Technology Scaling and Roadmap for 22nm CMOS logic and beyond

October 27, 2008

＠Dalian University of Technology

Hiroshi Iwai

Tokyo Institute of Technology

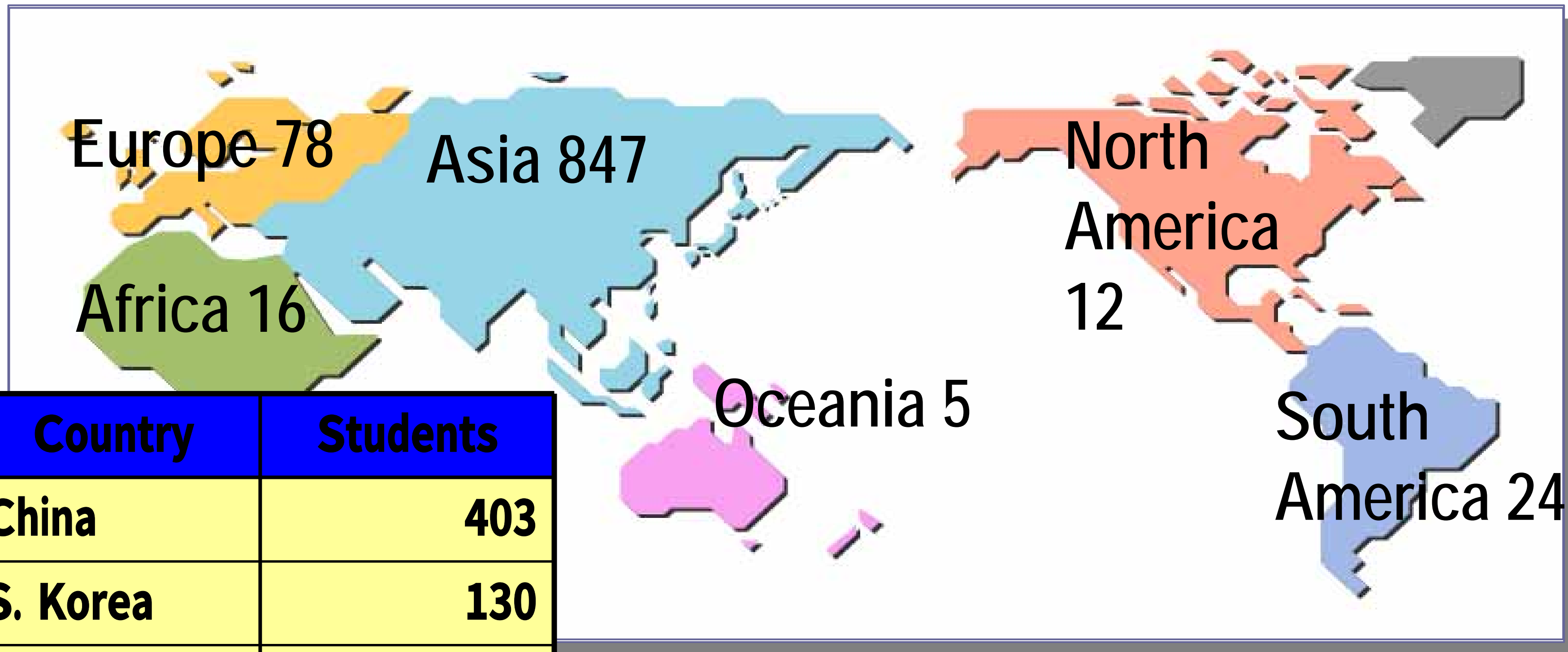
Founded in 1881, Promoted to Univ. 1929

Tokyo Institute of Technology
東京工業大学

Promoted to Univ. 1929

Total 10,000 students: 5,000 under graduate
5,000 graduate

International Students



Country	Students
China	403
S. Korea	130
Indonesia	64
Thailand	55
Vietnam	60
Malaysia	28

Total 982

(As of May. 1, 2005)

Outline

1. Scaling

2. ITRS Roadmap

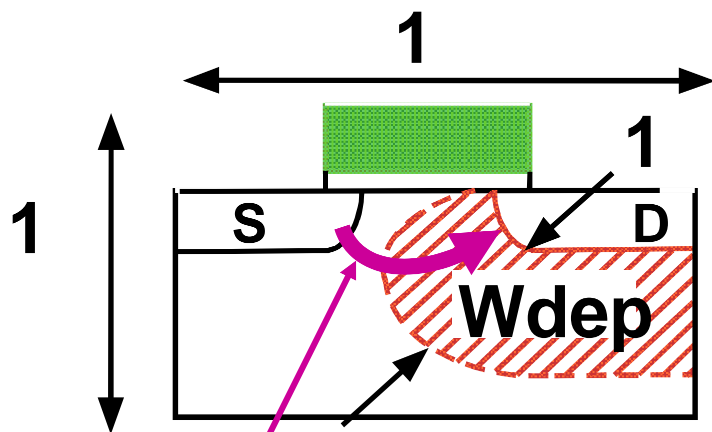
3. Voltage Scaling/ Low Power and Leakage

4. SRAM Cell Scaling

5. Roadmap for further future
   as a personal view

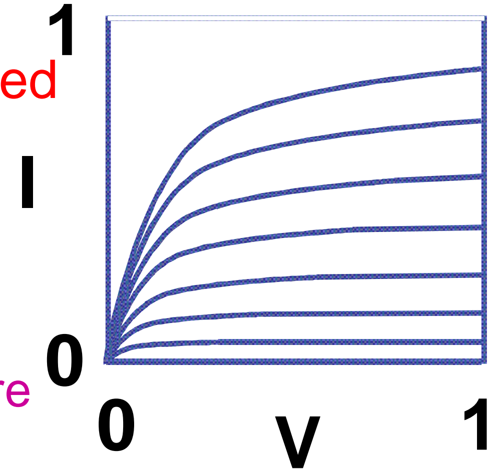
1. Scaling

Scaling Method: by R. Dennard in 1974



Wdep: Space Charge Region (or Depletion Region) Width

Wdep has to be suppressed
Otherwise, large leakage between S and D



Leakage current

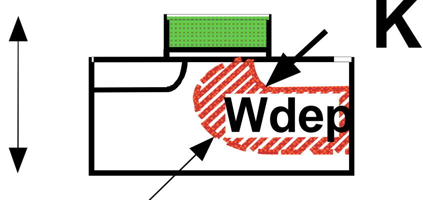
Potential in space charge region is high, and thus, electrons in source are attracted to the space charge region.

K=0.7 for example

X , Y , Z : K,	V : K,	Na : 1/K

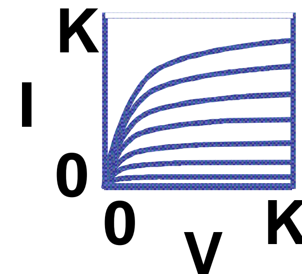
By the scaling, Wdep is suppressed in proportion, and thus, leakage can be suppressed.

→ Good scaled I-V characteristics



$$\text{Wdep} \propto \sqrt{V/Na}$$

: K



I : K

Downscaling merit: Beautiful!

Geometry & Supply voltage	L_g, W_g T_{ox}, V_{dd}	K	**Scaling K : K=0.7 for example**
Drive current in saturation	I_d	K	$I_d = v_{sat} W_g C_o (V_g - V_{th})$ C_o: gate C per unit area $\rightarrow W_g (t_{ox}^{-1})(V_g - V_{th}) = W_g t_{ox}^{-1}(V_g - V_{th}) = KK^{-1}K = K$
I_d per unit W_g	$I_d/\mu m$	1	I_d per unit $W_g = I_d / W_g = 1$
Gate capacitance	C_g	K	$C_g = \varepsilon_o \varepsilon_{ox} L_g W_g / t_{ox} \rightarrow KK/K = K$
Switching speed	τ	K	$\tau = C_g V_{dd} / I_d \rightarrow KK/K = K$
Clock frequency	f	1/K	$f = 1/\tau = 1/K$
Chip area	A_{chip}	α	α: Scaling factor $\rightarrow$ **In the past, $\alpha > 1$ for most cases**
Integration (# of Tr)	N	α/K^2	$N \rightarrow \alpha/K^2$ **$= 1/K^2$, when $\alpha=1$**
Power per chip	P	α	$fNCV^2/2 \rightarrow K^{-1}(\alpha K^{-2})K(K^1)^2 = \alpha$ **$= 1$, when $\alpha=1$**

$k = 0.7$ and $\alpha = 1$	$k = 0.7^2 = 0.5$ and $\alpha = 1$
Single MOFET	
Vdd → 0.7	Vdd → 0.5
Lg → 0.7	Lg → 0.5
Id → 0.7	Id → 0.5
Cg → 0.7	Cg → 0.5
P (Power)/Clock → $0.7^3 = 0.34$	P (Power)/Clock → $0.5^3 = 0.125$
τ (Switching time) → 0.7	τ (Switching time) → 0.5
Chip	
N (# of Tr) → $1/0.7^2 = 2$	N (# of Tr) → $1/0.5^2 = 4$
f (Clock) → $1/0.7 = 1.4$	f (Clock) → $1/0.5 = 2$
P (Power) → 1	P (Power) → 1

- The concerns for limits of down-scaling have been announced for every generation.

- However, down-scaling of CMOS is still the 'royal road'* for high performance and low power.

- Effort for the down-scaling has to be continued by all means.

---

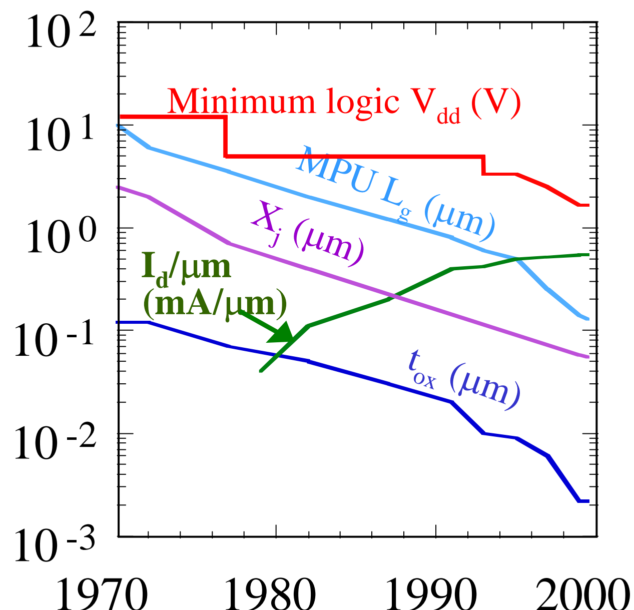
*Euclid of Alexandria (325BC?-265BC?)

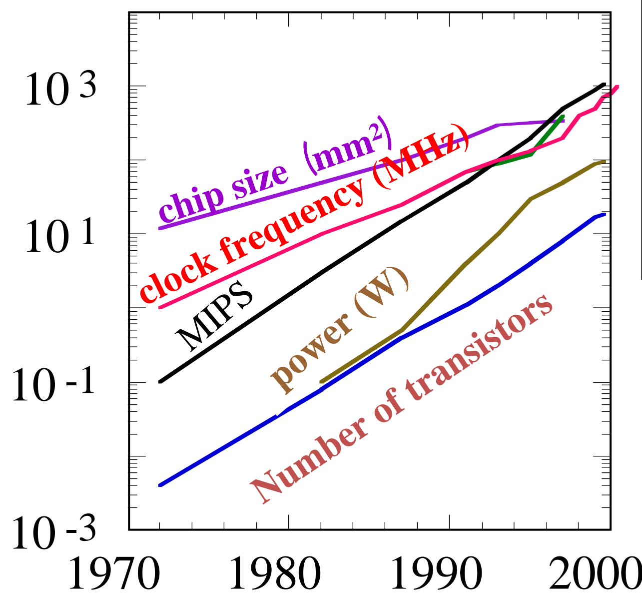
'There is no royal road to Geometry'

Mencius (Meng-zi), China (372BC?-289BC?)

孟子: 王道, 覇道 (Rule of right vs. Rule of military)

Actual past downscaling trend until year 2000





Past 30 years scaling

Merit: N, f increase

Demerit: P increase

V_{dd} scaling insufficient

⇩

Additional significant increase in

I_d, f, P

Source. Iwai and S. Ohmi, Microelectronics Reliability 42 (2002), pp.1251-1268

Change in 30 years

	Ideal scaling	Real Change		Ideal scaling	Real Change		Ideal scaling	Real Change
L_g	K	10^{-2}	I_d	K (10^{-2})	10^{-1}	f	$1/K(10^{2})$	10^3
t_{ox}	$K(10^{-2})$	10^{-2}	I_d/µm	1	10^1	P = $f\alpha NCV^2$	$\alpha(10^{1})$	10^5
V_{dd}	$K(10^{-2})$	10^{-1}	N	$\alpha/K^2(10^{5})$	10^4			
A_{chip}	α	10^1						

Vd scaling insufficient, α increased ➞ N, Id, f, P increased significantly

- Now, power and/or heat generation are the limiting factors of the down-scaling

- Supply voltage reduction is becoming difficult, because Vth cannot be decreased any more, as described later.

- Growth rate in clock frequency and chip area becomes smaller.

2. ITRS Roadmap (for 22 nm CMOS logic)

ITRS Roadmap does change every year!

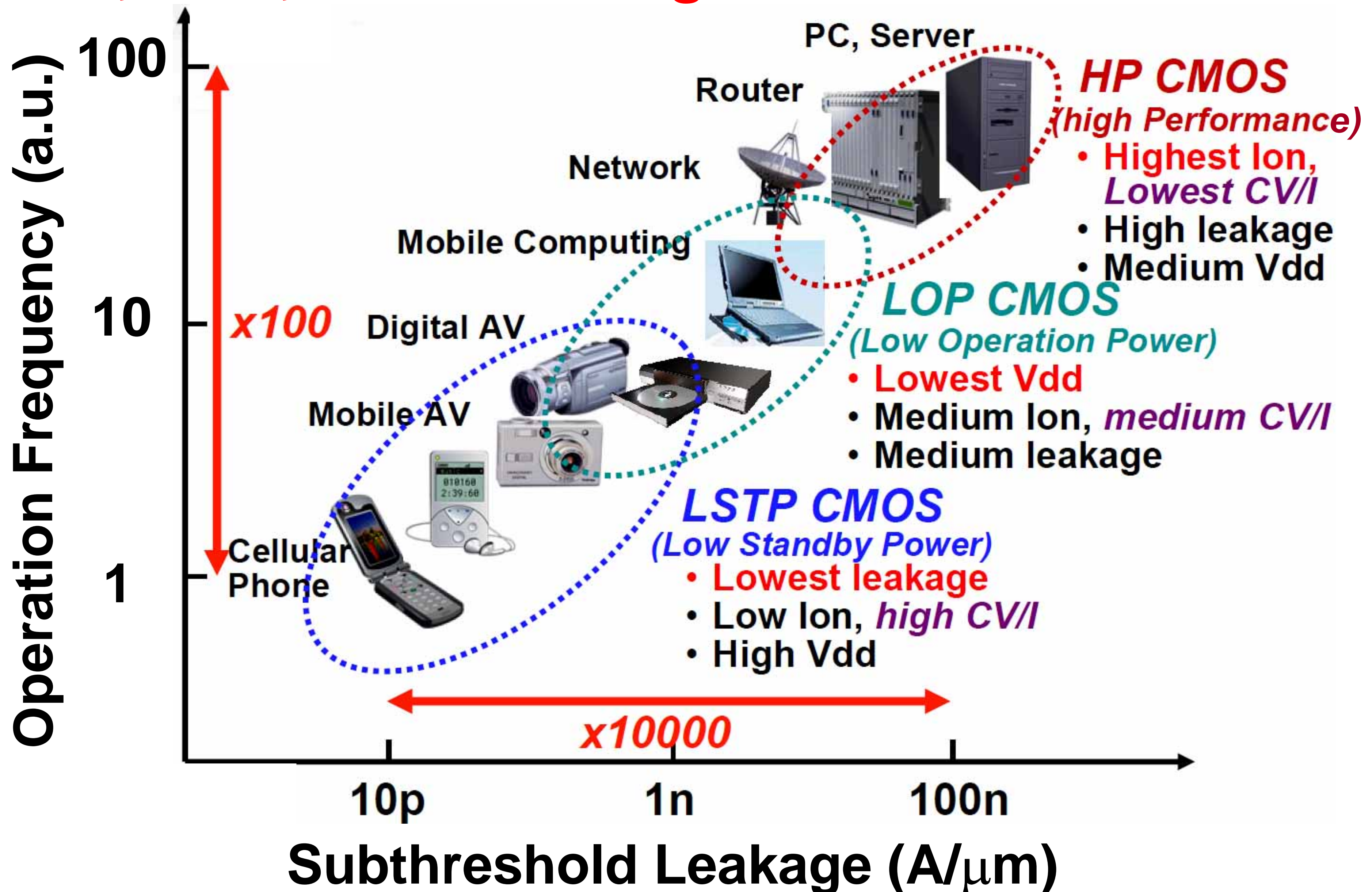
2007 Edition	2003 Edition
2006 Update	2002 Update
2005 Edition	2001 Edition
2004 Update	2000 Update

http://www.itrs.net/reports.html

The current latest version: ITRS 2007 Edition

ITRS 2008 Update will be published on the web at the end of Dec 2008 or Jan. 2009

HP, LOP, LSTP for Logic CMOS

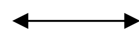
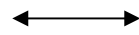
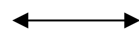
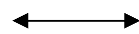


Source: 2007 ITRS Winter Public Conf.

What does '22 nm' mean in 22 nm CMOS Logic?

'XX nm CMOS Technology
Commercial Logic CMOS products

Technology name	Starting Year
45 nm	2007
32 nm	2009?
22 nm	2011?~ 2012?
16 nm	2013?~ 2014?

