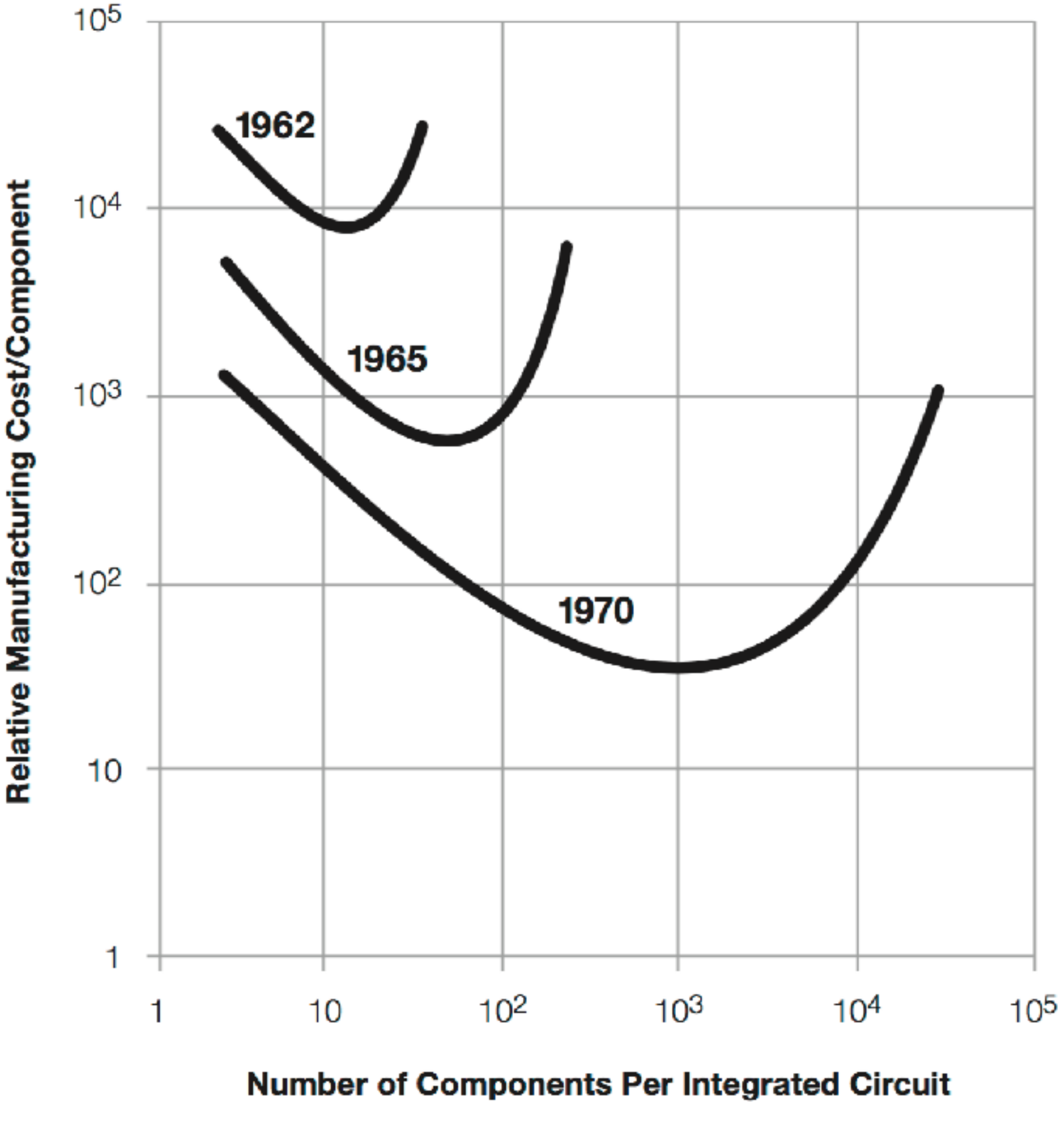
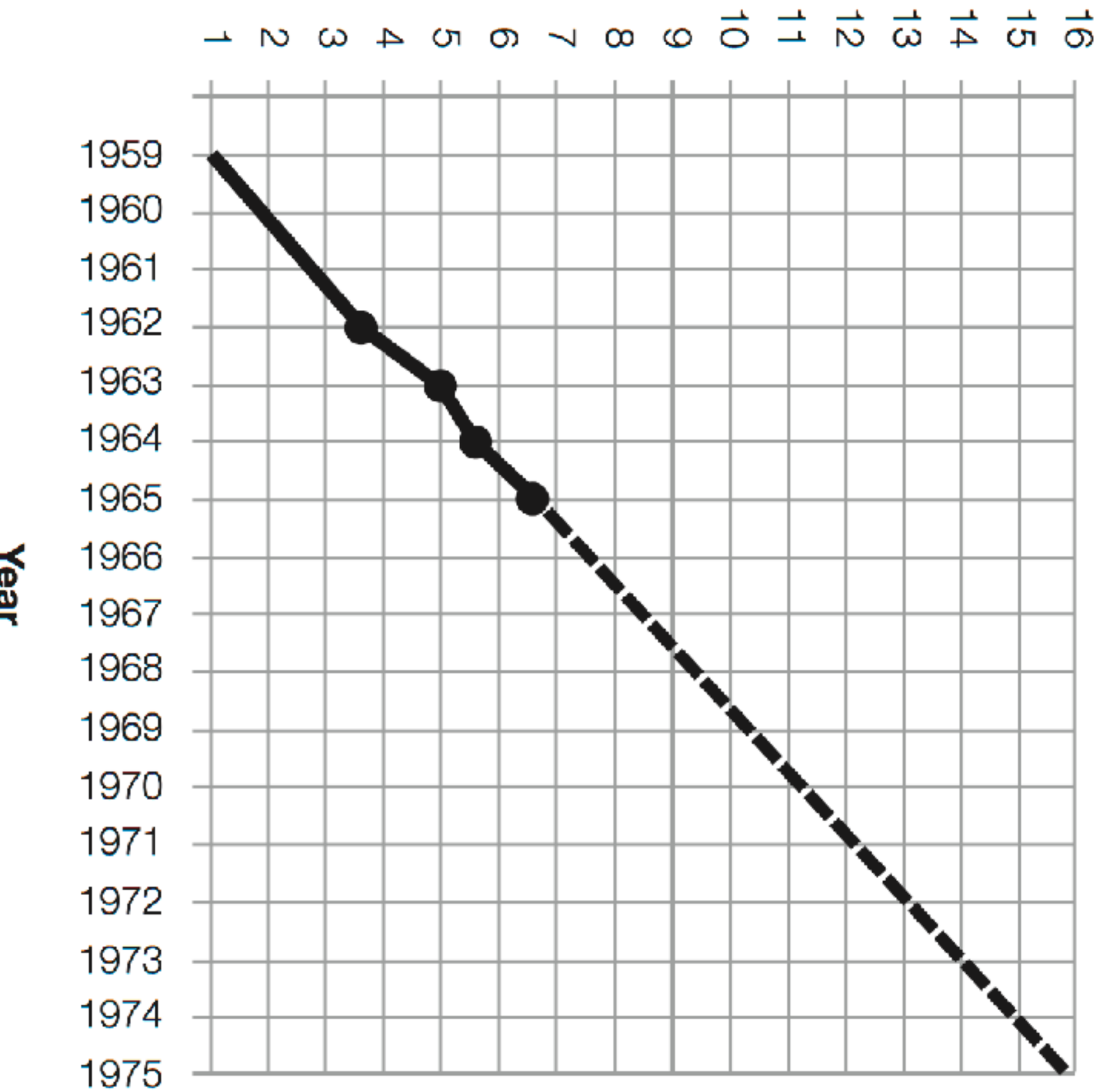
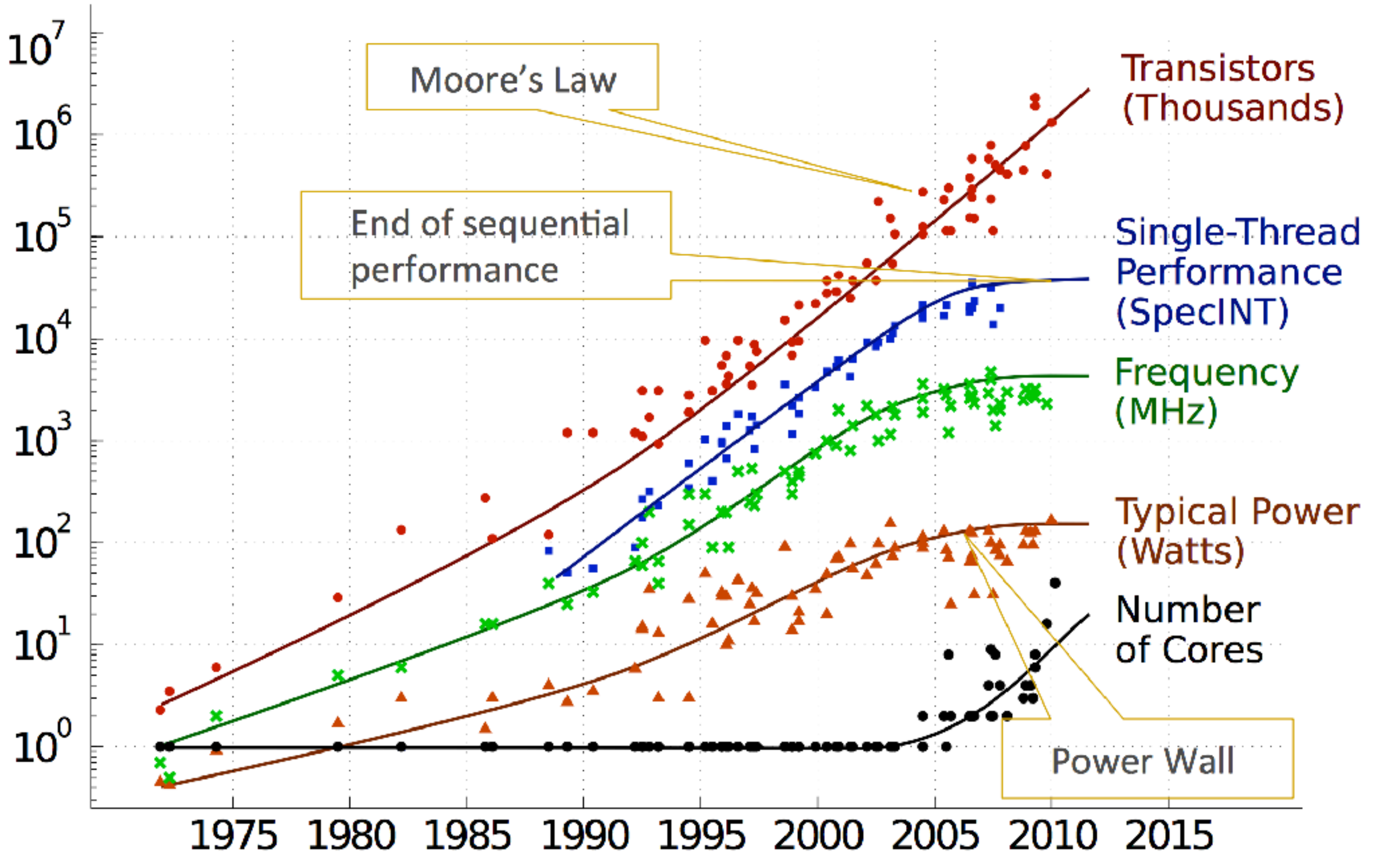
Dark silicon, Shiny objects, and the uncertain future

Moore’s Law

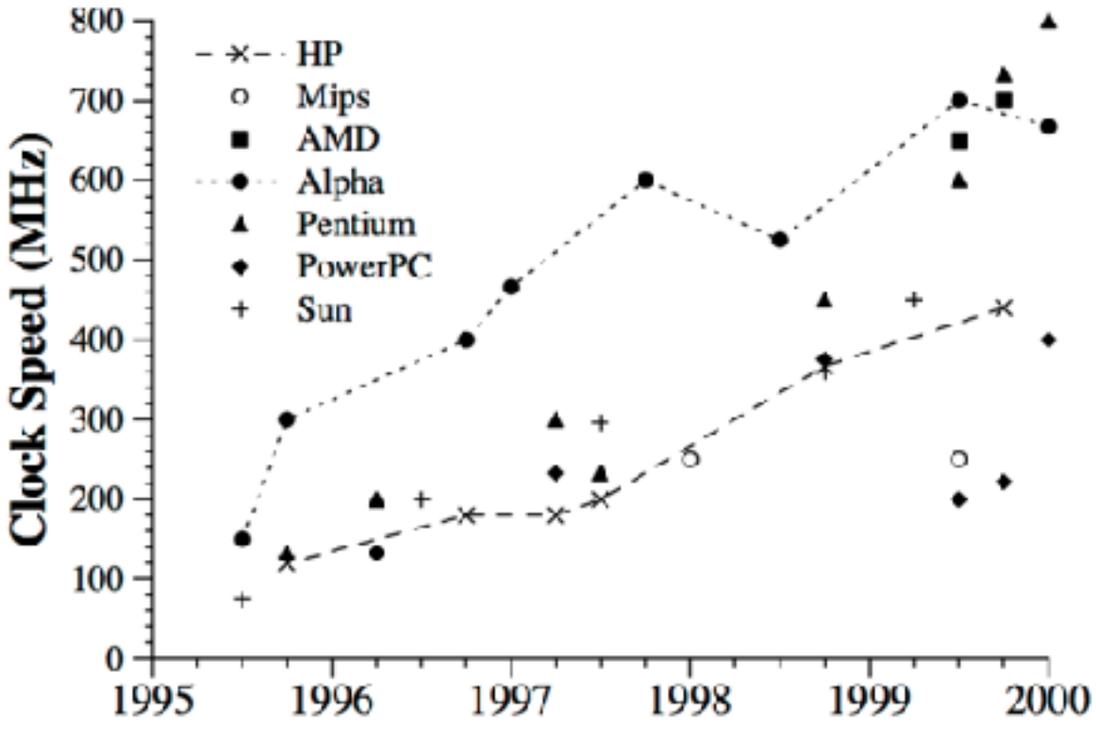

Relative Manufacturing Cost/Component
10^5
10^4
10^3
10^2
10
1
1962
1965
1970
1
10
10^2
10^3
10^4
10^5
Number of Components Per Integrated Circuit

