Where could we go from here? – The next phase of computing in HEP

Sverre Jarp
CERN
ACAT 2011
5 September 2011

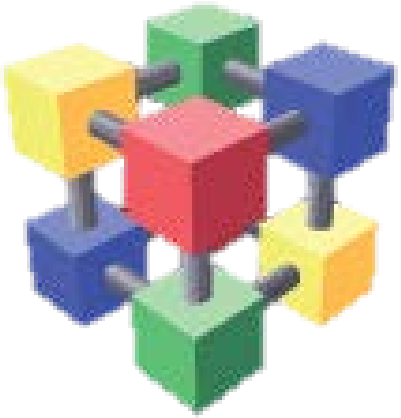
Overview

- Our current success: The World-wide LHC Computing Grid
- Some points from the past
- Megatrends and buzz from the Web
- What to propose for HEP Computing tomorrow?

With input from my CERN openlab colleagues, Ian Bird, René Brun, Pere Mato, Alberto Pace, Bernd Panzer-Steindel and others

CERN openlab

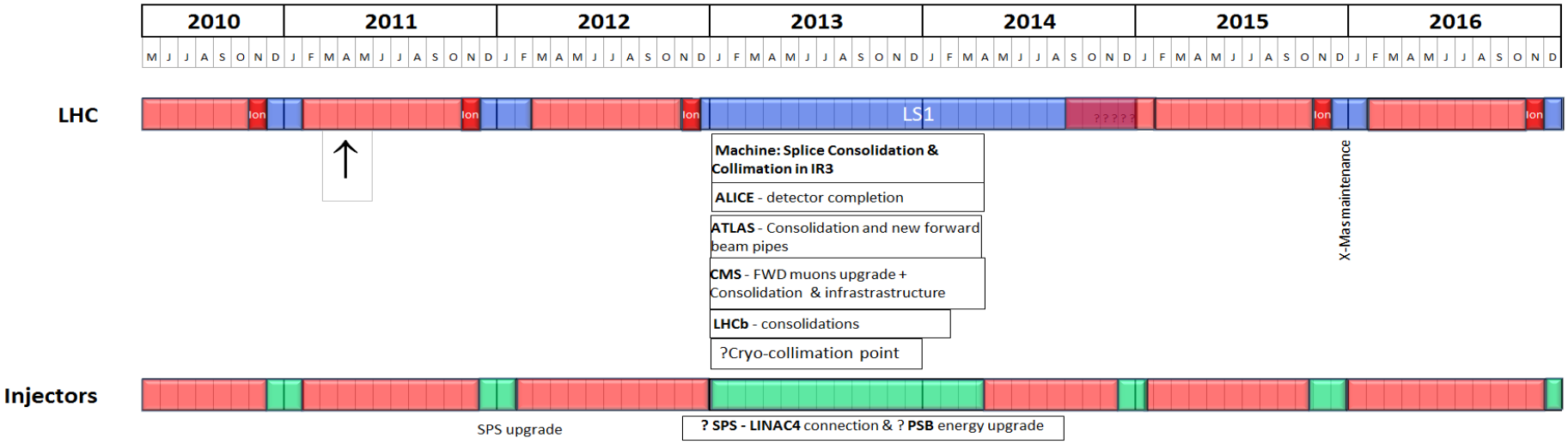
W-LCG

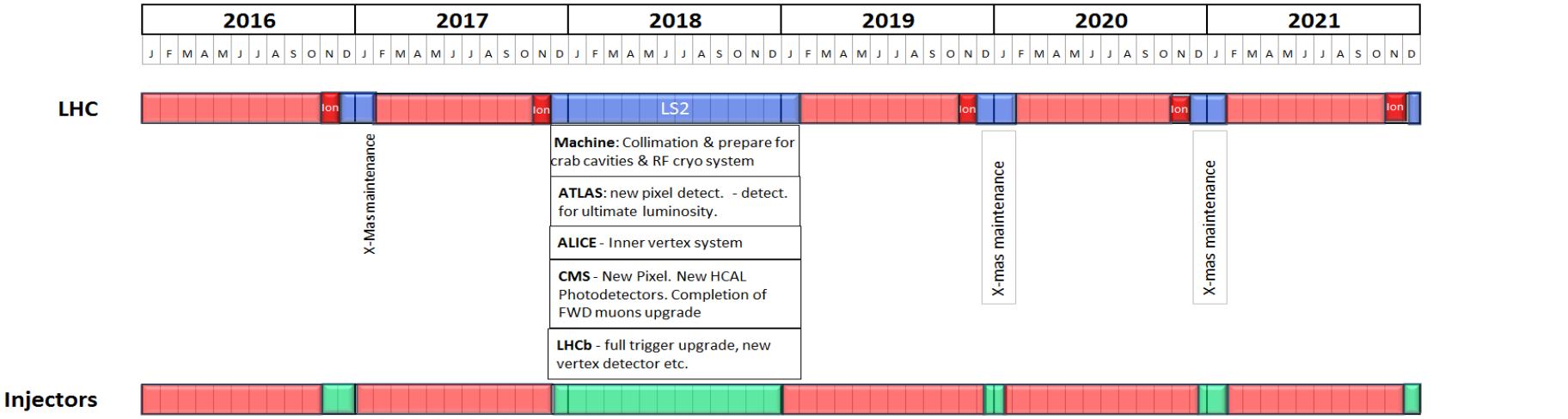


LHC schedule

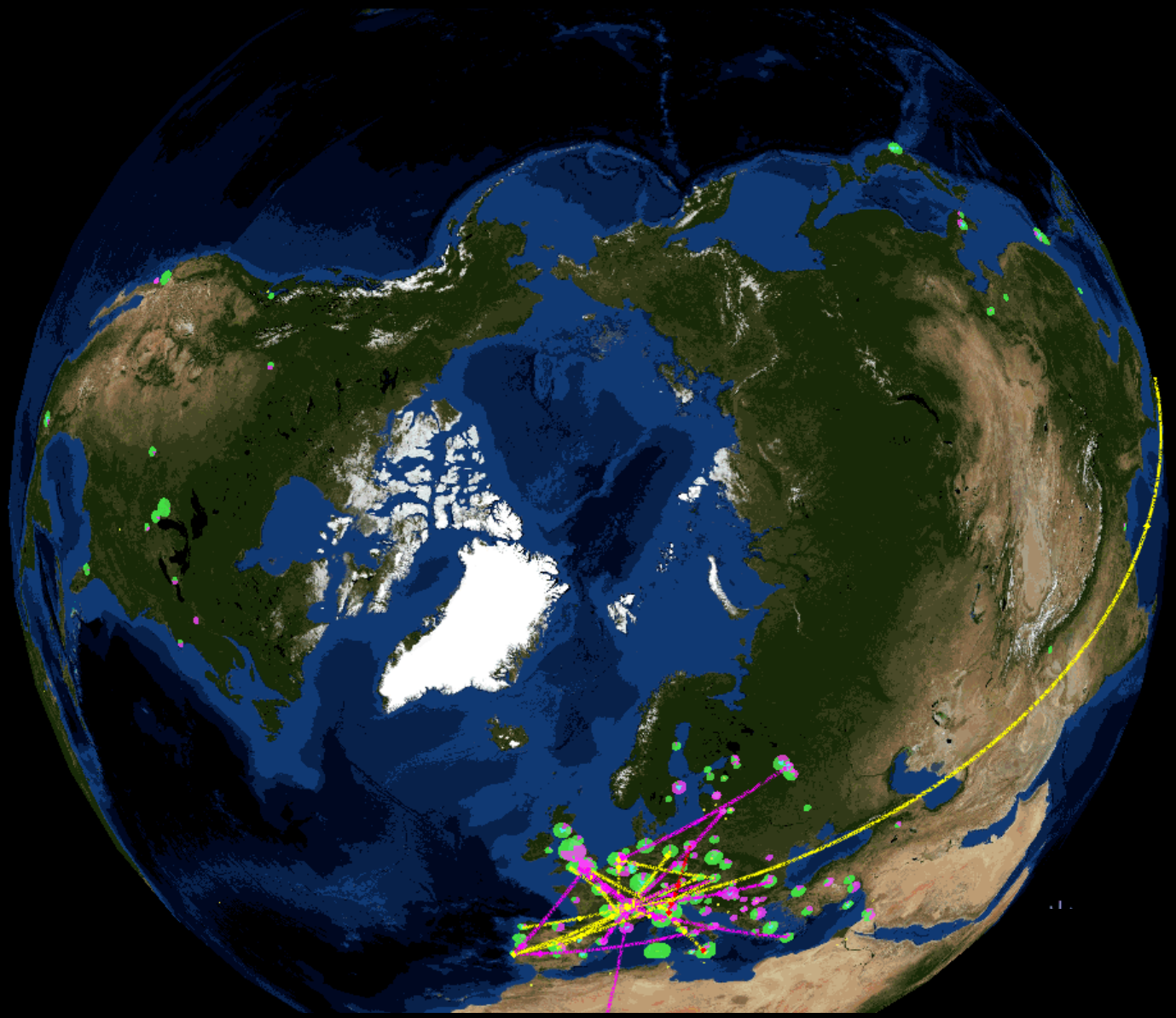
New rough draft 10 year plan

Not yet approved!





The World-wide LHC Computing Grid



The success of W-LCG

- Today:

