Moore's Law at 50:
Are we planning for retirement?
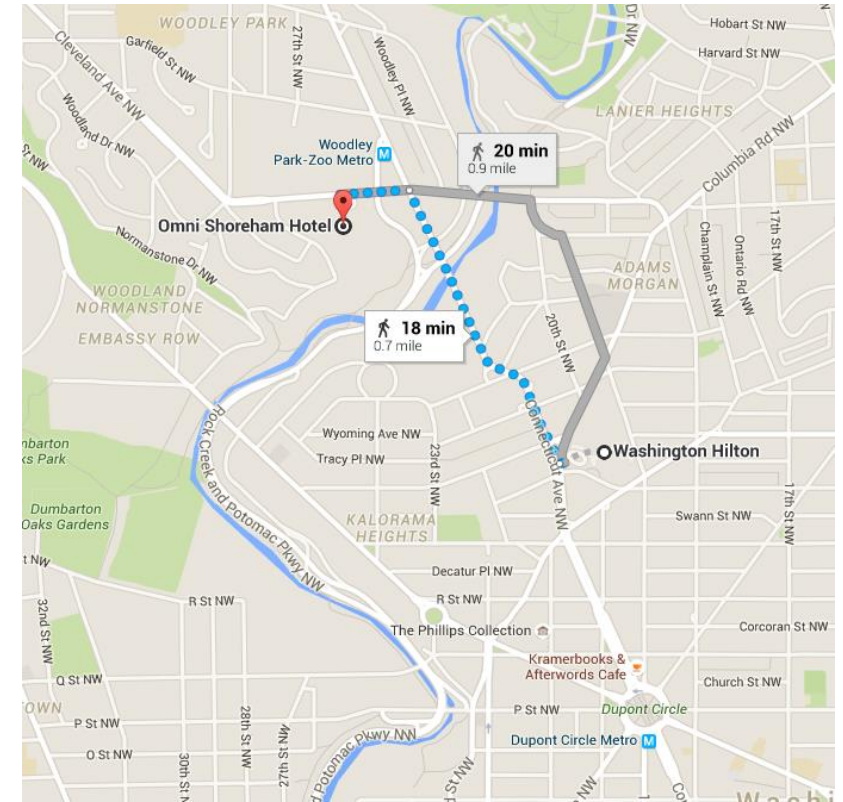
Greg Yeric

ARM

60th Anniversary of IEDM…

1st Annual Technical Meeting on Electron Devices: 1955
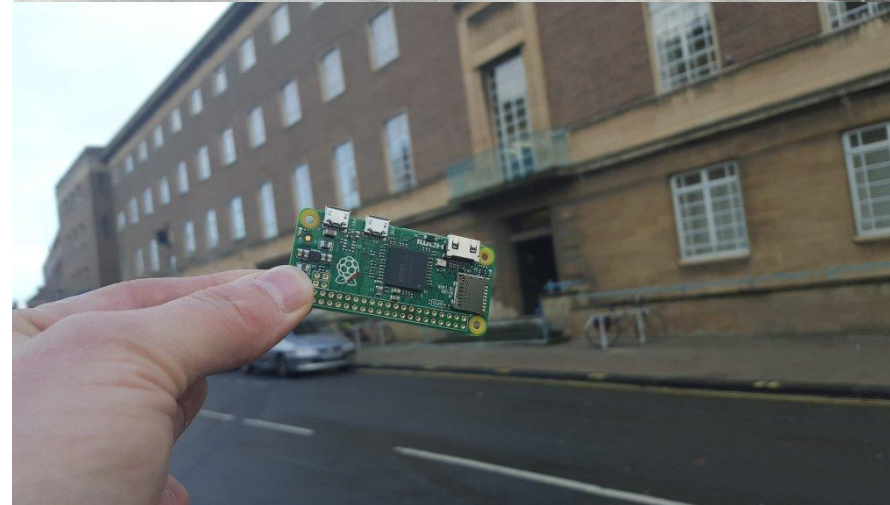

…and last in D.C. ?

IEDM 2015

"58 years on" *

Single-word nickel delay line (Science Museum, London)

Punched cards
3.3kB memory
3mS multiply
> 15 kW
~ £ 20,000

1080p video
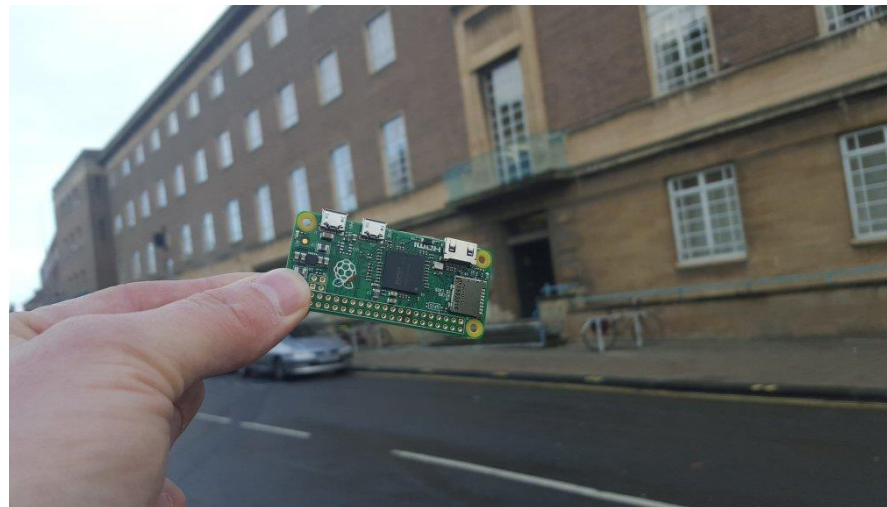512 MB memory
1 GHz CPU
0.6W
£ 4

*http://imgur.com/a/cafly

IEDM 2015

2015: Buy a magazine, get a free computer

or, just be a 7th grader
In the UK

raspberrypi.org

1080p video
512 MB memory
1 GHz CPU
0.6W
£ 4

2015: 50 years of an exponential

Session 15: Process and Manufacturing Technology – Moore and More

Data: https://en.wikipedia.org/wiki/Transistor_count

Transistor count

Double every 2 years

The Question

Moore's Law: 1965

Greg: 1965

Age 35

MIT Technology Review

NEWS & ANALYSIS · FEATURES · VIEWS · MULTIMEDIA · DISCUSSIONS · TOPICS · POPULAR: 50 SMARTEST COMPANIES 2015

TWELVE TOMORROWS — Science Fiction Inspired by Today's New Technologies — Pre-Order

Want to go ad free?

FEATURED STORY · 1 COMMENT

The End of Moore's Law?

The current economic boom is likely due to increases in computing speed and decreases in price. Now there are some good reasons to think that the party may be ending.

By Charles C. Mann on May 1, 2000

IEDM 2015

Age 40

ZDNet

CLOUD   CXO   HARDWARE   MICROSOFT   STORAGE   INNOVATION   MOBILITY   MORE ▾   NEWSLETTERS

JUST IN   MICROSOFT COMMITS TO 10-YEAR SUPPORT LIFECYCLE FOR WINDOWS 10

Is Moore's Law dead at 40 or is this just a mid-life crisis?

Last week, Michael Kanellos published this FAQ on the 40th anniversary of Moore's law, which is famously known as the phenomenon that computer processing power will double every 18 months.? Actually, Gordon Moore only said that transistor count would double every 24 months and it was David House (a former executive of Intel) who extrapolated that performance would double every 18 months as a result of the increase in transistors.?

By George Ou for Real World IT | April 5, 2005 -- 01:48 GMT (18:48 PDT) | Topic: Processors

IEDM 2015

Age 45

TECHNOLOGY | INNOVATION, THE INTERNET, GADGETS, AND MORE. | DEC. 20 2005 3:15 PM

The End of Moore's Law

Microchips are getting smaller—and that's the problem.

By Adam L. Penenberg

Until recently, **Moore's Law**, the observation that the number of transistors on a microchip doubles every 18 months to two years, seemed a self-fulfilling prophecy. When Intel co-founder Gordon Moore issued his famous prediction 40 years ago, a chip could hold a few dozen transistors. Today, Intel can cram almost **1 billion transistors**, each of which is less than 100 nanometers in size, on a single microchip. (One nanometer is 1 millionth of a millimeter—the equivalent of about 10 hydrogen atoms.) The transistors on Intel's chips are so tiny that they're not visible to the naked eye. *

2015: Age 50

Moore's Law Hits 50, but It May Not See 60

Dr. Gordon Moore

/ ENTERPRISE

By Arik Hesseldahl | @ahess247 | EMAIL | ETHICS

April 15, 2015, 3:11 PM PDT

Handy chart from The Economist:

The Economist

Faith no Moore
Selected predictions for the end of Moore's Law

Cited reason:
- Economic limits
- Technical limits

Prediction issued ●---------■ Predicted end date

| 1995 | 2000 | 2005 | 2010 | 2015 | 2020 | 2025 | 2030 |

Gordon Moore, Intel — *1995* ●----------■ *2005*

G. Dan Hutcheson, VLSI Research — *1996* ●---■ *2003*

Isaac Chuang, IBM Research — *2000* ●-----------■ *2020*

Paolo Gargani, Intel — *2003* ●----------■ *2021*

Lawrence Krauss, Case Western, and **Glenn Starkman**, CERN — *2004* ●----------- *approx. 2600*

Gordon Moore, Intel — *2005* ●--------▬▬▬ *2015–25*

Michio Kaku, City College of NY — *2011* ●----▬ *2021–22*

Robert Colwell, DARPA; (fmr) Intel — *2013* ●----▬ *2020–22*

Gordon Moore, Intel — *2015* ●----■ *2025*

Sources: Press reports; *The Economist*

Economist.com

IEDM 2015

In the spirit of Moore's Law: Extrapolating a few points…

Faith no Moore
Selected predictions for the end of Moore's Law

Cited reason:
■ Economic limits ■ Technical limits

Prediction issued ●------■ *Predicted end date*

| | 1995 | 2000 | 2005 | 2010 | 2015 | 2020 | 2025 | 2030 |

Gordon Moore, Intel — *1995* ... ■ *2005*

G. Dan Hutcheson, VLSI Research — *1996* ... ■ *2003*

Isaac Chuang, IBM Research — *2000* ... ■ *2020*

Paolo Gargani, Intel — *2003* ... ■ *2021*

Lawrence Krauss, Case Western, and **Glenn Starkman**, CERN — *2004* ... *approx. 2600*

Gordon Moore, Intel — *2005* ... ■ *2015–25*

Michio Kaku, City College of NY — *2011* ... ■ *2021–22*

Robert Colwell, DARPA; (fmr) Intel — *2013* ... ■ *2020–22*

Gordon Moore, Intel — *2015* ... ■ *2025*

Sources: Press reports; *The Economist*
Economist.com

Moore's Law ends: 2029

Economist, April 2015

And here's the crunch: that minimum **cost per transistor has been rising since 28nm** chips hit the market five years or so ago. That is partly a result of decreasing yields, but also because of the escalating cost of the photolithography equipment needed to fabricate ever-smaller circuits. In short, the cost-effectiveness of chip manufacturing seems to have hit a sweet spot at about 28nm.
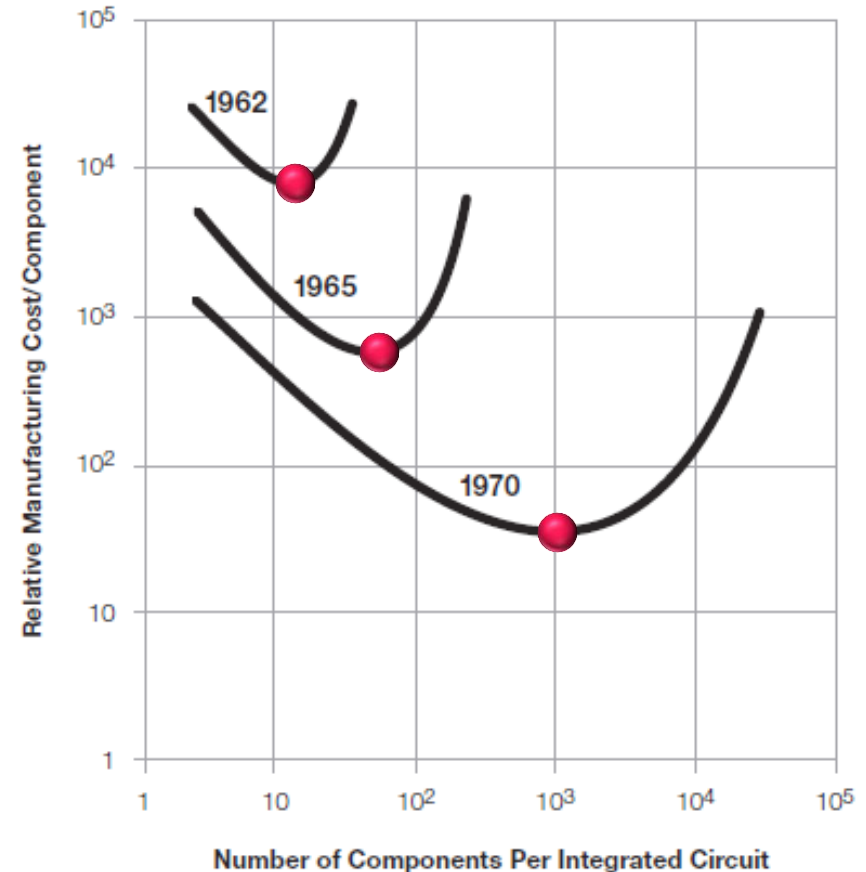
Data fueling the mantra

Technology	Area scale factor	Gates/mm^2 (KU)	Gate utilization (%)	Used gates/mm^2 (KU)	Parametric Yield Impact (Δ from Do yield)
90		637	86	548	97
65	0.57	1109	83	920	96
40	0.52	2139	78	1668	92

28nm: Optimal Balance of Cost and Power for 2015 Devices

Shrinking chips
Number and length of transistors bought per $

- Mid-range devices are highly sensitive to cost
- 28nm provides the most transistors per dollar

Under embargo until 6:00am GMT, February 11th 2014
- Not to be published without the consent of ARM

14NM Node

Crossover of Tra

Cost per Million Gates ($)
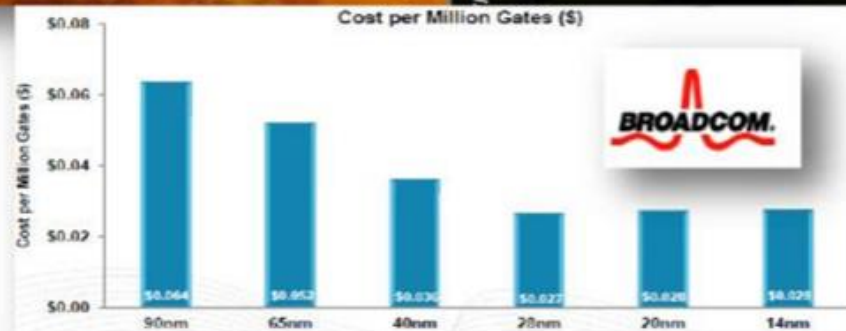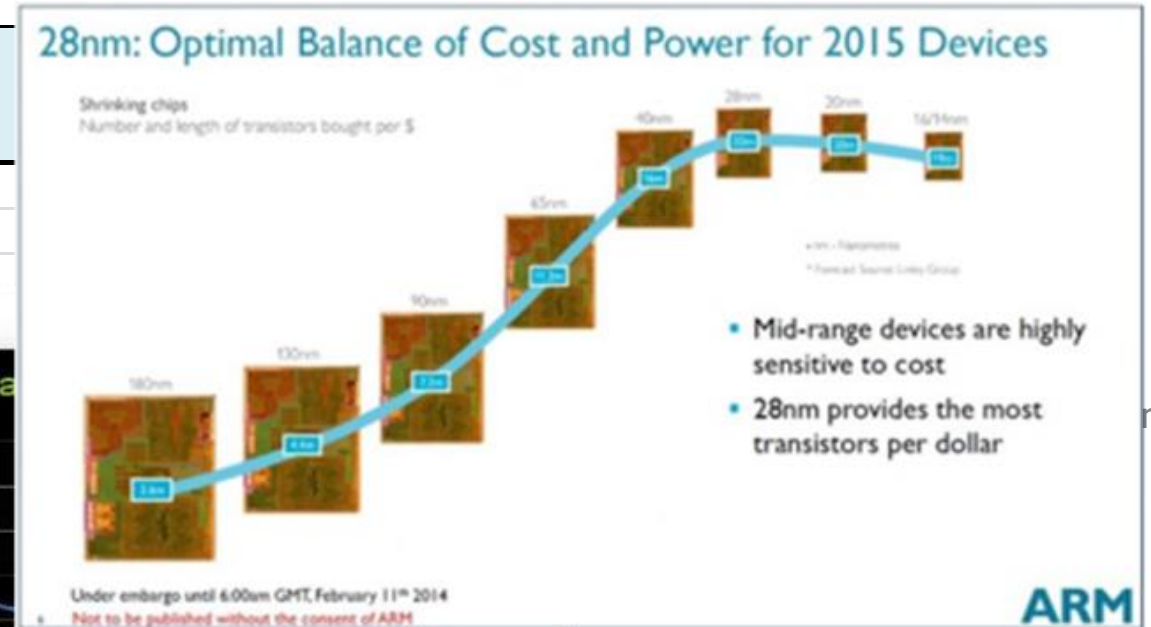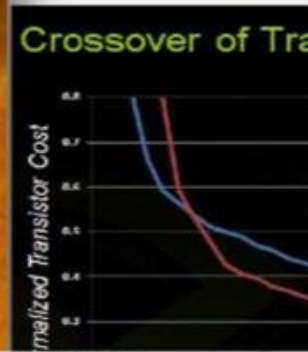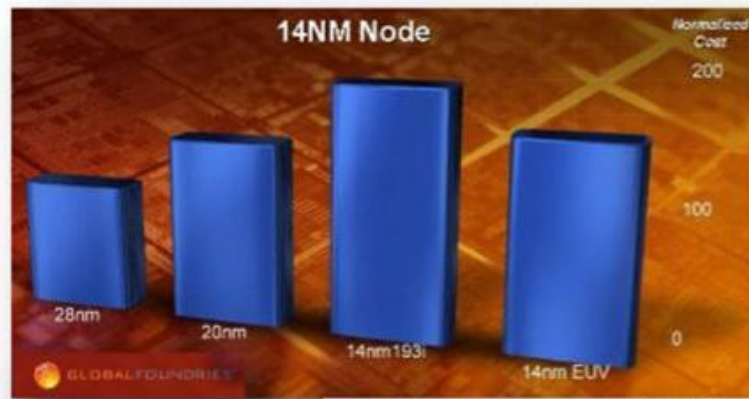
Sources: nVidea, ITPC, nov, 2011
Broadcom, IMEC, may 2012
GF, ISS, jan 2013

Data fueling the mantra

http://www.ibs-inc.net/

Technology	Area scale factor	Gates/ mm^2 (KU)	Gate utilization (%)	Used gates/mm^2 (KU)	Parametric Yield Impact (Δ from Do yield)	Actual used gates/mm^2 (KU)	Gates/wafer (MU)	Wafer price ($)	Wafer price (Δ)	Cost per million gate ($)
90		637	86	548	97	531	34009	1358		0.0399
65	0.57	1109	83	920	96	884	56554	1585	17%	0.0280
40	0.52	2139	78	1668	92	1535	98237	1899	20%	0.0193
28	0.54	3946	76	2999	87	2609	166982	2326	23%	0.0139
20	0.56	6992	65	4545	73	3318	212333	2981	28%	0.0140
16/14	0.56	12391	54	6691	61	4082	261222	4205	41%	0.0161

Alternate scenario:

<div style="border: 1px solid red; display: inline-block; padding: 4px;">NRE</div>

Technology	Area scale factor	Gates/ mm^2 (KU)	Gate utilization (%)	used gates/mm^2 (KU)	Parametric Yield Impact (Δ from Do yield)	Actual used gates/mm^2 (KU)	Gates/wafer (MU)	Wafer cost ($)	Wafer price (Δ)	Cost per million gate ($)	Mask set cost	Design Cost
90		637	86	548	97	531	34009	1811		0.0533	800,000	24,000,000
65	0.57	1109	84	932	96	894	57235	2177	16.8	0.0380	1,400,000	40,000,000
40	0.52	2139	82	1754	95	1666	106642	2712	19.7	0.0254	2,000,000	50,000,000
28	0.54	3946	83	3275	94	3079	197035	3500	22.5	0.0178	2,500,000	80,000,000
16/14	0.63	6263	81	5073	92	4668	298723	4375	25.0	0.0146	5,000,000	150,000,000
10	0.63	9942	79	7854	91	7147	457430	5906	35.0	0.0129	7,000,000	202,500,000
7	0.63	15781	77	12151	90	10936	699920	7383	25.0	0.0105	10,000,000	273,375,000

193i steppers for 7nm will be 50% faster than for 28nm

- \> 250 wafers per hour = < 15 seconds per wafer
- ~ 100 fields per wafer = < 0.15 seconds per die
- Chuck moves > 1m/second (develops >10G)
- 3 nm X and Y accuracy*

  *drop England onto the earth,
   align it to a precision of 12cm,
   7 times per second.

Etch is half of LE, and etch is improving also
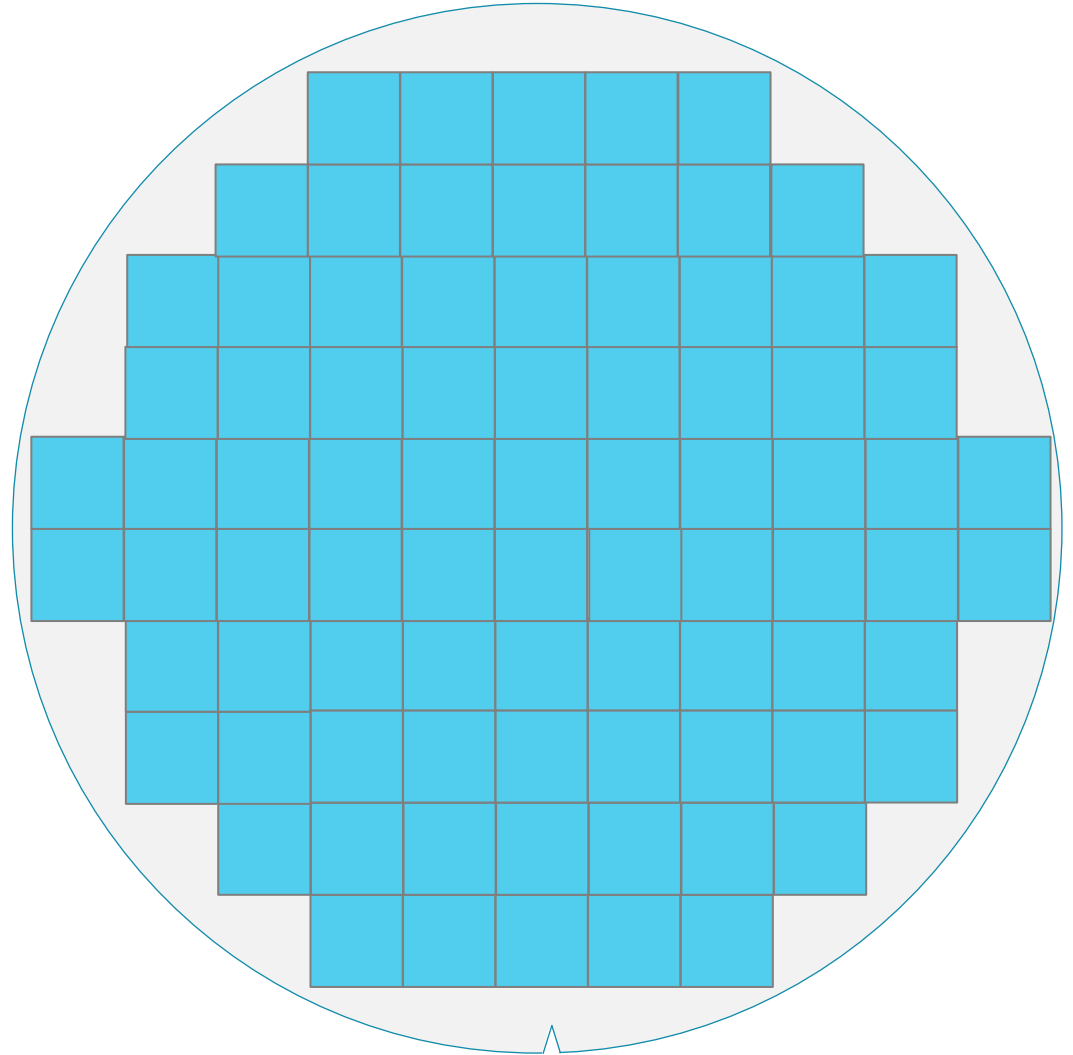
Mask cost reduction in the pipeline

Shot Reduction from New Manhattan

■ Shot count comparison

Nodes / MDP	Metal 1 #1	Metal 1 #2	Metal 1 #3
20nm /Conventional	635,005	678,823	2,286,066
20nm /MB-MDP	301,057	345,613	1,456,902
Reduction Rate	52.59%	49.09%	36.27%
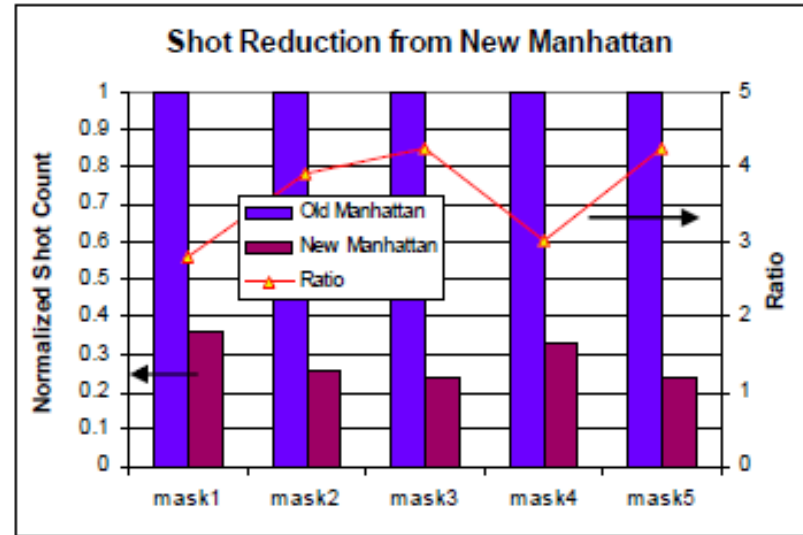
256k-APS (Aperture Plate System)

12

APS operated at -45 kV

5keV electron beam

Aperture Plate — *in-situ* adjustable in X,Y,Θ — Si

0V   3.3V   CMOS

Blanking Plate — with integrated CMOS electronics — Si

32μm   9 μm   4μm

200x reduction
⇩
20 nm

Slightly deflected beam blanked at last cross-over of projection optics
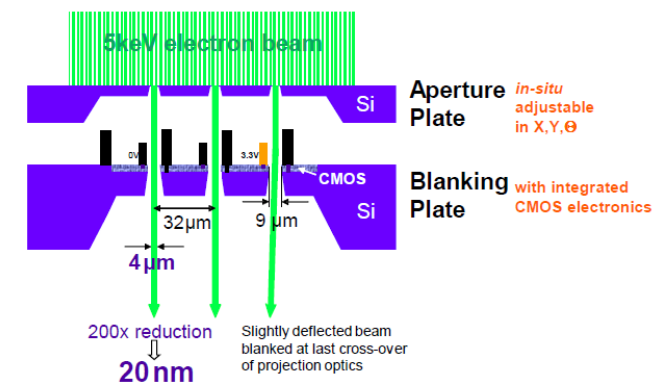
256k (k=1024) **= 262,144 programmable beams**
of
20nm beam size and 50keV beam energy
within
82μm x 82μm beam array field
at
6" mask substrate

IMS Nanofabrication

Tokyo Conference Center Shinagawa
June 26, 2012

SEMATECH Symposium Japan 2012

I E D M 2015

A cost per transistor scenario

IEDM 2015
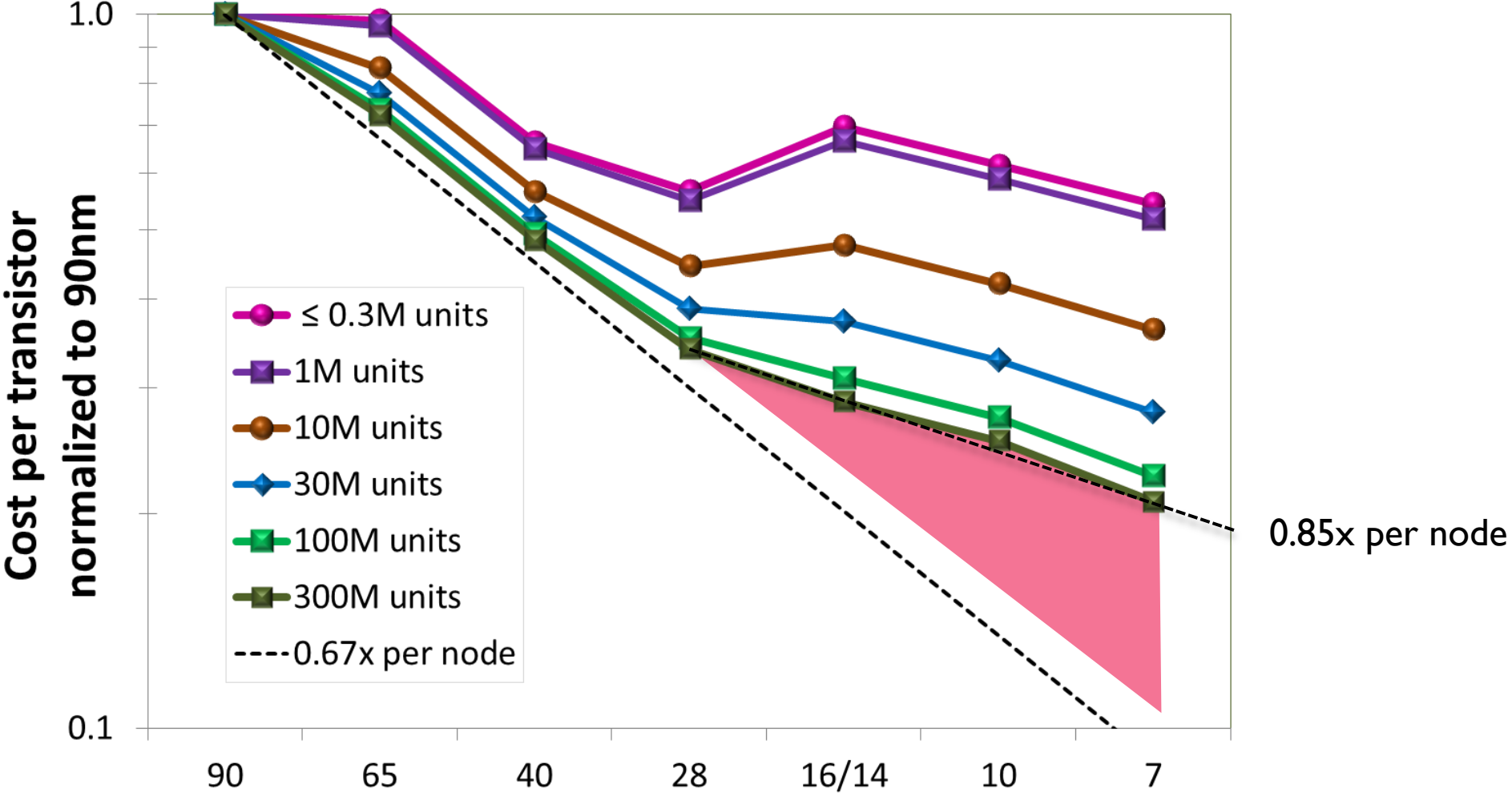
Flipping to cost per die (100mm^2 die)
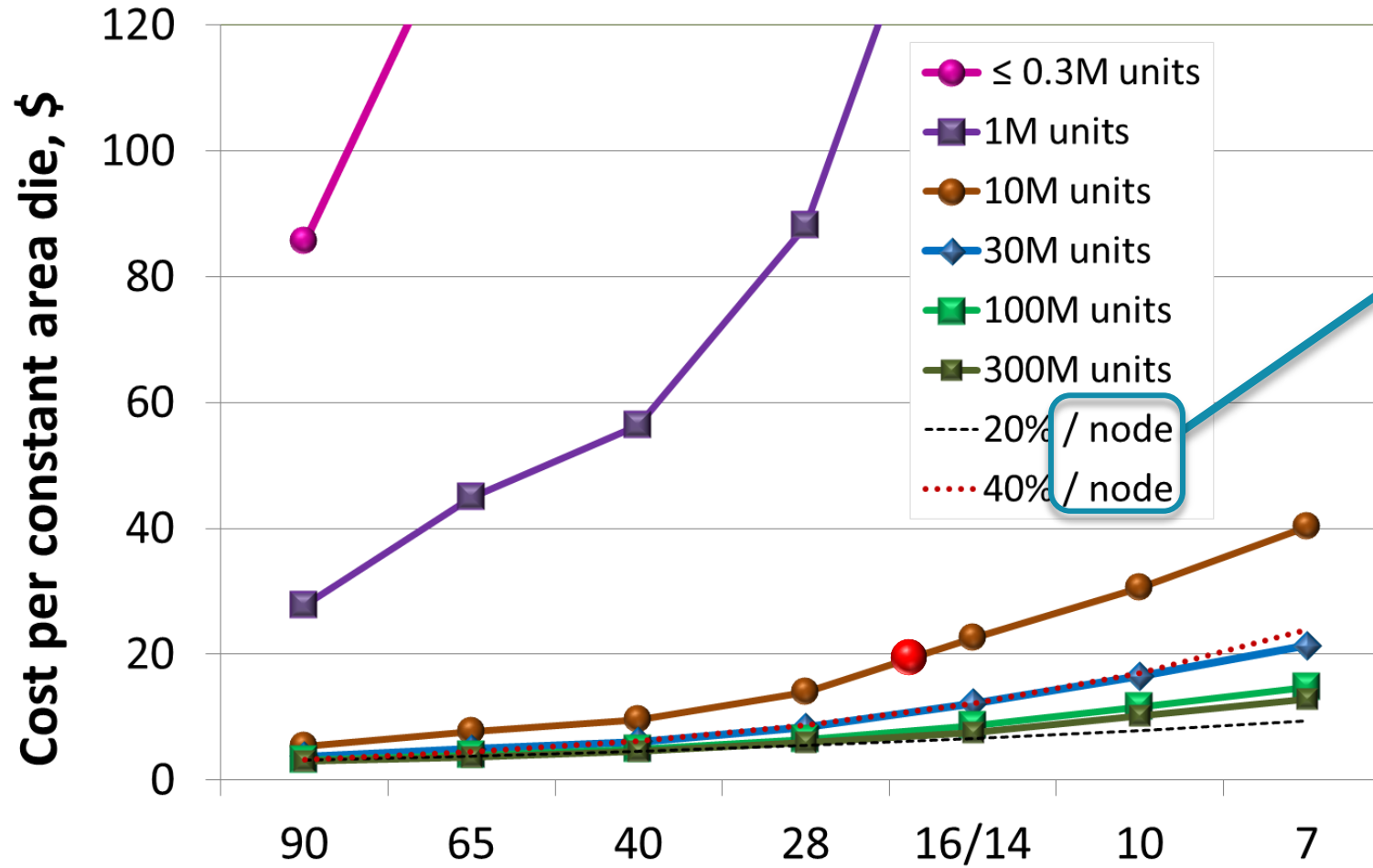
per node, not year
(typo in manuscript)

