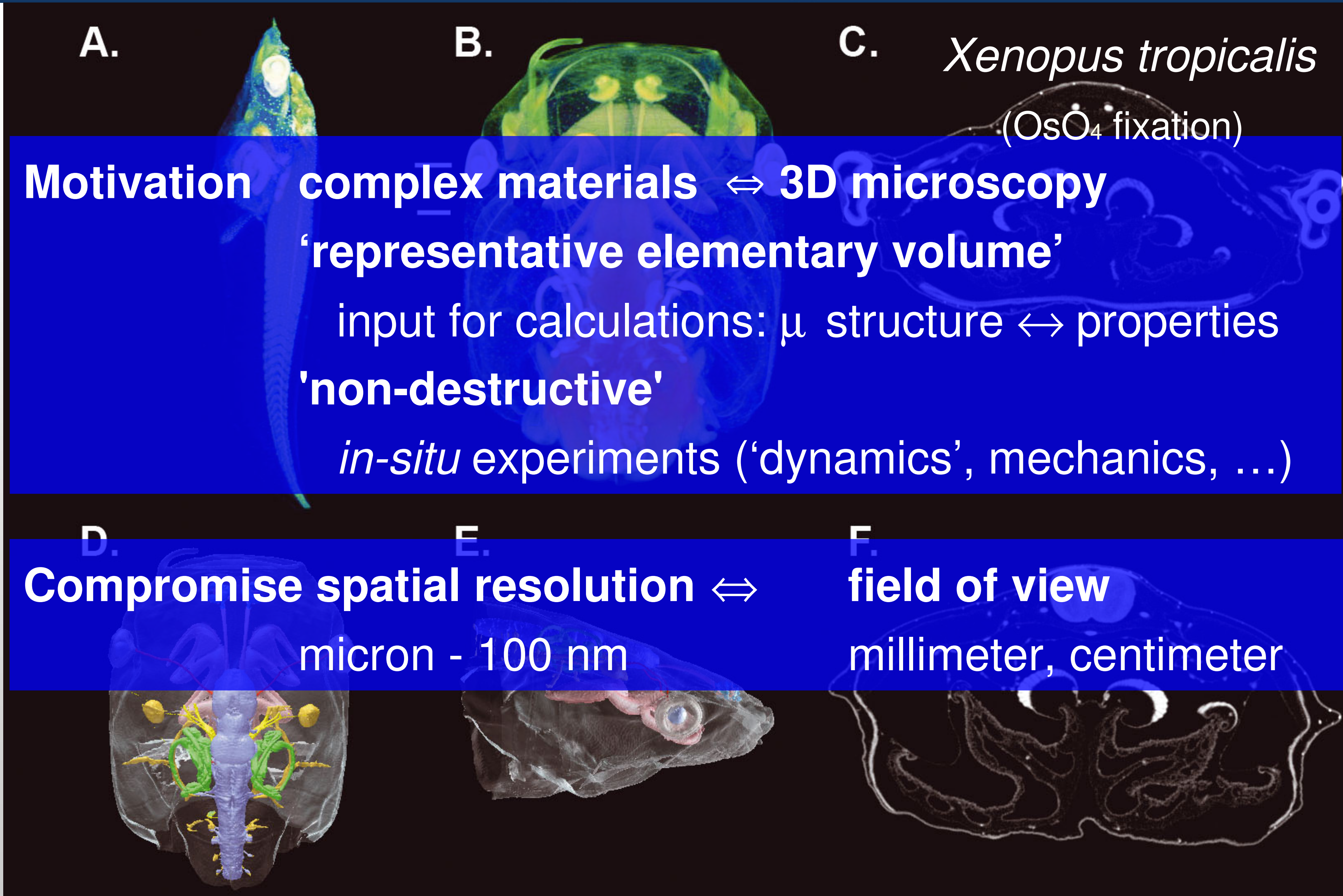
Tomography data processing and management

P. Cloetens

X-ray Imaging Group, ESRF

Computed Tomography



Motivation **complex materials ⇔ 3D microscopy**

‘representative elementary volume’

input for calculations: μ structure ↔ properties

'non-destructive'

