# Impact of Technology Evolution

Damien Querlioz

Chargé de recherche CNRS, https://sites.google.com/site/damienquerlioz/

*Centre de Nanosciences et de Nanotechnologies*
*(ex-Institut d'Electronique Fondamentale)*
*Université Paris-Saclay, CNRS,* ***Orsay***

# Microelectronics is changing!

HPCwire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

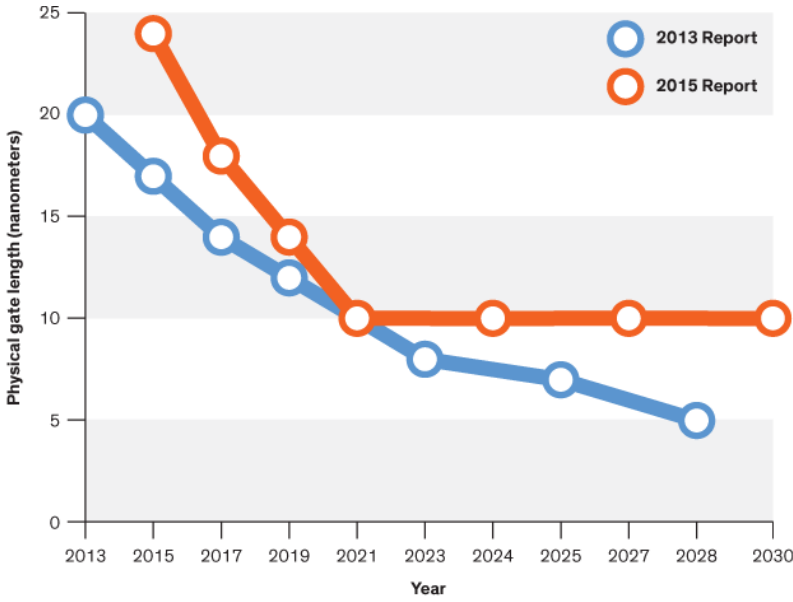
- Home
- Technologies
- Sectors
- Exascale
- Resources
- Specials
- Job Bank
- About


**Transistors Won't Shrink Beyond 2021, Says Final ITRS Report**

By Tiffany Trader

July 28, 2016

The final International Technology Roadmap for Semiconductors (ITRS) is now out. The highly-detailed multi-part report, collaboratively published by a group of international semiconductor experts, offers guidance on the technological challenges and opportunities for the semiconductor industry through 2030. One of the major takeaways is the insistence that Moore's law will continue for some time even though traditional transistor scaling (through smaller feature sizes) is




**Final** report of the ITRS (2015)

The whole model of which microelectronics industry has always worked is crumbling, and might soon come to an end!

*What now?*

# Outline of the class

- The issue of transistor variability
- Modern CMOS architectures and their implication for design
- What comes next?
- Getting more of the devices we have: *better than Worst Case* Design

# Outline of the class

- The issue of transistor variability
- Modern CMOS architectures and their implication for design
- What comes next?
- Getting more of the devices we have: *better than Worst Case* Design

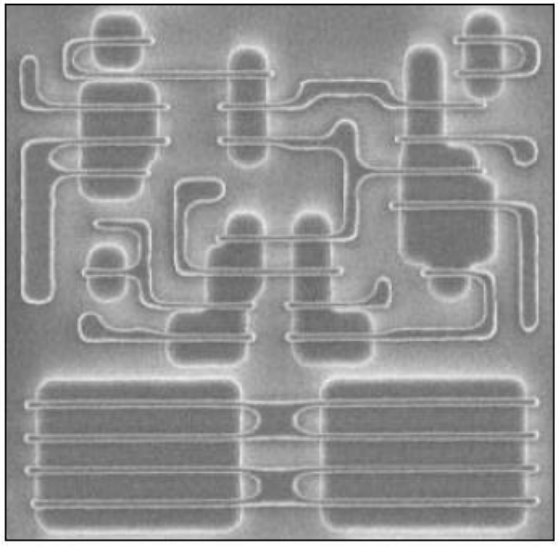
# The issue of device variability

- Since the 65nm node, variability has become a huge concern
  - Process: *each die is different*
  - Mismatch: *each transistor in a die is different*

# Restrictions to reduce variability

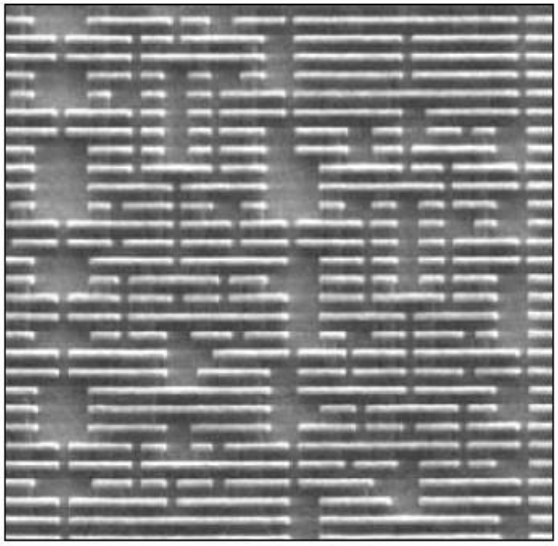
## Layout Restrictions 65nm to 32nm

### 65 nm Layout Style



- Bi-directional features
- Varied gate dimensions
- Varied pitches

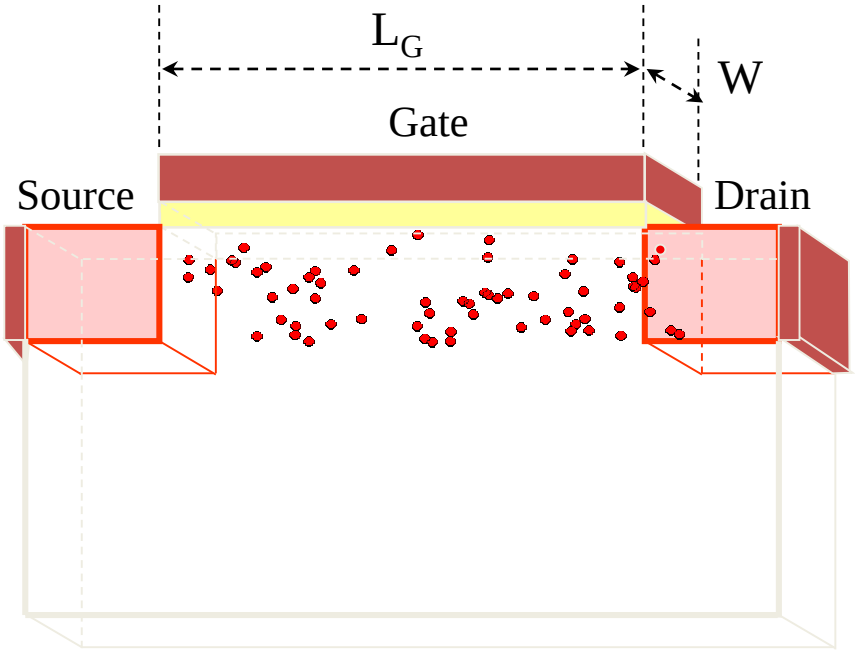
### 32 nm Layout Style



- Uni-directional features
- Uniform gate dimension
- Gridded layout

# But high variability becomes intrinsic

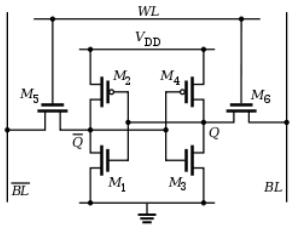
- Only a few dozens of dopants in the channel

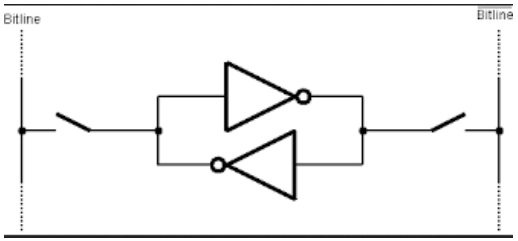


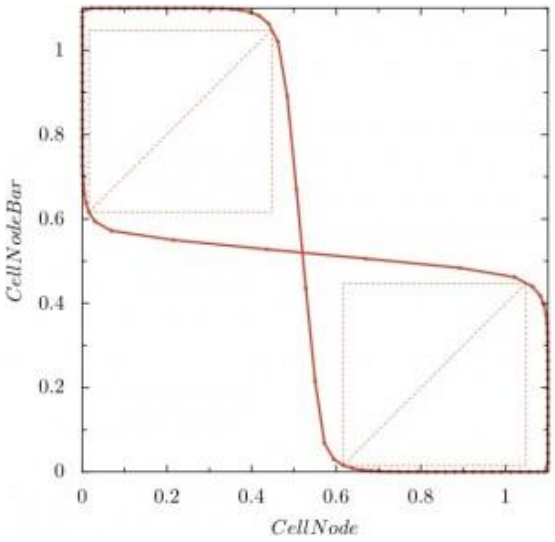

Variation in number of dopants causes unavoidable mismatch between transistors

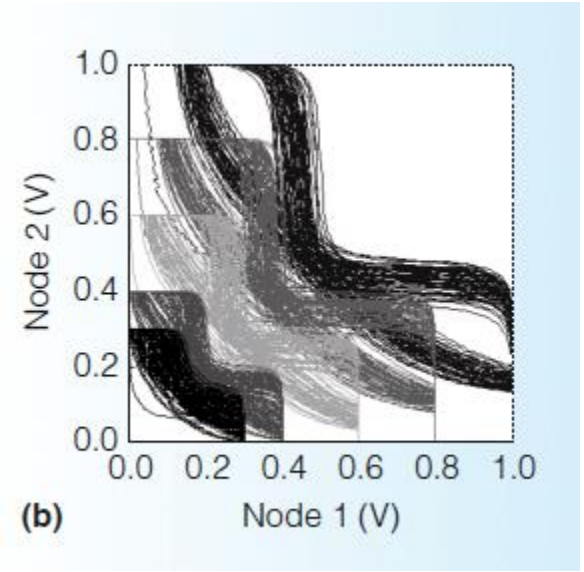
# The issue with variability

- SRAM cell (e.g. cache in microprocessors)