20nm / $20
Break-even is
10M units

Cost pressure on gate count
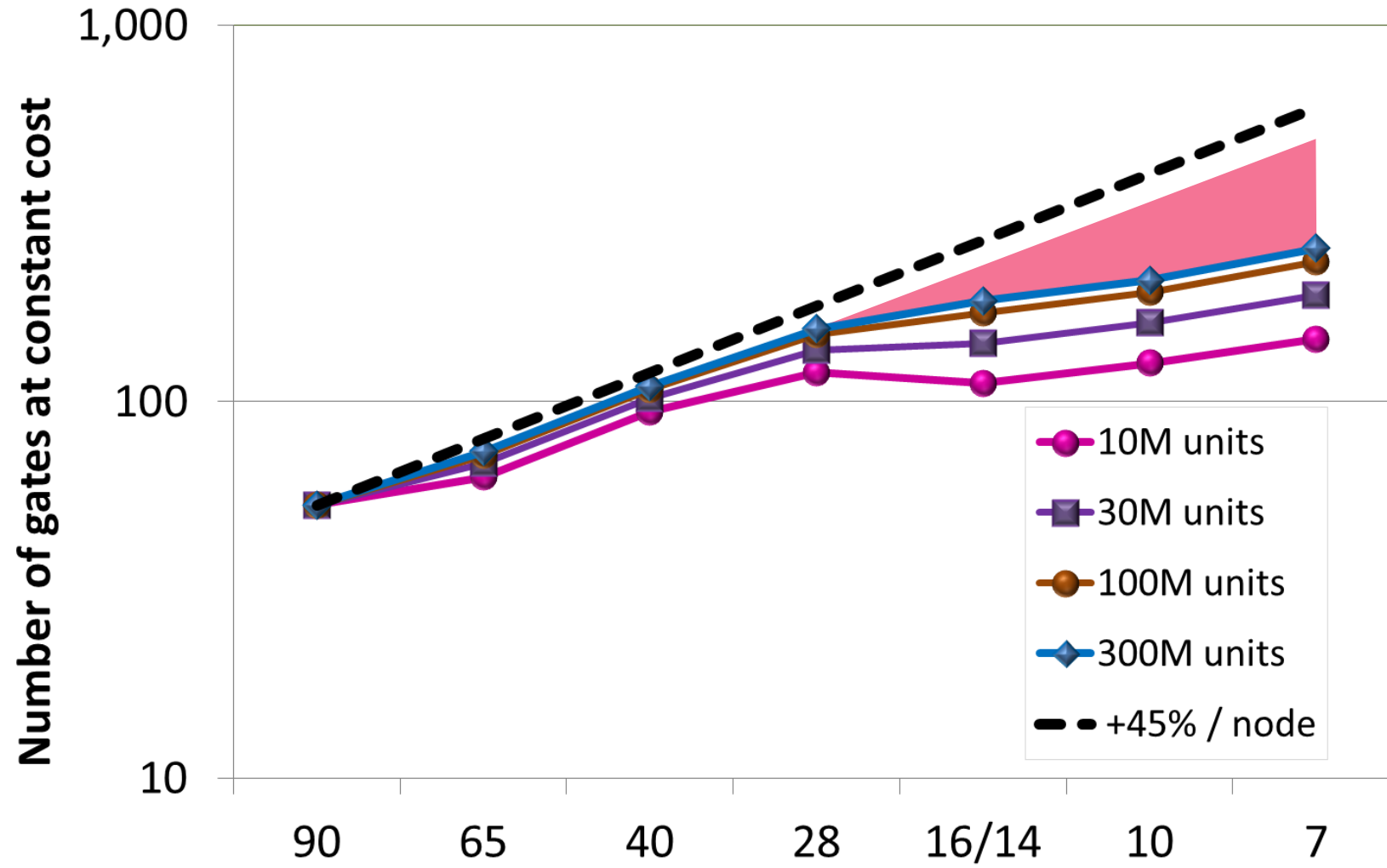
The Question is digital

Greg Yeric, ARM

2015
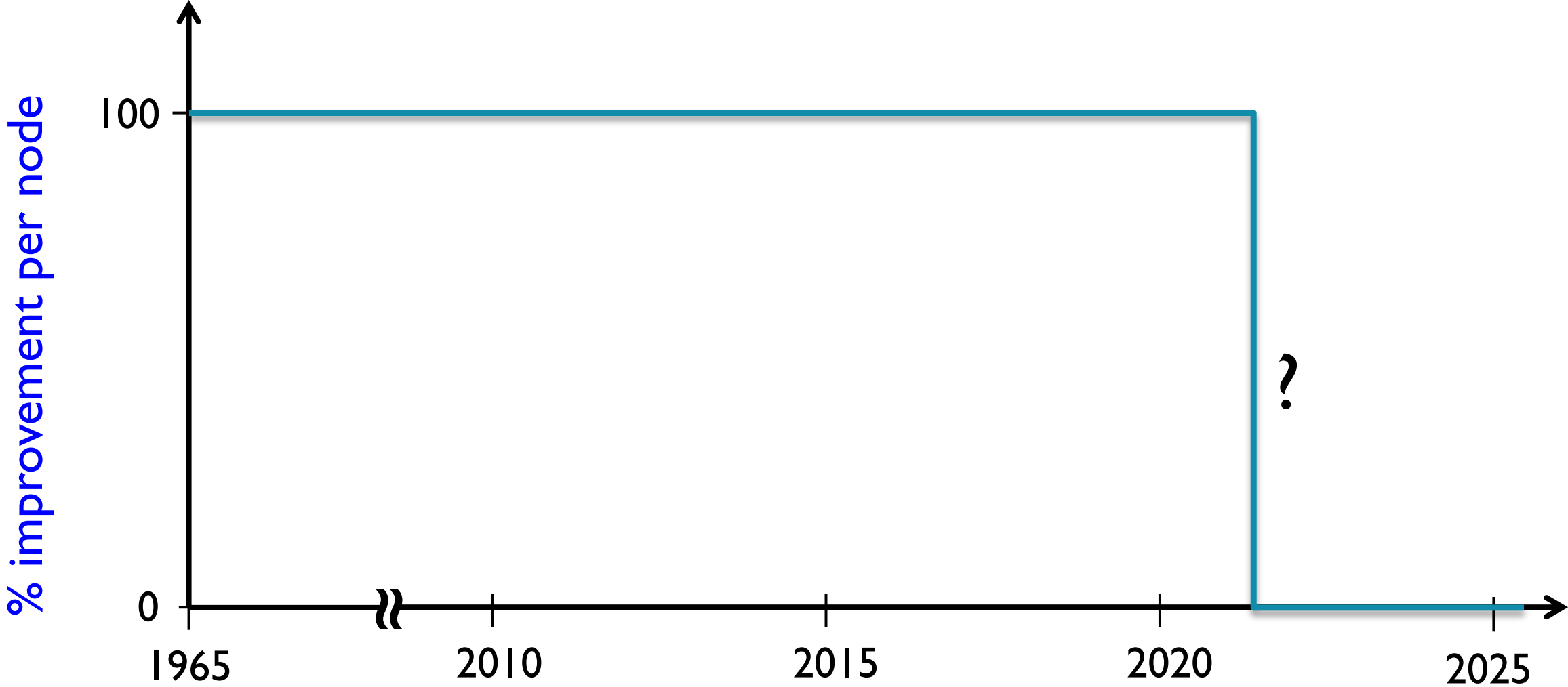
The Answer is analog

Can we do more with less (Moore)?

Legend:
- 10M units
- 30M units
- 100M units
- 300M units
- +45% / node

Y-axis: Number of gates at constant cost (10, 100, 1,000)
X-axis: 90, 65, 40, 28, 16/14, 10, 7
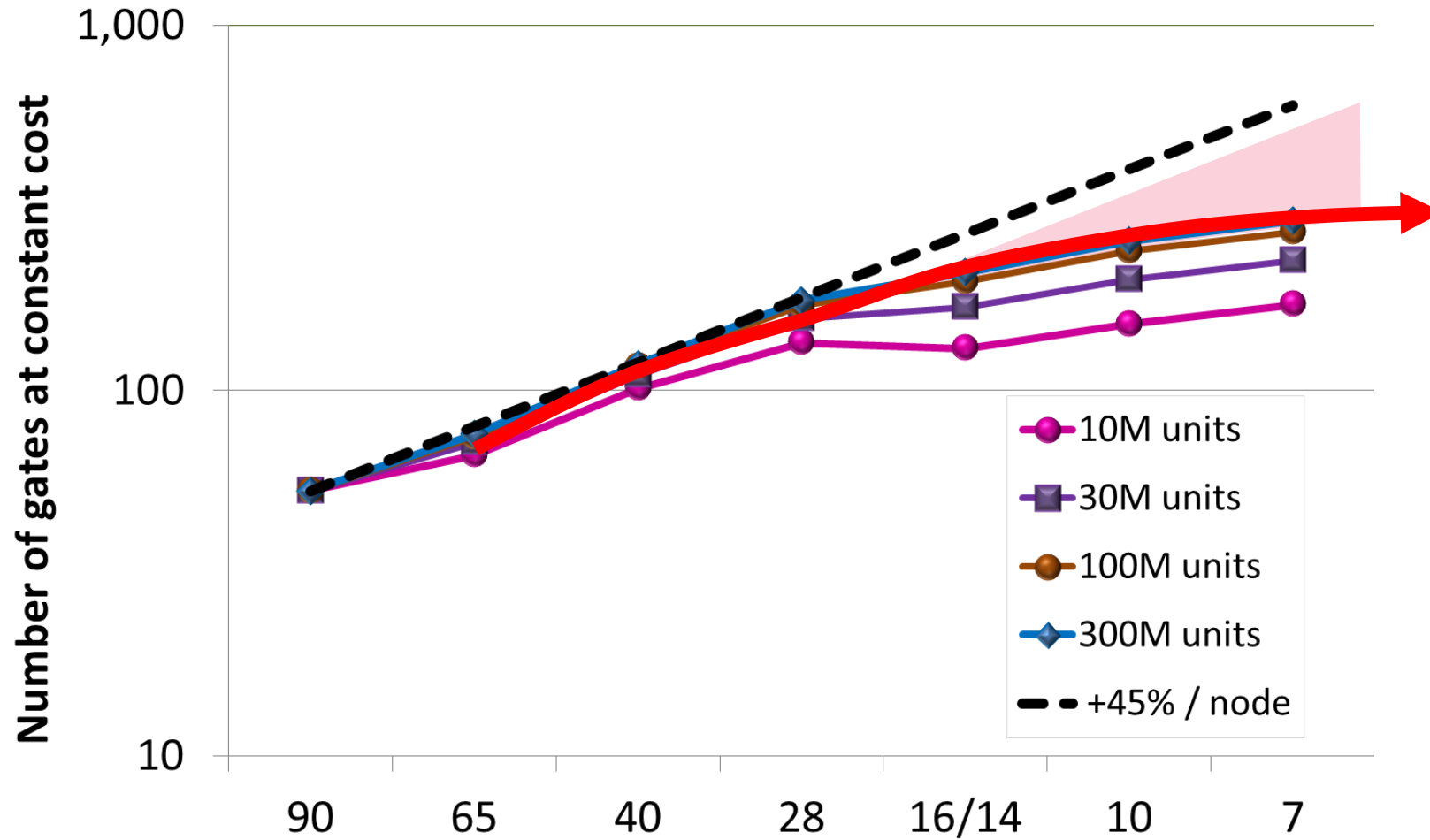
Today's systems: There is not just one problem

Can we do more with less (Moore)?

Number of gates at constant cost

- 10M units
- 30M units
- 100M units
- 300M units
- +45% / node

X-axis: 90, 65, 40, 28, 16/14, 10, 7
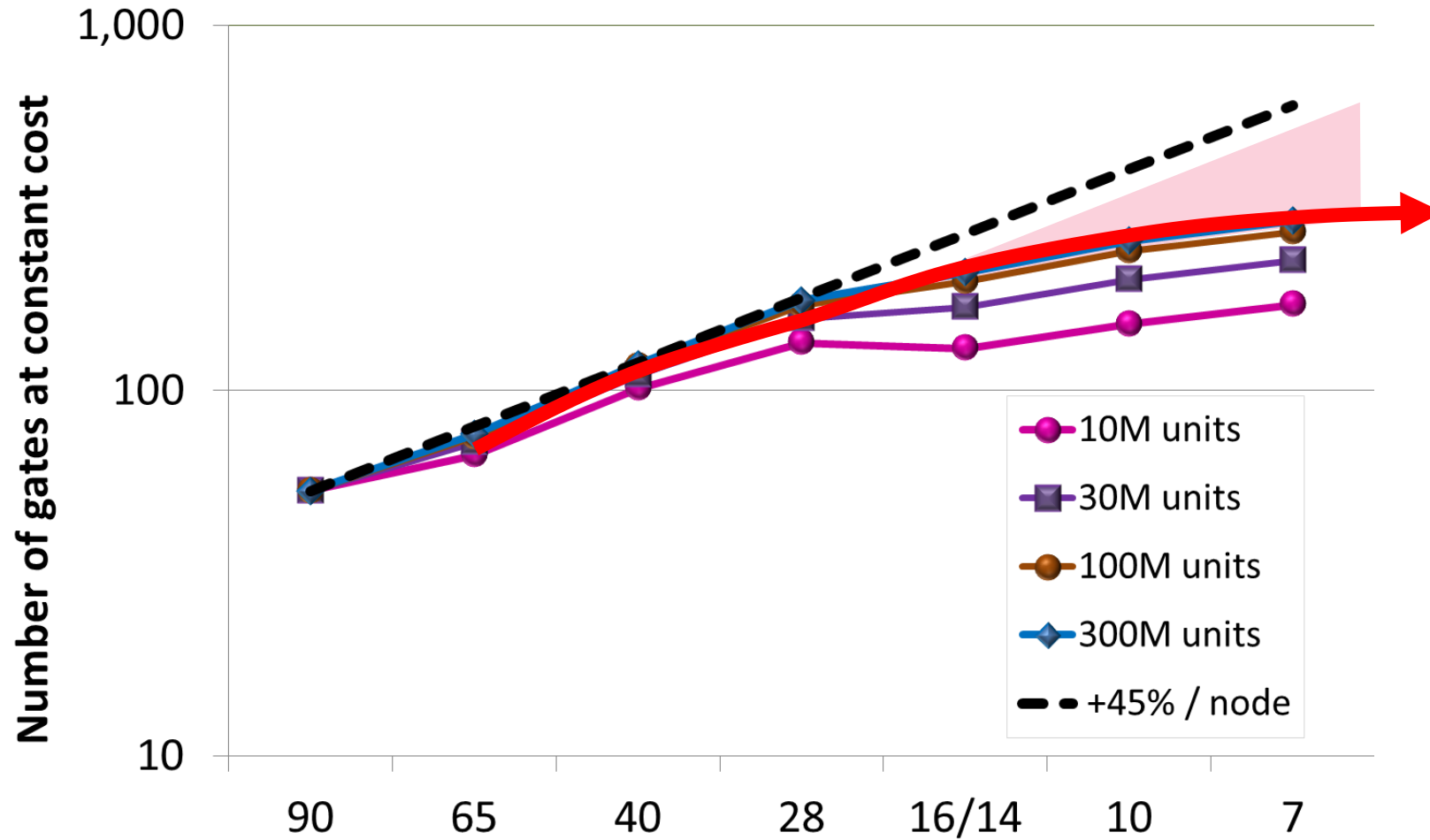
There are Three Problems

Cost

…and performance …

in a power budget

Electricity
Bill

Cooling
Bill

Battery
Capacity

Touch
Temperature

Form
Factor

Energy
Harvesting

Number of gates at constant cost (y-axis): 10, 100, 1,000

x-axis: 90, 65, 40, 28, 16/14, 10, 7

- ● 10M units
- ■ 30M units
- ● 100M units
- ◆ 300M units
- - - +45% / node
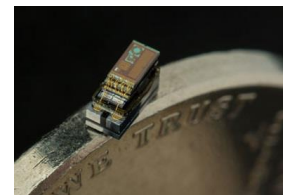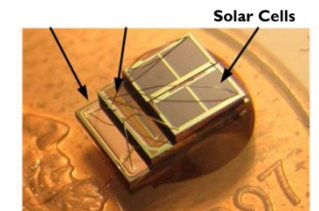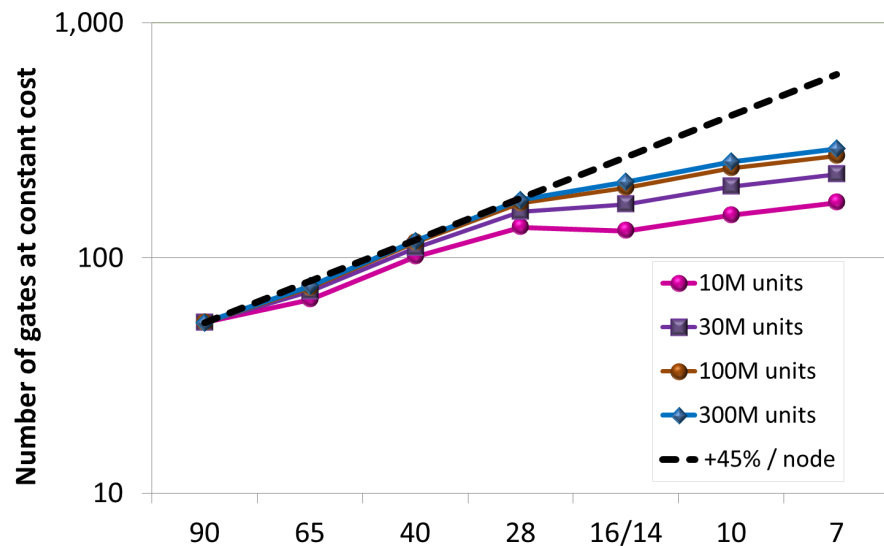
System scaling today:

1. Everything has a power budget,
2. But wants as much performance as possible

2009

2014

Display	5x
Camera	4x
Connectivity	20x
Sensors	3x
Video	34x
CPU	17x
GPU	40x
Memory Bandwidth	16x
Moore's Law	5-6x

Motorola Droid (2009)
By Creative Tools, Watermark removed by User:Ainali [CC BY 2.0 (http://creativecommons.org/licenses/by/2.0)], via Wikimedia Commons

R. Aitken, 2015 IEDM short course

Samsung Galaxy S5 (2014)
By GalaxyOptimus (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

With more transistors, improve performance within an energy budget:

- Multi-core parallelism
- Out-of-Order
- Branch prediction
- Pre-fetching
- Cache hierarchy complexity
- Multi-Threading

JOHN L. HENNESSY   DAVID A. PATTERSON

COMPUTER ARCHITECTURE
A Quantitative Approach

Interdependent chain to make end product

Logic cells

memory
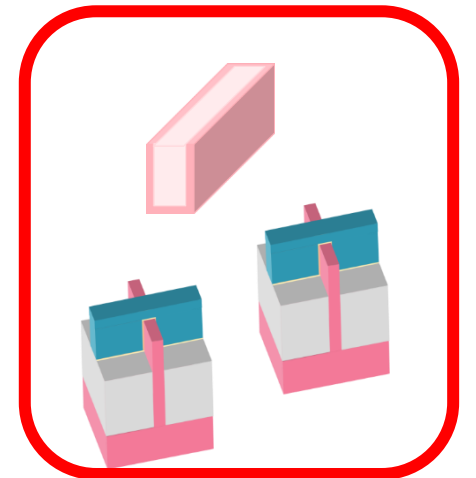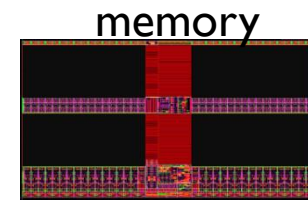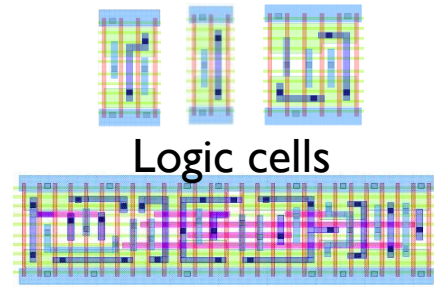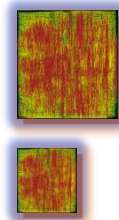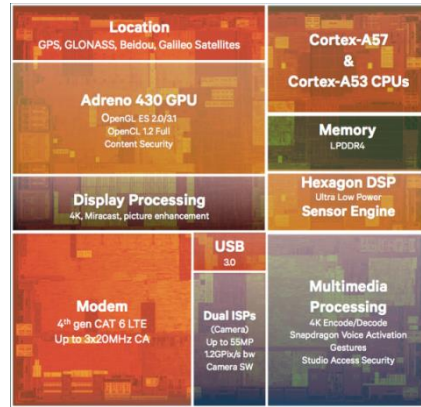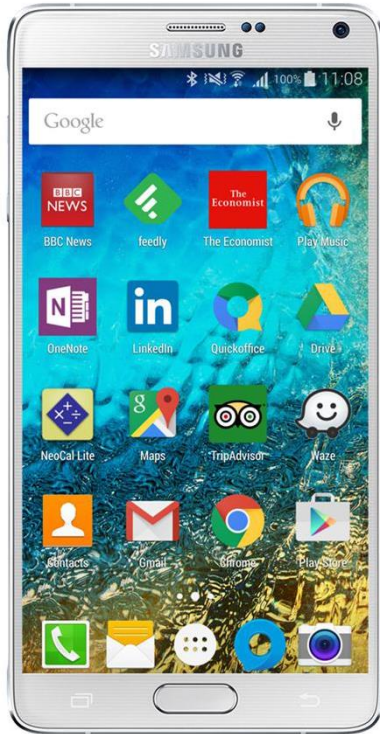
Next Generation Transistors at IEDM:
What you think they look like
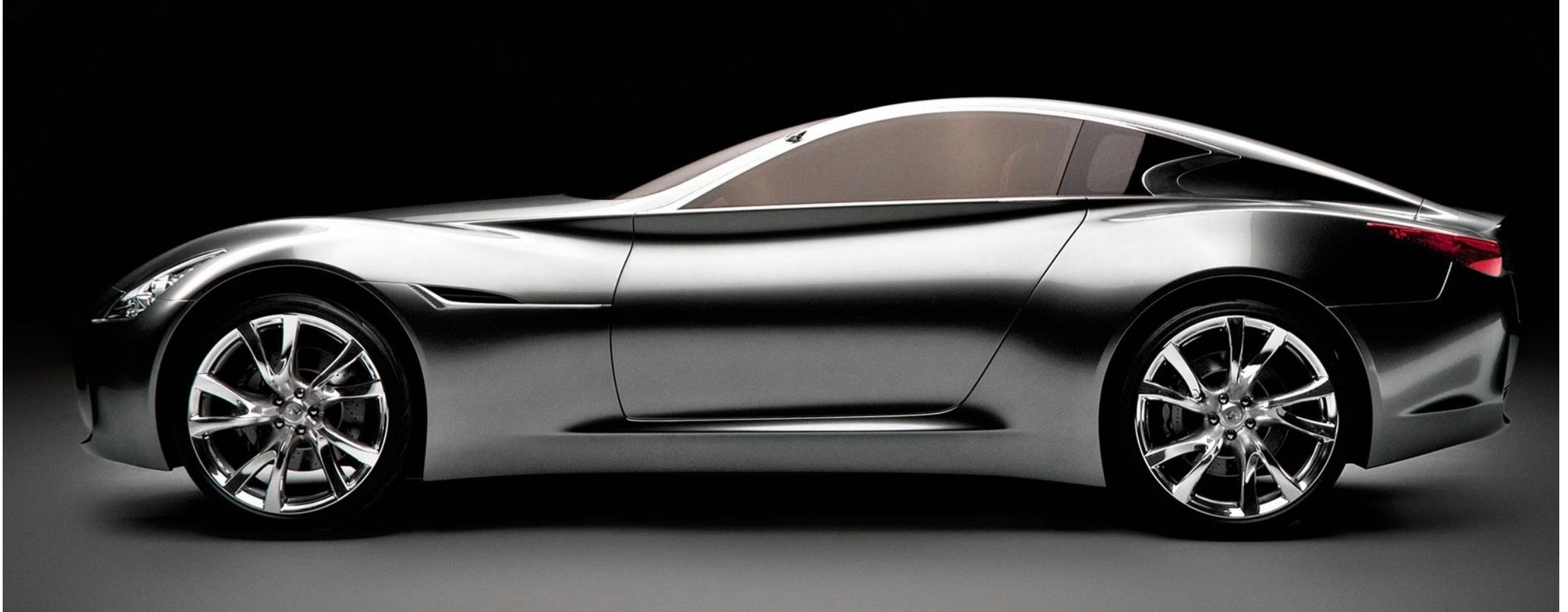
2.1   First Demonstration of Ge Nanowire CMOS Circuits:

2.4   Germanium-based Transistors for Future High Performance and Low Power Logic Applications

IEDM 2015

Next Generation Transistors at IEDM:
What they actually look like
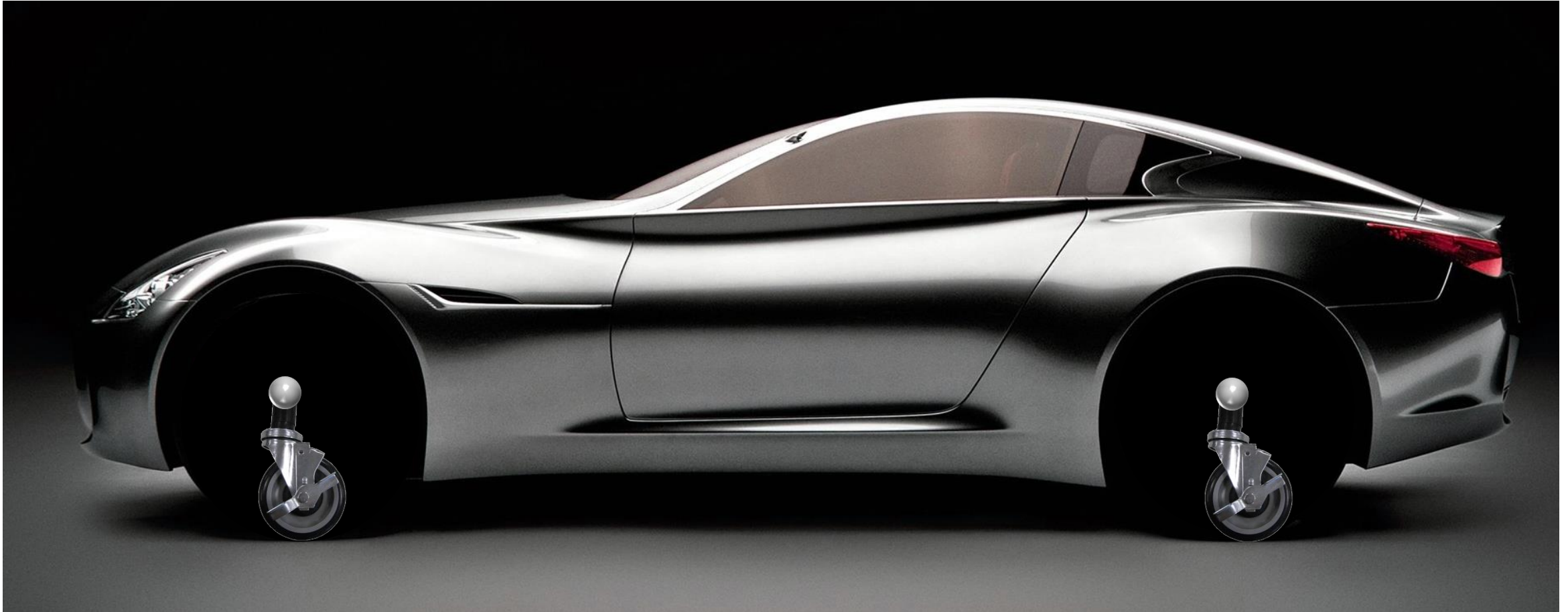
Evening Panel: Emerging Devices – Will they solve the bottlenecks of CMOS?

The shrinking FET

Fig. 6. Breakdown of intrinsic & extrinsic switching capacitances for typical post-28nm simple logic gates (INV and NAND), post-layout parasitic extraction and simulated with SPICE.