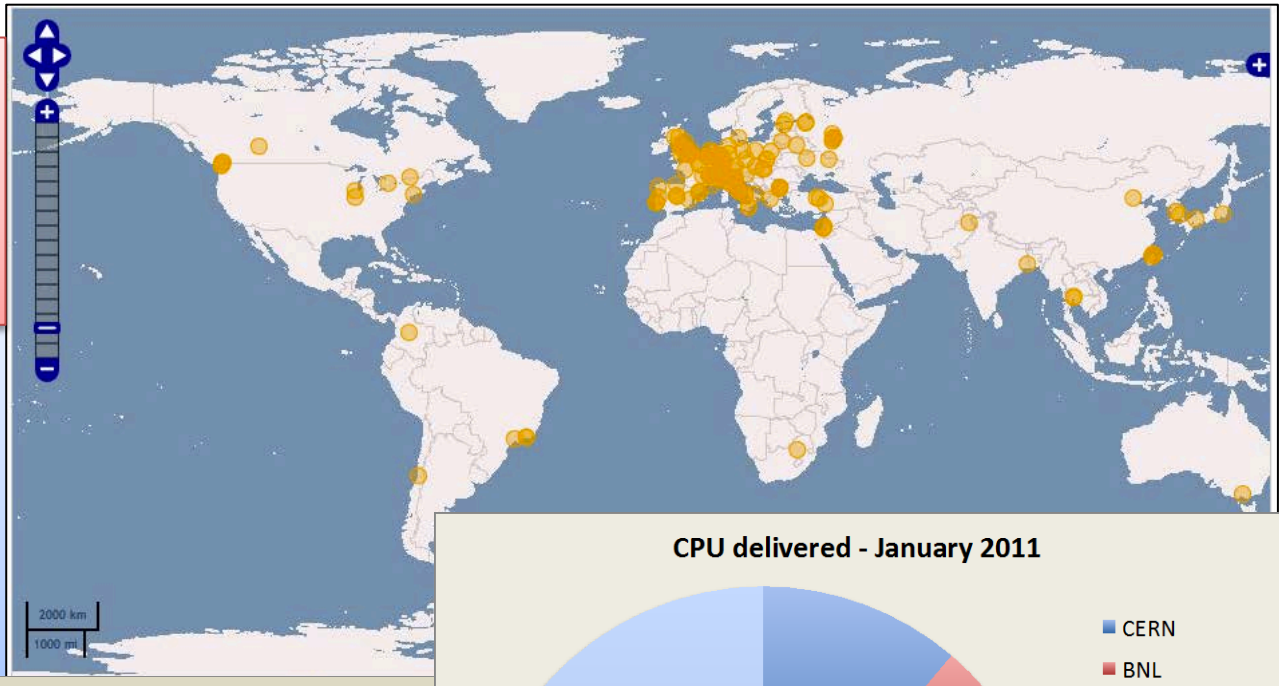
>140 sites

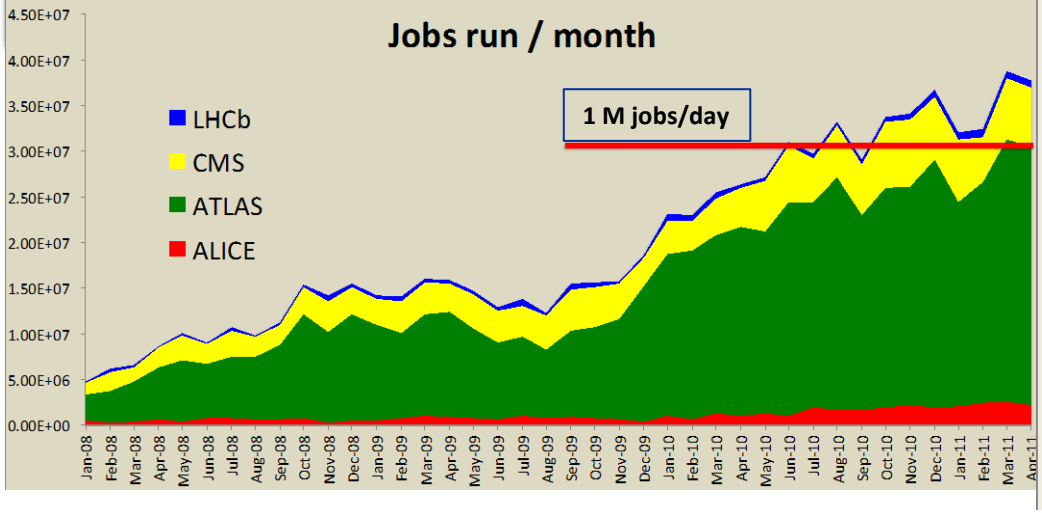
>250'000 CPU cores

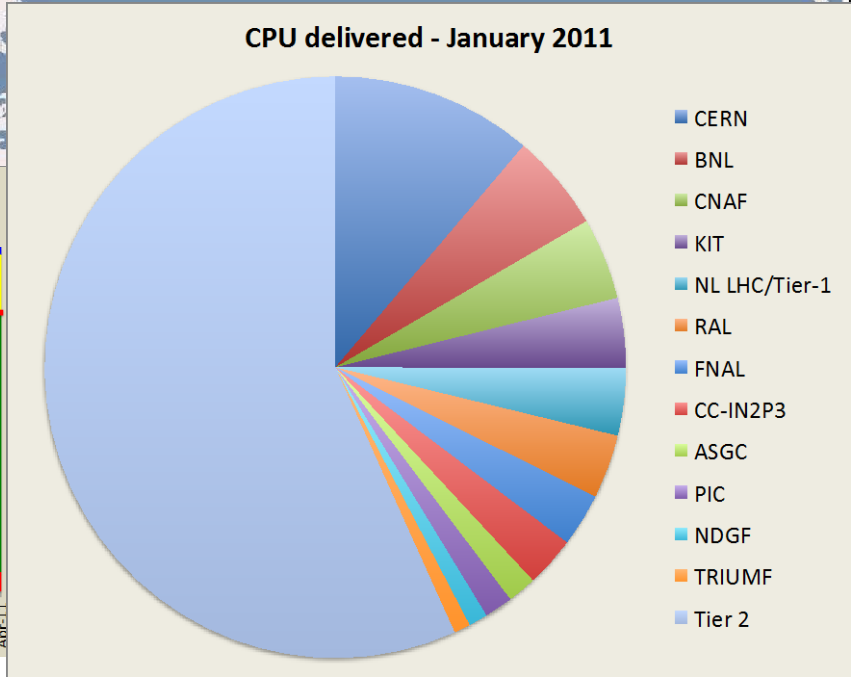
>150 PB disk space

34 countries:

Australia, Austria, Belgium, Brazil, Canada, China, Czech Rep, Denmark, Estonia, Finland, France, Germany, Hungary, Italy, India, Israel, Japan, Rep. Korea, Netherlands, Norway, Pakistan, Poland, Portugal, Romania, Russia, Slovenia, Spain, Sweden, Switzerland, Taipei, Turkey, UK, Ukraine, USA.







The layers in use

- Grid middleware; pilot jobs
- Ethernet [1 and 10 Gbits; LAN and WAN; ip v4]
- High-density tapes