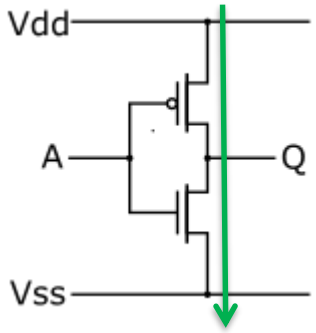

Rogenmoser et al, IEEE Micro

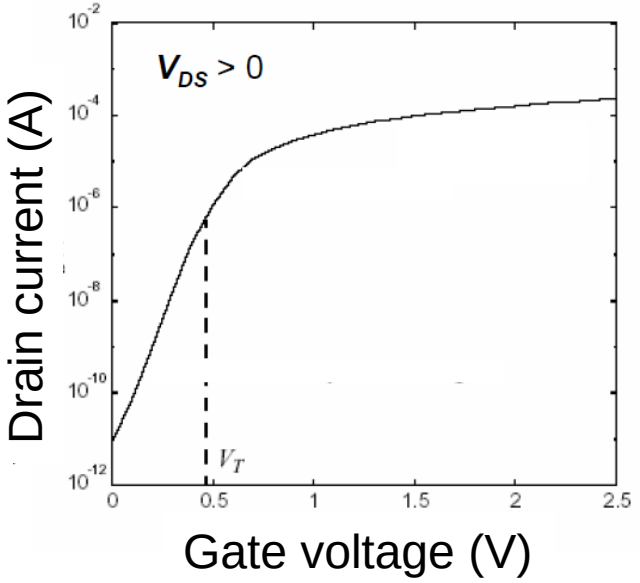
No variability
Noise is not a big concern

Variability
Noise is a big concern if we scale voltage!!!

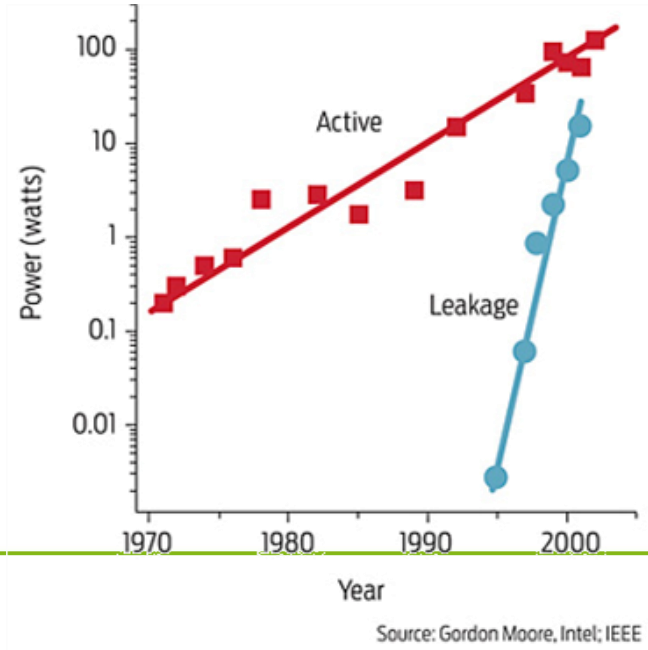
# Other severe concern: leakage

- Transistors leak




Leakage used to be a non issue



# But now…

- Static and dynamic power are similar
- Static can be dominant for low usage

Leakage prevents us from reducing threshold voltage and therefore supply voltage too much

# Implications

- Due to variability & leakage, conventional CMOS progress has become stuck at 30nm

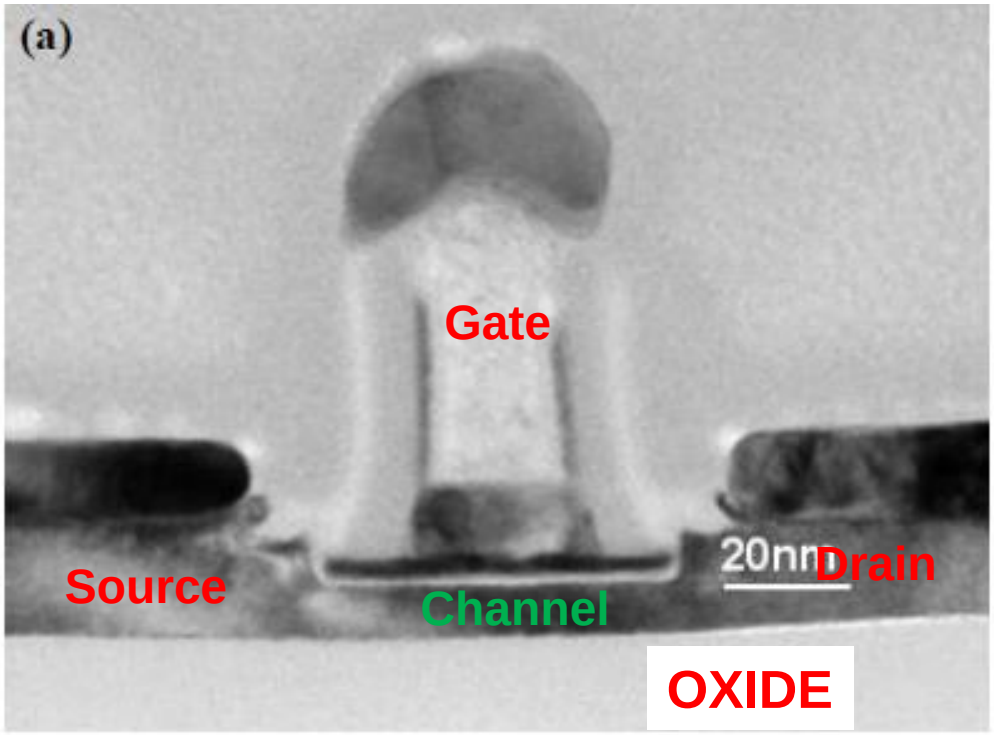
- For <30nm technology nodes, to increase energy efficency, new architectures of transistors have been introduced

# Outline of the class

- The issue of transistor variability
- Modern CMOS architectures and their implication for design
- What comes next?
- Getting more of the devices we have: *better than Worst Case* Design

# Fully Depleted SOI (FDSOI)

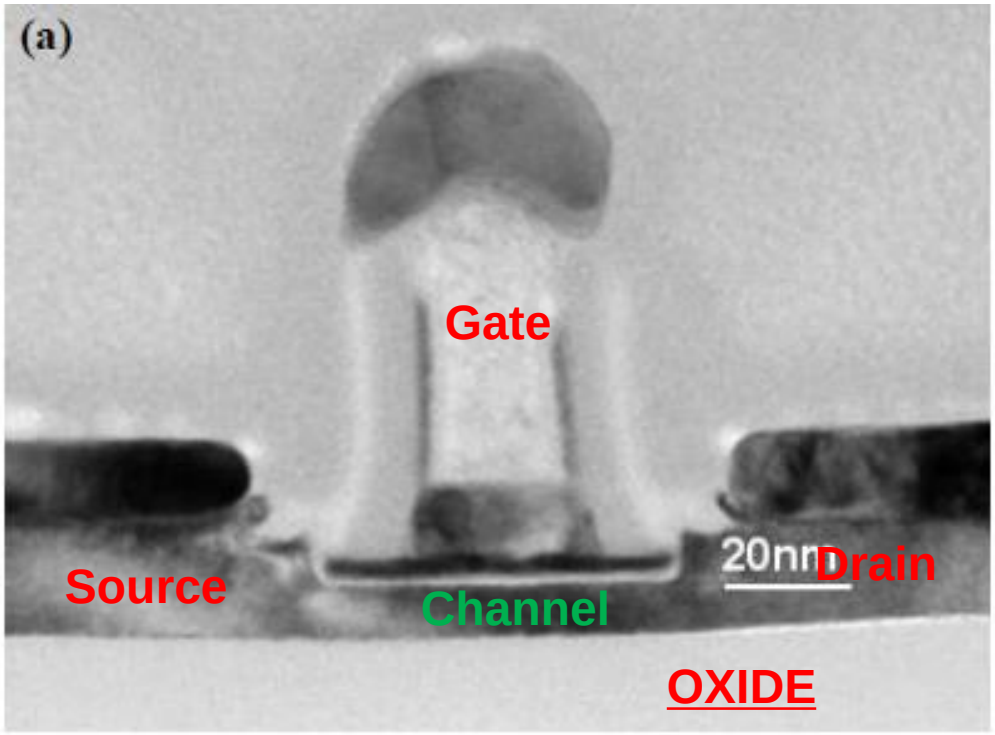
- ST Microelectronics since 28nm
- Now Samsung and Globalfoundries




