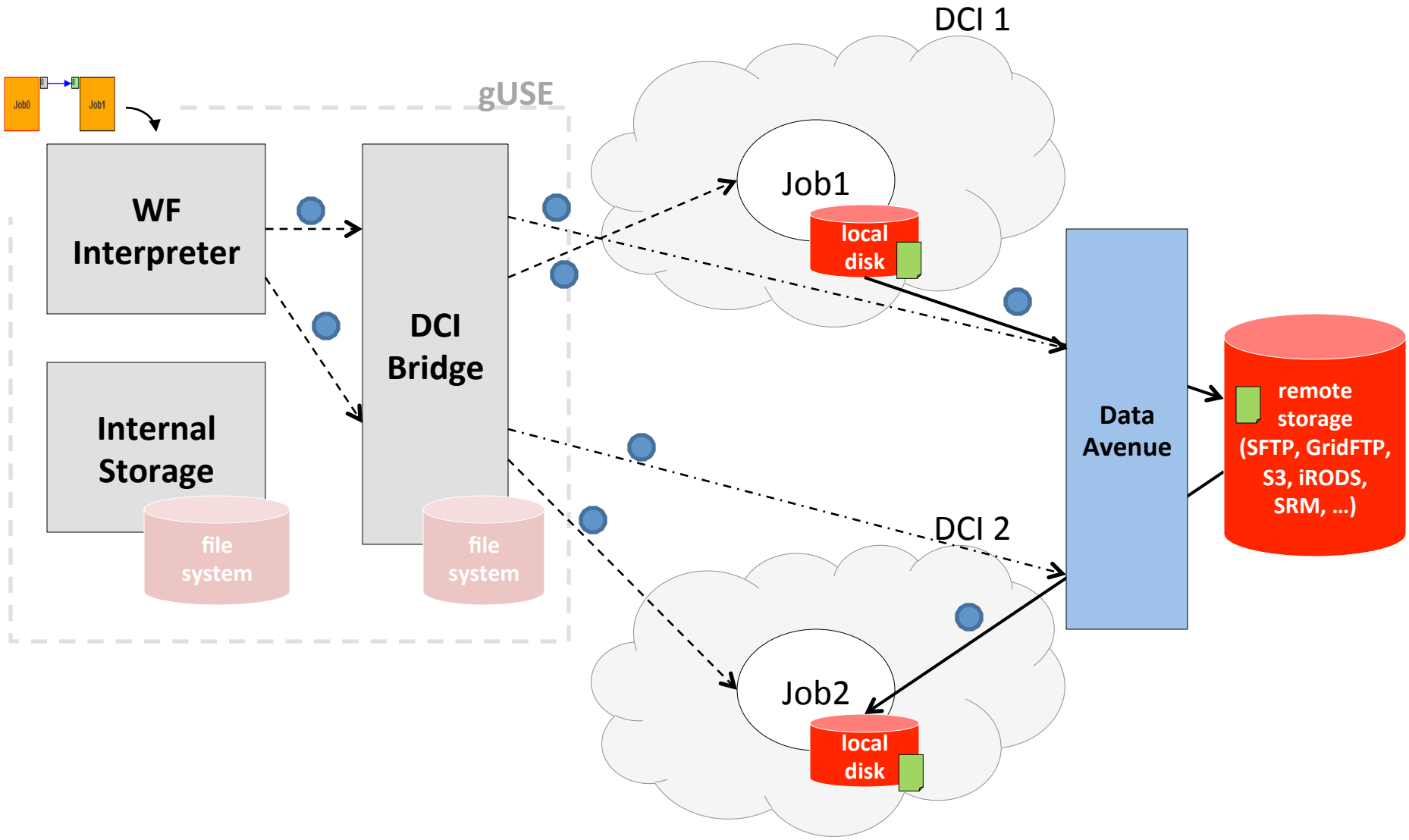
gUSE Data Staging

Ákos Hajnal, István Márton, Peter Kacsuk

The workflow

- The workflow to illustrate data exchange between jobs – potentially - running in different DCIs
- Job1 has output only, Job2 has input only, connected via a “channel”

Data exchange via **data avenue**-managed remote storage



Steps

1. WFI schedules Job1 and sends to DCI Bridge for submission
2. DCI Bridge requests an HTTP alias, *alias1* for the output of Job1
3. DCI Bridge submits Job1 to DCI 1
4. When Job1 is complete, its output is uploaded via Data Avenue alias *alias1* to the remote storage (wrapper)
5. WFI schedules Job2 and sends to DCI Bridge
6. DCI Bridge requests an HTTP alias for the input of Job2, *alias2* (stored in step 4)
7. DCI Bridge submits Job2 to DCI 2 for execution
8. Job2 (wrapper) downloads its input via Data Avenue alias *alias2* and executes

Pros:

- DCIs require no pre-installed tools still can access a wide range of storage resources (including cloud storages) only HTTP (curl)
- DCIs can cooperate as Data Avenue offers access to the same storage
- Credentials are not delegated to worker nodes
- Workflow is portable (separated data access)

Cons:

- Data Avenue can be a bottleneck on massive, concurrent use

A Scalable Data Avenue Service Architecture

To facilitate auto-scaling by acting on *Load* calculated by mod_cluster