Karlsruhe Institute of Technology

Dynamic Provisioning of Opportunistic Resources for HEP

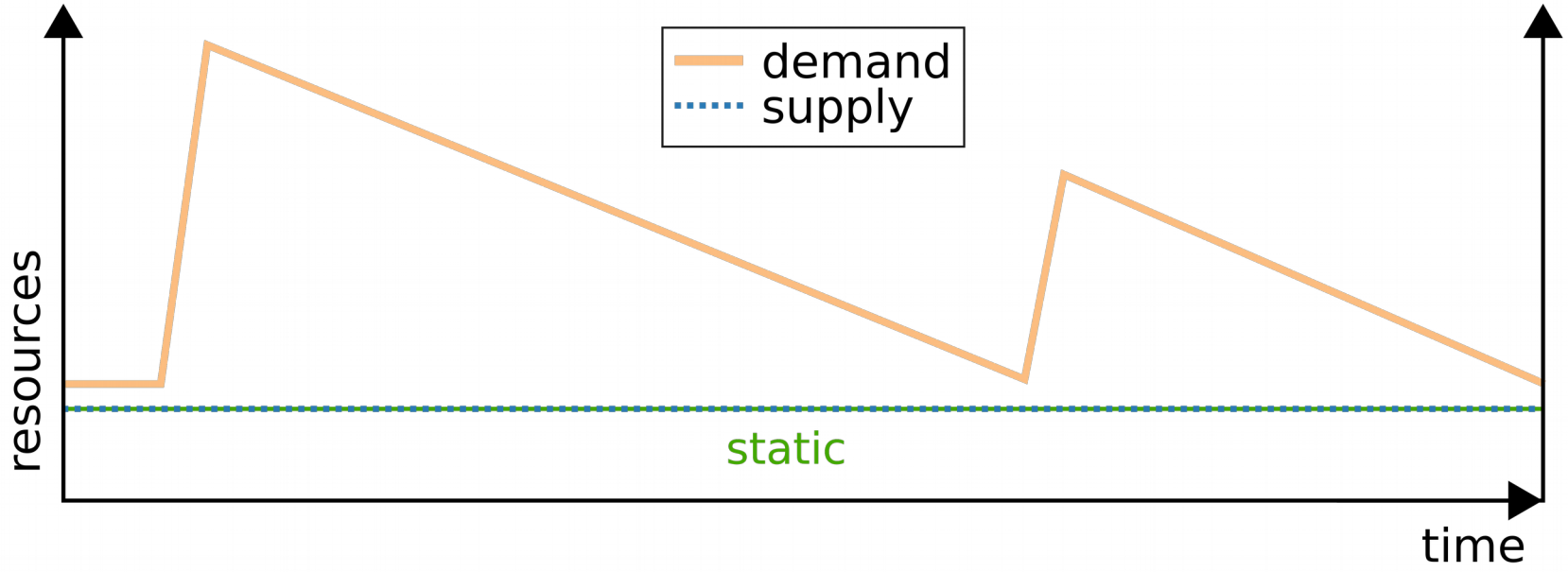
Matthias Jochen Schnepf, Christoph Heidecker, Manuel Giffels, Günter Quast

Terascale Meeting 2017

SCC / ETP

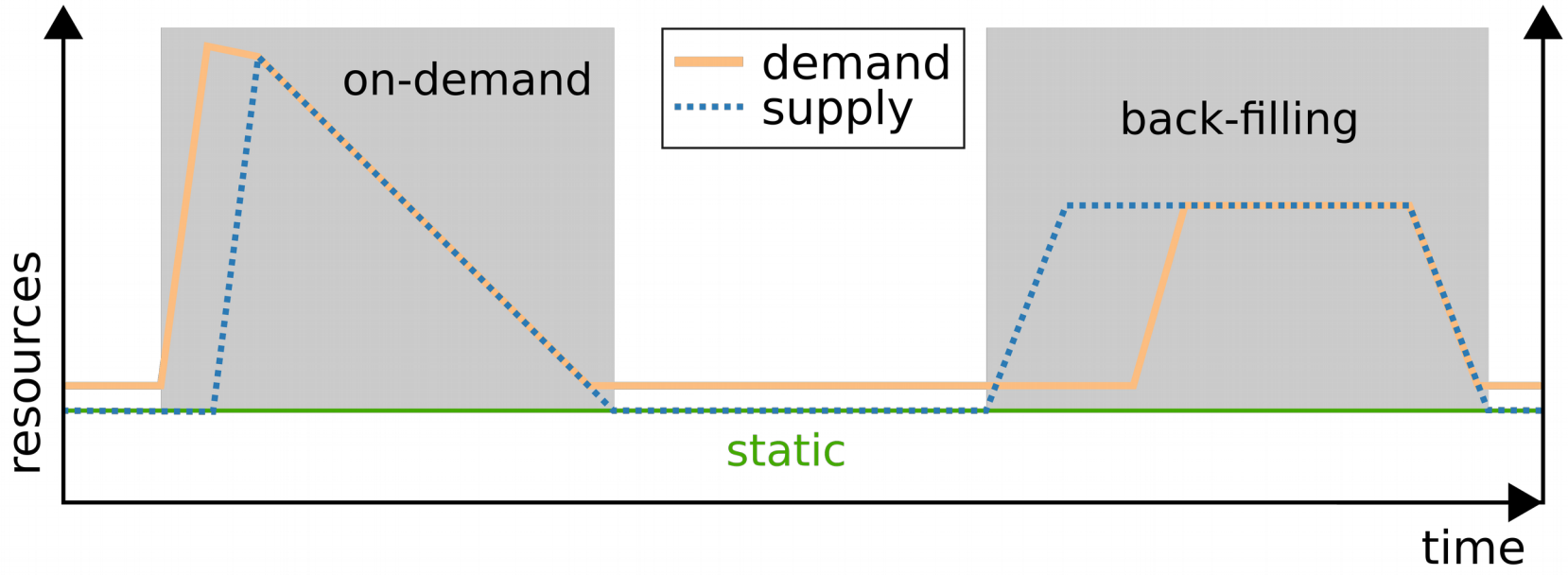
Resource Provisioning

- Static WLCG resources
- Limited job throughput



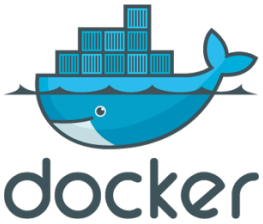
Resource Provisioning

- Add opportunistic resources
  - On-demand booking for peak loads
  - Back-filling of unused resources (cycle stealing)
- Improved job throughput