ITRS (Likely in 2008 Update)
for High Performance Logic

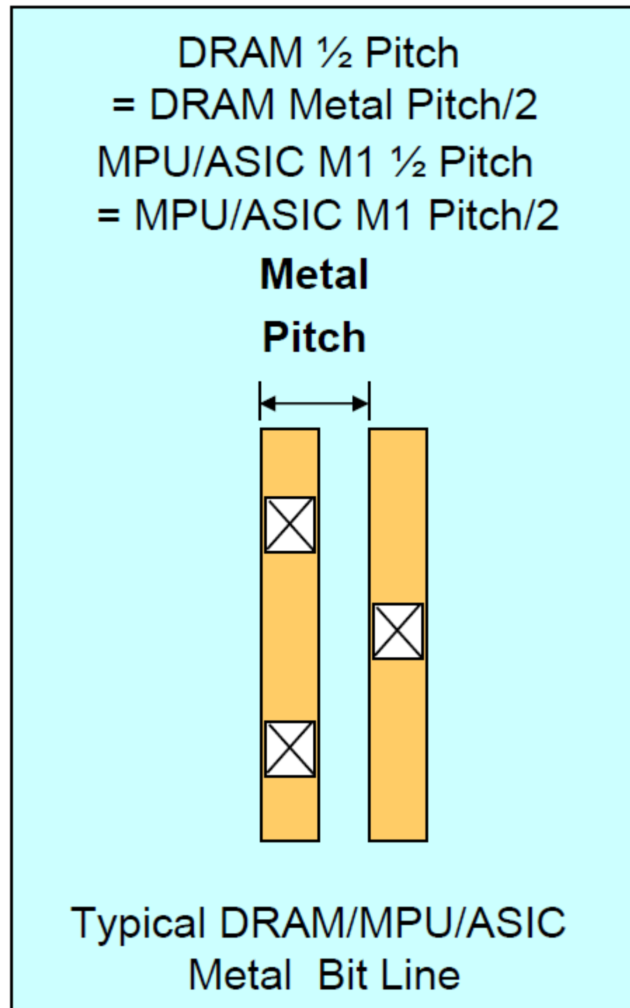
Year	Half Pitch (1st Metal)	Physical Gate Length
2007	68 nm	32 nm
2008	59 nm	29 nm
2009	52 nm	27 nm
2010	45 nm	24 nm
2011	40 nm	22 nm
2012	36 nm	20 nm
2013	32 nm	18 nm
2014	29 nm	16 nm

Source: 2008 ITRS Summer Public Conf.

'XX nm' CMOS Logic Technology:
- In general, there is no common corresponding parameter with 'XX nm' in ITRS table, which stands for 'XX nm' CMOS.

Definition of the Half Pitch

Logic 1st Metal Half Pitch



Source: 2008 ITRS Summer Public Conf.

What does '22 nm' mean in 22 nm CMOS Logic?

8μm → 6μm → 4μm → 3μm → 2μm → 1.2μm → 0.8μm → 0.5μm

- Originally, 'XX' means lithography resolution.
- Thus, 'XX' was the gate length, and half pitch of lines
- 'XX' had shrunk 0.7 in 3 years in average (0.5 in 6 years) those days.
- 'XX' value deviated among companies: example:**1.5μm, 1.2μm, 1μm**

→ 350nm → 250nm → 180nm → 130nm → 90nm → 65nm → 45nm

- 'XX' values were established by NTRS* and ITRS with the term of 'Technology Node**' and 'Cycle***' using typical 'half pitch value'.

  *NTRS: National Tech. Roadmap, **Term 'Technology Node' is not used now.

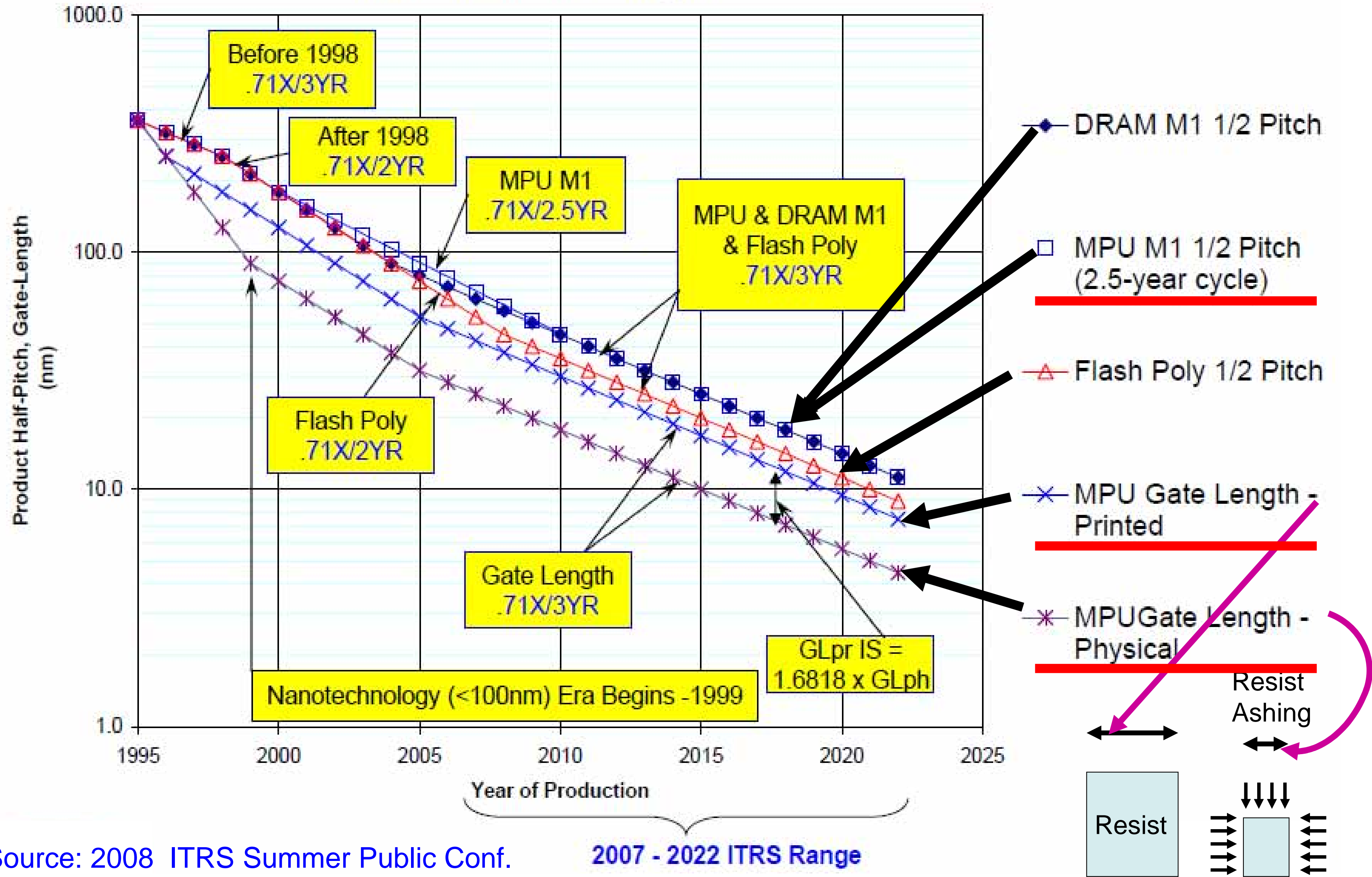
  ***Cycle: Period or year for which the half pitch becomes X0.71.

- The gate length of logic CMOS became smaller with one or two generations from the half pitch, and 'XX' names ahead of generations have been used for logic CMOS.

- Memory still keeps the half pitch as the value of 'XX'

→ 32nm → 22nm → 16nm → 11nm → 8nm?? → 5.5nm ??

For example, Typical Half Pitches at ITRS 2007

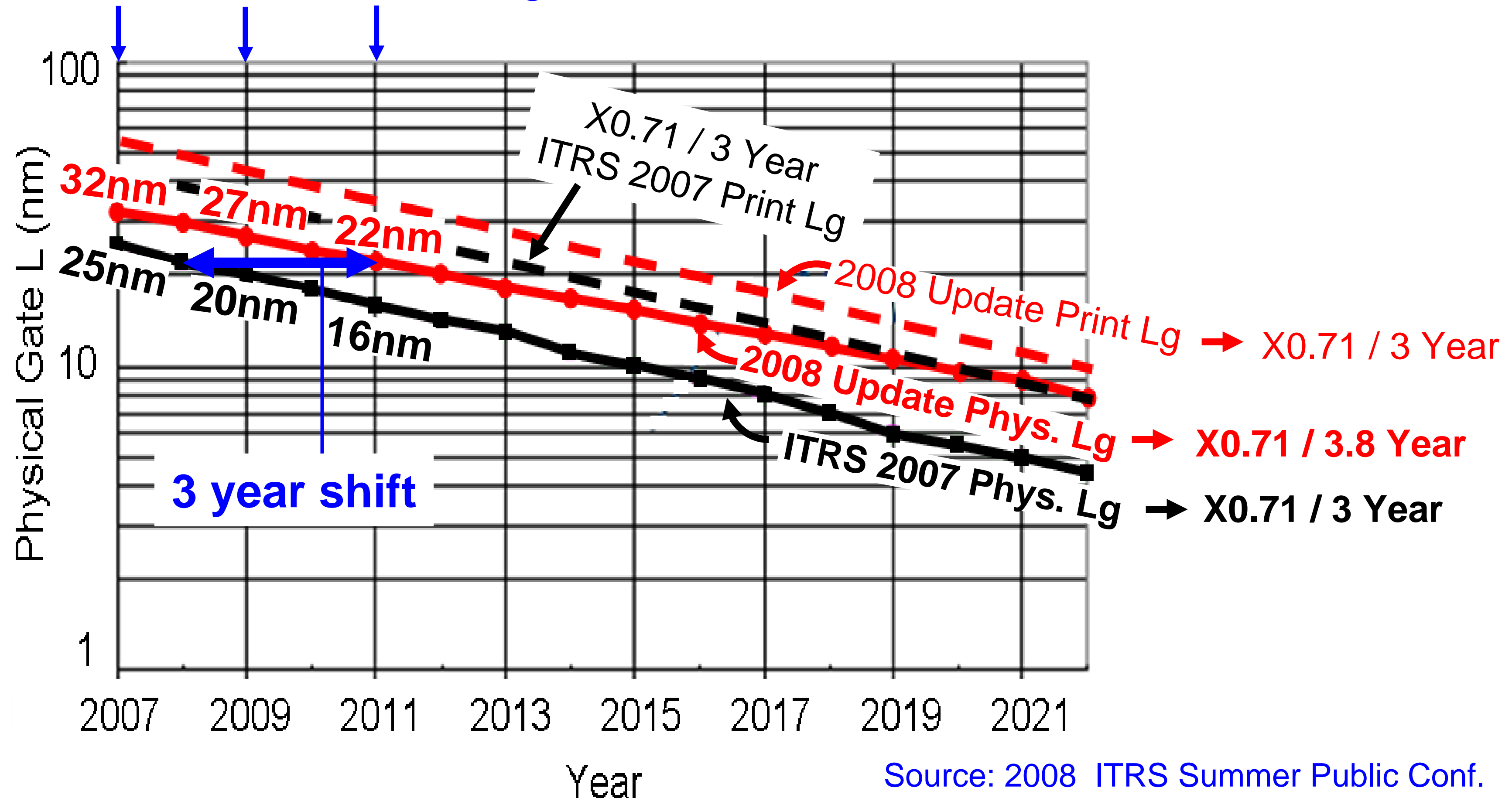


Source: 2008 ITRS Summer Public Conf.

Physical gate length in past ITRS was too aggressive.

The dissociation from commercial product prediction will be adjusted.

Physical gate length of High-Performance logic will shift by 3-5 yrs.



Source: 2008 ITRS Summer Public Conf.

EOT and Xj shift backward, corresponding to Lg shift

EOT: 0.55 nm → 0.88 nm, Xj: 8 nm → 11 nm @ 22nm CMOS

Likely in 2008 Update

Correspond to 22nm

Source: 2008/ ITRS Summer Public Conf.

Year of Production	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2007 MPU/ASIC Lg (nm)	25	23	20	18	16	14	13	11	10	9	8	7	6.3	5.6	5	4.5
2008 MPU/ASIC Lg (nm)	32	29	27	24	22	20	18	17	15	14.0	12.8	11.7	10.7	9.7	8.9	8.1
Shift/Interpolate Formua	2005	intrp	intrp	intrp	intrp	2009	2010	intrp	intrp	2012	intrp	intrp	intrp	intrp	intrp	intrp
EOT w/3E20 poly, bulk MPU (nm)	1.2	0.71	0.54	0.41												
EOT w/3E20 poly, bulk MPU (nm)	1.3	1.2	1.2	1	0.68	0.54	0.41									
EOT w/metal gate, bulk MPU (nm)		0.9	0.75	0.65	0.55	0.50										
EOT w/metal gate, bulk MPU (nm)			1.0	0.95	0.88	0.75	0.65	0.60	0.53	0.5						
Drain Ext. X_j bulk MPU (nm)	12.5	11	10	9	8	7										
Drain Ext. X_j bulk MPU (nm)	11	11	11	11	11		9	8.5	7.7	7						

Likely in 2008 Update

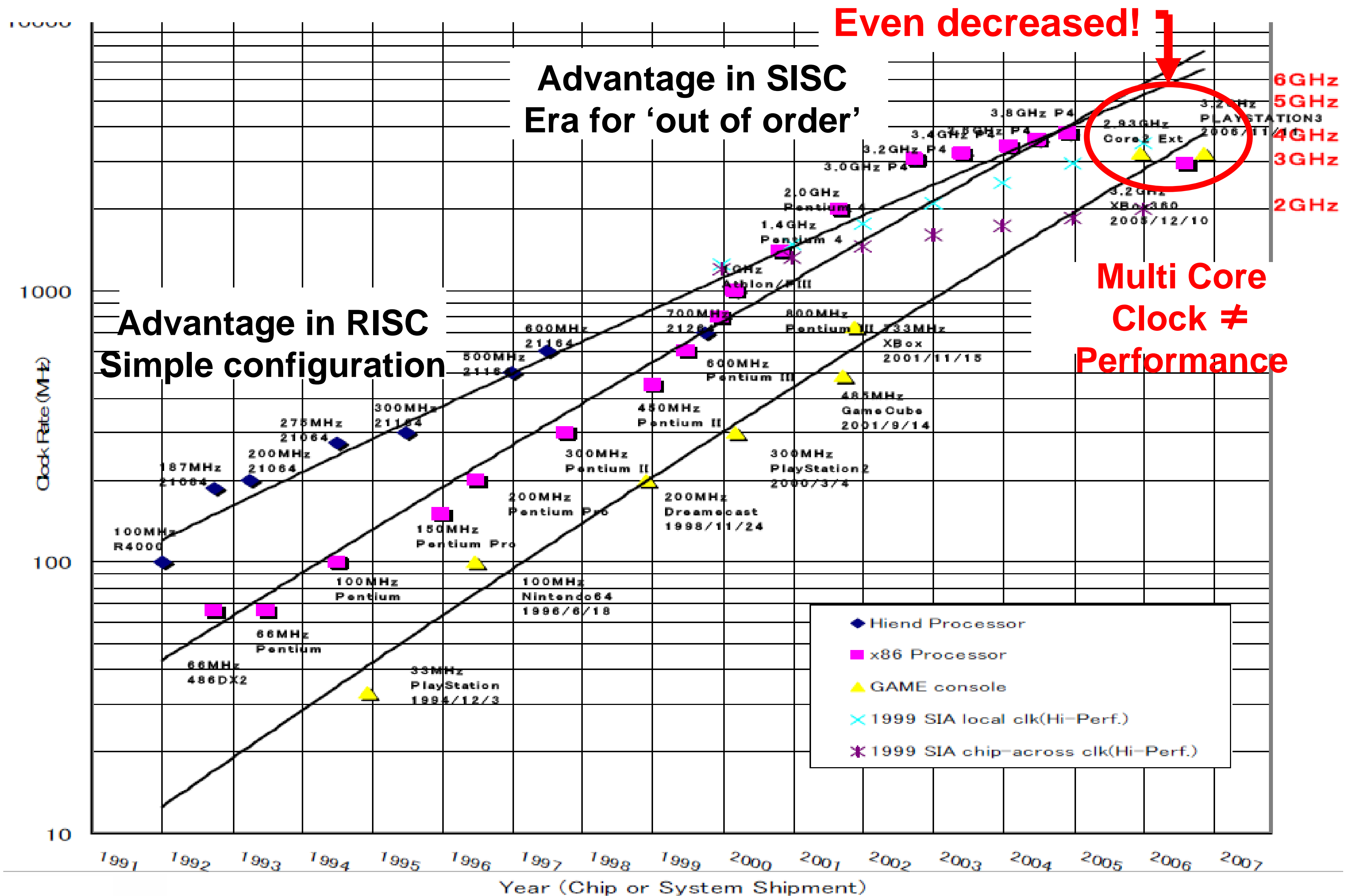
Likely in 2008 Update

Likely in 2008 Update

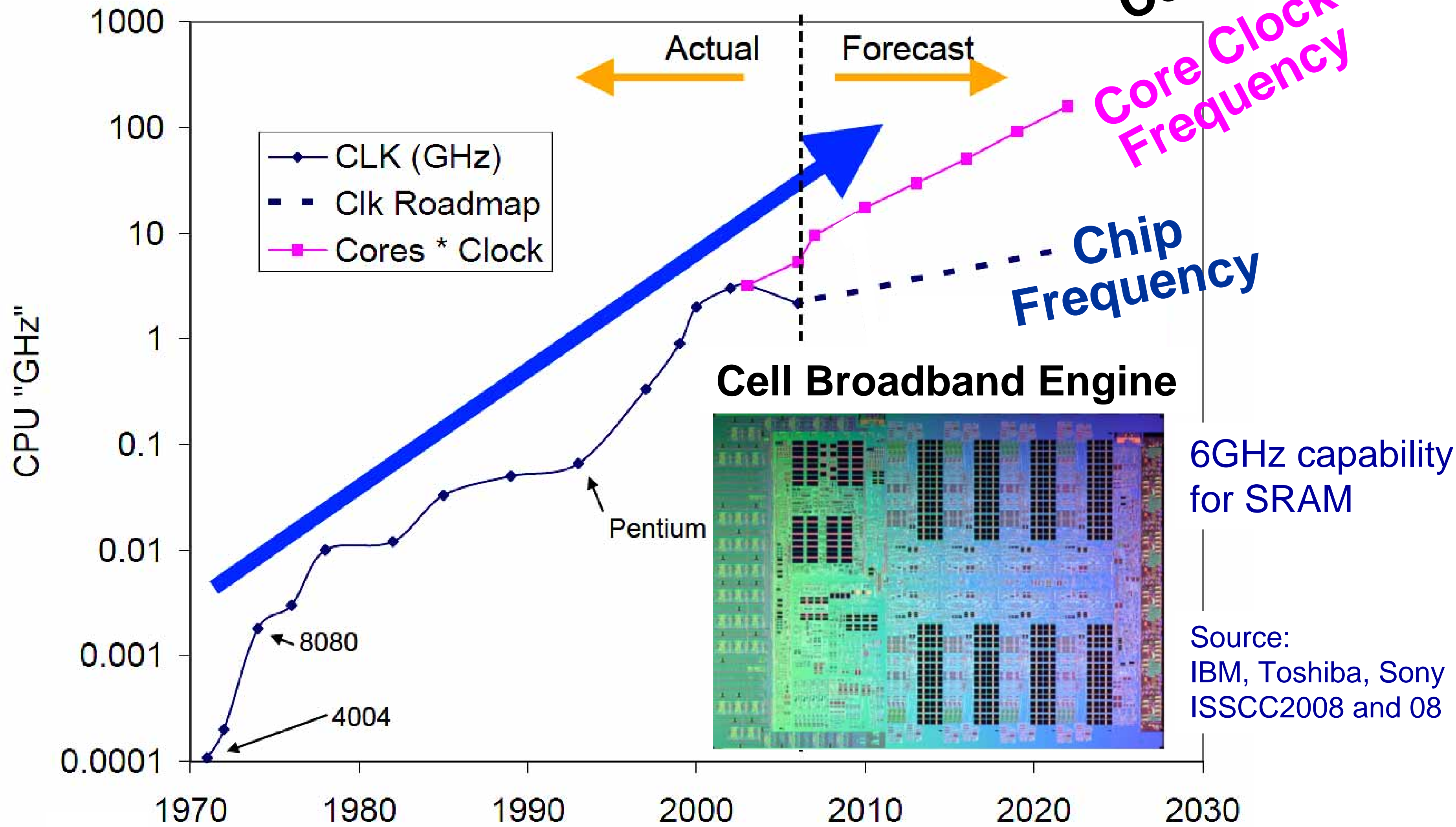
non-steady trend corrected

filled in for metal gate EOT for 2009/10 based on latest conference presentations

