Karlsruhe Institute of Technology

Large Scale Data Facility

Storage services for Data Intensive Science

Jos van Wezel, Rainer Stotzka

In close collaboration with:
Institute for Data Processing and Electronics
Institute of Toxicology and Genetics
Institute for Applied Computer Science

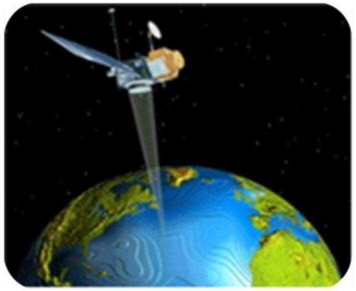
STEINBUCH CENTRE FOR COMPUTING - SCC

Outline

- Large scale data
  - examples and challenges
- Experience at SCC with large scale data
  - Steinbuch Centre for Computing
- The LSDF project
  - current status and implementation
- Data management
  - data placement / replica management
  - meta data handling: the need for a tailored approach

The data challenge

- (Imaging) Science produces tons of data
  - growth is exponential
  - need analysis and storage workflows
  - need integrated compute services
- Data that cannot be found (in a few seconds) is nonexistent data
  - accessibility increases the data value
  - simple access (input and retrieval) increases acceptance by communities
- Care for valuable ‘old’ data
  - needed for reprocessing
  - to track changes over time
  - analysis by others (verification)
  - legal issues



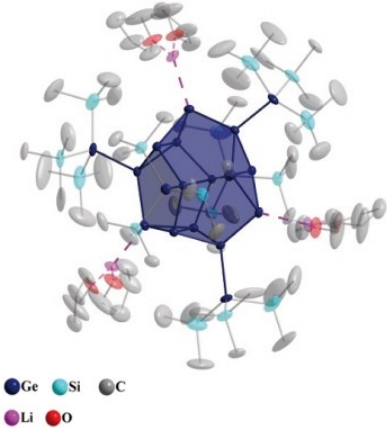
The Large Scale Data Facility Project

- Address the needs of data intensive science
  - offer data intensive computing
  - tools and infrastructure
    - where focus is on the data
    - and a tight integration of data storage and processing is ensured
- Started in 2009 using screens from ITG
- Today involving several KIT institutes
  - SCC, IPE, ITG, IAI, ANKA, ...
  - Cooperation with BioQuant of Univ. Heidelberg
    - State wide (Baden-Wuerttemberg) storage of scientific data

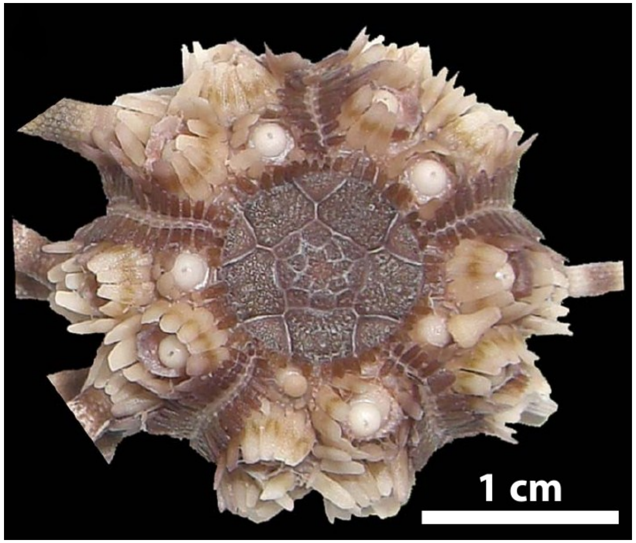
SCC – Steinbuch Centre for Computing

- What qualifies SCC to operate the LSDF?
  - Central IT services for KIT
    - networking, data protection and archiving
  - Several compute clusters each with several 1000 cores
    - HPC clusters
  - Node in BW grid project
    - Storage
    - Compute nodes
  - World wide Help desk
  - On-Call operators
    - 24x7 operation
  - Operate the GridKa T1
    - Storage, computing and networking for the LHC

