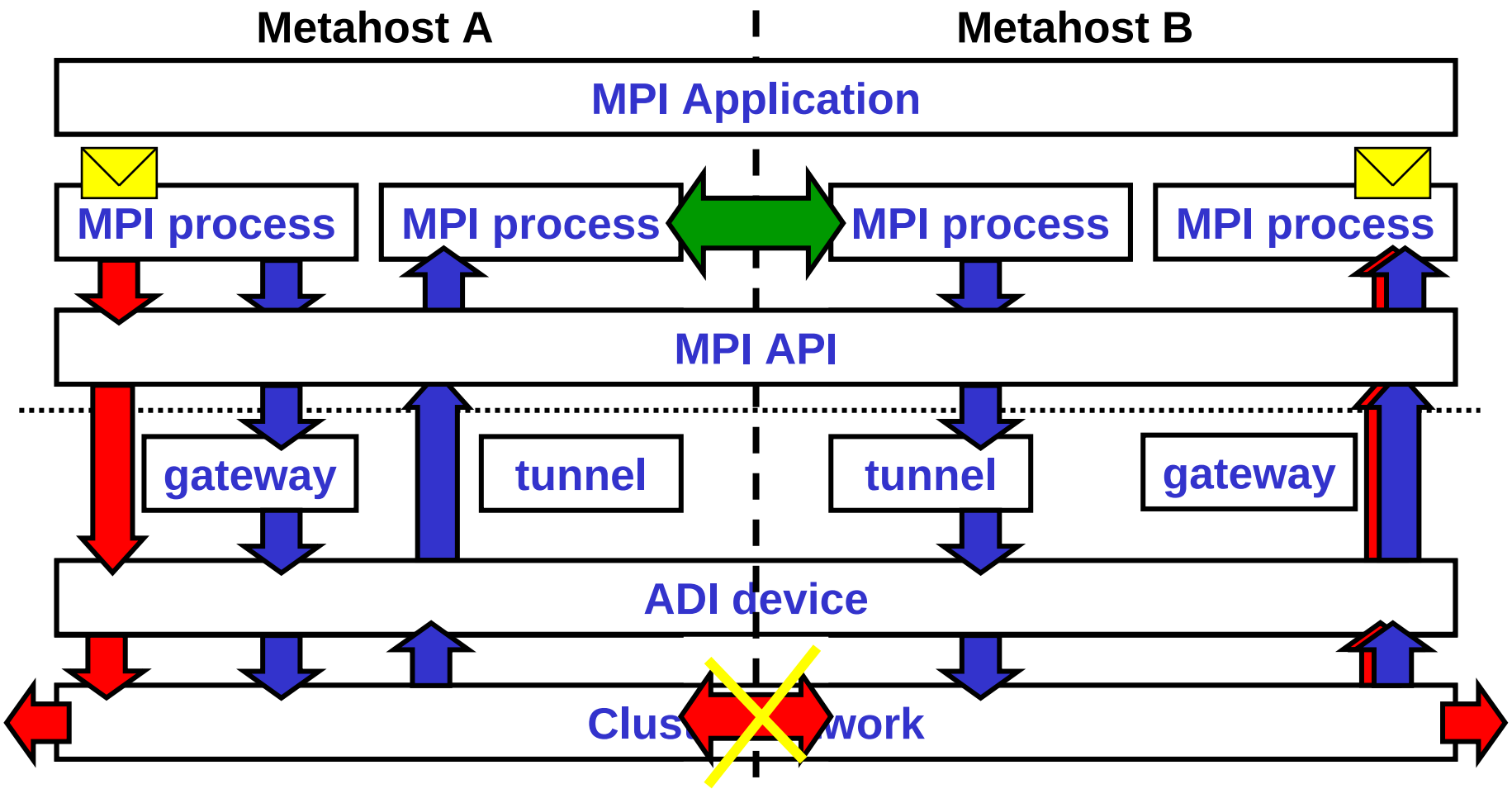
Architecture of MetaMPICH



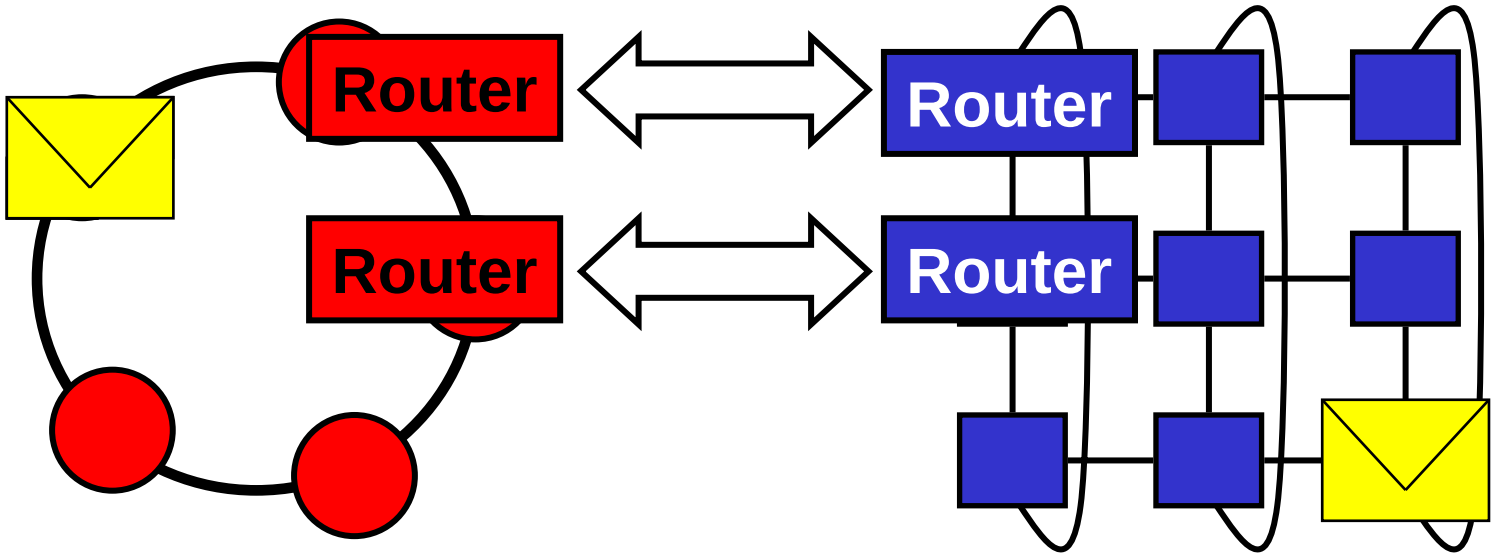
Architecture of MetaMPICH



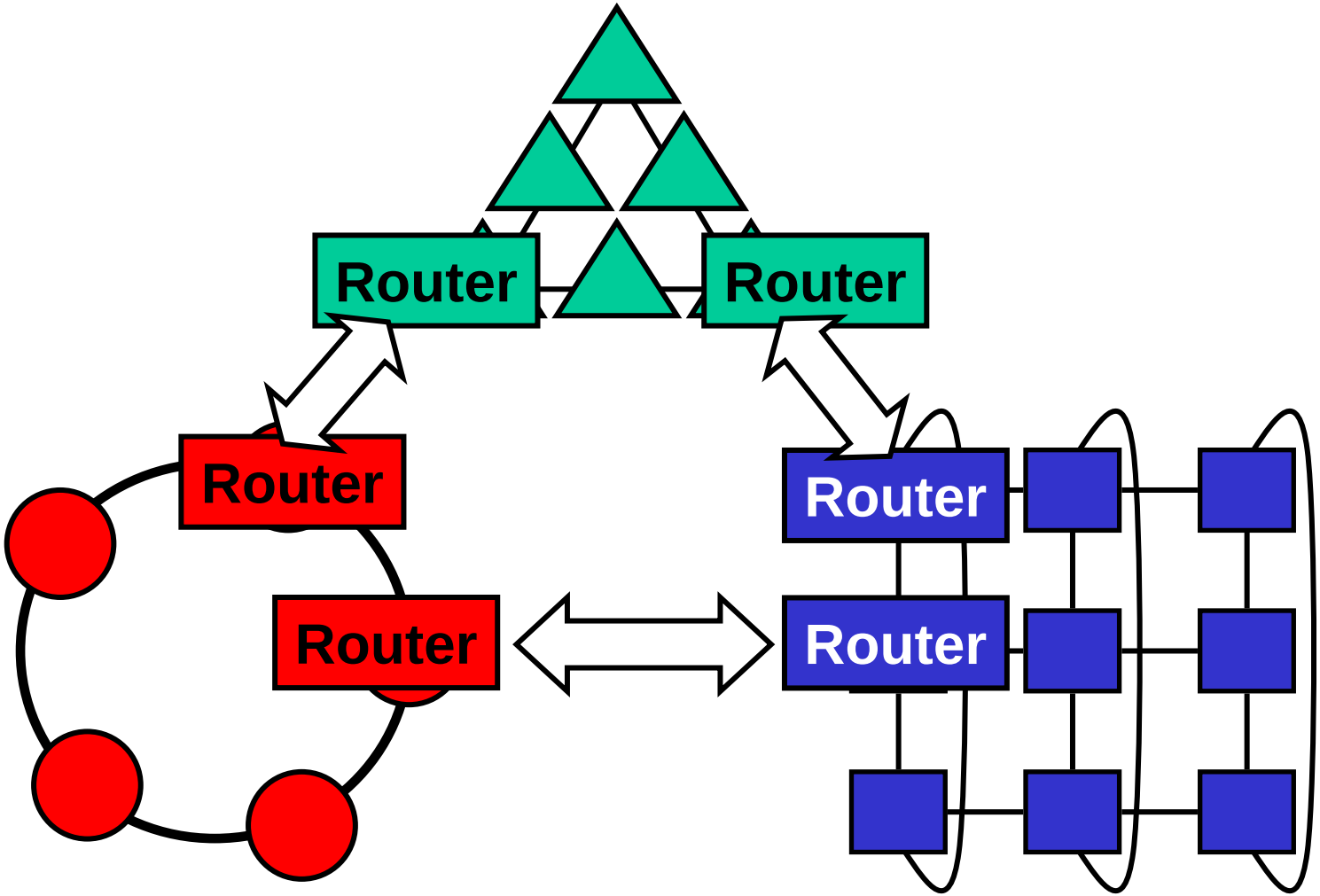
Architecture of MetaMPICH



Architecture of MetaMPICH

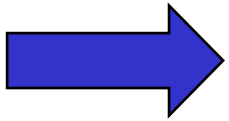


Architecture of MetaMPICH



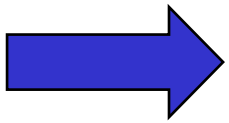
Metacomputing

Advantage



Many **more MPI processes** by using the remote computational power

Disadvantage

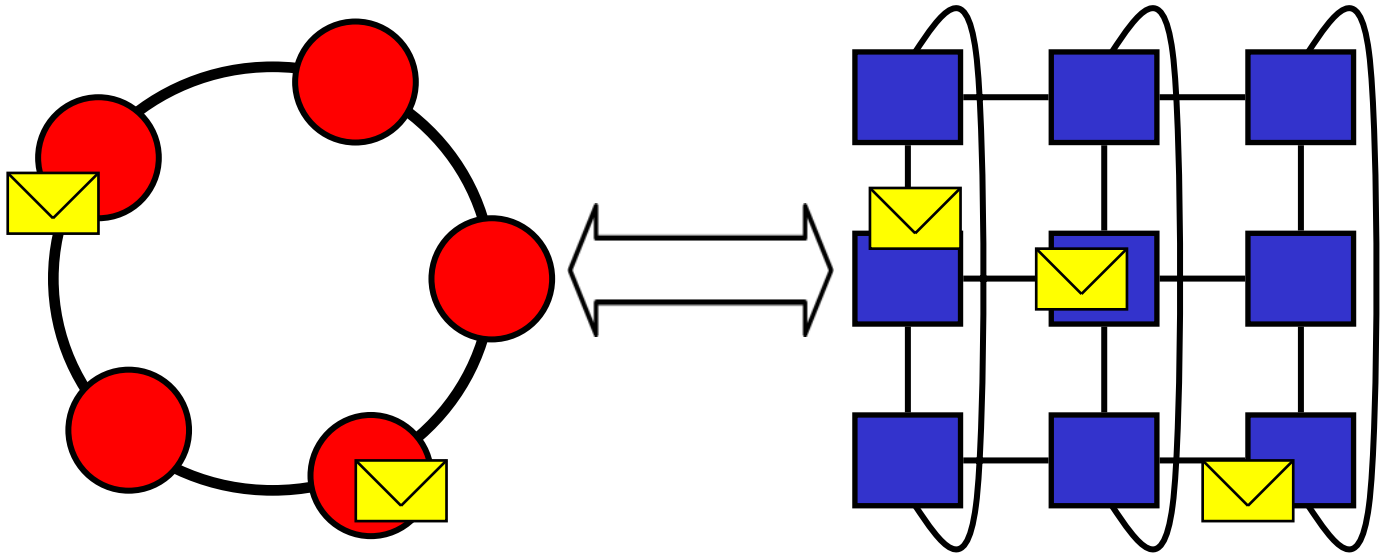


Inter-cluster communication is the system's **bottleneck**

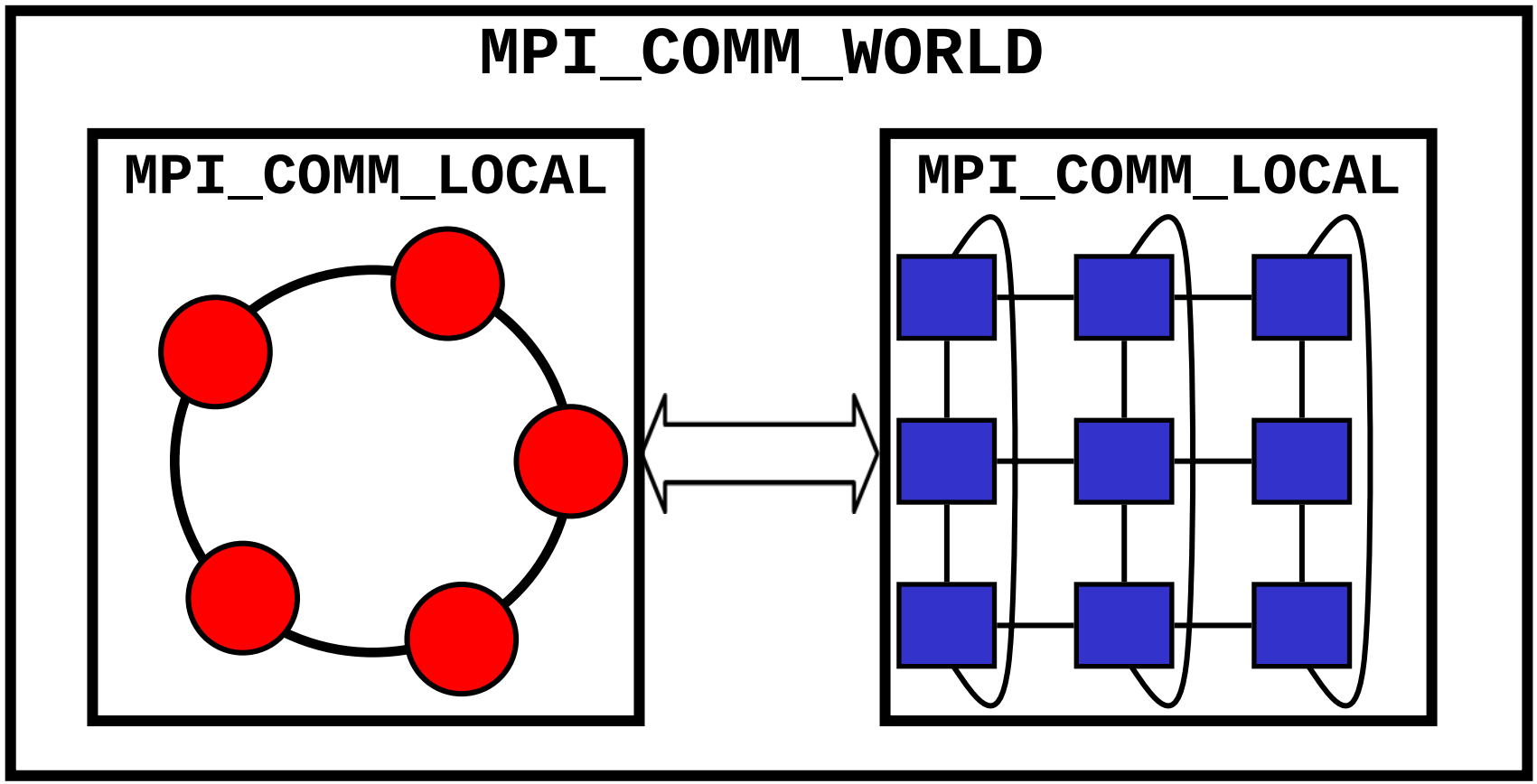
Existing applications **cannot benefit** from the ***fast*** **internal** networks ☹

Process Grouping

Division of the algorithmic problem:

