Volunteer Crowd Computing and Federated Cloud developments

David Wallom
Overview

• Crowd Computing through Climateprediction.net
• Cloud Computing with the EGI Federated Cloud
• Utilising cloud resources for Crowd Computing
The world’s largest climate modeling facility

>300,000 registered volunteers, ~40,000 active, 127M model-years
Climate modelling with distributed computing

- Disadvantages:
  - Limited diagnostics & resolution.
  - You make all your mistakes in public.

- Advantages:
  - Effectively unlimited ensemble size.
  - You make all your mistakes in public.
• What has distributed computing allowed us to do that we would not have done otherwise?
Unlimited ensemble size: exploring uncertainties in climate predictions

Results of the BBC Climate Change Experiment: Rowlands et al, *Nature Geosci.*, 2012
The weather@home regional modelling project
(with Microsoft Research, the Risk Prediction Initiative and Environment Guardian)

• High impact weather events are typically rare and unpredictable.
• They also involve small scales.
• Resolution provided by nested regional model.
• Compare numbers of events in the 1960s and 2000s to explore trends.

• Modify boundary conditions to mimic counter-factual “world that might have been”.

This map shows your climate model running
Now running regional model date and time (UTC): 02/12/1966 20:40

climatemodel.net

Use keyboard keys to change view
C - Clouds & surface
T - Temperature
H - Rain & Snow
P - Pressure
V - Potential vorticity
G - Show/Hide grid
W - Global/Regional views
H - Help & more options

Current view:
Temperature (°C)

Temperature is one of the clearest indicators of climate change. Watch how it changes over the model day and year.
And that grew into…

- **Weather@home - Regional climate modeling projects**
  - Weather@home 2014: the causes of the UK winter floods
  - Weather@home ANZ 2013: the causes of recent heatwaves and drought in Australia and New Zealand
  - Weather@home Climate Accountability: the causes of extreme heat in the Western US
  - Weather@home ACE-Africa
  - Weather@home: High Resolution 2003 European Heatwave
  - Forest Mortality, Economics, and Climate in Western North America (FMEC)
What is the role of increased greenhouse gas levels in UK autumn/winter flood events?

South Oxford on January 5th, 2003

Photo: Dave Mitchell
What is the role of increased greenhouse gas levels in UK autumn/winter flood events?

South Oxford on July 24th, 2007
What is the role of increased greenhouse gas levels in UK autumn/winter flood events?

South Oxford on July 24th, 2007

Oxfordshire has flooded before so how do we quantify the role of human influence?
An accidental spin-off: the Pall et al experiment

- Aim: to quantify the role of increased greenhouse gases in precipitation responsible for 2000 floods.
- Challenge: relatively unlikely event even given 2000 climate drivers and sea surface temperatures (SSTs).
- Approach: large (multi-thousand-member) ensemble simulation of April 2000 – March 2001 using forecast-resolution global model (90km resolution near UK).
- Identical “non-industrial” ensemble removing the influence of increased greenhouse gases, including attributable SST change, allowing for uncertainty.
- Still operated as a historical experiment
Risks of floods in the Pall et al ensemble

A flood that happened – and one that did not

→

40% decrease in risk

←

100% increase in risk

Pall et al and Kay et al, 2011
David Cameron 'very much suspects' climate change is behind recent storms
Prime minister concedes increase in abnormal weather events could be linked to global warming

Rowena Mason, political correspondent
The Guardian, Wednesday 8 January 2014 14.38 GMT

Severe flooding of the Thames brings swans on to a caravan park at Chertsey, Surrey, on Wednesday. Photograph: Peter Macdiarmid/Getty Images
Weather@home 2014: the causes of the UK winter floods

Generated at peak over 1.5TB data per day in returning models

 Academic paper quality needed 120k runs for increased statistics
The changing mode of Climateprediction.net

- Original focus on multi-decade prediction: challenging results, and challenging experiment.
- Increasing interest in event attribution
- Timeliness of results and processing becomes important, Science as a Service?
- Can we operate in a forecast mode for likelihood of extreme events?

