A Grid-based Infrastructure for Distributed Retrieval

F. Simeoni[1] **L. Candela**[2] G. Kakaletris[3] M. Sibeko[4]
P. Pagano[2] G. Papanikos[3] P. Polydoras[3] Y. Ioannidis[3]
D. Aarvaag[4] F. Crestani[1]

[1]Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK
{fabio.simeoni, f.crestani}@cis.strath.ac.uk

[2]Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – CNR, Pisa - Italy
{candela, pagano}@isti.cnr.it

[3]Computer Science Department, University of Athens – Athens, Greece
{g.kakaletris, g.papanikos, p.polydoras, yannis}@di.uoa.gr

[4]Fast Search & Transfer ASA, Oslo, Norway
{mads.sibeko, dagfinn.aarvaag}@fast.no

Research trend

e-Science scenarios (multidisciplinary and co-operative) face novel challenges

- highly-evolving requirements
- large scale resources and players distribution
- heterogeneity

. . . making standard development approaches often too "expensive" (and not sustainable)

- "from-scratch" development of ad-hoc solutions
- HW investment (even if intermittently needed)

The "magic" formula to reduce costs

sharing & reuse

Research trend

e-Science scenarios (multidisciplinary and co-operative) face novel challenges

- highly-evolving requirements
- large scale resources and players distribution
- heterogeneity

. . . making standard development approaches often too "expensive" (and not sustainable)

- "from-scratch" development of ad-hoc solutions
- HW investment (even if intermittently needed)

The "magic" formula to reduce costs

sharing & reuse

Research trend

e-Science scenarios (multidisciplinary and co-operative) face novel challenges

- highly-evolving requirements
- large scale resources and players distribution
- heterogeneity

. . . making standard development approaches often too "expensive" (and not sustainable)

- "from-scratch" development of ad-hoc solutions
- HW investment (even if intermittently needed)

The "magic" formula to reduce costs

sharing & reuse

[Grid-based] Infrastructures as enabling technologies

A physical and organisational structure based on the principle of resource sharing to serve one or more communities and support their operation

- current-generation focuses on low-level resources, e.g. network, storage, computing
- next-generation builds on top of the previous to extend the vision into application domains, e.g. retrieval services

The impact is potentially non-trivial

- co-ordinated sharing of these resource may invalidate cost analysis of current solutions, i.e. “expensive” solutions may be outsourced to the infrastructure
- novel solutions tackling classic problems (e.g. Information Retrieval) in novel ways can be realised

[Grid-based] Infrastructures as enabling technologies

A physical and organisational structure based on the principle of resource sharing to serve one or more communities and support their operation

- current-generation focuses on low-level resources, e.g. network, storage, computing
- next-generation builds on top of the previous to extend the vision into application domains, e.g. retrieval services

The impact is potentially non-trivial

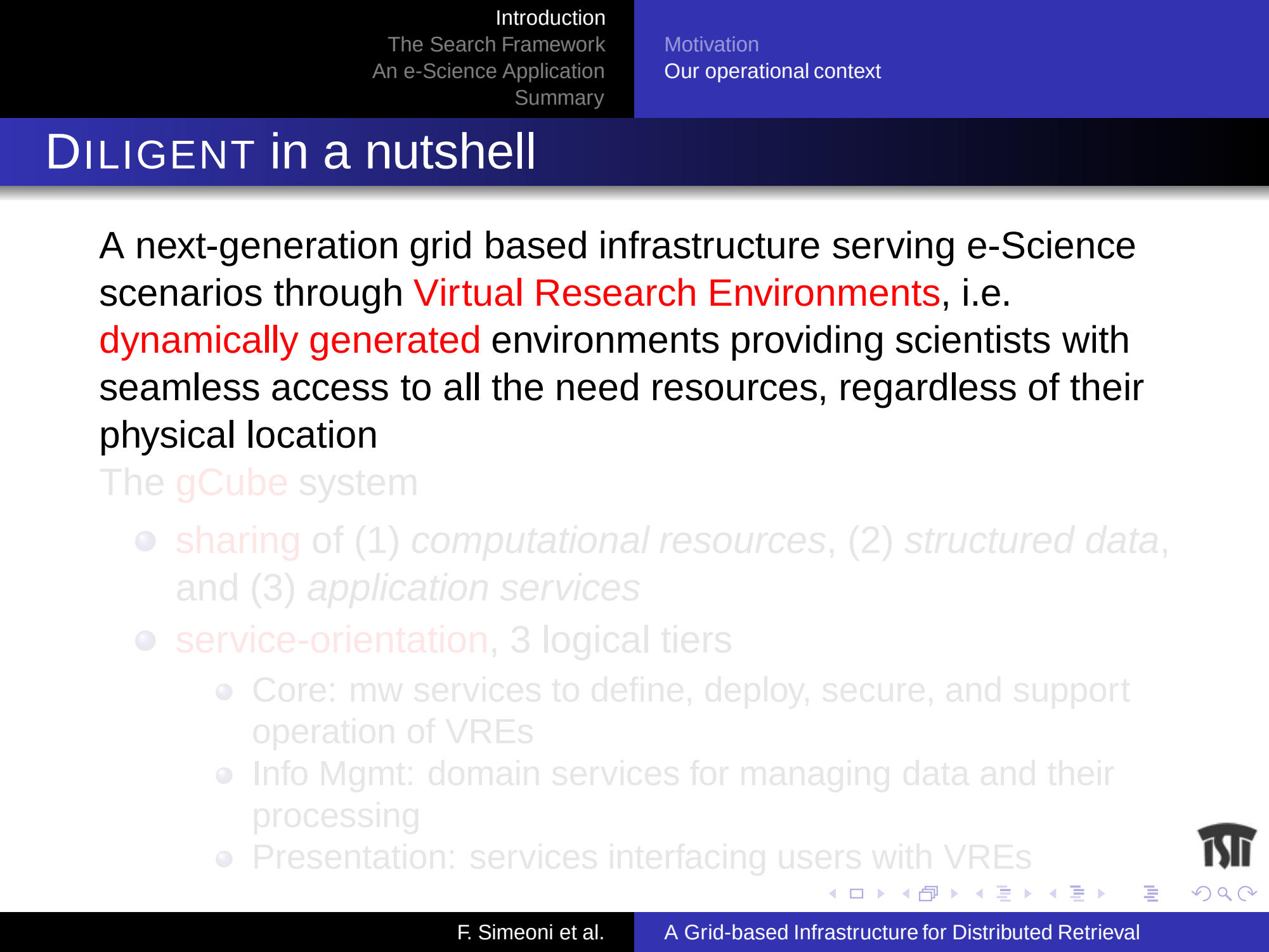
- co-ordinated sharing of these resource may invalidate cost analysis of current solutions, i.e. “expensive” solutions may be outsourced to the infrastructure
- novel solutions tackling classic problems (e.g. Information Retrieval) in novel ways can be realised

