“Big Data” at the European spallation source

Jon Taylor
Head DMSC

CREMLIN WP2 Workshop on Big data Management 15-16 Feb 2017

Summary

- ESS users, instrument suite and scientific work-flow

- Instrument control architecture
- Data reduction

- Data analysis (software per technique, infrastructure)

- Data Policy and Management

- ***Issues and foreseen solutions regarding the ever growing data volume***

A green field site is a real benefit for implementing a good data / software architecture.

Prologue

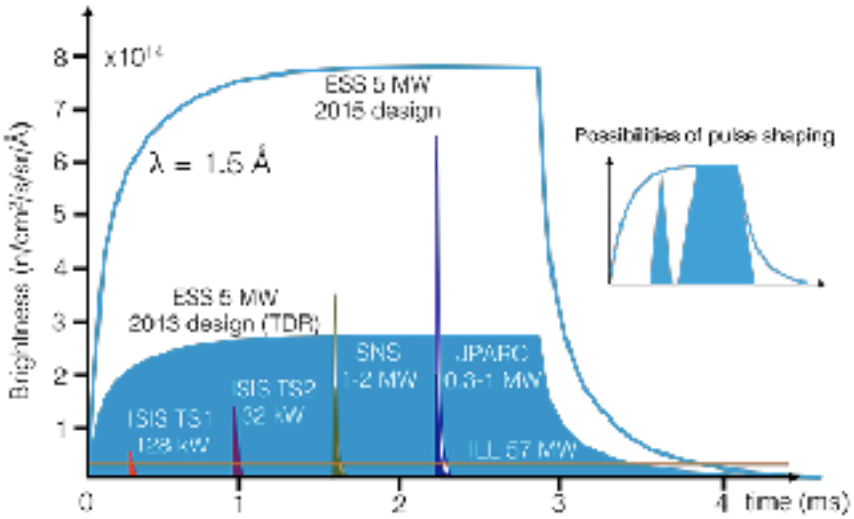
Data / software / architecture / instrument must be viewed as a unifiable system

We use apache Kafka for data streaming

The rest is just big data

ESS

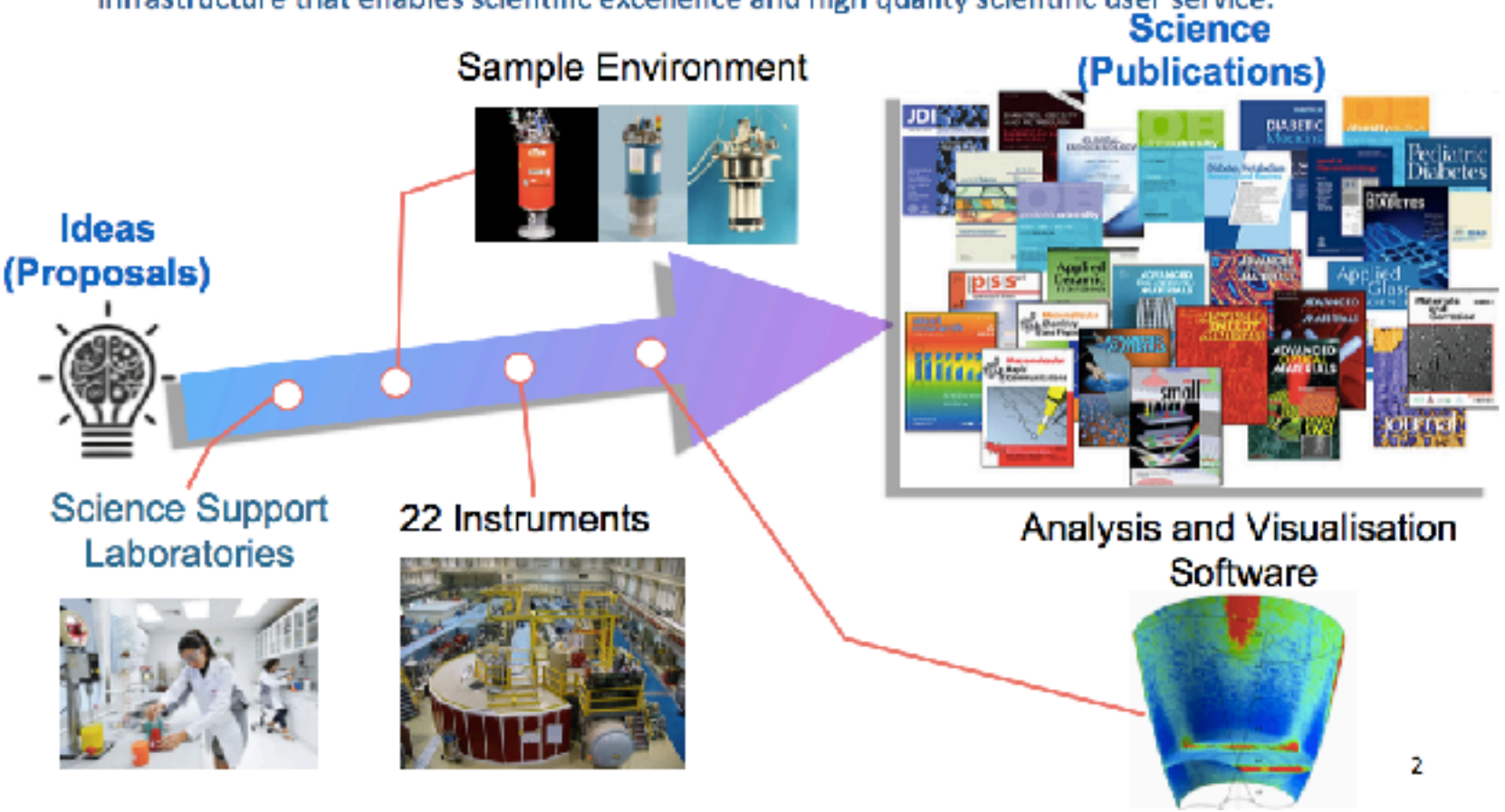
- High power long pulse spallation neutron source
- 5MW
- New scientific possibilities
- User programme 2023
- 1843kEuro
- 17 European partners



