The Technology Roadmap

ECE 260B / CSE 241A Guest Lecture

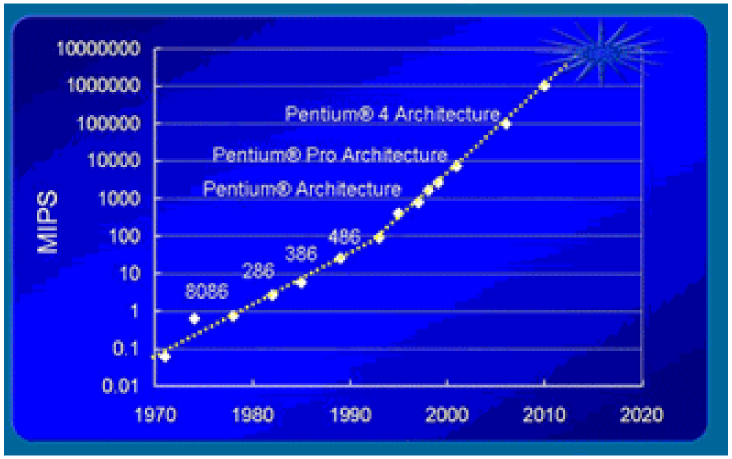
Andrew B. Kahng

Professor of CSE and ECE, UC San Diego

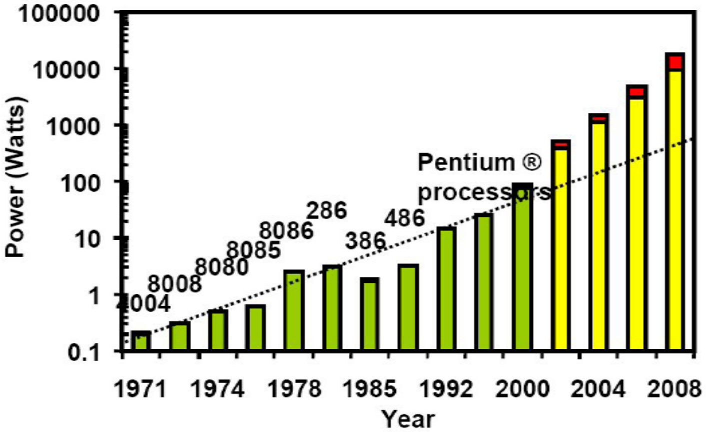
abk@ucsd.edu

http://vlsicad.ucsd.edu/

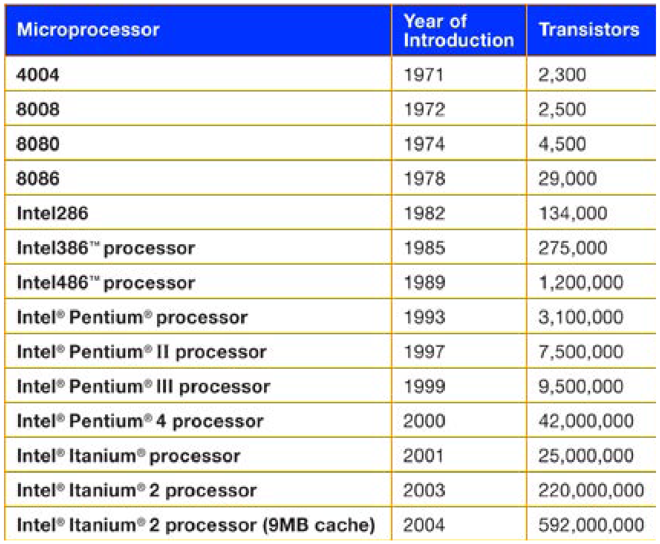
Semiconductor Technology Trends



Performance

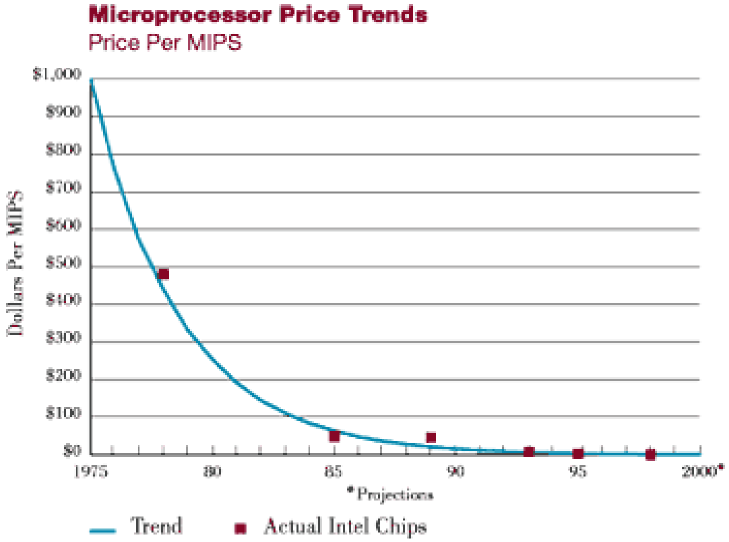


Power



Microprocessor	Year of Introduction	Transistors
4004	1971	2,300
8008	1972	2,500
8080	1974	4,500
8086	1978	29,000
Intel286	1982	134,000
Intel386™ processor	1985	275,000
Intel486™ processor	1989	1,200,000
Intel® Pentium® processor	1993	3,100,000
Intel® Pentium® II processor	1997	7,500,000
Intel® Pentium® III processor	1999	9,500,000
Intel® Pentium® 4 processor	2000	42,000,000
Intel® Itanium® processor	2001	25,000,000
Intel® Itanium® 2 processor	2003	220,000,000
Intel® Itanium® 2 processor (9MB cache)	2004	592,000,000

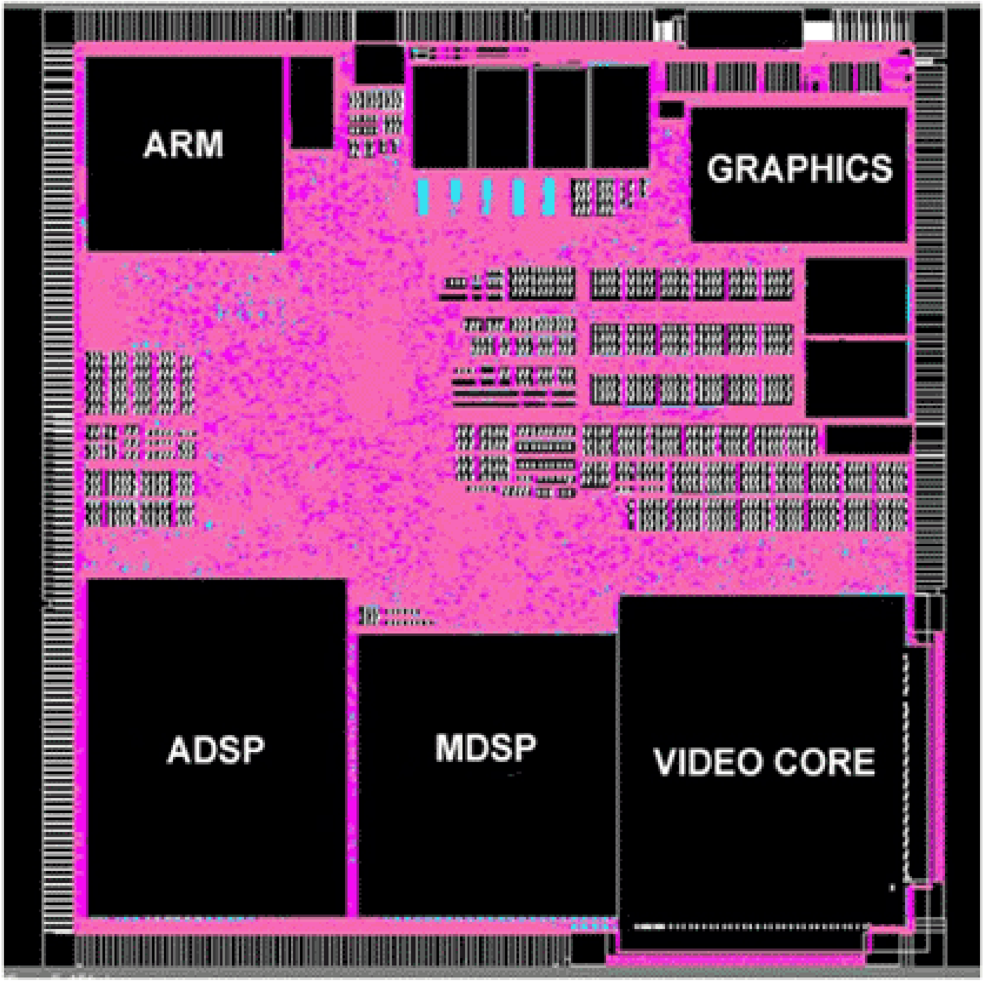
Integration



Cost

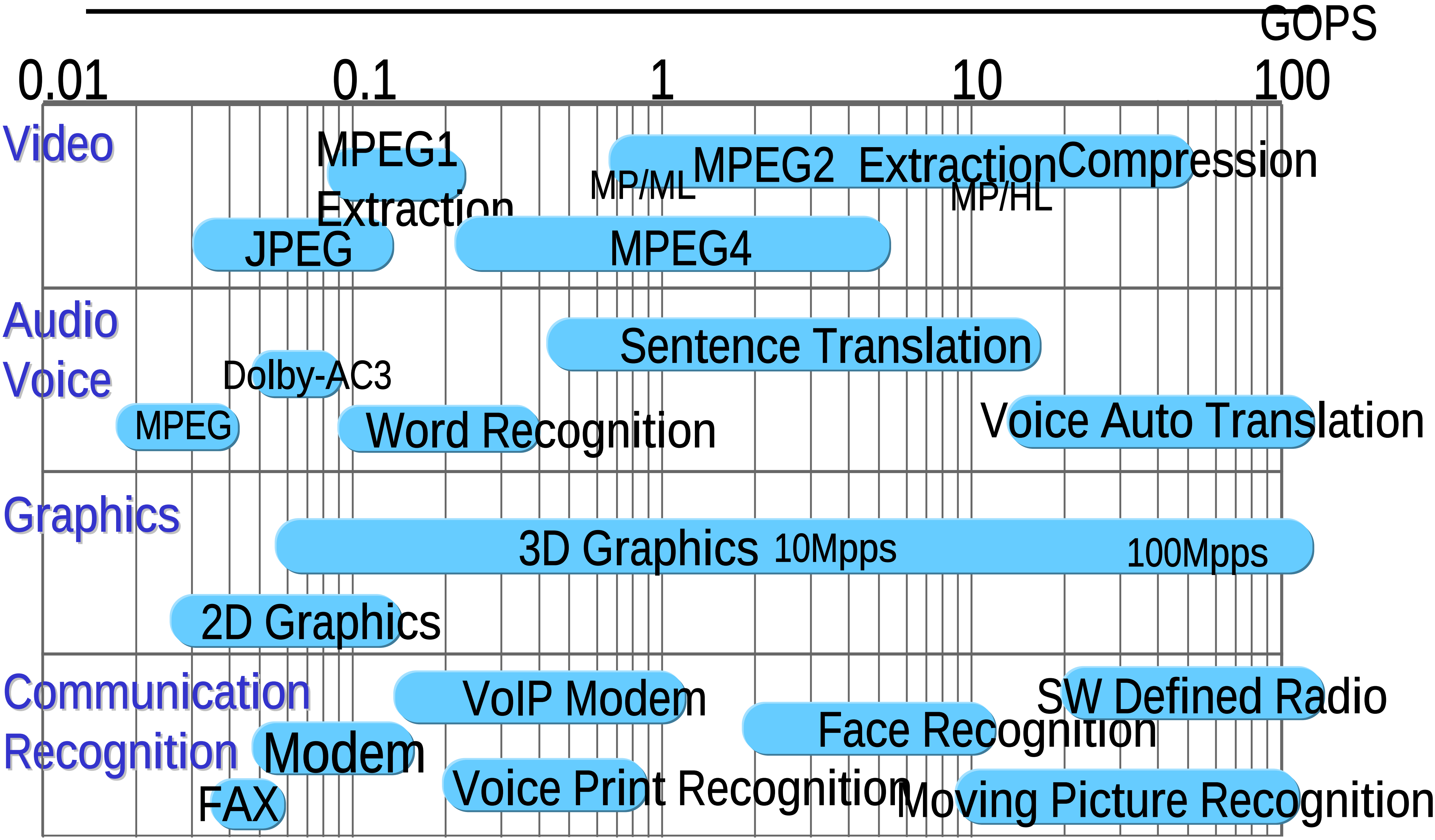
Figures courtesy Intel

What Drives Semiconductor Technology?



Modern cellphone chip: 2+ processors, modem, graphics and video engines, DSPs in 8mm x 8mm

What Does the IC Do?

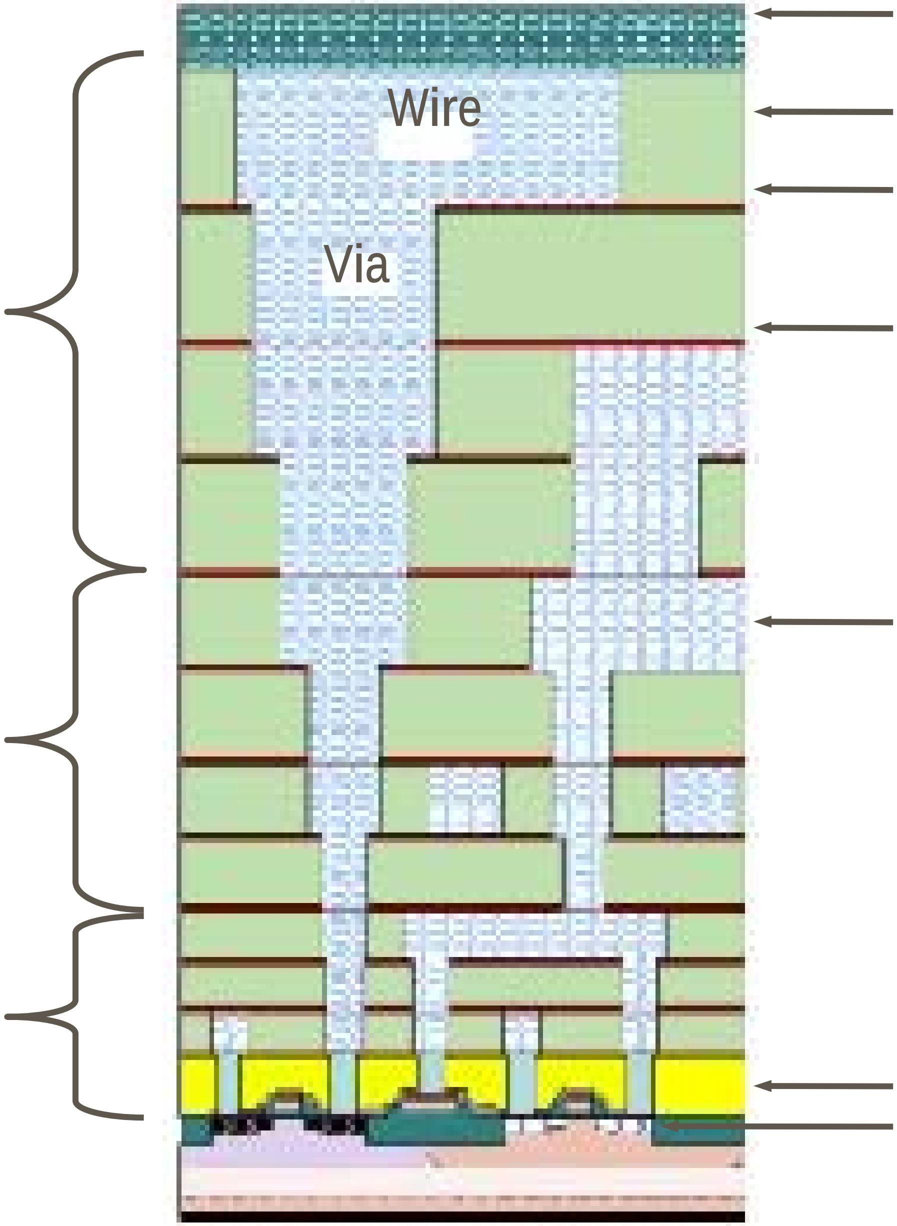


Required performance for multimedia processing (GOPS: Giga Operations Per Sec)

2007 ITRS SOC Consumer-Stationary Driver: 220 TFlops on a single chip by 2022

How Is It Connected?

SEMATECH Prototype BEOL (“back end of the line”) metal stack, 2000



Passivation

Dielectric

Etch Stop Layer

Dielectric Capping Layer

Copper Conductor with
Barrier/Nucleation Layer

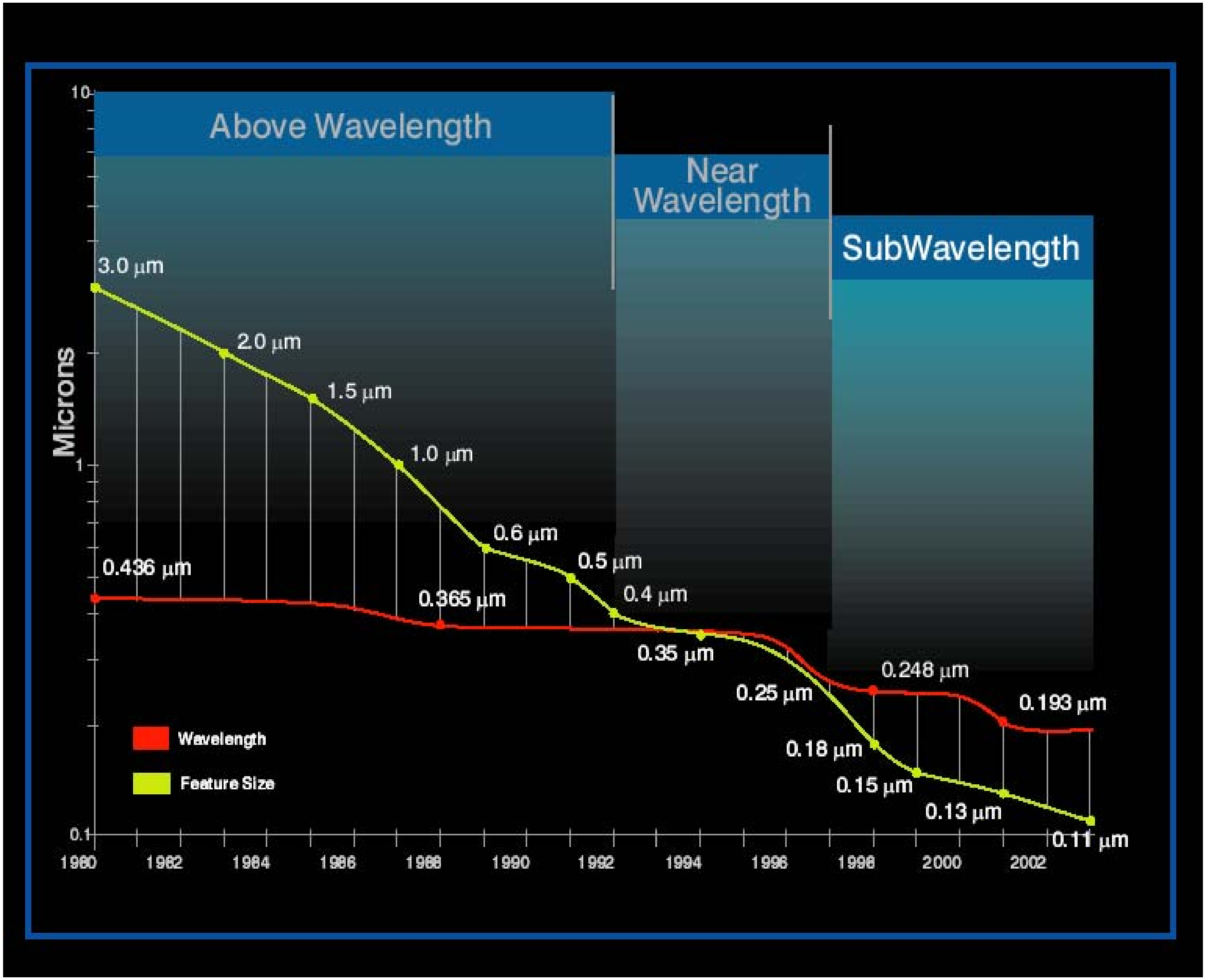
Pre Metal Dielectric
Tungsten Contact Plug

Global (up to 5)

Intermediate (up to 4)

Local (2)

How Is It Manufactured?

