2005 International Conference on Characterization and Metrology for ULSI Technology
Dallas, TX
March 15, 2005

MOSFET Scaling Trends, Challenges, and Potential Solutions Through the End of the Roadmap: A 2005 Perspective

Peter M. Zeitzoff
Howard R. Huff

SEMATECH

Accelerating the next technology revolution.

Outline

- Introduction
- MOSFET scaling and its impact
- Front-end approaches and solutions
- Non-classical CMOS
- Summary

Introduction

- IC Logic technology: following Moore's Law by rapidly scaling into deep submicron regime
  - Increased speed and function density
  - Lower power dissipation and cost per function
- But the scaling results in major MOSFET and process integration issues, including
  - Simultaneously maintaining satisfactory I_{on} (drive current) and I_{leak}
  - High gate leakage current for very thin gate dielectrics
  - Control of short channel effects for very small transistors
  - Etc.
- Potential solutions & approaches
  - Material and process (front end): high-k gate dielectric, metal gate electrodes, strained Si, …
  - Structural: non-classical CMOS device structures
- This talk gives an updated perspective from the 2003 International Technology Roadmap for Semiconductors (ITRS)

Metrology and Characterization Issues

- Dimensional scaling: meeting metrology requirements for accuracy and precision becomes increasingly challenging
  - Example: electrical and physical measurement of Tox < 1.2 nm
  - Another example: CD measurement
    - Line edge (and width) roughness is increasingly critical
- Potential solutions (high-k, metal gate electrodes, strained Si, non-classical CMOS) raise significant metrology and characterization challenges

Key Overall Chip Parameters for High-Performance Logic, from 2003 ITRS

Year of Production	2003	2004	2005	2006	2007	2008	2009	2010	2012	2013	2015	2016	2018
MPU Physical Gate Length (nm)	45	37	32	28	25	22	20	18	14	13	10	9	7
On-chip local clock	2,976	4,171	5,204	6,783	9,285	10,972	12,369	15,079	20,065	22,980	33,403	39,683	53,207
Allowable Maximum Power													
High-performance with heatsink (W)	149	158	167	180	189	200	210	218	240	251	270	288	300
Cost-performance (W)	80	84	91	98	104	109	114	120	131	138	148	158	168
Functions per chip at production (million transistors [Mtransistors])	153	193	243	307	386	487	614	773	1,227	1,546	2,454	3,092	4,908

– Rapid scaling of L_g is driven by need to improve transistor speed

- Clock frequency, functions per chip (density) scale rapidly, but allowable power dissipation rises slowly with scaling

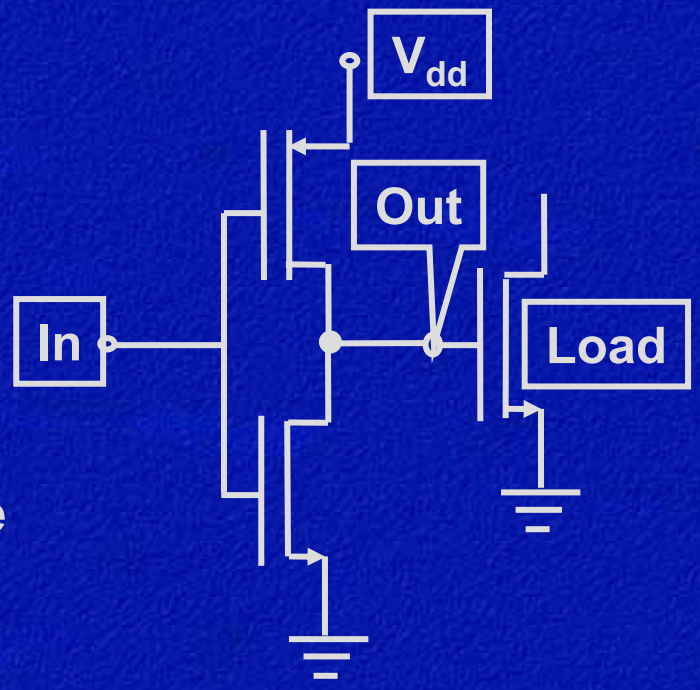
Outline

- Introduction
- ➢ MOSFET scaling and its impact
- Front-end approaches and solutions
- Non-classical CMOS
- Summary

Device Scaling Approach: 2003 ITRS

- Simple models capturing essential MOSFET physics→embedded in a spreadsheet
  - Room T, nominal devices assumed
  - Key parameters include: L_g, T_{ox}, V_{dd}, V_t, series parasitic resistance, drive current, leakage current, gate capacitance, subthreshold slope, etc.
- Using spreadsheet, MOSFET parameters are iteratively varied to meet ITRS targets for either
  - Scaling of transistor speed OR
  - Scaling for specific, low levels of leakage current

MOSFET Intrinsic Performance Parameter

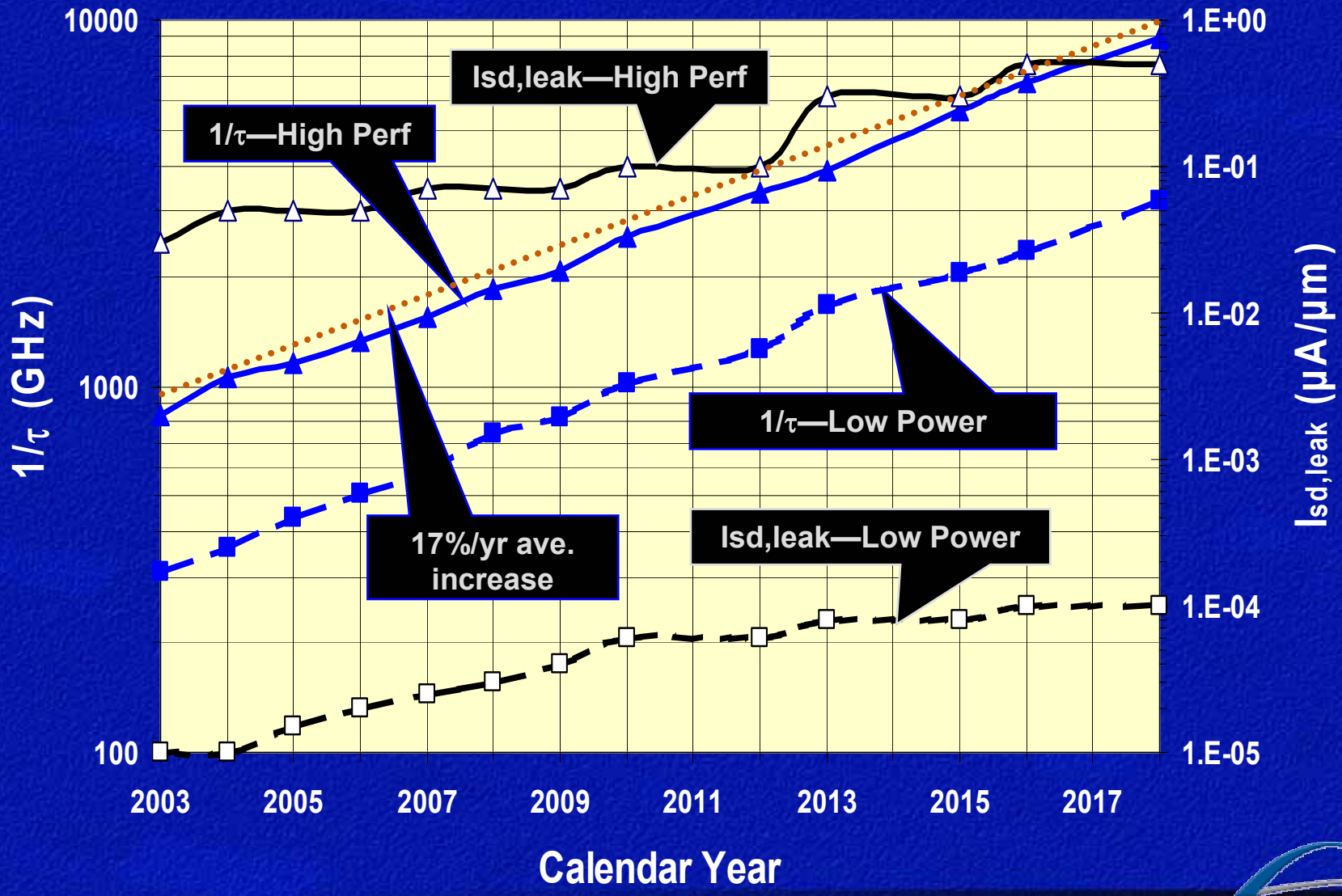


- Transistor intrinsic delay, τ
  - $\tau \sim C V_{dd}/(I_{on})$
    - I_{on} units: μA/μm
    - $C \sim C_L$
      - Gate dominated case: appropriate for local, dense logic
      - $C \sim C_L = C_{gate} \sim C_{ox} * L_g + C_{parasitic}$
      - $Cox \sim \varepsilon_{ox}/T_{ox}$
      - τ is the delay for a load consisting of one transistor's gate capacitance; shortest logic delay possible
- Transistor intrinsic switching frequency = $1/\tau$
  - Good metric for transistor performance
  - To maximize $1/\tau$, maximize I_{on}

Different Applications→Different ITRS Drivers

- **High-performance chips** (MPU, for example)
  - Driver: maximize chip speed→maximize transistor performance
    - Goal of ITRS scaling: $1/\tau$ increases at ~ 17% per year, historical rate
      - Must maximize I_{on}
      - Consequently, I_{leak} is relatively high
- **Low-power chips** (mobile applications)
  - Driver: minimize chip power (to maximize battery life) →minimize I_{leak}
    - Goal of ITRS scaling: specific, low level of I_{leak}
  - Consequently, $1/\tau$ is considerably less than for high-performance logic

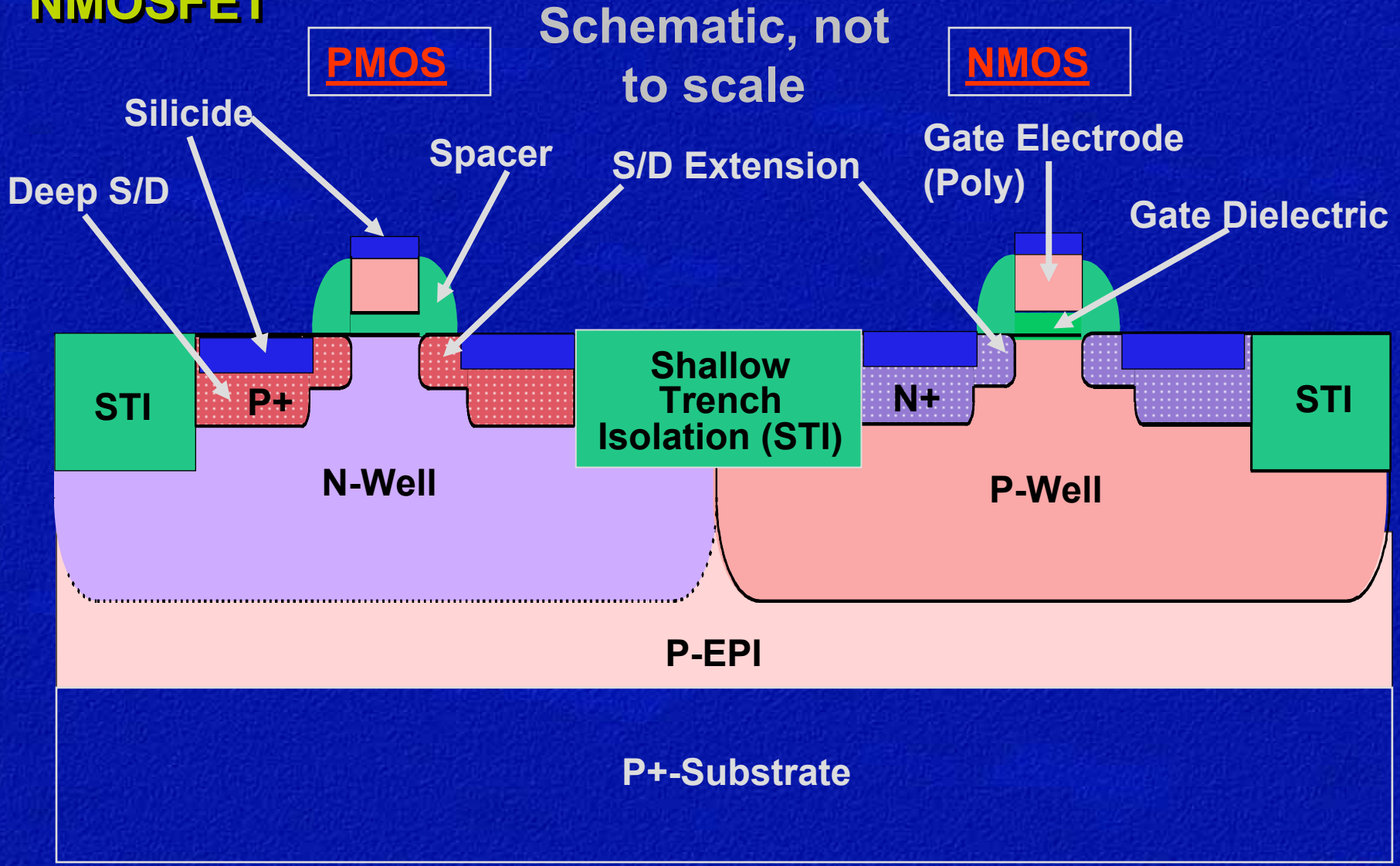
1/τ and Isd,leak scaling for High-Performance and Low-Power Logic. Data from 2003 ITRS.



Outline

- Introduction
- MOSFET scaling and its impact
- ➢ Front-end material and processing approaches and solutions
- Non-classical CMOS
- Summary

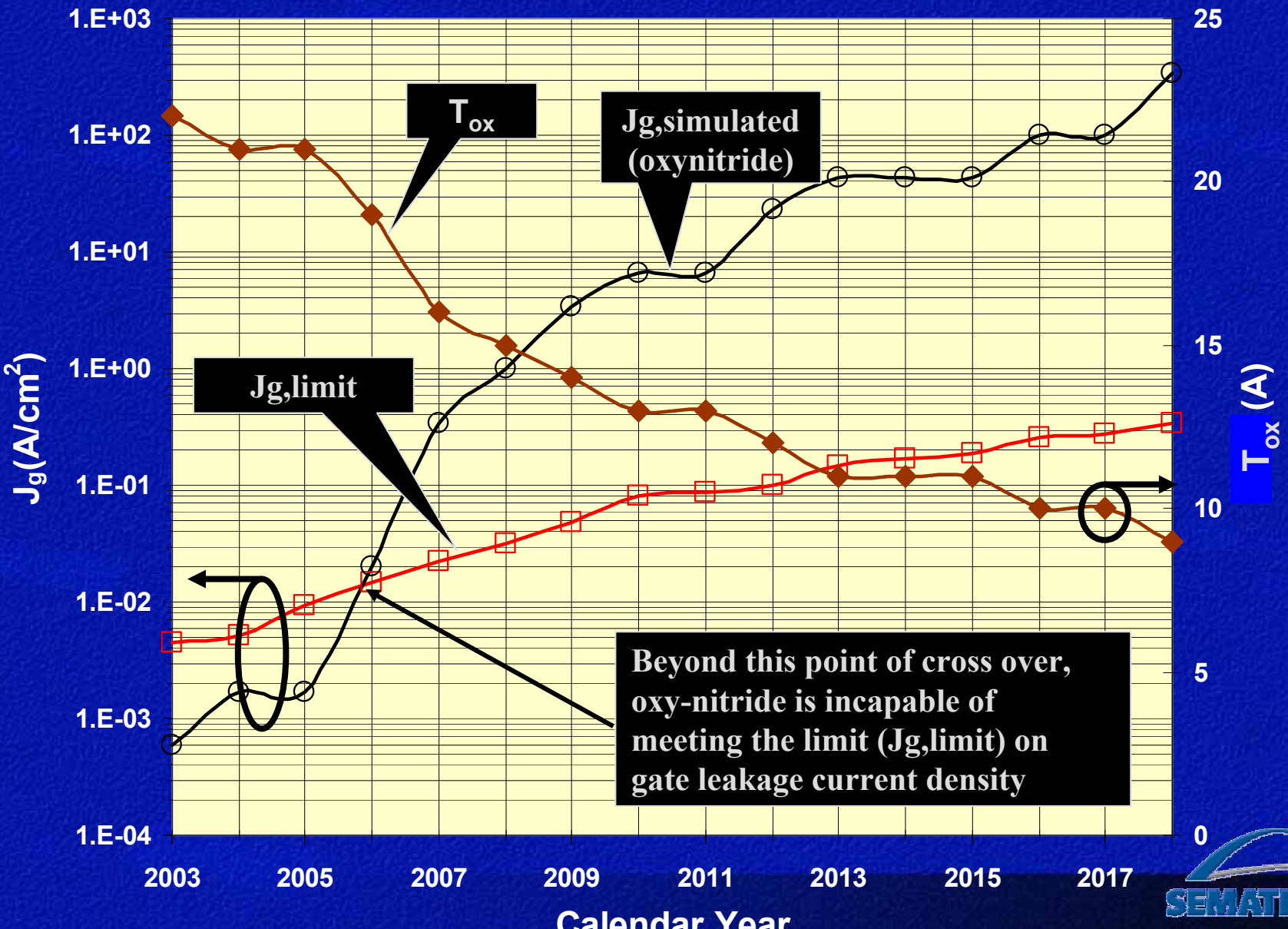
Simplified Cross Section of a Typical PMOSFET and NMOSFET



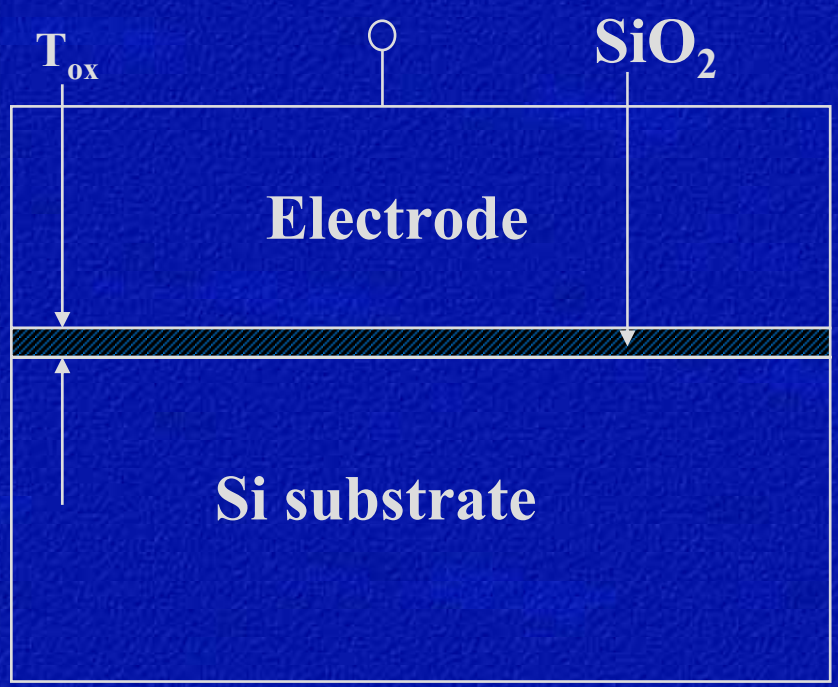
Difficult Transistor Scaling Issues

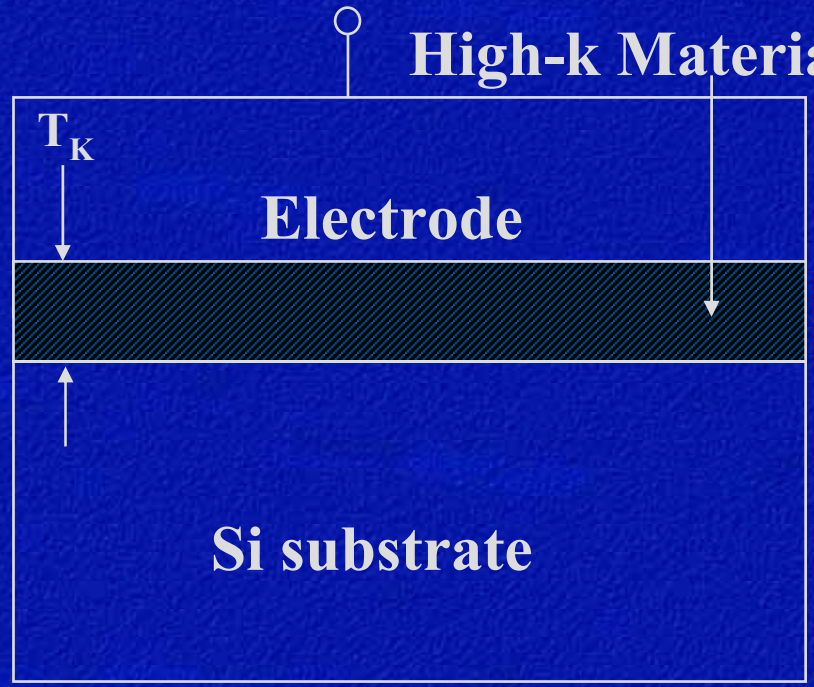
- Previously discussed scaling results involve determining the required transistor characteristics and performance to meet key scaling targets
  - Assumption: highly scaled MOSFETs with required characteristics can be successfully fabricated
- With scaling, increasing difficulty in meeting transistor requirements without significant technology innovations
  - High gate leakage
    - Direct tunneling increases rapidly as T_{ox} is reduced
    - Potential solution: high-k gate dielectric
  - Polysilicon depletion in gate electrode→increased effective T_{ox}, reduced I_{on}
  - Need for enhanced channel mobility
  - Etc.

For Low-Power Logic, Gate Leakage Current Density Limit Versus Simulated Gate Leakage due to Direct Tunneling. Data from 2003 ITRS.



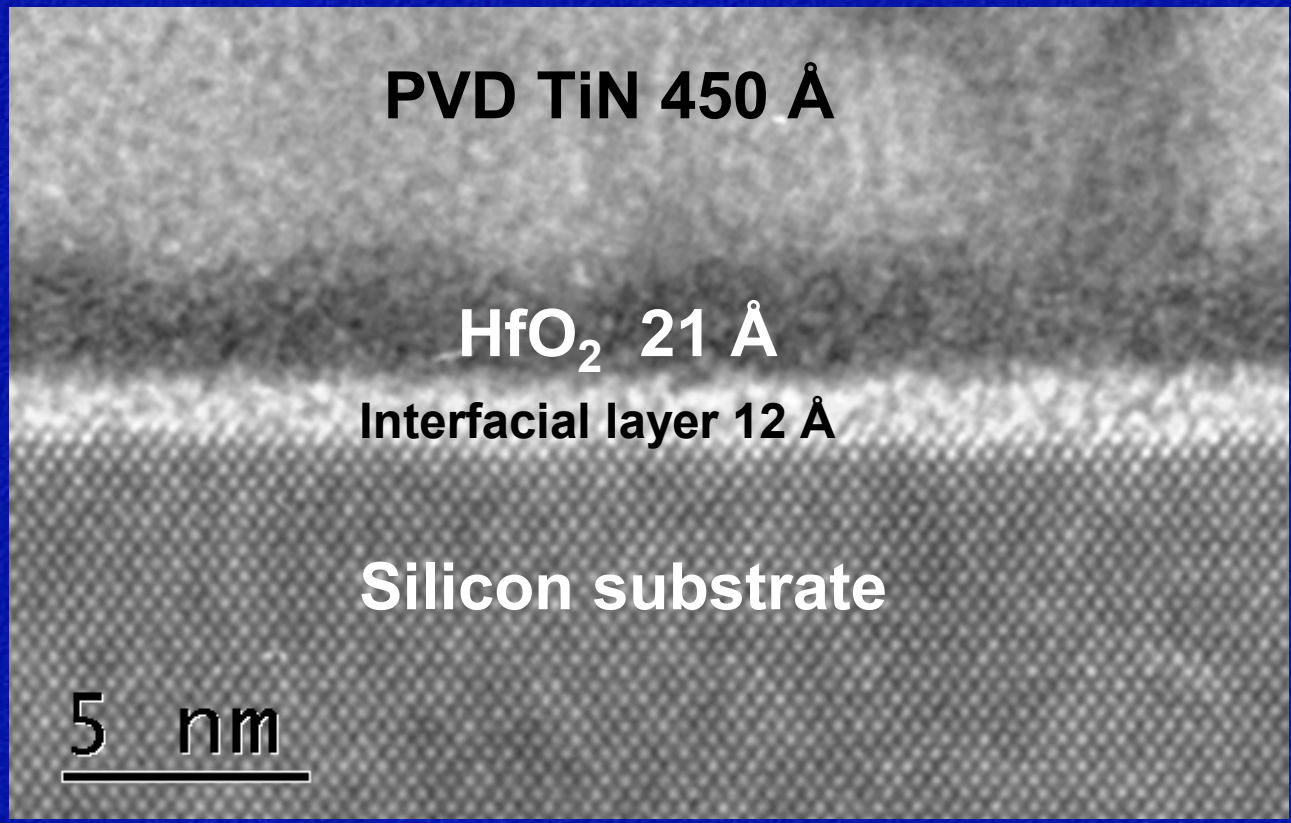
High K Gate Dielectric to Reduce Direct Tunneling





- Equivalent Oxide Thickness = EOT = T_{ox} = T_K * (3.9/K), where 3.9 is relative dielectric constant of SiO_2 and K is relative dielectric constant of high K material
  - C = Cox = ε_{ox}/T_{ox}
  - To first order, MOSFET characteristics with high-k are same as for SiO2
- Because $T_K > T_{ox}$, direct tunneling leakage much reduced with high K
  - If energy barrier is high enough
- Candidate materials: LaO_2/HfO_2/ ZrO_2(K~15 - 30); Hf, Zr-SiO_4 (K~12 - 16); others
  - Major materials, process, integration issues to solve

MOCVD HfO2 TEM (EOT = 0.95 nm) (HfO2 on HF-last, N2O-750°C Pre-Deposition Anneal)



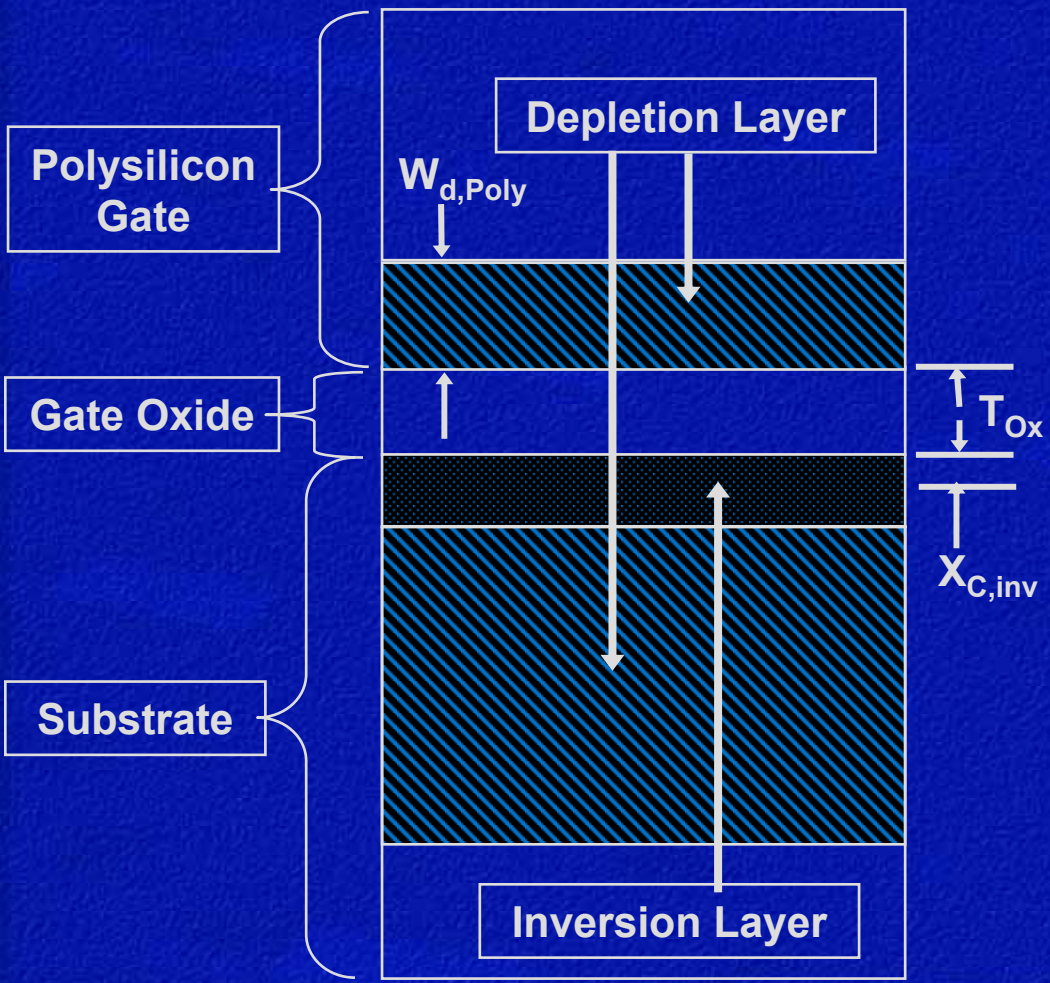
- Effective k for above dielectric stack ≈ 13.5
- k for interfacial layer could be significantly greater than SiO_2 indicating reaction or intermixing of HfO_2 film with interfacial SiO_2

Avinash Agarwal et al., Alternatives to SiO_2 as Gate Dielectrics for Future Si-Based Microelectronics, 2001 MRS Workshop Series (2001)

Difficult Transistor Scaling Issues

- With scaling, increasing difficulty in meeting transistor requirements without significant technology innovations
  - High gate leakage
    - Direct tunneling increases rapidly as T_{ox} is reduced
  - Polysilicon depletion in gate electrode→increased effective electrical T_{ox}, reduced I_{on}
    - Potential solution: metal gate electrodes
  - Need for enhanced channel mobility
  - Etc.

