Kepler scientific workflow orchestrator as a tool to build the computational workflows.

Marcin Płóciennik (PSNC), Sandro Fiore(CMCC), Tomasz Żok(PSNC)

eosc-hub.eu

@EOSC_eu

EOSC-hub receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 777536.

Outline

- Introducing Kepler
- Example components
- Execution modes
- Case studies
  - Fusion
  - Multi-model climate data analysis case study
- Future directions

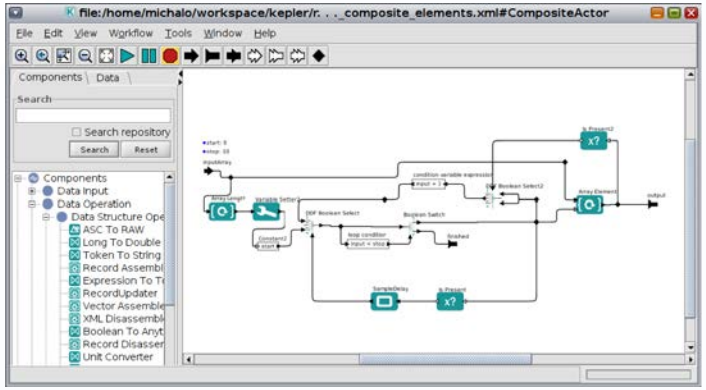
Introducing Kepler (1)

- Scientific Workflow System, initiated 2003
- Builds upon the open-source Ptolemy II framework
- Allows scientists to visually design and execute scientific workflows
- Actor-oriented model with directors acting as the main workflow engine (data driven)
- Enables different models of computation
- Workflows are saved as XML files - can easily be shared/published
- Kepler is supported by the NSF-funded Kepler/CORE team, which spans several of the key institutions that originated the Kepler project: UC Davis, UC Santa Barbara, and UC San Diego

www.kepler-project.org

Introducing Kepler (2)

- Kepler is an Open Source project (BSD License)
- Over 450 build in components
- Can be easily extended it by creating new elements
- Customised versions available – suits (e.g. bioKepler)
- Used across disciplines: Ecology, Engineering, Geology, Physics, Bioinformatics, Biology, Nuclear Fusion, Astrophysics, Nanotechnology, …
- Nested workflows
  - Multi level, mixed directors
  - Thousand of actors, loops, …
- Kepler can support users with:
  - building and executing workflows
  - executing tasks locally
  - executing tasks within distributed environments



Example components (1)

- Mathematical,
  - All kind of operations arithmetic, geometric, linear algebry, random number
  - Components, actors like R i Matlab
- Visualisation: ArrayPlotter, Bar Graph, Parallel Coordinate Plot , … , GML Displayer, XYPlotter , TimedPlotter,…. ImageDisplay, …. ImageJ,
- Data operation:
  - Strings, arrays, …
  - GeoData: GIS, GDAL
  - iRODS, SRB
  - Metadata: EML, ADN, Darwin Core
  - Data Access Protocol (DAP) 2.0
  - DataTurbine
- Database access
  - Oracle, MySQL, local or remote MS Access, DB2, MS SQL Server,PostgreSQL, MySQL, or Sybase SQL

