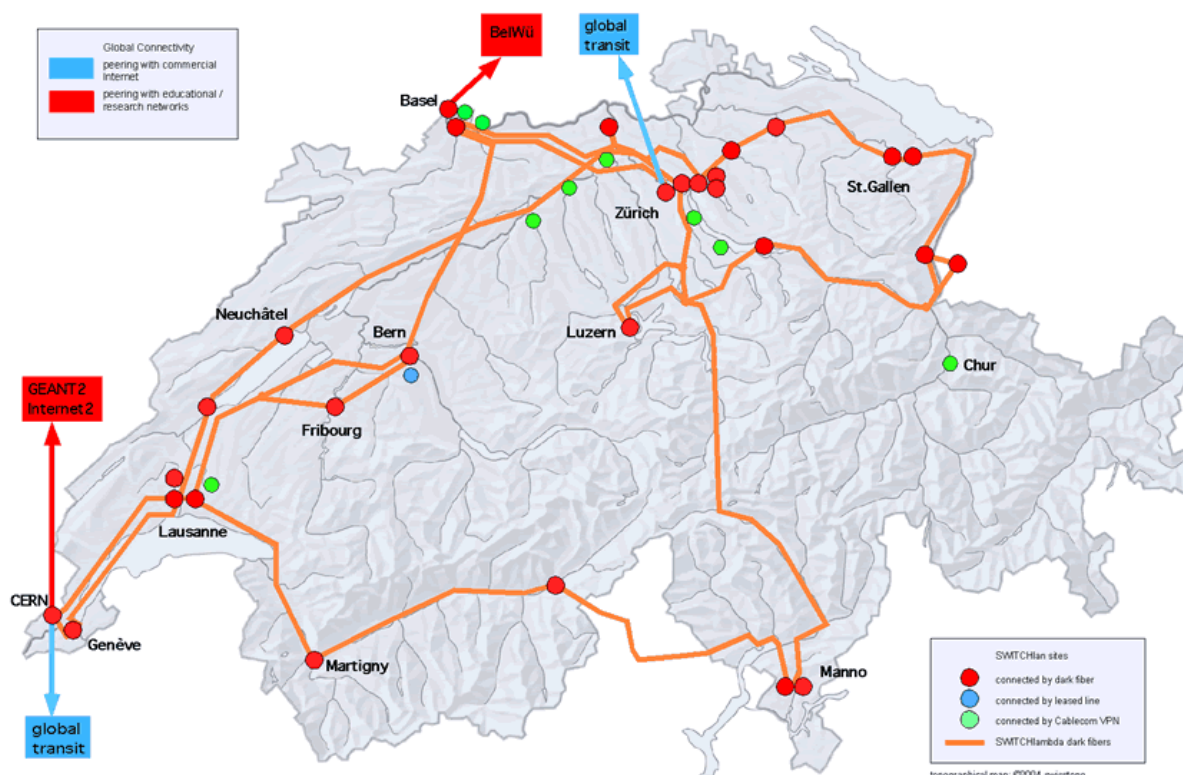


The SwissGrid Initiative



Peter Kunszt

Manager Swiss Grid Initiative

Biomed Grid School Varenna, July, 2006

Peter Kunszt



Doctorate in Theoretical Physics from
the **University of Bern**



Building the Science Database of the
Sloan Digital Sky Survey,
Johns Hopkins University Baltimore



EU Grid Projects, leading data management
middleware development
CERN, Geneva

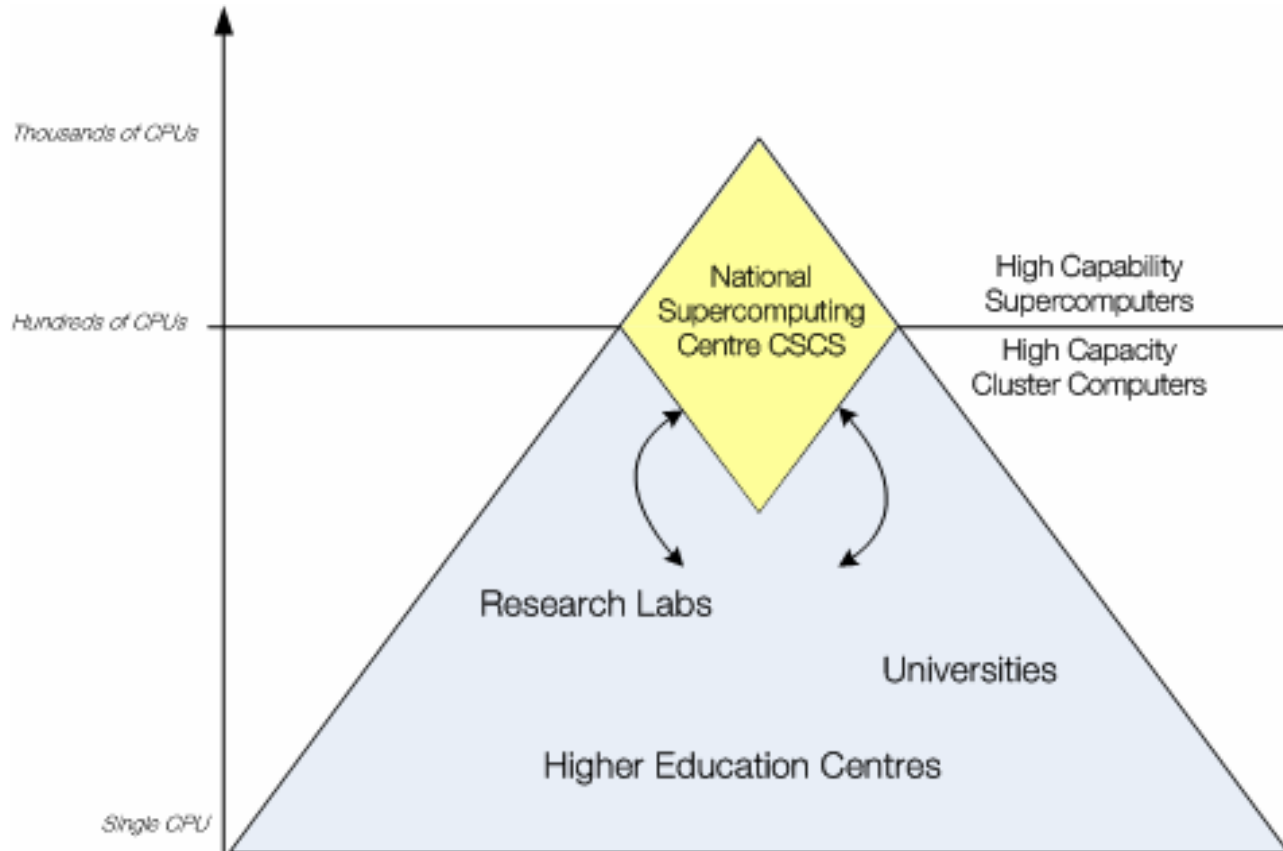


Manager Swiss Grid Initiative,
Swiss National Supercomputing Centre CSCS
Manno



BioInfoGrid Varenna, P. Kunszt





Content

Swiss Grid Initiative

Swiss Involvements in Grid Projects

Case Study: Swiss Bio Grid



Swiss Grid Initiative

Effort taking care of **coordinating and supporting** national Grid projects.

