Part II

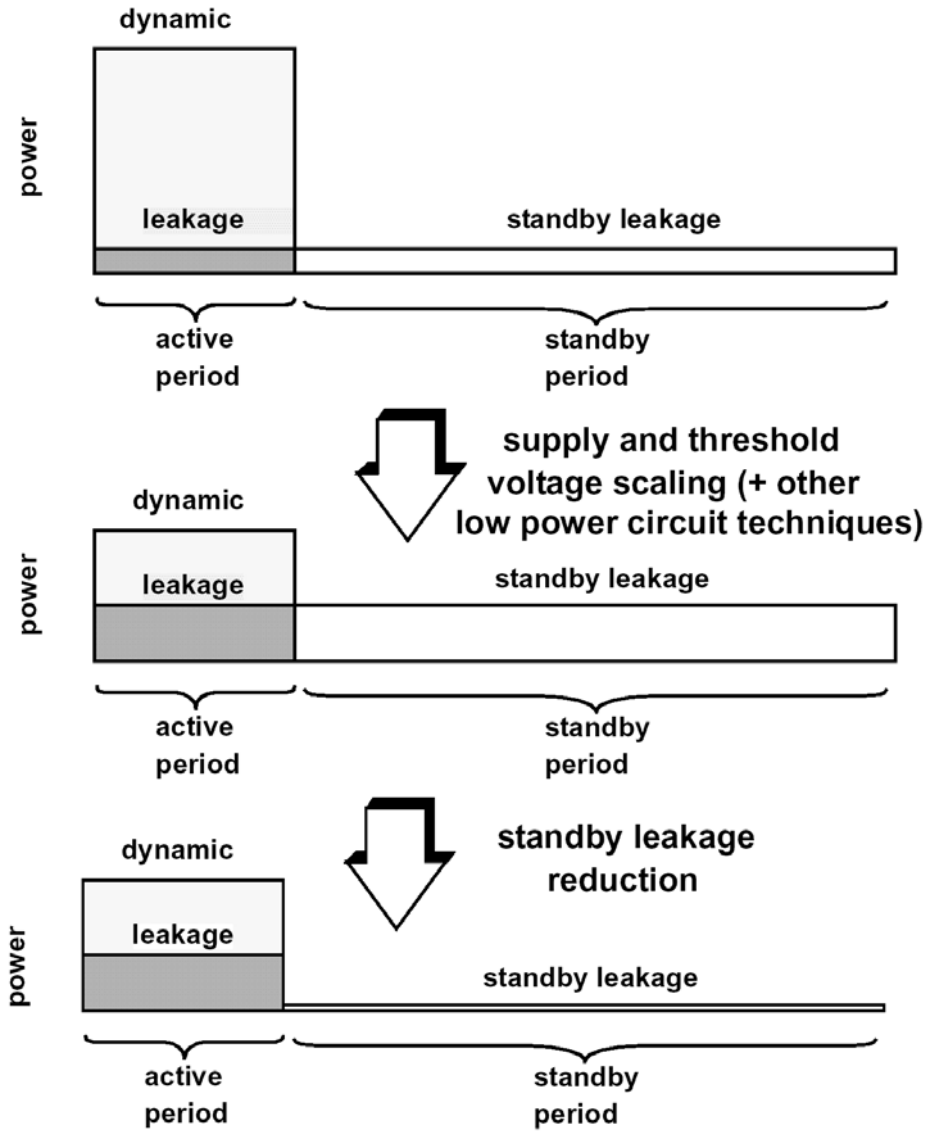
Leakage Reduction Techniques

Leakage Reduction Mechanisms

$$I_{leakage} = I_0 \exp^{\frac{\left(V_{gs} - V_{t0} - \gamma V_s + \eta V_{ds}\right)}{nVth}} * \left(1 - \exp^{\frac{\left(-V_{ds}\right)}{Vth}}\right)$$

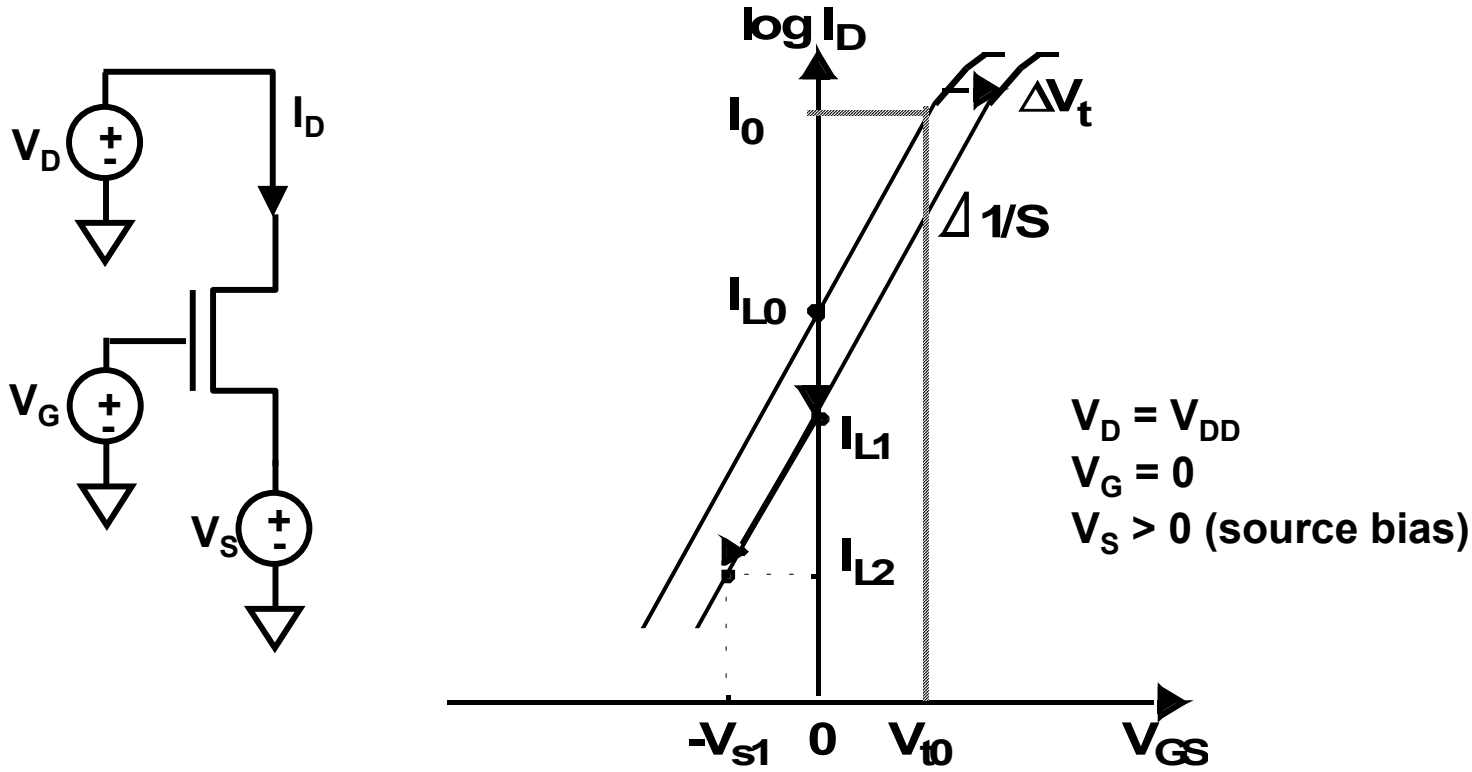
- **Increase V_t**
  - **dual threshold Voltage/ MTCMOS/ VTCMOS**
- **Increase V_S**
  - **source biasing, self reverse biasing, stack effect**
- **Decrease V_G**
  - **Super cut-off CMOS**
- **Decrease V_{DS}**
  - **not practical (CMOS output full rail)**

Standby and Active Leakage



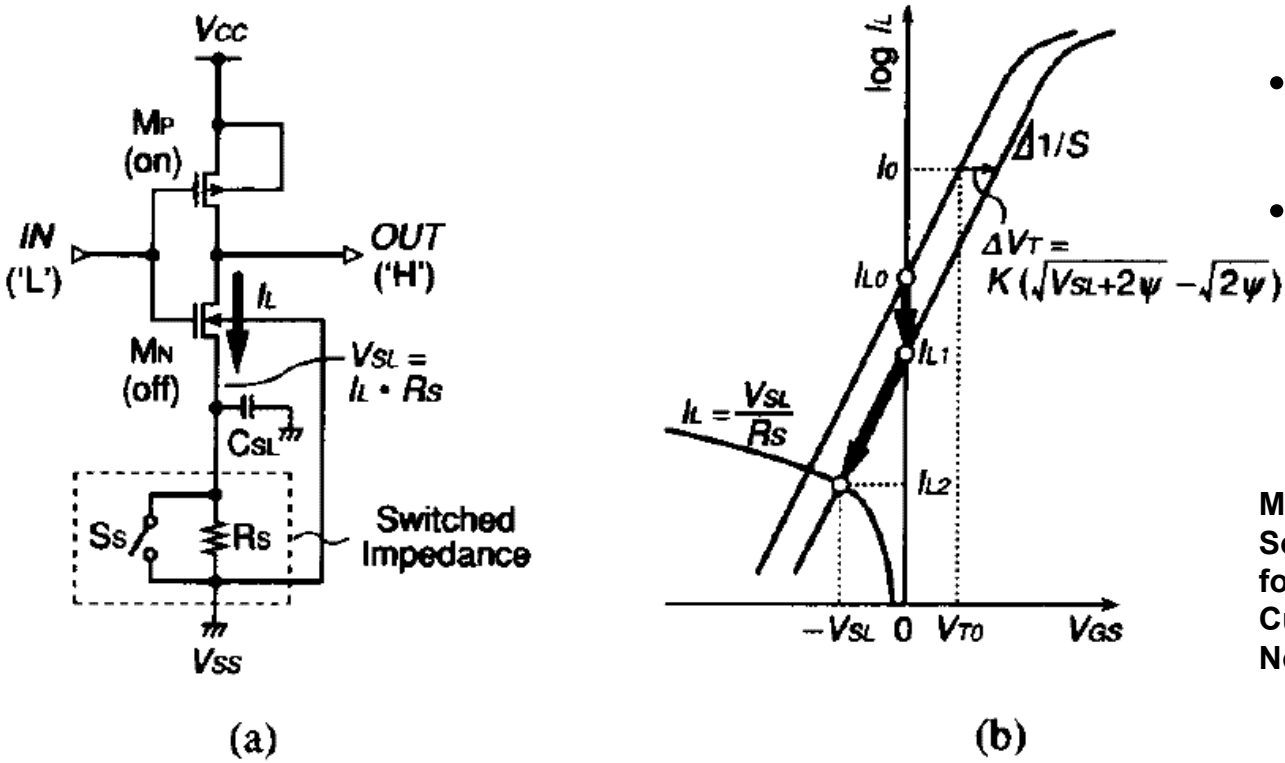
- V_t scaling causes exponential increase in leakage currents
- Dynamic power reduced with supply scaling
- Standby periods can be long (Burst Mode operation- cell phone, pager)
- Standby leakage problem more immediate
- Active leakage control can become important too

Source Biasing Principle



- V_t shift due to body effect γ
- V_{GS} becomes negative
- switched source impedance, self reverse biasing, stack effect

Switched Source Impedance

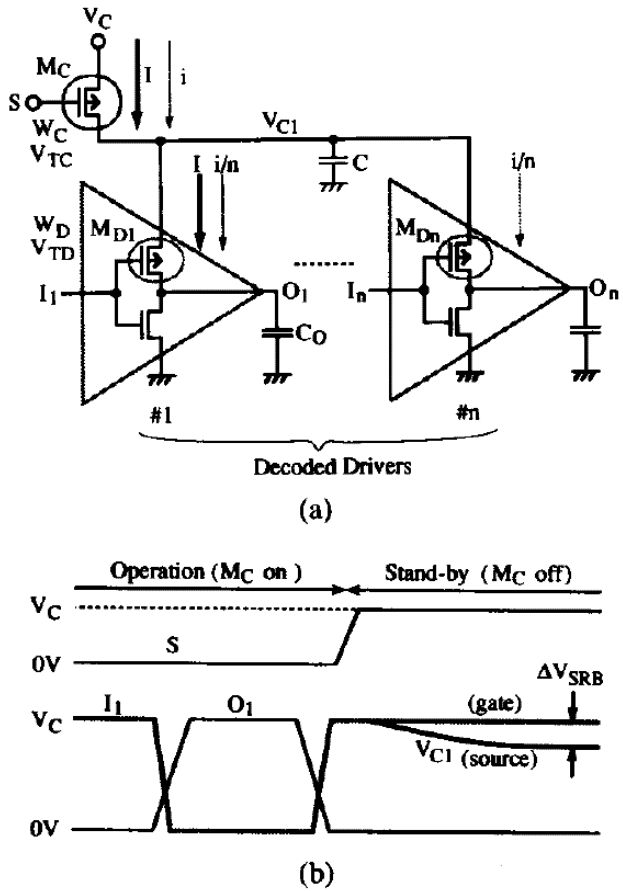


- R as source impedance
- Estimate >1000X reduction in standby leakage

M. Horiguchi, et al., "Switched-Source-Impedance CMOS Circuit for Low Standby Subthreshold Current Giga-Scale LSI's," JSSC November 1993.

Fig. 1. Principle of switched-source-impedance CMOS circuit: (a) schematic circuit diagram, (b) mechanism of subthreshold-current reduction.

Self Reverse Biasing



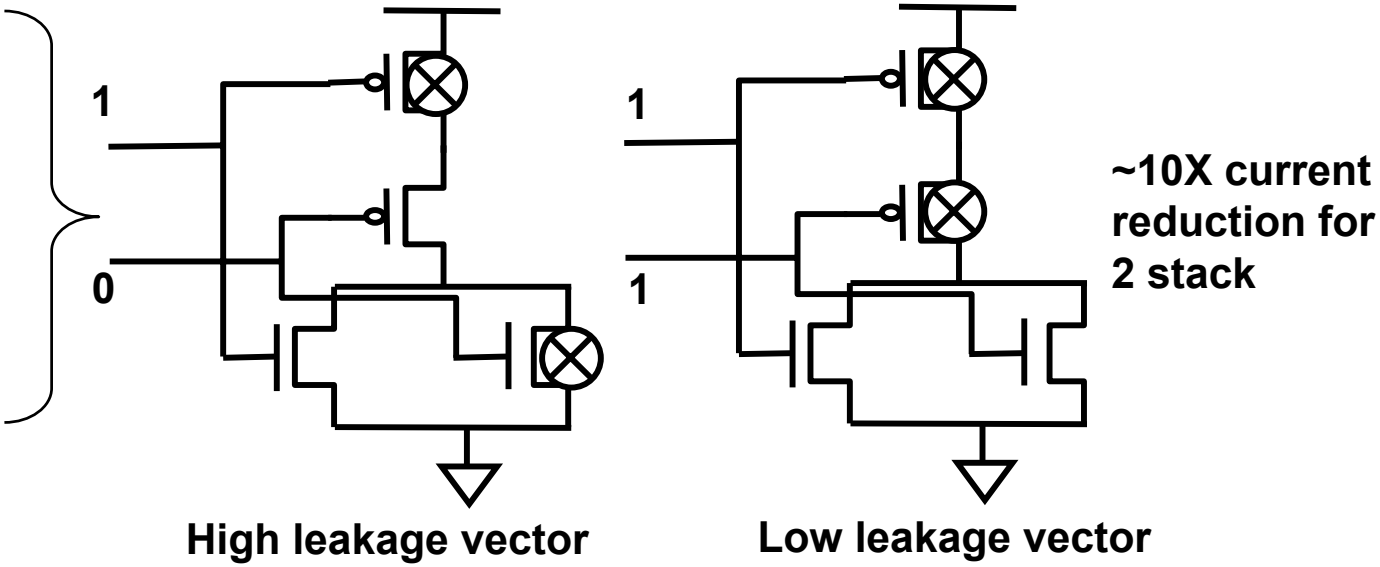
- Off device as source impedance
- Applied to DRAM decoded drivers
- Only one driver turns on (small M_C needed)
- M_{D1}-M_{D2} reverse biased during standby state
- 1000X reduction

Fig. 3. Subthreshold-current-reduced Decoded-Driver by self-reverse biasing. W_C: Gate width of M_C. V_{TC}: Threshold voltage of M_C. ΔV_{SRB}: Self-reverse biasing voltage at the steady state. W_D: Gate width of M_C. V_{TD}: Threshold voltage of M_D. I: Operation current. i: Subthreshold current. n: Number of Decoded-Drivers connected to one M_C. C_O: Output load of Decoded-Driver. C: Parasitic capacitance of common source.

