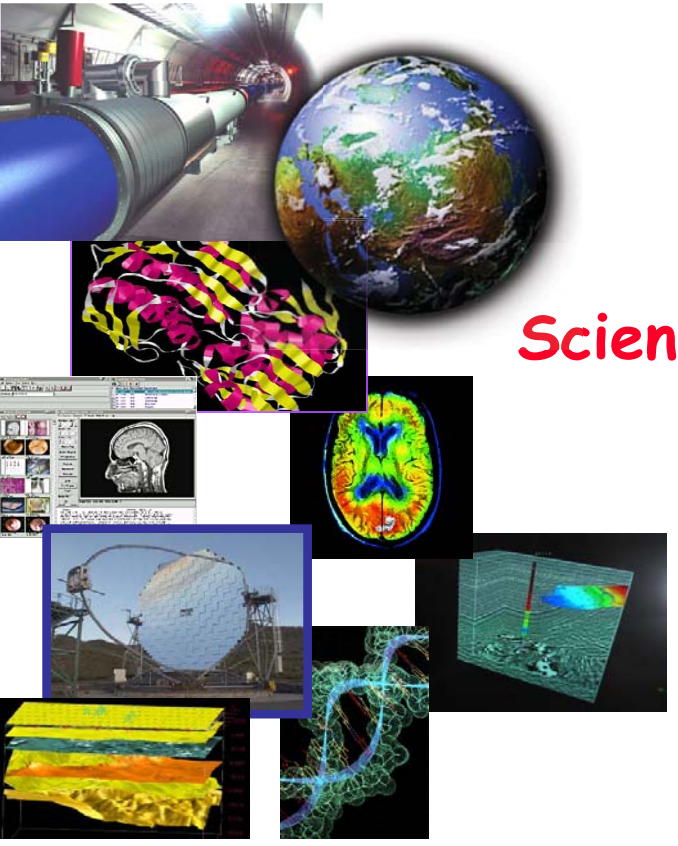


Sciences on the GRID

Chris Jones
CERN


CERN
where the Web was born…
Is most actively engaged on an international and multi-disciplinary development of software, networking and infrastructure…
From the Web to the GRID…

The Vision of the GRID…

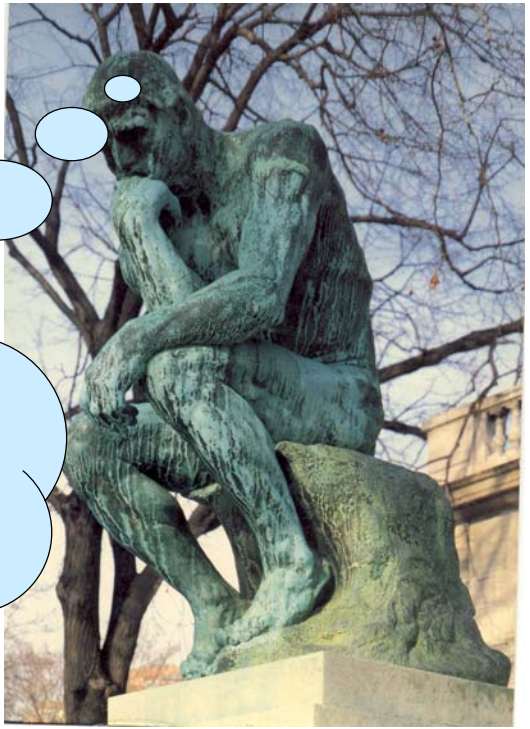
Extends the capabilities of DISTRIBUTED COMPUTING

Beyond those of the Web

to truly WORLDWIDE and intimate interconnection of

COMPUTERS and all their RESOURCES and SERVICES

Why are we developing GRID?



These physicists are never happy!

Isn't the Web enough?

CERN - The Large Hadron Collider 4 detectors

Requirements for data analysis

Storage –

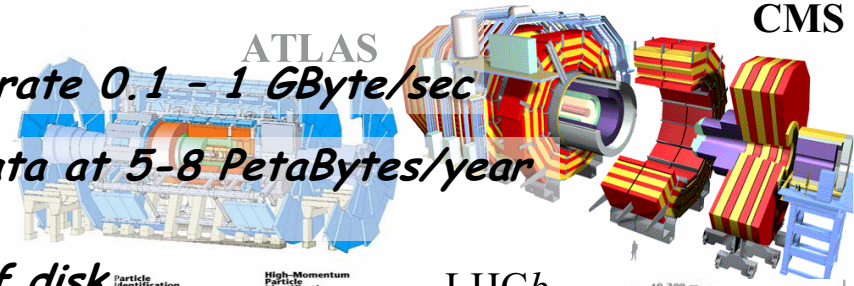
Raw recording rate 0.1 – 1 GByte/sec

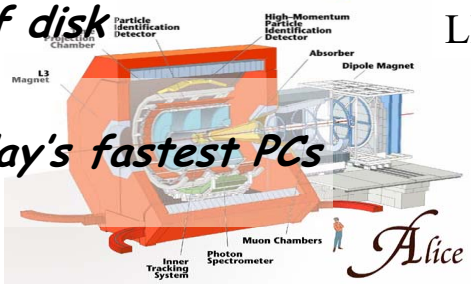
Accumulating data at 5-8 PetaBytes/year (plus copies)

