ICR The Institute of Cancer Research

Research Data Storage Platform

robert.horton@icr.ac.uk

The Royal Marsden Hospital

Founded in 1851 as the Free Cancer Hospital to Classify Tumours, research the causes and find new treatments.
ICR Became a separate entity in 1909.

Institute of Cancer Research - at a glance

Top 4 global cancer research organisation

Top-ranked UK academic institution (REF)

20 drug candidates discovered since 2005

More than 1,000 staff

£161.9m income
£110.0m expenditure

Athena SWAN Silver Award

Awarded Athena SWAN Silver

More than 900 scientific papers

Partnerships with 163 different companies

Top UK university for invention income

141 research students
143 MSc students

Making the discoveries

Our strategy to defeat cancer

The ICR and The Royal Marsden delivered a joint strategy covering the next five years

Our vision
We will overcome the challenges posed by cancer's complexity, adaptability and evolution through scientific and clinical excellence, innovation and partnership

First pillar: Unravelling cancer's complexity

Comprehend the full complexity of cancer by **harnessing the power of new technologies and Big data**

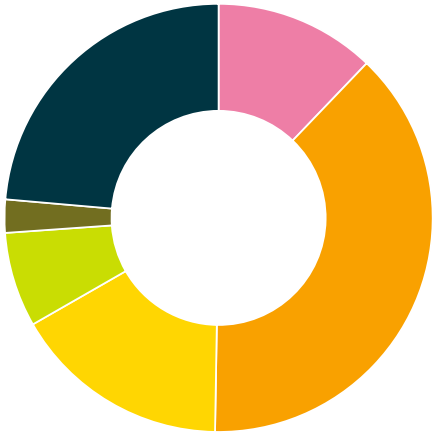
Substantial and diverse funding

2016

- Total income £162m
- HEFCE 12% based on research excellence
- Grant income 38%
- Legacies and donations 7%
- Invention income from our discoveries 16%

-> Services not centrally funded

Total incoming resources 2016



12%
Higher Education Funding Council for England

38%
Research grants

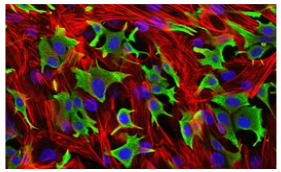
16%
Royalty income

7%
Legacies and donations

3%
Investment and tuition fees

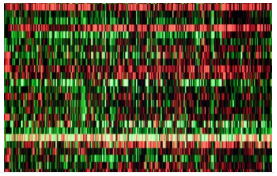
24%
Sale of part of our future royalty stream

ICR Research Divisions



Breast Cancer Research

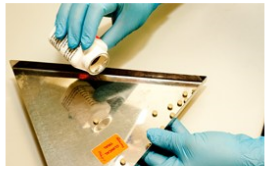
Cancer Therapeutics



Genetics and Epidemiology

Radiotherapy and Imaging

Cancer Biology



Clinical Studies

Molecular Pathology

Structural Biology

Scientific Computing Service

- HPC
  - ~2000 cores, 2PB Lustre storage
- Storage
  - RDS Platform
  - Future: Archive, Data Transfer Service
- Scientific Software
- Edge / User support
  - Instrument connectivity, Staging areas etc

History of Data Storage

- 2006 – Pandorica
  - SGI CXFS 3 Tier system based on 2 sites
- 2011 – Gallifrey
  - Quantum StorNext 2 Tier (Disk + tape) system
  - Issues caused by retrieving small files from tape
- 2016 – RDS…

