# Si MOSFET Roadmap for 22nm and beyond

December 16, 2009

Jadavpur University @ Kolkata, India

Hiroshi Iwai

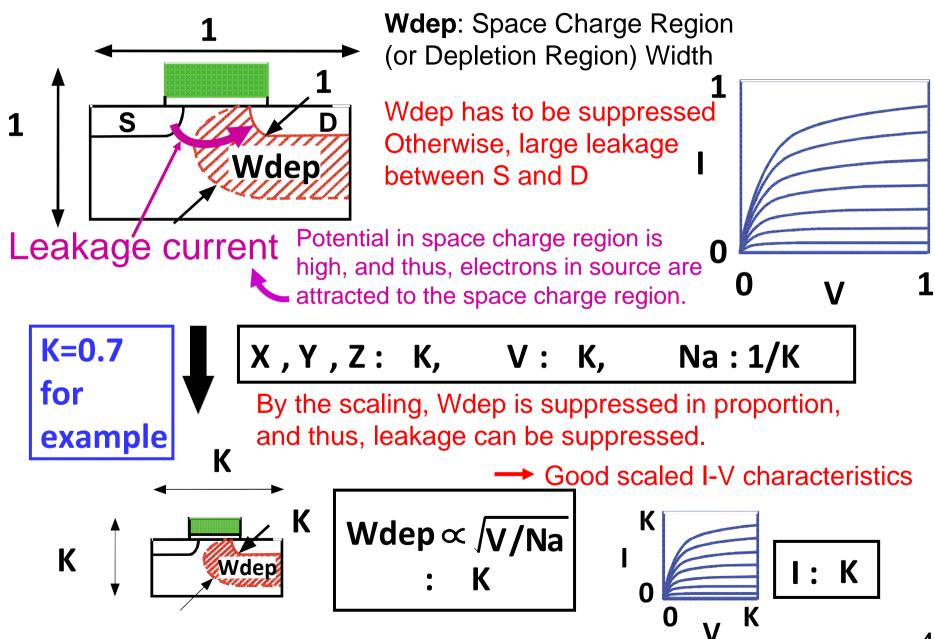
Tokyo Institute of Technology

## **Outline**

- 1. Scaling
- 2. ITRS Roadmap
- 3. Voltage Scaling/ Low Power and Leakage
- 4. SRAM Cell Scaling
- 5.Roadmap for further future

## 1. Scaling

## Scaling Method: by R. Dennard in 1974



## Downscaling merit: Beautiful!

Geometry & Supply voltage	$L_{g}$ , $W_{g}$ $T_{ox}$ , $V_{dd}$	К	Scaling K: K=0.7 for example
Drive current in saturation	l <sub>d</sub>	K	$I_{d} = V_{sat}W_{g}C_{o}(V_{g}-V_{th})$ $C_{o}: gate C per unit area$ $W_{g}(t_{ox}^{-1})(V_{g}-V_{th}) = W_{g}t_{ox}^{-1}(V_{g}-V_{th}) = KK^{-1}K=K$
I <sub>d</sub> per unit W <sub>g</sub>	I <sub>d</sub> /μm	1	$I_d$ per unit $W_g = I_d / W_g = 1$
Gate capacitance	C <sub>g</sub>	K	$C_g = \varepsilon_o \varepsilon_{ox} L_g W_g / t_{ox}$ $\longrightarrow$ KK/K = K
Switching speed	τ	К	$\tau = C_g V_{dd} / I_d \longrightarrow KK/K = K$
Clock frequency	f	1/K	$f = 1/\tau = 1/K$
Chip area	A <sub>chip</sub>	α	$\alpha$ : Scaling factor $\longrightarrow$ In the past, $\alpha$ >1 for most cases
Integration (# of Tr)	N	$\alpha/K^2$	N $\rightarrow \alpha/K^2 = 1/K^2$ , when $\alpha=1$
Power per chip	Р	α	fNCV <sup>2</sup> /2 $\longrightarrow$ K <sup>-1</sup> ( $\alpha$ K <sup>-2</sup> )K (K <sup>1</sup> ) <sup>2</sup> = $\alpha$ = 1, when $\alpha$ =1

2 Generations	$k$ = 0.7 $^2$ =0.5 and $\alpha$ =1
Single MOFET	
	Vdd → 0.5
	Lg → 0.5
	ld → 0.5
	Cg → 0.5
	P (Power)/Clock
	$\rightarrow 0.5^3 = 0.125$
	$\tau$ (Switching time) $\rightarrow$ 0.5
Chip	
	N (# of Tr) $\rightarrow$ 1/0.5 <sup>2</sup> = 4
	f (Clock) $\rightarrow$ 1/0.5 = 2
	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.

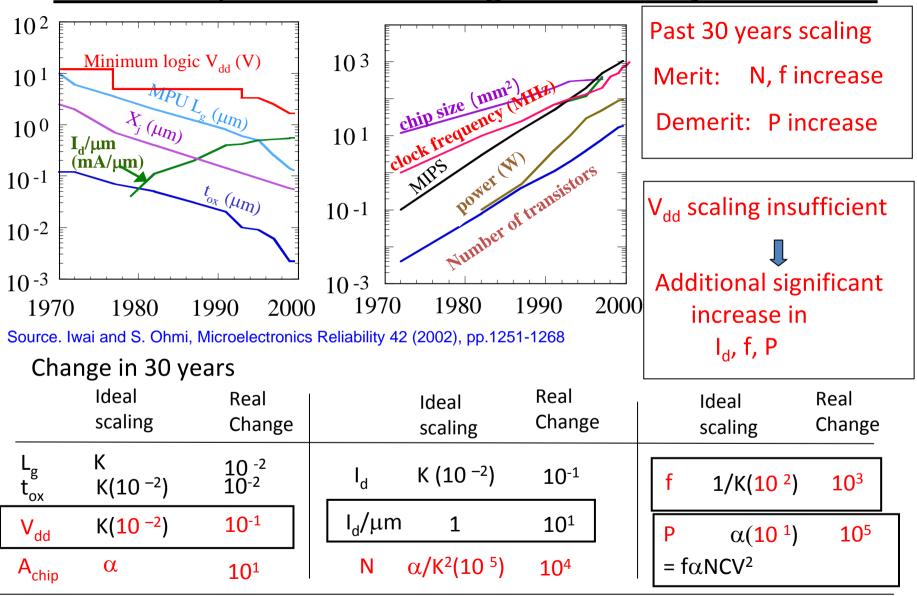
"Euclid of Alexandria (325BC?-265BC?)

'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



Vd scaling insufficient,  $\alpha$  increased  $\rightarrow$  N, Id, f, P increased significantly

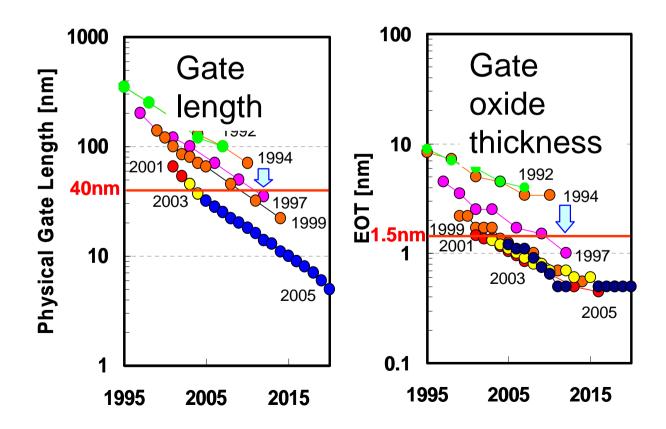
- Now, power and/or heat generation are the limiting factors of the down-scaling
- Supply voltage reduction is becoming difficult, because Vth cannot be decreased any more, as described later.
- Growth rate in clock frequency and chip area becomes smaller.

## 2. ITRS Roadmap (for 22 nm CMOS logic)

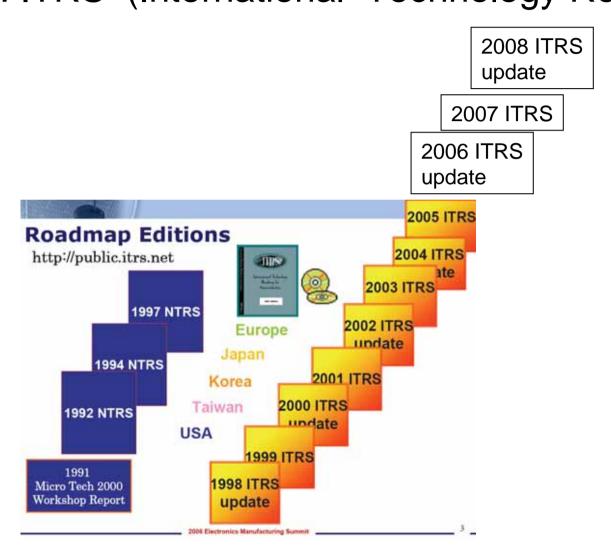
What is a roadmap? What is ITRS?

