Advanced Process Technologies

Prof. Adam Teman

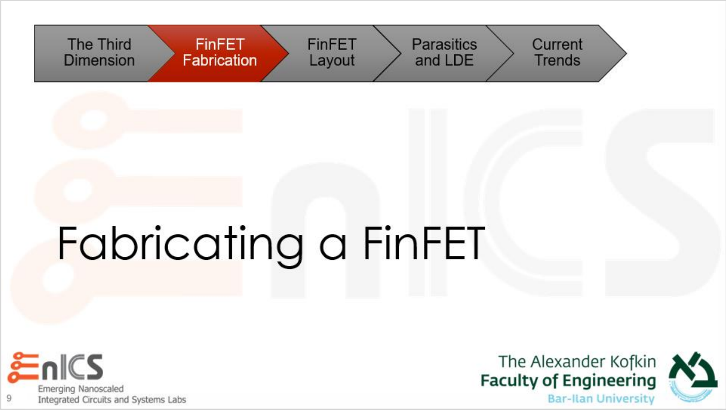
16 March 2022

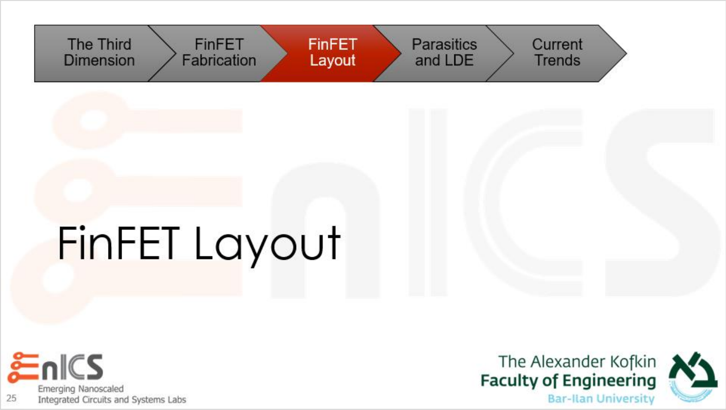
Special credit to **Alvin Loke** and **Or Nahum** for the material and wonderful explanations!

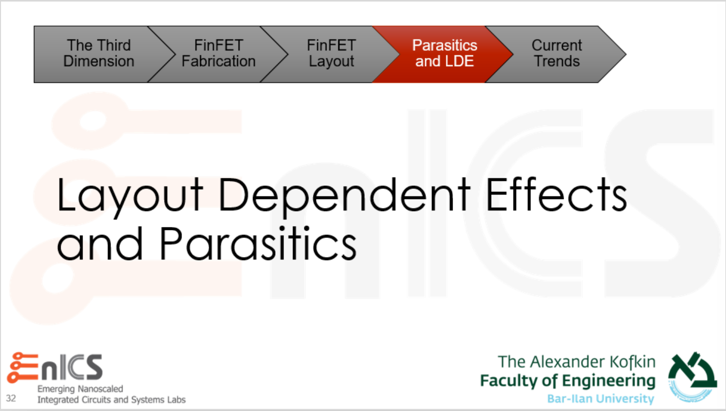
EnICS
Emerging Nanoscaled Integrated Circuits and Systems Labs

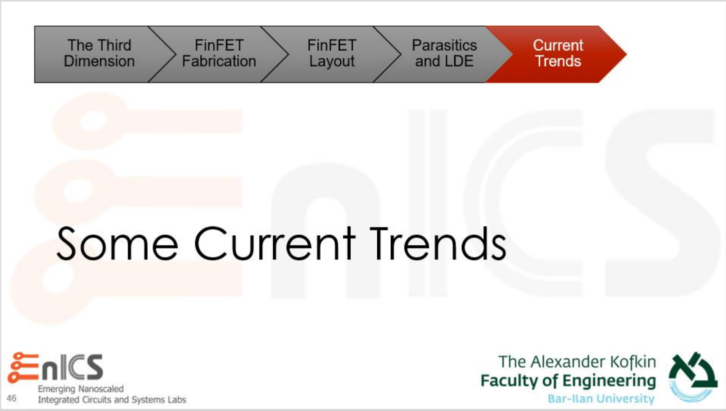
The Alexander Kofkin
Faculty of Engineering
Bar-Ilan University

Lecture Overview





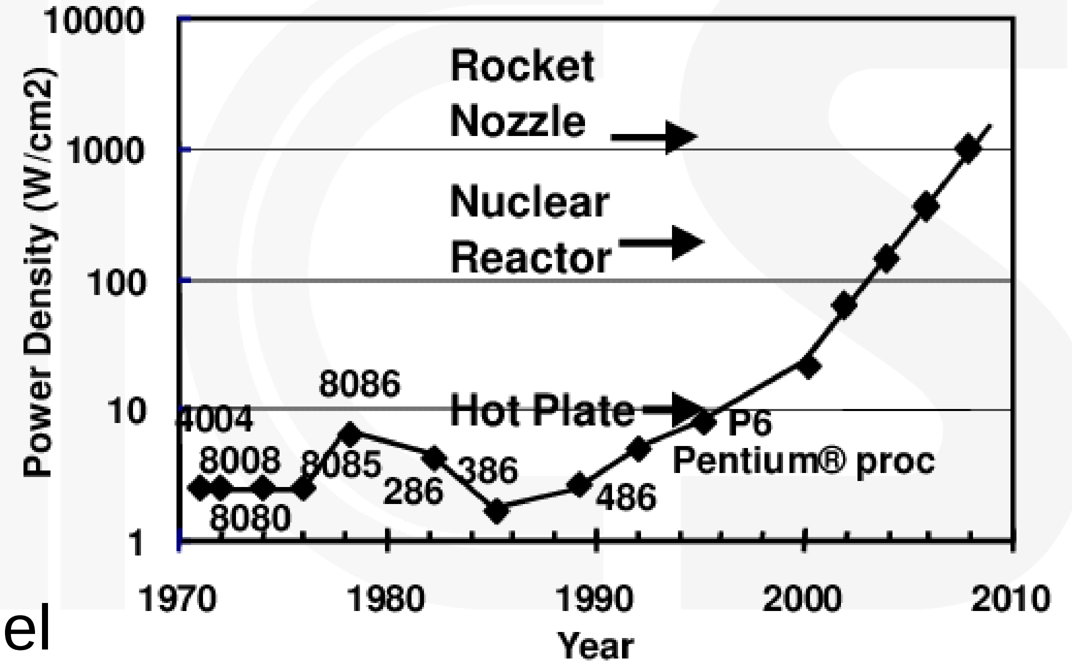




Moving to the third dimension

The breakdown of Dennard's Law

- **Robert Dennard (1974) observed that continuous scaling is enabled by:**
  - Scaling the transistor dimensions by 30%
  - Scaling the supply voltage by 30%
- **This results in:**
  - 2X area reduction
  - 30%-40% increase in frequency
  - Constant power density
- **However:**
  - Voltage wasn't scaled according to the model
  - Frequency was raised faster than the model
  - Leakage, due to VT scaling, wasn't taken into account
- **This led to the infamous power crisis**



Source: S. Borkar (Intel)

The problem with voltage scaling

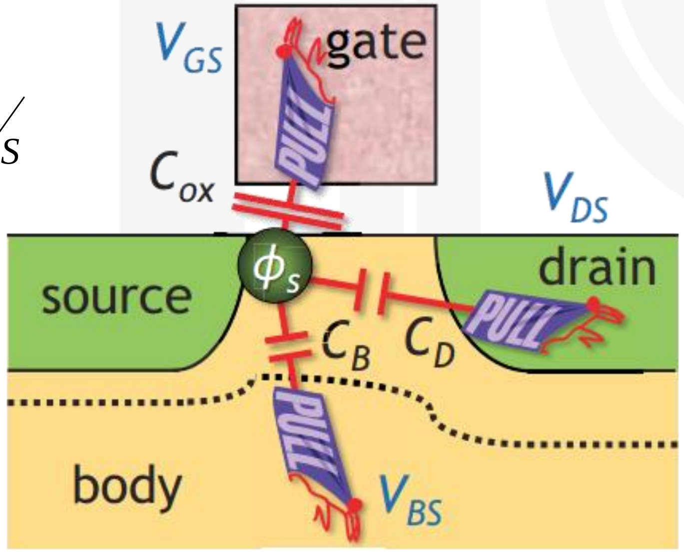
- In order to keep **constant power density**, the **supply voltage** needs to be scaled
- However, this also requires scaling of the **threshold voltage**
- But the **off current** of a transistor is **exponentially dependent on VT**

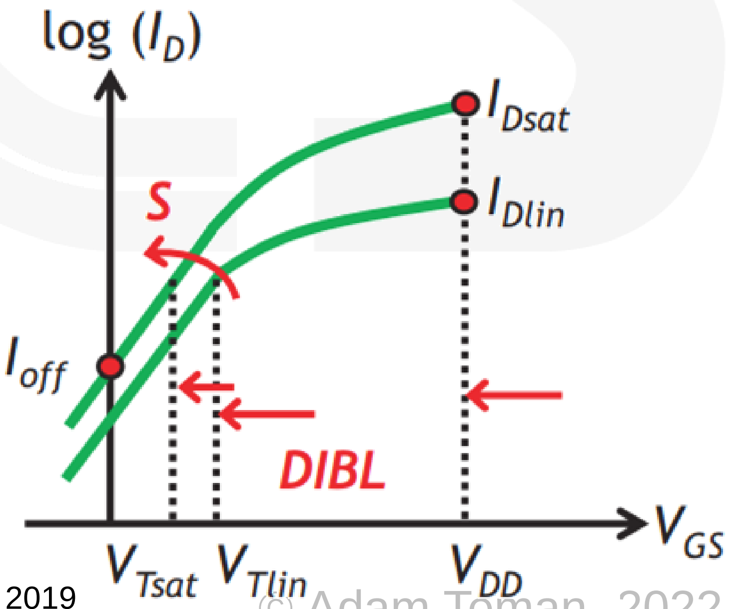
$$PD = \frac{f \cdot C \cdot V_{\mathrm{DD}}^2}{A} = \frac{S \cdot S^{-1} \cdot S^2}{S^2}$$

$$I_{\mathrm{sat}} = \mu_{\mathrm{n}} C_{\mathrm{ox}} \frac{W}{L} \left(V_{\mathrm{DD}} - V_{\mathrm{T}}\right)^2$$

$$I_{\mathrm{off}}\left[\mathrm{nA}\right] = 100 \frac{W}{L} e^{-V_{\mathrm{T}} / n\phi_T} = 100 \frac{W}{L} 10^{-V_{\mathrm{T}} / S}$$

- This is limited by both **subthreshold swing** and **DIBL**.

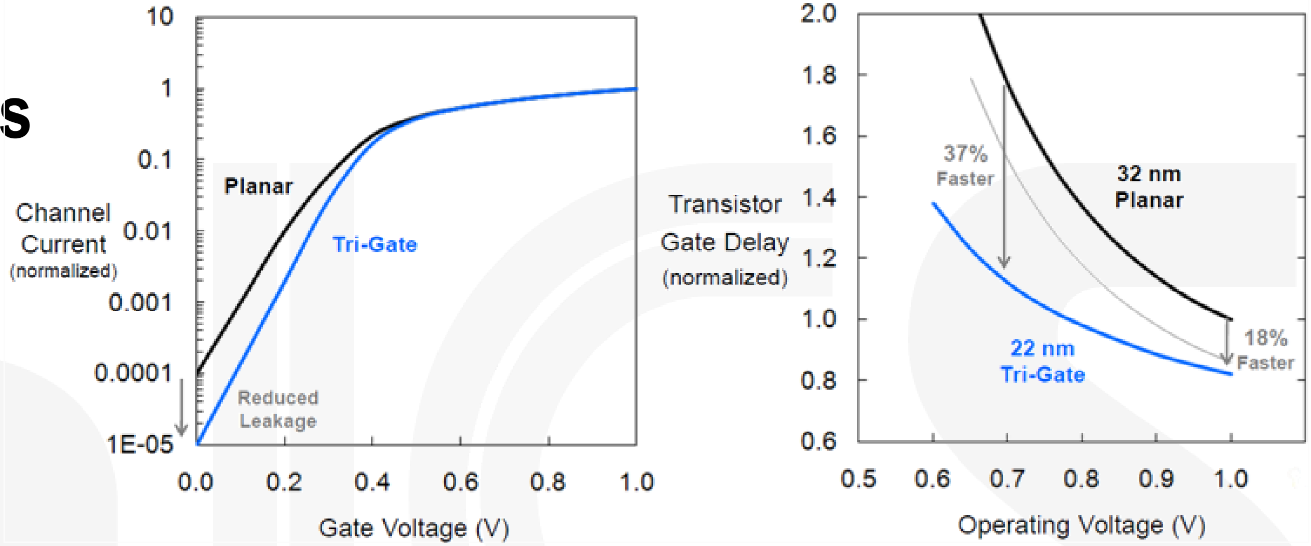




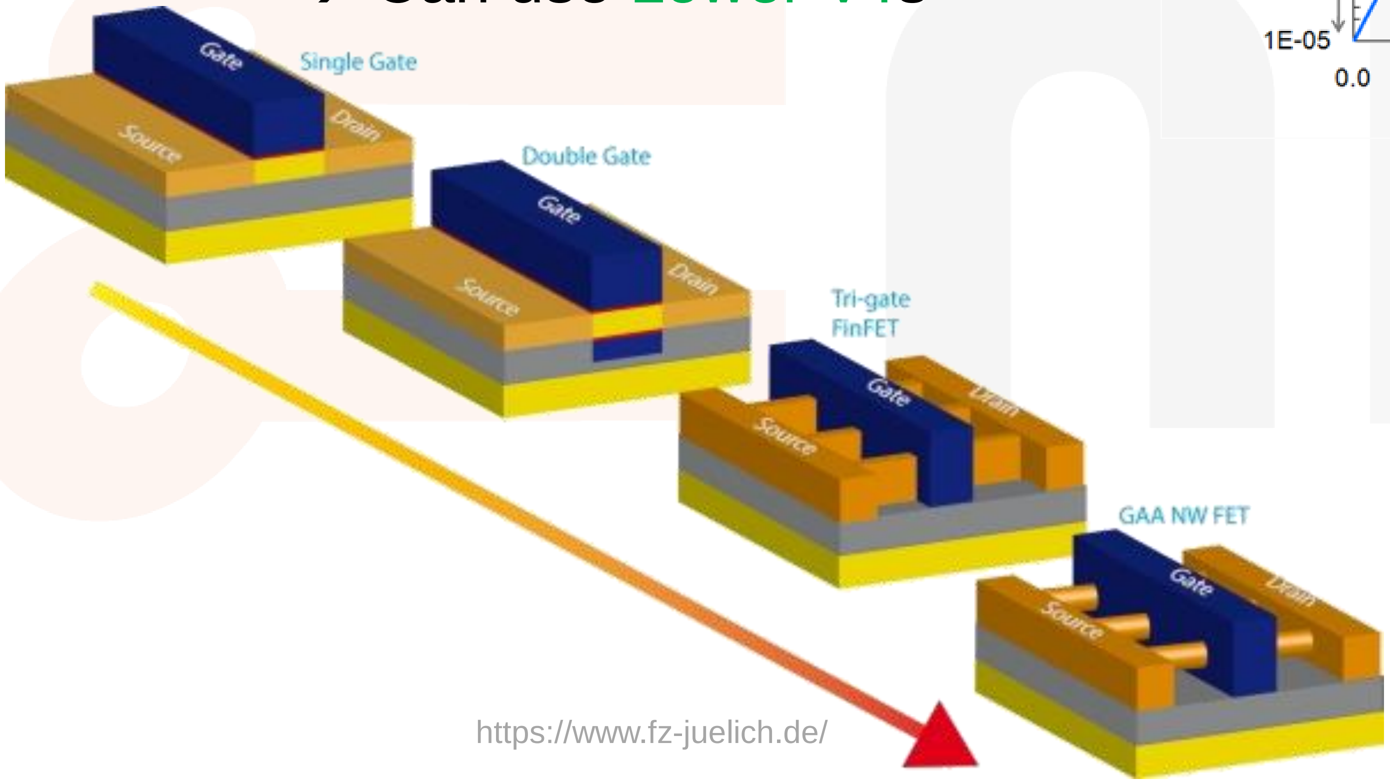
Loke, CICC 2019

The Multi-Gate Solution

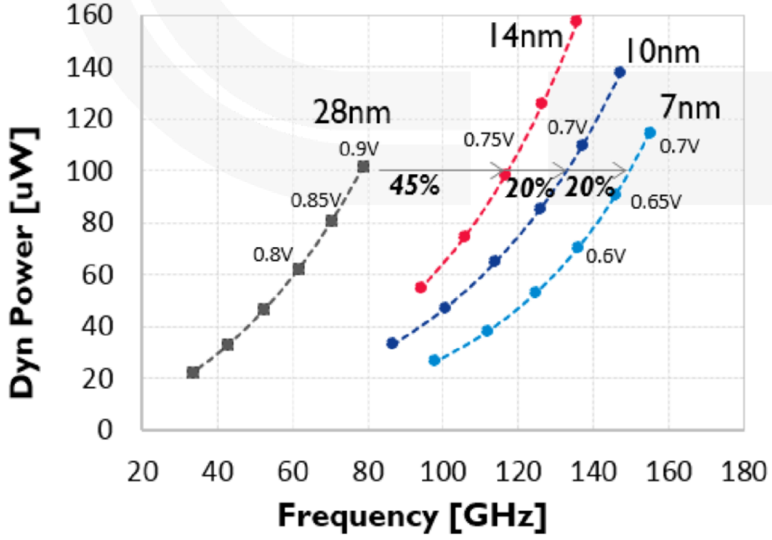
- **Moving to the third dimension enables better channel control**
  - → Smaller subthreshold slope
  - → Can use Lower VTs