- Castor, xrootd; ROOT files (physics data)
- Relational databases (metadata)
- AFS
- NAS w/RAID6
- Batch system [LSF, or similar]
- SHIFT architecture
- Multicore (multiprocessing) C++ frameworks
- gcc/Linux
- Scalar SSE (hardware vectors)
- x86_64

CERN openlab

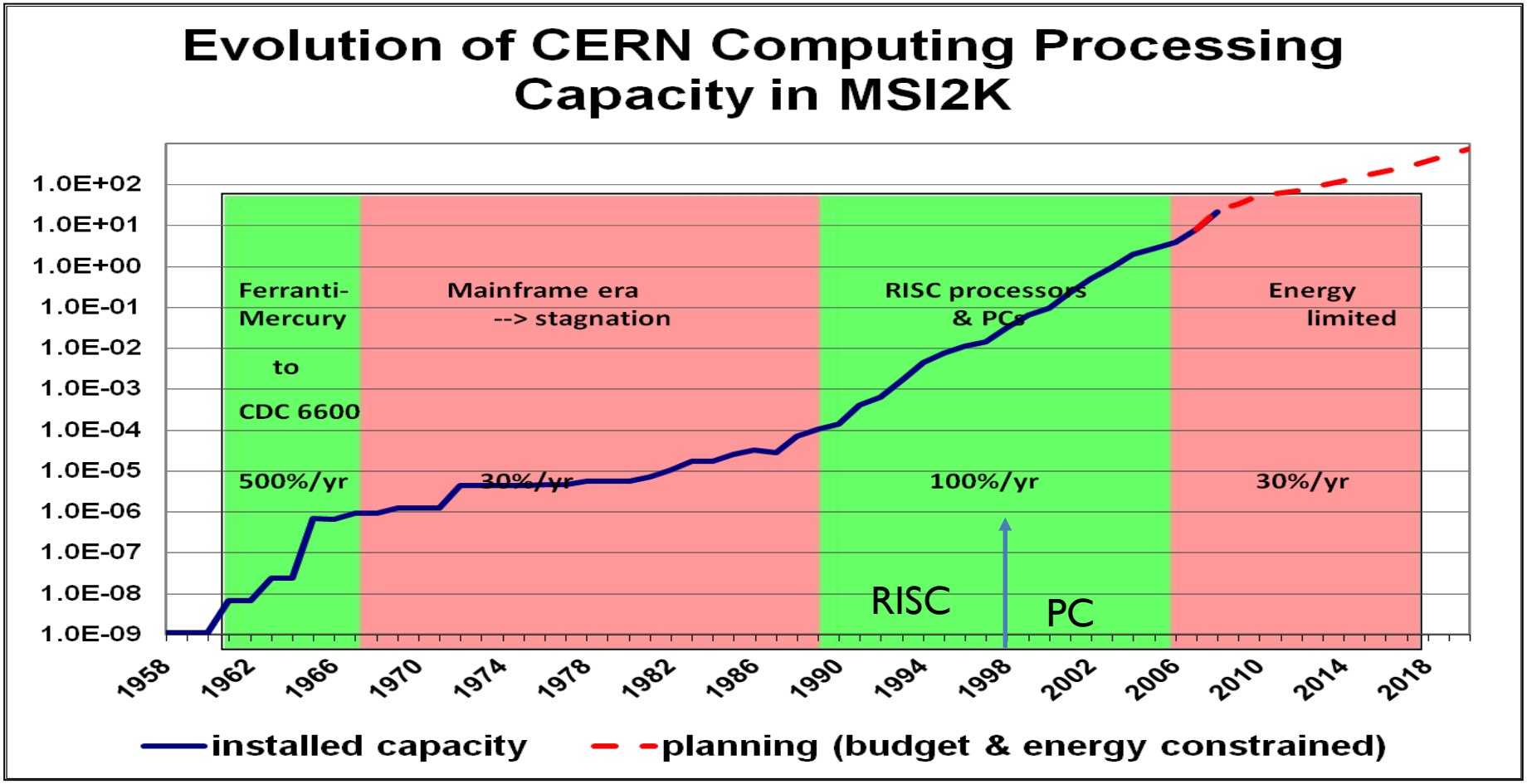
Some points from the past

Archaic CPUs

- As "stupid" as 50 years ago
- Still based on the Von Neumann architecture
- Primitive "machine language"
- Ferranti Mercury:
  - Floating-point calculations
    - Add: 3 cycles; Multiply: 5 cycles
- Today:
  - Programmers need to hide the latency in exactly the same way