Roadmap: Prediction of future technologies

ITRS: International Technology Roadmap for Semiconductors made by SIA (Semiconductor Industry Association with Collaboration with Japan, Europe, Korea and Taiwan)



## 1992 -1997:NTRS (National Technology Roadmap)1998 - : ITRS (International Technology Roadmap)

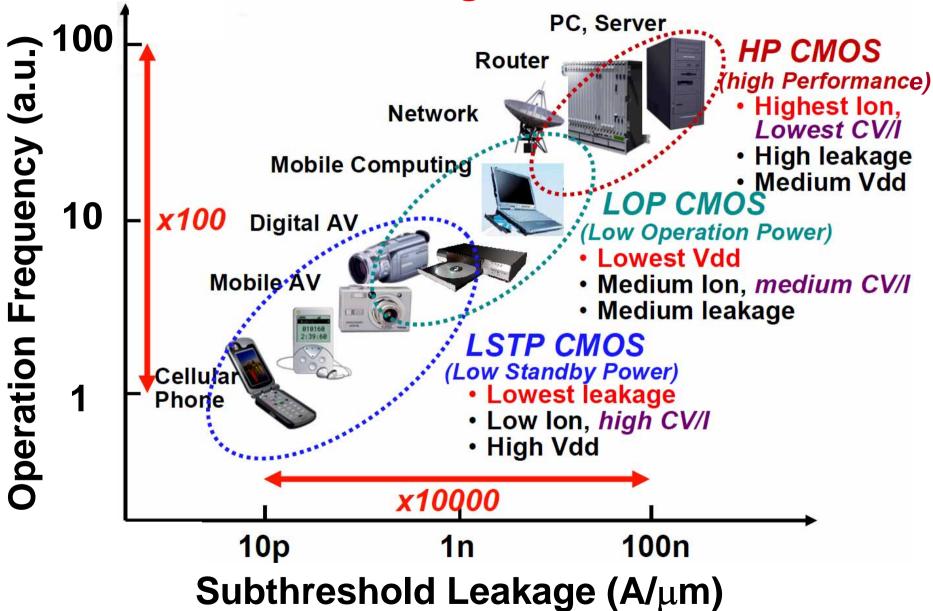


## ITRS Roadmap does change every year!

```
2007 Edition2003 Edition2006 Update2002 Update2005 Edition2001 Edition2004 Update2000 Update
```

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

## HP, LOP, LSTP for Logic CMOS



Source: 2007 ITRS Winter Public Conf.

## What does '45 nm' mean in 45 nm CMOS Logic?

#### **'XX nm CMOS Technology**

Commercial Logic CMOS products

#### ITRS 2008 Update

for High Performance Logic

Technology name	Starting Year		Year	Half Pitch (1 <sup>st</sup> Metal)	Physical Gate Length
45 nm	2007	<b>←</b>	2007 2008	68 nm 59 nm	32 nm 29 nm
32 nm	2009?	<b>─ ←</b> →	2009 2010	52 nm 45 nm	27 nm 24 nm

### 'XX nm' CMOS Logic Technology:

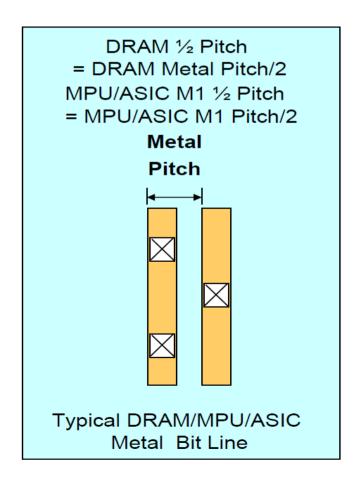
- In general, there is no common corresponding parameter with 'XX nm' in ITRS table, which stands for 'XX nm' CMOS.

## What does '45 nm' mean in 45 nm CMOS Logic?

#### $8\mu m \rightarrow 6\mu m \rightarrow 4\mu m \rightarrow 3\mu m \rightarrow 2\mu m \rightarrow 1.2\mu m \rightarrow 0.8\mu m \rightarrow 0.5\mu m$

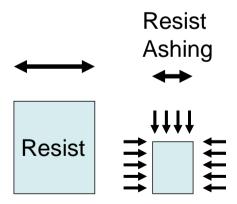
- Originally, 'XX' means lithography resolution.
- Thus, 'XX' was the gate length, and half pitch of lines
- 'XX' had shrunk 0.7 in 3 years in average (0.5 in 6 years) those days.

Logic 1<sup>st</sup> Metal Half Pitch



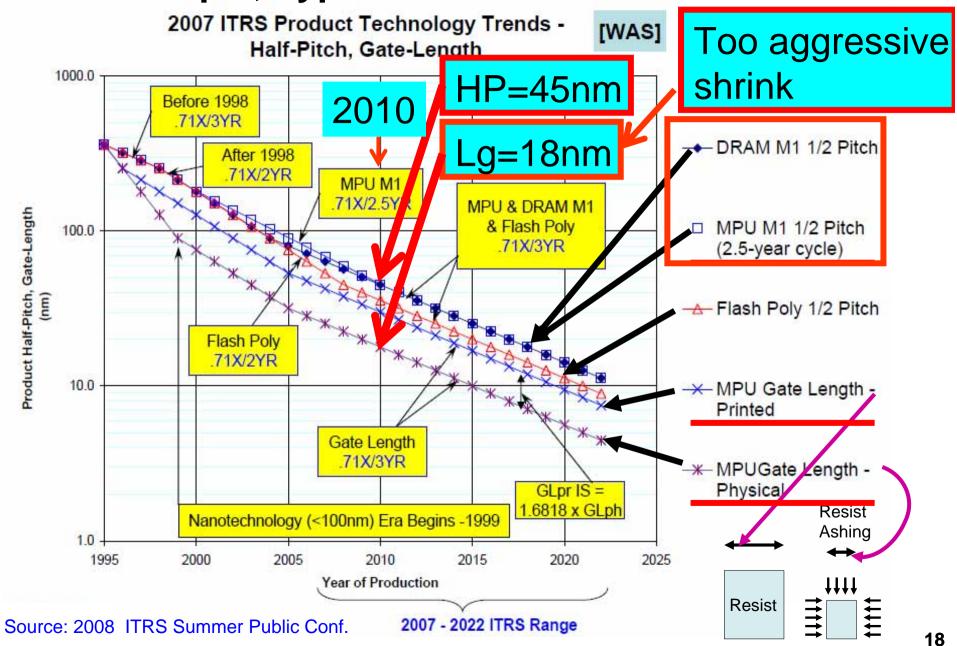
## What does '45 nm' mean in 45 nm CMOS Logic?

- $\rightarrow$  350nm  $\rightarrow$  250nm  $\rightarrow$  180nm  $\rightarrow$  130nm  $\rightarrow$  90nm  $\rightarrow$  65nm  $\rightarrow$  45nm
- -'XX' values were established by NTRS\* and ITRS with the term of 'Technology Node\*\*' and 'Cycle\*\*\*' using typical 'half pitch value'.
- 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'

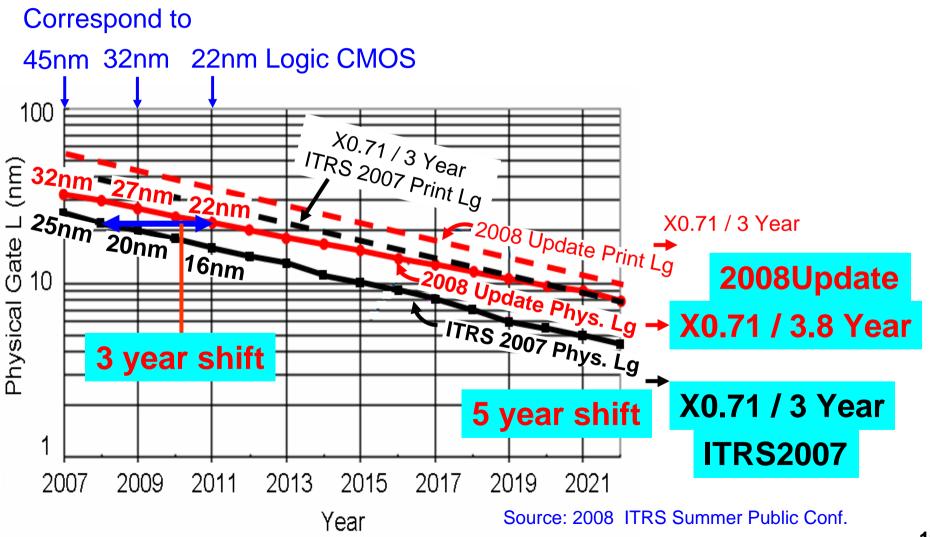
## For example, Typical Half Pitches at ITRS 2007



