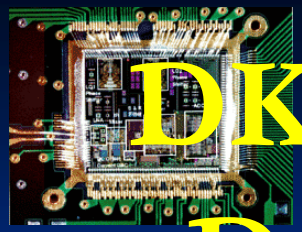
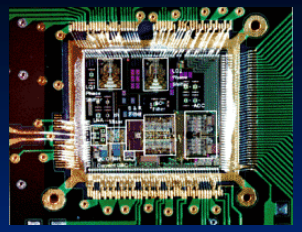


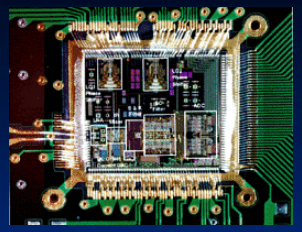
# DKDT: A Performance Aware Dual Dielectric Assignment for Tunneling Current Reduction

**Saraju P. Mohanty**
Dept of Computer Science and Engineering
University of North Texas
smohanty@cs.unt.edu
http://www.cs.unt.edu/~smohanty/



# Outline of the Talk

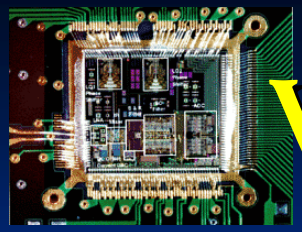
- Introduction
- Why Dual-K and Dual-T
- Related Work
- DKDT Assignment Algorithm
- Cell Characterization for DKDT
- Conclusions



# Why Low-Power ?

**Motivation**: Extending battery life ……


Source: Power Integrations Inc



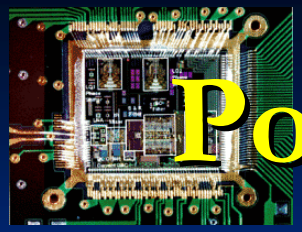
# Why Low-Power ? ……..

## Battery Lifetime

## Cooling and Energy Costs

## Environmental Concerns

## System Reliability



# Power Dissipation in CMOS

## Total Power Dissipation

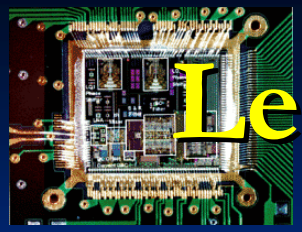
### Static Dissipation

- Sub-threshold current
- **Tunneling current**
- Reverse-biased diode Leakage
- Contention current

