COMMIT/

Smart Cyber Infrastructure for Big Data processing

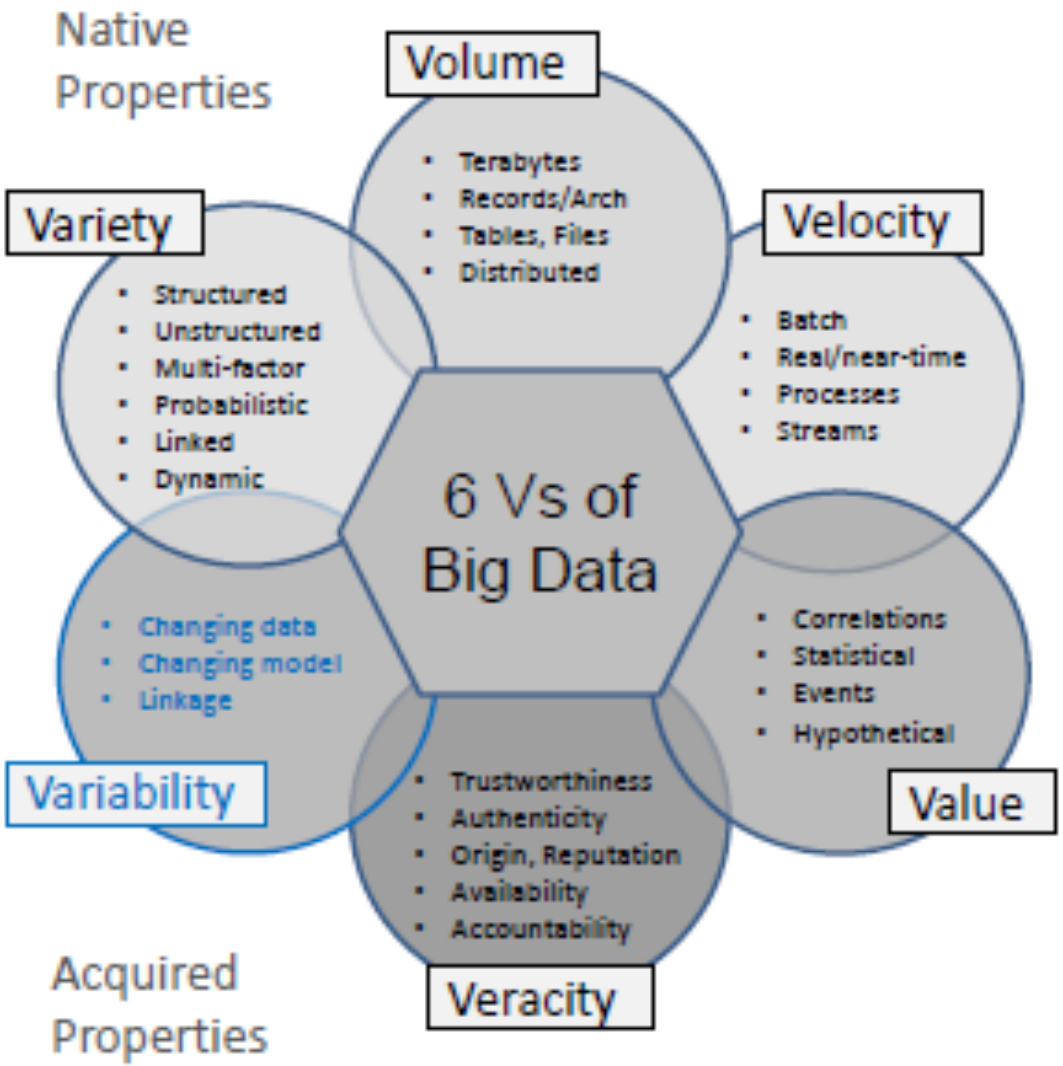
Dr. Paola Grosso
Email: p.grosso@uva.nl
URL: http://staff.science.uva.nl/~grosso

UNIVERSITEIT VAN AMSTERDAM

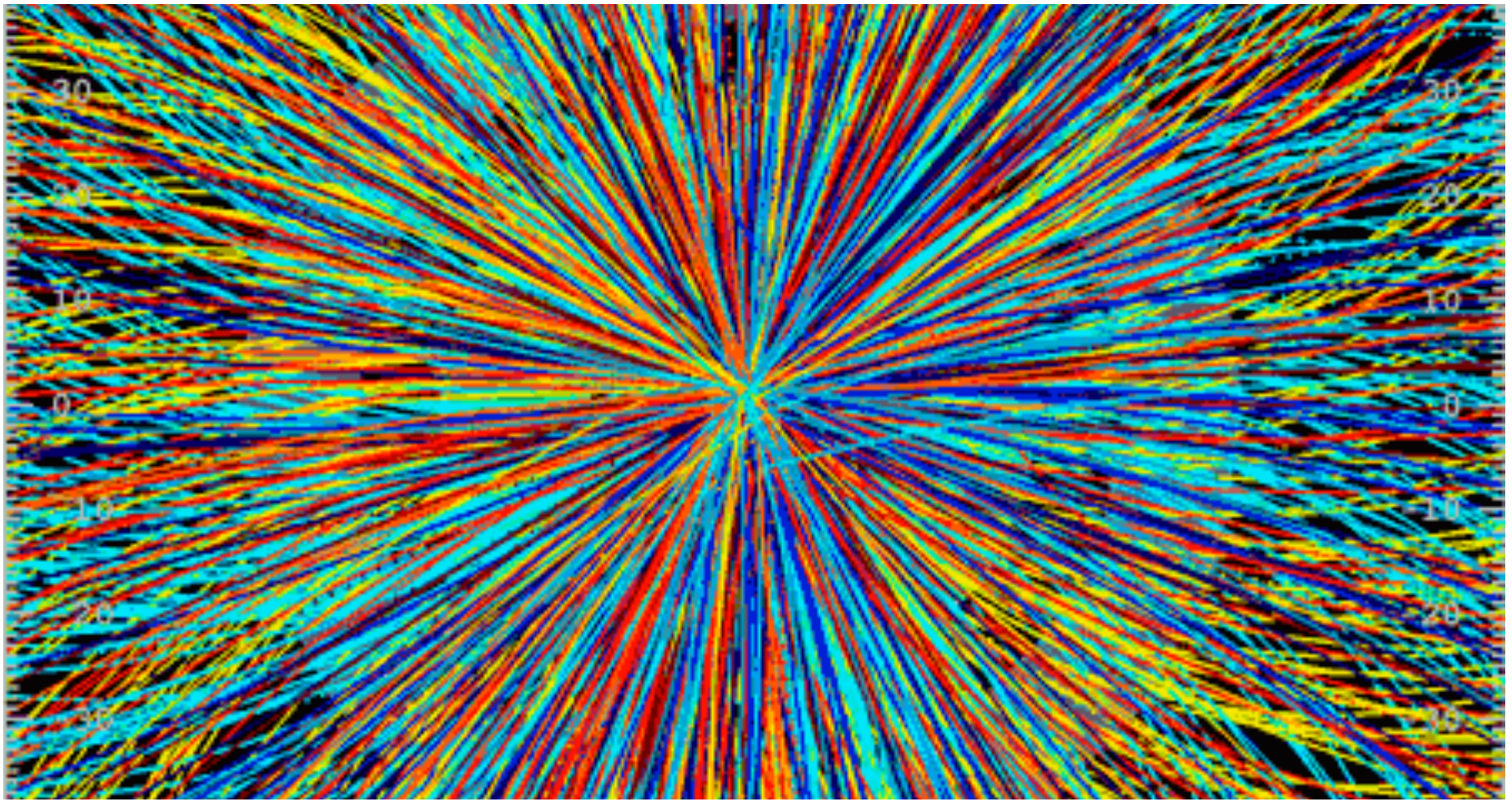
System and Network Engineering

Addressing Big Data Issues in Scientific Data Infrastructure
Y. Demchenko, P. Membrey, P. Grosso and C.de Laat
In: First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 International Conference on Collaboration Technologies and Systems (CTS 2013), 2013

