imense

the GRID as a catalyst for new business

Dr David Sinclair

david.sinclair@imense.com

http://www.imense.com

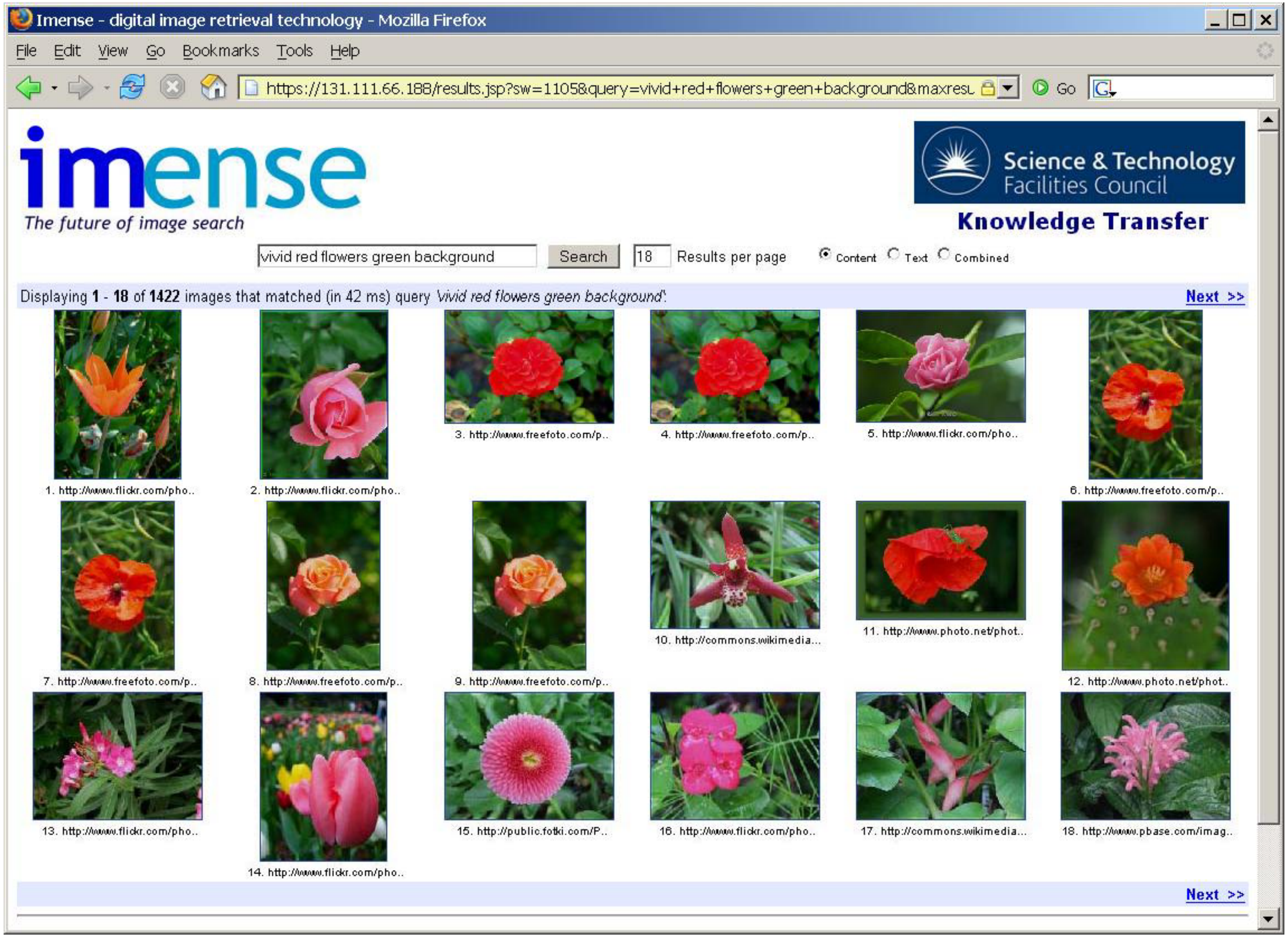
© imense Ltd

What does imense do?

- Sell image retrieval technology:
- Image content analysis
  - High reliability classifiers for visual content
    - Shapes, scenes, faces, colours etc.
- Indexing and query language provision
  - NLP parser for text query over image content
- Image search portal
  - Search photo-hosted and commercial images

imense

“vivid red flowers green background”


Imense - digital image retrieval technology - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
imense
The future of image search
Science & Technology Facilities Council
Knowledge Transfer

https://131.111.66.188/results.jsp?sw=1105&query=vivid+red+flowers+green+background&maxresu

vivid red flowers green background [Search] 18 Results per page ● Content ○ Text ○ Combined

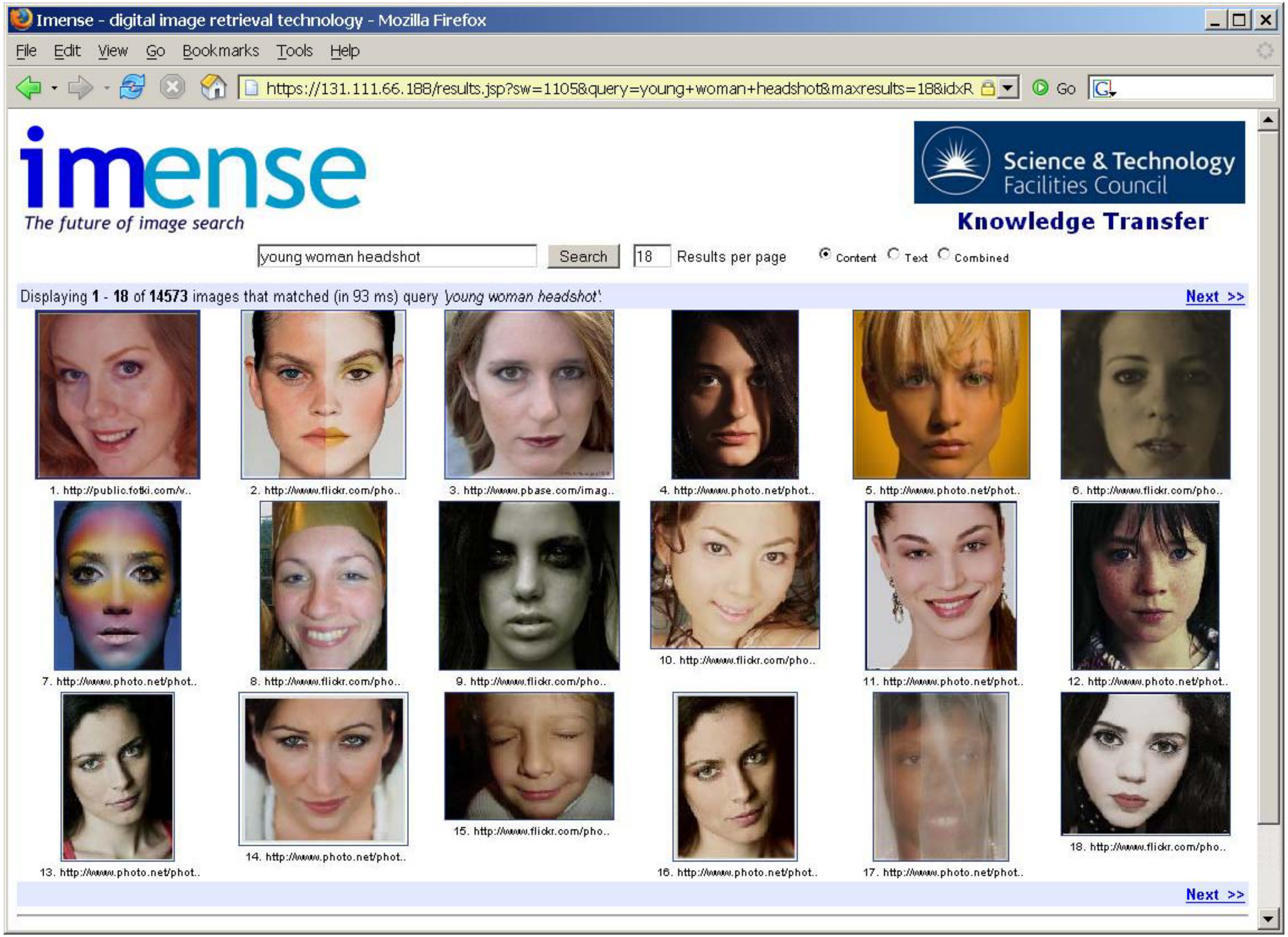
Displaying **1** - **18** of **1422** images that matched (in 42 ms) query *'vivid red flowers green background'* Next >>

