Towards Efficient Workflows with STFC and CCP4 Cloud

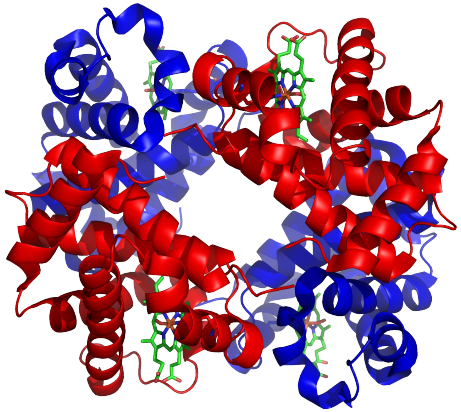
Ville Uski, CCP4, STFC

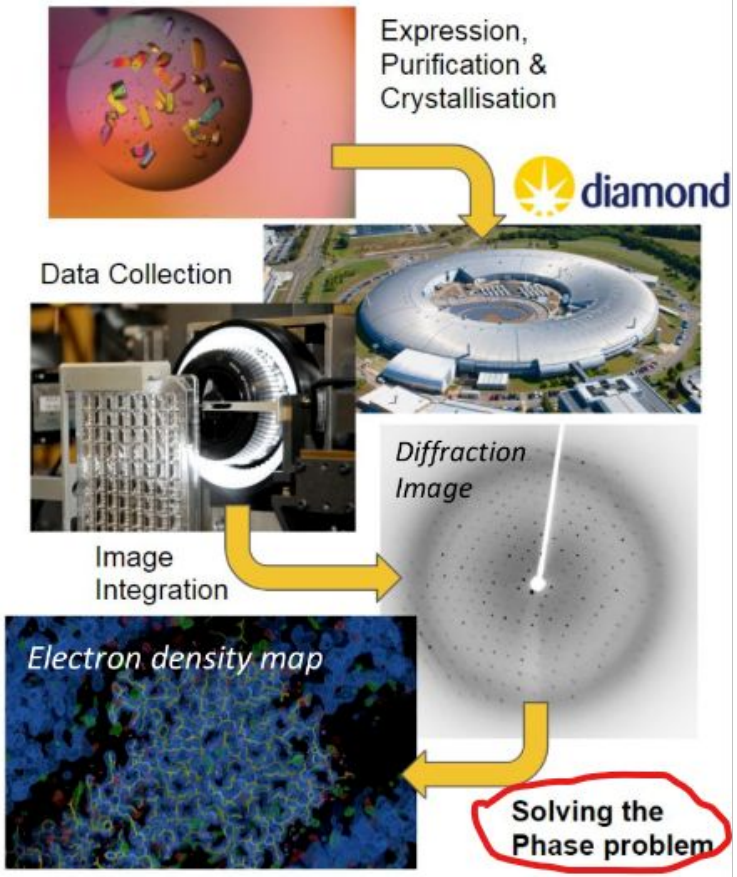
IRIS collaboration meeting, 2nd July 2025

CCP4

Science and Technology Facilities Council

Macromolecular crystallography (MX)





- the gold standard for determining 3D structures of proteins and other biological molecules.
- enables insights into how proteins function, how drugs bind, and how to design better therapeutics.

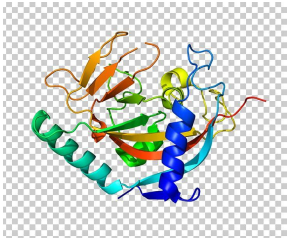
CCP4

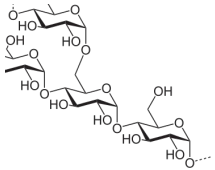
UKRI Science and Technology Facilities Council

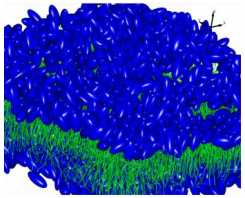
Software for Macromolecular X-Ray Crystallography since 1979

It is used for solving structures of macromolecules: proteins, carbohydrates, lipids, RNAs, DNAs.

Implement many complementary algorithms for all stages of the structure solution.