Historic plot of CERN's computing resources

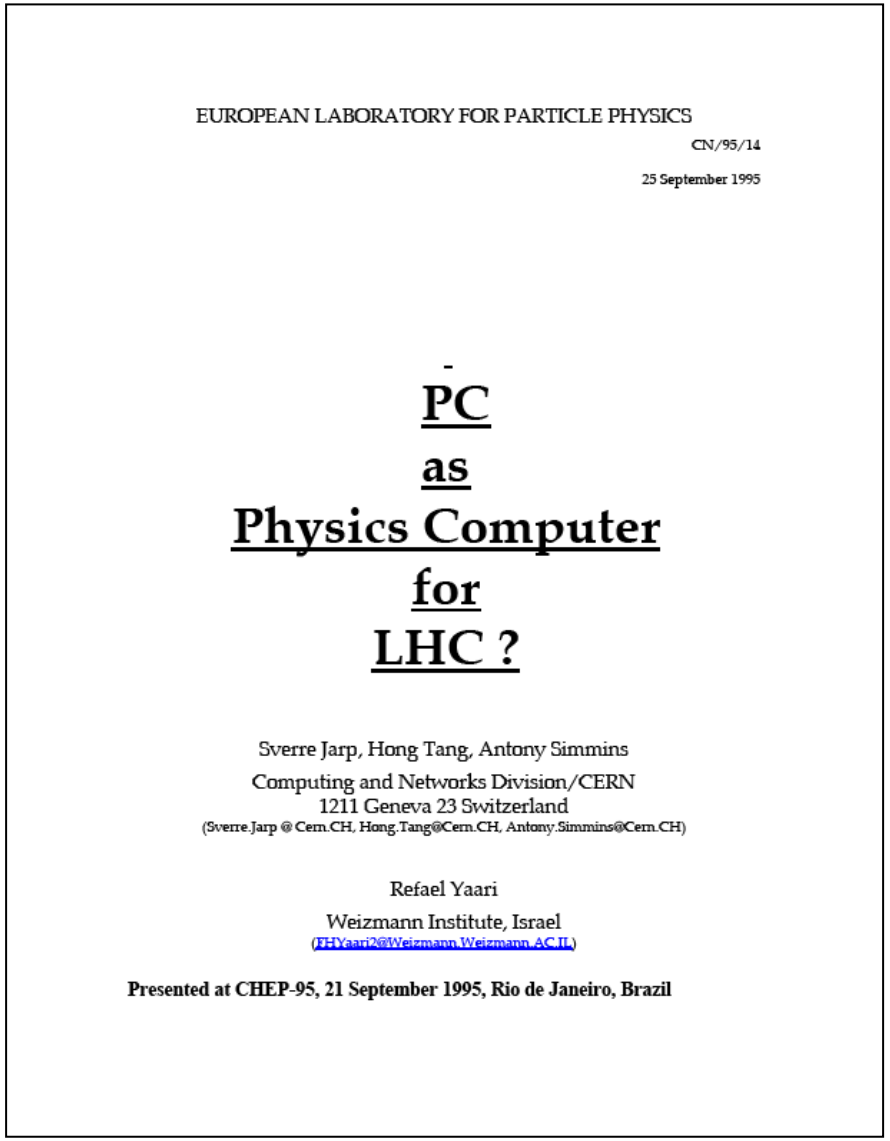
- 50+ years in review:



Plot by Les Robertson, ex-CERN

The move to PCs

- The paper at CHEP95 in Rio
- My unique correct prediction?
- What did it take to move?



EUROPEAN LABORATORY FOR PARTICLE PHYSICS

CN/95/14

25 September 1995

PC as Physics Computer for LHC ?

Sverre Jarp, Hong Tang, Antony Simmins
Computing and Networks Division/CERN
1211 Geneva 23 Switzerland
(Sverre.Jarp @ Cern.CH, Hong.Tang@Cern.CH, Antony.Simmins@Cern.CH)

Refael Yaari
Weizmann Institute, Israel
(FHYaari2@Weizmann.Weizmann.AC.IL)

Presented at CHEP-95, 21 September 1995, Rio de Janeiro, Brazil

CERN openlab

Trends and buzz

The mega-trends?

- Phones
  - Soon, there is one for every inhabitant on earth
    - 1'650'000'000 expected sold this year
- Smart-phones
  - Approaching one billion devices
    - 480'000'000 this year; CAGR: 60%
- Tablets
  - 50'000'000 with CAGR of 200%
- In comparison:
  - Netbooks/Notebooks (200'000'000)
  - Desktops (150'000'000)
  - Servers (10'000'000) with 55 BUSD in revenue

Buzz on the Web

- A small collection:

Intel Unveils Groundbreaking Tri-Gate Transistors for 22nm Chips

ARM vs Intel: the next processor war begins

AMD Fusion - The Future of Computing

Freescale eyes post-PC era

IBM uncloaks 20 petaflops BlueGene/Q super:
Lilliputian cores give Brobdingnagian oomph

Most phones shipped in 2015 will be smart-phones

Patent Wars: Google Buys Motorola Mobility for $12.5 Billion

300 Chinese fabless companies are springing up across the country

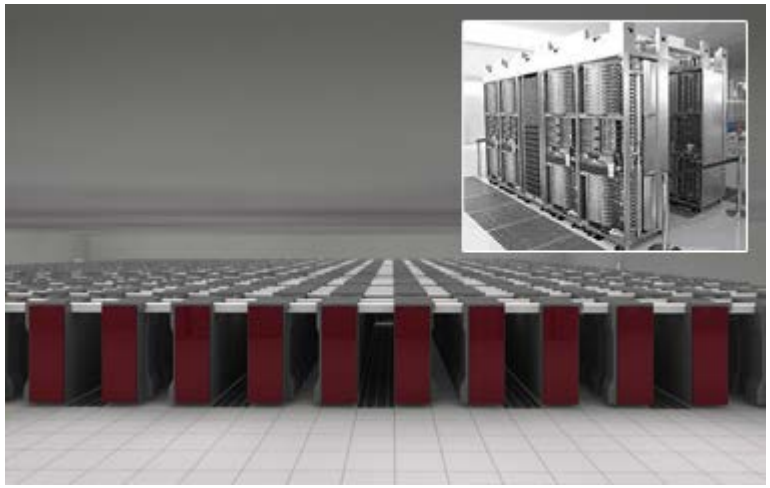
RAM prices set to hit 'free fall'

