The Green Computing Observatory: from instrumentation to ontology

Cécile Germain-Renaud[1], Fredéric Fürst[2], Gilles Kassel[2], Julien Nauroy[1], Michel Jouvin[3], Guillaume Philippon[3]

1: Laboratoire de Recherche en Informatique, U. Paris Sud, CNRS, INRIA

2: Université Picardie Jules Verne

3: Laboratoire de l'Accélérateur Linéaire, CNRS-IN2P3

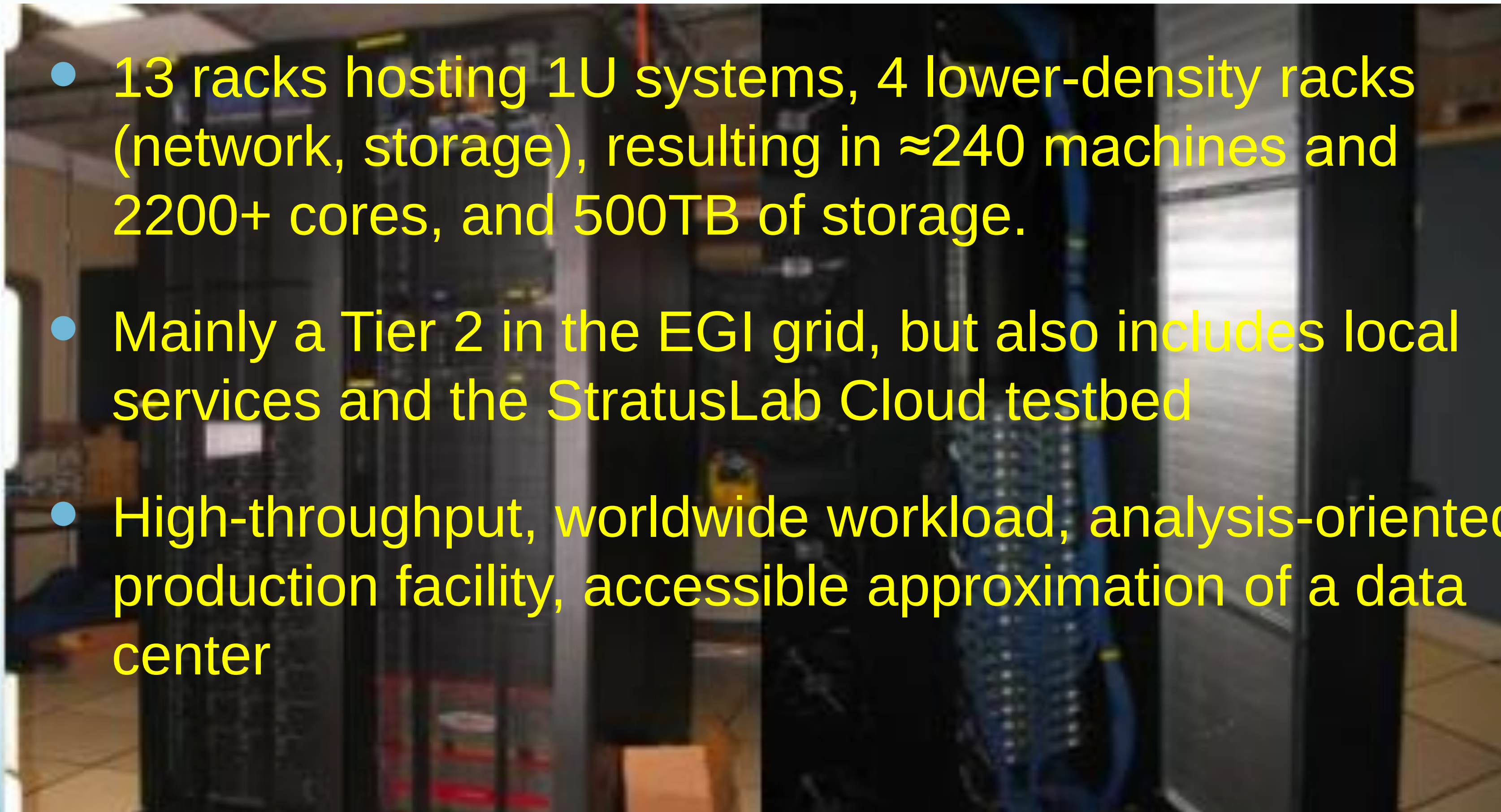
GCO: a Digital Curation approach

- Establish long-term repositories of digital assets for current and future reference
  - Continuously monitoring a large computing facility
- Tackling the good data creation and management issues, and prominently interoperability,
  - Formal mainstream ontology, standard-aware
- Providing digital asset search and retrieval facilities to scientific communities through a gateway
  - Files in XML format
  - Available from the Grid Observatory portal

With the support of

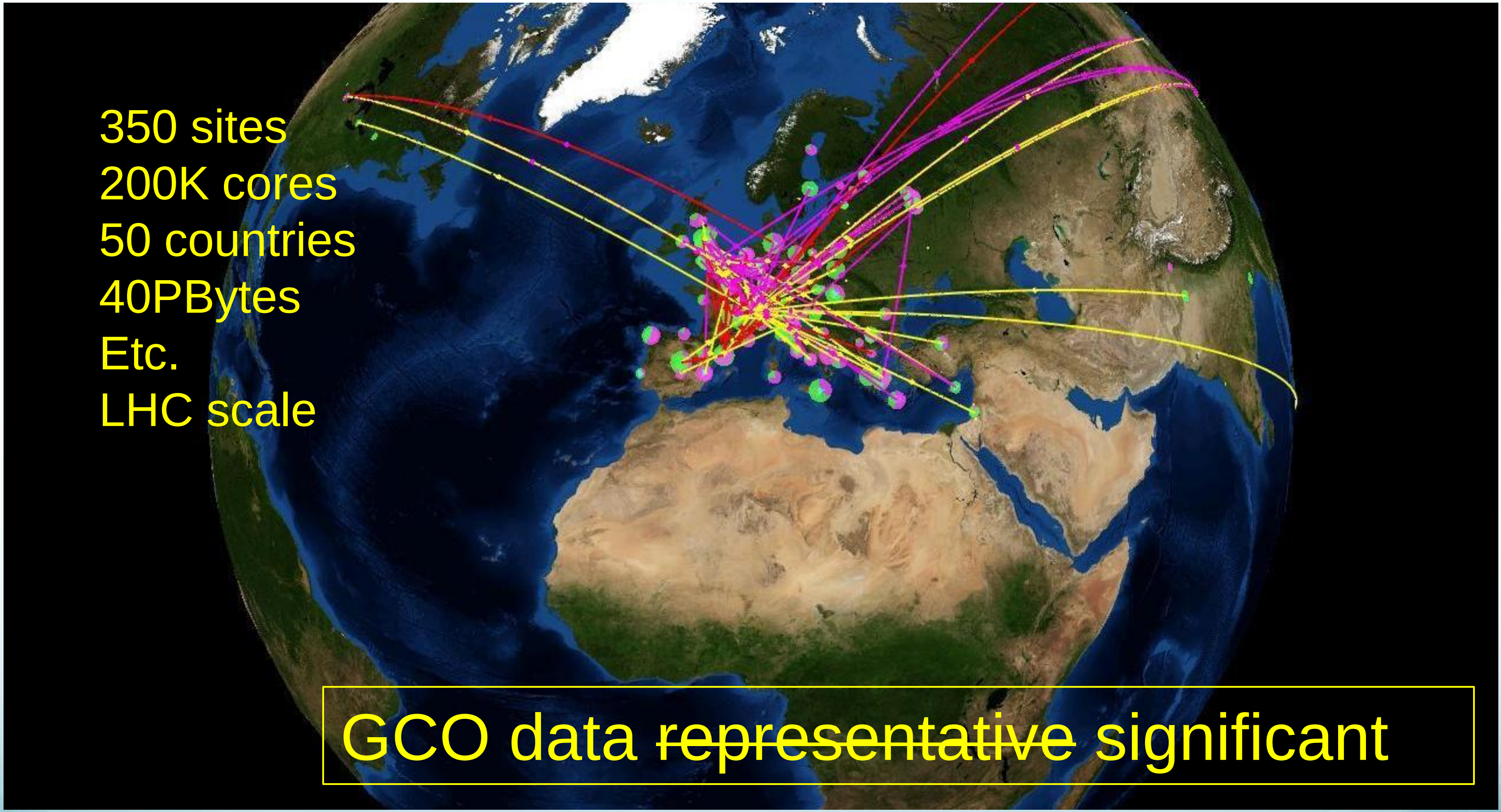
- France Grilles – French NGI member of EGI
- EGI-Inspire (FP7 project supporting EGI)
- INRIA – Saclay (ADT programme)
- CNRS (PEPS programme)
- University Paris Sud (MRM prog

The GRIF-LAL computing room



- 13 racks hosting 1U systems, 4 lower-density racks (network, storage), resulting in ≈240 machines and 2200+ cores, and 500TB of storage.
- Mainly a Tier 2 in the EGI grid, but also includes local services and the StratusLab Cloud testbed
- High-throughput, worldwide workload, analysis-oriented production facility, accessible approximation of a data center

