Workflow as the Methodology of Science:

An Application Perspective

Shantenu Jha
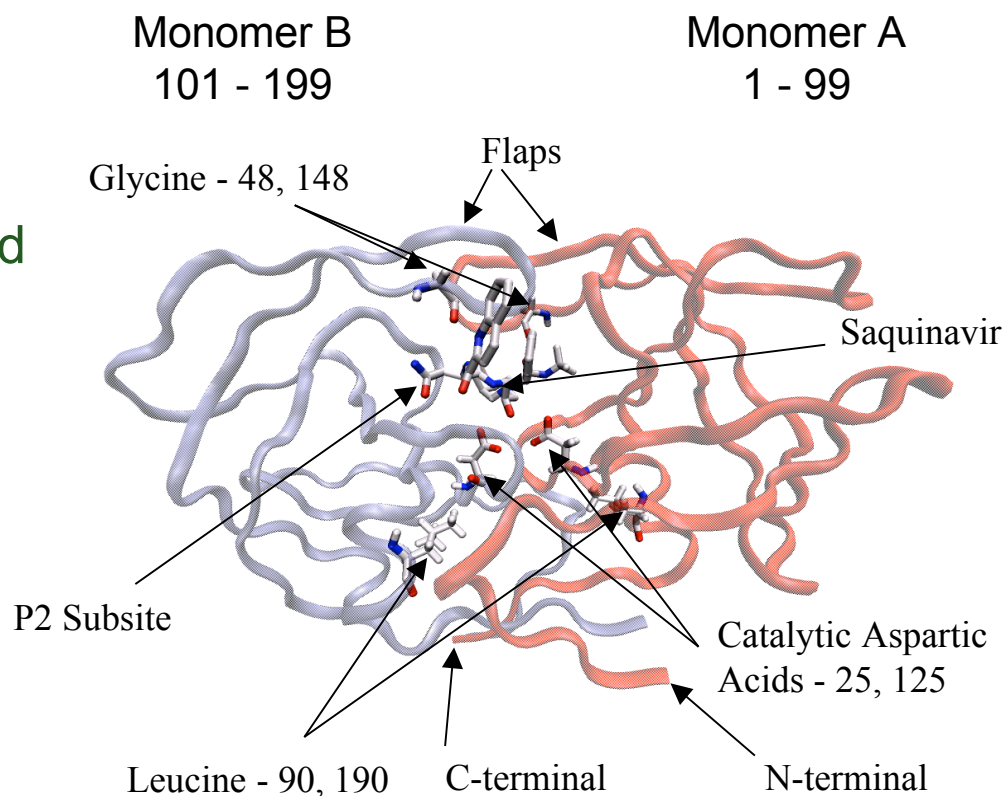
Centre for Computational Science, UCL

WORKS06,   Paris 2006

HIV-1 Protease

- Enzyme of HIV responsible for protein maturation
- Target for Anti-retroviral Inhibitors
- Example of Structure Assisted Drug Design
- 8 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