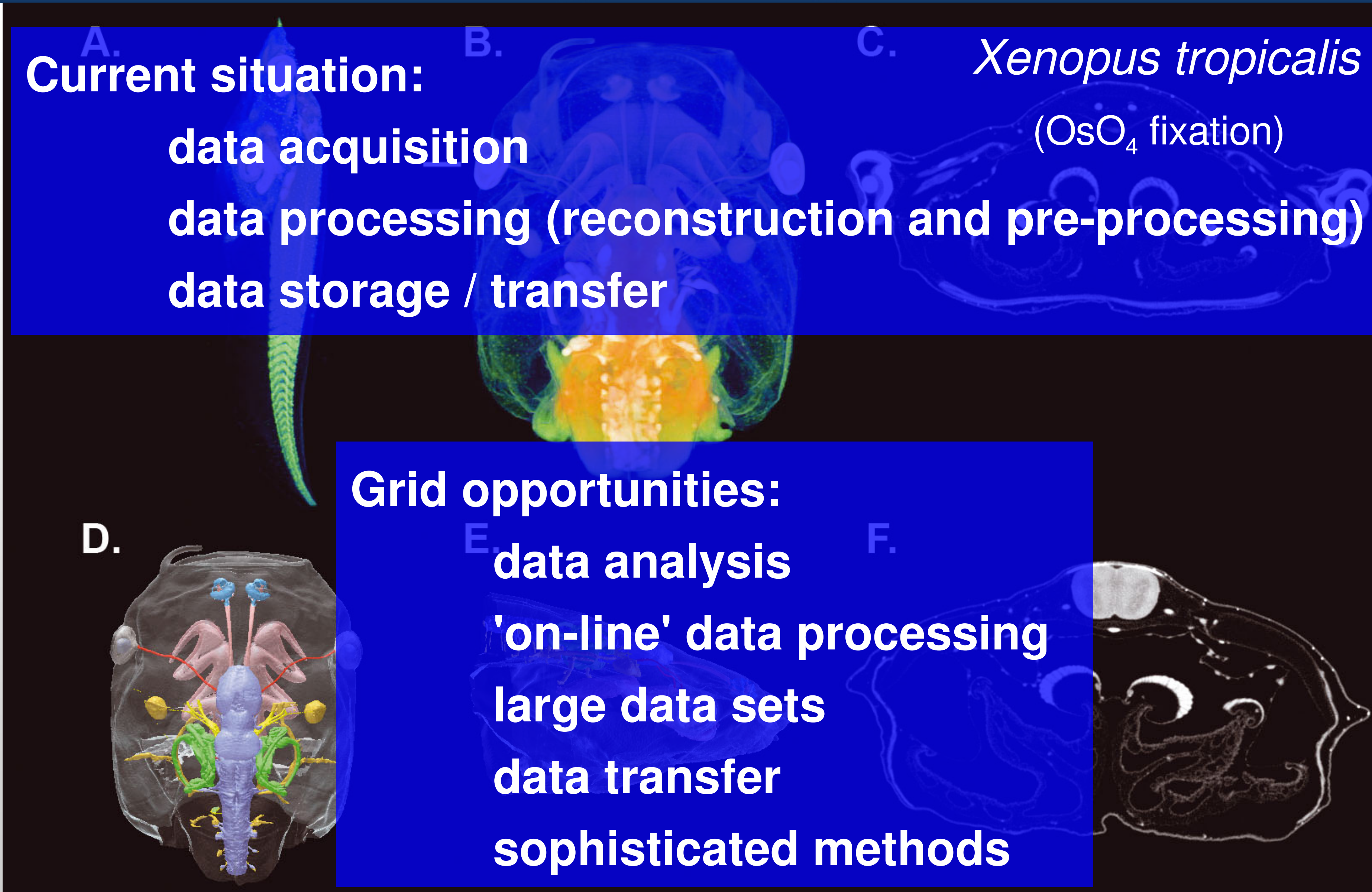
in-situ experiments (‘dynamics’, mechanics, …)

Compromise spatial resolution ⇔ field of view

micron - 100 nm | millimeter, centimeter

N Pollet (ibaic,u-psud), R Boistel (namc,u-psud), P Cloetens

Computed Tomography



Current situation:

- data acquisition
- data processing (reconstruction and pre-processing)
- data storage / transfer

Xenopus tropicalis

(OsO_4 fixation)

Grid opportunities:

- data analysis
- 'on-line' data processing
- large data sets
- data transfer
- sophisticated methods