Silicon on Insulator wafer

Weber et al, IEDM 2008

# Fully Depleted SOI (FDSOI)



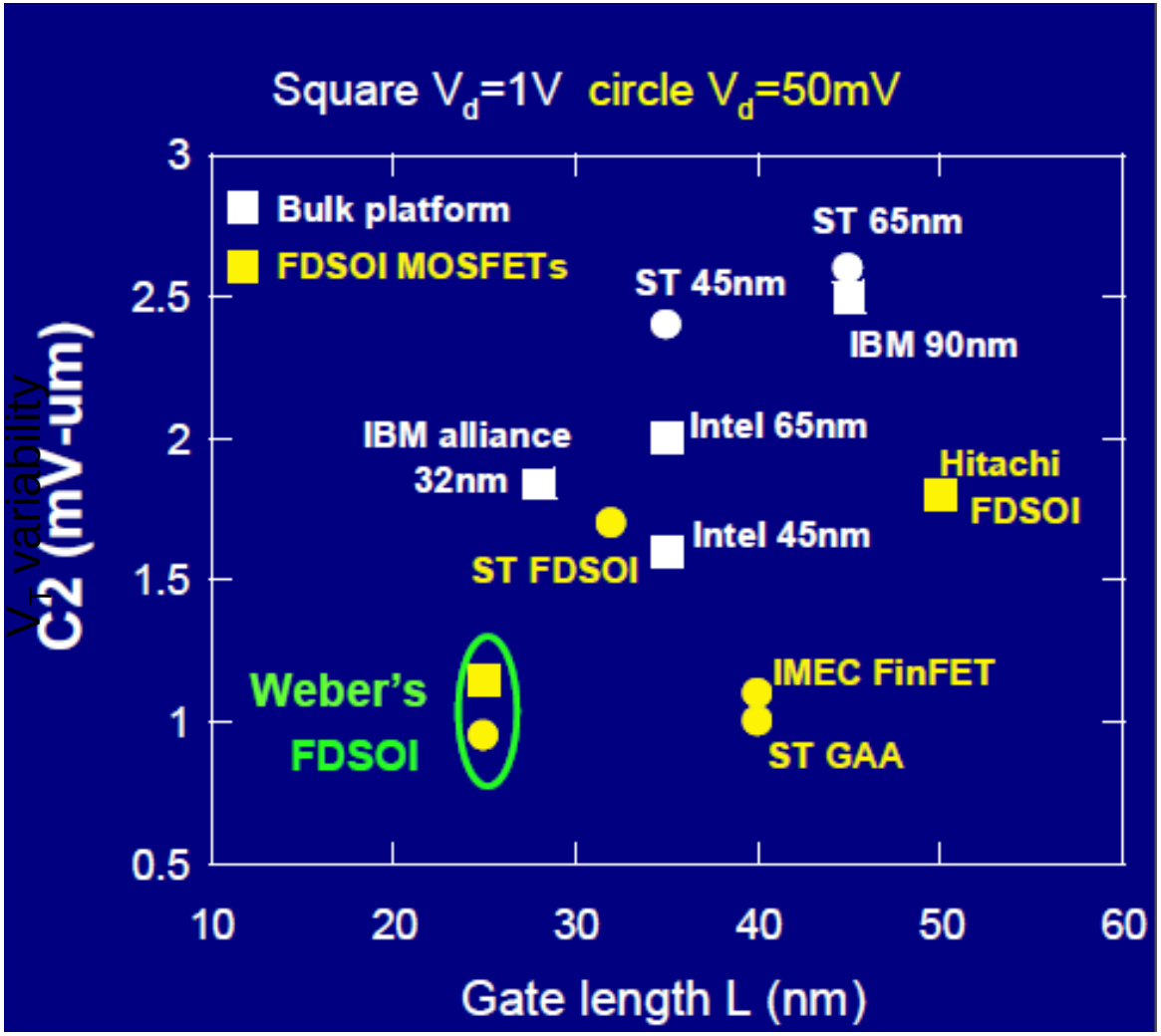

Channel is
created physically

- Low leakage

- No need for
channel doping!

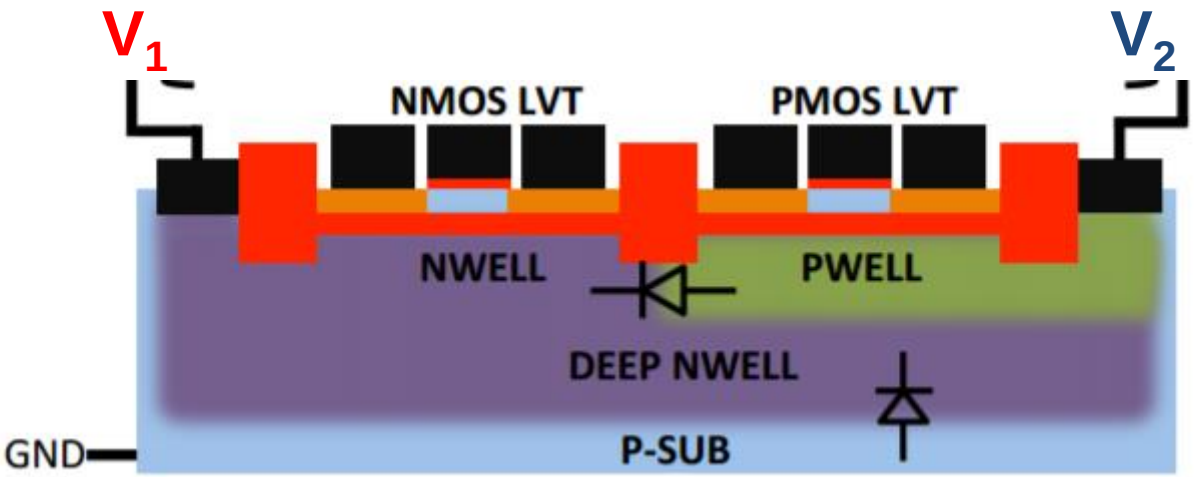
Weber et al, IEDM 2008

# No channel doping = low variability




Weber et al, IEDM 2008

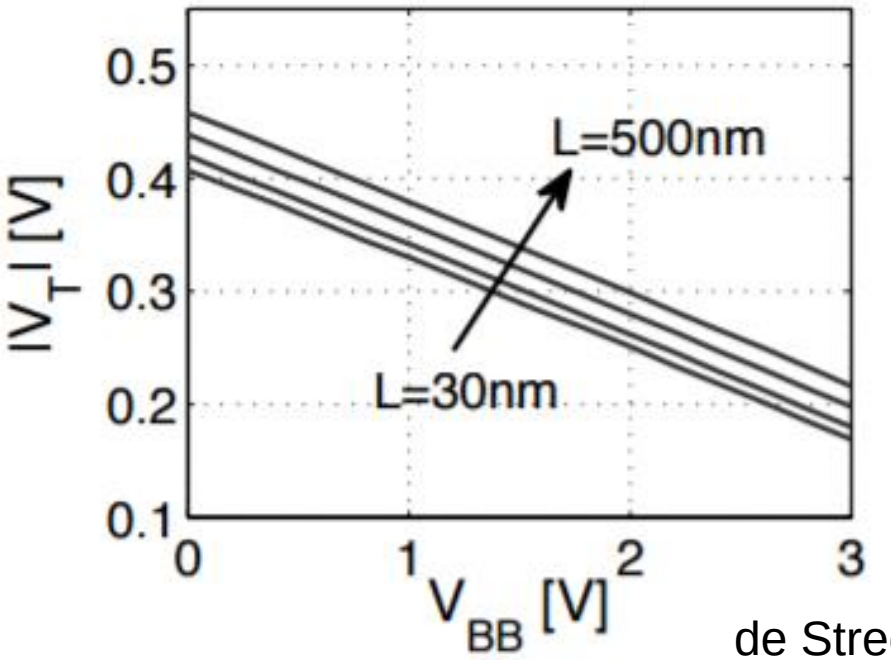
# FDSOI: opportunitunities for design




de Streel et al, JLPEA 2014

- Silicon behind channel can act as a « second gate »

# Imact of back biaising




de Streel et al, JLPEA 2014

(Effect already exists with bulk, but much more intense with FDSOI)

Why is it useful?