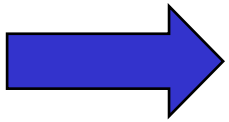


Process Grouping

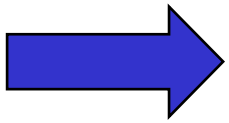


Metacomputing

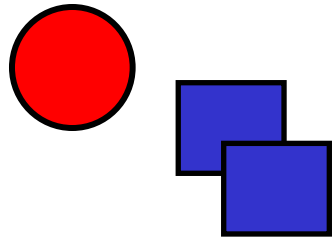
Pay Attention to Load Balance !!!

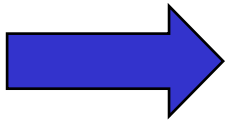


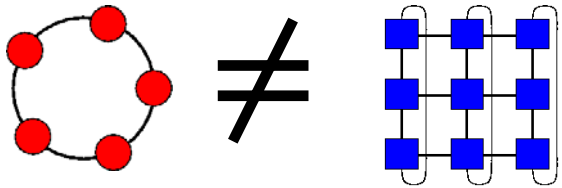
Clusters do NOT need to be identical !!!



- Different CPU Power
- Different Number of Nodes
- Different Internal Networks

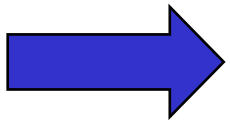




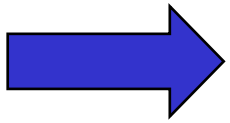


Metacomputing

Challenges:



Inter-Metahost Communication Bottleneck

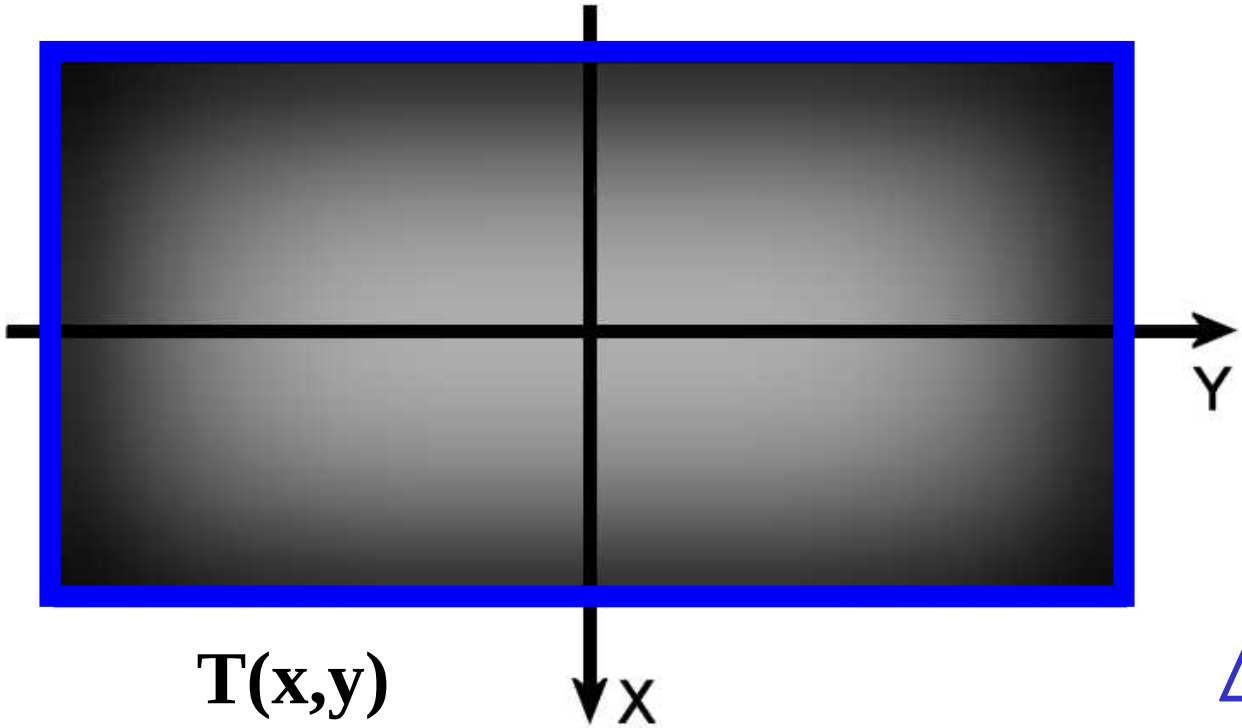


Heterogeneous System: Load Balance !!!

Table of Contents

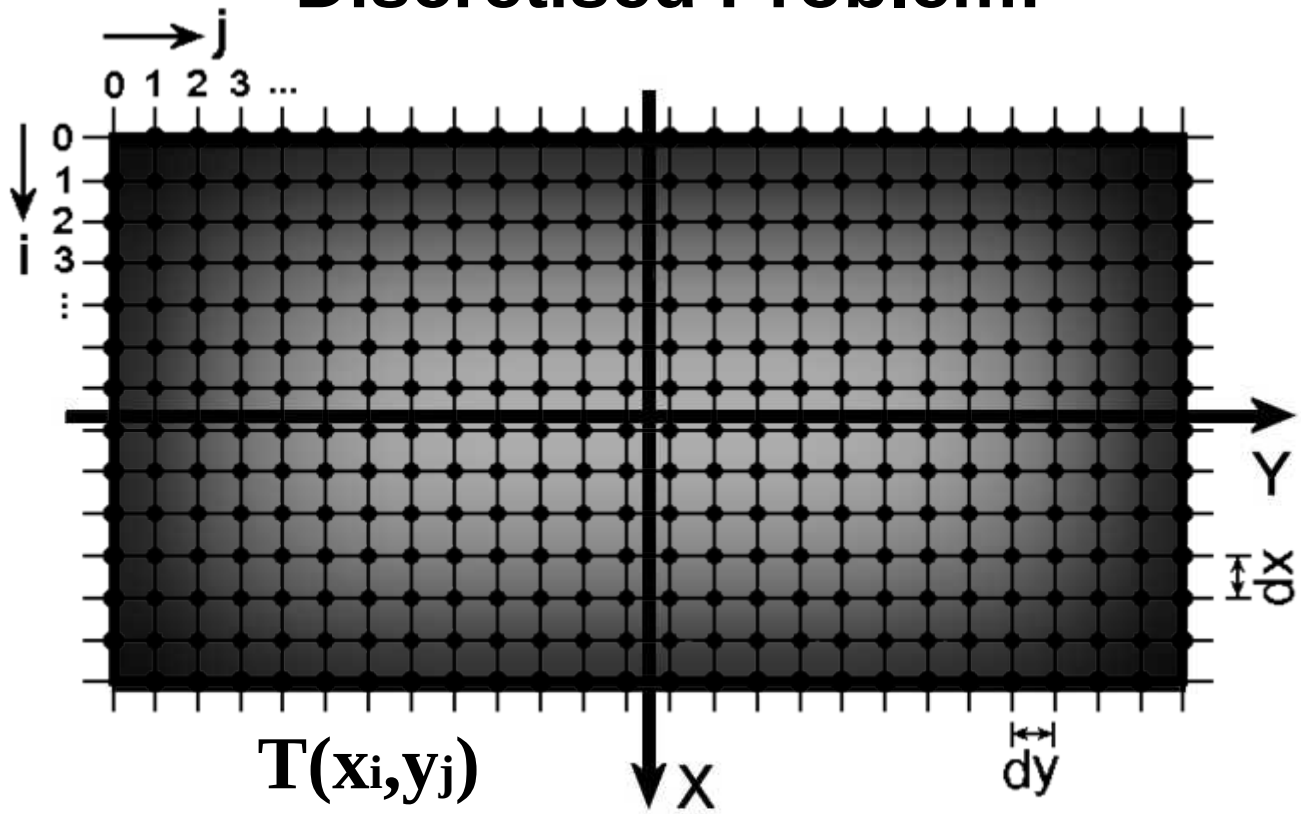
Boundary Value Problems

Distribution of Temperature in a Plate:



