What does the Future Hold?

Prof. Kavita Bala and Prof. Hakim Weatherspoon

CS 3410, Spring 2014

Computer Science

Cornell University

Announcements

Final Project

Demo Sign-Up via CMS.

sign up Tuesday, May 13th

or Wednesday, May 14th

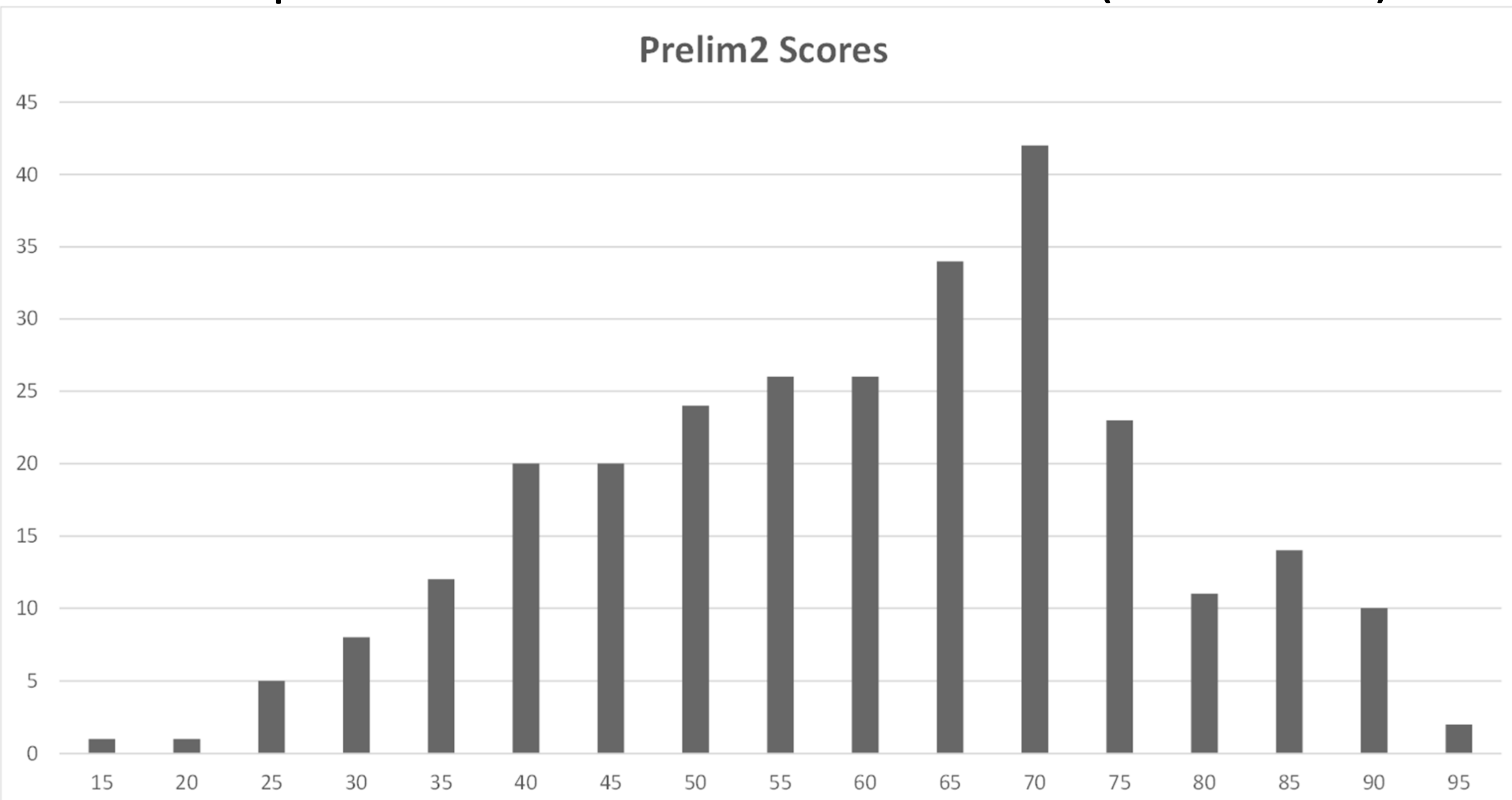
CMS submission due:

- Due 6:30pm Wednesday, May 14th

Announcements

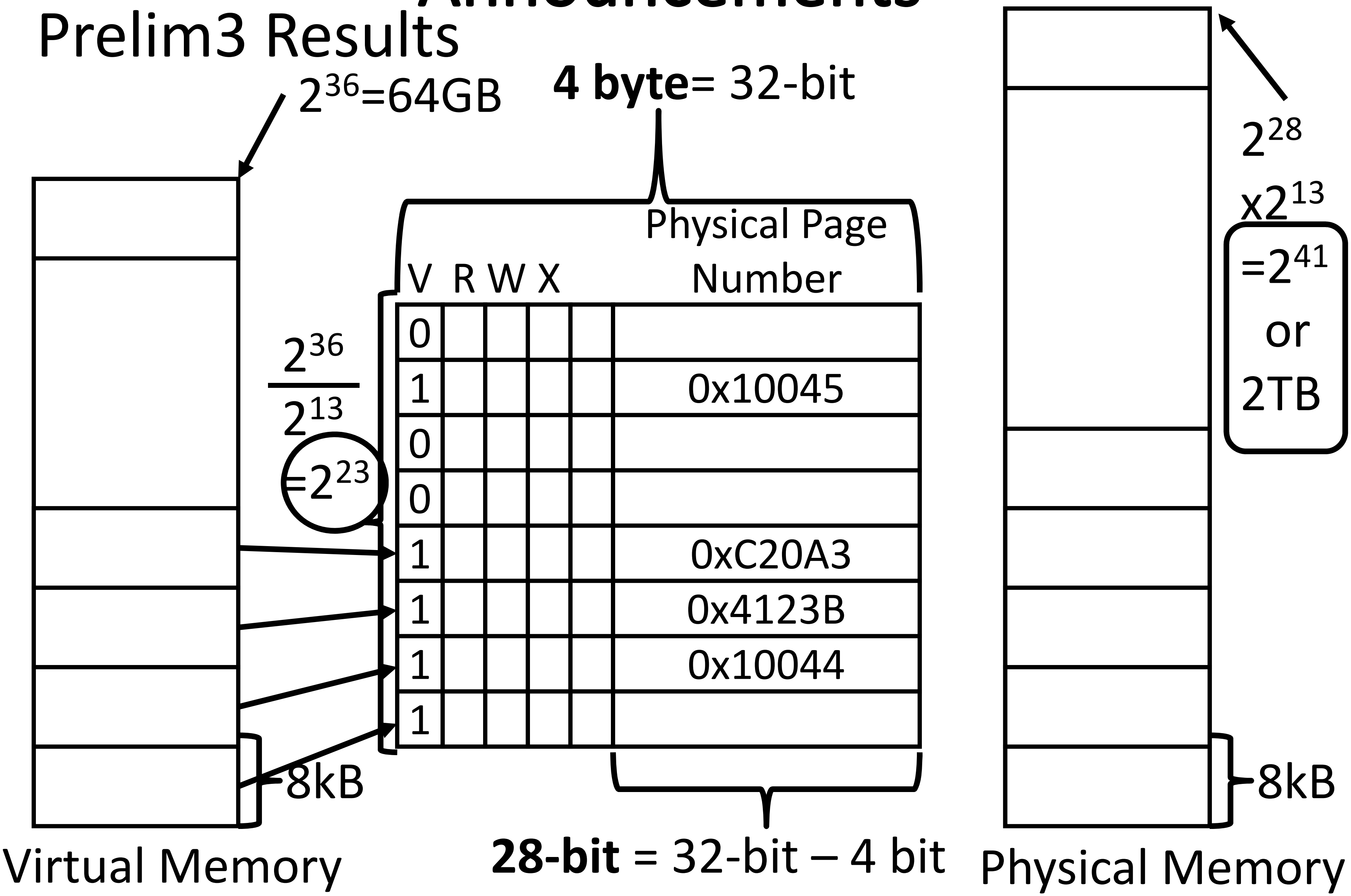
Prelim3 Results

- Mean 58 ± 16.2 (median 60), Max 93
- Pickup in Homework Passback Room (216 Gates)