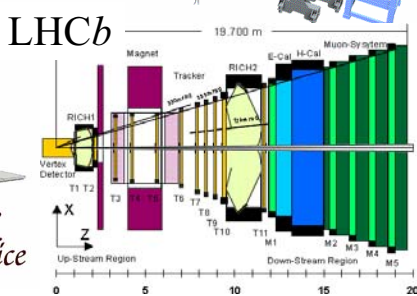
10 PetaBytes of disk

Processing –

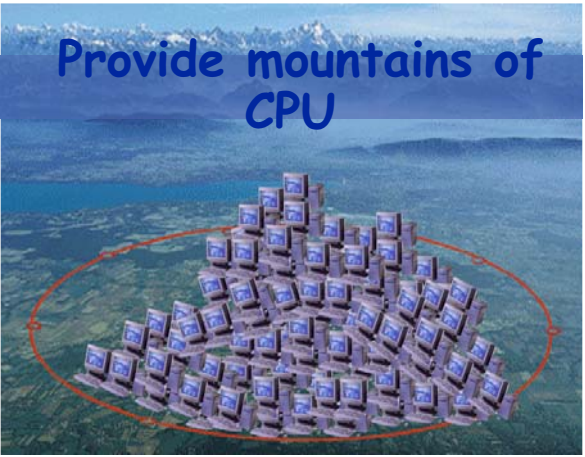
100,000 of today's fastest PCs







LHC Computing



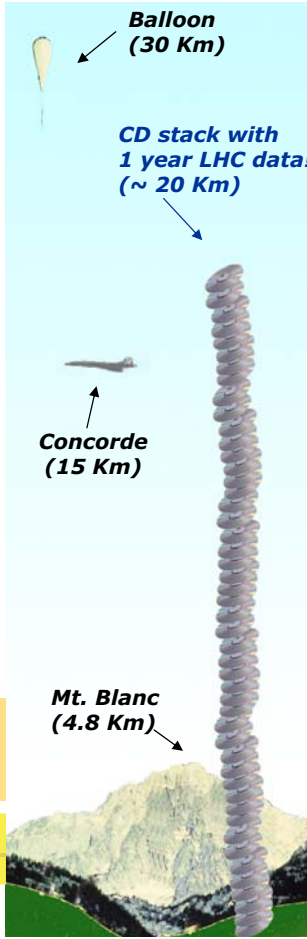
Annual data storage:

12-14 PetaBytes/year

For LHC computing, some 100 Million SPECint2000; 100K of today's PCs

Calibration, Reconstruction, Simulation, Analysis, Fabric and Grid software

Millions of lines of "interfaced" code

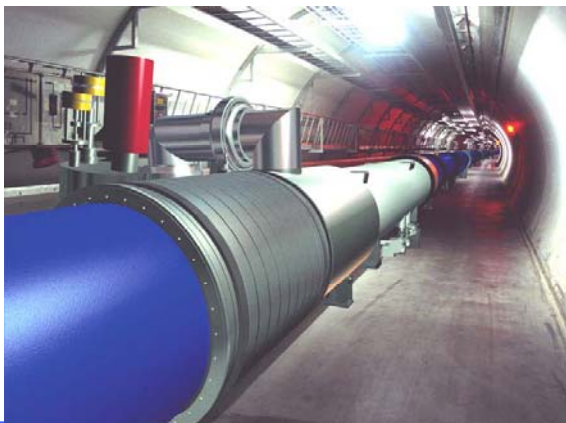


CERN and the GRID

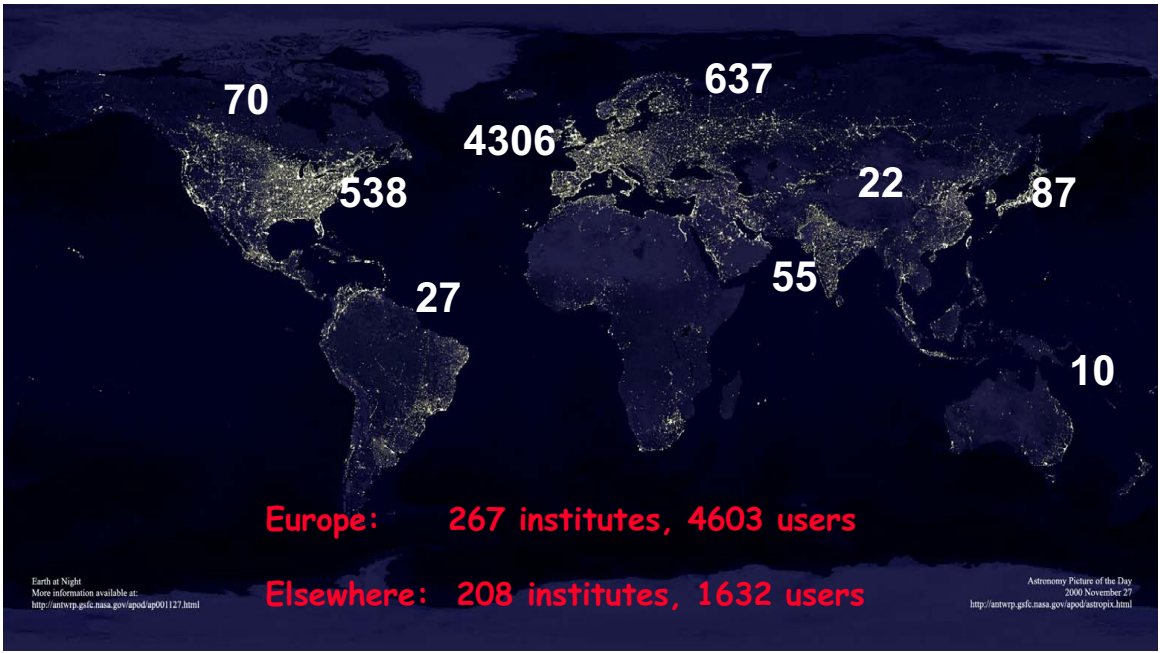
CERN cannot physically install 100 000 computers for LHC

