POLITECNICO DI MILANO

NECST laboratory

Parallelism in wonderland:

are you ready to see how deep the rabbit hole goes?

Class introduction

- A bird's eye view on the 2014 class -

Marco D. Santambrogio: marco.santambrogio@polimi.it

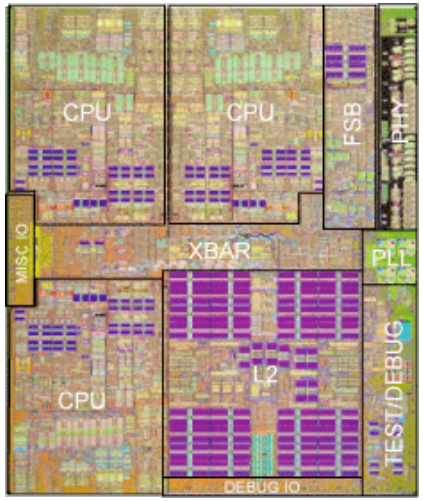
Simone Campanoni: xan@eecs.harvard.edu

CHANGE

One core to rule them all

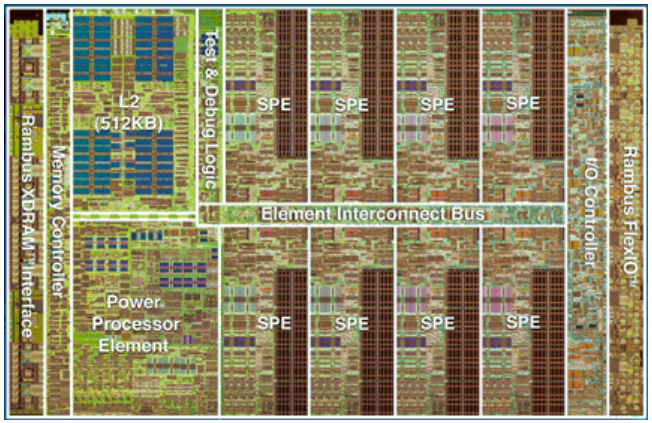
Xenon: XBOX360

- Three symmetrical cores
  - each two way SMT-capable and clocked at 3.2 GHz



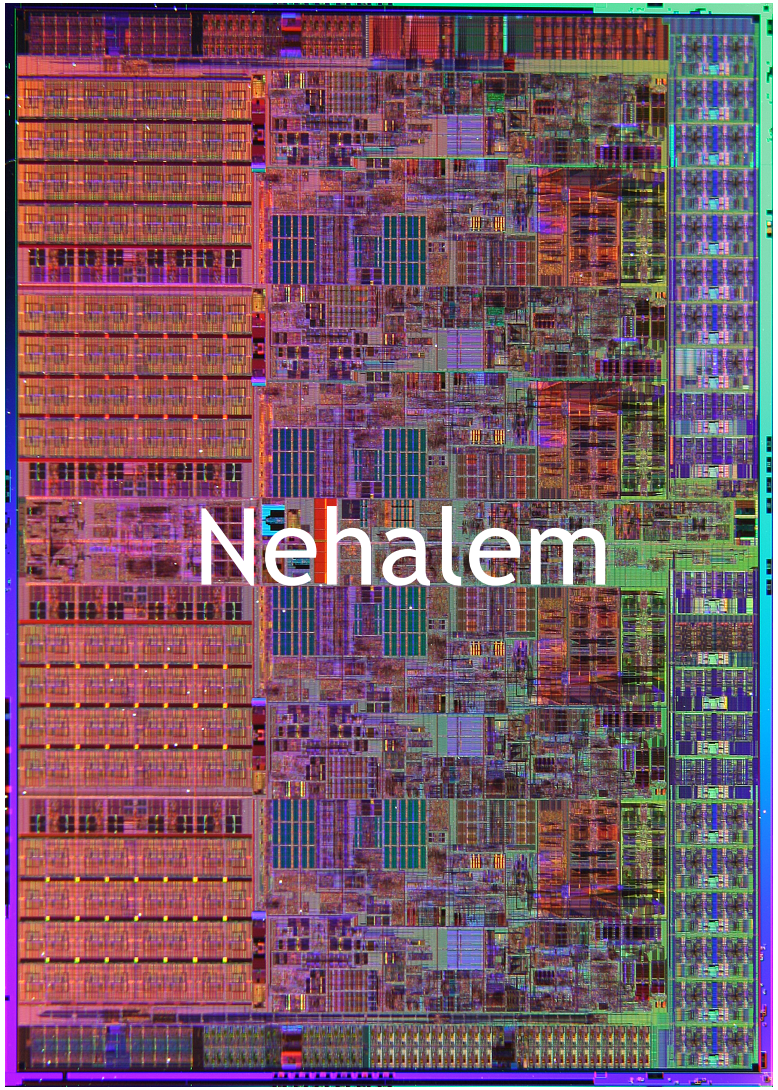
Cell: PS3

- Cell is a heterogeneous chip multiprocessor
  - One 64-bit Power core
  - 8 specialized co-processors



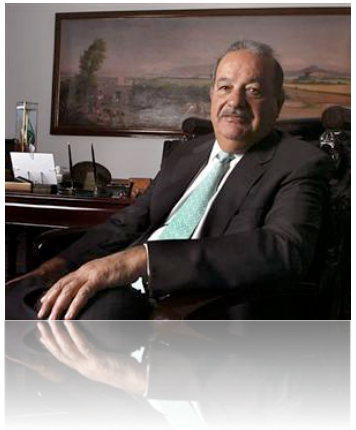
Technology constantly on the move!