DILIGENT in a nutshell

A next-generation grid based infrastructure serving e-Science scenarios through Virtual Research Environments, i.e. dynamically generated environments providing scientists with seamless access to all the need resources, regardless of their physical location

The gCube system

- sharing of (1) *computational resources*, (2) *structured data*, and (3) *application services*
- service-orientation, 3 logical tiers
  - Core: mw services to define, deploy, secure, and support operation of VREs
  - Info Mgmt: domain services for managing data and their processing
  - Presentation: services interfacing users with VREs



DILIGENT in a nutshell

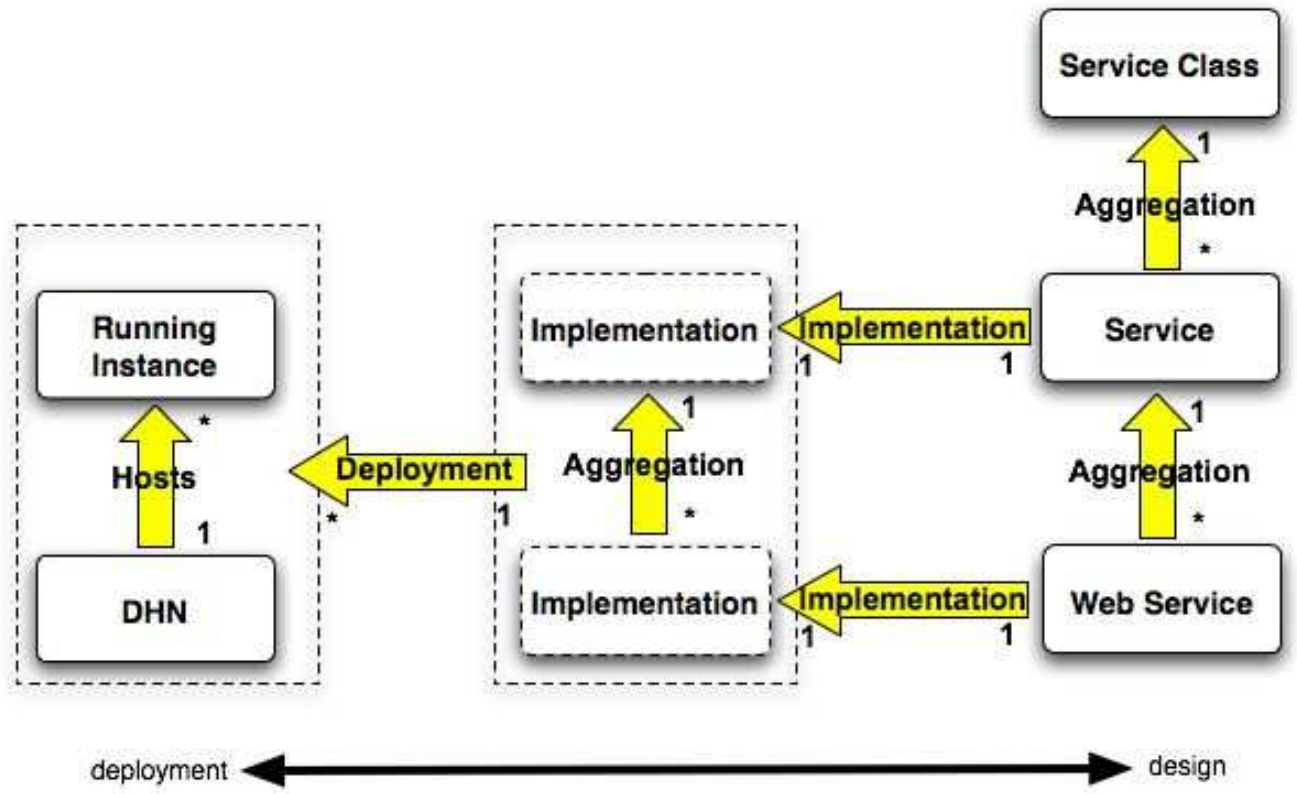
A next-generation grid based infrastructure serving e-Science scenarios through Virtual Research Environments, i.e. dynamically generated environments providing scientists with seamless access to all the need resources, regardless of their physical location

