Linux and Genomics

Two Revolutions

Canada's Michael Smith Genome Sciences Centre

Martin Krzywinski

Yaron Butterfield

Acknowledgements

asano, jennifer : astakhov, vadim : astakhova, tamara : astell, caroline : babakaiff, ryan : bala, miruna : barbazuk, stephen : barber, sarah : baross, agnes : bilenky, mikhail : bosdet, ian : brooks-wilson, angela : brown-john, mabel : butterfield, yaron : chan, susanna : chan, andy : chand, steve : chang, elbert : charest, david : charters, anita : chittaranjan, suganthi : chiu, gordon : chiu, readman : chow, william : chuah, eric : chun, hye-jung elizabeth : collins, jennifer : coughlin, shaun : crisostomo, lulu : del rio, luis : delaney, allen : despres, chantale : dhalla, noreen : farnoud, noushin : featherstone, ruth : field, matthew : fjell, chris : flibotte, stephane : freeman, doug : go, anne : gorski, sharon : griffith, obi : griffith, malachi : guan, jun : guin, ranabir (ran) : halaschek-wiener, julius : hanson, robyn : harrison, isabel : hassel, maik : hirst, martin : holt, robert : hou, claire : huang, peter : hume, lynn : jang, carrie : jendo, aga : jones, steven : khattra, jaswinder : kirkpatrick, robert : krzywinski, martin : kwong, letticia : leach, stephen : lee, darlene : lee, stephanie : lee, lisa : leung, amy : leung, derek : li, bernard : li, yvonne : liao, nancy : lin, keven : liu, jerry : ma, kevin : marcadier, julien : marra, marco : masson, amara : mathewson, carrie : matsuo, corey : mayo, mark : mayo, michael : mcconechy, melissa : mcdonald, helen : mckay, sheldon : melnyk, brianna : messervier, steve : mills, courtney : missirlis, perseus : moksa, michelle : montgomery, stephen : moore, richard : moran, josh : morin, gregg : morin, ryan : narain, abhishek : nayar, tarun : novik, karen : o'connor, katie : oliveira, lisa : olson, teika : oveisi, mehrdad : palmquist, diana : pandoh, pawan : peloso, irene : persaud, deryck : petrescu, anca : pleasance, erin : prabhu, anna-liisa : pugh, trevor : qadir, mohammed : quayle, adrian : robertson, neil : robertson, gordon : roger, jennifer : rogers, sean : ruschkowski, sharon : ruzanov, peter : saeedi, parvaneh : santos, joseph roy : schein, jacquie : schnerch, angelique : schoeffel, kirk : shin, heesun : siddiqui, asim : sipahimalani, payal : sleumer, monica : smailus, duane : sohaib ali, mohammad : stott, jeff : tai, isabella : teague, kevin : tremblay, ashley : tsai, miranda : tsang, eddy : varabei, dmitry : vardy, jill : varhol, richard : warren, rené : wilson, gary : with, sheila : wong, david : wong, kim : wright, charlotte : wye, natasja : yang, george : yuen, wendy : zeng, thomas : zhao, yongjun

JAMES T. KIRK
IN "MIRROR, MIRROR"

two revolutions

- **revolution [n.]**
  - drastic and far-reaching way of thinking and behaving
  - paradigm shift
- promise beneficial change
  - increased quality of life for under-represented
  - independence, broader rights, protection from predatory elements
- who are the (a) under-represented, and (b) predatory elements?
  - in computers
    - under-represented: **the OS user**
    - predators: corporate bodies with monopolies, marketing, FUD
    - my ability to meaningfully participate in development of close-source software is negligible
  - in science
    - under-represented: **individuals**
    - predators: restrictive patents/licenses, large privately-funded centers bereft of public accountability, academic cliques isolated from social context
    - my ability to grasp new findings, evaluate their impact and foresee implications is negligible

No, the whole point of Linux isn't that it's free, and never has been.

The point with Linux is that I didn't have a good OS on my machine ... So I wrote one.

Linus Torvalds
comp.os.minix, 20 Dec 1992

revolution 1 | linux

- UNIX on Intel CPU
  - free OS on cheap hardware
    - performance per unit cost is steadily dropping
    - "free" vs "absolutely free" – what is the real cost?
    - in our lab, computer hardware costs are negligible compared to personnel and laboratory equipment
    - my workstation costs ~ 1/50th of my salary
  - line between user and developer base blurred
  - "Beowulf" clusters are ubiquitous
    - cheap, limitless potential, free toolkits
- community encourages users to experiment
  - communities, not cliques
  - anyone can contribute
    - code, document, test, mirror
- rewards innovation and sharing
  - does not rely on scarcity tokens to drive its economy
  - skill and meaningful contribution is rewarded independent of socioeconomic status, reputation, or charisma

> Ten years out, in terms of actual hardware costs you can **almost** think of hardware as being free – I'm **not saying it will be absolutely free**.
>
> *B. Gates*
> emphasis is mine

revolution 2 | genomics

- genomics, [n.]
  - study of organisms in terms of their genomes
- offers insight to fundamental building blocks of living systems
  - structure, biochemistry, evolution
  - organism as network – system biology
  - understand basic processes in a living cell
- genomics is the branch of golden era of biology
  - less classification, more integrated descriptive and predictive models
  - periodic table of elements, quantum mechanics
- applications are the fruits
  - accelerate drug development by identifying target pathways
  - recognize variants contributing to health and resistance to disease
  - generate diagnostics for early detection of cancer
  - understand reasons for individual drug resistance

> It should be possible to understand the difference between a "bag of molecules" and a biological system.
>
> *F. Collins et al.*
> Nature (2003) 422: 835-847

common philosophy

programming	science
kernel	knowledge
patches	papers
code repository	literature
hacker	scientist
gatekeeper	editor
software	mindware

improvement of the "human condition"

Timo Hannay, Thought Experiment: Science as an Open Source Project
www.oreillynet.com/pub/a/network/2002/07/17/timo.html

common process

programming	science
new code should not crash system	new conclusions backwards-compatible with existing knowledge
add features or elegance	increase range power of explanation and prediction; reduce complexity of fundamental laws
robust and logical	in accordance to observed data
properly documented	reproducible
major rewrites require damn good code	extraordinary claims require extraordinary evidence
advancement	

Timo Hannay, Thought Experiment: Science as an Open Source Project
www.oreillynet.com/pub/a/network/2002/07/17/timo.html

embracing openness

- success in science requires open source principles
  - examine, verify and extend knowledge
  - knowledge must be publicly accessible to promote discovery
    - many eyes make all bugs shallow
  - scientists have created many open source tools
    - communication and collaboration
    - manipulate, munge, analyze large data sets
    - agree, adopt and improve

- integration of science into society requires open source principles
  - scientific process is not suitable for handling emergent ethical, legal and social issues (ELSI)
  - it is impossible to “test” wide-scale social models
  - social harm caused by misguided policies cannot be reversed
  - keeping science and its products open allows everyone to participate in ELSI discussions
    - experts from social sciences, law, humanities, cultural anthropology
    - public forums

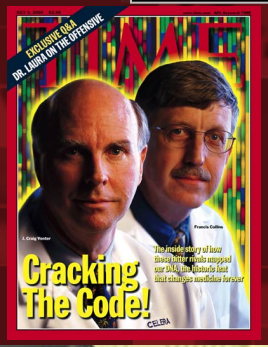
public effort triumphs

- public assembly of human genome done on UCSC centicluster
  - June 2000
  - 100 800-MHz P3 Linux boxes

UCSC kilocluster
1024 Linux nodes

I thought it would help to get as much information about genes and the genome into the public domain to help discourage people from patenting it wholesale.