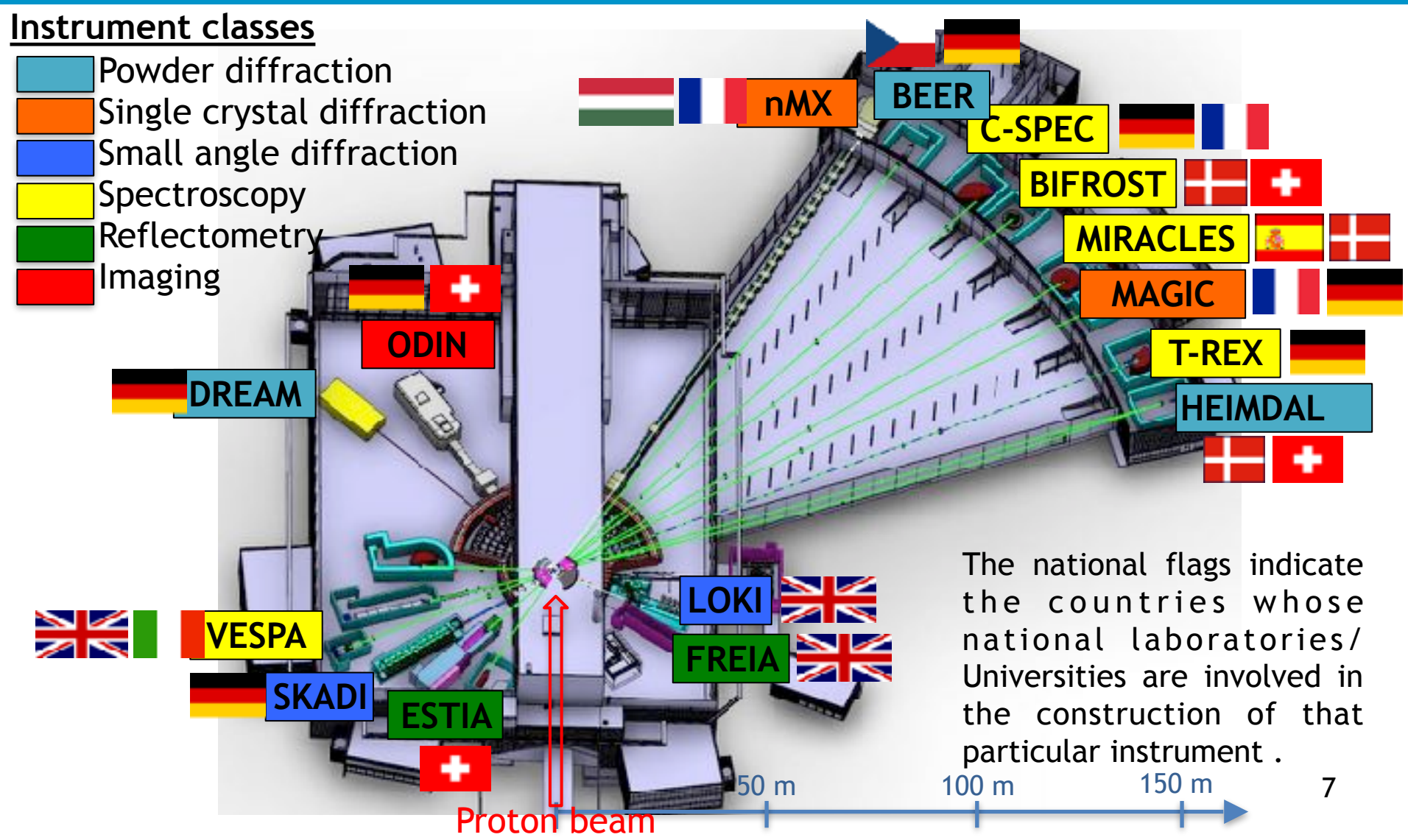
ESS construction site

Neutron scattering systems project

NSS Scope: 22 **public** instruments by 2028 together with a technical and scientific support infrastructure that enables scientific excellence and high quality scientific user service.



Common DAQ Control & reduction



The national flags indicate the countries whose national laboratories/ Universities are involved in the construction of that particular instrument .

ESS instrument suite

EUROPEAN SPALLATION SOURCE

Instrument Class	Instrument	Costbook (M€)	Upgrade (M€)	Performance target (@ 2MW)
Large Scale Structures	LOKI (Broad band SANS)	12.19	3.0	5 x D22 & 20 x SANS2D
	SKADI (General Purpose SANS) *(+SONDE funds)*	11.50	3.0	4 x D22
	ESTIA (Focusing Reflectometer)	11.80	4.6	• Conventional mode: ~ 100 x D17 • High intensity mode: 1cm² samples = seconds
	FREIA (Liquids Reflectometer)	*13.2*	5.0	30 x FIGARO, INTER
Diffraction	DREAM (Bispectral powder diffractometer)	13.66	5.1	> 10 x POWGEN or WISH
	HEIMDAL (Hybrid diffractometer)	13.55	3.7	~ 50 x GEM, ~ 8 x new POLARIS
	MAGIC (magnetism single crystal diffractometer)	13.10	1.9	• Cold: > 100 x worlds best, • Thermal: 1mm³ crystals = 10 min
	NMX (Macromolecular crystallography)	11.67	2.5	> 10 x LADI & Biodiff
Engineering & Industrial	BEER (Engineering diffractometer)	14.99	9.3	world leading in strain scanning, unique flexibility
	ODIN (multi-purpose imaging)	11.60	5.8	world leading for high resolution, > 10 x best for TOF methods
Spectroscopy	BIFROST (extreme environment spectrometer)	13.45	2.4	> 10 x THALES & MACS
	C-SPEC (cold chopper spectrometer)	16.50	2.4	2 - 6 x IN5
	T-REX (bispectral chopper spectrometer)	16.85	3.1	3 x 4-SEASONS, 3 x IN5
	VESPA (vibrational spectroscopy)	*12.0*	2.9	10 x VISION (ΔE = 130 meV)
	MIRACLES (backscattering spectrometer)	13.53	1.7	2 x BASIS and DNA

ESS users

	ESS[a]	ILL[b]	ESRF[c]	ISIS[d]	SNS[e]
Instruments (incl. CRGs/3)	22	31	35.6	30	18.3
Beam days for user programme	3520	3899	4404	*2678*	*2525*
per instrument	160	126	124	*89*	*138*
Number of operational days	200	158	225	*130*	*184*
Number of experiments	1173	785	1358	*766*	*681*
Average Experiment Length [days]	3.0	5.0	3.2	3.5	*3.7*
Local Contacts / Instrument	2.7	2.2	3.4	2.2	2.6
Instrument Support Staff / Instr.	1.0	1.0	1.6		0.9
Sample Environment Staff / Instr.	0.9	0.5		0.6	0.9
Other Tech. Support / Instrument	4.9	3.6	4.1	3.0	6.5
Total Staff / Instrument	**9.5**	**7.2**	**9.0**	**5.9**	**10.8**
Total Staff / Experiment	**0.18**	**0.28**	**0.24**	**0.23**	**0.29**

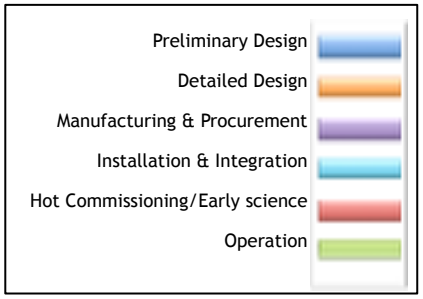
Neutron Beam Instrument Schedule
Draft awaiting council agreement

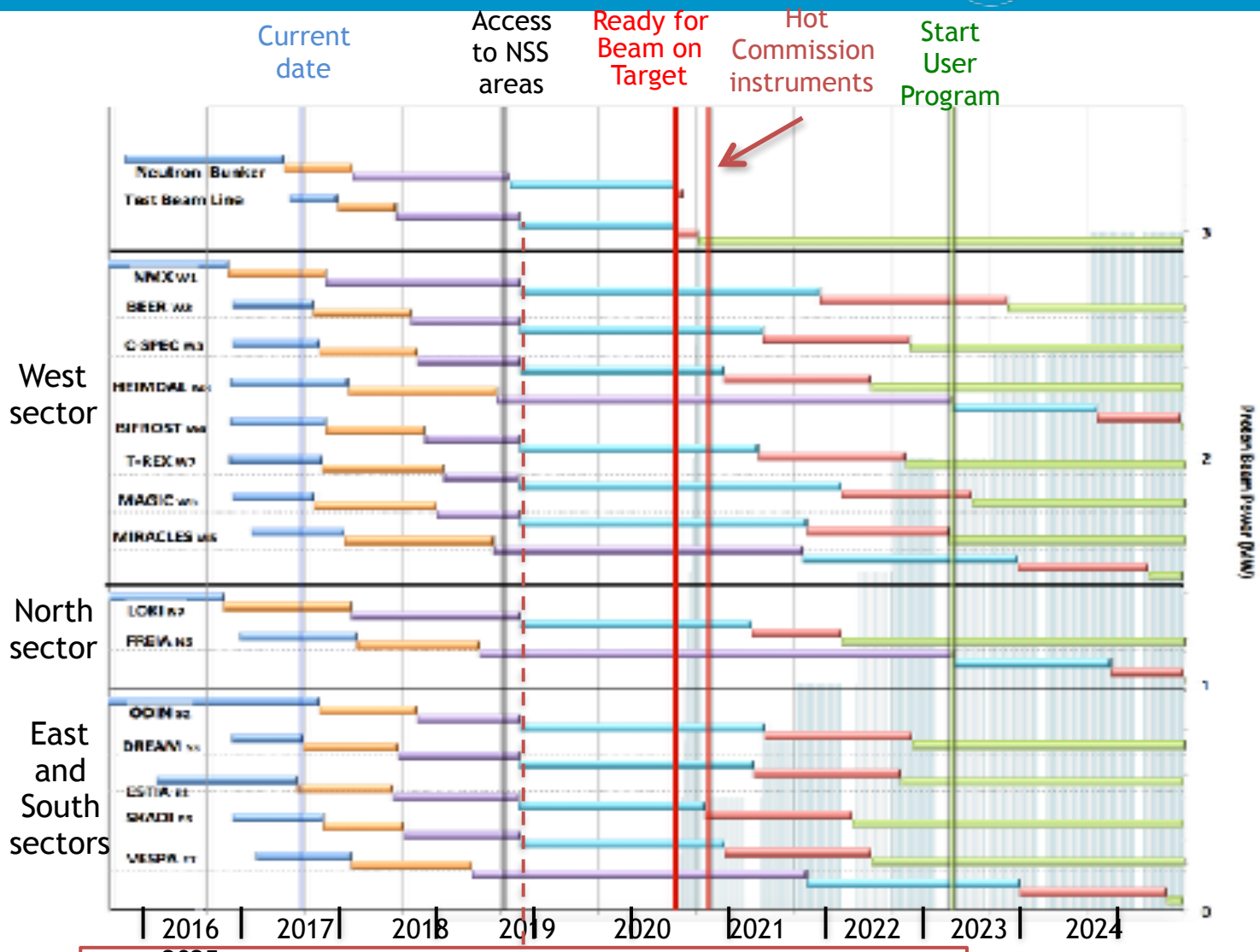
EUROPEAN SPALLATION SOURCE

Early bldg access dates;