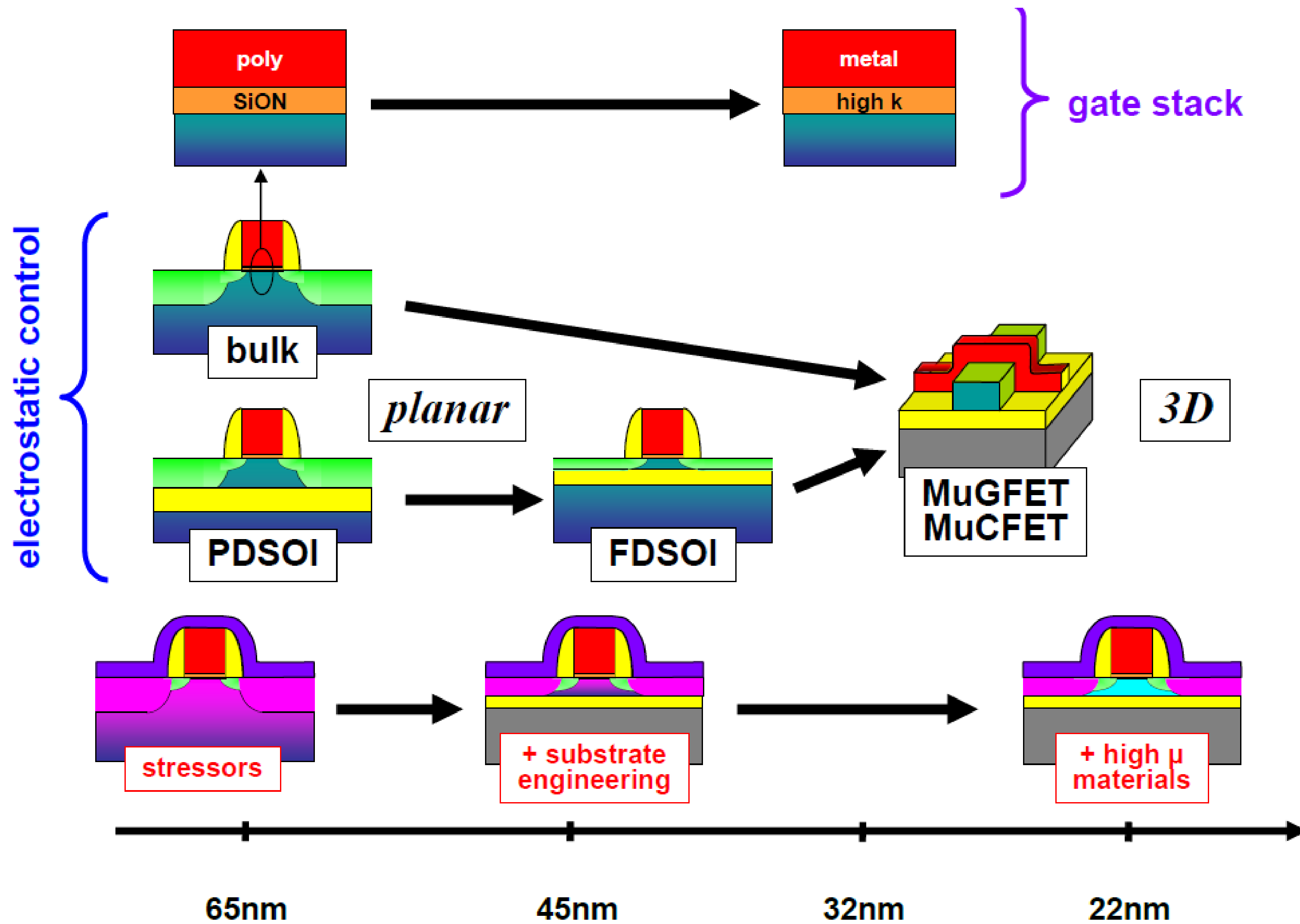
Clock frequency does not increase aggressively anymore.



Source: Mitsuo Saito, Toshiba

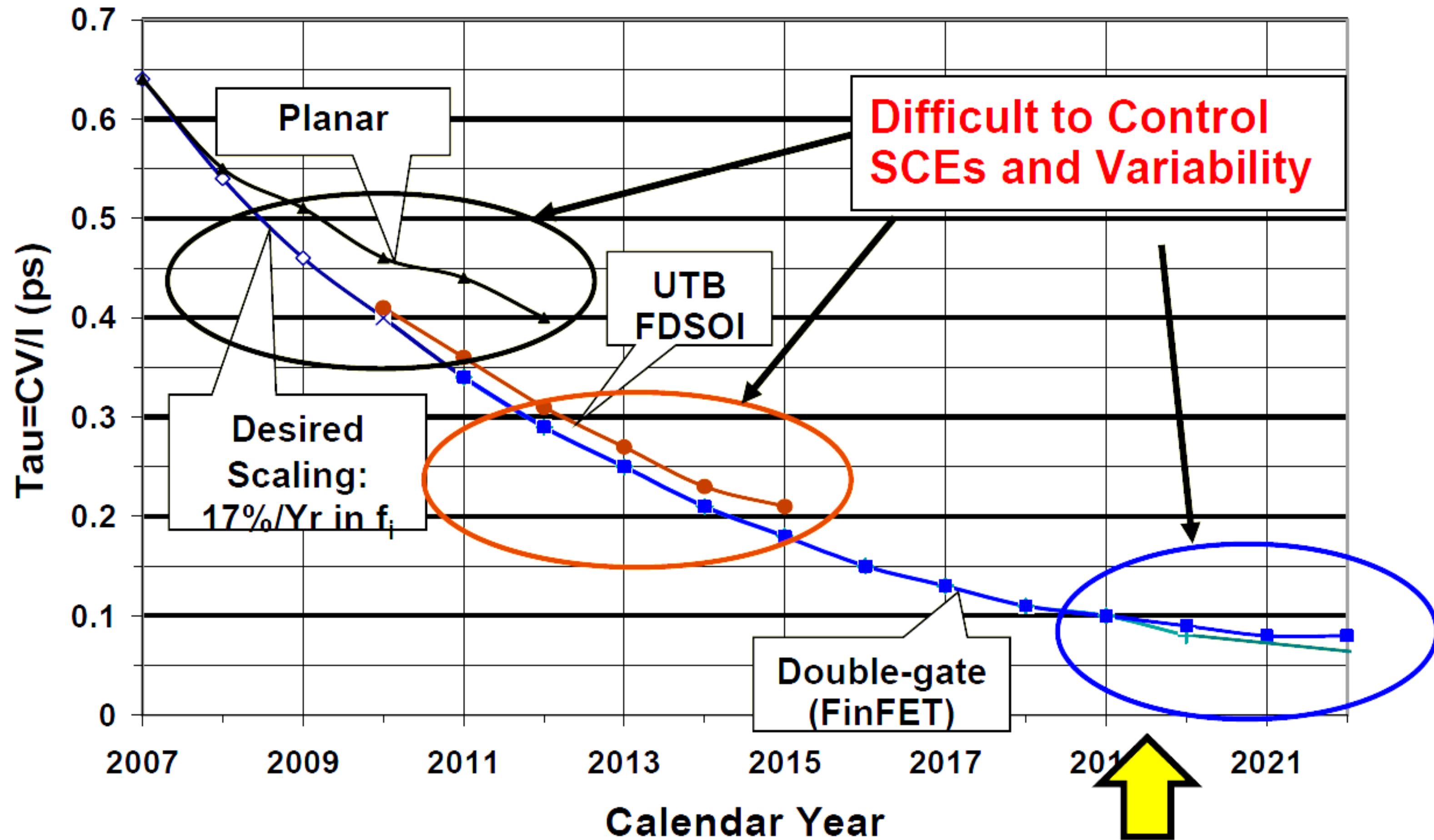

MPU "GHz" by "Cores" ITRS2007
Continued?
Actual
Forecast
CLK (GHz)
Clk Roadmap
Cores * Clock
Core Clock Frequency
Chip Frequency
Cell Broadband Engine
6GHz capability for SRAM
Source:
IBM, Toshiba, Sony
ISSCC2008 and 08
CPU "GHz"
1000
100
10
1
0.1
0.01
0.001
0.0001
1970
1980
1990
2000
2010
2020
2030
Pentium
8080
4004
Source: 2007 ITRS Winter Public Conf.

Structure and technology innovation (ITRS 2007)



Source: 2008 ITRS Summer Public Conf.

Technology innovation described in ITRS 2007

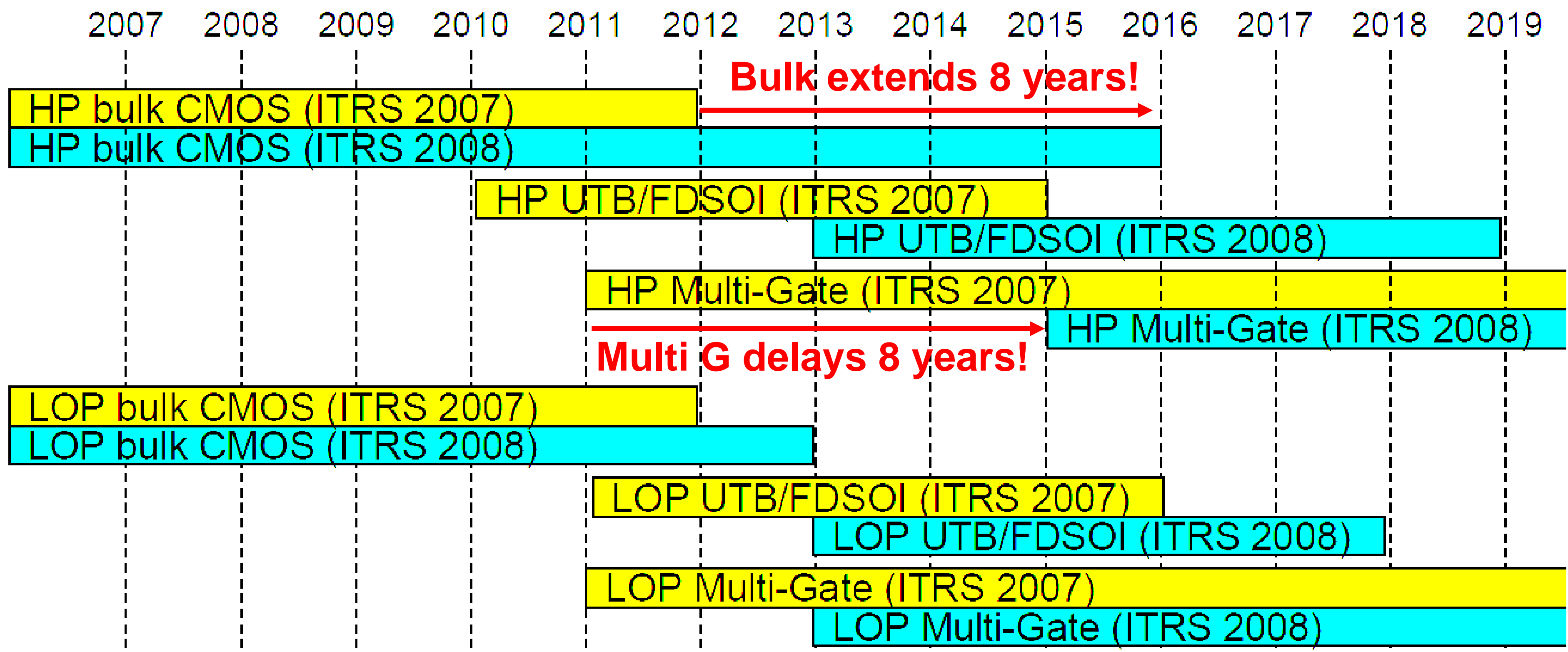


Alternative material (Ge, III-V) and structure (Nanowire) in channel region.

Source: 2007 ITRS Winter Public Conf.

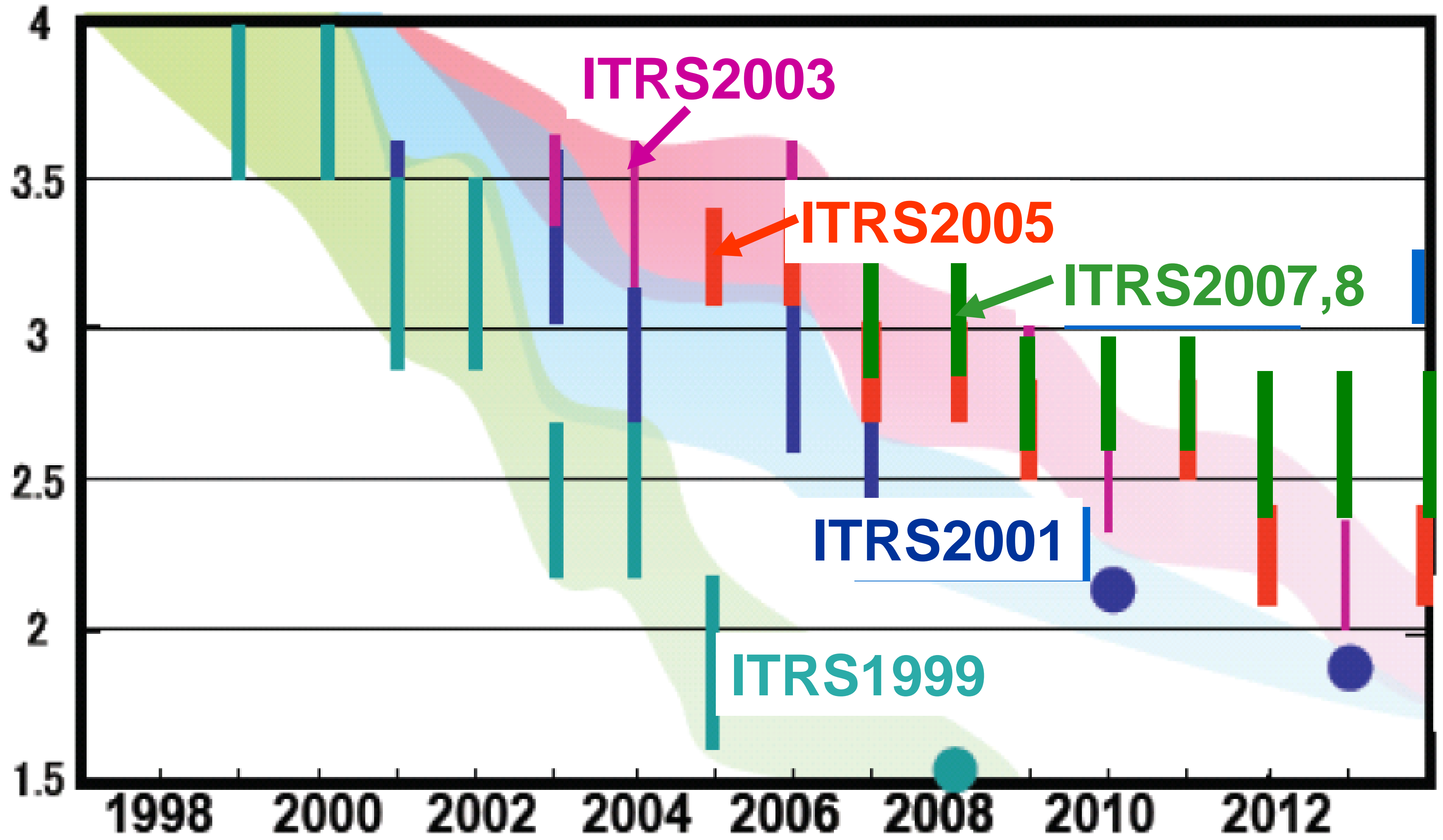
Timing of CMOS innovations shifts backward.

Bulk CMOS has longer life now!



Source: 2008 ITRS Summer Public Conf.

Historical Transition of ITRS Low-k Roadmap



Source: 2008 ITRS Summer Public Conf.

