Open-Source Software in the Scientific World

Case study: Triana Software

IAN J.TAYLOR
Cardiff University, UK
Center for Computation & Technology,
Louisiana University, USA

ADINA RIPOSAN
Military Technical Academy,
Contact Net Ltd,
Bucharest, Romania

May 18-19, 2007

eLiberatica 2007
Brasov, Romania

TRIANA

Workflows Environment

May 18-19, 2007

eLiberatica 2007
Brasov, Romania

Open-Source scientific environment for workflow composition:

=> more than 500 applications are developed based on Triana in support to scientific groups around the world

- Can integrate within a number of different distributed environments

  => for allowing true heterogeneous computing across different *Grids* and *distributed paradigms*

- Can specify distributed course-grained service workflows

http://www.trianacode.org/

Some Examples of Domains

- Gravitational wave data analysis (GridOneD)
- Radio astronomy (with Manchester)
- Astrophysical simulations (Cactus)
- Data mining (DIPSO, Data mining Grid)
- Biodiversity Problems (Bdworld)
- Galaxy visualization
- Audio processing and distributed music information retrieval (MIR)
- Distributed peer-to-peer simulations (NRL and AgentJ)
- Grid-enabled medical simulations GEMSS)
- Environmental science (INFERNO)
- E-Health (Contact-Net)

Projects

- GEO 600 project - http://www.geo600.uni-hannover.de/
  (gravitational wave data analysis)
- GridOneD - www.gridoned.org
  (distributed computing - P2P and Grid)
- GridLab - http://www.gridlab.org/
- DataMiningGrid - http://www.datamininggrid.org/
- BiodiversityWorld - http://www.bdworld.org/
- GEMSS - http://www.ccrl-nece.de/gemss/index.html
  (Grid-enabled medical simulations)
- ReSC - http://www.resc.rdg.ac.uk/projects.php
- DART - http://www.mrsdart.com
  (Digital Audio Retrieval using Triana)

Triana environment is used for problem solving and orchestrating flows of operations/services

=> *fine-grained dataflow applications*

=> *course-grained distributed workflow system*

Workflow is implicit in scientific algorithms that specify:

- a series of *inter-dependent operations* to be executed,
- connecting such algorithms in a series of *derivations*,

=> when aggregated perform some higher-level task

Workflows can be:

- simple => contain a few components
- complex => logic-based support can be integrated to make intelligent decisions about the dynamic evolution of the particular workflow

