Adapting to the GRID

(OR BETTER: ADAPTING TO THE NEXT REVOLUTION…)

TOMMASO BOCCALI
INFN PISA

Outline

- Not a real talk, just (random) thoughts for the discussion
- Look at the past to be prepared for the future:
  - Even if you carefully wrote your experiment software stack to have the minimum dependencies
  - Even if you tried to assume as little as possible
- Sometimes (IT) the scenario changes around you in unexpected ways
- Can/How to adapt? At which cost?

Lesson from the past: LEP

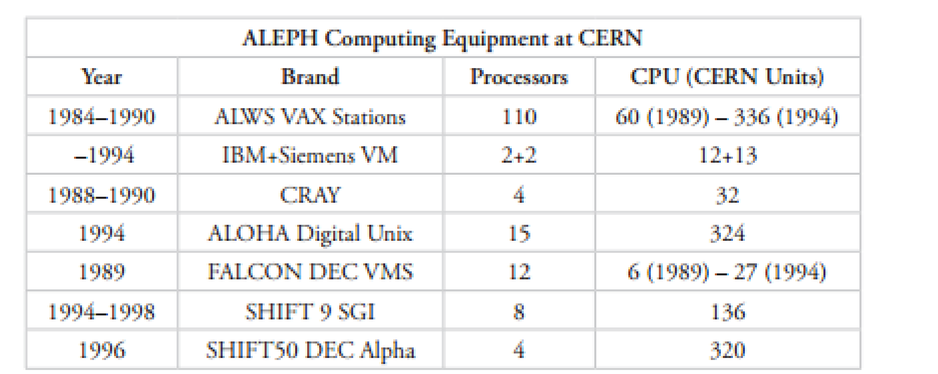
- LEP started taking data in the late 80s
- The (CERN) computing scenario was populated by VAX/VMS, if not even CERN/VM - VXCERN
- Fortran was the language for scientific programming
- Mainframes were the most common computing paradigm, eventually with batch submission systems; single user VAX workstation starting to appear
- Tapes were the common storage solutions, with small Disks for stagein / work areas

High quality / price systems

Low failure rate

By the end of LEP (~2000)

- The **same software and library** stack had been ported to various Unix systems (DEC, HP, SG, IBM ….) before, and eventually to Linux (RH Linux)
- The computing platform moved from mainframes (for ALEPHers: CRAY, shift3, shift9, shift50, ….) to farms (lxbatch)
- Still all local: resources present outside CERN, but not integrated with the central system



ALEPH Computing Equipment at CERN

Year	Brand	Processors	CPU (CERN Units)
1984–1990	ALWS VAX Stations	110	60 (1989) – 336 (1994)
–1994	IBM+Siemens VM	2+2	12+13
1988–1990	CRAY	4	32
1994	ALOHA Digital Unix	15	324
1989	FALCON DEC VMS	12	6 (1989) – 27 (1994)
1994–1998	SHIFT 9 SGI	8	136
1996	SHIFT50 DEC Alpha	4	320

Typical assumptions for a "late LEP" analysis

- You worked at CERN only (or XXX only, but in a single place)
- You could split the computing task in jobs, and use LSF (NQS, ….) to submit to batch systems
  - ○ Hand crafted scripts submitting via bsub, typically
- The jobs were landing on nodes with identical setup
- You had all the data "available" (either Disk, Tapes)
  - ○ If not, **you** were making sure you had all the data you need
- You had local (=site level) Disk/Tape where to store the results

- Failures due to "problematic machines", broken tapes/disks, failed library loads etc very limited
  - ○ (<)% level of failures, usually cured by manual resubmission

The big barn model …

- Whatever resources you need
  - Stuff them in a single place, with a LAN security model
  - They better be uniform, even more from the setup point of view
  - If you need more resources, buy more / enlarge the barn / hire more staff
  - Nothing out of direct control
    - Network paths predictable/ reliable
    - Single source of SW repositories area (either works or not, globally)
    - …

