Bridging Cloud and HPC for Scalable Event-driven Processing of AI Workloads

Estíbaliz Parcero, Sergio Langarita, Germán Moltó

Instituto de Instrumentación para Imagen Molecular (I3M), Universitat Politècnica de València, Valencia, Spain.
esparig@i3m.upv.es

Bridging Cloud and HPC for Scalable Event-driven Processing of AI Workloads

- Bridging Cloud and HPC ⇦ **What** we want
- Scalable Event-driven Processing ⇦ **How** we run workloads
- AI Workloads ⇦ **Why** we need that bridge

Index

- AI workloads
- Bridging Cloud and HPC
- Scalable Event-driven Processing
- General Architecture
- Technical aspects of the integration
- The use case
- The workflow
- Results
- Conclusions

AI workloads

The number of cloud-native machine learning tools is rapidly increasing, offering more options and capabilities to AI engineers:

RAY
Flyte
H2O.ai
polyaxon
Kubeflow
MLRun
FEAST
KServe
mlflow
itwinai
TensorFlow Extended

AI workloads

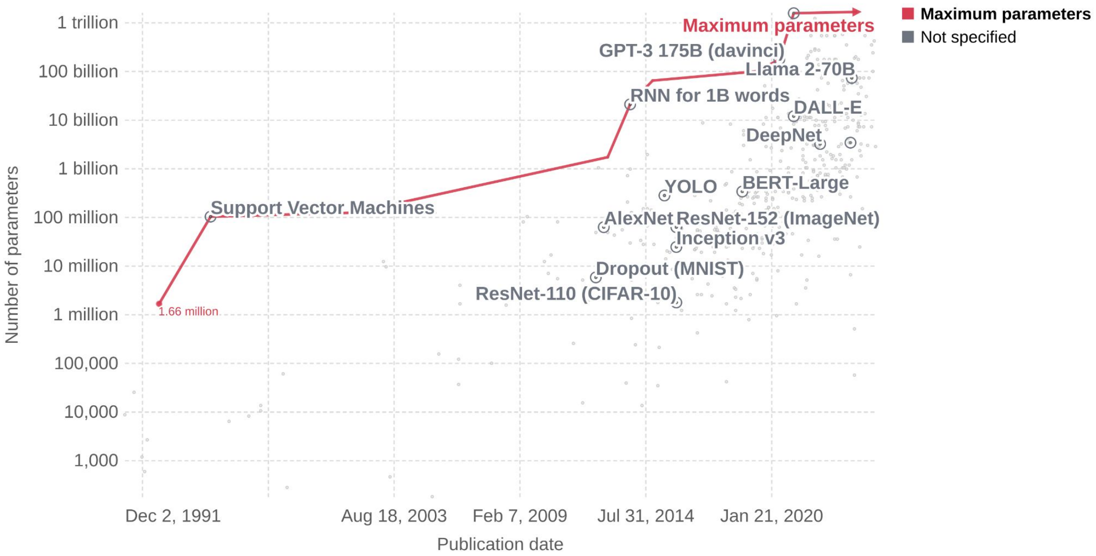
AI engineers use Cloud-native ML tools:

- **Accessibility and Cost-Effectiveness**: Provides necessary compute, storage, and services
- **Scalability**: Handles large datasets and complex models
- **Flexibility**: From programming languages, or pre-trained models to custom algorithms
- **Speed and Efficiency**: Speeds up building, training, and deploying models
- **Integration and Interoperability**: Integrates with other tools and services

AI workloads

Parameters in notable artificial intelligence systems

Parameters are variables in an AI system whose values are adjusted during training to establish how input data gets transformed into the desired output; for example, the connection weights in an artificial neural network.



Data source: Epoch (2024)

OurWorldinData.org/artificial-intelligence | CC BY

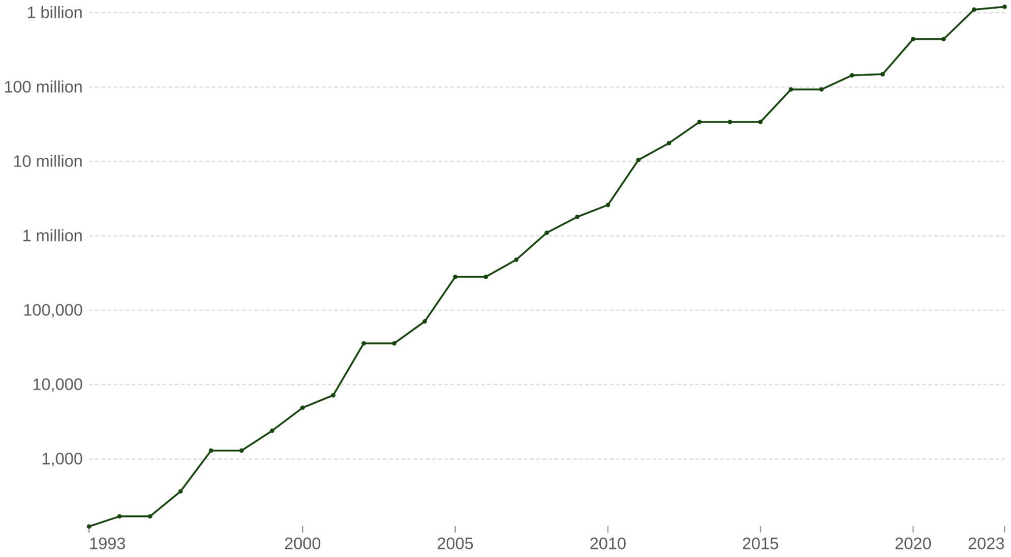
Note: Parameters are estimated based on published results in the AI literature and come with some uncertainty. The authors expect the estimates to be correct within a factor of 10.

- Exponential growth in the number of parameters
- Training tasks take a lot of resources
- Also Inference, eg:
  - LLMs
  - Diffusion models
  - GANs

AI workloads

Computational capacity of the fastest supercomputers

The number of floating-point operations[1] carried out per second by the fastest supercomputer in any given year. This is expressed in gigaFLOPS, equivalent to 10^9 floating-point operations per second.



Data source: Dongarra et al. (2023)

OurWorldinData.org/technological-change | CC BY

[1] **Floating-point operation**: A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

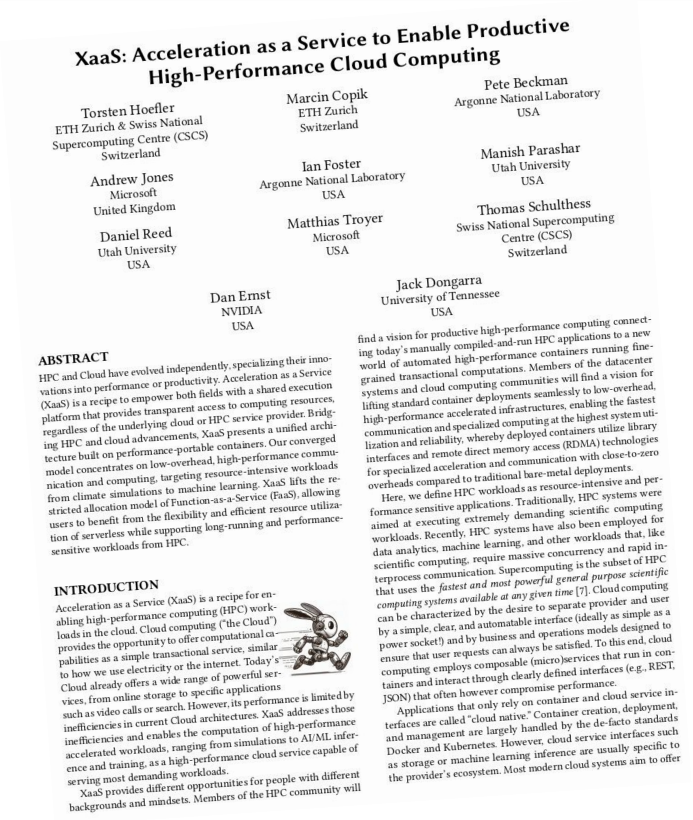
- Exponential growth in computational capacity
- Available on HPC Clusters

AI workloads

How to keep benefiting from Cloud ML tools and leverage the HPC resources?

Bridging Cloud and HPC seems a good idea

Bridging Cloud and HPC



XaaS: Acceleration as a Service to Enable Productive High-Performance Cloud Computing

Torsten Hoefler
ETH Zurich & Swiss National
Supercomputing Centre (CSCS)
Switzerland

Marcin Copik
ETH Zurich
Switzerland

Pete Beckman
Argonne National Laboratory
USA

Andrew Jones
Microsoft
United Kingdom

Ian Foster
Argonne National Laboratory
USA

Manish Parashar
Utah University
USA

Daniel Reed
Utah University
USA

Matthias Troyer
Microsoft
USA

Thomas Schulthess
Swiss National Supercomputing
Centre (CSCS)
Switzerland

Dan Ernst
NVIDIA
USA

Jack Dongarra
University of Tennessee
USA

ABSTRACT

HPC and Cloud have evolved independently, specializing their innovations into performance or productivity. Acceleration as a Service (XaaS) is a recipe to empower both fields with a shared execution platform that provides transparent access to computing resources, regardless of the underlying cloud or HPC service provider. Bridging HPC and cloud advancements, XaaS presents a unified architecture built on performance-portable containers. Our converged model concentrates on low-overhead, high-performance communication and computing, targeting resource-intensive workloads from climate simulations to machine learning. XaaS lifts the restricted allocation model of Function-as-a-Service (FaaS), allowing users to benefit from the flexibility and efficient resource utilization of serverless while supporting long-running and performance-sensitive workloads from HPC.

INTRODUCTION

Acceleration as a Service (XaaS) is a recipe for enabling high-performance computing (HPC) workloads in the cloud. Cloud computing ("the Cloud") provides the opportunity to offer computational capabilities as a simple transactional service, similar to how we use electricity or the internet. Today's Cloud already offers a wide range of powerful services, from online storage to specific applications such as video calls or search. However, its performance is limited by inefficiencies in current Cloud architectures. XaaS addresses those inefficiencies and enables the computation of high-performance accelerated workloads, ranging from simulations to AI/ML inference and training, as a high-performance cloud service capable of serving most demanding workloads.

XaaS provides different opportunities for people with different backgrounds and mindsets. Members of the HPC community will find a vision for productive high-performance computing connecting today's manually compiled-and-run HPC applications to a new world of automated high-performance containers running fine-grained transactional computations. Members of the datacenter systems and cloud computing communities will find a vision for lifting standard container deployments seamlessly to low-overhead, high-performance accelerated infrastructures, enabling the fastest communication and specialized computing at the highest system utilization and reliability, whereby deployed containers utilize library interfaces and remote direct memory access (RDMA) technologies for specialized acceleration and communication with close-to-zero overheads compared to traditional bare-metal deployments.

Here, we define HPC workloads as resource-intensive and performance sensitive applications. Traditionally, HPC systems were aimed at executing extremely demanding scientific computing workloads. Recently, HPC systems have also been employed for data analytics, machine learning, and other workloads that, like scientific computing, require massive concurrency and rapid interprocess communication. Supercomputing is the subset of HPC that uses the *fastest and most powerful general purpose scientific computing systems available at any given time* [7]. Cloud computing can be characterized by the desire to separate provider and user by a simple, clear, and automatable interface (ideally as simple as a power socket!) and by business and operations models designed to ensure that user requests can always be satisfied. To this end, cloud computing employs composable (micro)services that run in containers and interact through clearly defined interfaces (e.g., REST, JSON) that often however compromise performance.

Applications that only rely on container and cloud service interfaces are called "cloud native." Container creation, deployment, and management are largely handled by the de-facto standards Docker and Kubernetes. However, cloud service interfaces such as storage or machine learning inference are usually specific to the provider's ecosystem. Most modern cloud systems aim to offer

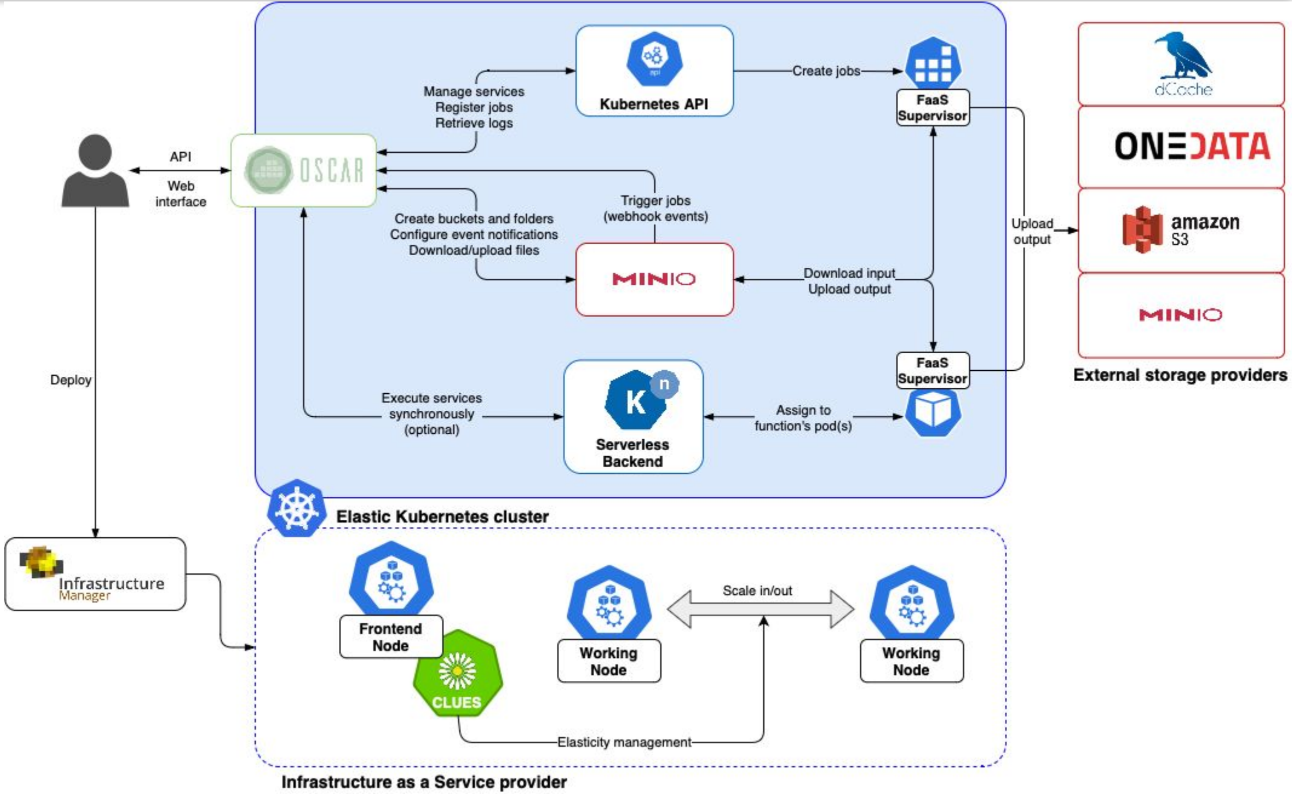
- Hoefler et al., 2024, highlight the importance of bridging Cloud and HPC for resource-intensive workloads like **climate simulations** or **machine learning** processing
- Some key ideas:
  - Leverage containers
  - Improve communication
  - Enable access to data (I/O)

