CENG 3420
Computer Organization & Design

Lecture 02: Basis

Bei Yu
CSE Department, CUHK
byu@cse.cuhk.edu.hk

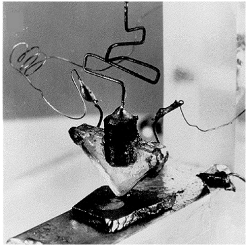
(Textbook: Chapters 1 & 2.4)

Spring 2022

Computer History

The Evolution of Computer Hardware

When was the first transistor invented?



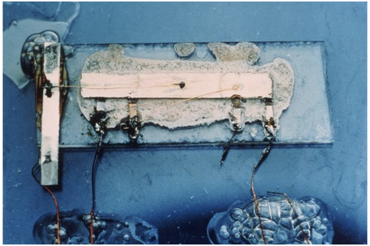
(a)

(b)

(a) 1947, bi-polar transistor, by John Bardeen et al. at Bell Laboratories; (b) UNIVAC I (Universal Automatic Computer): the first commercial computer in USA.

The Evolution of Computer Hardware

When was the first IC (integrated circuit) invented?



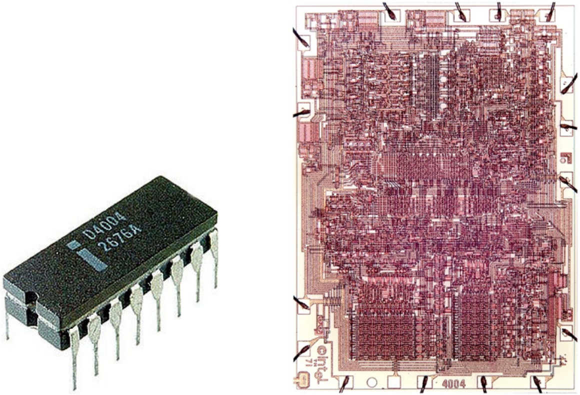
(a)

(b)

(a) 1958, by Jack Kilby@Texas Instruments, by hand. Several transistors, resistors and capacitors on a single substrate. (b) IBM System/360, 2MHz, 128KB – 256KB.

The Evolution of Computer Hardware

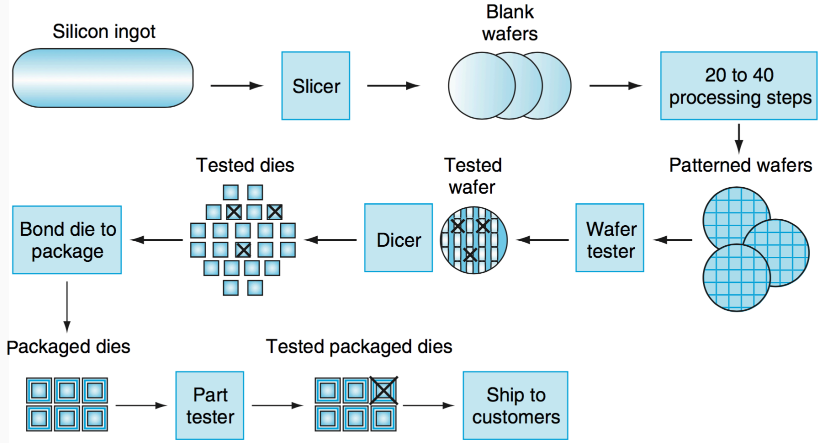
When was the first Microprocessor?



(a) (b)

1971, Intel 4004.

The IC Manufacturing Process

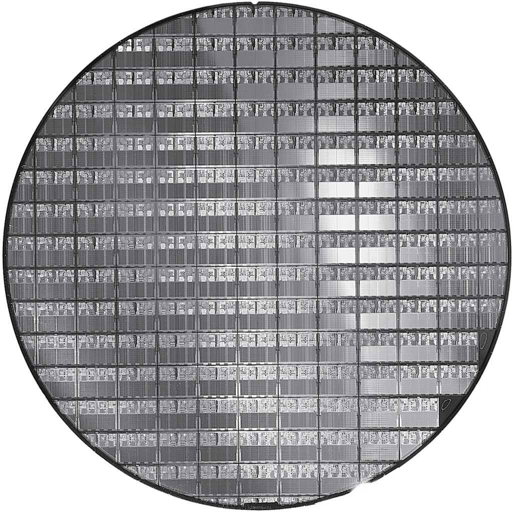


Yield

Proportion of working dies per wafer

Check this: https://youtu.be/d9SWNLZvA8g?list=FLELqiXCJQW-jcijW8ZAbA8w

AMD Opteron X2 Wafer



300*mm* wafer, 117 chips, 90*nm* technology.