EGI grid: very large non-profit distributed system

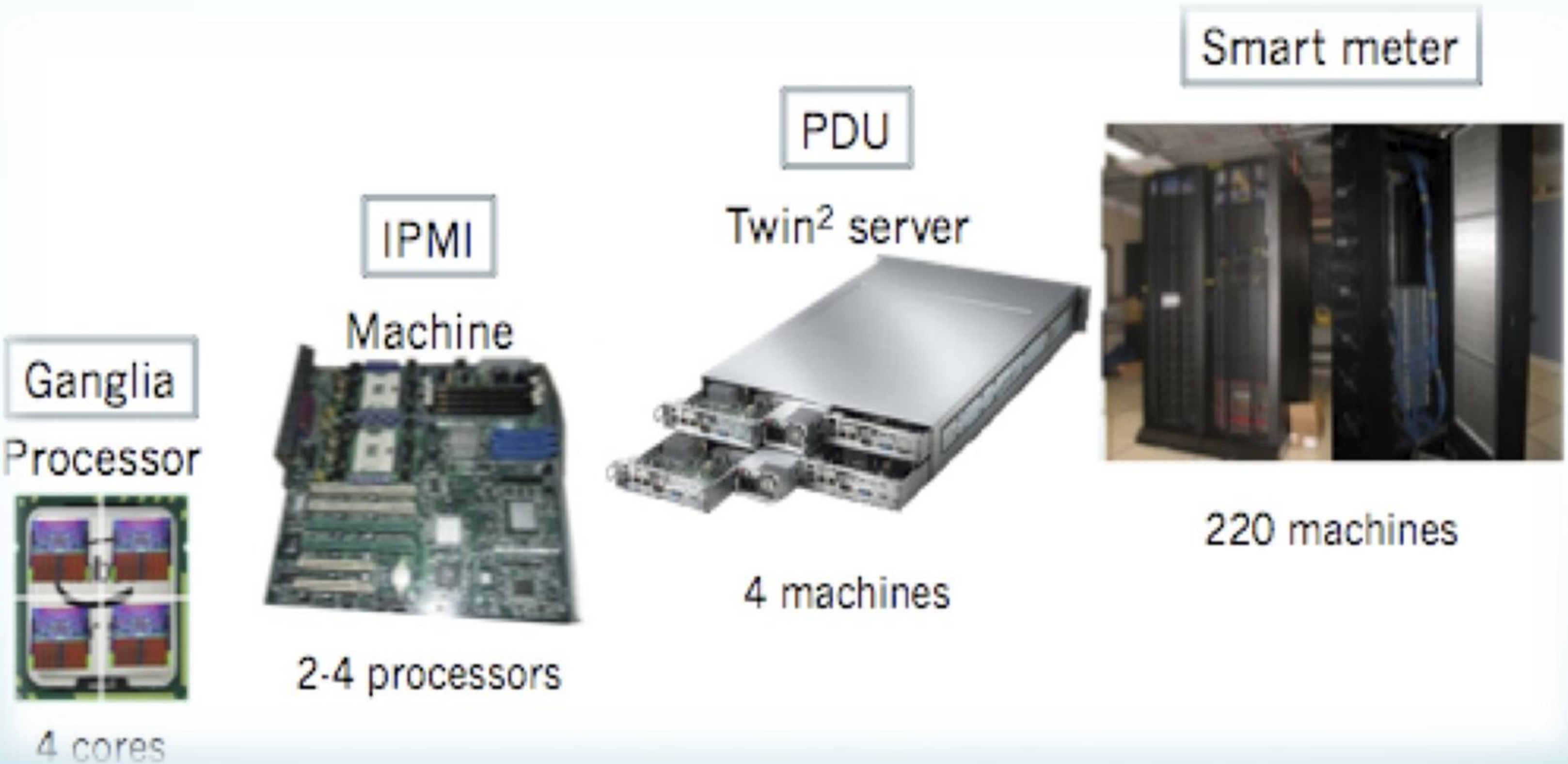


350 sites
200K cores
50 countries
40PBytes
Etc.
LHC scale

GCO data ~~representative~~ significant

Sensors

Grid Observatory



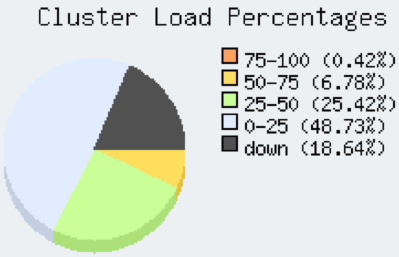
2GBytes/day at 1 minute sampling period

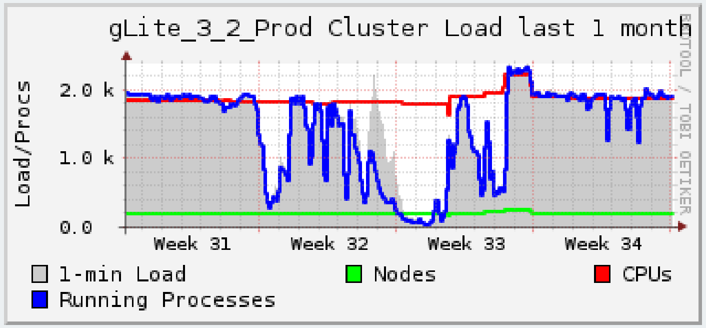
Overview of gLite_3_2_Prod

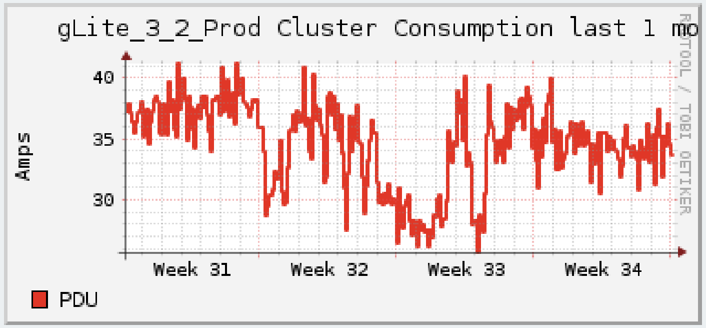
Cores Total : 2210
Hosts up : 192
Hosts down : 44

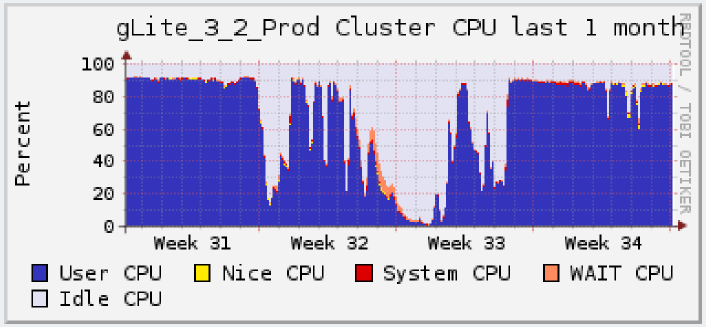
Average Load :
1 minute 33%
5 minutes 34%
15 minutes 34%

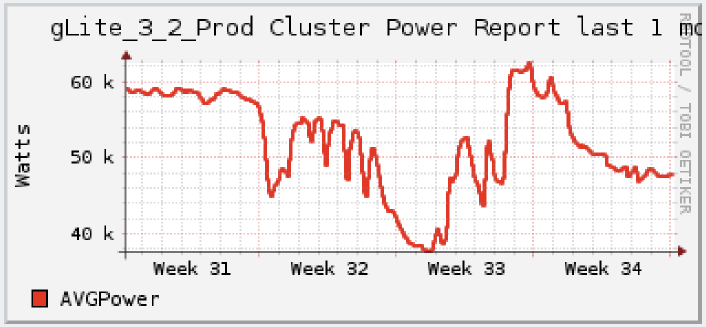
Localtime : 2011-08-29 02:29

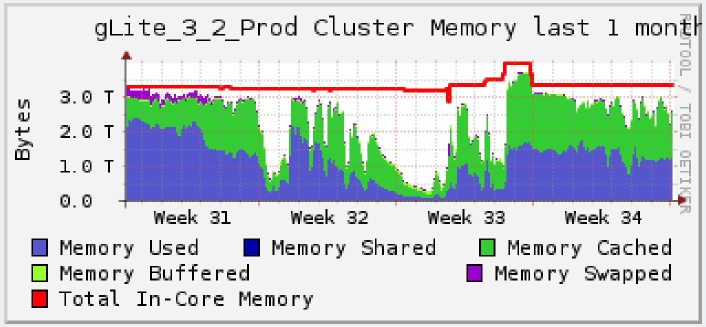


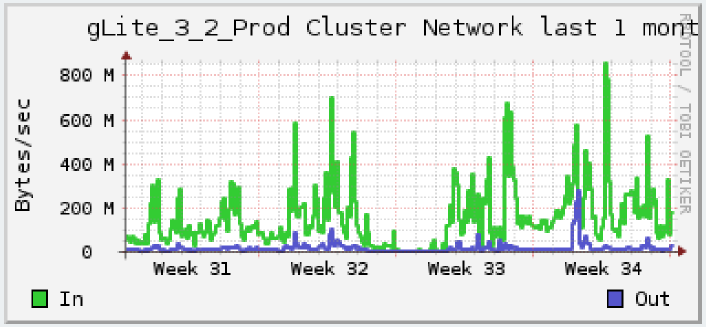












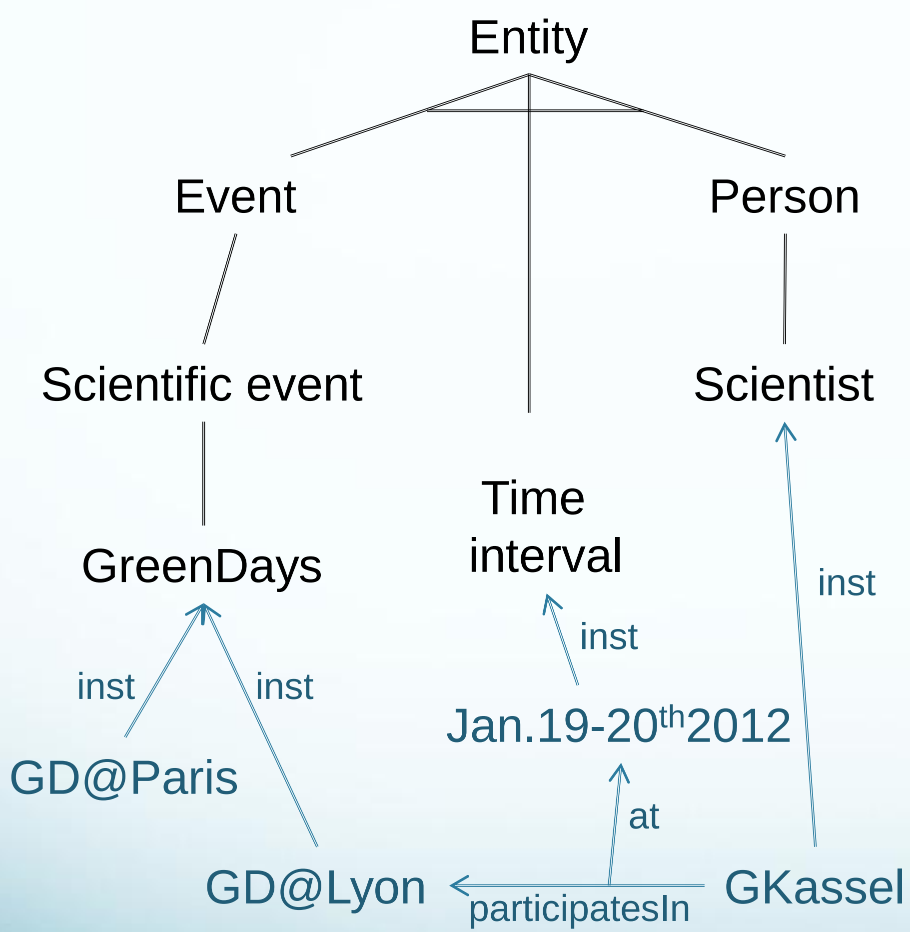
Information model

- There is no standard for
  - The output of the physical sensors
  - The integration of computational usage and physical sensors’ output
- There are standards for
  - OS information: Ganglia
  - Virtual Machine definition: OVF
  - Centralized statistics publication: SDMX (Statistical Data and Metadata Exchange). Successful experience of porting to a Linked Data model.

