Project acronym: **DATAZERO**

Project full title:

**DATAcenter with Zero Emission and Robust management using renewable energies**



# D3.2: How to aggregate information

Author: S. Caux

Version: 1.1 (11/06/2018)

Date: D3.2 T0+33 : june 2018

**Deliverable information**

| | |
|---|---|
| Deliverable number | D3.2 |
| Contractual date of delivery | 30/03/2018 (~~M30~~ M33) |
| Actual date of delivery | 30/06/2018 |
| Title of deliverable | How to aggregate information |
| Dissemination level | Restricted to consortium and ANR members |
| WP contributing to the deliverable | WP3 |
| Author | S. Caux |
| Co-authors | R Roche, D Hissel, P Stolf, G DaCosta, A Sayah, JM Pierson, L Philippe, JM Nicod, G Rostirolla, B Celik |

**Revisions**

| Version | Date | Author | Comments |
|---|---|---|---|
| 0.1 | 01/06/2017 | S. Caux | Skeleton |
| 1.0 | 01/10/2017 | S Caux, A Sayah | Link to 3.1, formats, two aggregation levels and other WP links |
| 1.1 | 24/05/2018 | G Rostirolla, J Lecuivre, R Roche | Document review |

**Abstract**

The aim of this deliverable is to give recommendations on :

- Specifying the links and interactions between the different modules in the system, specify the format of data exchanged between the different modules recalling D3.1 information.
- Defining the time and space aggregations, pertinent indicators to exploit data linked to scenarios of WP5 requesing Sources and IT models at system level for different scenarios at specific scales

As a whole, this document provides an overview of the operation of the system. The detailed algorithms of the modules will be further studied in WP4. This document complements the D3.1.

**Keywords**

Data aggregation, Time scale, System level, Module interactions

# Table of contents

**1. Global System Structure** 5

1.1 General Context: 5

1.2 Objectives

**2. Spatial Scale Aggregation** 6

2.1. Elements and Group of Elements at Power Side 6

PS infrastructure description 7

PS PDM activities 8

2.2. Elements and Group of Elements at IT Side 8

IT infrastructure description 8

ITDM activities 9

**3. Time Scales aggregation** 9

**4. Data format** 11

4.1 DataZero Information System 11

4.2 GUI System 11

4.3 Data, Messages and Aggregation 11

**5. Conclusion** 13

**Acronyms**

| | |
|---|---|
| AC | Alternating current |
| DC | Direct current |
| Cooling | Heat, ventilation and air conditioning |
| I | Current |
| IT | Information technology |
| ITDM | IT decision module |
| ITS | IT system |
| NM | Negotiation module |
| PDM | Power decision module |
| PS | Power system |
| RES | Renewable energy sources |
| UPS | Uninterruptible power supply |
| V | Voltage |

# 1. Global System Structure

## 1.1 General context:

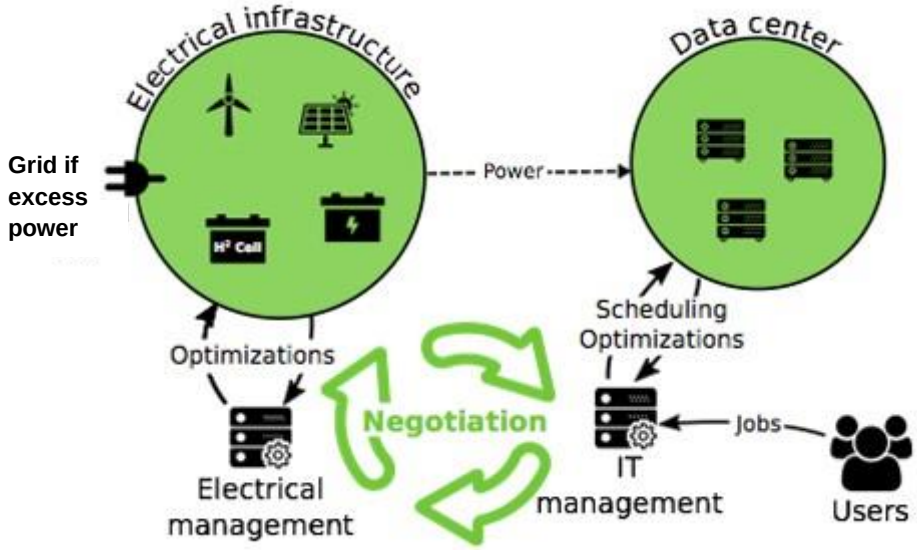


**Fig. 1: DataZero modules and interactions.**

The system includes several subsystems named '*modules*' as in D3.1 document. A synthetic overview of the messages exchanged between the different modules is provided in Fig. 1 of this document.