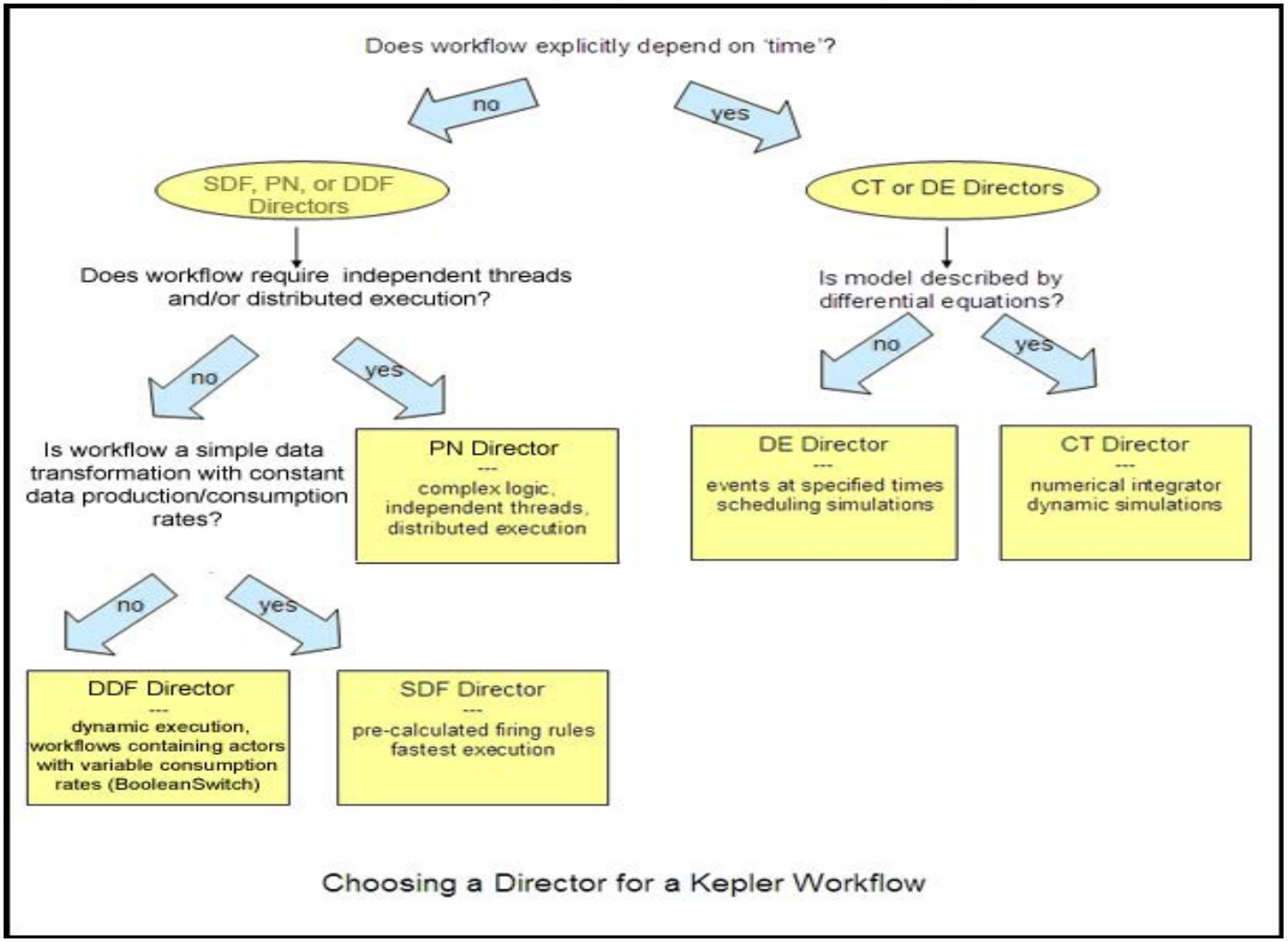
Example components (2)

- Other components:
  - web-services(SOAP, REST)
  - XML processing
- Specific modules:
  - GAMESS Input generator / molecule selector/BABEL , …
  - Sensor Processing and Acquisition Network (SPAN)
  - BioKepler (over 100 actors)
- Provenance Module
- Actor Languages
  - Java
  - C/C++
  - Python
  - Fortran

Provenance and reproducibility

- The Kepler Provenance add-on module suite provides allows the recording of workflow execution history
- Execution details are recorded into a database KeplerData/modules/provenance directory
- This feature is leveraged by modules such as Reporting and the Workflow Run Manager, which provides a GUI to manage and share past workflow runs and results

