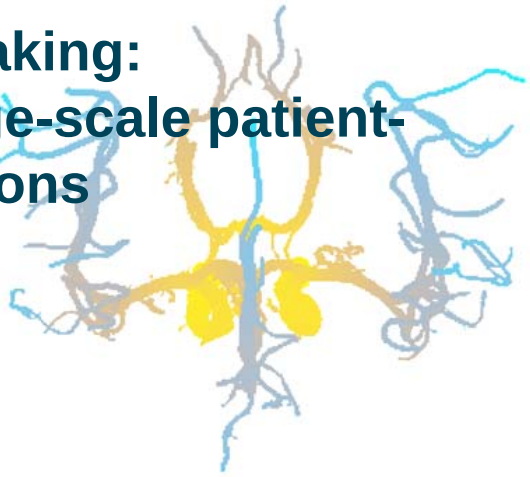
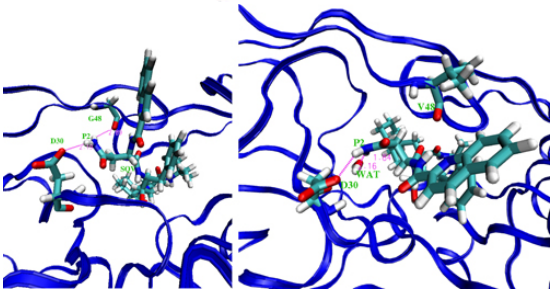
# Life or death decision-making: The medical case for large-scale patient-specific medical simulations



**Steven Manos**
Centre for Computational Science
University College London
London, U.K.

s.manos@ucl.ac.uk





**RIKEN Symposium, Tokyo, Japan, March 13th - 14th, 2008.**

## Overview

- What is patient-specific medical simulation?
- Clinical computing
- Computational infrastructure requirements
  - Grid middleware, the Application Hosting Environment
  - Cross-site runs and distributed computing
  - Advance reservation
  - Urgent computing
- **Case study I**: HIV/AIDS drug design
- **Case study II**: Treating neuro-vascular pathologies
- What's needed to make patient-specific medical computing a reality

## Patient-specific medicine

- 'Personalised medicine' - use the patient's genetic profile to better manage disease or a predisposition towards a disease
- Tailoring of medical treatments based on the characteristics of an individual patient

## Why use patient-specific approaches?

- Treatments can be assessed for their effectiveness with respect to the patient before being administered, saving the potential expense of ineffective treatments

## Patient-specific medical-simulation

- Use of genotypic and or phenotypic simulation to customise treatments for each particular patient, where computational simulation can be used to predict the outcome of courses of treatment and/or surgery



## What is grid computing?

*"Distributed computing performed transparently across multiple administrative domains"*

Any production grid should be:

- **Stable**
- **Persistent**
- **Usable**

It must provide easy access to many different types of resources from which to pick and choose those required.

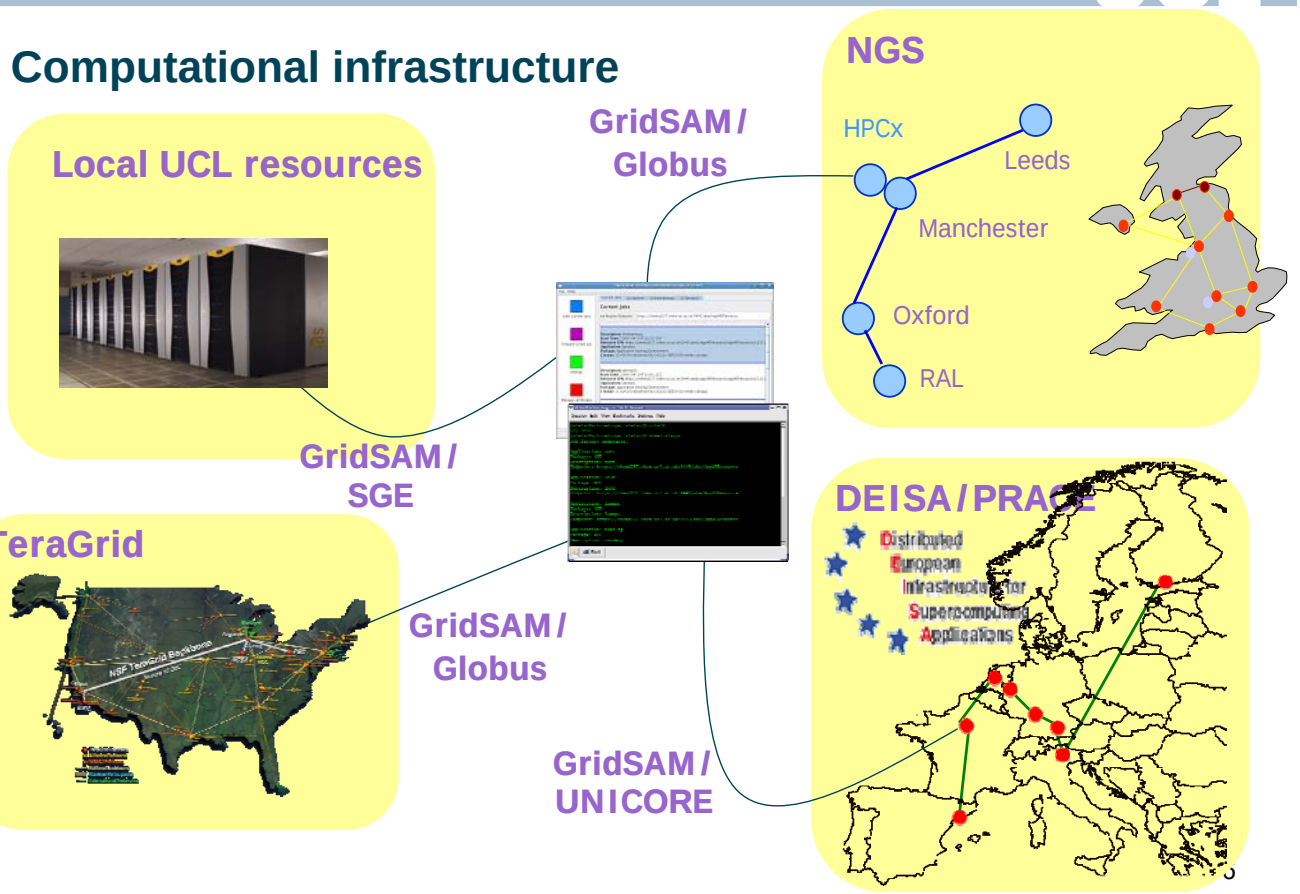
It is debatable whether many grids in operation today fit this definition

## What is clinical (grid) computing?

- Computational experiments integrated seamlessly into current clinical practice
- Clinical decisions influenced by patient specific computations: turnaround time for data acquisition, simulation, post-processing, visualisation, final results and reporting.
- Fitting the computational time scale to the clinical time scale:
  - Capture the clinical workflow
  - Get results which will influence clinical decisions: 1 day? 1 week?
- Development of procedures and software in consultation with clinicians
- On-demand availability of storage, networking and computational resources