Jim Kent



Venter, Collins
TIME July 3 2000

Venter, Clinton, Collins

www.cbse.ucsc.edu/UCSC_Kilokluster.html, biomedical.ucsc.edu/Computation.html

ACTGs of genomics

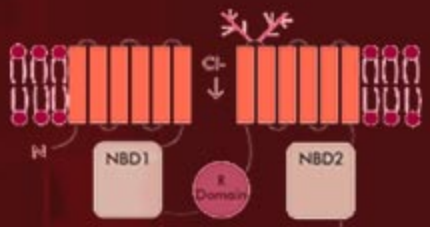
transcription

translation

proteins

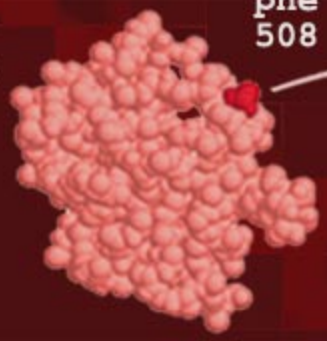
> chr 7, 159 Mb

...
ATTATGCCTGGCACCA
TTAAAGAAAATATCAT
CTTTGGTGTTTCCTAT
GATGAATATA...



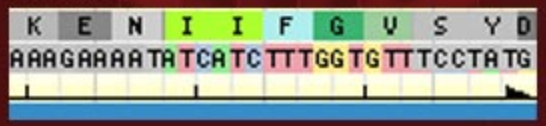
proteins participate
in biochemical pathways
and mediate processes

