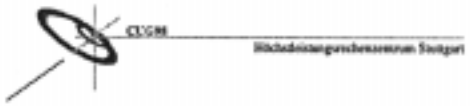
Clustering T3Es for Metacomputing Applications

Michael M. Resch, Thomas Beisel, Dirk Rantzau, Holger Berger, Katrin Bidmon, Edgar Gabriel, Rainer Keller

CUG'98, Stuttgart



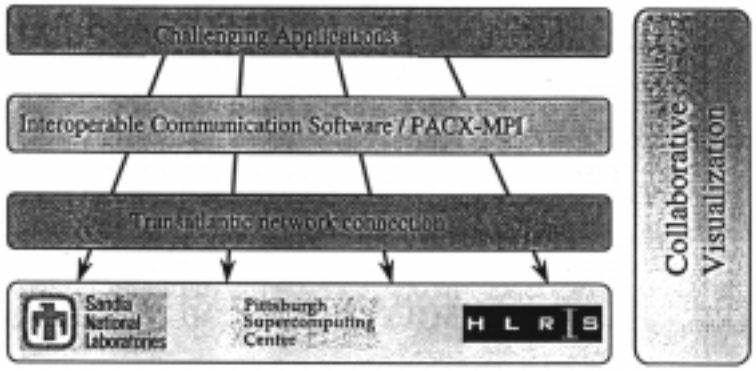
H L R S

Contents

- A Basic Concept
- Message Passing Software
- Applications
- Distributed Visualization

H L R S

G-WAAT
Global Wide Area Application Testbed



Höchstleistungsrechenzentrum Stuttgart

H L R S

Goals of PACX-MPI

- Coupling of MPPs
- No change in code
- No extensions to MPI
- Usage of vendor implemented MPI for internal communication
- Usage of standard protocol for external communication
- Implementation according to applications' needs
  no limitation in problem size or machine size

➡

Höchstleistungsrechenzentrum Stuttgart

H L R S

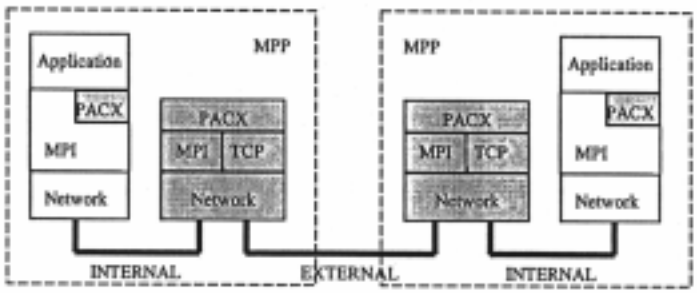
Development of PACX-MPI

- 1995:
  PACX-MPI 1.0 is developped to couple an Intel Paragon and a Cray-YMP via HiPPI.
- 1996:
  PACX-MPI 2.0 is developped to couple two machines of the same type (2 T3Es or 2 Paragons).
- 1997/1998:
  PACX-MPI 3.0 is developped to extend the number of machines involved and to allow heterogeneous clusters of MPPs.

Startup file

#machine	nodes	protocol	start-up command
host1	100	tcp	
host2	100	tcp	(rsh host2 mpirun -np 102 ./exename)
host3	100	tcp	(rsh host3 mpirun -np 102 ./exename)
host4	100	tcp	(rsh host4 mpirun -np 102 ./exename)

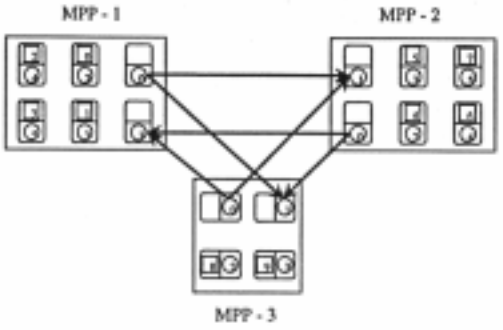
Interoperable Communication Software / PACX-MPI