#### 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.



## EOT and Xj shift backward, corresponding to Lg shift

EOT:  $0.55 \text{ nm} \rightarrow 0.88 \text{ nm}$ , Xj:  $8 \text{ nm} \rightarrow 11 \text{ nm}$  @ 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 (nn )	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/ <u>3E20 poly</u> , bulk MPU (nm)	1.3	1.2	1.2	1		0.54	0.41				Lik	ely	in 20	800	Upd	ate
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		1.0	0.95	0.88	0.75	0.65	0.60	0.53	0.5	Lik	ely	in 20	800	Upd	ate
Drain Ext. X <sub>j</sub> bulk MPU (nm)	12.5	11	10	9	8	7										
Drain Ext. X <sub>j</sub> bulk MPU (nm)	11	11	11	11	11		9	8.5	7.7	7	Lik	ely	in 20	800	Upd	ate

non-steady trend corrected

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

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

#### **'XX nm CMOS Technology**

Commercial Logic CMOS products

#### ITRS (Likely in 2008 Update)

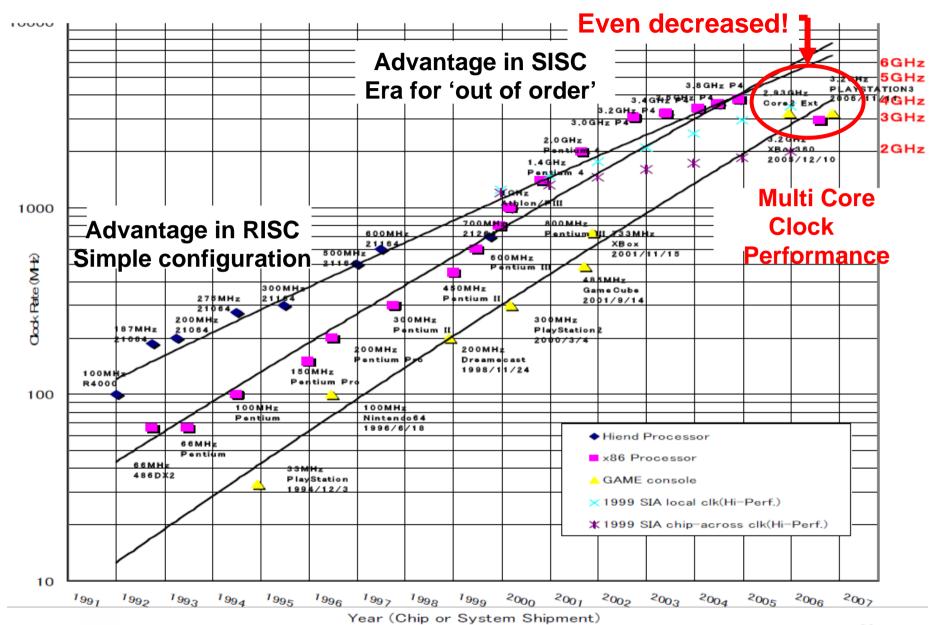
for High Performance Logic

Technology name	Starting Year		Year	Half Pitch (1 <sup>st</sup> Metal)	Physical Gate Length
45 nm	2007	<b>─</b>	2007	68 nm	32 nm
	_00.		2008	59 nm	29 nm
32 nm	2009?	<u> </u>	2009	52 nm	27 nm
			2010	45 nm	24 nm
22 nm	2011?~	<b></b>	2011	40 nm	22 nm
	2012?		2012	36 nm	20 nm
16 nm	2013?~		2013	32 nm	18 nm
	2014?	<b>←</b>	2014	29 nm	16 nm

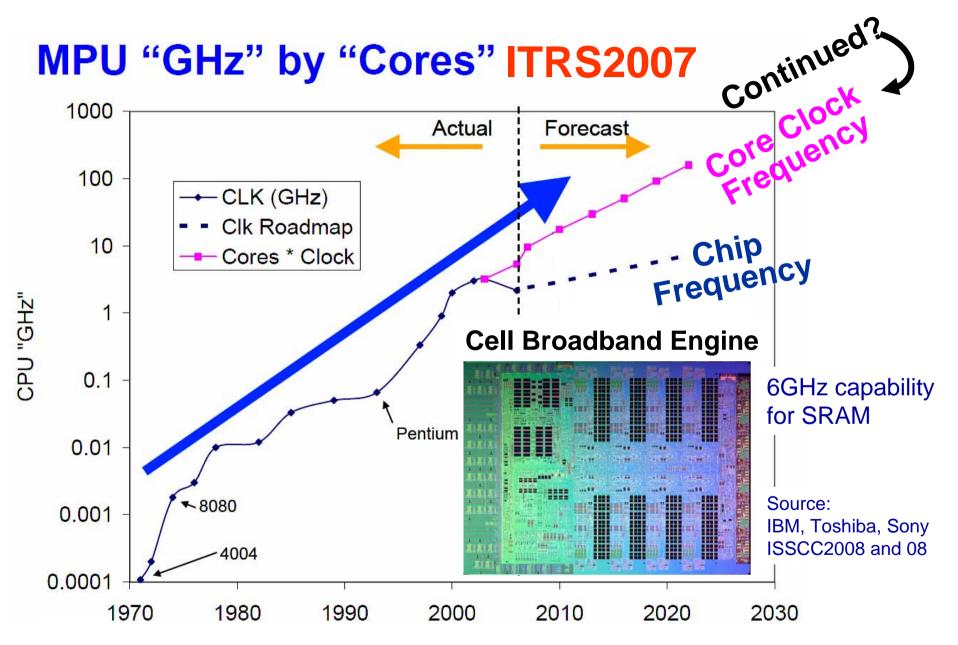
Source: 2008 ITRS Summer Public Conf.

From ITRS2008 Update, maybe XX nm stands for the physical Gate length

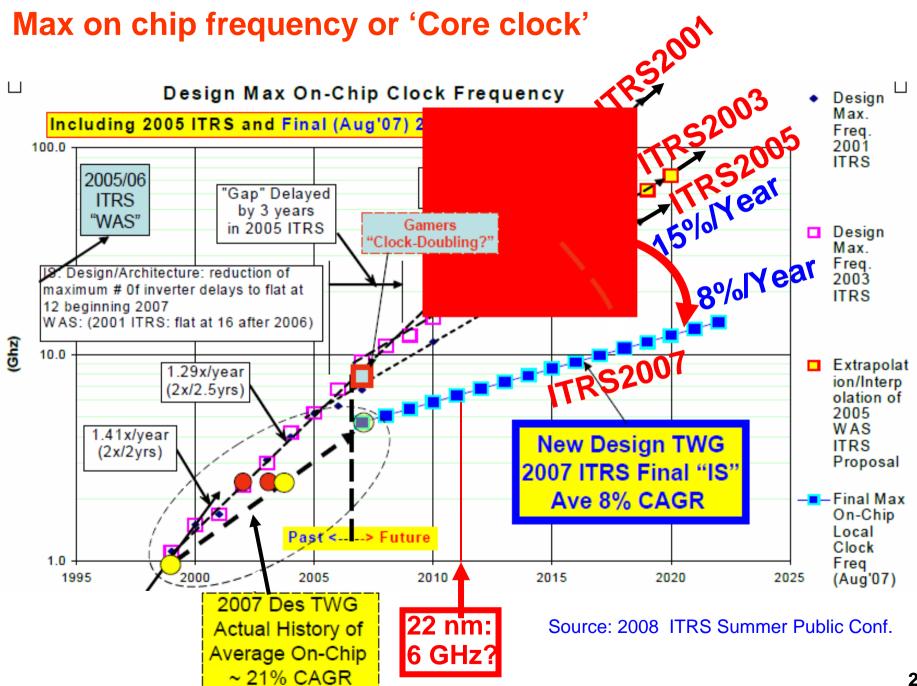
#### Clock frequency does not increase aggressively anymore.