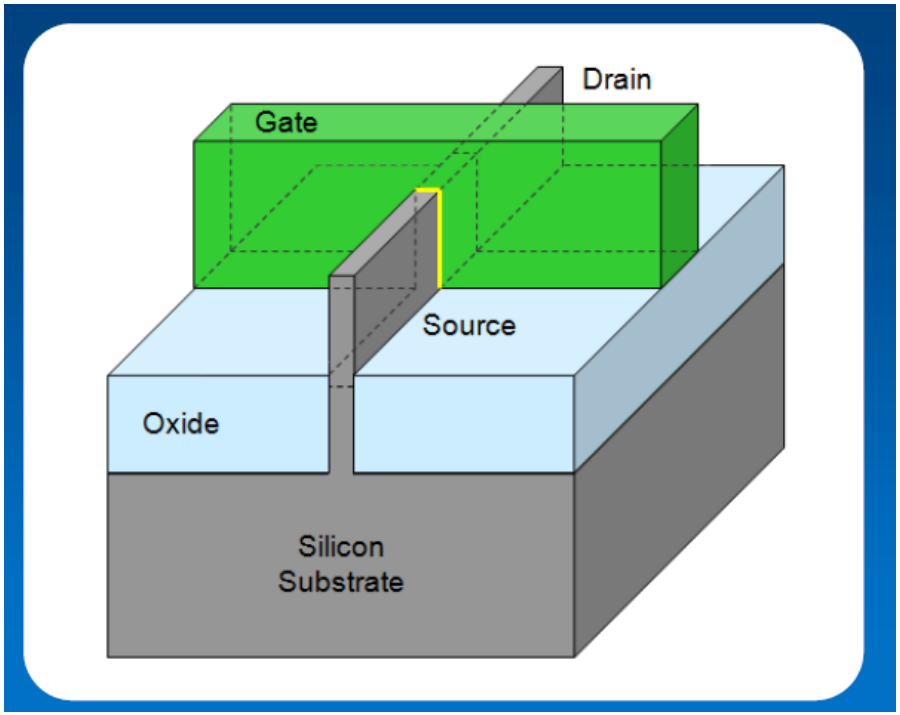
# How would you use it for a MCU?

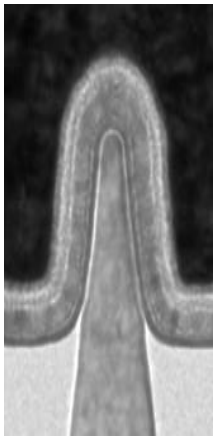
- *Class activity*

- ST microelectronics claims -50% energy reduction typically possible

# The other option : FinFET

- Intel: since 22nm
- Also TSMC, Samsung, Gloablfoundries…



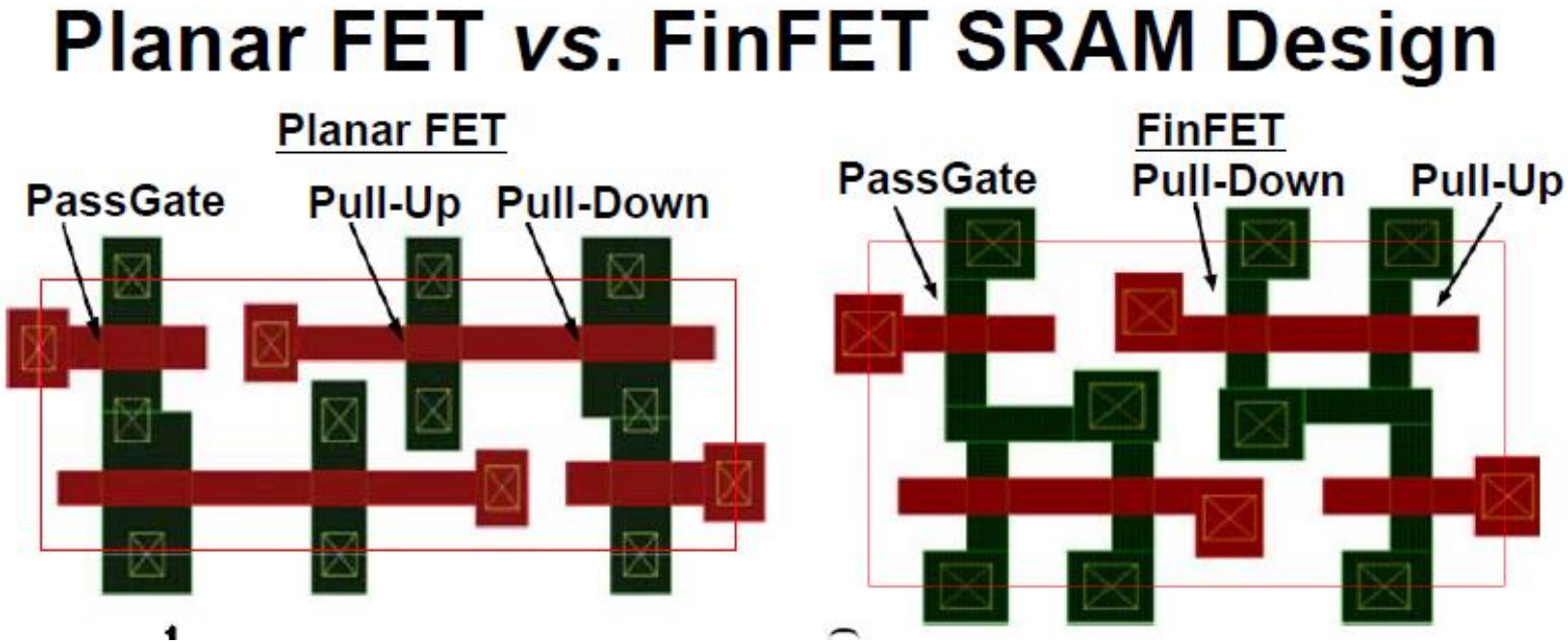



Burried channel surrounded by gate!

Undoped channel

# Impact of FinFET for design

- W is quantized
- Small area overhead in layout




Nuo Xu, UC Berkeley

# FinFET vs. FDSOI

- FDSOI +
  - Back biasing
  - Design very similar to bulk
- FinFET +
  - No need for expensive SOI substrate
  - Scales better

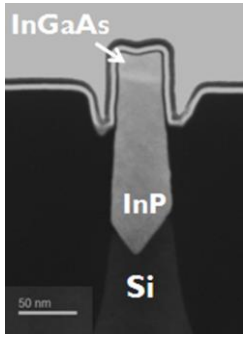
Severe competition!

# Outline of the class

- The issue of transistor variability
- Modern CMOS architectures and their implication for design
- What comes next?
- Getting more of the devices we have: *better than Worst Case* Design

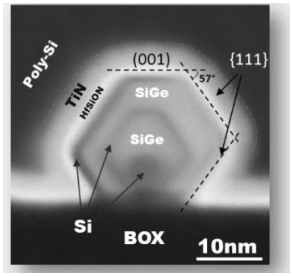
# Beyond FinFETs and FDSOIs?

- **Many ideas:**
- III-V FinFETs



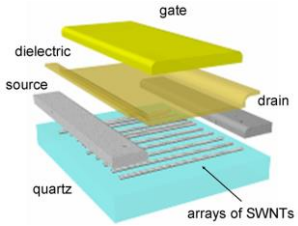

IMEC

- Nanowire FETs




LETI

- Nanotube FETs




UIUC

# Beyond FinFETs and FDSOIs?

**But for now, leakage, perf or variability of these solutions is always terrible!**

- Considerable development cost would be necessary for them to succeed
- There are also fundamental reasons that limit the performance of these alternative devices
  - *Will be seen in the* ***Nanoarchitecture class***

# Beyond FinFETs and FDSOIs?

***It is very possible that MOSFET progress stops very very soon!***

HPCwire

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

- Home
- Technologies
- Sectors
- Exascale
- Resources
- Specials
- Job Bank
- About