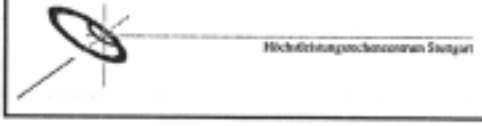


Interoperable Communication Software / PACX-MPI

H L R S

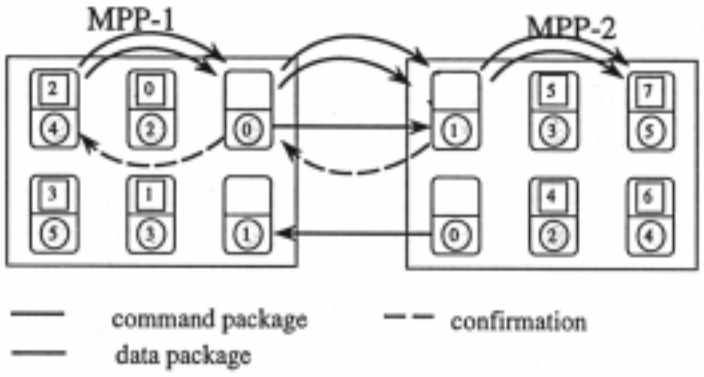
Concept of PACX-MPI 3.0





H L R S

Point-to-point communication in PACX-MPI 3.0

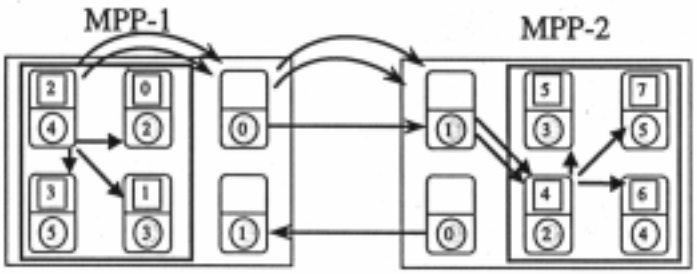


Sending a message from global node 2 to global node 7

Data conversion

- Conversion performed into the smallest supported data-format by sender and receiver
- Problem with MPI_PACK / MPI_UNPACK:
  data will be converted for internal commmunication too,
  since target unknown while packing the data.
- Data conversion only available as compile option

Global Communication in PACX-MPI 3.0



command package — local broadcast

data package

Broadcast from global node 2

MPI calls available

- Initialization and environment control:
  - MPI_Init,MPI_Finalize, MPI_Abort, MPI_Comm_rank, MPI_Comm_size
- Point-to-point communication:
  - MPI_Send, MPI_Recv, MPI_Bsend
  - MPI_Isend, MPI_Irecv
- Collective operations:
  - MPI_Barrier, MPI_Bcast, MPI_Reduce, MPI_Allreduce

Comparison with other tools I

		MPICH	PACX-MPI	PACX-MPI2	MPI-GLUE	PLUS	PVMPI
Hardware	Homogeneous Clusters	yes	yes	yes	yes	yes	yes
	Heterogeneous Clusters	yes	yes	yes	yes	yes	yes
Functionality	MPI-functions	full	full	full	full	pt2pt	pt2pt
	Optimization for Metacomputing	no	yes	yes	no	no	no
Operational	Fire wall support	no	yes	yes	no	yes	yes
	Encryption	no	yes	yes	no	yes	yes
	Compression	no	yes	yes	no	yes	yes
Applications	Homogeneous Applications	yes	yes	yes	yes	no	no
	Heterogeneous Applications	no	no	yes	no	yes	yes
	Applications available	1000s	C++	++	none	>1	>1