Log^2 of the Number of Components Per Integrated Function





Data collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten



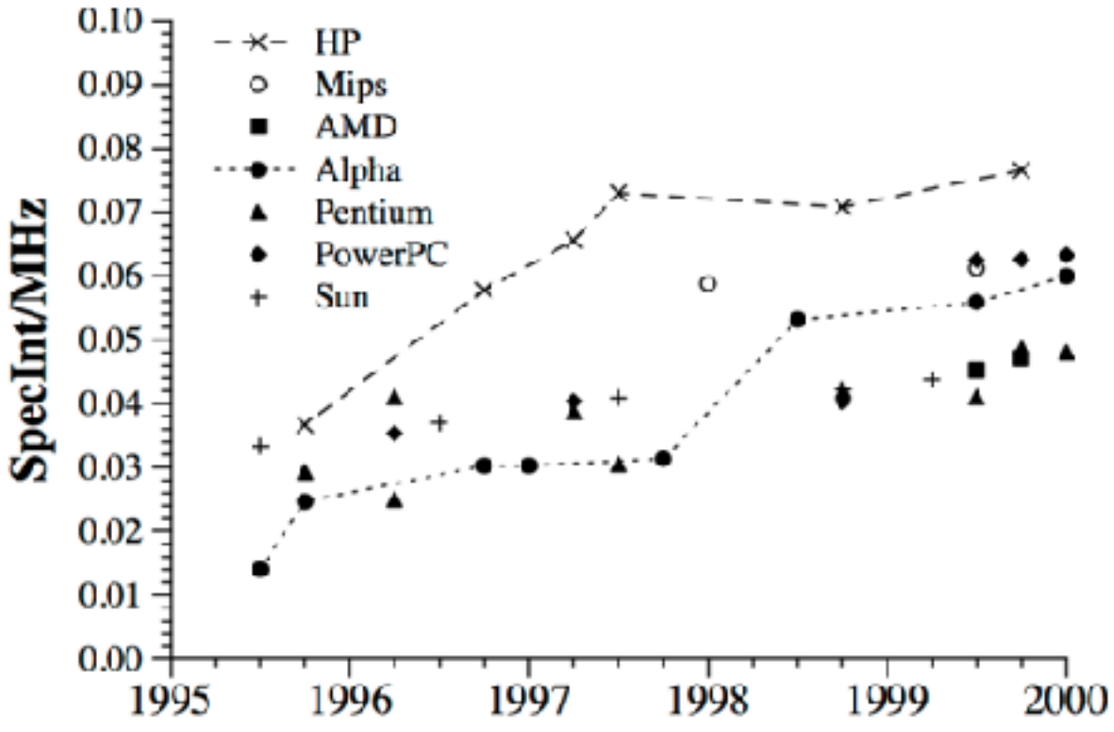
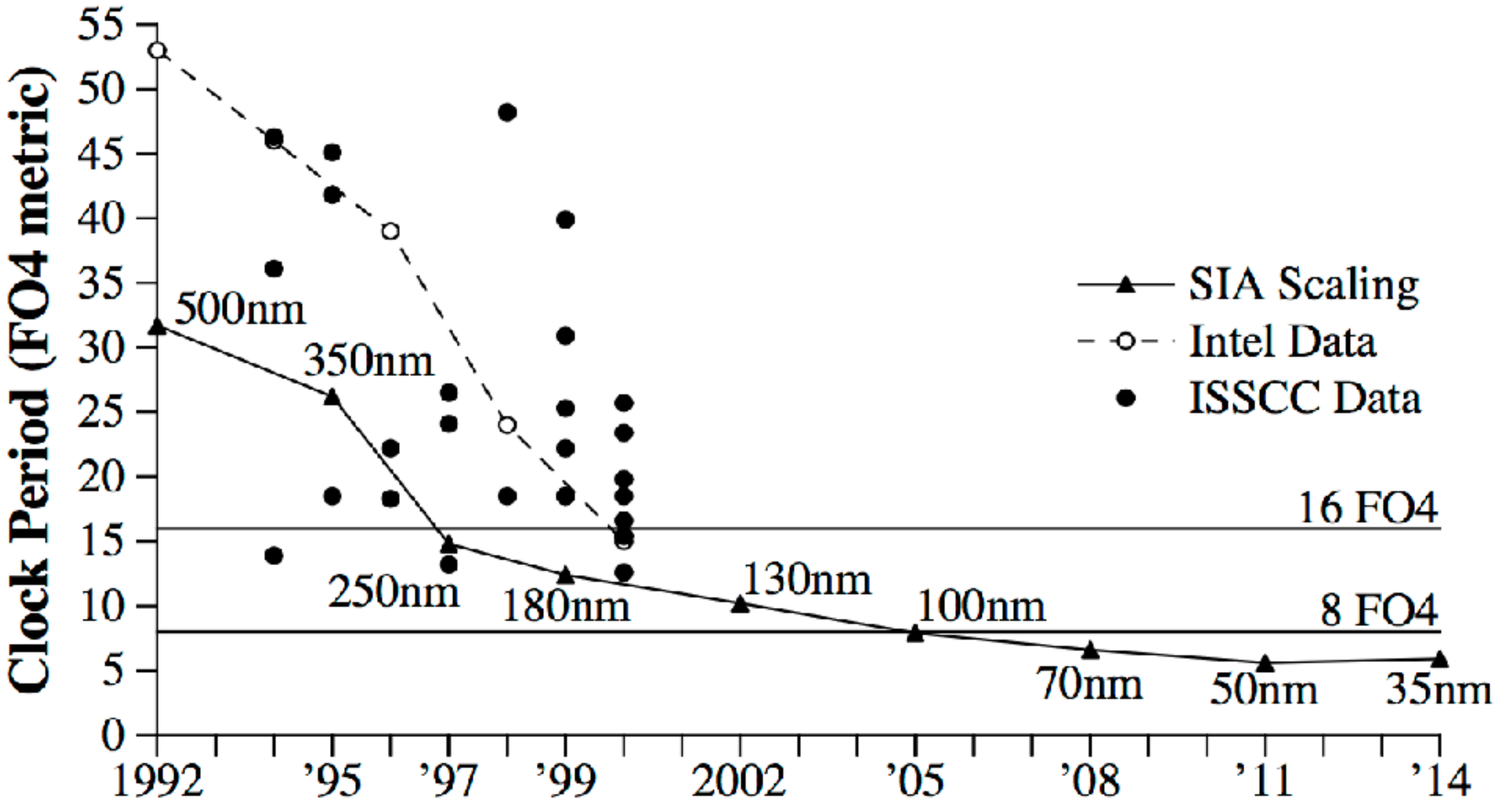
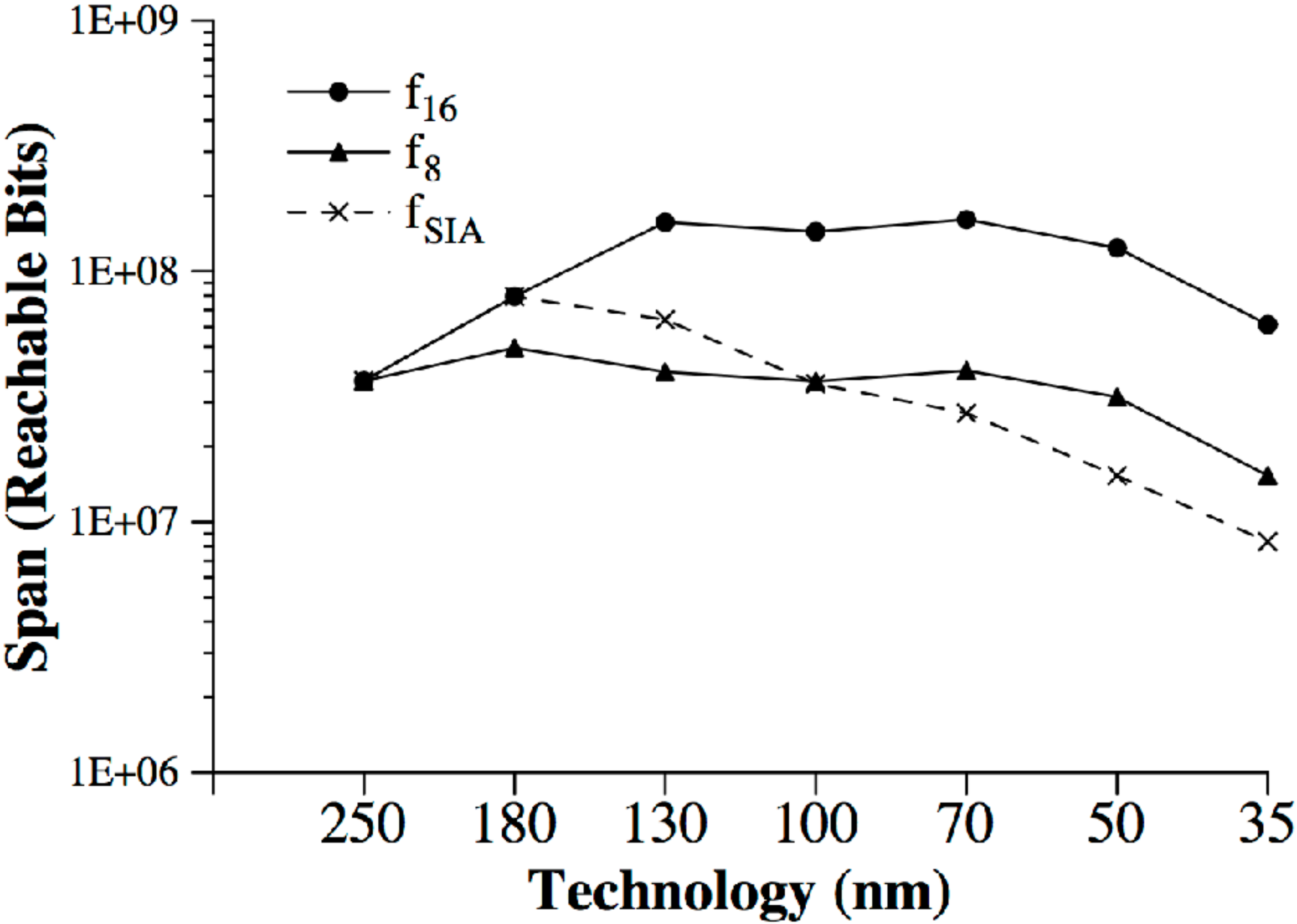
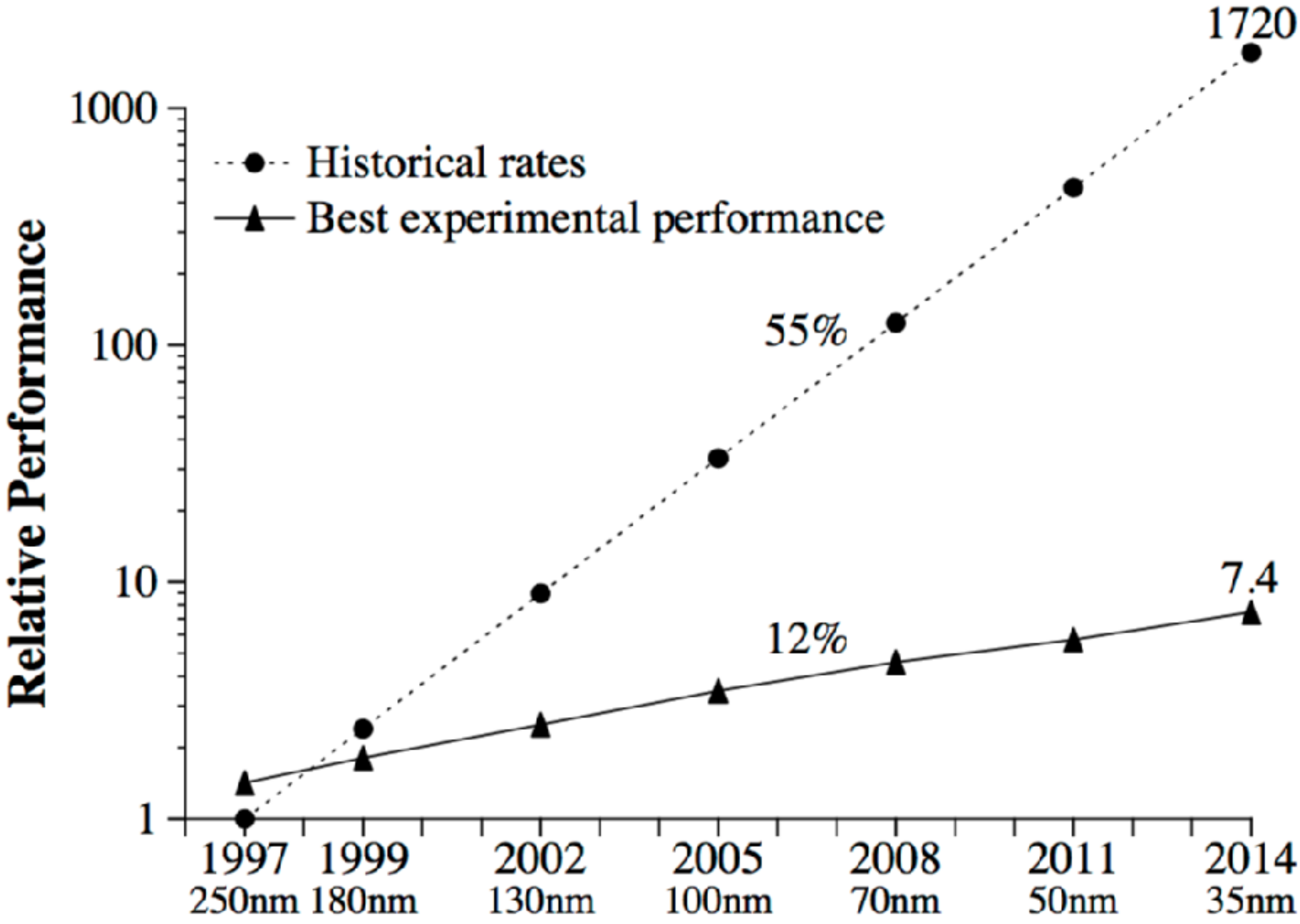
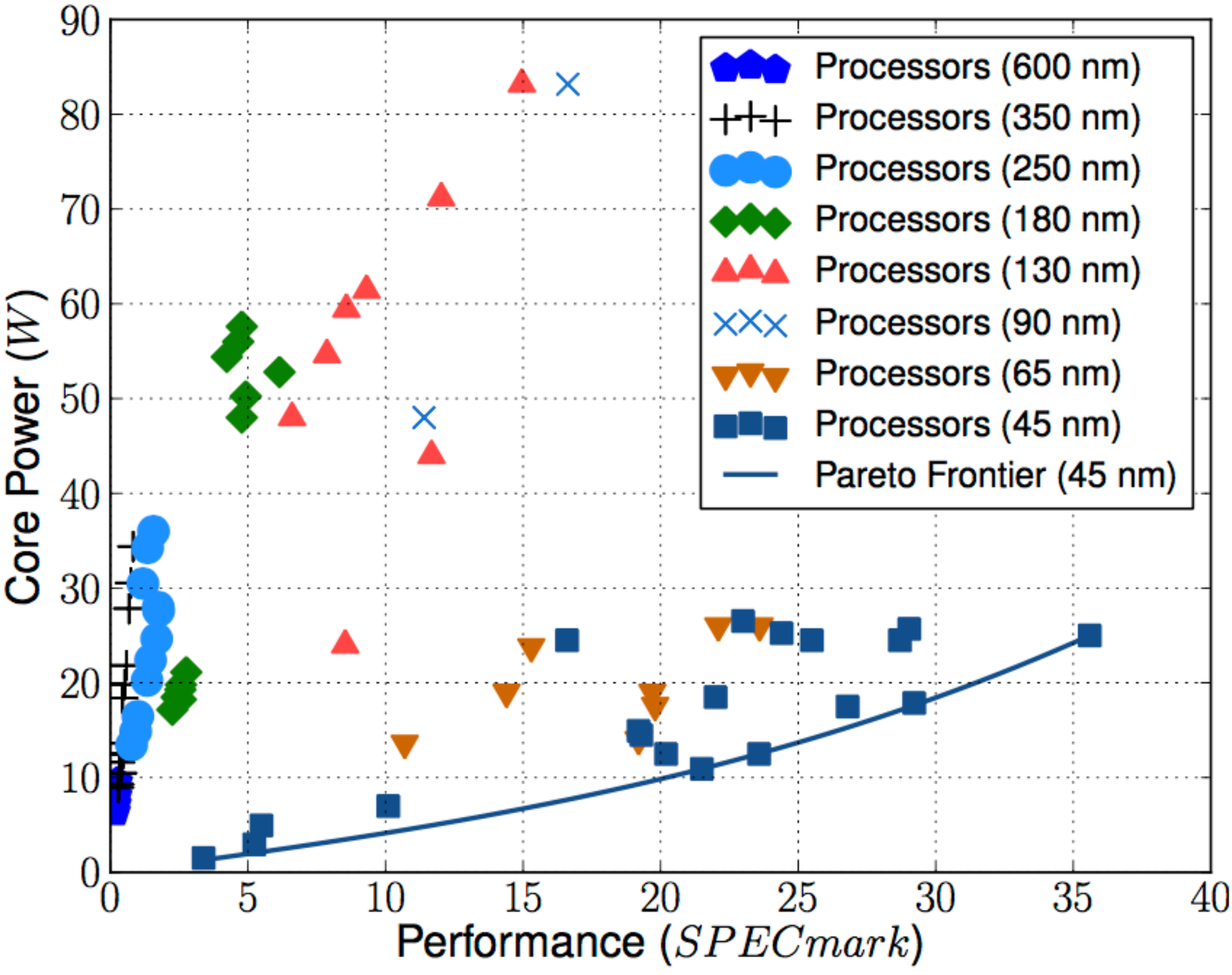


