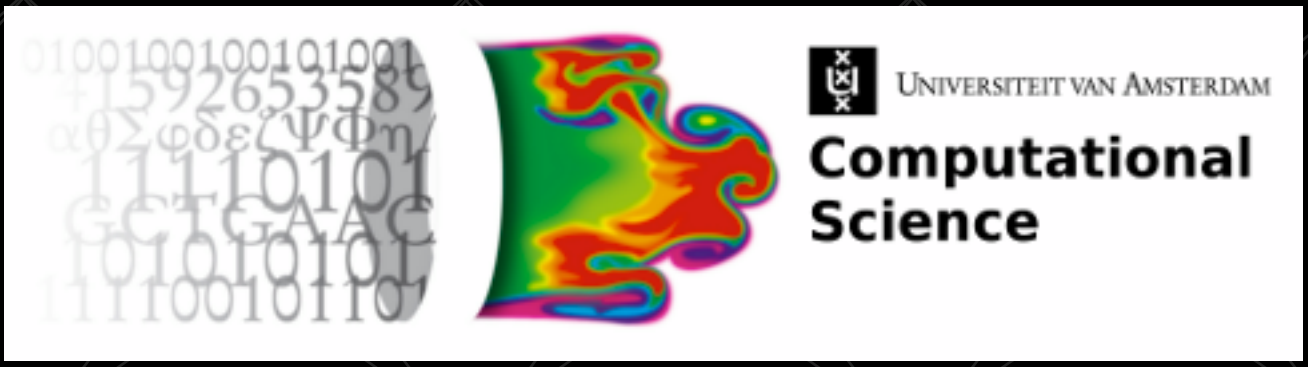


Why a Curriculum in eScience

Adam Belloum
Computational Science group
Institute of Informatics
University of Amsterdam
a.s.z.belloum@uva.nl

February 9 2012

WS-VLAM

Outline

- Introduction & background
  - From Science to e-Science
  - Grids / e-infrastructure / Clouds
- Motivations
  - What did we learn from VL-e (2004-2009)
- e-ScienceLand proposal

What is e-Science?

- "e-Science is about **global collaboration** in key areas of science, and the **next generation of infrastructure** that will enable it."

  John Taylor Director General of Research Councils
  Office of Science and Technology

- **global collaboration**
  → Virtual Organization
- **next generation of infrastructure**
  → e-Infrastructure

What is e-Infrastructure ?

- "**Grid** is a software infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and **resources**"

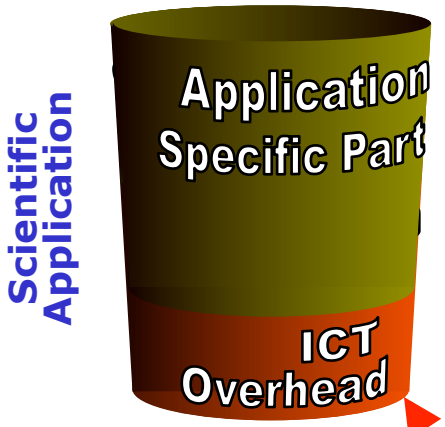
  Foster, Kesselman and Tuecke

  →**Resources:** not only computers but also data storage resources and specialized facilities

- **"Long term goal** is to develop the middleware **services** that allow scientists to routinely build the infrastructure for their Virtual Organisations"