- Point of contact for all Grid Projects, users and administrators
- Point of support for all Grid users and administrators
- Exchange of ideas, knowledge and resources in Switzerland
- Representation of Swiss Academic Research Interests
 - In Europe
 - Globally
 - Towards the Industry



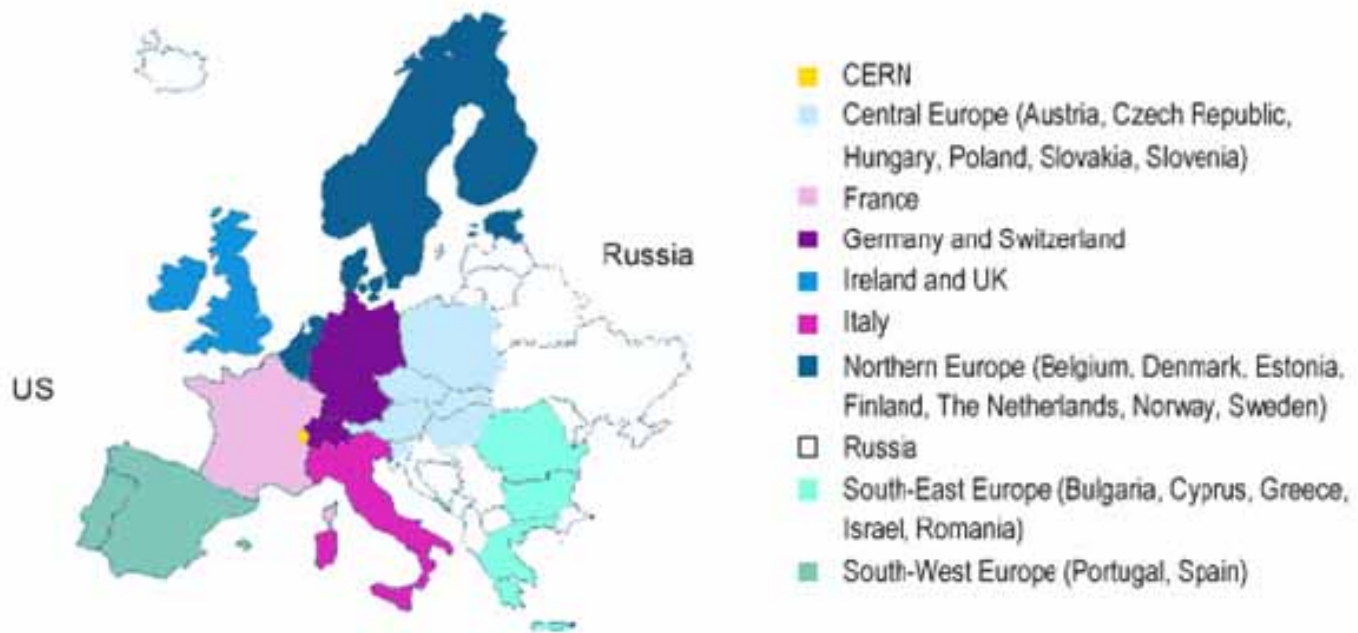
Swiss Grid Initiative Status

- In the process of establishing a legal entity:
Swiss National Grid Association SwiNG
- Working groups for e.g.
 - Funding and political support
 - Infrastructure
 - Applications
 - Security
- Next steps:
 - Establish structure and governance of the association
 - Sustained infrastructure and support

