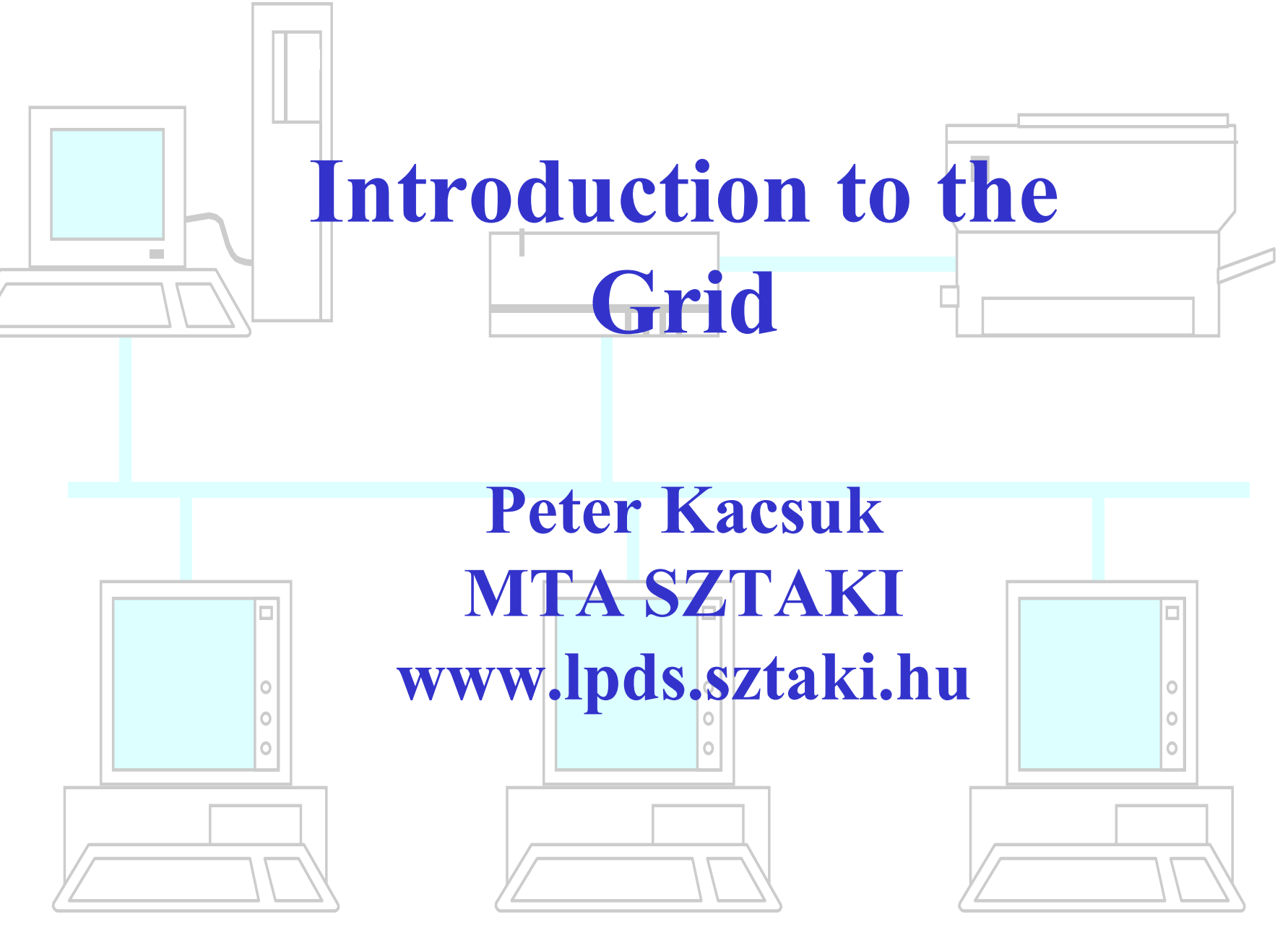


# Introduction to the Grid

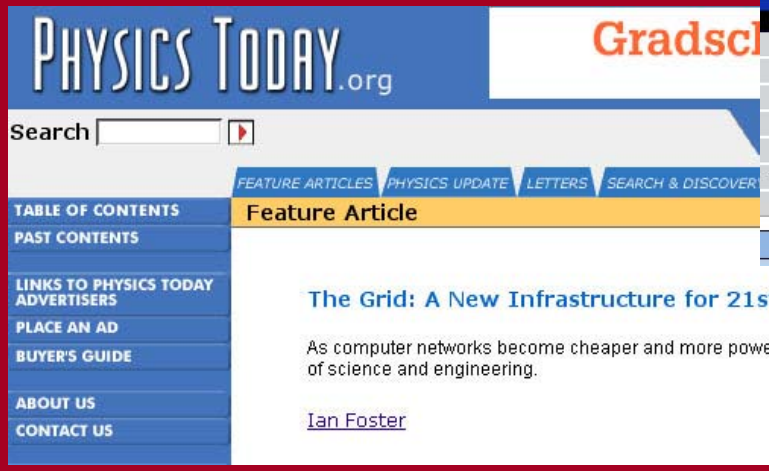
Peter Kacsuk
MTA SZTAKI
www.lpds.sztaki.hu

# *Agenda*

- From Metacomputers to the Grid
- Grid Applications
- Job Managers in the Grid - Condor
- Grid Middleware – Globus
- Grid Application Environments

# Grid Computing in the News





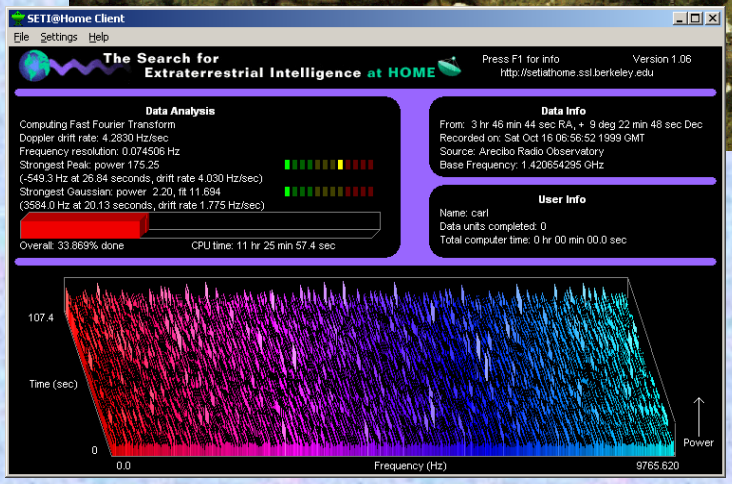





Credit to Fran Berman

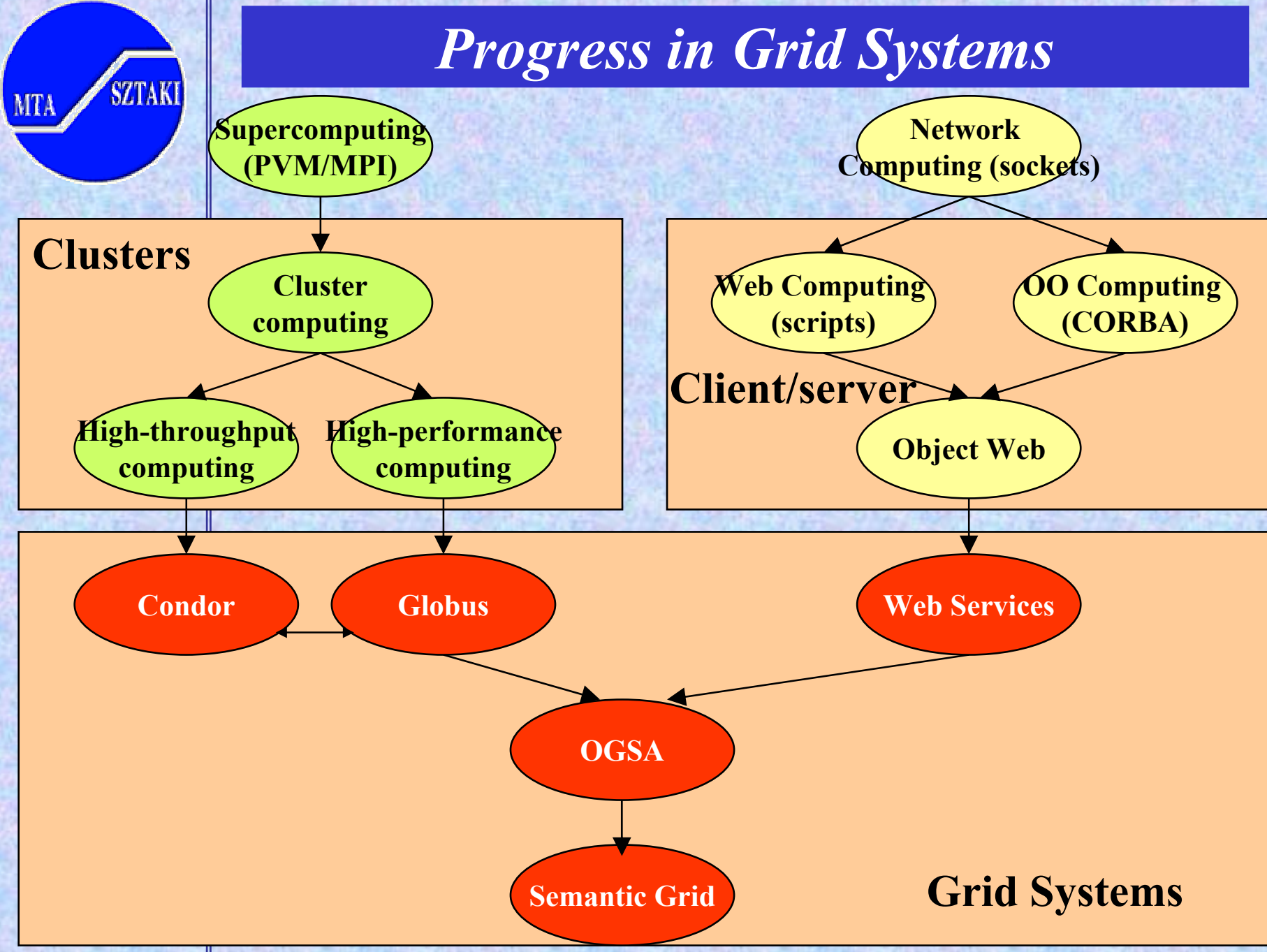
# *Real World Distributed Applications*

- SETI@home
  - 3.8M users in 226 countries
  - 1200 CPU years/day
  - 38 TF sustained (Japanese Earth Simulator is 40 TF peak)
  - 1.7 ZETAflop over last 3 years (10^21, beyond peta and exa …)
  - Highly heterogeneous: >77 *different* processor types

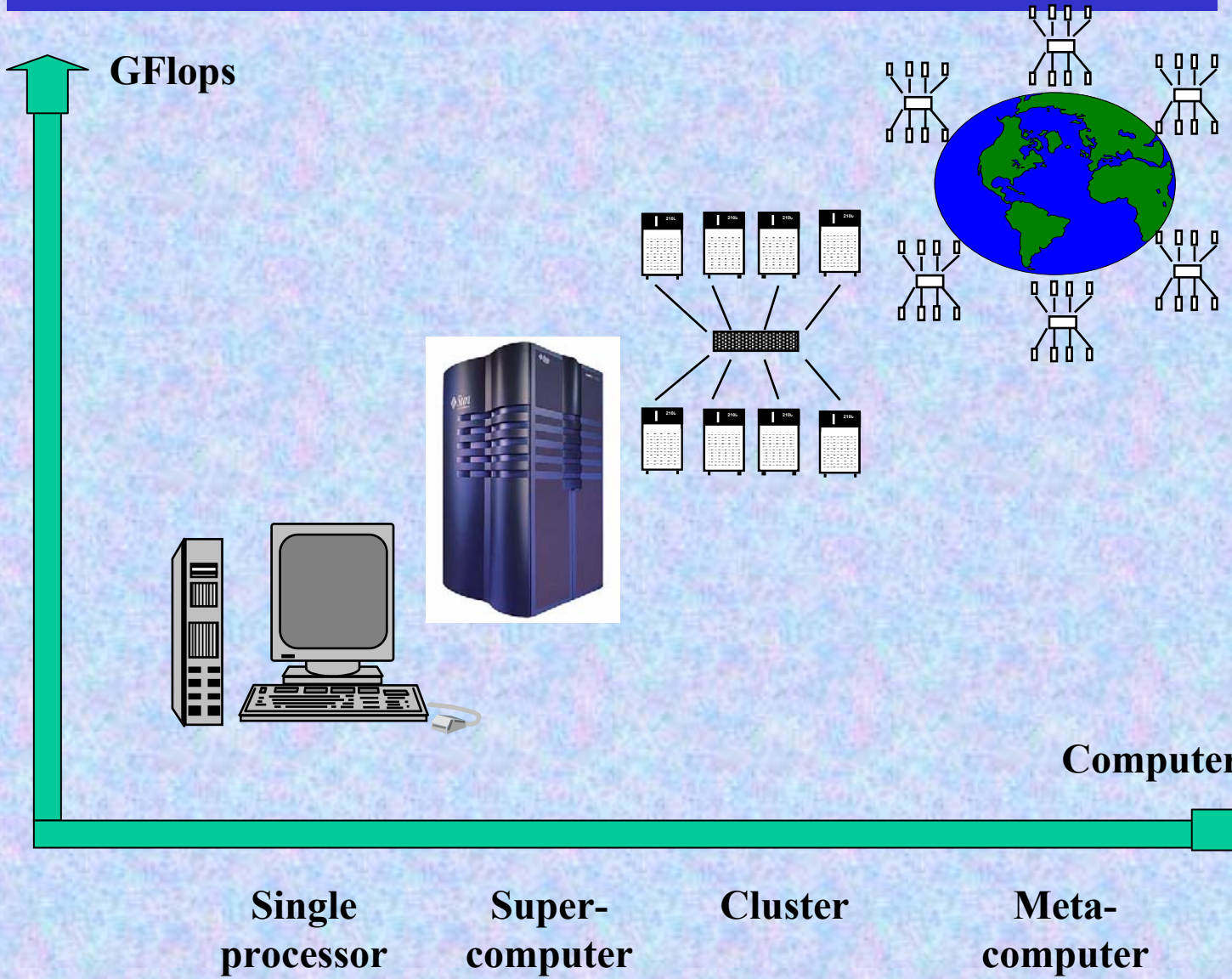





Credit to Fran Berman


MTA SZTAKI
Progress in Grid Systems
Supercomputing (PVM/MPI)
Network Computing (sockets)
Clusters
Cluster computing
Web Computing (scripts)
OO Computing (CORBA)
Client/server
High-throughput computing
High-performance computing
Object Web
Condor
Globus
Web Services
OGSA
Semantic Grid
Grid Systems

# *Progress to the Grid*



# *A metaszámítógépek megalkotásának eredeti motivációi*

- Az un. nagy kihívást jelentő problémák megoldása heteket sőt hónapokat vesz igénybe még a szuperszámítógépeken is

- Különböző szuperszámítógépeket és klasztereket kellett összekapcsolni távolsági hálózatokkal annak érdekében, hogy a fenti problémákat ésszerű időn belül meg lehessen oldani

