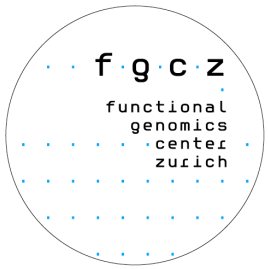
B-Fabric

Data Management in Life Sciences - Analysis and Storage

Marco Schmidt

Functional Genomics Center Zurich, Switzerland

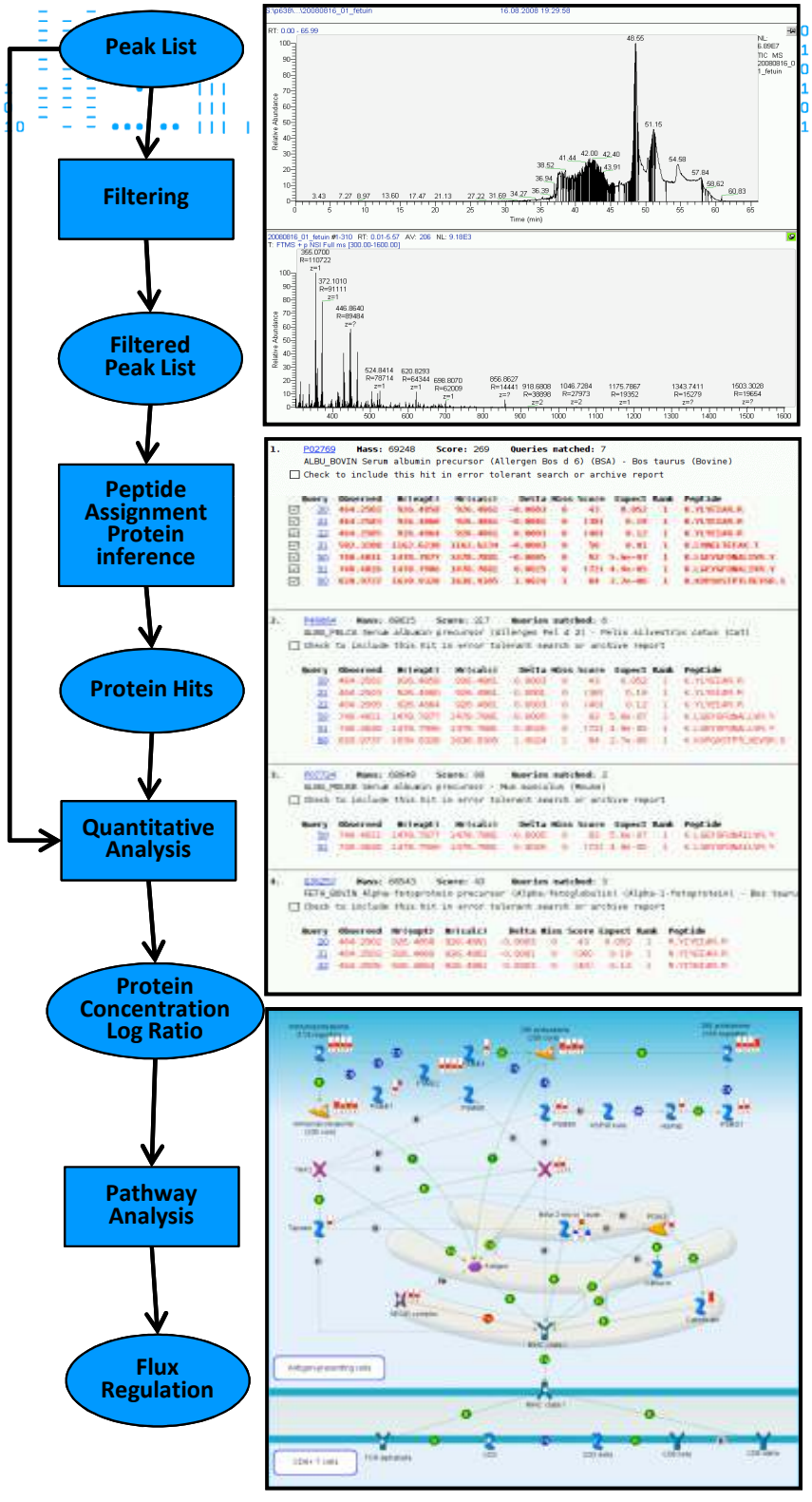


schmidt@fgcz.ethz.ch

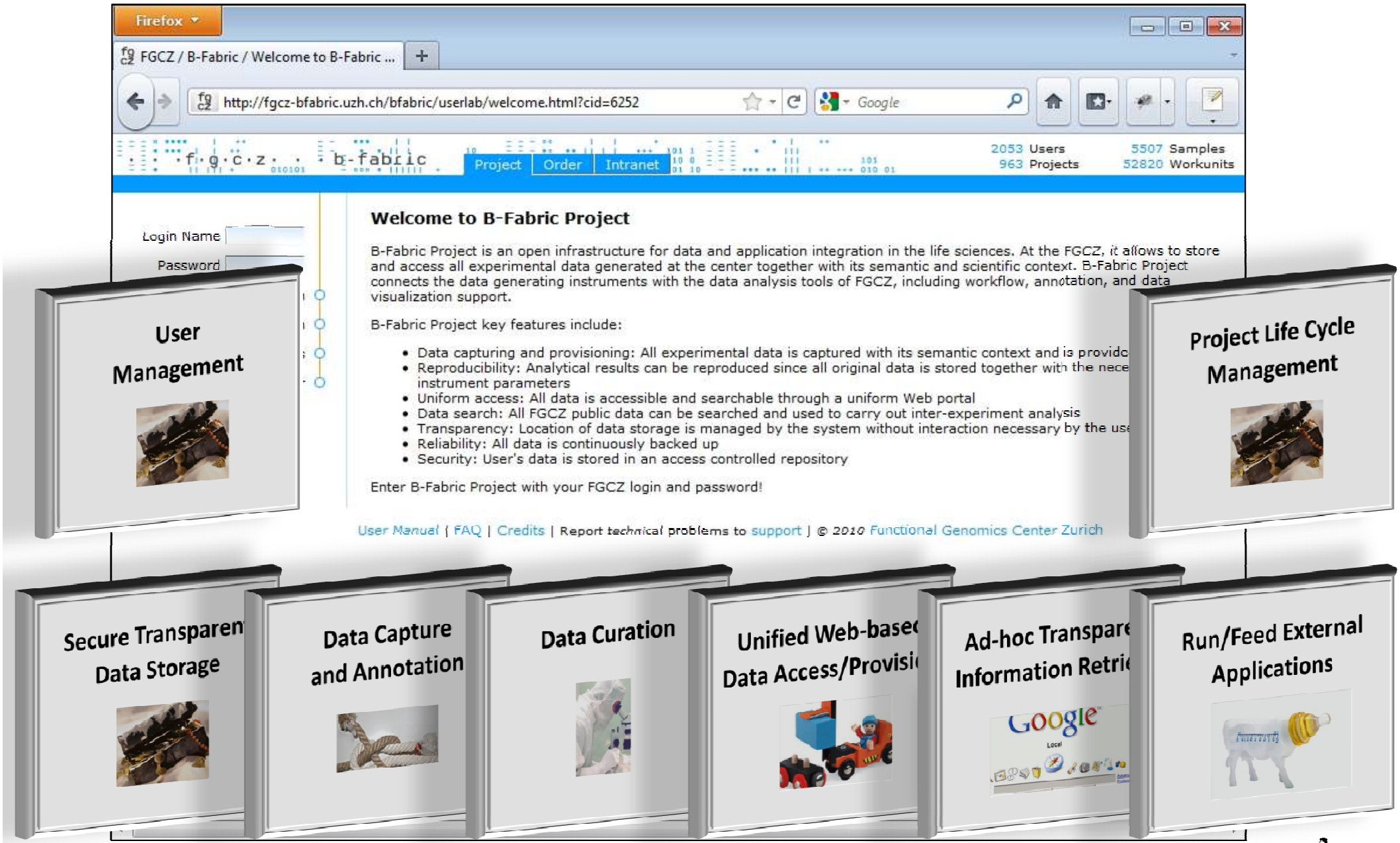
November 23th, 2011

Motivation for Integrative Data Management

- Observation
  - data lies around: huge volumes, often unstructured, inherently distributed, usually file-based
  - heterogeneous systems
  - applications with no or poor interfaces
  - no or weak interaction within instruments/applications
  - processes shredded in scripts & command line tools
  - experts needed to handle the workflows
- Consequences
  - no reuse of research results
  - no reproducibility/tracking of research
  - no semantic search
  - no data quality assurance
- Required
  - Data management system linking together all relevant data and applications