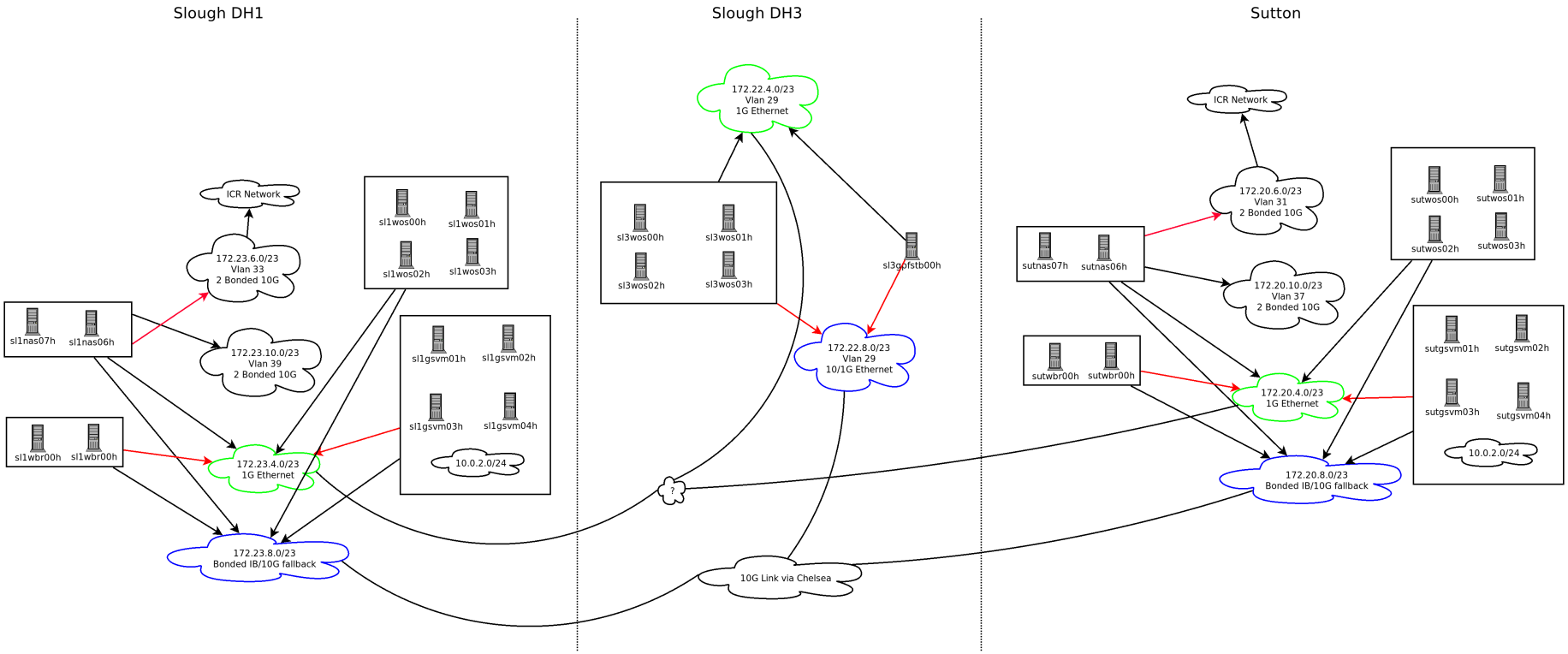
RDS **Service** - User Requirements

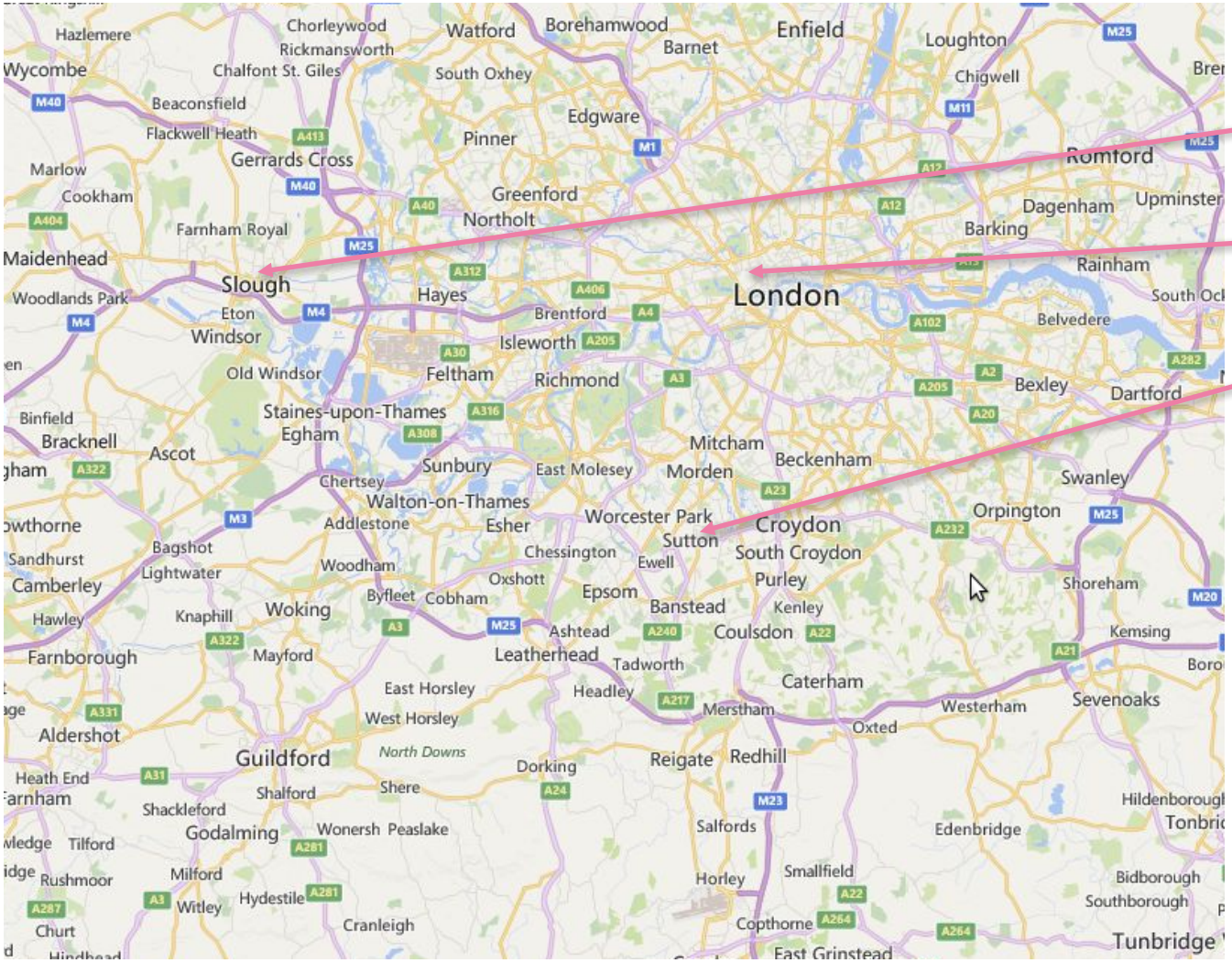
Collected from researcher workshops:

- 6PiB storage, expandable to growing research need; minimum 20PiB capacity
- High level redundancy ensuring robust solution
- Cost effective and competitively priced solution (-> vendors offered two Tiers)
  - Rapid access to data held in Tiers
  - Ability for researchers to manage data transfers between Tiers
  - Single namespace
  - Direct access to data in Tier 2
- Ability to protect against accidental loss of datasets

RDS Architecture

- DDN GS7Ks at Sutton and Slough providing 2PiB of mirrored storage
- DDN WOS using 3 site GOA policy (Sutton and Slough DH1 / DH3)
- GS Bridge to manage data between tiers
  - DMAPI handler enables transparent reads of migrated files
  - Mover uses policy engine rules to move data between tiers
- Total of 8 MediaScaler servers
  - DDN's package of CTDB / Samba / NFS
- Networking within each site (mostly) FDR InfiniBand
- 10G links between sites
- Daily snapshots (but no backup)


Slough DH1
Slough DH3
Sutton
ICR Network
172.23.6.0/23
Vlan 33
2 Bonded 10G
172.23.10.0/23
Vlan 39
2 Bonded 10G
172.23.4.0/23
1G Ethernet
172.23.8.0/23
Bonded IB/10G fallback
sl1nas07h
sl1nas06h
sl1wbr00h
sl1wbr00h
sl1wos00h
sl1wos01h
sl1wos02h
sl1wos03h
sl1gsvm01h
sl1gsvm02h
sl1gsvm03h
sl1gsvm04h
10.0.2.0/24
172.22.4.0/23
Vlan 29
1G Ethernet
172.22.8.0/23
Vlan 29
10/1G Ethernet
sl3wos00h
sl3wos01h
sl3wos02h
sl3wos03h
sl3gofstb00h
10G Link via Chelsea
?
ICR Network
172.20.6.0/23
Vlan 31
2 Bonded 10G
172.20.10.0/23
Vlan 37
2 Bonded 10G
172.20.4.0/23
1G Ethernet
172.20.8.0/23
Bonded IB/10G fallback
sutnas07h
sutnas06h
sutwbr00h
sutwbr00h
sutwos00h
sutwos01h
sutwos02h
sutwos03h
sutgsvm01h
sutgsvm02h
sutgsvm03h
sutgsvm04h
10.0.2.0/24


Slough
Chelsea
Sutton
Hazlemere
Chorleywood
Rickmansworth
Watford
Borehamwood
Barnet
Enfield
Loughton
Chigwell
Wycombe
Chalfont St. Giles
South Oxhey
Beaconsfield
Flackwell Heath
Edgware
Pinner
Gerrards Cross
Marlow
Cookham
Greenford
Northolt
Romford
Dagenham
Barking
Upminster
Farnham Royal
Maidenhead
Slough
Woodlands Park
Eton
Windsor
Hayes
Brentford
Isleworth
London
Rainham
Belvedere
Old Windsor
Feltham
Richmond
Bexley
Dartford
Binfield
Bracknell
Staines-upon-Thames
Egham
Ascot
Sunbury
East Molesey
Mitcham
Morden
Beckenham
Swanley
Chertsey
Walton-on-Thames
Addlestone
Esher
Worcester Park
Croydon
Orpington
Sutton
South Croydon
Chessington
Ewell
Purley
Sandhurst
Bagshot
Lightwater
Woodham
Oxshott
Camberley
Byfleet
Cobham
Epsom
Banstead
Kenley
Shoreham
Hawley
Knaphill
Woking
Ashtead
Coulsdon
Kemsing
Farnborough
Mayford
Leatherhead
Tadworth
Caterham
East Horsley
Headley
Merstham
Westerham
Sevenoaks
West Horsley
Oxted
Aldershot
Guildford
North Downs
Dorking
Reigate
Redhill
Heath End
Farnham
Shalford
Shere
Shackleford
Godalming
Wonersh
Peaslake
Salfords
Edenbridge
Tilford
Rushmoor
Milford
Hydestile
Witley
Horley
Smallfield
Bidborough
Southborough
Churt
Cranleigh
Copthorne
Tunbridge
East Grinstead