Beyonce Pregnant: New Twitter Record Set

Situation in HPC

3 years: 8x

- Mid 2008
  - Roadrunner: 1'026 Teraflops
- Mid 2011
  - Fujitsu "K" computer/ Sparc: 8'162 Tflops
- 2012
  - IBM Blue-Gene/Q: ~20 Peta-flops w/100'000 Power processors
    - 1.6 million cores



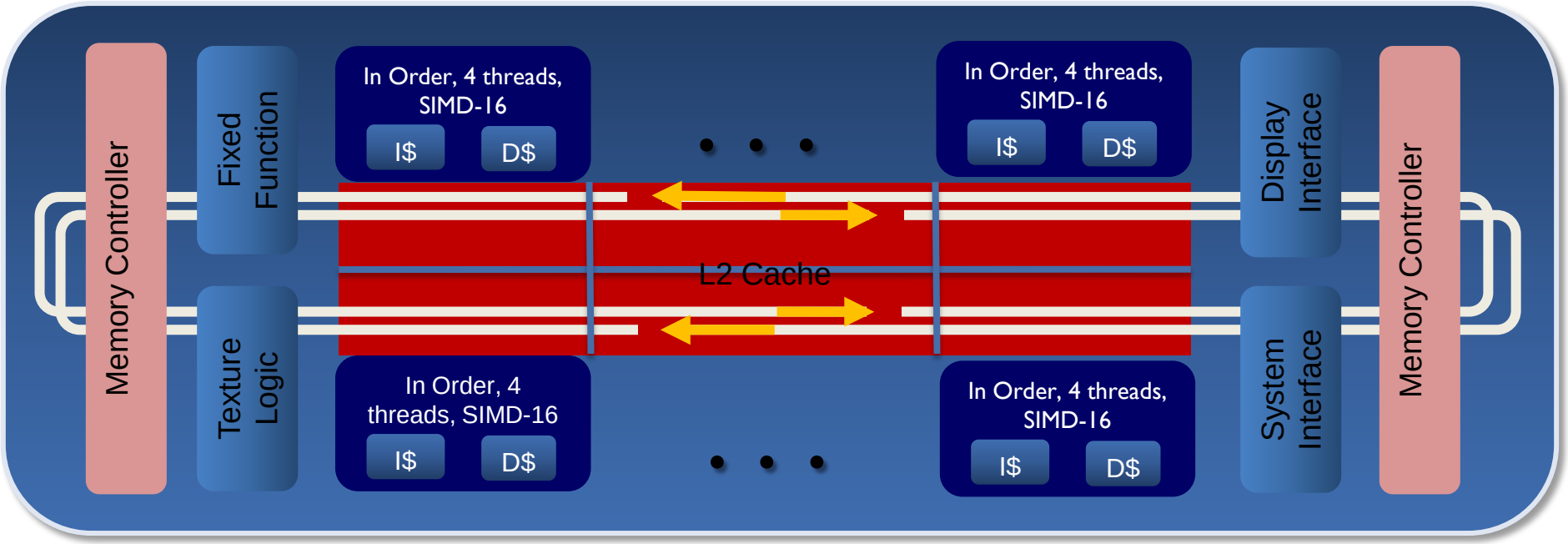
Steady march towards Exascale (2x per year) needing programs able to handle 1'000'000'000 threads

Continued PC era

- In spite of the smart-phone revolution, there is still (lots of) money in PCs
- Intel is betting on multiple horses:
  - Xeon
  - Atom
  - MIC
  - SSC

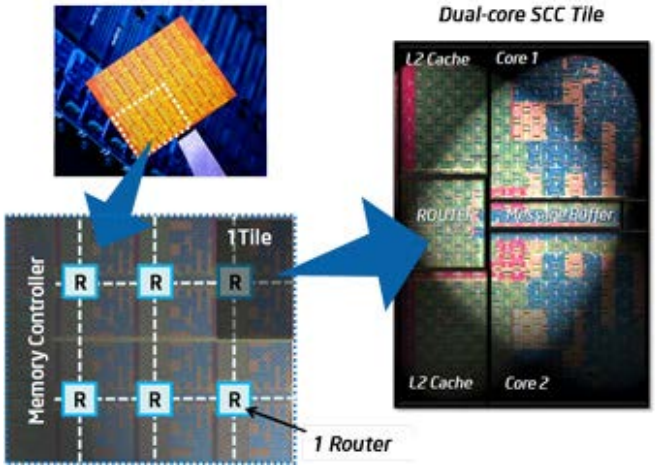
Intel's Many Integrated Core

- MIC Architecture:
  - Announced in May 2010
  - Based on the x86 architecture, 22nm (in 2012?)
  - Many-core (> 50 cores) + 4-way multithreaded + 512-bit vector unit
  - Limited memory: A few Gigabytes