N Pollet (ibaic,u-psud), R Boistel (namc,u-psud), P Cloetens

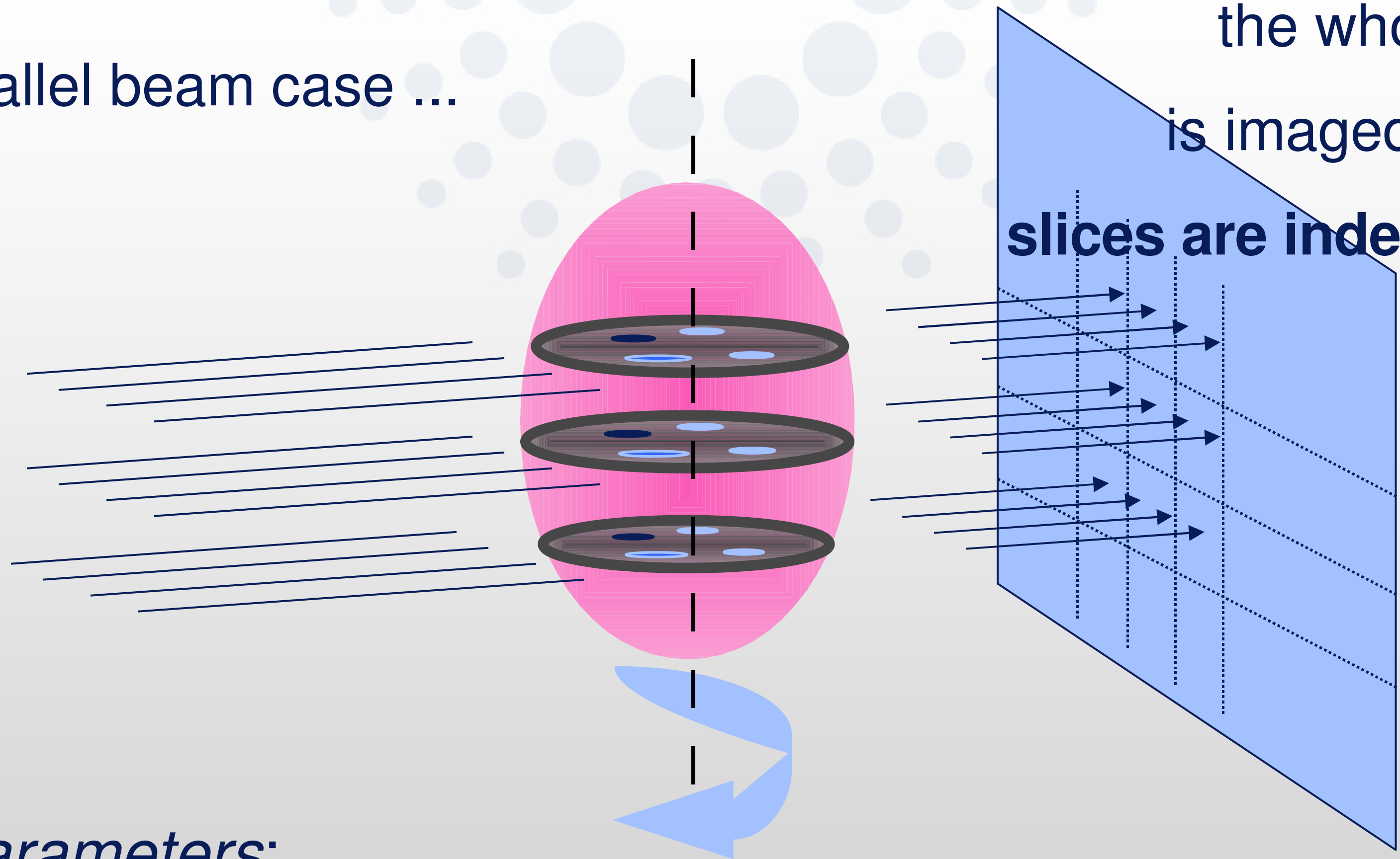
Synchrotron based tomography

Parallel beam case ...

the whole object

is imaged at once,

slices are independent



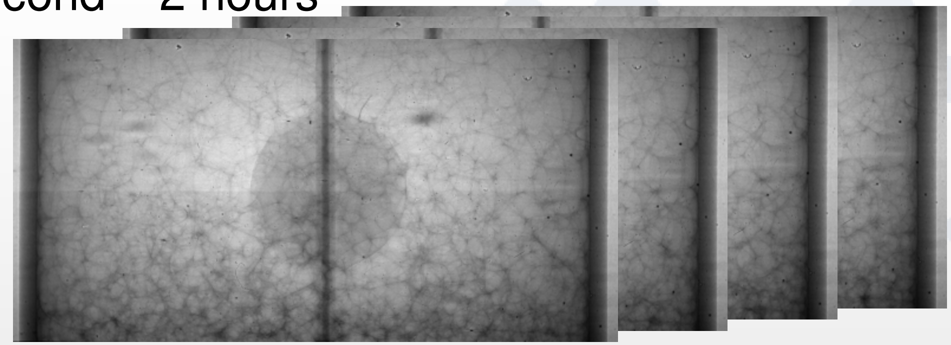
New *parameters*:

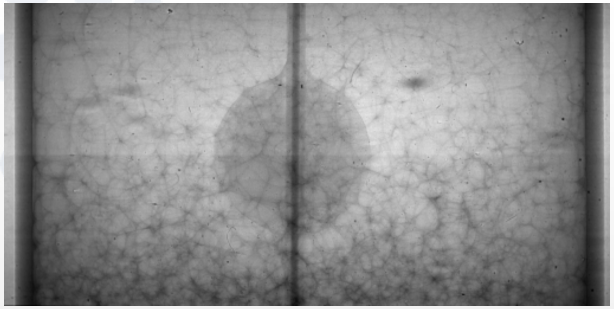
- time, *in-situ* experiments
- distance: holotomography, 3DXRD
- energy: edge CT, XANES, fluorescence

} **angles are independent**

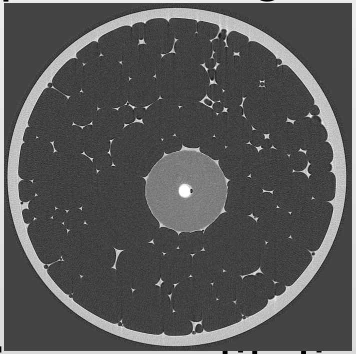
Micro-tomography work-flow

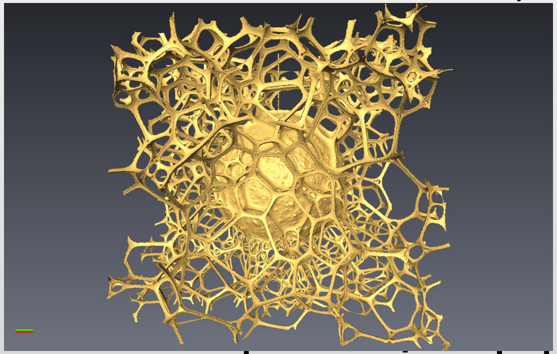
- **Data acquisition**
  ~ 1 second – 2 hours



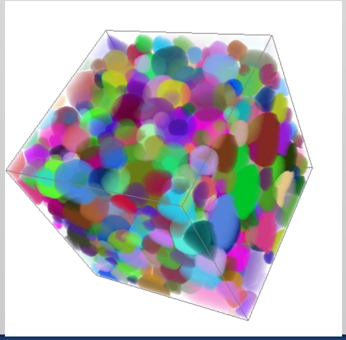


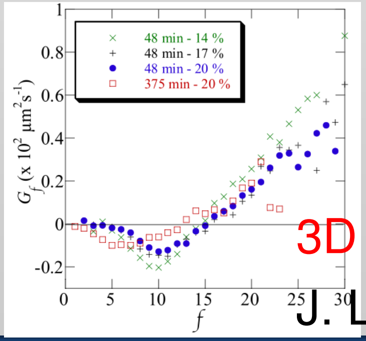
- **Data processing** (pre-processing and 3D-reconstruction)
  1 hour – 6 months