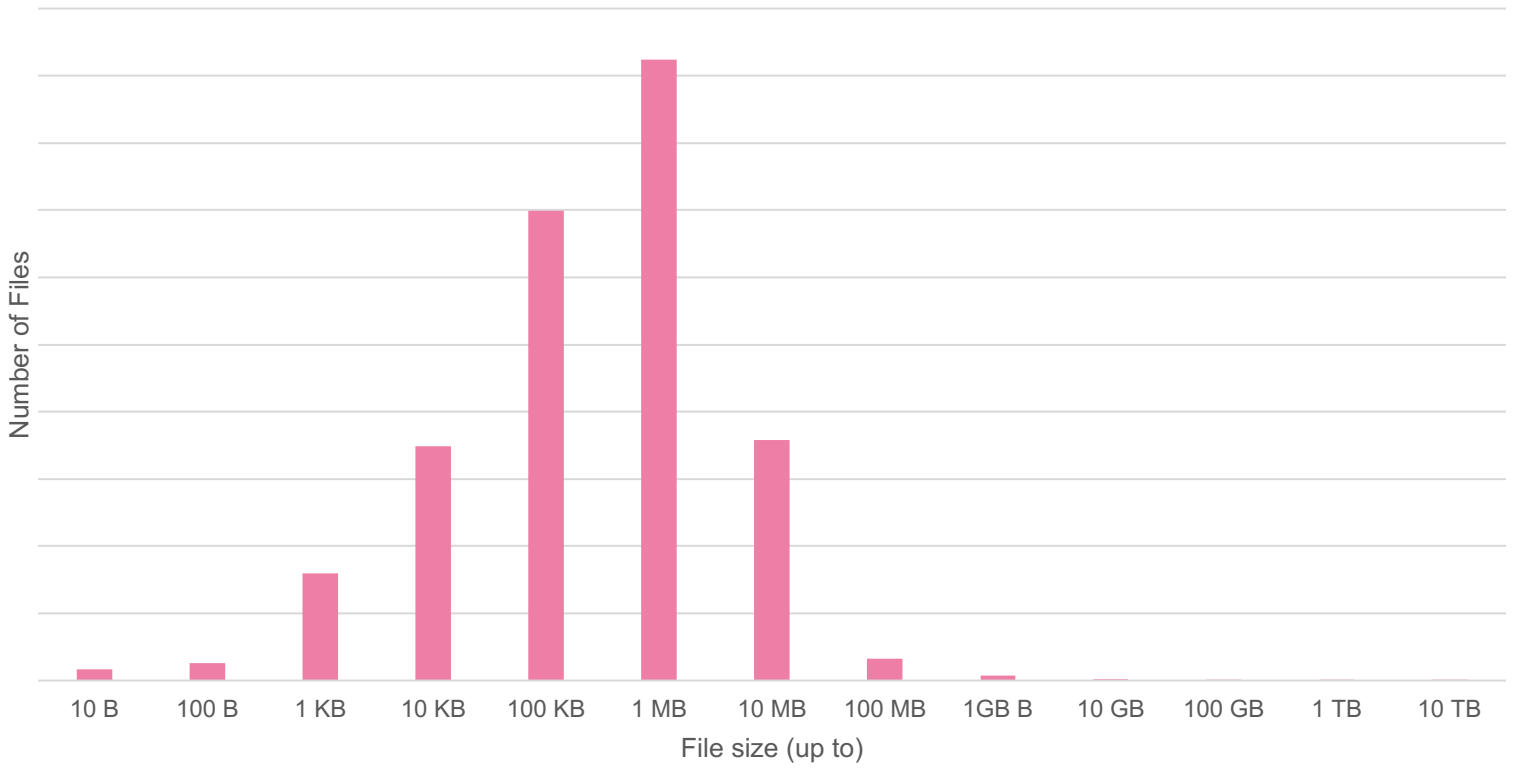
SpectrumScale Performance

- For large files - write performance limited by 10Gb/s inter-site links
  - (but need some network tuning first!)
- Can do better with read performance by setting readReplicaPolicy=fastest
- Small file performance less good (particularly via SMB)
- Mirroring working well
  - We have lost inter site link on a few occasions but recovery has been pretty smooth
  - Takes us ~2 hours for policy scan (100m files) + time needed to resync data
- “Double Size” files causes some confusion.

