

Design and Implementation of a Service-integrated Session Layer for Efficient Message Passing in Grid Computing Environments

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

ISPDC 2008, Krakow, Poland



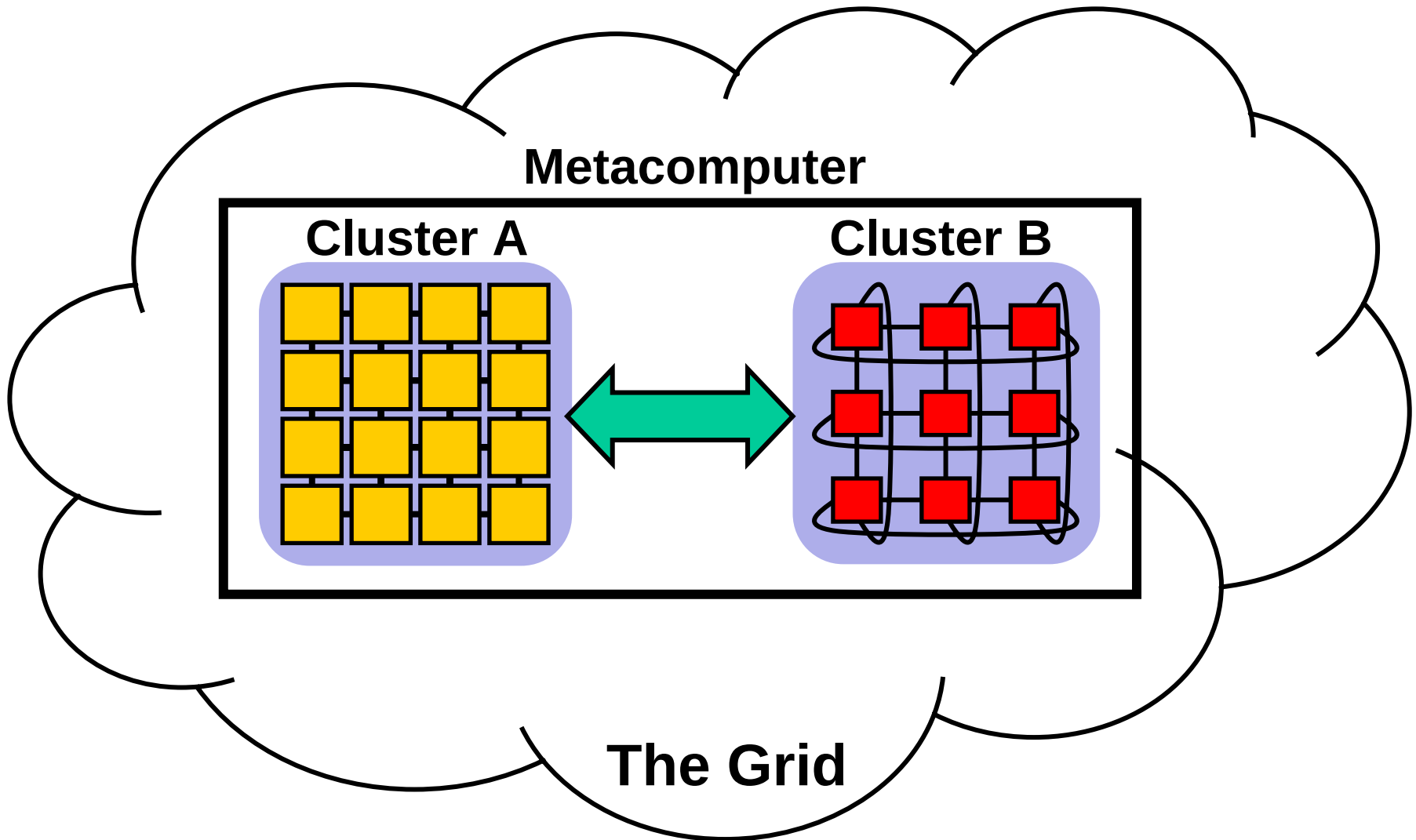
LEHRSTUHL FÜR BETRIEBSSYSTEME

Univ.-Prof. Dr. habil. Thomas Bemerl



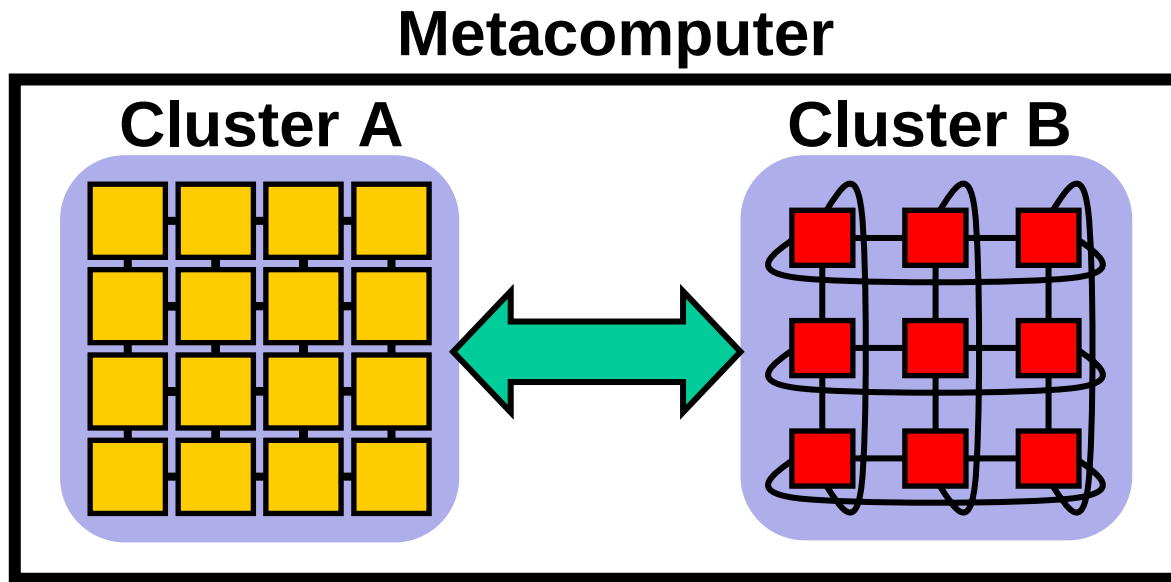
- **Metacomputing / Grid-enabled MPI**
- **The ISI Session Layer**
- **The Integrated Services**
- **Integrating ISI into MetaMPICH**
- **Conclusions**

- **Example: Cluster of Clusters**



- **Dedicated Cluster Networks**

InfiniBand, Myrinet, SCI, Quadrix, ...

