OKA Service Definition

Simulation is strategic, offering competitive advantages to industries and advancing scientific research. With the explosion of data and the rise of artificial intelligence, it is becoming an essential part of our daily lives. Efficiently running an HPC infrastructure is complex and often lacks the proper tools to track down and get insights on how the users are behaving and how the cluster is responding to the demand.

UCit has packaged its HPC and machine learning expertise into a software suite designed to help HPC systems operate more effectively: **OKA**.

OKA, the platform for efficient HPC.

OKA Core

Boost your On-Prem HPC efficiency

OKA Cloud

Boost your Cloud HPC efficiency

OKA Predict

Improve cluster efficiency thanks to AI-trained custom Predictors

License Analysis Module

Get full visibility on your HPC license usage

OKA features the right tool for each of your clusters' area of optimization, it is composed of 4 distinct and complementary products to address them all: **OKA Core**, **OKA Cloud**, **OKA Predict**, and the **License Analysis Module.**

OKA Core

OKA Core is the platform dedicated to On-Premise clusters - it is composed of 3 main features sets: Performance, Energy, and Financials. **When buying an OKA Core license, you get access to all functionalities and modules contained in these three feature sets.**

Performance

OKA Core

Right at the backbone of OKA, **OKA Core - Performance** provides an extensible platform that presents the state of your HPC infrastructure through simple and comprehensible dashboards. Whether you need high level KPIs to report the cluster usage, or low-level information to track down the origin of an issue, OKA gives you the right level of details to:

- **Easily identify areas of optimization for your clusters and** resources in one single HPC platform.
- **Quickly diagnose and understand issues** using platform-wide data cross-checks and zoom-ins.
- **Build advanced KPI reporting** with customizable and exportable dashboards.
- Now with **multi-cluster views** and **GPU metrics**.

OKA Core contains the following tools:

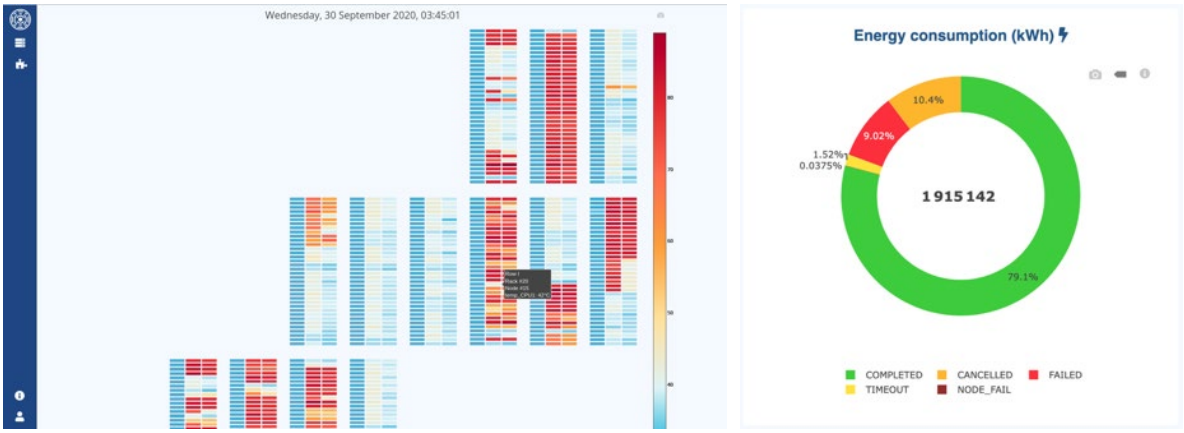
- **Job Status** – To spot wasted resources due to failed jobs (displays number of jobs and core-hours consumed per job status)
- **Load** – To understand overall resource allocation (displays allocated cores through time, and number of jobs allocated per node)
- **Throughput** – To analyze QoS and job submission patterns (displays submission frequency, slowdown, interarrival...)
- **Resources** – To determine jobs typologies and their consumption (displays number of cores & core-hours, memory and nodes consumed by the jobs)
- **Consumers** – To do advanced cross-checks for in-depth workload analysis (allows grouping of jobs per Group, User, JobName, Queue/Partition, QoS, Parallel Environment. For each, displays number of cores & core-hours, execution & waiting time, slowdown…)
- **Concurrent users** – To spot abnormal behaviors among users (displays active users per period)
- **Congestion/Contention –** provides a day-to-day update of the cluster status (Optimal, Acceptable, Contention, Congestion) based on resources needs and delivered computing power, and jobs life cycle for each day. It helps to identify if the cluster is correctly sized and configured, or if upgrades should be performed or if additional/external resources could be beneficial.

