This project has received funding from the European Union's Horizon 2020 research and innovation program through grant agreement 801101.

MAESTRO
DATA ORCHESTRATION

Maestro Project Introduction

François Tessier
Swiss National Supercomputing Centre, ETH Zurich, Lugano, Switzerland

PADAL Workshop 2019
Bordeaux, France

Context

Complex workflows or frameworks in various scientific domains have increasing I/O needs

Institution	Scientific domain	Workflows	Data size (real & projection)
European Centre for Medium-Range Weather Forecasts (ECMWF)	Weather Forecast	Ensemble forecasts, data assimilation,...	25PB/year (2025: 350PB/year)
Paul Scherrer Institute (PSI)	Synchrotron imaging	X-ray spectroscopy, high resolution microscopy,...	10-20PB/year
Cherenkov Telescope Array (CTA)	Astrophysics	Gamma Rays & Cosmic Sources,...	25PB/year

- Workloads with specific needs of data movement
  - Big data analysis, machine learning, checkpointing, in-situ, co-located processes, …
  - Multiple data access patterns (model, layout, data size, frequency)

MAESTRO
DATA ORCHESTRATION

Context

- But the ratio “I/O performance” / “computing power” is decreasing!

Criteria	2007	2017	Relative Inc./Dec.
Name, Location	BlueGene/L, USA	Sunway TaihuLight, China	N/A
Theoretical perf.	596 TFlops	125,436 TFlops	✕ 210
#Cores	212,992	10,649,600	✕ 50
Total Memory	73,728 GB	1,310,720 GB	✕ 17.7
Memory/core	346 MB	123 MB	÷ 2.8
Memory/TFlop	124 MB	10 MB	÷ 12.4
I/O bw	128 GBps	288 GBps	✕ 2.25
I/O bw/core	600 kBps	27 kBps	÷ 22.2
I/O bw/TFlop	214 MBps	2.30 MBps	÷ 93.0

- Mitigating the I/O bottleneck from an hardware perspective leads to an increasing complexity and a diversity of the multiple tiers
  - Node-local storage (PCIe, SATA)
  - Burst buffers like Cray DataWarp, DDN Infinite Memory Engine

Context

- But the ratio "I/O performance" / "computing power" is decreasing!

NEW!

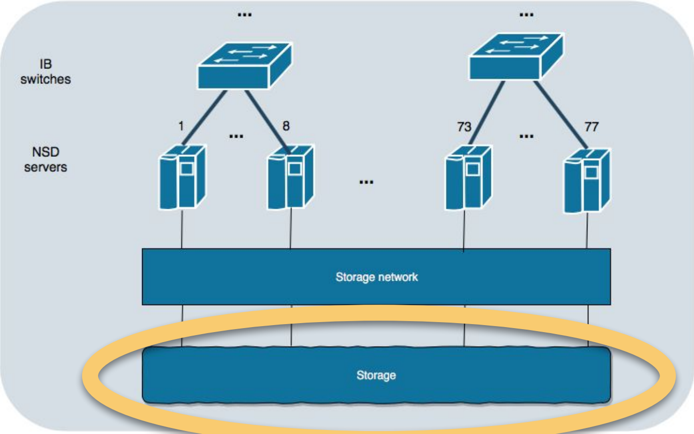
System Specs	TITAN	SUMMIT	FRONTIER
Peak Performance	27 PF	200 PF	>1.5 EF (**✕ 7.5**)
Storage	32 PB, 1 TB/s Lustre file-system	250 PB, 2.5 TB/s GPFS	**2-4x** performance and capacity of Summit's I/O subsystem. Frontier will have near node storage like Summit.