> alter molecules
> transport molecules
> catalyze reactions

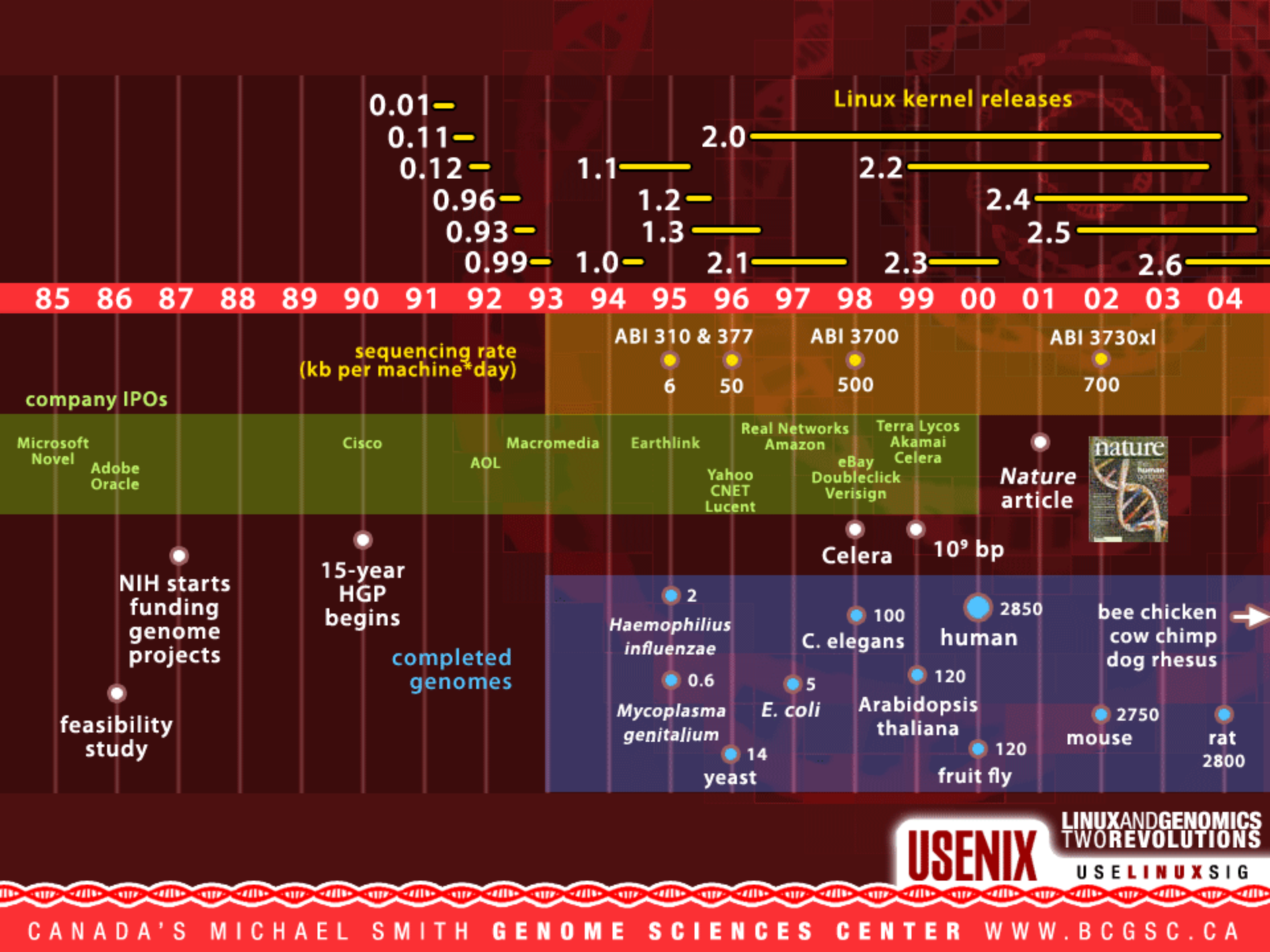


membrane transporter

normal



cystic fibrosis

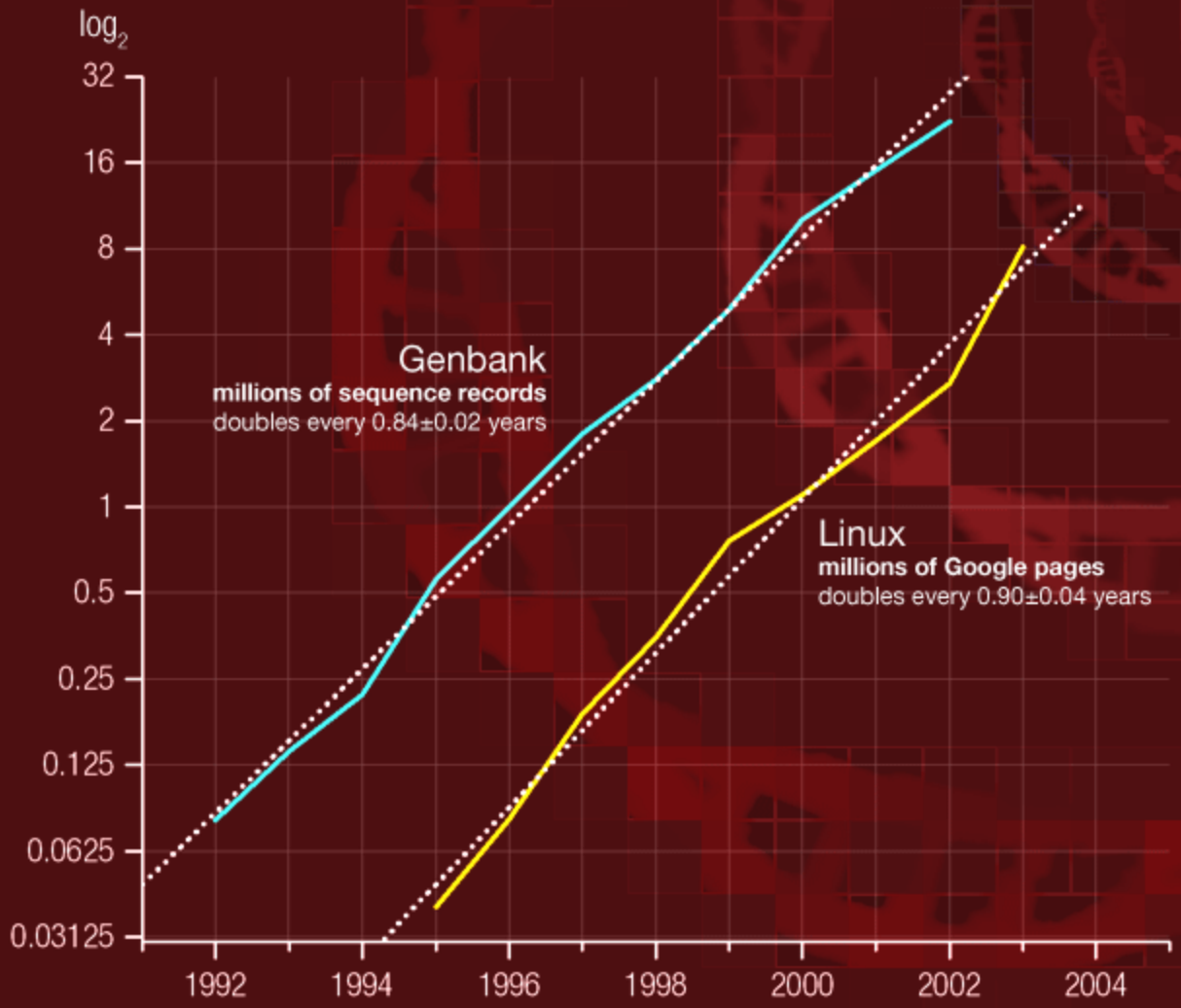


USENIX

LINUX AND GENOMICS TWO REVOLUTIONS

USELINUXSIG

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTER WWW.BCGSC.CA


log2
32
16
8
4
2
1
0.5
0.25
0.125
0.0625
0.03125
1992
1994
1996
1998
2000
2002
2004
Genbank
millions of sequence records
doubles every 0.84±0.02 years
Linux
millions of Google pages
doubles every 0.90±0.04 years

Scientific progress will be maximized by early, open and continuing access to large data sets.

...rewarding and protecting the interests of scientists who wish to share their data with the community in such a generous manner.

Francis Collins

A vision for the future of genomics research, Nature (2003) 422: 835-847

genomics data is public

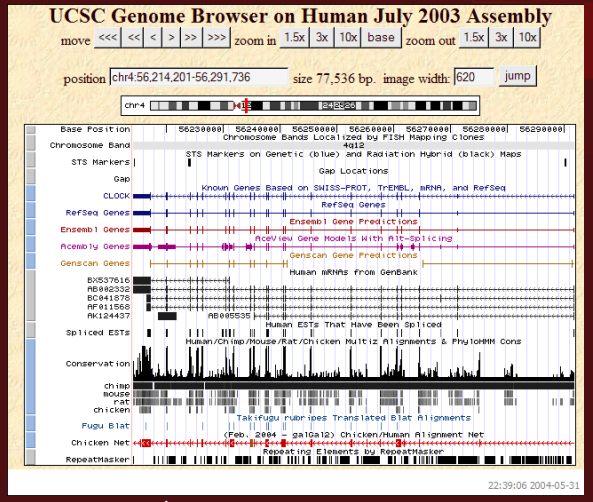
- the net started at CERN
  - physicists needed to share documents and large datasets
  - genome centers and scientists have been networked from the start
- public data release is a mandate of sequencing centers
  - data is submitted to Genbank, a public repository of sequence information, as soon as it is collected
  - "Bermuda Principles" ratified during a 1996 sequencing conference, call for **automatic and rapid** release of primary sequence data to the public domain
  - internet is the natural forum