This part D3.2 is 'system relevant' and should propose some aggregated data also linked to Human Machine Interface, Web monitoring and Metrics indicators.

On this purpose:

- a specific part is isolated on the Power Sources side due to real time control (only the proxy part should send messages to PDM to know constraints and possible Power profiles)
- on the IT side, another specific part is dedicated to DCWoRMS, SimGrid or OpenStack software responsible for real time task placements (only external information are delivered to other modules, internal specific management stays inside)
- on the middleware side, ITDM, PDM and NM modules are able to send data, and subscribe to the message queues, requesting them for further processing.

## 1.2 Objectives:

1. Identify/List decision variables/indicators linking both levels (internal data to negotiation data through decision modules).
   - Define hierarchy based on spatial scale, datacenter/rack/processor and Renewable Energy Sources/Fuel Cells both linked to IT and Power sources sizing and architecture.
   - Define hierarchy based on time scale instantaneous/average/annual variables, time varying on different time steps for different objectives OPEX/CAPEX, Control, Ageing and so on.

2. Manage heterogeneity mainly due to time scales (several time window sizes and sampling rates). Separate data for offline (scheduling and GUI) from online decision (control, effective IT placement/effective power dispatching)

3. Define the data Aggregation/Format used by Control/Dispatching requests and communication protocol define specifically (eg: socket and proxy for control, JSON and ActiveMQ for Decision Modules, GUI/HMI requirements). All the power coming from several different electrical sources are summed in a DC bus, even if EATON infrastructure may be scaled with 4x275kW Uninterruptible Power Supply (UPS Eaton's module). Virtual Machine management is in charge of IT scheduling and placement on physical machines also driven by ePDU in several racks containing different kinds of machine technologies.

# 2. Spatial Scale Aggregation

From component to system. A complete description is provided in D3.1 document and D2.4 (IT, power and sources profiling). This is also expected to serve as a first step towards defining the algorithms running in each module, which will be further developed and completed in WP4.

## 2.1. Elements and Group of Elements at Power Side

The role of the Power System (PS) is to provide the electric power necessary to match the datacentre power demand, which includes IT load, cooling system load (HVAC ventilation and air conditioning). Each component is assumed to be controlled by a local controller in charge of:

- Measuring several parameters such as current, voltage, State of Charge, temperature and alerts;
- Receiving set points (i.e., commands in terms of current or power);
- Operating the component so it reaches these set points whenever possible.