Tasks with inter-dependencies expressed and handled by a computation flow

⇒ the chain of elementary tasks are not necessarily linear

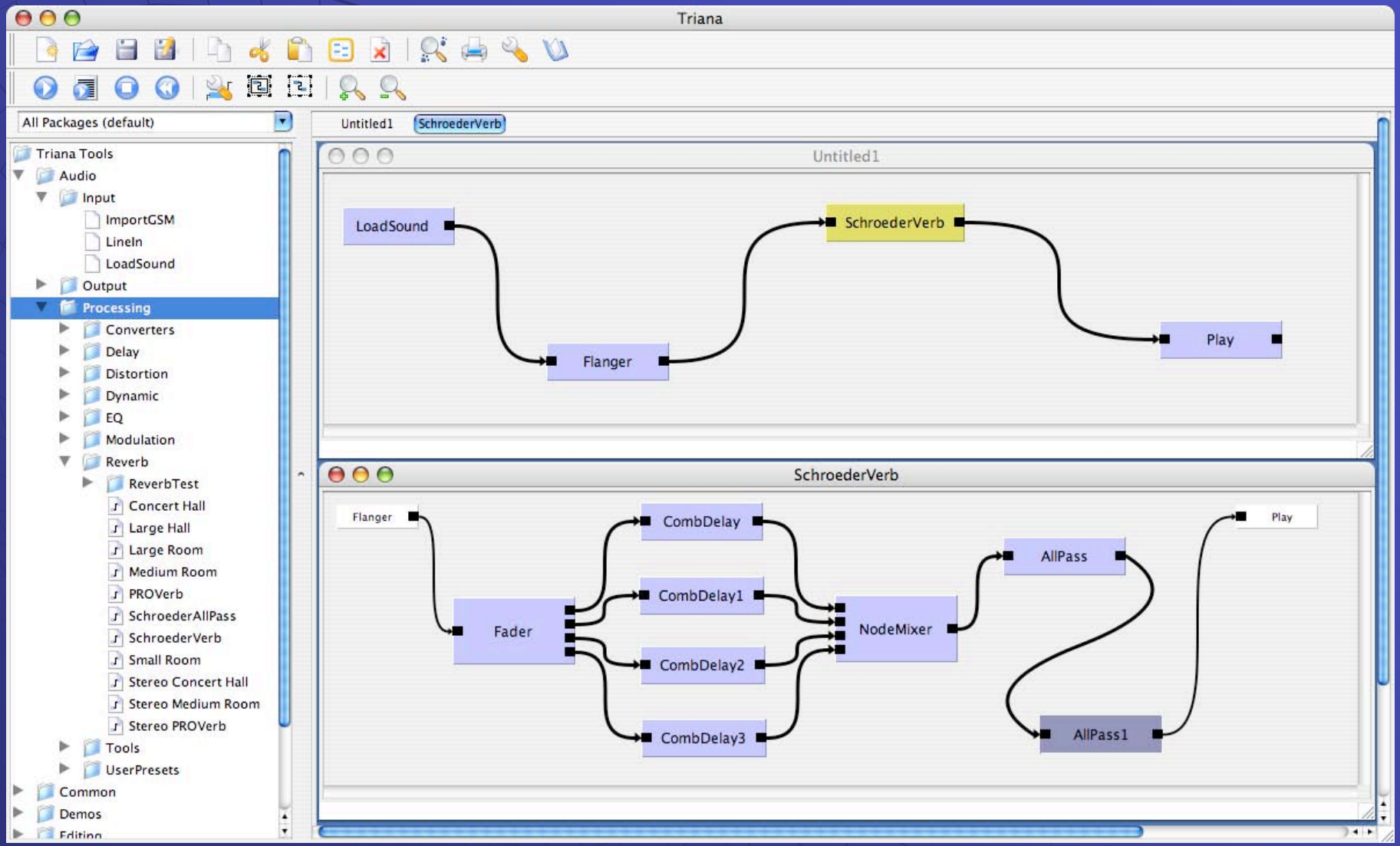
⇒ request a graph of interconnected tasks

Workflow management is data driven:

- the scientific experiments need to process large datasets
- the scheduler responsible for distributing the computational load should take into account the input dataset as well as the workflow graph topology

