Hartree Centre

Science & Technology Facilities Council

Tackling industry challenges with HPC, Big Data and Cognitive Computing

Iain Bethune

Technical Programme Manager

STFC Hartree Centre

iain.bethune@stfc.ac.uk

@iainbethune @HartreeCentre

Who are we?

STFC's high performance computing, data analytics and cognitive technology centre

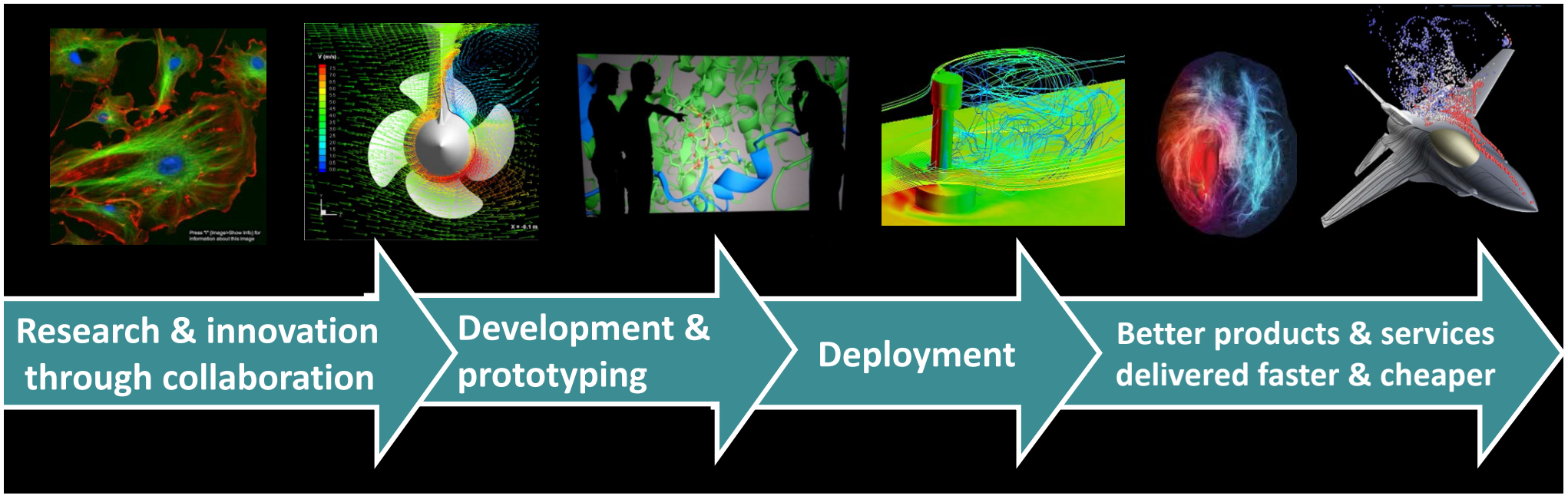
Provides businesses and researchers with access to powerful technologies, facilities and scientific computing expertise

Hartree Centre

Science & Technology Facilities Council

Our mission

To transform UK competitiveness by facilitating and accelerating the adoption of high performance computing, data-centric computing and cognitive technologies.

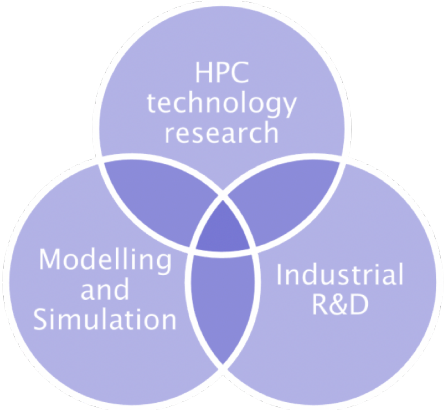


Hartree Centre Clients

Hartree Centre
Science & Technology Facilities Council

A Brief History

- Hartree Centre builds on STFC's long heritage and expertise in High Performance Computing
- 2012 focus on economic impact through software & modelling



- 2015 major investment in collaborative research (IROR)
  - Additional focus on data centric and cognitive computing
  - Embedded IBM Research Centre
  - Extended industrial & scientific reach
- 50 staff + 30 IBMers

Hartree Centre
Science & Technology Facilities Council

Our machines

Intel platforms

Bull Sequana X1000 | 846 SKL nodes + 840 KNL

Lenovo NeXtScale | 8,192 cores

Lenovo System x iDataPlex system | 2048 cores

Intel Xeon Phi | Knight's Corner

IBM Big Data Analytics cluster | 288TB

IBM data centric platforms

IBM Power8 + NVLink + Tesla P100 ”Paragon”

IBM Power8 + Nvidia K80 “Panther”