- Monolith (D02) ’19 — 7 Feb
- North (D03) ‘19 — 1 Nov
- East (D01-E) ’20 — 6 Jan
- South (D01-S) ‘20 — 2 Mar
- West (4 areas)
  - ➢ Guides (D03-W) — 1 Nov ‘19
  - ➢ Guides (E02-2) ’19 — 15 Oct
  - ➢ Guides (E02-1) — 27 Apr ‘19
  - ➢ Caves (E01) — “

Provisionally aligned with Accelerator planning





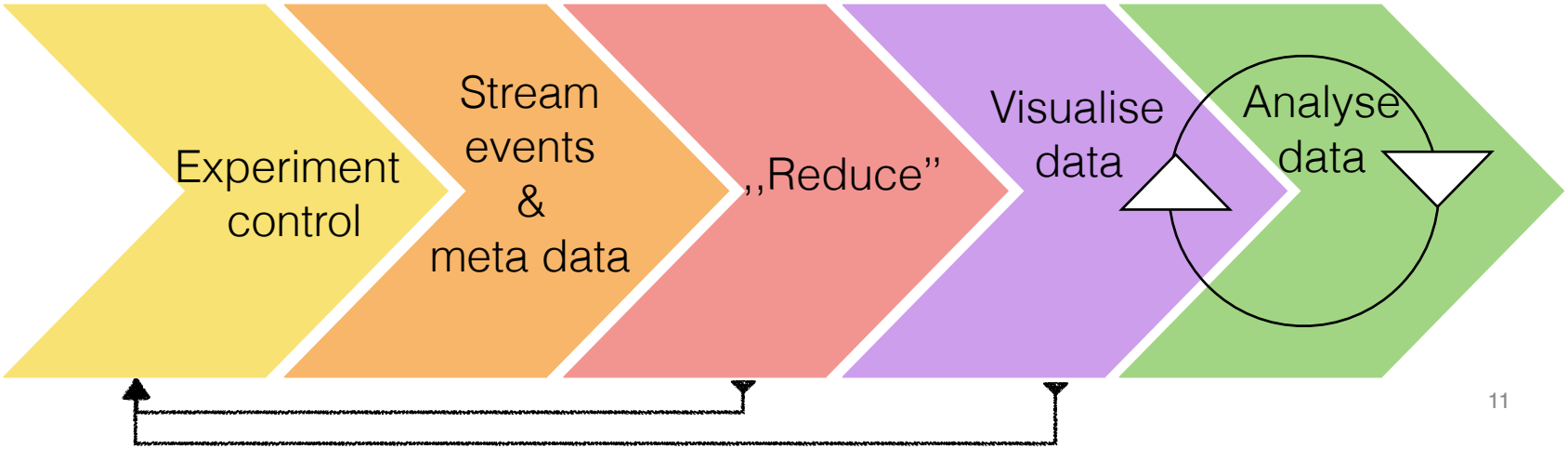
Installation for most instruments begins in Q2/2019 with in-monolith optics
Access to other “instrument construction areas” will be later (see top left panel)

The scientific computing challenge

1. Understanding (even accessing) data is the key barrier to scientific publication and impact

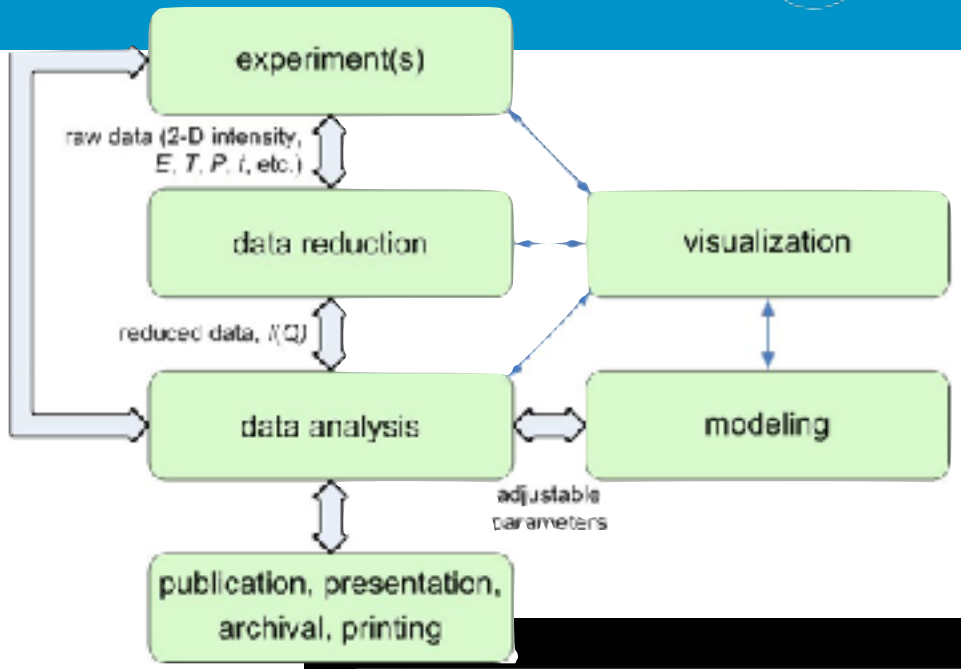
2. ESS will generate a lot of data

3. Workflow and architecture has to promote science



What is an experiment…

- Planning
- Setup
- Acquisition
- reduction & re-reduction
- analysis
  - keep going until either the coffee or beam time runs out



Centralised scientific computing

- Lessons learned from other facilities
  - All facilities moving in this direction.
- Single point failures
- Curation of software & data
- Planning & strategy
- User programme efficiency & impact
- Collaborative development

Data Management and Software Centre

Provide world leading scientific software and scientific computing support for neutron scattering at ESS

Scientific Software development.

- ➢ The ESS experiment control system
- ➢ Data acquisition software.
- ➢ Data correction software.
- ➢ Data visualization software.
- ➢ Software to model and analyze experimental data

DMSC offices located at COBIS. Copenhagen University north campus

Data centre operations.

- ➢ Store & catalogue ESS neutron datasets.
- ➢ Provide ESS users remote access to their data
- ➢ Compute provisioning for live data correction, visualization and analysis software during and after experiments.

User programme support (operations phase)

- ➢ Provide support & assistance to ESS users for data treatment and data analysis.

DMSC construction phase strategy

- Collaboration.
- In-kind effort from domain experts
- European funded initiatives

Science & Technology Facilities Council
ISIS

NEUTRONS FOR SOCIETY

Jülich Centre for Neutron Science

Heinz Maier-Leibnitz Zentrum

PAUL SCHERRER INSTITUT
PSI

Science and Innovation with Neutrons in Europe 2020

A European H2020 project

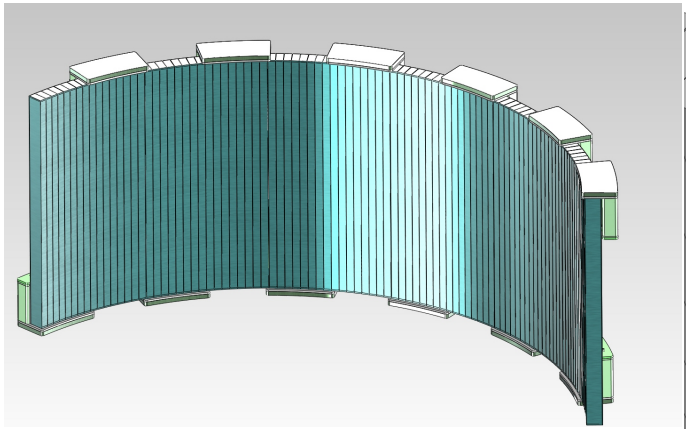
brightness

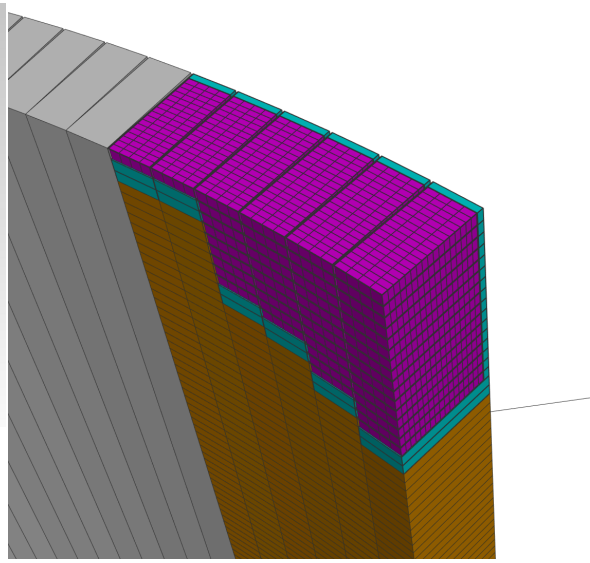
In-kind funded work packages

- Data streaming - STFC active
- Data reduction and experiment control STFC active
- Experiment control PSI active
- Data curation PSI active
- Imaging data treatment PSI approved at PSI
- Reflectometry / QENS / Engineering diffraction data treatment FZJ agreed by FZJ