Batch System

- HTCondor
  - Dynamic integration of worker nodes
  - Supports to run jobs inside container
    - Singularity
    - Docker
  - Send jobs to other HTCondor pools



HTCondor
High Throughput Computing

Example of Opportunistic Resources

- KIT Desktop Cloud
  - Institute desktops with docker container
  - Back-filling to utilize unused resources (cycle stealing)

- HPC ForHLRII@KIT
  - HPC centre supports singularity (testing)
  - Resources allocated on-demand (back-filling)

SCC
Steinbuch Centre for Computing

- Helix Nebula Science Cloud
  - EU project to provide computing resources via VMs
  - On-demand resource provisioning

HELIX NEBULA
THESCIENCECLOUD

- HPC NEMO
  - Shared HPC center supports virtualization
  - On-demand resource allocation via batch system

NEMO
bwForCluster

Example of Opportunistic Resources (2)

remote

- Access to
  - Grid storage via Grid protocols
  - Institute file server via XRootD (testing)

local

- Access to
  - Grid storage via Grid protocols
  - Institute file server via NFS

SCC
Steinbuch Centre for Computing

HELIX NEBULA
THESCIENCECLOUD

NEMO
bwForCluster

Example of Opportunistic Resources (2)

remote

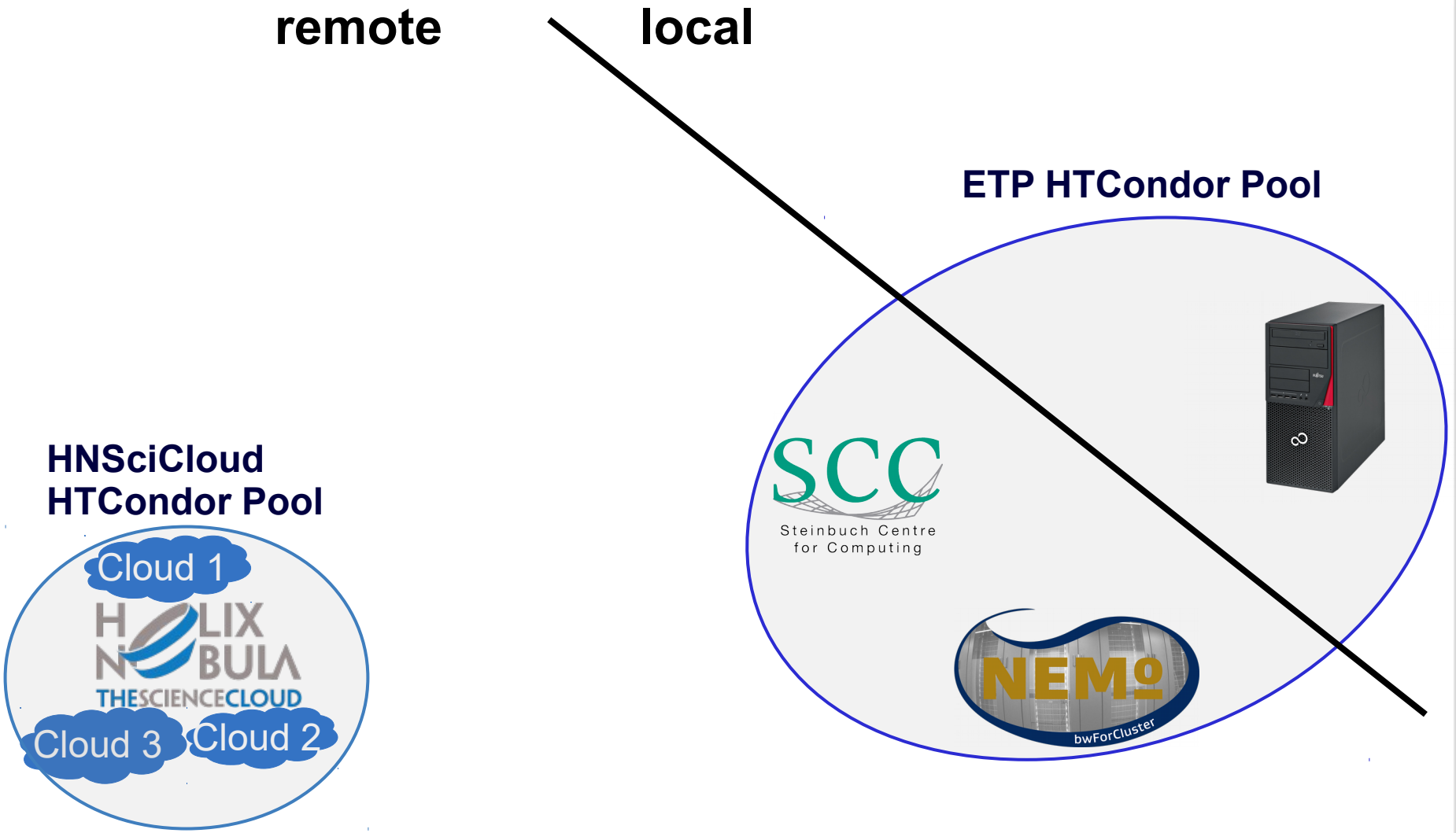
- Access to
  - Grid storage via Grid protocols
  - Institute file server via XRootD (testing)