- Num of transistors not limiting factor
  - Currently ~ 1 billion transistors/chip
  - Problems:
    - Too much Power, Heat, Latency
    - Not enough Parallelism
- 3-dimensional chip technology?
  - Sandwiches of silicon
  - “Through-Vias” for communication
- On-chip optical connections?
  - Power savings for large packets
- The Intel® Core™ i7 microprocessor (“Nehalem”)
  - 4 cores/chip
  - 45 nm, Hafnium hi-k dielectric
  - 731M Transistors
  - Shared L3 Cache - 8MB
  - L2 Cache - 1MB (256K x 4)



Reasons to attend this class…

World's richest people



- **Carlos Slim**, $73B
- 1°: 2013, 2012, 2011, 2010
- 2°: 2008
- 3°: 2009, 2007, 2006

- **Bill Gates**, $54B
- 2°: 2013, 2012, 2011, 2010
- 3°: 2008
- 1°: 2009, 2007, 2006, …, 1995

- **Sergey Brin**, $22.8B
- 21°: 2013

Things to avoid

- Wikipedia
  - http://www.wikipedia.org/
- Wikipedia
  - http://www.wikipedia.org/
- Wikipedia
  - http://www.wikipedia.org/

Something against Wikipedia? Why?

The World's Billionaires

Edited by Luisa Kroll, Matthew Miller and Tatiana Serafin 03.11.09, 6:00 PM EDT

The richest people in the world have gotten poorer, just like the rest of us. This year the world's billionaires have an average net worth of $3 billion, down 23% in 12 months. The world now has 793 ...naires, down from 1,125 a year ago.

After slipp... ...rs, the U.S. is regaining its domi... ...f wealth. Americans account for ...f the list's slots, up seven and ...m last year, respectively. Bill Gates ...gained his title as the world's richest man. ...st year's No. 1, saw his fortune decline $25 ...es of Berkshire Hathaway fell nearly 50% in 12 ... Mexican telecom titan Carlos Slim Helú maintains his ...t in the top three but lost $25 billion.

More ...

Billionaire Bust

Last year the world had 1,125 billionaires. Today there are 793. How $1.4 trillion vanished.

Outline

- Class Topics: how to turn lead into gold
- Class organization
- End?...

Topics

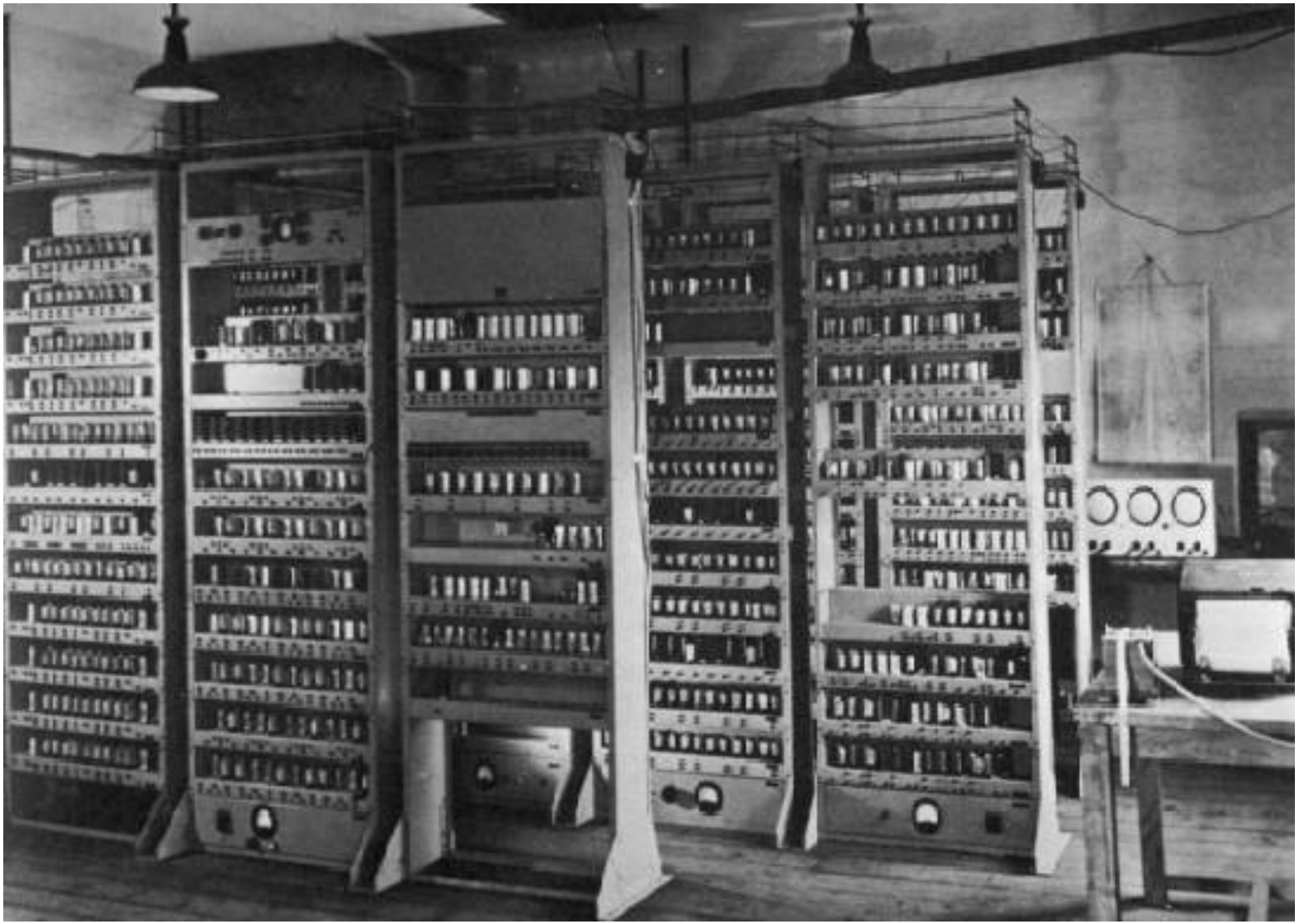
- Computer Architecture
  - Internal Parallelism in processors:
    - Pipelining
    - Instruction level parallelism inside processors
    - How to get more performance? Superscalar processors
  - Process level parallelism
    - Thread-level
    - Multiprocessor architectures
  - Cache memories in the multicore era