Scalable Event-driven Processing

- **OSCAR**, an open source platform for serverless event-driven computing

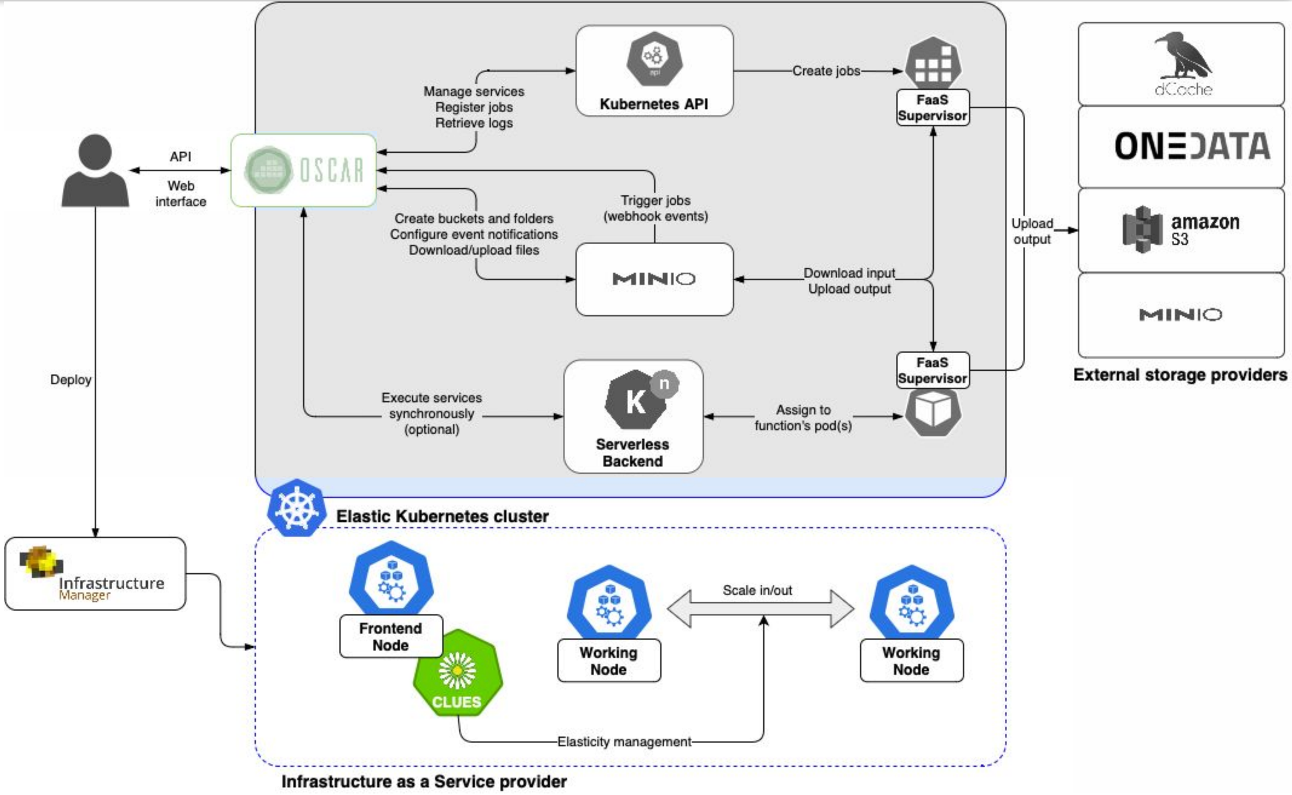
- **OSCAR** is a cloud-native tool, so all the previous benefits apply

https://oscar.grycap.net/



Scalable Event-driven Processing

Multi-Cloud Support:
Provision OSCAR clusters on on-premises, public, and federated Clouds

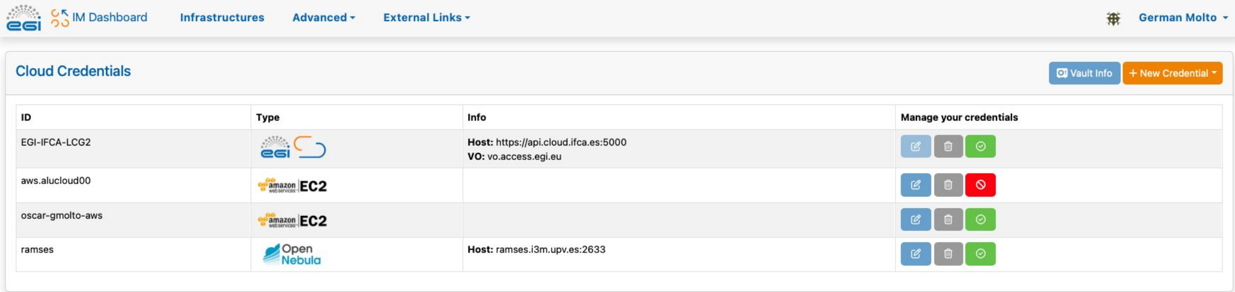


Scalable Event-driven Processing

Multi-Cloud Support:
Provision OSCAR clusters on on-premises, public and federated Clouds

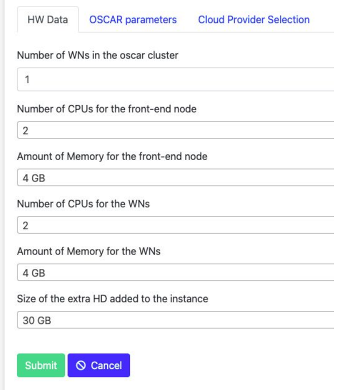
Easily deployed via **IM** web interface:

- Add infra credentials
- Configure (step by step) your elastic OSCAR cluster



Cloud Credentials

ID	Type	Info	Manage your credentials
EGI-IFCA-LCG2	egi	**Host:** https://api.cloud.ifca.es:5000 **VO:** vo.access.egi.eu	
aws.alucloud00	amazon EC2		
oscar-gmolto-aws	amazon EC2		
ramses	Open Nebula	**Host:** ramses.i3m.upv.es:2633	



Number of WNs in the oscar cluster: 1

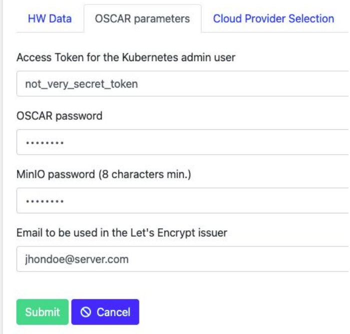
Number of CPUs for the front-end node: 2

Amount of Memory for the front-end node: 4 GB

Number of CPUs for the WNs: 2

Amount of Memory for the WNs: 4 GB

Size of the extra HD added to the instance: 30 GB

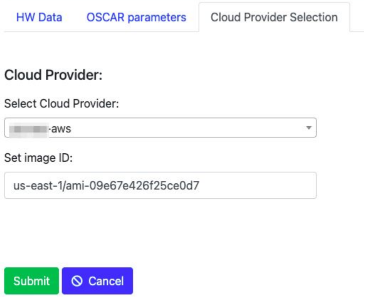


Access Token for the Kubernetes admin user: not_very_secret_token

OSCAR password: ••••••••

MinIO password (8 characters min.): ••••••••

Email to be used in the Let's Encrypt issuer: jhondoe@server.com



Cloud Provider:

Select Cloud Provider: [redacted]-aws

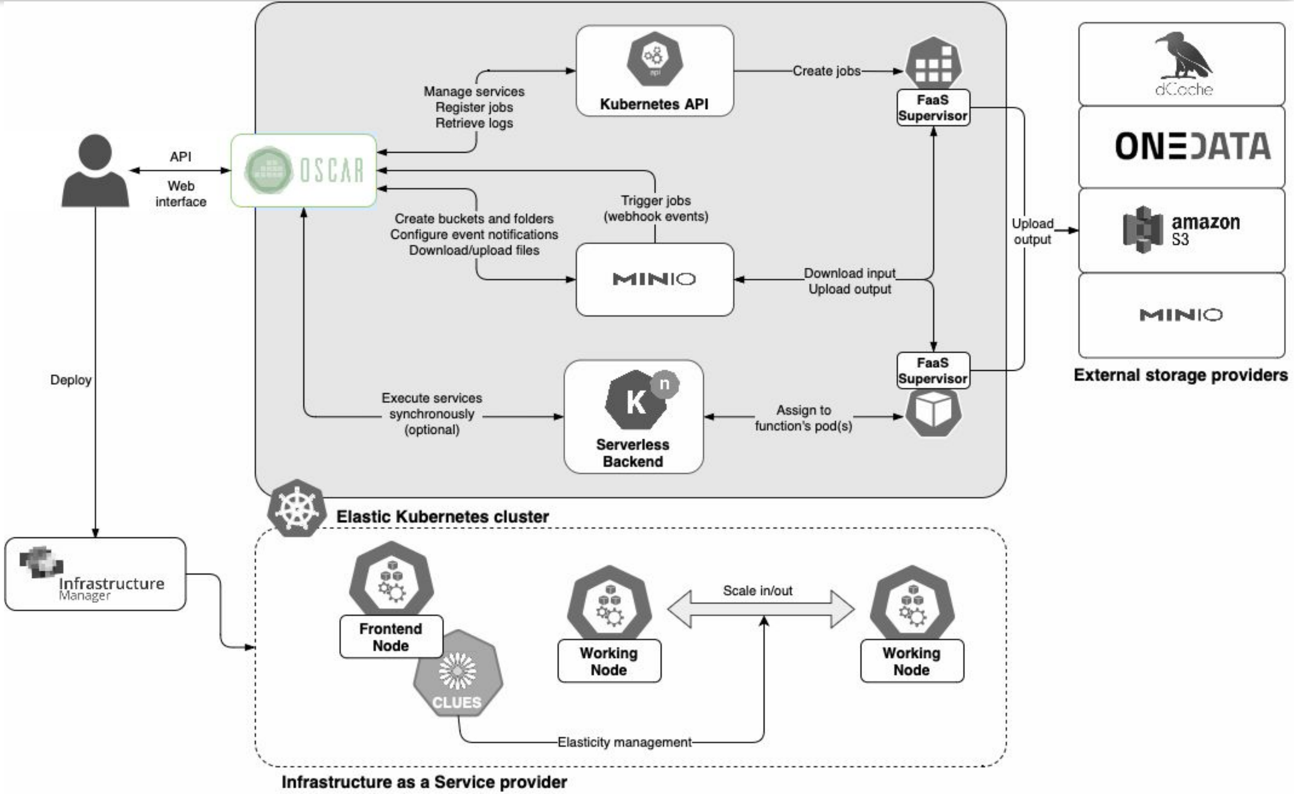
Set image ID: us-east-1/ami-09e67e426f25ce0d7

https://im.egi.eu

Scalable Event-driven Processing

Flexible Interfaces:

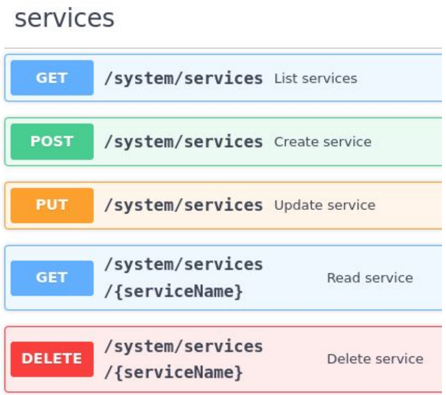
- API REST
- Web Interface
- Command Line Interface



Scalable Event-driven Processing

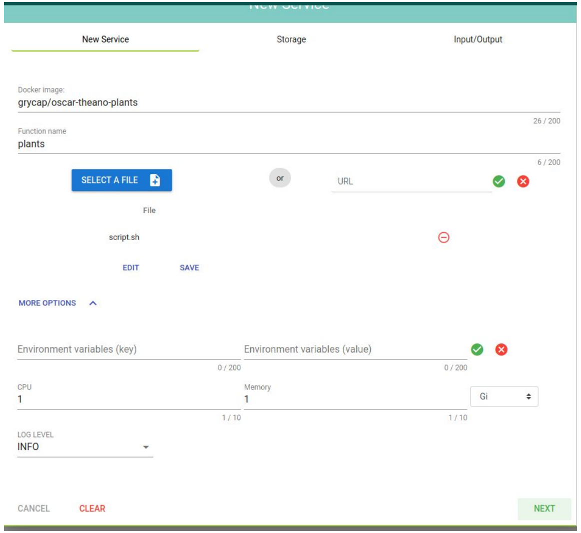
Flexible Interfaces:

- API REST
- Web Interface
- Command Line Interface



services

GET	/system/services	List services
POST	/system/services	Create service
PUT	/system/services	Update service
GET	/system/services/{serviceName}	Read service
DELETE	/system/services/{serviceName}	Delete service