B-Fabric - The FGCZ Approach to Project and Data Management

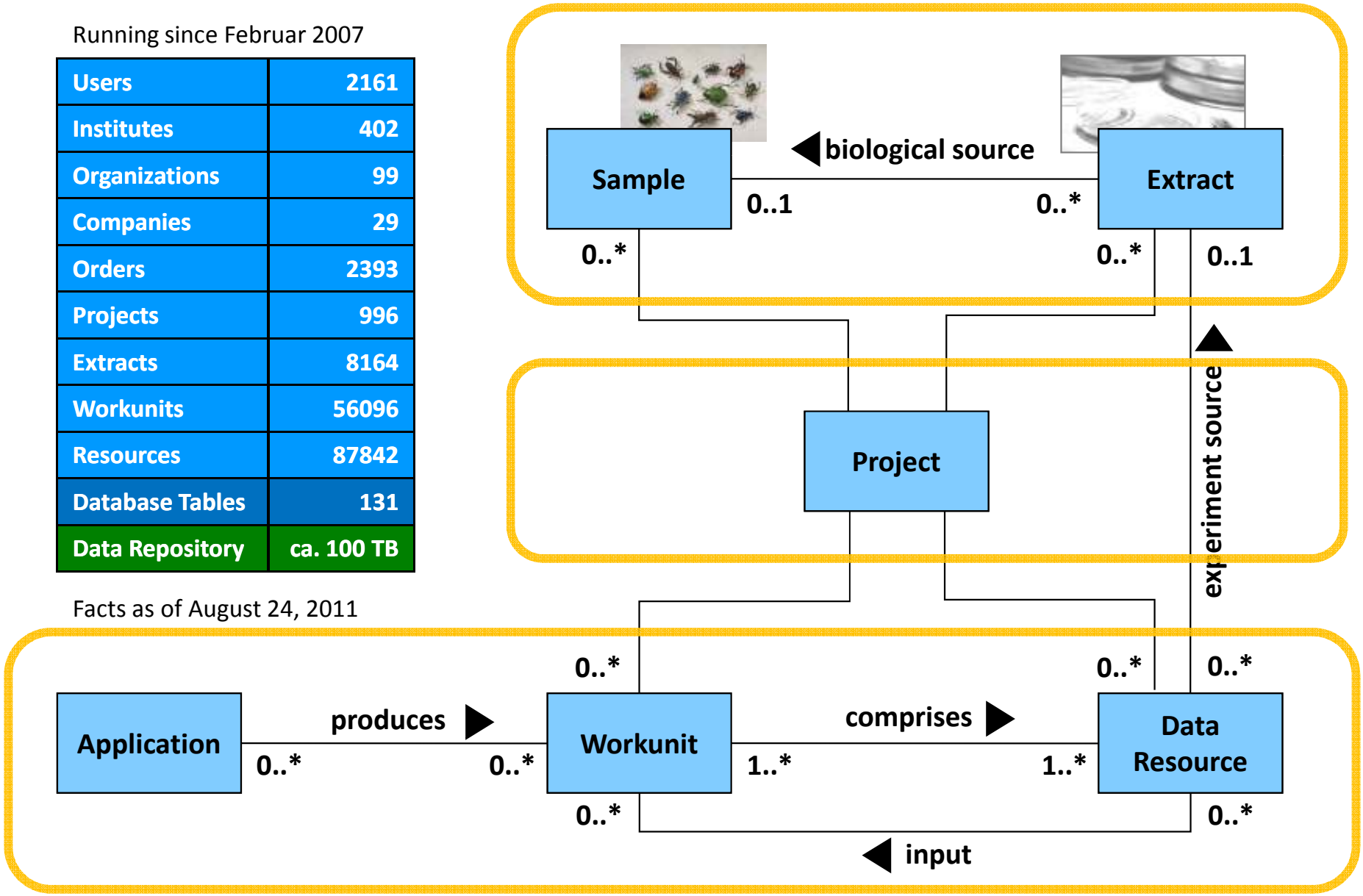


B-Fabric Deployment @ FGCZ

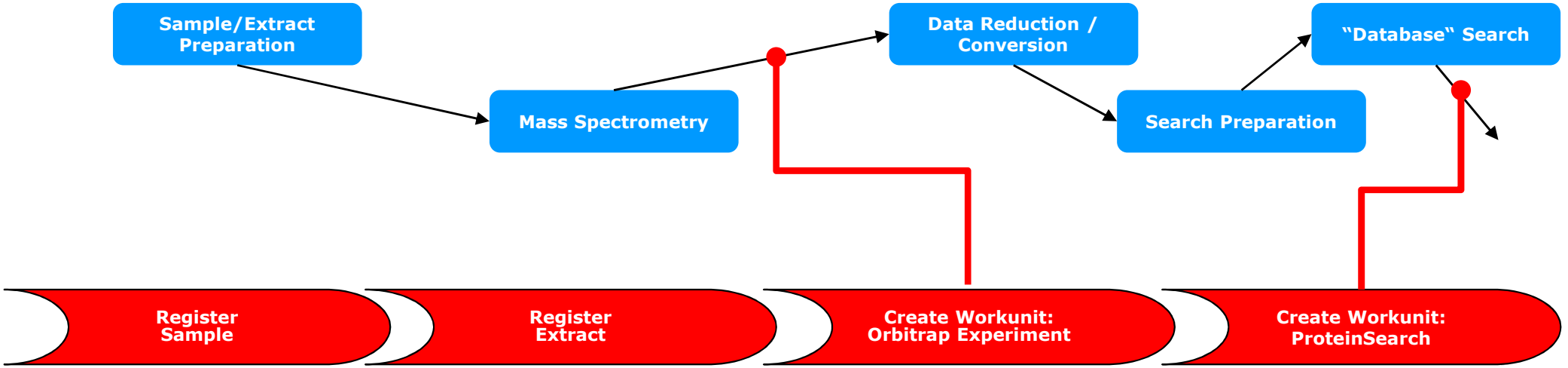
Running since Februar 2007

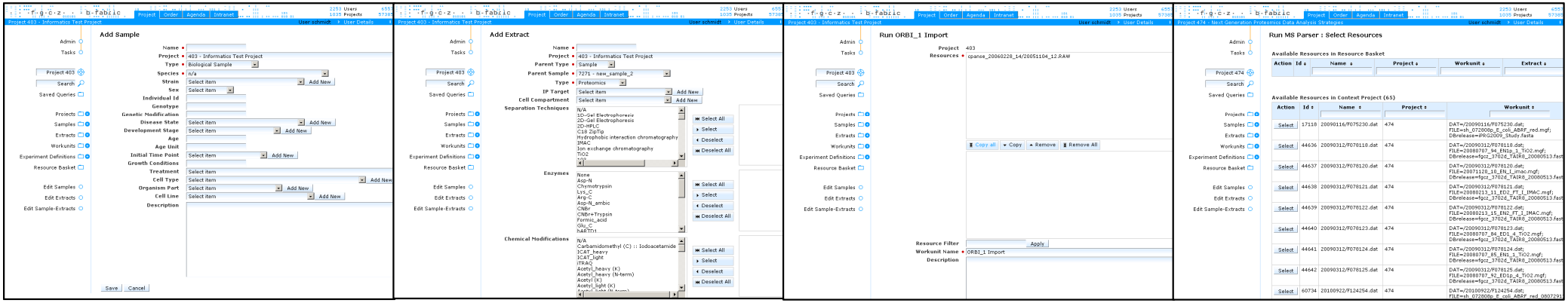
Users	2161
Institutes	402
Organizations	99
Companies	29
Orders	2393
Projects	996
Extracts	8164
Workunits	56096
Resources	87842
Database Tables	131
Data Repository	ca. 100 TB