Höchstleistungsrechenzentrum Stuttgart

H L R S

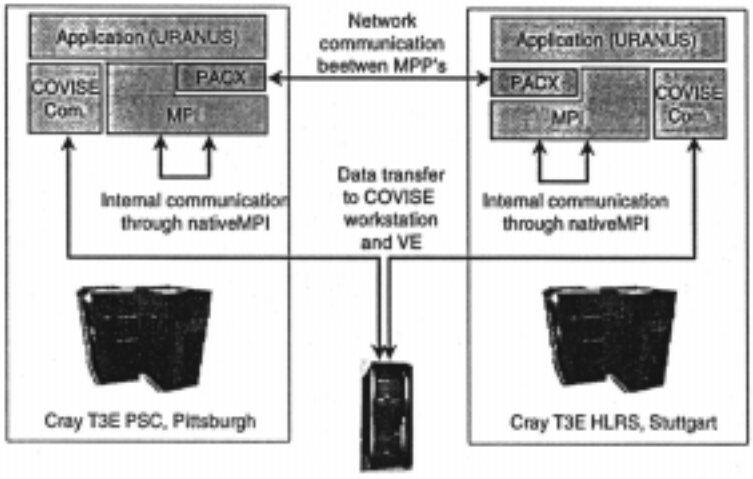
Comparison with other tools II

Tool	Application scenario
MPICH	Homogeneous programs on NOWs
PACX-MPI	Homogeneous applications on heterogeneous clusters of MPPs and PVPs
PACX-MPI2	Homogeneous and heterogeneous applications on all clusters
MPI-Glue	Research tool for homogeneous applications
PLUS	Heterogeneous applications using only send/recv on all kinds of clusters
PVMPI	Heterogeneous applications using only send/recv on all kinds of clusters

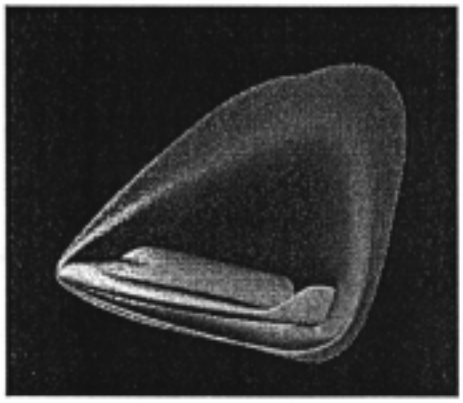
Höchstleistungsrechenzentrum Stuttgart

H L R S

Interactive Metacomputing: Software Architecture



Applications



Challenging Applications

URANUS:
Numerical Simulation of the reentry phase of a space vehicle

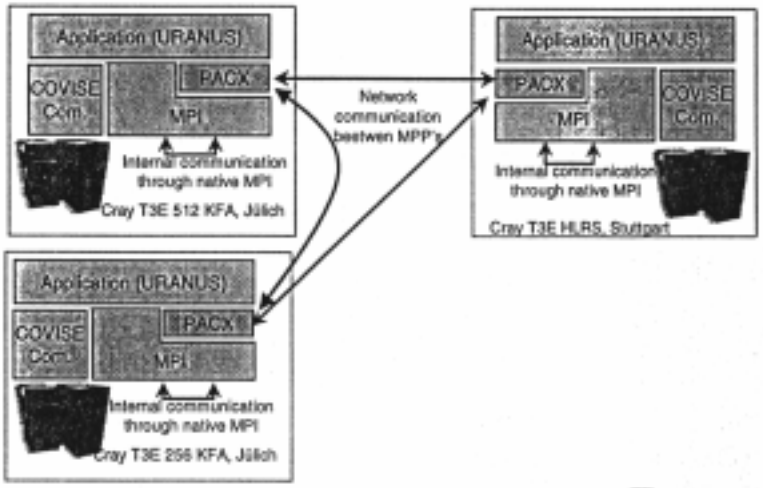
Developed by IRS
Parallelized by HLRS
Adapted for Metacomputing by HLRS

Supercomputing'97
1.7 million cells on 760 nodes

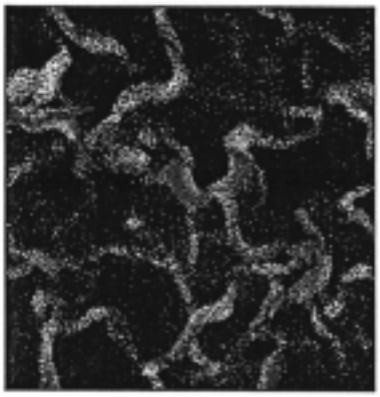
Results for URANUS

	without PACX 1 x 128 Nodes	with PACX 2 x 64 Nodes
URANUS normal	102.4 sec 272.2 sec	156.7 sec 508.5 sec
URANUS adapted	91.2 sec 269.2 sec	150.5 sec 487.6 sec
URANUS MP	-	116.7 sec 460.4 sec