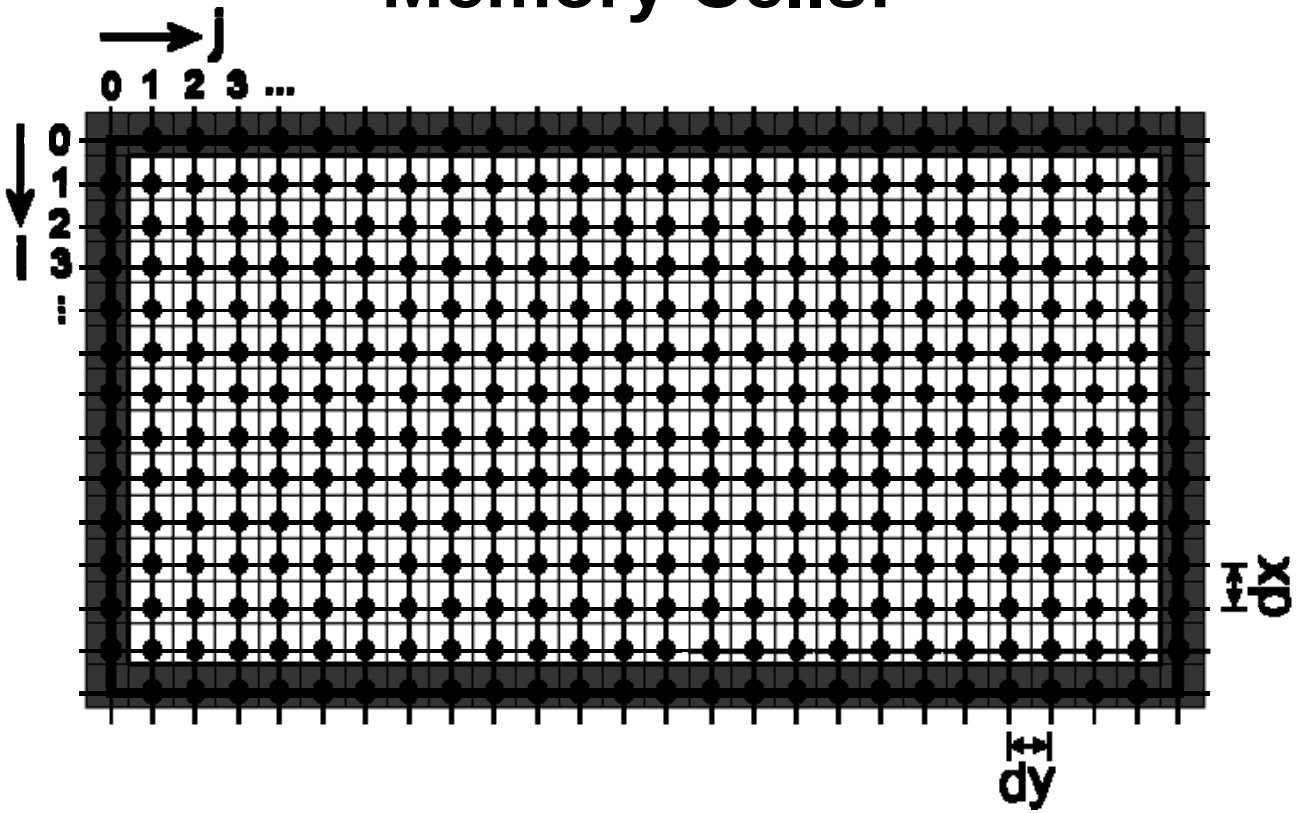
Boundary Value Problems

Discretised Problem:



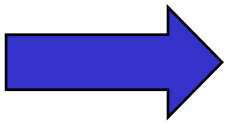
Boundary Value Problems

Memory Cells:



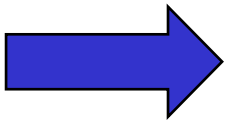
Boundary Value Problems

Discretising the Problem:



big and sparsely populated linear equation systems

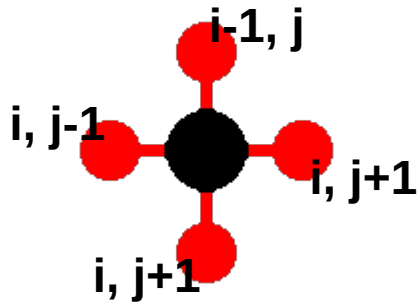
Iterative Solver:



simple example: Jacobi- or Gauss-Seidel method

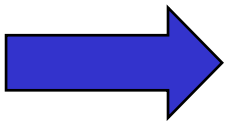
Boundary Value Problems

Iteration Rule:



$$T_{[i][j]} = 0.25 \left(T_{[i-1][j]} + T_{[i+1][j]} + T_{[i][j-1]} + T_{[i][j+1]} \right)$$

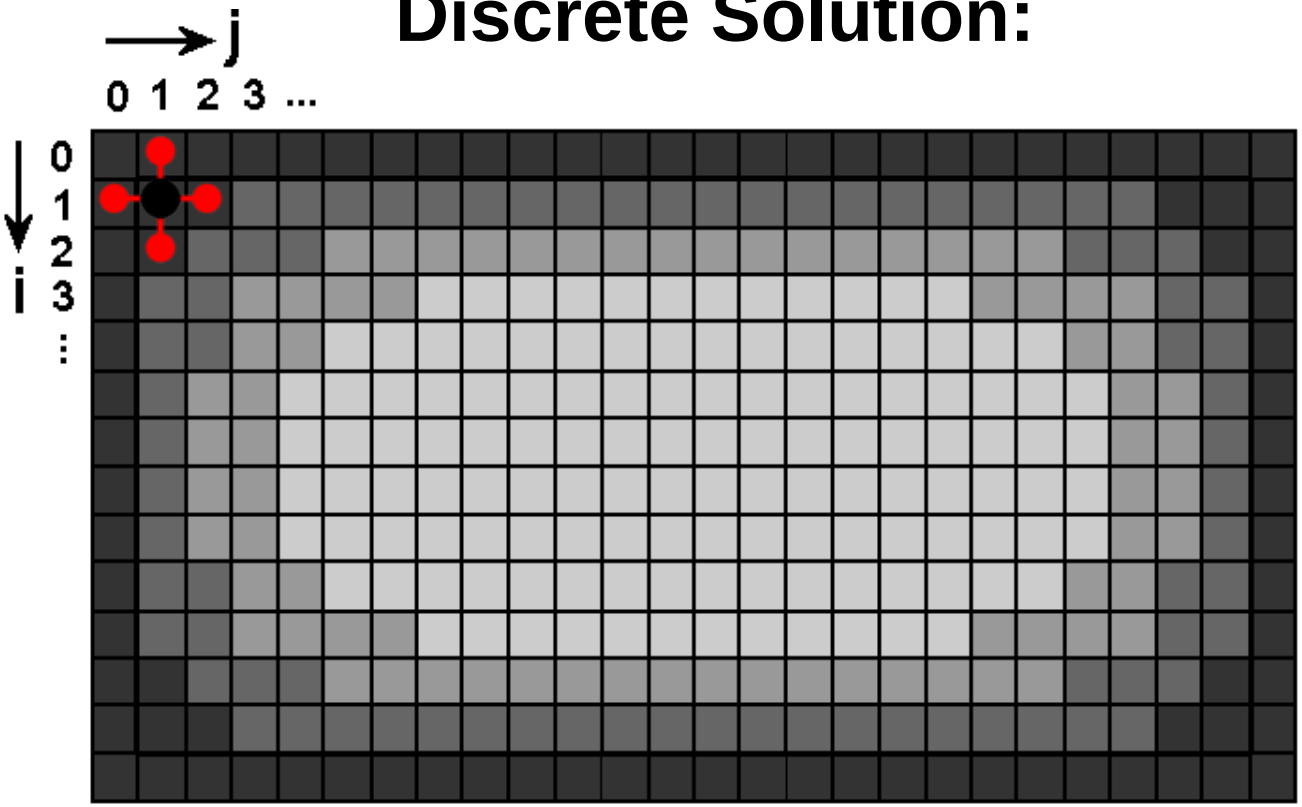
Large number of iterations over the grid



Approximation for the solution

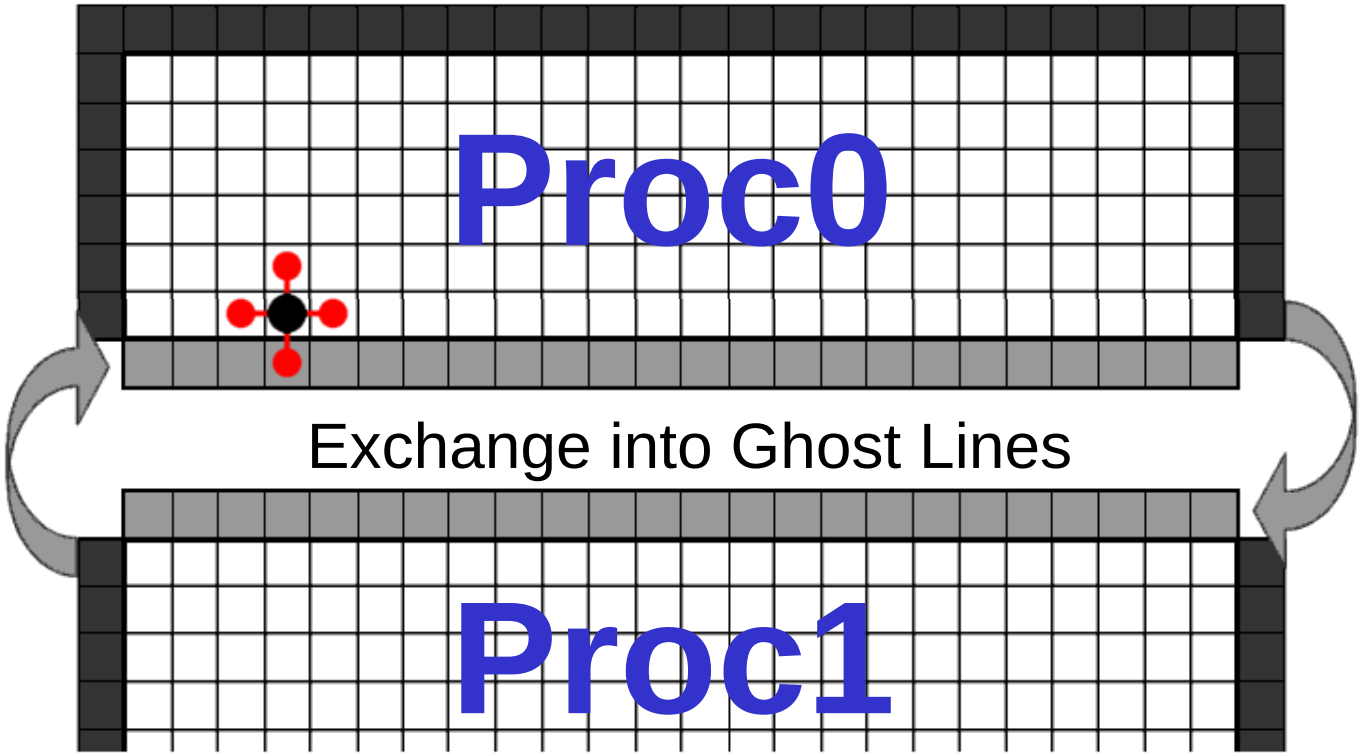
Boundary Value Problems

Discrete Solution:



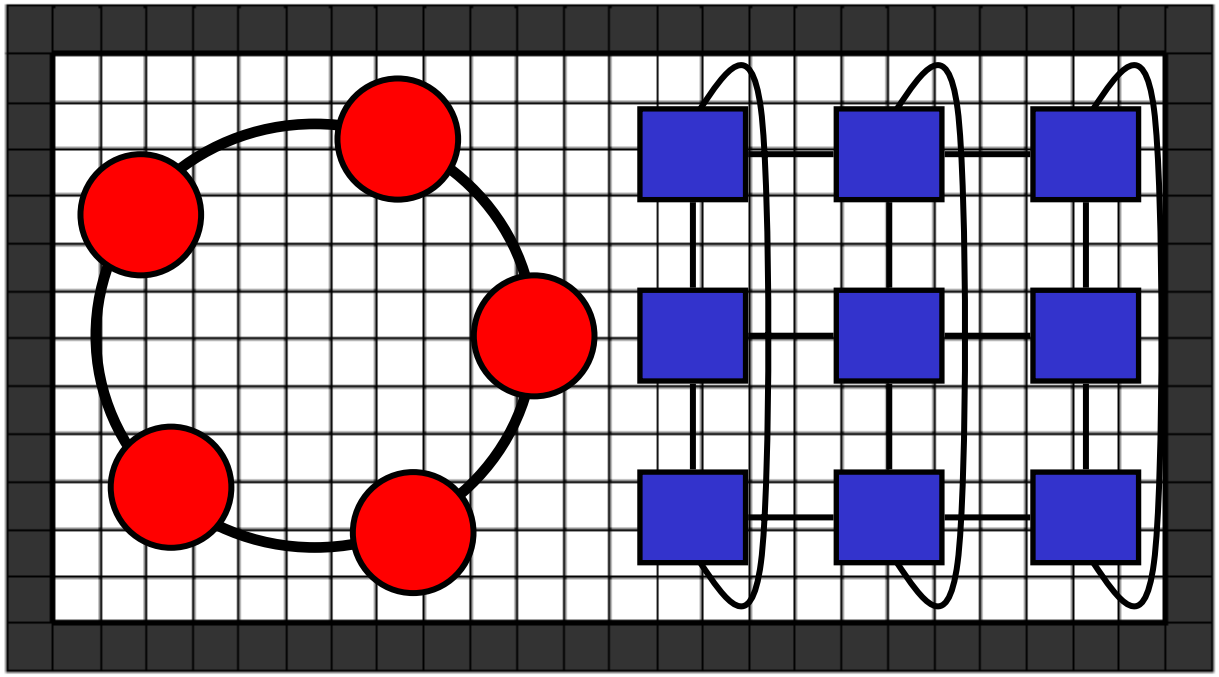
Parallelisation

Domain Decomposition:



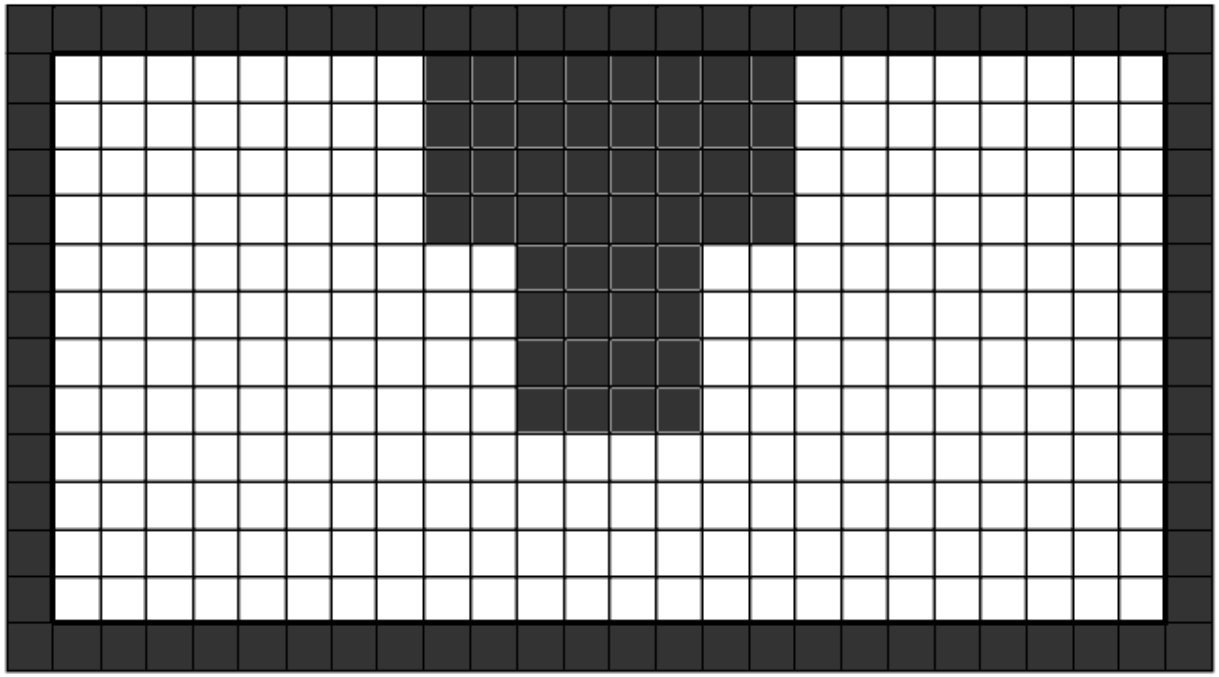
Using a Metacomputer

Grouping of Processes?



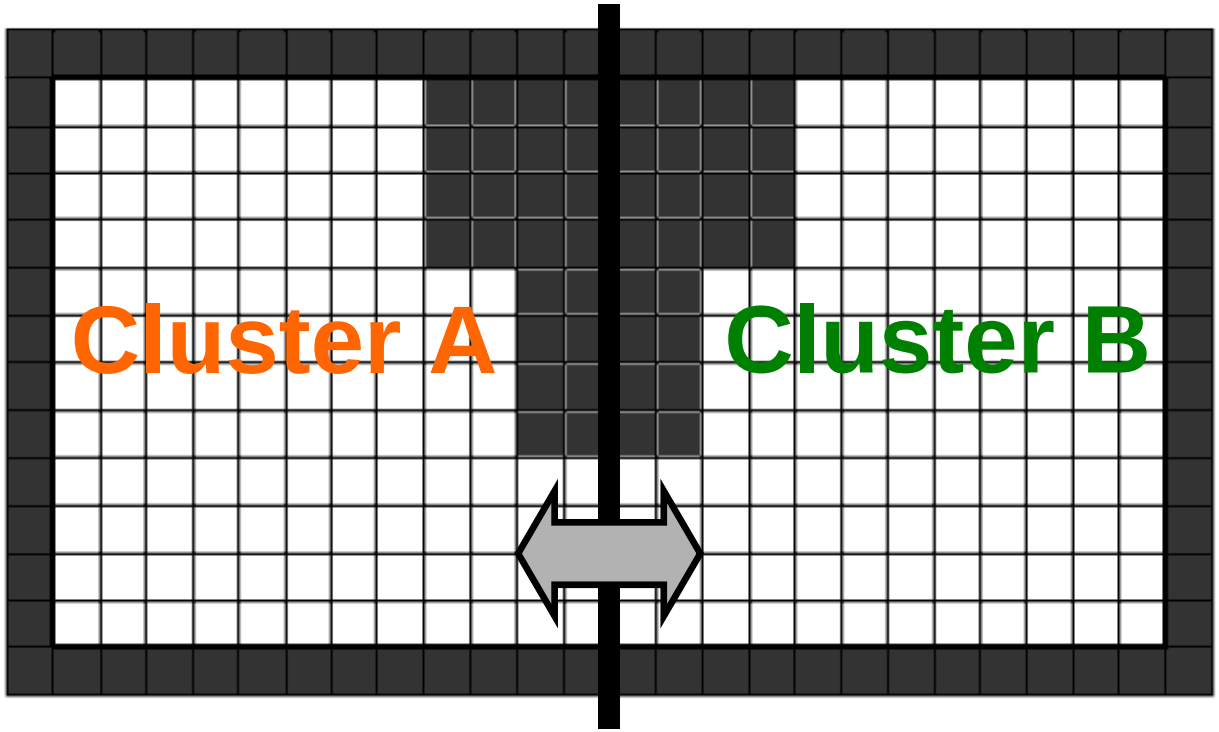
Using a Metacomputer

Inner Boundaries:



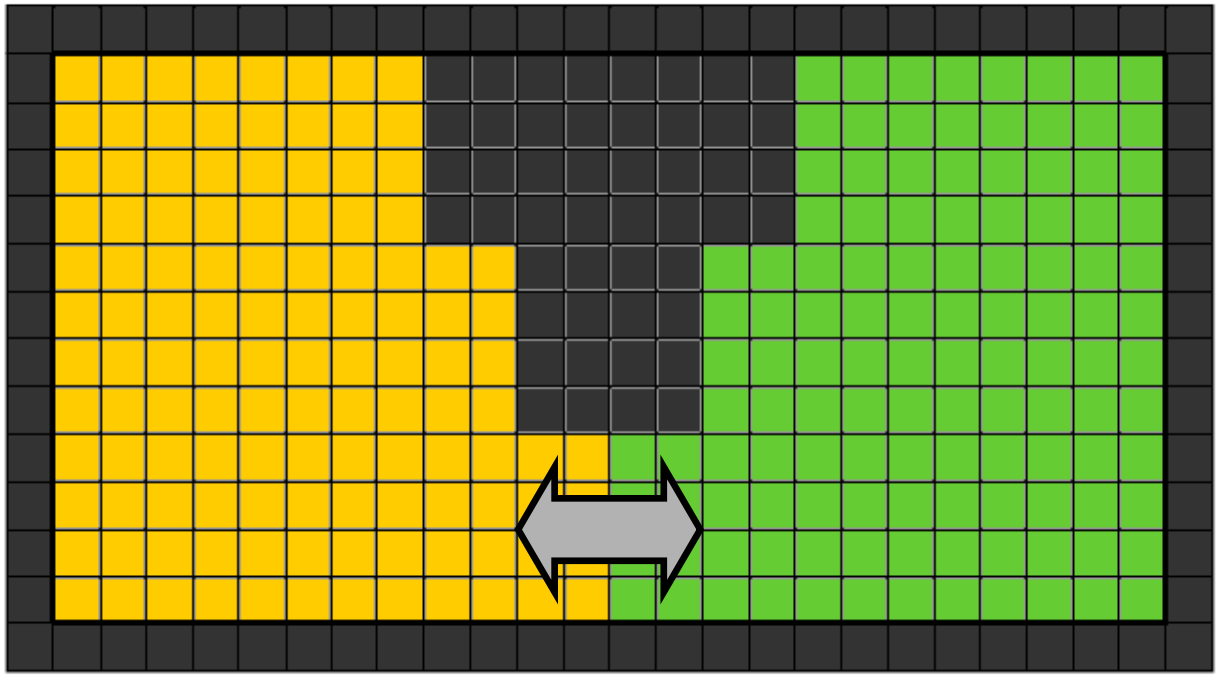
Using a Metacomputer

Reduced Communication:



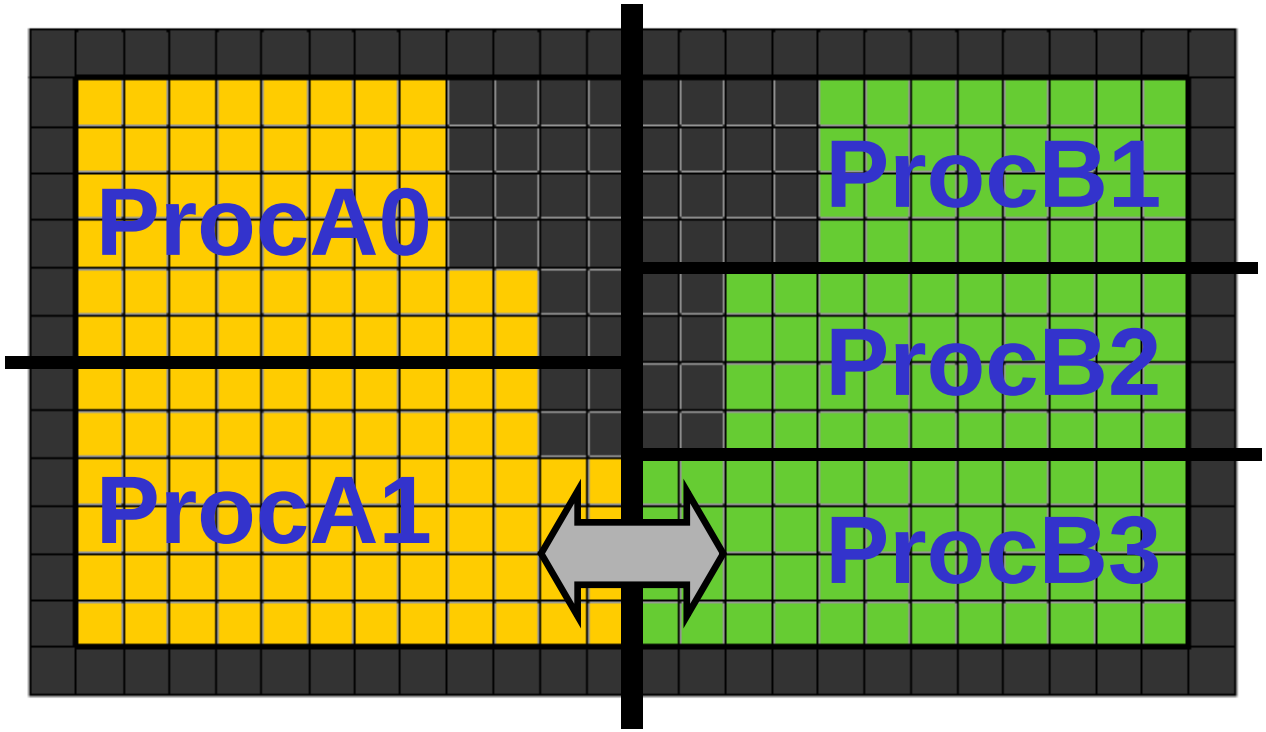
Using a Metacomputer

Load Balance?