GCO: a Digital Curation approach

- Establish long-term repositories of digital assets for current and future reference
  - Continuously monitoring a large computing facility
- Tackling the good data creation and management issues, and prominently interoperability,
  - **Formal mainstream ontology, standard-aware**
- Providing digital asset search and retrieval facilities to scientific communities through a gateway
  - Files in XML format
  - Available from the Grid Observatory portal

You said “ontology”!



$\forall x$ ScientificEvent(x) $\rightarrow$ Event(x) $\wedge$ $\exists y,t$ participatesInAt(y,x,t)

$\forall x$ GreenDays(x) $\rightarrow$ ScientificEvent(x)

GreenDays(GD@lyon)

Scientist(GKassel)

participatesInAt(GKassel, GD@lyon, January19-20th2012)

```
<owl:Class rdf:ID="GreenDays">
<rdfs:subClassOf rdf:resource="#Scientific Event
"/>
…
</owl:Class>
<GreenDays rdf:ID= " GD@Lyon" />
<owl:ObjectProperty rdf:ID="participatesIn">
<rdfs:domain rdf:resource="#Person"/>
<rdfs:range rdf:resource="#Event"/>
</owl:ObjectProperty>
```

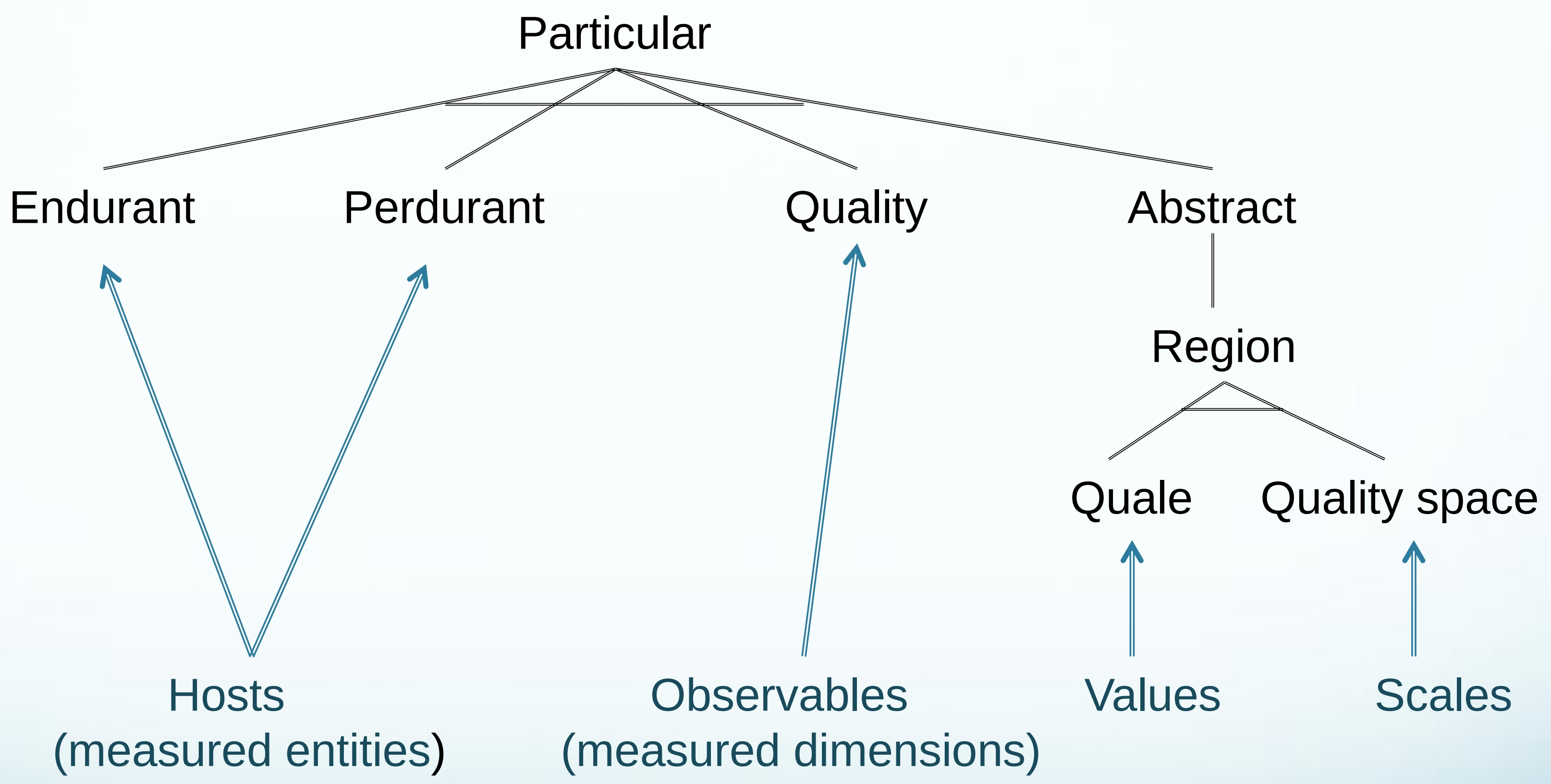
Why use ontologies in GCO?

- Our purpose
  - To clarify the ***semantics*** of data
    - To get a ***computational*** model
  - To define an ***ontological semantics*** for the XML schema
- Our approach
  - To define a ***semantically transparent*** ontology
    - To reuse the ***foundational*** DOLCE[1] ontology
  - To use the OntoSpec[2] methodology (**modularity + high expressiveness**) which integrates the OntoClean[1] methodology

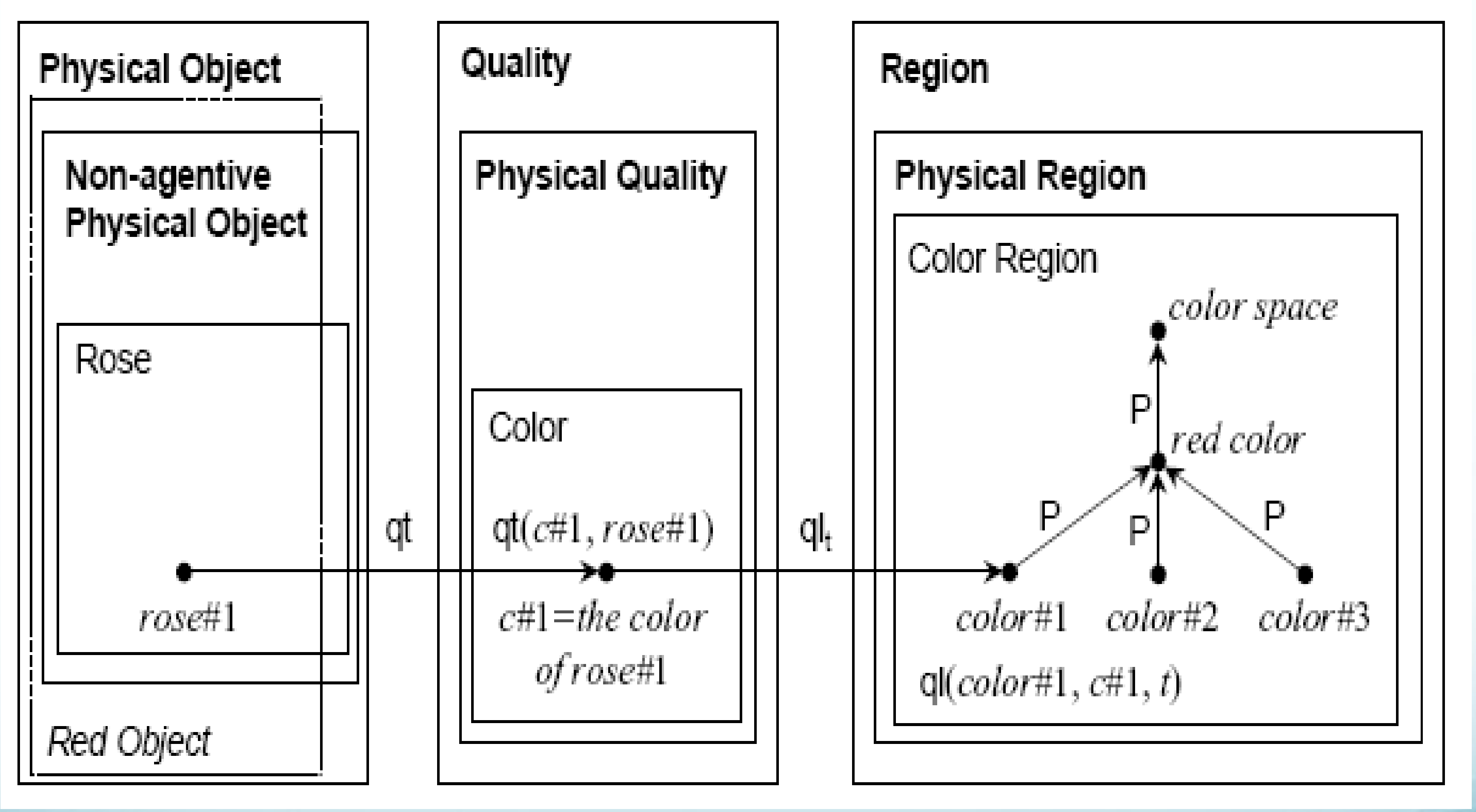
[1] Laboratory for Applied Ontology: http://www.loa.istc.cnr.it/