Energy-efficient & Emerging Tech

Maxeler FPGA system

ARM 64-bit platform

Clustervision novel cooling demonstrator

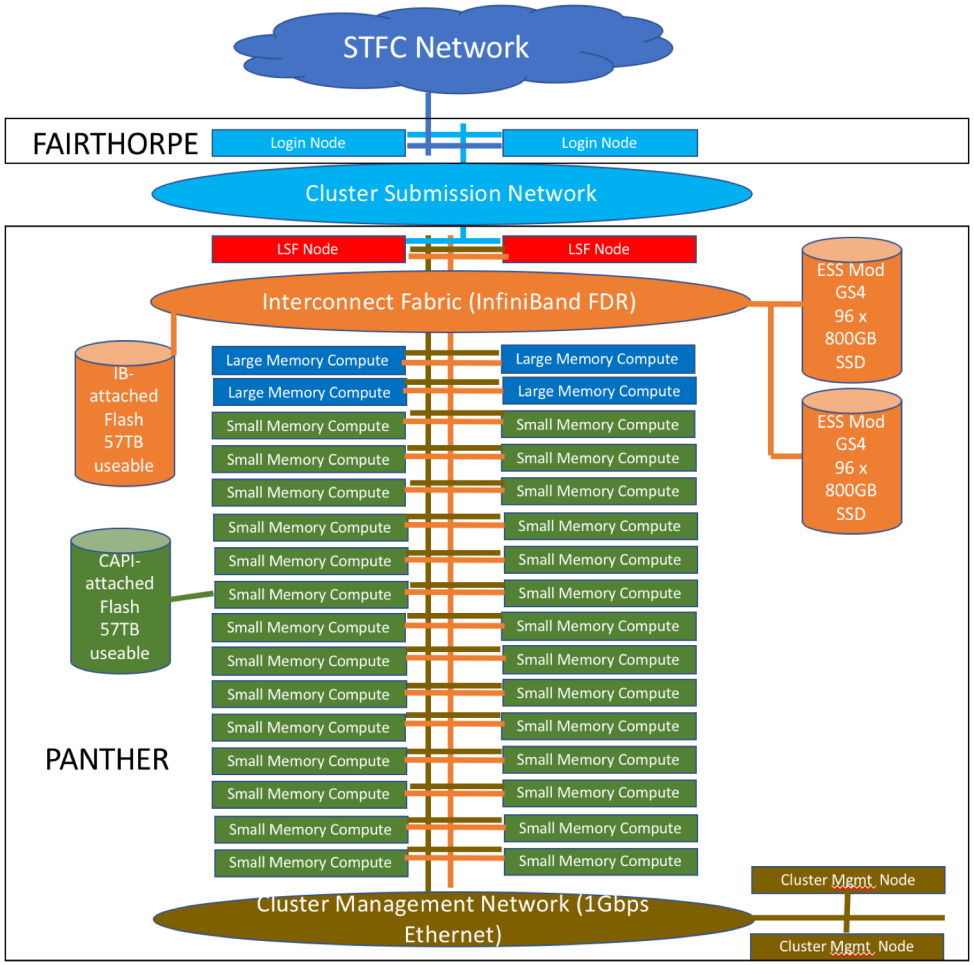
Academic HPC platform (Tier 2)

JADE Nvidia DGX-1

Hartree Centre

Science & Technology Facilities Council

Panther: Current Architecture



SERVER SPECIFICATIONS

Login Node
- S822LC (8335-GTA)
- Bare metal RHEL 7.2 LE
- 16 x 3.3 GHz cores
- 1TB RAM
- 2 x 1TB 7.2k RPM drives
- 2 x nVidia K80 GPUs

LSF Node
- S822LC (8335-GCA)
- Bare metal RHEL 7.2 LE
- 16 x 3.3 GHz cores
- 512GB RAM
- 2 x 1TB 7.2k RPM drives

Large Memory Compute
- S822LC (8335-GTA)
- Bare metal RHEL 7.2 LE
- 16 x 3.3 GHz cores
- 1TB RAM
- 2 x 1TB 7.2k RPM drives
- 2 x nVidia K80 GPUs

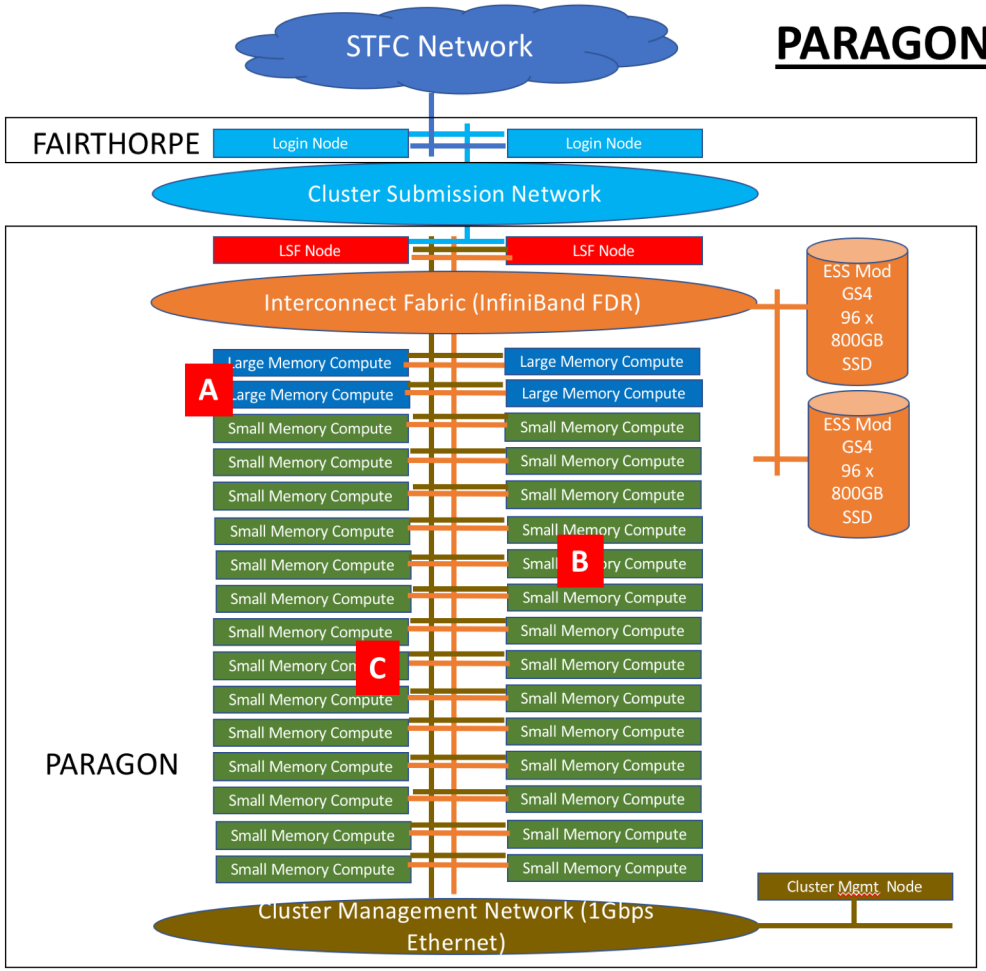
Small Memory Compute
- S822LC (8335-GTA)
- Bare metal RHEL 7.2 LE
- 16 x 3.3 GHz cores
- 512GB RAM
- 2 x 1TB 7.2k RPM drives
- 2 x nVidia K80 GPUs

Cluster Mgmt. Node
- S822LC (8335-GCA)
- Bare metal RHEL 7.2 LE
- 16 x 3.3 GHz cores
- 512GB RAM
- 2 x 1TB 7.2k RPM drives

- 32x compute nodes
- LSF Data Manager – stage to/from CDS
- LSF Batch job scheduler