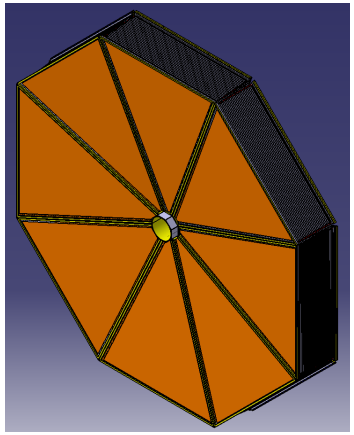
The detector geometry challenge

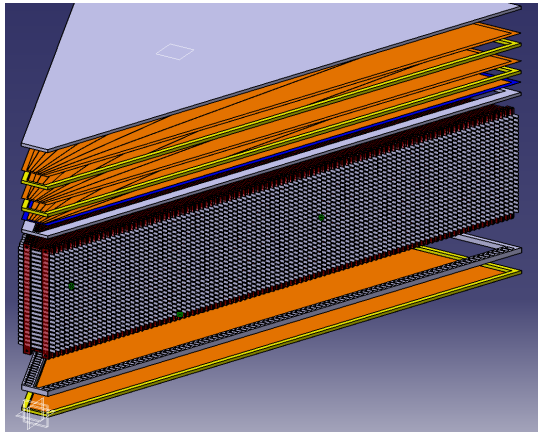
- 3 key issues
  - Shape
  - Position
  - Number



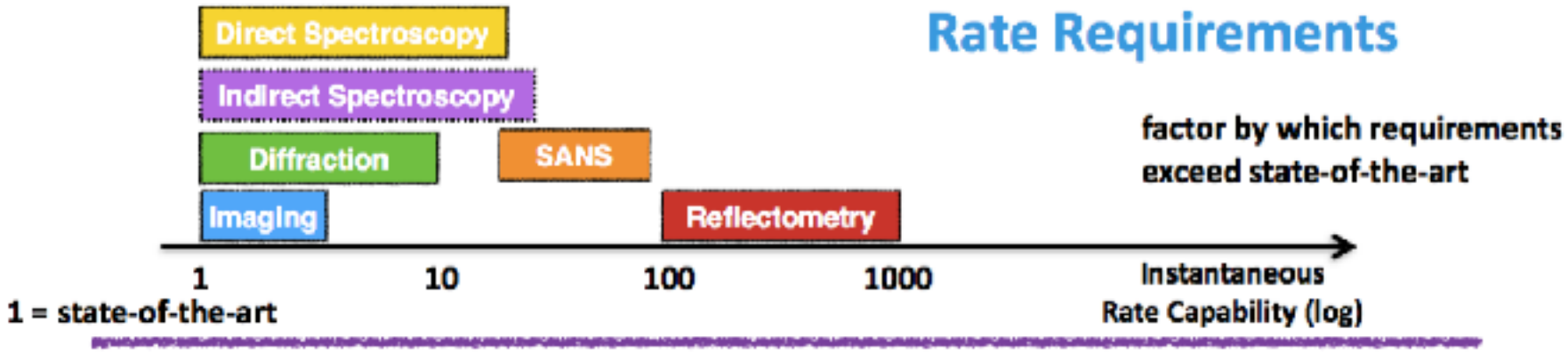


- 10B Detectors are not geometrically simple to describe
- Channel count is increased to deal with the rate
- Each instrument is different
- Instrument detectors are not always static
- They all flow gas…





Data rate challenge



The detector rate challenge presents the same challenge to DMSC

Data rate

- Event rates calculated from instrument proposal data
- Event data is collected as a list of timestamps for each spectrum
- High data rate creates a long list
- ESS average rate is ~8GB / min

	Event rate corrected for sample	MB/s/MW	1 minute collection GB (1MW)
LOKI	1.E+06	20	1.2
SKADI	0.E+00	20	1.2
Estia	4.E+07	500	30
Friea	6.E+05	10	0.6
Dream	9.E+06	100	6
Hiemdal	1.E+07	150	9
BEER	2.E+05	3	0.18
Magic	4.E+07	500	30
NMX	7.E+03	0.1	0.006
ODIN	-	150	9
CSPEC	3.E+06	40	2.4
TREX	7.E+05	10	0.6
MIRICLES	3.E+05	4	0.24
BIFROST	7.E+7	500	30
VESPA	3.E+05	4	0.24

To solve the data rate challenge

- Processing has to keep up with the experiment
- Optimise software, hardware and network
- Use the event streaming method of acquisition and live data treatment.
- Use automated reduction
- Use distributed computing SAAS
- Improve Mantid performance

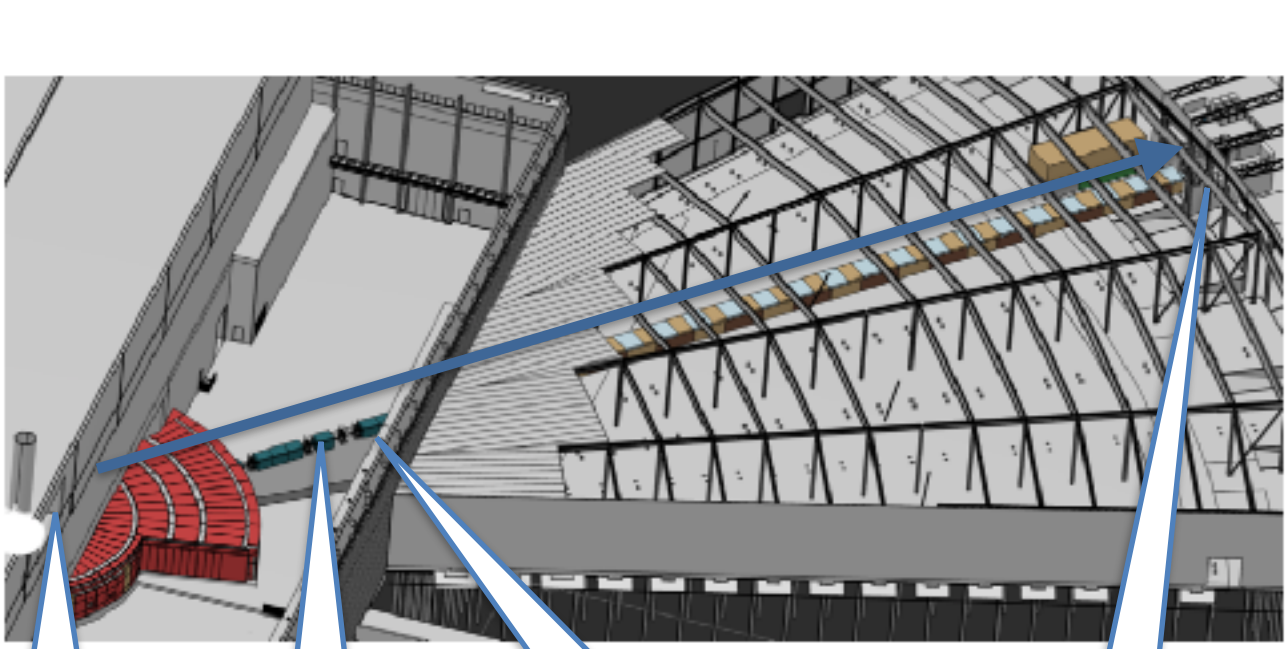
Data acquisition & control

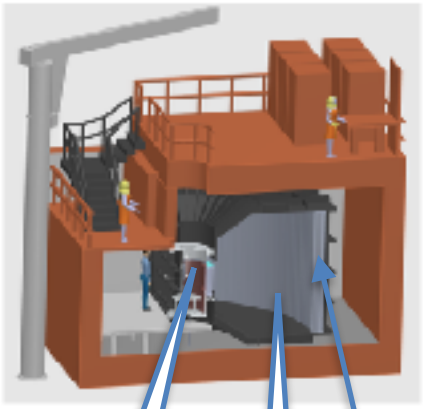
- DAQ synced to accelerator 14Hz pulse
- Event mode data collection
- All meta data timestamped from ESS central timing system
  - Timing system is delay compensated
- Neutron event data is aggregated with metadata on a frame by frame basis
- 71ms Latency is key
- Frames of data on a pub / sub network
- Vetos and meta data filtering are performed in software
- Requires high performance network, software and distributed computing
- Negates requirements for dedicated DAQ electronics