1. http://www.flickr.com/pho..
2. http://www.flickr.com/pho..
3. http://www.freefoto.com/p..
4. http://www.freefoto.com/p..
5. http://www.flickr.com/pho..
6. http://www.freefoto.com/p..
7. http://www.freefoto.com/p..
8. http://www.freefoto.com/p..
9. http://www.freefoto.com/p..
10. http://commons.wikimedia...
11. http://www.photo.net/phot..
12. http://www.photo.net/phot..
13. http://www.flickr.com/pho..
14. http://www.flickr.com/pho..
15. http://public.fotki.com/P..
16. http://www.flickr.com/pho..
17. http://commons.wikimedia...
18. http://www.pbase.com/imag..

Next >>

imense

“young woman headshot”


Imense - digital image retrieval technology - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
Go
Science & Technology Facilities Council
Knowledge Transfer

https://131.111.66.188/results.jsp?sw=1105&query=young+woman+headshot&maxresults=18&idxR

imense
The future of image search

young woman headshot [Search] 18 Results per page ● Content ○ Text ○ Combined

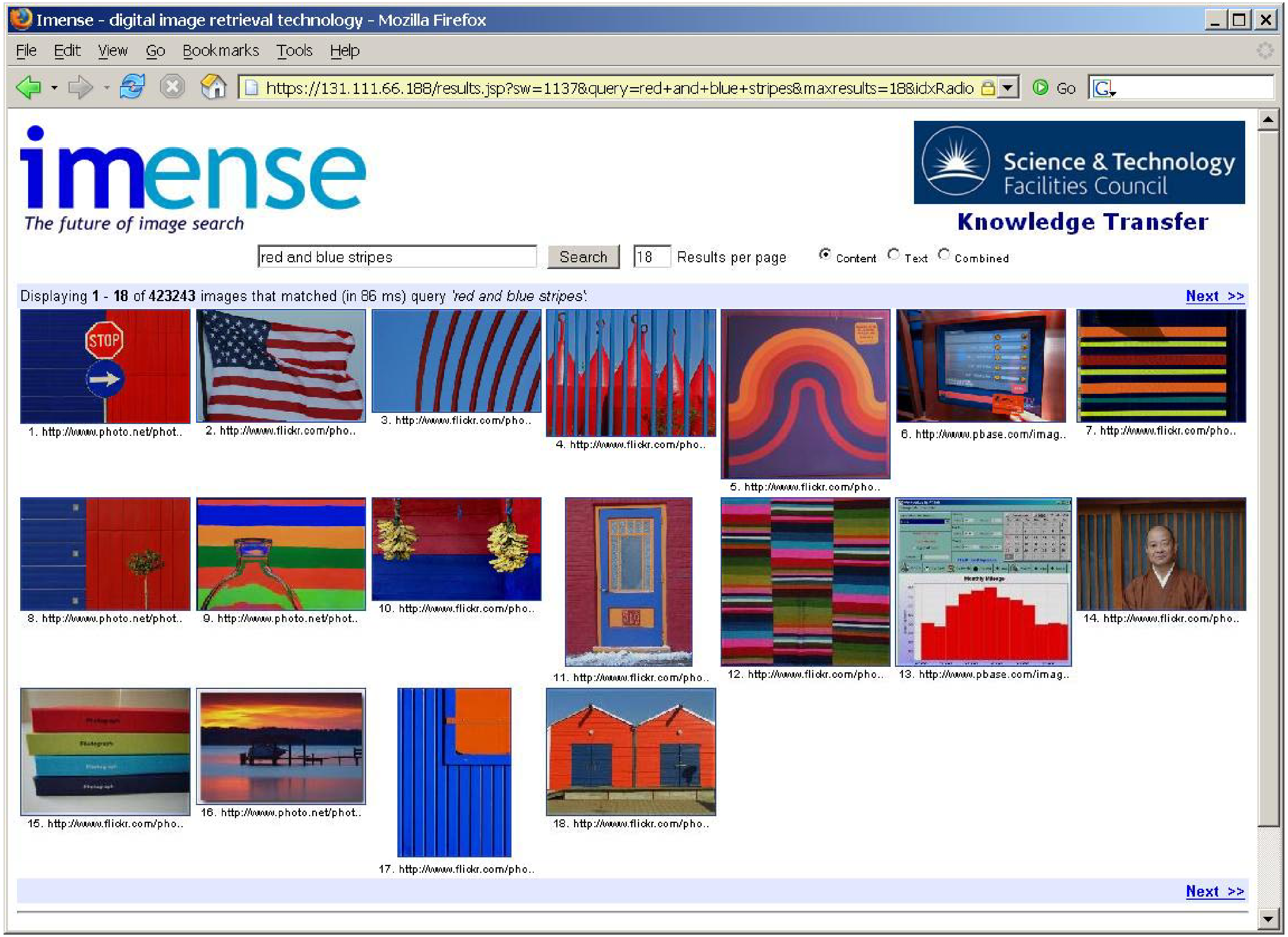
Displaying **1** - **18** of **14573** images that matched (in 93 ms) query *'young woman headshot'*. **Next >>**

1. http://public.fotki.com/v..
2. http://www.flickr.com/pho..
3. http://www.pbase.com/imag..
4. http://www.photo.net/phot..
5. http://www.photo.net/phot..
6. http://www.flickr.com/pho..
7. http://www.photo.net/phot..
8. http://www.flickr.com/pho..
9. http://www.flickr.com/pho..
10. http://www.flickr.com/pho..
11. http://www.photo.net/phot..
12. http://www.photo.net/phot..
13. http://www.photo.net/phot..
14. http://www.photo.net/phot..
15. http://www.flickr.com/pho..
16. http://www.photo.net/phot..
17. http://www.photo.net/phot..
18. http://www.flickr.com/pho..