- Compiler
  - Thread Level Parallelism
    - Different sources
    - Declaring vs. inferring TLP
  - Implications of TLP
    - Practical benefits and disadvantages
    - Performance variation
    - Communication
    - Developing

How to turn lead into gold

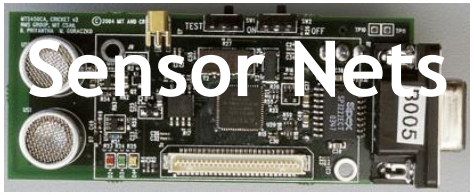
The Alchemist (1771), Joseph Wright of Derby
(Derby Museum and Art Gallery, Derby, UK)

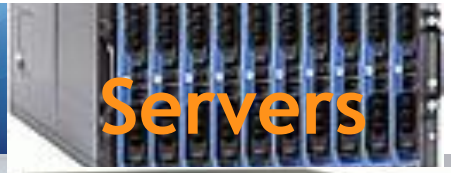
Computing Devices Then…

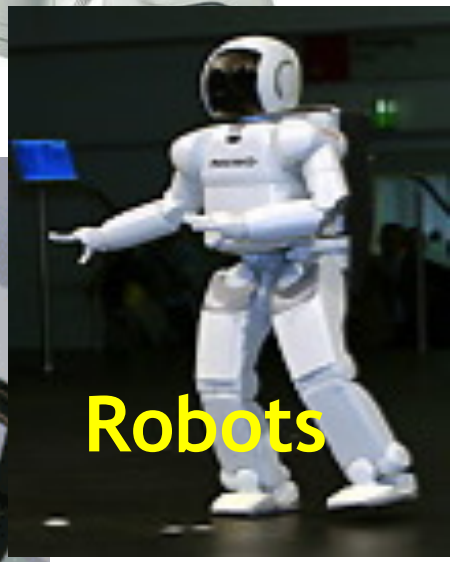


EDSAC, University of Cambridge, UK, 1949

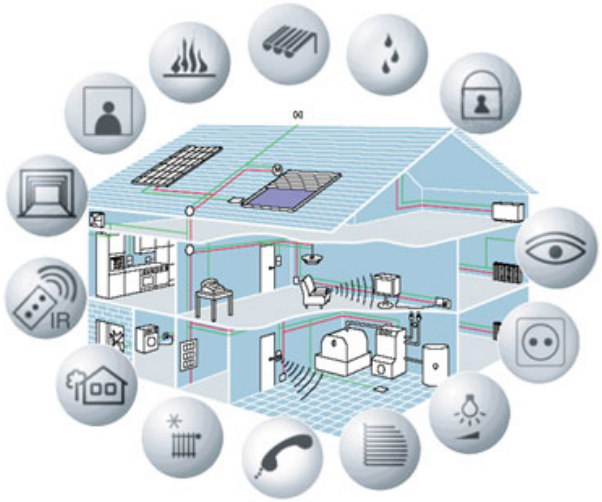
Computing Devices Now





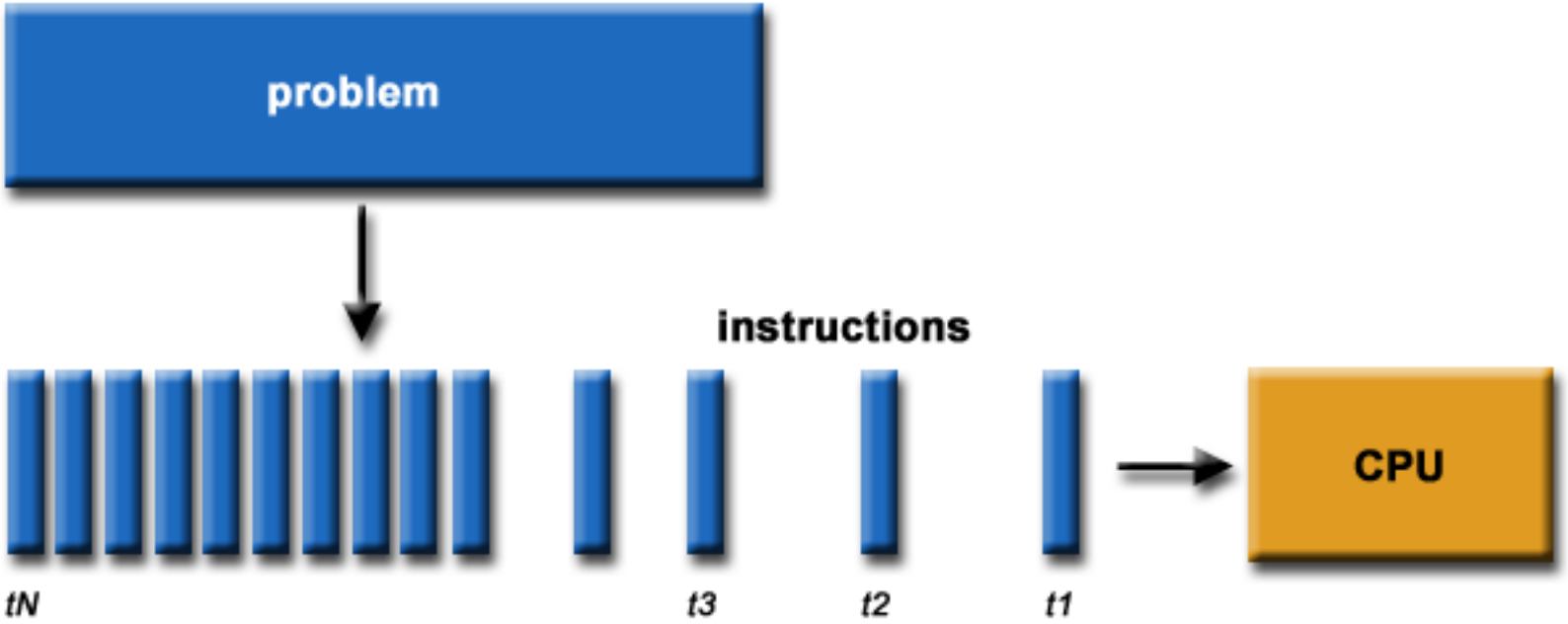


But also…



"Traditional" Computation

- Software is written for serial computation
  - It has to be executed on a single computer having a single Central Processing Unit (CPU)
  - A problem is broken into a discrete series of instructions
    - Instructions are executed one after another
    - Only one instruction may execute at any moment in time.



