18-600 Foundations of Computer Systems

Lecture 27: "Future of Computing Systems"

John P. Shen & Zhiyi Yu (content from Randy Bryant)

December 7, 2016

Electrical & Computer ENGINEERING

Carnegie Mellon University

Moore's Law Origins

April 19, 1965

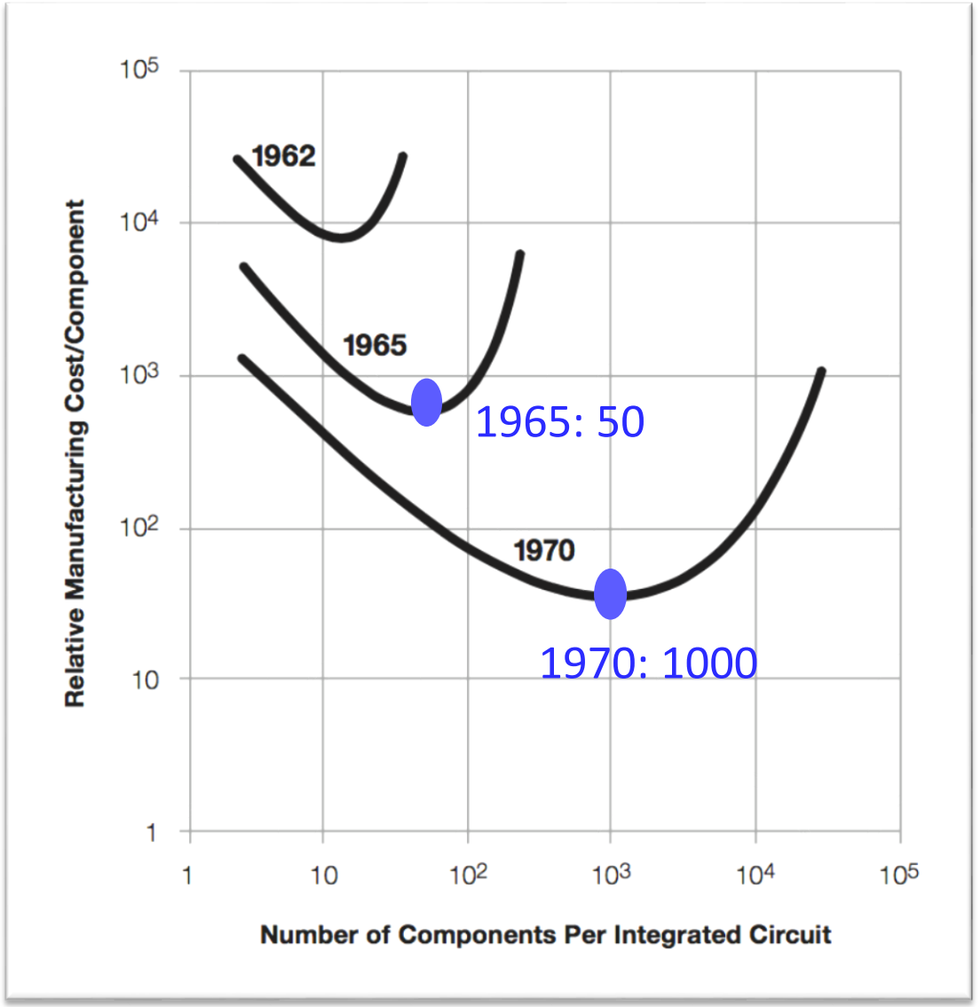
Cramming more components onto integrated circuits

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

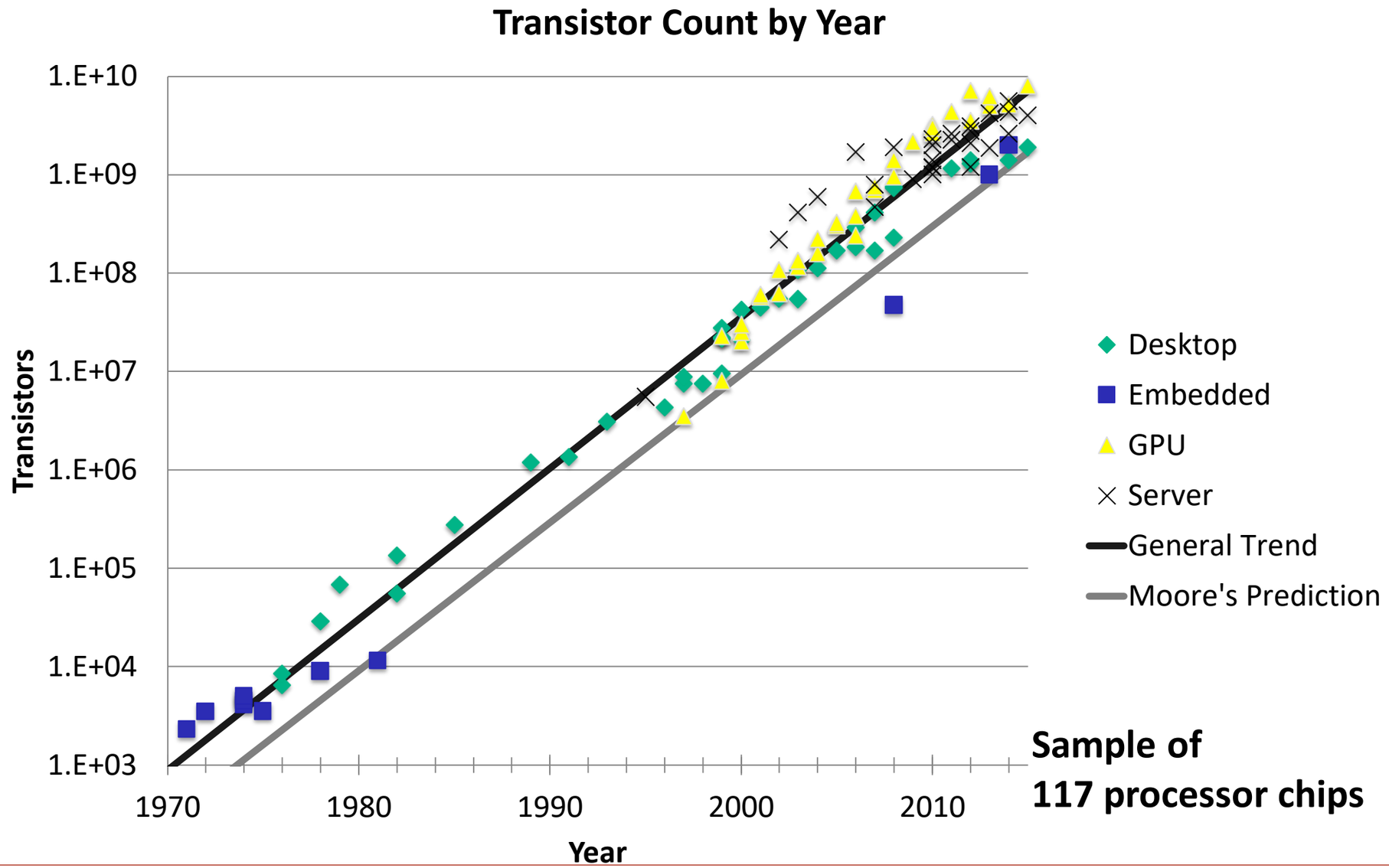
Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.

Moore's Law Origins

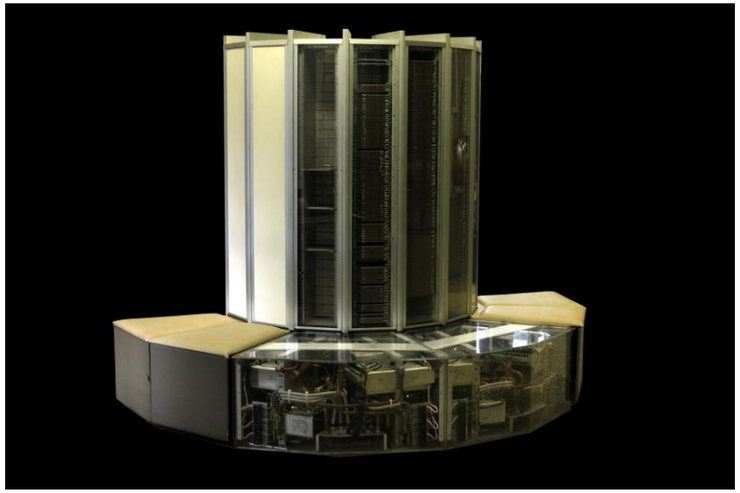


- **Moore's Thesis**
  - Minimize price per device
  - Optimum number of devices / chip increasing 2x / year
- **Later**
  - 2x / 2 years
  - "Moore's Prediction"

Moore's Law: 50 Years



What Moore's Law Has Meant



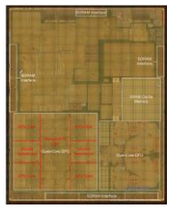
- **1976 Cray 1**
  - 250 M Ops/second
  - ~170,000 chips
  - 0.5B transistors
  - 5,000 kg, 115 KW
  - $9M
  - 80 manufactured