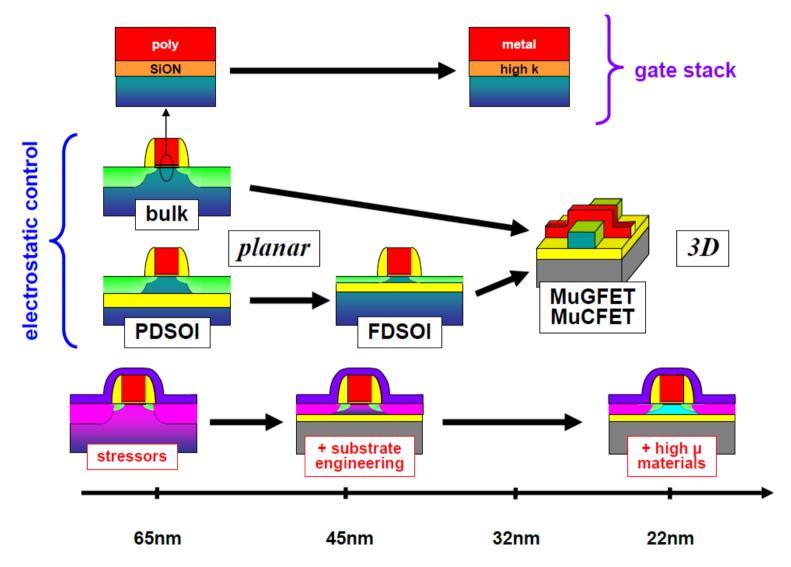
Source: Mitsuo Saito, Toshiba



Source: 2007 ITRS Winter Public Conf.



#### Structure and technology innovation (ITRS 2007)

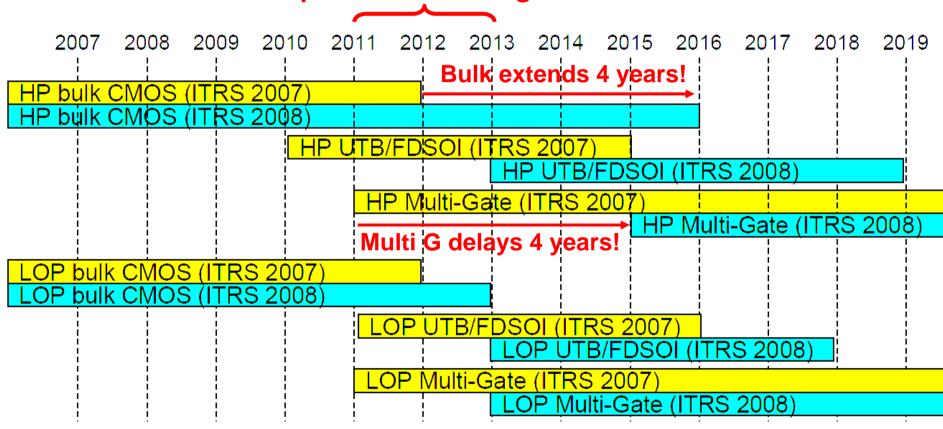


Source: 2008 ITRS Summer Public Conf.

#### Timing of CMOS innovations shifts backward.

#### **Bulk CMOS has longer life now!**

#### **Correspond to 22nm Logic CMOS**



Source: 2008 ITRS Summer Public Conf.

## Wafer size (ITRS 2007)

#### **Correspond to 22nm**

Year of Production	2007	2008	2009	2010	2011	2012	2013	2014	2015
MPU/ASIC Metal 1 (M1) ½ Pitch (nm)(contacted)	68	59	52	45	40	36	32	28	25
MPU Physical Gate Length (nm)	25	23	20	18	16	14	13	11	10
MPU High-Performance Total Chip Area(mm²)	310	246	195	310	246	195	310	246	195
MPU High-Performance Active Transistor Area(mm²)	31.7	25.1	20.0	31.7	25.1	20.0	31.7	25.1	20.0
General Characteristics * (99%	Chip Yield)	•			•				
Maximum Substrate Diameter (mm)—High-volume Production (>20K wafer starts per month)**	300	300	300	300	300	450	450	450	450

Source: ITRS 2007

Maybe delay??

## ITRS2008 Low-k Roadmap Update

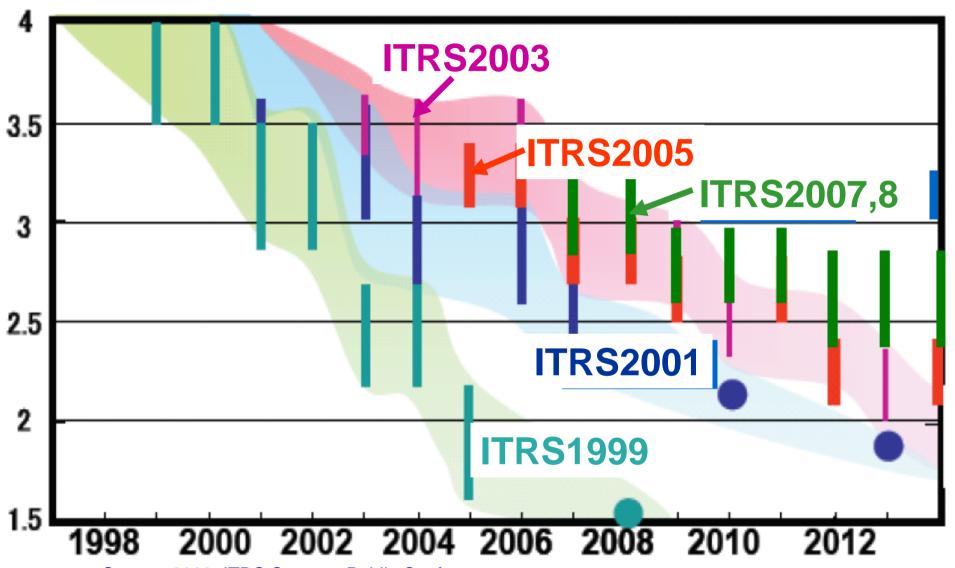
#### **Correspond to 22nm Logic**

		Near-term		•			
ITRS	Year of Production	2008	2009	2010	2011	2012	2013
2007	Interlevel metal insulator – effective dielectric constant (κ)	2.7-3.0	2.5-2.8	2.5-2.8	2.5-2.8	2.1-2.4	2.1-2.4
Update 2008	Interlevel metal insulator – effective dielectric constant (κ)	2.9-3.3	2.6-2.9	2.6-2.9	2.6-2.9	2.4-2.8	2.4-2.8
ITRS 2007	Interlevel metal insulator – bulk dielectric constant (κ)	2.3-2.7	2.1-2.4	2.1-2.4	2.1-2.4	1.8-2.1	1.8-2.1
Update 2007	Interlevel metal insulator – bulk dielectric constant (κ)	2.5- <u>2.8</u>	2.3- <u>2.</u> 6	2.3 <mark>-2.</mark> 6	2.3- <u>2.6</u>	2.1-2.4	2.1- <u>2.4</u>

Source: 2008 ITRS Summer Public Conf.

k value increases by 0.1 ~ 0.3

## **Historical Transition of ITRS Low-k Roadmap**



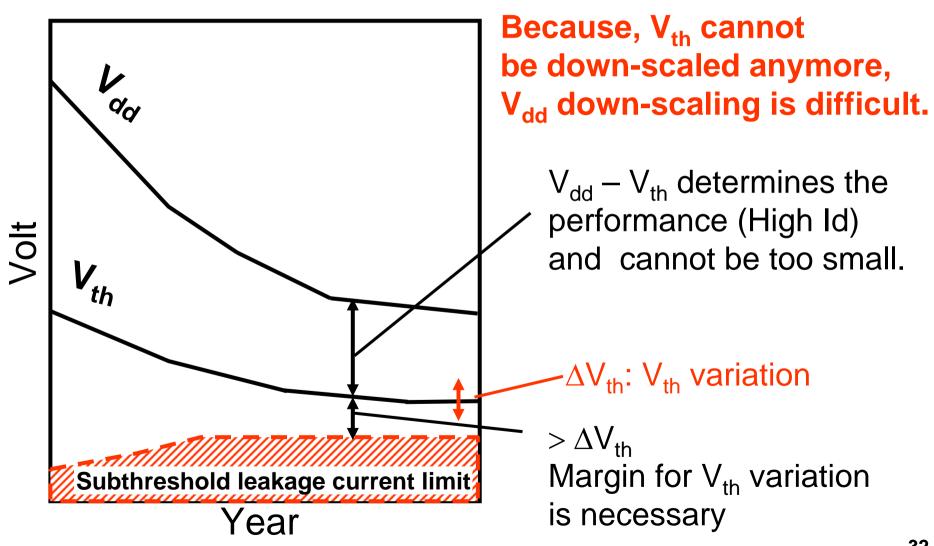
Source: 2008 ITRS Summer Public Conf.