IMEC: IEDM 2013

Intrinsic FET capacitance is a minority!
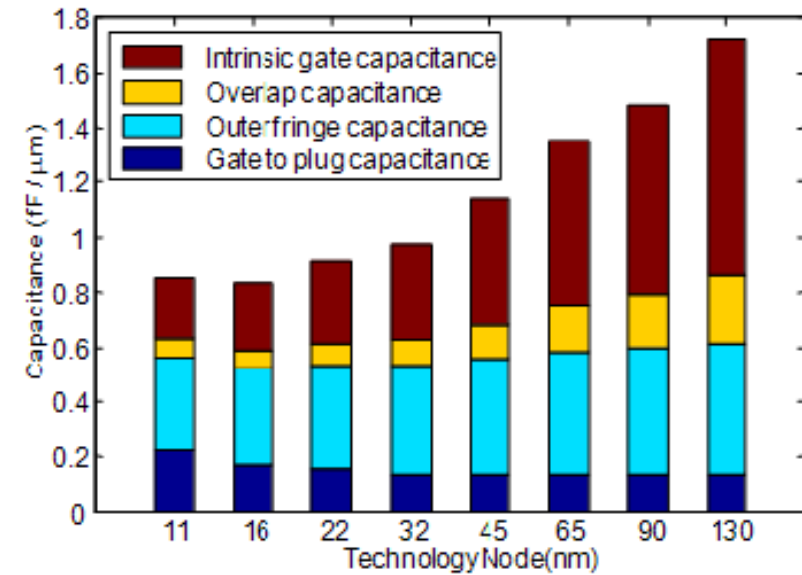
HS Wong, et al., SISPAD 2009

IEDM 2015

The shrinking FET
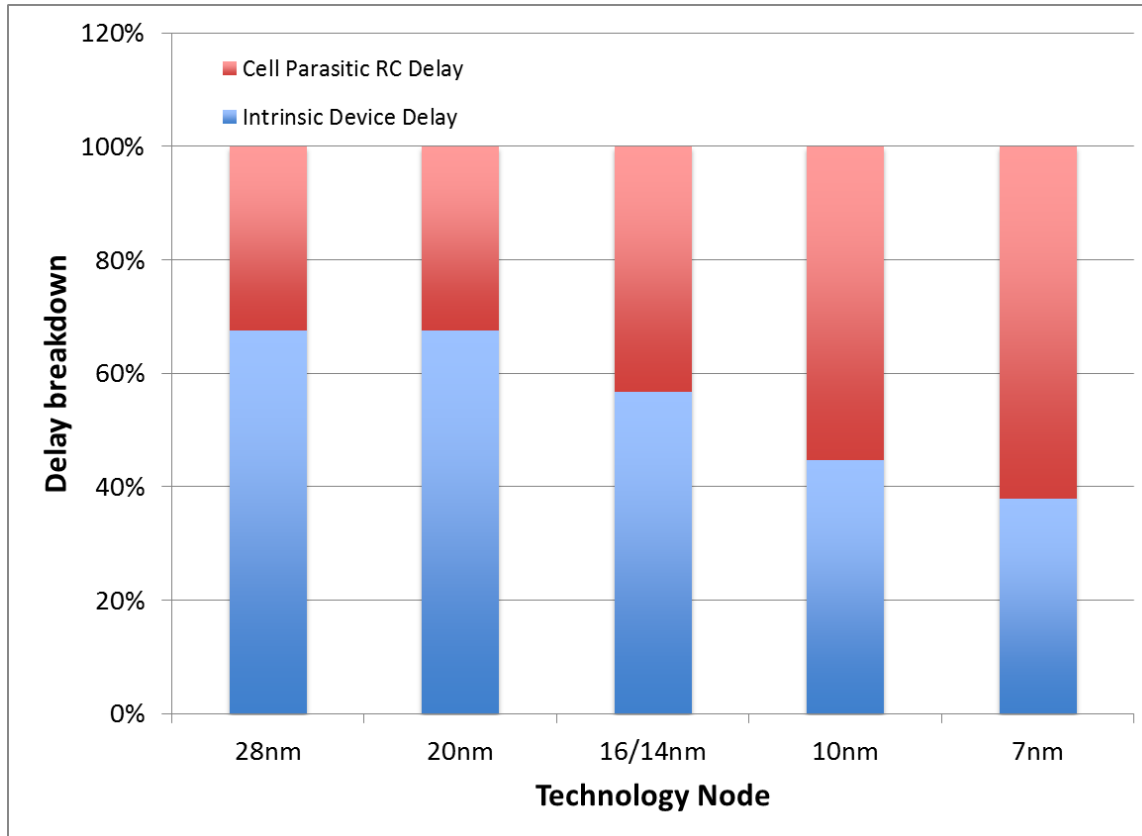
Electrostatics for gate length reduction:
- previously beneficial in reducing C
- now beneficial in reducing R_C

This is now larger than this
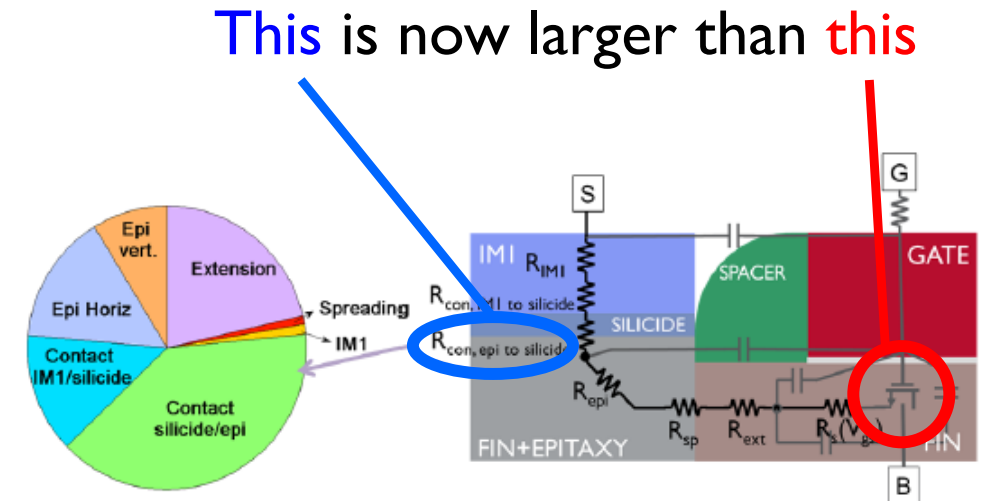
Fig. 12. Transconductance gm of an n-type 10nm-Node FinFET as a function of gate bias evaluated with the processing assumptions from the

21.7 1.5×10^9 Ωcm^2 Contact Resistivity on Highly Doped Si:P Using Ge Pre-amorphization and Ti Silicidation

Non-scaling parasitics

μ-enhanced, big fin height FET

CV^2f

C

TFET

CV^2f

V/I

C
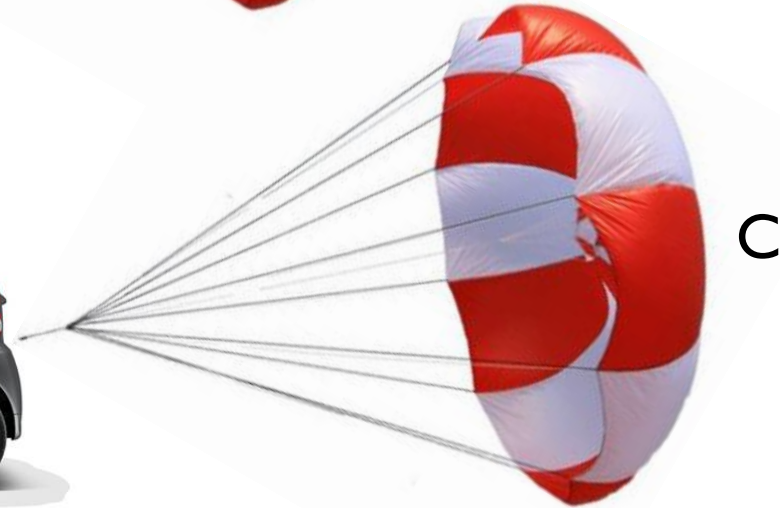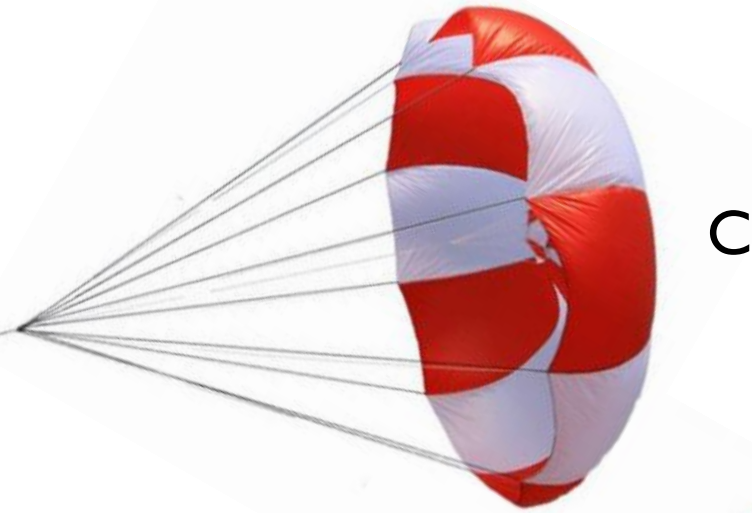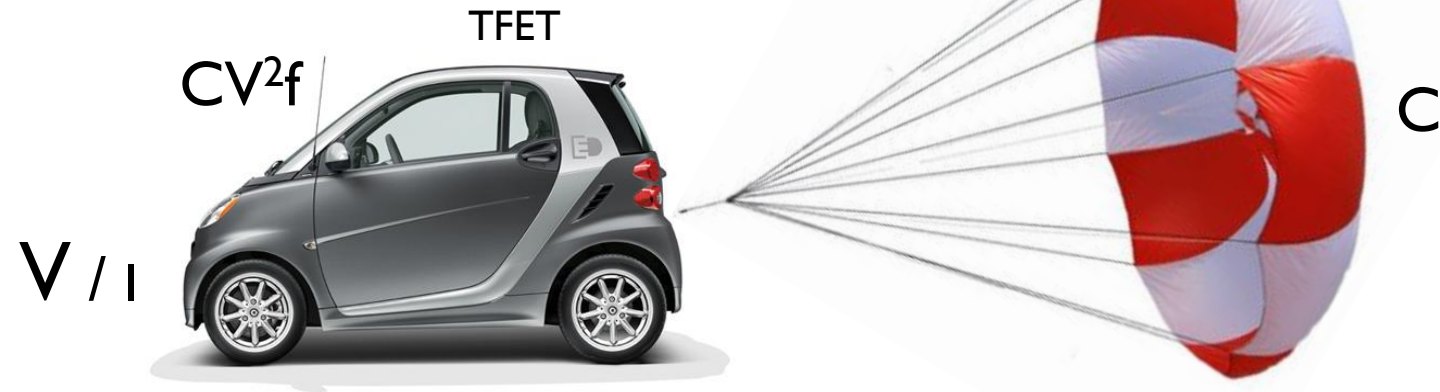
Session 22: Nano Device Technology – **Steep Slope Transistors**

22.2   First Foundry Platform of Complementary Tunnel-FETs in CMOS Baseline Technology for Ultralow-Power IoT Applications

22.6   Sub-60mV-Swing Negative-Capacitance FinFET without Hysteresis

FinFET performance scaling secret: "3D factor"

Planar → 1st generation FinFET → 2nd generation FinFET:
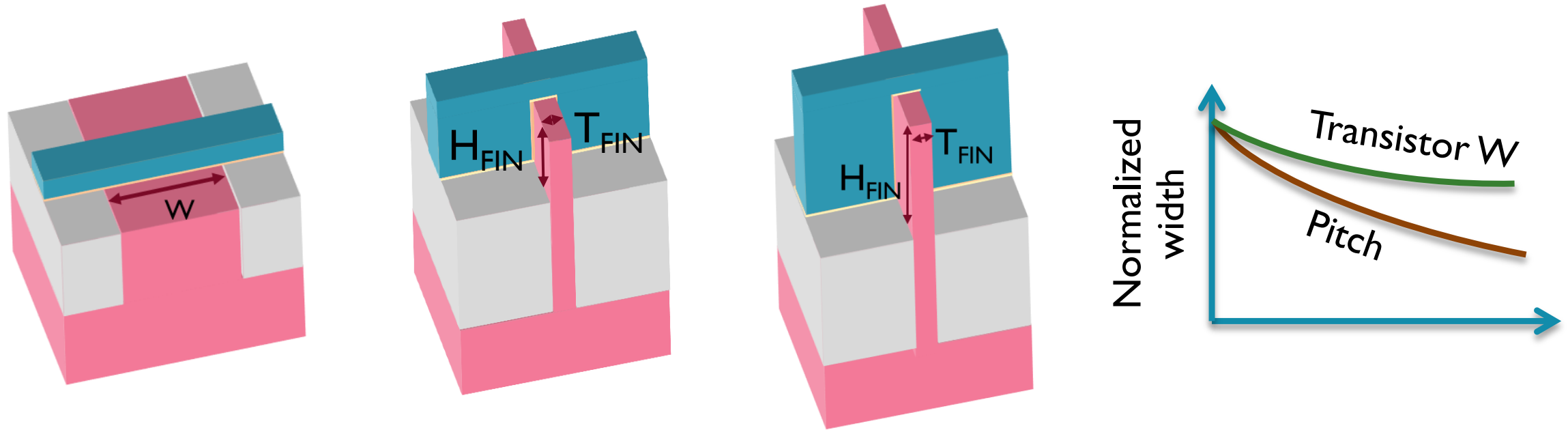
- Part of performance is improved electrostatics
- But part is simply more folded width
- "fin depopulation": Can't design with one-fin devices
- More 3D = more parasitics

Gate All-Around Horizontal Nanowire
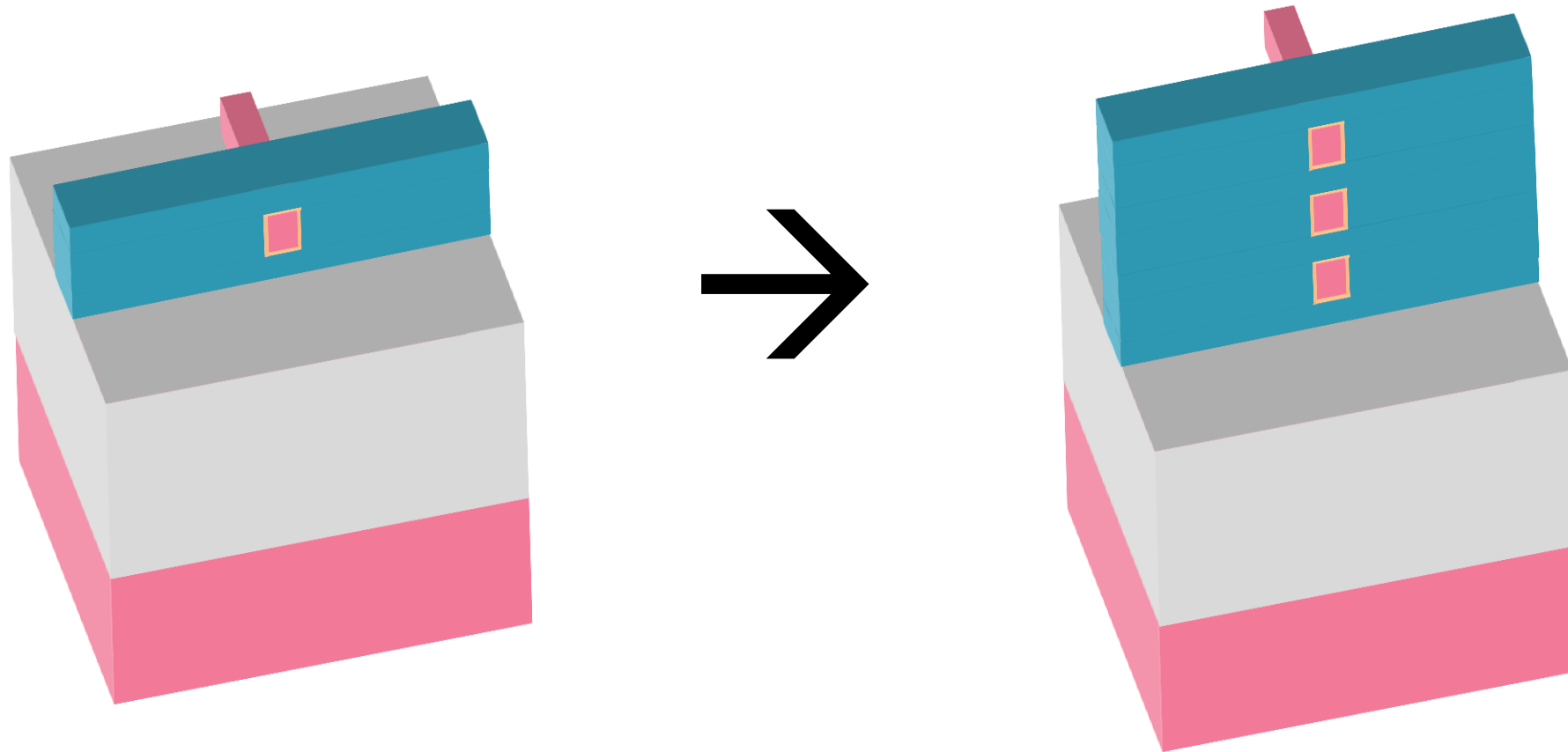
IEDM 2015

Chasing fin height, or back to 2D?
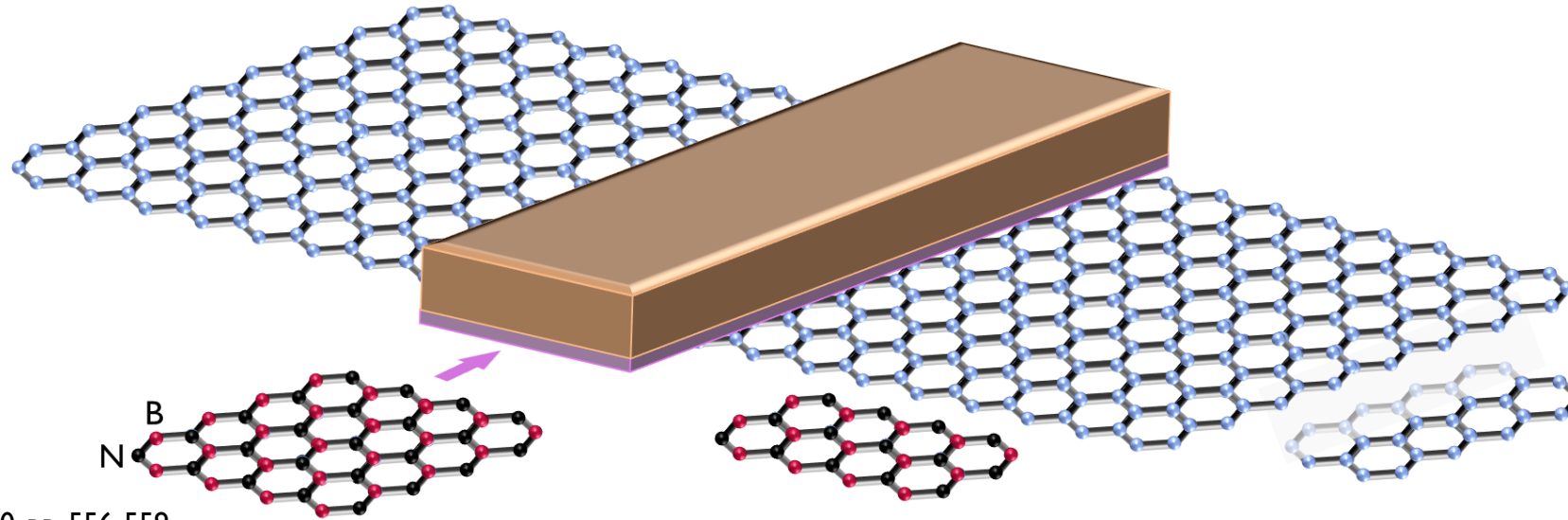
Yang, et al., GLSVLSI 2010, pp 263-268

B
N

Meric, et al. IEDM 2010, pp. 556-559

Hunt, et al. *Science*, June 2013, v340, n6139

Liu, et al., Nature Nanotechnology 8, 2013, pp 119-124

12.1   Phonon-Limited Performance of Single-Layer, Single-Gate Black Phosphorus n- and p-type Field-Effect Transistors,

12.3   Designing Band-to-Band Tunneling Field-Effect Transistors with 2D Semiconductors for Next-Generation Low-Power VLSI,

12.4   How Good is Mono-Layer Transition-Metal Dichalcogenide Tunnel Field-Effect Transistors in sub-10 nm? -

12.7   Understanding the Nature of Metal-Graphene Contacts: A Theoretical and Experimental Study,

Real V scaling: to the mV's:

μ-enhanced, big fin height FET

CV²f

C

TFET

CV²f

mV / I

C
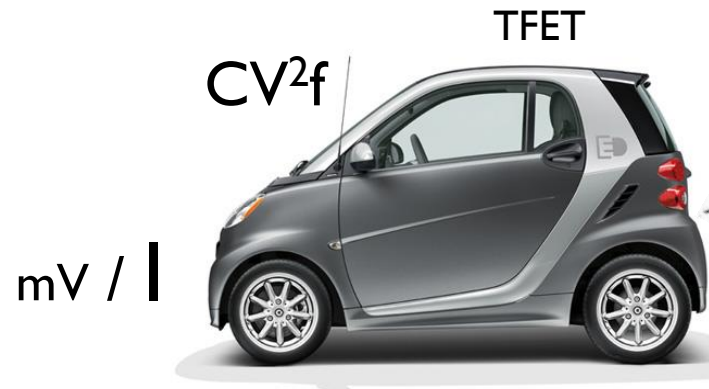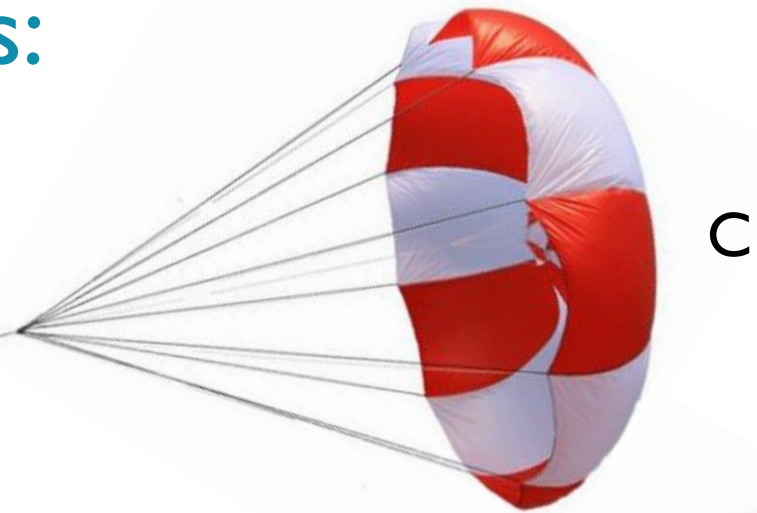
32.5   Spintronic Majority Gates

32.6   Spintronic Logic Circuit and Device Prototypes

150˚C?
Power networks?
Signal integrity?

2015

Greg Yeric, ARM

Moore's Law and Dennard scaling

"shrinking dimensions on an integrated structure makes it possible to operate the structure at higher speed for the same power per unit area" [1]

Table 1
Scaling Results for Circuit Performance

Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	$1/\kappa$
Doping concentration N_a	κ
Voltage V	$1/\kappa$
Current I	$1/\kappa$
Capacitance $\epsilon A/t$	$1/\kappa$
Delay time/circuit VC/I	$1/\kappa$
Power dissipation/circuit VI	$1/\kappa^2$
Power density VI/A	1

Also from The Economist: September 2015

Compactness is less important for another fast-growing branch of the information-technology business, cloud computing. In cloud-service providers' cavernous data centres, space is not at a premium, the way it is inside the latest iPhone. What increasingly matters most to cloud providers is energy efficiency: how much power their racks of servers consume, and how they can keep them sufficiently cool to ensure that their chips do not fry.

Fortunately, one of the corollaries of Moore's law is that the energy efficiency of transistors follows the same exponential law, doubling around every two years. And like the law itself, it's not quite dead yet.

2015

Dark Silicon

We get more transistors, we just can't afford to turn them all on

Power Density
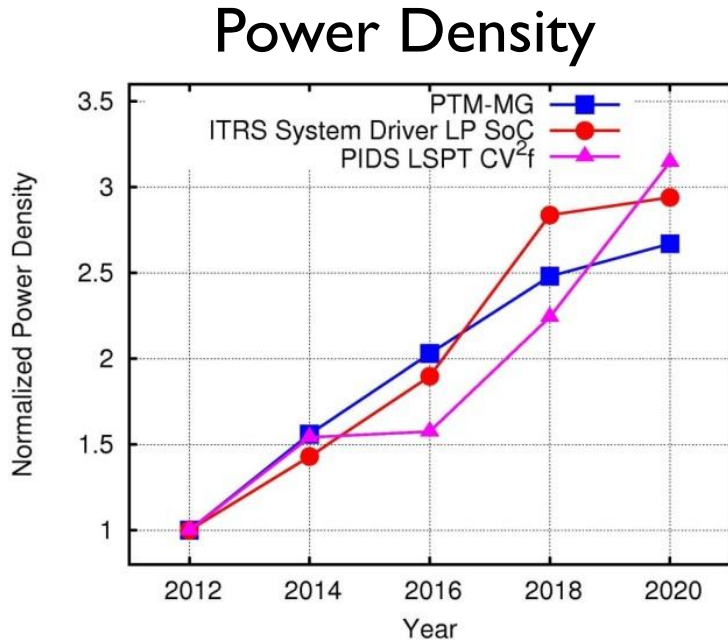
Sinha, Cline, Yeric, Chandra, Cao*, ISLPED 2012

5nm	80%
7nm	75%
10nm	56%
14/16nm	45%
20nm	33%

28nm: 0
(reference)
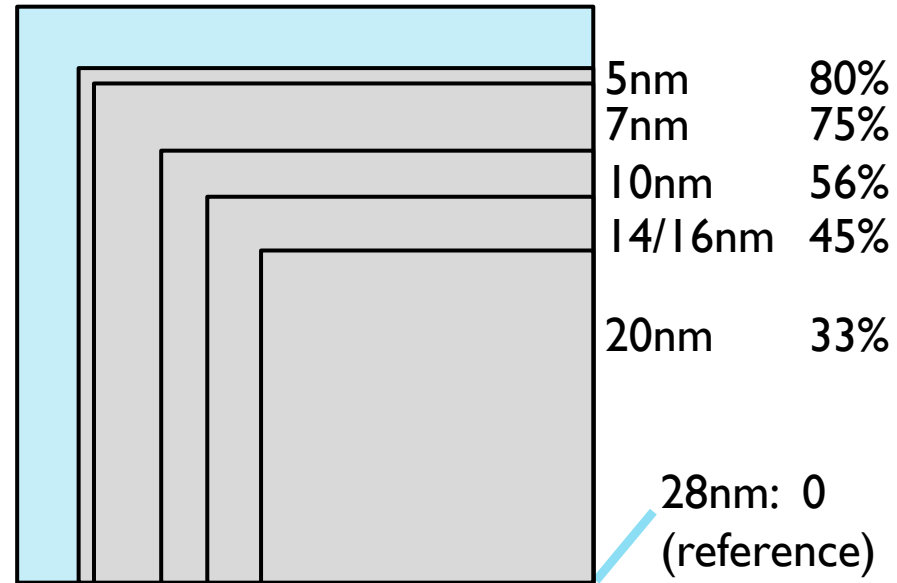
Design effort reduces effect on products:
- Power and clock gating
- Dynamic Voltage/Frequency Scaling (DVFS)
- Multi-core
- Memory assist

} Spend transistors to buy power

Can we get more out of less (Moore)?
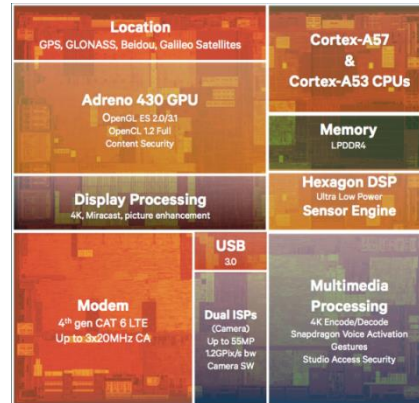
Design-Technology
Co-Optimization
(DTCO)

Example: Qualcomm 810

Location
GPS, GLONASS, Beidou, Galileo Satellites

Cortex-A57
&
Cortex-A53 CPUs

Adreno 430 GPU
OpenGL ES 2.0/3.1
OpenCL 1.2 Full
Content Security

Memory
LPDDR4

Display Processing
4K, Miracast, picture enhancement

Hexagon DSP
Ultra Low Power
Sensor Engine

USB
3.0

Modem
4th gen CAT 6 LTE
Up to 3x20MHz CA

Dual ISPs
(Camera)
Up to 55MP
1.2GPix/s bw
Camera SW

Multimedia
Processing
4K Encode/Decode
Snapdragon Voice Activation
Gestures
Studio Access Security

Logic cells

memory

Gate pitch, logic density, and frequency

SP&R

SP&R

Table 1: Summary of Process Variations Benchmarked

Process Version	Relative Poly Pitch	Other Features
P1	1	-
P2	1.076	-
P3	1.153	-
P4	1.076	2 pitch smaller low drive flops

15% larger gate pitch

M. Frederick, IEDM 2014

P1 Slowest
P1 Fastest
P2 Slowest
P2 Fastest
P3 Slowest
P3 Fastest
P4 Slowest
P4 Fastest

Area

Performance

Figure 6: Block area versus performance results for all available devices for each process. ...d possible using the slowest and fastest

5% smaller block area

Strain volume, contact-gate capacitance
(and complicated litho issues)

Gate pitch, logic density, and frequency

- Inverse Moore Scaling!