## Computational infrastructure



## Computational infrastructure: Application Hosting Environment

- Making computing power available to non-technical people
- Need to utilize resources from globally distributed grids
  - Administratively distinct
  - Running different middleware stacks
- Wrestling with middleware can't be a limiting step for scientists
- Need tools to hide complexity of underlying grids



## Computational infrastructure: Application Hosting Environment

- Applications are stateful WSRF services
- Lightweight hosting environment for running applications on grid resources and on local resources
- Community model: expert user installs AHE, shares applications with others
- Simple clients with very limited dependencies

## Computational infrastructure: Application Hosting Environment

- Applications not jobs
  - Application could consist of a coupled model, parameter sweep, steerable application, or a single executable
- AHE supports single site jobs, multisite MPIg jobs, and single and multisite steerable jobs
- We use “application” to denote a higher level concept than a job
  - In AHE terminology, an application may require running multiple jobs
- Architecturally, the AHE is a portal, where the interface is a rich client, not a web browser
  - Of course, AHE services can be used behind a Web portal, if you like



## Computational infrastructure: Cross-site runs

MPIg is the next version of MPICH-G2

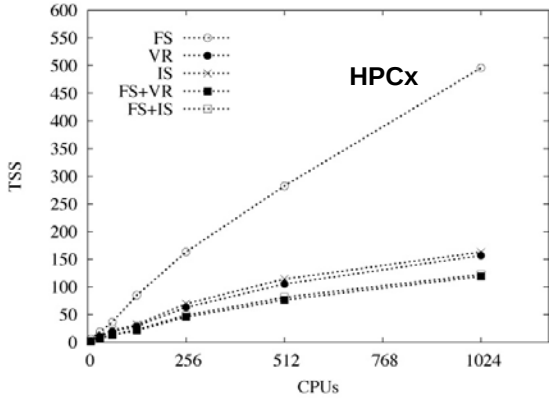
•Some problems won’t fit on a single machine, and require the RAM/processors of multiple machines on the grid.

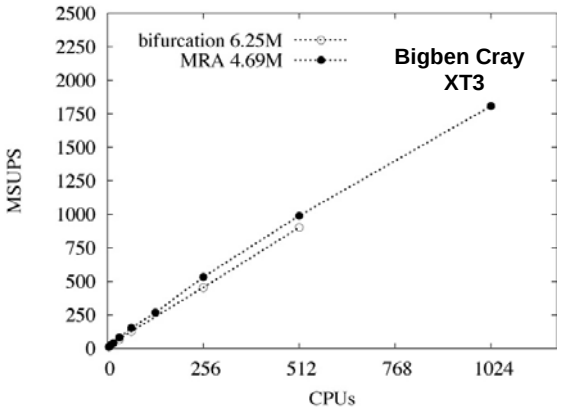
• MPIg allows for jobs to be turned around faster by using small numbers of processors on several machines - essential for clinician

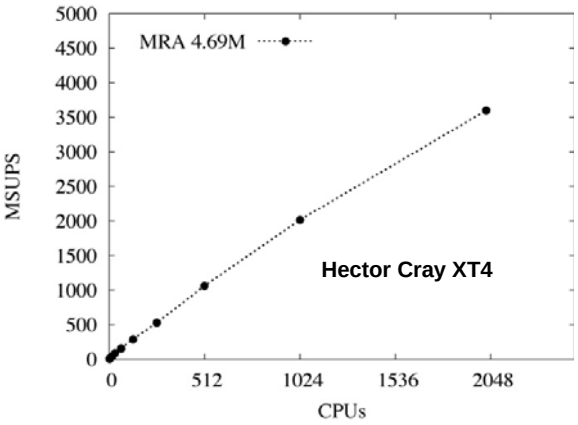
• MPIg uses a true threaded model for overlapping communication and computation, so with appropriate programming, latencies between sites can be effectively hidden.

| Site | Intra-machine (ms) | | Inter-machine (ms) | |
|---|---|---|---|---|
| | Median | $\sigma$ | Median | $\sigma$ |
| TeraGrid | 0.025 | 0.16 | 60 | 40 |
| LONI | 0.083 | 0.38 | 9.7 | 7.5 |

# Computational infrastructure: Cross-site runs II

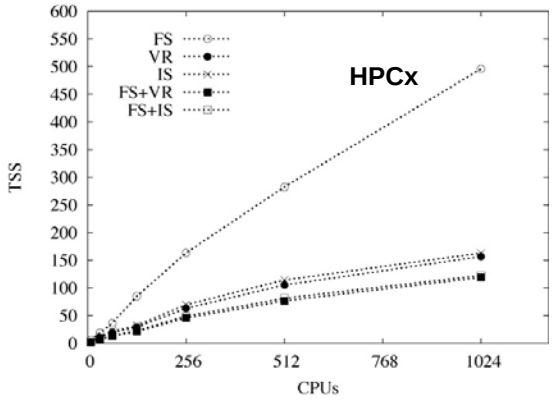


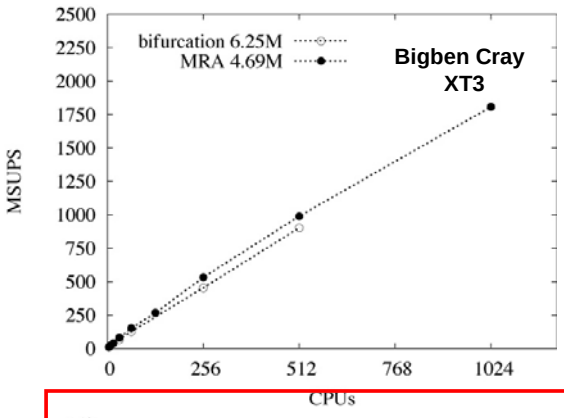



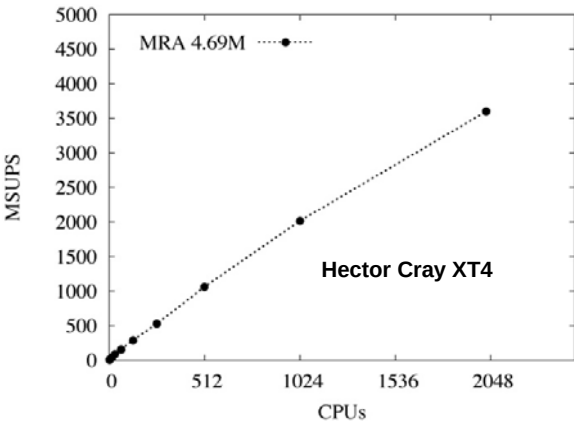



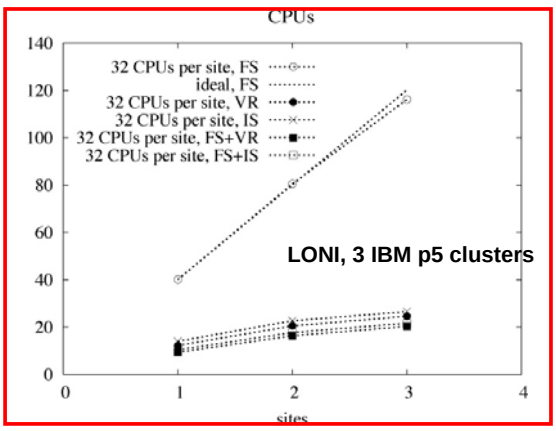



# Computational infrastructure: Cross-site runs II



## Computational infrastructure: Advanced reservations I

- HARC - Highly Available Resource Co-Allocator

- What is ***Co-allocation***?
- Process of reserving multiple resources for use by a single application or “thing” – but ***in a single step...***
- (Synonym for ***Co-scheduling***)