## Roadmap towards 22nm technology and beyond

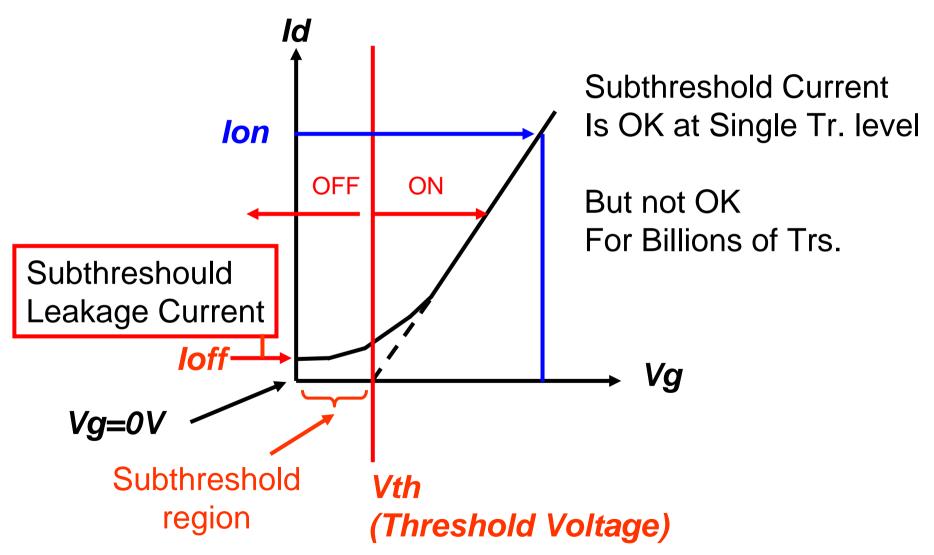
- Physical gate length downsizing rate will be less aggressive.
- Corresponding to the above, performance increase would slow down Clock frequency, etc.
- Introduction of innovative structures UTB SOI and DG delayed, and bulk CMOS has longer life than predicted by previous ITRS roadmaps.

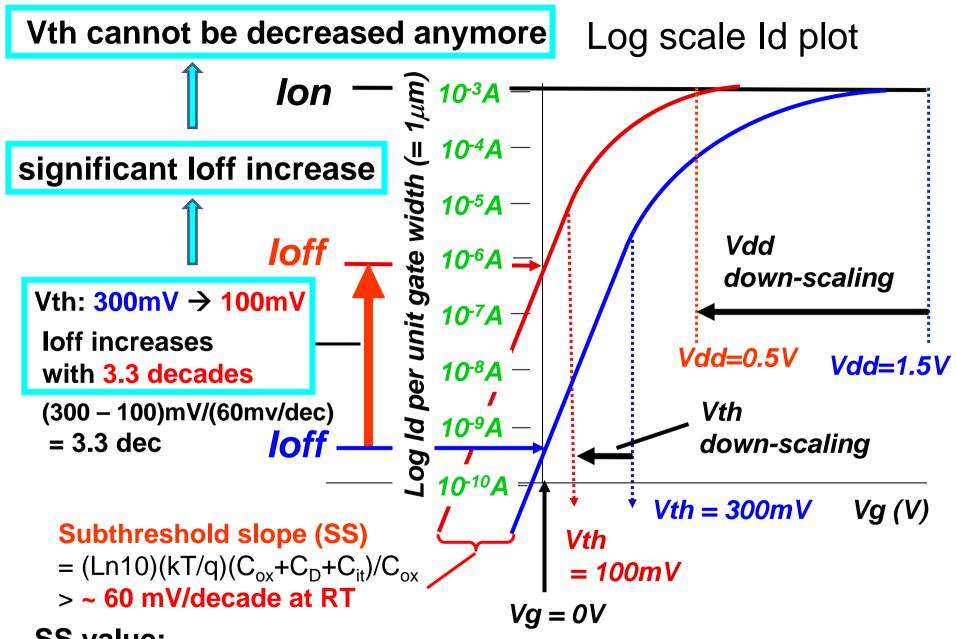
# 3. Voltage Scaling/ Low Power and Leakage

## Difficulty in Down-scaling of Supply Voltage: Vdd



## Subtheshold leakage current of MOSFET



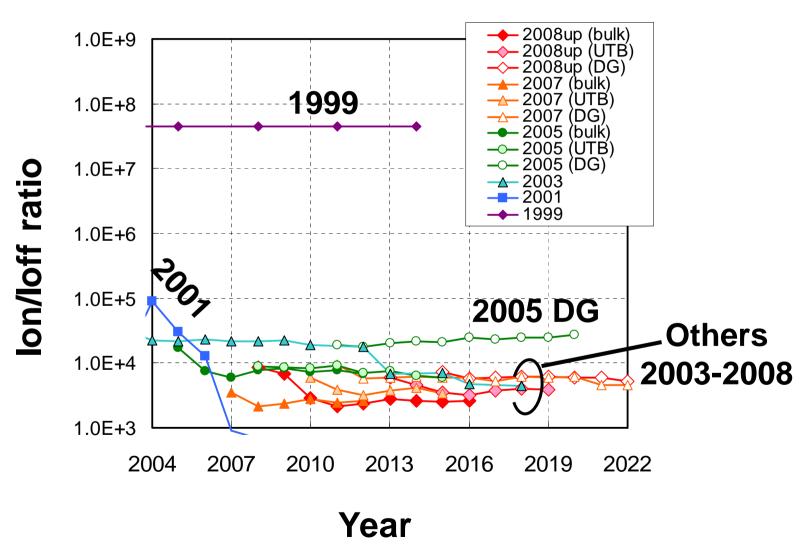


#### SS value:

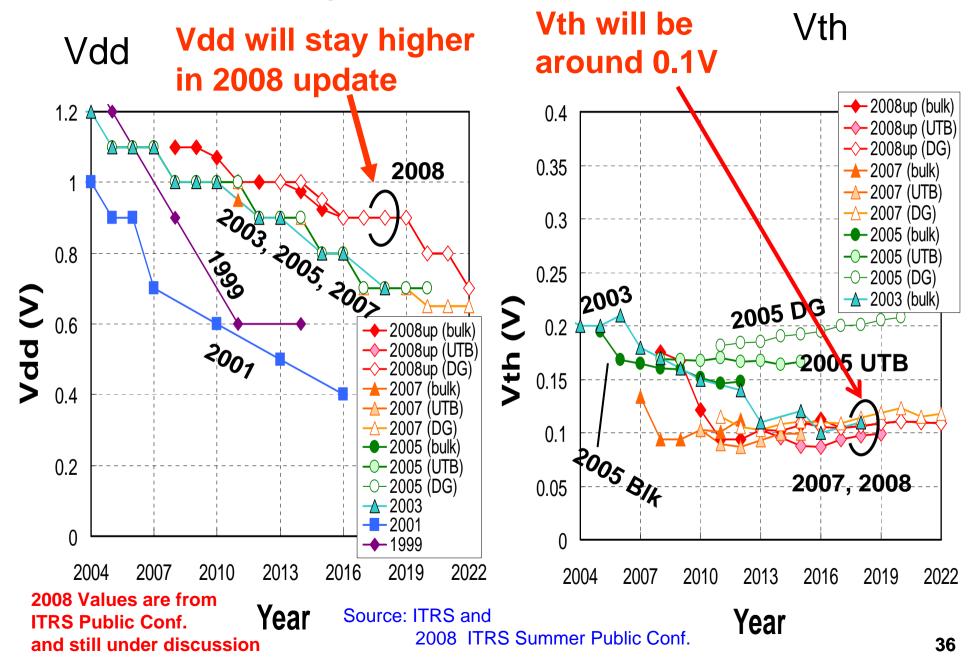
Constant and does not become small with down-scaling

## **ITRS for HP logic**

#### Ion/Ioff ratio



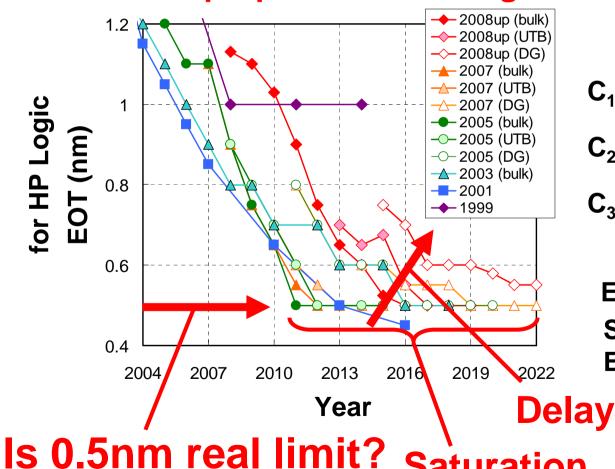
## ITRS for HP logic



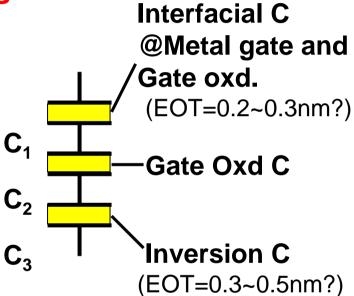
#### Improper down-scaling

Could we squeeze technologies for ultimate CMOS scaling?