Next >>

imense

"red and blue stripes"



Imense - digital image retrieval technology - Mozilla Firefox

https://131.111.66.188/results.jsp?sw=1137&query=red+and+blue+stripes&maxresults=18&idxRadio

File Edit View Go Bookmarks Tools Help
Go
imense
The future of image search
Science & Technology Facilities Council
Knowledge Transfer

red and blue stripes Search 18 Results per page ● Content ○ Text ○ Combined

Displaying **1** - **18** of **423243** images that matched (in 86 ms) query *'red and blue stripes'*. **Next >>**

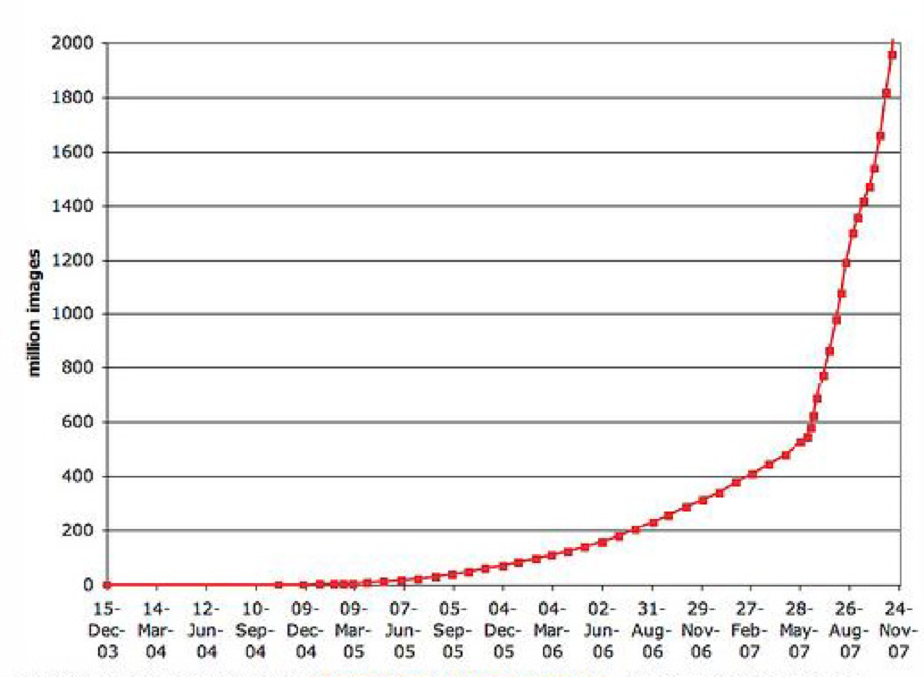
1. http://www.photo.net/phot..
2. http://www.flickr.com/pho..
3. http://www.flickr.com/pho..
4. http://www.flickr.com/pho..
5. http://www.flickr.com/pho..
6. http://www.pbase.com/imag..
7. http://www.flickr.com/pho..
8. http://www.photo.net/phot..
9. http://www.photo.net/phot..
10. http://www.flickr.com/pho..
11. http://www.flickr.com/pho..
12. http://www.flickr.com/pho..
13. http://www.pbase.com/imag..
14. http://www.flickr.com/pho..
15. http://www.flickr.com/pho..
16. http://www.photo.net/phot..
17. http://www.flickr.com/pho..
18. http://www.flickr.com/pho..

Next >>

How did imense get here?

- Founded in 2004 by: Dr David Sinclair and Dr Chris Town on the back of a DTI grant for R&D.
  - Equipment: 1 laptop, 1 3.2GHz desktop PC, 300Gig of disk.
  - Processed and indexed ~50,000 images.
  - ~5x10^6 images uploaded to flickr by early 2005.
    - 2.5 years processing with or available resources…
  - Google had ~200,000 servers in 2005.
- Mini PIPPS grant awarded carrying GRID access 2006.
  - imense indexed 3.5x10^6 images over the course of the grant
  - Indexing time roughly a month on a small fraction of the GRID.
  - 300x10^6 images uploaded to flickr by mid 2007.
- Angel investment of 535k generated in July 2007 as a consequence of demonstrating that our indexing technology scaled.
  - imense has now indexed 10x10^6 images.
  - 2x10^9 images now uploaded to flickr
  - 15+x10^9 images on the internet?
  - We still need the GRID for Internet scale image retrieval.
  - Google has 450,000+ servers (estimated 2008).
- Full PIPPS grant to look at improving access to the GRID

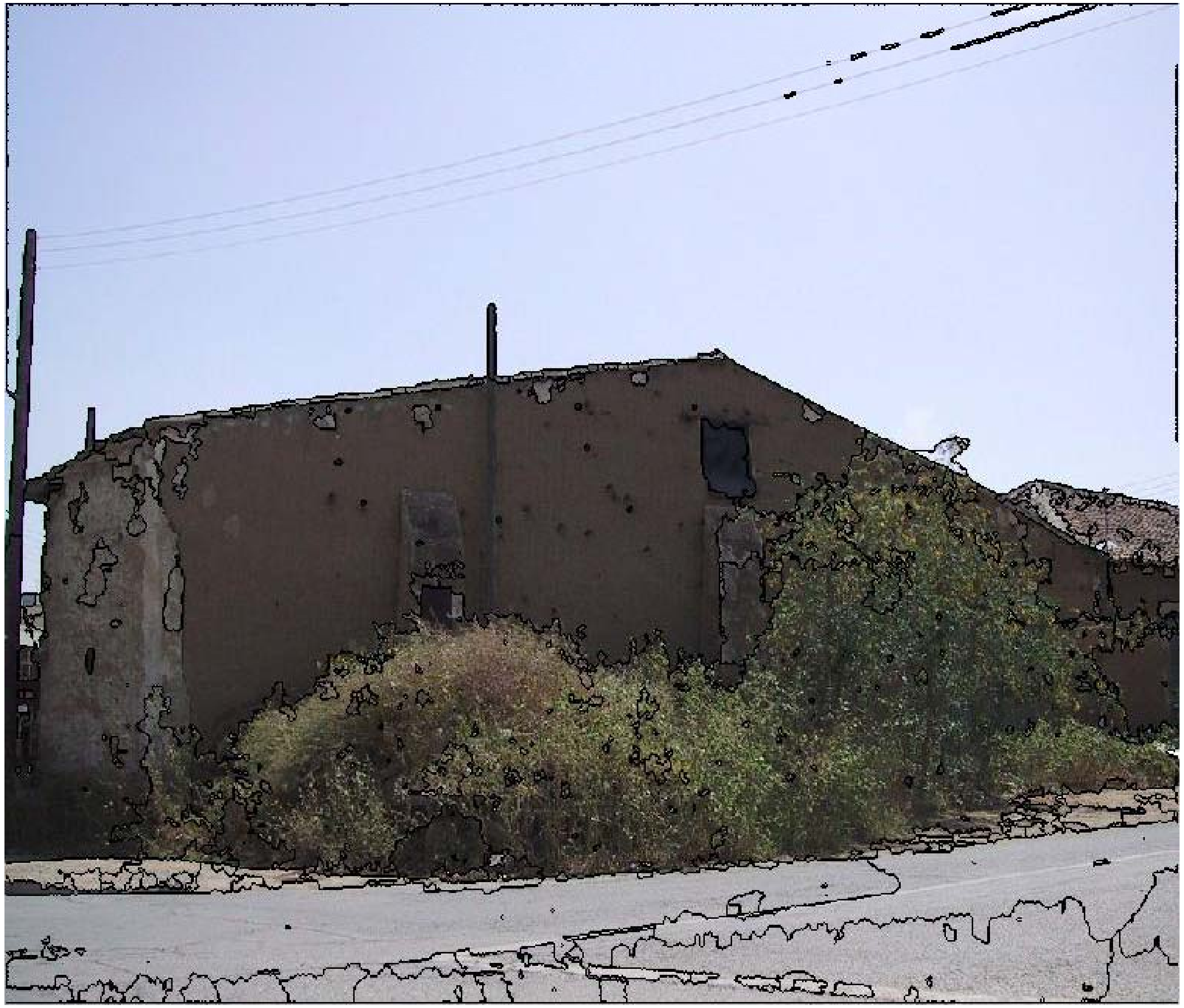
Internet scale image search?