- **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

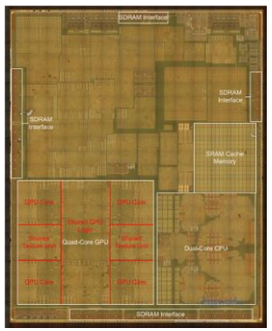
Visualizing Moore's Law to Date

If transistors were the size of a grain of sand

Intel 4004
1970
2,300 transistors

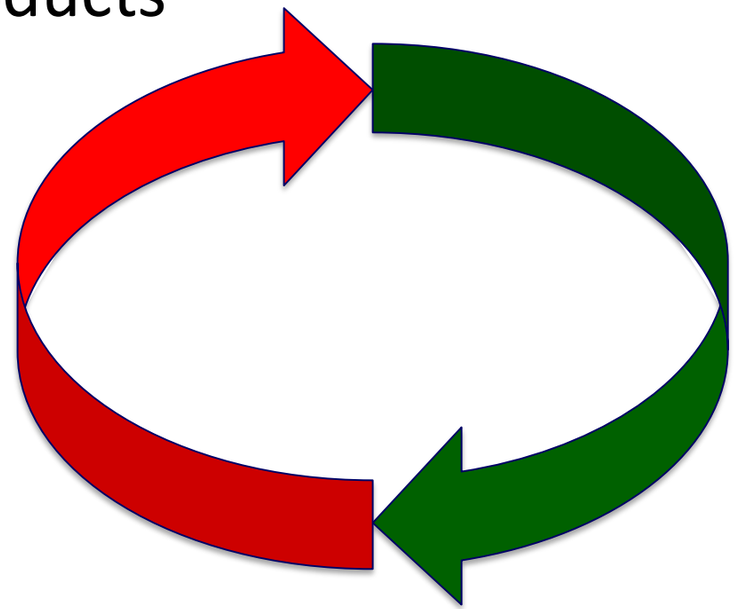
0.1 g

Apple A8
2014
2 B transistors



88 kg

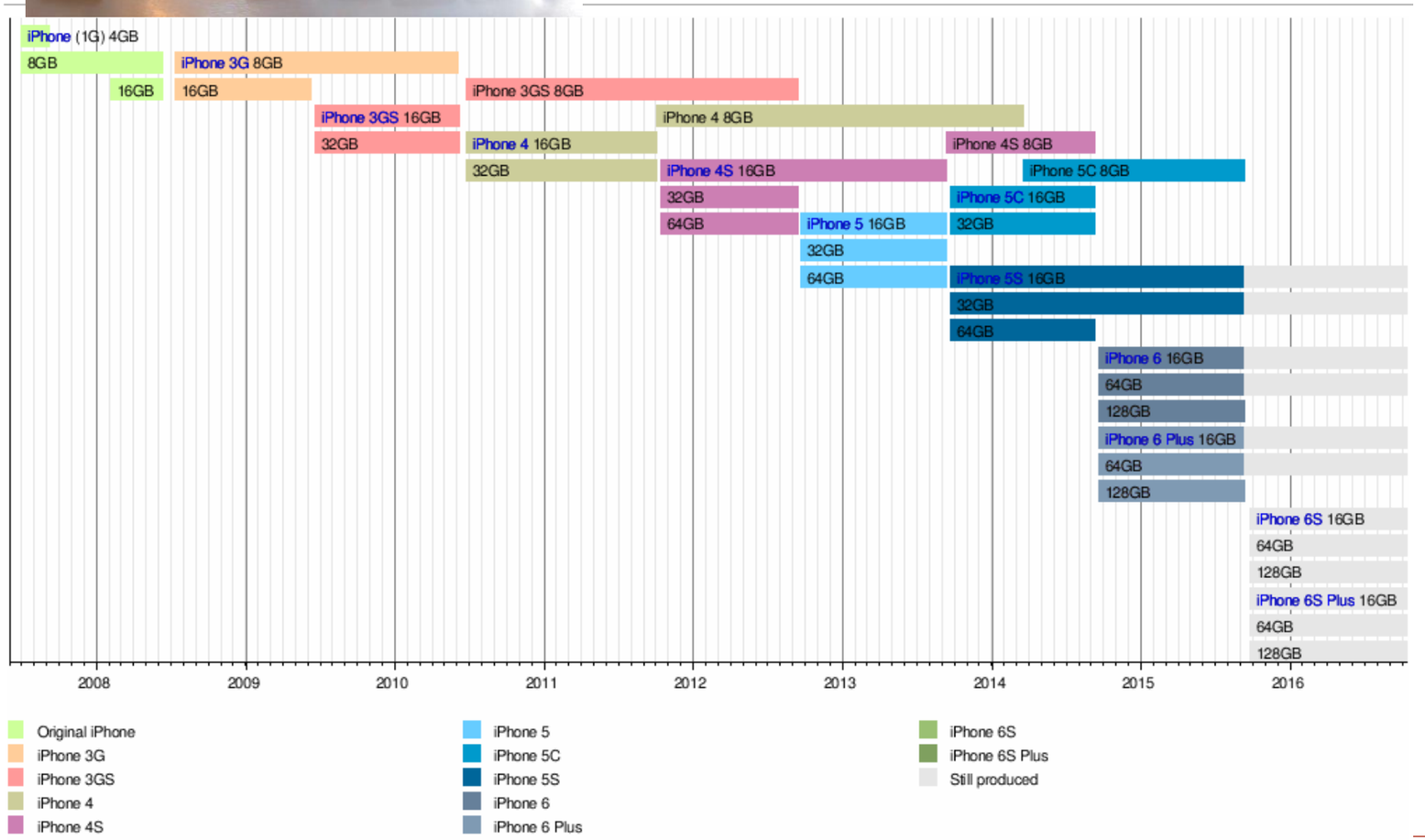
Moore's Law Economics



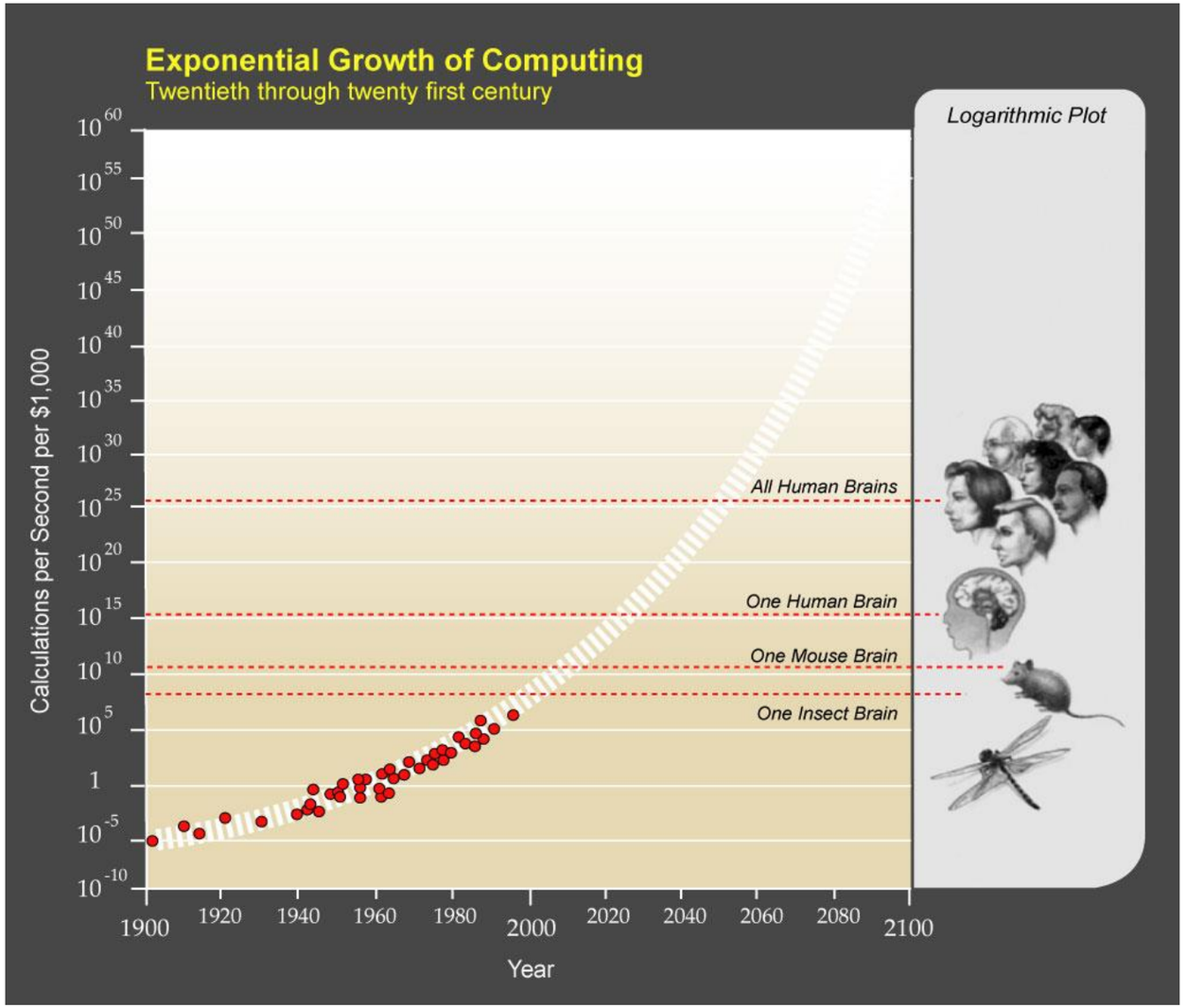
Consumer products sustain the $300B semiconductor industry

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**

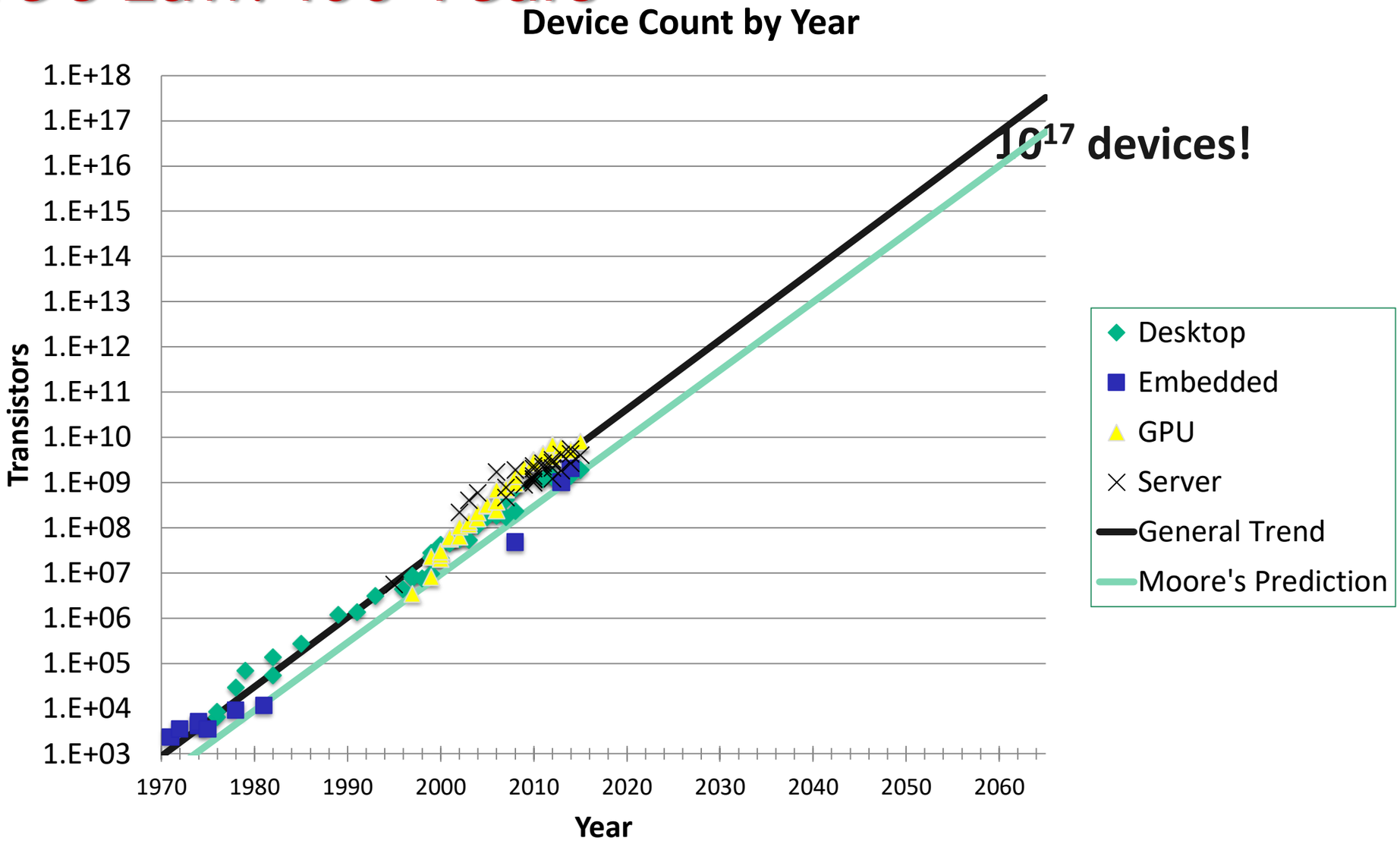
- **2065 Consumer Product**

- Portable
- 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



Visualizing 10^{17} Devices

If devices were the size of a grain of sand

0.1 m^3

3.5 X 10^9 grains

1 million m^3

0.35 X 10^{17} grains

Increasing Transistor Counts

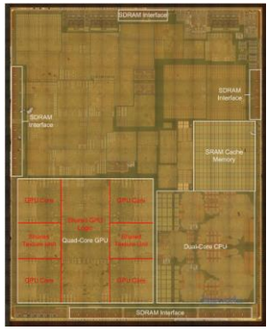
1. **Chips have gotten bigger**
   - 1 area doubling / 10 years
2. **Transistors have gotten smaller**
   - 4 density doublings / 10 years

Will these trends continue?