- Can reserve the resources:
  - For the same time:
    - Meta-computing, large MPIg/MPICH-G2 jobs
    - Distributed visualization
    - Booking equipment
  - Or some coordinated set of times
    - Computational workflows



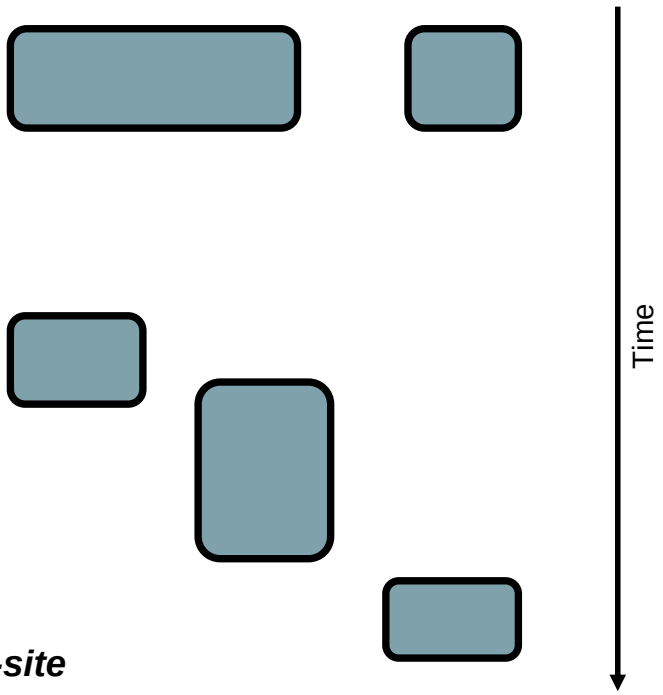
## Computational infrastructure: Advanced reservations II

### Meta-computing job

- 32 procs. on santaka
- 8 procs. on kite1
- both from 1-2pm

### Workflow job

- 1024 procs. on HPCx from 2-3pm
- 8 procs on vizws00 from 2.45-5pm
- 128 procs on ducky from 5.15-6pm




***Demonstrated reserved cross-site***

***Runs at SC07***

# Computational infrastructure: Advanced reservations III

```
$ harc-reserve \
       -c tg-login2.sdsc.edu/32 \
       -c grid-hg.ncsa.teragrid.org/32 \
       -s 2008-03-15T14:00GMT -d 4:00
tg-login2.sdsc.edu/1200268572
grid-hg.ncsa.teragrid.org/smanos.1870
```

```
$ harc-reserve \
       -c zeke.loni.org/32 \
       -c bluedawg.loni.org/32 \
       -c ducky.loni.org/32 \
       -s 2008-03-15T14:00GMT -d 4:00
zeke.loni.org/l3f1n03.191.r
bluedawg.loni.org/l1f1n01.493.r
ducky.loni.org/l2f1n01.425.r
```

Also available via the HARC API - can be easily built into Java applications.

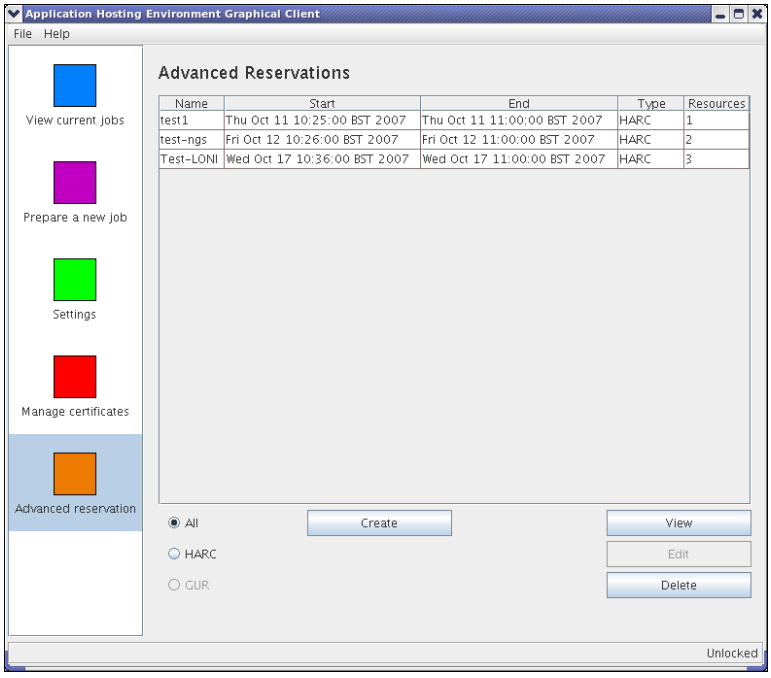
Deployed on a number of systems

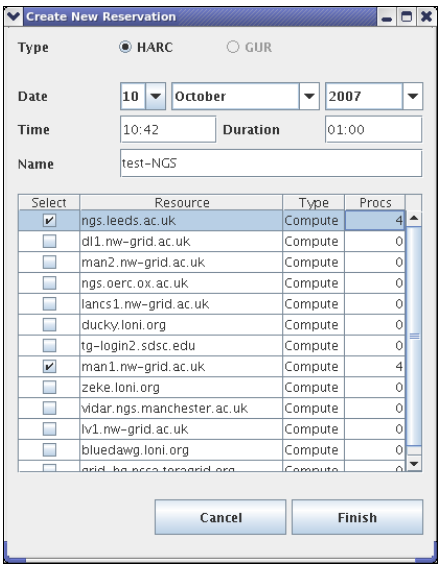
- LONI
- TeraGrid
- HPCx
- North West Grid (UK)
- National Grid Service - NGS (UK)



# Computational infrastructure: Advanced reservations IV

Creating HARC reservations in the AHE

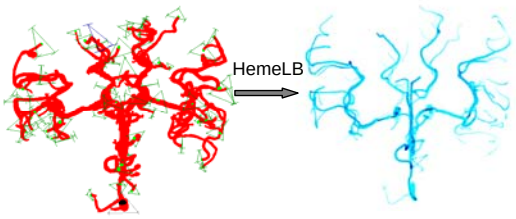


## Computational infrastructure: Urgent computing I

- Applications with dynamic data and ***result deadlines*** are being deployed

- Late results are useless
  - Wildfire path prediction
  - Storm/Flood prediction
  - Influenza modeling




- Some jobs need priority access "**Right-of-Way Token**"



## Computational infrastructure: Urgent computing II

- "Next-to-run" status for priority queue
  - wait for running jobs to complete
- Force checkpoint of existing jobs; run urgent job
- Suspend current job in memory (kill -STOP); run urgent job
- Kill all jobs immediately; run urgent job


**SPRUCE** Special PRiority and Urgent Computing Environment

Not only reserving or gaining access to computational resources, but can also be emergency access to bandwidth, for example.