Facts as of August 24, 2011

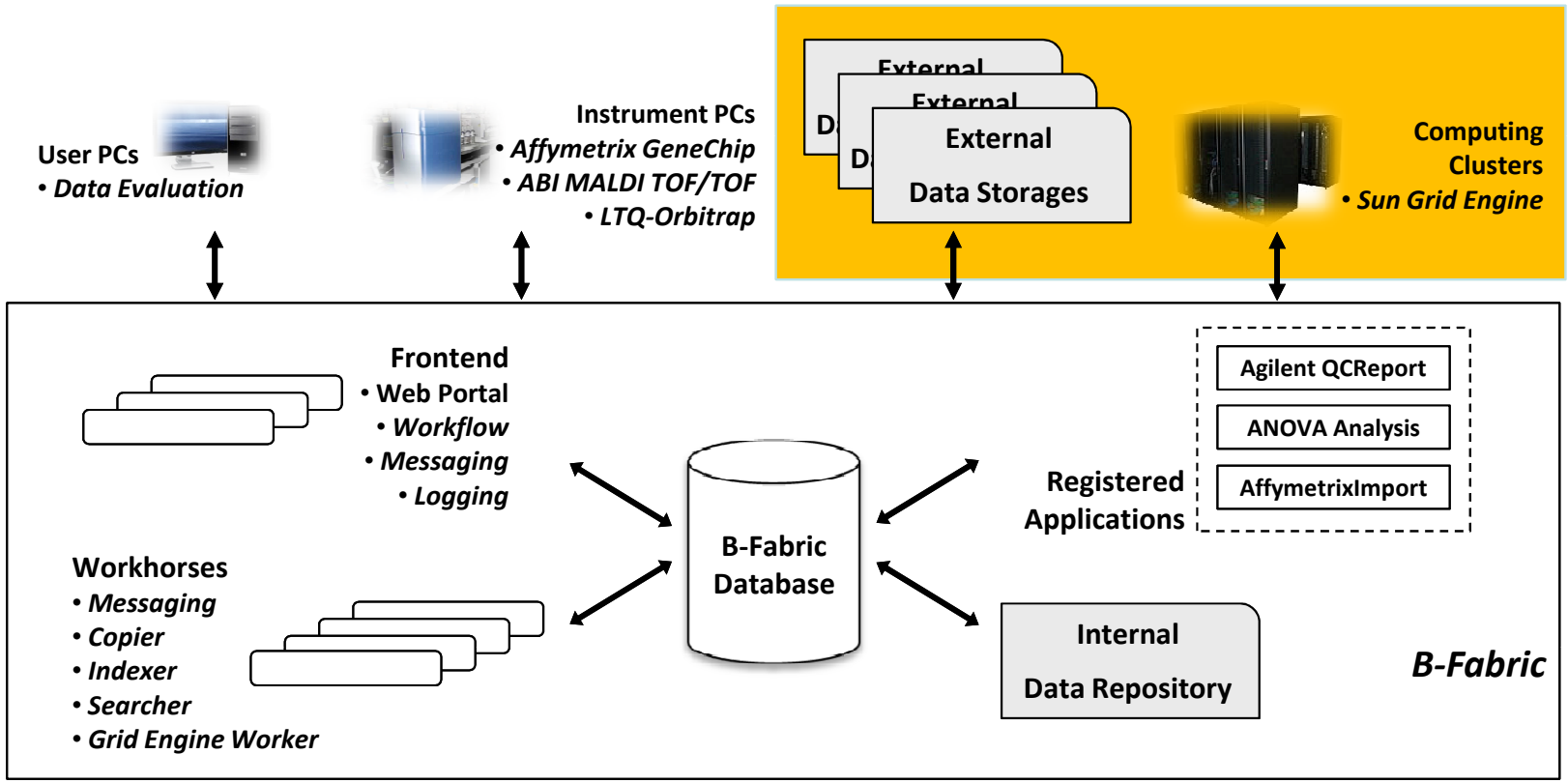


B-Fabric Philosophy: Be generic enough to capture