Figure 1: Processor clock rates and normalized processor performance (SpecInt/Clock rate), 1995-2000.

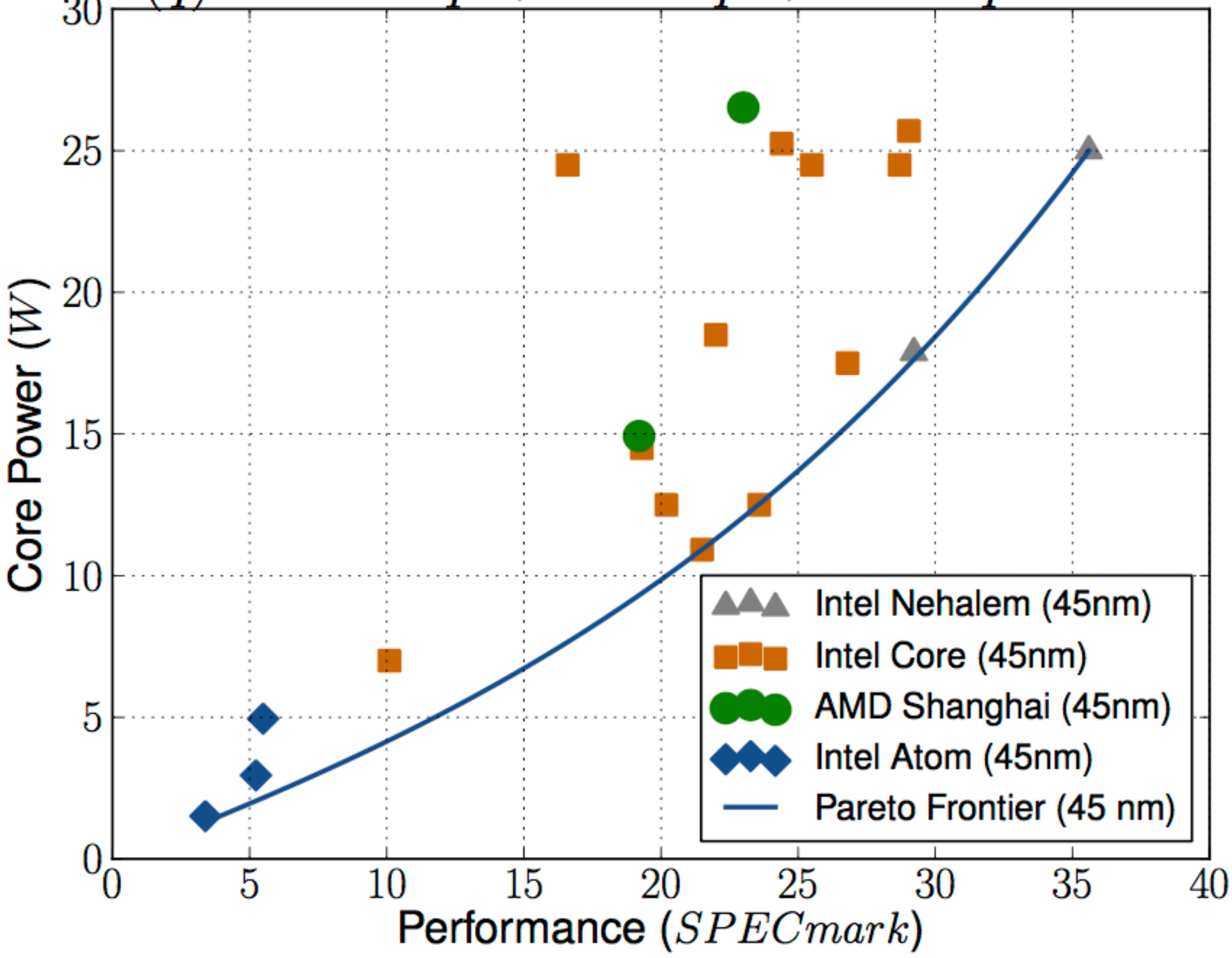

Clock Period (FO4 metric)
55
50
45
40
35
30
25
20
15
10
5
0
1992
'95
'97
'99
2002
'05
'08
'11
'14
500nm
350nm
250nm
180nm
130nm
100nm
70nm
50nm
35nm
16 FO4
8 FO4
SIA Scaling
Intel Data
ISSCC Data


1E+09
1E+08
1E+07
1E+06
Span (Reachable Bits)
f_{16}
f_8
f_{SIA}
250
180
130
100
70
50
35
Technology (nm)


Historical rates
Best experimental performance
Relative Performance
1000
100
10
1
1720
55%
7.4
12%
1997
250nm
1999
180nm
2002
130nm
2005
100nm
2008
70nm
2011
50nm
2014
35nm

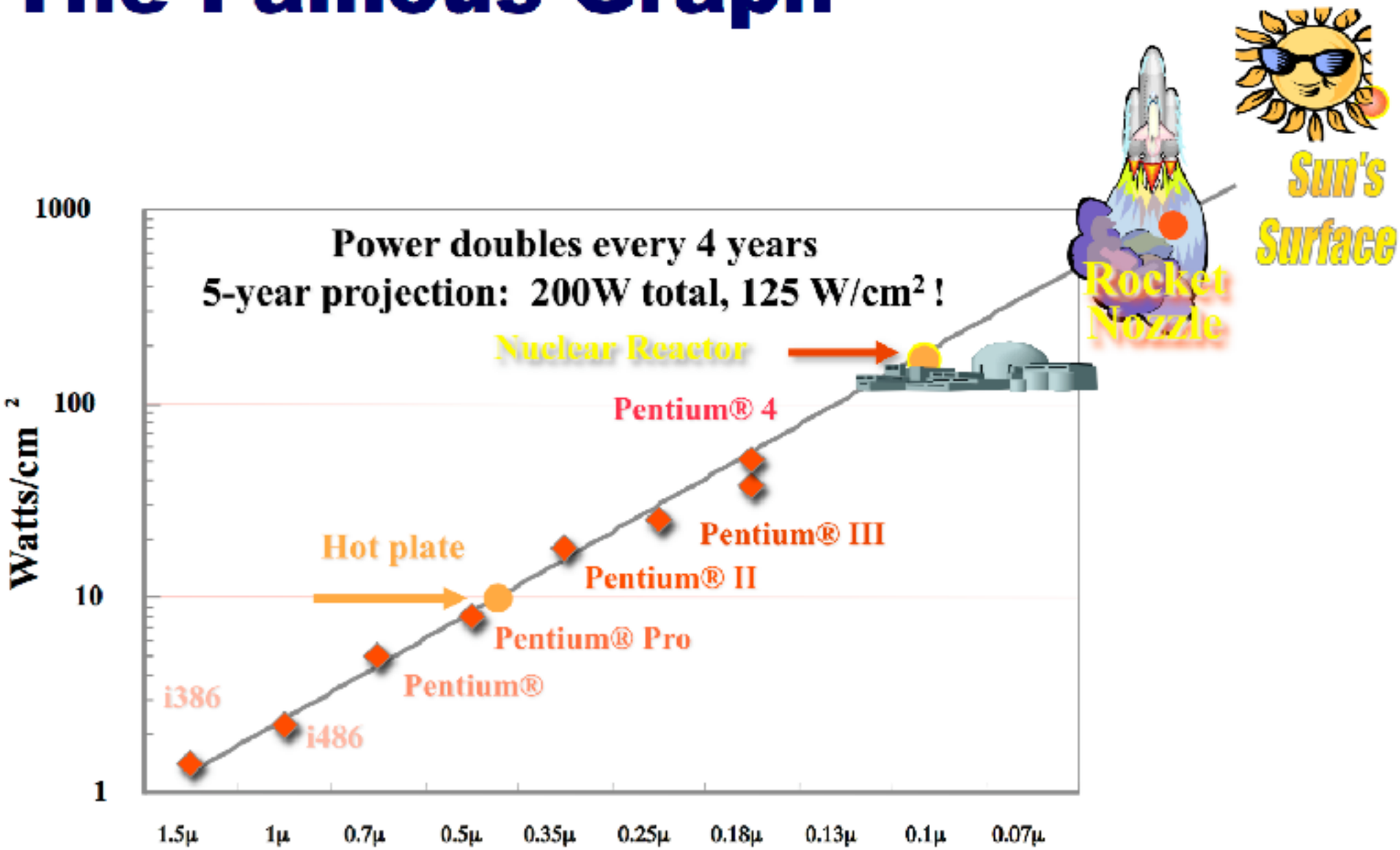

Core Power (W)
Performance (SPECmark)
0
10
20
30
40
50
60
70
80
90
0
5
10
15
20
25
30
35
40
Processors (600 nm)
Processors (350 nm)
Processors (250 nm)
Processors (180 nm)
Processors (130 nm)
Processors (90 nm)
Processors (65 nm)
Processors (45 nm)
Pareto Frontier (45 nm)

$$P(q) = 0.0002q^3 + 0.0009q^2 + 0.3859q - 0.0301$$



Dennard Scaling

The Famous Graph



From "New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies" – Fred Pollack, Intel Corp. Micro32 conference key note - 1999.