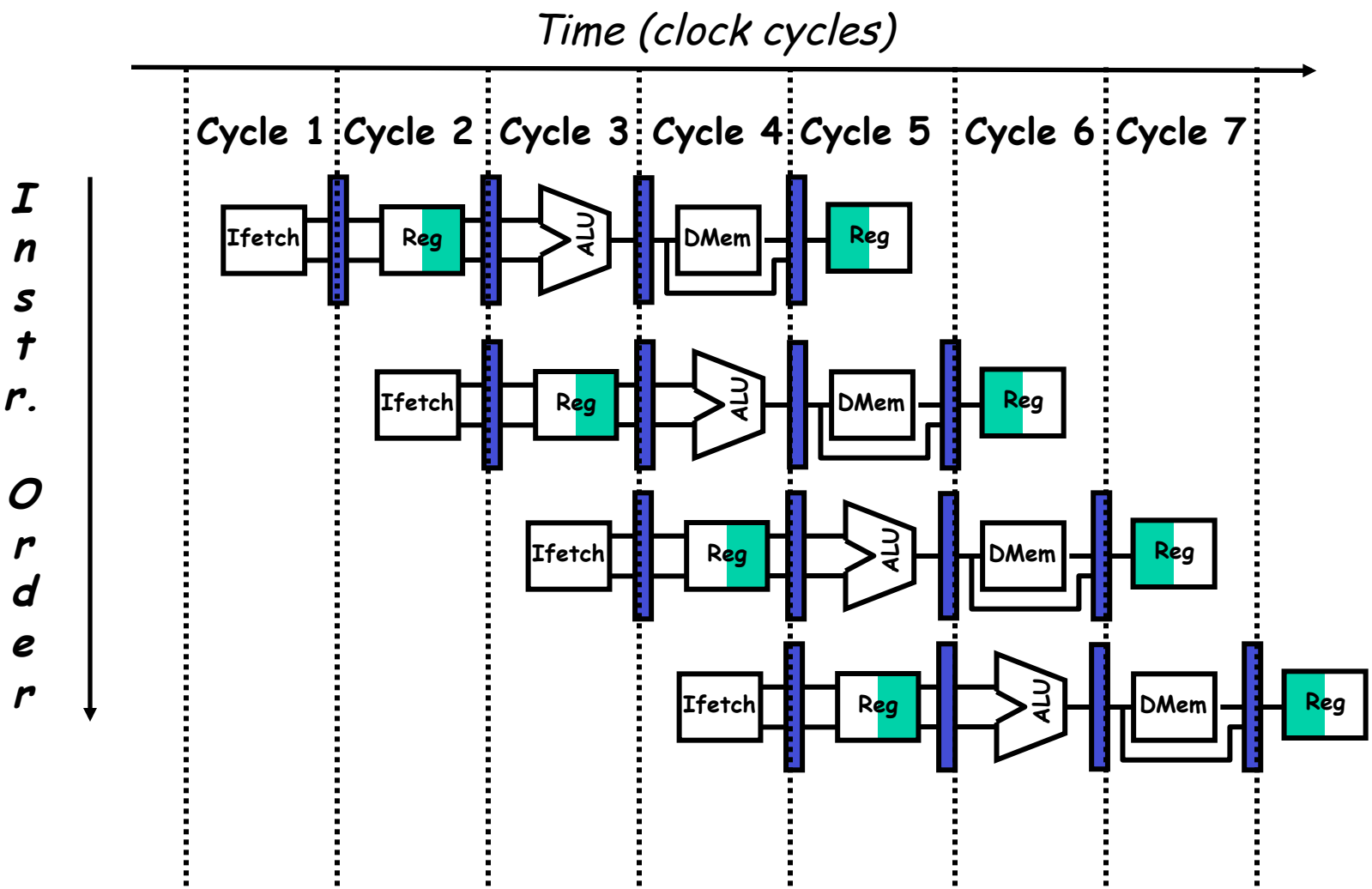
But the world is *"parallel"*

- Events are happening simultaneously
  - Many complex, interrelated events happening at the same time, yet within a sequence:

- Some examples:
  - Galaxy formation
  - Planetary movement
  - Tectonic plate drift
  - Rush hour traffic
  - Automobile assembly line
  - Building a space shuttle
  - Ordering a hamburger at the drive through

Parallel?: which kind of parallelism?

Pipelined Instruction Execution



Limits to pipelining

- Maintain the von Neumann "illusion" of one instruction at a time execution

- Hazards prevent next instruction from executing during its designated clock cycle
  - Structural hazards: attempt to use the same hardware to do two different things at once
  - Data hazards: Instruction depends on result of prior instruction still in the pipeline
  - Control hazards: Caused by delay between the fetching of instructions and decisions about changes in control flow (branches and jumps).