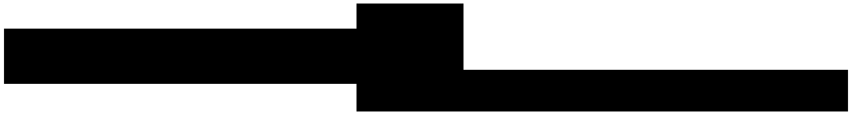


- **Sub-wavelength optical lithography**

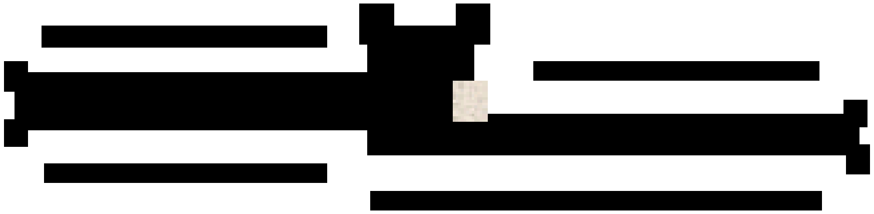
Slide courtesy of Numerical Technologies, Inc.

(Mask Shapes Used in Lithography)

Desired Pattern on wafer

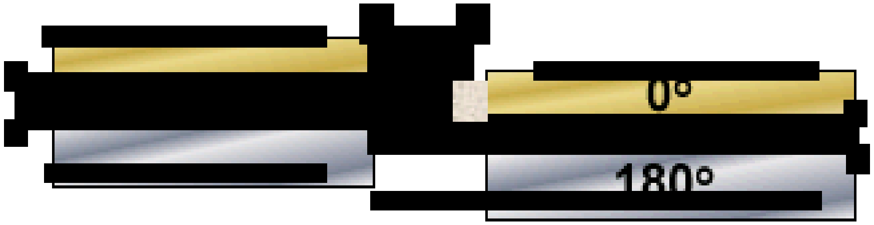


Actual Mask Pattern

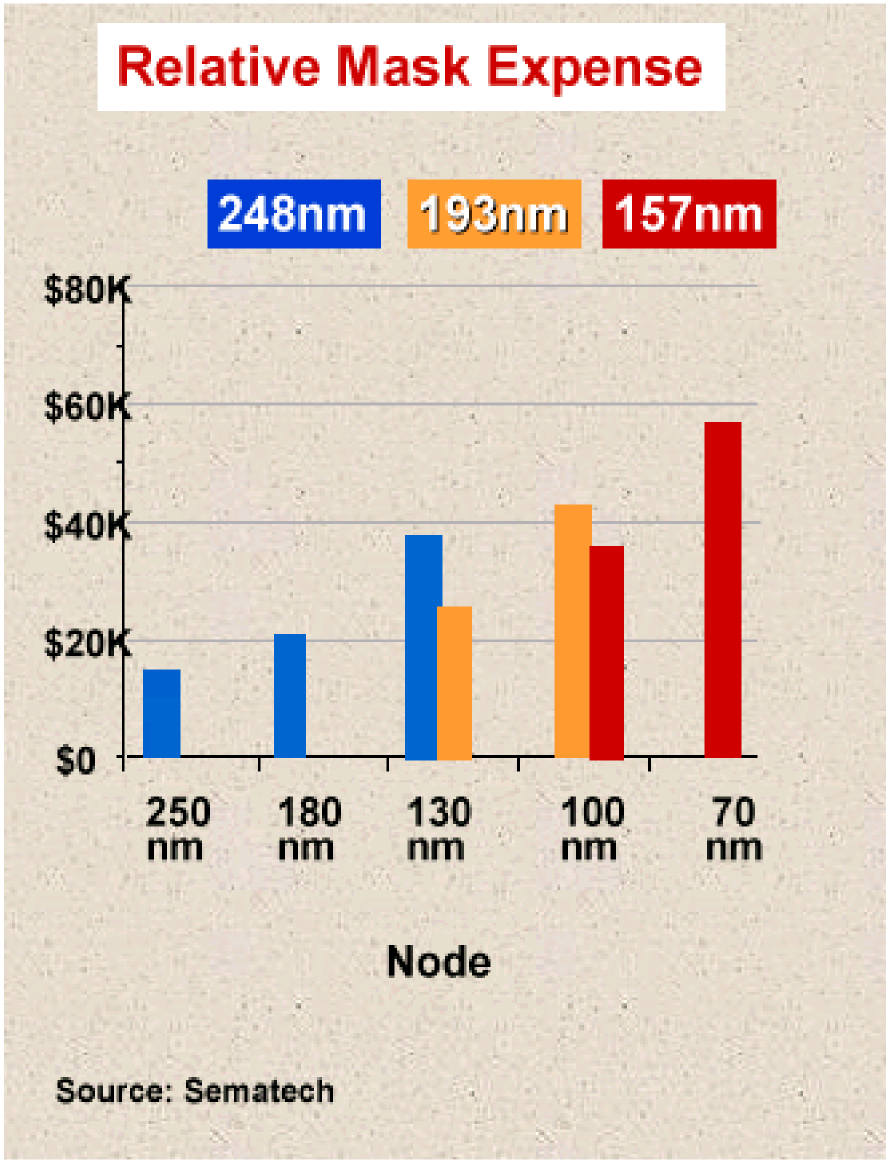


OPC Optical Proximity Correction

Multilevel Mask



PSM Phase Shift Mask



Source: Sematech

Many Interesting Technology Trends

- **Lithography**
  - Minimum feature size scales by 0.7x every three (two?) years
  - Add another pair of layers: last generation's chip = this generation's module
- **Interconnect delay doesn't scale well**
  - Dominates system performance
  - Coupling gets worse → timing uncertainty and design guardband
- **Multiple clock cycles needed to cross chip**
  - whether 3 or 15 not as important as "multiple" being > 1
- **How does manufacturing process enter into picture?**
  - Lower-permittivity dielectrics → organics to aerogels to air gaps
  - Copper interconnects → resistivity, reliability
  - Planarization → more layers are stackable

Many Interesting Design Challenges Result

- **Manufacturability (chip can't be built)**
  - antenna rules
  - minimum area rules for stacked vias
  - CMP (chemical mechanical polishing) area fill rules
  - layout corrections for optical proximity effects in subwavelength lithography
- **Signal integrity (chip fails timing constraints)**
  - crosstalk induced errors
  - timing dependence on crosstalk
  - IR drop on power supplies
- **Reliability (chip fails in the field)**
  - electromigration on power supplies
  - hot carrier effects on devices
  - wire self-heating effects on clocks and signals

Slide courtesy of Dr. Lou Scheffer, Cadence

SRC* Grand Challenges (~2005)

1. Extend CMOS to its ultimate limit
2. Support continuation of Moore's Law by providing a knowledge base for CMOS replacement devices
3. Enable Wireless/Telecomm systems by addressing technical barriers in design, test, process, device and packaging technologies
4. Create mixed-domain transistor and device interconnection technologies, architectures, and tools for future microsystems that mitigate the limitations projected by ITRS
5. Search for radical, cost effective post NGL patterning options
6. Provide low-cost environmentally benign IC processes
7. Increase factory capital utilization efficiency through operational modeling
8. Provide design tools and techniques which enhance design productivity and reduce cost for correct, manufacturable and testable SOC's and SOP's
9. Enable low power and low voltage solutions for mobile/battery conserving applications through system and circuit design, test and packaging approaches.
10. Enable very low cost components
11. Provide tools enabling rapid implementation of new system architectures

*** = Semiconductor Research Corporation, which funds a large portion of semiconductor-related U.S. academic research.**
My point: See the big picture!

Today’s Agenda

- What is the semiconductor roadmap?
- Connections game: Why do we care?
- Aspects of the Design roadmap
- Aspects of the System Drivers roadmap and the Overall Roadmap Technology Characteristics (ORTCs)
- More Than Moore

Background

- **Have written the IC physical design roadmap since 1996**
- **Chair / co-chair of U.S. and International Design Technology Working Groups since 2000**
- **Responsible for two chapters in the International Technology Roadmap for Semiconductors (ITRS), http://public.itrs.net/**
  - **Design chapter:** roadmaps for the EDA industry
  - **System Drivers chapter:** roadmaps for product classes that consume high-value silicon and drive semiconductor technology

What is the Semiconductor Roadmap?

- **Something you need to read !**
- **Enabling mechanism for Moore's Law**
  - Synchronizes <u>many</u> industries to "clock" of technology nodes = A Very Big Picture !
  - Lithography, Interconnect, Assembly and Packaging, Test, Design, …
- **<u>Technology</u> roadmap (not business roadmap)**
- **Structured as <u>requirements</u> + <u>potential solutions</u>**
- **Highly complex and interconnected**
  - 1000+ people worldwide produce new edition each odd-numbered year, and update in even
  - Many contradictions (predict vs. require, etc.)

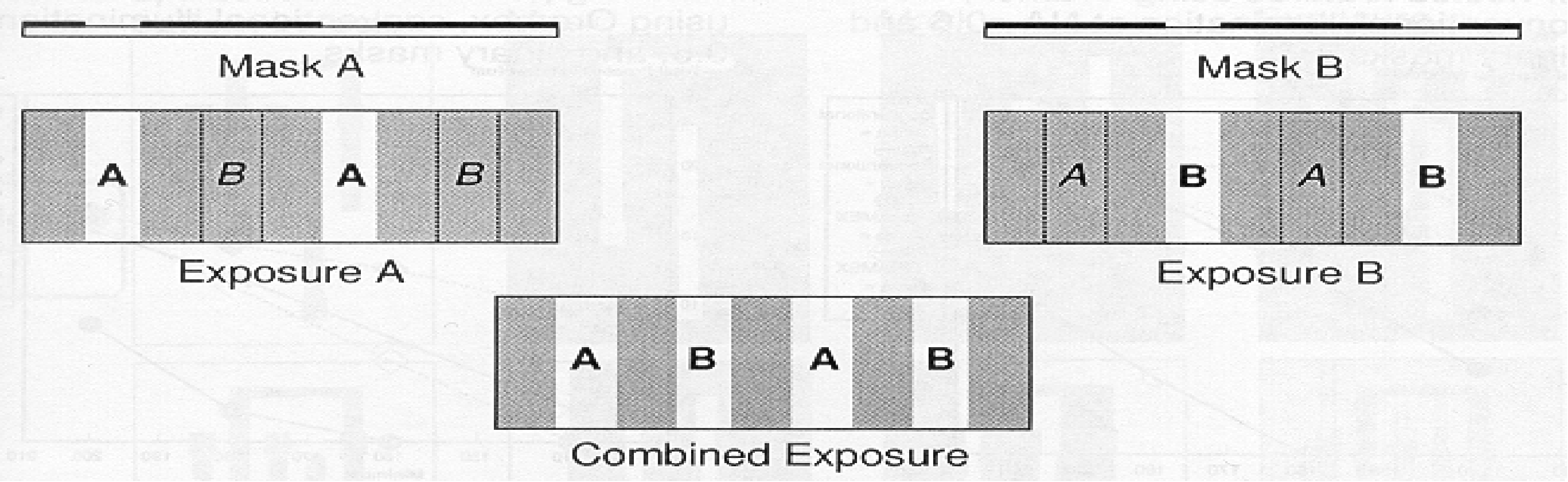
Today’s Agenda

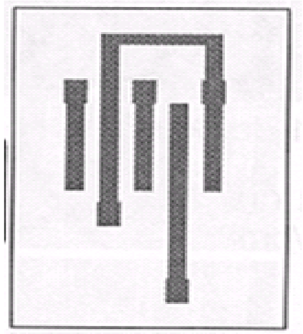
- What is the semiconductor roadmap?
- Connections game: Why do we care?
- Aspects of the Design roadmap
- Aspects of the System Drivers roadmap and the Overall Roadmap Technology Characteristics (ORTCs)
- More Than Moore

Lithography Roadmap (January 2009)

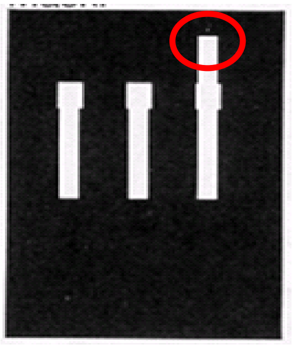
Year of Production	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*
DRAM ½ pitch (nm)	**52**	**45**	**40**	**36**	**32**	**28**	**25**
CD control (3 sigma) (nm) [B]	**5.4**	**4.7**	**4.2**	**3.7**	**3.3**	**2.9**	**2.6**
Contact in resist (nm)	**57**	**50**	**44**	**39**	**35**	**31**	**28**
Contact after etch (nm)	**52**	**45**	**40**	**36**	**32**	**28**	**25**
Overlay [A] (3 sigma) (nm)	**10.3**	**9.0**	**8.0**	**7.1**	**6.4**	**5.7**	**5.1**
Flash							
Flash ½ pitch (nm) (un-contacted poly)	**40**	**36**	**32**	**28**	**25**	**23**	**20**
CD control (3 sigma) (nm) [B]	**4.2**	**3.7**	**3.3**	**2.9**	**2.6**	**2.3**	**2.1**
Contact in resist (nm)	**44**	**39**	**35**	**31**	**28**	**25**	**22**
Contact after etch (nm)	**40**	**36**	**32**	**28**	**25**	**23**	**20**
Overlay [A] (3 sigma) (nm)	**13.2**	**11.8**	**10.5**	**9.4**	**8.3**	**7.4**	**6.6**
MPU							
MPU/ASIC Metal 1 (M1) ½ pitch (nm)	**52**	**45**	**40**	**36**	**32**	**28**	**25**
MPU gate in resist (nm)	**41**	**35**	**31**	**28**	**25**	**22**	**20**
*MPU physical gate length (nm) **	**29**	**27**	**24**	**22**	**18**	**17**	**15**
*Gate CD control (3 sigma) (nm) [B] ***	**3.0**	**2.8**	**2.5**	**2.3**	**1.9**	**1.7**	**1.6**
Contact in resist (nm)	**64**	**56**	**50**	**44**	**39**	**35**	**31**
Contact after etch (nm)	**58**	**51**	**45**	**40**	**36**	**32**	**28**
Overlay [A] (3 sigma) (nm)	**13**	**11**	**10.0**	**8.9**	**8.0**	**7.1**	**6.3**
Chip size (mm^2)							
Maximum exposure field height (mm)	**26**	**26**	**26**	**26**	**26**	**26**	**26**
Maximum exposure field length (mm)	**33**	**33**	**33**	**33**	**33**	**33**	**33**
Maximum field area printed by exposure tool (mm^2)	**858**	**858**	**858**	**858**	**858**	**858**	**858**
Wafer site flatness at exposure step (nm) [C]	**48**	**42**	**37**	**33**	**29**	**26**	**23**
Number of mask levels MPU	**35**	**35**	**35**	**35**	**37**	**37**	**37**
Wafer size (diameter, mm)	**300**	**300**	**300**	**450**	**450**	**450**	**450**

Double Patterning Lithography (DPL)