Docker image: grycap/oscar-theano-plants 26 / 200

Function name plants 6 / 200

script.sh

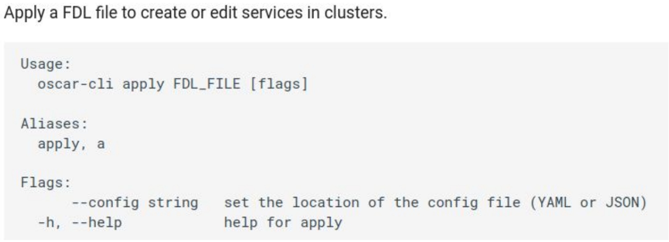
Environment variables (key) 0 / 200

Environment variables (value) 0 / 200

CPU 1 1 / 10

Memory 1 1 / 10 Gi

LOG LEVEL INFO



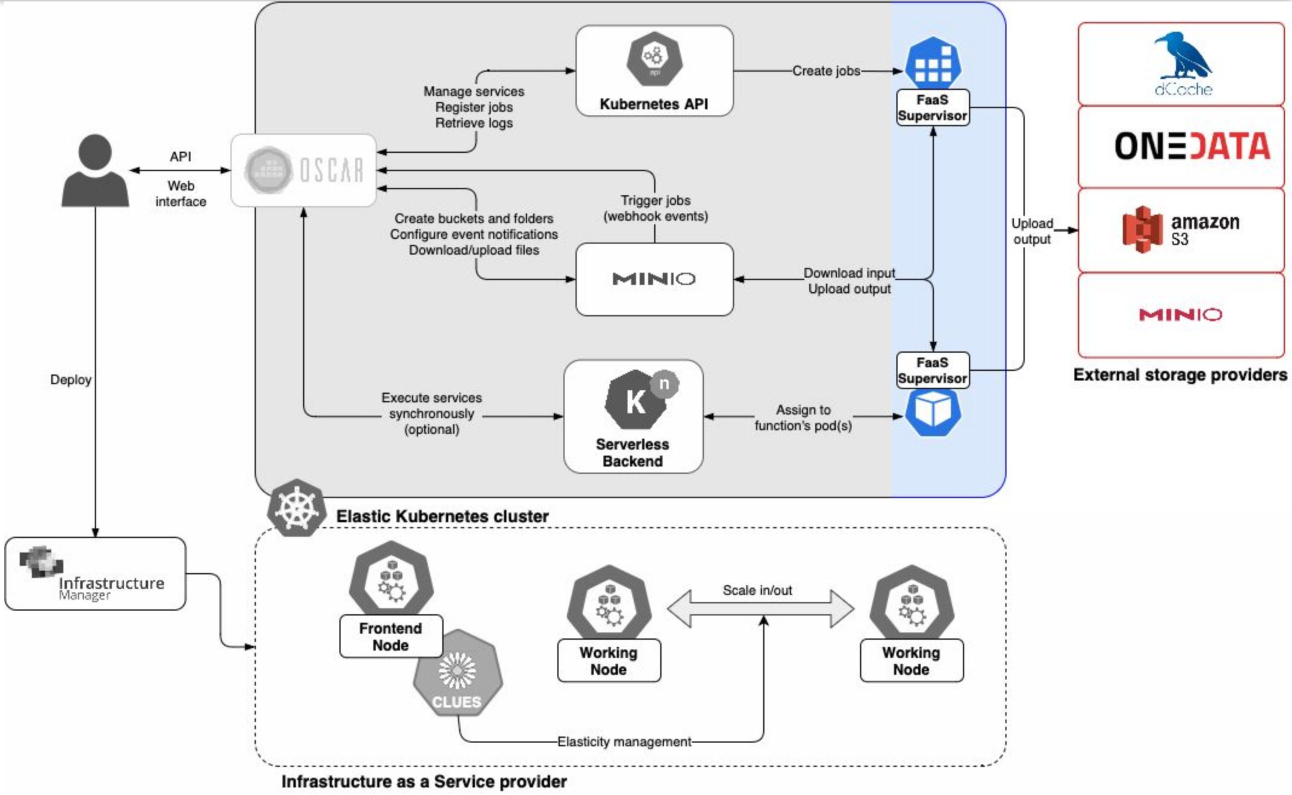
Apply a FDL file to create or edit services in clusters.

```
Usage:
  oscar-cli apply FDL_FILE [flags]

Aliases:
  apply, a

Flags:
      --config string   set the location of the config file (YAML or JSON)
  -h, --help            help for apply
```

Scalable Event-driven Processing

Support for Multiple Storage Back-ends:

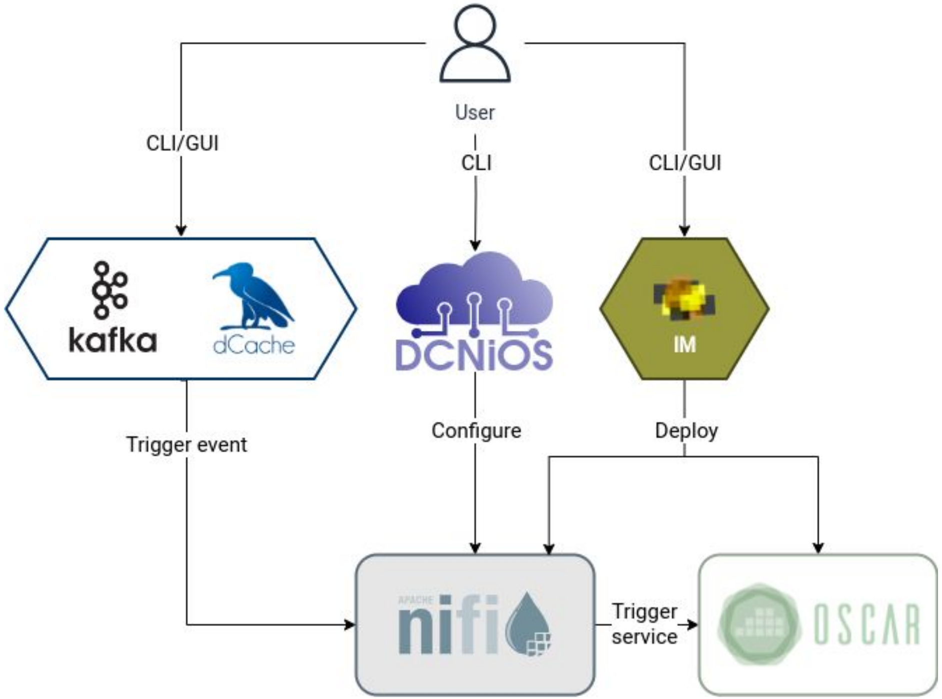
- MinIO, Amazon S3, OneData…



Scalable Event-driven Processing

Support for Multiple Storage Back-ends:

- MinIO, Amazon S3, Onedata…
- Any Storage Service using WebDAV protocol via DCNiOS

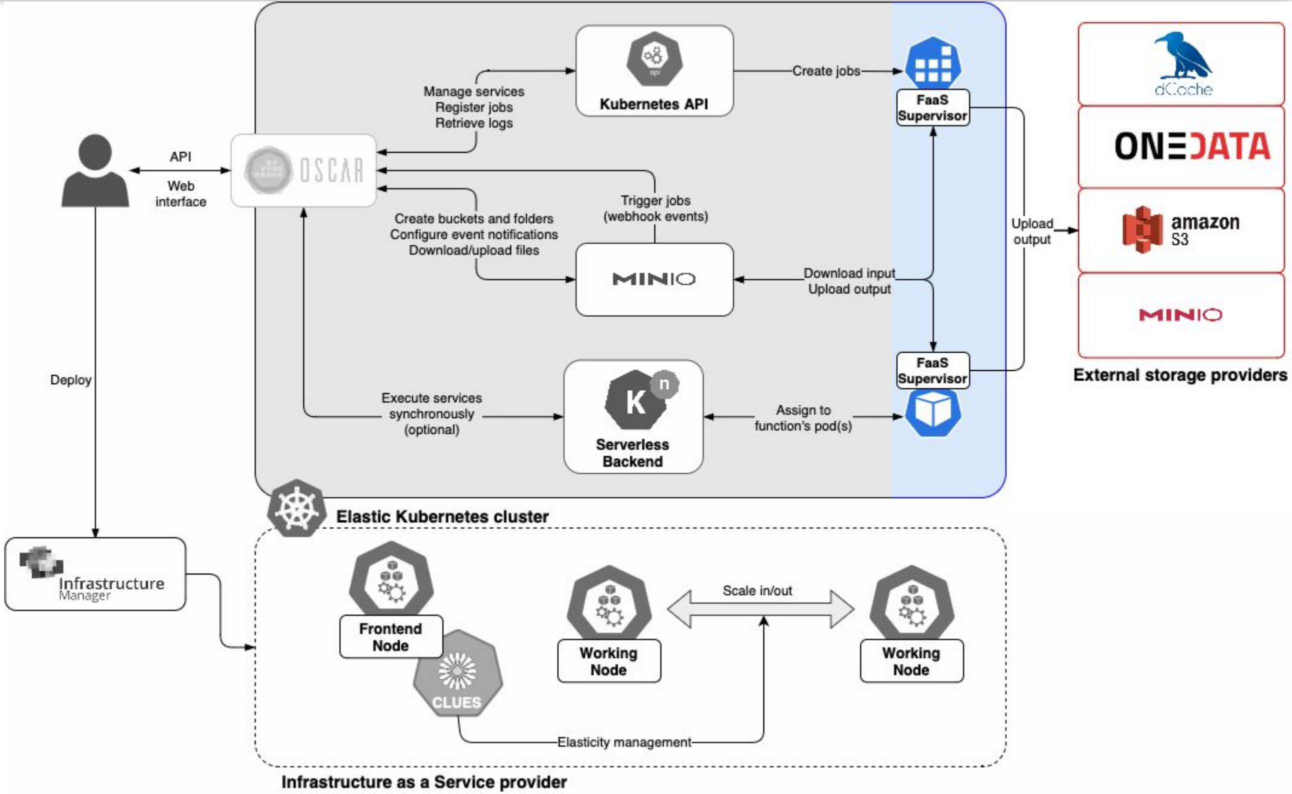


https://github.com/interTwin-eu/dcnios

Scalable Event-driven Processing

Support for Multiple Storage Back-ends:

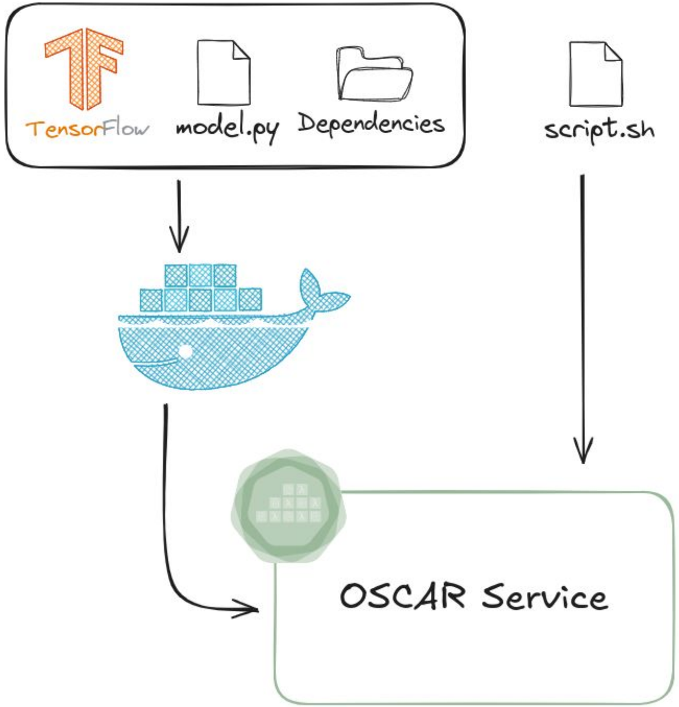
- MinIO, Amazon S3, Onedata…
- Any Storage Service using WebDAV protocol via DCNiOS
- Decoupled FaaS Supervisor is the I/O manager



Scalable Event-driven Processing

Container-based services:

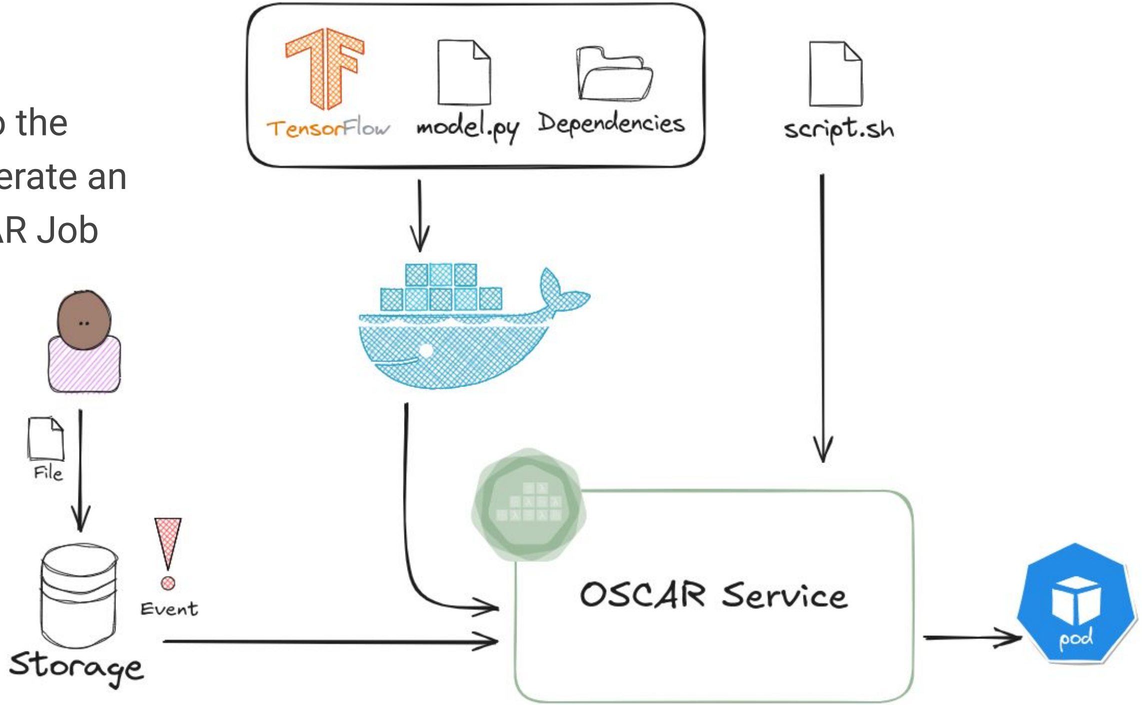
- Customized runtime environments



Scalable Event-driven Processing

Event-driven processing:

- When the user uploads data to the selected storage, that will generate an event that will trigger an OSCAR Job

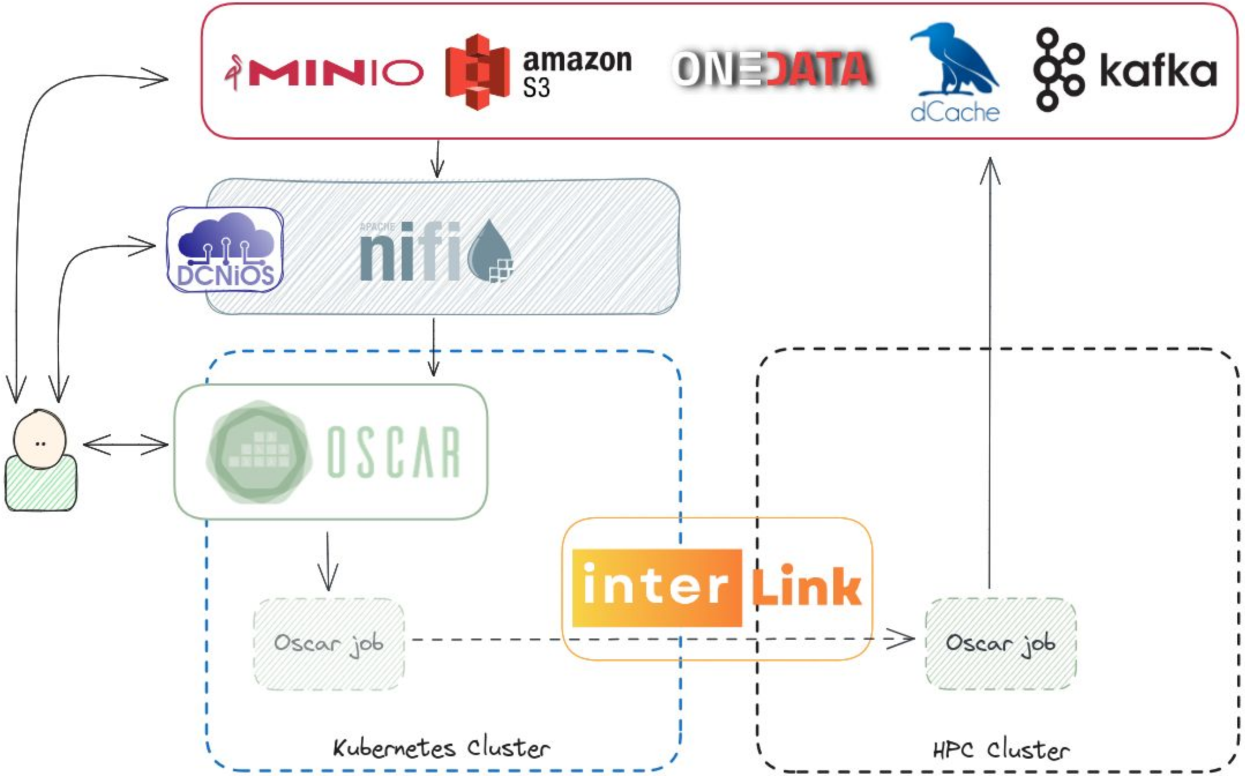


General Architecture

interLink:

- Developed by INFN, provides an abstraction for the execution of a Kubernetes pod on any remote host supporting containers
- Provides a gateway from OSCAR to delegate pod executions into HPC supercomputers

https://github.com/interTwin-eu/interlink



Technical aspects of the integration

Changes in OSCAR for the integration with interLink:

- The pod needs to be assigned to the Virtual Kubelet node so interLink can offload it (via node selector)
- The pod needs to include some annotations for SLURM (HPC system)
- When running in the Cloud, the FaaS Supervisor is mounted as a volume in the pod. To be able to be exported by interLink to HPC, the FaaS Supervisor is passed as a sidecar
- Environment variables in the pod have to be codify (base64) for interLink to be able to export them
- Once in HPC, decodify the environment variables to run the job

All these changes are transparent to the user

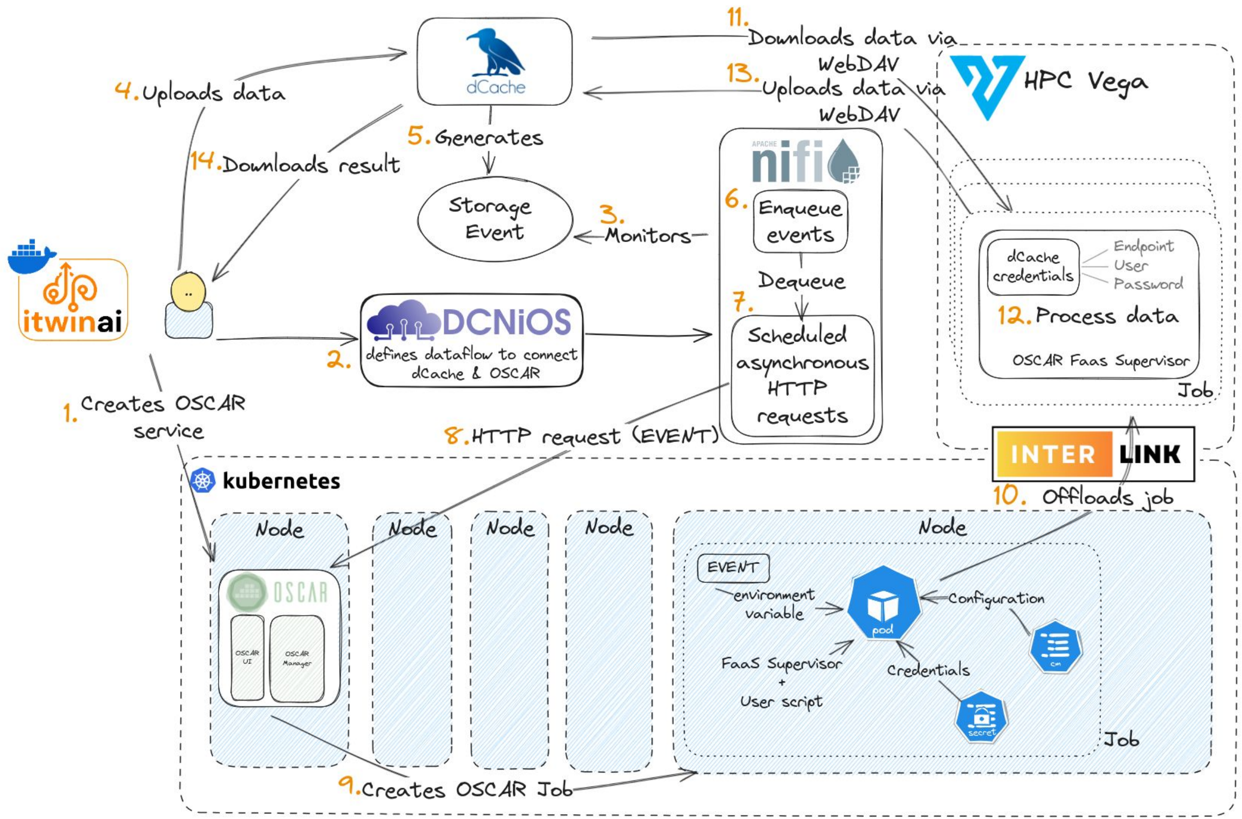
The use case

- Execution of an AI pipeline using itwinai (ML tool developed by CERN)
- The AI pipeline is configured to run the inference of a 3DGAN pretrained model to simulate particles in the HL-LHC (CERN)
- Data input from dCache, a distributed storage system for storing and retrieving huge amounts of data

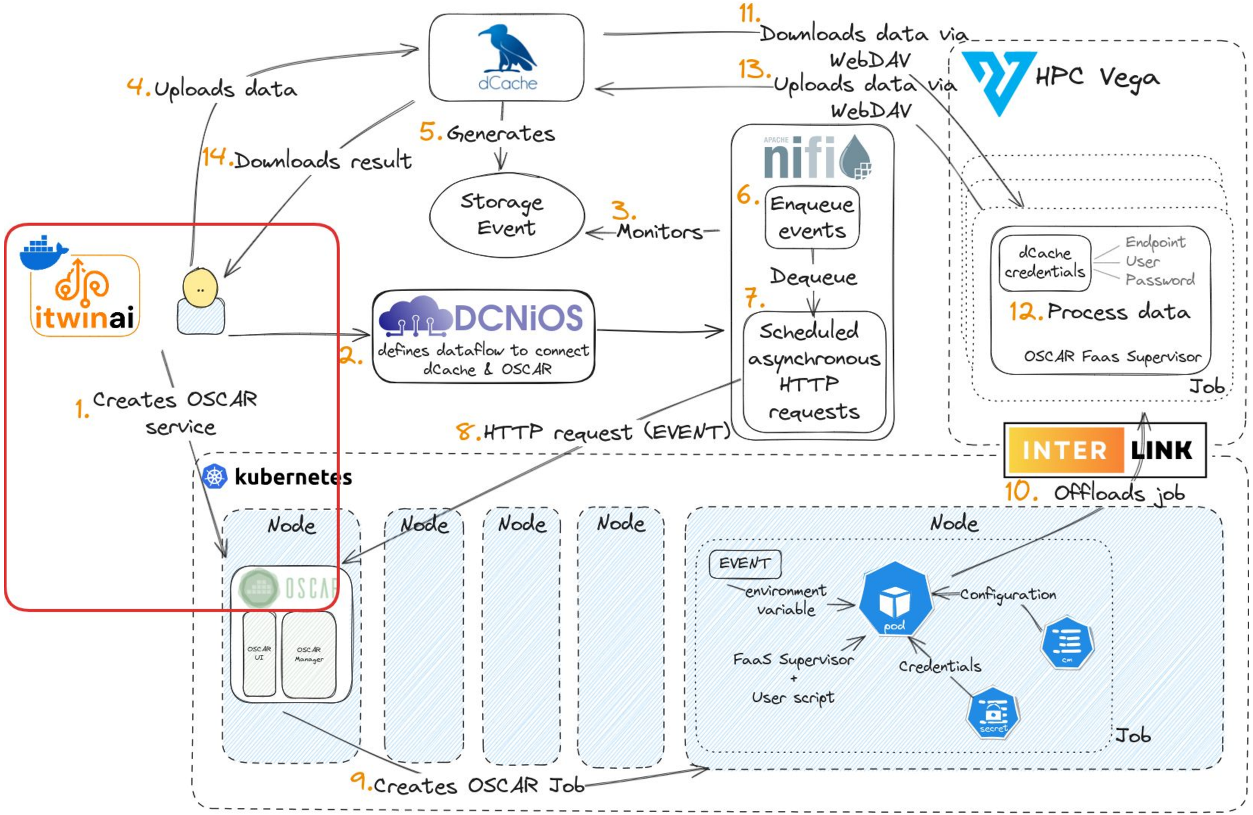
https://github.com/interTwin-eu/itwinai/

