Lecture 27:
“Future of Computing Systems”

John P. Shen & Zhiyi Yu (content from Randy Bryant)
December 7, 2016
Moore’s Law Origins

April 19, 1965

Cramming more components onto integrated circuits

With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip

By Gordon E. Moore
Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.
Moore’s Law Origins

- **Moore’s Thesis**
  - Minimize price per device
  - Optimum number of devices / chip increasing 2x / year

- **Later**
  - 2x / 2 years
  - “Moore’s Prediction”
Moore’s Law: 50 Years

Transistor Count by Year

1.0E+03 1.0E+04 1.0E+05 1.0E+06 1.0E+07 1.0E+08 1.0E+09 1.0E+10


Sample of 117 processor chips

- Desktop
- Embedded
- GPU
- Server
- General Trend
- Moore’s Prediction
What Moore’s Law Has Meant

- **1976 Cray 1**
  - 250 M Ops/second
  - ~170,000 chips
  - 0.5B transistors
  - 5,000 kg, 115 KW
  - $9M
  - 80 manufactured

- **2014 iPhone 6**
  - > 4 B Ops/second
  - ~10 chips
  - > 3B transistors
  - 120 g, < 5 W
  - $649
  - 10 million sold in first 3 days
What Moore’s Law Has Meant

- **1965 Consumer Product**
- **2015 Consumer Product**

Apple A8 Processor
2 B transistors
Visualizing Moore’s Law to Date

If transistors were the size of a grain of sand

Intel 4004
1970
2,300 transistors

Apple A8
2014
2 B transistors

Intel 4004: 2,300 transistors, 0.1 g

Apple A8: 2 B transistors, 88 kg
Moore’s Law Economics

Consumer products sustain the $300B semiconductor industry

Better Products → Sales $$ → New Technology → Product Design → Capital + R&D Investment

Product Design

New Technology

Capital + R&D Investment

Sales $$

Better Products
What Moore’s Law Has Meant

9 generations of iPhone since 2007
What Moore's Law Could Mean

Kurzweil, *The Singularity is Near*, 2005
What Moore’s Law Could Mean

- 2015 Consumer Product
  - Portable
- 2065 Consumer Product
  - Low power
  - Will drive markets & innovation
Requirements for Future Technology

- **Must be suitable for portable, low-power operation**
  - Consumer products
  - Internet of Things components
  - Not cryogenic, not quantum

- **Must be inexpensive to manufacture**
  - Comparable to current semiconductor technology
    - $O(1)$ cost to make chip with $O(N)$ devices

- **Need not be based on transistors**
  - Memristors, carbon nanotubes, DNA transcription, ...
  - Possibly new models of computation
  - But, still want lots of devices in an integrated system
Moore’s Law: 100 Years

Device Count by Year


- **General Trend**
- **Moore’s Prediction**
- **Desktop**
- **Embedded**
- **GPU**
- **Server**

10^{17} devices!
Visualizing $10^{17}$ Devices

If devices were the size of a grain of sand

- $0.1 \text{ m}^3$, $3.5 \times 10^9$ grains
- $1\text{ million m}^3$, $0.35 \times 10^{17}$ grains
Increasing Transistor Counts

1. Chips have gotten bigger
   ▪ 1 area doubling / 10 years

2. Transistors have gotten smaller
   ▪ 4 density doublings / 10 years

Will these trends continue?
Chips Have Gotten Bigger

- Intel 4004
  - Year: 1970
  - Number of Transistors: 2,300
  - Dimension: 12 mm$^2$

- Apple A8
  - Year: 2014
  - Number of Transistors: 2 B
  - Dimension: 89 mm$^2$

- IBM z13
  - Year: 205
  - Number of Transistors: 4 B
  - Dimension: 678 mm$^2$
Chip Size Trend

Area by Year

Year

Area (mm²)


