

Advancing Cutting-Edge Biological Research with a High-Throughput UNICORE Workflow

UNICORE Summit 2011

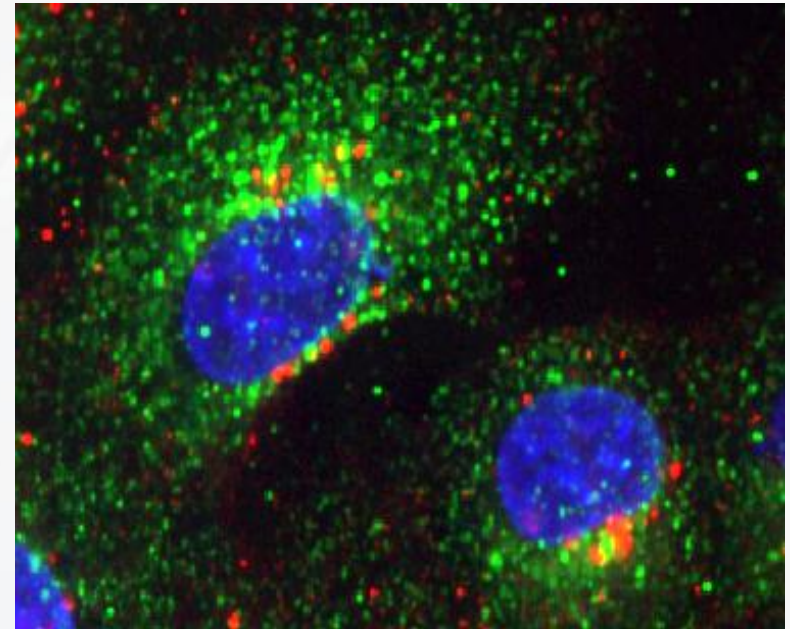
Richard Grunzke
Ralph Müller-Pfefferkorn
Technische Universität Dresden
Center for Information Services and
High Performance Computing

Using the UNICORE Grid Technology in biological research

- Biological Research
- Motivating Choices
- Architecture
- Workflow
- Summary & Outlook

Biological Research

- Max Planck Institute of Cell Biology and Genetics (MPI-CBG), Dresden
- Research of molecular mechanisms of absorption and transportation of molecules in cells
- Parts of cells are highlighted to examine the cellular processes
 - Genomes are prepared/modified
 - Reaction creates light emissions
 - Pictures are taken
 - Image analyses by Software
- Collinet et al. :
Systems survey of endocytosis by multiparametric image analysis;
Nature, 464(7286): 243-249,
March 2010.



Automatic Microscopy

- Automatic microscopes
 - Plates with up to 384 wells (small compartment)
 - Each well contains an sub-experiment
 - Several pictures (e.g. at different depth or different positions) are taken
 - Hundreds of plates per experiment
 - Millions of images (up to 10 MB) of different stages of the biological processes
- Three-dimensional, time-resolved movies (up to several GB) under way



Computing Demands

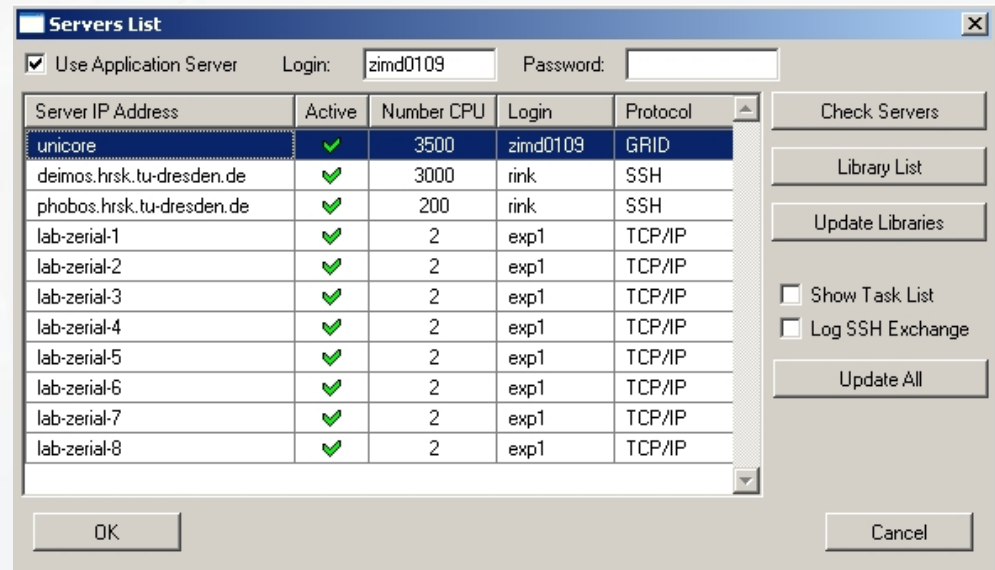
- Many short independent single-core jobs running in farming mode
 - Analysis takes about 10 to 30 minutes per images
 - Batches of tens of thousands of jobs submitted at once
 - Analysis exceeds computational capabilities of MPI-CBG
 - Analysis with MPI-CBG analysis software on ZIH clusters
 - In 2009 and 2010 largest project at ZIH clusters
 - Research possibilities limited by computational power
 - Ever growing demand
- Ideal for Grid Computing