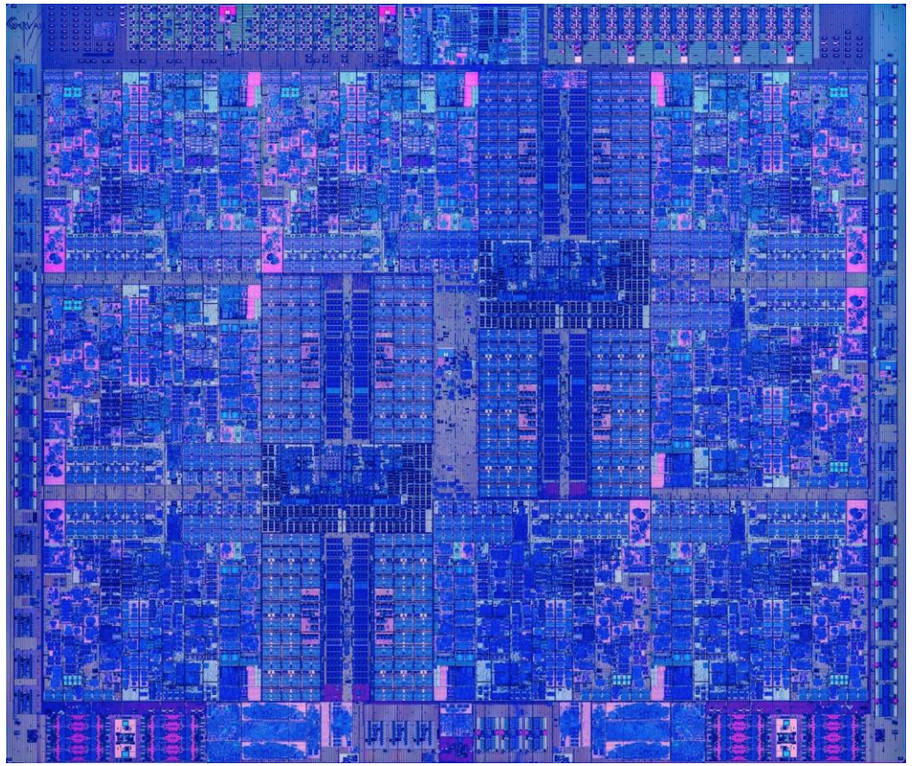
Chips Have Gotten Bigger

Intel 4004
1970
2,300 transistors
12 mm²

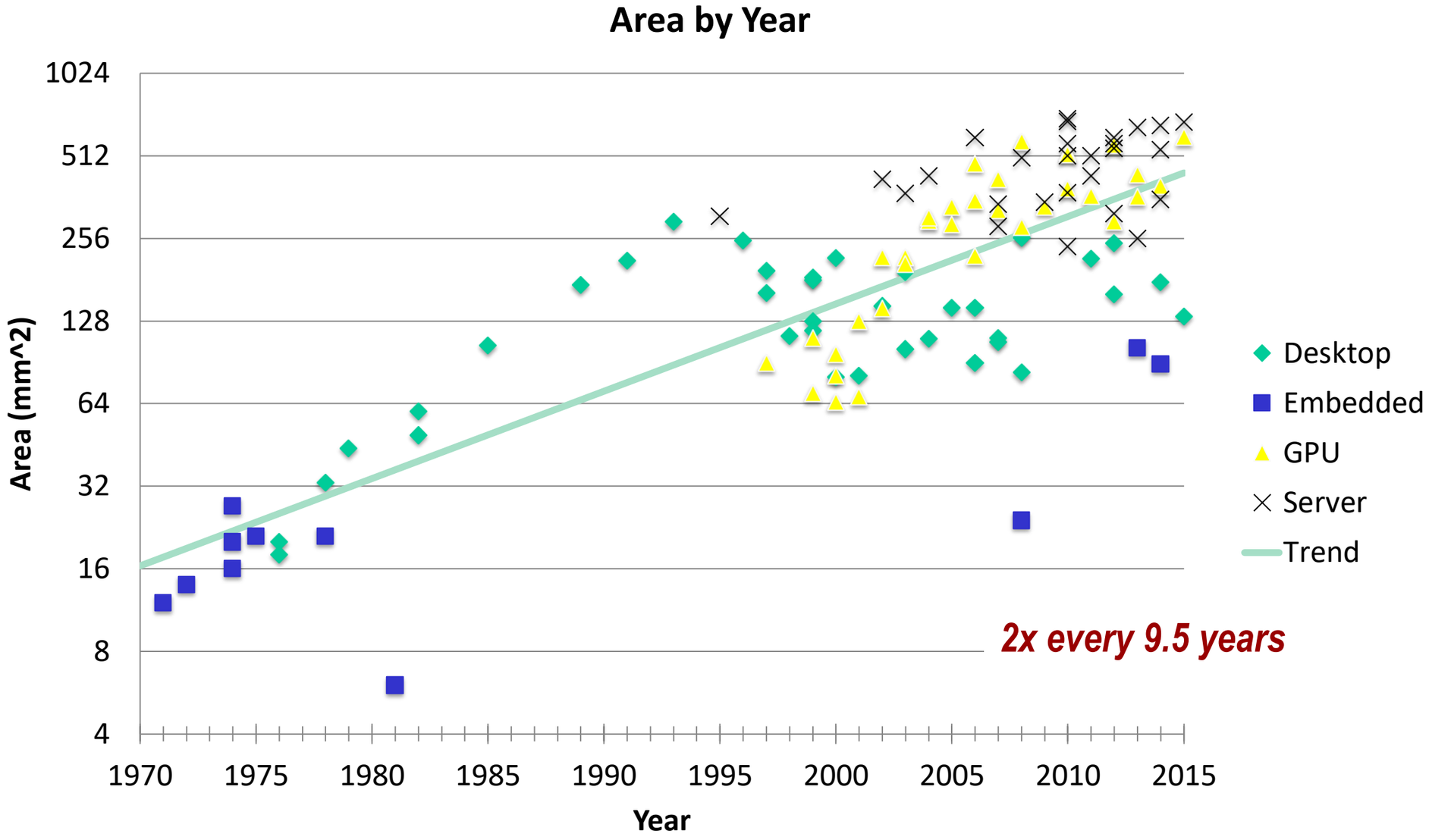
Apple A8
2014
2 B transistors
89 mm²



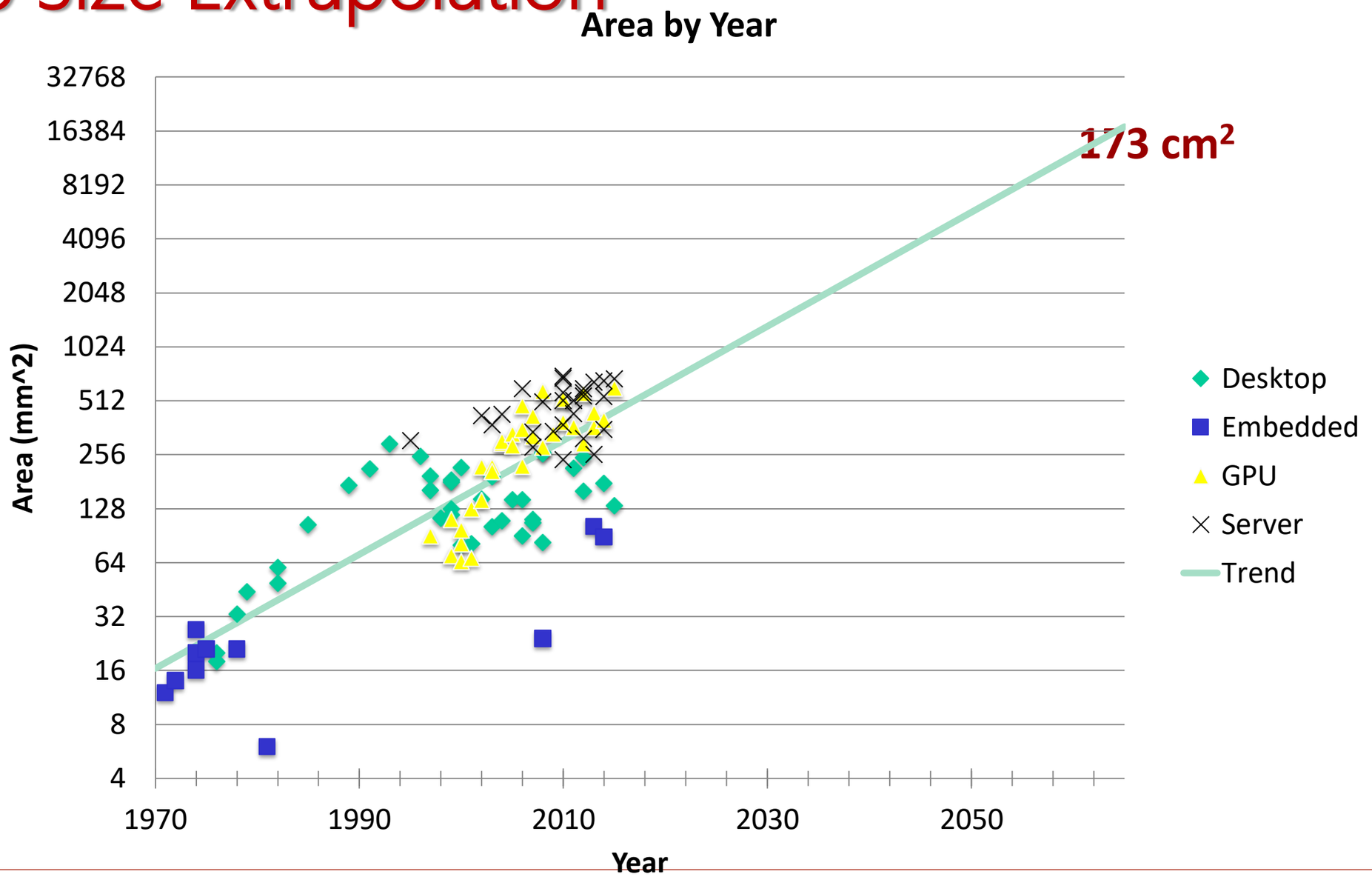
IBM z13
205
4 B transistors
678 mm²



Chip Size Trend



Chip Size Extrapolation



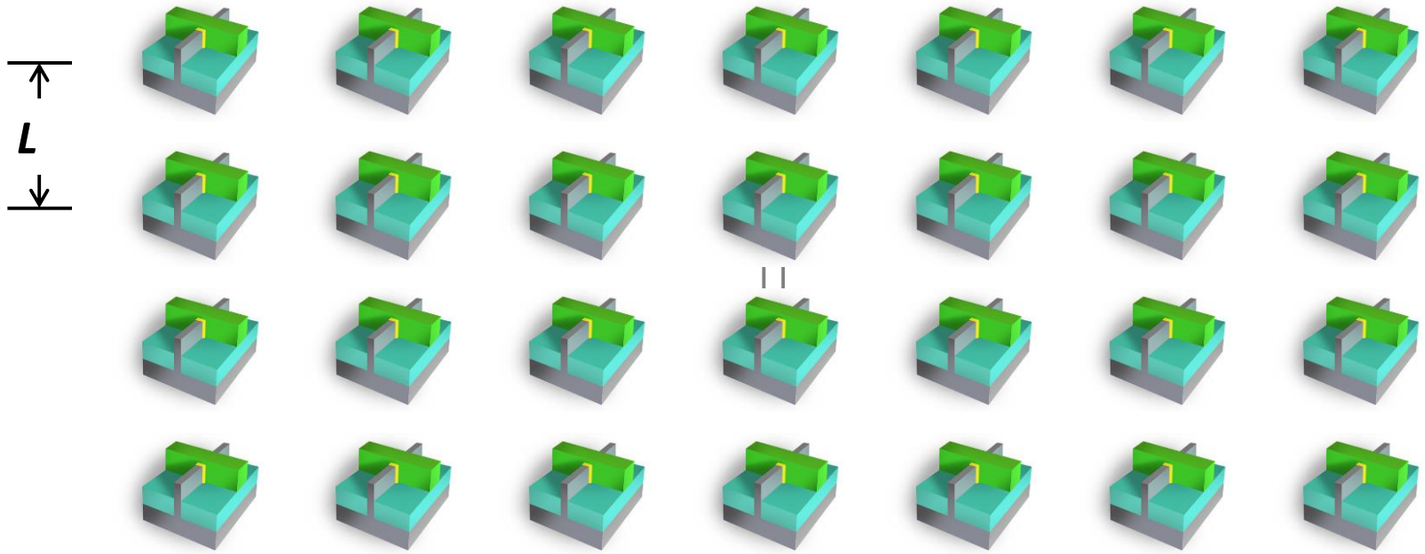
Extrapolation: The iPhone 31s

Apple A59
2065
10^{17} transistors
173 cm^2

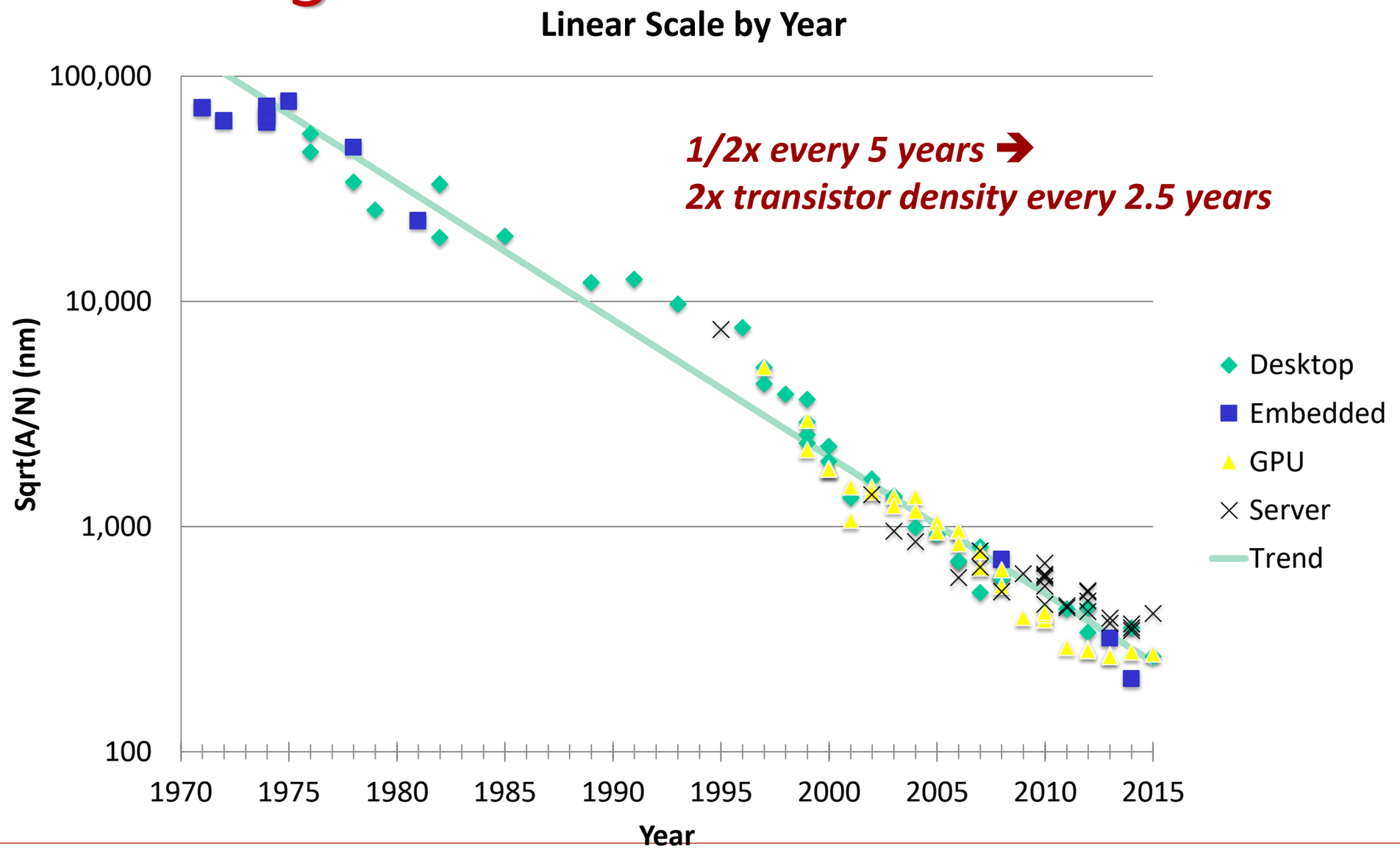
Transistors Have Gotten Smaller

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

$$L = \sqrt{A / N}$$



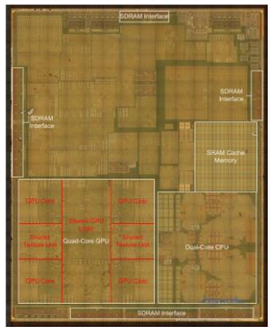
Linear Scaling Trend



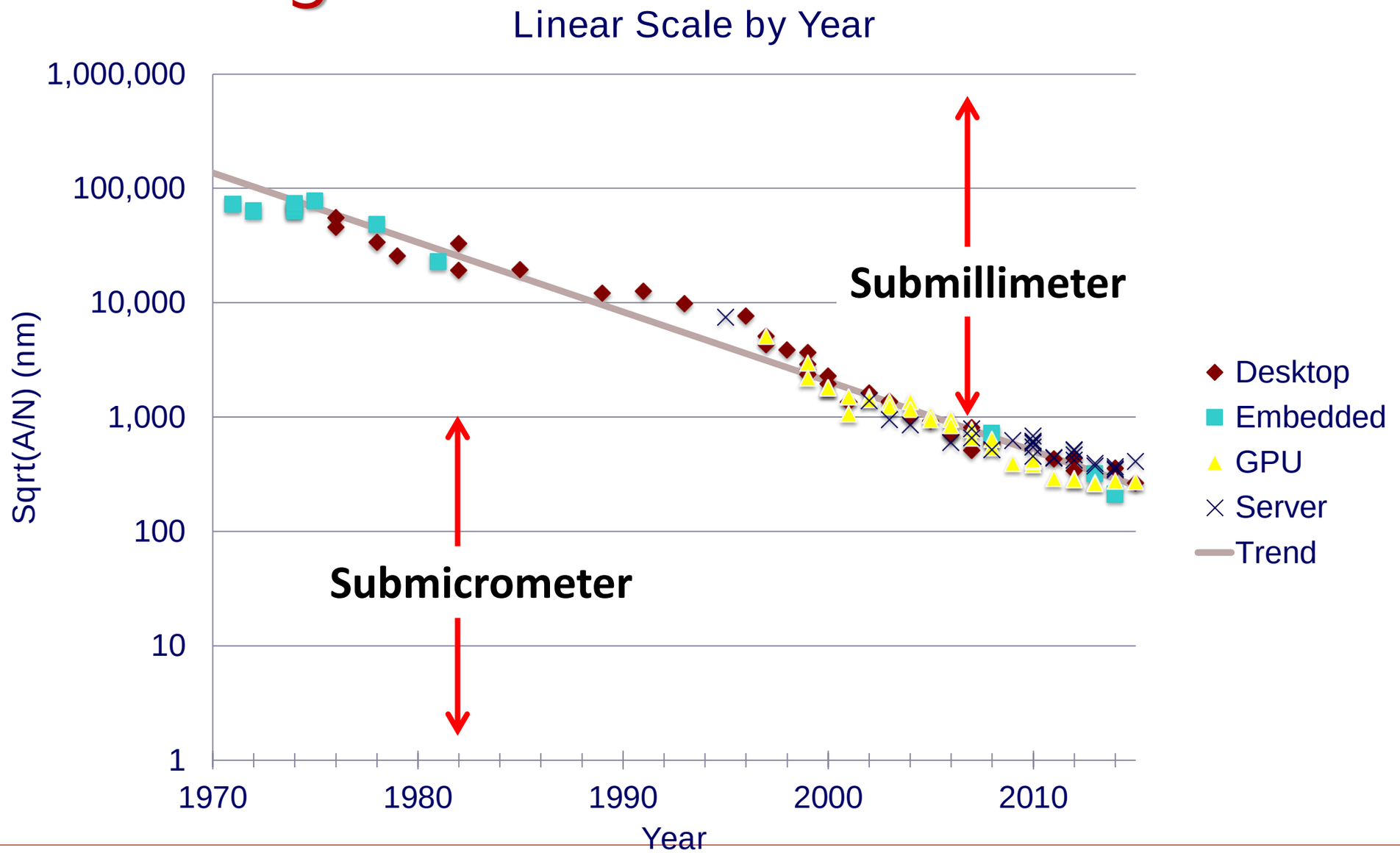
Decreasing Feature Sizes

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

Apple A8
2014
2 B transistors
***L* = 211 nm**



Linear Scaling Trend

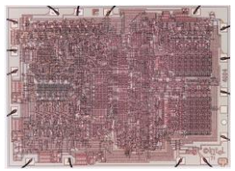


Submillimeter Dimensions

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

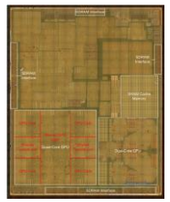
2µm: E coli bacterium length

10^{-6} 1 micrometer (µm)