Job Status

OKA Core -Energy and -Financials complement OKA's -Performance features with cost and energy metrics.

OKA Core - Energy provides a view on your clusters' overall energy consumption and environmental impact. Appropriate configuration of a *Data Enhancer* can help evaluation of carbon footprint of various workloads:

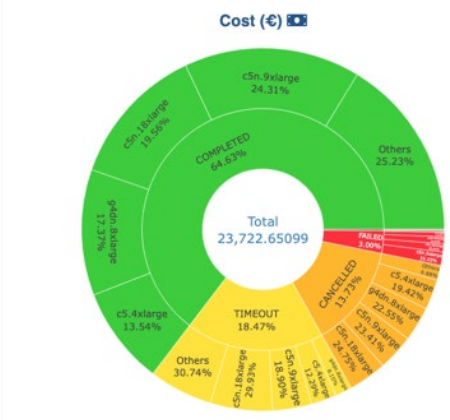
- **Energy** – OKA uses EAR (Energy Aware Runtime) and/or scheduler-reported energy data, when available, to report energy consumption at both cluster and job level.
- **Carbon** – Carbon footprint can be evaluated after setting up specific *data enhancers*, and provides carbon emission metrics at many levels.



Energy consumption analysis for the cluster and the jobs

OKA Core - Financials enhances OKA with precise and manipulable cost metrics. These cost metrics are available in most of OKA's toolsets and modules.

- **Cost propagation** – After setting up Costs information, this dimension becomes available for all OKA Core modules. You can either define the core-hour cost of your cluster, of have more fine-grained control on the way to report your costs (e.g., per queue and account) through the creation of a custom Data Enhancer.



Costs per job status

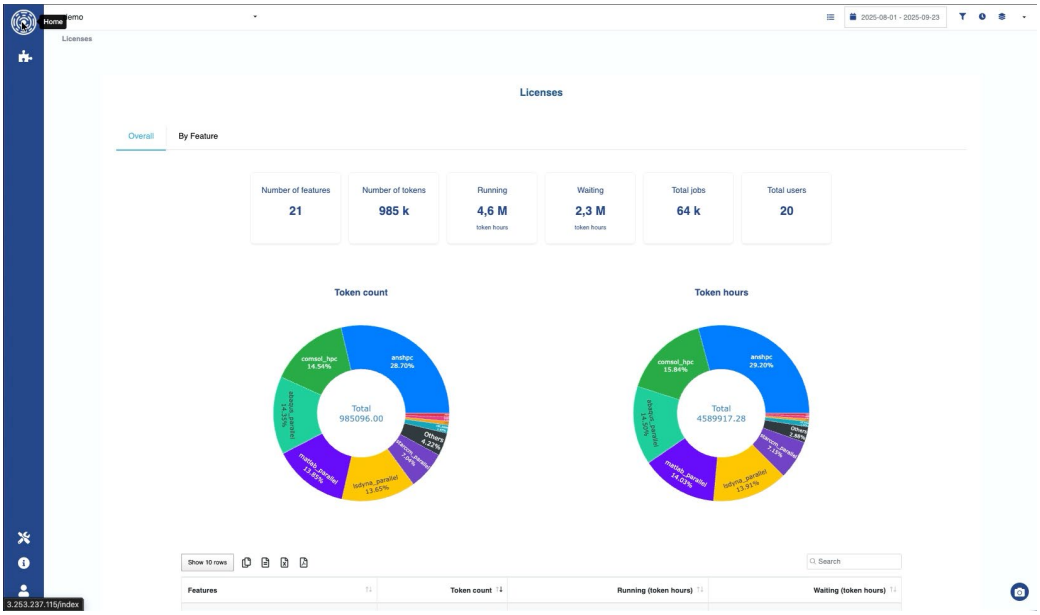
License Analysis Module

In many HPC environments, software licenses represent up to 30% of total cost of ownership. Yet license usage often remains opaque, poorly attributed, and difficult to act on. OKA's **License Analysis Module** brings clarity by turning scheduler-reported license data into actionable insights.

