AVALON
Algorithms and Software Architectures for Distributed & High Performance Computing Platforms

Christian Perez
LIP, ENS Lyon

2014, September 18

Agenda

Team Members

Avalon Research Activities

Overview of Some Research Activities

- Measuring and Modeling Energy Consumptions
- Scientific Applications and multi-Clouds
- Modeling Scientific Applications With Software Components
- Large Scale Data management
- Two European Projects

Conclusion

Avalon Members @ August 1st, 2014

Faculty Members (8)
(4 INRIA, 1 CNRS, 2 UCBL, 1 ENSL)

- Eddy Caron, MCF ENS Lyon, HDR (80%)
- Frédéric Desprez, DR INRIA, HDR (30%)
- Gilles Fedak, CR INRIA
- Jean-Patrick Gelas, MCF UCBL
- Olivier Glück, MCF UCBL
- Laurent Lefèvre, CR INRIA, HDR
- Christian Perez, DR INRIA, HDR, Project leader
- Frédéric Suter, CR CNRS

PhD students (7)

- Maurice-Djibril Faye, ENS-Lyon / Université Gaston Berger (Sénégal)
- Sylvain Gault, MapReduce, INRIA
- Anthony Simonet, MapReduce, INRIA
- Vincent Lanore, ENSL
- Arnaud Lefray, SEED4C, ENSIB
- Daniel Balouek, CIFRE New Generation SR
- Violaine Villebonnet, INRIA

Engineers (3+4+1)

- Simon Delamare, IR CNRS (80%)
- Jean-Christophe Mignot, IR CNRS (20%)
- Matthieu Imbert, INRIA SED (40%)

- François Rossigneux, XCLOUD
- Guillaume Verger, SEED4C
- Yulin Zhang Huaxi, SEED4C

- *Laurent Pouilloux (IPL Héméra)*

Postdoc / Temporary Researcher

- Jonathan Rouzaud-Cornabas, CNRS
- Marcos Dias de Asuncao, Inria

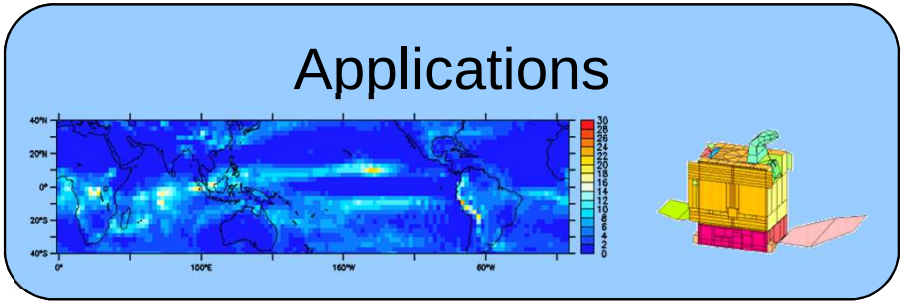
Temporary Teacher-Researcher

- Ghislain Landry Tsafack, UCBL

Assistant

- Evelyne Blesle, INRIA

Avalon: Research Activities



?

CPU/data-intensive Scientific Applications

- From “simple” to code coupling
  - Structure complexity
  - “New” forms of interactions (MR)

Computing platforms

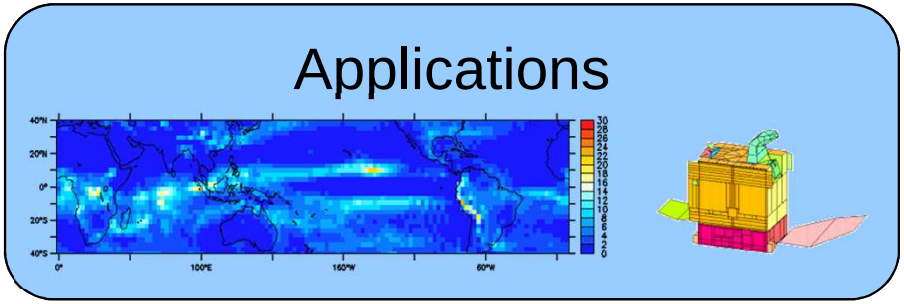
- Different characteristics
  - Performance, energy, size, cost, reliability, QoS, etc.
- Hybridization
  - Sky computing, HPC@Cloud, Exascale, Spot instance

Objectives

- Expressiveness simplicity
- Application portability
- Resource specific optimizations
  - Elastic resource management
  - Energy consumption

Super-computers (Exascale)	Grids (EGI)	Desktop Grids	Clouds (IaaS, PaaS)
Large scale	Heterogeneity	Volatility	On demand

Avalon: Research Activities



Programming Abstractions

Algorithmics

Application & Resource Models

Resource Abstractions

CPU/data-intensive Scientific Applications

- From "simple" to code coupling
  - Structure complexity
  - "New" forms of interactions (MR)

Computing platforms

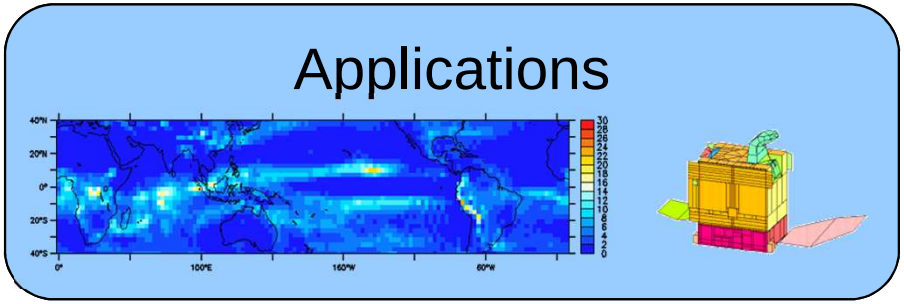
- Different characteristics
  - Performance, energy, size, cost, reliability, QoS, etc.
- Hybridization
  - Sky computing, HPC@Cloud, Exascale, Spot instance

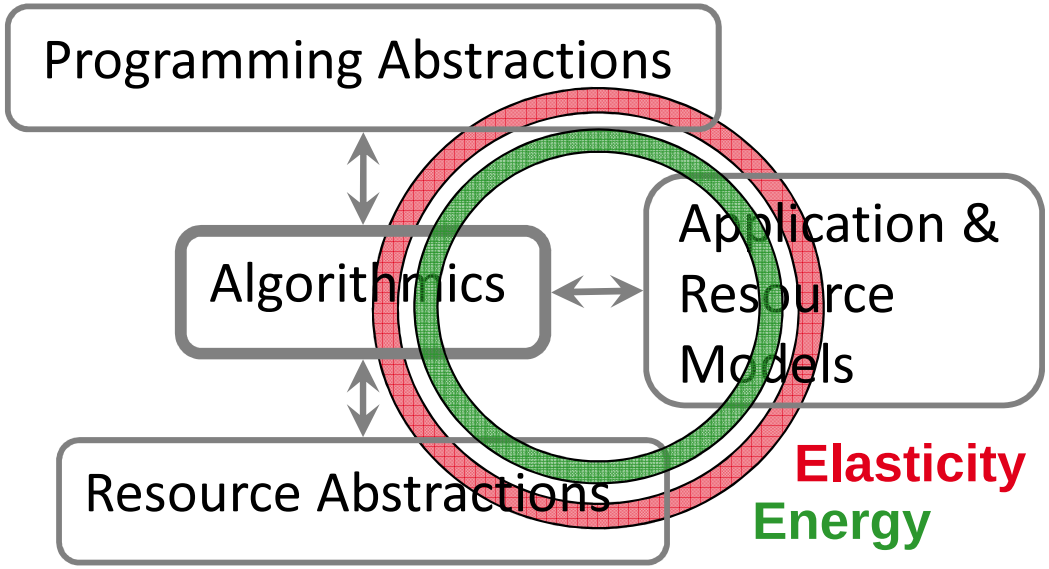
Objectives

- Expressiveness simplicity
- Application portability
- Resource specific optimizations
  - Elastic resource management
  - Energy consumption

Super-computers (Exascale)	Grids (EGI)	Desktop Grids	Clouds (IaaS, PaaS)
Large scale	Heterogeneity	Volatility	On demand

Avalon: Research Activities





CPU/data-intensive Scientific Applications

- From "simple" to code coupling
  - Structure complexity
  - "New" forms of interactions (MR)

Computing platforms

- Different characteristics
  - Performance, energy, size, cost, reliability, QoS, etc.
- Hybridization
  - Sky computing, HPC@Cloud, Exascale, Spot instance

Objectives

- Expressiveness simplicity
- Application portability
- Resource specific optimizations
  - Elastic resource management
  - Energy consumption

Super-computers (Exascale)	Grids (EGI)	Desktop Grids	Clouds (IaaS, PaaS)
Large scale	Heterogeneity	Volatility	On demand

Avalon: Four Research Axes

Energy Application Profiling and Modeling

J.-P. Gelas, O. Glück, L. Lefèvre, J.-C. Mignot

- Large Scale Energy Consumption Analysis for Physical and Virtual Resources
- Energy Efficiency of Next Generation Large Scale Platforms

Data-intensive Application Profiling, Modeling, and Management

F. Desprez, G. Fedak, F. Suter,

- Performance Prediction of Parallel Regular Applications
- Modeling Large Scale Storage Infrastructure
- Data Management for Hybrid Computing Infrastructures

Resource Agnostic Application Description Model

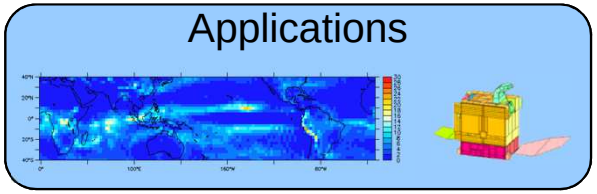
E. Caron, L. Lefèvre, C. Pérez

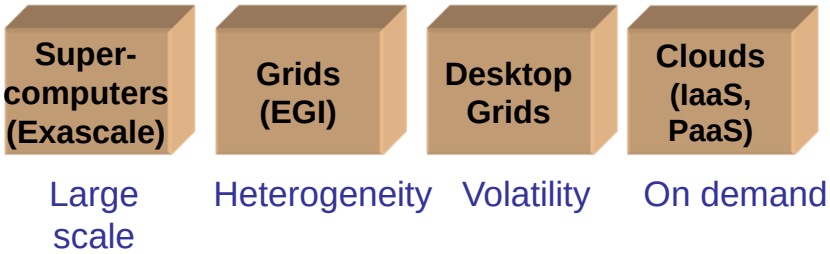
- Moldable Application Description Model
- Dynamic Adaptation of the Application Structure

Application Mapping and Scheduling

E. Caron, F. Desprez, L. Lefèvre, C. Pérez, F. Suter

- Application Mapping and Software Deployment
- Non-Deterministic Workflow Scheduling
- Security Management in Cloud Infrastructure