Submicrometer Dimensions

10^{-6} 1 micrometer

400-700nm: Visible light wavelengths



211nm: Apple A8 linear scale

10^{-7}

30nm: Minimum cooking oil smoke particle diameter

10^{-8}

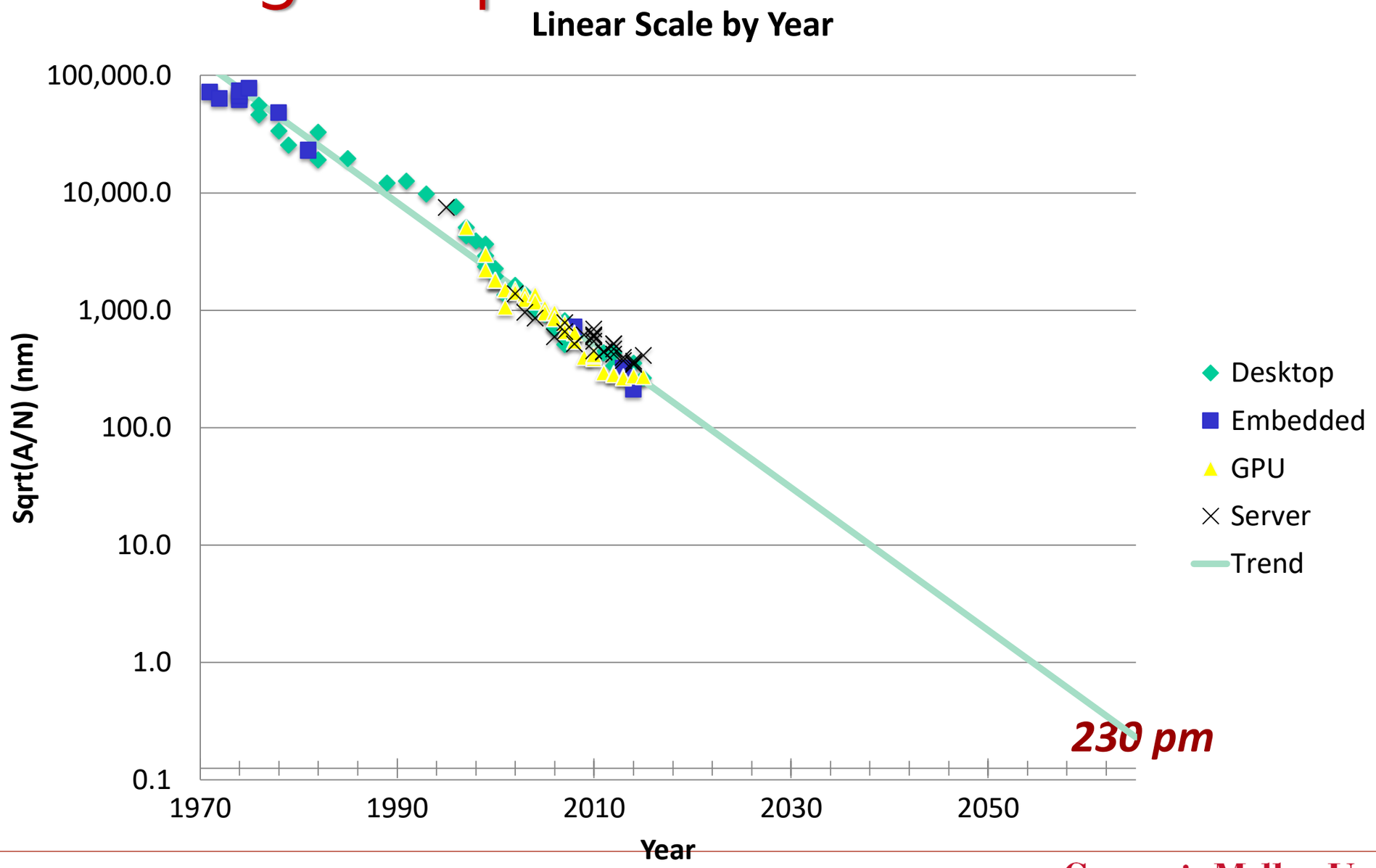
9nm: Cell membrane thickness

2nm: DNA helix diameter

10^{-9} 1 nanometer (nm)

1nm: Carbon nanotube diameter

Linear Scaling Extrapolation

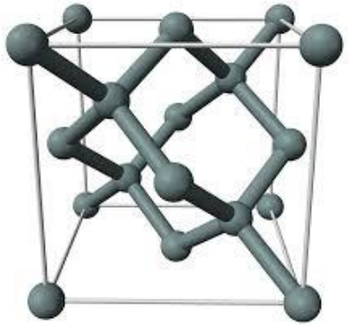


Subnanometer Dimensions

10^{-9} 1 nanometer (nm)

1nm: Carbon nanotube diameter

543pm: Silicon crystal lattice spacing

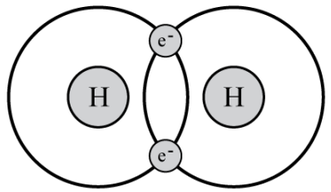


230pm: 2065 linear scale projection

10^{-10}

74pm: Spacing between atoms in hydrogen molecule

53pm: Electron-proton spacing in hydrogen (Bohr radius)



10^{-11}

2.4pm: Electron wavelength (Compton wavelength)

10^{-12} 1 picometer (pm)

Reaching 2065 Goal

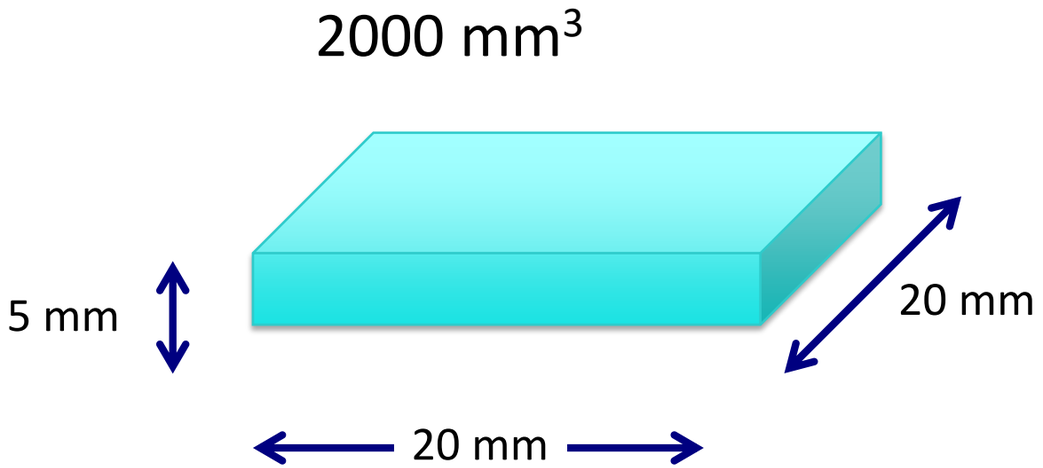
- **Target**
  - 10^{17} devices
  - 400 mm^2
  - L = 63 pm

- **Is this possible?**

No!

Not with 2-d fabrication

Fabricating in 3 Dimensions



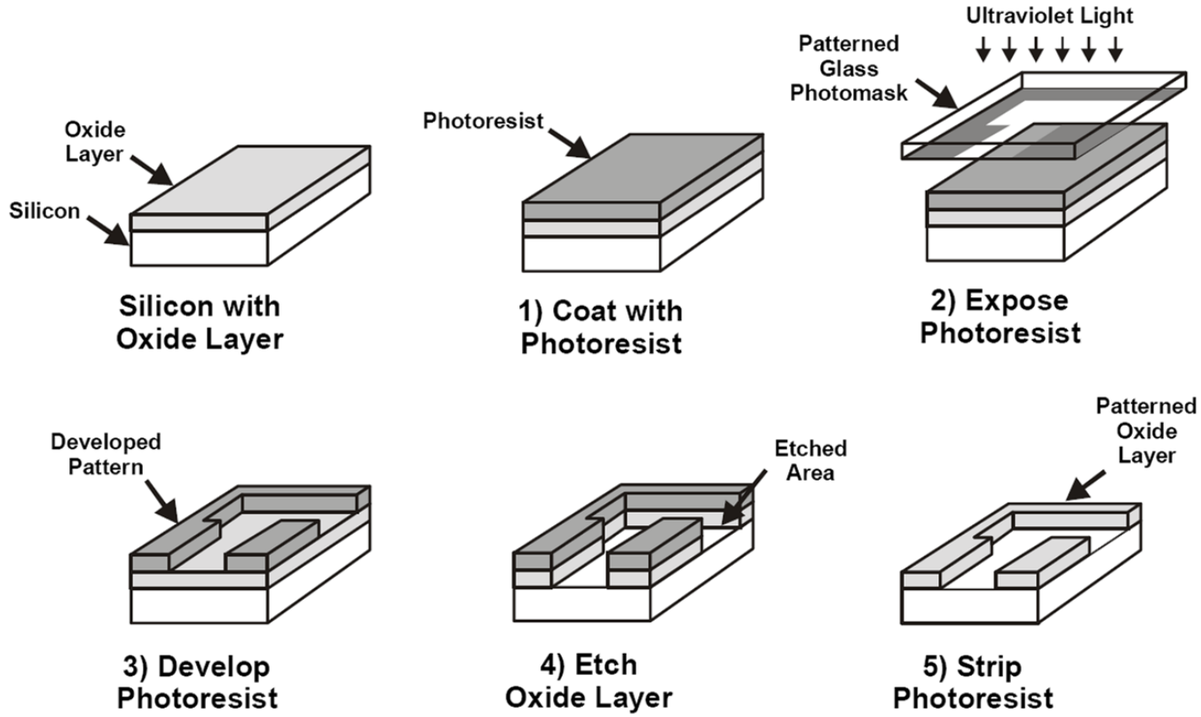
- **Parameters**
  - 10^{17} devices
  - 100,000 logical layers
    - Each 50 nm thick
    - ~1,000,000 physical layers
      - To provide wiring and isolation
  - *L* = 20 nm
    - 10x smaller than today

2065 mm³

3D Fabrication Challenges

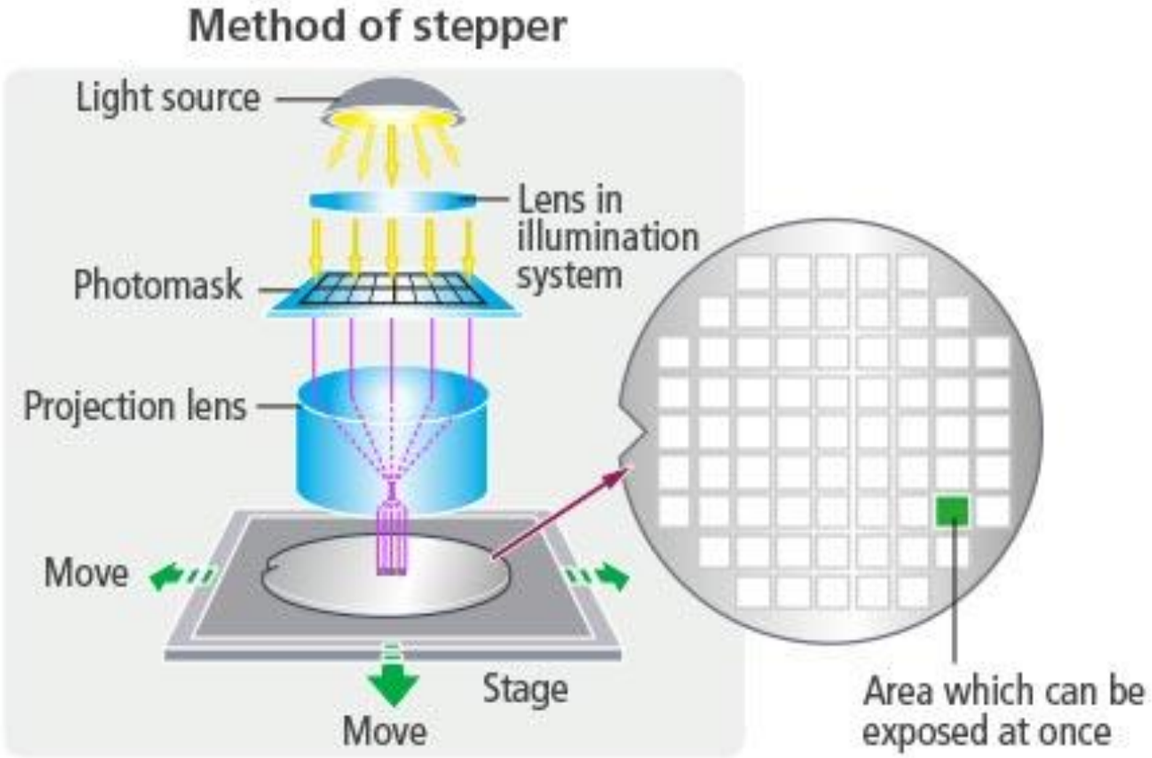
- **Yield**
  - How to avoid or tolerate flaws
- **Cost**
  - High cost of lithography
- **Power**
  - Keep power consumption within acceptable limits
  - Limited energy available
  - Limited ability to dissipate heat

Photolithography



- Pattern entire chip in one step
- Modern chips require ~60 lithography steps
- Fabricate *N* transistor system with O(1) steps

Fabrication Costs

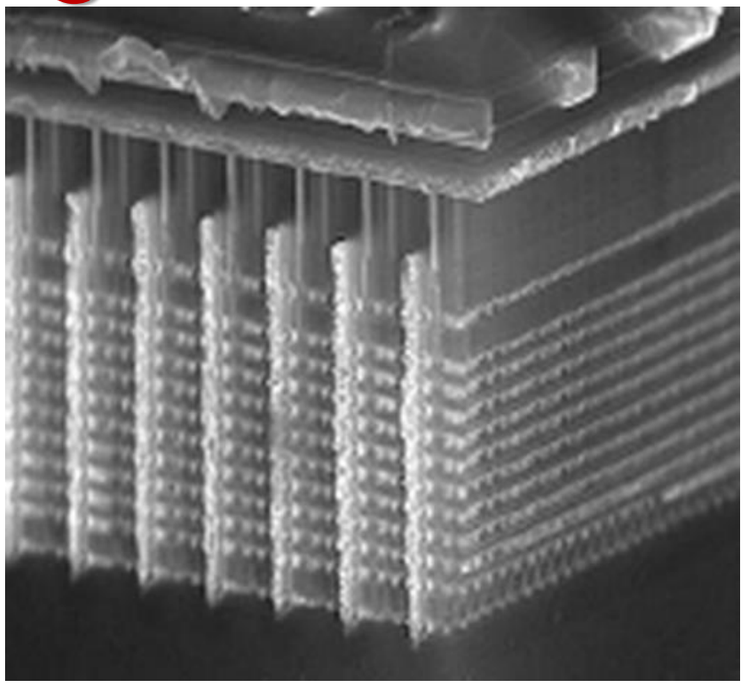


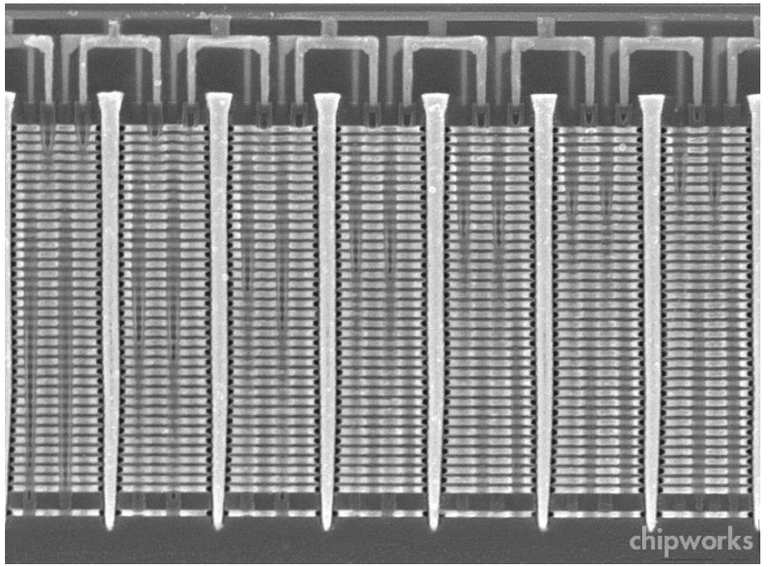
- **Stepper**
  - Most expensive equipment in fabrication facility
  - Rate limiting process step
    - 18s / wafer
  - Expose 858 mm^2 per step
    - 1.2% of chip area