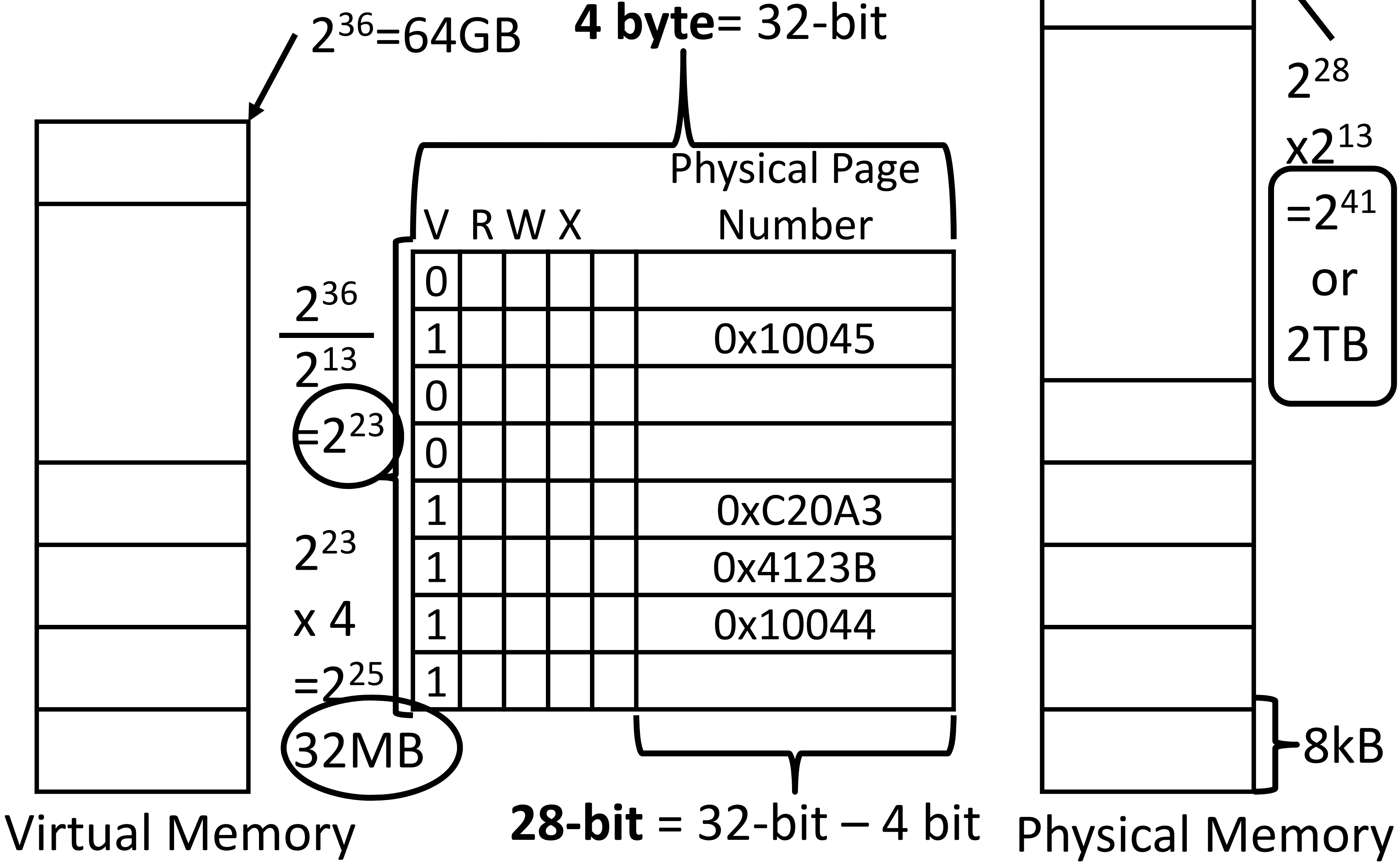
Announcements

Prelim3 Results



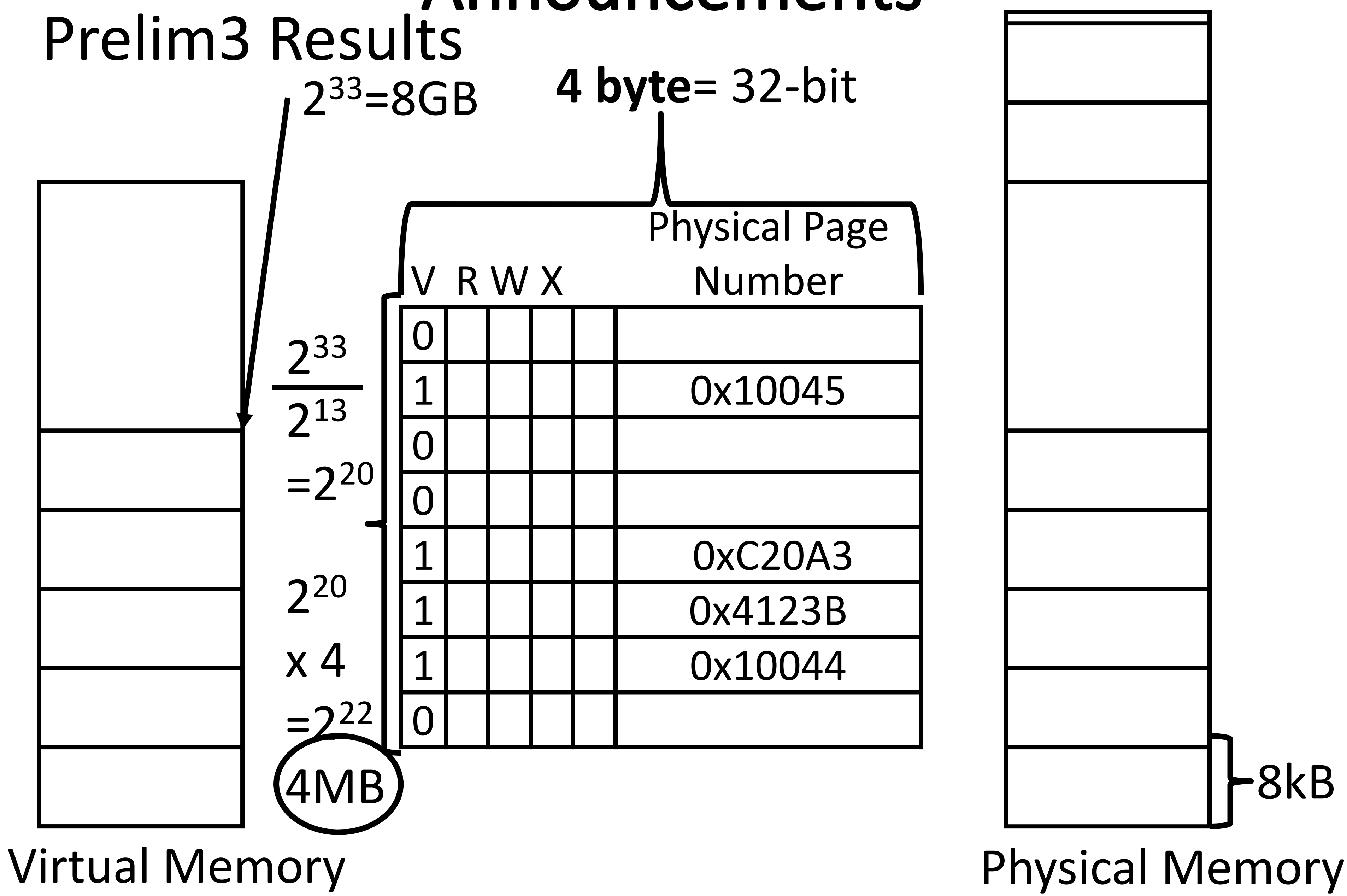
Announcements

Prelim3 Results

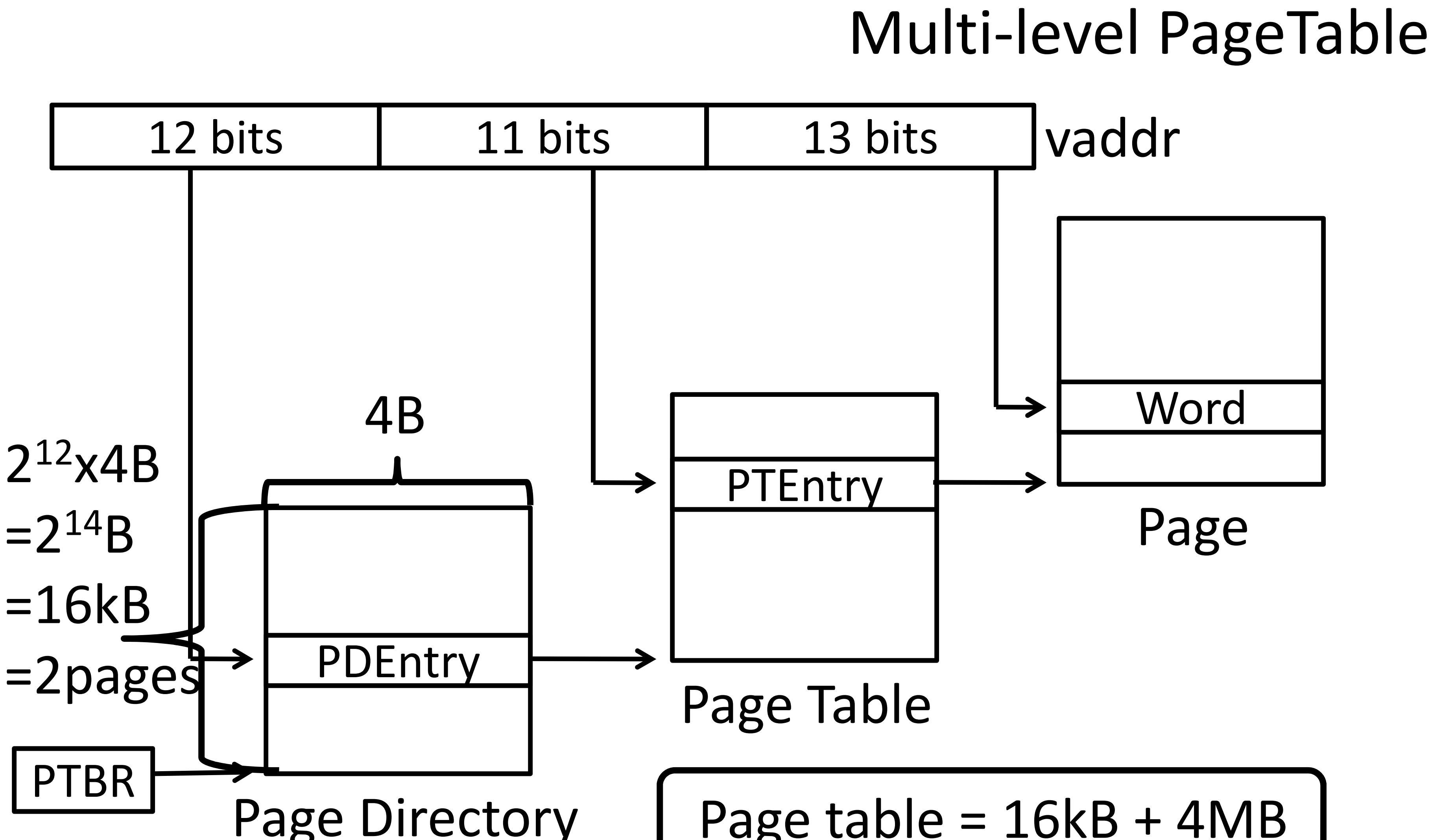


Announcements

Prelim3 Results



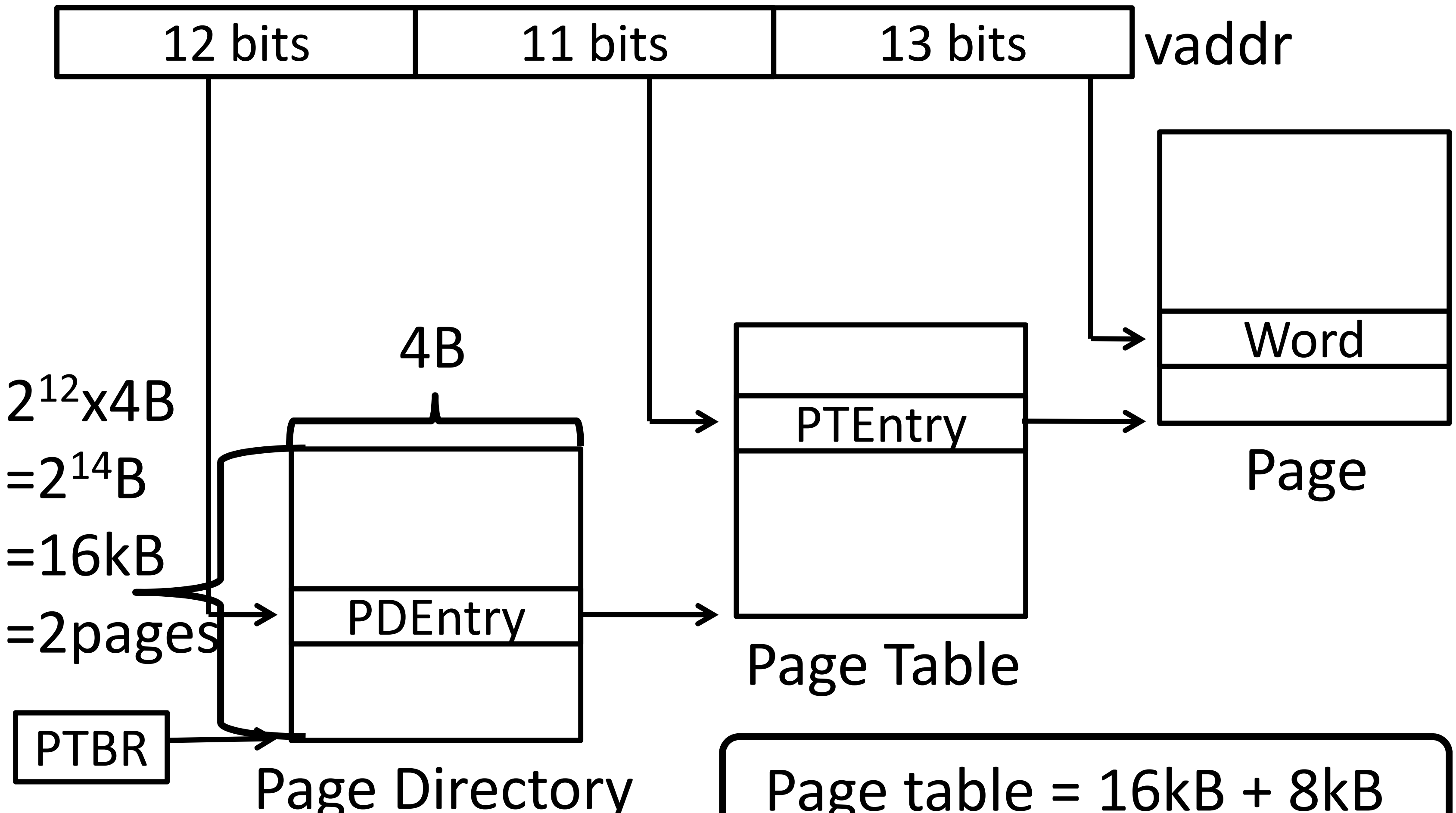
Announcements



Page table = 16kB + 4MB

Announcements

Multi-level PageTable



Page table = 16kB + 8kB

Announcements

How to improve your grade?

Submit a course evaluation and drop lowest in-class lab score

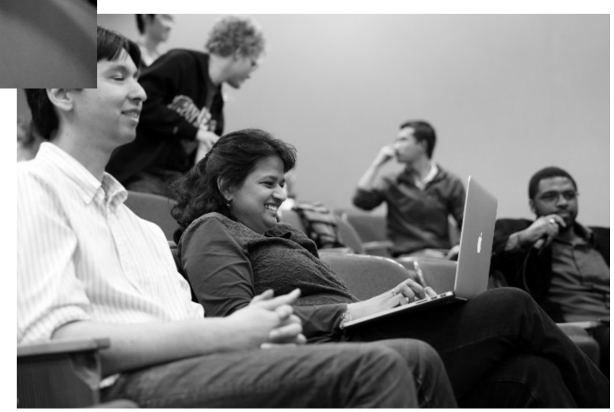
- To receive credit, Submit before Monday, May 12th

Announcements

CacheRace Games Night was great!