Nor the required number of disks…

But…
we have many collaborators…

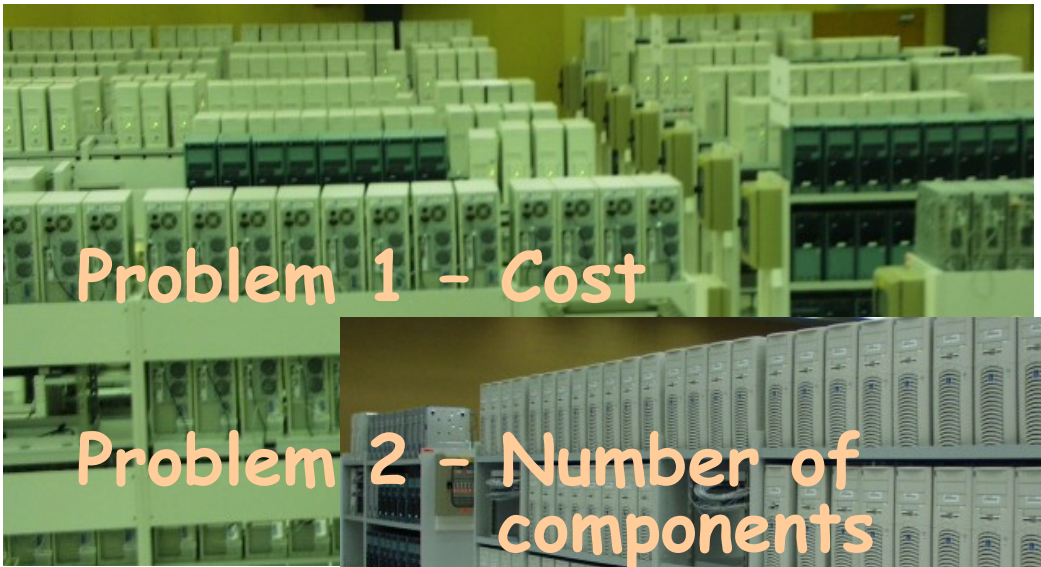


CERN's Users and Collaborating Institutes



CERN's Users and Collaborating Institutes

Europe: 267 institutes, 4603 users
Elsewhere: 208 institutes, 1632 users

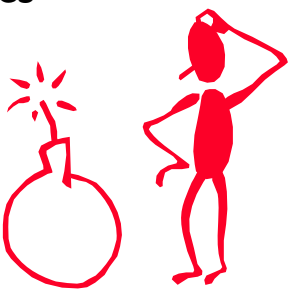


Problem 1 – Cost

Problem 2 – Number of components

LCG: The LHC Computing Grid Project

- a geographically distributed computing facility
- for a very large user population of independently-minded scientists
- with independent ownership/management of the different nodes
- each with different access and usage policies
- and serving multiple user communities



