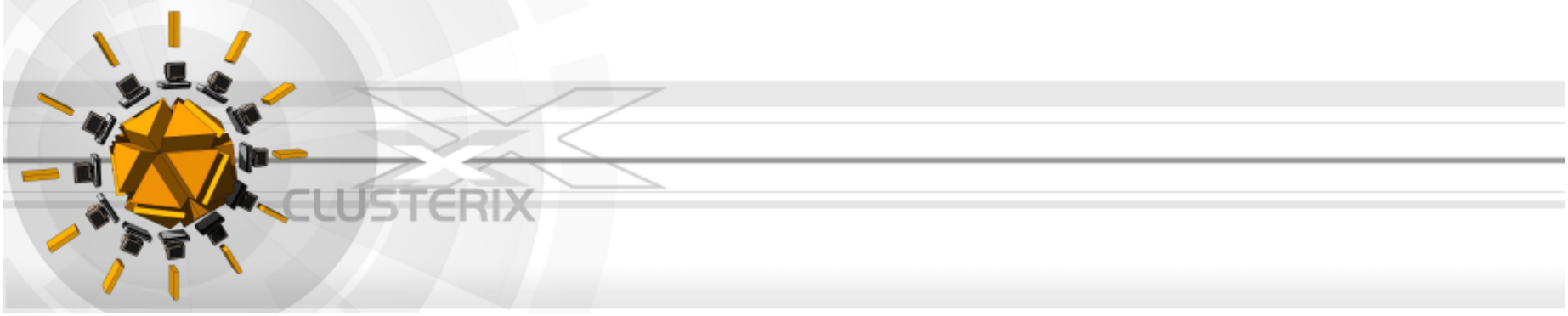


Using CLUSTERIX - National CLUSTER of LInuX Systems:

Roman Wyrzykowski*, Norbert Meyer**

*Czestochowa University of Technology
**Poznań Supercomputing and Networking Center

Outline

- CLUSTERIX status, goals and architecture
- Pilot installation & network infrastructure
- CLUSTERIX middleware
  - Technologies and architecture
  - Dymamic cluster attachment
  - User account management
- Pilot applications
- An example of running applications in CLUSTERIX
- Final remarks

Current Status

- project started on January 2004
- the entire project lasts 32 months with two stages:
- - research and development – finished in Sept. 2005
- - deployment – starting in Oct. 2005, till June 2006
- 12 members – Polish supercomputing centers and MANs
- total budget – 1,2 milion Euros
- 53 % funded by the consortium members, and 47 % - by the Polish Ministry of Science and Information Society Technologies

Partners

- **Częstochowa University of Technology (coordinator)**
- Poznań Supercomputing and Networking Center (PNSC)
- Academic Computing Center CYFRONET AGH, Kraków
- Academic Computing Center in Gdańsk (TASK)
- Wrocław Supercomputing and Networking Center (WCSS)
- Technical University of Białystok
- Technical University of Łódź
- Marie Curie-Skłodowska University in Lublin
- Warsaw University of Technology
- Technical University of Szczecin
- Opole University
- University of Zielona Góra

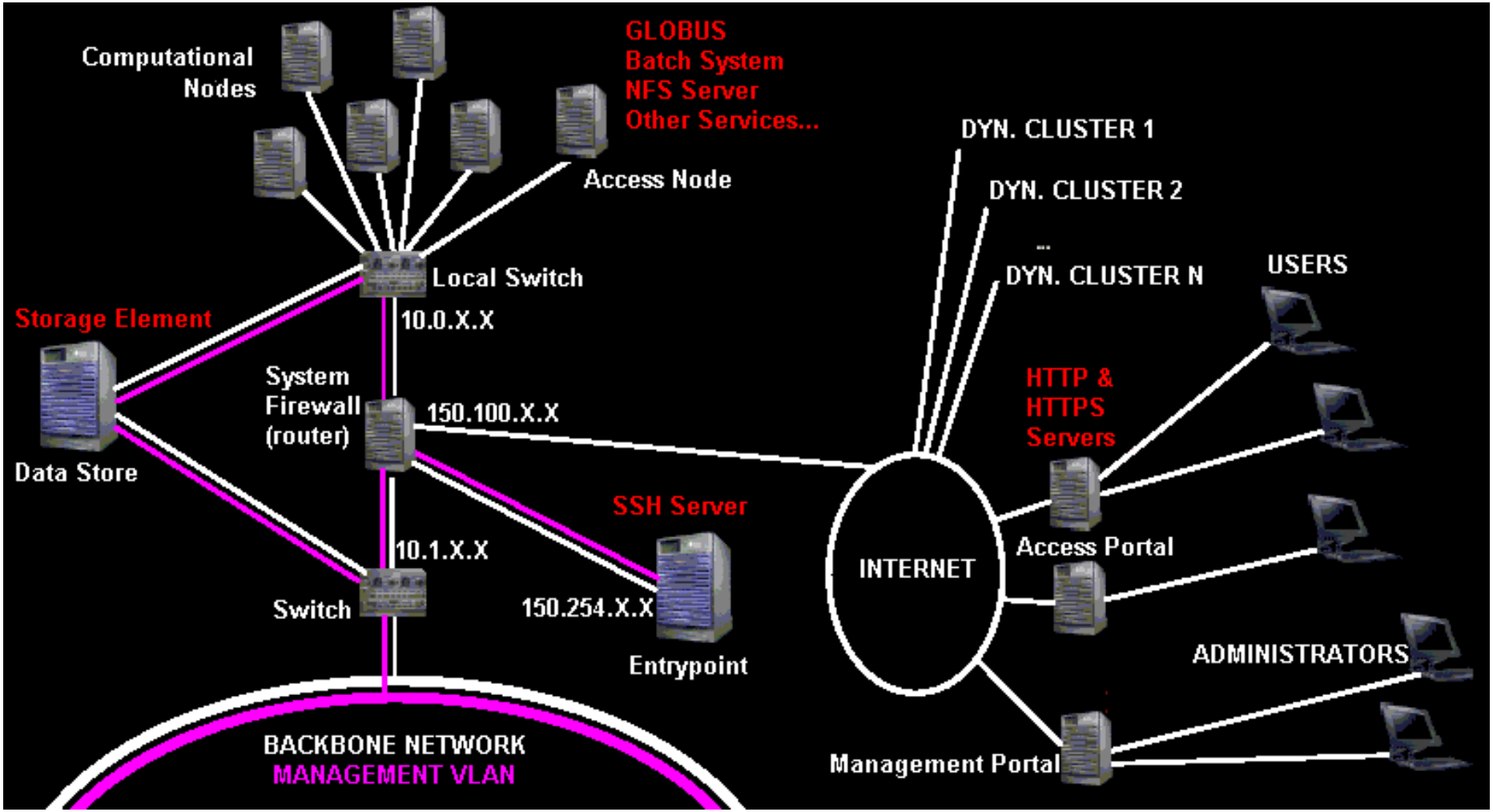
CLUSTERIX Goals

- to develop mechanisms and tools that allow the deployment of a **production Grid environment**
- basic infrastructure consists of local LINUX clusters with 64-bit architecture located in geographically distant independent centers connected by the fast backbone provided by the Polish Optical Network PIONIER
- existing and newly built LINUX clusters are dynamically connected to the basic infrastructure
- as a result, a **distributed PC-cluster** is built, with a dynamically changing size, fully operational and integrated with services delivered as the outcome of other projects

Added Values

- development of software capable of cluster management with dynamically changing configuration (nodes, users and available services); one of the most important factors is reducing the management overhead
- taking into consideration local policies of infrastructure administration and management, within independent domains
- new quality of services and applications based on the IPv6 protocols
- integration and making use of the existing services delivered as the outcome of other projects (data warehouse, remote visualization, …)
- integrated end-user/administrator interface
- providing required security in a heterogeneous distributed system
- production-class Grid infrastructure
- the resulting system tested on a set of pilot distributed applications developed as a part of the project

CLUSTERIX
Architecture

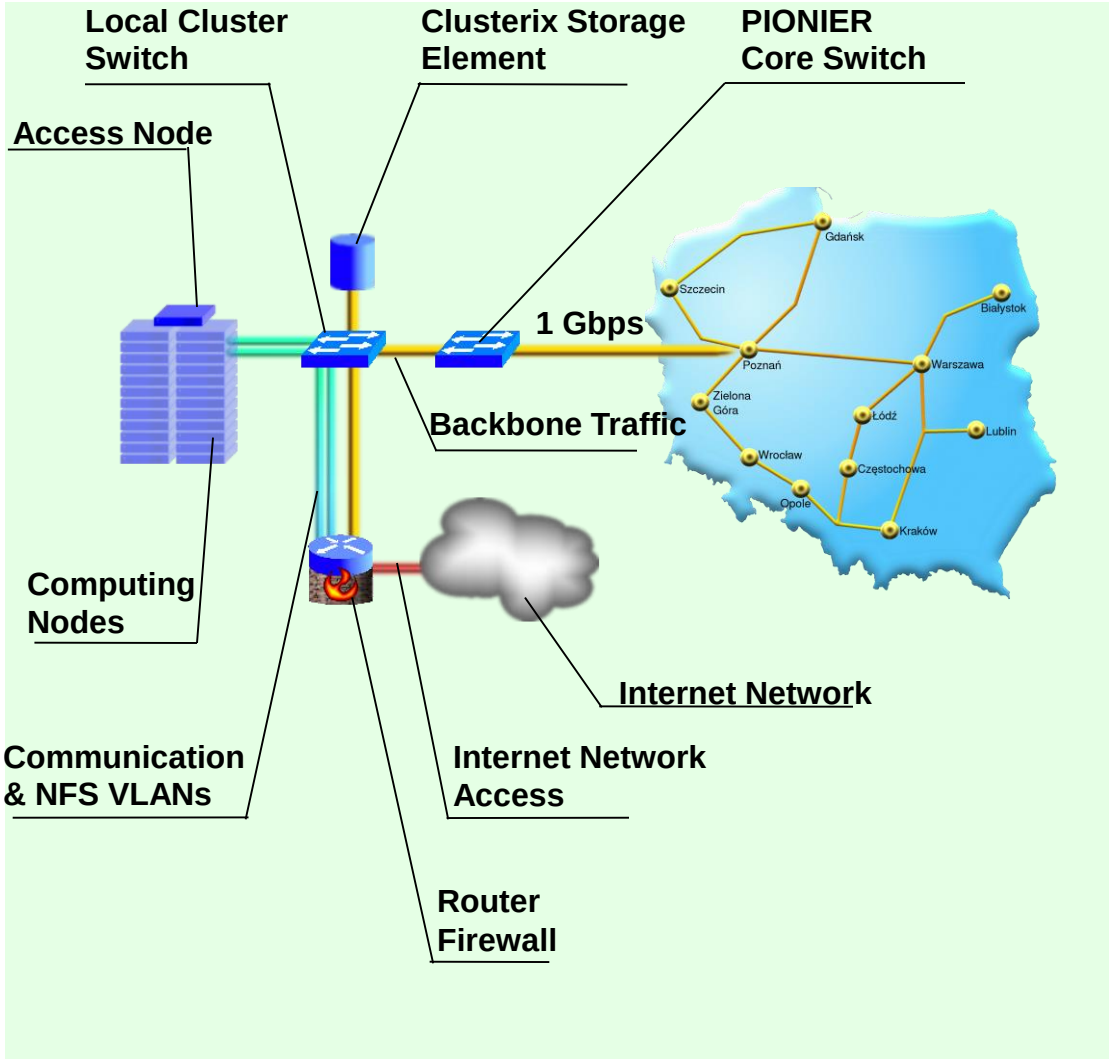


Pilot Installation

- 12 local clusters with 250+ IA-64 in the core
- Linux Debian, kernel 2.6.x
- PIONIER Network: 3000+ km of fibers with 10Gbps DWDM technology
- 2 VLANs with dedicated 1Gbps bandwidth for the CLUSTERIX network

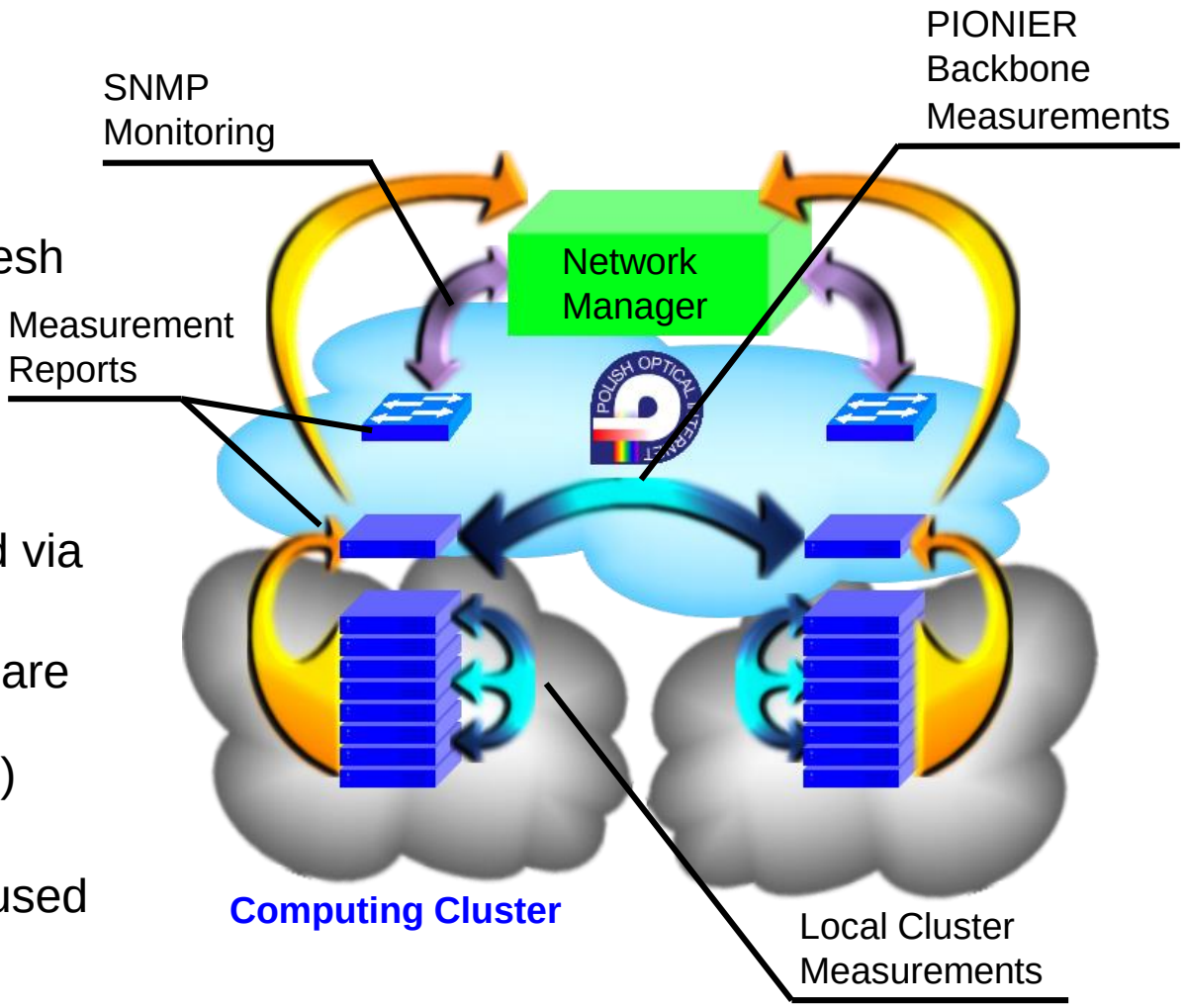
CLUSTERIX Network Architecture

- Communication to all cluster is passed through router/firewall
- Routing based on IPv6 protocol, with IPv4 for back compatibility feature
- Application and Clusterix middleware are adjusted to IPv6 usage
- Two 1 Gbps VLANs are used to improve management of network traffic in local clusters
  - Communication VLAN is dedicated to support nodes messages exchange
  - NFS VLAN is dedicated to support file transfer