- Winner: Team ***balabot***

Adwit Tumuluri and **Arjun Biddanda**

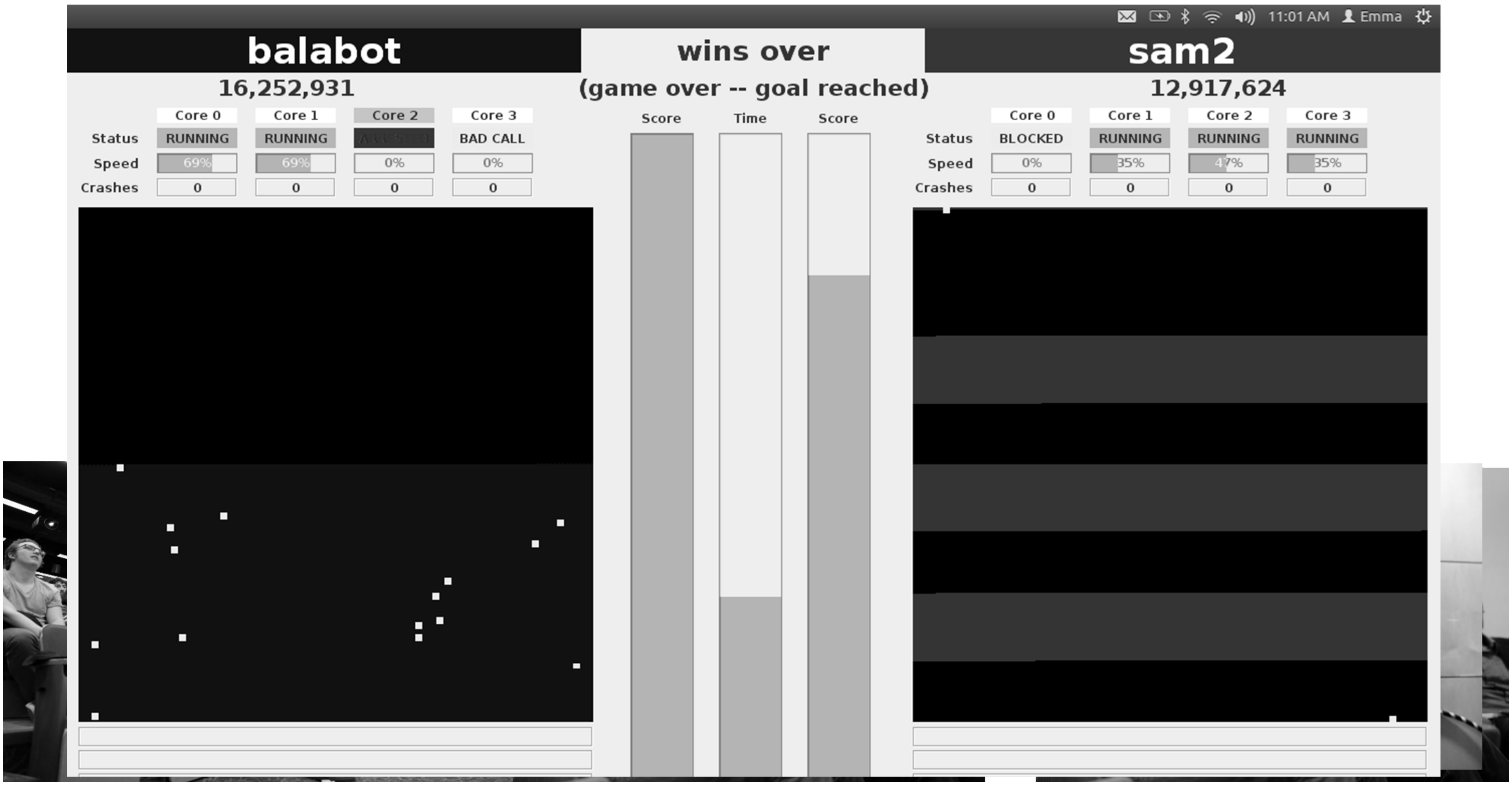


Announcements

CacheRace Games Night was great!

- Winner: Team ***balabot***

  Adwit Tumuluri and **Arjun Biddanda**



Announcements

CacheRace Games Night was great!

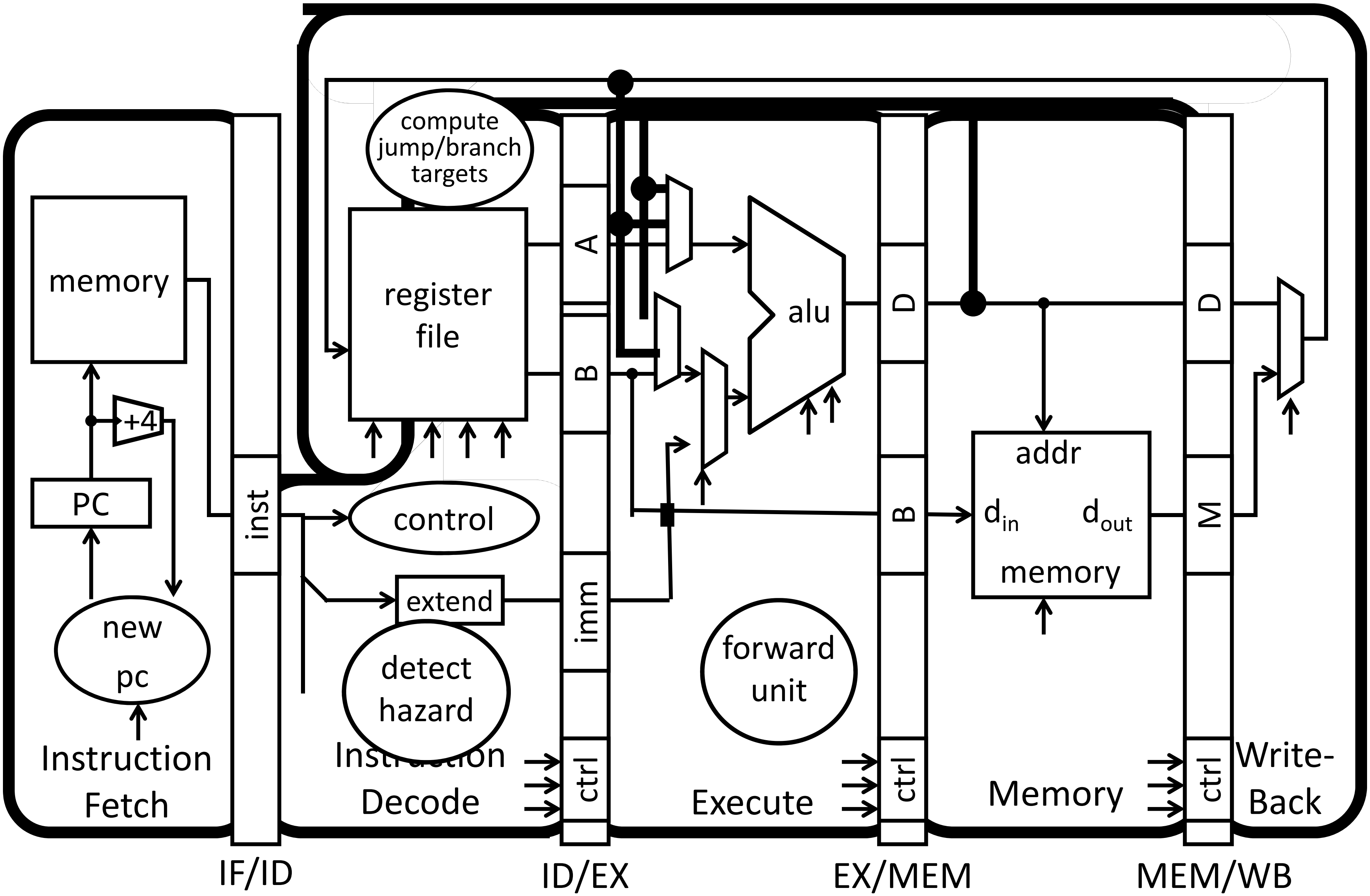
- Champion of Champions: 2014 vs 2011

balabot (2014) vs **hakimPeterspoon (2011)**

Big Picture about the Future

Big Picture

How a processor works? How a computer is organized?



What’s next?

More of Moore

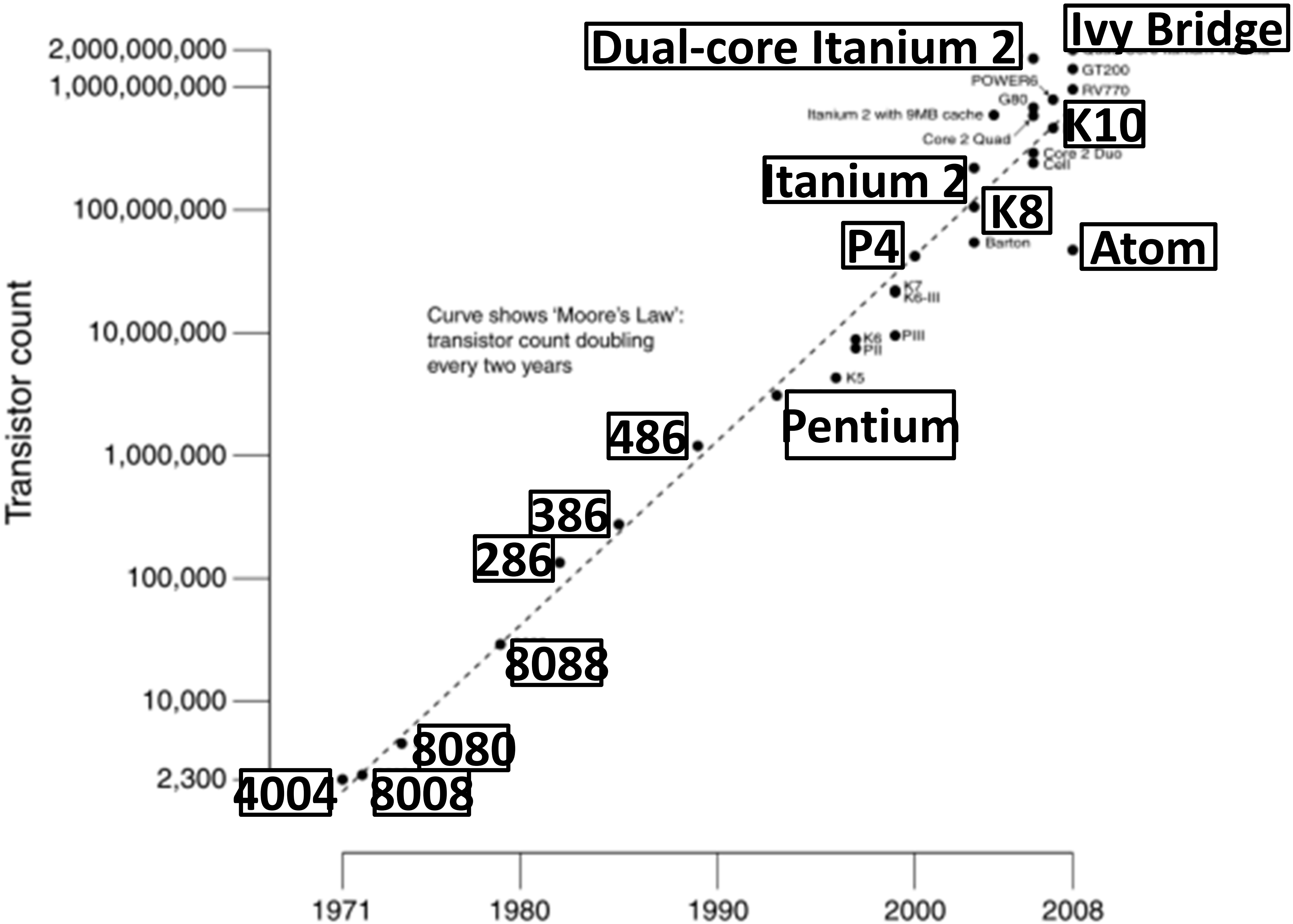
Moore’s Law

Moore’s Law introduced in 1965

- Number of transistors that can be integrated on a single die would double every 18 to 24 months (i.e., grow exponentially with time).

Amazingly visionary

- 2300 transistors, 1 MHz clock (Intel 4004) - 1971
- 16 Million transistors (Ultra Sparc III)
- 42 Million transistors, 2 GHz clock (Intel Xeon) – 2001
- 55 Million transistors, 3 GHz, 130nm technology, 250mm2 die (Intel Pentium 4) – 2004
- 290+ Million transistors, 3 GHz (Intel Core 2 Duo) – 2007
- 731 Million transistors, 2-3Ghz (Intel Nehalem) – 2009
- 1.4 Billion transistors, 2-3Ghz (Intel Ivy Bridge) – 2012



Transistor count
2,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
2,300
1971
1980
1990
2000
2008
Curve shows 'Moore's Law':
transistor count doubling
every two years
Dual-core Itanium 2
Ivy Bridge
GT200
POWER6
RV770
G80
Itanium 2 with 9MB cache
K10
Core 2 Quad
Core 2 Duo
Cell
Itanium 2
K8
Barton
P4
Atom
K7
K6-III
K6
PII
PIII
K5
Pentium
486
386
286
8088
8080
4004
8008

Why Multicore?