Integrated Circuit Cost

$$\text{Cost per die} = \frac{\text{Cost per wafer}}{\text{Dies per wafer} \cdot \text{Yield}}$$

$$\text{Dies per wafer} = \text{Wafer area} \ / \ \text{Die area}$$

$$\text{Yield} = \frac{1}{[1 + (\text{Defects per area} \cdot \text{Die area} \ / \ 2)]^2}$$

Nonlinear relation to area and defect rate

- Wafer cost and area are fixed
- Defect rate determined by manufacturing process
- Die area determined by architecture and circuit design

Impacts of Advancing Technology

Processor

- Logic capacity: increases about 30% per year
- Performance: 2× every 1.5 years

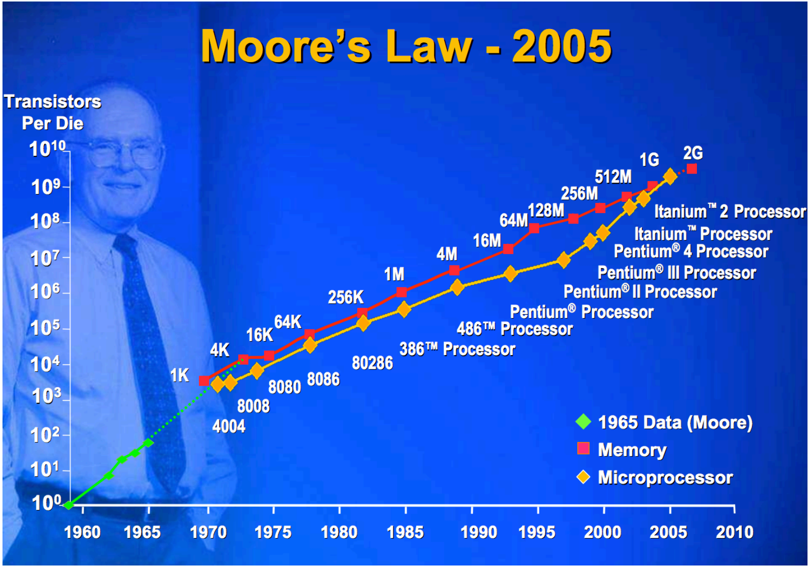
Memory

- DRAM capacity: 4× every 3 years, about 60% per year
- Memory speed: 1.5× every 10 years
- Cost per bit: decreases about 25% per year

Disk

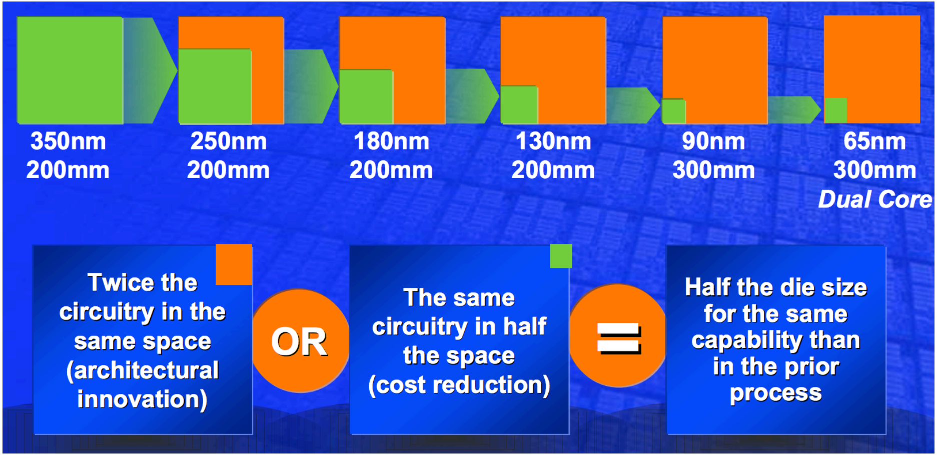
- Capacity: increases about 60% per year

Moore's Law for CPUs and DRAMs



From: "*Facing the Hot Chips Challenge Again*", Bill Holt, Intel, presented at Hot Chips 17, 2005.

Main driver: device scaling ...


350nm
200mm
250nm
200mm
180nm
200mm
130nm
200mm
90nm
300mm
65nm
300mm
Dual Core
Twice the circuitry in the same space (architectural innovation)
OR
The same circuitry in half the space (cost reduction)
=
Half the die size for the same capability than in the prior process

From: “*Facing the Hot Chips Challenge Again*”, Bill Holt, Intel, presented at Hot Chips 17, 2005.