Source: https://www.olcf.ornl.gov/frontier/

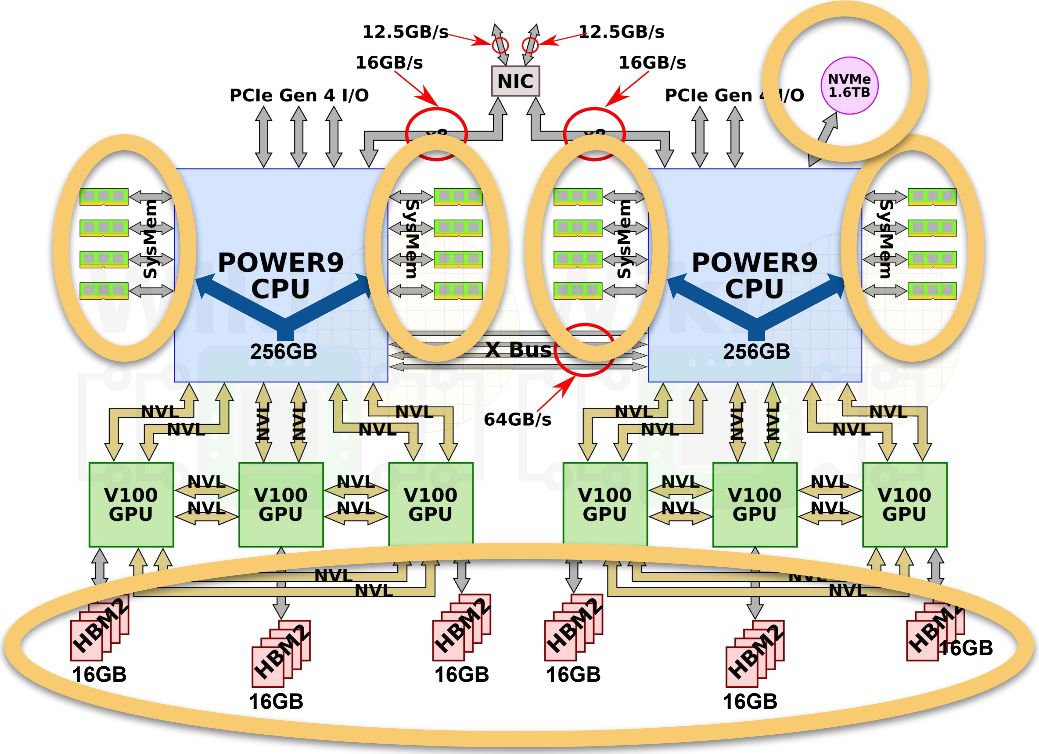
- Mitigating the I/O bottleneck from an hardware perspective leads to an increasing complexity and a diversity of the multiple tiers
  - Node-local storage (PCIe, SATA)
  - Burst buffers like Cray DataWarp, DDN Infinite Memory Engine

MAESTRO
DATA ORCHESTRATION

Hardware Architecture Examples: Summit



Source: https://www.olcf.ornl.gov



Source: https://fuse.wikichip.org

Today’s Shortcomings

Data Awareness

- HPC Software stack focusing on data processing
  - Optimised for filling the processing pipelines
  - Provide means for leveraging parallelism
- Lacking basic data handling at various levels of the stack
  - Lacking functionality for controlling data handling
  - Lacking (unified) semantics for guiding data transport

Memory Awareness

- Missing information about available memory/storage hardware and its characteristics
  - Lacking ability for making data transport decisions
  - Missing information leads to hardware-neutral decisions
- Challenging variety of data access methods
  - Example storage class memory: Block store, file system, object storage
- This becomes more critical with deeper memory and storage hierarchies

Maestro

- Maestro will build a **data and memory-aware middleware framework** that addresses the ubiquitous problems of data movement in complex memory hierarchies that exist at multiple levels of the HPC software stack.