The module allows HPC teams to:

- **Analyze real license consumption as reported by job schedulers**, not estimates or static allocations.
- **Identify waste and bottleneck**s by spotting idle tokens, over-subscribed licenses, and blocked workloads.
- **Attribute exact license costs to business dimensions** such as teams, projects, users, workloads, or accounts, enabling accurate P&L allocation and chargeback models.
- **Track usage trends over time** to support license sizing, renewal decisions, and vendor negotiations.
- **Build clean, configurable KPI dashboards and export reports** for internal reviews, finance teams, or management reporting.

By making license usage transparent, measurable, and attributable, the License Analysis Module helps organizations reduce unnecessary spend while ensuring critical workloads get the licenses they need.

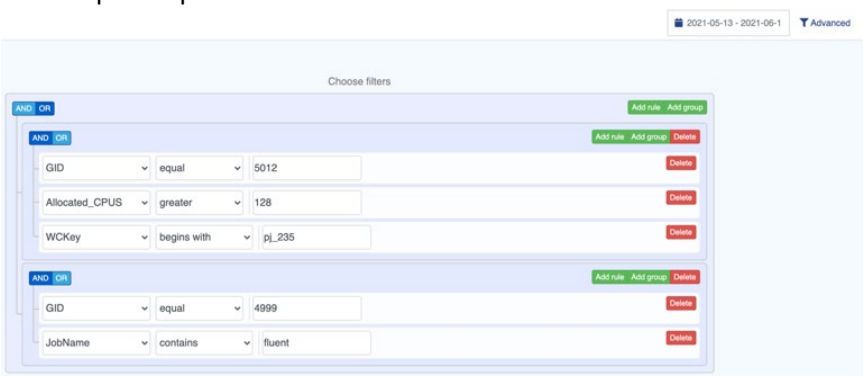


License Token dashboard

OKA's adaptability and flexibility

OKA is extensible, adaptable, and dynamic, and provides consistent interfaces and integrated tools:

- Each cluster view is accessible to authorized users and configurable through profiles.
- OKA's filtering capabilities allow to precisely select the workloads to analyze (for example, you could analyze separately all the jobs created by each of the research departments or concentrate on last summer's period where there have been strange job behaviors…).
- Commonly used filters can be saved and reused to quickly review the most important KPIs. It is even possible to train specific predictors on them.



OKA's advanced workload filtering

OKA can be extended and customized for your own uses-cases through Data Enhancers, which allow additional data sources to be plugged to OKA, such as the EAR database or hardware vendor tools which measure power and energy information or application specific logs.

You can discover more on OKA at https://oka.how or jump directly to https://doc.oka.how/ to understand how it works and how to start an evaluation.

OKA Cloud

OKA Cloud gathers OKA's functionalities in a specific package for Cloud Clusters.

Getting the most out of the flexibility of Cloud while containing costs is one of the primary concerns of Cloud HPC: but balancing performance and cost requires careful monitoring and adjustment of cluster configurations over time. How are your Cloud clusters used and how does it relate to your spendings? How can you make sure your resources are properly sized to ensure efficient operation while providing the right amount of flexibility for your fluctuating needs?

OKA Cloud provides an extensible platform that presents the state of your Cloud HPC infrastructure through simple and comprehensible dashboards. Whether you need high level KPIs to report the cluster usage, or low-level information to track down the origin of an issue, OKA Cloud gives you the right level of details.