… but then the whole landscape changed …

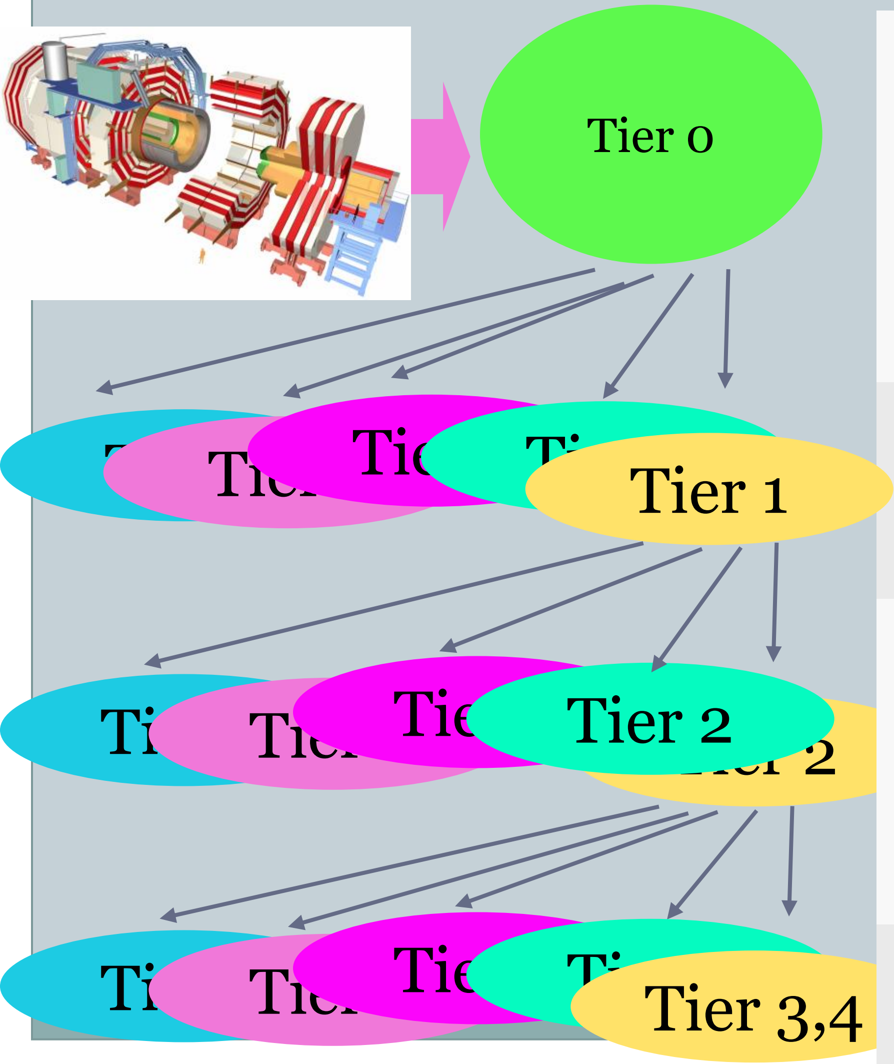
- LHC: away from this model, by no means for purely technical reasons
  - A big (huge) barn would have worked

- Political/Sociological decision
  - Funding Agencies preferring to build national IT infrastructure instead of paying checks to CERN
  - FAs preferring to build local expertise
  - Physicists preferring to compute @home
  - Easier dissemination to other sciences (or even industry…)
  - EU keen on financing such a dissemination effort

- How bigger the barn? Today CERN hosts (REBUS) for LHC
  - ~20% of CPU resources
  - ~15% of Disk resources
  - ~30% of tape resources

5x would have been enough …

MONARC …



CERN

Master copy of RAW data

Fast calibrations

Prompt Reconstruction

7 centers

A second copy of RAW data (Backup)