- Concern though that new experiments could become ‘volunteer/crowd limited’

- Science as a Service would imply measurable deliverables including time to results…
The EGI Federated Cloud
The EGI Federated Cloud, a federation of institutional private Clouds, offering Cloud Services to researchers in Europe and worldwide

A single cloud system able to

- Scale to user needs
- Integrate multiple different providers to give resilience
- Prevent vendor lock-in
• 12 countries provide 19 certified resources
  - Czech Republic, Germany, Greece, Hungary, Italy, Macedonia, Poland, Slovakia, Spain, Sweden, Turkey, United Kingdom

* Not shown on map
Federated Cloud Services

Tier 1: Reliable Infrastructure Cloud

Tier 2: General-purpose platform services

Tier 3: Platform as a Service

Tier 4: Zero ICT Infrastructures

Federated IaaS and STaaS Cloud

PaaS

DBaaS

HadoopaaS

VRE

Virtual eLaboratory

Secure storage

Key Mgmt Encryption ACL mgmt
• **Ecology** – BioVeL: Biodiversity Virtual e-Laboratory

• **Structural biology** – WeNMR: a worldwide e-Infrastructure for NMR and structural biology

• **Linguistics** – CLARIN: ‘British National Corpus’ service (BNCWeb)

• **Earth Observation** – SSEP: European Space Agency’s Supersites Exploitation Platform for volcano and earthquakes monitoring (Collaboration with Helix Nebula)

• **Software Engineering** – SCI-BUS: simulated environments for portal testing

• **Software Engineering** – DIRAC: deploying ready-to-use distributed computing systems

• **Software Engineering** – Catania Science Gateway Framework

• **Musicology** – Peachnote: dynamic analysis of musical scores

• **Earth Observation** – ENVRI: Common Operations of Environmental Research infrastructures (collaboration with EISCAT3D)

• **Geology** – VERCE: Virtual Earthquake and seismology Research

• **Ecology** – LifeWatch: E-Science European Infrastructure for Biodiversity and Ecosystem Research

• **High Energy Physics** – CERN ATLAS: ATLAS processing cluster via HelixNebula
New EGI FedCloud Communities since launch

- **Education** – Cranfield University distributed systems course
- **Cultural Heritage** – DCH-RP management of preservation services in the cloud
- **Hydrological Modelling** – Running Hydrological models to support real time analysis
- **Bioinformatics** – ELIXIR execution of the Ensamble application in the Federated Cloud environment
- **Systems implementations** – deployment of FTK developed tools and services and data preservation
- **Internet of Things** – Smart Grid systems investigation
- **Software Development** – deployment of research PaaS
- **RNA Sequencing** – deployment of analysis engines in the cloud
- **Physiological Modelling** – Calibration, scenario mapping and development

• Crowd and Cloud?
Where can the cloud help?

1. Volunteers – Short term demand response due to multiple models launched concurrently exceeding volunteer capacity
2. Volunteers – Demand for high resolution models exceeds standard volunteers systems [e.g. multicore coupled models]
3. Volunteers – Supporting 3 platforms is time consuming and difficult, distribute VMs
4. Servers – Large numbers of high resolution models generate vast amounts of data -> upload servers in the cloud with coupled analysis servers
5. Servers – Resilience of central core services becomes more important with global partners
Demand Response

- Distribute in the cloud multiple cloud volunteer BOINC clients
- Simplified since no console -> no graphics
- Must not put off or make it appear replacing volunteers
High resolution

- Next generation coupled models designed for multi/many core

- Some implementations of multicore support are platform dependant

- Current efficiency gained through utilising multi-core for separate models
Multi-Platform

- Supporting 3 platforms [Windows, Linux and Mac] becomes time consuming

- Simplifying the application creation procedure is key to supporting multiple models

- Distribution of apps as VMs or even containers?
Data Storage?

- Individual experiment can exceed 20TB
- New SciaaS applications could increase by OoM
- Need scalable storage services coupled with analysis capability to prevent undue movement of BIG DATA
Resilient core servers

- CPDN/WatH has global partners, launching work on their own timescales

- Core services must be reliable to support this

- Scale hybrid cloud from current VMWare located services to external cloud services
Server Infrastructure

- Oxford Volunteer Computing group supporting multiple proposals for Crowd Computing
- Some funders require ‘exploratory work’ to show promise before funding
- Connect cloud volunteers to cloud server stack to accelerate time to start
Conclusions

- CPDN moving towards SciaaS or at least ‘Data sets aaS’
- Federated cloud services are a reality across Europe through EGI
- Open standards give user confidence to build tools and services against known interfaces
- Cloud can support different aspects of Crowd Computing ‘lifecycle’