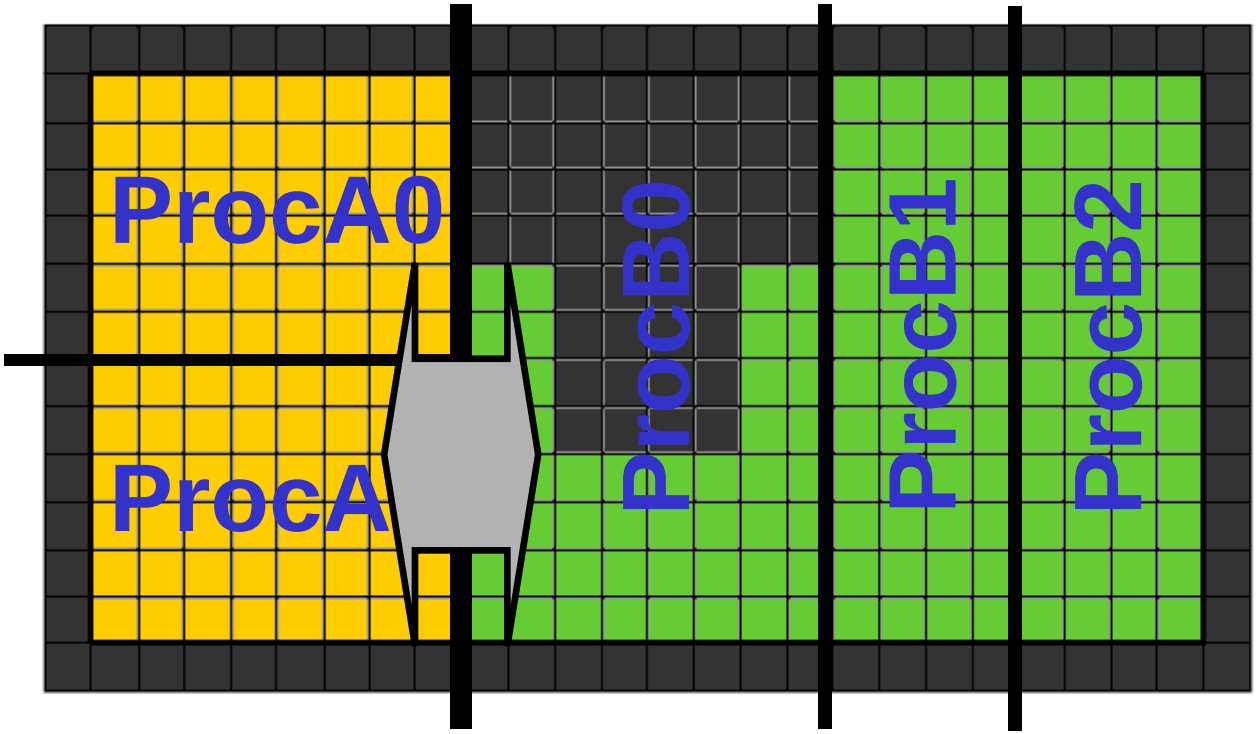
Using a Metacomputer

Domain Distribution onto Processes:

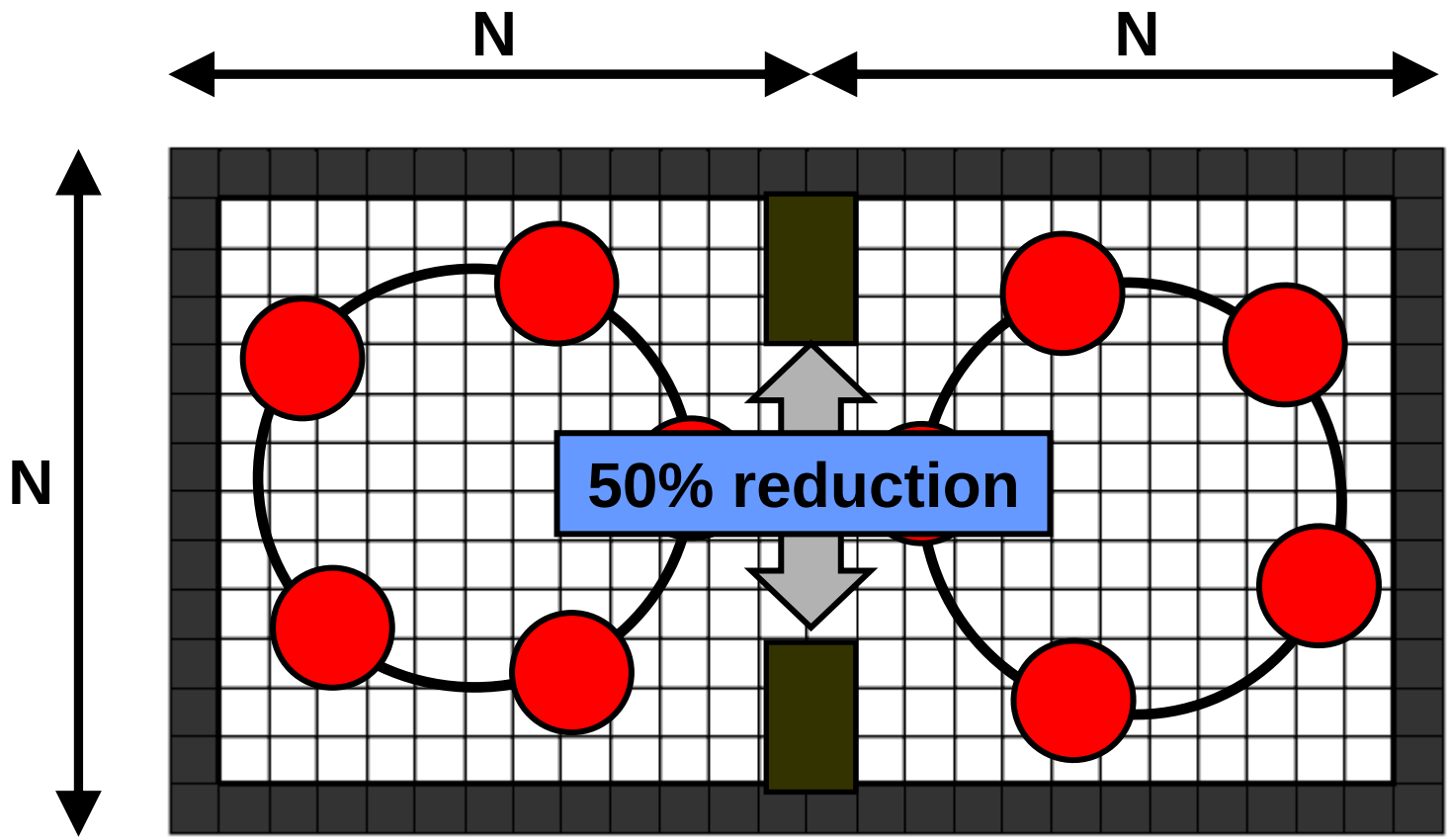


Using a Metacomputer

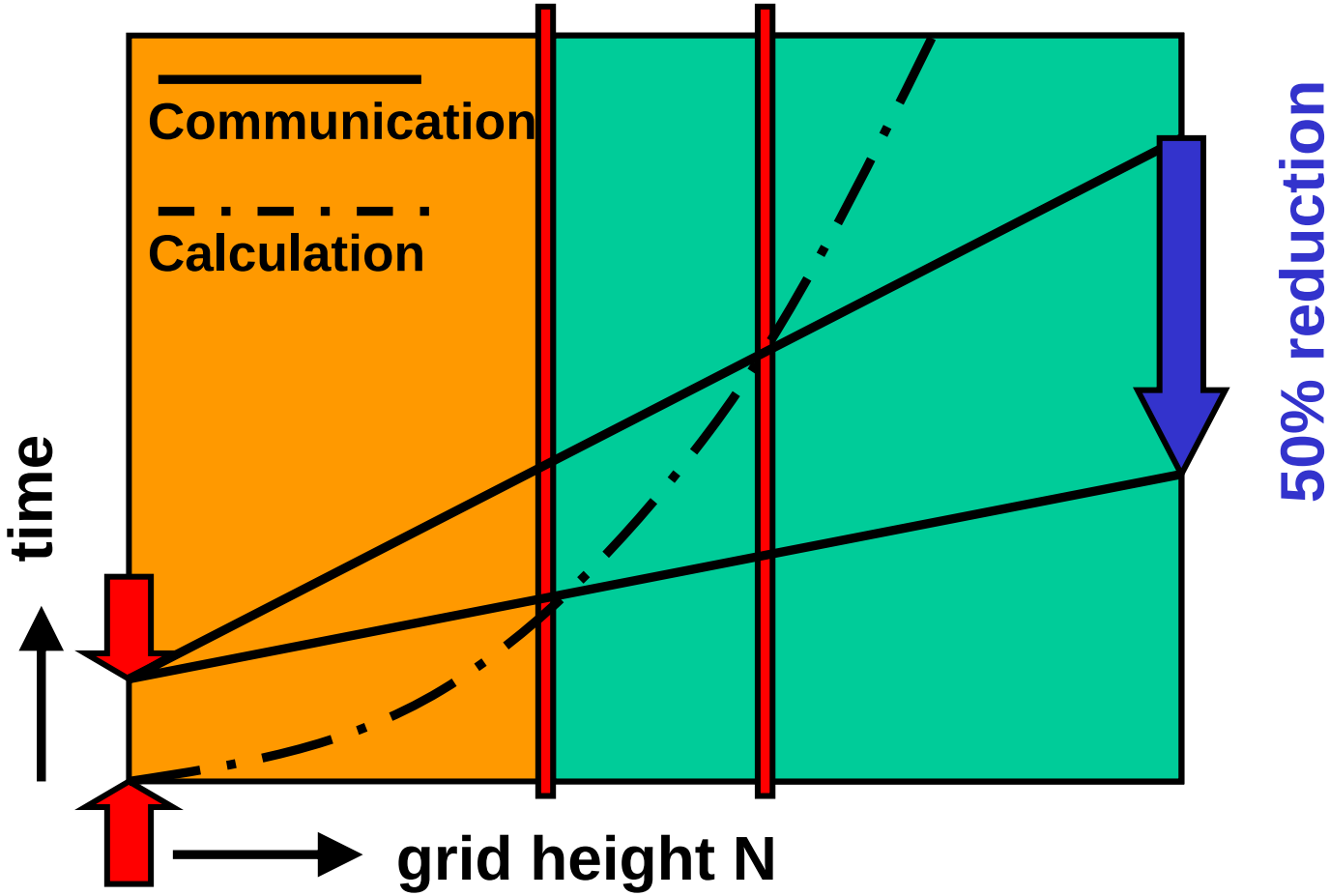
Load Balance VS Communication



Metacomputing?



Metacomputing?



Smart Partitioner

A *smart* partition scheme provides:

- **Load Balance *and***
- **Reduced Communication (if possible)**

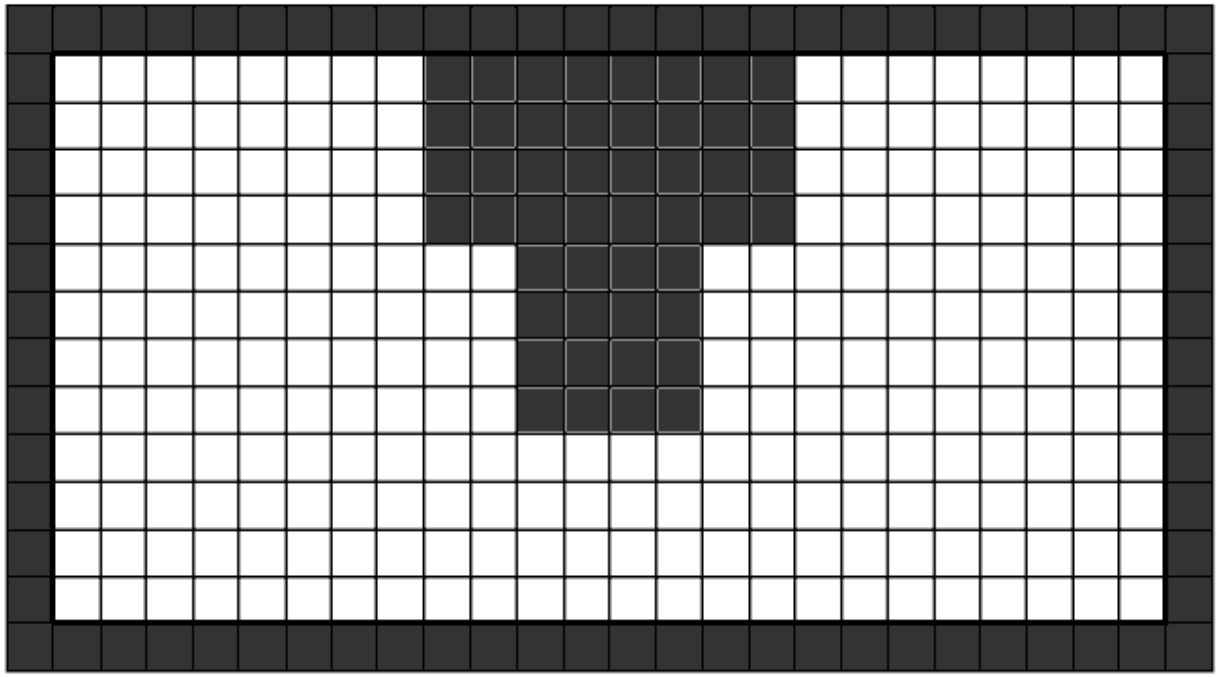
A *smart* partition scheme depends on:

- **the given boundary value problem**
- **the structures of the metacomputer**

Smart Partitioner (SmartPart)