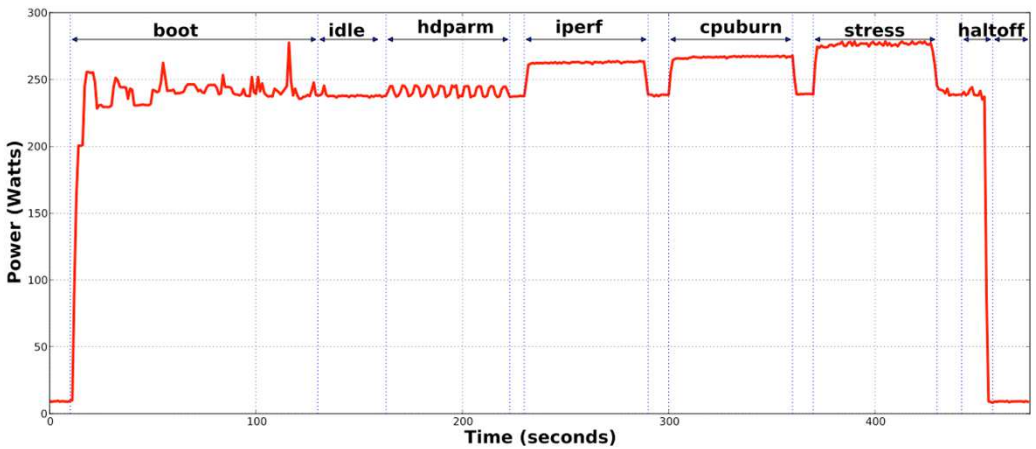


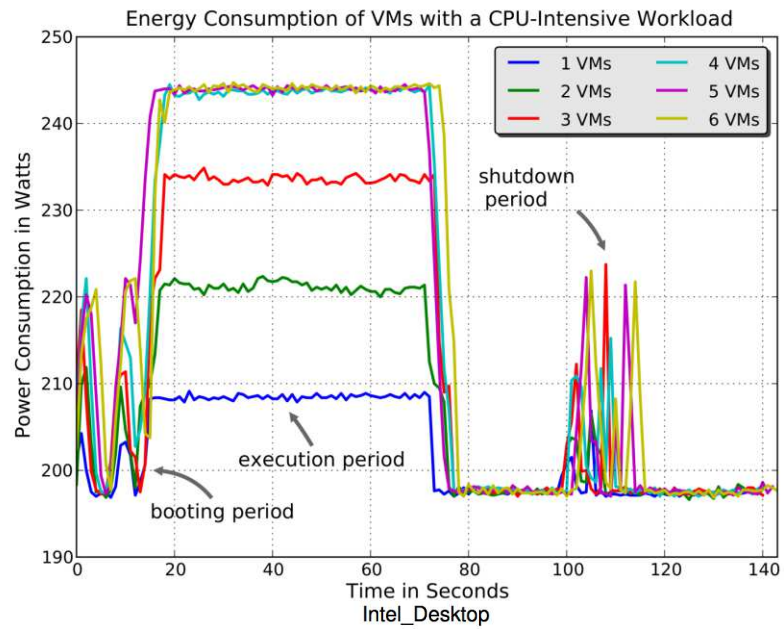


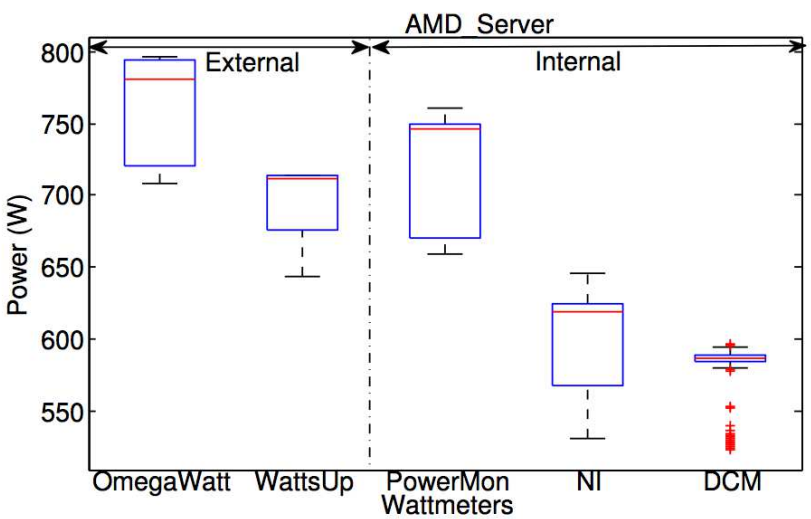
Measuring and Modeling Energy Consumptions

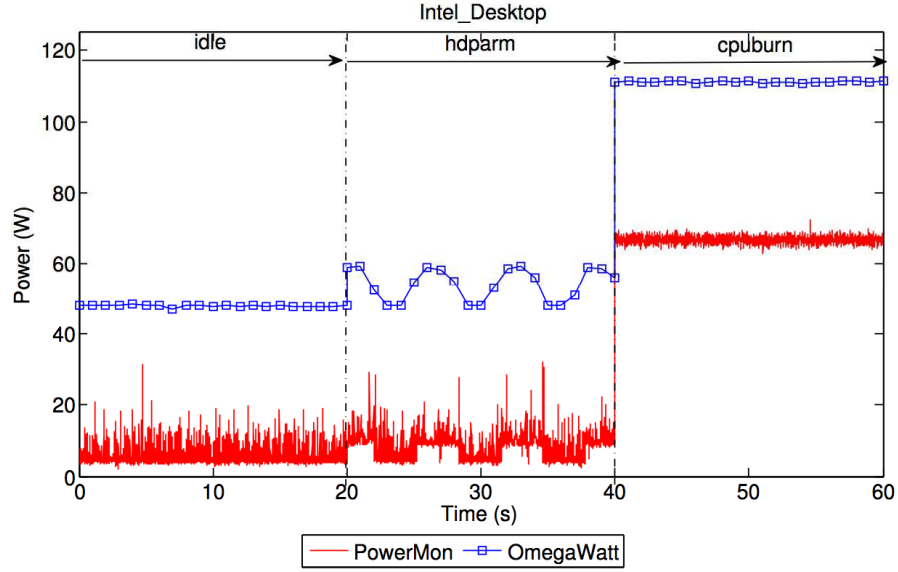
L. Lefevre, J.-P. Gelas, O. Gluck,
M. Diouri, G. Tsafack, A.-C. Orgerie, J.-C. Mignot

Profiling and Understanding Energy Consumption of Real Applications







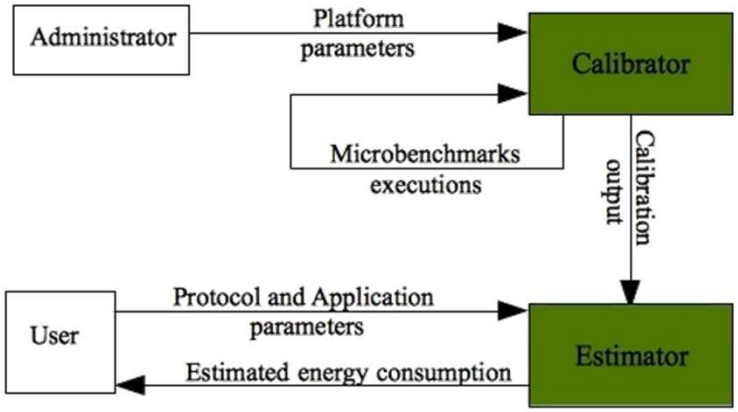


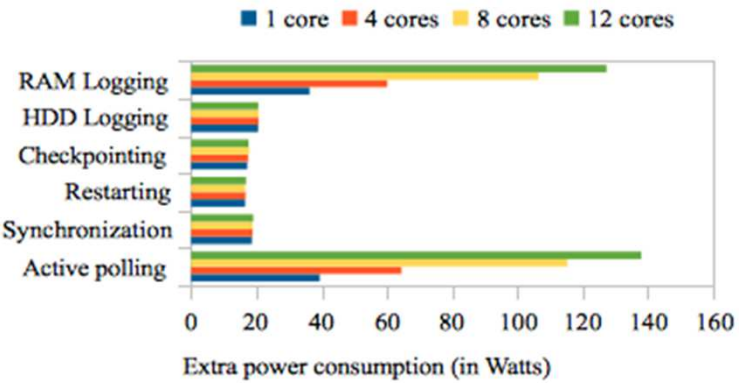
Energy Efficient Software in HPC

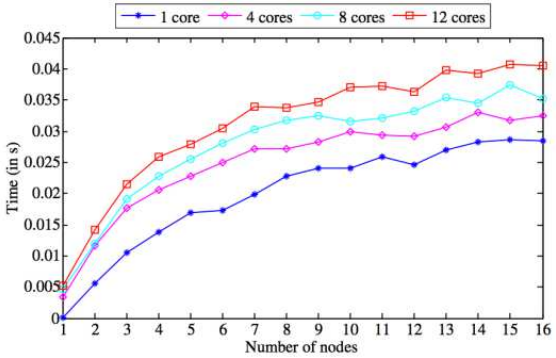
Two focus: fault tolerance and data broadcast

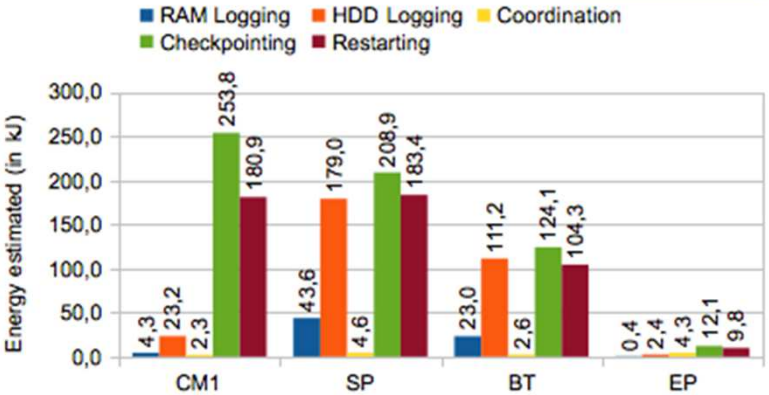
Help users to choose the best service

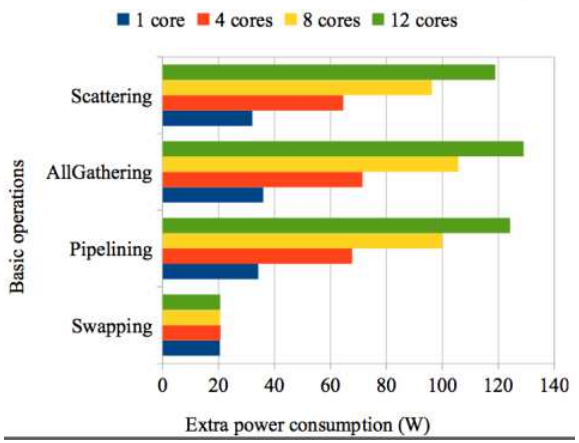
Applications on exascale infrastructures