The gCube system

- sharing of (1) *computational resources*, (2) *structured data*, and (3) *application services*
- service-orientation, 3 logical tiers
  - Core: mw services to define, deploy, secure, and support operation of VREs
  - Info Mgmt: domain services for managing data and their processing
  - Presentation: services interfacing users with VREs

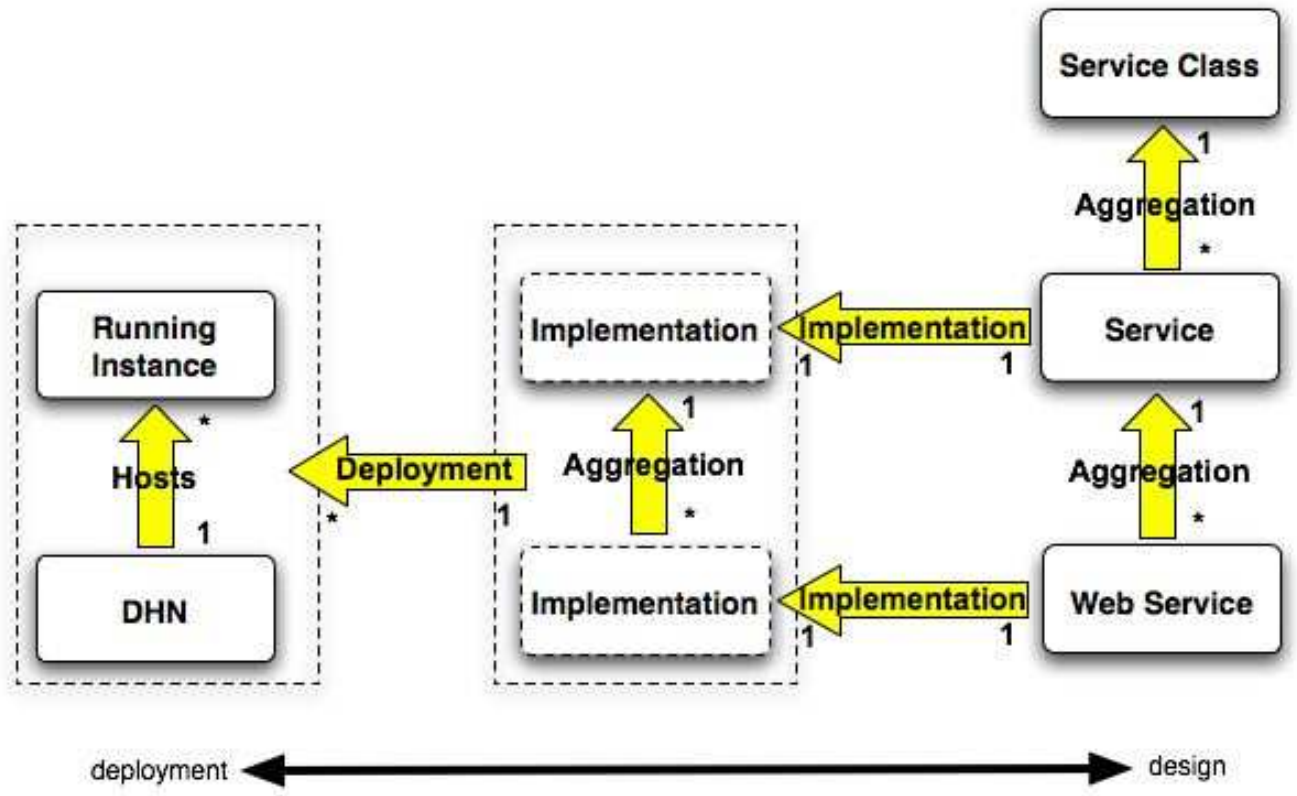
The Service Model

service classes, i.e. flat groupings of services within the same functional area (e.g. the *Index* class)



