Drug Discovery on the Azure Cloud

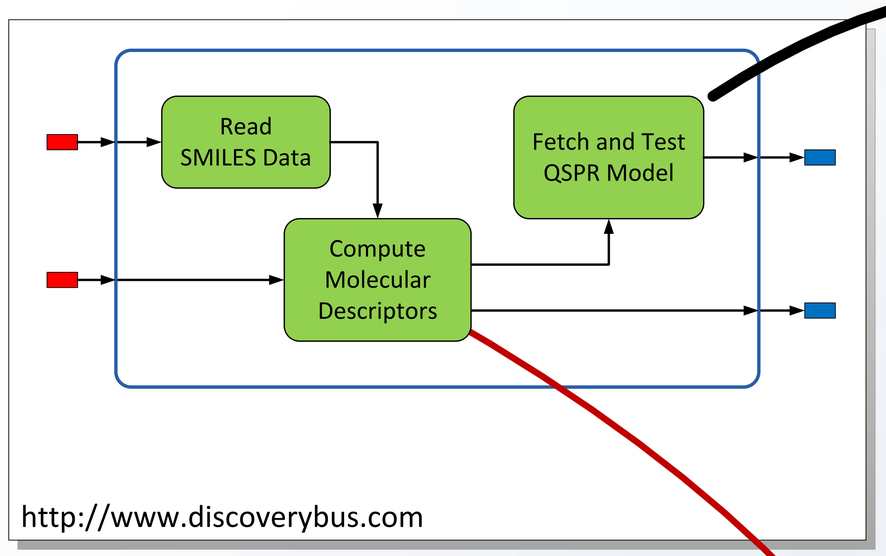
Paul Watson, Hugo Hiden, Simon Woodman, Jacek Cala {paul.watson, h.g.hiden, s.j.woodman, jacek.cala}@ncl.ac.uk

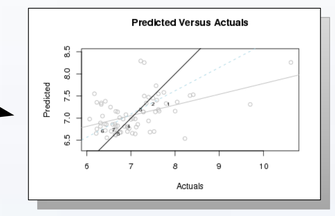
Newcastle University

The family of "kinase" enzymes are important biological targets since many are intimately connected to cell division and other important cell maintenance functions.

Scientist use a method known as **Quantitative Structure-Activity Relationships** (QSAR) to mine experimental data for patterns that relate the chemical structure of a drug to its kinase activity.

Molecular descriptors play the fundamental role in this process. These are numeric values that describe certain properties of molecules, such as electronic distribution and 3D shape configuration.



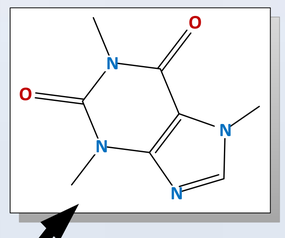


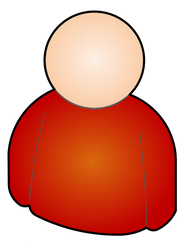
Discovery Bus is an autonomous QSAR system that allows exhaustive exploration of descriptor and model space, automated model validation and continuous updating.

Internally, Discovery Bus is a workflow engine designed specifically for drug discovery. It is, however, **difficult to parallelize processing on the Bus**.

Currently we run Discovery Bus on the Amazon EC2 cloud

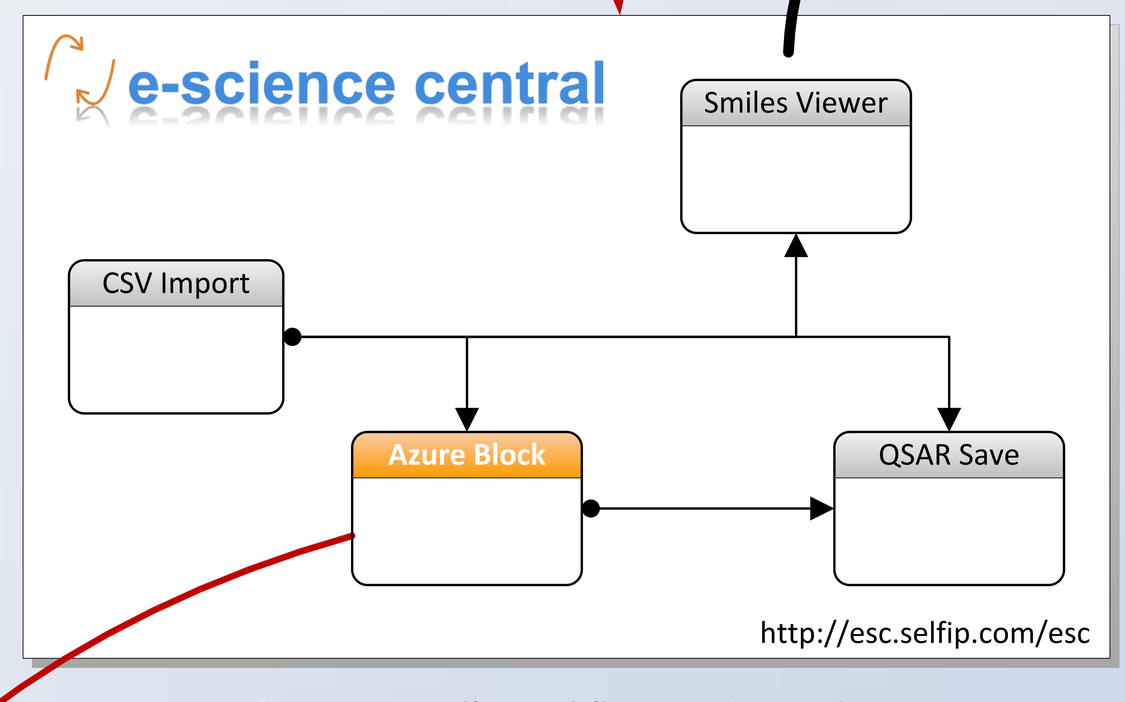
We combine workflows from Discovery Bus and e-Science Central to enable better parallelization



