Introduction to the Camtology group and the camont virtual organisation

Karl Harrison

University of Birmingham

UKI Monthly Operations Meeting
16th April 2009

Camtology group

- The camont virtual organisation (VO) was set up towards the end of 2006, and is used by the Camtology group
    - Aim at building a next-generation internet search engine
- Camtology is a joint venture between two Cambridge start-up companies
    - Imense founded by David Sinclair and Chris Town
        - Software developed to allow image retrieval based on image content
    - iLexIR founded by Ted Briscoe
        - Expertise in text analysis, mining, classification, search applications
- Grid-related activities carried out in collaboration with physicists from HEP groups at Cambridge and Birmingham
- Funding support from STFC through PIPSS (knowledge-transfer) programme
    - Two related STFC press releases issued
        - http://www.stfc.ac.uk/PMC/PRel/STFC/imense1.aspx
        - http://www.stfc.ac.uk/PMC/PRel/STFC/ImenseiLexIR.aspx

The camont virtual organisation

- The camont VO is hosted by the GridPP VOMS server
  - https://voms.gridpp.ac.uk:8443/voms/camont/webui/
- The VO is currently enabled on one WMS at RAL (lcgwms03) and at nine GridPP sites (Birmingham, Brunel, Cambridge, Durham, Glasgow, Lancaster, Oxford, Royal Holloway, RAL PPD)
- Small number of active members
  - VO manager: Andy Parker
  - Software manager: Frederic Brochu
  - Job preparation and submission: Karl Harrison, Mark Slater
  - Advice and suggestions: Jeremy Coles, Santanu Das, Mark Hayes

⇒ Grid jobs for camont VO are run by people with good understanding of Grid technology

⇒ Extensive testing performed for each new type of job before submitting in large numbers

Camtology activities on the Grid

- Imense has made significant use of the Grid over the past two years
  - Activity concentrated in bursts of a few weeks, with software development in between
  - Main Grid use has been for analysing image content
    - Record of almost 20 million stock photographs (around 6 TByte of data) analysed during 4-week period, November-December 2008
      - Up to 500 Grid jobs run in parallel (150-300 more common)
      - Information collected on Grid performance (surprisingly good!)
  - More recently have used Grid to search for images in selected domains (Universities, Government agencies, museums, etc)
    - Good demonstration of technology, although general quality of images obtained has been a bit disappointing
- iLexIR Grid activities just starting
  - Aim to analyse textual content of some 50000 scientific papers
    - Relatively modest processing requirements
  - Aim to create n-gram corpus based on around 10^{12} words from English-language web sites
    - Large processing requirements to achieve this in reasonable time
    - Details discussed in separate presentation