The Big Data Challenge



Big Science



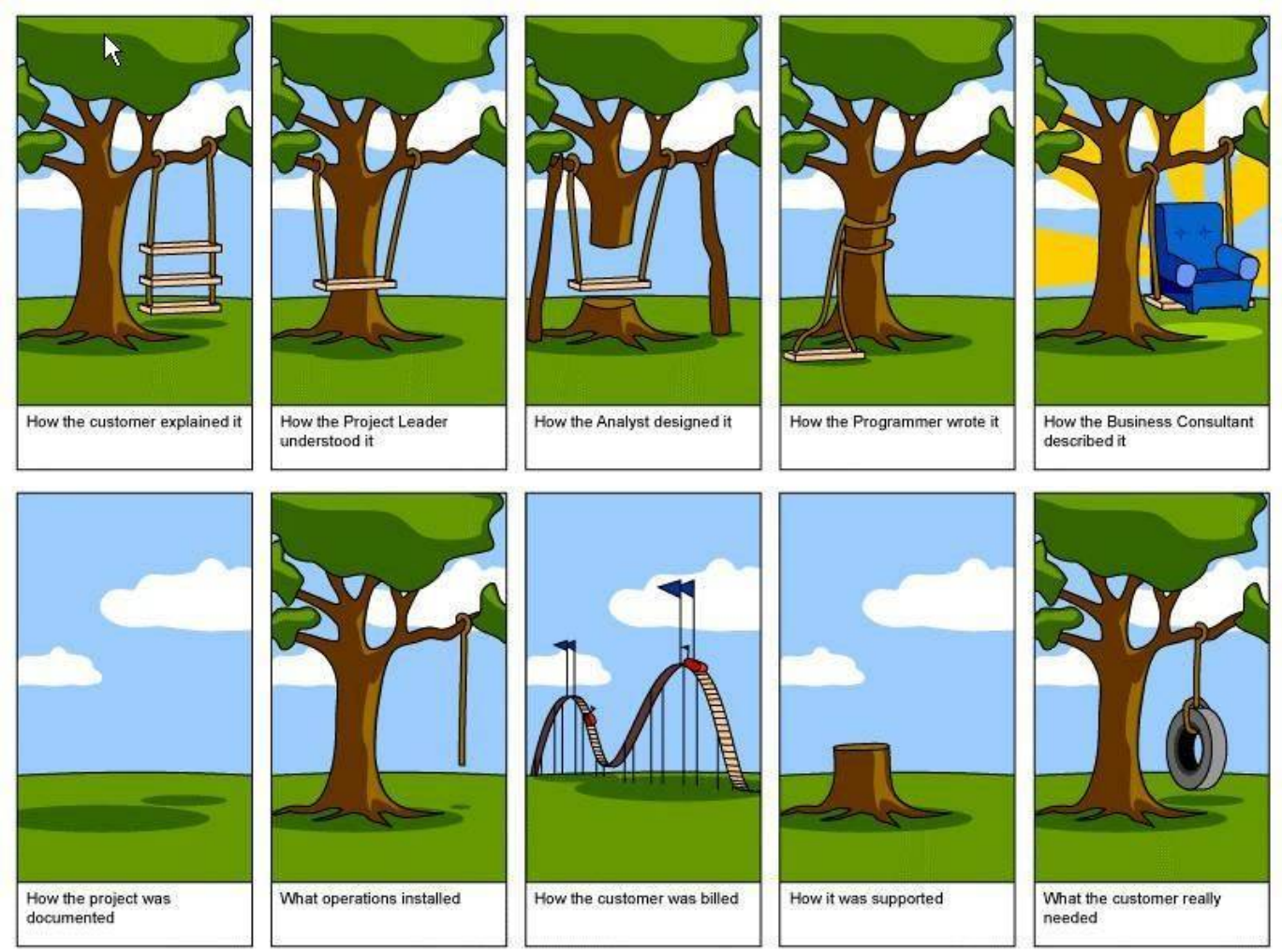
The virtualization opportunity

A changing relation between applications and infrastructures.

BEFORE	NOW
A fixed infrastructure with the application that molds to the infrastructure	Virtualization enables the infrastructure to adapt to the application

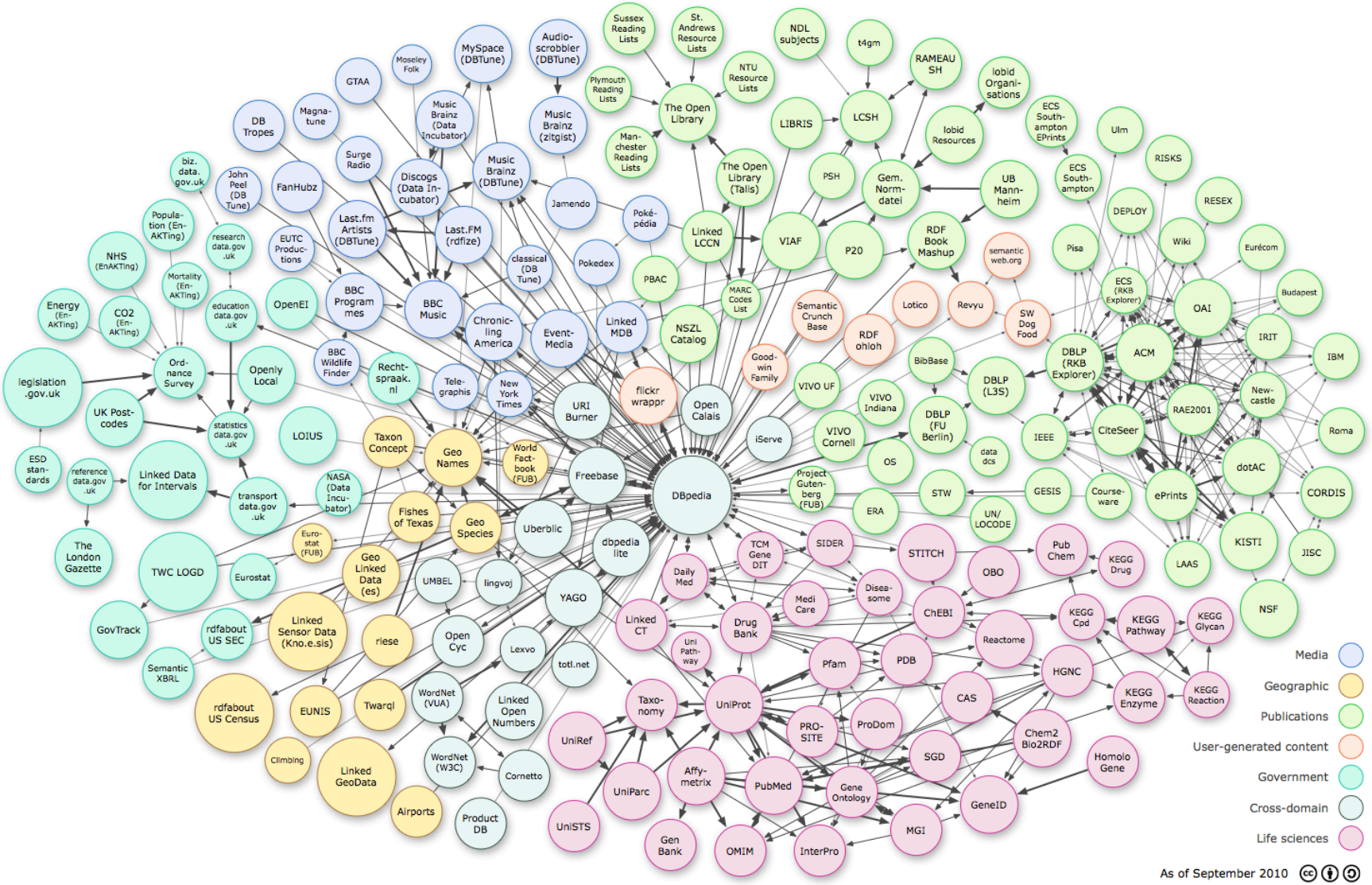
This talk provides an overview of the research done on this theme in our group (System and Network Engineering).