Roadmap towards 22nm technology and beyond

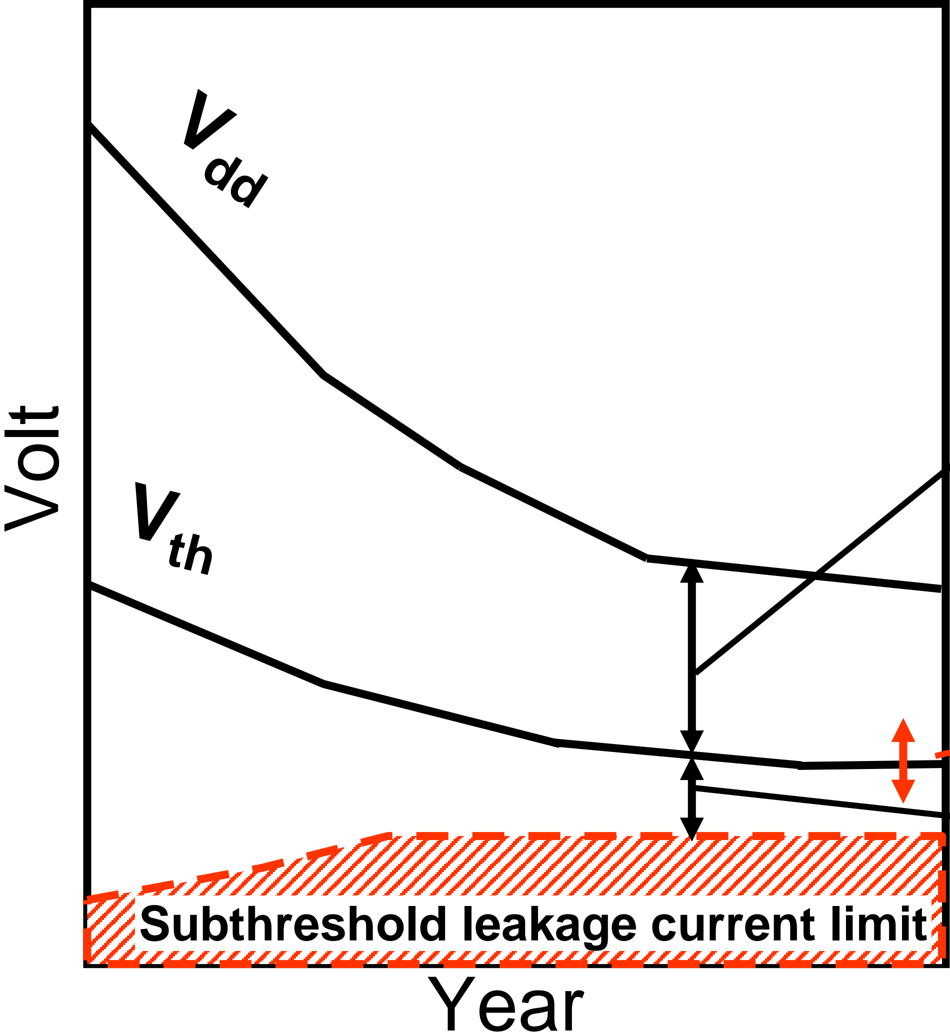
- Physical gate length downsizing rate will be less aggressive.

- Corresponding to the above, performance increase would slow down – Clock frequency, etc.

- Introduction of innovative structures – UTB SOI and DG delayed, and bulk CMOS has longer life than predicted by previous ITRS roadmaps.

3. Voltage Scaling / Low Power and Leakage

Difficulty in Down-scaling of Supply Voltage: Vdd



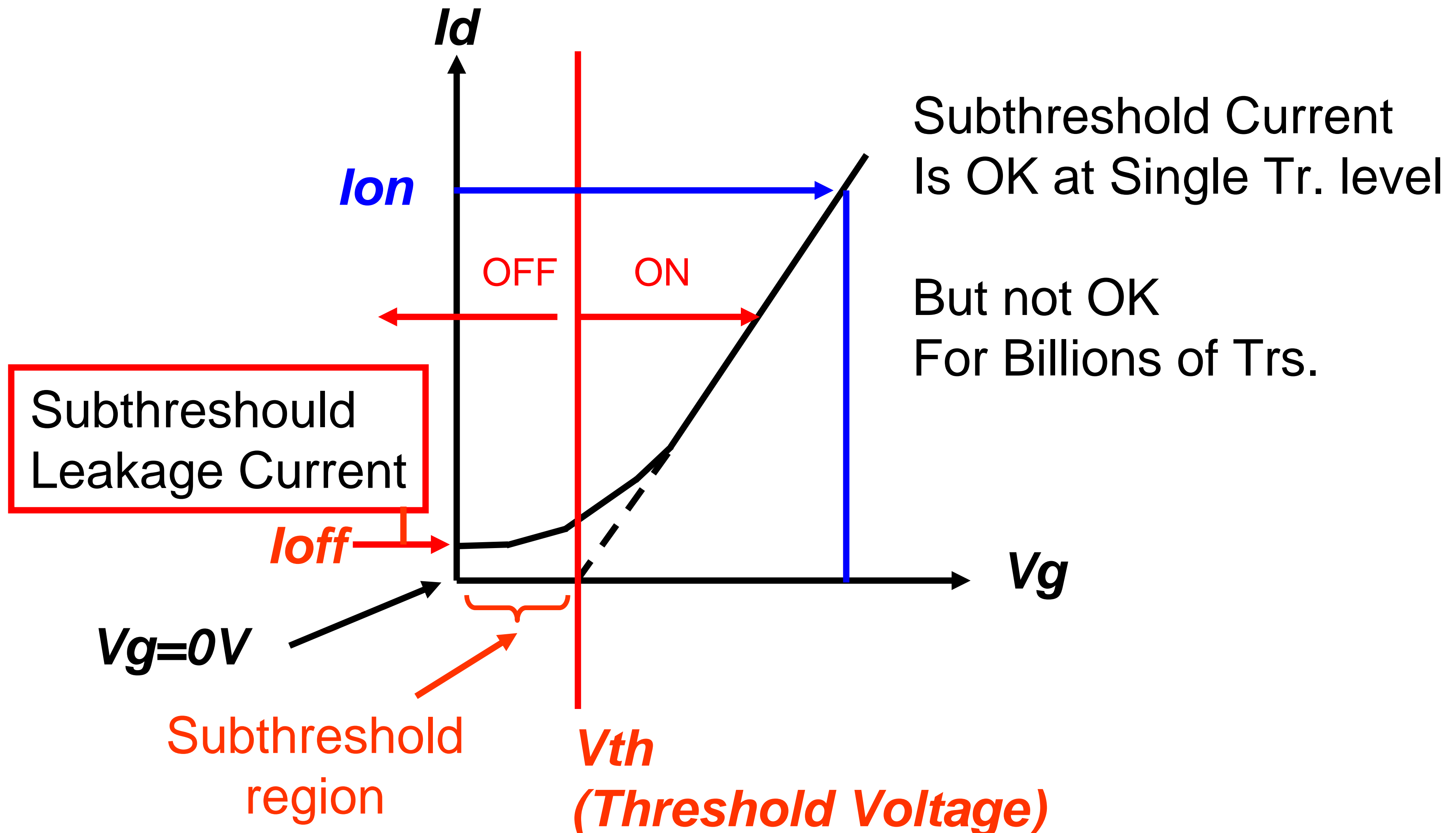
Because, V_{th} cannot be down-scaled anymore, V_{dd} down-scaling is difficult.

$V_{dd} – V_{th}$ determines the performance (High Id) and cannot be too small.

ΔV_{th}: V_{th} variation

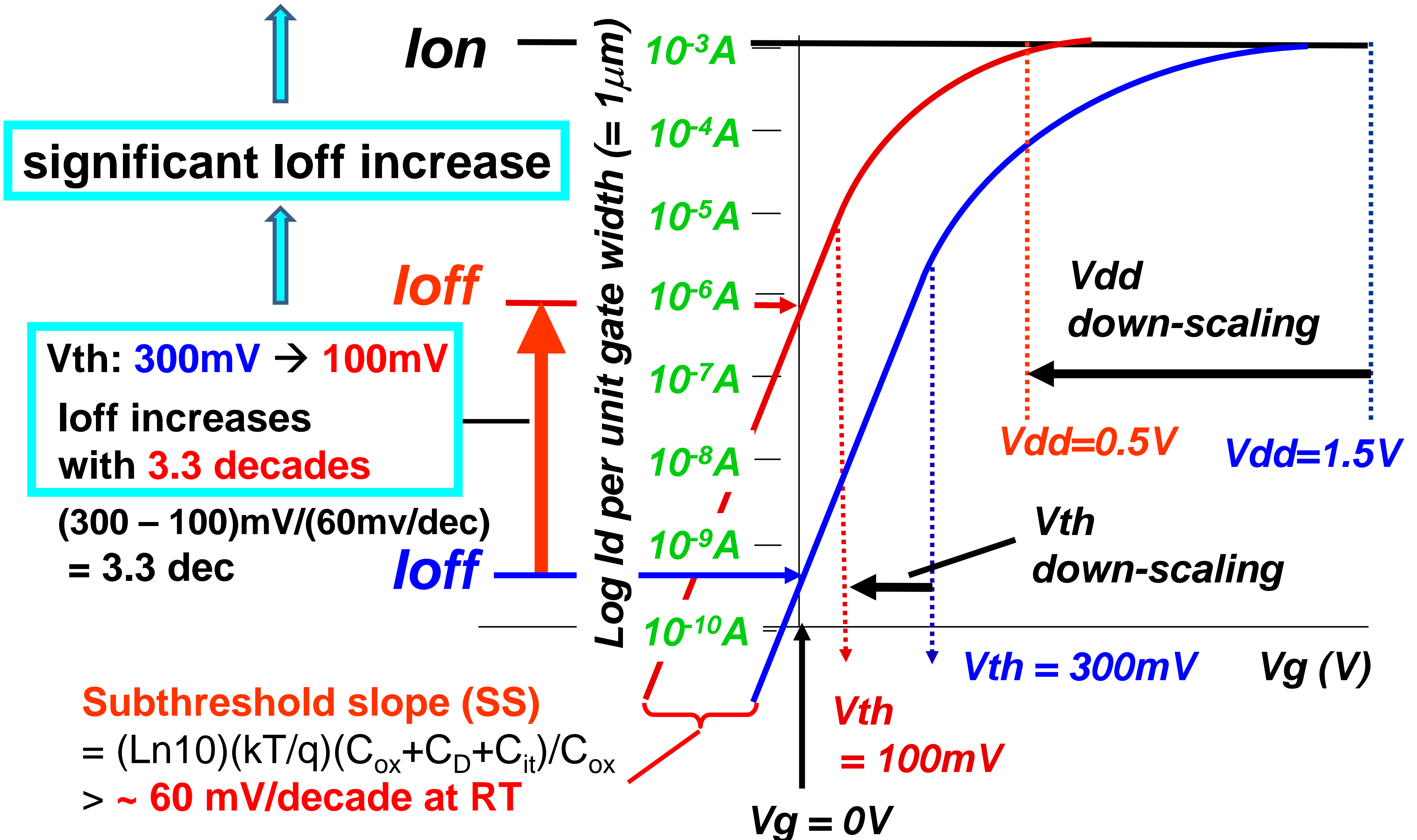
$> \Delta V_{th}$
Margin for V_{th} variation is necessary

Subtheshold leakage current of MOSFET



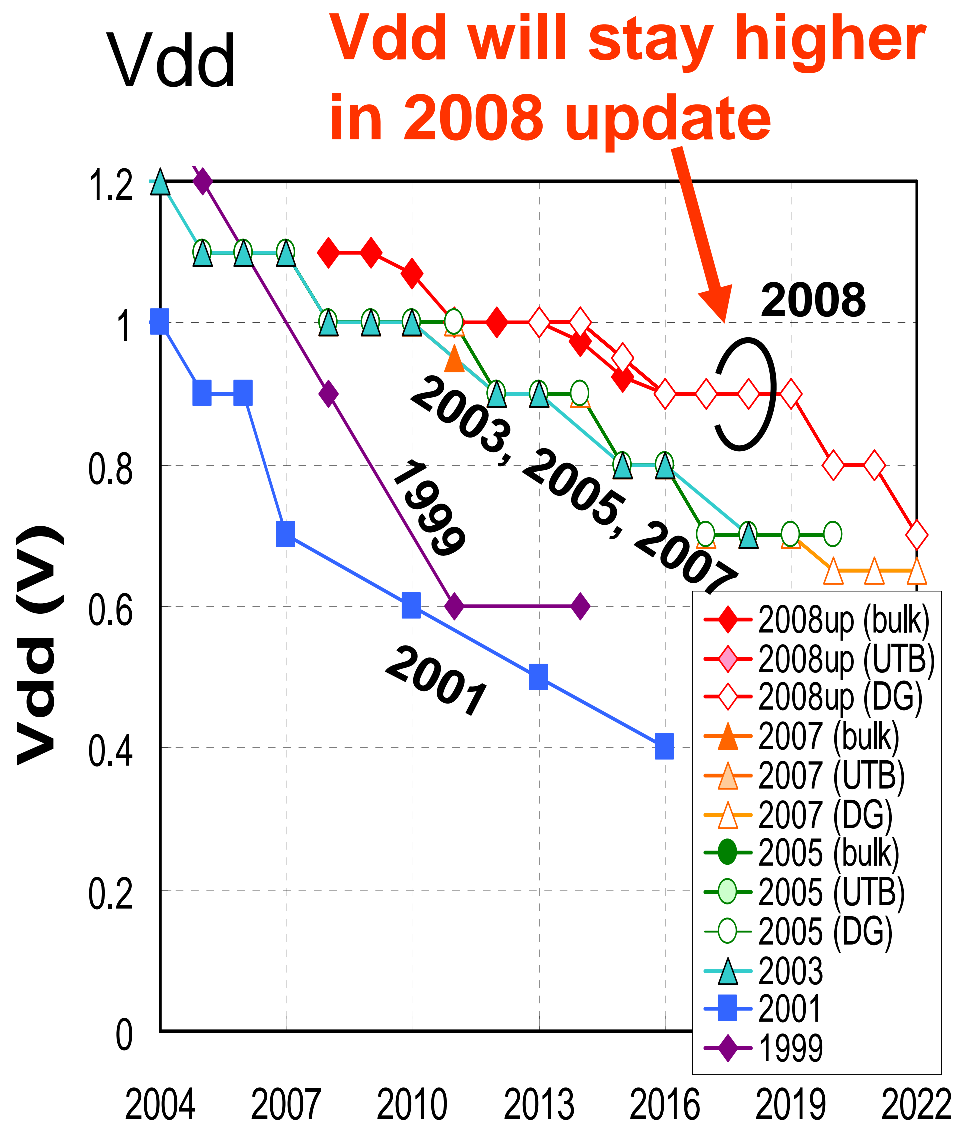
Vth cannot be decreased anymore

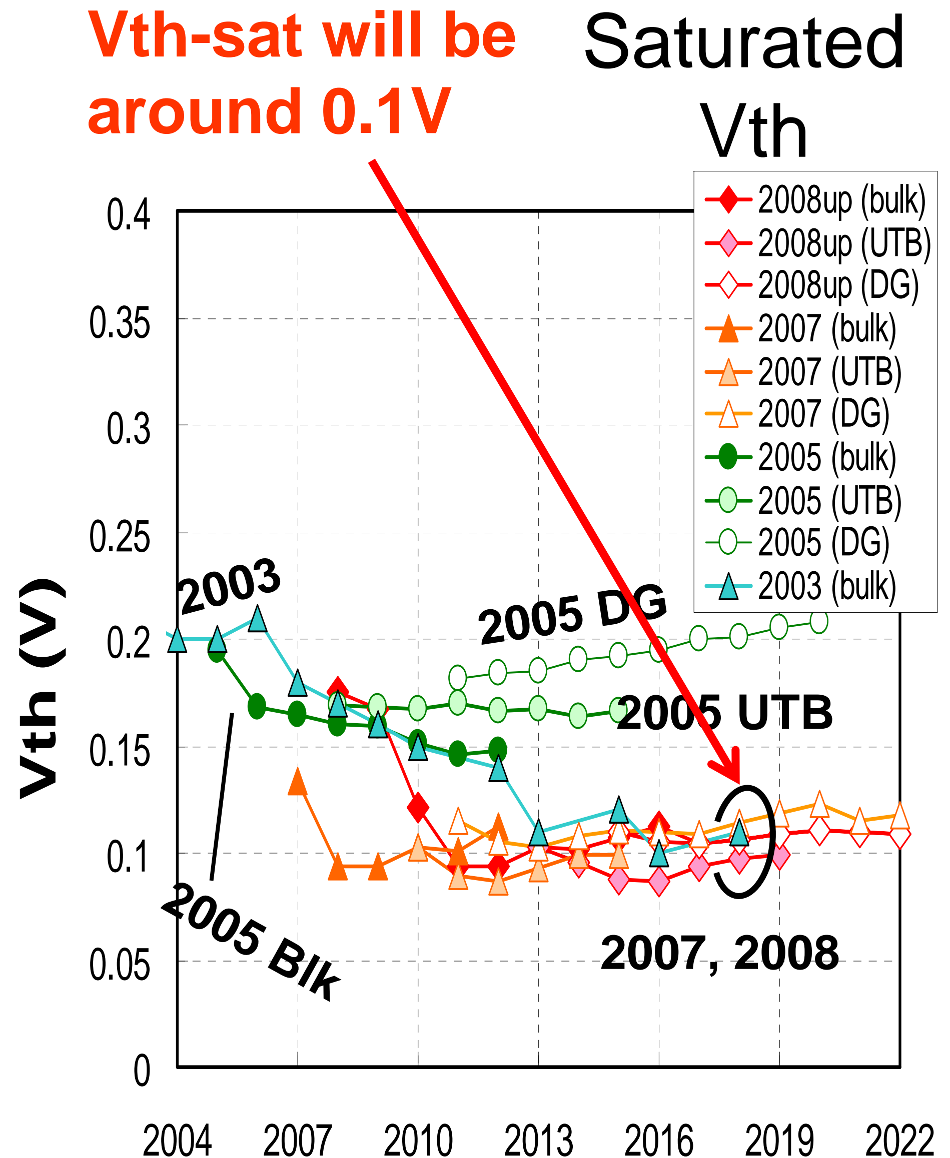
Log scale Id plot



SS value:
Constant and does not become small with down-scaling

ITRS for HP logic

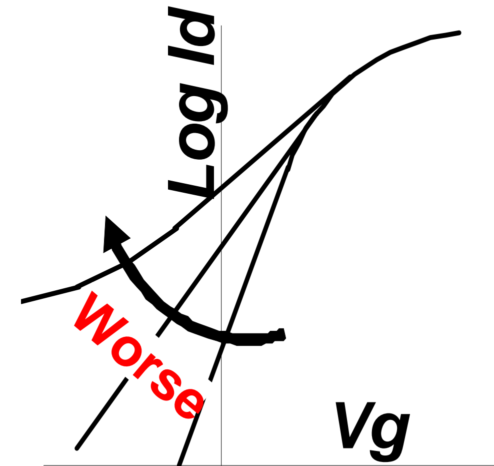




2008 Values are from ITRS Public Conf. and still under discussion

Source: ITRS and 2008 ITRS Summer Public Conf.