~50 centers

Analysis Activity

They are dimensioned to help ~ 50 physicists in their analysis activities

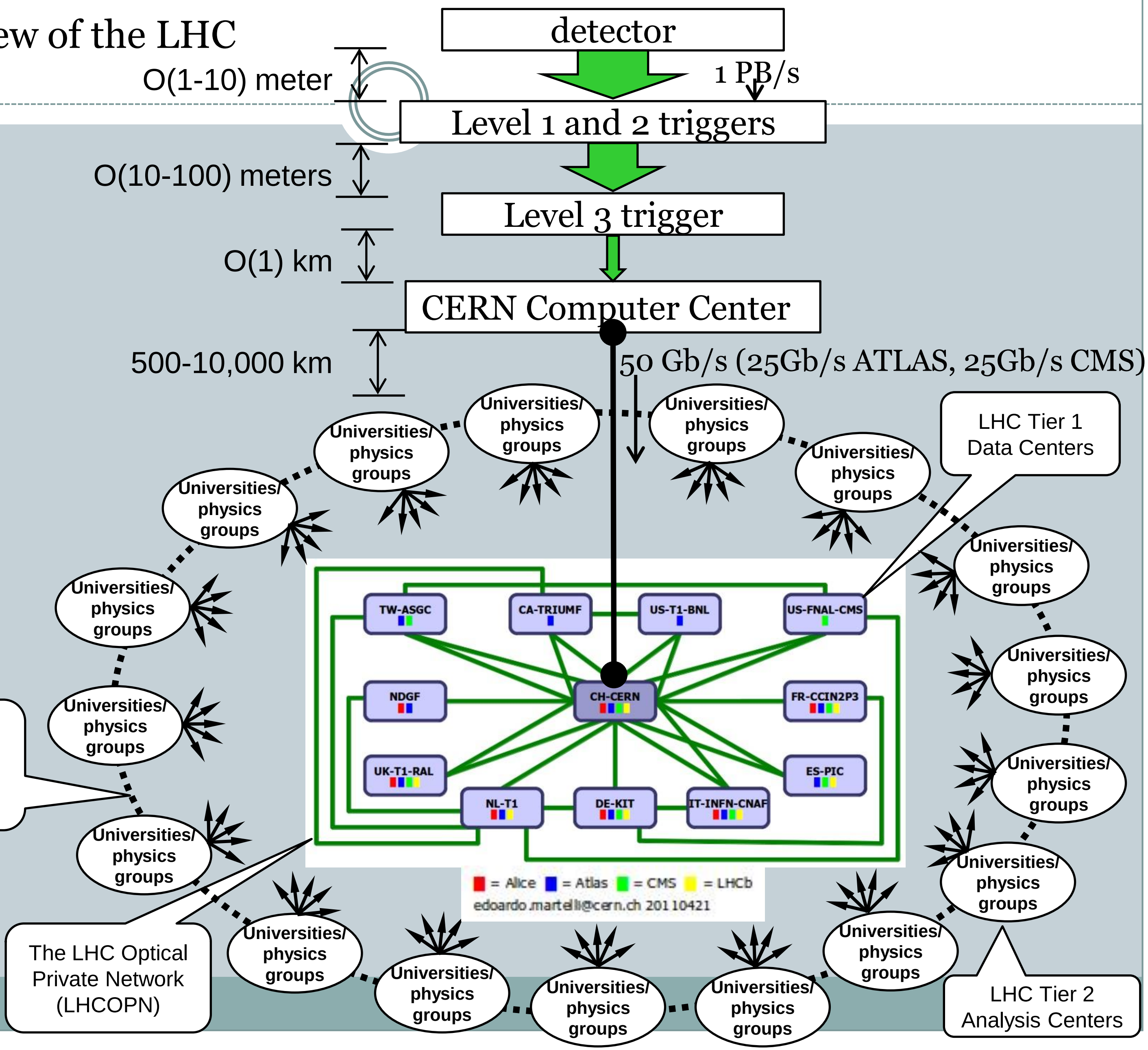
Anything smaller, from University clusters to your laptop

… and the network (complexity) …

A Network Centric View of the LHC

CERN →T1	miles	kms
France	350	565
Italy	570	920
UK	625	1000
Netherlands	625	1000
Germany	700	1185
Spain	850	1400
Nordic	1300	2100
USA – New York	3900	6300
USA - Chicago	4400	7100
Canada – BC	5200	8400
Taiwan	6100	9850

The LHC Open Network Environment (LHCONE)



The LHC Optical Private Network (LHCOPN)

This ✳ is intended to indicate that the physics groups now get their data wherever it is most readily available

A big change?