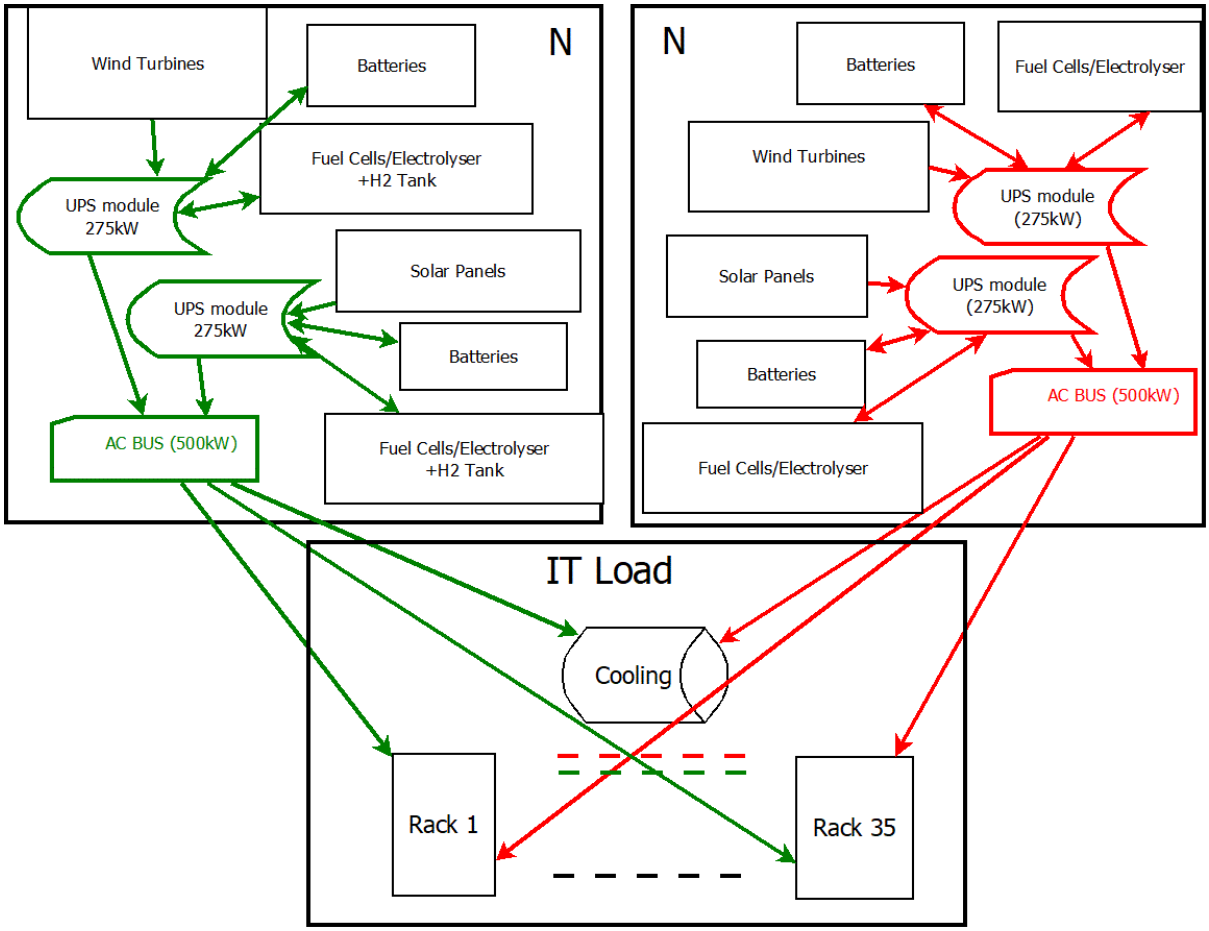


**Fig. 2: 2N Microgrid structure (cf D2.4 doc).**

## PS infrastructure description

As presented in Fig. 2, for the proposed 'breakthrough' architecture proposed in DataZero, the energy sources can be seen as either:

- The global power delivered to the DC bus, as a sum of all power sources, controlled in voltage and dispatched using EATON 9395 UPS modules;
- The power individually delivered by each kind of sources (battery stack, wind turbines farm and solar panels farm, Fuel Cell stack, etc.).

### Stack definition

Each kind of sources has to deliver a certain amount of power which is 'locally' composed by several 'elementary' components arranged in series and in parallel form to reach requested current and/or voltage (eg: solar panels farm made of several panels with several cells and local control, a battery stack arranged with batteries in series and parallel with local battery management system, etc.).

### Power source definition

In sources profiling (presented in D2.4 document), each source is characterized and behaviours are provided depending on time scales used, and can be generalized depending on the sizing defined.

### PS PDM activities

The power system (PS) is responsible for sources commitment, sources power and energy dispatching. It also sends and receives information coming from Negotiation Module and sends information to real time control systems responsible to respect these reference set points. We describe below what each system is responsible for:

#### PS control

Send set points (voltage, current) to each power source (PHIL or emulated)
Receive SoC, Power delivered, events (alarms, limitations…).

#### PDM references

Optimize the source commitments according to the global power envelope to deliver,
Send each power references.
Provide several power production profiles to negotiation depending on the current SoC, LOH and weather predictions,

#### Datacenter PS level

Control sources status (on/off) and power delivered one by one, and also globally (sum of each kind of sources) in a requested time window.
Allow to have aggregations by type of sources (Wind, solar …) and by unit source.

## 2.2. Elements and Group of Elements at IT Side

The 1MW (2N of 2x500kW) datacenter proposed is composed of racks of computing nodes arranged in the datacenter infrastructure. The power consumed is monitored:

- At each electrical node (UPS EATON);
- At each connection, an ePDU allows to measure and act on each server;
- At each processor and its cores, and also tasks associated with a CPU percentage and memory used (converted to electrical power consumption).

The IT elements (tasks, servers) are monitored and managed in OpenStack or simulated in DCWoRMS or SimGrid software. This IT side is made of racks containing machines based on various processor and memory technologies. Jobs, tasks and phases (for simulation and VM management) analysed in D2.4 deliver the information on their execution time and placement and allow to compute power consumption under power availability constraints. Connected to the ActiveMQ message BUS, the IT side also has internal information such as historic and prediction on job files, migration and so on.