Moore's law

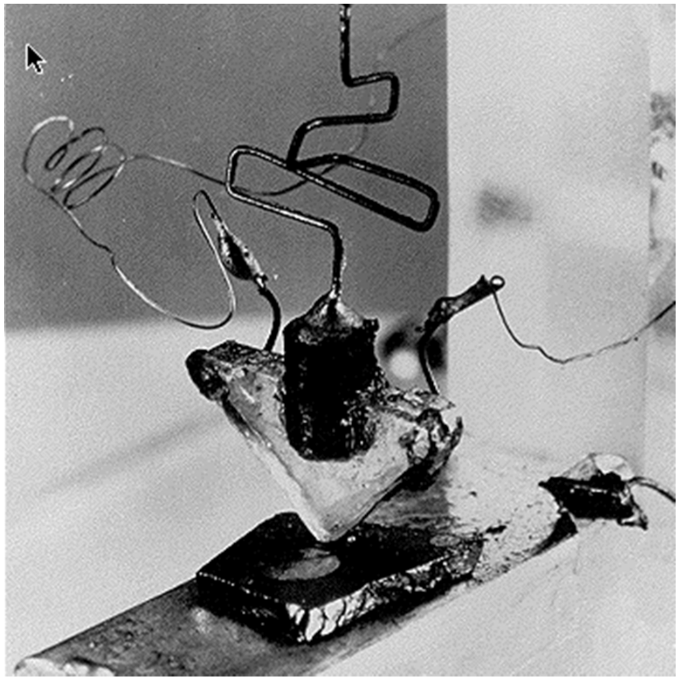
- A law about transistors
- Smaller means more transistors per die
- And smaller means faster too

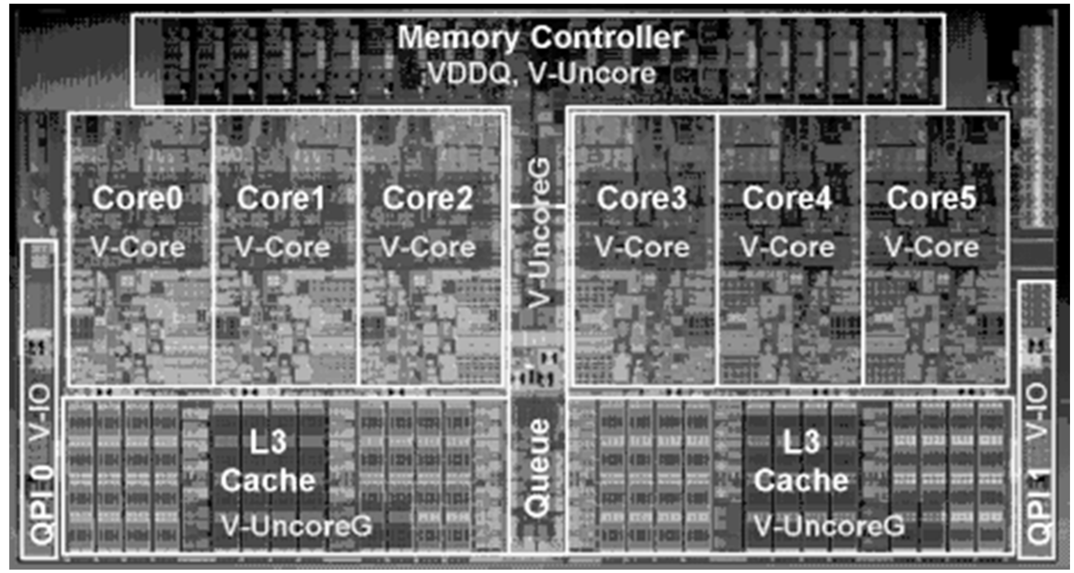
But: Power consumption growing too…

What to do with all these transistors?

Multi-core

Multi-core





http://www.theregister.co.uk/2010/02/03/intel_westmere_ep_preview/

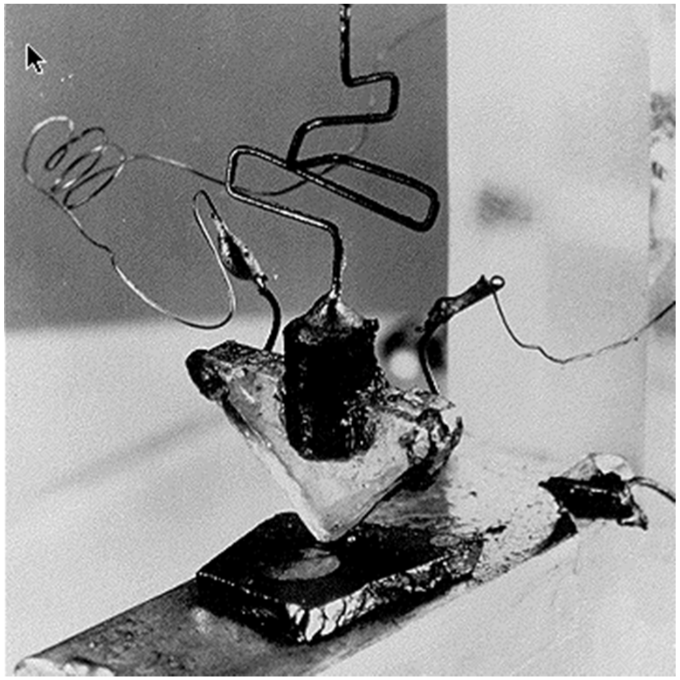
The first transistor

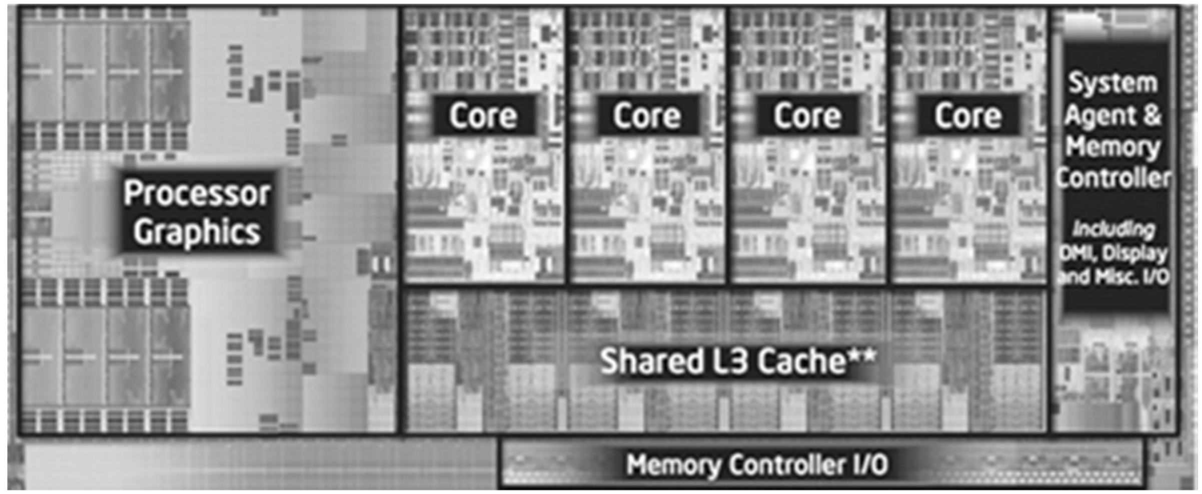
- on a workbench at

  AT&T Bell Labs in 1947
- Bardeen, Brattain, and Shockley

- An Intel Westmere
  - 1.17 billion transistors
  - 240 square millimeters
  - 32 nanometer: transistor gate width
  - Six processing cores
  - Release date: January 2010

Multi-core





http://forwardthinking.pcmag.com/none/296972-intel-releases-ivy-bridge-first-processor-with-tri-gate-transistor

The first transistor

- on a workbench at
  AT&T Bell Labs in 1947
- Bardeen, Brattain, and Shockley

- An Intel Ivy Bridge
  - 1.4 billion transistors
  - 160 square millimeters
  - 22 nanometer: transistor gate width
  - Up to eight processing cores
  - Release date: April 2012

What to do with all these transistors?

Cloud Computing

Cloud Computing

The promise of the Cloud

- *ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.*

NIST Cloud Definition

Cloud Computing

The promise of the Cloud

- *ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.*

NIST Cloud Definition

Cloud Computing

The promise of the Cloud

- *ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.* NIST Cloud Definition

Requires fundamentals in systems

- Computation
- Networking
- Storage

Cloud Computing

Large organizations ~~considering~~ using the cloud

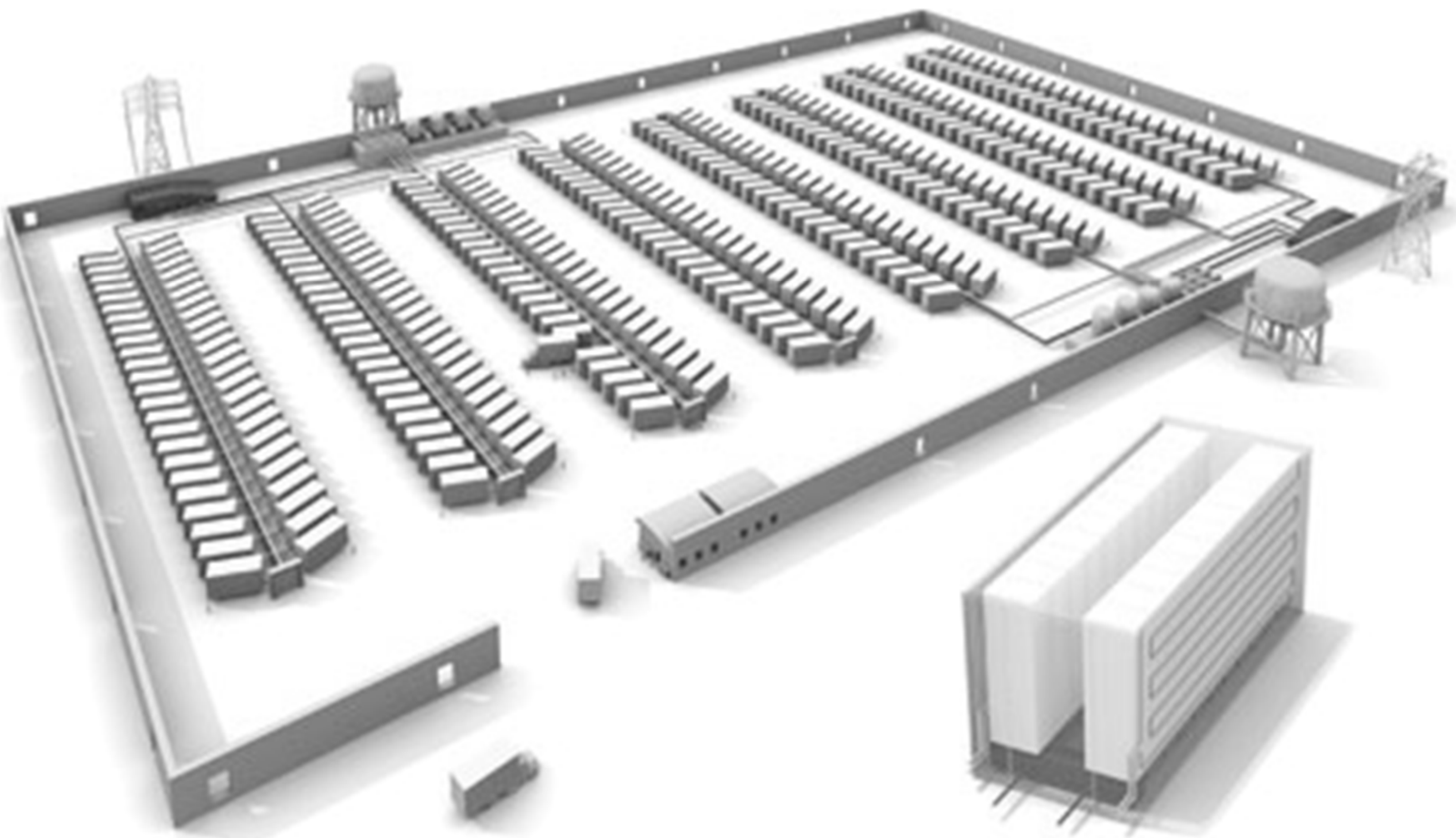
- New York Times
- Netflix
- Nintendo
- Cornell
- Library of Congress

The more data you have, the harder it is to move

- Switching providers entails paying for bandwidth *twice*
- Inhibits opportunistic migration

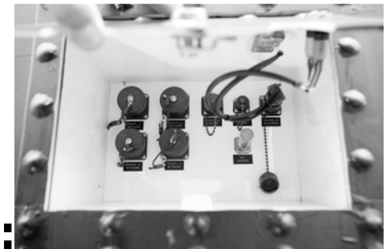
Cloud Computing

How hard is to program with a ExaByte of data?