  Tony Hey Director of UK e-Science Core Programme

From Science to e-Science

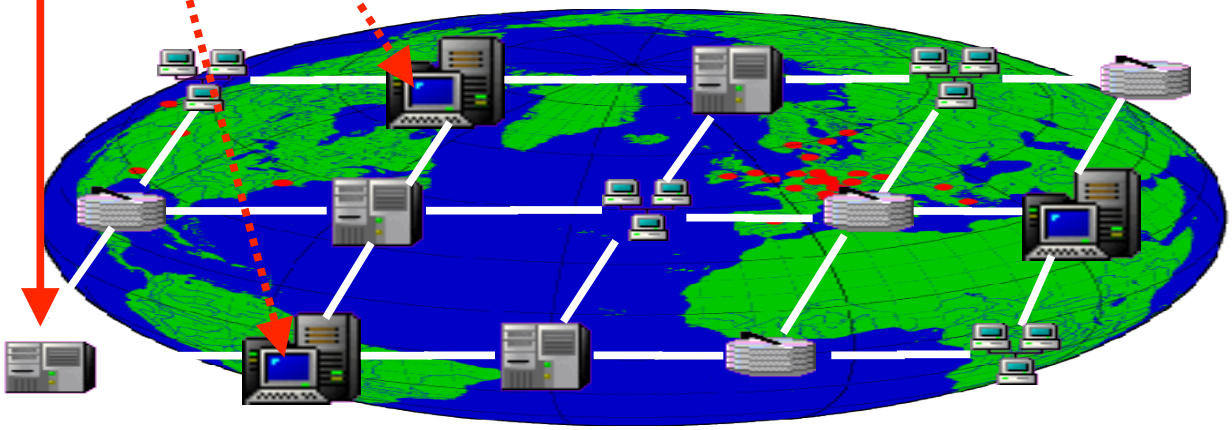


- **e-Science** *is building on achievements of* Grid and other supporting ICT…

➡ ***To stimulate scientific applications to adopt e-Science an infrastructure is needed!***

- **e-Science** frameworks are providing basic services
- which will hide computing resources and boost the
- development of data and computationally intensive
- e-Science applications on a large scale distributed
- infrastructure…

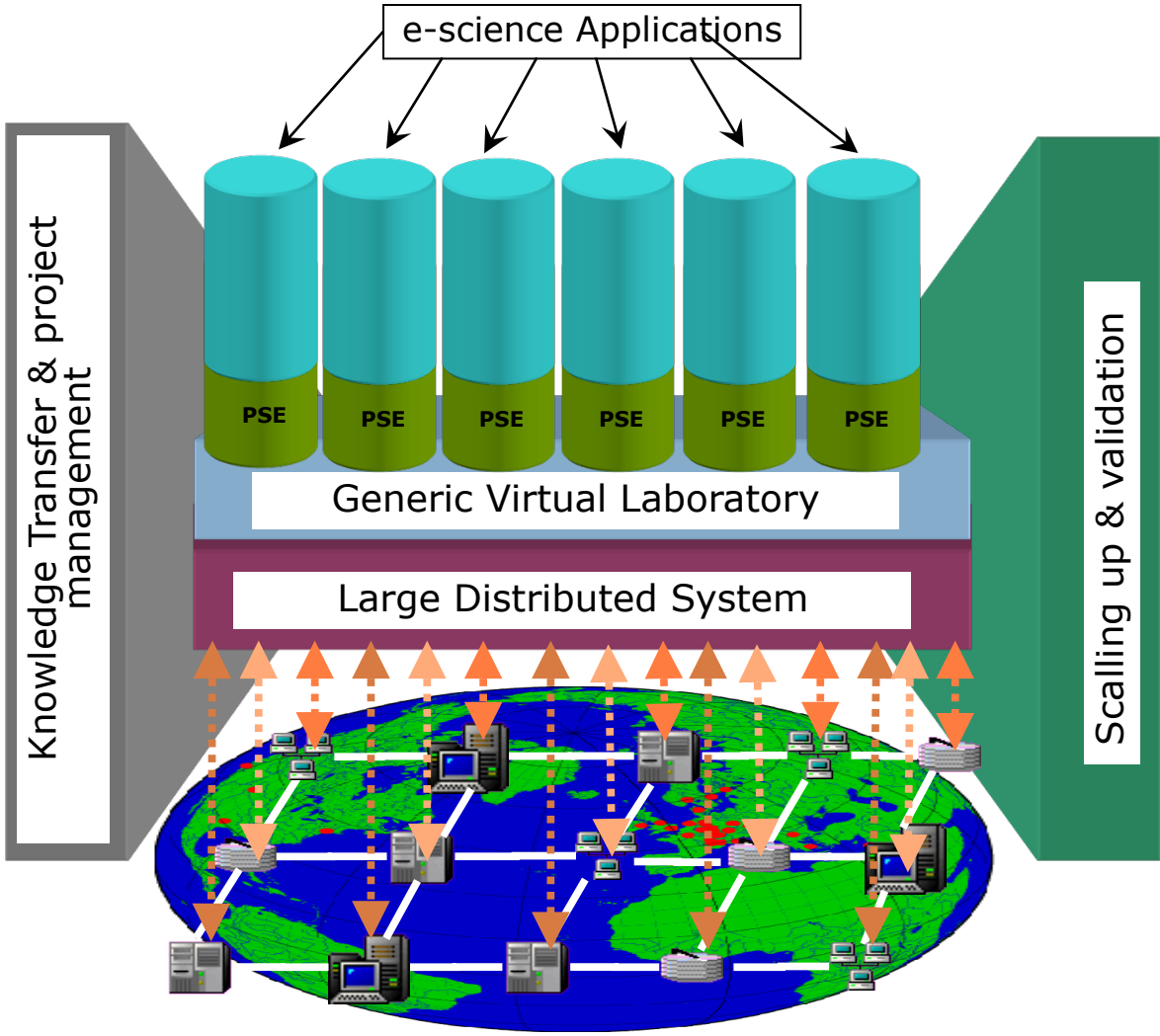
➡ ***Application scientist can increase his productivity, while focusing on his own science rather than on ICT problems***



