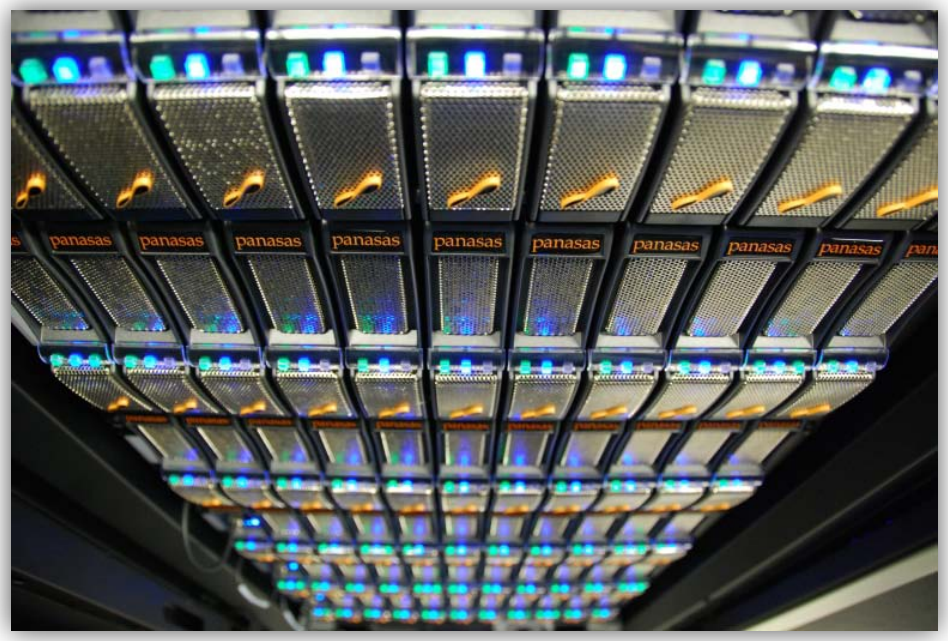
Science & Technology Facilities Council

CEDA Storage



Dr Matt Pritchard

Centre for Environmental Data Archival (CEDA)

www.ceda.ac.uk

- How we store our data

- NAS Technology
- Backup
- JASMIN/CEMS

CEDA Storage

Data stored as files on disk.

Data is migrated from media to media (early data will have moved six time)

Data is audited to make sure there is no silent corruption.

Storage Classes

- Primary : main archive copy of data - “Crown jewels”
- Secondary : 2nd copy of data - CEDA not primary copy
- Facilitative - CEDA merely helps redistribute data

Storage structure

- Logical path : an interface for users & services (this is the reference path for users)
  - /badc/N/X
  - /badc/N/Y
  - /neodc/M/Z
    - Filesets e.g. X exist within datasets e.g. N
      - Break up datasets into manageable chunks for backup etc.

- Physical path : unseen by users, the “real” path to the data
  - /archive/X
  - /archive/Y
  - /archive/Z

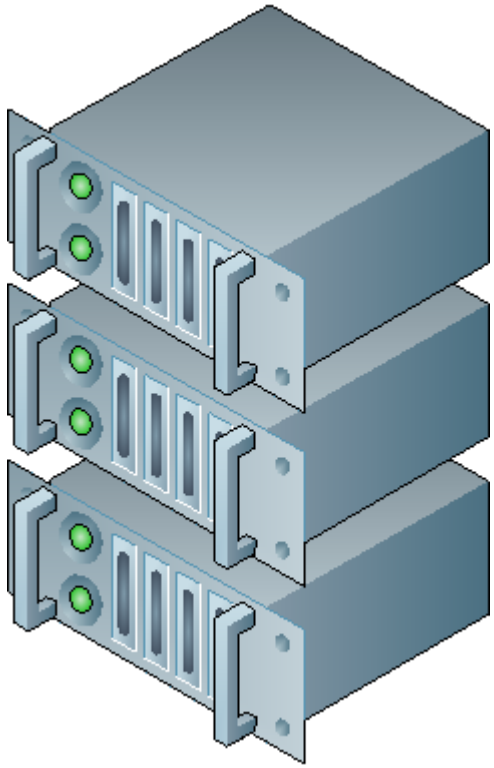
  Connected to logical path via symlinks in filesystem

Storage : Now

foo.badc.rl.ac.uk (NAS server)
/disks/foo1 (30 Tb)
/disks/foo2 (30 Tb)

bar.badc.rl.ac.uk (NAS server)
/disks/bar1 (20 Tb)
/disks/bar2 (20 Tb)

baz.badc.rl.ac.uk (NAS server)
/disks/baz1 (10 Tb)
/disks/baz2 (10 Tb)



System of symlnks and mountpoints builds virtual filesystem.