local

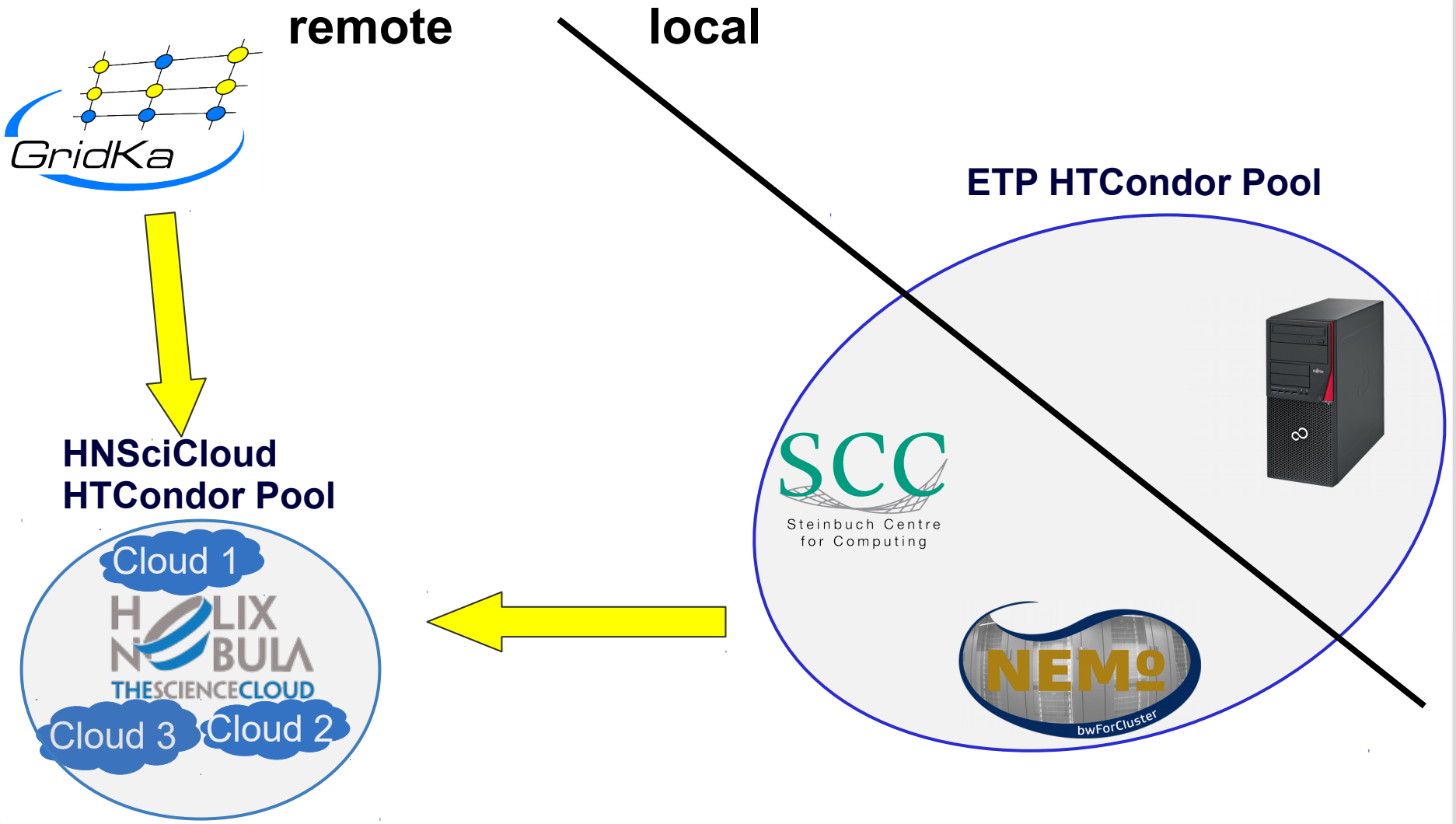
- Access to
  - Grid storage via Grid protocols
  - Institute file server via NFS

User set additional attribute for remote jobs

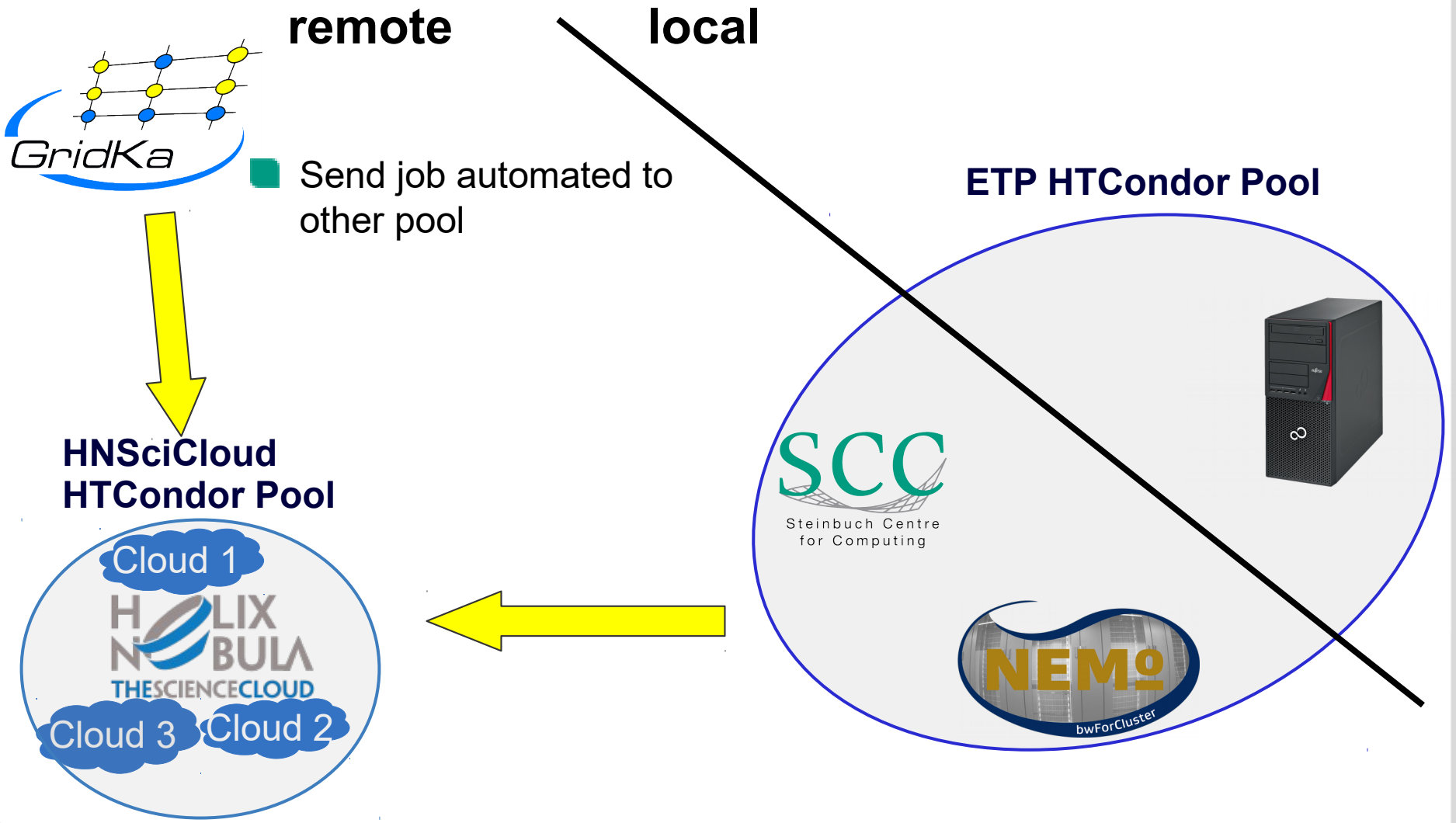
Example of Opportunistic Resources (3)