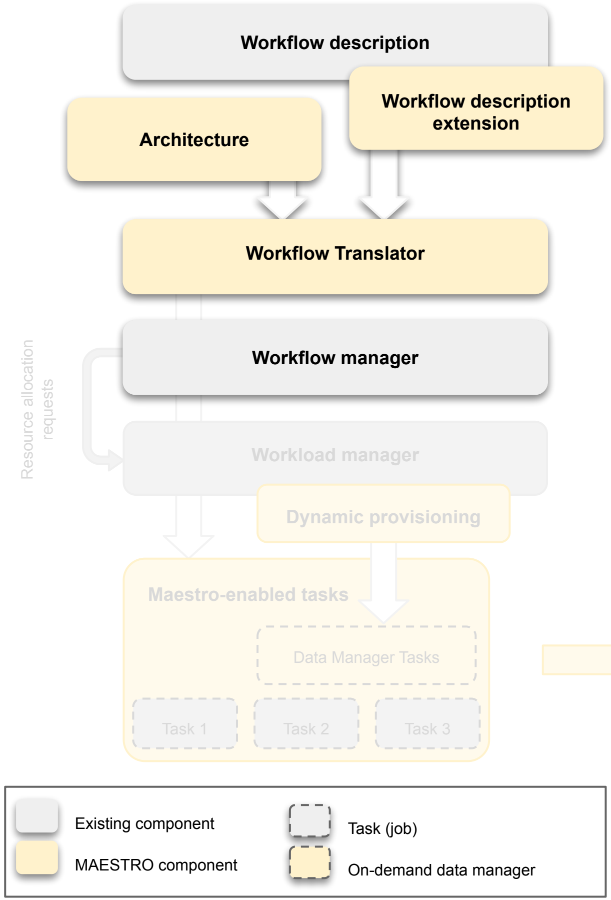
- 3-year European project, started in September 2018, involving partners from academia and industry