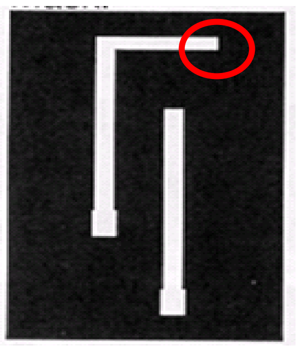


Desired pattern

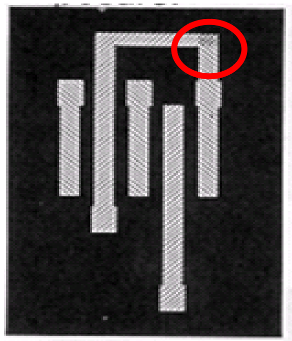


First Mask

+

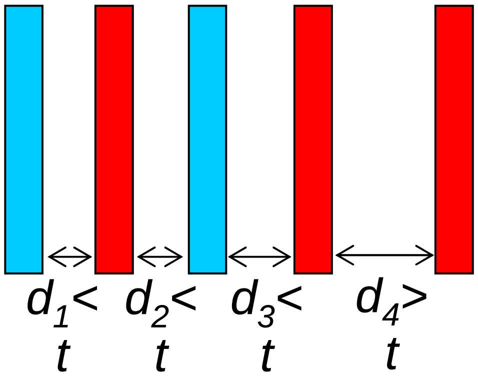


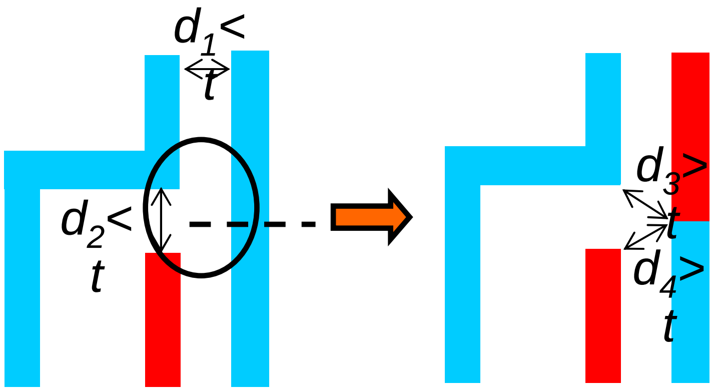
Second Mask



Combined exposure

DPL Layout Decomposition

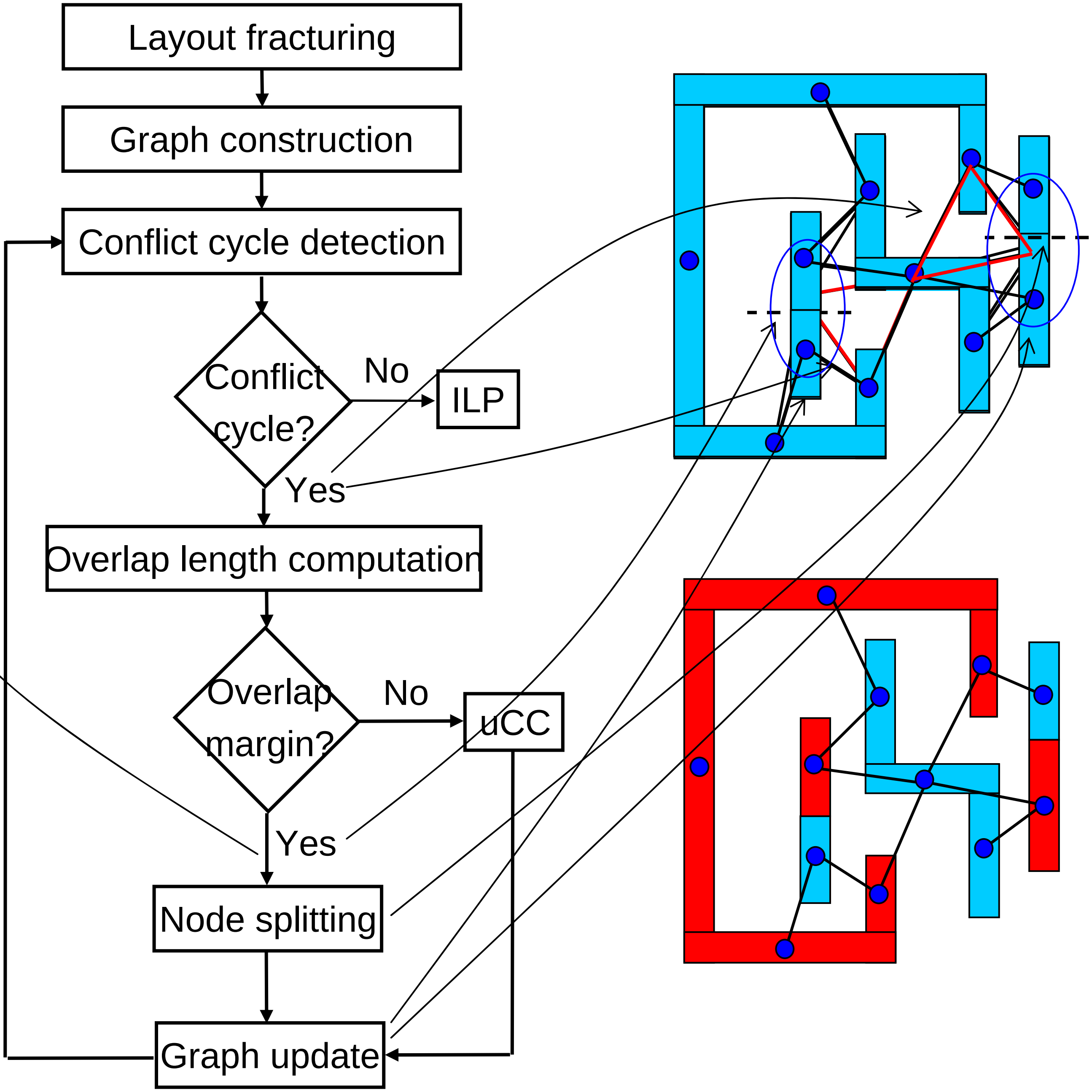




- **Two features are assigned opposite colors if their spacing is less than the minimum coloring spacing *t***
- **IF two features within minimum coloring spacing *t* cannot be assigned different colors**
  - THEN at least one feature is split into two or more parts
- **Pattern split increases manufacturing cost, complexity**
  - Line ends → corner rounding
  - Overlay error and interference mismatch → line edge errors → tight overlay control
  - **Optimization: minimize cost of layout decomposition**
  - Various “Graph Bipartization” engines from my group since 1998

Example DPL Layout Decomposition Flow

- **Layout fracturing**
  - Polygons → rectangles
- **Graph construction**
- **Conflict cycle (CC) detection**
- **Overlap length computation**
  - If there is a feasible dividing point → node splitting
  - Otherwise, report an unresolvable conflict cycle (uCC)
- **Graph updating**
- **ILP based DPL color assignment**



Layout fracturing

Graph construction

Conflict cycle detection

Conflict cycle?

No

ILP

Yes

Overlap length computation

Overlap margin?

No

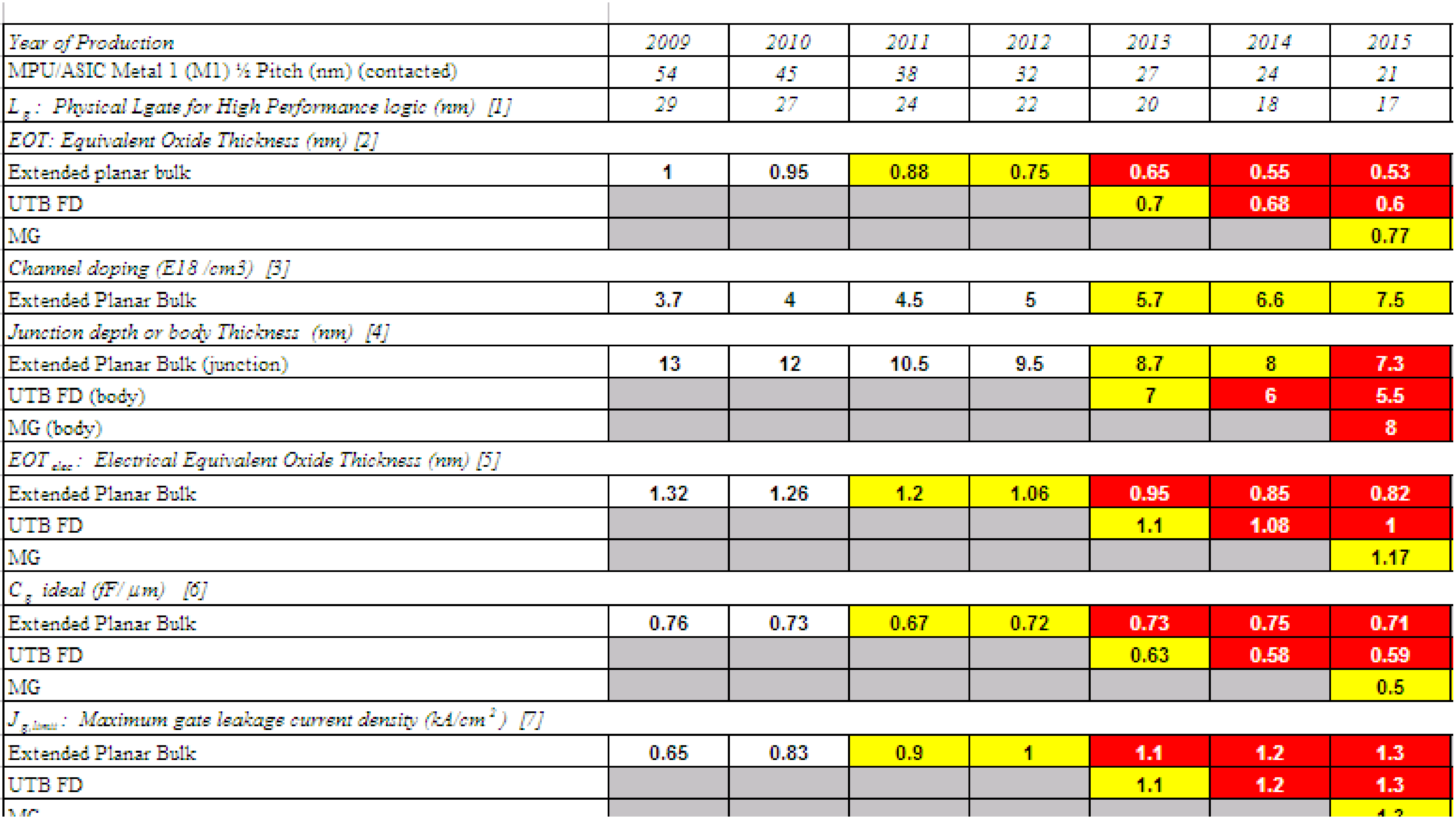
uCC

Yes

Node splitting

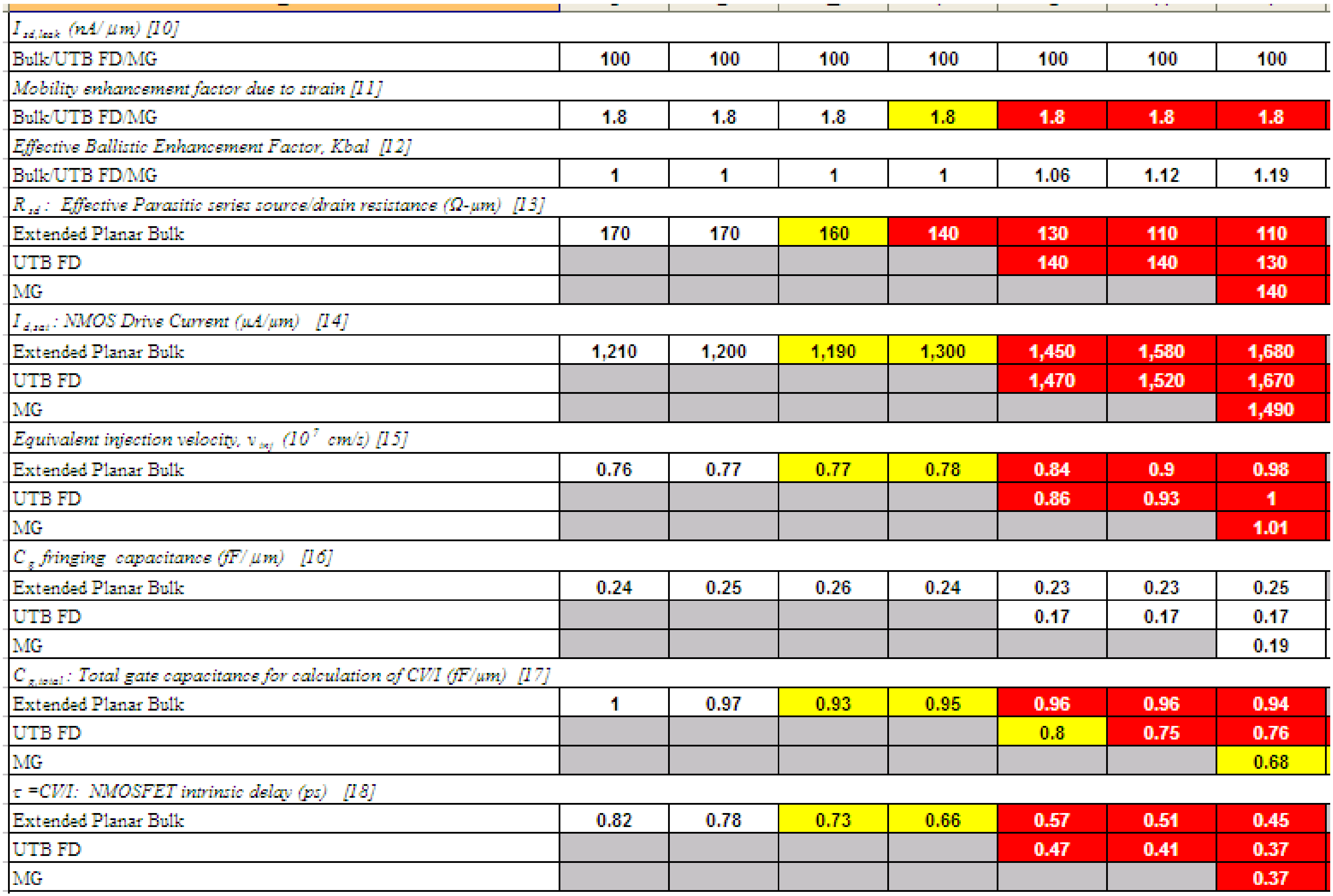
Graph update

Process Integration, Device Structures Roadmap (December 2009) – HIGH PERFORMANCE



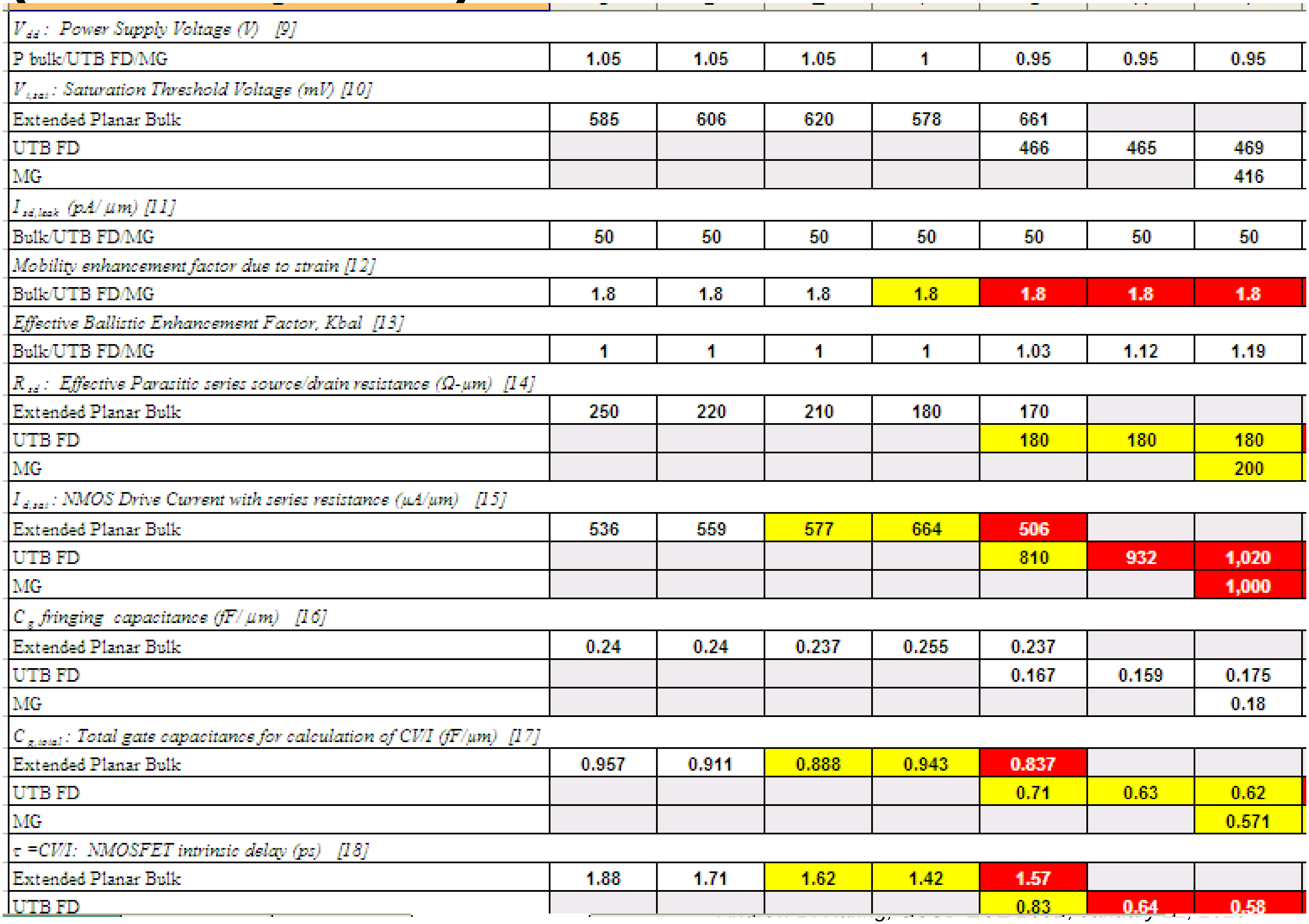
Year of Production	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*
MPU/ASIC Metal 1 (M1) ½ Pitch (nm) (contacted)	*54*	*45*	*38*	*32*	*27*	*24*	*21*
L_g: Physical Lgate for High Performance logic (nm) [1]	*29*	*27*	*24*	*22*	*20*	*18*	*17*
EOT: Equivalent Oxide Thickness (nm) [2]							
Extended planar bulk	1	0.95	0.88	0.75	0.65	0.55	0.53
UTB FD					0.7	0.68	0.6
MG							0.77
Channel doping (E18 /cm3) [3]							
Extended Planar Bulk	3.7	4	4.5	5	5.7	6.6	7.5
Junction depth or body Thickness (nm) [4]							
Extended Planar Bulk (junction)	13	12	10.5	9.5	8.7	8	7.3
UTB FD (body)					7	6	5.5
MG (body)							8
EOT_{elec}: Electrical Equivalent Oxide Thickness (nm) [5]							
Extended Planar Bulk	1.32	1.26	1.2	1.06	0.95	0.85	0.82
UTB FD					1.1	1.08	1
MG							1.17
C_g ideal (fF/ μm) [6]							
Extended Planar Bulk	0.76	0.73	0.67	0.72	0.73	0.75	0.71
UTB FD					0.63	0.58	0.59
MG							0.5
$J_{g,limit}$: Maximum gate leakage current density (kA/cm^2) [7]							
Extended Planar Bulk	0.65	0.83	0.9	1	1.1	1.2	1.3
UTB FD					1.1	1.2	1.3
MG							1.3

Process Integration, Device Structures Roadmap (December 2009) – HIGH PERFORMANCE