[2] http://home.mis.u-picardie.fr/~site-ic/site/?lang=en

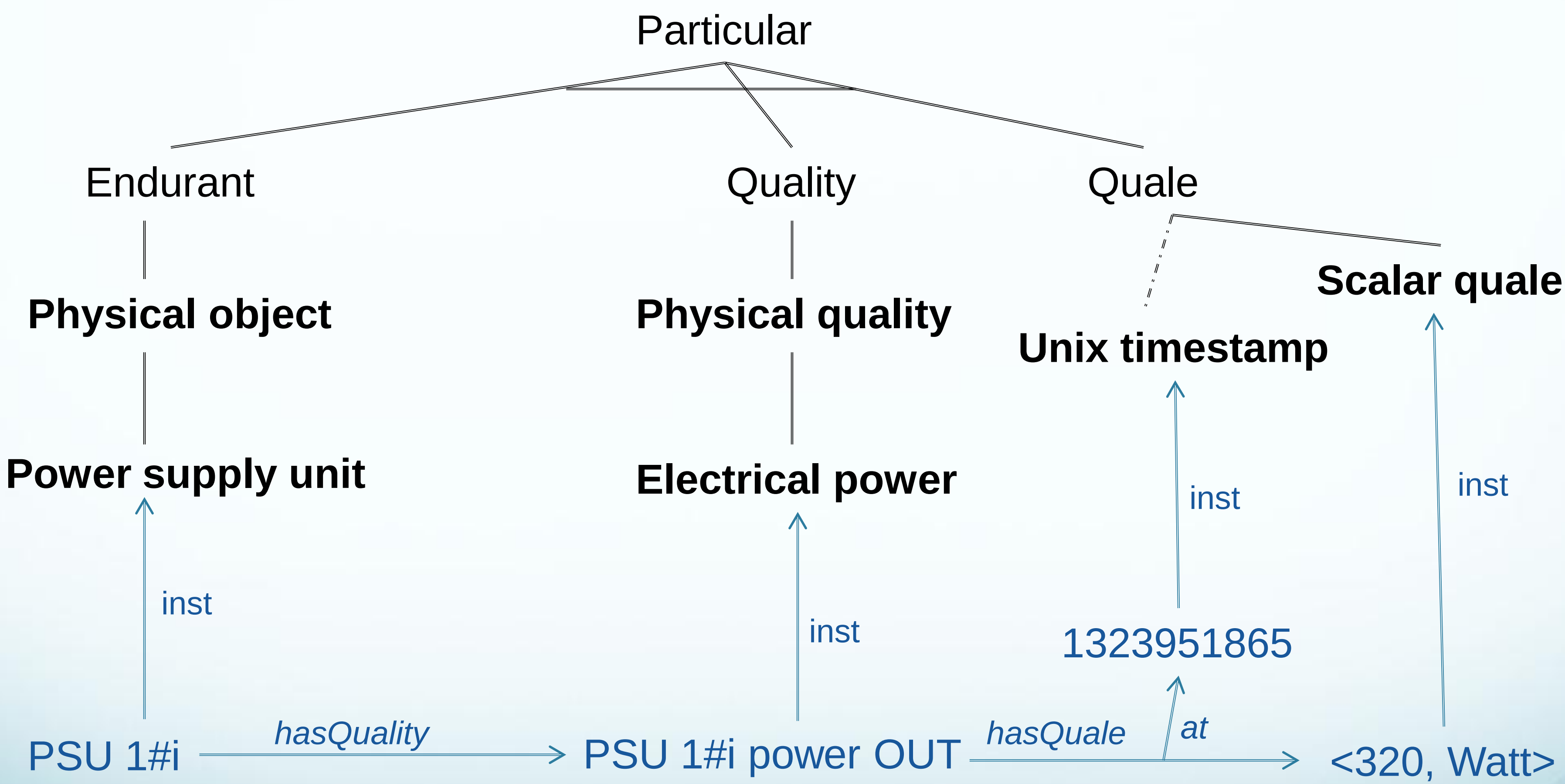
DOLCE and the measurement of entities



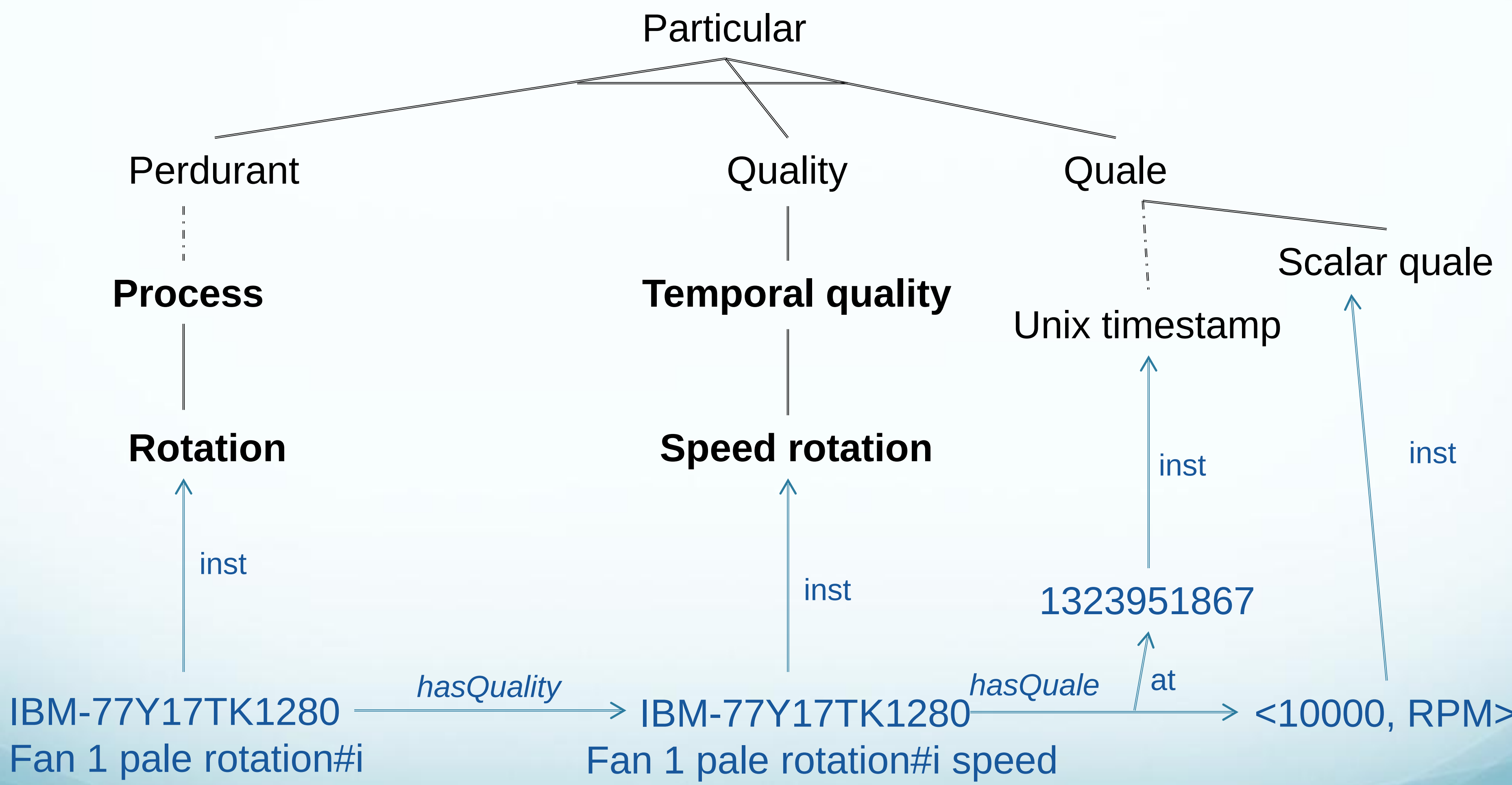
Qualities are inherent to their host



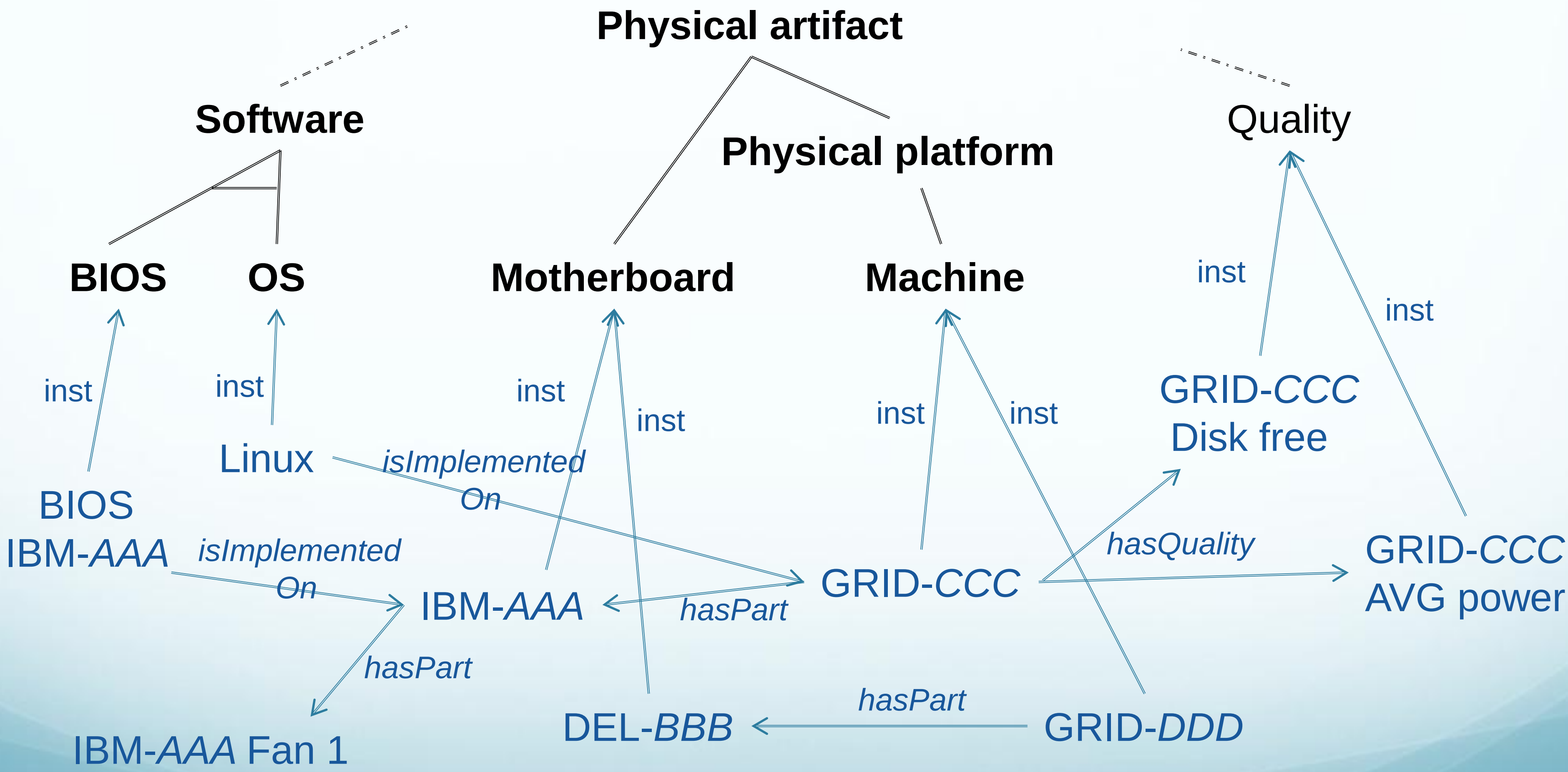
Example of a Physical object/quality



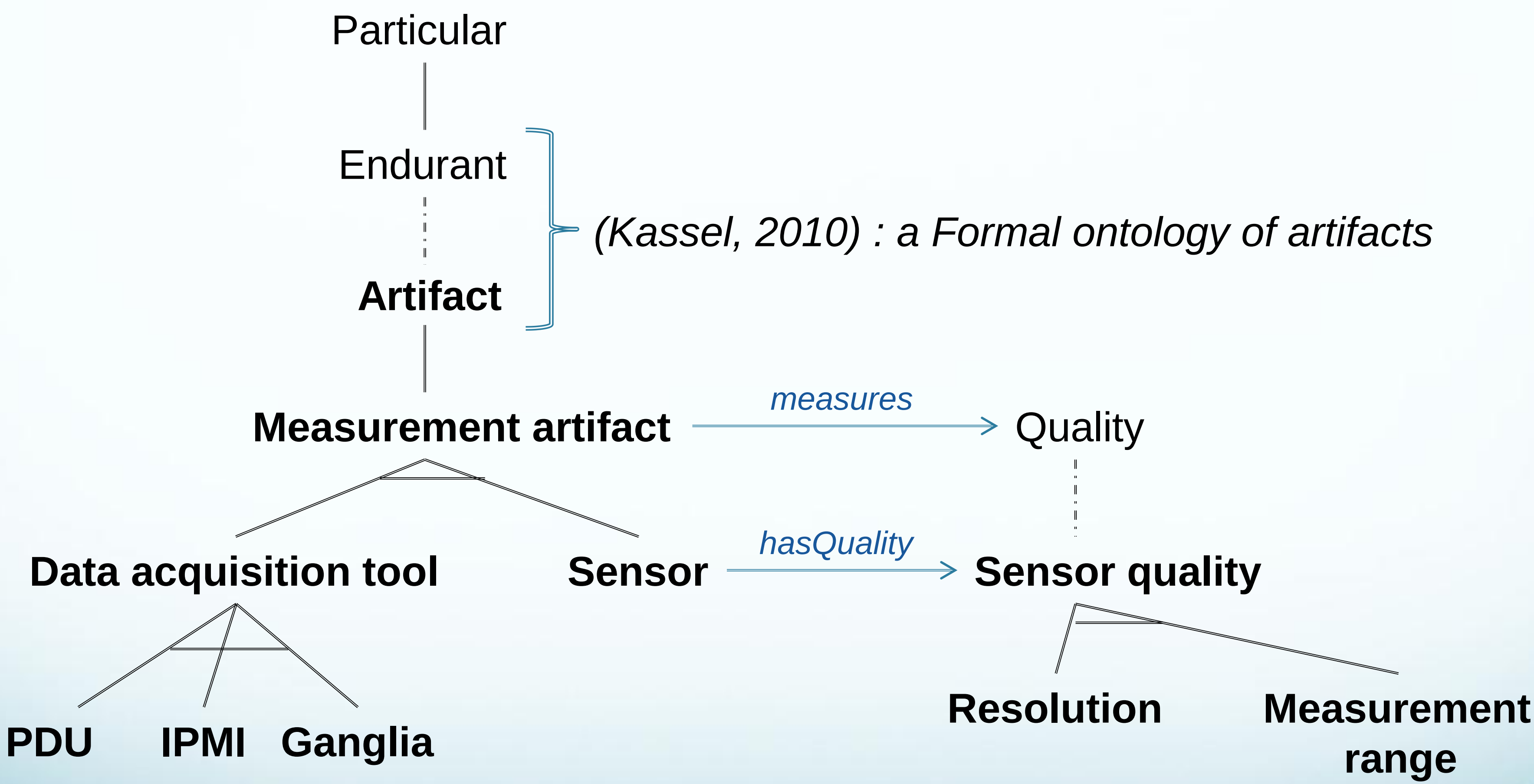