101101100101

DATA AWARENESS

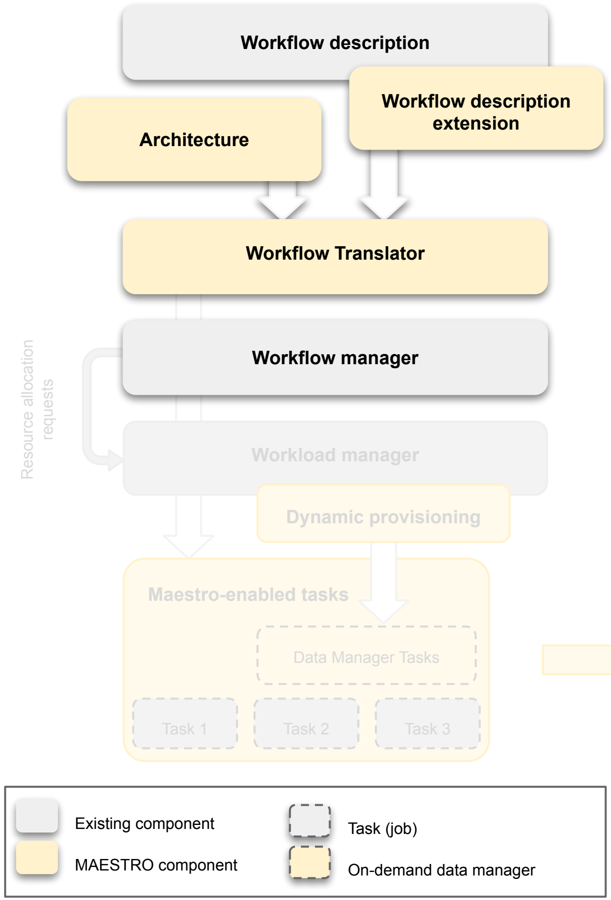
MEMORY AWARENESS

Design of the Maestro middleware



MAESTRO
DATA ORCHESTRATION

Design of the Maestro middleware



Architecture

Lustre parallel file-system
Near-compute SSD nodes (DataWarp)
Memory hierarchy

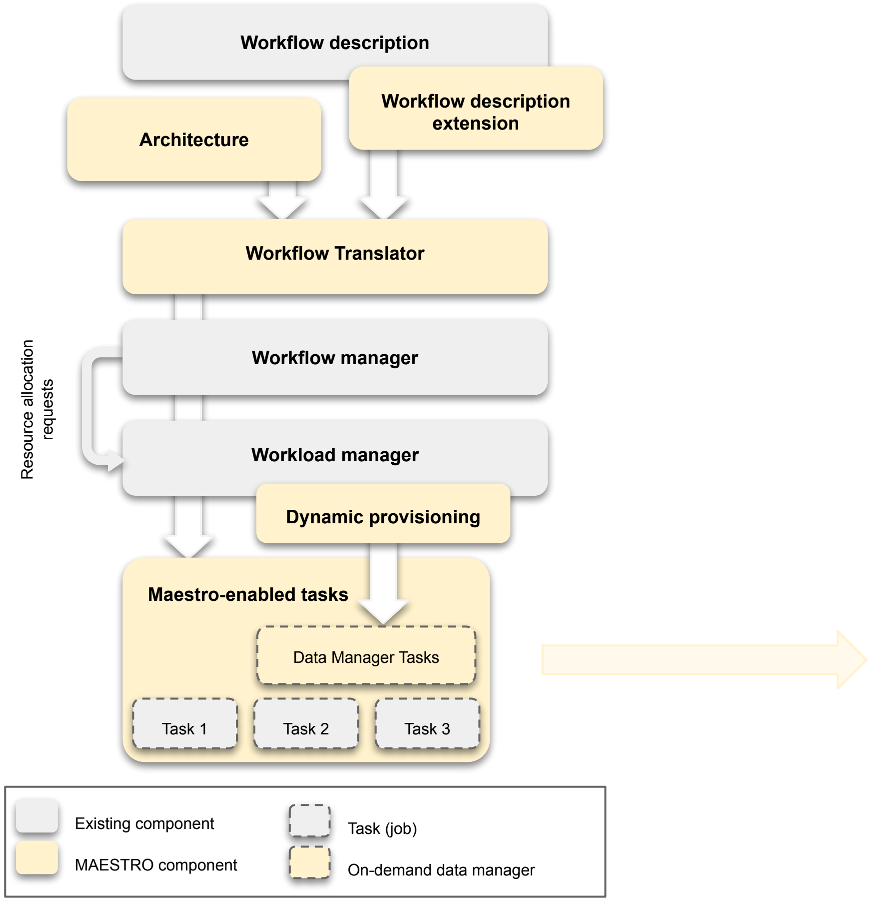
Maestro attributes for Task 1

```
maestro.workload.requirements.persistency: workflow_lifetime
maestro.workload.characteristic.metadata_intensive
maestro.workload.data_management: ‘file’
```

Translator

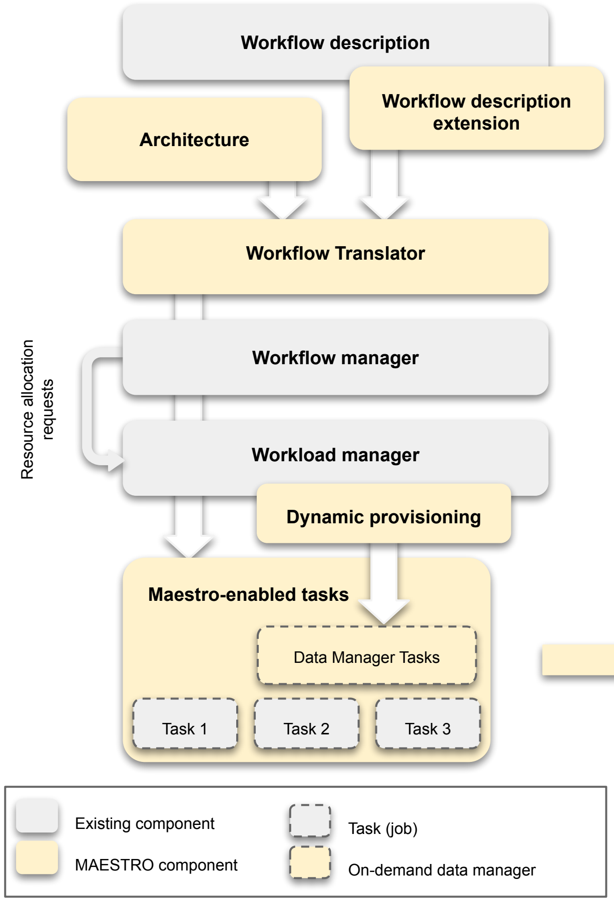
Task 1 close to intermediate storage nodes
Dynamically provisioned BeeGFS (extra task)
New graph of dependencies