Hartree Centre
Science & Technology Facilities Council

Paragon: Current Architecture



PARAGON 2017 EXTENSIONS

A 1 x new Large Memory Compute node added (to bring number to 5)

B 7 x new Small Memory Compute nodes added (to bring number to 35)

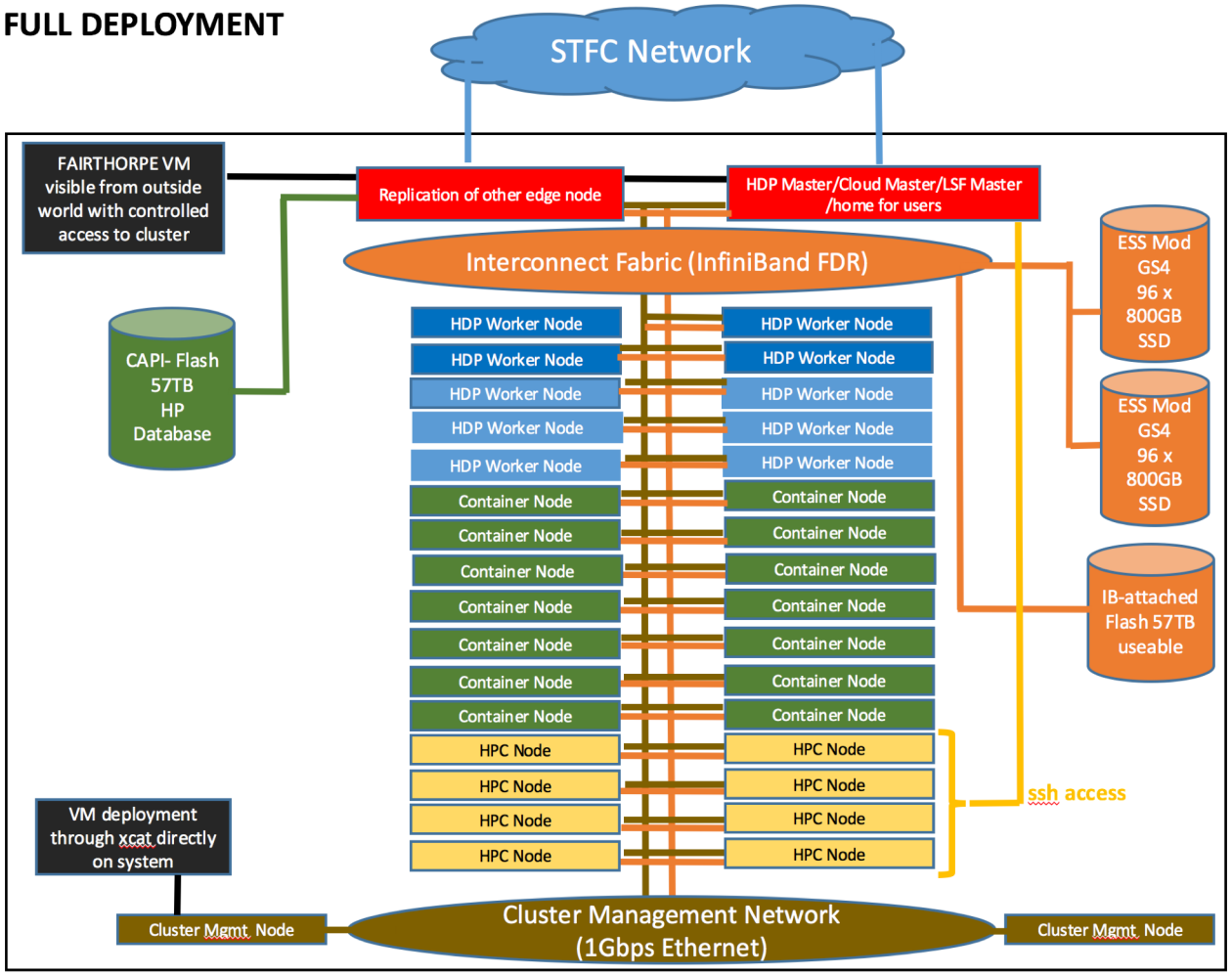
C 3.2TB NVMe PCIe flash added to all Compute Nodes (only included in Large Memory Compute nodes at initial deployment)

- 40x compute nodes
- NVIDIA Tesla P100 GPUs

Hartree Centre
Science & Technology Facilities Council

Towards a converged architecture for HPC, Big Data & Container/cloud workflows

PHASE 2
FULL DEPLOYMENT



- HDFS-GPFS adaptor
- HPC Nodes (LSF)
- Hortonworks Data Platform Worker Nodes (Yarn)
- Container Nodes (Kubernetes / IBM Cloud Private)

- VMs for web-services

Hartree Centre
Science & Technology Facilities Council

IROR: Innovation Return on Research

- £115m investment by BEIS (2015-2020)
- Collaborative R&D programme with IBM Research
- Developing ‘Digital Assets’ to benefit key sectors of UK Economy
  - Chemistry & Materials, Life Sciences, Science/Engineering/Manufacturing
- Projects developed with major UK companies
  - Sectoral approach -> wider benefits

Hartree Centre
Science & Technology Facilities Council

Chemistry

- Parameterisation Engine
- Chemistry Workflow for HPC
- Cognitive Accelerator for Chemistry
- Consumable Computing for Chemistry
- Particle Dynamics Modeller

Life Sciences

- Membrane Binding & Permeation Assay
- Electronically Coarse Grained Biomolecular Simulation Method
- Omics analytics Catalogue
- HPC omics workflows as a service
- Precision Agriculture

Sci,Eng,Man

- Multiscale & Multiphysics Code Coupler
- Cognitive Uncertainty Quantification Toolkit
- Simulation Methods for Engineering Applications
- Intelligent Manufacturing

Enabling Technologies

- Code analysis & generation for HPC
- Cognitive Technical Computing as a Service
- Machine Learning at Scale
- Cognitive Knowledge APIs
- Cognitive Transportation Technology
- Cybersecurity at Scale
- Data Technology Platform

