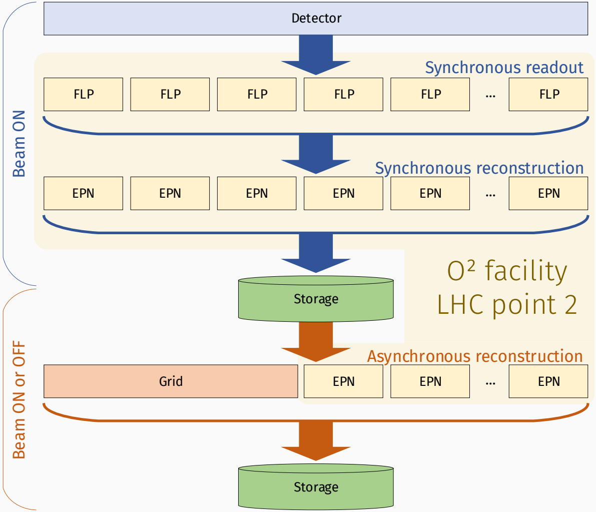
O^2 Control

A Control and Configuration System for the ALICE O^2 Facility

Teo Mrnjavac
CERN EP-AID-DA

November 21, 2018

The ALICE Online-Offline computing system



- Multiprocess **data flow and processing** framework
- **100,000s of processes,** ~2000 machines
- **Synchronous and asynchronous** (grid-like) workflows

O² Control: goals

"Just run some processes in a network…"

O² Control: goals

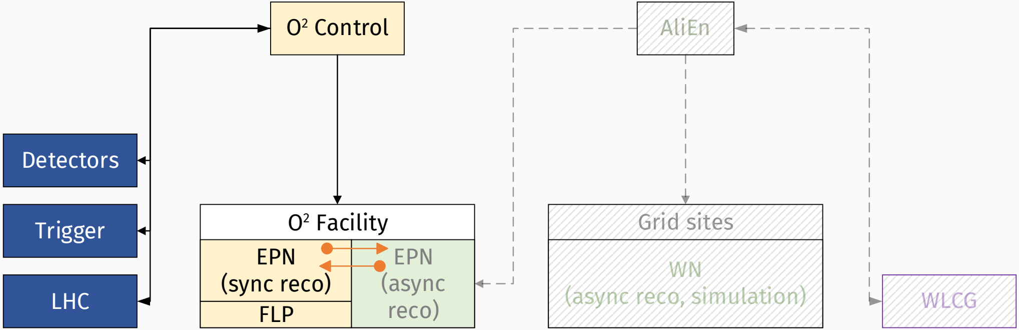
“Just run some processes in a network...”

- **Manage the lifetime** of thousands of processes in the O² facility:
    - allocation of cluster resources,
    - deployment, configuration and teardown of multiple workflows,
    - high degree of autonomy.
- **Minimize waste of beam time** by reusing running processes and avoiding restarts.
- Interface with LHC, trigger, DCS, bookkeeping and other systems.
- Ensure fair and efficient resource allocation between **synchronous and asynchronous** tasks.

O² Control: target improvements

- Improved flexibility & latency:
    - **no workflow redeployment** when excluding/including a detector from data taking,
    - **recover** from process and server crashes,
    - **reconfigure** processes without restart,
    - **scale** EPNs during data taking (e.g. as luminosity decreases in a fill).
- Next gen web-based GUIs with SSO & **revamped design**.
- Take advantage of modern developments in computing.

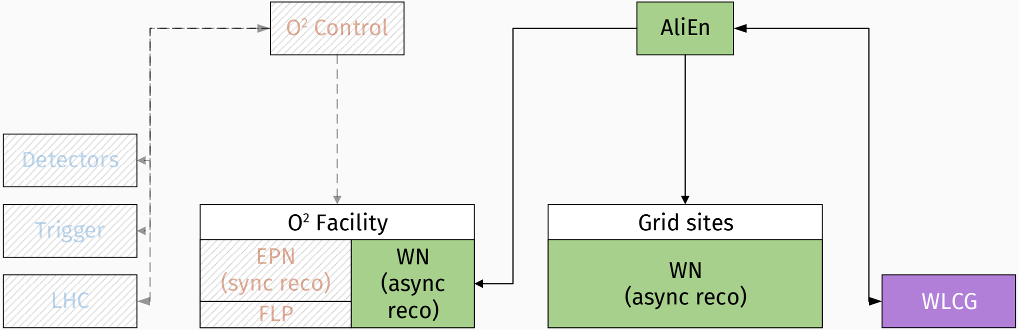
O² Control: synchronous operation



O² Control can mark a node as synchronous or asynchronous.