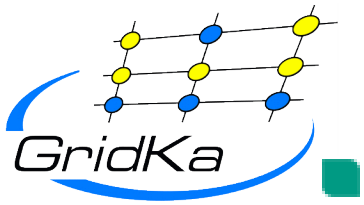
Example of Opportunistic Resources (3)



Example of Opportunistic Resources (3)



Example of Opportunistic Resources (3)

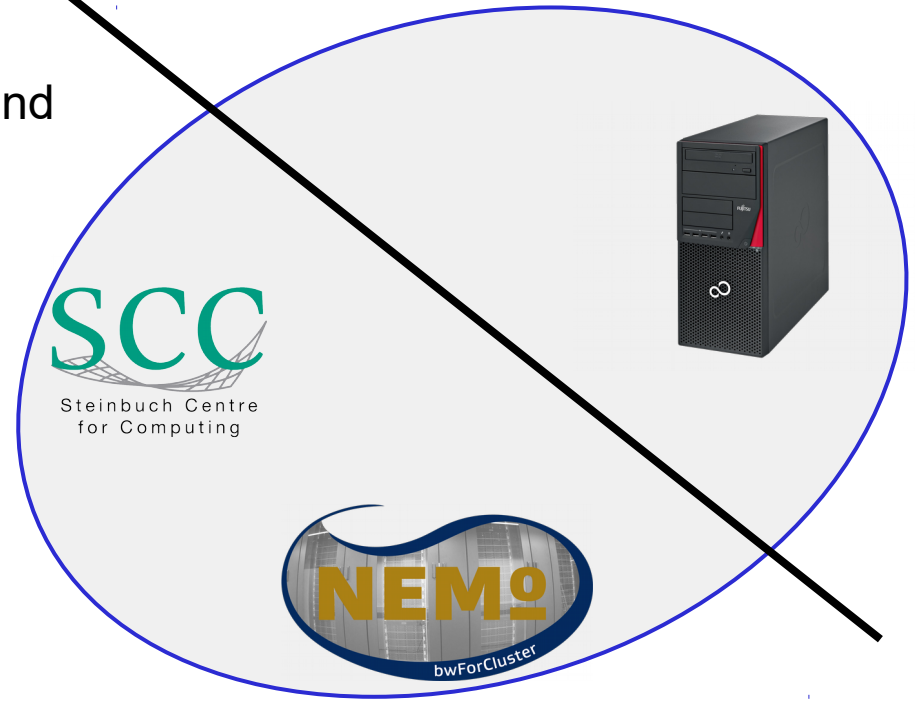


remote **local**

- Send job automated to other pool
  - HTCondor copy job and submit to other pool (job routing)
  - Pools are connected and send job to other pool (flocking)

HNSciCloud HTCondor Pool

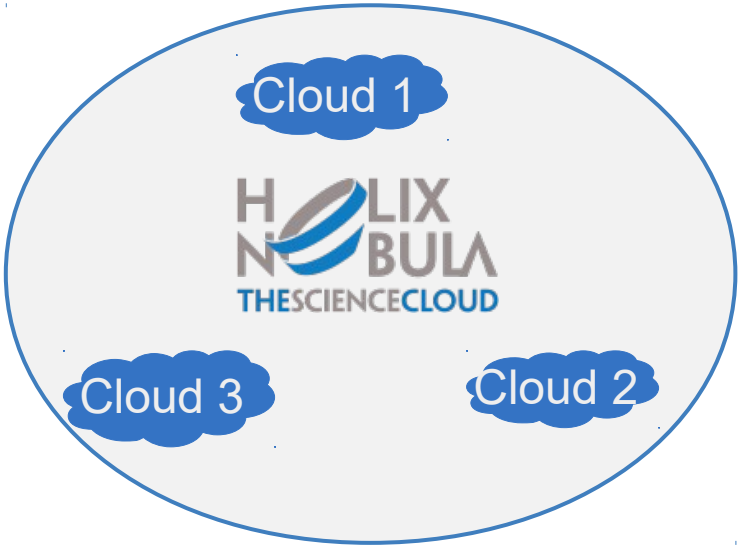
ETP HTCondor Pool



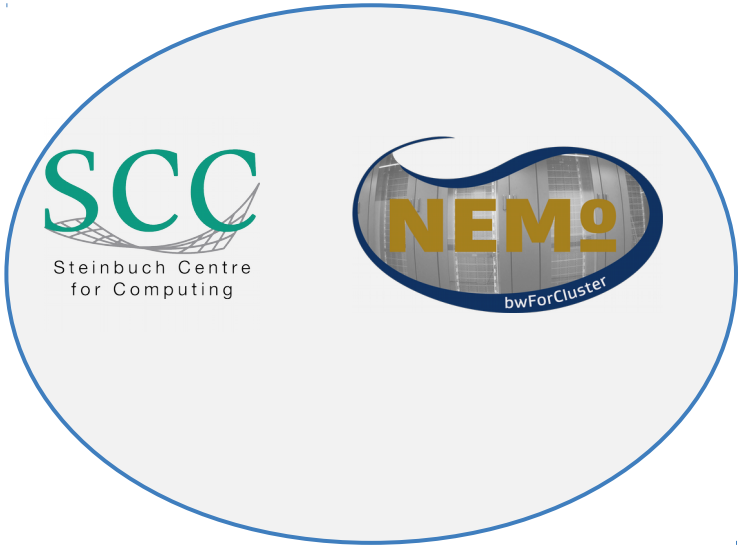
Example of Opportunistic Resources (4)