Mission and strategy

To achieve its mission the VL-e project has set itself the following aims:

- create scientific prototypes of application-specific e-Science environments,
- develop a methodology for re-usable ICT components,
- **scaling up** & validating in 'real-life applications',
- build up and **transfer knowledge** on effectively realising application-specific e-Science environments.



BigGrid Infrastructure

BiG Grid
the dutch e-science grid

NIKHEF pdp

Nikhef (NDPF)

2500	processor cores
2000	TByte disk
160	Gbps network

SARA (GINA+LISA)

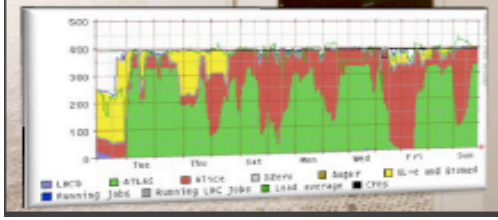
4800	processor cores
1800	TByte disk
2000	TByte tape
160	Gbps network

RUG-CIT (Grid)

120	processor cores
8 800	GByte disk
10	Gbps network

Philips Research Ehv

1600	processor cores
100	TByte disk
1	Gbps network



A quick summary

- Possibility to re-use across scientific domains
- Possibility to use a large computing and storage power

- Highly complex distributed infrastructure
- Multi-disciplinary working environments:
  - Scientists, Scientific programmers, application developers, middleware developers, system admins.

Quick summary of concepts exposed to e-Scientists

- Distributed computing
  - Clusters, Grids, and Clouds
- Data management
  - Find the data location, stage in/out the data
  - Follow the data lifecycle (provenance)

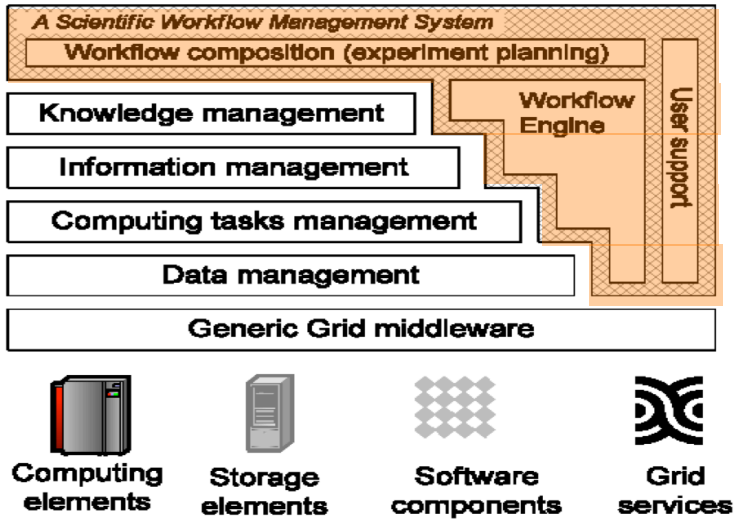
  →we are talking about GBs →TBs

- Security
  - Authentication / Authorization / Accounting

System-Level Science

"An approach to scientific investigations which, besides of analyses of individual phenomena, **integrates** different, **interdisciplinary sources** of knowledge about a **complex system**, to acquire understanding of the **system as a whole**."