Titan tech boom, randy katz, 2008

Cloud Computing



Datacenters are becoming a commodity

Order online and have it delivered

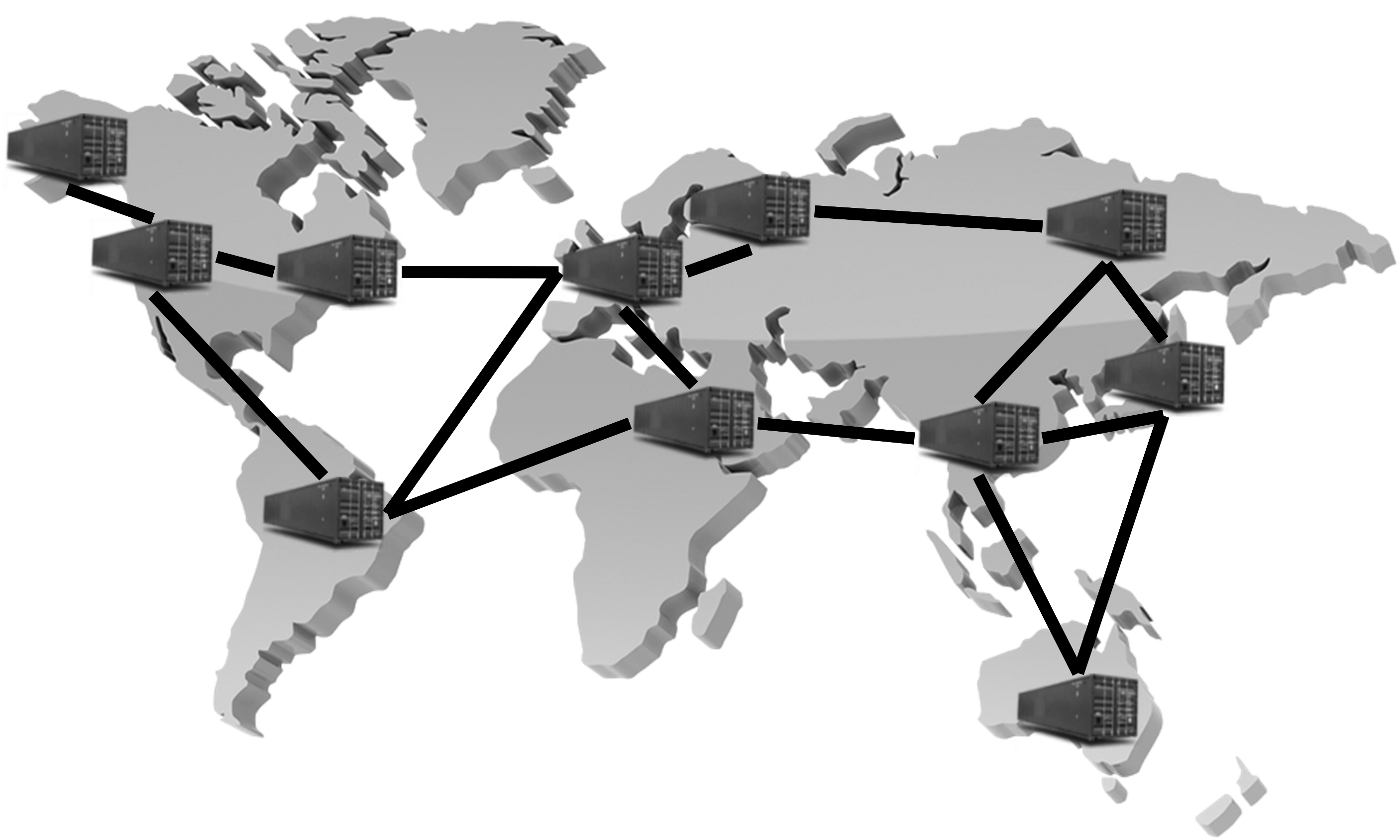
- Datacenter in a box: already set up with commodity hardware & software (Intel, Linux, petabyte of storage)
- Plug data, power & cooling and turn o
  - typically connected via optical fiber

such datacenters

Cloud Computing = Network of Datacenters

Cloud Computing

- How to optimize a global network of data centers?



Cloud Computing = Network of Datacenters

Cloud Computing

Vision

The promise of the Cloud

- A computer utility; a commodity
- Catalyst for technology economy
- Revolutionizing for health care, financial systems, scientific research, and society

However, cloud platforms today

- Entail significant risk: vendor lock-in vs control
- Entail inefficient processes: energy vs performance
- Entail poor communication: fiber optics vs COTS endpoint

Example: Energy and Performance

Why don't we save more energy in the cloud?

No one deletes data anymore!

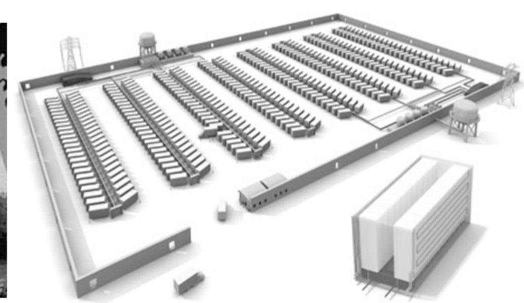
- Huge amounts of seldom-accessed data

Data deluge

- Google (YouTube, Picasa, Gmail, Docs), Facebook, Flickr
- 100 GB per second is faster than hard disk capacity growth!
- Max amount of data accessible at one time << Total data

New scalable approach needed to store this data

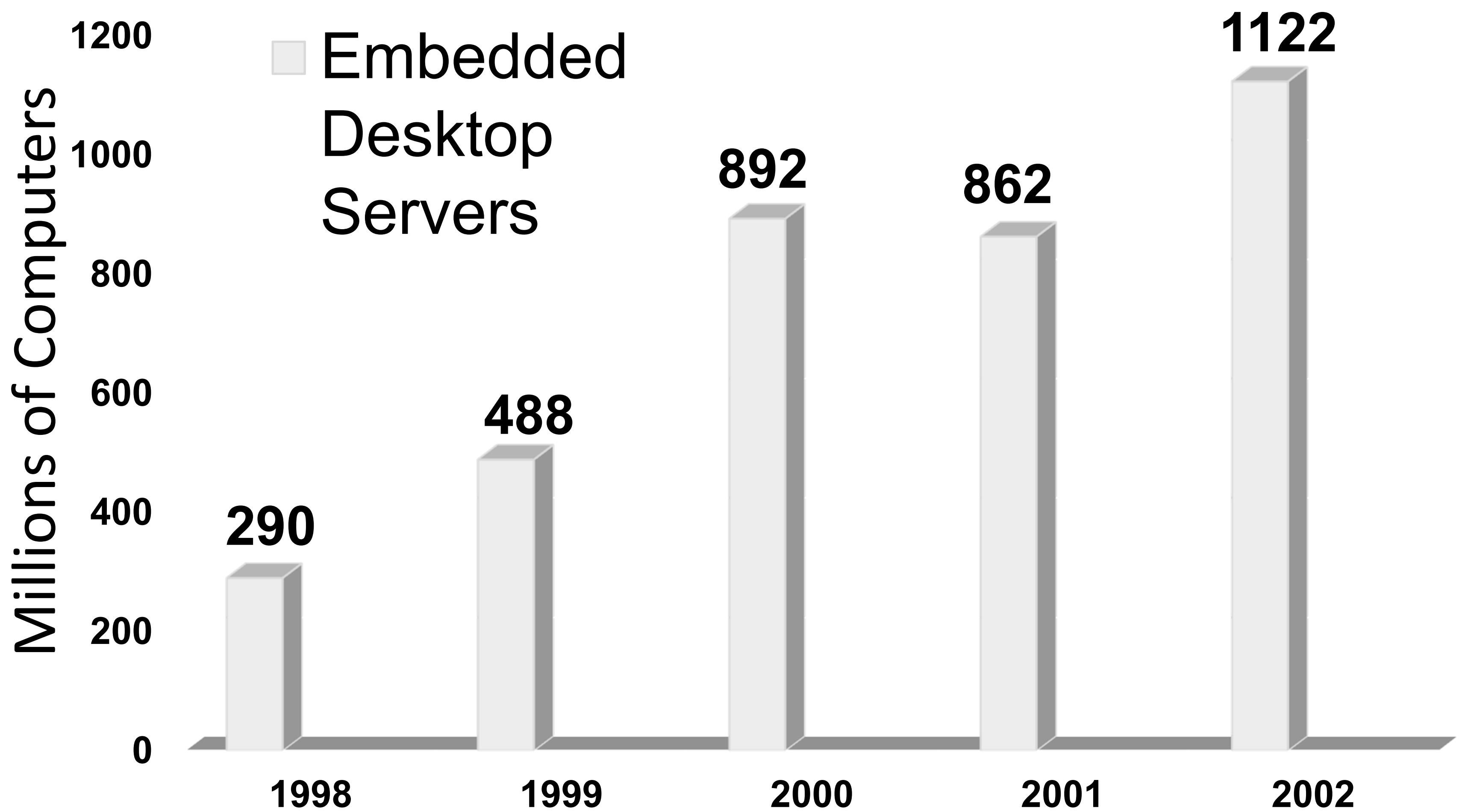
- Energy footprint proportional to number of HDDs is *not* sustainable



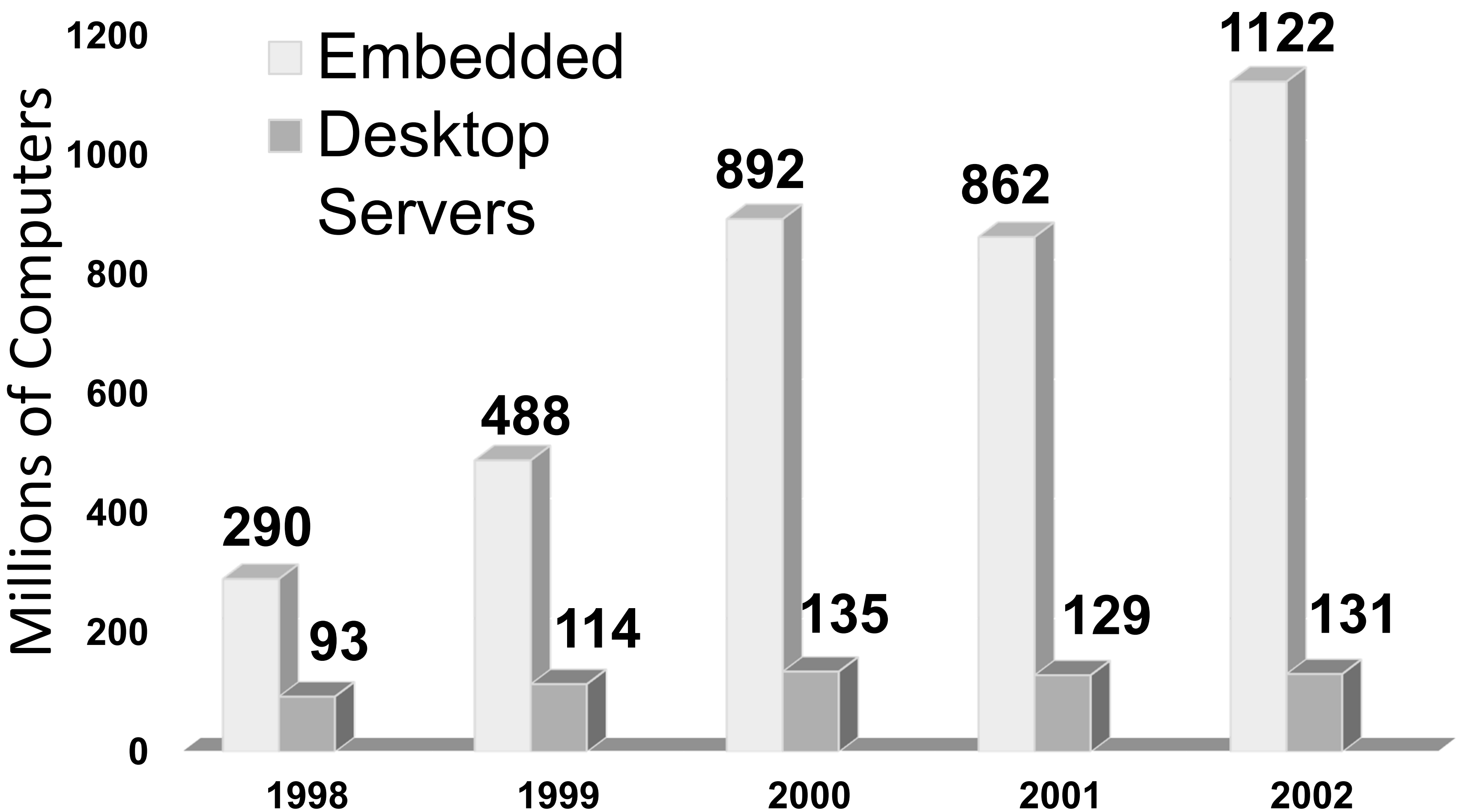
What to do with all these transistors?

Embedded Processors

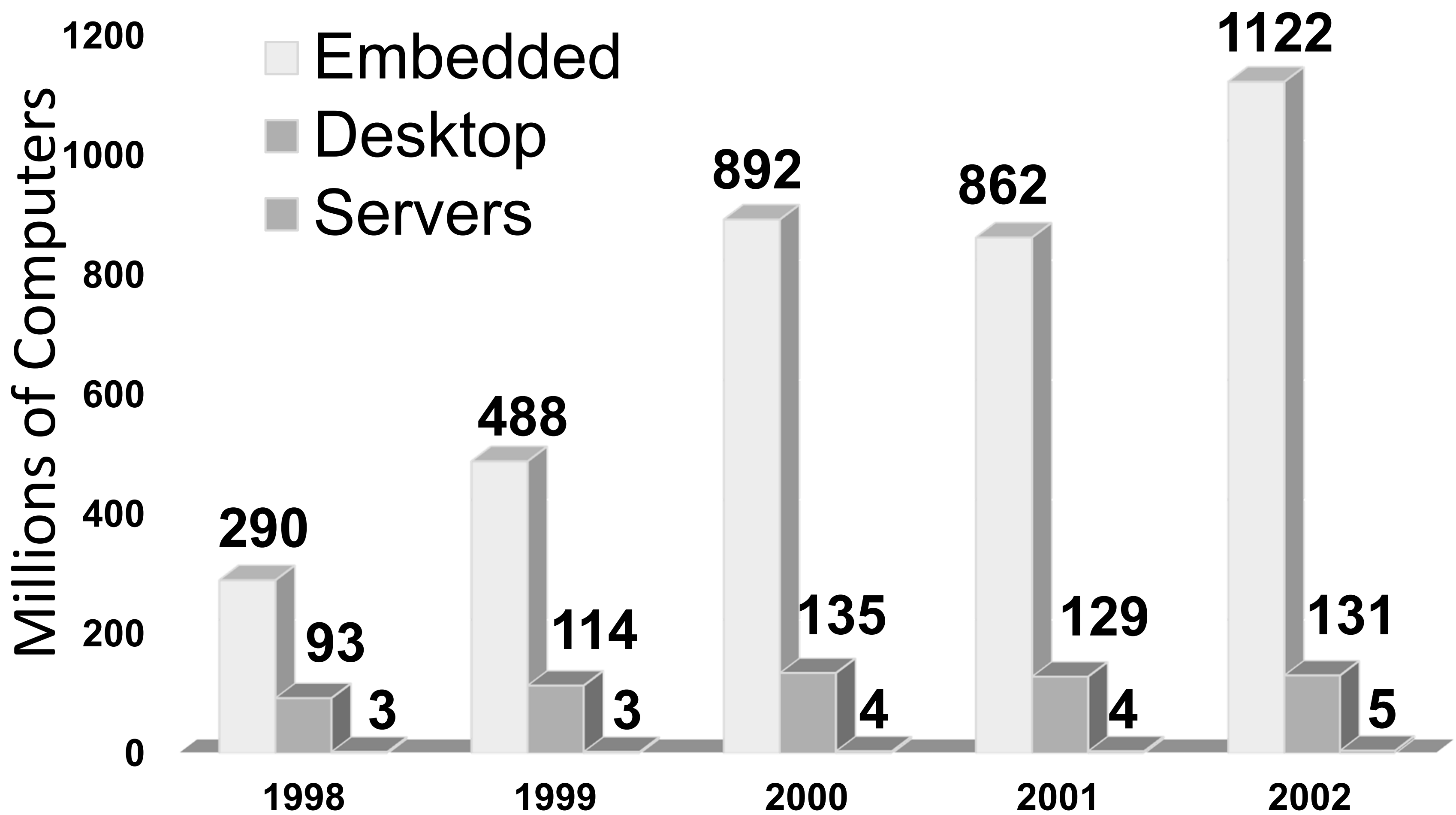
Where is the Market?