remote

HNSciCloud HTCondor Pool



ETP HTCondor Pool

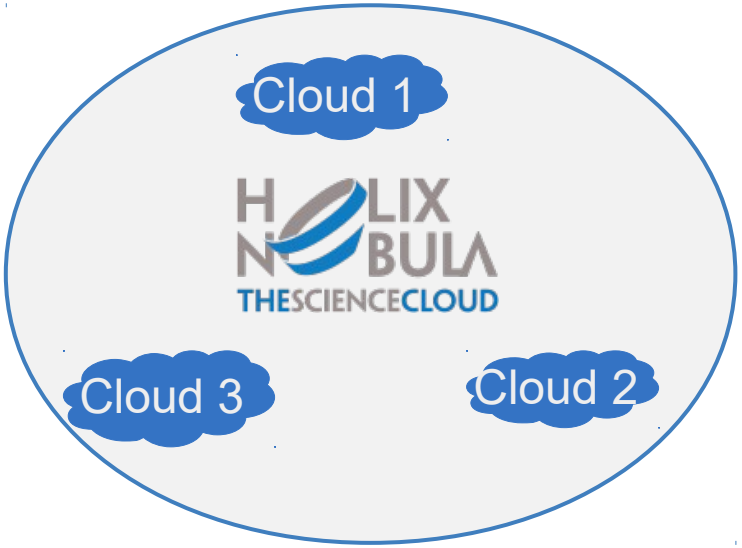


Example of Opportunistic Resources (4)

