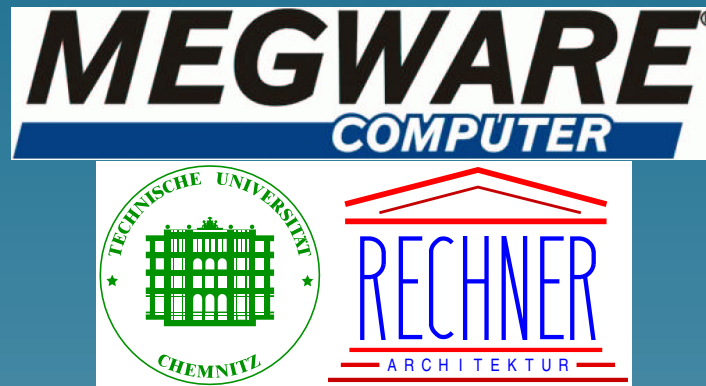


Cluster-of-Clusters-Grid — erste Erfahrungen



vertreten durch
Mario Trams

Technische Universität Chemnitz
Professur Rechnerarchitektur

ZKI – Arbeitskreis Supercomputing
18. Oktober 2002, Chemnitz

- Projektrahmen
- Projektziel
- CoC – Klassifizierung
- existierende Lösungen and Ansätze
- Schlussfolgerungen
- weitere Arbeiten

Titel:

**Entwicklung eines Cluster-of-Clusters(CoC)
basierten Leistungsserver-Prototypen**

- Verbundprojekt zwischen Megware GmbH
und TU-Chemnitz
(Fak.f.Informatik / Professur Rechnerarchitektur)
- gefördert durch SMWA
(Sächsisches Ministerium für
Wissenschaft und Arbeit)

Projektziel

Kopplung von Clustern:

- weniger im Sinne von Metacomputing, sondern im Sinne sich verändernder/wachsender Cluster

Kopplung von Clustern:

- weniger im Sinne von Metacomputing, sondern im Sinne sich verändernder/wachsender Cluster

⇒ gezielte Ausnutzung von Gegebenheiten wie ...

- ★ gemeinsame Administrationsdomäne
- ★ enge Nachbarschaft und damit Verfügbarkeit von SANs als Inter-Cluster-Connection

Kopplung von Clustern:

- weniger im Sinne von Metacomputing, sondern im Sinne sich verändernder/wachsender Cluster

⇒ gezielte Ausnutzung von Gegebenheiten wie ...

- ★ gemeinsame Administrationsdomäne
- ★ enge Nachbarschaft und damit Verfügbarkeit von SANs als Inter-Cluster-Connection

... um dadurch effizientere Lösung als Gridansätze zu bekommen

Kopplung von Clustern:

- weniger im Sinne von Metacomputing, sondern im Sinne sich verändernder/wachsender Cluster

⇒ gezielte Ausnutzung von Gegebenheiten wie ...