https://www.realworldtech.com/intel-22nm-finfet/

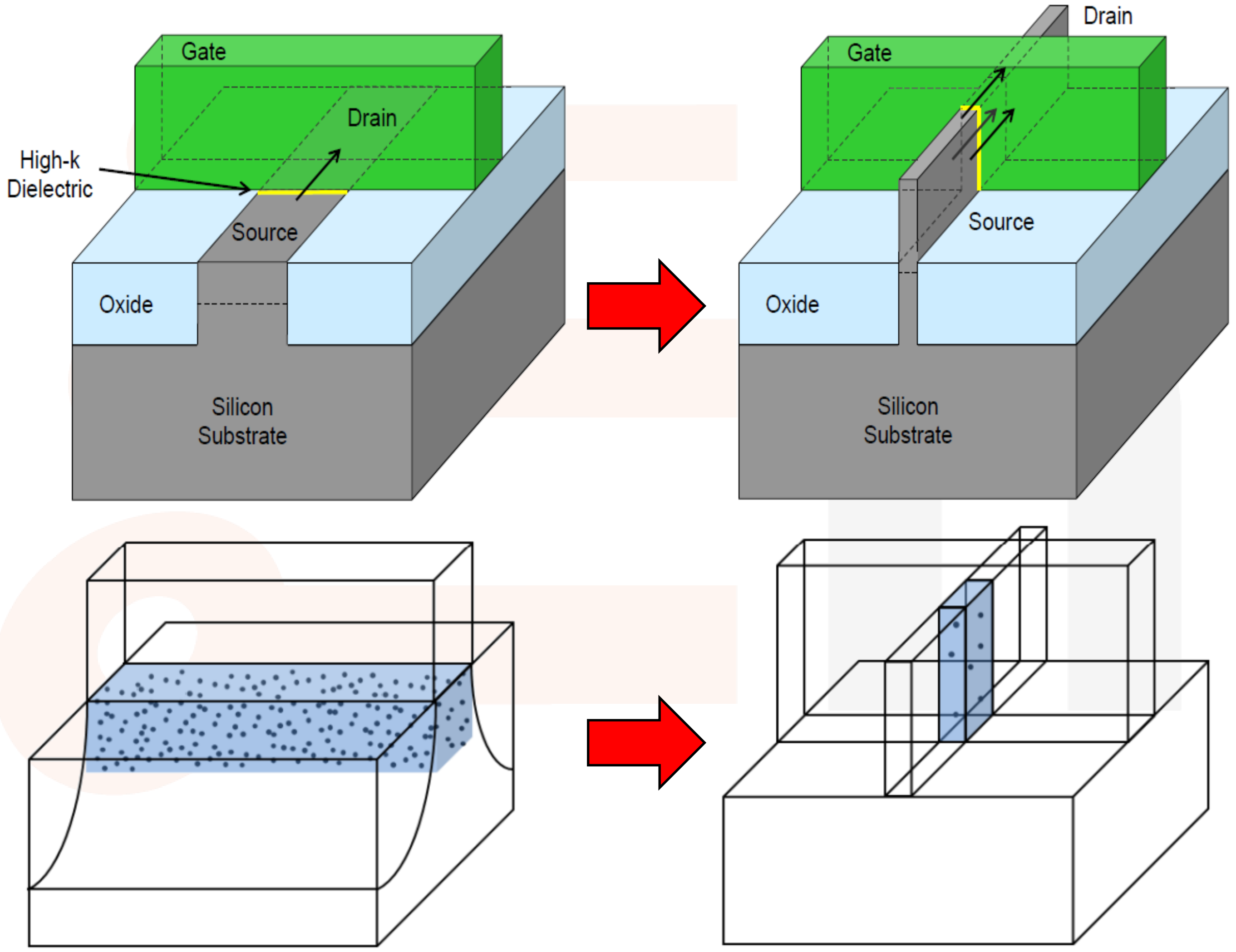


https://www.fz-juelich.de/



M. G. Bardon (IMEC) *ICICDT* (2015)

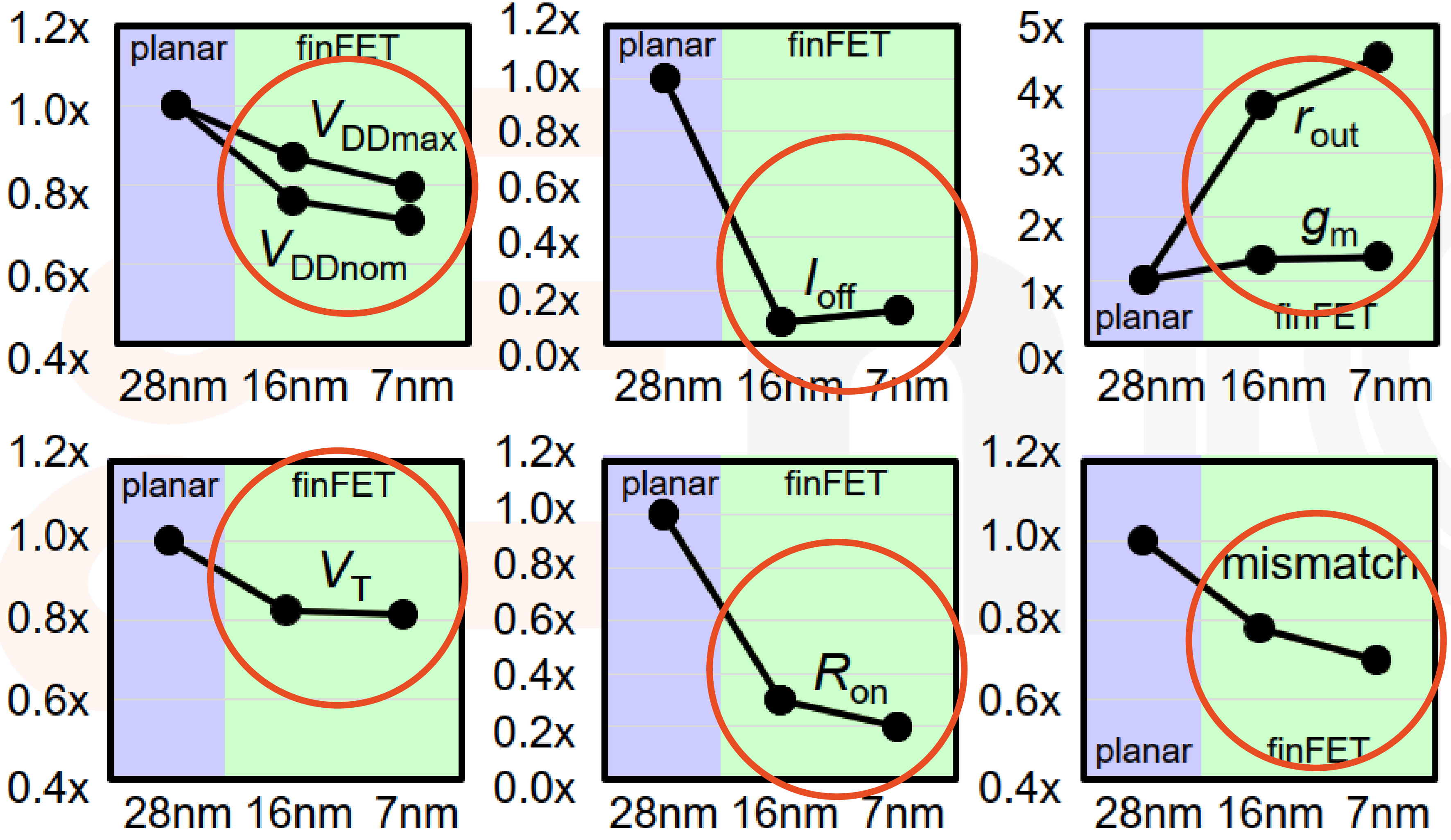
Introducing the FinFET



Planar transistor **FinFET transistor**

- **3D Structure**
  - More I_{on} & gm per area
- **Fully depleted channel**
  - Less DIBL
  - Less RDF mismatch
  - Negligible body effect
- **Quantized channel width**
- **Problematic Parasitics**
  - High S/D resistance
  - High S/D coupling to gate

Main Parameters



Loh, Mediatek
Loke, BCICTS

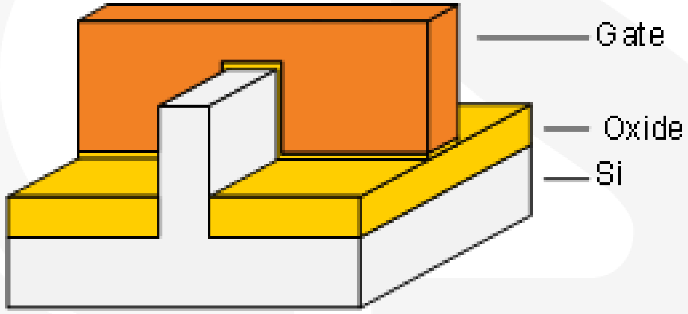
- **Lower power**
- **Lower leakage**
- **Higher intrinsic gain**
- **Better switch**
- **Better mismatch**
- **Smaller area**

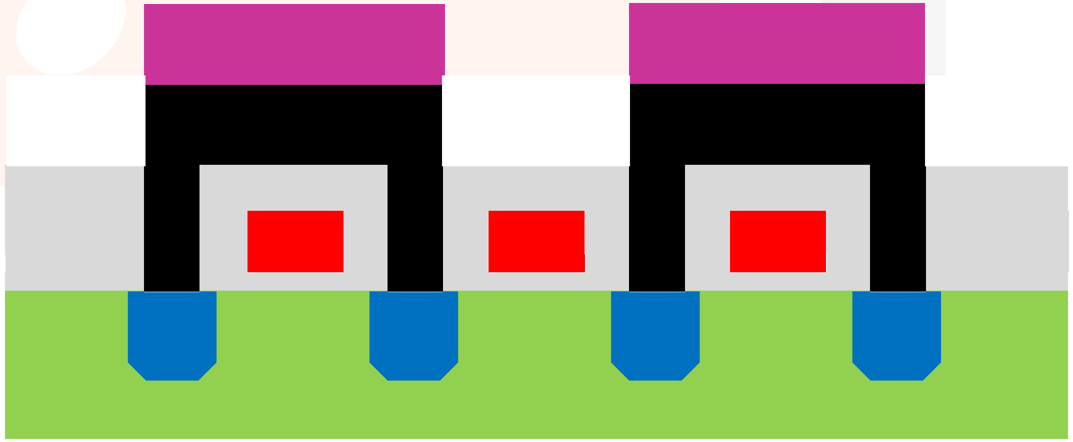
- **Summary:**
  Improved PPA

Fabricating a FinFET

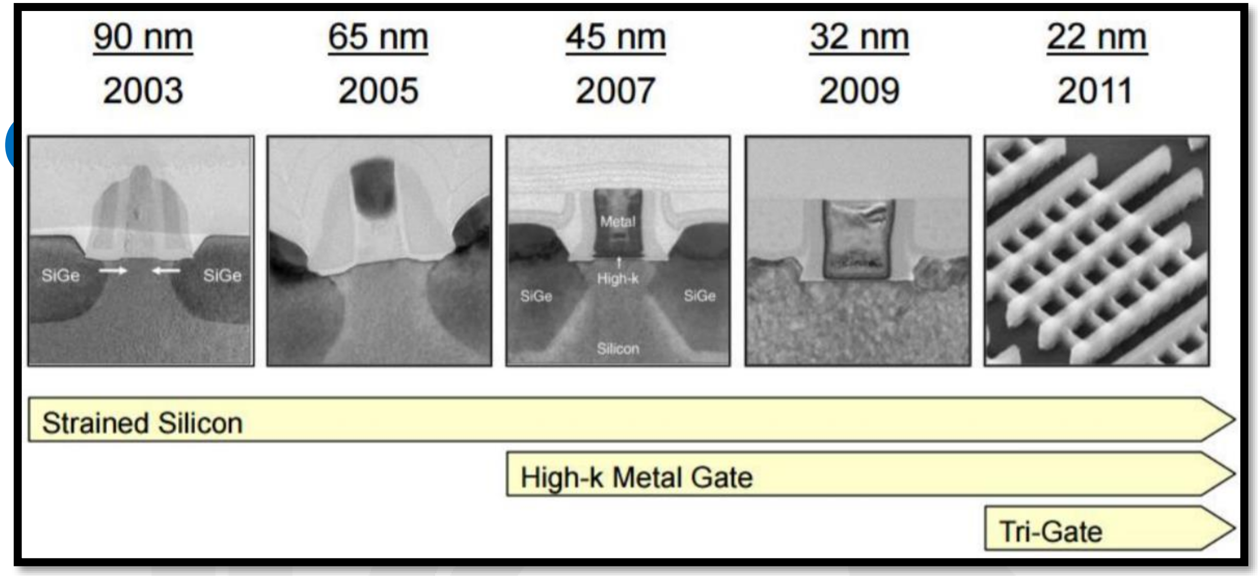
Planar vs. FinFET Fabrication

- **We'll start by remembering how planar CMOS is constructed.**
  - We patterned the poly gate, oxide and diffusions using a self aligned process
  - Now the backend layers can be made by **litho**, **etch**, **clean**, **deposit**, **polish**
- **But how do we make Fins *above* the substrate plane?**
  - We *carve* them out of the substrate!
  - And then cover them with oxide and poly.

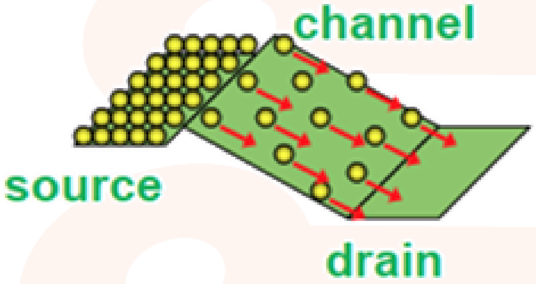




The roads to higher p

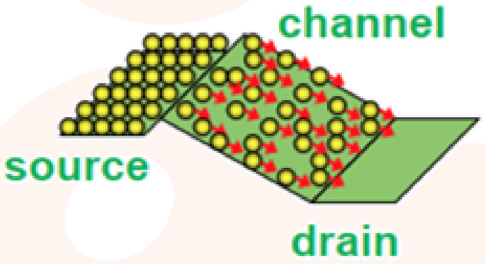


Increase μ – move carriers faster



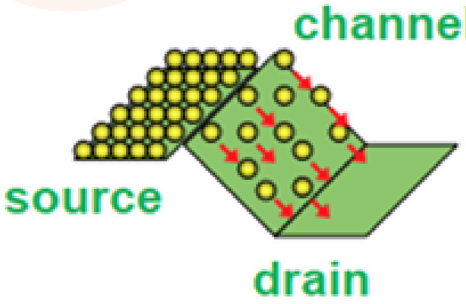
strain engineering

Increase C_{ox} – move more carriers



high-*K* dielectric
metal gate

Decrease *L* – steepen the hill

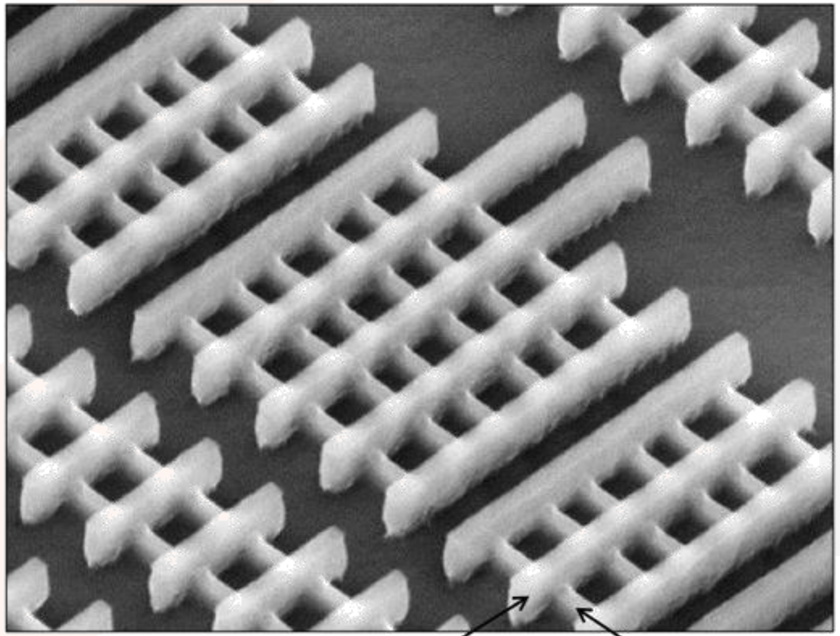


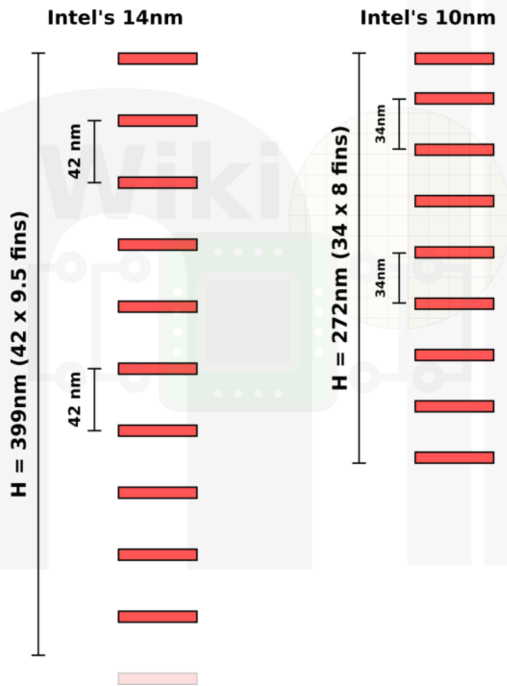
lithography
scaling

Technology Innovation	Foundry Debut	Reason Required
Mechanical stressors	40nm	Mobility boost for more FET drive & higher I_{on}/I_{off}
HKMG replacement gate	28nm (HK-first) 20nm (HK-last)	Higher C_{ox} for more FET drive & channel control
Multiple-patterning	20nm	Sub-80nm pitch lithography without EUV (13.5nm λ = 193 nm)
Complex middle-end-of-line	20nm	Contact FET diffusion & gate with tighter CPP

Fin and Gate Patterning

- **The first critical step of the FinFET process is to make very thin fins at a very tight pitch and build tightly pitched gates.**





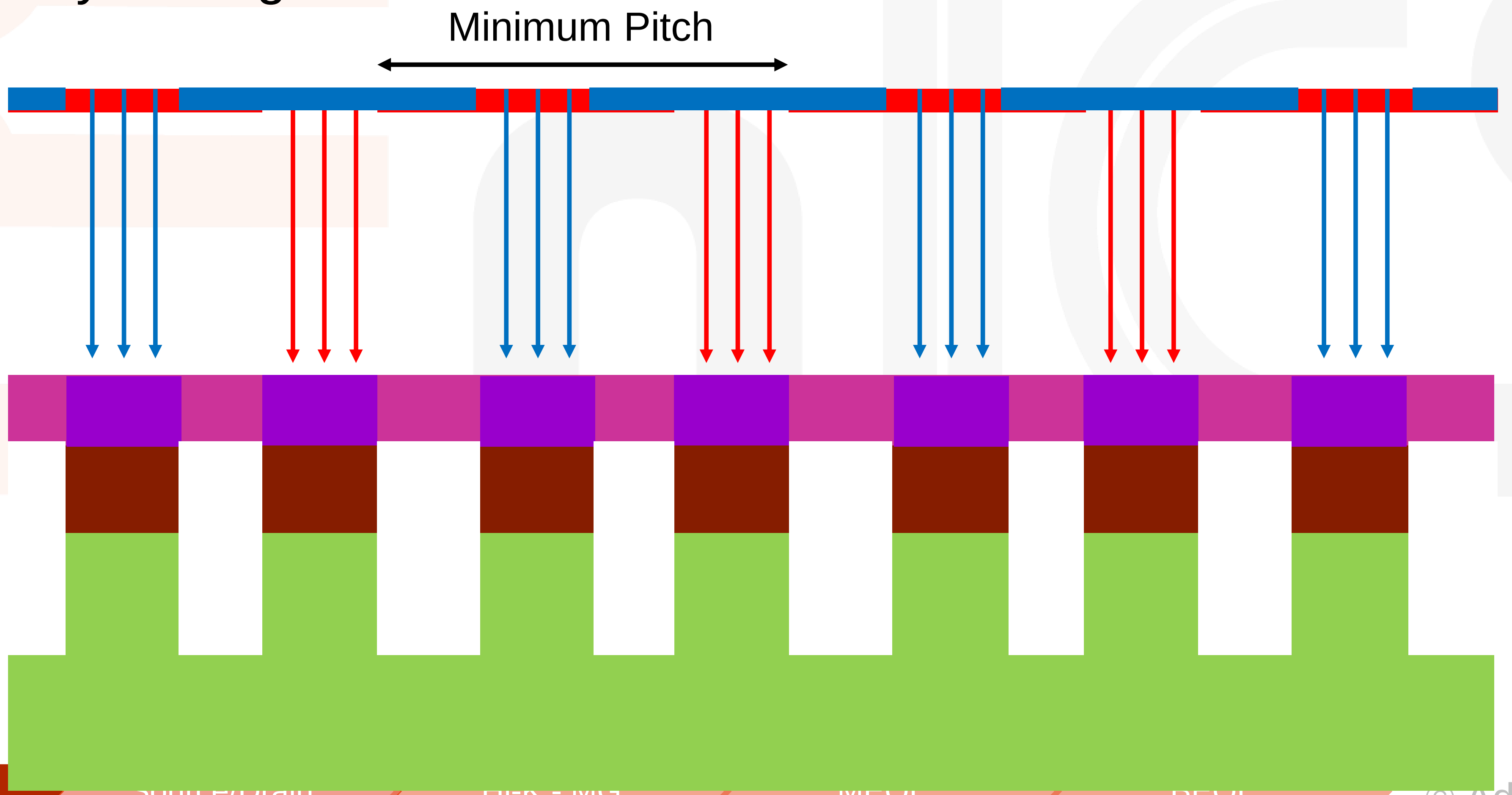
- Immersion Litho
- Phase shifting
- Double patterning (a.k.a. LE-LE)
- Self-aligned Patterning (SADP/SAQP)

- **How can we do this with 193nm litho?**

$$CD = \frac{k \cdot \lambda}{n \sin \alpha}$$

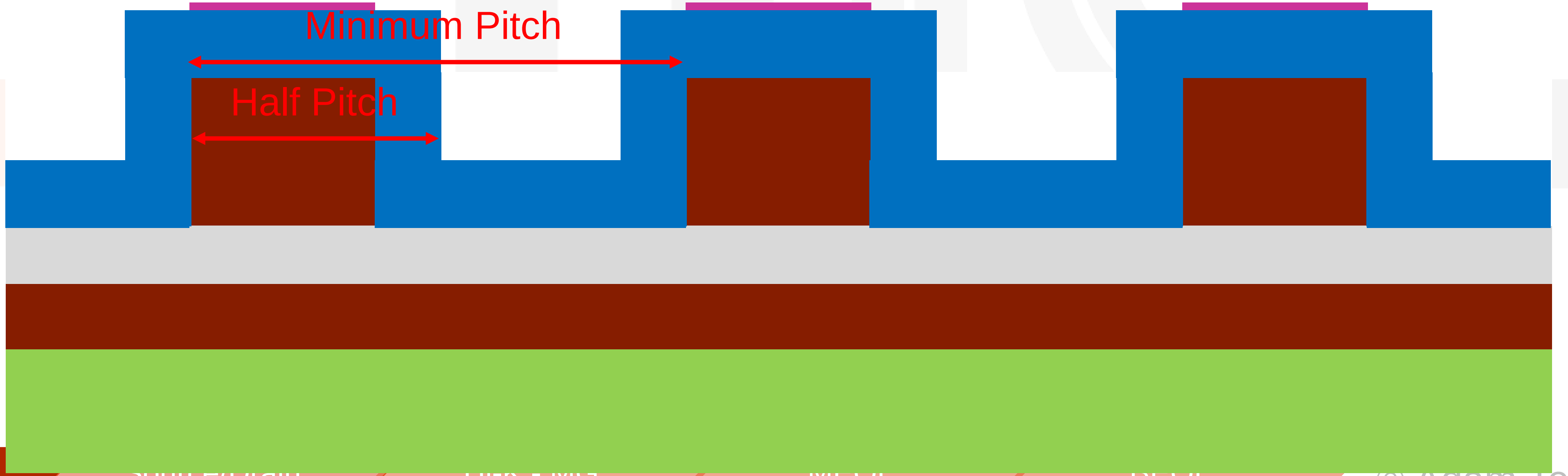
Double Patterning (LE-LE)

- **Introduce "colored" masks for pitch splitting**
  - Limited by misalignment between masks.