What you might like to have seen…

reliable

available

powerful

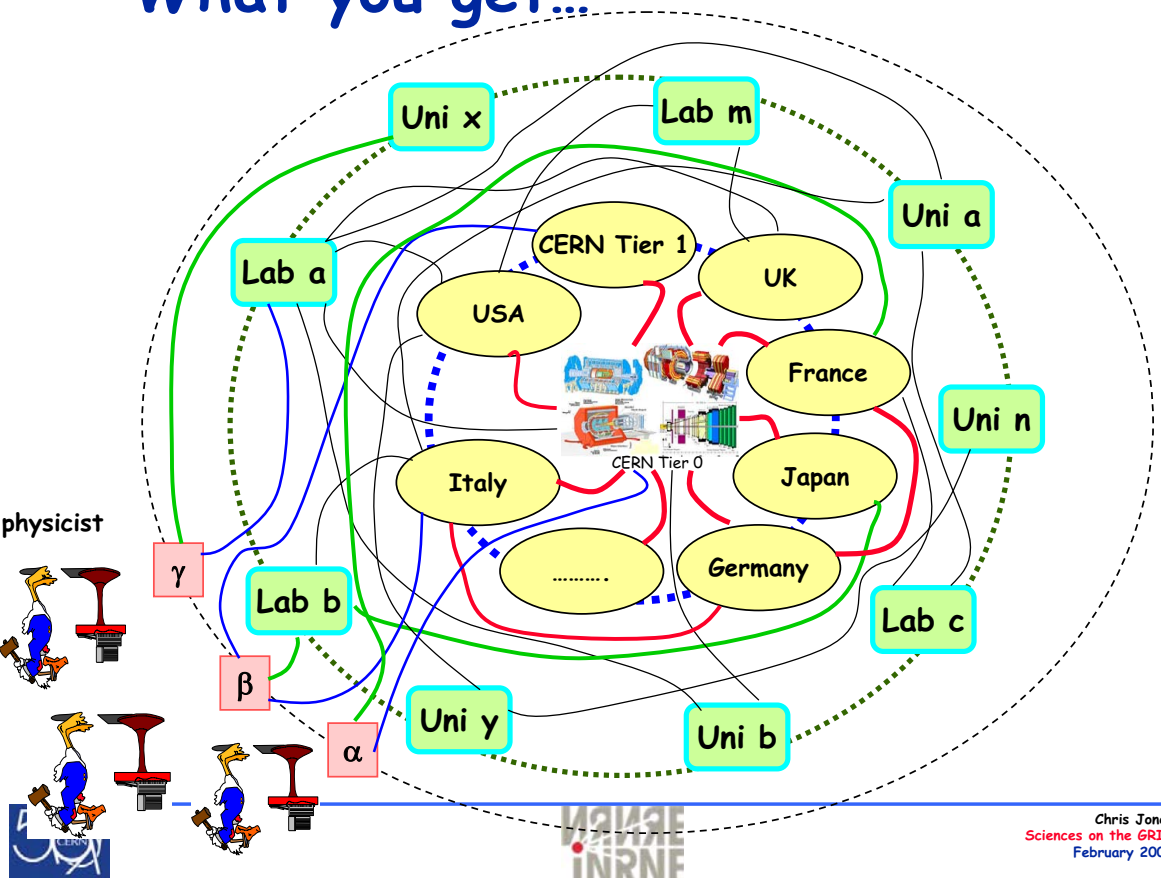
calm

cool

easy to use

…. and nice to look at

What you get…



les.robertson@cern.ch

The Bottom Line

If the LHC experiments are to analyse the data they collect ..

This worldwide GRID of computers has to work by the standards of this community

Which is a serious statement

The LHC Worldwide GRID has to Work!

The Wellcome Trust Genome Campus

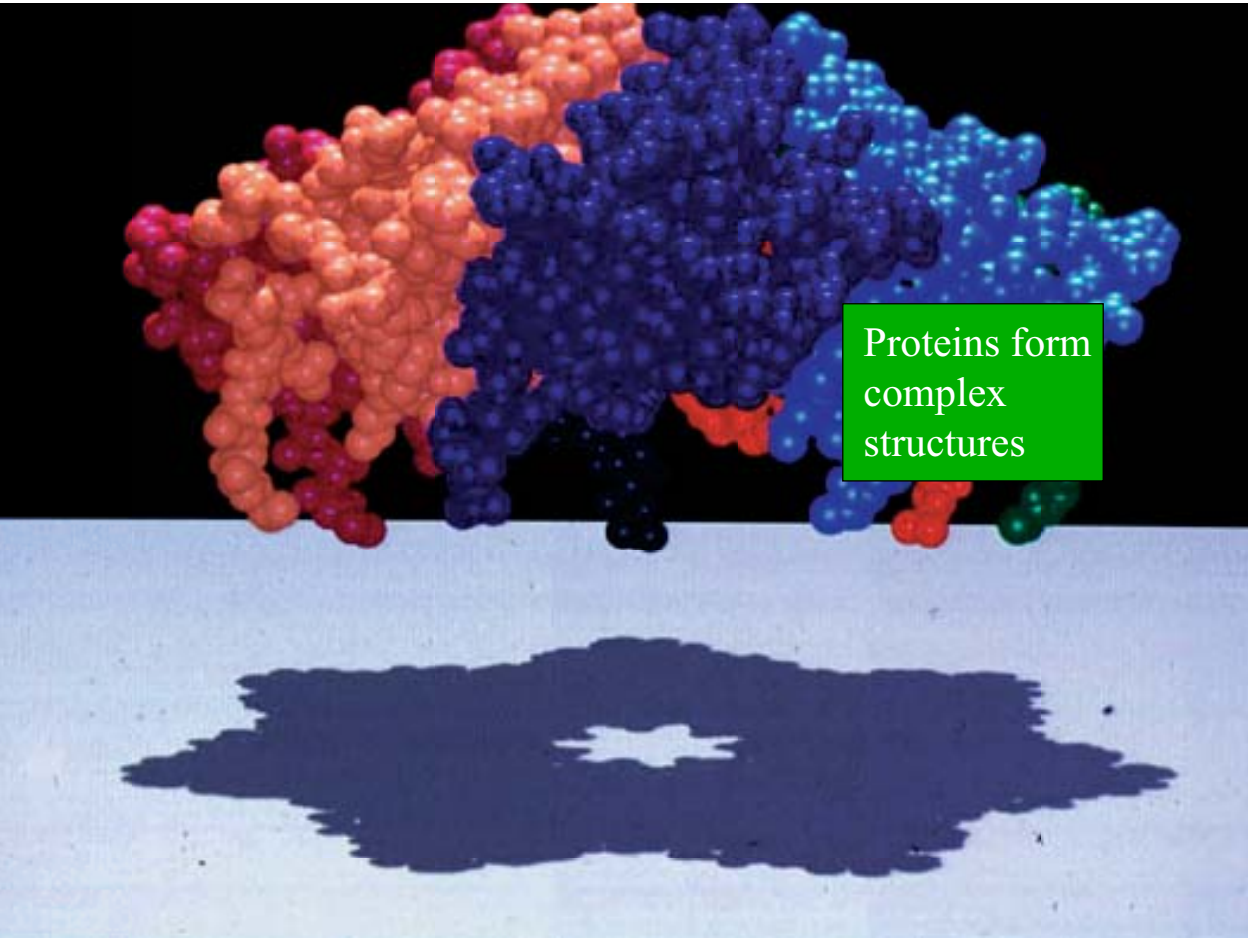
European Bioinformatics Institute - EMBL