If a node is used for **synchronous** processing, O² Control stays in charge.

O² Control: asynchronous operation



When O² Control assigns a node to **asynchronous** operation, it launches a pilot job to set up a Grid-like asynchronous execution environment.

O² Control can reclaim these resources if necessary.

The requirements

- In order to satisfy the described use cases, O^2 Control:
  - is a **distributed system** in charge of the O^2 Facility, with full knowledge and control over its resources,
  - implements a reliable, distributed **state machine** mechanism to represent the aggregated state of the constituent O^2 processes of an O^2 workflow,
  - allows reconfiguration and reuse of running O^2 processes as often as possible and **avoids process restarts**,
  - allows simultaneous operation of multiple **asynchronous and synchronous** workflows, with easy reallocation of resources between them,
  - reacts promptly to user input, and handles events from LHC, trigger, detectors and the cluster itself with a high degree of **autonomy**.

Managing a cluster with Apache Mesos

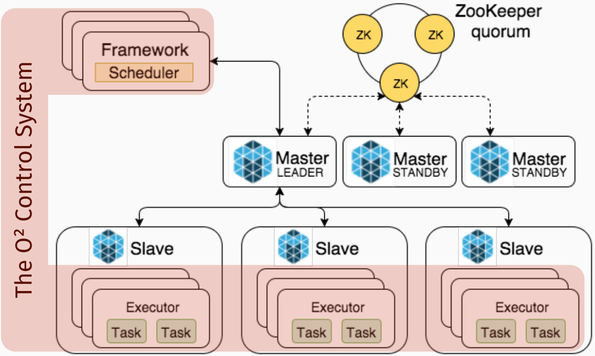
“Program against your datacenter like it’s a single pool of resources.”

Managing a cluster with Apache Mesos

"Program against your datacenter like it's a single pool of resources."

- We implement the **O² Control System** as a distributed application, using **Apache Mesos** as toolkit.
- Mesos acts as a unified **distributed execution environment** which streamlines how O² Control manages its components, resources and tasks inside the O² farm.

The Apache Mesos architecture



- Apache Mesos components on every host.
- Scales to **10,000s of nodes**.
- Open source, commercial support.
- Benefits for O² Control:
  - **knowledge** of what runs where,
  - **resource management** (ports, ...),
  - **transport** for control messages,
  - task event **notification** (dead, ...),
  - ...

A **framework**: a distributed application for Mesos, it has a **scheduler** and one or more **executors**.

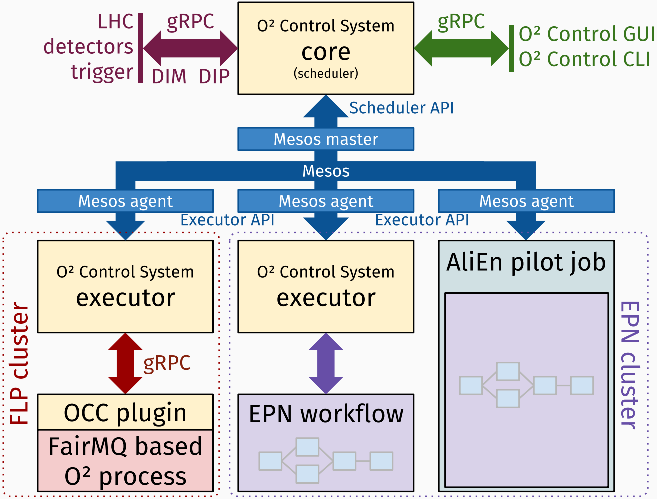
The Mesos **master** sends **offers** to the scheduler. Mesos **slaves** then deploy executors to run **tasks**.