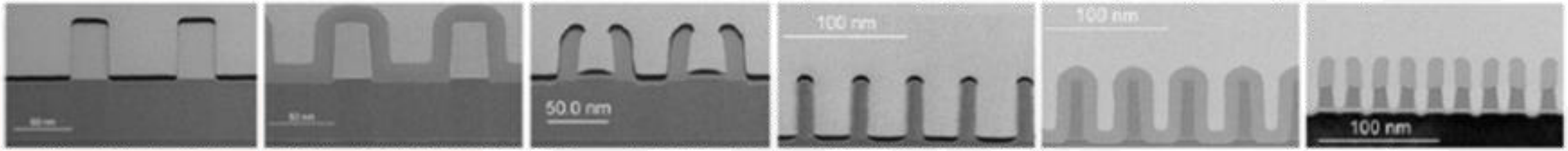
Self-Aligned Double Patterning (SADP)

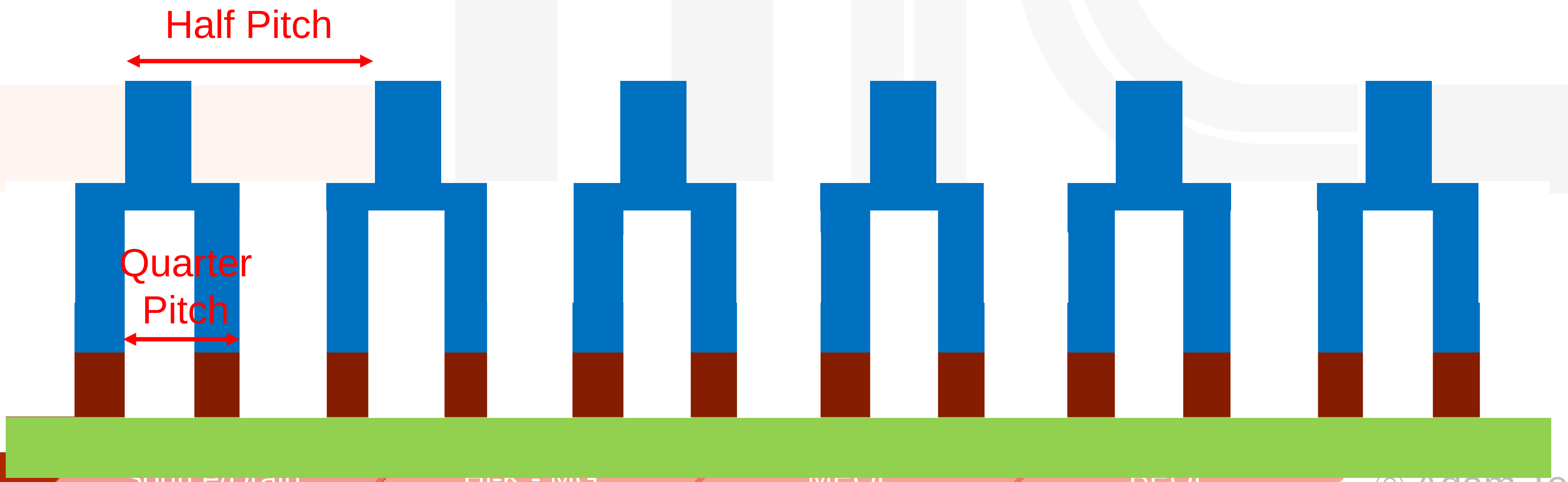
- **We can achieve the same resolution using a *self-aligned* technique**
  - Start by constructing "mandrels" with the minimum litho pitch.
  - Next, create spacers around the mandrels.
  - Remove the hard mask and the spacers are at half the pitch.
  - *Note that Line Edge Roughness is correlated, reducing L variation!*



Self-Aligned Quadruple Patterning (SADP)

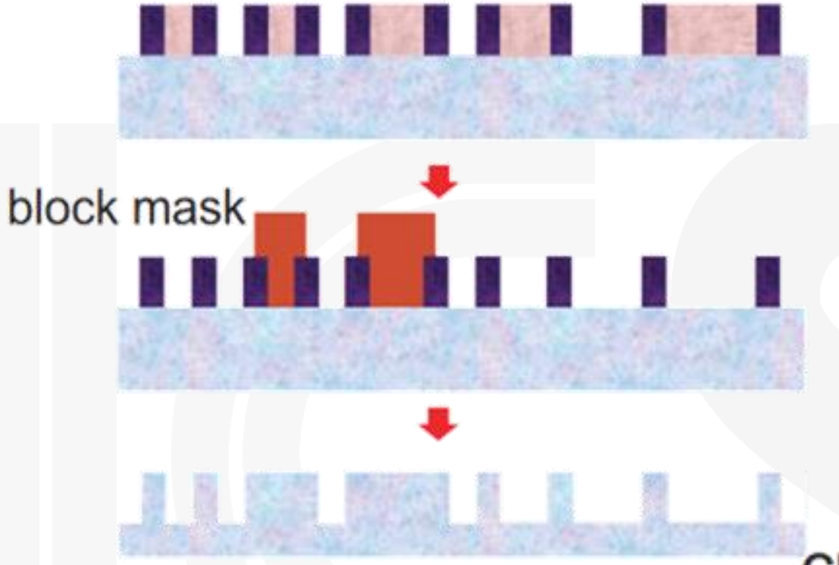
- **And just repeat the process to double the resolution again.**
  - Just remember that this cost us another few masks and many steps.



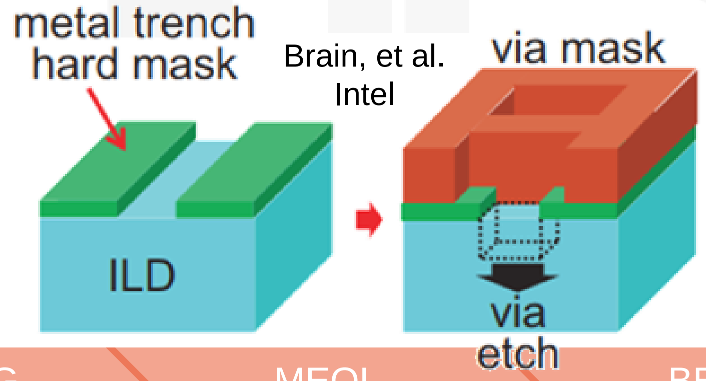


Additional Advanced Litho Points

- **You can only print one line width!**
  - To get multiple line widths, change mandrel width and spacing.
- **Each multi-patterned layer is unidirectional.**
  - No more wrong-way routes or jogs!
- **Use orthogonal cut masks to break patterns.**
  - Orthogonal cut mask according to CD.
  - Eliminate corner rounding on fins/gates.
- **Self-aligned vias**
  - Via etch only at intersection of trench & via masks

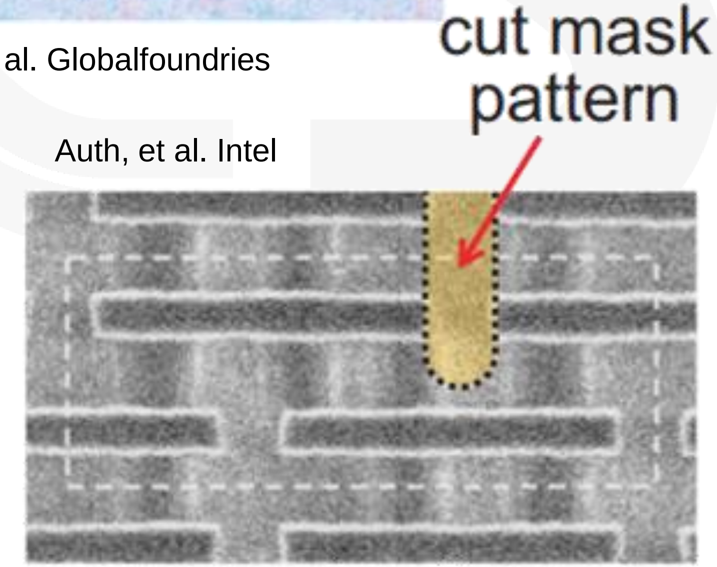


Woo, et al. Globalfoundries



Brain, et al. Intel

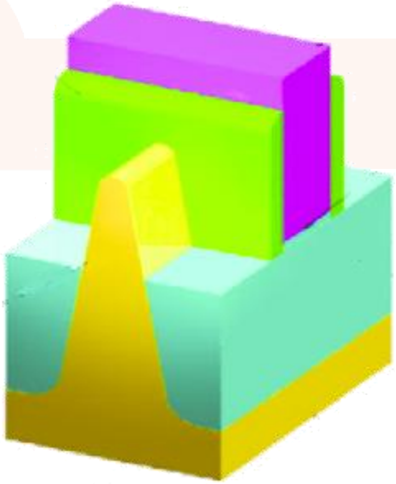
Auth, et al. Intel



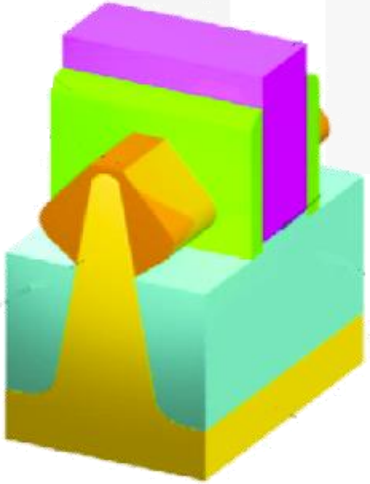
Source/Drain Regions

- **We have seen how to make the "array" of fins and gates.**
- **But we have a few problems with the S/D regions:**
  1. They are really small and delicate → hard to contact to.
  2. They have high resistance.
  3. We need to apply stressors to improve mobility.
- **Therefore, we build an epitaxial area on the fins**

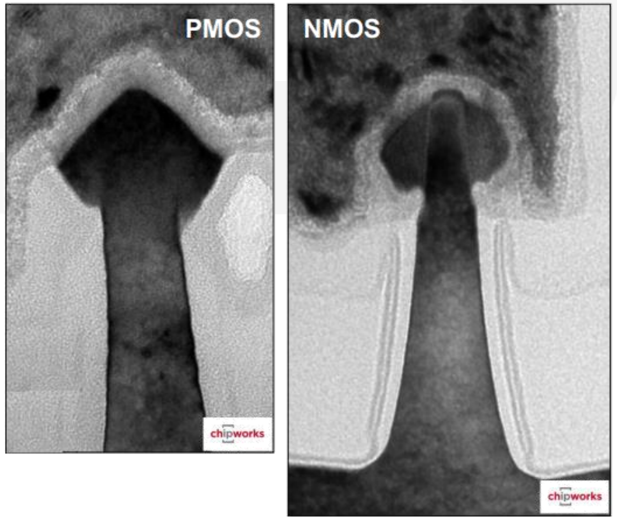
Start by building a spacer between the gate and diffusion



Then grow an epi layer of a stressor on the exposed fin

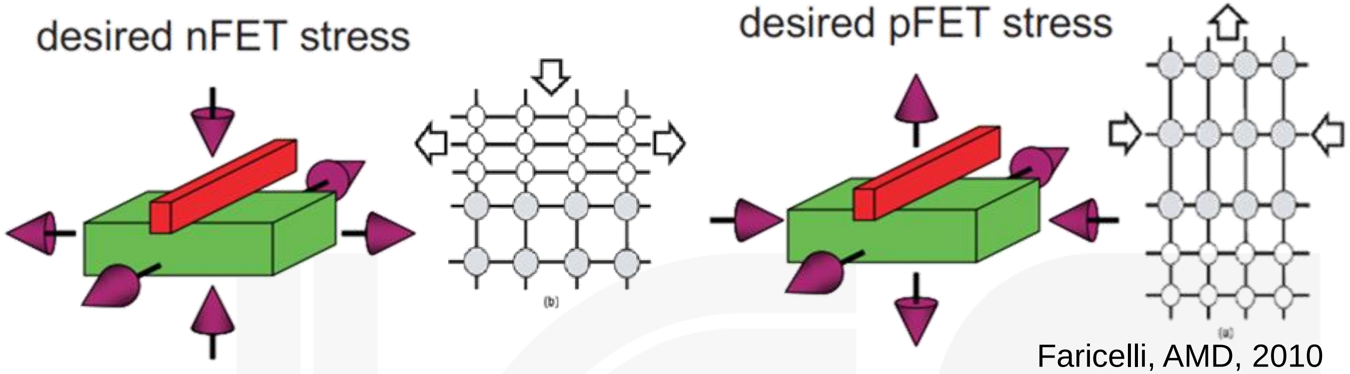


The result is a diamond shape. Note that a PMOS diamond is much "cleaner" than an NMOS



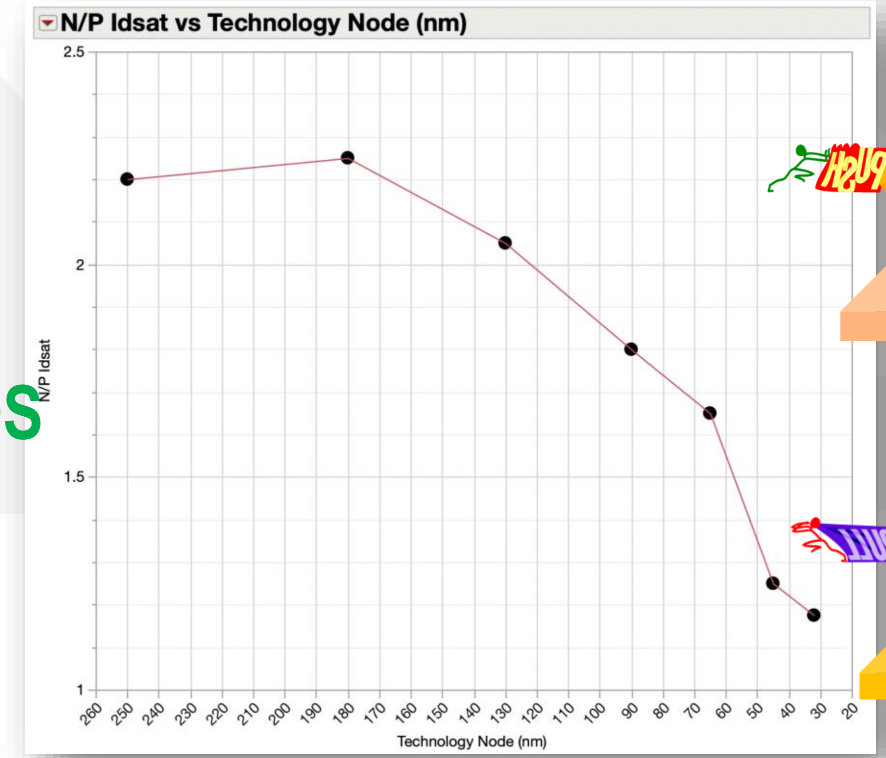
A note about Stress

- **A silicon channel is piezoresistive**
  - Lattice strain affects mobility.
  - Tensile stress improves NMOS, compressive stress for PMOS
  - Depends on lattice orientation: (100) fin top vs. (110) fin sidewall

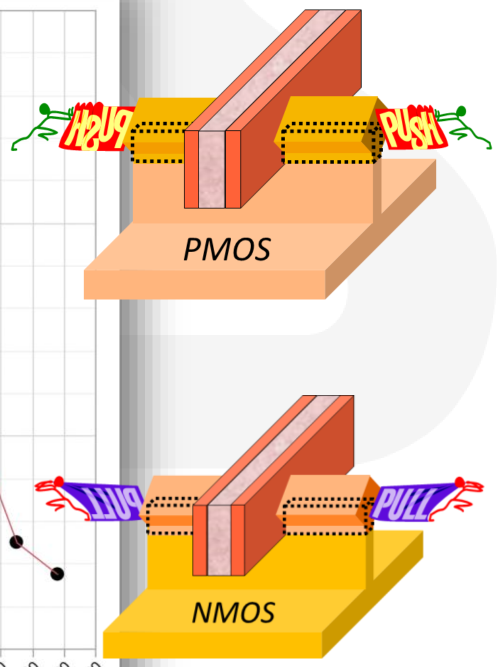


Faricelli, AMD, 2010

- **Stress has been more effective for PMOS**
  - This has caused beta (N/P) ratio to fall to about unity at 7nm.
  - Expected to change for nanosheets



Nahum, Intel



Loke, CICC, 2019

High-K – Metal Gate (HKMG)

- **Two problems with the transistor gate:**
  1. Thinning Oxide → Gate Leakage
  2. Polysilicon → high resistance, poly diffusion
- **Therefore, at 45nm-28nm move to HKMG**
  - High-K dielectrics enable thicker oxides
  - Metal gates improve resistance, EOT
- **However:**
  - High-K Materials have lower Energy Bandgap
  - Metal gates are sensitive to high temperatures during annealing

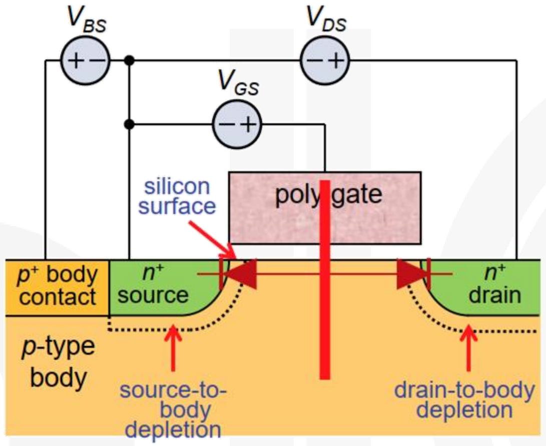
Gate dielectric Material	Dielectric constant (k)	Energy bandgap Eg (eV)	Conduction band offset ΔEc (eV)	Valence band offset ΔEc(eV)
SiO_2	3.9	9	3.5	4.4
Al_2O_3	8	8.8	3	4.7
TiO_2	80	3.5	1.1	1.3
ZrO_2	25	5.8	1.4	3.3
HfO_2	25	5.8	1.4	3.3

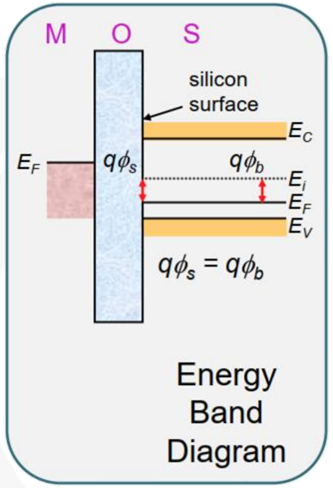
Touati, et al., J. New Technol. Mater.

Metal Gate and VT adjustment

- **VT of a transistor is primarily set by:**
  - The workfunction difference between gate and substrate
  - The doping in the junction
  - The backgate capacitance
- **In older technologies**
  - VT adjustment was achieved through channel implants
  - However, random doping fluctuations caused huge variation
- **FinFETs have intrinsic channels**
  - Therefore, RDF has basically "gone away"
  - VT adjustment is done through the workfunction of the metal gate

Flatband Condition ($V_{GS}=V_{FB}$)

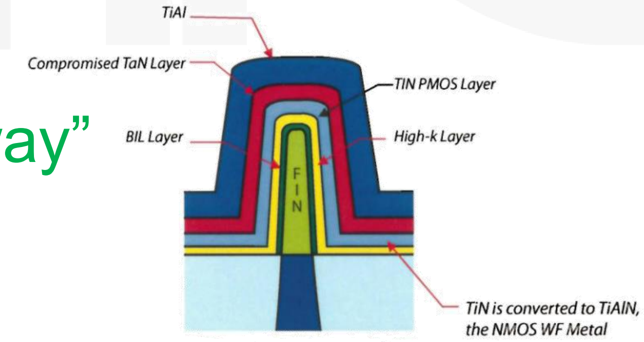


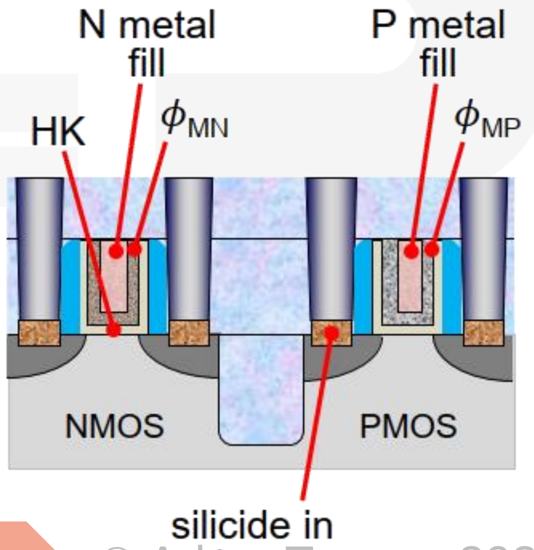


The workfunction of a material is the energy gap between the Fermi energy level and the vacuum.