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

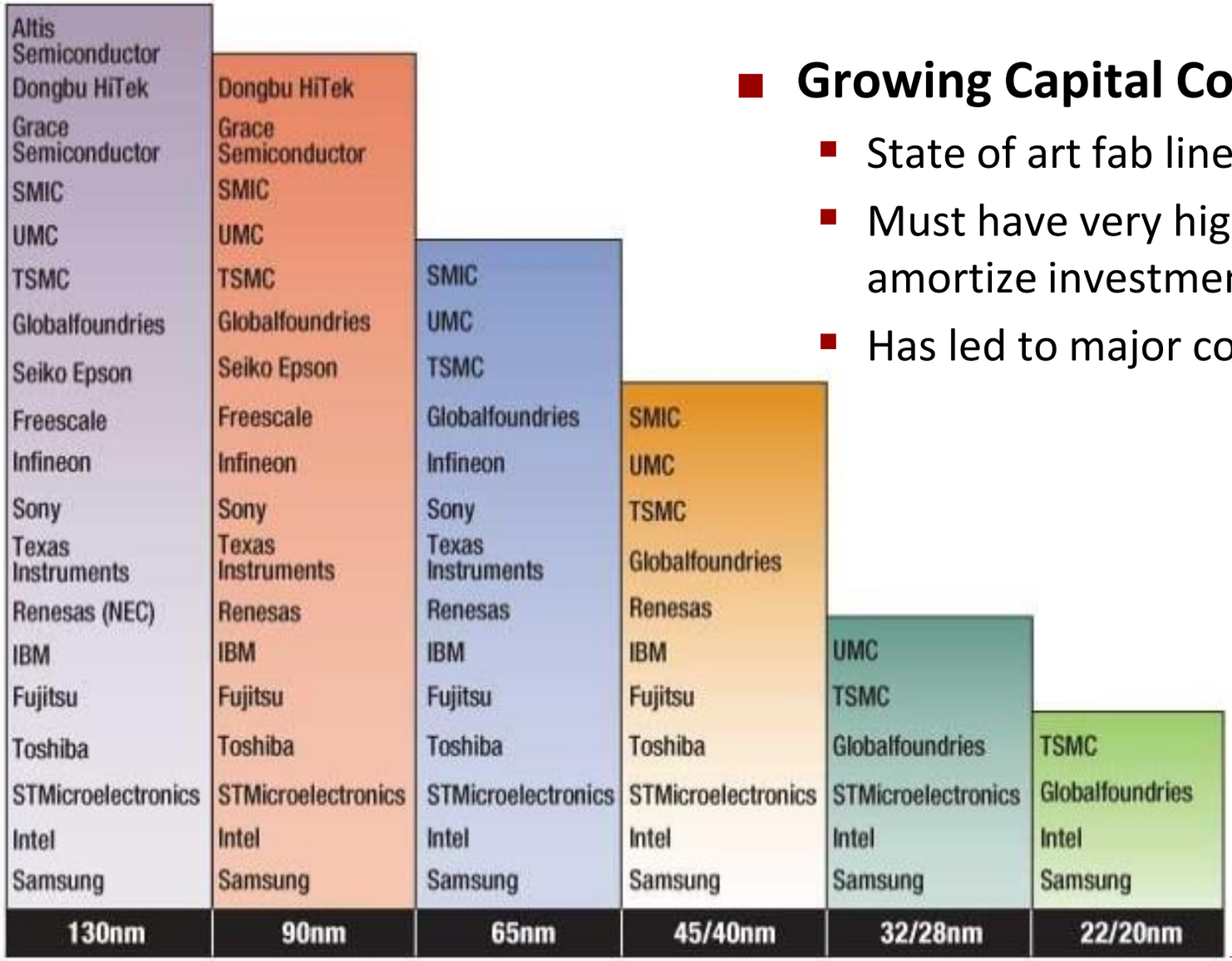
Meeting Power Constraints

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

Can we increase number of devices by 500,000x without increasing power requirement?

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

Challenges to Moore's Law: Economic



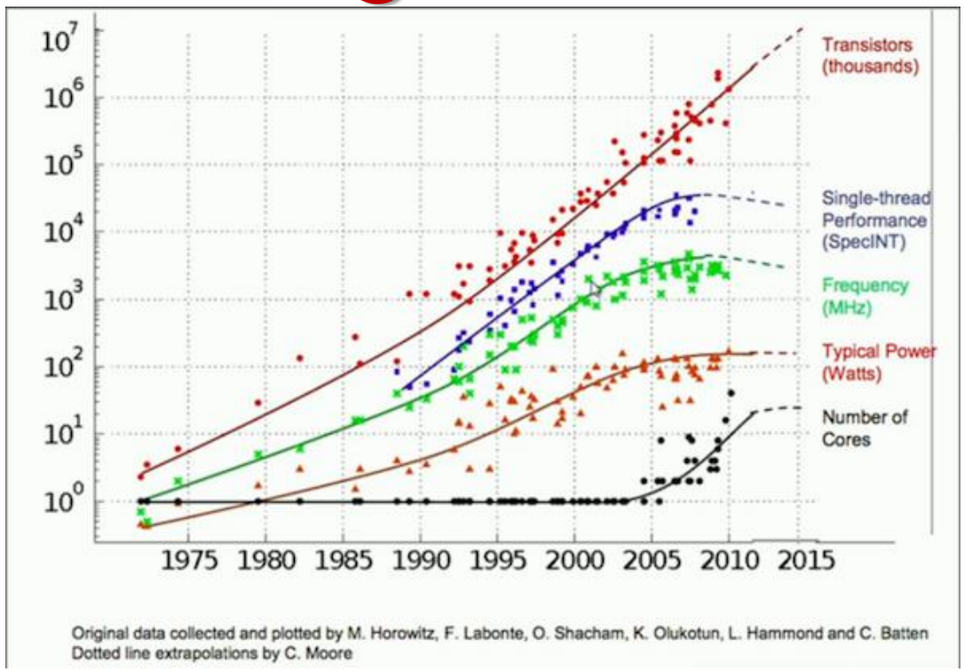
- **Growing Capital Costs**
  - State of art fab line ~$20B
  - Must have very high volumes to amortize investment
  - Has led to major consolidations

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?

18-600 Foundations of Computer Systems

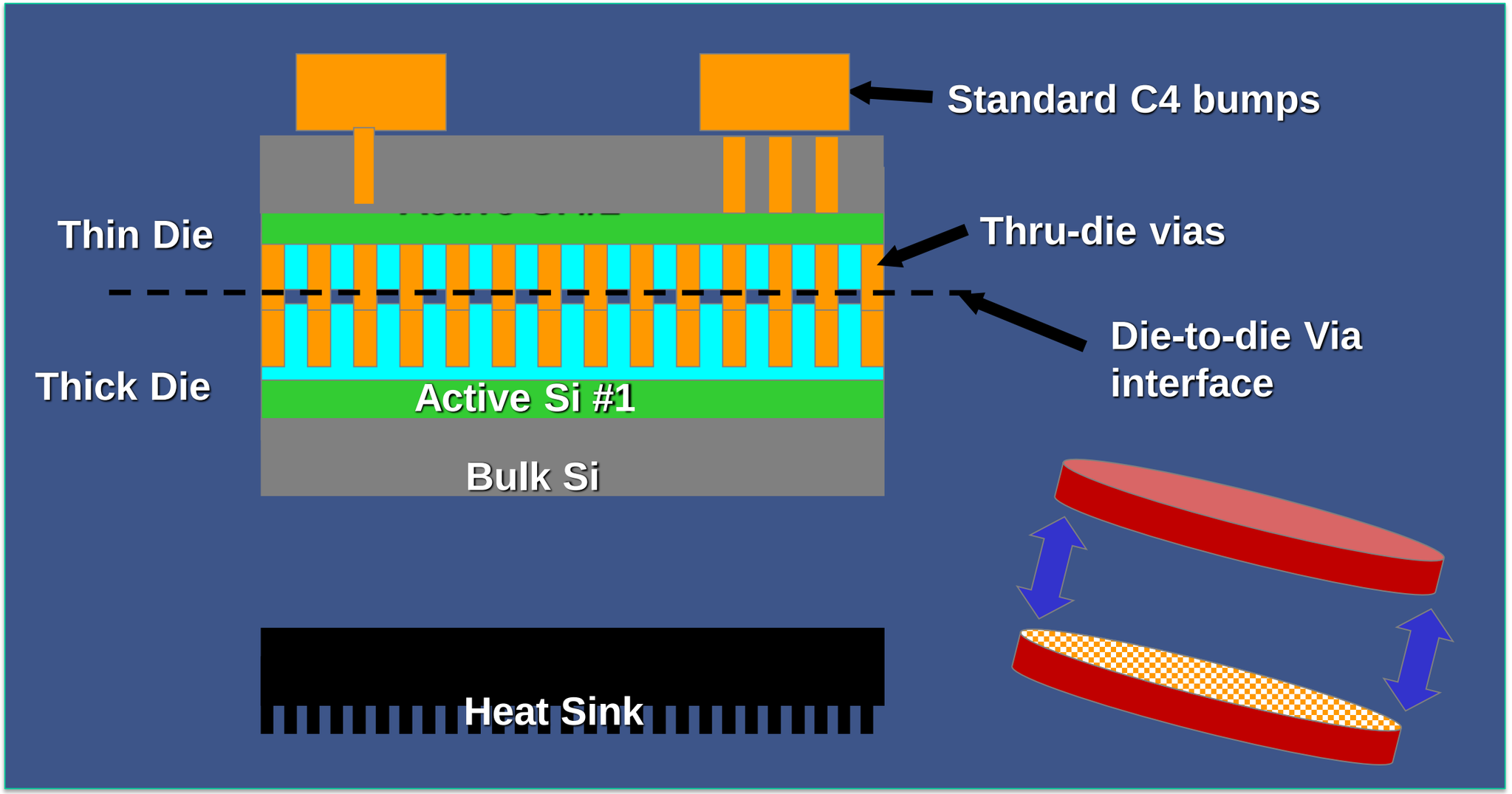
"Computing Systems Mega-Trends 2015-2025"

John P. Shen
December 7, 2016

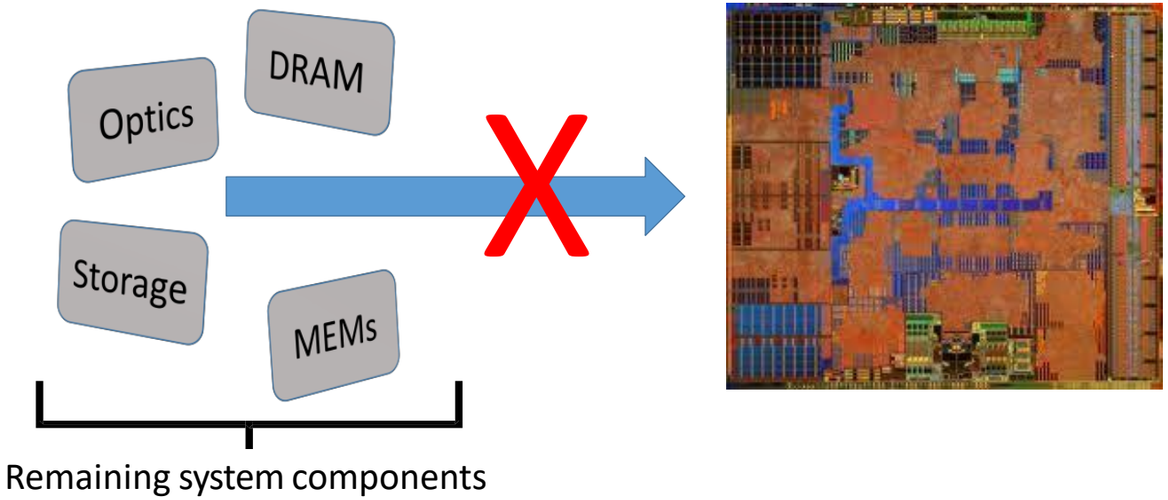
- **Silicon Technology**
- **Mobile Devices**
- **Software Development**
- **Cloud Infrastructure**

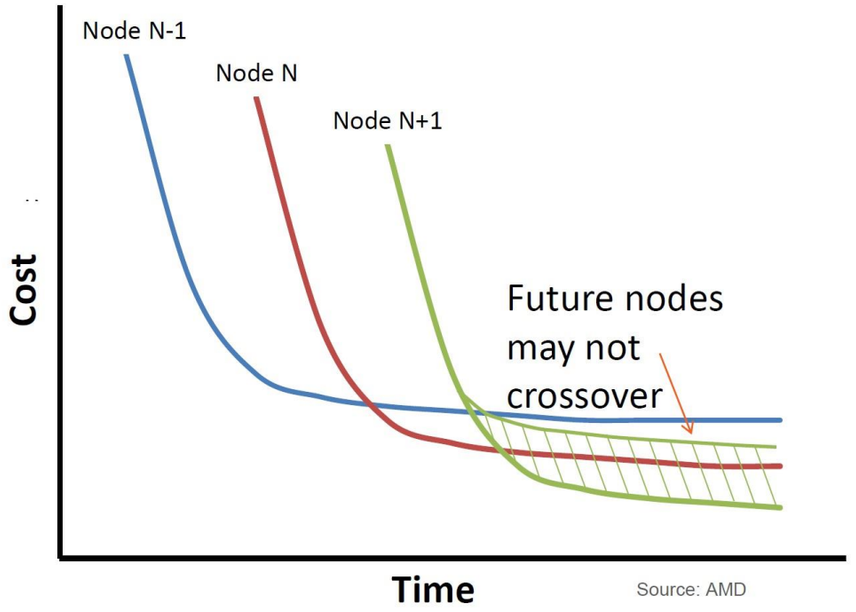
Electrical & Computer ENGINEERING

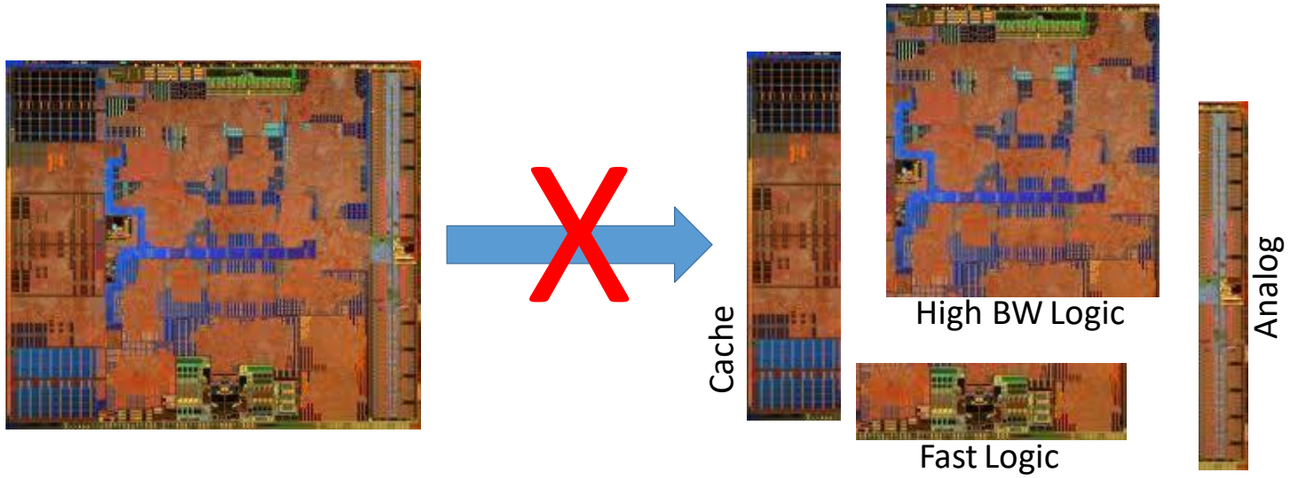
Silicon Technology: Potentials of 3D Die Stacking