### Dynamic Dissipation

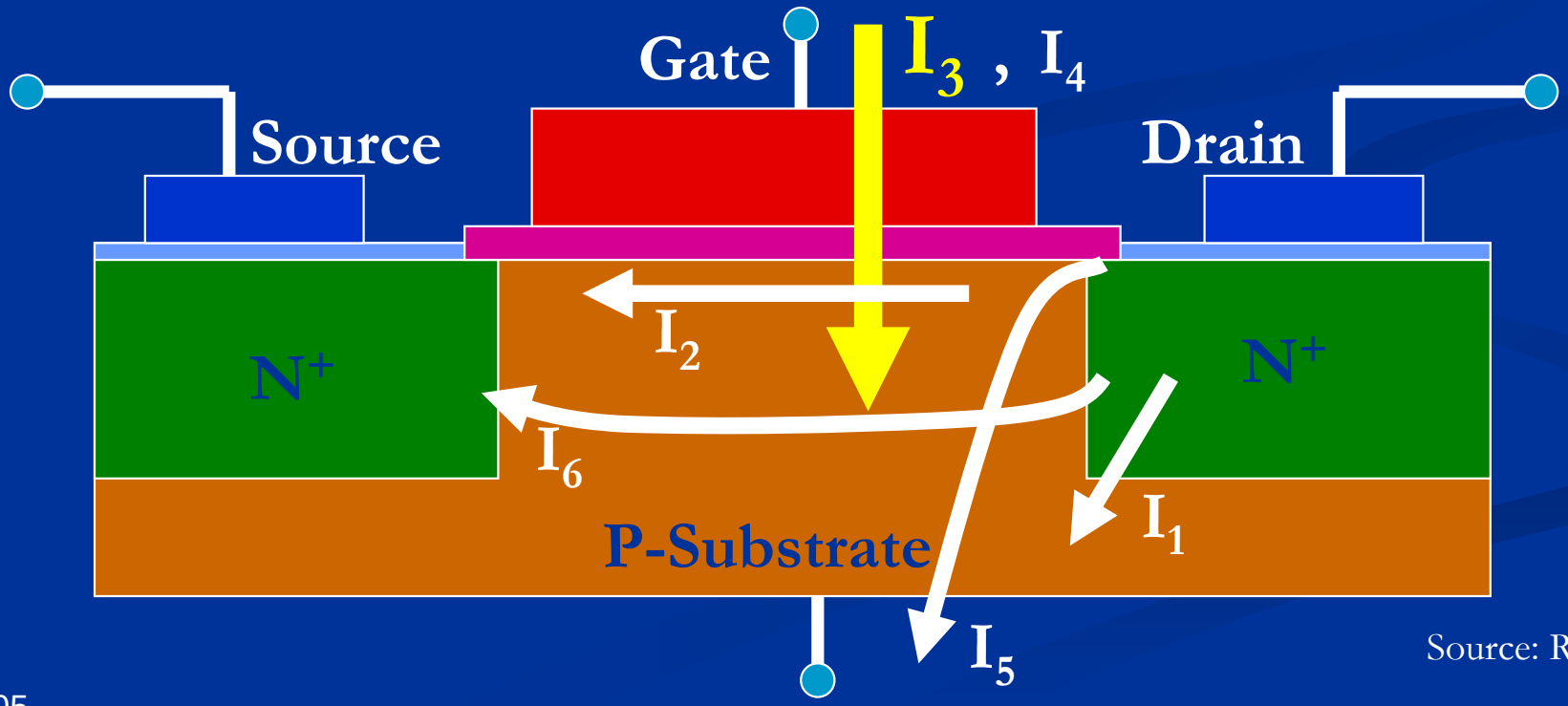
- Capacitive Switching
- Short circuit

Source: Weste and Harris 2005

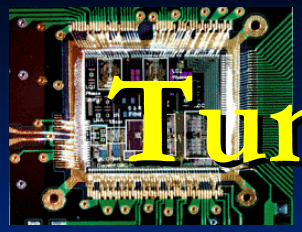


# Leakages in Nanometer CMOS

$I_1$ : reverse bias pn junction (both ON & OFF)
$I_2$ : subthreshold leakage (OFF )
**$I_3$ : oxide tunneling current (both ON & OFF)**
$I_4$ : gate current due to hot carrier injection (both ON & OFF)
$I_5$ : gate induced drain leakage (OFF)
$I_6$ : channel punch through current (OFF)