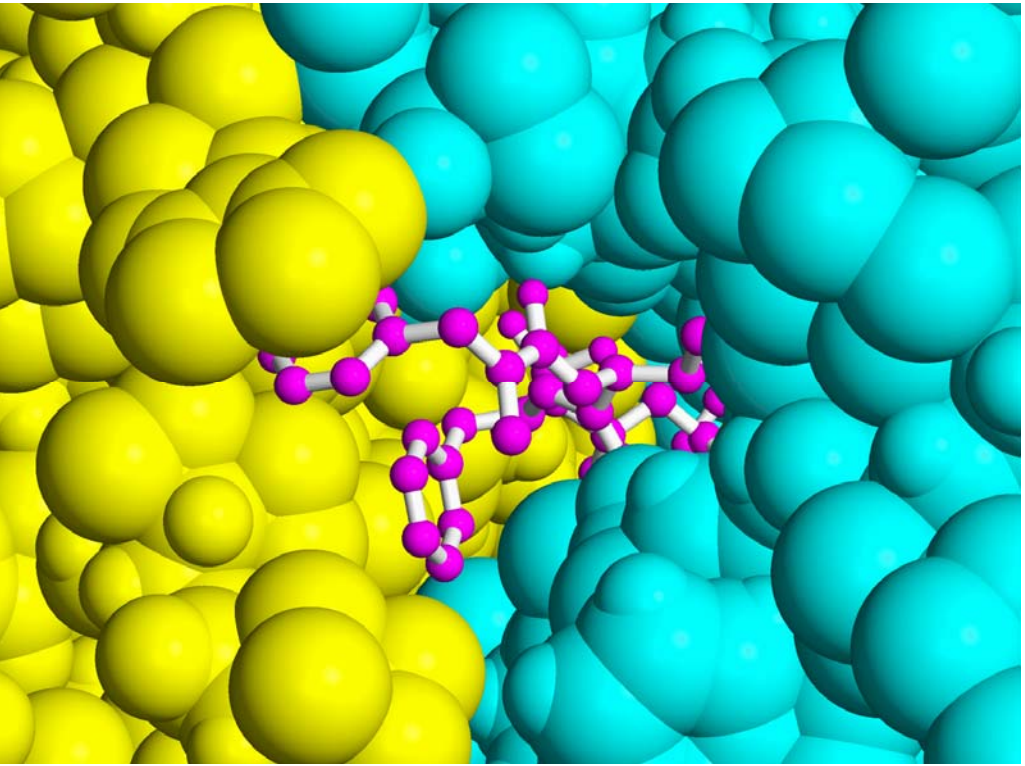
DNA sequence and genes…..

cacaattacttccacaaatgcagtt
gaagcttctactcttcttgcatagg
taacctgagtcggagcagttttcct
cgtggcttcatctttggtgctggat
cttcagcataccaatttgaaggtgc
agtaaacgaaggcggtagaggacca
agtatttgggataccttcacccata
aatatccagaaaaaaataagggatgg
aagcaatgcagacatcacggttgc

make proteins

MRNSYRFLASSLSVVVSLLLIPEDVCEKII
GGNEVTPHSRPYMVLLSLDRKTICAGALIA
KDWVLTAAHCNLNKRSQVILGAHSITREEP
TKQIMLVKKEFPYPCYDPATREGDLKLLQL
TEKAKINKYVTILHLPKKGDDVKPGTMCQV
AGWGRTHNSASWSDTLREVNITIIDRKVCN
DRNHYNFNPVIGMNMVCAGSLRGGRDSCNG
DSGSPLLCEGVFRGVTSFGLENKCGDPRGP
GVYILLSKKHLNWIIMTIKGAV

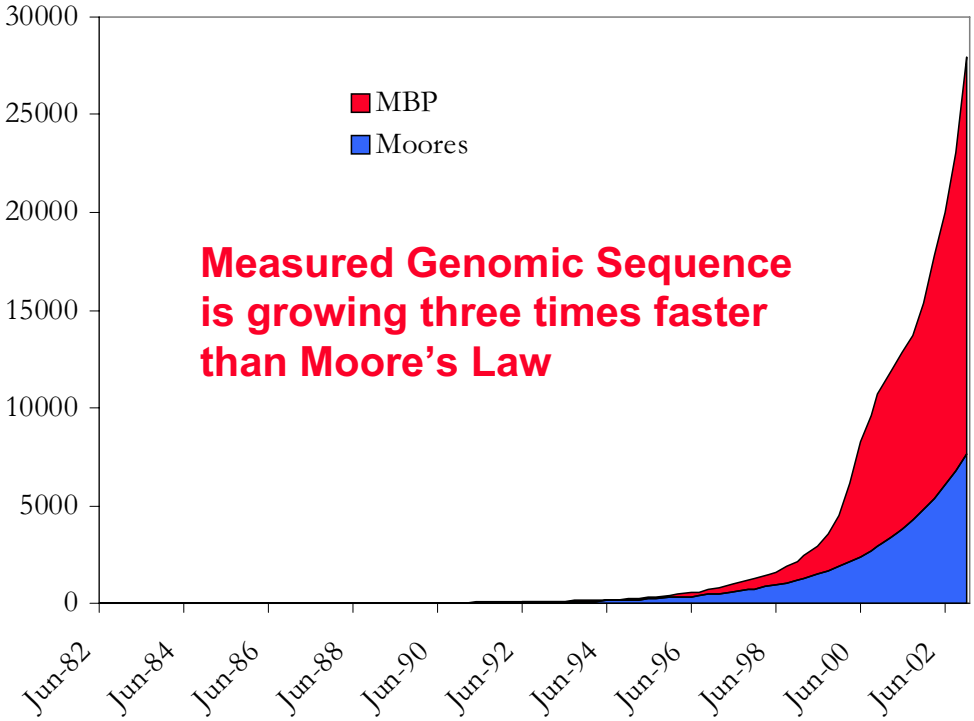

Proteins form
complex
structures



molecules interact

PDB code 1DIF
HIV-1 Protease/Inhibitor Complex
A79285 (Difluoroketone)