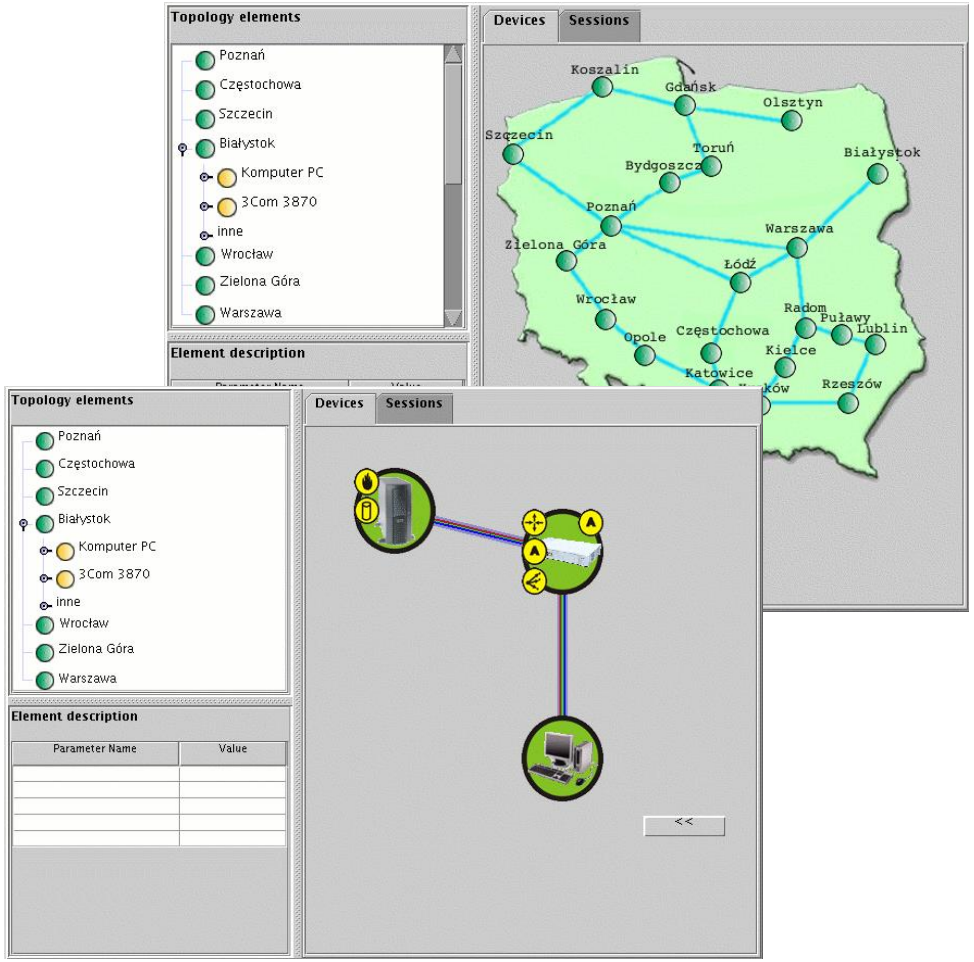
Active Network Monitoring

- Measurement architecture
  - Distributed 2-level measurement agent mesh (backbone/cluster)
  - Centralized control manager (multiple redundant instances)
  - Switches are monitored via SNMP
  - Measurements reports are stored by manager (forwarded to database)
  - IPv6 protocol and addressing schema is used for measurement



Graphical User Interface

- GUI
  - Provides view of network status
  - Gives a look at statistics
  - Simplifies network troubleshooting
  - Allows to configure measurement sessions
  - Useful for topology browsing



Middleware in CLUSTERIX

- the software developed is based on Globus Toolkit 2.4 plus web services - with Globus 2.4 available in Globus 3.2 distribution

  - this makes the created software easier to reuse

  - allows for interoperability with other Grid systems on the service level

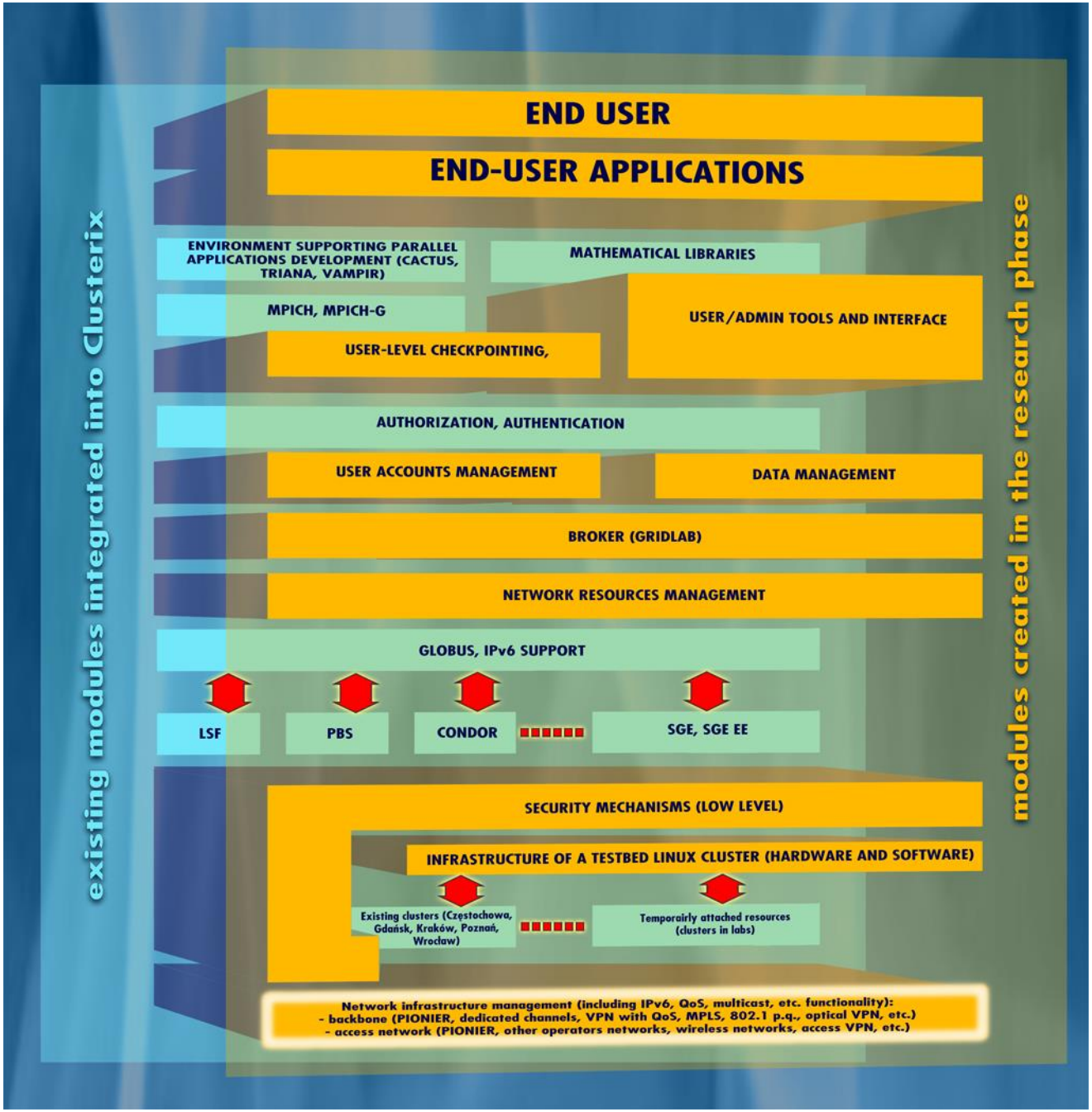
- *Open Source* technology, including LINUX (Debian, kernel 2.6.x) and batch systems (Open PBS/Torque, SGE)

  - open software is easier to integrate with existing and new products

  - allows anybody to access the project source code, modify it and publish the changes

  - makes the software more reliable and secure

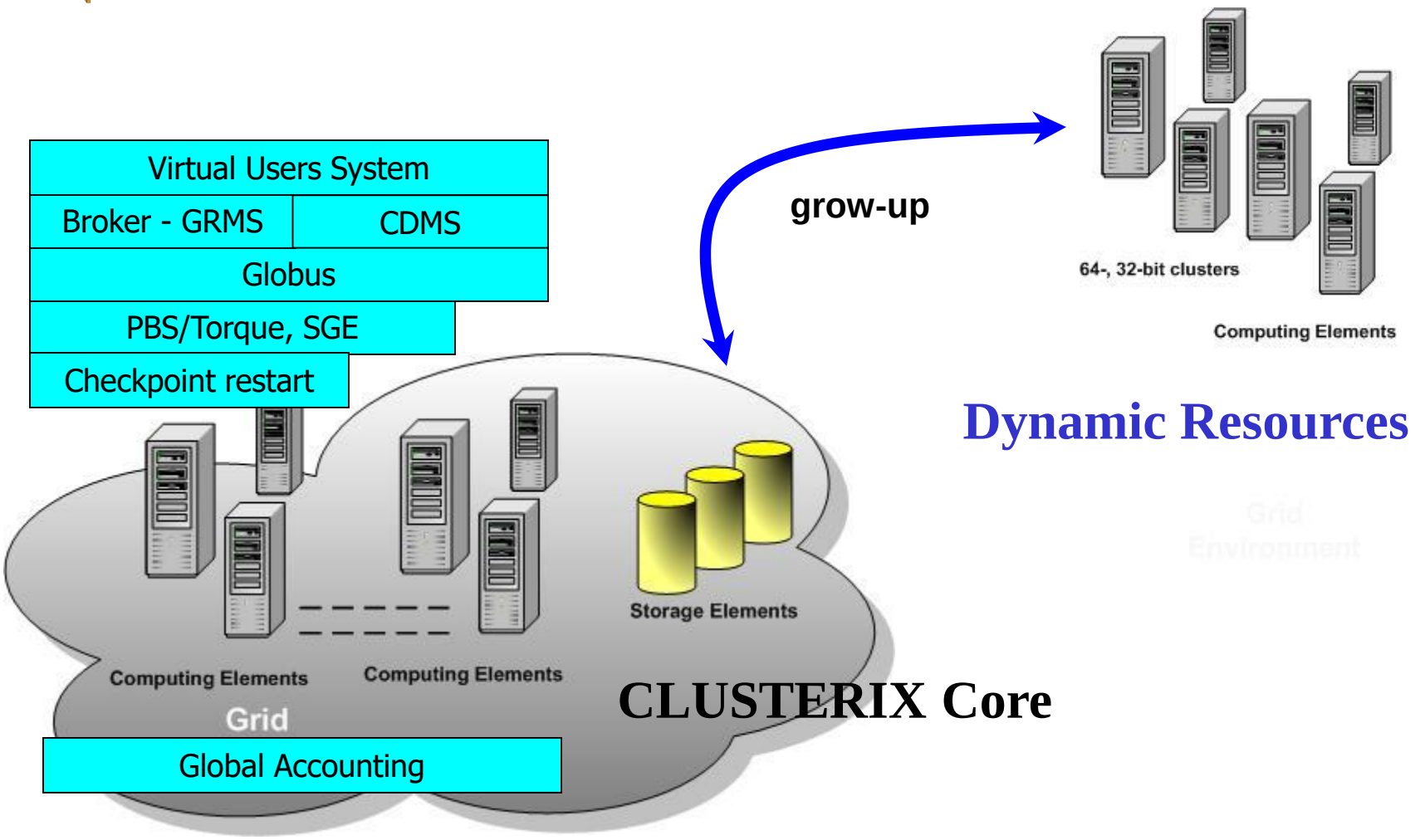
- existing software will be used extensively in the CLUSTERIX project, e.g., *GridLab* broker, Virtual User Account (SGIgrid)


existing modules integrated into Clusterix
END USER
END-USER APPLICATIONS
ENVIRONMENT SUPPORTING PARALLEL APPLICATIONS DEVELOPMENT (CACTUS, TRIANA, VAMPIR)
MATHEMATICAL LIBRARIES
MPICH, MPICH-G
USER/ADMIN TOOLS AND INTERFACE
USER-LEVEL CHECKPOINTING,
AUTHORIZATION, AUTHENTICATION
USER ACCOUNTS MANAGEMENT
DATA MANAGEMENT
BROKER (GRIDLAB)
NETWORK RESOURCES MANAGEMENT
GLOBUS, IPv6 SUPPORT
LSF
PBS
CONDOR
SGE, SGE EE
SECURITY MECHANISMS (LOW LEVEL)
INFRASTRUCTURE OF A TESTBED LINUX CLUSTER (HARDWARE AND SOFTWARE)
Existing clusters (Częstochowa, Gdańsk, Kraków, Poznań, Wrocław)
Temporarily attached resources (clusters in labs)
Network infrastructure management (including IPv6, QoS, multicast, etc. functionality):
- backbone (PIONIER, dedicated channels, VPN with QoS, MPLS, 802.1 p.q., optical VPN, etc.)
- access network (PIONIER, other operators networks, wireless networks, access VPN, etc.)
modules created in the research phase