- Transistor performance can be traded for area reduction

- Perceived value of performance vs. area is dependent on application targets

Table 1: Summary of Process Variations Benchmarked

Process Version	Relative Poly Pitch	Other Features
P1	1	-
P2	1.076	-
P3	1.153	-
P4	1.076	2 pitch smaller low drive flops

15% larger gate pitch

M. Frederick, IEDM 2014

SP&R

Area

Performance

P1 Slowest
P1 Fastest
P2 Slowest
P2 Fastest
P3 Slowest
P3 Fastest
P4 Slowest
P4 Fastest

5% smaller block area

Figure 6: Block area versus performance results for all available devices for each process. ...d possible using the slowest and fastest
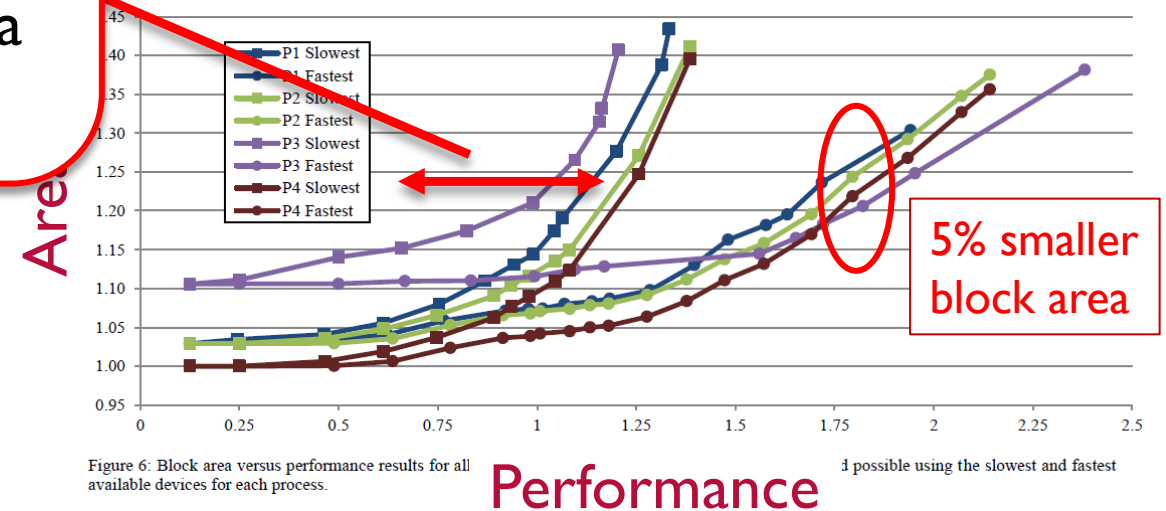
Strain volume, contact-gate capacitance (and complicated litho issues)

BEOL cost: Good news for FEOL (?)
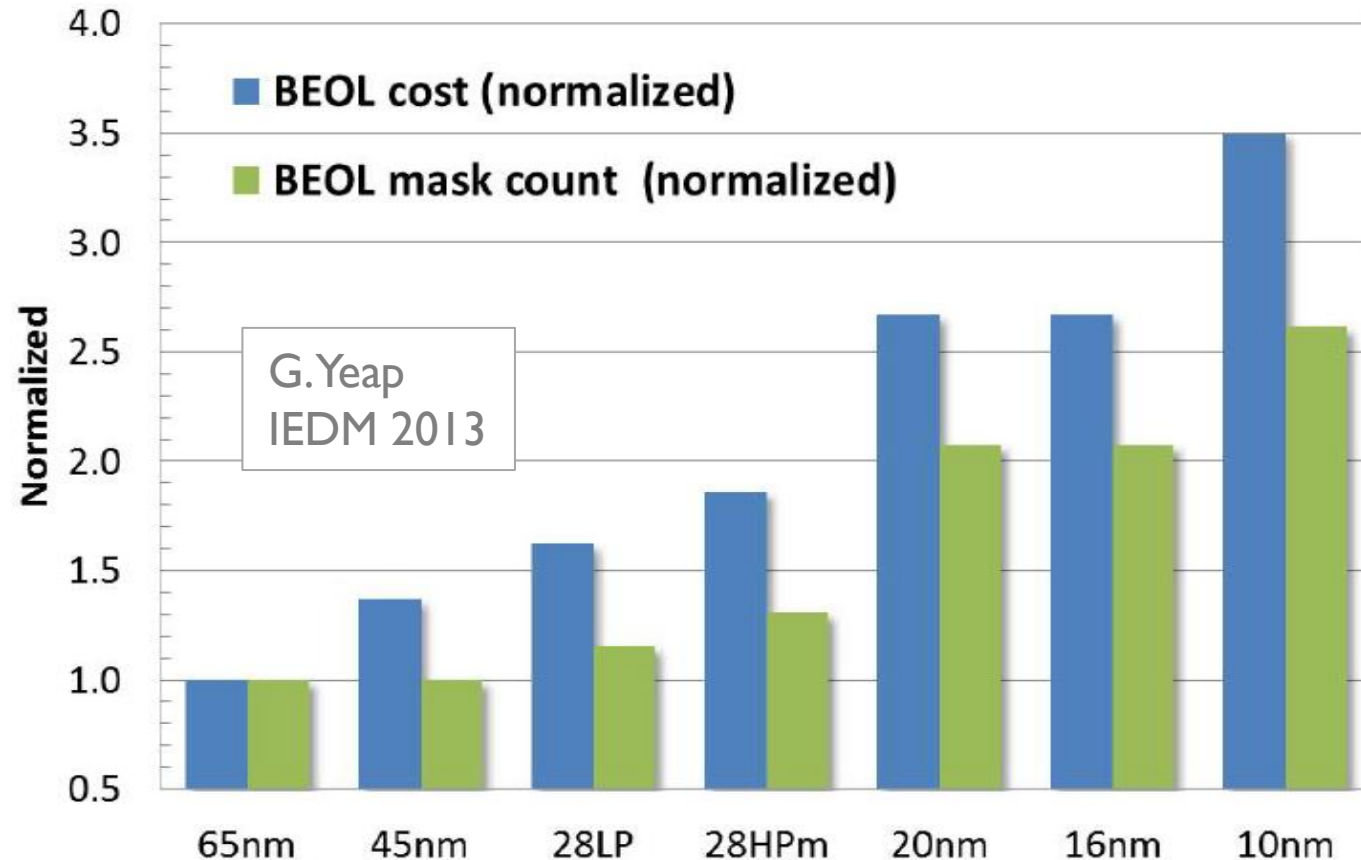
Retirement planning advice:

diversify your investments

Can we support multiple gate pitches on one die?

SP&R

SP&R

memory

GPU

Greg Yeric, ARM
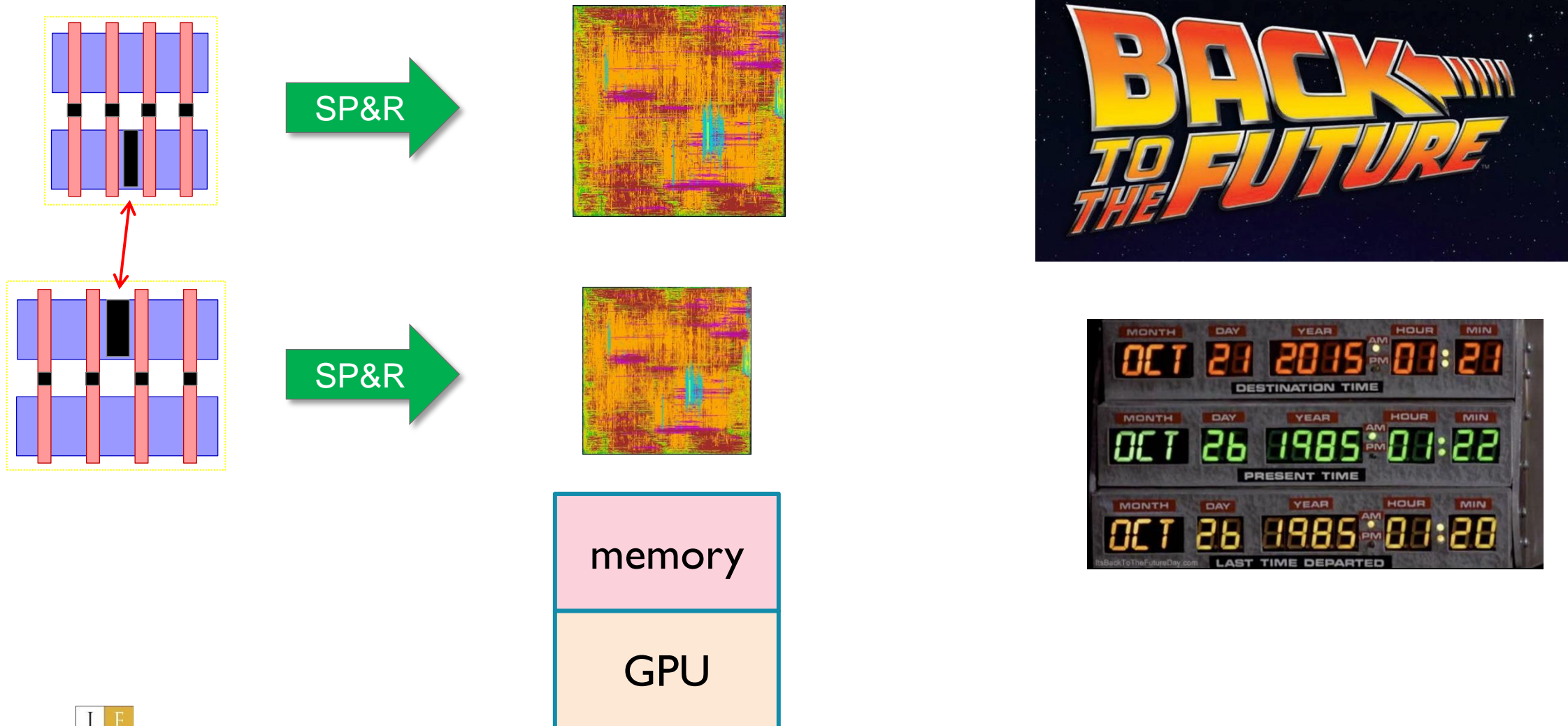
The lost half node

Deep sub-λ lithography has forced regularity. This has cost density.
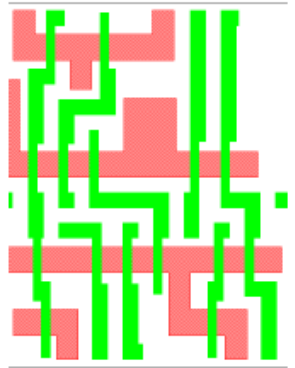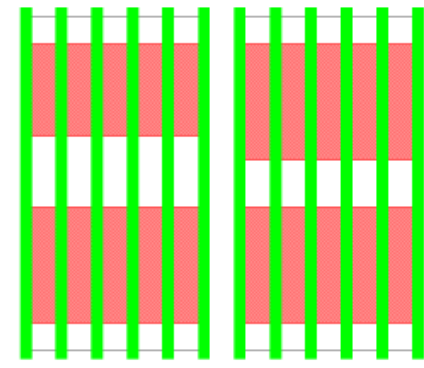
65nm
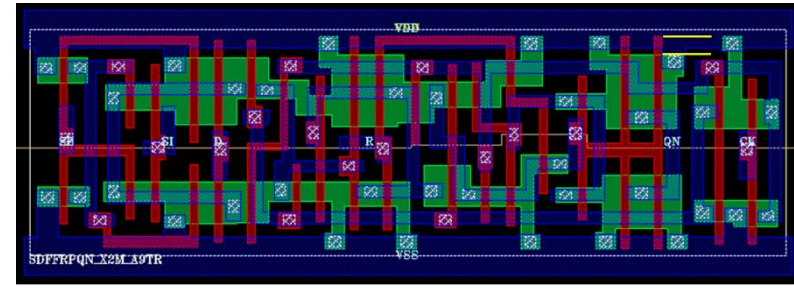
32nm
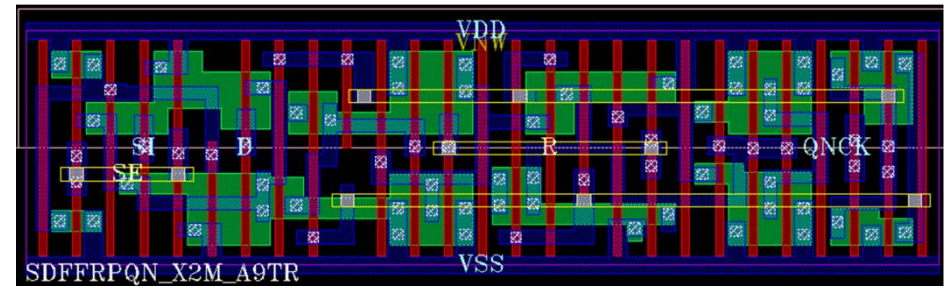
Liebmann, Pietromonaco, SPIE_8684_12

Middle Of Line (MOL)

- Cost bump: 28 → 16/14
- Adds RC
- Adds gate efficiency pressure

~8 masks at 7nm
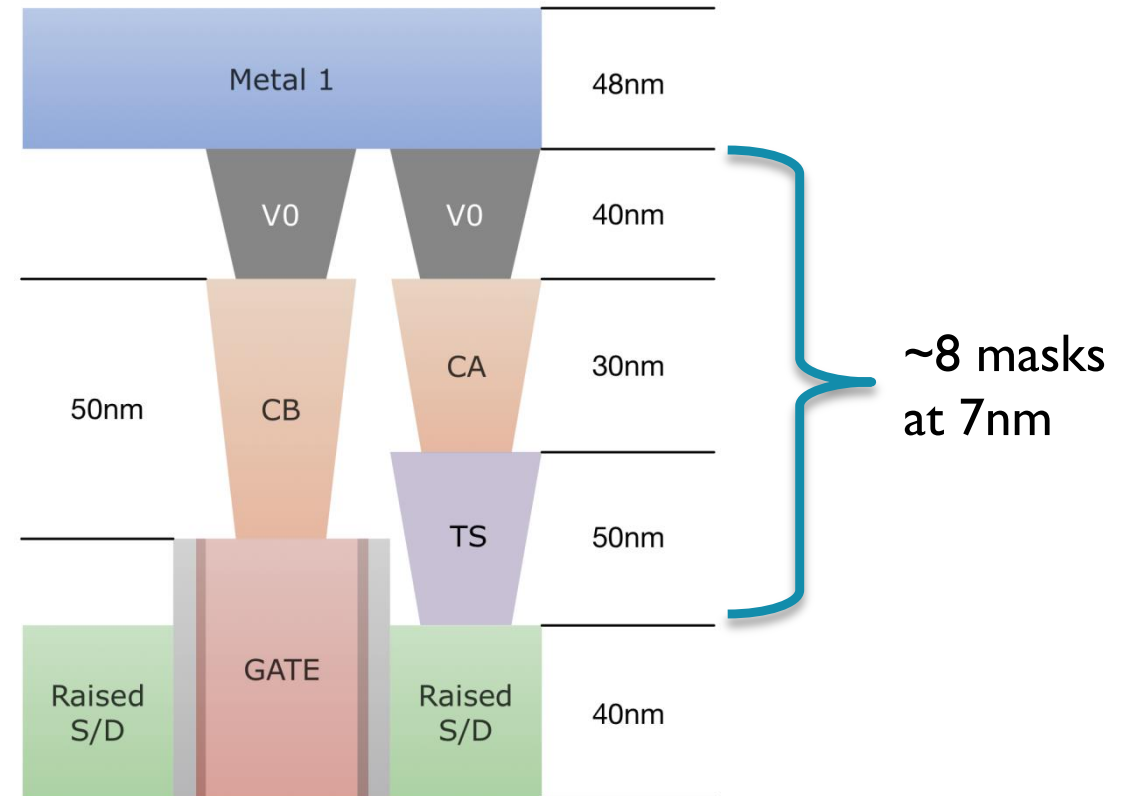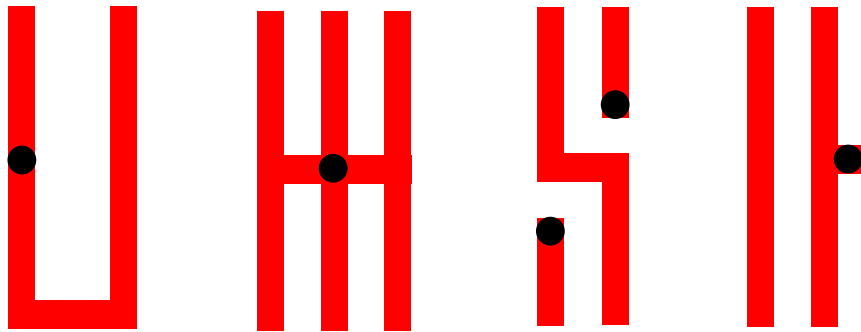
Middle Of Line (MOL)

- Anytime you can avoid a V0, you free up routing resources, and reduce chip size
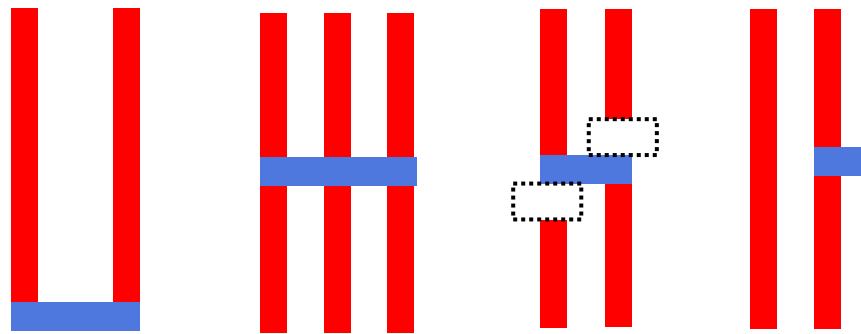
■ 1st pass gate

▬ 2nd pass gate (does not have to be ALD WF, just a metal)

Metal 1 — 48nm

V0   V0 — 40nm

CA — 30nm

CB   50nm

TS — 50nm

GATE

Raised S/D   Raised S/D — 40nm

~8 masks at 7nm
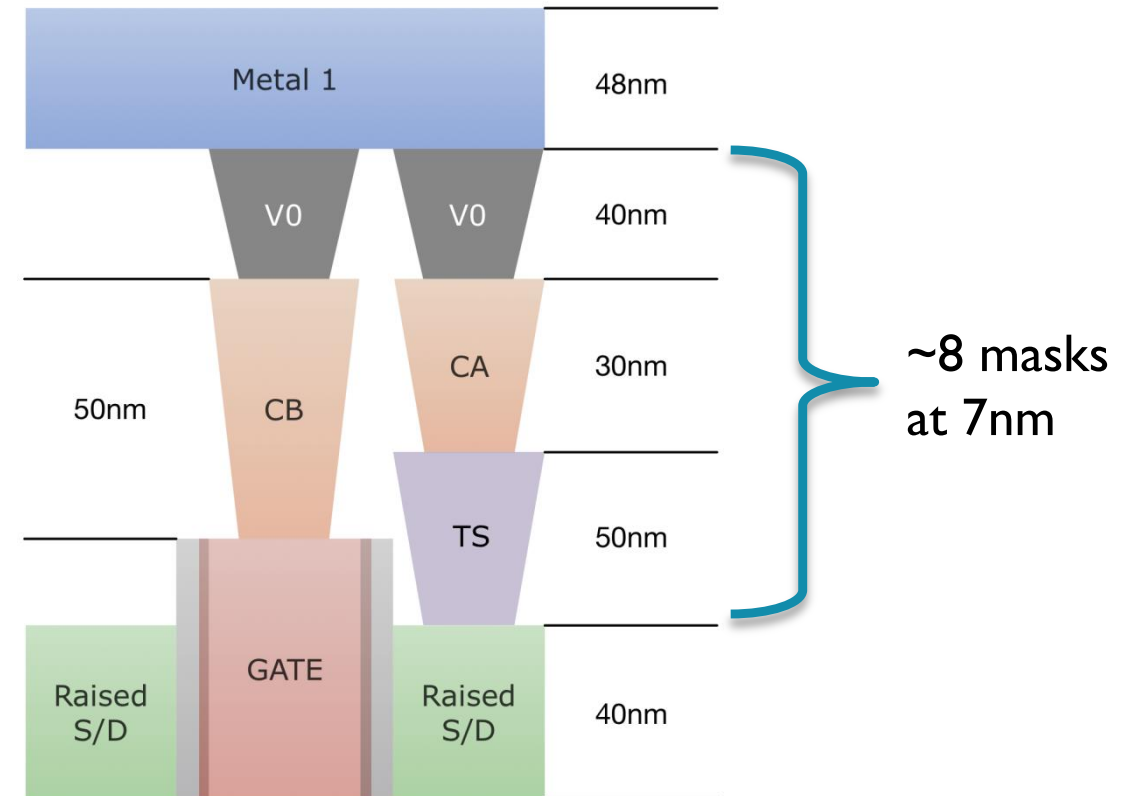
Expectations: V_T options

~ 2 Moore's Law Nodes

More V_T options = lower power products

Chip Power

Pleak (mW)

Fmax (Mhz)

Legend: RAM — Std Cell — Fmax

X-axis: baseline, Backfill 1, Backfill 2, Backfill 3
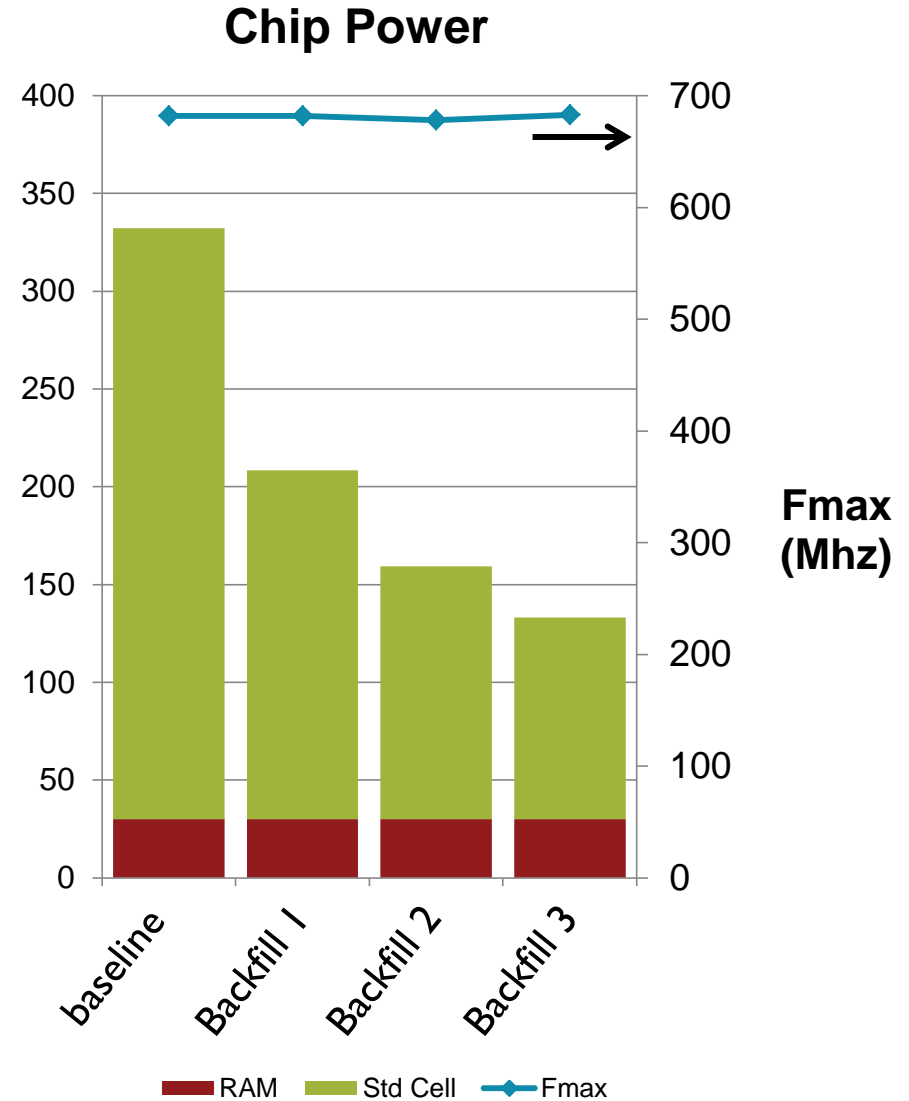
Device options = reduced design cost

Target Frequency vs Leakage Power

Area Efficiency vs Power Efficiency
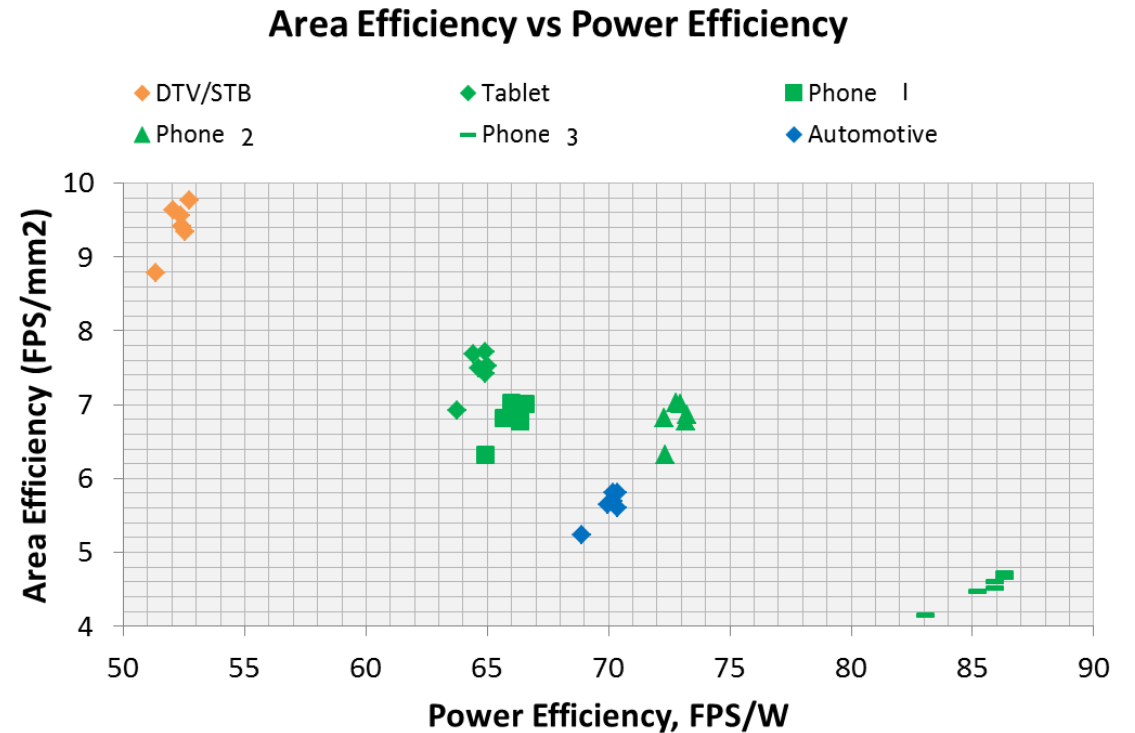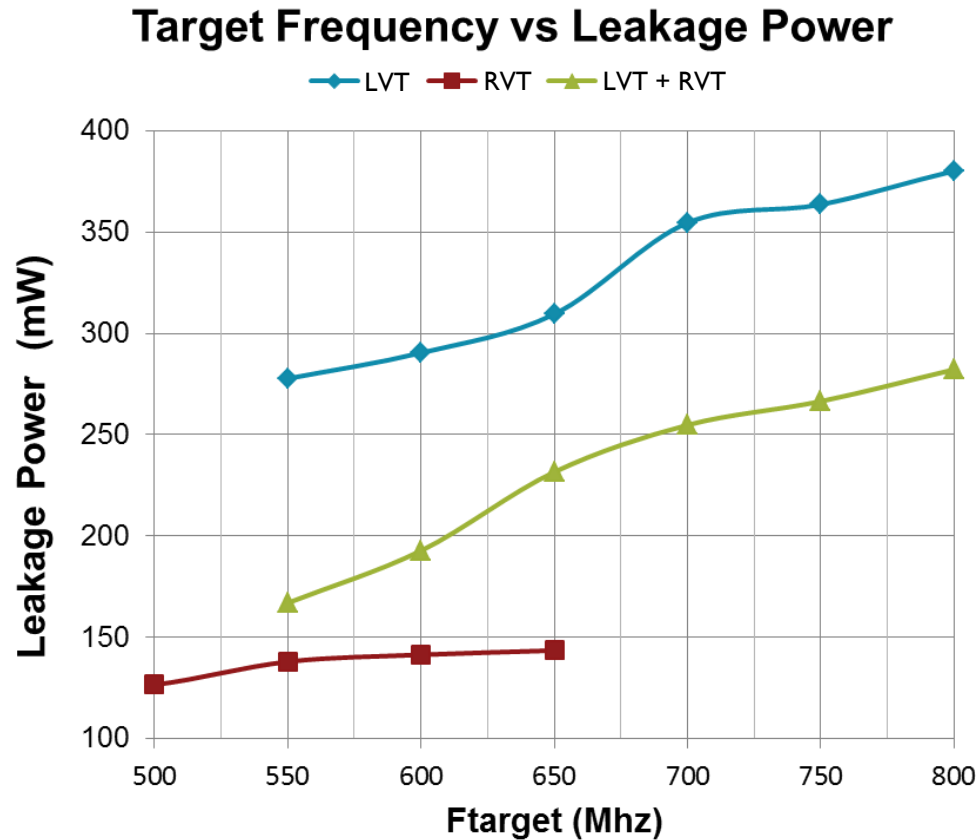
Performance vs. Leakage Power

The best products will come from
the fastest possible transistor
AND
the most energy efficient transistor

The range is important

16/14nm
20nm
10nm
7nm
5nm

Normalized Leakage Power

Normalized Performance
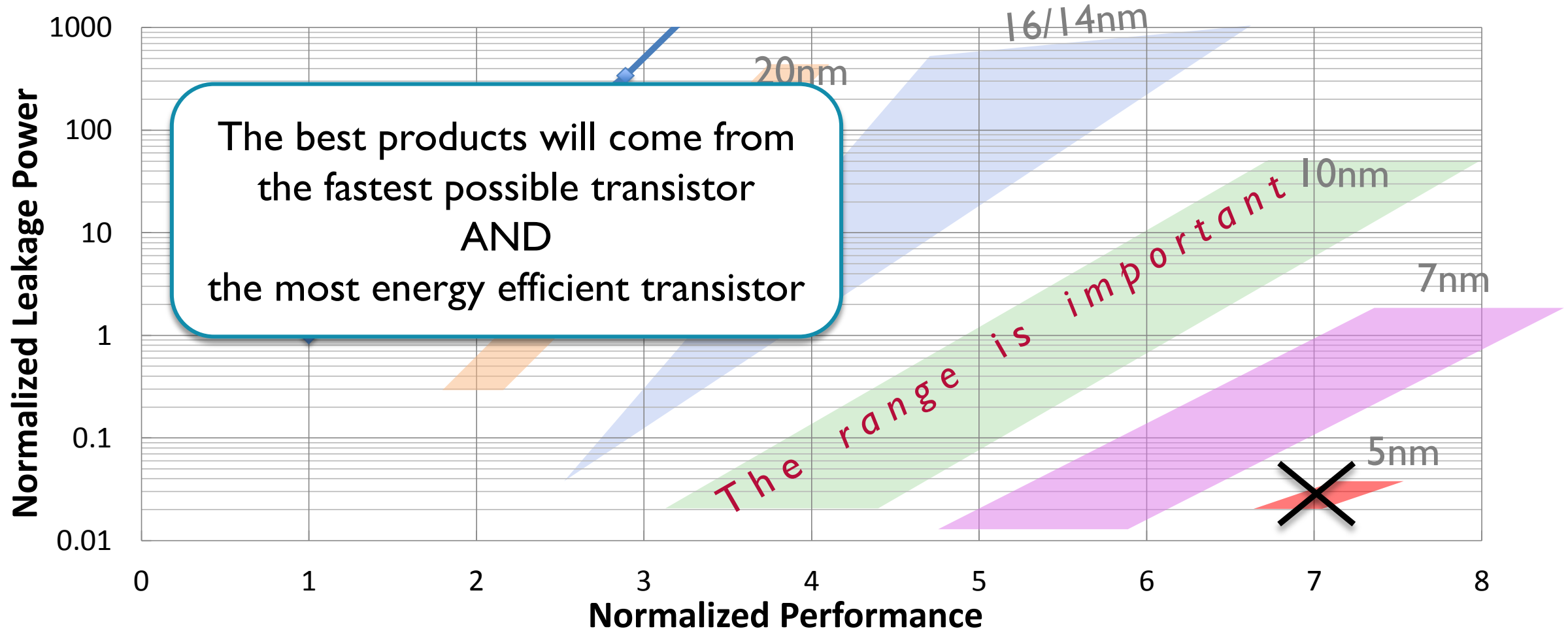
What FET heterogeneity can be cost effective?
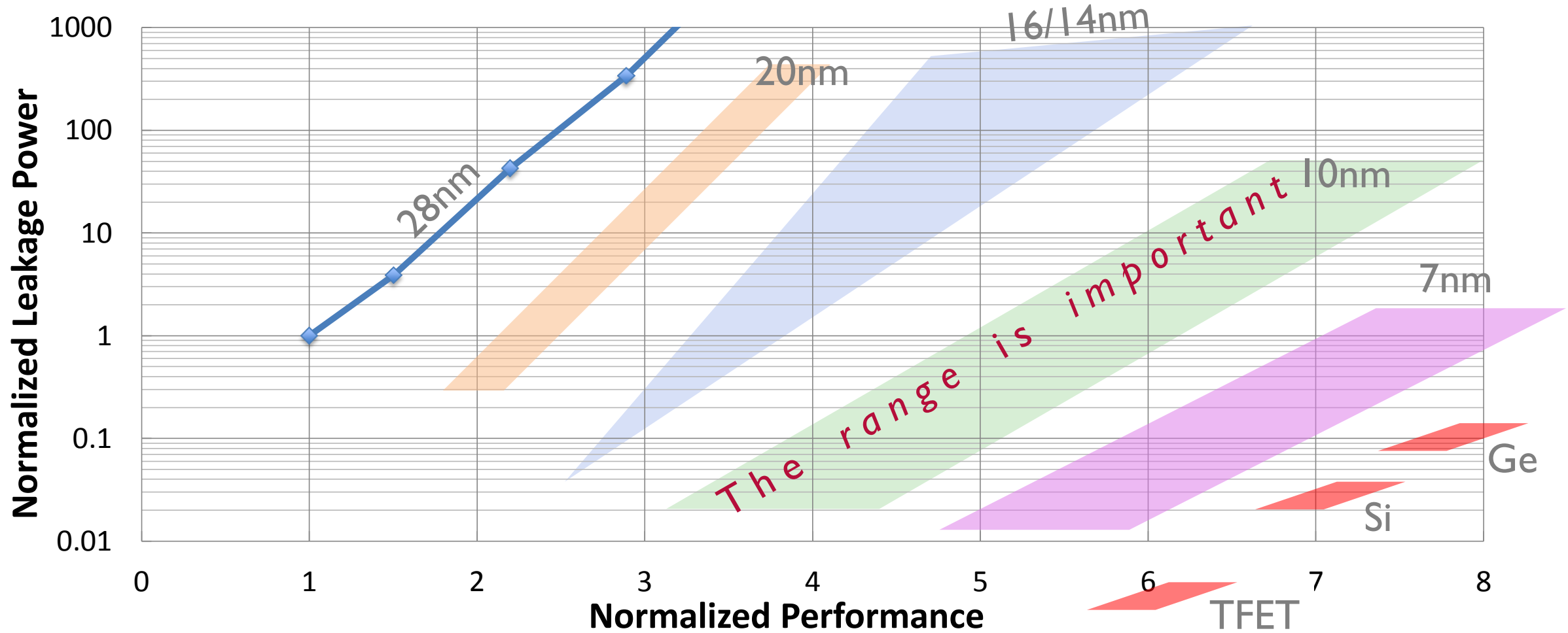
18.1   Energy-Delay Performance Optimization of NEM Logic Relay

An R_C equivalent to V_T?