www.ncbi.nlm.nih.gov/ www.ebi.ac.uk/embl

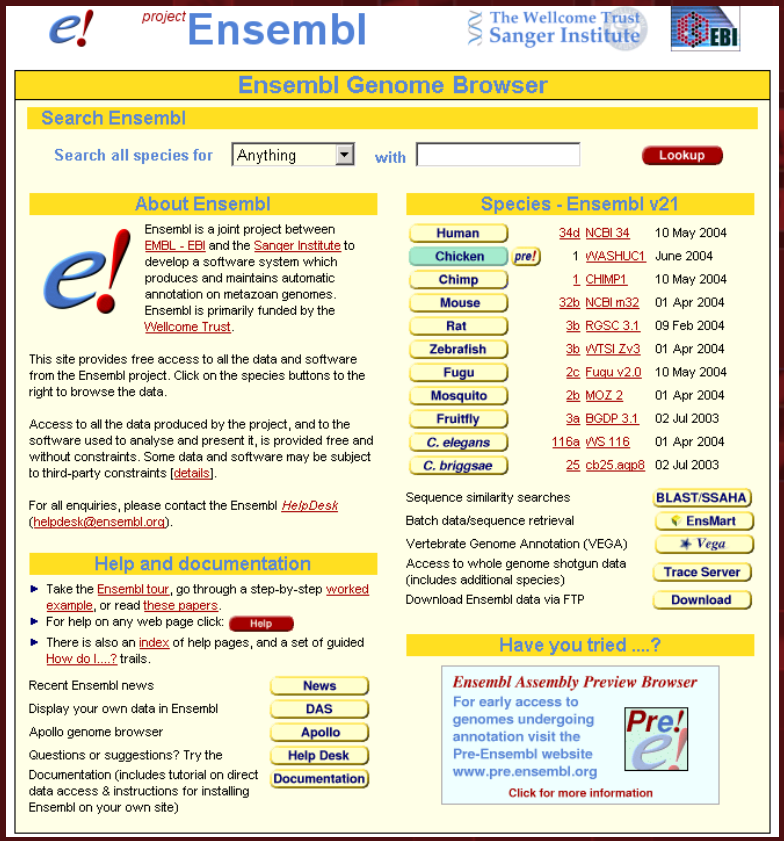
www.gene.ucl.ac.uk/hugo/bermuda.htm

data viewers are public

- openly available genome browsers
  - permit public mining of information
  - visualization of annotations
- Ensembl and UCSC browsers run on MySQL
  - entire data set can be downloaded freely
  - you can become an Ensembl mirror
    - ~100 Gb, MySQL/Perl



genome.ucsc.edu



www.ensembl.org

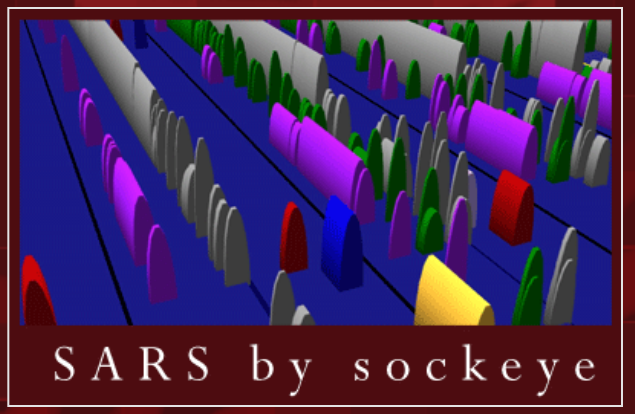
genomics community is online

- Lincoln Stein, CSHL (NY, USA)
  - CGI.pm, GD.pm
  - generic genome browser
  - genome knowledge base
  - generic model organism database construction set
    - modular and extensible framework for storing biological information
  - distributed sequence annotation system (DAS)
    - XML web service for contributing sequence information
  - "How Perl Saved The Human Genome Project"
- Jim Kent, UCSF (CA, USA)
  - Gigassembler algorithm assembly algorithm
    - open source, Linux cluster
  - public effort beat Celera by 3 days
- Ewan Birney, Wellcome Trust (UK)
  - Ensembl browser
  - BioPerl bundle

SARS – linux called to action

- sudden acute respiratory syndrome
  - first case in 1999
  - 8,100 cases, 800 deaths since Nov 2002
- our center was the first to publicly release the full coronavirus sequence assembly
  - 30kb ~ 1/100,000 size of human genome
  - assembled on 8-way Linux machine
  - published on Zope/Apache web server
  - sequence data used to characterize virus and search for potential therapies

SARS affects everyone
www.stileproject.com

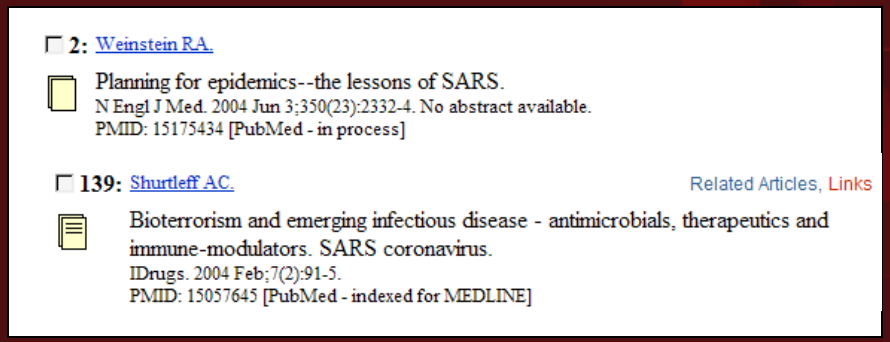


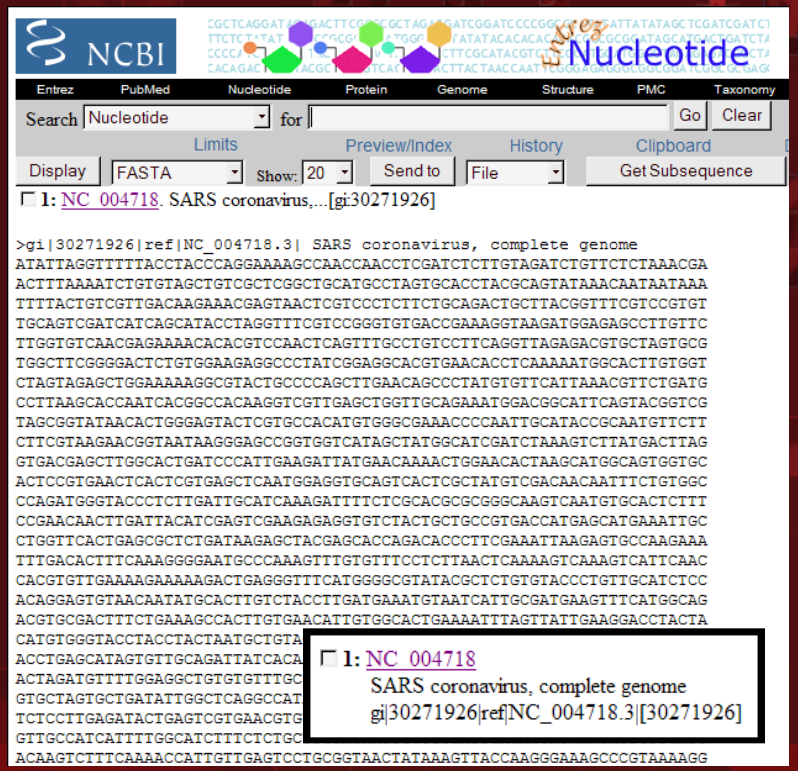
sockeye 3D genome browser

www.who.int/csr/sars/country/table2004_04_21/en/

open data - SARS genomes in Genbank

- complete genomic sequence was immediately submitted to Genbank
- 293 sequences from 151 different strains of SARS in Genbank database
- currently >1,600 SARS-related publications indexed by Pubmed





www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=30271926
www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=227859

Slashdot

News for Nerds. Stuff that matters.

faq
code
awards
journals
subscribe
older stuff
rob's page
preferences
submit story

Canadian Lab Unravels SARS With A Beowulf Cluster

Posted by timothy on Sunday April 13, @03:21PM

from the you-knew-they'd-come-in-handy dept.