Dynamic Clusters Attachment: Goals

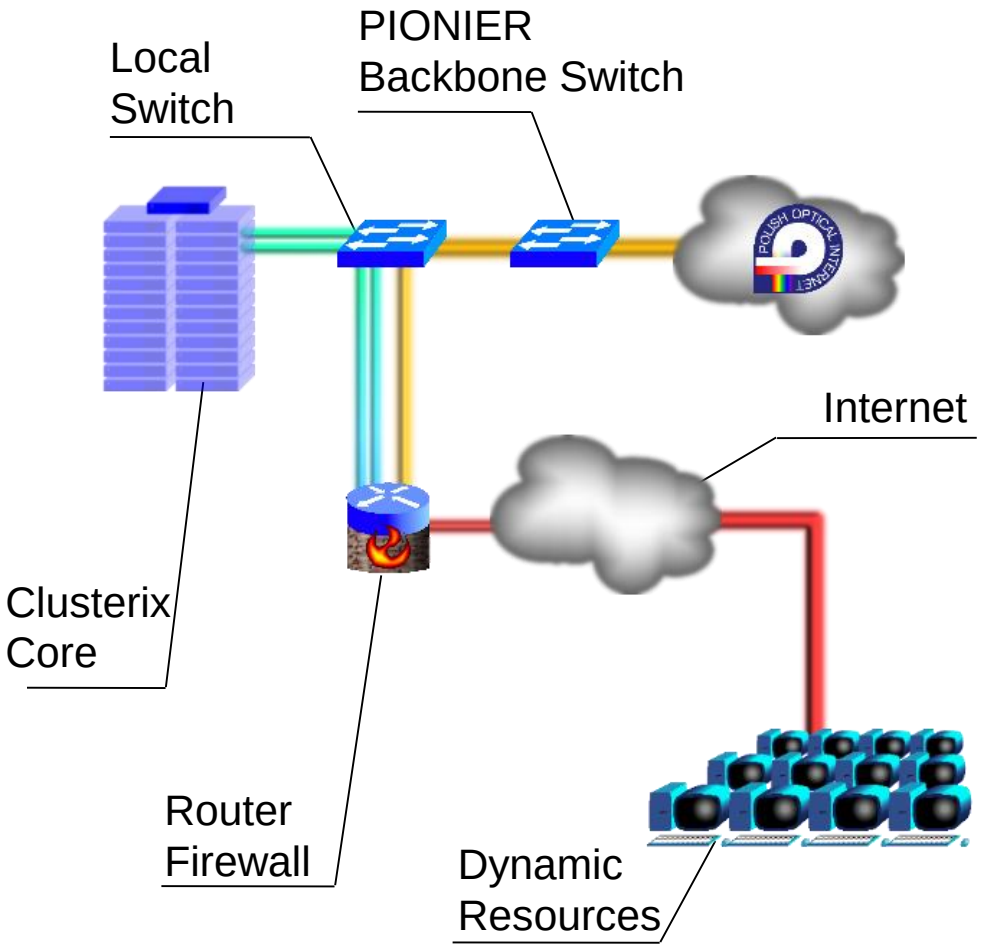
- Dynamic (external) clusters can be easily attached to CLUSTERIX core in order to:
  - Increase computing power with new clusters
  - Utilize external clusters during nights or non-active periods
  - Make CLUSTERIX infrastructure scalable

Integrating Dynamic Clusters

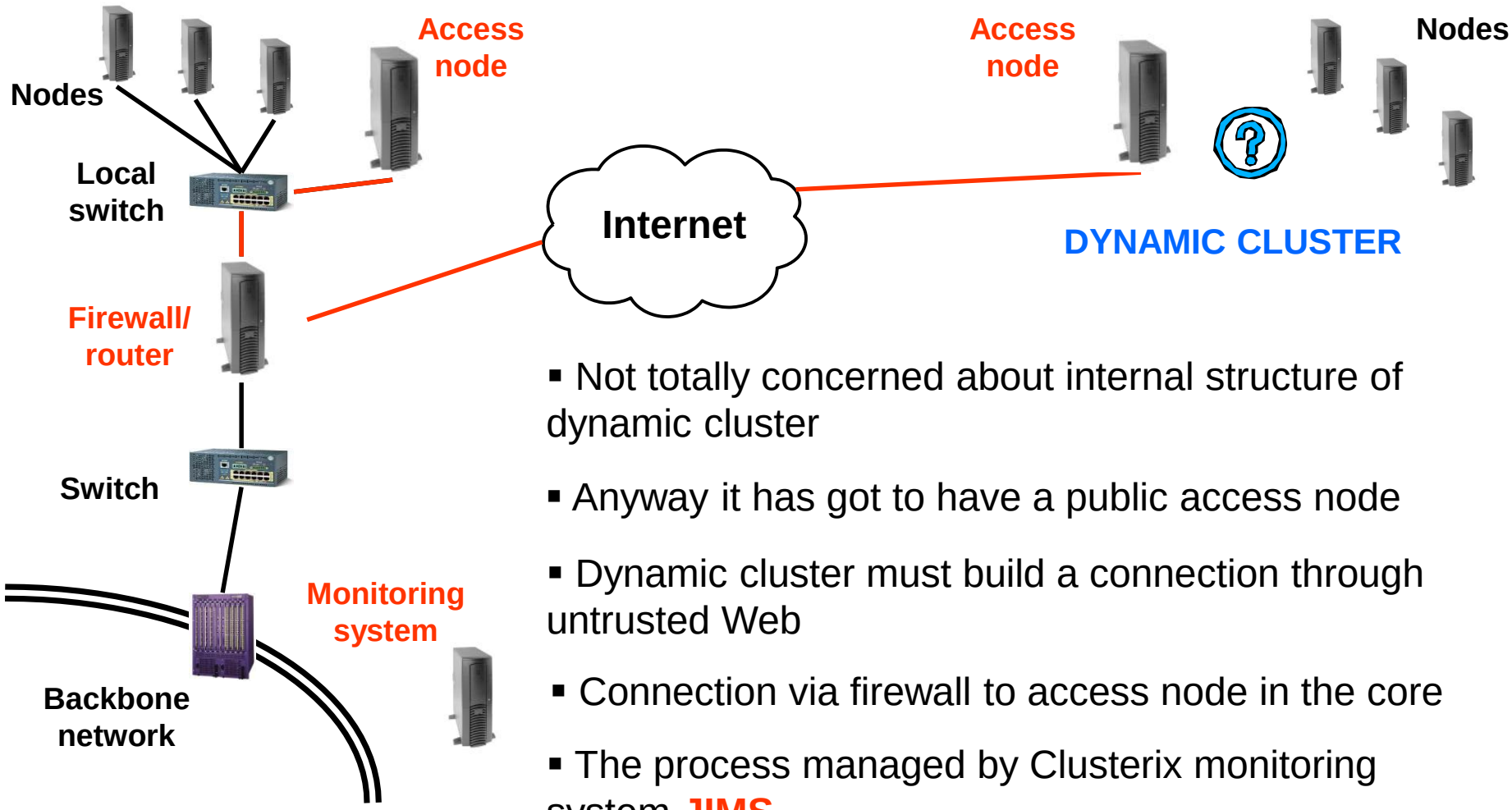


Dynamic Cluster Attachment: Architecture

- Requirements needs to be checked against new clusters
  - Installed software
  - SSL certificates
- Communication through router/firewall
- Monitoring System will automatically discover new resources
- New clusters serve computing power on regular basis



Connecting Dynamic Cluster

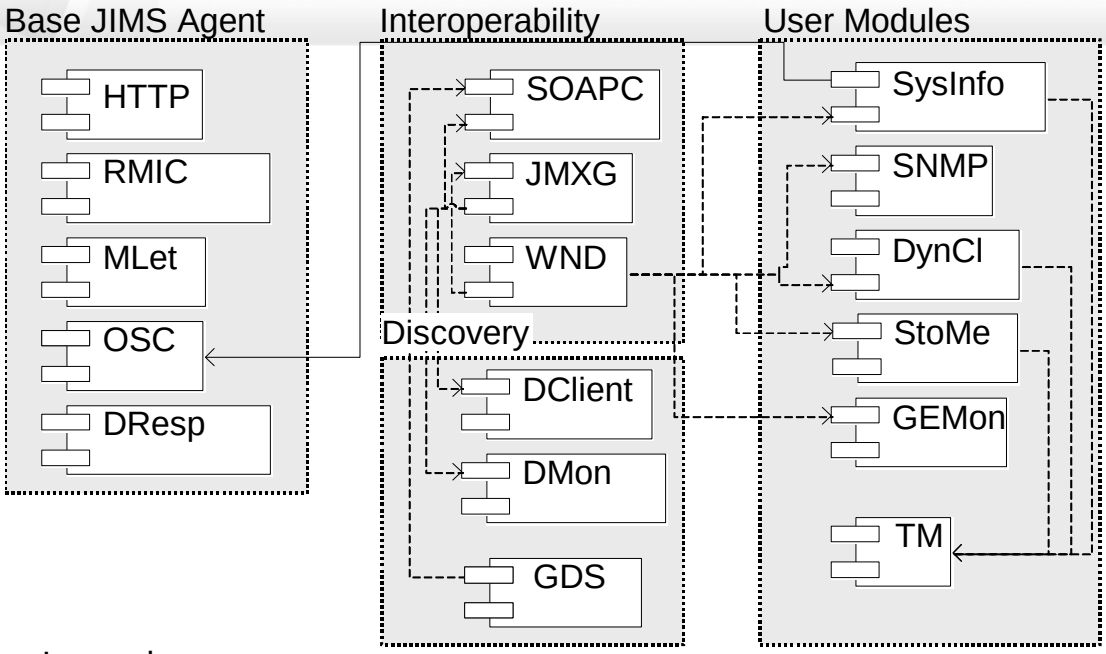


- Not totally concerned about internal structure of dynamic cluster
- Anyway it has got to have a public access node
- Dynamic cluster must build a connection through untrusted Web
- Connection via firewall to access node in the core
- The process managed by Clusterix monitoring system **JIMS**

JIMS & Dynamic Clusters

- **JIMS** - the JMX-based Infrastructure Monitoring System
- Additional module implementing functionality necessary for supporting Dynamic Clusters
- Support for Dynamic Cluster installation through Web Service, with secure transmission and authentication (SSL/GSI)
- Support for Broker notification about following events:
  - Dynamic Cluster added
  - Dynamic Cluster removed
- System managed through JIMS Manager (GUI), by administrator or automatically, using dedicated service in Dynamic Cluster

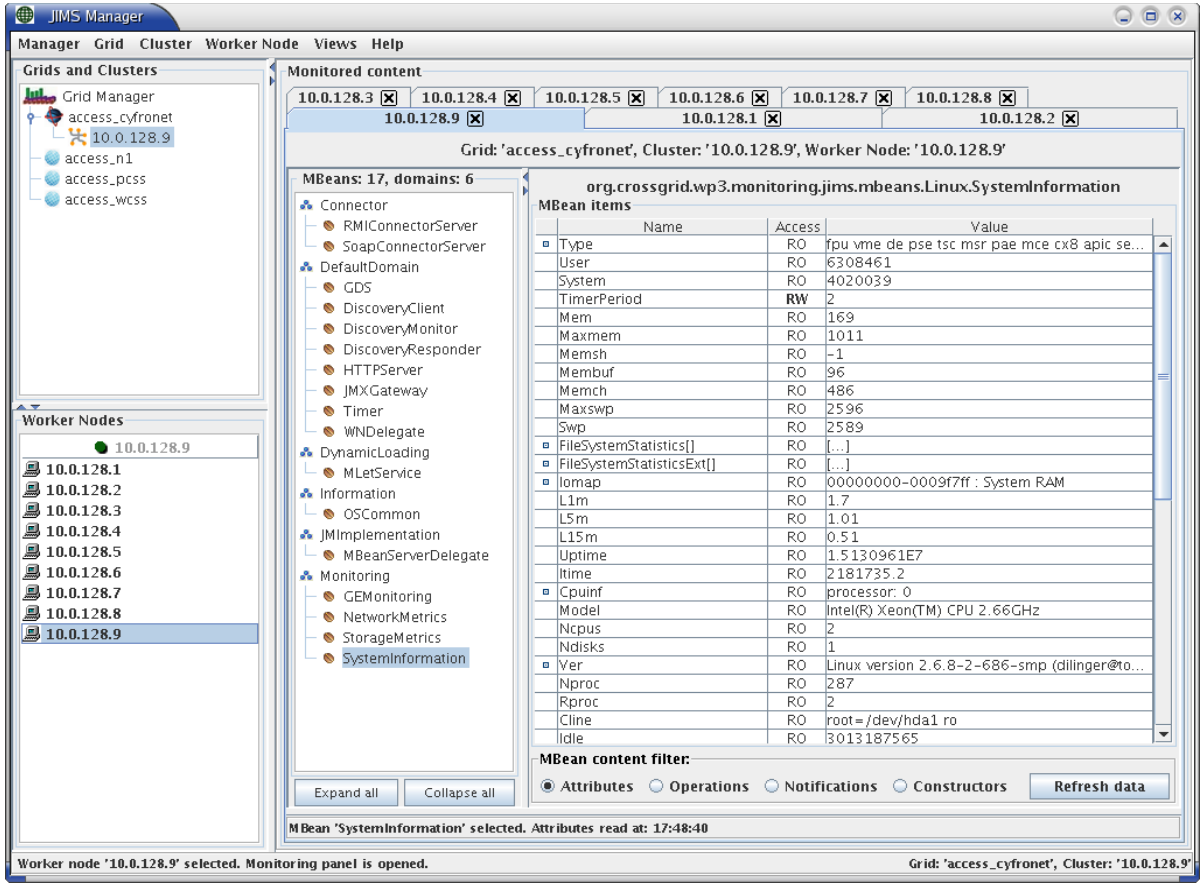
JIMS Dynamic Cluster Module



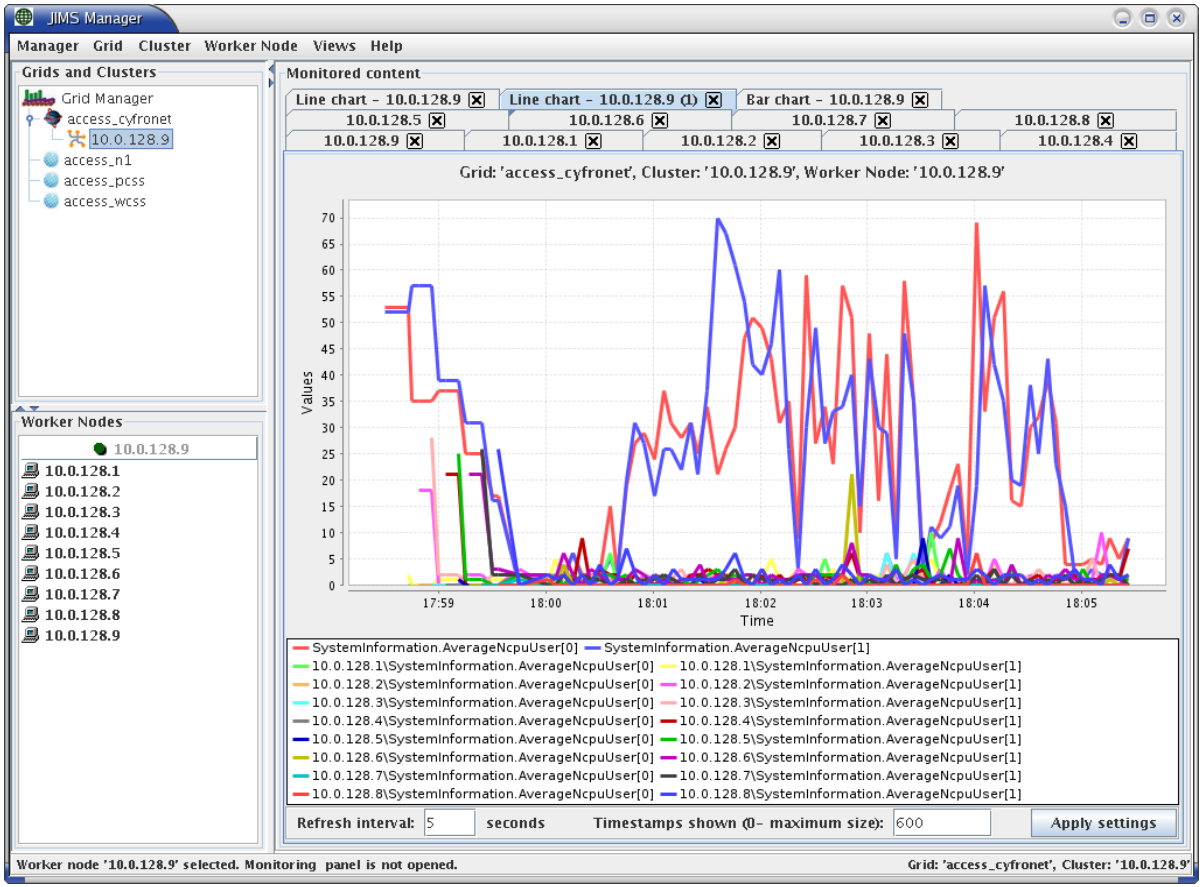
Legend:

HTTP	- HTTP Server	DClient	- Discovery Client
RMIC	- RMI Connector	DMon	- Discovery Monitor
MLet	- MLet Service	GDS	- Global Discovery Service
OSC	- Operating System Common	SysInfo	- System Information
		SNMP	- SNMP Proxy
DResp	- Discovery Responder	DynCl	- Dynamic Cluster Module
SOAPC	- SOAP Connector	StoMe	- Storage Metrics
JMXG	- JMX Gateway	GEMon	- Grid Engine Monitoring
WND	- Worker Node Delegate	TM	- JMX Timer

JIMS Management Application (1)



JIMS Management Application (2)



Growing-up

- core installation:

  - 250+ Itanium2 CPUs distributed among 12 sites located across Poland

- ability to connect dynamic clusters from anywhere (clusters from campuses and universities)

  - peak installation with 800+ CPUs (4,5 Tflops) – not automatic procedure yet

User Account Management:

Problems with Wide Area Cluster

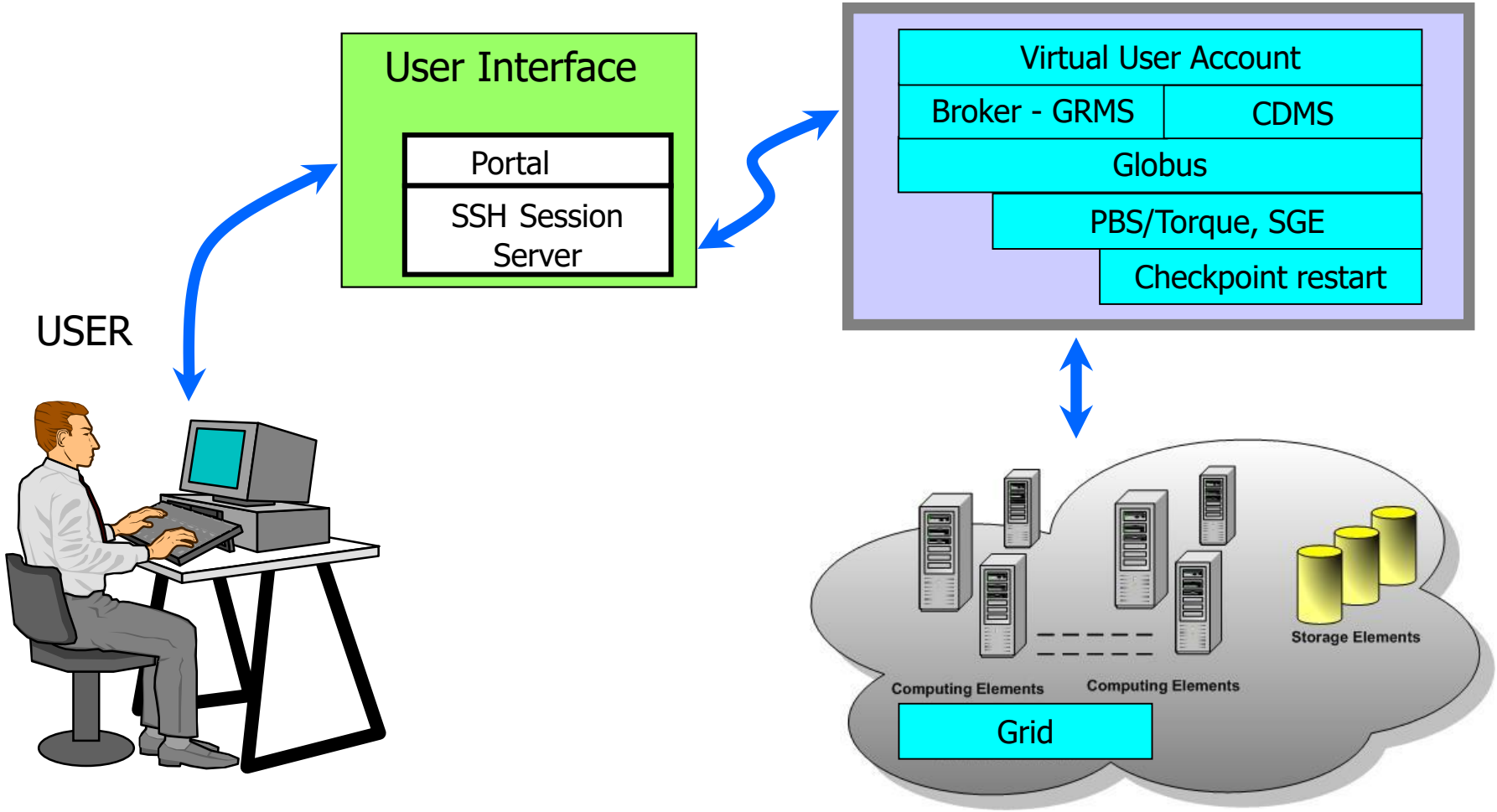
Local cluster ≠ wide area cluster !!!

The main (from our perspective) problems are:

- User accounts problems
- Global resource accounting
- Queuing system incompatibility
- File transfer problems
- ....

The integration of the person into the system in a seamless and comfortable way is paramount to obtain maximal benefit.

Task execution in CLUSTERIX



User Account Management: Requirements

User

To be able to submit jobs to remote resource
To have summary information about resources used

Admin

To have full knowledge of who is using resources

VO (Virtual Organization) manager

To have summary information about its users

Site manager

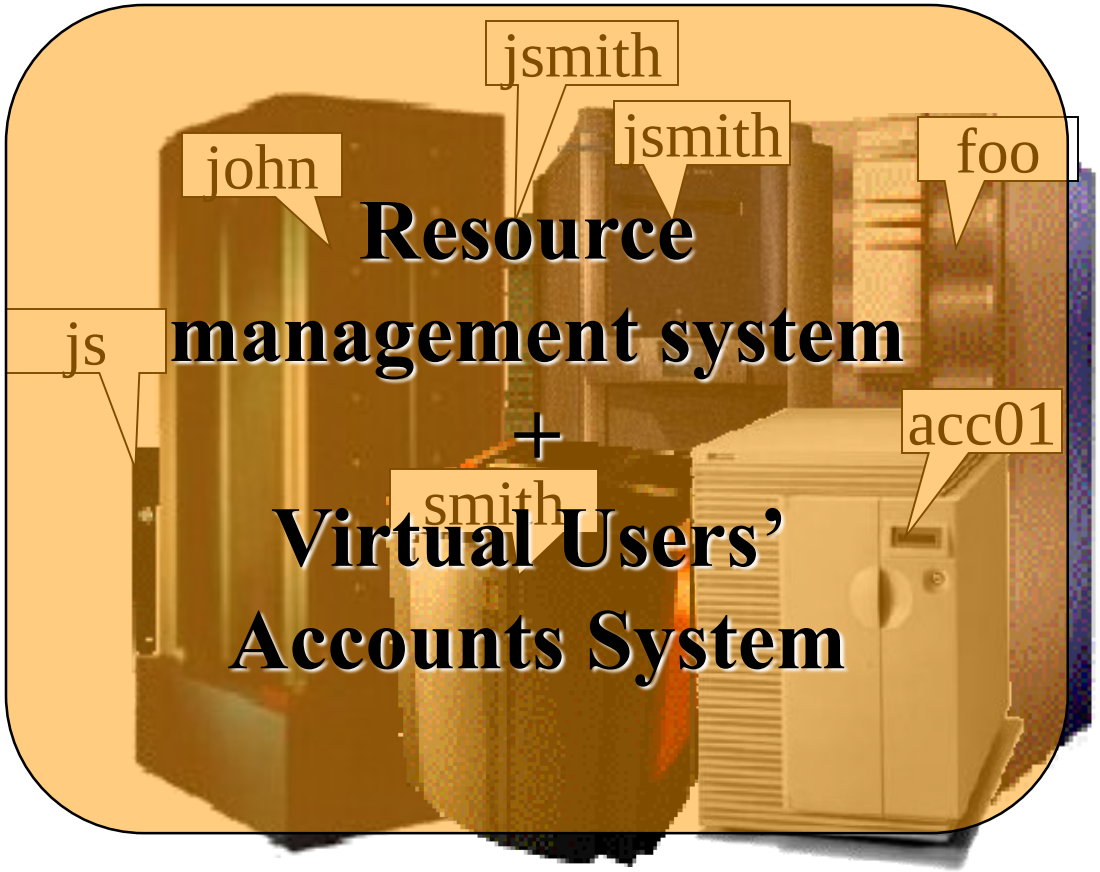
To have summary information about its machines

Requirements (cont.)

- Enabling the user to access all required Grid resources regardless of physical location
  - trivial in testbeds
  - hard to reach in production Grid environment
- Taking into consideration all local (domain) policies regarding security and resource management
- Decreasing the time overheads of user account management
- Enabling distributed accounting, i.e. retrieval of information about resource usage in a distributed environment, with many different and independent domain policies
- Maintaining an adequate security level

No 'Virtual' Accounts

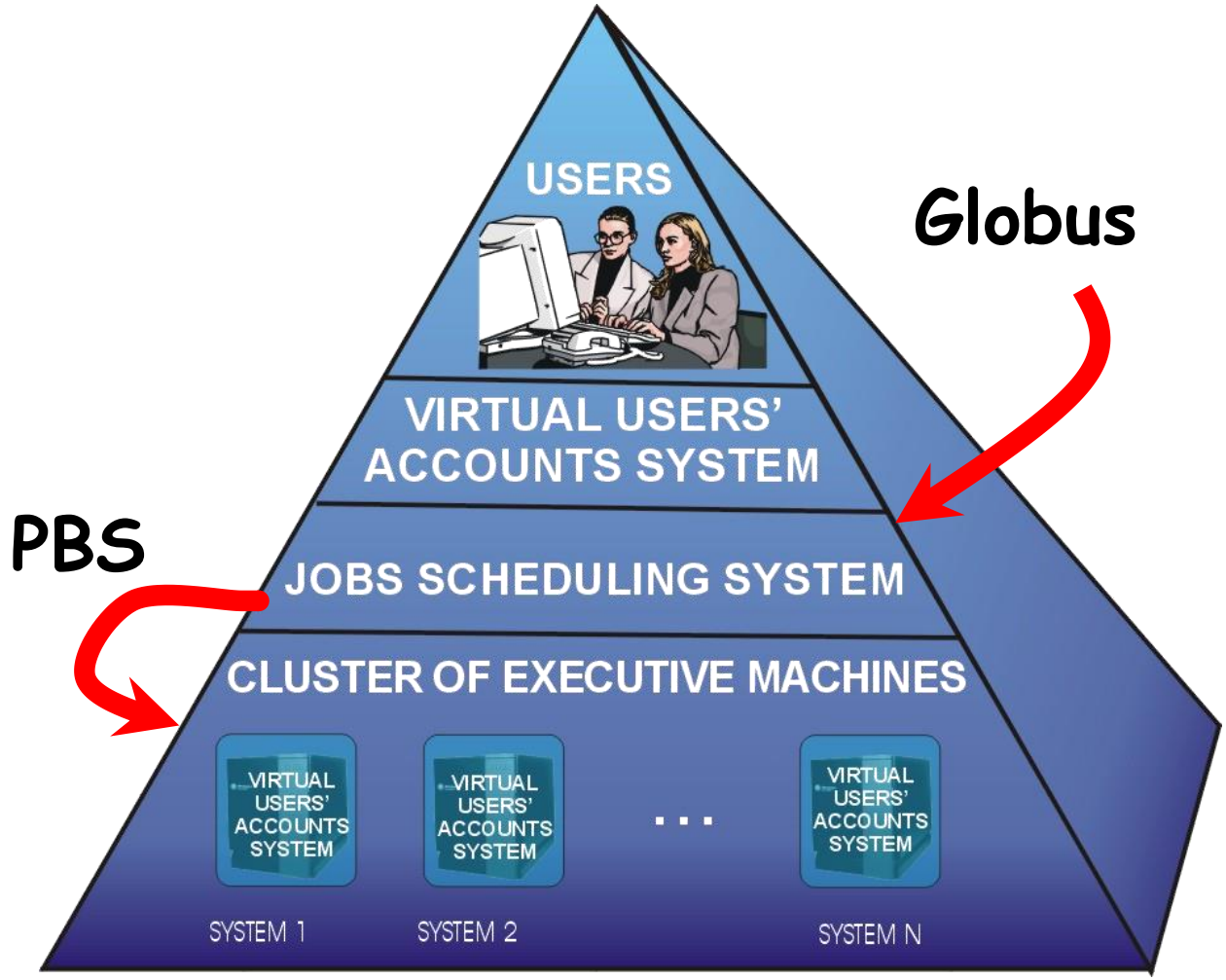
So far the user has had to apply for account on each machine



Virtual User System

- VUS is an extension of the system that runs users' jobs to allow running jobs without having an user account on a node.
- The user is authenticated, authorized and then logged on a 'virtual' account (one user per one account at the time).
- The history of user-account mapping is stored, so that accounting and tracking user activities is possible.