Storage Demands

- One large data set has i.e. about 18 TB in 8 million files
 - Currently about 100 TB stored in NFS at MPI-CBG
 - Access to many small files in NFS slow
 - Need more, more, more storage space
 - Simple metadata stored in directory name
 - Some more metadata in extra files for each image
- Performant storage system with metadata support

User Demands

- Image analysis controlled by Graphical User Interface
 - Windows based workstation
 - Biologists are not computer scientists
 - No deep interest in computing infrastructure
 - It just has to work
 - Just want to click
 - Focus on biology
- Integration into their work environment and GUI



Based on demands

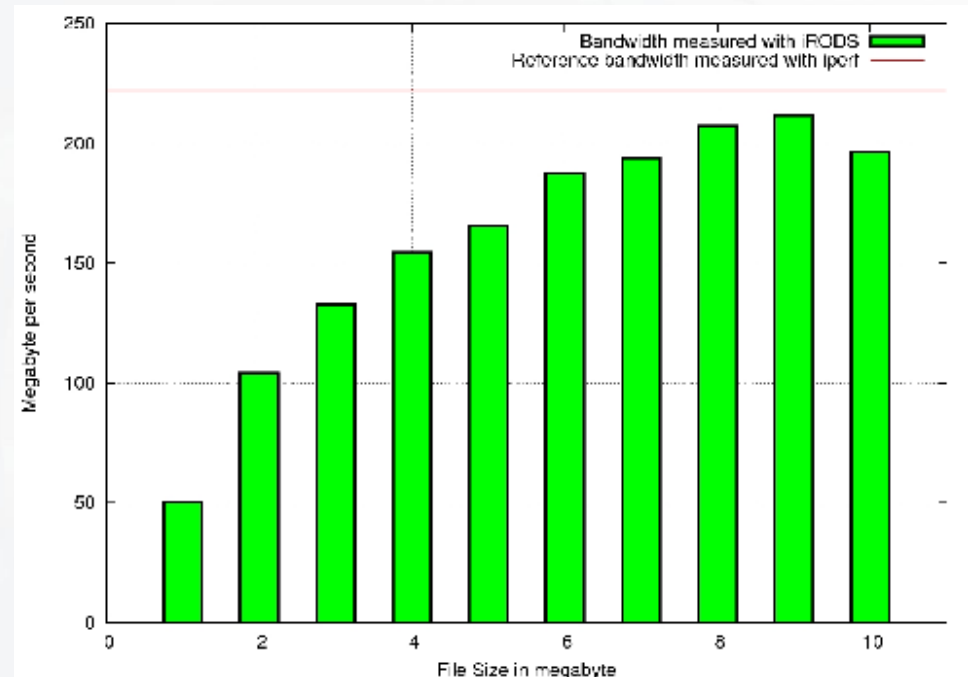


- UNICORE for job submission
 - Clients for Windows
 - Easy to deploy and maintain
 - Easy integration of heterogeneous systems
 - Mature
 - Great support and continuous development