Triana Workflows

- multimodal, multimedia -

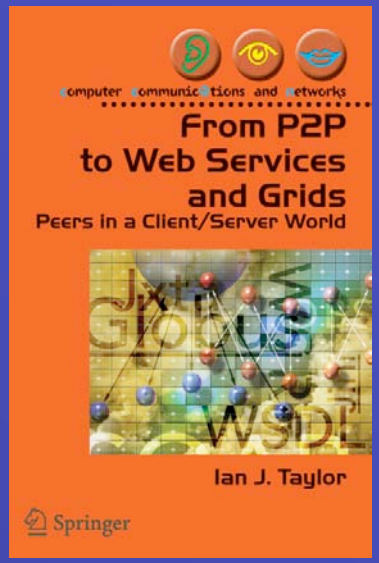


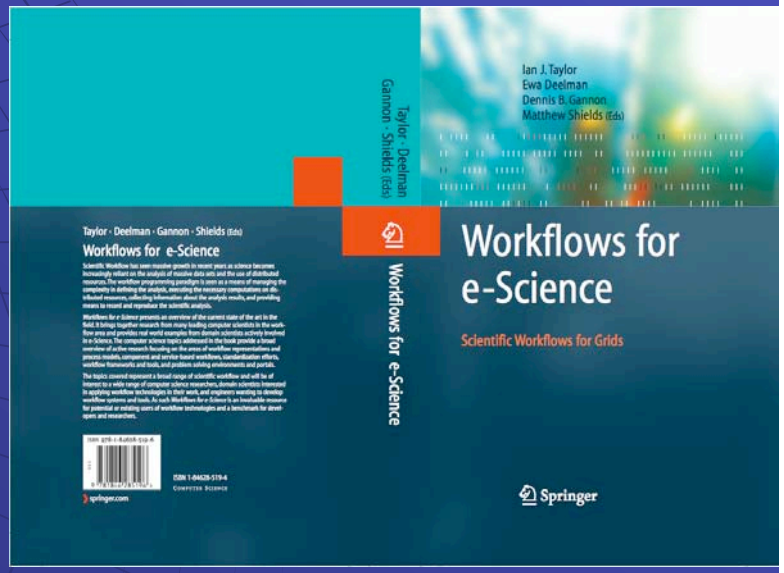
Based on Triana => *The Alchemist Infrastructure*

- A new paradigm in search and discovery of distributed resources, based on:
  - multimodal workflows
  - the coupling of metadata fusion & social tagging with the more traditional index-based search techniques

Information sources:

- http://www.trianacode.org/
- http://www.wspeer.org/
- http://www.trianacode.org/p2ps/





- WSPeer has been Triana's Web Services toolkit for the past three years and many projects have used this combination to specify their distributed course-grained service workflows

  http://www.wspeer.org/

  WSPeer - existing middleware which provides:

  - a SOAP messaging layer (using Web Services / WS-RF)
  - within a P2P network that supports a super-peer topology of *rendezvous* or *advert caching* peers

    =>to support the scalability of the discovery & access to information

    =>to cache application-specific data, scientific data & metadata

    (not just discovery information)

- P2PS has been used as the underlying P2P environment during this time

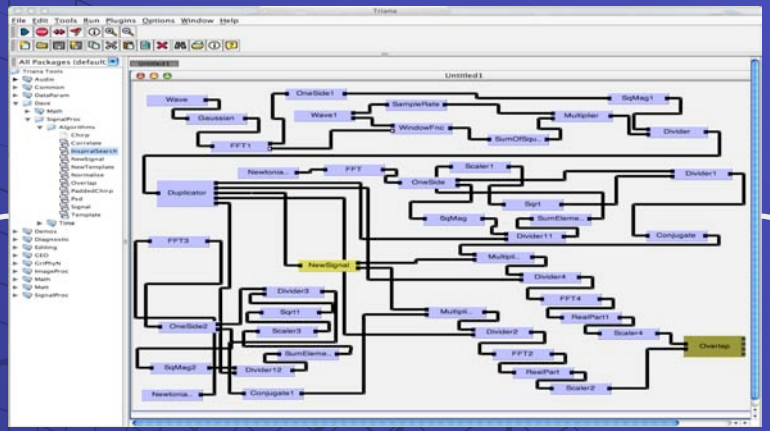
  http://www.trianacode.org/p2ps/

- Triana provides a number of different bindings to underlying middleware and therefore a number of possible modes of operation.
  - Triana on the one hand has a full binding to Java GAT interface, capable of invoking tools and services such as Condor, GridFTP, GRAM, etc.
  - On the other has integrated with service-based middleware, such as Web Services, WS-RF, Jxta and P2PS.
  - It also has the capability to dynamically wrap applications remotely behind Web Services interfaces so that existing software can be easily integrated.