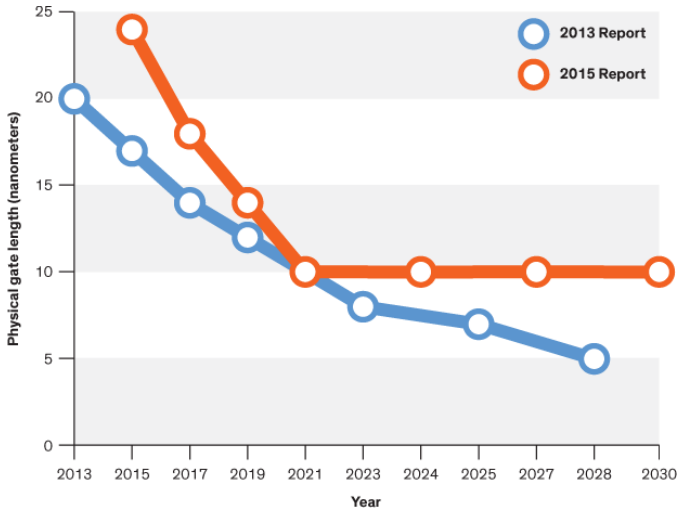

**Transistors Won't Shrink Beyond 2021, Says Final ITRS Report**

By Tiffany Trader

July 28, 2016

The final International Technology Roadmap for Semiconductors (ITRS) is now out. The highly-detailed multi-part report, collaboratively published by a group of international semiconductor experts, offers guidance on the technological challenges and opportunities for the semiconductor industry through 2030. One of the major takeaways is the insistence that Moore's law will continue for some time even though traditional transistor scaling (through smaller feature sizes) is

**Final** report of the ITRS (2015)



# Other technological developments

- Beyond truly replacing CMOS, many technological developments can **improve** on it
  - 3-D integration
  - Embedding novel memories
  - Integration with sensors…

**"More than Moore" approaches**

*These medium/long term developments will be studied in the* ***Nanoarchitecture class***

# Outline of the class

- The issue of transistor variability
- Modern CMOS architectures and their implication for design
- What comes next?
- Getting more of the devices we have: *better than Worst Case* Design

# New research direction

- If the technology is not going to make much progress…

- *can we do <u>more</u> with the technology we have?*

# The *cost* of worst case design

- Let's think about it…

# Two big strategies for « Better-than-worst-case » design

- Detect and correct errors
  - Example here: RAZOR
- Accept an approximate result

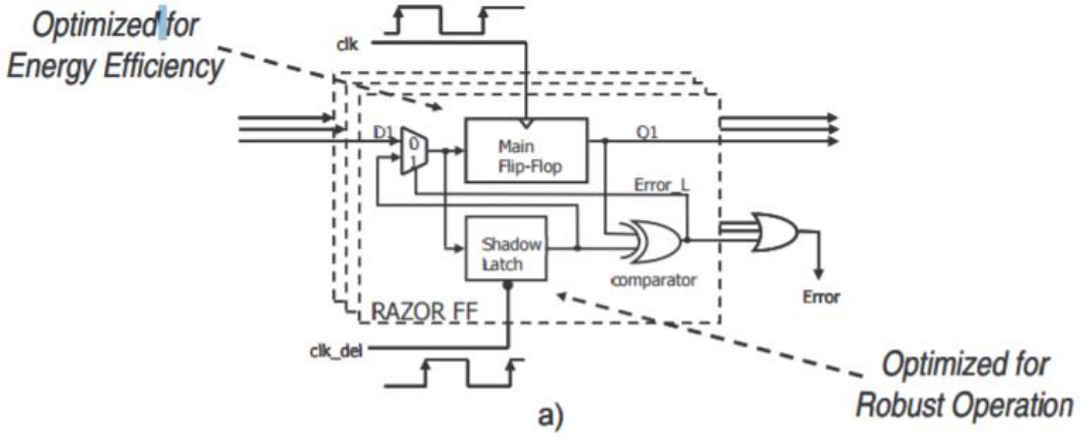
# The RAZOR system: idea

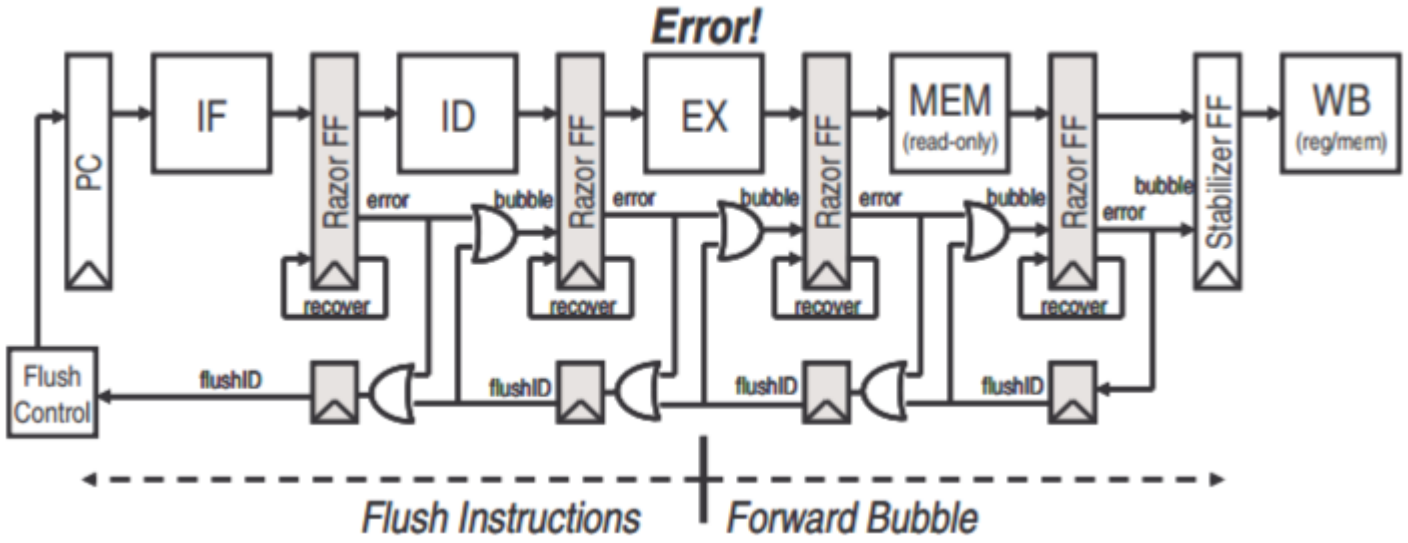
- I want to design a low power system at *e.g.* 100MHz
- *Usually*: I take the lowest supply voltage so that the circuit will always work at 100MHz in the *worst case situation*

- *Here: I take a supply voltage* so that the circuit will *typically* work at 100MHz

*And I find a way to detect if an operation did not have time to finish so that it can be flushed*

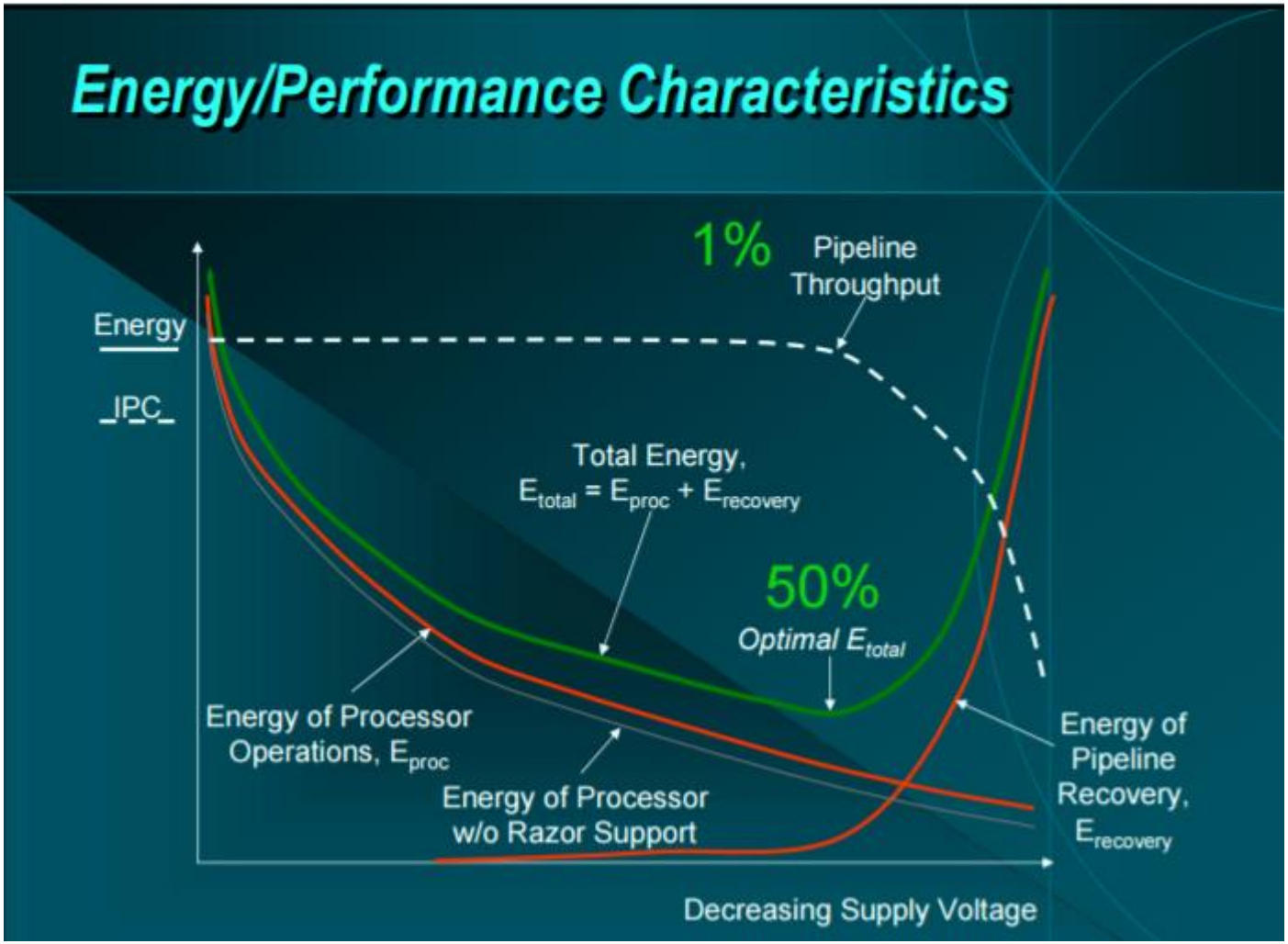
# The RAZOR system




Ernst et al ASPDAC 2005

# Results




Ernst et al ASPDAC 2005

# Optimization

- Also possible to adjust voltage in real time based on number of errors!