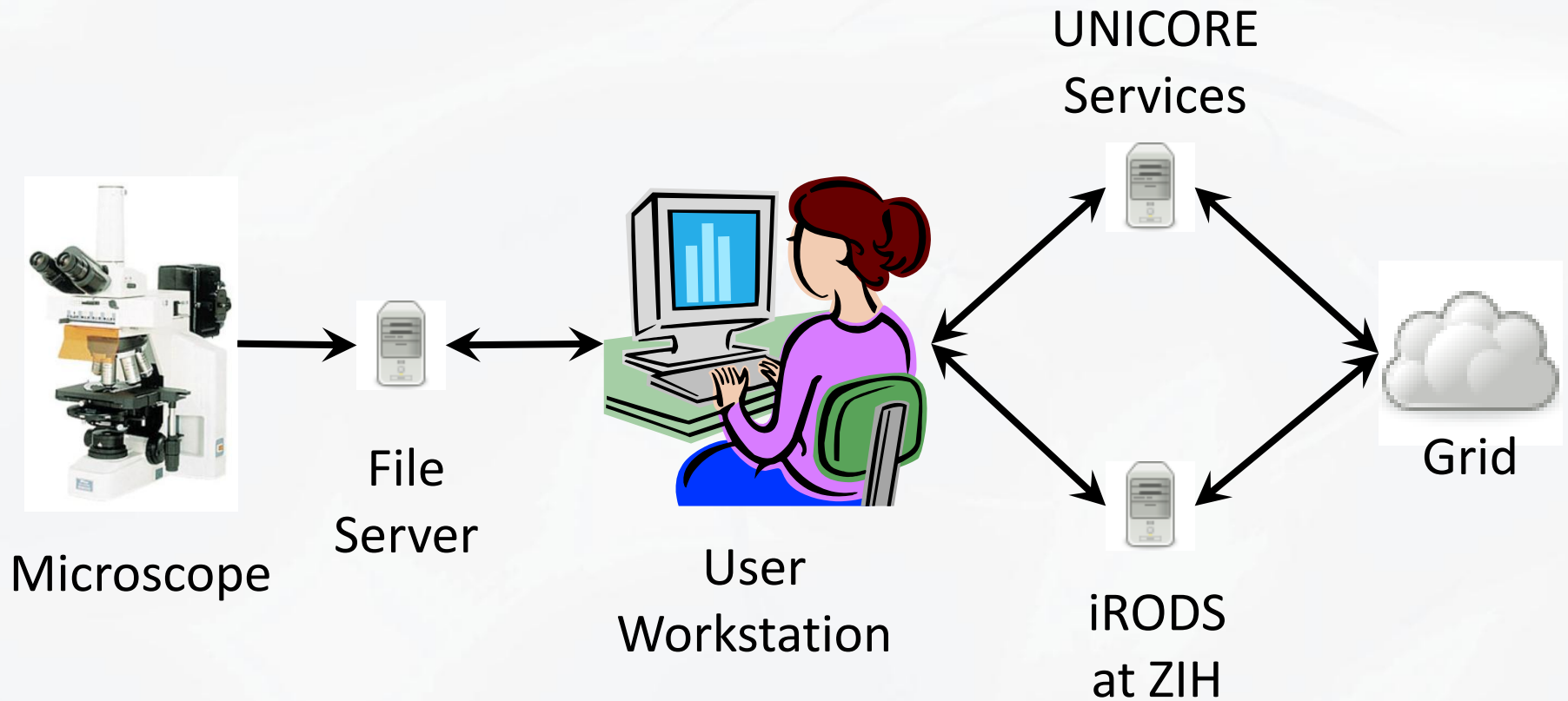
Based on demands

- iRODS for data management

- Efficient data transfers
- Flexibility with rules
- Maturity
- Metadata capabilities