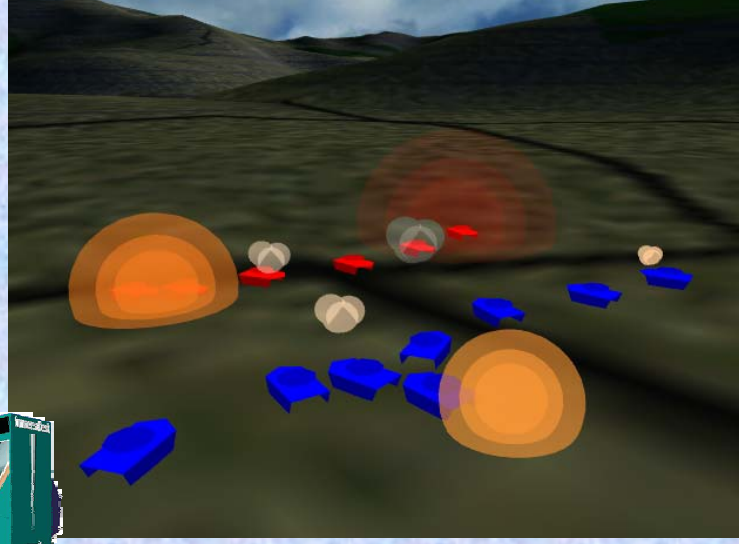
# *A metaszámítógép eredeti jelentése*

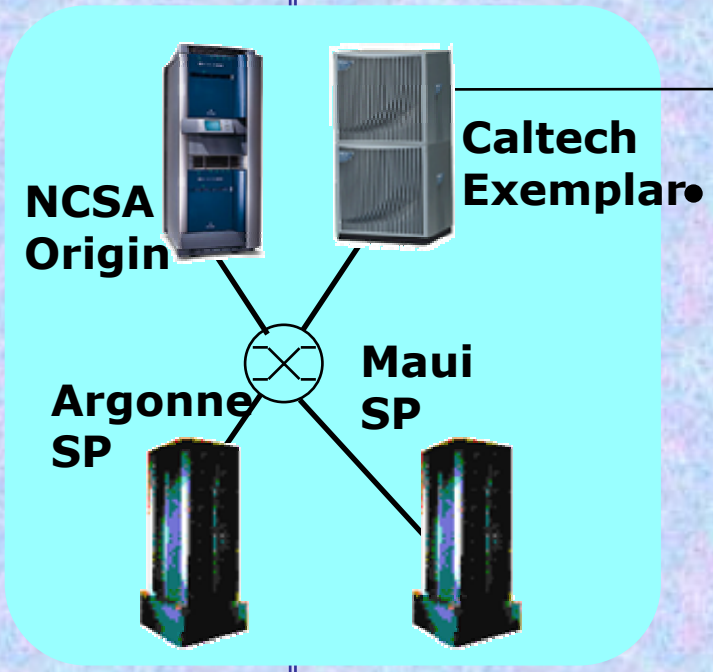
Metaszámítógép = Szuperszám. technológia + Távolsági hálózat

## A metaszámítógép eredeti célja

- **Nagyobb teljesítményt** elérni, mint az egyedi szuperszámítógépek/klaszterek tudnak biztosítani

# *Distributed Supercomputing*






Issues:

- **Resource discovery, scheduling**
- **Configuration**
- **Multiple comm methods**
- **Message passing (MPI)**
- **Scalability**
- **Fault tolerance**

**SF-Express Distributed Interactive Simulation:**
**Caltech, USC/ISI**

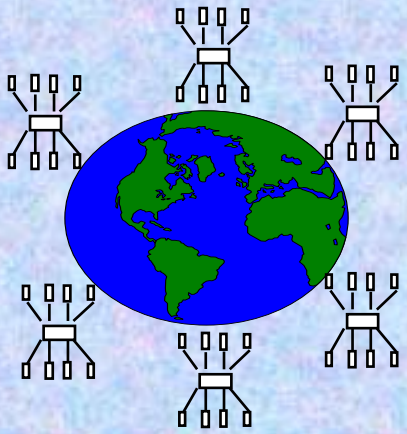
# *Második lépés a metaszámítógépek felé*

Szuperszám. technológia

WAN technológia

Elosztott rendszerek

Metaszám.gépek



# *Mi is a metaszámítógép?*

- A metaszámítógép olyan számítógépek együttese, amelyek
  - heterogének minden szempontból
  - földrajzilag elosztottak
  - távolsági hálózattal vannak összekötve
  - egyetlen komputer képét alkotják (SSI)
- Metaszámítás jelentése:
  - hálózat alapú
  - elosztott **szuperszámítógép technológia**

# *További motivációk a metaszámítógépek megalkotására*

- A távolsági hálózattal elérhető számítási és egyéb erőforrások hatékonyabb kihasználása

- Különböző számítógépeket kell távolsági hálózattal összekötni a szabad ciklusok kihasználására
- Különböző speciális készülékeket kell távolsági hálózattal összekötni kollaboratív munka biztosításához

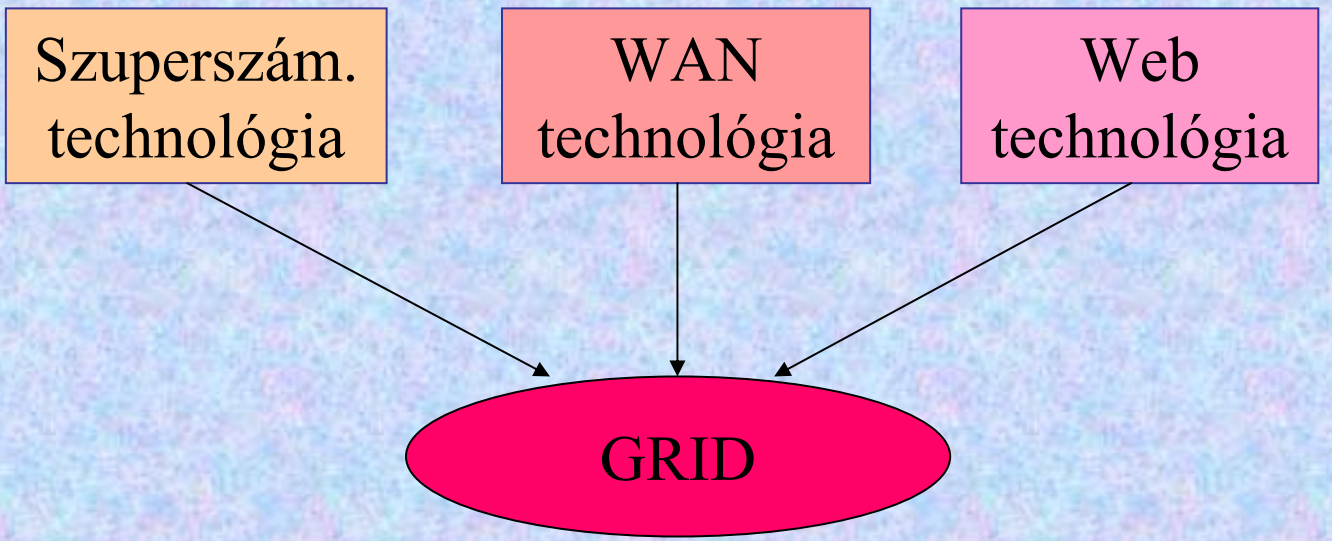
# *Motivációk a GRID infrastruktúra megalkotására*

- Olyan számítási és adatfeldolgozási GRID-et létrehozni, amely hasonlóan széleskörű, mint az információ elérése a weben

- Bármely számítógépet/készüléket célszerű összekötni távolsági hálózattal, hogy univerzális feldolgozó kapacitást nyerjünk

- **Grid** = általánosított metaszámítógép technológia

# *A GRID megalkotásához vezető technológiák*



# *Mi is a GRID?*

- A GRID olyan számítógépek, tárolóegységek és egyéb készülékek együttese, amelyek
  - heterogének minden szempontból
  - földrajzilag elosztottak
  - távolsági hálózattal vannak összekötve
  - egyetlen komputer képét alkotják (SSI)
- Az általánosított metaszámítási technológia (GRID) jelentése:
  - hálózat alapú
  - elosztott **adatfeldolgozási technológia**

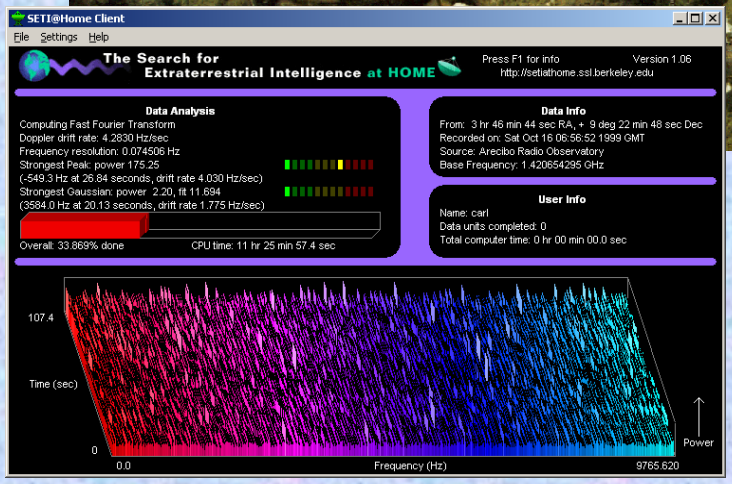
# *A GRID alkalmazási területei*

- High throughput computing
- Virtuális laboratórium
  - Kollaboratív tervezés
- Adat intenzív alkalmazások
  - adatbányászat, részecskefizika
- Földrajzi információs rendszerek

- Távoli jelenlét (tele-immersion)
- Vállalati rendszerek

# Real World Distributed Applications

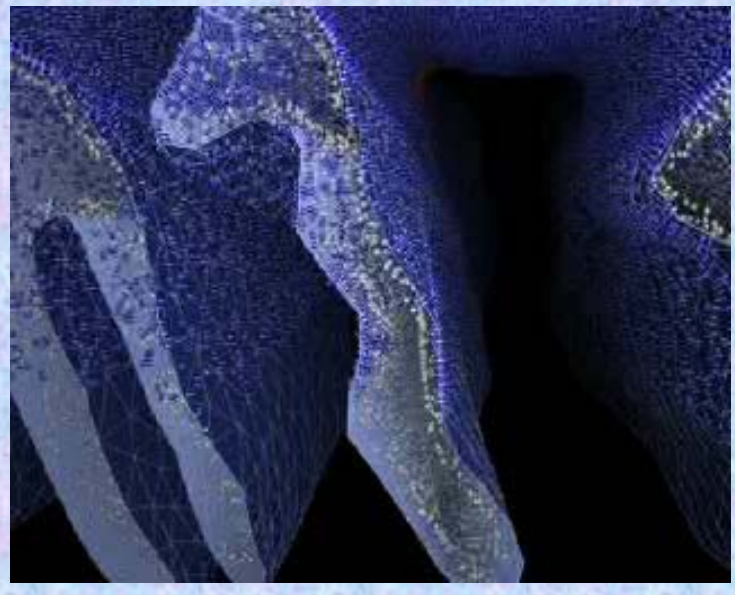
- SETI@home
  - 3.8M users in 226 countries
  - 1200 CPU years/day
  - 38 TF sustained (Japanese Earth Simulator is 40 TF peak)
  - 1.7 ZETAflop over last 3 years (10^21, beyond peta and exa …)
  - Highly heterogeneous: >77 *different* processor types



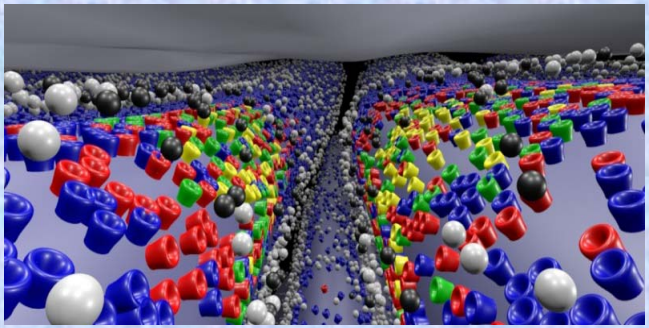
# Minimal Communication Applications

## MCell -- Simulation of neuromuscular synaptic transmission

- Uses Monte Carlo diffusion and chemical reaction algorithm in 3D to simulate complex biochemical interactions of molecules
- Molecular environment represented as 3D space in which trajectories of ligands against cell membranes tracked
- Ultimate Goal: A complete molecular model of neuro-transmission at level of entire cell

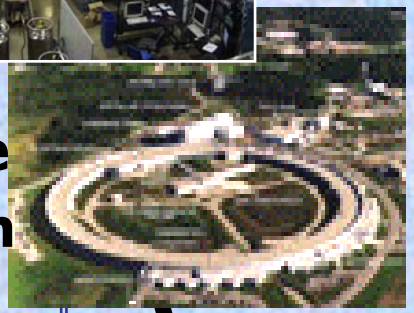


MCell Animation



# Online Instruments

Advanced Photon Source

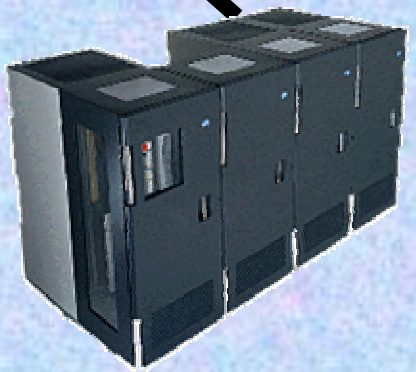


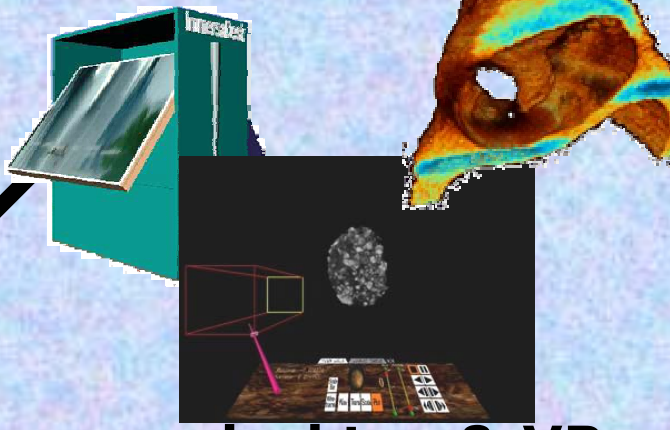
eal-time
ollection

wide-area
dissemination

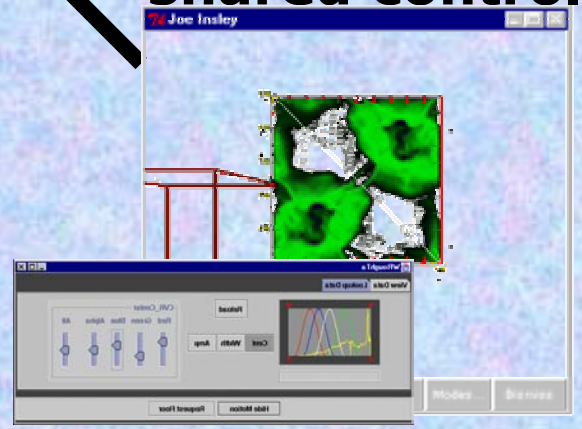
archival
storage

tomographic
reconstruction





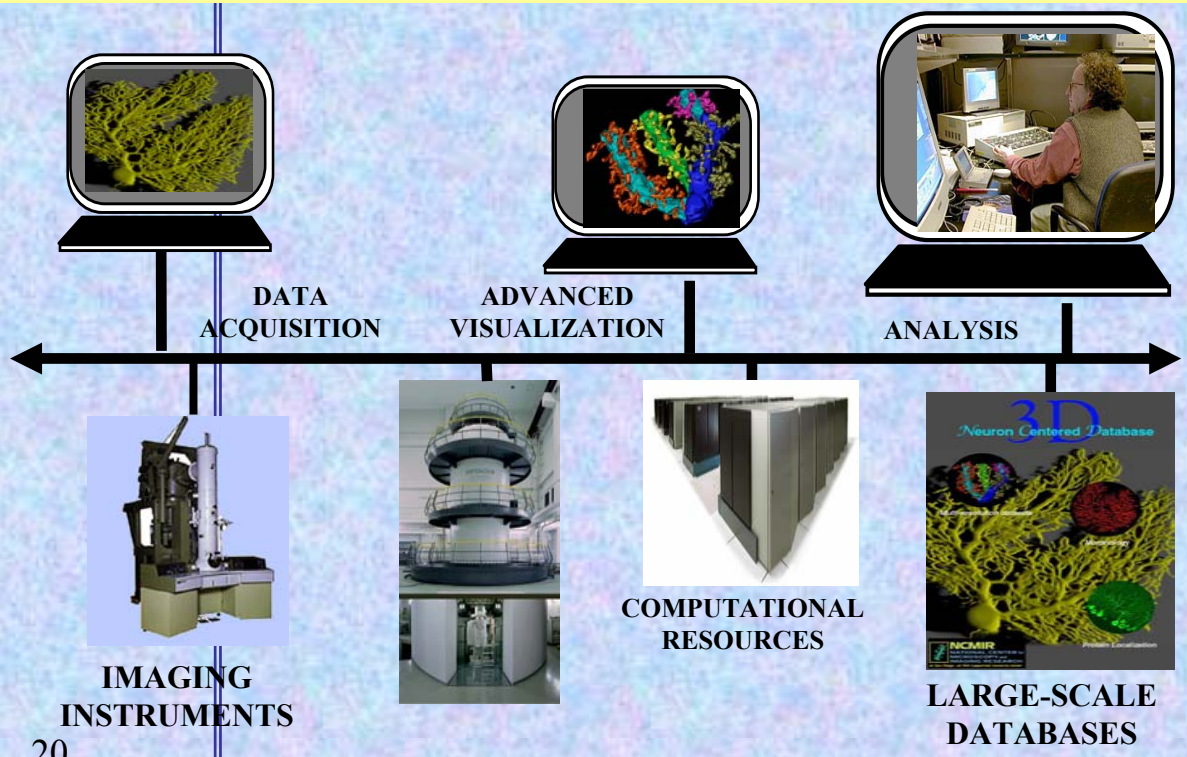
desktop & VR
clients with
shared controls