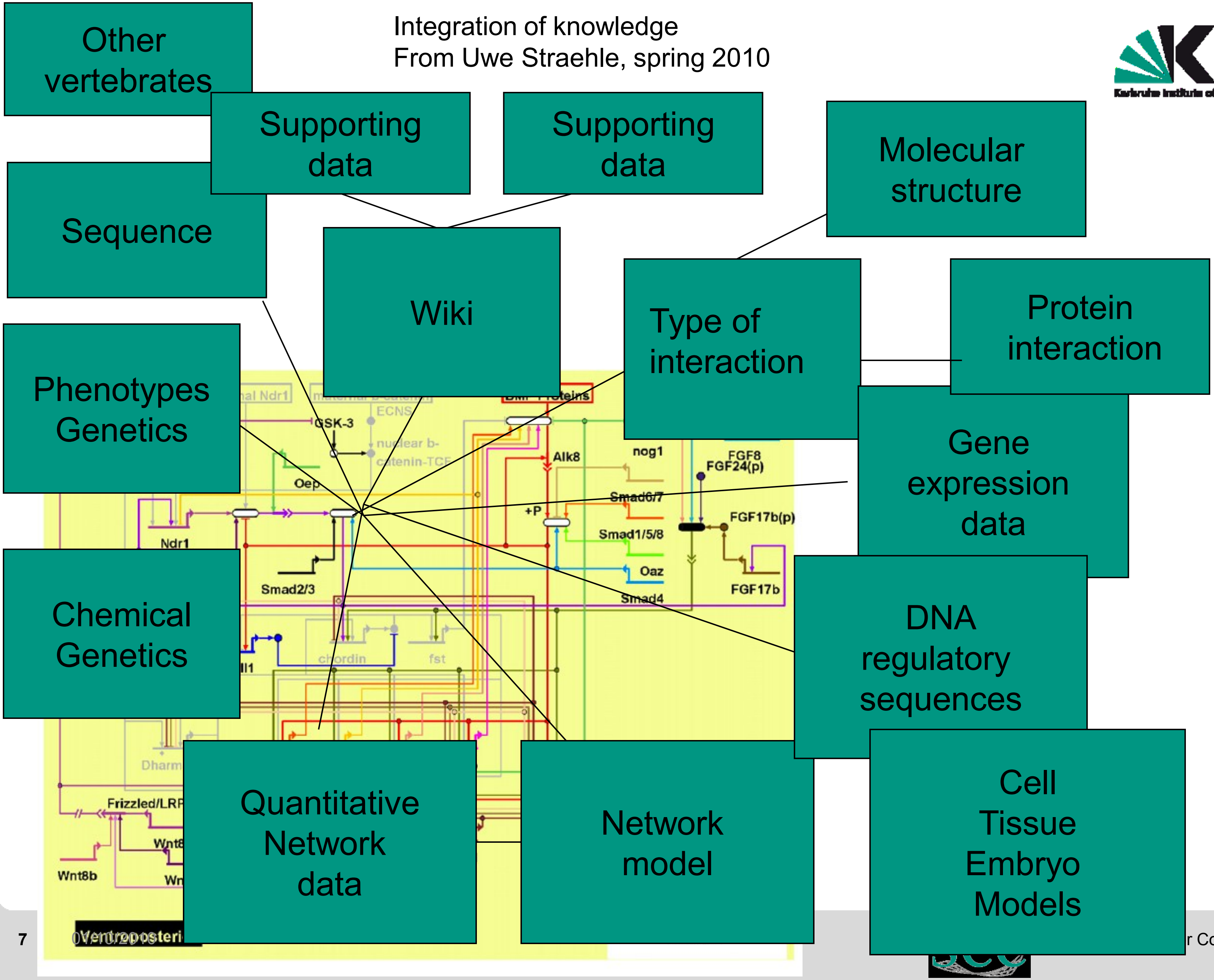
Existing Large Scale Data



- LHC Large Hadron Collider
  - Produces 10 PB per year in 2010
    - The GridKa Tier 1 at SCC stores 1 PB/year
    - Currently 5 PB data stored
- ANKA - ÅNgströmquelle KArlsruhe / ISS
  - Tomography and other beam line experiments
  - 60 TB raw data + 3 times processed data = 240 TB/year - 1 PB/year (2013)
- Immunogenetics Institute Charité Berlin
  - Computer tomography of sea urchins
  - several hundreds of TB
- ITG - KIT
  - High Throughput Microscopy
  - estimated 1-2 PB/year
- BioQuant - Univ. of Heidelberg
  - High Throughput Microscopy
  - Genome sequencing, Electron-microscopy
  - estimated 1-2 PB/year