https://www.dcache.org/

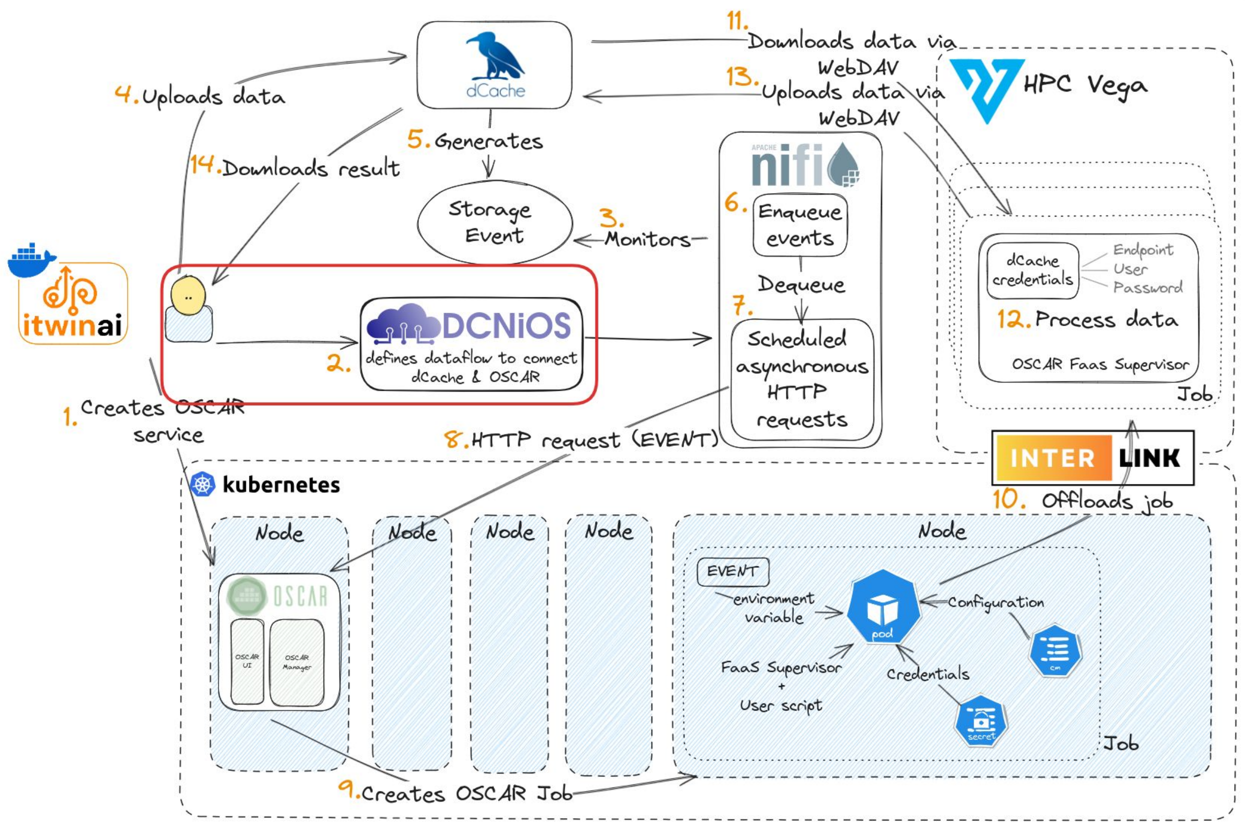
The workflow



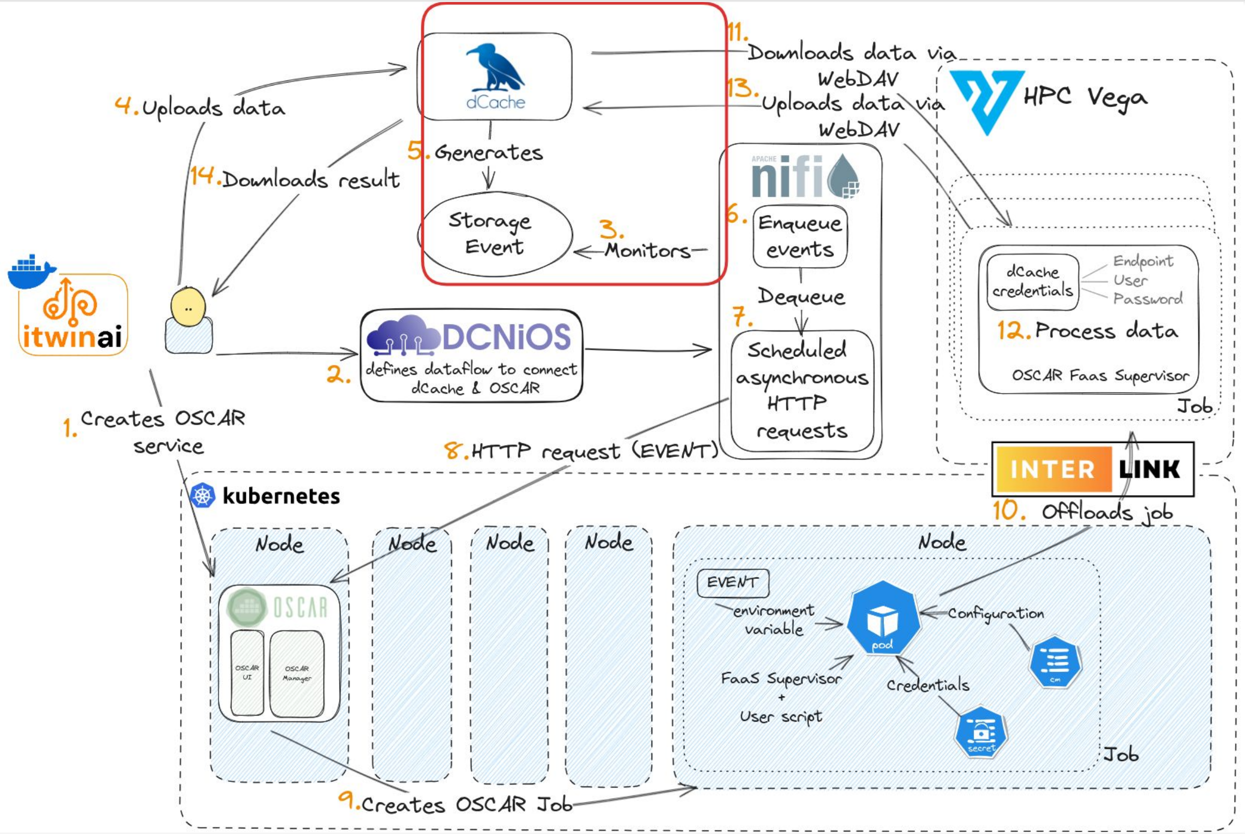
The workflow



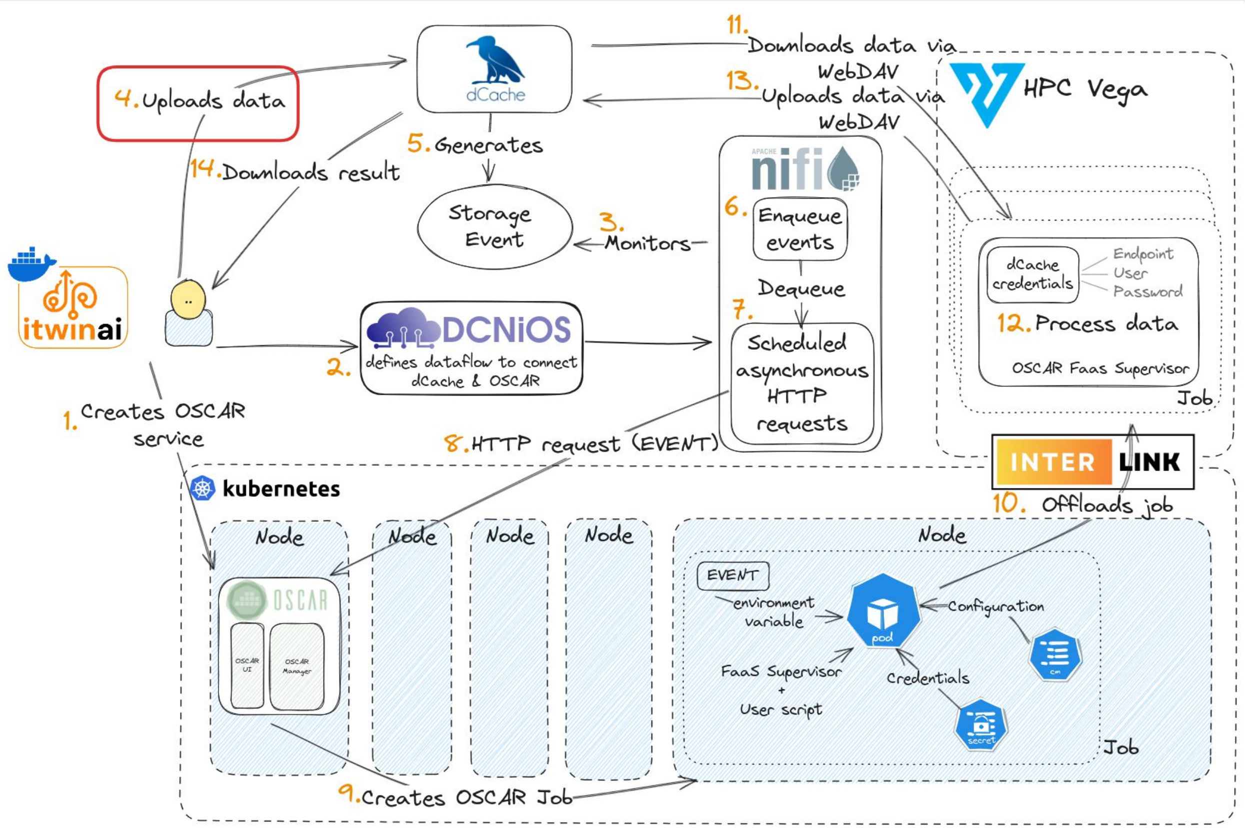
The workflow



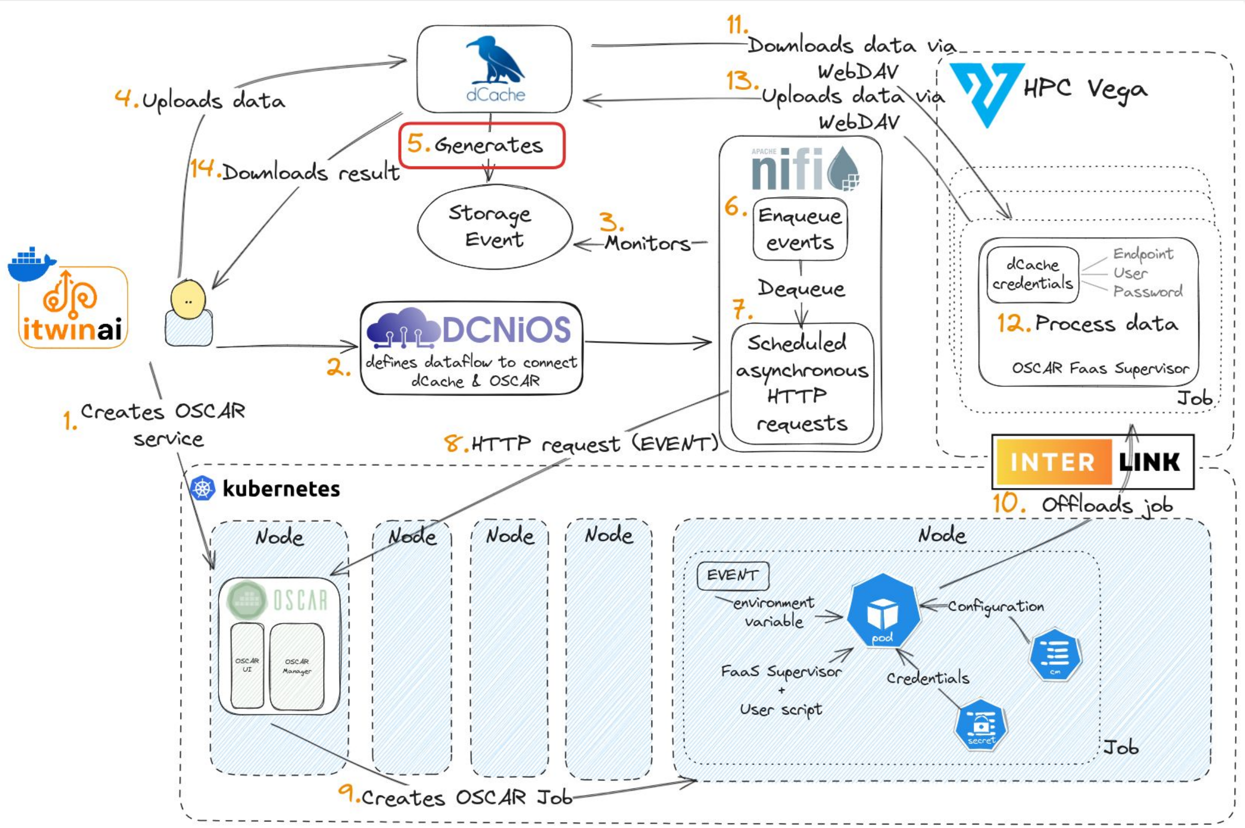
The workflow



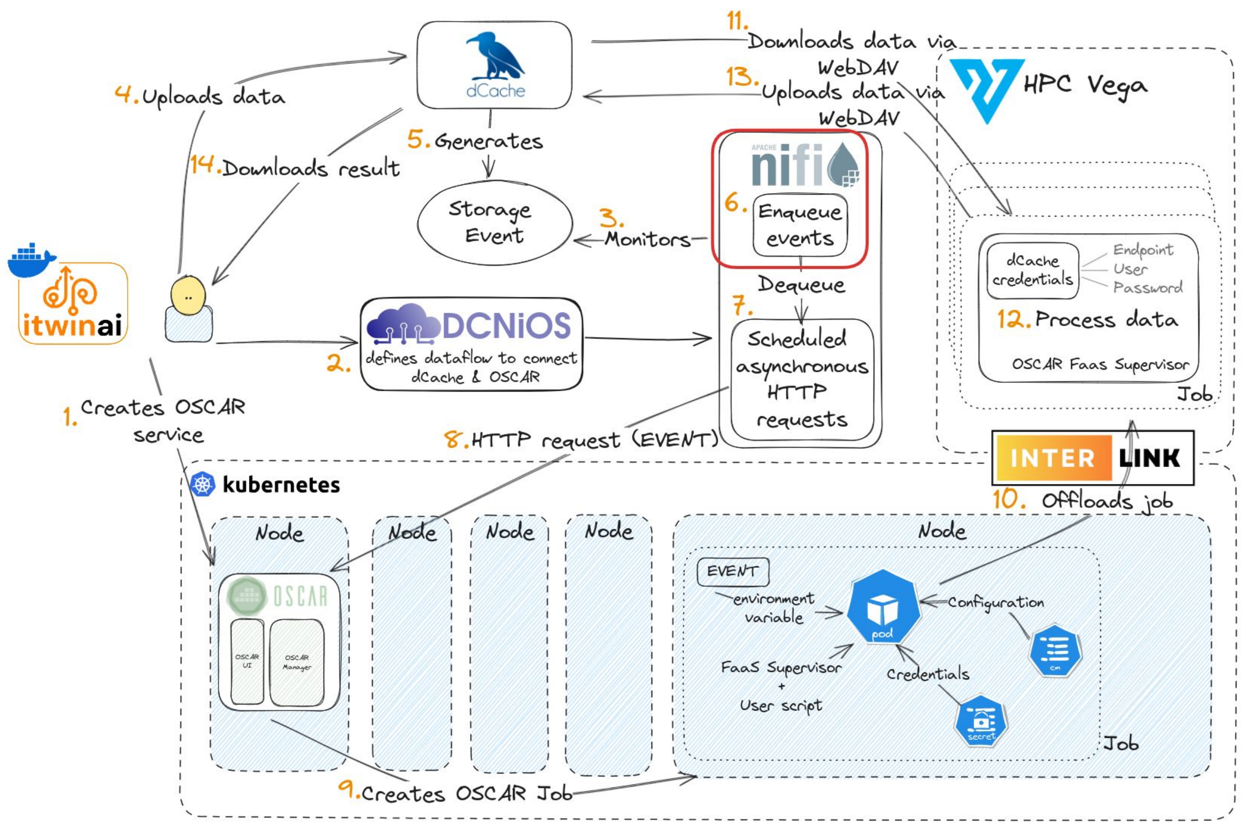
The workflow



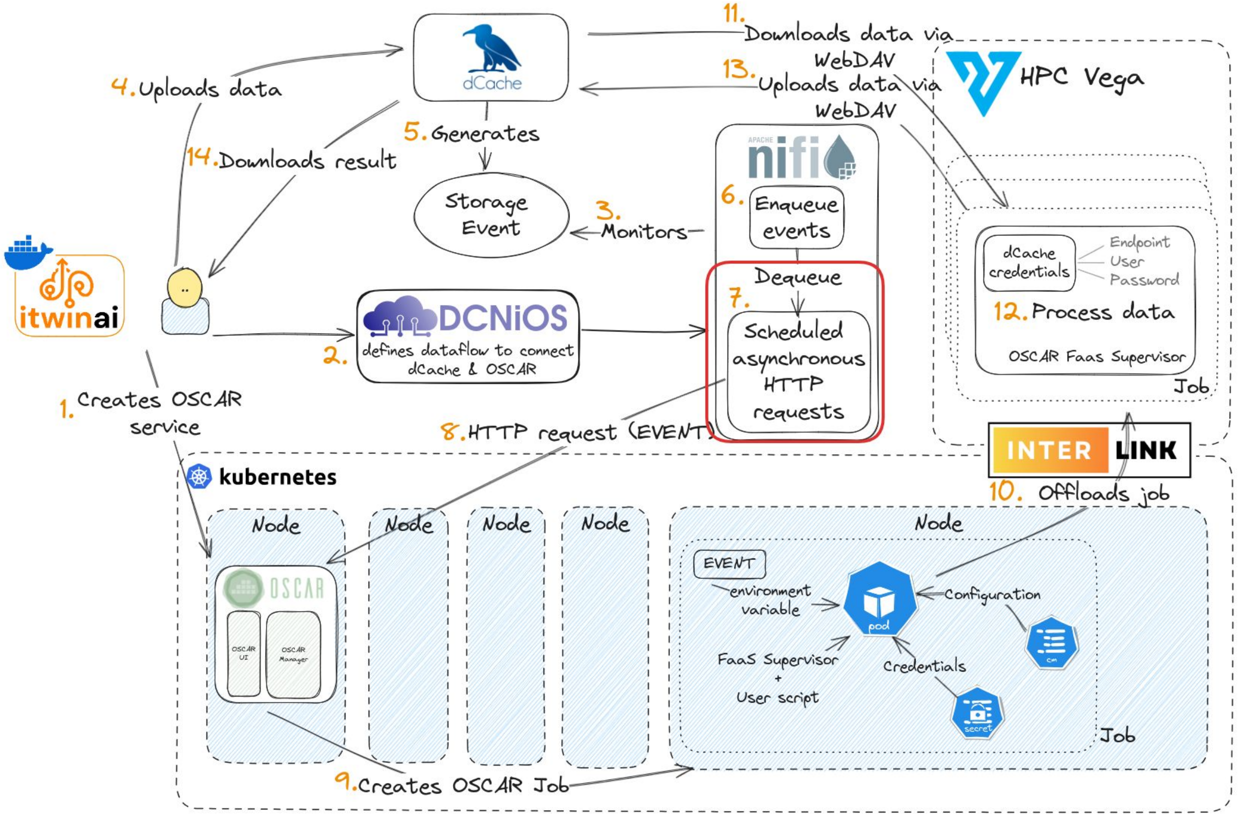
The workflow



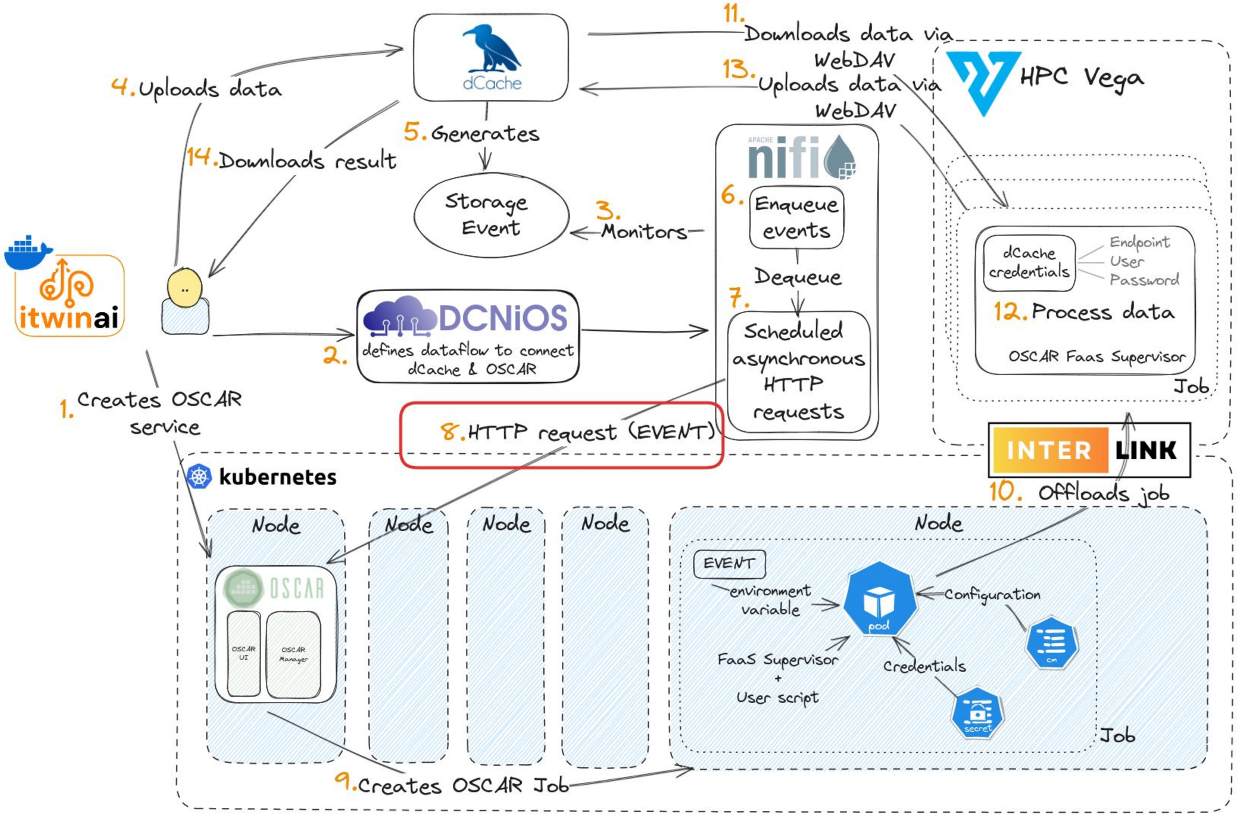
The workflow



The workflow



The workflow



dCache

HPC Vega

11. Downloads data via WebDAV

13. Uploads data via WebDAV

4. Uploads data

14. Downloads result

5. Generates

Storage Event

3. Monitors

nifi

6. Enqueue events

Dequeue

7. Scheduled asynchronous HTTP requests

dCache credentials

Endpoint

User

Password

12. Process data

OSCAR Faas Supervisor

Job

itwinai

DCNiOS

2. defines dataflow to connect dCache & OSCAR

1. Creates OSCAR service

8. HTTP request (EVENT)

INTER LINK

10. Offloads job

kubernetes

Node

Node

Node

Node

Node

OSCAR

OSCAR UI

OSCAR Manager

EVENT

environment variable

pod

Configuration

cm

FaaS Supervisor + User script

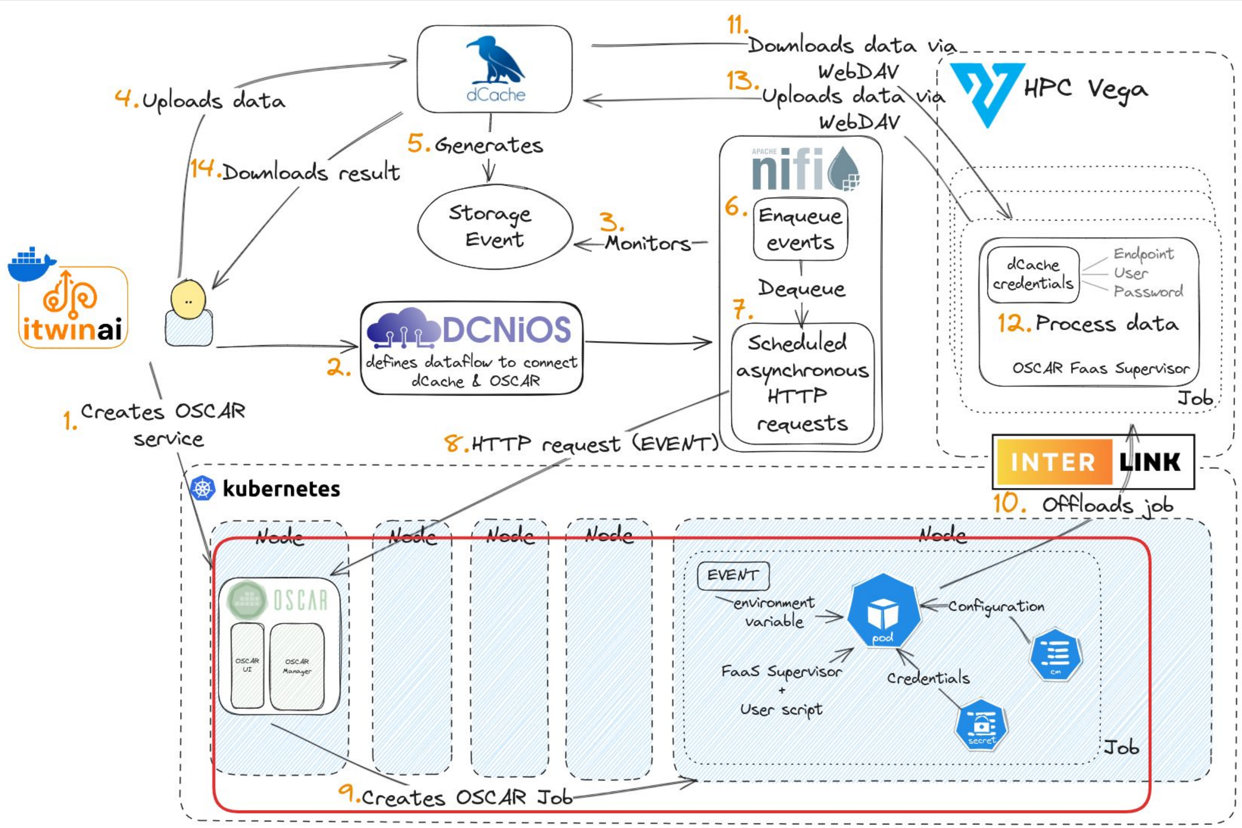
Credentials

secret

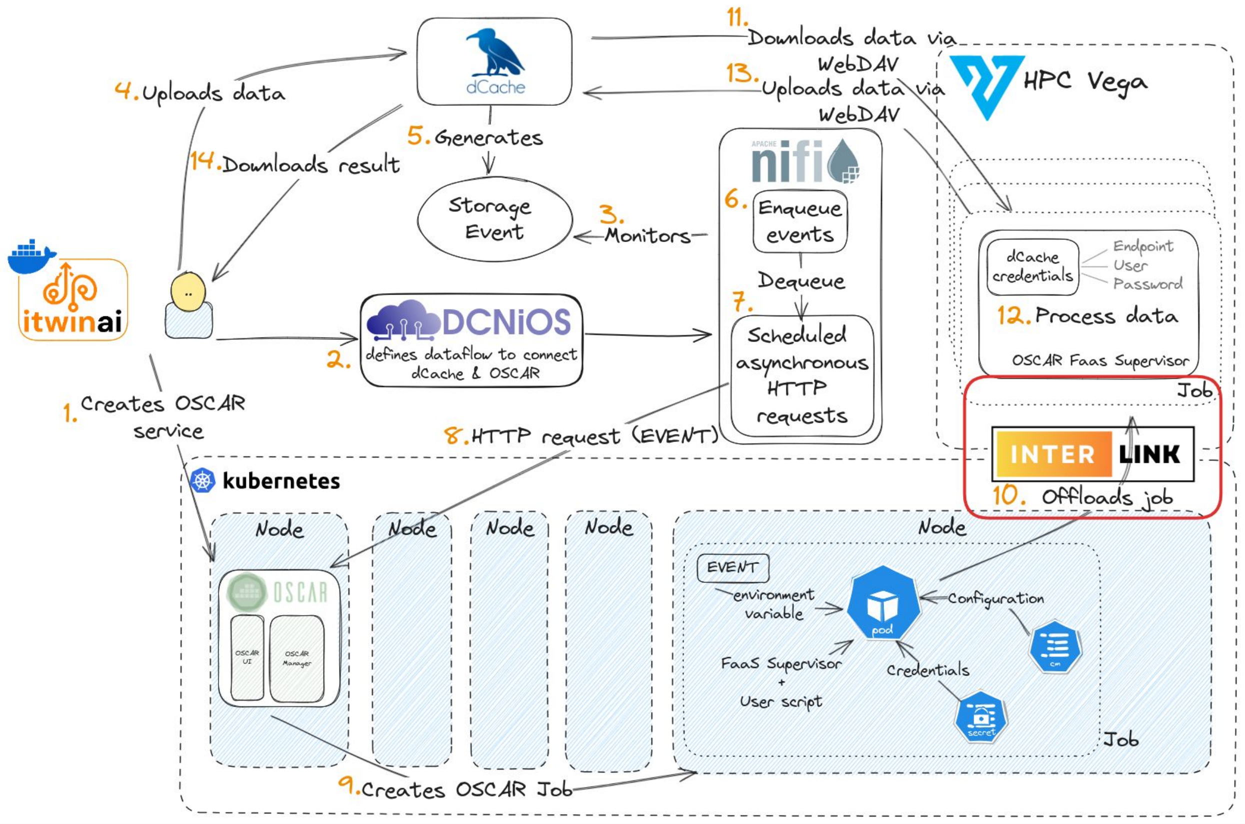
Job

9. Creates OSCAR Job

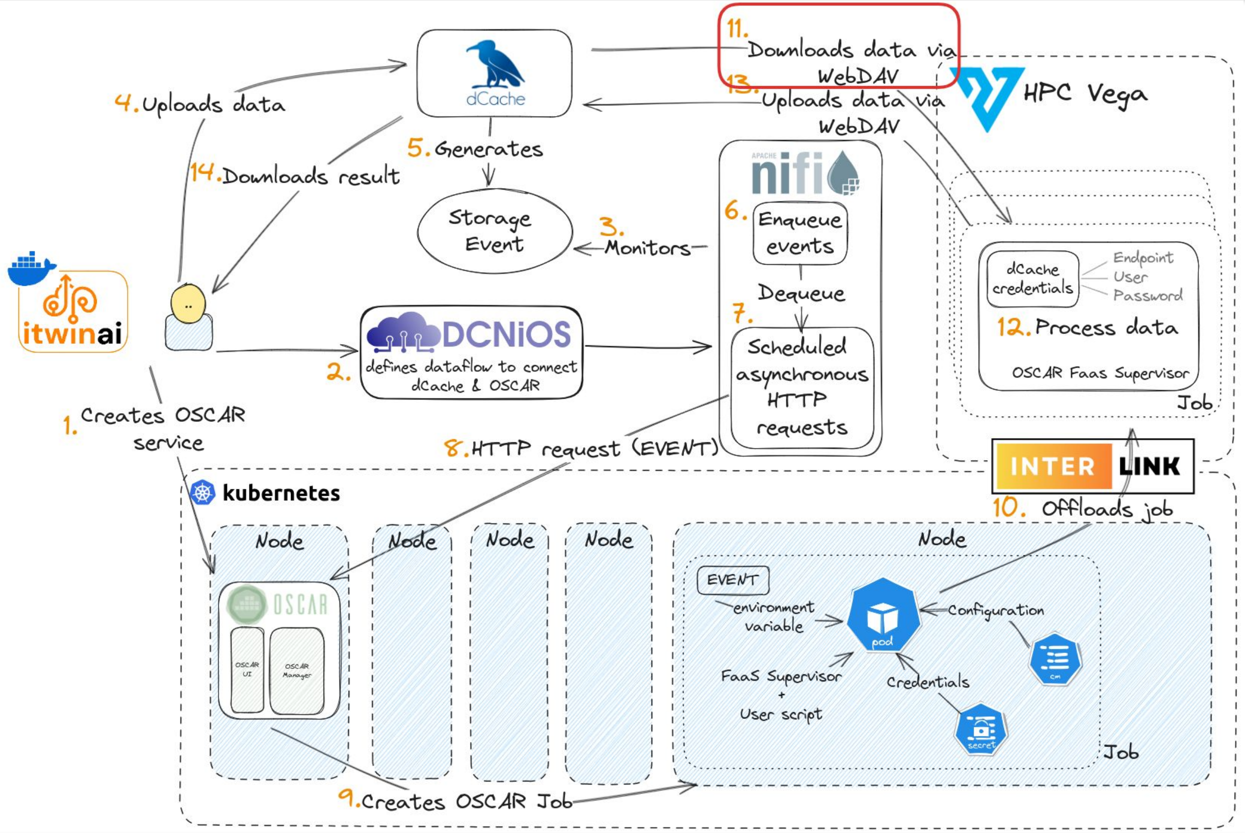
The workflow



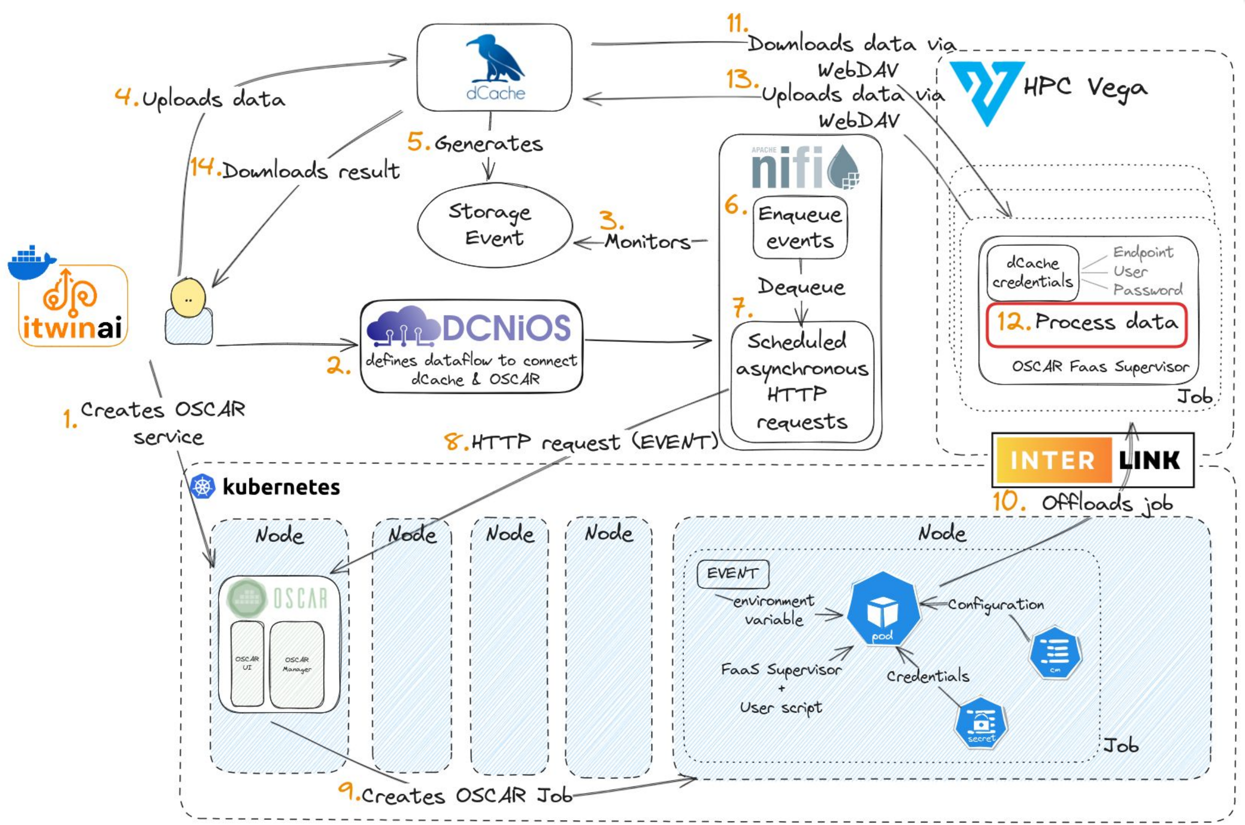
The workflow



The workflow



The workflow



dCache

HPC Vega

itwinai

kubernetes

1. Creates OSCAR service

2. DCNiOS defines dataflow to connect dCache & OSCAR

3. Monitors

4. Uploads data

5. Generates

Storage Event

6. Enqueue events

Dequeue

7. Scheduled asynchronous HTTP requests