Three Limitations to Moore's Law



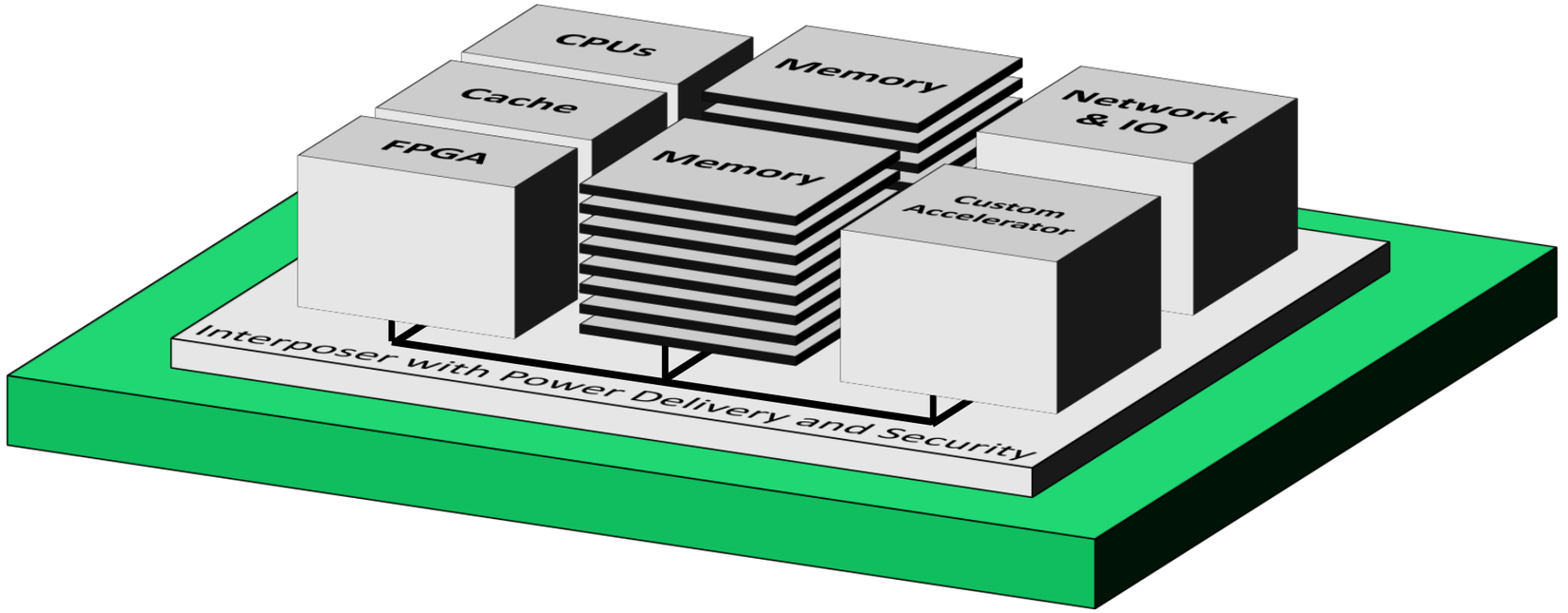




These limitations will make it very challenging to continue integrating systems

[Bryan Black, 2015, AMD]

Strategic Vision (Stacked System)



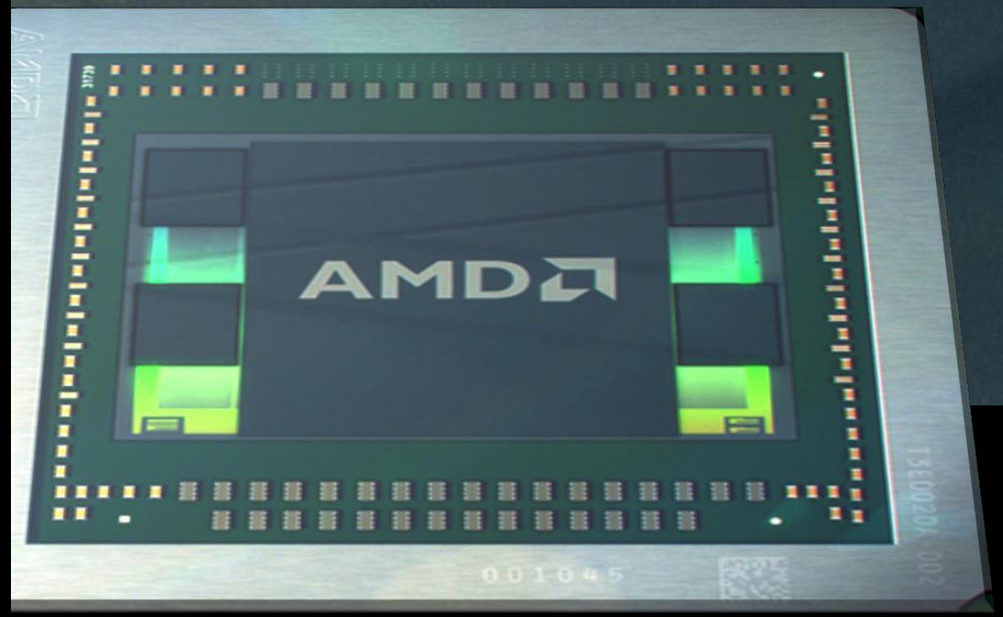
- The Stacked System model integrates dies from disparate technologies using a combination of 2.5D and 3D technology
- This construction model enables:
  - Disparate die integration to improve form factors and reduce system overheads
  - Die splitting to reduce process node complexity and cost
- Results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

[Bryan Black, 2015, AMD]

The Road to

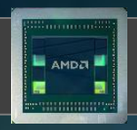
“FIJI”

Featuring Die Stacking and HBM Technology



[Bryan Black, 2015, AMD]

"Fiji" Chip



DETAILED LOOK

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

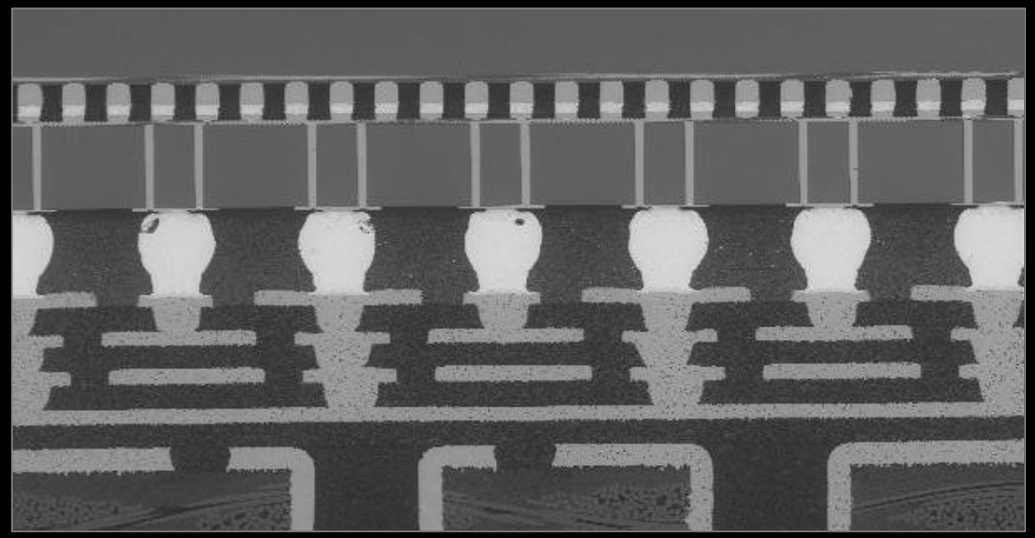
- First high-volume interposer
- First TSVs and µBumps in the graphics industry
- Most discrete dies in a single package at 22
- Total 1011 sq. mm.

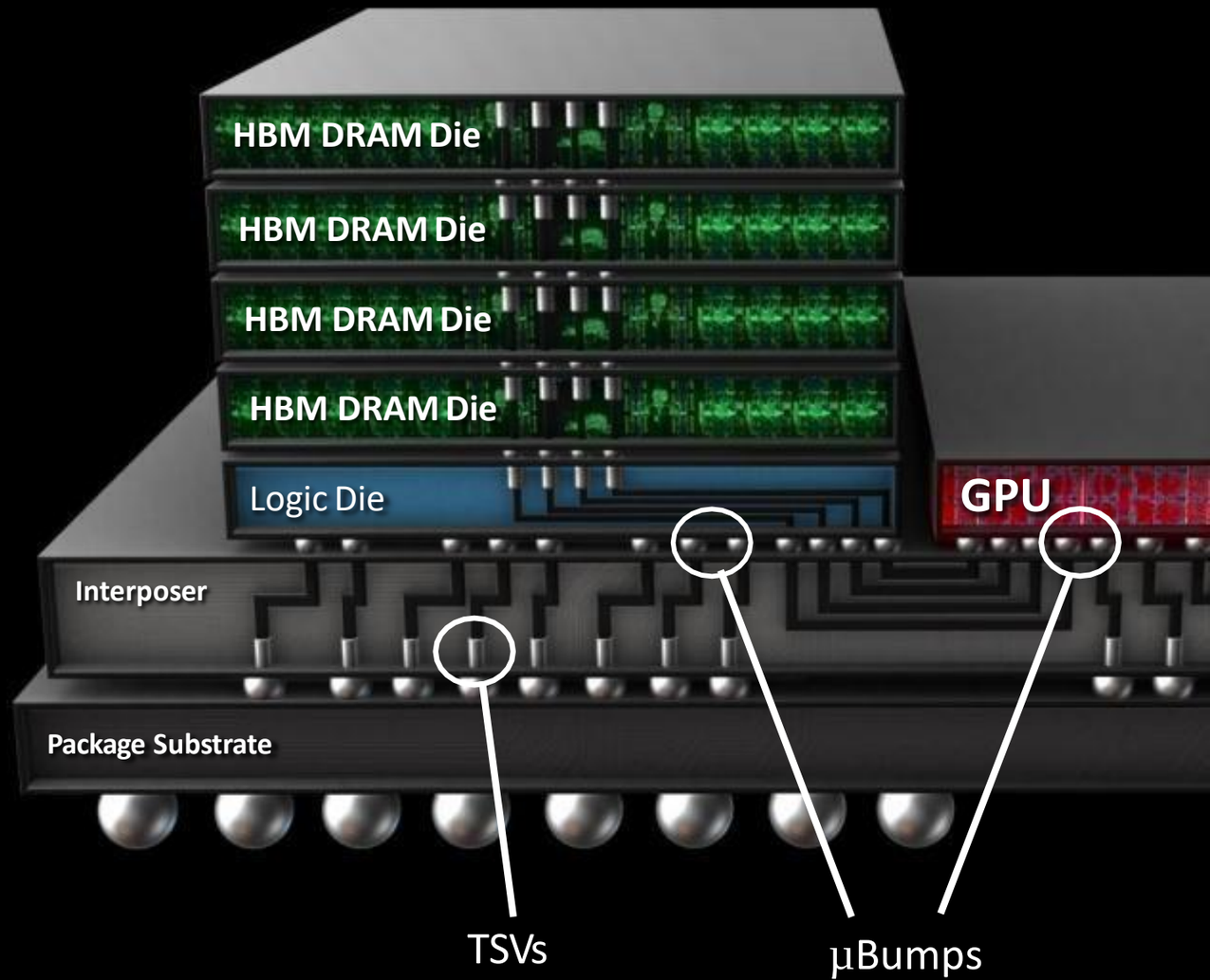
- Graphics Core Next Architecture
- 64 Compute Units[14]
- 4096 Stream Processors
- 596 sq. mm. Engine

[Bryan Black, 2015, AMD]

DIE STACKING TECHNOLOGY

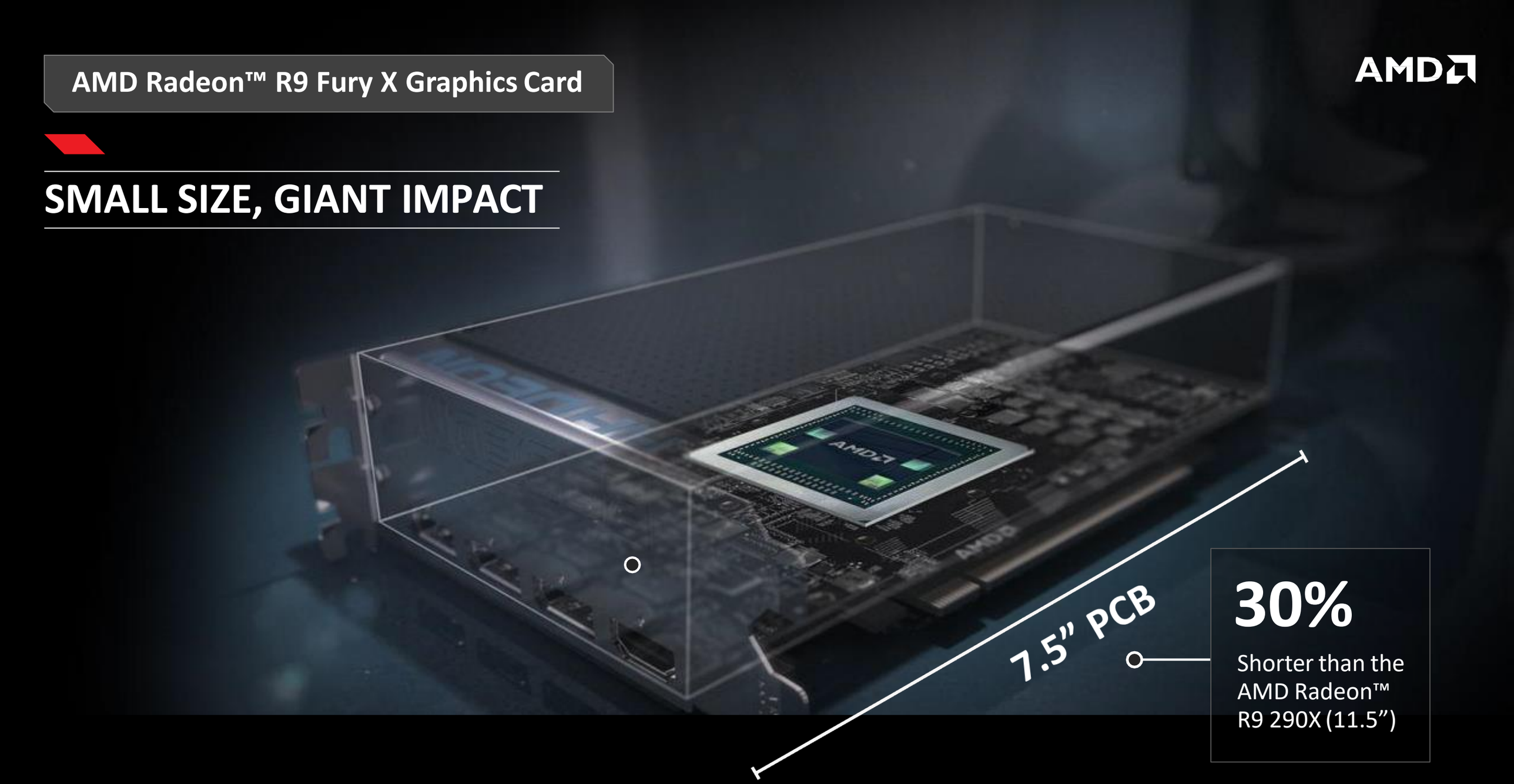
- Die stacking facilitates the integration of discrete dies
- 8.5 years of development by AMD and its technology partners