- **Data analysis** (extracting quantitative parameters, simulations)
  1 month – ∞





3D Growth Law

J. Lambert et al, *PRL* **99**, 058304 (2007)

Data acquisition

- FReLoN (Fast Read-out Low Noise)

  CCD (2k x 2k camera, '80 Mpixels/s', '160 MB/s')
- SPEC control (same macros on ID19, ID22NI, BM05, ...; Bliss)

  but Device Servers for camera, ...
- Highly optimized

  Step-by-step acquisition

  or

  Continuous acquisition
- Storage: directly to NICE file servers over 1Gb / 10 Gb link

  normal scans and network OK

  local buffer on fast disks

  fast tomography (~70 MB/s – 120 MB/s)
- Raw data: 16 GB / sample (64 GB for holo-tomography)

  1 TB / day in user experiment is feasible

Data processing

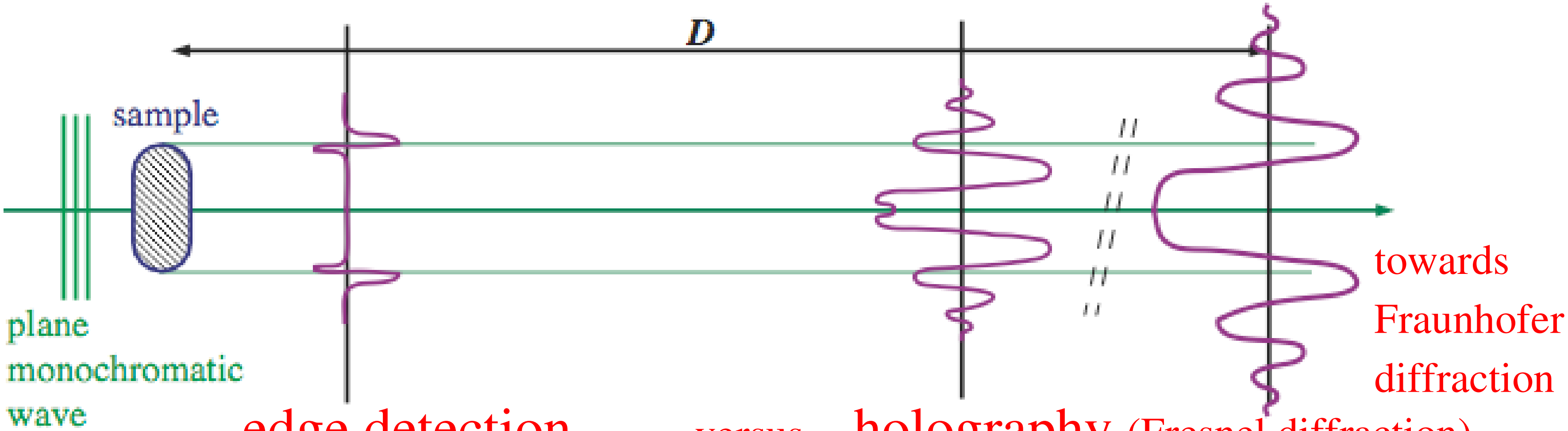
PyHST (SciSoft, A Mirone, C Ferrero) + Condor High Throughput Computing (CS, G Förstner)

- Filtered back-projection implementation
- Optimized for total through-put not for a single scan

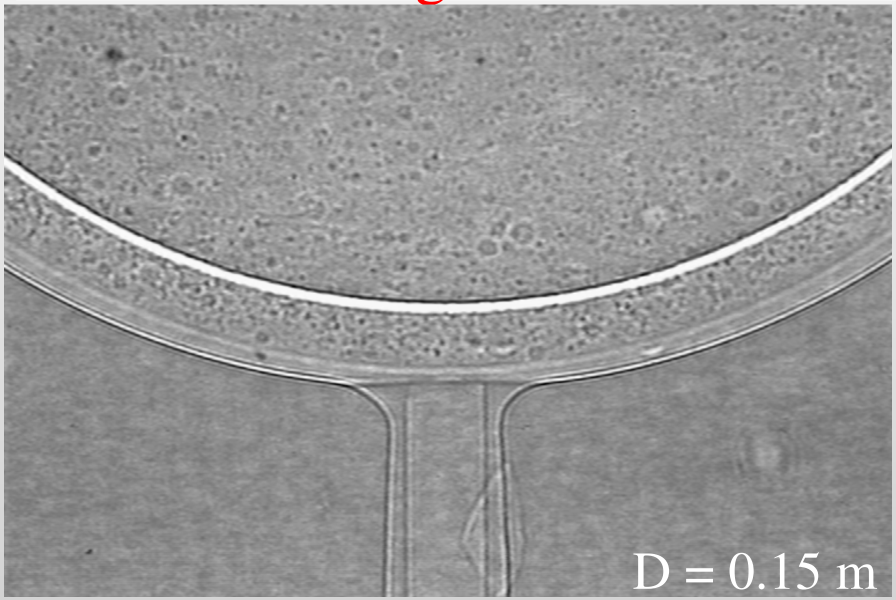
  parallelization in 1 GB sub-volumes not individual slices