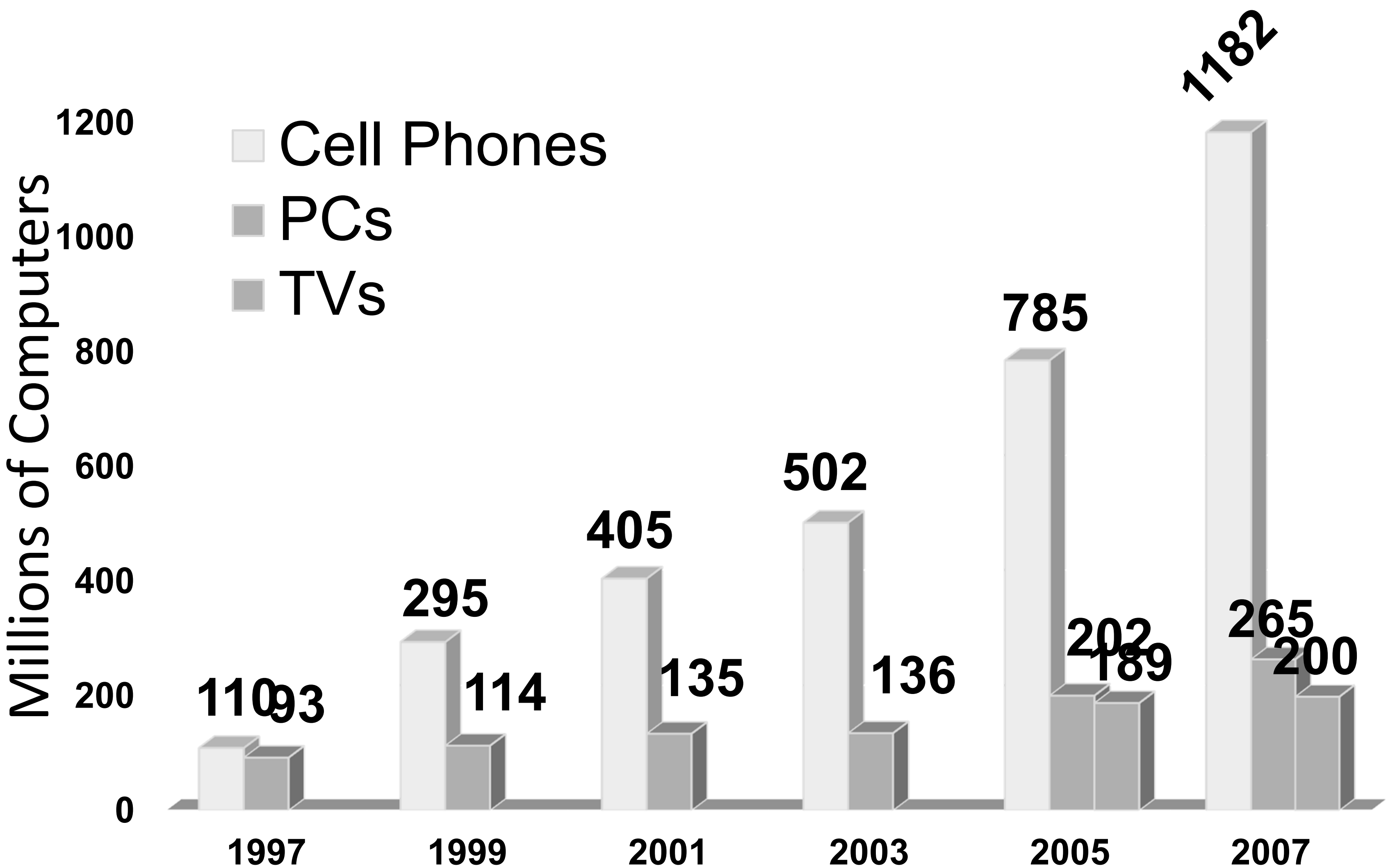
Where is the Market?



Where is the Market?



Where is the Market?




151st St to US-69 S
KIOWA
ARROWHEAD
LINDENWOOD
Arrival
Menu
Turn In


at&t
android and
me
SAMSUNG

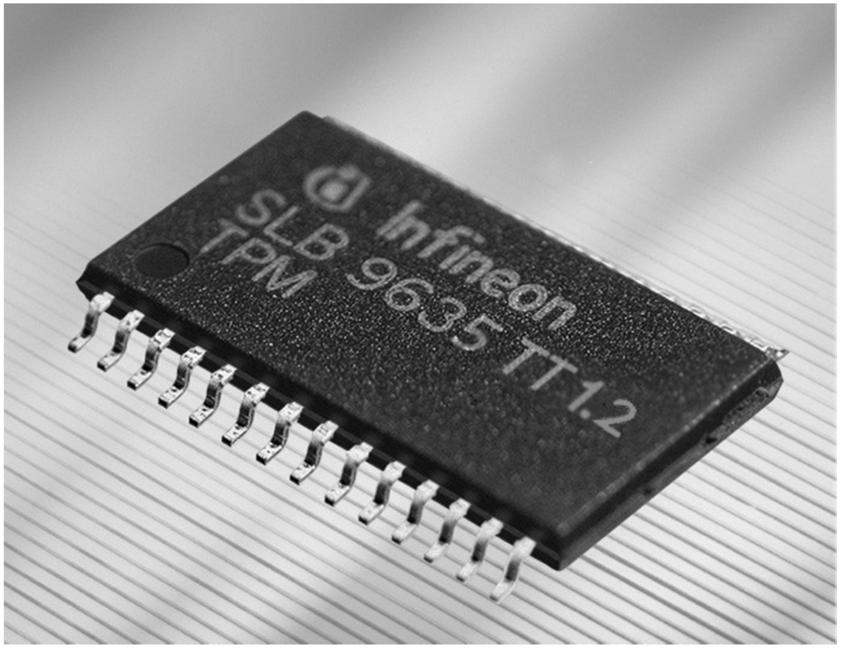
Where to?

Sr

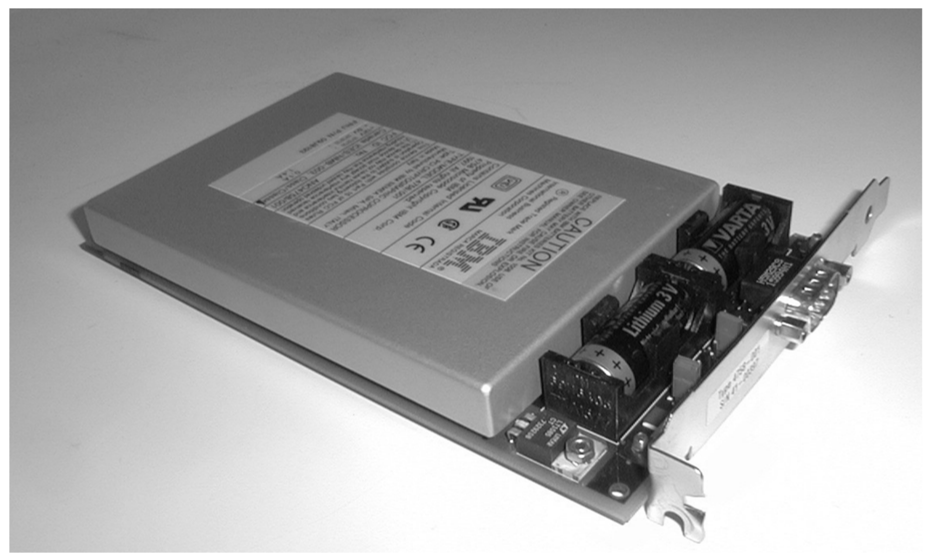
The soft contact lens encapsulates the electronics
The sensor can detect glucose levels in tears
The chip and antenna wirelessly receives power and sends data to the users mobile phone
Network Node
ZEEVO

Security?

Cryptography and security…



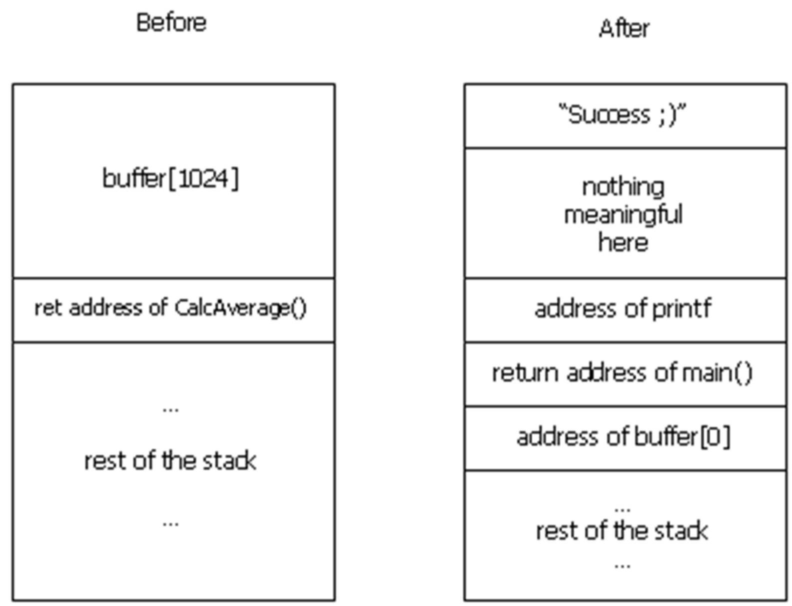
TPM 1.2



IBM 4758
Secure Cryptoprocessor

Security?

Stack Smashing…



What’s next?

Moore’s Law

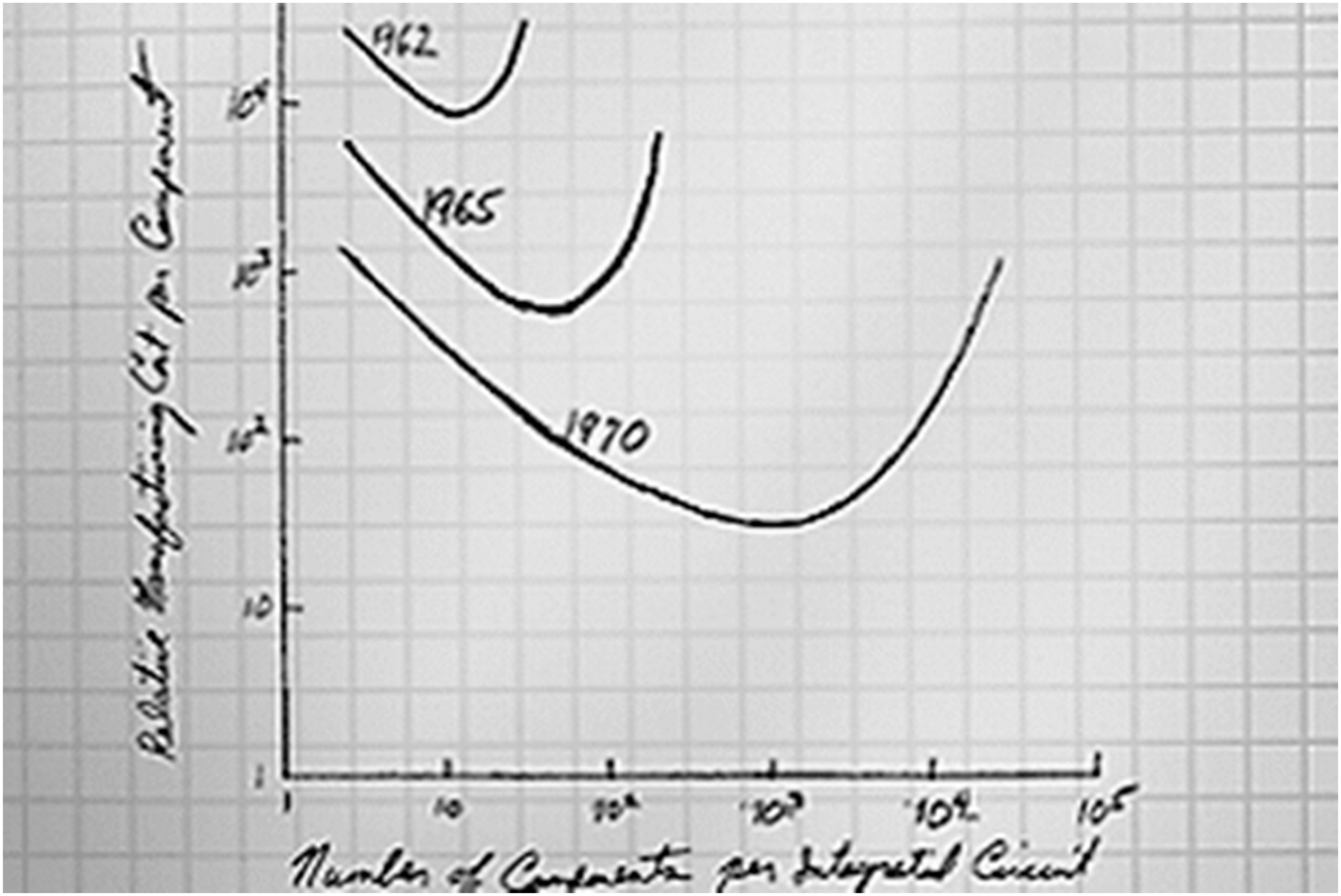
Moore’s Law introduced in 1965

- Number of transistors that can be integrated on a single die would double every 18 to 24 months (i.e., grow exponentially with time)