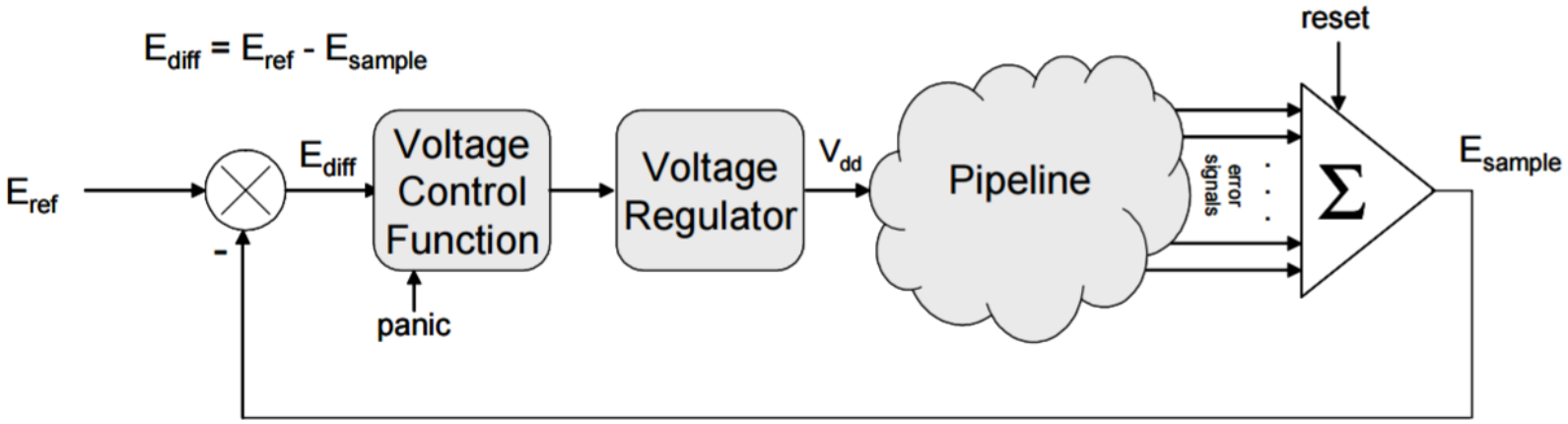


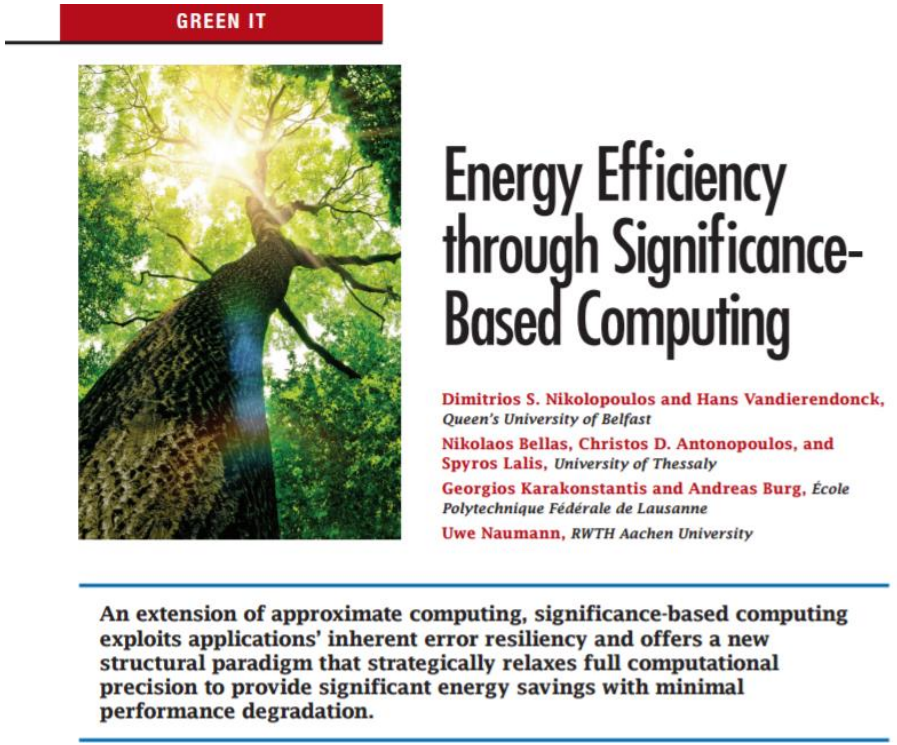

**Figure 6. Supply Voltage Control System**

# Approximate computing

- Let’s think about it…
- Do we *really* need absolutely exact result for evrything?
- The **cost** of exactness

# A current trend toward *approximate*

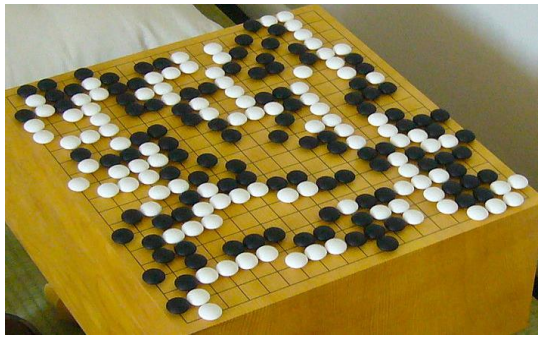

IEEE Spectrum 2015




Computer 2014

# The other motivation for going approximate computing

Lee Sedol (brain)



**20 Watt**

AlphaGo (CPU+GPU with tree seach and deep neural networks)

**>250 000 Watt**

The brain seems to have something very special about energy efficiency

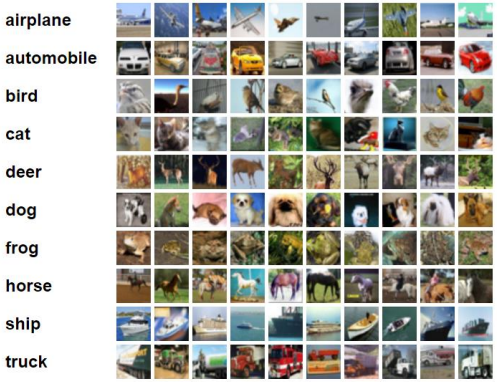
*Is approx. computing one of the keys?*

# Biggest candidates for approx. computing

Tasks that the brain does well!