Polysilicon Depletion and Substrate Quantum Effects



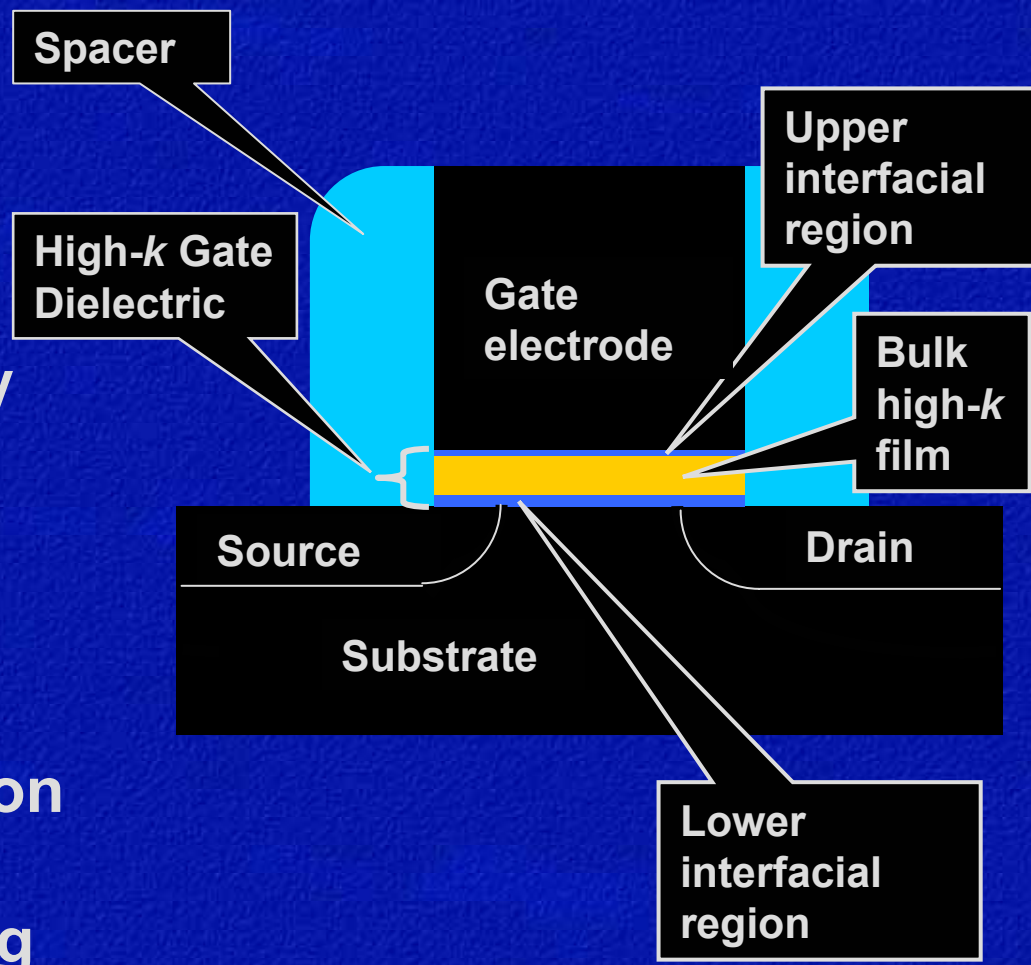
- $T_{ox,electric} = T_{ox} + (K_{ox}/K_{si})^* (W_{d,Poly} + X_{C,inv})$
  - $K_{ox} = 3.9$
  - $K_{si} = 11.9$
- $T_{ox,electric} = T_{ox} + (0.33)^* (W_{d,Poly} + X_{C,inv})$
  - $W_{d,Poly} \sim 1/(\text{poly doping})^{0.5}$ → increase poly doping to reduce $W_{d,Poly}$ with scaling
  - But max. poly doping is limited → can't reduce $W_{d,Poly}$ too much
- Fermi Level pinning with high-k
- Poly depletion and $X_{C,inv}$ become more critical with T_{ox} scaling
  - Eventually, poly will reach its limit of effectiveness

Metal Gate Electrodes

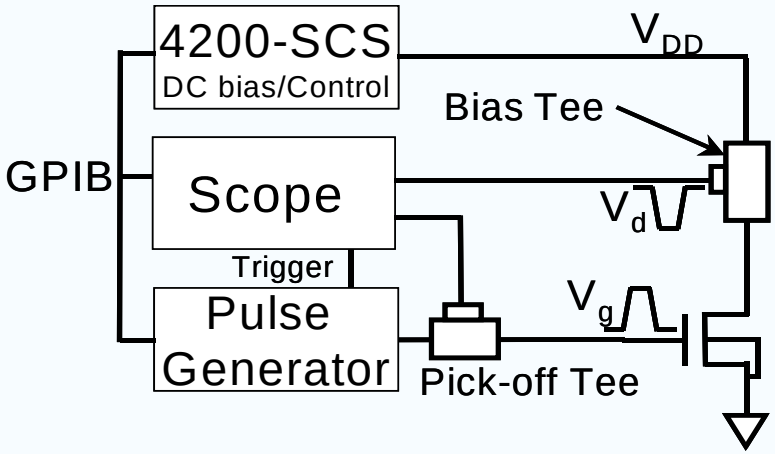
- Metal gate electrodes are a potential solution when poly "runs out of steam": probably implemented at 65 nm tech. generation (2007) or beyond
  - No depletion, very low resistance gate, no boron penetration, compatibility with high-k
  - Issues
    - Different work functions needed for PMOS and NMOS==>2 different metals may be needed
      - Process complexity, process integration problems, cost
    - Etching of metal electrodes
    - New materials: major challenge

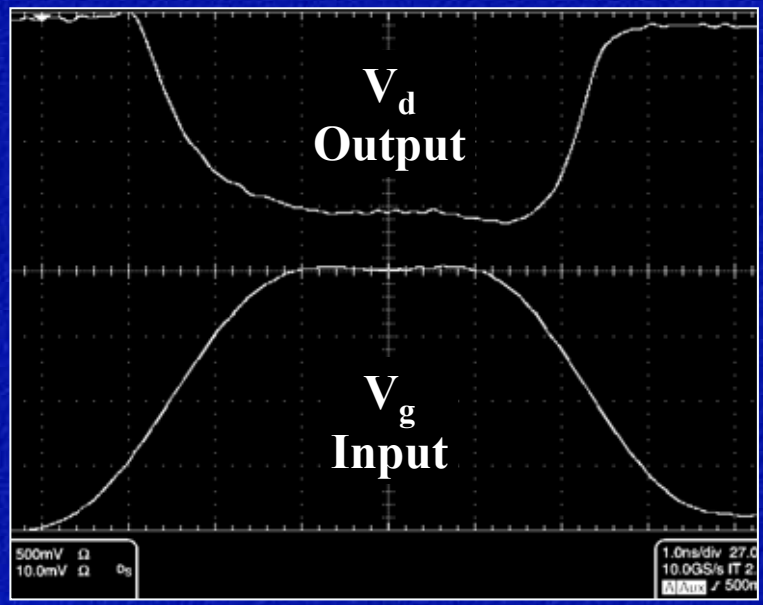
Advanced Gate Stack: Key Metrology and Characterization Challenges

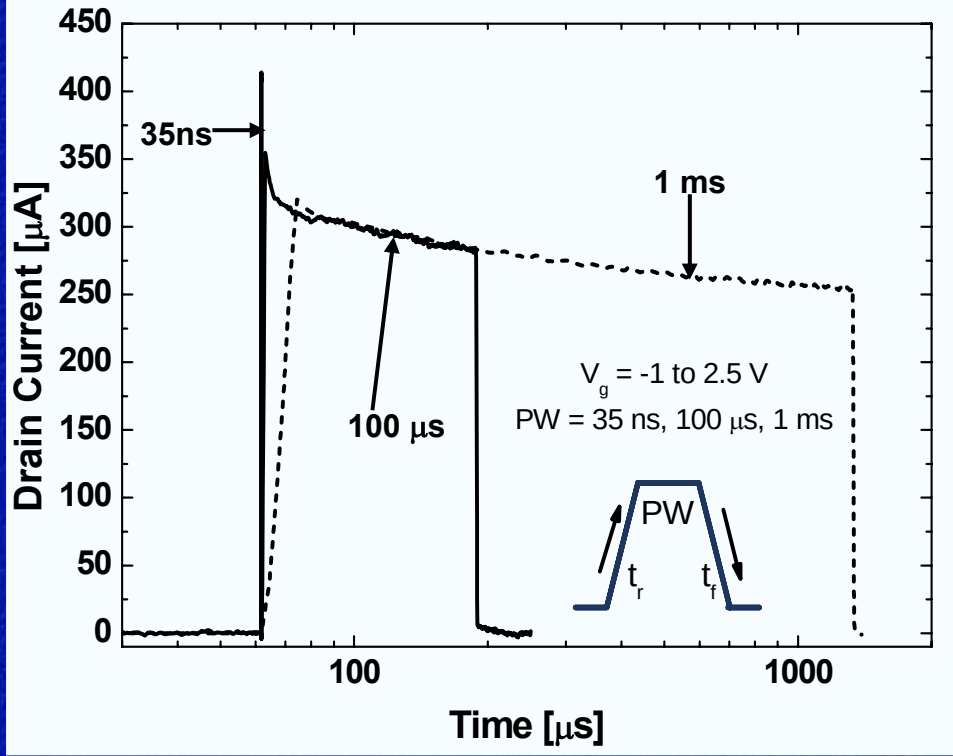
- Transient charge trapping in high-k bulk
  - Characterizing charge trapping
  - Extracting mobility: $\mu_{eff}=(LI_d)/(WQ_{inv}V_d)$
  - Determining V_t, V_{FB} from C-V
  - Fast pulse measurements help
- Charge in high-k & Interaction between metal gate and high-k: unambiguous determination of ϕ_m
- Gate leakage➔determining EOT from C-V



Example: Fast Transient Electron Trapping with Pulse Measurements on High-k Gate Dielectric







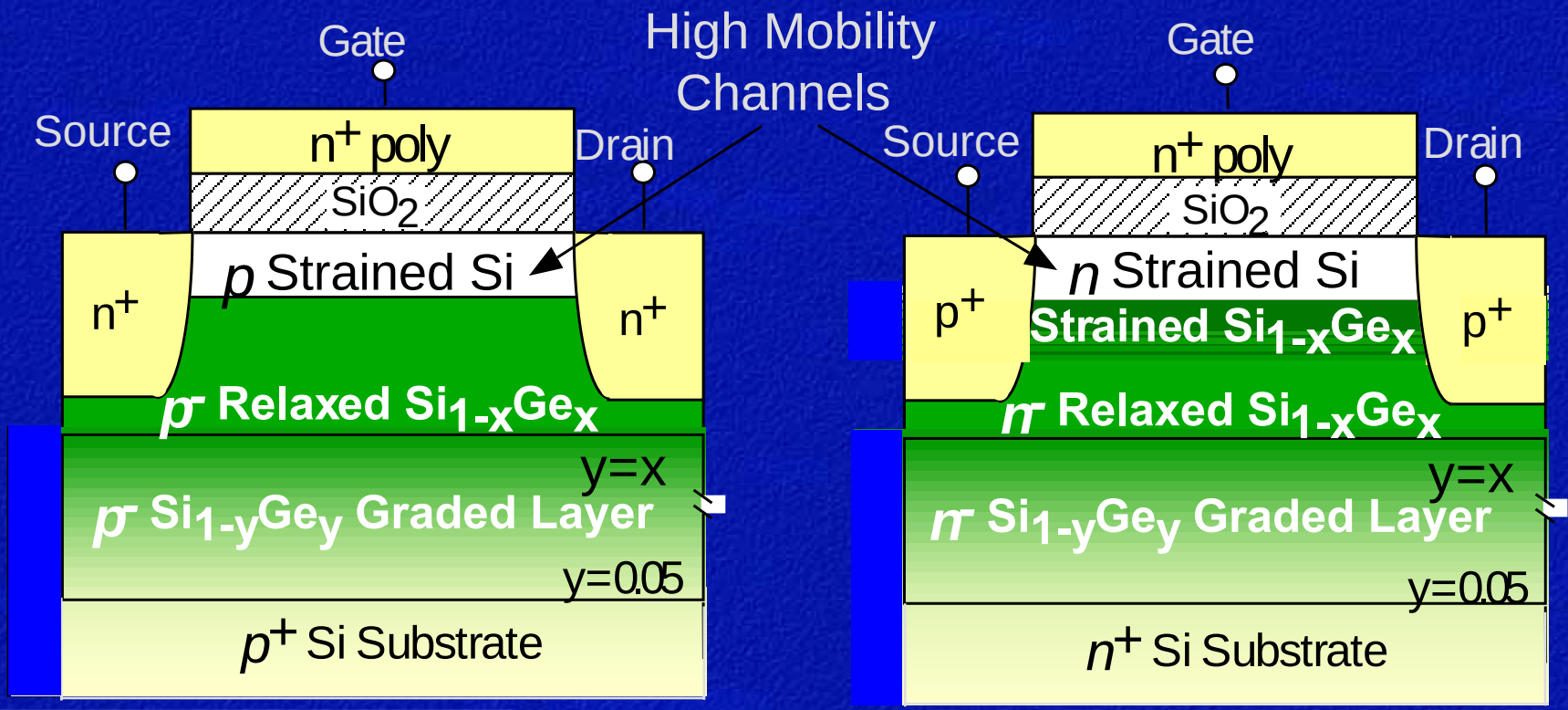
Significant trapping occurs within few μsec

C. Young, SSDM 2004

Difficult Transistor Scaling Issues

- With scaling, increasing difficulty in meeting transistor requirements
  - High gate leakage
    - Direct tunneling increases rapidly as T_{ox} is reduced
  - Polysilicon depletion in gate electrode → increased effective T_{ox}, reduced I_{on}
  - Need for enhanced channel mobility
    - Potential solution: strained Si channels
  - Etc.

Band Engineered MOSFETs: Strained MOSFET Structures



(J. Welser, J.L. Hoyt, and J.F. Gibbons, IEDM, 1992, pp. 1000-1003.)

Courtesy of J. Hoyt - MIT

(K. Rim, J. Welser, S. Takagi, J.L. Hoyt, and J.F. Gibbons, IEDM, 1995, pp. 517-520.)

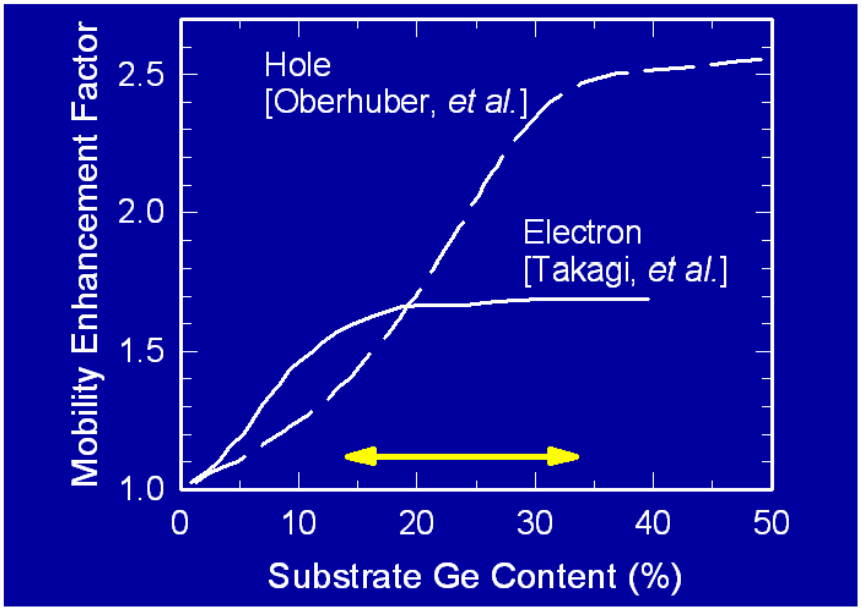
+ **Increased effective mobility, increased I_{on}**

- **Difficult integration issues: manufacturability, thermal stability, simultaneous optimization of both PMOS and NMOS, defects, leakage**
- **Compatibility with ultra-thin body SOI**
- **Cost**

Strained Si Device Structures

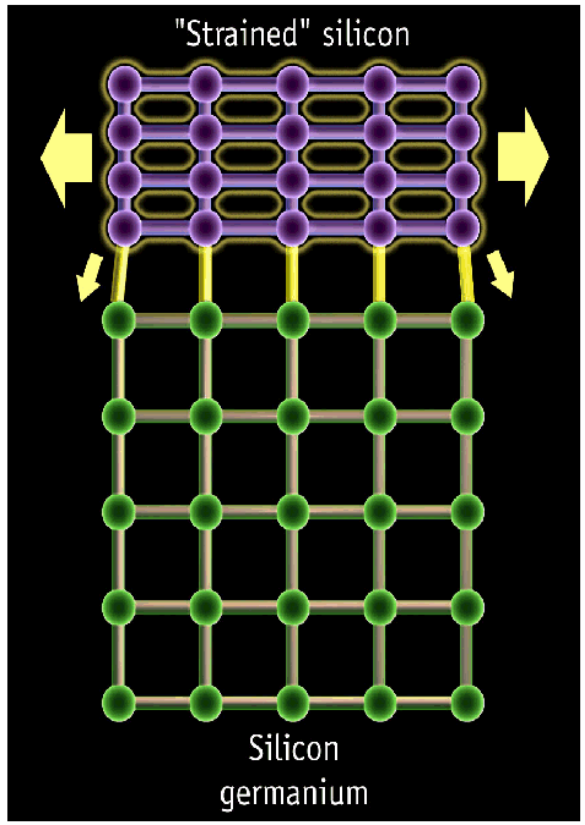
Courtesy of Patricia Mooney (IBM Corp.) From P. M. Mooney et al., presented at the American Physics Society Meeting, Austin, TX, March 3-7, 2003.

modified band structure of Si under biaxial tensile strain ==> enhanced mobility



need relaxed $Si_{1-x}Ge_x$ with $0.15<x<0.35$

Strained Si on SiGe



Alternate Approach: Uniaxial Process Induced Stress

NMOS: uniaxial tensile stress from stressed SiN film

PMOS: uniaxial compressive stress from sel. SiGe in S/D

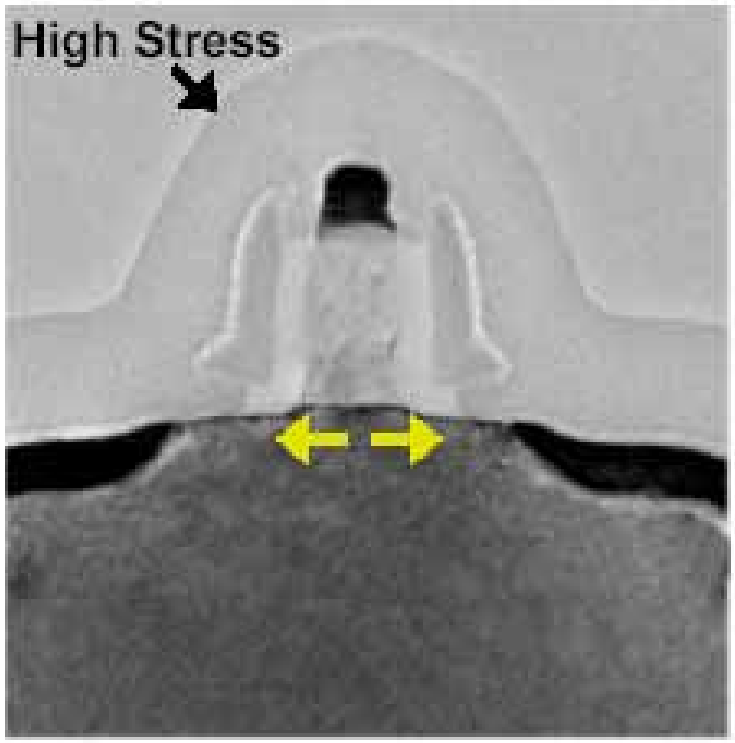


Fig. 3 TEM of NMOS transistor showing high tensile stress nitride overlayer.

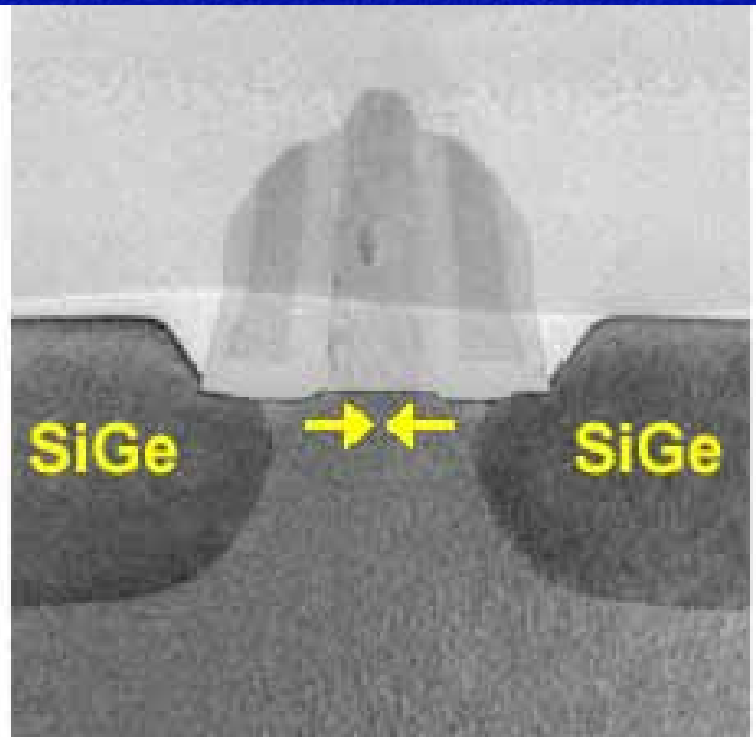


Fig. 4 TEM of PMOS showing SiGe heteroepitaxial S/D inducing uniaxial strain.

From K. Mistry et al., "Delaying Forever: Uniaxial Strained Silicon Transistors in a 90nm CMOS Technology," *2004 VLSI Technology Symposium*, pp. 50-51.