any relevant data





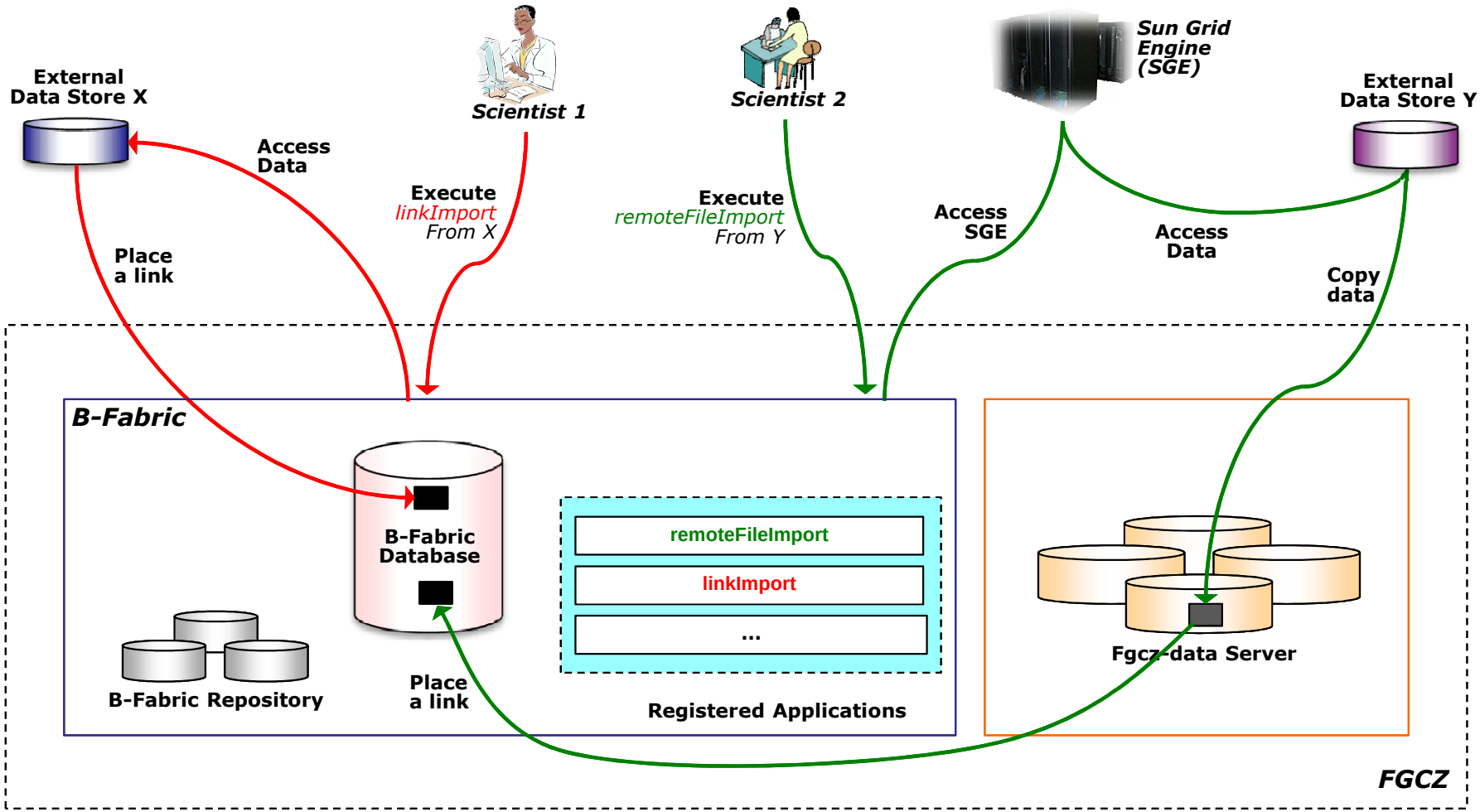
A little deeper look into the B-Fabric Architecture