T.Kawahara, et al., "Subthreshold Current Reduction for Decoded-Driver by Self-Reverse Biasing," JSSC November 1993.

Stack Effect By Vector Activation

Activate with predetermined input vector stored in latch



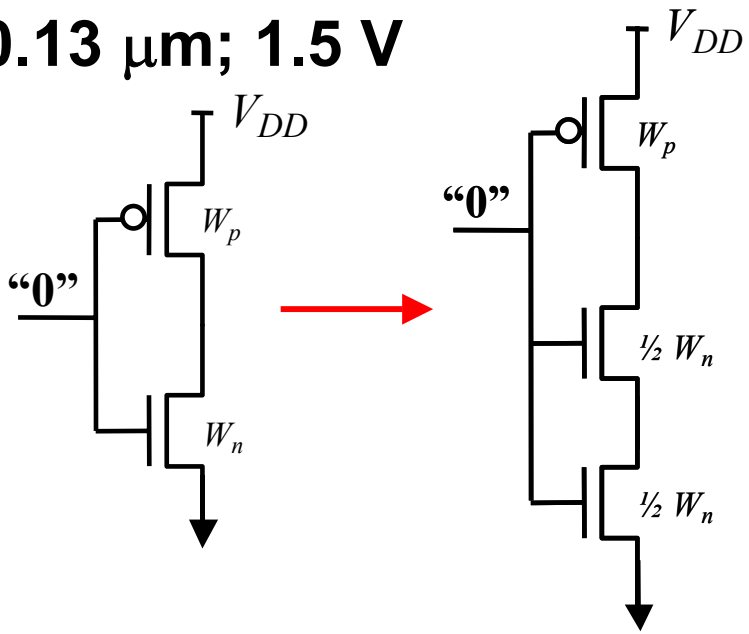
Eg. 32-bit static CMOS Kogg-Stone adder ~2X reduction in total leakage current

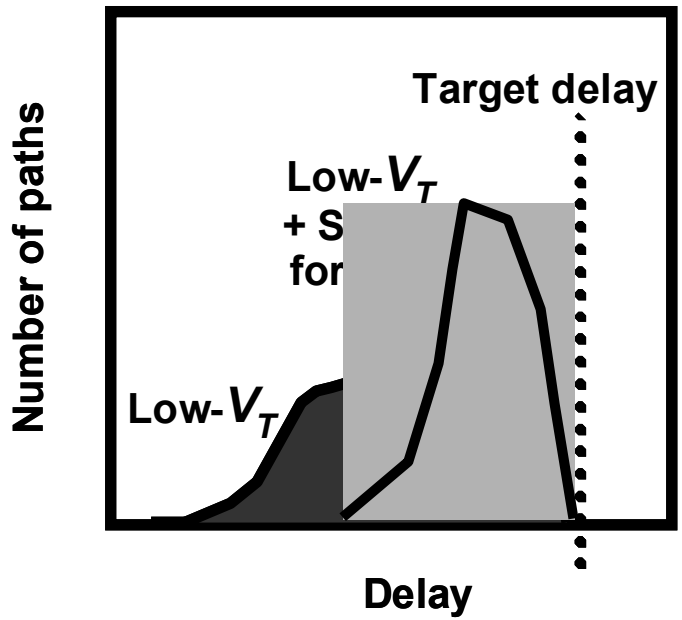
- limited by number of stacks available
- proper choice of activating vector (NP-hard algorithm -> use of heurestics)
- internal node settling time can be long
- single stacks are still HIGH leakage

Y. Ye, "A New Technique for Standby Leakage Reduction in High-Performance Circuits" VLSI Symposium, 1998.
M. Johnson, "Models and algorithms for bounds on leakage in CMOS circuits," IEEE TCAD, June 1999.

Stack Forcing Principle

0.13 μm; 1.5 V





10-30X leakage reduction
~100% higher delay

Force low-V_T stacks in non-critical paths to reduce leakage

Stack Forcing Effectiveness

32-bit μP instruction decode block 0.13 μm; 1.5 V

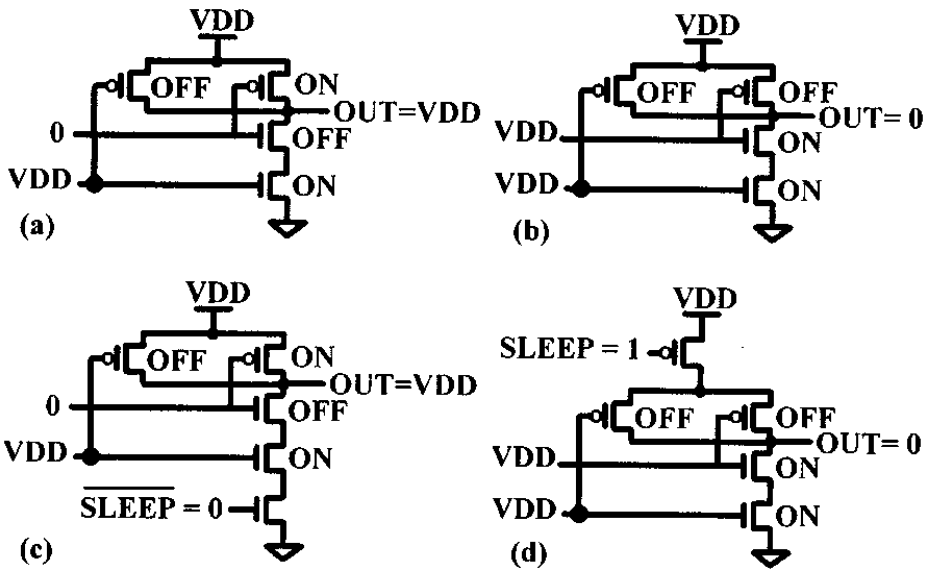
Frequency of operation:	1.0 GHz		
Active power @ 10% activity:	45.9 mW		
All Low-Vt leakage:	39.1 mW		
Dual-Vt leakage:	9.0 mW	High-Vt usage:	94.2%
Forced stack in low-Vt:	13.2 mW	Forced stack usage:	70.2%

Leakage power reduction

4.3X with dual-V_t, 3X with stack forcing

S. Narendra, S. Borkhar, D. Antoniadis, A. Chandrakasan, "Scaling of Stack Effect and its Application for Leakage Reduction," ISLPED August 2001.

Leakage Control Stack Devices



conventional

leakage controlled gates

- Single V_t leakage reduction mechanism
- Insertion of extra stack devices (in addition to vector activation)
- Sleep devices can be shared among several gates
- Gives further 35% - 90% reduction compared to state dependence alone
- Boils down to single V_t version of MTCMOS (to be discussed)

M. Johnson, et. al., "Leakage Control With Efficient Use of Transistor Stacks in Single Threshold CMOS," IEEE TVLSI, Feb 2002.

Dual V_t CMOS

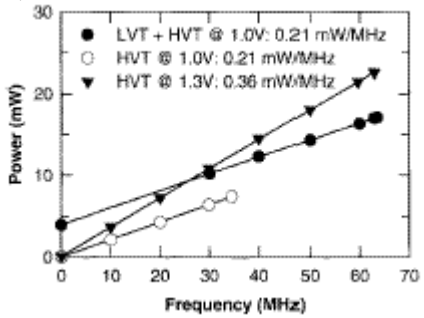
$$I_{leakage} = I_0 \exp^{\frac{\left(V_{gs} - V_{t0} - \gamma V_s + \eta V_{ds}\right)}{nVth}} * \left(1 - \exp^{\frac{\left(-V_{ds}\right)}{Vth}}\right)$$

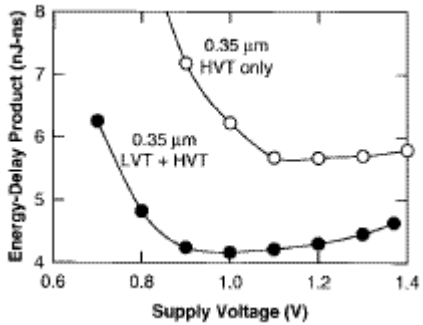
- **Dual V_t more effective at reducing leakage currents than source biasing**
- **Multiple threshold technologies more common**
- **For S=85 mV/Decade**
  each 255mV shift = 3 orders of magnitude reduction
- **Low V_t device= fast, high leakage**
- **High V_t device= slow, low leakage**
- **Achieves both Active and Standby leakage reduction**

Dual V_t Gate Partitioning

A simple approach: Use Low V_t cells for time-critical paths to improve performance

- **W. Lee, "A 1V Programmable DSP for Wireless Communications," JSSC, Nov. 1997**





- **RN. Rohrer, "A 480 Mhz RISC uProcessor in a .12um Leff CMOS Technology with Copper Interconnects", JSSC Nov. 1998.**

  Use of LVT in 4% of standard cells yield 6.5% performance improvement

- **T. Yamashita, "A 450 Mhz 64b RISC Processor Using Multiple threshold Voltage CMOS," ISSSC Feb 2000**

  LVT + HVT improves performance by 12.5% (all LVT causes standby current to be so large as to cause thermal runaway)

Dual V_t Optimization

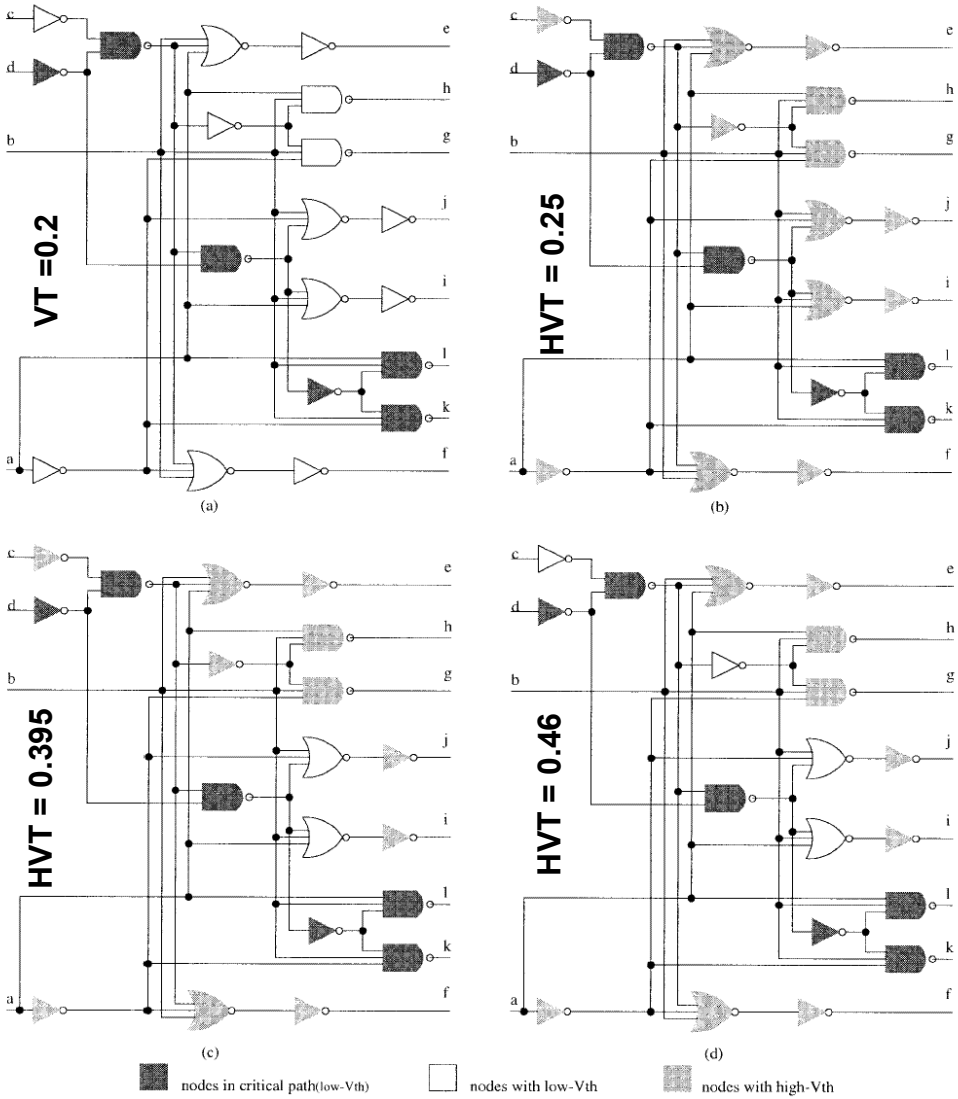


Fig. 4. An example circuit. (a) Original circuit $V_{dd} = 1$ V, $V_{th_1} = 0.2$ V. (b) $V_{th_2} = 0.25$ V. (c) $V_{th_2} = 0.395$ V. (d) $V_{th_2} = 0.46$ V.