Strained Si: Metrology and Characterization Challenges

- Measuring strain distribution with high spatial resolution in deep sub-micron structures
  - Possible approaches
    - X-ray diffraction (XRD)
    - Raman spectroscopy
    - Convergence Beam Electron Diffraction (CBED)
    - Electron Diffraction Contrast (EDC)

Outline

- Introduction
- Scaling and its impact
- Front end approaches and solutions
  - Non-classical CMOS
- Summary

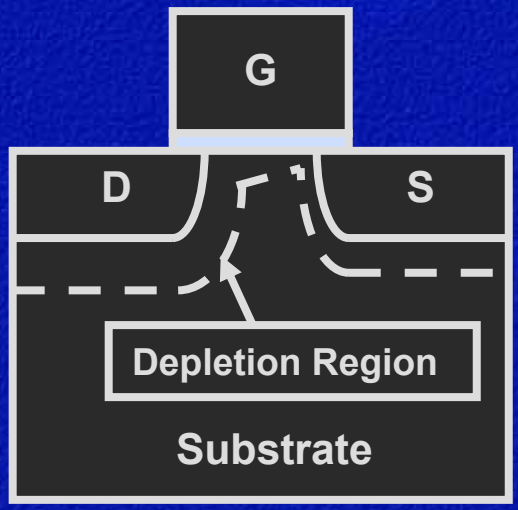
Limits of Scaling Planar, Bulk MOSFETs

- 65 nm tech. generation (2007, L_g = 25nm) and beyond: increased difficulty in meeting all device requirements with classical planar, bulk CMOS (even with material and process solutions: high K, metal electrodes, ….)
  - Control of SCE
  - Impact of quantum effects and statistical variation
  - Impact of high substrate doping
  - Control of series S/D resistance ($R_{series,s/d}$)
  - Others

- Alternative device structures (non-classical CMOS) may be utilized
  - Ultra thin body, fully depleted: single-gate SOI and multiple-gate transistors

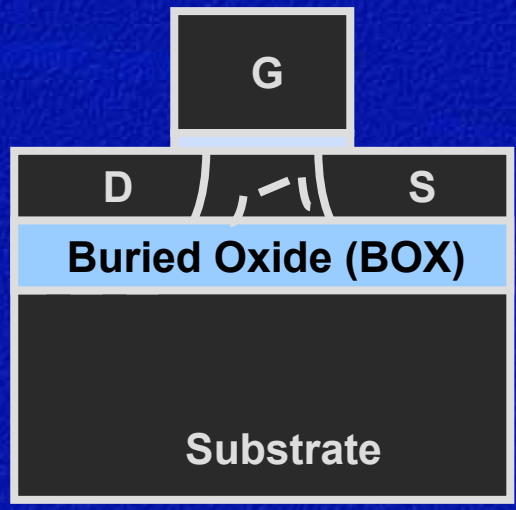
Transistor Structures

Planar Bulk



+ Current solution
+ Wafer cost / availability
- SCE scaling difficult
- High doping effects and Statistical variation
- Parasitic junction capacitance

Partially Depleted SOI

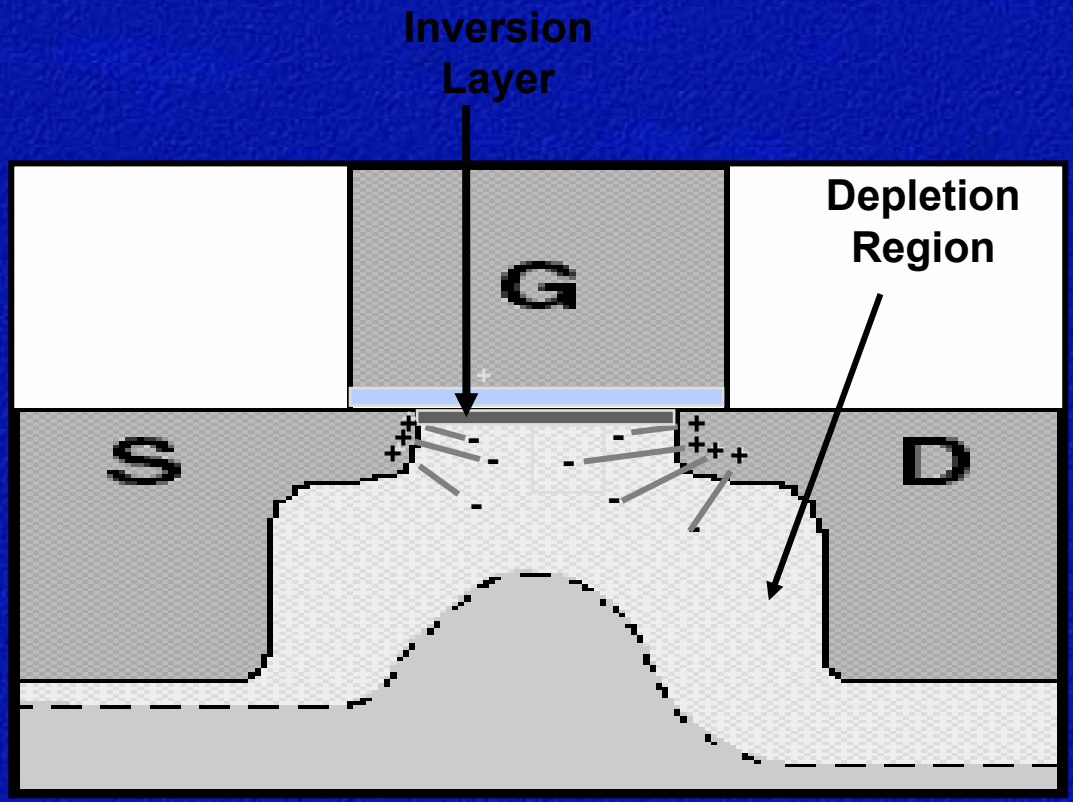


+ Lower junction cap
+ F.B. performance boost
- F.B. history effect
- SCE scaling difficult
- Wafer cost/availability

REFERNCES

1. *P.M. Zeitzoff, J.A. Hutchby and H.R. Huff, MOSFET and Front-End Process Integration: Scaling Trends, Challenges, and Potential Solutions Through The End of The Roadmap, International Journal of High-Speed Electronics and Systems,* ***12****, 267-293 (2002).*
2. *Mark Bohr, ECS Meeting PV* ***2001-2****, Spring, 2001.*

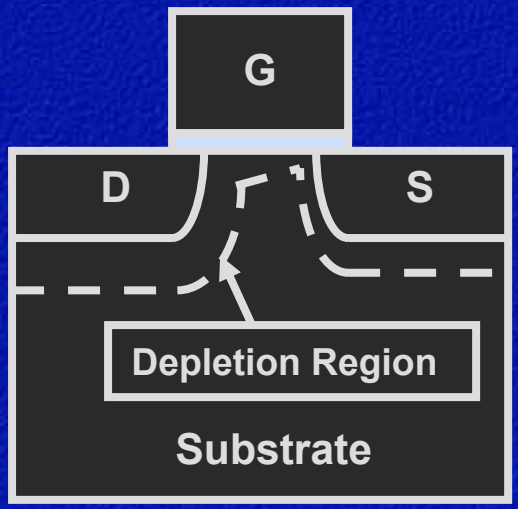
Schematic cross section of planar bulk, UTB SOI, and DG SOI MOSFET



Bulk MOSFET

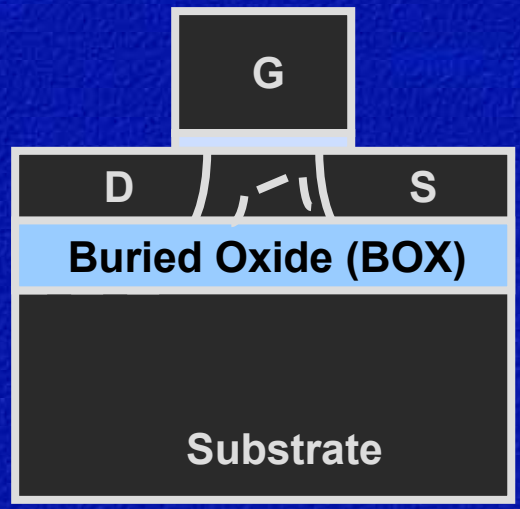
Transistor Structures

Planar Bulk



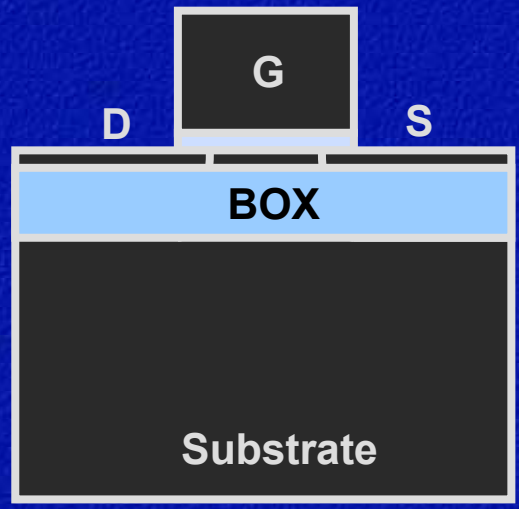
+ Current solution
+ Wafer cost / availability
- SCE scaling difficult
- High doping effects and Statistical variation
- Parasitic junction capacitance

Partially Depleted SOI



+ Lower junction cap
+ F.B. performance boost
- F.B. history effect
- SCE scaling difficult
- Wafer cost/availability

Fully Depleted SOI



+ Lower junction cap
+ Light doping possible
- SCE scaling difficult
- High $R_{series,s/d}$ → elevated S/D
- Sensitivity to Si thickness (very thin)
- Wafer cost/availability

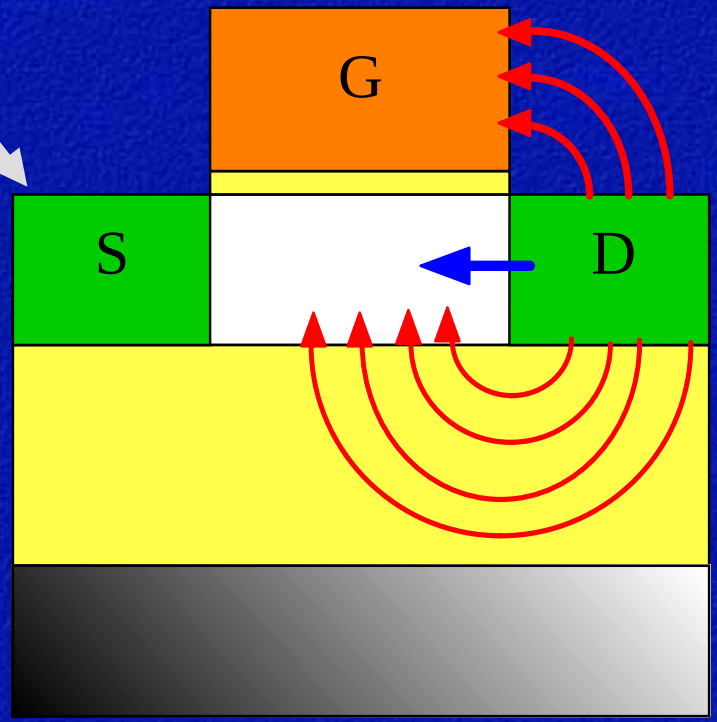
REFERNCES

1. *P.M. Zeitzoff, J.A. Hutchby and H.R. Huff, MOSFET and Front-End Process Integration: Scaling Trends, Challenges, and Potential Solutions Through The End of The Roadmap, International Journal of High-Speed Electronics and Systems,* ***12****, 267-293 (2002).*
2. *Mark Bohr, ECS Meeting PV* ***2001-2****, Spring, 2001.*

Field Lines for Single and Double-Gate MOSFETs

E-Field lines

To reduce SCE's, aggressively reduce Si layer thickness



Single-Gate SOI

Courtesy: Prof. J-P Colinge, UC-Davis

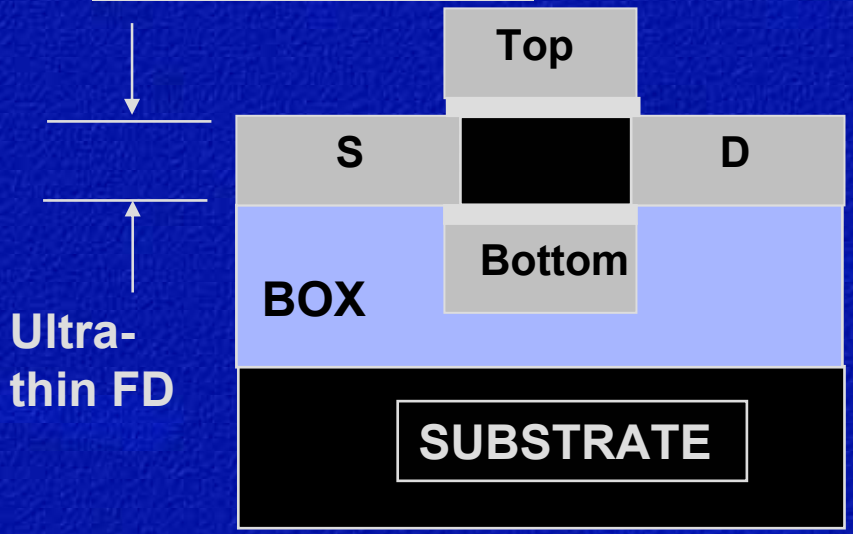
Double Gate Transistors

REFERENCES

1. P.M. Zeitzoff, J.A. Hutchby and H.R. Huff, MOSFET and Front-End Process Integration: Scaling Trends, Challenges, and Potential Solutions Through The End of The Roadmap, International Journal of High-Speed Electronics and Systems, ***12****, 267-293 (2002).*

2. Mark Bohr, ECS Meeting PV **2001-2***, Spring, 2001.*

Double-Gate SOI:



+ ***Enhanced scalability***
+ **Lower junction capacitance**
+ **Light doping possible, with near-midgap metal gate**
+ **~2x drive current**
- **~2x gate capacitance**
- **High $R_{series,s/d}$ → raised S/D**
- **Complex process**

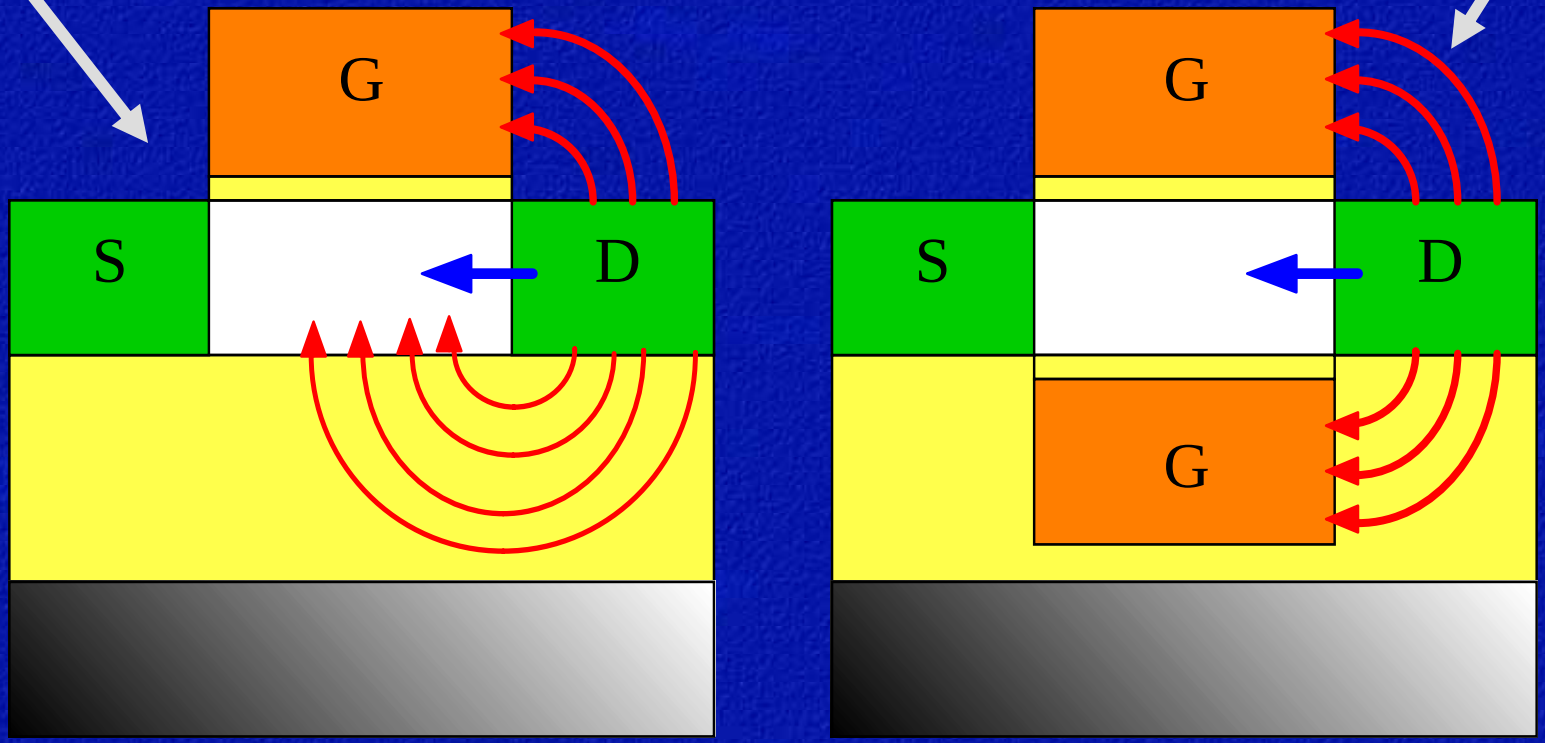
Summary: more advanced, optimal device structure, but difficult to fabricate, particularly in this SOI configuration

Field Lines for Single and Double-Gate MOSFETs

E-Field lines

To reduce SCE's, aggressively reduce Si layer thickness

Double gates electrically shield the channel

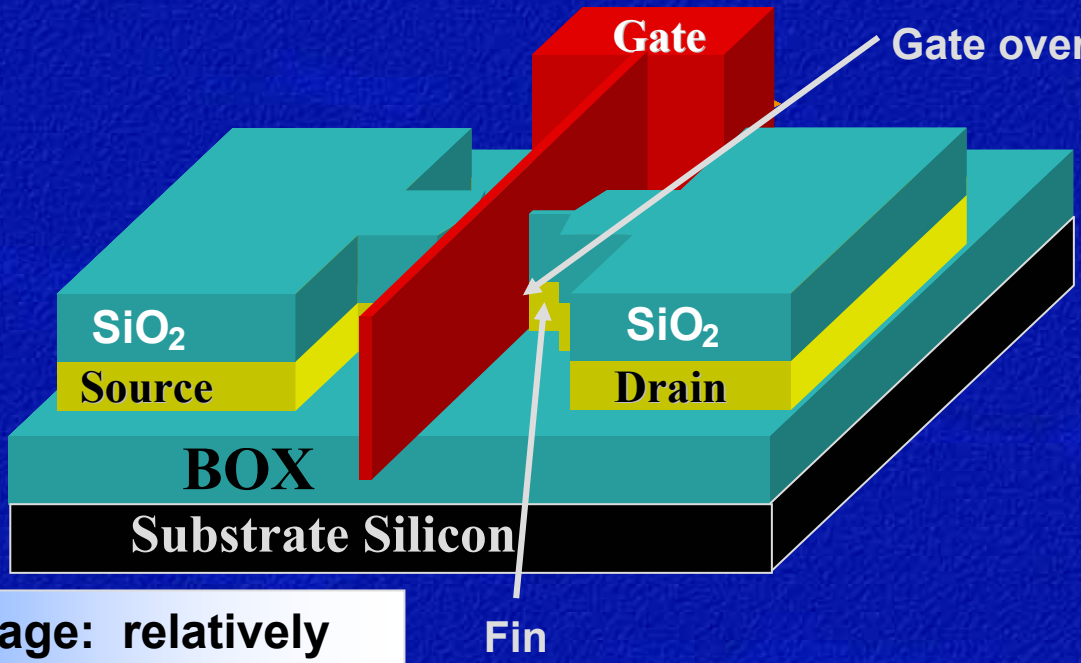


Single-Gate SOI

Double-Gate

Courtesy: Prof. J-P Colinge, UC-Davis

Other Double-Gate Transistor Structures (FinFET)

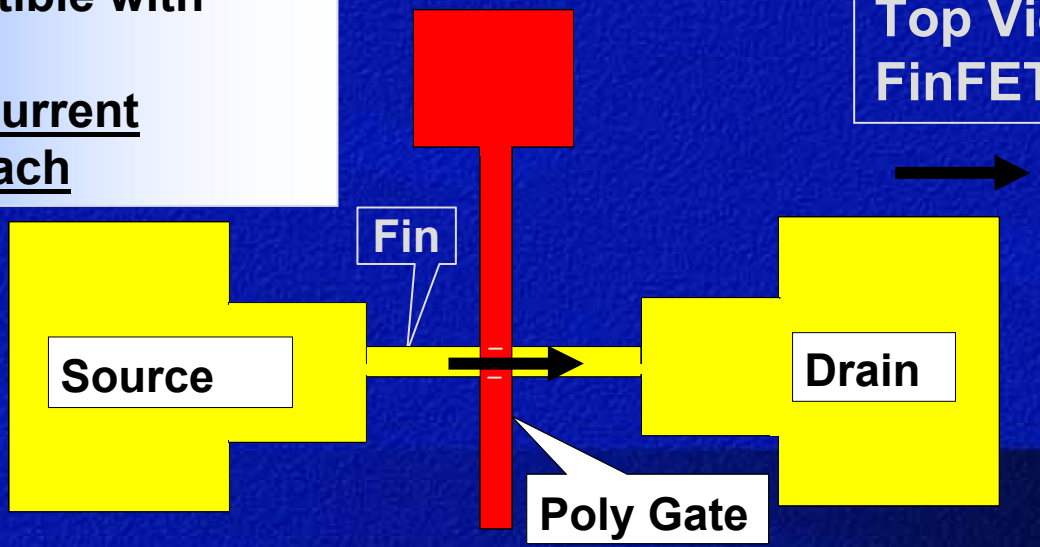


Perspective view of FinFET. Fin is colored yellow.

Courtesy: T-J. King and C. Hu, UC-Berkeley

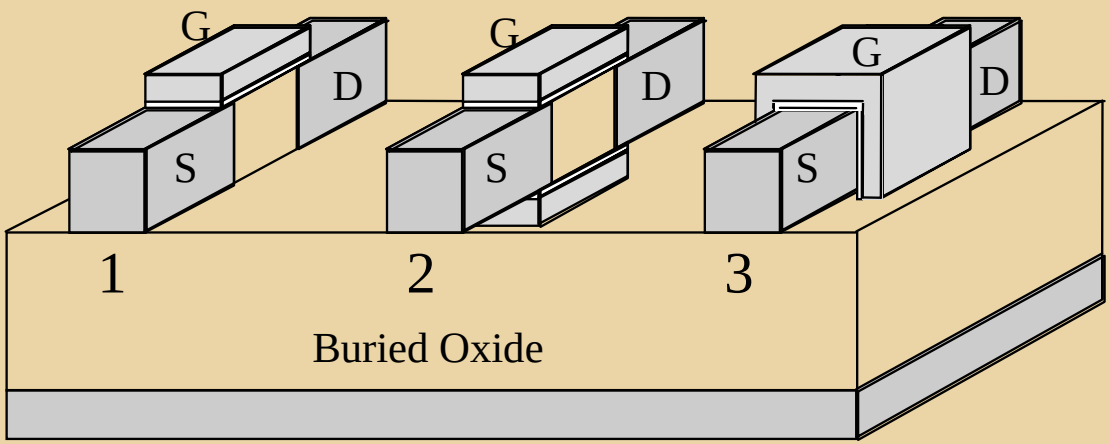
Key advantage: relatively conventional processing, largely compatible with current techniques→current leading approach

Top View of FinFET



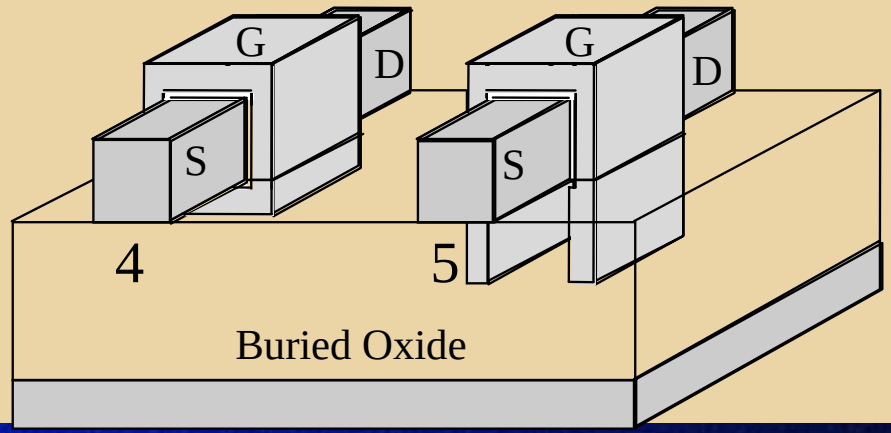
Arrow indicates current flow direction

Types of Multiple-Gate Devices



Courtesy: Prof. J-P Colinge, UC-Davis

1: Single gate
2: Double gate
3: Triple gate
4: Quadruple gate (GAA)
5: Πgate



Metrology and Characterization Challenges for Non-Classical CMOS

- C-V measurements for thin, fully-depleted Si
- Single-gate SOI: measurement of very thin body thickness, 10 nm and less
- Multiple-gate
  - Measurement of fin height and width, high AR
  - Measuring roughness of vertical fin edges
  - Measuring high- k film thickness on vertical fin edges
  - Measuring 2D and 3D doping profiles in thin fins

Outline

- Introduction
- Scaling and its impact
- Front end approaches and solutions
- Non-classical CMOS
- ➢ Summary

Summary

- Rapid transistor scaling will continue through the end of the Roadmap
  - Transistor performance will improve rapidly, but leakage will be hard to control
  - Many technology innovations will be needed in relatively short time to enable this rapid scaling
    - Front-end potential solutions include high-k gate dielectric, metal gate electrodes, and enhanced mobility through strained silicon
      - High-k needed first for low-power (mobile) chips in ~ 2006
    - Structural potential solutions: non-classical CMOS
    - *The technology innovations will raise significant challenges for metrology and characterization*
- Non-classical CMOS and front-end solutions being pursued in parallel, and will likely be combined in the ultimate, end-of-Roadmap device
  - L_g < 10nm MOSFETs expected by the end of the Roadmap in 2018