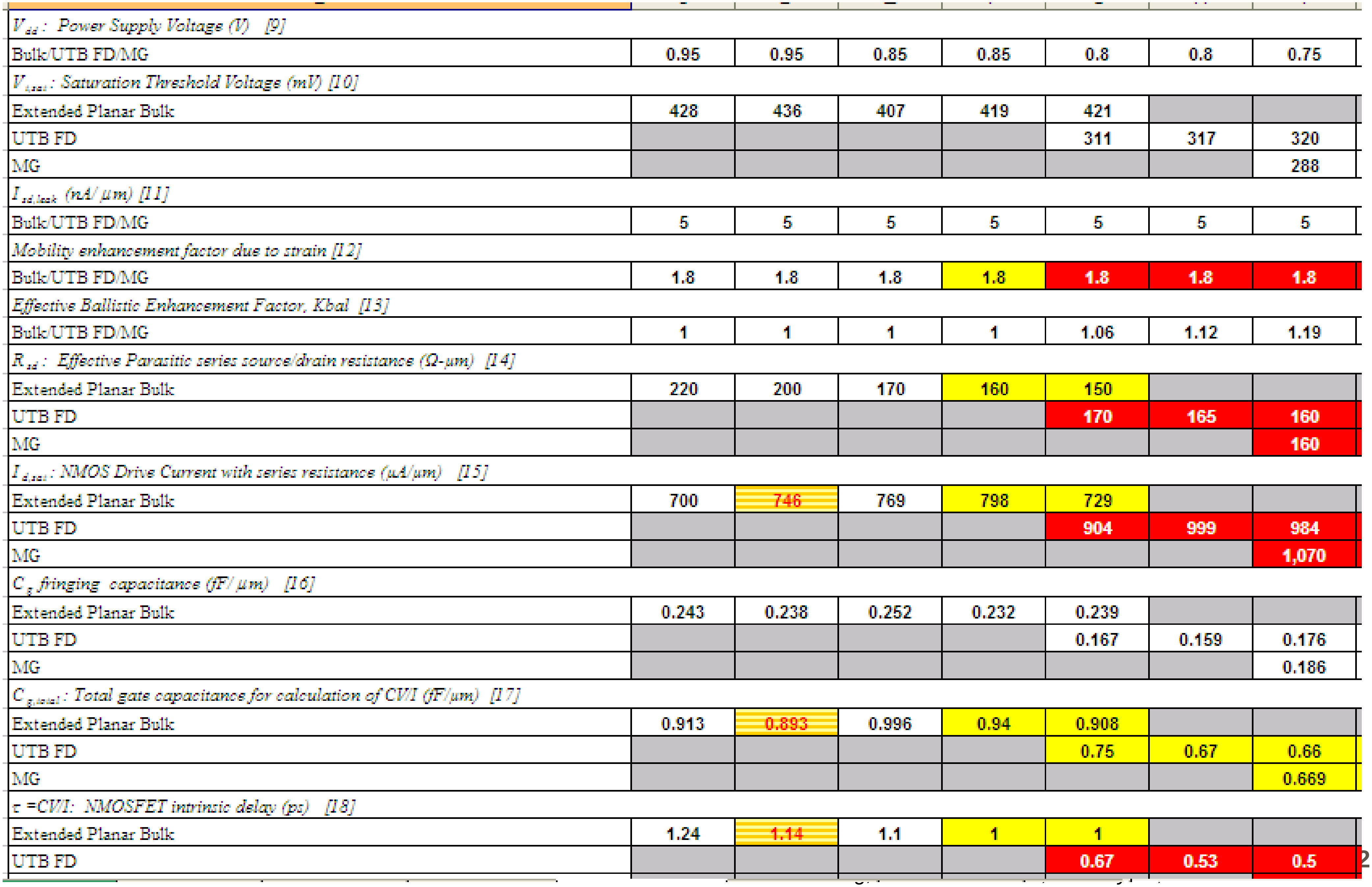
$I_{sd,leak}$ (nA/ µm) [10]							
Bulk/UTB FD/MG	100	100	100	100	100	100	100
Mobility enhancement factor due to strain [11]							
Bulk/UTB FD/MG	1.8	1.8	1.8	1.8	1.8	1.8	1.8
Effective Ballistic Enhancement Factor, Kbal [12]							
Bulk/UTB FD/MG	1	1	1	1	1.06	1.12	1.19
R_{sd} : Effective Parasitic series source/drain resistance (Ω-µm) [13]							
Extended Planar Bulk	170	170	160	140	130	110	110
UTB FD					140	140	130
MG							140
$I_{d,sat}$: NMOS Drive Current (µA/µm) [14]							
Extended Planar Bulk	1,210	1,200	1,190	1,300	1,450	1,580	1,680
UTB FD					1,470	1,520	1,670
MG							1,490
Equivalent injection velocity, v_{inj} (10^7 cm/s) [15]							
Extended Planar Bulk	0.76	0.77	0.77	0.78	0.84	0.9	0.98
UTB FD					0.86	0.93	1
MG							1.01
C_g fringing capacitance (fF/ µm) [16]							
Extended Planar Bulk	0.24	0.25	0.26	0.24	0.23	0.23	0.25
UTB FD					0.17	0.17	0.17
MG							0.19
$C_{g,ideal}$: Total gate capacitance for calculation of CV/I (fF/µm) [17]							
Extended Planar Bulk	1	0.97	0.93	0.95	0.96	0.96	0.94
UTB FD					0.8	0.75	0.76
MG							0.68
τ =CV/I: NMOSFET intrinsic delay (ps) [18]							
Extended Planar Bulk	0.82	0.78	0.73	0.66	0.57	0.51	0.45
UTB FD					0.47	0.41	0.37
MG							0.37

Process Integration, Device Structures Roadmap (December 2009) – LOW STANDBY POWER



V_{dd} : Power Supply Voltage (V) [9]							
P bulk/UTB FD/MG	1.05	1.05	1.05	1	0.95	0.95	0.95
$V_{t,sat}$: Saturation Threshold Voltage (mV) [10]							
Extended Planar Bulk	585	606	620	578	661		
UTB FD					466	465	469
MG							416
$I_{sd,leak}$ (pA/ µm) [11]							
Bulk/UTB FD/MG	50	50	50	50	50	50	50
Mobility enhancement factor due to strain [12]							
Bulk/UTB FD/MG	1.8	1.8	1.8	1.8	1.8	1.8	1.8
Effective Ballistic Enhancement Factor, Kbal [13]							
Bulk/UTB FD/MG	1	1	1	1	1.03	1.12	1.19
R_{sd} : Effective Parasitic series source/drain resistance (Ω-µm) [14]							
Extended Planar Bulk	250	220	210	180	170		
UTB FD					180	180	180
MG							200
$I_{d,sat}$: NMOS Drive Current with series resistance (µA/µm) [15]							
Extended Planar Bulk	536	559	577	664	506		
UTB FD					810	932	1,020
MG							1,000
C_g fringing capacitance (fF/ µm) [16]							
Extended Planar Bulk	0.24	0.24	0.237	0.255	0.237		
UTB FD					0.167	0.159	0.175
MG							0.18
$C_{g,total}$: Total gate capacitance for calculation of CV/I (fF/µm) [17]							
Extended Planar Bulk	0.957	0.911	0.888	0.943	0.837		
UTB FD					0.71	0.63	0.62
MG							0.571
τ =CV/I: NMOSFET intrinsic delay (ps) [18]							
Extended Planar Bulk	1.88	1.71	1.62	1.42	1.57		
UTB FD					0.83	0.64	0.58

Process Integration, Device Structures Roadmap (December 2009) – LOW OPERATING POWER



V_{dd}: Power Supply Voltage (V) [9]							
Bulk/UTB FD/MG	0.95	0.95	0.85	0.85	0.8	0.8	0.75
$V_{t,sat}$: Saturation Threshold Voltage (mV) [10]							
Extended Planar Bulk	428	436	407	419	421		
UTB FD					311	317	320
MG							288
$I_{sd,leak}$ (nA/ µm) [11]							
Bulk/UTB FD/MG	5	5	5	5	5	5	5
Mobility enhancement factor due to strain [12]							
Bulk/UTB FD/MG	1.8	1.8	1.8	1.8	1.8	1.8	1.8
Effective Ballistic Enhancement Factor, Kbal [13]							
Bulk/UTB FD/MG	1	1	1	1	1.06	1.12	1.19
R_{sd}: Effective Parasitic series source/drain resistance (Ω-µm) [14]							
Extended Planar Bulk	220	200	170	160	150		
UTB FD					170	165	160
MG							160
$I_{d,sat}$: NMOS Drive Current with series resistance (µA/µm) [15]							
Extended Planar Bulk	700	746	769	798	729		
UTB FD					904	999	984
MG							1,070
C_g fringing capacitance (fF/ µm) [16]							
Extended Planar Bulk	0.243	0.238	0.252	0.232	0.239		
UTB FD					0.167	0.159	0.176
MG							0.186
$C_{g,total}$: Total gate capacitance for calculation of CV/I (fF/µm) [17]							
Extended Planar Bulk	0.913	0.893	0.996	0.94	0.908		
UTB FD					0.75	0.67	0.66
MG							0.669
τ =CV/I: NMOSFET intrinsic delay (ps) [18]							
Extended Planar Bulk	1.24	1.14	1.1	1	1		
UTB FD					0.67	0.53	0.5

Comments

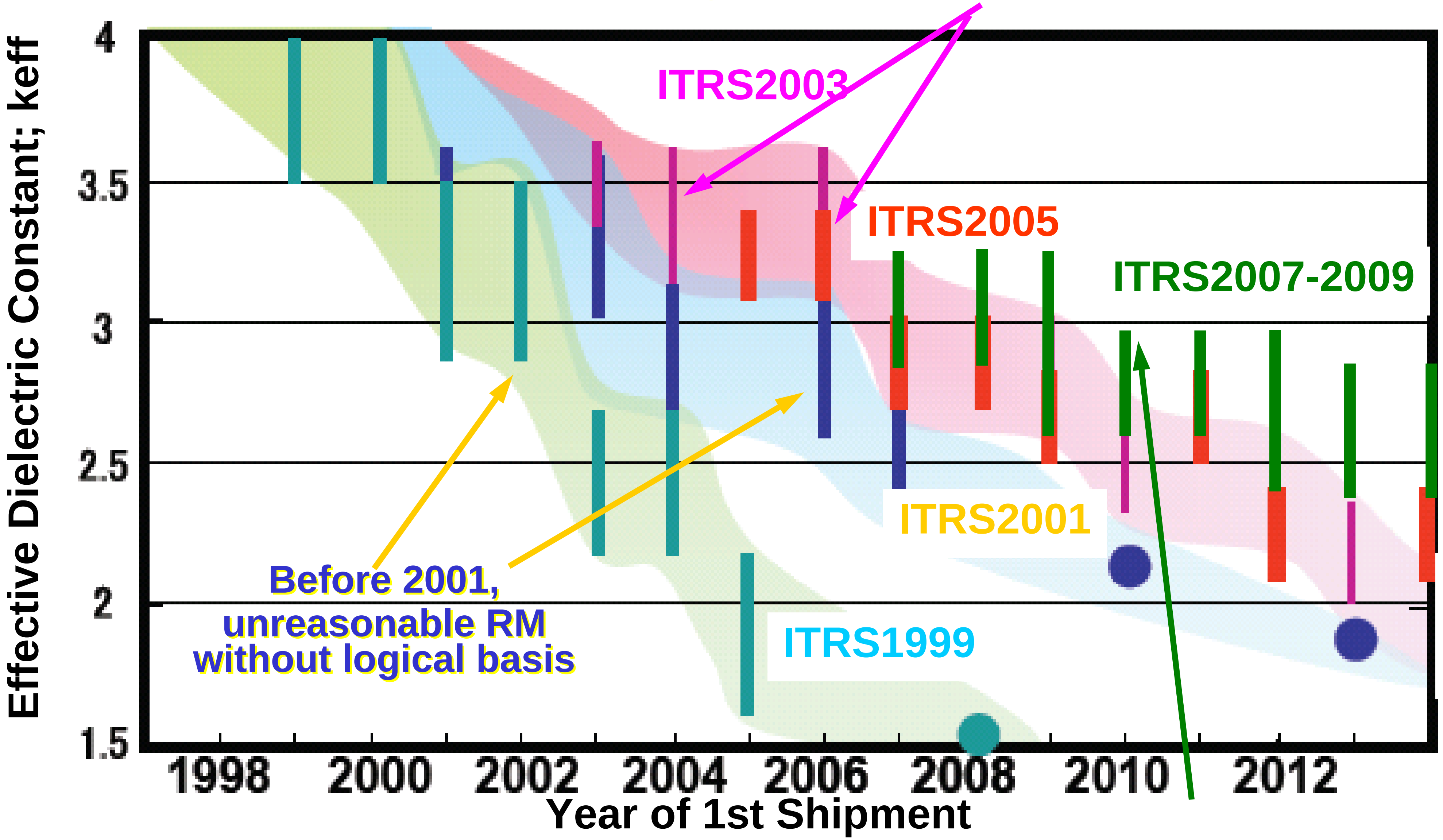
- **LSTP subthreshold leakage requirement of 50 pA/μm used to be 1 pA/μm in early 2000's !**
- **HP scaling of CV/I is now 13%/year, instead of historical 17%/year, based on Design input that the extra speed wasn't usable because of power limits**
- **HP, LSTP correspond to G and LP process flavors from major foundries**
- **2009 LOP roadmap *increased* VDD especially in long-term years; this is wrong from design and product viewpoint, and is likely to be corrected in 2010**
  - LOP roadmap might also go away in light of previous comment

Interconnect Roadmap (January 2009)

Year of Production	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*
MPU/ASIC Metal 1 ½ Pitch (nm)(contacted)	*52*	*45*	*40*	*36*	*32*	*28*
Number of metal levels (includes ground planes & passive devices)	**12**	**12**	**12**	**12**	**13**	**13**
Total interconnect length (m/cm^2) – Metal 1 and five intermediate levels, active wiring only [1]	**2000**	**2222**	**2500**	**2857**	**3125**	**3571**
FITs/m length/cm^2 × 10^{-3} excluding global levels [2]	**2.5**	**2.3**	**2**	**1.8**	**1.6**	**1.4**
Interlevel metal insulator – effective dielectric constant (κ)	**2.6-2.9**	**2.6-2.9**	**2.6-2.9**	**2.4-2.8**	**2.4-2.8**	**2.4-2.8**
Interlevel metal insulator – bulk dielectric constant (κ)	**2.3-2.6**	**2.3-2.6**	**2.3-2.6**	**2.1-2.4**	**2.1-2.4**	**2.1-2.4**
Copper diffusion barrier and etch stop – bulk dielectric constant (κ)	**3.5-4.0**	**3.5-4.0**	**3.5-4.0**	**3.0-3.5**	**3.0-3.5**	**3.0-3.5**
Metal 1 wiring pitch (nm)	**104**	**90**	**80**	**72**	**64**	**56**
Metal 1 A/R (for Cu)	**1.8**	**1.8**	**1.8**	**1.8**	**1.9**	**1.9**
Barrier/cladding thickness (for Cu Metal 1 wiring) (nm) [3]	**3.7**	**3.3**	**2.9**	**2.6**	**2.4**	**2.1**
Cu thinning at minimum pitch due to erosion (nm), 10% × height, 50% areal density, 500 µm square array	**9**	**8**	**7**	**6**	**6**	**5**
Conductor effective resistivity (µΩ cm) Cu Metal 1 wiring including effect of width-dependent scattering and a conformal barrier of thickness specified below	**3.80**	**4.08**	**4.30**	**4.53**	**4.83**	**5.20**
Interconnect RC delay (ps) for 1 mm Cu Metal 1 wire, assumes width-dependent scattering and a conformal barrier of thickness specified below	**1465**	**2100**	**2801**	**3491**	**4555**	**6405**
Line length (µm) where 25% of switching voltage is induced on victim Metal 1 wire by crosstalk [4]	**89**	**82**	**78**	**64**	**57**	**49**
Total Metal 1 resistance variability due to CD erosion and scattering (%)	**30**	**30**	**31**	**32**	**32**	**31**
Intermediate wiring pitch (nm)	**104**	**90**	**80**	**72**	**64**	**56**

History: Low-k Roadmap Evolution

Since 2003, based on wiring capacitance calculation of three kinds of dielectric structures and validated against publications



Before 2001, unreasonable RM without logical basis

2009 decreased max bulk k by 0.1 - no significant change on k_{eff} in 2009

Comments

- AR is important
- Thickness control (planarization by CMP) spec implies large interconnect RC variation
- Current processes often have thick-metal on top two layers (above “global”)
- Leading-edge designs (clock, analog) will often “staple” (superpose) traces on multiple layers to reduce resistance
- M1 pitches show that “foundry X nm process” is often not a true X nm process in the ITRS sense – rather, more in a marketing sense

Packaging Roadmap (January 2009)

Year of Production	*2009*	*2010*	*2011*	*2012*	*2013*
Cost per Pin Minimum for Contract Assembly (Cents/Pin)					
Low-cost, hand-held and memory	**.24-.46**	**.23-.44**	**.22-.42**	**.21-.40**	**.20-.38**
Cost-performance	**.63-1.70**	**.60-1.20**	**.57-.97**	**.54-.92**	**.51-.87**
High-performance	**1.64**	**1.56**	**1.48**	**1.41**	**1.34**
Harsh	**0.24–1.90**	**0.23–1.54**	**.22-1.81**	**.21 - 1.71**	**.20 - 1.63**
Maximum Power (Watts/mm^2)					
Hand held and memory (Watts)	**3**	**3**	**3**	**3**	**3**
Cost-performance (MPU)	**0.9**	**0.96**	**1.13**	**1.11**	**1.1**
High-performance (MPU)	**0.46**	**0.47**	**0.52**	**0.51**	**0.48**
Harsh	**0.2**	**0.22**	**0.22**	**0.24**	**0.25**
Package Pin count Maximum					
Low-cost	**160–850**	**170–900**	**180–950**	**188–1000**	**198–1050**
Cost performance	**660–2801**	**660–2783**	**720- 3061**	**720–3367**	**800–3704**
High performance (FPGA)	**4620**	**4851**	**5094**	**5348**	**5616**
Harsh	**425**	**447**	**469**	**492**	**517**
Minimum Overall Package Profile (mm)					
Low-cost, hand held and memory	**0.3**	**0.3**	**0.3**	**0.3**	**0.3**
Cost-performance	**0.65**	**0.65**	**0.65**	**0.5**	**0.5**
High-performance	**1.4**	**1.2**	**1.2**	**1**	**1**
Harsh	**0.8**	**0.8**	**0.7**	**0.7**	**0.7**

Test (Burn-In) Roadmap (January 2009)

Year of Production	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*
Clock input frequency (MHz)	**400**	**400**	**400**	**400**	**400**	**400**	**400**
Off-chip data frequency (MHz)	**75**	**75**	**75**	**75**	**75**	**75**	**75**
Power dissipation (W per DUT)	**600**	**600**	**600**	**600**	**600**	**600**	**600**
Power Supply Voltage Range (V)							
High-performance ASIC / microprocessor / graphics processor	**0.5–2.5**	**0.5–2.5**	**0.5-2.5**	**0.5–2.5**	**0.5–2.5**	**0.5–2.5**	**0.5–2.5**
Low-end microcontroller	**0.7–10.0**	**0.5–10**	**0.5–10**	**0.5–10**	**0.5–10**	**0.5–10**	**0.5–10**
Mixed-signal	**0.5–500**	**0.5–500**	**0.5–500**	**0.5–500**	**0.5–500**	**0.5–500**	**0.5–1000**
Maximum Number of Signal I/O							
High-performance ASIC	**384**	**384**	**384**	**384**	**384**	**384**	**384**
High-performance microprocessor / graphics processor / mixed-signal	**128**	**128**	**128**	**128**	**128**	**128**	**128**
Commodity memory	**72**	**72**	**72**	**72**	**72**	**72**	**72**
Maximum Current (A)							
High-performance microprocessor	**450**	**450**	**450**	**450**	**450**	**450**	**450**
High-performance graphics processor	**200**	**200**	**200**	**200**	**200**	**200**	**200**
Burn-in Socket							
Pin count	**3000**	**3000**	**3000**	**3000**	**3000**	**3000**	**3000**
Pitch (mm)	**0.3**	**0.2**	**0.2**	**0.2**	**0.2**	**0.2**	**0.1**
Power consumption (A/Pin)	**4**	**5**	**5**	**5**	**5**	**5**	**5**
Wafer Level Burn-In							
Maximum burn-in temperature (ºC)	**175±3**	**175±3**	**175±3**	**175±3**	**175±3**	**175±3**	**175±3**
Pad Layout – Linear							
Minimum pad pitch (µm)	**65**	**65**	**65**	**65**	**65**	**65**	**50**
Minimum pad size (µm)	**50**	**50**	**50**	**50**	**50**	**50**	**40**
Maximum number of probes	**70k**	**70k**	**70k**	**70k**	**70k**	**70k**	**140k**
Pad Layout – Periphery, Area Array							
Minimum pad pitch (µm) *1	**80**	**80**	**80**	**80**	**80**	**80**	**60**
Minimum pad size (µm)	**35**	**35**	**35**	**30**	**30**	**30**	**25**
Maximum number of probes	**150k**	**150k**	**150k**	**150k**	**150k**	**150k**	**300k**
Power consumption (KW/wafer – Low-end microcontroller, DFT/BIST SOC *2)	**5**	**5***Andr*	**10**	**10**	**10**	**10**	**15**

Today's Agenda

- What is the semiconductor roadmap?
- Connections game: Why do we care?
- Aspects of the Design roadmap
- Aspects of the System Drivers roadmap and the Overall Roadmap Technology Characteristics (ORTCs)
- More Than Moore