O^2 Control components

`https://github.com/AliceO2Group/Control`

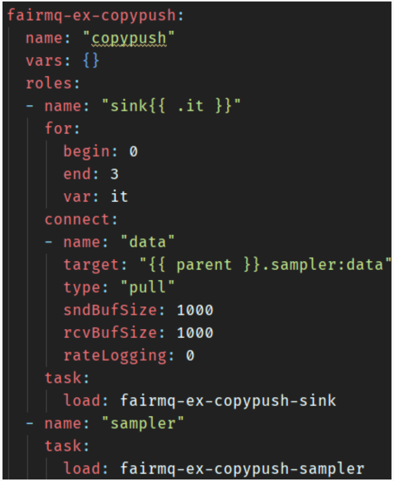
- O^2 Control currently (v0.1) consists of:
  - O^2 Control core (incl. Apache Mesos scheduler) GO MESOS GRPC
  - O^2 Control executor GO MESOS GRPC
  - O^2 Control and Configuration FairMQ plugin (`FairMQPlugin_OCC`) C++ FAIR GRPC
  - O^2 Control and Configuration library (`libOCC`) C++ GRPC
  - O^2 Control and Configuration CLI utility (`coconut`) GO GRPC
  - the web-based O^2 Control GUI node JS GRPC

O² Control overview



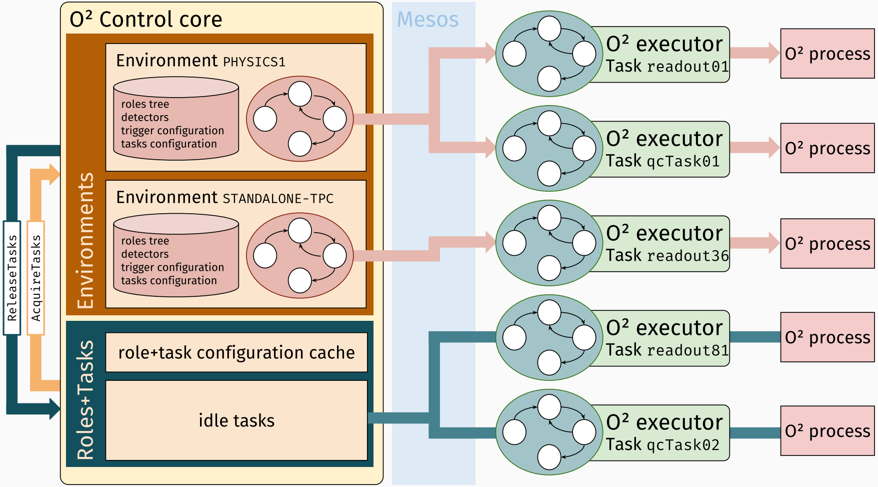
Workflows, roles and tasks

- Concepts:
  - **task** - the basic unit of control, generally 1 process
  - **role** - a node in the control tree, aggregates child roles and ultimately tasks
  - **workflow** - the in-memory control tree of an environment, made of roles which drive tasks
- Workflow templates generate workflows of tasks
  - Stored in O² Configuration (currently YAML, will switch to Consul backend)
  - Variables, iterators, internal references
  - Expanded into a workflow and associated with an environment