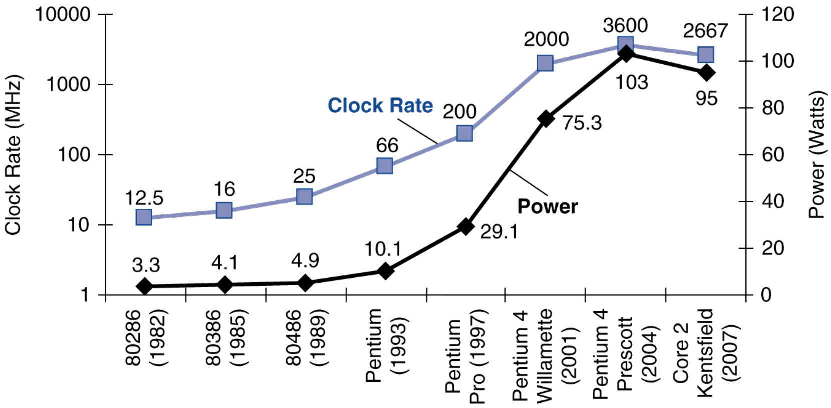
Technology Scaling Road Map (ITRS)

Year	2004	2006	2008	2010	2012
Feature size (nm)	90	65	45	32	22
Intg. Capacity (BT)	2	4	6	16	32

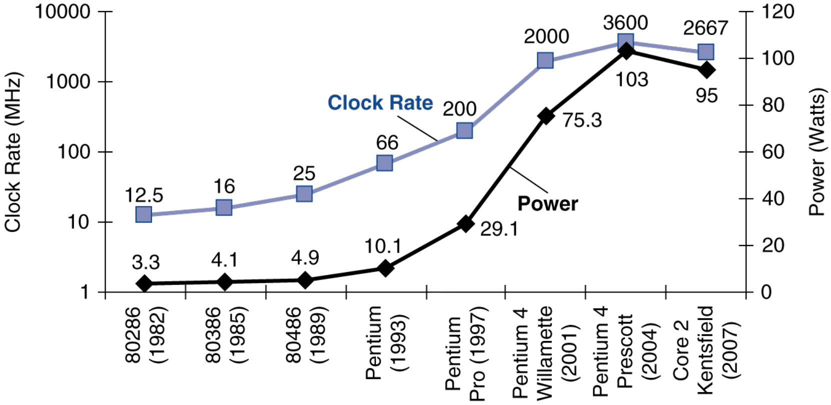
Fun facts about 45nm transistors

- 30 million can fit on the head of a pin
- You could fit more than 2,000 across the width of a human hair
- If car prices had fallen at the same rate as the price of a single transistor since 1968, a new car today would cost about 1 cent

Highest Clock Rate of Intel Processors



Highest Clock Rate of Intel Processors



What if the exponential increase had kept up? Why not?

- Due to process improvements
- Deeper pipeline
- Circuit design techniques

Power Issue

$$\text{Power} = \text{Capacitive load} \cdot \text{Voltage}^2 \cdot \text{Frequency}^1$$

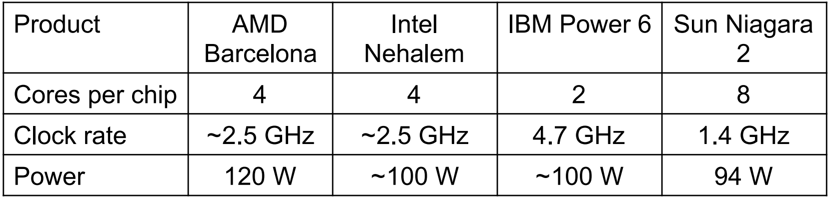
Example

For a simple processor, if capacitive load is reduced by 15%, voltage is reduced by 15%, maintain the same frequency, how much power consumption can be reduced?