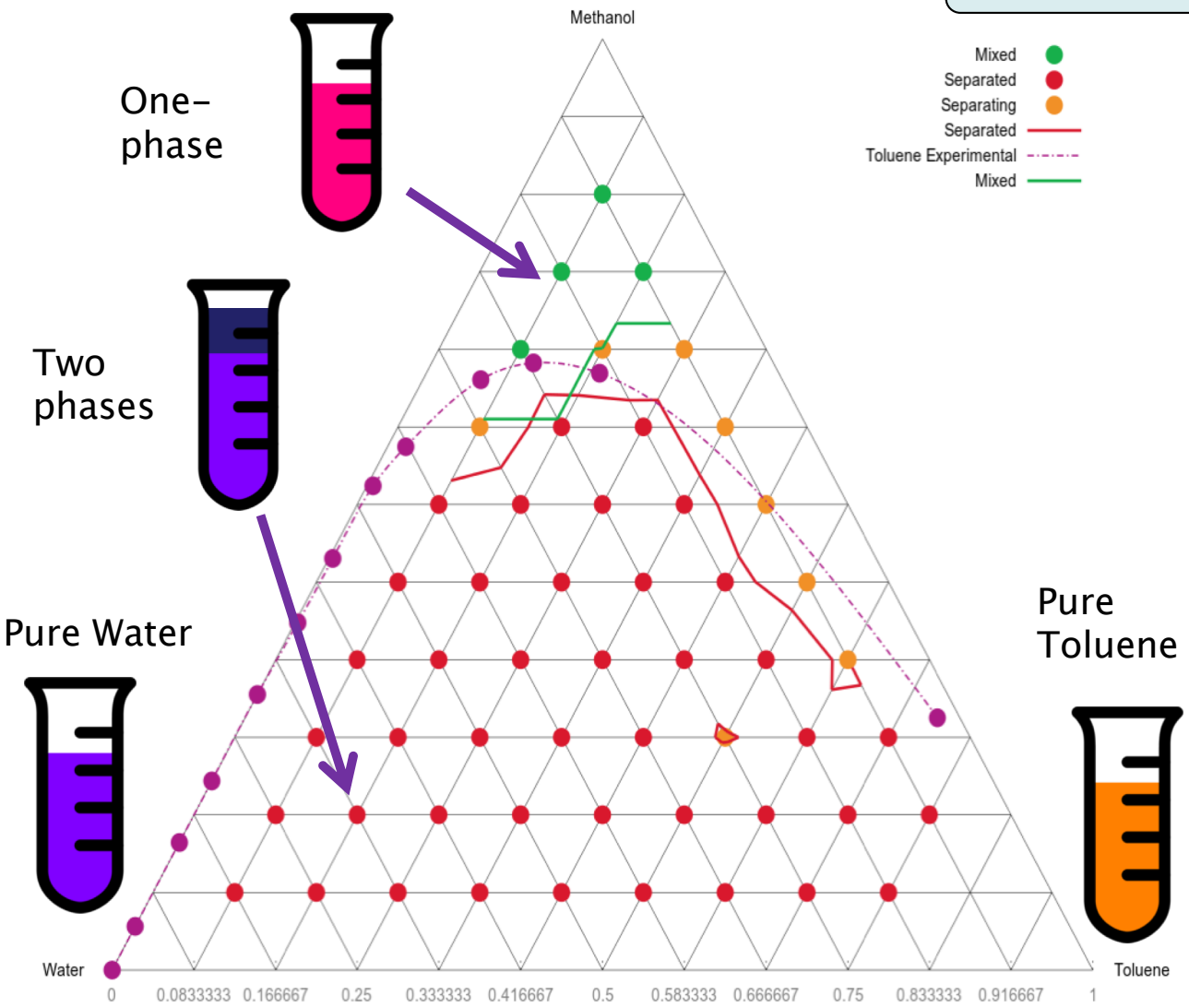

Converting physical experiments to a simulation workflow
Reducing the innovation cycle and time to market
Putting the tools in the hands of the formulation chemist
Ch
L
S
S
E
M
ET
Unilever – Computer Aided Formulation

Hartree Centre
Science & Technology Facilities Council

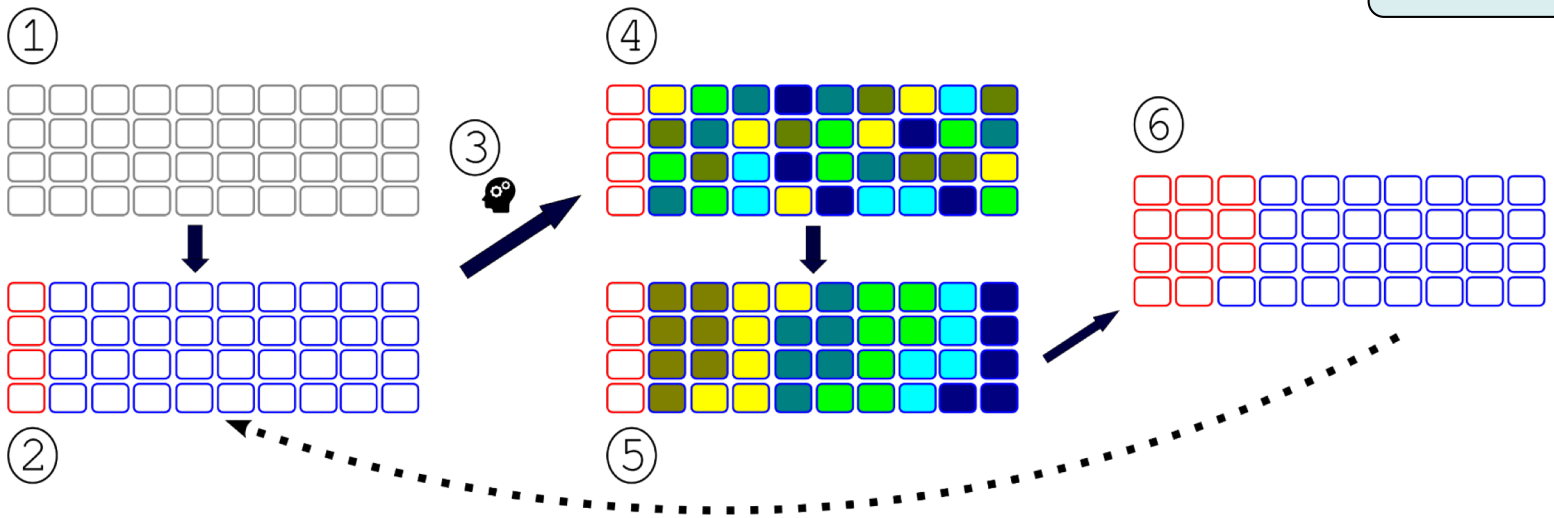
Ch | L S | S E M | ET

Ternary Phase Diagram

- Simulations of solutions of three different molecule types:
  1. Water
  2. Methanol
  3. Toluene
- Solutions will become either phase separate into two liquids or remain mixed