Integration of knowledge
From Uwe Straehle, spring 2010



Other vertebrates

Supporting data

Supporting data

Molecular structure

Sequence

Wiki

Type of interaction

Protein interaction

Phenotypes Genetics

Gene expression data

Chemical Genetics

DNA regulatory sequences

Quantitative Network data

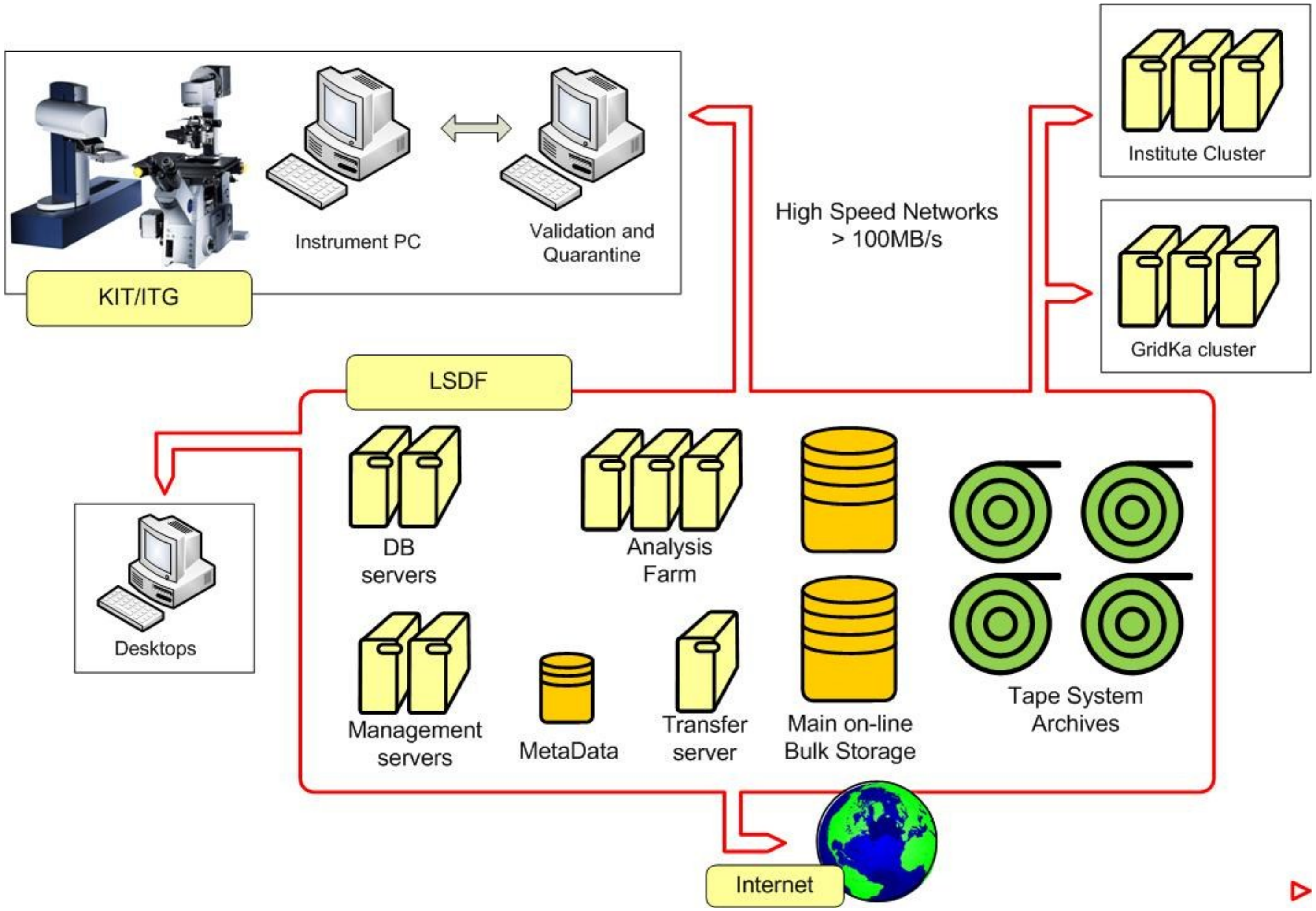
Network model

Cell Tissue Embryo Models

Integrative tools: prerequisites

- Large scale storage facilities
  - handling of large data sets
- Bioinformatics tools to retrieve, analyse and model data
  - interactive environments
  - graphical Wiki
  - mathematical tools
  - virtual reality tools for animation of networks,
  - 3D and 4D models
- Search engines
- Everything we did not think of before