OKA Cloud gives you all the tools to proactively address the biggest challenges of Cloud cluster management:

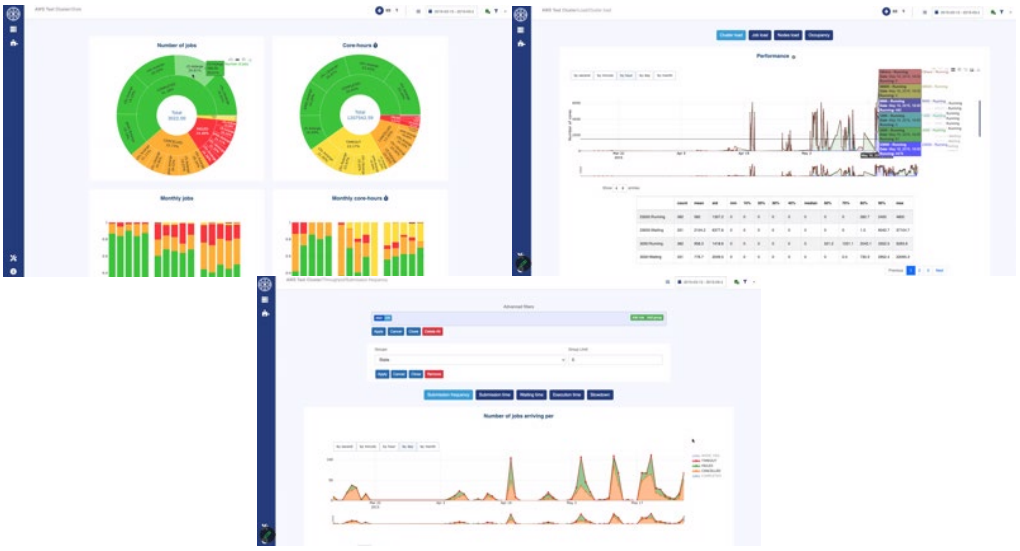
1. **Establish an insightful relationship between usage and spending**: examine resource consumption across the various machines and instance types you are using, but also across users, workloads, etc. Traditional AWS environments provide the costs associated with AWS Services but do not help identifying the workloads which could be optimized. OKA digs on the job scheduler logs to establish this relationship.

2. **Identify resource waste and its sources**: easily zoom in and out waste-related metrics to tackle sources of misuse (reserved vs used memory, congestion/contention states...). Pinpoint fully-reserved nodes that are only partially used, understand where the inefficiency lays (CPU, memory…) to match resource allocation to actual needs. AWS groups instances in "families" which provide the same Memory per Core ratio and job failures can result for insufficient memory allocation when a model becomes too large. Finding the best compromise is a source of direct savings.

3. **Understand and optimize the costs related to elasticity management** by maintaining a pool of resources available so that jobs (and the associated resources such as software licenses) don't wait too much for additional instances to start. Analyzing startup/shutdown + cooldown times helps to configure the baseline and proactively schedule resource boot accordingly.

Below, you will find a detail of all the available tools in OKA Cloud:

PERFORMANCE TOOLS: to analyze all essential cluster KPIs in one single, dedicated platform.



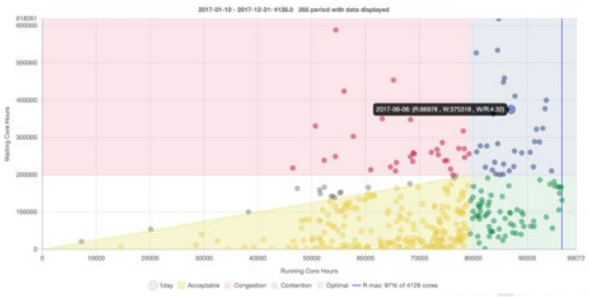
Snapshots of the State, Load and Throughput tools

- **State** – To spot wasted resources due to failed jobs (displays number of jobs and core-hours consumed per job status)
- **Load** – To understand overall resource allocation (displays allocated cores through time, and number of jobs allocated per node)
- **Throughput** – To analyze QoS and job submission patterns (displays submission frequency, slowdown, interarrival...)
- **Resources** – To determine jobs typologies and their consumption (displays number of cores & core-hours, memory and nodes consumed by the jobs)

Consumers – To do advanced cross-checks for in-depth workload analysis (allows grouping of jobs per Group, User, JobName, Queue/Partition, QoS, Parallel Environment. For each, displays number of cores & core-hours, execution & waiting time, slowdown…)

Concurrent users – To spot abnormal behaviors among users (displays active users per period)

Congestion/Contention – To visualize a day-to-day update of the cluster status (Optimal, Acceptable, Contention, Congestion) based on resources needs and delivered computing power, and jobs life cycle for each day. It helps to identify if the cluster is correctly sized and configured, or if upgrades should be performed or if additional/external resources could be beneficial.