Smarter Simulations by Machine Learning

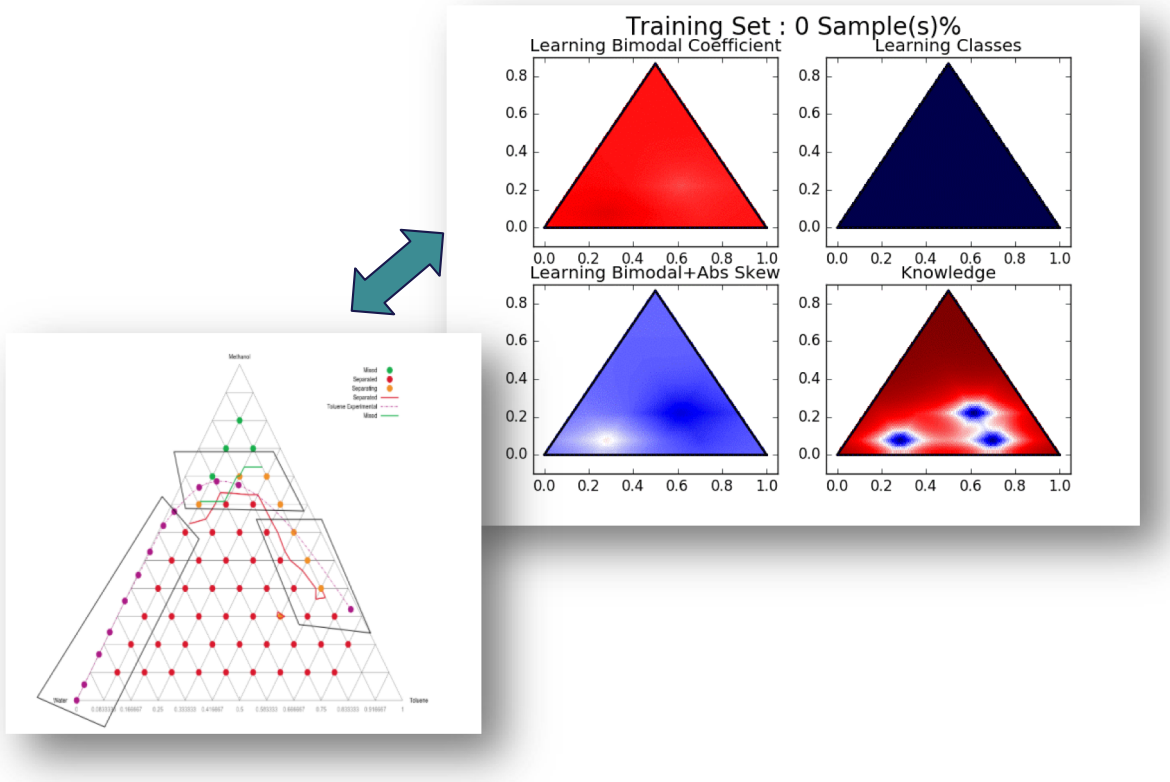


- Speedup attained in two different ways:
  - Choosing the calculations intelligently (**less calculations**)
  - Interpolating between the calculations performed in a sophisticated manner
- Gaussian process used for both tasks
  - Bayesian operation, so it **knows what it doesn't know**
- Maximum entropy sampling used to choose next calculations
  - Minimizes redundant information from calculations
- Non-parametric interpolation increases the resolution by 6 orders of magnitude

Getting HPC to “work smart, not hard”

Ch L S S E M ET

- Typically HPC development is focused on increased speed.
  - The fastest calculation is the one which you don’t run!
- We can use machine learning to make better decisions on **which simulations give the most value**
- We can use machine learning to **improve resolution of information**



Cognitive workflow uses 1/3 of the calculations to achieve 4 orders of magnitude resolution increase