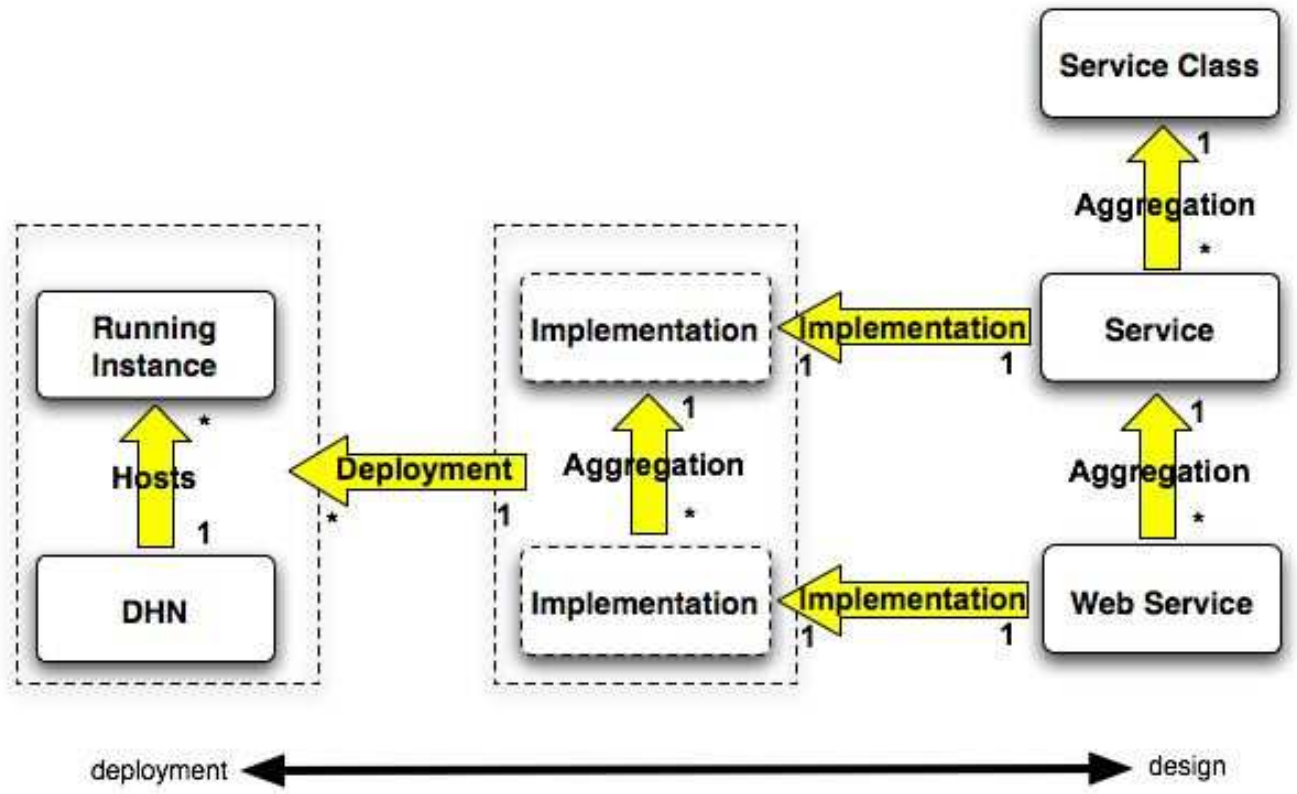
The Service Model

services, i.e. abstract instances of service classes (e.g. the *LookupService* in the Index class)