Triana, the GAT and the GAP

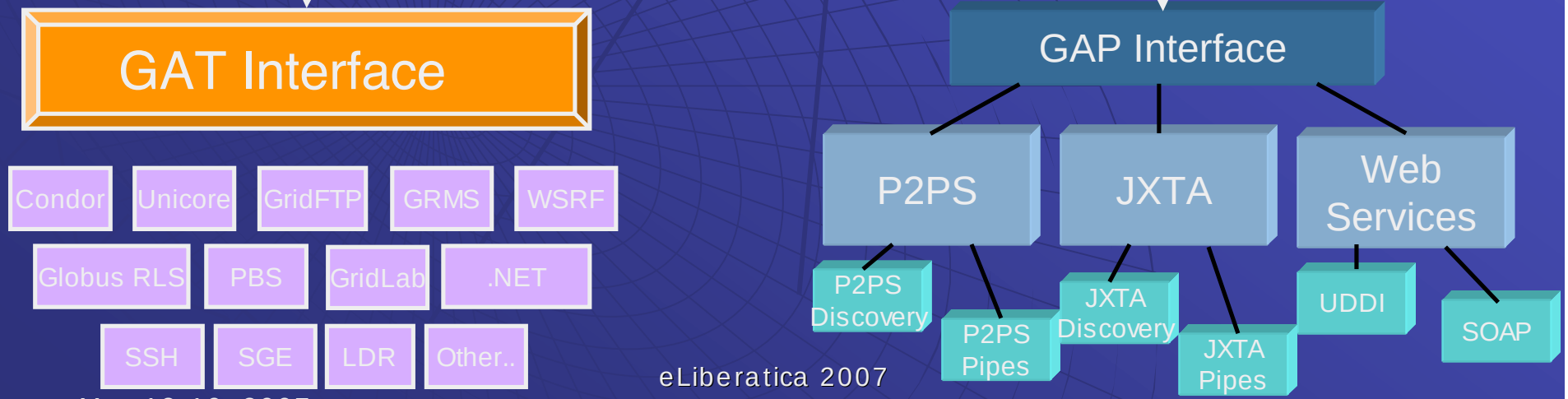
Grid Computing:

Job Submission,
File services
A Graphical Grid Computing Environment or Portal



Service Based Computing:

Deployment, discovery and communication with distributed services e.g. P2P and (GSI) Web services



Triana Focus

- Two core underlying focuses:
  - Interactive graphical programming of the distributed tasks - complex editing
    - Intuitive drag/drop flexible editing - copy/paste services, wizards for creating tools/toolboxes, user interfaces, adding nodes and multi-level grouping.
    - Has been used as a “graphical editor” for other languages, e.g. DAG, VDLx (DAX in progress).
  - Heterogeneous workflows - Bridge the gap between different distributed environments
    - Use cross-environment interfaces
    - led to integration with GAT (pre SAGA), GAP

Triana allows to spontaneously create and run data analysis algorithms on the data at its source.

- Such a component-based object-oriented approach allows scientists to easily create new algorithms that conform to an agreed and defined set of data types and can adapt to different internal parameters

- Makes it easy to create individual user interfaces for each component to allow the modification of its internal parameters

- Graphical approach

  => REUSABILITY of existing units

  => Allows simple type-safe orchestration of data analysis pipelines on-the-fly without the need for code-level reconfiguration.

Triana enables project scientists to design and create systems than connect together a number of software components.

ALCHEMIST

Workflows for Data Discovery

May 18-19, 2007

eLiberatica 2007
Brasov, Romania

The Alchemist framework:

Project at Cardiff University

Domain-independent workflows & search mechanism, built on a generic P2P (Peer-to-Peer) architecture, supporting:

- distributed database queries
- complex search algorithms based on *workflows*
  - composed as a collection of Peer-to-Peer overlays, Grid-based services and distributed workflows

⇒ Uses industry standards such as *Web services* and *SOAP for messaging*

⇒ Alchemist framework & tools are *extensible & interoperable*

Alchemist framework is an alternate approach to the classic Internet search engines

built on top of decentralised technologies

allowing users to proactively push information into a decentralised "search database"

using standardised Web Services interfaces,

- developed within the business & Grid computing communities,
- and hosted on a Peer-to-Peer (P2P) infrastructure

The system already interfaces with existing Grid middleware (e.g. Globus) through Web Services interfaces

Distributed P2P database framework

For decentralizing metadata & data

- Based on Web Services technologies
- Unstructured P2P i.e. in the style of super-peers but
  - Allows different overlays to be created (data caching) - dynamic grouping
  - Allows different caching policies (replication, forwarding overlays etc) for groups
  - Allows sophisticated Grid-style security sign-on, delegation
- Based on existing technologies
  - Triana, P2PS and WSPeer

A framework providing a P2P layer for supporting:

- pluggable *network discovery & caching overlays*
- the ability to execute *distributed workflows*