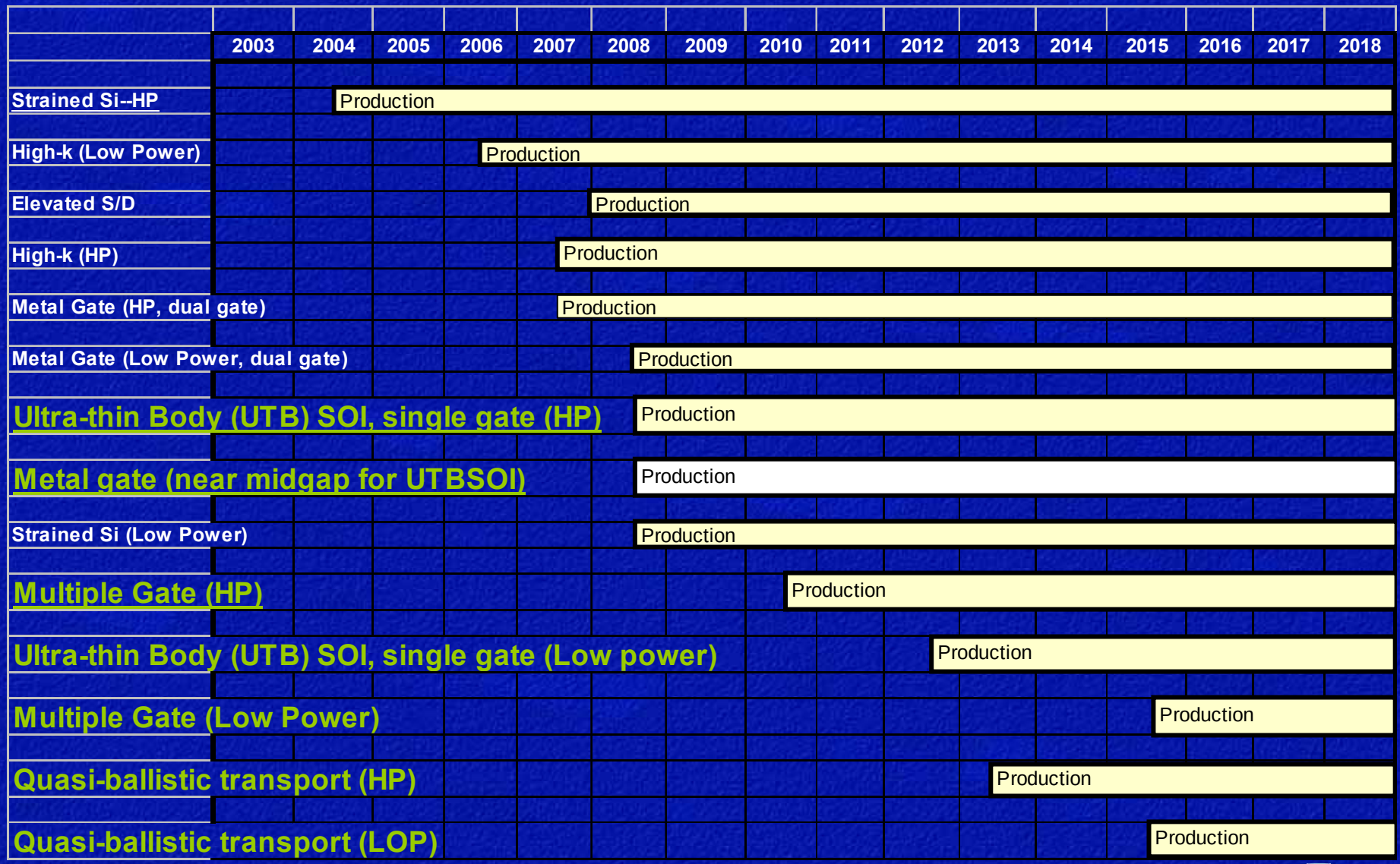
BACK-UP

Potential Solutions for Power Dissipation Problems, High-Performance Logic

- Due to high leakage, static power dissipation is a special challenge
- Increasingly common approach: multiple transistor types on a chip→multi-Vt, multi-Tox, etc.
  - Only utilize high-performance, high-leakage transistors in critical paths—lower leakage transistors everywhere else
  - Improves flexibility for SOC
- Electrical or dynamically adjustable V_t devices (future possibility)
- Circuit and architectural techniques: pass gates, power down circuit blocks, etc.
- Improved heat removal, electro-thermal modeling and design

Timeline of Projected Key Technology Innovations from '03 ITRS, PIDS Section

This timeline is from PIDS evaluation for the 2003 ITRS



International Technology Roadmap for Semiconductors (ITRS)

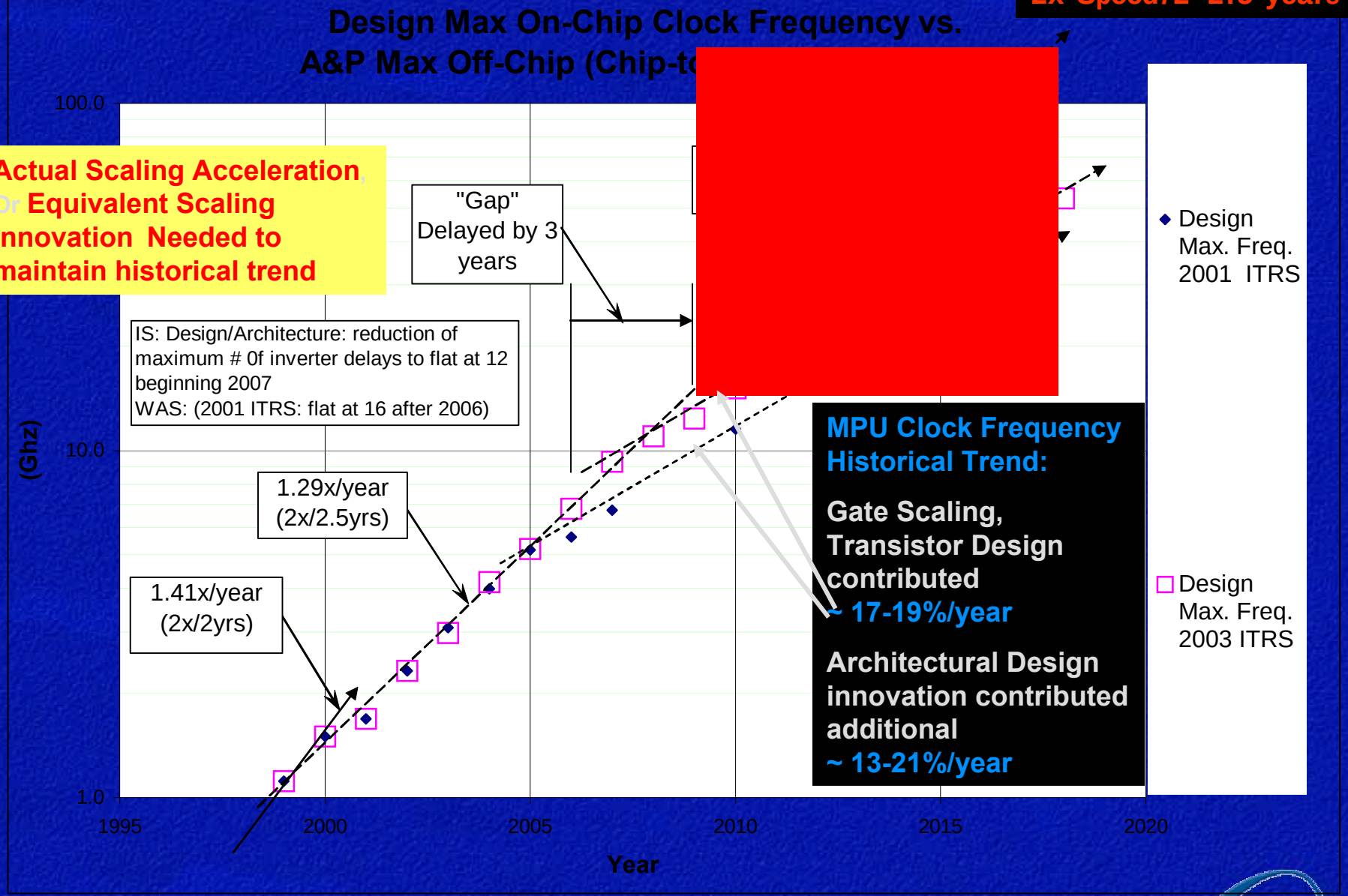
- Industry-wide, fully international effort to map IC technology generations for the next 15 years
  - For each technology generation
    - Projects targets for technology characteristics and requirements
    - Assesses key needs and gaps
    - Lists potential solutions
  - Provides common reference for semiconductor industry: device manufacturers, equipment and materials vendors, researchers
    - Useful for planning
    - Focus: stimulating needed R&D, not intended to restrict research
    - Enabling factor in continuing to follow Moore's Law
  - Much of this talk is based on the 2003 ITRS (formally presented in Dec., 2003)

Typical Technology Requirements Table: High-Performance Logic. Data from 2003 ITRS.

Year in Production	Units	2003	2004	2005	2006	2
Physical Lgate (High Performance)	nm	45	37	32	28	
EOT (Equivalent Oxide Thickness)	A	13	12	11	10	
Gate Poly Depletion & Inversion-Layer Thickness	A	8	8	7	7	
Inversion Gate Dielectric Thickness Value	A	21	20	18	17	
Maximum Gate Leakage Limit	A/cm^2	2.2E+02	4.5E+02	5.2E+02	6.0E+02	9.3
Power Supply Voltage	V	1.2	1.2	1.1	1.1	
Saturation Threshold Voltage	V	0.21	0.20	0.20	0.21	0
Source/Drain Subthreshold Off-State Leakage Drain Current	uA/um	0.03	0.05	0.05	0.05	0
Effective NMOS Current Drive	uA/um	980	1110	1090	1170	1
Sub-threshold Slope Adjustment Factor (Full Depletion/Dual-Gate Effects)(0-1)		1	1	1	1	
Mobility Enhancement Factor		1	1.3	1.3	1.4	
Effective Saturation Carrier Velocity Enhancement Factor		1	1	1	1	
Effective Parasitic Rsd	ohm-um	180	180	180	171	
Ideal NMOS Device Gate Capacitance	F/um	7.4E-16	6.4E-16	6.1E-16	5.7E-16	6.6
Parasitic Fringe/Overlap Capacitance	F/um	2.4E-16	2.4E-16	2.4E-16	2.3E-16	2.2
NMOS Device Time Constant	ps	1.20	0.95	0.86	0.75	0
Relative Performance Improvement (compared to 2003)		1.00	1.26	1.39	1.60	1
Nominal Gate Delay (NAND Gate)	ps	30.24	23.94	21.72	18.92	1
NMOS Device Static Power Dissipation due to Drain & Gate Leakage	Watts/um	3.96E-07	6.60E-07	6.05E-07	6.05E-07	8.4
NMOS Device Power Delay Product	Joules/um	1.41E-15	1.27E-15	1.03E-15	9.66E-16	1.0

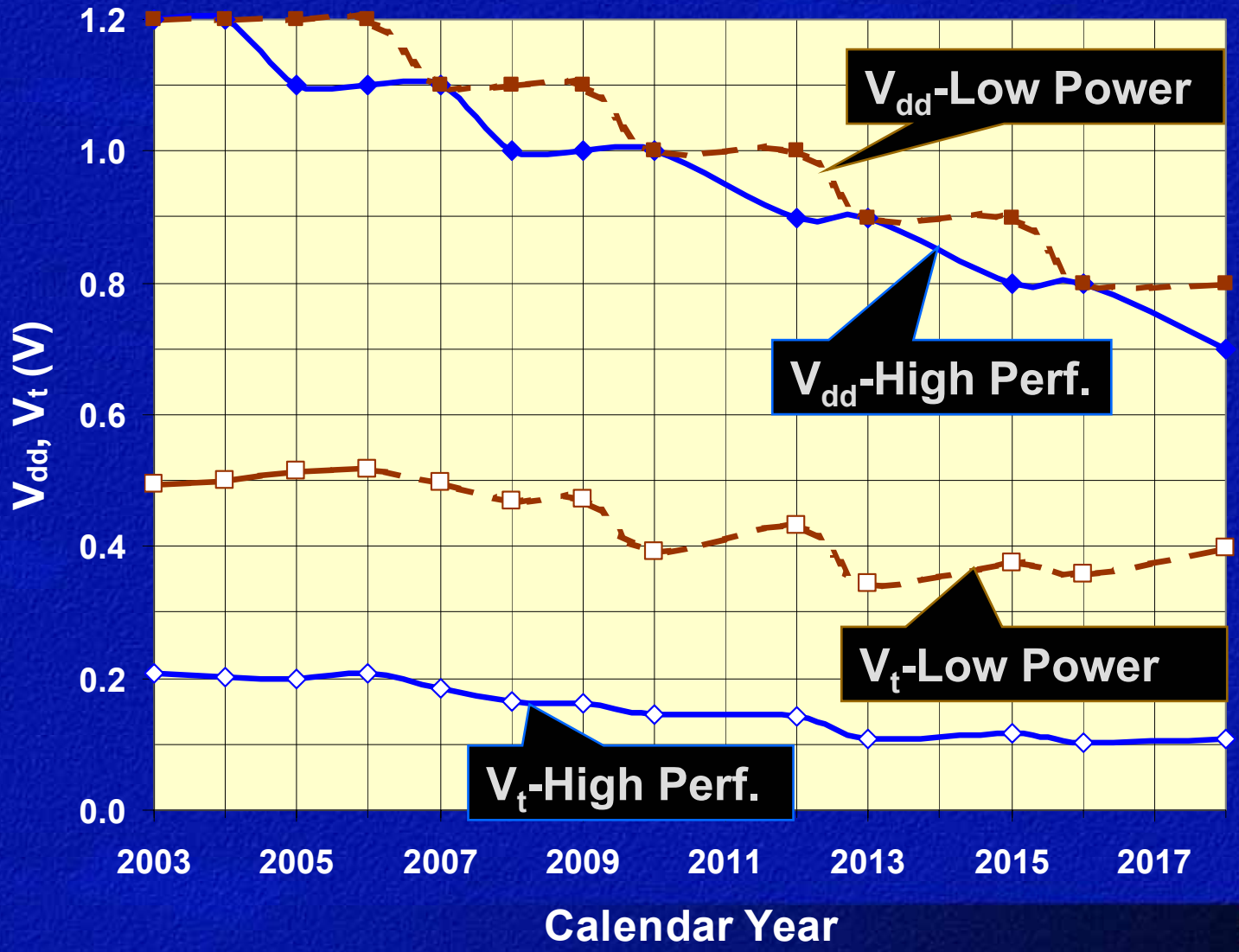
Chip Frequency Scaling, Data from 2003 ITRS

Goal: Increase Speed by **2x Speed/2-2.5 years**



Courtesy: Alan Allan, Intel

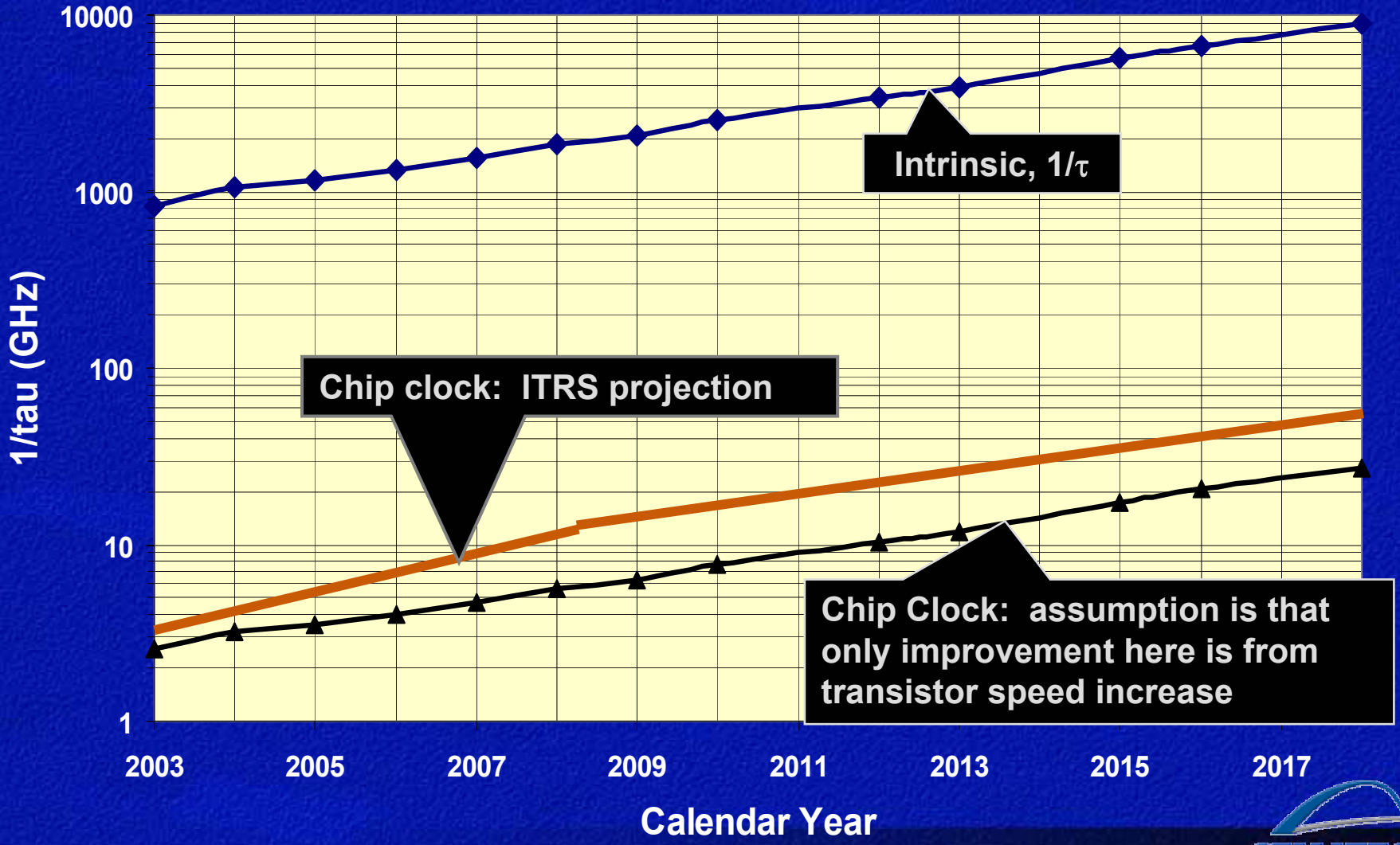
ITRS Projections of V_{dd} and V_t Scaling. Data from 2003 ITRS.



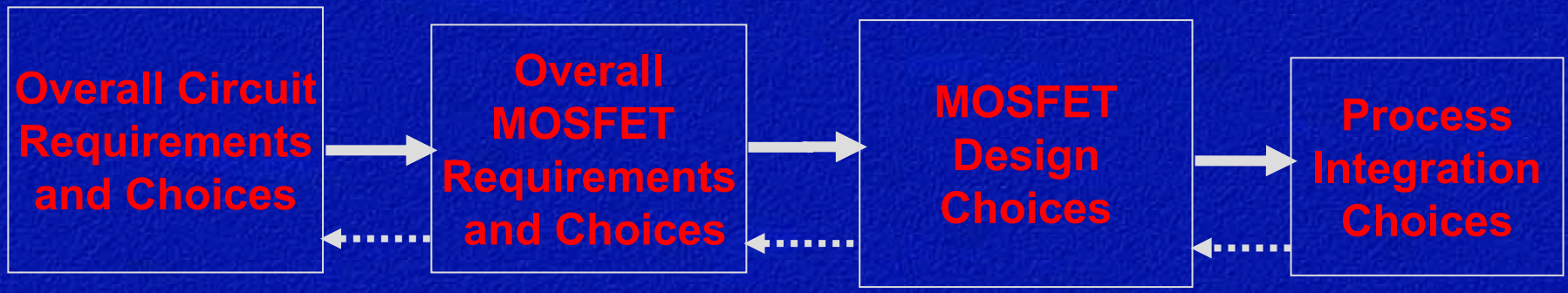
Key MOSFET Scaling Results, 2003 ITRS: Performance and Leakage

- High-performance logic
  - Average 17%/yr improvement in $1/\tau$ is attained
  - $I_{sd,leak}$ is high, particularly for 2007 and beyond → chip static power dissipation scaling is an issue
- Low-power logic
  - Very low $I_{sd,leak}$ target is met
    - $I_{gate,leak}$ is also very low: difficult to meet this→drives need for high-k gate dielectric
  - $1/\tau$ is considerably lower than for high-performance, but close to 17%/yr improvement in $1/\tau$ is still attained
- ITRS MOSFET targets are chosen to drive the technology scaling→pretty aggressive

Frequency scaling: Transistor Intrinsic, Fanout-3 NAND Gate, Chip Clock for High-Performance Logic. Data from 2003 ITRS.



Hierarchy of IC Requirements and Choices



- Chip Power
- Chip Speed
- Functional Density
- Chip Cost
- Architecture
- Etc.

- V_{dd}
- MOSFET Leakage
- MOSFET Drive current
- Parasitic series resistance
- Transistor size
- V_t control
- Reliability
- Etc.

- T_{ox}, L_g, x_j, R_s
- Channel engineering
- Oxynitride or High K gate dielec.
- Classical Planar Bulk or Non-classical CMOS Structures
- Etc.

- Thermal processing
- Overall process flow
- Process modules
- Material properties
- Boron penetration
- Etc.

Key Overall Chip Parameters for High-Performance Logic, from 2001 ITRS

		Near Term							Long Term		
Calendar Year		2001	2002	2003	2004	2005	2006	2007	2010	2013	2016
DRAM Half Pitch	nm	130	115	100	90	80	70	65	45	35	22
Physical Gate Length, L_g	nm	65	53	45	37	32	28	25	18	13	9
Nominal Power Supply Voltage (Vdd)	V	1.2	1.1	1.0	1.0	0.9	0.9	0.7	0.6	0.5	0.4
Maximum on-chip local clock frequency	GHz	1.7	2.3	3.1	4.0	5.2	5.6	6.7	11.5	19.4	28.8
Allowable maximum power dissipation, with heatsink	W	130	140	150	160	170	180	190	218	215	288
Number of transistors per chip	Millions of transistors	276	348	439	553	697	878	1106	2212	4424	8848

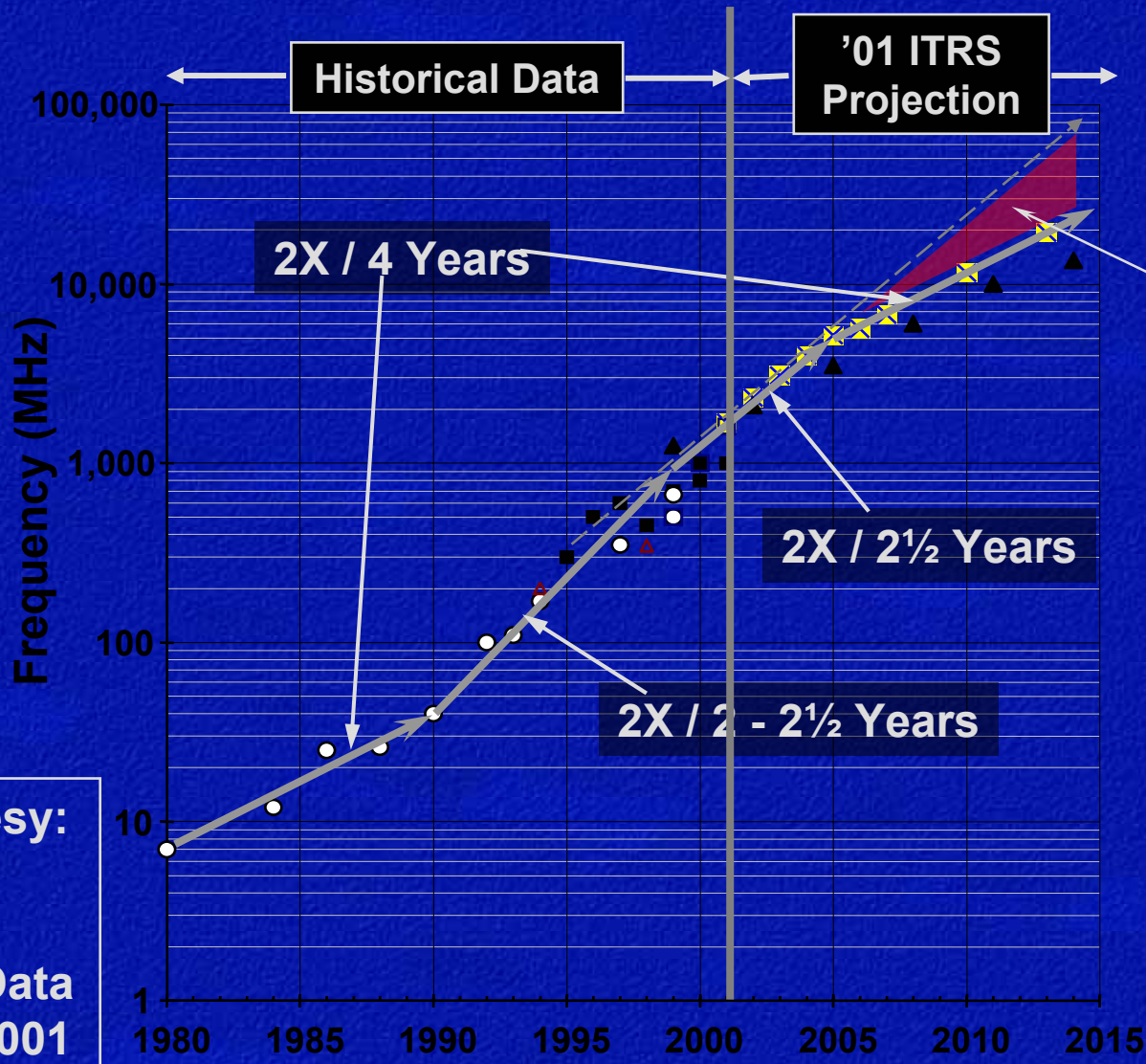
- The DRAM half pitch and L_g are drivers of IC technology scaling, including lithography
- Technology generations (in red) defined by DRAM half pitch
  - This is a dense feature: drives functional density and Litho. and Etch
  - Reduction factor of 0.7X ~ $1/\sqrt{2}$ between generations (130nm in 2001, 90nm in 2004, 65nm in 2007, etc.)
  - Three years between generations
- Gate length (L_g) ≤ 0.5 X DRAM half pitch
  - These are isolated features
  - Rapid scaling of L_g is driven by need to improve transistor speed