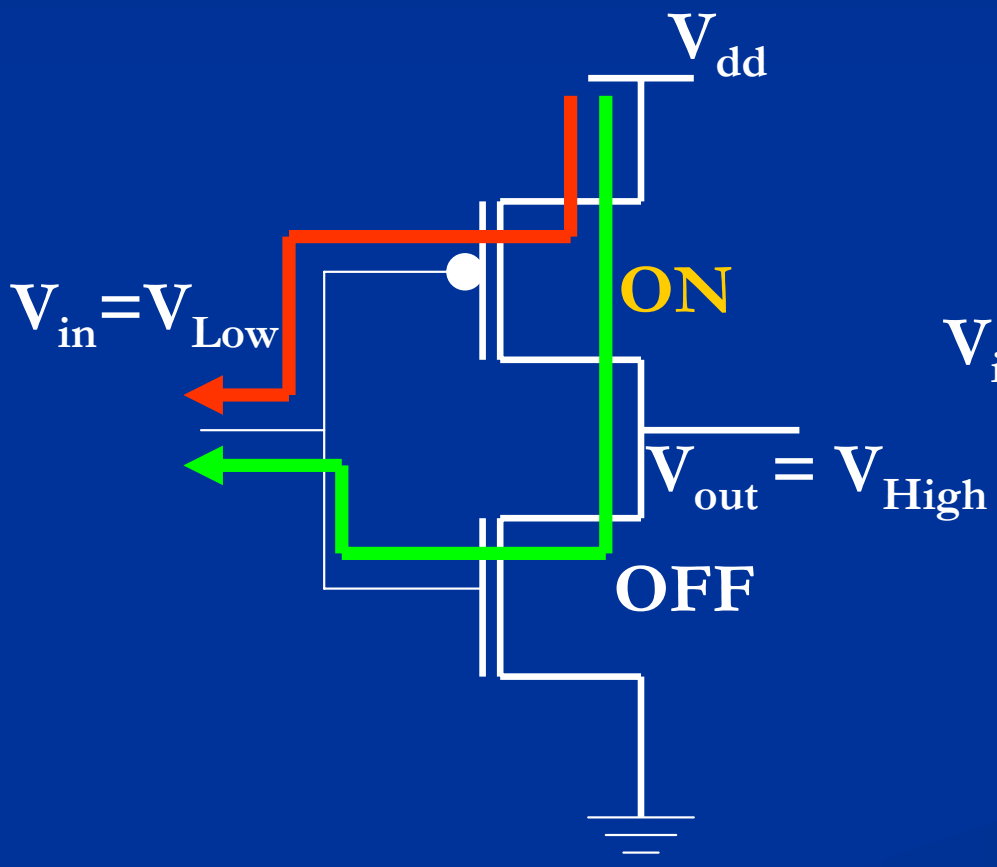


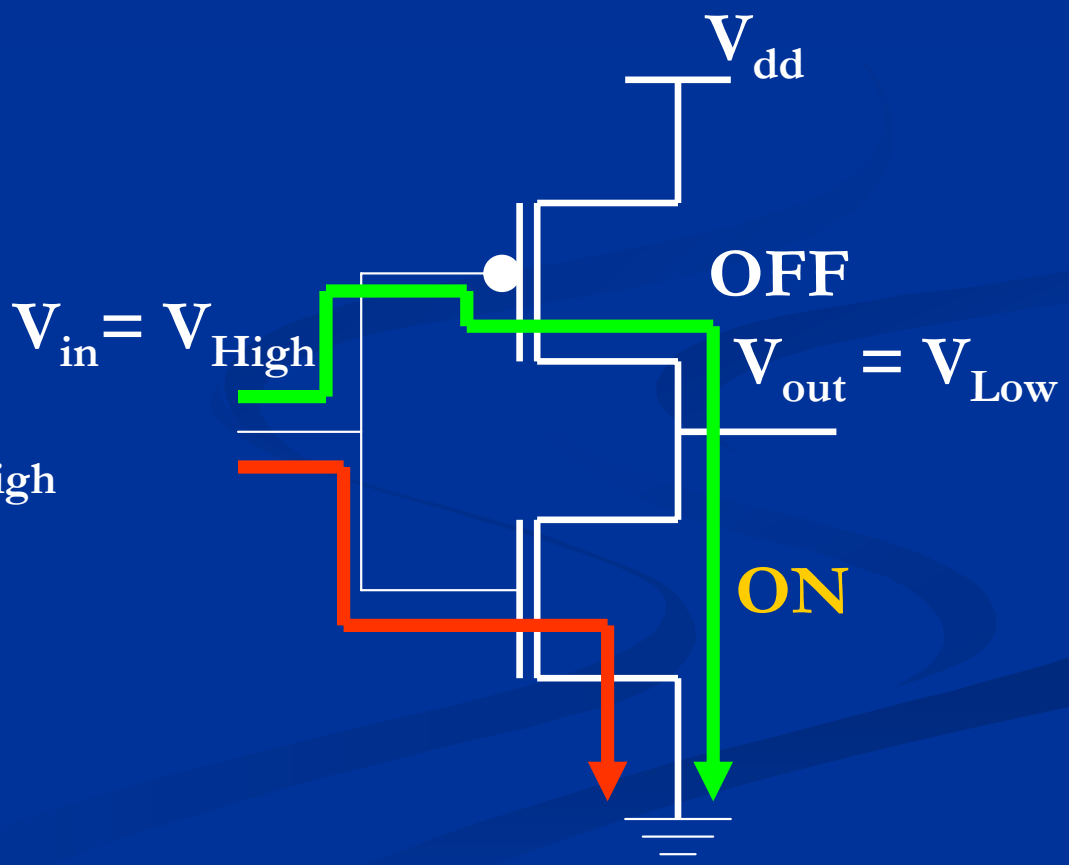

Source: Roy 2003

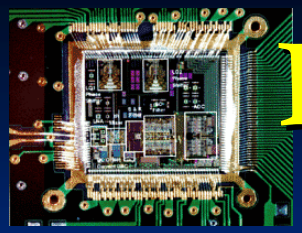


# Tunneling paths in an