Dynamic overlays can be created *on-the-fly* for the particular application

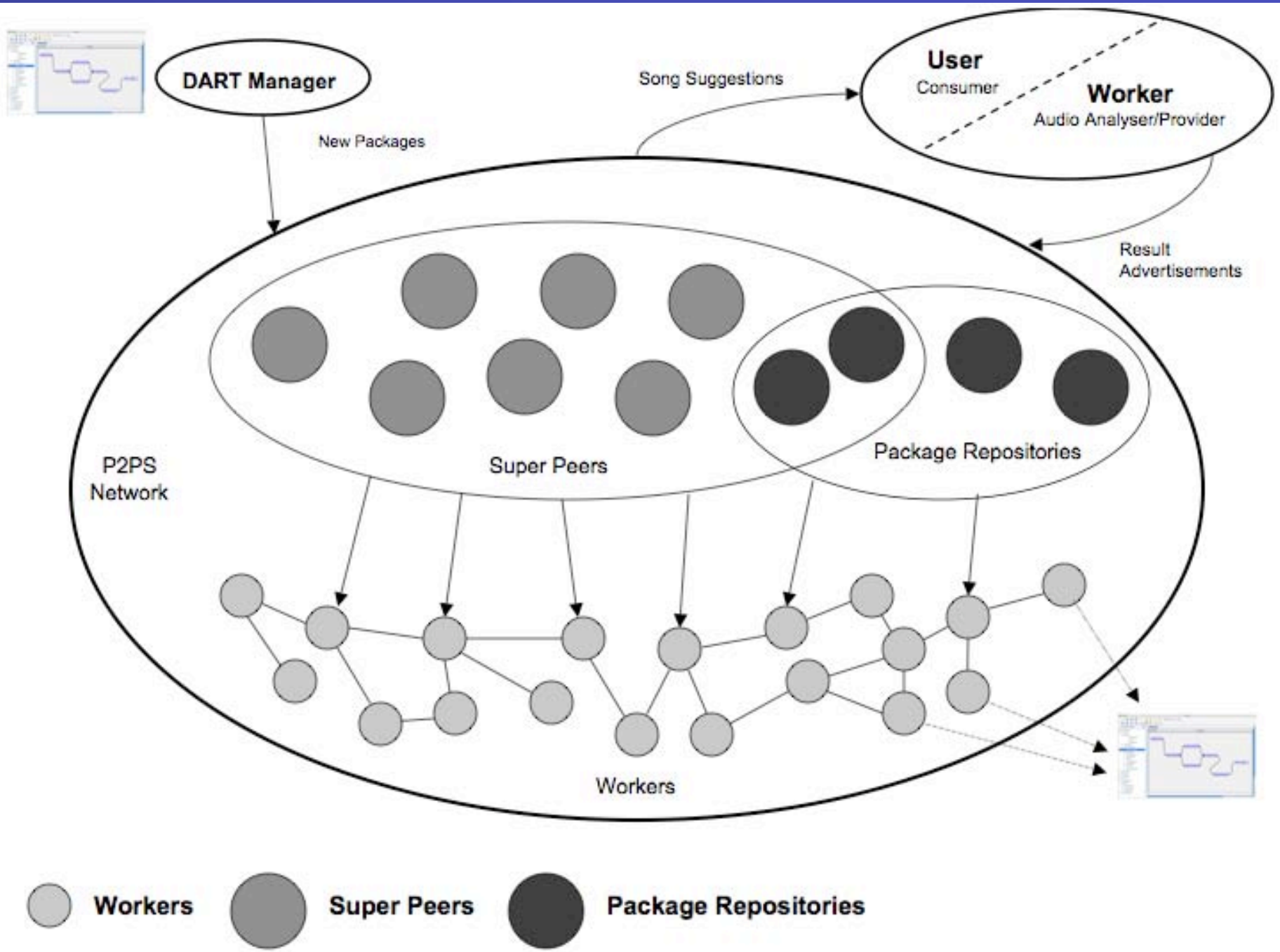
=> deployed onto the peers through the use of P2P groups:

- Security services for participating overlay
- Membership services, Group services