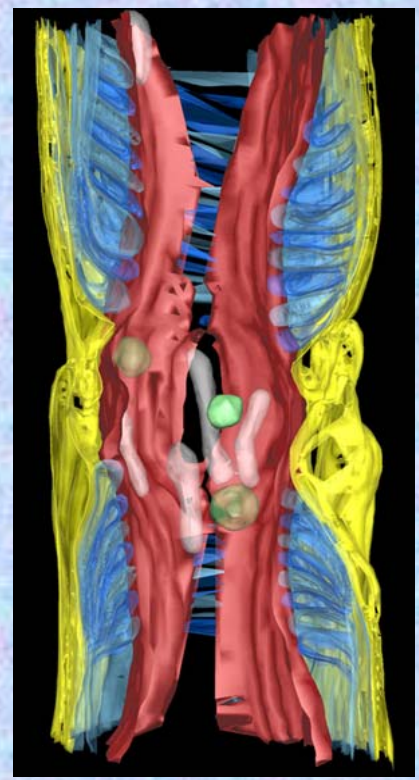


DOE X-ray source grand challenge: ANL, USC/ISI, U.Chicago

# Telescience – *Collaborative Engineering*

- Links computation and data management to unique, expensive instrumentation
- Requires advanced visualization tools for segmentation and analysis of the data
- Provides critical database of biological structure info for neuroscientists



**3D Model of the Node of Ranvier**

# *Az "álom" alkalmazás*

Távoli víz. Los Angelesben

Távoli víz. és megfigyelés Berlinből

Távoli megfigyelés és monitorozás Ferihegyről

Előző szimuláció eredményeinek víz.-ja egy bécsi kávéházból

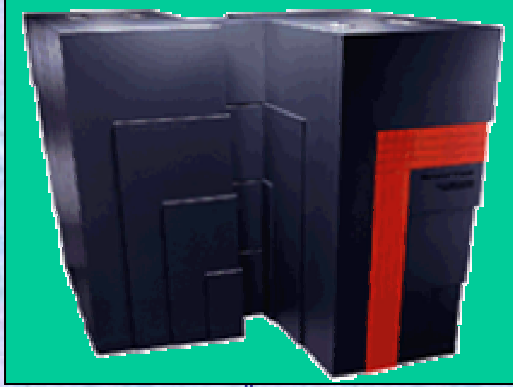
Ekvi-felületek

HDF5

http

Globus

T3E: Jülich (Németo.)

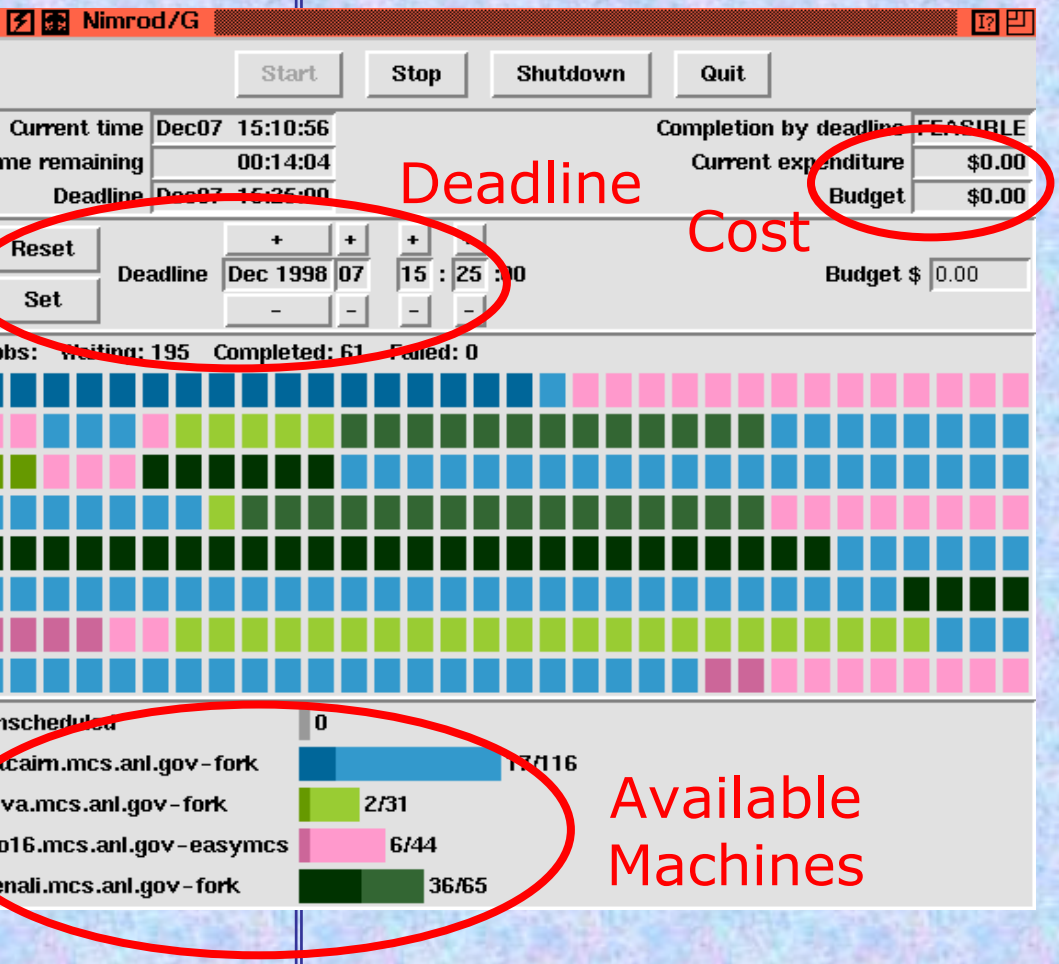


Origin: NCSA (USA)

1. Cactus web portálról indító szimuláció (Róma)

Grid kiterjesztésű Cactus futása elosztott szuperszámítógépeken

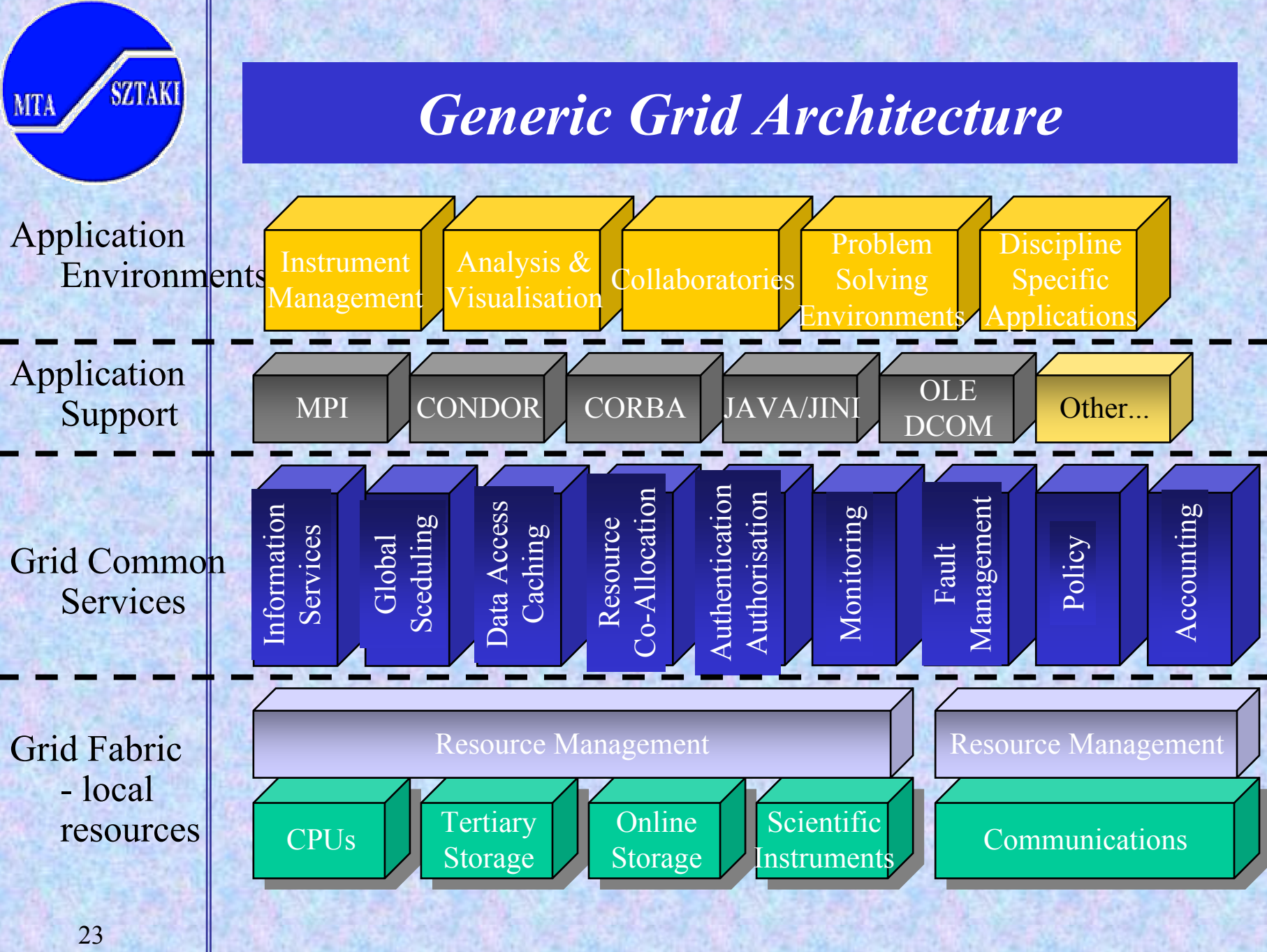
# High-Throughput Computing




- **Schedule many independent tasks**
  - **Parameter studies**
  - **Data analysis**
- **Issues:**
  - **Resource discovery**
  - **Data Access**
  - **Scheduling**
  - **Reservatation**
  - **Security**
  - **Accounting**
  - **Code management**

**Nimrod-G: Monash University**

# Generic Grid Architecture



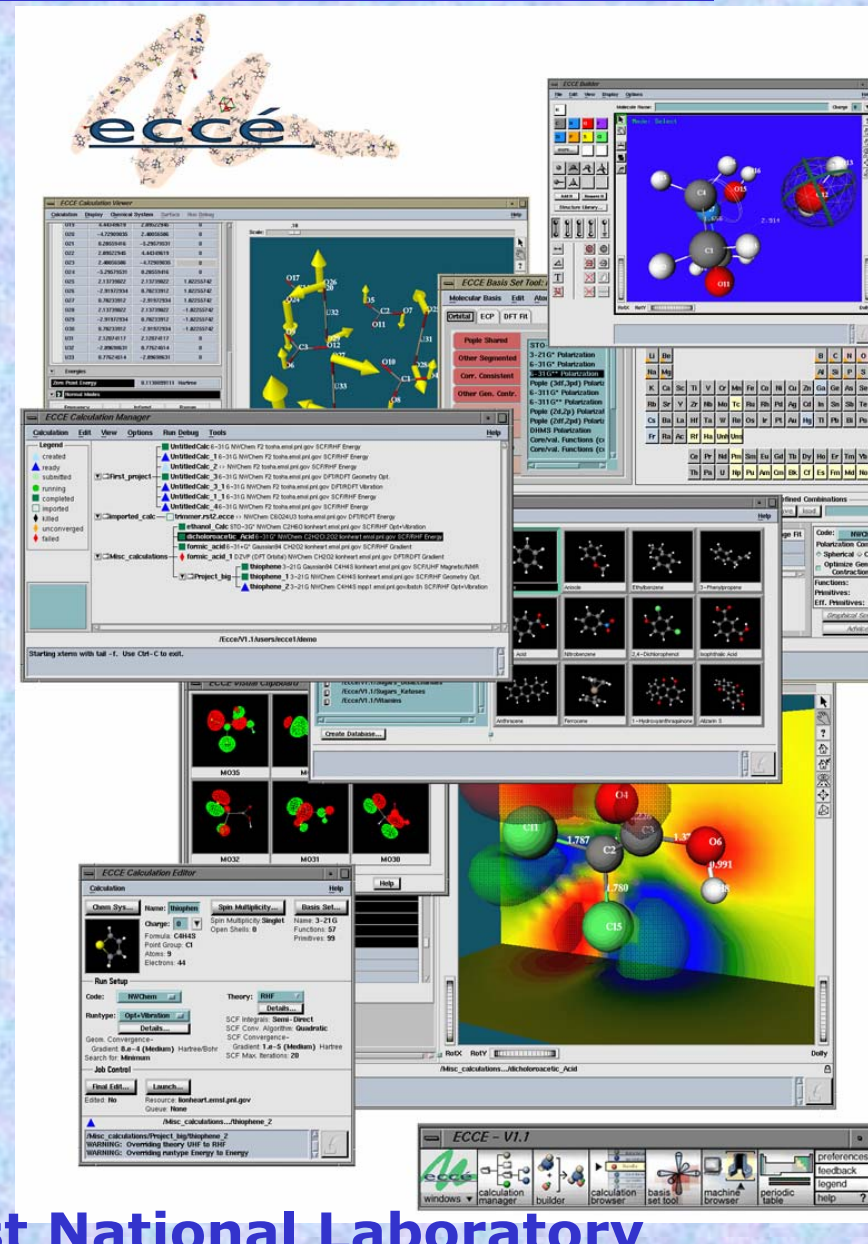
# Problem Solving Environments

## Examples:

- Problem solving env. for computational chemistry
- Application web portals

## Issues:

- Remote job submission, monitoring, and control
- Resource discovery
- Distributed data archive
- Security
- Accounting

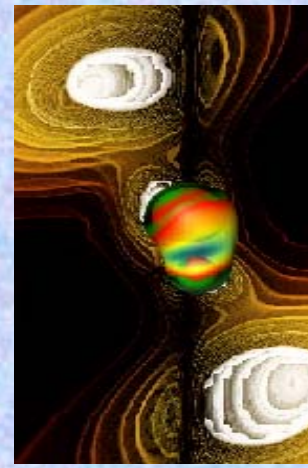


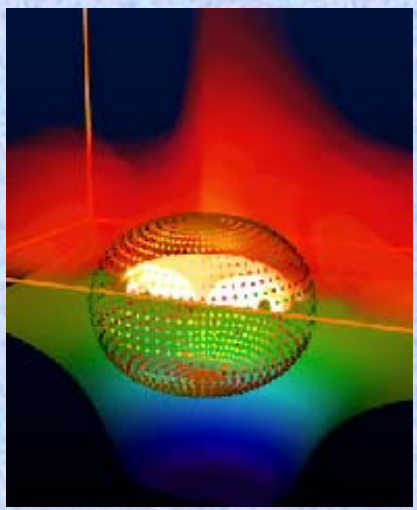


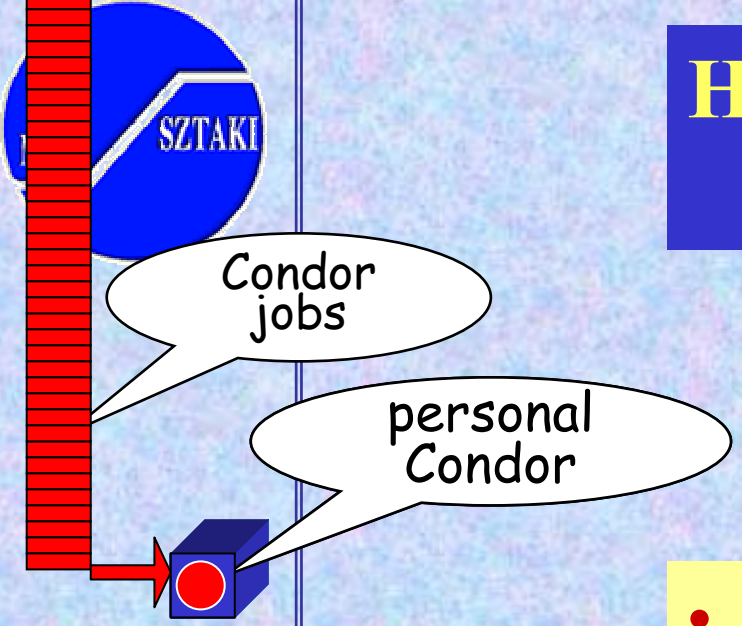

ECCE': Pacific Northwest National Laboratory

# Cactus

- Black Holes (prime source for GW) – from Misner BH collisions to grazing BH collisions with initial spin and momentum (Brandt-Bruegmann initial data)
- Gravitational Waves (Evolution of Brill Waves, collapse of pure GW, investigation of critical amplitude, i.e. when do black holes form)
- Neutron Stars (NASA Neutron Star Grand Challenge, GR hydrodynamics, neutron stars colliding to black holes,...)