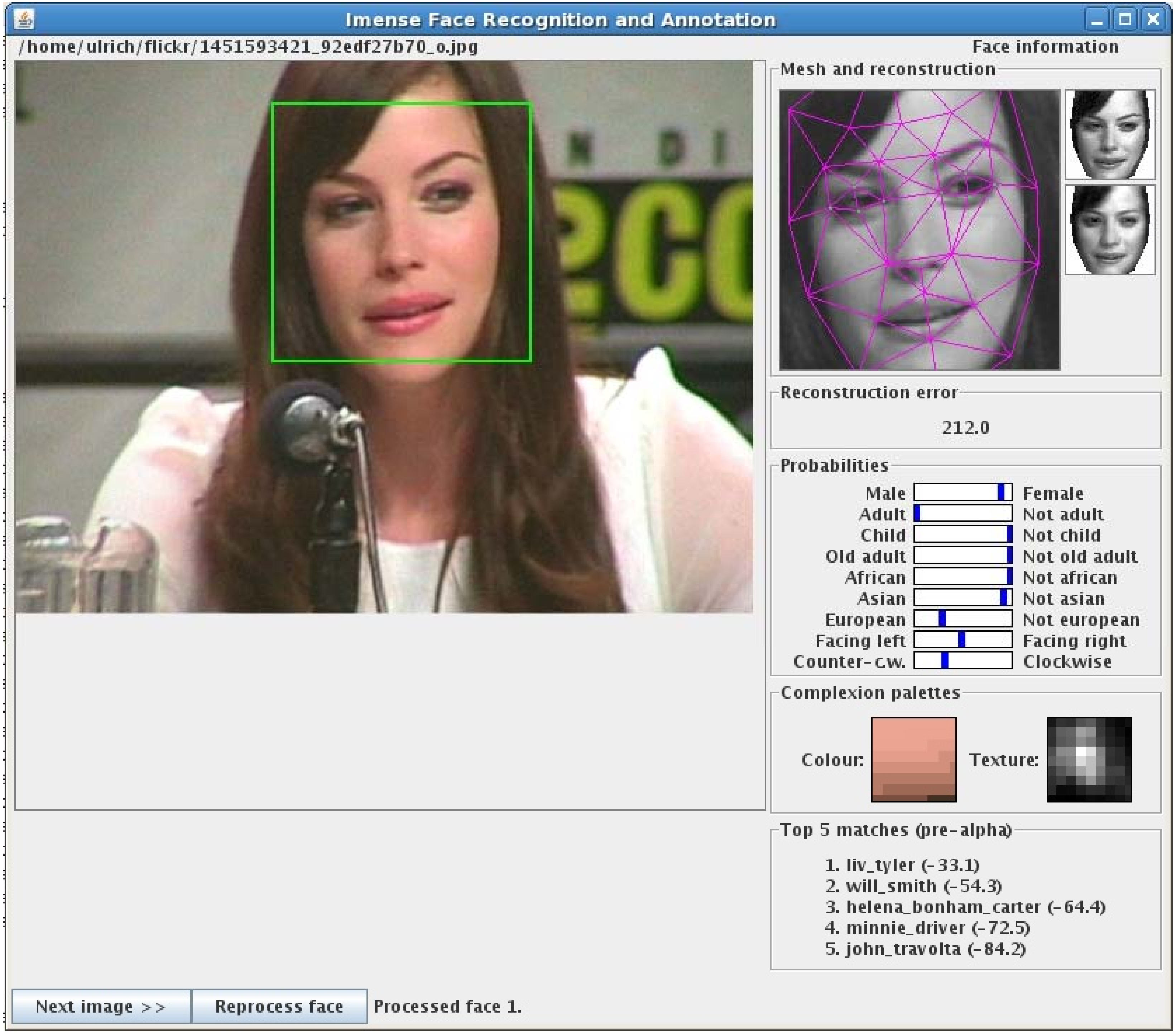
2,000,000,000 images
uploaded to flickr by Jan 2008

- There are more then 10x10^9 images hosted on the Internet!
  - 4439 cpu-years to index.
    - Or 16 days on a 100,000cpu cluster!
  - 300kBytes/ image 3x10^12 Bytes
    - $0.10/Gig $300,000 just for bandwidth to move the images about.
- We still need grid computing.

imense technology



Segmentation 1 second per image.



/home/ulrich/flickr/1451593421_92edf27b70_o.jpg

Face information

Mesh and reconstruction

Reconstruction error

212.0

Probabilities

Male	Female
Adult	Not adult
Child	Not child
Old adult	Not old adult
African	Not african
Asian	Not asian
European	Not european
Facing left	Facing right
Counter-cw.	Clockwise

Complexion palettes

Colour: Texture:

Top 5 matches (pre-alpha)

1. liv_tyler (-33.1)
2. will_smith (-54.3)
3. helena_bonham_carter (-64.4)
4. minnie_driver (-72.5)
5. john_travolta (-84.2)

Processed face 1.

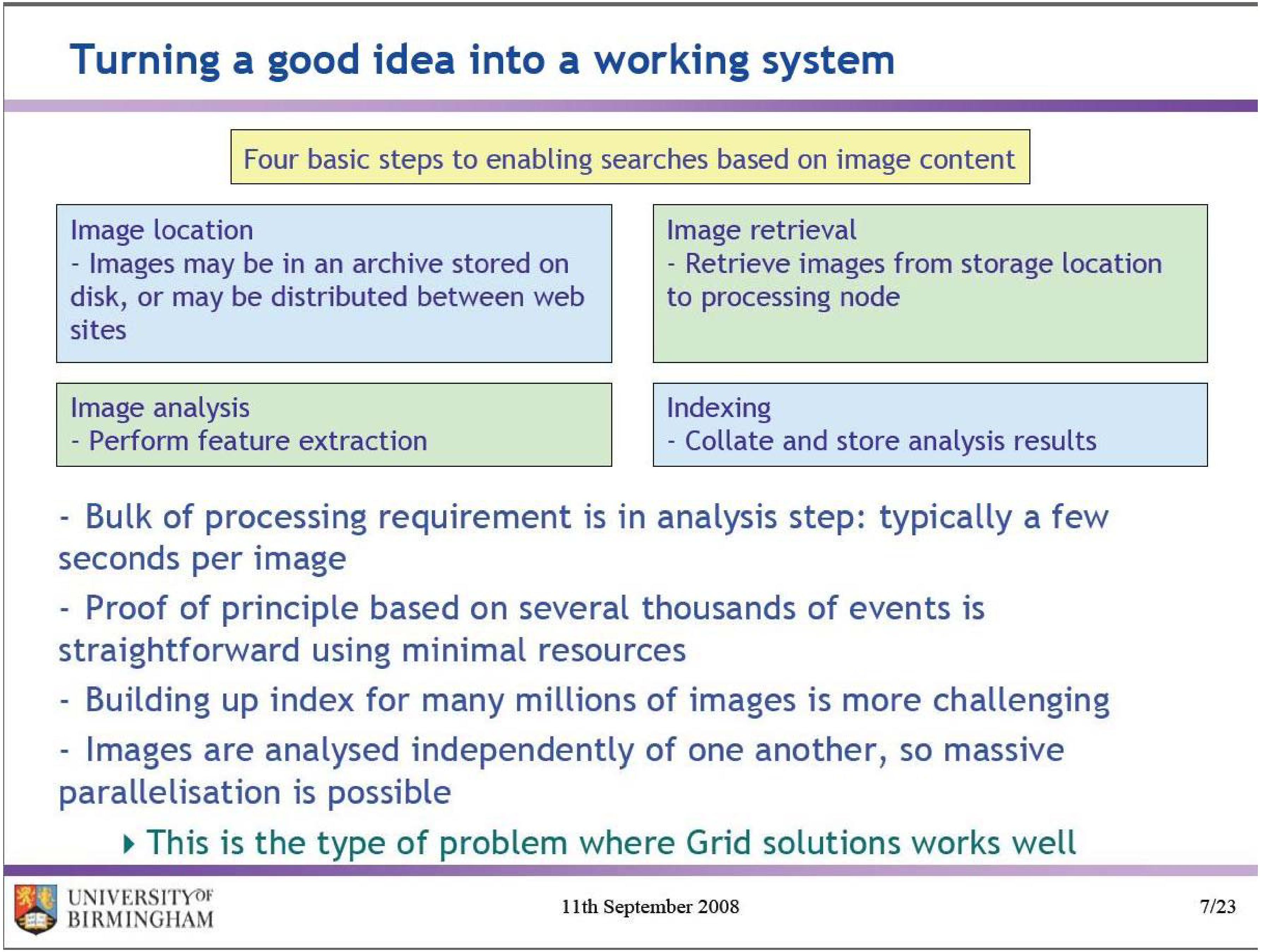
imense

problem

- Index all of the images on PhotoHosting websites (not even the whole web)
  - – Rough estimate 10,000,000,000 images: (flickr.com, photobucket.com, webshots.com etc.)
  - – 14sec per image…..4439 cpu_years.
    - 16days on a 100,000cpu cluster!
  - – 300kBytes/ image 3x10^12 Bytes
    - $0.05/Gig $150,000 just for bandwidth to move the images about.