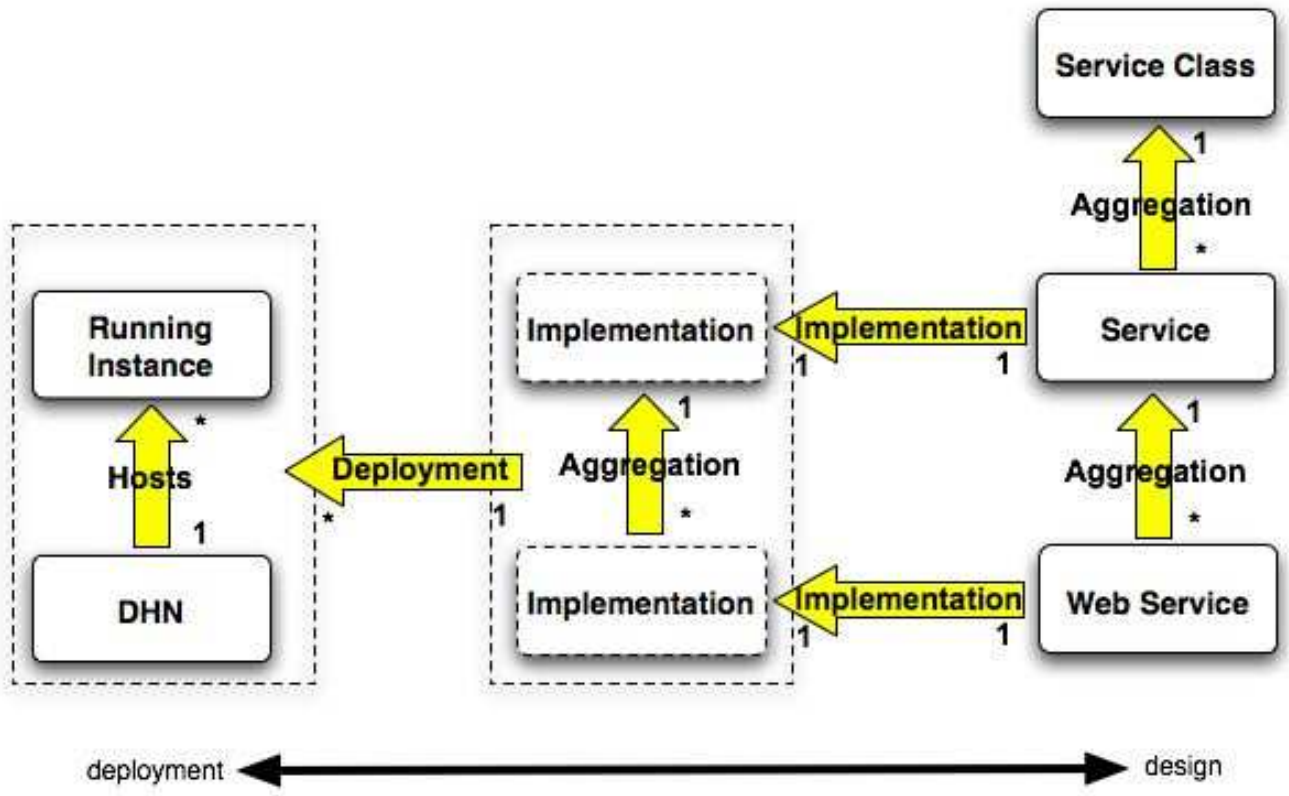
The Service Model

web services, i.e. entry-points to concrete implementations of abstract services (e.g. the *LookupFactoryService*)



The Service Model

running instances, i.e. service implementations dynamically deployed on available hosting nodes



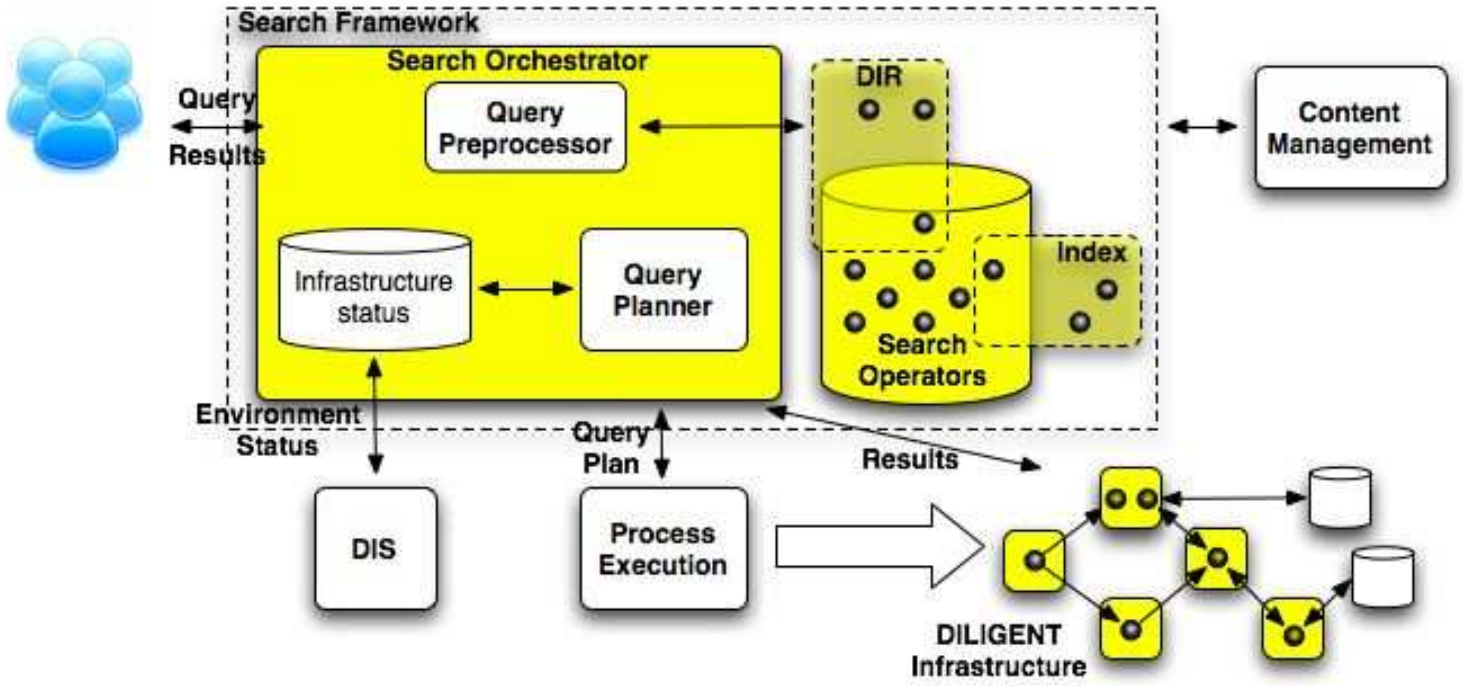
Goal

Equip the gCube system with a search framework (not a "simple" search engine!)

- modular
  - diverse, autonomous, and pluggable elements
- suited to maximise infrastructure exploitation and support
  - state-of-the-art IR solutions implementations on-board
  - optimal search elements (and resources) consumption (avoid mis-utilisation and misuse)
  - capabilities to capture complex application scenarios by combining information retrieval and data processing procedures
  - operational framework for experimentation with novel IR technologies and solutions