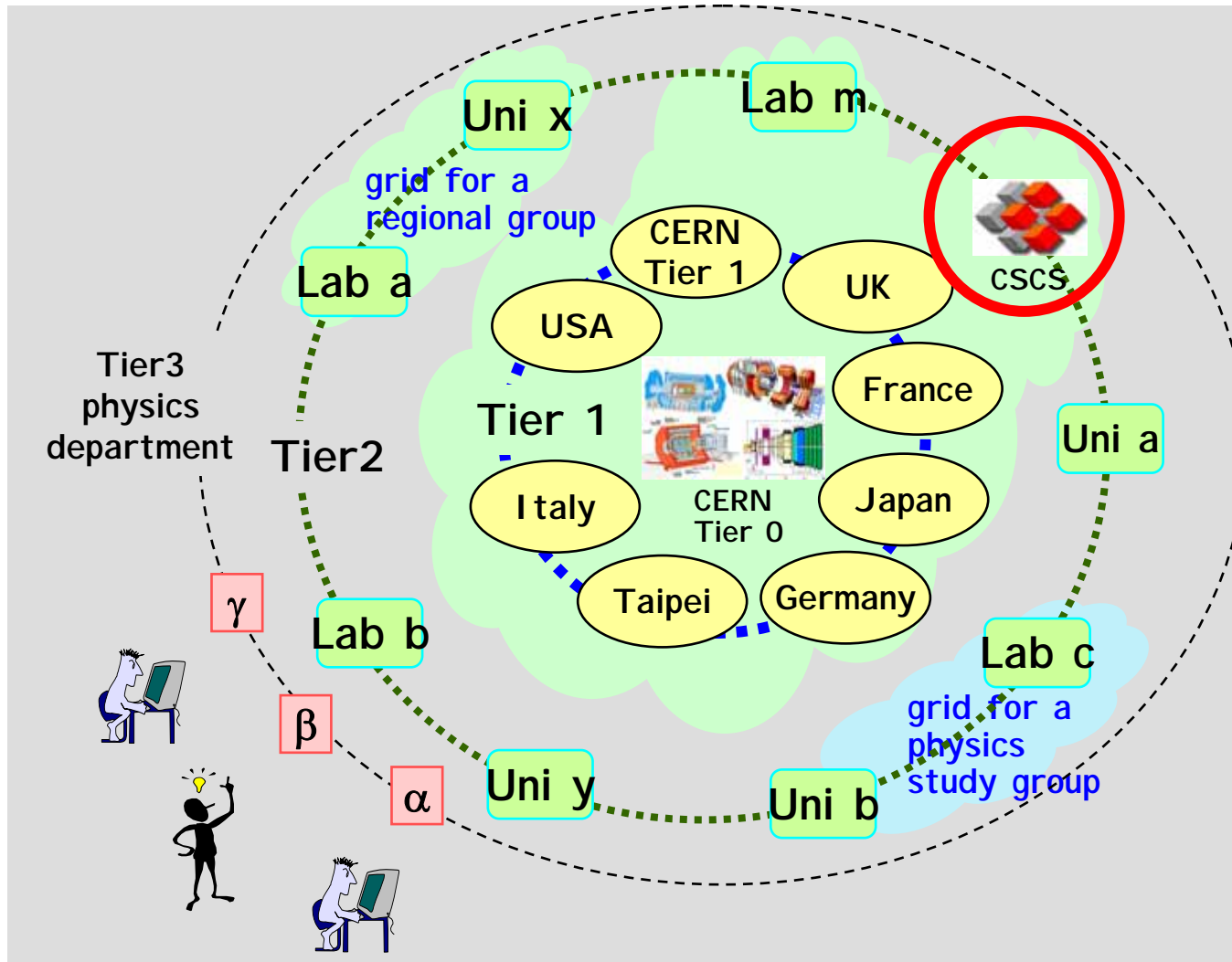


EGEE and LCG

European Grid Infrastructure for Enabling E-science
Teaming up with the D-Grid in the DECH Federation



Tier-ed model



Swiss Partners

CSCS – Swiss Supercomputing Centre

SWITCH – Swiss Research & Edu Network

CSCS

- SA1, NA2, NA3, NA4
- Is an LCG Tier2 Site
- Support for Region and all of EGEE
- Analysis of Physics Data
- Biomed, Comp.Chemistry, EO applications
- Training, Education, Public Relations

SWITCH:

- JRA1
- Security Middleware: Next generation of Grid Certificates by integrating Shibboleth and PKI



SEPAC

SEPAC stands for South European Partnership for Advanced Computing

- SPACI consortium
 - University of Lecce
 - University of Calabria
 - Hewlett-Packard
- CILEA
- CSCS
- ETHZ
- UNIZH



SEPAC Project Scope

- Infrastructure and Technology oriented collaboration
- Exploration of technology and interoperability
- Application portfolio being built

- Building on another Grid Portal: the Grid Resource Broker from the Univ. of Lecce



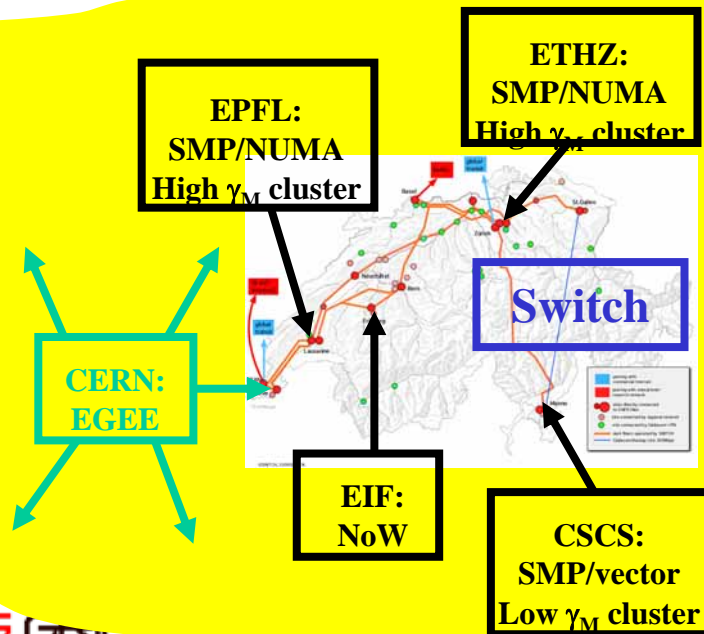
Intelligent Scheduling System ISS

- Partners: CSCS, EPFL, EIA-FR
- Provide a middleware service allowing optimal placement and scheduling of applications on the Grid – **submit to the most suited computer architecture based on resource and application monitoring**
- Research-oriented project, exploiting new ideas for a scheduling approach (2 PhDs)



ISS Details

- Cost function includes monitoring data on machine status and application behaviour. Usage of Γ model. See <http://pleiades1.epfl.ch/~rgruber/projects/iss.pdf>
- Monitoring Data on machines and applications delivered by application monitoring and the service itself
- Actual job submission through existing Grid middleware



First Testbed : EPFL
Mechanics departement
machines (clusters &
single CPU machines)