Problem #1: how can infrastructures expose their enhanced capabilities?



Semantic models

Open Linked Data



The Semantic Web

- RDF - Resource Description Framework - provides a way to categorize information:
  - resources are described by URIs;
  - triples define the relations between resources:

- OWL – Web Ontology Language - has stronger support for classes, attributes and constraints
  - Operations (unions, intersections, complements, cardinality constraints)

Information models

"One of the **main ingredients** *in the design, implementation and operation of cloud computing infrastructures is the* **information model**. *This information model must describe both the physical infrastructure and its virtualization aspects"*

Information model

An information model describes resources at a conceptual layer.

Data model

A data model describes protocols and implementation details, based on the representation of concepts and their relations provided by the information model.

A Semantic-Web Approach for Modeling Computing Infrastructures
M.Ghijsen, J. van der Ham, P. Grosso, C.Dumitru, H. Zhu, Z. Zhao and C. de Laat
Elsevier Journal of Computers and Electrical Engineering – 2013

INDL

An effort started in 2010 (in parallel with our involvement in the FP7 projects Geysers and NOVI).

The goal was to capture the concept of virtualization in computing infrastructures and to describe the storage and computing capabilities of the resources.

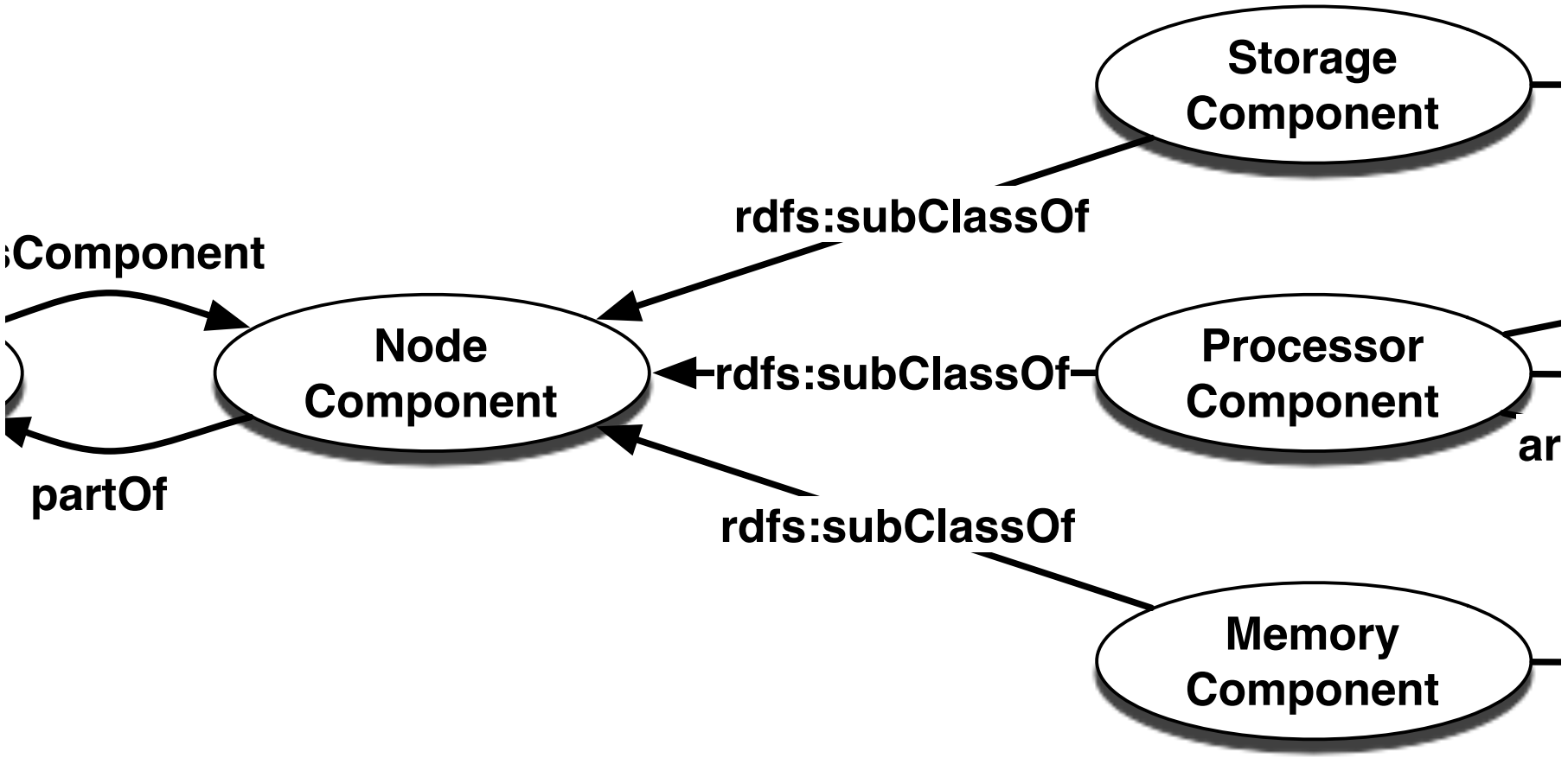
A key feature is the decoupling of virtualization, connectivity and functionalities.

It is built upon the NML ontology, an OGF standard.
It uses the nml:node concept as basic entity to describe resources in computing infrastructures.

It can be used as:

- a stand-alone model (i.e. without any network descriptions),
- in combination with NML by importing the NML ontology into the INDL definition.