IEDM 2015
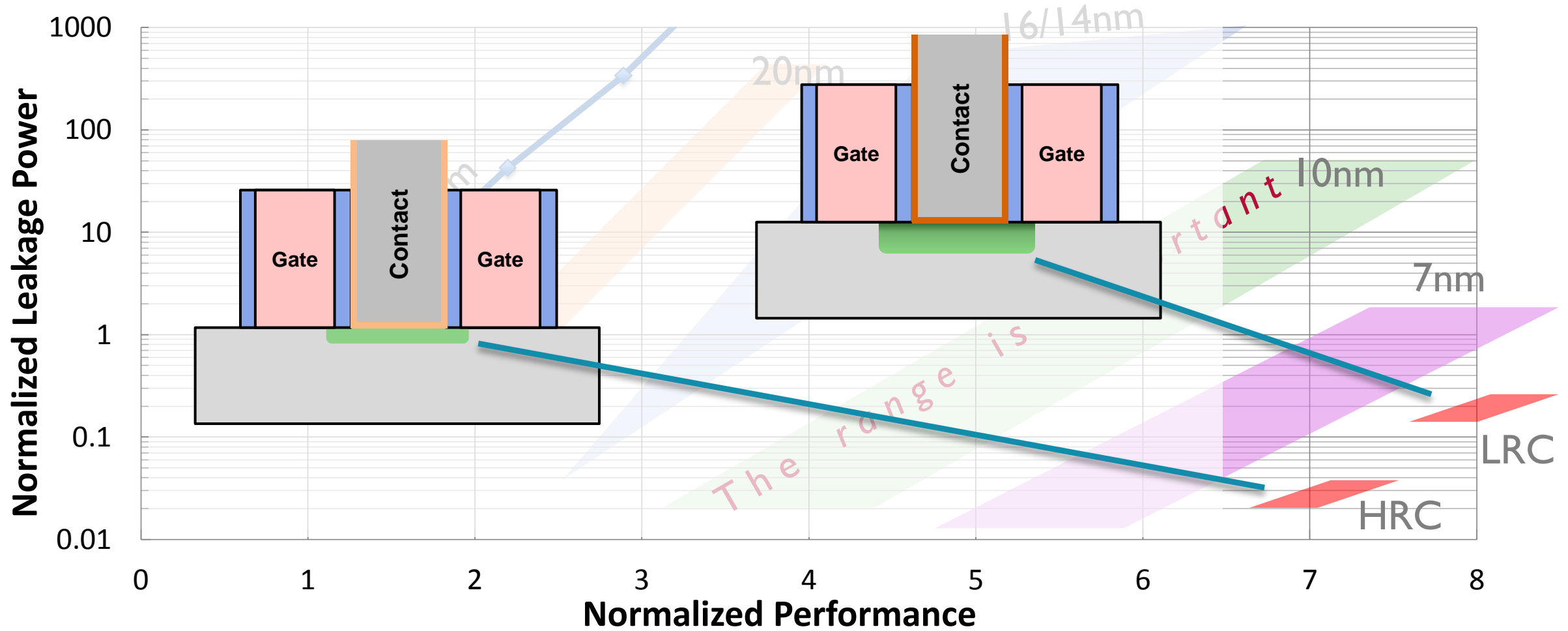
Lots of curvy M1 shapes: The bane of lithography.

http://www.chipworks.com/blog/technologyblog/2012/07/31/samsung-32-nm-technology-looking-at-the-layout/
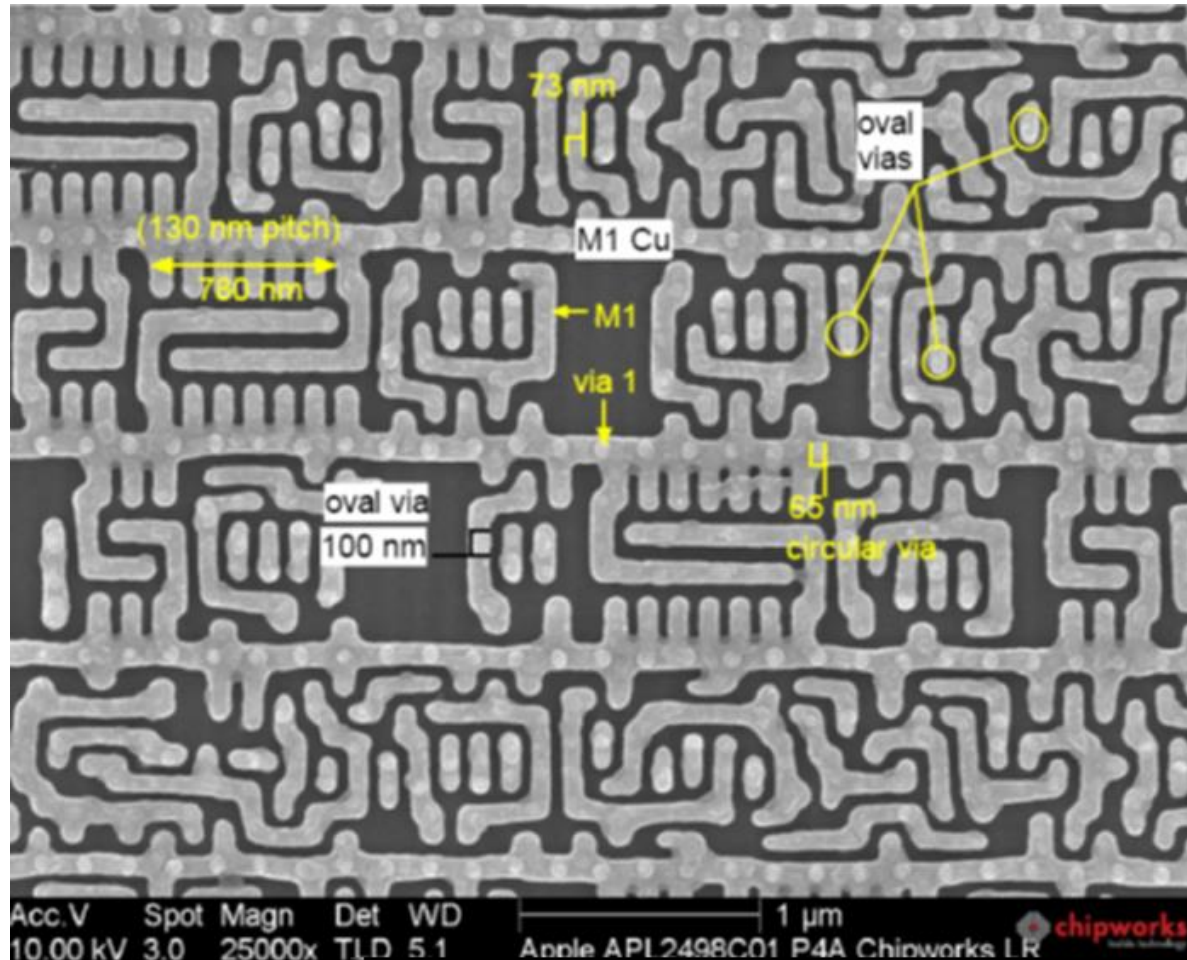
What does that have to do with transistors?

What does that have to do with transistors?

S/D plane

Gate plane

Gate

S/D

S/D plane
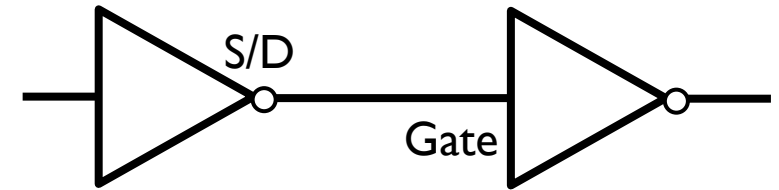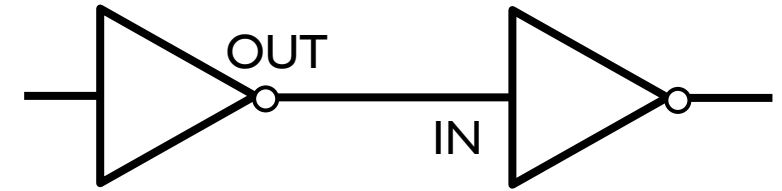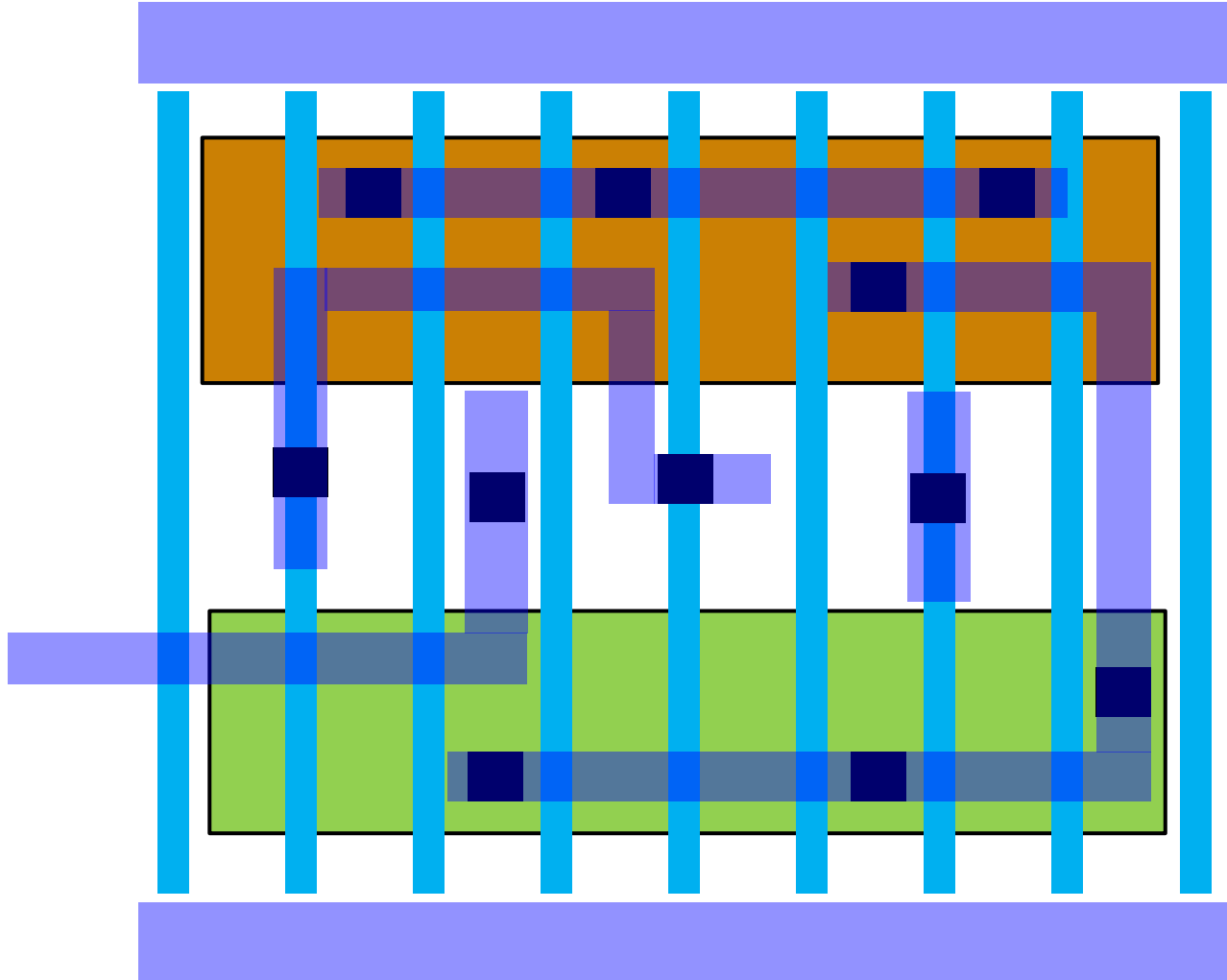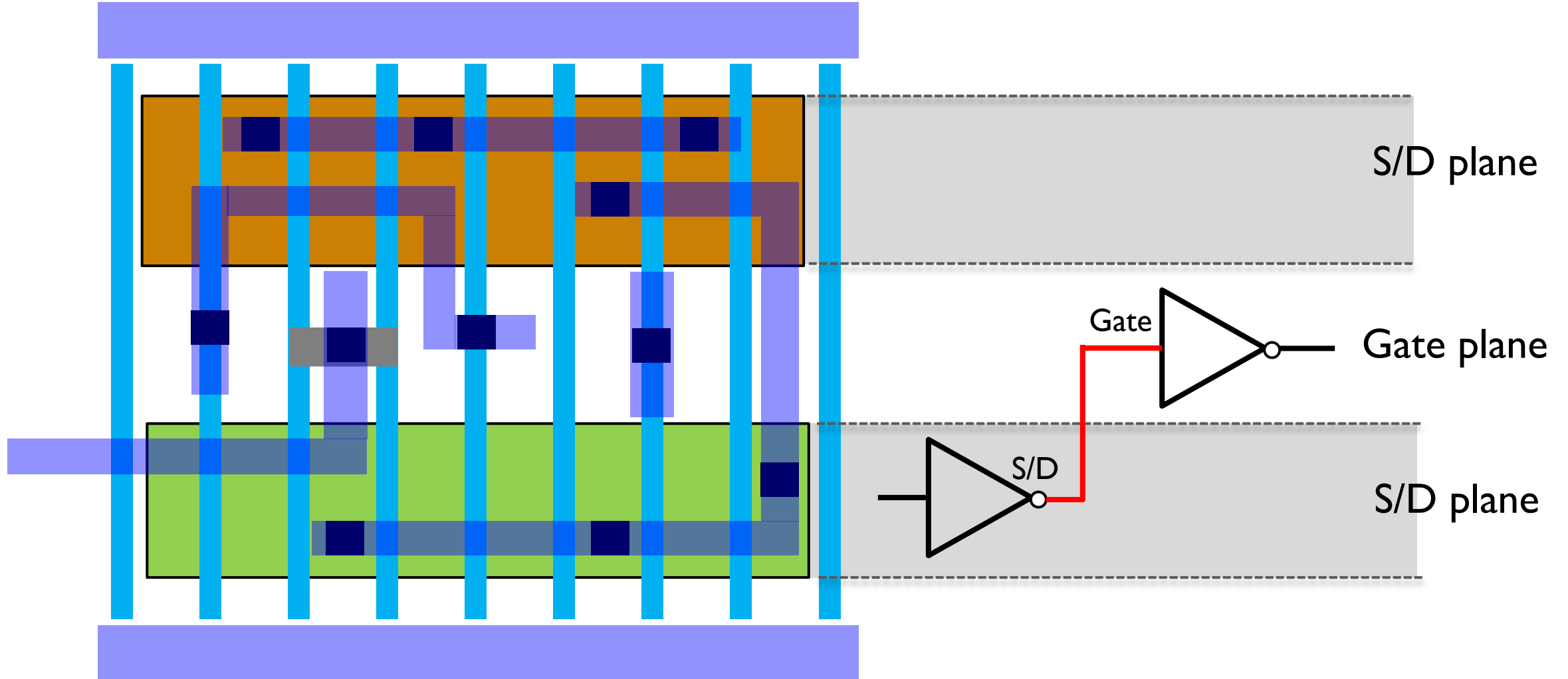
A potential improvement: gate contact over active

S/D

Gate
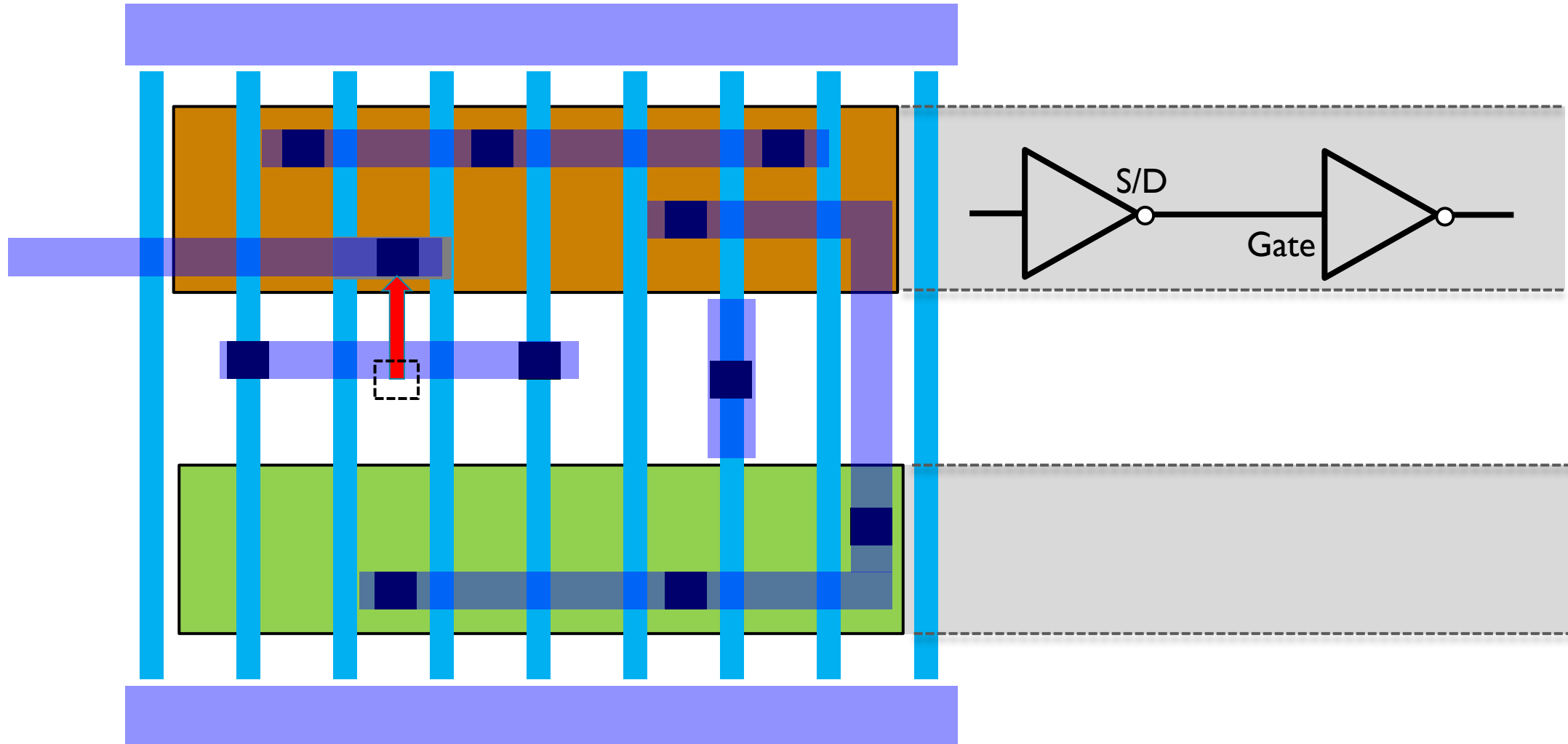
Plus performance improvement (or area shrink)

Wider devices = improved PPA
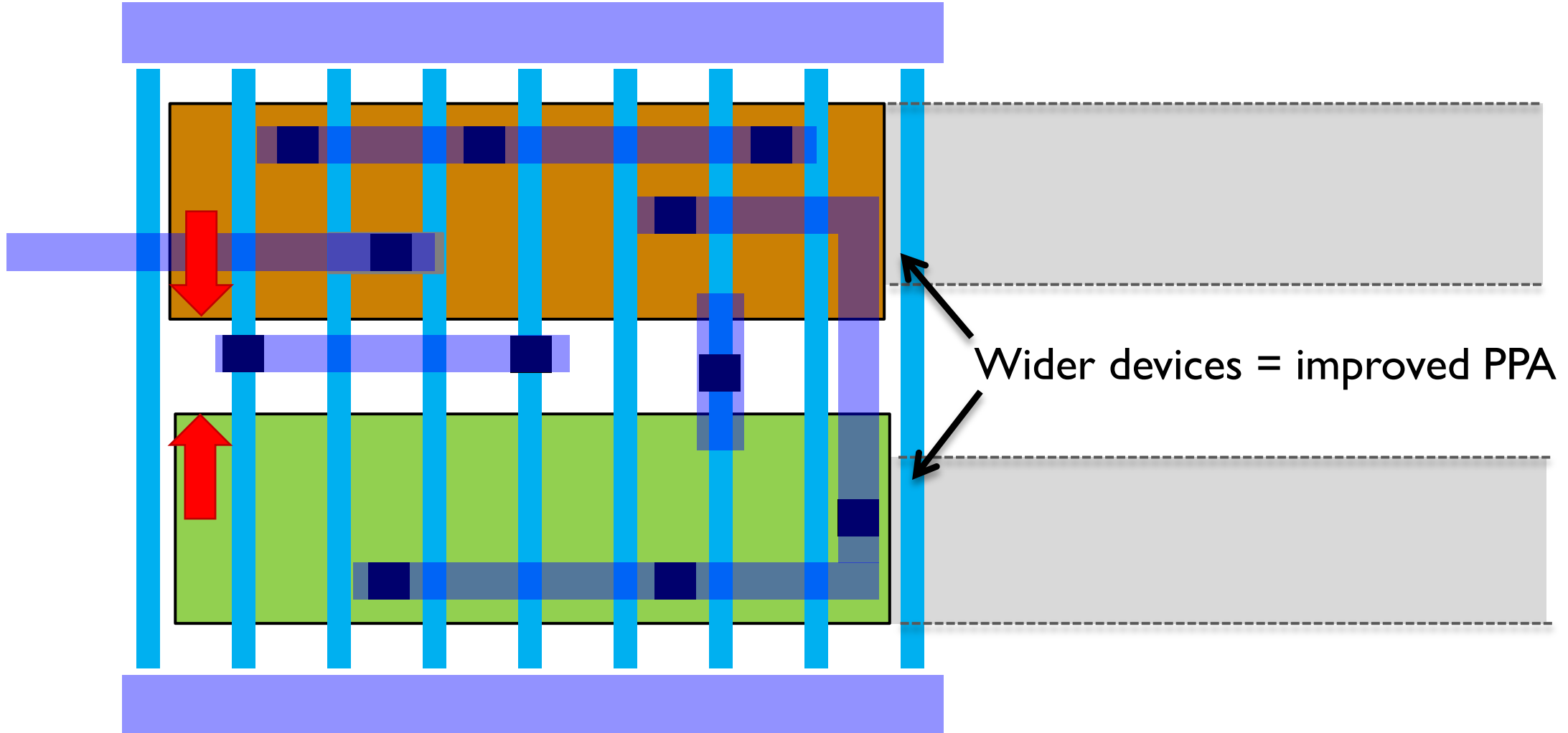
Another important scaling law:

$$\text{Variation} \propto \frac{A_{VT}}{\sqrt{W \cdot L}}$$

95% of device engineer's time

M. Pelgrom

Slowest	Typical	Fastest
(SS)	(TT)	(FF)
3σ		3σ

95% of design engineer's time
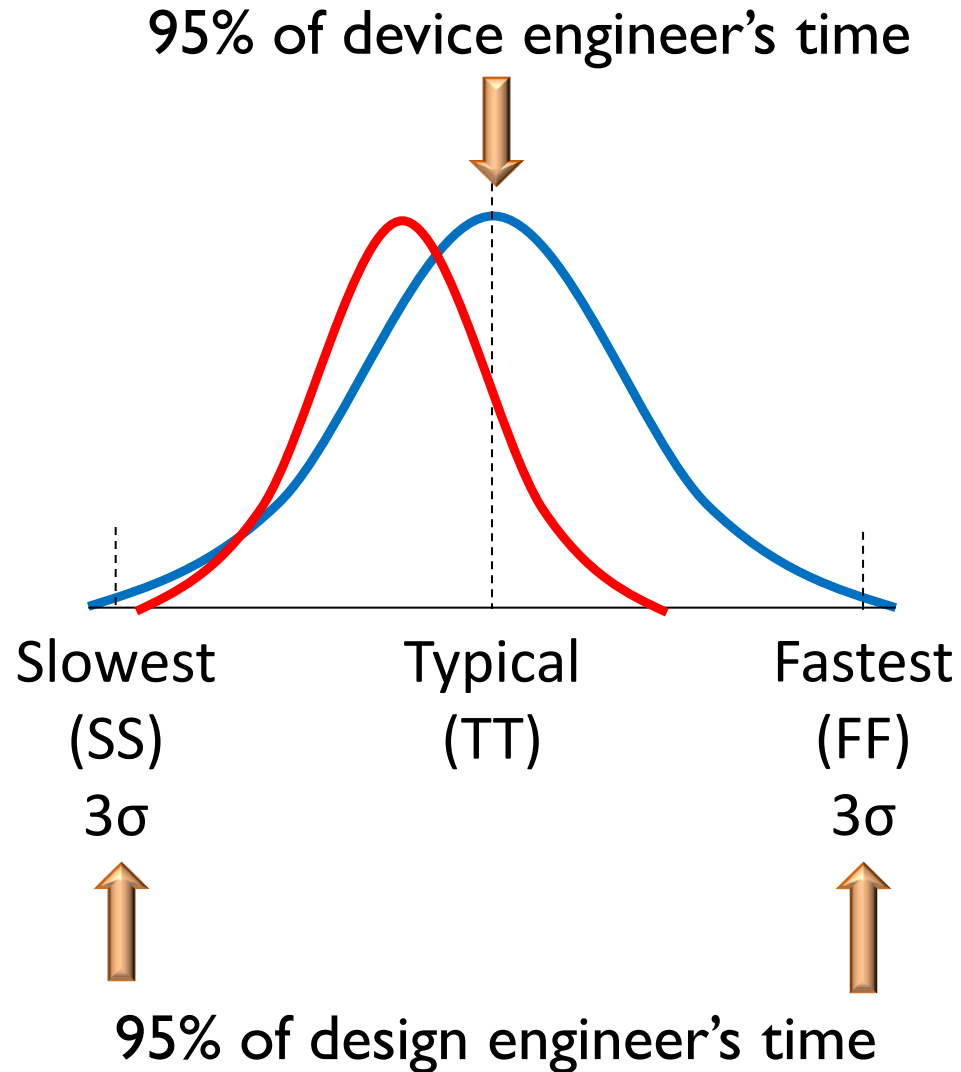
Red process faster than blue process!!

Investing in reducing variability may be more helpful than mobility

IEDM 2015

IEDM 2015 and variation

Line Edge Roughness.   Work Function variation.   Random Dopant Fluctuation

Session 11: Circuit Device Interaction – **CMOS Scaling and Circuit/Device Variability**

15.3   Novel Wafer-Scale Uniform Layer-by-Layer Etching Technology for Line-Edge-Roughness Reduction and Surface-Flattening of 3-D Ge Channels,

21.2   Variation Improvement for Manufacturable FINFET Technology

Session 20: Characterization, Reliability and Yield – **Transistors Ageing, Variability and the Impact on Circuit Design**

20.5   Technology Scaling and Reliability: Challenges and Opportunities

20.8   Implications of Variability on Resilient Design (Invited), R. Aitken, ARM

IEDM 2015

The irony of reliability limits to Moore's Law

IEDM 2014

Will Reliability Limit Moore's Law?

Anthony S. Oates

TSMC Ltd., 168 Park Ave. 2, Hsinchu Science Park, Hsinchu, Taiwan 30075; aoates@tsmc.com

Abstract– Up to the present time reliability has not limited the rapid evolution of Si process technologies. However, the near future will bring a continual stream of innovations in transistor architecture and gate dielectric and interconnect materials. Maintaining historical high levels of reliability in this environment will be challenging. In this paper we discuss the reliability issues that have the potential to limit the future pace of technology progress.

minimal. PBTI in NMOS has been added as a issue to be monitored [3], while other transis mechanisms have been readily optimized. Nov to the FinFET transistor architecture is unde indications pointing to successful implementat: quence of the introduction of FinFETs to reli pears to be relatively benign since the structure new mechanisms of degradation [4]. However, the confined fin geometry is exacerbated comp
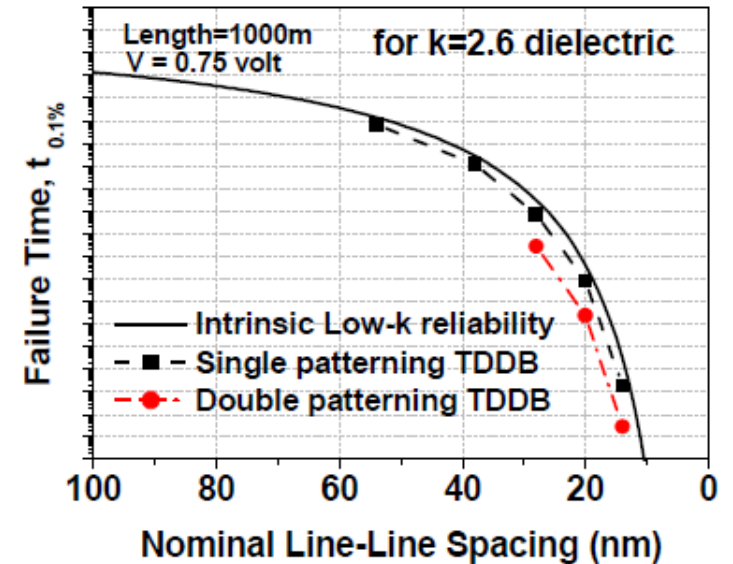
Fig. 4: Use condition low-k failure time prediction for L=1000m interconnect line length at V=0.75Volt.

The answer is yes:
- Electromigration (already)
- Soft Errors, not just in memory (interleaving, ECC), but in logic, especially HPC
- RTN, BTI, etc.:   the defects don't scale

2015

Dennard warned us about electromigration

Table 2
Scaling Results for Interconnection Lines

Parameter	Scaling Factor
Line resistance, $R_L = \rho L / Wt$	κ
Normalized voltage drop IR_L/V	κ
Line response time $R_L C$	1
Line current density I/A	κ

….and Moore's Law is paying the price

Electromigration today (16/14) materially affects Moore's Law scaling entitlement

Maximum Current Limits → fan out limits → more buffering

We now spend significant transistors (and power) on EM

2015

© 2015 ARM   Greg Yeric, ARM

Device to circuit scaling: Summary

- Moore's Law scaling is slowing down
    - BEOL lithography cost is growing, with no near-term reset (EUV, DSA, etc.)
    - Cost per chip pressure, may include pressure on transistors/chip (chip area)
- Dennard scaling pressure is as bad or worse, and is often interchangeable
    - Within existing device electrostatics, invest in variation reduction
    - Contacts becoming a key limiter, and are a key focus of next generation transistors
        (via and wire R also increasing, adding R to C in general trend)
    - New devices that can offer reduced V without sacrificing I are ultimately needed
- DTCO to pull in the S-curve
    - As costs increase, and/or node timing slows, more radical changes become viable
    - Added device flexibility: heterogeneity, gate patterning flexibility, voltage swing, etc.
        → A Moore's Law Node may gained or lost
- Reliability and Yield: How much longer can we stay the course?