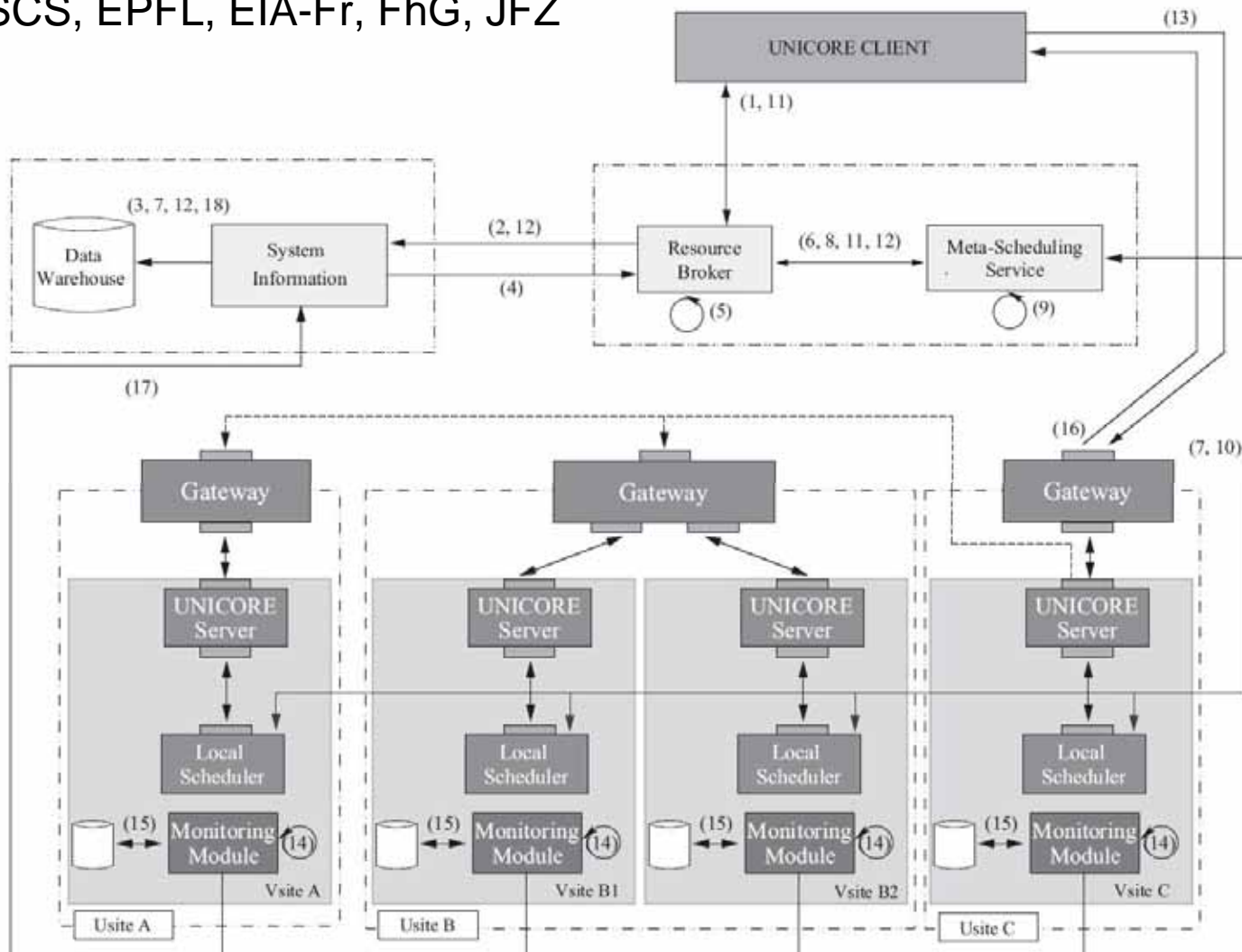
Second Testbed : Whole
EPFL

Third Testbed : EPFL +
CSCS + EIA-Fr + ETHZ
machines



Exemple : Integration of ISS into VIOLA/MSS/UniCORE Environment

Team : CSCS, EPFL, EIA-Fr, FhG, JFZ



And There Are More...

... Swiss Involvements in Grids

- CoreGrid: EPFL, CSCS, EIA-FR
- KnowARC project: University of Geneva
- DILIGENT: University of Basel
- EMBRACE: University of Lausanne, SIB
- Computational Chemistry Grid: University of Zurich
- ...



Middleware

Using existing stacks wherever possible

- Mixed approach – end to end application integration
- Development together with existing projects
 - EGEE: security development (SWITCH, JRA3)
 - Extension of existing middleware (CSCS, NorduGrid LSF integration; gLite metadata catalog)
- New development (research)
 - ISS
 - With GUP (Austrian Grid): TUBITS high-throughput submission

• ...



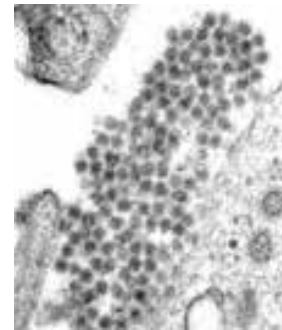
Swiss Bio Grid Applications

Usage Patterns of different Applications

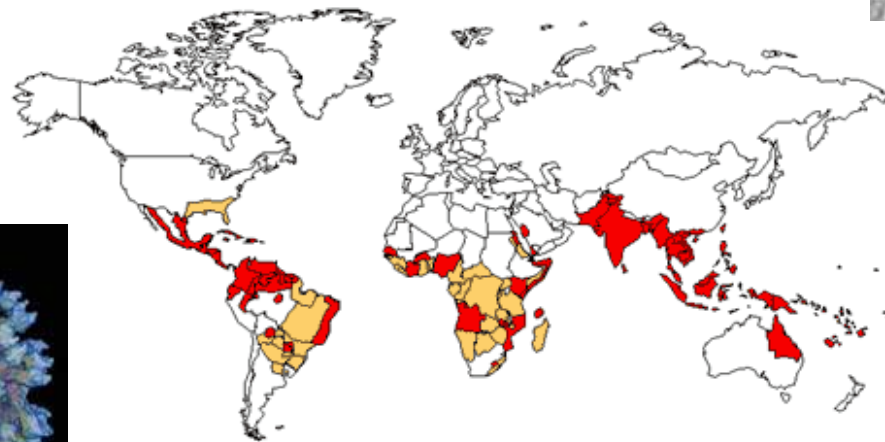
- Identified three classes of applications
 - Short CPU jobs (Docking)
 - Medium CPU + data exchange (Proteomics Pipelining)
 - Data intensive (Mass Spectrometry MS; Systems Biology)
- Strategy: Address them in sequence, find commonalities
 - Dengue docking project (see next slides) – concluded
 - swissPIT (Protein Identification Toolbox) Project – in progress
 - Data project – bio data distribution project – in development



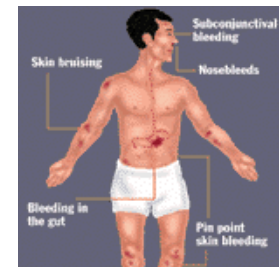
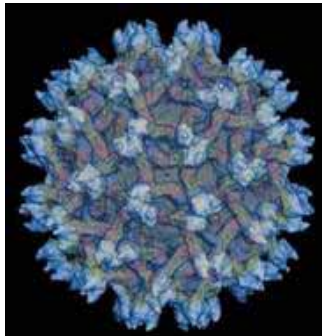
Orphan Diseases: Dengue



World Distribution of Dengue



- Areas infested with *Aedes aegypti*
- Areas with *Aedes aegypti* and dengue epidemic activity



CDC



What does it take to make a drug?