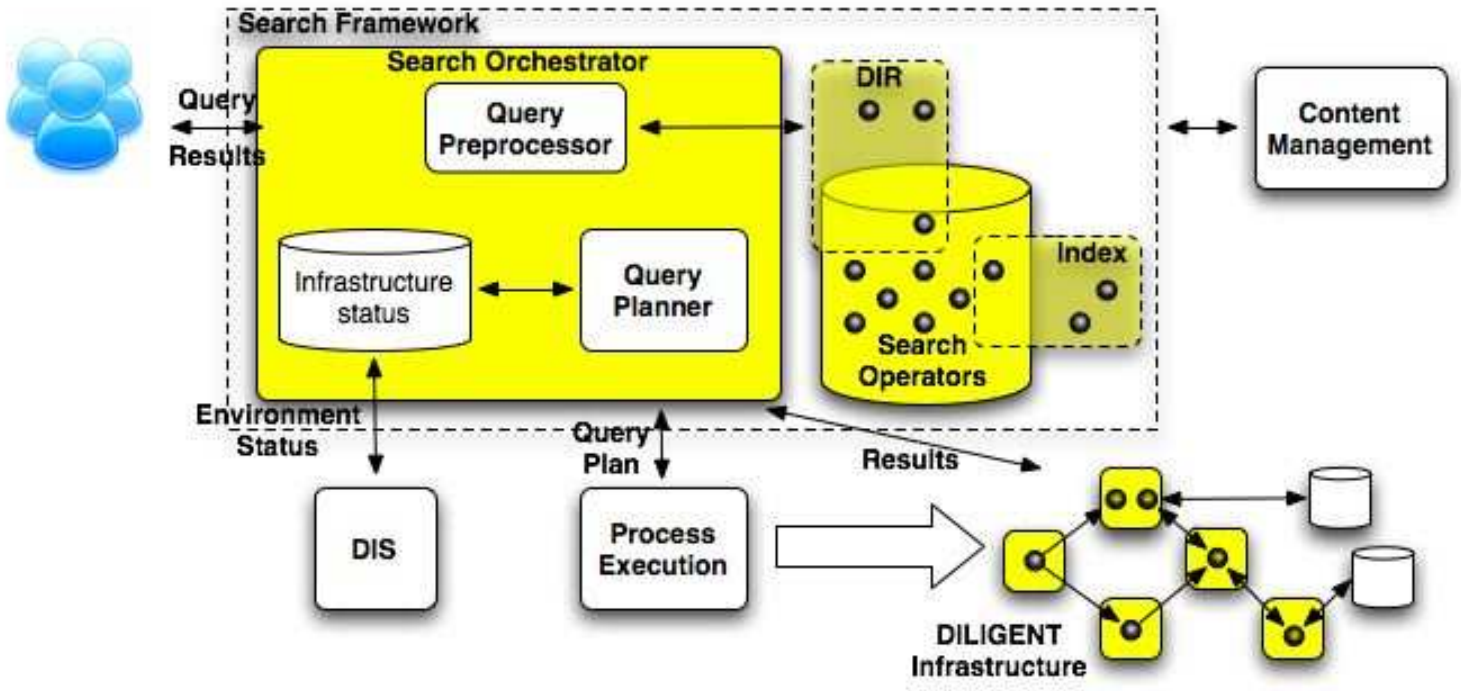
Overall View

Search Orchestrator: manages the whole query process by relying on available resources and operators



