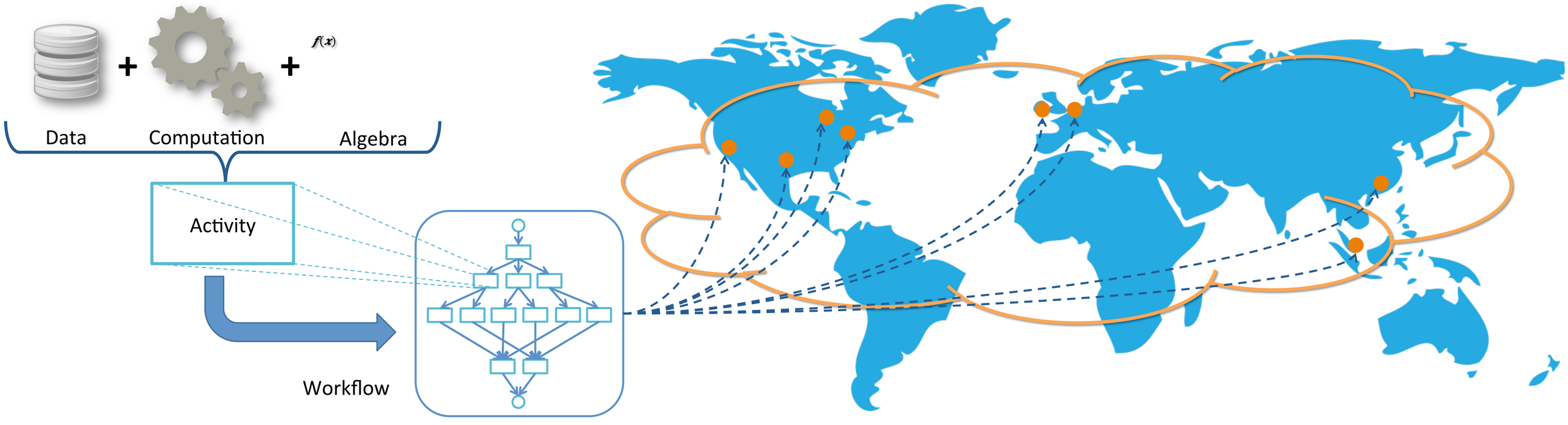
Z-CloudFlow: Enabling Data-Intensive Scientific Workflows on Geographically Distributed Clouds

J. Liu, E. Pacitti, P. Valduriez, M. Mattoso



Goals

- Develop a framework for the efficient execution of scientific workflows in the cloud
- Leverage the cloud infrastructure capabilities for handling and processing big data
- Validate using synthetic benchmarks and real-life scientific applications

Challenges

- What strategies to use for **big data storage and transfer**?
- How to **couple task and dataset scheduling** to **minimize data transfers**?
- How to realize **workload balancing** between data centers to avoid bottlenecks?

Approach

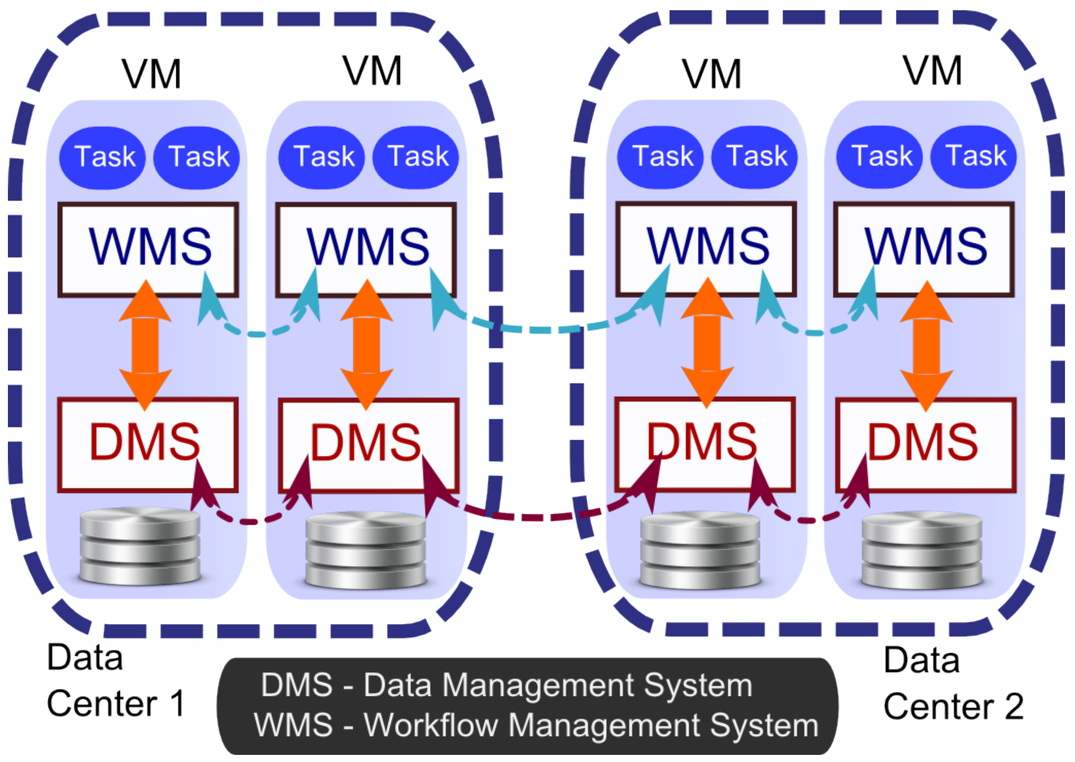
- Adapt workflow processing to the multisite cloud environment
  - Exploit multisite cloud capabilities
- Adopt an algebraic approach for specifying workflows
  - Eases parallelization, optimization and scheduling
- Process workflow execution plans efficiently by optimizing data transfers during execution
  - Rely on an efficient distributed storage layer
- Build an appropriate cloud storage framework addressing the challenges of multisite clouds

Validation (Microsoft Azure)

- Azure for research Award
- 280K compute hours per year

Motivation

- Big Data: **Petabytes** or more
- **Multi-site** clouds
- Complex processing → MapReduce is **not** enough!



Participants

- INRIA, Kerdata Project-Team, Rennes Bretagne – Atlantique Research Center
- INRIA, Zenith Project-Team, Sophia Antipolis – Méditerranée Research Center
  - Patrick Valduriez
  - Esther Pacitti
  - Marta Mattoso
  - Ji Liu
- Microsoft Research

http://www.msr-inria.fr/projects/z-cloudflow-data-workflows-in-the-cloud/

Microsoft Research - Inria JOINT CENTRE

informatics mathematics Inria

UFRJ

UNIVERSITÉ MONTPELLIER 2 SCIENCES ET TECHNIQUES