Data Storage and Processing for High Throughput RNAi Screening

Julian M. Kunkel, Thomas Ludwig (Inst. Informatik, Uni. HD), Marc Hemberger (BioQuant),
Gloria Torralba, Eberhard Schmitt, Michael Hausmann, Volker Lindenstruth (KIP, Uni. HD),
Nigel P. Brown, Reinhard Schneider (Uni. HD / EMBL)

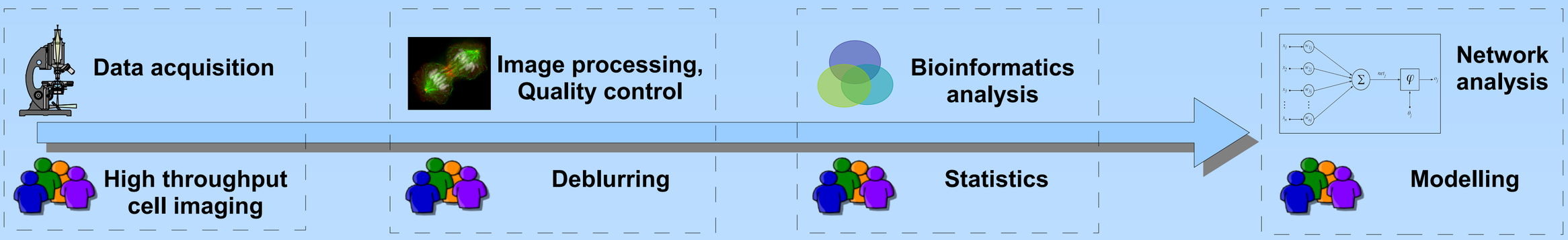
Abstract

The *ViroQuant* project on Systems Biology of Virus-Cell Interactions will develop models for cellular networks responsible for virus replication and spread, using high-content RNAi screening, high-resolution microscopy and image analysis of very large data sets.

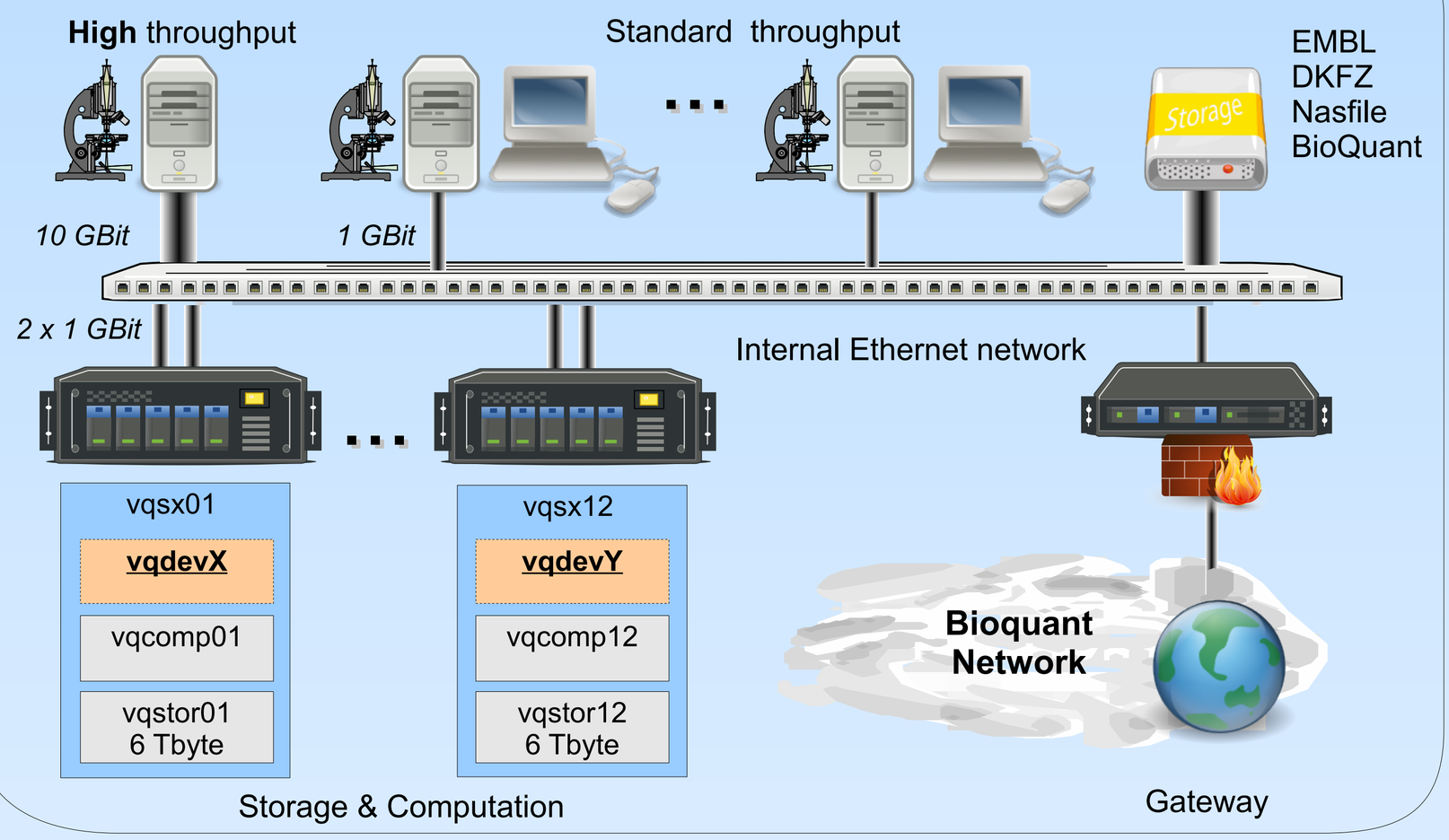
The data storage and processing infrastructure is designed to provide a fast parallel and distributed processing environment supporting data capture, data access, analysis, and collaboration of disparate research groups within the project. The design requires massive storage capability ensuring reliability and durability of the data combined with high throughput and processing.