Desktop
Embedded
GPU
Server
Trend

2x every 9.5 years
Chip Size Extrapolation

Area by Year

Area (mm^2)

Year

1970 1990 2010 2030 2050

2016 (J.P. Shen)
Extrapolation: The iPhone 3G

Apple A59
2065
$10^{17}$ transistors
173 $cm^2$
Transistors Have Gotten Smaller

- Area $A$
- $N$ devices
- Linear Scale $L$

$$L = \sqrt{\frac{A}{N}}$$
Linear Scaling Trend

1/2x every 5 years ➔
2x transistor density every 2.5 years
Decreasing Feature Sizes

Intel 4004
1970
2,300 transistors
$L = 72,000$ nm

Apple A8
2014
2 B transistors
$L = 211$ nm
Submillimeter Dimensions

10^{-3}  1 millimeter (mm)

- 500μm: Length of amoeba

10^{-4}

- 72μm: Intel 4004 linear scale
- 50μm: Average size of cell in human body

10^{-5}

- 10μm: Thickness of sheet of plastic food wrap
- 5μm: Spider silk thickness
- 2μm: E coli bacterium length

10^{-6} 1 micrometer (μm)
Submicrometer Dimensions

- $10^{-6}$ 1 micrometer
  - 400-700nm: Visible light wavelengths
- $10^{-7}$
  - 211nm: Apple A8 linear scale
  - 30nm: Minimum cooking oil smoke particle diameter
- $10^{-8}$
  - 9nm: Cell membrane thickness
  - 2nm: DNA helix diameter
- $10^{-9}$ 1 nanometer (nm)
  - 1nm: Carbon nanotube diameter

12/07/2016 (J.P. Shen)
Linear Scaling Extrapolation

Linear Scale by Year

Sqrt(A/N) (nm)

Year

Desktop
Embedded
GPU
Server
Trend

230 pm
Subnanometer Dimensions

10^{-9}  1 nanometer (nm)

1 fm  Carbon nanotube diameter
543 pm  Silicon crystal lattice spacing
230 pm  2065 linear scale projection

10^{-10}  

74 pm  Spacing between atoms in hydrogen molecule
53 pm  Electron-proton spacing in hydrogen (Bohr radius)

10^{-11}  


10^{-12}  1 picometer (pm)

2.4 pm  Electron wavelength (Compton wavelength)
Reaching 2065 Goal

- **Target**
  - $10^{17}$ devices
  - $400 \text{ mm}^2$
  - $L = 63 \text{ pm}$

- **Is this possible?**

  **No!**

  Not with 2-d fabrication
Fabricating in 3 Dimensions

**Parameters**

- $10^{17}$ devices
- 100,000 logical layers
  - Each 50 nm thick
  - ~1,000,000 physical layers
    - To provide wiring and isolation
- $L = 20$ nm
  - 10x smaller than today
3D Fabrication Challenges

- **Yield**
  - How to avoid or tolerate flaws

- **Cost**
  - High cost of lithography

- **Power**
  - Keep power consumption within acceptable limits
  - Limited energy available
  - Limited ability to dissipate heat
- Pattern entire chip in one step
- Modern chips require ~60 lithography steps
- Fabricate $N$ transistor system with $O(1)$ steps
Fabrication Costs

- Stepper
  - Most expensive equipment in fabrication facility
  - Rate limiting process step
    - 18s / wafer
  - Expose 858 mm$^2$ per step
    - 1.2% of chip area
Fabrication Economics

Currently
- Fixed number of lithography steps
- Manufacturing cost $10–$20 / chip
  - Including amortization of facility

Fabricating 1,000,000 physical layers
- Cannot do lithography on every step

Options
- Chemical self assembly
  - Devices generate themselves via chemical processes
- Pattern multiple layers at once
Samsung V-Nand Flash Example

- Build up layers of unpatterned material
- Then use lithography to slice, drill, etch, and deposit material across all layers
- ~30 total masking steps
- Up to 48 layers of memory cells
- Exploits particular structure of flash memory circuits
Meeting Power Constraints