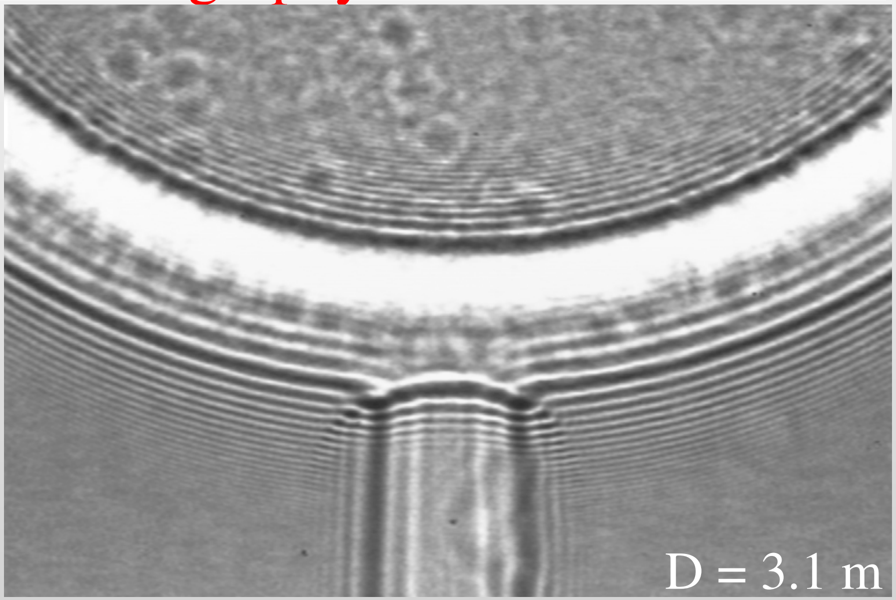
  20 dedicated cpu's (+ 140 common cpu's)
- Improved through-put by including several pre-processing steps:

  motion correction, basic ring correction, hot pixel correction, distortion correction, ...
- User interacts with Octave + ImageJ and terminal
- Special pre-processing with Octave (phase retrieval) or Matlab/IDL
- Volume reconstructed data ≃ raw data ! (no data reduction)



edge detection versus holography (Fresnel diffraction)





each edge imaged independently

no access to phase, only to border

$\sqrt{\lambda D} \approx a$

$\lambda = 0.7$ Å

50 µm

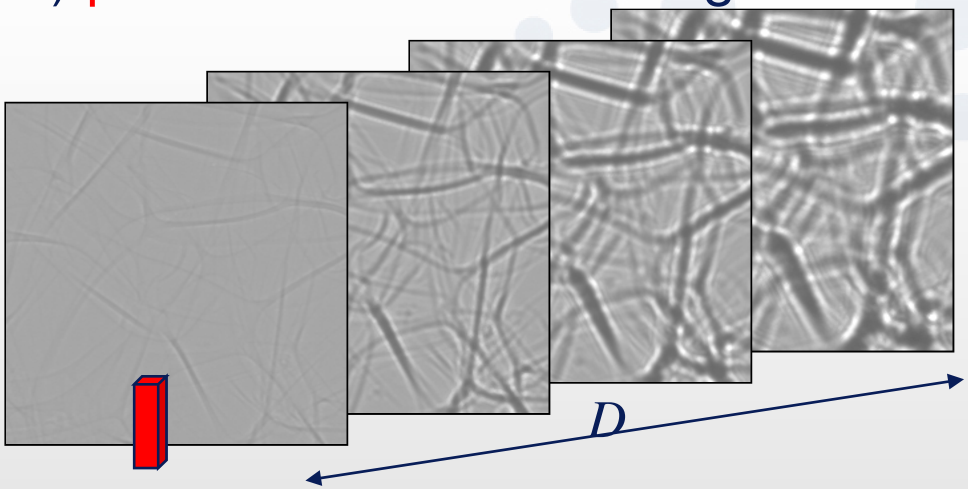
access to phase, if recorded at ≠ D's

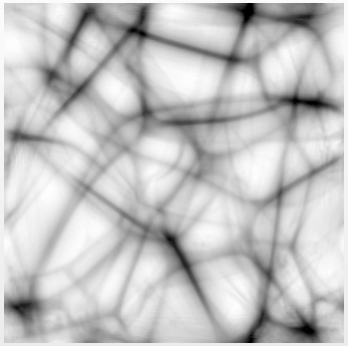
defocused *image*

$\sqrt{\lambda D} \ll a$

Holo-tomography

1) phase retrieval with images at different distances

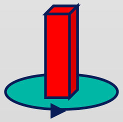


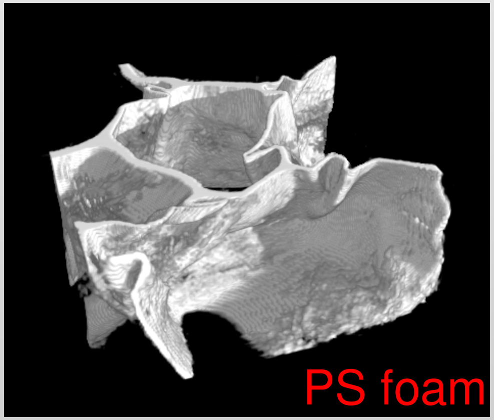


Phase map

2) tomography: repeated for ~ 1000 angular positions

3D distribution of δ or the electron-density
improved resolution
straightforward interpretation
processing