Systems hosting data access services mount all storage filesystems in order to “see” entire structure.

Scaled to < 1 Pb, but difficult to manage so many individual storage servers & mounted filesystems.

Centre for Environmental Data Archival

CEDA Data

Project	Type	Current volume (Tb)
NEODC	Earth Observation	300
BADC	Atmospheric Science	350
CMIP5	Climate Model	350
	Total	1000 Tb = 1 Pb

Backup

StorageD

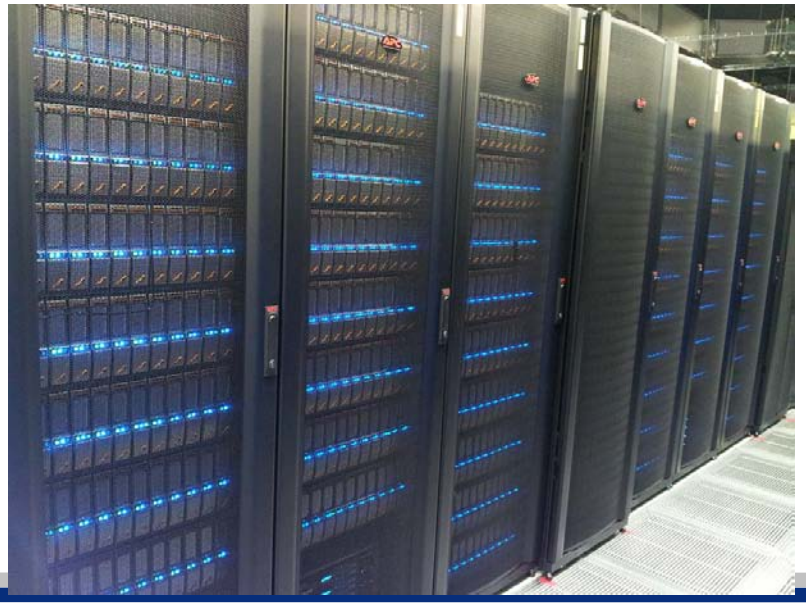
- Tape-based backup storage solution provided by STFC e-science centre
  - Filesets marked for backup
  - Rsynced to StorageD cache
  - Written to tape
  - Secondary tape copy made & kept off site

Secondary online storage

- Some datasets mirrored using rsync to secondary online storage (for rapid recovery)