Kepler Directors



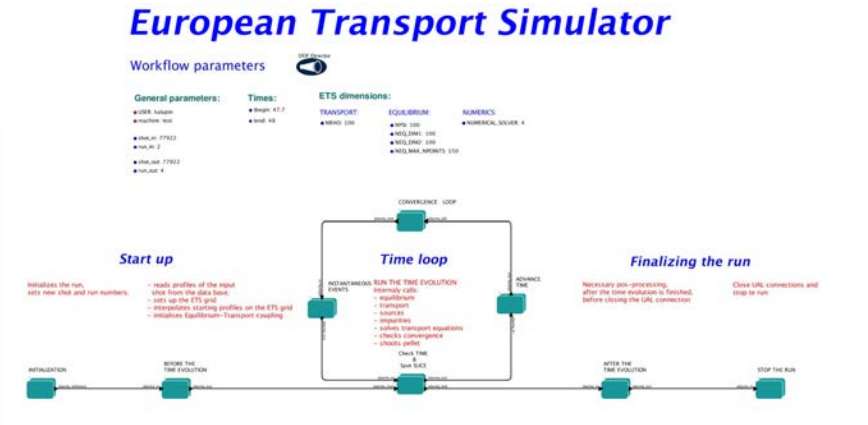
Choosing a Director for a Kepler Workflow

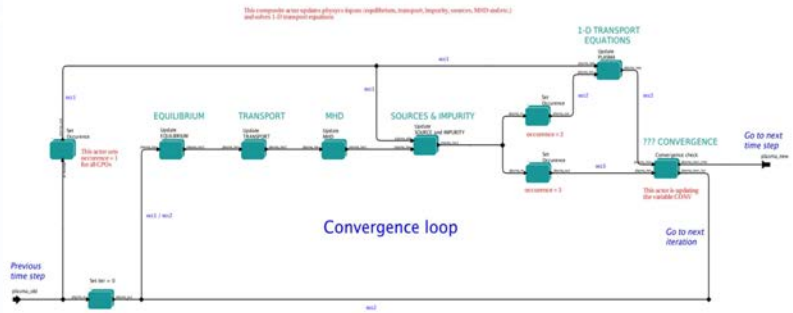
Execution modes

- Remote and external execution components
  - ExternalExecution
  - SSH
  - Grid: Globus, UNICORE, QCG, Nimrod, Serpens
  - Cloud: Amazon, OpenNebula, PaaS (e.g. INDIGO-DC orchestrator) etc.
  - Hadoop
- Can be run as a job itself
  - Docker images available (general, domain specific)
  - Running on HPC via uDocker (Docker images) within user space

Case study - Fusion

- EUROfusion WP CD work on validated suite of simulation tools for ITER plasma
- Whole simulation platform for cross validation between different fusion devices (Kepler as w-f engine)
- Individual codes in Fortran, C++, C, Python, Java and also Matlab
- Many very complex multi level workflows has been developed in Kepler: European Transport Solver (ETS), Turbulence-transport, Equilibrium reconstruction and MHD
- Easy exchange of modules of the same kind, in order to optimize the physics complexity versus the performance (CPU time)





Kepler ecosystem- Fusion

- Production testbed
- 90+ users and workflow developers and support team
- Whole versioning system: versions of the workflows, the actors, kepler, the provenance results, central management
- Running codes in the context of different computational resources (HPC – Marconi@CINECA, Cloud)
- Multi-level parallelism(multi Kepler, multi actors, MPI)
- Whole ecosystem:
  - User friendly interface for configuration an runs
  - Visualisation tools and libraries
  - Profilers
  - Tools that automatically include the physics codes written in different languages