Hartree Centre
Science & Technology Facilities Council

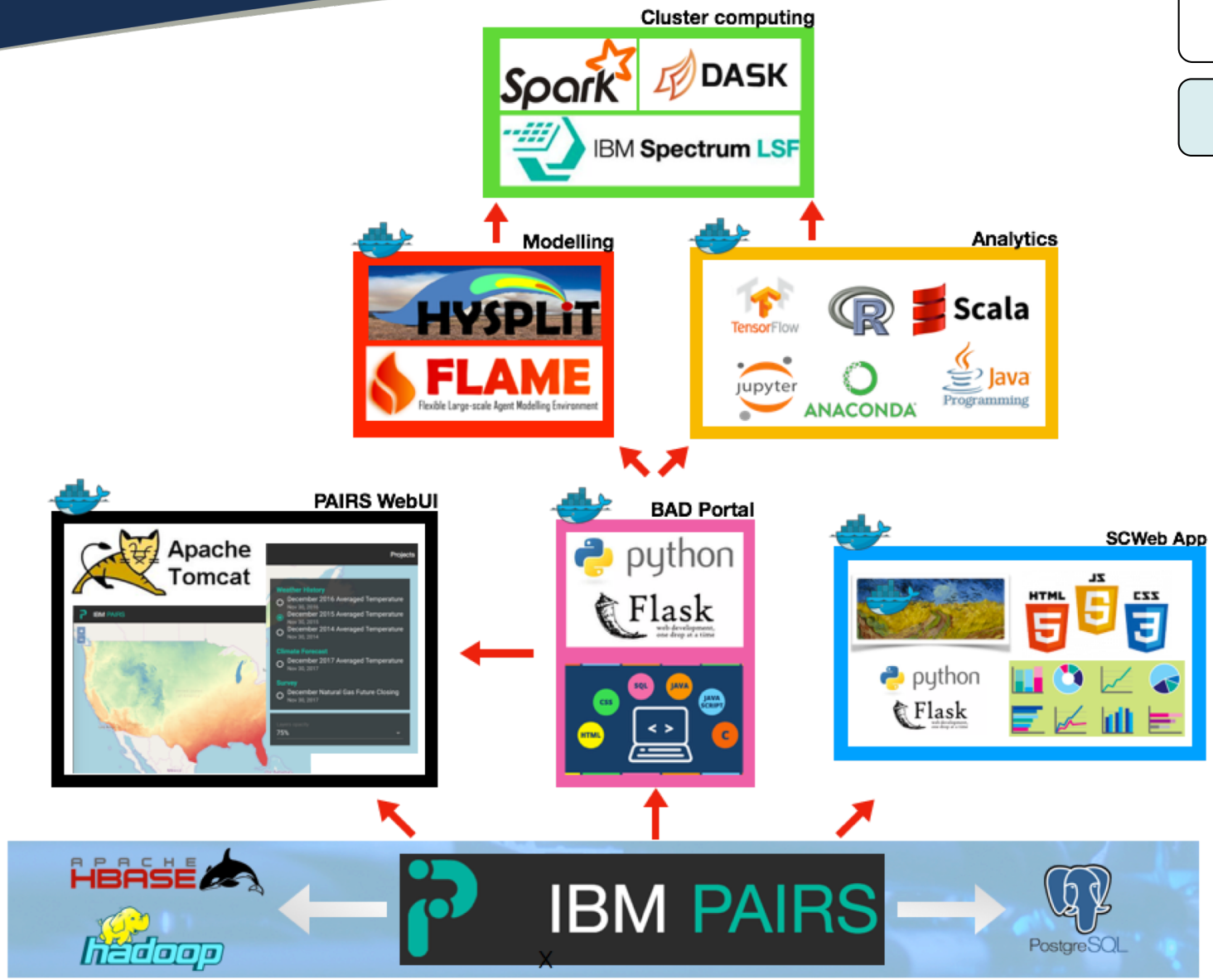
Ch | L S | S E M | ET

Smart Crop Protection using Big Agricultural Data

- Large quantities of geospatial data exist
  - Public (open data) / private
  - Static / dynamic
  - Raster / vector
- Can we store, access and query the data?
- Can we use the data for agriculture applications?
- PoC with Rothamsted Research
  - Moth migration

30% crop loss worldwide due to pests, pathogens and weeds!

ROTHAMSTED RESEARCH 175


Cluster computing
Spark
DASK
IBM Spectrum LSF
Modelling
HYSPLIT
FLAME
Flexible Large-scale Agent Modelling Environment
Analytics
TensorFlow
Scala
jupyter
ANACONDA
Java Programming
PAIRS WebUI
Apache Tomcat
BAD Portal
python
Flask
SCWeb App
HTML
python
Flask
APACHE HBASE
hadoop
IBM PAIRS
PostgreSQL