Node components

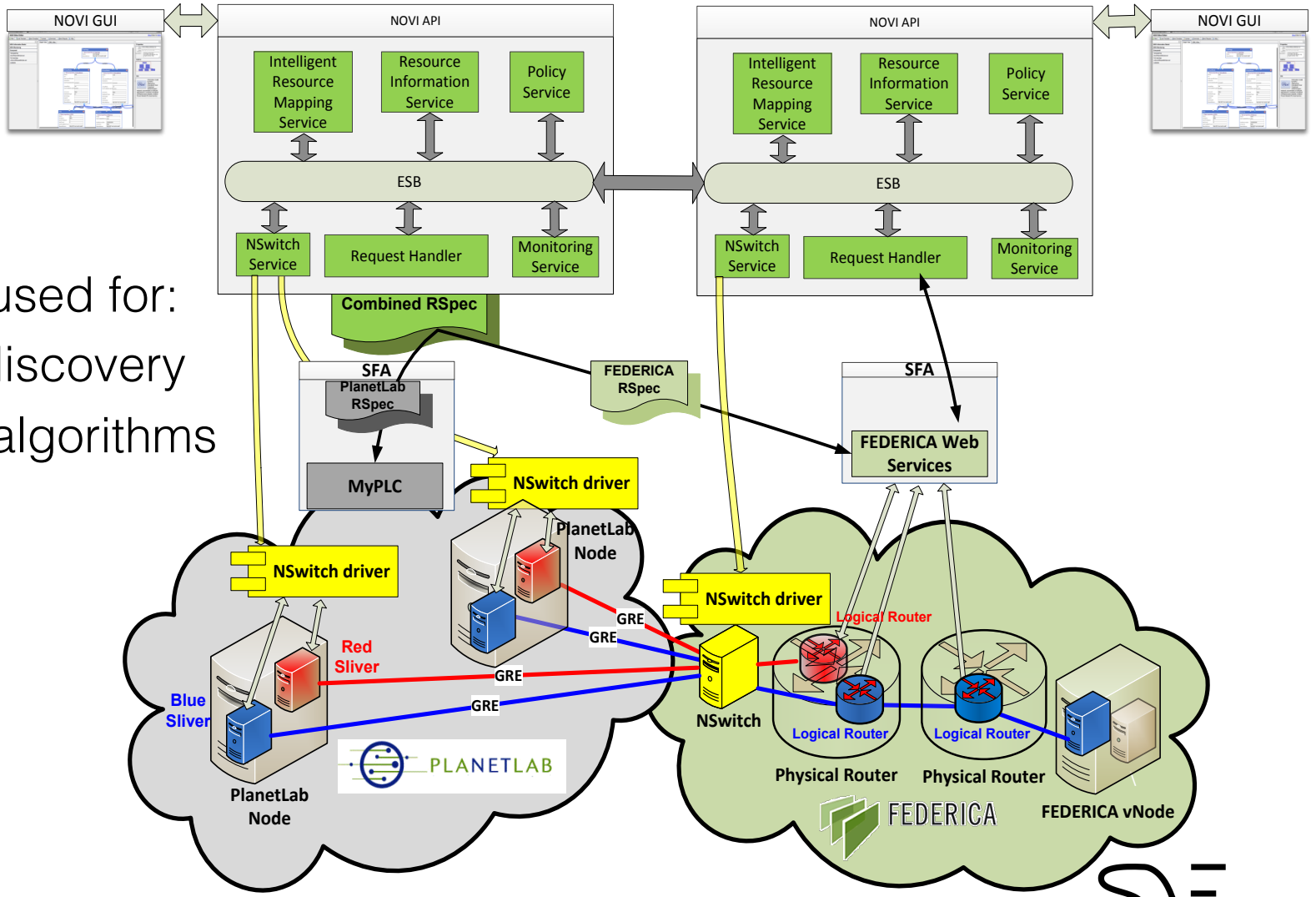


Resource discovery and allocation for federated virtualized infrastructures
C. Pittaras, C. Papagianni, A. Leivadeas, P. Grosso, J. van der Ham, S. Papavassiliou
To appear in: Future Generation Computer Systems 2014

NOVI Federation

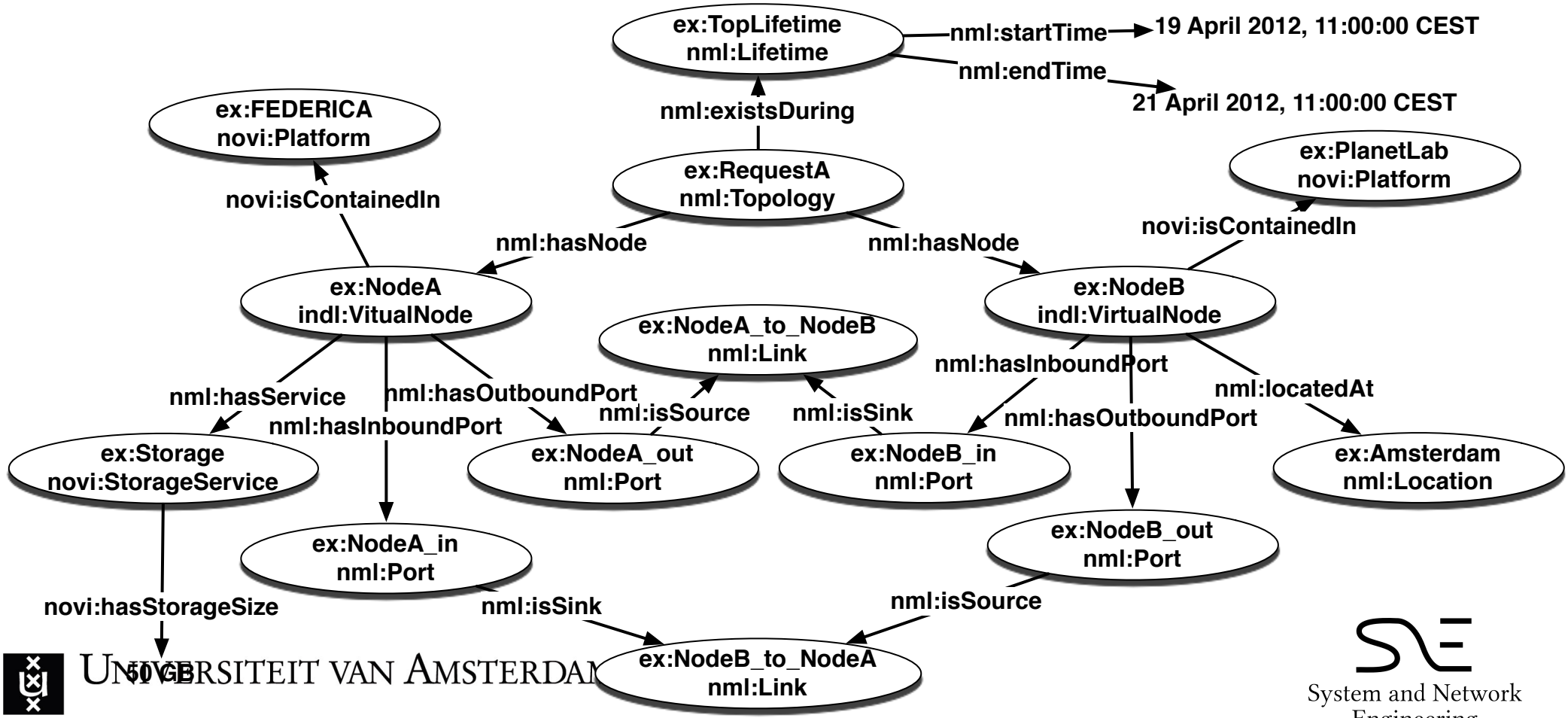
Our model is used for:

- Resource discovery
- Embedding algorithms



INDL use in NOVI

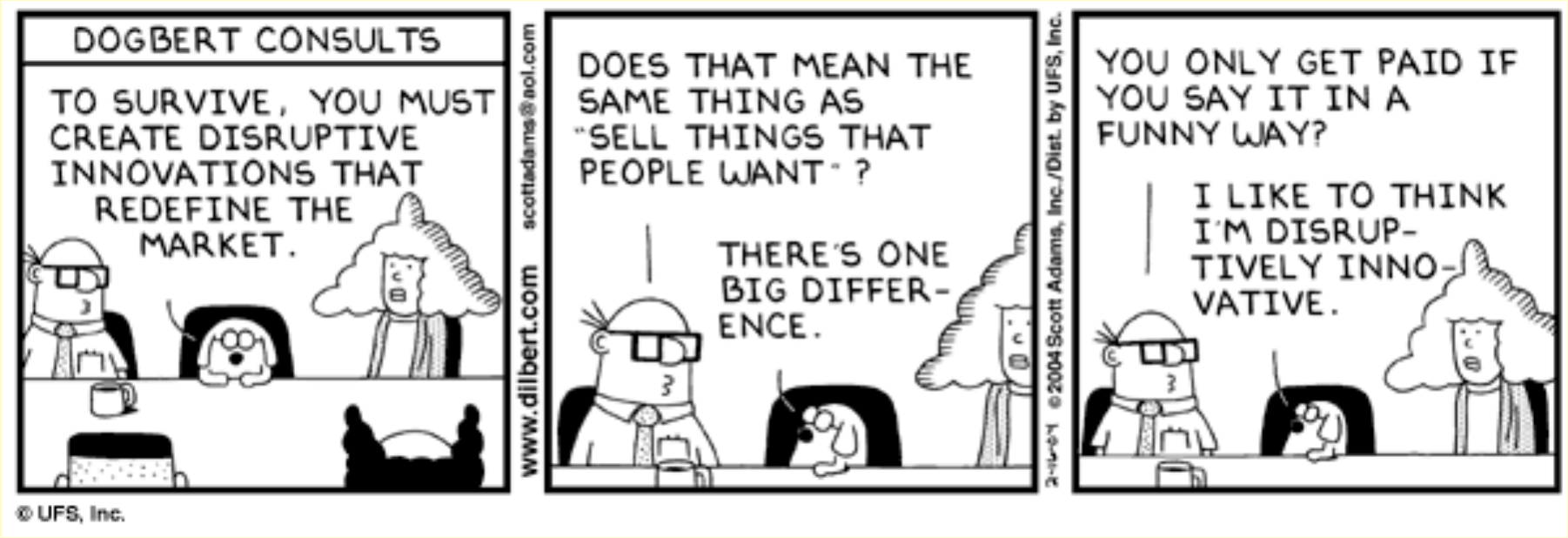
- Two nodes in the NOVI federation:



UNIVERSITEIT VAN AMSTERDAM

System and Network Engineering

Problem #2: what can people actually do?