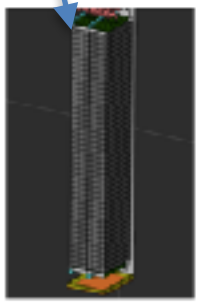
TOF Neutron scattering instrumentation

EUROPEAN SPALLATION SOURCE

The neutron energy is encoded in its *Time of Flight*







Neutron source

Neutron choppers

Neutron guide

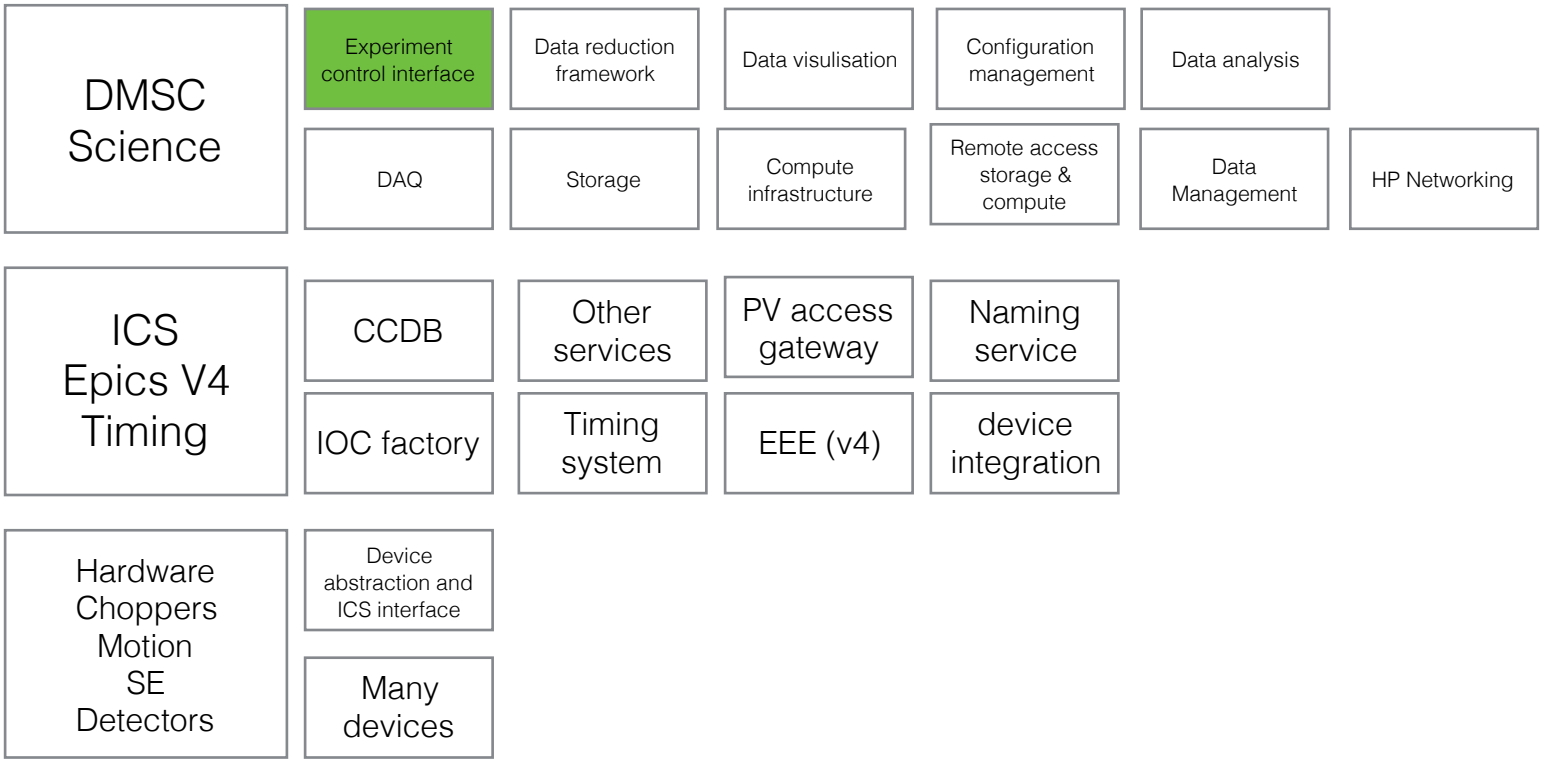
Neutron instrument

Sample position

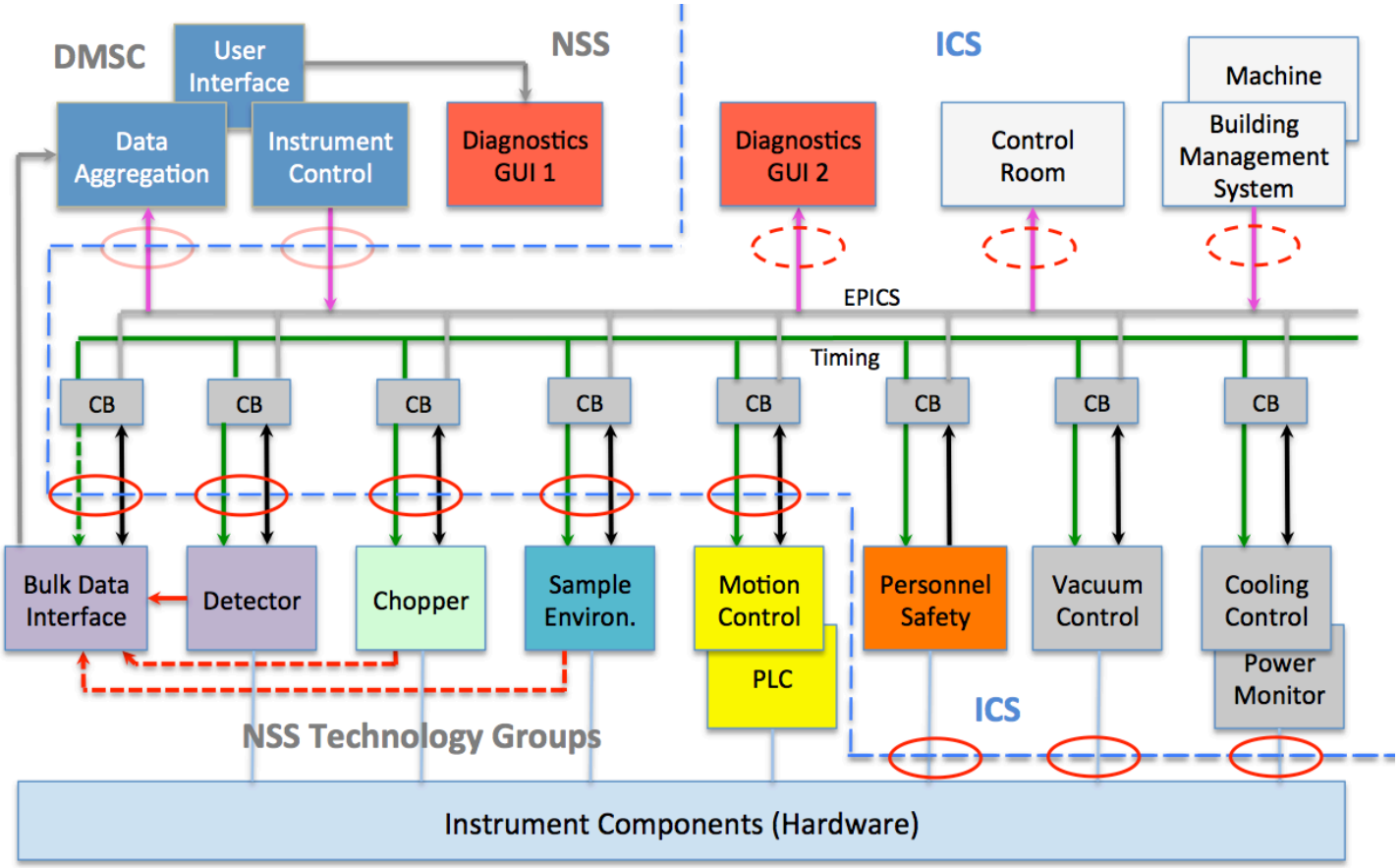
Detector tank 3D position voxel dets

Experiment control

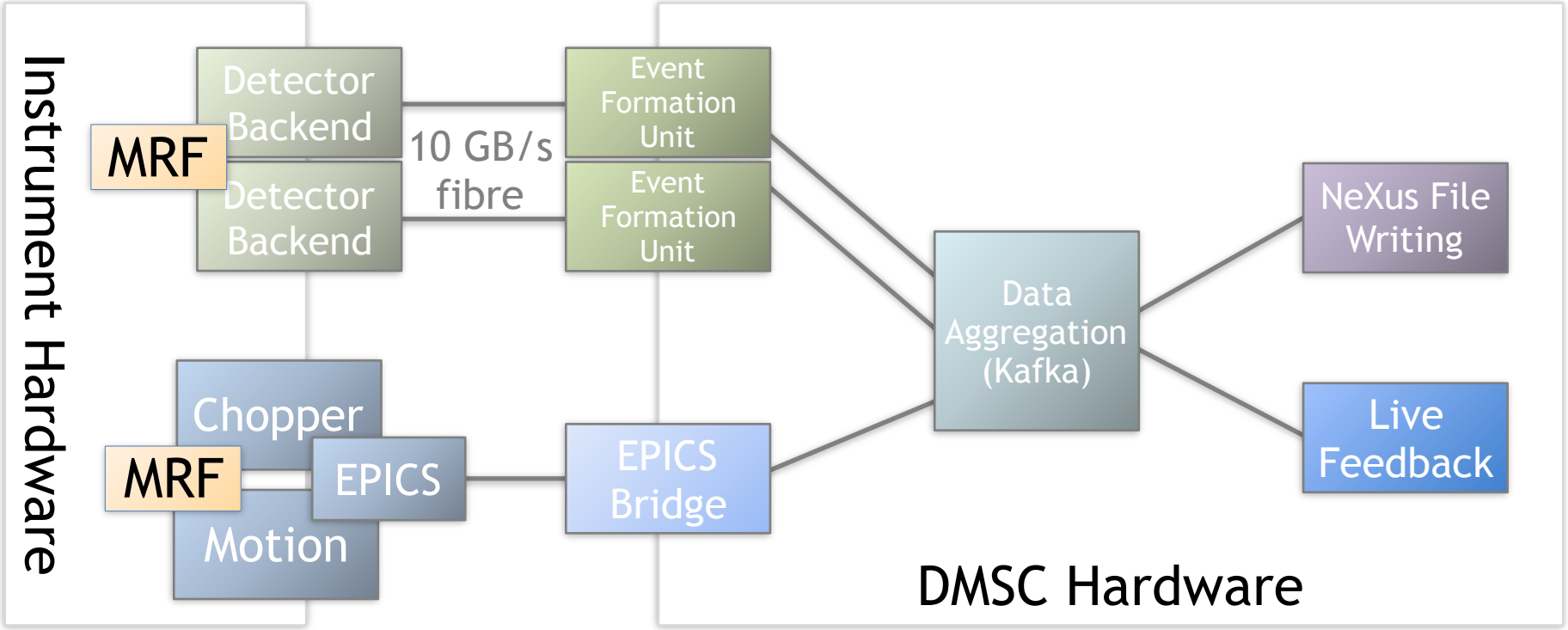
- Device abstraction layer is EPICS
- Experiment control layer will be python based
- utilising a number of tools to deploy epics control onto instruments
- Experiment control at ESS will be a beyond the current state of the art.
    - Instrument science cases are based on complex instrumentation
    - Complex modes of data acquisition WFM,RMM
- This complexity has to be useable.
- There is no off the shelf solution that fits all the ESS edge cases