# Computational infrastructure: Urgent computing III

- Deployed and Available on TeraGrid -
  - UC/ANL
  - NCSA
  - SDSC
  - NCAR
  - Purdue
  - TACC
- Other sites
  - LSU
  - Virginia Tech
  - LONI

TeraGrid Deployment Status

| SITE | CONFIGURATION | POLICY | SCHEDULER | PROCESSORS | CONTACT |
|---|---|---|---|---|---|
| **UC/ANL IA64** usage guide | Intel IA-64, 62 nodes | elevated-priority next-to-run pre-emption | Torque/Moab | 124/124 (Production) | Ti Leggett |
| **UC/ANL IA32** usage guide | Intel IA-32, 96 nodes | elevated-priority next-to-run | Torque/Moab | 192/192 (Production) | Joe Insley |
| **NCSA Mercury** usage guide | Intel IA-64, 631 nodes | next-to-run | Torque/Moab | TBD/1262 (Production) | Peter Enstrom |
| **Purdue Lear** usage guide | Dell EM64T Cluster, 512 nodes | next-to-run | PBS Pro | 1024/1024 (Upgrading) | Preston Smith |
| **SDSC Datastar** usage guide | IBM P series, 272 (8-way) P655+ and 6 (32-way) P690 | next-to-run | LoadLeveler/Catalina | 2368/2368 (Production) | Tony Vu |
| **SDSC OnDemand** usage guide | Rocks cluster, 64 nodes | elevated-priority next-to-run | SGE | 256/256 (Pre-production) | DJ Choi |
| **TACC Lonestar** usage guide | Dell PowerEdge 1955, 1460 nodes | next-to-run | LSF | 16/5840 (Production) | Bill Barth |
| **NCAR Frost** usage guide | Single-rack BG/L, 1024 nodes | next-to-run | Cobalt | 2048/2048 (Production) | Jason Cope |

***Demo at SC07 using TACC Lonestar***



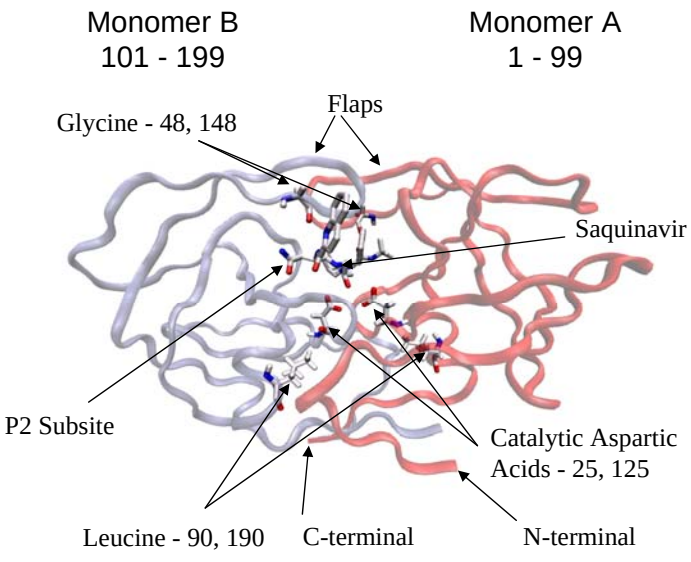
# Case study I : Patient-specific HIV drug therapy

## HIV-1 Protease is a common target for HIV drug therapy

- Enzyme of HIV responsible for protein maturation
- Target for Anti-retroviral Inhibitors
- Example of Structure Assisted Drug Design
- 9 FDA inhibitors of HIV-1 protease

**So what's the problem?**

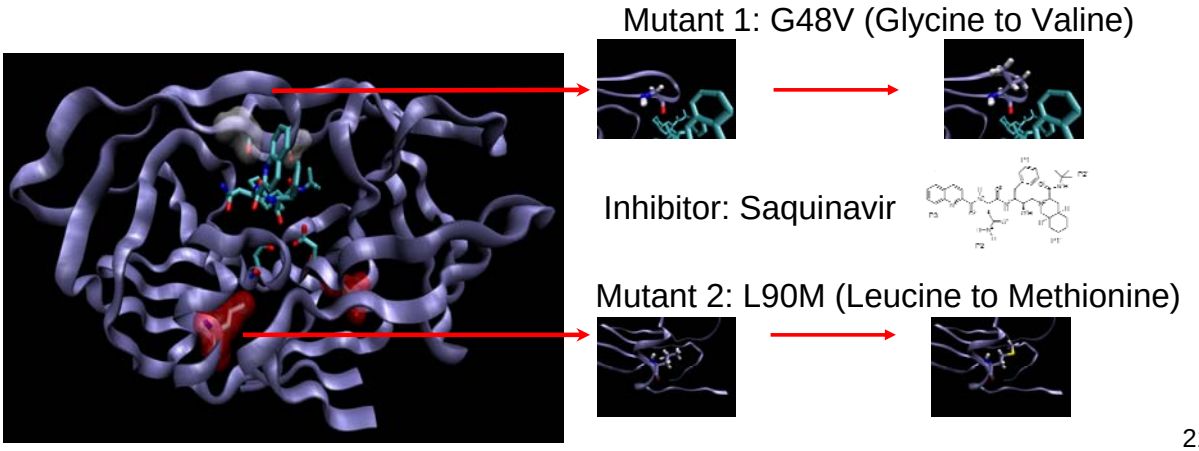
- Emergence of drug resistant mutations in protease
- Render drug ineffective
- Drug resistant mutants have emerged for all FDA inhibitors



# HIV-1 Protease

**AIMS:**

- Study the differential interactions between wild-type and mutant proteases with an inhibitor
- Gain insight at molecular level into dynamical cause of drug resistance
- Determine conformational differences of the drug in the active site
- Calculate drug binding affinities