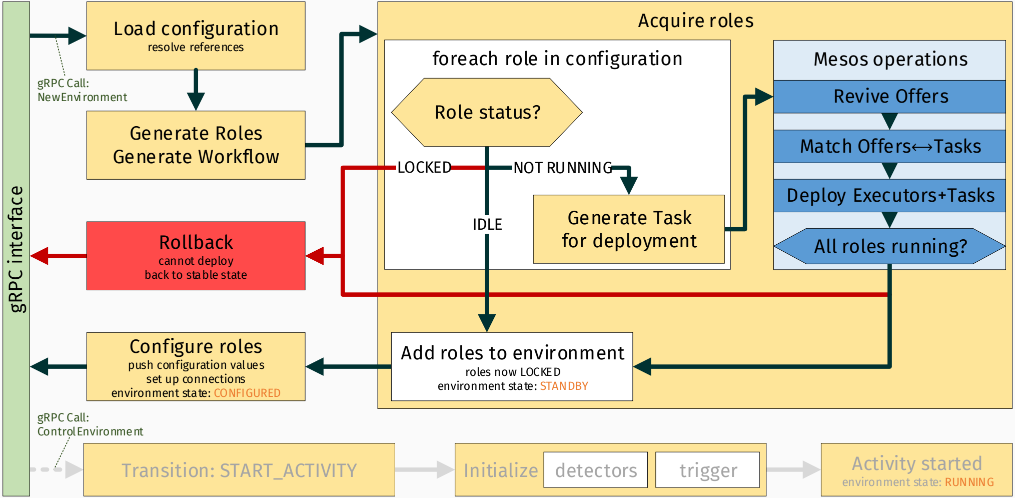


```yaml
fairmq-ex-copypush:
  name: "copypush"
  vars: {}
  roles:
  - name: "sink{{ .it }}"
    for:
      begin: 0
      end: 3
      var: it
    connect:
    - name: "data"
      target: "{{ parent }}.sampler:data"
      type: "pull"
      sndBufSize: 1000
      rcvBufSize: 1000
      rateLogging: 0
    task:
      load: fairmq-ex-copypush-sink
  - name: "sampler"
    task:
      load: fairmq-ex-copypush-sampler
```

O² Control role management

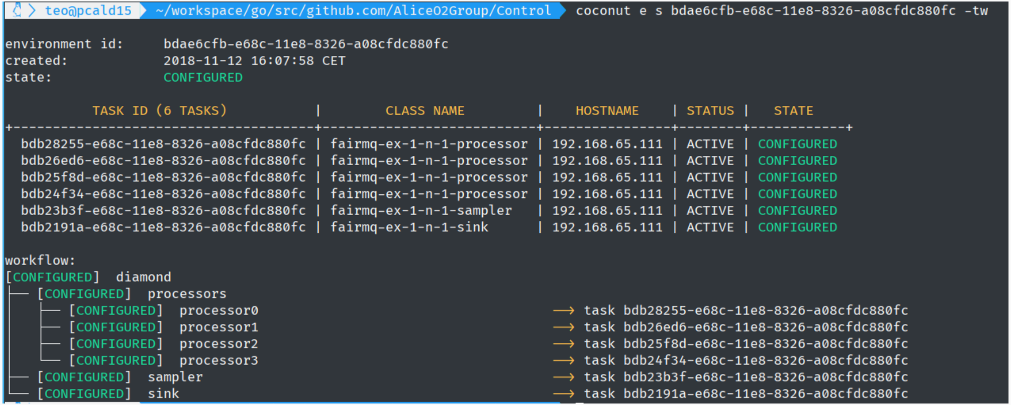


Example: create new environment



Environments in coconut

- coconut, the **co**ntrol and **con**figuration **ut**ility can be used to deploy, query and trigger transitions in environments.