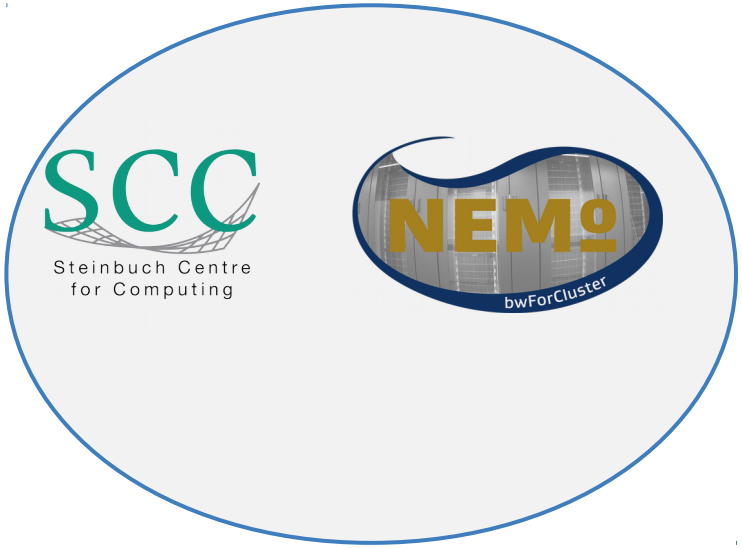
remote

Resources need to be scheduled

HNSciCloud HTCondor Pool

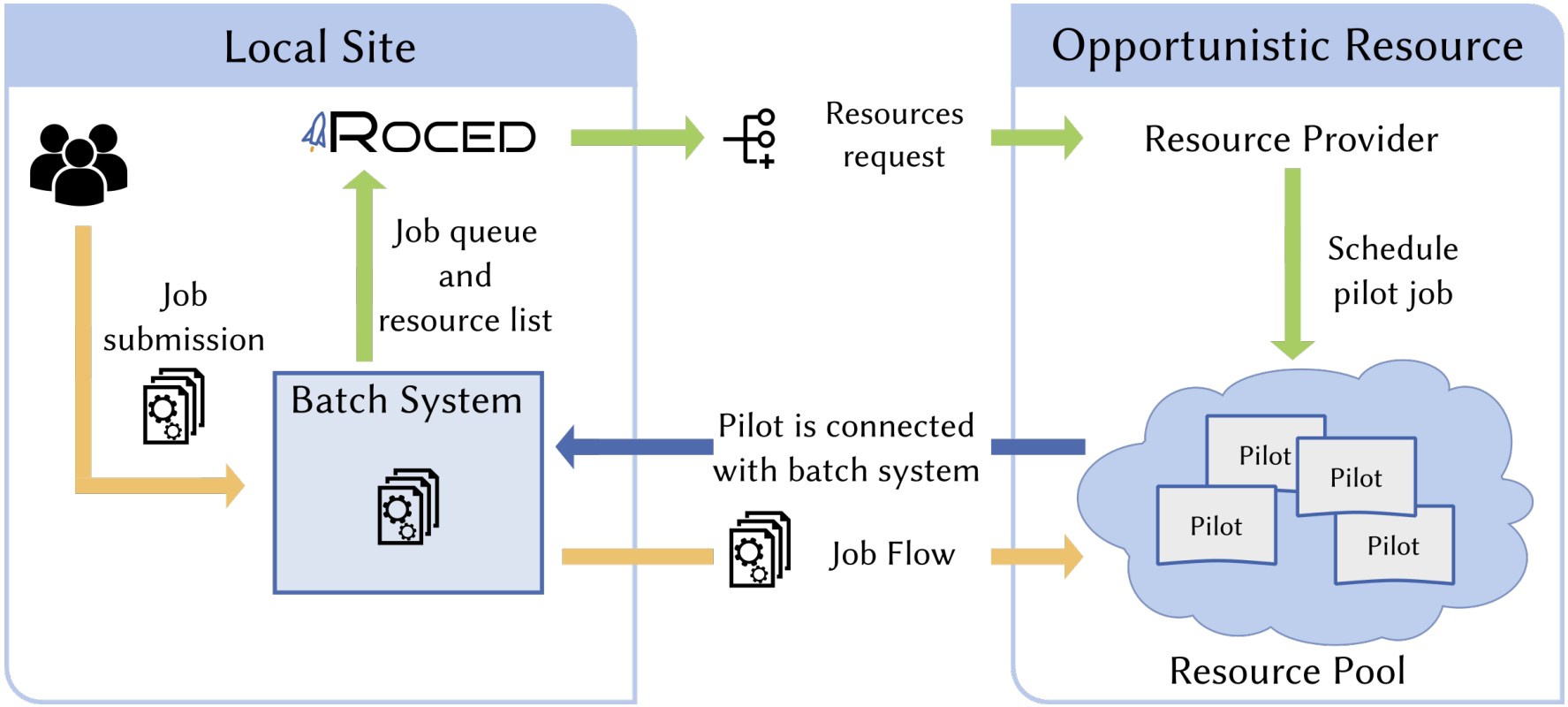


ETP HTCondor Pool



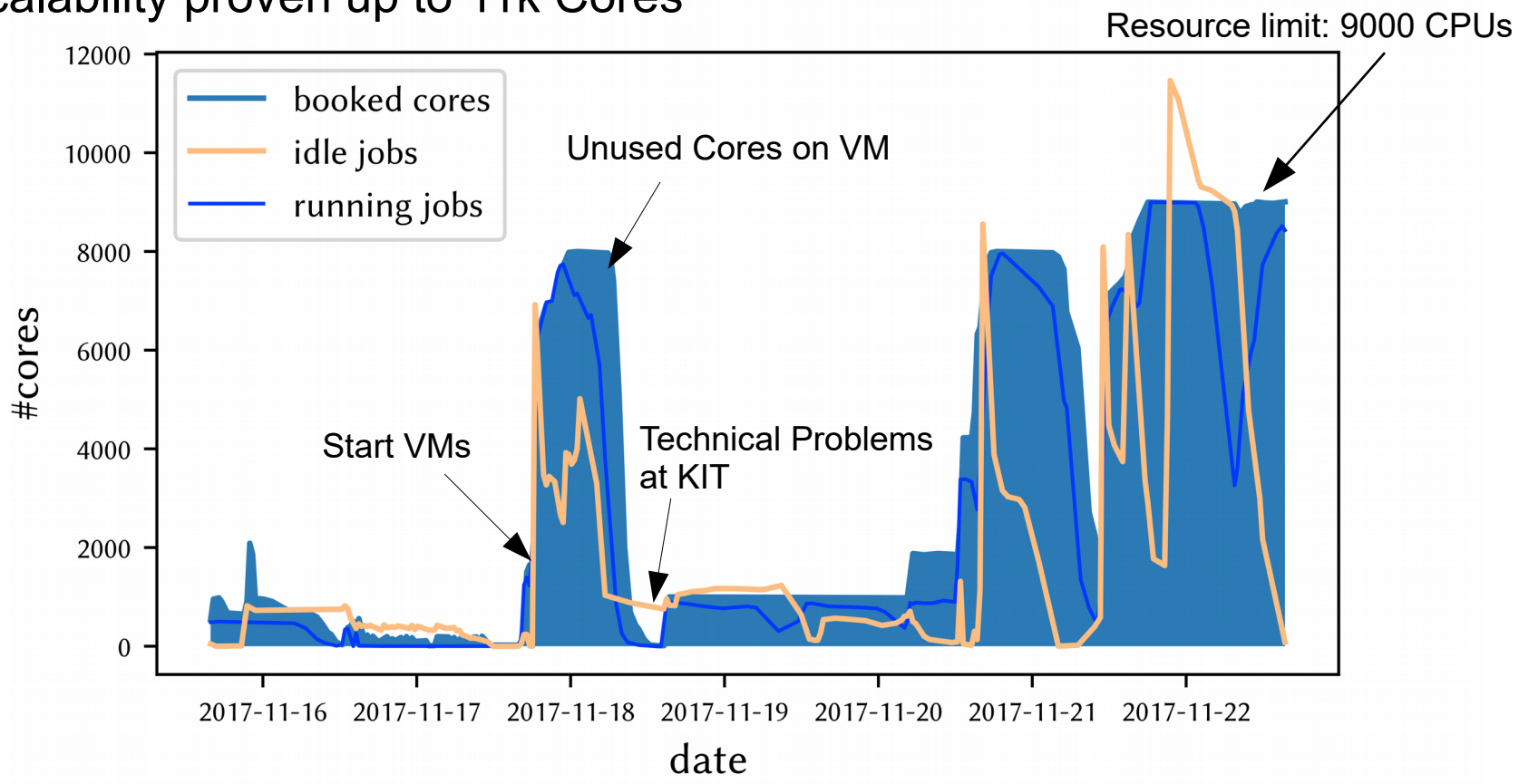
Resource Scheduler: ROCED

- Lightweight management solution developed at KIT
- Support for multiple batch systems and resource providers
- https://github.com/roced-scheduler/ROCED



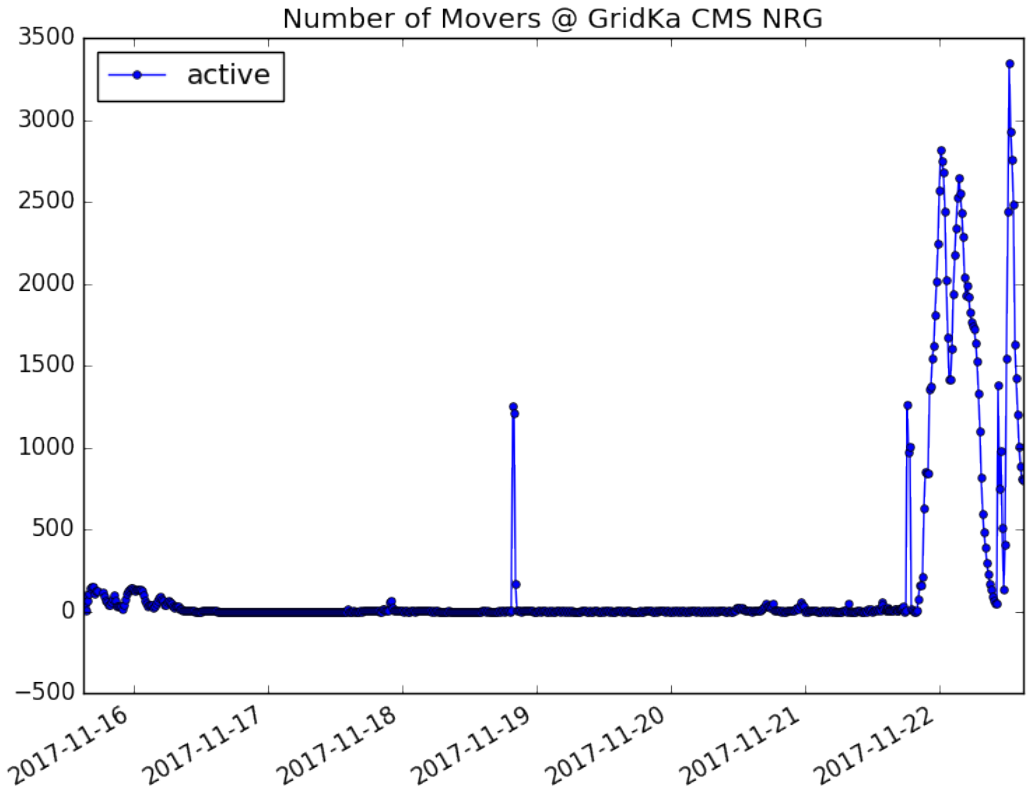
HPC NEMO Usage

- Dynamic on-demand provisioning of VMs (10 Cores, 55GB RAM)
- Maximum walltime of VM set to 1 day
- Scalability proven up to 11k Cores