- 2 B transistors
- 2 GHz operation
- 1—5 W

- 64 B neurons
- 100 Hz operation
- 15—25 W
  - Liquid cooling
  - Up to 25% body’s total energy consumption

Can we increase number of devices by 500,000x without increasing power requirement?
Challenges to Moore’s Law: Economic

- **Growing Capital Costs**
  - State of art fab line ~$20B
  - Must have very high volumes to amortize investment
  - Has led to major consolidations
Dennard Scaling

- Due to Robert Dennard, IBM, 1974
- Quantifies benefits of Moore’s Law

**How to shrink an IC Process**
- Reduce horizontal and vertical dimensions by $k$
- Reduce voltage by $k$

**Outcomes**
- Devices / chip increase by $k^2$
- Clock frequency increases by $k$
- Power / chip constant

**Significance**
- Increased capacity and performance
- No increase in power
End of Dennard Scaling

■ What Happened?
  ▪ Can’t drop voltage below ~1V
  ▪ Reached limit of power / chip in 2004
  ▪ More logic on chip (Moore’s Law), but can’t make them run faster
    ▪ Response has been to increase cores / chip
Final Thoughts

- Compared to future, past 50 years will seem fairly straightforward
  - 50 years of using photolithography to pattern transistors on two-dimensional surface

- Questions about future integrated systems
  - Can we build them?
  - What will be the technology?
  - Are they commercially viable?
  - Can we keep power consumption low?
  - What will we do with them?
  - How will we program / customize them?
"Computing Systems Mega-Trends 2015-2025"

John P. Shen
December 7, 2016

- Silicon Technology
- Mobile Devices
- Software Development
- Cloud Infrastructure
Silicon Technology: Potentials of 3D Die Stacking

- Standard C4 bumps
- Thru-die vias
- Die-to-die Via interface

Diagram showing Thin Die and Thick Die with connections through Bulk Si and Heat Sink.
Three Limitations to Moore’s Law

These limitations will make it very challenging to continue integrating systems

[Bryan Black, 2015, AMD]
Strategic Vision (Stacked System)

- The Stacked System model integrates dies from disparate technologies using a combination of 2.5D and 3D technology.
- This construction model enables:
  - Disparate die integration to improve form factors and reduce system overheads.
  - Die splitting to reduce process node complexity and cost.
- Results in an interesting business model opportunity.

[Bryan Black, 2015, AMD]
The Road to

“FIJI”

Featuring Die Stacking
and HBM Technology

[Bryan Black, 2015, AMD]
“Fiji” Chip

DETAILED LOOK

- First high-volume interposer
- First TSVs and µBumps in the graphics industry
- Most discrete dies in a single package at 22
- Total 1011 sq. mm.

- Graphics Core Next Architecture
- 64 Compute Units
- 4096 Stream Processors
- 596 sq. mm. Engine

- 4GB High-Bandwidth Memory
- 4096-bit wide interface
- 512 GB/s Memory Bandwidth

[Source: Bryan Black, 2015, AMD]
Die stacking facilitates the integration of discrete dies

8.5 years of development by AMD and its technology partners

[DIE STACKING TECHNOLOGY]

[Bryan Black, 2015, AMD]
SMALL SIZE, GIANT IMPACT

30%
Shorter than the AMD Radeon™ R9 290X (11.5”)

Board shot shown for illustration purposes only. Final board design may differ.

[ Bryan Black, 2015, AMD ]
Mobile Devices: What’s the Next Epoch?

