UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

High Performance Computing in the Life Sciences

or

It came for us, it’s coming for you

HPC questions

Who am I and why are you listening to me?

What is High Performance computing?

What are the challenges of HPC?

What happens when biomedical research meets HPC?

How we are meeting the challenges?

A Scientific Computing CV

Peter Maccallum BSc PhD

Head of IT and Scientific Computing

Cancer Research UK Cambridge Research Institute

A Scientific Computing CV

Peter Maccallum BSc PhD

Head of IT and Scientific Computing

Cancer Research UK Cambridge Research Institute

(Relevant) previous employers

- Edinburgh Parallel Computing Centre
- European Bioinformatics Institute

High Performance Computing

Fast processors

- …or multiple processors
- …or multiple cores

High memory

- …or shared memory
- …or memory interconnect

Fast interconnect

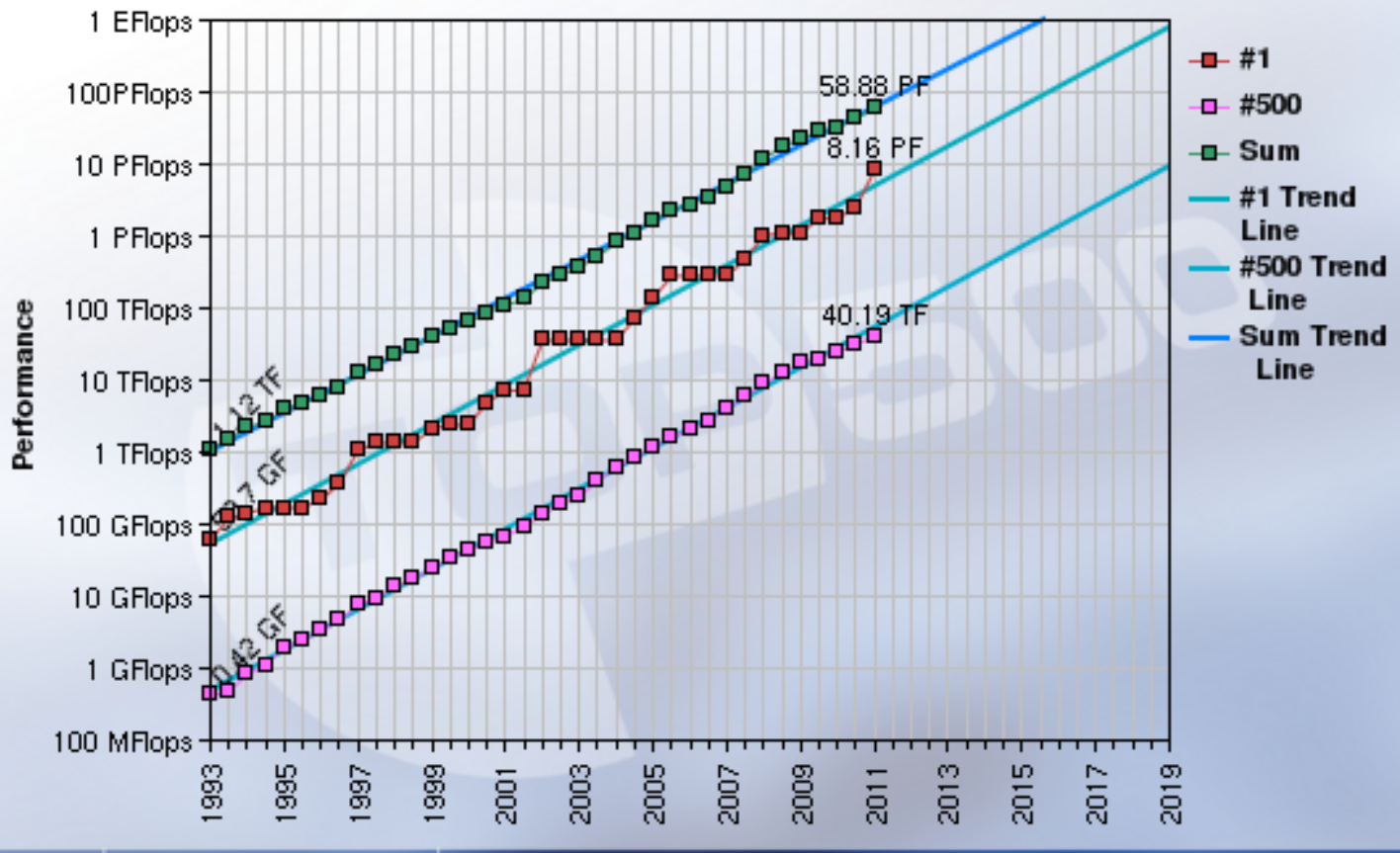
Fast disk