Interacting with GridPP

- I have unashamedly stolen the next few slides for Karl Harrison who has contributed greatly to our work with GridPP.



Turning a good idea into a working system

Four basic steps to enabling searches based on image content

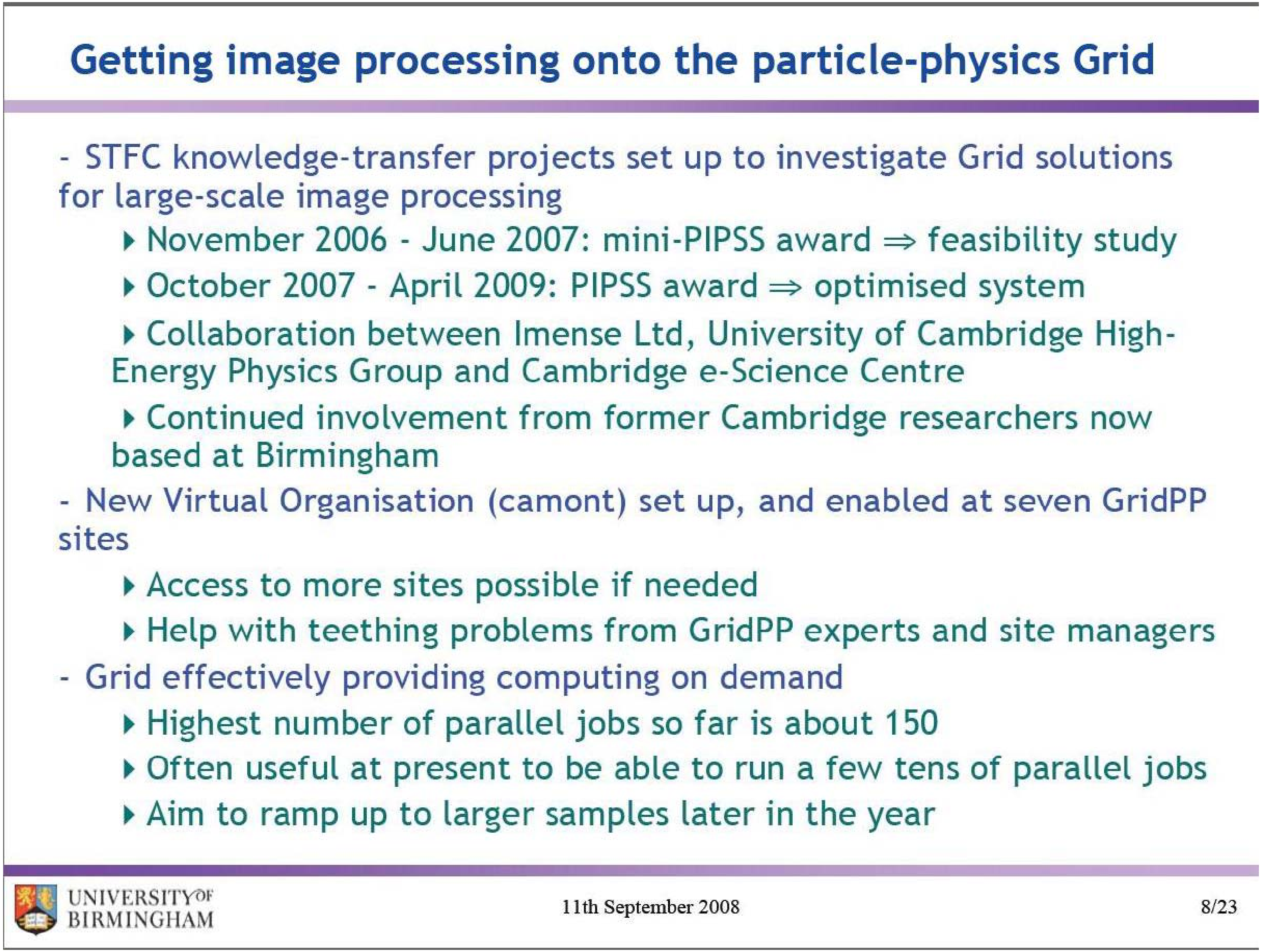
Image location
- Images may be in an archive stored on disk, or may be distributed between web sites

Image retrieval
- Retrieve images from storage location to processing node

Image analysis
- Perform feature extraction

Indexing
- Collate and store analysis results

- Bulk of processing requirement is in analysis step: typically a few seconds per image
- Proof of principle based on several thousands of events is straightforward using minimal resources
- Building up index for many millions of images is more challenging
- Images are analysed independently of one another, so massive parallelisation is possible
  - ▸ This is the type of problem where Grid solutions works well



Getting image processing onto the particle-physics Grid

- STFC knowledge-transfer projects set up to investigate Grid solutions for large-scale image processing
  - November 2006 - June 2007: mini-PIPSS award ⇒ feasibility study
  - October 2007 - April 2009: PIPSS award ⇒ optimised system
  - Collaboration between Imense Ltd, University of Cambridge High-Energy Physics Group and Cambridge e-Science Centre
  - Continued involvement from former Cambridge researchers now based at Birmingham
- New Virtual Organisation (camont) set up, and enabled at seven GridPP sites
  - Access to more sites possible if needed
  - Help with teething problems from GridPP experts and site managers
- Grid effectively providing computing on demand
  - Highest number of parallel jobs so far is about 150
  - Often useful at present to be able to run a few tens of parallel jobs
  - Aim to ramp up to larger samples later in the year