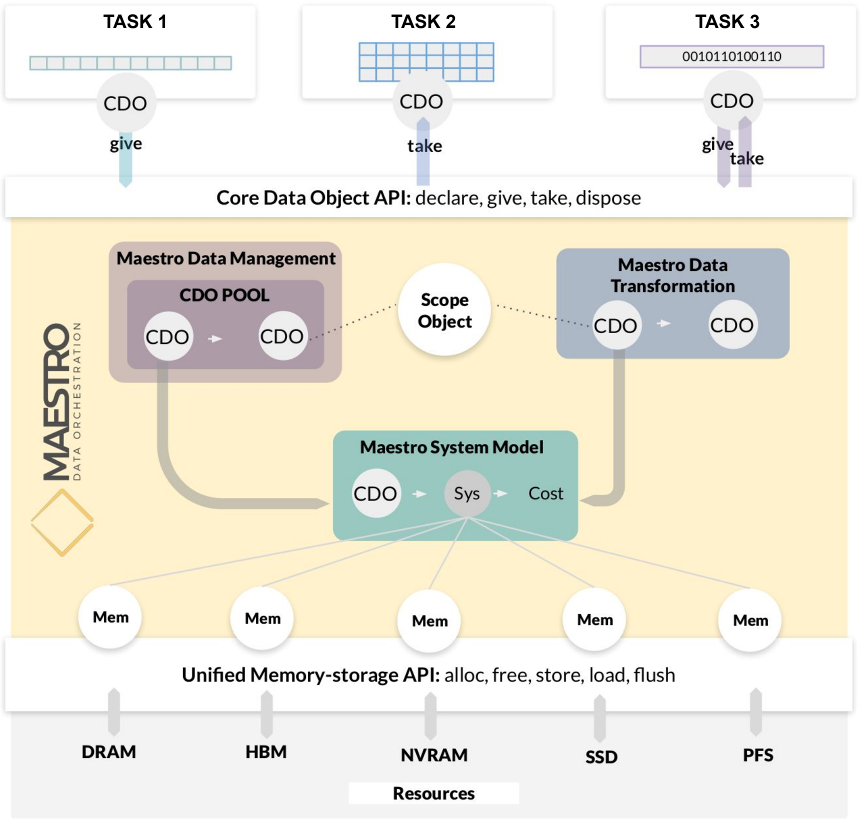
Design of the Maestro middleware



MAESTRO
DATA ORCHESTRATION

Design of the Maestro middleware





MAESTRO
DATA ORCHESTRATION

Design of the Maestro middleware

CDO
CDO (Core Data Object)
It is at the heart of Maestro's design and is used to encapsulate data and metadata. Supports dependencies.

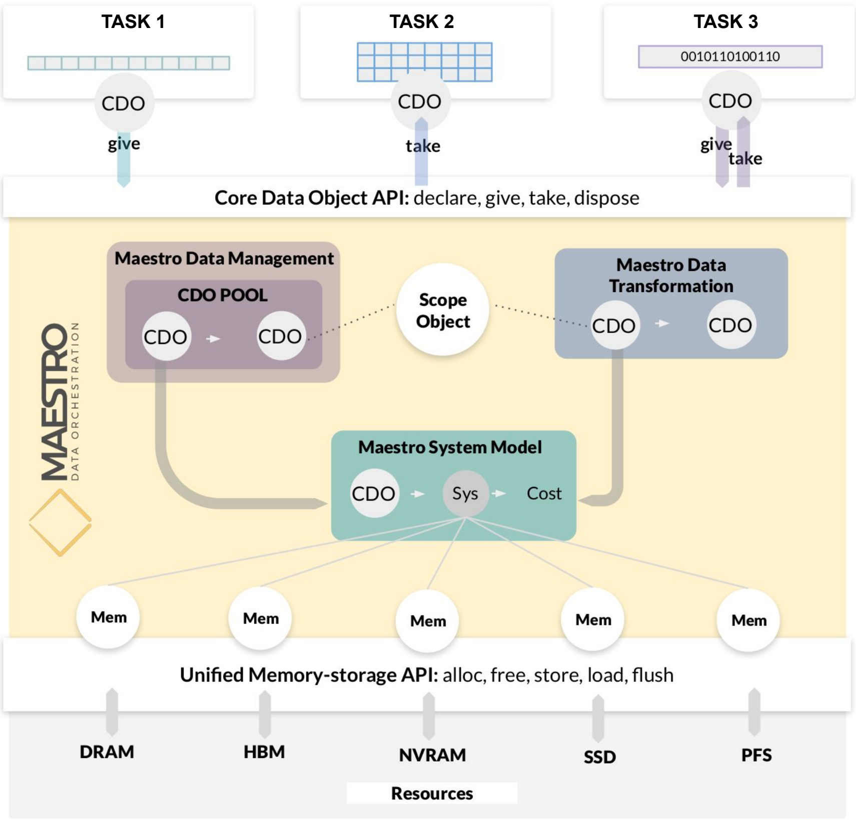
GIVE
Applications give CDOs to the management pool. Maestro manages the data.