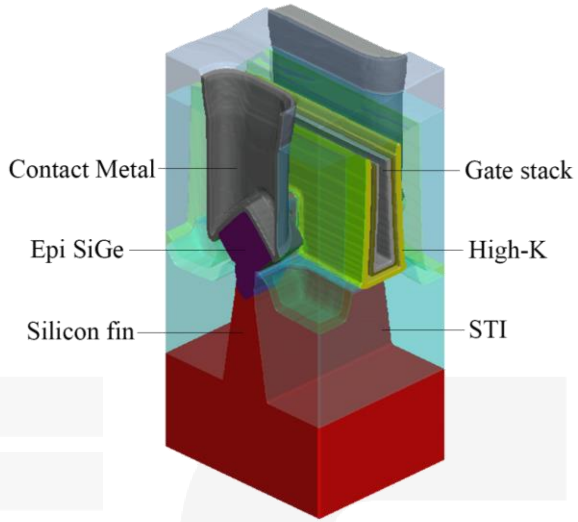
$$f(E) = \frac{1}{e^{(E-E_F)/kT} + 1}$$

Fermi-Dirac distribution function

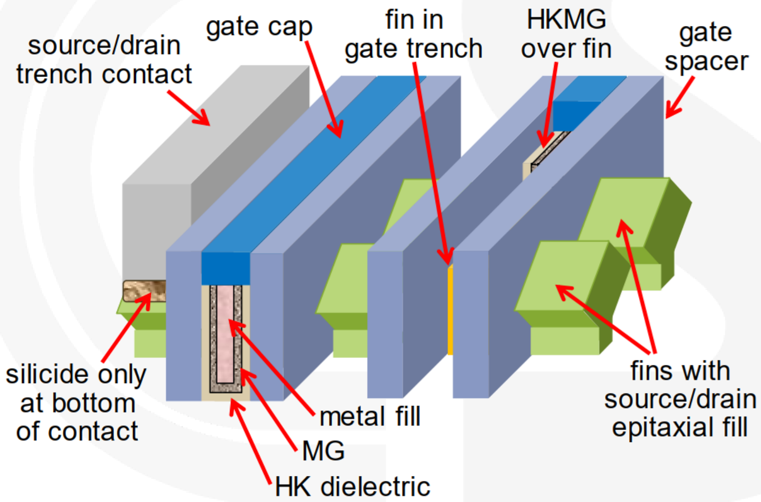


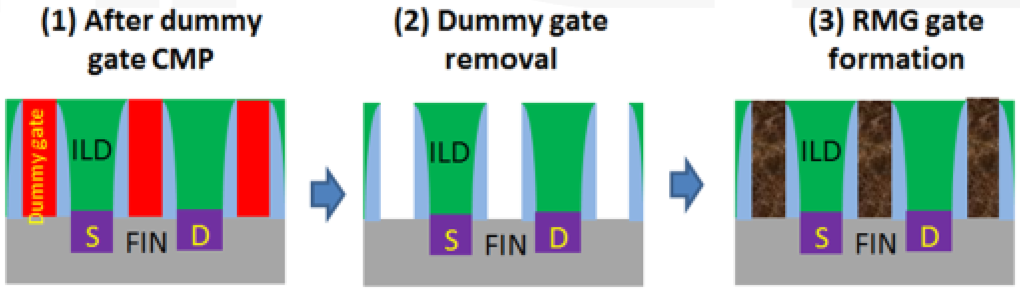


Replacement Metal Gate (RMG)



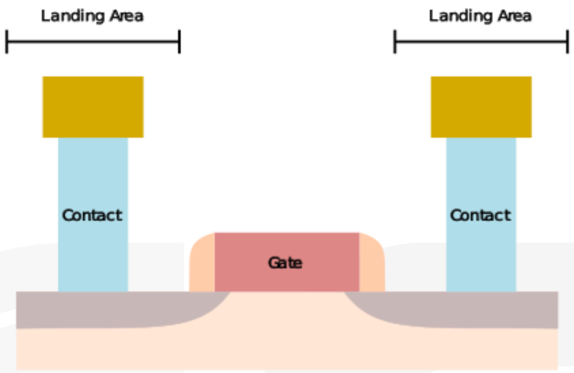
- **Metal Gates are sensitive to high-temperature process steps required for S/D engineering (epi, stress, etc.)**
- **Therefore, a "gate-last" approach is used:**
  - Form the gates with polysilicon. Also known as a dummy gate.
  - After S/D formation, etch poly gates.
  - Partially fill in with barrier and workfunction metal.
  - Fill in with low resistance metal.
- **This also impairs the Silicide, increasing S/D resistance**



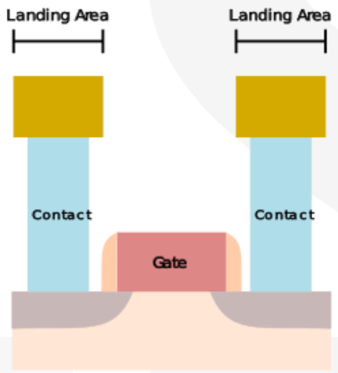


Jacob, et al., Globalfoundries

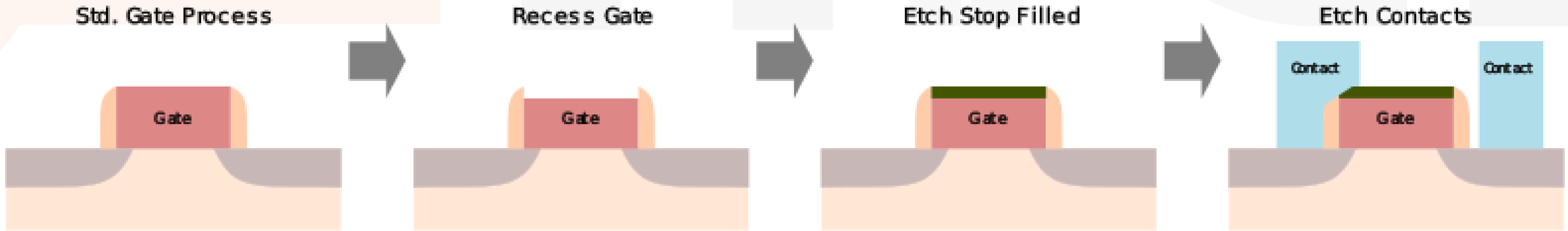
Middle-end-of-the-line (MEOL)



- **In older technologies, making contacts was easy**
  - The contacted gate pitch (CGP) was large enough to place gates and contacts next to each other.
  - But at tight gate pitches, any misalignment can short the contact to the gate.



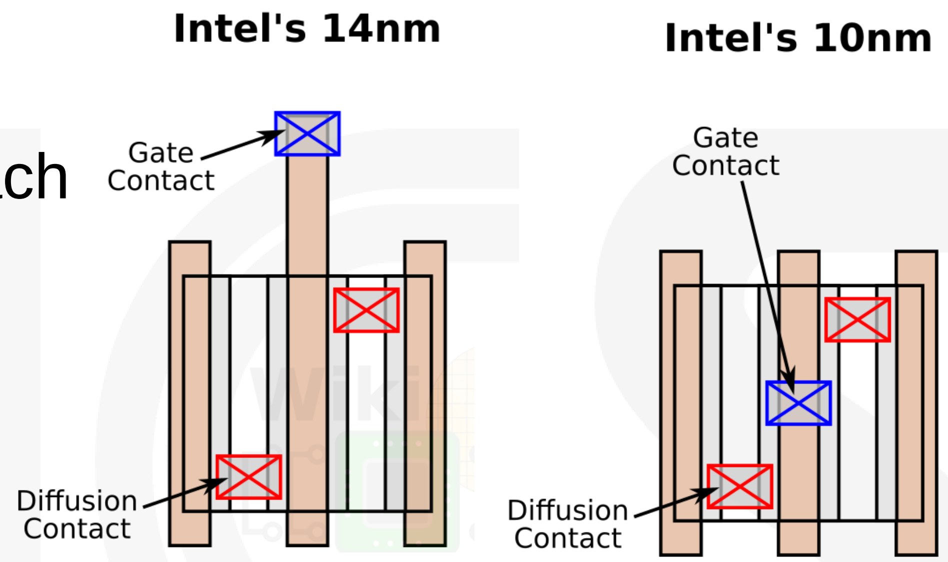
- **Therefore, Self-Aligned Contacts (SAC) were introduced**
  - A dielectric cap is added on top of the gate so that if the contact overlaps the gate, no short occurs.

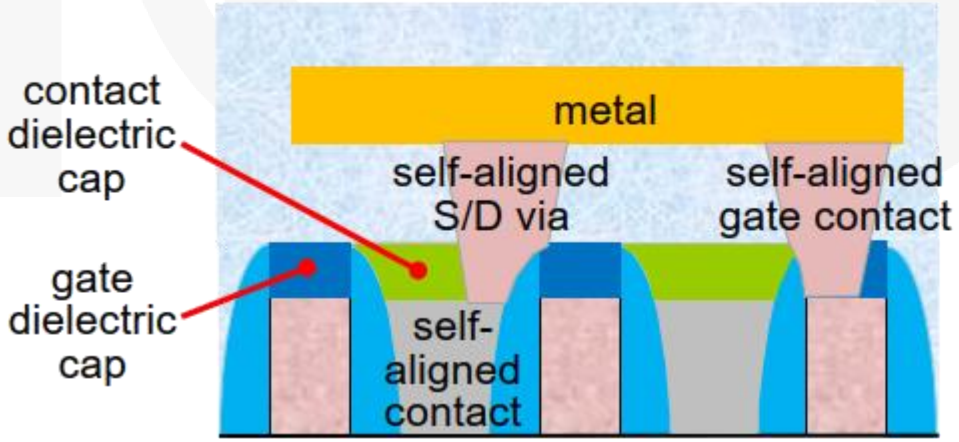


wikichip

Middle-end-of-the-line (MEOL)

- **SAC process is more complex in FinFET**
  - RMG is harder to recess than gate-last approach
- **Contact Over Active Gate (COAG) is desired**
  - Introduced by Intel in 10nm process
- **The MEOL process now includes many steps**
  - Cap gate and create SAC
  - Cap contact and create COAG
  - Create additional VIA0 to go through caps
- **Disadvantages:**
  - High Gate to S/D Contact capacitance
  - High S/D, MEOL & lower BEOL resistance

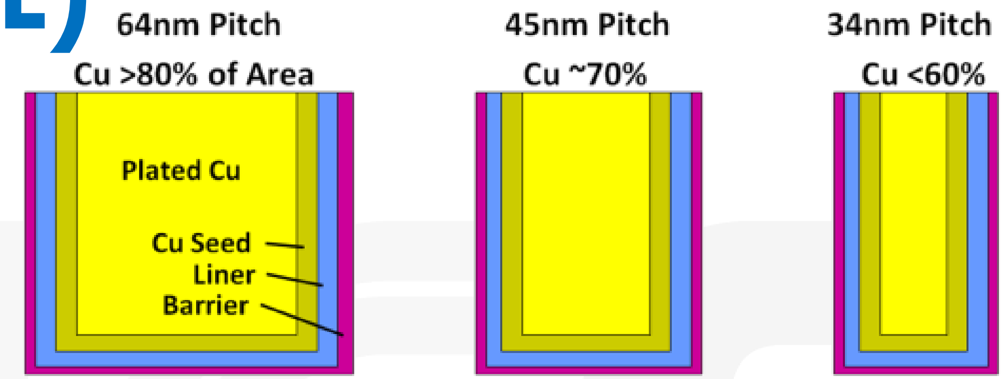




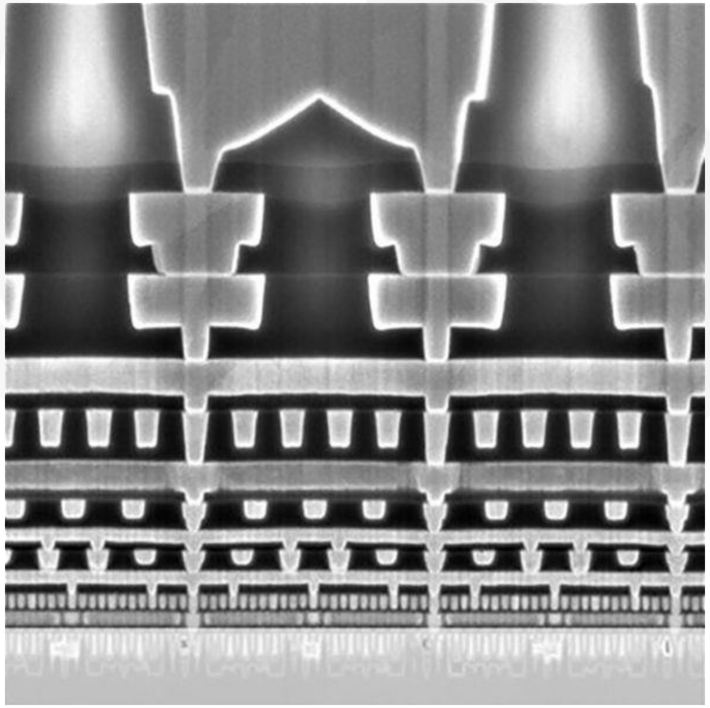
Yang, Qualcomm
Auth, et al. Intel

Backend-of-the-line (BEOL)

- Copper interconnect replaced aluminum
  - Lower resistance, Improved electromigration
  - Dual Damascene Process
- However, a barrier and liner are required
  - Plated copper area is reduced, increasing resistance
- At 10nm Intel started using Cobalt on M0 & M1
  - 5-10X Electromigration, 2X resistance

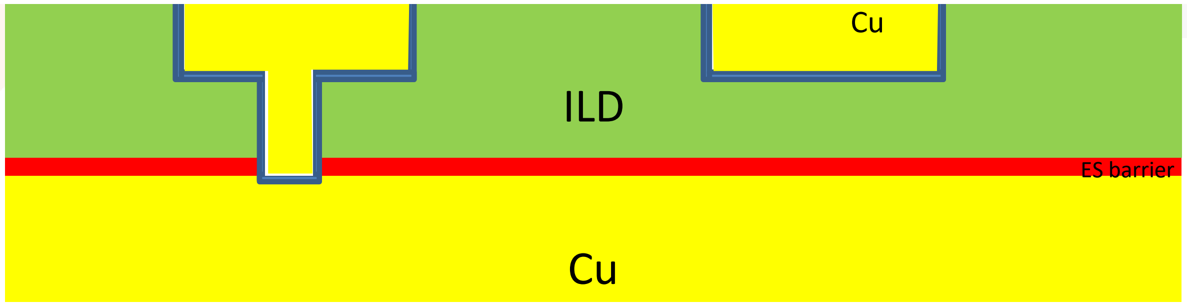


Jacob, et al., Globalfoundries



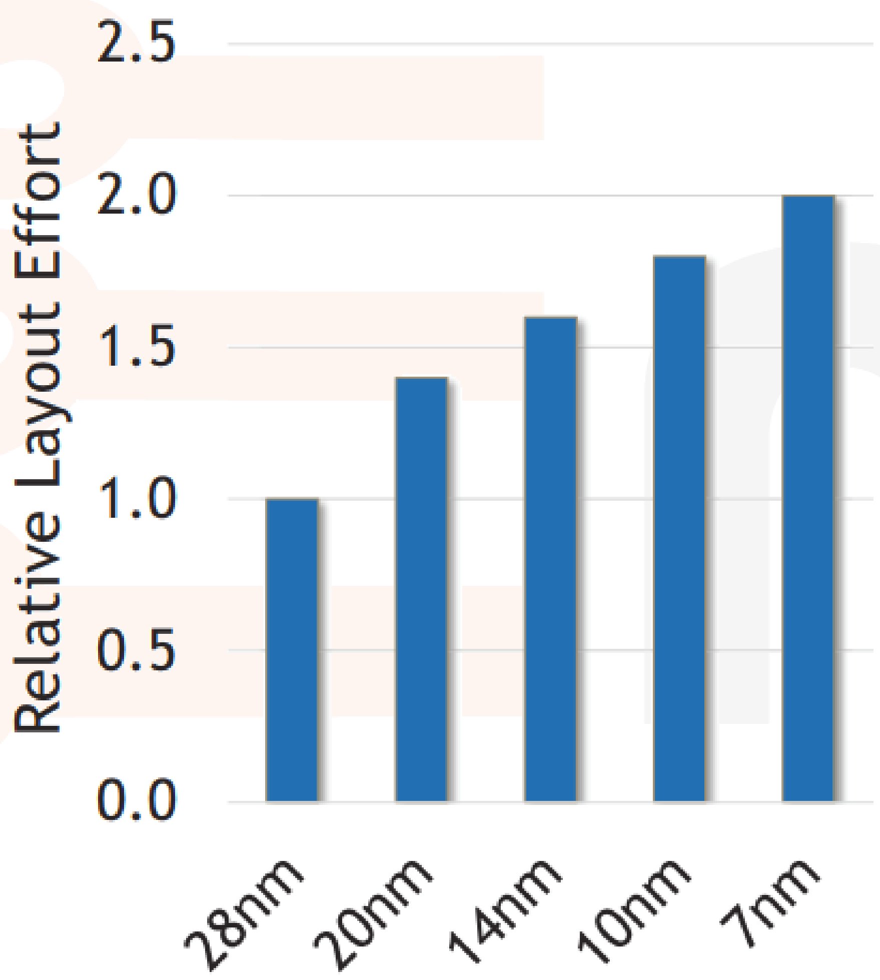
Intel 10nm metal interconnects cross section