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



Interfacial C **Metal gate** (Quantum eff) High-k oxd Inversion C (Quantum eff) Si

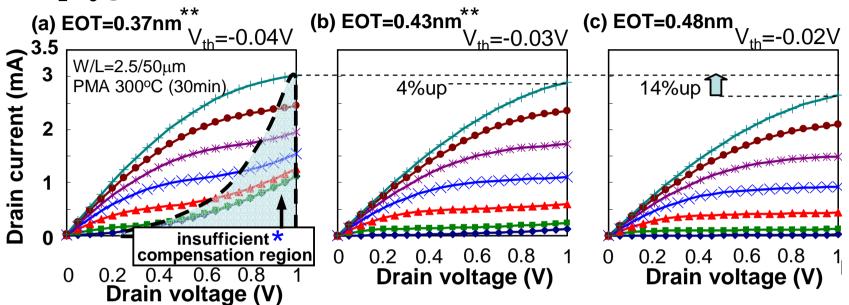


 $EOT(C_1) + EOT(C_3) > 0.5nm$ Small effect to decrease EOT(C<sub>2</sub>) beyond 0.5nm?

**Saturation** 

#### **EOT<0.5nm** with Gain in Drive Current is Possible

#### La<sub>2</sub>O<sub>3</sub> 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

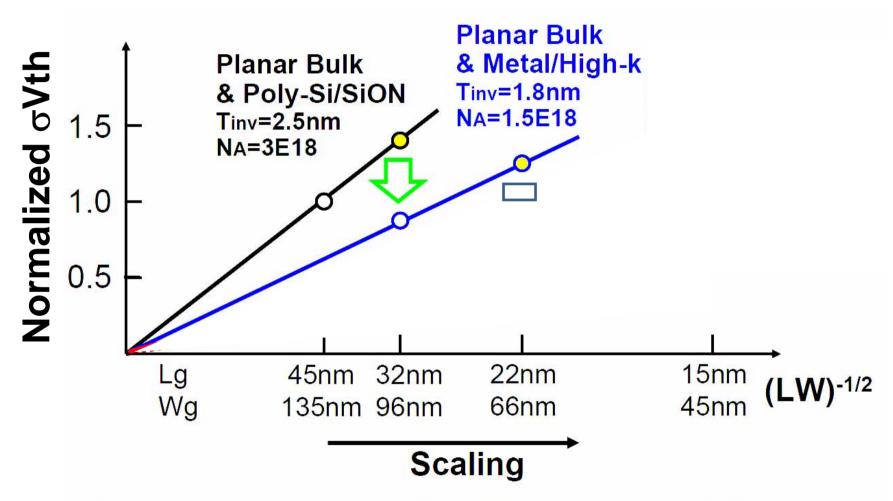
\* Because Lg is very large (2.5μm), gate leakage is large in case (a). The gate leakage component was subtracted from measured data for case (a). However, if we make small gate length, the gate leakage current should become sufficiently small to be ignored compared with Id as we verified with SiO<sub>2</sub> gate before (Momose et al.,IEDM 1994). The gate leakage could be suppressed by modifying material and process in future.

\*\* Estimated by Id value

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

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

## Random Variability Reduction Scenario in ITRS 2007

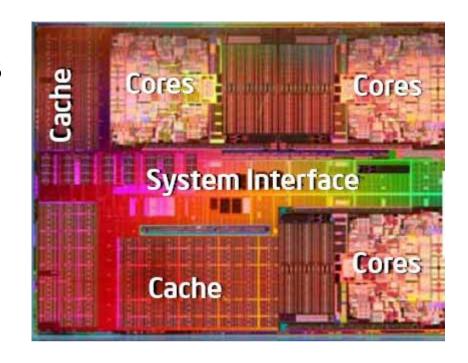


Assumption: Random dopant fluctuation is Main source of Random Variability: Line width roughness of Lg and Wg is not considered in this source: 2007 TES WINTER PUBLIC CONT.

### 4. SRAM cell scaling

## Intel® Xeon® 7400 Series (Dunnington)

45 nm high-k 6 cores 16MB shared L3 cache



Source: Intel Developer Forum 2008

#### Cache occupies huge area

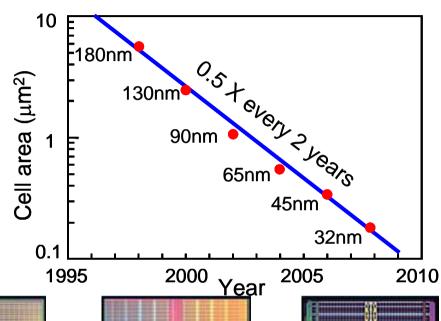
- → Cell size of SRAM should be minimized
- → However, Isd-leak should be minimized
  - → Vth are often designed to be higher than Min. logic Vth
  - → Lg are often designed to be larger than Min. logic Lg

#### Intel's **SRAM** test chip trend

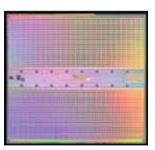
Source: B. Krzanich, S. Natrajan, Intel Developer's Forum 2007 <a href="http://download.intel.com/pressroom/kits/events/idffall\_2007/Briefing">http://download.intel.com/pressroom/kits/events/idffall\_2007/Briefing</a> Silicon&TechManufacturing.pdf

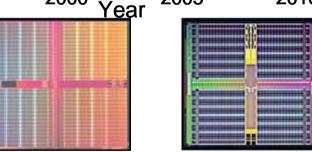
## SRAM down-scaling trend has been kept until 32nm and probably so to 22nm











Technology
Cell size
Capacity
Chip area
Functional Si

90 nm Process
1.0 μm²cell
50 Mbit
109 mm²
February '02

65 nm Process
0.57 μm²cell
70 Mbit
110 mm²
April '04

45 nm Process
0.346 μm²cell
153 Mbit
119 mm²
January '06

32 nm Process
0.182 μm²cell
291 Mbit
118 mm²
September '07

#### 22 nm technology 6T SRAM Cell: Size = $0.1\mu$ m

Source: <a href="http://www-03.ibm.com/press/us/en/">http://www-03.ibm.com/press/us/en/</a> pressrelease/24942.wss

Announced on Aug 18, 2008

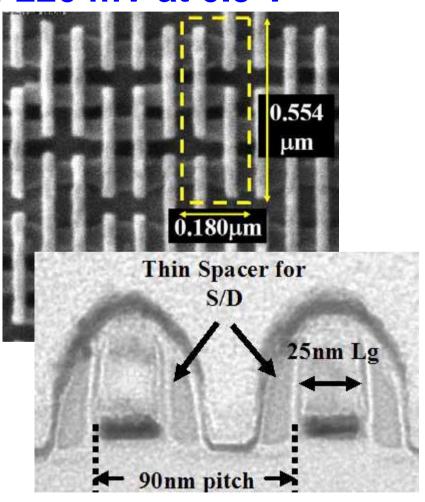
Consortium: IBM (NYSE), AMD, Freescale, STMicroelectronics, Toshiba and the College of Nanoscale Science and Engineering (CNSE)

## **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

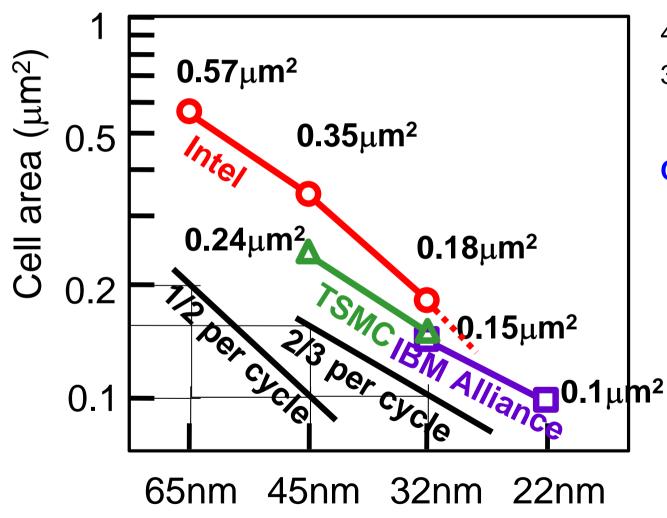
## Static noise margin of 220 mV at 0.9 V



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

#### Cell size reduction trends

1/2 or 2/3 per cycle?