Greg Yeric, ARM

2015: Age 50

<re/code>

NEWS ▼ | REVIEWS | VIDEO | PODCASTS | VOICES | WRITERS

Moore's Law Hits 50, but It May Not See 60

Dr. Gordon Moore

/ ENTERPRISE

Intel

By Arik Hesseldahl | @ahess247 | EMAIL | ETHICS

April 15, 2015, 3:11 PM PDT

Number of gates at constant cost

- 10M units
- 30M units
- 100M units
- 300M units
- +45% / node

90 65 40 28 16/14 10 7

Power Density

- PTM-MG
- ITRS System Driver LP SoC
- PIDS LSPT CV^2f

Normalized Power Density

2012 2014 2016 2018 2020

Year

I E D M 2015

Greg Yeric, ARM

Architecture-Technology Interactions

Applications www.vogella.com

Home, Contacts, Phone, Browser, ...

Application Framework

Managers for Activity, Window, Package, ...

Libraries	**Runtime**
SQLite, OpenGL, SSL, ...	Dalvik VM, Core libs

Linux Kernel

Display, camera, flash, wifi, audio, IPC (binder), ...

Logic cells

memory

IEDM 2015
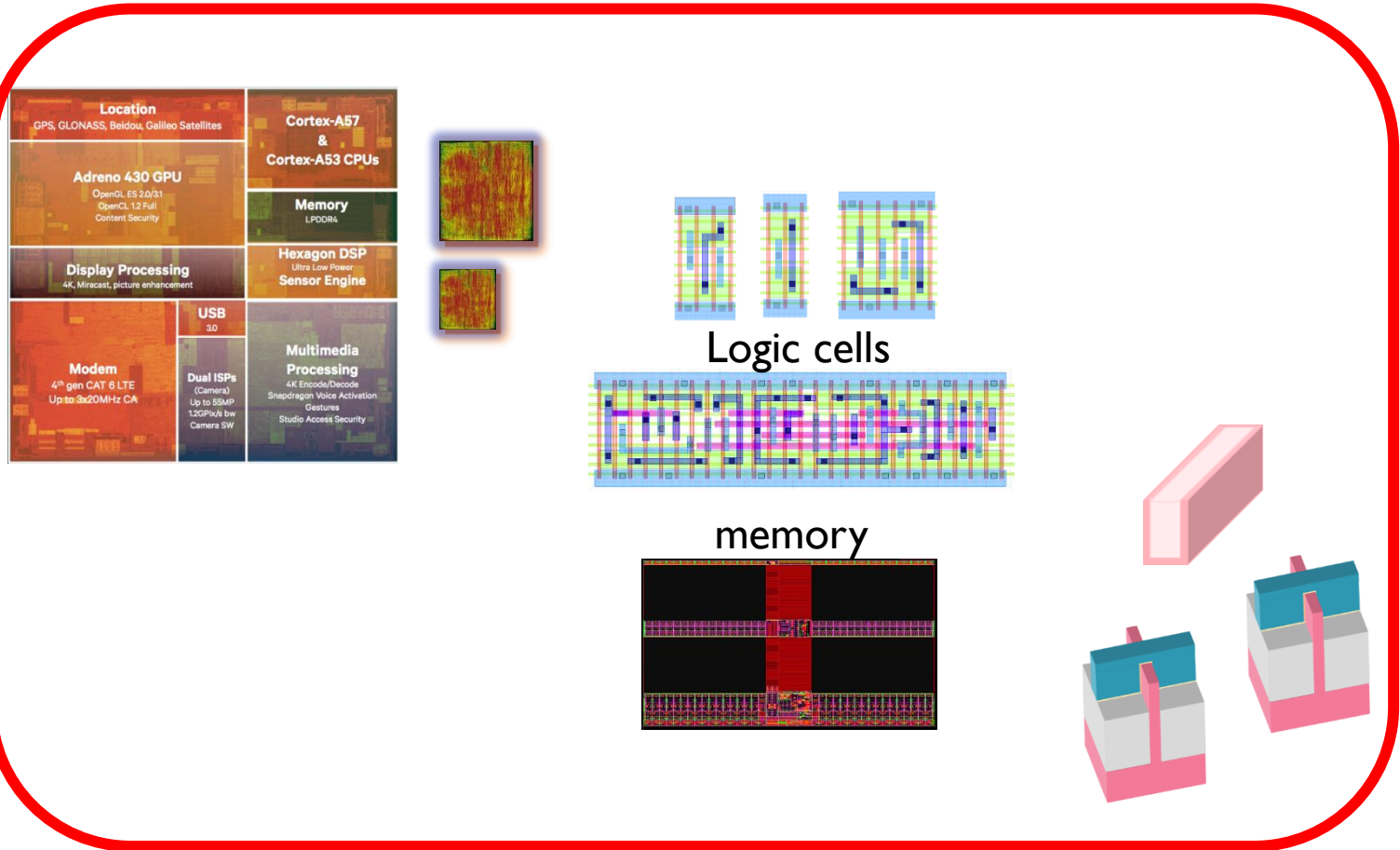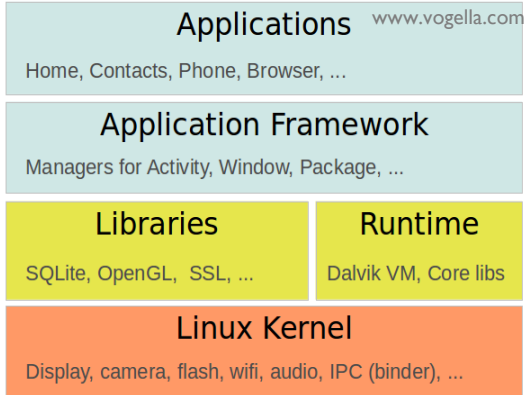
Heterogeneous Multi Core in the Dark Silicon era

Deca/10-Core CPU Architecture

2.5GHz	2.0GHz	1.4GHz
eXtreme Performance	Best Perf/Power Balance	Best Power Efficiency

A72	A72

A53	A53
A53	A53

A53	A53
A53	A53

L2 Cache | L2 Cache | L2 Cache

MediaTek Coherent System Interconnect (MCSI)

AXI Memory Bus
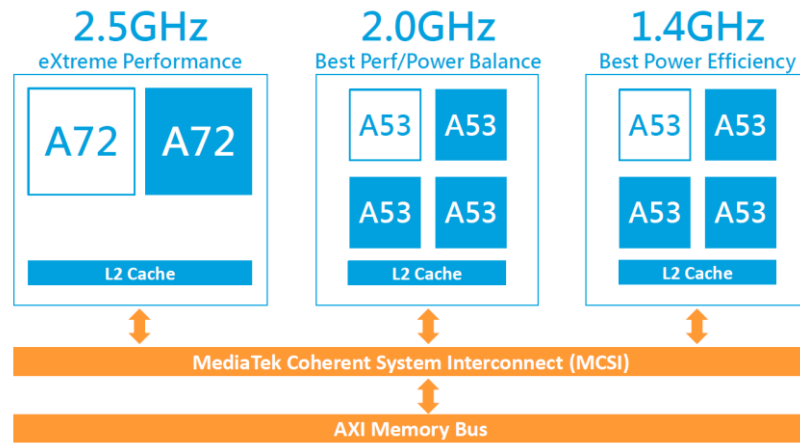
Source: anandtech

World's 1st Integrated Cortex-M4
- ❑ Clock speed up to 364MHz
- ❑ Dedicated SRAM size of 512KB
- ❑ Isolated low power domain
- ❑ Direct access to DRAM

Open Platform – More Differentiation
- ❑ ARM-based, friendly to developers
- ❑ Rich sample code and technical support
- ❑ Complete tool chain

Low-Power Sensor Hub

AP Processor | Modem
Audio/Sensor DSP
Audio HW & Peripherals

Low-Power MP3 Playback

Speech Enhancement

NB/WB/super WB
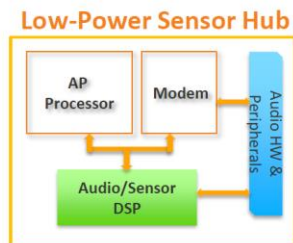
Power gain from Tri-cluster CPU architecture

	Dual-cluster power consumption	Tri-cluster power consumption	Improvement
FB launch	0.385W	0.318W	17%
FB Read	0.139W	0.084W	40%
FB Message	0.157W	0.101W	36%
FB scroll	0.217W	0.152W	30%
Beauty Plus	0.487W	0.378W	23%
Temple run launch	0.378W	0.316W	17%
Temple run play	0.303W	0.199W	34%
voice call	0.204W	0.121W	41%
Web Page loading	0.655W	0.627W	5%
Web Page Browsing	0.326W	0.273W	17%
Youtube HD	0.256W	0.156W	39%
Video Record	0.289W	0.197W	32%
Video Playback	0.113W	0.067W	41%
Homescreen idle	0.050W	0.026W	48%
Gmail	0.104W	0.061W	42%

x DoU Profile

30% OFF

MEDIATEK

13 | Confidential

Source: phonearena

Current trend:
Lower utilization for higher efficiency

Future trend:
More dedicated accelerators

Wire scaling modifies Moore's Law results

Legend:
- INV_X1
- INV_X2
- INV_X3
- INV_X4
- INV_X5
- INV_X8
- INV_X16

Y-axis: Wire length / Freq. (um/GHz)

X-axis: 28nm, 20nm, 16/14nm, 10nm, 7nm
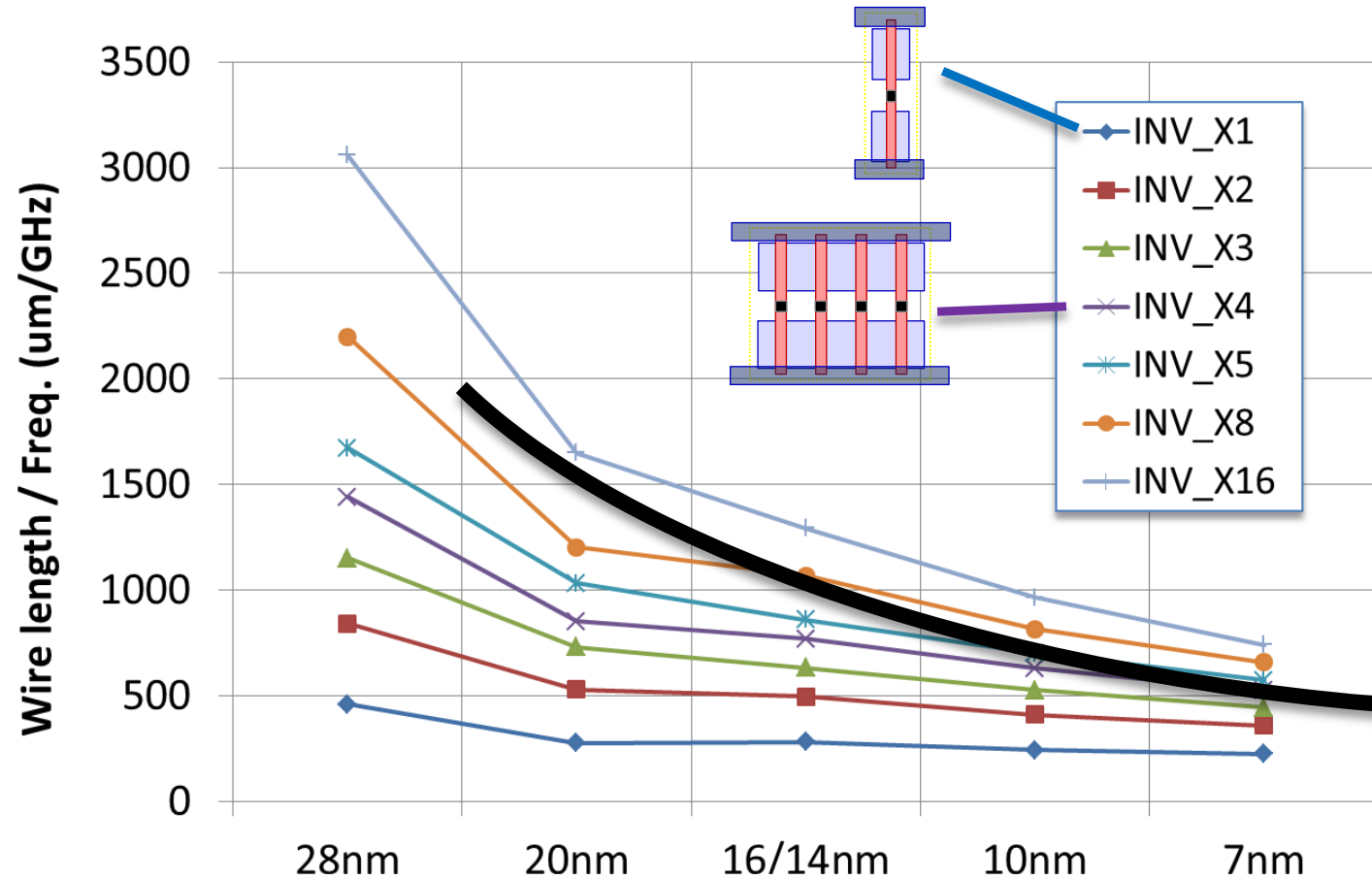
- Insert more transistors to maintain performance

- Add to power problems

- Problems are not just lateral, but more and more vertical

http://www.zyvexlabs.com/EIPBNuG/2005MicroGraph.html
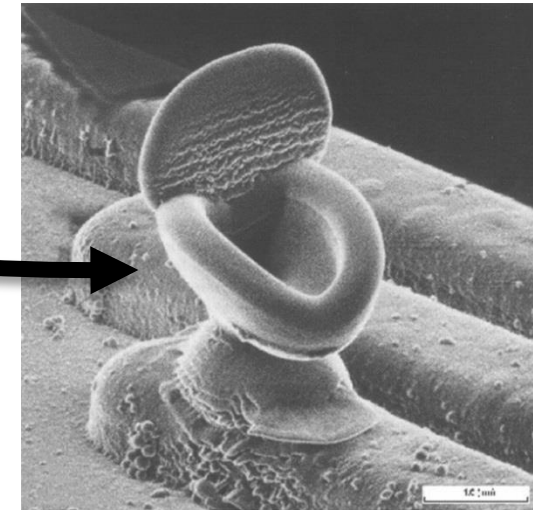
15.5   Present Status and Future Prospects of Nano-Carbon Interconnect Technologies for LSIs

Evening Panel: Is there a potential for a revolution in on-chip interconnect?

3DIC: More than Moore (?)

"It may prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected" [1]

3DIC and Moore's Law

10^5

10^4

1962

10^4

1965

10^3

1970

10^2

10

1

1   10 10^2 10^3 10^4 10^5

Relative Manufacturing Cost/Component

Number of Components Per Integrated Circuit

VIRTEX 3D UltraSCALE

48 16.3G GTs   2.6 Tbps of Bandwidth   1456 I/O

89 Mbits of RAM   4.4M Logic Cells

XCVU440
✓ Largest in the Industry by 4X
✓ A full generation ahead
✓ 50M equivalent ASIC gates

3x 100GE MACs

Xilinx.com
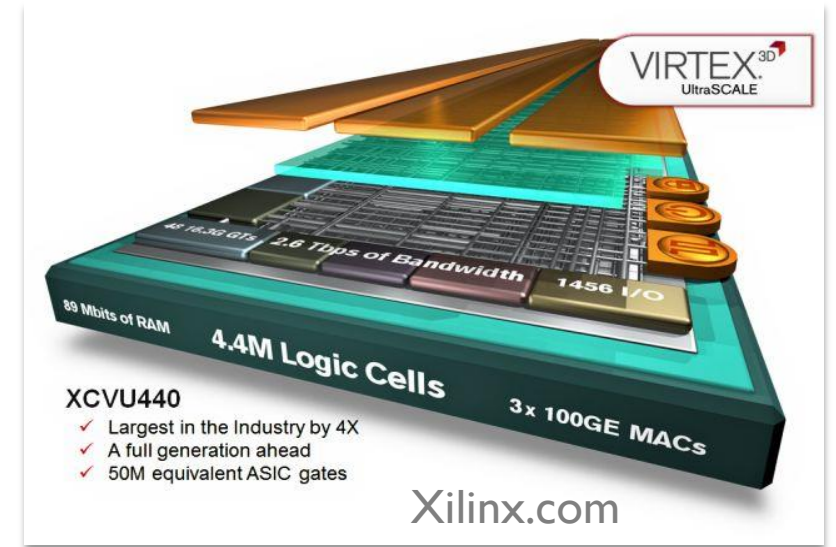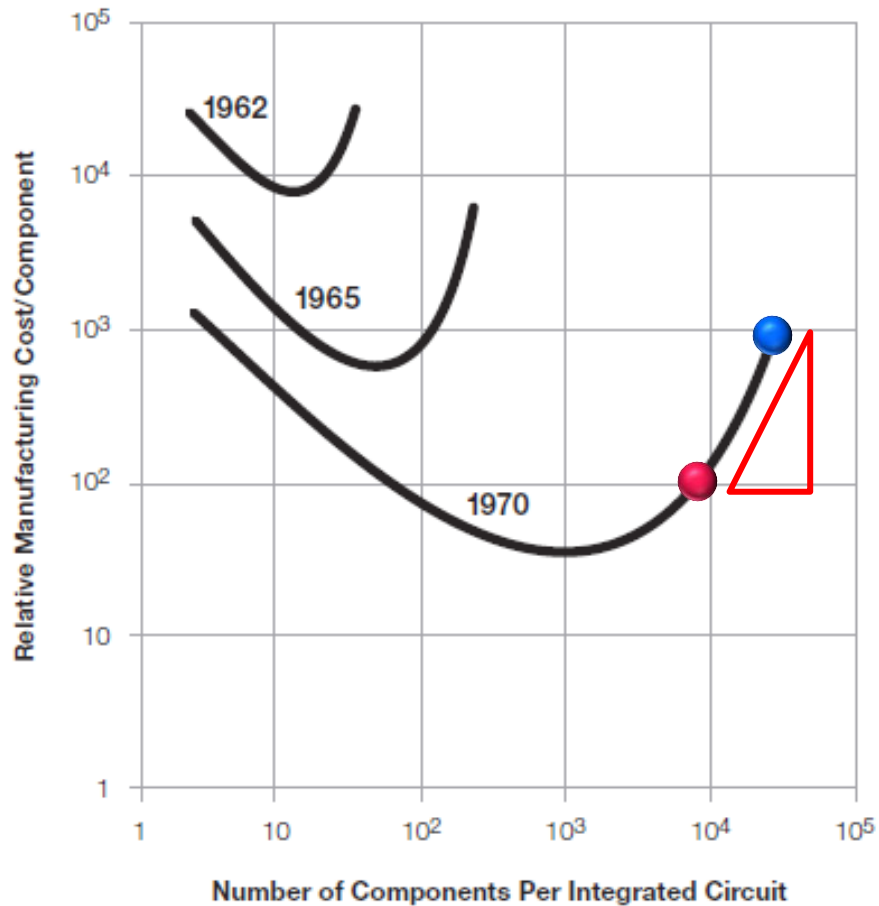
3DIC cost + 3 x Cost(🔴)

Cost(🔵)

(plus defect clustering advantage)

3D folding granularity

2D
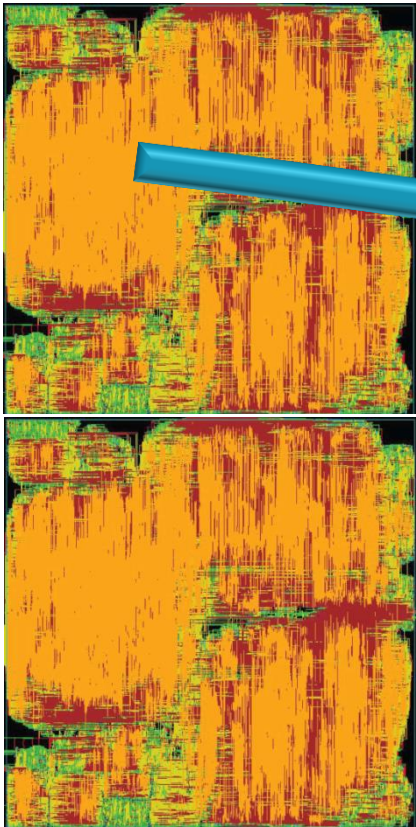
Core / block folding
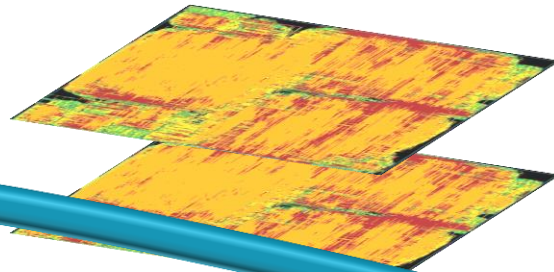
Gate-level folding

Monolithic 3D

Chang, Acharya, Sinha, Cline, Yeric, Lim; ISLPED 2015
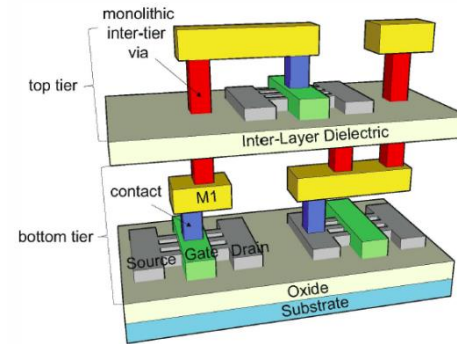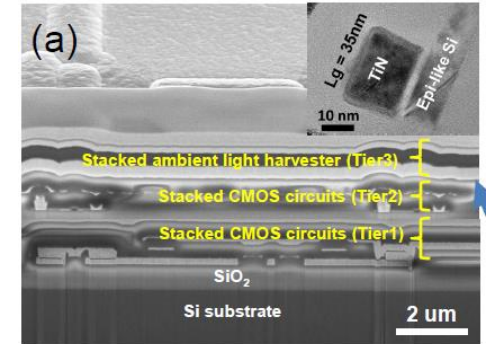
Power/Performance: ~One Moore's Law Node

3DIC early adopters

micron.com

VIRTEX 3D UltraSCALE

89 Mbits of RAM

46 16.3G GTs   2.6 Tbps of Bandwidth   1456 I/O

4.4M Logic Cells

3x 100GE MACs

XCVU440
- ✓ Largest in the Industry by 4X
- ✓ A full generation ahead
- ✓ 50M equivalent ASIC gates

Xilinx.com

Conventional **Back-illuminated CMOS image sensor**

Newly-developed **Stacked CMOS image sensor**
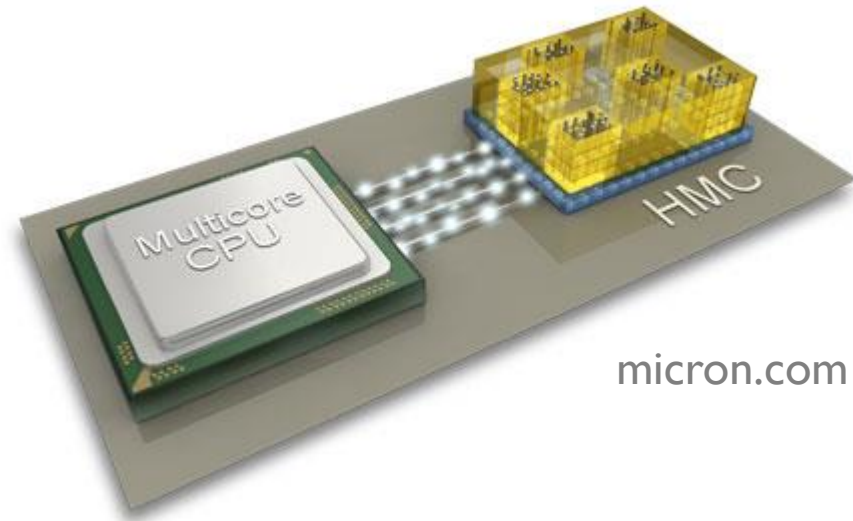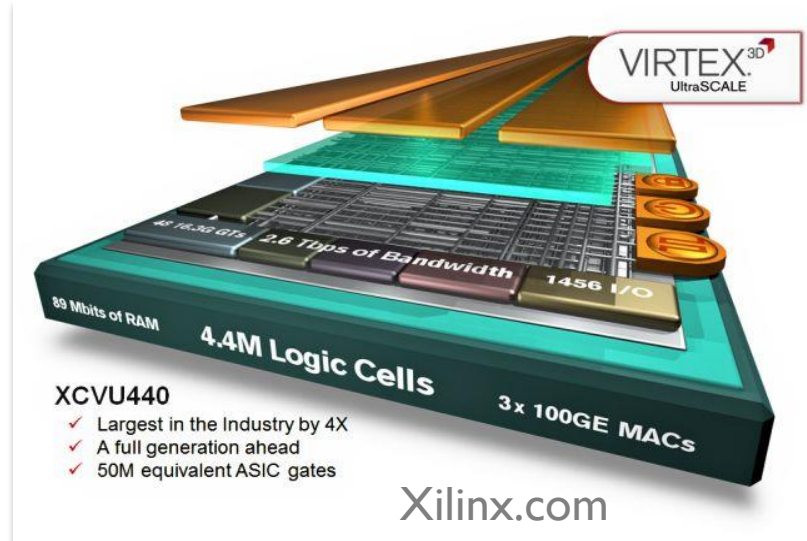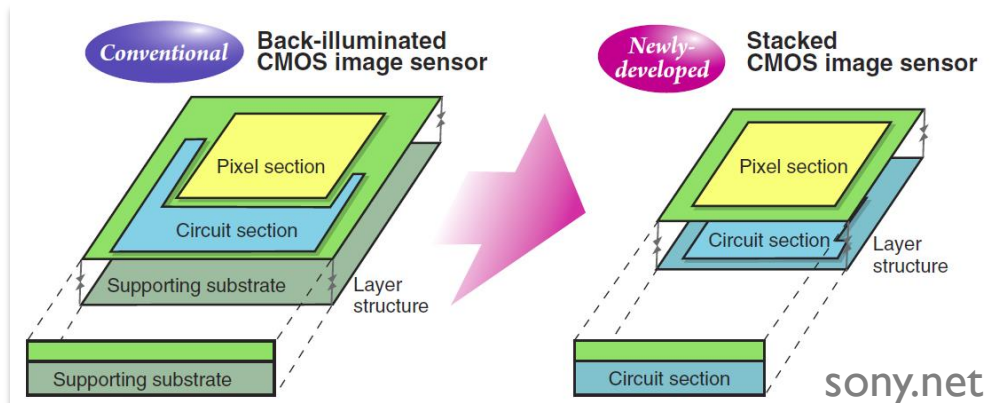
sony.net

CMOS Image Sensor and 3DIC: Virtuous circles

Smart Phones

3DIC
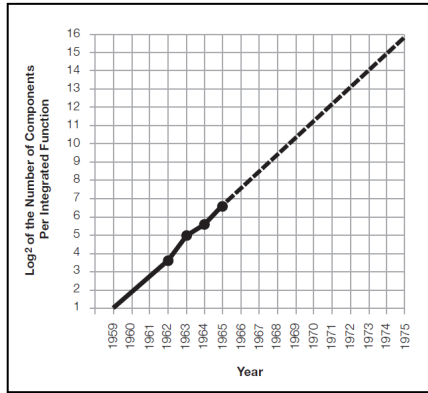
CIS
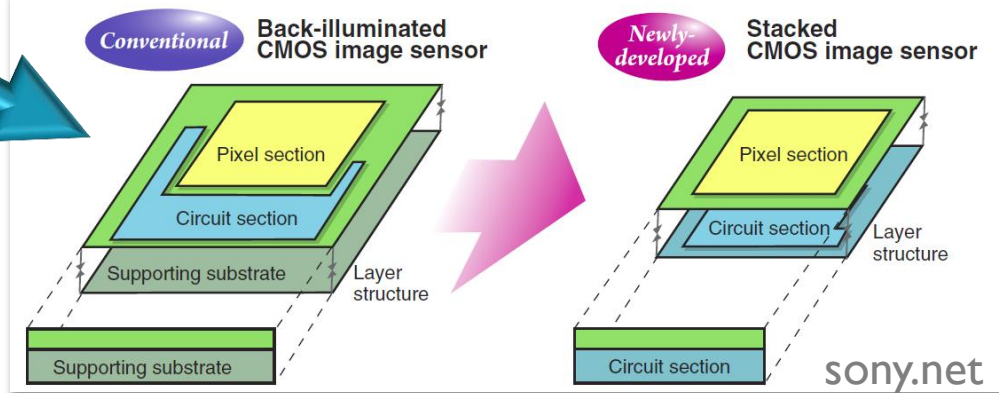
Demand

Heterogeneous SoC

Parallelizable functions:
Favors efficiency over raw performance

High performance CMOS

Analog devices not optimized in CMOS, RF may prefer FDSOI

Leading edge feature density wasted?

GPU

Big core

Big core

Display

Little core

Little core

GPS Radios I/O sensor
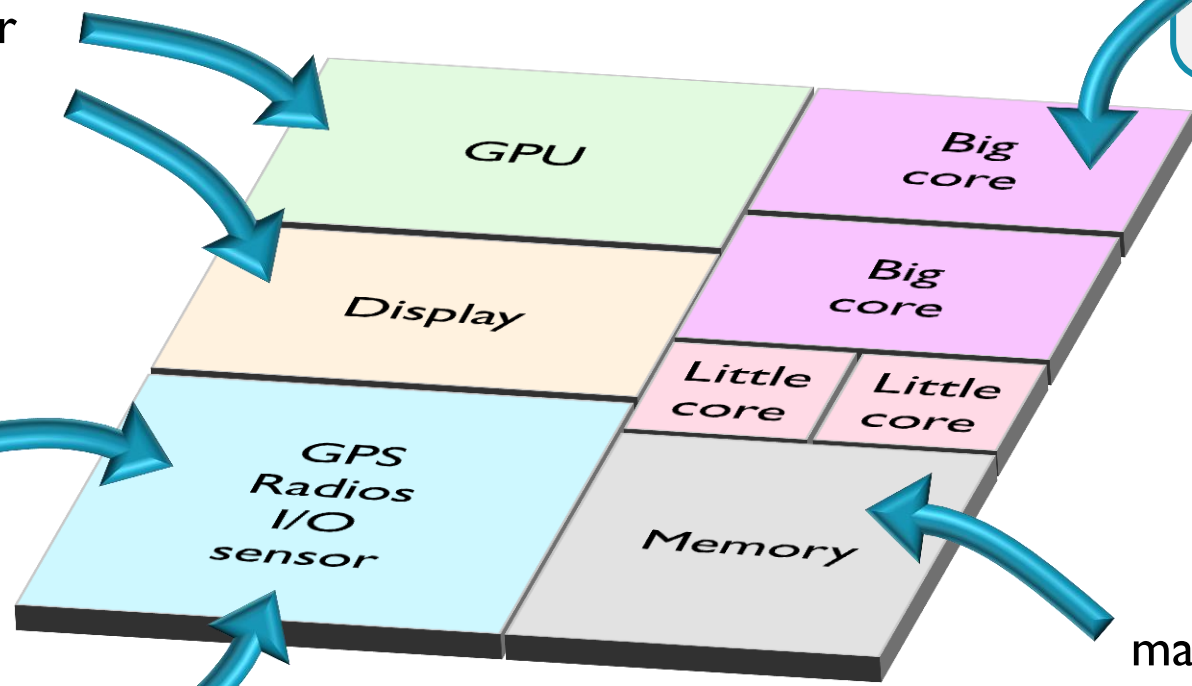
Memory

matching, leakage

Always on sensor
Leakage / performance

Now…. think along product portfolios, and not just one chip

IEDM 2015

Modern von Neumann computers

INPUT

Central Processing Unit

Arithmetic Logic Unit

Central Control Unit

OUTPUT

MEMORY

PROGRAM

"Stored Program Computer"
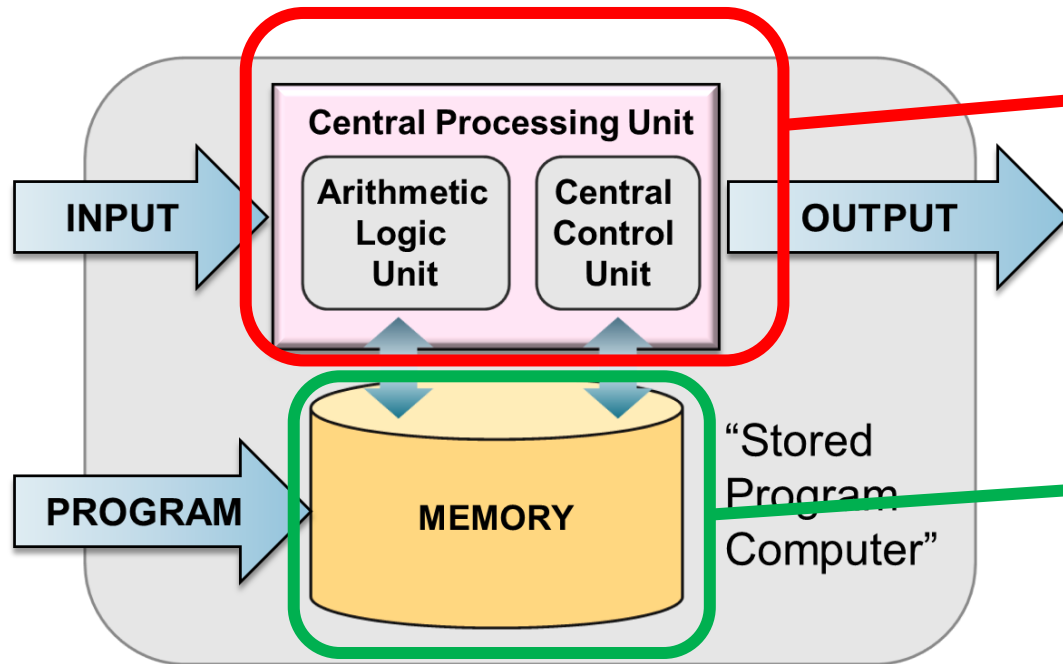
What software thinks memory looks like

Processor

CPU

SUPER FAST
SUPER EXPENSIVE
TINY CAPACITY

PROCESSOR REGISTER

CPU CACHE
LEVEL 1 (L1) CACHE
LEVEL 2 (L2) CACHE
LEVEL 3 (L3) CACHE

FASTER
EXPENSIVE
SMALL CAPACITY

EDO, SD-RAM, DDR-SDRAM, RD-RAM and More...