Monomer B
101 - 199

Monomer A
1 - 99

Flaps

Glycine - 48, 148

Saquinavir

P2 Subsite

Catalytic Aspartic
Acids - 25, 125

Leucine - 90, 190   C-terminal   N-terminal

WORKS06,   Paris 2006

RealityGrid

Molecular Dynamics Simulations of 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 by computing free energy (difference) values
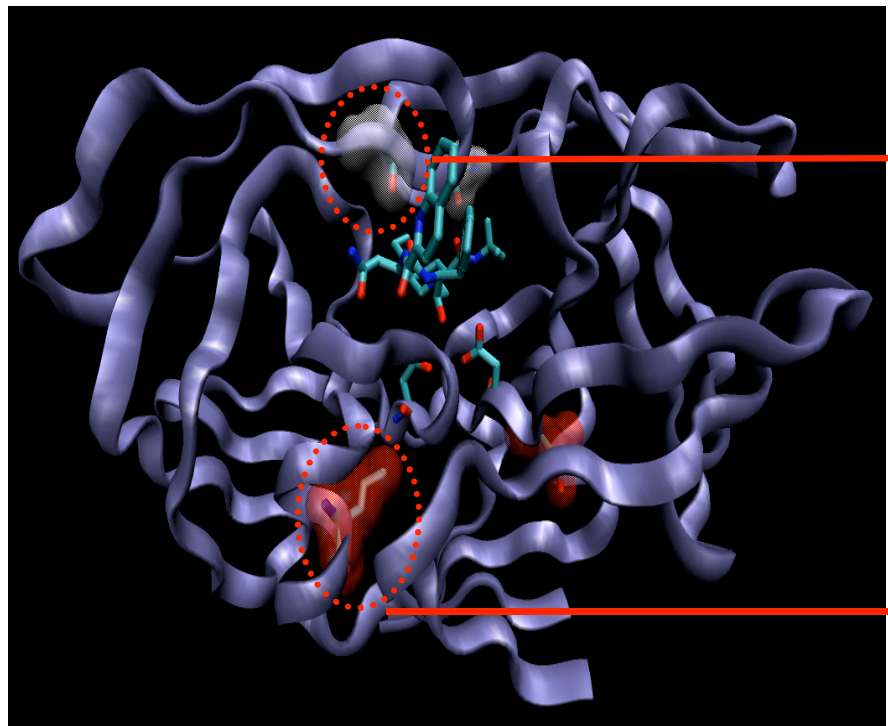
Mutant 1: G48V (Glycine to Valine)
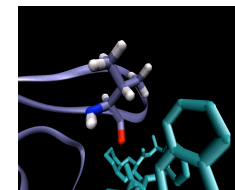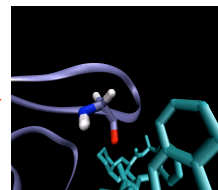
Inhibitor: Saquinavir
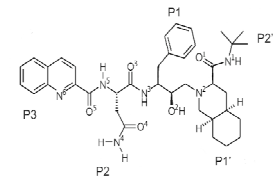
Mutant 2: L90M (Leucine to Methionine)

WORKS06, Paris 2006

UCL

RealityGrid
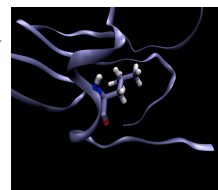
Compute Free Energy: Thermodynamic Integration