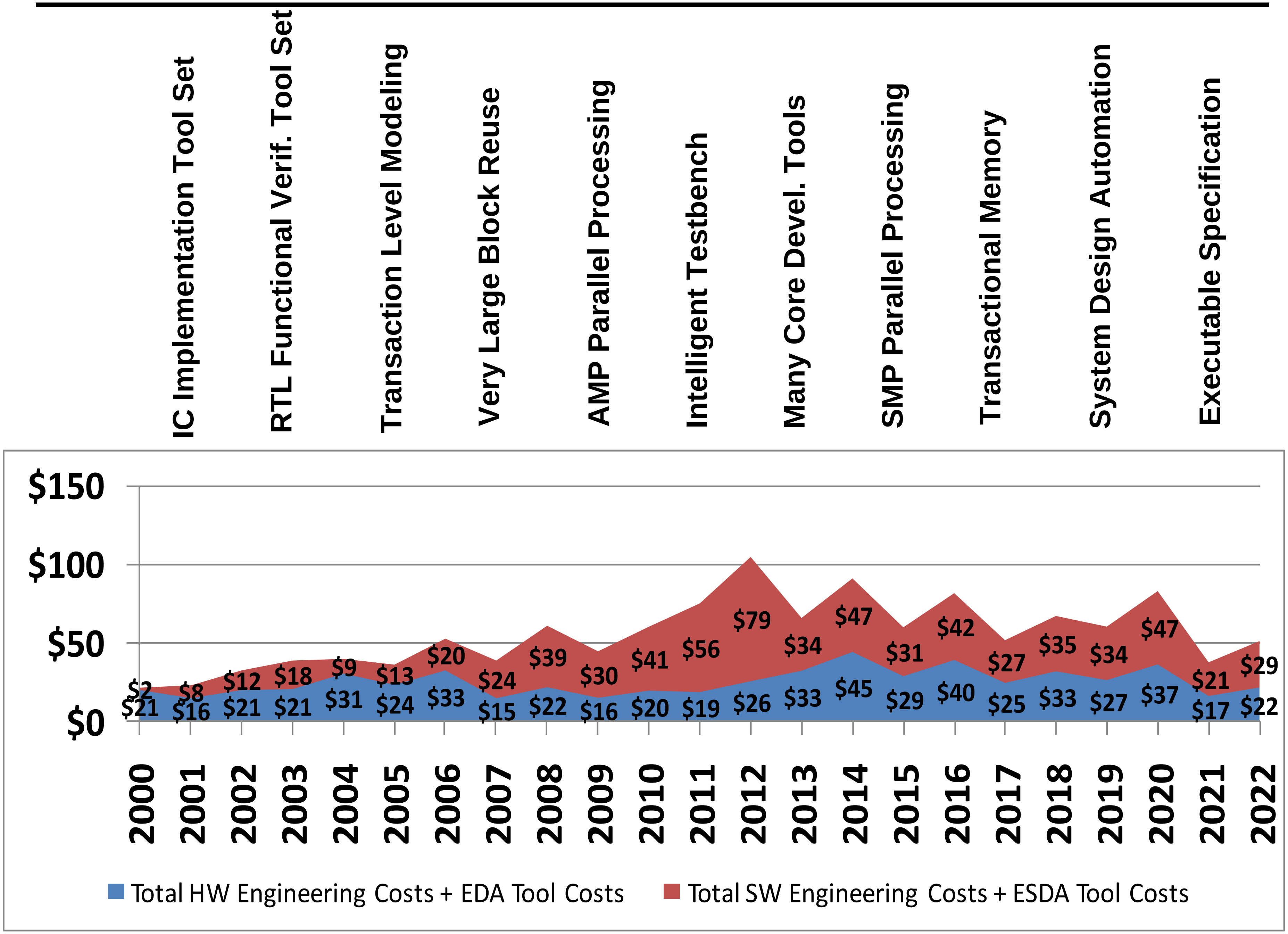
Silicon Complexity Challenges

- Silicon Complexity = impact of process scaling, new materials, new device/interconnect architectures
- Non-ideal scaling (leakage, power management, circuit/device innovation, current delivery)
- Coupled high-frequency devices and interconnects (signal integrity analysis and management)
- Manufacturing variability (library characterization, analog and digital circuit performance, error-tolerant design, layout reusability, static performance verification methodology/tools)
- Scaling of global interconnect performance (communication, synchronization)
- Decreased reliability (SEU, gate insulator tunneling and breakdown, joule heating and electromigration)
- Complexity of manufacturing handoff (reticle enhancement and mask writing/inspection flow, manufacturing NRE cost)

System Complexity Challenges

- System Complexity = exponentially increasing transistor counts, with increased diversity (mixed-signal SOC, …)
- Reuse (hierarchical design support, heterogeneous SOC integration, reuse of verification/test/IP)
- Verification and test (specification capture, design for verifiability, verification reuse, system-level and software verification, AMS self-test, noise-delay fault tests, test reuse)
- Cost-driven design optimization (manufacturing cost modeling and analysis, quality metrics, die-package co-optimization, …)
- Embedded software design (platform-based system design methodologies, software verification/analysis, codesign w/HW)
- Reliable implementation platforms (predictable chip implementation onto multiple fabrics, higher-level handoff)
- Design process management (team size / geog distribution, data mgmt, collaborative design, process improvement)

ITRS Design Cost Chart 2009 ($M)



System-Level Design and Software

Hardware design productivity is growing appropriately

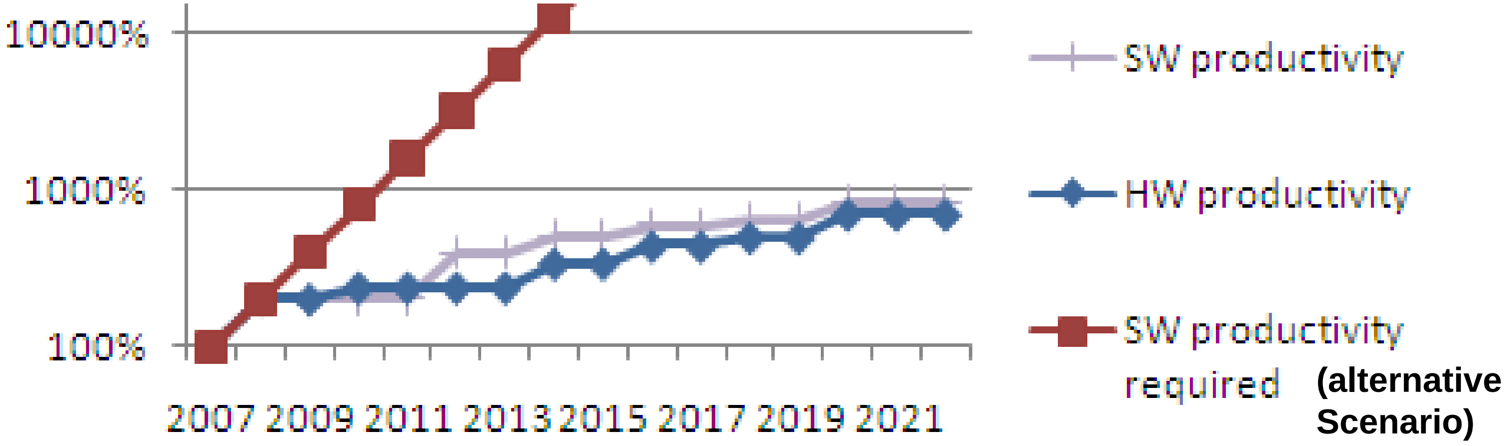
- Requirements correspond roughly with solutions
- Innovations pacing properly (transistors / designer / year)

Large gap in software productivity possibly opening up

- If hardware accelerators are heavily leveraged, problem mitigated
- Otherwise, possibly 100X gap can affect memory size, other

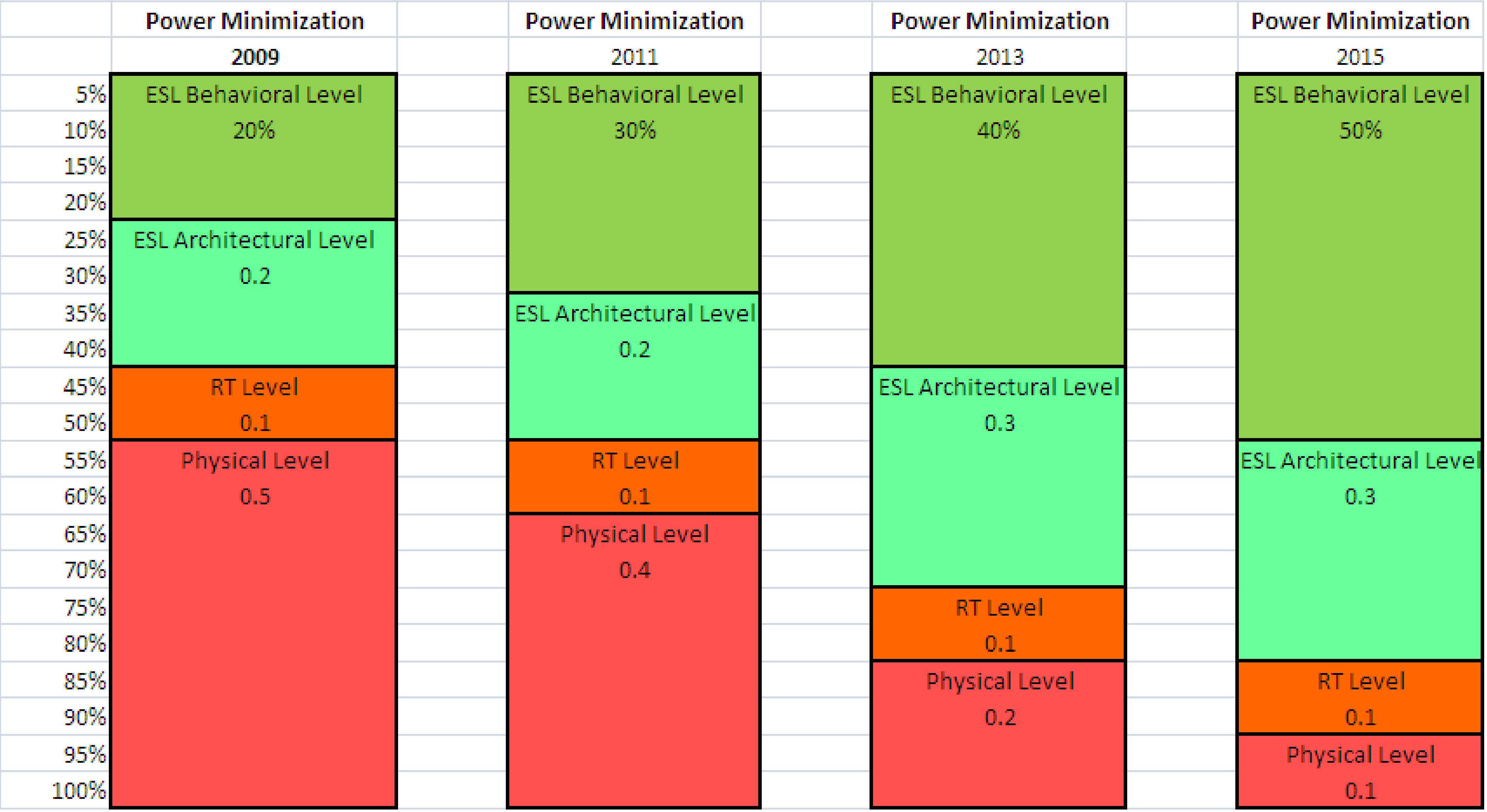
2009 ITRS adds new parameters accordingly

- Hardware design productivity requirement
- Software design productivity requirement



(alternative Scenario)

Future Impact of (System-Level, SW/HW) Design on Power



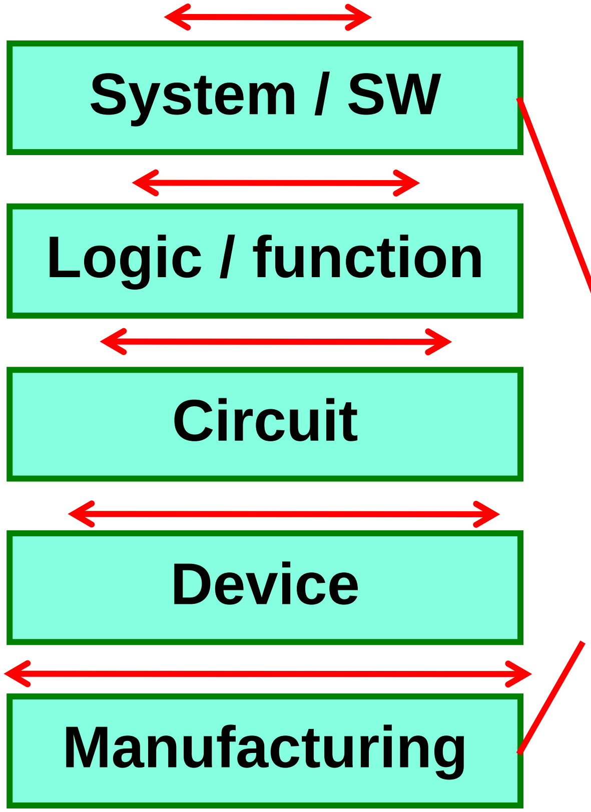
Impact of Design on “Sigma” (Variability)

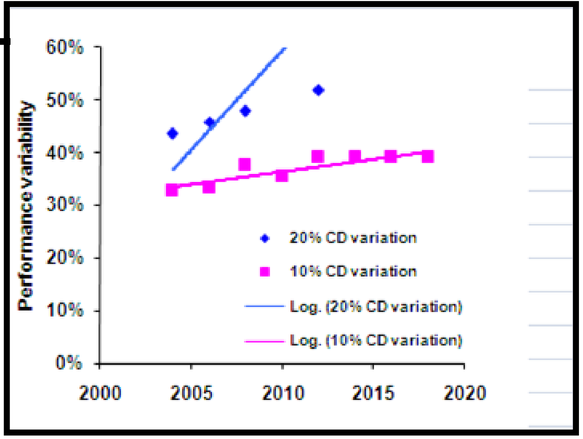
Goal

- Quantify “how many sigmas” design can “reduce”
- ITRS 2005: CD 3σ tolerance changed from 10% → 12% per Design guidance

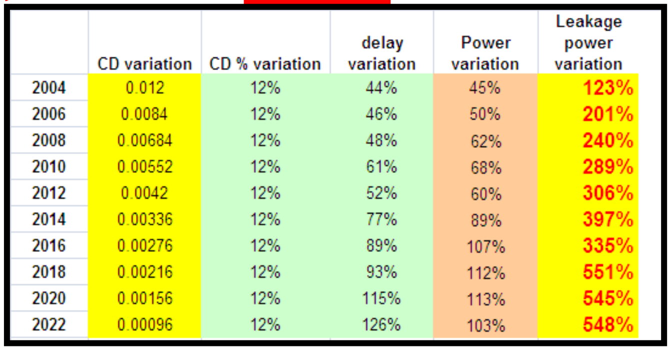
Approach

- Inventory of design techniques / tools
- Match inventory to parameters or correlations in model
- Use variability model to capture “delta” in sigmas
- See work of S. Nassif et al., IBM ARL



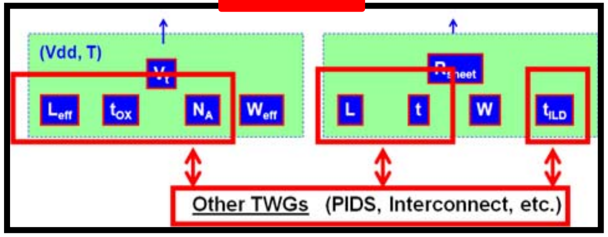


Check overall variation



	CD variation	CD % variation	delay variation	Power variation	Leakage power variation
2004	0.012	12%	44%	45%	123%
2006	0.0084	12%	46%	50%	201%
2008	0.00684	12%	48%	62%	240%
2010	0.00552	12%	61%	68%	289%
2012	0.0042	12%	52%	60%	306%
2014	0.00336	12%	77%	89%	397%
2016	0.00276	12%	89%	107%	335%
2018	0.00216	12%	93%	112%	551%
2020	0.00156	12%	115%	113%	545%
2022	0.00096	12%	126%	103%	548%

Use variability model

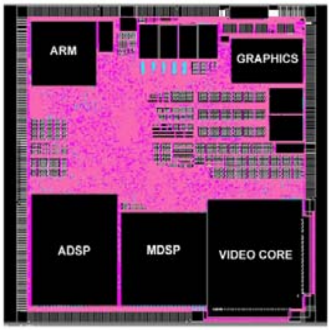


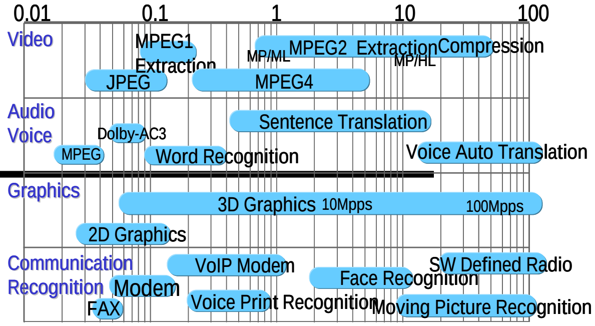
Inputs (manufacturing)

Today’s Agenda

- What is the semiconductor roadmap?
- Connections game: Why do we care?
- Aspects of the Design roadmap
- Aspects of the System Drivers roadmap and the Overall Roadmap Technology Characteristics (ORTCs)
- More Than Moore

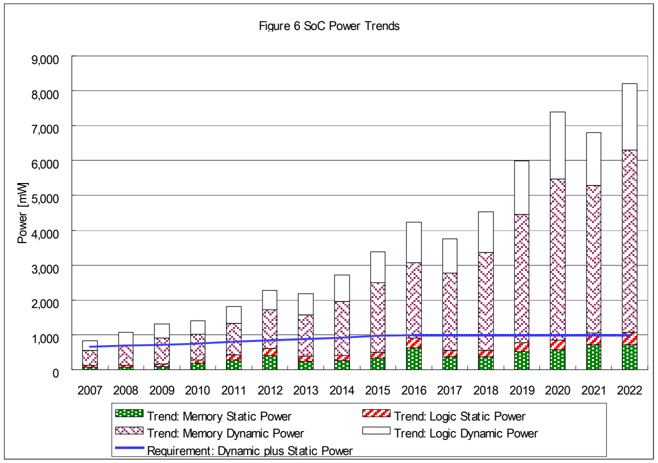
Consumer Driver



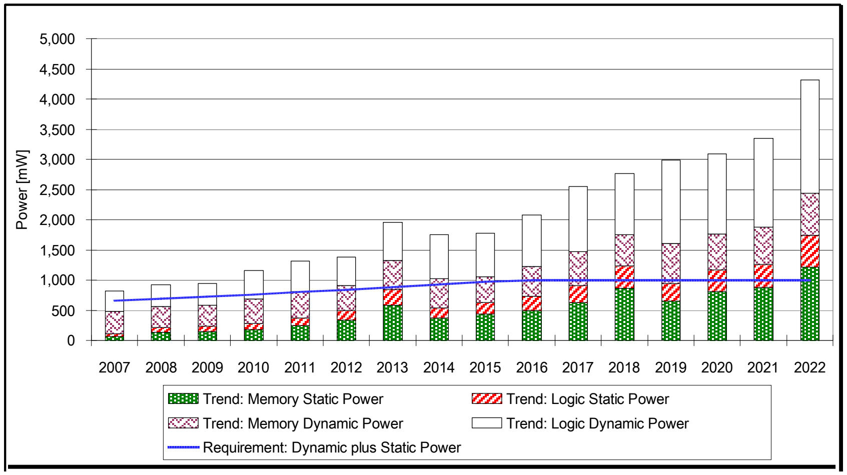


- **Two flavors: Portable (baseband processor) and Stationary (GPU)**
- **2008: Updated with realistic dynamic power**
  - Memory dynamic power 10X less than modeled previously

8 W max total (2022)



4.3 W max total (2022)



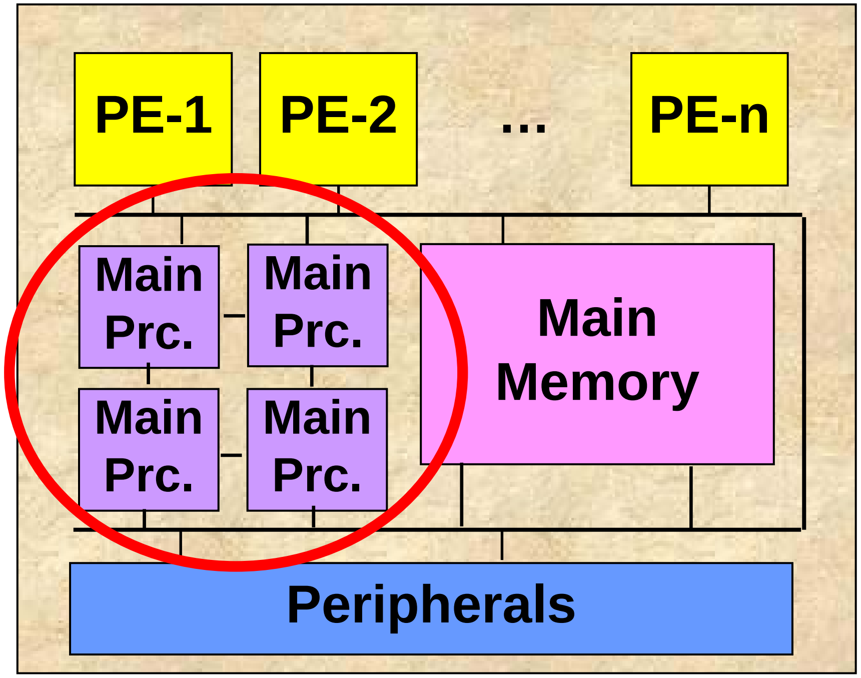
- **2009: Total power budget reduced 1W → 0.5W**
- **Future: “wireless” driver with RF/A/MS requirements**
- **Future: more specific parameters for Test roadmap**
  - #clocks, #power domains, #unique cores, #IOs, etc.