(Network) services for Big Data applications

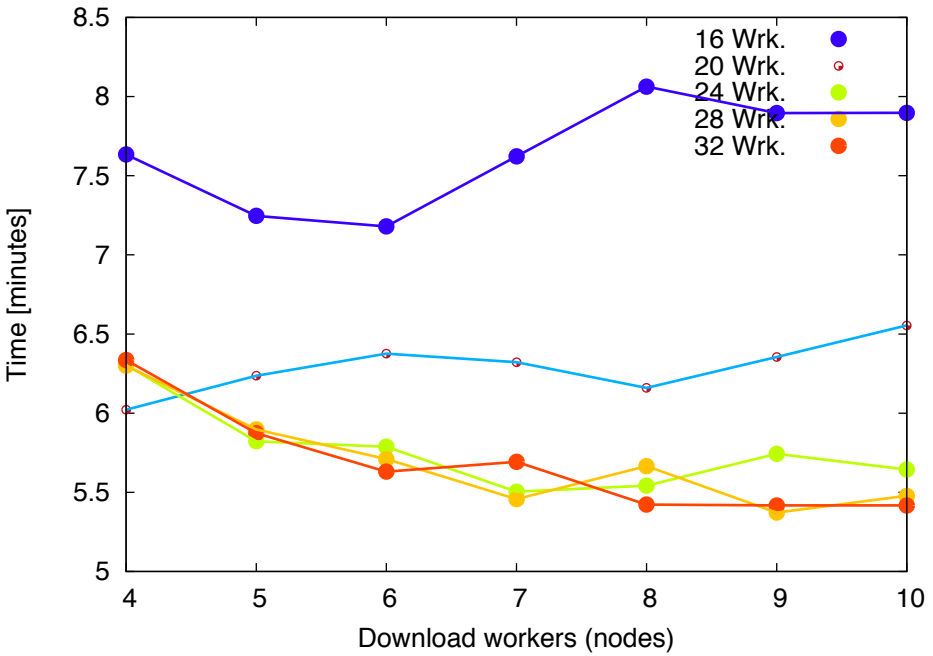
Automated advanced capabilities

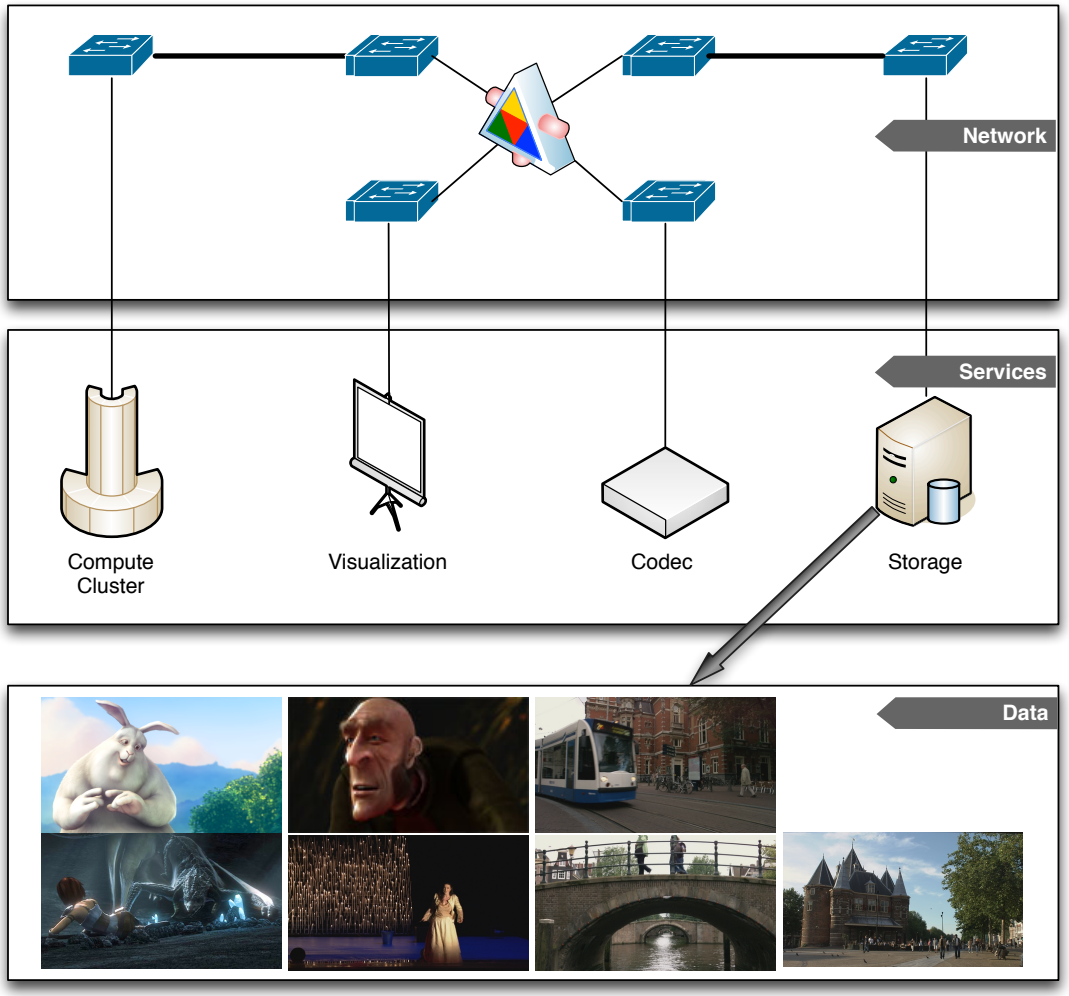
Such as:

- intelligent resource mapping,
- policy-driven access and resource allocation,
- context aware resource discovery,
- transparent data plane connectivity and
- monitoring of combined user slices and substrate resources across domains

HyperFlow

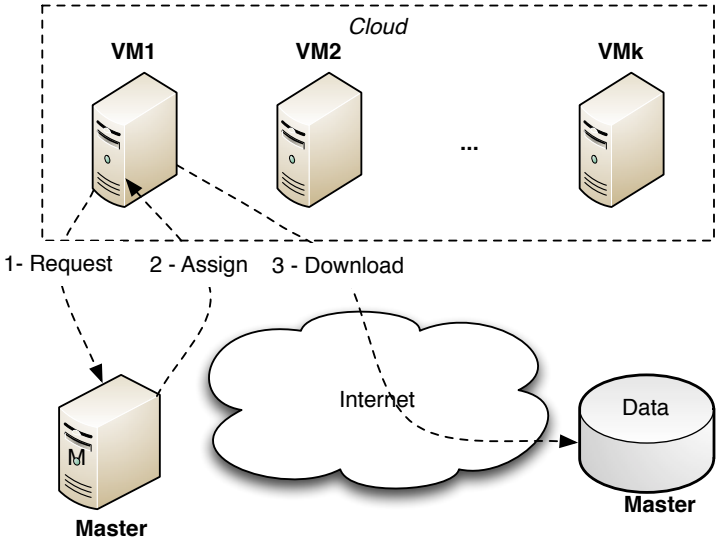
Encoding times improve as the end nodes are connected via dynamic lightpaths