Application Scenario: Metacomputing II



Applications



Challenging Applications

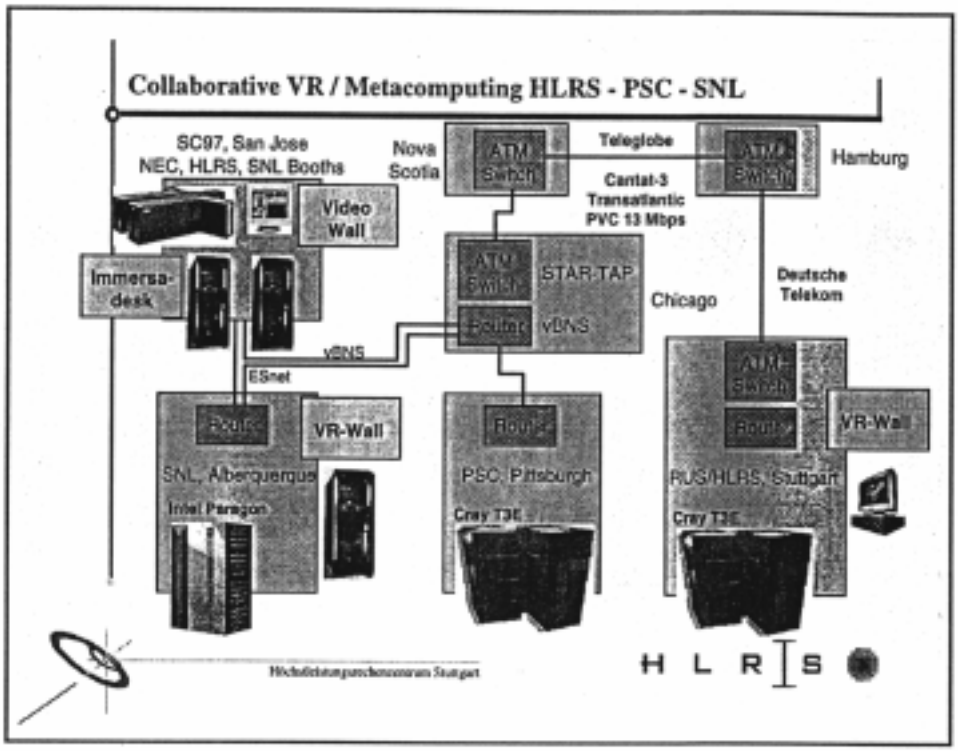
P3T-DSMC:
Direct Simulation
Monte Carlo Code

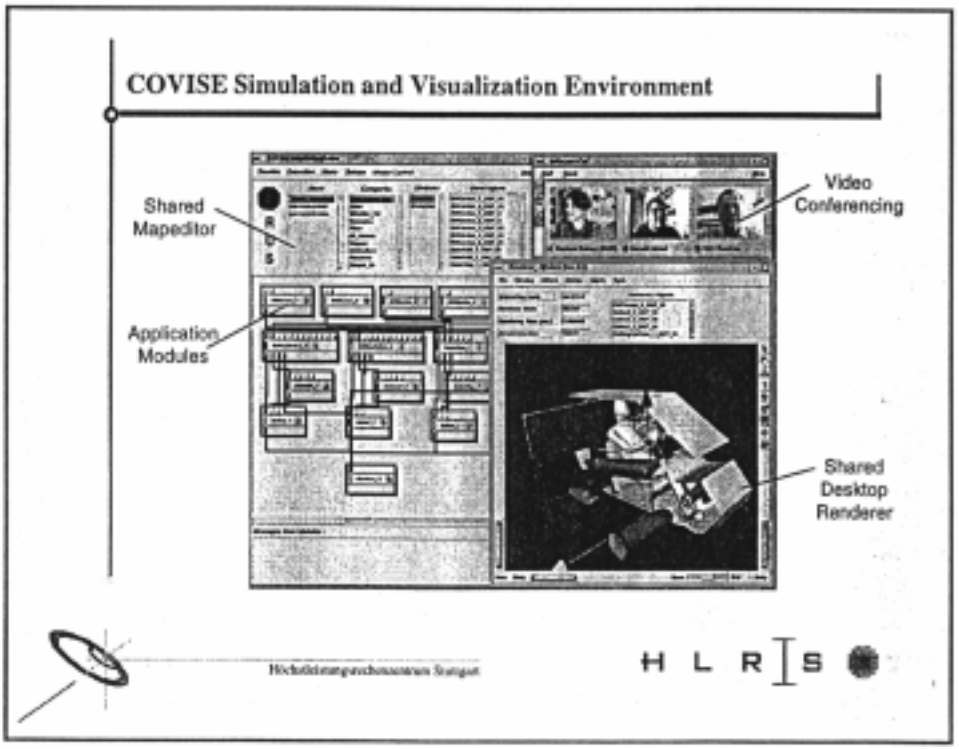
Parallelized by ICA
Adapted for
Metacomputing by
ICA/HLRS

