Science, Technology and Innovation

Projects

€20

A Smarter Internet

Making sense of linking data

GRID in Europe

Innovation to change the world

Lighting the way

New improved network for large-scale R&D

Leading Dissemination for EU Research & Development

Per Öster Ludek Matyska Dai Davies Wendy Hall Keith Jeffery

British Publishers media with influence

Data deluge dammed? Detection, discussion, and deliverance

Professor Keith Jeffery, ***President of ERCIM, explains that despite the many difficult issues that have arisen with the massive deluge of data accompanying Grids, Clouds and the exponential rise in Internet usage, there are solutions out there that will deliver us from these demands***

The Data Deluge Concept originated in the late 1990s when it was realised that deployment of information technology – especially detectors, computers, storage, networks – was becoming widespread due to ever increasing performance benefits for unit cost. The classical scientific/research method of research review followed by observation / hypothesis / experiment / simulate / compare and then publish was being enhanced by availability of data. However, this technology availability also permitted researchers to control experiments remotely, intercommunicate in cooperative research, interoperate their software and data with other researchers, and develop cooperatively publications and research proposals. From this realisation the e-Science (or e-Research) concept was initiated in the UK. Although based on the Grid concept from the USA essentially limited to metacomputing (shared connected supercomputers), the UK idea was to virtualise (hide from the researcher) not only the computation, data storage, networking and detectors but also the heterogeneity of information and knowledge.

The basic idea behind e-Science was that the end-user did not know or care where computing was done, or from where the information came as long as the quality of service or service level agreement was appropriate. Clearly this concept applies to industry, commerce and government as well as research. To achieve such virtualisation there were (and are) several challenges to be overcome. These are: Insufficient representativity, Insufficient expressivity, Insufficient resilience and Insufficient dynamic flexibility. The European Commission set up the Next Generation Grids Expert Group which, in a series of three reports, defined progressively a roadmap for successful development and utilisation of Grids technology. Large European projects led to the well-used EGEE (Enabling Grids for e-Science) software (GLite) and the specification for EGI (European Grid Initiative).

The Data Deluge grew. The massively increased amount of research data was succeeded quickly by massive amounts of data from industrial and commercial production – from control systems, finance, healthcare and telecommunications, for example. As end-users took up social networking utilising Web2.0 technologies, there was a vast increase in data being transmitted and stored, used and re-used (like videos) especially with YouTube, Facebook and smart phones becoming more and more popular.

The Future Internet and Internet of Things concepts envision billions of connected nodes each capable of data processing (including storage and transmission). Mobile technology permits IT-supported actions 'on the move and in the open'. The complexity of the networks demanded new solutions

to avoid bottlenecks: P2P (Peer to Peer) architectures were developed. The complexity of IT systems supporting enterprises increased requiring expensive system administration and management functions just to keep operational. This formed another bottleneck. The Next Generation Grids Expert Group of the EC had already anticipated these problems in 2005 (the report was published January 2006) and suggested a solution – SOKUs.

SOKUs (Service-Oriented Knowledge Utilities) – essentially intelligent self-describing, self-composing, self-managing services – should relieve many of the systems administration and management tasks and permit the end-user to initiate their required business process which in turn automatically discovers and utilises the required resources of computation, storage, communication, detectors, information and knowledge.

Clouds emerged. Some large companies with excessive hardware resources, whose full capacity was rarely used, provided a virtualised service (based on virtual machines and multi-tenancy of their hardware resources) to customers on a 'pay as you go' basis. This had great advantages in allowing customers to have access to massive resources without capital expenditure and with fine-grained cost control. The concepts of IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service) emerged. Currently Clouds are used for occasional excessive IT use, for business continuity and disaster recovery and for testing out new IT solutions. However full ASP (Application Service Provider) offerings are becoming available providing the opportunity for an organisation to outsource totally its IT.

The barriers to this are that

- Clouds offerings are proprietary and one attempt at standardisation failed so customers are wary of vendor lock-in;
- Business models and the economic benefits of Cloud approaches are not well-understood;
- There are issues of legislation and policy concerning the management of data in other countries.

A solution may lie in Grids of Clouds, since Clouds inherit many properties of Grids (resource sharing and dynamic management, virtualisation of resources) using SOKUs.

> "The basic idea behind e-Science was that the end-user did not know or care where computing was done, or from where the information came as long as the quality of service was appropriate"

Issues remain. The key ones are common to all proposed solutions to the data deluge. These are:

Representation of real world

- Hierarchies are insufficient yet most IT is based on them; we need fully connected cyclic graphs;
- Rich semantics are required both to describe the real world and to permit services to behave autonomously;
- Domain ontologies are required to provide supportive metadata for interoperability and for 'intelligent' services;
- Well-formed richly-structured syntax is needed to permit tractable programming;
- Business processes (because they evolve) need general code and execute-time binding of data and should be configured dynamically by representative metadata.

Management of global state

- Local state and 'active interfaces' between local stateful systems;
- Transactions need to be complex, multi-level, local but with open-ended properties using messaging interfaces.

Completeness

- Dealing with incomplete information for decision-making implies probability, fuzziness, learning systems.

Certainty

- Dealing with uncertain information also requires probability, fuzziness, learning systems.

Optimisation

- New paradigms are required when no global state exists so local states are optimised and then optimised more widely with negotiation;
- This implies partitioning, approximation, and much use of metadata.

Trust has aspects of

- Representation of business organisations and their policies;
- Contracts and proposals;
- Service level agreements;
- Links to both security and privacy.

Security

- System availability and continuity under attack;
- Prevention of unauthorised system access;
- Authentication and authorisation – global or connected local systems;

Privacy: with aspects of

- Openness of personal data to data subject (right to correctness);
- Security of personal data to others (right to privacy).

Unacceptable Use

- With different cultural modes internationally how are ethics in the future Internet defined? How are ethical standards maintained: for example what does a spam transaction look like? How do we prevent the hijacking of systems by 'adult, political, racist... transactions'?

Deliverance from the problems – via solutions to those issues – are expected to be developed over the next years. Without fast-paced research we cannot keep up with the exponential demands caused by the data deluge. Within the EU, through national programmes of research – and the EC Framework programme – there are relevant active projects. ERCIM member organisations, together with industrial and academic partners, are at the forefront of tackling these issues. ★