Storage : JASMIN/CEMS

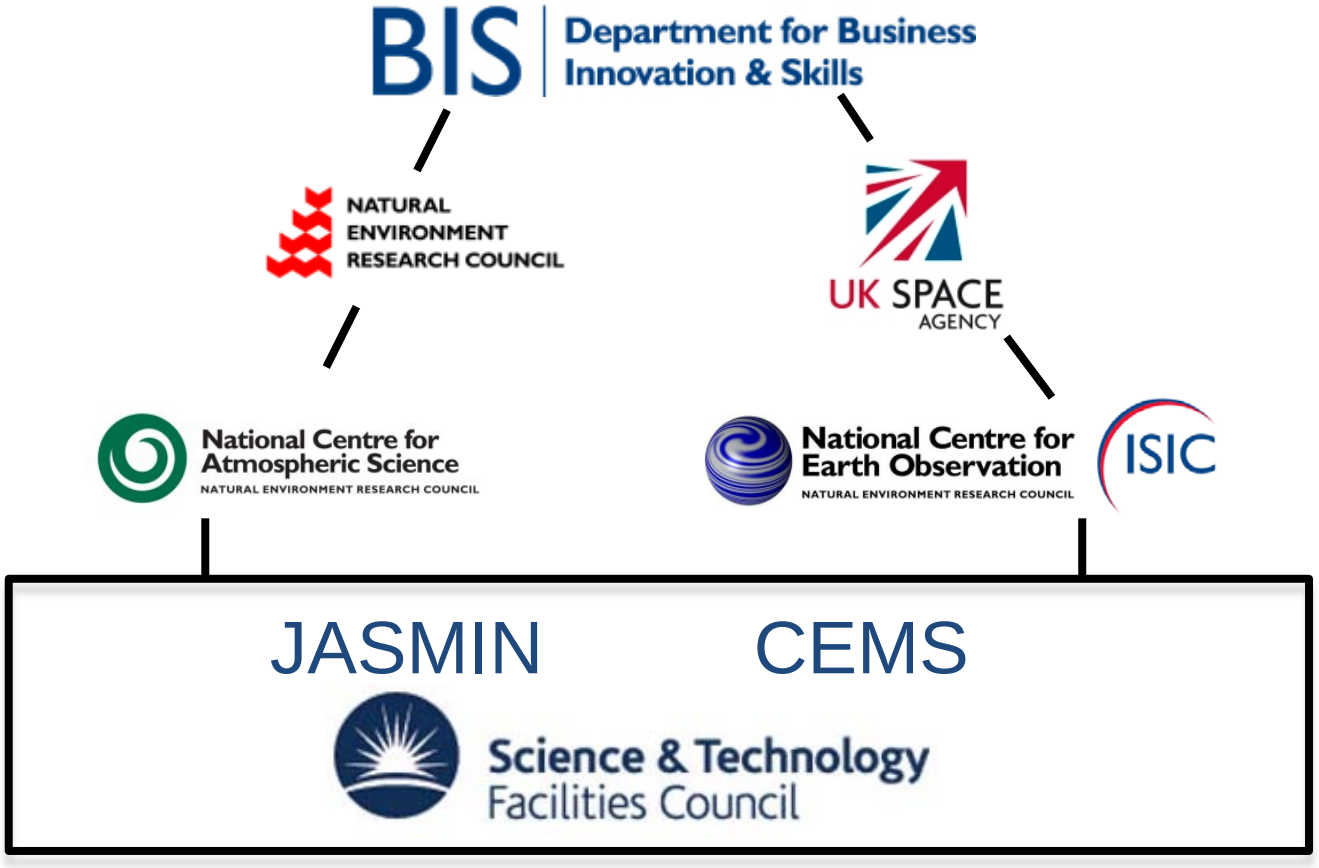
- Storage blades arranged into bladesets (with 1+ director blade)
- Director blades respond to data request (share out load among cluster)
- Parallel access, high bandwidth



- Single namespace can appear as one huge filesystem
- Reality : break up into logical chunks, expandable into free space
- Vastly reduced number of filesystems to mount!

e-Infrastructure

e-Infrastructure Investment



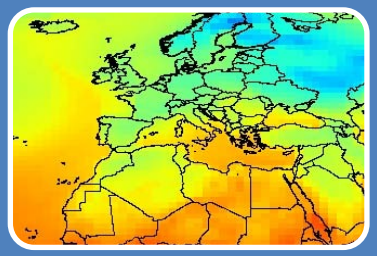
JASMIN/CEMS Data

Project	JASMIN	CEMS
NEODC Current		300
BADC Current	350	
CMIP5 Current	350	
CEDA Expansion	200	200
CMIP5 Expansion	800	300
CORDEX	300	
MONSooN Shared Data	400	
Other HPC Shared Data	600	
User Scratch	500	300
Totals	3500 Tb	1100 Tb

JASMIN functions

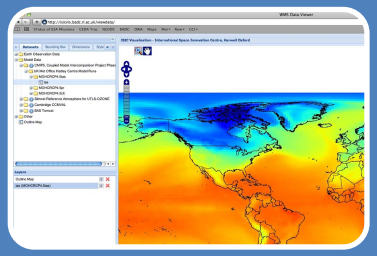
CEDA data storage & services

- Curated data archive
- Archive management services
- Archive access services (HTTP, FTP, Helpdesk, ...)