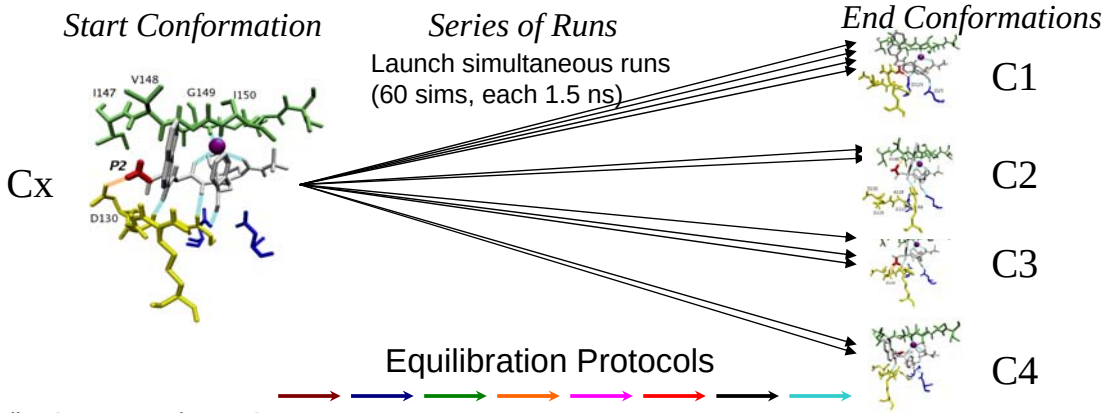





# HIV-1 Protease

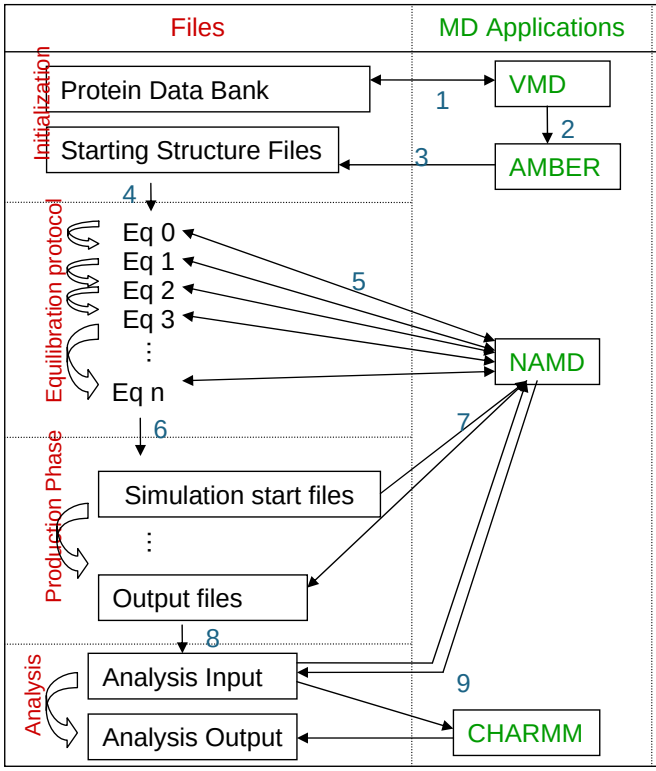
Compute intensive MD is well suited for an supercomputing grid

- Uses the NAMD MD code
- Simulate each system many times from same starting position
- Each run has randomized atomic energies fitting a certain temperature
- Allows conformational sampling




*S.K. Sadiq, S. Wan and P.V. Coveney, Biochemistry, **46**, 14865-14877 (2007)*

# Simulation Workflow




Processes

1. Strip out relevant pdb information
2. Incorporate mutations
3. Ionize and solvate to build system
4. Static Equilibration files are built according to variable protocol; output feeds into input of next equilibration
5. Each step of the chained equilibration protocol runs sequentially
6. End equilibration output serves as input of the production run
7. Production run
8. Output files of simulation are used as input for analysis
9. Analysis returns files containing required data for end user



# HIV-1 Protease

**Constructing Workflows with the AHE**

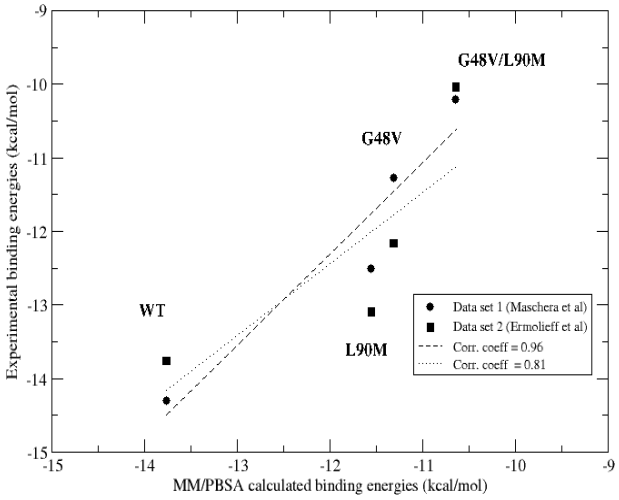
- AHE developed as part of OMII/EPSRC funded projects
- AHE used as middleware to automate the large number of MD simulations required for HIV-1 protease study
- Simulations launched across internationally distributed supercomputers
- By calling command line clients from Perl script complex workflows can be achieved
- Easily create chained or ensemble simulations
- e.g. MD equilibration protocol implemented by:
  - ahe-prepare → prepare a new simulation for the first step
  - ahe-start → start the step
  - ahe-monitor → poll until step complete
  - ahe-getoutput → download output files
  - repeat for next step

# HIV-1 Protease

## Binding of saquinavir to wildtype and resistant HIV-1 proteases L90M and G48V/L90M

**Thermodynamic decomposition**

- explains the distortions in enthalpy/entropy balance caused by the L90M and G48V mutations
- absolute drug binding energies are in **excellent agreement** (1 – 1.5kcal/mol) with experimental values

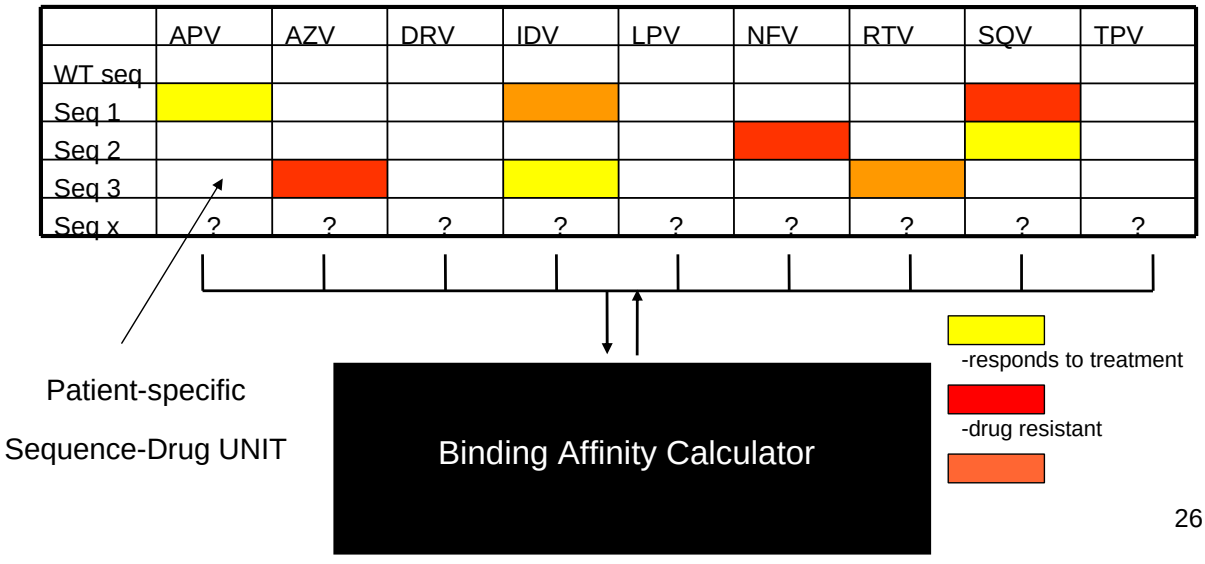



*I. Stoica, S. K. Sadiq, P. V. Coveney, "Rapid and Accurate Prediction of Binding Free Energies for Saquinavir-Bound HIV-1 Proteases", Journal of the American Chemical Society. (doi=10.1021/ja0779250,url: http://pubs.acs.org/cgi-bin/abstract.cgi/jacsat/2008/130/i08/abs/ja0779250.html)*