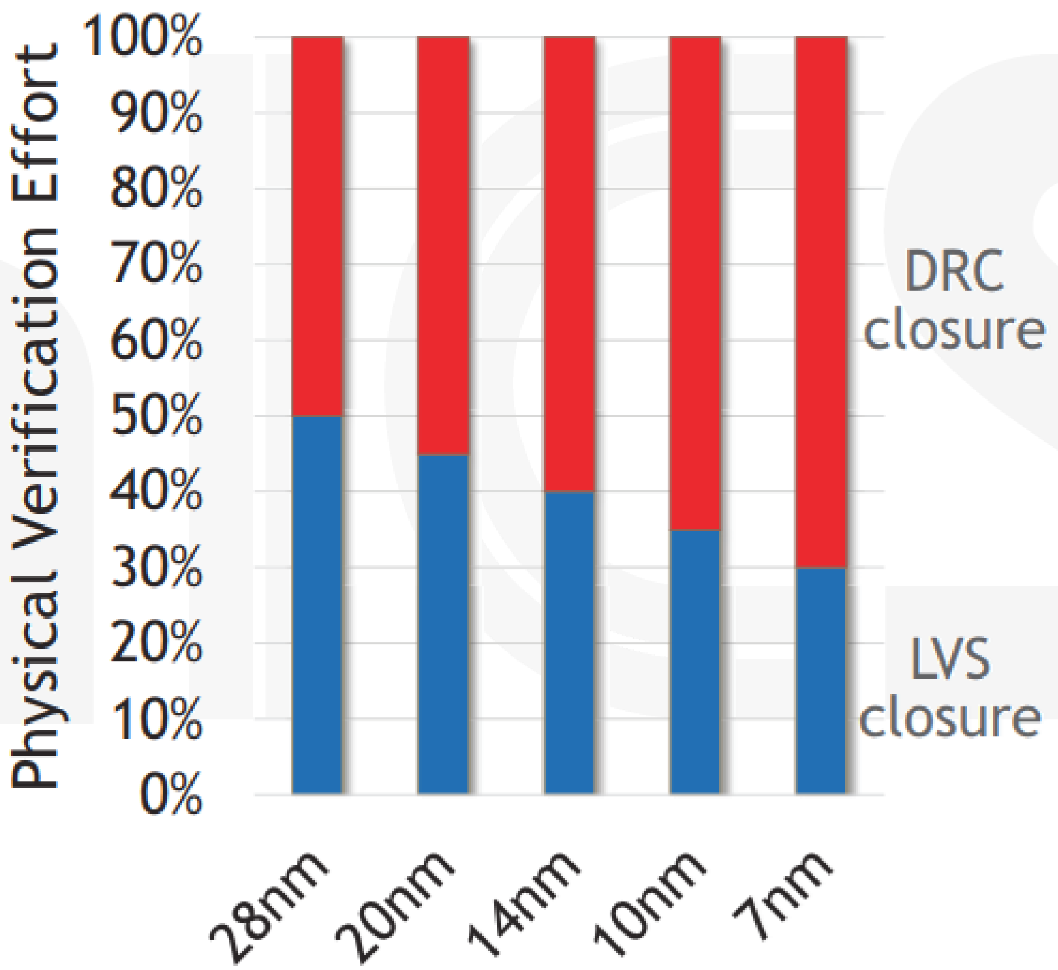
wikichip



FinFET Layout

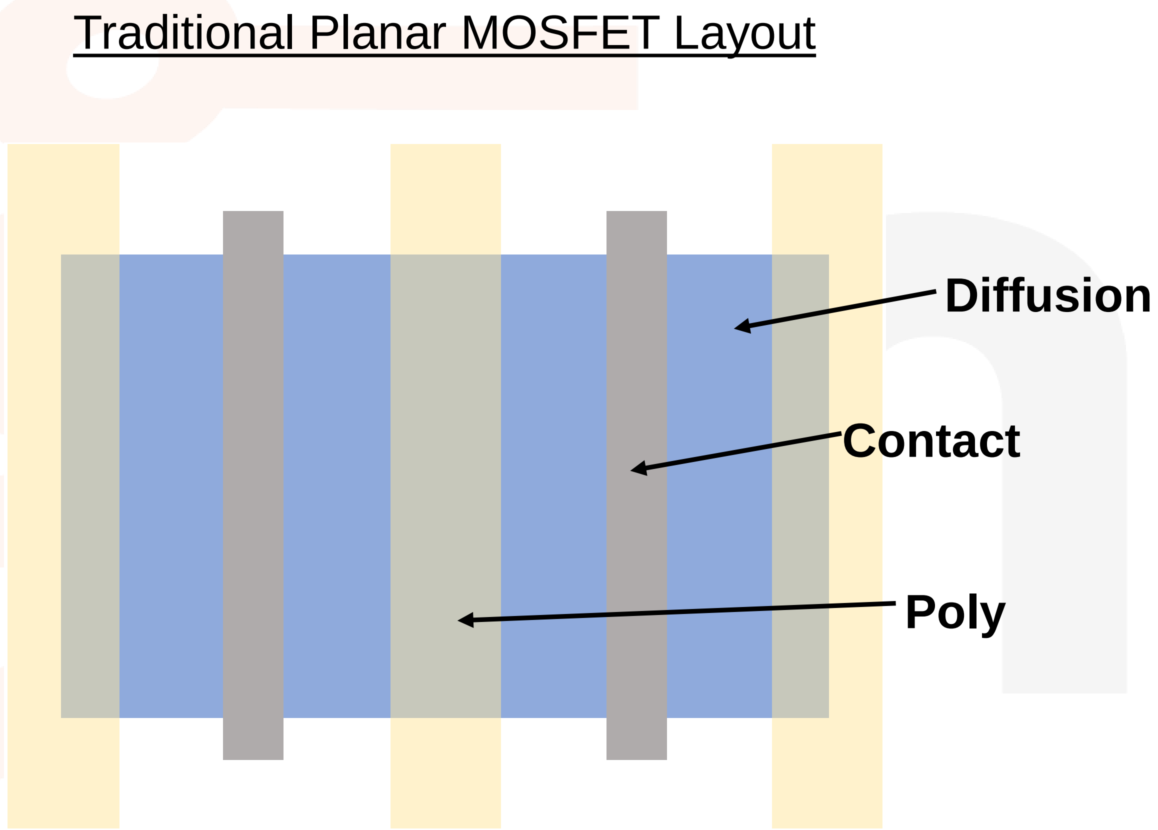
Layout Complexity

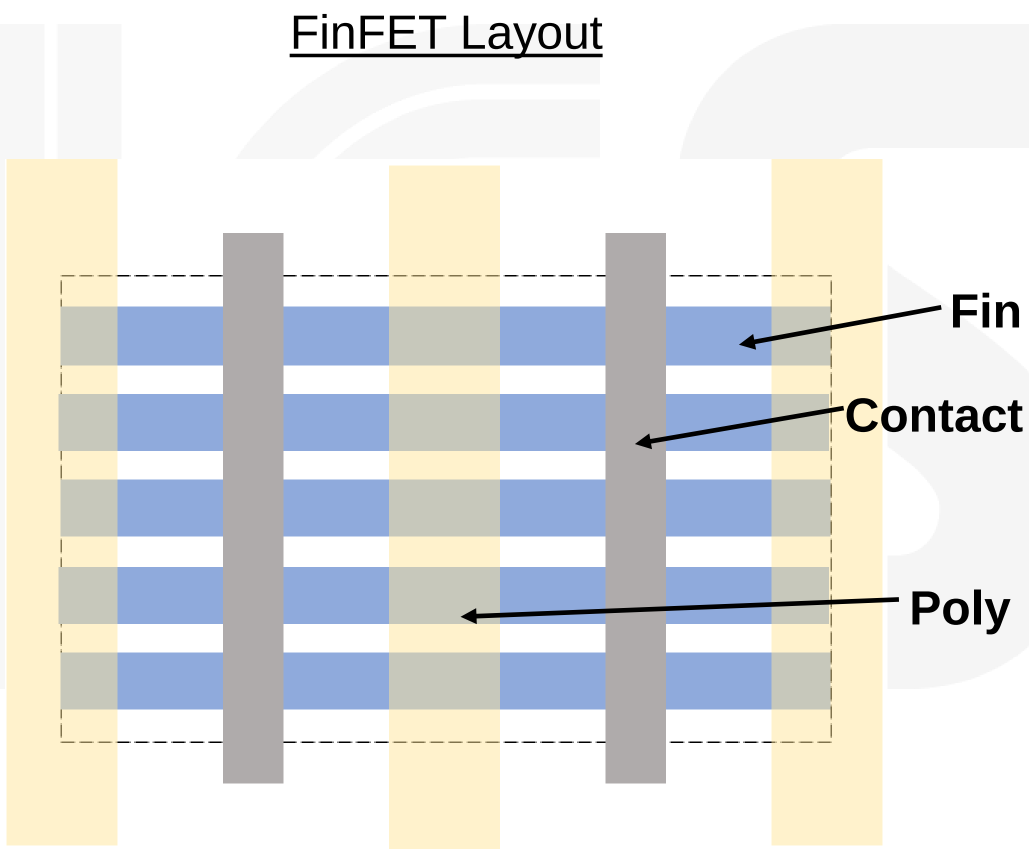




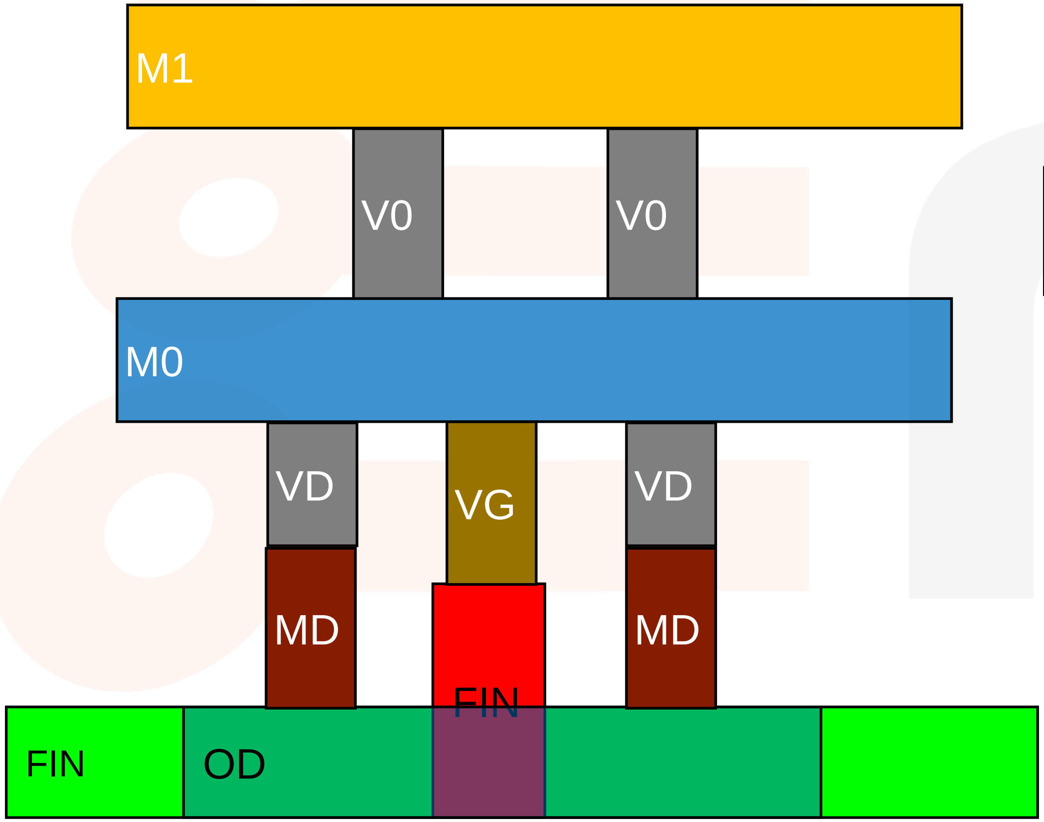
Leary, Qualcomm

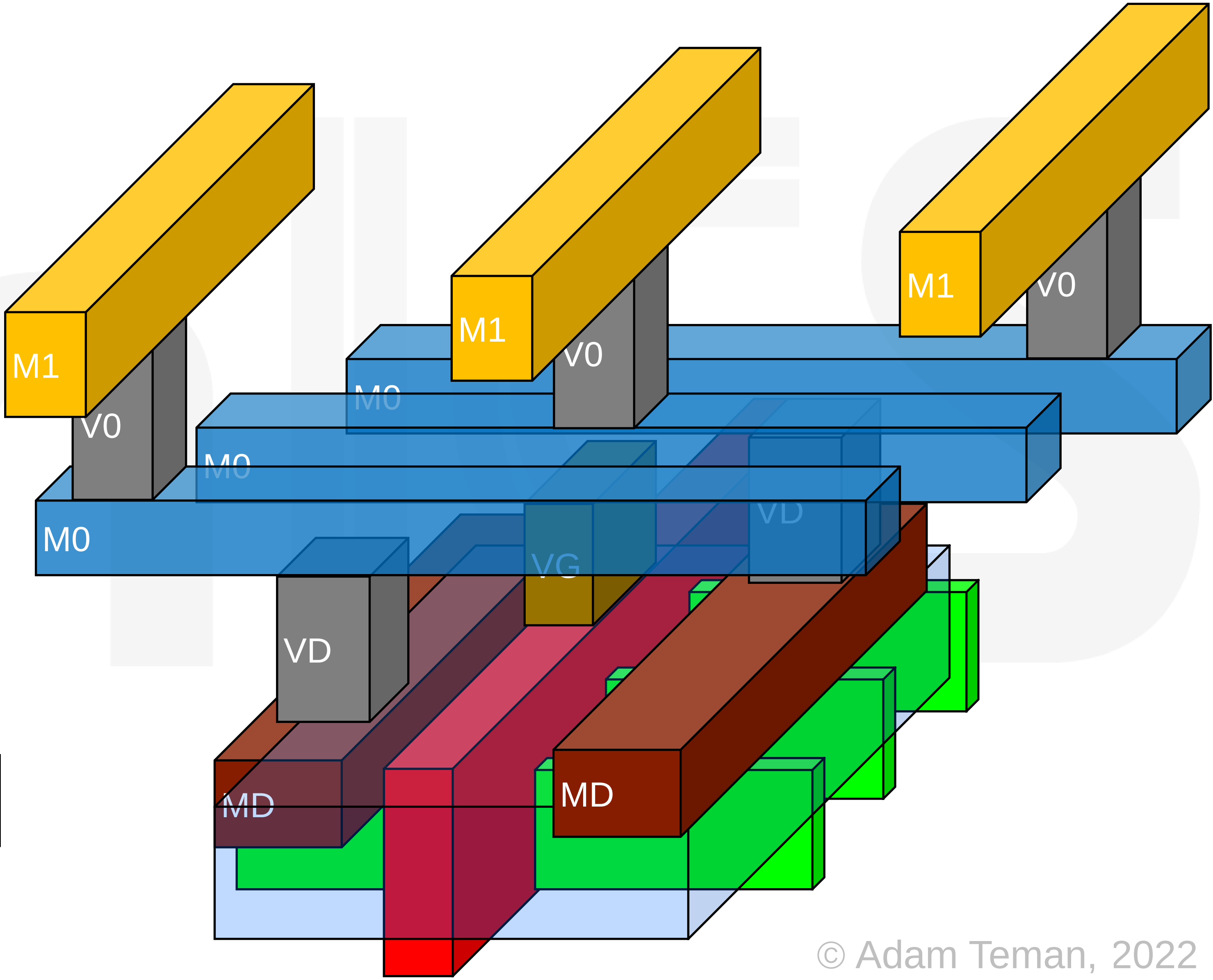
Planar vs. FinFET



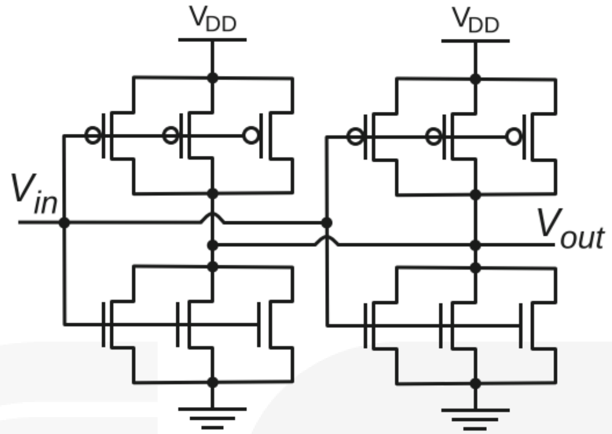


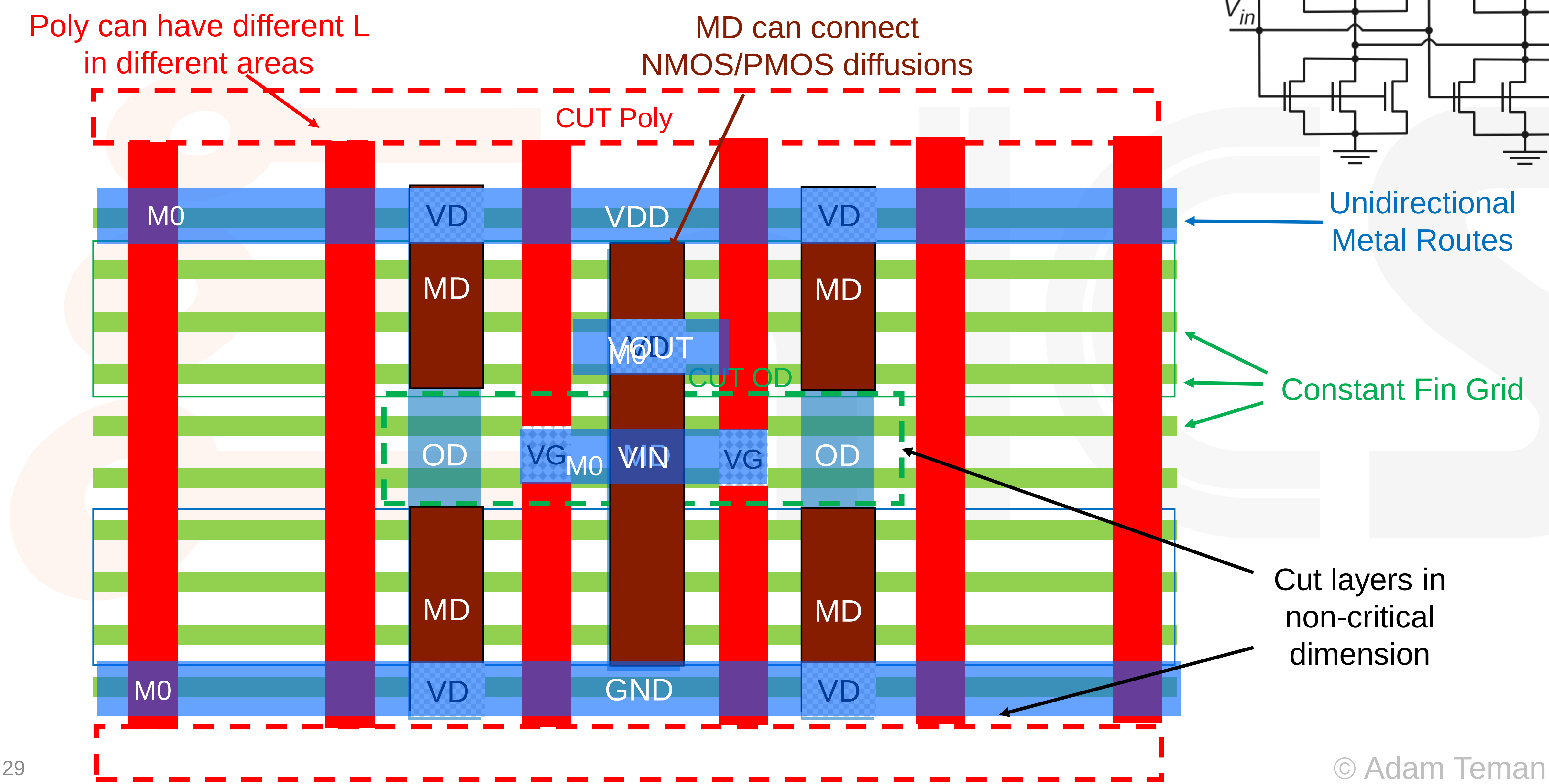
FinFET Stack (5nm)





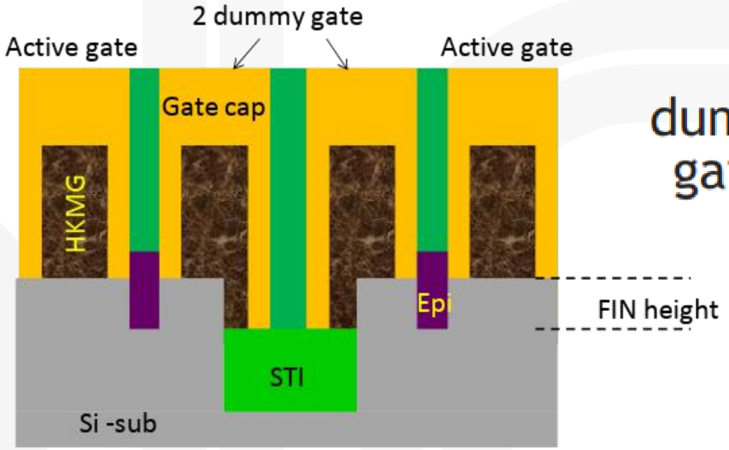
Layout in FinFET Technologies



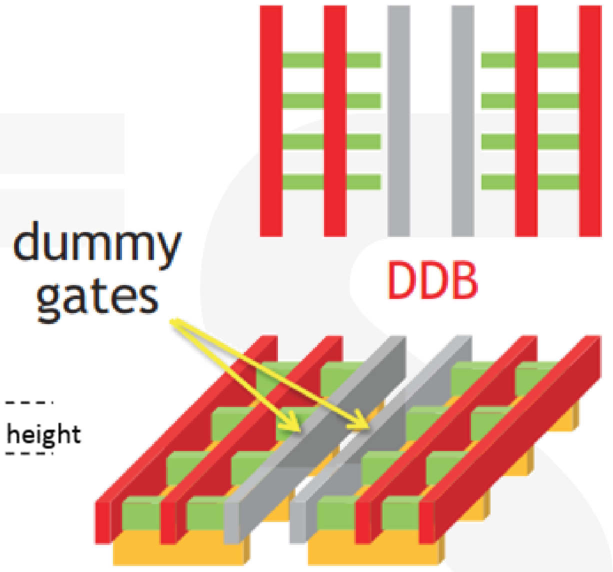


Diffusion Breaks

- **A diffusion break is required between two active areas:**
  - Blocks Epitaxial growth
  - Provides “back wall” for stressors
  - “Double Diffusion Break” (DDB) is done with two dummy gates and STI

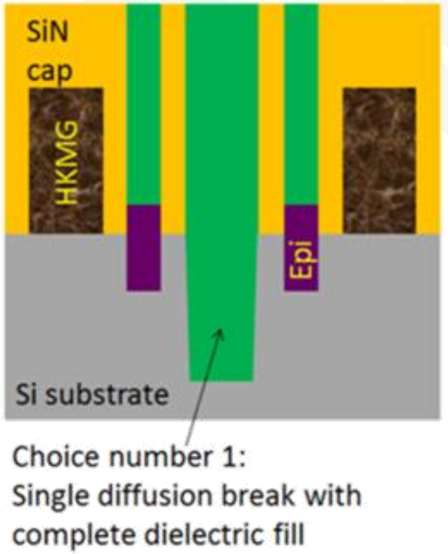


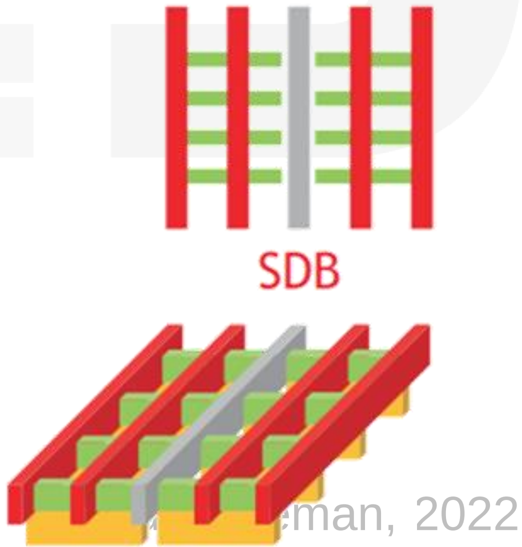
Jacob, et al., Globalfoundries



Yang, et al., Qualcomm

- **This wastes a lot of area**
  - Instead use only one dummy gate
  - “Single diffusion break” (SDB) saves area but is complex to process



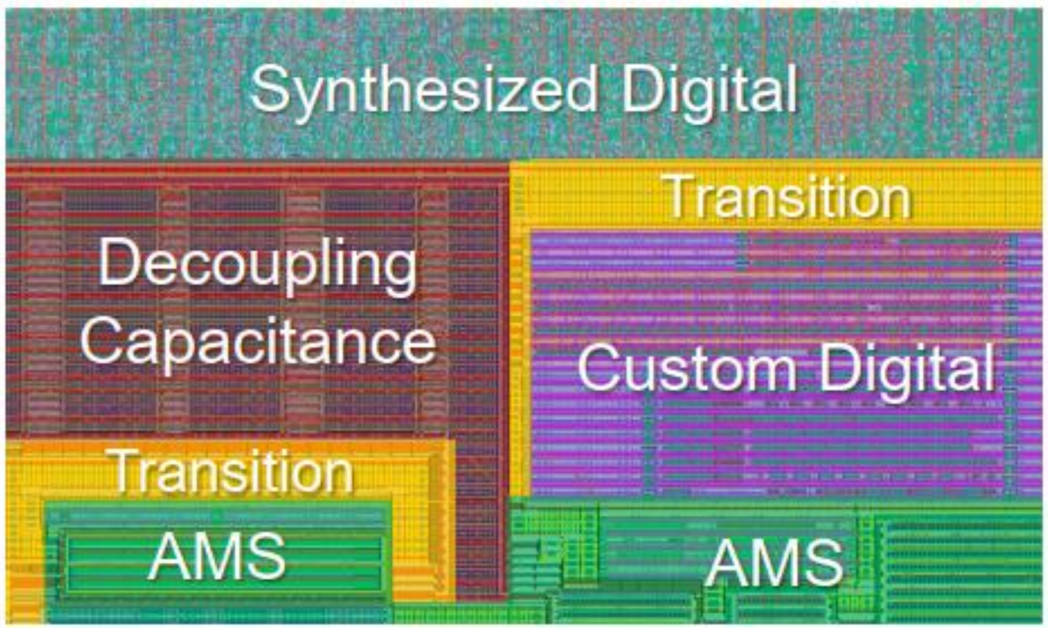


man, 2022

Density & Floorplan Considerations

- **Critical process steps are extremely sensitive to pattern density & loading**
  - 1000s of DRCs, many very tough to pass, increasingly restrictive & foreign
- **DRCs reduce unmodeled long-range systematic & random variation**
  - → iterative rework of smaller cells
  - Area, perimeter, gradient
  - Contacts, vias, cuts, tight-pitch metal
  - Larger checking windows
  - Density union of multiple metal levels
- **Floorplanning more tedious & bloated**
  - More dummy gates, well taps, guard rings
  - Wasteful transitions between different device types & pattern densities

Example Mixed-Signal Floorplan



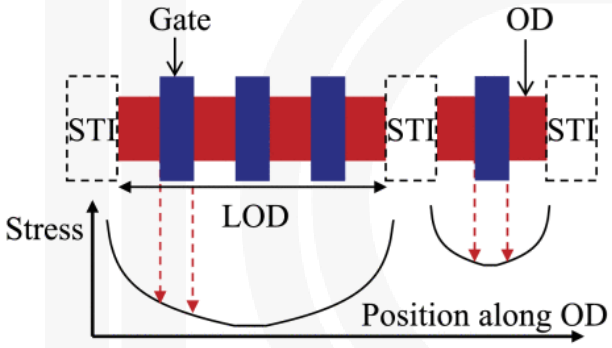
Yang, et al., Qualcomm

Layout Dependent Effects and Parasitics

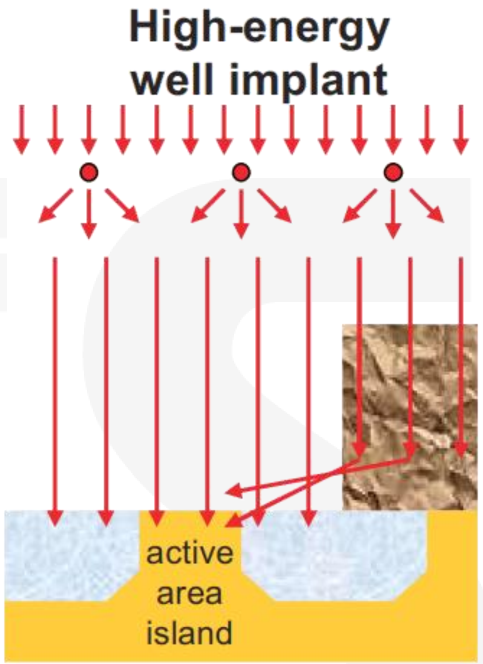
EnICS
Emerging Nanoscaled Integrated Circuits and Systems Labs
The Alexander Kofkin Faculty of Engineering
Bar-Ilan University

Layout Dependent Effects (LDEs)

- **Fabricated device characteristics have shown a high dependency on layout features for several generations:**
  - Well Proximity Effects (WPE)
  - Length of Diffusion (LOD)
  - Oxide-to-Oxide Spacing (OSE)
- **FinFETs further introduce LDEs:**
  - Stress LDEs more significant due to stronger stressors
  - Gate cut stress
  - HKMG LDEs
- **All of this causes more pre- to post-layout simulation differences**



