Running Large Workflows in the Cloud

Paul Watson
School of Computing Science & Digital Institute
Newcastle University, UK
Paul.Watson@ncl.ac.uk

The team: Jacek Cala, Hugo Hiden, Simon Woodman, David Leahy
With thanks to:
-Vlad Sykora, Martyn Taylor, Christophe Poulain, Savas Parastatidis
-Microsoft External Research & the EU (Venus-C) for their financial support
The Promise of Clouds

- Reduced capital expenditure
- Reduced Energy
- Scalability
- “On-demand resources”
The Problem with Clouds

Building scalable, reliable, secure systems on cloud infrastructure is still hard

- deep IT skills
- bespoke
- on-going management costs
- lock-in

Realising cloud advantages is beyond most who could benefit
Cloud Options

Cloud Infrastructure:
Storage & Compute

App 1

App n

Users

Cloud Platform

Cloud Infrastructure:
Storage & Compute

App 1

....

App n

Users
Science as a Service for users & programs

Cloud Platform for developers

Portable: run on
- Azure
- Amazon
- Private Cloud
Used for variety of research:
- activity recognition
- spectroscopy
- driving analysis
- chemistry...
Chemists want to know:

**Q1. What are the properties of this molecule?**

**Q2. What molecule would have aqueous solubility of 0.1 μg/mL?**
Answering the Question by performing experiments

Small molecule + Protein → Experiment → 3.5 nM

..... time consuming, expensive, ethical Issues
An alternative to experimentation: QSAR

Quantitative Structure Activity Relationship
- predict properties based on similar molecules

\[ \text{Activity} \approx f(\text{quantifiable structural attributes, e.g. } \#\text{atoms, logp, shape, ...}) \]
Generating the models - Discovery Bus (Leahy et al)

Model Generation

Data

Model- Builders

Models

New Data or Model-Builders

www.openqsar.com

New/ Improved Models
Increasing amounts of data for model building...

**CHEMBL**: data on 622,824 compounds, collected from 33,956 publications

**WOMBAT**: data on 251,560 structures, for over 1,966 targets

**WOMBAT-PK**: data on 1230 compounds, for over 13,000 clinical measurements

- More models
- Better models
- × est. 5 years to process new datasets on existing server
JUNIOR Project Aim

Use Azure to generate models in weeks not years
Potential for parallelism...

1. Separate Training & Test Data
2. Calculate Descriptors from Structures
3. Combine Descriptors
4. Filter Descriptors
5. Select Descriptors + Responses
6. Build & Test Models Independently:
   - Multiple Linear Regression
   - Neural Network
   - Partial Least Squares
   - Classification Trees
7. Select Best Models
8. Add to Model Database
Approach #1
Minimal change to Discovery Bus

Diagram:
- Discovery Bus Co-ordinator
- Amazon
- App API
- Security
- Workflow Enactment
- Analysis Services
- Social Networking
- Provenance
- Metadata
- Storage
- Processing
- e-Science Central
- Azure
Succeeded in reducing time from (est.) 5 years to 3 weeks, but…

![Throughput Graph](image-url)

**Throughput (Requests/min)**

- X-axis: Azure Nodes
- Y-axis: Throughput (Requests/min)
Approach #2

• run entirely within Azure
  – through e-Science Central on Azure
7 Workflows with 67 blocks
Big run:
- 460K workflow invocations
- 4.4 million service invocations
Approach #2
Early Results

- Reduces time to 9 days
  - 5yrs → 22 days → 9 days
  - but room for improvement....
Scalability

![Graph showing scalability with execution time vs. number of workflow engines. The graph compares ideal and actual performance.](image-url)
• *Discovery Bus* exemplifies a good Cloud pattern
  – large, variable, bursty requirements
• e-Science Central is a scalable, secure, portable cloud platform for Azure (and Amazon, and Private Clouds)
• next steps
  – optimize large workflow scheduling
  – automatically adapt #workers