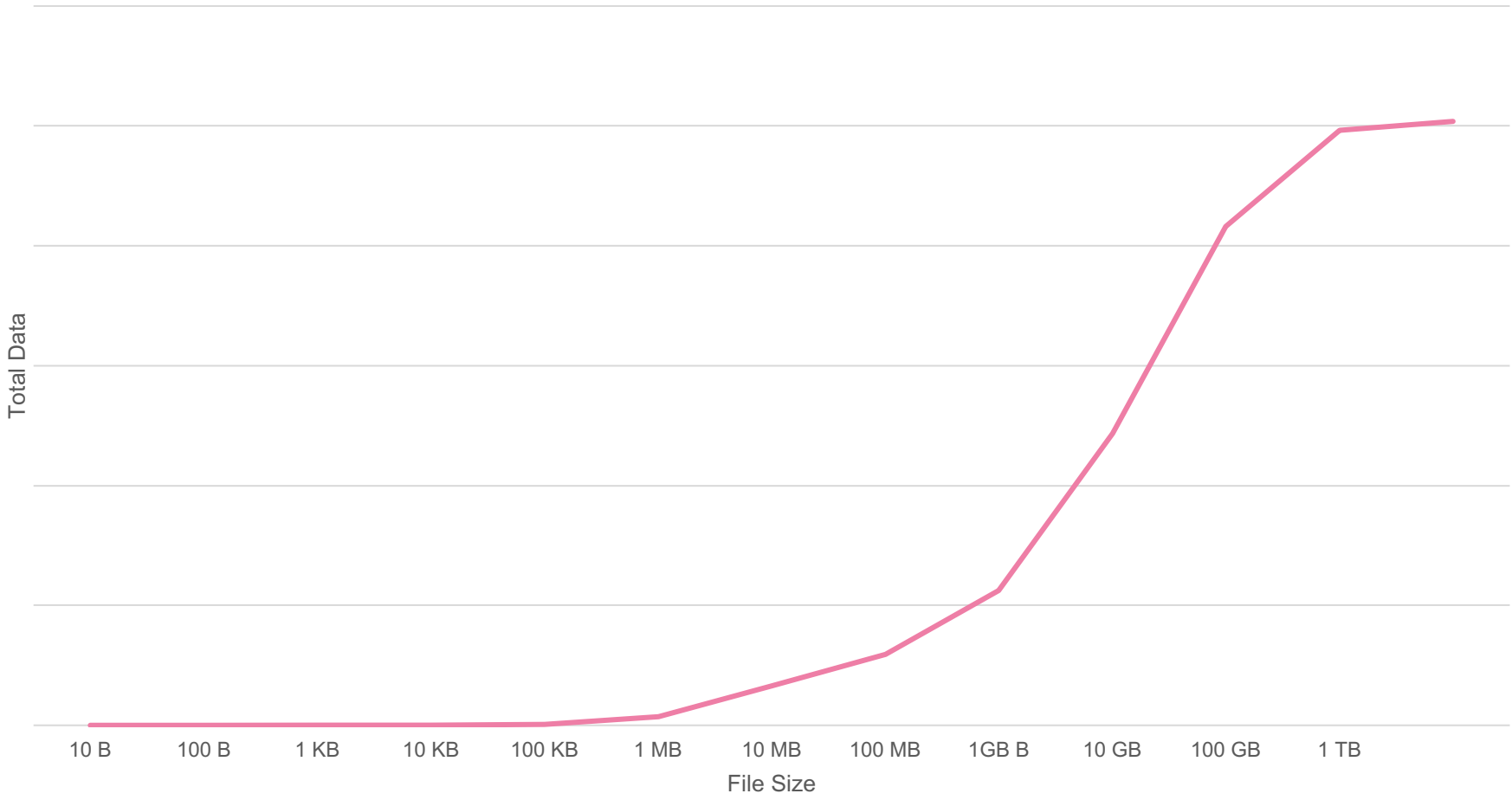
WOS Bridge Performance

- Significantly slower than GPFS data - ~100MB/s
  - …but can scale fairly well (in the right circumstances)
- Sometimes stops working at all
  - DDN have identified some issues with WOS but problems currently ongoing
  - Unfortunate interaction with NFS causes all MediaScaler nodes to fall over!
- Difficult to work out which data should be migrated just based on access times etc
  - Given performance difference we really want all active large files on GPFS