  - **Epoch #1**
  - **Win**: Personal Computer (portable but not real mobile)
  - **Cat5**: Personal Computer (not real mobile)

- **Laptop**: 1992 → 2007 → 2014 →
  - **Epoch #2**
  - **Win95**: Laptop PC (mobile but not real PC)
  - **Wifi**: Smartphone

- **Netbook**: 2007 → 2014 →
  - **Epoch #3**
  - **UX**: Mobile Computer
  - **Cheap**: Laptop PC and Smartphone

- **Tablet**: 2007 → 2014 →
  - **Epoch #3**
  - **UX**: Laptop PC and Smartphone
  - **Not PC**: Mobile Personal Computing (replaces laptop & smartphone)

- **Thin Laptops + Smart Phones**: 2014 →
  - **Win8**: Laptop PC and Smartphone
  - **4G**: Mobile Personal Computing (replaces laptop & smartphone)

- **PC 4.0 + Accessories**: 2014 →
  - **Win10**: Mobile Personal Computing (replaces laptop & smartphone)
  - **5G**: Mobile Personal Computing (replaces laptop & smartphone)

**Epochs**:
- **Epoch #1**: 1981 → 1992
- **Epoch #2**: 1992 → 2007
- **Epoch #3**: 2007 → 2014
- **Epoch #4?**: 2014 →
## Seamless Roaming Experience Across 5+ Screens

<table>
<thead>
<tr>
<th>PC</th>
<th>smartphone</th>
<th>tablet</th>
<th>living room</th>
<th>dashboard</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="logo.png" alt="Apple" /></td>
<td><img src="logo.png" alt="iPhone" /></td>
<td><img src="logo.png" alt="iPad" /></td>
<td><img src="logo.png" alt="Apple TV" /></td>
<td><img src="logo.png" alt="CarPlay" /></td>
</tr>
<tr>
<td>Mac computers</td>
<td>iPhone</td>
<td>iPad</td>
<td>Apple TV</td>
<td>CarPlay</td>
</tr>
<tr>
<td><img src="logo.png" alt="Google" /></td>
<td><img src="logo.png" alt="Android" /></td>
<td><img src="logo.png" alt="Android tablets" /></td>
<td><img src="logo.png" alt="Google TV, Chromecast" /></td>
<td><img src="logo.png" alt="Open Automotive Alliance" /></td>
</tr>
<tr>
<td>Chrome browser</td>
<td>Android</td>
<td>Android tablets</td>
<td>Google TV, Chromecast</td>
<td>Open Automotive Alliance</td>
</tr>
<tr>
<td><img src="logo.png" alt="Microsoft" /></td>
<td><img src="logo.png" alt="Windows Phone" /></td>
<td><img src="logo.png" alt="Windows 8" /></td>
<td><img src="logo.png" alt="Xbox" /></td>
<td>Windows in the Car</td>
</tr>
<tr>
<td>Windows, Office</td>
<td>Windows Phone</td>
<td>Windows 8</td>
<td>Xbox</td>
<td>Windows in the Car</td>
</tr>
<tr>
<td><img src="logo.png" alt="Amazon" /></td>
<td><img src="logo.png" alt="Online store" /></td>
<td><img src="logo.png" alt="Soon to come?" /></td>
<td><img src="logo.png" alt="Kindle Fire" /></td>
<td><img src="logo.png" alt="Fire TV" /></td>
</tr>
<tr>
<td>Online store</td>
<td>Soon to come?</td>
<td>Kindle Fire</td>
<td>Fire TV</td>
<td></td>
</tr>
</tbody>
</table>

Carnegie Mellon University
From Two Screens to Multiple Screens

Computers

Tablets + E-readers

(Smart)phones

Televisions & Game Consoles

Car displays

Gadgets & Wearables
Software Development: Dominant Mobile Platforms
The Android & iOS Duopoly Continues to Strengthen

By Unit Volume

By Device Profits

Data: Sameer Singh, Gartner, IDC, vendor data, VisionMobile estimates

Data: McKinsey, Asymco, Canaccord, VisionMobile estimates

(Caribbean Mellon University)
Dominant Mobile Platforms 2014

Most Popular Platform by Country

iOS or Android dominate every market

Source: Developer Economics Q1 2014
Development Languages and Tools

**iOS, Android and WP Have a Language Lock-In on Most of Their Developers**

Primary language use by platform indicates levels of developer loyalty.