PHYSICAL MEMORY
RAMDOM ACCESS MEMORY (RAM)

FAST
PRICED REASONABLY
AVERAGE CAPACITY

SSD, Flash Drive

SOLID STATE MEMORY
NON-VOLATILE FLASH-BASED MEMORY

AVERAGE SPEED
PRICED REASONABLY
AVERAGE CAPACITY

Mechanical Hard Drives

VIRTUAL MEMORY
FILE-BASED MEMORY

SLOW
CHEAP
LARGE CAPACITY

▲ Simplified Computer Memory Hierarchy
Illustration: Ryan J. Leng

http://www.bit-tech.net

2015

Greg Yeric, ARM

Modern von Neumann computers

All semiconductor memories have all three problems

Application trends demand exponential commodity memory.

Processor

CPU	SUPER FAST SUPER EXPENSIVE TINY CAPACITY
PROCESSOR REGISTER	
CPU CACHE LEVEL 1 (L1) CACHE LEVEL 2 (L2) CACHE LEVEL 3 (L3) CACHE	FASTER EXPENSIVE SMALL CAPACITY

EDO, SD-RAM, DDR-SDRAM, RD-RAM and More...

PHYSICAL MEMORY — RAMDOM ACCESS MEMORY (RAM) — FAST PRICED REASONABLY AVERAGE CAPACITY

SSD, Flash Drive

SOLID STATE MEMORY — NON-VOLATILE FLASH-BASED MEMORY — AVERAGE SPEED PRICED REASONABLY AVERAGE CAPACITY

Mechanical Hard Drives

VIRTUAL MEMORY — FILE-BASED MEMORY — SLOW CHEAP LARGE CAPACITY

▲ Simplified Computer Memory Hierarchy
Illustration: Ryan J. Leng

http://www.bit-tech.net

IEDM 2015

© 2015 ARM   Greg Yeric, ARM

DRAM scaling at the S curve

DRAM up to half of system power

    Still consumes power when not doing anything

2-3 "known" nodes left

DRAM density chart (Density(Mb) vs Year), showing 4X/3yr, 2X/3yr, and 2X/4yr scaling trends with nodes from 1Mb (1985) to 16Gb (beyond 2020).

Processor	CPU — PROCESSOR REGISTER	SUPER FAST SUPER EXPENSIVE TINY CAPACITY
	CPU CACHE — LEVEL 1 (L1) CACHE, LEVEL 2 (L2) CACHE, LEVEL 3 (L3) CACHE	FASTER EXPENSIVE SMALL CAPACITY
EDO, SD-RAM, DDR-SDRAM, RD-RAM and More...	PHYSICAL MEMORY — RAMDOM ACCESS MEMORY (RAM)	FAST PRICED REASONABLY AVERAGE CAPACITY
SSD, Flash Drive	SOLID STATE MEMORY — NON-VOLATILE FLASH-BASED MEMORY	AVERAGE SPEED PRICED REASONABLY AVERAGE CAPACITY
Mechanical Hard Drives	VIRTUAL MEMORY — FILE-BASED MEMORY	SLOW CHEAP LARGE CAPACITY

▲ Simplified Computer Memory Hierarchy
Illustration: Ryan J. Leng

http://www.bit-tech.net

IEDM 2015

NAND: MLC and Vertical

Park et al.
JSSC v50 n1 2015

...but for how much longer?

Jung H. Yoon, flashmemorysummit.com

▲ Simplified Computer Memory Hierarchy
Illustration: Ryan J. Leng
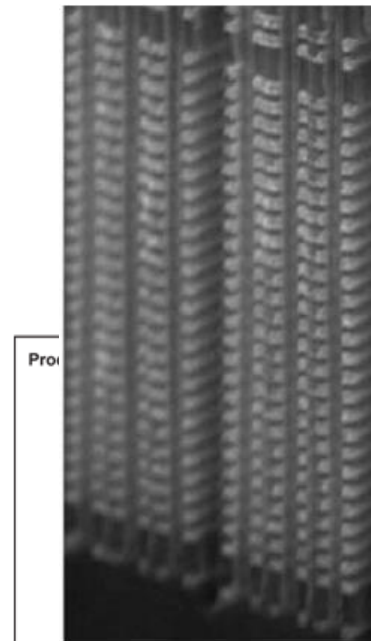
NAND Flash is a horrible NVM technology: Cost, Power, Speed, Endurance
Until you consider the alternatives

SRAM cache scaling

SRAM DENSITY - 16nm vs 28nm

Simplified Computer Memory Hierarchy
Illustration: Ryan J. Leng

http://www.bit-tech.net
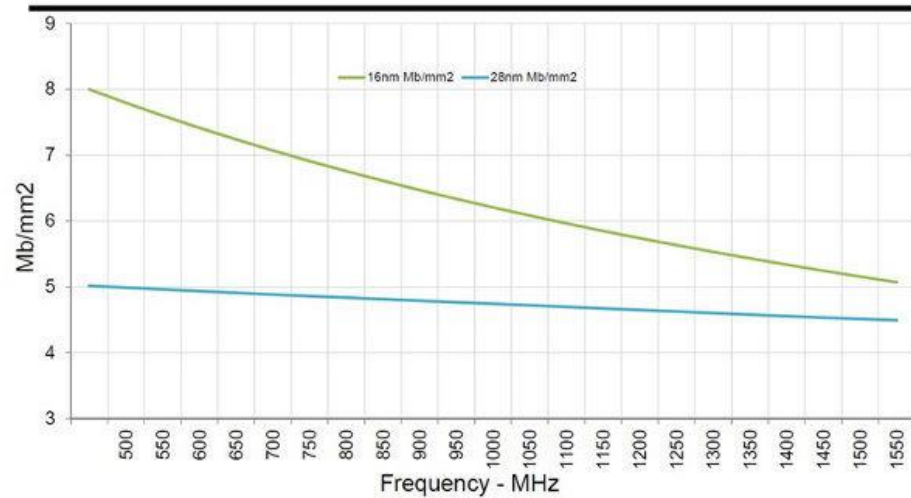
- The trend is for smaller bank sizes and more overhead (<50% efficiency)
- Bigger transistors, more transistors 6T → 8T
- Often the voltage scaling bottleneck

Wanted: Super Memory

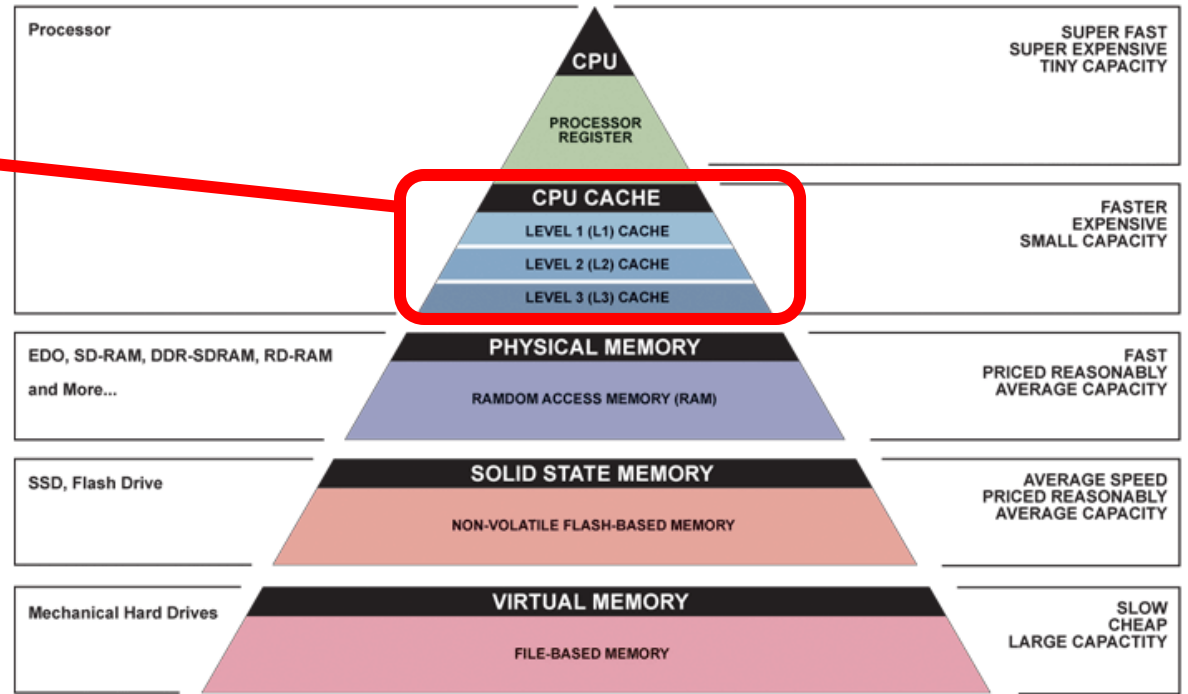|	SRAM	DRAM	NAND	SuperMem
Area (F^2)	~120	4-6	<4 (eff)	<=6
Write Speed	<300ps	10ns	50ns+	<1ns
Read Speed	<300ps	10ns	10ns	<1ns
Leakage	High	Low	0	0
Active Power	"Low"	Low	High (write)	Low (1pJ/bit)
Nonvolatile	No	No	Yes	Yes
High Voltage	No	No	Yes	No
Logic process	Yes	No	No	Yes
Endurance	Infinite	Infinite	10^5	>10^15

- Super memory needs to be able to scale!
- Bringing denser, faster, and/or lower power memory to compute can greatly offset scaling problems
- A fast, high endurance NVM would enable new paradigms in persistent compute

Memory technology at IEDM 2015

3.3   A Floating Gate Based 3D NAND Technology with CMOS Under Array

3.6   Crystalline-as-Deposited ALD Phase Change Material Confined PCM Cell for High Density Storage Class Memory,

7.7   Distribution Projecting the Reliability for 40 nm ReRAM and beyond based on Stochastic Differential Equation,

10.6   Programming-Conditions Solutions Towards Suppression of Retention Tails of Scaled Oxide-Based RRAM,

10.1   Non Volatile Memory Evolution and Revolution

26.1   Fully Functional Perpendicular STT-MRAM Macro Embedded in 40 nm Logic for Energy-efficient IOT Applications
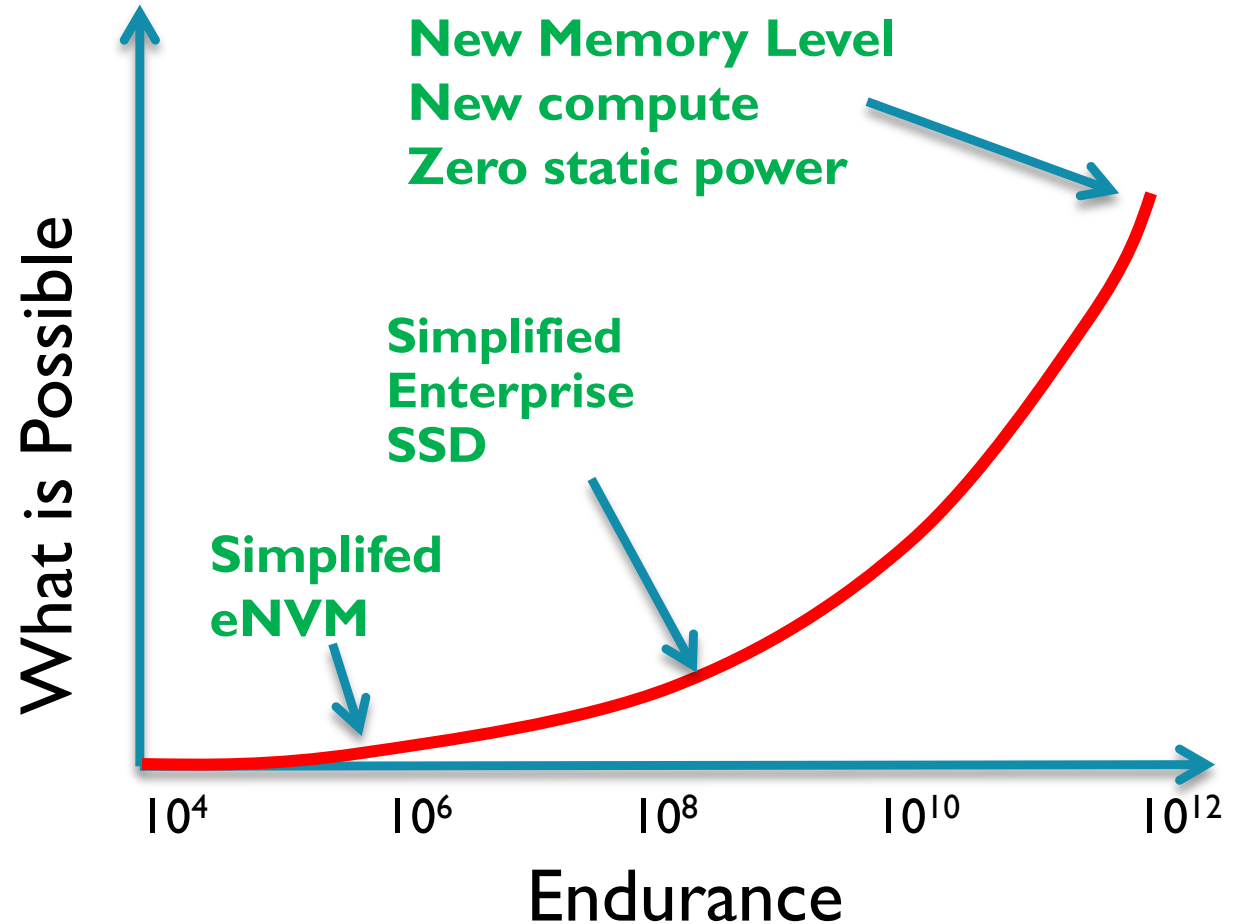
26.2   Systematic Optimization of 1 Gbit Perpendicular Magnetic Tunnel Junction Arrays for 28 nm Embedded STT- MRAM and Beyond,

26.4   Solving the Paradox of the Inconsistent Size Dependence of Thermal Stability at Device and Chip-level in Perpendicular STT-MRAM

26.7   A Novel Bi-stable 1-Transistor SRAM for High Density Embedded Applications,

2015

The contenders

- Filamentary RRAM:
  - Stochastic variability
  - Endurance
  - Scalability (filaments don't scale)
- PCM:
  - Variability
  - Endurance
- MRAM:
  - Power/speed tradeoff (including read margin)
  - Disturb
  - Cost

**New Memory Level
New compute
Zero static power**

**Simplified
Enterprise
SSD**

**Simplifed
eNVM**

What is Possible

10^4 10^6 10^8 10^{10} 10^{12}

Endurance

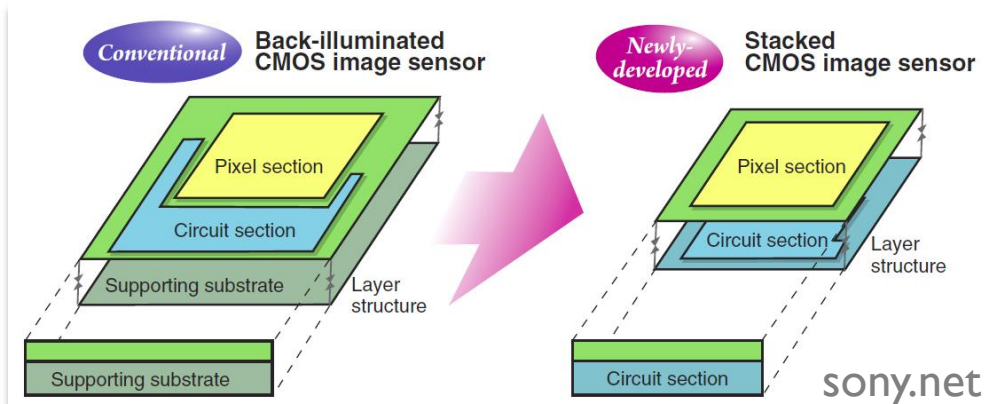# Age 50: Children

Greg Yeric, ARM

Children of Moore's Law

- Solar Panel
- Sensor
- Modem

$5000
(or $149/month over 5 years)

$370

The city of Philadelphia reduced weekly trash collections from 17 to 3
The city of Barcelona estimates $4B savings over 10 years

Child of Moore's Law: Photovoltaics

- $0.5 on Alibaba:
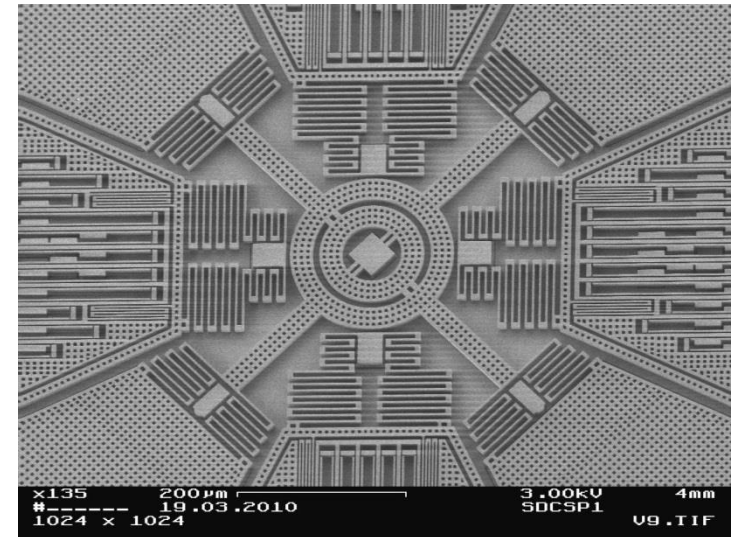
Child of Moore's Law:
MEMS sensors

Chris Wasden at 2014 MEC, via semiwiki

- MEMs can further benefit from VLSI "ization" (ecosystem standards, etc.)

1T SENSORS IN 10 YEARS

Year	Unit Price	Units Sold	Industry Revenues	Developed Population	MEMS Rev/ Person	MEMS Unit/ Person
2005	30.000	46,666,667	5,000,000,000	4,000,000,000	1.25	0.01
2010	15.000	466,666,667	7,000,000,000	4,000,000,000	1.75	0.12
2015	1.800	8,333,333,333	15,000,000,000	4,000,000,000	3.75	2.08
2020	0.216	138,888,888,889	30,000,000,000	4,000,000,000	7.50	34.72
2025	0.026	1,388,888,888,889	60,000,000,000	4,000,000,000	15.00	347.22

2007 the average cost of an accelerometer sensor was $3. In 2014, the average was 54 cents

IEDM 2015

Future sensors: Also children of Moore's Law progress
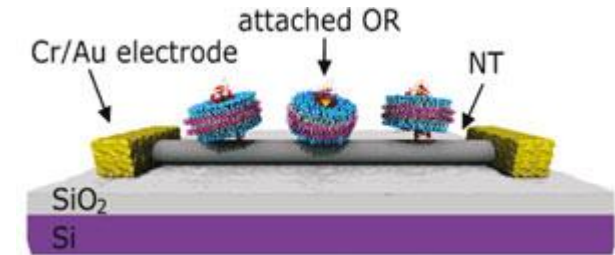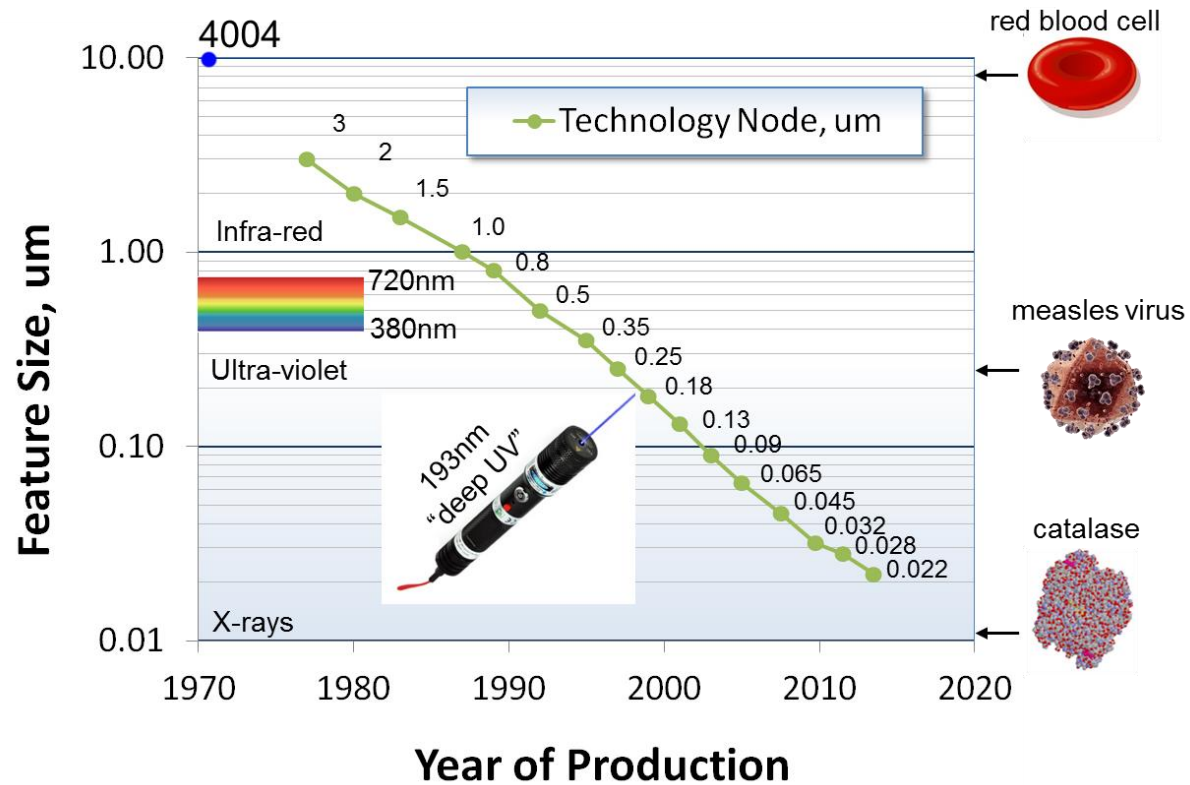
Adamant technologies, e.g.

Moore's Law progress created the IoT

1. Vanishing edge node compute
2. Ubiquity of smart phones
   - A thing-to-person hub
3. Ubiquity of Internet, and plummeting cost to connect to it
4. Plummeting cost of (and richer set of) sensing
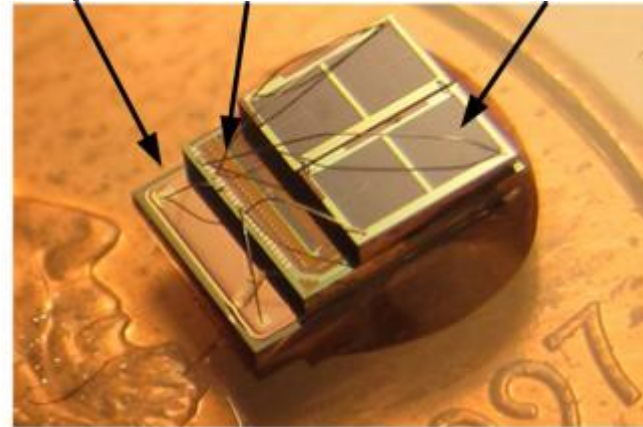5. Reduced cost of energy harvesting

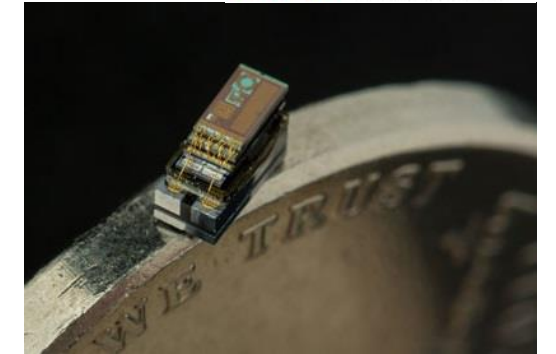Digital World

IoT

People

Physical World
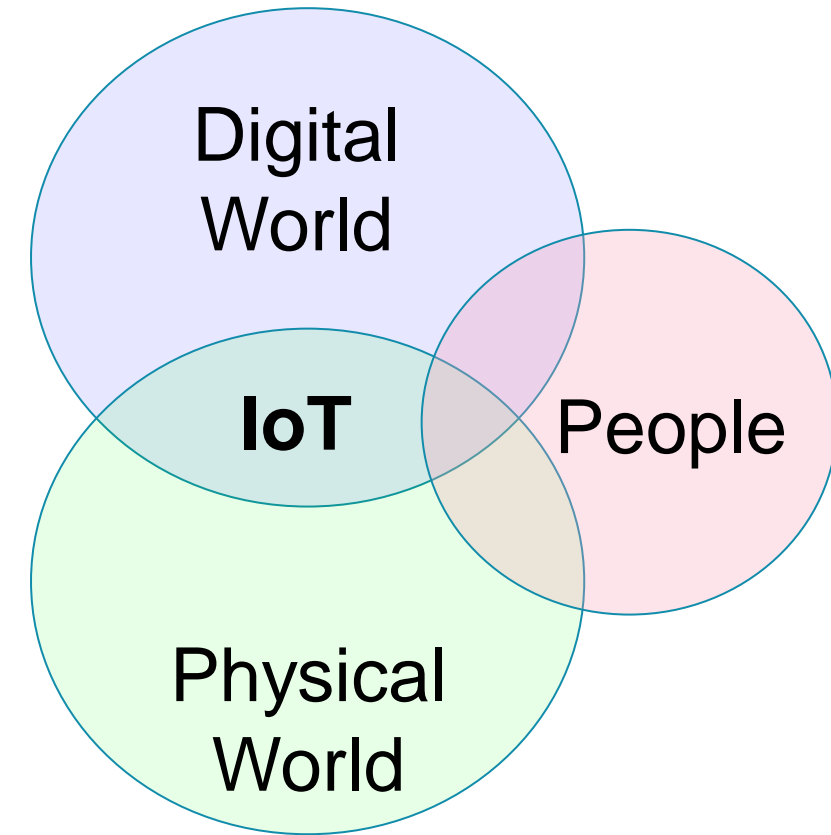
reference 31

Example IoT: Michigan Micro Mote

- 3DIC for form factor
- Energy Harvesting →
- Near Threshold operation
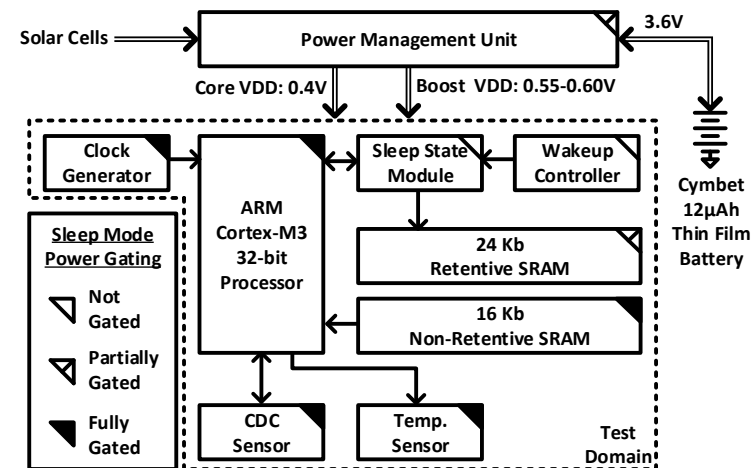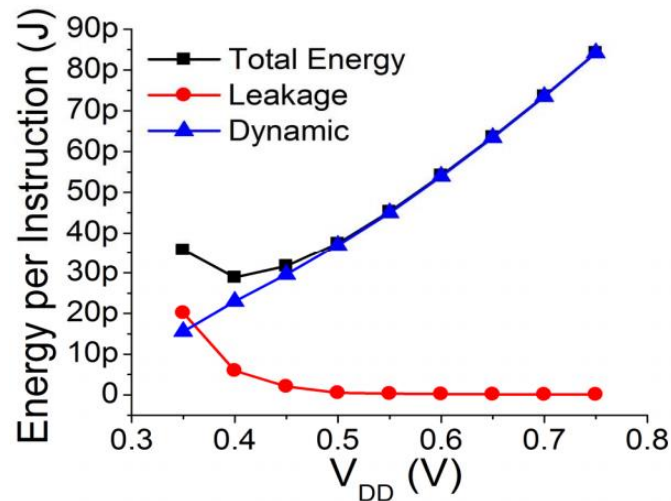   - Variation 10x more important

Battery   **Processor**   **Solar Cells**
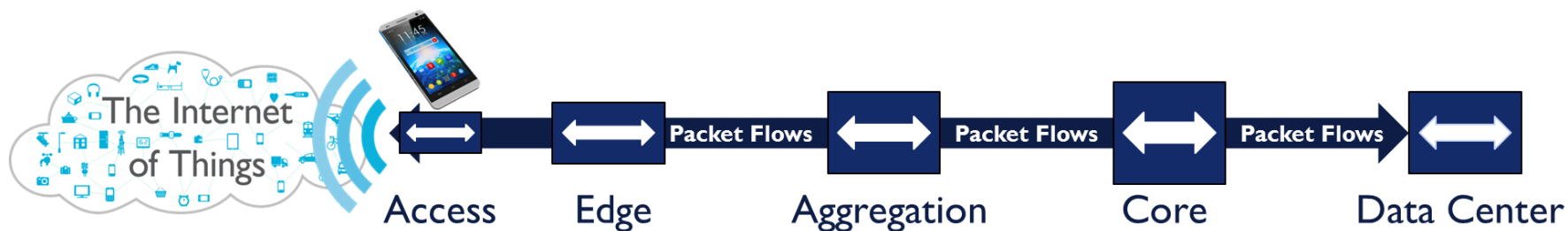
G. Chen et al., ISSCC, 2010.