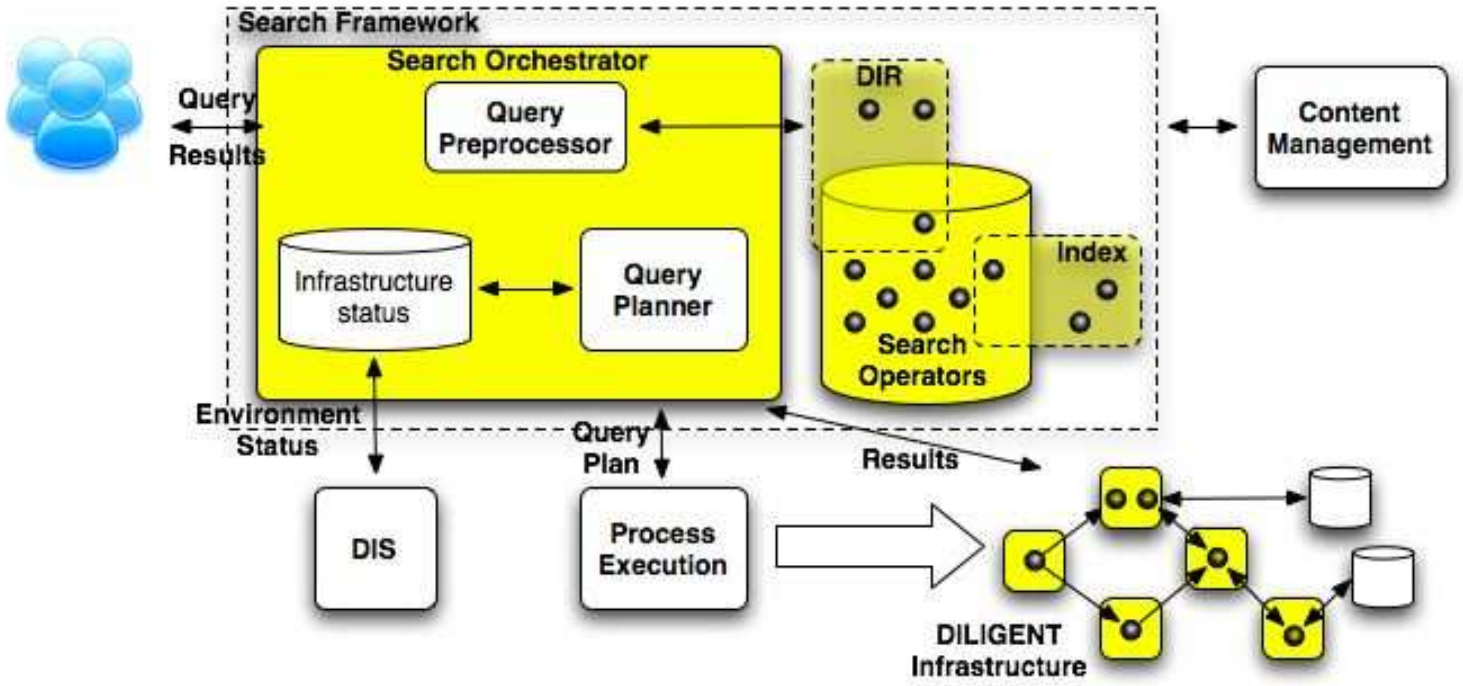
Overall View

Index: provides for indices enhancing query response



Overall View

DIR (Distributed Information Retrieval): provides for searching across multiple sources



Search Services

A complex service class consisting of

- A Search Orchestrator of Information Retrieval processes
  - Consolidates query and environment information
  - Prepares and plans retrieval execution (outsourced to grid)
- Numerous “worker” services provide data processing features, e.g.
  - XML processing (joiners, sorters, transformers, filterers)
  - Lookups (Index, etc.)
  - Results merging (DIR)
- A data transport mechanism (ResultSet)
  - Overcomes environment restrictions towards performance
  - Standardizes information exchange among IR workers

Index Services

An Index is a service that replies to queries by relying on a data structure (full index). Different structures imply different indexes

- Full Text, Forward, VA files and Geographical Index (R-Tree)
- *delta files* to manage replication

Different services for different roles:

- *Manager*, manages and represents one specific index
- *Updater*, consumes data source and generates delta files
- *Lookup*, replicates the index for queries lookup

For possible organisations:

- *All-in-one*, minimise latency
- *Lookup separated*, ideal for query intensive scenarios
- *Updater separated*, ideal for intense feeding scenarios
- *One Service per role*, updater and lookup replication

DIR Services

Three service classes:

- *Content Source Description*, generates and maintains summary descriptions of each source, e.g. partial indices
  - cooperative (index) vs uncooperative (query sampling)
- *Content Source Selection*, limits the routing of queries to the “best” target sources
  - various weighting algorithms
- *Data Fusion*, derive a total order of result sets produced by different data sources
  - various merging algorithms

Different services for different roles

- *access*, serves requests by relying on the current RI state
- *monitor*, observes of the environment for changes
- *factory*, creates the state of the RIs

ImpECt

ImpECt goal is to support Earth Observation community in managing the reports production task

- have seamless access to heterogeneous information sources
  - Environmental Reports, workshops proceedings, science papers, presentations
  - High level geophysical products, DEMs (Digital Elevation Models) and added value maps

from

- ESA EO web portal (`www.eoportal.org`)
- ESA Grid on-demand (`eogrid.esrin.esa.int`)
- Geonetwork (`www.fao.org/geonetwork`)
- Web Map Servers
- European Environment Agency (`www.eea.eu.int`)
- NASA CEOS IDN (`idn.ceos.org`)
- Medspiration (`www.mespiration.org/products`)

ImpECt

ImpECt goal is to support Earth Observation community in managing the reports production task

- have seamless access to heterogeneous information sources
  - Environmental Reports, workshops proceedings, science papers, presentations
  - High level geophysical products, DEMs (Digital Elevation Models) and added value maps