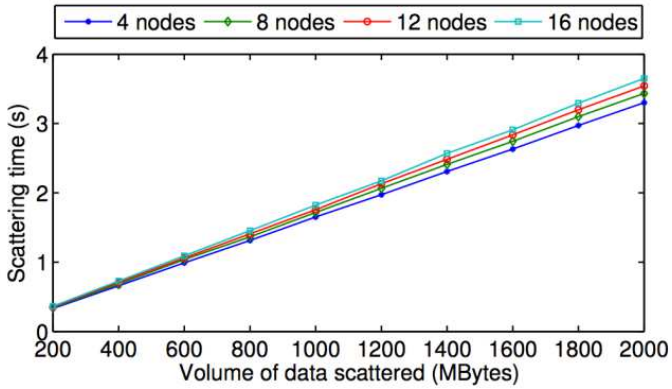


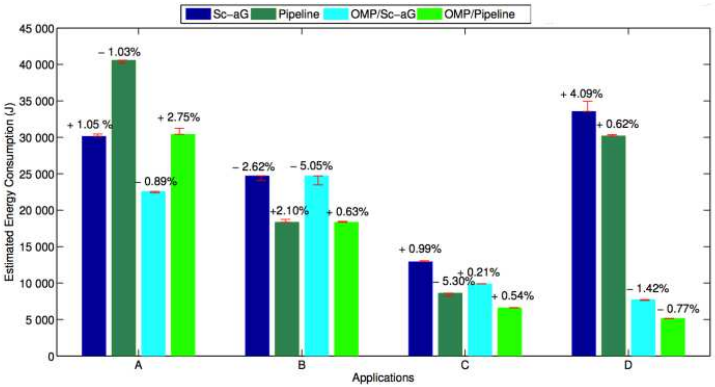












M. Diouri, Olivier Glück, Laurent Lefevre, and Franck Cappello. **"ECOFIT: A Framework to Estimate Energy Consumption of Fault Tolerance Protocols during HPC executions"**, *CCGrid2013, the 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, Delft, the Netherlands, May 13-16, 2013

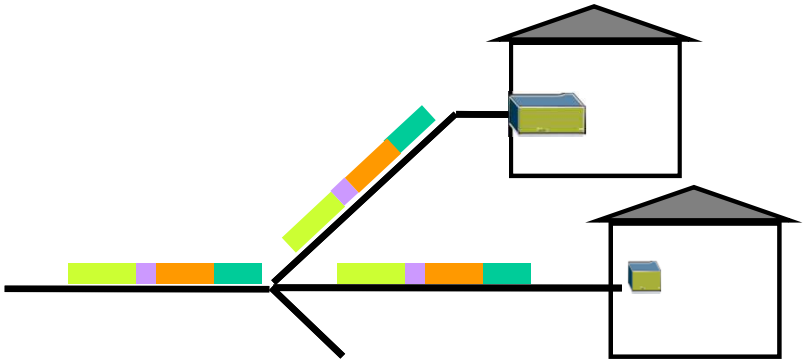
Joint Laboratory for Petascale Computing

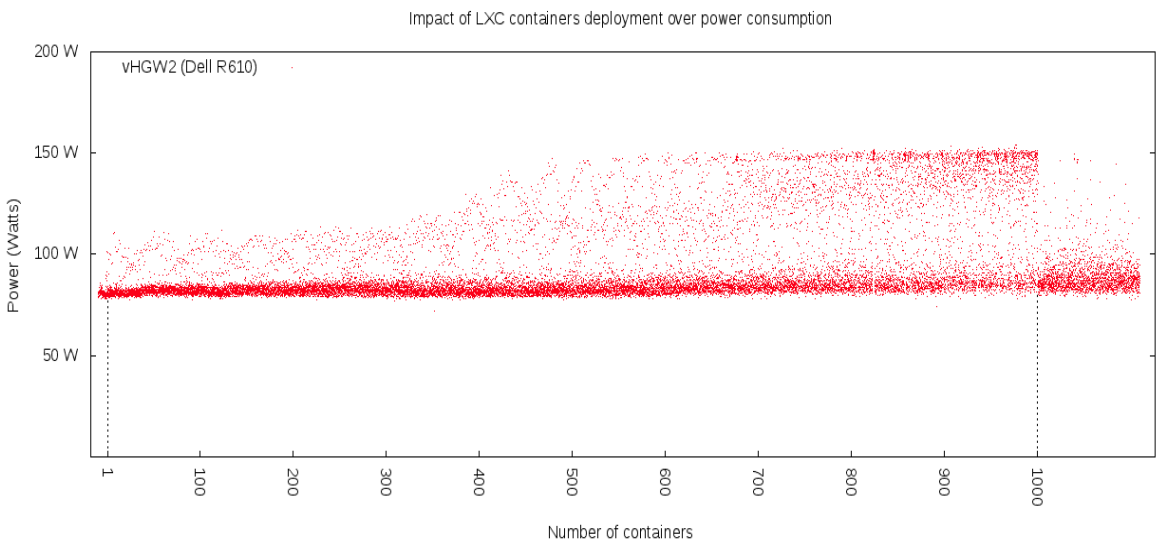
Virtual Home Gateway (vHGW)

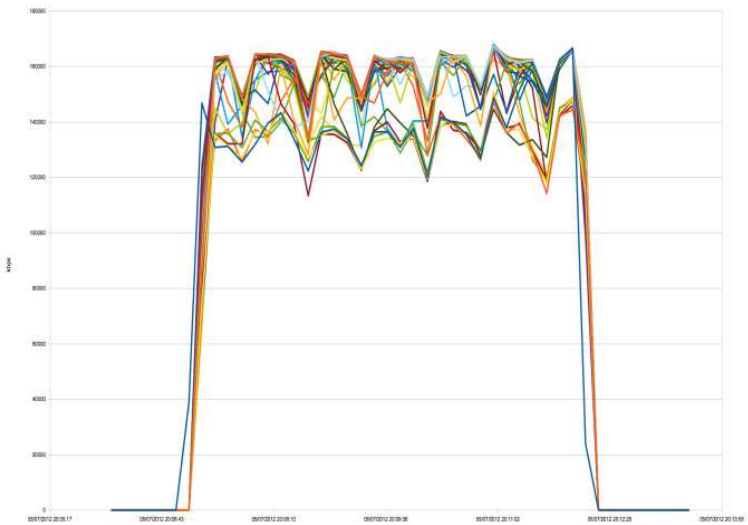
Within GreenTouch project (1000 factor)

Virtualizing home gateway services to reduce energy consumption at the last mile

- Combining with quasi passive CPE
- Taking care of Quality of Service
- Evaluating energy usage reduction
- Studying consolidation effects





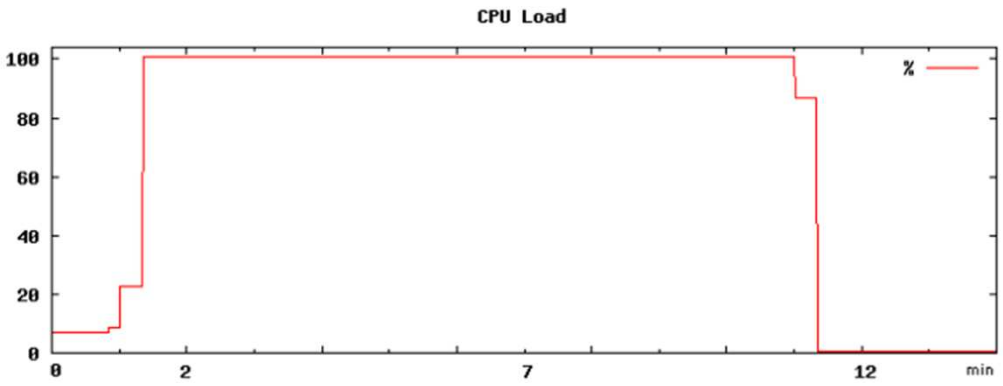


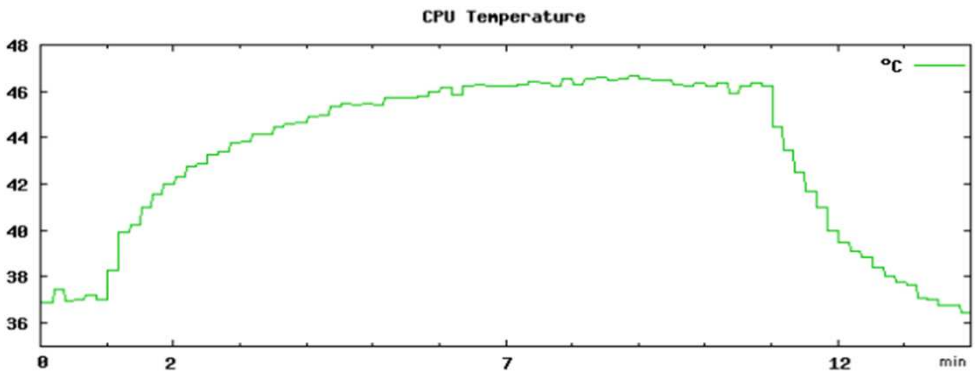
DataCenter (1/2)

Energy-aware layer for DC automation with direct knowledge of resources

- Smart allocation of tasks (consolidation)
- Dynamic profiling of the hardware
- Smart management of resources (on/off)

Challenge: Align supply with demand on-the-fly by using the power/energy data as input information for central software to perform actions