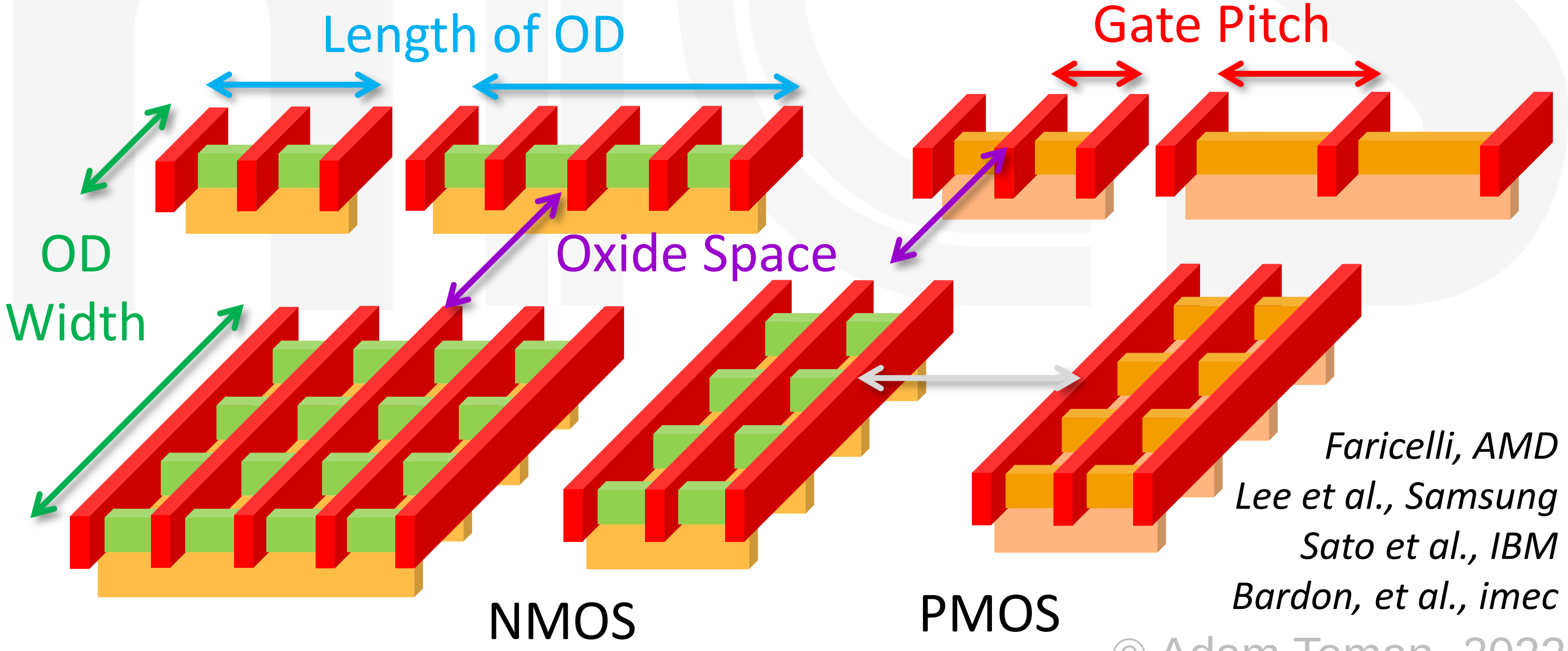
Ou, et al., IEEE TCAD, 2016



Faricelli, AMD, 2010

Stress LDEs in FinFETs

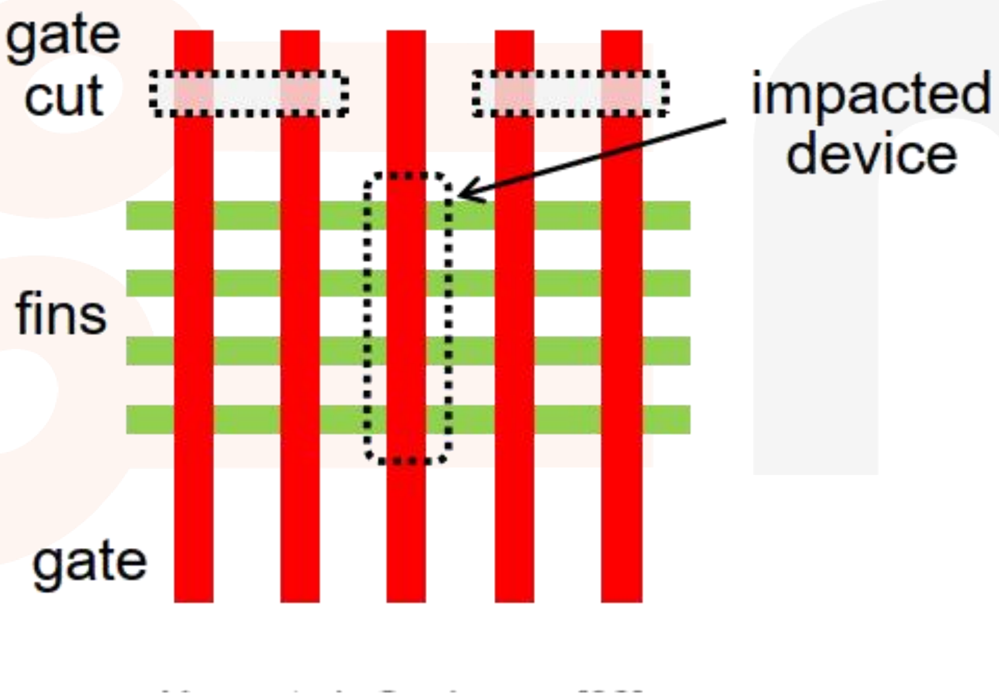
- **Stress LDEs are caused by:**
  - Longer diffusions (OD Length)
  - Wider diffusions (I_D/fin not constant vs. # fins)
  - Oxide spacing
  - Gate pitch
- **Models capture $\Delta\mu$ & ΔV_T (some effects as early as 130nm)**

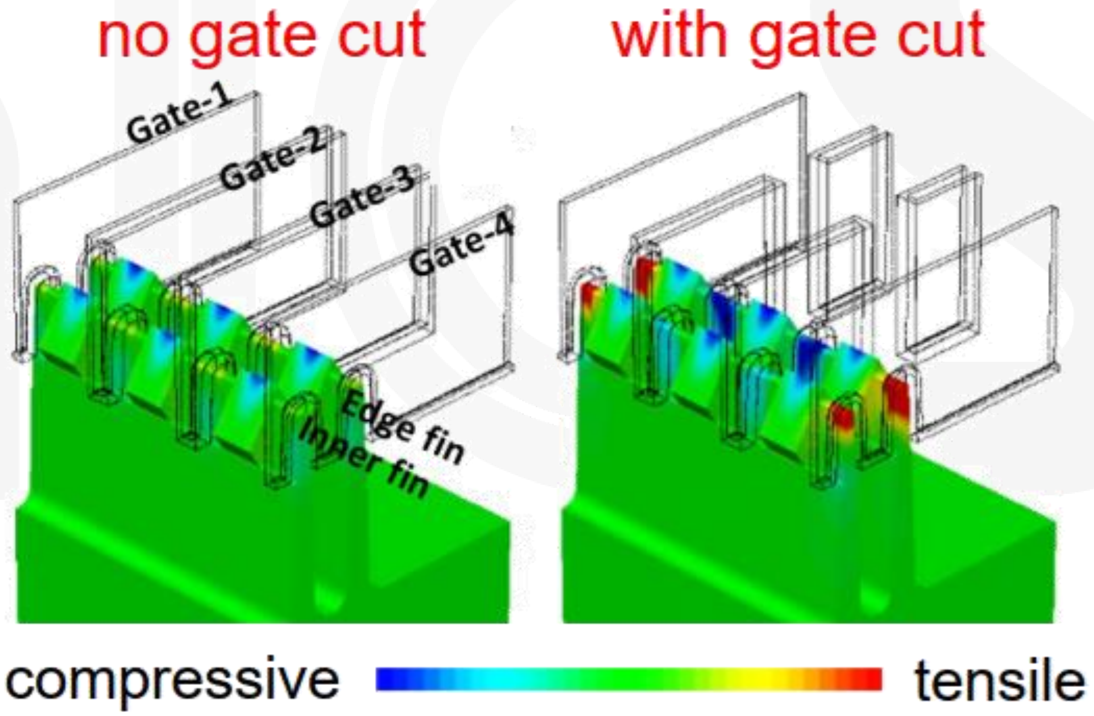


Faricelli, AMD
Lee et al., Samsung
Sato et al., IBM
Bardon, et al., imec

Gate Cut Stress LDE

- **Gate cut disrupts mechanical support of continuous gate & stress near cut**
  - → $\Delta\mu$ & ΔV_T, modeled in post-layout netlist starting in 16/14nm

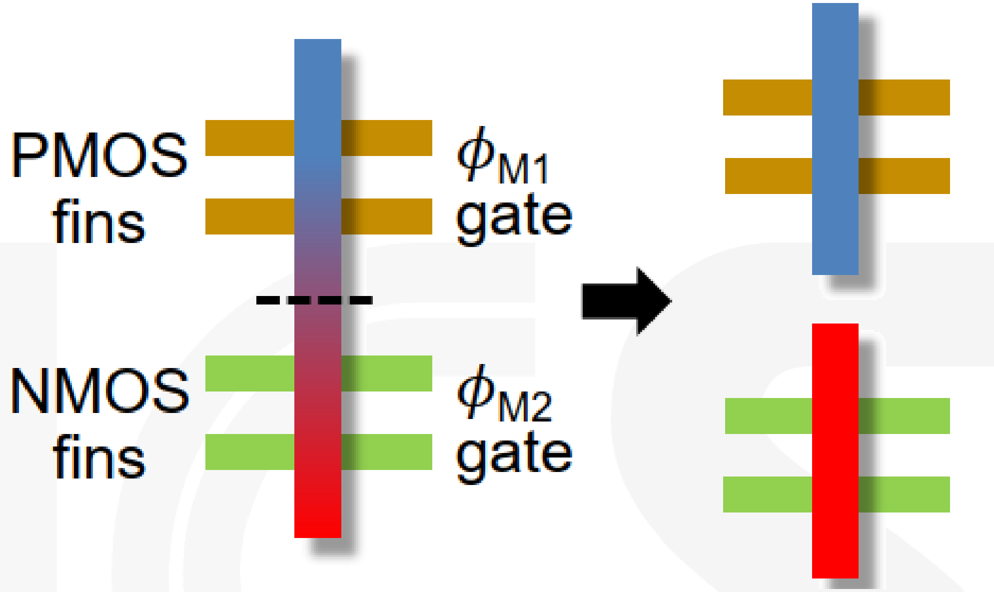




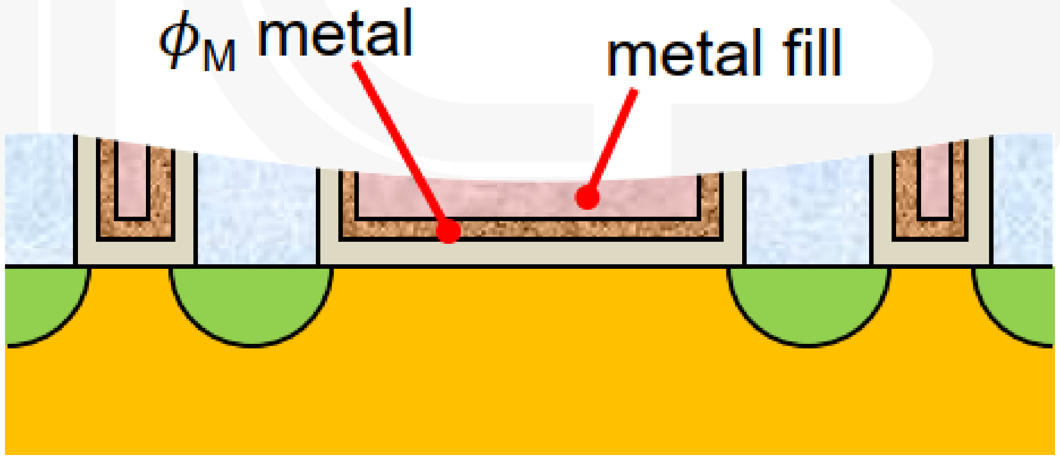
Yang et al, Qualcomm

HKMG LDEs

- **Metal Boundary Effect**
  - ΔV_T near border of different Φ_M due to Interdiffusion of Φ_M.
  - Mitigated with gate cut but costs area
  - Models capture $\Delta\mu$ & ΔV_T.
- **Density Gradient Effect (DGE)**
  - Gate density gradients
    - → ΔV_T & variation from RMG CMP dishing
  - Φ_M influenced by metal fill & sidewall Φ_M
  - Not modeled, contained with DRC



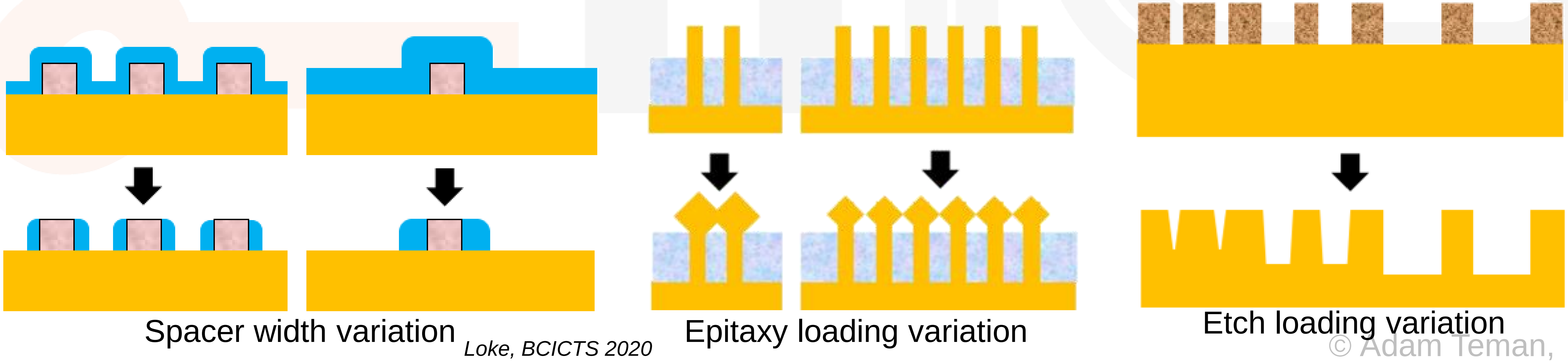
Hamaguchi et al., Toshiba



Yang et al., Qualcomm

Process Loading Variation

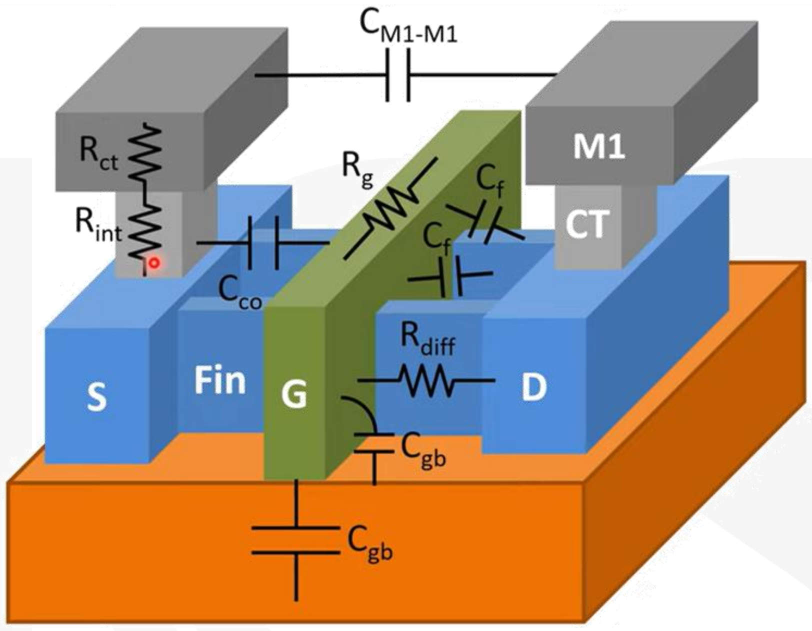
- **Local pattern density modulates deposition rate, etch rate/profile & CD**
  - Deposition loading: Spacer width variation (gate and metal CD)
  - Epitaxy loading: S/D volume variation (S/D resistance & channel stress)
  - Etch loading: Depth/profile variation (Lgate, fin & metal height)
  - Rapid Thermal Annealing (device variation)
  - Chemical Mechanical Polishing (variation in STI, poly, RMG, MEOL & BEOL)



Loke, BCICTS 2020

Parasitic Resistance and Capacitance

- **High Resistance:**
  - Contacts, Metal Gate, Low Metals
- **High Capacitance:**
  - Tight metal pitches
  - S/D trench contacts & gate form vertical plate capacitors



Cap	Planar	FinFET
C_f	1X	2.0X
C_{co}		1.8X
C_{gb}		1.2X

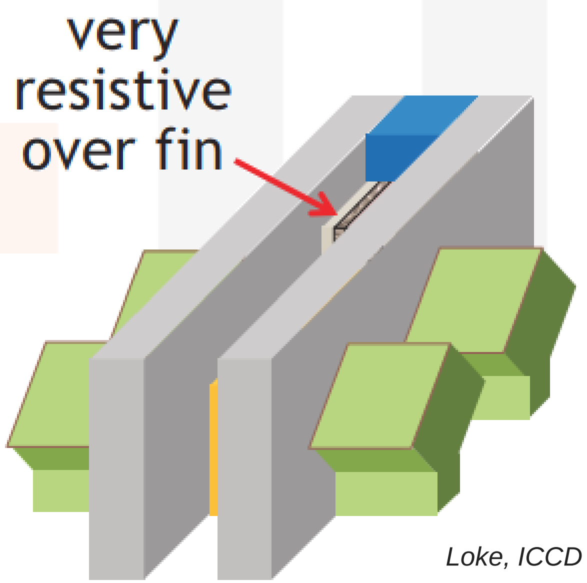
[Lawrence Loh, ISSCC’18]

Res	Planar	FinFET
R_g	1X	1.5-2.5X
R_{ct}		1.5X
R_{int}		1.3X
R_{diff}		1.3X

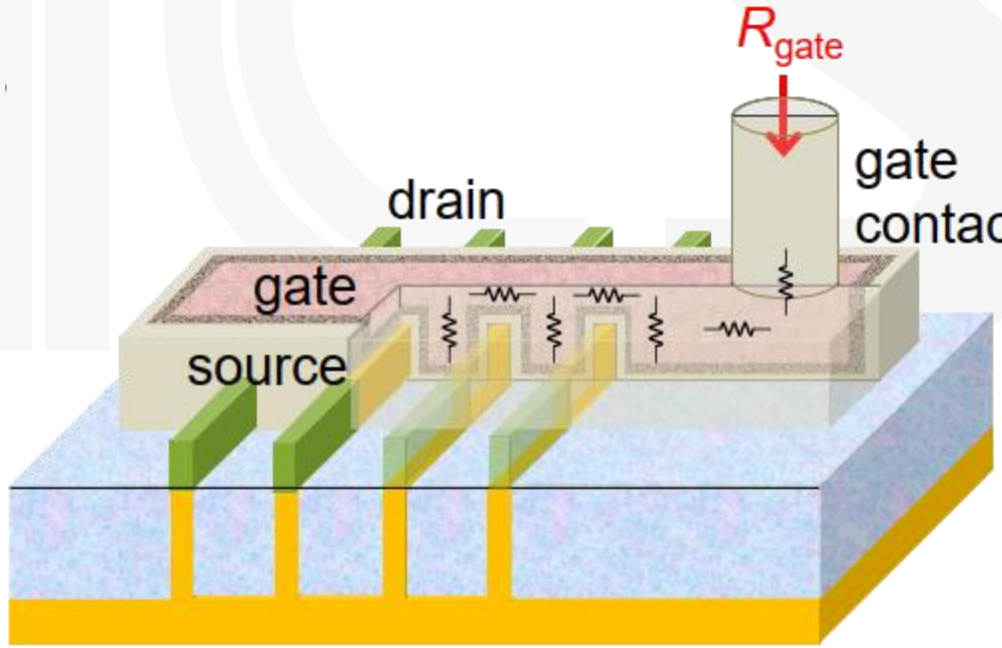
Gate Resistance

- **Metal gate *should have* lower resistance than poly gate**
  - But gates are very thin.
  - Workfunction Metals have high resistance.
- **Lower resistance metal filled on top of workfunction metal**
  - But still, this is very little conductive metal

- **Another point**
  - It's tough to make thick oxide I/Os!