8. HTTP request (EVENT)

9. Creates OSCAR Job

10. Offloads job

11. Downloads data via WebDAV

12. Process data

13. Uploads data via WebDAV

14. Downloads result

dCache credentials

Endpoint

User

Password

OSCAR Faas Supervisor

Job

INTER LINK

Node

Node

Node

Node

Node

OSCAR

OSCAR UI

OSCAR Manager

EVENT

environment variable

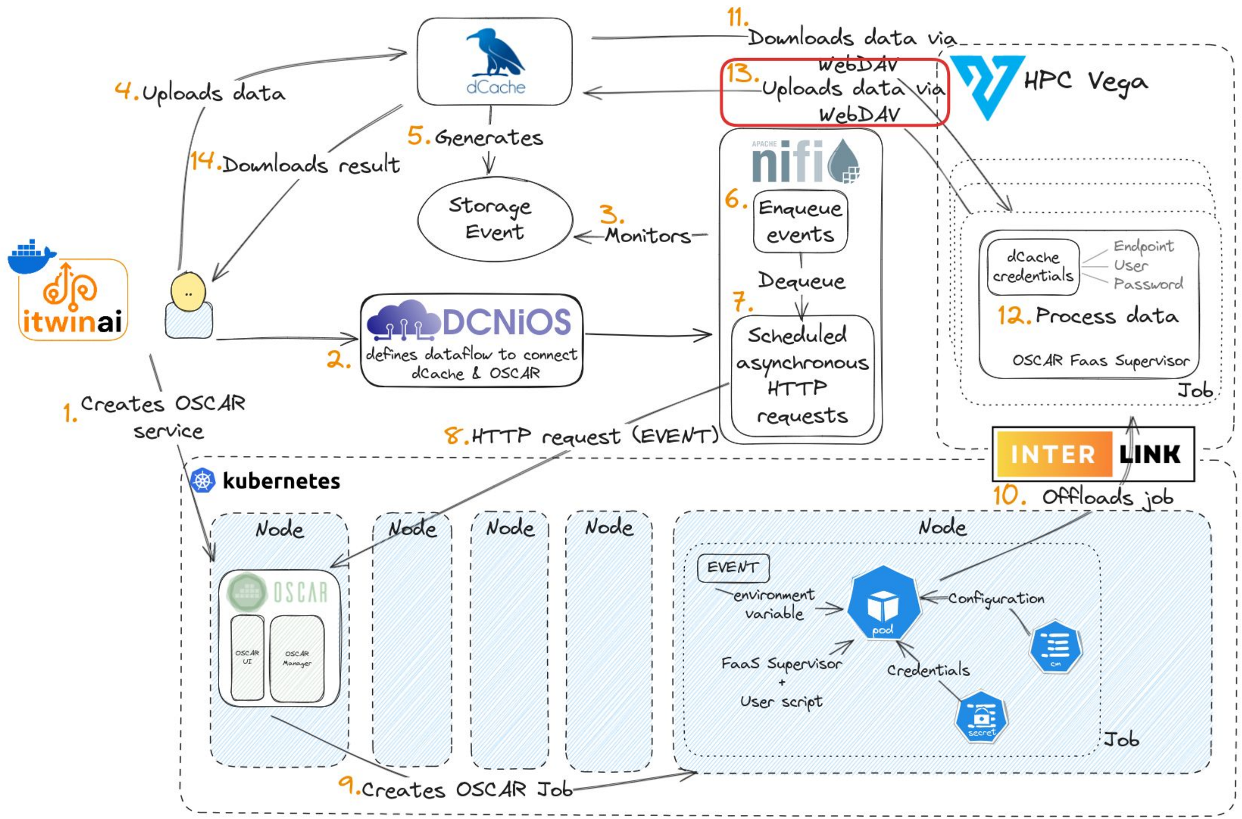
FaaS Supervisor + User script

Configuration

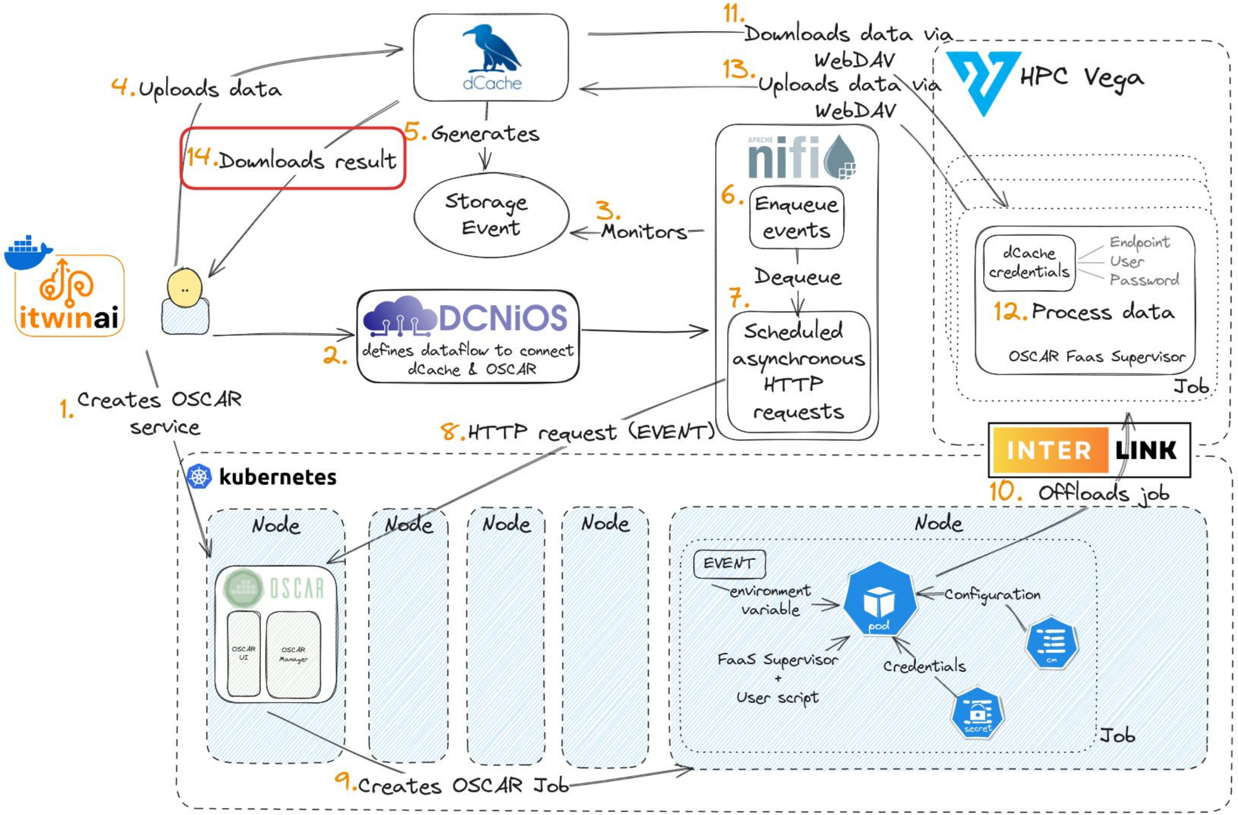
Credentials

Job

The workflow



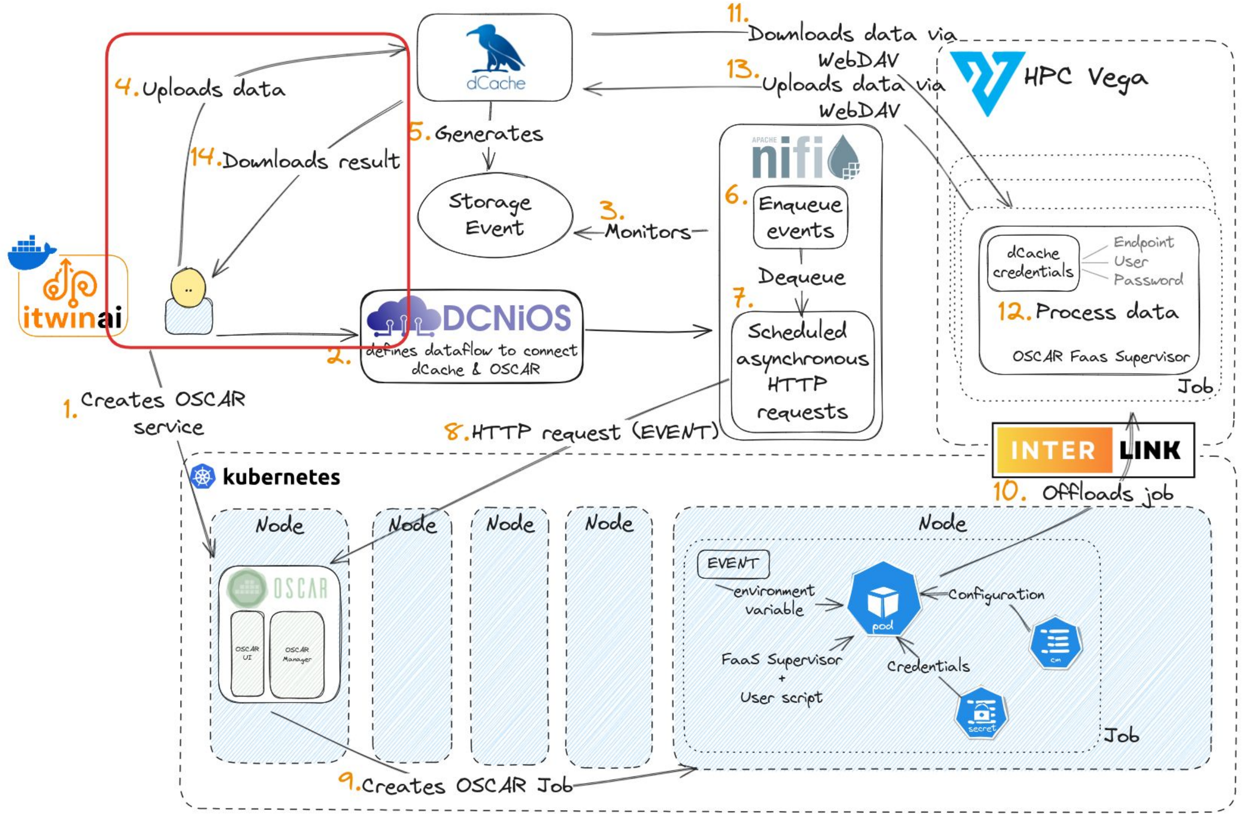
The workflow



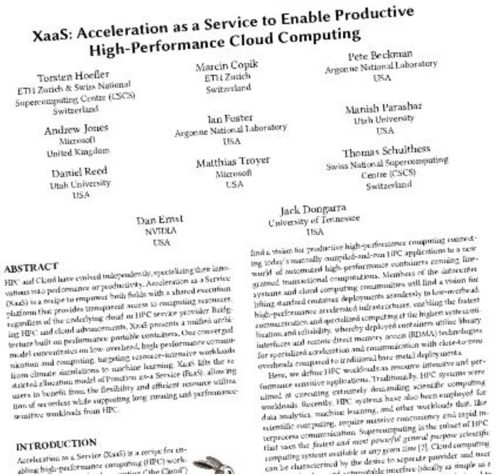
The workflow

This image is a creative interpretation based on a descriptive prompt and does not represent an actual scene or object. It is a product of imagination and artistic expression.

The workflow



Results



XaaS: Acceleration as a Service to Enable Productive High-Performance Cloud Computing

Torsten Hoefler
ETH Zurich & Swiss National Supercomputing Centre (CSCS)
Switzerland

Marcin Copik
ETH Zurich
Switzerland

Pete Beckman
Argonne National Laboratory
USA

Andrew Jones
Microsoft
United Kingdom

Ian Foster
Argonne National Laboratory
USA

Manish Parashar
Utah University
USA

Daniel Reed
Utah University
USA

Matthias Troyer
Microsoft
USA

Thomas Schulthess
Swiss National Supercomputing Centre (CSCS)
Switzerland

Dan Ernst
NVIDIA
USA

Jack Dongarra
University of Tennessee
USA

ABSTRACT