Inverter

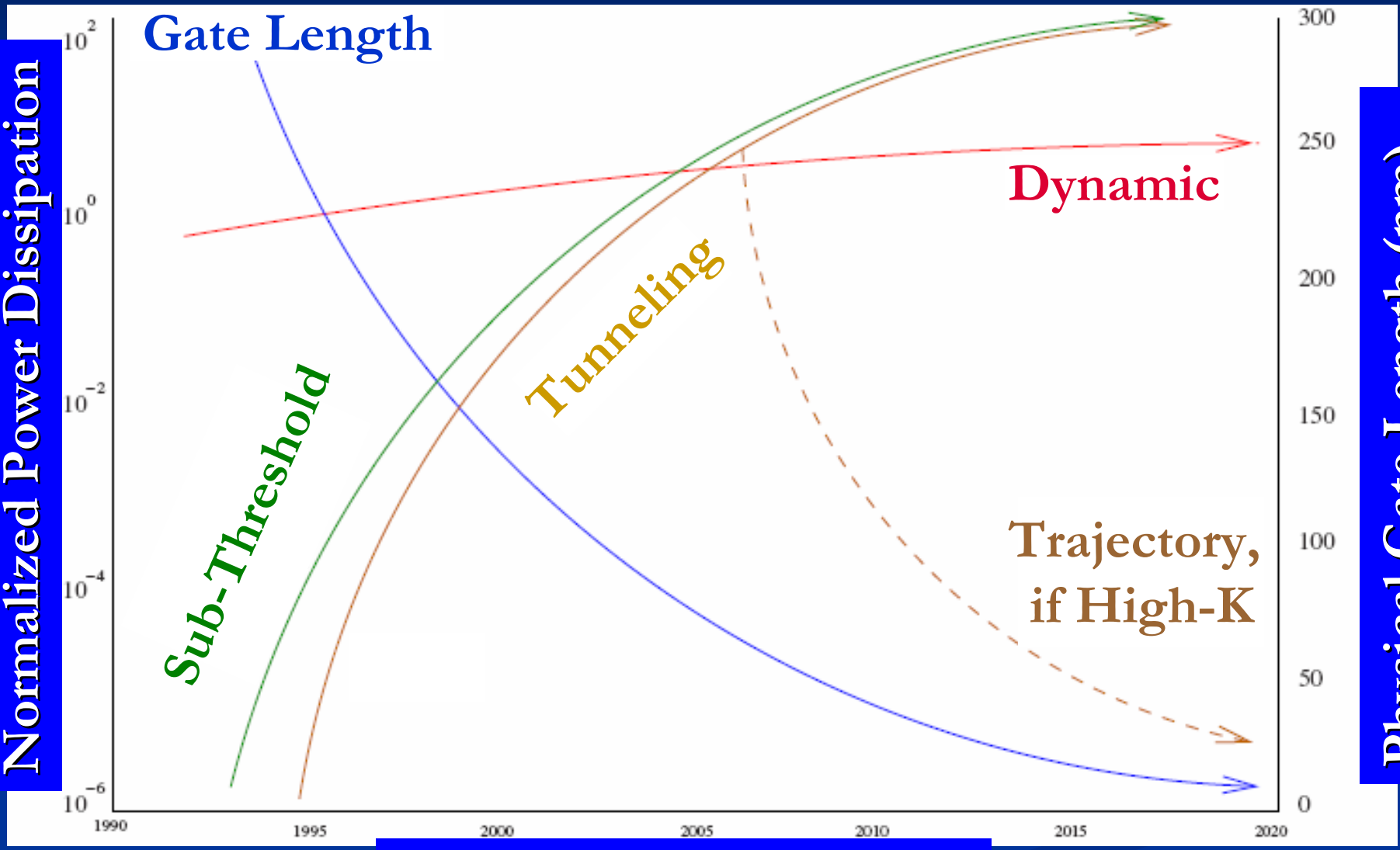
- **Low Input** : Input supply feeds the tunneling current.
- **High Input** : Gate supply feeds the tunneling current.