#### **Functional Si**

65nm Apr.2004

45nm Jan.2006

32nm Sep.2007



**TSMC** 

#### **Conference (IEDM)**

45nm Dec.2007

32nm Dec.2007

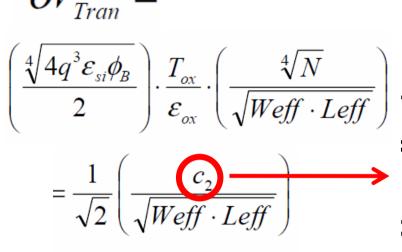


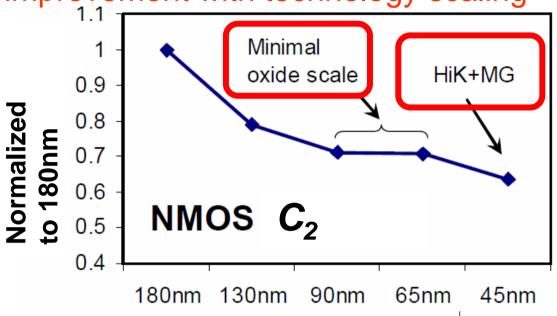
#### **Conference (IEDM)**

32nm Dec.2007

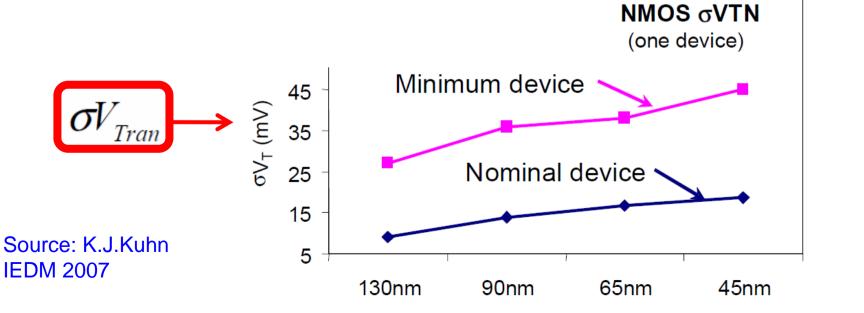
Press release 22nm Aug.2008

## NMOS Mismatch Coefficient ( $C_2$ ) improvement with technology scaling



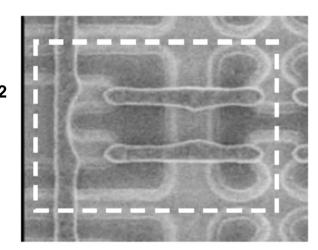


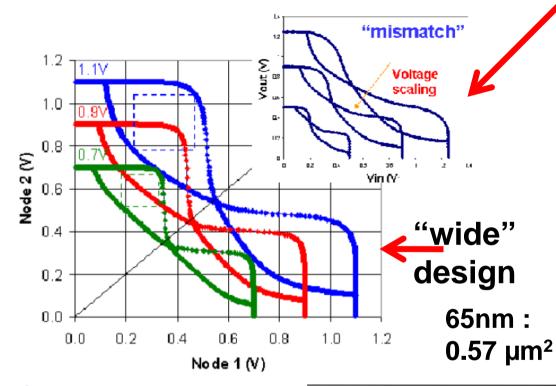
46

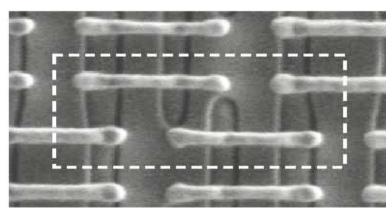


Mismatch improvement by layout (Intel)

"tall" design 90nm :1.0 µm<sup>2</sup>



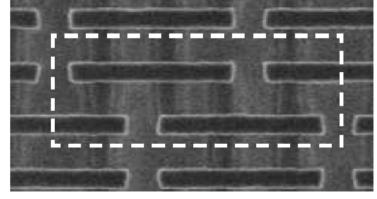




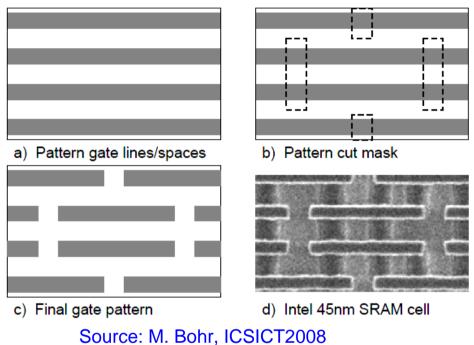
Source: K. J. Kuhn IEDM2007 Tech. Dig. pp.471

"wide" design (Square endcaps)

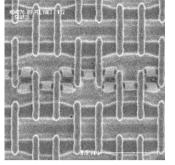
45nm 0.346 µm<sup>2</sup>



#### **Double patterning for square endcap**

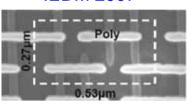


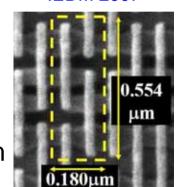
Cell evolution is similar



TSMC 45nm **IEDM 2007** 

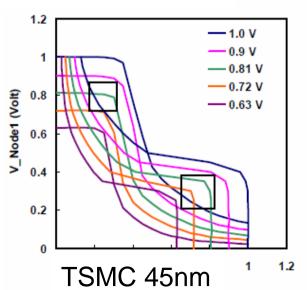
TSMC 32nm **IEDM 2007** 

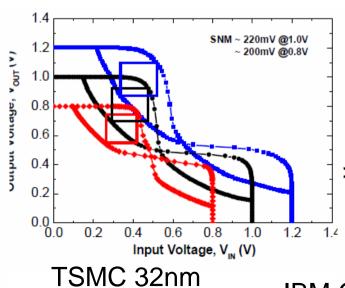


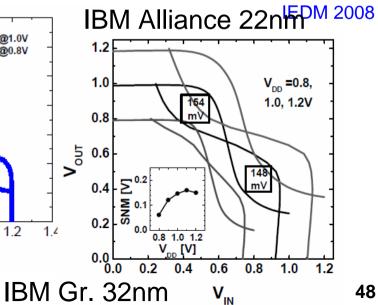


IBM Alliance 32nm

**IEDM 2004** 







48

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

#### Nehalem(Intel) 2,4 or 8 Cores

Voltage/Frequency Partitioning

DDR Vcc

Core Vcc

Uncore Vcc

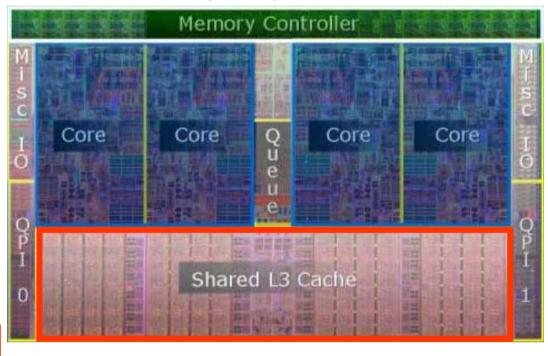
Dynamic Power Management

#### 8T SRAMCell

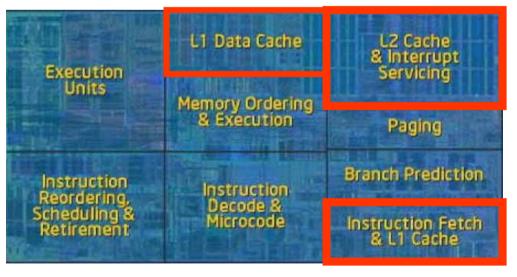
32kB L1 I -cache 32kB L1 D-cache 256kB L2 -cache

#### **6T SRAMCell**

8 MB L3 cache



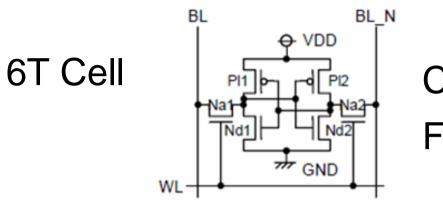
Chip



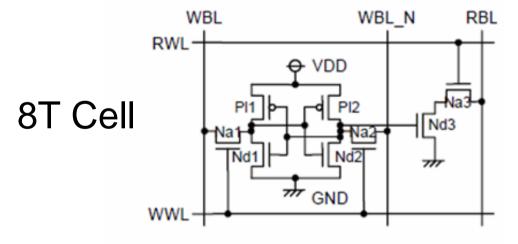
Core

Source: Intel Developer Forum 2008

#### 6T and 8T Cell



Cell size is small For high density use



Source: Morita et. al, Symp. on VLSI Circ. 2007

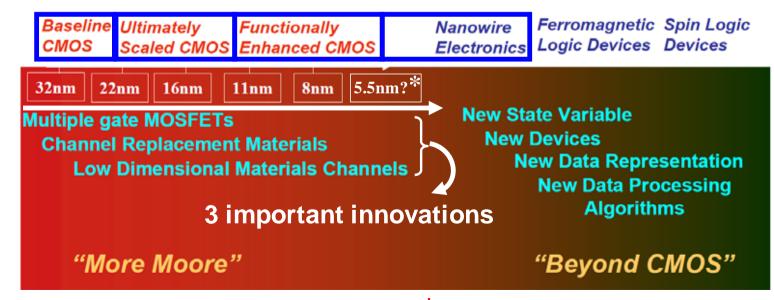
Add separate read function

Cell size increase 30%

For low voltage use

## 5. Roadmap for further future as a 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 downscaling limit, in some year between 2020-30 (H. Iwai, IWJT2008).
- -Even 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
  - 1. Nanowire/tube MOSFETs
  - 2. Alternative channel MOSFETs (III-V, Ge)
- Other Beyond CMOS devices are still in the cloud.