Amad writes *"A Canadian Genetics Research Lab in BC, Canada has used a Li*
sequence the genetic code of the virus linked to SARS. This lab is the first to cra
the public. You can read an article about the discovery, or check out the lab."

B.C. team unravels SARS

By Charlie Anderson
Staff Reporter

in quotes

Michael Smith Genome Sciences Centre, and Yaron Butterfield, assistant co-
RS project at the B.C. Cancer Agency.

SARS cases

Singapore: 140 cases, nine deaths
Taiwan: 23 cases, 0
U.S.: 166

Open-Source Desktop Publishing: Scribus

LINUX JOURNAL

LINUX JOURNAL READERS' CHOICE 2003

GENOME SCIENTISTS REPORT:

How we sequenced the SARS virus in five days

Virtual security zones

Highly available cluster management with OSCAR

Secure IMAP mail servers

Disaster plans for your web site

Design C++ programs for security

1U servers

JUNE

TOP STORY

SARS BUSTERS

How the virus crackers got started.

By Lila MacLellan

expecting only more of the same.
PR agents estimate they responded to 150

USENIX

LINUXANDGENOMICS
TWOREVOLUTIONS

USELINUXSIG

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTER WWW.BCGSC.CA

public feedback

Subject: transcription starting position of genes in TOR2

Dear BCGSC:
I downloaded your sequences of TOR2. Thank you very much for your great work about this.

JZ, Palo Alto, CA

Subject: SARS draft sequence

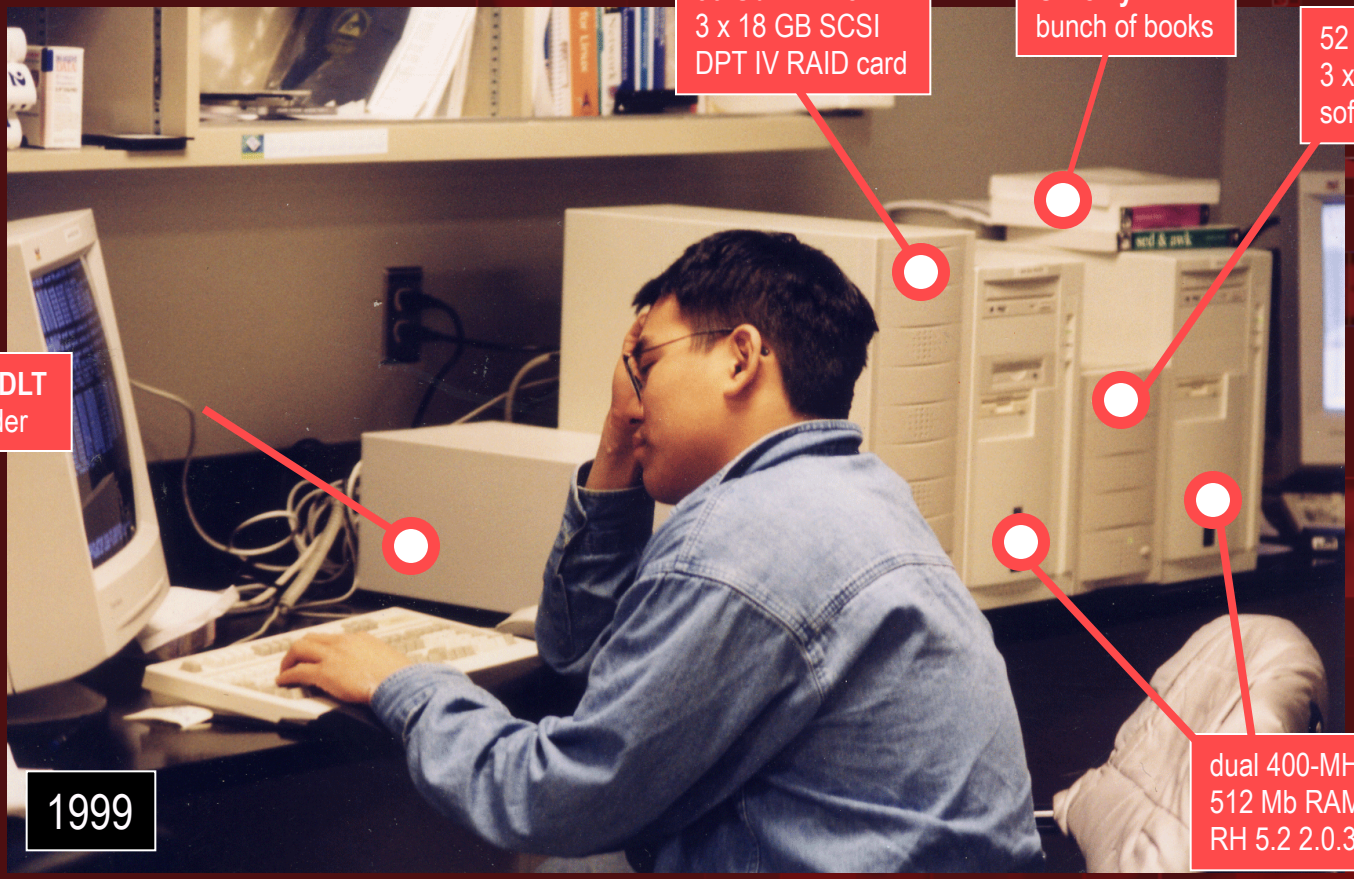
To all involved, CONGRATULATIONS! We will have a PCR-based SARS diagnostic up and running by next week thanks to YOU.

MJM, Madison, WI

Subject: You have to be NUTS!