- What if a LEP Experiment would like to turn today/then to the GRID, with a similar (although smaller) MONARC Setup?
- If you take the ideal GRID model, not so much of an effort
  1. LSF configuration language -> jdl
  2. bsub -> glite-wms-job-submit
  3. rfcp -> lcg-cp
  4. All machines similar (all using EMI3, for example)
  5. Global filesystem: rfdir -> LFN via LFC

… but in practice …

- Failure rate (random) 1% -> 30%
  - Glitches in the WAN / WAN saturation
  - Computing machines with glitches due to local factors (locally shared disk problems, installation peculiarities, local infrastructure problems, e.g.)

- Unique logical FS, but access performance wildly different LAN/WAN; you need to send jobs to data
  - Need an efficient data placing

- GRID MW not scaling for LHC needs, substituted by experiment MW
  - Metadata catalogs
  - WMS -> pilots
  - glite-wms-job-submit -> submission to frameworks taking care of (smart) resubmission upon failures
  - LFC -> experiment location DBs
  - SRM limited to tapes, Xrootd/Httpd upcoming
  - Simple SRM transfers -> trial and error approach to get 100% of the files at destination

Reality for LHC …

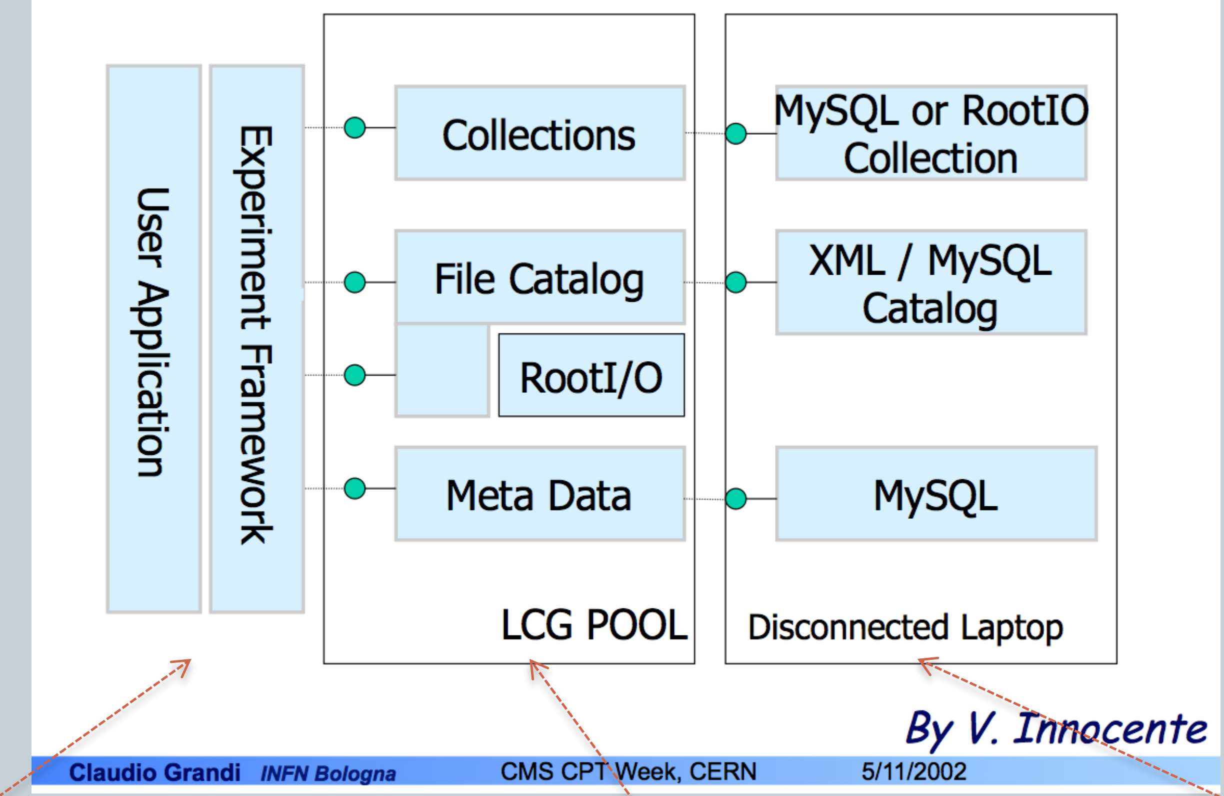
- … Is a huge effort from the (luckily manpower rich) collaborations
  - Development
    - Catalogs (metadata, location)
    - Data movement
    - WM
  - Operations
    - user/workflow/site support
  - Monitoring, monitoring, monitoring …