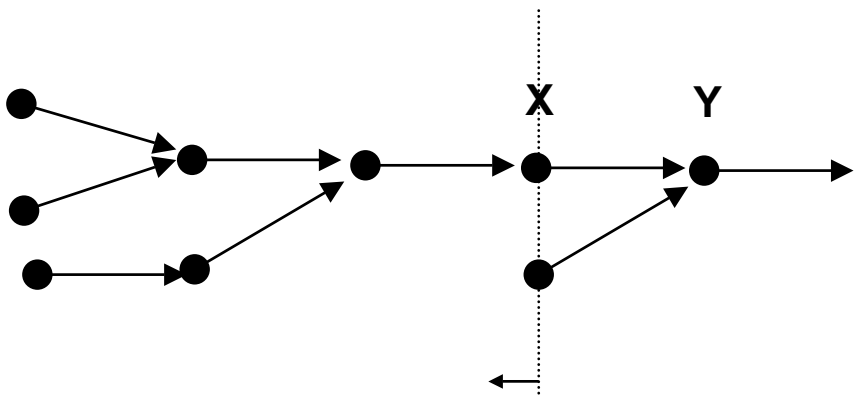
- **Initially assume all LVT (for best performance)**
- **Some non-critical gates can be made HVT**
- **Choice of high V_t determines mixture**
- **Proposal for breadth first search algorithm to assign optimal high V_t value.**

L. Wei, "Design and Optimization of Dual Threshold Circuits for Low-Voltage Low-Power Applications," TVLSI, March 1999.

A Dual V_t Partitioning Algorithm

L. Wei, "Design and Optimization of Dual-Threshold Circuits for Low-Voltage Low-Power Applications," TVLSI, March 1999

- Initially all low V_t
- For each node (gate) in graph calculate
  - Arrival time, Departure time, Propagation delay, graph level
- From output to input (back tracing level-by-level)
  - Determine slack availability for each node in level
  - Gates with enough slack set to high V_t
  - level = level -1
- Simulate and reiterate with other V_t choices



ΔX (slack)= $T_{maxarrv}(Y) - T_{dep}(X) + \Delta Y$
If $\Delta X > 0$, X can be made high V_t

Update graph with new T_p, T_{dep}, ΔX
Move to next node/ level

Advanced Dual V_t Optimization

- Gate level dual V_t -> transistor level dual V_t
  L. Wei, "Mixed-V_{th} (MVT) CMOS Circuit Design Meth. for Low Power Appl," DAC 1999
  - Improved leakage reduction
  - More involved partitioning algorithm (traverse transistors level by level)
- Combine dual V_t with transistor sizing:
  S. Sirichotiyakul, D. Blaauw, "Stand-by Power Minimization through Simultaneous V_t Selection and Circuit Sizing, " DAC 1999
  - High V_t to low V_t with same sizing can be too fast
  - Low V_t increases node capacitance seen by crossing paths
  - Use "Dominant Leakage State" + probability to estimate total leakage
  - Too complex to optimize: use heurestic approach
  - 1) Choose some V_t low for performance 2) resize circuit to win back area 3) repeat

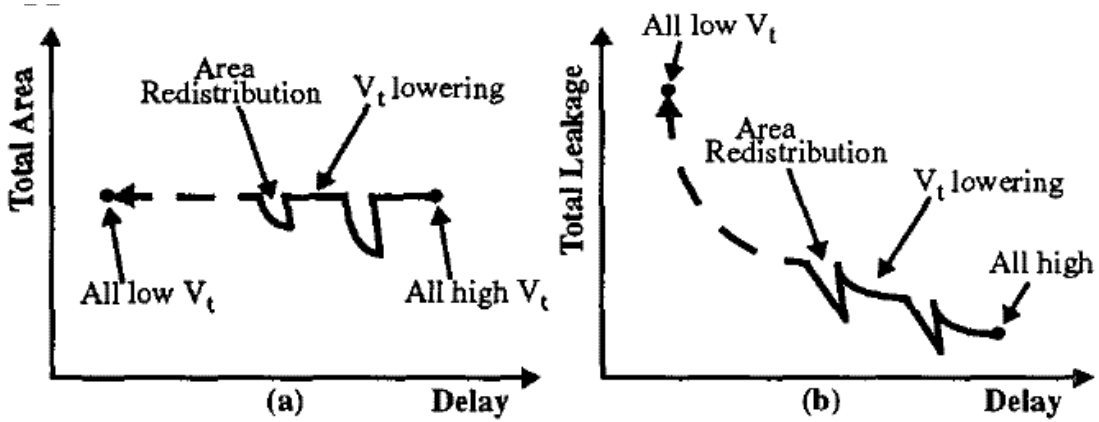


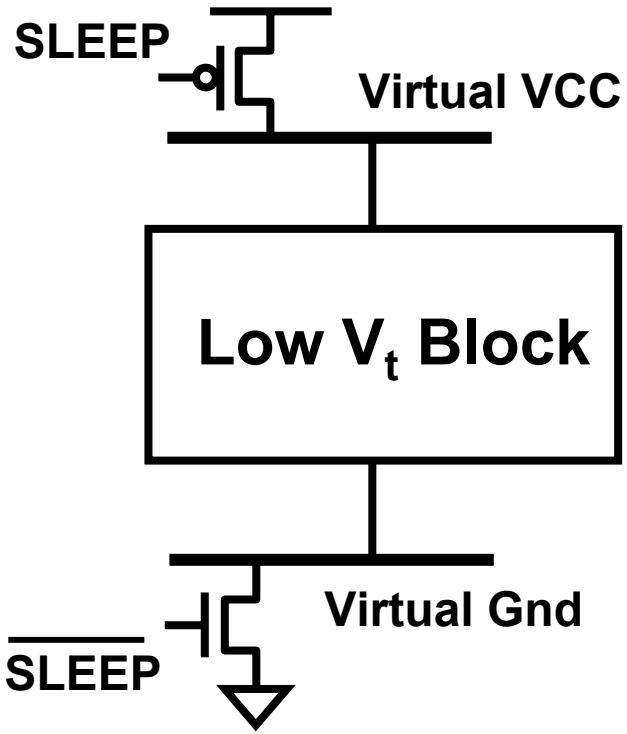
Figure 4. V_t selection and redistribution of area, two views

CAD for Dual V_t Optimization

- Leakage reduction principle simple
- Difficult to optimally choose parameters

  V_{th}, V_{tl}, V_{DD} ,device selection, transistor sizing
- Need to develop fast, efficient CAD tools

MTCMOS Principle



Single polarity sleep device sufficient for combinational logic block

Active Mode

Low V_t circuit operation (or combined)

Standby Mode

Disconnect power supplies through High V_t devices

- **For S=85 mV/Decade, ΔV_t = 225mV**
  ~1000X reduction

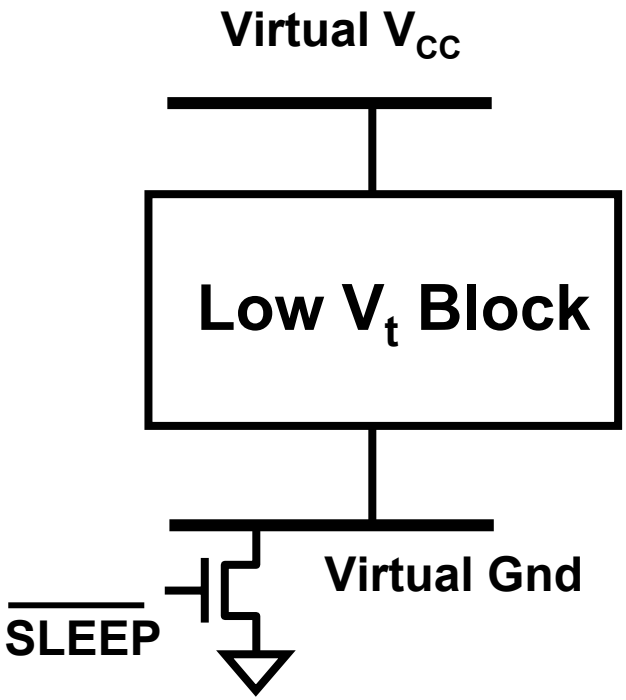
Use of LVT sleep with +/- gate (Super Cut-off/ Multi-Voltage CMOS, M. Stan, ISLPED 1998)

For fine grain sleep control

Sequential circuits must retain state

S. Mutoh, et. al., "1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS," JSSC August 1995.

MTCMOS Sleep Sizing



Virtual Ground Bounce

Gate drive decreases

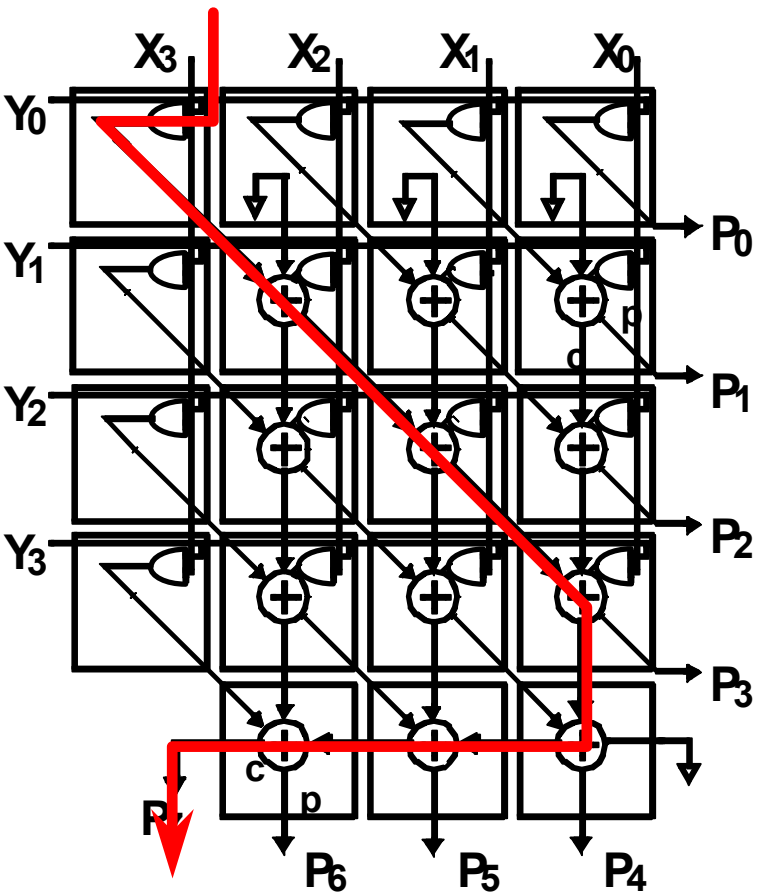
Body effect increases V_t

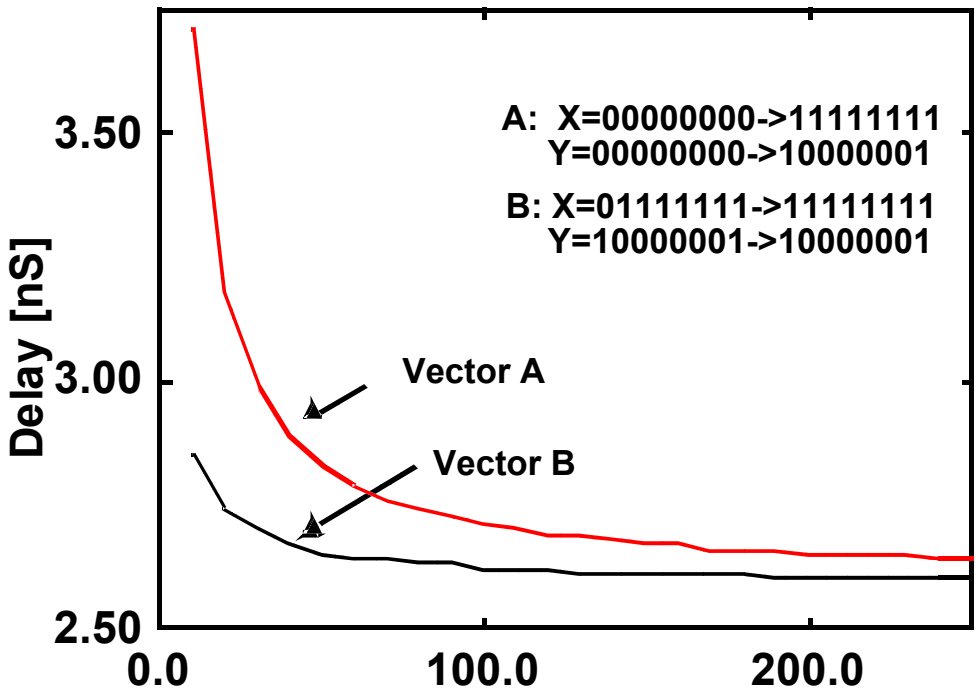
Reverse conduction noise concerns

Main Design Issue in MTCMOS

Properly size sleep transistor

Input Vector Impact

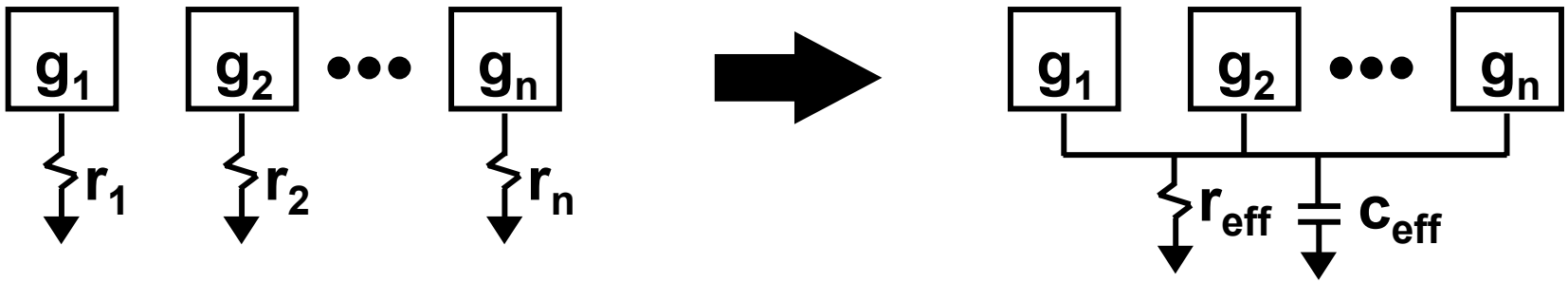




Vector	CMOS Delay	% Degr (W/L=5.4%)	% Degr (W/L=18%)
A	2.58 ns	15.4%	4.6%
B	2.59 ns	4.7%	1.6%

J. Kao, et al., "Transistor Sizing Tool and Methodology for For Multi-Threshold CMOS Technology," DAC 1997

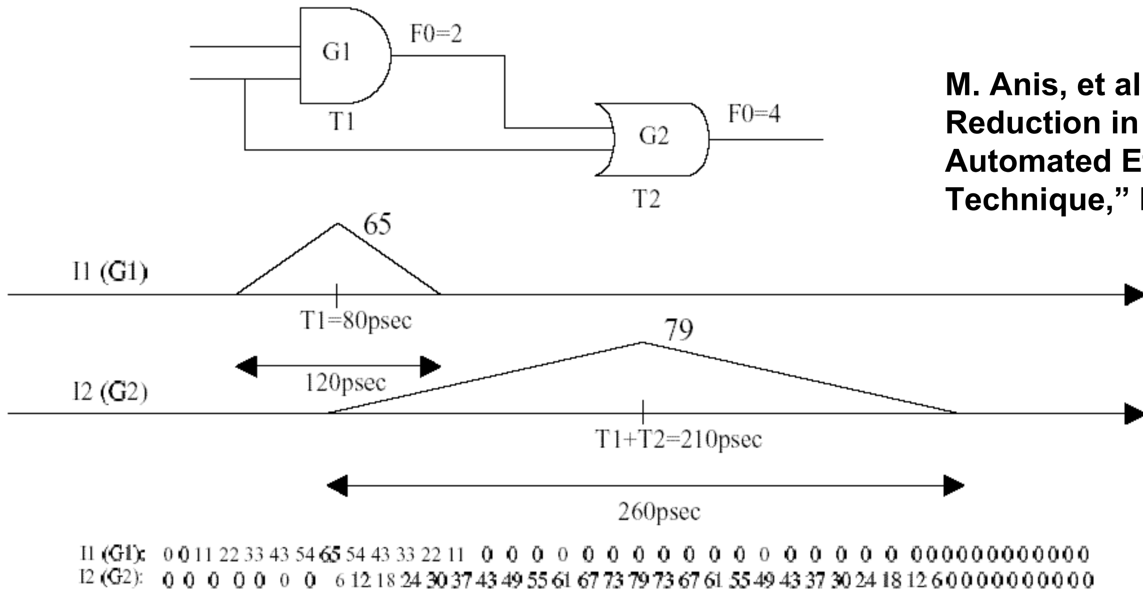
Hierarchical Sizing Approach



- Compute effective sleep resistor for each gate
  - Sets Maximum Gate Degradation
  - Overall delay is guaranteed
- Mutual exclusive gates can share common sleep transistor
- Applied at multiple hierarchical levels

J. Kao, et al., "MTCMOS Hierarchical Sizing Based on Mutual Exclusive Discharge Patterns," DAC 1998

Gate Clustering



M. Anis, et al. ,"Dynamic and Leakage Power Reduction in MTCMOS Circuits Using an Automated Efficient Gate Clustering Technique," DAC 2002.