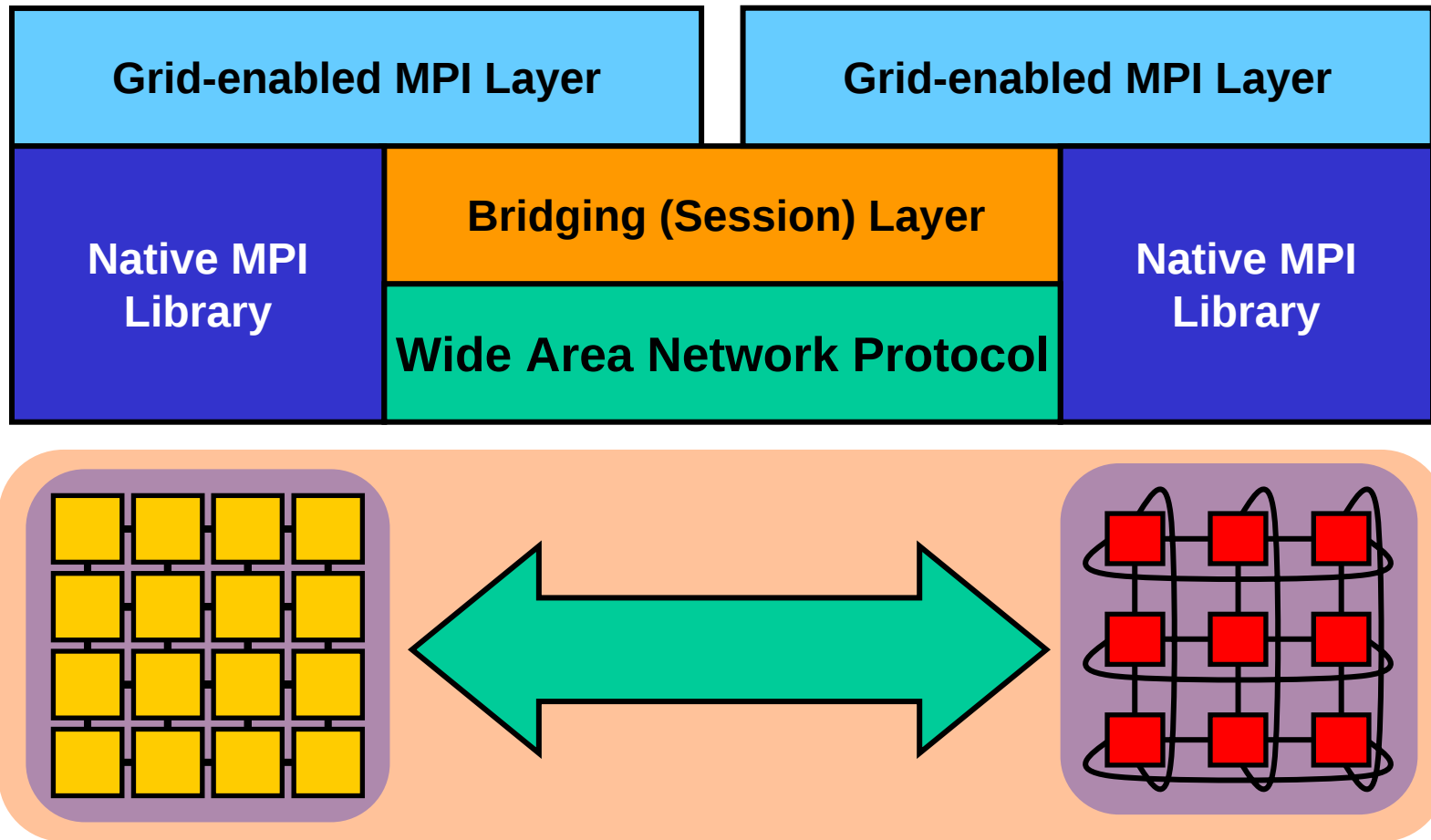


- **Interlinking Networks**

Wide-Area-Networks, (Local-Area-Networks)

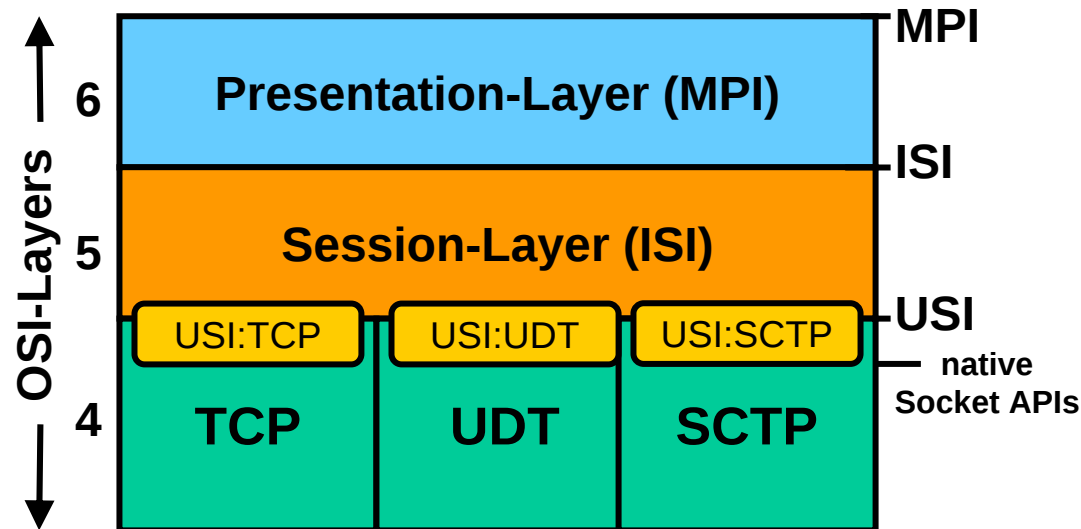
- Heterogeneous Communication Structures!
- Need: **Topology-aware** Message-Passing System!
- Metacomputing and **Grid-enabled** MPI-libraries:
 - **GridMPI**
 - **MPICH-G2**
 - **PACX-MPI**
 - **MetaMPICH**
 - ...
- Support for **local high-performance networks**
 - These libraries in turn often rely on **site-native** MPI libraries (like a *Vendor-MPI*)

- A Bridging Layer for inter-Site Communication



- **Transport Protocols for Wide-Area-Networks**
 - Kernel-Level: **TCP, SCTP**
 - User-Level: **UDT, PSocketS, ...**
(rely internally on UDP or TCP, again)
- **Why additional User-Level Transport Protocols?**
- **TCP is not always the best choice!**
 - TCP is designed for **general purpose**
 - **slow congestion** control **recovery** mechanism
 - problems especially in **high-bandwidth** but **high-latency** WANs

- **Key Task:** establishing all point-to-point connections between the distant processes (**on demand!**)
- **Overview:**



- **A Uniform Socket Interface: USI**
 → for **encapsulating** the **native Socket-APIs**

- Questions of **Session Configuration**:
 - Which USI module should be used?
 - How to connect to a remote process?
 - How many other processes exist?
 - ...
- An ISI **Domain Configuration**:

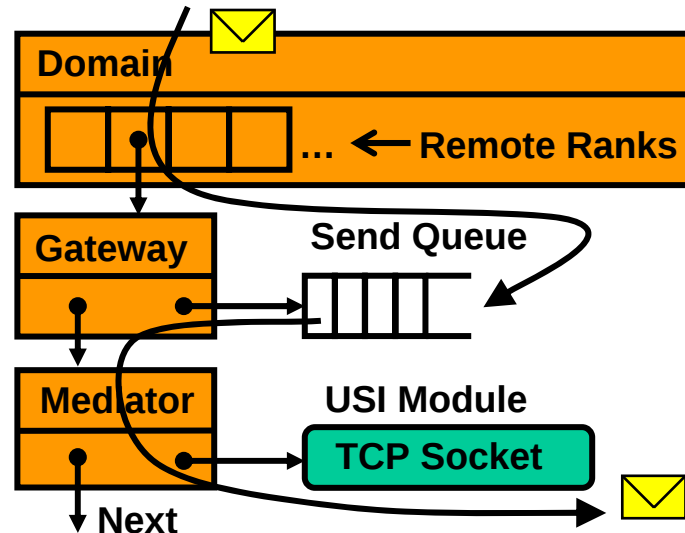
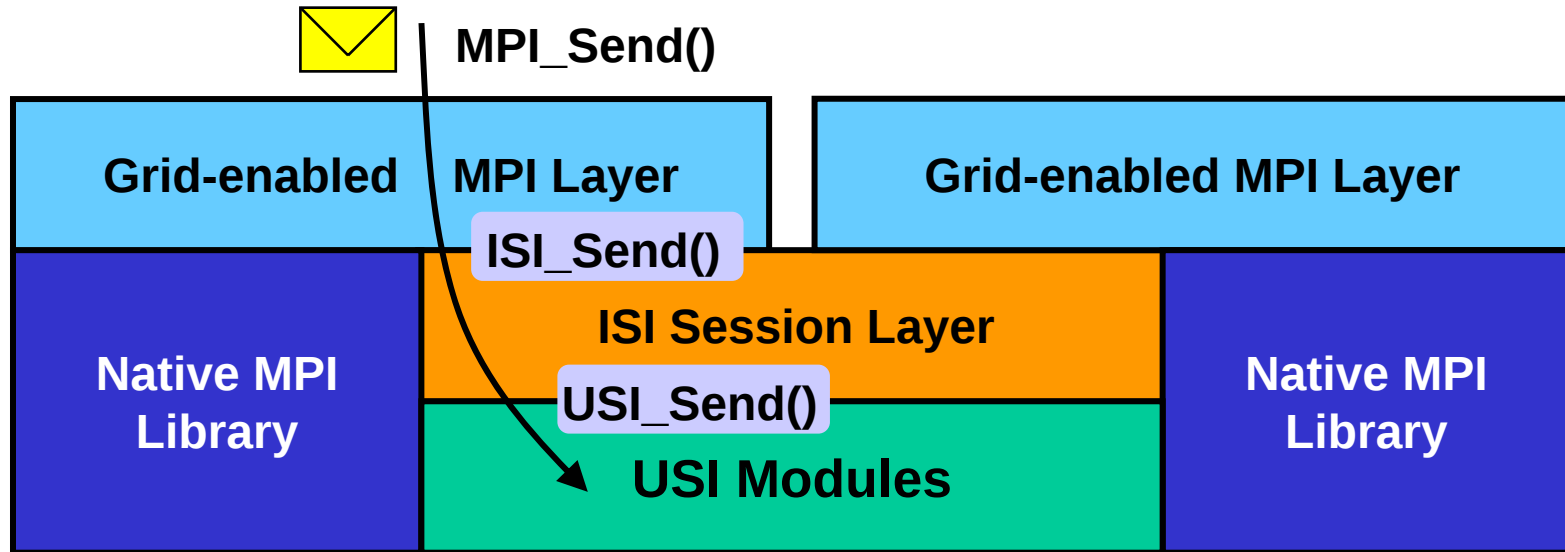
```
<?xml version="1.0"?>

<domain identifier = "my_domain">
  <size> 32 </size>

  <process identifier = "pd-01">
    <rank> 0 </rank>
  ...
</domain>
```

- How to **retrieve a domain configuration?**
 - configurations must be **passed to each process** within an ISI session
 - different environments demand **different retrieve methods**:
 - processes can contact a configuration server
 - the configurations are read from a local file
 - ...

• Handling of Payload Transfers



- **Handling of Payload Transfers**
 - ISI works **strictly header-less**
 - posted requests are buffered in **FIFO queues**
 - payload is transferred by ***Mediator* service objects**
 - **multiple payload connections** between two processes are possible!
- **Which of these connections should be used?**