‘Fast’ … ‘High’ … these are relative terms

Cray T3D
1993
19 Gflops

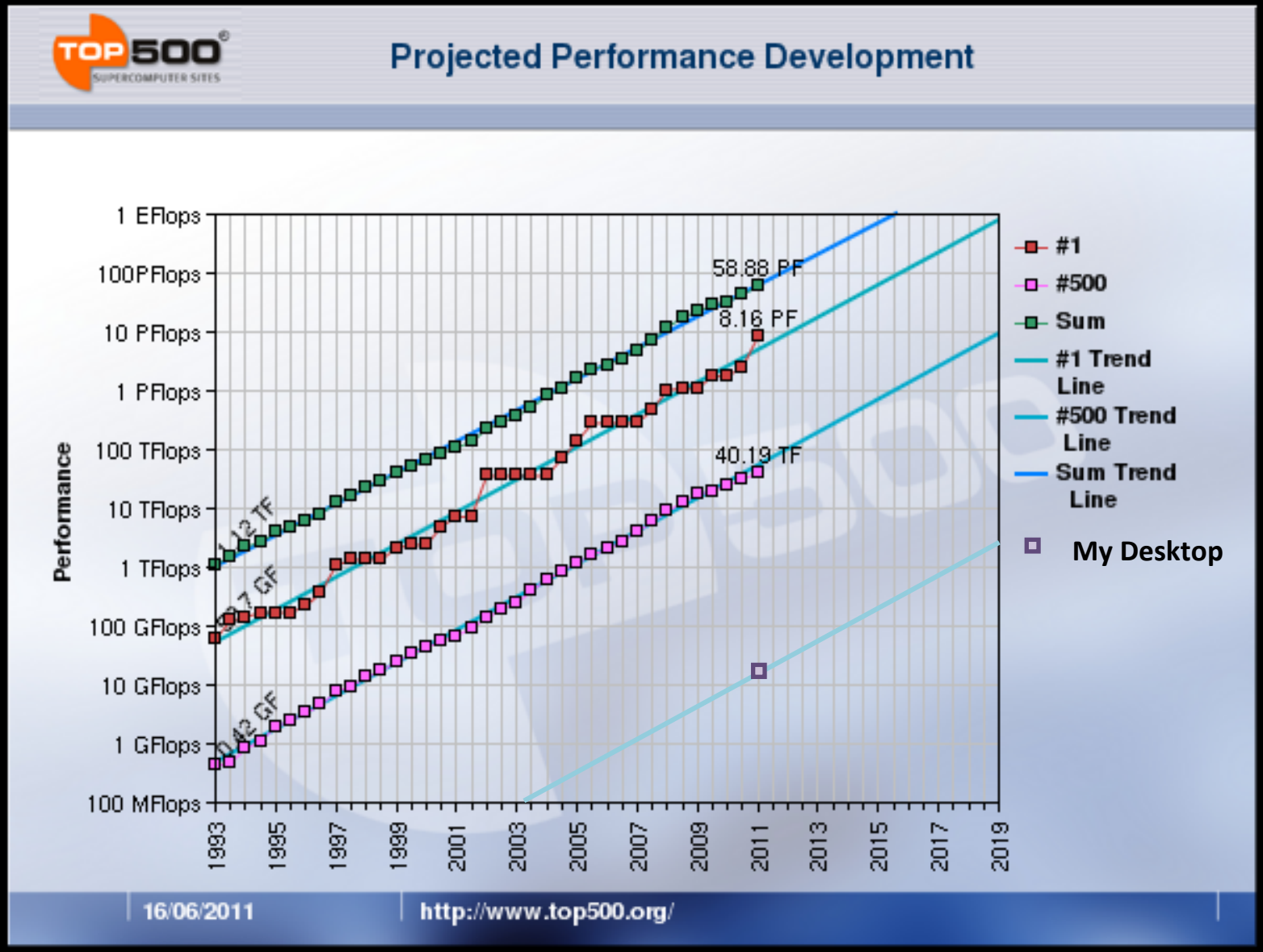
TOP500 SUPERCOMPUTER SITES

Projected Performance Development



16/06/2011 http://www.top500.org/

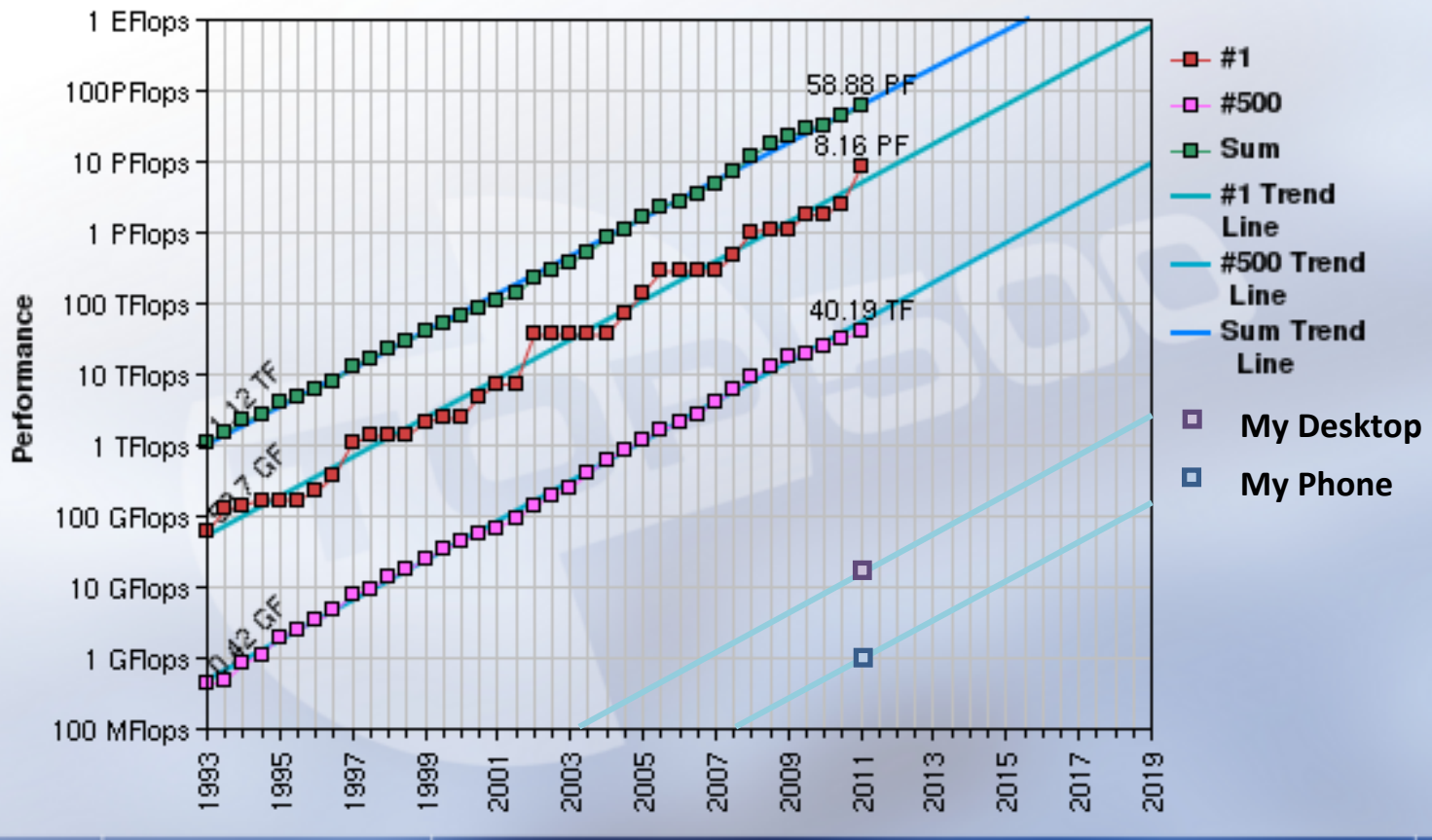
Dell Optiplex
2011
~10 Gflops


TOP500
SUPERCOMPUTER SITES
Projected Performance Development
Performance
1 EFlops
100 PFlops
10 PFlops
1 PFlops
100 TFlops
10 TFlops
1 TFlops
100 GFlops
10 GFlops
1 GFlops
100 MFlops
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