SS (Subtheshold Slope) becomes worse in the following cases



1. Improper down-scaling

Ex. When T_{ox}, W_{dep}, or V_{dd} is not scaled

2. High impurity doping in channel or substrate

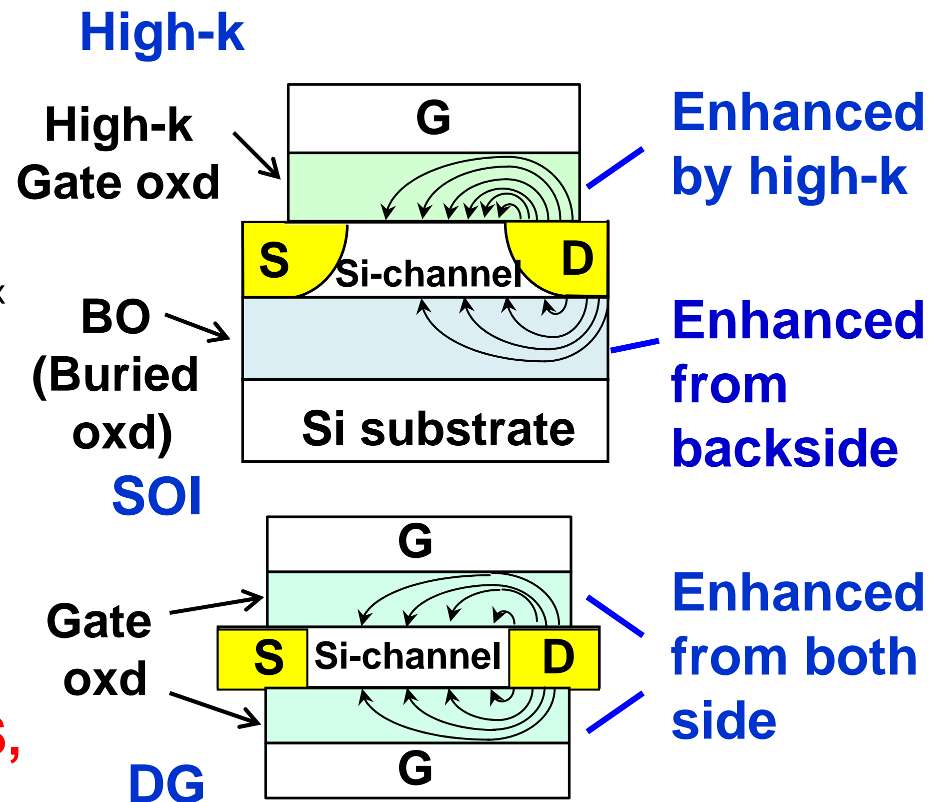
High impurity Conc.
→ C_D increase
→ SS increase

SS = (Ln10)(kT/q)(C_{ox}+C_D+C_{it})/C_{ox}

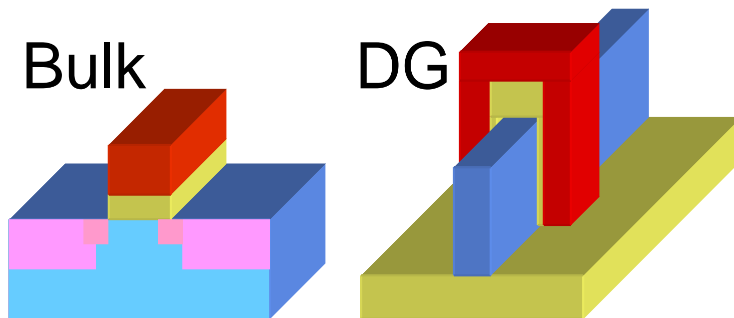
3. Enhanced Drain-Electric-field penetration through oxide

Ex. High-k, SOI,
Multi-gate (Double gate: DG)

DG and SOI often show better SS, but be careful!



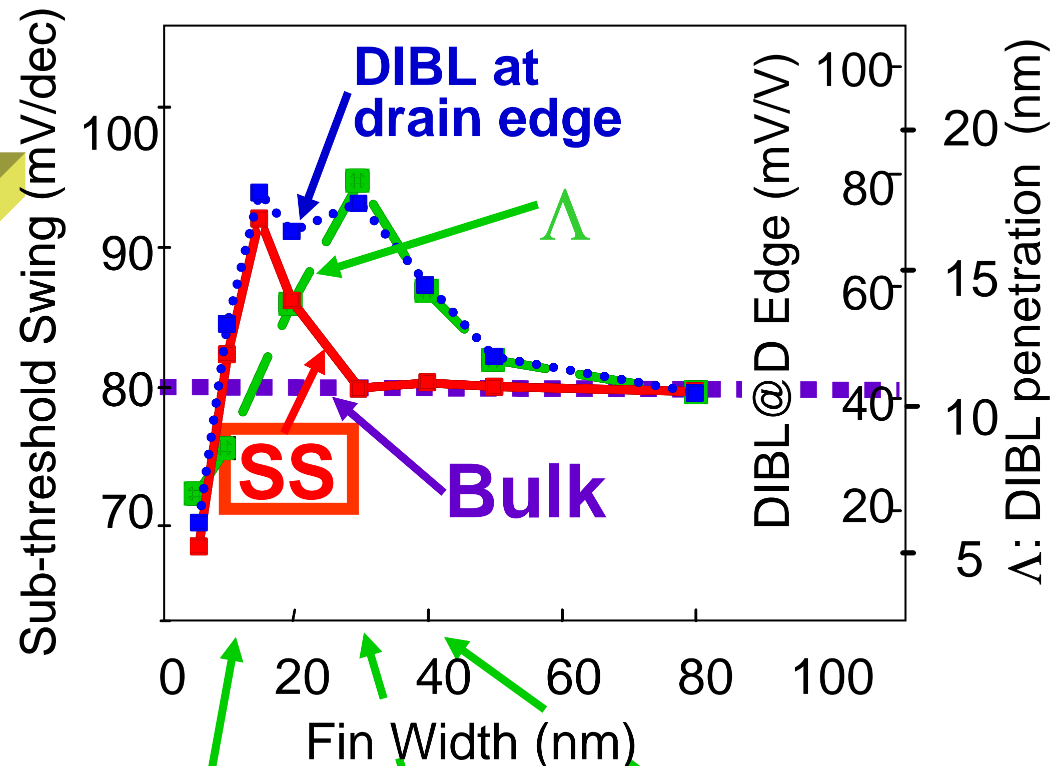
Comparison of Bulk and DG



Same parameter condition for both

(2006 ITRS Bulk parameters are used for both Bulk and DG)

Lg=16nm, tox(EOT)=0.5nm,
Dopant@Channel=$8.1 \times 10^{18} cm^{-2}$



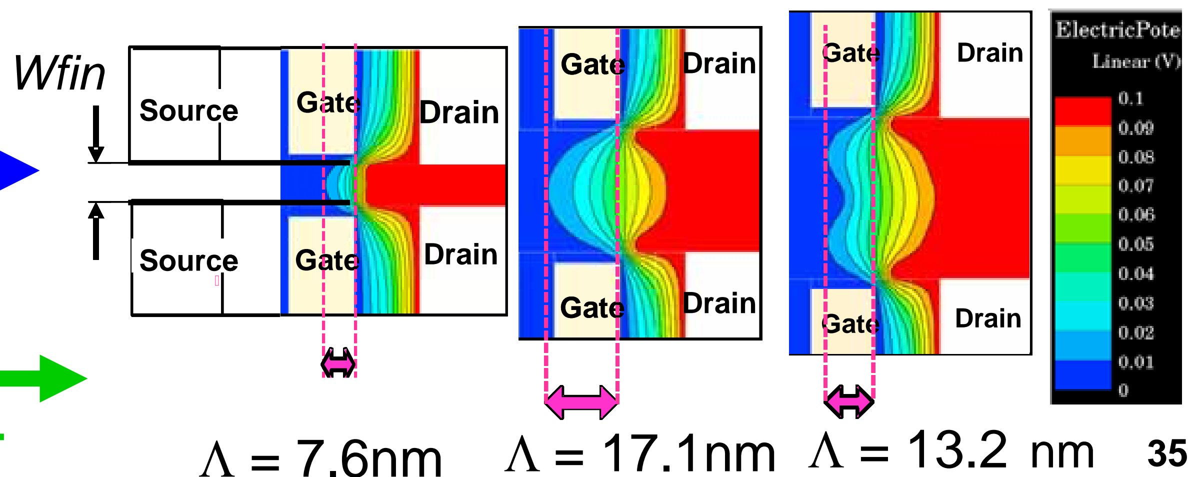
Source: ECS Fall Meeting, Oct 2008, Honolulu,
Y. Kobayashi, A. B. Sachid, K. Tsutsui, K. Kakushima,
P. Ahmet, V. Ramgopal Rao and H. Iwai.

DIBL: Drain Induced Barrier Lowering

$$\left.\frac{\partial V(x,y)}{\partial Vd}\right|_{Vd=1V}$$

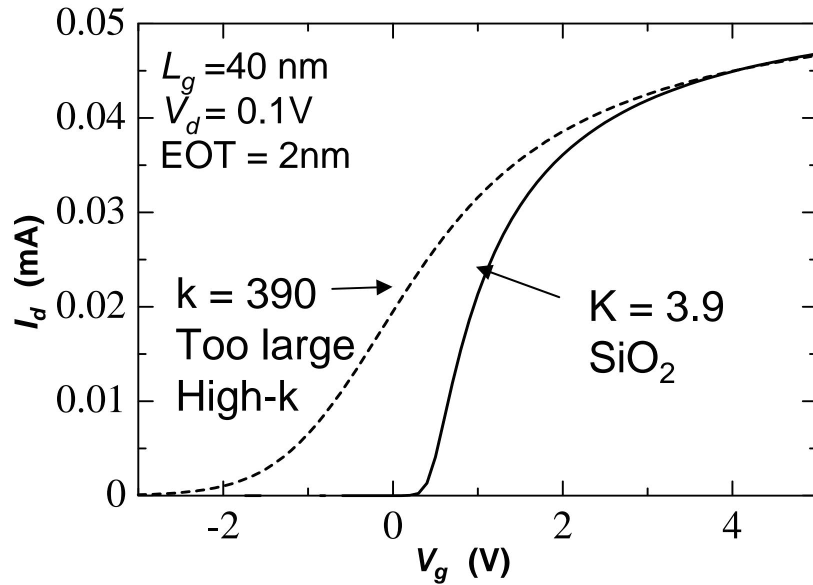
Λ: Penetration Depth of DIBL

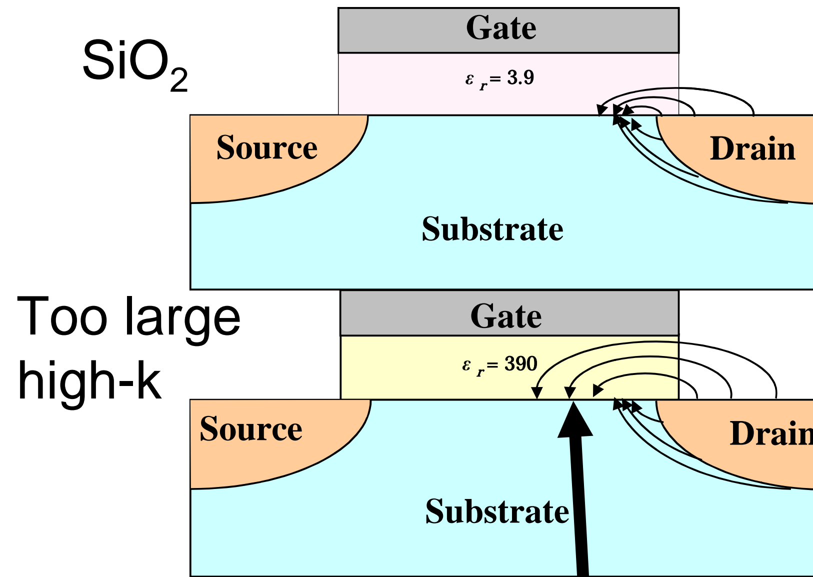
Wfin = 10.7 nm *Wfin* = 30 nm *Wfin* = 40 nm

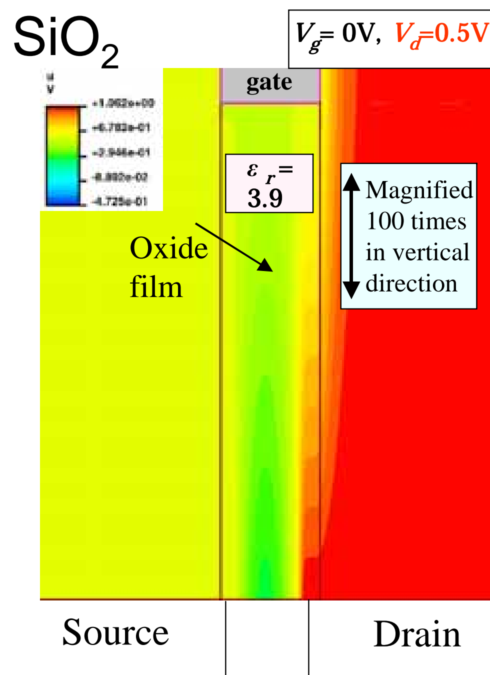


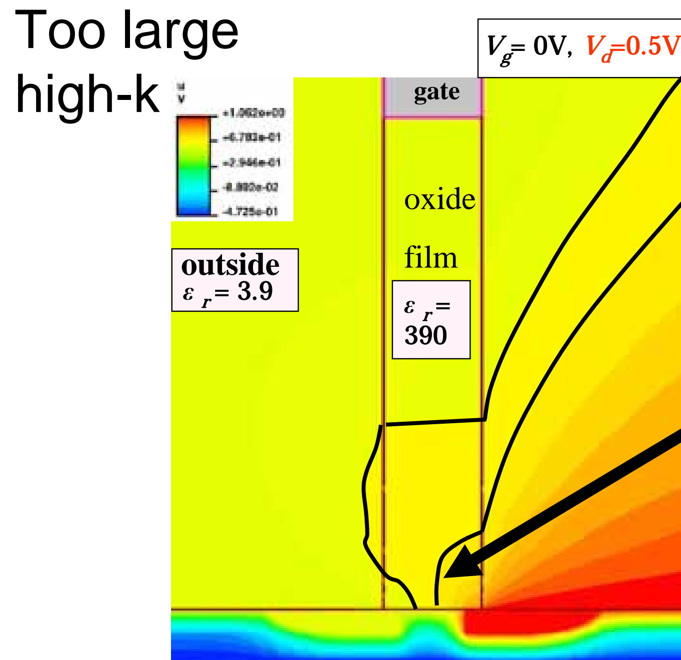
Λ = 7.6nm Λ = 17.1nm Λ = 13.2 nm

Comparison of High-k and SiO_2 MOSFETs









Penetration of lateral field from Drain through high-k causes significant short channel effects

R. Fujimura, M. Takeda, K. Sato, S. Ohmi, H. Ishiwara, and H. Iwai, ECS Symp. on ULSI Process Integration II, Volume 2001-2, pp.313-323, 2001,

V_{dd} will stay higher than predicted by previous ITRS roadmaps.

Solution towards Low V_{dd}

Effort to reduce $I_{sd\text{-}leak}$ and increase $I_{d\text{-}sat}$ is important

- Scaling: Proper down-scaling
  - Introduction of Next generation high-k, S/D etc.
  - CD* variation control by lithography and etching techniques

*** CD: Critical dimension**

- Structure: Bulk → UTB-SOI → DG → Nanowire
- Variation: Proper scaling by new tech. – High-k, litho. Etc.

  V_{th} adjustment by V_{sub} control
- Circuit techniques: Dynamic and local Multi-V_{dd}, etc.