Data intensive scientific computing

- Global / regional datasets & models
- High spatial, temporal resolution
- Private cloud

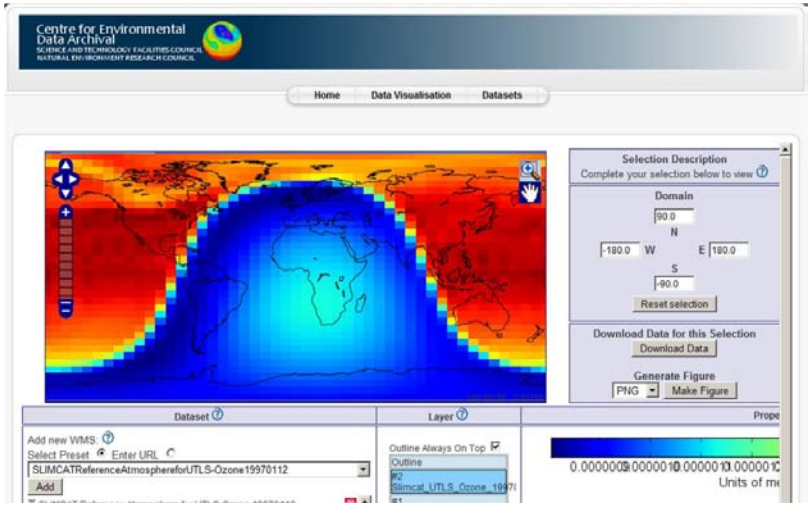


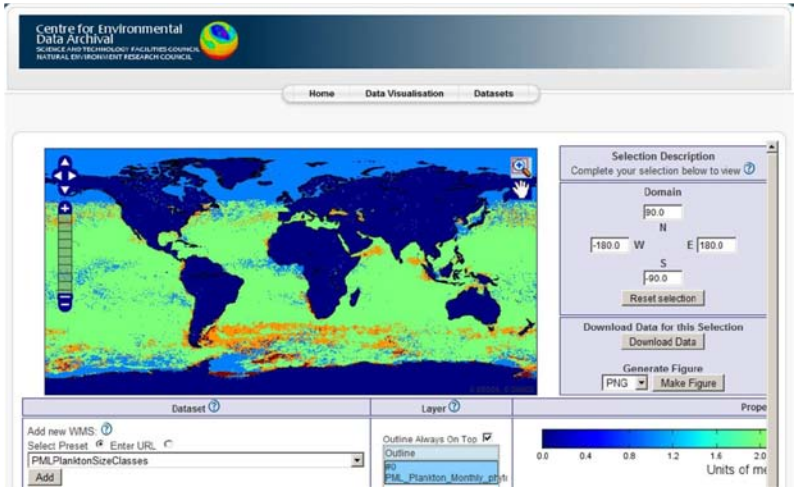
Flexible access to high-volume & complex data for climate & earth observation communities

- Online workspaces
- Services for sharing & collaboration

Use cases

- Processing large volume EO datasets to produce:
  - Essential Climate Variables
  - Long term global climate-quality datasets
- EO data validation & intercomparisons
  - Evaluation of models relying on the required datasets (EO datasets & in situ) and simulations) being in the same place