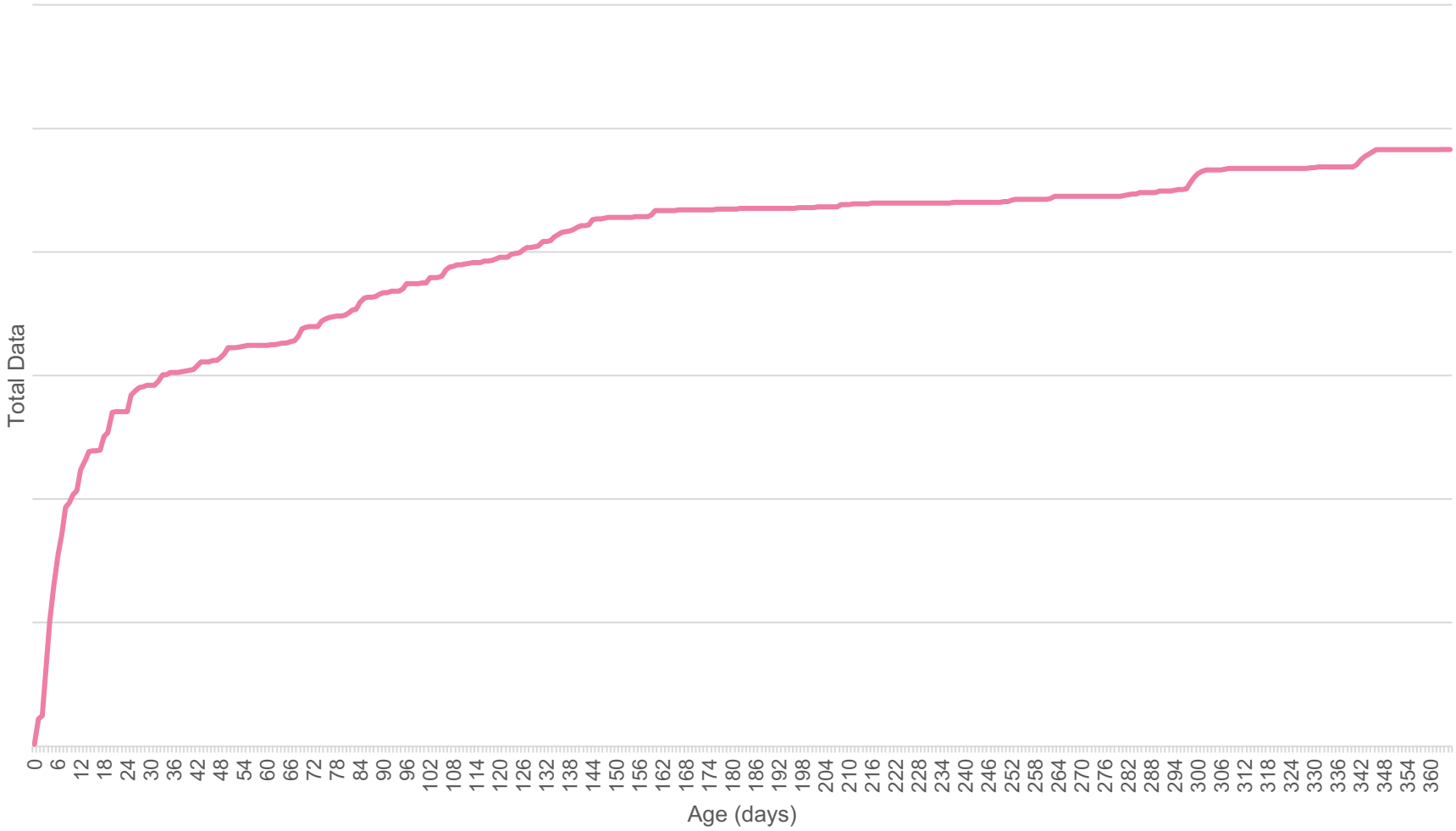
Number of files by size



Total data by file size



Total data by access time

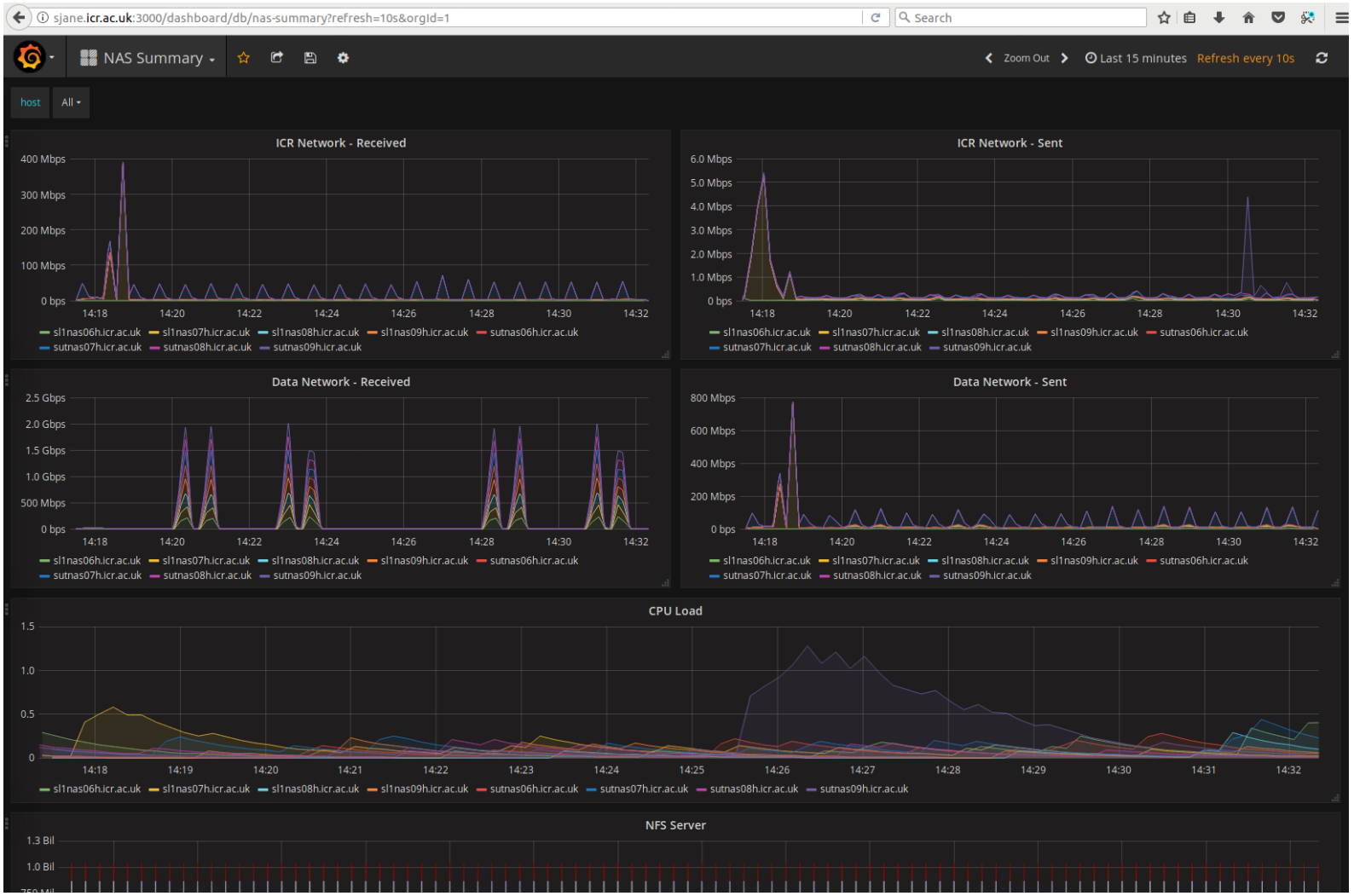


Monitoring

Currently using:

- Icinga2 to check health of system
  - Spectrum Scale Plugins from IBM work well
  - Monitoring NFS / Samba / CTDB more tricky
- Telegraf + InfluxDB + Grafana for metric storage / analysis
  - Haven’t got any Spectrum Scale specific metrics in (yet)

Grafana

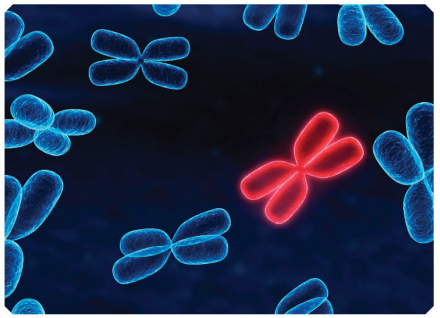

sjane.icr.ac.uk:3000/dashboard/db/nas-summary?refresh=10s&orgId=1
Search
NAS Summary
Zoom Out
Last 15 minutes
Refresh every 10s
host
All
ICR Network - Received
400 Mbps
300 Mbps
200 Mbps
100 Mbps
0 bps
14:18
14:20
14:22
14:24
14:26
14:28
14:30
14:32
sl1nas06h.icr.ac.uk
sl1nas07h.icr.ac.uk
sl1nas08h.icr.ac.uk
sl1nas09h.icr.ac.uk
sutnas06h.icr.ac.uk
sutnas07h.icr.ac.uk
sutnas08h.icr.ac.uk
sutnas09h.icr.ac.uk
ICR Network - Sent
6.0 Mbps
5.0 Mbps
4.0 Mbps
3.0 Mbps
2.0 Mbps
1.0 Mbps
0 bps
Data Network - Received
2.5 Gbps
2.0 Gbps
1.5 Gbps
1.0 Gbps
500 Mbps
0 bps
Data Network - Sent
800 Mbps
600 Mbps
400 Mbps
200 Mbps
0 bps
CPU Load
1.5
1.0
0.5
0
14:18
14:19
14:20
14:21
14:22
14:23
14:24
14:25
14:26
14:27
14:28
14:29
14:30
14:31
14:32
NFS Server
1.3 Bil
1.0 Bil