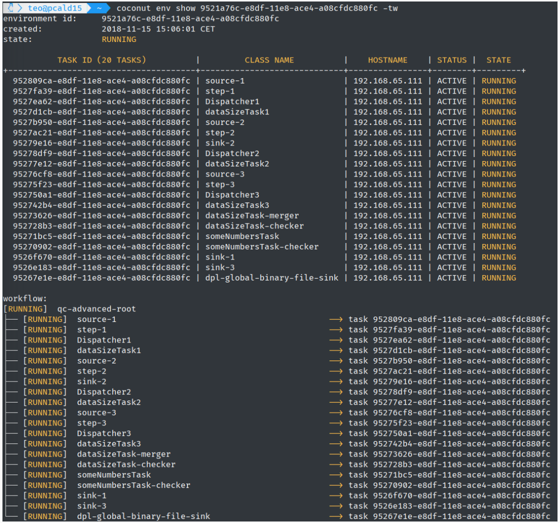
```
 > teo@pcald15 > ~/workspace/go/src/github.com/AliceO2Group/Control > coconut e s bdae6cfb-e68c-11e8-8326-a08cfdc880fc -tw

environment id:     bdae6cfb-e68c-11e8-8326-a08cfdc880fc
created:            2018-11-12 16:07:58 CET
state:              CONFIGURED

          TASK ID (6 TASKS)           |        CLASS NAME         |    HOSTNAME    | STATUS |    STATE
+-------------------------------------+---------------------------+----------------+--------+------------+
  bdb28255-e68c-11e8-8326-a08cfdc880fc | fairmq-ex-1-n-1-processor | 192.168.65.111 | ACTIVE | CONFIGURED
  bdb26ed6-e68c-11e8-8326-a08cfdc880fc | fairmq-ex-1-n-1-processor | 192.168.65.111 | ACTIVE | CONFIGURED
  bdb25f8d-e68c-11e8-8326-a08cfdc880fc | fairmq-ex-1-n-1-processor | 192.168.65.111 | ACTIVE | CONFIGURED
  bdb24f34-e68c-11e8-8326-a08cfdc880fc | fairmq-ex-1-n-1-processor | 192.168.65.111 | ACTIVE | CONFIGURED
  bdb23b3f-e68c-11e8-8326-a08cfdc880fc | fairmq-ex-1-n-1-sampler   | 192.168.65.111 | ACTIVE | CONFIGURED
  bdb2191a-e68c-11e8-8326-a08cfdc880fc | fairmq-ex-1-n-1-sink      | 192.168.65.111 | ACTIVE | CONFIGURED

workflow:
[CONFIGURED]  diamond
├── [CONFIGURED]  processors
│   ├── [CONFIGURED]  processor0                          ⟶ task bdb28255-e68c-11e8-8326-a08cfdc880fc
│   ├── [CONFIGURED]  processor1                          ⟶ task bdb26ed6-e68c-11e8-8326-a08cfdc880fc
│   ├── [CONFIGURED]  processor2                          ⟶ task bdb25f8d-e68c-11e8-8326-a08cfdc880fc
│   └── [CONFIGURED]  processor3                          ⟶ task bdb24f34-e68c-11e8-8326-a08cfdc880fc
├── [CONFIGURED]  sampler                                 ⟶ task bdb23b3f-e68c-11e8-8326-a08cfdc880fc
└── [CONFIGURED]  sink                                    ⟶ task bdb2191a-e68c-11e8-8326-a08cfdc880fc
```