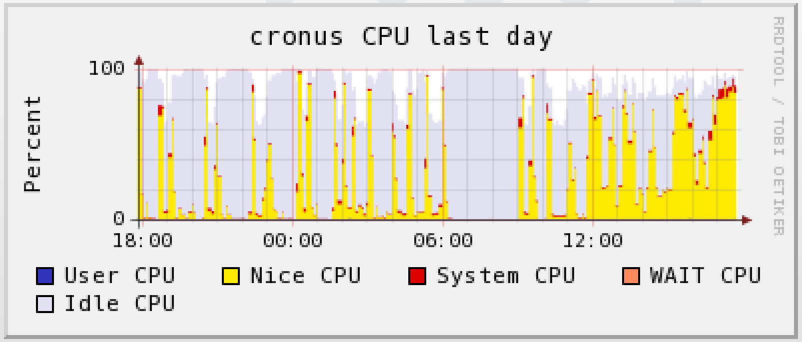


P.Cloetens et al., Appl. Phys. Lett. 75, 2912 (1999)

JP Guigay, M Langer, R Boistel, P Cloetens, Opt. Lett., 32, 1617 (2007)

Data processing: limitations on parallelization

- Data Input – Output (file servers)
  - Too many reconstruction jobs blocks processing and acquisition (!)
  - Maximum 20 PyHST jobs on dedicated cluster cronus (no flocking)



- Memory intensive jobs (holo-tomography)
  - Only 4 or 8 GB / physical machine on NICE -> 1-2 GB / CPU
  - Performance collapses on swapping
  - In practice: only 8 GB machines used, 1 job per physical machine

Data storage / transfer

- Tomography produces huge amounts of data (raw + reconstructed)
- Short term: NICE file servers
  - Case ID19 (~ 80 TB)

    inhouse: 8TB (87% full)
    inhouse1: 8TB (85% full)
    external: 8.5 TB (83% full)
    external1: 6 TB (89% full)
    external2: 6 TB (82% full)

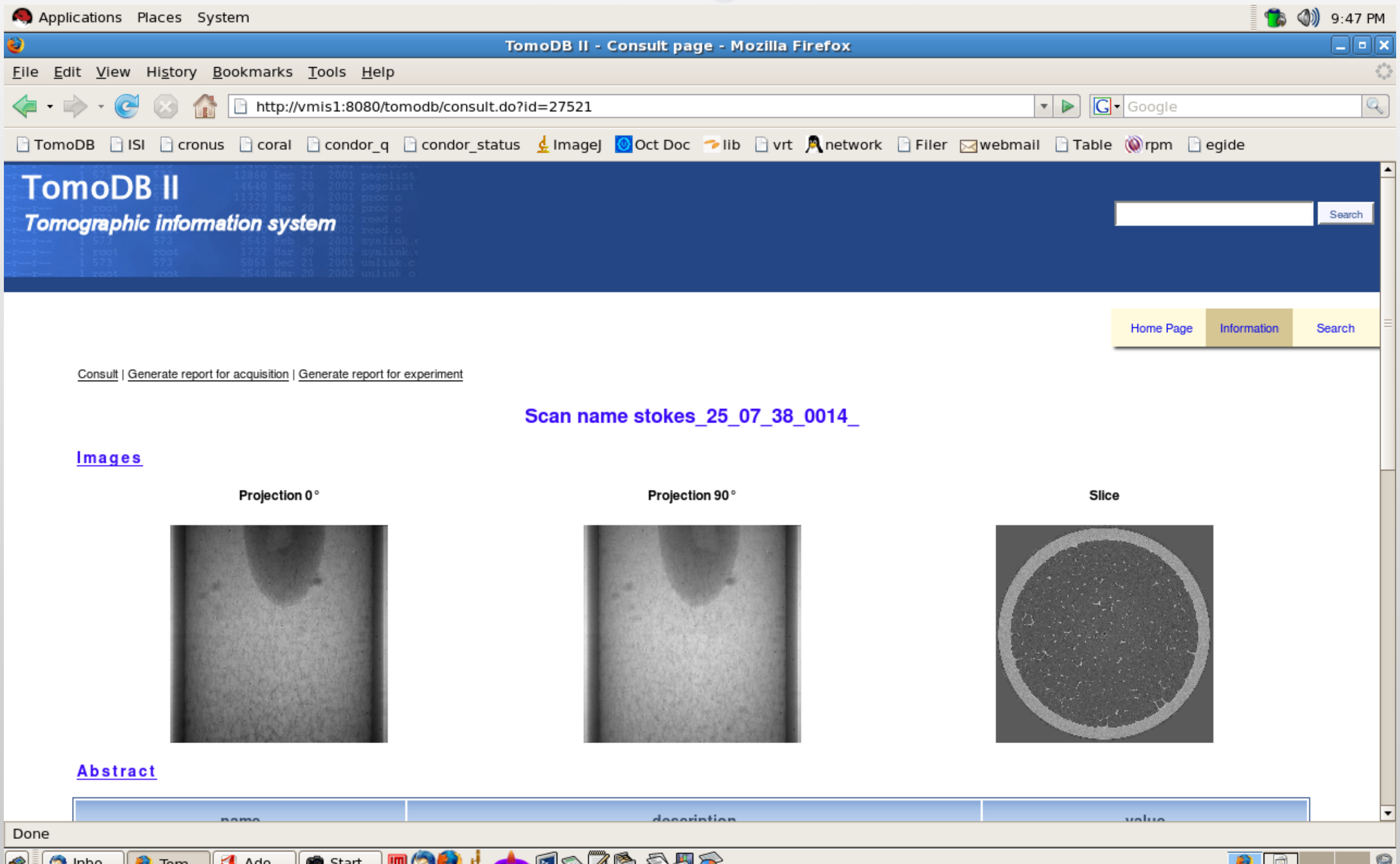
    paleo: 8TB (83% full)
    lamino: 4TB (99% full)
    graintracking: 7 TB (99% full)
- Medium term:

  6 months central backup on tape
  many data restorations from central backup (man power)
- Long term:

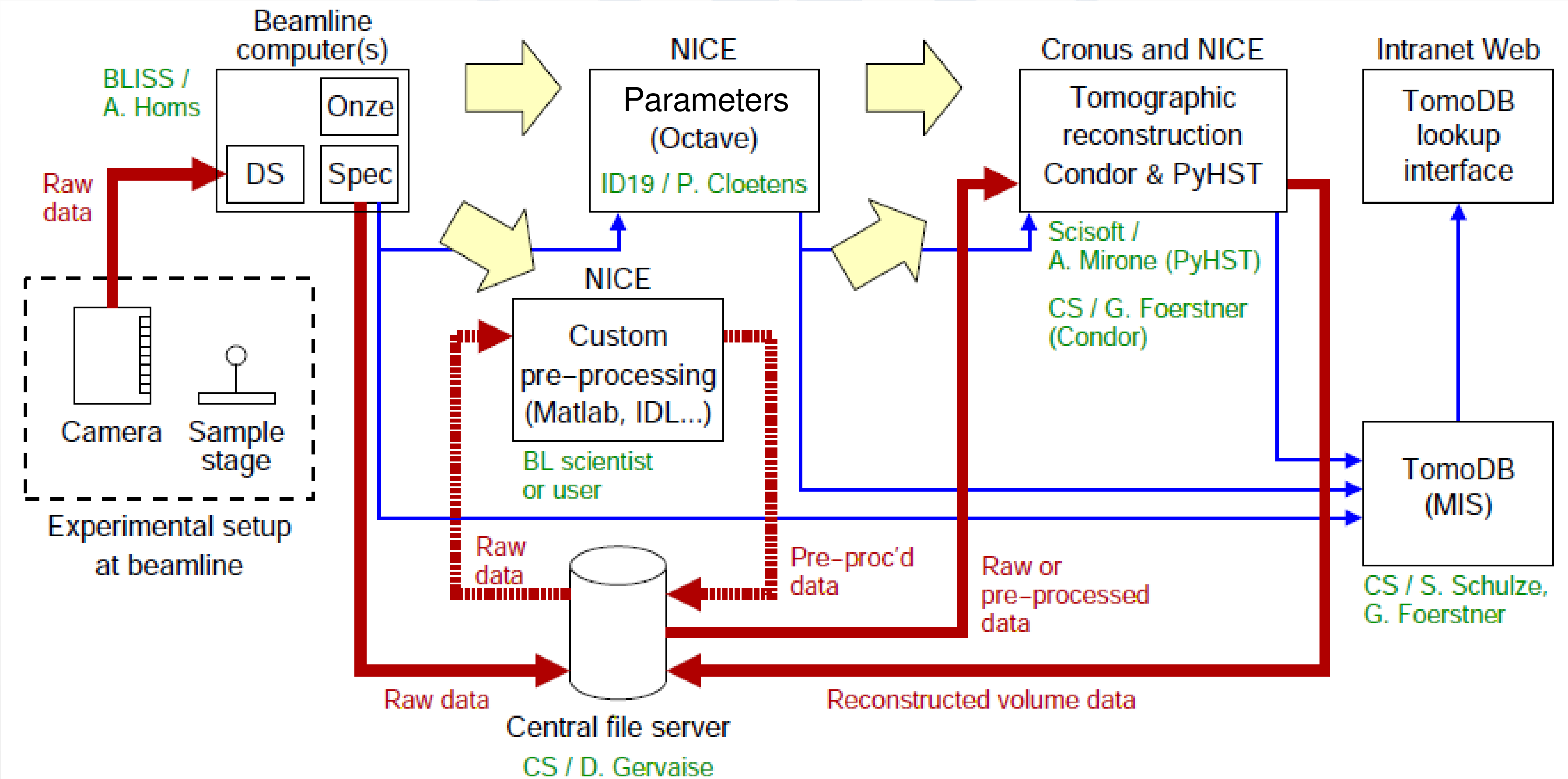
  external hard disks / typically only reconstructed data
- NICE also used for long term storage in practice (congestion)

Database

- TomoDB II database (S. Schulze, MIS)
- Most experimental conditions stored
- Not all data processing parameters
- Only available through intranet



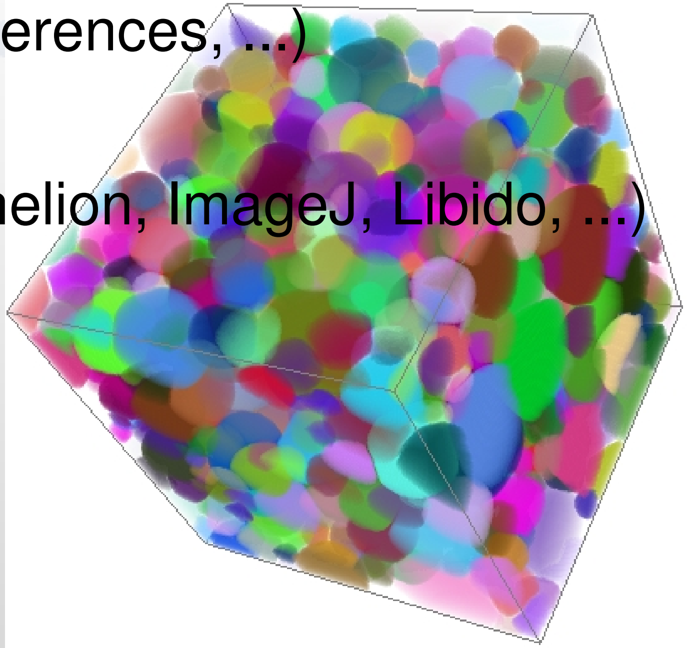
Data Flow



courtesy: T. Weitkamp

Grid opportunities: data analysis

- From simple segmentation
  over parameter extraction (local thickness, volume, …)
  to simulations (finite elements, finite differences, ...)
- Presently left to user
  Several dispersed tools (VGStudio, Aphelion, ImageJ, Libido, ...)