VUS in Action



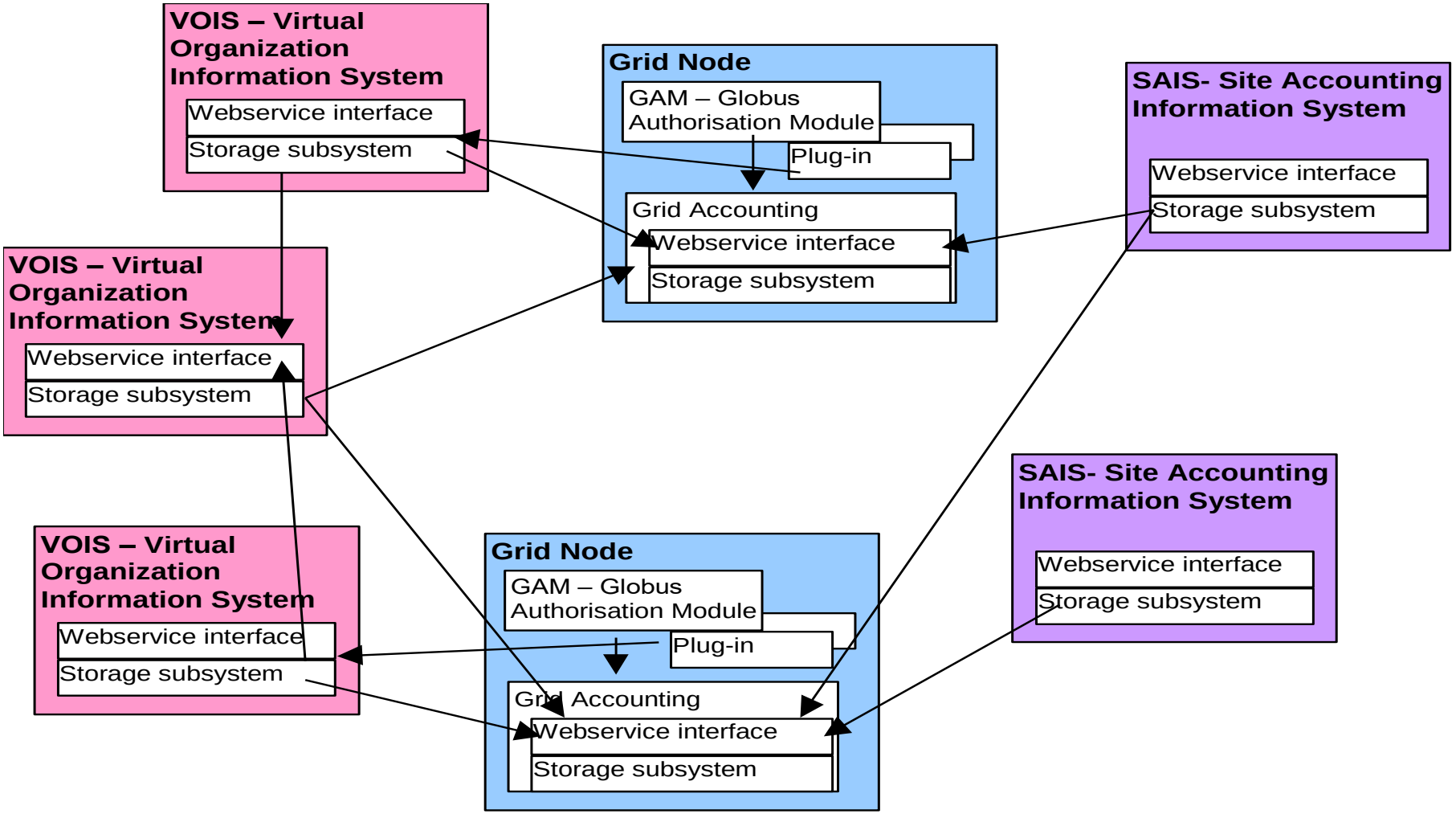
User Accounts in Globus

- Every user has to be added to the grid-mapfile
  - grid-mapfile tends to be very long
- grid-mapfile includes user and not VO
  - Frequent changes to grid-mapfile
- It is recommended that every user should have his/her own account
  - User needs to contact many admins
- There is no accounting support

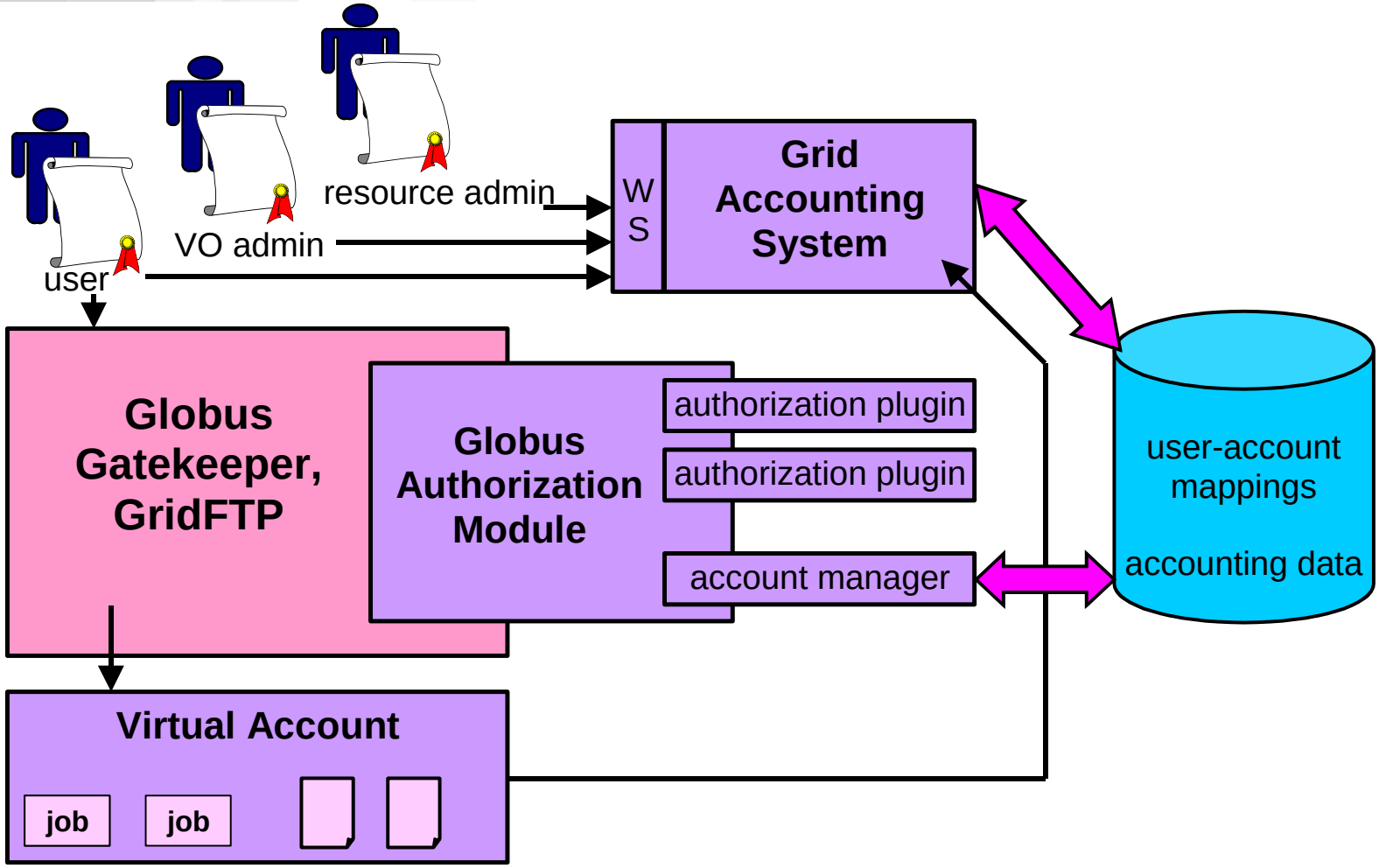
VUS in Globus

- Globus authentication library is replaced
  - impacts gatekeeper, gridftpserver and MDS
- Account service to keep full accounting information
- VO server to keep user list and VO summary accounting
- Each VO can have its own pool of accounts (with different permisions)
- Site server to keep machine list and site summary accounting

Architecture of the System



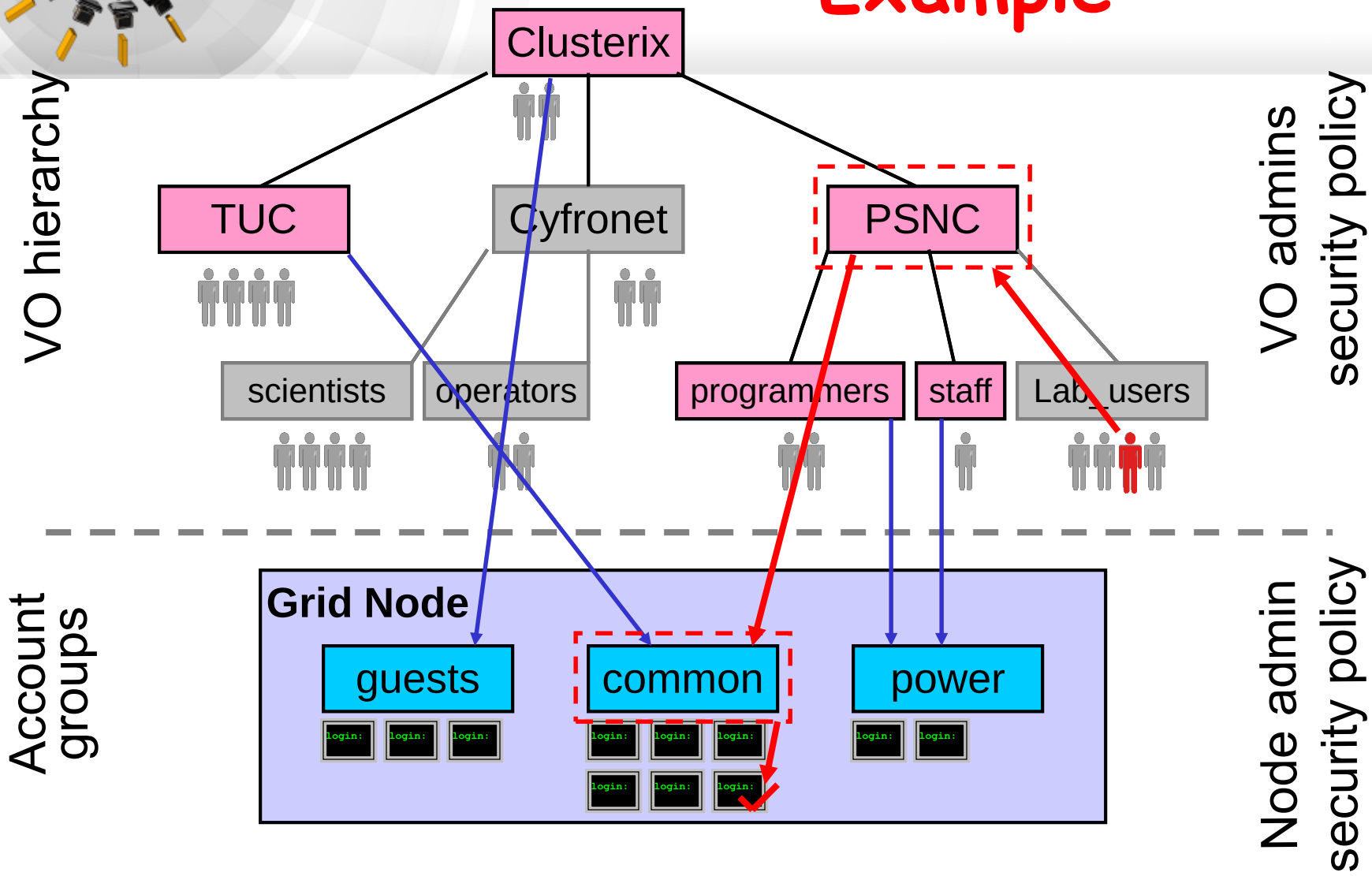
VUS on Grid Node



Plug-in Modules

- Accept all users listed in the grid-mapfile
  - backwards compatibility
- Accept all users that are members of VOs
- *Ban users* assigned to local ban list
- Ask Remote Authorisation System to accept or reject request
- Accept all users with certificate name matching a certain pattern (/C=PL/O=Grid/O=PSNC/*)

VOIS Authorization - Example

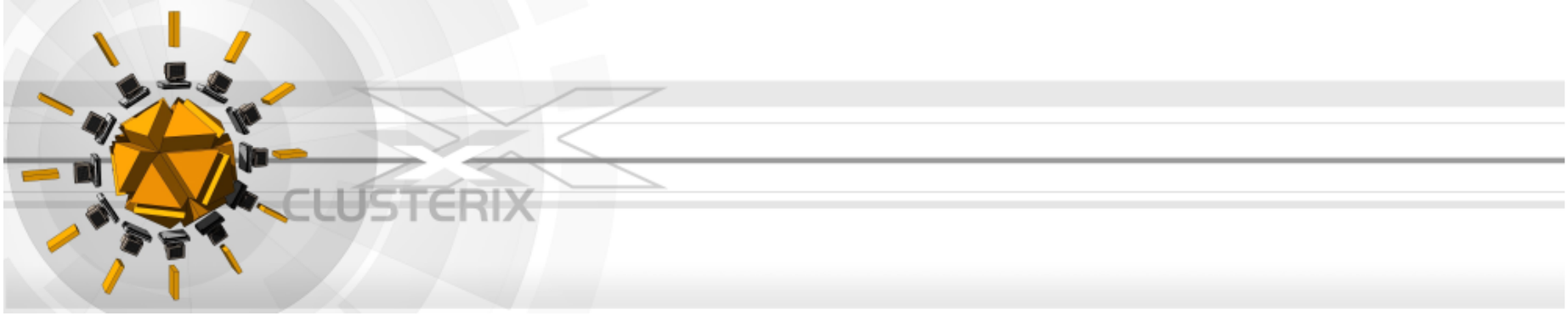


User Account Management: Conclusions

- Allows to introduce the production Grid
  - Dynamic changeable environment
  - First step towards Grid economy
- Keeps local and global policies
- Decreases management (administration overheads)
- Stores standard and non-standard resource usage information
- Supports different *Grid players* : user, resource owner, organization manager

Pilot Applications

- selected applications are developed for experimental verification of the project assumptions and results, as well as to achieve real application results
- running both HTC applications, as well as large-scale distributed applications that require parallel use of one or more local clusters (meta-applications)
- two directions:
  - adaptation of existing applications for Grids
  - development of new applications



SELECTED SCIENTIFIC APPLICATIONS

(out of ~30)