LSDF birds view



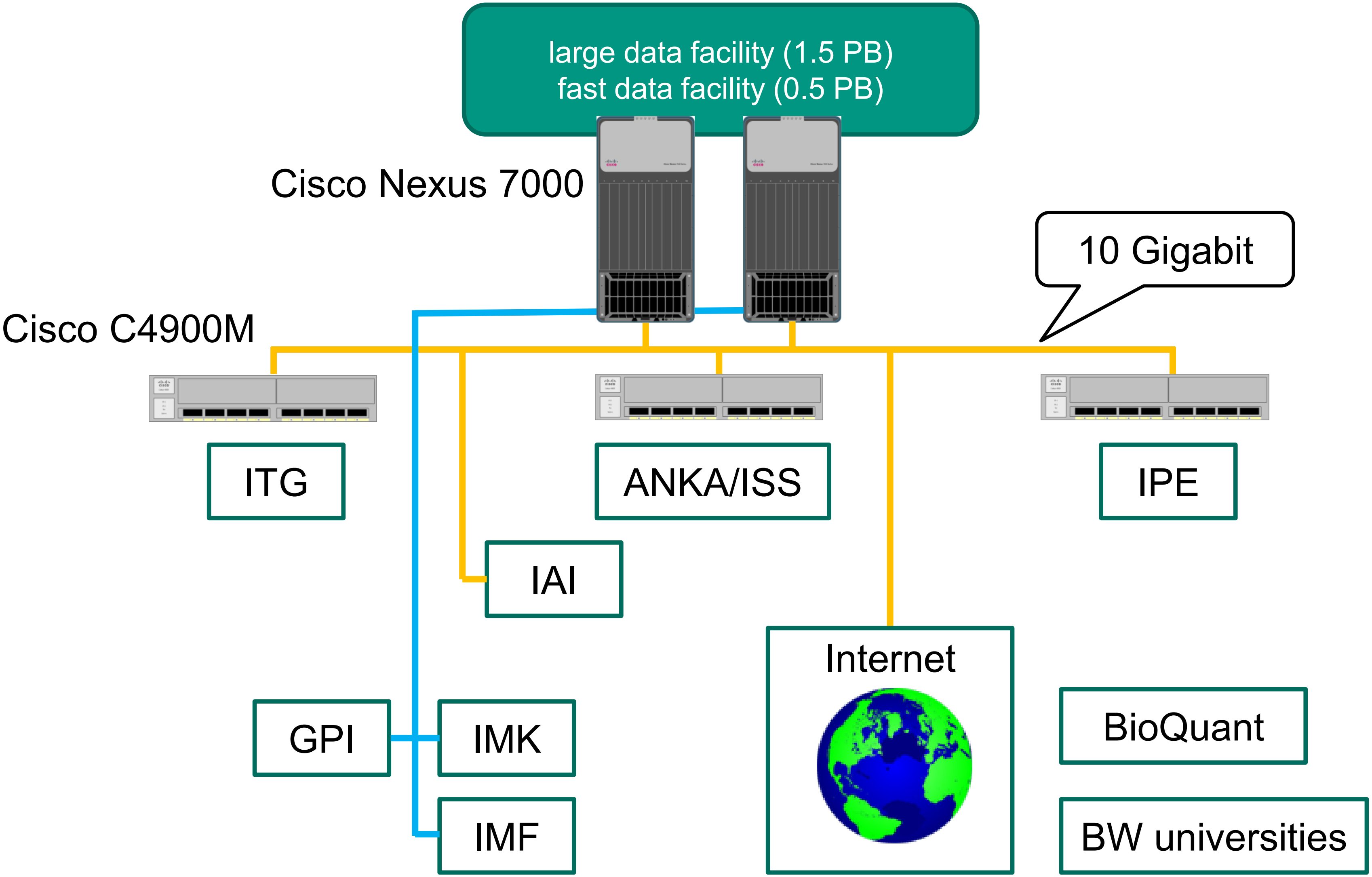
Computing

- Experiments should be able to process data locally
  - i.e. where the data is located
  - 15 days to transfer 1 PB over ideal 10Gb/s link
  - dedicated cluster
- 58 nodes with 8 cores, 36GB memory
  - directly attached to storage (GPFS)
- Hadoop environment
  - 110 TB HDFS, Hadoop native filesystem
  - required for Hadoop workflows
  - extreme scalability on commodity hardware
  - available from the Cloud environment OpenNebula
  - users can deploy own dedicated data-processing VMs
  - reliable, highly flexible, and very fast to deploy