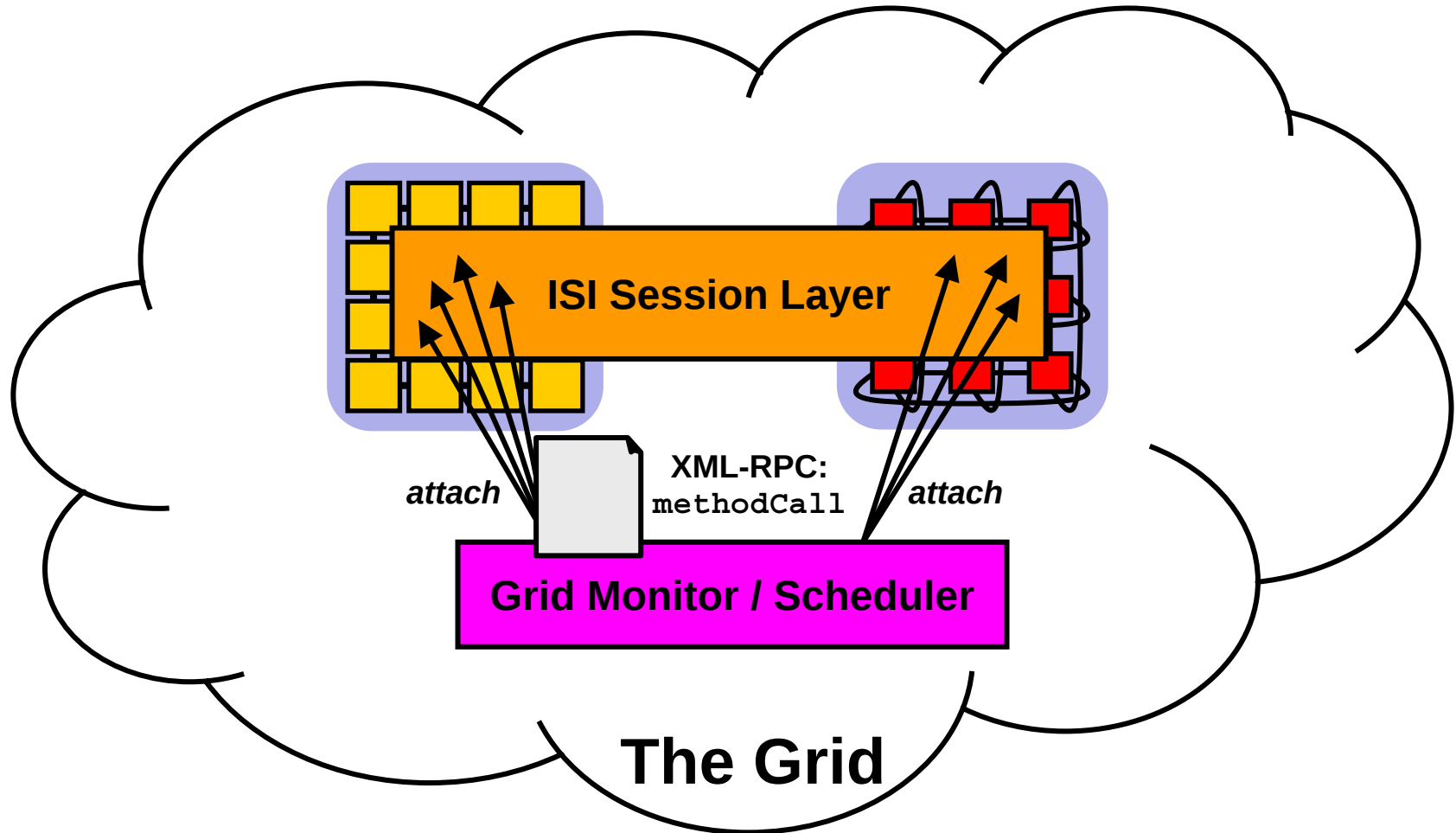
- ISI Service **Entities**

- C++ programmed **Objects**
- Objects are **triggered frequently** (Round-Robin or Lottery-Scheduling)
- Service instances can be **created** at **startup-time** and even at **run-time**

- ISI Service **Interface**

- objects can be **accessed** and **controlled** via **Remote Procedure Calls** (XML-RPC)
- sessions can be **monitored** and even **re-configured** at run-time

- Interaction with the Grid Environment



- **Different Types of Service Objects**
 - ***Acceptors:***
for accepting incoming connections
 - ***Mediators:***
for payload transfers
 - ***Surveyors:***
for probing a connection's quality
 - ***Supervisors:***
for acting as layer-internal daemons
 - ***Executors:***
for handling service inquiries

- Creating Services at **Startup-Time**

ISI Domain configuration:

```
<?xml version="1.0"?>
...
<gateway rank = "0">

    <acceptor>
        <module>TCP</module>
        <address>123.123.123.100:55555</address>
        <attribute key="timeout">120</attribute>
    </acceptor>

...
<gateway rank = "1">

    <supervisor name="LazyAutoConnect">
        <module>TCP</module>
        <address>123.123.123.101:55555</address>
    </supervisor>

...
```

- **Creating / Adding Services at Run-Time**

XML-RPC Method Calls:

- *AddNewAcceptor*
- *AddNewSupervisor*
- *AddNewMediator* (→ new Payload Connection)
- *SwitchPayloadConnection*
- *DeletePayloadConnection*
- ...

- **Retrieving Run-Time Information**

XML-RPC Method Calls:

- *ReturnStatistics*
- *ReturnBandwidth*
- *ReturnLatency*
- ...