- **Platform Priority**: % of all developers
  - **iOS**: 58% (32% Android, 12% BlackBerry 10, 9% Windows Phone)
  - **Android**: 42% (10% BlackBerry 10, 13% Windows Phone)
  - **Windows Phone**: 10% (8% Android, 16% BlackBerry 10)
  - **BlackBerry 10**: 3% (8% Android, 11% Windows Phone)

- **Programming Languages**:
  - **Objective-C**: iOS
  - **C/C++**: Android
  - **C#**: Windows Phone
  - **Java**: iOS
  - **HTML/CSS/JavaScript**: BlackBerry 10
  - **Visual Development Tool**: Others
  - **ActionScript**: Others

Source: Developer Economics: State of Nation Q3 2014 | www.DeveloperEconomics.com/go | Licensed under CC BY ND | Copyright VisionMobile
Development Language Popularity

HTML5 TOPS LANGUAGE RANKING FOR MOBILE APP DEVELOPERS
The languages most developers use are not always the ones they use most

% of developers using

HTML/ CSS/ JavaScript  42%

Java  38%

C/C++  26%

Objective - C  24%

C#  23%

Visual Tool  17%

JavaScript/ TypeScript  15%

PHP  11%

Python  7%

Action Script  3%

Ruby  3%

Lua  3%

RedMonk top 20 ranking based on GitHub projects and StackOverflow tags

VisionMobile top 12 ranking of most popular primary languages

Most Widely Used Languages (in mobile)

Most Actively Used Languages (in mobile)

Java  26%

Objective-C  17%

HTML/CSS/JavaScript  17%

C#  17%

C/C++  13%

Visual Development Tool  10%

(Java/JavaScript/Type) Script  9%

PHP  8%

Python  8%

Action Script  7%

Lua  7%

VisionMobile top 12 ranking of most popular primary languages

Sources: Developer Economics: State of Nation Q1 2014 | www.DeveloperEconomics.com/go | Licensed under CC BY ND | Copyright VisionMobile
Disruptive Trends in Web-Based SW Development

1) WebGL
   • Built-in OpenGL ES 2.0 APIs in a generic web browser
   • Enables the rendering of interactive 3D and 2D graphics within compatible web browsers without any plug-ins

2) WebRTC
   • Universal voice calling, video calls and P2P data connections in a generic web browser – no plug-ins required
   • Forthcoming W3C standard – already supported by Google Chrome and Mozilla Firefox

3) Backend as a Service (BaaS)
   • Developer-friendly, all-in-one cloud solutions that require minimal installation or maintenance

[Antero Taivalsaari, Nokia, 2014]
BaaS – Commonly Provided Features

- Cloud Data Storage
- Push Notifications
- 3rd Party Data Integration
- User Management
- Cross-Platform Support
- Versioning, analytics, etc.

[Antero Taivalsaari, Nokia, 2014]
Example Feature Set: Parse.com

Parse Data
Store your app's data in the cloud. No servers necessary.

Parse Hosting
A powerful web presence without all the hassle.

Parse Push
Creating, scheduling, and segmenting push notifications just got a whole lot easier.

Parse Social
Make your app social. Instantly.

Cloud Code
Run custom app code in the Parse Cloud. Say goodbye to servers.

Parse was recently acquired by Facebook

[Antero Taivalsaari, Nokia, 2014]
Cloud Infrastructure: Current Mega-Trends

The computing cloud ecosystem is maturing and several trends are becoming evident and dominant.