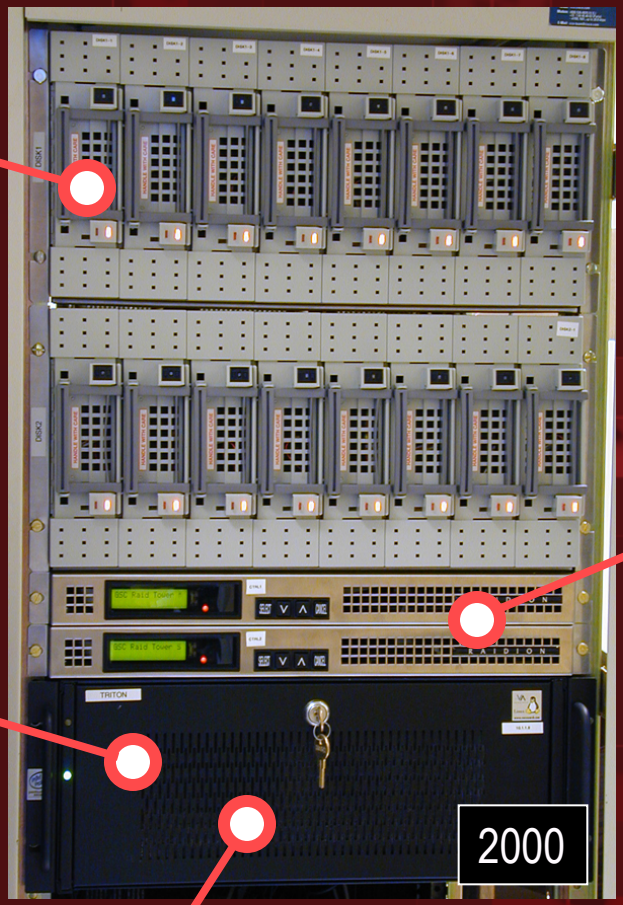
My daughter doesn't think its such a good idea to have the gene sequencing for the new coronavirus on the internet. I don't either! There should have been a better way! You must be crazy!

linux at the Genome Sciences Centre



linux at the Genome Sciences Centre

Raidion RAID-5
2 x 8 x 36 GB
400 GB



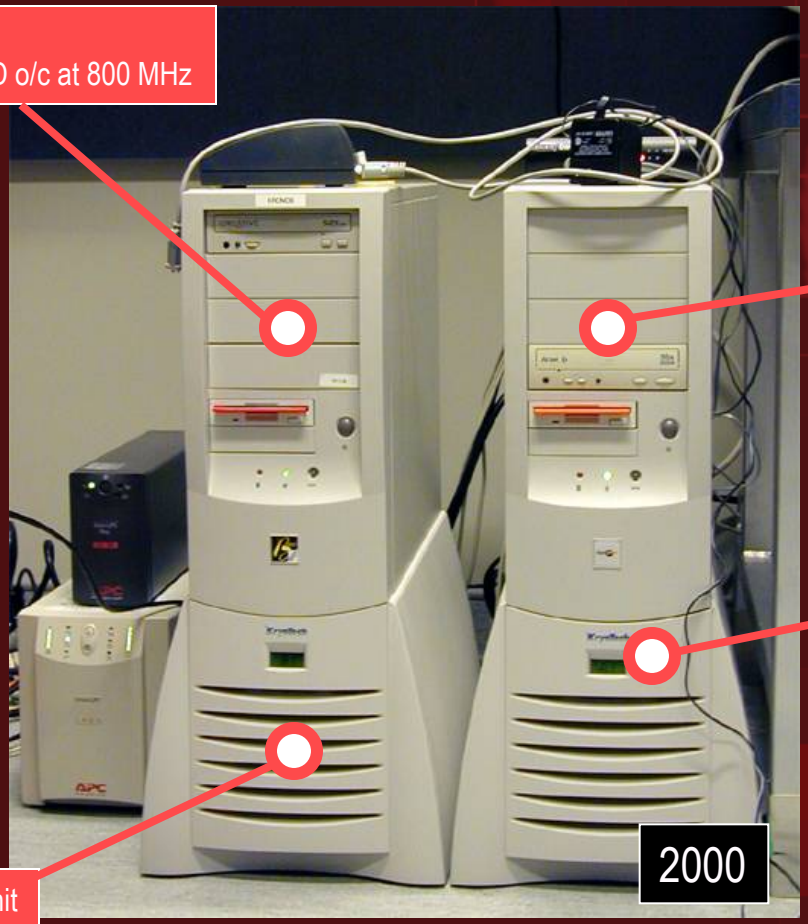
RAIDION.u2w
RAID controllers

VA VAR server 900
VA 6.0.3 2.0.36

```
4:41pm up 393 days, 6:48, 9 users, load average: 0.37, 0.29, 0.63
```