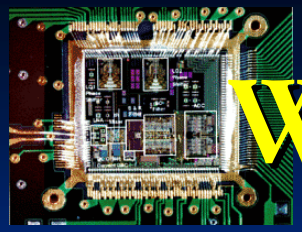





# Power Dissipation Trend




Source: Hansen 2004

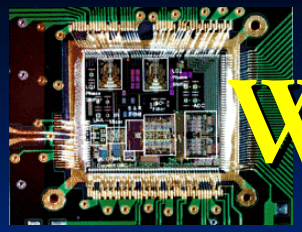


# Why Dual-K and Dual-T ?

- Gate oxide tunneling current $I_{gate}$ [Kim2003, Chandrakasan2001] (k is a experimentally derived factors):

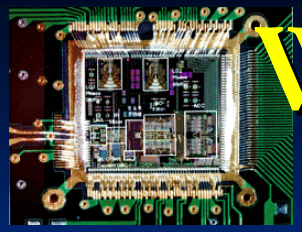
$$I_{gate} \propto (V_{dd} / T_{gate})^2 \exp(-k\, T_{gate} / V_{dd})$$

- **Options for reduction of tunneling current :**
  - Decreasing of supply voltage $V_{dd}$ (will play its role)
  - Increasing gate $SiO_2$ thickness $T_{gate}$ (opposed to the technology trend !!)