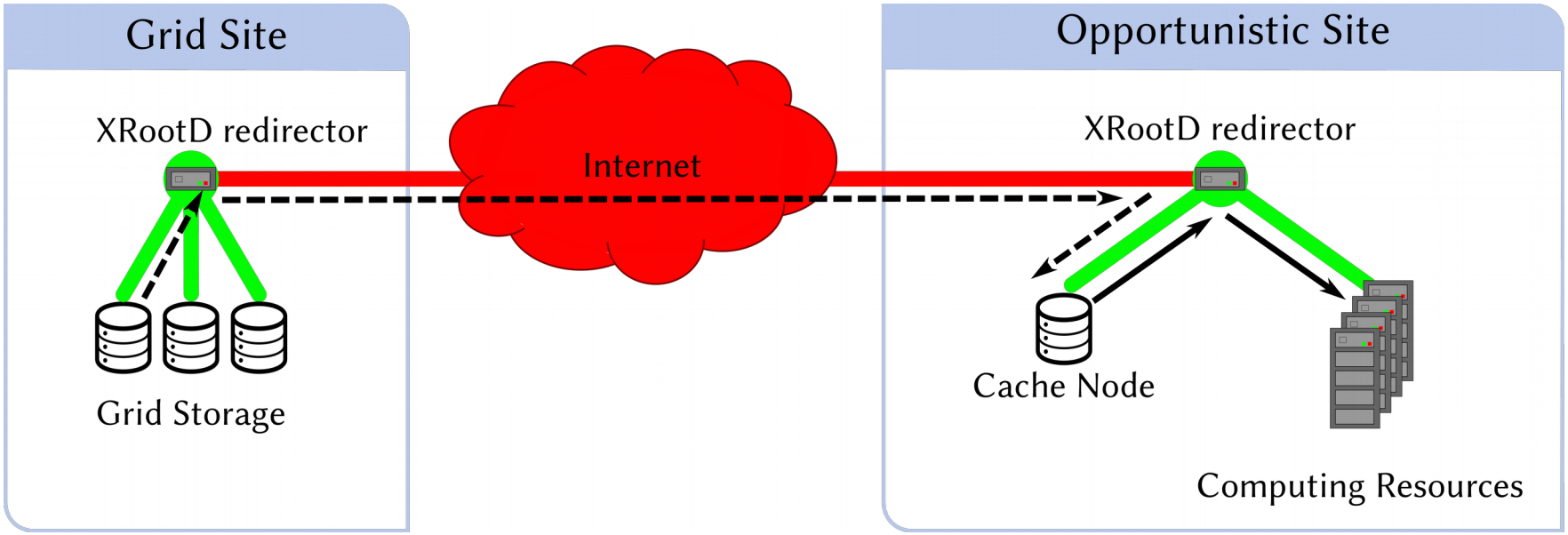
Storage Systems

- Number of concurrent file server accesses
- Storage systems have to scale with computing resources
- Additional network traffic between computing site and storage site



Further Development

- Improve Scheduling with information about
  - Bandwidth capacity of resources
  - Network usage of jobs
- XRootD-Caching at opportunistic sites