V_{dd} and V_t Device Scaling Issues

- We need to scale V_{dd} down rapidly with the technology generations
  - To keep dynamic power dissipation (~V_{dd}^2) within acceptable bounds
  - For reliability, control of short channel effects (SCE), general device scaling
- $1/I_{sd,leak}$ exp. dependent on V_t
- I_{on} strongly dependent on gate overdrive, (V_{dd}-V_t)
- Also, $V_{dd} \geq 2\ V_t$ for circuit functionality

- Scaling requires key tradeoffs between I_{on} and $I_{sd,leak}$, V_{dd} and V_t
  - Tradeoff choices driven by application needs

Historical Data and 2001 ITRS Projections for Chip Clock Frequency, High-Performance Logic



Actual Scaling Acceleration, Or Equivalent Innovation Needed to maintain historical trend

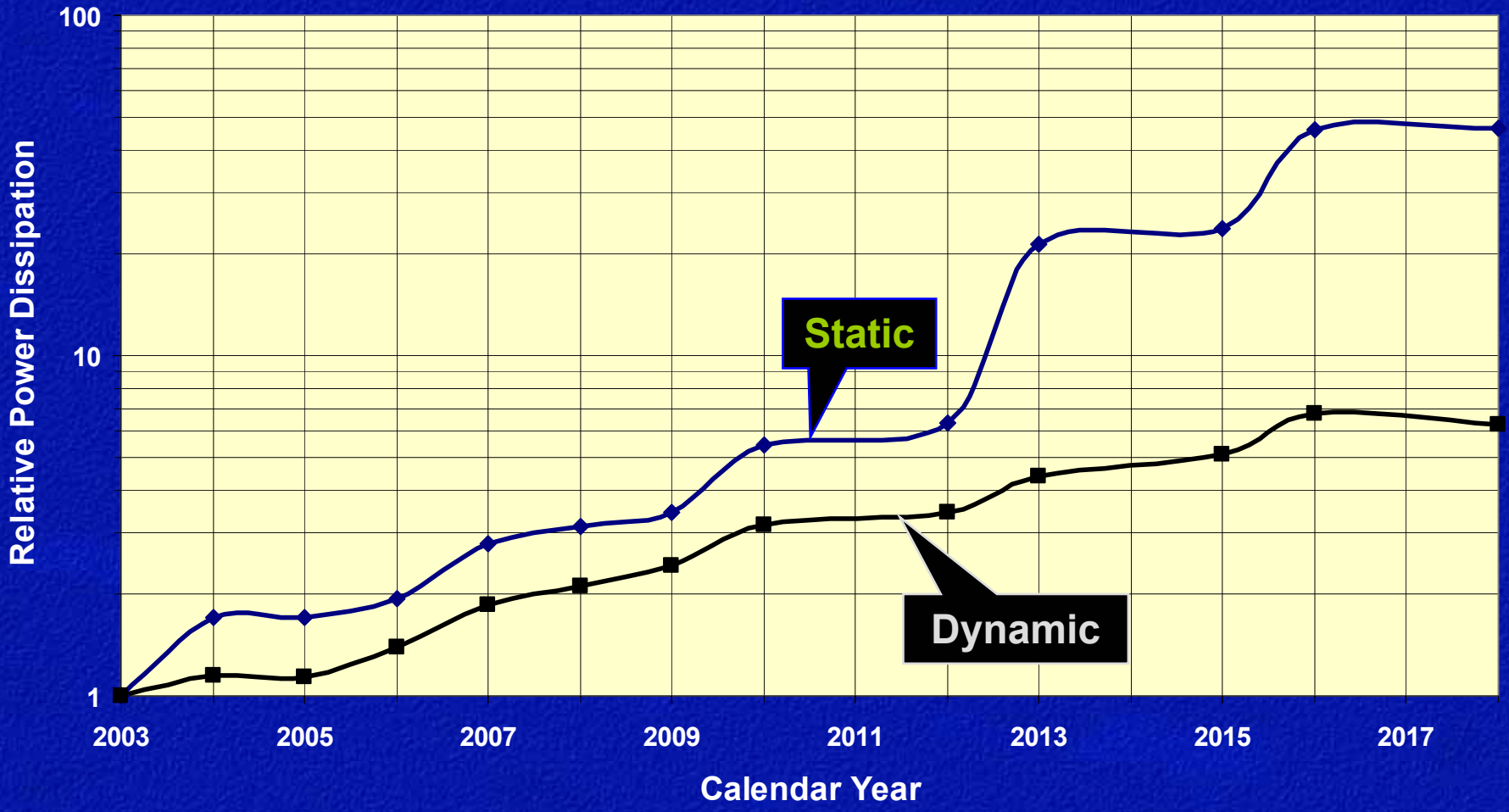
MPU Clock Frequency Historical Trend:

Transistor scaling has contributed ~ 17-19%/year

Architectural Design innovation contributed additional ~ 21-13%/year

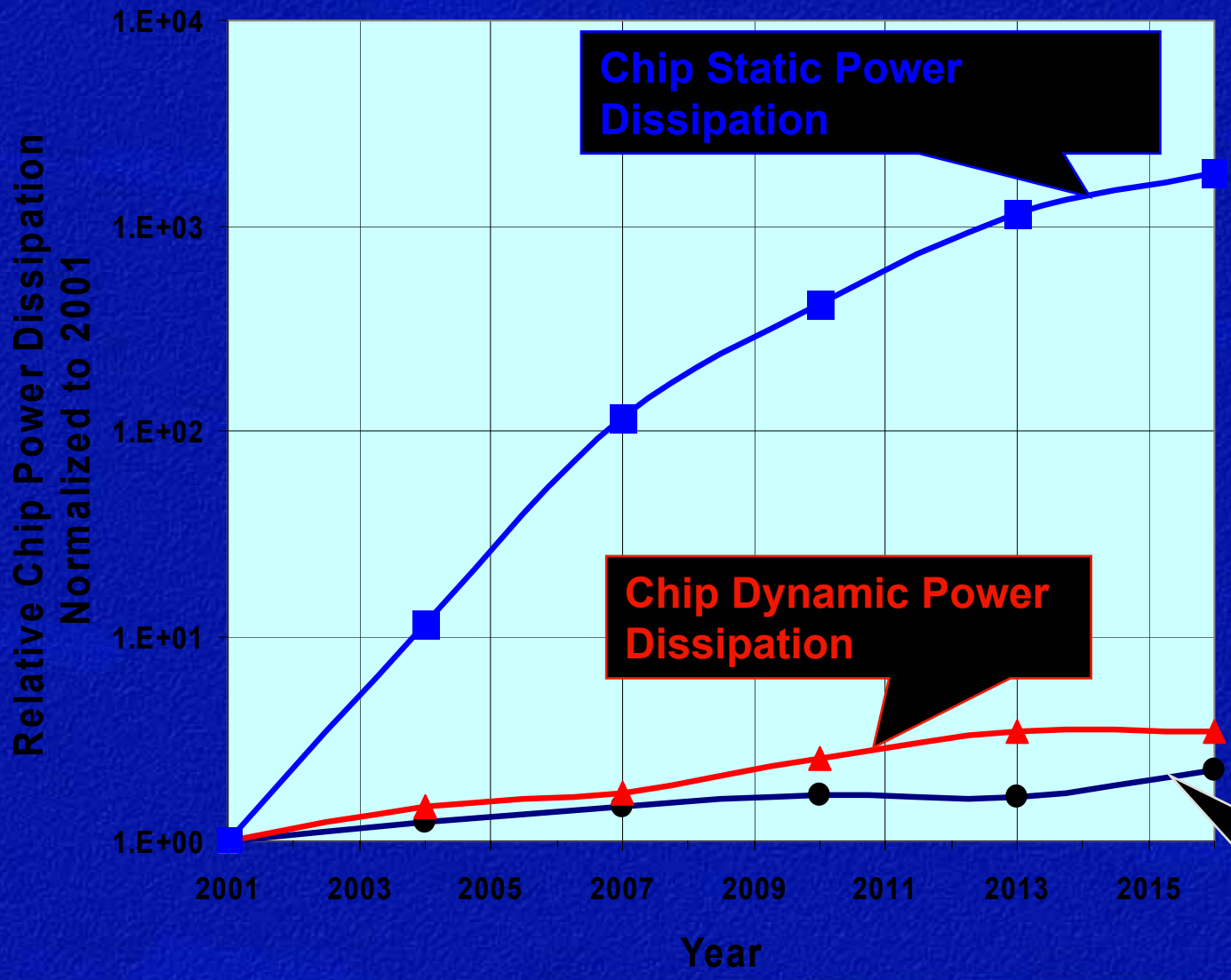
Courtesy: Alan Allan, Intel. Data from 2001 ITRS.

Potential Problem with Static Power Dissipation Scaling: High-Performance Logic



Assumption, to make a point re P_{static}: all transistors are high performance, low V_t type

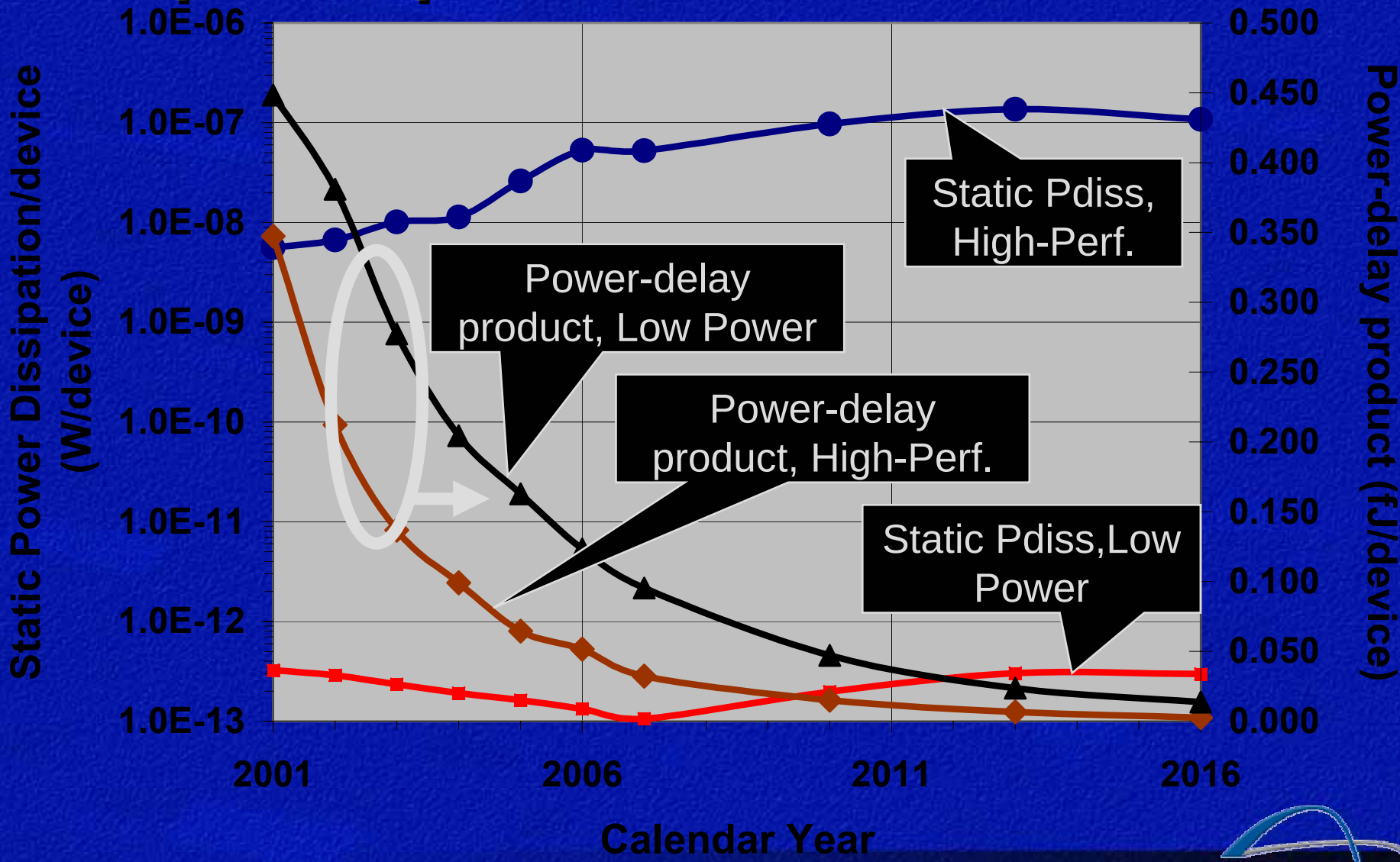
Relative Chip Power Dissipation, High Performance



Assumptions:

- Only one type of (high I_{on}, high I_{leak}) transistor, and no power reduction techniques.
- Simple scaling
- This is an unrealistic scenario, only meant to clearly illustrate static power dissipation issues

ITRS Projected Scaling of Power Dissipation per Device



Impact of Key MOSFET Parameters on Chip Power Dissipation

- $P_{total} = P_{dynamic} + P_{static}$
  - $P_{dynamic} = C_{active} V_{dd}^2 f_{clock}$
    - With scaling, C_{active} and f_{clock} increase rapidly
    - To keep $P_{dynamic}$ within tolerable limits, reduce V_{dd} with scaling
    - Reduce V_{dd} for reliability, SCE, general device scaling reasons, also
  - $P_{static} = N_{off} W I_{leak} V_{dd}$
    - With scaling, N_{off} increases rapidly, but V_{dd} and W scale down
    - To keep P_{static} within tolerable limits, constrain increase of I_{leak} with scaling

Solutions for Power Dissipation Problems, High-Performance Logic

- Increasingly common approach: multiple transistor types on a chip→multi-V_t, multi-T_{ox}
  - Only utilize high-performance, high-leakage transistors in critical paths—lower leakage transistors everywhere else
  - Improves flexibility for SOC
- Electrical or dynamically adjustable V_t devices (future possibility)
- Circuit and architectural techniques: pass gates, power down circuit blocks, etc.

Summary: MOSFET Scaling

- MOSFET scaling is the "raw material" for designers to improve chip performance, control power dissipation
  - MOSFET scaling projected to scale at historical ~17% per year in "raw" speed improvement for high-performance logic
  - Design and architectural innovation has contributed about as much, but is expected to slow down in the future: continued MOSFET speed improvement is critically important
- MOSFET scaling goals are critically important
  - High-performance logic emphasizes speed at the expense of high leakage and static power dissipation
  - Low-power logic emphasizes low leakage at the expense of speed
- Static power dissipation is a growing problem for high-performance logic, and there are numerous approaches to dealing with it

High-K Issues

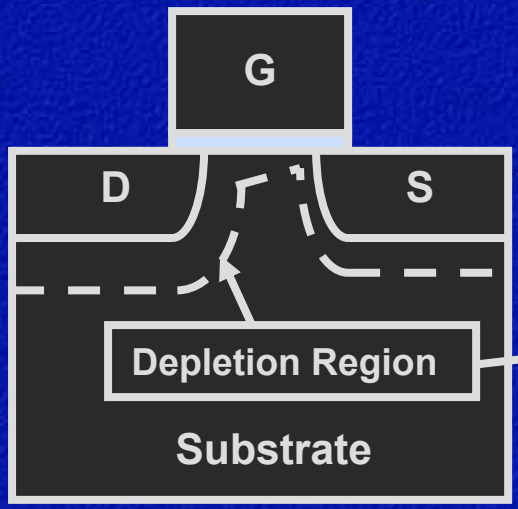
- Process integration
  - Thermal stability of high-k material
    - Retain high-k performance with planar CMOS flow (S/D anneal, etc.,) challenge
  - Chemical, electrical compatibility with polysilicon
    - Boron penetration
    - PMOS V_t
    - Metal electrode may be required
  - Interface with Si substrate and gate electrode
    - Deposition / post process anneals $\Rightarrow$ thin SiO_2-like layer
- Interface properties: D_{it}, Q_f, $\mu = \mu$(interfacial "SiO_2")
- Charges and charge trapping in high-k: V_t control and instability
- Mobility degradation
- Leakage, reliability
- New material: major challenge

Polysilicon Limitations

- Polysilicon depletion
  - Increases effective electrical T_{ox} → reduces inversion charge & I_{on}
  - More of a problem as T_{ox} is scaled → Poly doping must increase with scaling
- PMOSFETs: B penetration through very thin oxides
  - Oxy-nitrides & reduction of DT effective now
- Compatibility with high-k
- Gate resistance of very thin gates (even with silicide)

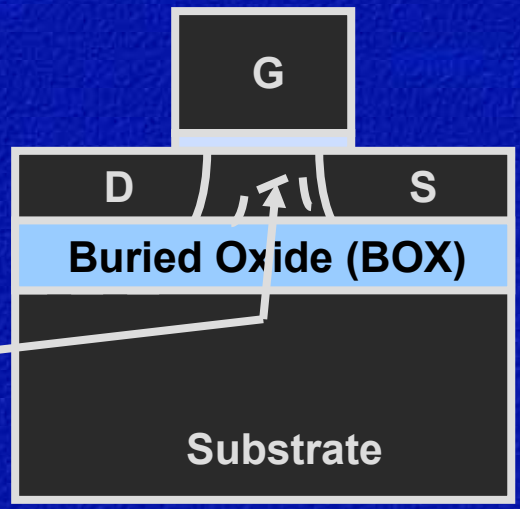
Transistor Structures

Planar Bulk



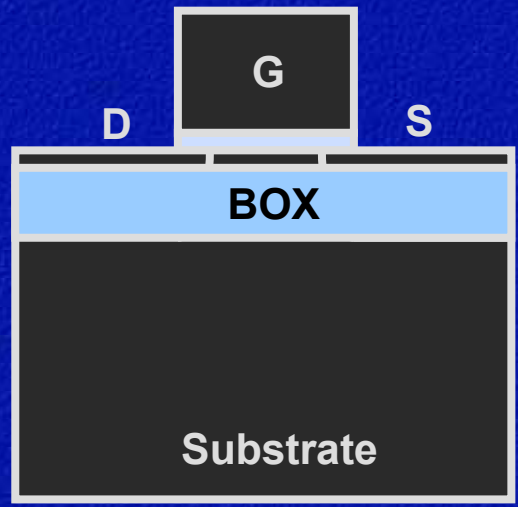
+ Current solution
+ Wafer cost / availability
- SCE scaling difficult
- High doping effects and Statistical variation
- Parasitic junction capacitance

Partially Depleted SOI



+ Lower junction cap
+ F.B. performance boost
- F.B. history effect
- SCE scaling difficult
- Wafer cost/availability

Fully Depleted SOI



+ Lower junction cap
+ Light doping possible
- SCE scaling difficult
- High $R_{series,s/d}$ → elevated S/D
- Sensitivity to Si thickness (very thin)
- Wafer cost/availability

REFERNCES

1. *P.M. Zeitzoff, J.A. Hutchby and H.R. Huff, MOSFET and Front-End Process Integration: Scaling Trends, Challenges, and Potential Solutions Through The End of The Roadmap, International Journal of High-Speed Electronics and Systems,* ***12****, 267-293 (2002).*
2. *Mark Bohr, ECS Meeting PV* ***2001-2****, Spring, 2001.*

Non-Classical CMOS Summary

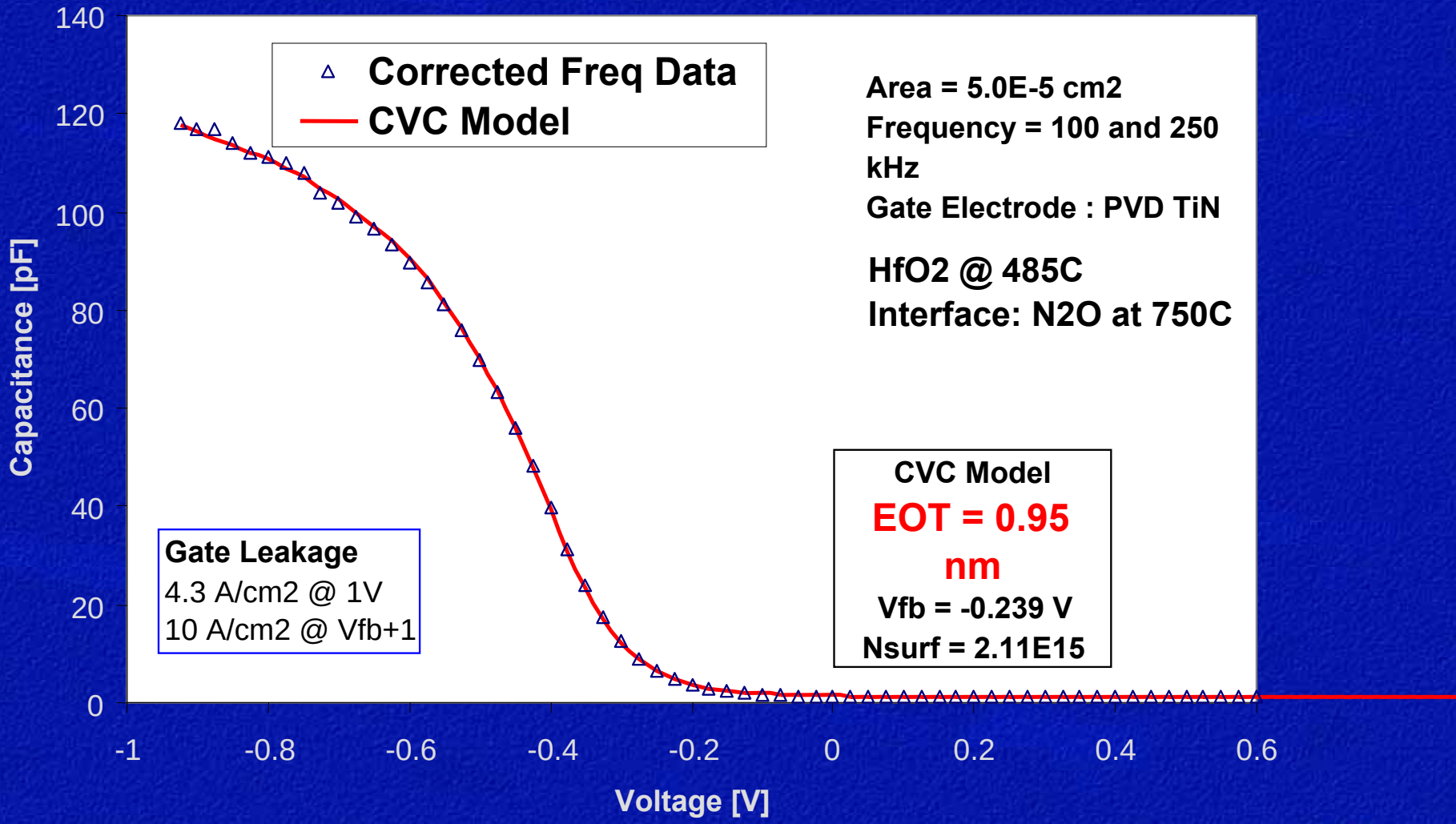
- Below L_g = 25nm or so, planar bulk CMOS may not scale effectively
  - Ultra-thin body, single-gate SOI and (eventually) multiple-gate, ultra-thin body MOSFETs are more optimal from a device point of view than planar bulk CMOS. Key issues:
    - Effectiveness of planar bulk CMOS scaling in this regime
      - Working **but suboptimal** 8nm devices reported in literature
    - Finding effective solutions to difficult processing issues for SOI and multiple-gate
  - Ultimate MOSFET (Lg < 10nm) likely to be multiple-gate with high-k, metal gate electrodes, strained Si, etc.
    - Such devices will require metal electrodes with near-midgap work functions
      - Tuning of work function of single metal gate material may be feasible

High-k Gate Dielectric Candidates and Key Issues

- Modest k (<10)
  - Al_2O_3
    - Negative charge, complicated defect structure
- Medium k (10-25)
  - Group IV Oxides - ZrO_2, HfO_2
    - Low crystallization temperature
  - Group III Oxides - Y_2O_3, La_2O_3,…
    - Charge
  - Silicates - (Zr, Hf, La, Y, ..) SiO_4
    - Lower k if too dilute
  - Aluminates - (Zr, Hf, La, Y, ..)•Al_2O_3
    - Charge issue, complicated defect structure
- High k (≥ 25)
  - Ta_2O_5 , TiO_2
    - Low-barrier height

C. M. Osburn and H.R. Huff, Spring ECS abst. # 366

MOCVD HfO_2 CV Curve (EOT = 0.95 nm)



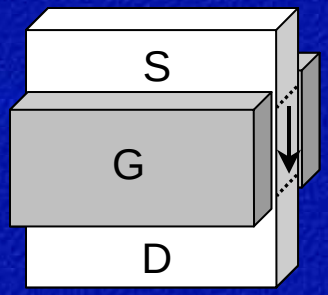
Avinash Agarwal et al., (Alternatives to SiO_2 as Gate Dielectrics for Future Si-Based Microelectronics, 2001 MRS Workshop Series (2001)

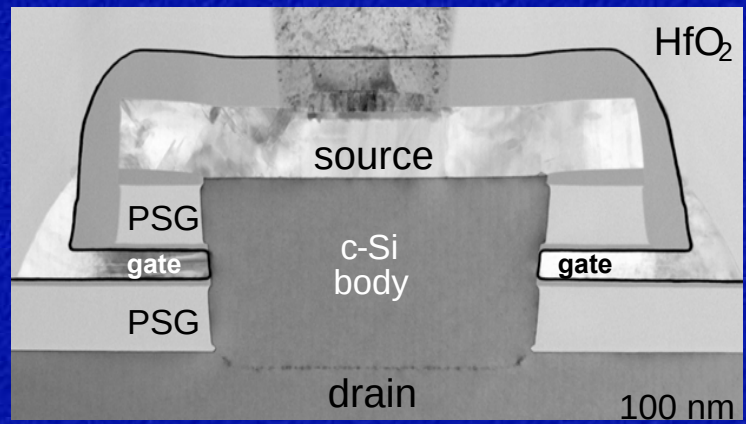
Difficult Transistor Scaling Issues

- With scaling, increasing difficulty in meeting transistor requirements
  - High gate leakage
    - Direct tunneling increases rapidly as T_{ox} is reduced
    - Potential solution: high-k gate dielectric
  - Poly depletion in gate electrode→increased effective T_{ox}, reduced I_{on}
    - Potential solution: metal gate electrode
  - Need for enhanced channel mobility
    - Potential solution: strained Si channels
  - Etc.