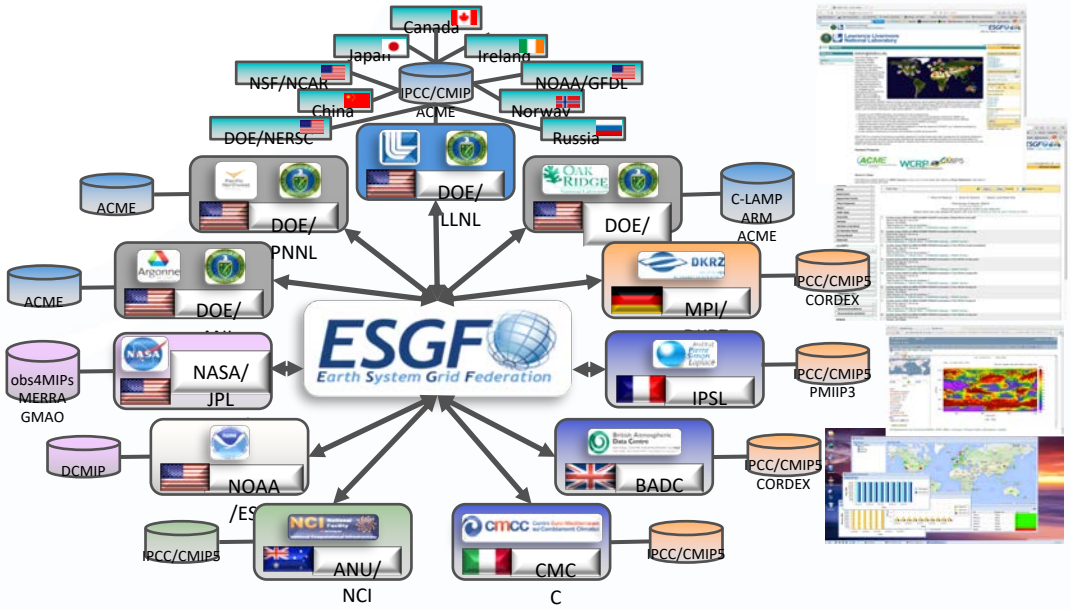
Multi-model climate data analysis case study

- H2020 INDIGO-DataCloud (2015-2017)
- This case study proposed in INDIGO by CMCC was mainly related to the multi-model climate data analysis
- It was directly connected to the Coupled Model Intercomparison Project (CMIP) and to the Earth System Grid Federation (ESGF) infrastructure
- Besides CMCC, several partners (PSNC, UPV, INFN, LIP) contributed in the case study from different point of views (infrastructure, portal, WfMS, cloud technology, etc.)

