

A Message Passing Library for inhomogeneous coupled Clusters

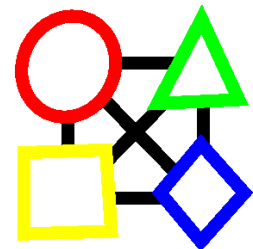
Martin Pöppe, Silke Schuch, Thomas Bemmerl

`martin@lfbs.rwth-aachen.de`

`http://www.mp-mpich.de`

RWTH Scalable
Computing
Aachen

Lehrstuhl für Betriebssysteme
RWTH Aachen





Overview

Introduction:

- Metacomputing environments
 - Clusters of Clusters

MetaMPICH

- MetaMPICH – a MPI for Metacomputers
- Performance
- Applications

Conclusion and Outlook



Clusters at the LfBS

4 Node Dual PII-450

- SCI D320

8 Node Dual PIII-850

- SCI D330

1 Node Quad Xeon-450

8 Node Dual Xeon-2.4

- SCI D340 2D Torus

**What to do with all these
Clusters together?**

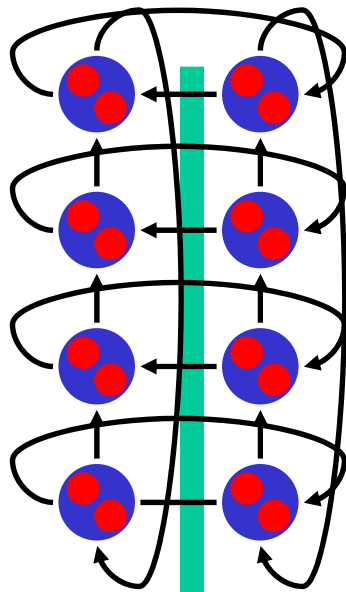




Connecting Clusters: Standard Ethernet

Using standard Networks generates bottlenecks!