- Faster data analysis can significantly increase publication rate
- Faster data analysis can significantly decrease data storage
- Users could perform their analysis with ESRF resources
- Users can make specialized code available

Grid opportunities: 'on-line' data processing

- Presently: on-line reconstruction of sample slice
- On-line reconstruction of volumes
  - multi-scale approach (e.g. zoom tomography)
  - *in-situ* experiments (result defines what to do next)
- Avoid processing time >> acquisition time
  - extended stay of user
  - post-processing by user or beamline staff

Grid opportunities: large data sets

- Nano-imaging:

  improve spatial resolution with reasonable field of view

  → increase number of pixels, scales as **N^3**
- CCD's used today are only 2K x 2K

  2K x 2K → 16 GB

  4K x 2K → 64 GB

  4K x 4K → 128 GB

  Projection within the Upgrade Program:

  8K x 8K → 1 TB for a single scan !

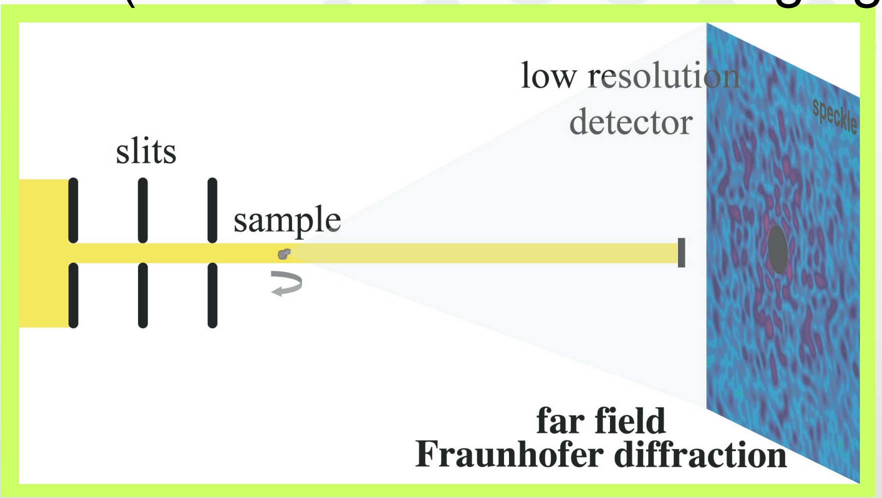
Grid opportunities: data transfer and archiving

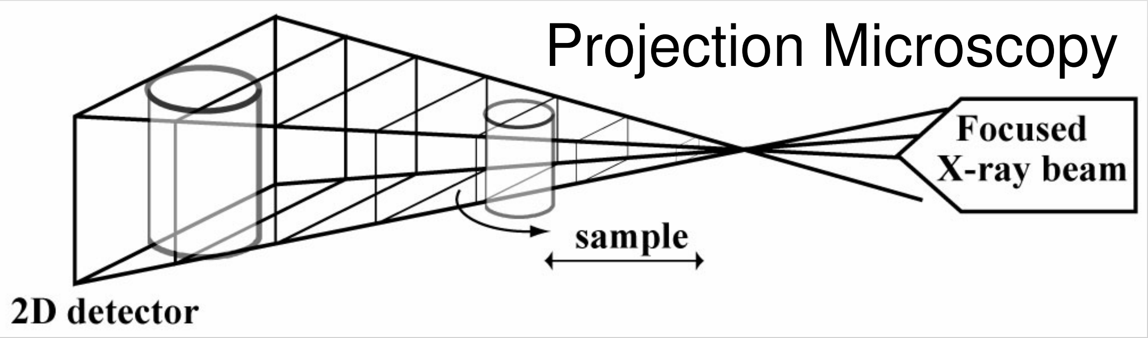
- External hard disk solution is impractical and not secure
- **Reliable** and fast data transfer over network
- Long term archiving for inhouse data is under discussion

Grid opportunities: sophisticated methods

- Nano-scale imaging:

CDI (Coherent Diffraction Imaging)





- Very computer intensive methods (CPU power, memory usage)

The Nano-Imaging end-station ID22NI

Projection Microscopy - PXM

electron density Dose efficient, fast ; magnified holotomography

Scanning Transmission Microscopy - STXM

X-ray Fluorescence: element distribution Slow ; trace elements

Grid opportunities: Iterative Phase Retrieval

Fluotomography

Chemical identification

Magnified holotomography

Electron density mapping

CPU bound

Iterative minimization of cost functional

Conclusions

• Tomography is a three steps process: data acquisition, data processing (reconstruction) and data analysis. The amount of data involved is huge. There is no data reduction before the data analysis.

• Parallelization is straightforward (independent slices or angles).
It is limited in practice by data input/output.

• Possible grid opportunities:
- data analysis
- data transfer
- more sophisticated methods of pre-processing, reconstruction