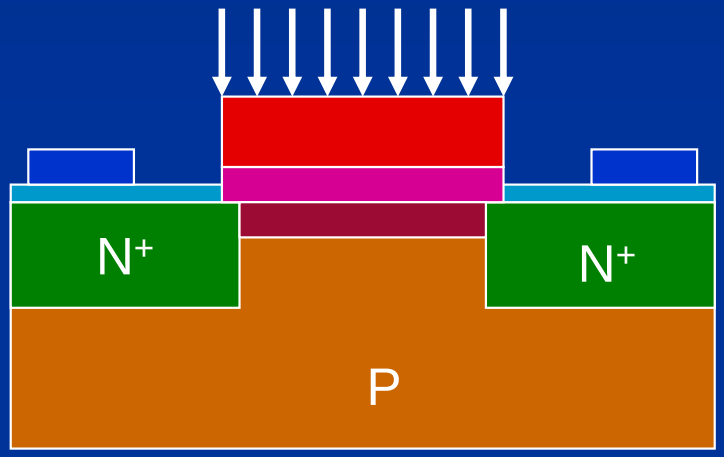


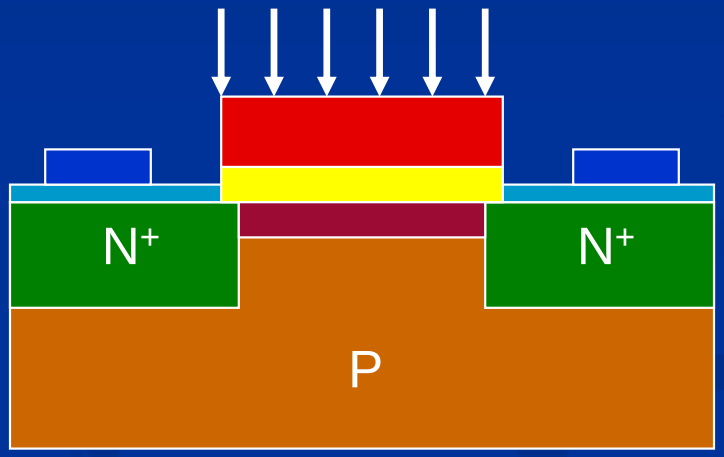
# Why Dual-K and Dual-T ?

We believe that use of multiple dielectrics (denoted as $K_{gate}$) of multiple thickness (denoted as $T_{gate}$) will reduce the gate tunneling current significantly while maintaining the performance.



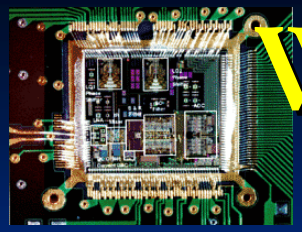
# Why Dual-K and Dual-T ? (Low $K_{gate}$ Vs High $K_{gate}$)



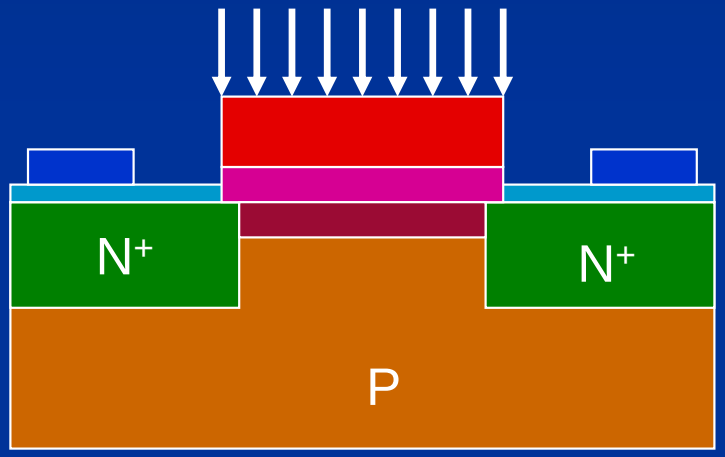



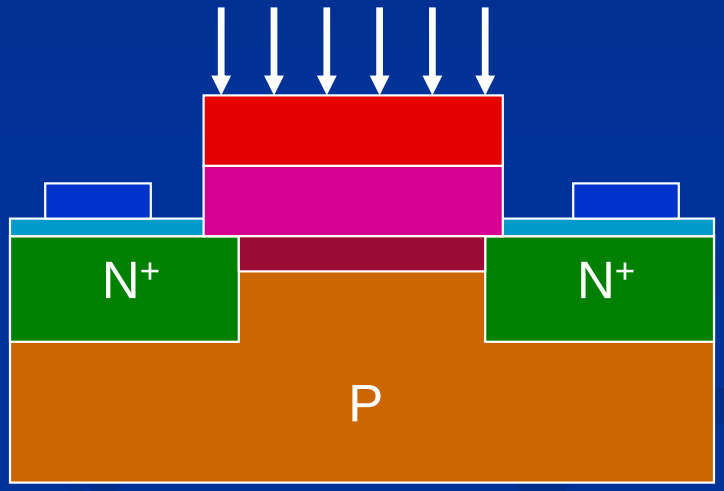

Low $K_{gate}$ → Larger $I_{gate}$, Smaller delay

High $K_{gate}$ → Smaller $I_{gate}$, Larger delay



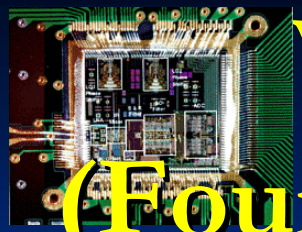
# Why Dual-K and Dual-T ? (Low $T_{gate}$ Vs High $T_{gate}$)