The Genomic Data Explosion



Courtesy:
Graham Cameron

Moore's Law

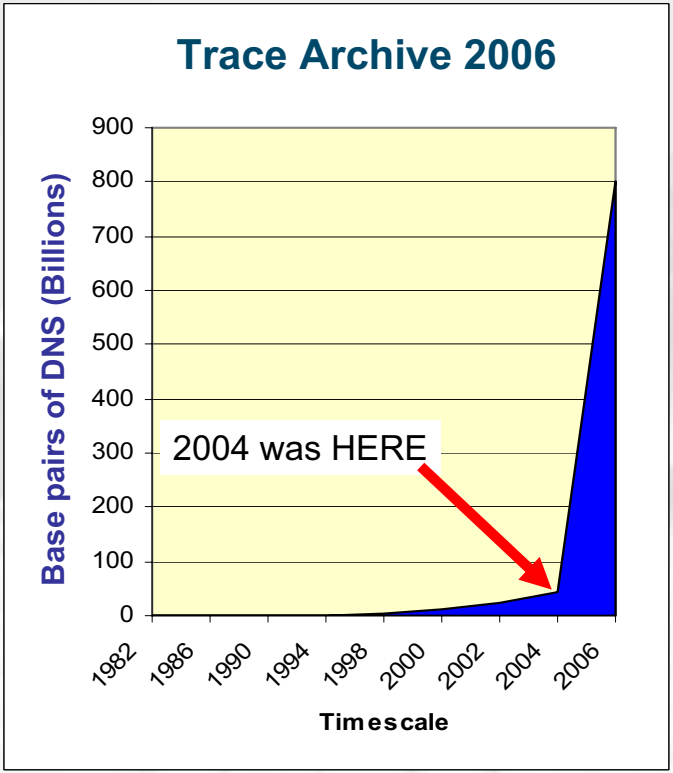
- In the 1960's Intel co-founder Gordon Moore made statements about transistor density relating to predictions in the increasing rate of CPU speeds.

- Over the following years this became folklore and roughly translated into :-

CPU performance approximately doubles every 18 months

- In the past two decades technological improvements have kept this idea alive.

- ***However, over the past ten years data acquisition rates in Life Sciences have increased far ahead of Moore's Law***



TRACE ARCHIVE

Genetic Sequence Database
Wellcome Trust Sanger Institute AND
National Institute of Health
(synchronised databases)

a collection of all publicly available DNA sequences.