Storage

- 2 high grade disk systems
  - 550 TB [Data Direct Networks]
  - 1,2 PB – 3 PB (2011) [IBM]
  - Fibre Channel attached
- Dedicated storage servers
- Tape backend for archive and backup
- GPFS on top of each storage system
  - exported as GPFS, NFS, CIFS (Windows native)
- Directly attached to processing cluster
- **Networking**
  - 10 Gb/s dedicated redundant backbone
  - 10 Gb/s dedicated links to some partners, 1Gb/s others

LSDF Network



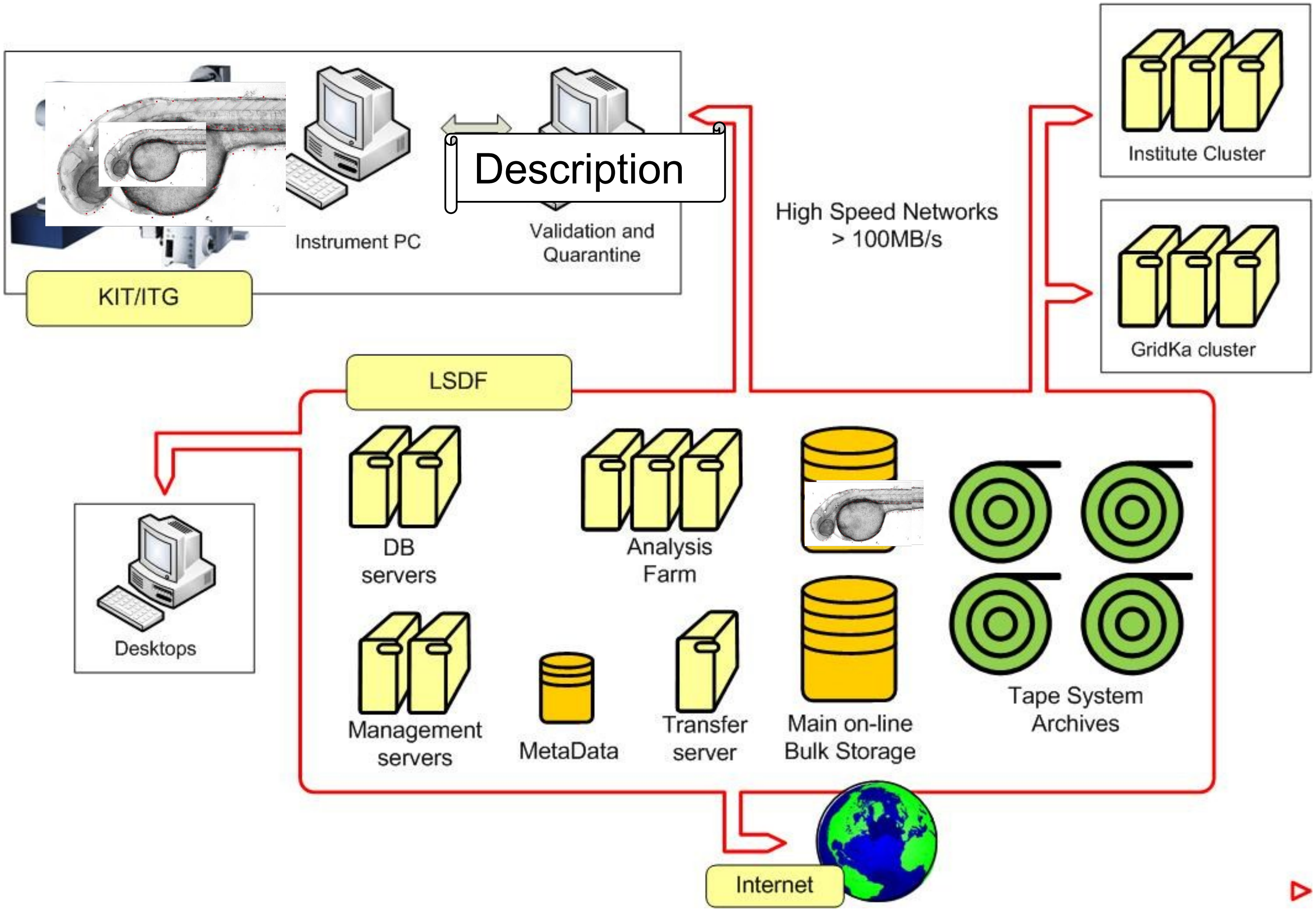
LSDF: Provided services

- Large scale storage and world wide secure access to data
  - assure transparent access over diverse storage technologies
  - mask technology changes
  - integration in world wide accepted authentication methods
- Added value and tools to process data
  - Archival of data
  - Name space federation
  - Meta-data management and tools (see next talk of Rainer Stotzka)
- Development and deployment of community specific services
  - honour community specific techniques and systems
  - integrate existing methods and tools

