Lessons learnt
building OGSA-DAI
EGC 2005

Malcolm Atkinson
Director

www.nesc.ac.uk

15th February 2005
Contents

• Why invest in shared software
  • Facilitating Applications
  • Facilitating Production use
  • Improving code quality
• The Data Bonanza
• OGSA-DAI
• International Collaboration
  • Foundation for economic high-quality e-Infrastructure
• Summary & Take Home Message
Conflicting Views?

- Governments, EU Commission, ... 
  - *Shared* e-Infrastructure will transform 
    - Economy 
    - Society 
  - Stimulate creativity and innovation 
  - Improve our diagnoses, research, decisions, designs & businesses 

- Researchers, ... 
  - Want to pursue *their* particular goals 
  - Want *no change* if it doesn’t help them 
  - Want *new facilities* ASAP, if their research needs it 
  - Want *convenient, easy to understand* and use facilities 
  - Want *long-term commitment* to support 
  - Want reliability & performance 
  - Prefer to pay as little as possible
Conflicting views?

- **Resource providers**
  - Fund providers
  - Institutions hosting developers and operations
  - Specific missions
  - Must demonstrate *they* have delivered
    - Better than the other providers
  - Want best value for money
    - But in their current time scales
  - Politically unable to give long-term commitment
    - With some exceptions?

- **Technology and Service vendors**
  - Profit and business survival informs their decisions
  - Risk averse
  - Incremental approach – where is the business *this year*
  - Distinctive business models
    - Inform their view on standards
Eternal Triangle

Applications

all want reliability, dependability, security, performance, long-term stability

Developers

Operations

How do you balance innovation and safe engineering
Eternal Triangle

All want reliability, dependability, security, performance, long-term stability

Want familiar trustworthy code
No distractions
With just the additions crucial to their goals
⇒ Many e-Infrastructures

Developers

Applications

Operations
Eternal Triangle

Applications

Operations

Developers

Want
Familiar trusted tools & libraries
Few & stable deployment contexts
Cost of testing & maintenance dominates
Many customers / version of code
⇒ Few e-Infrastructures

all
want reliability, dependability, security, performance, long-term stability

Many e-Infrastructures
Mix & Match model
Eternal Triangle

Applications

 Developers

 Operations

 Many e-Infrastructures
 Mix & Match model

 Want
 Familiar trusted tools & systems
 Stability
 Cost of Systems
 Administration dominates
 Operator error dominates loss of production
 ⇒
 One e-Infrastructure

 Prefer just one e-Infrastructure
 Testing & maintenance limit innovation

 all want reliability, dependability, security, performance long-term stability
Eternal Triangle

Applications

Developers

Operations

prefer just one e-Infrastructure
Testing & maintenance limit innovation

One e-Infrastructure
Stable with good management tools

all want reliability, dependability, security, performance long-term stability

Compromise: One e-Infrastructure – Select services & libraries
Agreed & Simple APIs / mappings wrap all common functions

Many e-Infrastructures Mix & Match model
Data Bonanza
Generating & Storing Data gets Easier

- More, faster & cheaper digital devices
  - Higher resolution
  - Greater deployment
  - Faster duty cycles
  - Do they produce required metadata?
- Larger, faster & cheaper storage technology
  - Economic to store (multiple copies of) primary data
  - Economic to store derived data
  - Crucial to store sufficient good quality metadata
- Digital Communications - higher bandwidth & cheaper
  - Practical to access and copy remote data
  - Latency not decreasing - get what you need in a few trips
Curation and Publishing Data

- Invest in preserving data
  - High guarantees that an observation will not be lost
- Invest in organising data
  - Efficient access for popular queries
  - Registration and description
  - Provenance records
  - Metadata (how to interpret this data)
  - Annotation (related data & comments)
- Invest in publishing data
  - Obligation from funders - democratisation
  - Expensive and technically hard

Collaboration necessary

Creative scientific contribution

Recognition? Attribution?