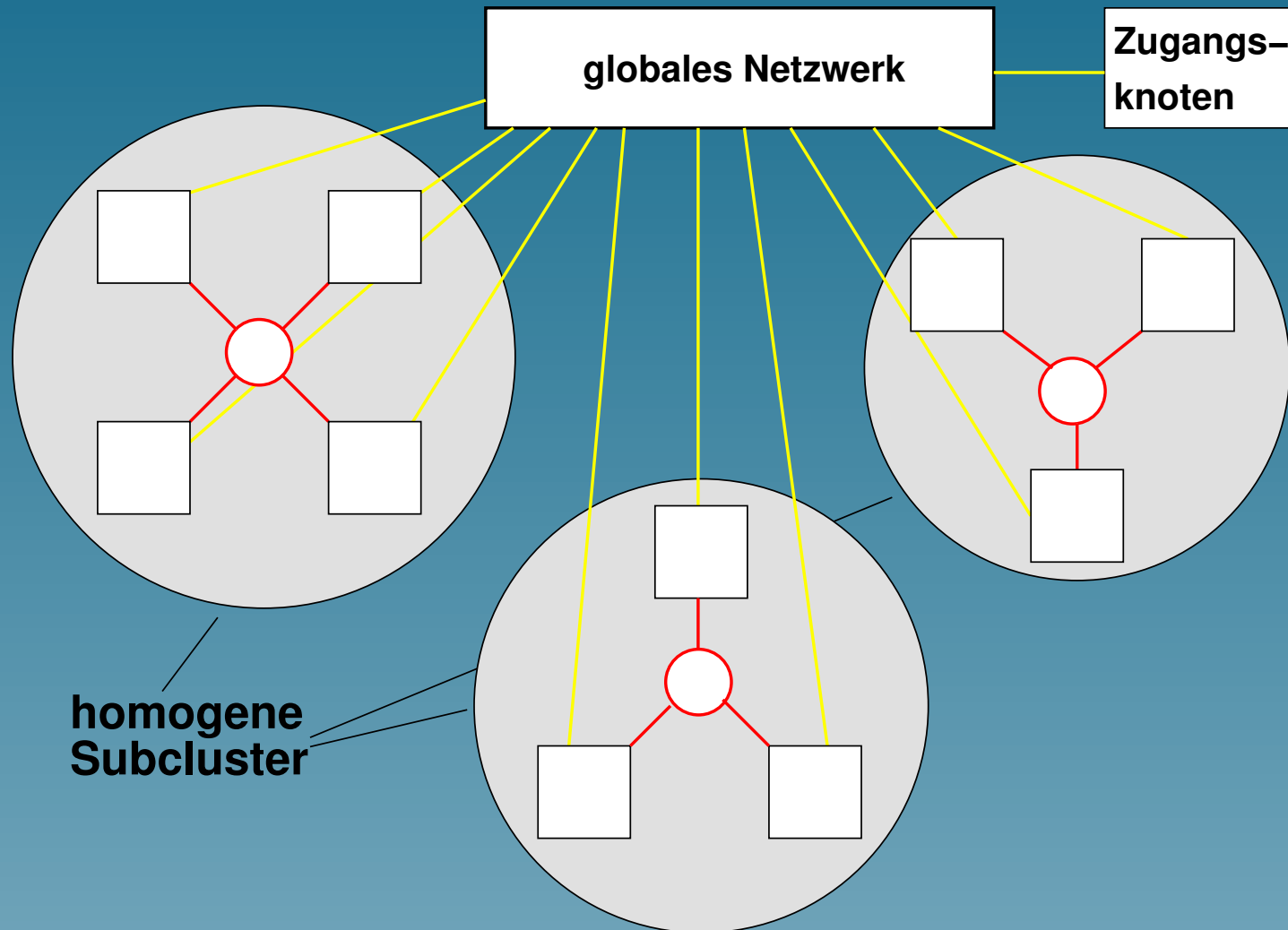
- ★ gemeinsame Administrationsdomäne
- ★ enge Nachbarschaft und damit Verfügbarkeit von SANs als Inter-Cluster-Connection

... um dadurch effizientere Lösung als Gridansätze zu bekommen

- schlussendlich das zur Verfügung stellen einer allumgreifenden MPI-Schnittstelle