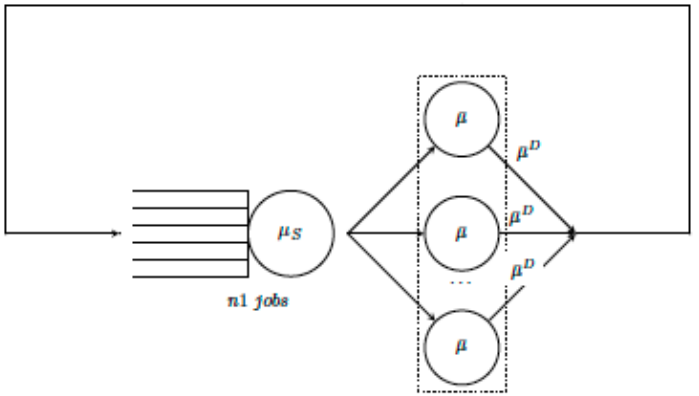


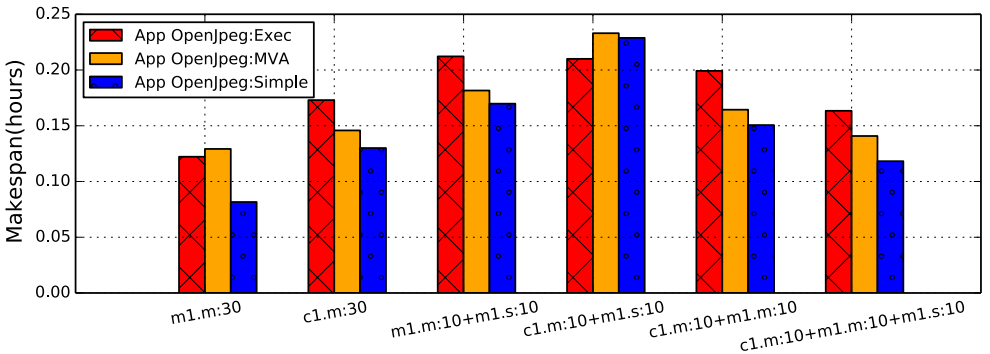


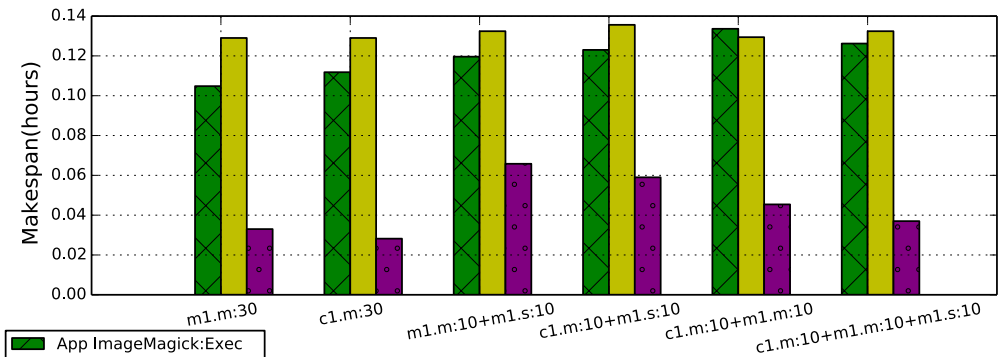
C. Dumitru, Z. Zhao, P. Grosso and C. de Laat
HybridFlow: Towards Intelligent Video Delivery and Processing Over Hybrid Infrastructures
(In CTS 2013))

A queueing model approach



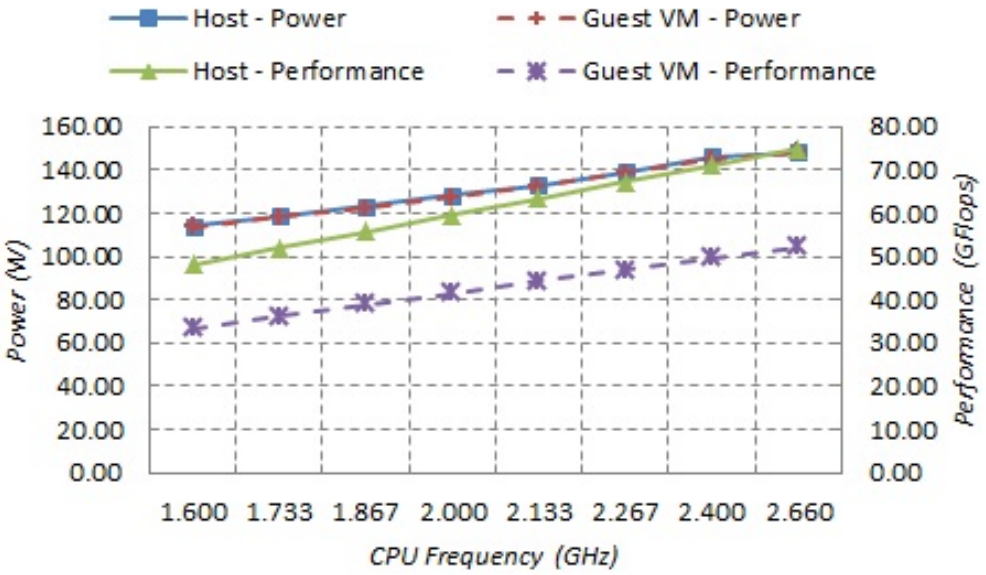






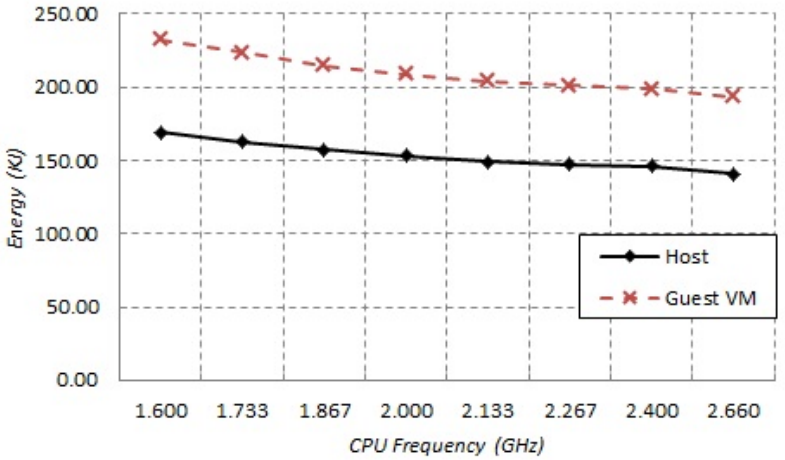
A Queueing Theory Approach to Pareto Optimal Bags-of-Tasks Scheduling on Clouds
C. Dumitru, A. Oprescu, M. Zivkovic, R. v/d Mei, P. Grosso and C.de Laat
Submitted to Europar2014

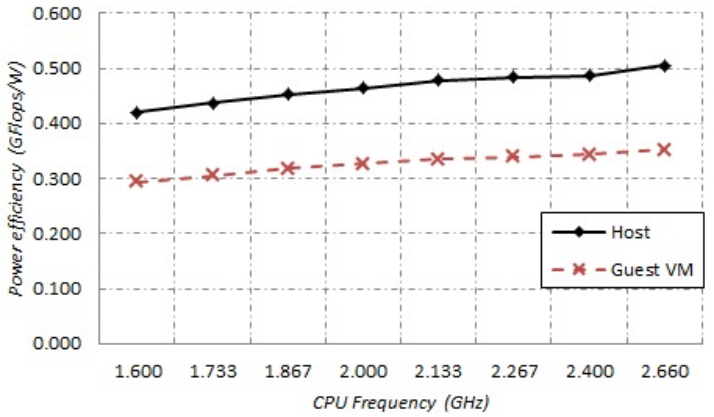
Energy saving in clouds



Quantifying the energy performance of VMs is the first step toward energy-aware job scheduling.