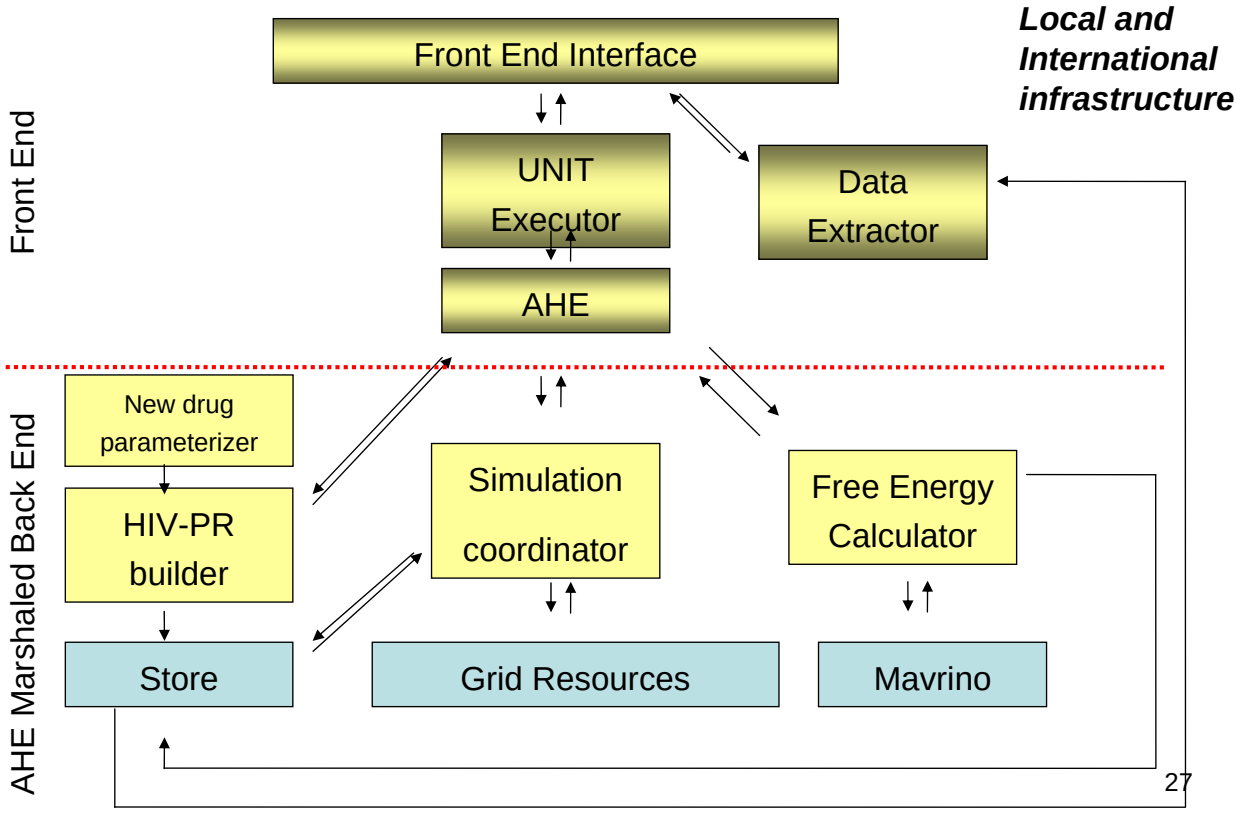


# High-throughput Patient-Specific Binding Affinity Calculations (BAC)

- **Input:** patient genotype (MRC Clinical Trials Unit's HIV/AIDS database)
- **Output:** resistance profile for all FDA-approved inhibitors



# Automating Binding Affinity Calculations



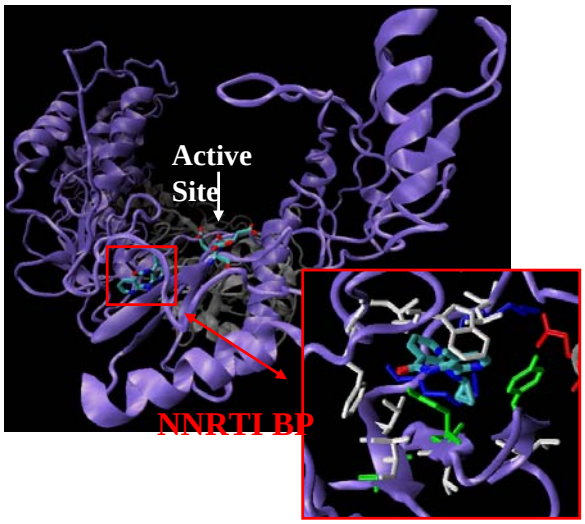



# HIV-1 Reverse Transcriptase

**Extending the BAC**

Aim to incorporate another critical HIV enzyme – Reverse Transcriptase

- 5 times bigger than Protease
- Target for two types of drugs: NRTIs and NNRTIs
- Initially concentrating on the allosteric NNRTI class
- Three FDA approved NNRTIs: Nevirapine, Efavirenz & Delavirdine




NNRTIs create a binding pocket which doesn't exist in the apo structure (seen in the picture to the right).

We use the same techniques applied to HIV-protease to measure the drugs' binding affinity.

# Constructing workflows with GSEngine and AHE

- ViroLab - a virtual laboratory for decision support in viral diseases treatment.
- GSEngine (previously named VLEngine), a Ruby based run time environment which can be used to script workflows and experiments
- Data acquisition, data pre-processing, simulation, post-processing, visualisation, can be generically scripted.
- Object-oriented, so parts of it can be reused
  - Expert users can develop own modules using the Eclipse development environment
  - Basic users can use and recombine pre-written modules
- This recently combined with AHE, meaning that large scale grid computing tasks can be seamlessly integrated into the workflow.