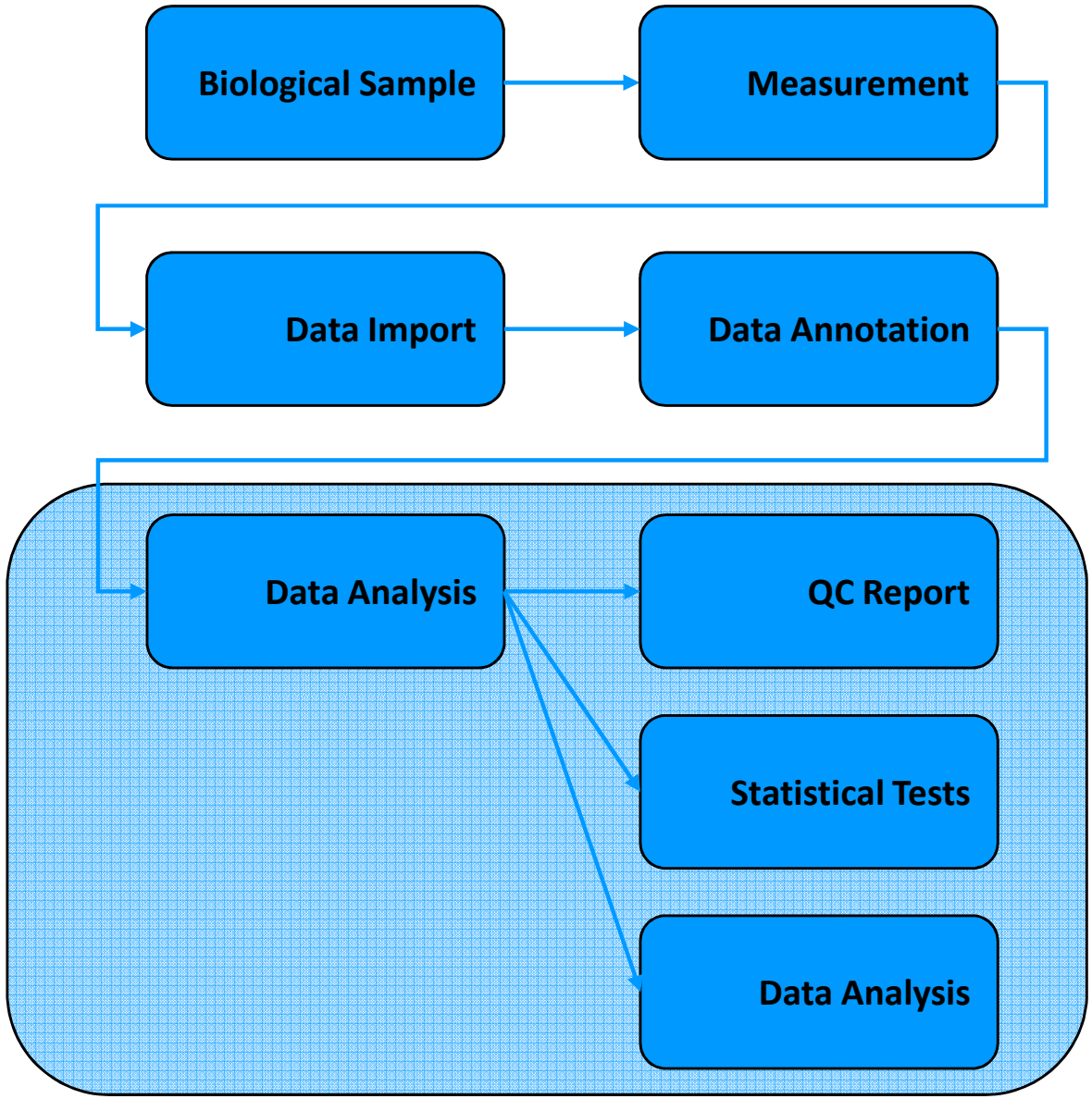
Importing Data

- Generic approach with independant external storages
  - Any protocoll can be handled (scp, http, smb, …)
  - Link import
    - Files are just linked to B-Fabric and stay where they are on the external storage
  - Physical file import
    - Files are physically copied to a target storage
    - Target storage can be any data storage accessible to B-Fabric
    - Links to the files are created in B-Fabric
  - a „storage supervisor“ is responsible for the communication between storage and B-Fabric
- After import, the data has to be annotated !!