Assumption on predictability of current discharge patterns…

- **Cluster gates that have partially overlapping discharge currents**
- **Group such that Current Budget (to ensure V_x gives 5% degredation) is maintained**
- **Use operations research clustering techniques**
  - **Bin Packing**
    - **assign gate (I_j) to bins so that total current < I_{max} and #bins minimized**
  - **Set Partitioning**
    - **Extension of BP including cost function related to routing complexities**

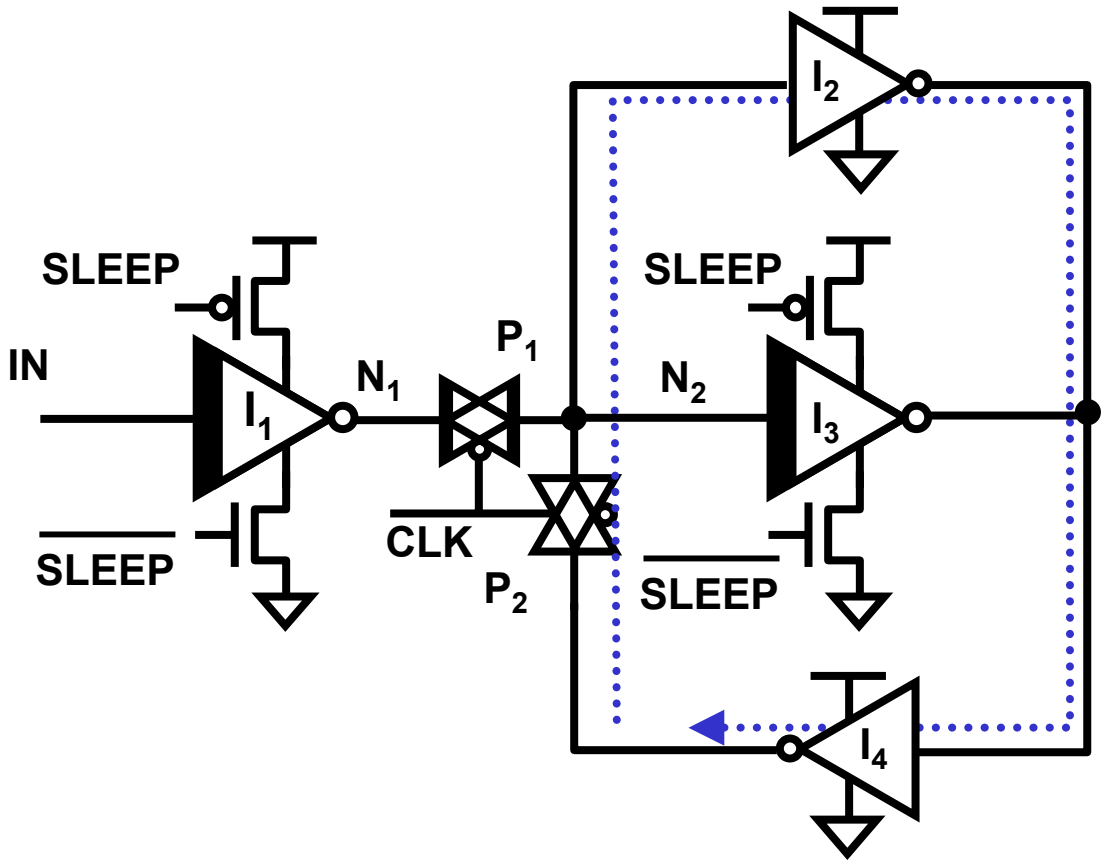
MTCMOS Sleep Sizing TBD

- Need for improved sleep transistor sizing algorithms
- Static, functional timing techniques to better characterize MTCMOS discharge patterns
- Apply ideas from similar CAD research on power /gnd noise
  - S. Bobba, I.N. Hajj, "Estimation of Maximum Current Envelope for Power Bus Analysis and Design," Intl. Symp on Physical Design, April 1998
  - G. Bai, S. Booba, I.N. Hajj, "Static Timing Analysis Including Power Supply Noise Effect on Propagation Delay in VLSI Circuits," DAC 2001
  - Y. Jiang, K. Cheng, "Dynamic Timing Analysis Considering Power Supply Noise Effects," Int. Symp. on Quality of Electronic Design, March 2000
  - F. Najm, "Survey of Power Estimation Techniques in VLSI Circuits," TVLSI Dec. 1994
  - H. Kriplani, F. Najm, I.N. Hajj, "Pattern Independent Maximum Current Estimation in Power and Ground Buses of CMOS VLSI circuits: Algorithms, Signal Correlations, and their Resolution," TCAD, August 1995
  - S. Chowdhury, J. Barkatullah, "Estimation of Maximum Currents in MOS IC Logic Circuits," TCAD, June 1990.
  - + many others …

MTCMOS Sequential Circuits

- **MTCMOS Combinational Circuits**
  - **Simple operation**
  - **Difficulty in sizing/ distributing sleep transistors**
- **MTCMOS Sequential Circuits**
  - **Virtual power/ gnd disconnected during sleep**
    - **Nodes will float**
  - **Techniques needed to maintain state**
    - **Need always powered circuits**
    - **Must avoid sneak leakage paths**

Basic MTCMOS Latch



- **Use of always powered CMOS gates**

S. Mutoh, et. al., "1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS," JSSC August 1995.

Balloon Flip Flop

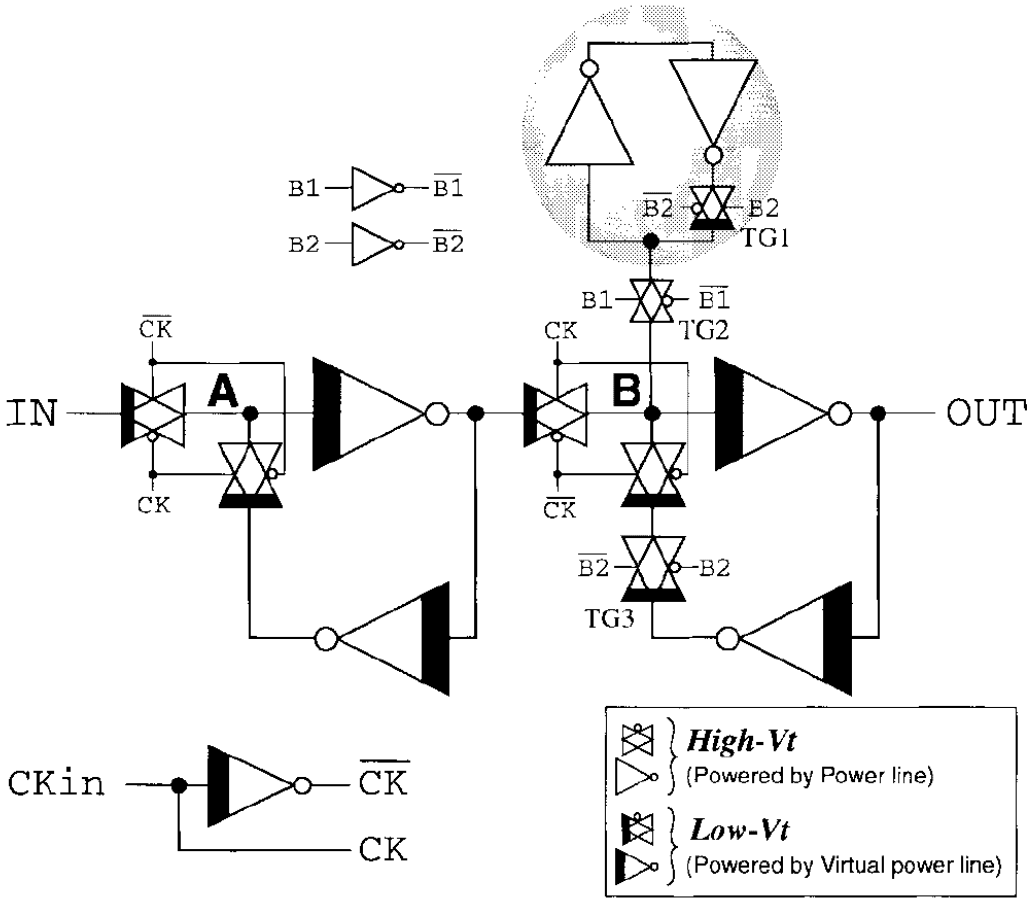


Fig. 9. A balloon circuit applied to a DFF circuit (clock-dependent type).

- **HVT storage "balloon" decoupled from LVT logic**
- **LVT blocks can share common virtual pwr/gnd**
- **Elimination of sneak leakage paths**
- **Complicated signalling**

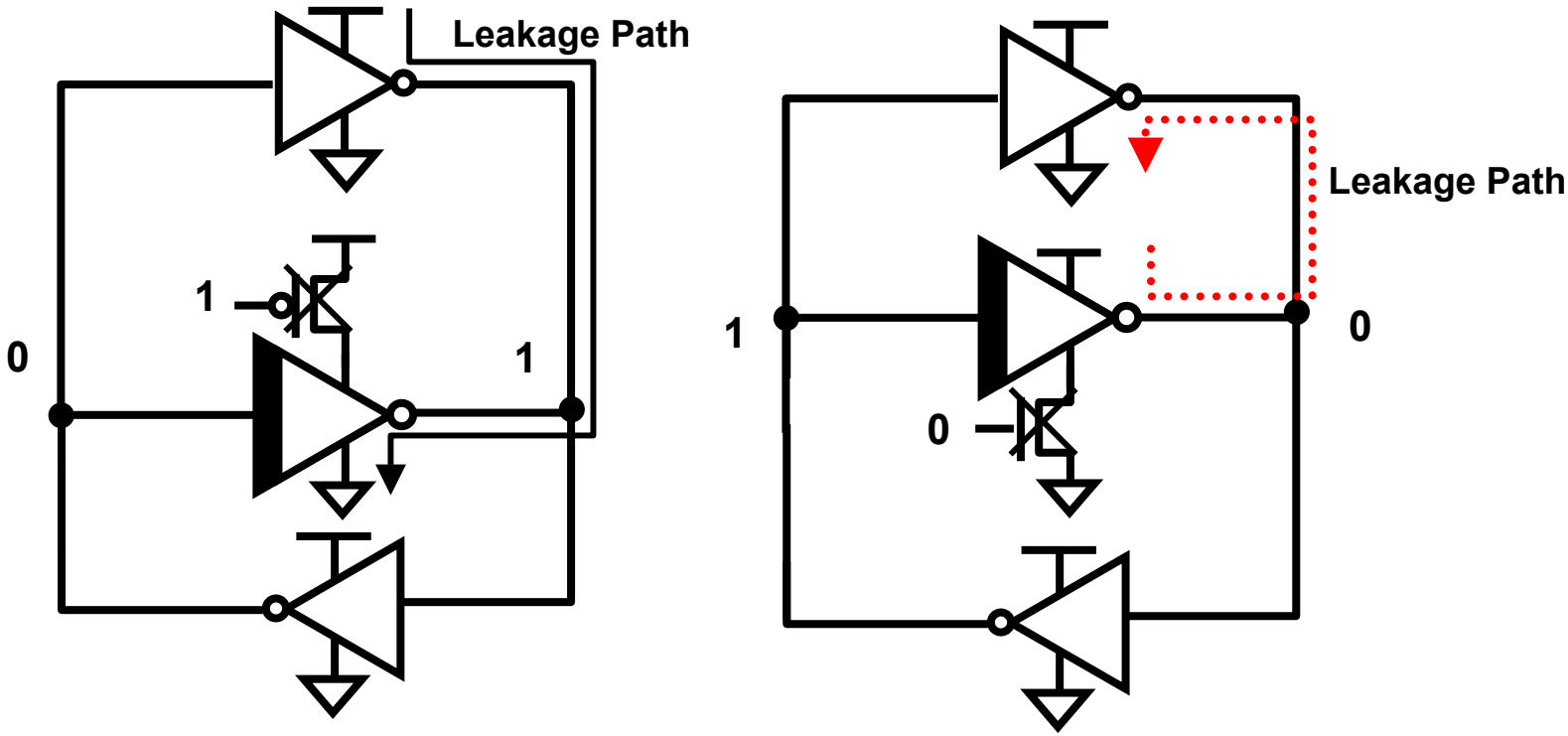
S. Shigematsu, et. al., "A 1-V High Speed MTCMOS Circuit Scheme for Power-DOWN Application Circuits," JSSC June 1997

Sneak Leakage Paths

- Sneak paths from MTCMOS/ CMOS interaction
- Leakage currents from V_{CC} to ground without passing through off high V_t device
- Need to utilize:
  - both polarity sleep devices
  - local sleep (non shared sleep devices)
  - novel structures

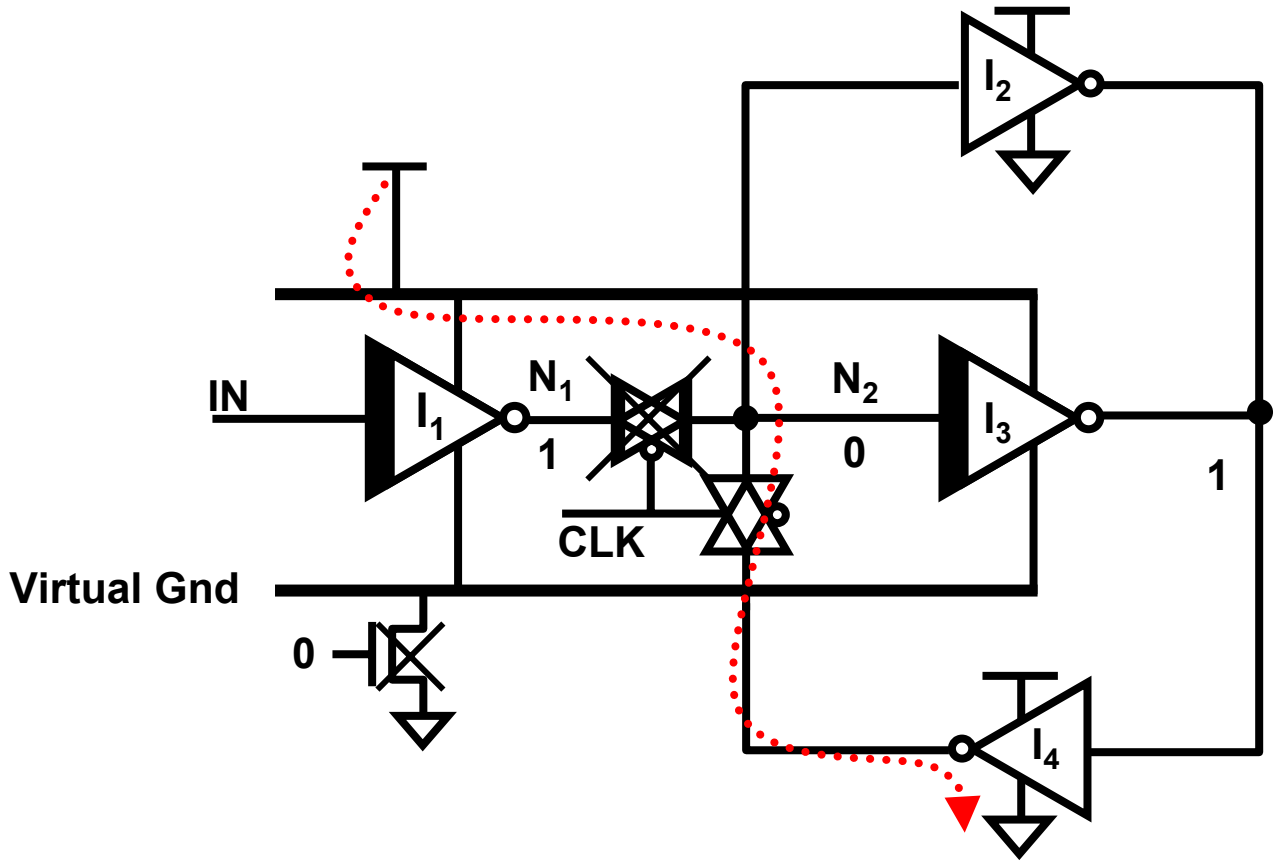
J. Kao, A. Chandraksan, "MTCMOS Sequential Circuits," ESSCIRC September 2001.