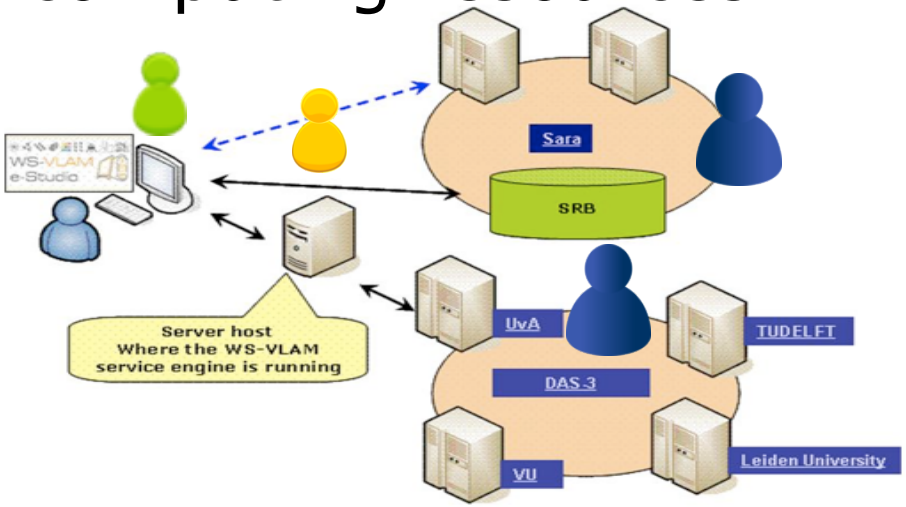
Foster, I., Kesselman, C., *Scaling system-level science: Scientific exploration and its implications*. IEEE Computer 39 (11) 2006

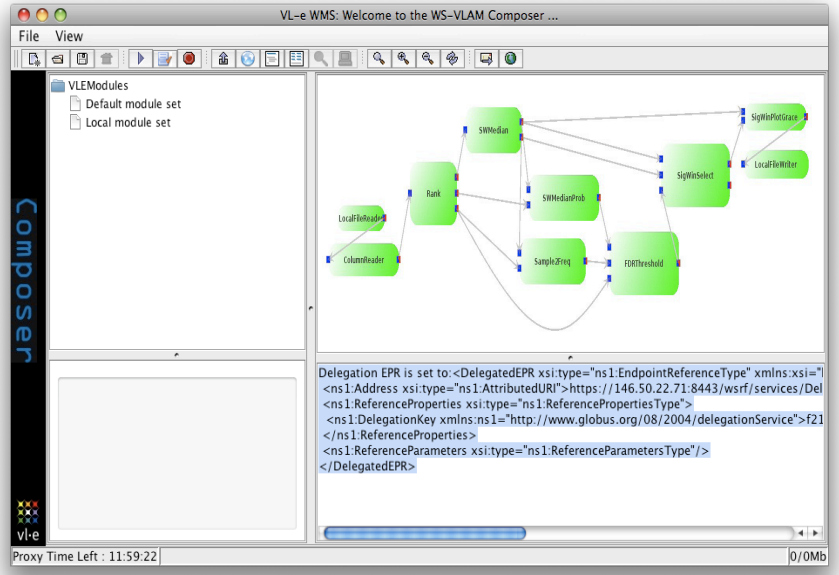


"Improving Interoperability, Sustainability and Platform Convergence in Scientific and Scholarly Workflow" Report of the NSF/Mellon Workshop on Scientific and Scholarly Workflow. Oct 3-5, 2007, Baltimore, MD.

Workflow management system

- **Workflow management system** is a computer program that manages the execution of a workflow on a set of computing resources.





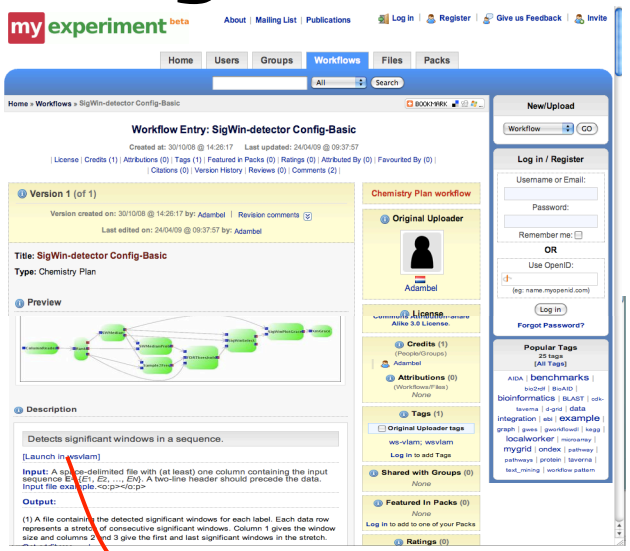
The user interface of the WS-VLAM a workflow management system developed in the VL-e project to execute application workflow on geographically distributed computing resources

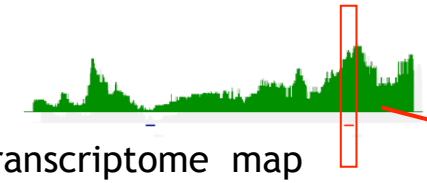
Deployed as service on DAS3, and BigGrid Clusters