[1] here we only consider dynamic power, but not static power

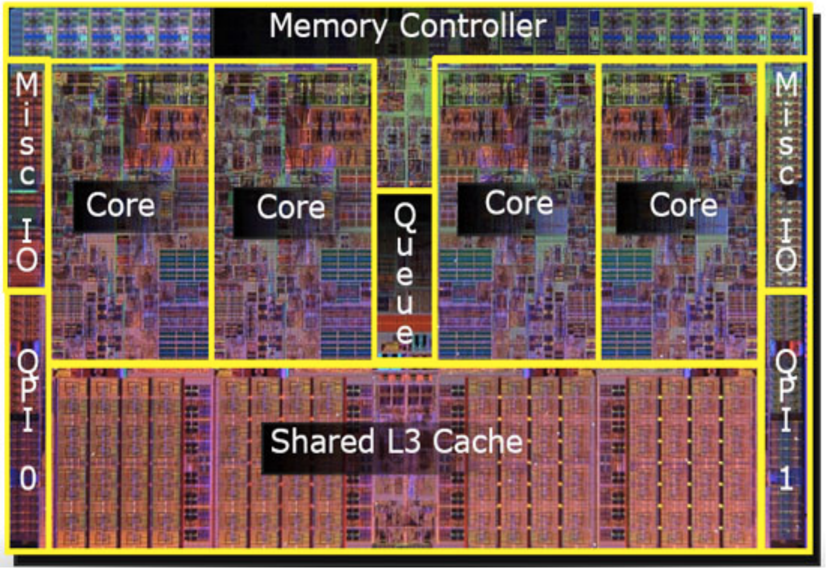
A Sea Change Is at Hand

- The power challenge has forced a change in the design of microprocessors
- Since 2002 the rate of improvement in the response time of programs on desktop computers has slowed from a factor of 1.5 per year to less than a factor of 1.2 per year
- As of 2006 all desktop and server companies are shipping microprocessors with multiple processors – cores – per chip
- Plan of record is to add two cores per chip per generation (about every two years)



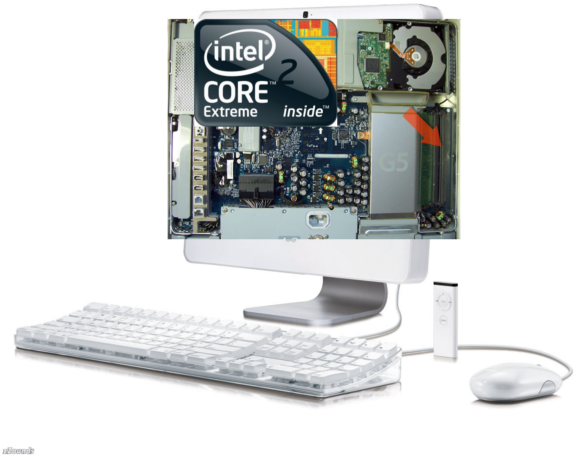
Product	AMD Barcelona	Intel Nehalem	IBM Power 6	Sun Niagara 2
Cores per chip	4	4	2	8
Clock rate	~2.5 GHz	~2.5 GHz	4.7 GHz	1.4 GHz
Power	120 W	~100 W	~100 W	94 W

Intel Core i7 Processor



45nm technology, 18.9mm x 13.6mm, 0.73billion transistors, 2008

A Computer



Desktop computers

Designed to deliver good performance to a single user at low cost usually executing 3rd party software, usually incorporating a graphics display, a keyboard, and a mouse

Other Classes of Computers

Servers

Used to run larger programs for multiple, simultaneous users typically accessed only via a network and that places a greater emphasis on dependability and (often) security

Supercomputers

A high performance, high cost class of servers with hundreds to thousands of processors, terabytes of memory and petabytes of storage that are used for high-end scientific and engineering applications.

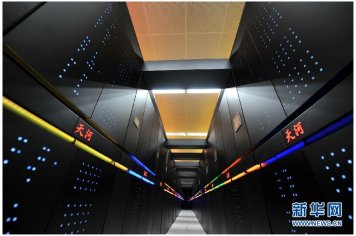
Embedded computers (processors)

A computer inside another device used for running one predetermined application

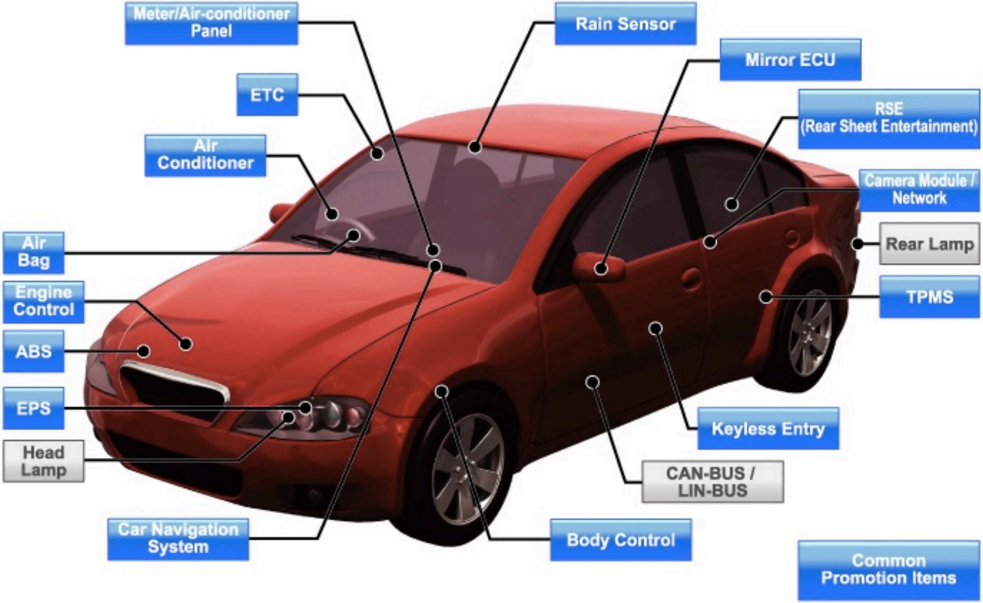
Supercomputers

Tianhe-2 (MilkyWay-2)