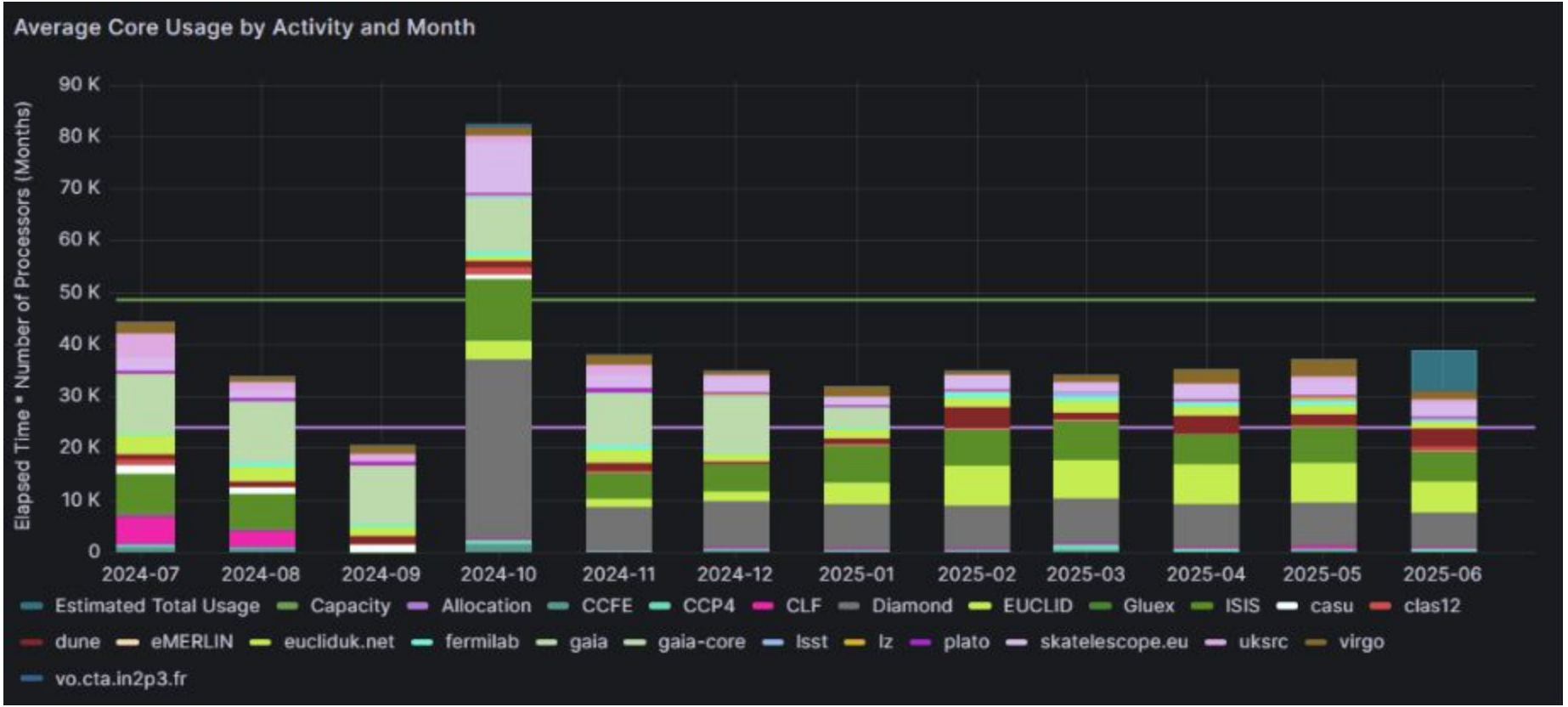




CCP4 is supported by the Biotechnology and Biological Sciences Research Council (BBSRC) and the industrial income.

Science and Technology Facilities Council

Biotechnology and Biological Sciences Research Council



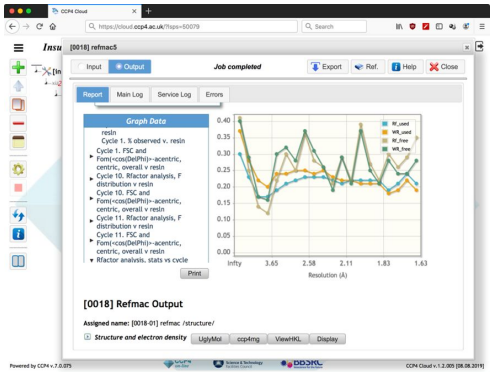
CCP4

Science and Technology Facilities Council

The Evolving Landscape of MX

- AlphaFold and multi-crystal experiments have made structure prediction more accessible
- But modern beamlines collect data faster than traditional software can handle
- High throughput requires scalable, automated solutions.

CCP4 Cloud



- Web-based platform for solving macromolecular structures from X-ray data.
- Built on CCP4 suite with cloud-native, automated workflows.
- ccp4.cloud.ac.uk is an instance of CCP4 Cloud run by us

CCP4

Science and Technology Facilities Council

CCP4 Cloud: Number crunchers

When users start new tasks (or workflows) on the CCP4 Cloud front-end, tasks are sent to run on ‘number crunchers’.