Citation? Responsibility?
Multi-dimensional Growth

- The Number of Data Collections
  - Grows rapidly
- The Size of each Data Collection
  - Grows rapidly
- The Complexity of each Data Collection
  - Grows rapidly & autonomously
- The Interdependencies between Collections
  - Grow rapidly
- The User communities
  - Grow rapidly - dispersed, diverse & mingling
Data Integration is Everything

- Motivation
  - No business or research team is satisfied with one data resource
- Data Curation Expertise Human Centred
- Integration Human centred
- Domain-specialist driven
- Dynamic specification of combination function
- Iterative processes

Sources inevitably heterogeneous
- Revised request minutes later
- Revised request after months of thought
- Time-varying content, structure & policies
- Robust, stable steerable integration services
  - Higher-level services over multiple resources
  - Fundamental requirements for (re)negotiation

Federation or Virtualisation preceding integration

Scientific Insight needed here

or kit of integration tools to be interwoven with an application?
This is the motivation for & home of OGSA-DAI

- Identify the recurrent requirements
- Provide one infrastructure that meets them
- Wide use enables a robust, reliable and supported set of facilities
- Steadily increase power of facilities
- Steadily raise the level of abstraction
- Standardise & achieve multi-national investment

- Robust, stable steerable integration services
- Higher-level services over multiple resources
- Fundamental requirements for (re)negotiation

Federation or Virtualisation preceding integration or kit of integration tools to be interwoven with an application?
OGSA-DAI is one of the Grid Middleware Centre Projects

Collaboration between:
- EPCC
- IBM (+ Oracle in phase 1)
- National e-Science Centre
- Manchester University
- Newcastle University

Project funding:
- OGSA-DAI, 2002-03,
  - £3.3 million from the UK Core e-Science funding programme
- DAIT (DAI Two), 2003-06
  - £1.3 million from the UK e-Science Core Programme II

"OGSA-DAI" is a trade mark

Funded by UK’s Department of Trade & Industry + Engineering & Physical Sciences Research Council as part of the e-Science Core Programme

Thanks to Mario Antonioletti for these EPCC slides
Geographically Distributed Team

EPCC Team, Edinburgh

NeSC, Edinburgh

ESNW, Manchester

Neresc, Newcastle

IBM Development Team, Hursley
Communication vital

Web Site
Support Desk
Training

Access Grid meetings
Telephone conferences
Face-to-face meetings

Twiki
IRC
Bugzilla

Email
Mailing Lists
Deliver data back to the client.

Deliver data to a third party.

Deliver data another GDS.

And there’s a lot more that you can do …
Perform Document → Grid Data Service → Result Data Resource → Response Document

- Grid Data Service
Developers encouraged to roll their own – many do

Predefined Activities

- Developers encouraged to roll their own – many do
  - fileManipulation
  - FileAccess
  - directoryAccess
  - fileWriting
  - relationalResourceManager
  - sqlBulkLoadRowset
  - sqlUpdateStatement
  - sqlStoredProcedure
  - sqlQueryStatement
  - xmlCollectionManagement
  - xmlResourceManagement
  - xPathStatement
  - xUpdateStatement
  - xQueryStatement
  - xmlResourceManagement
  - xmlCollectionManagement
  - sqlUpdateStatement
  - sqlStoredProcedure
  - sqlQueryStatement

- Delivered from URL:
  - DeliverFromFile
  - DeliverFromGFTP
  - DeliverFromGDT

- Delivered to URL:
  - DeliverToFile
  - DeliverFromURL
  - DeliverToGFTP
  - DeliverToGDT