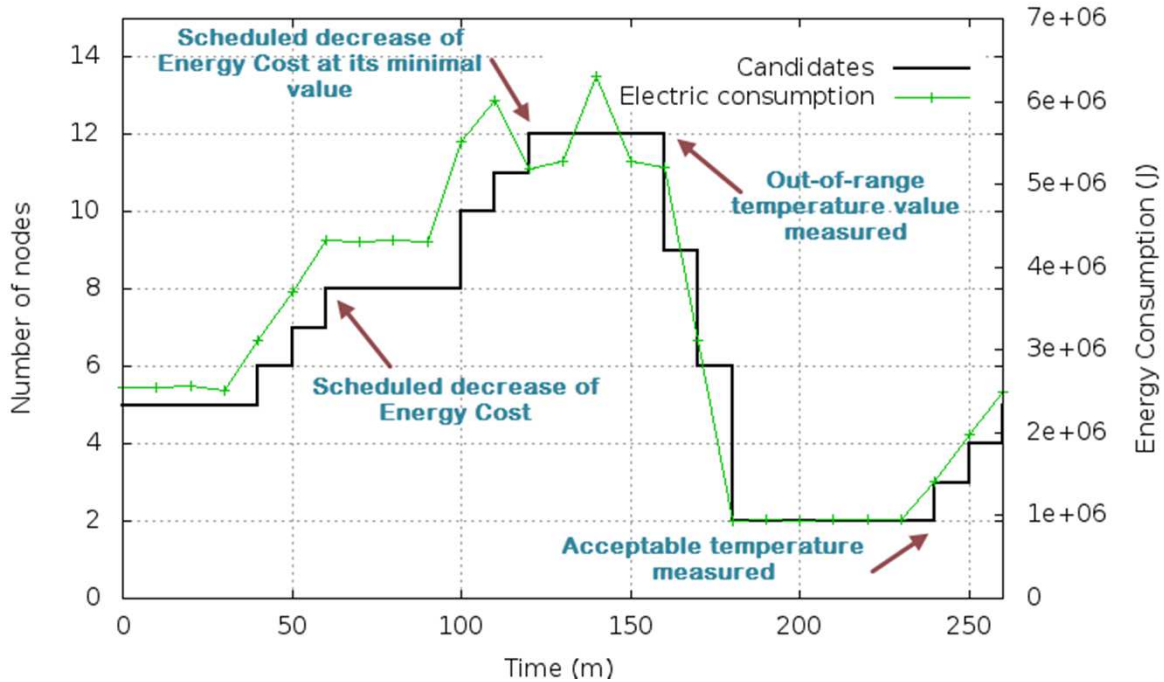




Most of the operations costs is dedicated to cooling

DataCenter (2/2)

Real-life experiments on the Grid'5000 platform on 1000+ jobs



Regulation of the infrastructure power consumption

- Schedule of energy provider
- Local conditions of temperature
- Exploitation incidents

Up to 25% of energy saving with minimal performance degradation

Power Measurement @ Grid'5000 [Hemera/G5K]

What users need

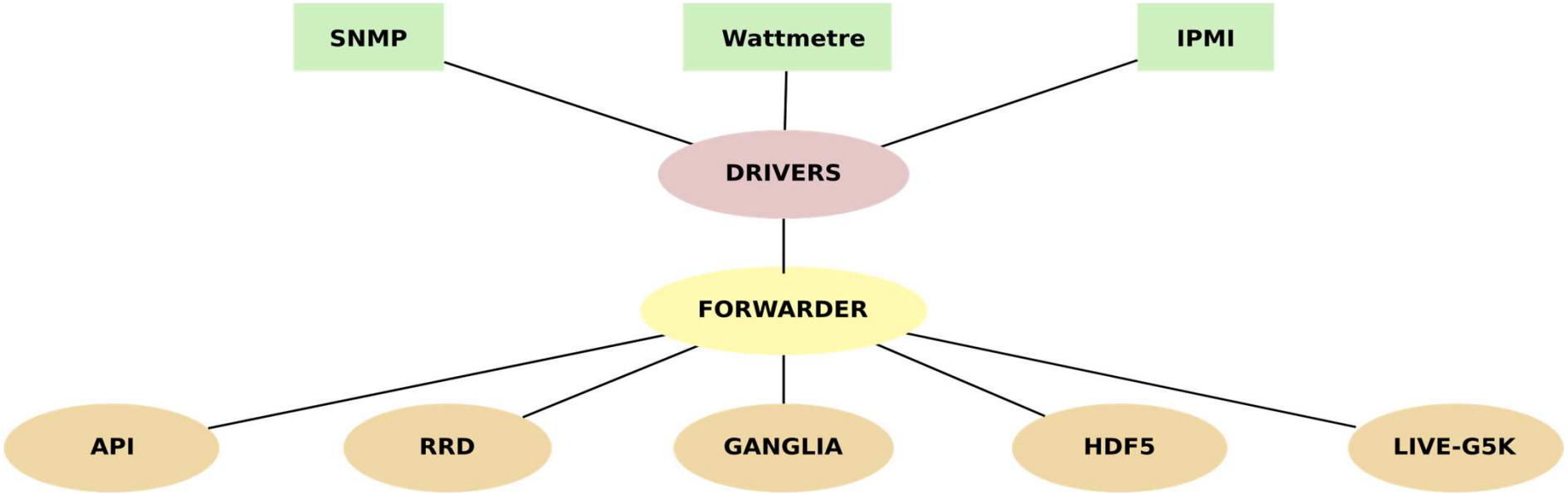
- Live visualization of the experiment
- Use instantaneous power consumption in your application
- Access data post mortem

Only available on Lyon site

Kwapi Architecture (Soon in production) [Hemera/Grid'5000]

Uniformization with one VM by site

- API : allow to retrieve instantaneous data
- RRD : store data (but temporal resolution decrease with time)
- GANGLIA : push the data on the Grid'5000 supervision service
- HDF5 : store data without and provide an interface to retrieve post-mortem data
- LIVE : allow to follow in live your experiment



Kwaapi: User Tools

Available on Lyon, Rennes, Nancy, Reims

Real-time data

- curl energy.<site>:5000/probes

Live visualization

- https://intranet.grid5000.fr/supervision/<site>/energy/last/minute/

Post mortem data retrieval

- curl http://energy.<site>:12000/timeseries/?job_id=XXXXXX

More information

- https://www.grid5000.fr/mediawiki/index.php/Kwapi

Mapping (Scientific) Applications onto multi-Clouds

J. Rouzaud-Cornabas, F. Desprez, C. Perez, E. Caron, A. Lefray

Scientific Applications and multi-Clouds

Emergence of data-intensive science and Big Data

Still a lot of heavy computing

Tightly coupled applications (e.g. MPI)

- Executed on supercomputers
- Performance issues on Clouds
- 10-20% of scientific applications

Loosely coupled applications: Bag Of Tasks and Workflows

- Not suitable for supercomputers
- 80-90% of scientific applications
- Increasing number (and domains) of applications
- Dramatic increase of the quantity of data and compute

Using federated Clouds to run these applications

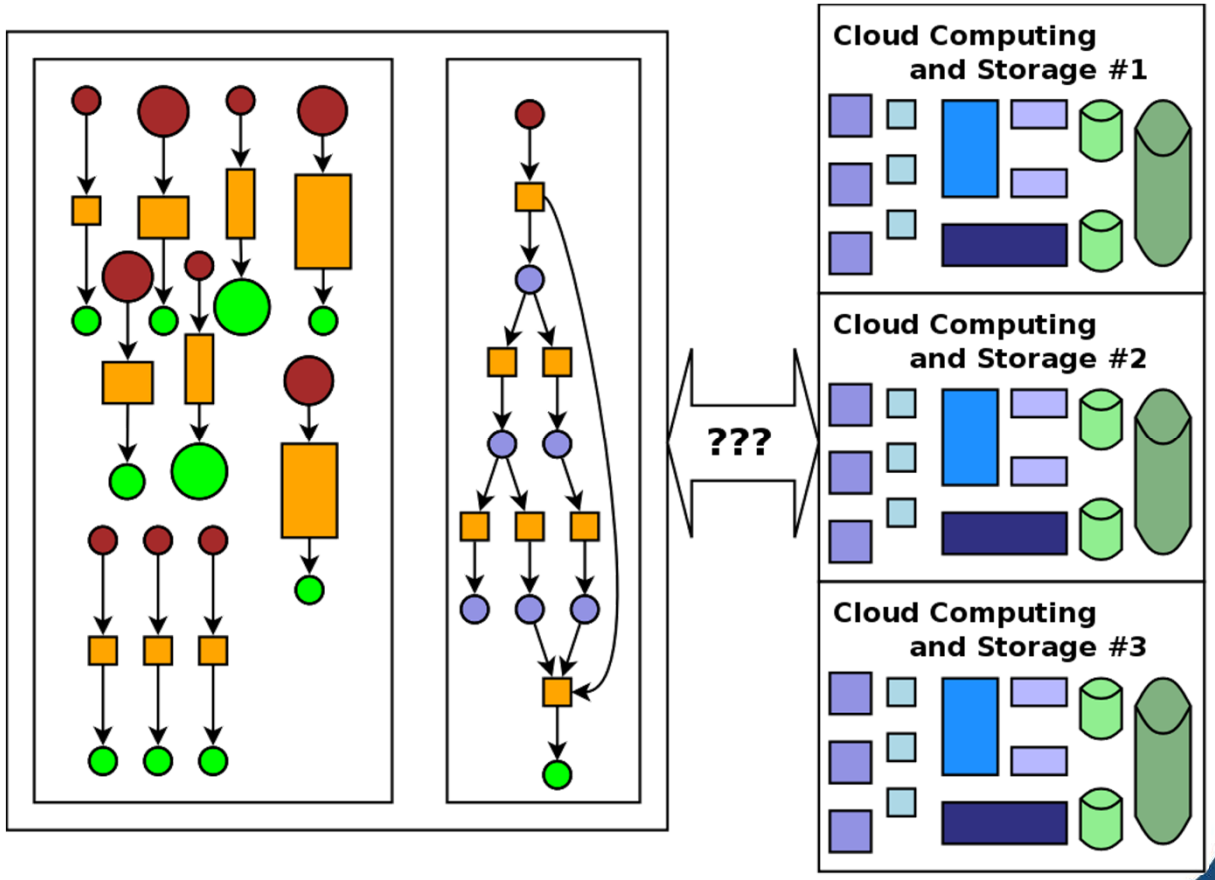
Application and multi-Clouds

Two steps

- Provisioning Virtual Machines
- Scheduling tasks in Virtual Machines

Related Work

- Only taking into account processor speed
- Homogenous, static, and reliable resources
- Do not take into account data



Application Model: Bag Of Tasks

x Tasks and *no* dependency between them but *a large number of parameters*

Three parameters (I, O and FLOPS) for tasks in BoT (impact task allocations)

- Homogenous
- Stochastic (uniform/bimodal/heavytail)

Different task arrival (impact on provisioning) models

- At the beginning
- Poisson
- Dependency and think time

Different objectives

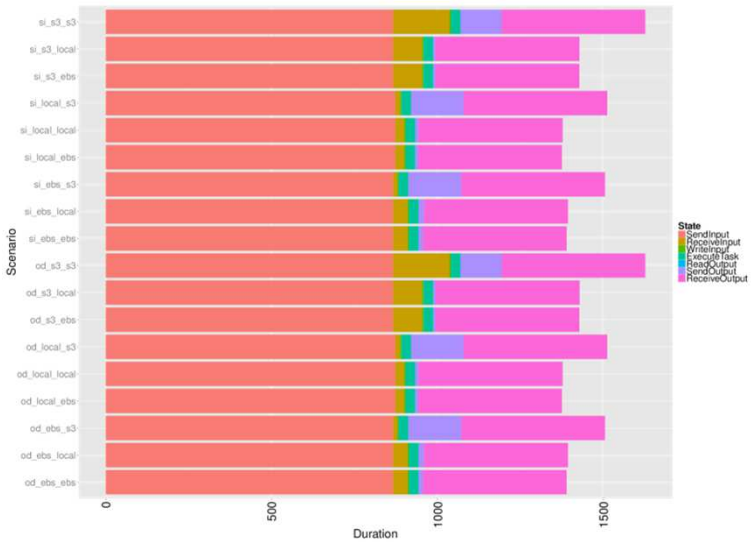
- Cost
- Performance
- Deadline
- Etc.