SCI-2D-Torus: $8\mu\text{s}$ / 250 Mbyte/s



8 port

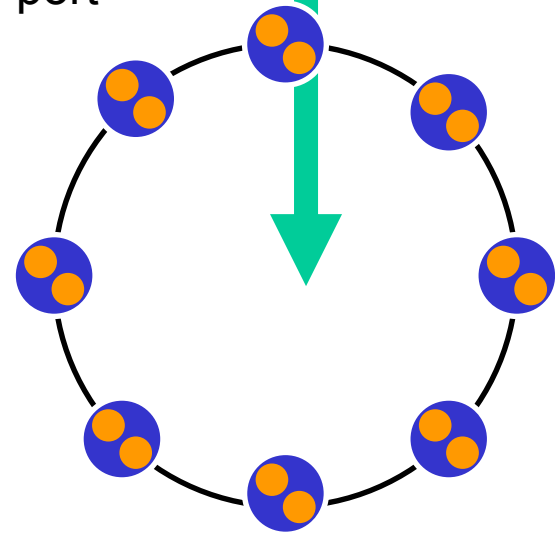
Ethernet Switch

1 port

$2 \cdot n \cdot m$ TCP connections

Ethernet Switch

8 port



SCI-ring

latency: $8\mu\text{s}$

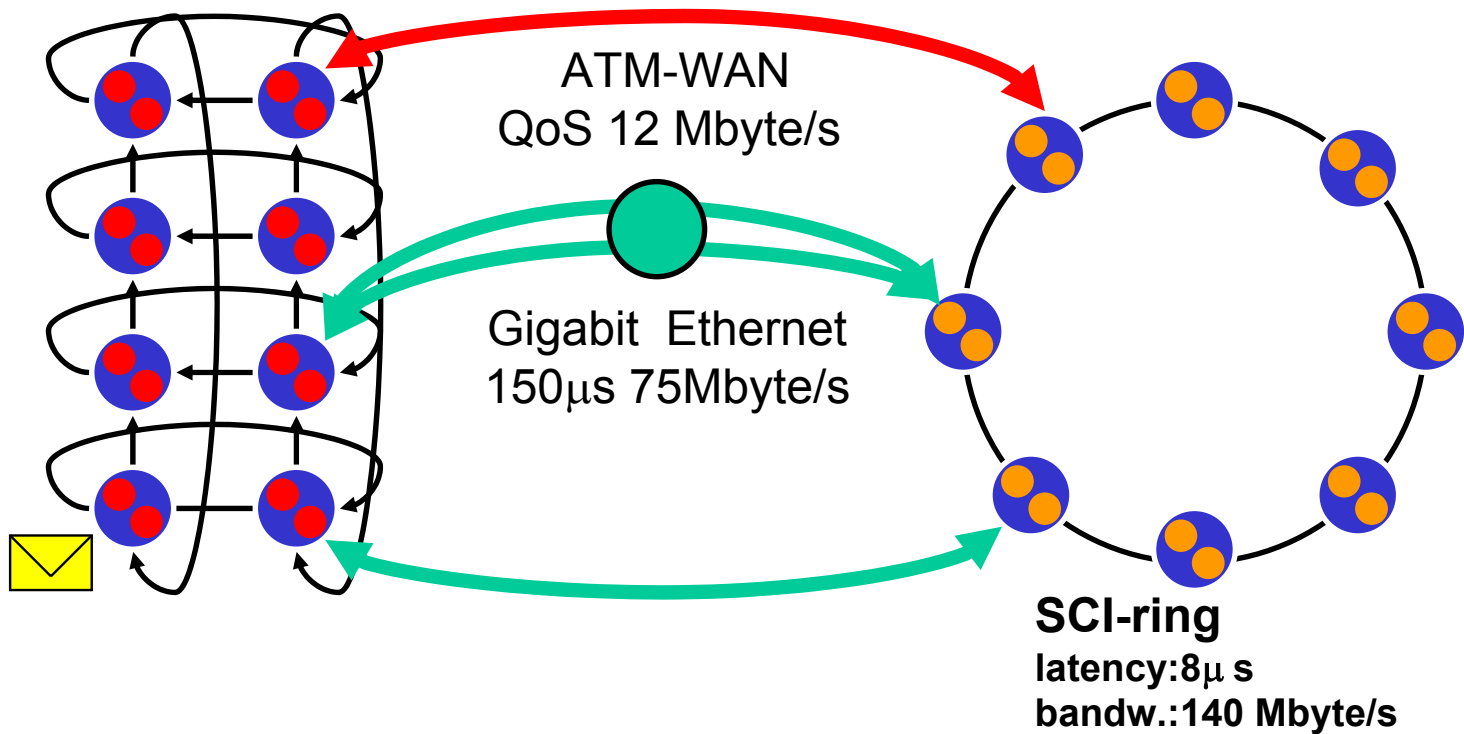
bandw.: 140 Mbyte/s



Dedicated Connections

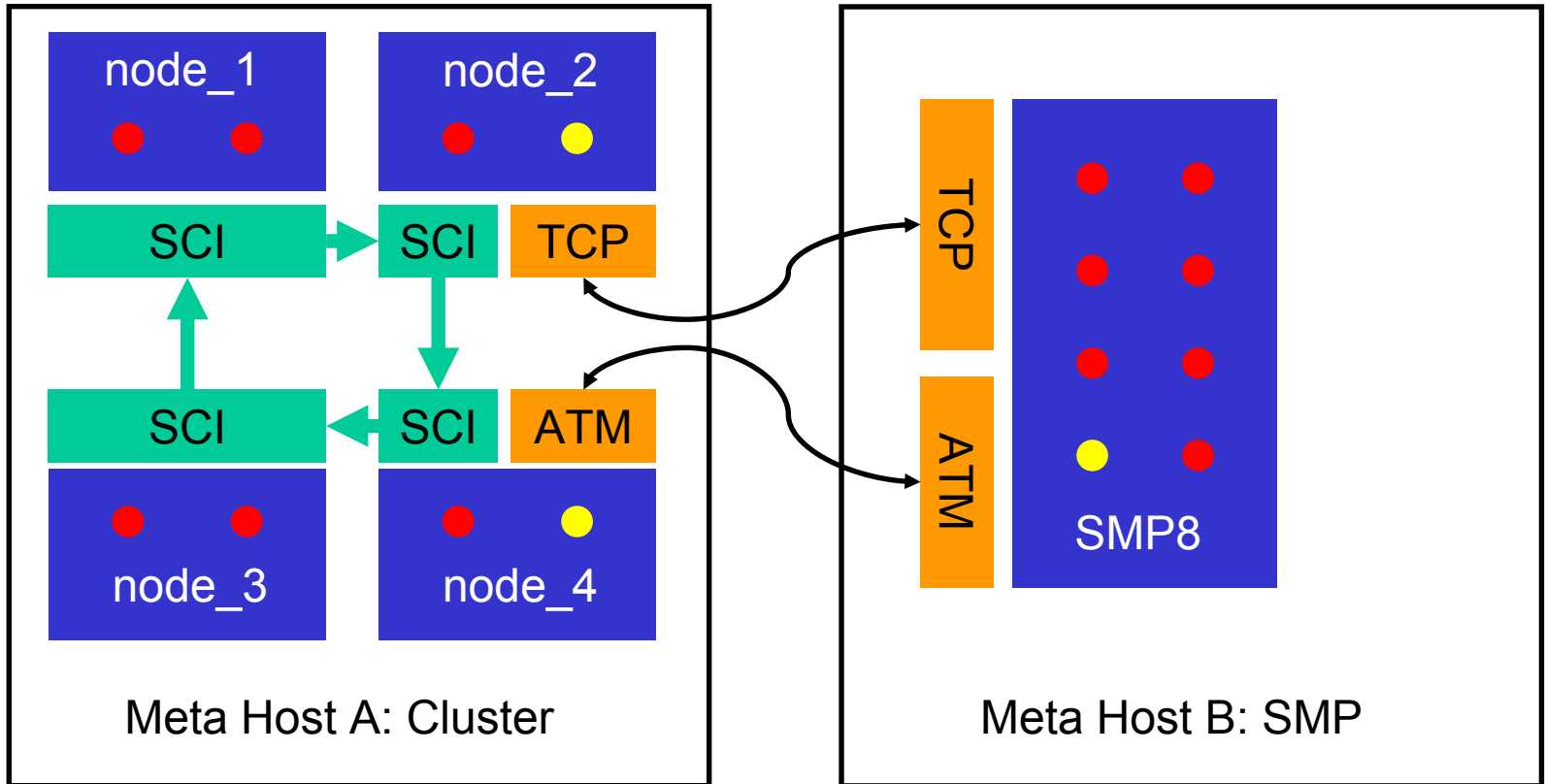
Router nodes increase scalability!

SCI-2D-Torus: $8\mu\text{s}$ / 250 Mbyte/s





Configuration: Example





Configuration Language (1)

```
METAHOST metahost_A {  
    EXECPATH=~martin/mp-mpich/examples/basic;  