Sharing/Reuse of things

- Workflow can be made available to entire community

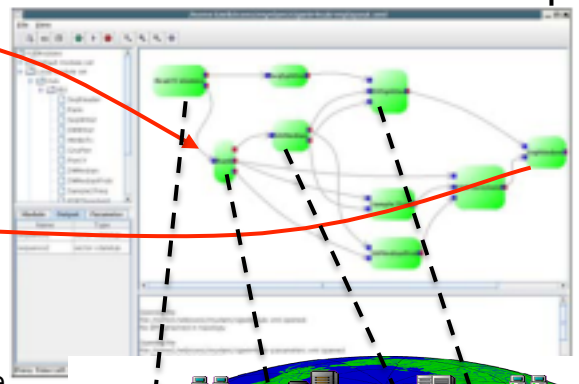
  (using Web 2.0 approach)
- Repositories, Databanks,

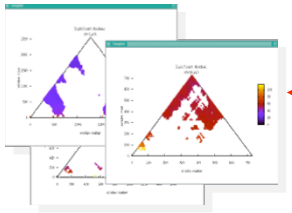




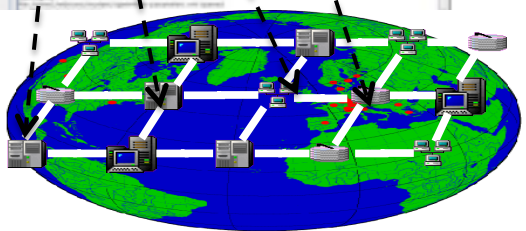
Human transcriptome map

WS-VLAM composer



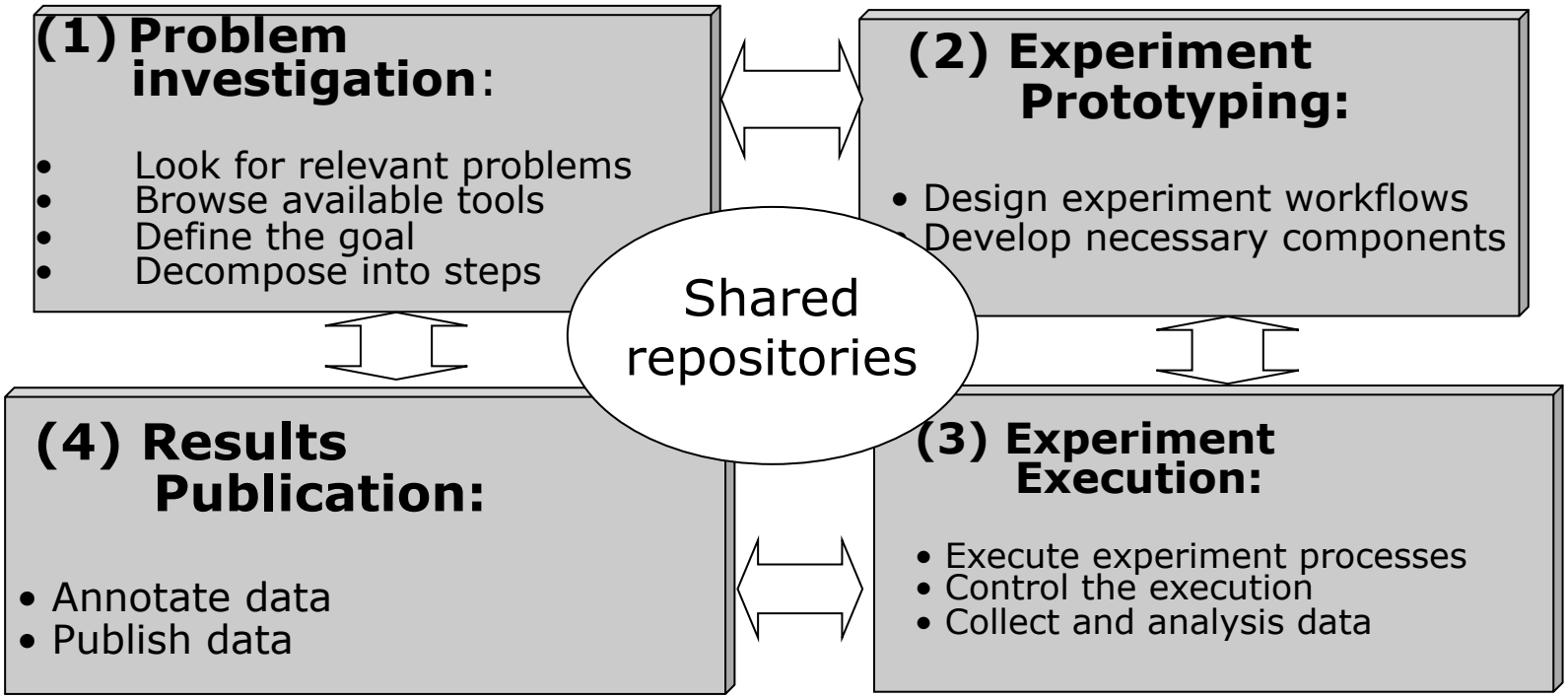


DNA curvature of the *Escherichia Coli* chromosome



Interactive Dataflow Driven workflow Engine
V. Korkhov, A. Wibisono, D. Vasyunin, A.S.Z. Belloum et al.