Case study challenges & issues

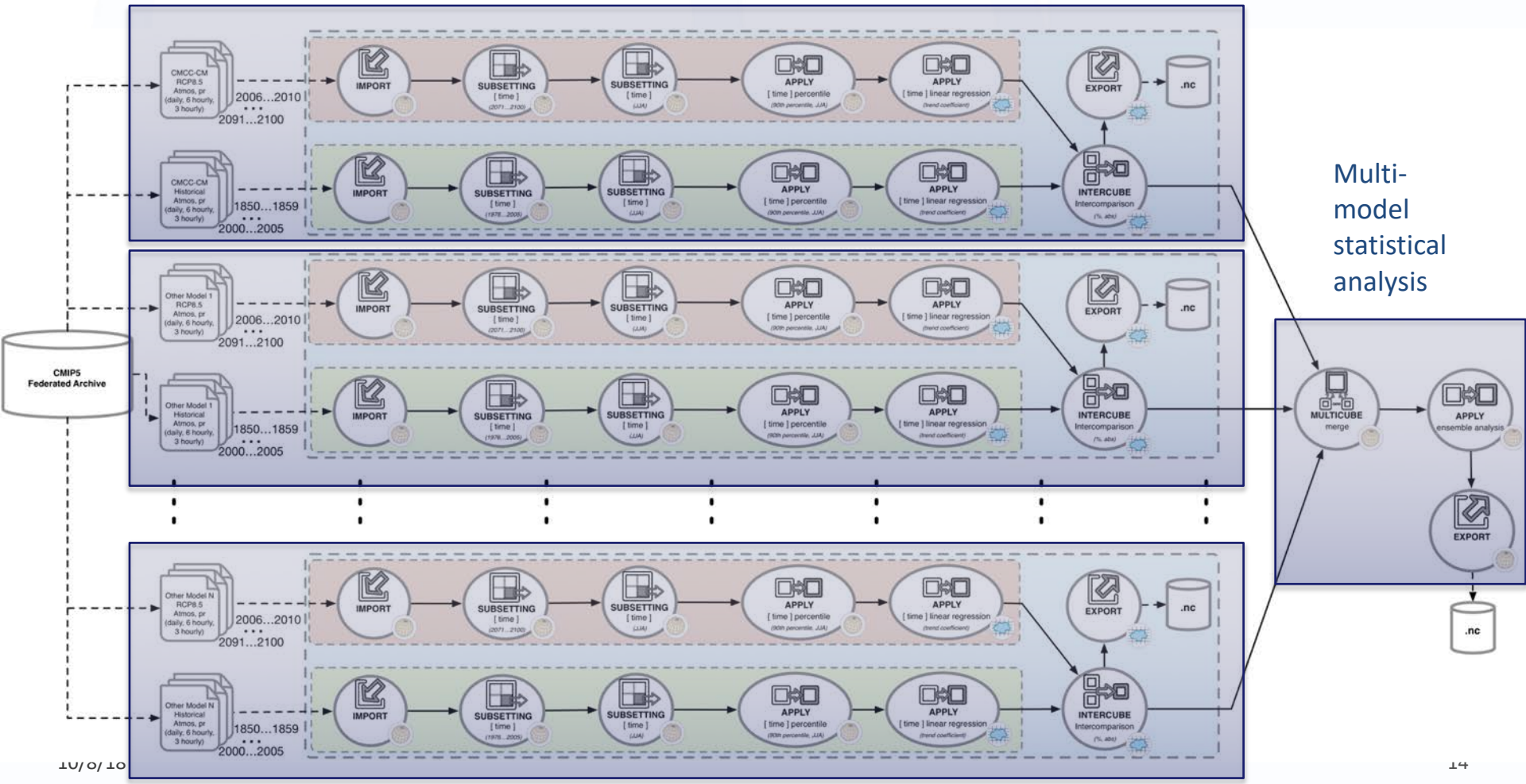
- Input data from multiple models are needed
- Data distribution inherent in the infrastructure
- Data download is a big barrier for end-users (download can take from several days to weeks!)
- Current infrastructure mainly for data sharing
- Data analysis mainly performed using client-side & sequential approaches
- Complexity of the data analysis needs more robust end-to-end support

ESGF Infrastructure and the CMIP5 Federated data Archive



High-level view („workflow”) of the multi-model experiment on precipitation trend analysis