Software is apparently less of a problem

- ALEPH example just shown by M.Maggi:
  - Assume something easy: POSIX access for ALEPH, just TFile::Open for CMS
    - For future and potentially unpredictable access patterns, we can assume things like Parrot/Fuse will exist
  - Leave catalog access / data discovery / outside of the software internals
  - Our Software stack (the algorithmic code) is already now capable to run virtually anywhere there is a POSIX system supported by GCC
- If you
  - Stay low enough with assumptions, basing yourself on "stable" ground (e.g. POSIX)
  - Use SW you can recompile
  - Decouple completely the algorithmic and data access part

Example (circa 2002) from CMS

- CMS pre 2000 was not GRID aware
  - Local machines in institutions + a cluster at CERN



User Application

Experiment Framework

Collections

File Catalog

RootI/O

Meta Data

LCG POOL

MySQL or RootIO Collection

XML / MySQL Catalog

MySQL

Disconnected Laptop

By V. Innocente

Claudio Grandi *INFN Bologna* CMS CPT Week, CERN 5/11/2002

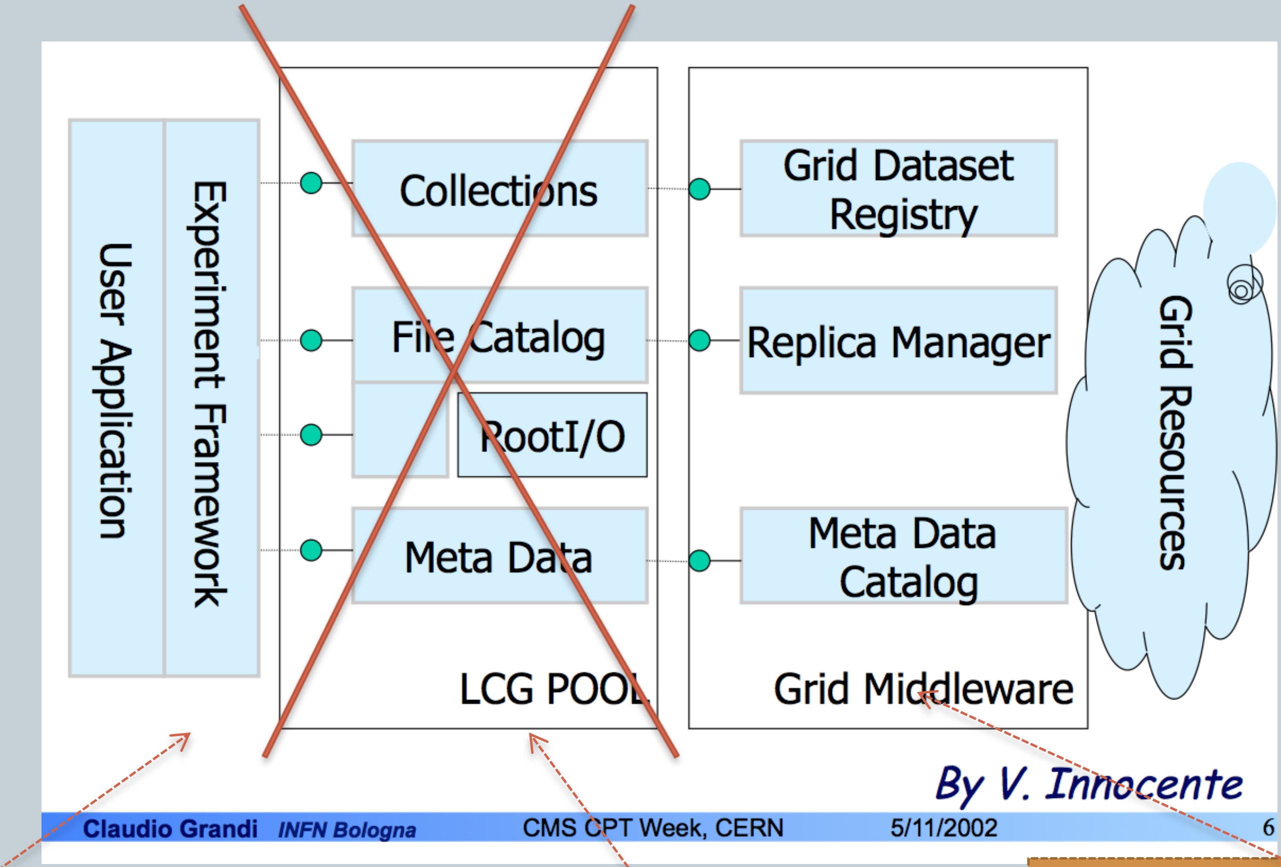
Experiment stuff

The “stable” layer

Data access part

.. To GRID ...

ROOT



By V. Innocente

Claudio Grandi INFN Bologna CMS CPT Week, CERN 5/11/2002 6

Experiment stuff

The “stable” layer

Data access part, gridified

What is the “stable ground”?

- LEP:
  - POSIX, CERNLIB, GEANT3, (BOS, ORACLE, FORTRAN) …
- LHC:
  - POSIX, ROOT, GEANT4, (ORACLE, *CC), …

  - The fewer the better