SimGrid Cloud Broker

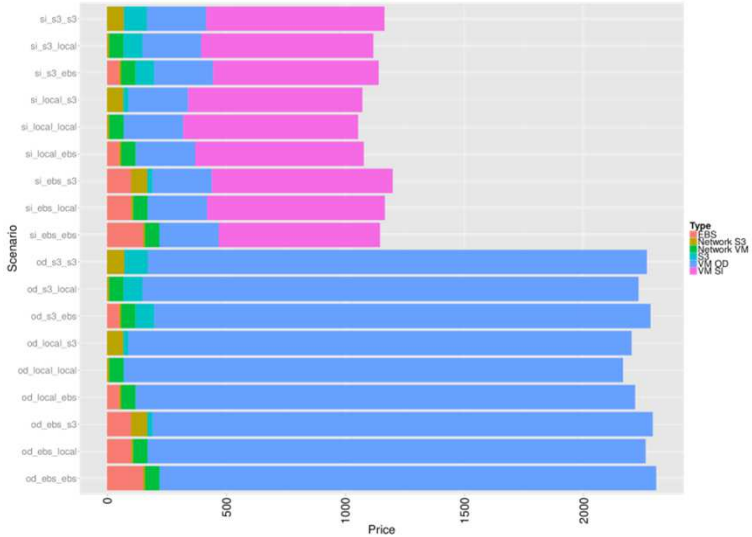
Multi-clouds environment (Not only EC2/S3 API) (ANR SONGS project)

- Can simulate any public or private Cloud (need to implement performance models)
- All AWS implemented
  - All AWS regions, instance types (resources and prices), On-demand and spot instances, S3 and EBS Storage, Accounting of resources usage (Network, Compute, Storage), Spot Instances (random, file, model dynamic price policies)
  - Resources performance models based on information given by Amazon and extracted from scientific papers

Examples

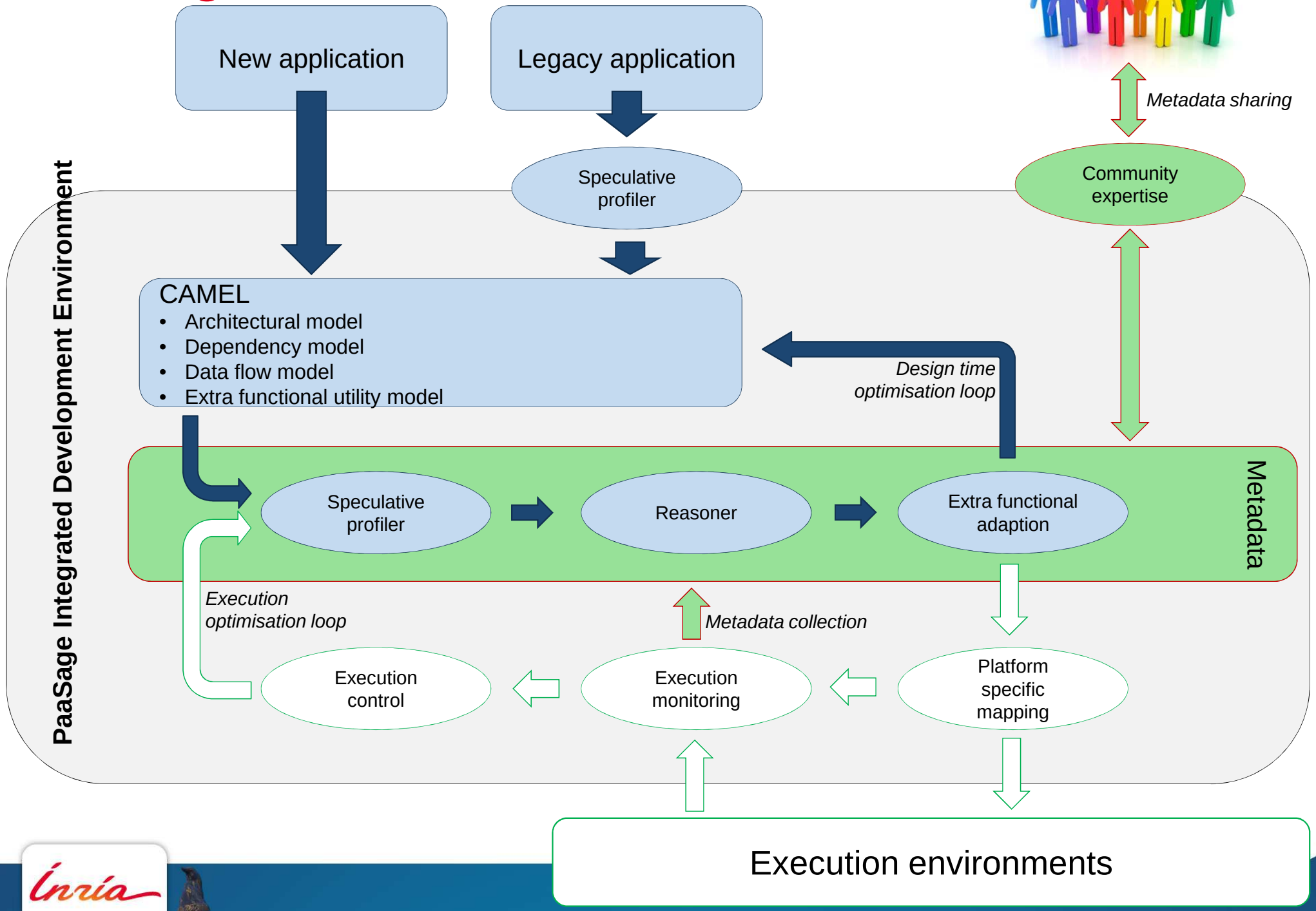


Impact of Storage Policies on Completion Time



Impact of Storage Policies on Billing

PaaSage Overview



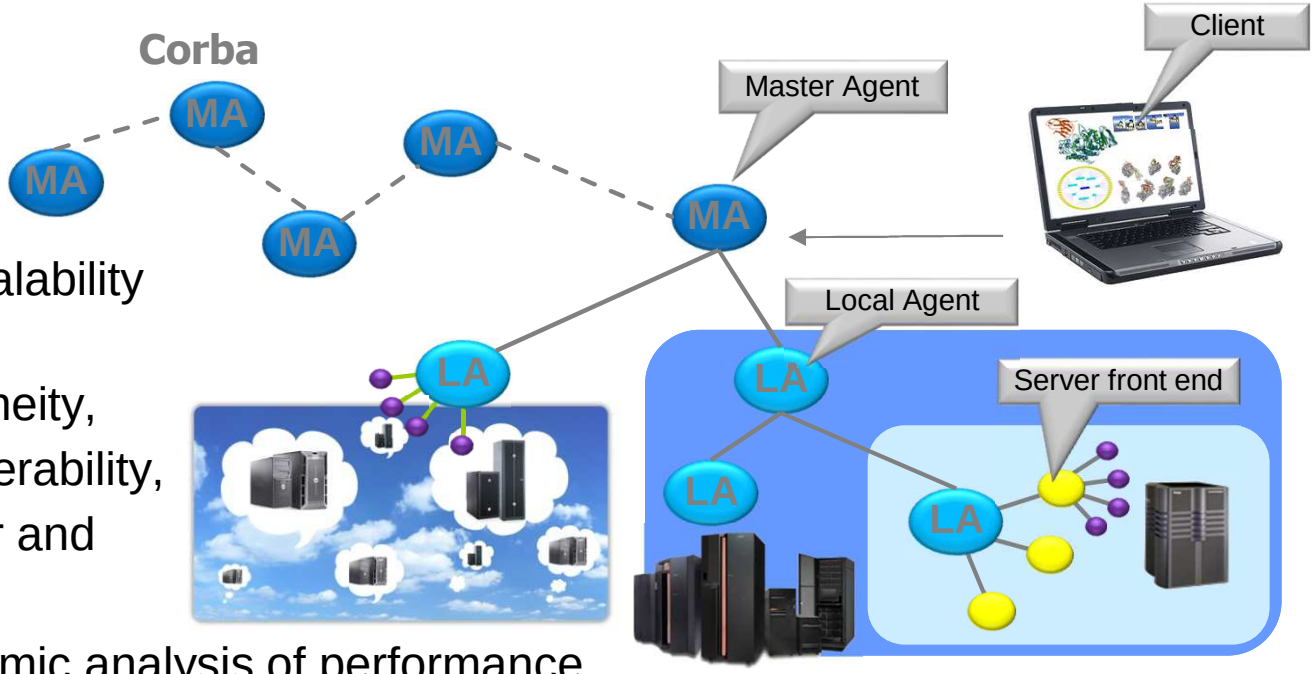
DIET as a Multi-Cloud Manager with security support

Context

- Development of a toolbox for deploying application services providers with a hierarchical architecture for scalability

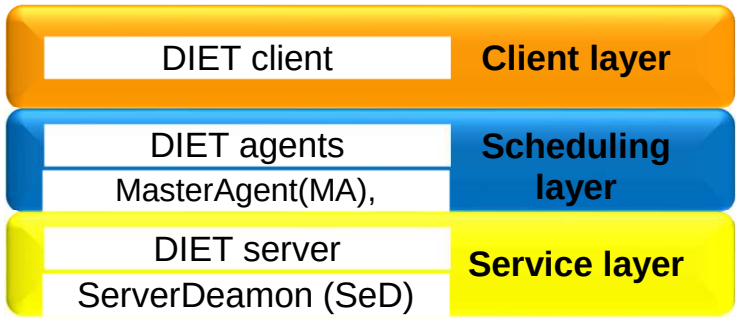
Main Research Issue

- security, scheduling, heterogeneity, automatic deployment, interoperability, high performance data transfer and management, monitoring, fault tolerance, static and dynamic analysis of performance, …