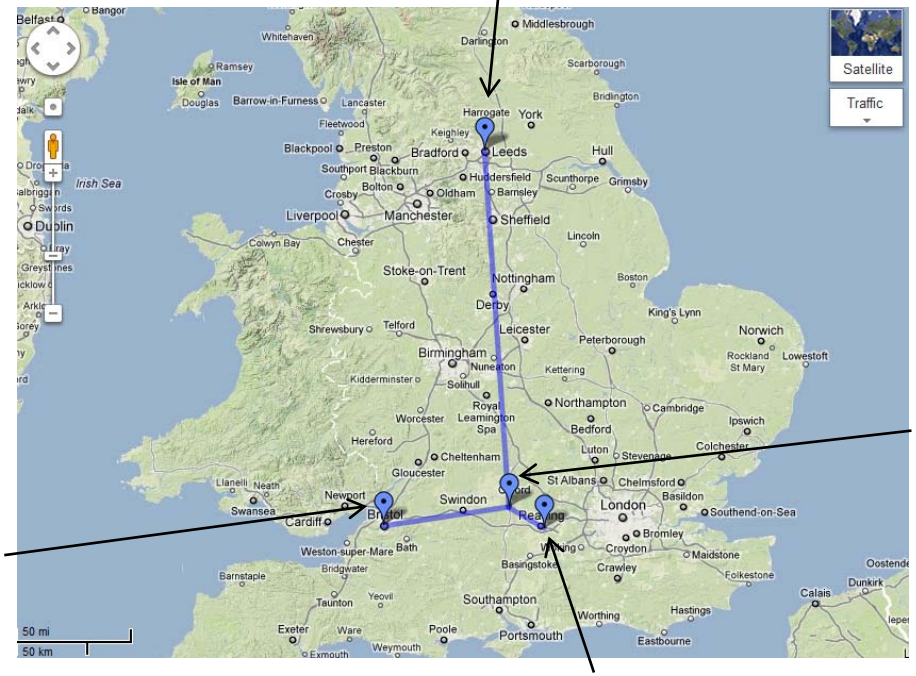


Use cases

- User access to 5th Coupled Model Intercomparison Project (CMIP5)
  - Large volumes of data from best climate models
  - Greater throughput required
- Large model analysis facility
  - Workspaces for scientific users. Climate modellers need 100s of Tb of disk space, with high-speed connectivity
    - UPSCALE project
      - 250 Tb in 1 year
      - PRACE supercomputing facility in Germany (HERMIT)
      - Being shipped to RAL at present
      - To be analysed by Met Office as soon as available
      - Deployment of VMs running custom scientific software, co-located with data
      - Outputs migrated to long term archive (BADC)