Workflows – Processing pipeline

- Experimental data acquisition
- Raw data copied to LSDF
  - data-placement tools enforce rules
  - data is physically moved to most appropriate storage
  - initial metadata, tagging is provided
  - preprocessing workflows
- Scientists accessing data
  - search/locate
  - processing workflows on LSDF facilities
  - download and process data externally
  - define access rights for collaborators, or public access
- Data is archived
  - post validation
  - legal binding

Simplified Workflow



Handling large scale data

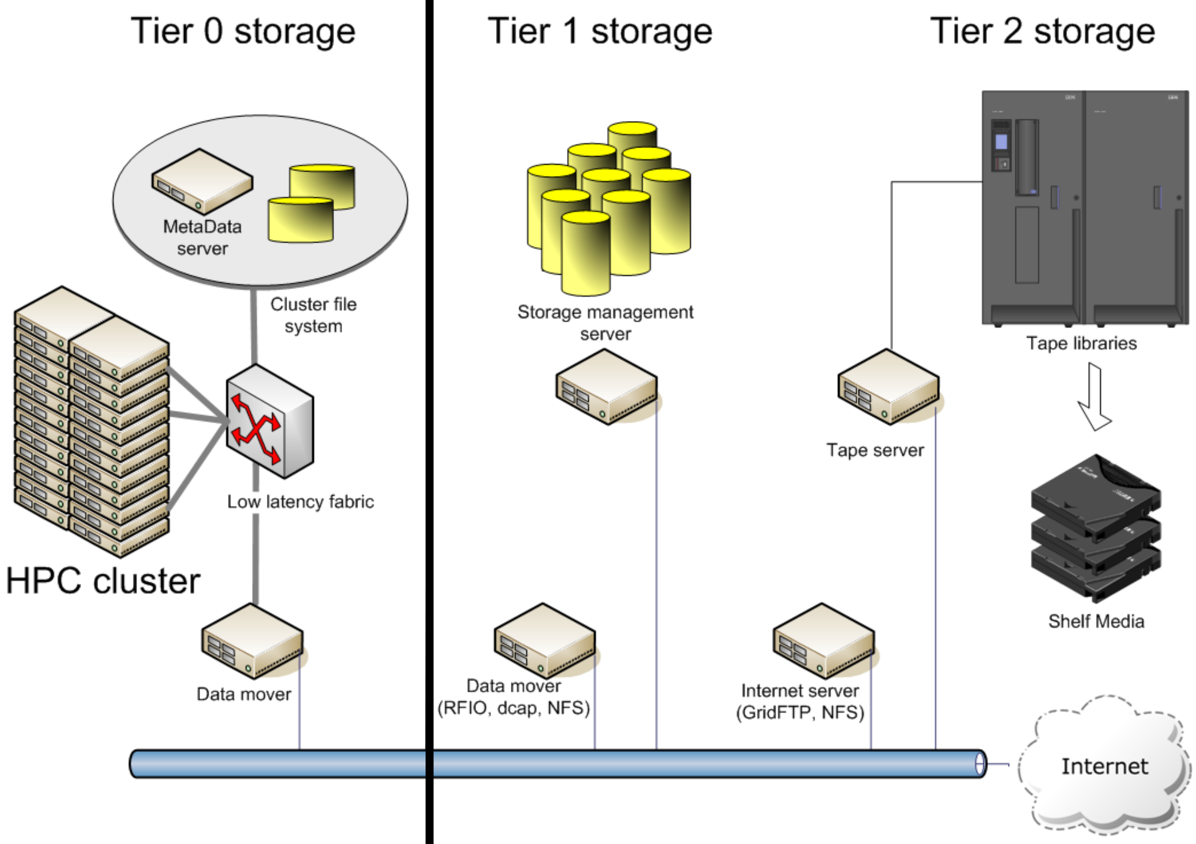
- Equipment and staffing
  - 24 x 7
- Finding the data: Metadata
  - Payload data and Meta data follow different paths
  - Needs to be stored and kept up to date with data
  - Metadata schema is highly project-dependent
  - Presupposes the use of a project metadata DB
  - More on meta-data in following talk of Rainer Stotzka
- Data placement and workflows
  - tiered data storage