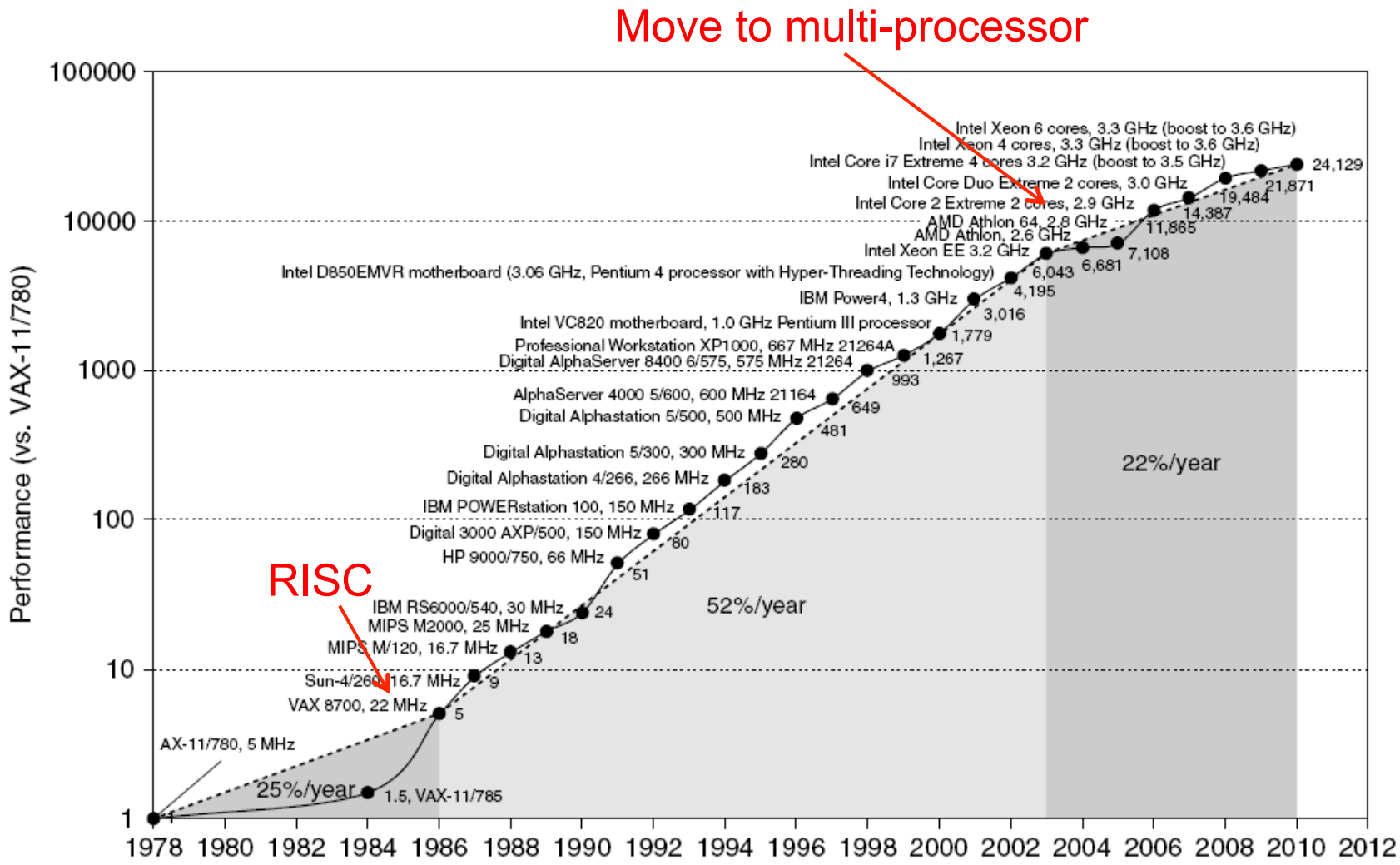
Progression of ILP

- 1st generation RISC – pipelined
  - Full 32-bit processor fit on a chip ->issue almost 1 IPC
    - Need to access memory 1+x times per cycle
  - Floating-Point unit on another chip
  - Cache controller a third, off-chip cache
  - 1 board per processor -> multiprocessor systems
- 2nd generation: superscalar
  - Processor and floating point unit on chip (and some cache)
  - Issuing only one instruction per cycle uses at most half
  - Fetch multiple instructions, issue couple
    - Grows from 2 to 4 to 8 …
  - How to manage dependencies among all these instructions?
  - Where does the parallelism come from?
- VLIW
  - Expose some of the ILP to compiler, allow it to schedule instructions to reduce dependences

Modern ILP

- Dynamically scheduled, out-of-order execution
- Current microprocessor fetch 10s of instructions per cycle
- Pipelines are 10s of cycles deep
  - many 10s of instructions in execution at once
- Grab a bunch of instructions, determine all their dependences, eliminate dep's wherever possible, throw them all into the execution unit, let each one move forward as its dependences are resolved
- Appears as if executed sequentially
- On a trap or interrupt, capture the state of the machine between instructions perfectly
- Huge complexity

Single Processor Performance



Current Trends in Architecture

- Cannot continue to leverage Instruction-Level parallelism (ILP)
  - Single processor performance improvement ended in 2003
- New models for performance:
  - Data-level parallelism (DLP)
  - Thread-level parallelism (TLP)
  - Request-level parallelism (RLP)
- These require explicit restructuring of the application… one of the reasons behind this joint CA-Compiler class

When all else fails - guess

- Programs make decisions as they go
  - Conditionals, loops, calls
  - Translate into branches and jumps (1 of 5 instructions)
- How do you determine what instructions for fetch when the ones before it haven't executed?
  - Branch prediction
  - Lot's of clever machine structures to predict future based on history
  - Machinery to back out of mis-predictions
- Execute all the possible branches
  - Likely to hit additional branches, perform stores

  ⇒speculative threads

  ⇒What can hardware do to make programming (with performance) easier?

Have we reached the end of ILP?