Sneak Leakage Path From Parallel Combinations



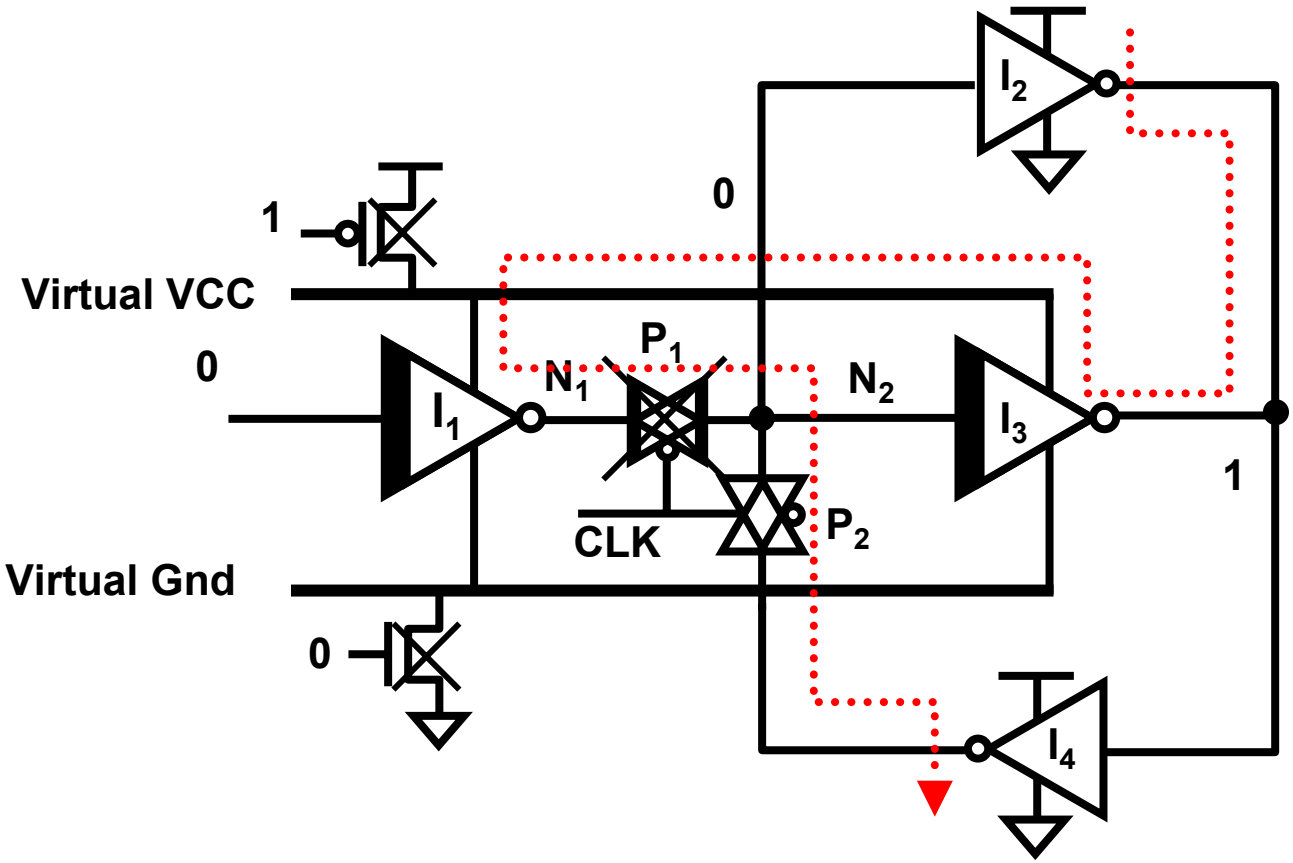
- Need for both polarity high V_t sleep devices

Sneak Leakage Path Through Low Threshold Passgate



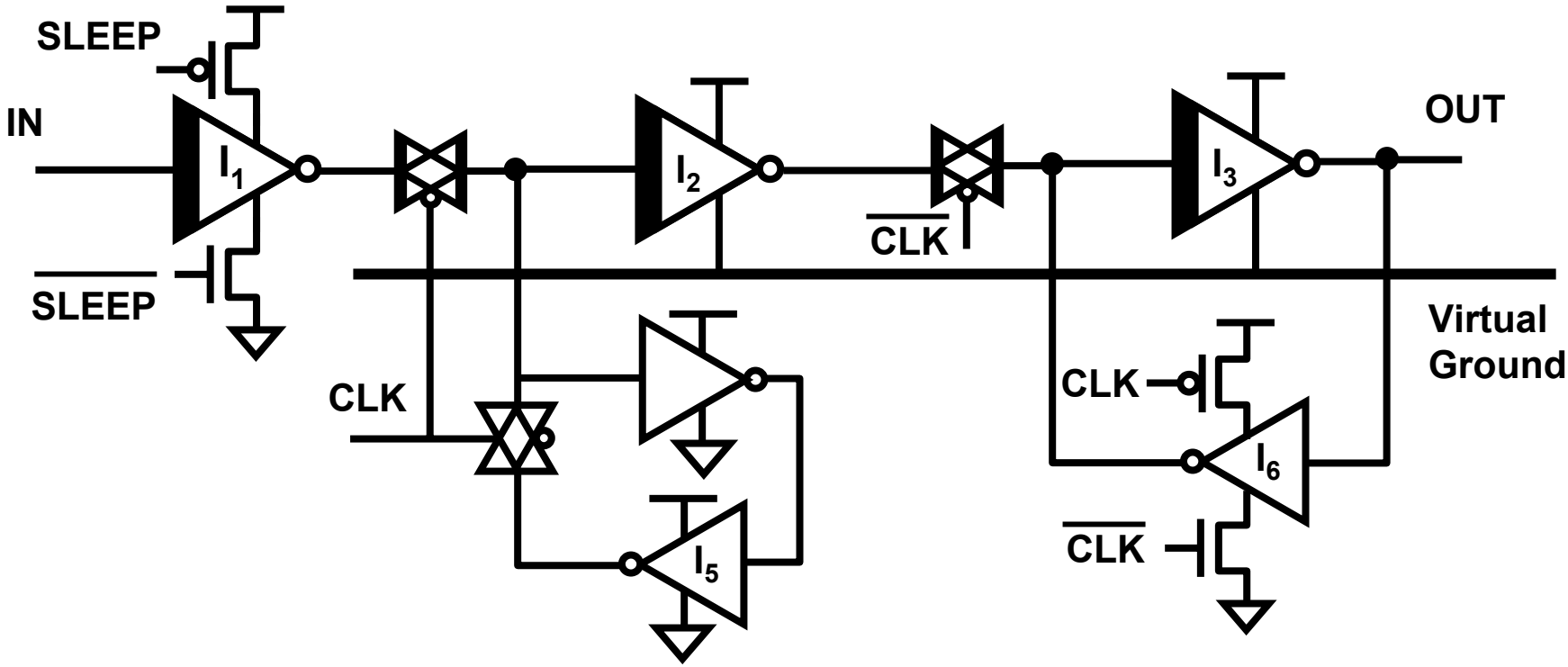
- Need for both polarity high V_t sleep devices

Sneak Leakage Path From Reverse Conduction Paths



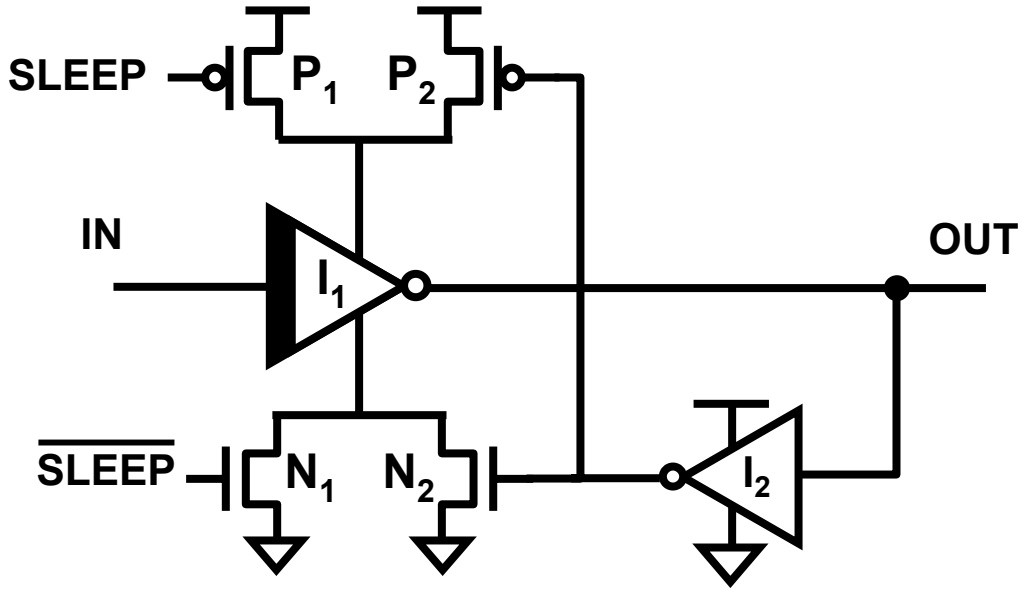
- Need for localized, non shared, high V_t sleep devices

Improved MTCMOS Flip Flop



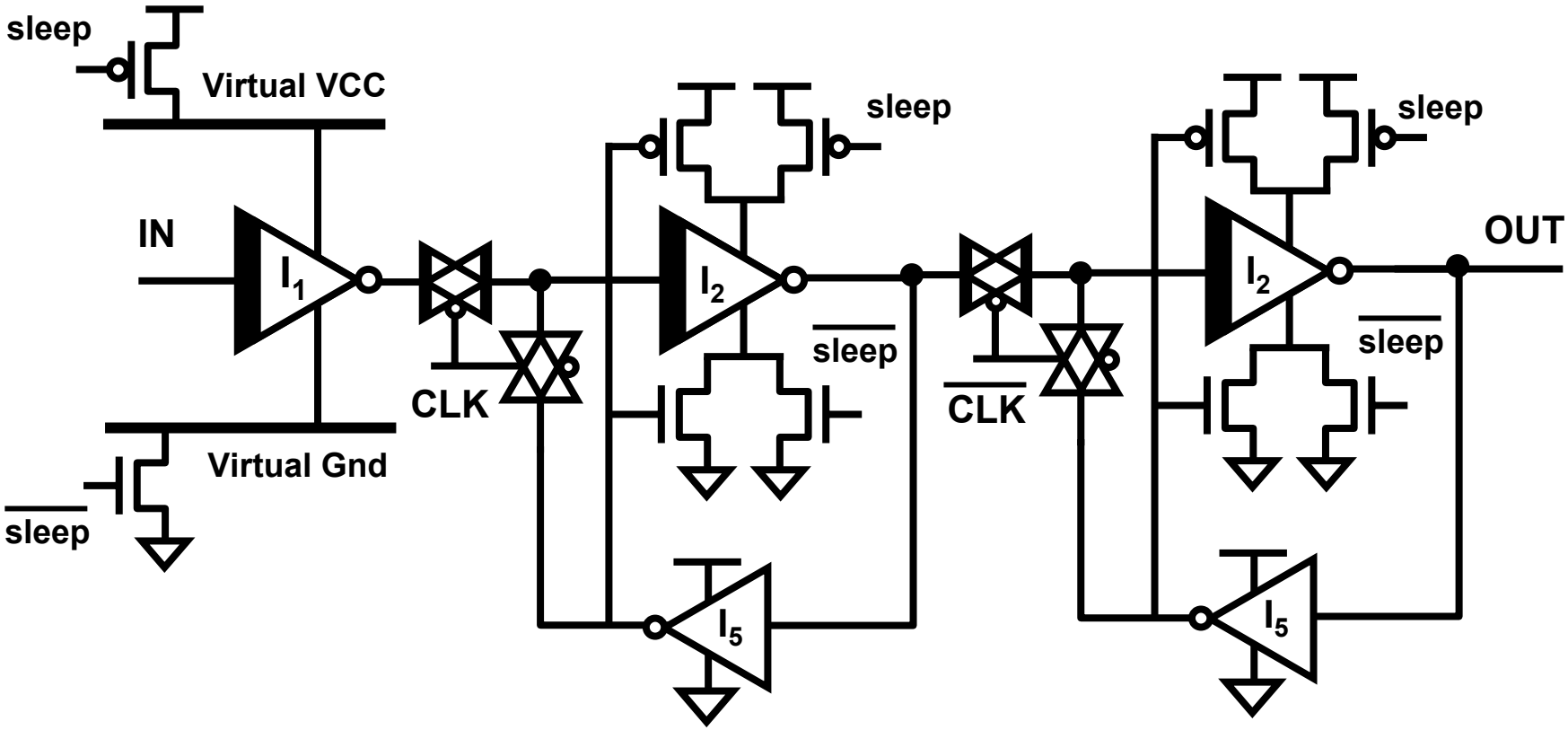
- **Careful consideration of sneak leakage paths yields improved implementation**

Leakage Feedback Gate



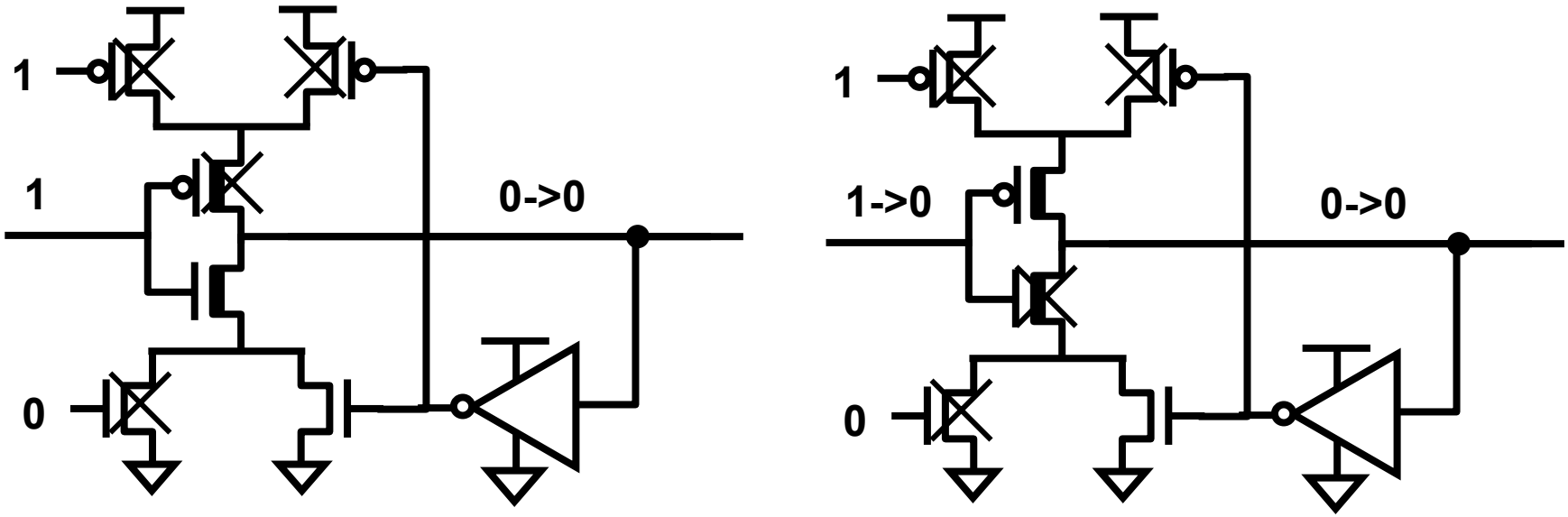
- Sufficient if either VCC or VSS path is cutoff
- Proper cutoff path yields actively driven output
- Low V_t operation + actively driven low leakage state
  - directly imported into CMOS structures