Supercomputing'97
world record simulating
1.8 billion particles on
1024 nodes

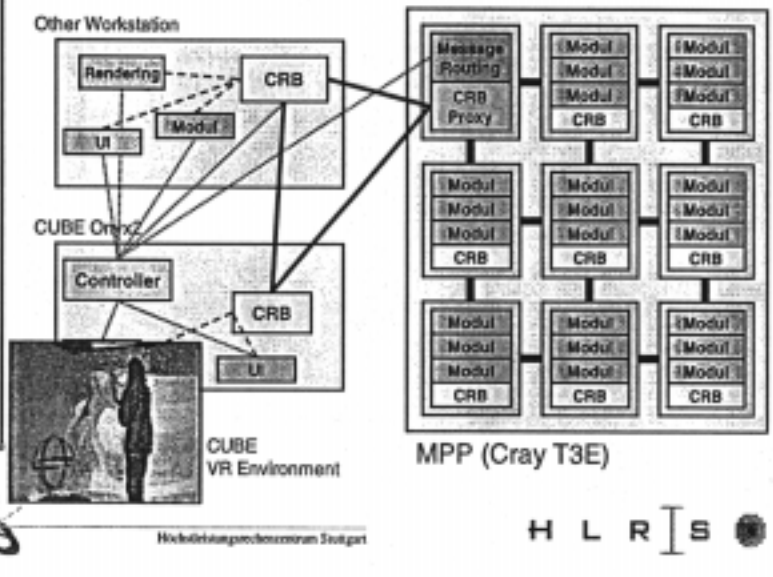
DSMC - Direct Simulation Monte Carlo

Particles/CPU	without PACX 1 x 60 Nodes	with PACX 2 x 30 Nodes
1953	0.05 sec	0.28 sec
3906	0.10 sec	0.31 sec
7812	0.20 sec	0.31 sec
15625	0.40 sec	0.40 sec
31250	0.81 sec	0.81 sec
125000	3.27 sec	3.30 sec
500000	13.04 sec	13.41 sec


Collaborative VR / Metacomputing HLRS - PSC - SNL
SC97, San Jose
NEC, HLRS, SNL Booths
Nova Scotia
ATM Switch
Teleglobe
Hamburg
Cantat-3
Transatlantic
PVC 13 Mbps
Video Wall
Immersa-desk
ATM Switch
STAR TAP
Router
vBNS
Chicago
Deutsche Telekom
ESnet
VR-Wall
SNL, Albequerque
Intel Paragon
PSC, Pittsburgh
Cray T3E
RUS/HLRS, Stuttgart
Cray T3E
H L R S


COVISE Simulation and Visualization Environment
Shared Mapeditor
Application Modules
Video Conferencing
Shared Desktop Renderer
H L R S

CoviseMP for Parallel Computing



H L R S

Conclusions

- Clustering of T3Es is possible
- We can solve bigger problems
- Monte Carlo methods are much better suited for metacomputing
- We need to improve coupled distributed visualization

H L R S