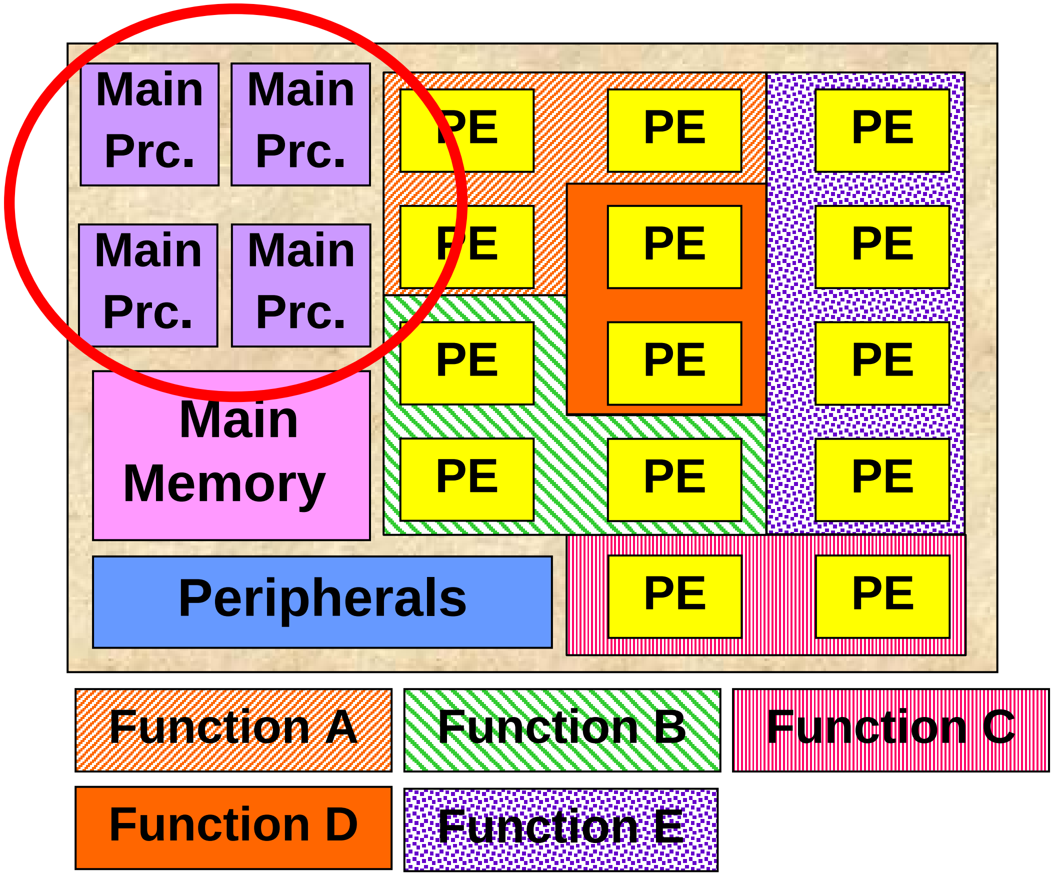
SOC Consumer Portable Architecture Model

- #Main Processors grows to 2, 4 and beyond

- Power budget reduced to 0.5W

- Die size reduces slowly to 44mm^2

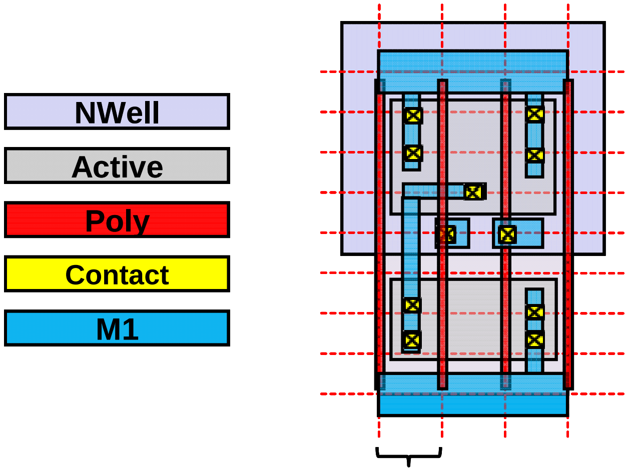




ORTCs: A-Factor Models (= Heart of ITRS)

(Area = A-factor × F^2)

■ Logic: A-factor = 175



M2 pitch ($P_{M2} \approx 1.25P_{M1}$)

Contacted-poly pitch ($P_{Poly} \approx 1.5P_{M1}$)

NAND2 Area

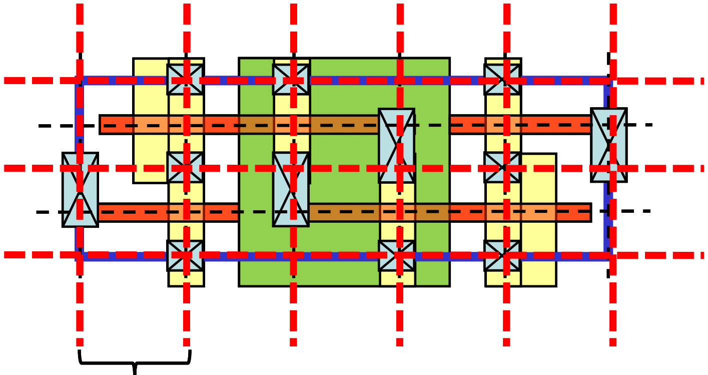
$= 3\ P_{Poly} \times 8\ P_{M2}$

$\approx\ (3 \times 1.5\ P_{M1}) \times (8 \times 1.25\ P_{M1})$

$= 45\ (P_{M1})^2$

$= 180\ F^2 \rightarrow$ **175** F^2

■ SRAM: A-factor = 60



M1 pitch (P_{M1})

Contacted-poly pitch ($P_{Poly} \approx 1.5P_{M1}$)

SRAM Bitcell Area

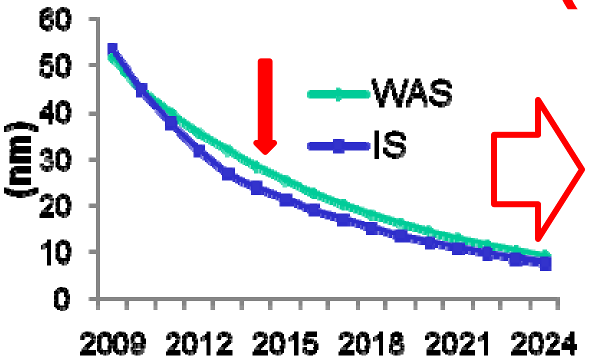
$= 2\ P_{Poly} \times 5\ P_{M1}$

$= 3\ P_{M1} \times 5\ P_{M1} = 15\ (P_{M1})^2$

$= 15\ (2\ F)^2 =$ **60** F^2

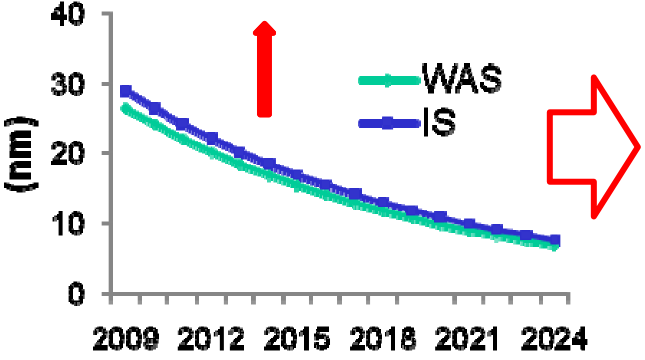
New MPU Density/Power/Frequency Roadmap

M1 Half-Pitch (F)



Decrease P_{dyn} and P_{leak}

Physical L_{gate} (L)



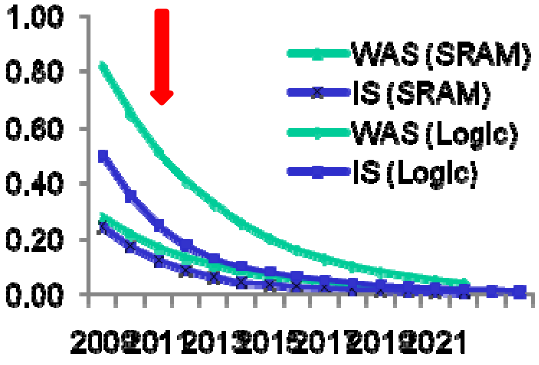
Increase P_{dyn} , decrease P_{leak}

A-Factor (A)

Logic: ~320 (WAS)→ 175 (IS)

SRAM: ~100 (WAS) → 60 (IS)

Unit cell size



Growth of #Tr

2x / 3 year (WAS)

→ 2x / 2 year (IS)

up to 2013

Die size reduction

310mm^2 (WAS)

→ 260mm^2 (IS)

#core/die, #tr/core

12.2% / year (WAS)

→ 18.9% / year (~2013, IS),

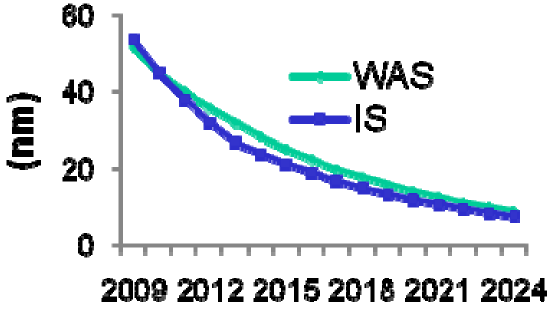
→ 12.2% / year (2014~, IS)

Increased P_{dyn} and P_{leak}

Design Pacing, Challenges Unabated

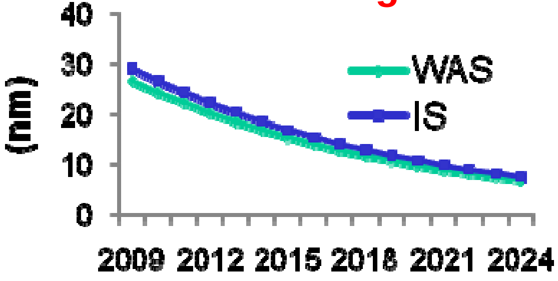
- 2009: Lgate and M1 HP scaling updates change Drivers

M1 Half Pitch



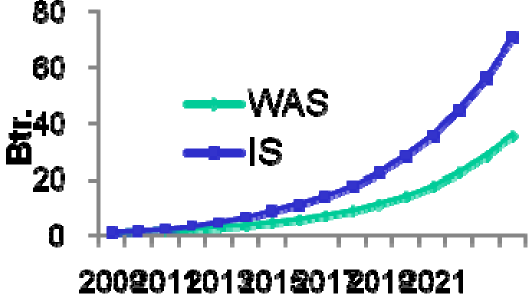
2 year delay, but faster scaling
0.7x / 3yr → 0.7 / 2yr (~2013), 0.7x / 3yr (2014~)

Physical L$_{gate}$



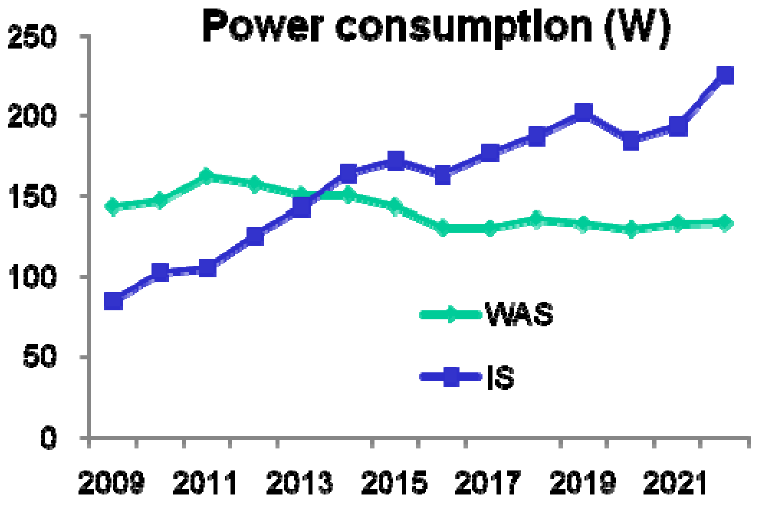
1 year shift

#Tr per die



New A-factors
Faster M1 half pitch reduction

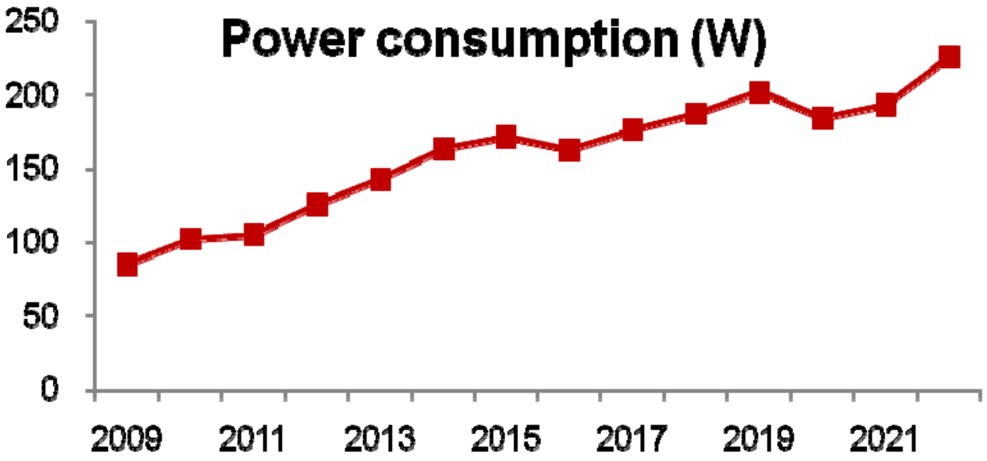
Updated MPU model (power)

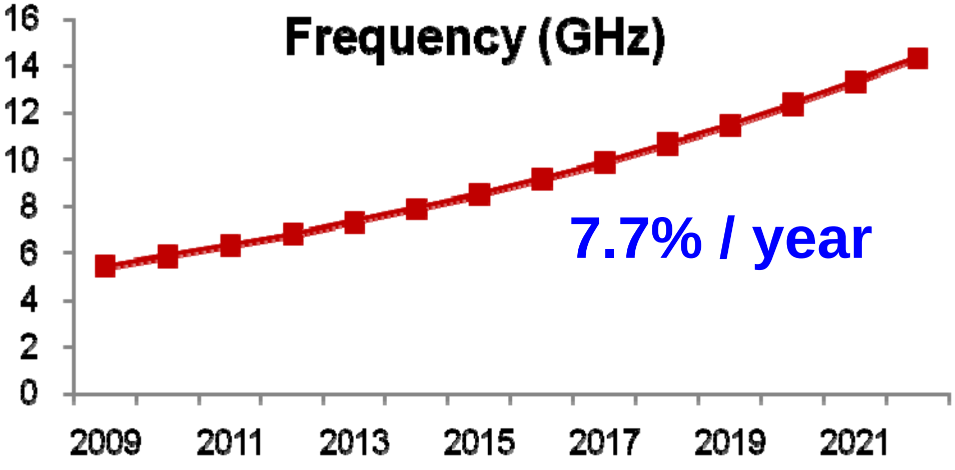


Frequency-Power Envelope Remains Critical System Issue

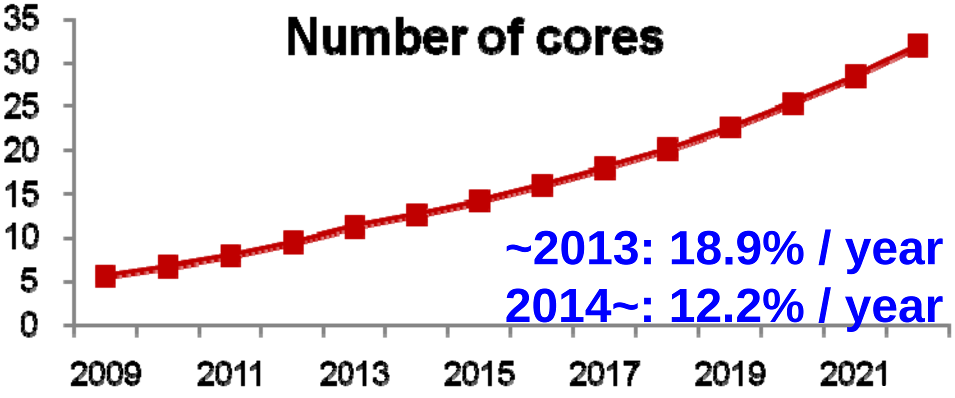
- **Current priorities**
  - Power #1 goal
  - Frequency slowdown
  - Multicore enables tradeoff
  - Point of this slide: ITRS gives a “best-guess” tradeoff