58.88 PF
8.16 PF
40.19 TF
1.12 TF
59.7 GF
0.42 GF
#1
#500
Sum
#1 Trend Line
#500 Trend Line
Sum Trend Line
My Desktop
16/06/2011
http://www.top500.org/

HTC Desire
2011
1 GHz…

TOP500 SUPERCOMPUTER SITES

Projected Performance Development

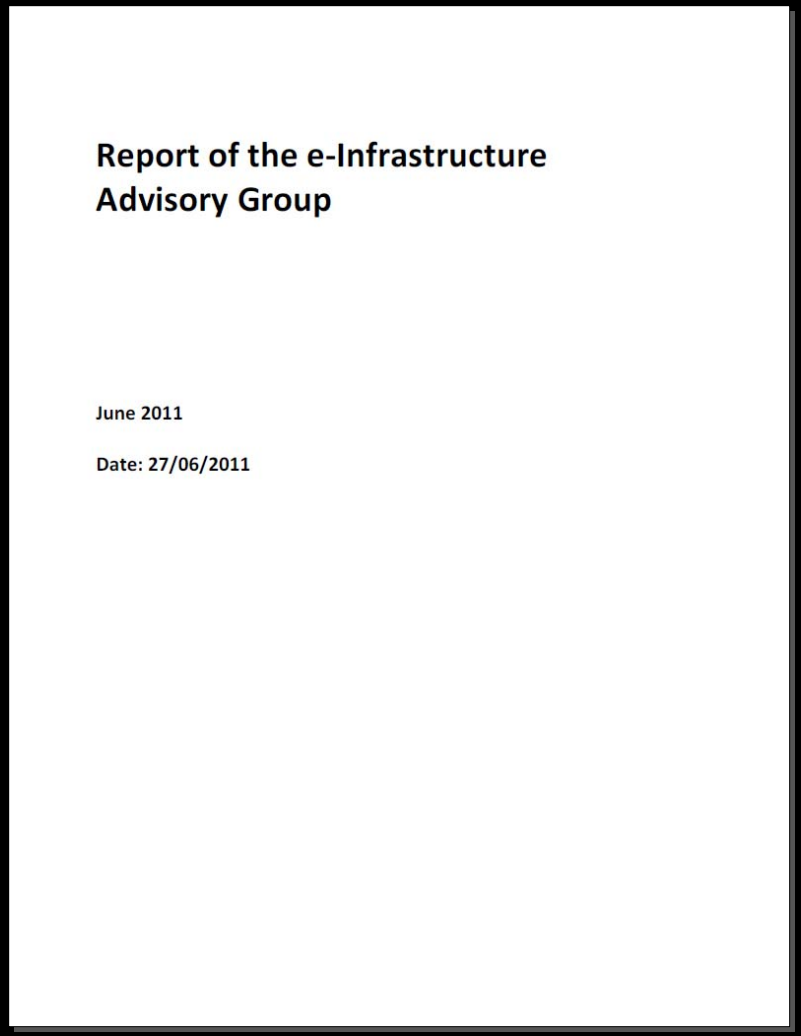


16/06/2011 http://www.top500.org/

‘e-Infrastructure’

2011 UK BIS/RCUK Review identified these key areas:

- Networks
- People & skills
- Data
- Computation
- Software
- Authentication & Security