Another concern for Low V_{dd} besides $I_{sd\text{-}leak}$ increase

→ Huge power loss for voltage conversion to such low V_{dd}

4. SRAM cell scaling

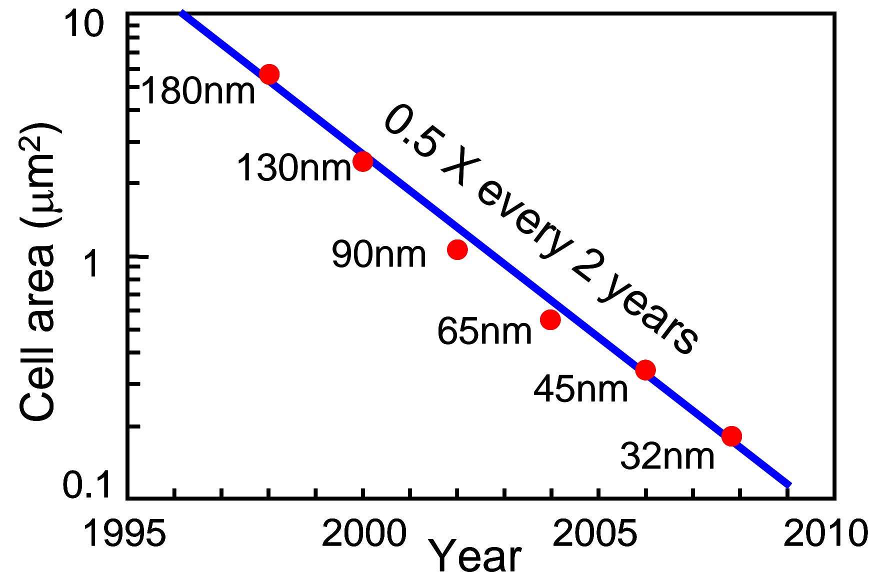
Intel's **SRAM** test chip trend

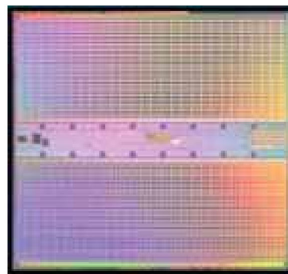
Source: B. Krzanich, S. Natrajan, Intel Developer's Forum 2007
http://download.intel.com/pressroom/kits/events/idffall_2007/Briefing
Silicon&TechManufacturing.pdf

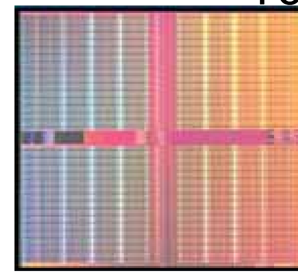
Process name	Lithography	1st production
P1264	65nm	2005
P1266	45nm	2007
P1268	32nm	2009
P1270	22nm	2011

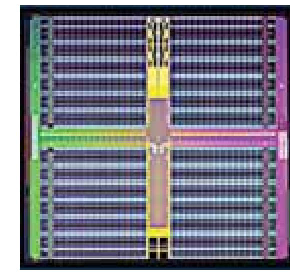
Only schedule has been published

SRAM down-scaling trend has been kept until 32nm and probably so to 22nm









Technology	**90 nm Process**	**65 nm Process**	**45 nm Process**	**32 nm Process**
Cell size	**1.0 μm²cell**	**0.57 μm²cell**	**0.346 μm²cell**	**0.182 μm²cell**
Capacity	50 Mbit	70 Mbit	153 Mbit	291 Mbit
Chip area	109 mm²	110 mm²	119 mm²	118 mm²
Functional Si	February '02	April '04	January '06	September '07

22 nm technology 6T SRAM Cell: Size = 0.1μm

Source: http://www-03.ibm.com/press/us/en/pressrelease/24942.wss

Announced on Aug 18, 2008

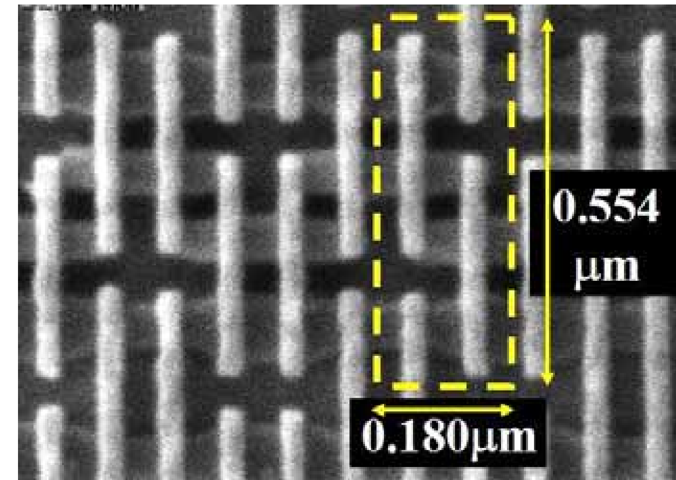
Consortium: IBM (NYSE) , AMD, Freescale, STMicroelectronics, Toshiba and the College of Nanoscale Science and Engineering (CNSE)

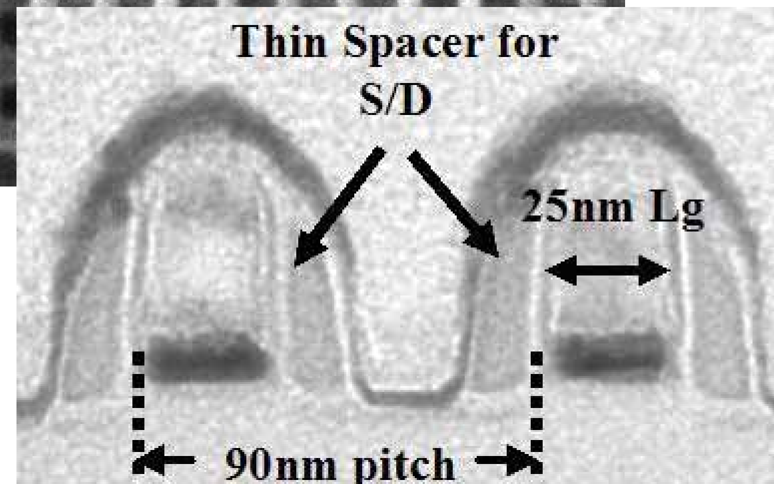
0.1μm cell size is almost on the down-scaling trend

New technologies introduced

- High-NA immersion lithography
- High-K metal gate stacks
- **25 nm gate lengths**
- Thin composite oxide-nitride spacers
- Advanced activation techniques
- Extremely thin silicide
- Damascene copper contacts

Static noise margin of 220 mV at 0.9 V

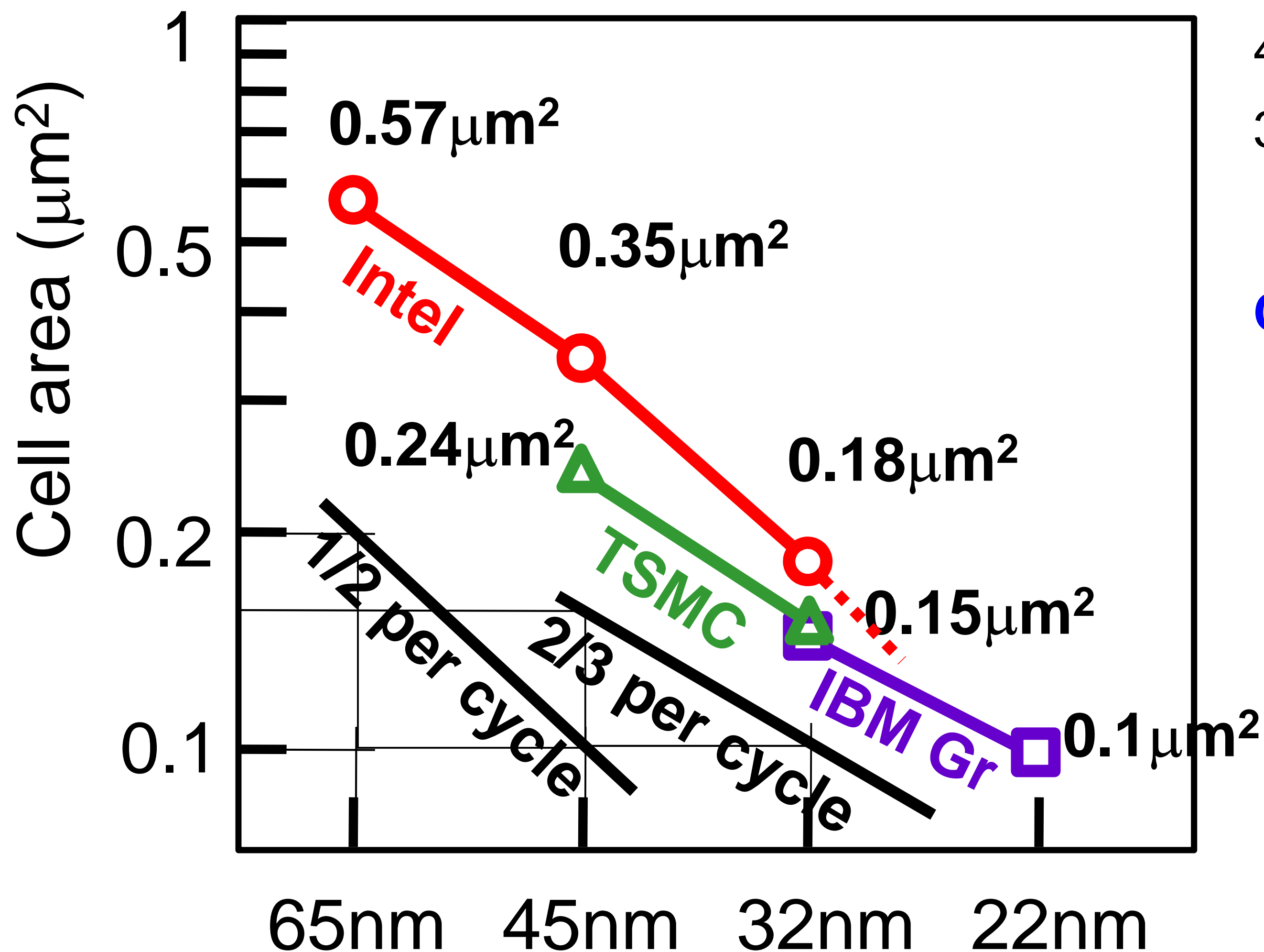




Source: IEDM2008 Pre-conference Publicity http://www.btbmarketing.com/iedm/

Cell size reduction trends

1/2 or 2/3 per cycle?



Intel

Functional Si

65nm Apr.2004

45nm Jan.2006

32nm Sep.2007

TSMC

Conference (IEDM)

45nm Dec.2007

32nm Dec.2007

IBM Gr. (Consortium)

Conference (IEDM)

32nm Dec.2007

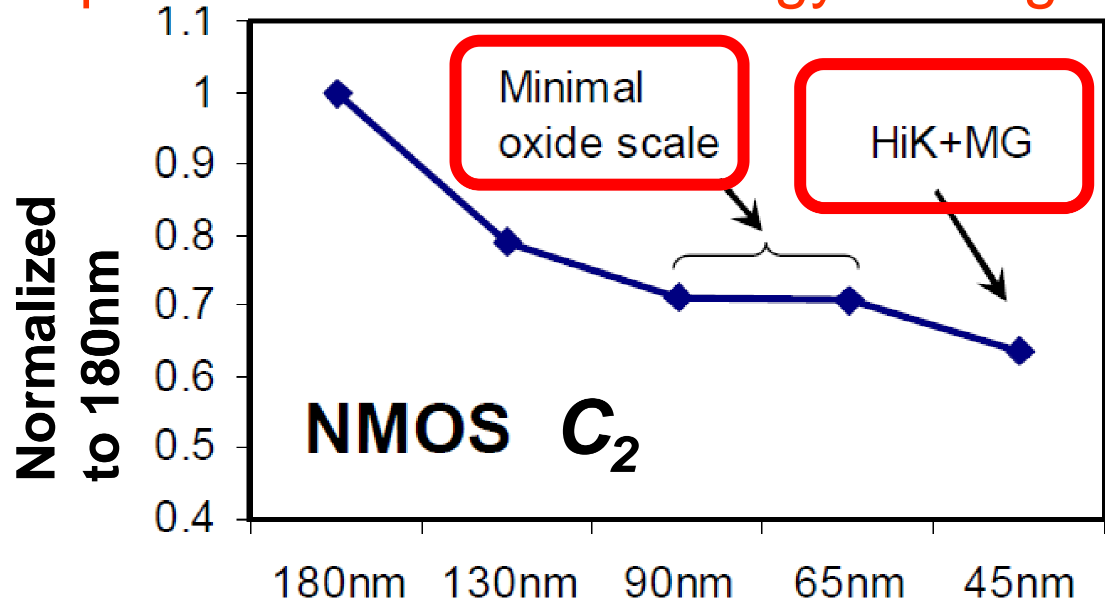
Press release

22nm Aug.2008

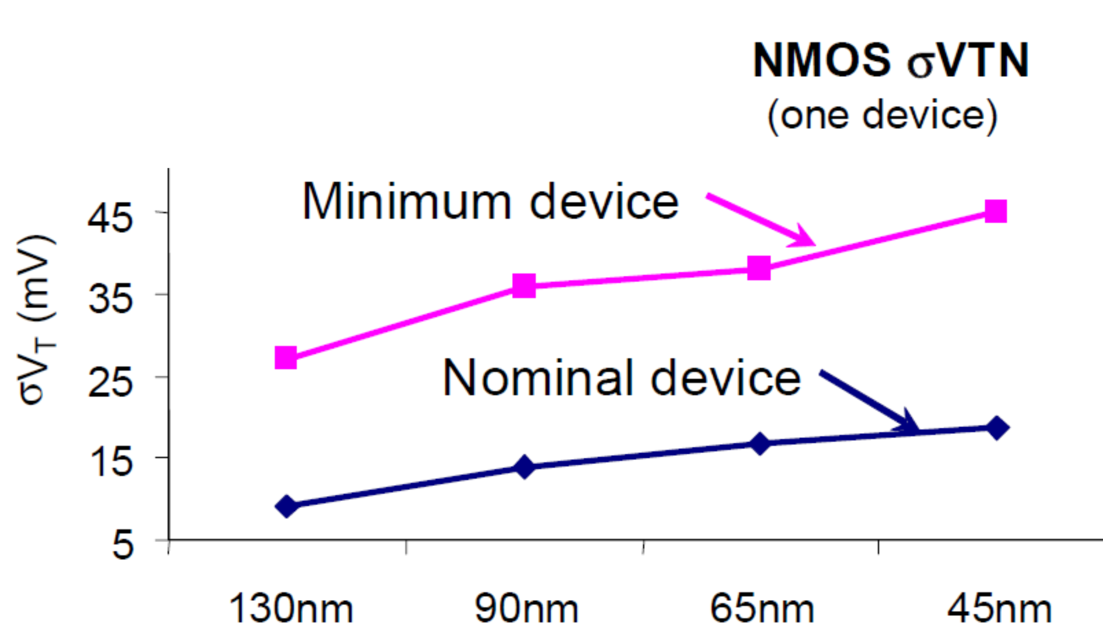
NMOS Mismatch Coefficient (C_2) improvement with technology scaling

$$\sigma V_{Tran} = \left(\frac{\sqrt[4]{4q^3\varepsilon_{si}\phi_B}}{2}\right)\cdot\frac{T_{ox}}{\varepsilon_{ox}}\cdot\left(\frac{\sqrt[4]{N}}{\sqrt{Weff\cdot Leff}}\right)$$

$$= \frac{1}{\sqrt{2}}\left(\frac{c_2}{\sqrt{Weff\cdot Leff}}\right)$$

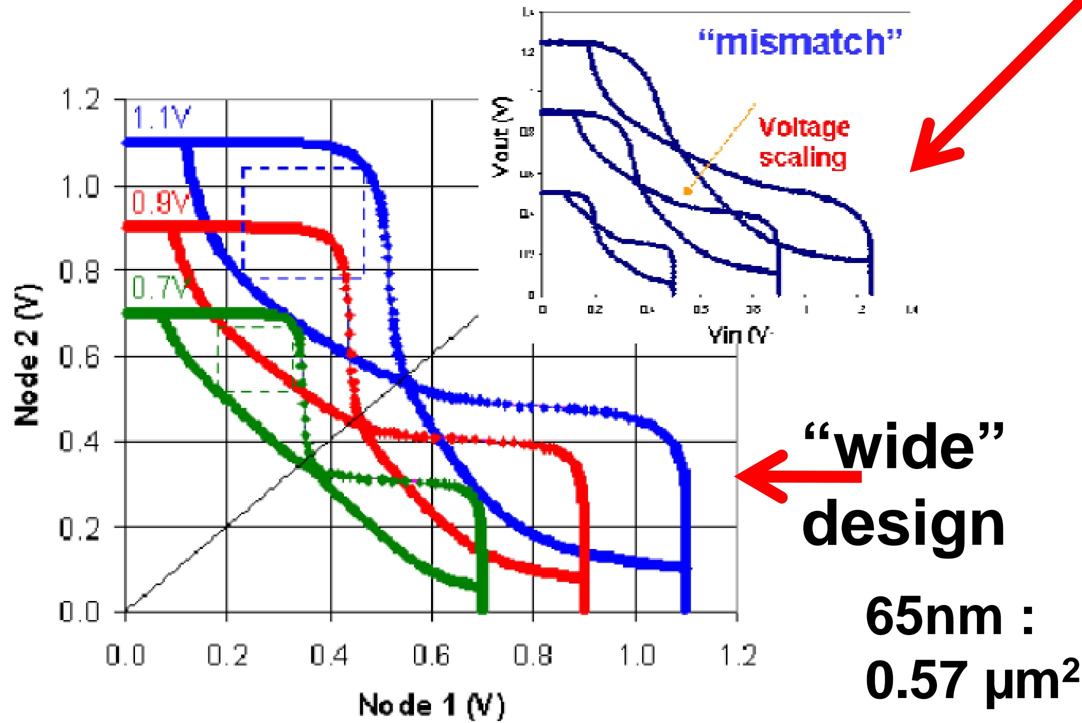


σV_{Tran}



Source: K.J.Kuhn IEDM 2007

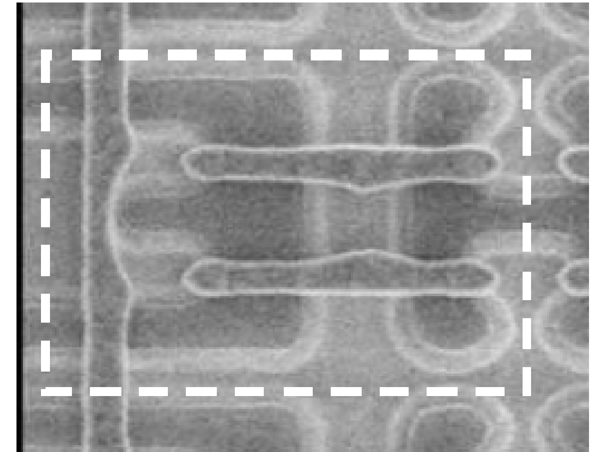
Mismatch improvement by layout (Intel)