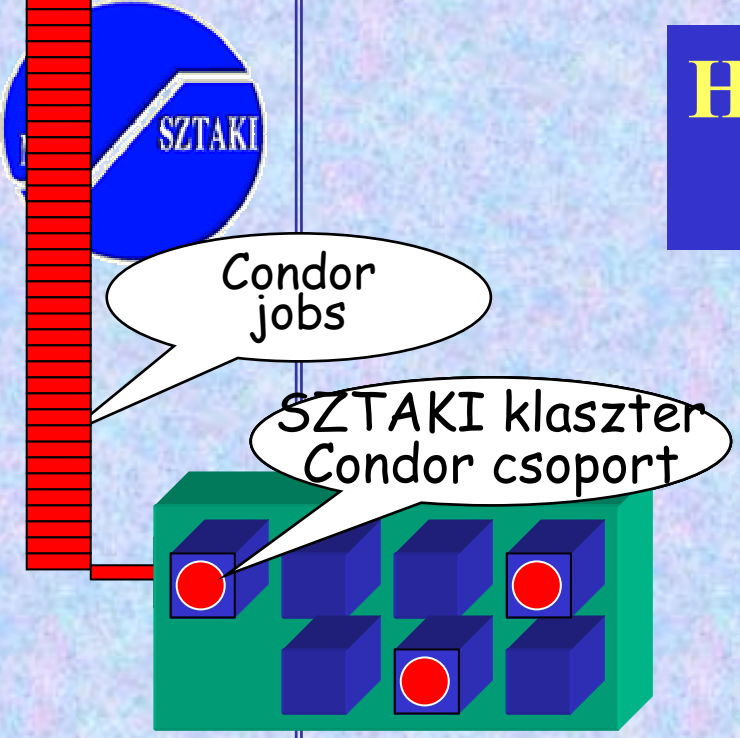





# High throughput Computing: Condor

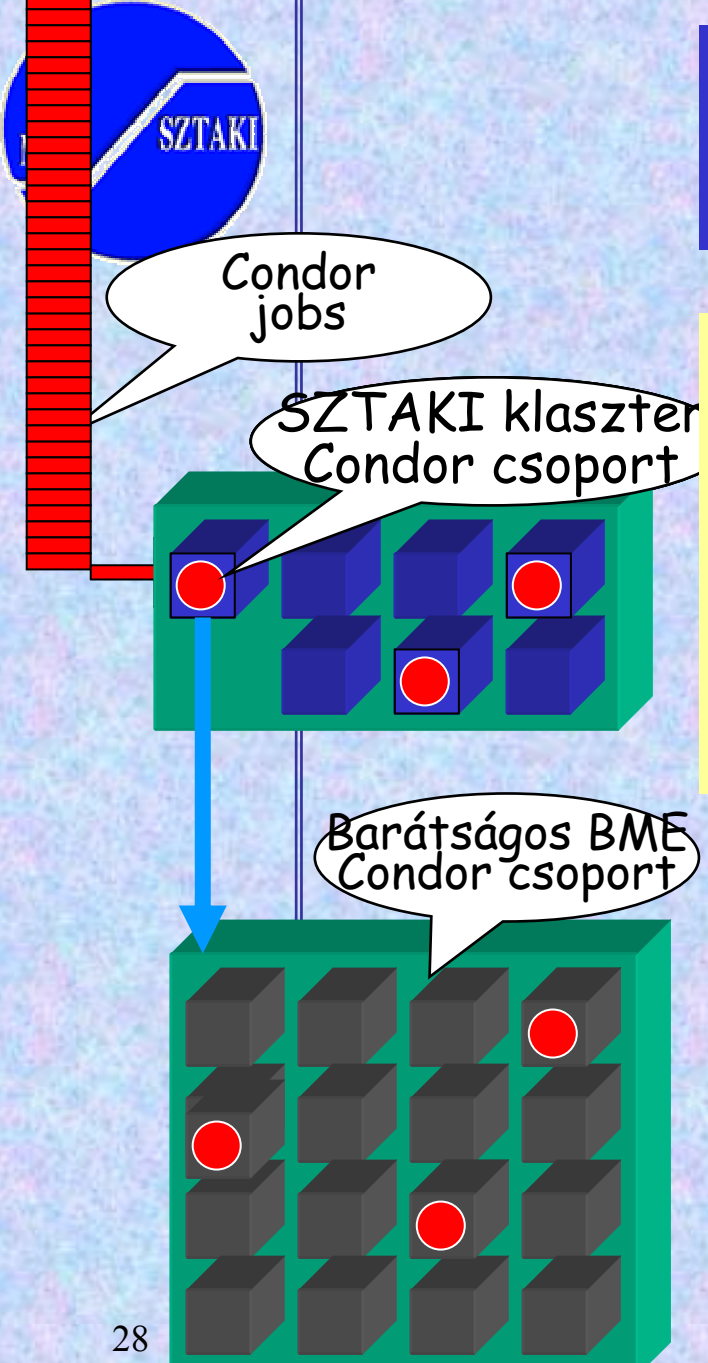
- Cél: A gridben lévő számítógépek szabad ciklusainak kihasználása
- Megvalósítási lépések (1): A személyes PC v. munkaállomás átalakítása személyes Condor géppé

# High throughput Computing: Condor



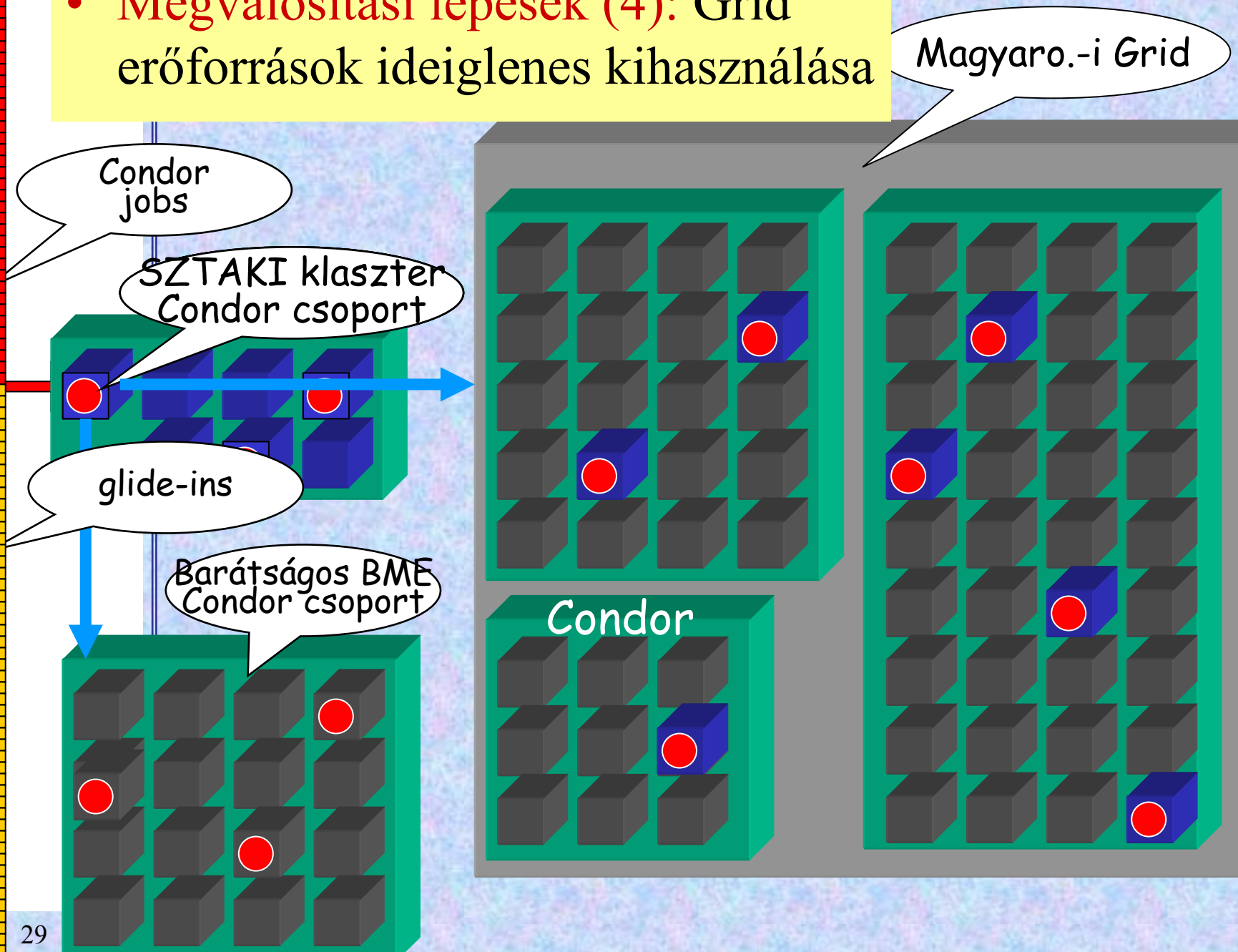

- Megvalósítási lépések (2): Intézeti Condor csoport létrehozása

# High throughput Computing: Condor



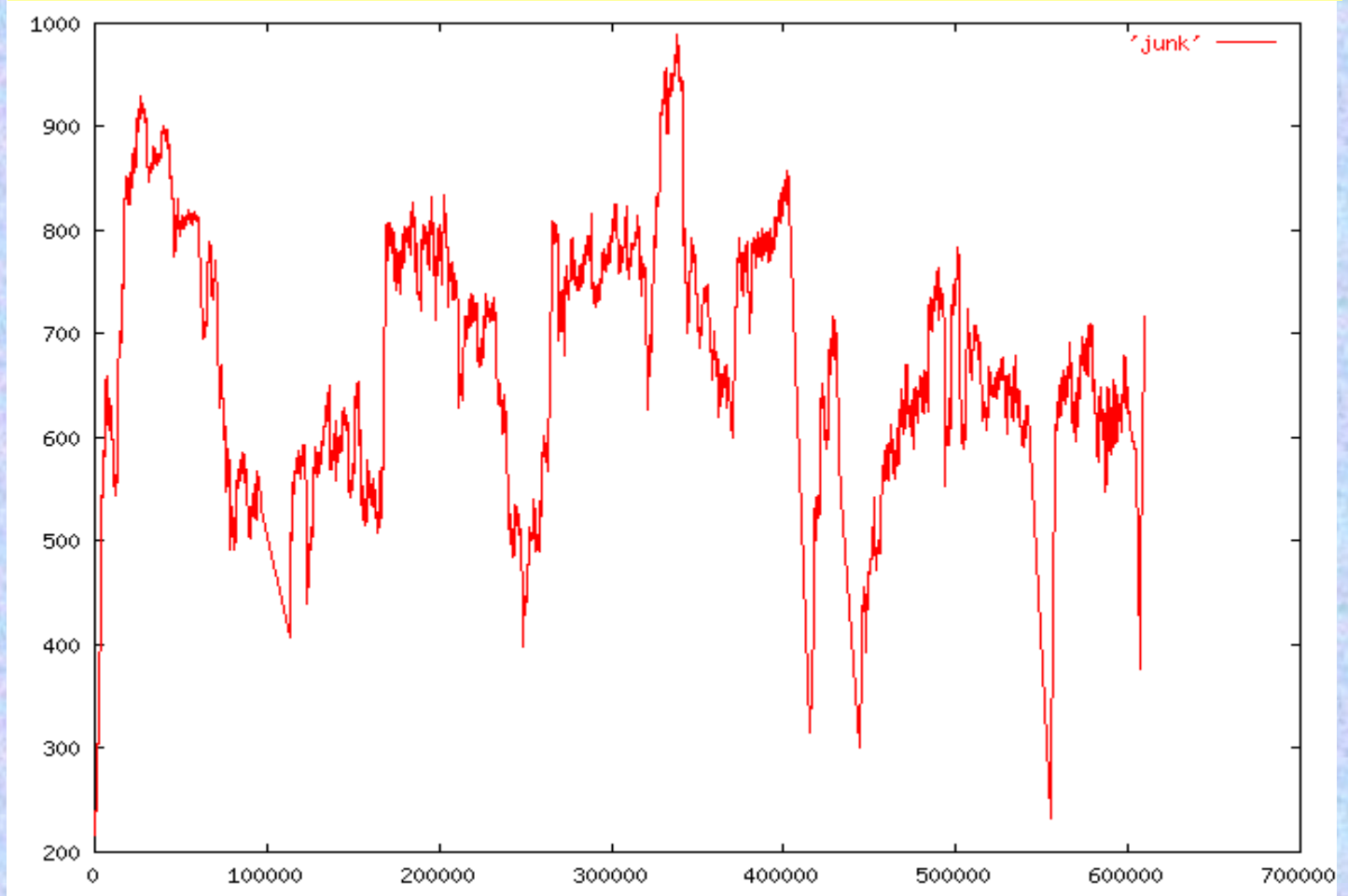

- Megvalósítási lépések (3): Intézeti Condor csoport összekapcsolása más "barátságos" Condor csoportokkal.

- Megvalósítási lépések (4): Grid erőforrások ideiglenes kihasználása



# *NUG30 kvadratikus allokálási probléma*

- Megoldva 7 nap alatt 10.9 év helyett
- Az első 600K másodperc …




Processzorok száma

# *A NUG30 kísérletben alkalmazott számítógépek*

**Barátságos Condor csoportok:**

-- the main Condor pool at U. of Wisconsin (600 processors)

-- the Condor pool at Georgia Tech. U. (190 Linux boxes)

-- the Condor pool at UNM (40 processors)

-- the Condor pool at U. of Columbia (16 processors)

-- the Condor pool at Northwestern U. (12 processors)

-- the Condor pool at NCSA (65 processors)

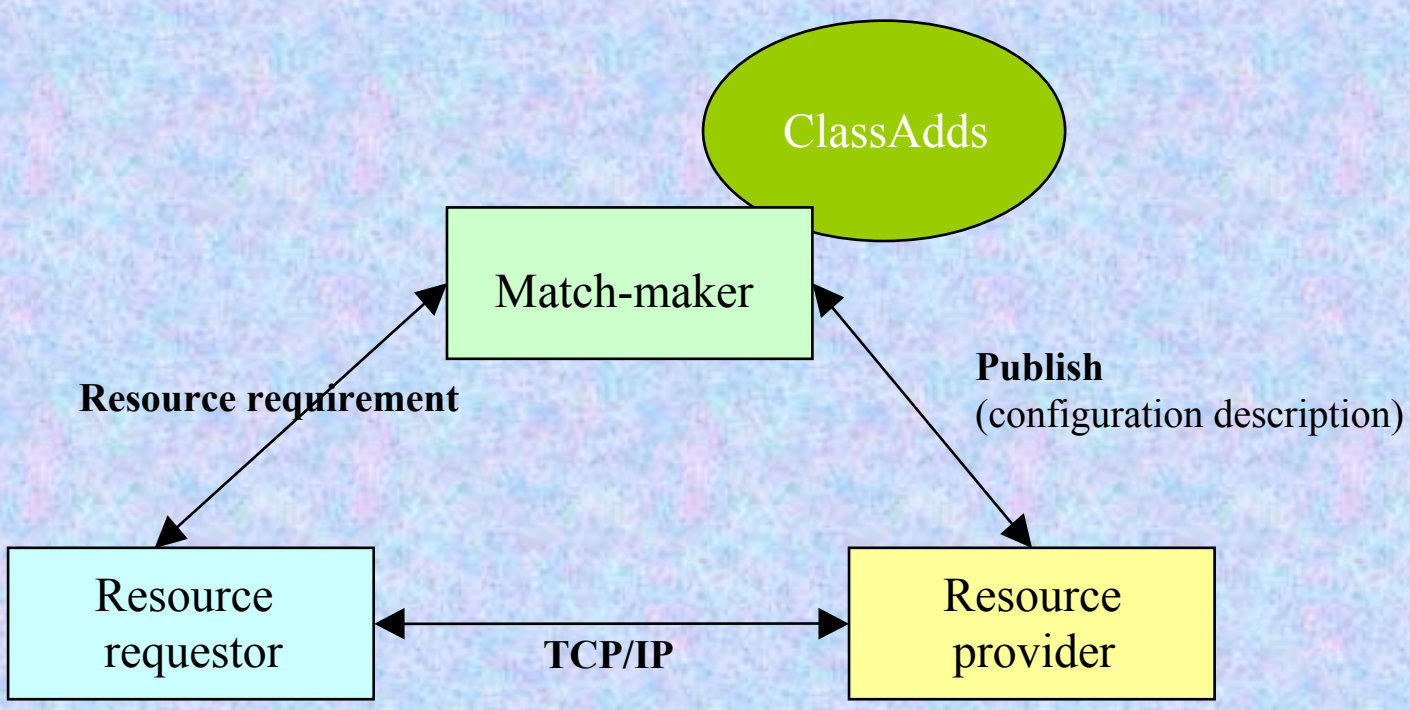
-- the Condor pool at INFN (Olaszo.) (200 processors)