### IT infrastructure description

In order to clarify the datacenter infrastructure, which is composed of several racks, containing machines of various technologies, in the following items we present the definition of machine and rack.

Machine definition

A machine (computer) contains several cores and memory capacity. It also includes frequency capabilities and the corresponding power consumption characteristics, along with those for on/off and migration states. All these elements are known and provided by IT and servers profiling (D2.4).

Rack definition

A rack is composed by several machines corresponding to the datacenter internal architecture defined. The rack power consumption is the sum of the machines dynamic and idle power consumptions, also described in D2.4.

## ITDM activities

Jobs are managed and machines are allocated constituting the IT activity which is also monitored and metrics are computed at the IT system level. Messages are thus regularly exchanged (observed then memorised for historic data and/or predicted based on models). We present below the messages concerning the IT jobs and resources:

Job

- Job and new job status (stop, active, migrating, finished…) taken into account in ITDM.
- Job placement information.
- Power consumption due to its own activities on a given machine (CPU, memory, etc.).

IT resources

- **machine level**
  Machine activity depends of its own status (boot, shutdown, waiting, working), processor frequency, used cpu percentage and thus power consumption.
- **rack level**
  Rack activity is defined by its own on/off states and also consumption of the machines it contains. Job status and internal statistics on jobs realised are also available metrics that can be observed (stored by log process and sent to GUI and/or monitoring interface).
- **Datacenter IT level**
  The global activity contains an aggregation of all racks included in the datacenter architecture.

# 3. Time Scales aggregation

In the same way we can go from a component to the whole system, the time scale can also be increased. The chosen time scale depends on possible experiments and scenarios (defined in WP5). As presented in Fig. 3, the time can be:

- sampling time of 1s to show IT tasks and Power commitment during few minutes;
- sampling time of 1h to show ITDM and PDM scheduling and dispatching during few days;
- sampling time of days/weeks: to show the overall indicators (PUE, QoS) for 1 year of the datacenter usage.

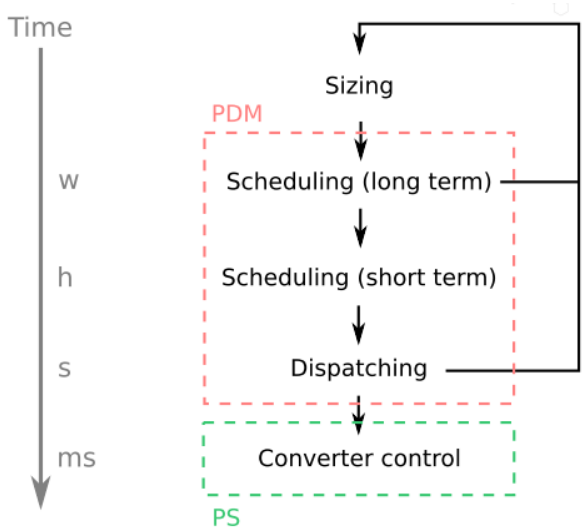


**Fig. 3: Time scale granularity.**

At this point it is also important to highlight that real time instantaneous models are involved in short time (*ms*) power control and IT management corresponding to a specific profiling, models in a given window with a given sampling time (several time granularity) correspond also to an average modelling. Finally cost and long term consideration correspond also to specific formulation and behaviour profiling presented in D2.4 in the models section.

Very short term messages have to be managed internally (OpenStack management, PHiL and proxy-based real time control) and do not provide a heavy data flow to exchange.
ITDM, PDM, NM have to exchange data observed/memorized or predicted, in a regular time interval which could be for instance every 15mn for optimization part, considering the scenario were the sampling time is in seconds and in the case of cost-to-use considerations every 1h/1day for the study on optimization sizing part for example.
The messages exchanged contain this time-objectives information, IT and PS profiling provide different messages on this purpose. If electrical information is available every 15mn, at HMI-GUI level a time-aggregation corresponding to an average (probably maximum) computation has to be implemented to exploit data observed. IT side also provides machine, racks and global level status information as mentioned previously.

# 4. Data format

For control purposes, no aggregation is possible either on detailed IT placement or Power electronics management. This occurs due to sampling and on/off (duty cycle) switching time which is usually less than 1ms, similarly to the IT placement where tasks are not aggregated.