- Multiple processor easily fit on a chip
- Every major microprocessor vendor has gone to multithreading
  - Thread: loci of control, execution context
  - Fetch instructions from multiple threads at once, throw them all into the execution unit
  - Intel: hyperthreading,
  - Concept has existed in high performance computing for 20 years (or is it 40? CDC6600)
- Vector processing
  - Each instruction processes many distinct data
  - Ex: MMX
- Raise the level of architecture – many processors per chip

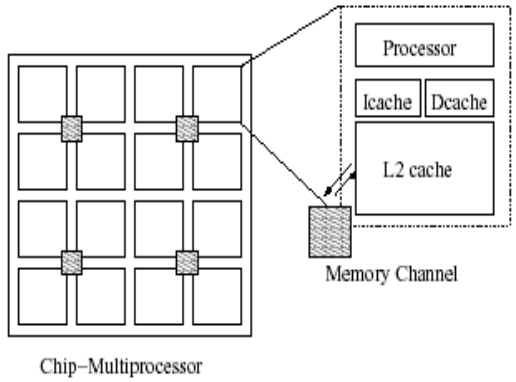
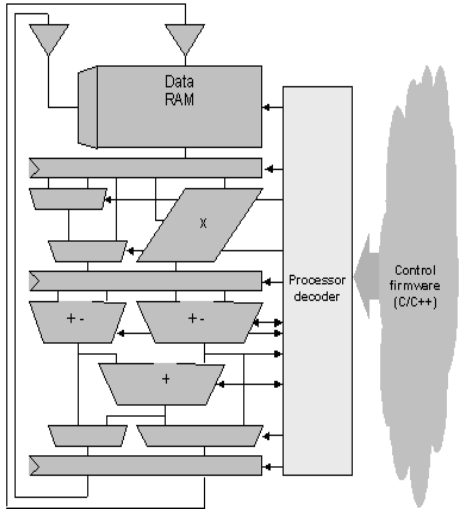


Figure 1. Chip-multiprocessor model.



Tensilica Configurable Proc

Classes of Computers

- Personal Mobile Device (PMD)
  - e.g. start phones, tablet computers
  - Emphasis on energy efficiency and real-time
- Desktop Computing
  - Emphasis on price-performance
- Servers
  - Emphasis on availability, scalability, throughput
- Clusters / Warehouse Scale Computers
  - Used for “Software as a Service (SaaS)”
  - Emphasis on availability and price-performance
  - Sub-class: Supercomputers, emphasis: floating-point performance and fast internal networks
- Embedded Computers
  - Emphasis: price

Computer Architecture Topics: Beyond ILP

- ILP architectures (superscalar, VLIW...):
  - Support fine-grained, instruction-level parallelism;
  - Fail to support large-scale parallel systems;
- Multiple-issue CPUs are very complex, and returns (as far as extracting greater parallelism) are diminishing
  - extracting parallelism at higher levels becomes more and more attractive.
- A further step: ***process- and thread-level parallel architectures.***
- To achieve ever greater performance: ***connect multiple microprocessors in a complex system.***

Computer Architecture Topics: Parallel Architectures

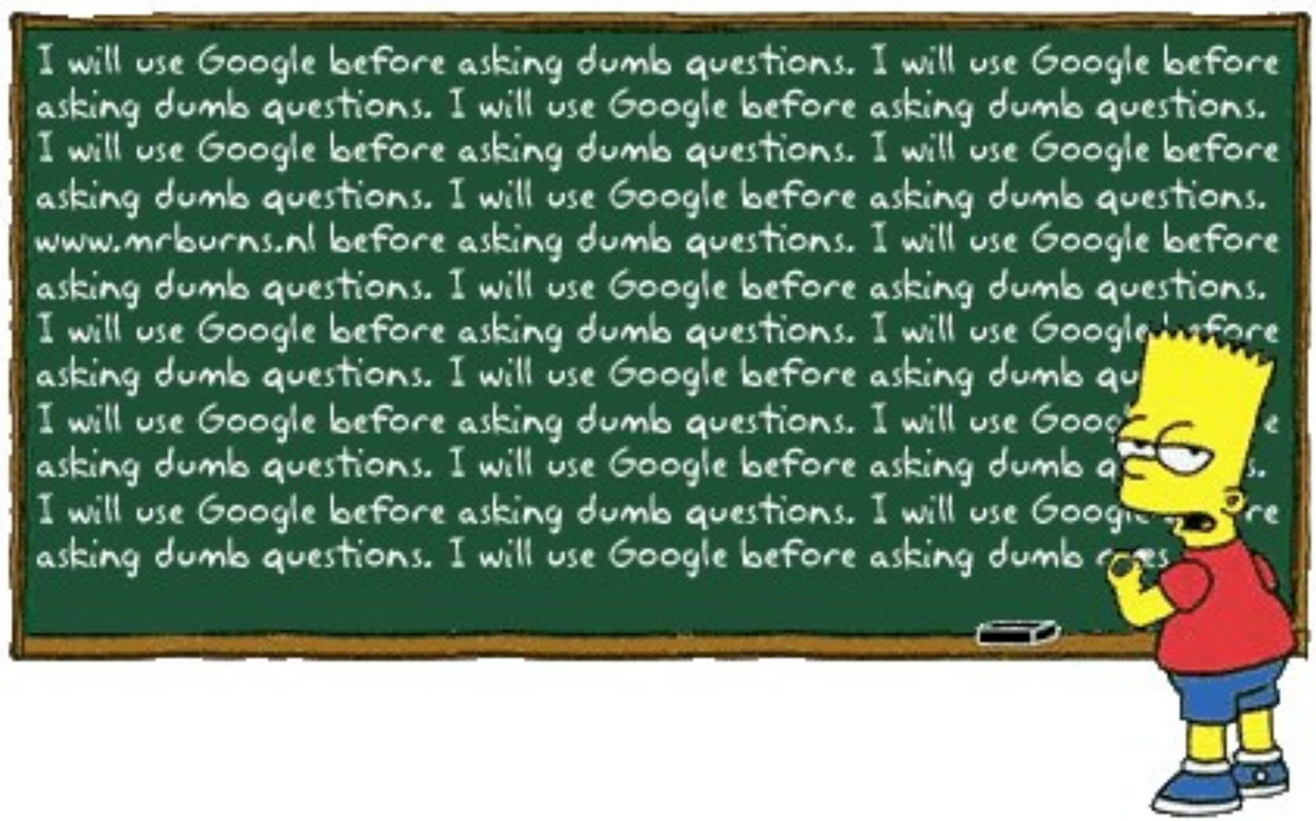
- Definition: “A parallel computer is a collection of processing elements that cooperates and communicate to solve large problems fast"