TAKE
When an application takes a CDO, Maestro relinquishes all control of the data.

Scope Object
SCOPE OBJECT
Captures information about scope, size, access relations and schedules of the data to enable efficient movement and/or transformation

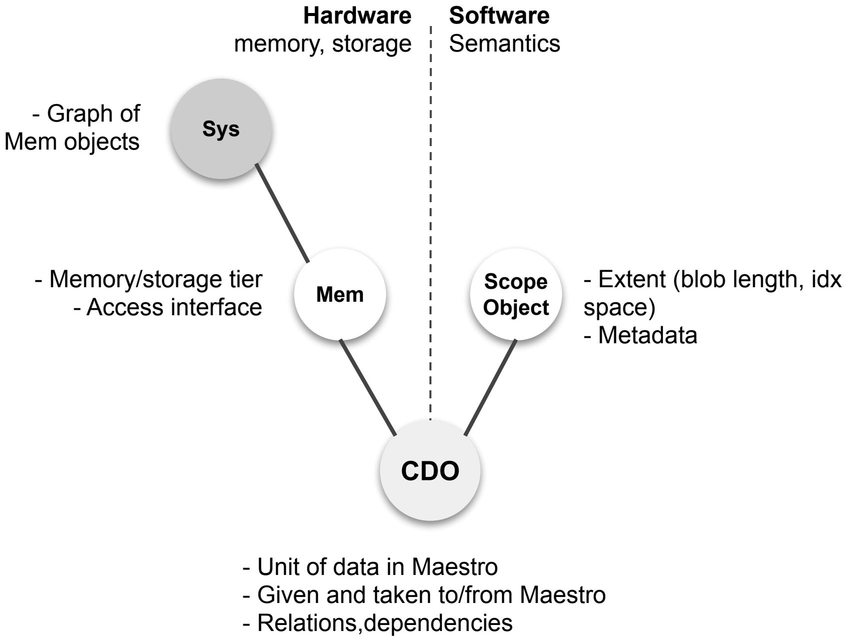
MAESTRO SYSTEM MODEL
Computes the cost of moving, transforming or copying data a CDO