Example of a temporal object/quality



Other kinds of Physical objects: Motherboards and Machines



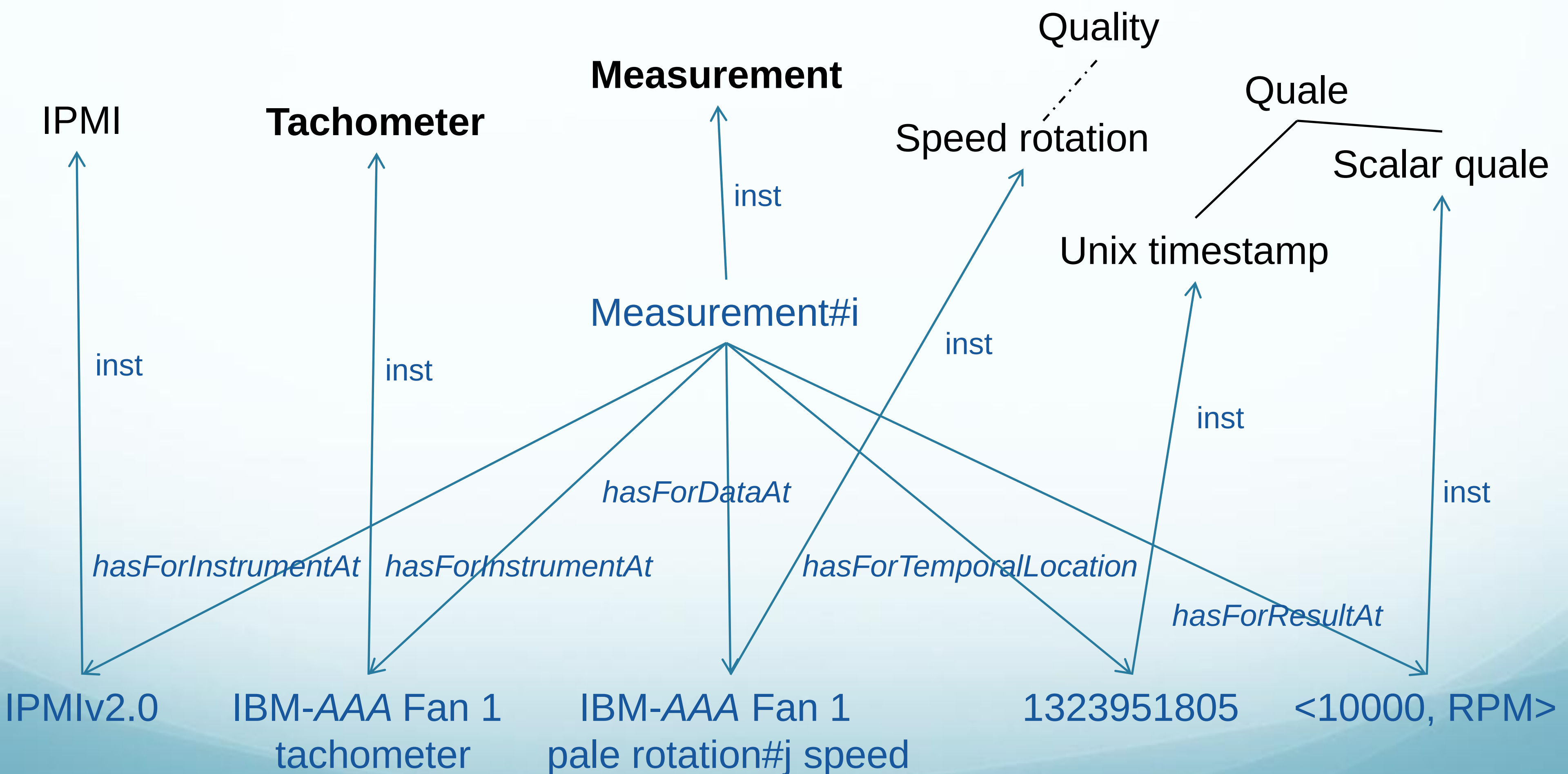
Measurement tools



Ontological semantics of tuples (1/2)

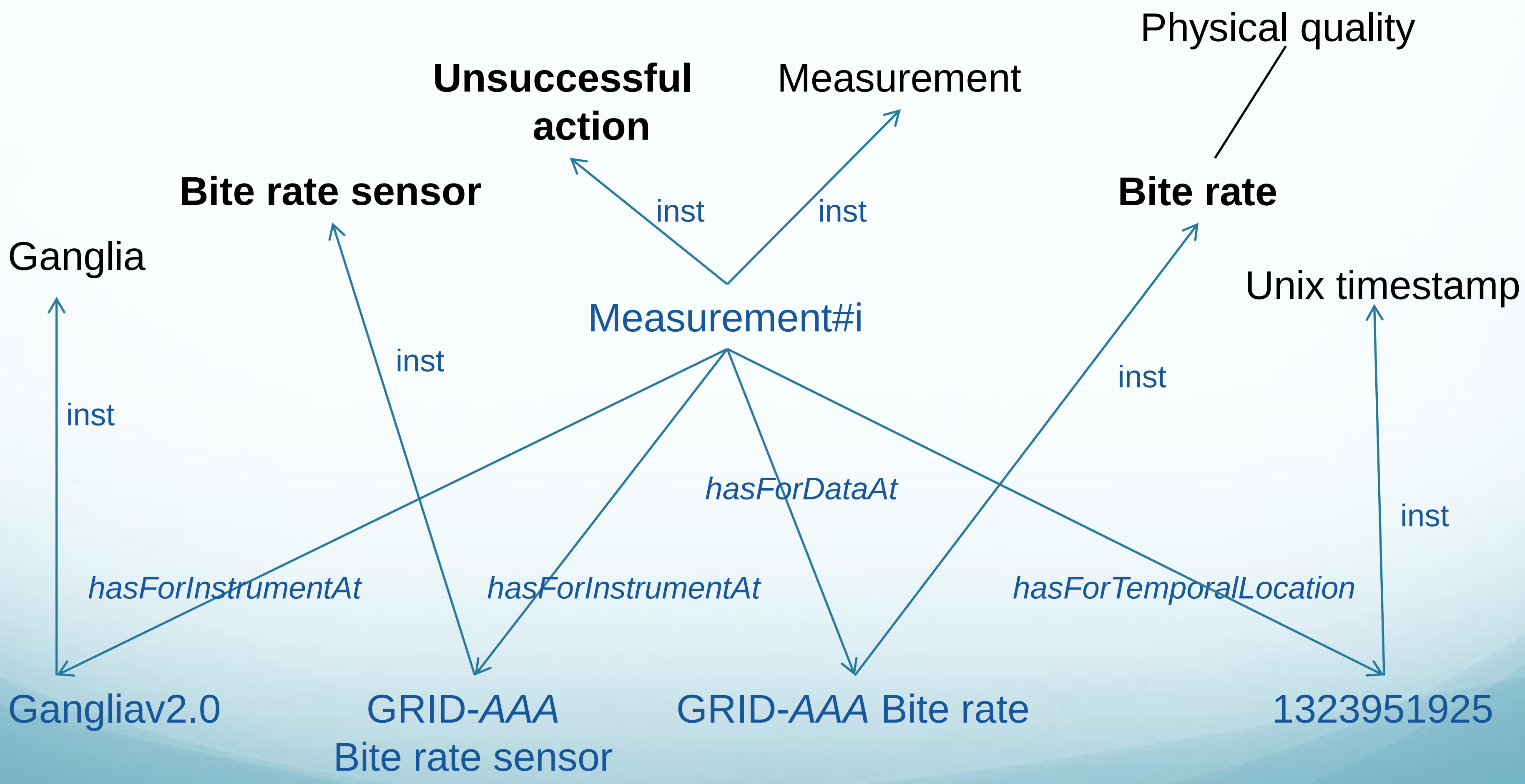
Grid Observatory

("IPMI", "FAN1 TACH", 1323951805, 10000)