- Image / Video

- Machine Learning /Artificial Intelligence

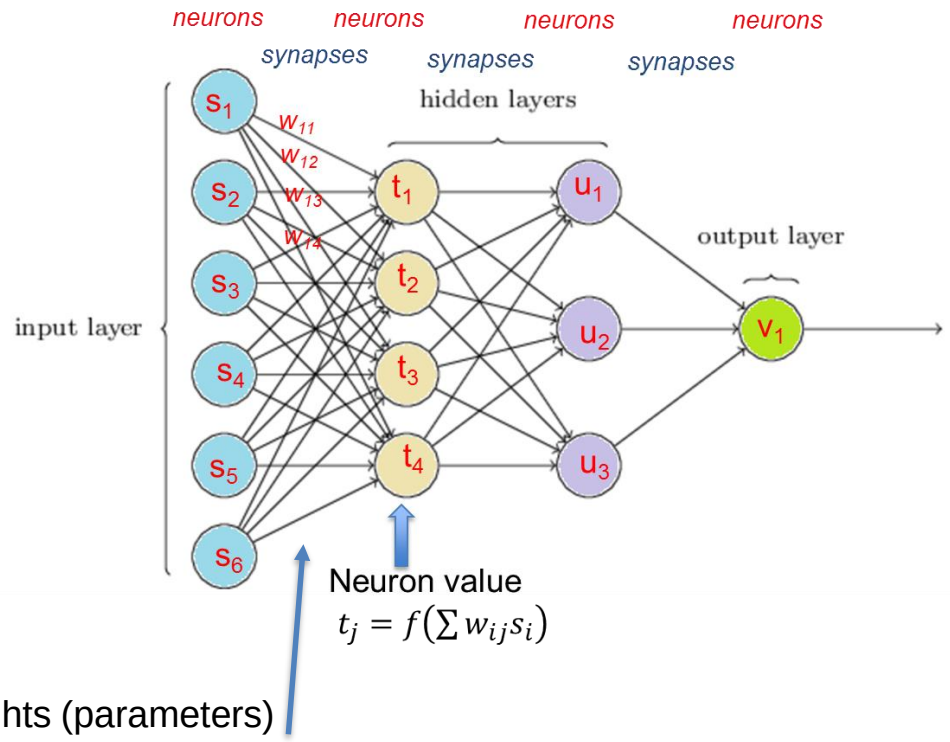


# Three ideas for approx. computing

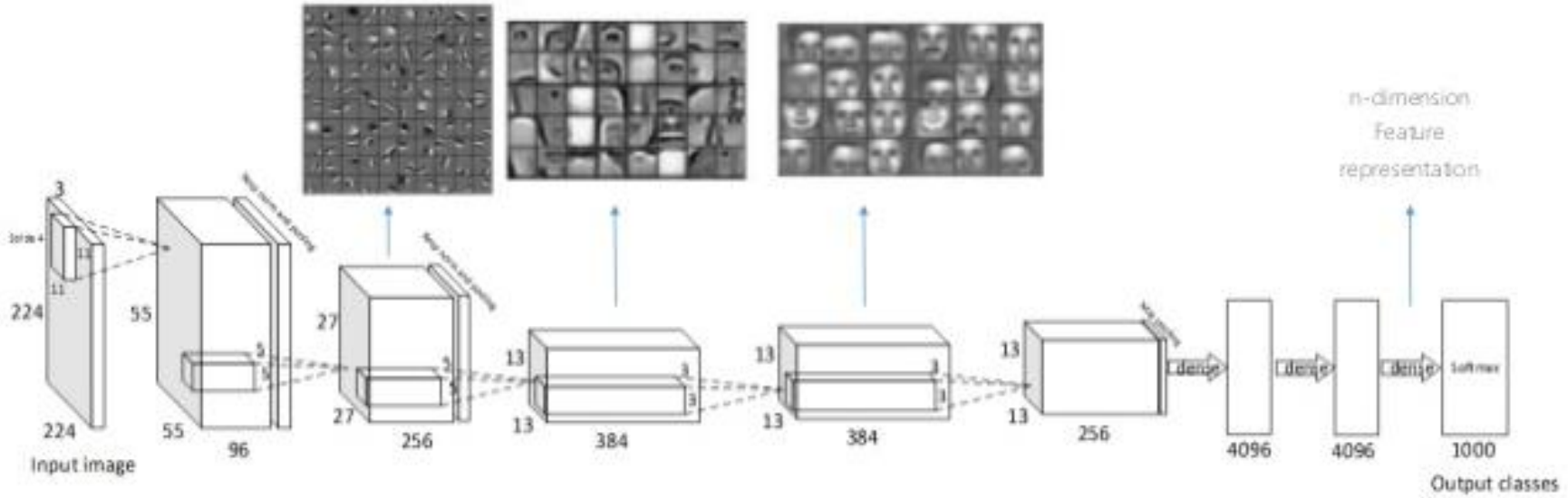
- Idea 1: Simplifying algorithm
- Idea 2: Replacing Floating Point by Fixed Point
- Idea 3: Approximate circuits

# Example: neural networks

- **Flagship algorithm of modern Artificial Intelligence**
- Takes lot of power on CPU/GPU



# Modern neural networks



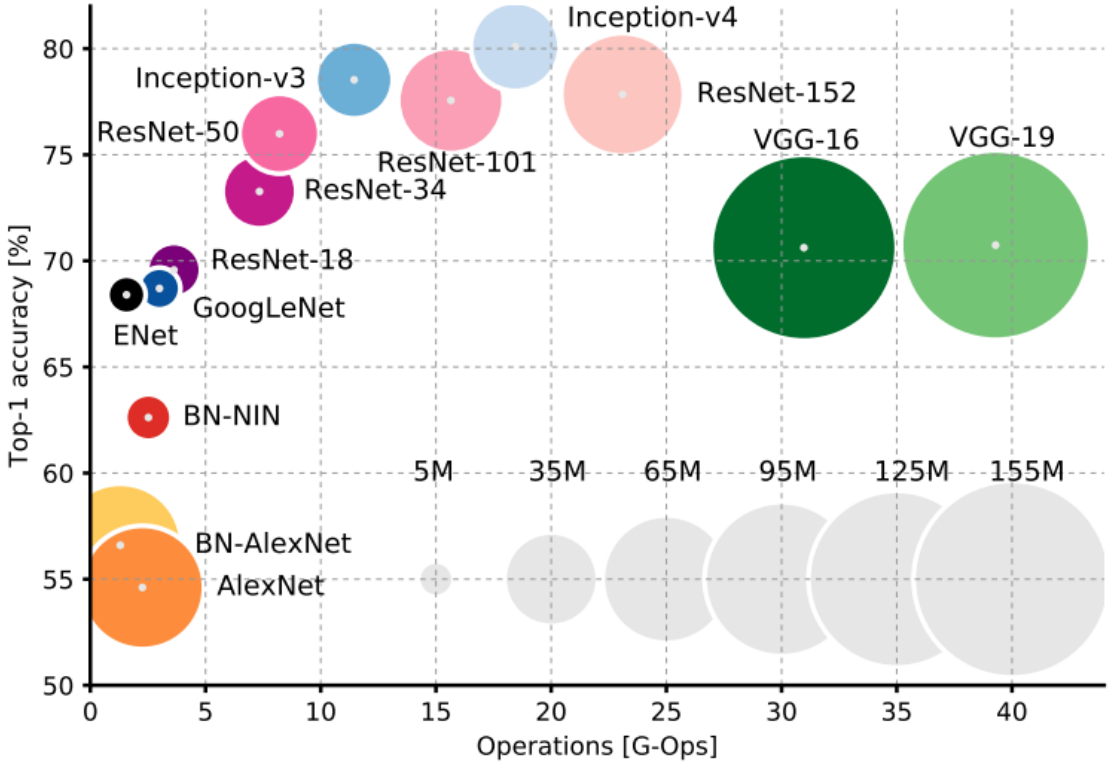

[Krizhevsky, Sutskever, Hinton'12]

Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Ng, "Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks", 11

**Many many parameters (can be hundreds of millions!!!)**

# Idea 1 simplifying algorithm

## How complicated should the neural network be?

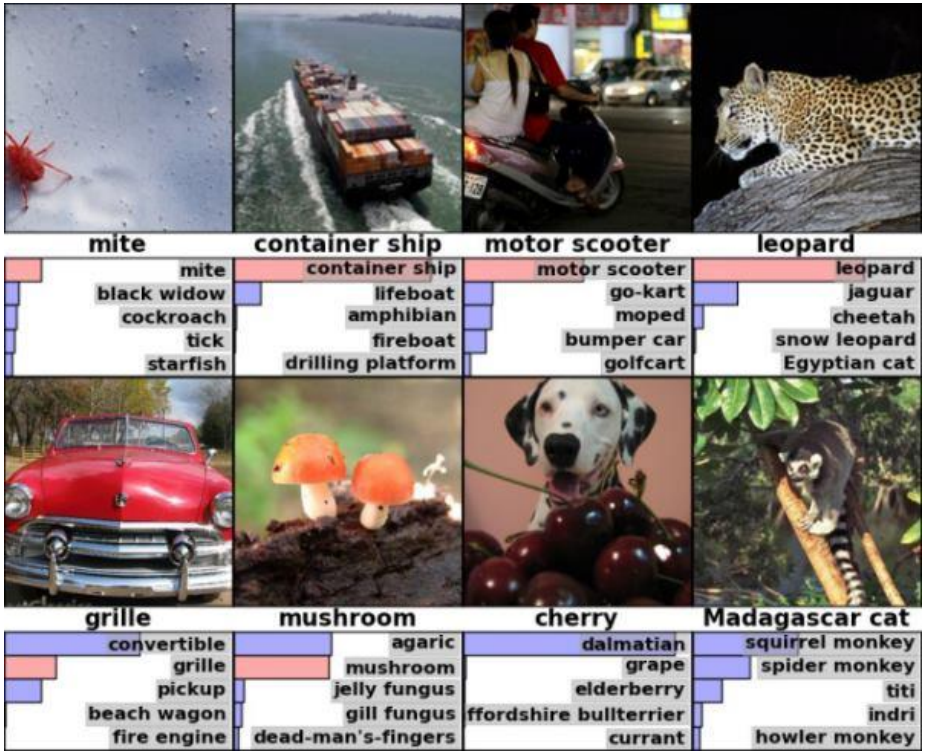



AN ANALYSIS OF DEEP NEURAL NETWORK MODELS FOR PRACTICAL APPLICATIONS

**Alfredo Canziani & Eugenio Culurciello**
Weldon School of Biomedical Engineering
Purdue University
{canziani,euge}@purdue.edu

**Adam Paszke**
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw
a.paszke@students.mimuw.edu.pl

# Example of AlexNet




PREDICTIONS ON IMAGENET. FROM "*KRIZHEVSKY A., SUTSKEVER I., HINTON. G.E. (2012) IMAGENET CLASSIFICATION WITH DEEP CONVOLUTIONAL NEURAL NETWORKS*".