**Grid erőforrások:**

-- Origin 2000 at NCSA

-- Origin 2000 at Argonne National Lab

# The Condor model




**Your program moves to resource(s)**

**Security is a serious problem!**

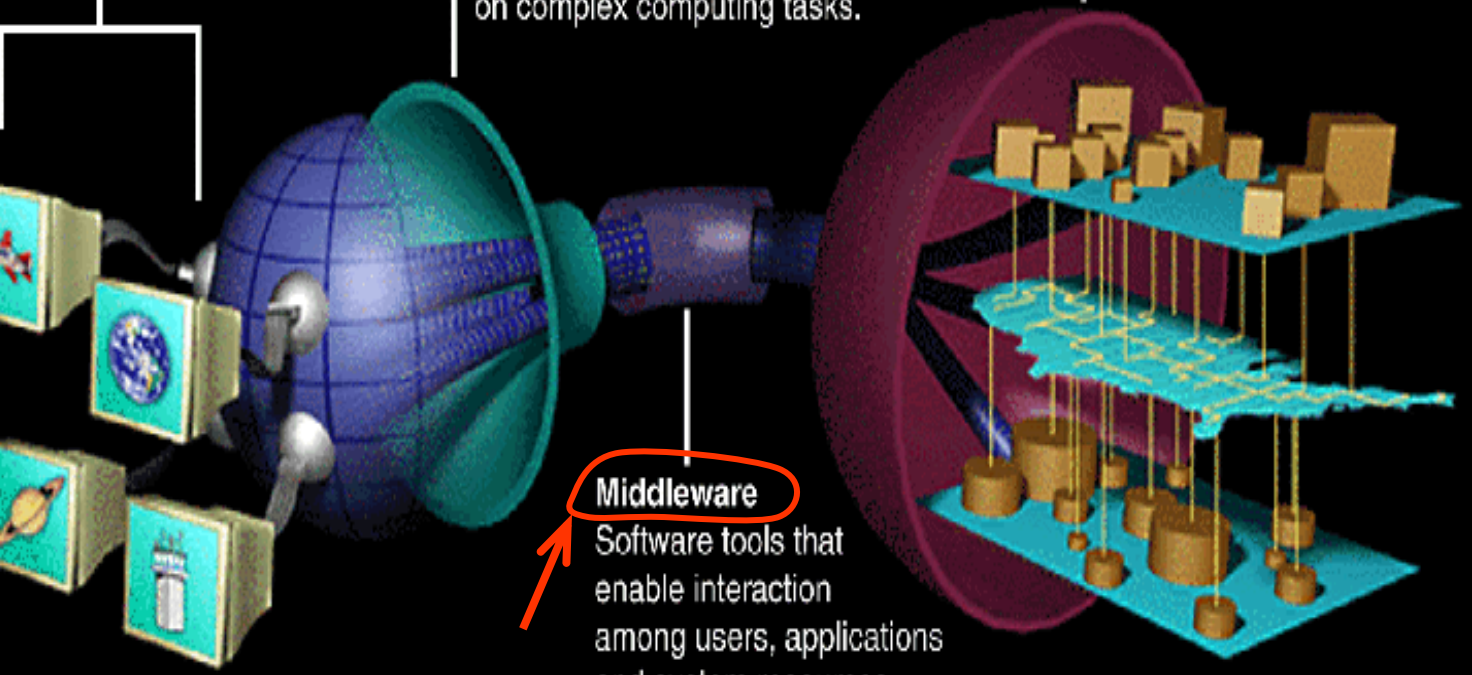
# *GRID/metaszámítási rendszerek létrehozásának komponensei*

- Programozási modell
- Architekturális komponensek
- Middleware
- Programozási környezet
- Ütemezés és erőforrás kezelés
- Kommunikációs rendszerek és protokollok
- Rendszer problémák megoldása (pl. biztonsági kérdések)

**System Users**
Scientists and engineers
using computation to
accomplish Lab missions.

**Intelligent Interface**
A knowledge-based environment
that offers users guidance
on complex computing tasks.

**Cluster Operating System**
The software which coordinates
the interplay of computers,
networks and storage.



**Middleware**
Software tools that
enable interaction
among users, applications
and system resources.

**Supercomputing**
Heterogeneous collection
of high-performance
computer hardware and
software resources.

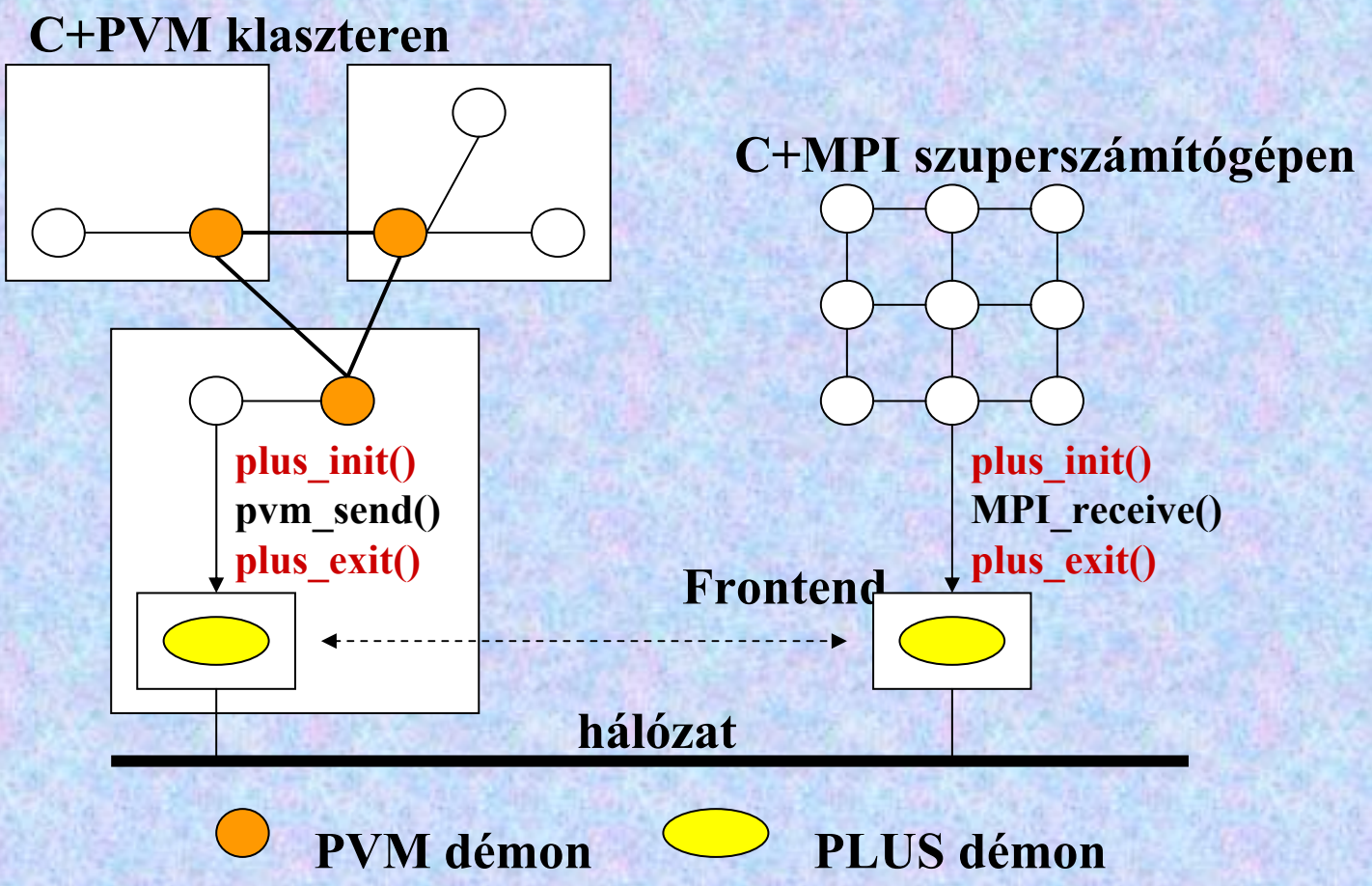
**Networking**
The hardware and
software that permits
communication among
distributed users and
computer resources.

**Mass Storage**
A collection of devices
and software that allow
temporary and long-term
archival storage of
information.


# *Programozási modellek*

- **Üzenet alapú**
  - PVM, MPI, PVMPI, Globus, PLUS
- **Makro adatfolyam (data-flow)**
  - Webflow, Legion
- **Objektum-orientált modellek**
  - Java, CORBA, DCOM, Legion, Globe
- **Intelligens mobil agensek**
  - applet, servlet

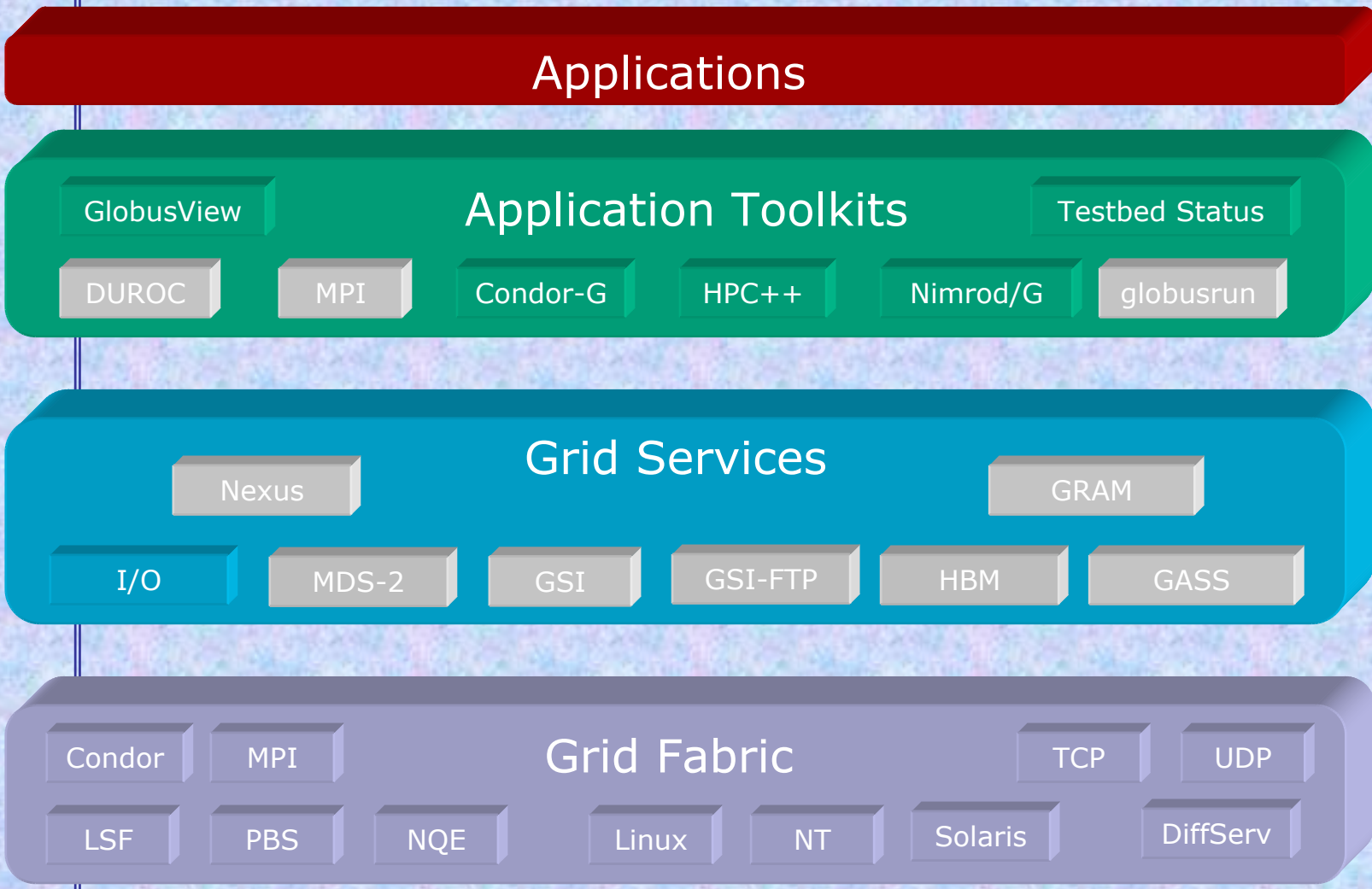
# Processz kommunikáció
# PVM és MPI között PLUS segítségével



# *Middleware koncepciók*

- Hol helyezkedik el a middleware?
- A middleware célja:
  - Egy radikálisan heterogén környezetet virtuálisan homogén rendszerré alakítani
- Három fő koncepció:
  - Toolkit (mix-and-match) modell
    - Globus
  - Objektum-orientált modell
    - Legion, Globe
  - Internet-www modell
    - Webflow

# Globus Layered Architecture



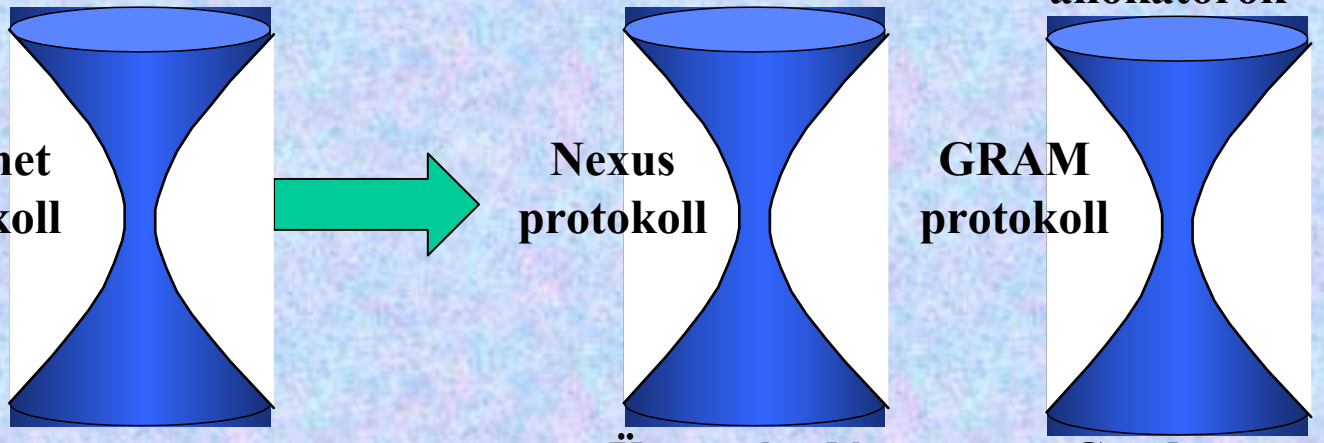
# A Globus homokóra koncepciója

## Magasszintű szolgáltatások

TCP, FTP, HTTP, stb.

MPI, CC++, HPC++, CORBA, stb.