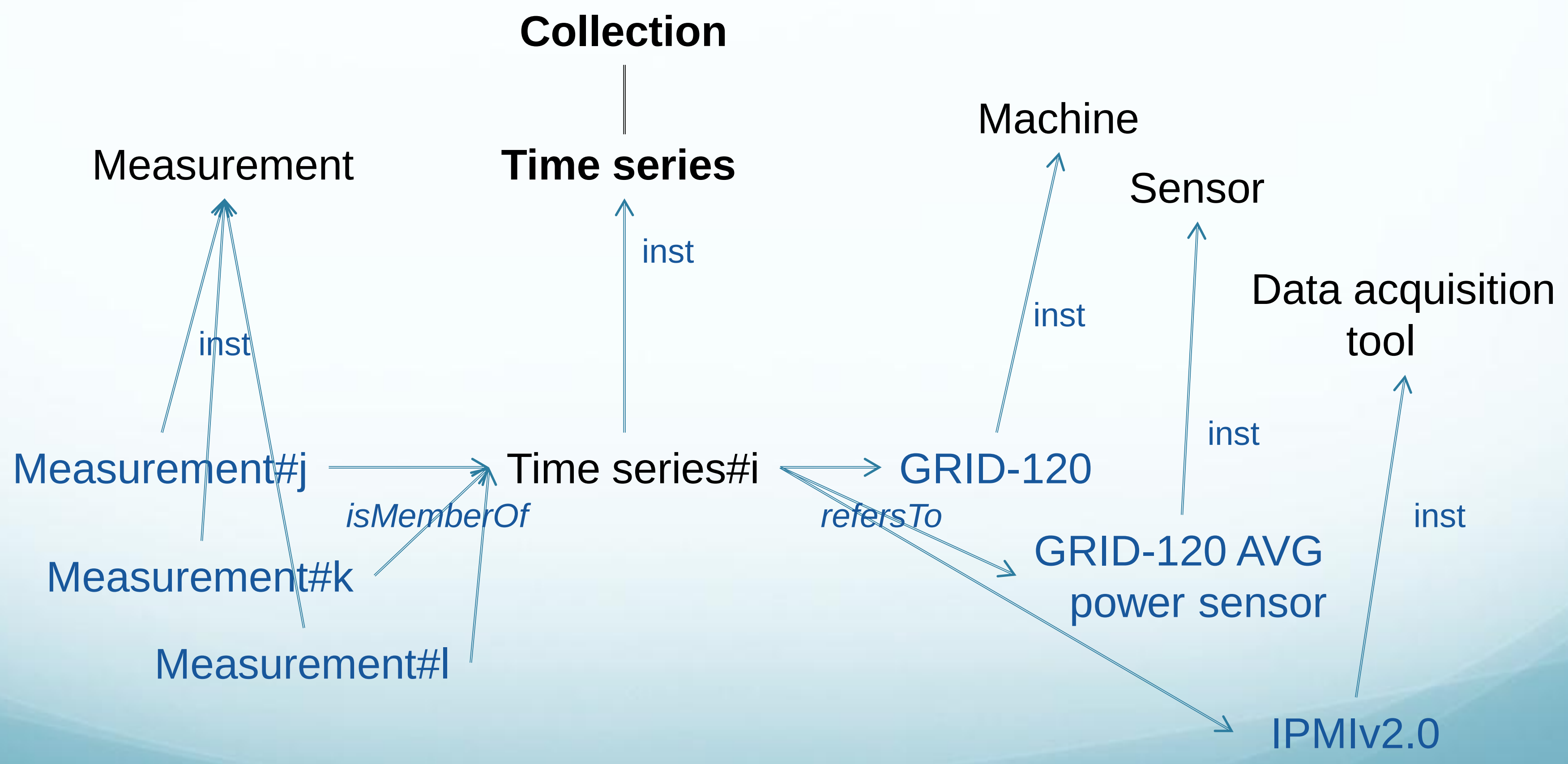
Ontological semantics of tuples (2/2)

("Ganglia", "Byte_in", 1323951925, NaN)



Ontological semantics of time series

```
<timeseries machineID="1" machineName="grid120" instrumentID="2">
    <a t="1325942960" v="320.00" />
    <a t="1325943020" v="310.00" />
    <a t="1325943080" v="320.00" />
</timeseries>
```



GCO: a Digital Curation approach

Grid Observatory

- Establish long-term repositories of digital assets for current and future reference
  - Continuously monitoring a large computing facility
- Tackling the good data creation and management issues, and prominently interoperability,
  - Formal mainstream ontology, standard-aware
- Providing digital asset search and retrieval facilities to scientific communities through a gateway
  - **Files in XML format**
  - **Available from the Grid Observatory portal**

Ontology-compatible XML Format: Why XML ?

- Interchange format
  - Easy to define your own syntax (DTD, XSD)
- Easy to manipulate
  - Manipulation languages (XSLT, XPath, XQuery…)
  - Lots of available libraries and tools (libXML, databases)
- Easy to extend
- Drawbacks:
  - Parsing can be quite slow
  - Libraries are not adapted to parsing gigabytes of data

Acquisitions

- Currently, 220+ machines are monitored
  - Data (time series) and metadata is acquired
- Machine = motherboard + middleware
  - “middleware” refers to the OS or the hypervisor
- Time series always refer to a machine
  - Helps uniting different acquisitions
    - e.g.: processor temperature connected to processor usage
  - Forces evaluating the acquisition context
    - e.g.: chassis temperature depends on power consumption
- Metadata is decisive to interpret the acquisitions

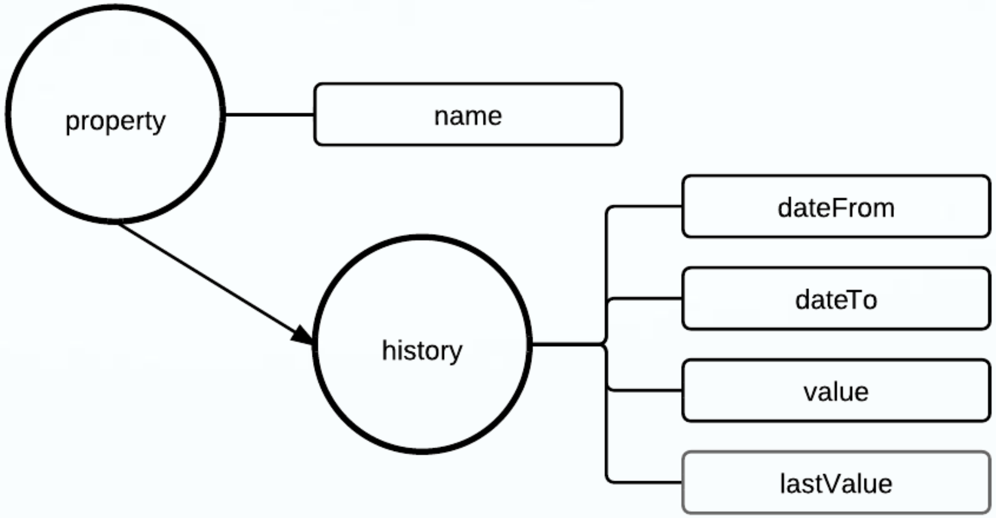
Acquisition

- ~25 time series from Ganglia
  - CPU usage, memory usage, network traffic, etc
- 30 to 50 time series from IPMI
  - Temperatures, fans speed, voltages, power consumption
  - Not all are relevant! e.g.: “Drive 1 Status”
  - Some give erroneous values! e.g.: “MCH Temp ” = -1° ?
- 1 time series for each power outlet
  - May be shared by multiple machines
- Metadata is acquired with the same tools

Different types of acquisition

- Identity properties: never change
  - e.g.: motherboard information, middleware version
  - Used to fully identify the entity
- Slowly mutable properties: can sometimes change
  - e.g.: IP address, Firmware version
  - Keep a history of the modifications
- Time series: values can constantly change
  - e.g.: CPU activity, power consumption
  - Keep track of every value and acquisition date

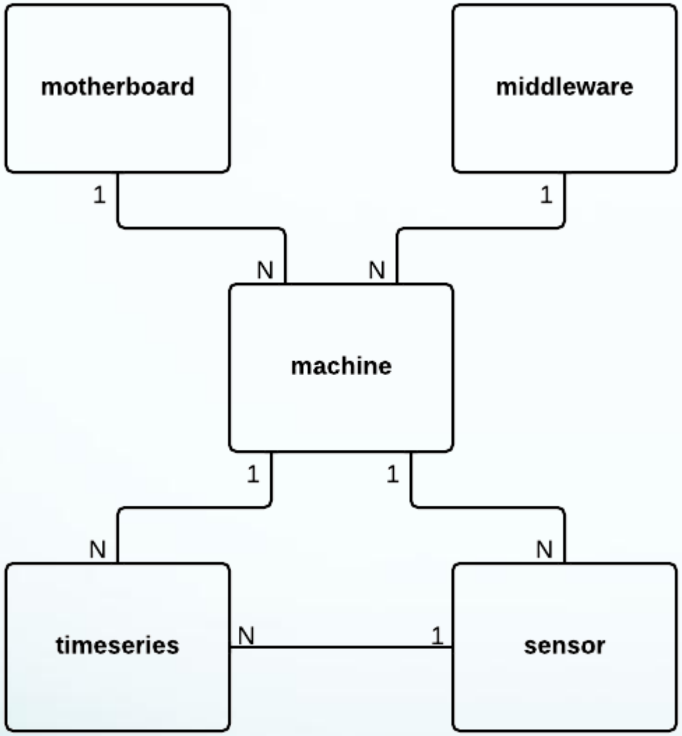
Slowly mutable properties



- Slowly mutable properties include:
  - The last known value
  - A history of previous values
- Gap every time a value changes, e.g.:
  - [t1, t2] = X
  - [t3, t4] = Y
  -]t2, t3[= ?