Intel's Single Cloud Computer

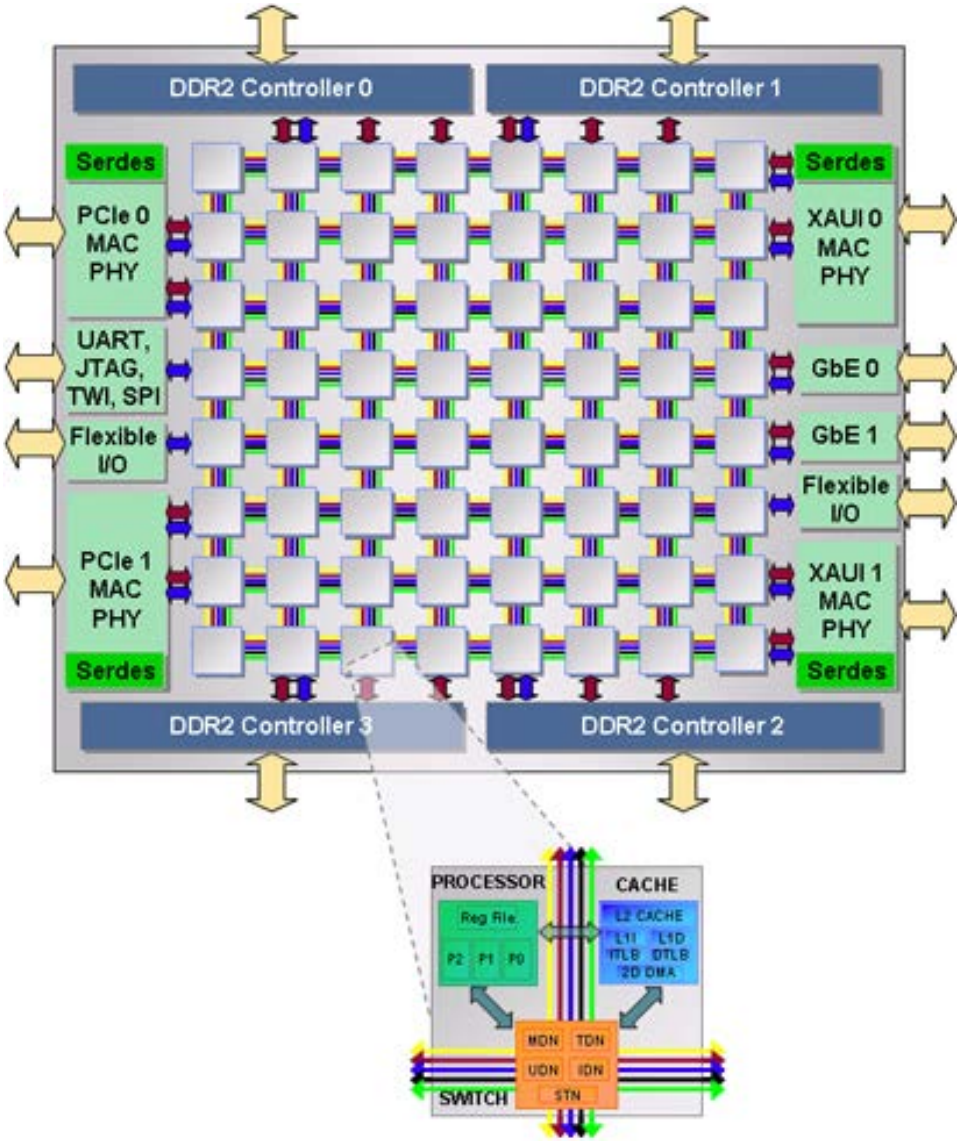
- Chip with 48 cores:
  - scalable to much bigger core counts
  - 24 "tiles" with two IA cores per tile
  - A 24-router mesh network with 256 GB/s bisection bandwidth
  - No cache coherence
  - Hardware support for message-passing



Successfully tested in CERN openlab this summer

Tilera

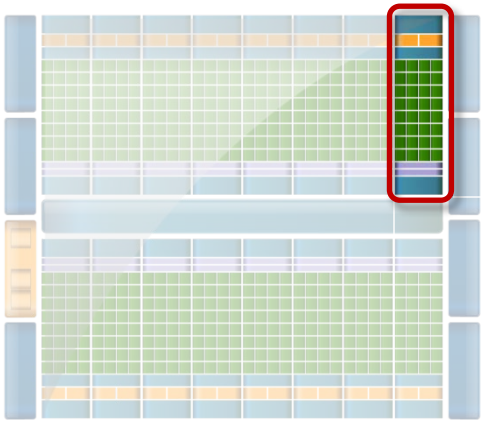
- Currently: 64 way
- This year:
  - Tilera will present the TILE-Gx™ family of processors hitting the 100-core milestone

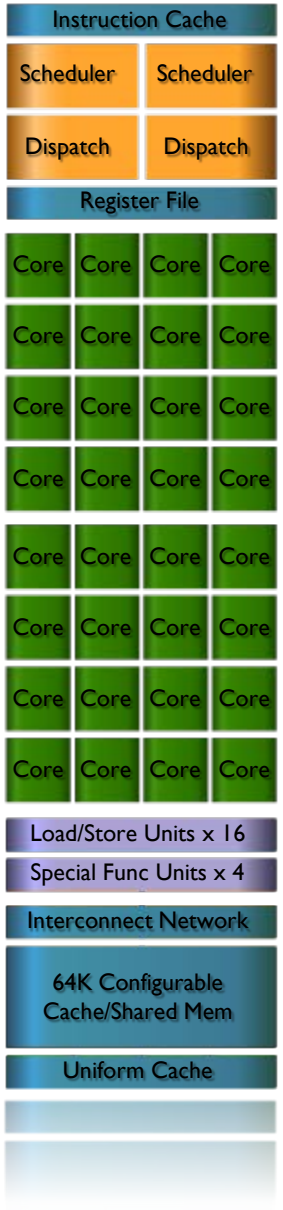


Nvidia GPUs

- Streaming Multiprocessing architecture
- Teraflops (DP) per card:
  - Fermi → Keppler → Maxwell

- But only a few Gigabytes of memory



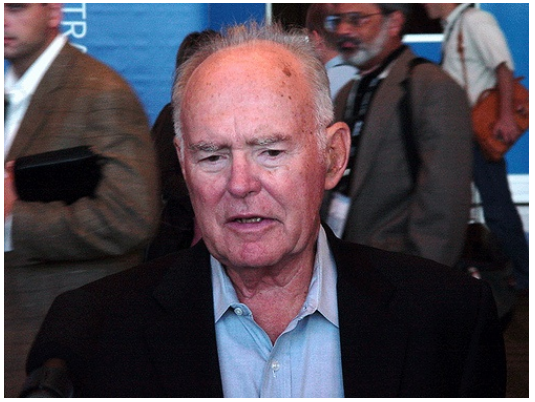


Lots of interest in the HEP on-line community

Evaluation of likelihood functions on CPU and GPU devices (Y.Sneen-Lindal/CERN openlab)

Adapted from Nvidia

Intel roadmap



	2008	2010	2012	2014	2016	2018	2020
Process	45 nm	32 nm	22 nm	14 nm	10 nm	7 nm	5 nm
Frequency scaling	15 %	10 %	8 %	5 %	4 %	3 %	2 %
Vdd scaling	- 10 %	- 7.5 %	- 5 %	-2.5 %	- 1.5 %	- 1 %	- 0.5 %
Transistor density	1.75x	1.75x	1.75x	1.75x	1.75x	1.75x	1.75x