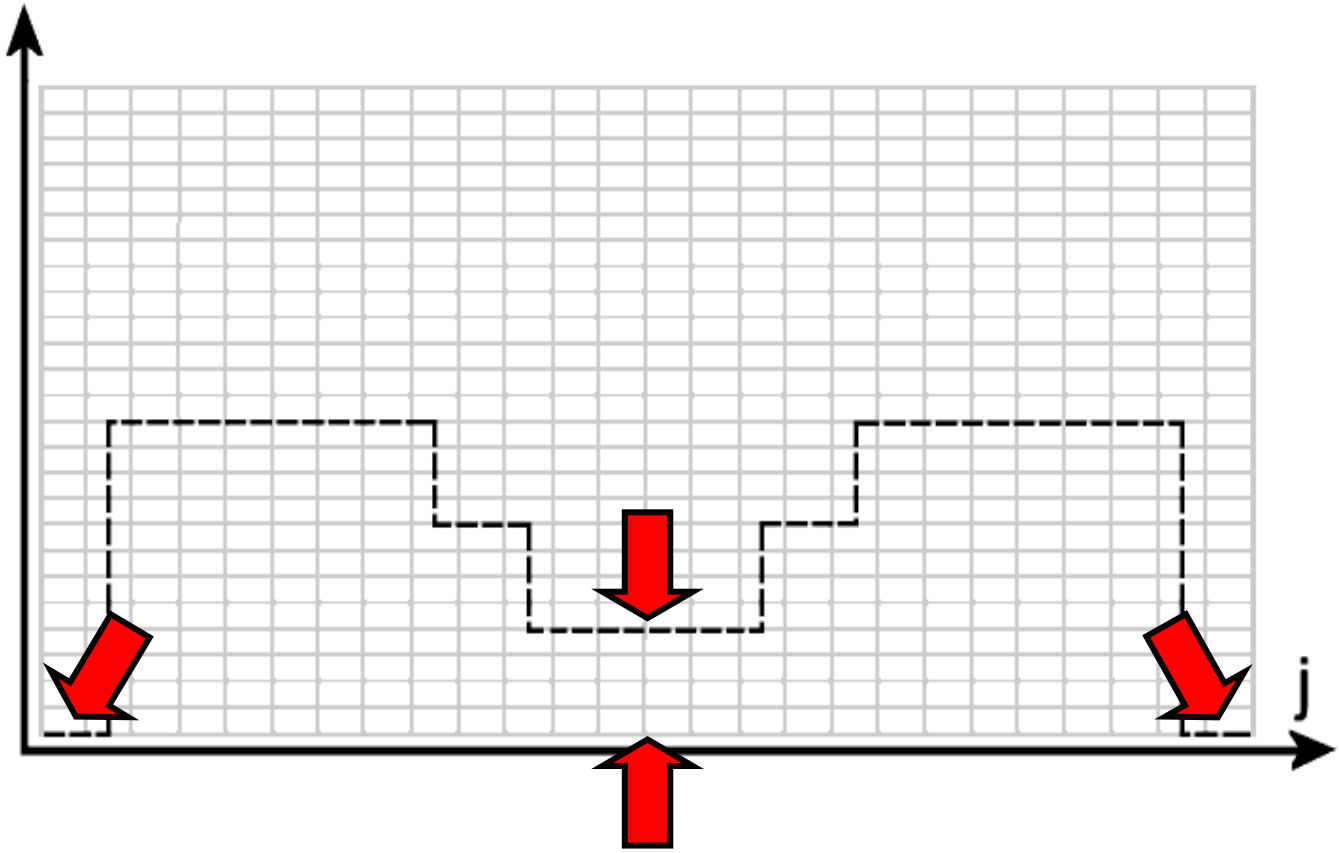
Smart Partitioner

Determination of Cut Metrics:



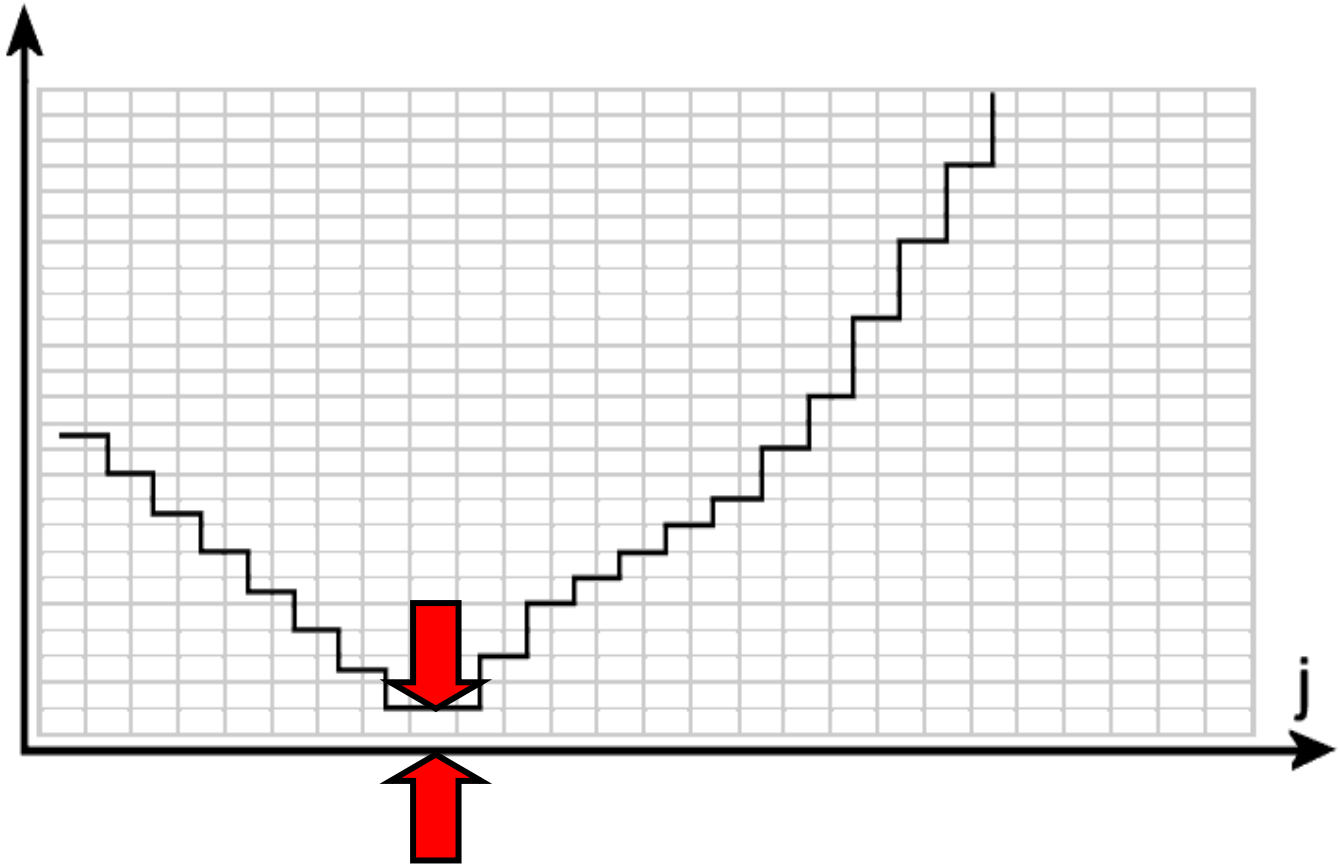
Smart Partitioner

Communication Metric:



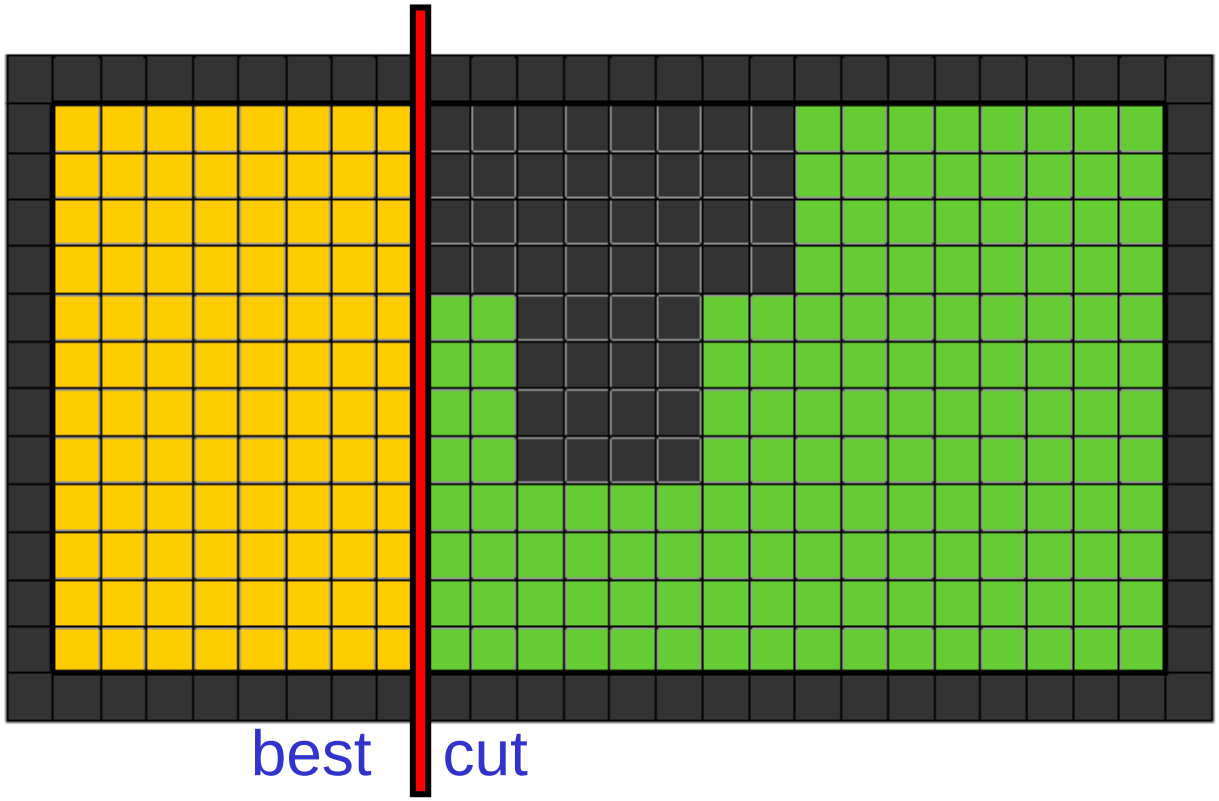
Smart Partitioner

Load Balance Metric:



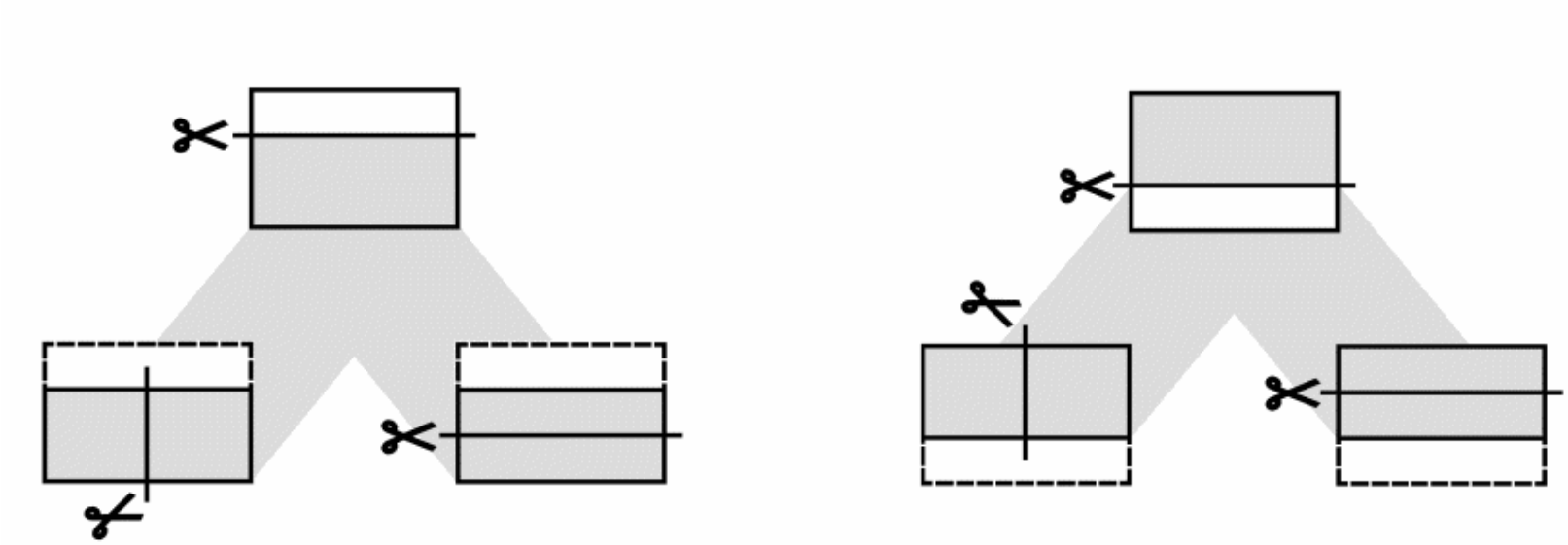
Smart Partitioner

Superposition:



Smart Partitioner

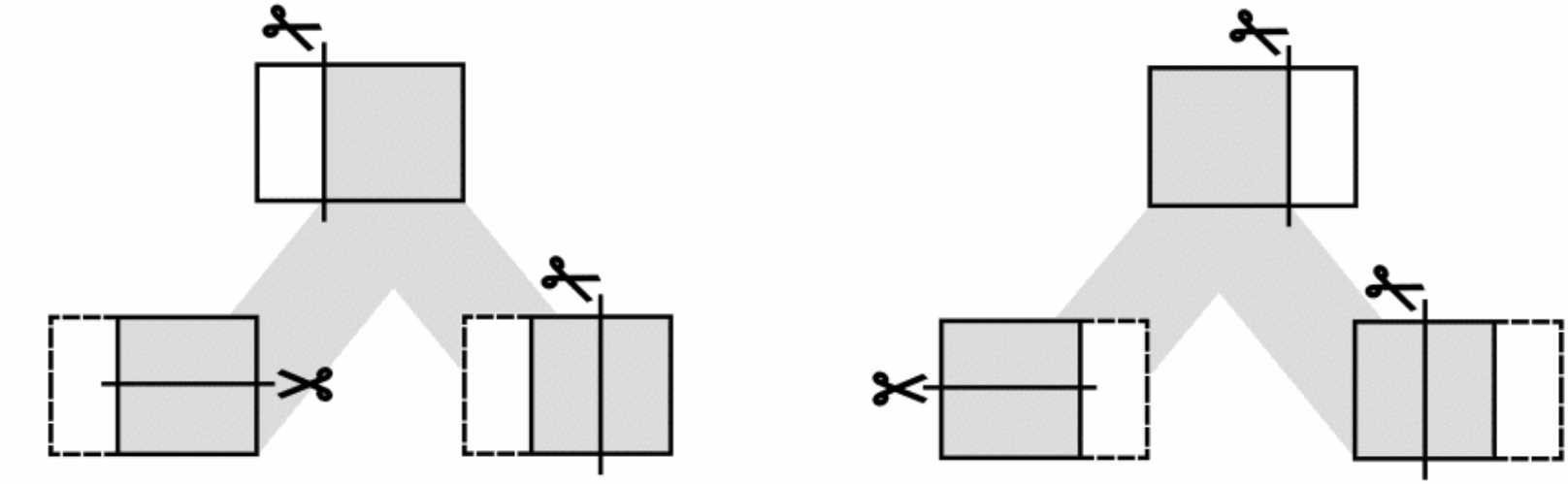
Decomposition Patterns:



three metahosts / first cut: horizontal

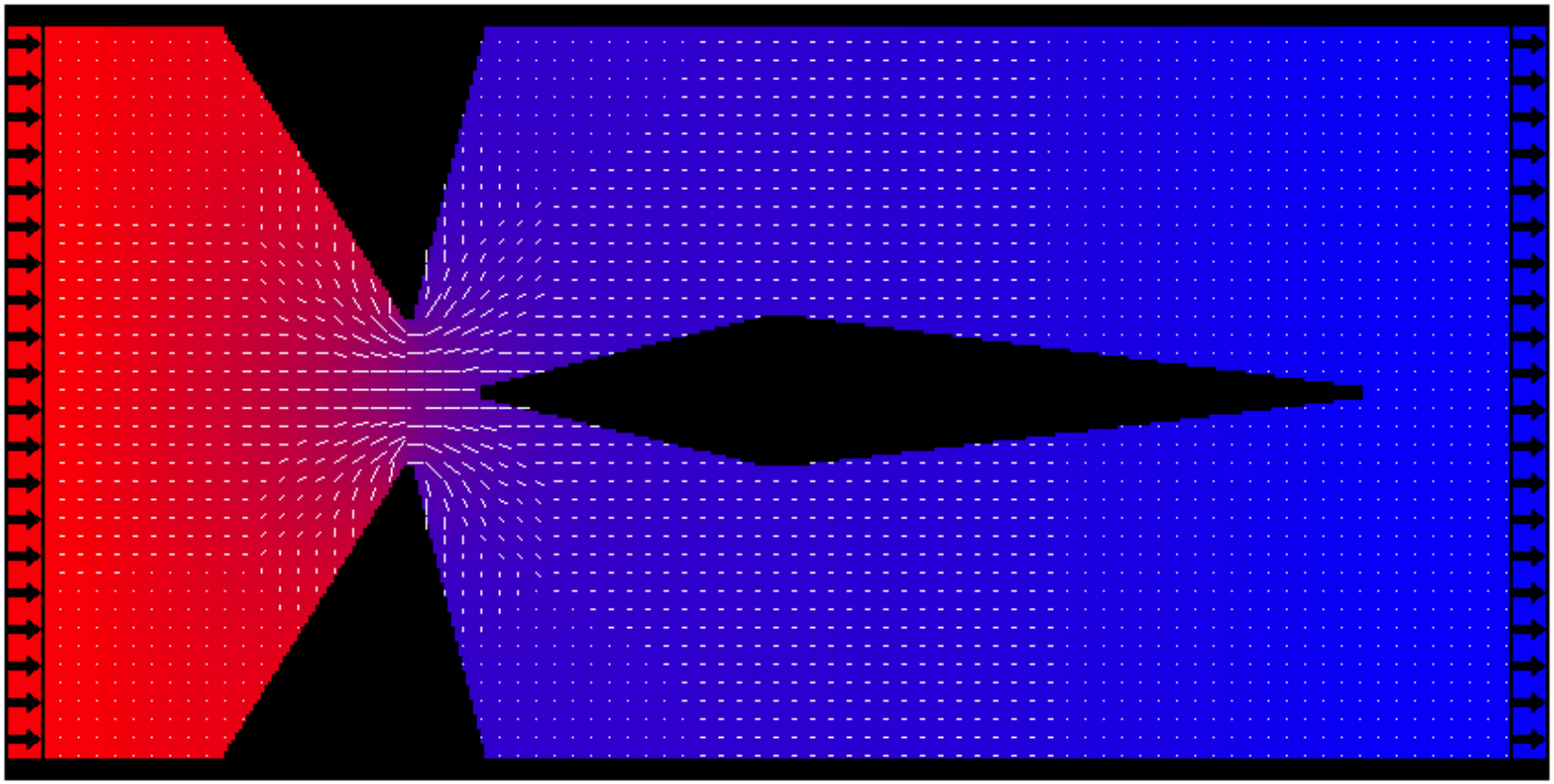
Smart Partitioner

Decomposition Patterns:

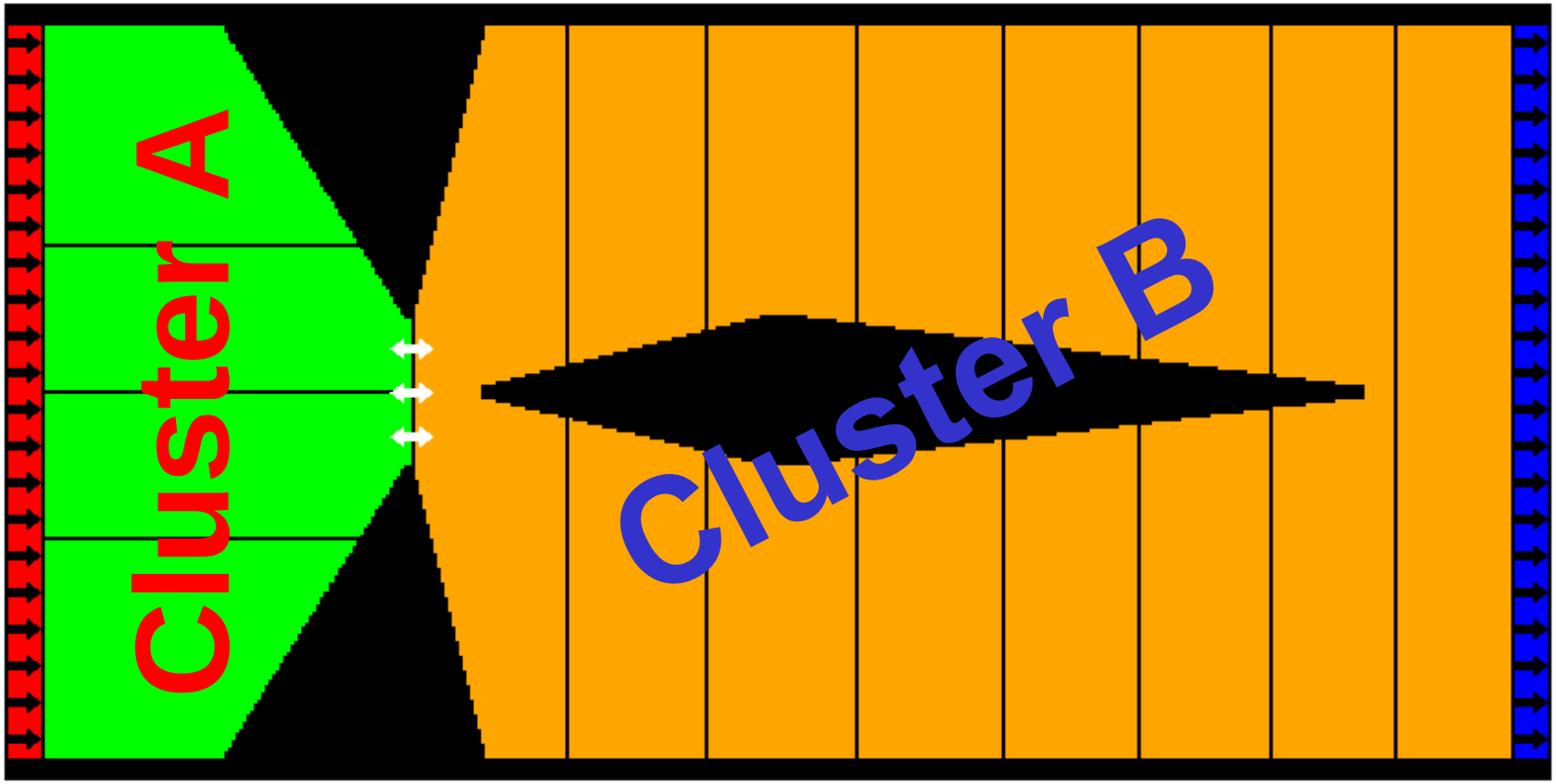


three metahosts / first cut: vertical

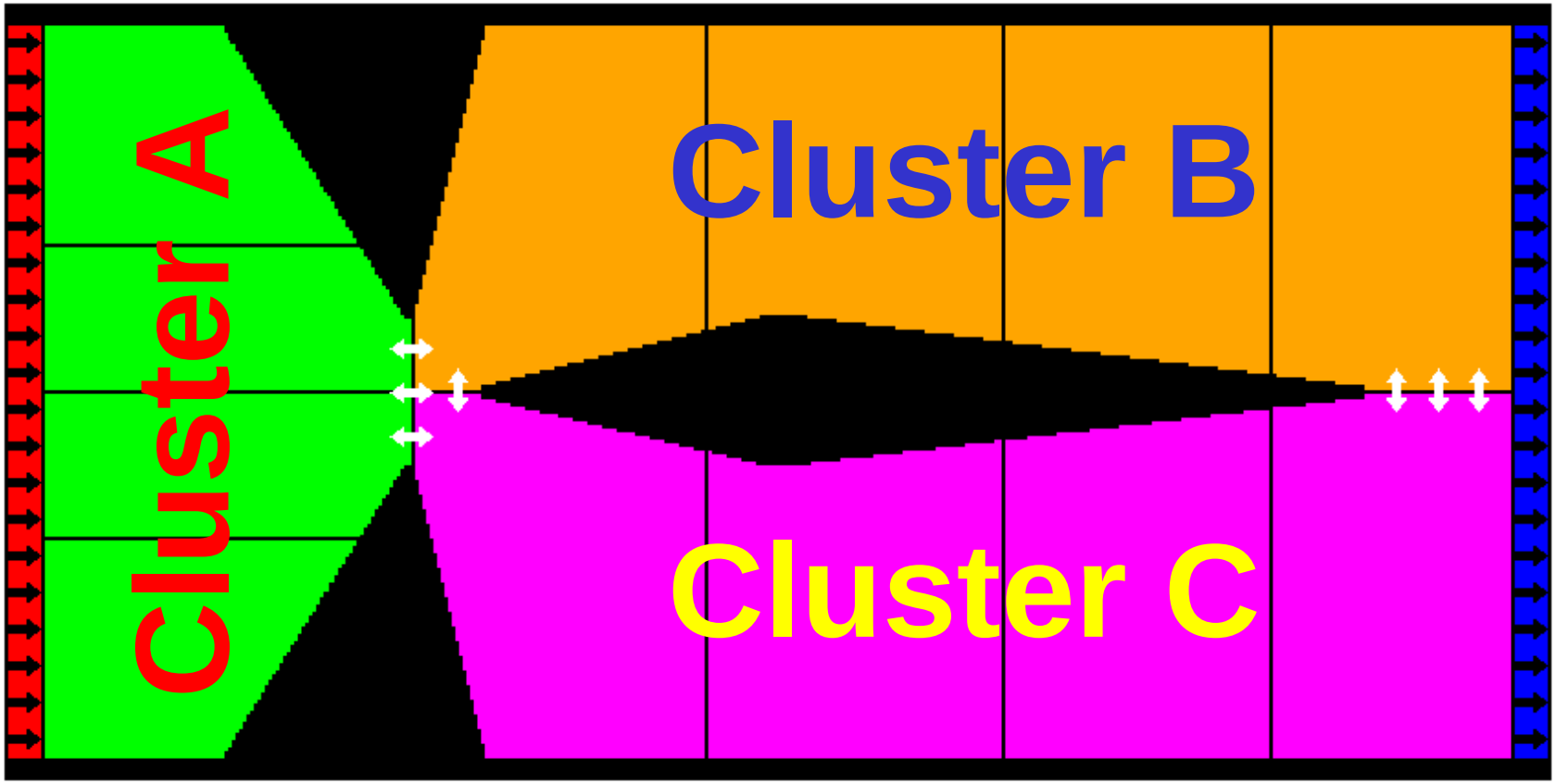
Example: CFD Simulation



Example: CFD Simulation

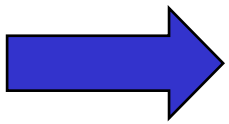


Example: CFD Simulation



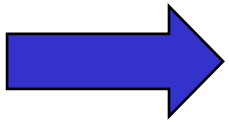
Adaptation Layer

Smart Partitioner

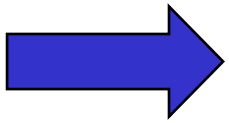


Optimal decomposition scheme for *your* problem on *your* metacomputer

How to use it in your application ?