Power = Capacitance-per-component X
Components X
Activation X
Frequency X
Voltage^2

Linear mode of operation: Voltage prop Frequency

$$P = 1/2CV^2fA \quad ; \; V \text{ pprop } f$$

Dennard:

"We can keep power consumption constant"

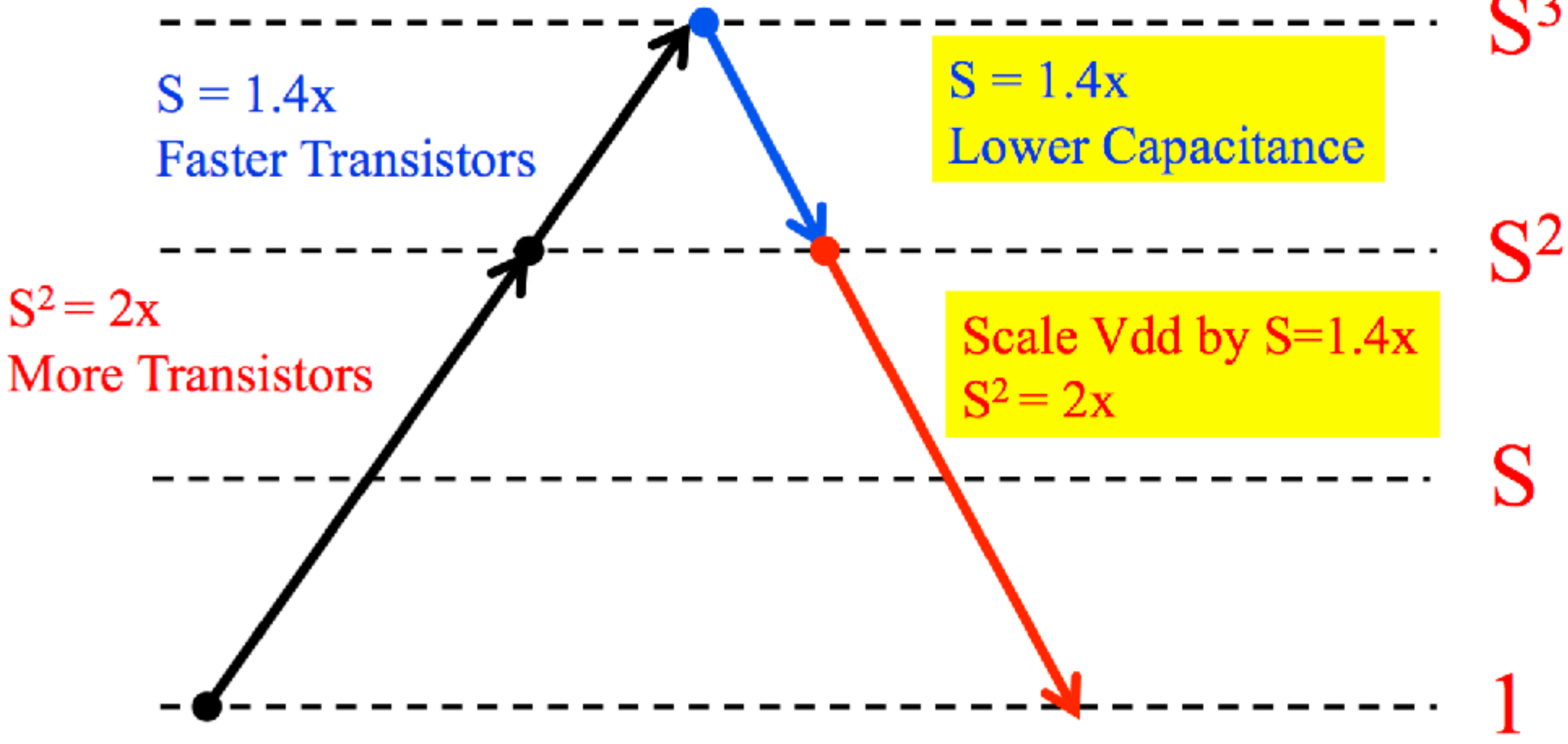
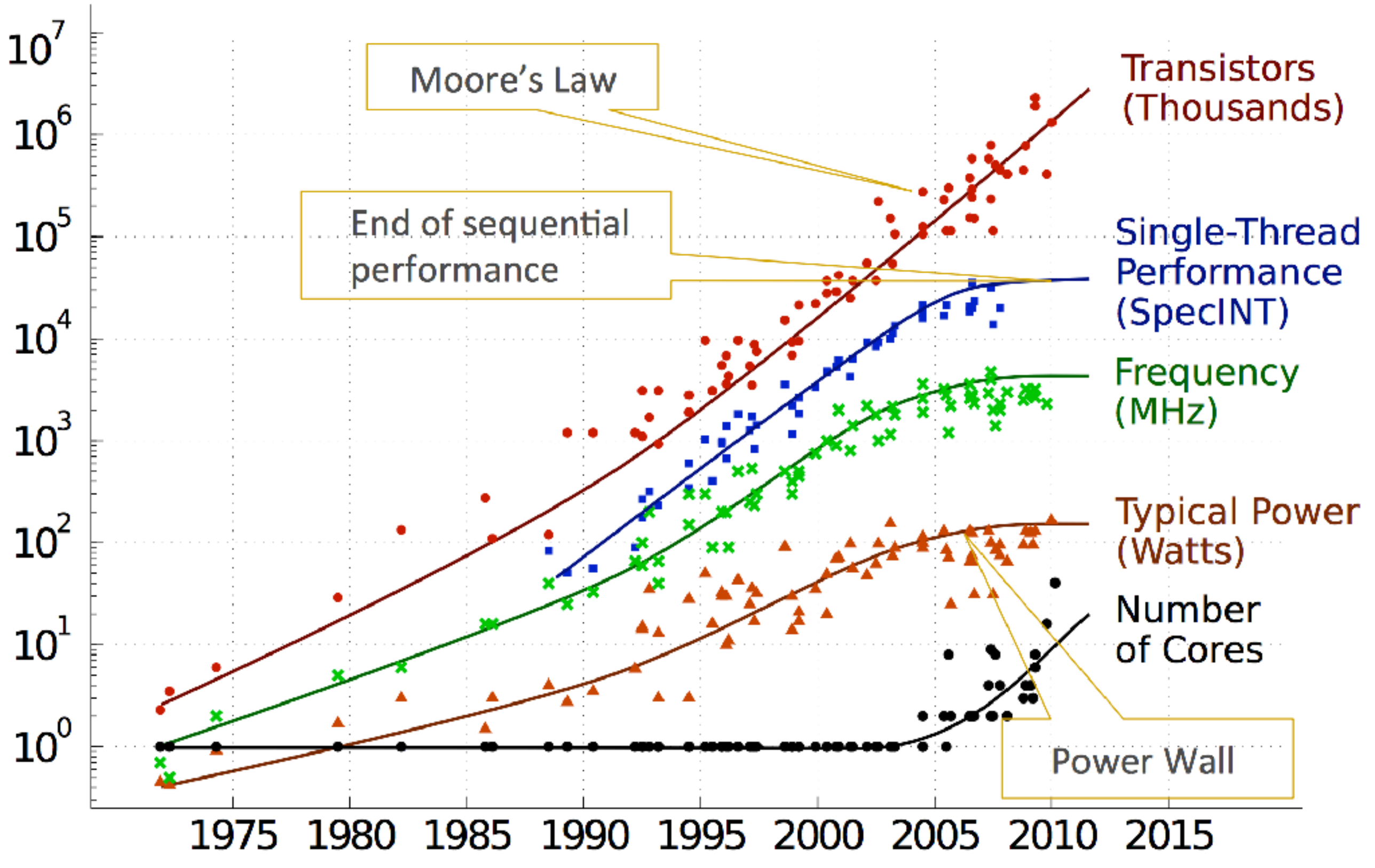


Table 1. Dennard vs. post-Dennard (leakage-limited) scaling.[1] In contrast to Dennard scaling,[5] which held until 2005, under the post-Dennard regime, the total chip utilization for a fixed power budget drops by S^2 with each process generation. The result is an exponential increase in dark silicon for a fixed-sized chip under a fixed area budget.

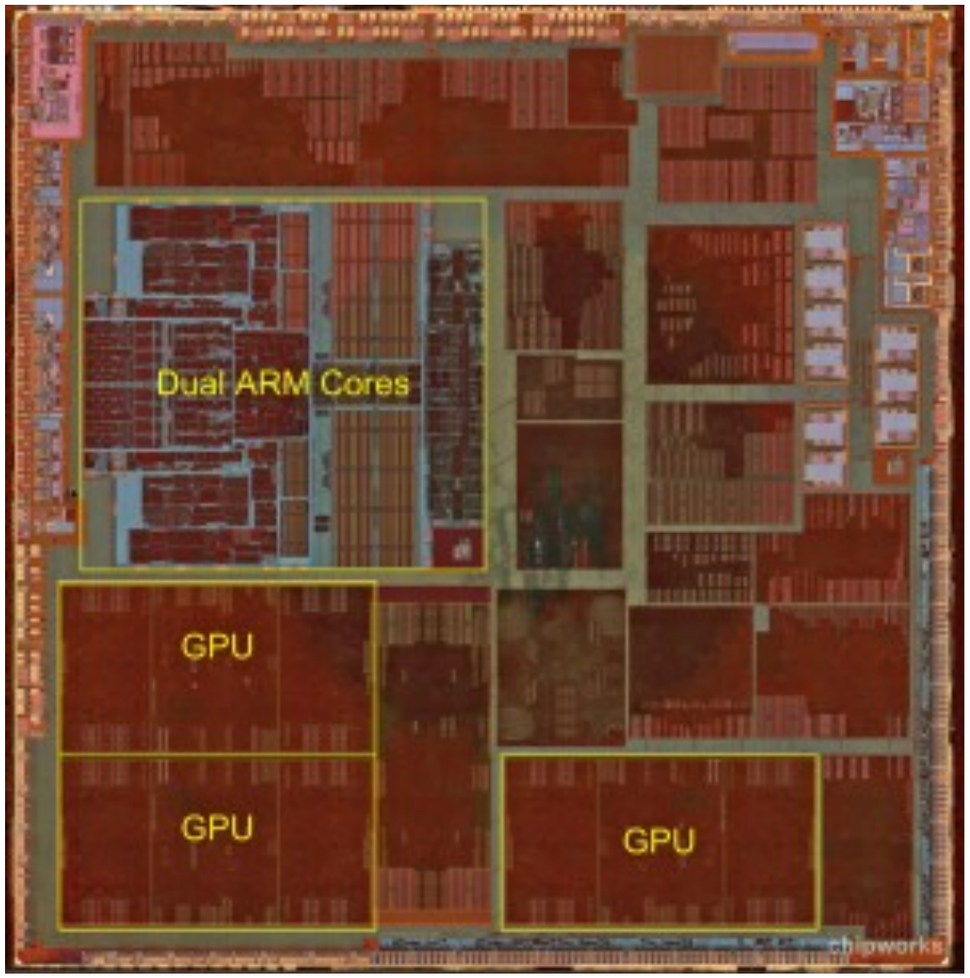
Transistor property	Dennard	Post-Dennard
Δ Quantity	S^2	S^2
Δ Frequency	S	S
Δ Capacitance	$1/S$	$1/S$
V_{DD}^2	$1/S^2$	1
⇒ Δ Power = Δ $QFCV^2$	1	S^2
⇒ Δ Utilization = 1/Power	1	$1/S^2$



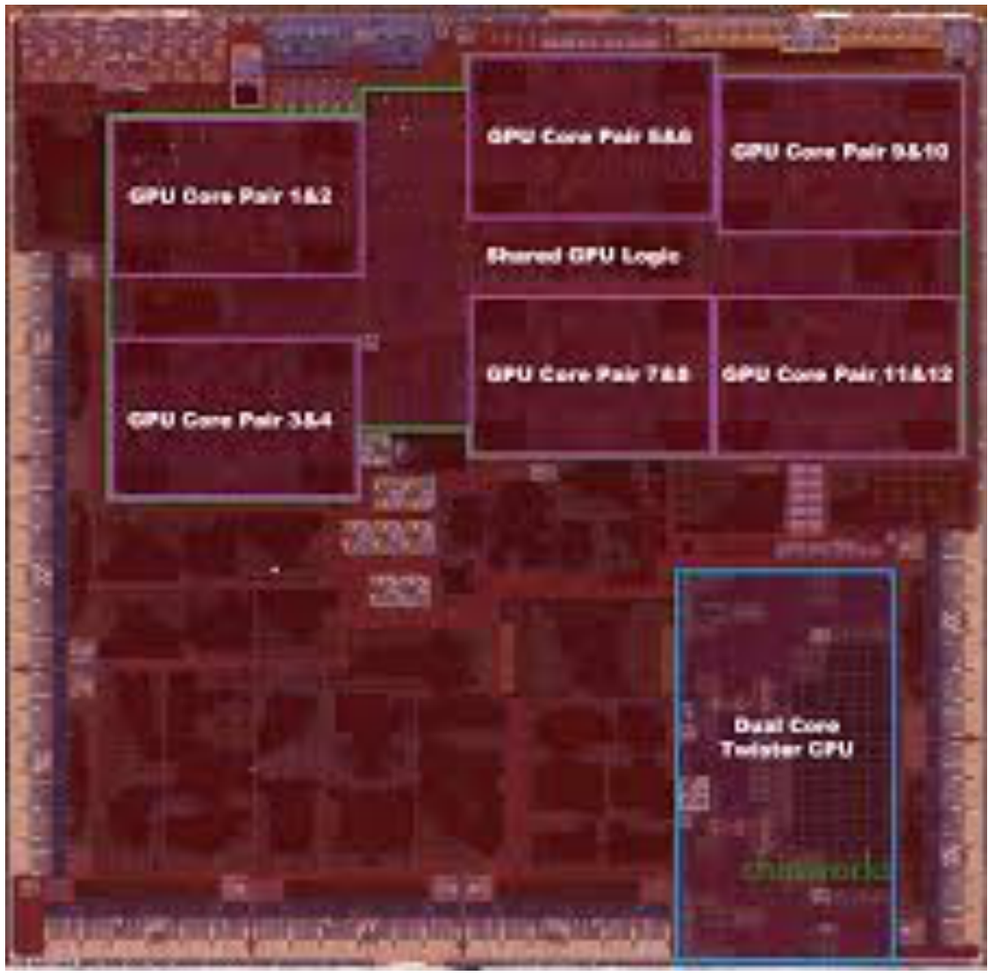
Data collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, C. Batten

What now?