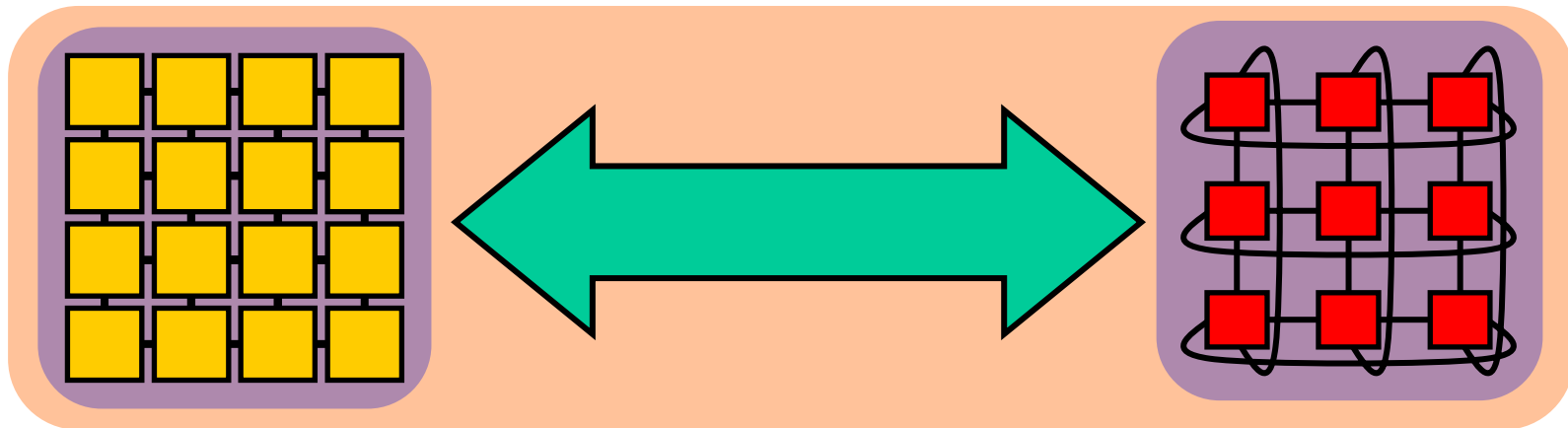
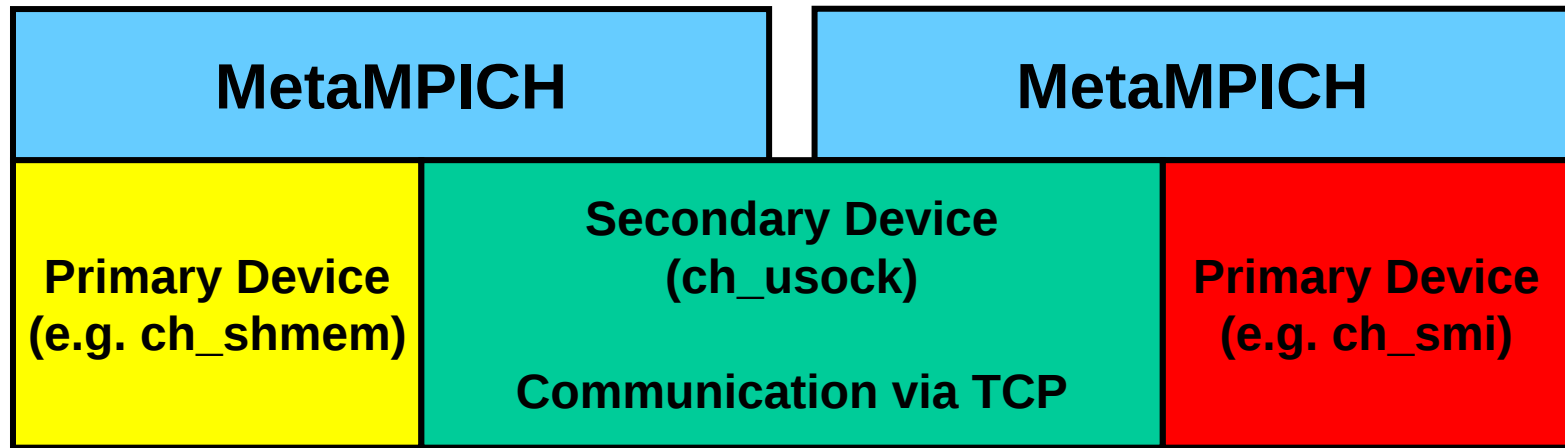
- **MetaMPICH**

- transparent MPI world on Metacomputers
- offers several adaptation/optimization features
- part of MP-MPICH