Workflows packages are dynamically propagated onto the network using an overlay of package repositories

=> decentralised layer of package repository cachers

(created by Alchemist's dynamic overlay mechanism)


DART Manager
New Packages
Song Suggestions
User
Consumer
Worker
Audio Analyser/Provider
Result
Advertisements
P2PS
Network
Super Peers
Package Repositories
Workers
Workers
Super Peers
Package Repositories

Built on existing well-tested technologies:

- Triana Workflow Environment http://www.trianacode.org/
  - Can specify distributed course-grained service workflows
  - Used in *radio astronomy, astrophysical simulations, gravitational wave analysis, data mining, biodiversity problems, grid-enabled medical simulations, environmental science, audio processing etc.*

- WSPeer - existing middleware http://www.wspeer.org/
  (Triana's Web Services toolkit) which provides
  - a SOAP messaging layer (using Web Services / WS-RF)
  - within a P2P network that supports a super-peer topology of *rendezvous* or *advert caching* peers

    =>to support the scalability of the discovery & access to information

    =>to cache application-specific data, scientific data & metadata

    (not just discovery information)

- P2PS - the underlying P2P environment http://www.trianacode.org/p2ps/

Alchemist framework can be customised for particular fields / user-environments

=> Biomedical research
=> Astrophysics research
=> Audio research

- Alchemist toolkit integrates applications, data providers, digital content, and algorithms

  => enables the simple composition of mixed-media queries for combinational searches

  => to interpret heterogeneous datasets in a logically defined order

  => to multiplex search results

  => produce rich metadata

- Graphical workflow builder => Alchemist provides a framework for specifying complex search algorithms, using a series of logical search steps

  => application developers do not need to write custom software algorithms from scratch

  => are able to create complex queries and data fusion techniques in a modular and pluggable fashion

Alchemist Workflows

and Data Discovery

for

Biomedical Research

May 18-19, 2007

eLiberatica 2007
Brasov, Romania

- Alchemist infrastructure provides an innovative mechanism for biomedical image data search and spectral data discovery
  - metadata generated from user and automatic annotations is fused
  - semantics can be included as part of the query

- Triana - as a scientific workflow environment - has been used in Bioinformatics projects as the *Grid-based problem-solving environment for collaborative exploration and analysis of global biodiversity patterns*

Within the context of Biomedical sciences, the Alchemist has a significant potential to support:

(i) distributed biomedical communities focused on a specific disease process

(ii) disease-oriented collaborative studies which share large datasets

(iii) integrative biology projects that need to analyse inter-related information at different levels
(e.g. clinical, cellular, molecular and genomic)

(iv) population-based studies (e.g. clinical trials in diabetes or cancer)

For biomedical images, current practice usually involves searching databases containing:

(i) patient data repositories

(ii) case-oriented reference atlases (i.e. dynamic information across spatial and temporal scales of abstraction)

(iii) training collections (documented biomedical images, either anonymised individual or averaged data and training datasets)

Biomedical audio-visual content and associated metadata can be discovered and retrieved, allowing also the visualization of particular regions of interest within the images, or anomalies in the patterns of spectral data.

While content-based retrieval is an active field of research, multiple modality search (using data from multiple sources and media, and data from different levels of biological organisation) can give deeper insights into the nature of biological entities and the processes they are involved in.

By using the Alchemist for disease-oriented studies,

P2P caching could support the search, selection and aggregation of similar biomedical and spectral datasets.

=> Caching, using rendezvous peers (as in WSPeer), can be adapted to cache similar requests for resources

=> store hits on nodes which are within closer proximity to the particular group of researchers interested in the file

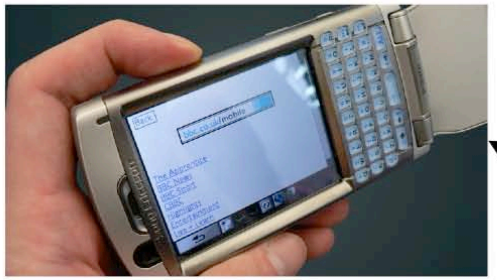
Mobile support is also provided for the Alchemist infrastructure

- Clinicians can benefit greatly from the possibility of using a mobile device to initiate search operations, retrieve medical audio/visual data and associated records, or operate data transfer between different static repositories, in a controlled and secure way.