  - Be prepared to maintain also that for the intended lifetime
    - Or, clearly, freeze all and go to VMs

LHC today: not a typical situation

- LHC for now has lived in a surprising stable environment
  - C++/Linux/x86 from the start (apart from a some very initial efforts)
  - GRID by very early design
  - At least since 2000, and realistically at least up to 2015 (in this time span, LEP had changed more than an handful of solutions)
  - We see definite trends
    - Linux will probably stay with us, but x86 -> x86, ARM, GPU, ???
    - GRID-> Cloud
    - Even if probably NONE of these is comparable to the local -> distributed transition

LEP: Cost "would have been"

- Cost would have been
  - Build N computing centers instead of one, which still means
    - More infrastructure cost and manpower costs
      - Manpower: (my) estimate is currently O(10-20) FTEs per T1, 2-5 per T2
      - Have Oracle at T1s
      - In the end, FAs chose to go this way for
  - Someone preparing the MW
    - But with EU/US specific funds
      - Difficult to estimate, but definitely exceeding 100MEur in the decade

Cost would be "now" …

- Completely different, of course …
- Direct resource cost is already compatible with zero for LEP experiments
  - Total ALEPH DATA + MC (analysis format) = 30 TB
  - ALEPH: Shift50 = 320 CernUnit. One of today's pizza box **largely** exceeds this
  - CDF data: O(10 PB), bought today for <400kEur
  - CDF CPU ~ 1MSi2k = 4 kHS06 = 40kEur

Zero!