Report of the e-Infrastructure Advisory Group

June 2011

Date: 27/06/2011

Networks

- Connectivity between processors
  - Gb Ethernet
  - InfiniBand
  - 10Gb Ethernet
  - NUMA memory architectures
- Connectivity to the outside world
  - JANET
  - Parallel NFS, GridFTP
  - The last 10 yards…

People & skills

Where do IT staff come from

- Desktop support – MCSE – Systems Architect
- Computing Science – Software Engineering – Architecture
- Research Science – Systems Administration – HPC

Some skills issues for HPC

- Breadth of understanding
  - Storage, networking, processing, advanced programming
- Parallel, multithreaded, distributed code
- Career paths

Data

Traditional problems were CPU driven

- Simulation
- Numerical analysis

Current drive to HPC has data as its source

- Sensors
- Image capture
- Internet

HPC storage

- High speed, low-latency storage
- High parallel throughput storage
- Scale-out storage

Computation

- HPC is driven by commodity computing
  - Vector processors, SIMD devices no longer mainstream
- Specialist considerations in commodity processors
  - Floating point performance
  - Multi-core architectures
  - GPU, hybrid GPU-CPU
- Constraints
  - Power
  - Memory bandwith, cacheing, I/O
  - Programmability– peak vs observed performance

Software

- Parallelism
  - Advanced topic in most programming courses
  - Multithreading is the barest minimum for multi-core
  - Compilers are clever, but not magic
- Scalability
  - Multi node, multi-core architectures
  - With multilevel caches
  - And add in I/O scalability...
- Is exascale programming even possible?
  - Will we have to re-write our science to match the scale-out algorithms?

Authentication and Security

- You own the computers
  - ...so you give out the passwords
  - A surprising number of systems are still run this way
  - But we need to federate systems, networks, storage
- Certificates/public key infrastructures
  - Based on personal certificates, private/public keys
  - Management overheads with trust hierarchies
- Federated authentication
  - I don't know you and I don't trust you
  - ...but I trust the organisation you work for
  - Eduroam, Shibboleth, JANET Moonshot

Cancer Research UK

Cambridge Research Institute

Cancer Research UK

Cambridge Research Institute

IT &
Scientific Computing

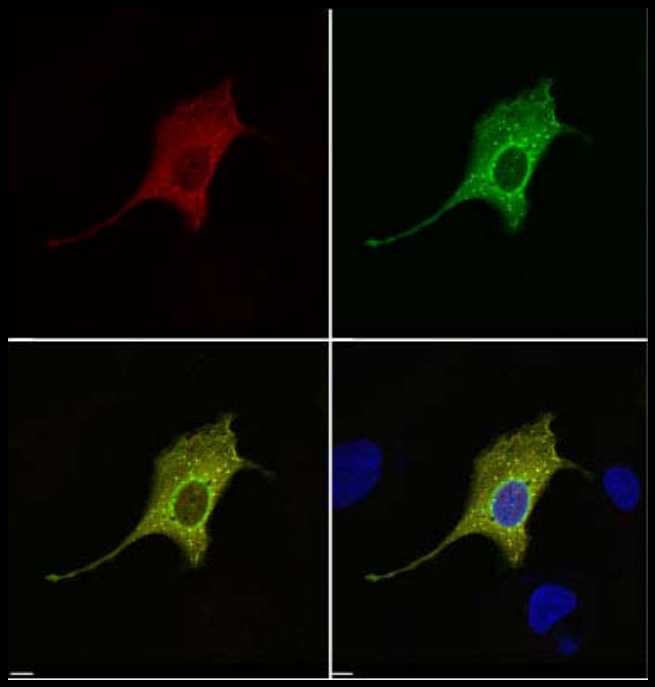
Microscopy

Confocal microscope
high definition, multidimensional image data

Depth – z-stacks generate 3D data

Time – time series of organelle migration

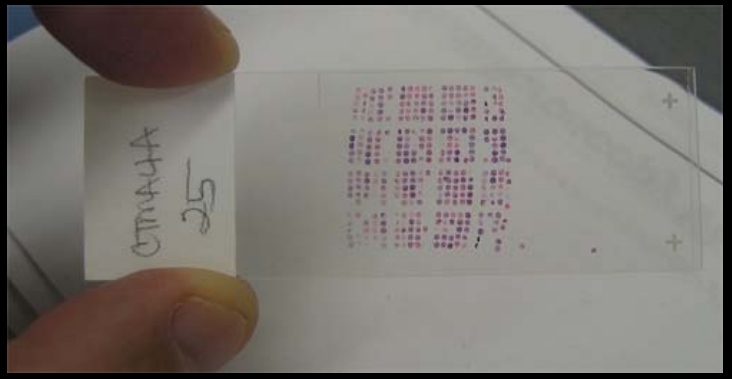
Colour – differential fluorescence to Identify location of molecules