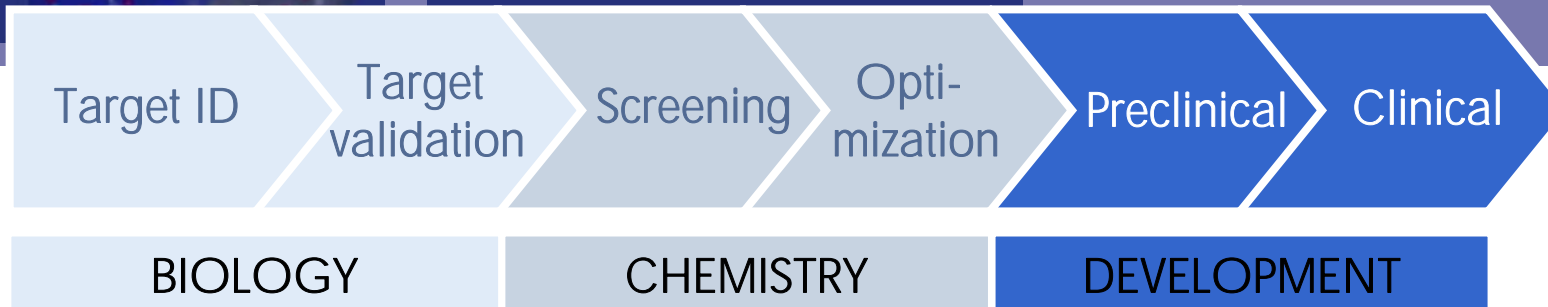
12 years of development, 802 mio US\$

(DiMasi, J.A. et al. (2003) *J Health Econ*, 22, 151-185).

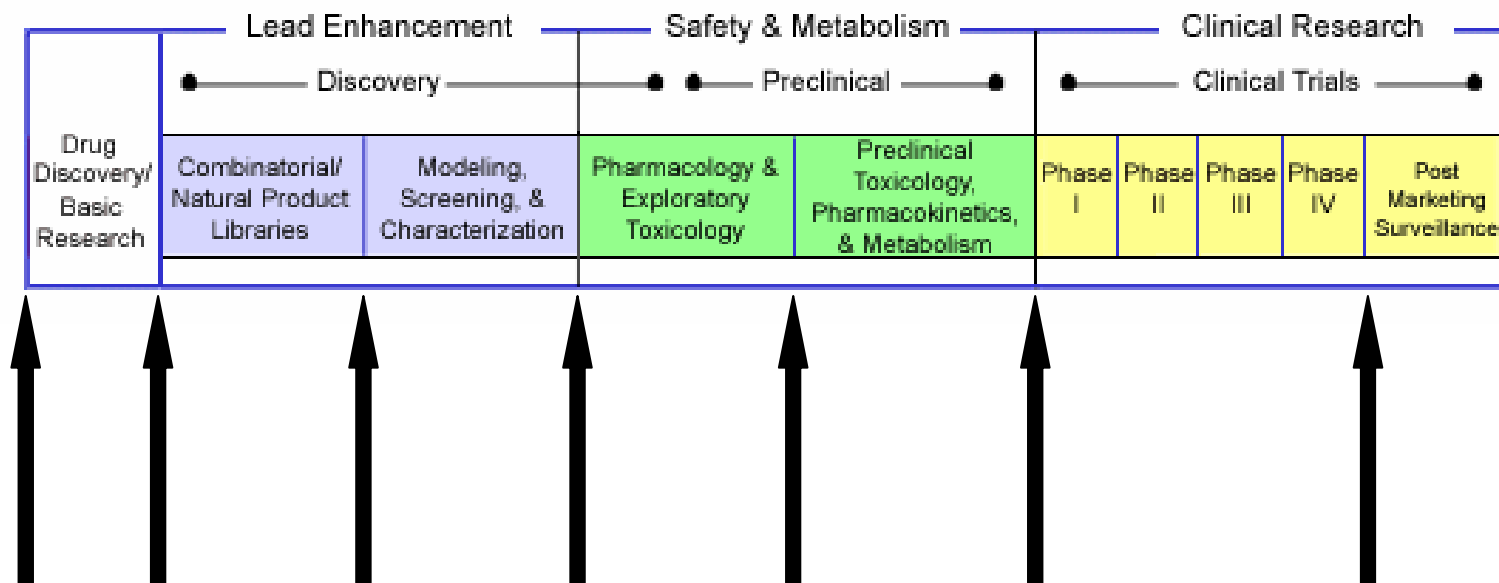
1 in 10'000 NCE becomes a product

(Heilman, R.D. (1995) *Qual Assur* 4(1) 75-9.)

,Only' 20 years of Patent – 8 years to make money



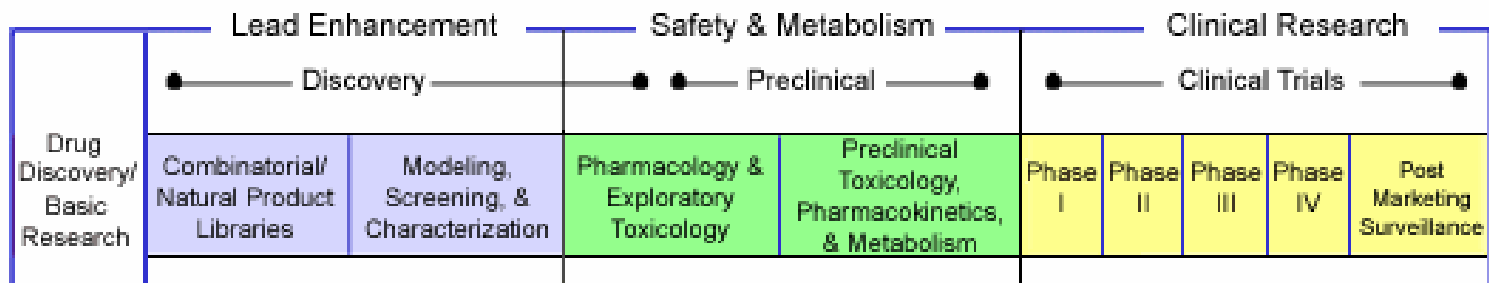
“In Silico” Drug Development



Bioinformatics, data mining, visualization, simulations, modeling, and many algorithms, databases