HPC and Cloud have evolved independently, specializing their innovations into performance or productivity. Acceleration as a Service (XaaS) is a recipe to empower both fields with a shared execution platform that provides transparent access to computing resources, regardless of the underlying cloud or HPC service provider. Bridging HPC and cloud advancements, XaaS presents a unified architecture built on performance portable containers. Our converged model concentrates on low-overhead, high-performance communication and computing, targeting resource-intensive workloads from climate simulations to machine learning. XaaS lifts the restricted allocation model of Function-as-a-Service (FaaS), allowing users to benefit from the flexibility and efficient resource utilization of serverless while supporting long-running and performance-sensitive workloads from HPC.

INTRODUCTION

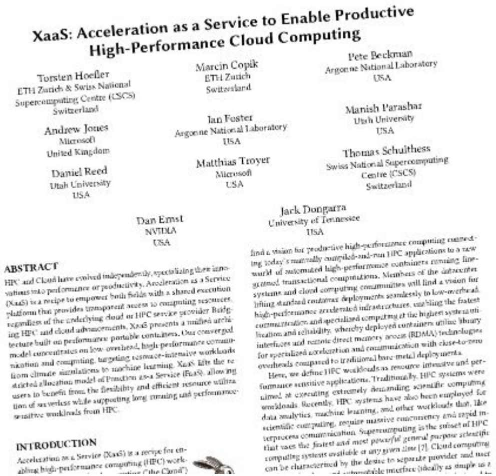
Acceleration as a Service (XaaS) is a recipe for enabling high-performance computing (HPC) work- [illegible] computing ("the Cloud")

find a vision for productive high-performance computing connecting today's manually compiled-and-run HPC applications to a new world of automated high-performance containers running fine-grained, transactional computations. Members of the datacenter systems and cloud computing communities will find a vision for lifting standard container deployments seamlessly to low-overhead high-performance accelerated infrastructures, enabling the fastest communication and specialized computing at the highest system utilization and reliability, whereby deployed containers utilize library interfaces and remote direct memory access (RDMA) technologies for specialized acceleration and communication with close-to-zero overheads compared to traditional bare-metal deployments.