- Over 3 million cores
- Power: 17.6 MW (24 MW with cooling)
- Speed: 33.86 PFLOPS (peta $= 10^{15}$)



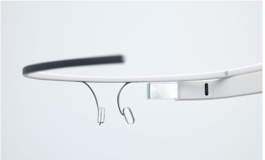
Embedded Computers in You Car



PostPC Era

Personal Mobile Device (PMD)

Battery-operated device with wireless connectivity

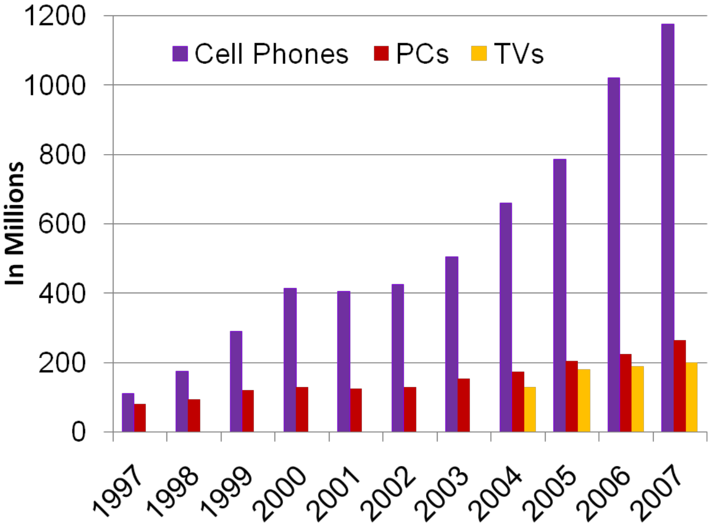


Warehouse Scale Computer (WSC)

Datacenter containing hundreds of thousands of servers providing software as a service (**SaaS**)

Growth in Cell Phone Sales (Embedded)

- embedded growth >> desktop growth
- Where else are embedded processors found?



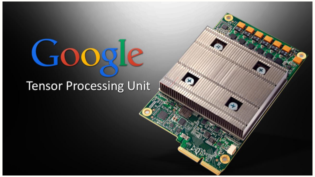
When Machine Learning Meets Hardware

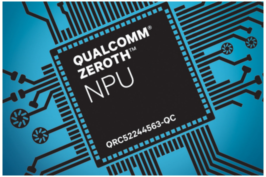
Convolution layer is one of the most expensive layers

- Computation pattern
- Emerging challenges

More and more end-point devices with limited memory

- Cameras
- Smartphone
- Autonomous driving





Convolutional Neural Network (CNN)

Autonomous drive

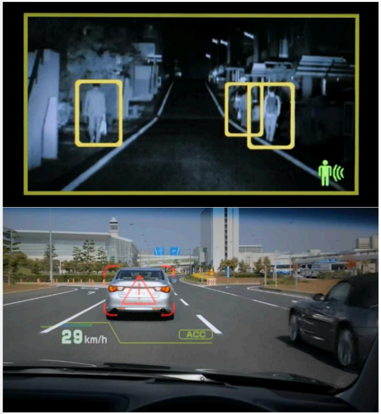
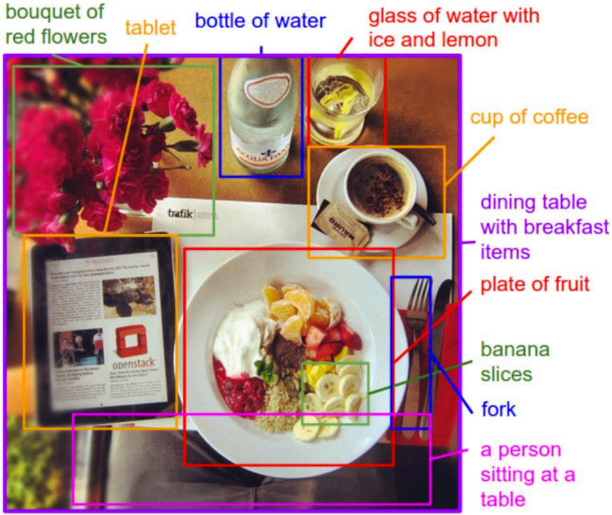
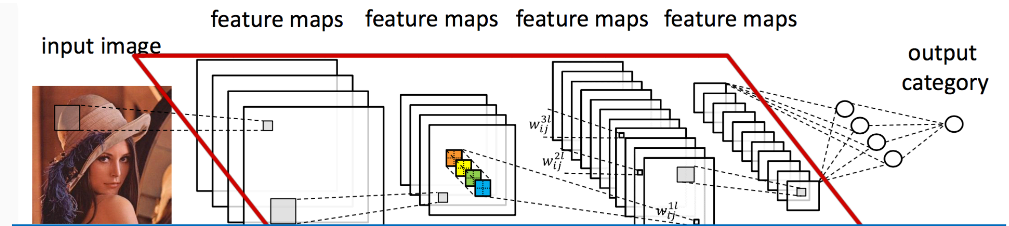