Tiered storage

- Data is moved to and from tiers that differ in storage function and quality

- Tdaq : Quarantine
  - Where data is produced
- T0 : Process
  - High bandwidth and low latency to CPUs
  - Scratch/Volatile
  - local disks of workers nodes
  - shared file system of clusters
- T1: Archive
  - Longer term online storage (disk)
  - Source and destination of LAN transfers
  - Serving data to some (read mostly) applications
- T2
  - Long term archive storage (tape)
  - Low latency scheduled access

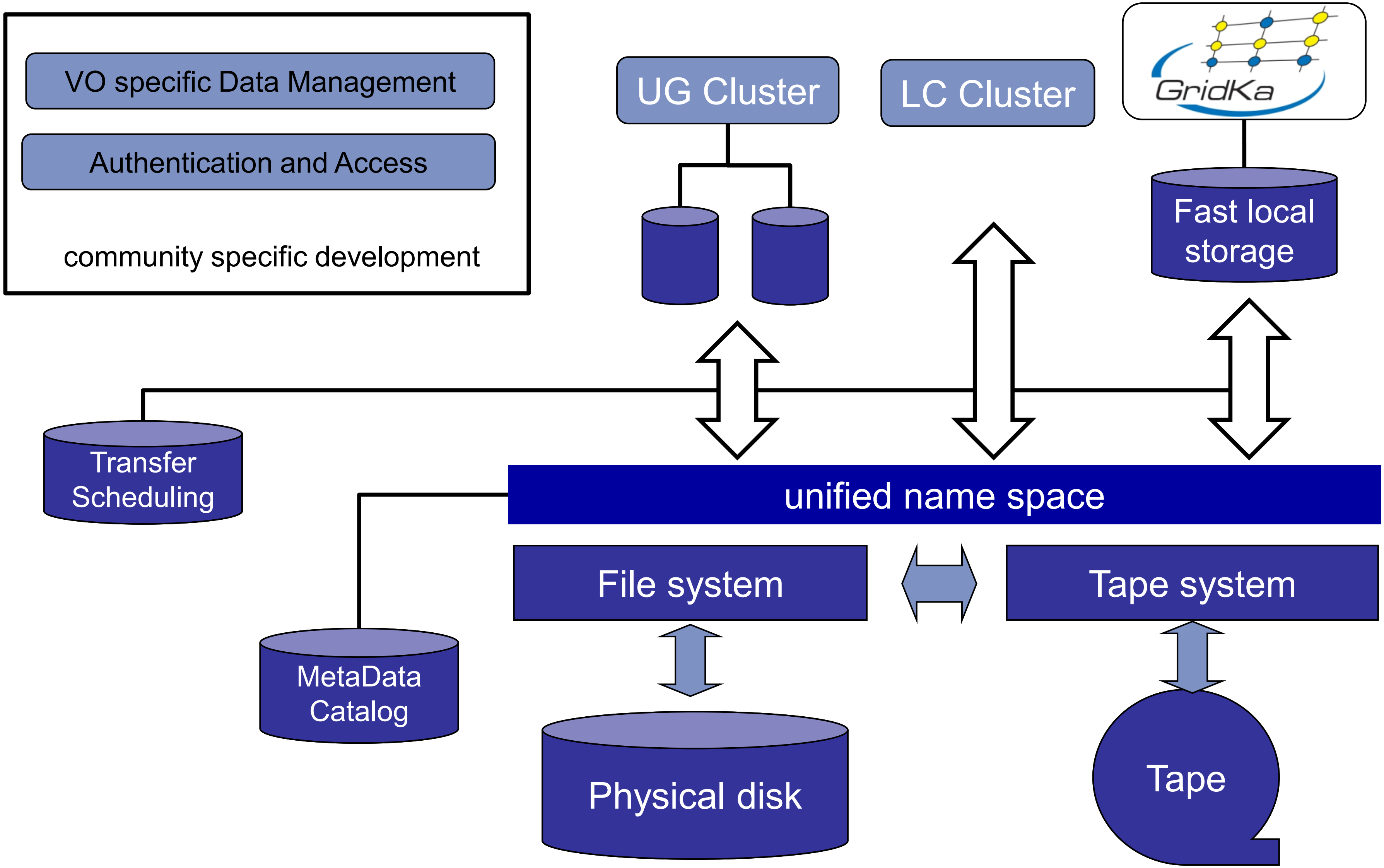
Storage Tiers



ITG/IMK/ANKA

SCC

Data placement



VO specific Data Management

Authentication and Access

community specific development

Software

- Data placement via iRods
  - Popular in science community
  - IBM is in IRODs collaboration.
  - interface to users with X509 authentication.
  - Not to be used as disk manager or transfer protocol, Not as data catalog, provenance, bookkeeping tools
  - Just a replica catalog, data tier migration manager
- LSDF will enable global access via DataBrowser tool
- Open protocols like FTP , HTTP,

Roadmap (excerpt)

- Hardware
  - Growing storage capacity
    - 2 PB in Q4 2010
    - 4 PB in 2011, switch to SONAS on IBM storage
    - 6 PB in 2012
  - Dedicated tape storage by Q4 2010
  - Improved network connectivity
    - Dedicated 10 Gb/s backbone for remaining KIT institutes by Q1 2011
  - 40 Gb/s Link to BioQuant/Heidelberg by Q2 2011
    - Initial support for IPv6 in 2011
  - Provide tape archiving for BioQuant
- Service
  - Enable direct access storage for first experiments, Q3 2010
  - iRods software operational, Q4 2010
  - Additional communities integrated in 2011
    - ANKA (synchroton radiation ring)
    - IMK (meteorology and climate research)

To conclude

- First hardware up and running
- First software services available
- First data stored
- First experimental data processed
- Focus on user requirements
  - added value services on top of large storage
- Many scientific communities interested and getting involved
- Actively pursued
  - grow beyond KIT, HGF and build international collaborations
  - toward an exabyte storage system
  - Scaling to terabit networks and exabyte storage must start today
  - involve new experiments