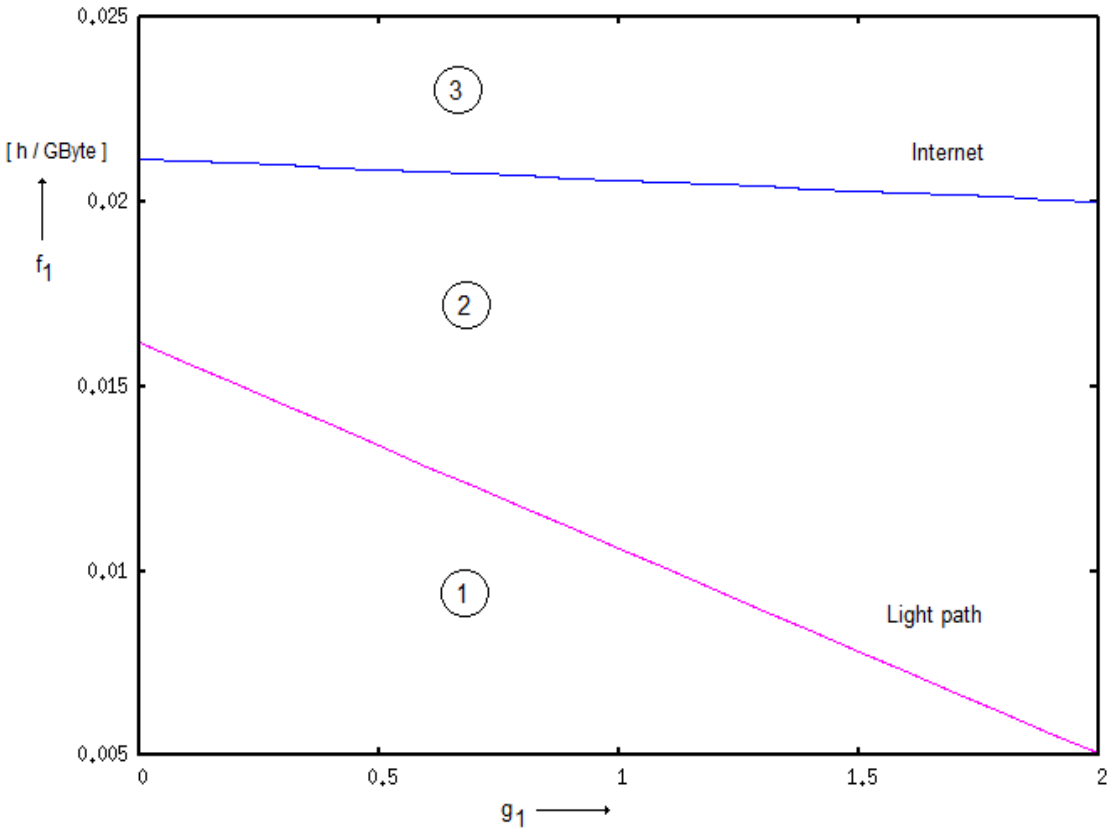
Q. Chen, P. Grosso, K. van der Veldt, C. de Laat, R. Hofman and H.Bal.
Profiling energy consumption of VMs for green cloud computing
In: International Conference on Cloud and Green Computing (CGC2011), Sydney December 2011





A decision framework for placement of applications in clouds that minimizes their carbon footprint
M. Makkes, A. Taal, A. Osseyran and P. Grosso
In: Journal of Cloud Computing: Advances, Systems and Applications 2013, Vol.2

Results



In region 1 the task should be performed locally, independently of the type of transport network.

In region 2 the task can be performed remotely provided that the connection is a light path.

In region 3 the task should be done remotely for both types of transport networks.

Given different network paths we can identify decision boundaries as function of the task complexity.

Problem #3: how can application control the infrastructure?

COMMIT/

InterClouds Operating System (ICOS)

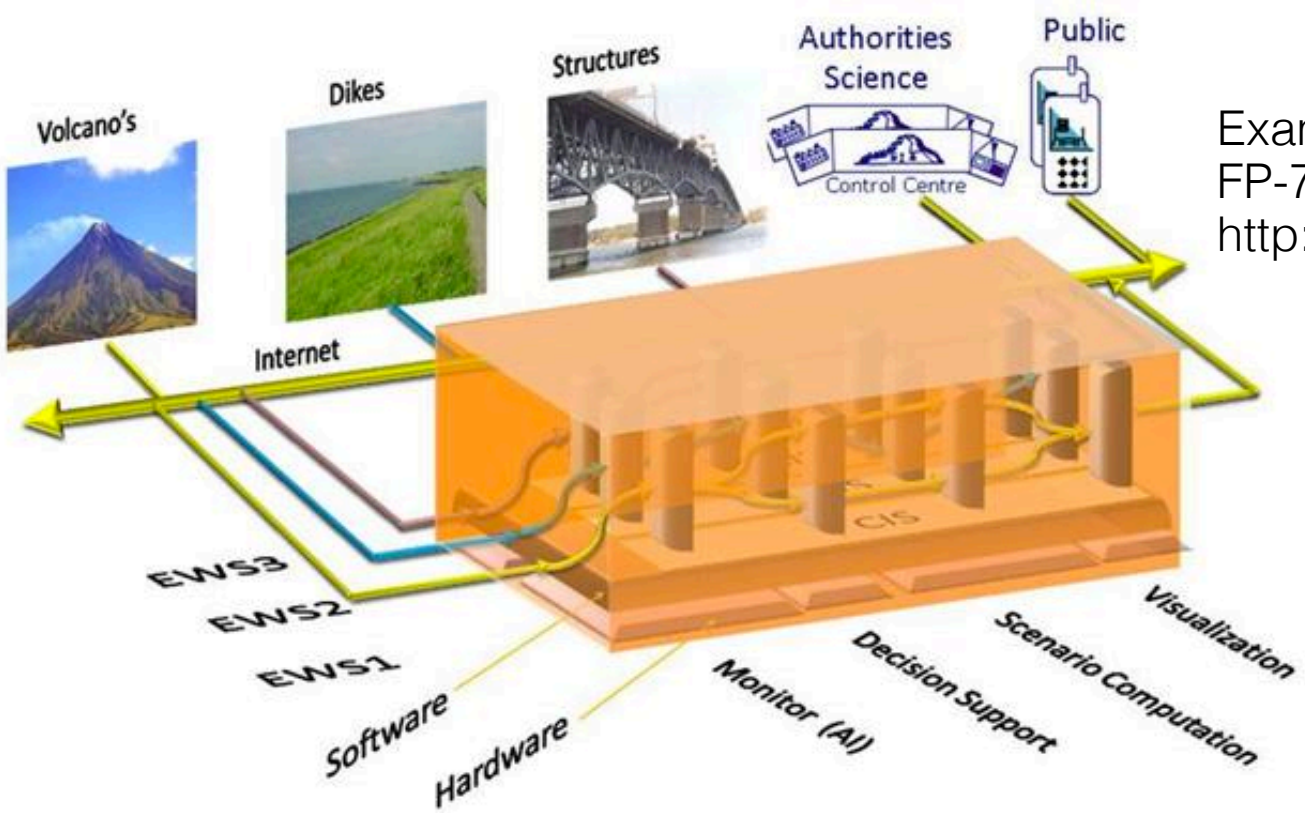
Universiteit van Amsterdam

System and Network Engineering

"*Internet factories: Creating application-specific networks on-demand*"
R. Strijkers, M. Makkes, C. de Laat and R. Meijer
In: Elsevier Journal of Computer Networks (2014)