- **MP-MPICH**

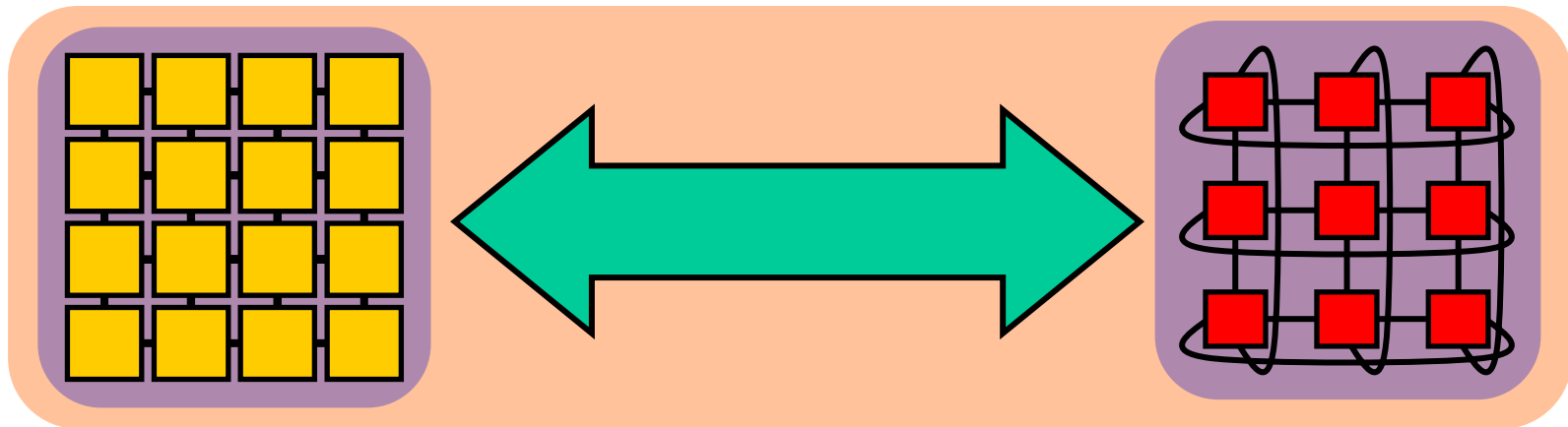
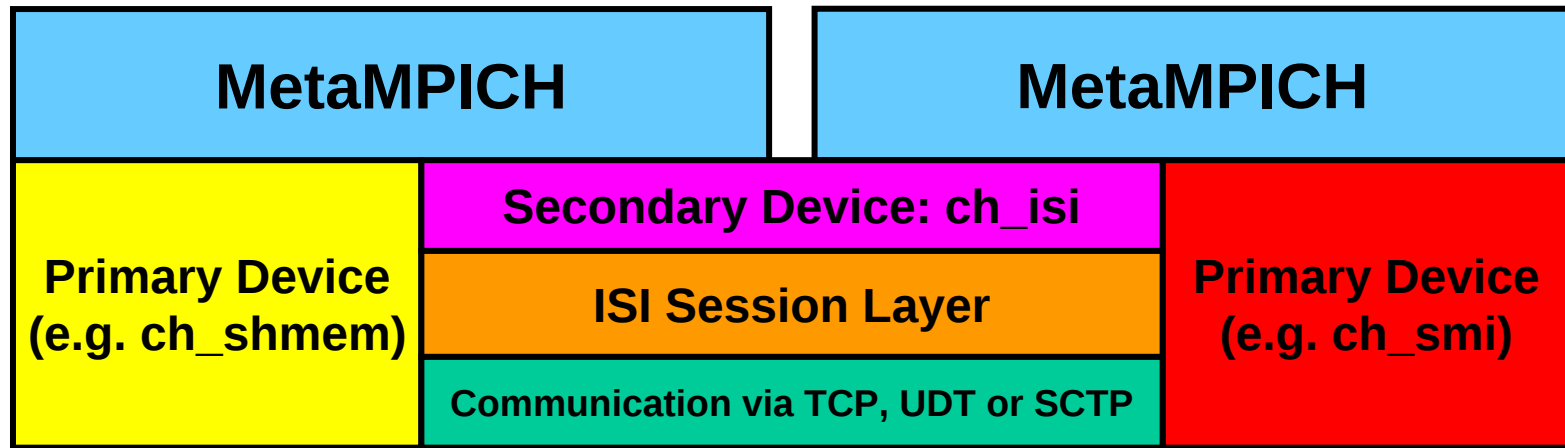
- Linux, Solaris, (Windows)
- communication devices for:
 - SCI (ch_smi)
 - Myrinet (ch_gm)
 - InfiniBand (ch_gen2)
 - Shared-Memory (ch_shmem)

- The Secondary Device / *ch_usock*



- **A New Device: *ch_isi***
 - a new **ADI2-device** for (MP-)MPICH
 - (very) **small layer** between MPI and ISI
 - supports **USI-based** communication via **TCP**, **UDT** and **SCTP**
 - offers the **ISI-based interface** for interacting with the (Grid-) environment
- *ch_isi* can now be used as a **secondary device** within MetaMPICH

- The Secondary Device / *ch_isi*



- ISI is a **session layer** for **message passing** especially via **Wide-Area-Networks**
- ISI supports **multiple transport protocols**
- ISI offers a flexible **service interface** to the **Grid environment**
- ISI has already successfully been **integrated** into **MetaMPICH**
- ISI is free and may be beneficial also for other **Grid-enabled MPI libraries**

Thank you for your attention!