Hartree Centre
Science & Technology Facilities Council

Ch | L S | S E M | ET

IBM PAIRS geospatial database platform

Hartree Centre
Science & Technology Facilities Council

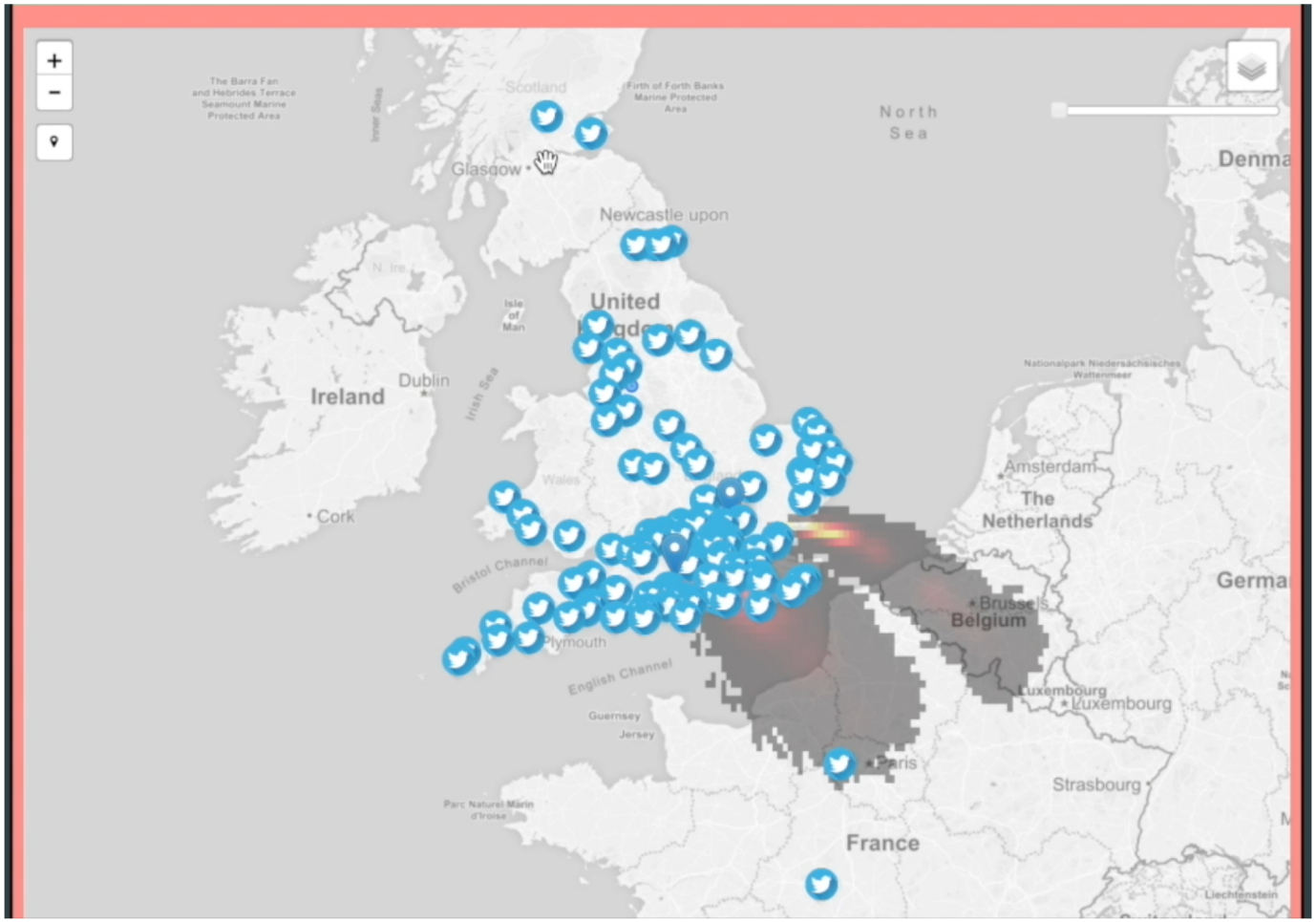
Ch | L S | S E M | ET

‘SCApp’ web app for farmers

Hartree Centre
Science & Technology Facilities Council

Ch | L S | S E M | ET

Agent-based modelling based on weather and Twitter-sourced sighting data



Hartree Centre
Science & Technology Facilities Council

Ch L S S E M ET

Jupyter (+Spark) analytics portal

jupyter Demo_LandSatelliteNDVI Last Checkpoint: Last Tuesday at 1:14 PM (autosaved)

Logout

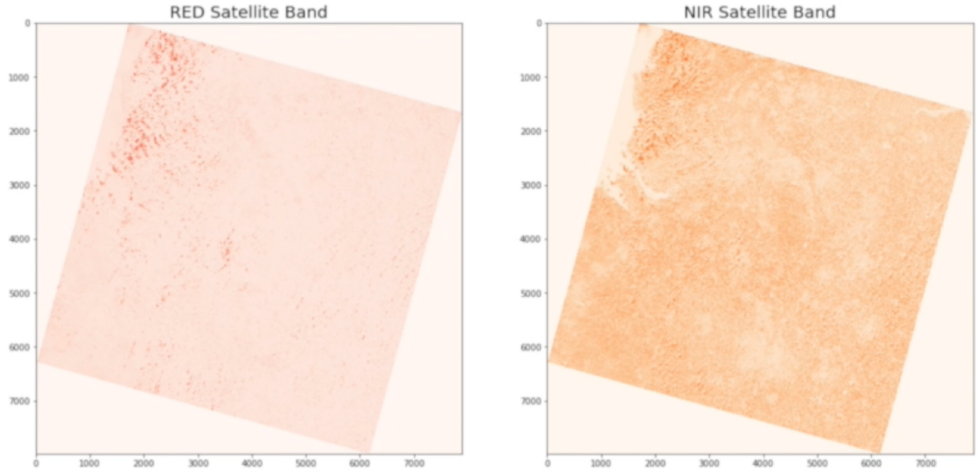
File Edit View Insert Cell Kernel Help Trusted Python 2

Code

1. NDVI Calculation from RED and NIR

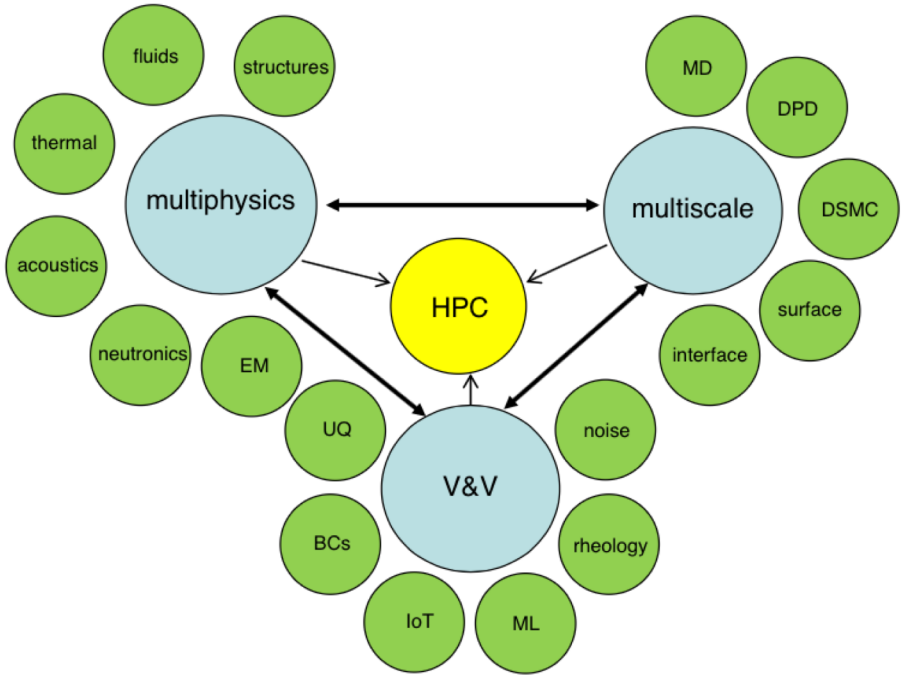
We first need to import the RED and NIR layer and convert the Geotiff into a numpy array. We can either import the data directly or get it from PAIRS instead.

In [8]:
```
#Read layers in from raw files:

def read_dat_layers(path,band):
    file_location = path+"LC08_L1TP_203023_20160504_20170325_01_T1_"+band+".TIF"
    ds = gdal.Open(file_location)
    return np.array(ds.GetRasterBand(1).ReadAsArray())

bandRed = read_dat_layers(path="ndvi/",band="B4")
bandNIR = read_dat_layers(path="ndvi/",band="B5")

#Explore what data looks like:
#print bandRed
plt.figure(figsize=(20,10))
plt.subplot(1, 2, 1)
plt.imshow(bandRed,cmap="Reds")
plt.title('RED Satellite Band',fontsize=20)
plt.subplot(1, 2, 2)
plt.imshow(bandNIR,cmap='Oranges')
plt.title('NIR Satellite Band',fontsize=20)
plt.show()
```



Virtual Product Design

- a.k.a. Digital Twinning
- Full product lifecycle
  - Design
  - Manufacture
  - Operation
  - Disposal