Controls layout

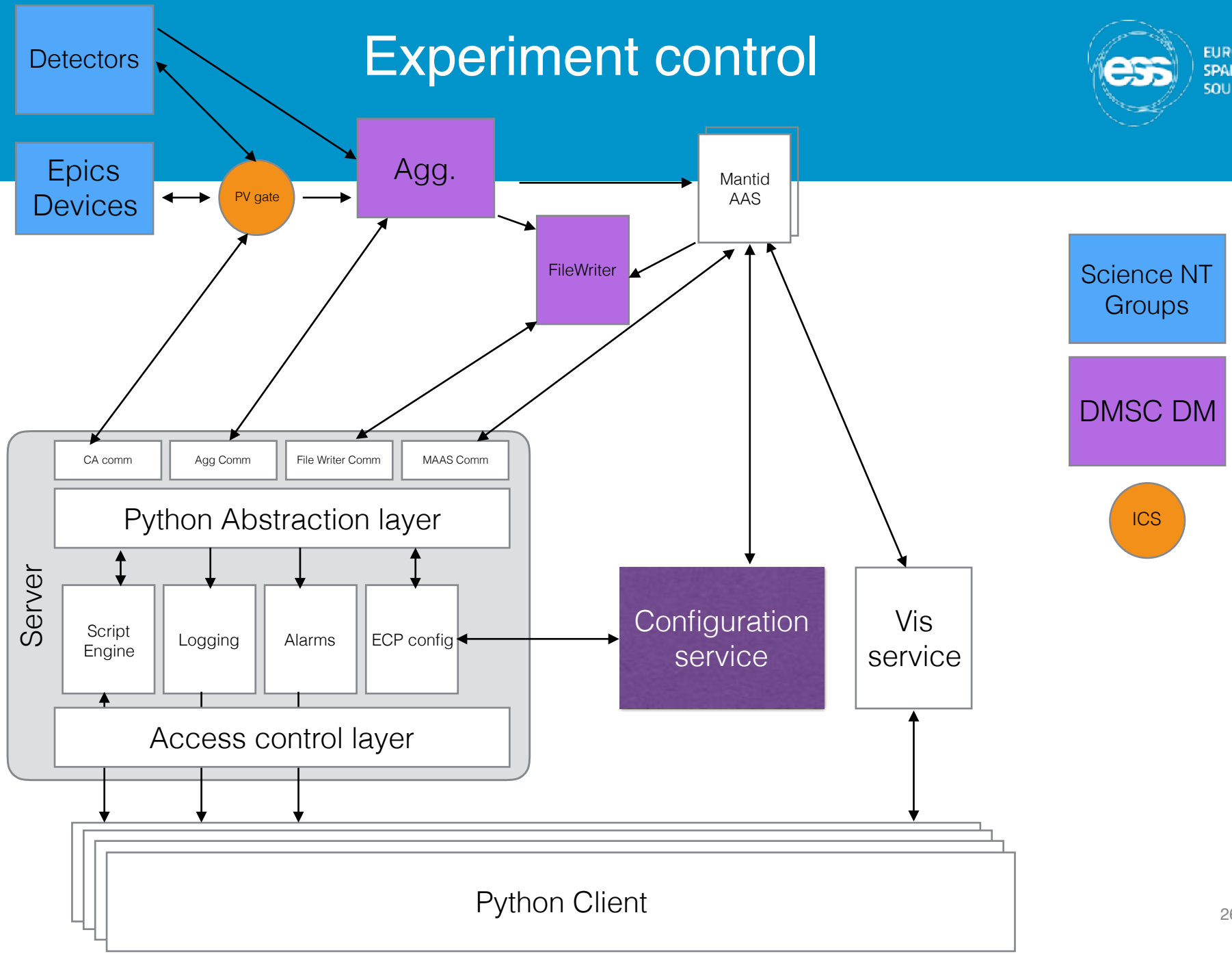


Readout Architecture

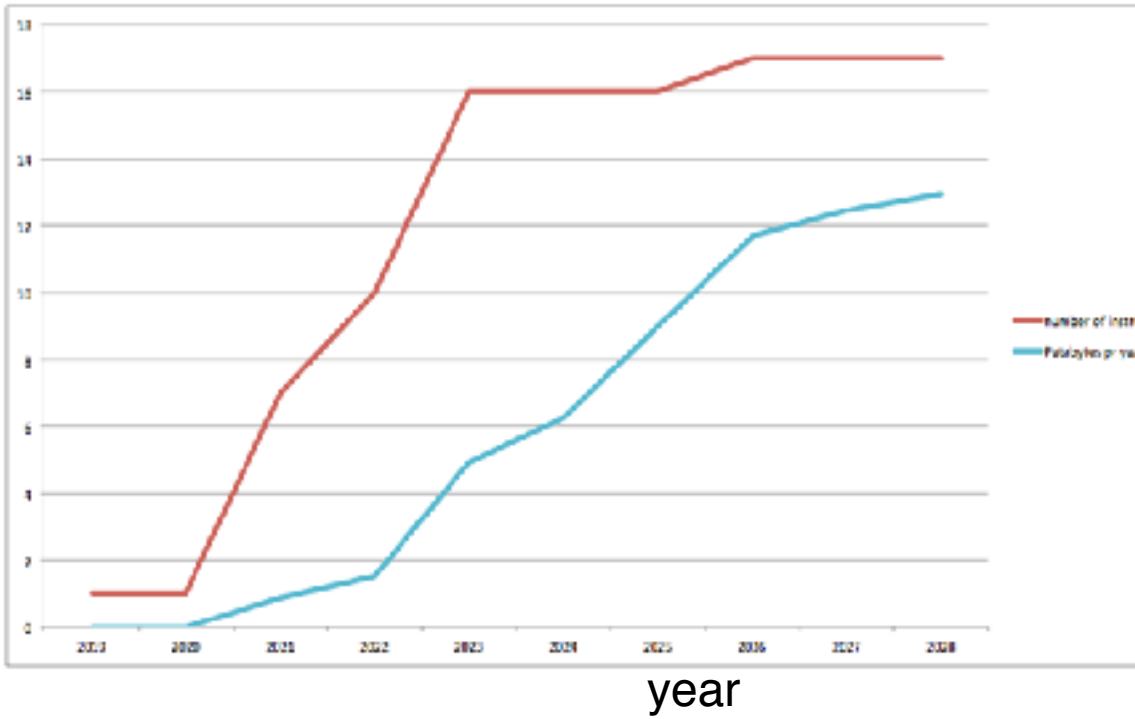


BrightnESS is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 676548

Experiment control



Infrastructure



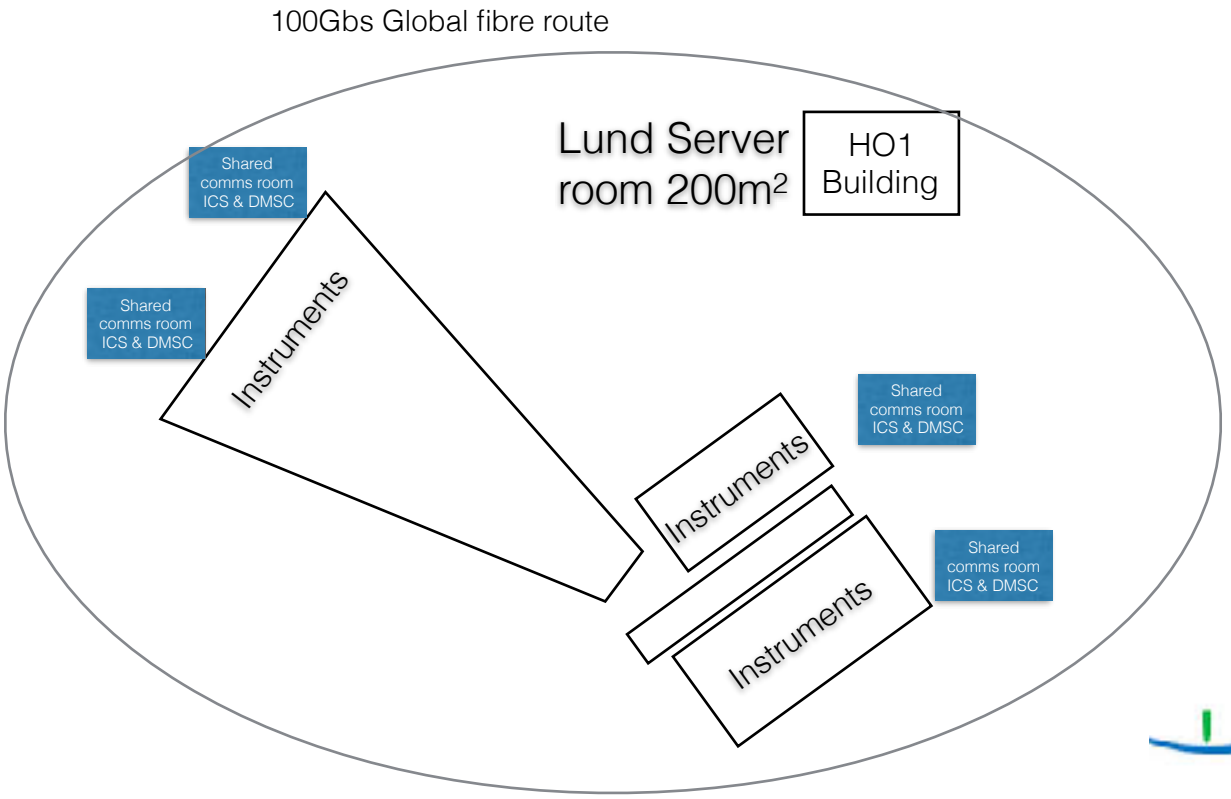
- Focus on UX at all stages of data lifecycle
- PFS in Lund and CPH
- Cluster based compute
- VM infrastructure for reduction and analysis software
- CI development
- Continuous deployment and testing

High availability infrastructure

Complexity metric

- Fast storage
- Fast network
- Data next to compute
- Design into facility

Infrastructure



- Focus on network capacity
  - Dual redundant 100Gbs link from instrument to server room
  - Instrument on virtual network that includes server room infrastructure
- Minimise hardware deployed on instrument

- Detector read out hardware on the instrument
- Experiment control server in H01
- Online data reduction in H01
- Online analysis in H01
- archival storage in CPH
- Offline reduction analysis and modelling in Copenhagen

Copenhagen server room 100m²

ESS data

ESS will generate large data files and larger data sets

Moves back towards a more centralised approach

Good data management & policy

Put the data next to the compute