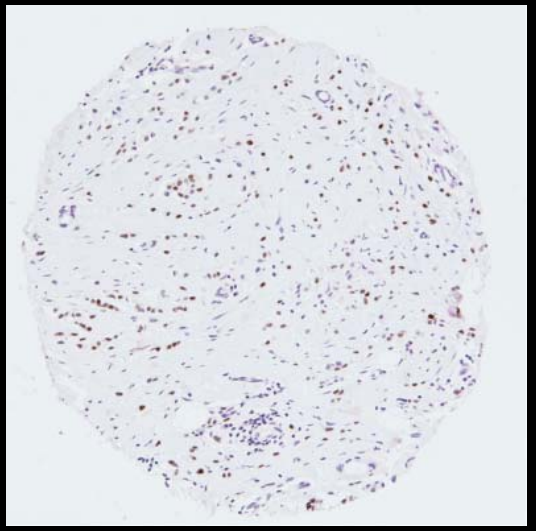
Histopathology

Robotic image scanner

Tissue microarrays – hundreds of samples per slide



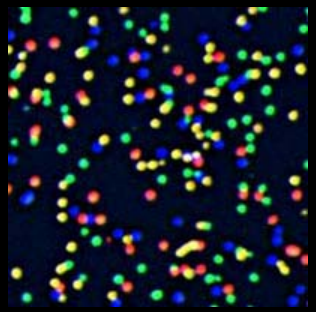
Each sample scanned at high resolution



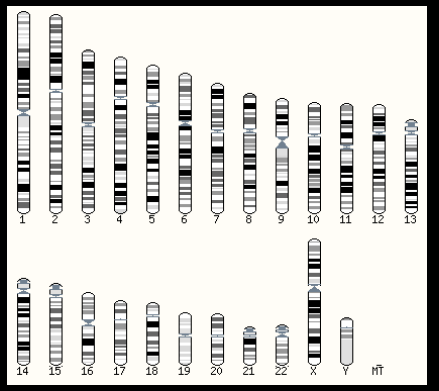
DNA sequencing

Illumina Genome Analyser

Parallel Sequencing of DNA fragments



Rapid sequencing of whole individuals,
Detailed studies of cellular processes



Networks at the CRI

Gb Ethernet

- in HP blade enclosures
- 4:1 oversubscription
- Would prefer IB…

100TB of sequence data

- …via sneakernet

People and skills at the CRI

- Traditional HPC route
  - Chemist
  - Physicist
  - Medical Physicist
- Computer Scientists
  - Engineer … computer scientist
  - Computer Scientist x3

Data at the CRI

Lustre

Parallel distributed FS

96TB

Ibrix

Scale out FS

450TB + + +

Computation at the CRI

Xeon Westmere

64x2x4 core HP BL460

768 cores in 60u

Xeon Sandy Bridge

16x2x2x6 core HP BL2/220c

384 cores in 10u

Platform LSF scheduler

Software at the CRI

- Codes
  - Image processing
  - Systems biology
  - Perl, Python, shell scripts…
- How do we optimise without becoming a software house?
  - Resource optimisation
  - Code optimisation
  - Algorithm design
  - Hardware selection
  - Frameworks
  - Training

Authentication and Security at the CRI

Option 1
Guest accounts + ssh

Option 2
JANET Moonshot
AD federation via GSS

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

High Performance Computing

Is a relative term

Today's HPC developments will become mainstream

Depends on people as much as technology

There's never a stable, commodity HPC platform

Asks us all some hard questions

Will our current approaches to software scale to take advantage of next generation HPC?