CoC Klassifikation

Typ 1: global verfügbares Netzwerk



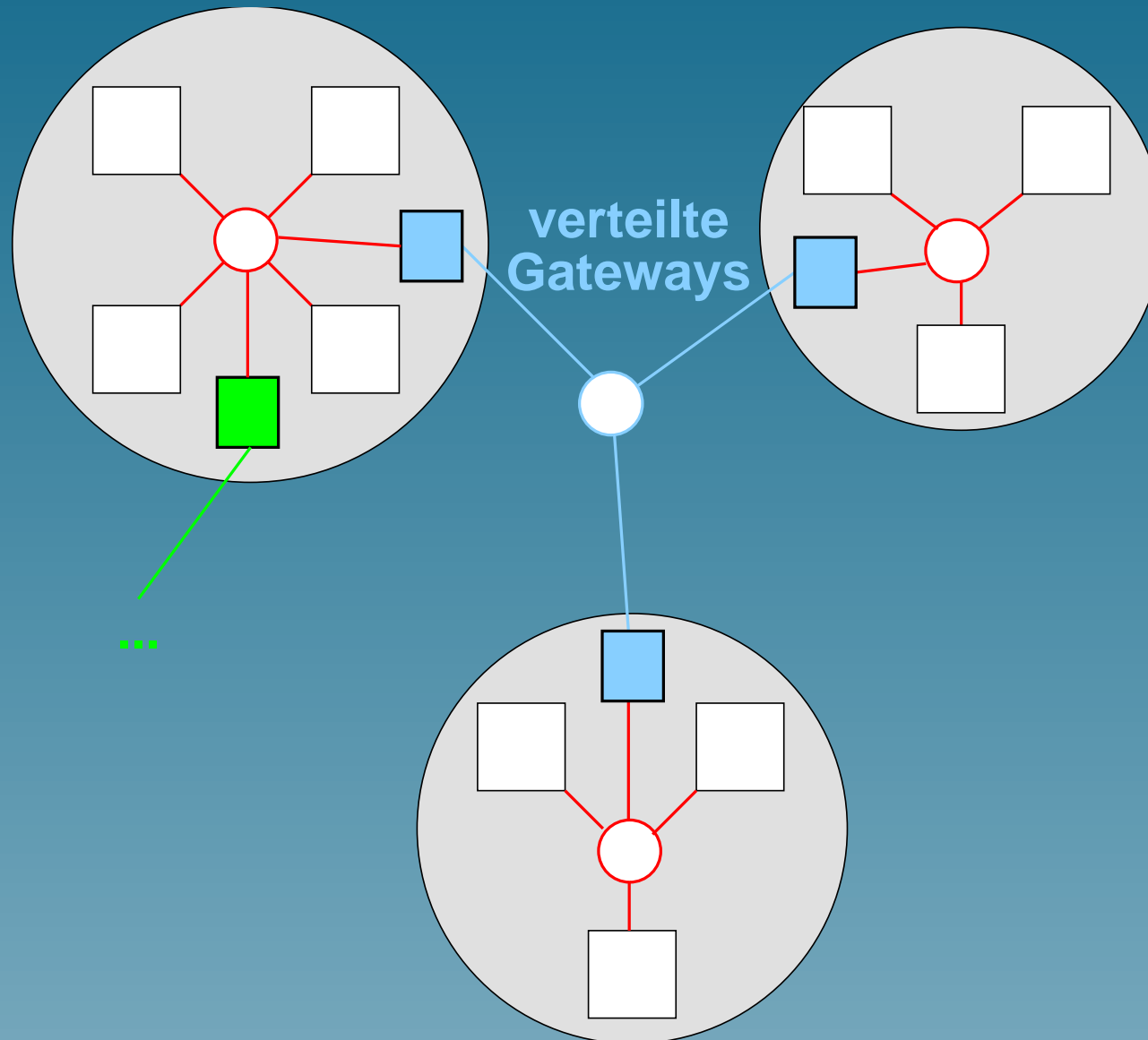
CoC Klassifikation (cont.)

Typ 1: global verfügbares Netzwerk

- ein generelles (Administrations-)Netzwerk — in der Regel Fast Ethernet;
mehrere eingebettete Subcluster mit eigenem High-Speed Netzwerk
- typischer Fall, der durch schrittweise Erweiterung eines Clusters entsteht
- direkte Peer-to-Peer Kommunikation beliebiger Knoten möglich

CoC Klassifikation (cont.)

Typ 2: Kopplung mittels verteilten Gateways



weggelassen: globales/Administrations-Netzwerk

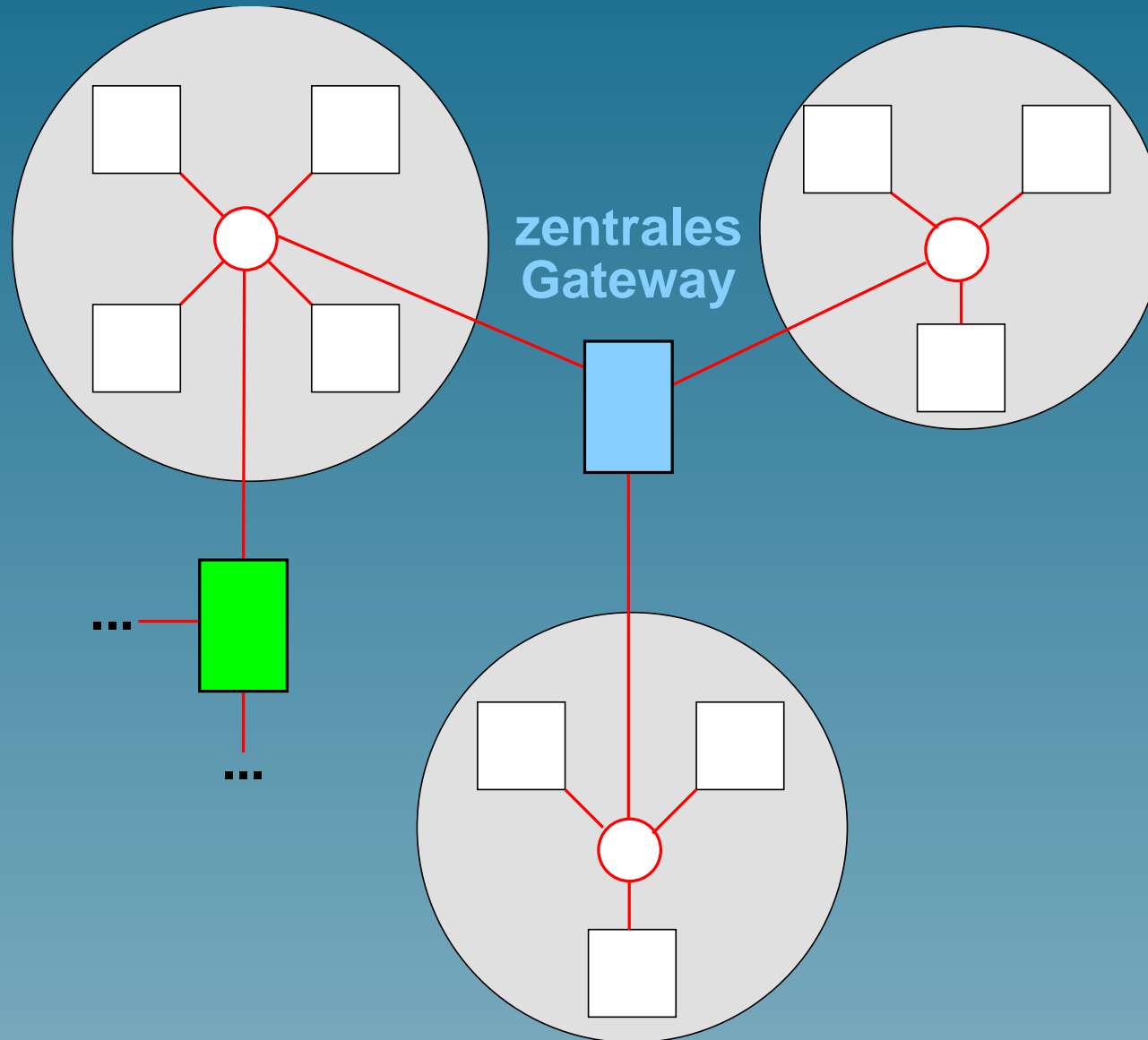
CoC Klassifikation (cont.)