Services Infrastructure



On the Workstation

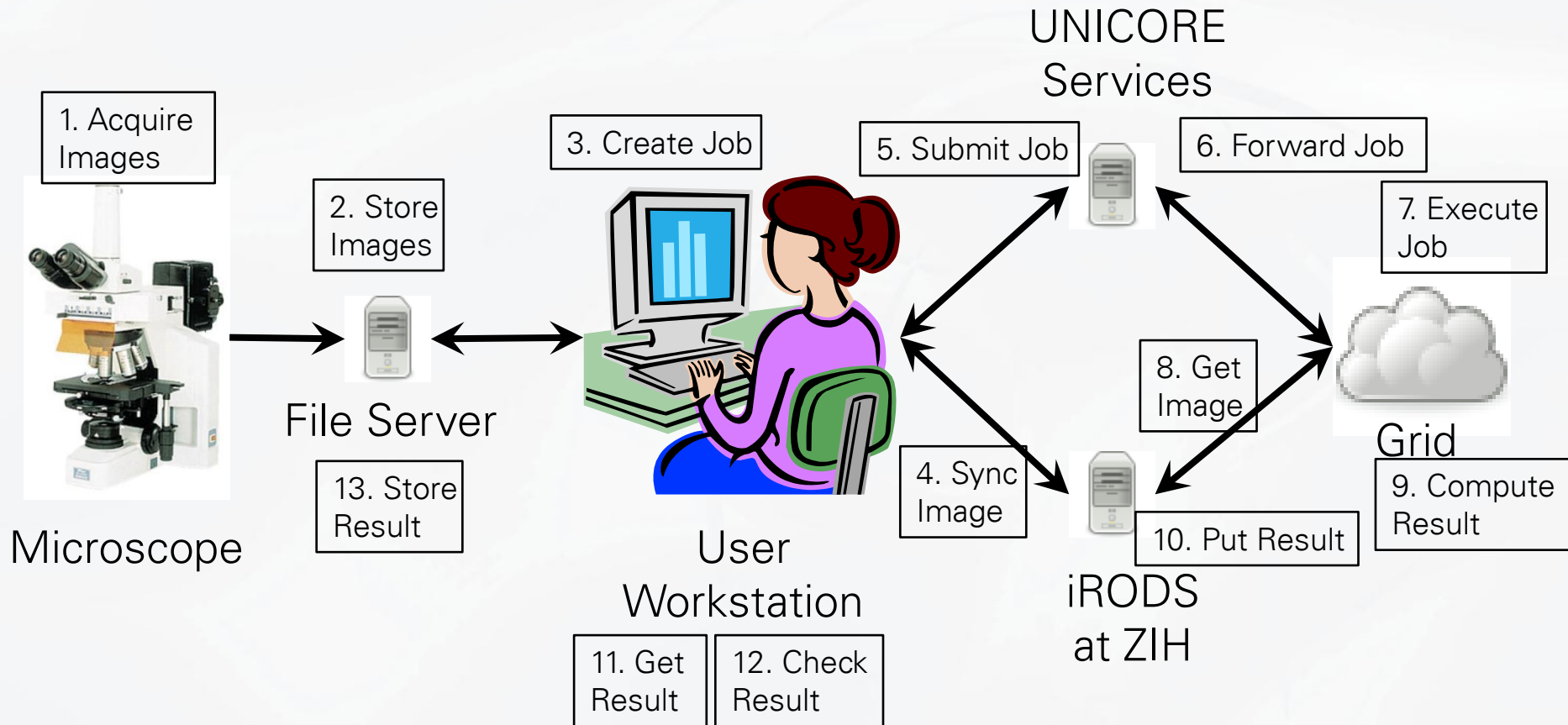
- Windows workstations at MPI-CBG
- MPI-CBG Analysis software Motion Tracking with UNICORE and iRODS support
- Workflow environment
 - Bin directory with scripts, UCC and URC
 - Input, output, jobs, conf and misc directory
 - Windows service for UCC
- Connected to file server which is connected to automatic microscope
- iRODS libraries used by Motion Tracking to access iRODS server

- UCC client on workstations at MPI-CBG
- Registry at ZIH
- Local UNICORE
 - Five UNICORE/X to handle TSIs on login nodes
 - One XUADB on each U/X server

Infrastructure Data Management

- IRODS server at ZIH
 - 16 cores with 2.6 GHZ, 32 GB
 - 10 GB Ethernet
 - 40 TB of storage capacity
- Turned out this is not enough storage capacity
 - Used as a data cache only
- iRODS client (lcommands) on clusters for accessing iRODS server

Logical Workflow



Summary and Outlook

● Summary

- Implemented UNICORE and iRODS based workflow for images analysis in Biology
- Advances cutting-edge biological research
- Integrated in biological environment
- Easy to use and transparent UNICORE workflow

● Outlook

- Integration of Space based approach planned when production version is ready
- Evaluation of UFTP protocol for data transfers