linux at the Genome Sciences Centre

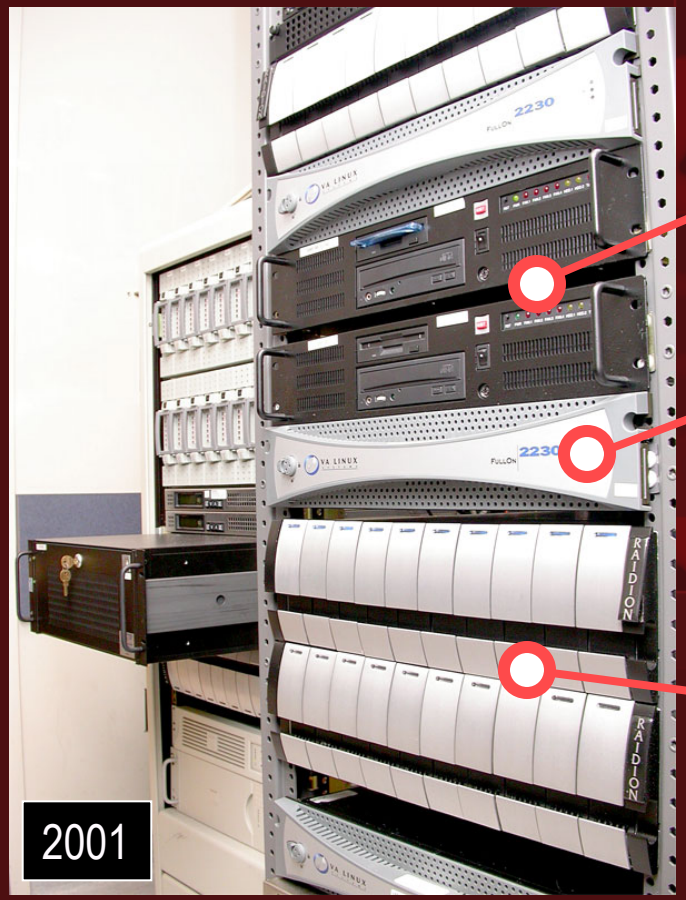
Kryotech
700 MHz AMD o/c at 800 MHz



Kryotech "SuperG"
800 MHz AMD o/c at 1 GHz

cooling unit
- 40 C

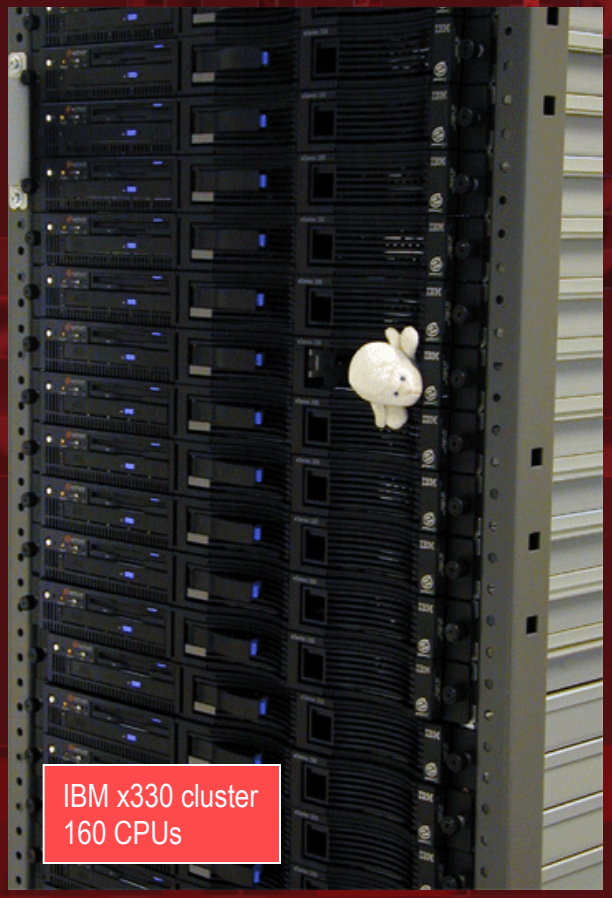
linux at the Genome Sciences Centre



dual 1 GHz P3
4 GB RAM

VA Linux 2230
dual 800 MHz P3
2 Gb RAM

Raidion RAID-5
3 x 10 x 72 Gb SCSI
Raidion.ha RAID controller
2 TB

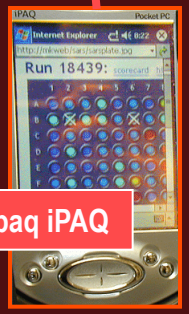


IBM x330 cluster
160 CPUs

linux at the Genome Sciences Centre

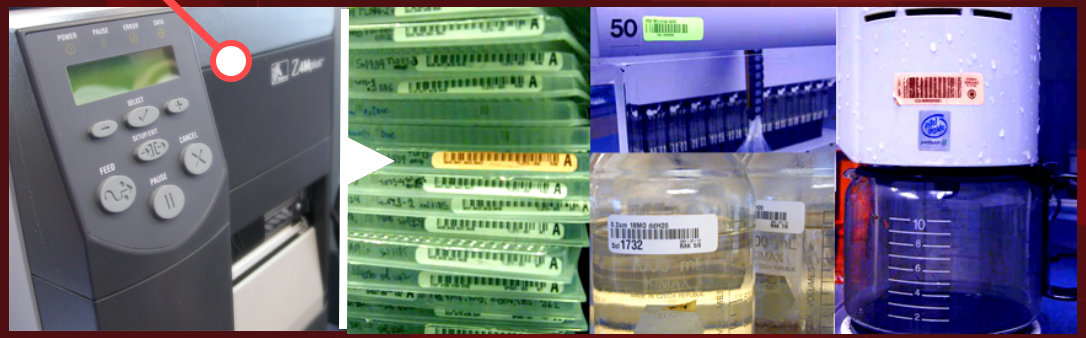
Zebra S600
ZPL printer

- wireless interface to ZPL printers for on-demand barcodes
  - Perl/Apache, Compaq IPAQ, Zebra S600
- MySQL database stores all events, objects and data