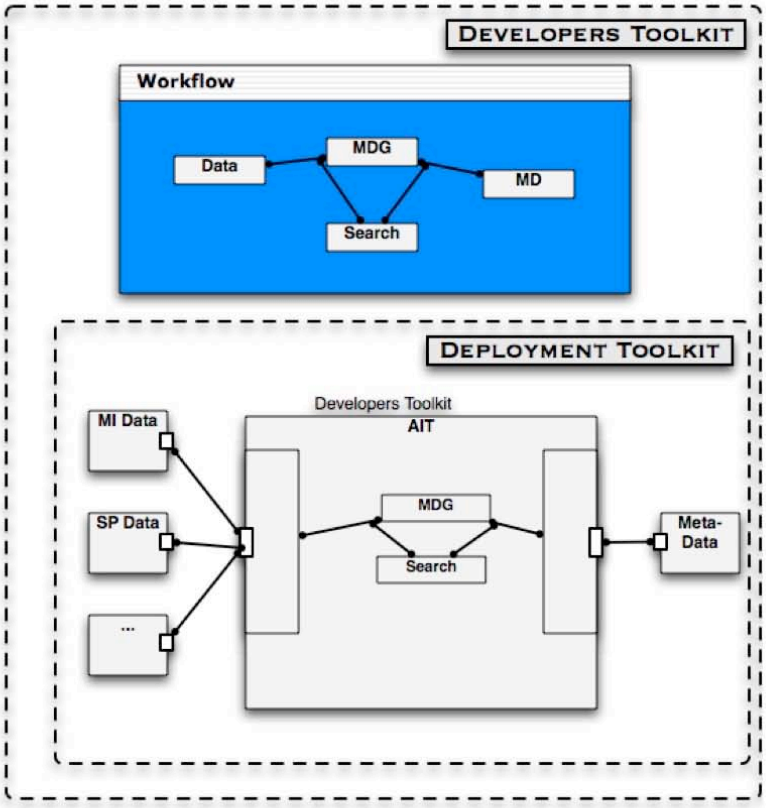
- We also address the possibility of using mobile devices for workflow management by the remote control of the enactment engine from a mobile user interface, for the data fusion of different kinds of data and metadata and the integration of algorithms designed under a common problem-solving environment

Scientific Workflow Management

For image and spectral data search and retrieval



Remote Workflow Control Via Web Services Interface



MI - Medical Image Data
SP - Spectral Data
... - any other

Alchemist Workflows

for Data Discovery

in

Diabetic Retinopathy

May 18-19, 2007

eLiberatica 2007
Brasov, Romania

Diabetic Retinopathy (DR):

=> All patients with diabetes are at risk of developing DR, and its progression to a sight-threatening stage is often not detected

=> Disease that accounts for c.80-90% of cases of blindness due to diabetes in the UK

SCOPE:

Early detection of pathologic mechanisms underlying diabetic retinopathy in research and clinical trial scenarios:

- Mechanisms for *imaging and spectral data discovery*
- Vertical and horizontal *biomedical data integration*

2 SCENARIOS:

- Early Detection and Prevention of Retinal Disease
- Investigational Drug Discovery

The study of DR (diabetic retinopathy):

Customised utilisation of Alchemist multimodal search / workflow system, spanning:

- fundamental Biomedical research
- routine clinical (screening) practices

Guided by:

- DRSSW (Diabetic Retinopathy Screening Service for Wales)
- DRU (Diabetes Research Unit, Llandough Hospital)

- High quality research data collected => factors associated with graded outcomes of DR
- Additional data currently being obtained from primary care settings
  - high-resolution retinal images
  - associated quantitative physiological, demographic and other variables

Alchemist system:

- pattern searches at multiple levels
- transformations of signal data

Thank you for attention !

For more information…

Adina.Riposan@contactnet.ro

Ian.J.Taylor@cs.cardiff.ac.uk