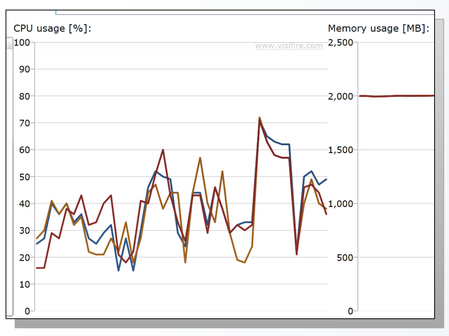


e-Science Central (e-SC) is a framework offering a generic web interface for managing scientific data, service and distributed processing

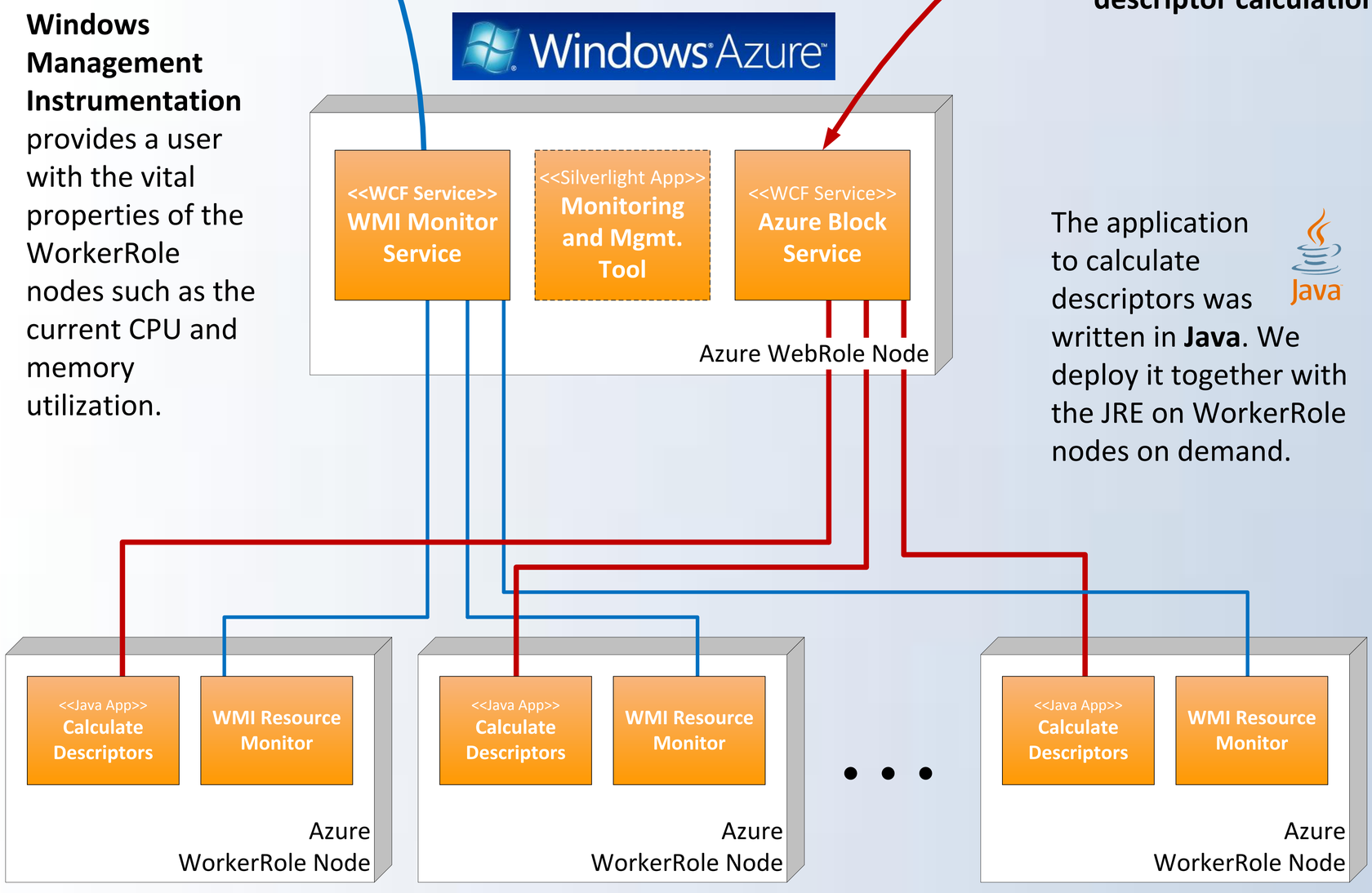
It allows scientists to collaborate and share results in a controlled manner, and gives them the ability to design and enact workflows



We use e-Science Central's workflow engine and a dedicated workflow block to delegate molecular descriptor calculation to Azure



Windows Management Instrumentation provides a user with the vital properties of the WorkerRole nodes such as the current CPU and memory utilization.



The application to calculate descriptors was written in **Java**. We deploy it together with the JRE on WorkerRole nodes on demand.

Predicting the properties of new structures requires significant computing resources. We use the Windows Azure cloud to speed up this process.

We were able to integrate these disparate technologies and provide scientist with a more effective tool.

In the future, we aim at correlating dynamic allocation of resources with cost prediction to offer users more efficient solution.

An administrator is allowed to change the number of WorkerRoles involved in descriptor calculation.