Scientific Programming, vol.15(3), pp. 173-188 (2007), ISSN 1058-9244

Complex Scientific experiments model

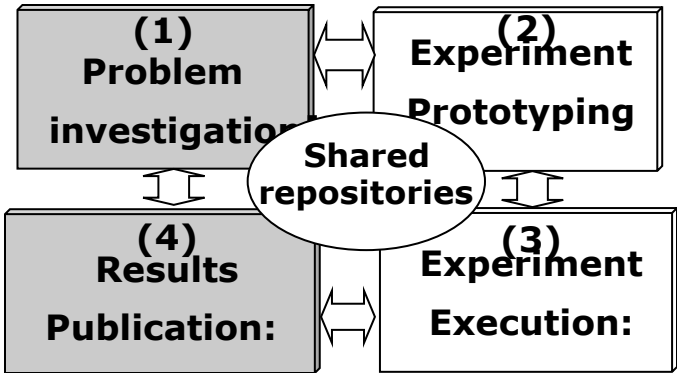


Collaborative e-Science experiments: from scientific workflow to knowledge sharing *A.S.Z. Belloum, et al., JULY/AUGUST, IEEE Internet Computing, 2011*

Quick summary of concepts exposed to e-Scientists

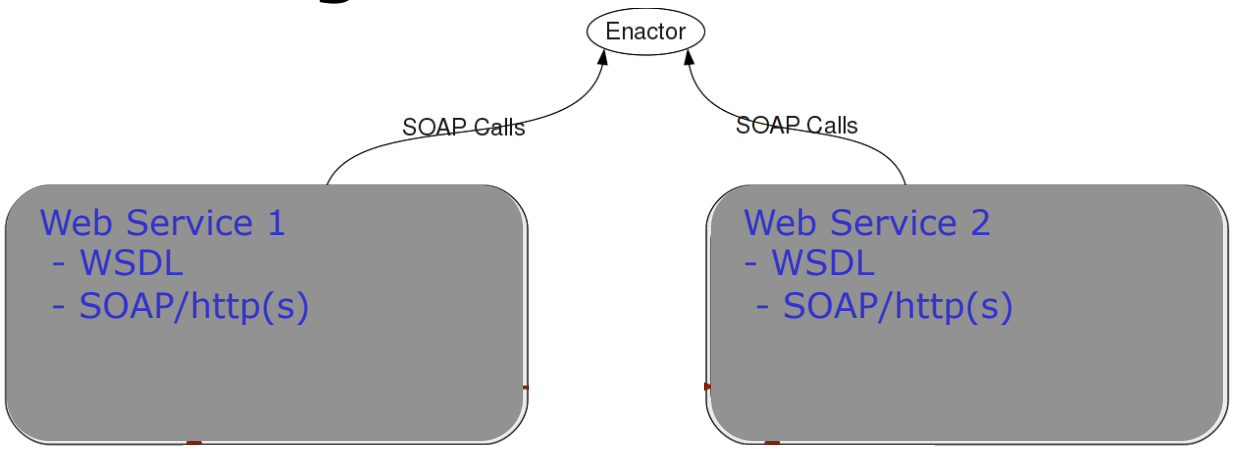
- Workflows
- Semantics
- Data management
- Social / Organizational