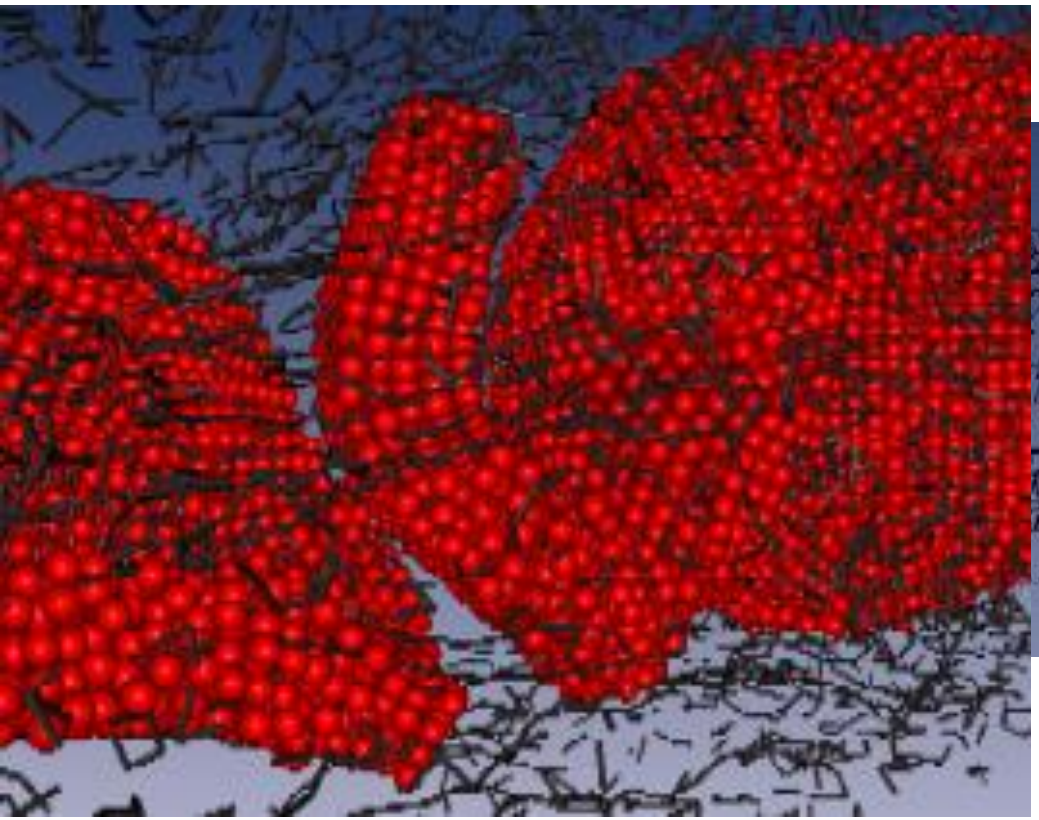
Białystok | Częstochowa | Gdańsk | Łódz | Lublin | Kraków | Opole | Poznań | Szczecin | Warszawa | Wrocław | Zielona Góra

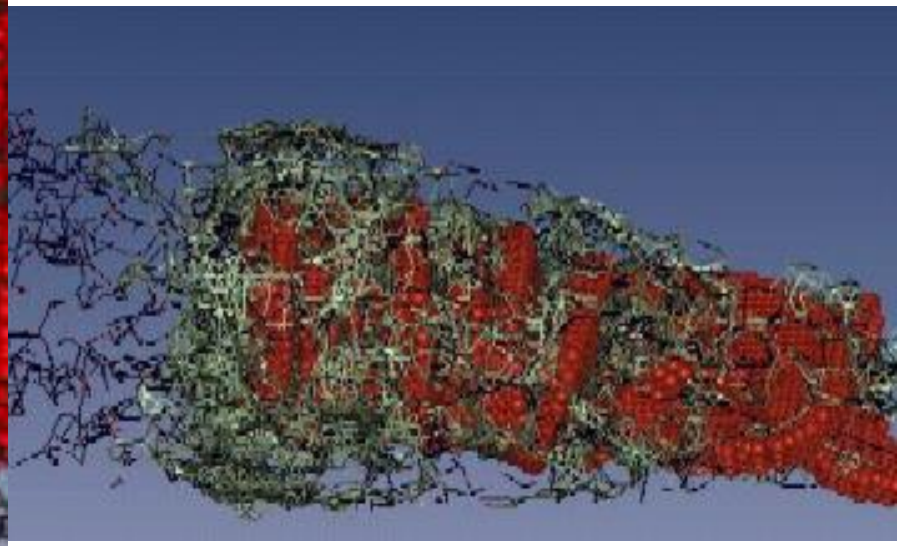
Large scale simulations of blood flow in micro-capillaries (discrete particle model)

W.Dzwinel, K.Boryczko

AGH, Institute of Computer Science

Clot formation due to fibrinogen

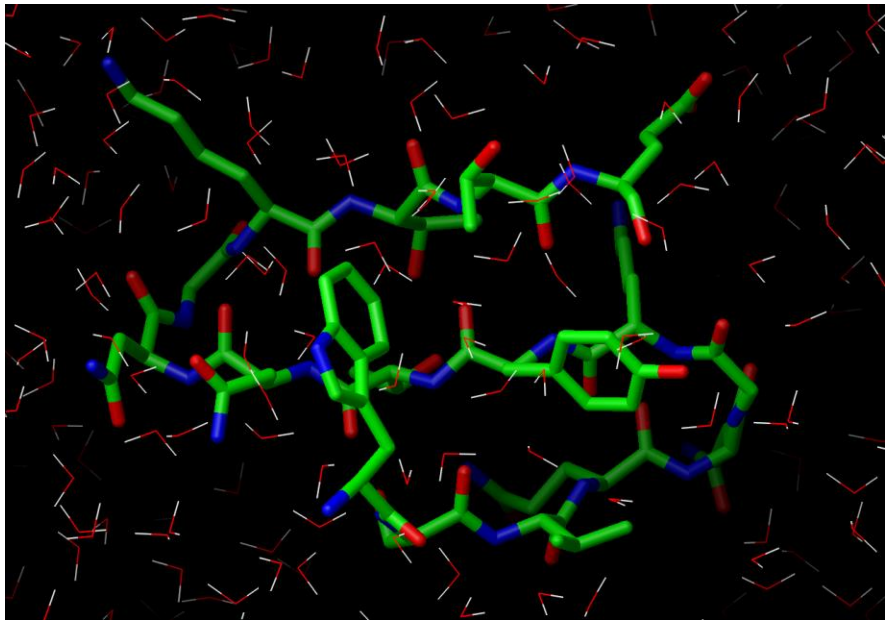


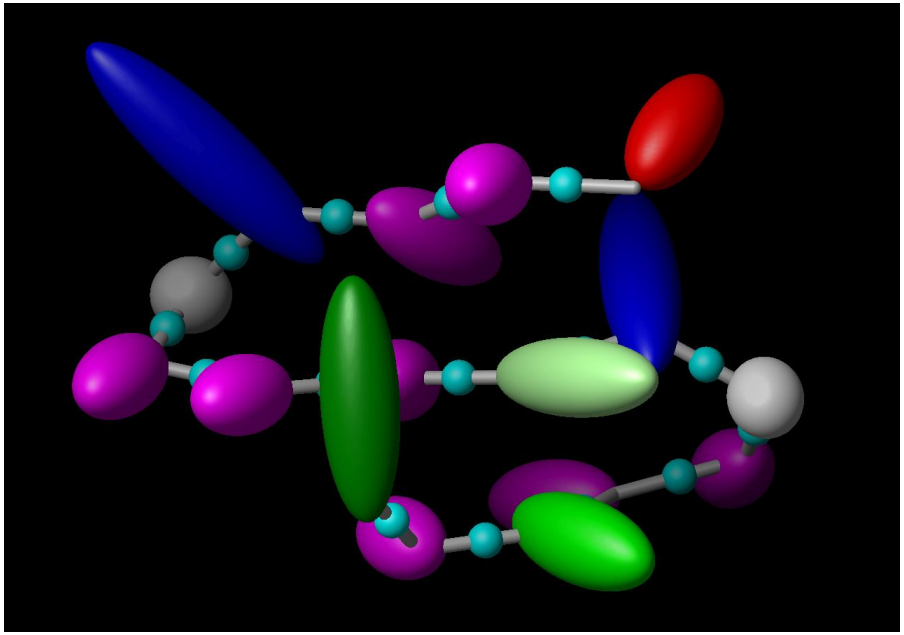


(5x10^6 particles, 16 processors used)

Prediction of Protein Structure

Adam Liwo, Cezary Czaplewski, Stanisław Ołdziej
Department of Chemistry, University of Gdansk

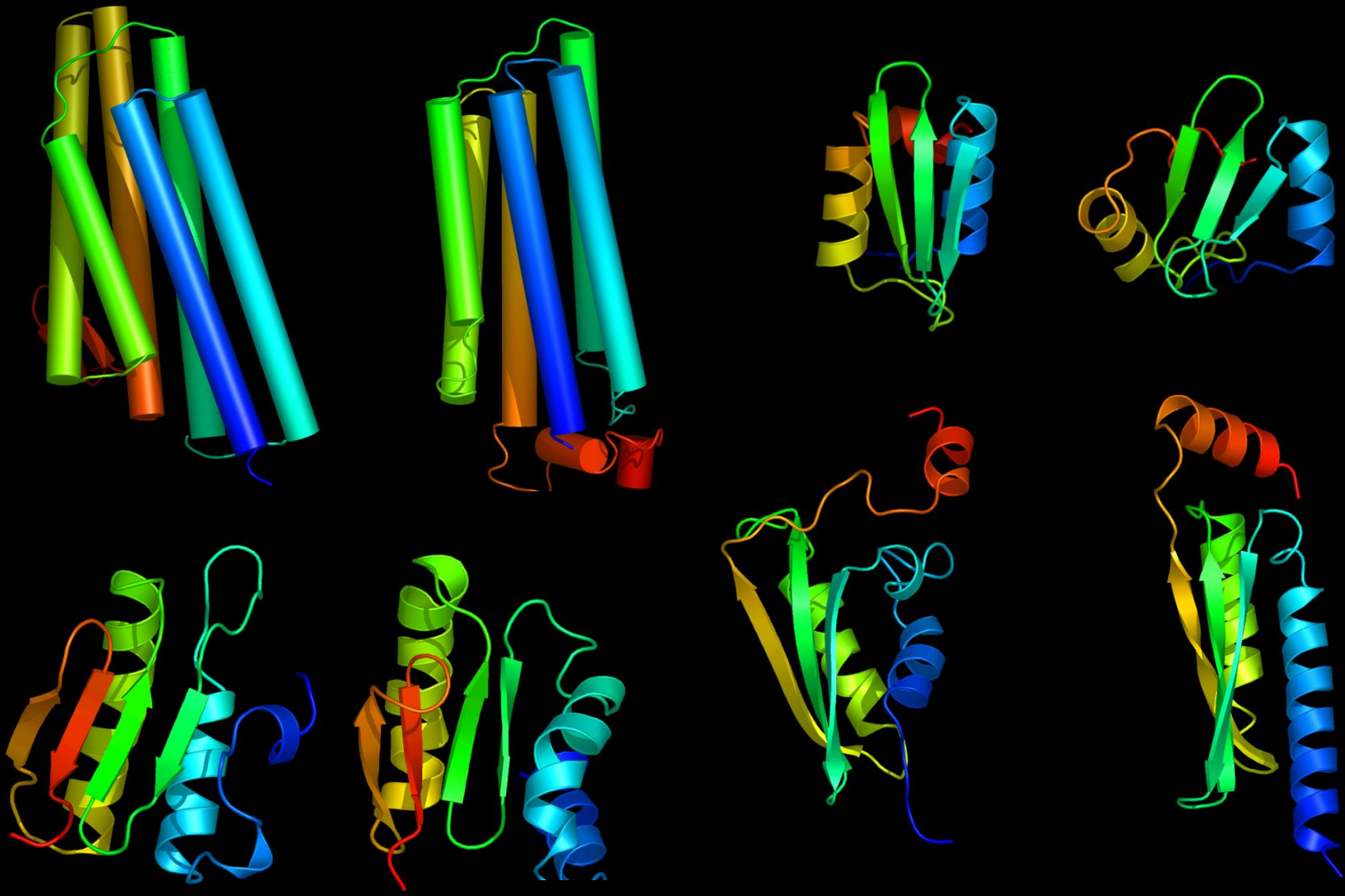




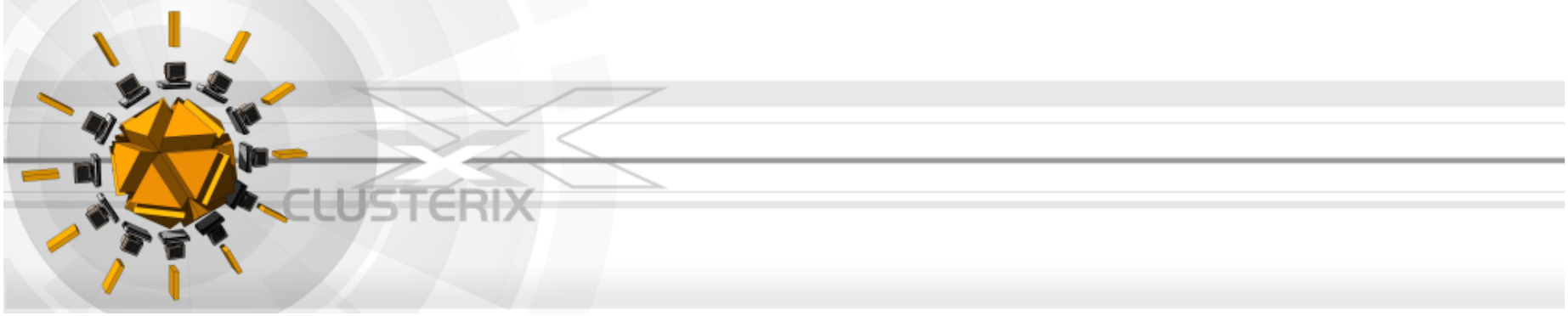
Selected UNRES/CSA results from 6th Community Wide Experiment on the