Specialize



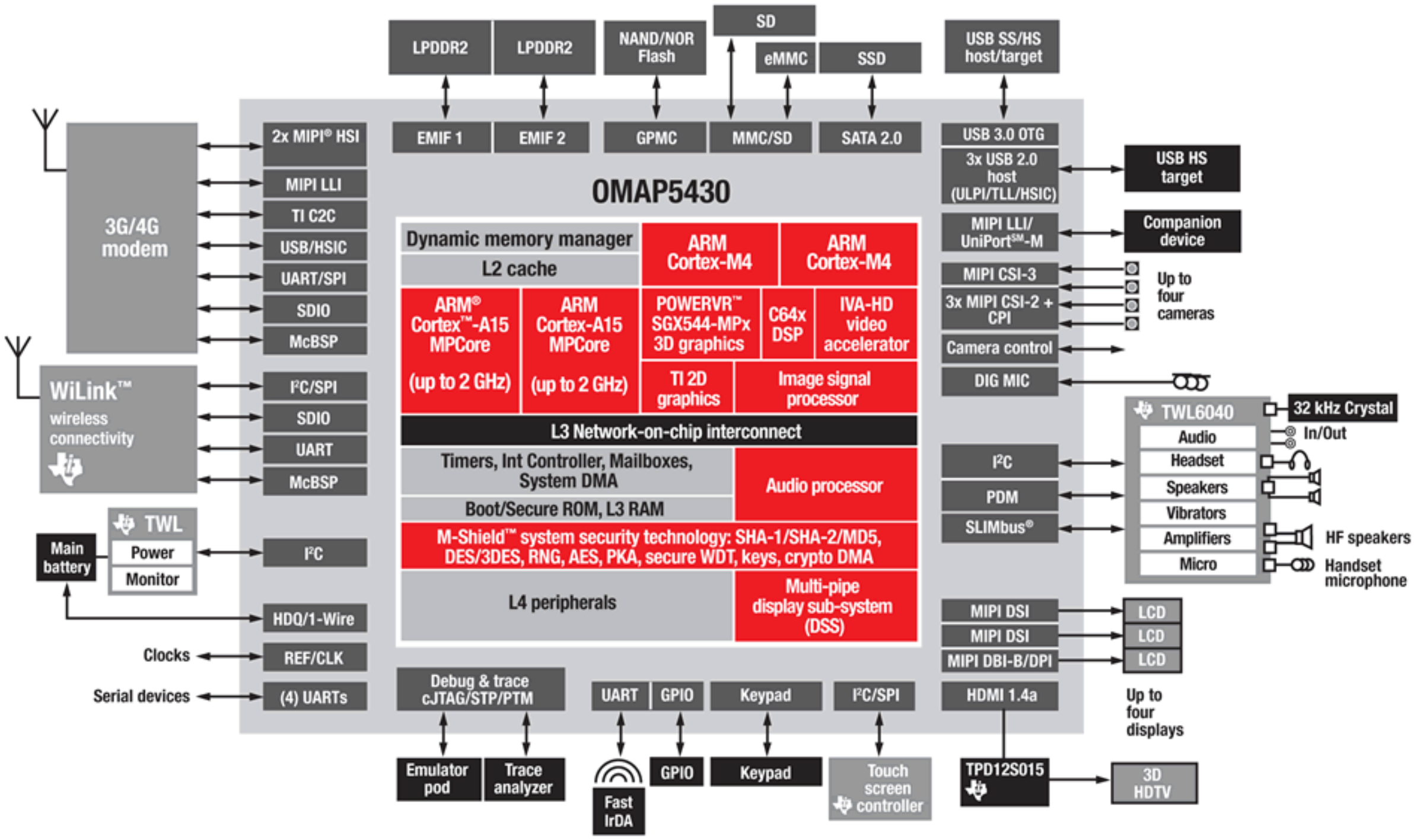
A6

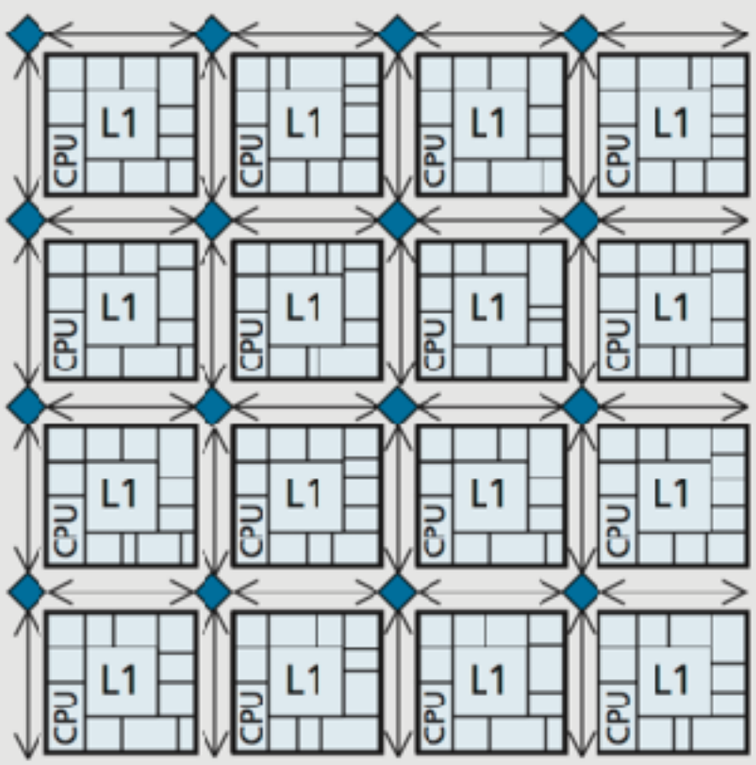


A9X

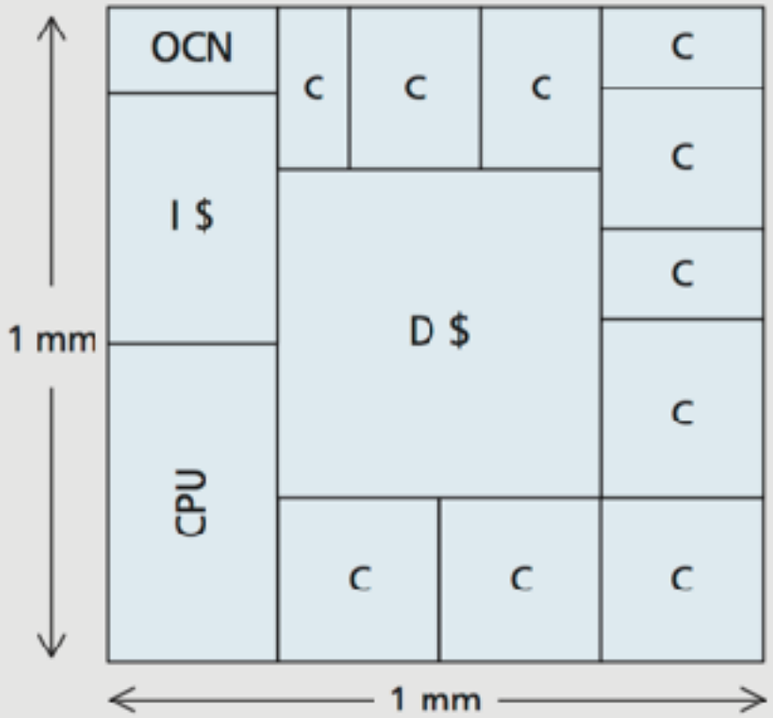
Specialization is the answer?

TI OMAP5430 SoC

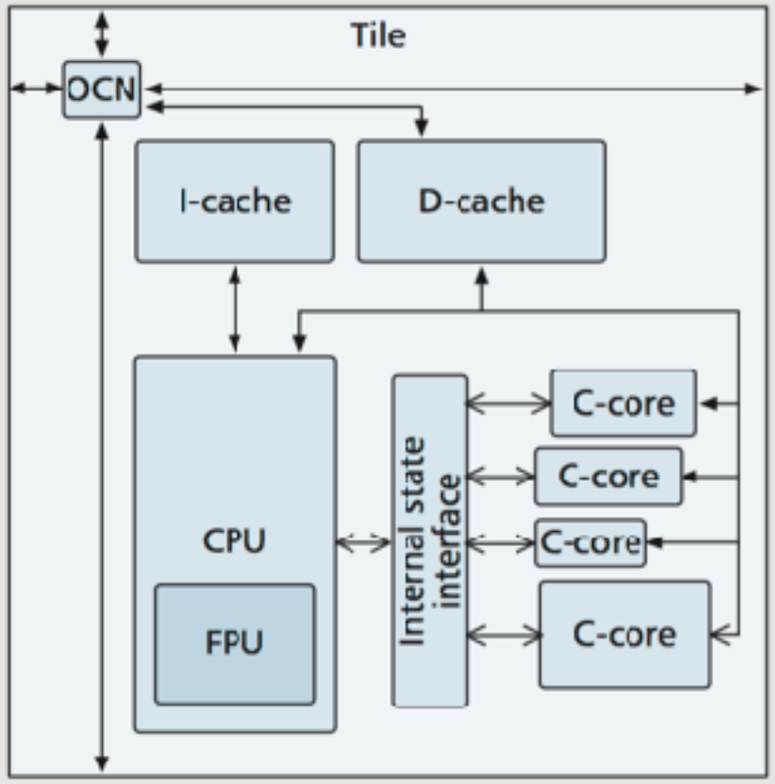



CPU
L1

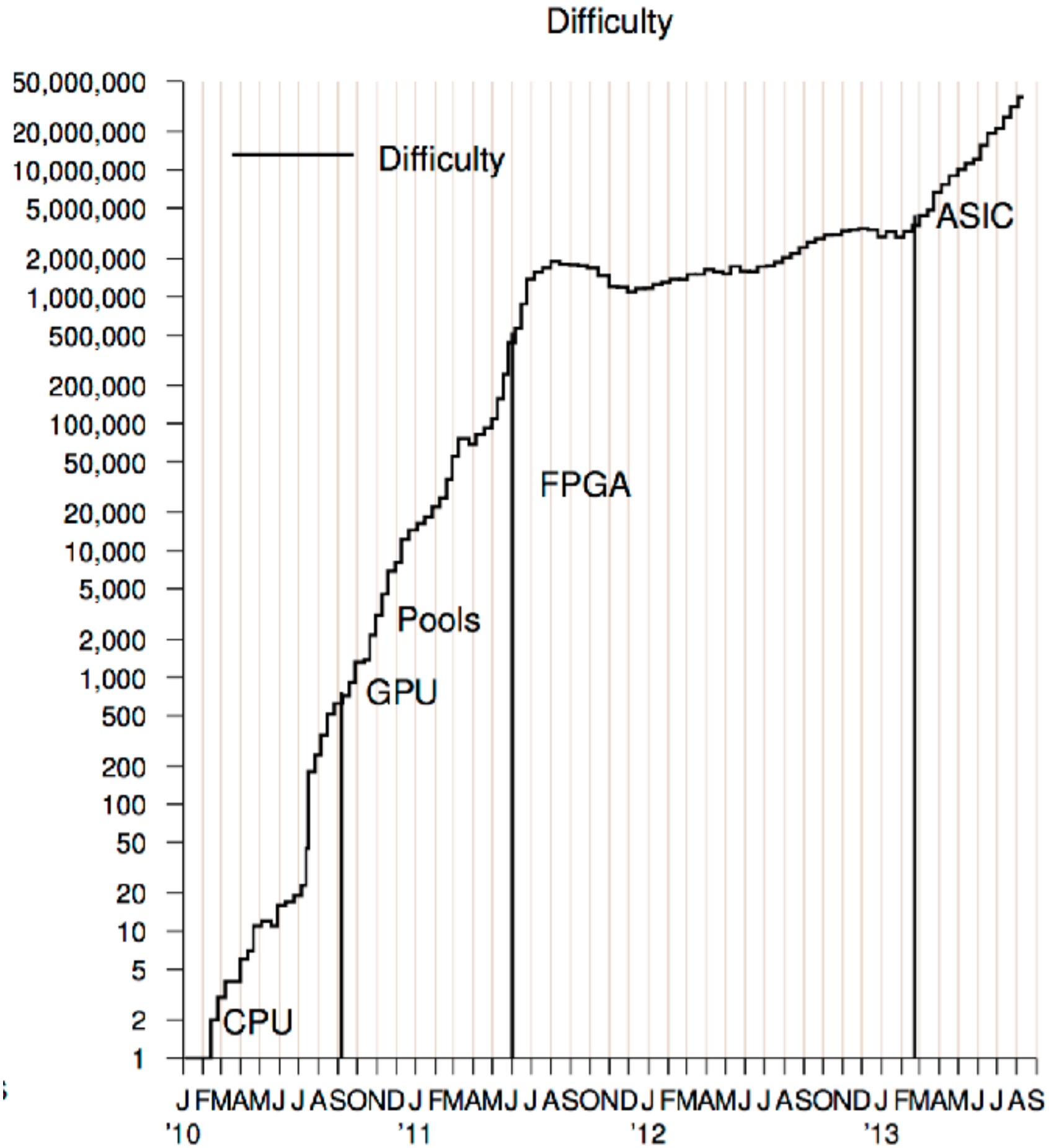
(a)