Filesystem Analysis Tool(s)

- Interested in tools for
  - Helping users see how much data they have
  - Understanding size / age profiles etc
  - Managing quotas
- Looked at
  - Starfish
  - ClarityNow
  - Perl
- In the long term, need to move to being more of a data management service

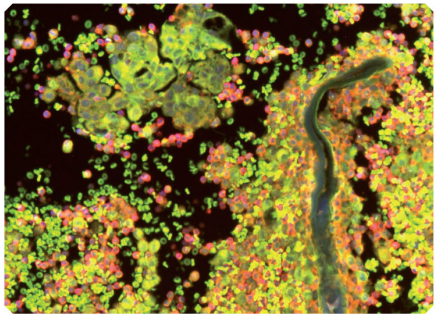
Unrivalled track record

ICR The Institute of Cancer Research



Making the discoveries that defeat cancer

ICR



One of the world's most influential cancer research institutes

ICR The Institute of Cancer Research

[Title slide with Mission Statement]

Making the discoveries that defeat cancer

ICR The Institute of Cancer Research

Our mission is to make the discoveries that defeat cancer.

The Institute of Cancer Research is one of the world's most influential cancer research institutes.

Making discoveries

Scientists and clinicians are working every day in our labs to make a real impact on cancer patients' lives.

Translational research

Our unique partnership with The Royal Marsden and bench-bedside approach means we create and deliver results in a way that other institutions cannot.

Our unrivalled track record

Whether in cancer biology, genetics, personalised therapies or new drug discoveries, our achievements speak for themselves.