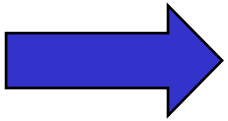
Adaptation Layer on top of MetaMPICH *your* applications can easily attach to



Communication Library (MetaComm)

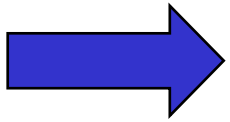
Adaptation Layer

MetaComm:



- additional communication functions
- can replace all explicit MPI functions

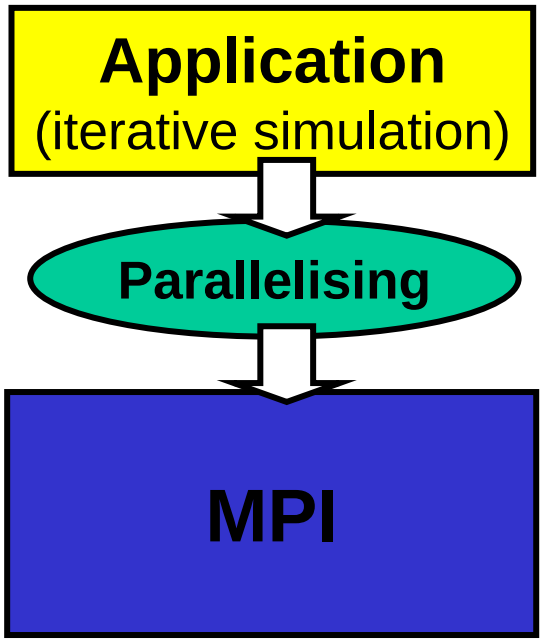
Optimised Communication:



- based on smart partition scheme
- performs *every* possible reduction
- metacomputer is still transparent

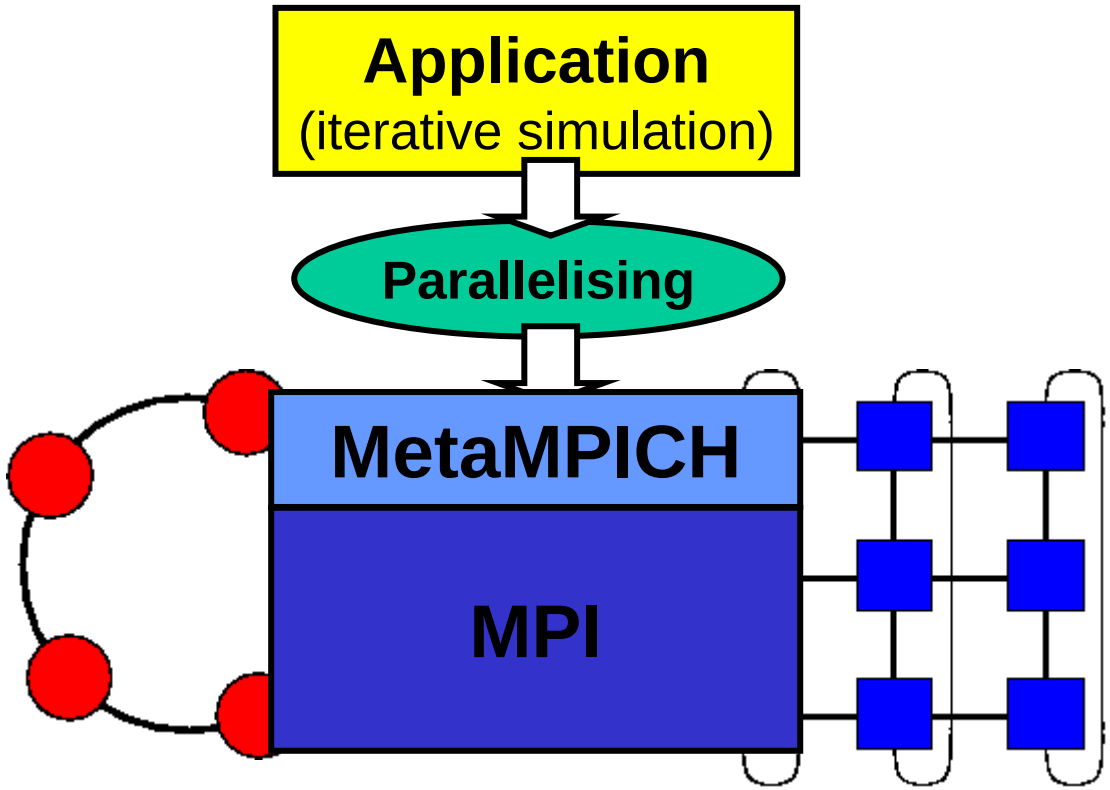
Adaptation Layer

***normal* case:**



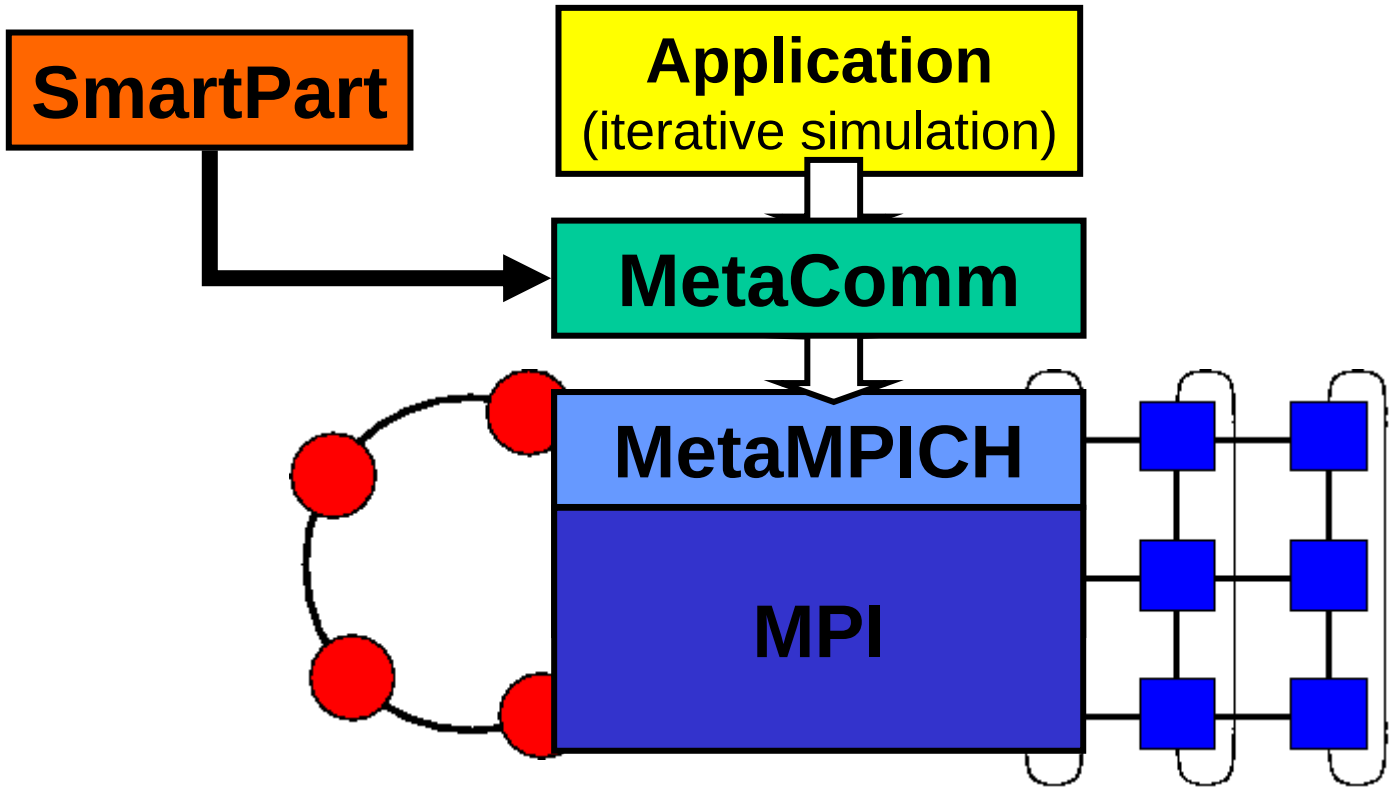
Adaptation Layer

Coupled Clusters:



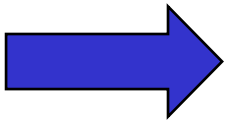
Adaptation Layer

transparent use → NO good performance ☹



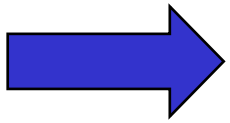
Conclusion And Outlook

Metacomputing And Coupled Clusters



- Software Library MetaMPICH
- Challenges of metacomputing

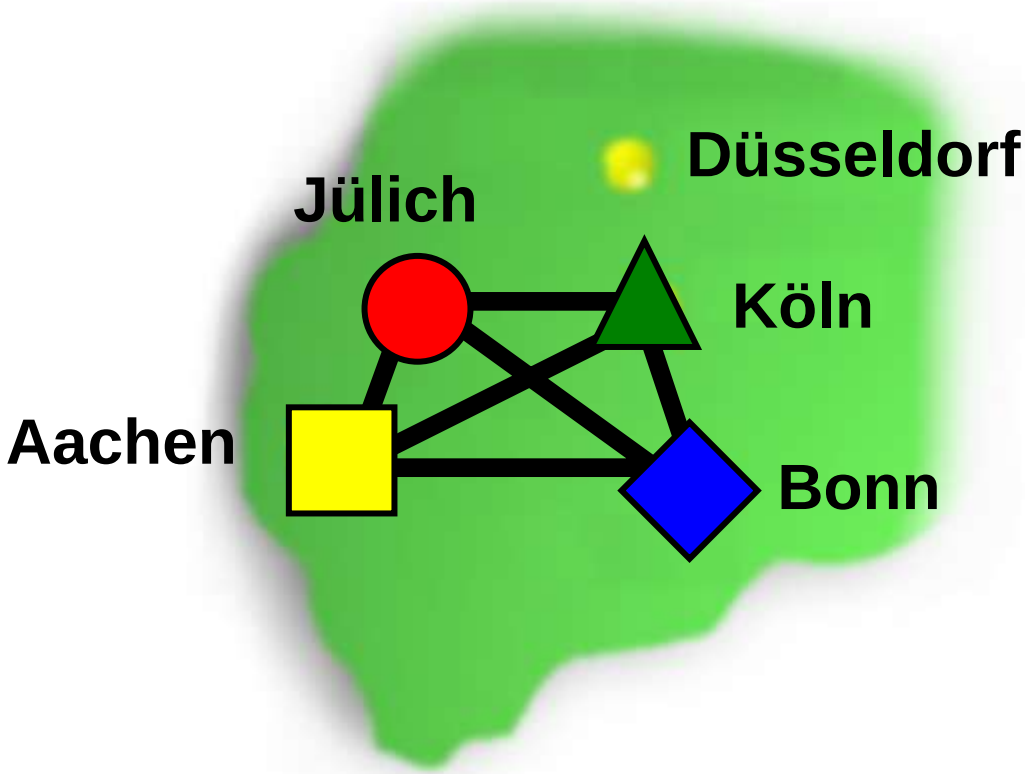
Optimising of Applications



- Algorithms on structured grids
- SmartPart and MetaComm

Conclusion And Outlook

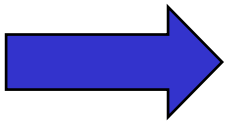
VIOLA-Project:



Conclusion And Outlook

Metacomputing *can* be a powerful way to increase performance.

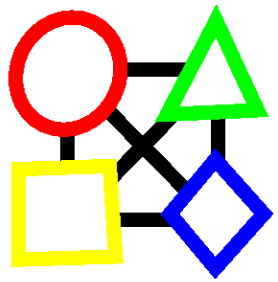
Metacomputer is a heterogeneous System !!!



You should always search for ways to optimise your applications !!!

Thank you for your attention!

Any Questions?



Lehrstuhl für Betriebssysteme
RWTH Aachen