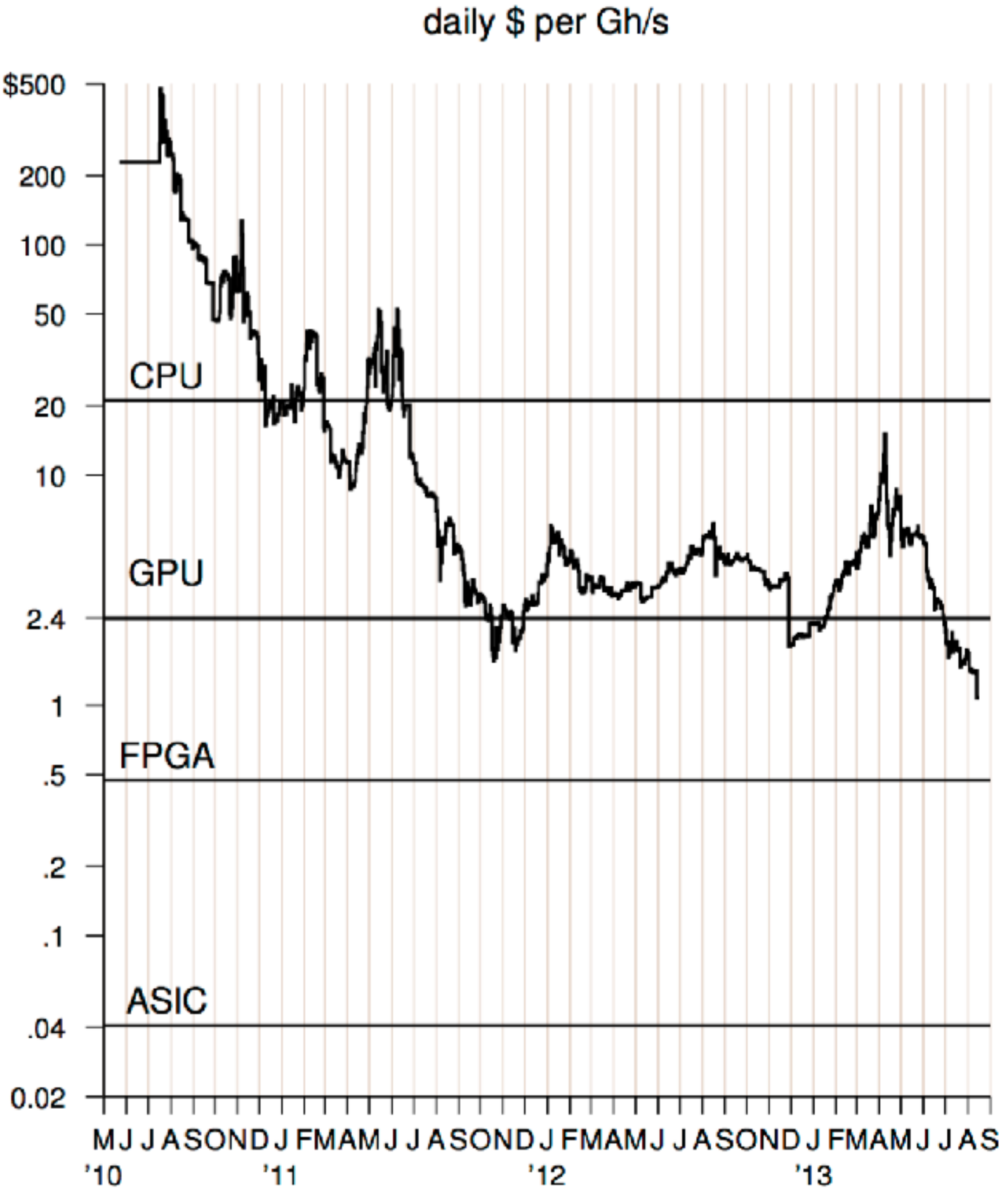

OCN
C
I $
D $
CPU
1 mm
1 mm

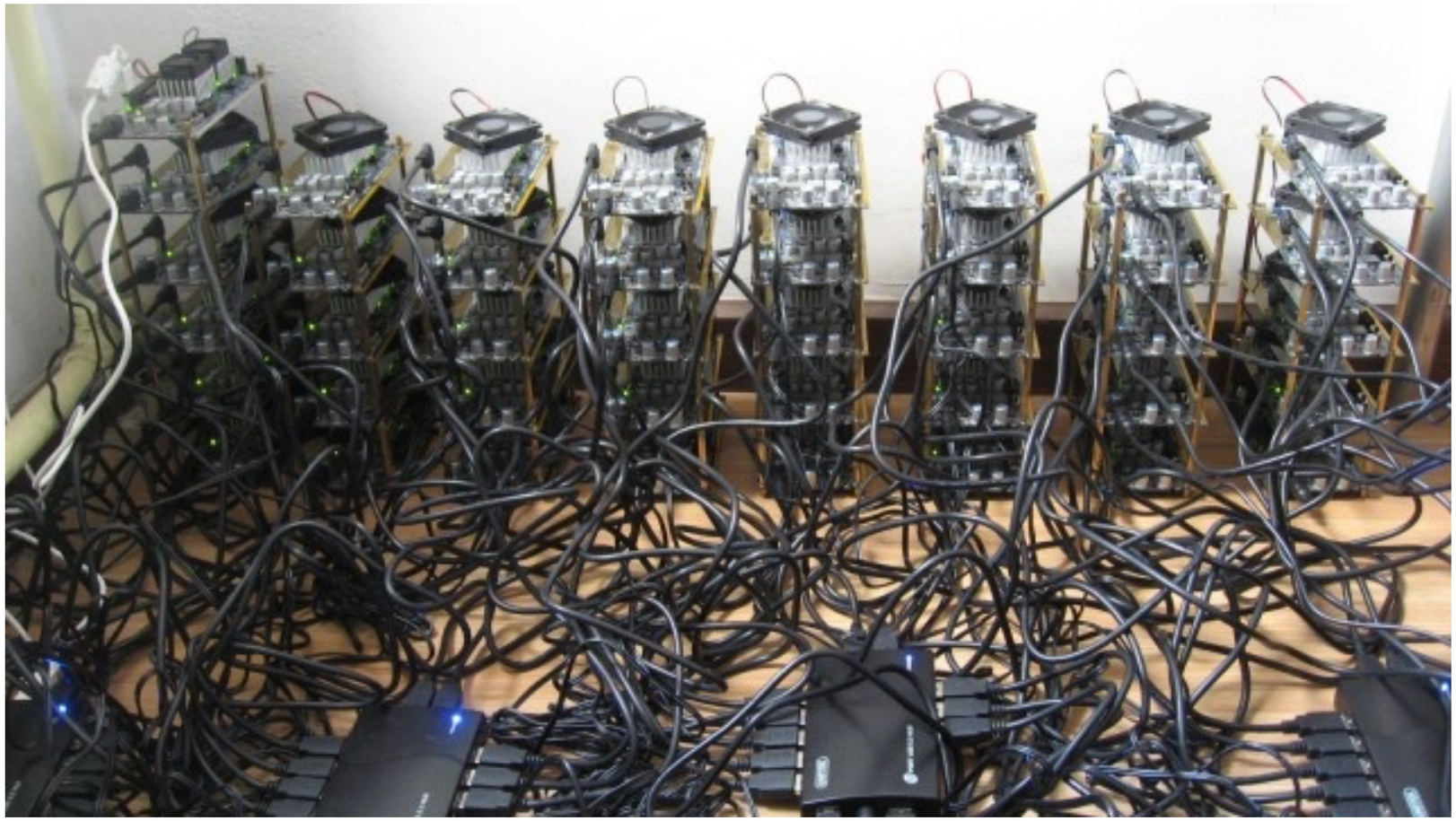
(b)

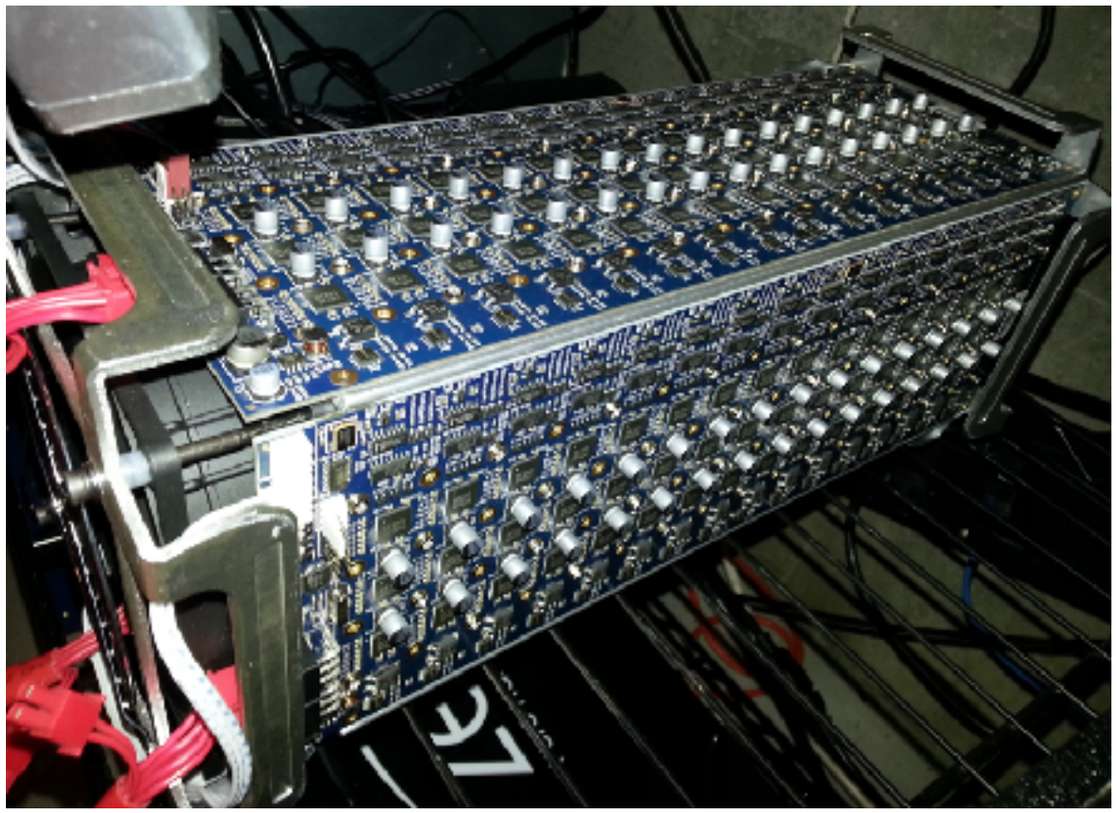

Tile
OCN
I-cache
D-cache
CPU
FPU
Internal state interface
C-core

(c)

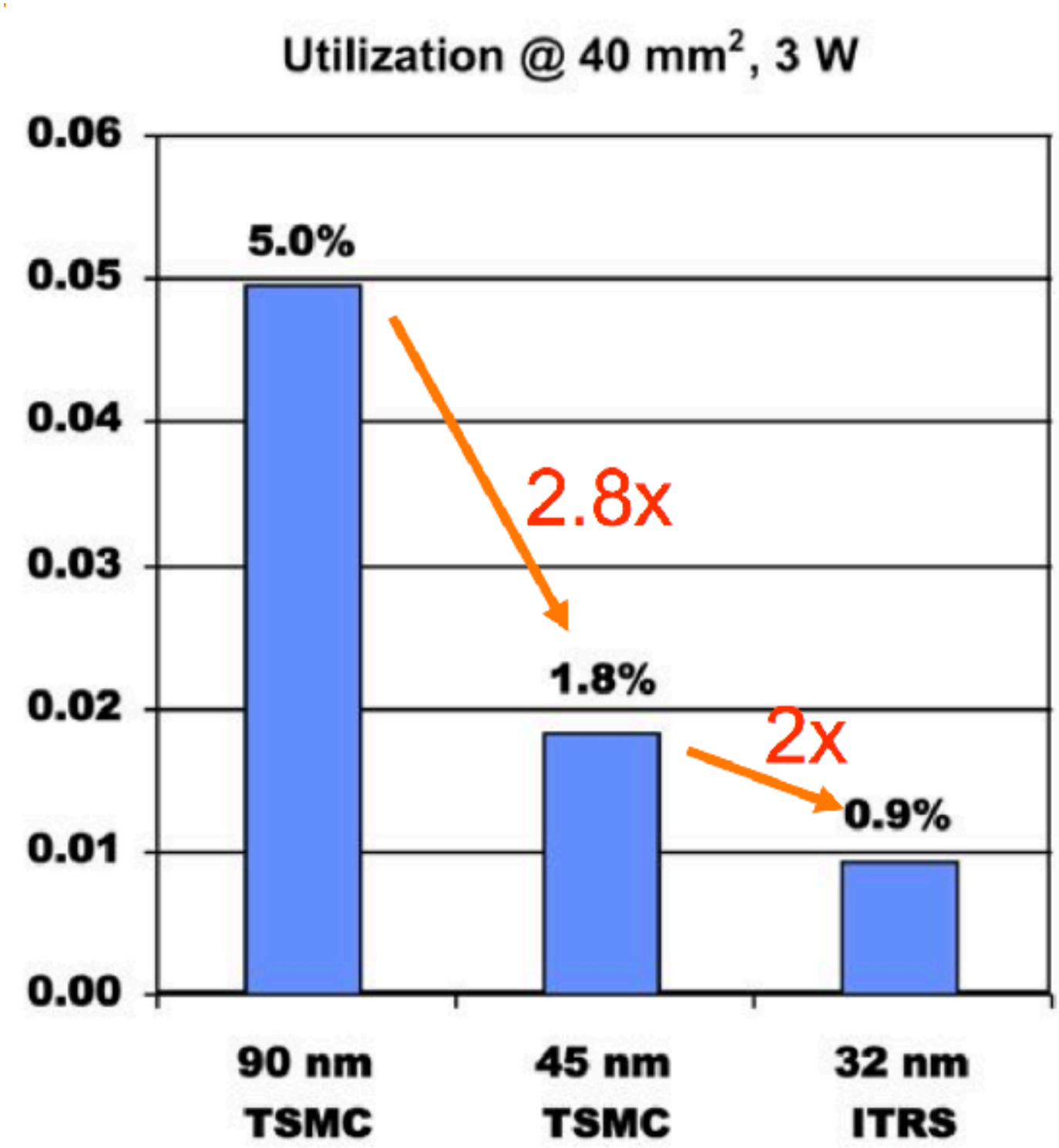

Difficulty
50,000,000
20,000,000
10,000,000
5,000,000
2,000,000
1,000,000
500,000
200,000
100,000
50,000
20,000
10,000
5,000
2,000
1,000
500
200
100
50
20
10
5
2
1
Difficulty
ASIC
FPGA
Pools
GPU
CPU
J FMAMJ J ASONDJ FMAMJ J ASONDJ FMAMJ J ASONDJ FMAMJ J AS
'10
'11
'12
'13


daily $ per Gh/s
$500
200
100
50
20
10
2.4
1
.5
.2
.1
.04
0.02
CPU
GPU
FPGA
ASIC
M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S
'10
'11
'12
'13