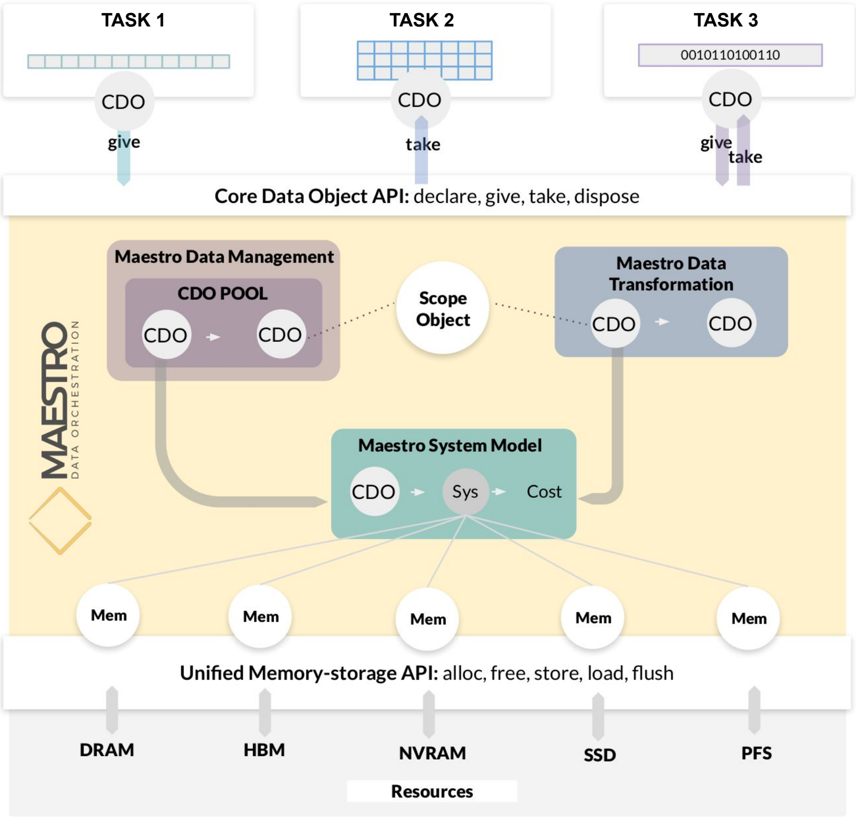
Sys
SYS
Interface to every memory level, enabling core functionality of that memory.



MAESTRO
DATA ORCHESTRATION

Design of the Maestro middleware



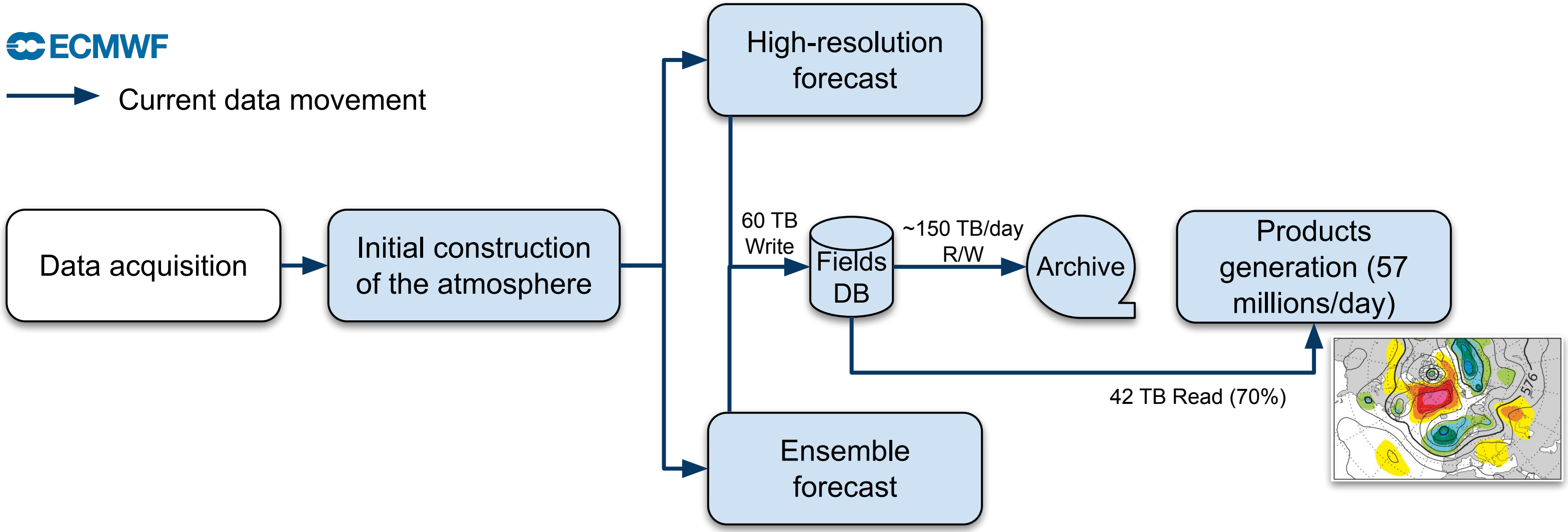


MAESTRO
DATA ORCHESTRATION

Co-Design Applications

- IFS numerical weather prediction system
  - Complex data processing and simulation system with multiple data producers and consumers
- Computational Fluid Dynamics plus in-situ analysis
  - Pipeline coupling multiple simulations plus data post-processing
- Electronic structure calculation library SIRIUS
  - Simulations involving GPU acceleration
- Global Earth Modelling system TerrSysMP
  - Coupled simulations