Leakage Feedback Flip Flop



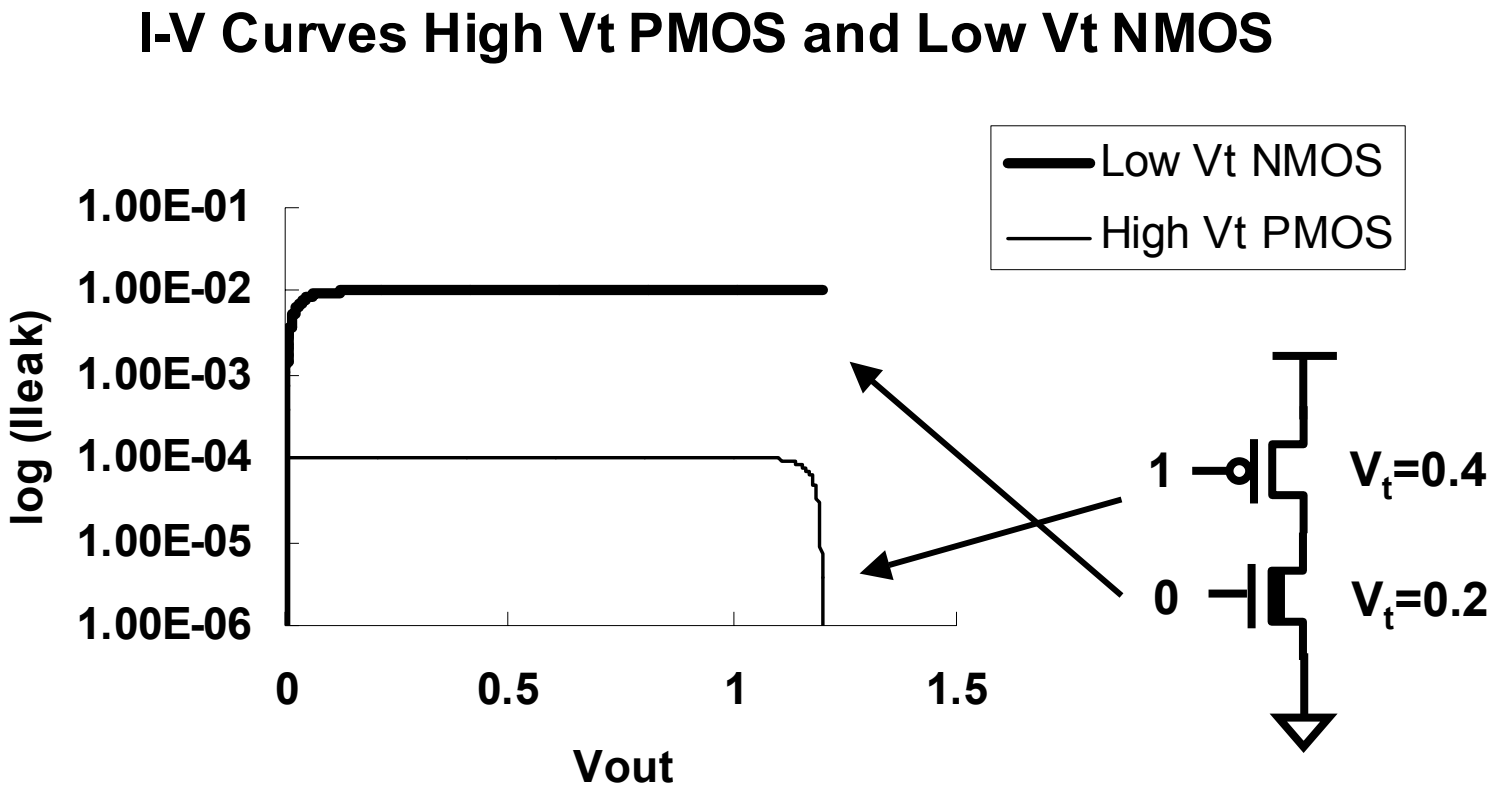
- **Virtually no extra loading -> performance is better than standard MTCMOS FF**
- **Same operation as a CMOS FF**

Leakage Feedback Effect

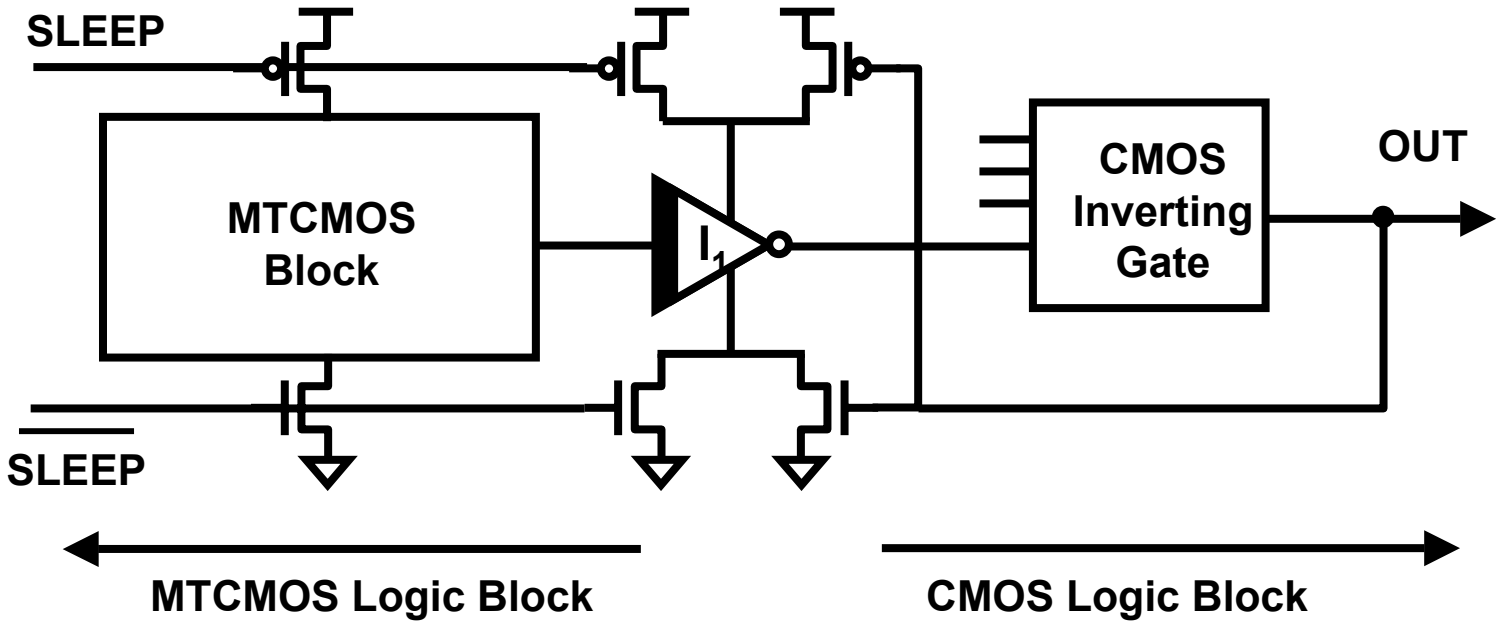


- Output data holds even if input floats
  - held by leakage mismatch
- Potential charge sharing if inputs change

Leakage Induced DC Operating Point

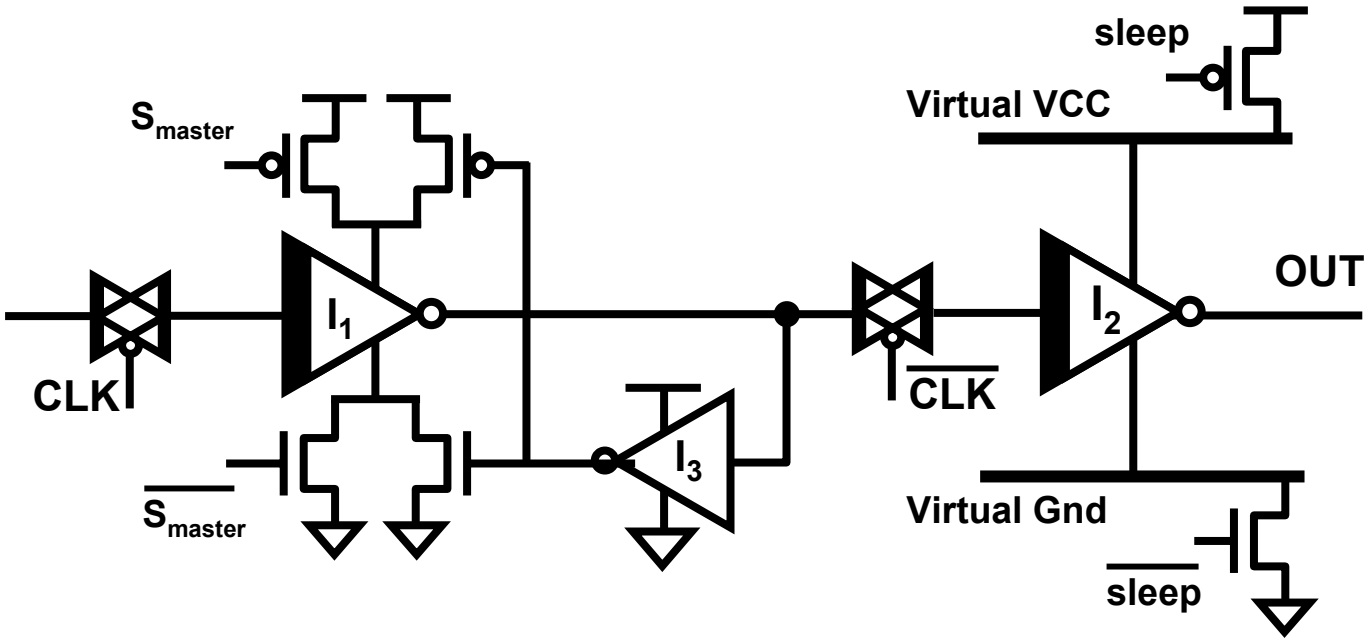


MTCMOS / CMOS Interface



- **Leakage feedback gate natural interface block between MTCMOS logic and CMOS logic**

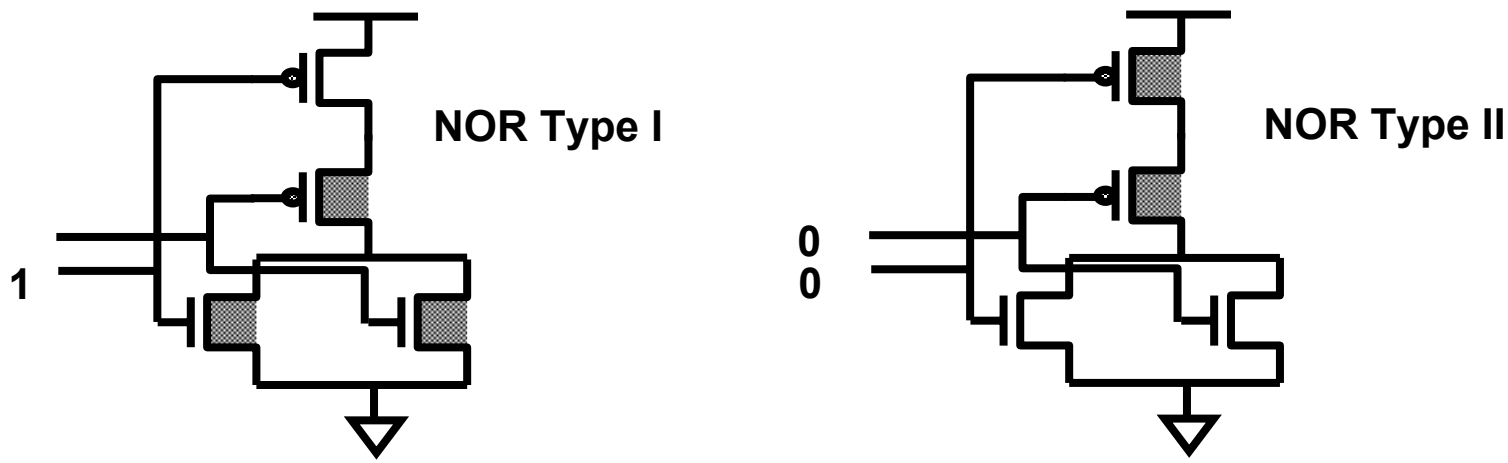
Dynamic Leakage Feedback FF



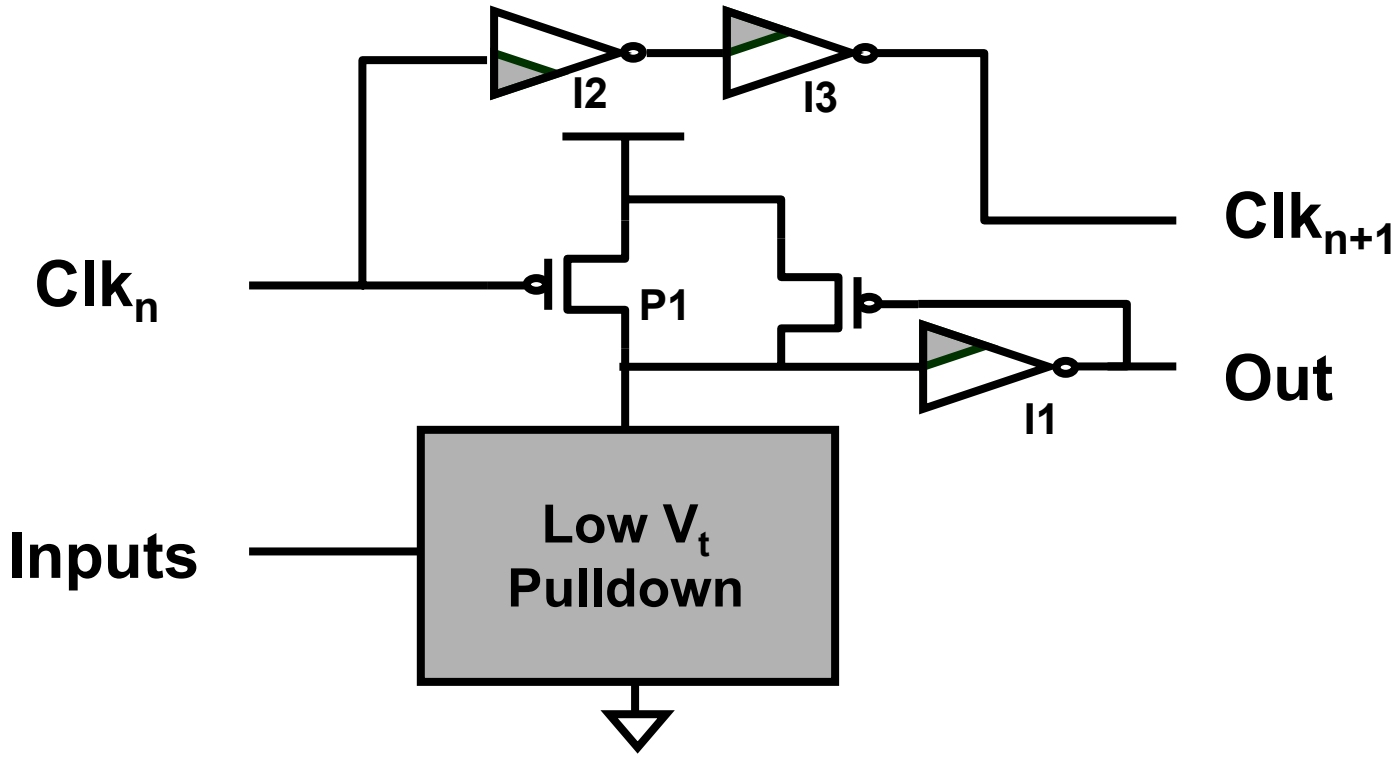
- Operates like standard dynamic FF during active mode
- Retains state during the standby mode (held by leakage)