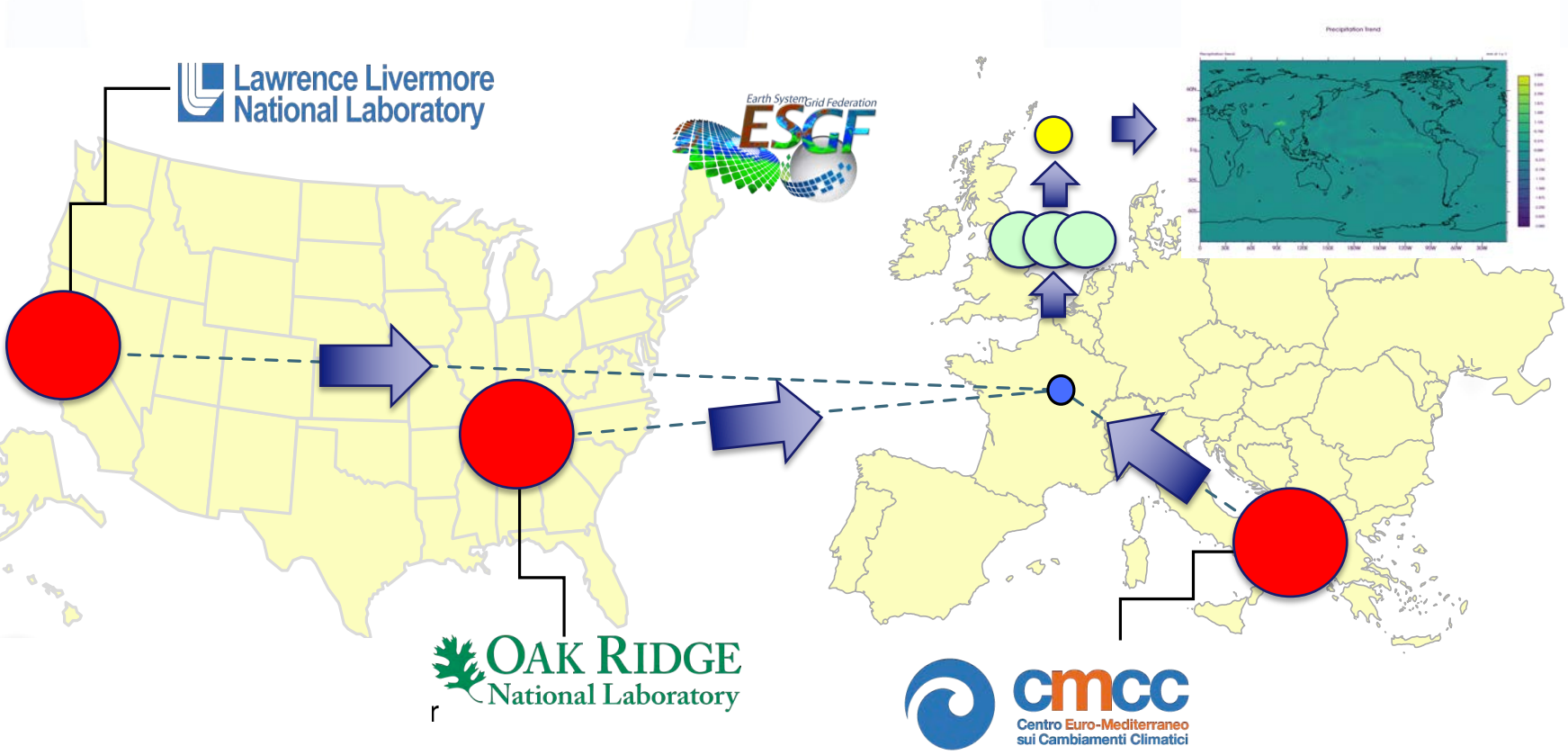
Single model precipitation trend analysis