Amazingly visionary

- 2300 transistors, 1 MHz clock (Intel 4004) - 1971
- 16 Million transistors (Ultra Sparc III)
- 42 Million transistors, 2 GHz clock (Intel Xeon) – 2001
- 55 Million transistors, 3 GHz, 130nm technology, 250mm2 die (Intel Pentium 4) – 2004
- 290+ Million transistors, 3 GHz (Intel Core 2 Duo) – 2007
- 731 Million transistors, 2-3Ghz (Intel Nehalem) – 2009
- 1.4 Billion transistors, 2-3Ghz (Intel Ivy Bridge) – 2012

Moore’s Law



Parallelism

Dennard scaling: power

Must exploit parallelism for performance

MIMD: multiple instruction, multiple data

- Multicore

SIMD: single instruction, multiple data

- GPUs

My slide from 2008

Is Moore’s law dead?

Some thoughts

Bob Colwell

Chief Architect Pentium

DARPA

Introduction

Bill Dally, Nvidia CTO

Talk

The Chip Design Game at the End of Moore's Law

Hot Chips, Aug 2013

Singularity

Approximate Computing

Better interfaces

- Brain interfaces

Specialized chips

- Make it programmable

More

Supercomputers

Petaflops: GPUs/multicore/100s-1000s cores


FEATURED POST

WORLD'S FIRST ARM-BASED SUPERCOMPUTER TO LAUNCH IN BA

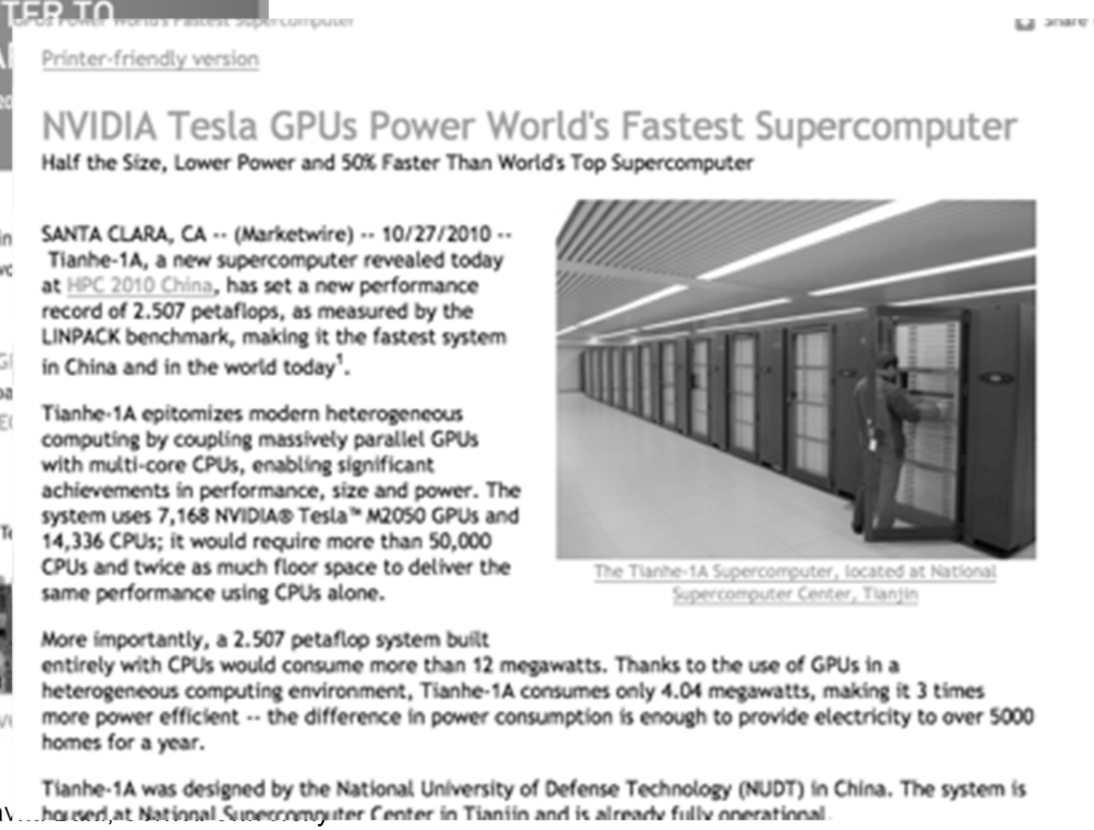
BY SUMIT GUPTA on Nov 14 20
Software, Supercomputing
1 COMMENT

The Barcelona Supercomputing Center (BSC) - Spain's national supercomputin
news today in the supercomputing world, by announcing plans to build the wo
ARM-based supercomputer.

BSC is planning to build the first ARM supercomputer, accelerated by CUDA G
scientific research. This prototype system will use NVIDIA's quad-core ARM-ba
on-a-chip, along with NVIDIA CUDA GPUs on a hardware board designed by SE
variety of scientific research projects.

In their search for more energy efficient architectures in supercomputers, BSC concluded that typical x86-based CPUs in today's supercomputers consume up to 40 percent of the system's total power. They've also realized that ARM CPUs are much more energy-efficient than x86 CPUs from Intel and AMD.

SECO Hardware Dev


Printer-friendly version

NVIDIA Tesla GPUs Power World's Fastest Supercomputer

Half the Size, Lower Power and 50% Faster Than World's Top Supercomputer

SANTA CLARA, CA -- (Marketwire) -- 10/27/2010 -- Tianhe-1A, a new supercomputer revealed today at HPC 2010 China, has set a new performance record of 2.507 petaflops, as measured by the LINPACK benchmark, making it the fastest system in China and in the world today[1].

Tianhe-1A epitomizes modern heterogeneous computing by coupling massively parallel GPUs with multi-core CPUs, enabling significant achievements in performance, size and power. The system uses 7,168 NVIDIA® Tesla™ M2050 GPUs and 14,336 CPUs; it would require more than 50,000 CPUs and twice as much floor space to deliver the same performance using CPUs alone.

The Tianhe-1A Supercomputer, located at National Supercomputer Center, Tianjin

More importantly, a 2.507 petaflop system built entirely with CPUs would consume more than 12 megawatts. Thanks to the use of GPUs in a heterogeneous computing environment, Tianhe-1A consumes only 4.04 megawatts, making it 3 times more power efficient -- the difference in power consumption is enough to provide electricity to over 5000 homes for a year.

Tianhe-1A was designed by the National University of Defense Technology (NUDT) in China. The system is housed at National Supercomputer Center in Tianjin and is already fully operational.