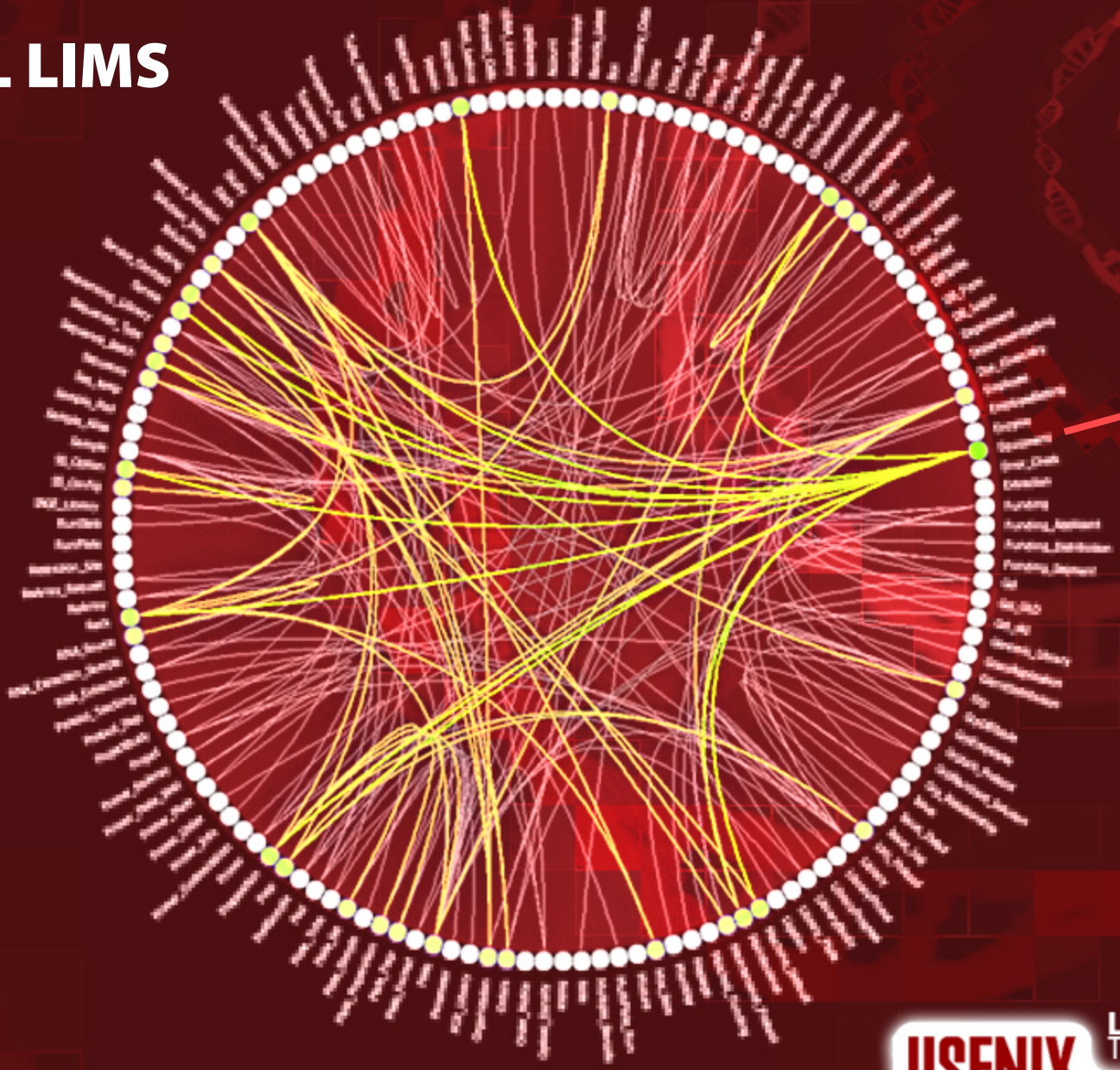


Compaq iPAQ

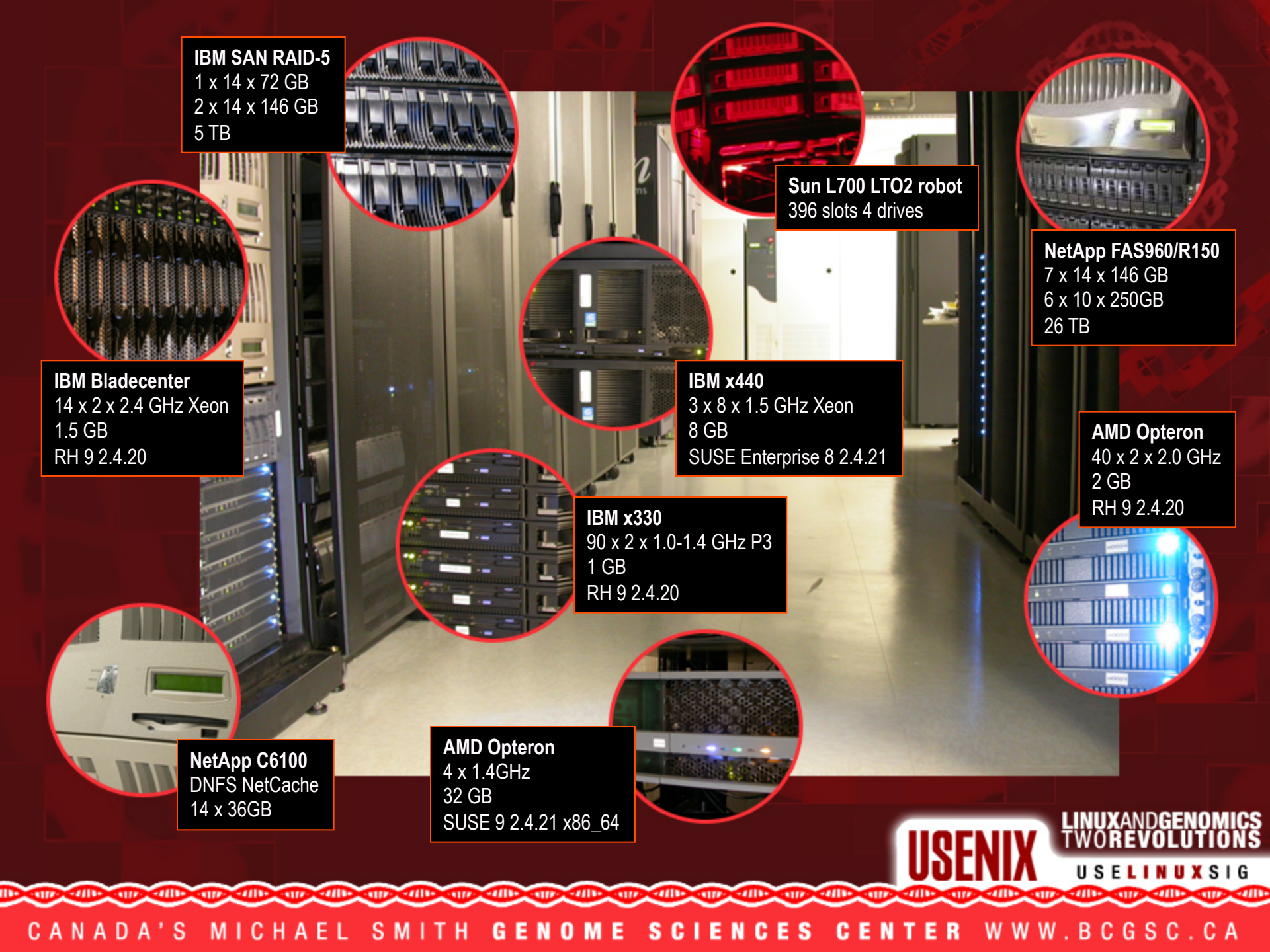
Zebra Z4Mplus
ZPL printer



MySQL LIMS



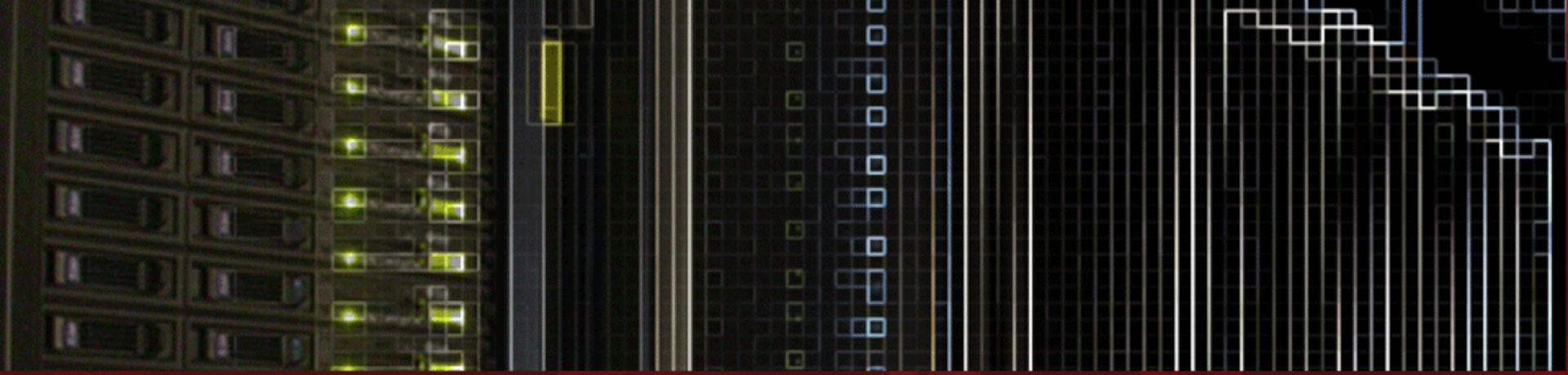
mkweb.bcgsc.ca/schemaball



MAKE NO LITTLE PLANS.
THEY HAVE NO MAGIC TO
STIR MEN'S BLOOD...

...MAKE BIG PLANS.

DANIEL BURNHAM
ARCHITECT



*nix | systems | db | programming | automation

www.bcgsc.ca/about/employment

talk slides mkweb.bcgsc.ca/sars/usenix