Loke, ICCD

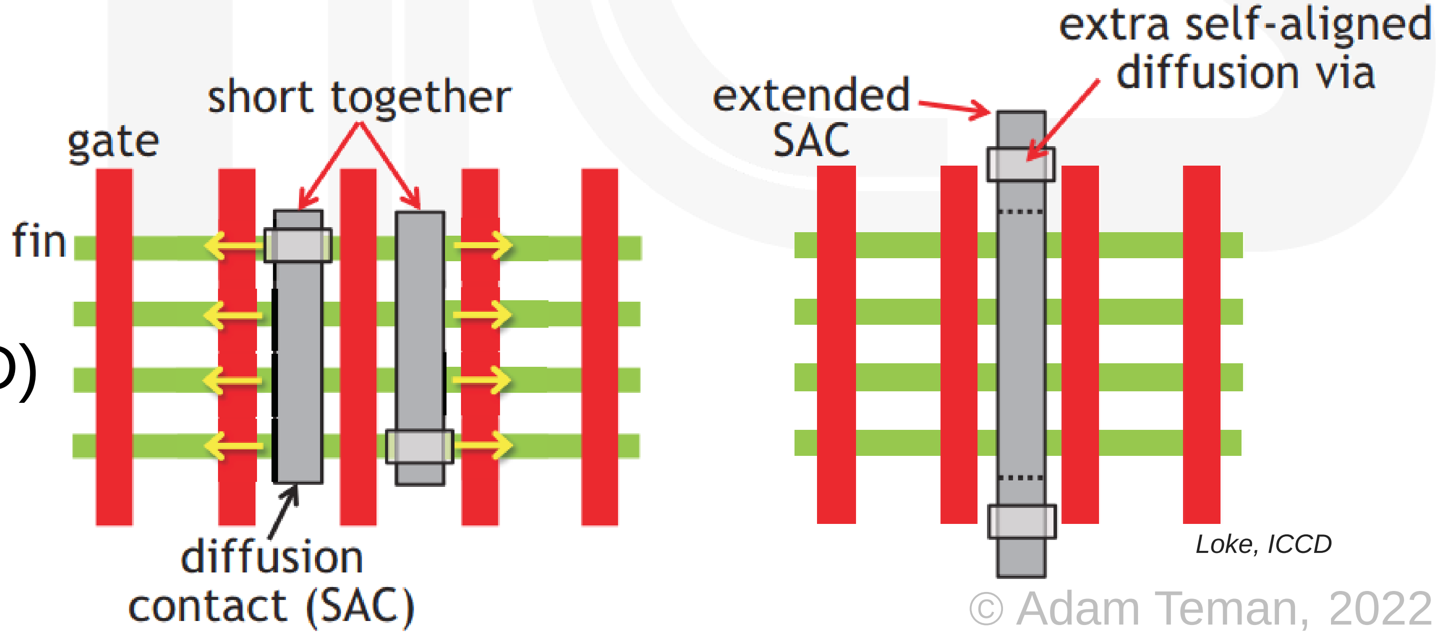


Wu & Chan, HKUST / Lee, Intel

Diffusion & MEOL Resistance

- **Traditionally:**
  - Try to share diffusions for smaller S/D capacitance
  - But S/D resistance is now a much bigger problem than capacitance
- **This becomes challenging for high-current circuits**
  - e.g., I/O drivers, clock buffers

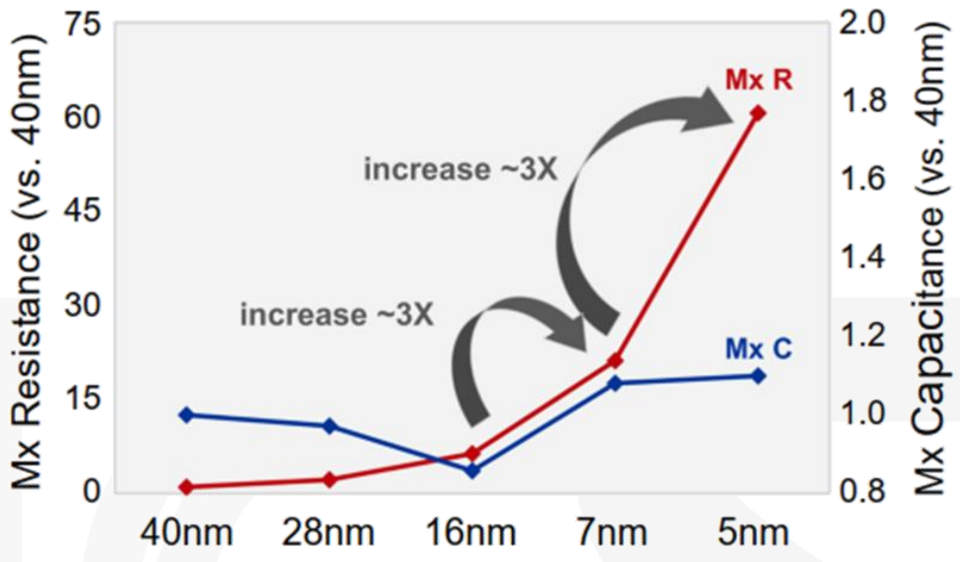
- **Therefore:**
  - Double-source layout
    → S/D $R_{contact}$ is halved
  - Extend self aligned contact (MD) to land extra diffusion via

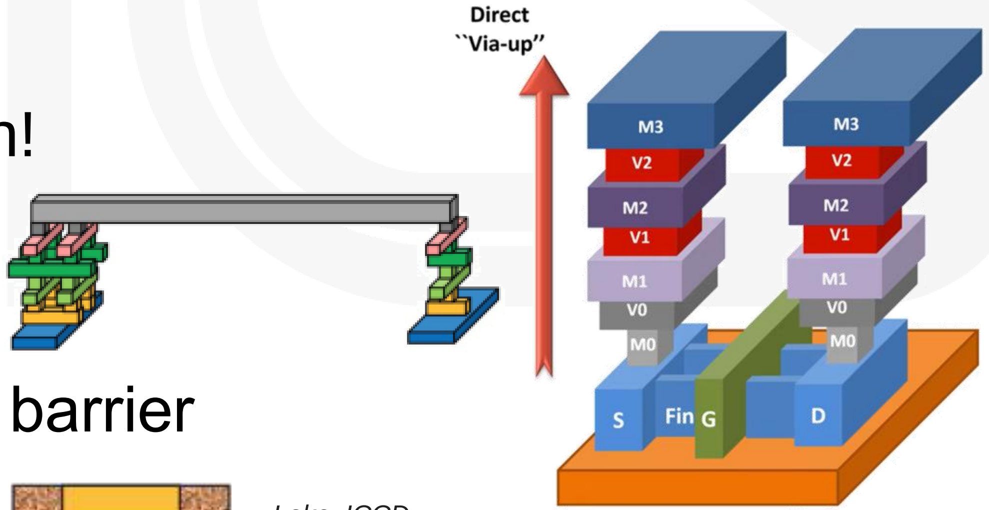


Loke, ICCD

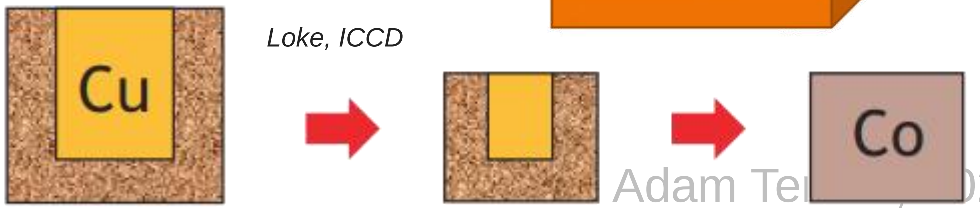
BEOL Resistance

- **Copper interconnect used for low resistivity**
  - However, copper diffuses into ILD.
  - Need barrier, liner and seed layers.
- **Local interconnect (MX) aggressive pitch scaling**
  - Dense logic routing→less die area & cost
  - But not much copper left in the wire…
  - 6X rise in resistance from 80nm to 48nm pitch!
- **Therefore:**
  - Remove seed, liner layers.
  - Use cobalt & ruthenium which don't require a barrier
  - Despite higher ρ – less wire resistance!
  - Use M3 and up for inter-cell routing.

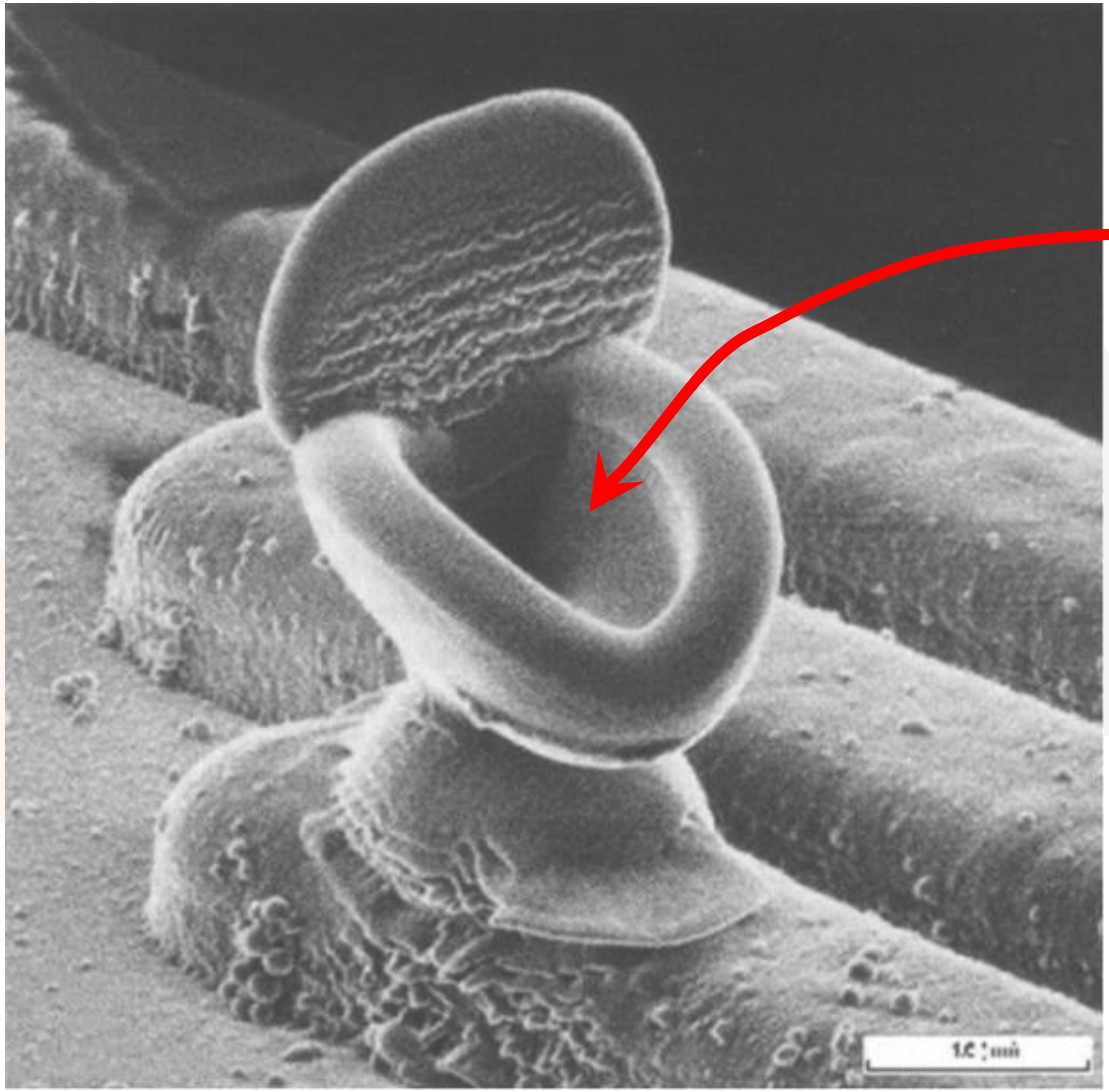




Loke, ICCD



Parasitics Summary…



Half of your performance from scaling is going here!

Courtesy of Greg Yeric, ARMTechCon 2016

https://madematics.com/2011/09/13/nanotoilet/

FinFET Node Models

- **FET models**
  - BSIM-CMG – based on channel surface potential, less equation fitting
  - *Target-based* for latest nodes, more silicon influence in mature nodes
  - Prone to model-vs.-silicon gap from increasing density & loading effects
- **BEOL models**
  - Electrical information provided, limited to no physical stack-up details
  - Less pessimistic corners for relaxed timing closure (customer pressure)
- **Usual reliability models (HCI, BTI, TDDB, EM)**
  - Vague allowable VDD, depends on application
- **Foundries extremely paranoid to protect their technology IP from competitors**
  - Process corner methodologies & many model parameters encrypted
  - Limited physical information available – even basic dimensions (e.g., L_{gate}) not real
  - CD bias & mask booleans to conceal process details (e.g., RMG flow for multiple VT)

Overcoming Process/Model Immaturity

- "**All models are wrong, but some are useful**"

  George Box (1919-2013), British Statistician

- **Corollary:**
  "**All simulations are wrong, but some are useful**"

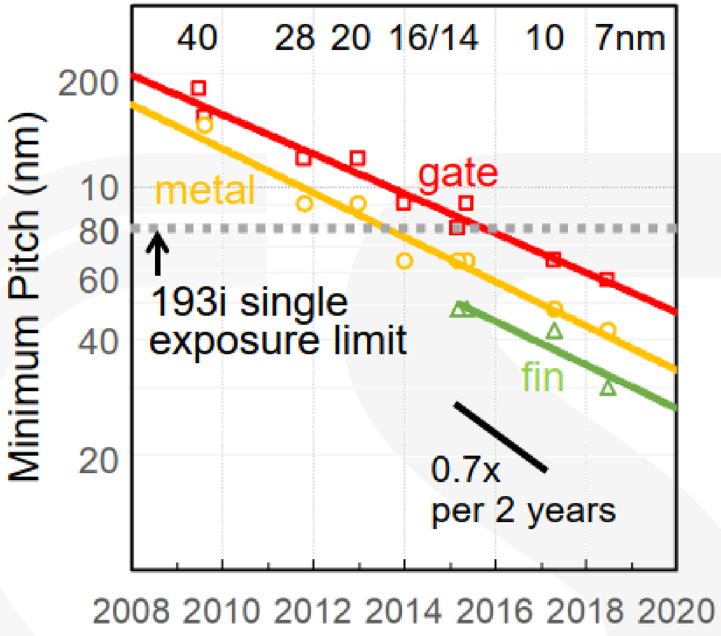
  Alvin Loke, NXP/TSMC

Layout Guideline	Reduced Exposure
Use continuous OD stress plateau	Stress LDEs (S/D epitaxy, STI)
Attach dummy FETs to OD ends	
Avoid single-diffusion break	
Use only one Φ_M in each gate	Metal boundary LDE
Avoid using gate as interconnect	Gate cut LDE
Contact gate on both sides	Gate resistance, Φ_M tuning to adjust V_T
Use groups of fewer fins	
Use double-source layout for high-I nets	S/D contact resistance & epitaxy
Extend S/D contact to land extra via	MEOL & S/D resistance
Do not push DRCs to limit	DRC updates

Some Current Trends

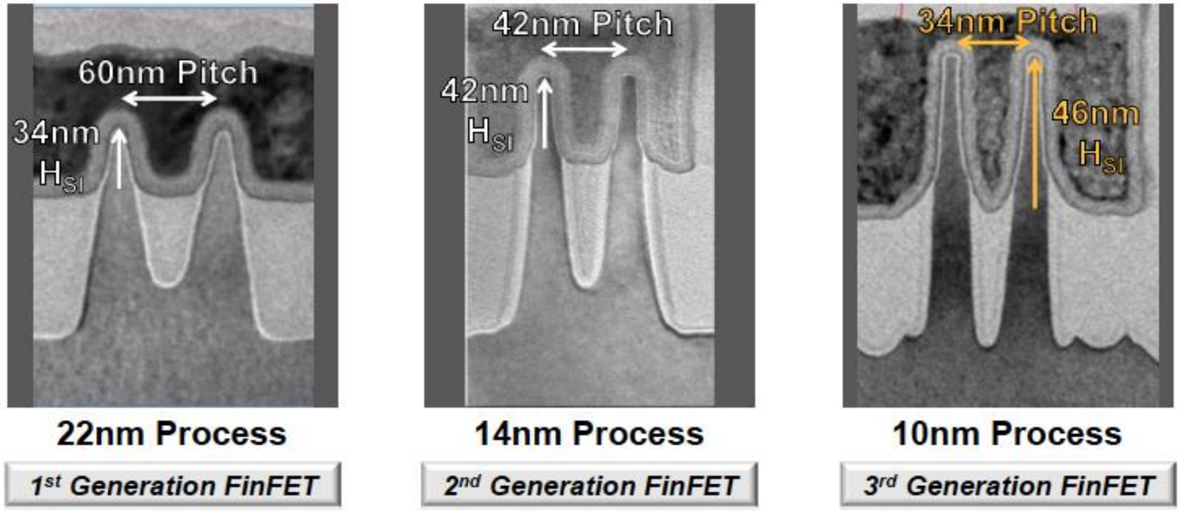
A note about process node naming

- **Scaling rate slower than 0.7x per node**
  - Node name just marketing number for equivalent PPA
  - Physical gate lengths haven’t scaled below ~14nm
  - 193i Single exposure pitch limit ~80nm
- **“Intra Nodes” on official scaling roadmaps**
  - Process optimizations for performance improvements and yield enhancements



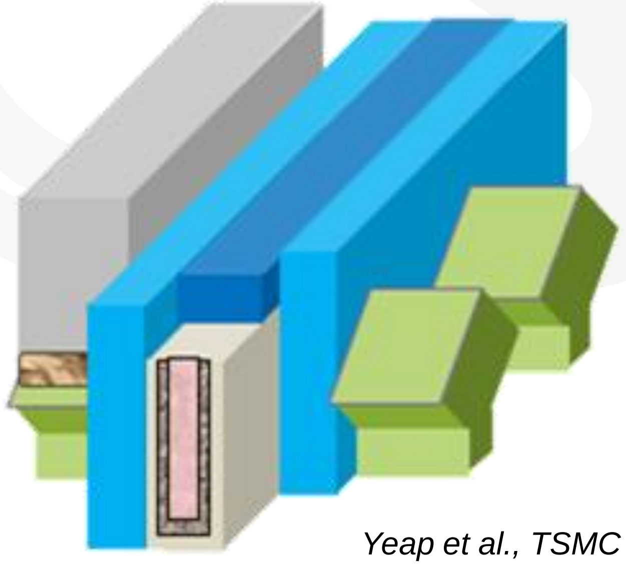
Loke, BCICTS

Zheng Guo et al. Intel



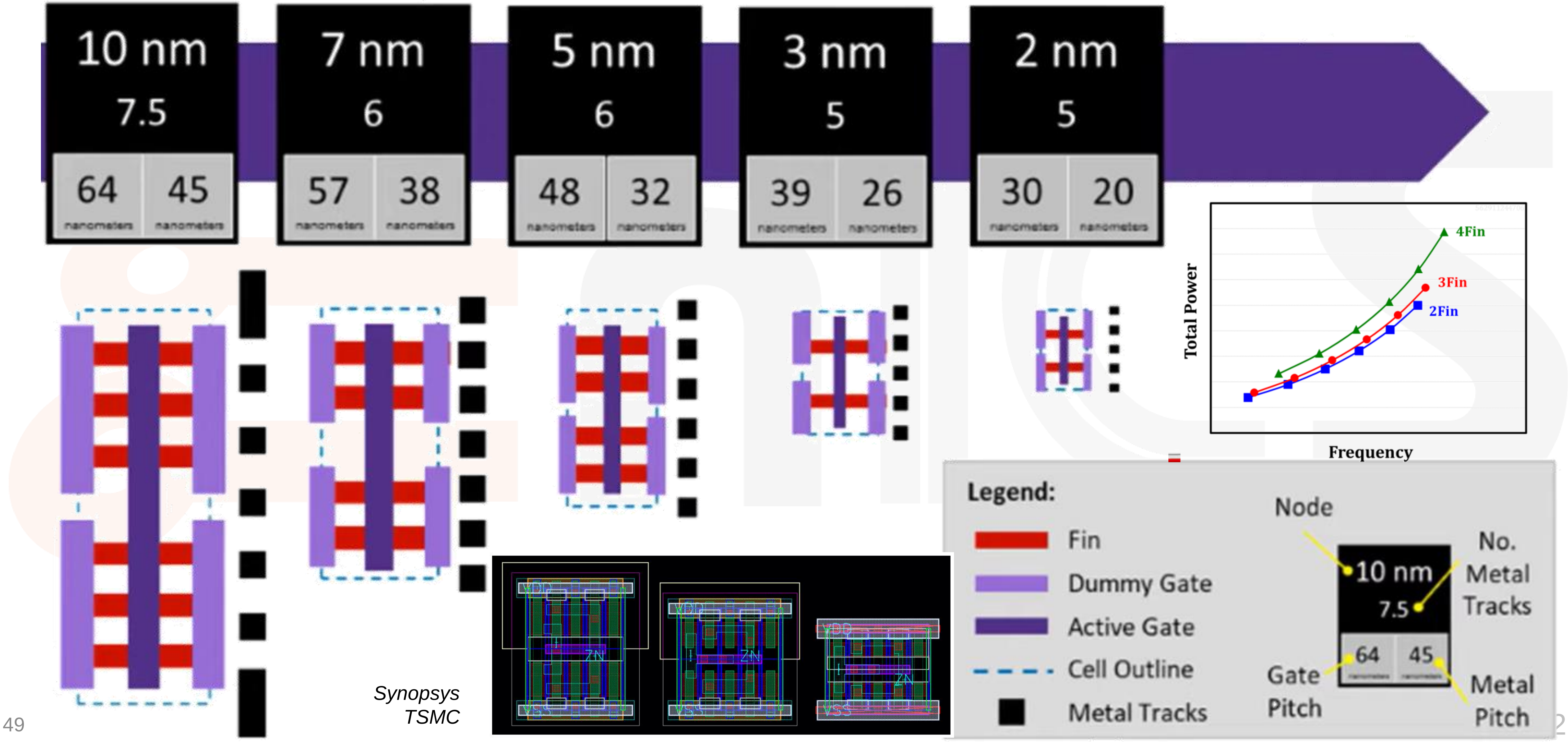
Squeezing Out What's Left in FinFETs

- **Really tough after 4 FinFET generations**
  - Realistically, never been any low hanging fruit with each new node
  - Process innovations & complexity for only incremental gain
  - +5% ring oscillator frequency is a big deal
- **Areas of development (no stones unturned)**
  - Short-channel control ➜ narrower fins, tradeoff vs. μ reduction
  - Channel mobility ➜ high-μ fin material, e.g., TSMC 5nm
  - EOT ➜ higher K, thinner & reliable gate dielectrics
  - Device variation ➜ fin uniformity & geometry control
  - Volumeless VT tuning using only HK dipoles
  - $R_{contact}$ ➜ contact resistivity (interface quality) & area
  - R_{gate} ➜ selective bottom-up HKMG deposition
  - C_{GS} & C_{GD} ➜ gate spacer K, air gap spacers
  - MEOL & BEOL resistance ➜ metal resistivity, R_{via}