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

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



LEHRSTUHL FÜR BETRIEBSSYSTEME

Univ.-Prof. Dr. habil. Thomas Bemmert



```
/home/carsten> mpirun -v -meta metaconf.cfg mlb.A.32
```

```
/home/carsten> mpirun -v -meta metaconf.cfg mlb.A.32
```

```
*** mpirun.meta: launching 32 processes on 2 metahosts: pd_cluster (16) , p4_cluster (16)
```

```
MLB: benchmark is exploring the environment...
```

```
MLB: class 0 (276.94 MB/s), group 0 ( 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 )
```

```
MLB: class 0 (276.94 MB/s), group 1 ( 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 )
```

```
MLB: class 1 (91.73 MB/s), group 2 ( 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22  
23 24 25 26 27 28 29 30 31 )
```

```
MLB: found meta groups: 0 / 1
```

```
MLB: inter-comm pair is: 0 <-> 16
```

```
MLB: benchmark is running with problem size A on 32 procs...
```

```
[1]+  Stopped          mpirun
```

```
/home/carsten>
```

```
/home/carsten>isish metaconf.domain.xml
```

```
[not attached]> Attach rank=0  
Attaching to rank 0 ... done.
```

```
[rank 0]>
```

```
/home/carsten>isish metaconf.domain.xml
```

```
[not attached]> Attach rank=0
Attaching to rank 0 ... done.
```

```
[rank 0]> ReturnMediatorList rank=16
MethodResponse from "0x8092768:-1": "ReturnMediatorList"
    "0x8097e80:0", "Mediator", "active", "TCP", "192.168.2.26:55555"
```

```
[rank 0]>
```

```
/home/carsten> isish metaconf.domain.xml
```

```
[not attached]> Attach rank=0
Attaching to rank 0 ... done.
```

```
[rank 0]> ReturnMediatorList rank=16
MethodResponse from "0x8092768:-1": "ReturnMediatorList"
    "0x8097e80:0", "Mediator", "active", "TCP", "192.168.2.26:55555"
```

```
[rank 0]> ReturnAddressList rank=16
MethodResponse from "0x8092768:-1": "ReturnAddressList",
    Address[0]: "TCP", " 192.168.2.26:55555",
    Address[1]: "SCTP", " 192.168.2.26:55560|134.130.62.26:55560",
```

```
[rank 0]>
```

```
/home/carsten>isish metaconf.domain.xml
```

```
[not attached]> Attach rank=0
Attaching to rank 0 ... done.
```

```
[rank 0]> ReturnMediatorList rank=16
MethodResponse from "0x8092768:-1": "ReturnMediatorList"
    "0x8097e80:0", "Mediator", "active", "TCP", "192.168.2.26:55555"
```

```
[rank 0]> ReturnAddressList rank=16
MethodResponse from "0x8092768:-1": "ReturnAddressList",
    Address[0]: "TCP", " 192.168.2.26:55555",
    Address[1]: "SCTP", " 192.168.2.26:55560|134.130.62.26:55560",
```

```
[rank 0]> AddNewMediator rank=16 address=1
MethodResponse from "0x8092768:-1": "AddNewMediator"
    Created new Mediators with ID "0x8097158:0"
```

```
[rank 0]>
```

```
/home/carsten>isish metaconf.domain.xml
```

```
[not attached]> Attach rank=0
Attaching to rank 0 ... done.
```

```
[rank 0]> ReturnMediatorList rank=16
MethodResponse from "0x8092768:-1": "ReturnMediatorList"
    "0x8097e80:0", "Mediator", "active", "TCP", "192.168.2.26:55555"
```

```
[rank 0]> ReturnAddressList rank=16
MethodResponse from "0x8092768:-1": "ReturnAddressList",
    Address[0]: "TCP", " 192.168.2.26:55555",
    Address[1]: "SCTP", " 192.168.2.26:55560|134.130.62.26:55560",
```

```
[rank 0]> AddNewMediator rank=16 address=1
MethodResponse from "0x8092768:-1": "AddNewMediator"
    Created new Mediators with ID "0x8097158:0"
```

```
[rank 0]> SwitchMediator rank=16 id=0x8097158:0
MethodResponse from "0x8092768:-1": "SwitchMediator"
    Successfully switched active Mediator to ID "0x8097158:0"
```

```
[rank 0]>
```

```
/home/carsten>fg 1
```

```
MLB: Speedup for transparent coupling: S = 123.799046 / 146.037569 = 0.848
```

```
MLB: Speedup with optimized algorithm: S = 123.799046 / 93.743662 = 1.321
```

```
*** mpirun.meta: Application terminated.
```

```
/home/carsten>
```