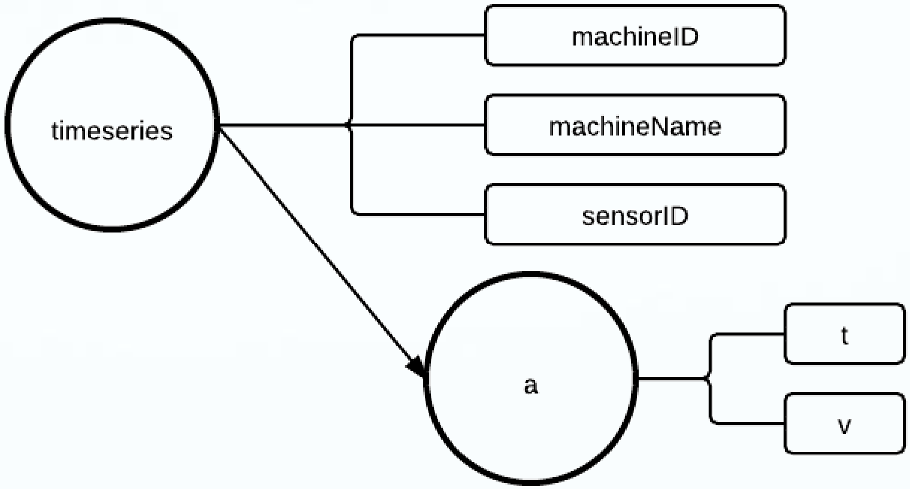
XML format and specificities

Grid Observatory



- A time series refers to a machine
  - A measurement needs a context !
- Hosts are not directly represented
  - Difference with the ontology
- “Middleware” accommodates virtual and non-virtual cases

Definition of a time series



- A time series is a collection of **a**cquisitions represented by a couple (**t**imestamp, **v**alue)
  - Acquisition frequency not guaranteed! bugs, slowdowns,…
- A time series refer to a machine and a sensor
- Typically, one time series per day

Example of time series

```xml
<timeseries machineID="1"
            machineName="grid120"
            instrumentID="2">
    <a t="1325942960" v="320.00" />
    <a t="1325943020" v="310.00" />
    <a t="1325943080" v="320.00" />
</timeseries>
```

Definition of a motherboard

- A motherboard is fully defined by:
  - Vendor (“DELL”)
  - Serial number (“CN0D61XP747510BN0926A00”)
- Other immutable features:
  - Manufacturer (“DELL”)
  - Manufacturing date (“Sun Nov 28 12:05:00 2010”)
  - Product name (“PowerEdge”)
  - Part number (“VJ0BMP0878”)
- Slowly mutable features:
  - Firmware revision (“1.27”)
  - IPMI version (“2.0”)

Example of motherboard

```
<motherboard ID="1" dateFrom="1325942960"
        name="Dell-CN0D61XP747510BN0926A00"
        vendor="Dell"
        product="PowerEdge"
        partNumber="VJ0BMP0878"
        serial="CN0D61XP747510BN0926A00"
        manufacturingDate="Sun Nov 28 13:09:00 2010">
    <property name="firmwareRevision">
        <history value="1.27" from="1325942960"
                to="1325943080" lastKnown="true" />
    </property>
    <property name="IPMIVersion">
        <history value="2.0" from="1325942960"
                to="1325943080" lastKnown="true" />
    </property>
</motherboard>
```

Definition of a middleware

- A middleware is fully defined by:
  - Type: OS or hypervisor
  - Product name and version (“SL 5.5”)
  - Kernel name and version (“Linux 2.6.18”)
- Other immutable features:
  - Architecture (e.g.: “x86”, “x86_64”)
- Slowly mutable features:
  - Any ? Information retrieved at the “running OS” level generally belong to the machine
    - e.g. hostname : one per machine, not one per middleware

Example of middleware

Grid Observatory

```
<middleware ID="1"
        hostname="grid120.lal.in2p3.fr"
        productName="SL"
        productVersion="release 5.5 (Boron)"
        kernelName="Linux"
        kernelVersion="2.6.18-238.12.1.el5"
        OSArchitecture="x86_64" />
```

Definition of a machine

- A machine is fully defined by:
  - Its hardware
  - Its middleware
- Changing the association = creating a new machine
  - For a given hardware, a different middleware can be used
- Immutable features:
  - Resources attribution: memory, cores (threads), etc
- Slowly mutable features:
  - Hostname (name + domain : “grid200.lal.in2p3.fr”)
  - IP address (“134.158.73.96”)

Example of machine

```
<machine ID="1"
         dateFrom="1325942960"
         motherboardInstanceID="1"
         middlewareInstanceID="1">
    <property name="name">
        <history value="grid120" from="1325942960"
                to="1325943080" lastKnown="true" />
    </property>
    <property name="domain">
        <history value="lal.in2p3.fr" from="1325942960"
                to="1325943080" lastKnown="true" />
    </property>
    <property name="IP Address">
        <history value="134.158.73.96" from="1325942960"
                to="1325943080" lastKnown="true" />
    </property>
</machine>
```

Definition of a sensor

Grid Observatory

- A sensor is fully defined by:
  - The machine it belongs to
  - Its acquisition tool (IPMI, Ganglia, PDU)
  - Its name inside the acquisition tool
- Slowly mutable features:
  - The unit of the measurement (Volt, Watt, RPM, etc)
  - Qualities of the measurement process
    - Resolution
    - Accuracy
    - Precision
    - Response time
    - etc

Example of sensor

```
<sensor ID="1" machineID="1"
        acquisitionTool="IPMI"
        name="AVG Power" unit="Watt" />
```

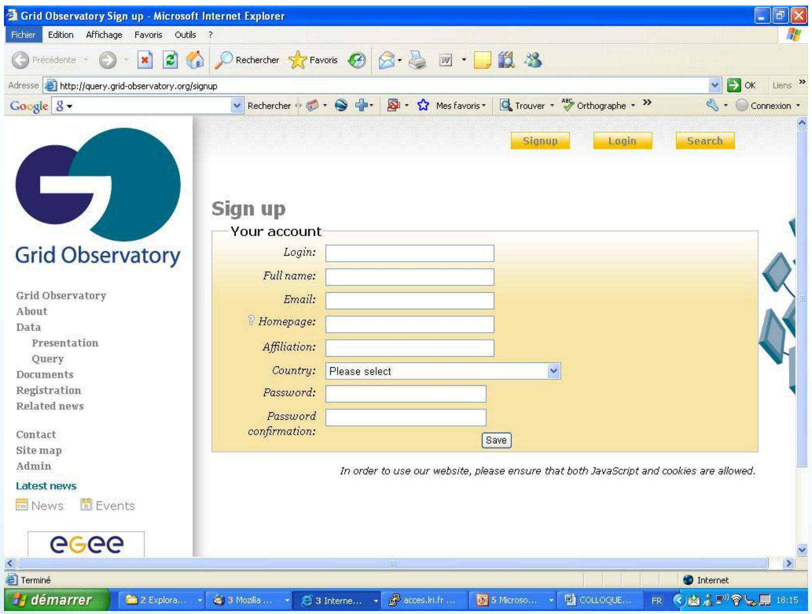
- Qualities are not yet completely defined

File representation

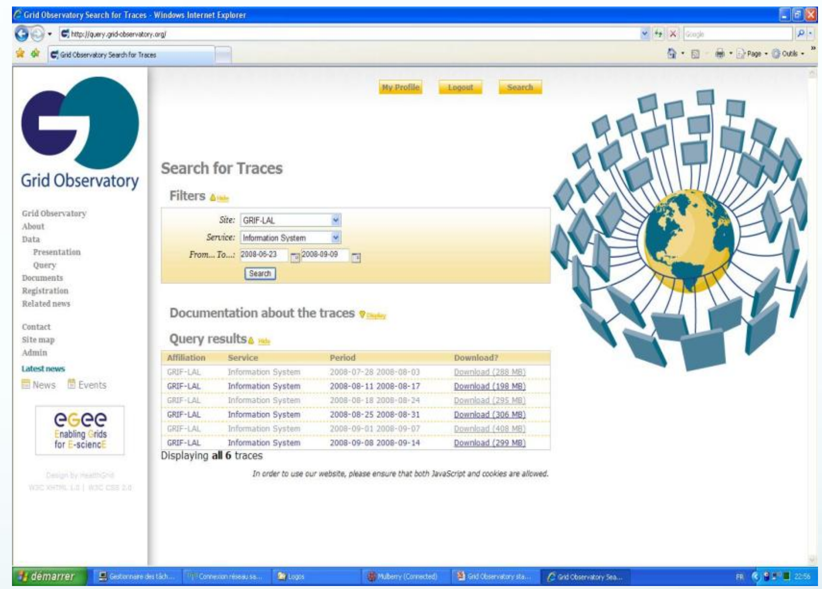
- One file for metadata, each week
  - e.g.: “metadata2012W03.xml”
  - Contains machines, hardware, middleware, sensors
  - Contents of week X fully represented in week X+1
  - Expected size :1/1000th of the time series size
- One file for time series :
  - Per day
  - Per machine
  - Per acquisition source
  - e.g.: <grid120>-<20120119>-IPMI.xml
  - 2,5MB per day for IPMI, 2MB for Ganglia, 50kB for PDU