The system has to be able to capture image data from the microscopes at rates of 1GB/s sustained. A typical human genome screen will collect of the order of 100,000 images (about 250G) every few days. We anticipate a total data production approaching 200TB over the project lifetime.

Idealized workflow



Simplified hardware architecture



Hardware specifications

In BioQuant a dedicated network infrastructure is provided for image acquisition from microscopes, data storage and later processing on a cluster environment. Initially, 5 standard microscopes are attached and configured to store their data on the cluster. A special high throughput microscope developed within ViroQuant requires a high constant rate of data to be distributed across the storage.

Gateways provide access to the data storage and computation resources. The cluster consists of homogeneous multicore nodes each handling computation as well as storage.

A single node is interconnected with the network by two Gigabit Ethernet network adapters bundled into a logical device allowing a data transfer rate of 200 Mbyte/s.

An Areca ARC-1220 8-Port SATA RAID controller equipped with eight 1 Tbyte disks manages 6 Tbyte of data in a redundant RAID-6 configuration. This setup delivers a sustained I/O-bandwidth of 200 Mbyte/s for writes and 300 Mbyte/s for read access. Therefore, the available network bandwidth matches the available I/O-bandwidth.

With increasing demands the scalable architecture will allow us to increase the number of nodes. A dozen of these nodes are provided in the initial setup.

Later in the project there may be direct connections to other collaborating institutions such as EMBL or DKFZ.

System environment

The storage and computation nodes run under Linux (Ubuntu 8.04, Hardy Heron). A XEN virtualisation environment running on all nodes provides shared resources to virtual machines (domains) associated with tasks. Domains can be instantiated on demand.

Computation domains (**vqcompN**) offer distributed image processing and analysis and support for software such as CellProfiler, R or MATLAB and languages like C++, perl and python. Additional nodes equipped with graphic cards for high performance computing is also envisaged. These nodes will use the C development environment provided by NVIDIA.

Storage domains (**vqstorN**) allow direct access to each machine's RAID-controller, the storage space of all machines is bundled into a coherent namespace. A software layer will mirror important data among pairs of storage domains to ensure durability.

One of the gateways hosts a special domain acting as a management host to monitor the state of the cluster and storage.

Additional development domains (**vqdevN**) provide a development environment for new image processing and bioinformatics tools. Matlab with necessary toolboxes is provided. The compile toolbox allows the software to run on compute nodes.

Data organization and bioinformatics

In the independent servers, a simple FUSE layer is used to compose independent remote directories into one. The raw microscope data and post-processing analyses are accessed through a dedicated file system.

The acquired (raw) data (image) is only readable by the users and can be written only by the microscopes. The derived metadata is put into a virtual directory i.e. <plate>/<well>/<metadata> for instance *cellcount*, and this metadata is stored in a mySQL database.

The virtual directories are available per Project, Experiment and Image, and the setup of required fields is done on a Project level, allowing the scientist to define the necessary fields for the pipeline.

Databases of genomic and associated information such as human gene loci with cross-references to synonyms, sequences, diseases, literature, and collected human and viral protein interaction data will be integrated into the system.

Tools for analysis of hit lists (genes up/down regulated) such as interaction network visualisation or interaction graph analysis will be added or implemented for the system as necessary, to complete the bioanalysis environment.

Current RNAi screening production workflow (Druggable screen example, status Feb. 2009)