- And a **bit** of
  - Distributed computing
  - Security



Usage of Web Services in e-science

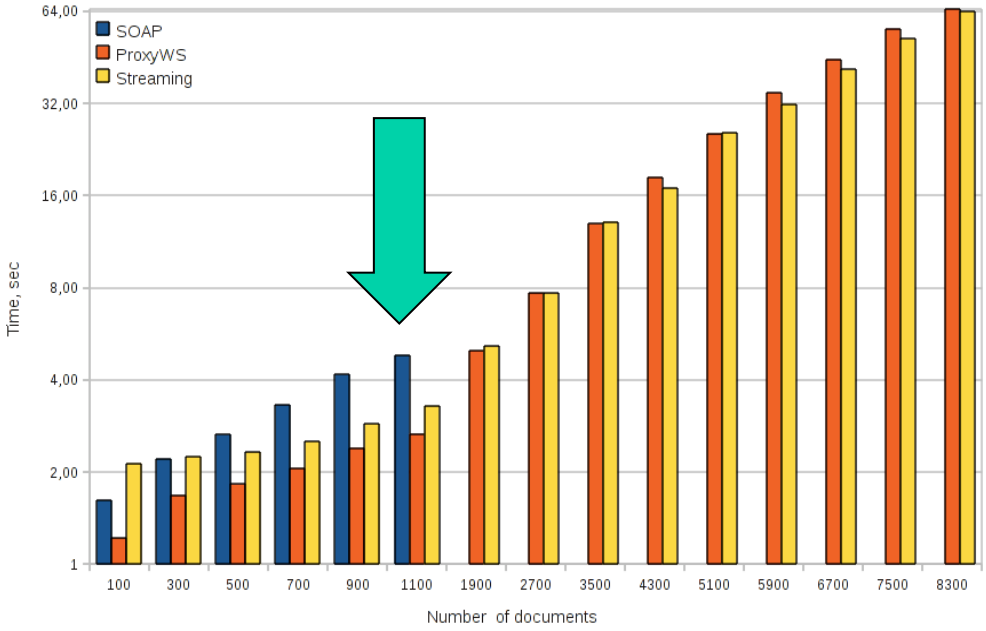
- In service orchestration, **all data is passed to the workflow engine** before delivered to a consuming WS
- Data transfers are made through **SOAP**, which **is unfit for large data transfers**



Enabling web services to consume and produce large distributed *datasets Spiros Koulouzis, et al. JAN/FEB, IEEE Internet Computing, 2012*

Indexing **Name Entry Recognition**

- SOAP failed to scale for more than 1100 documents.
- Proxying and streaming managed to cope with the data requirements of 8300 documents



Execution times for Search and NER (phase ii and iii)

List of applications ported to an e-Infrastructure

- sigWin detector *[Micro-Array Dept-UvA]*
- *Affymetrix Permutation* *[Micro-Array Dept-UvA]*
- *Omnimatch* *[UU/Leiden]*
- wave propagation *[TUE]*
- Blast *[AMC]*
- gut microbiota *[CWI]*
- Smart Infrastructure *[SNE-UvA]*
- Dynamic network control *[SNE-UvA]*
- GridSFEA, *[TU Munchen]*

More applications www.science.uva.nl/~gvlam/wsvlam/Applications

What did we learn from these use case

- *Getting the scientists to get a simple application (Hello-world) working on e-Infrastructure is **not straightforward***

  → (Mathematician, Biomedical-medical engineering, system Admin, software engineer)

  *→ **Step by Step** Introduction was needed in all cases*

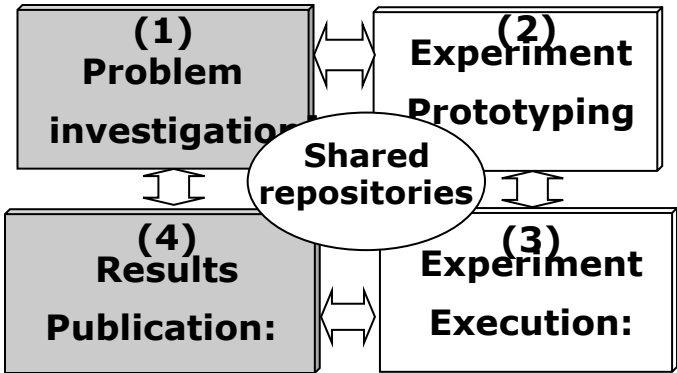
  *→ **Can be avoided** if the Scientists have had an **introductory** course(s) to e-infrastructure*

What did we learn from these use case

- *Moving a working application from the Application developer desktop to the e-Infrastructure is **not straightforward***

  → *Sometimes part of the application need to be rewrite it (**collaboration** between middleware and application developer is needed)*

  → *We can **speed up this phase** if the application developers have a **global idea** about how the infrastructure works*



UvA experience "MSc Grid Computing"

- *Is it a good experience Yes and No*

→ ***Yes**: we have introduced courses which prepare students to face problems related to e-Infrastructure, it works well. Students who followed the course managed to find their way easily either in companies or PhD students when it come to the using e-Infrastructure*

→ ***No**: Naming the entire program as Grid computing was too much, because the content is more generic (the name does not reflect the content), students were reluctant because of the Job market*

European Commission Plans INFRA-2012-3.3: Coordinating actions, …

- Objective 1: Develop strategies that will remove **obstacles** facing researchers … to enter the e-Science landscape.