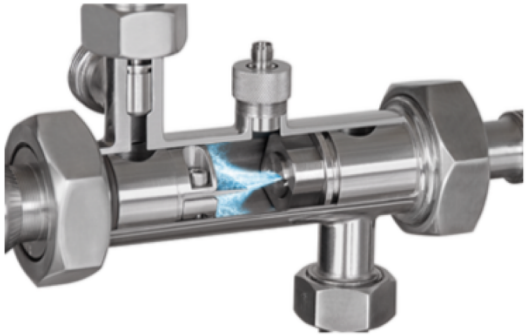


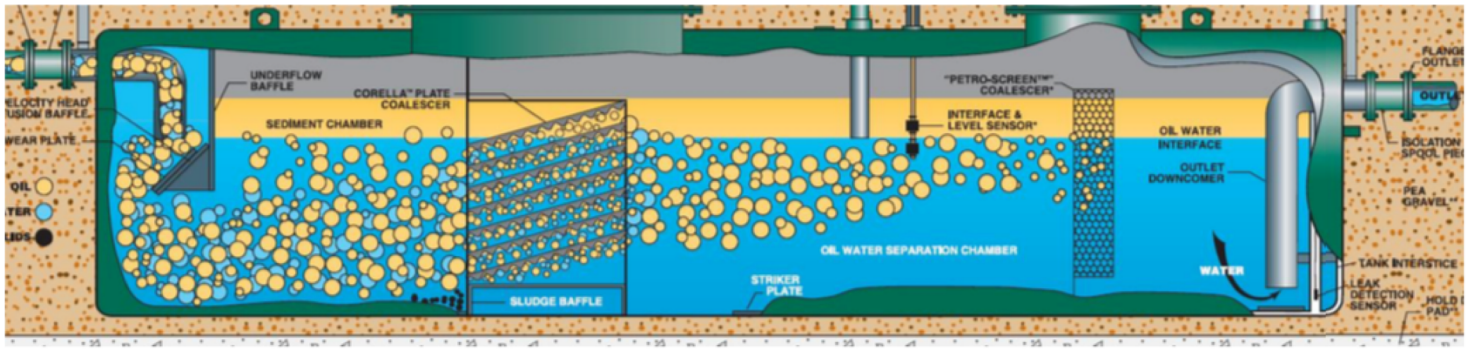
- Requires complex physics and understanding limitations

Mixing in industrial processes

Chemical engineering applications:

- oil and gas,
- food industry,
- personal-care products,
- paint and coating industry,
- effluent dispersion in wastewater.

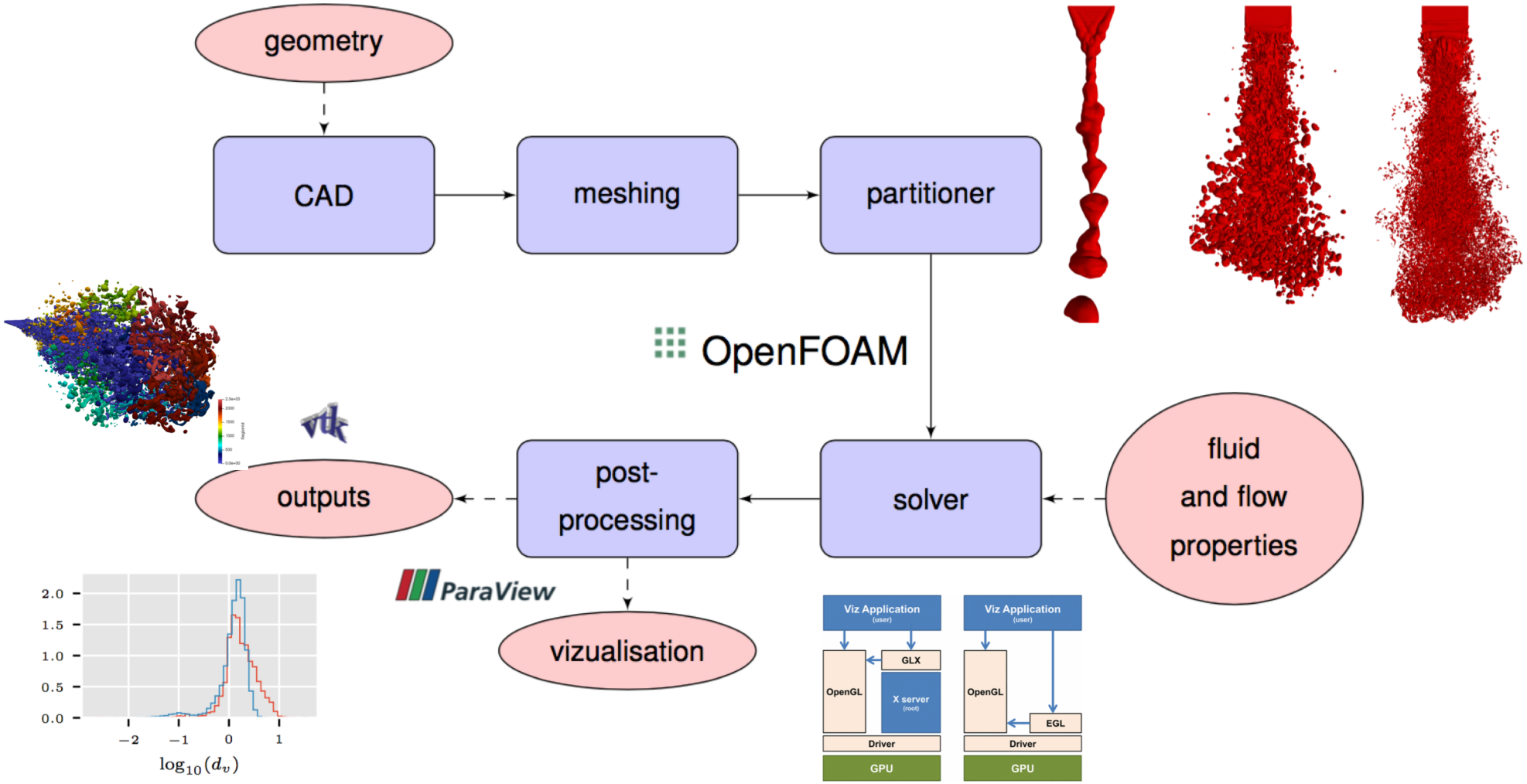




Hartree Centre
Science & Technology Facilities Council

Ch L S S E M ET

Classical HPC CFD workload



Thanks for listening!

Acknowledgments:
STFC & IBM Research Staff