[Bryan Black, 2015, AMD]

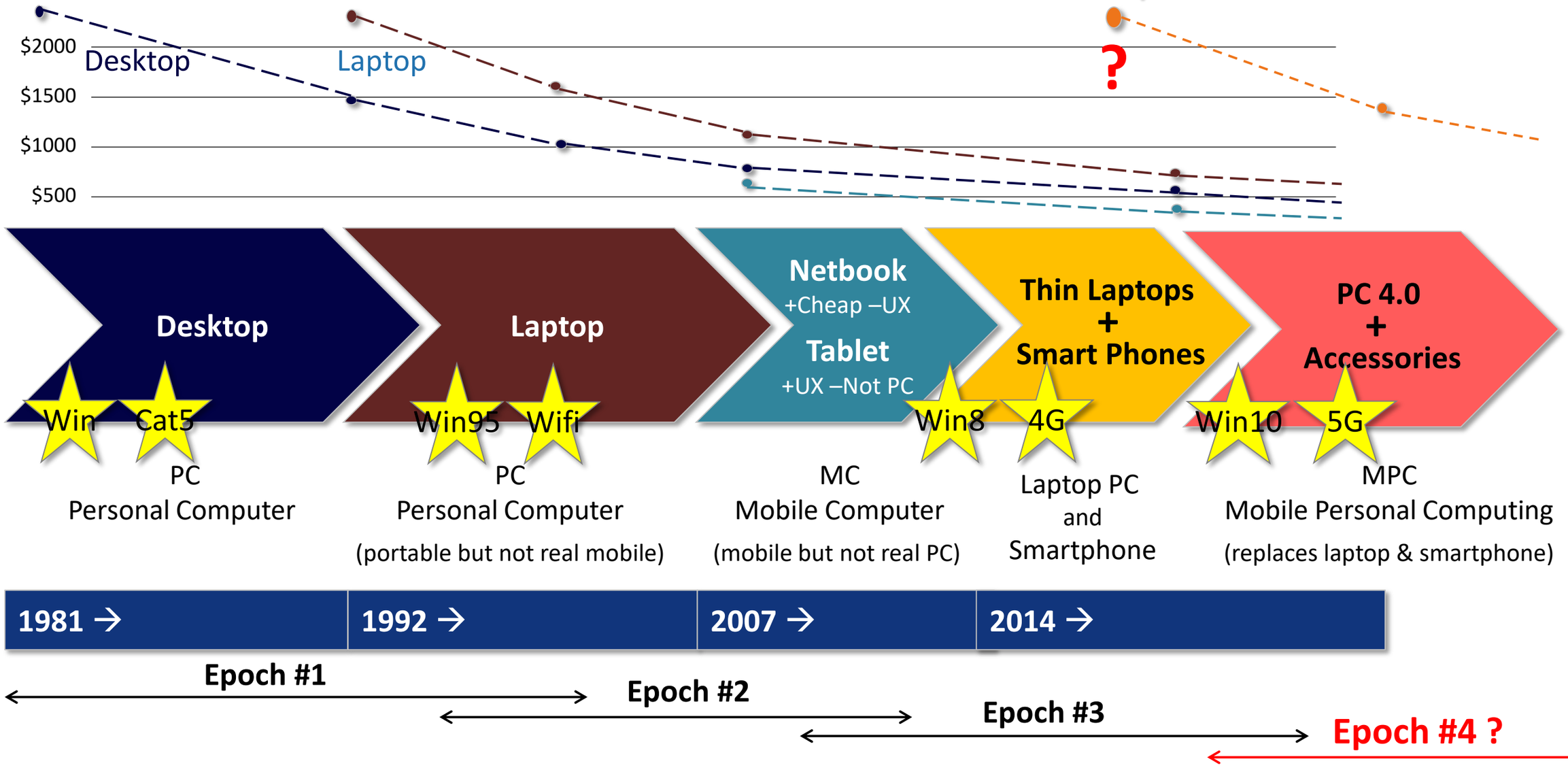
SMALL SIZE, GIANT IMPACT



[Bryan Black, 2015, AMD]

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

Mobile Devices: What's the Next Epoch?



Seamless Roaming Experience Across 5+ Screens

	PC	smartphone	tablet	living room	dashboard
Apple	Mac computers	iPhone	iPad	Apple TV	CarPlay
Google	Chrome browser	Android	Android tablets	Google TV, Chromecast	Open Automotive Alliance
Microsoft	Windows, Office	Windows Phone	Windows 8	Xbox	Windows in the Car
amazon	Online store	Soon to come?	Kindle Fire	Fire TV	?

From Two Screens to Multiple Screens

Computers

Tablets + E-readers

(Smart)phones

Televisions & Game Consoles

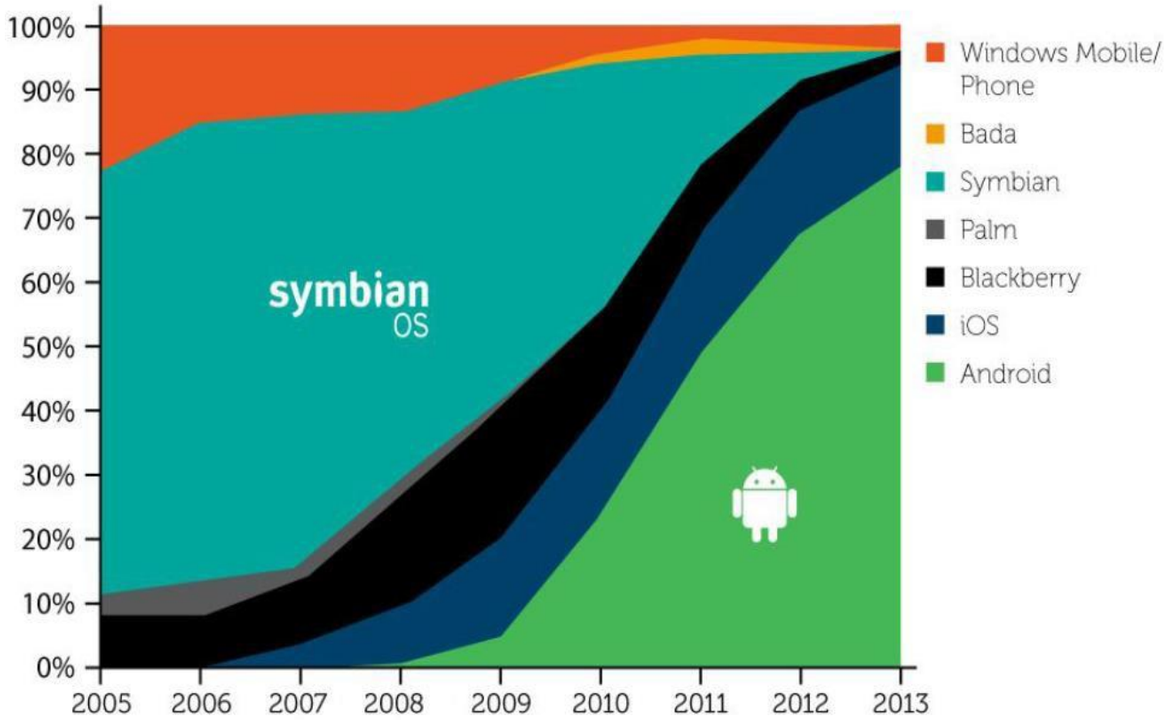
Car displays

Gadgets & Wearables

Software Development: Dominant Mobile Platforms

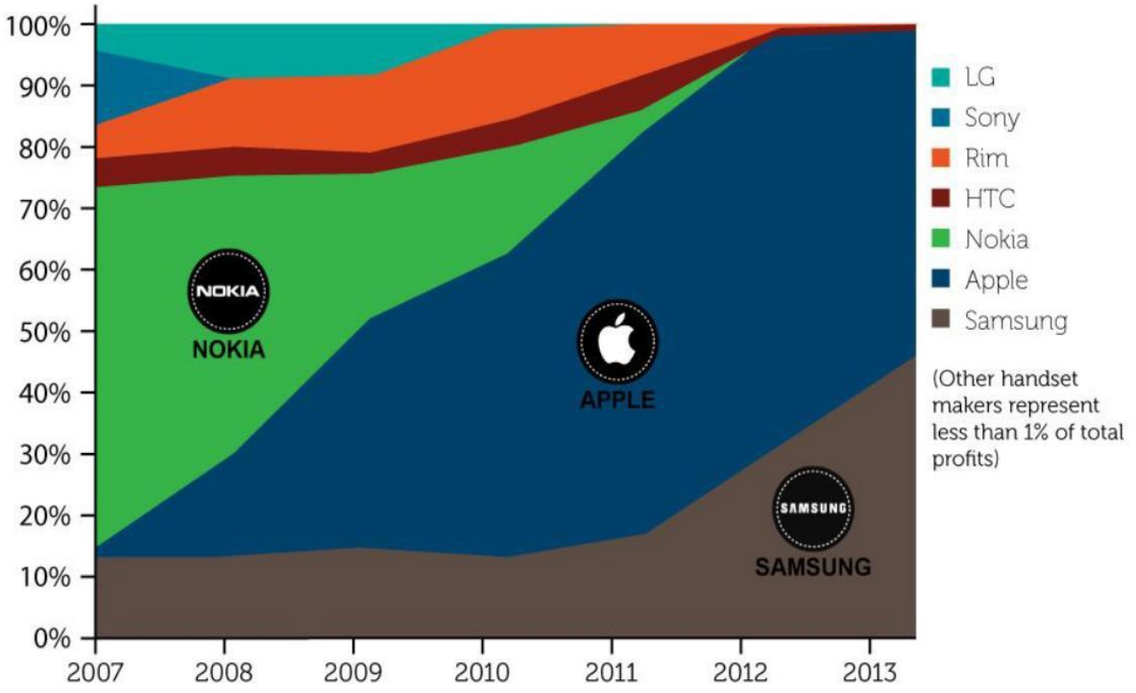
The Android & iOS Duopoly Continues to Strengthen