Here, we define HPC workloads as resource intensive and performance sensitive applications. Traditionally, HPC systems were aimed at executing extremely demanding scientific computing workloads. Recently, HPC systems have also been employed for data analytics, machine learning, and other workloads that, like scientific computing, require massive concurrency and rapid interprocess communication. Supercomputing is the subset of HPC that uses the fastest and most powerful general purpose scientific computing systems available at any given time [7]. Cloud computing can be characterized by the desire to separate provider and user

- Hoefler et al., 2024, highlight the importance of bridging Cloud and HPC for resource-intensive workloads like **climate simulations** or **machine learning** processing
- Some key ideas:
  - Leverage containers
  - Improve communication
  - Enable access to data (I/O)

Results



XaaS: Acceleration as a Service to Enable Productive High-Performance Cloud Computing

Torsten Hoefler
ETH Zurich & Swiss National Supercomputing Centre (CSCS)
Switzerland

Marcin Copik
ETH Zurich
Switzerland

Pete Beckman
Argonne National Laboratory
USA

Andrew Jones
Microsoft
United Kingdom

Ian Foster
Argonne National Laboratory
USA

Manish Parashar
Utah University
USA

Daniel Reed
Utah University
USA

Matthias Troyer
Microsoft
USA

Thomas Schulthess
Swiss National Supercomputing Centre (CSCS)
Switzerland

Dan Ernst
NVIDIA
USA

Jack Dongarra
University of Tennessee
USA

ABSTRACT

HPC and Cloud have evolved independently, specializing their innovations into performance or productivity. Acceleration as a Service (XaaS) is a recipe to empower both fields with a shared execution platform that provides transparent access to computing resources, regardless of the underlying cloud or HPC service provider. Bridging HPC and cloud advancements, XaaS presents a unified architecture built on performance portable containers. Our converged model concentrates on low-overhead, high-performance communication and computing, targeting resource-intensive workloads from climate simulations to machine learning. XaaS lifts the restricted allocation model of Function-as-a-Service (FaaS), allowing users to benefit from the flexibility and efficient resource utilization of serverless while supporting long-running and performance-sensitive workloads from HPC.

INTRODUCTION

Acceleration as a Service (XaaS) is a recipe for enabling high-performance computing (HPC) work- [illegible] ("the Cloud")

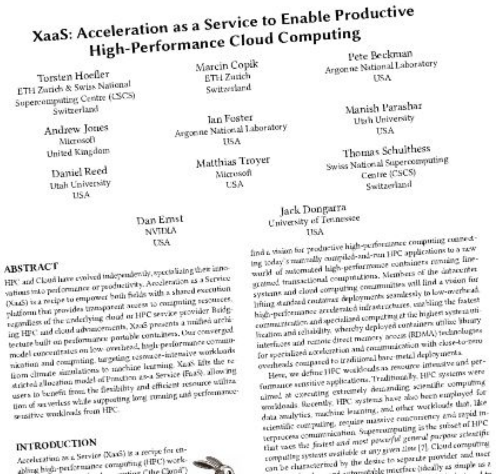
find a vision for productive high-performance computing connecting today's manually compiled-and-run HPC applications to a new world of automated high-performance containers running fine-grained, transactional computations. Members of the datacenter systems and cloud computing communities will find a vision for lifting standard container deployments seamlessly to low-overhead high-performance accelerated infrastructures, enabling the fastest communication and specialized computing at the highest system utilization and reliability, whereby deployed containers utilize library interfaces and remote direct memory access (RDMA) technologies for specialized acceleration and communication with close-to-zero overheads compared to traditional bare-metal deployments.

Here, we define HPC workloads as resource intensive and performance sensitive applications. Traditionally, HPC systems were aimed at executing extremely demanding scientific computing workloads. Recently, HPC systems have also been employed for data analytics, machine learning, and other workloads that, like scientific computing, require massive concurrency and rapid interprocess communication. Supercomputing is the subset of HPC that uses the fastest and most powerful general purpose scientific computing systems available at any given time [7]. Cloud computing can be characterized by the desire to separate provider and user

- Hoefler et al., 2024, highlight the importance of bridging Cloud and HPC for resource-intensive workloads like **climate simulations** or **machine learning** processing
- Some key ideas:
  - Leverage containers
  - Improve communication
  - Enable access to data (I/O)

We employ containers to be able to run the workload in the Cloud and in a HPC cluster

Results



XaaS: Acceleration as a Service to Enable Productive High-Performance Cloud Computing

Torsten Hoefler
ETH Zurich & Swiss National Supercomputing Centre (CSCS)
Switzerland

Marcin Copik
ETH Zurich
Switzerland

Pete Beckman
Argonne National Laboratory
USA

Andrew Jones
Microsoft
United Kingdom

Ian Foster
Argonne National Laboratory
USA

Manish Parashar
Utah University
USA

Daniel Reed
Utah University
USA

Matthias Troyer
Microsoft
USA

Thomas Schulthess
Swiss National Supercomputing Centre (CSCS)
Switzerland

Dan Ernst
NVIDIA
USA

Jack Dongarra
University of Tennessee
USA

ABSTRACT

HPC and Cloud have evolved independently, specializing their innovations into performance or productivity. Acceleration as a Service (XaaS) is a recipe to empower both fields with a shared execution platform that provides transparent access to computing resources, regardless of the underlying cloud or HPC service provider. Bridging HPC and cloud advancements, XaaS presents a unified architecture built on performance-portable containers. Our converged model concentrates on low-overhead, high-performance communication and computing, targeting resource-intensive workloads from climate simulations to machine learning. XaaS lifts the restricted allocation model of Function-as-a-Service (FaaS), allowing users to benefit from the flexibility and efficient resource utilization of serverless while supporting long-running and performance-sensitive workloads from HPC.

INTRODUCTION

Acceleration as a Service (XaaS) is a recipe for enabling high-performance computing (HPC) work-

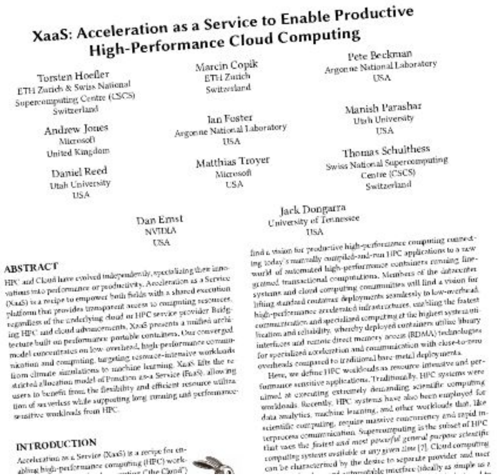
find a vision for productive high-performance computing connecting today's manually compiled-and-run HPC applications to a new world of automated high-performance containers running fine-grained, transactional computations. Members of the datacenter systems and cloud computing communities will find a vision for lifting standard container deployments seamlessly to low-overhead, high-performance accelerated infrastructures, enabling the fastest communication and specialized computing at the highest system utilization and reliability, whereby deployed containers utilize library interfaces and remote direct memory access (RDMA) technologies for specialized acceleration and communication with close-to-zero overheads compared to traditional bare-metal deployments.

Here, we define HPC workloads as resource intensive and performance sensitive applications. Traditionally, HPC systems were aimed at executing extremely demanding scientific computing workloads. Recently, HPC systems have also been employed for data analytics, machine learning, and other workloads that, like scientific computing, require massive concurrency and rapid interprocess communication. Supercomputing is the subset of HPC that uses the *fastest and most powerful general purpose scientific computing systems available at any given time* [7]. Cloud computing can be characterized by the desire to separate provider and user

- Hoefler et al., 2024, highlight the importance of bridging Cloud and HPC for resource-intensive workloads like **climate simulations** or **machine learning** processing
- Some key ideas:
  - Leverage containers
  - Improve communication
  - Enable access to data (I/O)

We employ interLink to establish communication between Cloud and HPC

Results



XaaS: Acceleration as a Service to Enable Productive High-Performance Cloud Computing

Torsten Hoefler
ETH Zurich & Swiss National Supercomputing Centre (CSCS)
Switzerland

Marcin Copik
ETH Zurich
Switzerland

Pete Beckman
Argonne National Laboratory
USA

Andrew Jones
Microsoft
United Kingdom

Ian Foster
Argonne National Laboratory
USA

Manish Parashar
Utah University
USA

Daniel Reed
Utah University
USA

Matthias Troyer
Microsoft
USA

Thomas Schulthess
Swiss National Supercomputing Centre (CSCS)
Switzerland

Dan Ernst
NVIDIA
USA

Jack Dongarra
University of Tennessee
USA

ABSTRACT

HPC and Cloud have evolved independently, specializing their innovations into performance or productivity. Acceleration as a Service (XaaS) is a recipe to empower both fields with a shared execution platform that provides transparent access to computing resources, regardless of the underlying cloud or HPC service provider. Bridging HPC and cloud advancements, XaaS presents a unified architecture built on performance portable containers. Our converged model concentrates on low-overhead, high-performance communication and computing, targeting resource-intensive workloads from climate simulations to machine learning. XaaS lifts the restricted allocation model of Function-as-a-Service (FaaS), allowing users to benefit from the flexibility and efficient resource utilization of serverless while supporting long running and performance-sensitive workloads from HPC.

INTRODUCTION

Acceleration as a Service (XaaS) is a recipe for enabling high-performance computing (HPC) work-

find a vision for productive high-performance computing connecting today's manually compiled-and-run HPC applications to a new world of automated high-performance containers running fine-grained, transactional computations. Members of the datacenter systems and cloud computing communities will find a vision for lifting standard container deployments seamlessly to low-overhead, high-performance accelerated infrastructures, enabling the fastest communication and specialized computing at the highest system utilization and reliability, whereby deployed containers utilize library interfaces and remote direct memory access (RDMA) technologies for specialized acceleration and communication with close-to-zero overheads compared to traditional bare-metal deployments.

Here, we define HPC workloads as resource intensive and performance sensitive applications. Traditionally, HPC systems were aimed at executing extremely demanding scientific computing workloads. Recently, HPC systems have also been employed for data analytics, machine learning, and other workloads that, like scientific computing, require massive concurrency and rapid interprocess communication. Supercomputing is the subset of HPC that uses the *fastest and most powerful* general purpose scientific computing systems available at any given time [7]. Cloud computing can be characterized by the desire to separate provider and user

- Hoefler et al., 2024, highlight the importance of bridging Cloud and HPC for resource-intensive workloads like **climate simulations** or **machine learning** processing
- Some key ideas:
  - Leverage containers
  - Improve communication
  - Enable access to data (I/O)

We employ FaaS Supervisor to enable data access from Cloud and HPC clusters

Results

- Access to HPC provides better resources and powerful GPUs
- We compared three different scenarios, Cloud-CPU, HPC-CPU, and HPC-GPU
- The processing time for each inference is stable
- HPC-CPU run the inference a ~ 30% faster than Cloud
- HPC-GPU run the inference a ~ 80% faster than Cloud

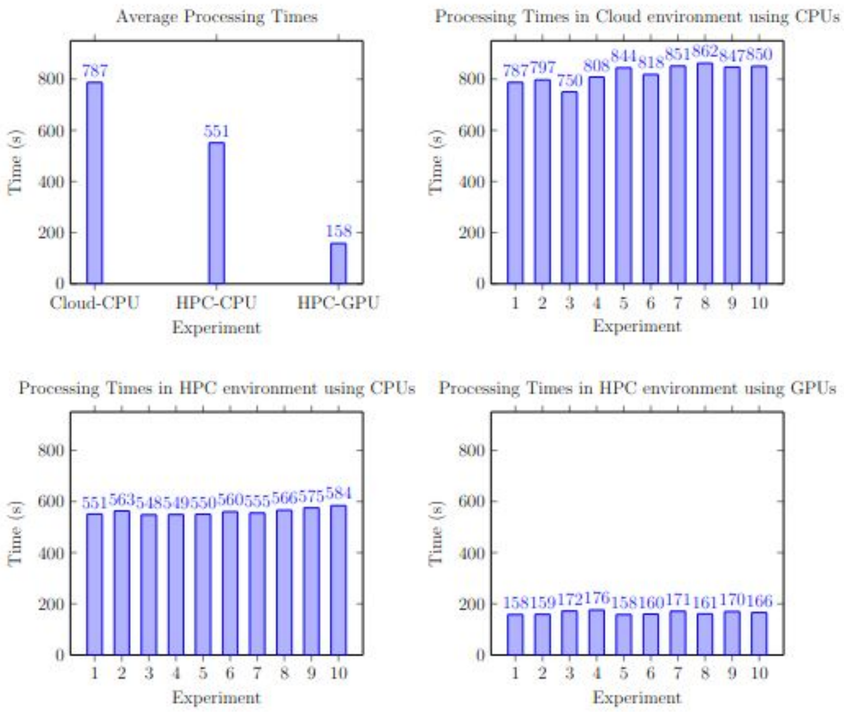


Figure 5: Process time of the itwinai inference use case showing total execution time for each experiment and environment type.

Conclusions

- In this work we introduced an architecture to facilitate the integration of cloud computing and HPC for the execution of compute-intensive AI model inference, leveraging the event-driven serverless computing paradigm

- This integration facilitates access to HPC resources from OSCAR/Kubernetes clusters, thus making HPC resources seamlessly accessible to a wider range of users and applications via automated offloading through interLink

- A successful use case integrating dCache, Apache NiFi, OSCAR, interLink, and itwinai to support generative AI 3DGAN neural network model has been achieved, demonstrating the benefits of the approach by exploiting remote GPUs from an HPC facility from an OSCAR cluster running on a cloud infrastructure

In summary, the proposed system demonstrates the potential for offloading compute-intensive tasks to an HPC supercomputer, marking a significant step forward in Bridging Cloud and HPC

Acknowledgements

We would like to thank the following organizations and projects:

- “An interdisciplinary Digital Twin Engine for science” (interTwin) that has received funding from the European Union’s Horizon Europe Programme under Grant 101058386
- Grant PID2020-113126RB-I00 funded by MCIU/AEI/10.13039/501100011033
- Project PDC2021-120844-I00 funded by MCIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR

Thanks for your attention!

Bridging Cloud and HPC for Scalable Event-driven Processing of AI Workloads

Estíbaliz Parcero, Sergio Langarita, Germán Moltó

Instituto de Instrumentación para Imagen Molecular (I3M), Universitat Politècnica de València, Valencia, Spain.
esparig@i3m.upv.es

OSCAR

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

GRyCAP