    MPIROOT=~martin/mp-mpich;  
    ENVFILE=~martin/metahost_a_env;  
    TYPE=ch_smi;  
    NODES= node_1 - node_4 2,  
           node_2 2 (192.168.0.1) ,  
           node_4 2 (ATM_PVC 0.0.42) ;  
}
```

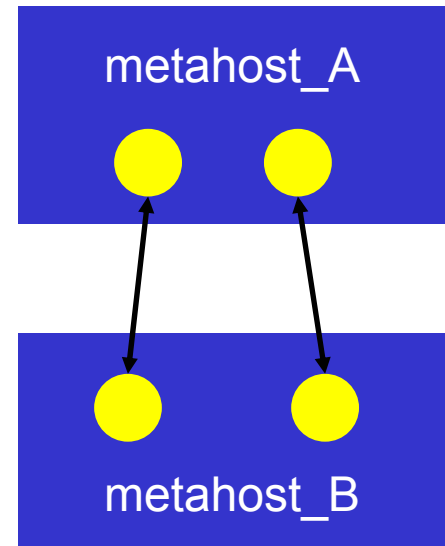


Configuration Language (2)

n meta hosts need $n*(n-1)$ communication relations:

```
PAIR metahost_A metahost_B 2
192.168.0.1 -> 192.168.0.2
ATM_PVC 0.0.42 -> ATM_PVC 0.0.42
```

```
PAIR metahost_B metahost_A 2
192.168.0.2 -> 192.168.0.1
ATM_PVC 0.0.42 -> ATM_PVC 0.0.42
```