NC’s are daemon processes on remote servers, submitting to a job scheduling system, such as Slurm.

cloud.ccp4.ac.uk uses several number crunchers

- ccp4 servers
- Virtual CPU cluster in STFC Cloud (RAL cloud)
- Virtual GPU cluster in STFC Cloud

From Manual Pipelines to Automated Workflows

CCP4 Cloud features:

- Predefined workflows for structure solution.
- Custom workflows via WScript for tailored experiments.
- Integration with beamlines and remote data collection.

Benefits:

- Reduces manual intervention.
- Enables reproducibility and scalability.
- Supports both novice and expert users.

Received: 19 February 2025 | Revised: 7 May 2025 | Accepted: 9 May 2025

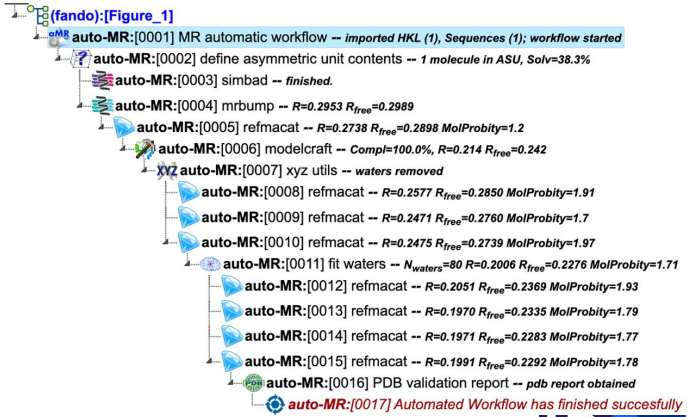
DOI: 10.1002/pro.70176

RESEARCH ARTICLE

THE PROTEIN SOCIETY WILEY

Project automation in CCP4 Cloud: Enabling customization and high-throughput efficiency

Eugene Krissinel | Maria Fando | Oleg Kovalevskiy | Ronan Keegan | Jools Wills | Andrey Lebedev | Ville Uski | Charles Ballard