Starting conformation

V

t

Check each for convergence

Combine and calculate integral

$\lambda=0.1$ → time

$\lambda=0.2$

$\lambda=0.3$

Seed successive simulations

$\lambda=0.9$

Use steering to launch, spawn and terminate λ- jobs

$\partial V/\partial \lambda$

λ

Run each job on grid resource
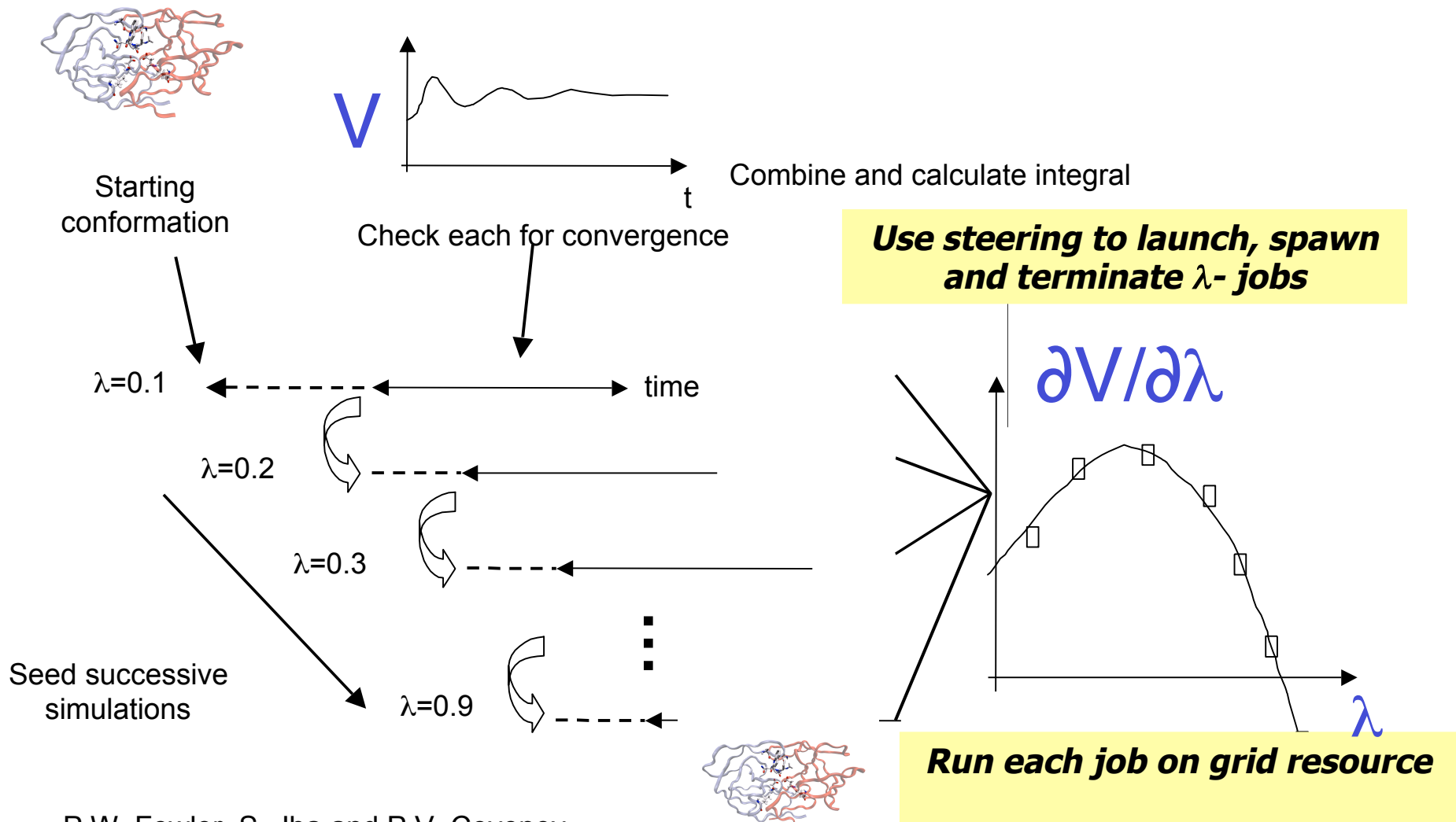
P.W. Fowler, S. Jha and P.V. Coveney,
Phil. Trans. R. Soc. A., **363**,1999-2015 (2005)

WORKS06,   Paris 2006

UCL

RealityGrid
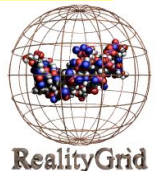
Other Computational Techniques

- **Replica Exchange Methods**
  Need for "intelligent" infrastructure to be coupled with analysis method

- **Ensemble MD**
  - Simulate each system many times from same starting position
  - Allows conformational sampling. Can't say how much *a priori.*