Erőforrás brókerek, Erőforrás ko-allokátorok



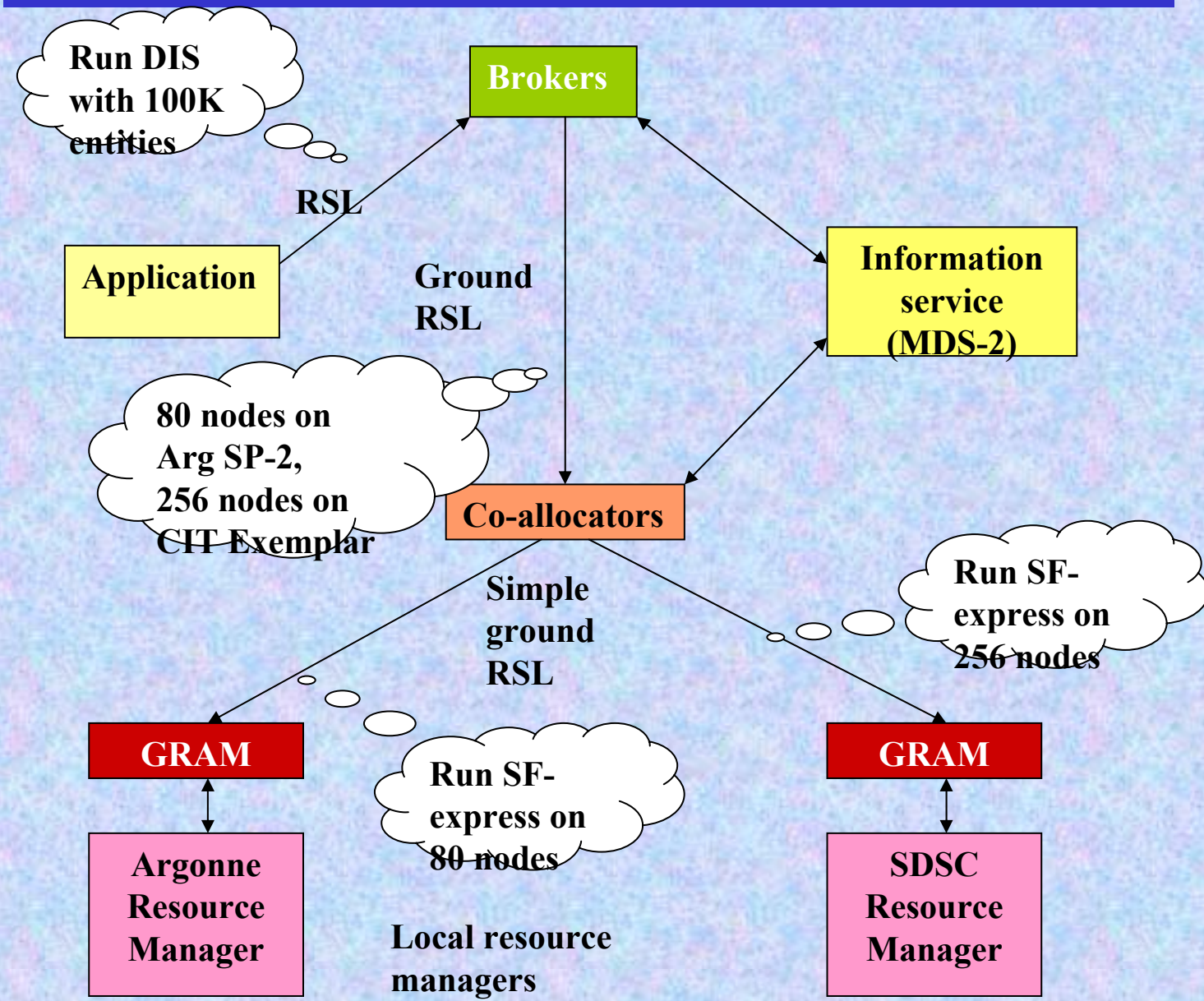

Ethernet, ATM, FDDI, stb.

Üzenetátadás, közös memória, IP, stb.

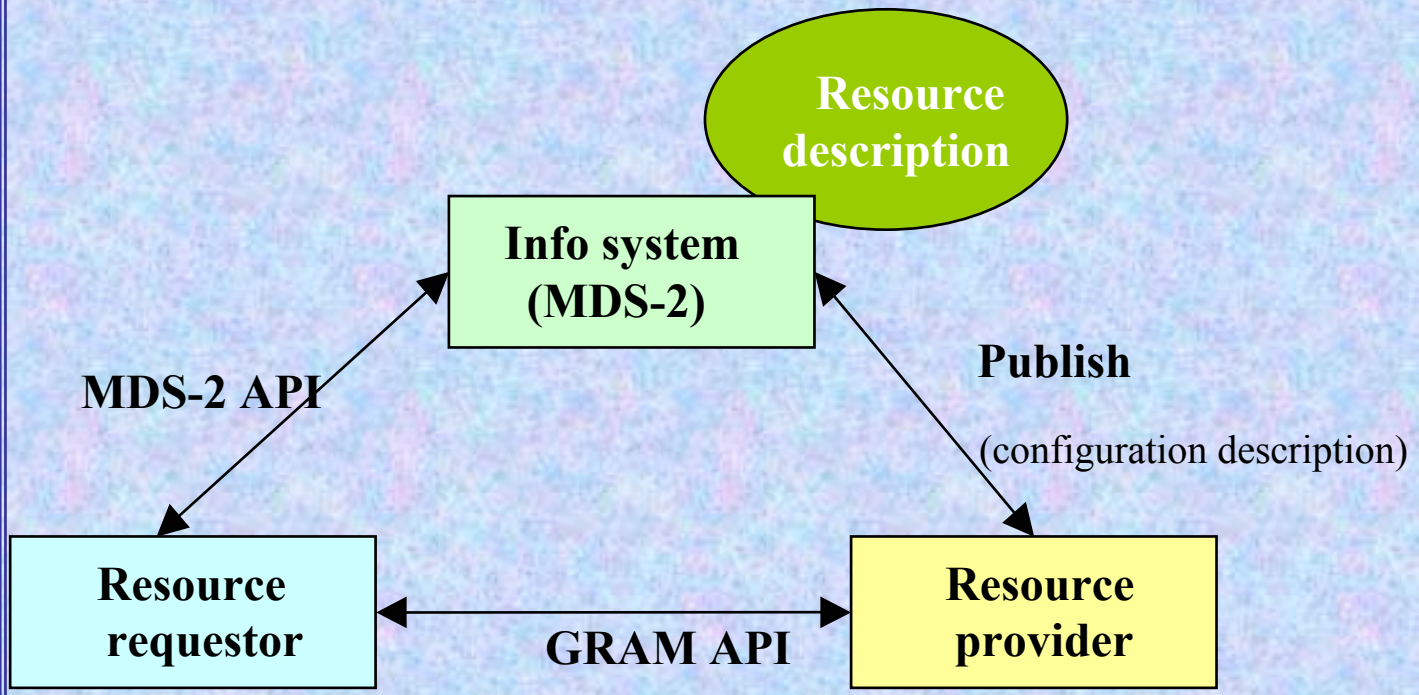
Condor, LSF, NQE, LoadLeveler, stb.

## Alacsonyszintű eszközök

# Globus hierarchical resource management architecture



# The Globus Model

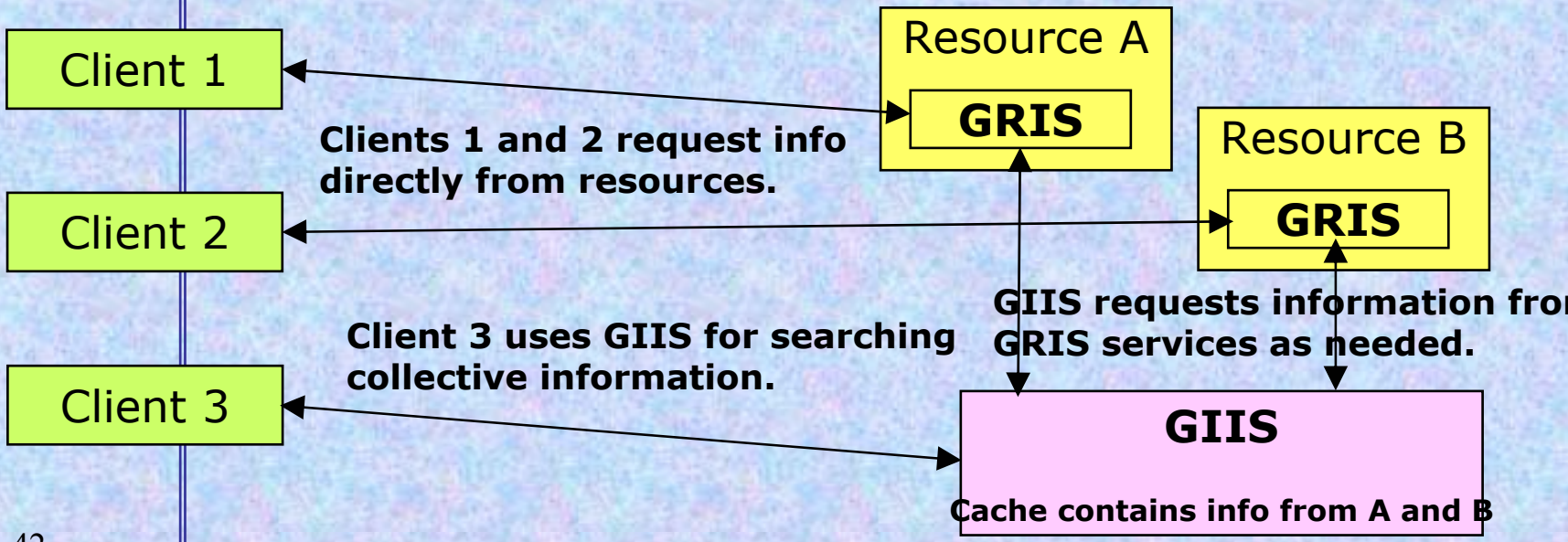



**Your program moves to resource(s)**

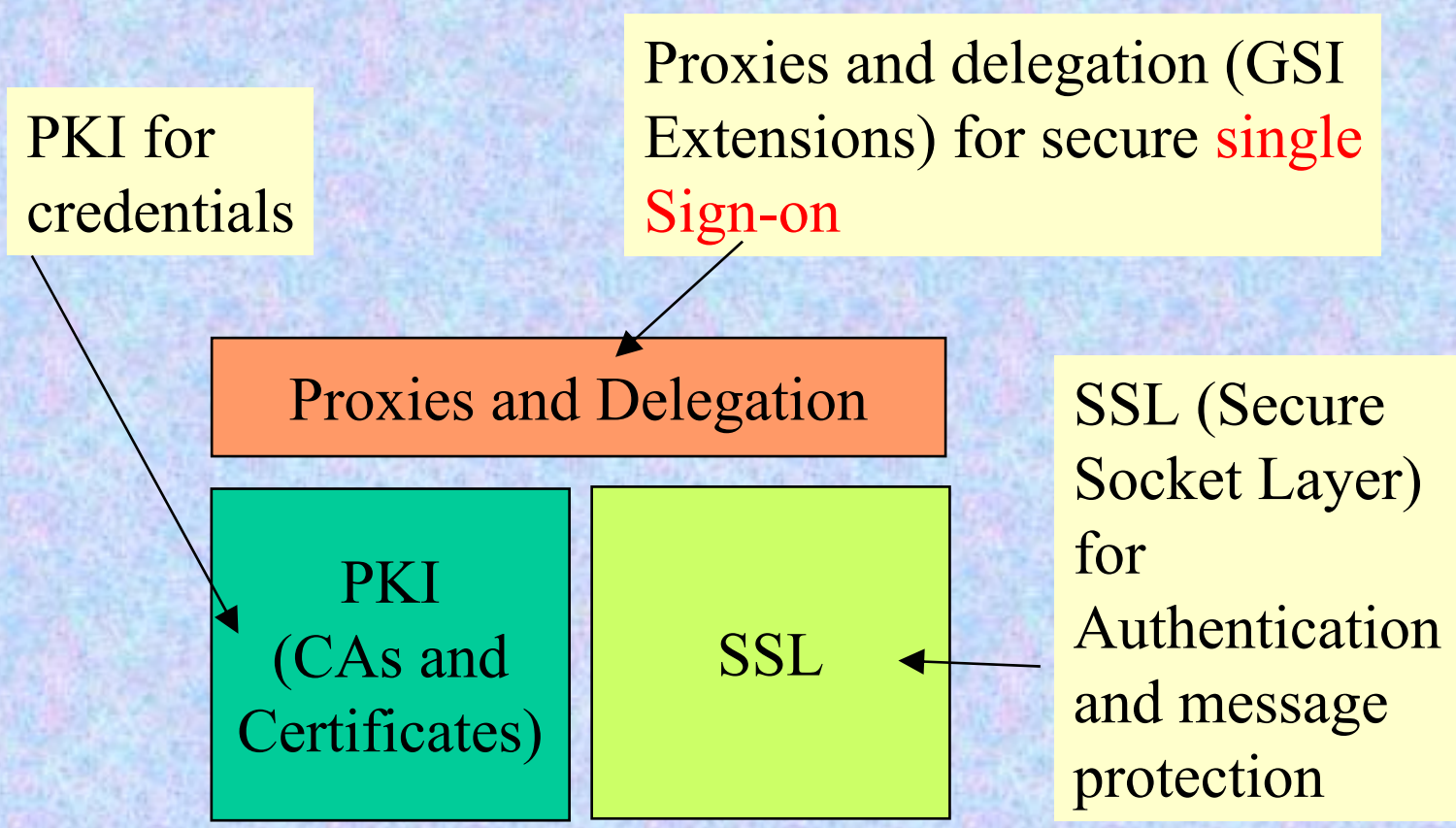
**Security is a serious problem!**

# *"Standard" MDS Architecture (MDS-2)*

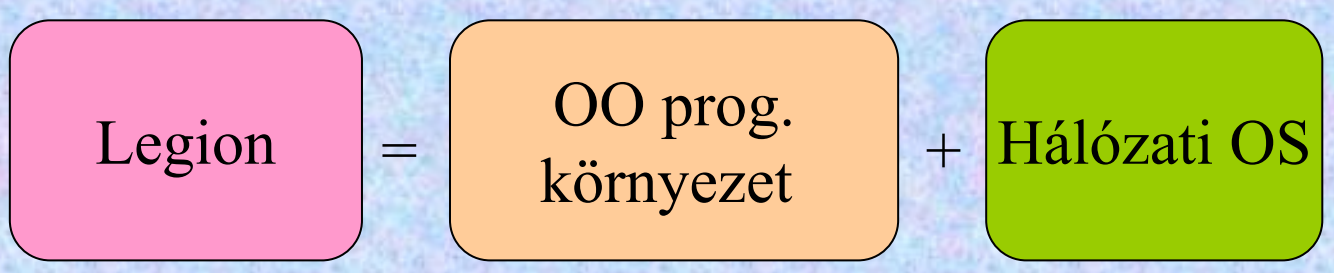
- Resources run a standard information service (**GRIS**) which speaks LDAP and provides information about the resource (no searching).
- **GIIS** provides a "caching" service much like a web search engine. Resources register with GIIS and GIIS pulls information from them when requested by a client and the cache is expired.
- **GIIS** provides the collective-level indexing/searching function.