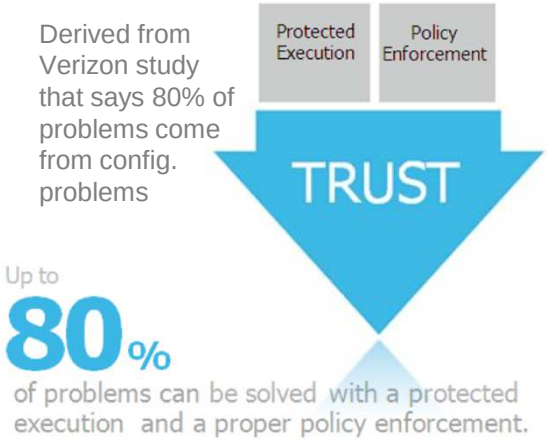
http://graal.ens-lyon.fr/DIET

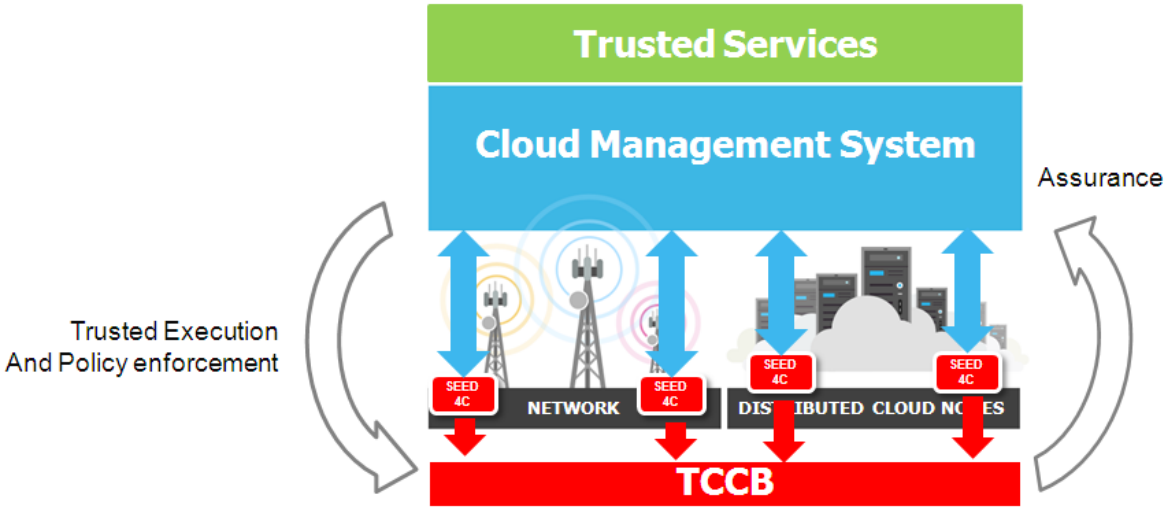
Validation: Large validation over Grid'5000.
DIET used case: The Decrypthon project - DIET was selected by IBM -
Collaborations: Celtic+, RNTL GASP, ACI GRID ASP, TLSE, ACI MD GDS, ANR LEGO, ANR GWENDIA, Grid'5000
Start'up: SysFera (created in march 2010).
Contact: Eddy.Caron@ens-lyon.fr
Avalon Team, Inria, LIP ENS Lyon
Web: http://graal.ens-lyon.fr/DIET



Seed4C Project motivation

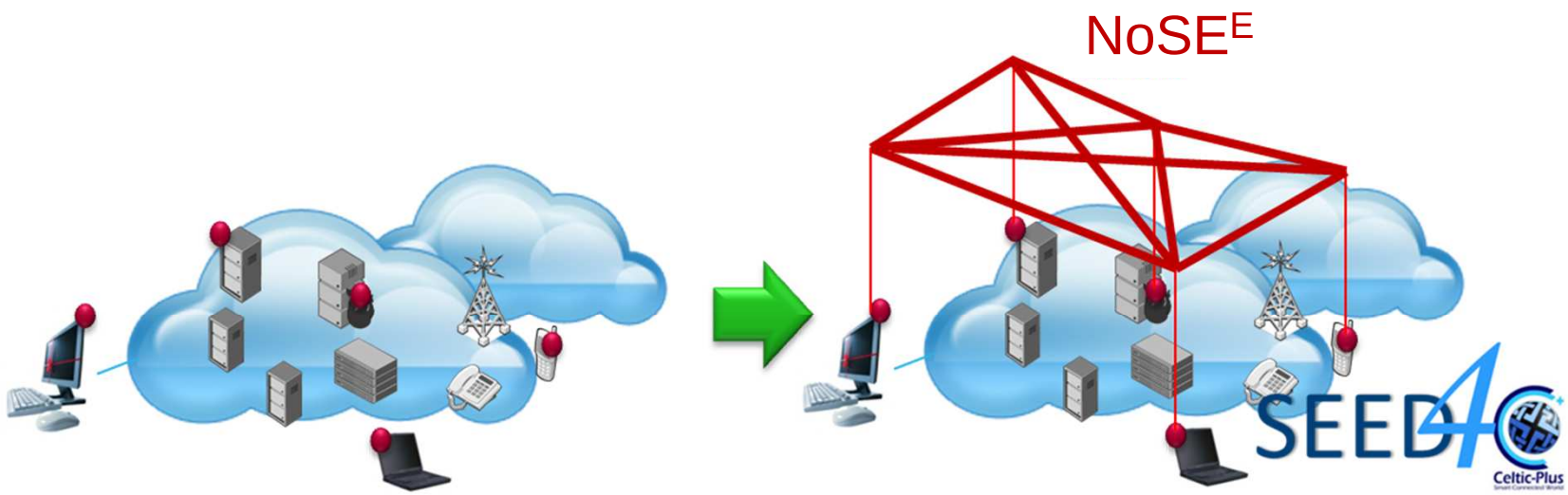
- Secure Embedded Element and Data protection for Cloud
- Can we get a Seed to build trusted Clouds?
- One that transforms the way we trust Cloud based Services
- Building a Trusted Cloud Computing Base (TCCB)
- A Cloud of minimal Trusted Computing Bases: the SEEDs.





Seed4C Project (con't)

- From isolated Security to coordinated Security



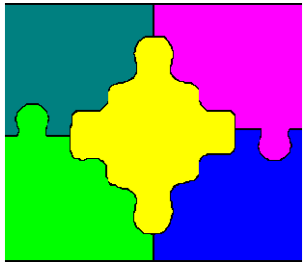
Isolated Security
by secure elements or not

Coordinated Security
by Network of Secure Elements Extended

Modeling Scientific Applications With Software Components

C. Perez, J. Bigot

Software Component



Technology that advocates for composition

- Old idea (late 60's)
- *Assembling* rather than *developing*

Many models

- Salome, CCA, CCM, Fractal, OGSi, SCA, …

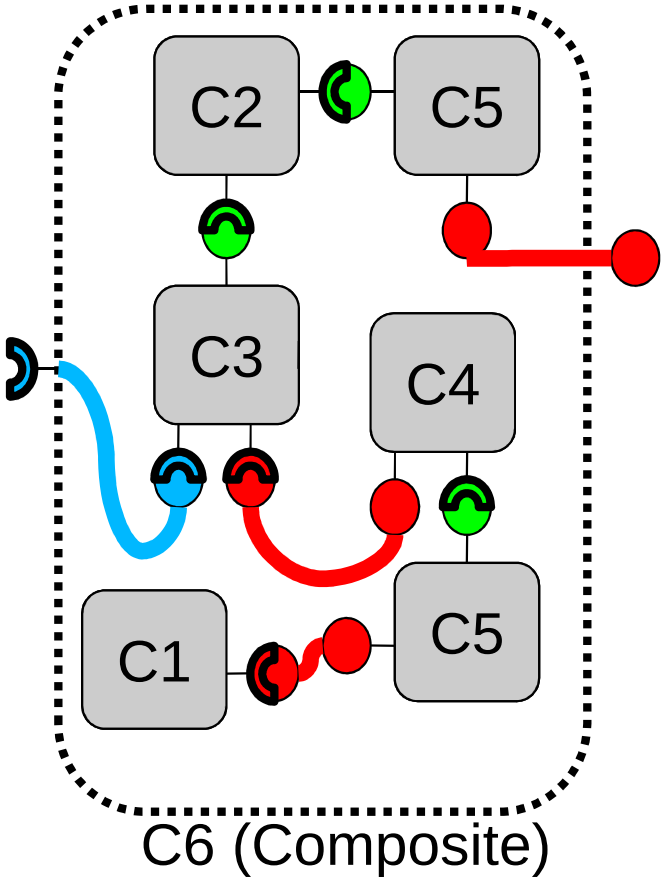
Pre-defined set of interactions

- Usually function/method invocation oriented

Provide communication abstraction

- (Limited) Language interoperability (~IDL)
- Network transparency (overhead?)

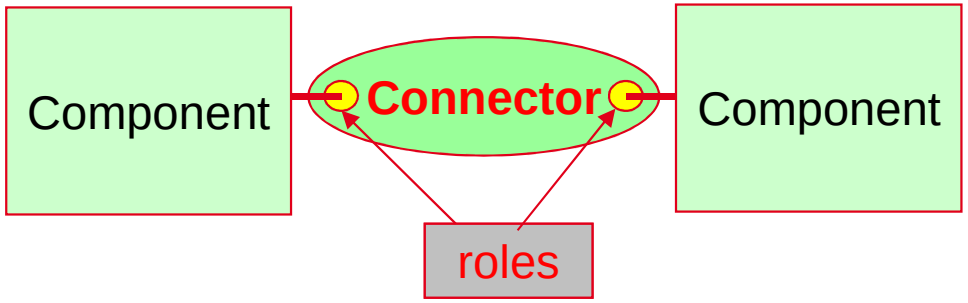
➢ ***Programming model vs execution model***



HLCM: High Level Component Model

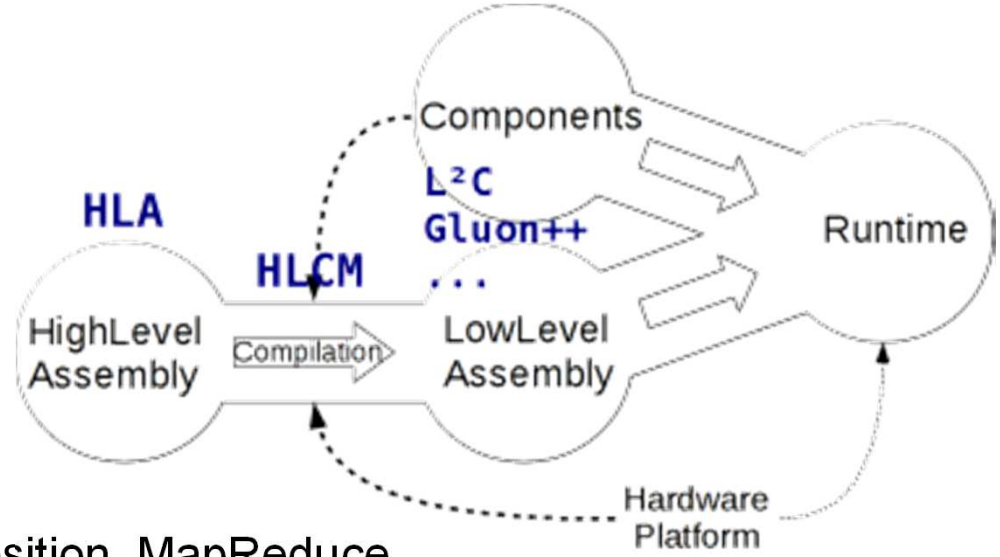
Major concepts

- Component model (hierarchical)
  - Primitive and composite
- Connector based
  - Primitive and composite
- Generic model
  - Support meta-programming (template *à la* C++)
- Currently static