Energy per Instruction (J) vs V_{DD} (V)

- Total Energy
- Leakage
- Dynamic

Solar Cells → Power Management Unit ← 3.6V

Core VDD: 0.4V   Boost VDD: 0.55-0.60V

Clock Generator

ARM Cortex-M3 32-bit Processor

Sleep State Module   Wakeup Controller

24 Kb Retentive SRAM

16 Kb Non-Retentive SRAM

Sleep Mode Power Gating

- Not Gated
- Partially Gated
- Fully Gated

CDC Sensor   Temp. Sensor

Test Domain

Cymbet 12μAh Thin Film Battery

2015

Greg Yeric, ARM

Energy Efficiency: Things you can do with 100pJ

- Run a Cortex®-M0 for 10 cycles
- Write one bit of flash
- Write ~300 bits of DRAM or SRAM
- Send ~5 bits across LPDDR4
- Transmit 2 bits of UWB data
- Transmit 0.02 bits over Bluetooth LE
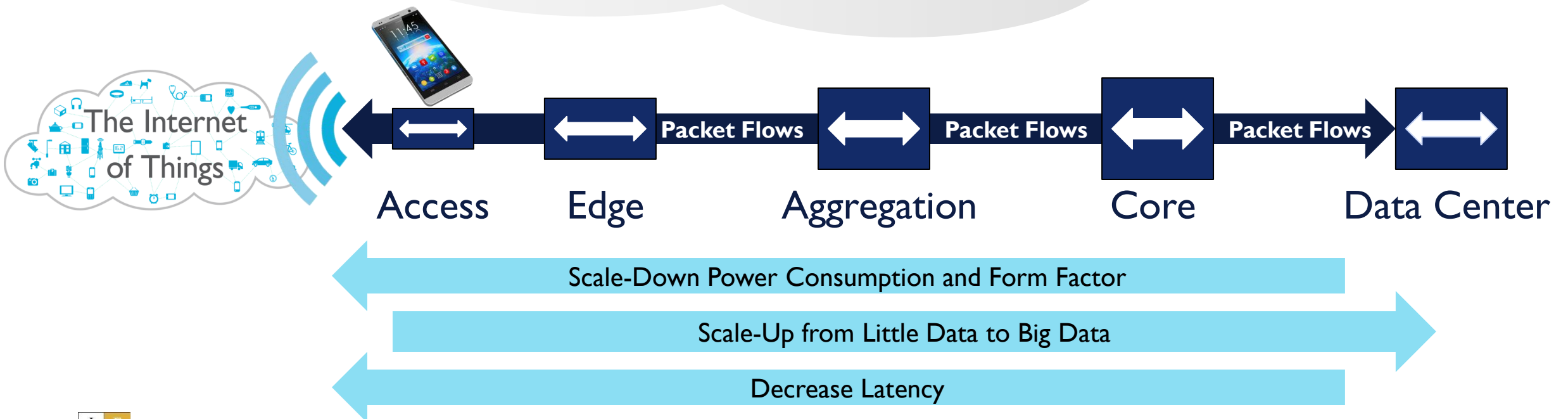- Drive an electric car 100fm (@1MJ/km) ~0.05% of the distance across Si atom

The IoT is an NVM problem

The Internet of Things — Access — Edge — Packet Flows — Aggregation — Packet Flows — Core — Packet Flows — Data Center
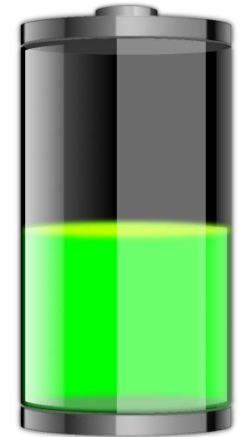
Energy costs to transmit, compute, and store data will define the shape of the IoT
VSLI Technology advancements will re-write the boundary conditions

IEDM 2015

Greg Yeric, ARM

Intelligent Flexible Cloud

- Applications run where the data is, independent of the network node
- Heterogeneous compute is distributed into the network
- ALL OF THIS COMMUNICATION MUST BE SECURE

The Internet of Things

Access   Edge   Aggregation   Core   Data Center

Packet Flows   Packet Flows   Packet Flows

Scale-Down Power Consumption and Form Factor

Scale-Up from Little Data to Big Data

Decrease Latency

IEDM 2015

Internet of Things at IEDM 2015

13.1 Ultra Low Power Sensor Platforms for Personal Health and Personal Environmental Monitoring

19.4 Free Form CMOS Electronics: Physically Flexible and Stretchable

13.5 Precision Mass Measurements in Solution Reveal Properties of Single Cells and Bioparticles

13.6 Fabrication and Analysis of SiN Nanopores for Direct DNA Sequencing

18.7 Output Enhancement of Triboelectric Energy Harvester by Micro-Porous Triboelectric Layer,

19.1 Flexible Electronics Manufacturing: Flexible Digital x-ray to Flexible Hybrid Electronics

19.5 Large Area Sensing Surfaces: Flexible Organic Printed Interfacing Circuits and Sensors

19.8 Flexible 2D FETs using hBN Dielectrics

Session 25: Circuit Device Interaction – **More than Moore – Value Added Technologies**

25.4 Low-Cost and TSV-free Monolithic 3D-IC with Heterogeneous Integration of Logic, Memory and Sensor Analogy Circuitry for Internet of Things,

25.5 New Devices for Internet of Things: A Circuit Level Perspective

25.7 An Integrated Silicon Photonics Technology for O-band Datacom,

2015

Child of Moore's Law: Image Recognition

- Everyone carries a camera → storage is "free" → Internet of Images
- Hardware scaling allows semi-affordable Machine Learning (ML):

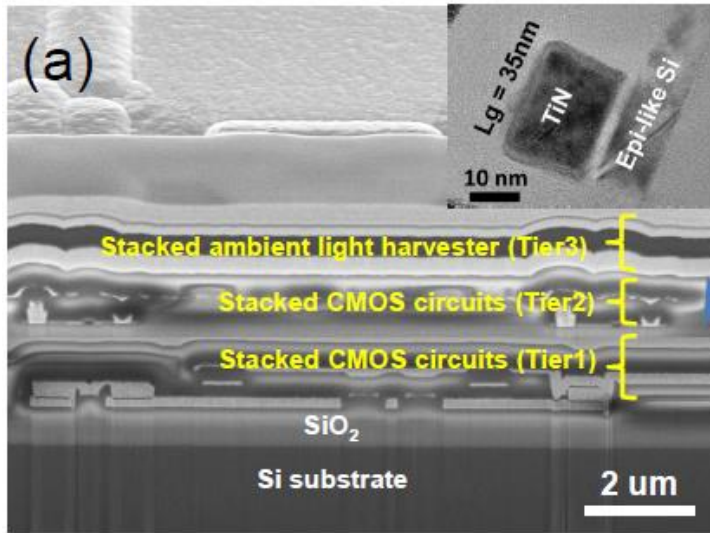(>900M operations/image)
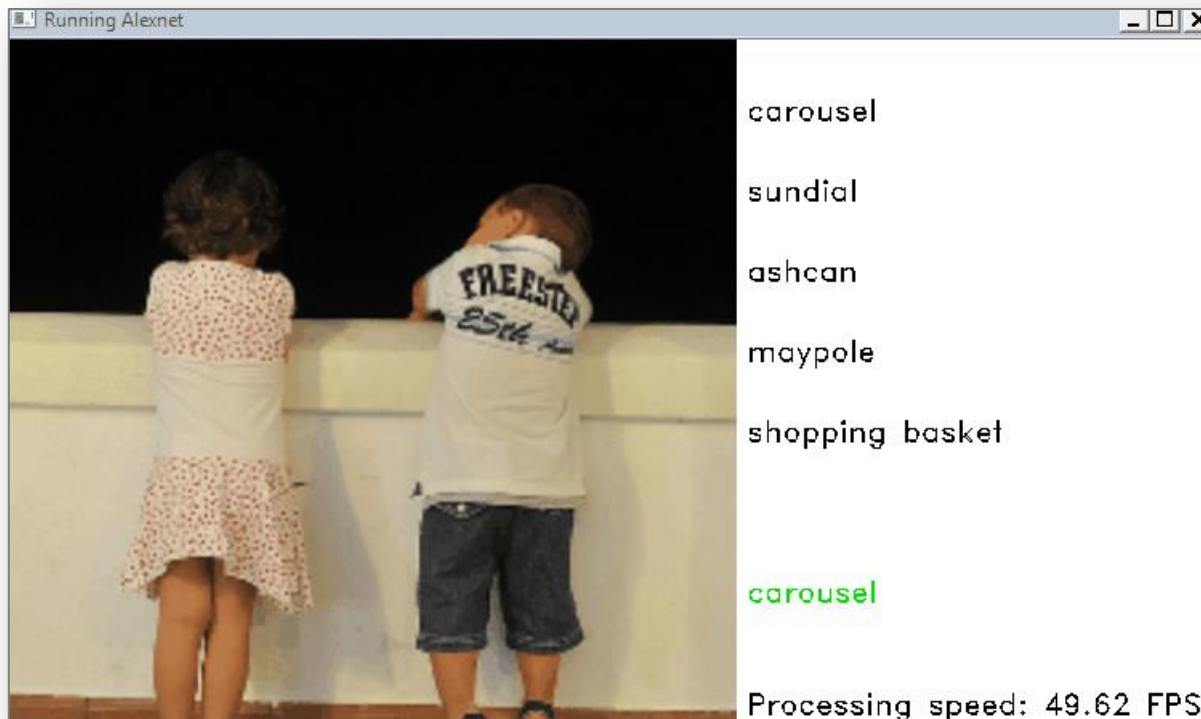
Machine Learning:

Use transistors instead of expert developers

"this is a face"
"this is a QR code"
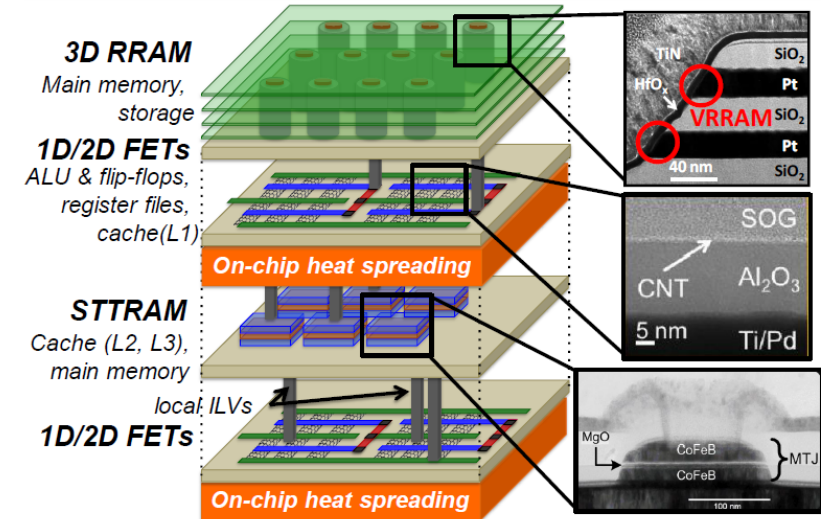
Putting it all together: Achieving the 100x

- Optimized layer processing
- New bandwidth paradigms
- New form factors

Shen et al., IEDM 2014

Neuromorphic computing at IEDM 2015

4.5   Memristive Based Device Arrays Combined with Spike Based Coding Can Enable Efficient Implementations of Embedded Neuromorphic Circuits

4.1   Brain-inspired Computing with Emerging Memories

4.4   Large-Scale Neural Networks Implemented with Non-volatile Memory as The Synaptic Weight Element: Comparative Performance Analysis

4.6   A Mixed-Signal Universal Neuromorphic Computing System

4.9   DARPA Neurocomputing

4.7   Oxide Based Nanoscale Analog Synapse Device For Neural Signal Recognition System

17.2   Investigation of the Potentialities of Vertical Resistive RAM (VRRAM) for Neuromorphic Applications,

17.7   Optimized Learning Scheme for Grayscale Image Recognition in a RRAM Based Analog Neuromorphic System
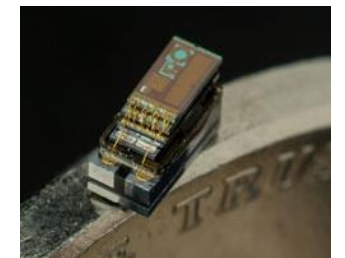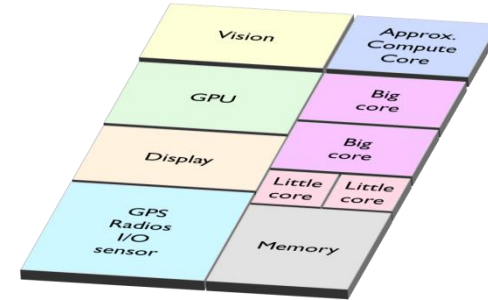
1.2 Quantum Computing in Si,

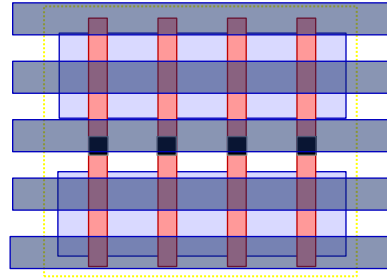# Past transistor pipeline

Disruptiveness

Academia

Industry

Ferromagnet   Ferromagnet

Ge, InGaAs

STI

Si

1971   2010   2015   2020   2025

Greg Yeric, ARM

Future Transistor pipeline

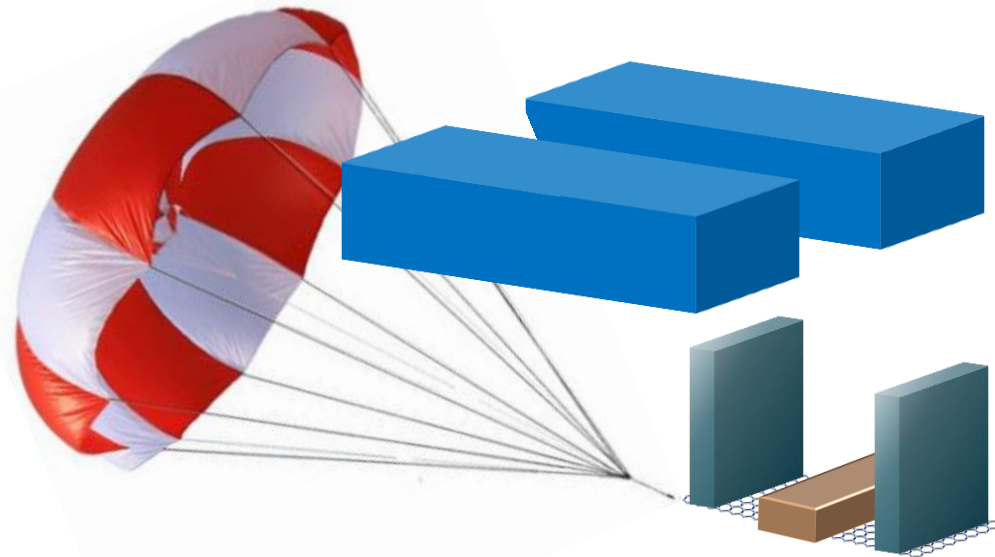Academia

Industry

Disruptiveness

1971   2010   2015   2020   2025

IEDM 2015
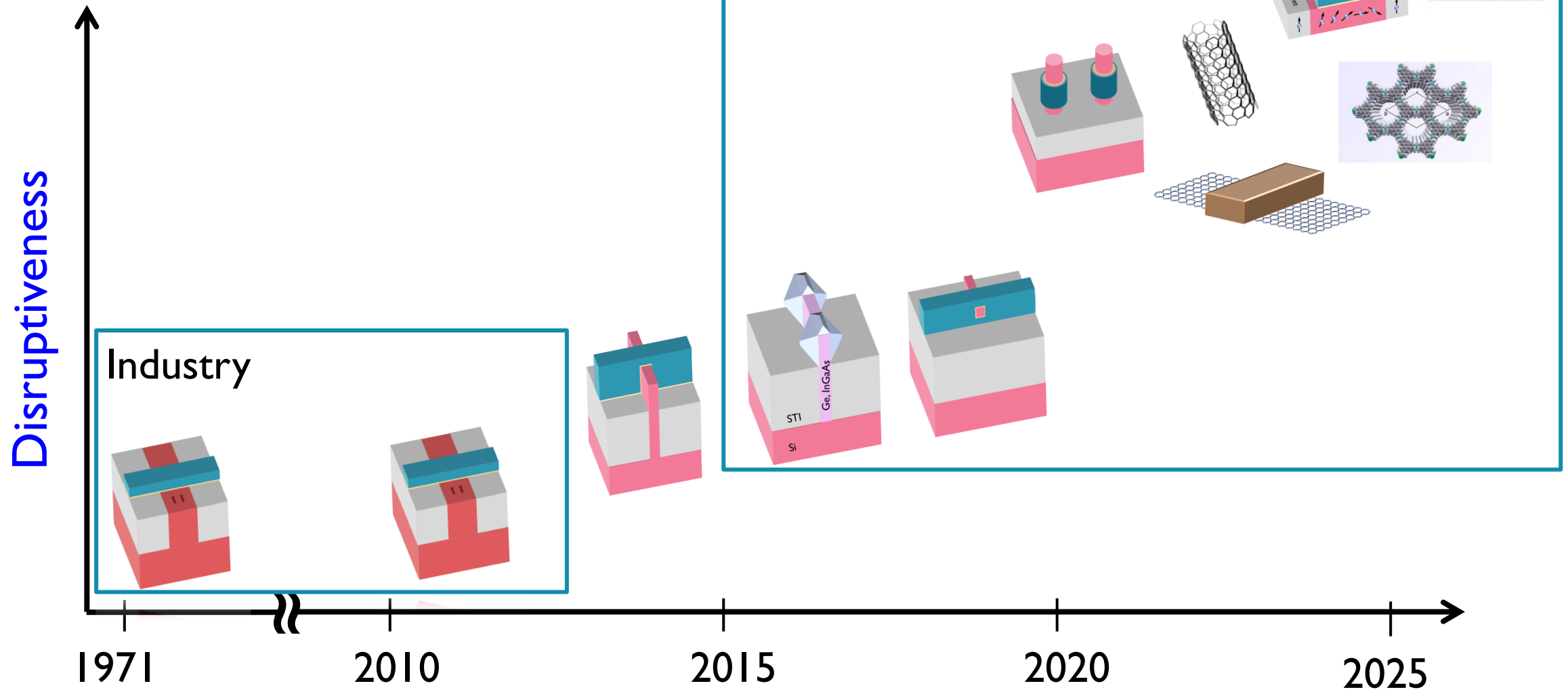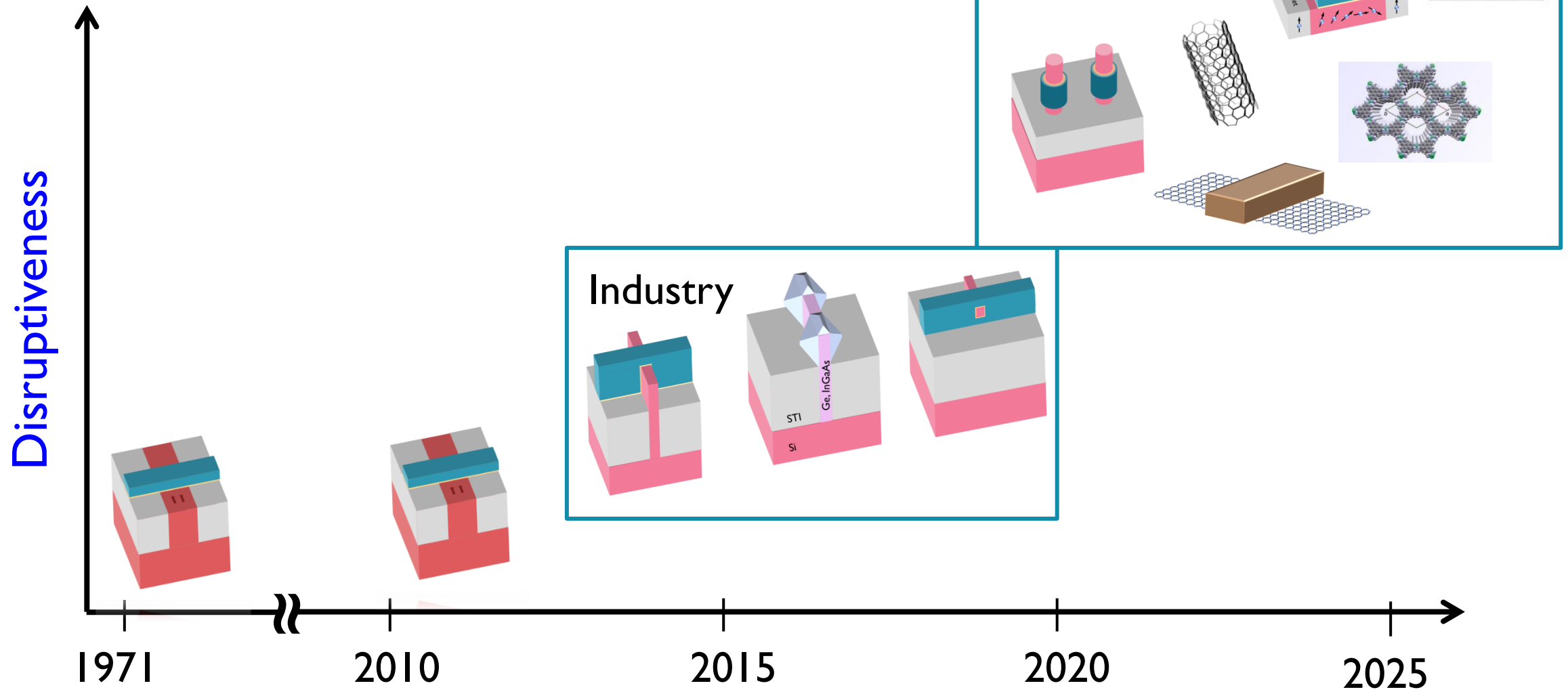
What can you do with a future transistor?

Needed: technology Benchmarking

(SRC ref #20)

ASU ASAP 7nm Predictive PDK

2015

© 2015 ARM   Greg Yeric, ARM
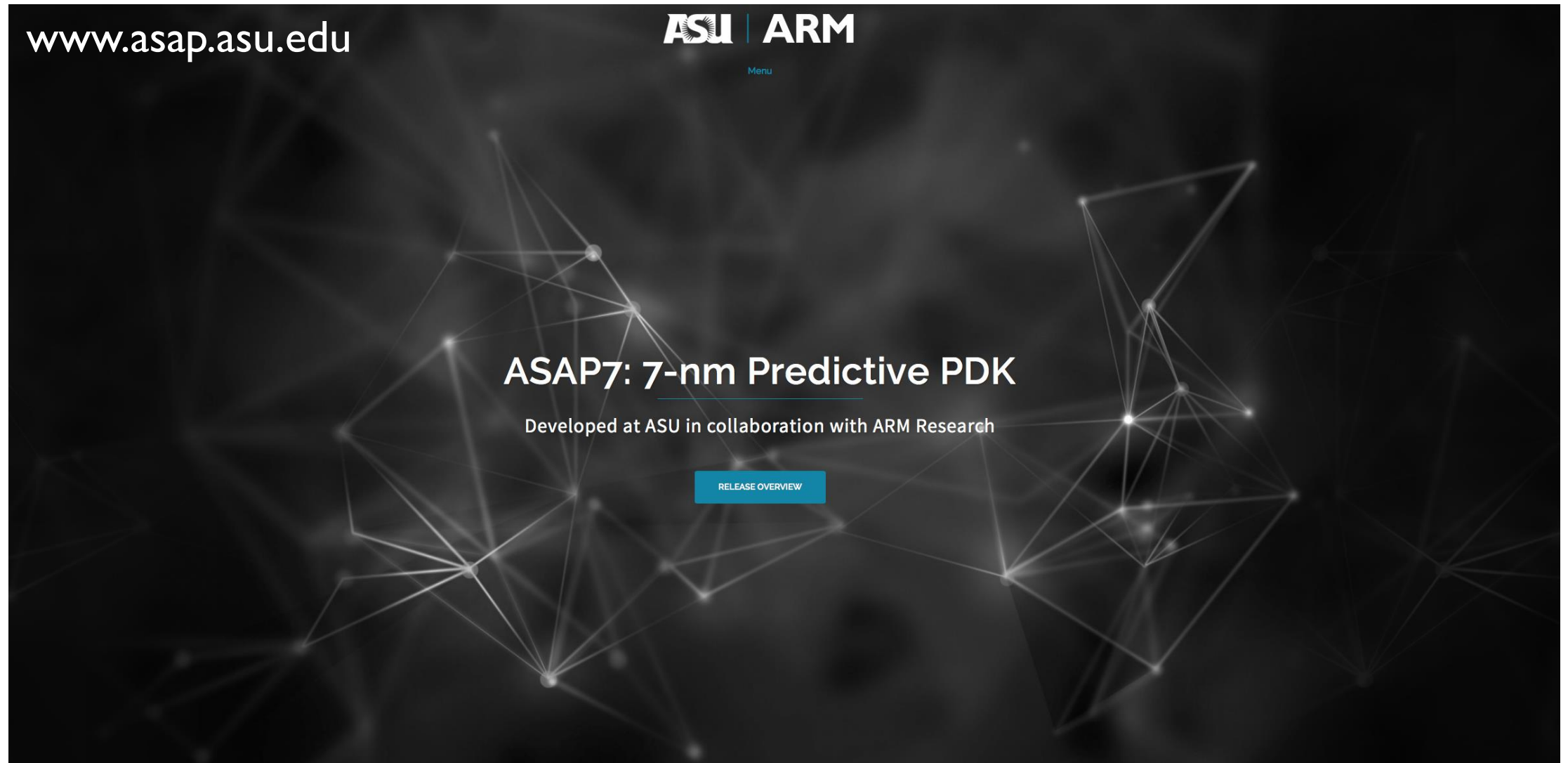
PDK enabling more accurate university research

www.asap.asu.edu

The ASAP 7nm Predictive PDK was developed at ASU in collabora

The ASAP 7nm PDK contains SPICE-compatible FinFET device models (BSIM-CMG),
Extraction Deck for the 7nm technology node. For more details regarding the technica
this process design kit is provided as an academic

Predictive Technology Model

- Introduction
- Latest Models
- Nano-CMOS
- Post-Silicon
- Interconnect
- Reliability

LATEST MODELS

Typical SPICE model files for each future generation are available here.

Attention: By using a **PTM** file, you agree to acknowledge both the URL of PTM: http://
publications in all documents and publications involving its usage.

New!

June 01, 2012:
PTM releases a new set of models for multi-gate transistors (**PTM-MG**), for both HP and LSTP
BSIM-CMG, a dedicated model for multi-gate devices.
Acknowledgement: PTM-MG is developed in collaboration with ARM.

Please start from models and param.inc.
- 7nm PTM-MG HP NMOS, HP PMOS, LSTP NMOS, LSTP PMOS
- 10nm PTM-MG HP NMOS, HP PMOS, LSTP NMOS, LSTP PMOS
- 14nm PTM-MG HP NMOS, HP PMOS, LSTP NMOS, LSTP PMOS
- 16nm PTM-MG HP NMOS, HP PMOS, LSTP NMOS, LSTP PMOS
- 20nm PTM-MG HP NMOS, HP PMOS, LSTP NMOS, LSTP PMOS

The entire package is also available here: PTM-MG

Saurabh Sinha, Lucian Shifren, Brian Cline, Greg Yeric.

Other students of the EEE598 Special Topics course:
Alan Sam, Nalim Gupta, Rohit Musalay, Akash Thakare, Srividhya Jambunathan, Ramana
Rao Pandeshwar, Jayesh Sohanlal, Chandrakanth Puttaswamygowda, Adesh Namekumar,
Varun Kaushik, Sanyogita Singh.

**Cells:**
Sai Chaitanya Reddy, Punit Shah, Anant Mithal.

Acknowledgements
Mentor Graphics: Tarek Ramadan (for excellent DRC and LVS training at ASU).

What's in a PDK?

1. Transistor models:
   BSIM-CMG FinFET transistor models:
   3 VTs and corners
2. Technology files for schematic capture/layout
3. DRC
4. LVS
5. Extraction

2015

PDK enabling more accurate university research

ww.asap.asu.edu

Release Overview

The ASAP 7nm Predictive PDK was developed at ASU in collaboration with ARM Research. The PDK is available

The ASAP 7nm PDK contains SPICE-compatible FinFET device models (BSIM-CMG), Technology files for Cadence Virtuoso, Desig
Extraction Deck for the 7nm technology node. For more details regarding the technical specifications of the PDK, please refer the P
ty and the resulting desi
ng the need for Mentor G
ense.

ributors

wrence T. Clark, Vinay V

th Sinha, Lucian Shifren

students of the EEE598
am, Nalim Gupta, Rohit
ndeshwar, Jayesh Soha
Kaushik, Sanyogita Sing

Sai Chaitanya Reddy, Punit Sh

Comes with the kit at download.

Related Publications
TBD

Acknowledgements
Mentor Graphics: Tarek Ramadan (for excellent DRC and LVS training at ASU).
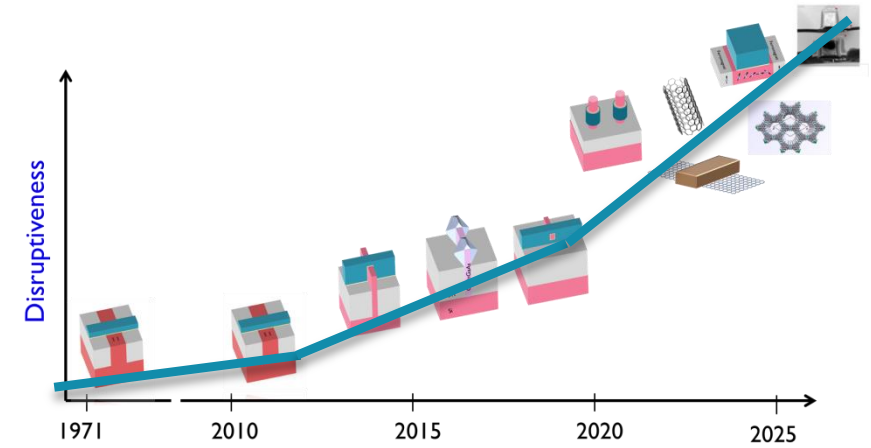
Summary

- Moore's Law and IEDM: "not free" on lorry to free in magazine
    - BUT, cost scaling is slowing down. There are no near term magic bullets.
    - Task/opportunity for technology investment: Pull in the curves
        - Fab / Tools / Process Integration / Device / Circuit (DTCO)
    - Dennard scaling is as challenged, and as important (FETs and wires)
- SoC: Continued informed innovation. Leverage 3DIC and novel memory
    - Energy efficiency vs. utilization → heterogeneity
    - Complex future technology choices need transistor-to-system benchmarking
- Breadth of future systems strains simple Moore's Law answers
    - Slowing fundamental scaling increases opportunity for radical change
    - Exciting new systems leverage the children of Moore's Law: Sensors, MEMS, ML
    - Systems must understand (and guide) underlying technology options
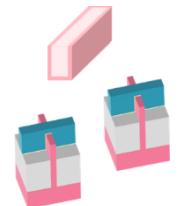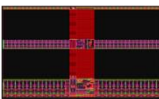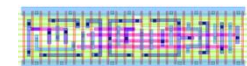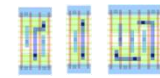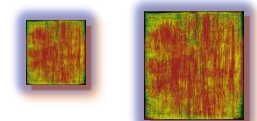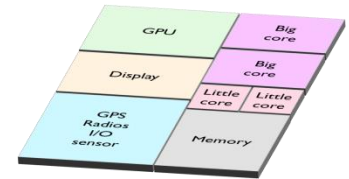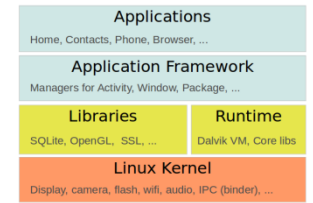
Moore's Law at 50, IEDM, and you: Summary

■ **You live in interesting times**
- New device physics in logic and memory
- New levels of design efficiency
- Tighter pipeline to new technologies

We
■ ~~You~~ **are in the right place**
- Circuits and Systems help define device choices

Together, at the system level, I expect Moore's Law level progress well past my own retirement