Example: Weather Prediction Workflow

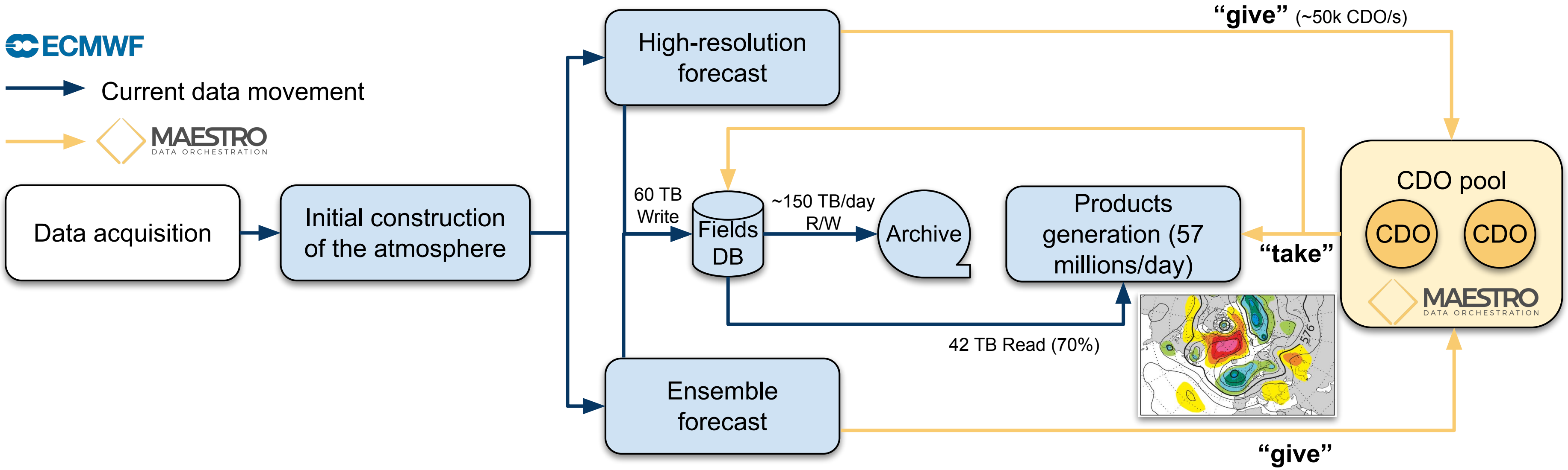


Today’s bottlenecks

- Data movement between forecast stages and product generation
- Irregular archiving of output from research workflows

MAESTRO DATA ORCHESTRATION

Example: Weather Prediction Workflow



Today's bottlenecks

- Data movement between forecast stages and product generation
- Irregular archiving of output from research workflows

Summary and Outlook

- Today's HPC (and HPDA) solutions lack data and memory awareness

- Maestro will develop a data and memory aware middleware
    - Abstractions based on data objects
    - Memory-aware data transport and placement in middleware

- Tag tasks with data-related information, tag data with metadata (ownership, location, size, and so on)

- Open for providing early access to technology

Project Schedule

- Requirements definition completed in August 2019
- Core design fully specified by April 2020
- Start application demonstration this autumn
- Project completion in August 2021

Conclusion

Thank you for your attention!

francois.tessier@cscs.ch

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Acknowledgment

- This work is part of the MAESTRO EU Project
- 3-year European project, started in September 2018
- **Middleware library that automates data movement across diverse memory systems**
- https://www.maestro-data.eu/

MAESTRO
DATA ORCHESTRATION

MAESTRO

DATA ORCHESTRATION

JÜLICH FORSCHUNGSZENTRUM · cea · appentra make code parallel · ETH zürich · CSCS · ECMWF · SEAGATE · CRAY