MetaMPICH: MPI for Meta-Clusters

- **Complete implementation of MPI-1.2**
 - Based on the MPICH implementation
- **Part of the MP-MPICH project:**
 - Windows 2K/XP, Solaris, Linux
- **Communication with routers via pseudo ADI devices `ch_gateway` / `ch_tunnel`**
 - Only marginal changes to ADI-device needed
- **Supported networks:**
 - `ch_smi` and `ch_shmem` internal
 - TCP and AAL 5 external

<http://www.mp-mpich.de>



Architecture of MetaMPICH

SMP Host

MPI Application

MPI process

MPI process

MPI process

MPI process

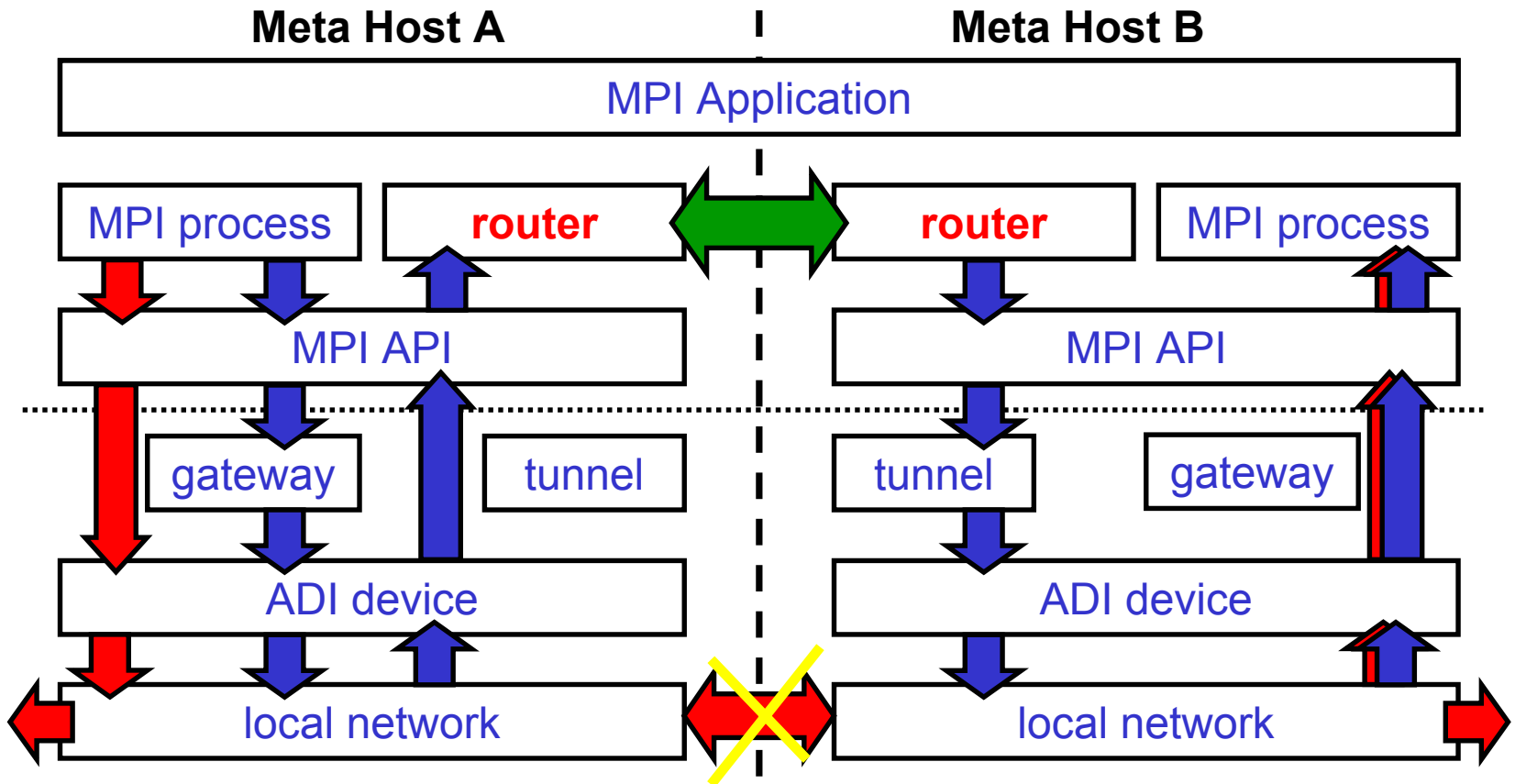
MPI API

ADI device

Cluster Interconnect

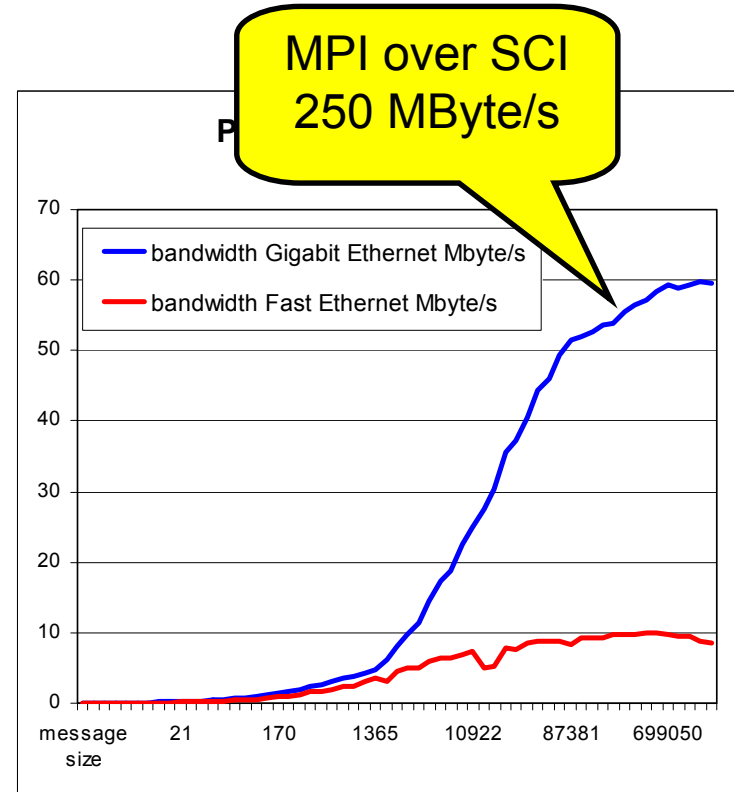
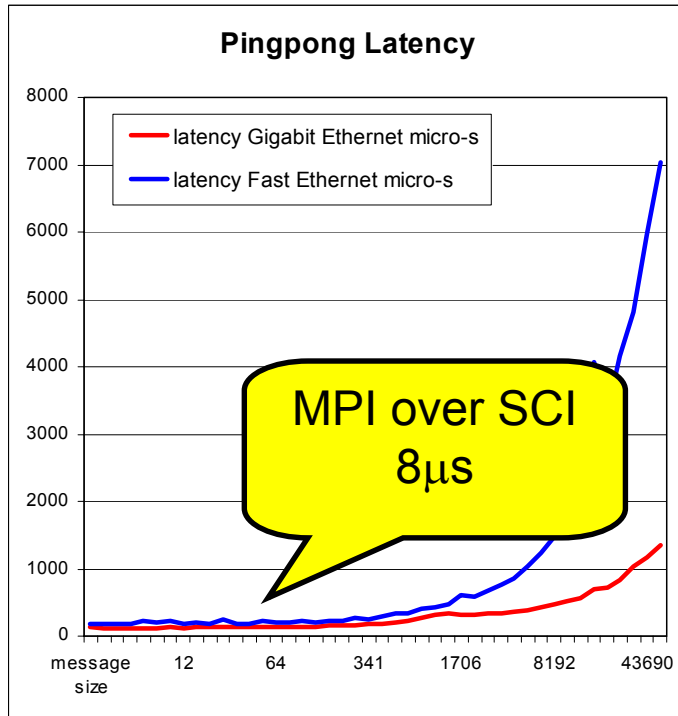


Architecture of MetaMPICH (2)