Moving towards requirement for a more centralised federated storage

Data policy FAIR compliant

- 3 year embargo
- Open afterwards catalogue search on meta data
- stored on fast disk for 5 years
- Archived onto tape.

Data management

- Chose a sensible file format (hdf is good)
- Save the correct data
- Catalogue it
- archive it
- Make it open
- Enforce the policy

NeXus

A common data format for neutron, x-ray and muon science.

Data management II

Not just allowing users access to their data

Automatic reduction and analysis requires efficient data management

Define what the data are

- Sample
- Calibration
- Background

User access to the correct data is important.

- Creates real operational issue is this is poorly implemented

Software development

- Collaborate on projects
- Engage with the community
- Don’t reinvent
- Use the right tools
- Be open

Jenkins

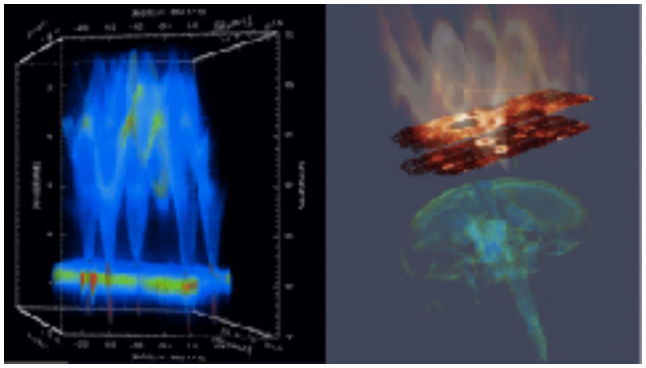
open source

The Mantid Project

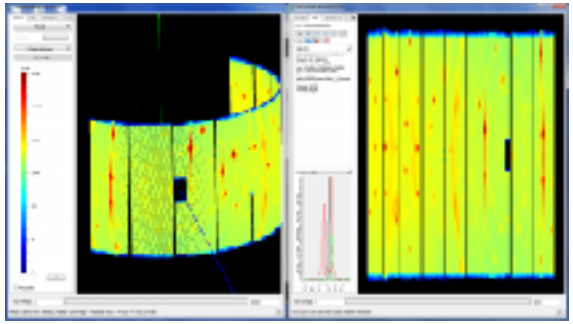
EUROPEAN SPALLATION SOURCE

MANTID

- Neutron specific data treatment framework
- Standardised beyond data format
- Event data capable
- Live view
- Complete instrument geometry
- nD data visualisation
- Data and software curation
- App based UI
- Python interace
- Jupyter notebook
- MPL graphing