For decision and monitoring: for several seconds average data should be sent. This can be done for 1 minute to 1 day. For overall indicators specific data has to be computed and sent, also used for sizing optimization part, and to prove the Datazero overall efficiency.

## 4.1 DataZero Information System

DataZero Information System has 2 main objectives using 2 main modules:

- Store all messages sent at system level, allowing post-analysis, replay same sequences and extract data specific for other module usage (eg : Power values and indicator, IT infrastructure composition…), this function is made in "Log Message Process" task registering all message intercepted in the ActiveMQ message bus.
- Store in structured tables the system activities, to easily extract information for GUI interface. These information can be requested in the past, current time or predicted for the future. This task is called "System Activity Recorder Process" and use topics and the specific data concerned. In addition to the log table that records all exchanges, different dedicated SQL/MariaDB tables complement the DataZero Information System with keys and data extracted from ActiveMQ messages, to answer the various queries coming, in particular, from GUI interface (production of electrical sources, environment, and state of machines or racks …).

The Database selected is MariaDB. It is one of the most popular database servers in the world, made by the original developers of MySQL and guaranteed to stay open source.

## 4.2 GUI System

The Graphic User Interface system consists of a web interface built to monitor several parameters in the overall datacenter. It was developed using the Angular_4 framework, using data provided in form of a web service, with *http* requests. The data exchanged follows the JSON format previously defined.

## 4.3 Data, Messages and Aggregation

There are various data exchanged not based on the same time and/or with the same sampling time. All data are not necessary for all modules. For example, PHIL real time control values use short time with high frequency loops only using references provided at a lower frequency corresponding to an average behaviour and separating them is sufficient.

This average is a kind of temporal aggregation (the specific average-model is included in WP2 profiling and the same applies for OpenStack local management). A specific socket for PHIL and Matlab is used to link them to the communication BUS. Internal OpenStack loops

are not extracted and only system values are communicated. The same applies when using DataCenter simulation tools as IT simulator.

## Temporal aggregation point of view:

Only the data useful for other modules are present on ActiveMQ message BUS, with a minimal time step of 1mn roughly fixed. Some modules related to GUI and/or Optimization and Decision modules may ask information aggregation in a certain time window with a specific length (several min, days, weeks…). Data collected every minute can be summed and considered as an average behaviour for the corresponding time window.

## Structural/Spatial aggregation point of view:

At IT level, machine, rack and datacenter profiling send their own values corresponding to their behaviour and models (presented on WP2 IT profiling). At power level, unit sources (1 SolarPanel, 1 FuelCell), a group of same type (1000m2 solar panels, packs of batteries…), or all primary sources and storages elements data are just summed. If an application (GUI or Decision Module for instance) asks for a certain kind of aggregation, the corresponding sum will be made according to the sampling time, average and window size requested.

## Overall DataZero infrastructure aggregation:

IT and Power infrastructure (description.it, description.power) are described and known in specific messages (see Appendix document summarising JAVA classes, JSON format, Data structure messages, also linked to D3.1).

Topics and messages concerned, according to D3.1:

- Sizing, spatial scale description:
  - **description.it (**IT_DESC_DC, IT_DESC_MACHINE, IT_DESC_RACK)
  - **description.power (**ELEC_DESC_DC, ELEC_DESC_BATTERY, ELEC_DESC_WIND, ELEC_DESC_VOLTAIC, ELEC_DESC_GRID)
  - **description.datacenter (**DZ_DESC_DC)

- Optimization, time scale description:
  - **activity.it.resources**
  - **activity.it.jobs**
  - **activity.power.sources**
  - **activity.power.environment**

# 5. Conclusion

This document presents two aggregations requested by different objectives linked to offline optimization, online management and monitoring.

- Spatial aggregation is defined linked to IT and PS architecture allowing to distinguish nodes/processors, stacks/racks, groups of elements and datacenter IT/PS parts.
- Time aggregation is explained and also linked to time-dependent profiling defined in D2.4 and other work packages: annual optimization sizing, offline optimization in a given short time window, monitoring passed/actual/predicted data both in IT and PS parts.

The elements presented receive and/or send data on the ActiveMQ communication bus defined and structured in D3.1 document. These messages contain all available data useful to operate such aggregation on demand. This means GUI can do its own average computation for graphical representation if necessary, PDM/ITDM can also manage core elements or groups of elements (in the same way minute/hour/week) to reach their own accuracy and decision time, allowing to solve their optimization problem and formulation.