2004

Base pairs 44,575,745,176
Sequences 40,604,319

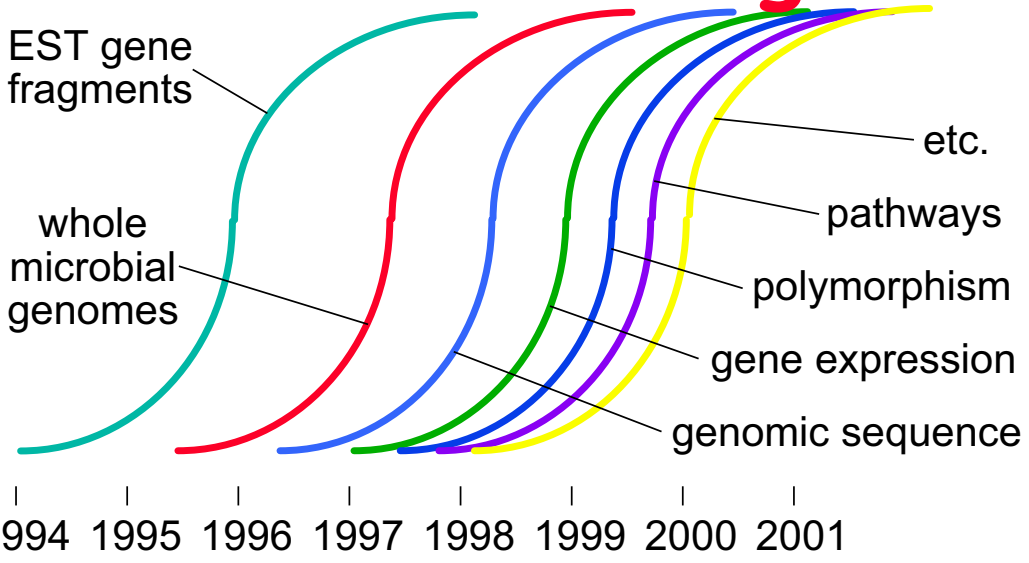
2006

Base pairs 800,000,000,000
Sequences 1,000,000,000

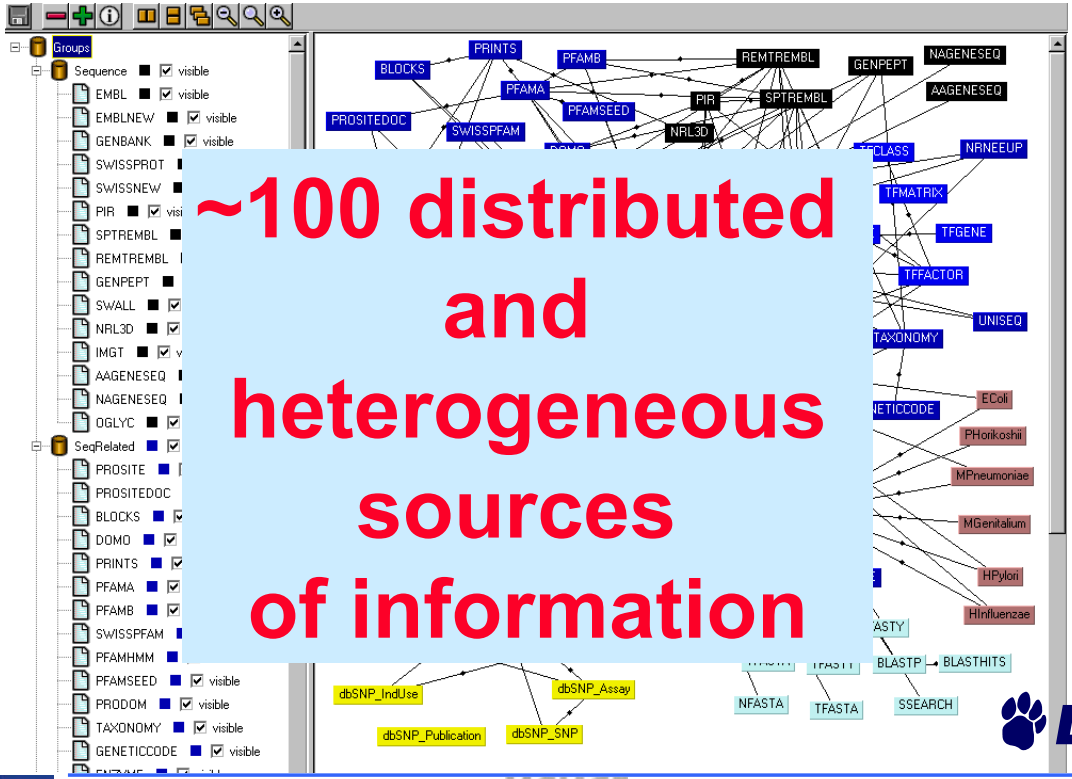
wellcome trust sanger institute

The Flood of Bioinformatics Data - Wave after Wave

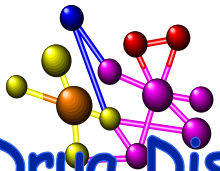
"The Data Deluge"



The World of a Bioinformatics Researcher

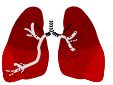


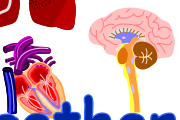
Drug Discovery – The Numbers Game



thousands

billions





dozens

polymorphism
patient records
epidemiology

secondary structure
tertiary structure

expression patterns
physiology

linkage maps
cytogenetic maps
physical maps

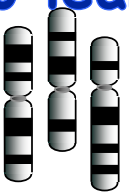
sequences
alignments

receptors
signals
pathways

Drug Discovery is bringing together information from these heterogeneous sources in order to learn something new

ATGCAAGTCCCT
AAGATTGCATAA
GCTCGCTCAGTT

billions



dozens

trillions

Two Sciences - Common Factors

Different sciences, different problems, – but common factors

Digital detectors…data explosion
Need to bring together resources from outside one institute
World-wide Collaboration, Virtual Organisations, Sharing

Distributed computational resources that work together
Faster networks that make new things possible.
Computing has become all pervasive
(but do not forget the "Digital Divide")…

We need effective
DISTRIBUTED COMPUTING…
…The GRID

Distributed Computing ?

The "Holy Grail" for more than 30 years…

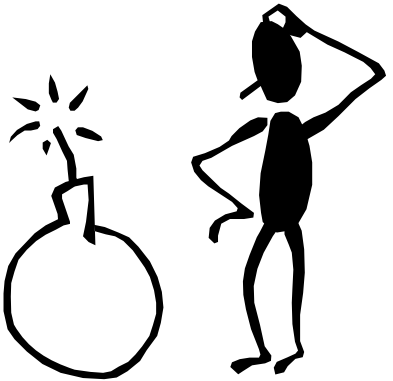
The Web was a huge advance…

Do we really manage Distributed Computing?

Are there big issues?

Do we have the necessary standards?

Are we happy?



The GRID

So is the GRID just the next step
in the inevitable progress towards

Truly Distributed Computing?

If you like!
Its one definition

But is there more to the GRID…?

And what is e-Science?

What do we mean by "e-Science"?

e-Science

science increasingly done through distributed global *collaborations* enabled by the internet

using very large data collections, terascale computing resources and high performance visualisation

Science de plus en plus accomplie avec des *collaborations* distribuées et globales, facilitée par l'internet

Utilisant des énormes collections de données, des ressources de computations « terascale », et des outils de visualisation puissants

Sir John Taylor, UK Head of Research Councils

What do we mean by "e-Science"?

Grid

new generation *information utility*

middleware, software and hardware to access, process, communicate and store huge quantities of data

infrastructure enabler for e-science

Utilité d'information nouvelle génération

Middleware, software et hardware pour l'accès, traitement, communication et stockage d'énormes quantités de données

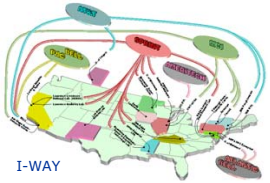
Le GRID facilite « e-science »

La science améliorée par l'informatique du GRID

The Grid - A Vision

A Vision of the future,

shared strongly by many.