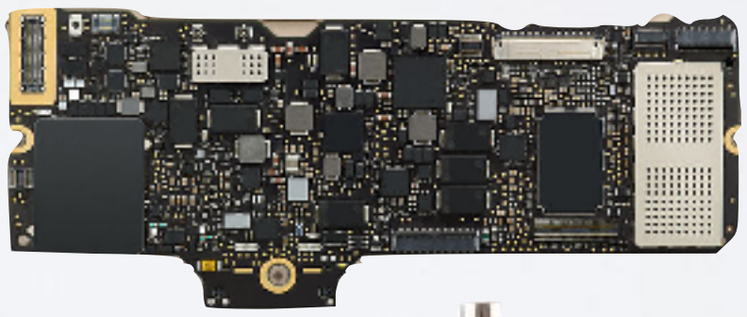


Do less

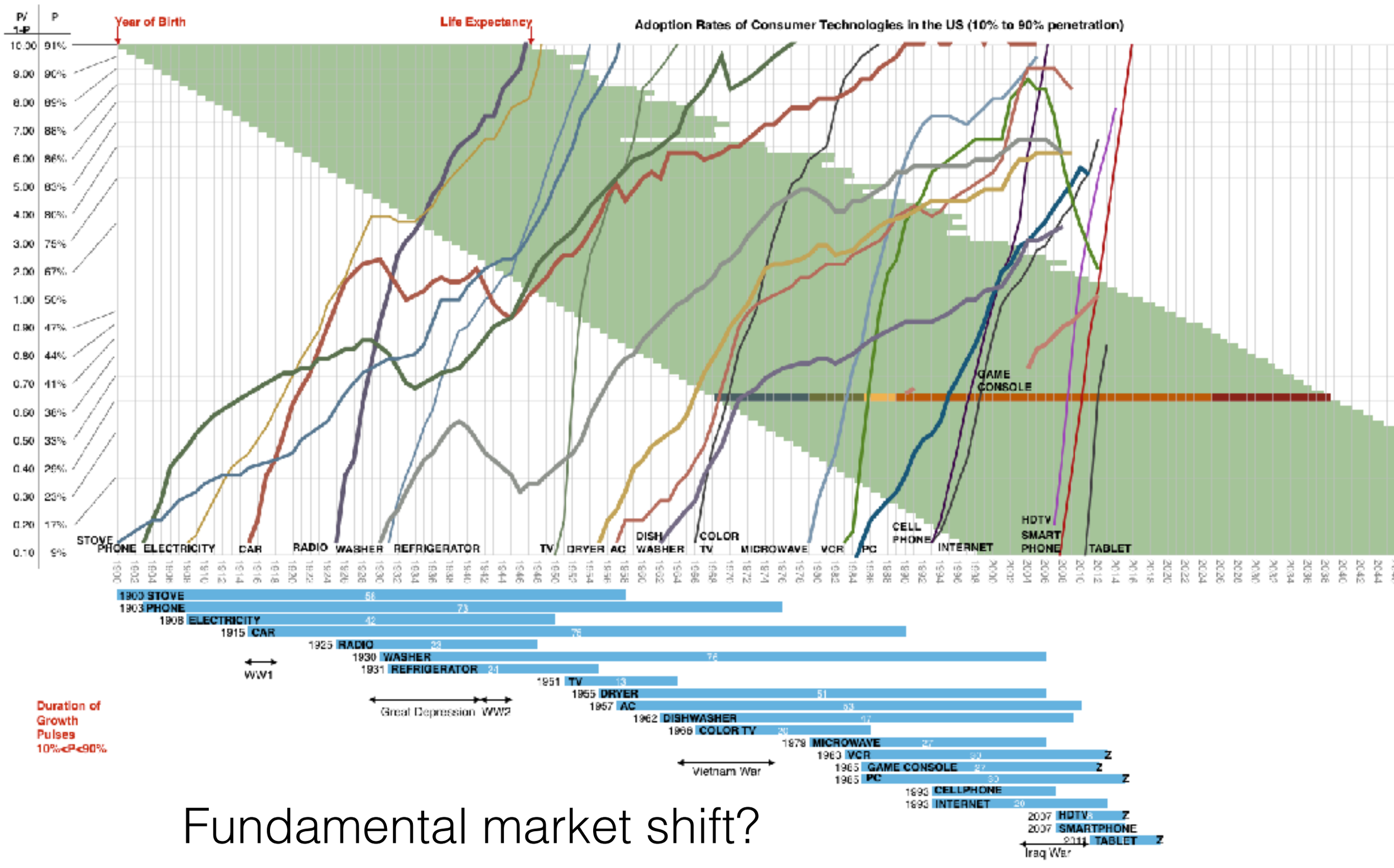

Utilization @ 40 mm², 3 W
0.06
0.05
0.04
0.03
0.02
0.01
0.00
5.0%
2.8x
1.8%
2x
0.9%
90 nm
TSMC
45 nm
TSMC
32 nm
ITRS

Which one is the iPhone and which one is the Macbook?

to scale

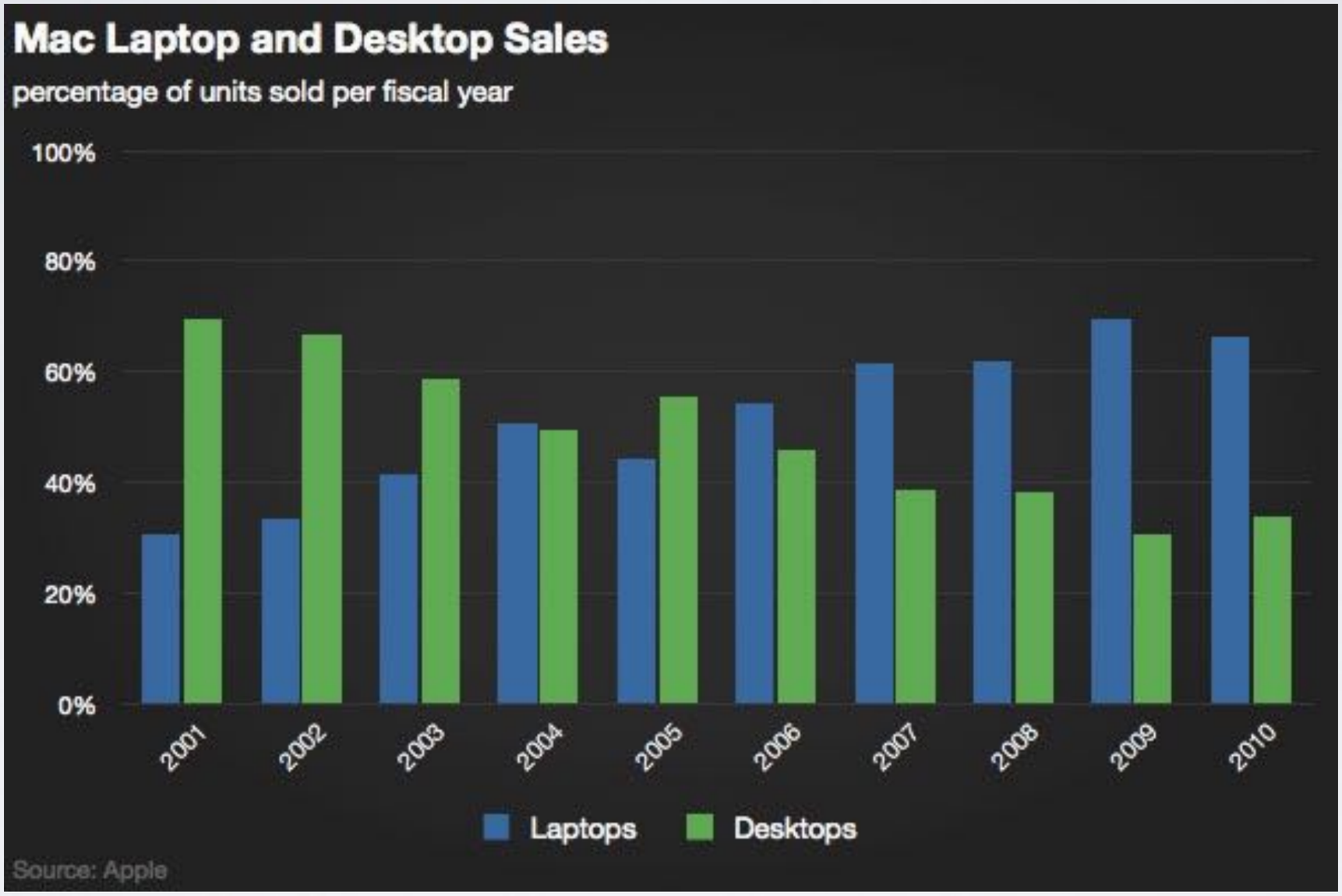


Do something else

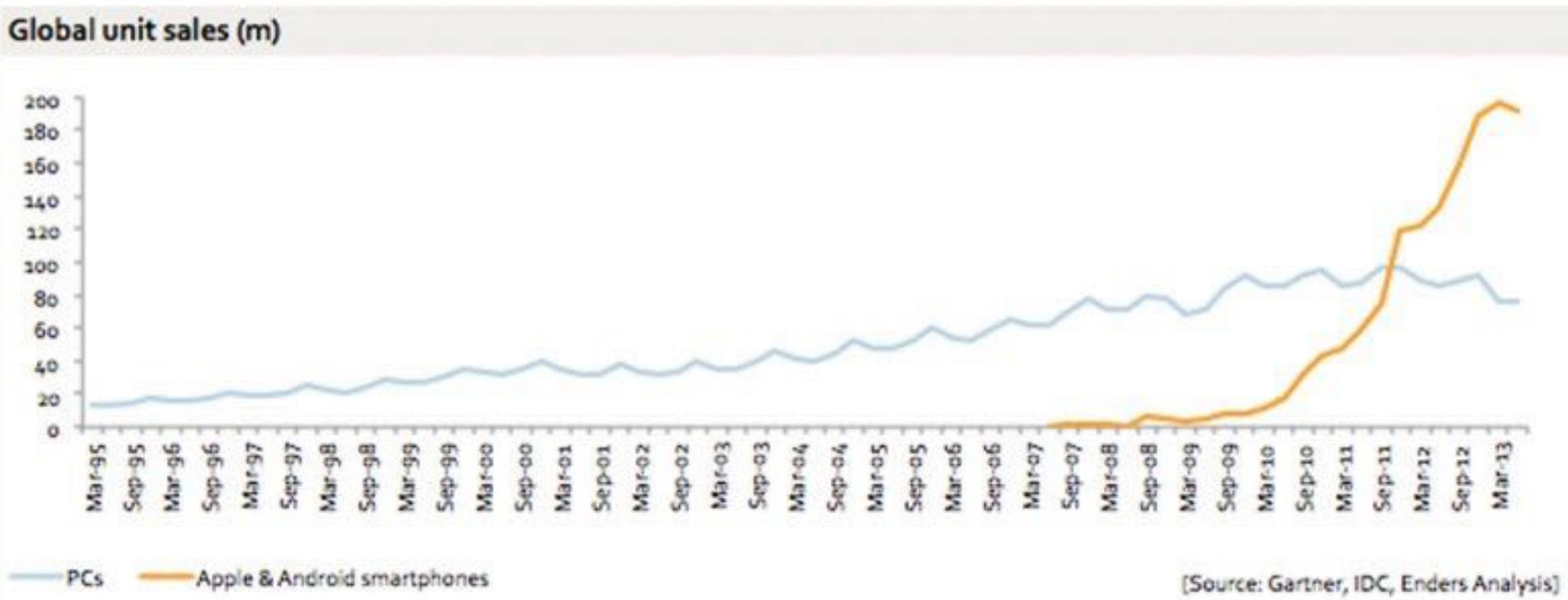


Fundamental market shift?

the death of the desktop?



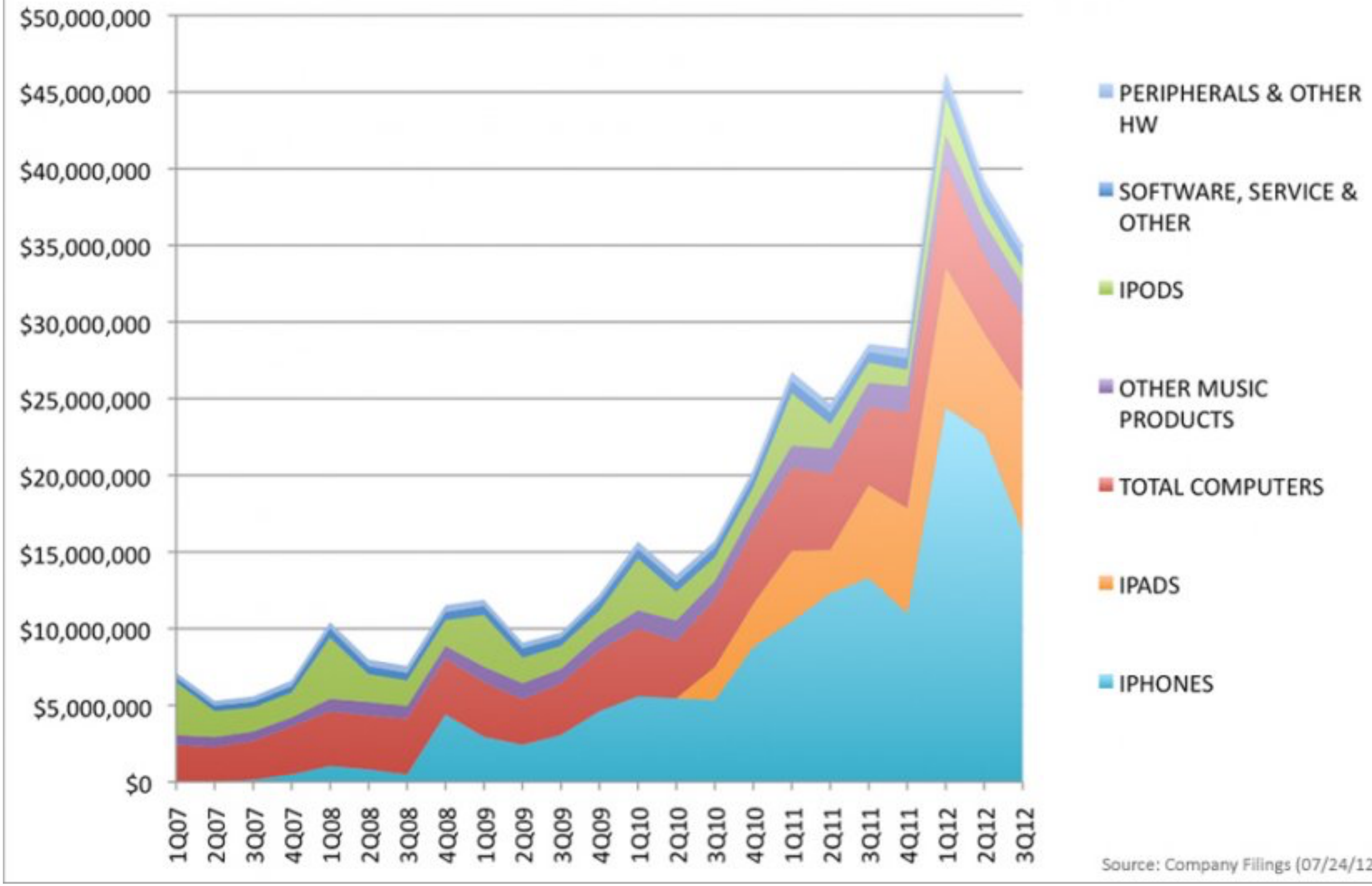
the death of the laptop?