# Case study II : Grid enabled neurosurgical imaging using simulation

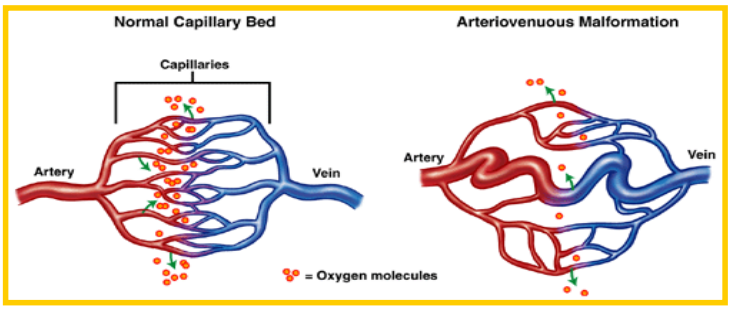
**The GENIUS project aims** to model large scale patient specific cerebral blood flow in clinically relevant time frames

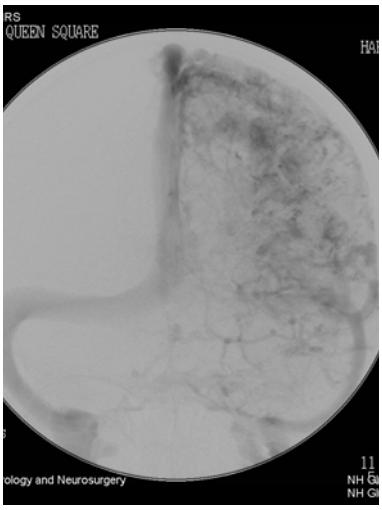
**Objectives**:

- To study cerebral blood flow using patient-specific image-based models.
- To provide insights into the cerebral blood flow & anomalies.
- To develop tools and policies by means of which users can better exploit the ability to reserve and co-reserve HPC resources.
- To develop interfaces which permit users to easily deploy and monitor simulations across multiple computational resources.
- To visualize and steer the results of distributed simulations in real time

*Yield patient-specific information which helps plan embolisation of arterio-venous malformations, aneurysms, etc.*

## Arterio-venous malformations (AVM)



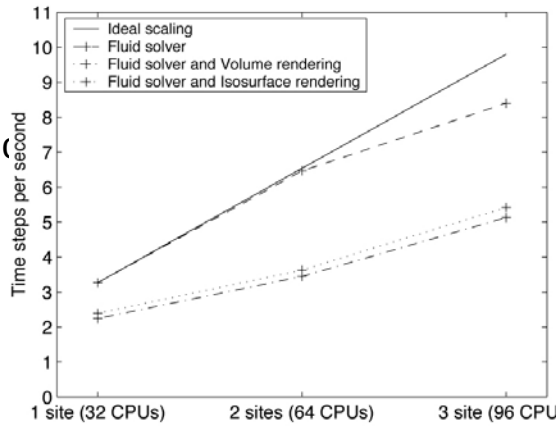






## Modeling vascular blood flow - HemeLB

Efficient fluid solver for modelling brain bloodflow called HemeLB:

- Uses the lattice-Boltzmann method
- Efficient algorithms for sparse geometries
- Machine-topology aware graph growing partitioning technique, to help minimise the issue of cross-site latencies
- Optimized inter- and intra-machine c
- Full checkpoint capabilities.

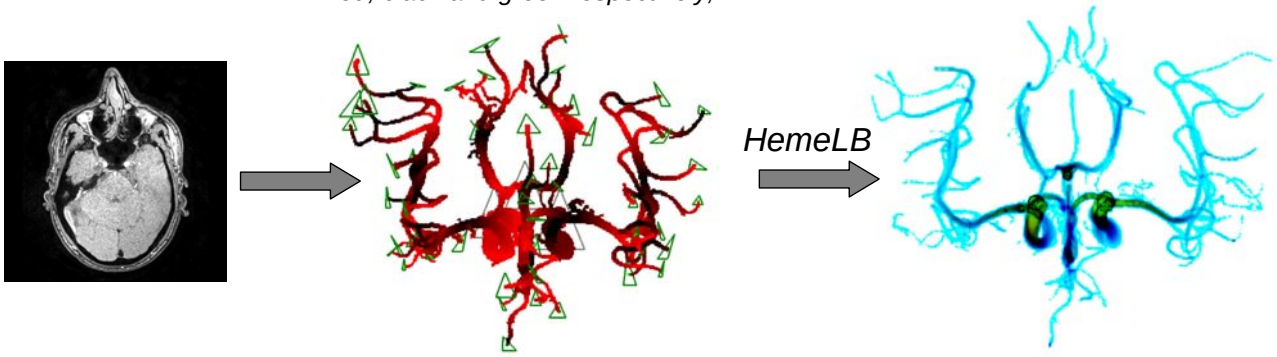


## Modelling and visualisation

- Convert DICOM slice data to 3D model, MRI or CT scan where the vasculature is of high contrast, 200 - 200 $\mu m$ resolution, $1000^3$ voxels
- Each voxel is a solid (vascular wall), fluild, fluid next to a wall, a fluid inlet or a fluid outlet
- Our current simulation has 3 inlets and ~50 outlets
- We apply an oscillating pressure at the inlet and an oscillating or constant one at the outlets
- Real-time *in-situ* visualisation of the data using streamlines, iso-surfacing or volume rendering

*Reconstruction and boundary condition set-up; fluid sites, inlet and outlet sites in red, black and green respectively;*

*Stationary von Mises stress flow field obtained with our ray tracer*

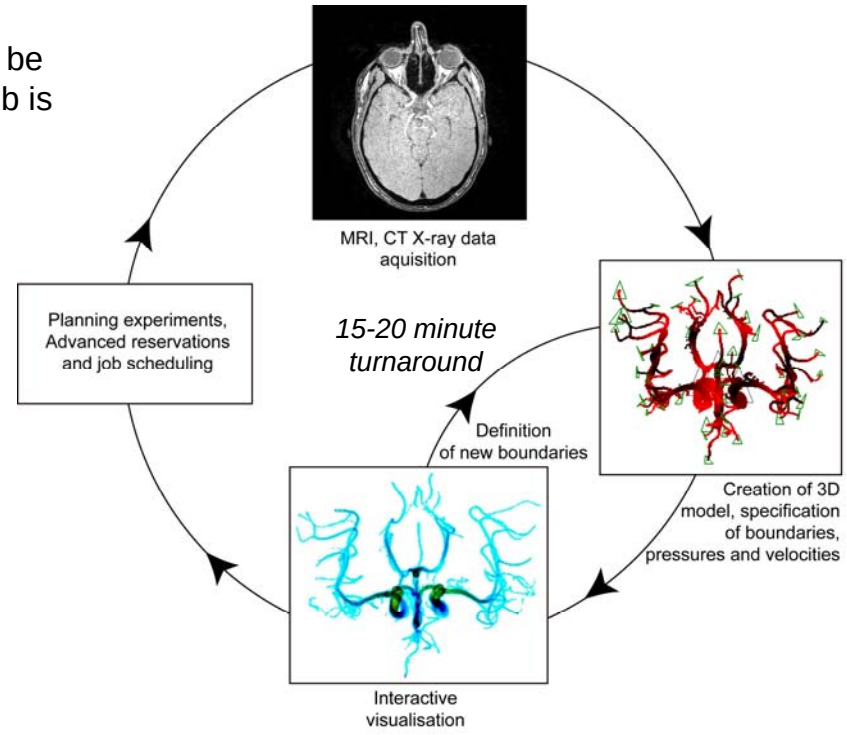



## Clinical work flow

Clinician's few of how things should work in the software environment

Clinician's shouldn't have to be concerned with *where* the job is running.. or *how*.