  - explore new techniques, integrate/develop new services

Thanks for listening

LSDF

The team behind LSDF at SCC:
Serguei Bourov
Ariel Garcia
Bruno Hoeft
Rainer Kupsch
Bernhard Verstege

Points to seed the discussion

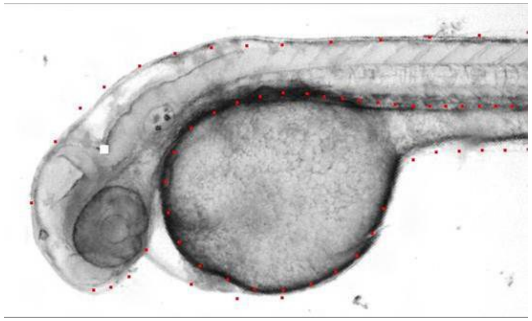
- data granularity
  - large data means large files.
  - data organisation is larger compartments is a must
- If tools do not exist or are not exactly what you like
  - software development is a time consuming process
  - software maintenance is a costly process
- data exchange
  - who wants data when?

Spare: Workpackages (proposal)

- Management (Programme Head, SCC LSDF coordinatior)
- Provisioning infrastructure, administration, maintenance
- Support
- Integration of Data Xfer technologies into Storage Middleware
- High throughput link to computing infrastructure
- Data and meta data organization
- Abstract Data Access Level API
- Tools and software components for applications
- Application/DAQ integration
- Data analysis workflows

Spare 1

- @ Institute of toxicology and genetics
  - fully automated microscopes
  - robot moves object to microscope
  - can potentially run 24*7
  - produce high resolution images (4 MB each)
  - over varying parameters (focus point, wavelength, ...)
  - 200k images per day, 2 TB/day
  - Estimated: 1 + PB/year in 2011, 6 PB/year in 2014
  - Raw data must be heavily analysed



Spare 2 Developments within LSDF

- Provisioning of storage and archives in exabyte scale
- Development of software technologies for distributed data management and archiving
- Development of efficient transport protocols from the experimental facilities, e.g. robotic microscopes, to the LSDF
- Development of technologies to handle the special requirements of experiment data (e.g. 3D image stacks)of various research communities (e.g. systems biology)
- Development of open standards and implementation across computing centre borders
- Provisioning of compute resources for data analysis
- Development and integration of data analysis services
- Specific support for users with data intensive applications
- Development of data and meta data models for specific user groups
- Optimized data organization for specific user groups

Spare 3 Tiered storage

- Keep storage areas independent