JASMIN locations

JASMIN-North
University of Leeds
150 Tb

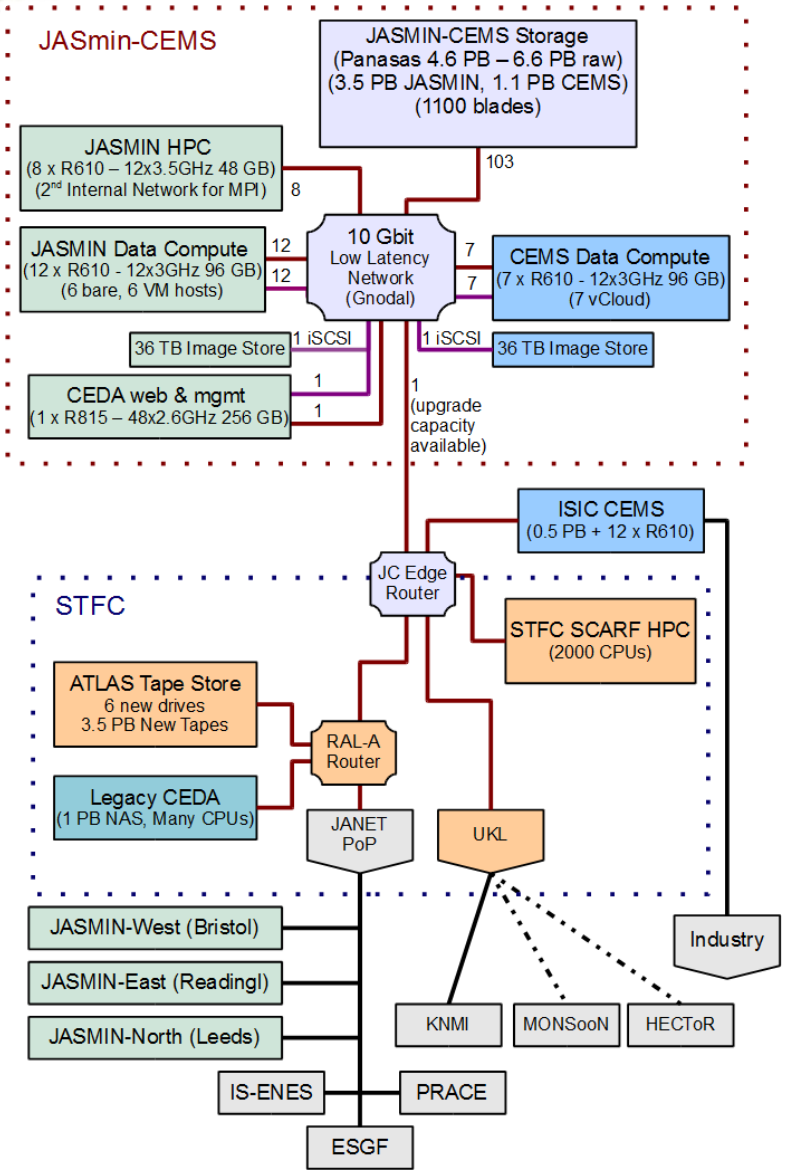


JASMIN-West
University of Bristol
150 Tb

JASMIN-Core
STFC RAL
3.5 Pb + compute

JASMIN-South
University of Reading
500 Tb + compute

JASMIN kit


JASmin-CEMS
JASMIN-CEMS Storage
(Panasas 4.6 PB – 6.6 PB raw)
(3.5 PB JASMIN, 1.1 PB CEMS)
(1100 blades)
JASMIN HPC
(8 x R610 – 12x3.5GHz 48 GB)
(2nd Internal Network for MPI)
8
103
JASMIN Data Compute
(12 x R610 - 12x3GHz 96 GB)
(6 bare, 6 VM hosts)
12
12
10 Gbit
Low Latency
Network
(Gnodal)
7
7
CEMS Data Compute
(7 x R610 - 12x3GHz 96 GB)
(7 vCloud)
36 TB Image Store
1 iSCSI
1 iSCSI
36 TB Image Store
CEDA web & mgmt
(1 x R815 – 48x2.6GHz 256 GB)
1
1
1
(upgrade
capacity
available)
ISIC CEMS
(0.5 PB + 12 x R610)
JC Edge
Router
STFC
STFC SCARF HPC
(2000 CPUs)
ATLAS Tape Store
6 new drives
3.5 PB New Tapes
RAL-A
Router
Legacy CEDA
(1 PB NAS, Many CPUs)
JANET
PoP
UKL
JASMIN-West (Bristol)
JASMIN-East (Readingl)
JASMIN-North (Leeds)
Industry
KNMI
MONSooN
HECToR
IS-ENES
PRACE
ESGF

JASMIN kit

JASMIN/CEMS Facts and figures

- *JASMIN:*
  - *3.5 Petabytes Panasas Storage*
  - *12 x Dell R610 (12 core, 3.0GHz, 96G RAM)Servers*
  - *1 x Dell R815 (48 core, 2.2GHz, 128G RAM)Servers*
  - *1 x Dell Equalogic R6510E (48 TB iSCSI VMware VM image store)*
  - *VMWare vSphere Center*
  - *8 x Dell R610 (12 core, 3.5GHz, 48G RAM) Servers*
  - *1 x Force10 S4810P 10GbE Storage Aggregation Switch*
  - 4 x Gnodal GS4008 10/40Gbe switched stack