- Delivered to Stream:
  - DeliverToStream

- Delivered to FTP:
  - DeliverToGFTP
  - DeliverToGDT

- Delivered to Files:
  - DeliverToFile
  - DeliverFromFile
  - DeliverToStream

- File and Directory Access:
  - fileAccess
  - directoryAccess
  - fileWriting

- File System Management:
  - fileAccess
  - directoryAccess
  - fileWriting

- Relational Resource Management:
  - relationalResourceManager

- XML Resource Management:
  - xmlResourceManagement

- XSLT Transformation:
  - xslTransform

- SQL Query:
  - sqlQueryStatement
  - sqlUpdateStatement
  - sqlStoredProcedure
  - sqlBulkLoadRowset

- XPath and XUpdate Statements:
  - xPathStatement
  - xUpdateStatement

- XML Collections and Resources:
  - xmlCollectionManagement
  - xmlResourceManagement

- Input/Output Streams:
  - inputStream
  - outputStream

- Compression/Archiving:
  - gzipCompression
  - zipArchive
Standardisation is important

GGF
- Arch: OGSA, CMM
- Data: INFOD, OREP, GSM, TM BoF, ADF BoF, DAIS
- ISP: CGS
- SRM: GRAAP, Policy

Other Standards Bodies
- OASIS: WS-DM, WS-RF, WS-N
- DMTF: CIM
- ANSI: SQL
- W3C: XQuery
- JCP: JDBC
- IETF: SNMP
- JCP: Address
- ????: WS Policy

http://forge.gridforum.org/projects/dais-wg
Example Projects Using OGSA-DAI

- Bridges: http://www.brc.dcs.gla.ac.uk/projects/bridges/
- BioSimGrid: http://www.biosimgrid.org/
- BioGrid: http://www.biogrid.jp/
- eDiaMoND: http://www.ediamond.ox.ac.uk/
- GeneGrid: http://www.qub.ac.uk/escience/projects.php#genegrid
- myGrid: http://www.mygrid.org.uk/
- ODD-Genes: http://www.epcc.ed.ac.uk/oddgenes/
- OGSA-DAI: http://www.ogsadai.org.uk
- N2Grid: http://www.cs.univie.ac.at/institute/index.html?project-80=80
- AstroGrid: http://www.astrogrid.org/
- GEON: http://www.geongrid.org/
- OGSA-WebDB: http://www.gtrc.aist.go.jp/dbgrid/
- FirstDig: http://www.epcc.ed.ac.uk/~firstdig/
- INWA: http://www.epcc.ed.ac.uk/
OGSA-DAI User Project classification

- Biological Sciences
  - Bridges
  - BioGrid
  - myGrid
  - GeneGrid

- Physical Sciences
  - AstroGrid

- Commercial Applications
  - ODD-Genes
  - BioSimGrid
  - eDiamond

- Computer Sciences
  - GEON
  - N2Grid
  - MCS
  - OGSA Web-DB
  - GridMiner
  - IU RGBench
OGSA-DAI Downloads

690 downloads since May 04
Actual user downloads not search engine crawlers
Does not include downloads as part of GT3.2 releases
Data from 13 December 04

Total of 966 registered users

<table>
<thead>
<tr>
<th>Release</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1.0 (Jan 03)</td>
<td>107</td>
</tr>
<tr>
<td>R1.5 (Feb 03)</td>
<td>110</td>
</tr>
<tr>
<td>R2.0 (Apr 03)</td>
<td>254</td>
</tr>
<tr>
<td>R2.5 (Jun 03)</td>
<td>294</td>
</tr>
<tr>
<td>R3.0 (Jul 03)</td>
<td>792</td>
</tr>
<tr>
<td>R3.1 (Feb 04)</td>
<td>655</td>
</tr>
<tr>
<td>R4.0 (May 04)</td>
<td>939</td>
</tr>
<tr>
<td>R5.0 (Dec 04)</td>
<td>138</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3323</strong></td>
</tr>
</tbody>
</table>
OGSA-DAI provides middleware tools to grid-enable existing databases
The OGSA-DAI Project Site:
  – http://www.ogsadai.org.uk

The DAIS-WG site:

OGSA-DAI Users Mailing list
  – users@ogsadai.org.uk
  – General discussion on grid DAI matters