Imbedded Dual V_t

- Logic gates have internal HVT and LVT devices
- No extra series HVT transistor required
- Suppose special case:
  - known sleep configuration
  - one transition direction more critical



Dual V_t Domino Gate

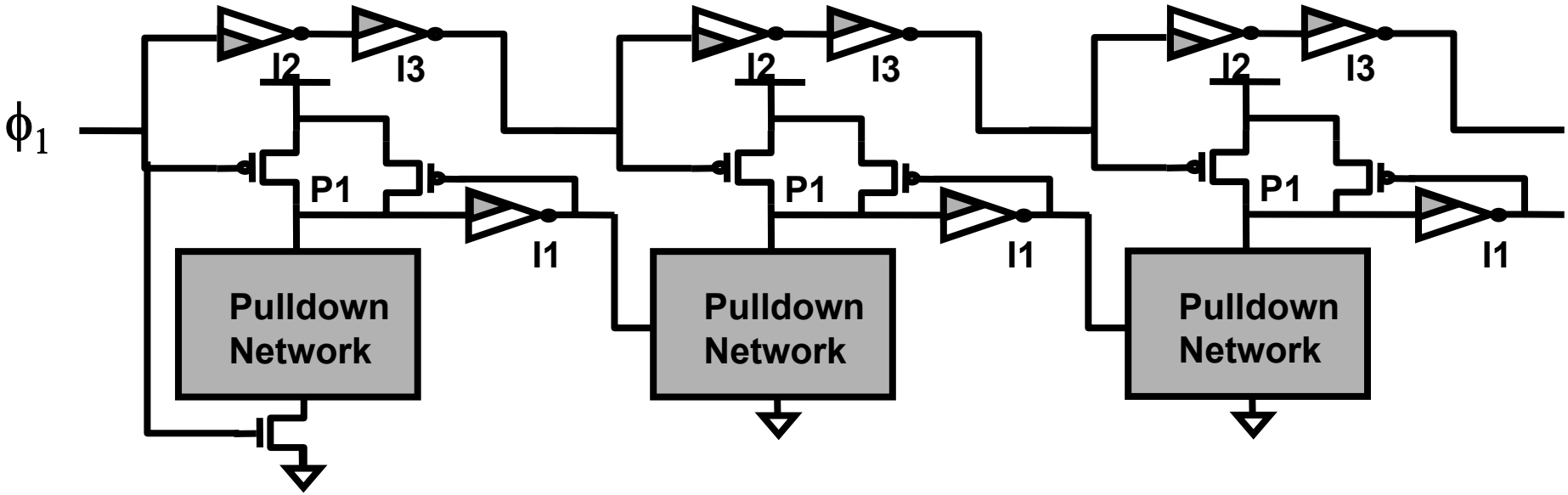


- Evaluate through LVT devices
- Precharge through HVT devices

J. Kao, "Dual Threshold Voltage Domino Logic," ESSCIRC September 1999.

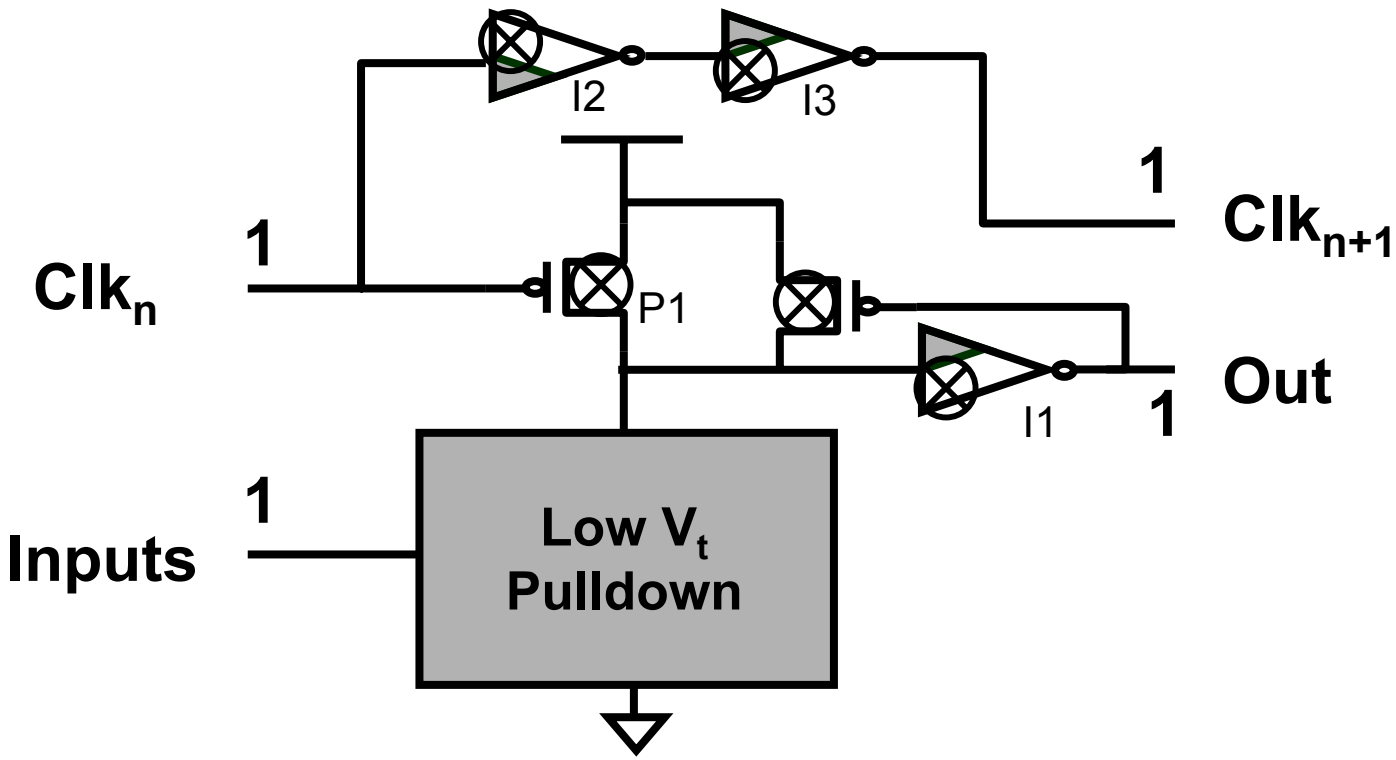
Clock Delayed Domino Logic

(ϕ_1 Pipeline Stage)



- Clock path matches evaluate path
- NMOS series transistors can be eliminated

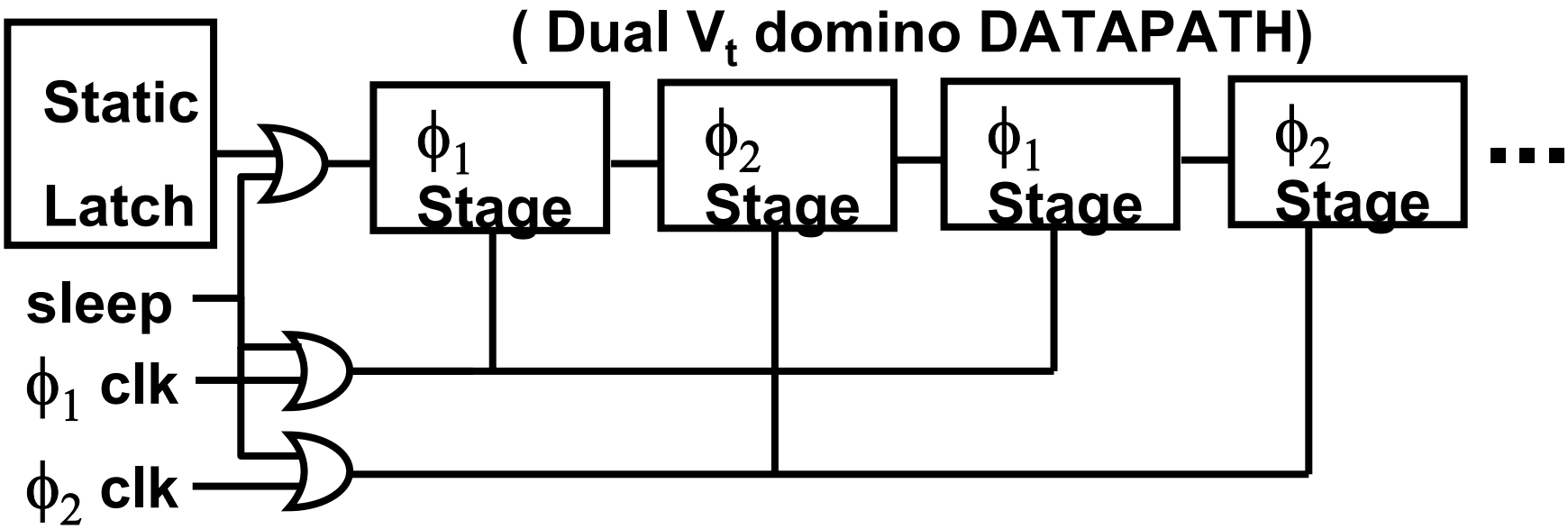
Leakage In Dual V_t Domino Gate



Sleep condition during evaluate mode

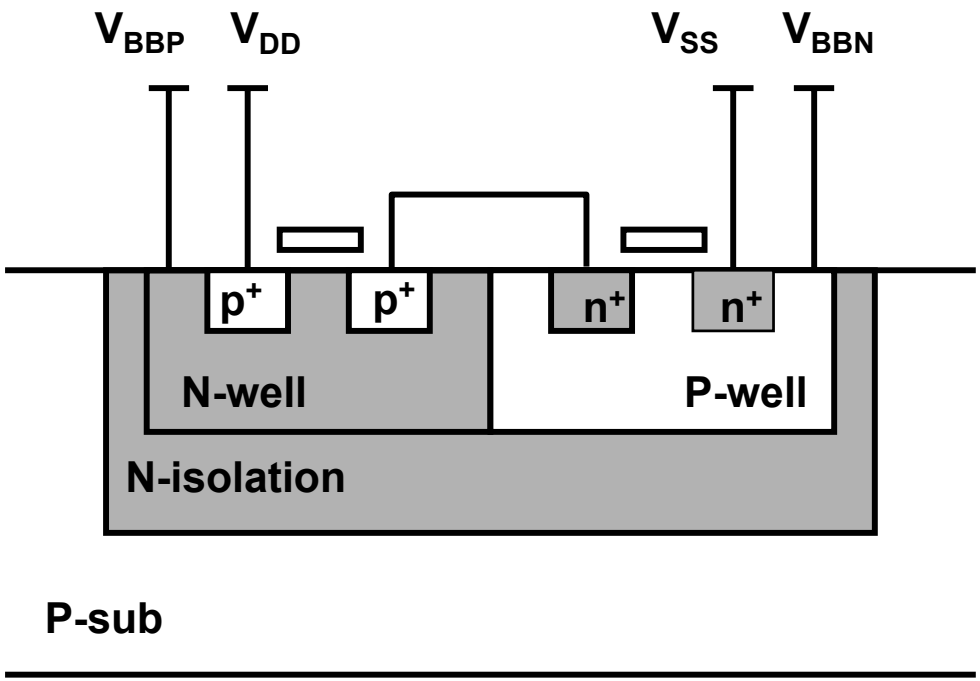
Sleep Mode Condition

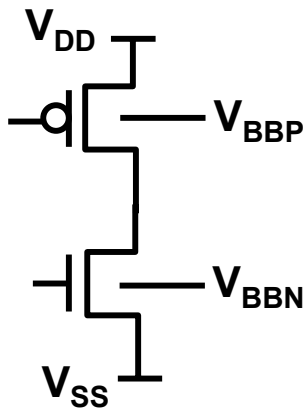
- Clear pipeline before sleep
- Put gates in EVALUATION mode
- Inputs to all gates must be high



Variable Threshold CMOS (VTCMOS)

- Body effect to change device V_t
- Standby leakage reduction with maximum reverse bias
- Triple well structure





Body Effect:

$$V_t = V_{t0} + \gamma\left(\sqrt{2\varphi_B - V_{BB}} - \sqrt{2\varphi_B}\right)$$

VTCMOS Example

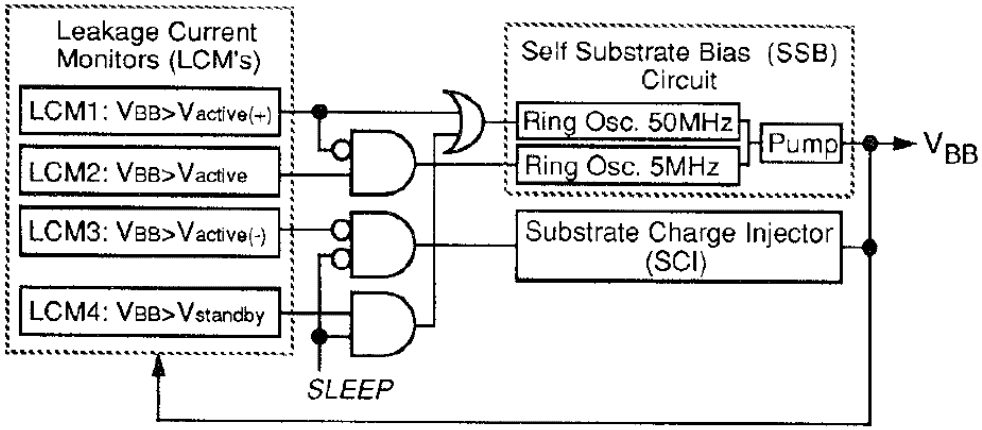


Fig. 3. VT block diagram.

T. Kuroda, et al, "A 0.9V, 150Mhz, 10mW, 4mm^2, 2-DCT Core Processor with Variable V_t Scheme, " JSSC Nov. 1996

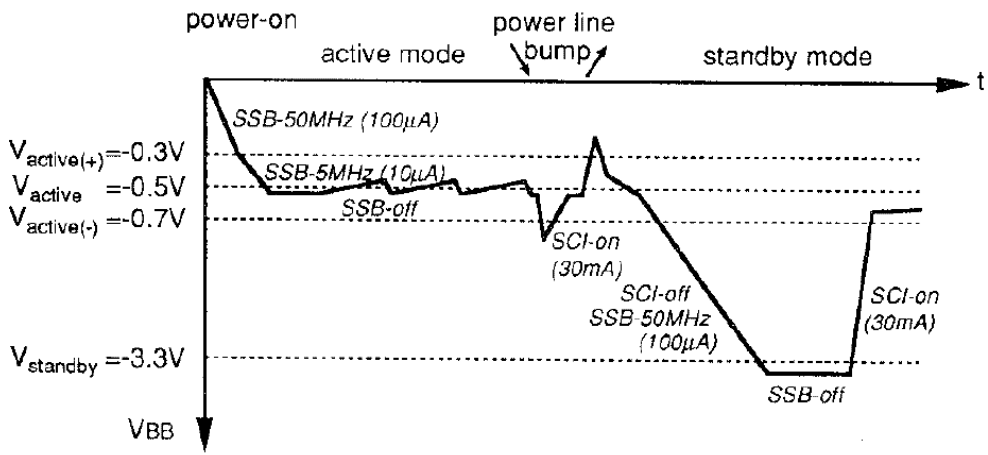


Fig. 4. Substrate-bias control in VT.