Quality Control in coconut



```
teo@pcald15  ~  coconut env show 9521a76c-e8df-11e8-ace4-a08cfdc880fc -tw
environment id:     9521a76c-e8df-11e8-ace4-a08cfdc880fc
created:            2018-11-15 15:06:01 CET
state:              RUNNING

          TASK ID (20 TASKS)           |         CLASS NAME          |    HOSTNAME    | STATUS |  STATE
+---------------------------------------+-----------------------------+----------------+--------+---------+
  952809ca-e8df-11e8-ace4-a08cfdc880fc | source-1                    | 192.168.65.111 | ACTIVE | RUNNING
  9527fa39-e8df-11e8-ace4-a08cfdc880fc | step-1                      | 192.168.65.111 | ACTIVE | RUNNING
  9527ea62-e8df-11e8-ace4-a08cfdc880fc | Dispatcher1                 | 192.168.65.111 | ACTIVE | RUNNING
  9527d1cb-e8df-11e8-ace4-a08cfdc880fc | dataSizeTask1               | 192.168.65.111 | ACTIVE | RUNNING
  9527b950-e8df-11e8-ace4-a08cfdc880fc | source-2                    | 192.168.65.111 | ACTIVE | RUNNING
  9527ac21-e8df-11e8-ace4-a08cfdc880fc | step-2                      | 192.168.65.111 | ACTIVE | RUNNING
  95279e16-e8df-11e8-ace4-a08cfdc880fc | sink-2                      | 192.168.65.111 | ACTIVE | RUNNING
  95278df9-e8df-11e8-ace4-a08cfdc880fc | Dispatcher2                 | 192.168.65.111 | ACTIVE | RUNNING
  95277e12-e8df-11e8-ace4-a08cfdc880fc | dataSizeTask2               | 192.168.65.111 | ACTIVE | RUNNING
  95276cf8-e8df-11e8-ace4-a08cfdc880fc | source-3                    | 192.168.65.111 | ACTIVE | RUNNING
  95275f23-e8df-11e8-ace4-a08cfdc880fc | step-3                      | 192.168.65.111 | ACTIVE | RUNNING
  952750a1-e8df-11e8-ace4-a08cfdc880fc | Dispatcher3                 | 192.168.65.111 | ACTIVE | RUNNING
  952742b4-e8df-11e8-ace4-a08cfdc880fc | dataSizeTask3               | 192.168.65.111 | ACTIVE | RUNNING
  95273626-e8df-11e8-ace4-a08cfdc880fc | dataSizeTask-merger         | 192.168.65.111 | ACTIVE | RUNNING
  952728b3-e8df-11e8-ace4-a08cfdc880fc | dataSizeTask-checker        | 192.168.65.111 | ACTIVE | RUNNING
  95271bc5-e8df-11e8-ace4-a08cfdc880fc | someNumbersTask             | 192.168.65.111 | ACTIVE | RUNNING
  95270902-e8df-11e8-ace4-a08cfdc880fc | someNumbersTask-checker     | 192.168.65.111 | ACTIVE | RUNNING
  9526f670-e8df-11e8-ace4-a08cfdc880fc | sink-1                      | 192.168.65.111 | ACTIVE | RUNNING
  9526e183-e8df-11e8-ace4-a08cfdc880fc | sink-3                      | 192.168.65.111 | ACTIVE | RUNNING
  95267e1e-e8df-11e8-ace4-a08cfdc880fc | dpl-global-binary-file-sink | 192.168.65.111 | ACTIVE | RUNNING

workflow:
[RUNNING]  qc-advanced-root
├── [RUNNING]  source-1                                           ⟶ task 952809ca-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  step-1                                             ⟶ task 9527fa39-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  Dispatcher1                                        ⟶ task 9527ea62-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  dataSizeTask1                                      ⟶ task 9527d1cb-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  source-2                                           ⟶ task 9527b950-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  step-2                                             ⟶ task 9527ac21-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  sink-2                                             ⟶ task 95279e16-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  Dispatcher2                                        ⟶ task 95278df9-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  dataSizeTask2                                      ⟶ task 95277e12-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  source-3                                           ⟶ task 95276cf8-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  step-3                                             ⟶ task 95275f23-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  Dispatcher3                                        ⟶ task 952750a1-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  dataSizeTask3                                      ⟶ task 952742b4-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  dataSizeTask-merger                                ⟶ task 95273626-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  dataSizeTask-checker                               ⟶ task 952728b3-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  someNumbersTask                                    ⟶ task 95271bc5-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  someNumbersTask-checker                            ⟶ task 95270902-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  sink-1                                             ⟶ task 9526f670-e8df-11e8-ace4-a08cfdc880fc
├── [RUNNING]  sink-3                                             ⟶ task 9526e183-e8df-11e8-ace4-a08cfdc880fc
└── [RUNNING]  dpl-global-binary-file-sink                        ⟶ task 95267e1e-e8df-11e8-ace4-a08cfdc880fc
```

- Example of a running workflow of DPL-based Quality Control tasks.
- The O² DPL (Data Processing Layer) has initial support for generating O² Control workflow templates.

First release: O² Control v0.1

- A tech preview release, with support for
    - multi-node workflows of FairMQ or DPL devices,
    - automatic port assignment,
    - runtime FairMQ device configuration via plugin.

Coming soon

- v0.2 including Control API and `coconut` improvements to enable progress on O² Control GUI.
- Further DPL integration.
- December 2018: InfoLogger integration, Control test cluster provisioning mechanism, run number generation.
- Early 2019: Consul, Bookkeeping, trigger, DCS integration.
- Also 2019: metrics collection, performance evaluation.