Image recognition

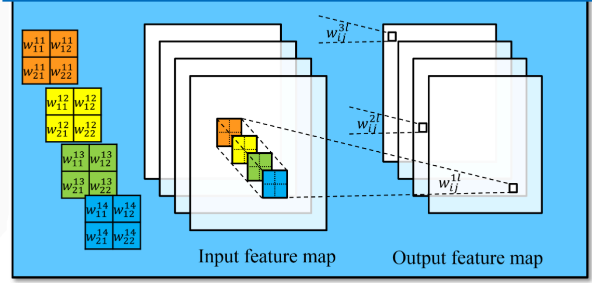


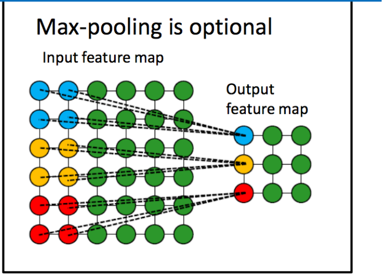
Bottleneck of CNN

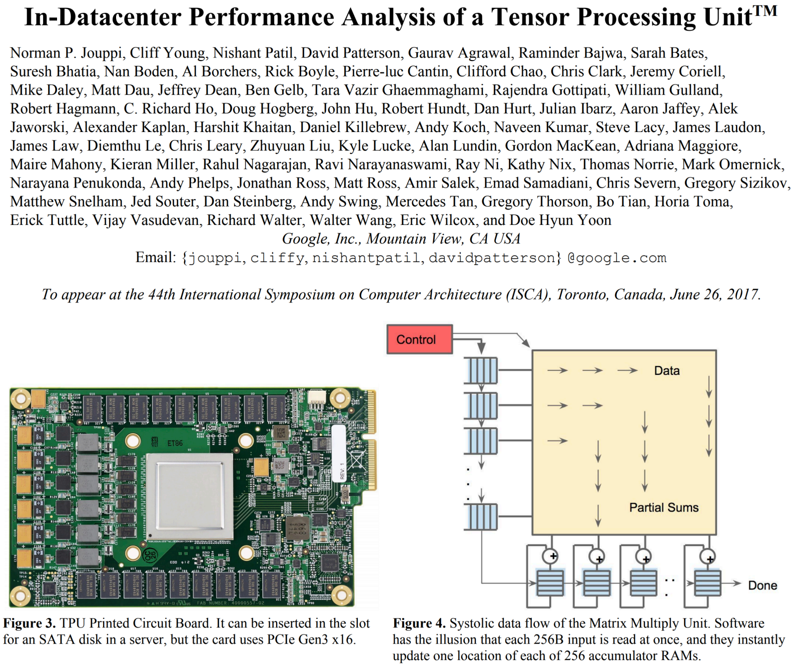


Convolutional layers account for over 90% computation

[1] A. Krizhevsky, etc. Imagenet classification with deep convolutional neural networks. NIPS 2012.
[2] J. Cong and B. Xiao. Minimizing computation in convolutional neural networks. ICANN 2014


w_{11}^{11} w_{12}^{11} w_{21}^{11} w_{22}^{11}
w_{11}^{12} w_{12}^{12} w_{21}^{12} w_{22}^{12}
w_{11}^{13} w_{12}^{13} w_{21}^{13} w_{22}^{13}
w_{11}^{14} w_{12}^{14} w_{21}^{14} w_{22}^{14}
w_{ij}^{3l}
w_{ij}^{2l}
w_{ij}^{1l}
Input feature map
Output feature map