ITRS figure edited by Iwai

#### Si nanowire FET with Semi-1D Ballistic Transport

Drain

**Merit of Si-nanowire** 

Reduction in loff (Isd-leak)

Source: Y. Lee., T. Nagata., K. Kakushima., K. Shiraishi, and H. Iwai, IWDTF 2008, Tokyo, November, 2008

#### Trade off

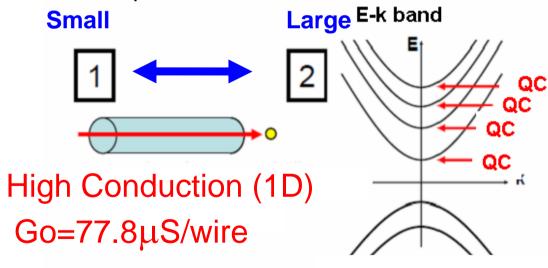
Carrier scattering probability

**Small** 

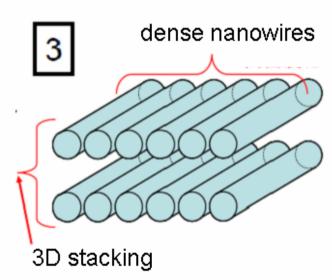
Large

# of quantum channel

**Increase in Ion (Id-sat)** 



Multiple quantum channel (QC) used for conduction



Good control of

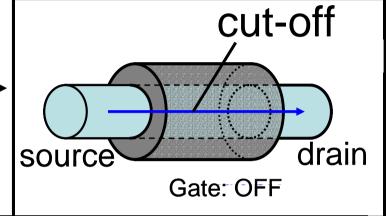
Isd-leak by

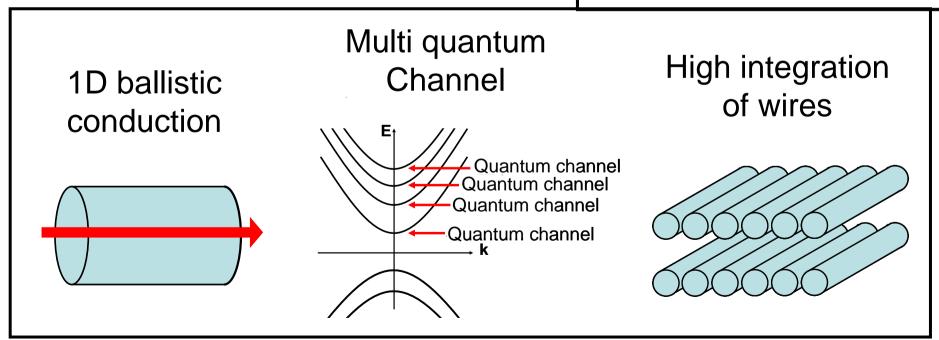
source surrounding gate

High-density lateral and vertical integration

## Si nanowire FET as a strong candidate Compatibility with after CMOS limitation

- 1. Compatibility with current CMOS process
- 2. Good controllability of I<sub>OFF</sub>
- 3. High drive current





# TEM image gate electrode 10nm sidewall wire

#### $L_0=160$ nm, $T_{ox}=3$ nm 10-3 $V_d = 1.0V$ $V_{q}$ - $V_{th}$ =1.2 $V_{\sim}$ current (A) 10.2 (step 0.2V) $V_d = 50 \text{mV}$ 20 15 O 10-13 <sub>a</sub>=160nm rain $T_{ox}$ =3nm $(A)10x10nm^2$ 10-15 0.2 0.4 0.6 0.8 1.0 -0.5 -1.0 Gate voltage (V) Drain voltage (V) Nice cut-off High drive

## Advantage of Si nanowire Large drive current

Spec. in 2019 by ITRS  $I_{ON}$ =2.3mA/ $\mu$ m  $L_g$ =11nm,  $T_{ox}$ =0.6nm

Our nanowire FET  $I_{ON}=0.25\text{mA/}\mu\text{m}$  (with 2010  $L_g=160\text{nm}$  Litho. tech.)  $T_{ox}=3.0\text{nm}$ 

With 2019 litho. tech.

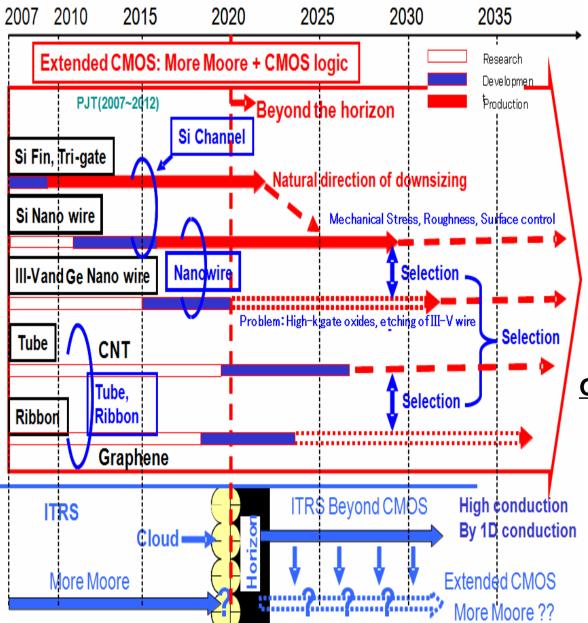
 $I_{ON}$ =2.3mA/ $\mu$ m will be obtained even with  $L_g$ =80nm and  $T_{ox}$ =1.5nm by the courtesy of Professor H.Iwai

#### Our roadmap for R &D

#### **Current Issues**

Source: H. Iwai, IWJT 2008

Si Nanowire



Control of wire surface property Source Drain contact Optimization of wire diameter

Compact I-V model

#### **III-V & Ge Nanowire**

High-k gate insulator
Wire formation technique

#### CNT:

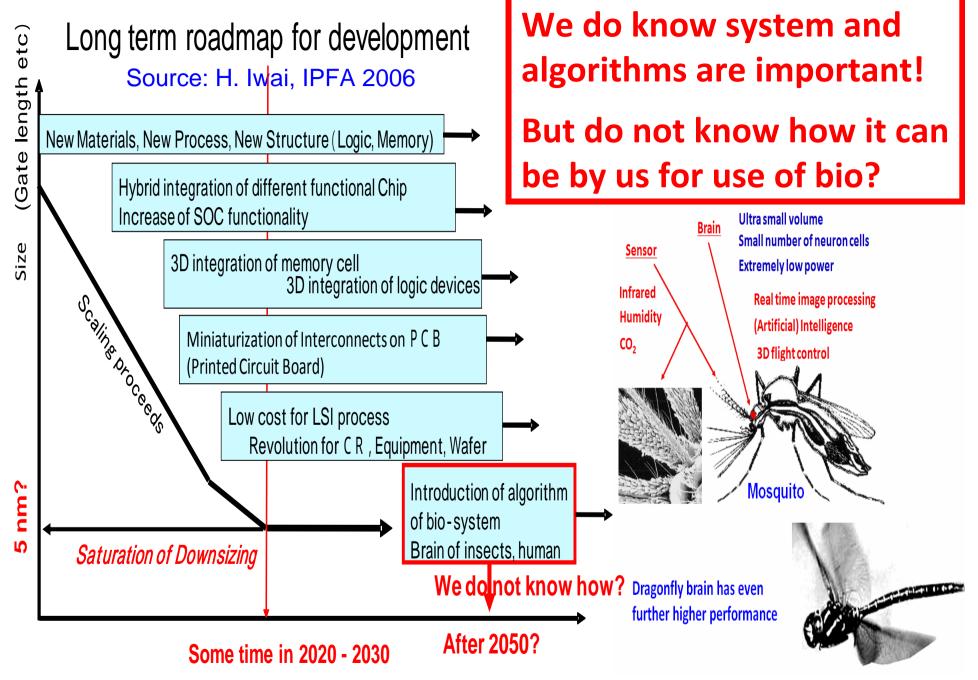
Growth and integration of CNT Width and Chirality control Chirality determines conduction types: metal or semiconductor

#### **Graphene:**

Graphene formation technique Suppression of off-current

Very small bandgap or no bandgap (semi-metal)

Control of ribbon edge structure which affects bandgap



## Thank you for your attention!