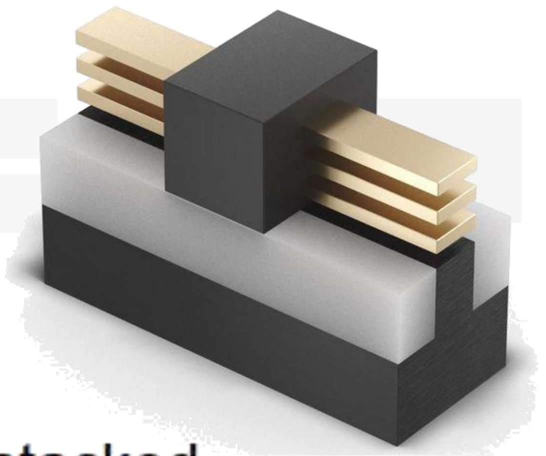
Yeap et al., TSMC
Cai, TSMC

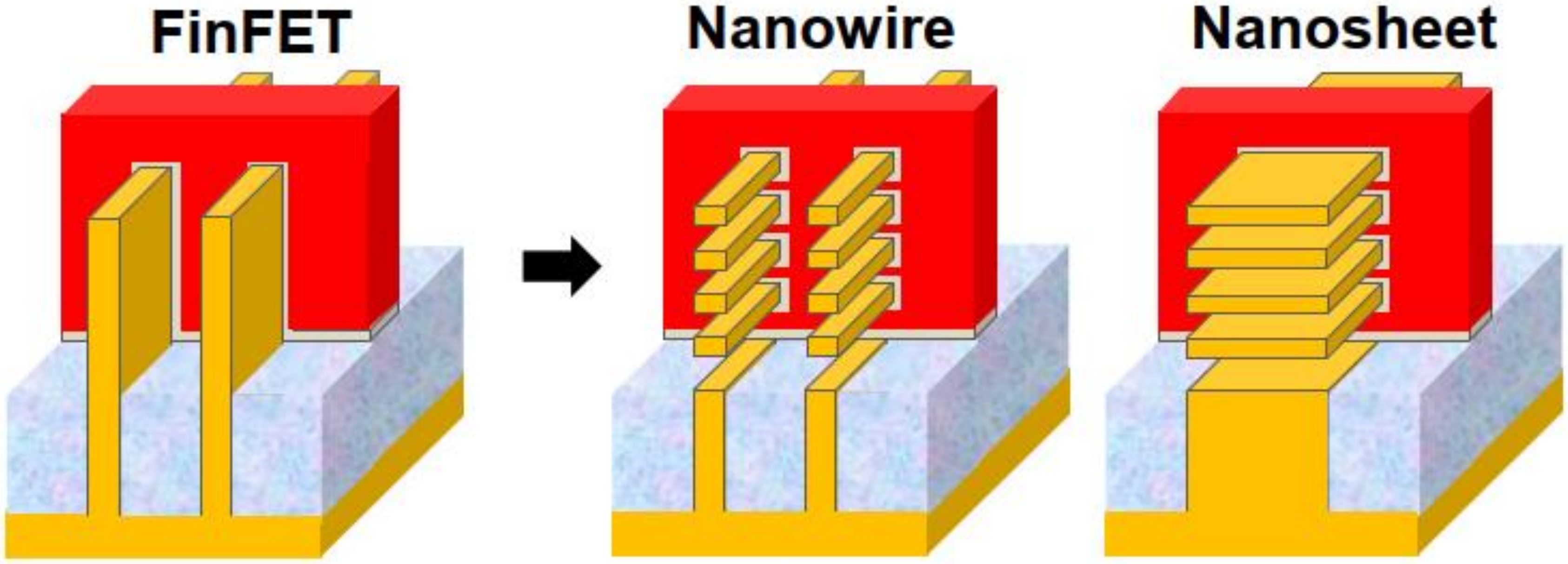
Vertical Scaling and Fin Depopulation



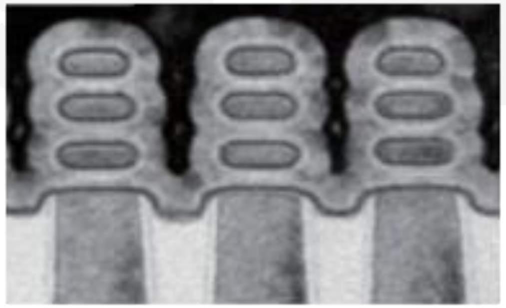
Migrating to Gate-All-Around (GAA)

- FinFET has poor short-channel control with further L_{gate} scaling
  - Need better short-channel control & more W_{eff} per die area
- Stacked GAA nanowires & nanosheets are promising
  - Nanowires offer better SCE, nanosheets offer better area scaling





stacked nanosheet demonstration

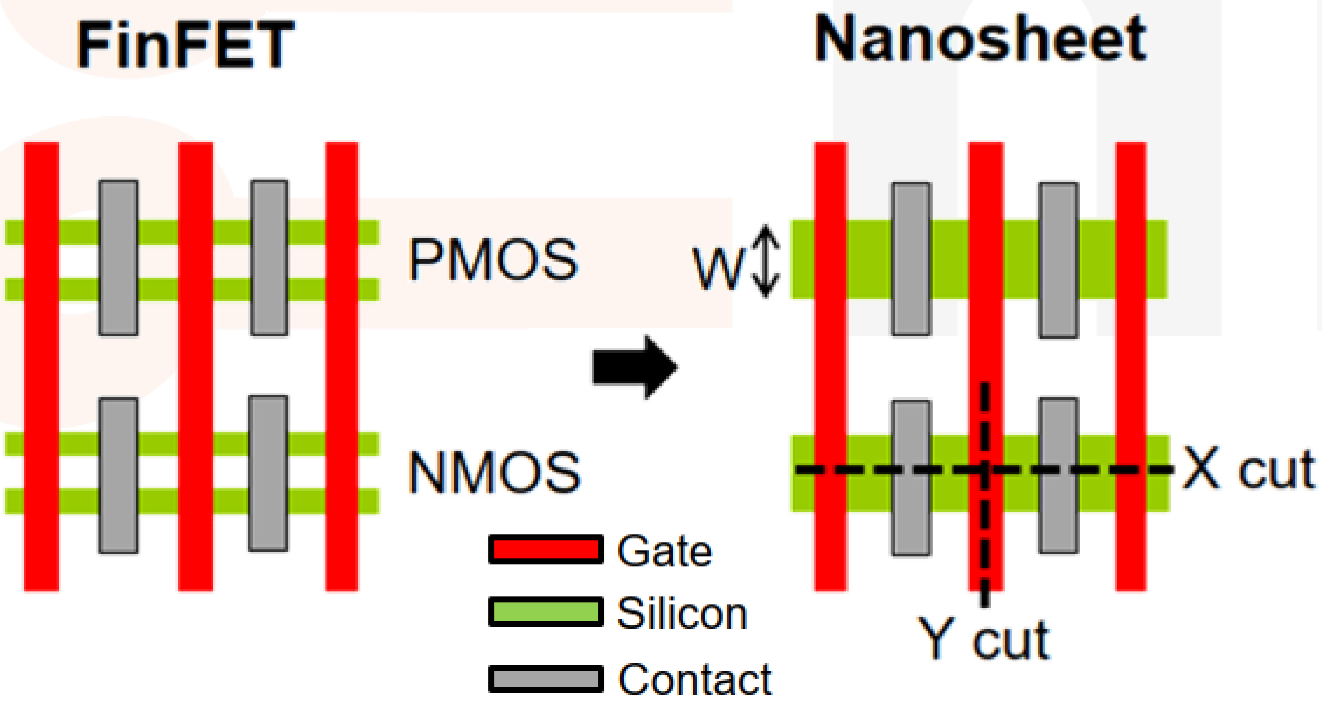


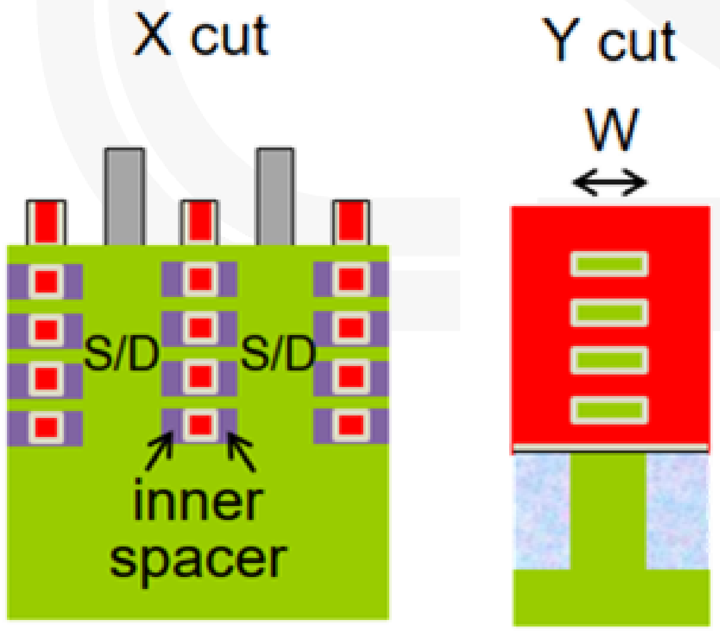
FinFET → Nanosheet

- **Current mostly flows along (100) surface**
  - Electrons ☺, Holes ☹→NMOS again stronger
- **Need inner spacers to reduce gate-to-S/D capacitance**
- **Disable parasitic planar FET below nanosheets**

Gate-All-Around Transistor Deployment			
AnandTech	Name	Process	Timeframe
Intel	RibbonFET	20A	2024
		18A	2025
TSMC	GAAFET	N2 / 2nm	EoY 2023?
Samsung	MBCFET	3GAE	2022
		3GAP	2023

https://technewsrooms.com/

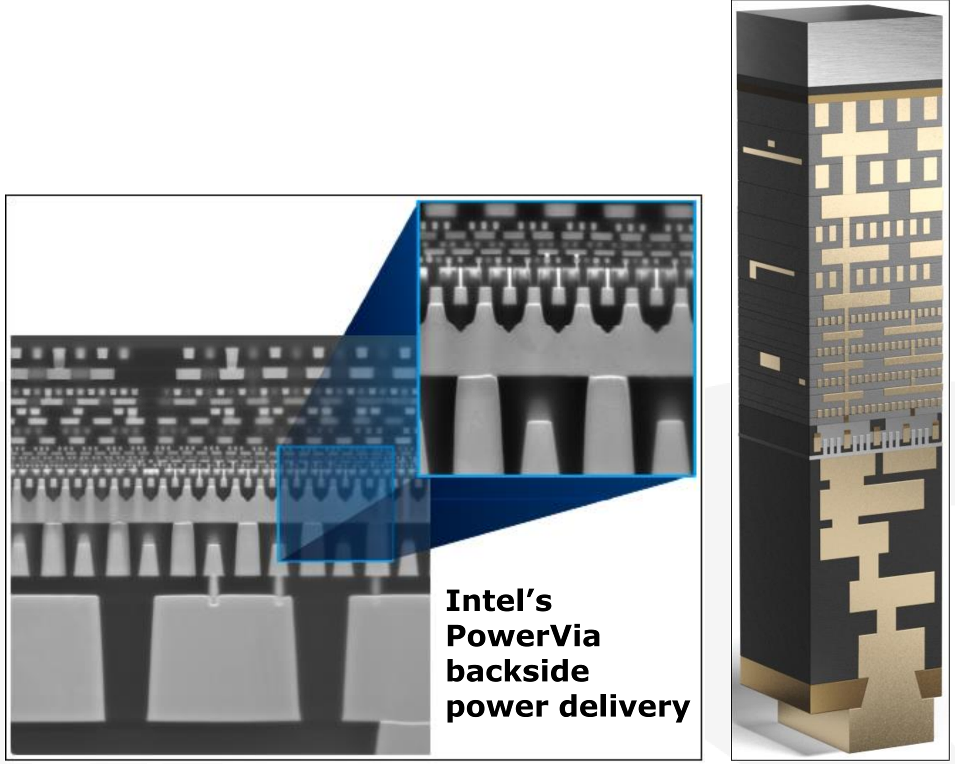


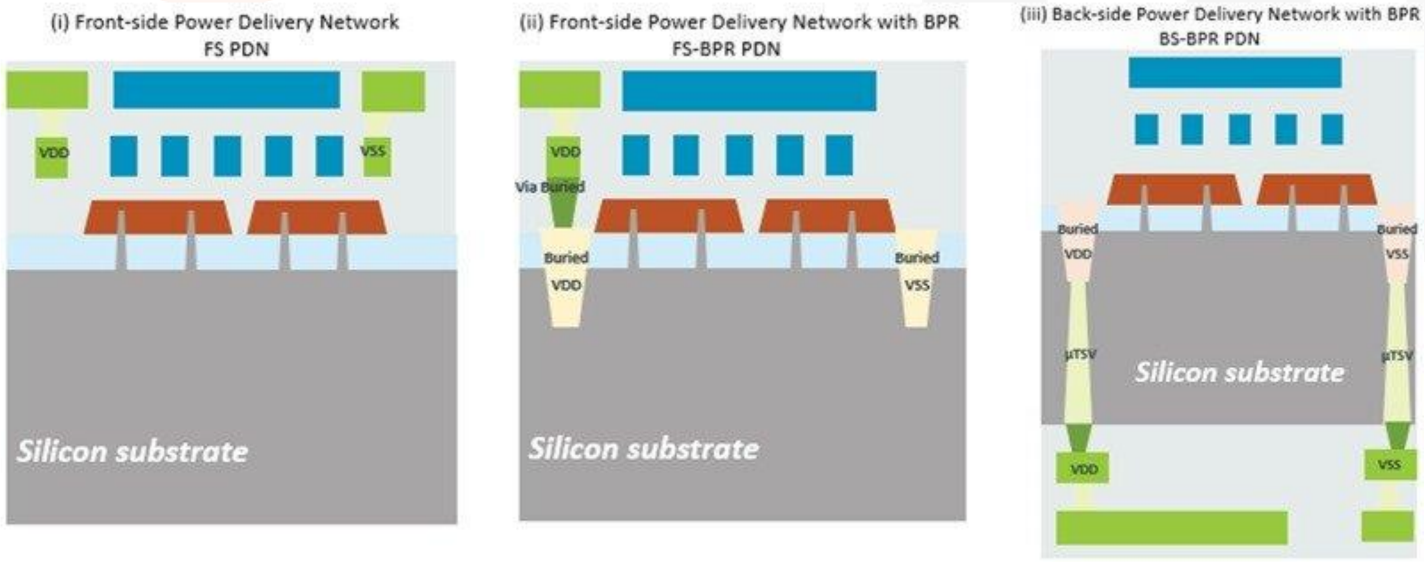


Loke, BCICTS
Cai, TSMC

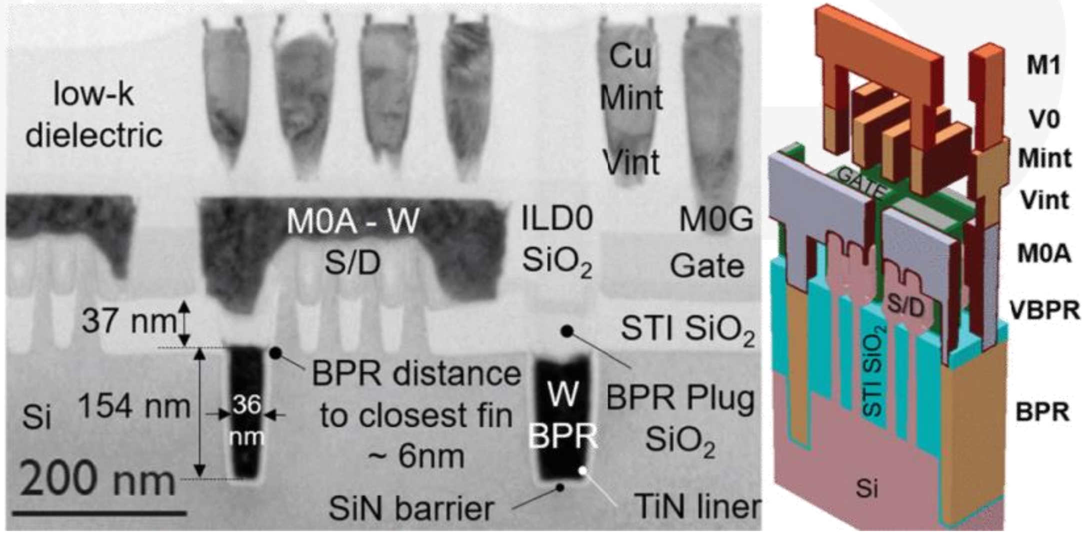
Backside power delivery

- **Burying the power rails *under* the transistors**
  - Power rails do not take up routing resources
  - →Reduced design area
  - →Improved IR drop and voltage droops
  - →Reduced power.





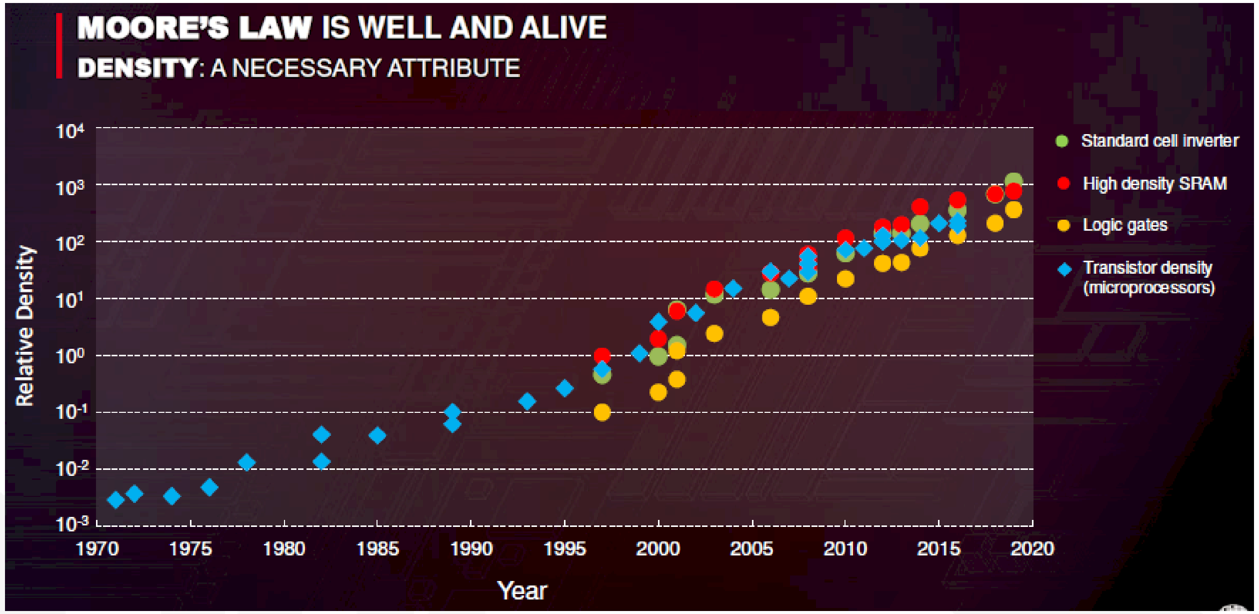
Prasad, et al., IEDM 2019



Gupta, et al., IEDM 2020

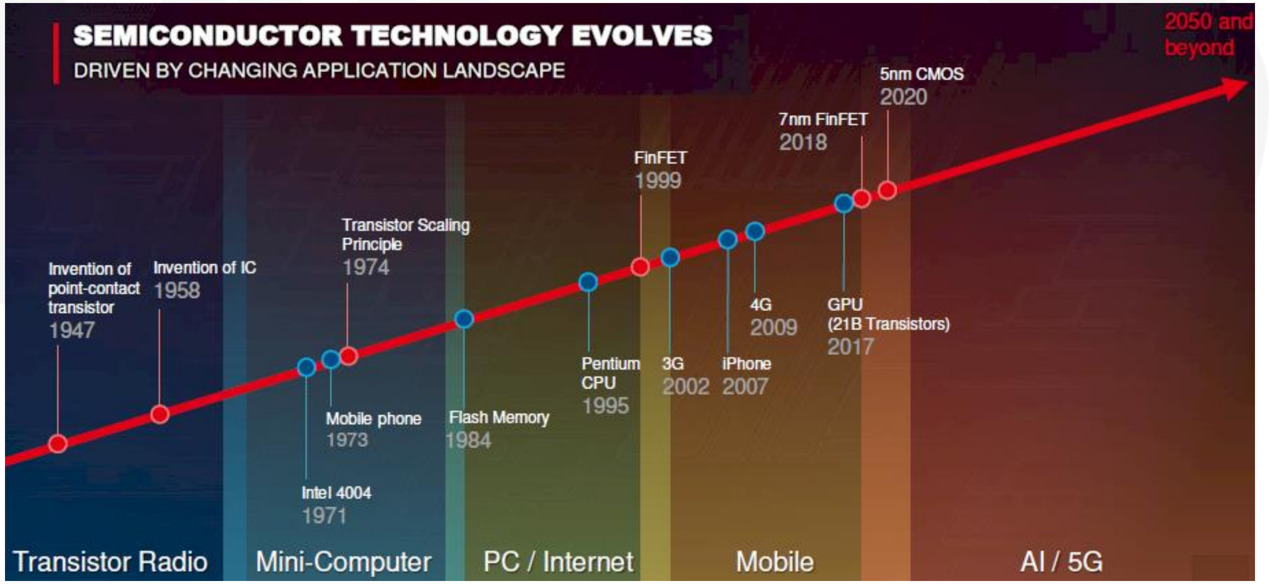
Conclusions

- "Moore's Law is well and alive.
  It's not slowing down.
  It's not even sick"
  Phillip Wong, TSCM, HotChips 2019



- SoC area scaling now driven primarily by **device innovation** & **DTCO**, less by feature size reduction.

- Understand & exploit technology for maximum PPA benefit & efficient design productivity



Wong, TSMC, HotChips, 2019

Main References

- Alvin Loke, et al., “Nanoscale FinFET Technology for Circuit Designers”, 2019 CICC Education Sessions, 2020 BCICTS, 2021 MTT-SCV
- Or Nahum, “FinFET Process Overview”, 2021
- www.halbleiter.org “Semiconductor Technology from A to Z”
- Jacob, et al., “Scaling Challenges for Advanced CMOS Devices” Globalfoundries