Other Structures of Interest

Vertical FET (one type of double-gate MOSFET)

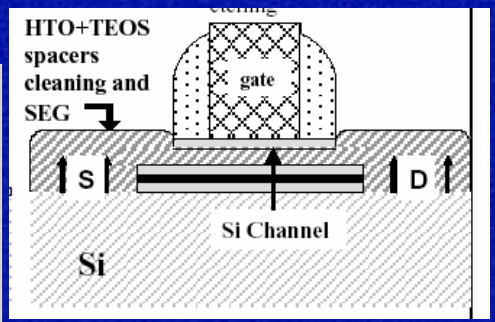


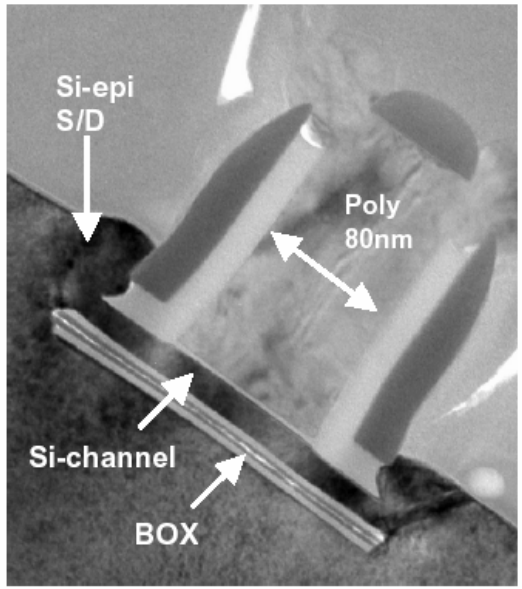


Agere '02

REF: Mark Bohr, ECS Meeting PV **2001-2**, Spring, 2001

Silicon on Nothing (SON): localized buried oxide (BOX)

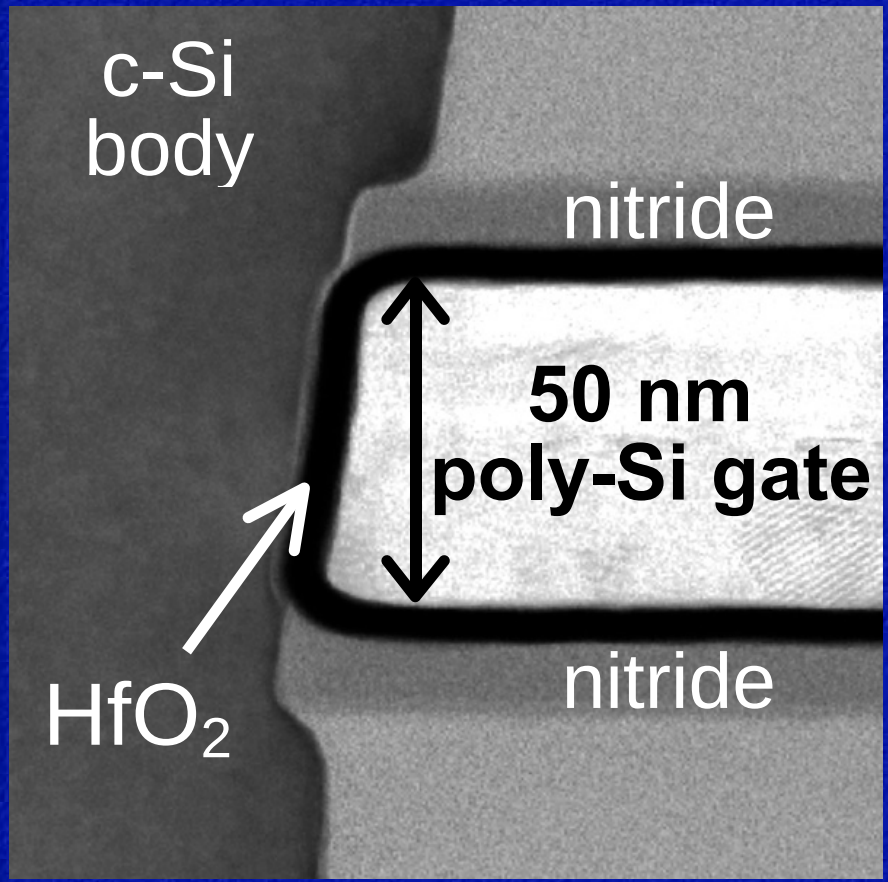


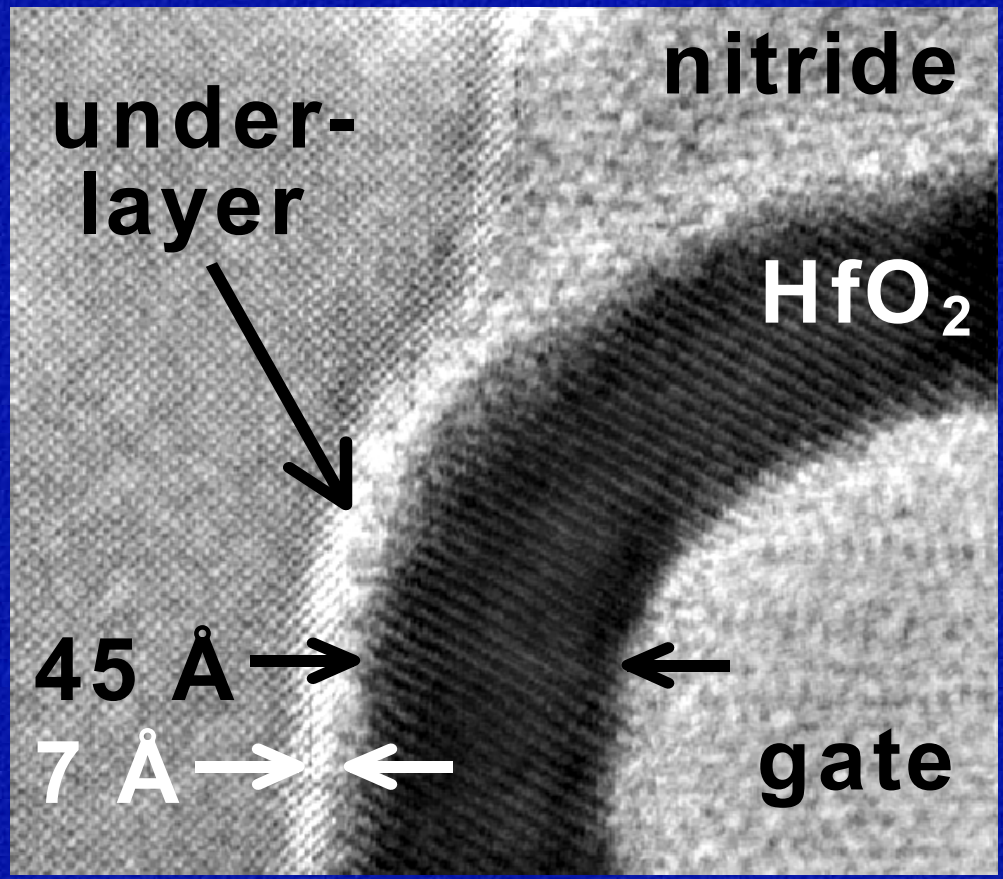


STM '01

REF: S. Monfray et al., '01 IEDM, p. 645.

Vertical Transistor Structure with High-k (Agere '02 IEDM)



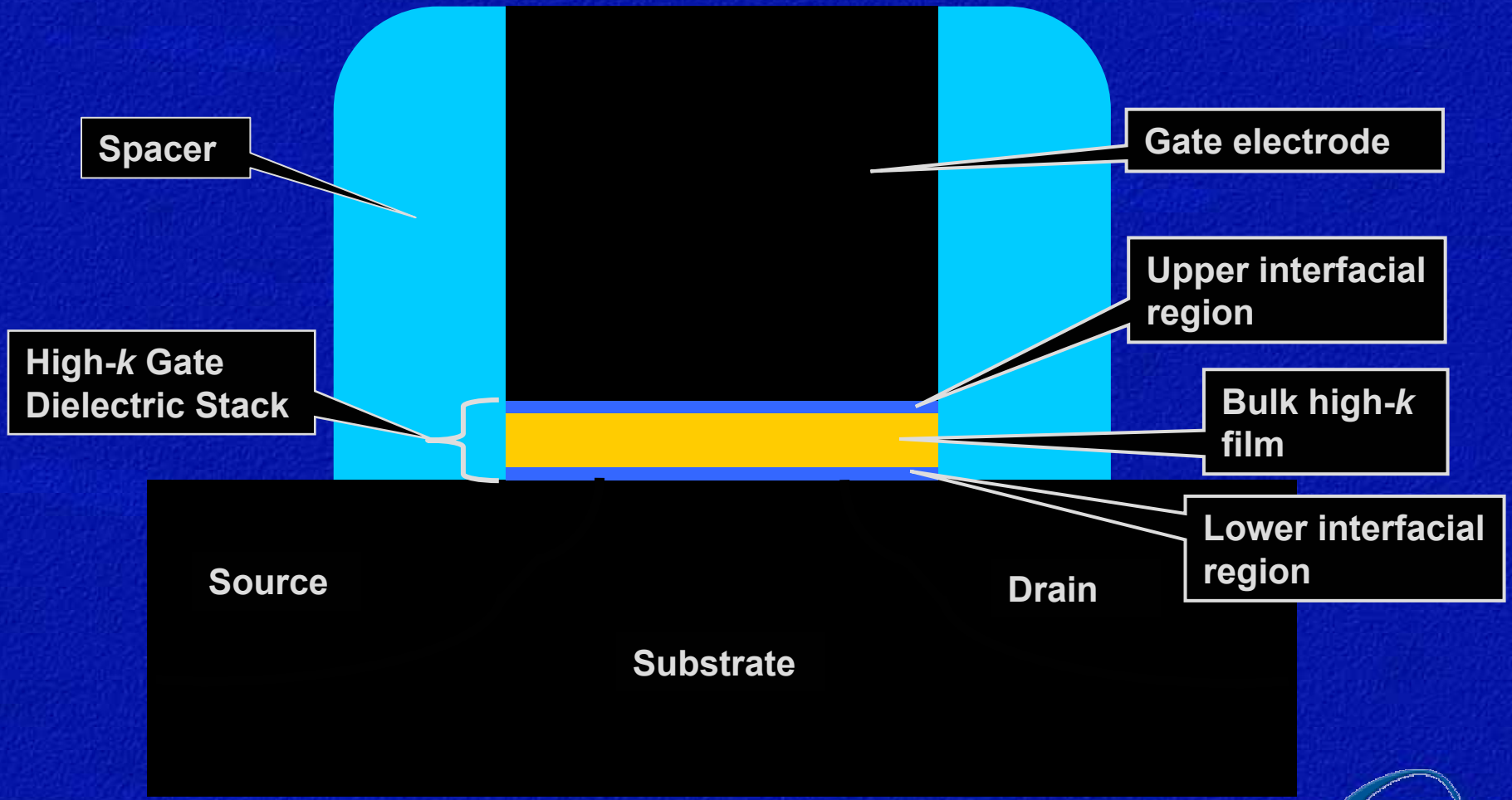


Jack Hergenrother et. al., 50 nm Vertical Replacement-Gate (VRG) nMOSFETs with ALD HfO_2 Gate Dielectrics, *Semiconductor Silicon/2002*, ECS **PV 2002-2**, 929-942 (2002)
Reproduced by permission of The Electrochemical Society, Inc.

Assumptions for All Logic Types

- All modeling is done for nominal devices, room T
- Models are simplified (spreadsheet-based), assume basic transistor functioning doesn't change
  - No dynamic Vt
  - S=85 mV/decade
  - $EOT_{electrical}$ = EOT + 0.8 nm/0.4nm→0.8 nm for poly gate, 0.4 nm for metal gate (in 2007 or beyond)
  - $Log(I_{sd,leak})$~-Vt/S
    - Gate leakage and junction leakage are related to $I_{sd,leak}$
  - $Id,sat \sim g_{m,eff}$ (Vdd-Vt)
  - Cideal = $\varepsilon ox/(EOT_{electrical})$; Cgate = Cideal + Cparasitic
  - τ=(Cgate Vdd)/(Id,sat)= intrinsic transistor delay
  - Parasitic Rs,d is included (20-30% of Vdd/Id,sat = Ron)
  - PMOS is like NMOS, except PMOS Id,sat is 40-50% of NMOS Id,sat
  - S/D junction capacitance is ignored in calculating

Simplified Cross-Section of High K Gate Dielectric Stack



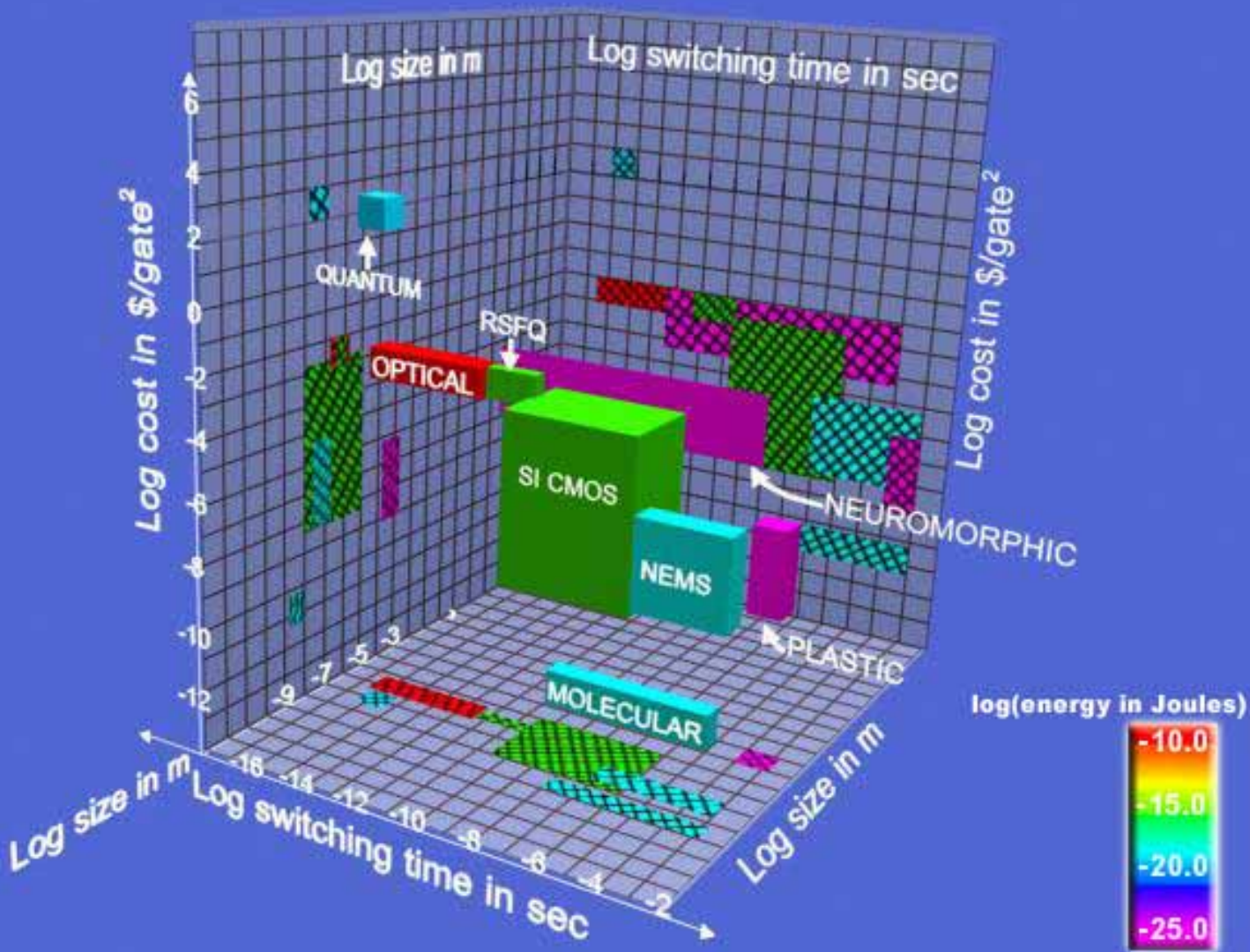
Process - Structure - Property Relation

- Crystallinity / polycrystallinity
  - Phase structure
    - Epitaxial alignment to substrate
  - Stoichiometry
  - Bond coordination
  - Morphology
  - Interfacial microroughness
- Retention of amorphicity by doping
- Mixed oxide phase separation
- Spatial inhomogeneity / periodicity in energy gap(s)

Why High-K (Dielectric Constant)?

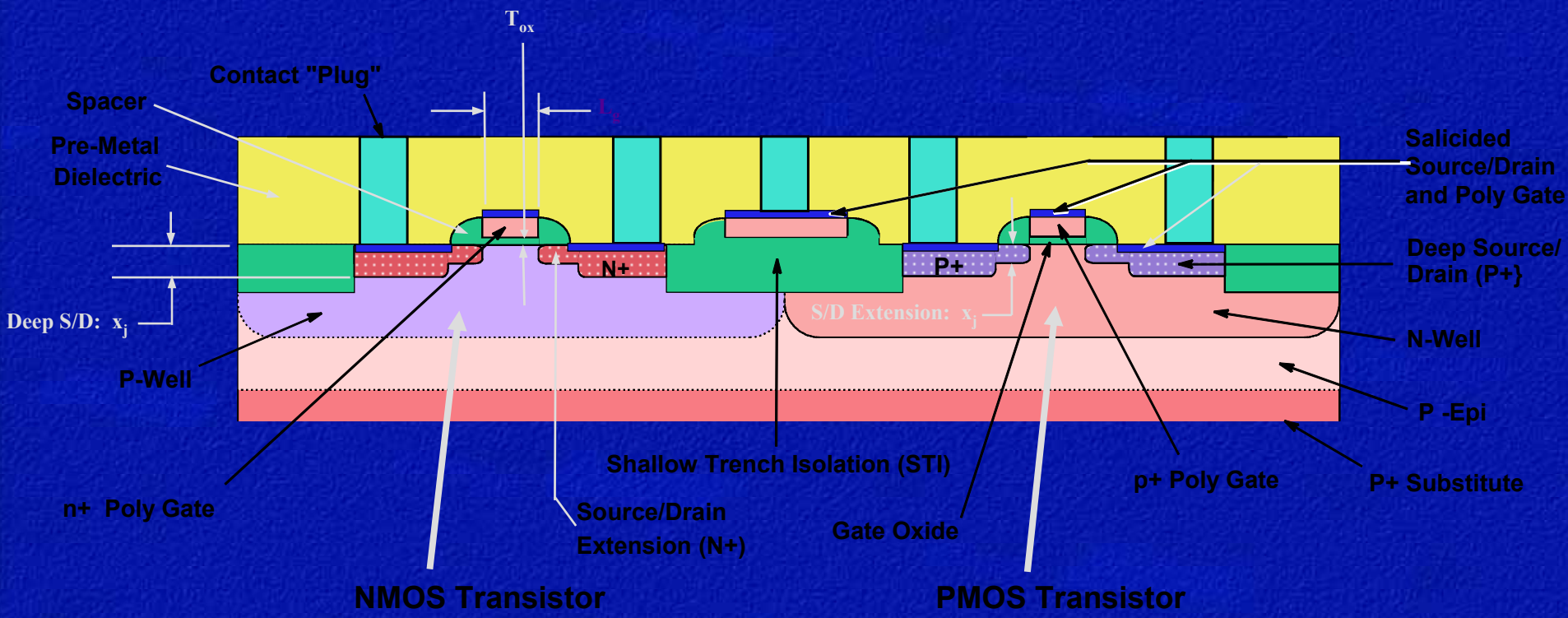
- Direct tunneling current depends on film *physical* thickness and barrier height (φ)
  - $I_{DT} \propto [\exp - \sqrt{[(2m^*q\,\varphi / (h/2\pi)^2]}\,[T_{phys}]$
- Transistor drive current depends on film *electrical* thickness
  - $I_{DSAT} = (w/2l)\,(3.9K_oA)\,(T_{EOT,INV})^{-1}\,\mu\,(V_G-V_T)^2$; $V_G \Rightarrow V_{DD}$
  - $T_{EOT} = T_{phys}$ x $(k_{SiO_2}/k_{high\ k})$
    - $k_{SiO_2} = 3.92$; $k_{high\ k} \approx 15 - 25$
- Increasing *k* increases I_{dsat} without increasing I_{DT}
  - Transistor performance improves *or* thickness may be increased (with increased *k*) to reduce gate leakage (direct tunneling) current without loss of transistor performance
- High-k gate dielectric proposed to obviate IC power concern while still achieving required gate electrode capacitive coupling with silicon
- High-k introduces new set of design constraints

Emerging Technology Parametrization



Simplified Cross Section of a Typical PMOSFET and NMOSFET

(Not to scale)

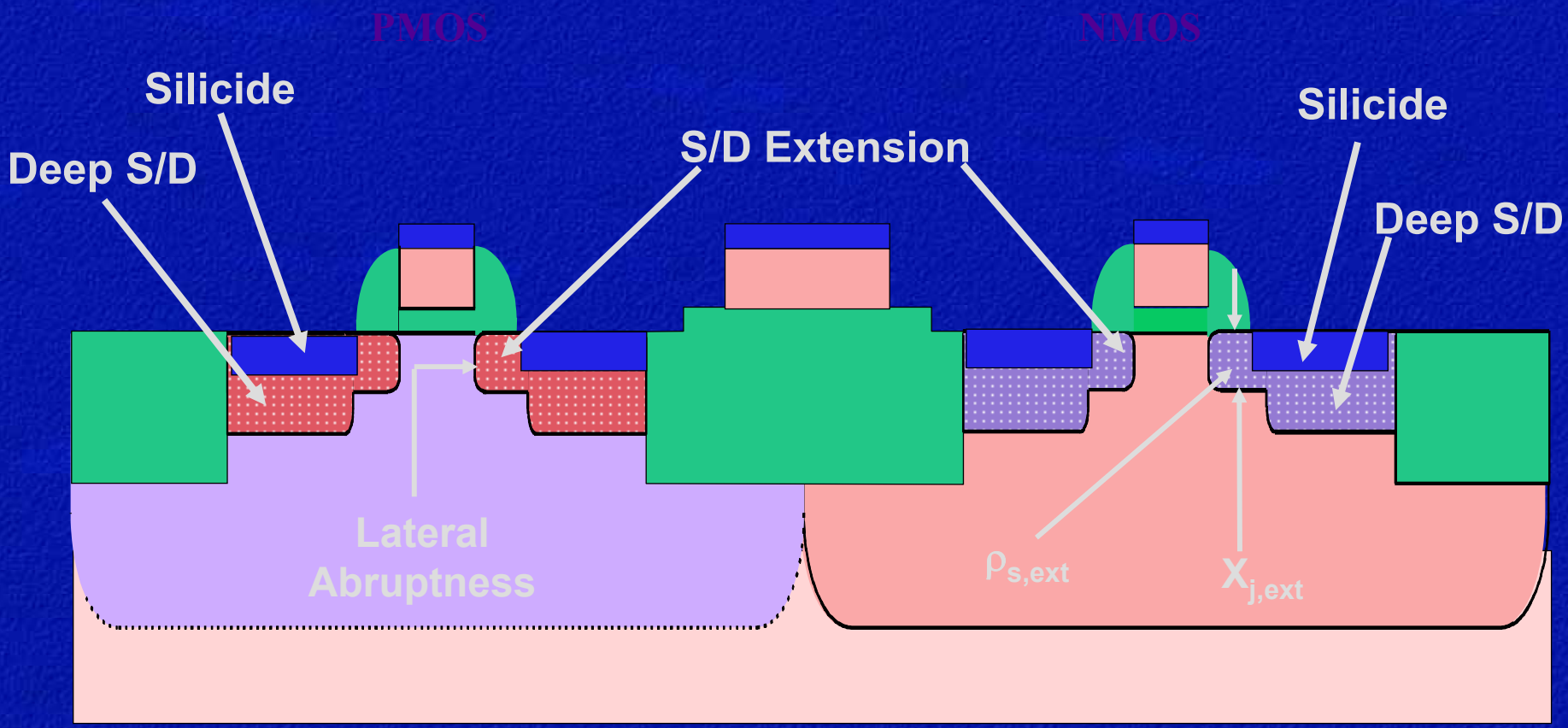


Difficult Transistor Scaling Issues

- With scaling, increasing difficulty in meeting transistor requirements
  - High gate leakage
    - Direct tunneling increases rapidly as T_{ox} is reduced
  - Poly depletion in gate electrode→increased effective T_{ox}, reduced I_{on}
  - Scaling S/D extension and deep S/D
    - High $R_{series,s/d}$ → reduced I_{on}
  - Etc.