- Move Storage & Computing to the Cloud SW Defined & Virtualized Data Center
- Multi-Device Usage Mobile Experience Multi-Domain Wireless Connectivity
- Smart Sensing & User Data Analytics Internet Integration of “Sensory Swarm”
Cloud Infrastructure: Potential Disruptions

The current cloud architecture can and will be disrupted as players begin to create new and better consumer experiences.

- **Shift Computing to the Cloud Edge:**
  - Off Load Core Network Bandwidth Demand
  - Reduce Service Delivery Latency to Users

- **Truly Seamless Mobile Experience:**
  - Seamless Cross-Device Cross-Domain UX
  - Unify both Broadband and Broadcast UX

- **Human Sensing for Societal Good:**
  - Deliver Real IOT Value to Mobile Users
  - Use both Eulerian and Lagrangian Sensing
The Big Picture: Enabling Real-Time Video Processing at the Edge in Software

- Bring the cloud to the edge by integrating video caching with CloudRAN (large pool of baseband processing connected to Remote Radio Heads by fiber)
  - Use real-time video transrating to optimize bandwidth (based on device capability)
  - Results in lower latency for video and mobile cloud computing, as well as more efficient usage of available spectrum and bandwidth.

Note that even the handsets are software defined systems!

Local Cloud/CDN + RAN

Requires a software defined system for PHY (baseband) and video translating

The Cloud

Requires cloud computing and high performance storage

HSPA+, MPEG2

TD-LTE, H.264 MP 5.1

FDD LTE, H.264 CBP 4.1

TD-LTE, H.264 BP 3.1
Computing Megatrends

Mobile Supercomputing

Emerging Killer Applications
Computing Megatrends

- **Leading-Edge Supercomputing**
  - Current TOP100 supercomputers are Petascale ($10^{15}$ FLOPS) systems
  - Challenges for next 5 years: push towards Exascale ($10^{18}$ FLOPS) systems
  - Must improve performance/power efficiency from 1 GF/W to 100 GF/W

- **Mobile Cloud Edge Computing**
  - Push towards cloud computing creates huge network bandwidth demands
  - Tension will result in federated and fragmented cloud computing models
  - Wireless edge of the cloud will be core to computing and communication

- **Personal Computing Experience**
  - Continuation of Moore’s law expected for at least two more process nodes
  - 100 GF/W technology can provide mobile supercomputers for mass market
  - Dealing with legacy SW and device installed base will be a huge challenge
Mobile Supercomputing

Mobile Supercomputers
- Improving performance/power efficiency to 100 GFLOPS/W will enable a Terascale (10^{12} FLOPS) mobile supercomputer with a 10W power budget.
- An airborne supercomputer capable of 100 TFLOPS can then be deployed in an UAV (e.g. the RQ-1 and MQ-1 Predator drone) with a 1KW power budget.

Architecture Innovations
- Dataflow driven execution model supported by powerful SW tool chain and programmable and extremely energy-efficient HW fabric will be essential.
- Current vertical/proprietary solutions will be horizontalized and commoditized.

Form Factor Innovations
- Extreme integration via 3D TSV die stacking of diverse technology dies, e.g. many-core processors, high-BW DRAMs and SSDs, FPGA, and power delivery.
Emerging Killer Applications

- **Real-Time Environmental Sensing and Processing**
  - Highly mobile and autonomous real-time data collection, data analytics, and data inferences, without having to off-load to some remote cloud infrastructure.
  - Example: real-time traffic, special events monitoring, human mobility behaviors.

- **Rapid Situational Deployment of Cloud Resource**
  - Swarms of mobile/airborne connected vehicles equipped with supercomputing can become a highly distributed platform for Sensing. Analytics, and Services.
  - Such swarms of connected vehicles can provide low latency and high bandwidth city-scale services by functioning as the mobile edge of the cloud infrastructure.

- **Swarm-of-Drones Infrastructure for Demanding Scenarios**
  - Swarm of collaborating drones can be rapidly deployed to provide wireless communication and Petascale (10^{15} FLOPS) supercomputing infrastructure.