Screening of compounds



Computational screening of small compounds to identify early drug candidates



Dengue Docking project

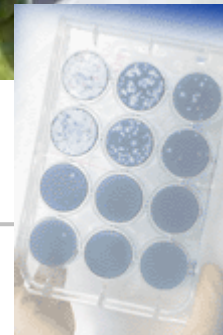
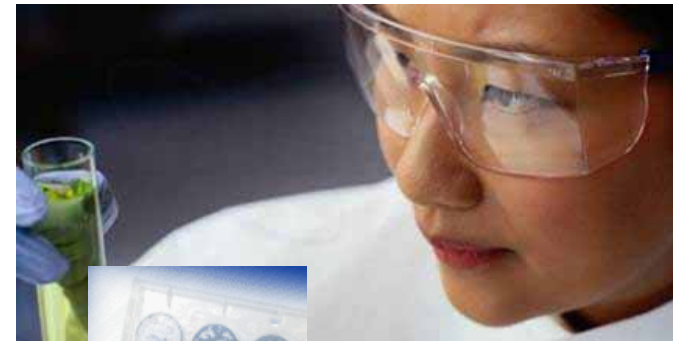
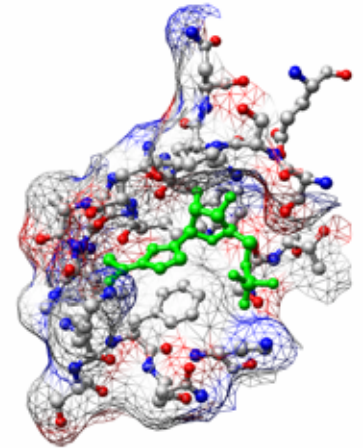
BIOZENTRUM
Universität Basel

**Proof of concept for
successful private-public
partnership**

**Biozentrum:
in silico docking**

**Novartis Institute for Tropical
Diseases:
In vitro/in vivo follow-up**

**Novartis:
drug development at cost**



NOVARTIS
INSTITUTE FOR
TROPICAL DISEASES

Dengue Docking project

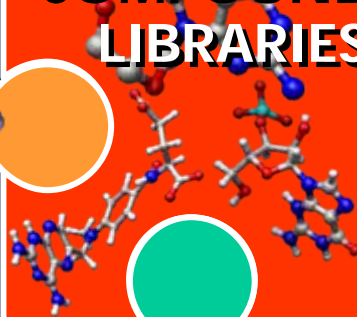
3D structure of targets

- NS5 Methyltransferase
- NS3 Protease
- GPE Envelope Glycoprotein
- NS3 Helicase

**TARGET
PROTEINS**



**COMPOUND
LIBRARIES**



**ALGORITHMS IT
INFRASTRUCTURE**

```
char* filename = argv[1];
int iterations = 100000;
/* fill the array of random numbers */
double numbers[ITERATIONS];
//double foo = 0;
for (int i = 0; i < ITERATIONS; i++)
{
    numbers[i] = (double)rand() / (double)RAND_MAX;
    //numbers[i] = foo++;
}
// write numbers to file
/* write the array to the file */
FILE* file = fopen ( filename, "w" );
if (myFile == 0)
{
    fprintf (err << "could not open file\n");
}
for (int i = 0; i < ITERATIONS; i++)
{
    fprintf (file, "%f\n", numbers[i]);
}
fclose (myFile);
```

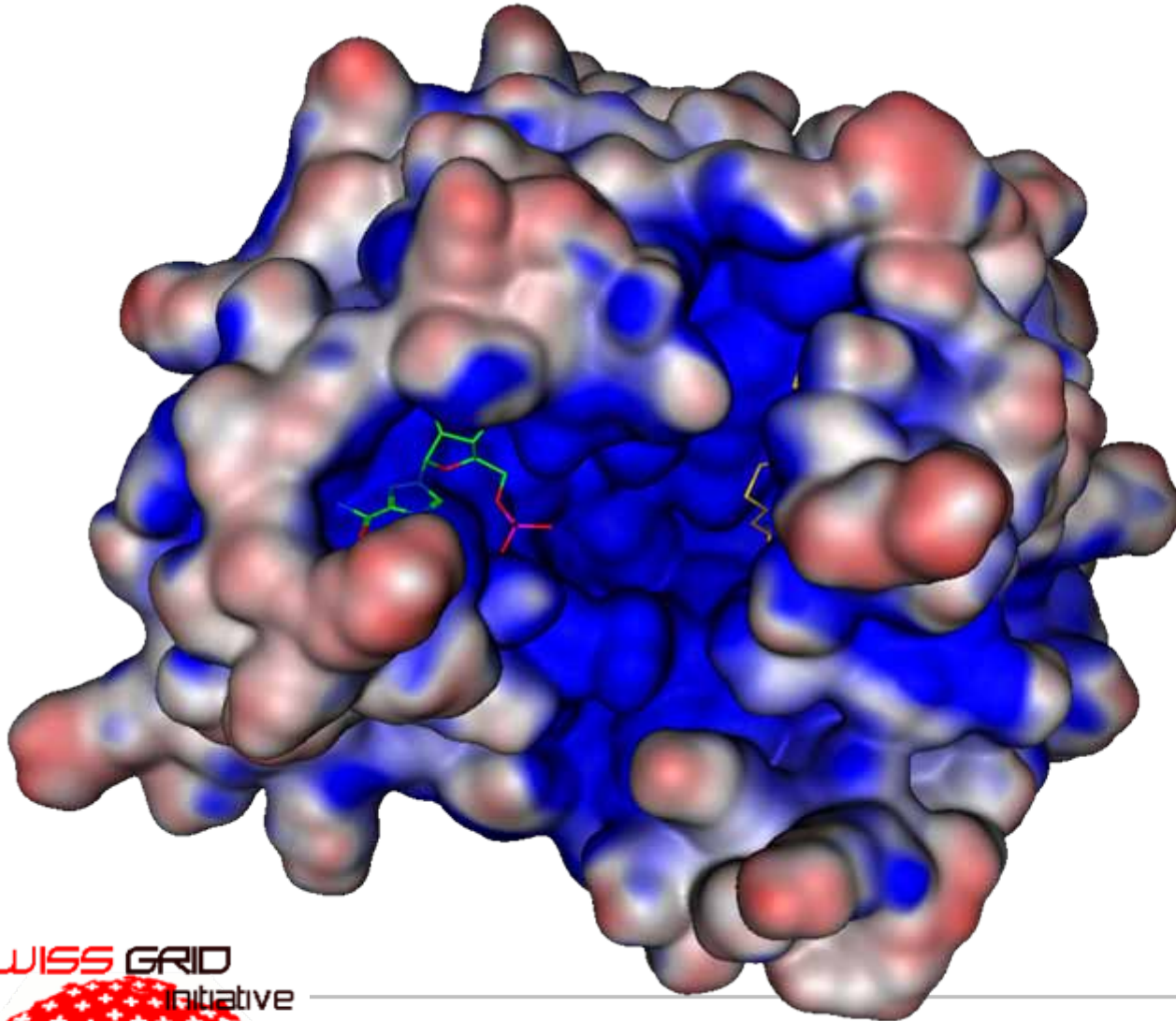


- NCI Diversity (2k)
- NCI DTP (200k)
- ZINC (2700k)