By Unit Volume



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

By Device Profits

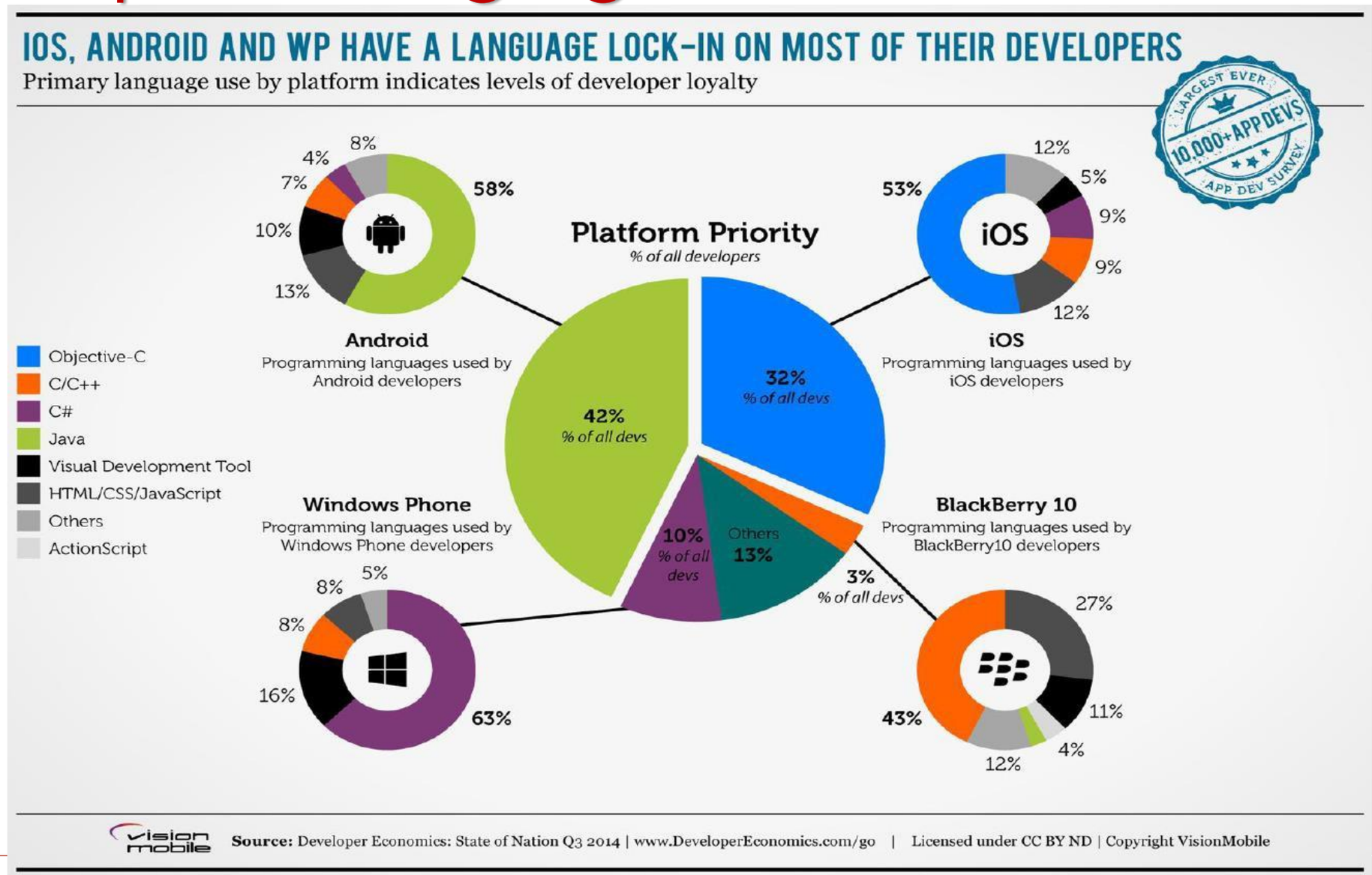


Data: McKinsey, Asymco, Canaccord, VisionMobile estimates

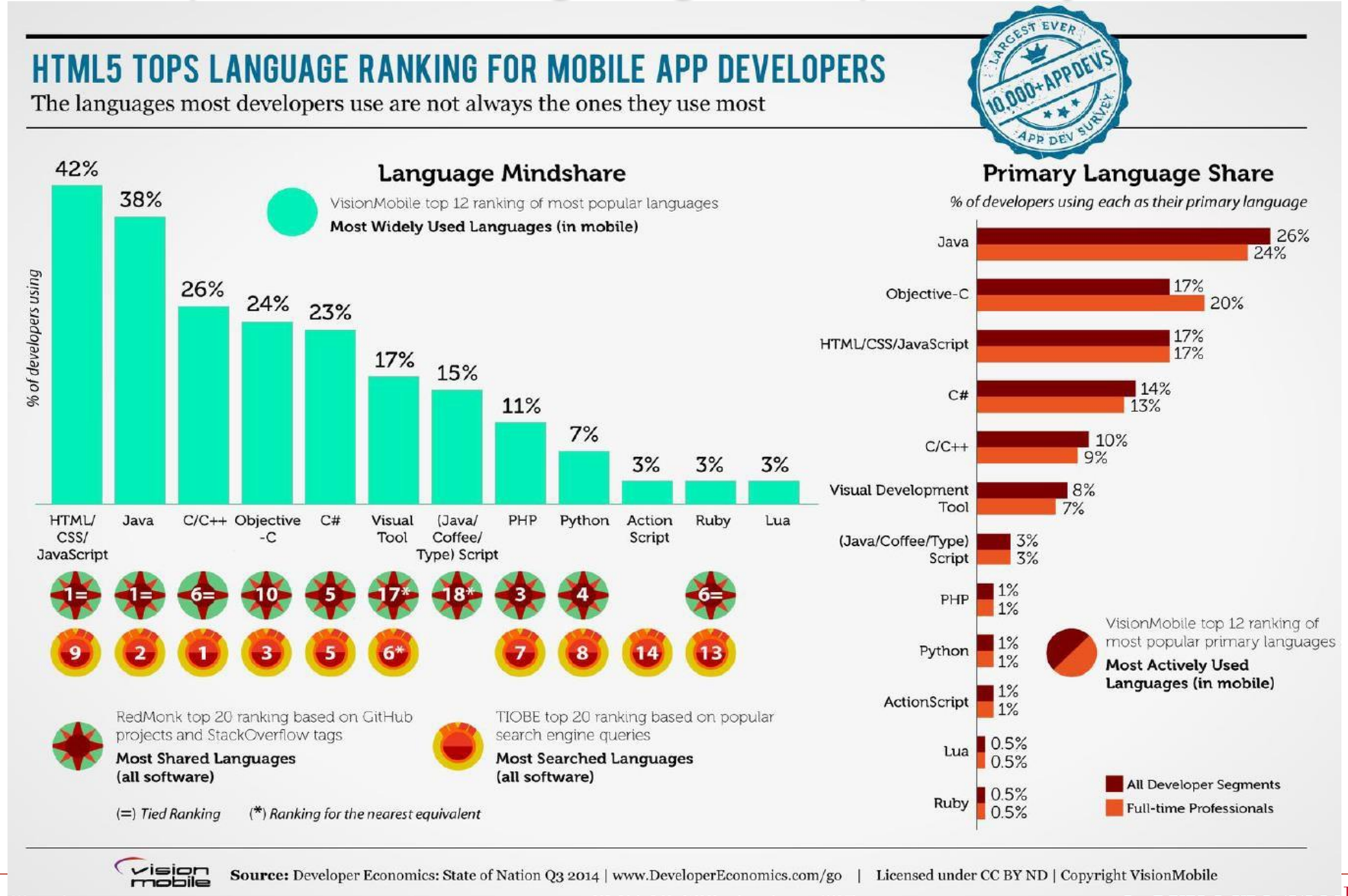
Dominant Mobile Platforms 2014

Most Popular Platform by Country

Development Languages and Tools



Development Language Popularity



Disruptive Trends in Web-Based SW Development

1) WebGL

Gaming

- 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

Communications

- 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)

Software development

- Developer-friendly, all-in-one cloud solutions that require minimal installation or maintenance

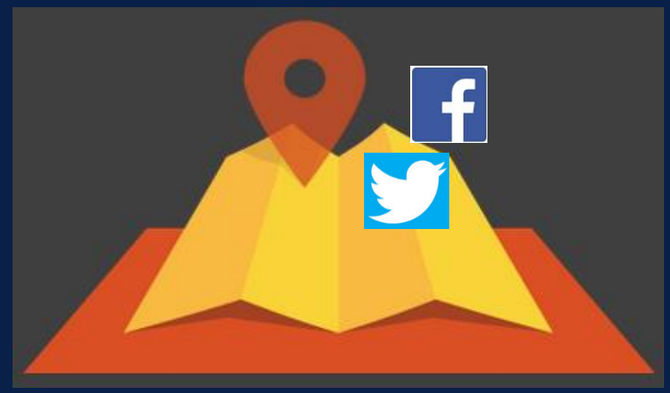
[Antero Taivalsaari, Nokia, 2014]

BaaS – Commonly Provided Features

Cloud Data Storage

Push Notifications

3rd Party Data Integration



User Management

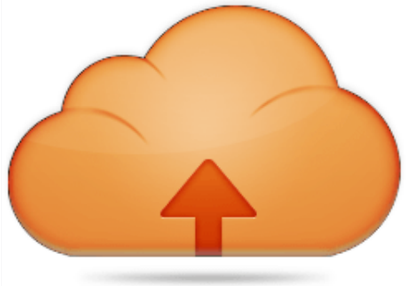
Cross-Platform Support

Versioning, analytics, etc.

[Antero Taivalsaari, Nokia, 2014]

Example Feature Set: Parse.com

Parse was recently acquired by Facebook

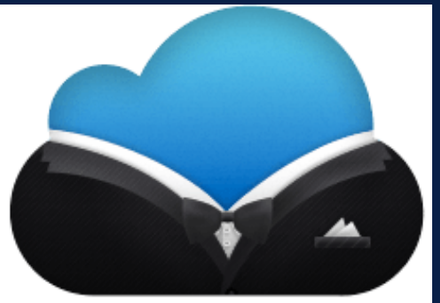


Parse Data

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.

One platform to rule them all

Parse Social

Make your app social. Instantly.

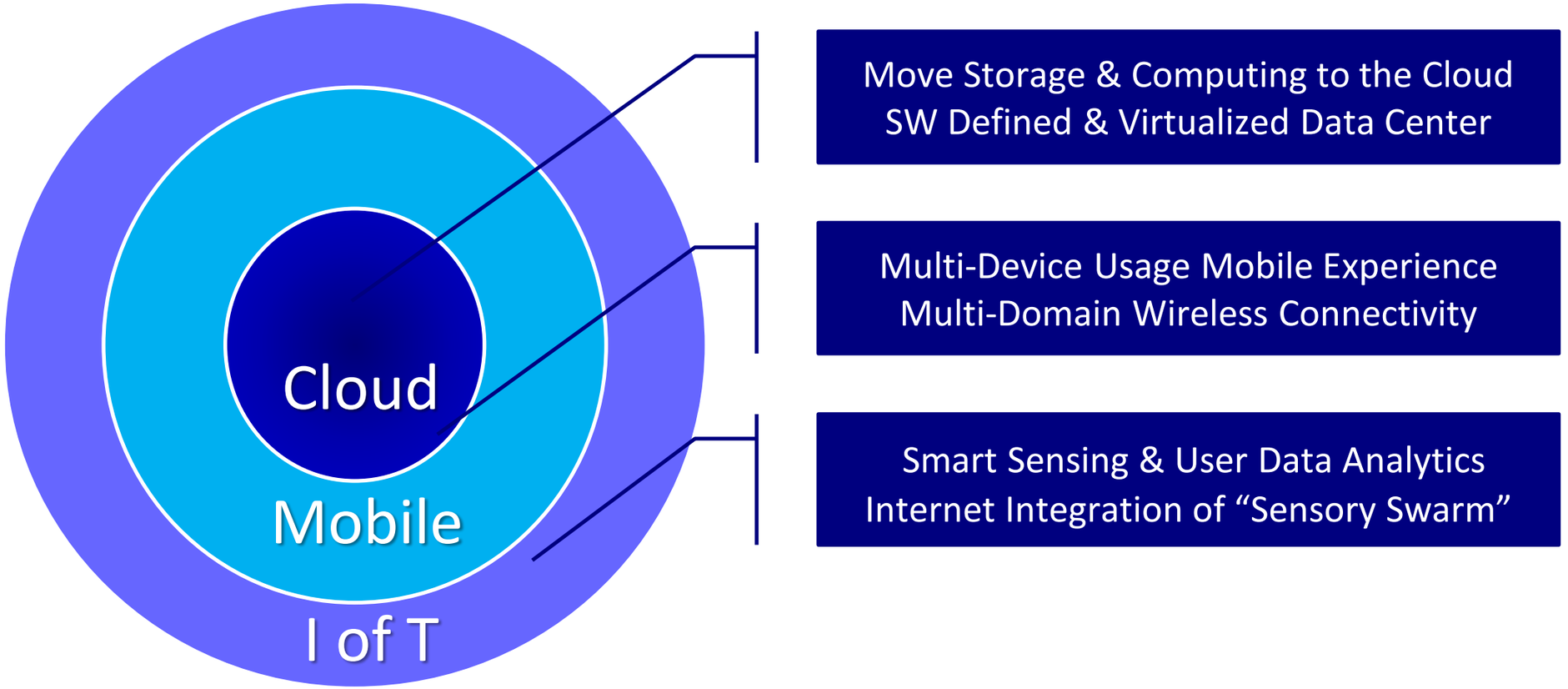
Cloud Code

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

[Antero Taivalsaari, Nokia, 2014]

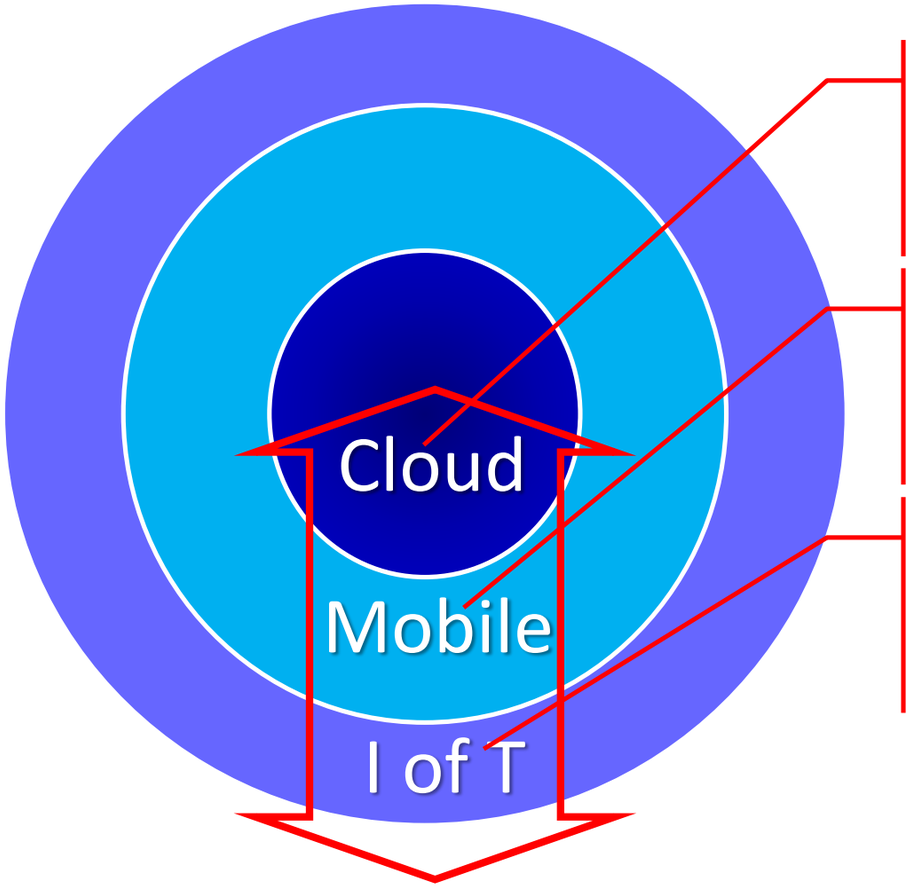
Cloud Infrastructure: Current Mega-Trends

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



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:

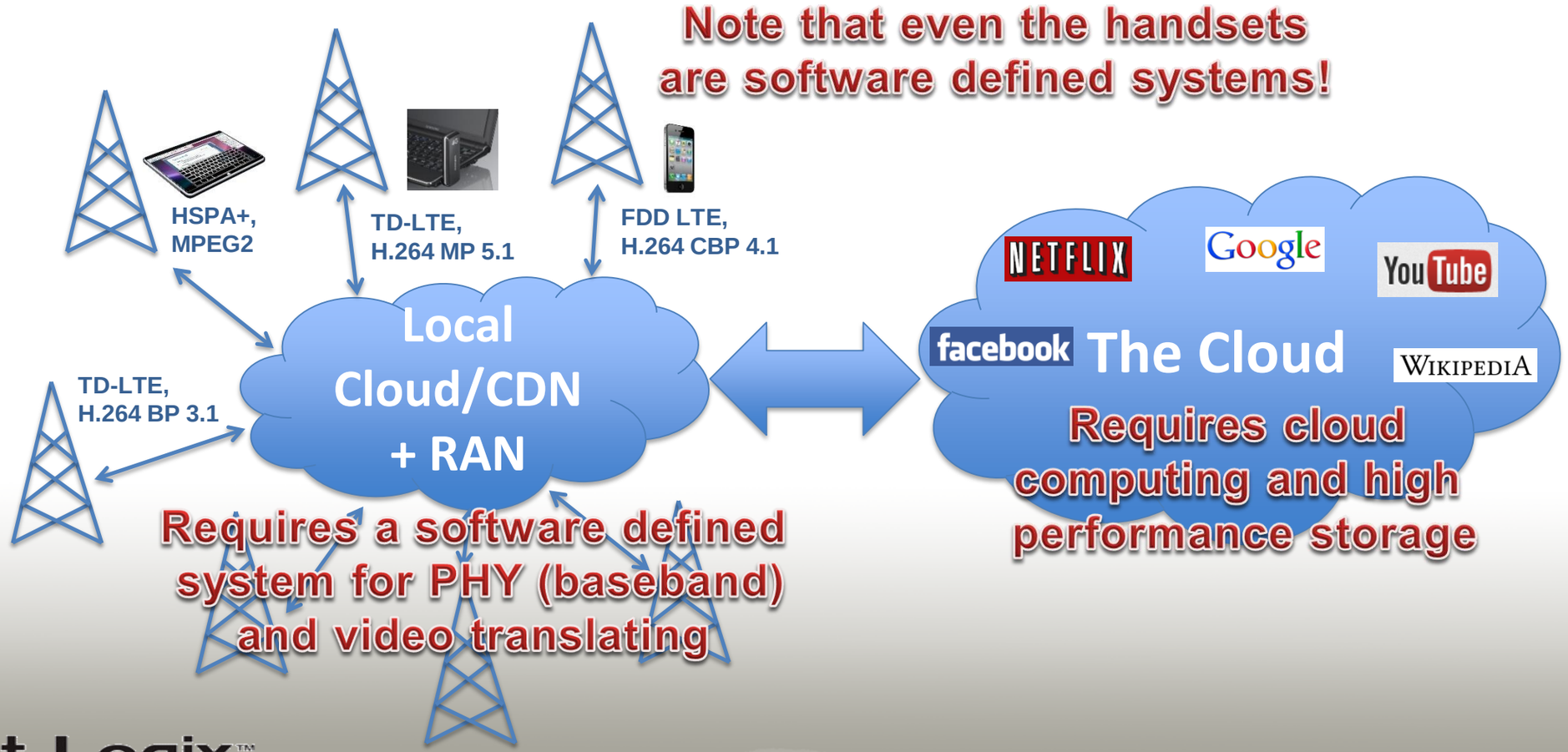
Seamless Cross-Device Cross-Domain UX
Unify both Broadband and Broadcast UX

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.**



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.