Multi-model statistical analysis

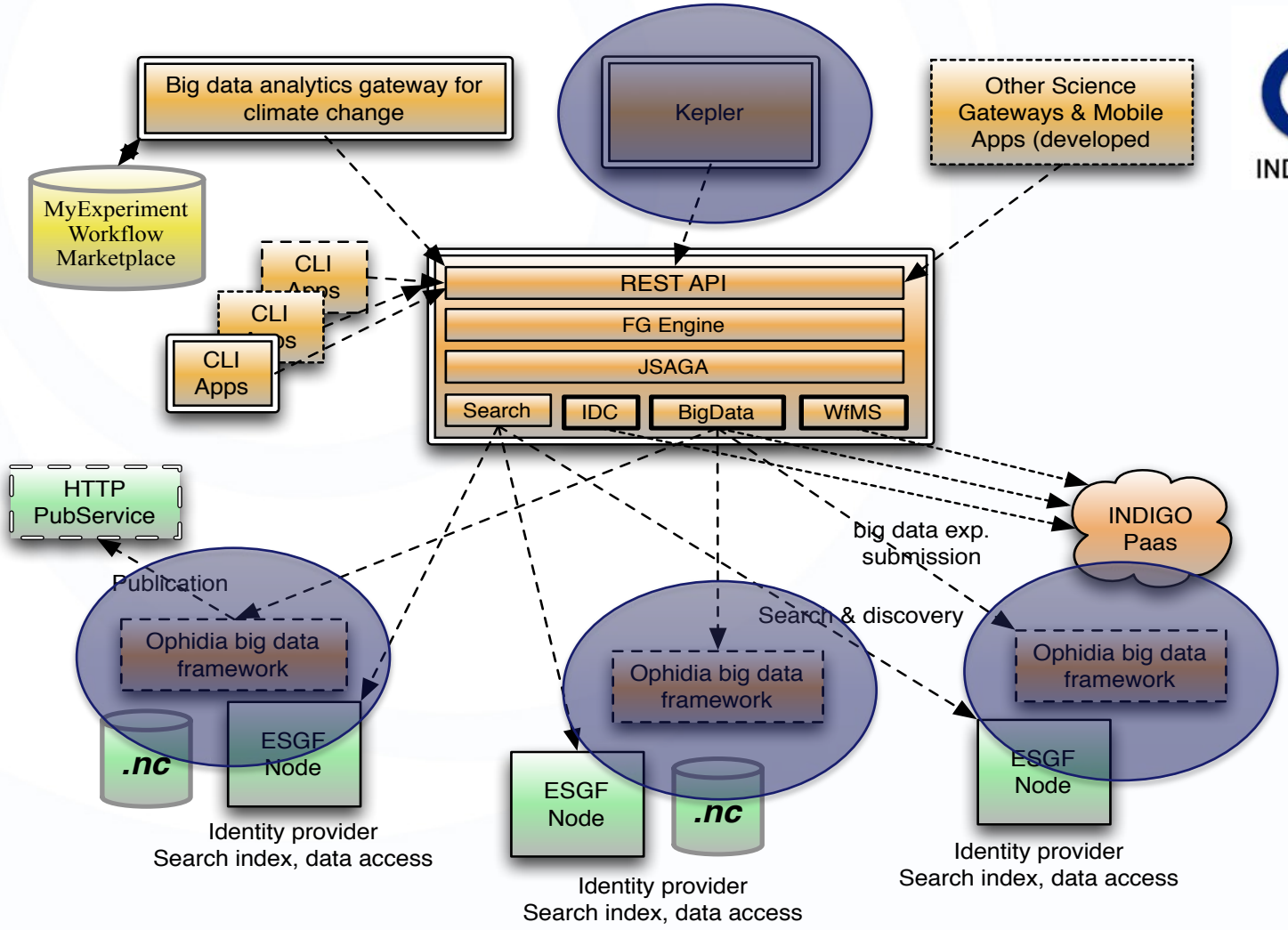
EU/US INDIGO testbed

Current scientific workflow for data analysis supported by ESGF (client-side)



Two-level workflow approach (Kepler-Ophidia) exploited in INDIGO (server-side)

Architectural solution
Running the multi-model experiment



Case study added value

- Paradigm shift from client- to server-side
- Intrinsic data movement reduction
- Lightweight end-user setup
- Re-usability of data, final/intermediate products, workflows, etc.
- Complements, extends and interoperates with the ESGF stack
- Time-to-solution reduction

Use case innovation and future work

- Provisioning of a core infrastructural piece (based on big data and cloud technologies) enabling large-scale data analysis
- Proof of concept level in INDIGO-DataCloud project
- Towards production-level approach in EOSC-hub

EOSC-hub

Services for the European Open Science Cloud

The EOSC-hub project mobilises providers from the **EGI Federation**, **EUDAT CDI**, **INDIGO-DataCloud** and major research e-infrastructures to jointly offer services, software and data for advanced data-driven research and innovation.