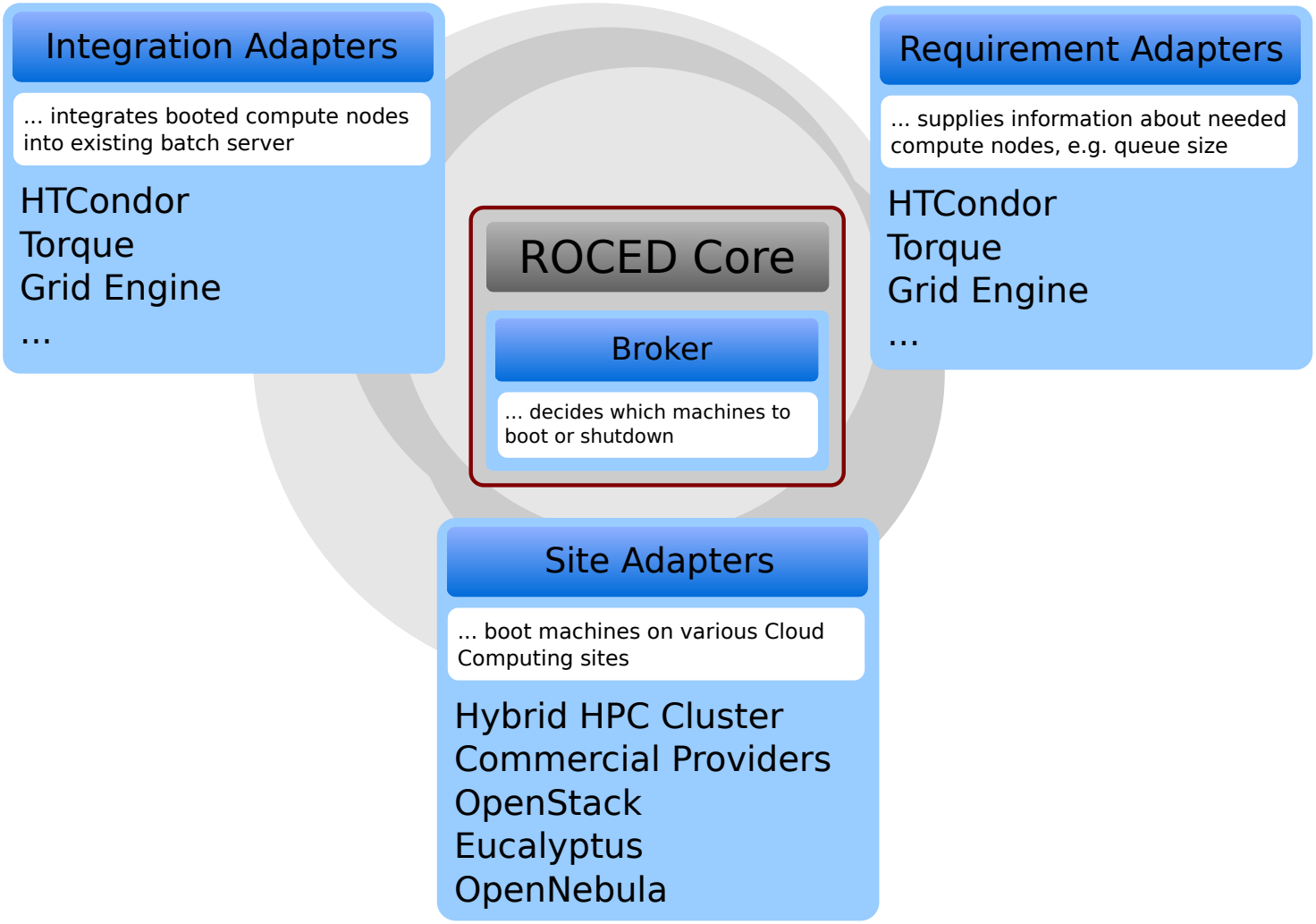


Conclusion

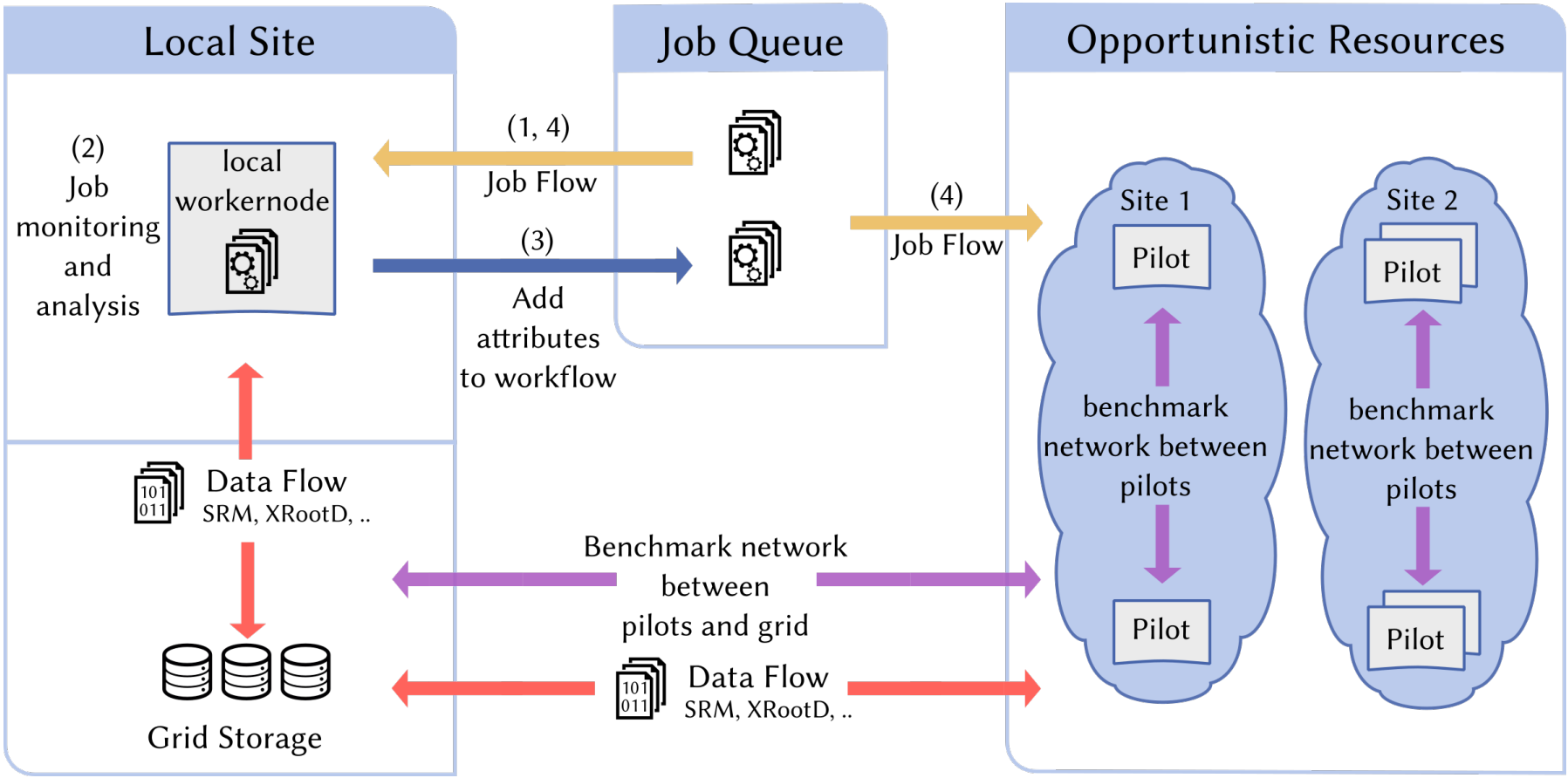
- Dynamic integration of opportunistic resources
  - Organization of resources
  - Opportunistic resources utilizing VMs and container technologies
  - On-demand provisioning of resources using ROCED
- Future research
  - Improve scheduling with information about network usage and utilation
  - XRootD-Caching

Backup

Backup


Integration Adapters
... integrates booted compute nodes into existing batch server
HTCondor
Torque
Grid Engine
...
ROCED Core
Broker
... decides which machines to boot or shutdown
Requirement Adapters
... supplies information about needed compute nodes, e.g. queue size
HTCondor
Torque
Grid Engine
...
Site Adapters
... boot machines on various Cloud Computing sites
Hybrid HPC Cluster
Commercial Providers
OpenStack
Eucalyptus
OpenNebula

Outlook


Local Site
Job Queue
Opportunistic Resources
(2)
Job
monitoring
and
analysis
local
workernode
(1, 4)
Job Flow
(3)
Add
attributes
to workflow
(4)
Job Flow
Site 1
Pilot
benchmark
network between
pilots
Pilot
Site 2
Pilot
benchmark
network between
pilots
Pilot
101
011
Data Flow
SRM, XRootD, ..
Benchmark network
between
pilots and grid
Grid Storage
101
011
Data Flow
SRM, XRootD, ..