- Objective 2: Build **trust** & **mutual understanding** between Research Infrastructures, e-Infrastructures and related authorities …

- Objective 3: Deliver an **e-Science curriculum for higher education** that defines the teaching of **data-based science**, **grid** and **High Performance Computing usage** as well as software application and tool development.

New curriculum …

- Curriculum should **identify** current barriers (**technical**, **social**, **organisational**) with respect to a greater adoption of e-infrastructure.

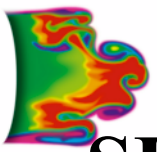
- Course components to be developed can be **incorporated** into current high educational programs

New curriculum …

- **Algorithms**, **modelling** and **simulation** software developed to solve research problems in the biological, physical and social sciences, in engineering, and in the humanities

- **Computer** and **information science** that develops and optimizes the advanced system hardware, software, networking, and data management components needed to solve computationally demanding problems

- **Computing infrastructure** that supports both the science and engineering problem-solving and the developmental computer and information science.

gSLM Service Delivery & Service Level Management in e-Infrastructures

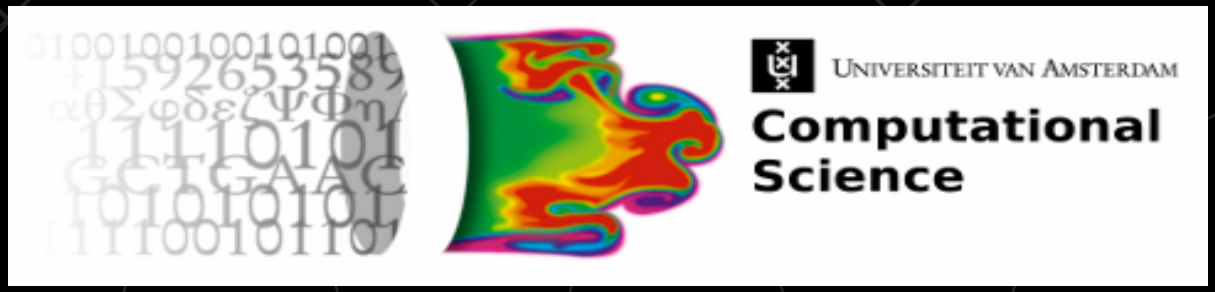
INFRA-2010-3.3: Coordination actions

- The gSLM project aims to help solve the **e-Infrastructure delivery problem**
- Bring together grid and ITSM experts and formulate new approaches to **Service Level Management** in grid

- Experts from both communities within the project
  - Grid operators, developers, policy and communication experts
  - Academics and accredited experts in ITSM approaches to service level management

References

1. J. Serrat, T. Szepieniec, **A.S.Z. Belloum**, J. Rubio-Loyola, O. Appleton, T. Schaaf, J. Kocot, gSLM: The Initial Steps for the Specification of a Service Management Standard for Federated e-Infrastructures, 8th IFIP International Conference on Research and Practical Issues of Enterprise Information Systems CONFENIS 2014, Hanoi, Vietnam
2. Szepieniec, J. Kocot, T. Schaaf, O. Appleton, M. Heikkurinen, **A.S.Z. Belloum**, J.S. -Fernandez, M. Metzker*, On Importance of Service Level Management in Grids*, In Proceedings of EuroPar 2011: Parallel Processing Workshops, Lecture Notes in Computer Science, vol. 7156, 2012, pp. 64-75, doi: 10.1007/978-3-642-29740-3_9.
3. R. Cushing, S. Koulouzis, **A.S.Z. Belloum**, M.T. Bubak, *Service Level Management for Executable Papers* Euro-Par 2011: Parallel Processing Workshops, Lecture Notes in Computer Science vol. 7156, 2012, pp. 116-123, doi: 10.1007/978-3-642-29740-3_14.



http://uva.computationalscience.nl/

http://www.vph-share.eu/

http://www.commit-nl.nl/

http://www.science.nl/~gvlam/wsvlam/

http://www.gslm.eu/