S/D Extension Issues

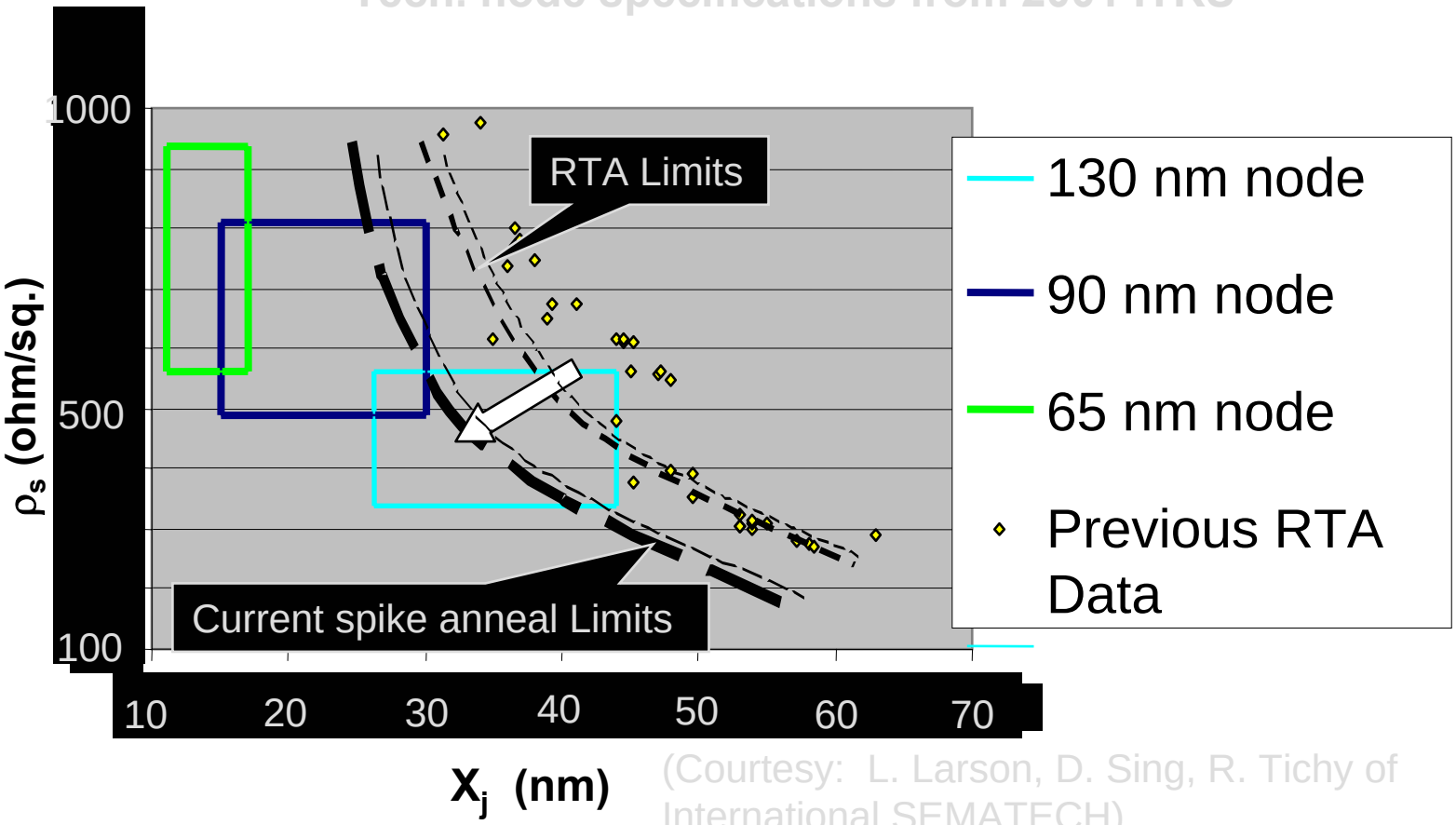
Schematic, not to scale



S/D Extension Issues

- Increasingly abrupt, shallow, heavily doped profiles required for successively scaled technologies
  - Needed for optimal devices, esp. to control short channel effects (SCE)
  - Difficult ρ_s-$x_{j,ext}$ tradeoffs, esp. for PMOS (B)➔difficult to control $R_{S/D,series}$

S/D Extension Solutions

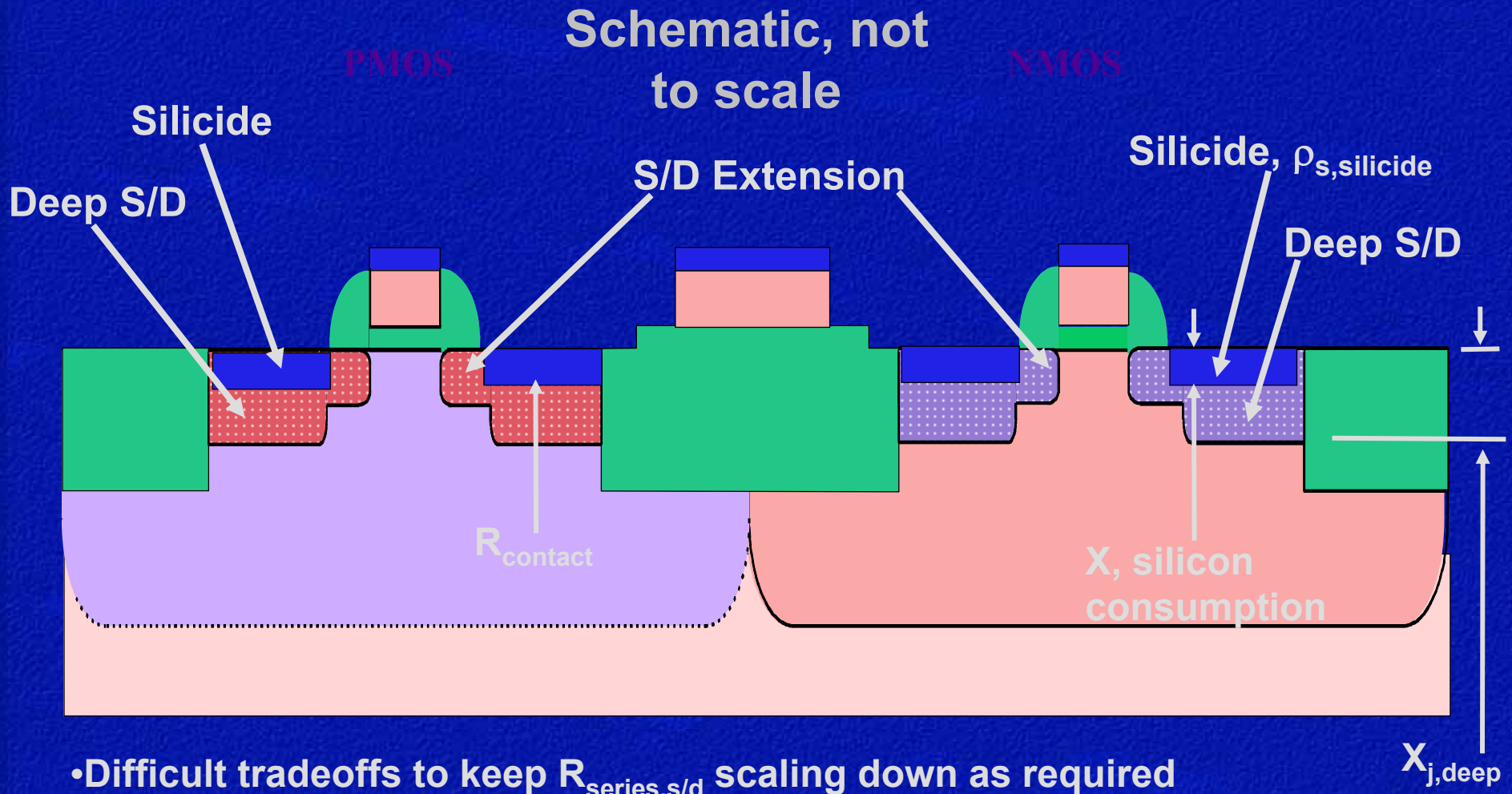


65 nm node and beyond: may require novel doping and annealing techniques

S/D Extension Potential Solutions

- Shorter range
  - Ultra-low energy implants (< 1 KeV, B)
  - Rapid Thermal Processing (RTP) and spike anneal: reduces DT & TED
  - Increase dose as much as possible==>reduced $R_{series,s/d}$
- Beyond 90 nm technology generation
  - Laser thermal annealing
  - Doped, selective epi
  - Co-implant
  - Others

Deep S/D & Silicide Issues

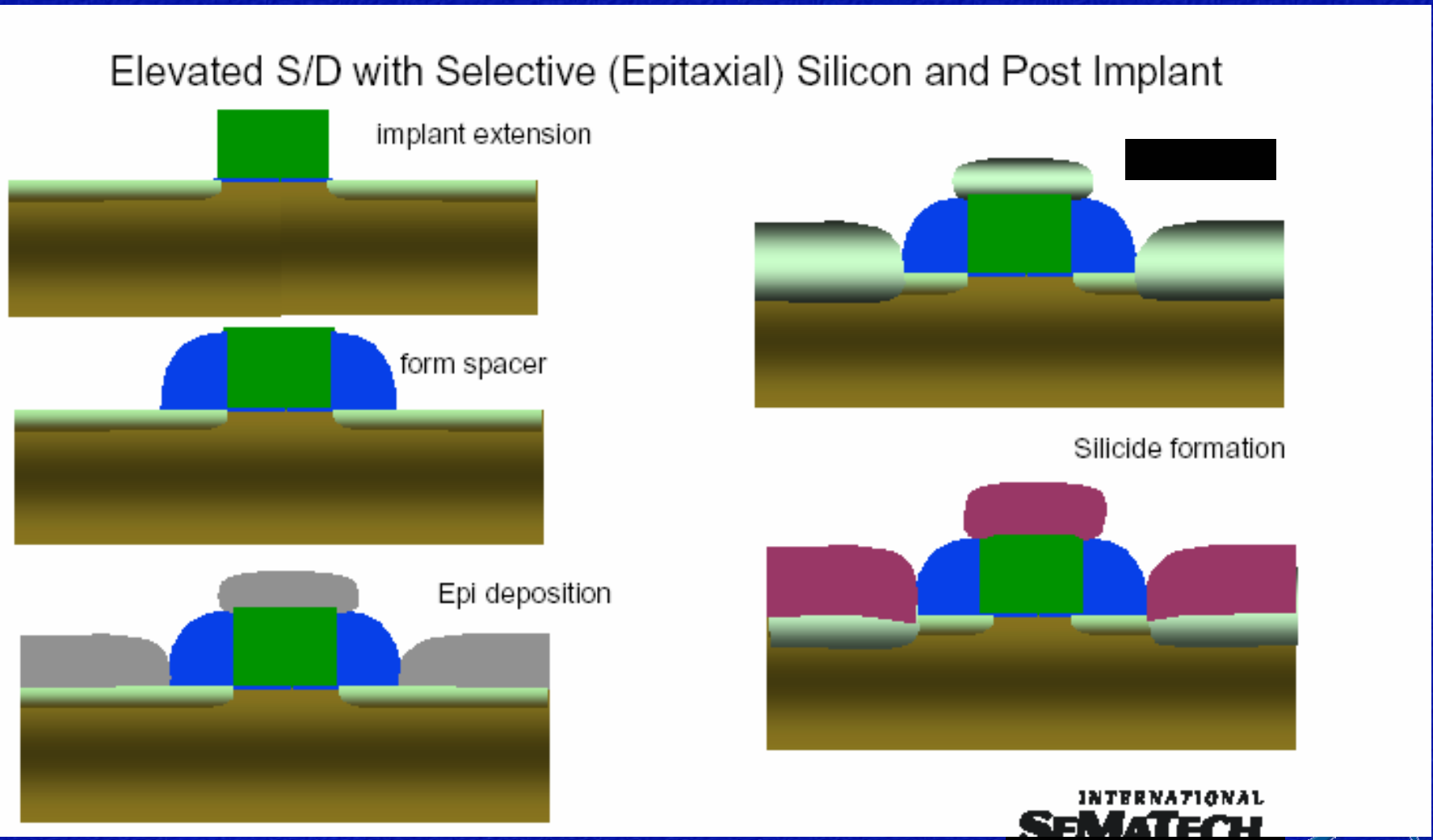


- Difficult tradeoffs to keep $R_{series,s/d}$ scaling down as required
  - $\rho_{s,silicide}$ must be minimized → X must be maximized
  - But X must be kept ≤ $X_{j,deep}/2$ to avoid excessive junction leakage
  - Also, $R_{contact}$ must be reduced with scaling

Deep S/D & Silicide Potential Solutions

- Through 90 or 65 nm generation, tradeoffs to get acceptable $R_{series,s/d}$ are possible
  - Change silicide to get better $\rho_{s,silicide}$ - X tradeoff: $TiSi_2 \rightarrow CoSi_2 \rightarrow NiSi$
- Potential long-range solutions
  - Elevated S/D: doped, selective epi
  - Reduced $R_{contact}$
    - Selective CVD silicide tailor Schottky energy barrier
    - Selective deposited metal

Elevated S/D

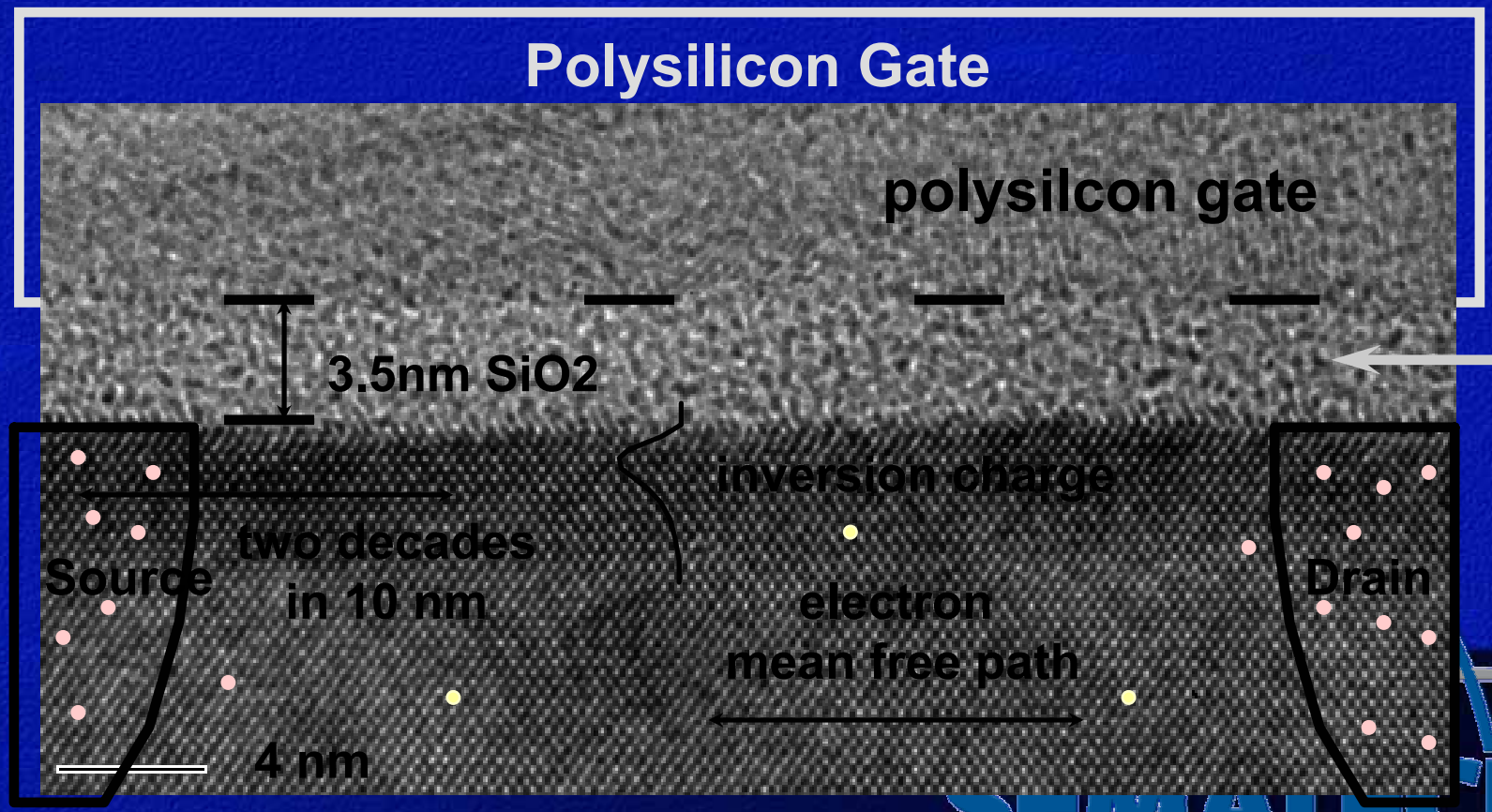


Courtesy: Eric Graetz, Infineon

Device Metrics

- Power
  - $P_{dynamic} = f_{clock}\ C_{load}\ V_{DD}^2$ and $P_{static} = N_{tr} W\ I_{leak}\ V_{DD}$
- Intrinsic transistor gate delay (speed)
  - $\tau = C_{load}\ V_{DD} / I_{DSAT}$
    - Maximum saturated drain current (I_{DSAT}): ideal, long-channel device
      - $I_{DSAT} = (W/2L_{phys})\ (3.9K_oA)\ (T_{EOT,INV})^{-1}\ \mu_{eff}\ (V_G - V_T)^2$
        - » W and L_{phys} device width and physical gate length
        - » $T_{EOT,INV}$ = equivalent oxide thickness in inversion
        - » μ_{eff} = mobility, generally determined for a long-channel device (g_m)
        - » $V_G - V_T$ = gate overdrive, where V_G is supply voltage (V_{DD}) applied to gate ($V_G \Rightarrow V_{DD}$) and V_T is threshold voltage
    - $C_{load} \sim (3.9K_oA)\ (T_{EOT,INV})^{-1} = \varepsilon_{ox} / T_{EOT,INV}$
    - Transconductance
      - $g_m = (W/L_{phys})\ (3.9K_oA)\ (T_{EOT,INV})^{-1}\ \mu_{eff}\ V_{DD}$
    - $T_{EOT} = (k_{high\ k} / k_{SiO2})\ T_{phys}$
    - S = Sub-threshold swing $\Rightarrow$ Inverse slope of log I_D versus V_G

High Resolution TEM Showing 30 nm Channel Length

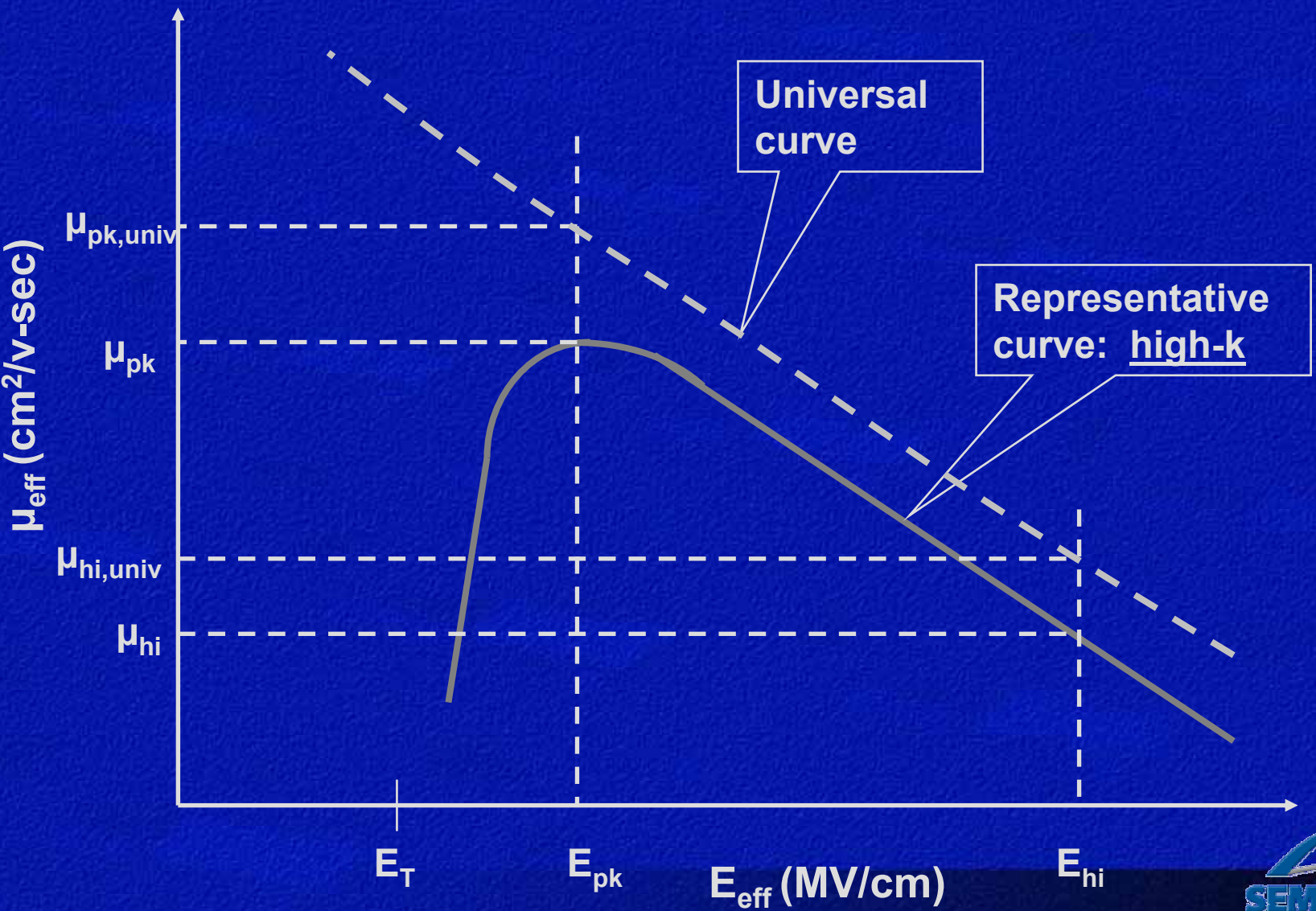


SEMATECH

Accelerating the next technology revolution.

Courtesy of Yoshi Nishi / Dick Chapman

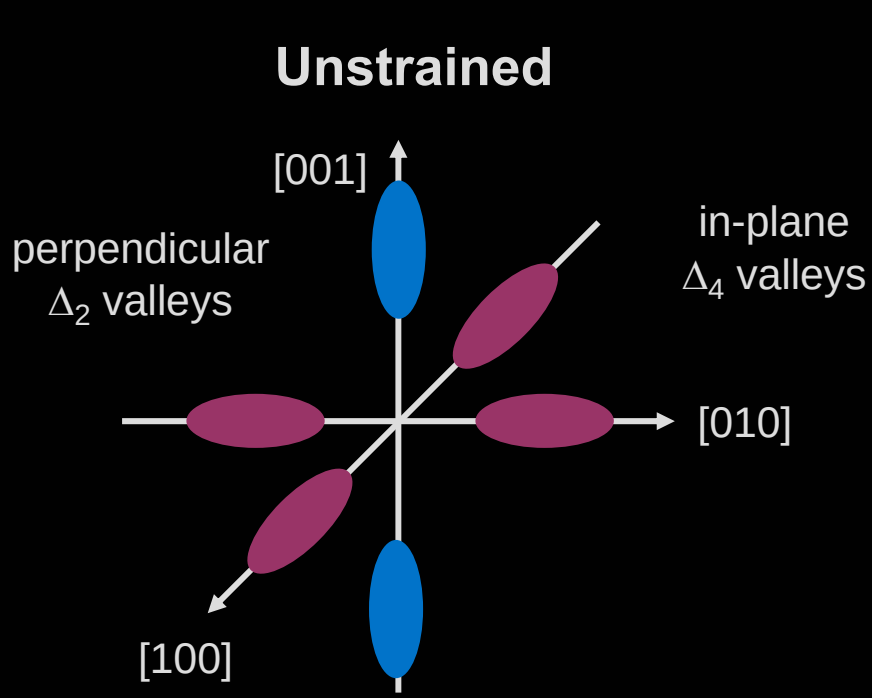
Representative Theoretical and Universal Mobility Curve

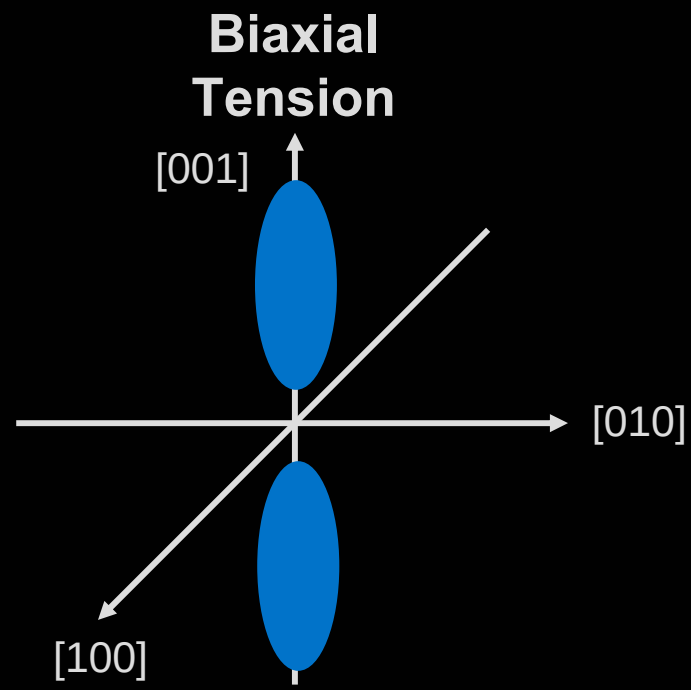


Mobility Considerations

- **Theoretical**
  - Low electric field
    - Unscreened (by inversion layer free carriers) ionized dopant scattering centers in silicon
  - High electric field
    - Acoustic phonons
    - Surface microroughness
      - H x L (where *H* is height of surface undulation and *L* is undulation correlation length)
    - Remote scattering due to high-k phonons
- **Experimental adders (not presently theoretically modeled)**
  - Interfacial and high-k bulk traps
  - N, Al and other elemental scatterers
  - Crystalline inclusions in amorphous high k gate dielect
  - Remote scattering due to gate electrode
- **Universal curve only considers high electric field contributions (extends to low electric field)**

Electron Transport in εMOS™



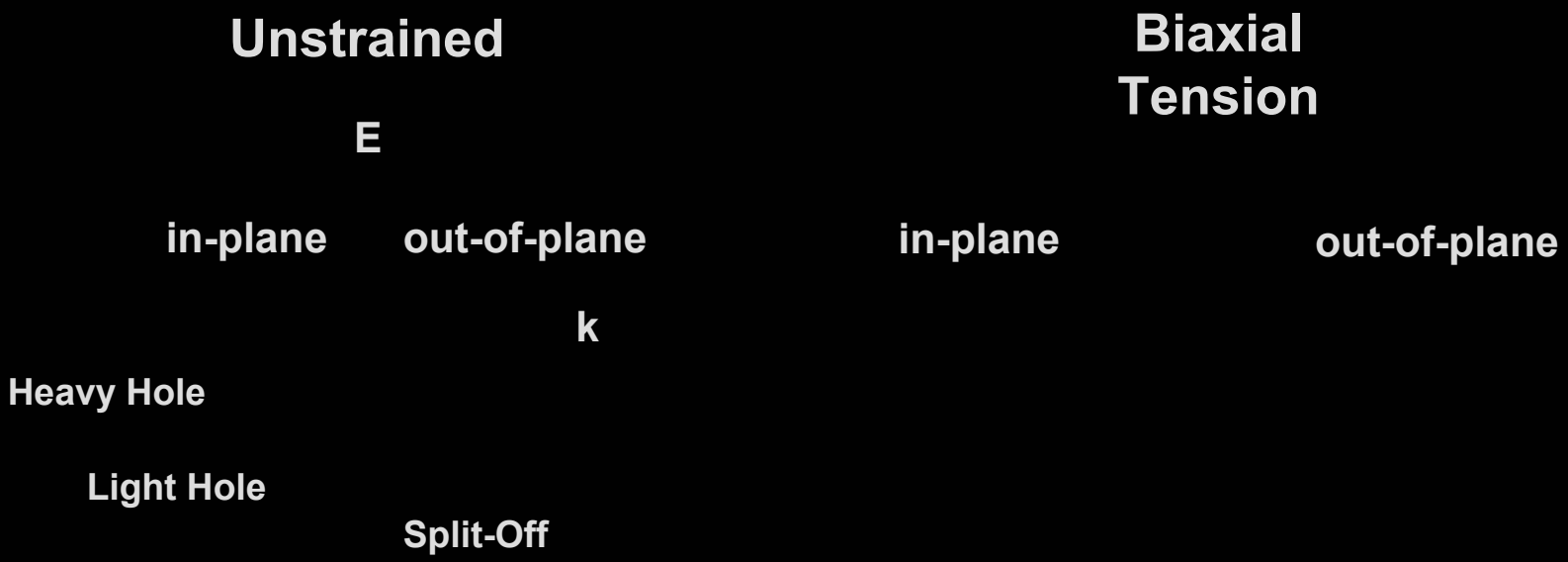


Tensile strain splits conduction band degeneracy

- **Reduced intervalley scattering**
- **Light in-plane effective mass**

Courtesy of Matt Currie
AmberWave Systems Corp.