How to

Get an account

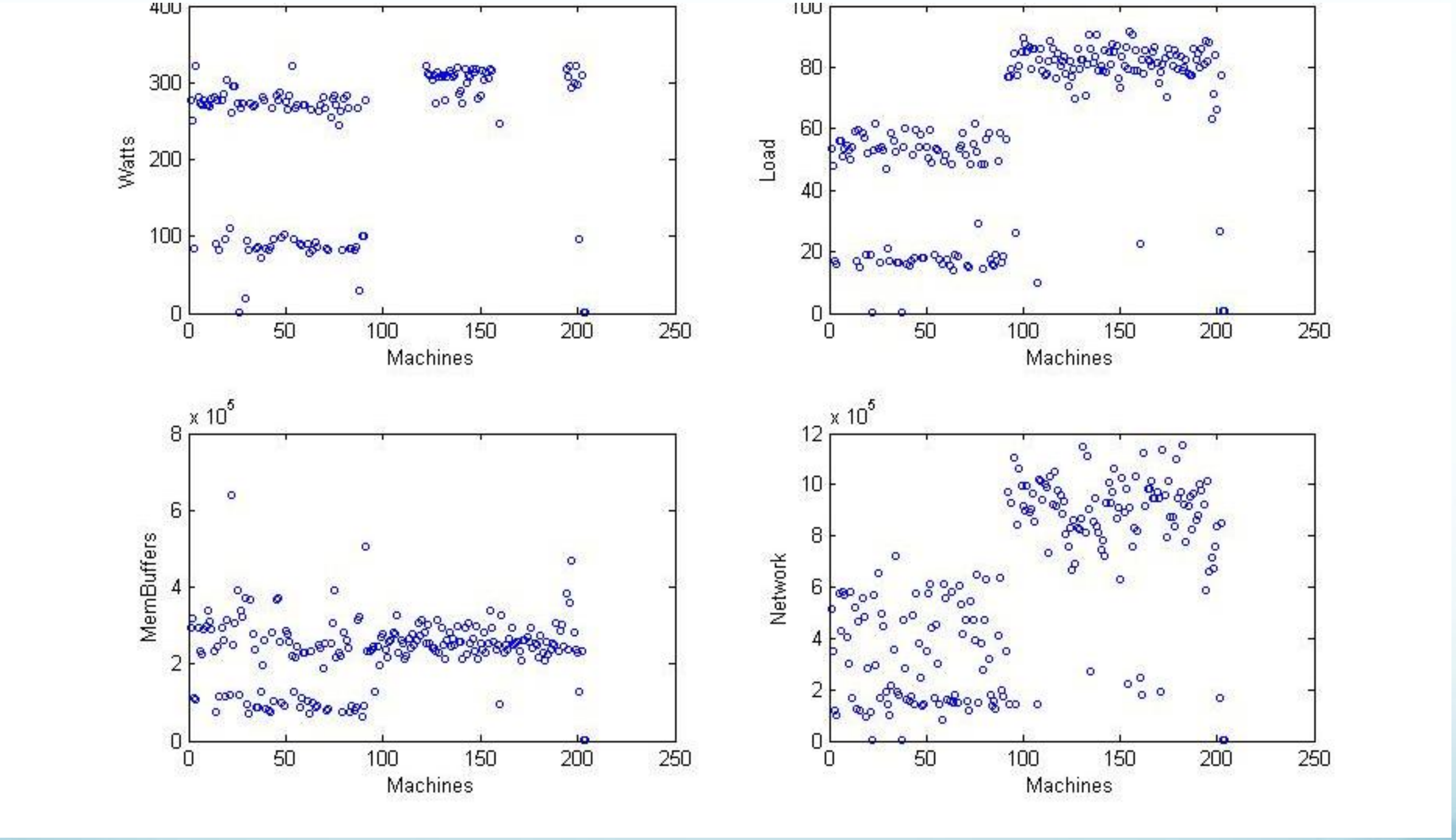


Download files



www.grid-observatory.org

Preliminary: different regimes

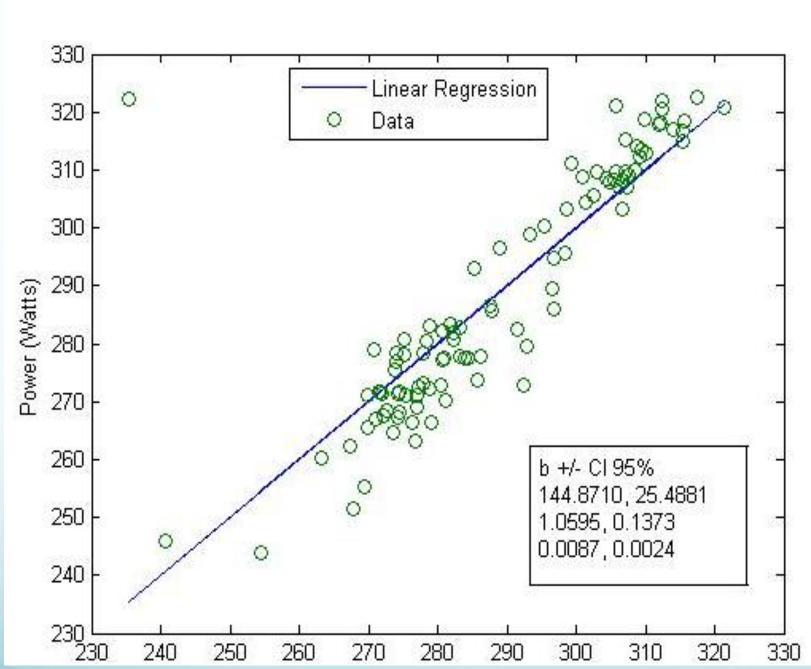


All are time averages

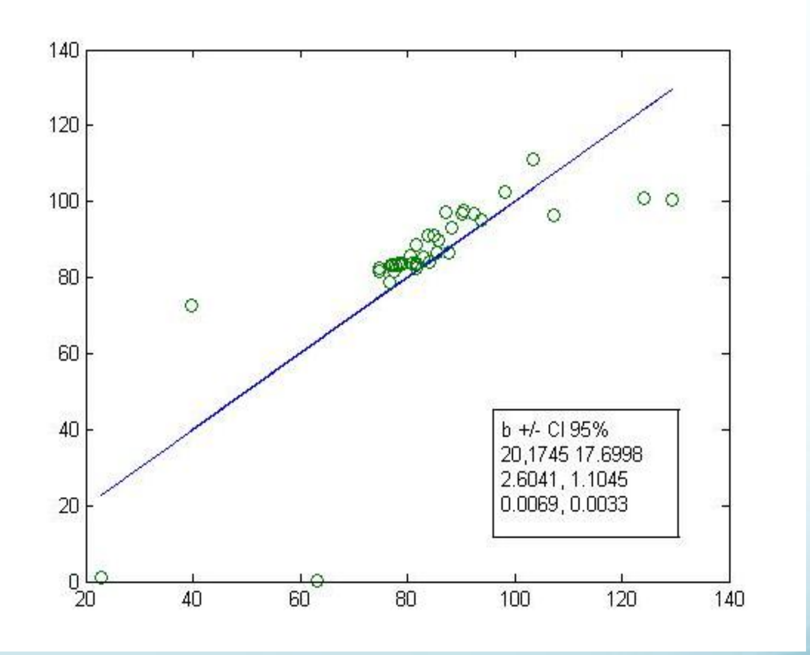
Preliminary: multivariate regression

Power as a function of load AND « fan speed »

Active machines



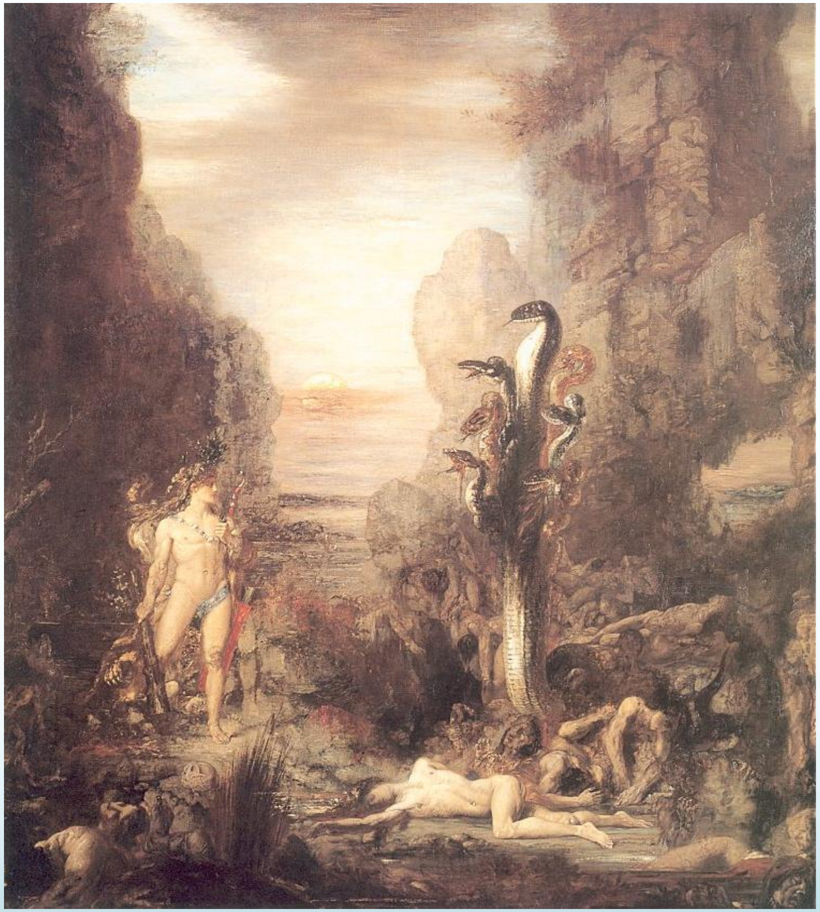
Idle machines



Status and Roadmap

- Acquisition of timeseries and metadata for IPMI, Ganglia, PDU and temperature are in production
- **Examples** of raw timeseries for IPMI, PDU and Ganglia released
- Metadata integration and temperature timeseries, stable XML schema V1 **Q1 2012**
- Monitoring Virtual Machines from the StratusLab platform Q4 2012
- Global energy consumption Q4 2012
- Also: rack monitoring

Conclusion



Discussion

- Même objectif général: publication, mise à disposition
- Tout le reste est différent !
  - Complémentaire
    - Basse fréquence : les conditions (charge-température) varient peu
  - Moins complémentaire
    - Multi-protocole, interopérable
    - Multi-senseurs, extensible sémantiquement
    - Contexte d'acquisition, charge : anti-confidentiel
    - Data curation
    - Format texte
    - Stockage illimité : sur la grille
    - Extensibilité interne à la grille EGI : protocole interne (ActiveMQ)
- Quelques pistes de standards pour l'interopérabilité
    - Multi –source pour l'acquisition : SDMX Statistical Data and Metadata Exchange
    - Multi-source pour les données : SDMX -> Linked Data