Formal support for OGSA-DAI releases
  – http://www.ogsadai.org.uk/support
  – support@ogsadai.org.uk

OGSA-DAI training courses
OGSA-DAI next steps
Platforms & Users

- Currently on OGSI (GT3)
  - Discontinue support when ~ R6
- Currently on WS-I+ (OMII1)
  - Will be in next release
  - Without asynchronous & Third-party data transfers
- Currently in Preview on WSRF (GT4)
  - Not yet a supported release ~GT4 release
- Users about equally divided
  - Some still use R3!
- Re-designed architecture
- Long list of requested features
- Many projects want long-term support commitments
DS (Mobius)

Registry

Registry2

Logging Service

Request TADD

Response TADD

DRAM

DSDL

initiateDataService()

0

Initiates/Manages

n

DS (DAIS)

DS (OGSA-DAI)

Single Service Session

Id - UUID

performRequest()

DS (Mobius)

DID

Type

Format

Txn

Local Store

Compute & storage resources

DR

DRs

Id – UUID

n
OGSA-DAI & Triangle

Applications

all want reliability, dependability, security, performance long-term stability

Developers

Mostly use client library Some use protocols No extra tools yet

Operations

No tools or interfaces yet Motivation for new architecture

One client library Increasingly important More abstraction needed

Compromise: One e-Infrastructure – Select services & libraries

Develop: Higher-level Client Library & Tools, more Integration, Operations support
OGSA-DAI team needs

- Agreed data naming system
  - OGSA effort - 3-level: human, abstract & physical address
- Addresses of state & data resources
  - WS-Addressing
- Life Time Management
  - WS-RF Resource LifeTime - “imported into OGSA-DAI”
- Properties
  - WS-Resource Properties - easily implemented look alike
- Error reporting
  - WS-BaseFaults
- Agreed Data Transport Abstraction
  - OGSA-Data Design + InfoD meeting at GGF13 Seoul

Most of all we need these standard with APIs only one of each!
International Collaboration
The Ultimate Challenge

- **Testing**
  - Large-scale, always on, distributed persistent infrastructure
  - Product space of platforms and external components
  - \{Oracle, DB2, MySQL, Postgres, ...\} \cup \{Xindice, eXist, ...\} \cup \{files, DFDL, semi-structured, indexed, text-mined, ...\}
  - \times \{java, J2EE, .NET, ...\} \times \{OGSI, WS-I+, WSRF, ...\}
  - Growing proportion of team effort - though mechanised

- **Maintenance**
  - Fixing bugs (<20%), Dealing with context changes (>20%)
  - Providing new required functions (~60%)
  - Better coding and testing can at best save 20%
  - Maintenance is a *life sentence*
  - No remission for good behaviour!

Grows to dominate costs and limit development
To Meet the Challenge 1

- Agree an Architecture: OGSA + NextGRID + ...
  - To partition the problem space
  - To raise the level of abstraction & discourse
  - To provide a framework for collaboration
    - Environments in which alternative solutions can perform roles
  - Incremental progress to agreement
    - Profiles

- Invest in APIs
  - Protect Application Developer investment
  - Protect Middleware subsystem investment
  - Clarify requirements
  - Specify semantics
To Meet the Challenge 2

- **Raise the Level of Abstraction**
  - Greater benefit for Application Developers
  - Greater benefit for other Middleware Developers
  - Easier comprehension for designers, implementers & exploiters
  - Opportunity for implementation improvement increases

- **Form ≤ 2 International Alliances / Consortia**
  - Agree on target e-Infrastructure function and properties
  - Safety of Open Source - future maintenance always possible
    - But only affordable through alliances
  - Agree partitioned R&D task:
    - Country X delivers A and Country Y delivers B
  - Incremental development of relationship
    - Compete → partnership → trust → mutual dependence
  - Avoid brittle dependency
    - Minimum functionality base platform in which subsystems can work
    - OGSA base profile a good candidate
To Meet the Challenge 3