JASMIN kit

JASMIN/CEMS Facts and figures

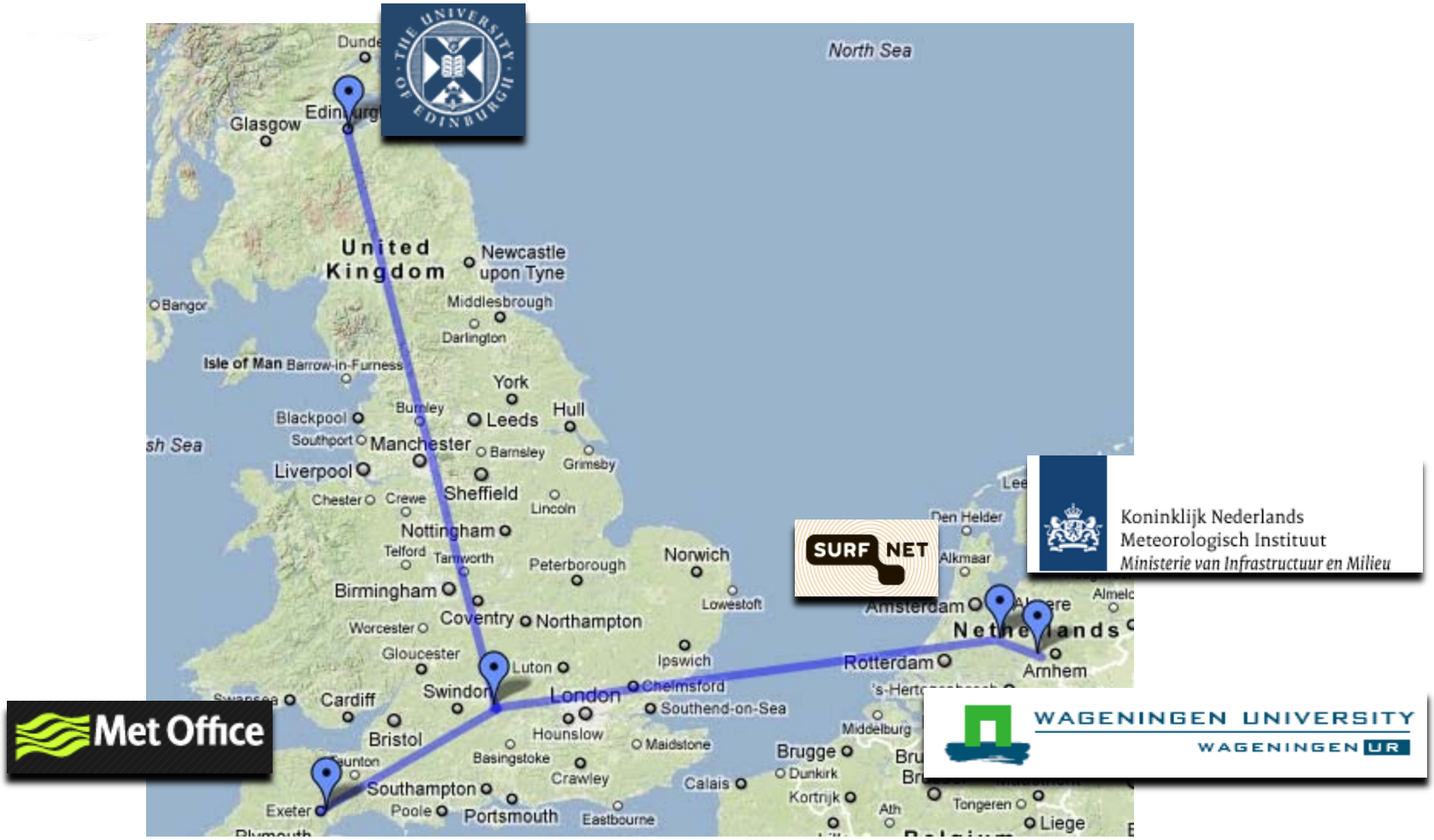
- *CEMS:*
  - *1.1 Petabytes Panasas Storage*
  - *7 x Dell R610 (12 core 96G RAM) Servers*
  - *1 x Dell Equalogic R6510E (48 TB iSCSI VMware VM image store)*
  - *VMWare vSphere Center + vCloud Director*

JASMIN kit

JASMIN/CEMS Facts and figures

- *Complete 4.5 PB (usable - 6.6PB raw) Panasas storage managed as one store, consisting of:*
  - *103 4U “Shelves” of 11 “Storage Blades”*
  - *1,133 (-29) “Storage Blades” with 2x 3TB drives each*
  - *2,266 3.5" Disc Drives (3TB Each)*
  - *103 * 11 * 1 -29 = 1,104 CPUs (Celeron 1.33GHz CPU w. 4GB RAM)*
  - *29 “Director Blades” with Dual Core Xeon 1.73GHz w.8GB RAM)*
  - *15 kW Power in / heat out per rack = 180 kW* ***(10-20 houses worth)***
  - *600kg per rack = 7.2 Tonnes*
  - ***1.03 Tb/s total storage bandwidth = Copying 1500 DVDs per minute***
  - ***4.6PB Useable == 920,000 DVD's = a 1.47 km high tower of DVDs***
  - *4.6PB Useable == 7,077,000 CDs = a 11.3 km high tower of CDs*

JASMIN links



http://www.ceda.ac.uk

http://www.stfc.ac.uk/e-Science/38663.aspx

Thank you!