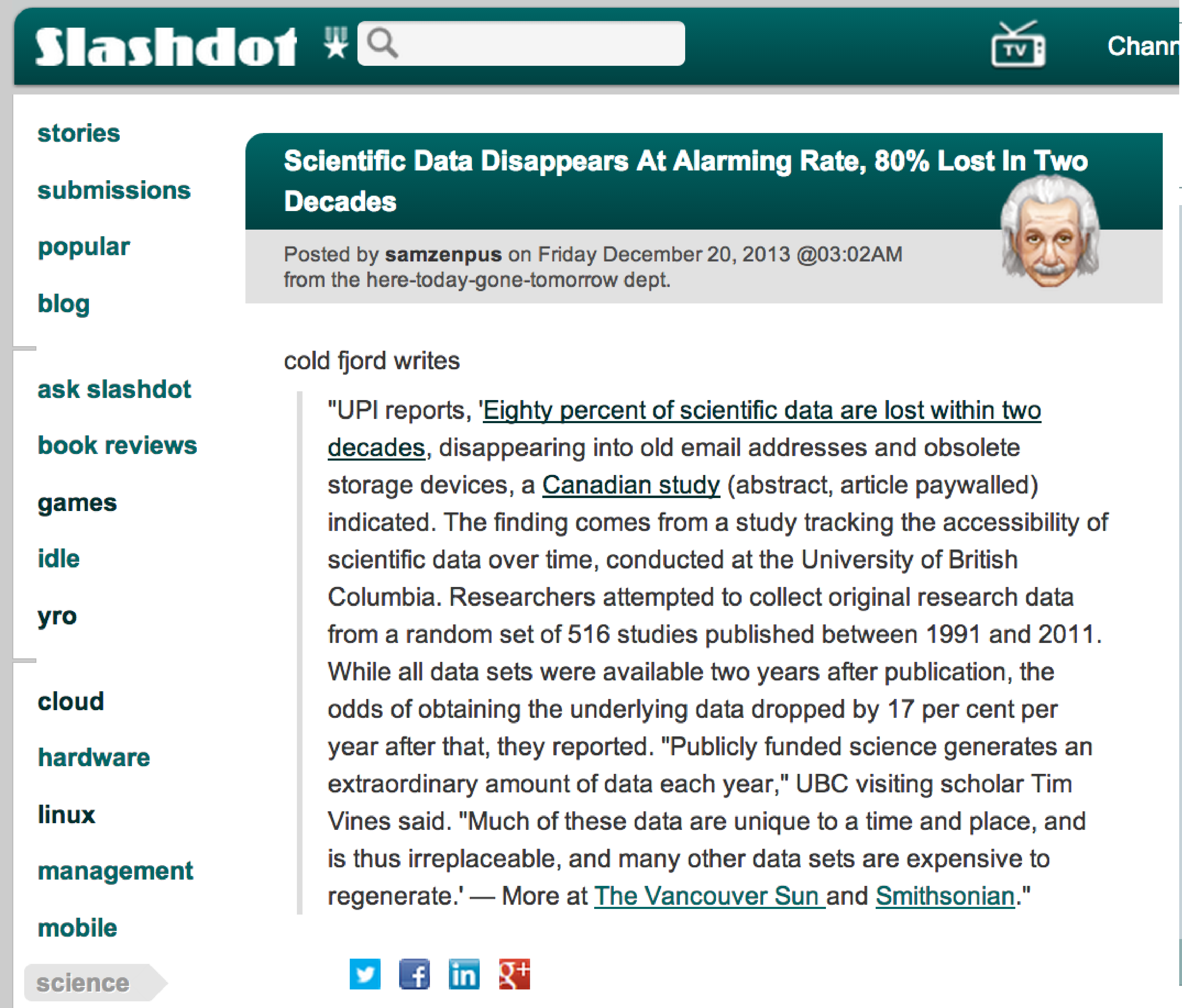
!=0, but decreasing fast

- Here the main problem is knowledge /support, clearly
  - Can you trust a "NP peak" 10 years later, when experts are gone?
- ALEPH reproducibility test (M.Maggi, by NO mean a DP solution) ~0.5 FTE for 3 months

Conclusions (?)

- After X years (10?), cost of resources is negligible to the human effort and the difficulty to gather old experiment wisdom

- Experiment Software is less of a problem wrt
  - Data management tools (Where is data? What is data?)
  - External foundation libraries
    - (and even there, you can always virtualize if nothing else works …)



Scientific Data Disappears At Alarming Rate, 80% Lost In Two Decades

Posted by **samzenpus** on Friday December 20, 2013 @03:02AM
from the here-today-gone-tomorrow dept.

cold fjord writes

> "UPI reports, 'Eighty percent of scientific data are lost within two decades, disappearing into old email addresses and obsolete storage devices, a Canadian study (abstract, article paywalled) indicated. The finding comes from a study tracking the accessibility of scientific data over time, conducted at the University of British Columbia. Researchers attempted to collect original research data from a random set of 516 studies published between 1991 and 2011. While all data sets were available two years after publication, the odds of obtaining the underlying data dropped by 17 per cent per year after that, they reported. "Publicly funded science generates an extraordinary amount of data each year," UBC visiting scholar Tim Vines said. "Much of these data are unique to a time and place, and is thus irreplaceable, and many other data sets are expensive to regenerate.' — More at The Vancouver Sun and Smithsonian."