Critical Assessment of Techniques for Protein Structure Prediction

December 4-8, 2004

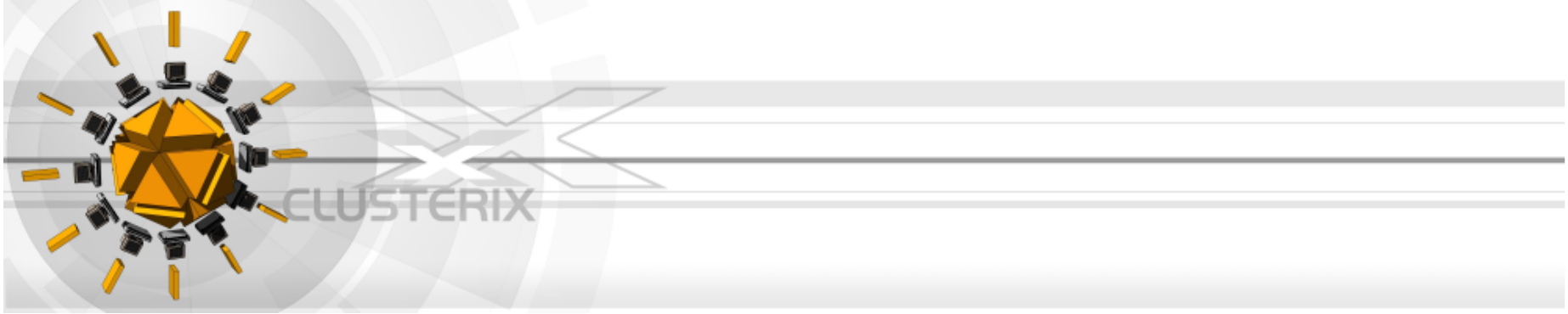


left - experimental structure, right - predicted structure

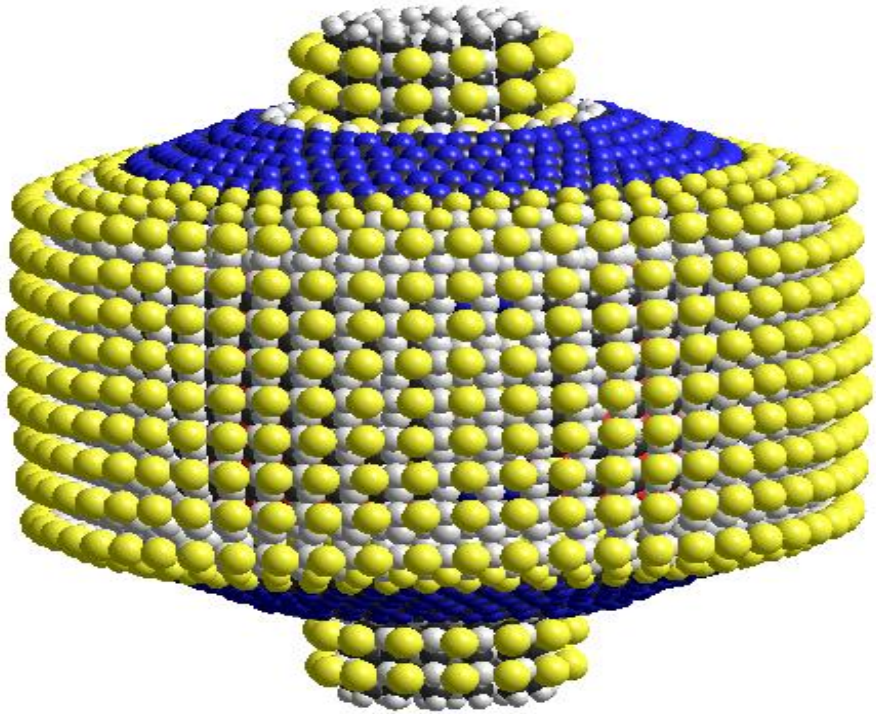


Nano-Engineering

Michał Wróbel, Aleksander Herman
TASK & Gdańsk University of Technology



XMD testing target:
a planetary gear device
containing 8297 atoms
(C, F, H, N, O, P, S and Si)
designed by
K. E. Drexler and R. Merkle



- XMD an Open Source computer package for performing molecular dynamics simulations of nano-devices and systems.

Flow simulations in Aeronautics

-

in-house HADRON code

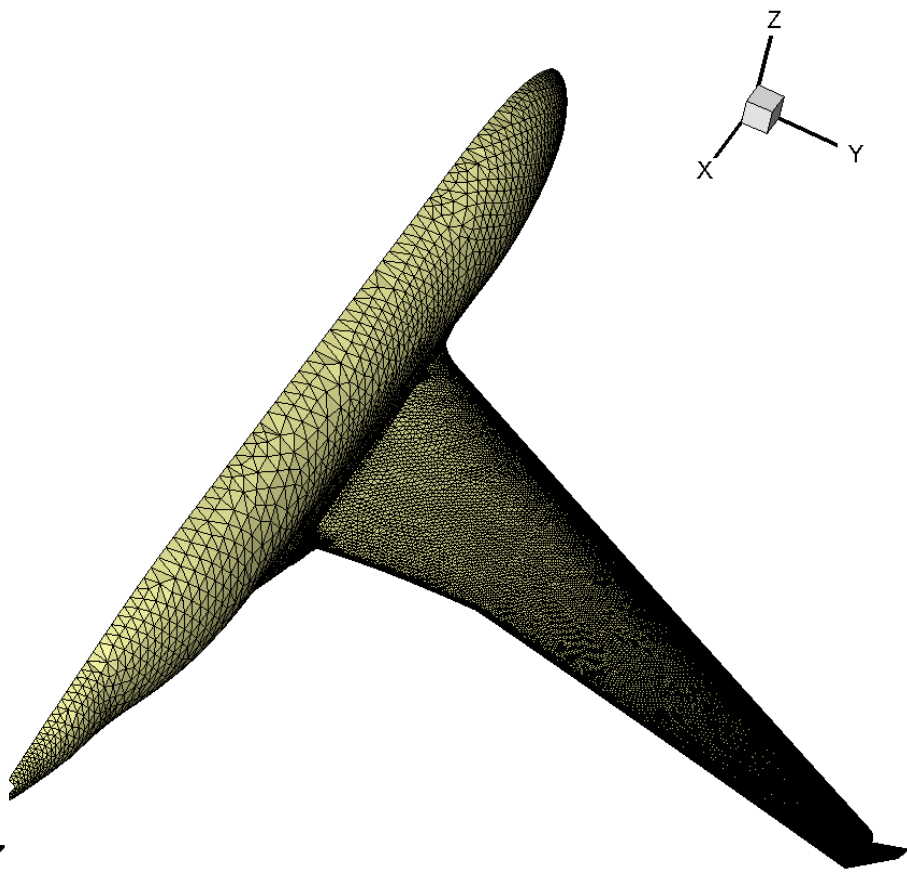
Jacek Rokicki

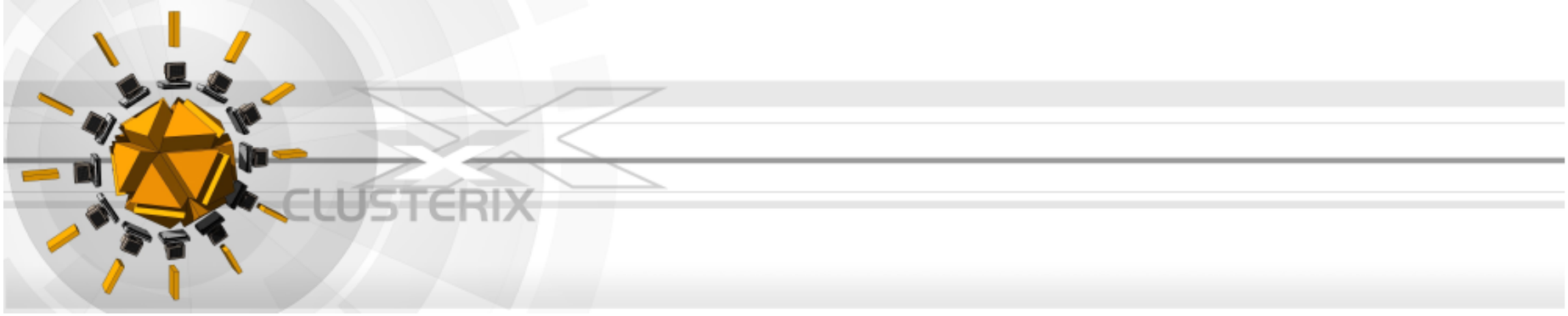
Warsaw University of Technology

Large 3D computational problems

Modeling and design of advanced Wing-tip devices

2003-2006





NuscaS

Czestochowa University of Technology

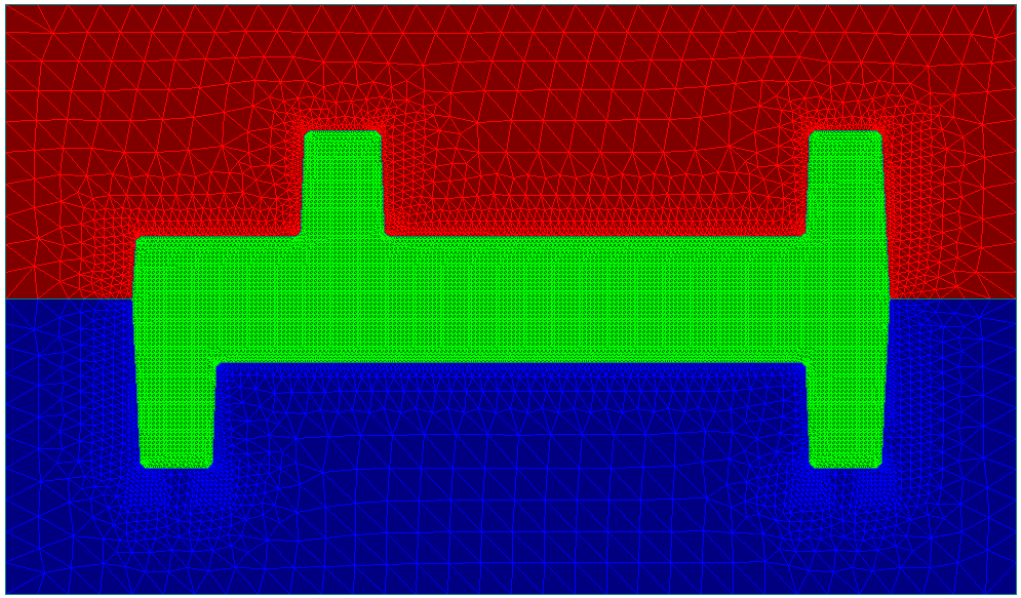
Tomasz Olas

Application areas:

different thermo-mechanic phenomena:

heat transfer, solidification, stress in thermo-elastic states, stress in thermo-elasto-plastic states, estimation of hot-tearing in casting, mechanical interactions between bodies, hot-tearing, damage, etc.

Finite Element Modeling of Solidification



FEM mesh and its partitioning

Different Scenarios of using Grid Resources

- Grid as the resource pool
  an appropriate computational resource (local cluster) is found via resource management system, and the sequential application is started there
- Parallel execution on grid resources
  (meta-computing application):
  - Single parallel application being run on geographically remote resources
  - Grid-aware parallel application - the problem is decomposed taking into account Grid architecture

MPICH-G2

- The MPICH-G2 tool is used as a grid-enabled implementation of the MPI standard (version 1.1)
- It is based on the Globus Toolkit used for such purposes as authentification, authorization, process creation, process control, …
- MPICH-G2 allows to couple multiple machines, potentially of different architectures, to run MPI applications
- To improve performance, it is possible to use other MPICH-based vendor implementations of MPI in local clusters (e.g. MPICH-GM)

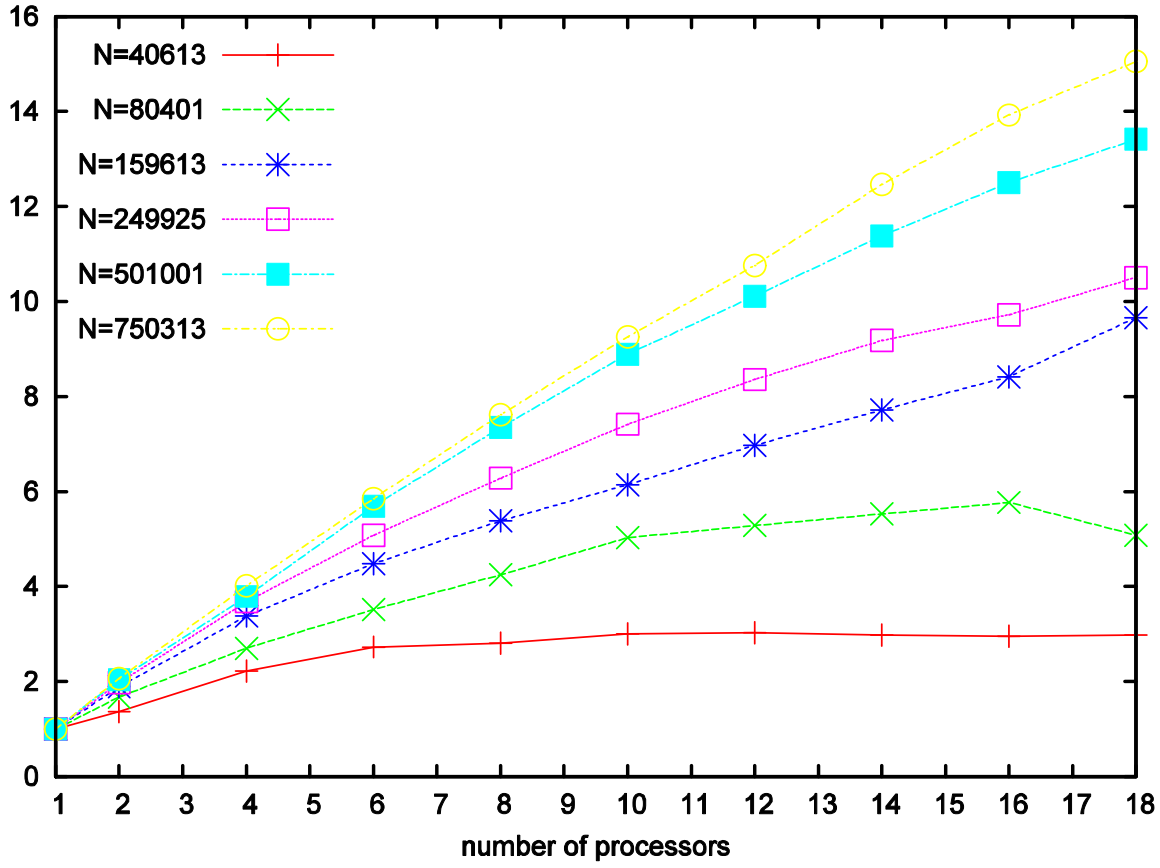
CLUSTERIX as a Heteregeneous System

- Hierarchical architecture of CLUSTERIX

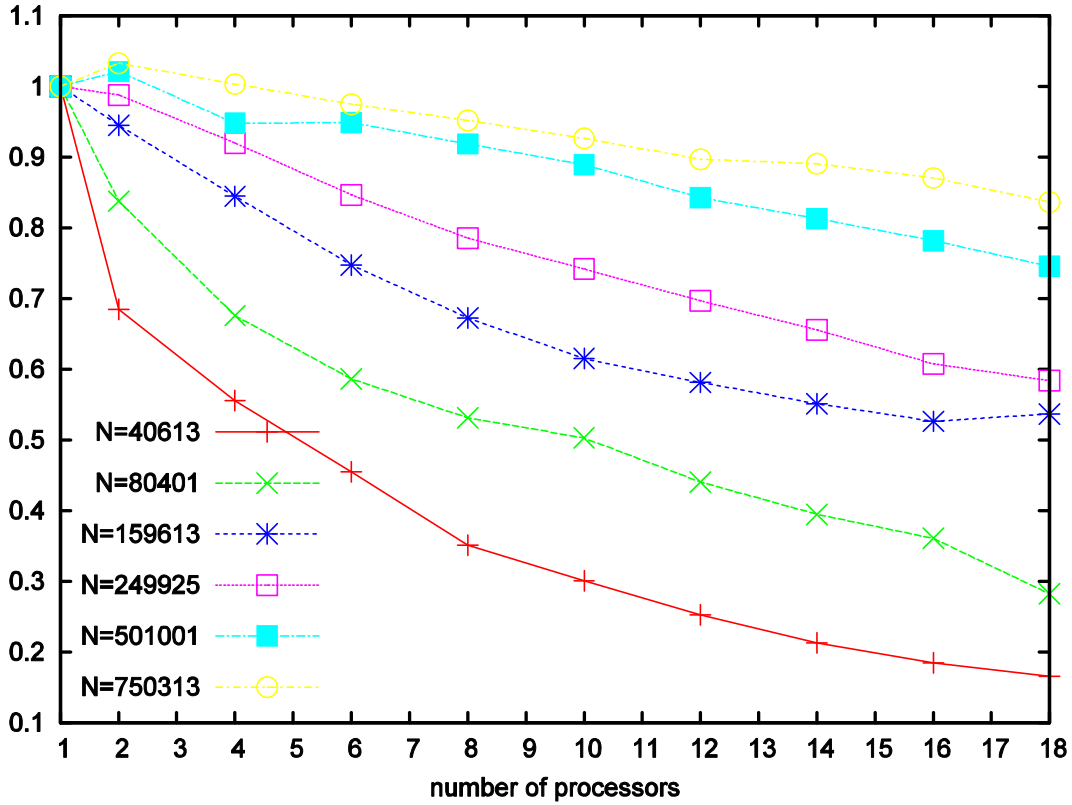
	latency	bandwitch
local (MPI)	104 μs	752 $\frac{Mb}{s}$
local (MPICH-G2)	124 μs	745 $\frac{Mb}{s}$
global (MPICH-G2)	10 ms	33 $\frac{Mb}{s}$