Start Conformation   *Series of Runs*   *End Conformations*

Equilibration Protocols
eq1 eq2 eq3 eq4 eq5 eq6 eq7 eq8

UCL

RealityGrid

Application Workflow

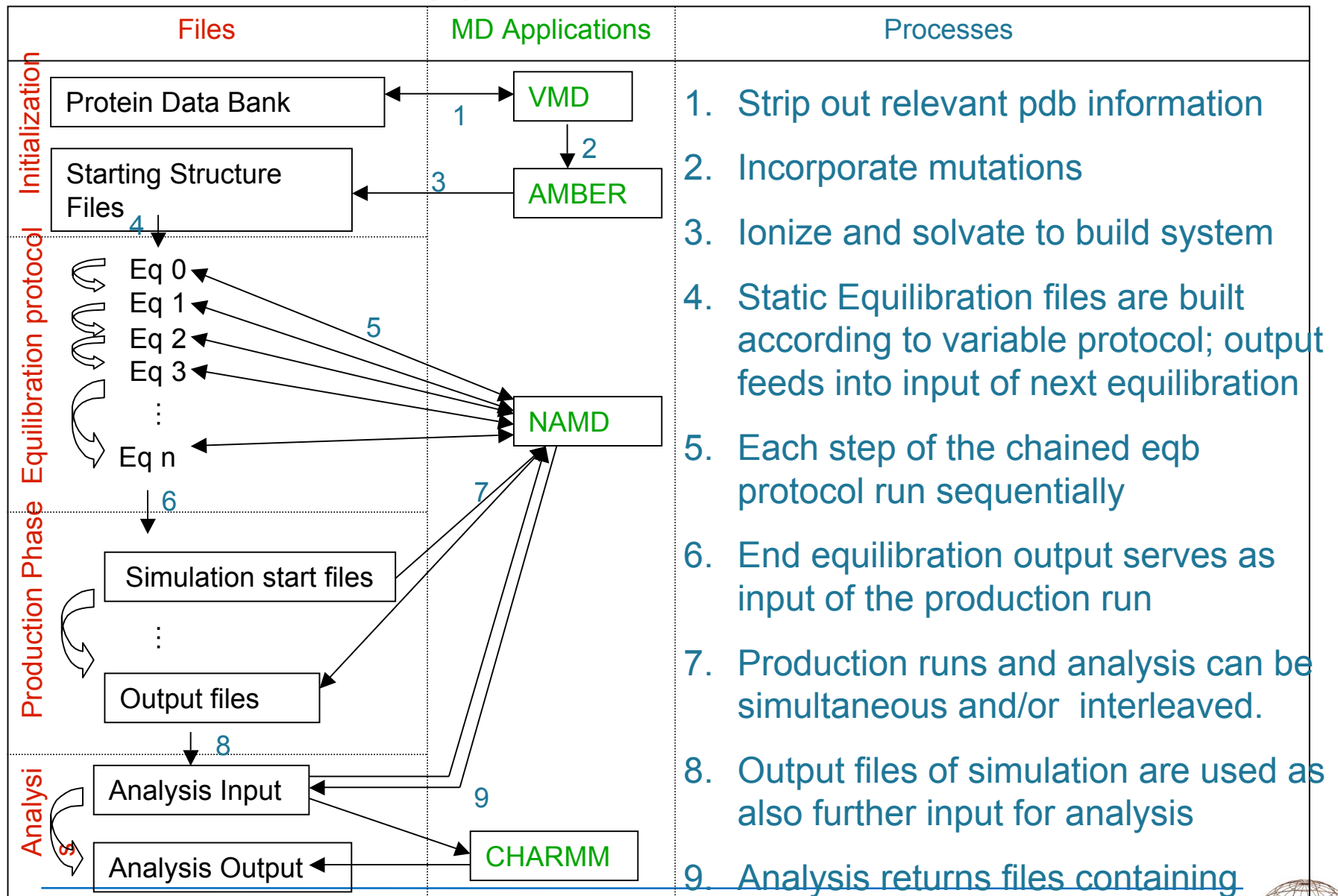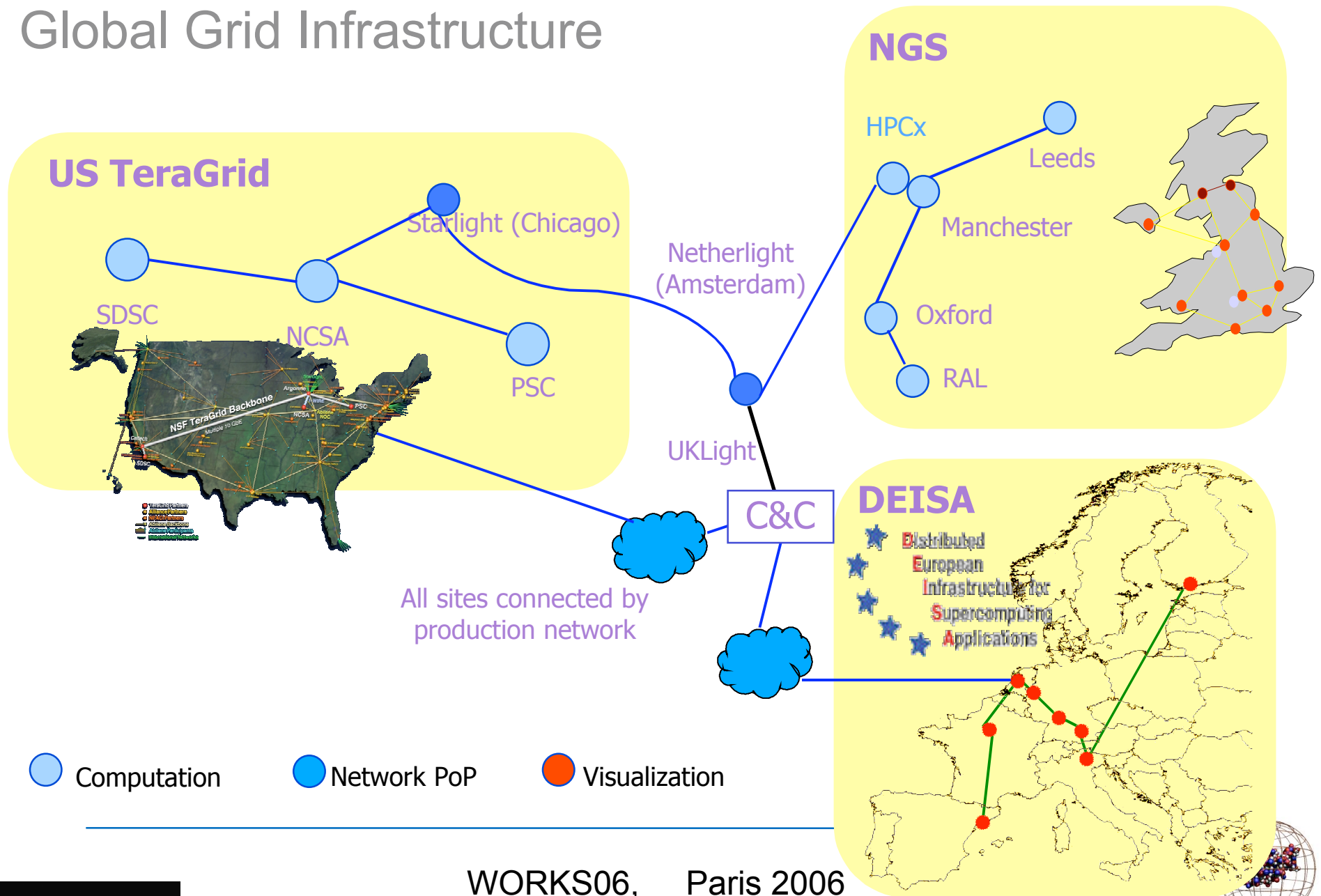
Files   **MD Applications**   **Processes**