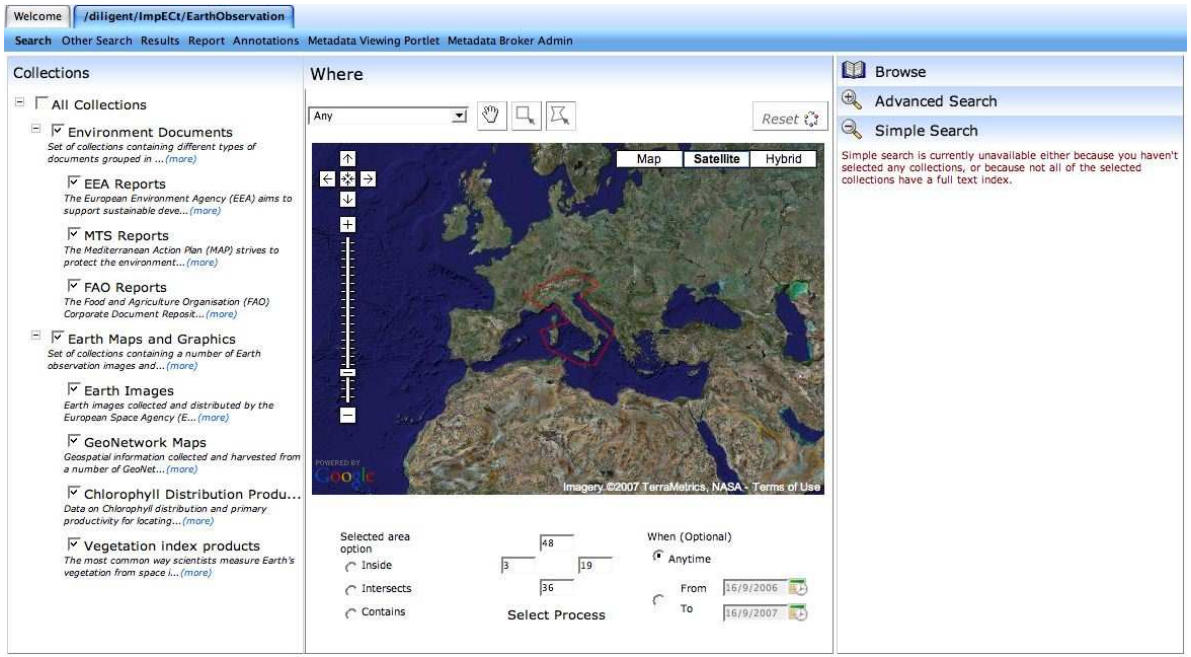
from

  - ESA EO web portal (`www.eoportal.org`)
  - ESA Grid on-demand (`eogrid.esrin.esa.int`)
  - Geonetwork (`www.fao.org/geonetwork`)
  - Web Map Servers
  - European Environment Agency (`www.eea.eu.int`)
  - NASA CEOS IDN (`idn.ceos.org`)
  - Medspiration (`www.mespiration.org/products`)

Search Facilities

By relying on the DILIGENT search framework the IMPECT heterogeneous data sources are seamlessly searchable

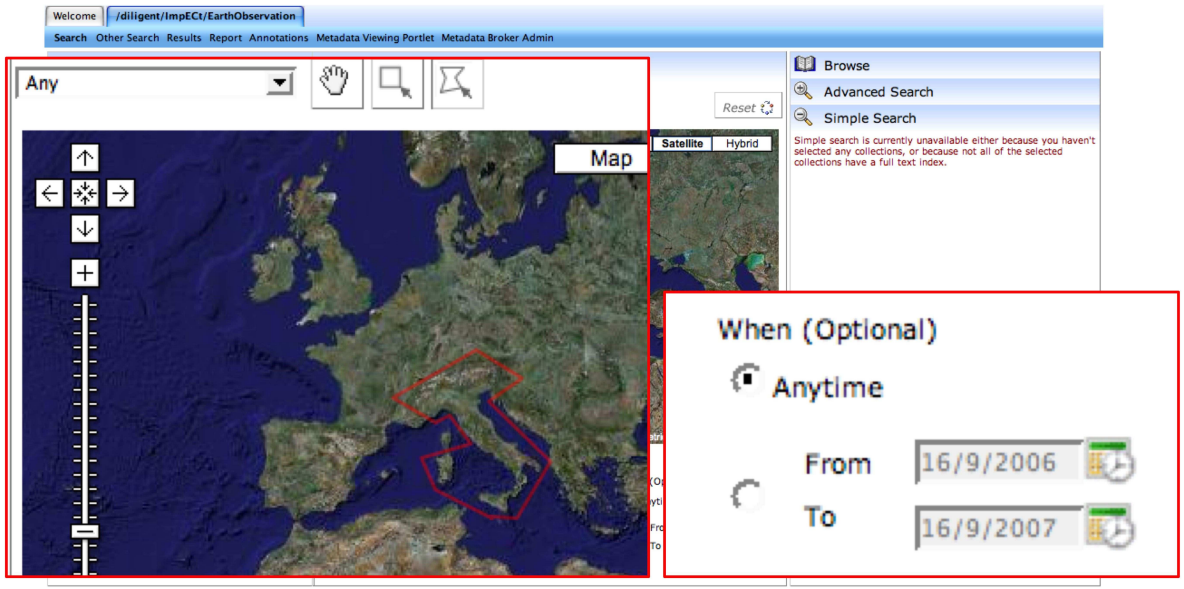
- TemporalRefiner and PolygonalRefiner refinement operators allow the user to refine the search results based on timestamp or polygonal shape



Search Facilities

By relying on the DILIGENT search framework the IMPECT heterogeneous data sources are seamlessly searchable

- TemporalRefiner and PolygonalRefiner refinement operators allow the user to refine the search results based on timestamp or polygonal shape



Summary

e-Science scenarios demand for infrastructure-oriented approaches to guarantee low-costs and sustainability

- the higher initial development cost than traditional ad-hoc solutions is well repaid by the long-term scale of adoption and maintenance

The gCube search framework:

- is equipped with a comprehensive set of state-of-the-art IR algorithms and techniques that can be easily reused
- has low adoption cost because the whole service is actually outsourced to the underlying infrastructure
- is open thus to guarantee the easy of use/adaptation in unexpected scenarios

Summary

e-Science scenarios demand for infrastructure-oriented approaches to guarantee low-costs and sustainability

- the higher initial development cost than traditional ad-hoc solutions is well repaid by the long-term scale of adoption and maintenance

The gCube search framework:

- is equipped with a comprehensive set of state-of-the-art IR algorithms and techniques that can be easily reused
- has low adoption cost because the whole service is actually outsourced to the underlying infrastructure
- is open thus to guarantee the easy of use/adaptation in unexpected scenarios

`http://www.diligentproject.org`
`http://www.gcube-system.org`