- **Need to track tradeoff**
  - Market vigilance
  - Yearly adjustment





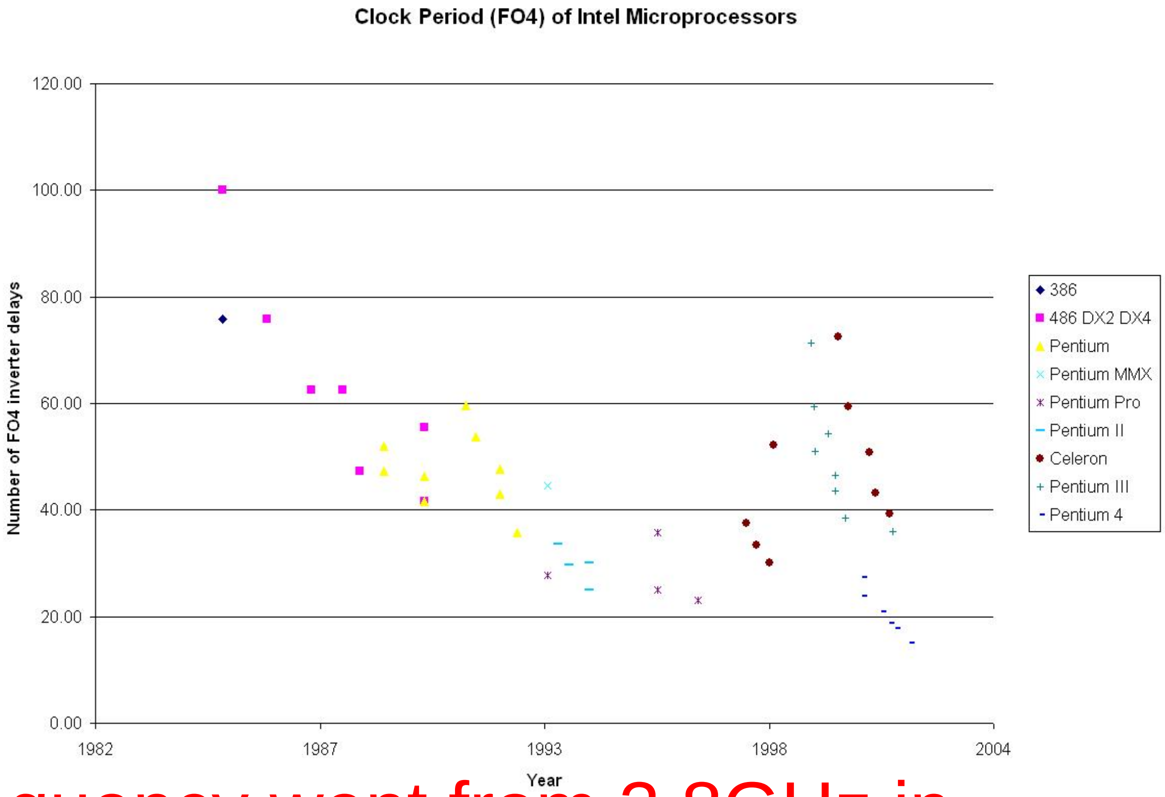
7.7% / year



~2013: 18.9% / year
2014~: 12.2% / year

History: μArchitecture Wakeup Call in 2001

- **Historical "Moore's Law" of 2X/node frequency increase came from two sources**
  - 1.4X from device: (PIDS 17%/year** improvement of CV/I)
  - 1.4X from "microarchitecture" (pipelining, etc.)
- **2001 ITRS: Clock period ≥ ~12 FO4 INV delays ≅ 200 × CV/I**
  - "Microarchitecture runs out of steam"
  - Frequency roadmap: 2X → 1.4X/node



****ITRS 2008: PIDS ITWG shifted to 13%/year CV/I per Design guidance**

MPU max on-chip clock frequency went from 3.8GHz in Pentium4 to 3.3GHz in Penryn – WHY?

History: Power Wakeup Call in 2007

- **Power is a hard limit**
  - E.g., 120W for the **desktop** platform
  - Previous ITRS allowed max chip power and max W/cm^2 power density to grow
  - Previous ITRS roadmapped the "power management gap" – but there can be no "gap" in actual products
- **"New Marketing" (2007): Utility = GOPS, not GHz**
  - …when we can't scale frequency due to power limit
- Frequency scaling for MPUs is function of: (1) multi-core roadmap, (2) hard limit on power, and (3) MPU architecture choices

2007 ITRS: ~1X Frequency Scaling for MPU

Crude Assumptions

Die Area:	1X / node **(current MPU model)**
Number of Cores:	2X / node **(current MPU model)**
Total $P_{dynamic}$:	1X / node **(NEW, CONSTRAINT)**
α (switch factor):	1X / node
Switched cap / mm^2:	1.15X / node **(Borkar/Intel, 2001 → reverify)**
Vdd:	0.95X / node **(historical ITRS)**
Total P_{static} :	1X / node **(high-k, #FO4s ↑, …)**

Implications

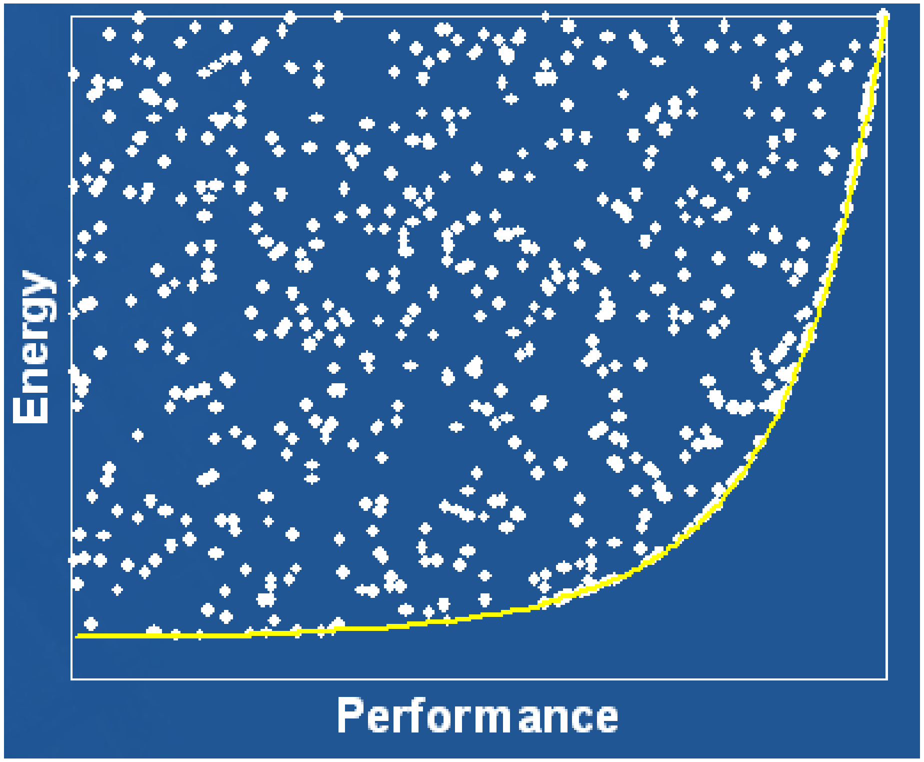
α x C x V_{dd}^2:	1.04X / node **(from above)**
Frequency:	**0.98 X / node** **($\alpha CV^2 f$ = 1X, P $\propto f^3$, 0.96 = 0.98^3)**
GOPS:	2X / node **(2X #cores, 1X frequency)**

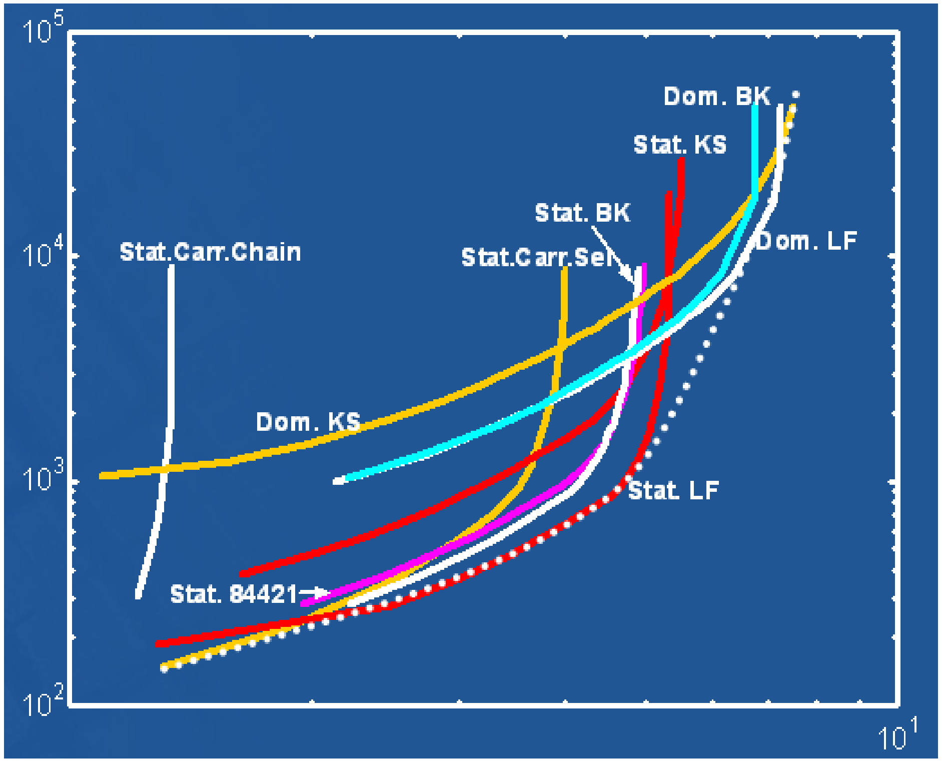
Your Thoughts on Frequency Scaling?

- **Why frequency might scale at < 0.98X / node**
  - Static power increases rapidly vs. dynamic power
  - Inter-die wires/logic not accounted for
- **Why frequency might scale at > 0.98X / node**
  - Number of FO4s in the clock period is increasing
    - Save power faster than we give up frequency, due to logic optimization
    - Static power can be better managed → can use more HVT, less LVT
  - High-k dramatically reduces I_{gate} (and improves subthreshold swing)
  - **Better opportunity for DVFS with multi-core (and heterogeneity)**
  - **Application, OS-driven power management**
  - Power budget may actually increase very gradually
  - Cores are smaller
  - Need to market new products
    - 2X cores, ≥ 1X frequency is value proposition for consumers

Energy-Delay Tradeoff Curve

- **Very little bang for the buck at extremes**
- **Shape of tradeoff curve, and location on curve, are relevant as MPU frequency backs away from limits of process**
  - E.g., more power reduction (logic, Vt) available when freq ↓
  - E.g., cubic relationship between power and frequency





Other Considerations

- Consider **reliability** as a constraint
- Consider **stacking / 3D** integration
- Consider **DVFS** impact on peak power, utility
- Consider **parallel SW** impact on utility
- Consider **frequency-power tradeoff** calibrated to standard ASIC/SOC implementation flows
- Adjust for **3-year** technology node timing
- Consider **server** platform vs. **desktop** platform

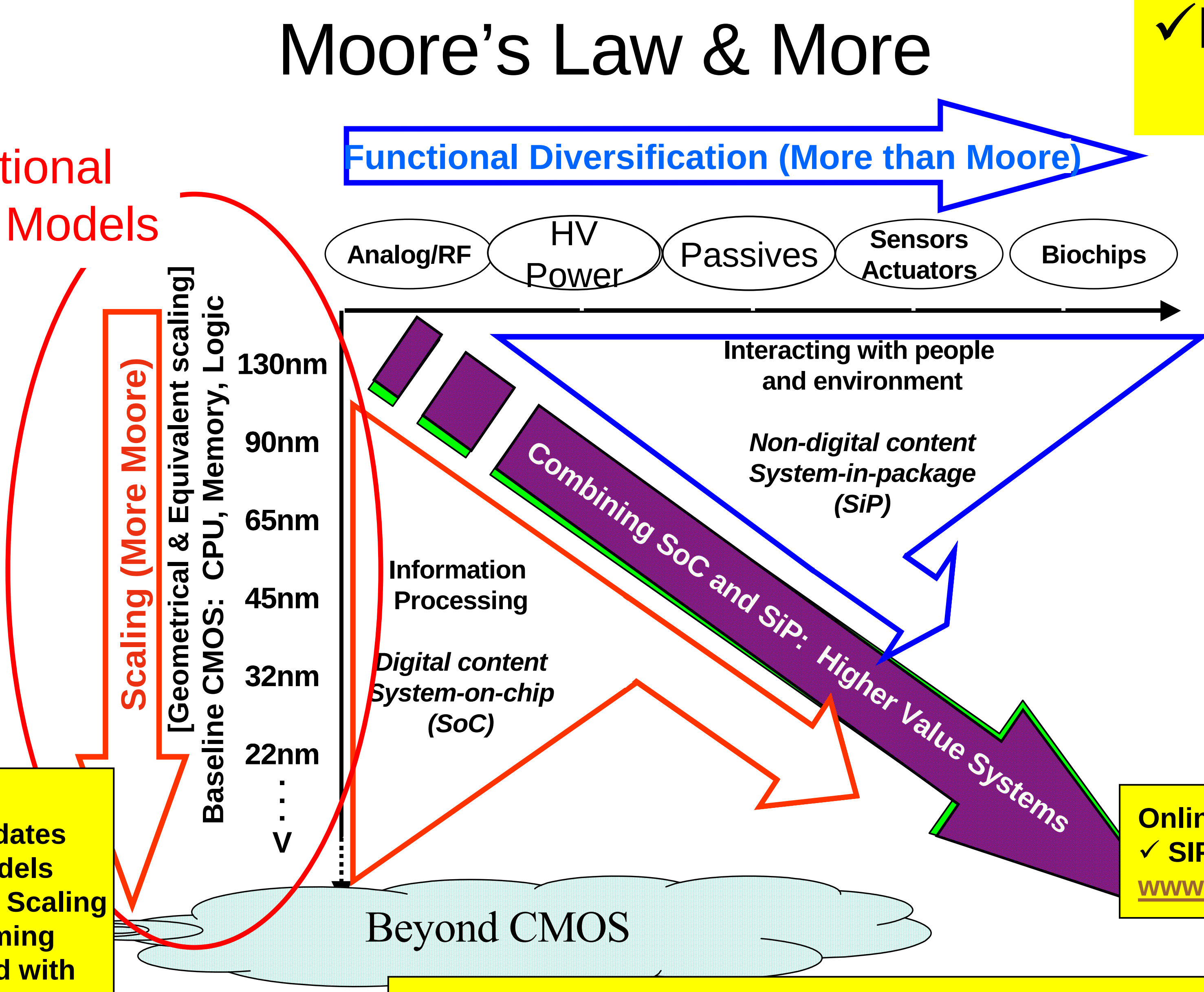
Today’s Agenda

- What is the semiconductor roadmap?
- Connections game: Why do we care?
- Aspects of the Design roadmap
- Aspects of the System Drivers roadmap and the Overall Roadmap Technology Characteristics (ORTCs)
- More Than Moore

“More Than Moore” (2007 ITRS)

✓New work In 2009

Traditional ORTC Models


Moore’s Law & More
Functional Diversification (More than Moore)
Analog/RF
HV Power
Passives
Sensors Actuators
Biochips
Scaling (More Moore)
[Geometrical & Equivalent scaling]
Baseline CMOS: CPU, Memory, Logic
130nm
90nm
65nm
45nm
32nm
22nm
. . . V
Interacting with people and environment
Non-digital content System-in-package (SiP)
Information Processing
Digital content System-on-chip (SoC)
Combining SoC and SiP: Higher Value Systems
Beyond CMOS

New in 2009:
- ✓ More than Moore “White Paper”
- ✓ More Commentary In ITWG Chapters

New in 2009:
- ✓ Survey updates to ORTC Models
- ✓ Equivalent Scaling Roadmap Timing Synchronized with PIDS and FEP

Online in 2008:
- ✓ SIP “White Paper” www.itrs.net/papers.html

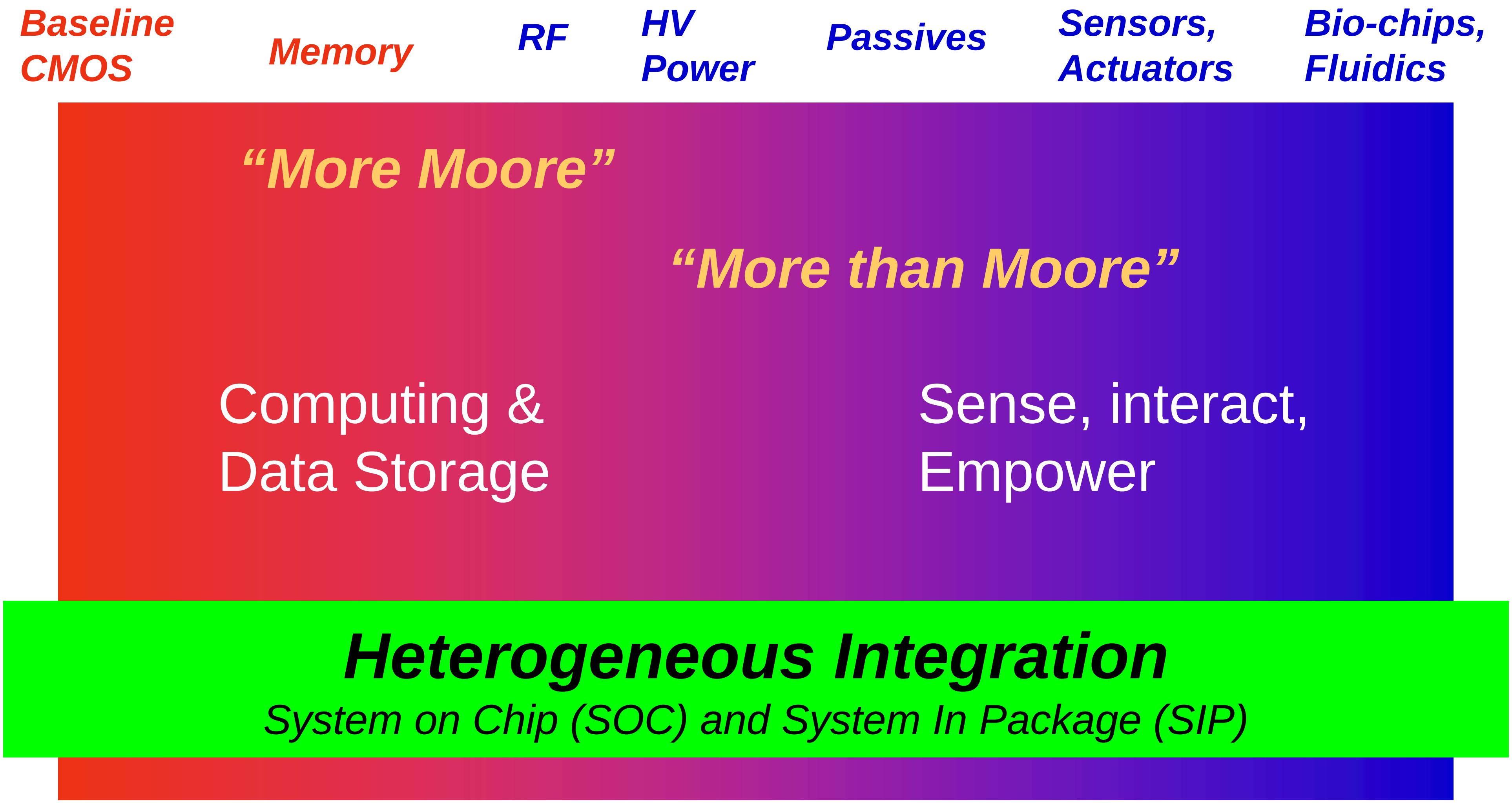
New in 2009:
- ✓ Research and PIDS transfer timing clarified
- ✓ Work underway to identify next storage element