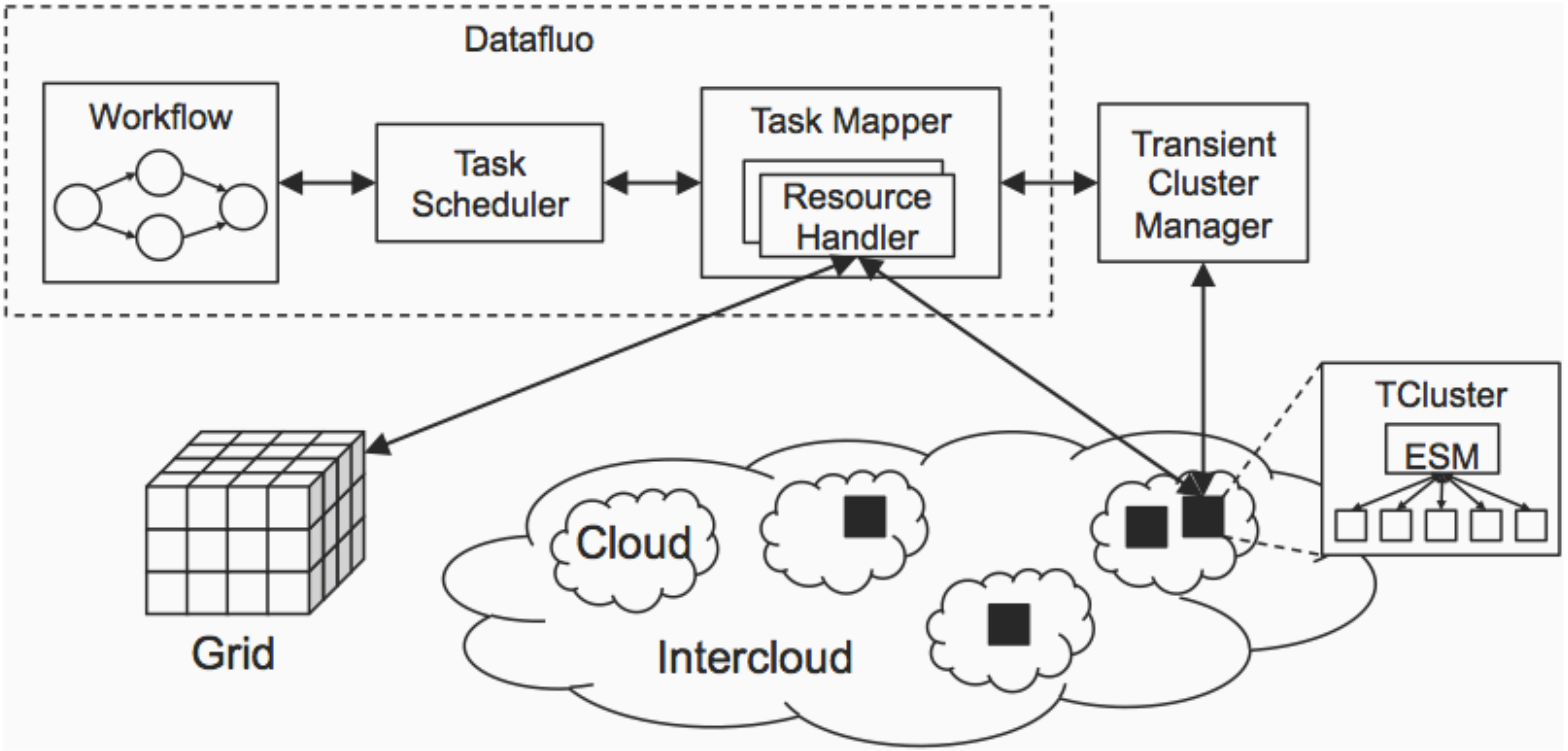
Interclouds and cyber physical systems

- Cloud federations have gained research attention
- Potential users are cyber-physical systems



Example:
FP-7 project Urbanflood
http://www.urbanflood.eu/

ICOS components



"Towards an operating system for intercloud,"
R. Strijkers, R. Cushing, M. Makkes, P. Meulenhoff, A. Belloum, C. de Laat and R. Meijer
In: 2013 IEEE International Conference on Cloud Computing Technology and Science (Cloudcom2013)

Networked Open Processes

Away from SWMS

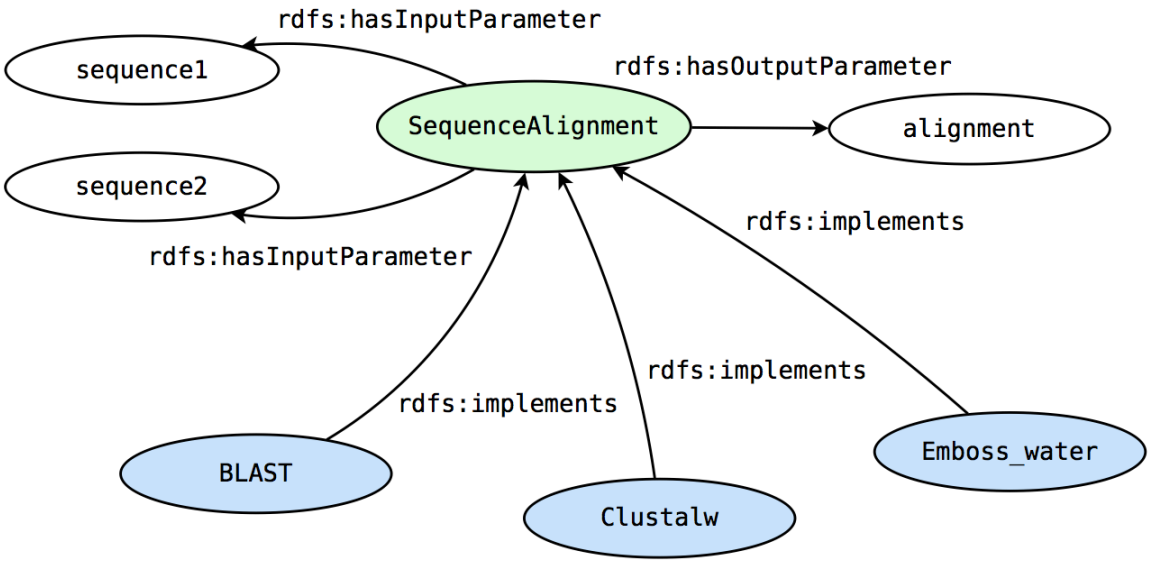
SWMS? Scientific Workflow Management Systems

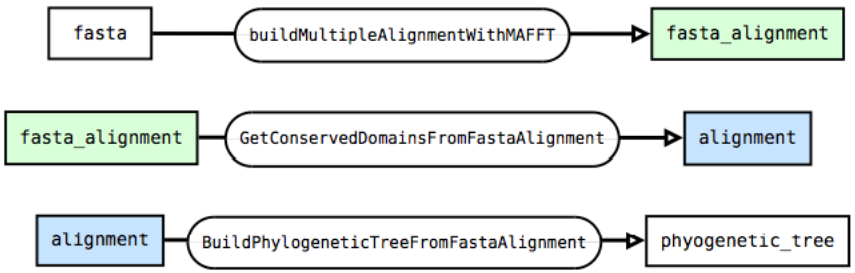
The next generation of distributed scientific computing will move towards open systems that can autonomously construct workflows using a global space of processes and minimal declarations by scientists to construct experiments.

Processes and services are semantically annotated in a way that makes it easy for discovering networks between processes.

Aided by semantics

"Beyond scientific workflows: Networked open processes"
R. Cushing, M. Bubak, A. Belloum, and C. de Laat
In: 2013 IEEE 9th International Conference on eScience (eScience)





Conclusions

Several ‘problems’ and an

- Semantic Web can be used to describe
  1. the infrastructure supporting Big Data processing (INDL)
  2. the processing making those processes (Open Network Processes).

- Big Data will require federative environments (InterClouds) where applicaions can fully exploit elasticity.

Want to know more?

- Contact me:

p.grosso@uva.nl
http://staff.science.uva.nl/~grosso/

- Our groups webpage:

http://sne.science.uva.nl/

Or see the publications listed in the following pages.