Initialization

Protein Data Bank → VMD

Starting Structure Files ← AMBER

Equilibration protocol

Eq 0
Eq 1
Eq 2
Eq 3
⋮
Eq n → NAMD

Production Phase

Simulation start files
⋮
Output files

Analysis

Analysis Input ←
Analysis Output ← CHARMM

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 eqb protocol run sequentially

6. End equilibration output serves as input of the production run

7. Production runs and analysis can be simultaneous and/or interleaved.

8. Output files of simulation are used as also further input for analysis

9. Analysis returns files containing required data for end user. Shared

WORKS06,   Paris 2006

UCL

RealityGrid

Global Grid Infrastructure

NGS

HPCx

Leeds

Manchester

Oxford

RAL

US TeraGrid

Starlight (Chicago)

SDSC

NCSA

PSC

Netherlight
(Amsterdam)

UKLight

C&C

All sites connected by
production network

DEISA

Distributed
European
Infrastructure for
Supercomputing
Applications

○ Computation   ● Network PoP   ● Visualization

WORKS06,   Paris 2006

RealityGrid

High Throughput Science: Requirements

• Want to calculate "many" binding affinities rapidly

• Want to invest time in the analysis from the many jobs; not to manage the many jobs/data/output or manually launch each simulation in a different manner…

•Flexible: Use different methods mid-stream (dynamic), e.g., exchange an "exact" calculation scheme with an "approximation" scheme.

• High Throughput: Use many resources. Thus interoperable across different grids.

RealityGrid

Application Perspective: Some Thoughts

• Reiterate requirements from WS : Provenance, multiple "ities"

• It is easy to underestimate the complexity of the analysis process (e.g., how to determine the best time to spawn next sim). Often this analysis is not algorithmic or programmatic -- hence often can't be automated.

• Analysis (real-time interactivity and steering) and process used place constraints on level of task parallelism

• Workflow methodology to use should not be an execution time decision, but a fundamental organizing principle when constructing application, like which algorithm to use?

• Workflow infrastructure as a value-added feature of Grids. Must contribute to convenience of managing distributed resources (makes compelling for application scientists).

• Any workflow - composition, mapping, execution - MUST be simple, lightweight, low-barrier-to-entry and easy to use.

• What is the right level at which to provide user control over workflow composition? Hide it under a "hosting environment"? Provide Programmatic interfaces?