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 13\textsuperscript{th}

or Wednesday, May 14\textsuperscript{th}

CMS submission due:

• Due 6:30pm Wednesday, May 14\textsuperscript{th}
Announcements

Prelim3 Results

- Mean 58 ± 16.2 (median 60), Max 93
- Pickup in Homework Passback Room (216 Gates)
Prelim3 Results

$2^{36} = 64\, \text{GB}$

$4\, \text{byte} = 32\, \text{-bit}$

Physical Memory

Virtual Memory

Physical Page Number

$28$-bit $= 32$-bit $- 4$ bit
**Announcements**

- $2^{36} = 64\text{GB}$
- 4 byte = 32-bit
- $2^{28}$
  - $\times 2^{13}$
  - $= 2^{41}$ or $2\text{TB}$

**Virtual Memory**

<table>
<thead>
<tr>
<th>V</th>
<th>R</th>
<th>W</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Physical Page Number**

- 0x10045
- 0xC20A3
- 0x4123B
- 0x10044

**Physical Memory**

- 8kB
- $2^{25}$
- 32MB

28-bit = 32-bit – 4 bit
Prelim3 Results

$2^{33} = 8\text{GB}$

$\frac{2^{33}}{2^{13}} = 2^{20}$

$2^{20} \times 4 = 2^{22}$

$4\text{MB}$

4 byte = 32-bit

<table>
<thead>
<tr>
<th>V</th>
<th>R</th>
<th>W</th>
<th>X</th>
<th>Physical Page Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td>0x10045</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>0xC20A3</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td>0x4123B</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td>0x10044</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Virtual Memory

Physical Memory

8kB
Multi-level Page Table

12 bits | 11 bits | 13 bits | vaddr

Page Directory

Page Table

Page

PTBR

Page table = 16kB + 4MB

2^{12} \times 4B
= 2^{14}B
= 16kB
= 2 pages

PTEntry

4B

Word

2^{12} \times 4B
= 2^{14}B
= 16kB
= 2 pages

Multilevel Page Table

Announcements
Multi-level Page Table

- 12 bits
- 11 bits
- 13 bits
- vaddr

- Page Directory
  - PDEntry
  - PTBR

- Page Table
  - PTEntry

- Word

- Page

- $2^{12} \times 4B = 2^{14}B = 16kB = 2\text{pages}$

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**

<table>
<thead>
<tr>
<th>Team</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>balabot</td>
<td>16,252,931</td>
</tr>
<tr>
<td>sam2</td>
<td>12,917,624</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Status</th>
<th>Core 0</th>
<th>Core 1</th>
<th>Core 2</th>
<th>Core 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>0%</td>
<td>69%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Crashes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**wins over**
(game over -- goal reached)

<table>
<thead>
<tr>
<th>Status</th>
<th>Core 0</th>
<th>Core 1</th>
<th>Core 2</th>
<th>Core 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>0%</td>
<td>25%</td>
<td>45%</td>
<td>25%</td>
</tr>
<tr>
<td>Crashes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
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
Curve shows ‘Moore’s Law’: transistor count doubling every two years.
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

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

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.  

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 on
  – 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 $\ll$ 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?

- **Embedded Desktop Servers**
  - 1998: 290
  - 1999: 488
  - 2000: 892
  - 2001: 862
  - 2002: 1122

Millions of Computers
Where is the Market?

Millions of Computers

- Embedded
- Desktop

Servers

Year

1998
1999
2000
2001
2002

290
488
892
862
1122

93
114
135
129
131

1000
900
800
700
600
500
400
300
200
100
0
Where is the Market?

- 1998: Embedded 290, Desktop 93, Servers 3
- 1999: Embedded 488, Desktop 114, Servers 3
- 2000: Embedded 892, Desktop 135, Servers 4
- 2001: Embedded 862, Desktop 129, Servers 4
- 2002: Embedded 1122, Desktop 131, Servers 5

Millions of Computers
Where is the Market?

- Cell Phones
- PCs
- TVs

<table>
<thead>
<tr>
<th>Year</th>
<th>Cell Phones</th>
<th>PCs</th>
<th>TVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>11093</td>
<td>114</td>
<td>135</td>
</tr>
<tr>
<td>1999</td>
<td>295</td>
<td>136</td>
<td>114</td>
</tr>
<tr>
<td>2001</td>
<td>405</td>
<td>136</td>
<td>135</td>
</tr>
<tr>
<td>2003</td>
<td>502</td>
<td>189</td>
<td>202</td>
</tr>
<tr>
<td>2005</td>
<td>785</td>
<td>265</td>
<td>200</td>
</tr>
<tr>
<td>2007</td>
<td>1182</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Where to?

Smart Dust…

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 user's mobile phone.
Security?

Cryptography and security...

TPM 1.2

IBM 4758
Secure Cryptoprocessor
Security?

Stack Smashing...

Before

- buffer[1024]
- ret address of CalcAverage()
- rest of the stack

After

- "Success ;)
- nothing meaningful here
- address of printf
- return address of main()
- address of buffer[0]
- rest of the stack
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

![Graph showing Relative Manufacturing Cost per Component over time with curves for 1962, 1965, and 1970. The x-axis represents Number of Components per Integrated Circuit, and the y-axis represents Relative Manufacturing Cost.]
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

Do you believe?

I WANT TO BELIEVE

© Kavita Bala 2008 Computer Science, Cornell University
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
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

- **146X**: Medical Imaging U of Utah
- **36X**: Molecular Dynamics U of Illinois
- **18X**: Video Transcoding Elemental Tech
- **50X**: Matlab Computing AccelerEyes
- **100X**: Astrophysics RIKEN
- **149X**: Financial simulation Oxford
- **47X**: Linear Algebra Universidad Jaime
- **20X**: 3D Ultrasound Techniscan
- **130X**: Quantum Chemistry U of Illinois
- **30X**: Gene Sequencing U of Maryland
GPUs for Neural Nets

Machine Learning using Deep Neural Networks
GPUs for Graphics, of course
What to do with all these transistors?

You could save the world one day?
Games
Graphics
Scientific Computing
Save the world?
Embedded Computing
Security
Cryptography
Cloud Computing
Smart Dust & Sensor Networks
Alan Turing’s Bombe
Used to crack Germany’s enigma machine

ENIAC - 1946
First general purpose electronic computer. Designed to calculate ballistic trajectories
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

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

And many more...
Thank you!

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

– Carl Sagan