Typ 2: Kopplung mittels verteilten Gateways

- ausgewählte Knoten der Subcluster haben Zugriff auf weiteres (High-Speed) Netzwerk
- lediglich diese Knoten der Subcluster können direkt miteinander kommunizieren
- keine direkte Peer-to-Peer Kommunikation beliebiger Knoten möglich
(... abgesehen von einem weiterhin existierenden, jedoch langsamen globalen Netzwerk)

CoC Klassifikation (cont.)

Typ 3: Kopplung mittels zentralem Gateway



Typ 3: Kopplung mittels zentralem Gateway

⇒ eigentlich Sonderfall von Typ 2

- denn: lediglich Umsetzung in drittes Medium entfällt bei Peer-to-Peer Kommunikation

Typ 3: Kopplung mittels zentralem Gateway

⇒ eigentlich Sonderfall von Typ 2

- denn: lediglich Umsetzung in drittes Medium entfällt bei Peer-to-Peer Kommunikation
- jedoch problematisch:
 - ★ sehr irdisch: Kabellänge der High-Speed Netze
 - ★ Ressourcenknappheit im Gateway
von Steckplätzen bis zu Bandbreitenproblemen
- erscheint dennoch interessant für kleinere Cluster

existierende (Teil-)Lösungen

High-Level Ansatz (aus Metacomputing/Grid-Bereich)

- MPICH-G2
- PACX-MPI
- METAMPICH
- IMPI

Middleware Ansatz

- Madeleine III
- VMI2

- Startup-Mechanismus auf Globus basierend
- Nutzung bestimmter MPIs für Intra-Cluster Kommunikation; MPICH-basierte MPIs jedoch momentan nicht möglich
- Inter-Cluster Kommunikation lediglich via Peer-to-Peer TCP/IP
- für CoC Typ 1 geeignet

Probleme:

- Anzahl von TCP/IP-Connections
 - ★ nicht beliebig skalierbar
 - ★ jedoch unproblematisch für kleinere Cluster

Probleme:

- Anzahl von TCP/IP-Connections
 - ★ nicht beliebig skalierbar
 - ★ jedoch unproblematisch für kleinere Cluster
- Probleme bei der Kopplung von Clustern mit privaten Subnetzen
 - ★ IP-Tunneling durch Zugangsknoten notwendig
 - ★ Abstimmung der IP-Adressen unter den Subclustern notwendig
 - ★ nicht so kritisch für CoC-Kontext

- ermöglicht ähnlich wie MPICH-G2 die Nutzung bestehender MPI-Bibliotheken für Intra-Cluster Kommunikation
 - Inter-Cluster Kommunikation via Gateway-Knoten
- ⇒ Peer-to-Peer Möglichkeit entfällt

- ermöglicht ähnlich wie MPICH-G2 die Nutzung bestehender MPI-Bibliotheken für Intra-Cluster Kommunikation

- Inter-Cluster Kommunikation via Gateway-Knoten

⇒ Peer-to-Peer Möglichkeit entfällt

- Gateways kommunizieren via TCP/IP oder ATM

⇒ leider keine Nutzung von High-Speed SANs möglich

- eignet sich für CoC Typ 1 und 2
(bedingt für Typ 3 ?)