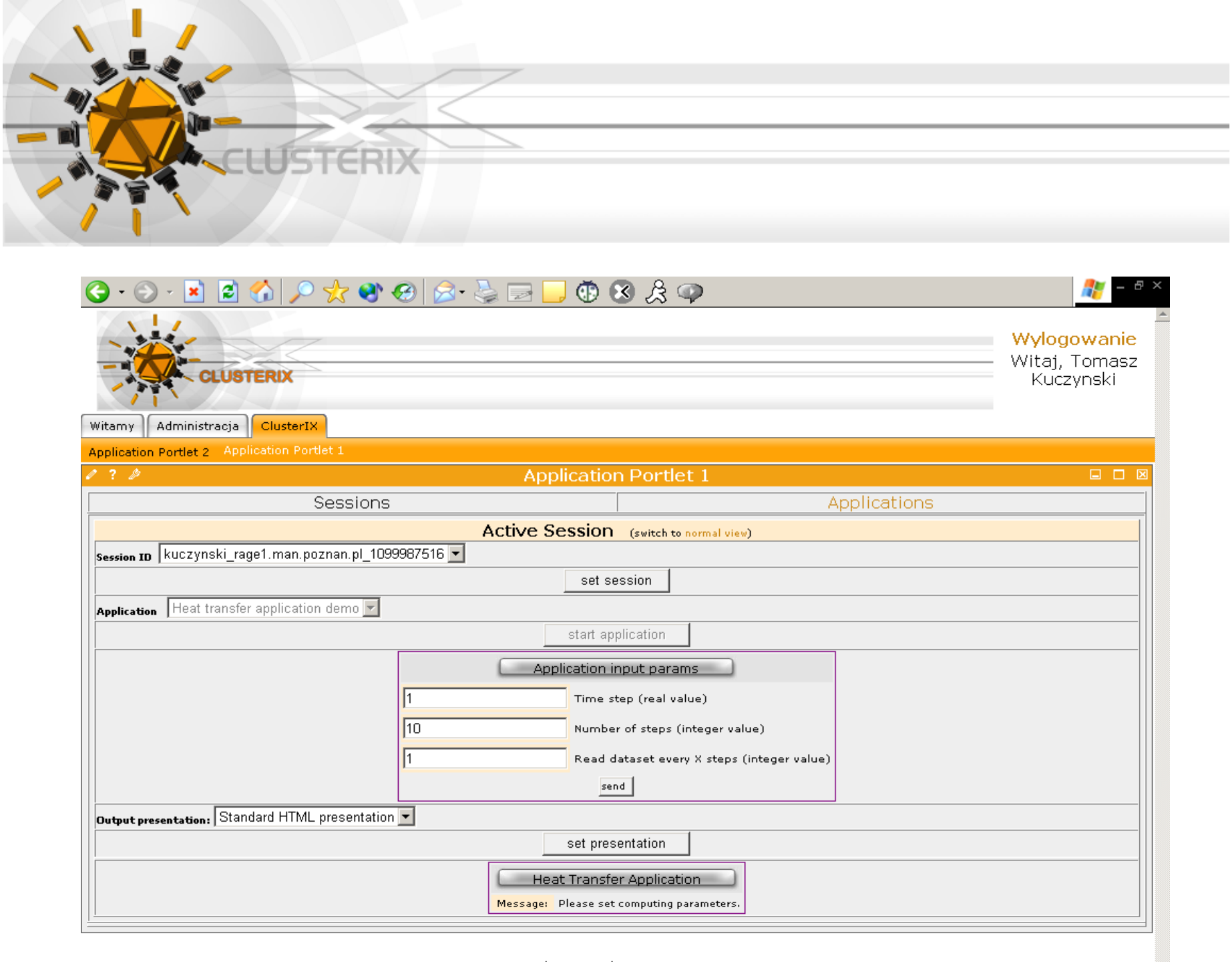
- It is not a trivial issue to adopt an application for its efficient execution in the CLUSTERIX environment
- Communicator construction in MPICH-G2 can be used to represent hierarchical structures of heterogenous systems, allowing applications for adaptation of their behaviour to such structures

Performance Results (1)



Performance Results (2)




CLUSTERIX
CLUSTERIX
Wylogowanie
Witaj, Tomasz Kuczynski
Witamy
Administracja
ClusterIX
Application Portlet 2
Application Portlet 1
Application Portlet 1
Sessions
Applications
Active Session
(switch to normal view)
Session ID
kuczynski_rage1.man.poznan.pl_1099987516
set session
Application
Heat transfer application demo
start application
Application input params
1
Time step (real value)
10
Number of steps (integer value)
1
Read dataset every X steps (integer value)
send
Output presentation:
Standard HTML presentation
set presentation
Heat Transfer Application
Message:
Please set computing parameters.

CLUSTERIX

Application input params

Run application — Command

send

Output presentation: Extended HTML presentation

set presentation

Output presentation parameters

870 — Start node number (for range details please see "Standard HTML presentation", default 0)

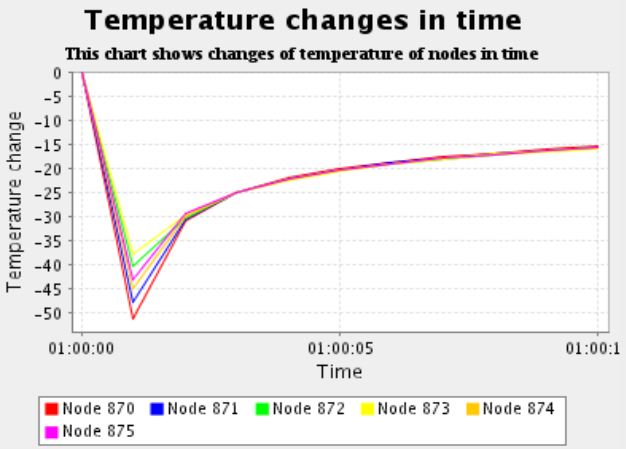
875 — End node number (for range details please see "Standard HTML presentation", default 0)

temperature changes in time — Chart type

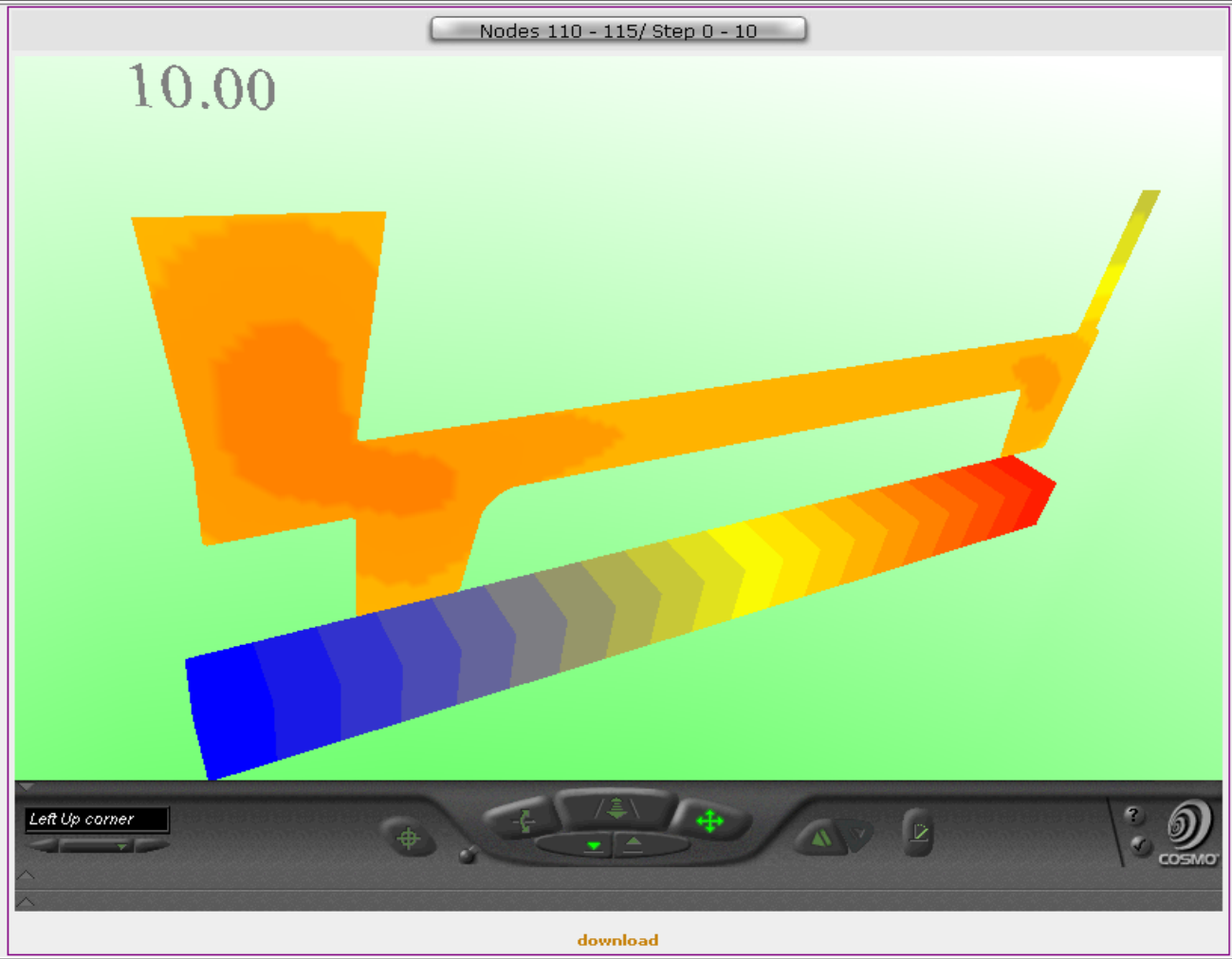
set

Nodes 870 - 875

Step	Time	Node 870 x: 0.085041 y: 0.05	Node 871 x: 0.0859094 y: 0.0521511	Node 872 x: 0.0880339 y: 0.0521417	Node 873 x: 0.0872211 y: 0.05	Node 874 x: 0.0894224 y: 0.05	Node 875 x: 0.091392 y: 0.0564499
0	0	1000	1000	1000	1000	1000	1000
1	1	948.711	952.393	959.769	962.151	955.172	957.066
2	2	917.795	921.919	929.63	932.534	925.231	927.613
3	3	892.88	896.955	904.523	907.463	900.285	902.725
4	4	870.953	874.907	882.263	885.139	878.156	880.555
5	5	850.918	854.742	861.875	864.665	857.892	860.224
6	6	832.239	835.937	842.848	845.549	838.985	841.245
7	7	814.612	818.188	824.883	827.496	821.136	823.325
8	8	797.847	801.306	807.791	810.319	804.159	806.276
9	9	781.818	785.165	791.447	793.892	787.925	789.972
10	10	766.437	769.676	775.76	778.126	772.345	774.326



CLUSTERIX



Final Remarks

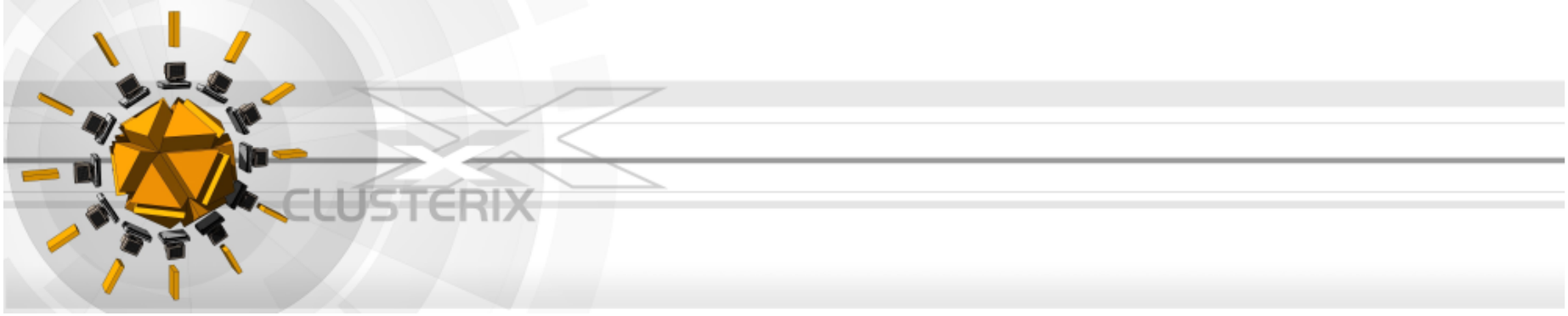
- At this moment, the first version of CLUSTERIX middleware is alraedy available
- Intensive testing of middleware modules and their interactions
- First experiences with running application in CLUSTERIX environment
- Demo at SC'05
- Extremely important for us:
  - to attract perspective users with new applications
  - to involve new dynamic clusters
  - training activities

Thank YOU !

www: https://clusterix.pl

Roman Wyrzykowski
roman@icis.pcz.pl

Norbert Meyer
meyer@man.poznan.pl


CLUSTERIX