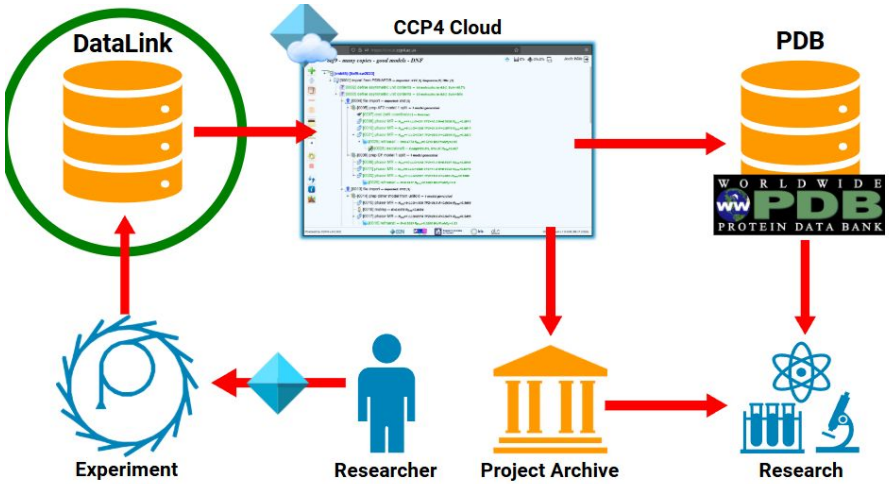
CCP4

Technology Facilities Council

DataLink: Connecting data sources with CCP4 Cloud

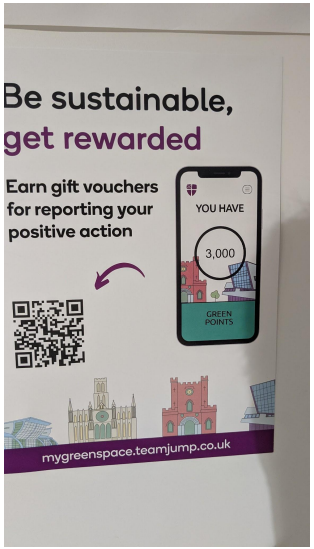
Jools Wills

Synchrotron	Location	Data Transfer	Depositions
aps-gmca	Chicago	**Globus**	21517
aps-lucat	Chicago	**Globus**	21517
aps-blocars	Chicago	**Globus**, hard disk	21517
aps-sercat	Chicago	**Globus**, hard disk	21517
aps-sbc-cat	Chicago	**Globus**, rsync, hard disk	21517
aps-necat	Chicago	**Globus**, sftp	21517
aps-imca-cat	Chicago	hard disk	21517
aps-lri-cat	Chicago	hard disk	21517
home source			11960
esrf	France/Europe	**Globus**, ftp, rsync	11124
diamond	UK	**Globus**	8343
sls	Switzerland	**Globus**, scp, rsync	6574
als	Berkeley	**Globus**, scp, rsync	6077
ssrf	Shanghai		5443
sser	San Francisco	**Globus**, scp, rsync	5182
nsls-ii	Brookhaven	**Globus**	4835
photon factory	Japan		3709
spring8	Japan	web portal	3637
bessy	Berlin	hard disk	2920
australian	Melbourne	sftp	2614
desy	Hamburg	scp, fastx3	1946
soleil	France	**Globus**	1793
pls-ii	Korea		1771
nsrrc	Taiwan	scp	1434
cls	Canada	hard disk, scp	1417
chess	Cornell	**Globus**, rsync, hard disk	970
max4	Sweden	**Globus**, sftp	761
alba	Spain	sftp	554
lnls	Brazil		535
elett	Italy	ftp	408



Resource usage patterns

- Peaks and idle times: Static cluster deployments often lead to either underutilization (wasted resources/cost) or insufficient capacity (jobs waiting)
- Most tasks run fast but complex workflows can lead to serious number crunching
- Motivation for dynamic resource allocation



ElastiCluster

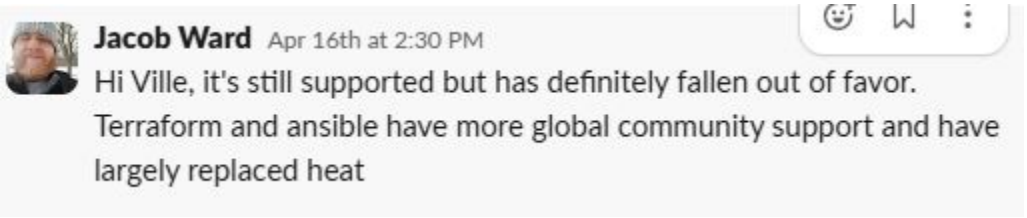
- github.com/elasticluster/elasticluster
- from University of Zurich
- creates ‘elastic’ HPC clusters of VMs on the cloud and configure them with Ansible.
- INI-style configuration file to define cluster templates
- Allows resizing, but this is slow and hard to automate

Auto-scaling with OpenStack Heat?

- “Heat is the core orchestration engine in OpenStack that manages the lifecycle of cloud resources, including virtual machines.”
- Heat templates define scaling policies that specify the conditions under which instances should be added or removed.

But it didn’t work.

Stack Name	Created	Updated	Status
test-stack	2 months, 2 weeks	Never	Create Failed
example-stack	3 years, 4 months	2 months, 2 weeks	Delete Failed



HashiCorp

Terraform

infrastructure-as-code tool from that enables the automation of infrastructure deployments and management

Packer

infrastructure-as-code tool that automates the creation of VM images

CCP4

UKRI Science and Technology Facilities Council

Slurm’s power saving parameters

- ResumeProgram: Path to a custom script executed when Slurm needs to power up a cloud node.
- SuspendProgram: Path to a custom script executed when Slurm decides an idle node can be powered down.
- ResumeTimeout & SuspendTimeout: Maximum time Slurm waits for script completion/node readiness.
- SuspendTime: Duration a node must be idle before Slurm considers it for suspension.

Auto-Scaling with Slurm and Terraform

‘Hybrid’ solution

- Terraform to provision and configure static infrastructure in OpenStack
- custom scripts that interact directly with OpenStack APIs for dynamic compute node management
- Packer to create and update base compute node images with the latest CCP4 suite

Autoscaling Workflow

1. CCP4 Cloud submits a new task to a number cruncher
2. Slurm needs resources; it executes ResumeProgram
3. ResumeProgram calls the OpenStack Nova API to create node
4. OpenStack provisions the VM. cloud-init configures a Slurm compute node, which registers itself with slurmctld via scontrol update
5. Job execution
6. Job completion
7. Suspend trigger
8. Instance deletion

Current Status

CCP4 Cloud has implemented automated workflows and allow custom ones

DataLink is in production

Terraform-powered auto-scaling cluster still in ‘dev’ stage but looks promising