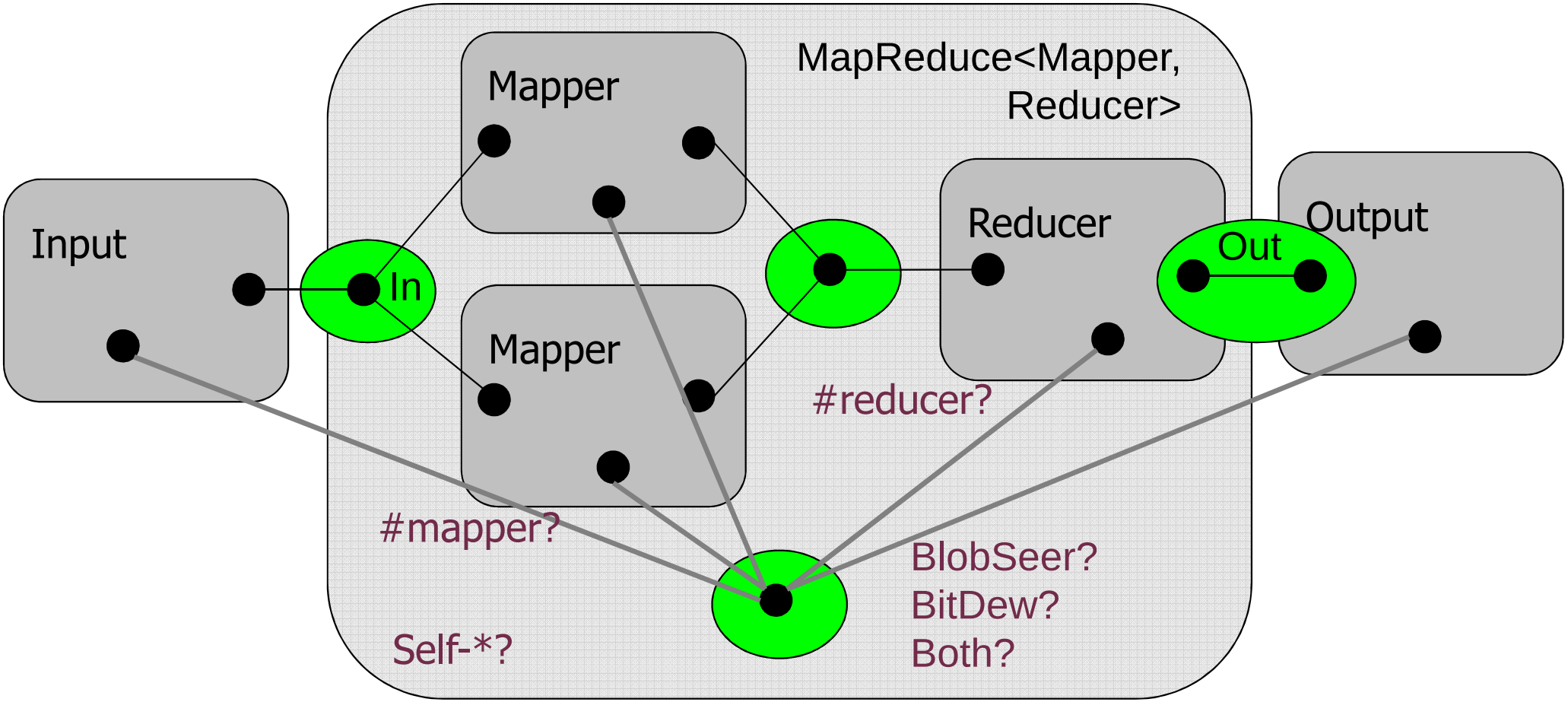
HLCMi: an implementation of HLCM

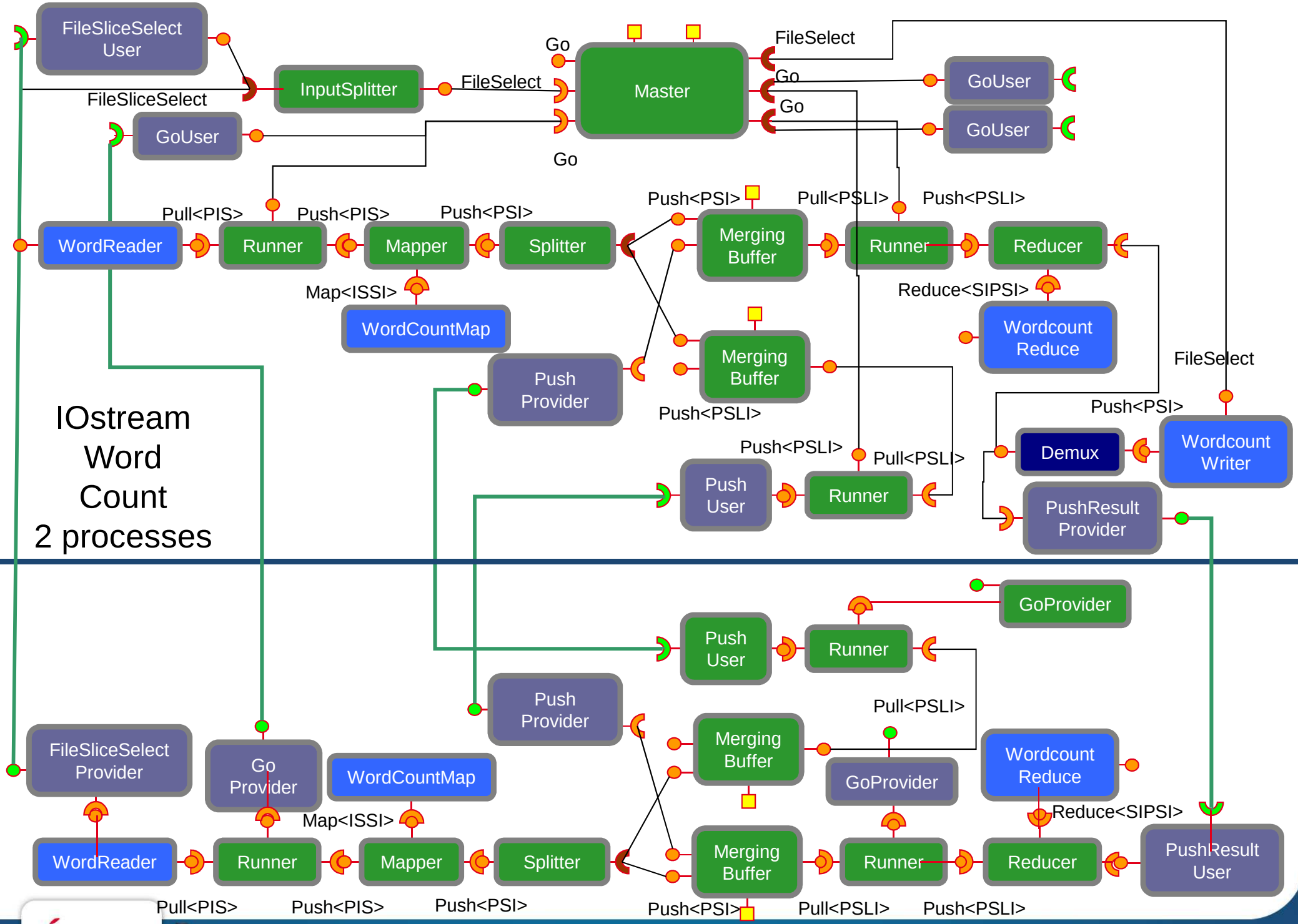
- Model-transformation based (EMF)
- Connectors
  - Use/Provide
  - Shared Data
  - Collective Communications
  - MxN
  - Some skeletons
    - Replication, Simple Domain Decomposition, MapReduce