Hole Transport in εMOS™



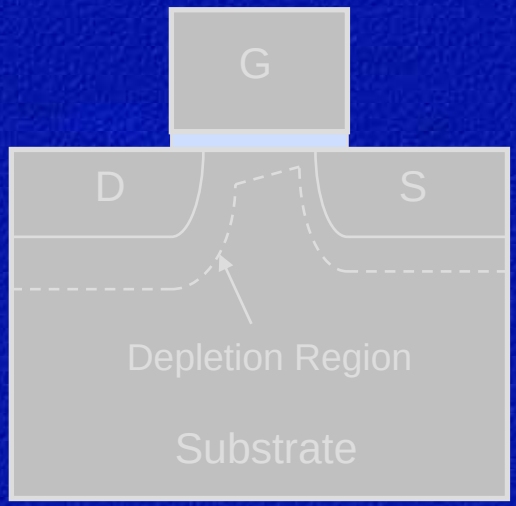
Tensile strain splits valence band degeneracy

- Reduced intervalley scattering
- Light in-plane effective mass

Courtesy of Matt Currie
AmberWave Systems Corp.

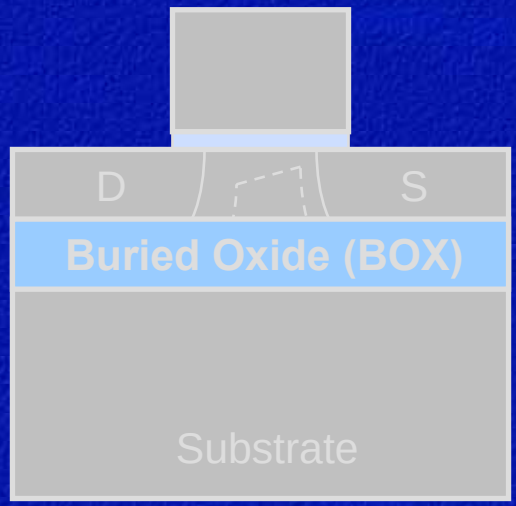
Transistor Structures

Planar Bulk



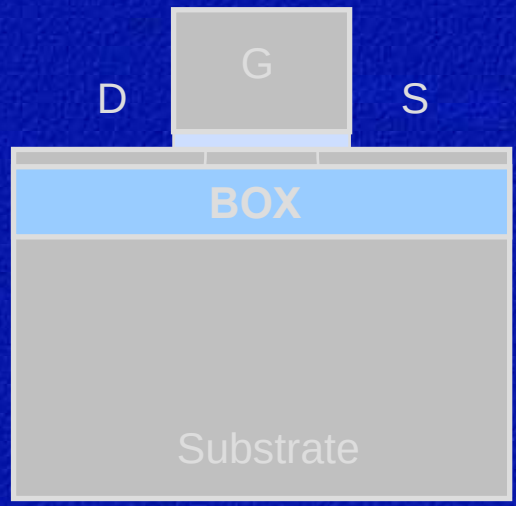
+ Wafer cost / availability
- SCE scaling difficult
- High doping effects and Statistical variation
- Parasitic junction capacitance

Partially Depleted SOI



+ Lower junction cap
+ F.B. performance boost
- F.B. history effect
- SCE scaling difficult
- Wafer cost/availability

Fully Depleted SOI

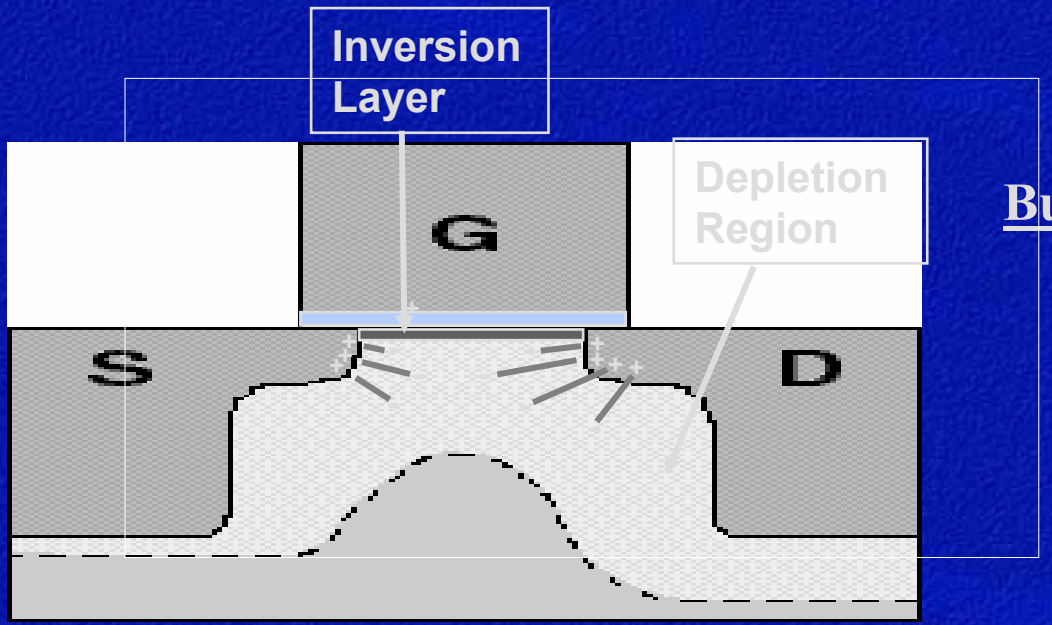


+ Lower junction cap
- SCE scaling difficult
- High $R_{series,s/d}$ → raised S/D
- Sensitivity to Si thickness (very thin)
- Wafer cost/availability

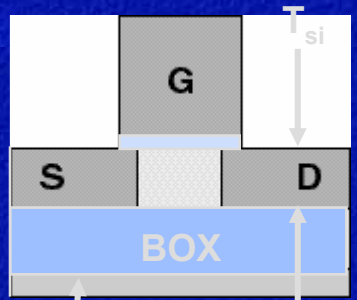
References:

1. P. Zeitzoff, J. Hutchby and H. Huff, to be pub. in Internat. Jour. Of High Speed Electronics and Systems
2. Mark Bohr, ECS Meeting PV **2001-2**, Spring, 2001

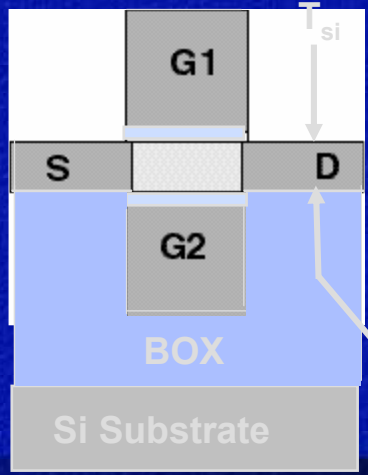
Schematic cross section of planar bulk, UTB SOI, and DG SOI MOSFET



Bulk MOSFET



Ultra-Thin Body SOI



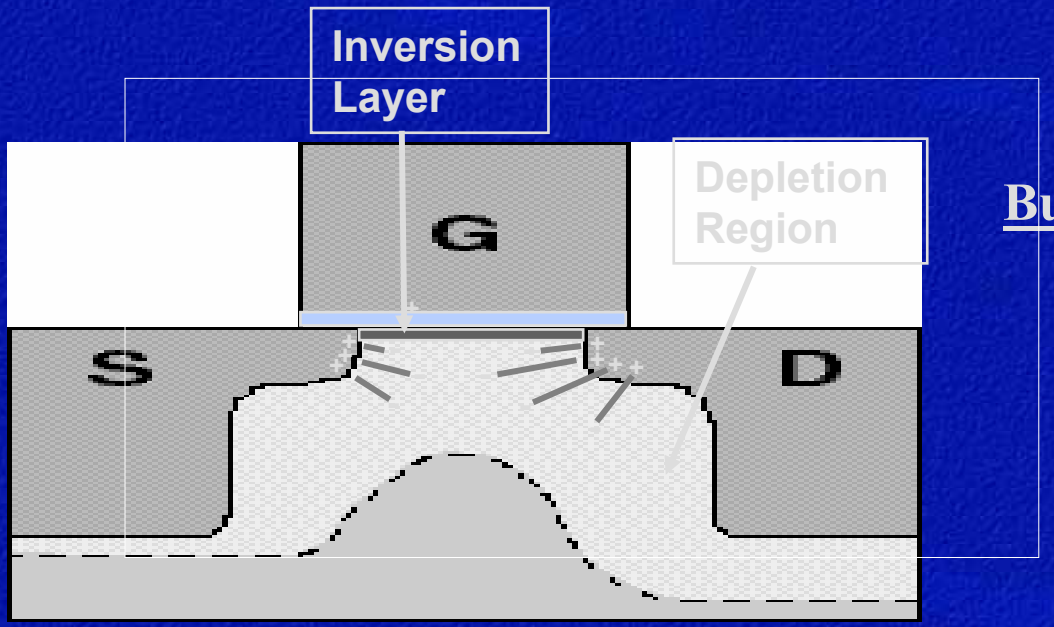
Double-Gate SOI MOSFET

Si Substrate

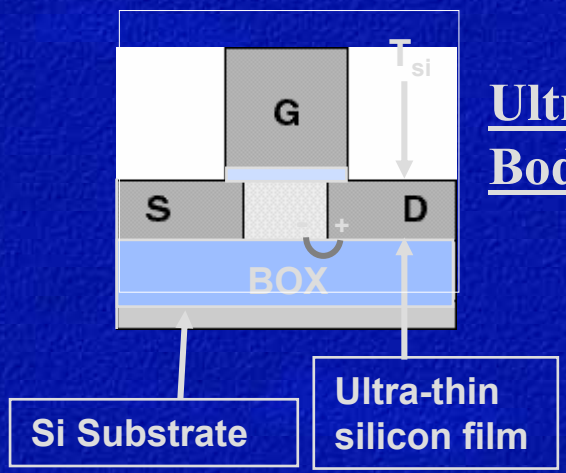
Ultra-thin silicon film

Ultra-thin silicon film

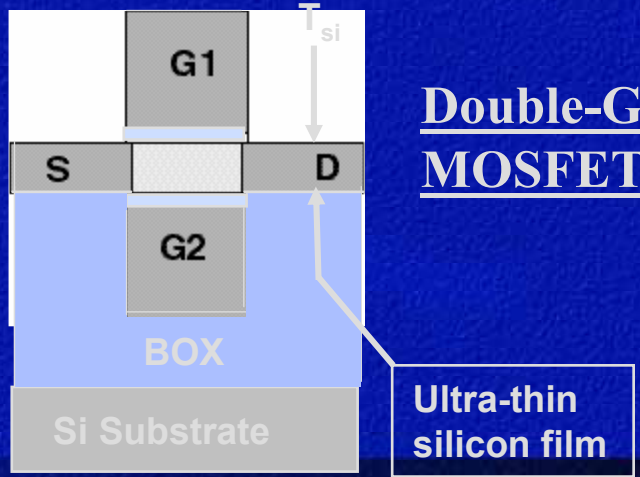
Schematic cross section of planar bulk, UTB SOI, and DG SOI MOSFET



Bulk MOSFET

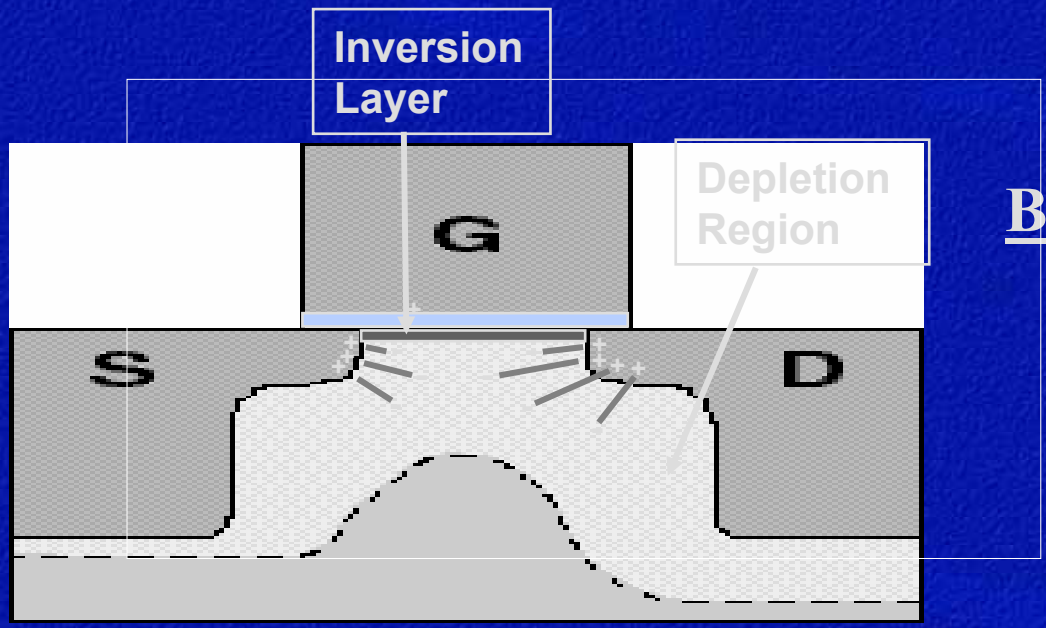


Ultra-Thin Body SOI

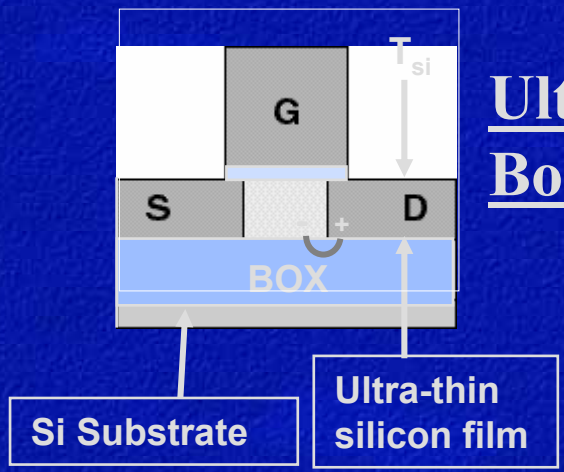


Double-Gate SOI MOSFET

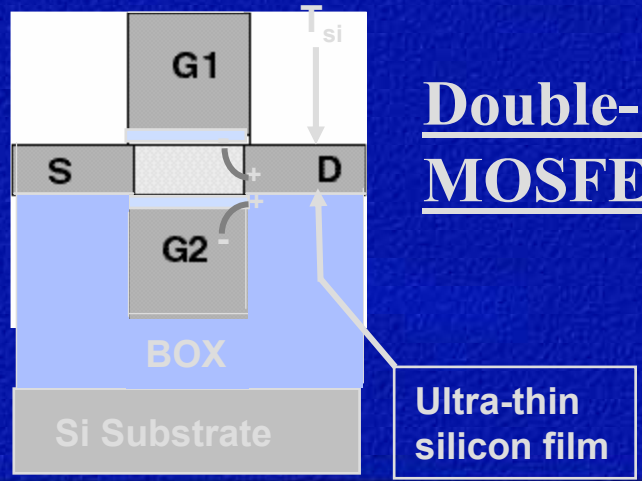
Schematic cross section of planar bulk, UTB SOI and DG SOI MOSFET



Bulk MOSFET



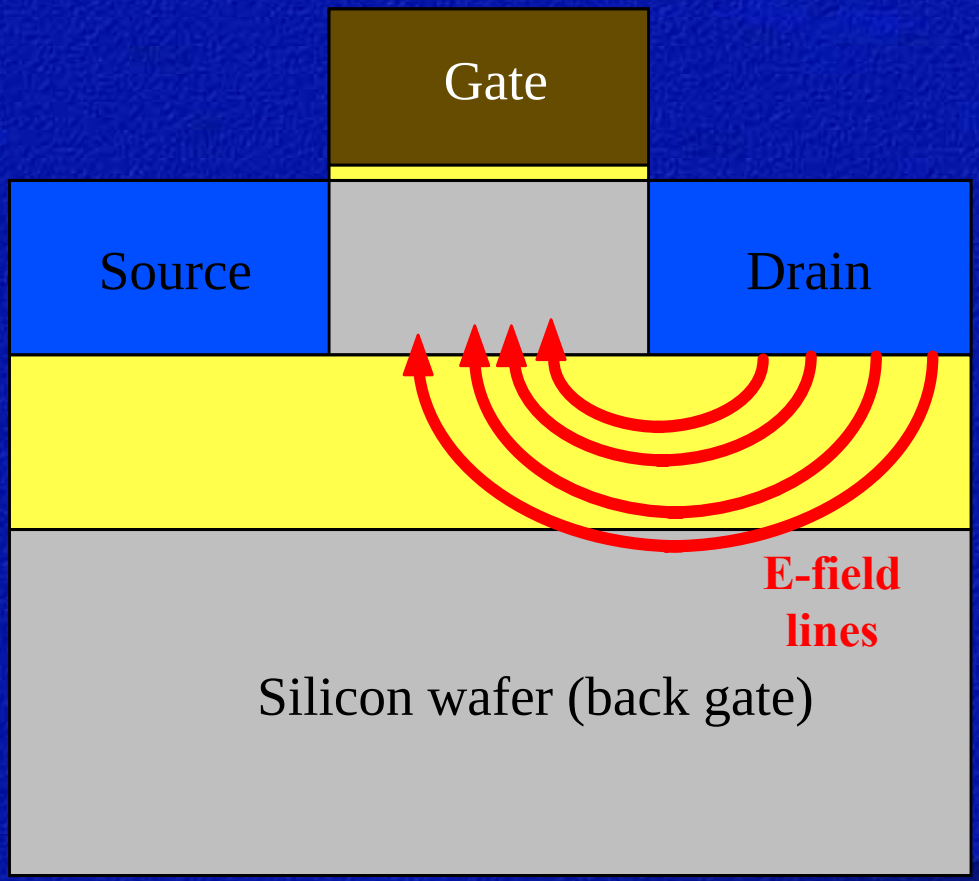
Ultra-Thin Body SOI



Double-Gate SOI MOSFET

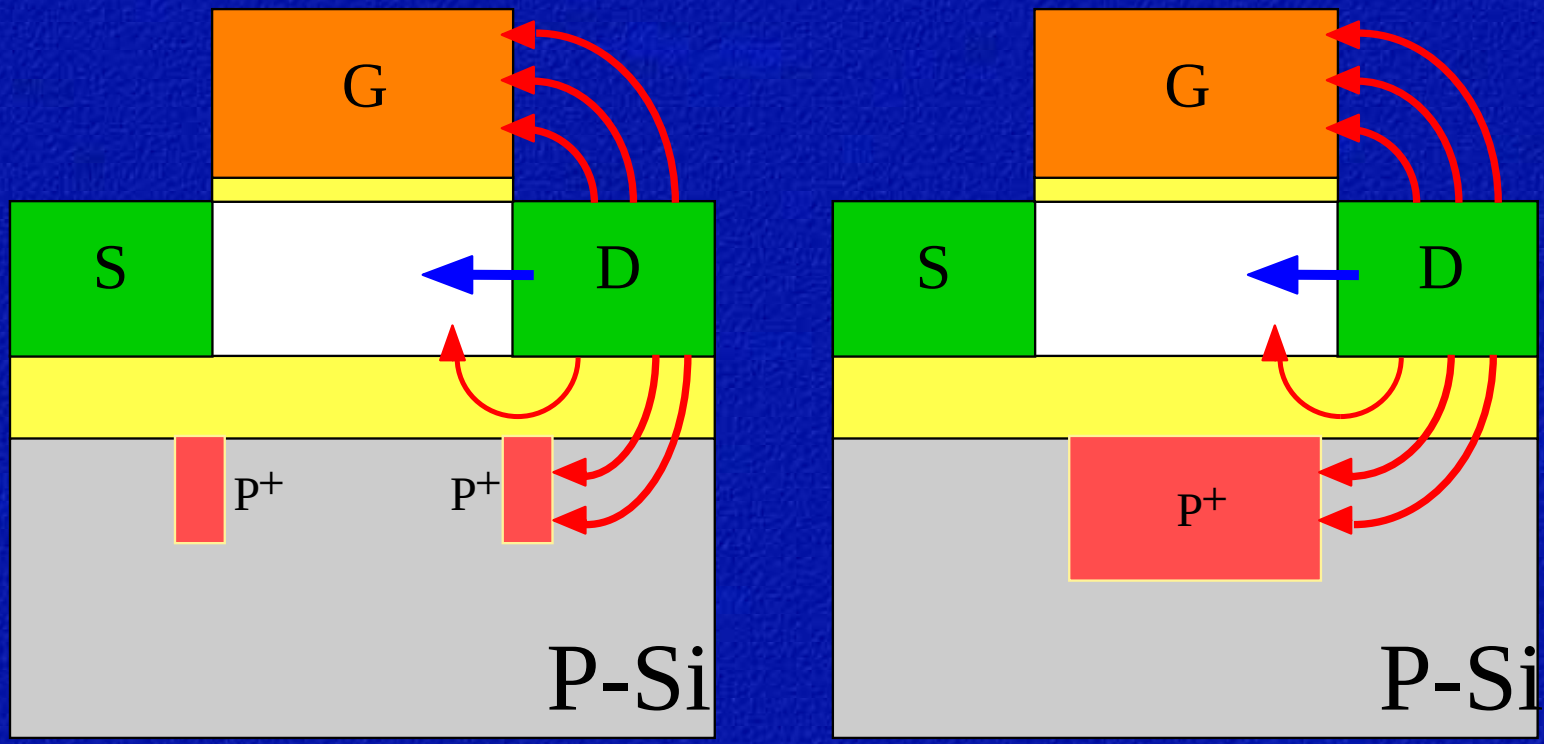
Short-channel MOSFET: DIBL

or: Drain-Induced Barrier Lowering



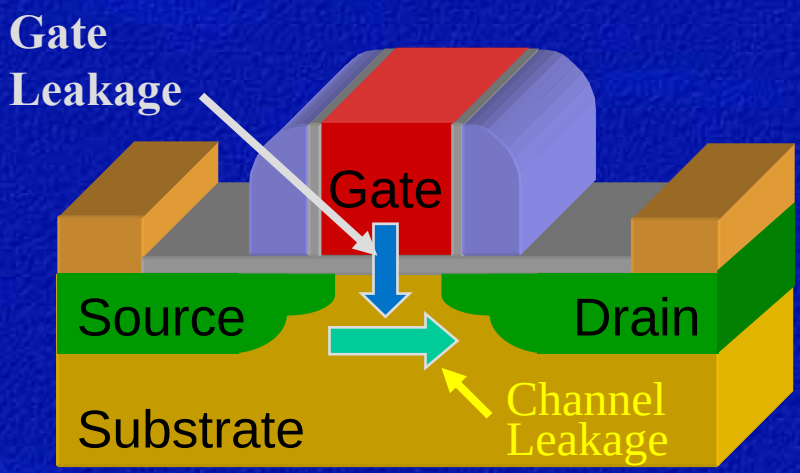
Electric field lines from the drain encroach on the channel region. Any increase of drain voltage decreases the threshold voltage (the "NPN" potential barrier between source and drain is lowered).

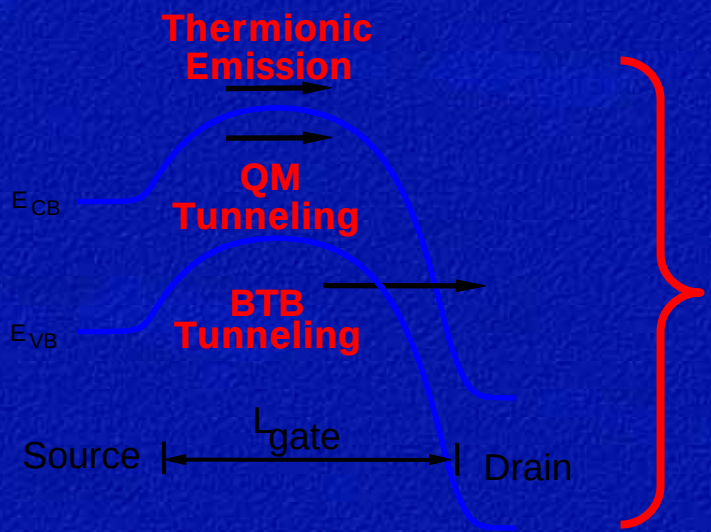
E-Field lines



Ground-plane SOI MOSFETs

Electrostatic Scaling - Channel Leakage (I_{off})





Sum = I_{off}

Channel Leakage