Ad-Hoc Coupling of External Data Resources



Data Analysis



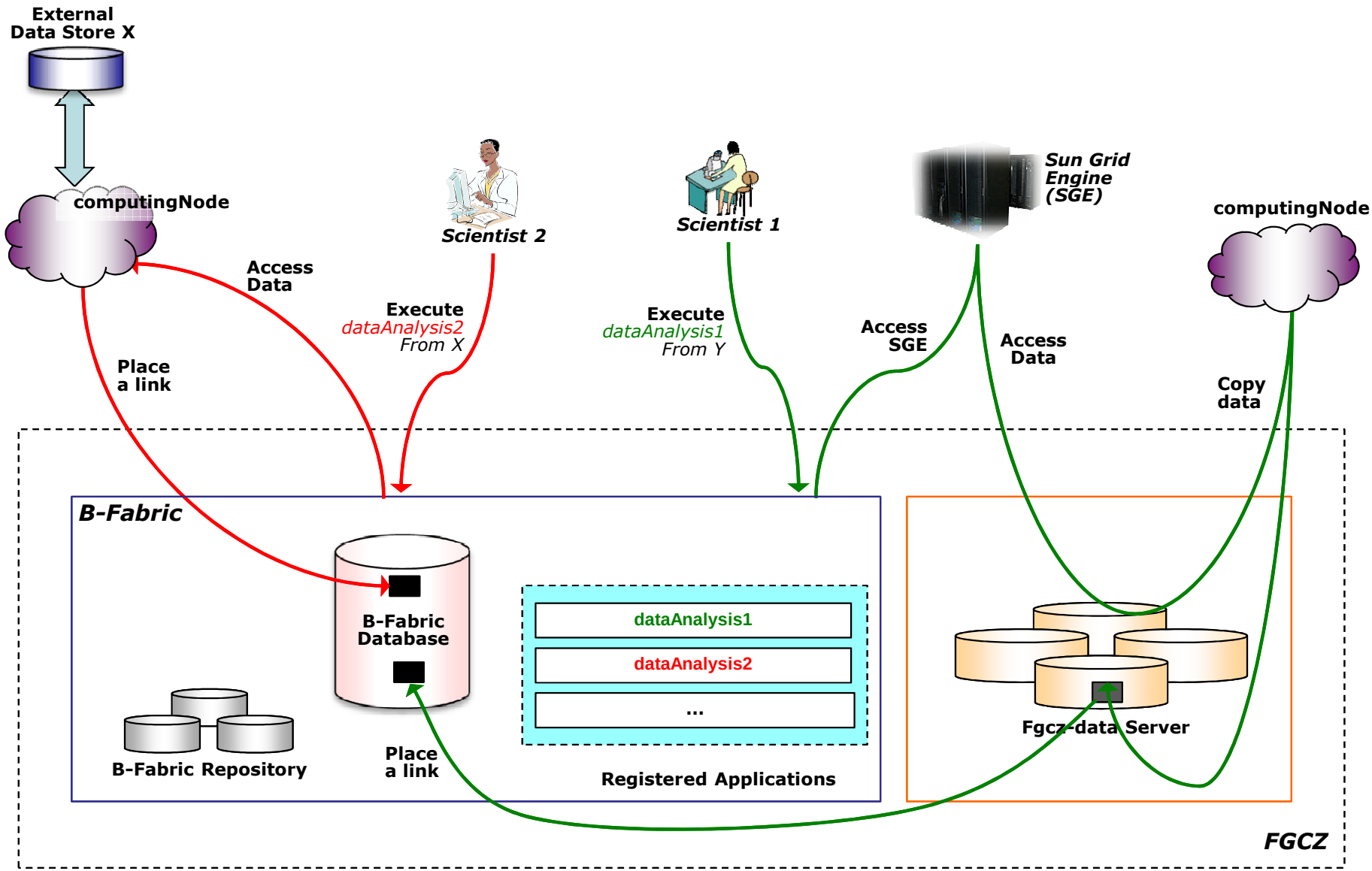
Goals of our B-Fabric based Data Analysis

- **cover 90% of the analysis tasks**
  - implementing pipelines for the remaining cases would be inefficient
- **analysis workflows must be robust**
  - use only well established, widely applicable analyses
- **analyses should be runnable by users**
  - sensible default parameters!
- **results should be trackable**
  - all information used to generate data is know to B-Fabric

B-Fabric Data Analysis Workflows

- Next-Generation Sequencing (NGS)
  - Read processing
  - Read mapping
  - Read quality control
  - Read & coverage visualization
  - RNA-seq: Differentially expressed genes
  - …
- Proteomics
  - Peptide & protein identification
  - Protein quantification
  - Post-translational modifications
  - …
- Microarray
  - Automated quality control
  - Differentially expressed genes
  - Affected GO categories and pathways
  - …