From a recent presentation by Shekhar Borkar/Intel

Progression in layers

Software infrastructure

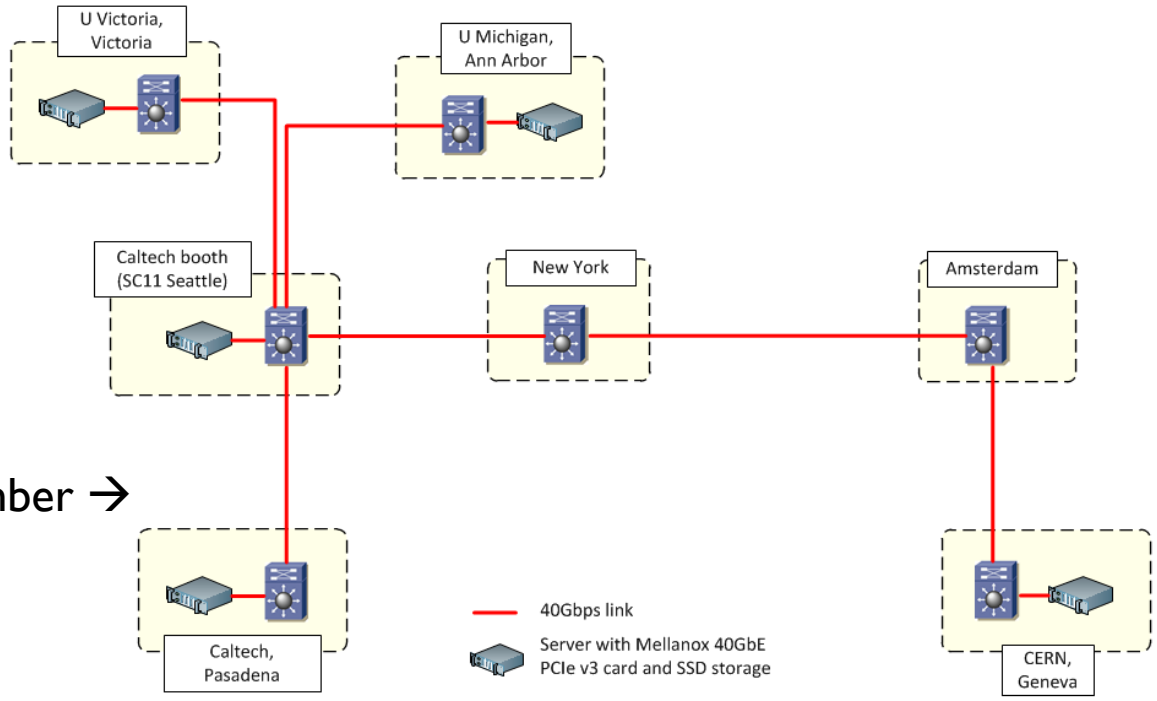
Site infrastructure

Grid/Cloud infrastructure

Site infrastructure

- I expect a lot to move transparently:
  - Virtualization of the resources
  - Higher speed networks
    - 40 GbE
    - 100 GbE

Planned demo at SC11 in November →

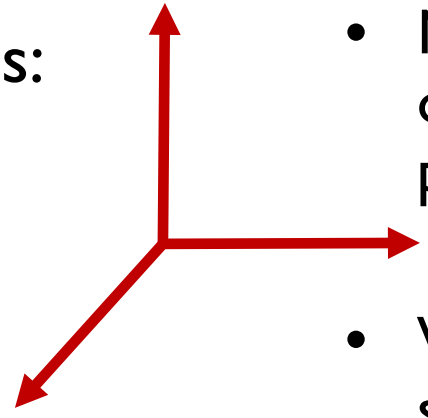


A proposal for the software

- Agile software:
  - Parallelism at all level
    - Events, tracks, vertices, etc.
    - Remove event separation (as proposed by René)
  - Built-in forward scalability
  - Compute-intensive kernels
  - Efficient memory footprint
  - Locality-optimised data layout
  - Broad programming talent

Seven multiplicative dimensions

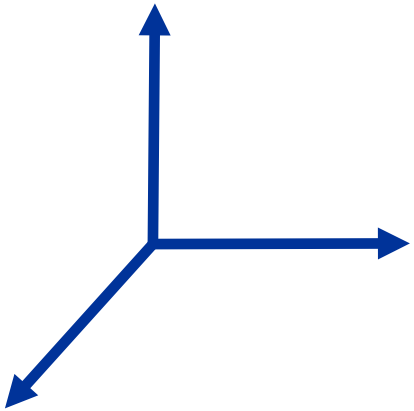
- First three dimensions:
  - Pipelining
  - Superscalar
  - Vector width/SIMD



- Need to understand overall hardware potential
- Where are we on the scale ?
  - 10% ?
  - 90% ?

- Next dimension is a "pseudo" dimension:
  - Hardware multithreading

- Last three dimensions:
  - Multiple cores
  - Multiple sockets
  - Multiple compute nodes