In-Datacenter Performance Analysis of a Tensor Processing Unit™

Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon

Google, Inc., Mountain View, CA USA

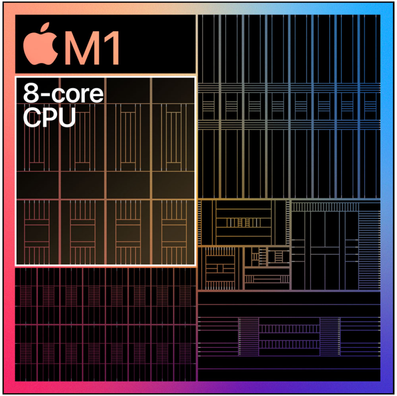
Email: {jouppi, cliffy, nishantpatil, davidpatterson} @google.com

To appear at the 44th International Symposium on Computer Architecture (ISCA), Toronto, Canada, June 26, 2017.

Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

Apple M1 Processor



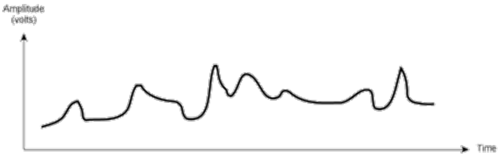
- 8-core CPU
- 8-core GPU
- 16-core Neural Engine

Numeral System

Analog vs. Digital

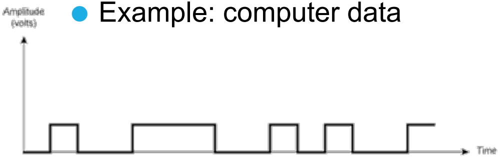
Analog Signal

- ❑ Vary in a smooth way over time
- ❑ Analog data are continuous valued
  - Example: audio, video



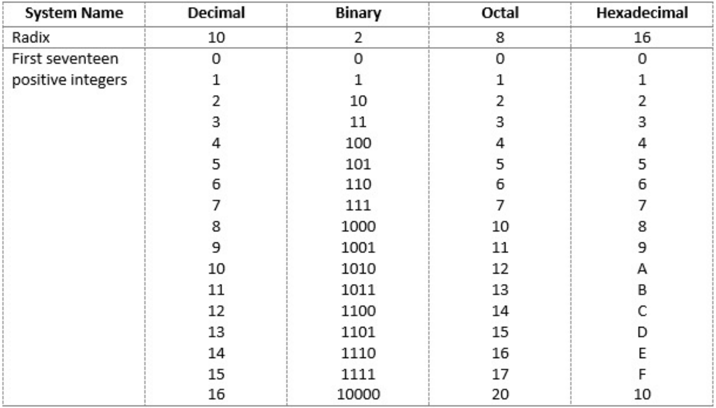
Digital Signal

- ❑ Maintains a constant level then changes to another constant level (generally operate in one of the two states)
- ❑ Digital data are discrete valued
  - Example: computer data



Number Systems

- An ordered set of symbols, called digits, with relations defined for addition, subtraction, multiplication, and division
- Radix or base of the number system is the total number of digits allowed in the number system
- Commonly used numeral systems



System Name	Decimal	Binary	Octal	Hexadecimal
Radix	10	2	8	16
First seventeen positive integers	0	0	0	0
	1	1	1	1
	2	10	2	2
	3	11	3	3
	4	100	4	4
	5	101	5	5
	6	110	6	6
	7	111	7	7
	8	1000	10	8
	9	1001	11	9
	10	1010	12	A
	11	1011	13	B
	12	1100	14	C
	13	1101	15	D
	14	1110	16	E
	15	1111	17	F
	16	10000	20	10

- In the 2009 film Avatar, Na'vi race employs an octal numeral system.

Conversion from Decimal Integer

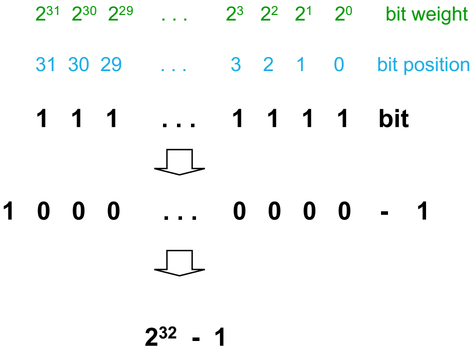
- Step 1: Divide the decimal number by the radix (number base)
- Step 2: Save the remainder (first remainder is the least significant digit)
- Repeat steps 1 and 2 until the quotient is zero
- Result is in reverse order of remainders