- VTCMOS principle applied to 4-mm^2 DCT core processor
- SSB increases V_t (more reverse bias)
- SCI decreases V_t (Standby -> Sleep)
- Leakage reduction
  - 0.1mA active -> 10nA sleep (2.8v ΔVBB)
  - 4 orders of magnitude
- Dynamically tunes V_t (by matching leakage current monitor) to minimize V_t variation

VTCMOS Pros/Cons

PROS:

- Significant standby leakage reduction
- Memory elements retain state
- No transistor sizing/ partitioning required
- **Dynamically tunable V_t during runtime**

CONS:

- Requires expensive triple well process
- Body factor decreases with scaling

Speed Adaptive V_t CMOS

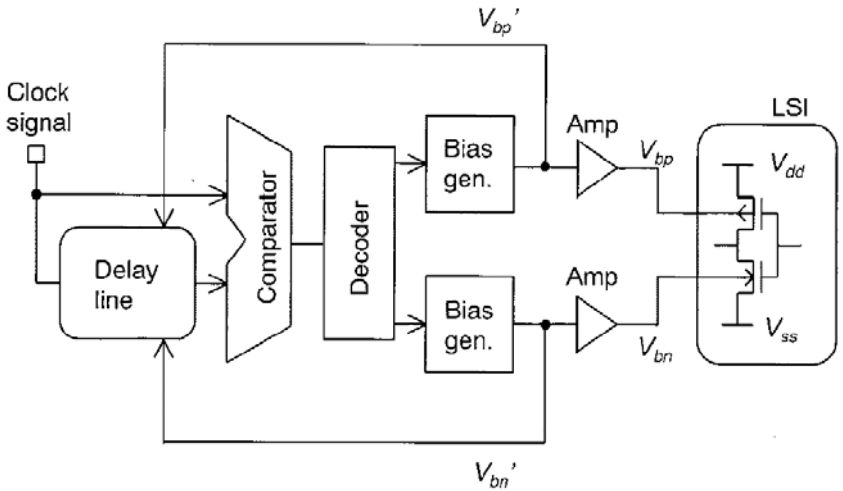
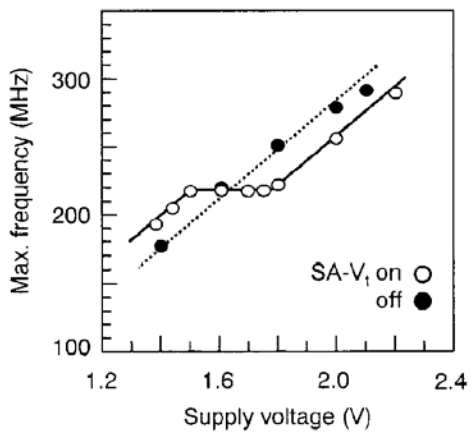


Fig. 2. Concept of SA-V_t CMOS scheme.

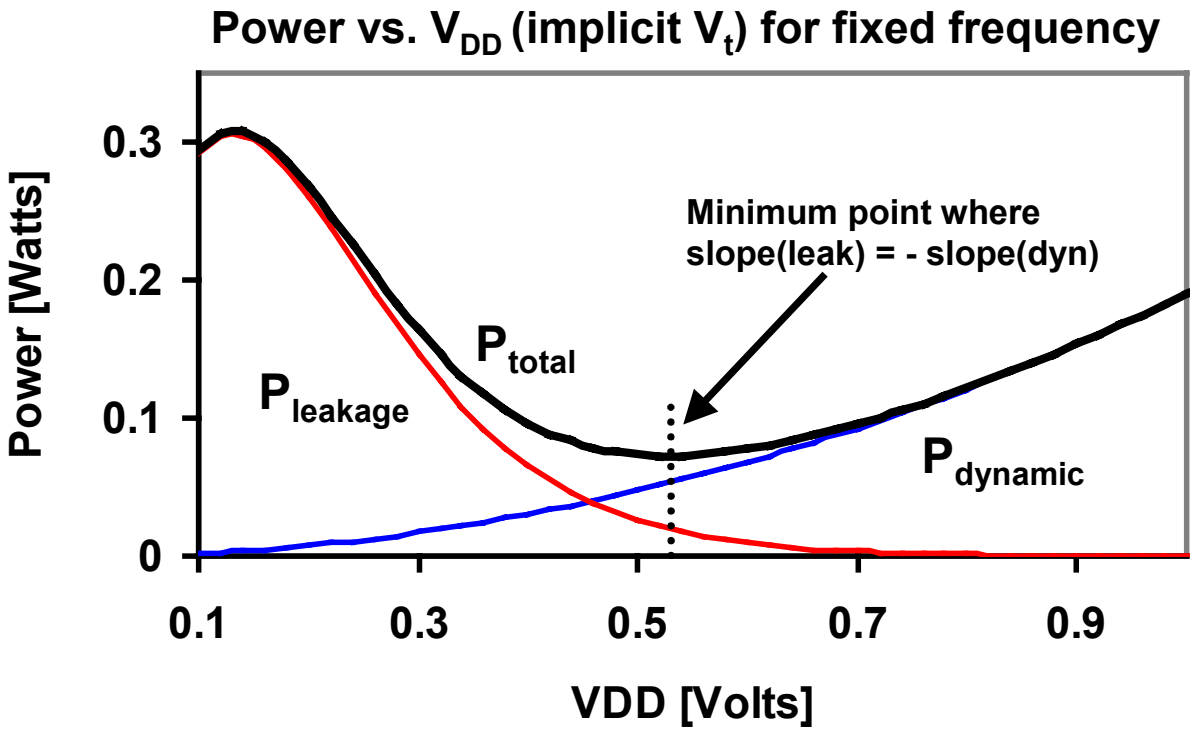


M. Miyazaki, et al, "A 1.2-GIPS/W uProc Using Speed-Adapative V_t CMOS with Forward Bias," JSSC Feb 2002.

- **Dynamically tune V_t so that critical path speed matched clock period**
- **Reduces chip-to-chip parameter variations**
- **Reverse bias:**
  - **Operate only as fast as necessary (reduces excess active leakage)**
- **Forward bias:**
  - **Speeds up slow chips**
- **Standby leakage with maximum reverse bias**
- **Also known as Adaptive Body Biasing (ABB)**

Adaptive Supply & Body Bias (ASB)

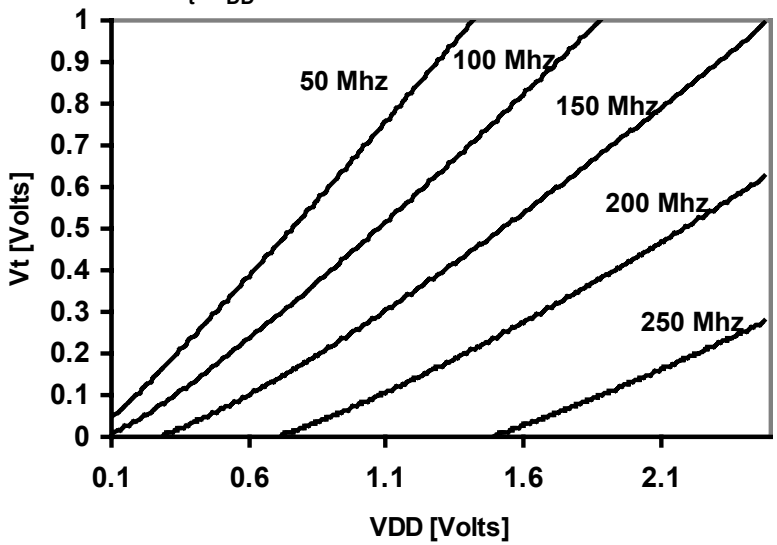
- Dynamically tune both V_{DD} & V_t as operating conditions change
- Trade-off between dynamic power (V_{DD} knob), leakage power (V_t)
- Minimize total ACTIVE power consumption
  (higher active leakage current at expense of lowering dynamic power)

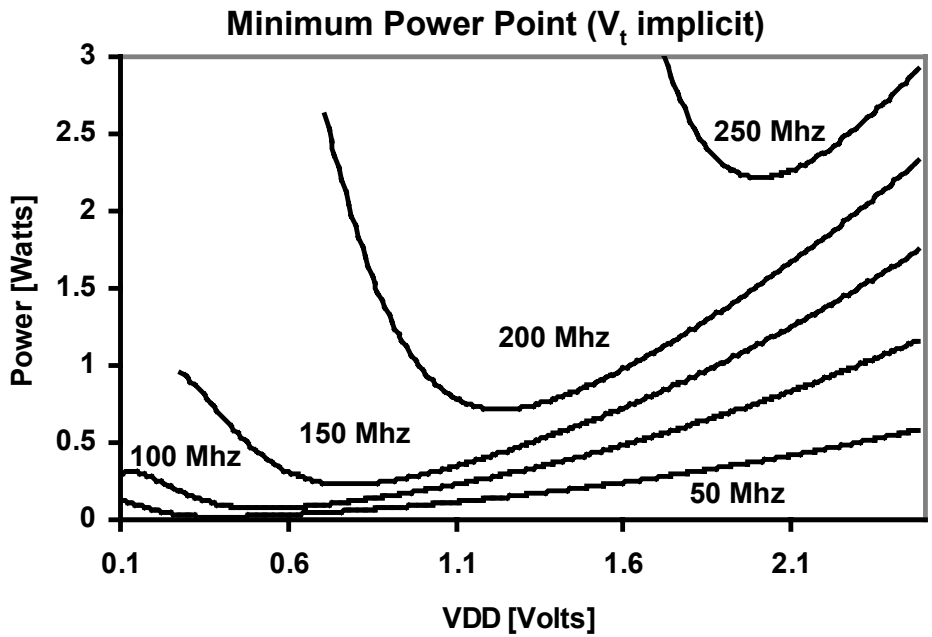


M. Miyazaki, et al, "A 175mV Multiply-Accumulate Unit using an Adaptive Supply Voltage and Body Bias (ASB) Architecture," ISSCC February 2002.

Optimal V_{DD}/V_T Selection

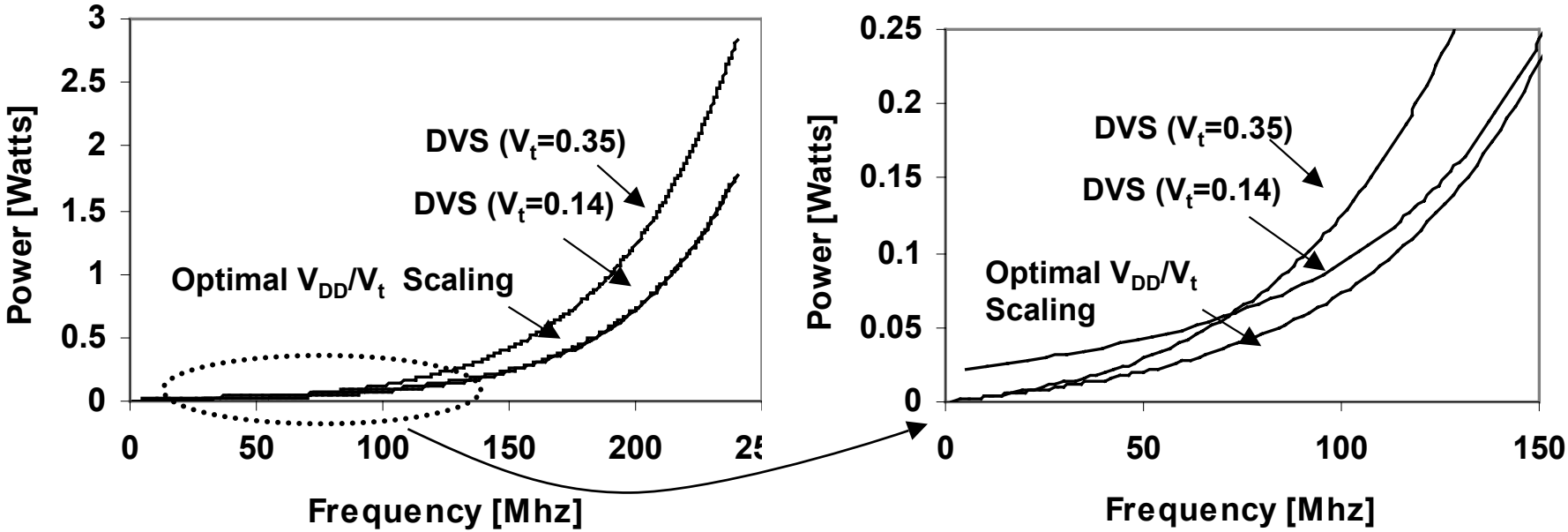
- Optimal V_{DD} & V_t target changes with operating conditions
  - e.g. Varying Workload
- Low frequencies high V_t more optimal
  - reduce leakage at expense of increased dynamic
- High Frequencies low V_{DD} more optimal
  - reduce dynamic at expense of increased leakage



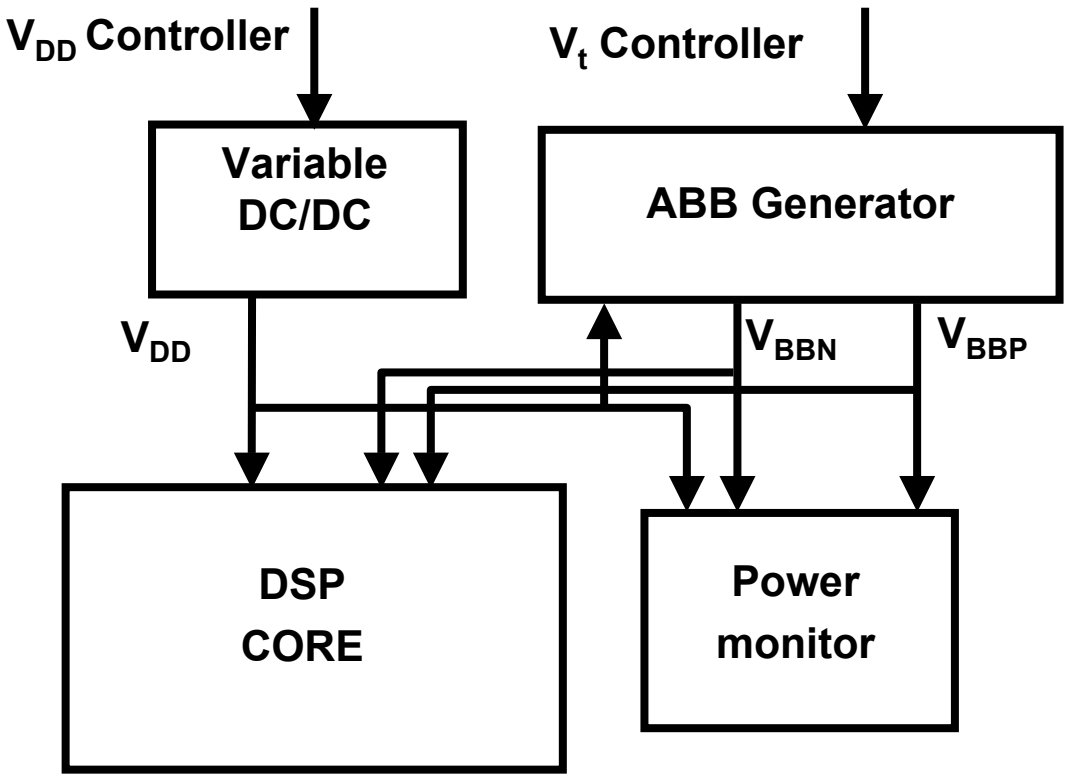


V_{DD}/V_T Optimization vs. DVS

- Dynamic voltage scaling ignores V_T influence
- DVS is sub-optimal over the frequency range



ASB Architecture



- Decouple V_{DD}/ V_t tuning loops
- ABB (Auto Body Biasing) generator chooses V_t based on V_{DD}/ Freq/ etc.
- Simple V_{DD} sweep to search minimum active power point
- Architecture ensures minimum power for any operating condition

M. Miyazaki, J. Kao, A. Chandrakasan, "A 175mV Multiply-Accumulate Unit using an Adaptive Supply Voltage and Body Bias (ASB) Architecture," ISSCC February 2002.

Summary

- Subthreshold leakage currents will grow exponentially
- Need to manage during STANDBY and ACTIVE
- Three main principles
  - Source Biasing
  - Multiple threshold voltage
  - Body biasing
- Need for CAD tools to model leakage currents
- Need for CAD tools to implement leakage reduction principles