# *Grid Security Infrastructure (GSI)*

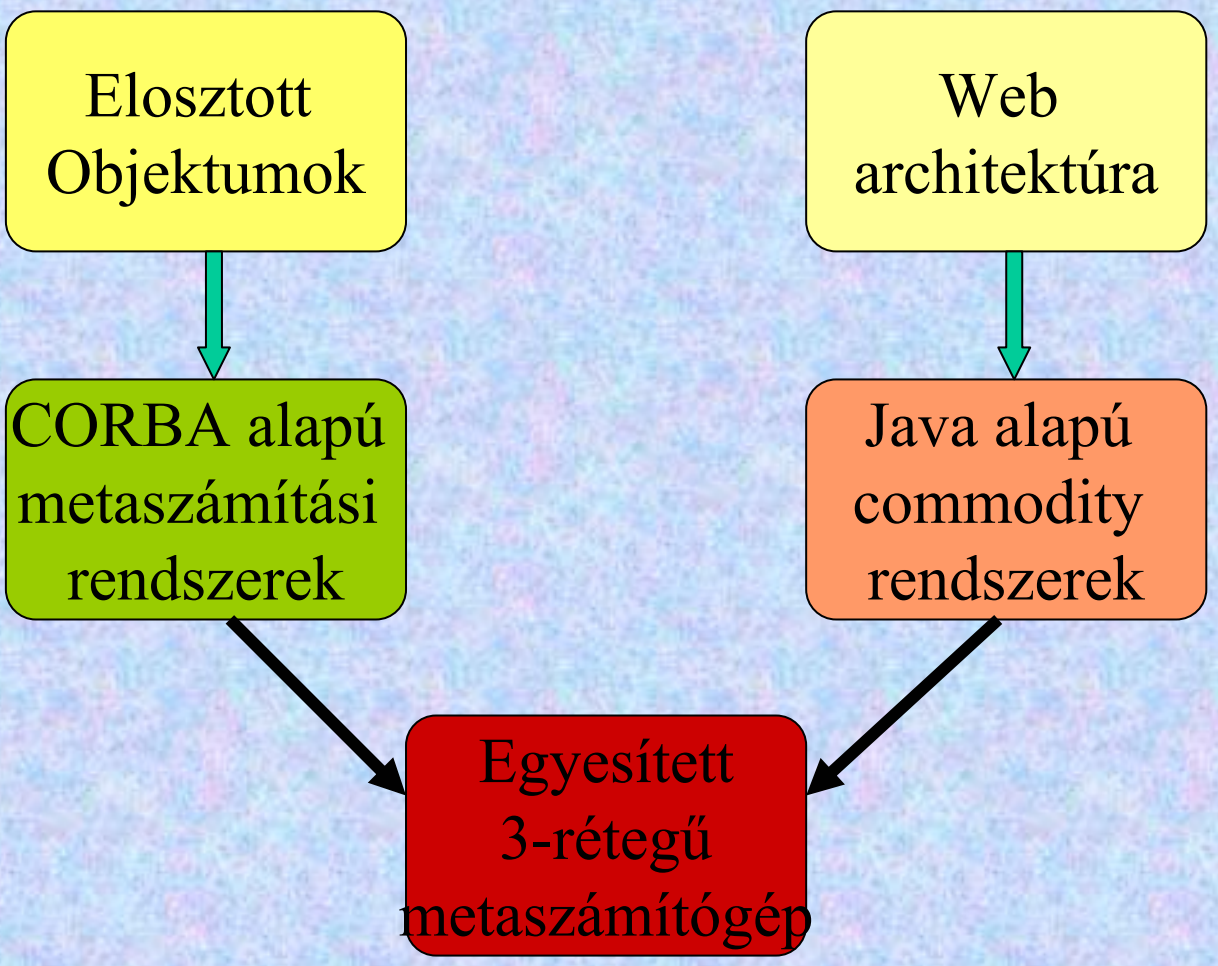


# *Legion*

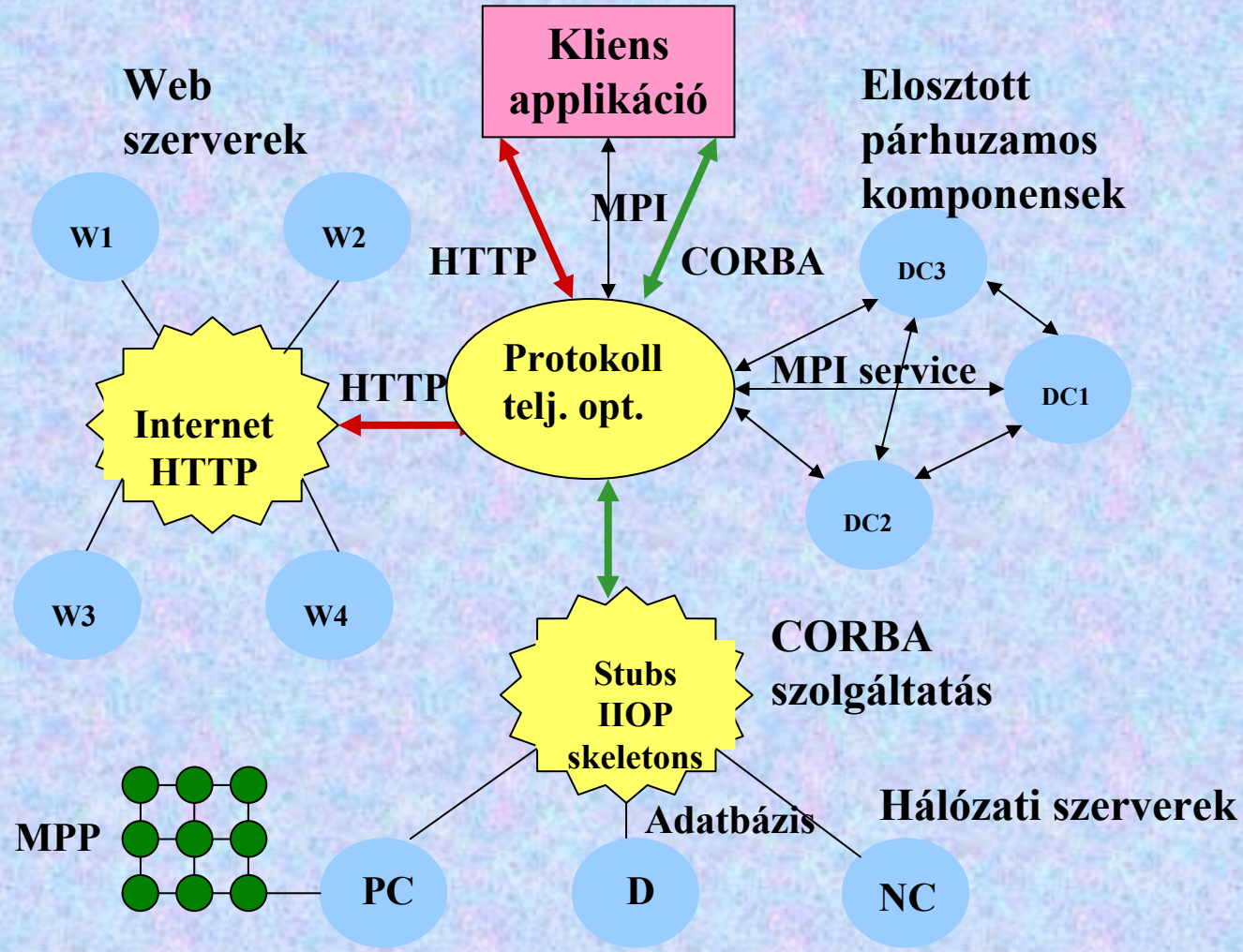



- Magasszintű funkciókat épít egy egyszerű **unifikált objektum modellre** alapozva.
  - Az objektumok processzek, amelyek a munkaelosztás, ütemezés és erőforráskezelés alanyai
- A software IC concepciót valósítja meg a **makro-dataflow** technikára alapozva

# *Commodity 3-rétegű architektúrák*



# Egyesített 3-rétegű metaszámítógép



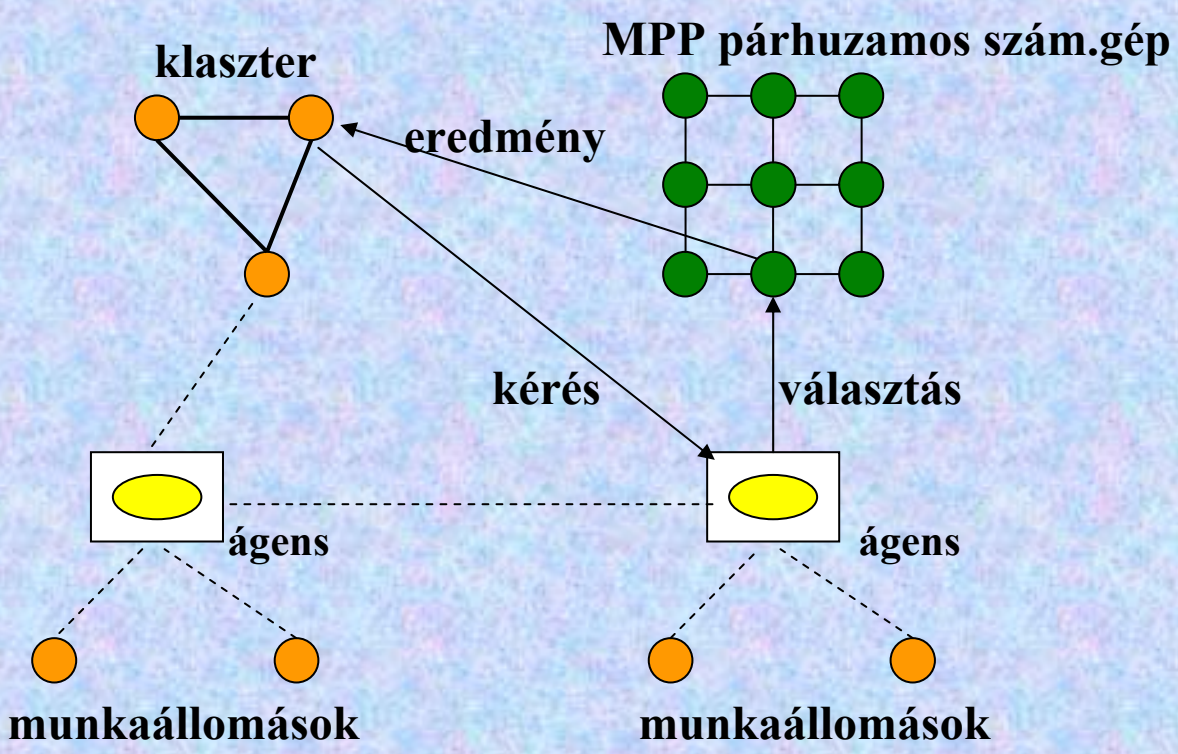
# *Programozási környezetek*

- **Toolkit alapú**
  - Cactus
  - Paderborn toolkit:
    - PLUS kommunikációs környezet
    - Resource and Service Description (RSD)
- **Integrált környezetek**
  - P-GRADE GRID verziója
- **Applikáció specifikus környezetek**
  - NetSolve

# *NetSolve*

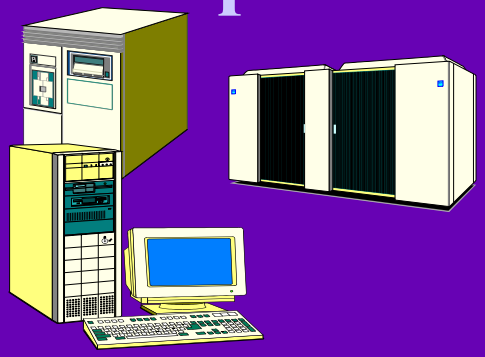
A GRID hálózatban rendelkezésre álló erőforrások kihasználása numerikus számításokhoz

Fő koncepció: kliens - szerver - ágens



# NetSolve

## Computational Resources



Hardware: Software:

Clusters Routines

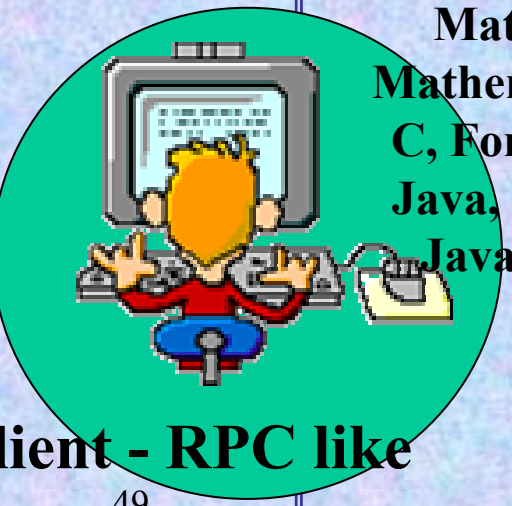
MPP Libraries

Workstations Applications

Globus,Condor, MPI,PVM

Reply

Choice



Matlab
Mathematica
C, Fortran
Java, Excel
Java GUI

lient - RPC like

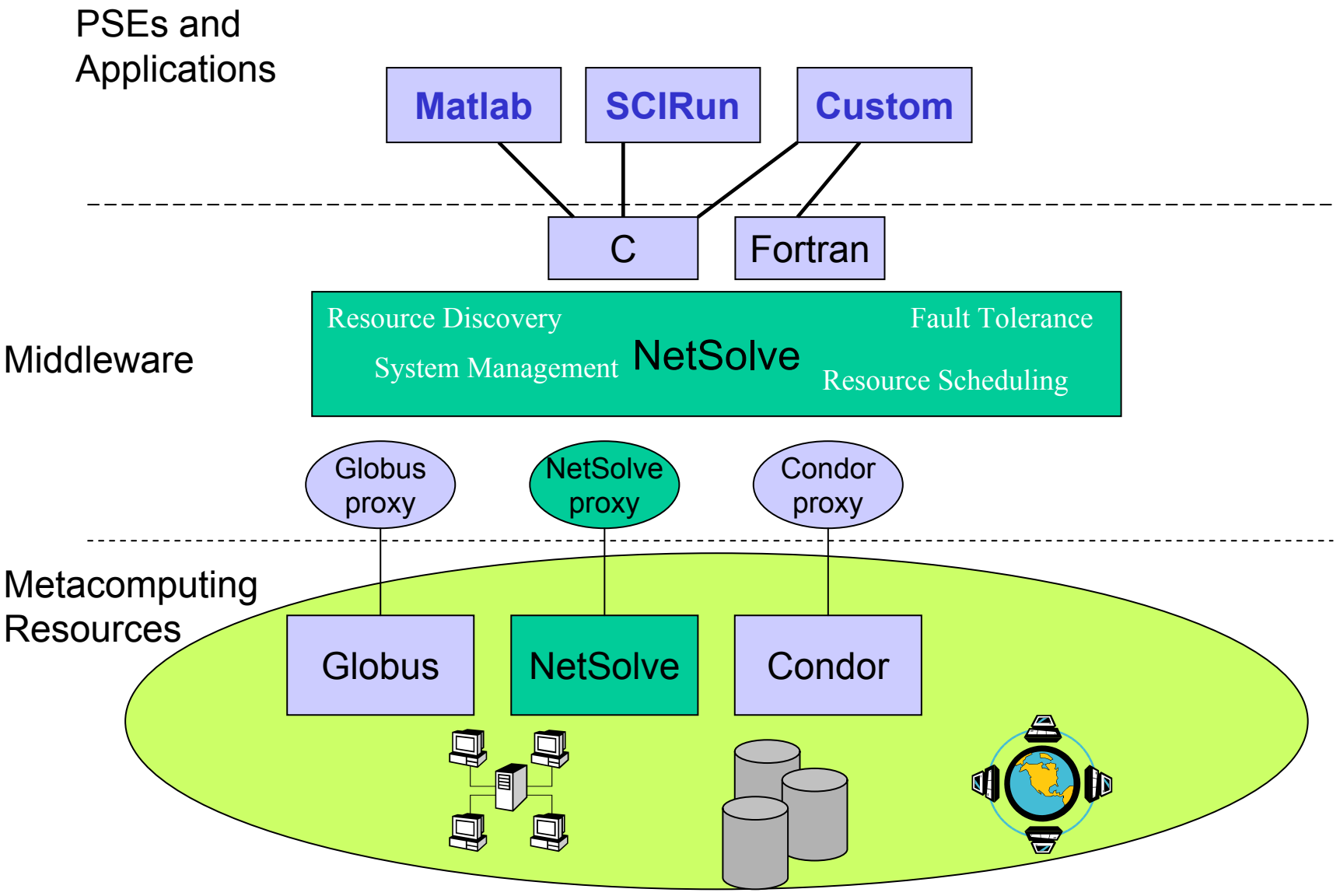
Request

Agent

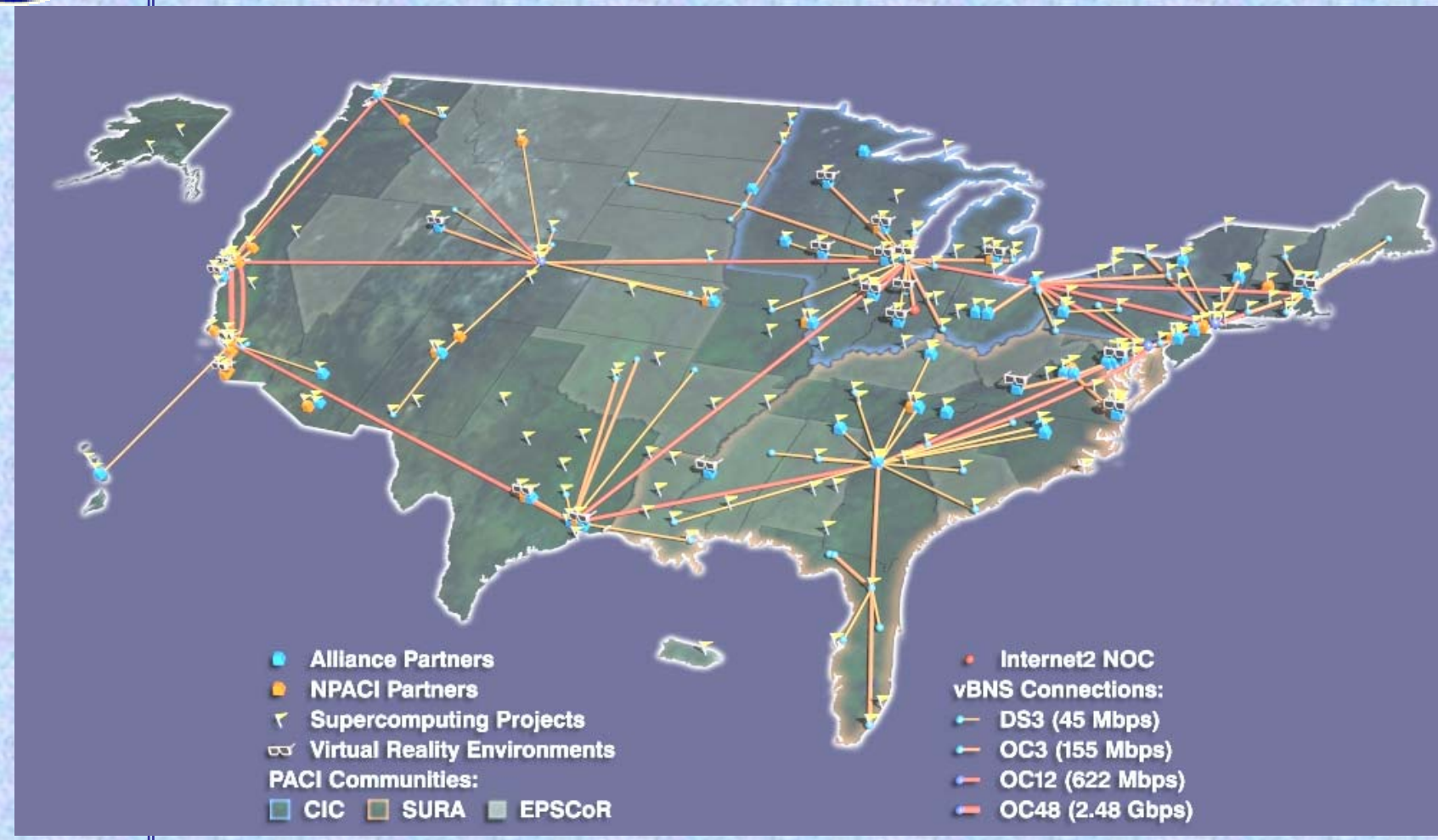
Scheduler

Database

# NetSolve Infrastructure

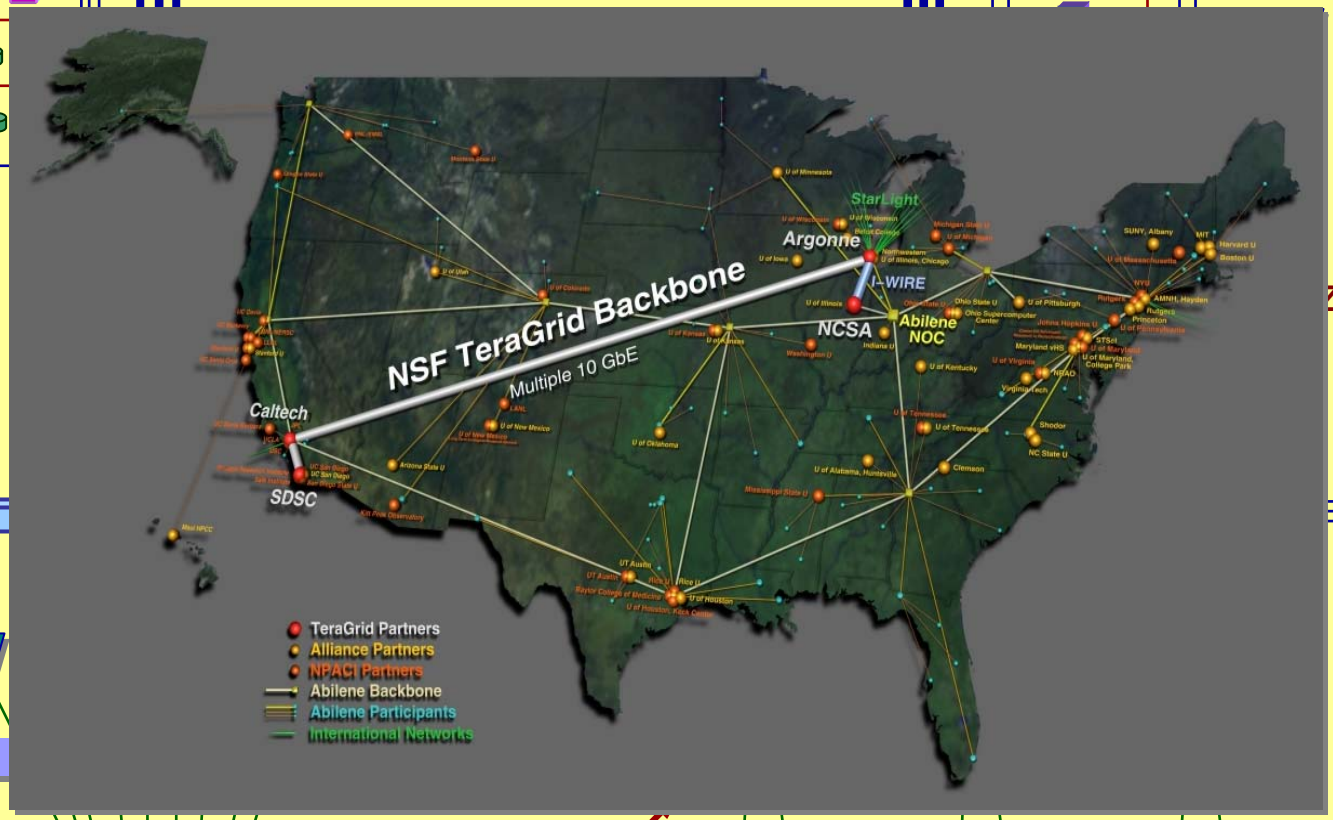


# *Hol tartott az USA 2000-ben?*




NSF National Technology Grid

# Hol tart ma az USA? - TeraGrid




56p HP X-Class
28p HP V2500
2p IA-32
HPSS

574p IA-32 Chiba City
128p Origin
HR Display & VR Facilities
HPSS

HPSS
176p IBM SP lue Horizon
Sun E10K

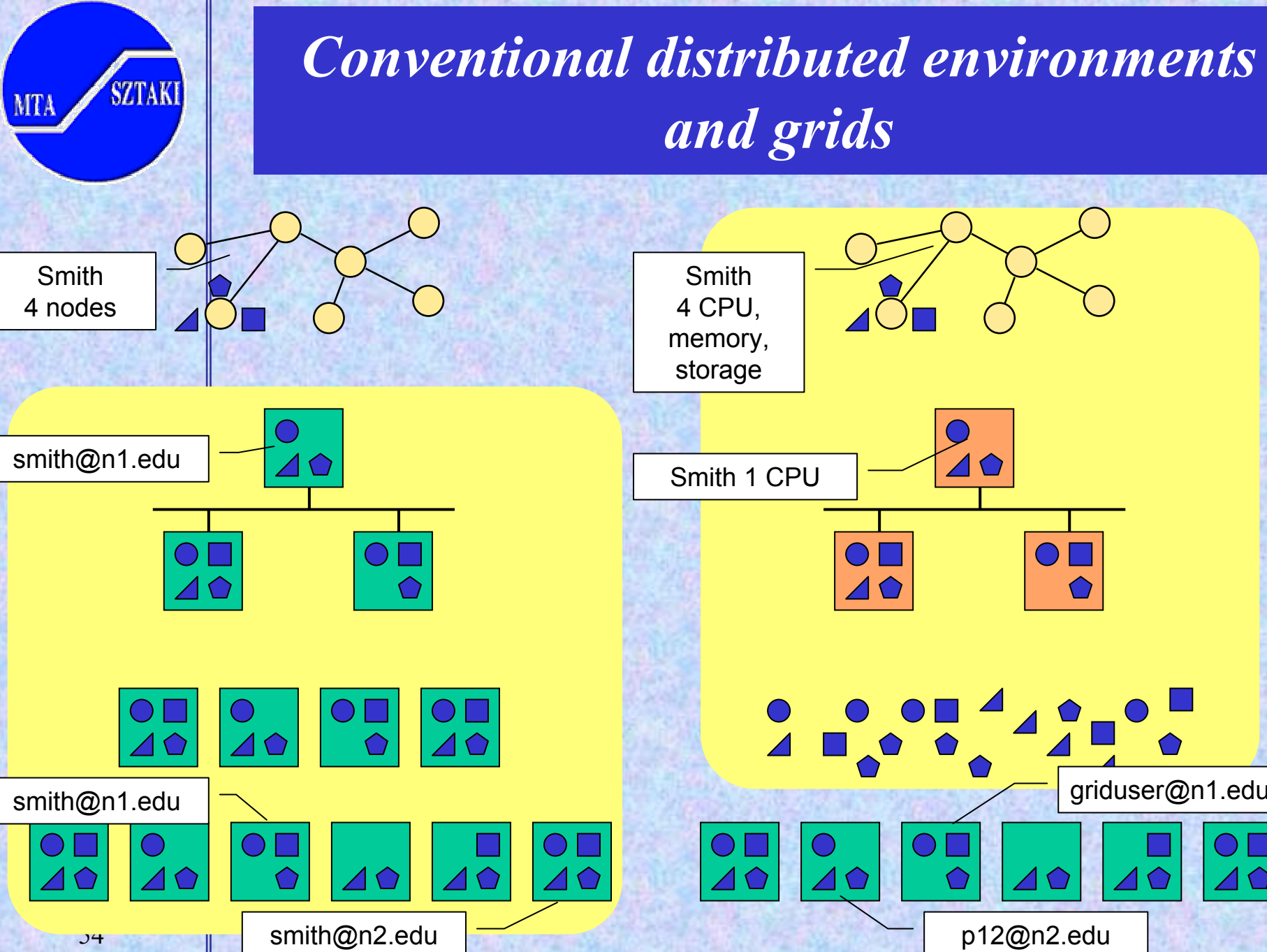
UniTre
1024p IA-
320p IA-
1500p Origin

**NCSA**: Compute-Intensive

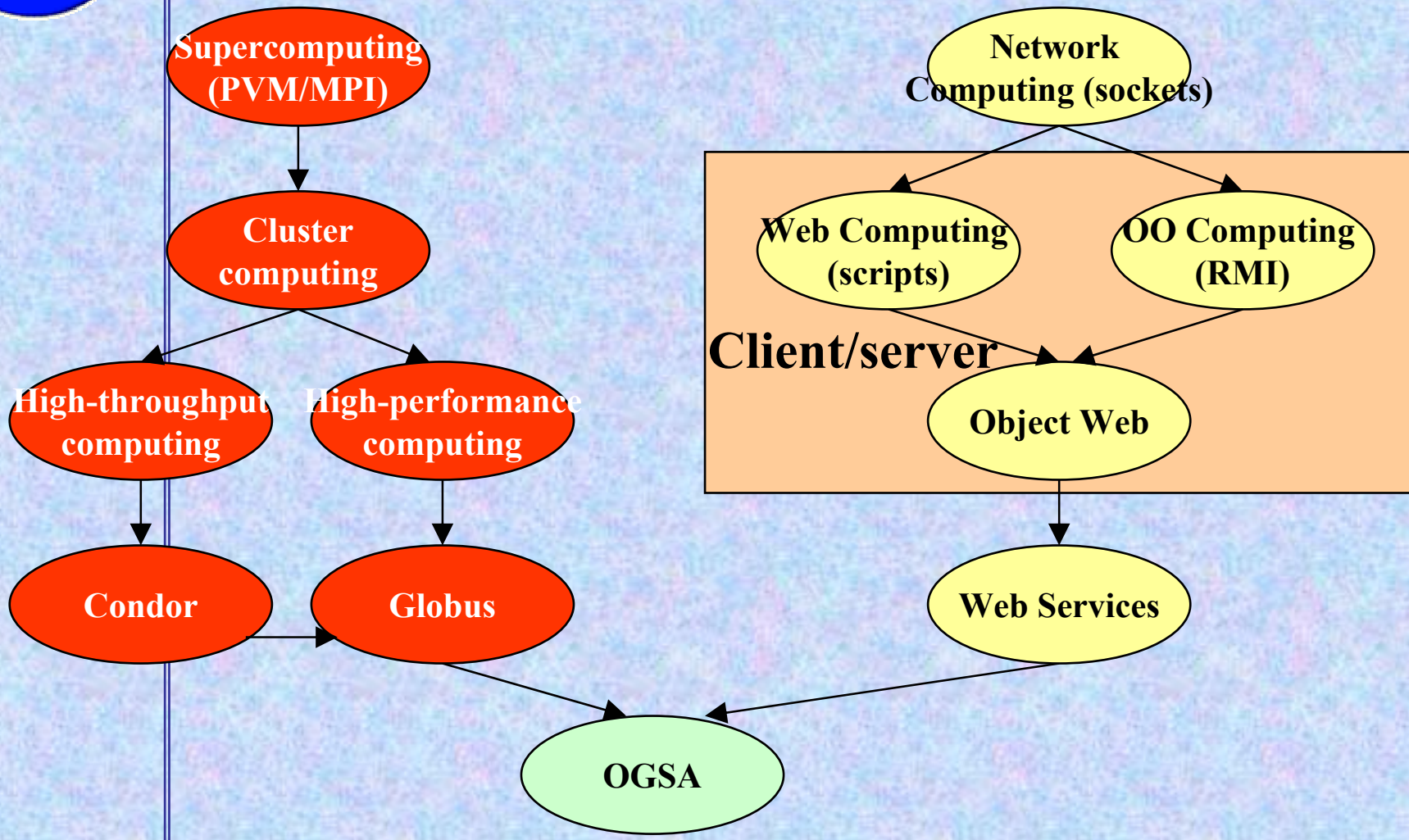
# *Fundamental grid functionalities*

- The essential grid functionalities are:
  - **Resource abstraction**
    - Physical resources can be assigned to virtual resource needs (matched by properties)
    - Grid provides a mapping between virtual resource needs and physical resources
  - **User abstraction**
    - User of the physical machine may be different from the user of the virtual machine
    - Grid provides a temporal mapping between virtual and physical users

# Conventional distributed environments and grids

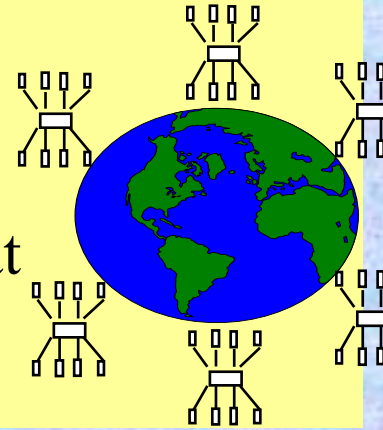


# Grid rendszerek fejlődése



# *Összefoglalás*

- A GRID/metaszámítógép egy új technológia, amely integrálja:
  - a szuperszámítógép technológiát
  - a távolsági hálózatok technológiáját
  - a WWW technológiát



- **A Grid egy új, az eletromos áramot elosztó hálózathoz hasonló infrastruktúra kialakulásához fog vezetni**
- **Ez az új infrastruktúra a www technológiához hasonlóan óriási hatással lesz az informatikai társadalom továbbfejlődésére**


# *Szuperszámítógépek, klaszterek és metaszámítógépek összehasonlítása I.*

| | Supercomputer | Cluster | NOW | Metacomputing system |
|---|---|---|---|---|
| Processing units (nodes) | Microprocessors | PCs, workstations | PCs, workstations | Supercomputers, clusters, PCs, workstations |
| Number of nodes | 100 - 1000 | 10 - 100 | 10 - 100 | 100 - 10000 |
| Communication network | Buses, switches | LAN | LAN | Internet |
| Node OS | Homogeneous | Typically homogeneous | Typically heterogeneous | Heterogeneous |
| Inter-node security | Nonexistent | Rarely required | Necessary | Necessary |

# Szuperszámítógépek, klaszterek és metaszámítógépek összehasonlítása II.

| | Supercomputer | Cluster | Metacomputing system |
|---|---|---|---|
| Programming models | • Shared memory<br>• Message passing | • Shared memory<br>• Message passing<br>• Peer-to-peer<br>• Client-server | • Message passing<br>• Client-server<br>• Code shipping<br>• Proxy computing<br>• Intelligent mobile agents |
| Programming language | • HPF<br>• (C/Fortran)+MPI | • HPF<br>• (C/Fortran)+MPI | • HPF<br>• (C/Fortran)+MPI<br>• Java/CORBA |
| Middleware | • No | • Limited forms | • Toolkit approach<br>• Three-tier commodity (Java/CORBA)<br>• Object-oriented |
| Programming environment | • Toolkit approach<br>• Integrated environment | • Toolkit approach<br>• Integrated environment | • Toolkit based<br>• Application specific<br>• Integrated environment |
| Resource allocation | • Mapping<br>• Load balancing | • Mapping<br>• Load balancing | • Resource manager |
| QoS | No | No | Yes |

# *Köszönöm a figyelmüket*

További információ: www.lpds.sztaki.hu