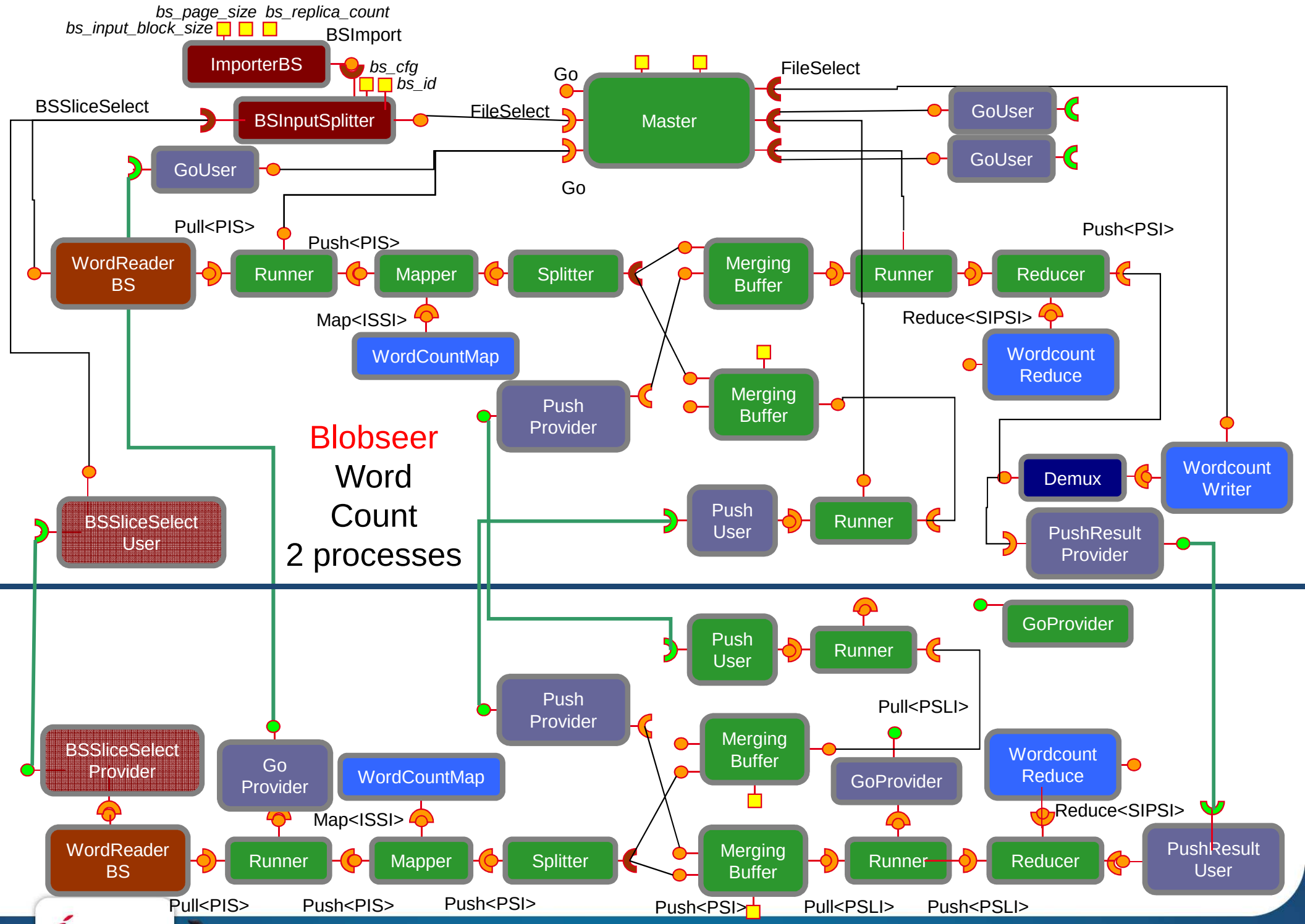


MapReduce Skeleton in HLCM

Component MapReduce<Component Map, Component Reduce>
exposes { In, Out }




FileSliceSelect User
InputSplitter
FileSliceSelect
FileSelect
Go
Master
FileSelect
Go
Go
GoUser
GoUser
GoUser
Go
WordReader
Pull<PIS>
Runner
Push<PIS>
Mapper
Push<PSI>
Splitter
Push<PSI>
Merging Buffer
Pull<PSLI>
Push<PSLI>
Runner
Reducer
Map<ISSI>
WordCountMap
Reduce<SIPSI>
Wordcount Reduce
Merging Buffer
Push Provider
Push<PSLI>
FileSelect
Push<PSI>
Wordcount Writer
Demux
Push<PSLI>
Pull<PSLI>
Push User
Runner
PushResult Provider
IOstream Word Count 2 processes
GoProvider
Push User
Runner
Push Provider
Pull<PSLI>
Merging Buffer
GoProvider
Wordcount Reduce
FileSliceSelect Provider
Go Provider
WordCountMap
Map<ISSI>
Reduce<SIPSI>
WordReader
Runner
Mapper
Splitter
Merging Buffer
Runner
Reducer
PushResult User
Pull<PIS>
Push<PIS>
Push<PSI>
Push<PSI>
Pull<PSLI>
Push<PSLI>


bs_page_size bs_replica_count
bs_input_block_size
BSImport
ImporterBS
bs_cfg
bs_id
BSSliceSelect
BSInputSplitter
FileSelect
Go
Master
FileSelect
GoUser
GoUser
GoUser
Go
Pull<PIS>
Push<PIS>
Push<PSI>
WordReader BS
Runner
Mapper
Splitter
Merging Buffer
Runner
Reducer
Map<ISSI>
Reduce<SIPSI>
WordCountMap
Wordcount Reduce
Push Provider
Merging Buffer
Blobseer
Word
Count
2 processes
Demux
Wordcount Writer
BSSliceSelect User
Push User
Runner
PushResult Provider
GoProvider
Push User
Runner
Push Provider
Pull<PSLI>
Merging Buffer
Wordcount Reduce
BSSliceSelect Provider
Go Provider
WordCountMap
GoProvider
Reduce<SIPSI>
Map<ISSI>
WordReader BS
Runner
Mapper
Splitter
Merging Buffer
Runner
Reducer
PushResult User
Pull<PIS>
Push<PIS>
Push<PSI>
Push<PSI>
Pull<PSLI>
Push<PSLI>

Large Scale Data management

G. Fedak, H. He, A. Simonet

BitDew: Large Scale Data Management

Set of services for high level data management on hybrid distributed platforms

- Data scheduler services
  - Steer distribution of data items according to *data distribution abstraction*
  - Affinity
- Multi-protocol and reliable file transfer service
  - Supports legacy protocols (ssh, http, ftp), P2P (bittorrent), Grid protocol (JSAGA, gftp), Cloud (Amazon S3, Dropbox)
- Decentralized data catalog: DHT, DLPT

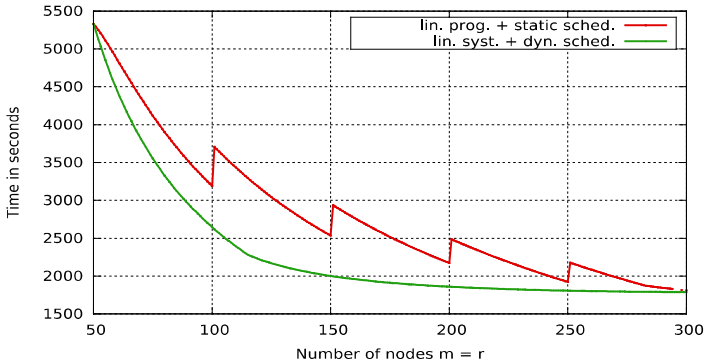
Some Applications Prototyped with Bitdew

- Distributed Checkpoint Server (Univ Paris XI, ANR Clouds@Home)
- Desktop Grid <-> Service Grid Bridge (FP7 EDGeS)
- Akratos: Decentralized, Social and Collaborative File Sharing (ADT INRIA)
- WukaStore: Hybrid and Configurable Cloud and Desktop Storage
- MapReduce for Desktop Grids (Internet deployment, n-faults resilience, decentralized result checking)

MapReduce for Large, Distributed, and Dynamic Datasets

Scheduling algorithms for optimizing shuffle phase

Execution time reduced by up to 47%!



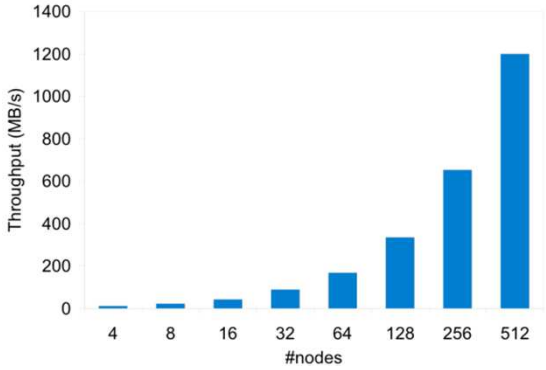
Time of map phase and shuffle w.r.t number of mappers and reducers

MapReduce runtime for

- Distributed over hybrid and widely distributed infrastructures
  - Cloud, Desktop PCs, sensors, smartphones…
- Dynamic, i.e. that grow or shrink during time, or partially unavailable because of infrastructure failures.

MapReduce/BitDew

- First implementation of MapReduce for Internet Desktop Grid
  - 2-level scheduler, latency hiding, p-failures resilient, collective communications
- Algorithm distributed result checking of intermediate
- MapReduce/ActiveData: incremental processing of dynamic datasets
- Storage on hybrid Cloud + Desktop PCs nodes
- Privacy computing on hybrid infrastructures using Information Dispersal Algorithms



Throughput of WordCount application on Grid'5000 (512 nodes) up to 2 TB

Active-Data:
A Programming Model for Data Life-Cycle Management

A *data life cycle model*

- Data management systems to expose data life cycle
- Well-formalized representation
  - Inspired by Petri Net

A *programming model* and a runtime environment

- Associate a code to each step of the data life cycle

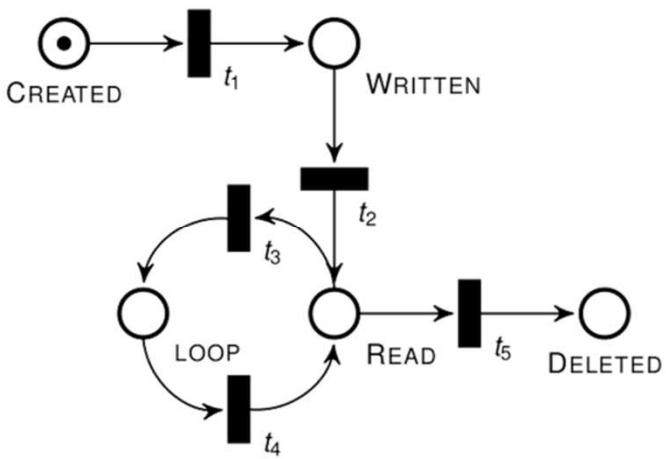


Figure: Representation of the *"Write-Once, Read-Many"* data life cycle.

```
TransitionHandler md5Handler = new ↘
    →TransitionHandler() {
  public void handler(LifeCycle lc, String ↘
      →transitionName, boolean isLocal) {
    MessageDigest md = ↘
        →MessageDigest.getInstance("MD5");
    String path = getPath(lc.getId());
    InputStream input = new ↘
        →FileInputStream(path);
    byte buffer[] = new byte[2048];

    int n = 0;
    while((n = input.read(buffer)) > 0)
      md.update(buffer, 0, n);

    byte[] digest = md.digest();
    BigInteger bigInt = new ↘
        →BigInteger(1,digest);
    String hash = bigInt.toString(16);
    while(hash.length() < 32 )
      hash = "0" + hash;

    OutputStream output = new ↘
        →FileOutputStream(path + ".md5");
    output.write(hash.getBytes());
    output.close();
  }
};
```

Conclusion

Efficient usage of resources (clouds, hpc, etc) require to master many aspects

Avalon team focuses on

- Energy Application Profiling and Modeling
- Data-intensive Application Profiling, Modeling, and Management
- Resource Agnostic Application Description Model
- Application Mapping and Scheduling

Research from theory to software development and experimental validation

- Diet, Simgrid, BitDew, HLCMi, L2C, etc.
- Access to research platform (Grid'5000) and production platforms through CC-IN2P3

Important involvement in international and national projects

- INRIA-UIUC-NCSA-ANL joint laboratory for petascale computing
- Green'Touch (2012-2015) -- *Reduce energy consumption in networks*

- PRACE-2IP (FP7 RI, 2011-2013) -- *Auto-tuning of component based applications for supercomputers*
- PaaSage (FP7 ICT, 2012-16) -- *Model Based Cloud Platform Upperware*
- Seed4C (Celtic-Plus, 2012-14) -- *Secured Embedded Element for Cloud*
- COST IC804 (2009-2013) -- *On Energy efficiency for large scale systems*

- ANR MapReduce (2010-2014) -- *Advanced data management, scheduling and algorithmic skeletons*
- ANR Songs (2012-16) -- *SimGrid for Clouds and High Performance Computation systems*
- FSN XLCLOUD (11-14) -- *Energy Efficient HPC as a Service (Openstack)*

Thank you

INVENTORS FOR THE DIGITAL WORLD