  Almasi and Gottlieb, Highly Parallel Computing, 1989

- The aim is to replicate processors to add performance vs design a faster processor.

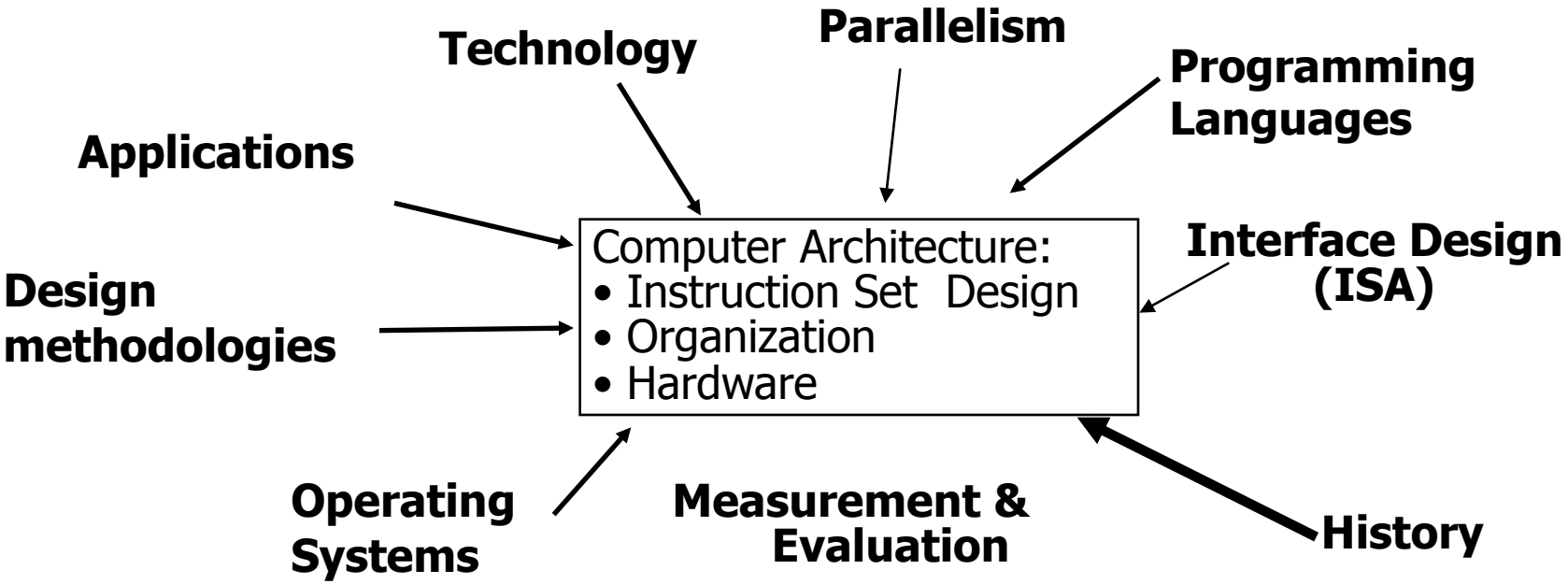
- Parallel architecture extends traditional computer architecture with a communication architecture
  - abstractions (HW/SW interface)
  - different structures to realize abstraction efficiently

Class Organization



Course Focus

- Understand design techniques, machine structures, technology factors, evaluation methods, design of computer architectures



Topic Coverage

- Textbook: Hennessy and Patterson, *Computer Architecture: A Quantitative Approach*, 4th Ed. (now there is the new 5th edition)
- Slides and papers
  - Can be found on Santambrogio's website (http://home.dei.polimi.it/santambr)

Lectures

- Lecture 1: intro - Wed 15, from 2pm to 3.30 pm (1.5h)

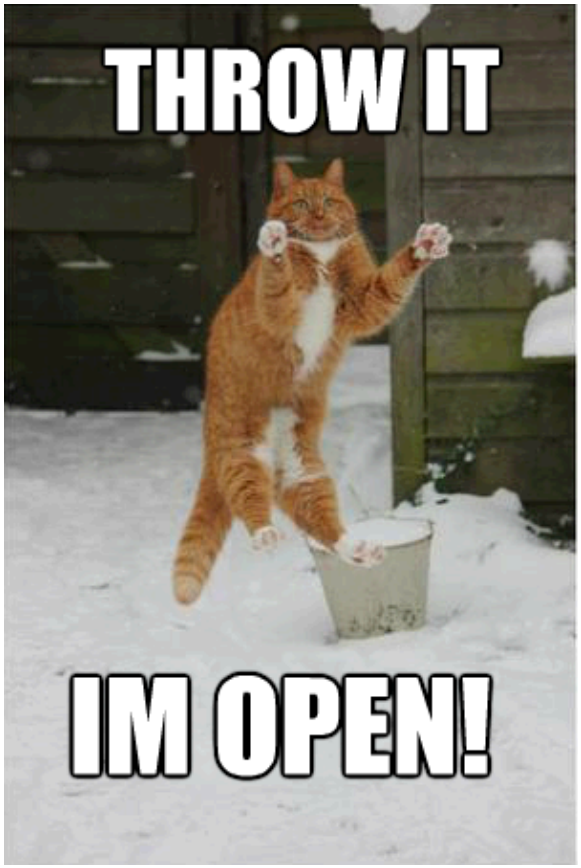
- Computer architecture (@ Aula Seminari Alessandra Alario - 4th Floor - Building 21)
  - Lecture 2 - Thr 16, from 10am to 12.30pm (2.30h - 4t)
  - Lecture 3 - Thr 16, from 2pm to 4pm (2h - 6t)
  - Lecture 4 - Fri 17, from 3.30pm to 6.30pm (3h – 9t)