BUSINESS INSIDER

Silicon Alley Insider Chart of the Day

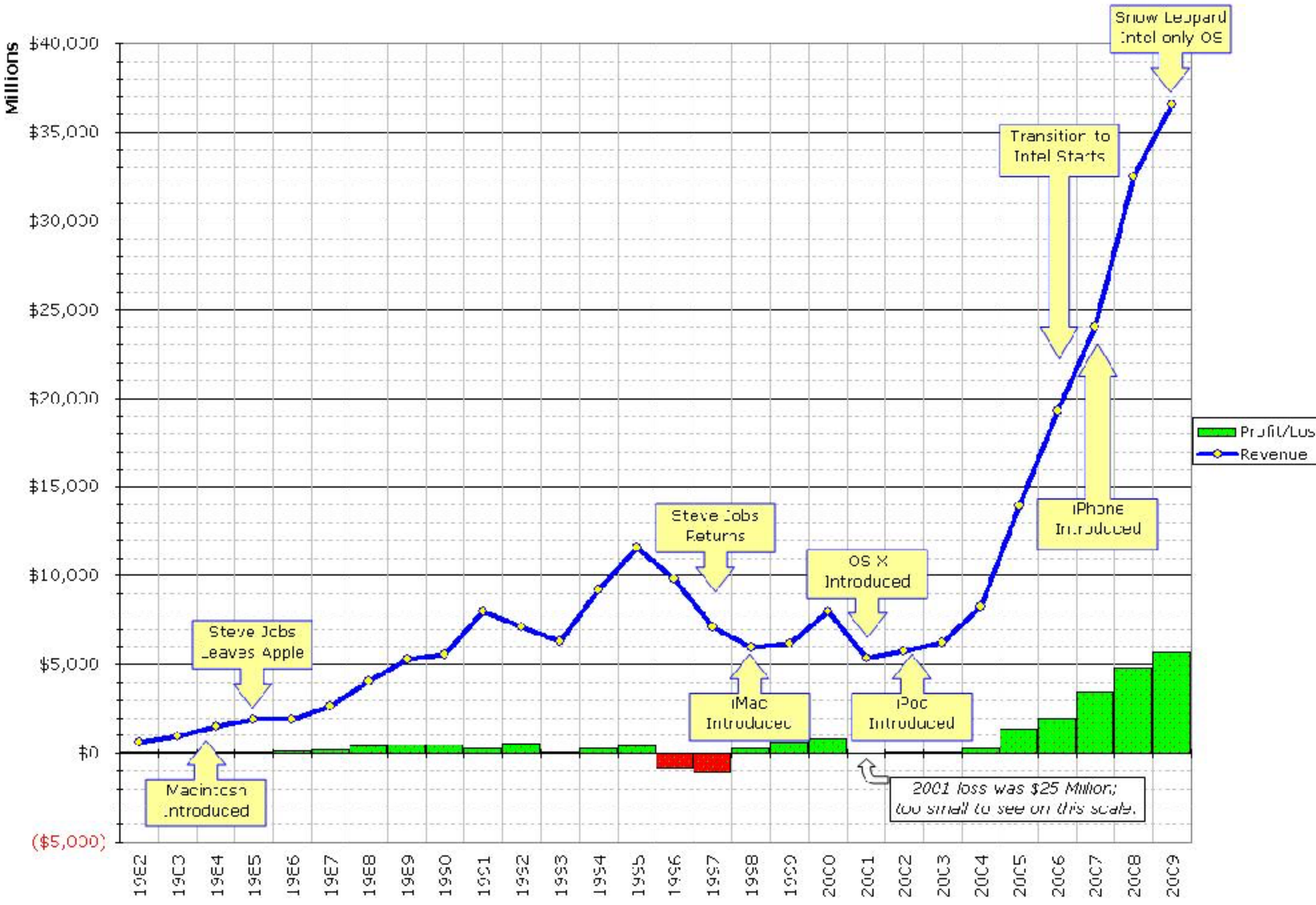
Apple's Quarterly Revenue By Product Category



Source: Company Filings (07/24/12

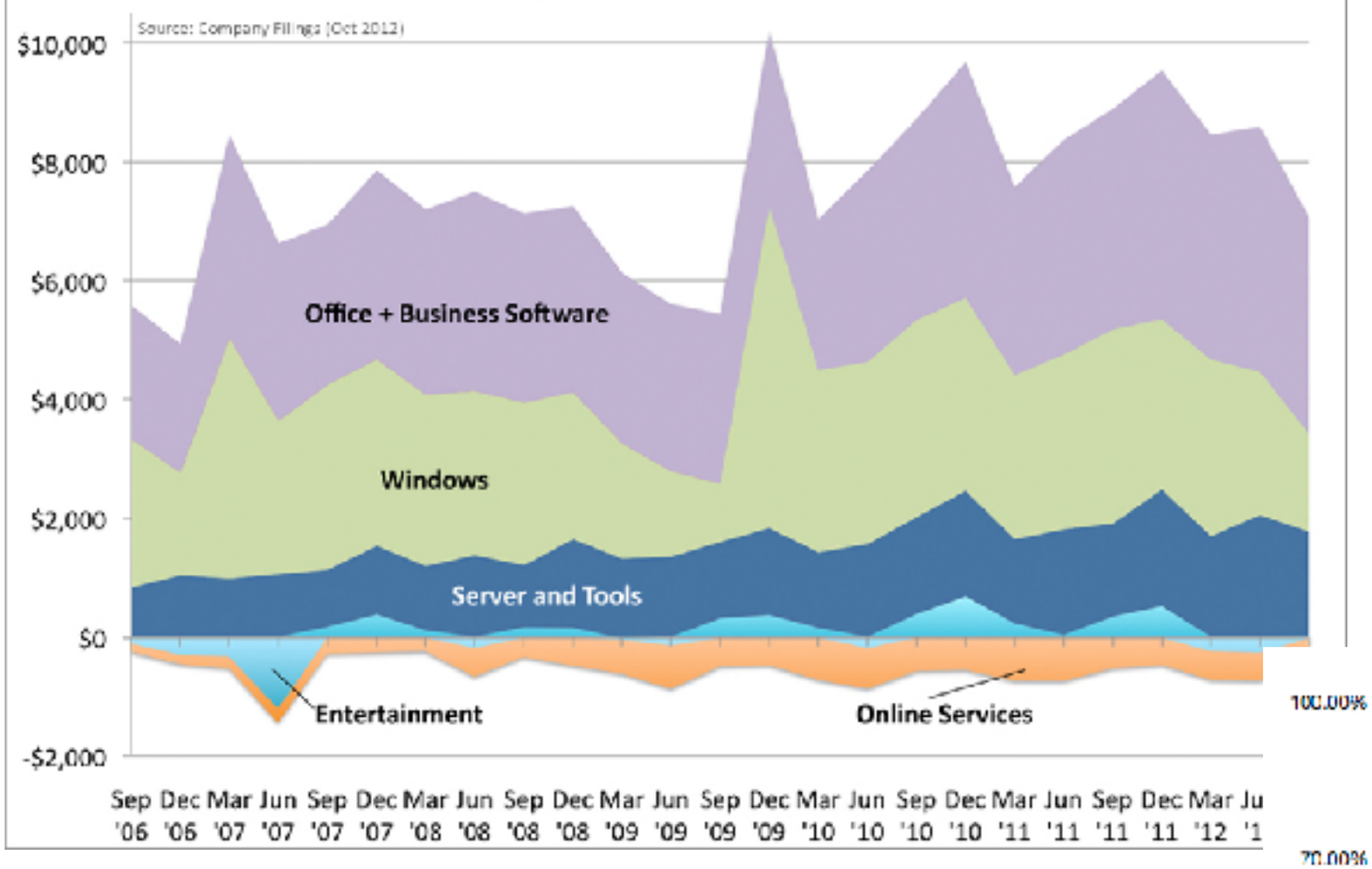
Apple Annual Gross Revenue and Profit/Loss

for a 28 year period from 1982 to 2009

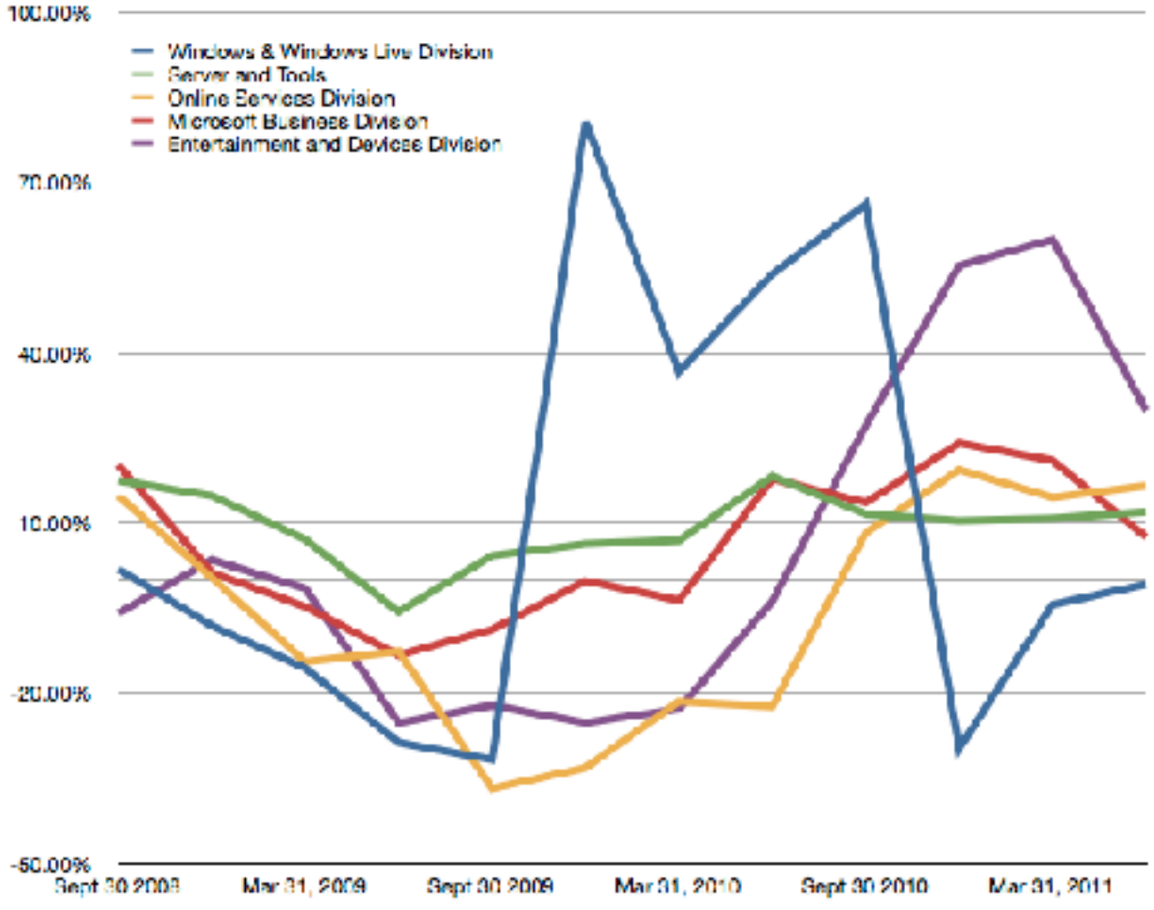


Silicon Alley Insider Chart of the Day

Microsoft's Operating Income (Loss) By Division



Year-on-Year Growth in Revenues by Division



Predicting the future is hard

- The future is more uncertain now than ever.
    - in 1998 Moore's law was supposed to end in a 3-5 years because of wavelength limitations
    - in 2001 I remember everyone thinking we'd scale pipelines out to 40-60 stages and push frequency up to 10-20Ghz, but then Denard Scaling ended
    - …. in 2003 the CTO of Intel told me we'd have 128 core Xeon chips in 10 years, but then Dark Silicon happened
    - …. in 2006
        - everyone thought parallel programming would be "solved" in 2-3 years, but hopefully this class has shown you that didn't happen
        - the iPhone happened, …but are we peeking? (half way there at least?)
        - Google started to use the term "the cloud" (1/4 of the way there?)