- ähnlicher Ansatz wie PACX-MPI:
Inter-Cluster Kommunikation via Router-Knoten
- ⇒ direkte Peer-to-Peer Möglichkeit entfällt hier ebenfalls
- basiert jedoch auf MPICH
 - Inter-Cluster Kommunikation via TCP/IP oder ATM
- ⇒ wie bei PACX-MPI leider keine Nutzung von High-Speed SANs für Inter-Cluster Komm. möglich

IMPI (Interoperable MPI)

- spezielles Protokoll für die Kommunikation verschiedener MPI-Implementationen untereinander
 - prinzipiell gleicher Ansatz wie PACX-MPI/METAMPICH:
Gateways/Router; kein direktes Peer-to-Peer
 - Inter-MPI Kommunikation primär via TCP/IP
 - IMPI muss direkt in jeweiliges MPI integriert werden
- ⇒ potentiell bessere Performance als z.B. PACX-MPI
- scheint für CoC jedoch ungeeignet

Madeleine III

- Kommunikationssystem der *Parallel Multithreaded Machine (PM2)* (Projekt aus Lyon)
 - Heterogenität wird durch Low-Level Kommunikationsbibliothek verborgen
 - Anbindung an (modifiziertes) MPICH via ADI-Device
 - bisher Unterstützung für TCP/IP, SCI und Myrinet
 - eignet sich für CoC Typ 1 und 3
- ⇒ Inter-Cluster Komm. sowohl direkt Peer-to-Peer als auch über zentrales Gateway möglich
verteilte Gateways nicht möglich

VMI 2

- VMI = Virtual Machine Interface
- vergleichbar mit Madeleine
- dynamisches Nachladen von Kommunikationsmodulen
- Channel-Bundling ist möglich
- MPICH ADI-Device für VMI 2 existiert
- derzeit Unterstützung für TCP/IP, VIA, Myrinet
- eignet sich für CoC Typ 1 und 2 (VMI 1???)

High-Level Ansatz

- alle High-Level Ansätze zeichnen sich durch TCP/IP für Inter-Cluster Komm. aus (durch die Natur der Sache bedingt)

⇒ keine optimale Ausnutzung der Gegebenheiten bei CoC

High-Level Ansatz

- alle High-Level Ansätze zeichnen sich durch TCP/IP für Inter-Cluster Komm. aus (durch die Natur der Sache bedingt)

⇒ keine optimale Ausnutzung der Gegebenheiten bei CoC

Middleware Ansatz

- erscheint besser geeignet für CoC da näher zur Hardware
- bisher besteht nirgends die Möglichkeit zur Benutzung nativer MPIs

mögliche weitere Arbeiten

- Nutzung von Vendor-supplied MPIs als Kommunikationsmedium in Middleware wie Madeleine oder VMI
- Verbesserung von Madeleine oder VMI:
 - ★ mehr Low-Level Devices an sich
 - ★ CoC Typ 2 Support in Madeleine

mögliche weitere Arbeiten

- Nutzung von Vendor-supplied MPIs als Kommunikationsmedium in Middleware wie Madeleine oder VMI
 - Verbesserung von Madeleine oder VMI:
 - ★ mehr Low-Level Devices an sich
 - ★ CoC Typ 2 Support in Madeleine
 - Anpassung etwa von PACX-MPI an abstrahierten Inter-Cluster Komm.Mechanismus ist denkbar
 - ★ z.B. TCP/IP oder ATM als Medium, sondern VMI oder Madeleine
- ⇒ Entkopplung vom benutzten Medium

Diverse Links

MPICH-G2:

<http://www.hpclab.niu.edu/mpi/>

PACX-MPI:

<http://www.hlrs.de/organization/pds/projects/pacx-mpi/>

METAMPICH:

<http://www.lfbs.rwth-aachen.de/~martin/MetaMPICH/>

IMPI:

<http://impi.nist.gov>

Madeleine III:

<http://www.ens-lyon.fr/~mercierg/mpi.html>

VMI 2:

<http://vmi.ncsa.uiuc.edu/>