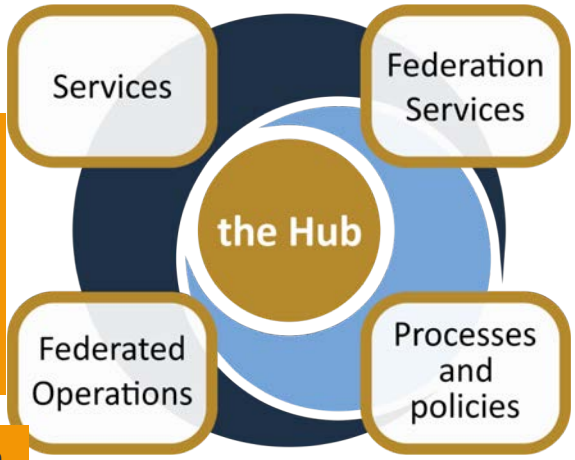
These resources are offered via the **Hub** – the integration and management system of the European Open Science Cloud, acting as a single entry point for all stakeholders.

EOSC-hub Mission

A federated integration and management system for EOSC

From the consortium AND from **external contributors**

- Data
- Applications & tools
- Baseline services (storage, compute, connectivity)
- Training, consultants



Usage according to **Principles of engagement**

- Marketplace
- AAI
- Accounting
- Monitoring

- Lightweight certification of providers
- SLA negotiation
- Customer Relationship Management

Based on **FitSM**

- Security regulations,
- Compliance to standards,
- Terms of use,
- FAIR implementation guidelines

Summary

- Kepler is a mature tool used to build and run the scientific computational workflows
- Ready to use components available
- Many case studies (large communities)
- Used within different computational context
- Multi-model climate data analysis case study developed
- In EOSC-hub Kepler offered within the production quality infrastructure
- Lesson learned
  - Providing fault tolerance is the most consuming part
  - Combining tasks in bunch of tasks

Publications

- Sandro Fiore et al 2017. **Big Data Analytics on Large-Scale Scientific Datasets in the INDIGO-DataCloud Project**. In Proceedings of the Computing Frontiers Conference (CF'17). ACM, New York, NY, USA, 343-348. DOI: https://doi.org/10.1145/3075564.3078884
- S. Fiore, M. Plociennik et al. **Distributed and cloud-based multi-model analytics experiments on large volumes of climate change data in the earth system grid federation eco-system**. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 2911–2918, IEEE, doi = 10.1109/BigData.2016.7840941
- Michal Owsiak, Marcin Plociennik,et al., **Running simultaneous Kepler sessions for the parallelization of parametric scans and optimization studies applied to complex workflows**, Journal of Computational Science, Available online 19 *December 2016, ISSN 1877-7503, http://dx.doi.org/10.1016/j.jocs.2016.12.005*.
- Marcin Plociennik, et al. **Two-level Dynamic Workflow Orchestration in the INDIGO DataCloud for Large-scale, Climate Change Data Analytics Experiments**, Procedia Computer Science, Volume 80, 2016, Pages 722-733, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2016.05.359. (http://www.sciencedirect.com/science/article/pii/S1877050916308341)
- Plociennik, M., Winczewski, S., Ciecielag, P., Imbeaux, F., Guillerminet, B., Huynh, P., Owsiak, M., Spyra, P., Aniel, T., Palak, B., Zok, T., Pych, W. and Rybicki, J. **Tools, methods and services enhancing the usage of the Kepler-based scientific workflow framework**, 2014 Procedia Computer Science, Vol. 29International Conference on Computational Science, pp. 1733-1744
- Płóciennik M., Zok T., Altintas I., Wang J., Crawl D., Abramson D., Imbeaux F., Guillerminet B., Lopez-Caniego M., Campos Plasencia I., Pych W., Ciecielag P., Palak B., Owsiak M., Frauel Y**., Approaches to Distributed Execution of Scientific Workflows in Kepler**, Fundamenta Informaticae, Volume 128, Issue 3, p. 281- 302

Contact:
Marcin Plociennik
marcinp@man.poznan.pl

EOSC-hub

eosc-hub.eu @EOSC_eu