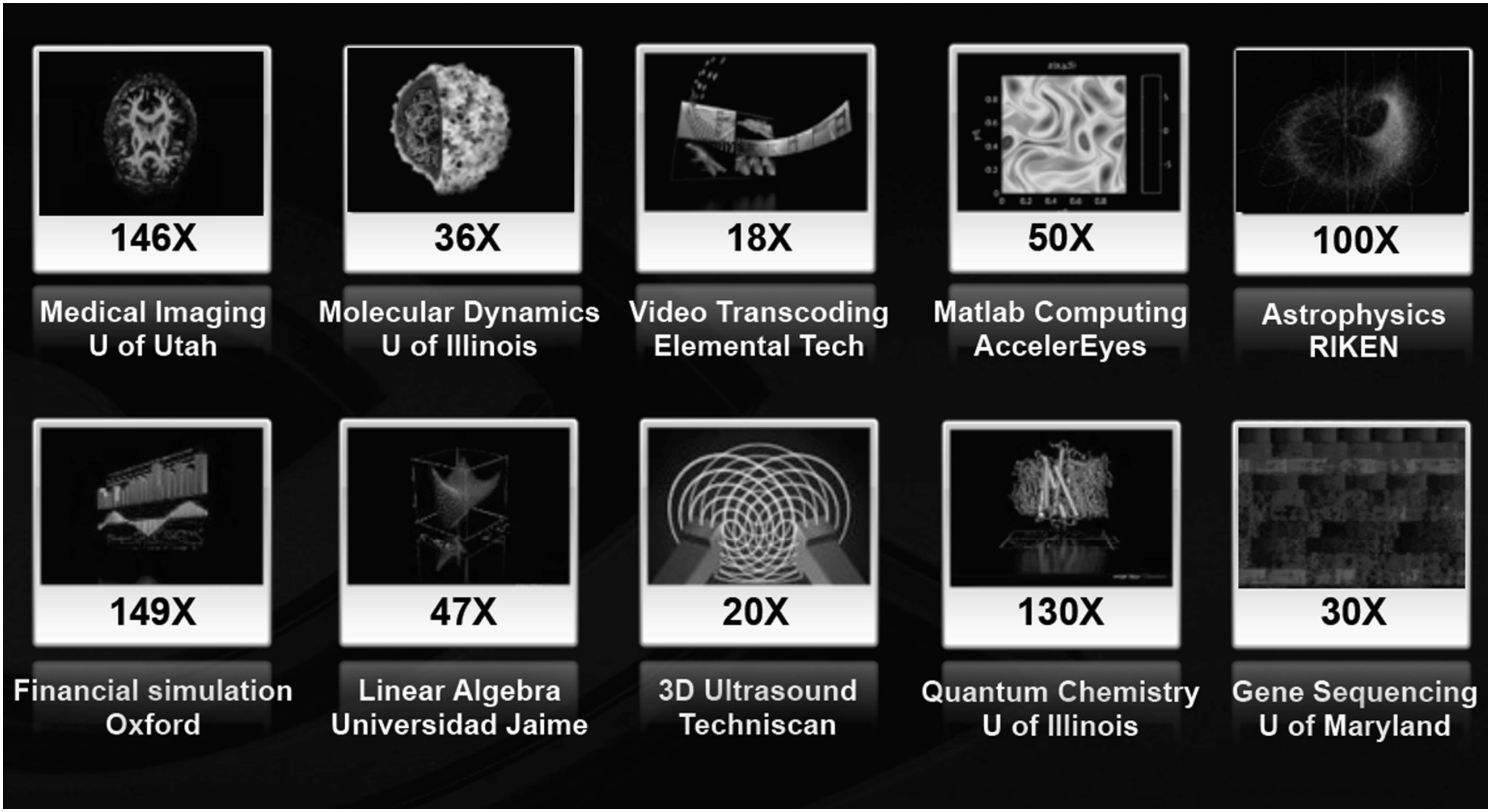
Petaflops

K computer

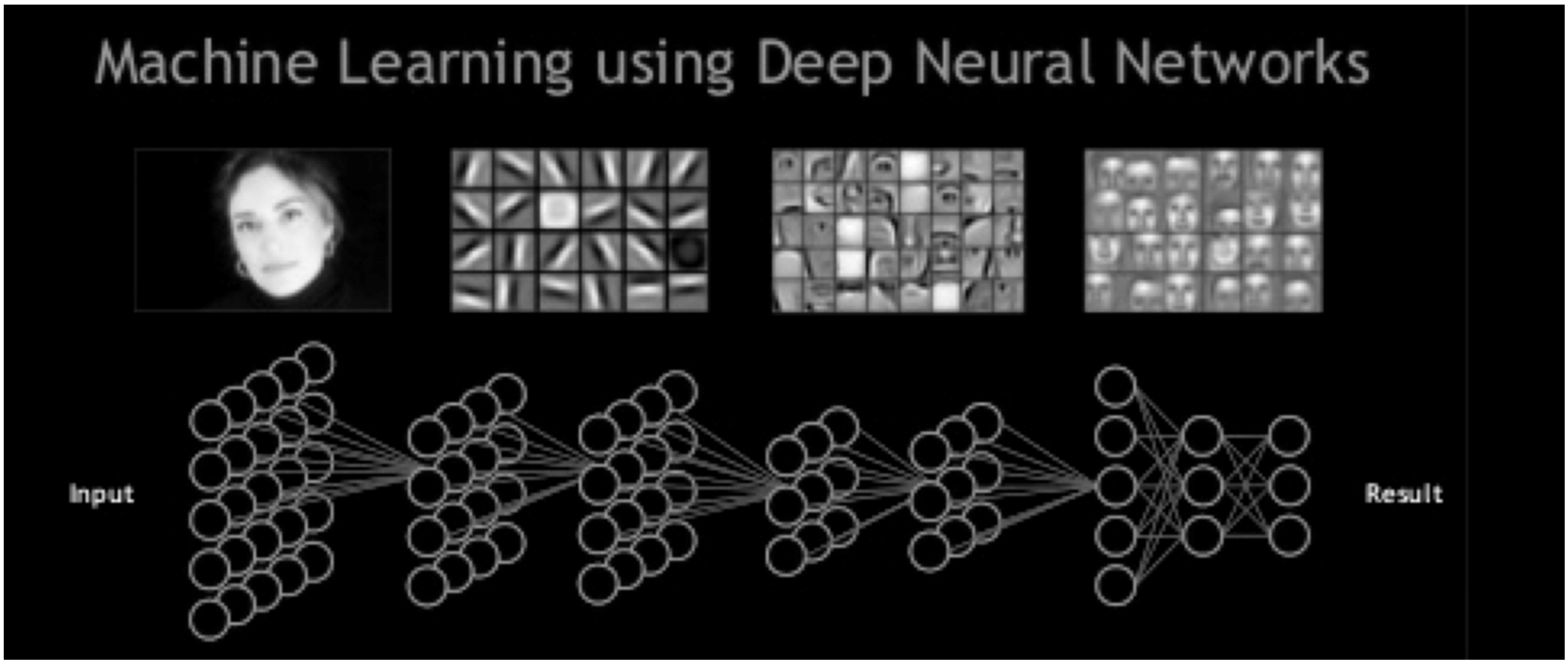
Japan and the rest of the world are faced with various problems that are hard to solve. The challenge for us to tackle is how to solve these issues promptly without further delay. To do this, we need to gather wisdom from around the world and accelerate our cutting-edge research in a variety of fields. Supercomputers will be crucial in achieving these goals. Fujitsu is striving to enable a prosperous future for the Earth and its peoples through the development of supercomputers.

One Fujitsu aim is to complete the development of the K computer by 2012 together with RIKEN, in accordance with the High

GPUs for Scientific Computing



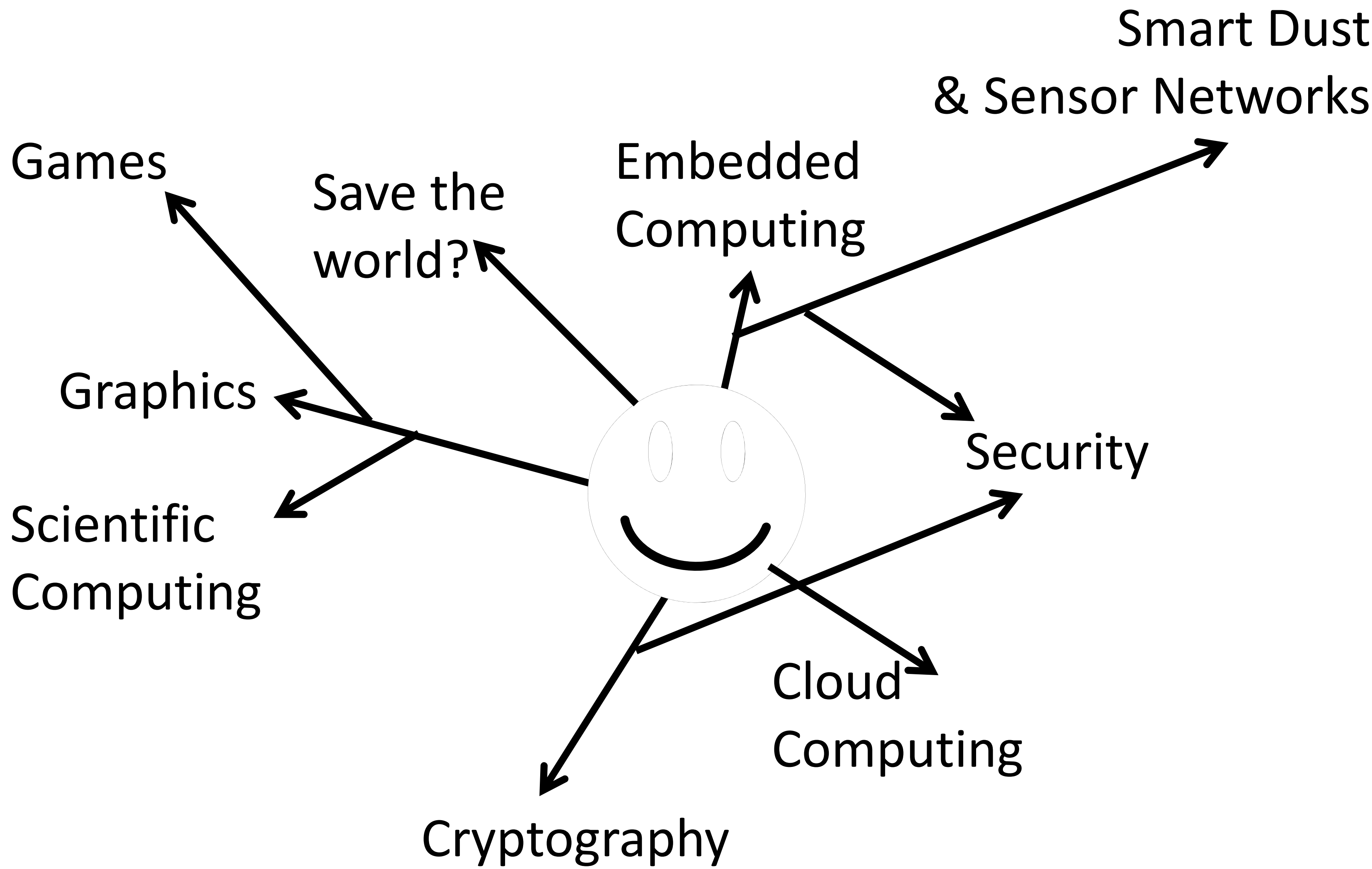
GPUs for Neural Nets

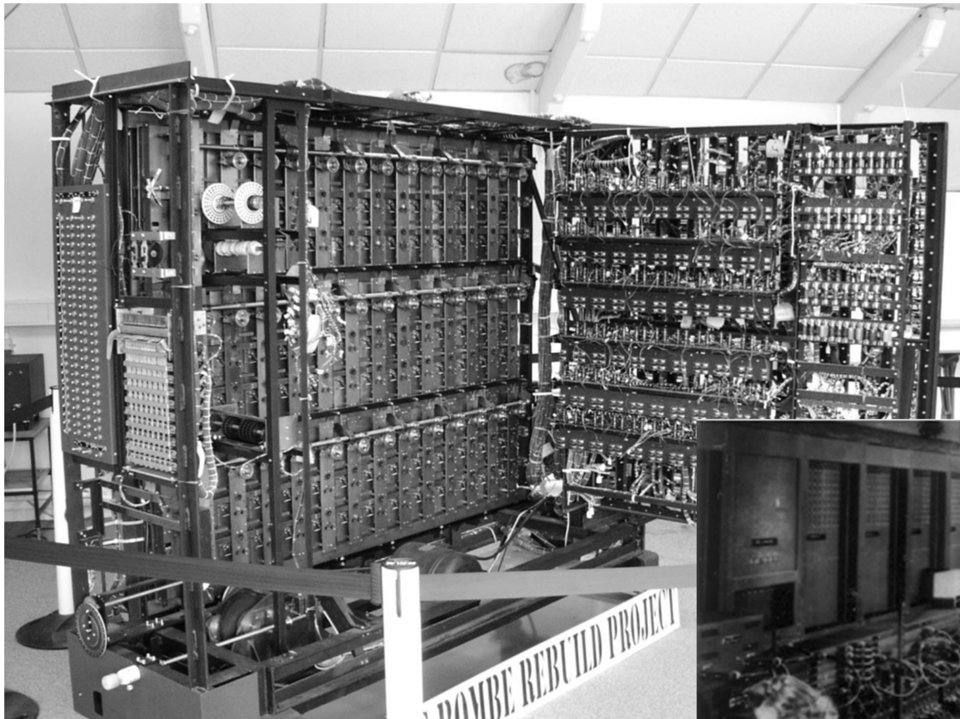


GPUs for Graphics, of course

What to do with all these transistors?

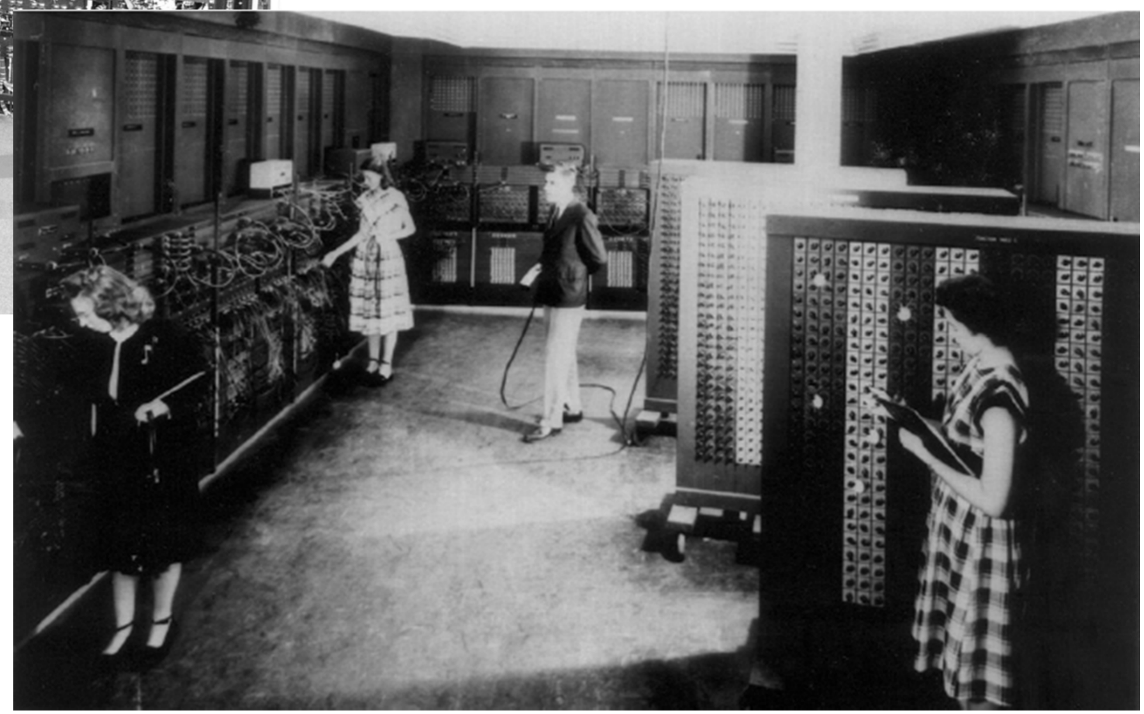
You could save the world one day?


Smart Dust
& Sensor Networks
Games
Save the
world?
Embedded
Computing
Graphics
Security
Scientific
Computing
Cloud
Computing
Cryptography



ENIAC - 1946

First general purpose electronic computer. Designed to calculate ballistic trajectories



Alan Turing's Bombe

Used to crack Germany's enigma machine

Survey Questions

Are you a better computer scientist and software engineering knowing “the low-level stuff”?

How much of computer architecture do software engineers actually have to deal with?

What are the most important aspects of computer architecture that a software engineer should keep in mind while programming?

Why?

These days, programs run on hardware...

… more than ever before

Google Chrome

→ Operating Systems

→ Multi-Core & Hyper-Threading

→ Datapath Pipelines, Caches, MMUs, I/O & DMA

→ Busses, Logic, & State machines

→ Gates

→ Transistors

→ Silicon

→ Electrons

Why?

Your job as a computer scientist will require knowledge the computer

Research/University

Cornell University
Faculty of Computing and Information Science

Industry

Government

Where to?

CS 3110: Better concurrent programming

CS 4410/4411: The Operating System!

CS 4420/ECE 4750: Computer Architecture

CS 4450: Networking

CS 4620: Graphics

~~CS 4821: Quantum Computing~~

MEng

5412—Cloud Computing, 5414—Distr Computing,

5430—Systems Secuirty,

5300—Arch of Larg scale Info Systems

And many more…

Thank you!

> If you want to make an apple pie from scratch, you must first create the universe.
>
> – Carl Sagan