- Compiler (@ Aula Seminari Alessandra Alario - 4th Floor - Building 21)
  - Lecture 5 - Mon 20, from 9am to 12pm (3h - 12t)
  - Lecture 6 - Tue 21, from 9am to 1pm (4h - 16t)
    - @ PT1 - Ground Floor - Building 20
  - Lecture 7 - Thr 23, from 9am to 1pm (4h - 20t)

- Exam/Papers' Assignment (@ Aula Seminari Alessandra Alario - 4th Floor - Building 21)
  - Papers' Assignment - Mon 20, from 2pm to 4pm
  - Exam - Fri 24, from 2pm to 6pm

Exams

- When: Fri 24, from 2pm to 6pm
- Where: @ Aula Seminari Alessandra Alario
  - 4th Floor - Building 21



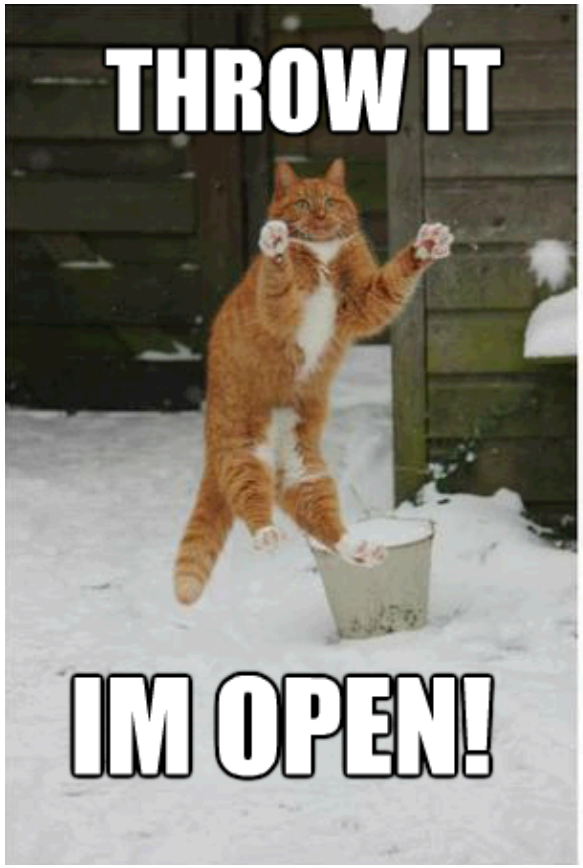
Exams

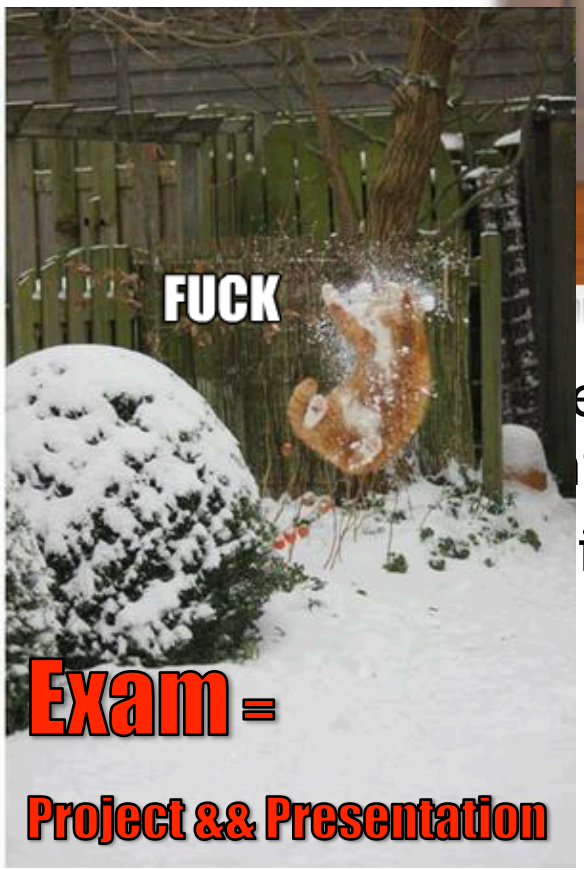
- When: Fri 24, from 2pm to 6pm
- Where: @ Aula Seminari Alessandra Alario
  - 4th Floor - Building 21

- What (**thrown**):
  - **Present and defend** to the rest of the class a research from a set of works we will briefly introduce on Monday
  - **Write a compiler pass** in ILDJIT to estimate an assigned type of parallelism of an assigned set of programs we will provide

Exams

- When: Fri 24, from 2pm to 6pm
- Where: @ Aula Seminari Alessandra Alario
  - 4th Floor - Building 21





e class a research
troduce on Monday

imate an assigned
of programs we

Contacts and Office Hours

- Simone Campanoni

Contact:

- email: xan@eecs.harvard.edu
- Gmail: simo.xan@gmail.com
- Skype: campanoni.simone

- Marco D. Santambrogio

Contact:

- email: marco.santambrogio@polimi.it
- skype: marco.santambrogio
- office: 3492 (DEI, first floor)

END?

Are you ready to see how deep the rabbit-hole goes?…