Source: K. J. Kuhn
IEDM2007 Tech. Dig. pp.471

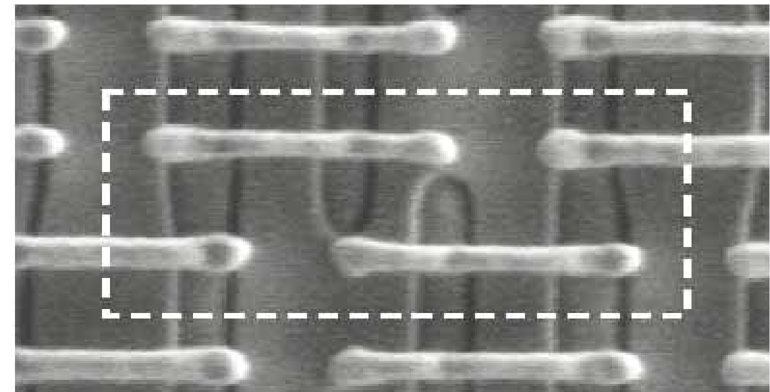
"tall" design

90nm :1.0 μm²



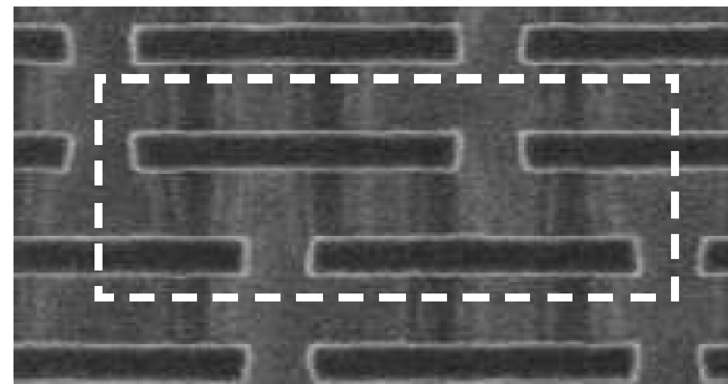
"wide" design

65nm : 0.57 μm²



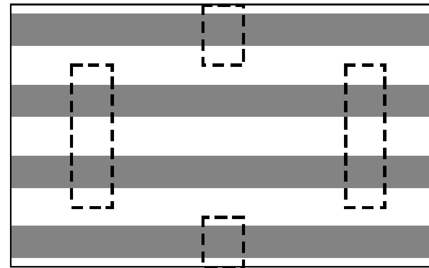
"wide" design
(Square endcaps)

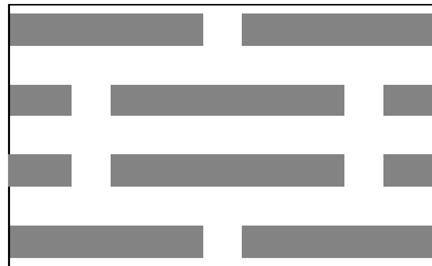
45nm 0.346 μm²

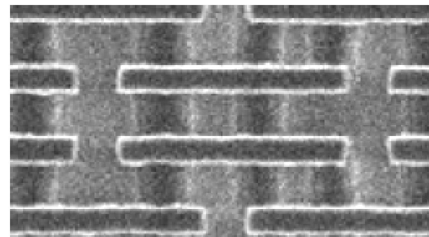


Double patterning for square endcap

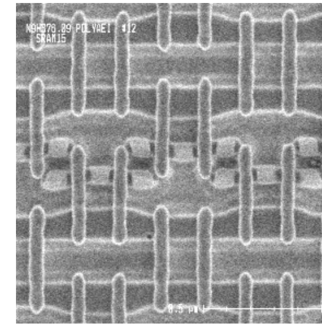
a) Pattern gate lines/spaces

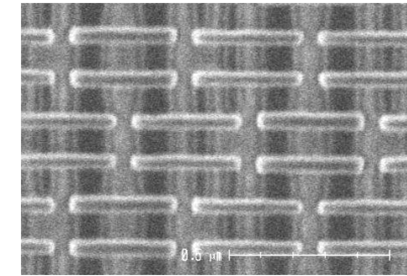

b) Pattern cut mask

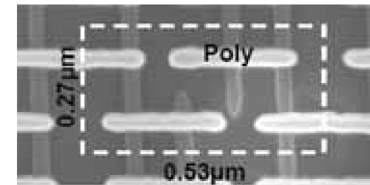

c) Final gate pattern


d) Intel 45nm SRAM cell

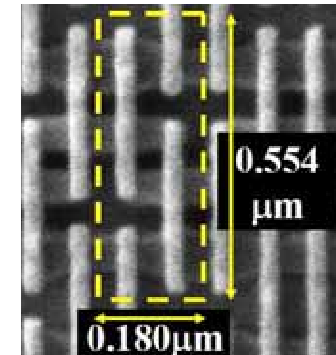
Source: M. Bohr, ICSICT2008


TSMC 45nm
IEDM 2007

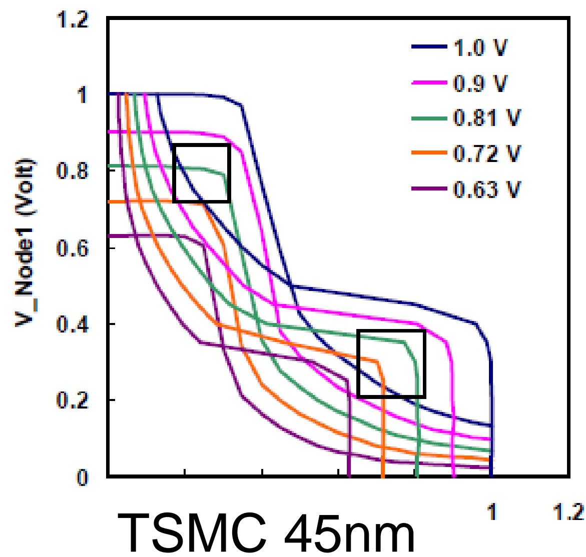

TSMC 32nm
IEDM 2007



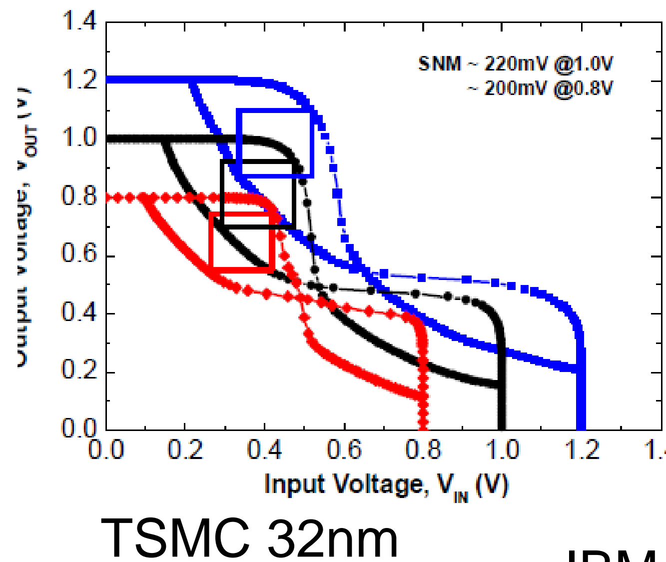
IBM Gr. 32nm
IEDM 2004



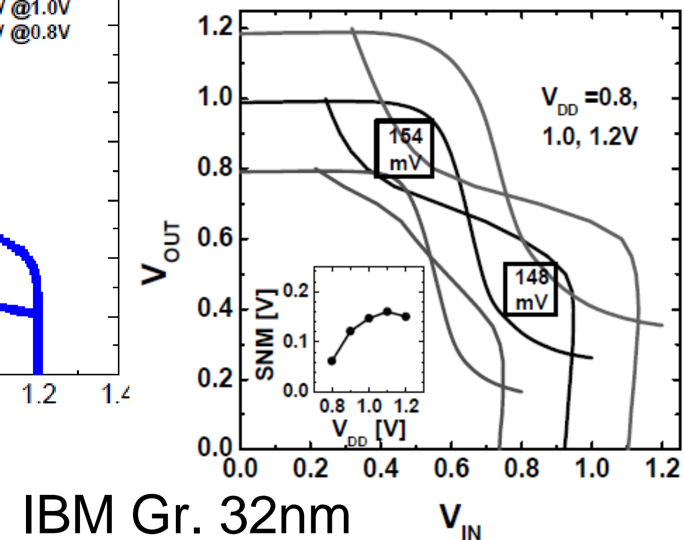
IBM Gr. 22nm IEDM 2008



TSMC 45nm



TSMC 32nm



IBM Gr. 32nm

Most Difficult part of SRAM down-scaling is Vdd down-scaling

Density of on-chip cache SRAM memory is high and thus, Vth cannot be down-scaled too much because of large Isd-leak

Also, under low Vdd, read- and write margin degrades, data retention degrade.

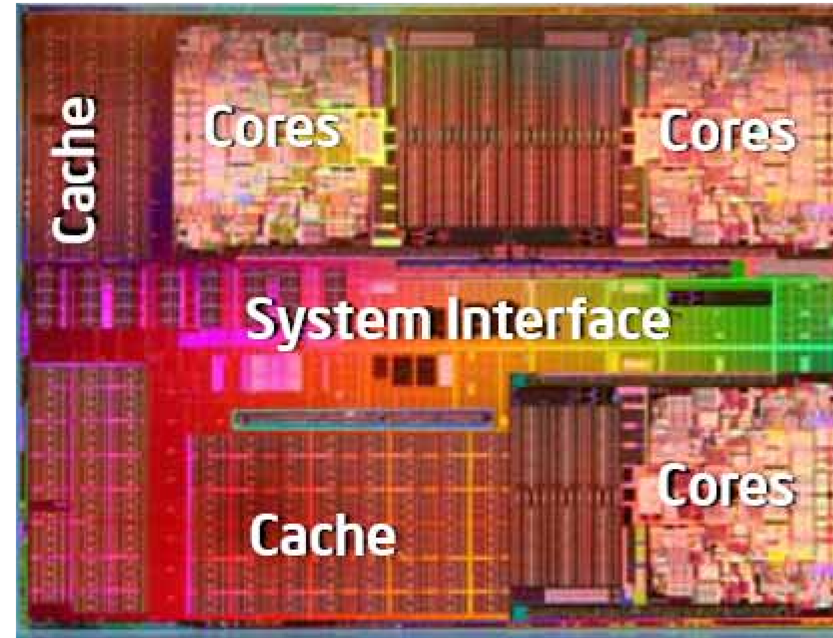
Thus, Vdd down-scaling is more severe in SRAM than logic part of the circuits

Intel® Xeon® 7400 Series (Dunnington)

Source: Intel Developer Forum 2008

45 nm high-k6 cores

16MB shared L3 cache



Cache occupies huge area

→ Cell size of SRAM should be minimized

→ Isd-leak should be minimized

→ Vth are often designed to be higher than logic Vth

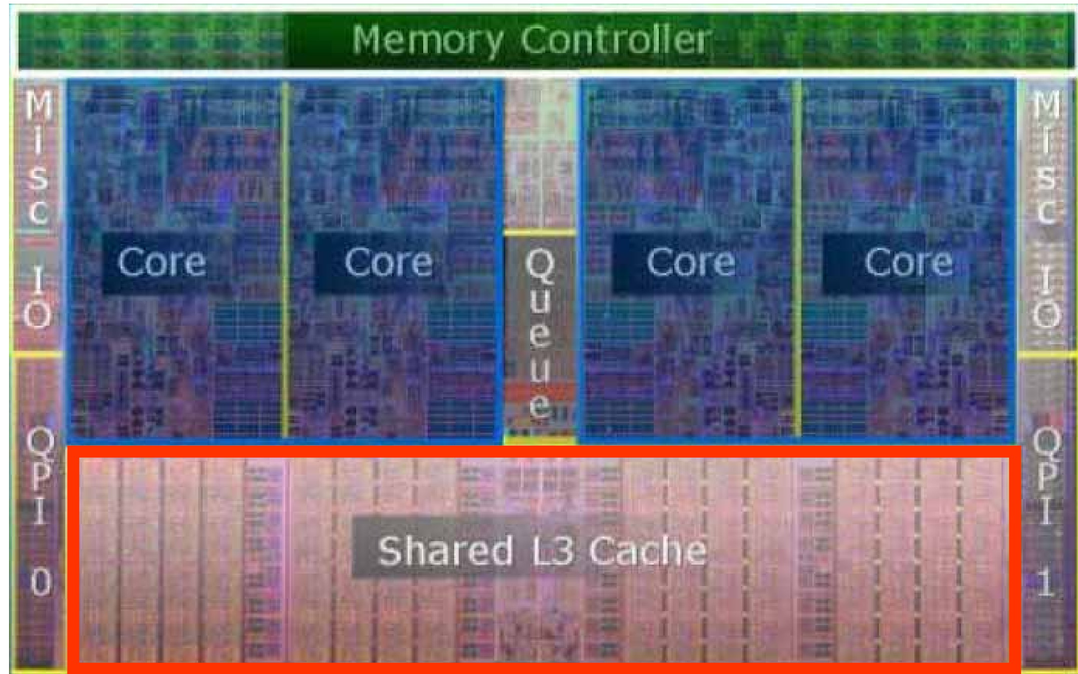
→ Lg are often designed to be larger than logic Lg