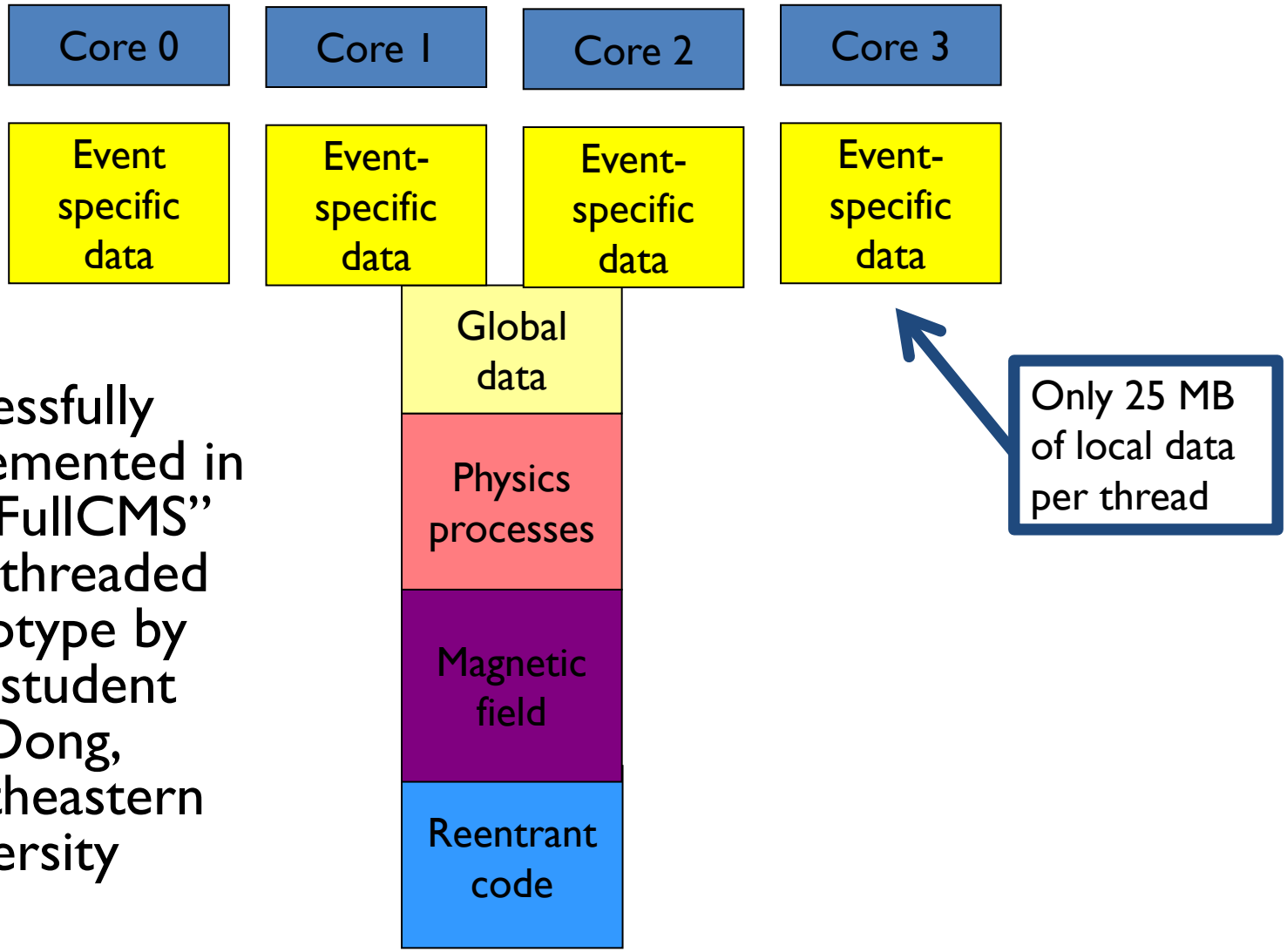
The holy grail: Forward scalability

- A program should be written in such a way that it extracts maximum performance from today's hardware
- In addition, on future processors, performance should scale automatically
- Additional CPU/GPU hardware, be it cores/threads or vectors, would automatically be put to good use
- Scaling would be as expected:
  - If the number of cores (or the vector size) doubled:
    - Scaling would also be 2x, and not just a few percent
- We cannot afford to "rewrite" our software for every hardware change!

Compute-intensive kernels

- Take the whole program and its execution behaviour into account
- Foster clear split:
  - Prepare to compute
  - Perform the heavy computation
  - Post-processing
- Consider exploiting the entire server

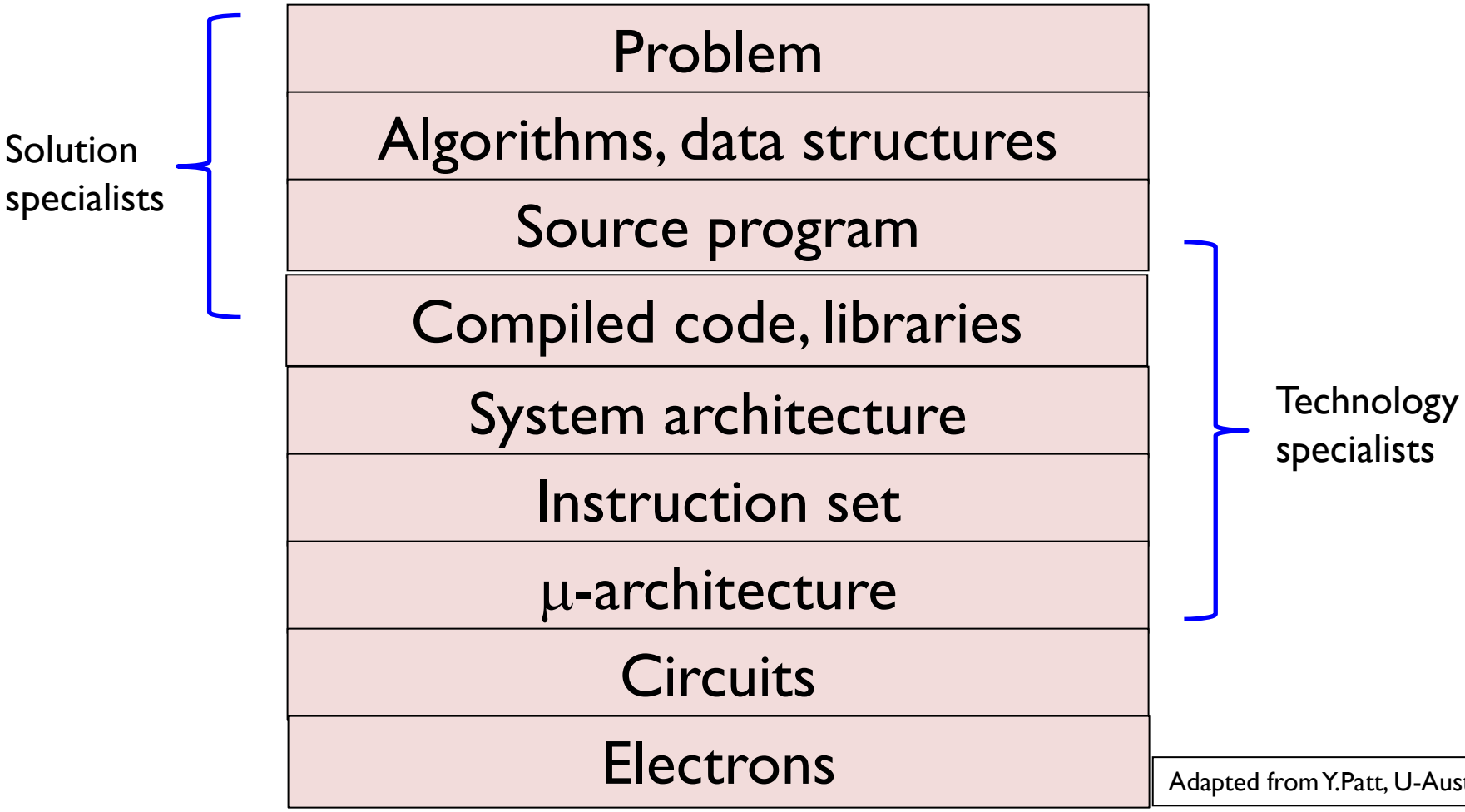
Efficient memory footprint



- Successfully implemented in the "FullCMS" multithreaded prototype by PhD student Xin Dong, Northeastern University

Broad programming talent

- The layers of computing:



Conclusions

- Our horizon should be the computing needs during the next one, two decades
- As we saw for LEP, there may be multiple phase transitions in computing
  - Transparent
  - Non-transparent
  - An agile software strategy will help us to take advantage of the new possibilities quickly

But, don't forget that others (not HEP) are in charge of the evolution!

Other sessions

- Building an Outsourcing Ecosystem for Science (K.Keahey)
- Integrating Amazon EC2 with the CMS Production Framework (A.M.Melo)
- Dynamic deployment of a PROOF-based analysis facility for the ALICE experiment over virtual machines using PoD and OpenNebula (B.Dario)
- The EOS disk storage system at CERN (A.J.Peters)
- Can 'Go' address the multicore issues of today and the manycore problems of tomorrow ? (S.Binet)
- Track finding using GPUs (C.Schmitt)
- Challenges in using GPUs for the reconstruction of digital hologram images. (P.Hobson)
- Offloading peak processing to Virtual Farm by STAR experiment at RHIC (J.Balewski)
- Computing On Demand: Analysis in the Cloud (A.Manafov)
- Multicore in Production: Advantages and Limits of the Multi-process Approach. (V.Tsulaia)
- Moving ROOT Forward. (F.Rademakers)
- Evaluation of likelihood functions on CPU and GPU devices (Y.Sneen-Lindal)
- Offloading peak processing to Virtual Farm by STAR/RHIC (J.Balewski)