EX: L02-1

- EX1: Convert 36_8 to binary value
- EX2: Convert 36_{10} to binary value

Unsigned Binary Representation

Hex	Binary	Decimal
0x00000000	0…0000	0
0x00000001	0…0001	1
0x00000002	0…0010	2
0x00000003	0…0011	3
0x00000004	0…0100	4
0x00000005	0…0101	5
0x00000006	0…0110	6
0x00000007	0…0111	7
0x00000008	0…1000	8
0x00000009	0…1001	9
	…	
0xFFFFFFFC	1…1100	$2^{32} - 4$
0xFFFFFFFD	1…1101	$2^{32} - 3$
0xFFFFFFFE	1…1110	$2^{32} - 2$
0xFFFFFFFF	1…1111	$2^{32} - 1$



2^{31} 2^{30} 2^{29} . . . 2^{3} 2^{2} 2^{1} 2^{0} bit weight

31 30 29 . . . 3 2 1 0 bit position

1 1 1 . . . 1 1 1 1 bit

1 0 0 0 . . . 0 0 0 0 - 1

2^{32} - 1

Signed Binary Representation

	2'sc binary	decimal
-2^3 =	1000	-8
$-(2^3 - 1)$ =	1001	-7
	1010	-6
	1011	-5
	1100	-4
	1101	-3
	1110	-2
	1111	-1
	0000	0
	0001	1
	0010	2
	0011	3
	0100	4
	0101	5
	0110	6
$2^3 - 1$ =	0111	7

complement all the bits

0101

and add a 1

0110

1011

and add a 1

1010

complement all the bits

EX: L02-2

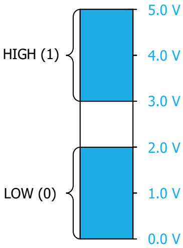
- For an n-bit signed binary numeral system, what’s the largest positive number and the smallest negative number?

Digital Signal Representation

- Active HIGH
  - High voltage means On

- Active LOW
  - Low voltage means On

Logic 0	Logic 1
False	True
Off	On
LOW	HIGH
No	Yes
Open switch	Closed switch



Addition & Subtraction

- Just like in grade school (carry/borrow 1s)

```
  0111        0111        0110
+ 0110      - 0110      - 0101
-------     ------      -------
```

- Two's complement operations are easy: do subtraction by negating and then adding

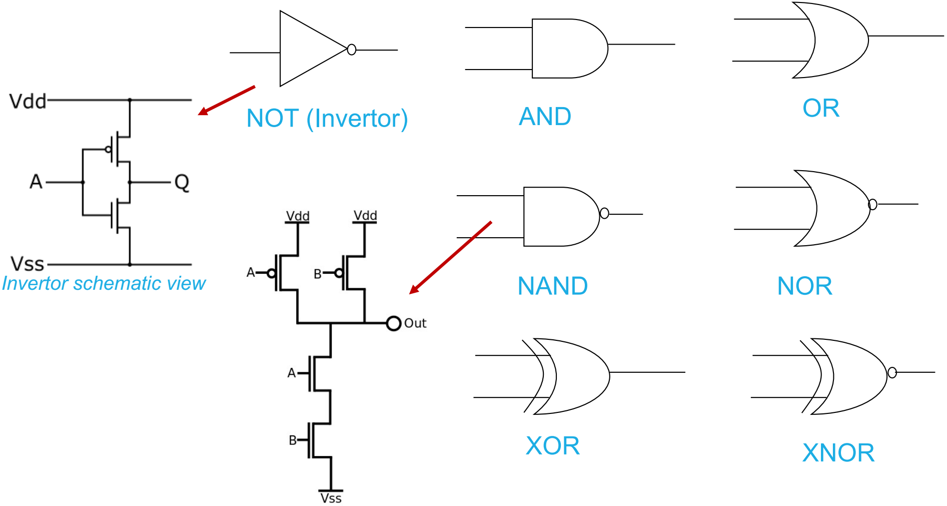
```
  0111    ->      0111
- 0110    ->    + 1010
------          -------
```

- Overflow (result too large for finite computer word). E.g., adding two n-bit numbers does not yield an n-bit number

```
  0111
+ 0001
------
```

Logic Gates

Logic Gates



NOT (Invertor)

AND

OR

Invertor schematic view

NAND

NOR

XOR

XNOR

❑ What is the schematic view of an AND gate?

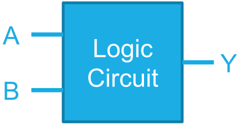
EX: L02-3

- Please draw NOR gate schematic view

Truth Table

- A means for describing how a logic circuit's output depends on the logic levels present at the circuit's inputs

- The number of input combinations will equal 2^N for an N-input truth table



Inputs		Output
A	B	Y
0	0	0
0	1	0
1	0	0
1	1	1

EX: L02-4

Determine the true table of a three-input AND gate