[Introduction] Heading

Activate second <List Level> for lighter grey subtitle style

1

Use [Introduction] from the ‘Slide Layouts’ to highlight three specific areas

2

3

[Title and Content]
Activate second <List Level> for lighter grey subtitle style

The first level is 19pt Arial and does not have bullets

- <Increase List Level> to activate the second style which does have a bullet style

[Biographies x1]
Activate second <List Level> for lighter grey subtitle style

Image size:
H=4cm
W=2.8cm

The first level is 19pt Arial and does not have bullets

- <Increase List Level> to activate the second style which does have a bullet style

[Biographies x2]
Activate second <List Level> for lighter grey subtitle style

Image size:
H=4cm
W=2.8cm

The first level is 19pt Arial and does not have bullets

- <Increase List Level> to activate the second style which does have a bullet style

[Biographies x3]
Activate second <List Level> for lighter grey subtitle style

Image size:
H=4cm
W=2.8cm

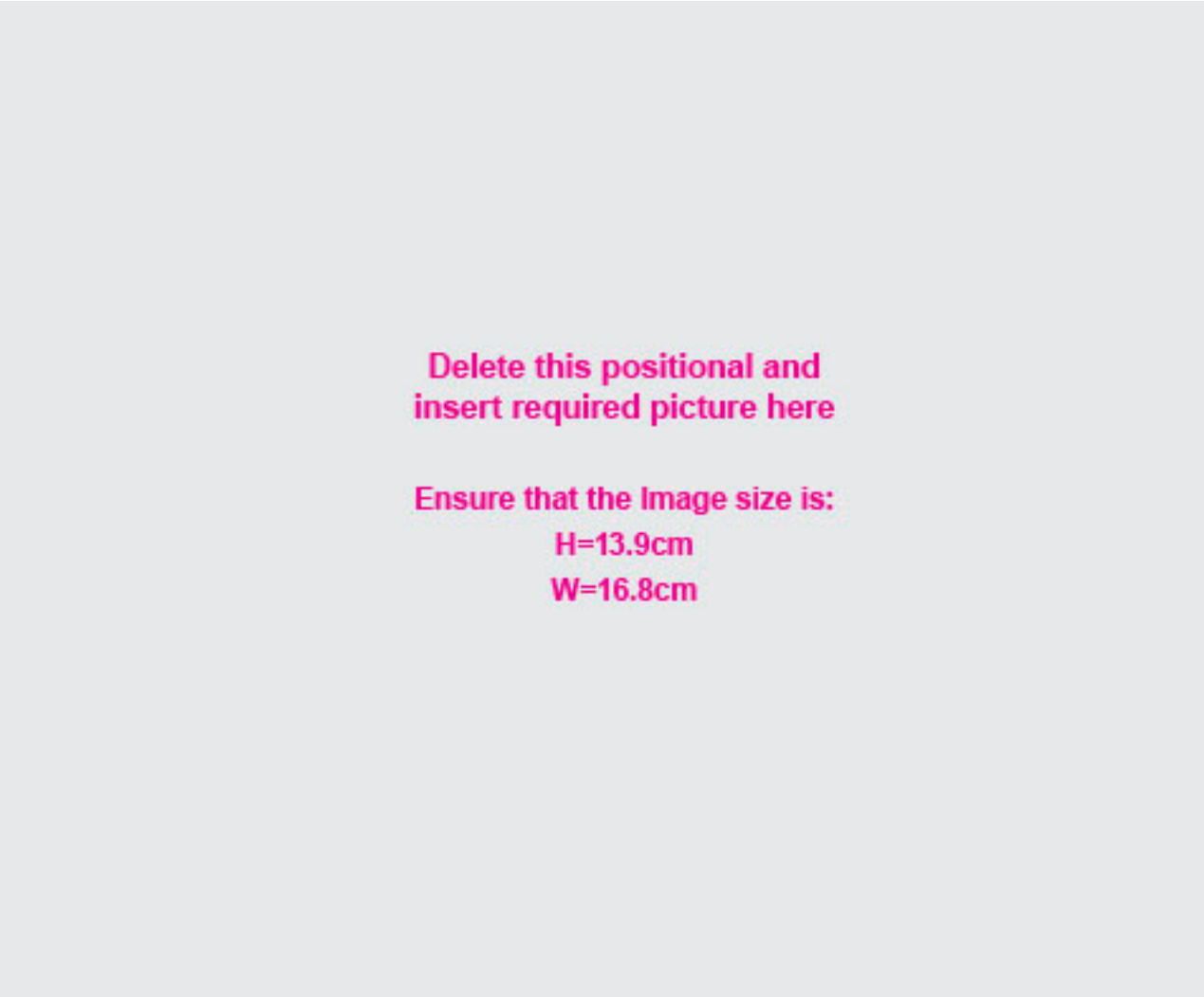
The first level is 19pt Arial and does not have bullets

- <Increase List Level> to activate the second style which does have a bullet style

[V1 Image with caption]
Second <List Level> for subtitle style

The first level is 19pt Arial and does not have bullets

- <Increase List Level> to activate the second style which does have a bullet style



[V2 Image with caption]
Second <List Level> for subtitle style

The first level is 19pt Arial and does not have bullets

- <Increase List Level> to activate the second style which does have a bullet style

[Image] Heading
Second <List Level> for subtitle style

Delete this positional and insert required picture here

Ensure that the Image size is:

H=13.9cm

W=24cm

[Title only] Heading

Second <List Level> for subtitle style

ICR The Institute of Cancer Research

[Author] 19pt Arial
Activate second <List Level> for Title
Department
Location
name@icr.ac.uk


ICR