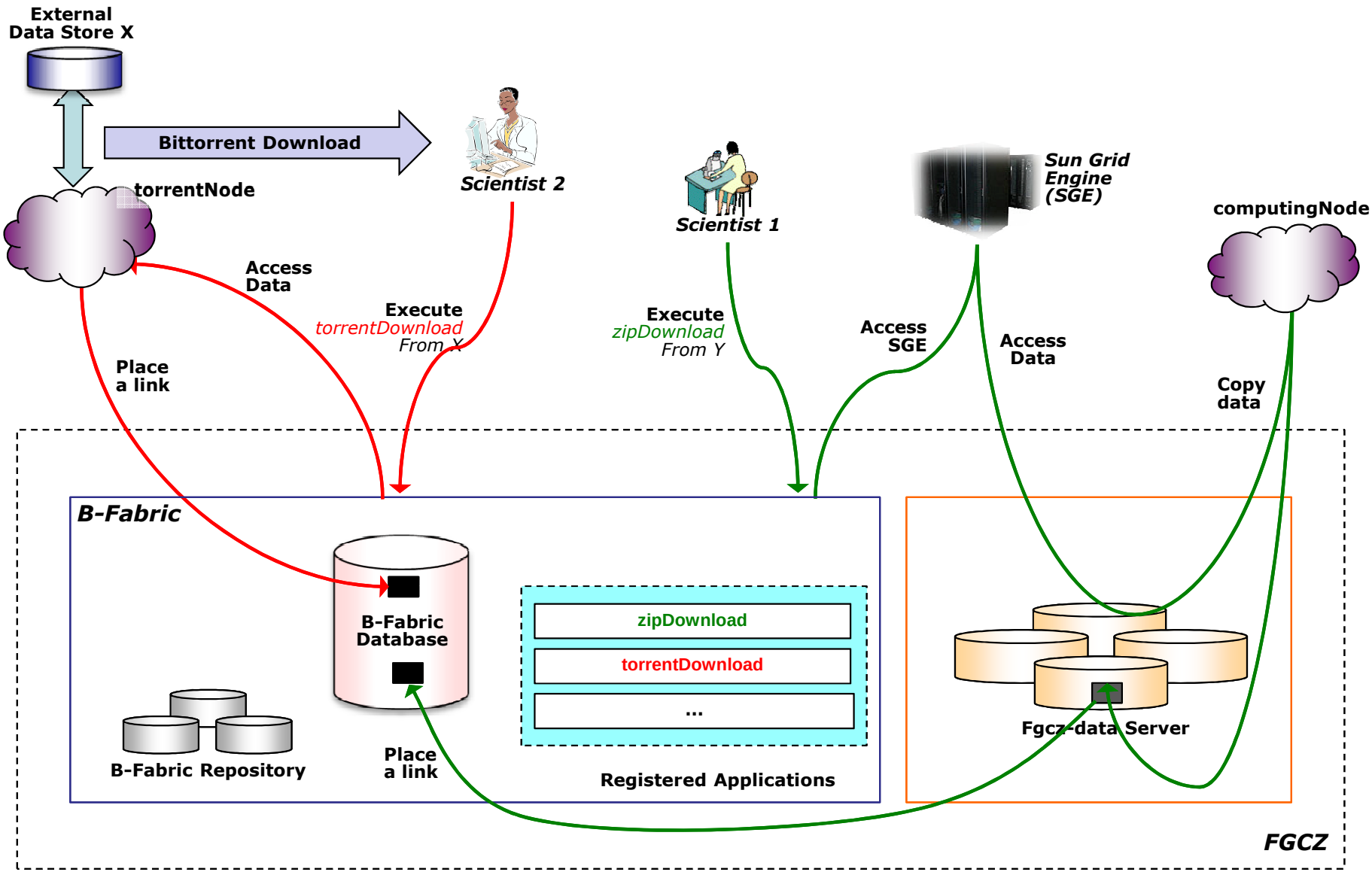
Ad-Hoc Executing of Data Analysis Applications



Exporting Data

- Scientist wants to „Download“ his data
- Could be done with „remoteFileImport“
  - add the storage of the scientist to B-Fabric & add an import application
  - in 99% the scientist wants the data on his „Desktop“
- External Storage has to provide a download mechanism
  - huge amounts of data (NGS: several TB)
  - http download direct from external storage with authentication
  - use applications to generate any kind of download needed !
- Example 1: Zip Download from FGCZ-data Storage
- Example 2: Bittorrent Download from FGCZ- data Storage

Ad-Hoc Executing of Data Analysis Applications



Wrap-Up: B-Fabric Benefits

- Trackability of results
- Easy web-based data access
- Fast access to relevant data
- Data reuse
- Reduced annotation work through automatic export to external marts
- Access-controlled data sharing
- Increased data quality
- Generation of reports etc.
- Reproducibility of research results

- Transparent management of users, projects, orders, …
- Ad-hoc addition of new services
- Task management (user guidance)
- Charging and Invoicing
- Tracking centers resources/capacities
- Central administration tasks automated (user registration/synchronization, door key request, …)

Reduced IT admin, scientists, secretary work

Improved service support/quality

How research centers/groups can benefit from B-Fabric?

- Request and run a project at FGCZ

- Have your own B-Fabric deployment: How?
  - Download B-Fabric, customize and run it!
    - www.bfabric.org
    - Requires a programmer to maintain and customize the system for specific needs
  - Rent an individual B-Fabric instance hosted elsewhere
    - Elsewhere could be «Informatikdienste» or FGCZ
    - Service and price model to be developed

- B-Fabric for Professors
  - To manage their PhD Students
  - PhD Students get their computer accounts with no need to go to the admin
  - PhD Students import and share all their relevant documents and data
  - Research becomes better documented and traceable
  - Not only secondary but also primary research data gets archived

Many thanks to all people having contributed to the development, testing, using, and supporting B-Fabric

Developers

- Fuat Akal
- Christian Decker
- Michael Fetzer
- Felix Knecht (Otego)
- Aleksander Markovic
- Lukas Marti
- Benedikt Thelen
- Can Türker

Alumni Developers

- David Altorfer
- Dieter Joho
- Haissam Mouhasseb
- Giacomo Pati (Otego)

Further Contributors

- Ralph Schlapbach
- Etzard Stolte

FGCZ External Application Developers

- Simon Barkow-Oesterreicher
- Remy Bruggmann
- Christian Panse
- Weihong Qi
- Hubert Rehrauer
- Marco Schmidt

Sponsors

- UZH / ETHZ (financiers of the FGCZ)
- SWITCH: «Generalizing B-Fabric towards an Infrastructure for Collaborative Research in Switzerland» (June 2009-May 2011)
- SYBIT: «Infrastructure for BATTLEX» (June 2010-December 2011)

Universität Zürich[UZH]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SWITCH
Serving Swiss Universities

SyBIT
SystemsX.ch
Biology IT