ParaView



ISIS

OAK RIDGE National Laboratory

EUROPEAN SPALLATION SOURCE

NEUTRONS FOR SCIENCE

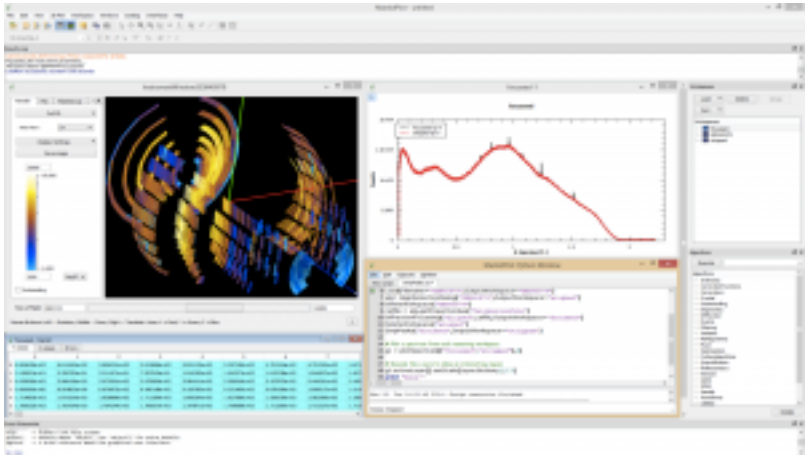
Tessella

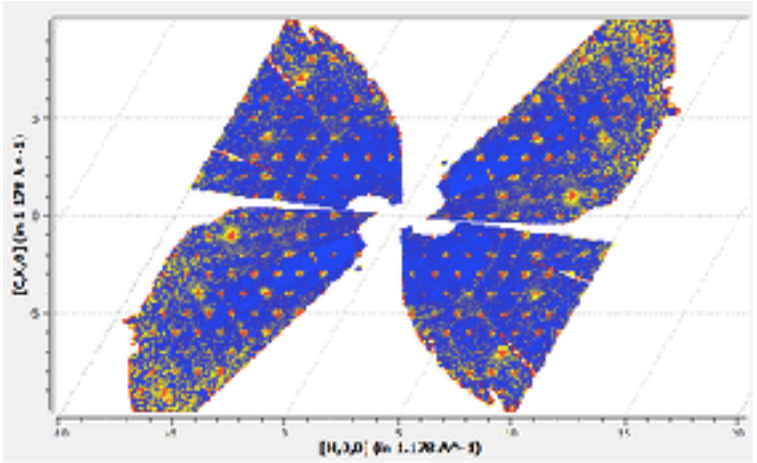
Ansto

PAUL SCHERRER INSTITUT PSI

Mantid project II

- C++ / Python (2&3) / Qt5
- ~35 developers across multiple sites
- single PM (Nick Draper)
- Collaborative governance model
- Provides live reduction & visualisation capability
- Event mode data processing
- Data processing history
- Strongly typed
- nD data types
- Histogram data
- Handles complex instrument geometry
- Feature driven development
- Abstract data processing
- Class and individual instruments interfaces and functionality





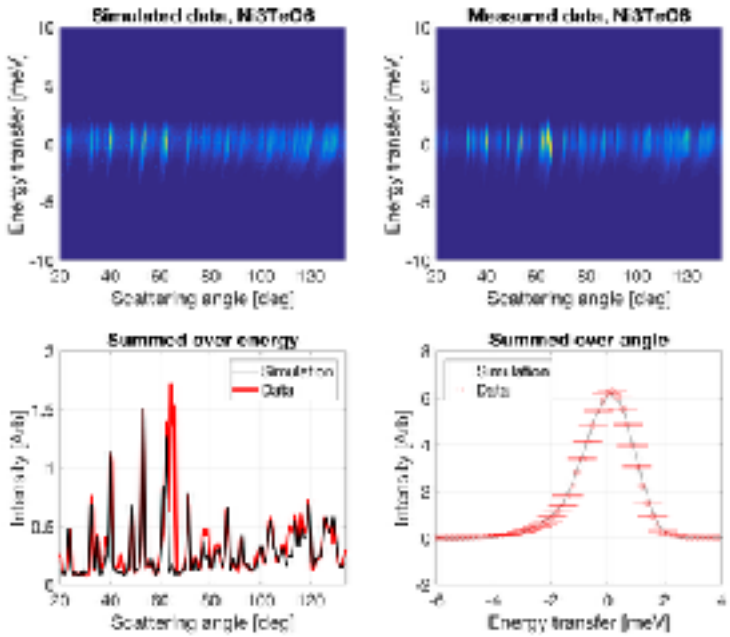
Mantid project Future developments

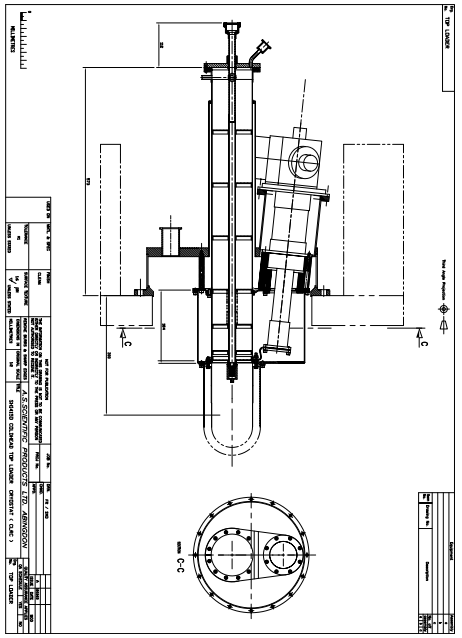
EUROPEAN SPALLATION SOURCE

- MPI compliant framework (ESS requirement)
- UI rewrite for QT5
  - Drop QTIPlot dependency
- Develop pythonic standard module for other projects to utilise
- MPL based plotting
- Greater McStas integration

McStas integration

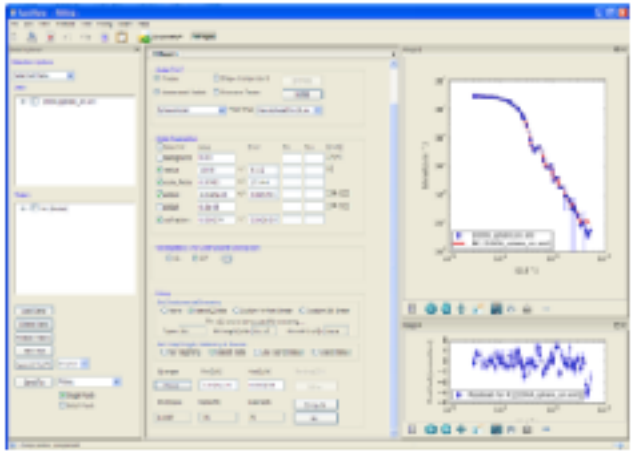
- Mads Bertelsen mads.bertelsen@gmail.com
- McStas Union component - model complex geometries
- Source + instrument + SE +sample
- Sample environment multiple scattering simulation
- Get the resolution and flux for free



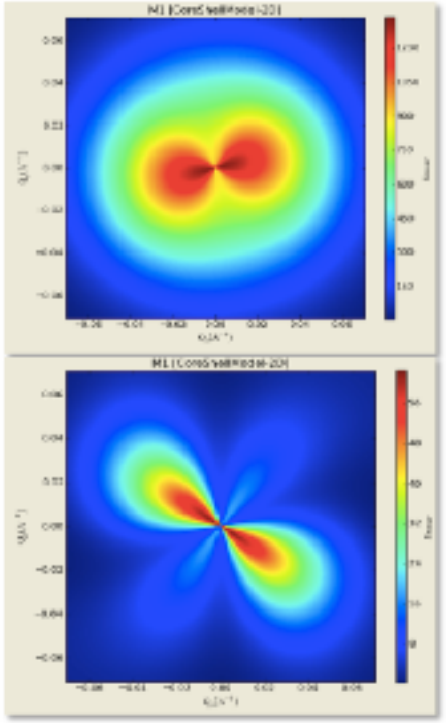


Data analysis SASView

- SASView
- Small angle scattering analysis
- Fitting and visualisation
- Photons and Neutrons



1D fitting screenshot.



CoreShell simulation with Micore parallel to the x-axis and Mishell oriented away. Top figure up,up bottom figure up,down

Data analysis

- Aim for live feedback
- Greater level of automation
  - Define / write APIs and interfaces
- Target codes
  - SASView
  - Fullprof / GSAS
  - Born Again - reflectometry framework
  - SpinW Magnetic excitations
- Community engagement is key

Concluding remarks

- Don’t compromise on systems that are difficult to fix in operations
- Collaborate
- Meta data storage is essential - but how is essentially an open question