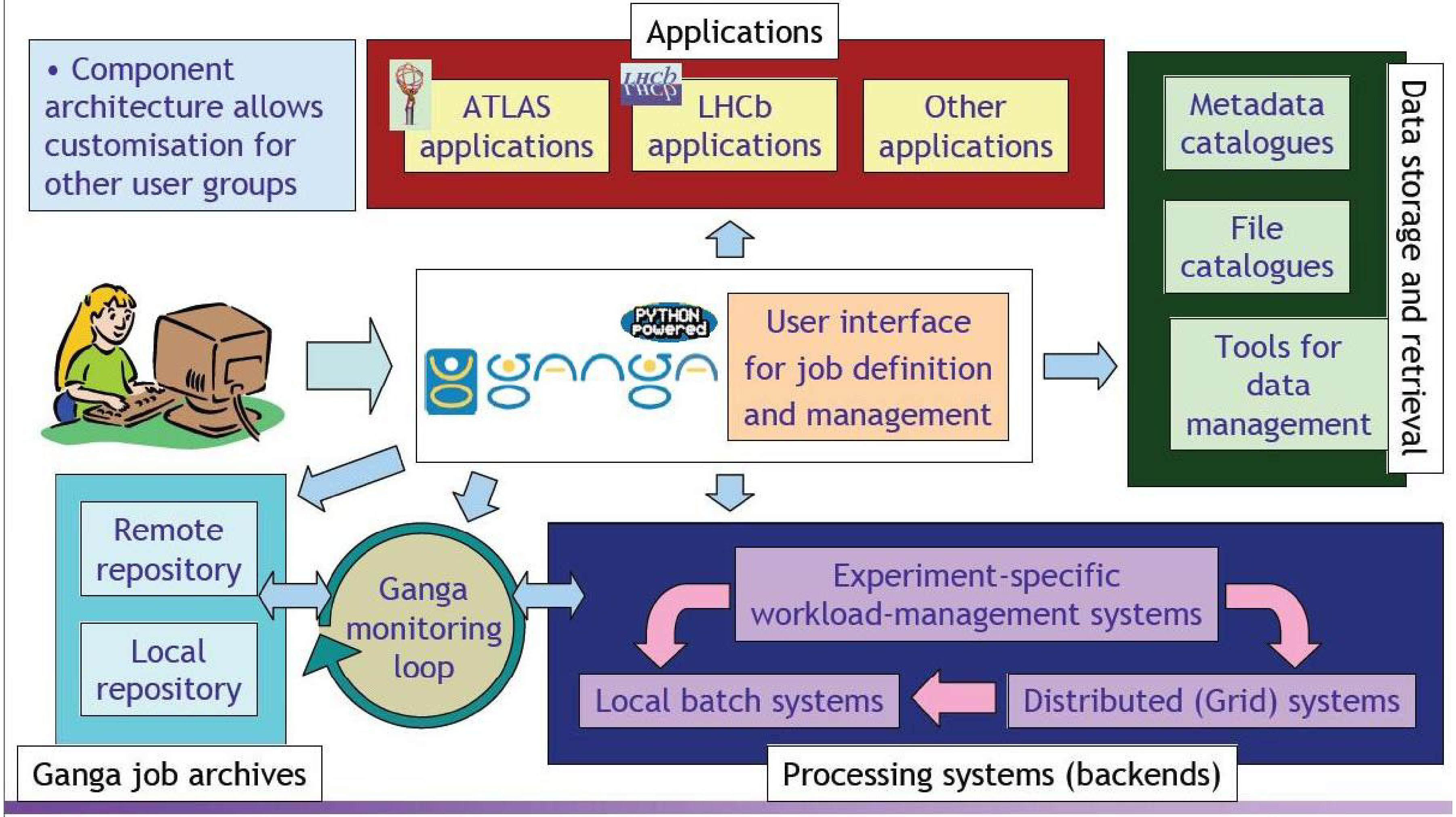
UNIVERSITY OF BIRMINGHAM

11th September 2008

8/23

Job-management sytem and Grid user interface

- Use Ganga system, developed to support particle-physics experiments (ATLAS and LHCb)

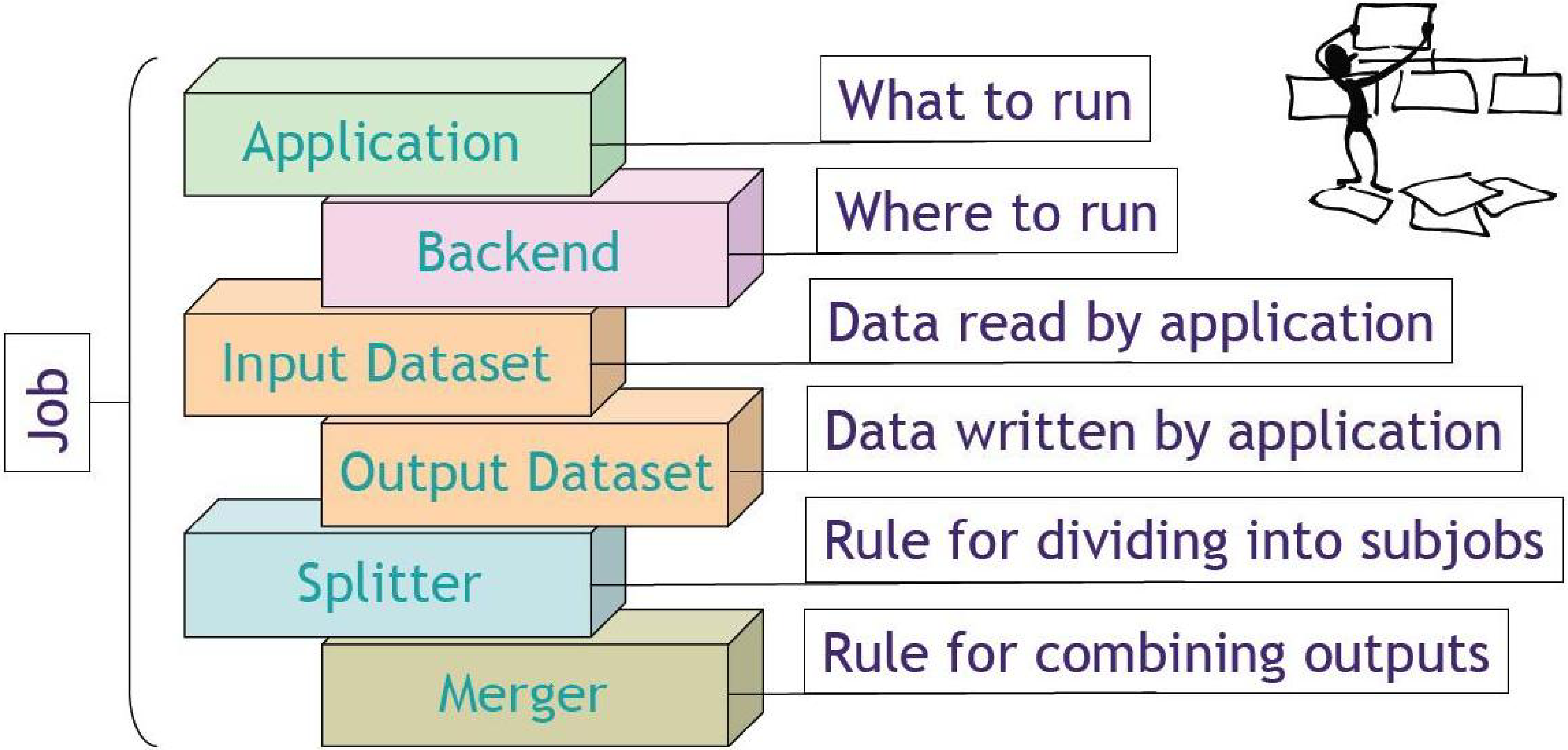


• Component architecture allows customisation for other user groups
Applications
ATLAS applications
LHCb applications
Other applications
PYTHON Powered
ganga
User interface for job definition and management
Metadata catalogues
File catalogues
Tools for data management
Data storage and retrieval
Remote repository
Local repository
Ganga monitoring loop
Experiment-specific workload-management systems
Local batch systems
Distributed (Grid) systems
Ganga job archives
Processing systems (backends)