I-WAY

NSF PACI's National Technology Grid

National and International Grid Testbeds

Such a COMMON VISION is a very powerful mechanism for launching major efforts…

NASA's Information Power Grid

Ian Foster

Not least because it attracts FUNDING

The Grid - a Source of Funding

EU – 400 Million Euros

UK – 300 Million Pounds

USA – roughly 1 Billion dollars

Plus other countries

Of the order of 2 Billion Dollars

Switzerland has been curiously discrete so far…

The Web

Was a response to the needs of a distributed collaborating community

And saved time and effort in fetching information from other places

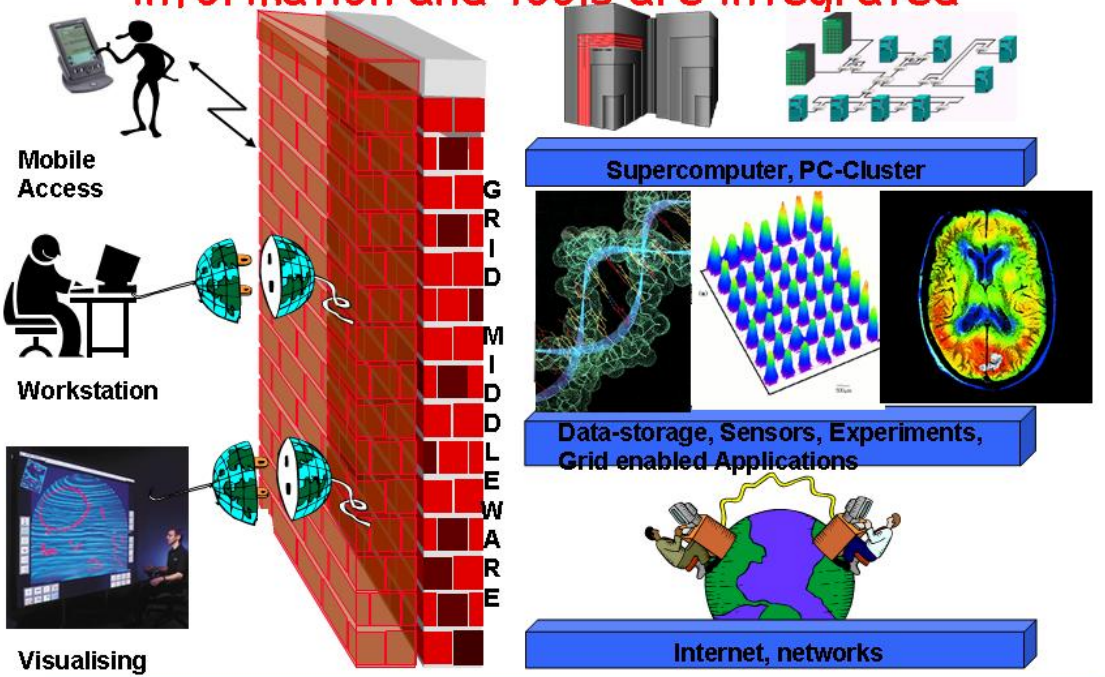
It made sharing information so much easier

That many NEW things became possible

Transparent access to information

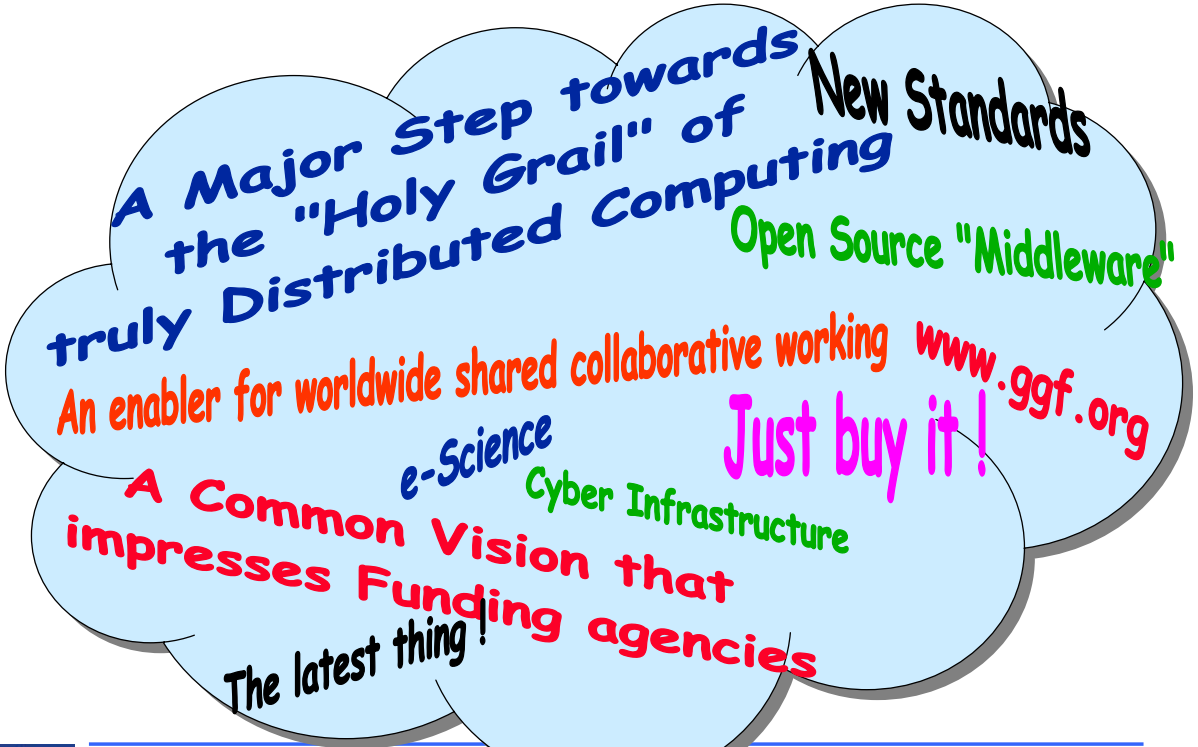
Independent of and removing barriers of space and time

Life Sciences and Health need more middleware for tailored environments in which all relevant information and tools are integrated



Hoffmann, Putzer, Reinefeld

The GRID – what is it?



The GRID: Summary so far:

A remarkably Common Vision

Attracting substantial funding and many projects

In extending distributed computing (beyond local site)

Virtual Organisations (VO's) and World Collaboration

Enhanced science

Enabling new and better ways of working