Really not so bad!

# Idea 2 moving from Floating Point to Low Precison Fixed Point

- Neural networks usally run in 32 bits Floating Points

- Is that really necessary?

# Neural networks with low precison

CIFAR-10

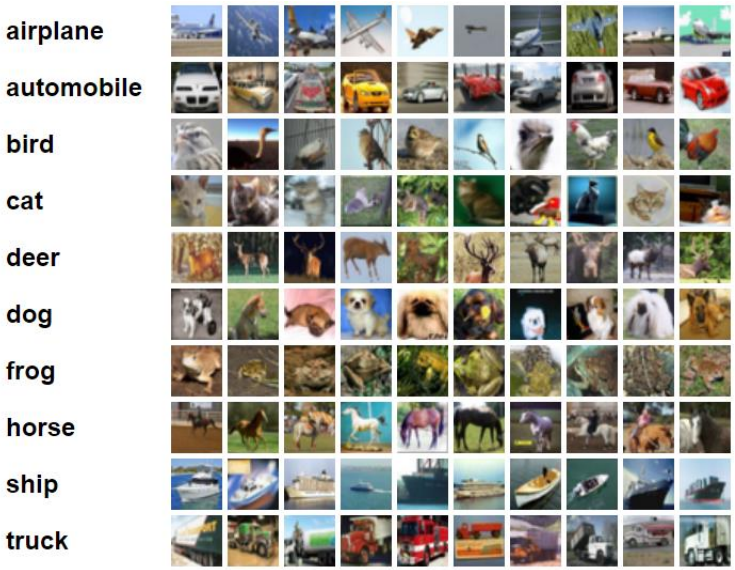



*Table 7.* CIFAR-10 classification error rate with different bit-width combinations

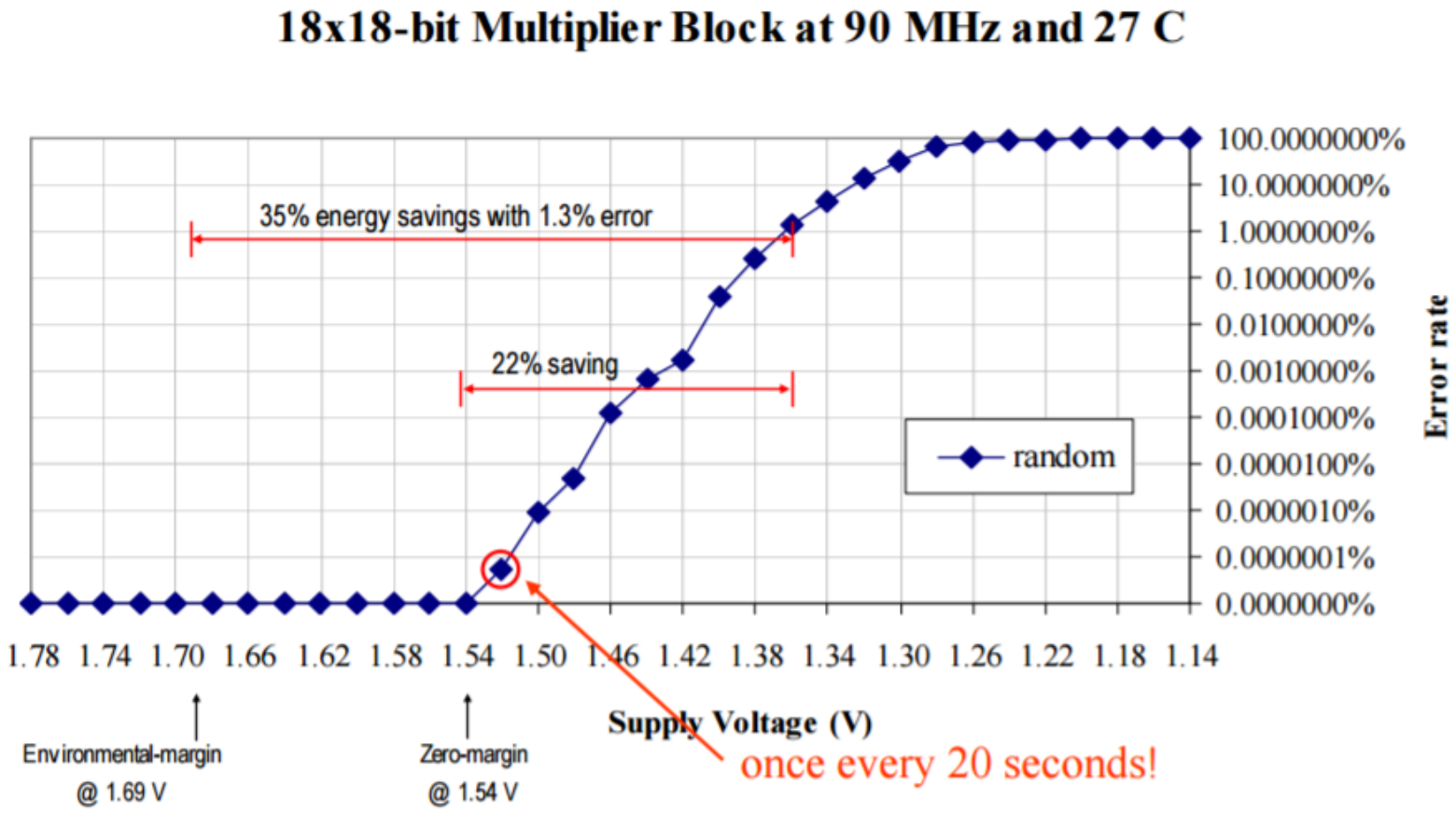
| Activation Bit-width | Weight Bit-width 4 | 8 | 16 | Float |
|---|---|---|---|---|
| 4 | 8.30 | 7.50 | 7.40 | 7.44 |
| 8 | 7.58 | 6.95 | 6.95 | 6.78 |
| 16 | 7.58 | 6.82 | 6.92 | 6.83 |
| Float | 7.62 | 6.94 | 6.96 | 6.98 |

Lin et al, ICML 2016

Machine learning applications are especially adapted to approximate computing

# Idea 3 Approximate circuits? Accepting incorrect least significant bits

Let's look at this graph…




Class activity : how to harness this?



Ernst et al ASPDAC 2005

# Need for associated ecosystem

## EnerJ: Approximate Data Types for Safe and General Low-Power Computation

Adrian Sampson Werner Dietl Emily Fortuna Danushen Gnanapragasam
Luis Ceze Dan Grossman

University of Washington, Department of Computer Science & Engineering
http://sampa.cs.washington.edu/

**Abstract**

Energy is increasingly a first-order concern in computer systems. Exploiting energy-accuracy trade-offs is an attractive choice in applications that can tolerate inaccuracies. Recent work has explored exposing this trade-off in programming models. A key challenge, though, is how to *isolate parts of the program that must be precise from those that can be approximated* so that a program functions correctly even as quality of service degrades.

We propose using type qualifiers to declare data that may be subject to approximate computation. Using these types, the system automatically maps approximate variables to low-power storage, uses low-power operations, and even applies more energy-efficient algorithms provided by the programmer. In addition, the system can statically guarantee isolation of the precise program component from the approximate component. This allows a programmer to control explicitly how information flows from approximate data to precise data. Importantly, employing static analysis eliminates

in data-centers. More fundamentally, current trends point toward a "utilization wall," in which the amount of active die area is limited by how much power can be fed to a chip.

Much of the focus in reducing energy consumption has been on low-power architectures, performance/power trade-offs, and resource management. While those techniques are effective and can be applied without software knowledge, exposing energy considerations at the programming language level can enable a whole new set of energy optimizations. This work is a step in that direction.

Recent research has begun to explore energy-accuracy trade-offs in general-purpose programs. A key observation is that systems spend a significant amount of energy guaranteeing correctness. Consequently, a system can save energy by exposing faults to the application. Many studies have shown that a variety of applications are resilient to hardware and software errors during execution [1, 8, 9, 19, 21–23, 25, 31, 35]. Importantly, these studies universally show that applications have portions that are more resilient and

Inventing a Computer Science where results are not always exact

# Industrial challenge of Better than Worst Case

- Class activity: how would industry feel about better than worst case design?

# Conclusions

- CMOS scaling is struggling
- New transistor architectures have been found and offer design opportunities
- It is unclear if even newer will emerge
- Significant improvements can be achieved with current technologies if we give up worst case design, but this raises considerabel challenges