© imense Ltd

Ganga job abstraction

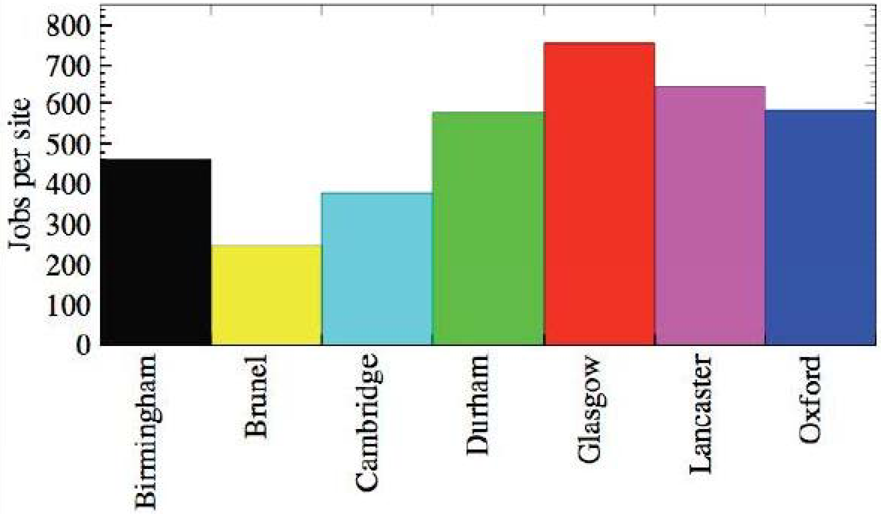
A job in Ganga is constructed from a set of building blocks, not all needed for every job



© imense Ltd

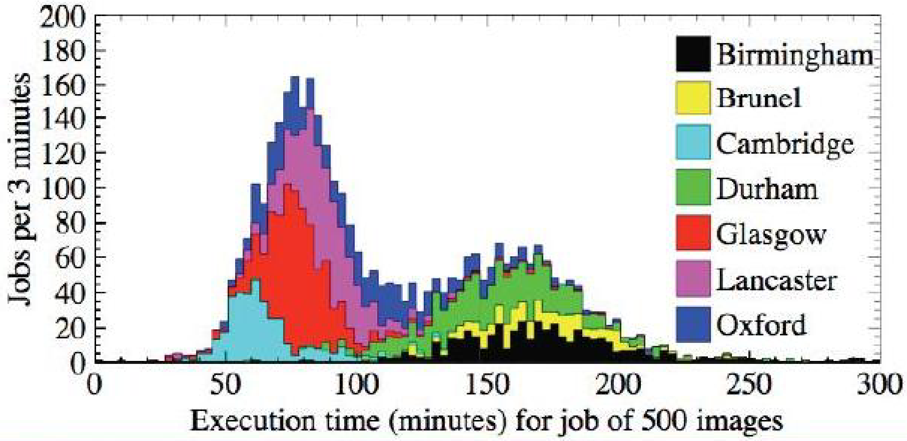
Job destinations and execution times

Results for 3638 jobs submitted over four-week period, July-August 2008



Destination chosen by Resource Broker of Workload Management System, based on minimum estimated waiting time

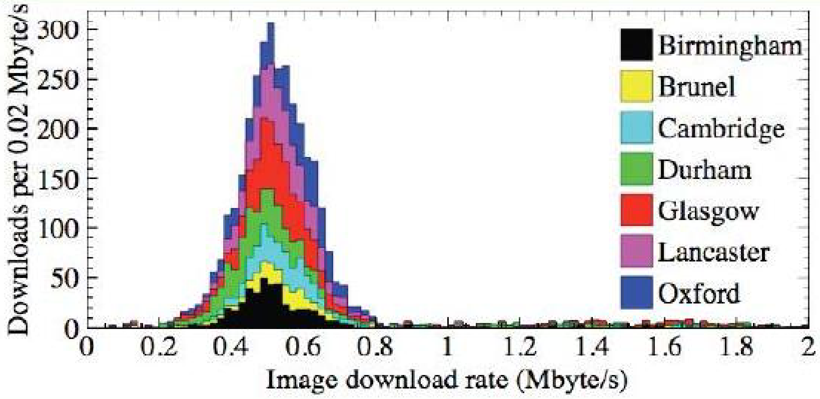
Significant differences in execution times reflect inhomogeneity of site resources



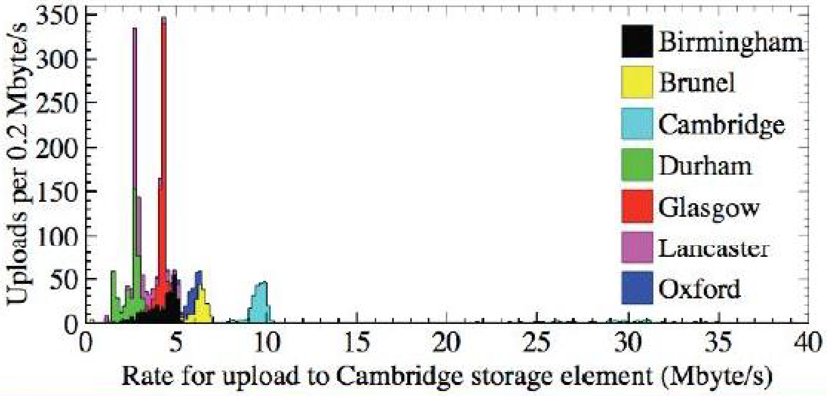
Data-transfer rates

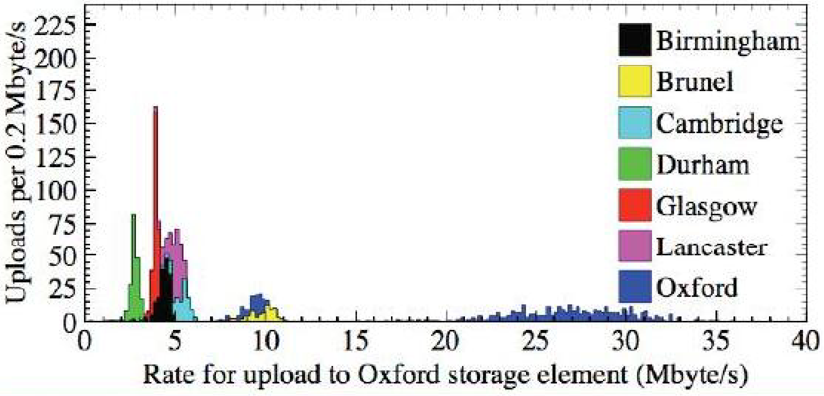
Results for 3638 jobs submitted over four-week period, July-August 2008

Image downloads from hosting site, using wget



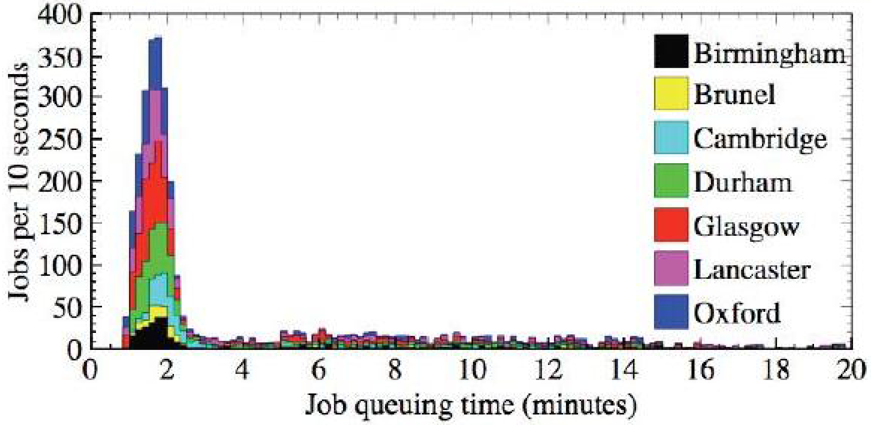
Upload of results to Grid storage elements, using globus-url-copy