Nehalem(Intel) 2,4 or 8 Cores

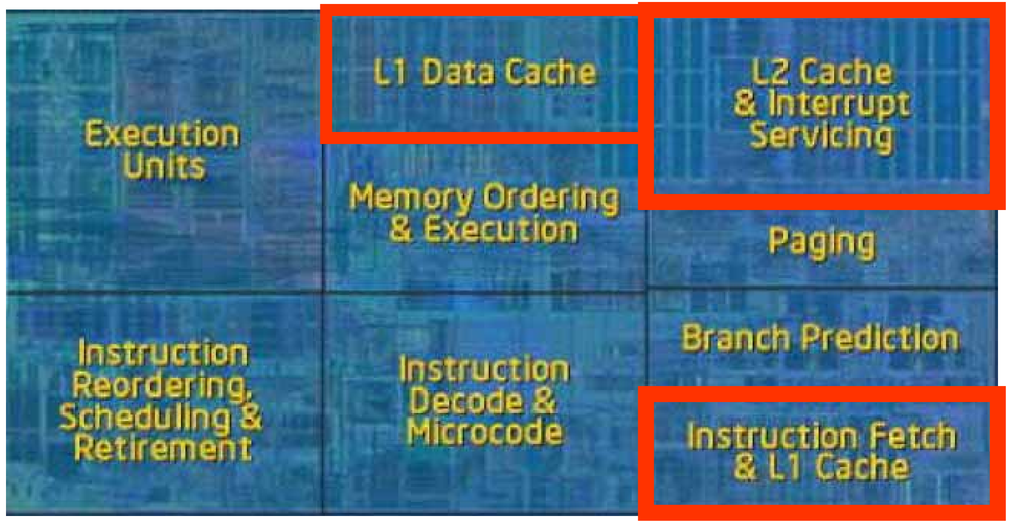
Voltage/Frequency Partitioning

- DDR Vcc
- Core Vcc
- Uncore Vcc

Dynamic Power Management



Chip



Core

8T SRAMCell

32kB L1 I -cache

32kB L1 D-cache

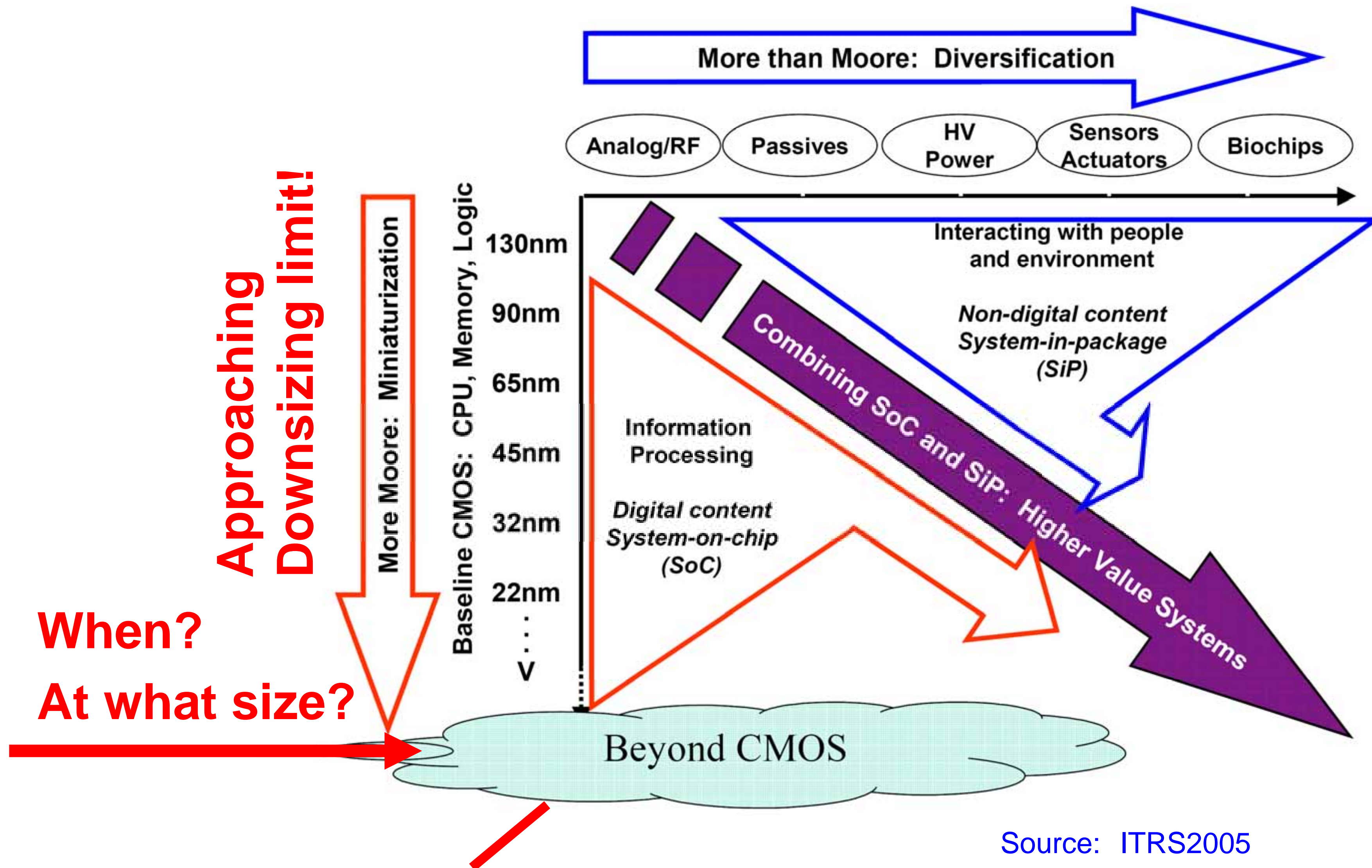
256kB L2 -cache

6T SRAMCell

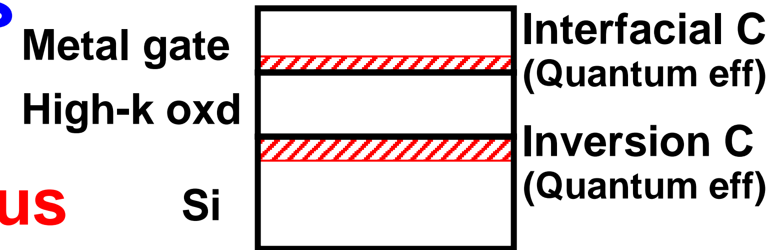
8 MB L3 cache

Source: Intel Developer Forum 2008

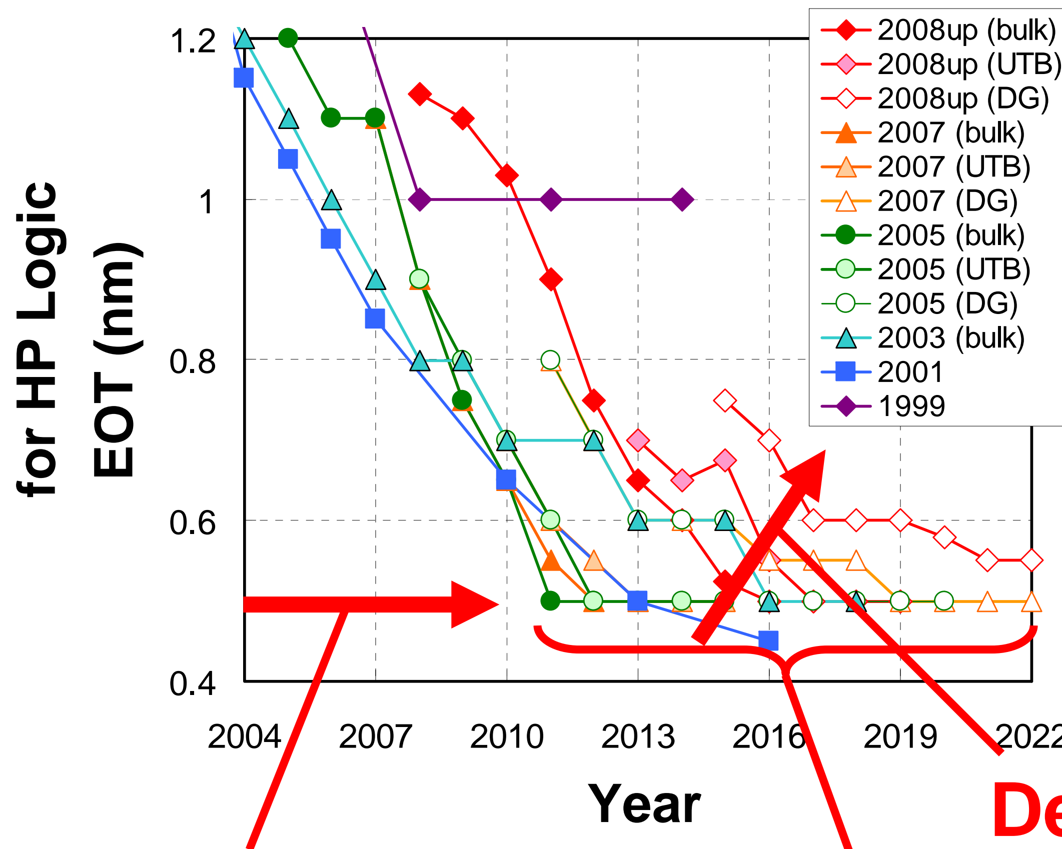
5. Roadmap for further future as a Personal View

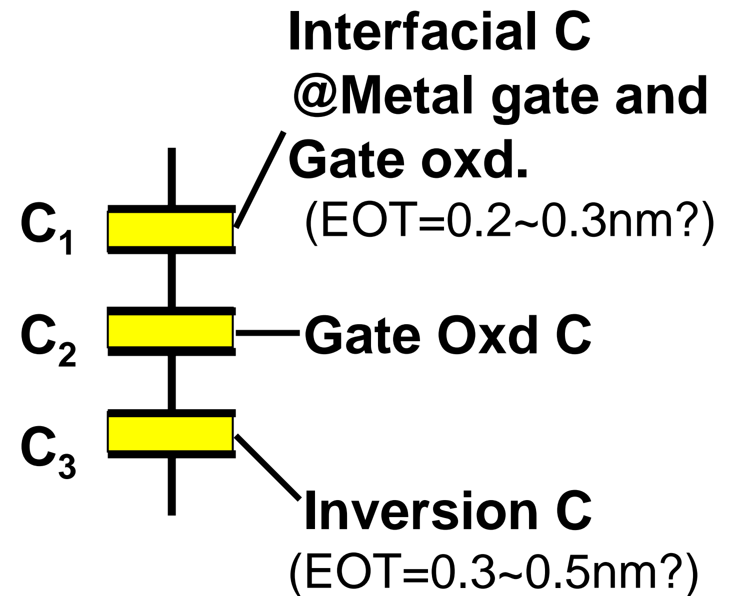

More than Moore: Diversification
Analog/RF
Passives
HV Power
Sensors Actuators
Biochips
Interacting with people and environment
Non-digital content System-in-package (SiP)
Approaching Downsizing limit!
More Moore: Miniaturization
Baseline CMOS: CPU, Memory, Logic
130nm
90nm
65nm
45nm
32nm
22nm
V
Information Processing
Digital content System-on-chip (SoC)
Combining SoC and SiP: Higher Value Systems
When?
At what size?
Beyond CMOS
What is waiting beyond the cloud?
Source: ITRS2005

Could we squeeze technologies for ultimate CMOS scaling?



Saturation of EOT thinning is a serious roadblock to proper down-scaling.





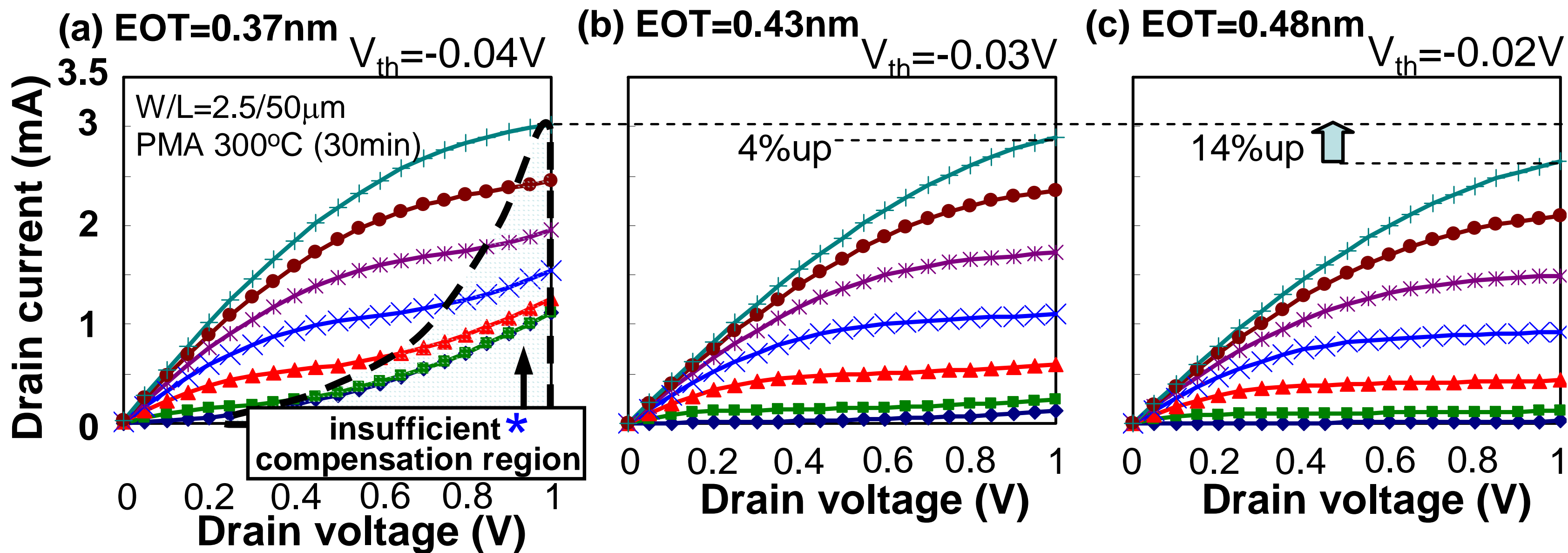
$EOT(C_1) + EOT(C_3) > 0.5nm$

Small effect to decrease $EOT(C_2)$ beyond 0.5nm?

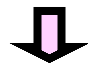
Is 0.5nm real limit?

EOT<0.5nm with Gain in Drive Current is Possible

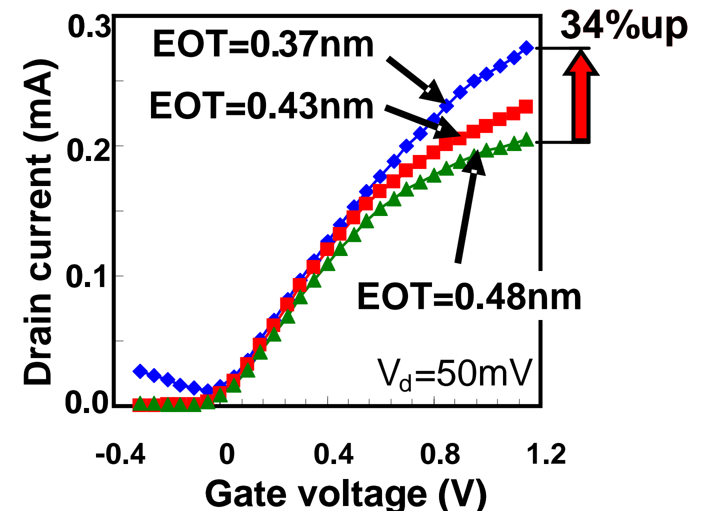
La_2O_3 gate insulator



EOT scaling below 0.5nm



Still useful for larger drain current



Source: K. Kakushima, K. Okamoto, K. Tachi, P. Ahmet, K. Tsutsui, N.i Sugii, T. Hattori, and H. Iwai, IWDTF 2008, Tokyo, November, 2008

* Because Lg is very large (2.5µm), gate leakage is large in case (a). The gate leakage component was subtracted from measured data for case (a). However, if we make small gate length, the gate leakage current should become sufficiently small to be ignored compared with Id as we verified with SiO_2 gate before (Momose et al.,IEDM 1994). The gate leakage could be suppressed by modifying material and process in future.

Thus, in future, maybe continuous development of new techniques could make more proper down-scaling possible.

It is difficult to say, but EOT and Vdd may become smaller than expected today.

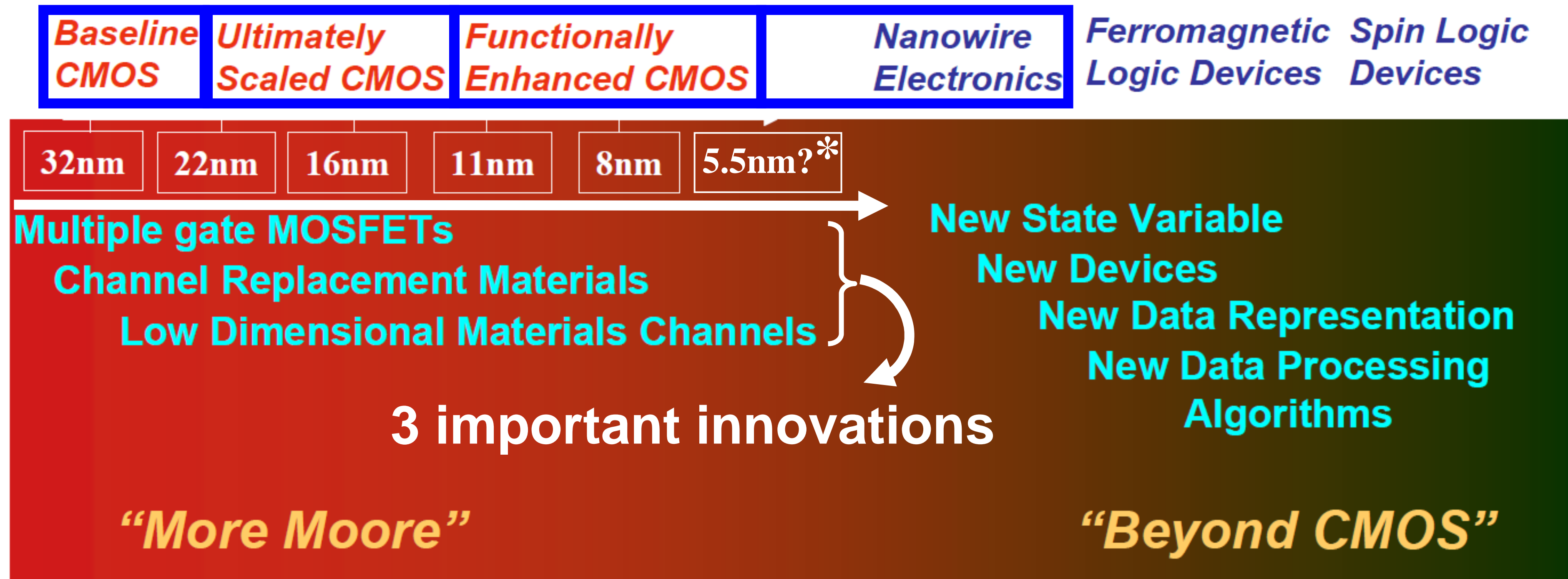
Personal view:

- There will be still 4~6 cycles (or technology generations) left until we reach 11 ~ 5.5 nm technologies, at which we will reach down-scaling limit, in some year between 2020-30.

- Effort and passion to pursue the down-scaling until then are important.

Probably, still CMOS and Si. Maybe, Ge,III-V, C

ITRS figure edited by Iwai



Source: 2008 ITRS Summer Public Conf.

*5.5nm? was added by Iwai

After reaching the down-scaling limit, we could still continue R & D, seeking sufficiently higher Id-sat under low Vdd.

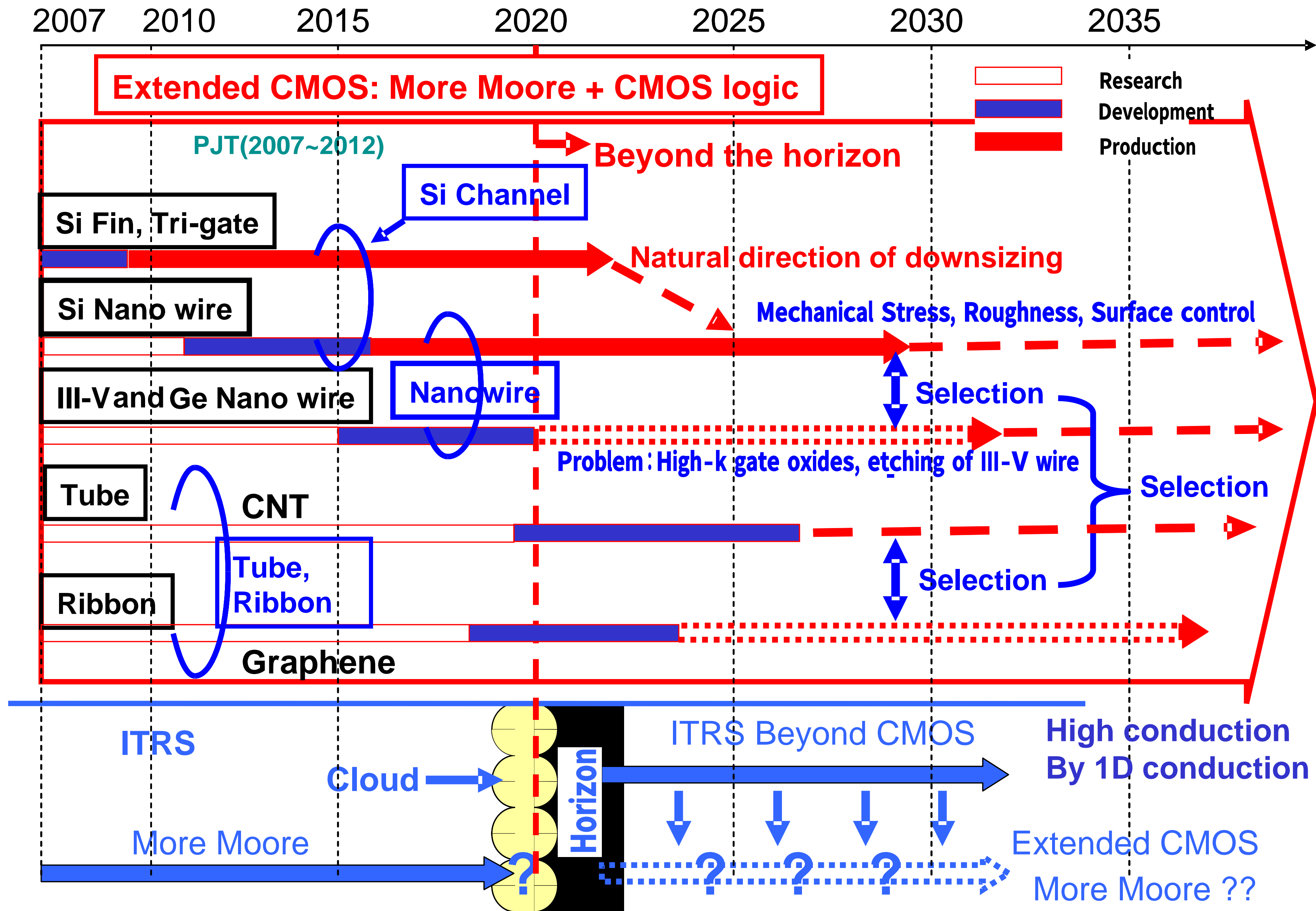
Two candidates have emerged for R & D for the above.

1. Nanowire/tube MOSFETs
2. Alternative channel MOSFETs (III-V, Ge)

Other Beyond CMOS devices are still in the cloud.

Our new roadmap

Source: H. Iwai, IWJT 2008



Thank you
for your attention!