- DOCK 5.1
- Autodock 3.05
- FlexX (SCAI/BioSolvIT)
- GLIDE(Schrödinger)



Dengue NS5 Methyltransferase



Current Achievements of GRID-enabled Dengue Docking

Completed Phase I SwissBioGrid

**Completed large-scale parameterization
test using**

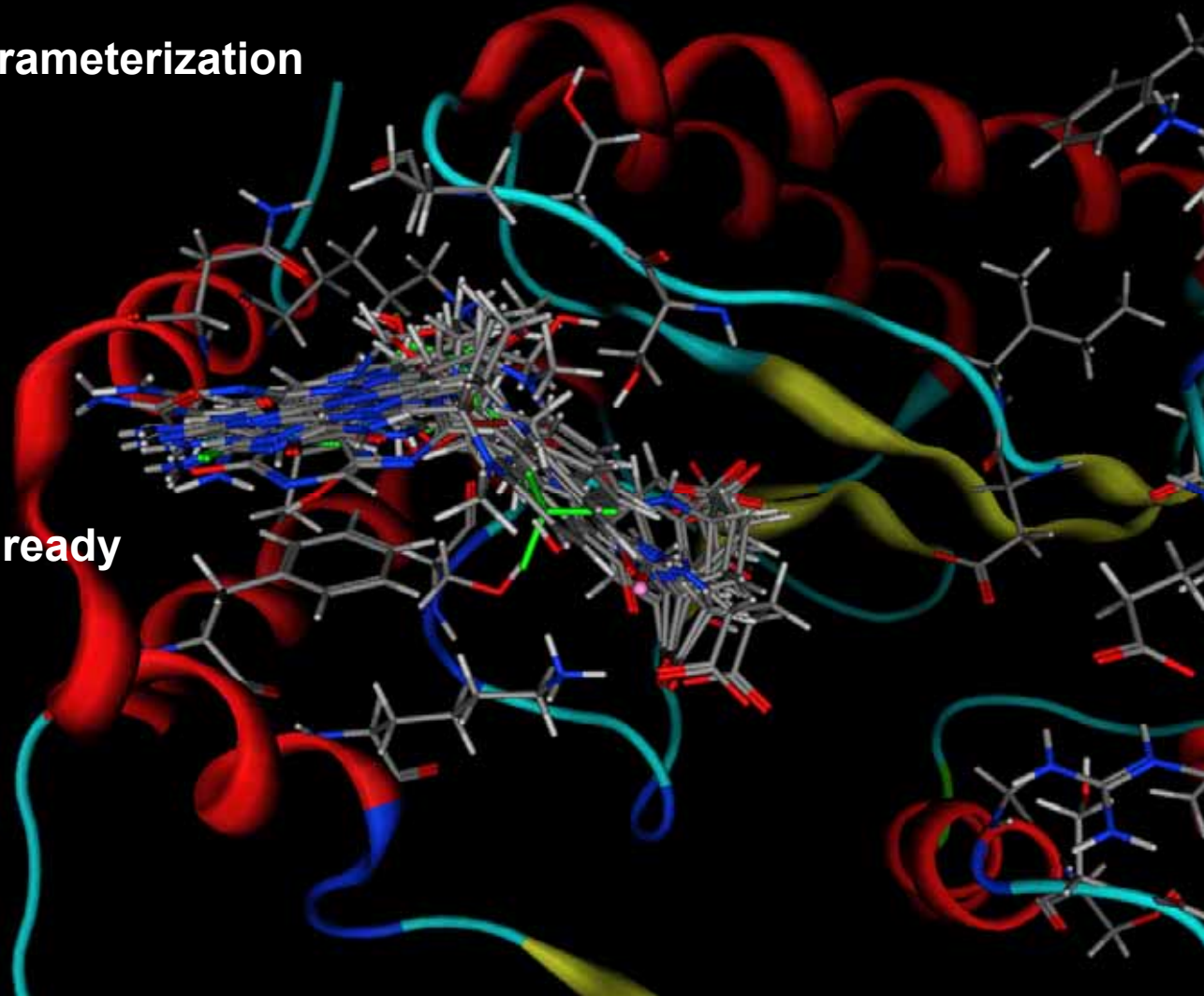
Autodock 3.0.5:

>500'000 docking runs,

>38'000h CPU time

***In vitro* testing of
predicted binders
is underway at NITD**

**Some initial candidates already
in next phase**



Some challenges in grid adoption

Compute resources are busy already

- Agree on dedicated compute time for grid projects
- PC Desktop grids: untapped resource
- Buy new clusters for your grid (not the idea)

Non-intrusiveness

- Firewall exceptions
- Non-intrusiveness on PC Desktop grids: application level

Application clearing:

- Security issues
- Numerical stability in heterogeneous environments

Data model in bioinformatics different from HEP

- Applications need access to large databases or data sets



Challenge: Heterogeneity

Very different resources at participating institutes

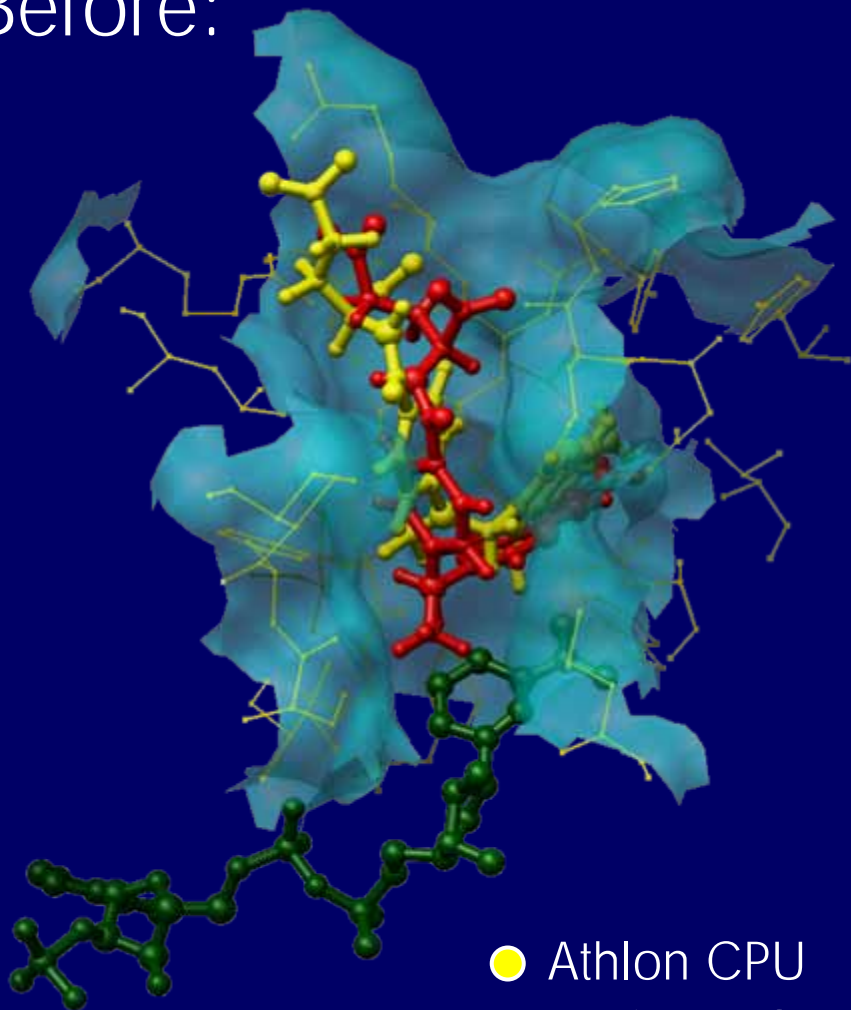
- Use 'standard' schedulers for clusters (Sun Grid Engine, LSF, PBS)
- Agree on a higher-level Grid scheduler
- Provide good documentation and bindings of the Grid scheduler to the predominant cluster schedulers
- Work on new bindings