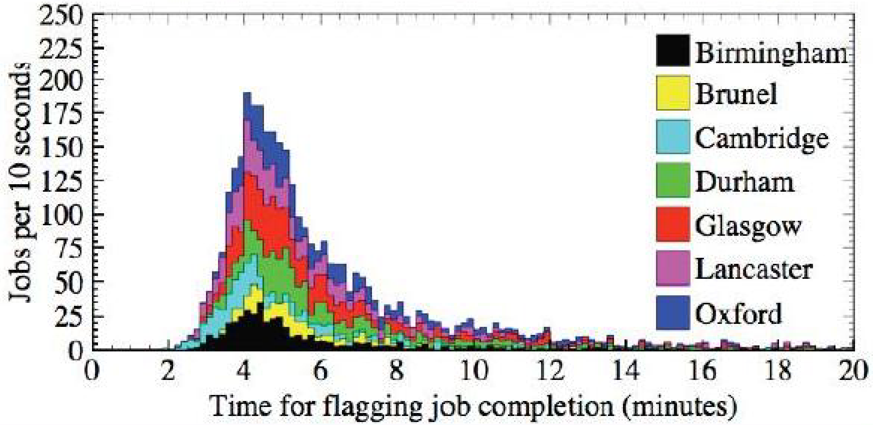


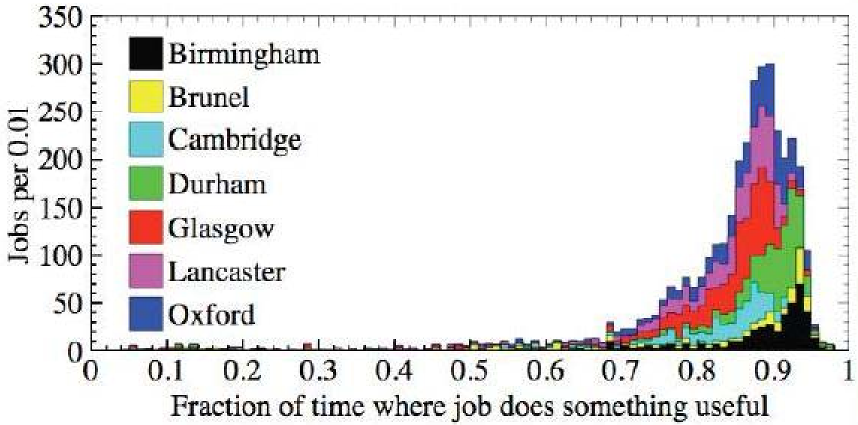


Grid overheads

Results for 3638 jobs submitted over four-week period, July-August 2008







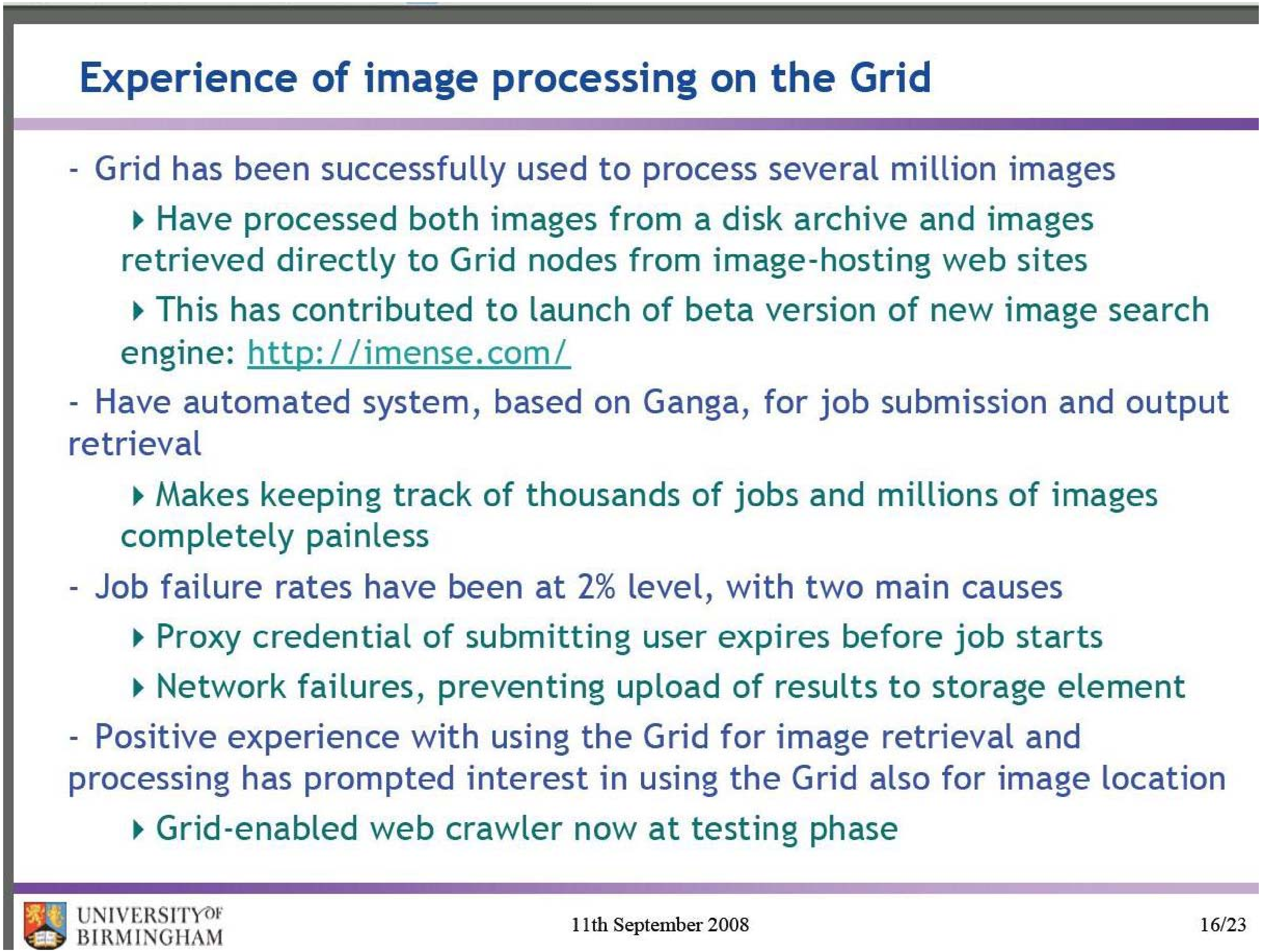
-Useful time is when job is downloading and processing images

- Grid overheads come from: startup time, system time for logging job completion, result upload and retrieval

- For jobs of 500 images, average start-to-finish time is 164 minutes, with 39 minutes spent on Grid overheads: 73% useful time

⇒ Timing distributions non-Gaussian, with long tails

⇒ Need to increase processing load to increase fraction of useful time



Experience of image processing on the Grid

- Grid has been successfully used to process several million images
  - Have processed both images from a disk archive and images retrieved directly to Grid nodes from image-hosting web sites
  - This has contributed to launch of beta version of new image search engine: http://imense.com/
- Have automated system, based on Ganga, for job submission and output retrieval
  - Makes keeping track of thousands of jobs and millions of images completely painless
- Job failure rates have been at 2% level, with two main causes
  - Proxy credential of submitting user expires before job starts
  - Network failures, preventing upload of results to storage element
- Positive experience with using the Grid for image retrieval and processing has prompted interest in using the Grid also for image location
  - Grid-enabled web crawler now at testing phase

thanks

- STFC for providing
  - Technical vote of confidence in our project
  - Facilities to demonstrate our technology
  - Expertise in GRID processing
  - money
- Personal thanks for invaluable support with our project to:
  - Prof Michael, A. Parker of Cambridge University, High Energy Physics Group.
  - Dr Alexander Efimov of Qi3
  - Dr Karl Harrison regrettably now of Birmingham University.