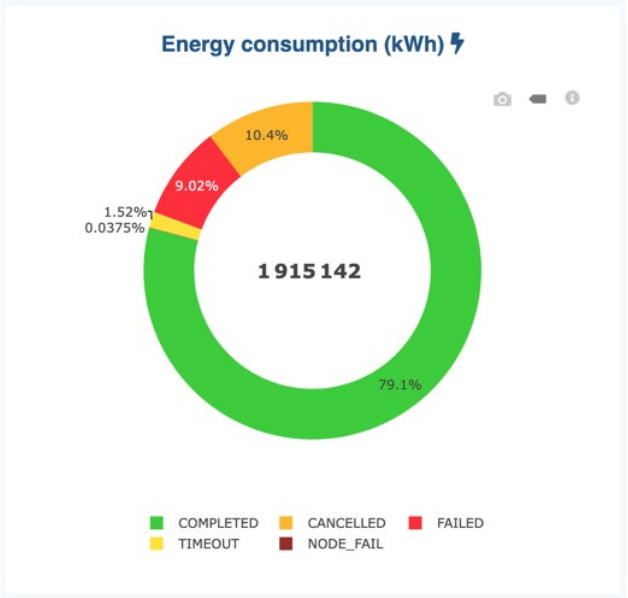
Snapshot of the Congestion/Contention quadrant plotting tool

ENERGY TOOLS: to get a view on your clusters' overall energy consumption and environmental impact. Appropriate configuration of a "Data Enhancer" can help evaluation of carbon footprint from the various workloads.

Energy – To observe the clusters/jobs energy consumption.

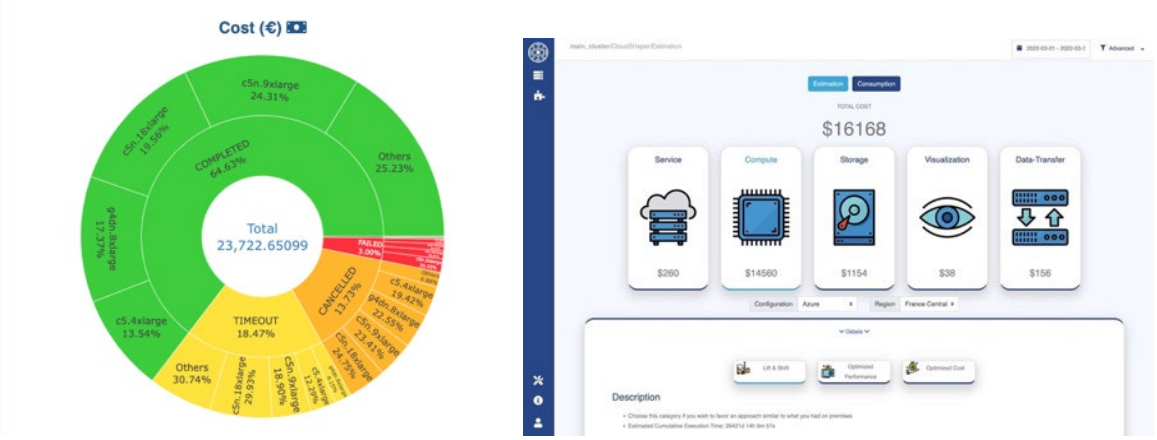
Carbon – To get a view of the carbon footprint of your clusters.

While Energy tools are less meaningful in a Cloud Environment where Data Center impact is integrated with the Services, deriving Carbon Footprint and improving wasted resources is becoming extremely important.



Energy consumption analysis for the cluster and the jobs

FINANCIAL TOOLS: to obtain precise & manipulable cost metrics and project the cost and size of your future hybrid HPC resources with dedicated shaper tools:



Snapshots of Cost and CloudSHaper

Cost propagation – You can either define the core-hour cost of your cluster, or have more fine-grained control on the way to report your costs (e.g., per queue and account, taking into account instance type…) through the creation of a custom Data Enhancer.