Here we are already quite advanced, can make good use of results of other projects – but still a long way to go!

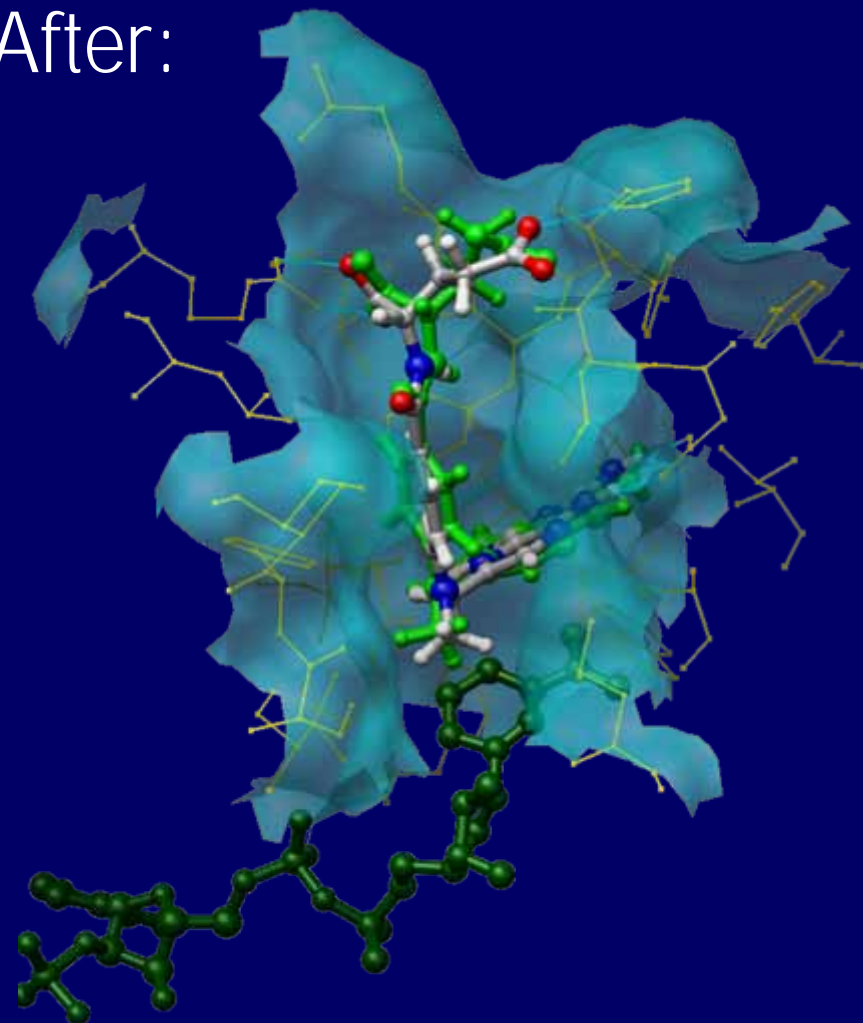


Challenge: Numerical stability

Before:



After:



- Athlon CPU
- Itanium2 CPU
- Bug fixed

Challenge: Security

Sensitive data, data safety

- Rely on standards for Authentication and Authorization
- Network data channel encryption
- Encryption of distributed data on storage
- Distributed keys and algorithms for retrieval (n of m schemes)

Not at all addressed yet; a lot of room for improvement



Challenge: Legacy

Licensed, proprietary, legacy code

- Solve the problem together with the software provider
- New licensing models for distributed computing (e.g. license servers don't scale)
- Legacy support
 - Recompilation if possible
 - Emulators
 - Virtual machines

Virtual Machines may be the way forward for many of these applications – but not production quality yet, lot of research to be done; also a lot of room for improvement



Challenge: User Interface

Users don't want to deal with Grid specifics

- Set up a Grid Portal
- Many portals exist, however application-specific interface for the users always needs to be built
- Proteomics Project addresses this: dedicated proteomics pipelining portal based on existing Grid portal technologies



Data model for bioinformatics

Bioinformatics often requires operations on large amounts of data (100s of GB)

Transparent versioning and provisioning of the correct data for the computation necessary

Local caching of large datasets reduces network traffic

DataProxy (ProtoGRID)

- Each job description requests data files.
 - Identified by message digest and size
 - Associated with a data-providing resource
- Data proxy transparently handles data requirements
 - Prior to execution, request or re-use from cache
 - unique interface to data for application.
- Cached data is purged when necessary



DataGrid implementation

Making use of Sybase AVAKI DataGrid

- Mature industrial strength system
- Not easy to set up
- Tests are in progress



Summary

Active and evolving Grid landscape in Switzerland

The Swiss Bio Grid demonstrated successfully the usefulness of Grids for the bioinformatics community in Switzerland



Links

SwissGrid Initiative:

<http://www.swiss-grid.org/> or

<http://www.gridinitiative.ch/>

Swiss Bio Grid

<http://www.swissbiogrid.ch/>

CSCS:

<http://cscs.ch/>