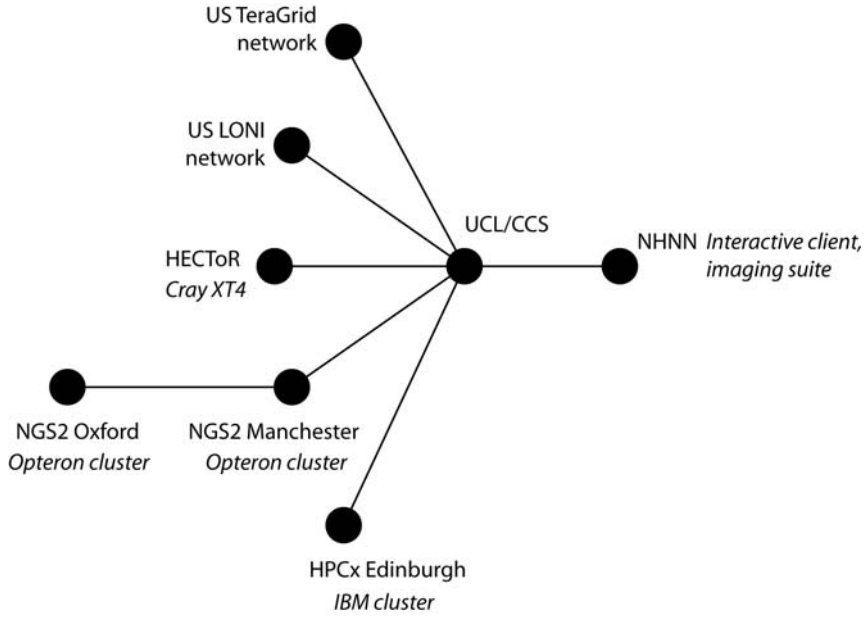
All the 'grid details' such as advance reservations, job launching, machine availability, etc. are hidden.



# Lightpath network

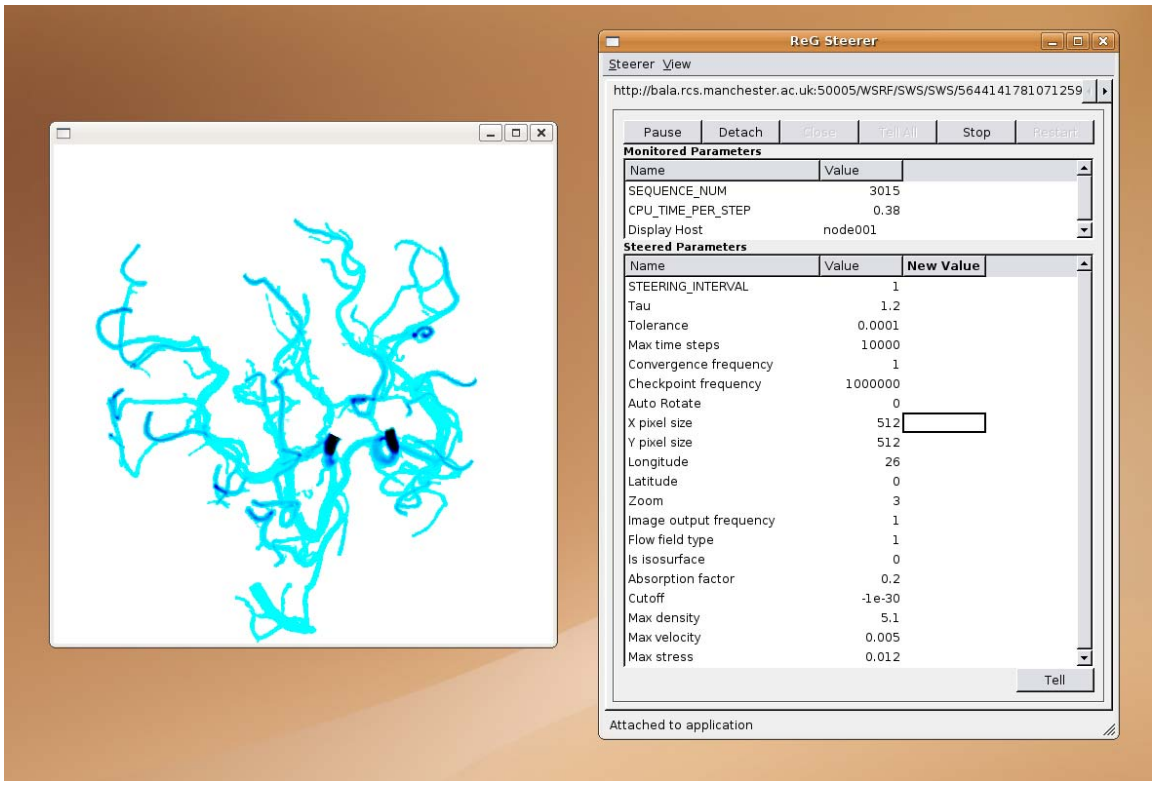
Lightpaths - dedicated national and international links to high-performance grid resources

All links are dedicated 1 Gb/s






# Real-time visualisation and steering



## VPH: Virtual Physiological Human


*Target outcomes:*

**Patient-specific computer models for personalised and predictive healthcare and ICT-based tools for modelling and simulation of human physiology and disease-related processes. Data integration and new knowledge extraction.**

- Several collaborative projects:
  - medical simulation environments for surgery;
  - prediction of disease/early diagnosis;
  - assessment of efficacy/safety of drugs

- Coordination and support actions:
  - enhancing security and privacy in modeling and simulation
  - international cooperation on health information systems based on Grid capabilities



## Concluding remarks I

- Clinical relevance of patient specific medicine
  - Both correctness and timeliness are important, fitting into current clinical practice
  - Batch-job submission won't work here

- Current emergency computing scenarios are far and few between (hurricane, earthquake simulations).
  - Successful patient-specific simulation techniques will likely have 1000's of cases. The level of compute time required will dwarf current resources.

- The cost, for example, HIV treatment, patient-specific response to 8 FDA approved drugs, 60,000 CPU hours, or 10 days of wall time (clinically acceptable time-frame).

- Economics of computational treatments
  - Using current available HPC resources, it would be impossible to conduct this day to day. Policies, who gets access? Do hospitals have in-house systems? Do supercomputers become public infrastructure? Much like utilities such as electricity?

## Concluding remarks II

- For widespread use, there are many moral, ethical and policy questions which need to be addressed
  - Resource availability
  - Data privacy, moving medical data around the grid, data anonymisation, data security, moving data to and from (often secure) hospital networks
- As such simulation becomes more widespread and embedded into the clinical process, markets will become available to supply the necessary resources, driving costs down.
- The hope is that the cost of simulation will be comparable or less than current medical treatments, saving money and time on ineffective treatments
- Ultimately, patient-specific computational data will sit side-by-side with traditional patient clinical records, further enhancing modern medical practice.

*"Distributed computing performed transparently across multiple administrative domains"*



## CTWatch Quarterly article, 17th of March 2008


Special issue on urgent computing

*"Life or death decision-making: The medical case for large-scale patient-specific medical simulations"*

*S. Manos, S. Zasada, P. V. Coveney*

**www.ctwatch.org**

*For more information… s.manos@ucl.ac.uk*