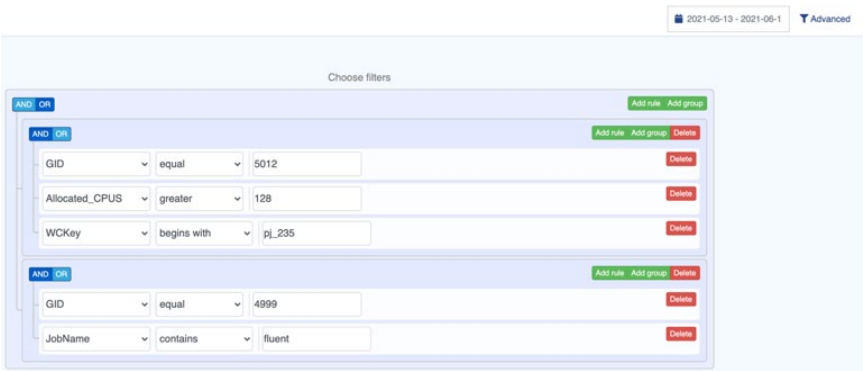
The following blog article https://ucit.fr/index.php/2023/08/08/integrating-the-oka-suite-with-your-aws-cluster/ describe a sample integration with Slurm for such creation. Similar integration can be done for other job schedulers such as LSF or PBS.

Note: OKA is primarily designed for Hybrid environments (on-premises, close-premises or public clouds) and some of the features are more relevant depending on the deployment context. For example, the Load and Congestion/Contention analyses were initially designed for on-premises HPC systems, where elasticity is physically constrained. Energy consumption is less a concern for Cloud Clusters, whereas Cost becomes a key driver in those environments.

OKA's adaptability and flexibility

OKA is extensible, adaptable, and dynamic, and provides consistent interfaces and integrated tools:

- Each cluster view is accessible to authorized users and configurable through profiles.
- OKA's filtering capabilities allow to precisely select the workloads to analyze (for example, you could analyze separately all the jobs created by each of the research departments or concentrate on last summer's period where there have been strange job behaviors…).
- Commonly used filters can be saved and reused to quickly review the most important KPIs. It is even possible to train specific predictors on them.



OKA's advanced workload filtering

OKA can be extended and customized for your own uses-cases through Data Enhancers, which allow additional data sources to be plugged to OKA, such as the EAR database or hardware vendor tools which measure power and energy information or application specific logs.

You can discover more on OKA at https://oka.how or jump directly to https://doc.oka.how/ to understand how it works and how to start an evaluation.

OKA Predict

OKA Predict takes your cluster usage optimization to the next level: allowing for real-time optimization of end-users' jobs' submissions. With OKA Predict, your team can build powerful Machine Learning predictors that will suggest or automatically apply the optimal job parameters to minimize resource use and time-to-results for end-users.

OKA Predict is a machine-learning tool to forecast jobs performance, costs and energy consumption. By integrating your job scheduler or submission portal with OKA Predict, you can:

- Improve Cluster productivity.
- Reduce waste of resources.
- Help End Users to get results Faster.

OKA Predict trains periodically on new data collected from the job scheduler or from additional logs. Its filtering functionalities make it possible to define workloads of interest to build specific predictors, and thus learn more precisely about the most frequent or most important cases. The submission parameters that OKA Predict includes as standard are the execution time and the RAM required for the job per node.
OKA Predict works in 3 phases, the training phase allows to build predictors, the prediction phase offers users parameters to specify for their jobs on the maximum execution time or on the memory required, or simply give feedback to users about their waiting or rendering time for tasks.

OKA Predict's Process

OKA Predict forecasts the following jobs' characteristics at submission time:

- State – detects the risk of a job failing or finishing in timeout.
- Execution Time – predicts the execution time of jobs to plan resources and get results faster.
- Memory – predicts how much memory should be requested.
- Waiting Time & Time to Result – get feedback on when your jobs will end.
- Energy – estimate the impact on the environment.