Performance – Pallas MPI-Benchmark



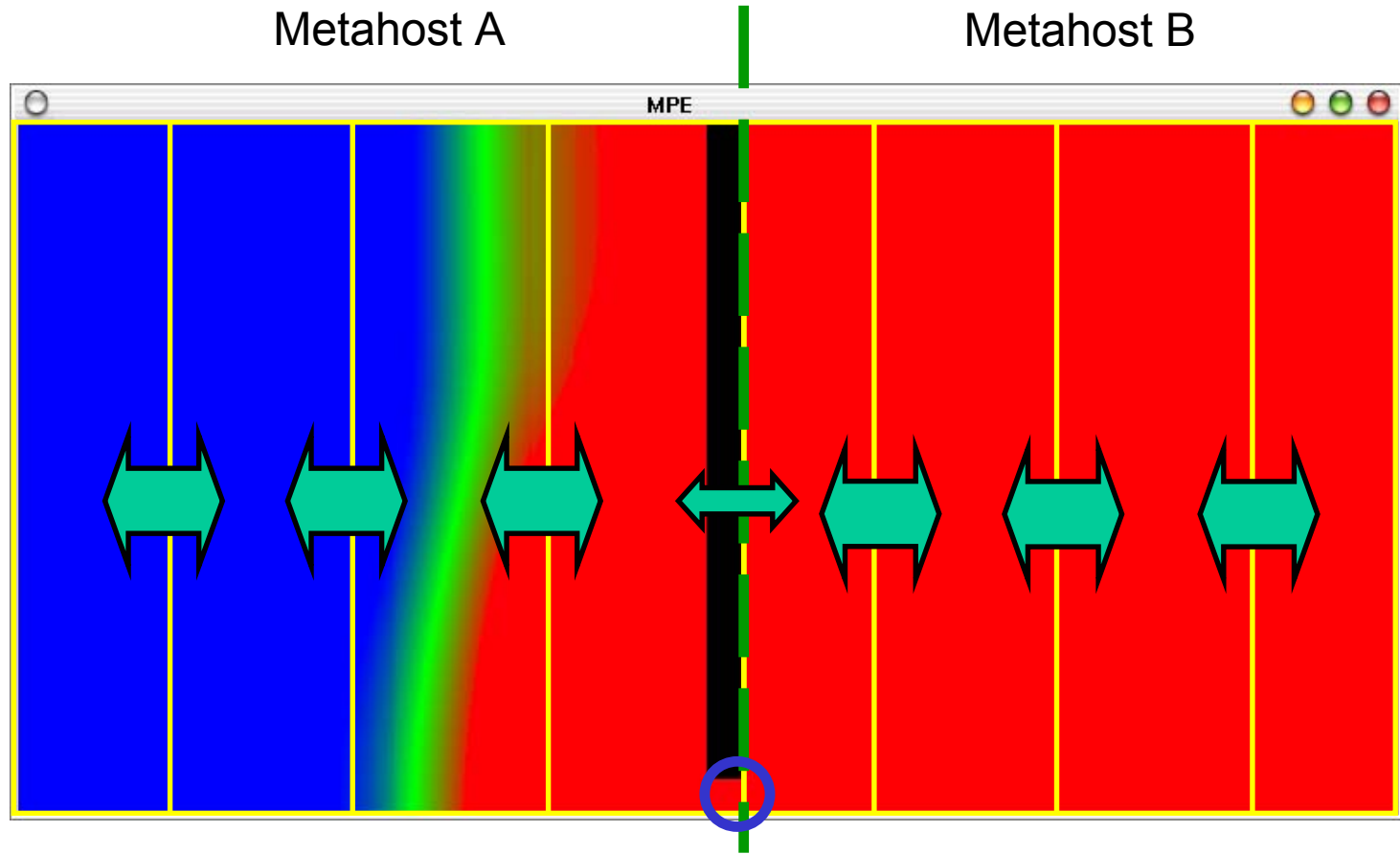


Applications

- **MPI applications will run out of the box**
 - ... the slow connection will slow it down!
- **MetaMPICH provides a new communicator:**
 - `MPI_COMM_LOCAL` for local meta host
 - Applications *can* use it for local communication
- **Scale will ..**
 - Applications with smart partitions
 - Weakly coupled simulations



Applications – Smart Partitions





Conclusion and Outlook

Features of MetaMPICH:

- MetaMPICH provides a MPI platform for coupled SCI-Clusters and SMP
- Communication structures can be configured very detailed, TCP and AAL5 supported

To do:

- Integration into grid resource management systems (e.g. SGE), own development is in progress
- Simple configurations done automatically
- New meta host interconnects – infiniband?