Source: 2009 ITRS - Executive Summary Fig 1

2007/08 ITRS “Moore’s Law and More”

Alternative Definition Graphic

[2009 – Unchanged]



Baseline CMOS

Memory

RF

HV Power

Passives

Sensors, Actuators

Bio-chips, Fluidics

“More Moore”

“More than Moore”

Computing & Data Storage

Sense, interact, Empower

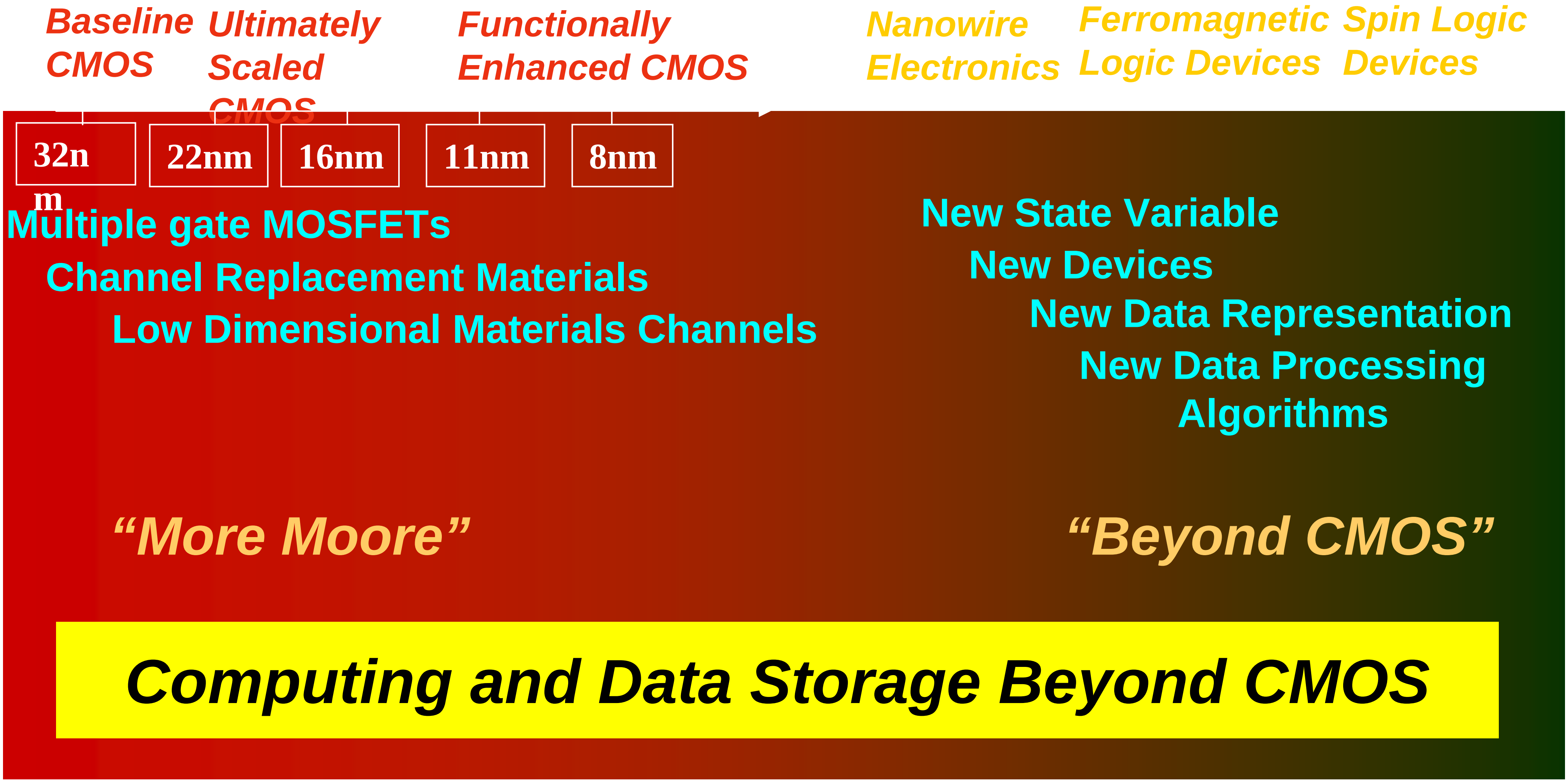
Heterogeneous Integration

System on Chip (SOC) and System In Package (SIP)

Source: ITRS, European Nanoelectronics Initiative Advisory Council (ENIAC)

2008 ITRS “Beyond CMOS” Definition Graphic

[2009 – Unchanged]



Baseline CMOS

Ultimately Scaled CMOS

Functionally Enhanced CMOS

Nanowire Electronics

Ferromagnetic Logic Devices

Spin Logic Devices

32nm

22nm

16nm

11nm

8nm

Multiple gate MOSFETs

Channel Replacement Materials

Low Dimensional Materials Channels

New State Variable

New Devices

New Data Representation

New Data Processing Algorithms

“More Moore”

“Beyond CMOS”

Computing and Data Storage Beyond CMOS

Source: Emerging Research Device Working Group

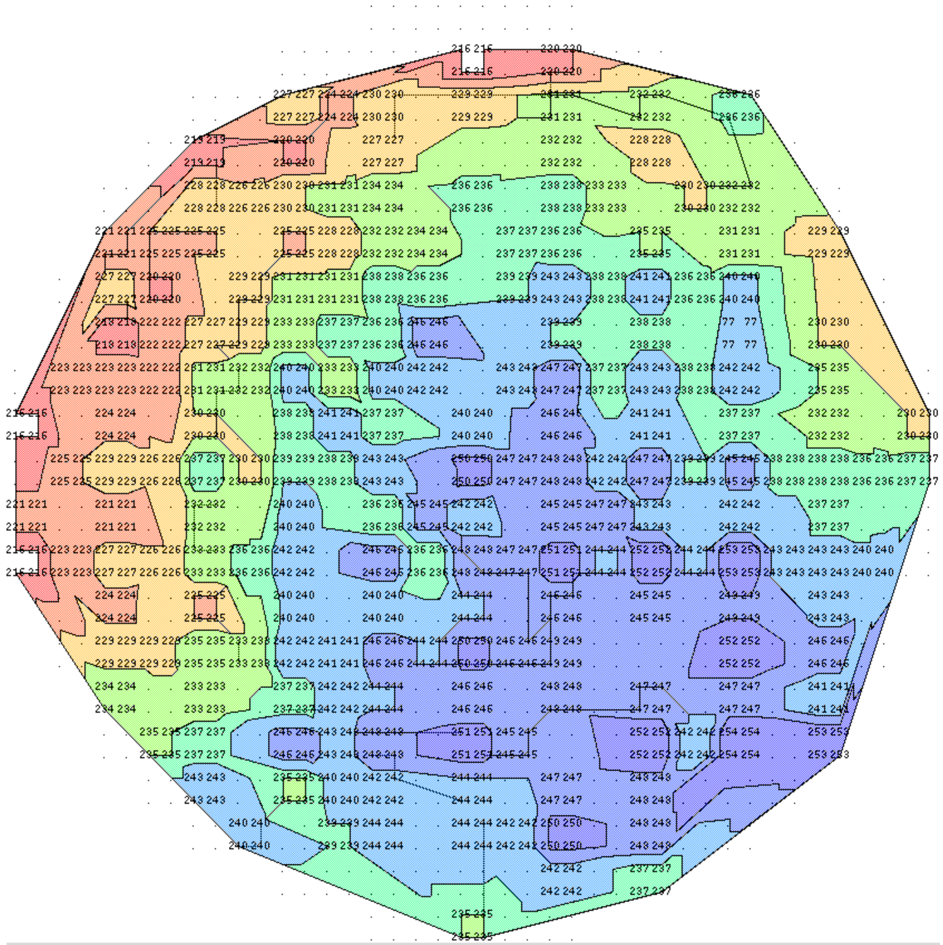
Recap

- What is the semiconductor roadmap?
- Connections game: Why do we care?
- Aspects of the Design roadmap
- Aspects of the System Drivers roadmap and the Overall Roadmap Technology Characteristics (ORTCs)
- More Than Moore

BACKUP

Problem: Uncontrollable Variation

- Chips don’t work as designed
- Loss of predictability →
  - Guardbands
  - Overdesign
  - Worse time to market, cost, power
  - Loss of product value



Across-wafer frequency variation → What performance spec for this chip?

Problem: Yield and Cost and Risk

- **Chips are thrown away**
- **Consider a cellphone chip selling 100M copies**
  - Design house pays $5K/300mm wafer in 90nm technology
  - 10mm x 10mm die size at 90nm → ~700 die/wafer
  - 90% vs. 95% yield
    - 630 vs. 665 good die per wafer
    - 158730 vs. 150370 wafers needed to meet the demand
    - $42M difference
- **What matters is *good die/wafer***
  - Not too slow, not too power-hungry….

Leakage Power

- Leakage power = unwanted current in transistors
- “Wasted power”
- Thought of as biggest potential roadblock to Moore’s Law
- Subthreshold leakage = biggest leakage component at operating temperatures (exponential dep)
- Back of envelope:
  - **30% of 100W power per uP is leakage**
  - **200M uP chips sold**
  - **100W-yr = 714 pounds of coal burned**
  - **10% leakage savings = 3W per uP**
  - **1W to cool per 1W dissipated**
  - **Saves (3 x 200M) x (714 / 100) x 2 = 8,568,000,000 pounds of coal per year (x2.86) = 24,504,000,000 pounds of CO2 per year**
  - **About 0.2% of total of USA or China**

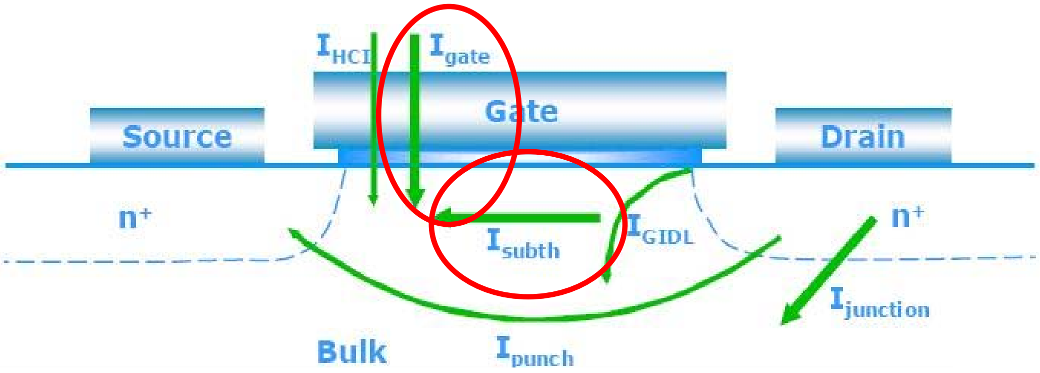


Figure courtesy Roy et al.

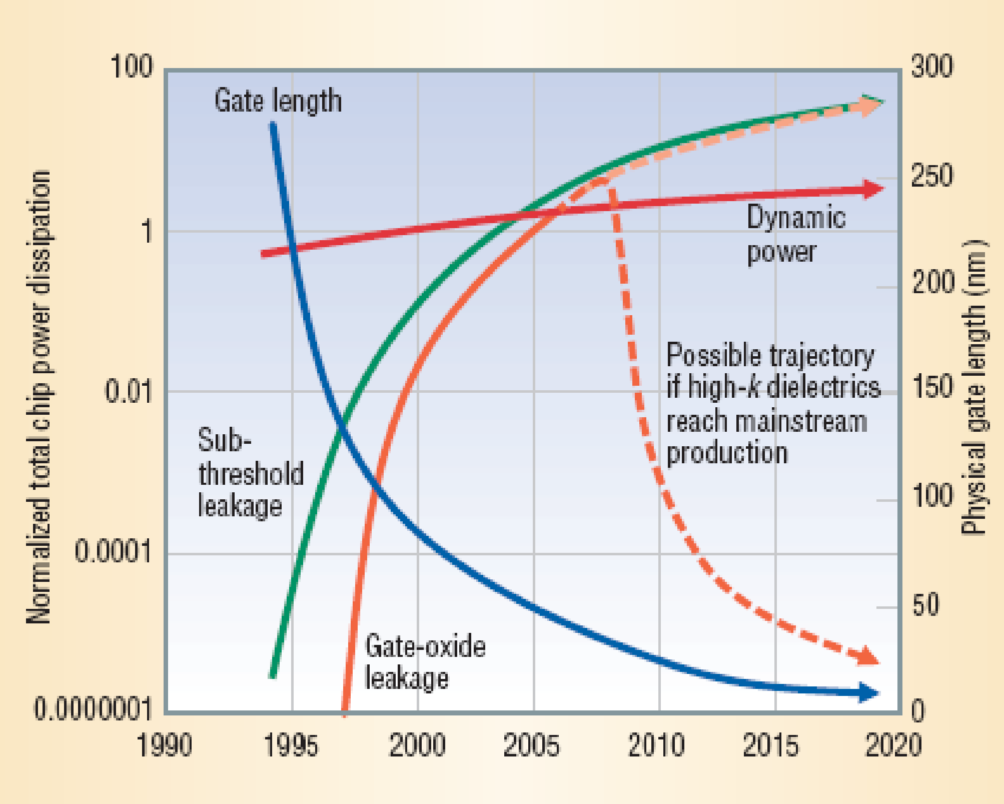
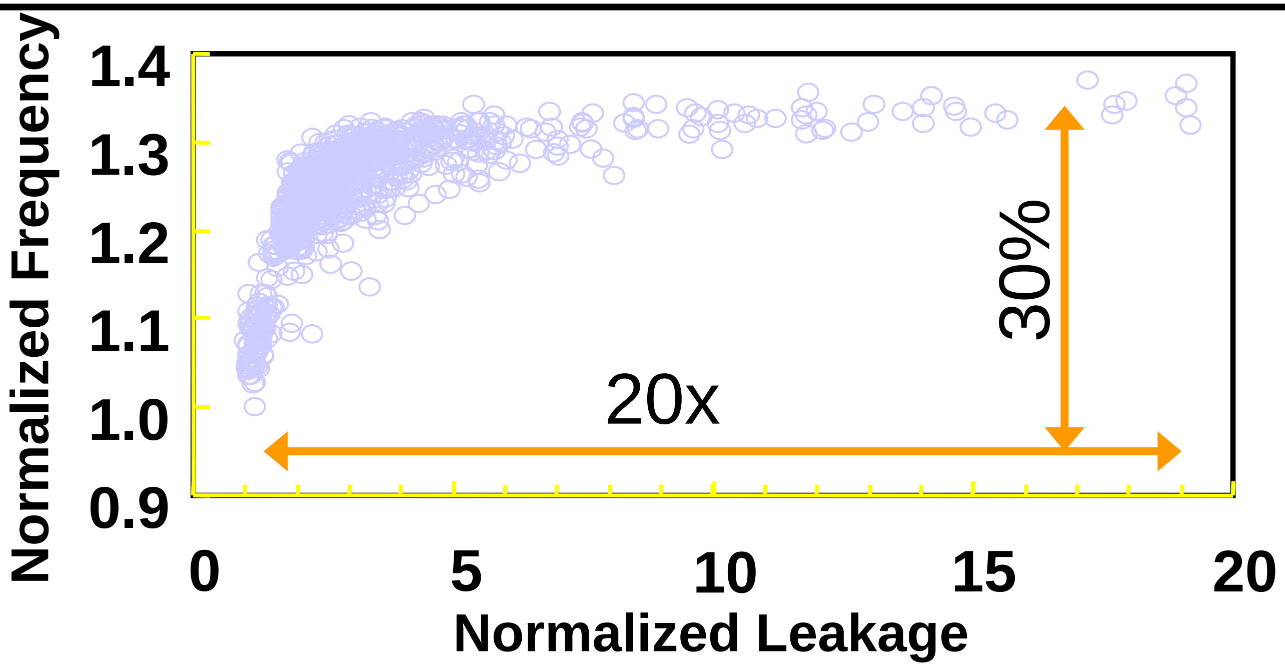


Figure courtesy Blaauw et al.

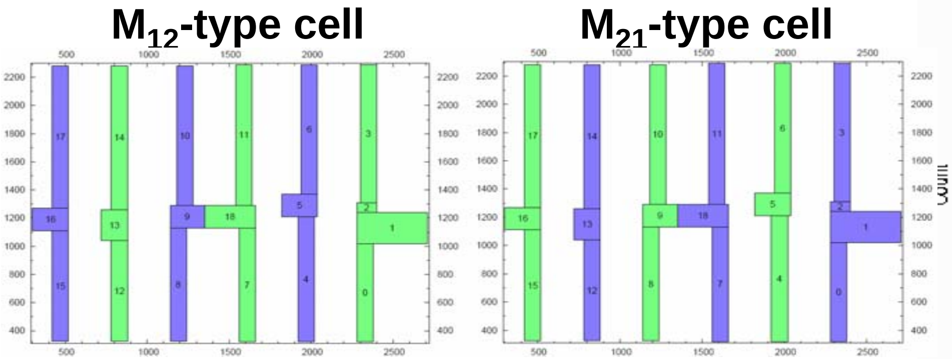
Leakage Power Variability



- **Leakage power variability**
  - Subthreshold leakage is exponential in almost everything (L, Vt, Tox, Temperature, Voltage..) → 5-20X variation is common
    - **Gate length (= “Lgate”, or “CD” – “critical dimension”) manufacturing variation is biggest source**
  - Power-limited yield loss
  - Problematic leakage power and ‘burn-in’ testing
- **Design must deal with this manufacturing-induced variation**

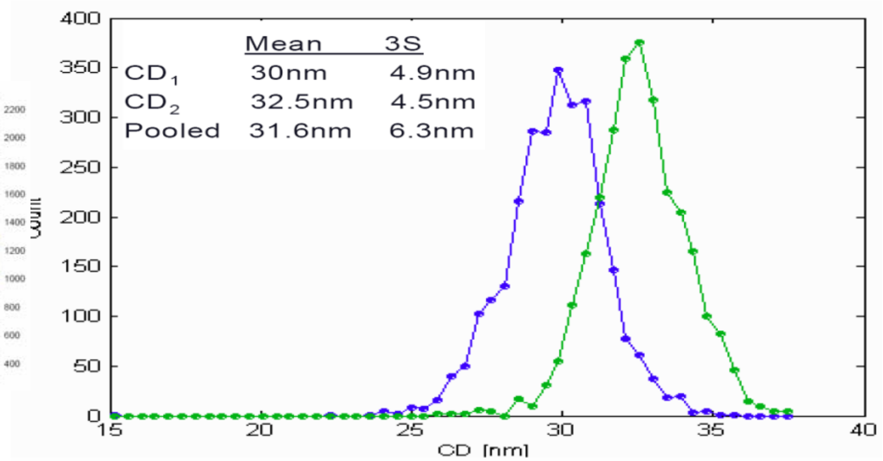
DPL Also Causes A “Bimodal” Problem…

- TWO CD distributions and TWO different colorings → TWO different timings



🟩 Gates from CD group1

🟦 Gates from CD group2



	Mean	3S
CD_1	30nm	4.9nm
CD_2	32.5nm	4.5nm
Pooled	31.6nm	6.3nm

- Is this really a problem?
  - Yes, I think so. (e.g., my 2008 SPIE Microlithography keynote)
  - In 2009 ITRS, CD mean difference in DPL is now roadmapped