- Desist from Starting from Scratch
  - “pencil sharpening” auto-distraction
  - Sort-term illusion of progress and success
  - Long-term - another body of software to maintain
  - Division of effort
  - Your legacy: Transition and translation problems
  - Your legacy: Another body of software to maintain
- Darwinian survival of the fittest
  - Doesn’t result in “best”
  - Expensive & slow
  - Some diversity and competition valuable
  - But don’t let it split users, developers, operations, training, …

Discard ego trips, “nationalism” & excess competition – wasteful and harmful
Concluding Remarks
Observations 1

- **E-Infrastructure**
  - Disruptive technology
    - Will change what we do and how we do it
  - Opportunity to reap major benefits
    - Is Europe prepared? Education, Education, Education
    - Will we focus effort? Critical mass. Don’t divide & conquer *ourselves*
  - *Education essential*

- **Must Collaborate Internationally**
  - To agree, build and operate e-Infrastructure
  - To give adequate support to our users
  - To afford maintenance and operations
  - To facilitate international research, business & decisions
Observations 2 - OGSA-DAI

- >30 staff years of effort, Release 5, coming soon R6
- 3 platforms, >1000 users, world-wide use, diversity
- Backed by standards effort - hard work!
- User community & User group
- Major investment in client-side API
  - Beneficial for users, application developers & training
  - Provides implementation options
  - Undergoing re-design
- Flexible and Extensible framework
  - Essential for applications
  - Contributors build using this - webDB, streaming data, ...
  - No contributor code shipped with releases yet 😞
- Diverse demand for new features
  - Diverse & multi-platform
  - Looking for reassurance about future support and maintenance
- Perhaps 25% of original OGSA-DAI vision built so far
Comments & Questions
Please
End of slide show

Reserve slides for questions
From *OGSA Status and Future*, Hiro Kishimoto and Ian Foster, GGF12
slide originally from Michael Behrens, DISA consultant
Use Cases & Applications

GRID COMPUTING
Distributed query processing
Collaboration
Persistent Archive
ASP
Data Centre
Multi Media

DISTRIBUTED COMPUTING

UTILITY COMPUTING

Core Services

VO
OGSA Self
OGSA-EMS
WS-DAI
Information
WSDM
Discover
GGF-UR
WS-
Namina
Privacy
Trust
GFD-C.16
WSRF-RP
WSRF-RL
Data Model
WSRF-RAP
WS-Security
SAML/XACML
X.509

Base Profile

WS-Addressing
HTTP(S)/SOAP
WSDL
CIM/JSIM
Data Transport

Provided by David Snelling (Fujitsu) and Mark Linesch (GGF & HP).
Why Invest in OGSA Architecture

- Integration
- Completeness
- Abstraction
- Cooperation

- OGSA partitions the e-Infrastructure implementation
- Encourages independent concurrent & coordinated
  - Development or evaluation of each part's standards
  - R&D on implementation of each part
- Promises assembly of the parts
- Basic profile provides context for concurrent R&D
  - Context for each M/W developer to build against
  - Reduced interdependence - each can deliver if others don’t

Focus effort on reaching minimum threshold that makes this work
Back OGSA more

- Invest effort in OGSA
  - Investigating, evaluating, contributing, commenting
  - Implementing profiles
  - Using it
- It is the ONLY show in town
  - Which offers
    - integration, completeness, abstraction
    - A foundation for collaboration
- UK focus on Data Design Team
- UK efforts in other design teams
  - EMS, Grid markets, JSDL, GSM, mySpace, ...
Use OGSA for Collaboration

- Big push to Reach OGSA Basic Profile
  - Sufficient platform, context & framework
  - For safely partitioning further R&D
- Agree a division of work
  - Upgrading / alternative trade-off components
  - New components
  - Higher level facilities
- Minimise duplication
- Maximise combined efforts to deliver
  - Function, Stability, Quality & Abstraction

Bury the egos, project competition & national pride silos