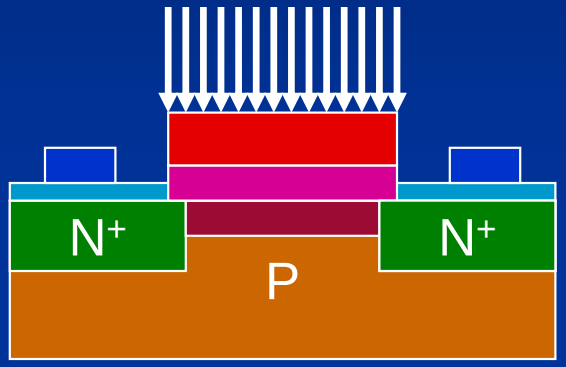




Low $T_{gate}$ → Larger $I_{gate}$, Smaller delay

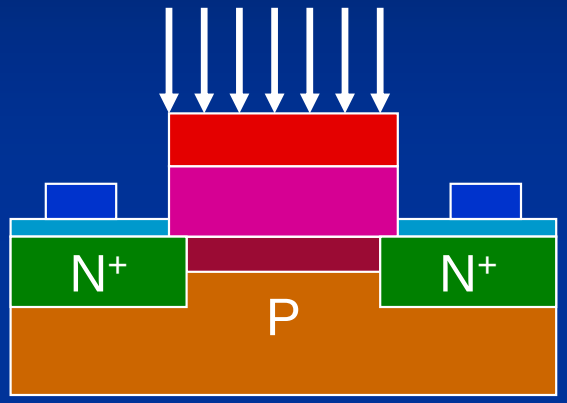
High $T_{gate}$ → Smaller $I_{gate}$, Larger delay



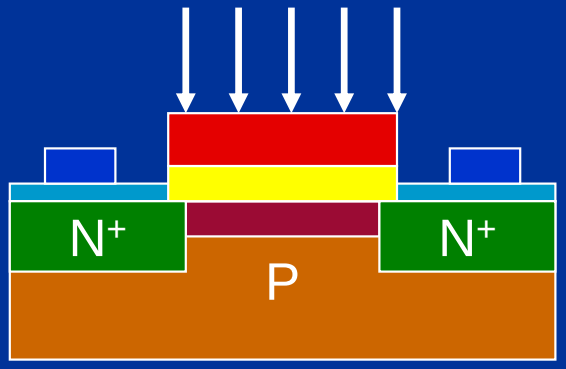
# Why Dual-K and Dual-T ? (Four Combinations of $K_{gate}$ & $T_{gate}$)



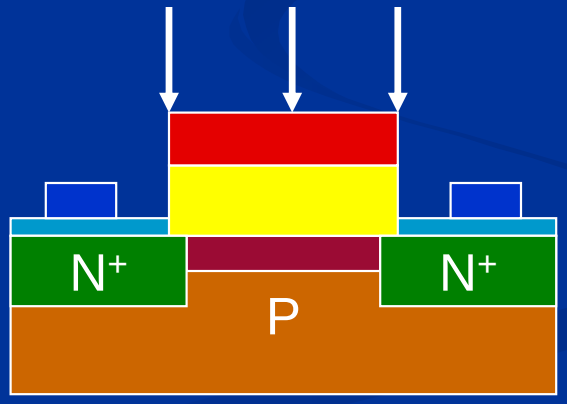

(1) $K_1T_1$




(2) $K_1T_2$



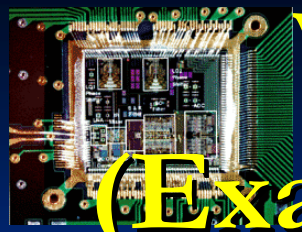

(3) $K_2T_1$




(4) $K_2T_2$

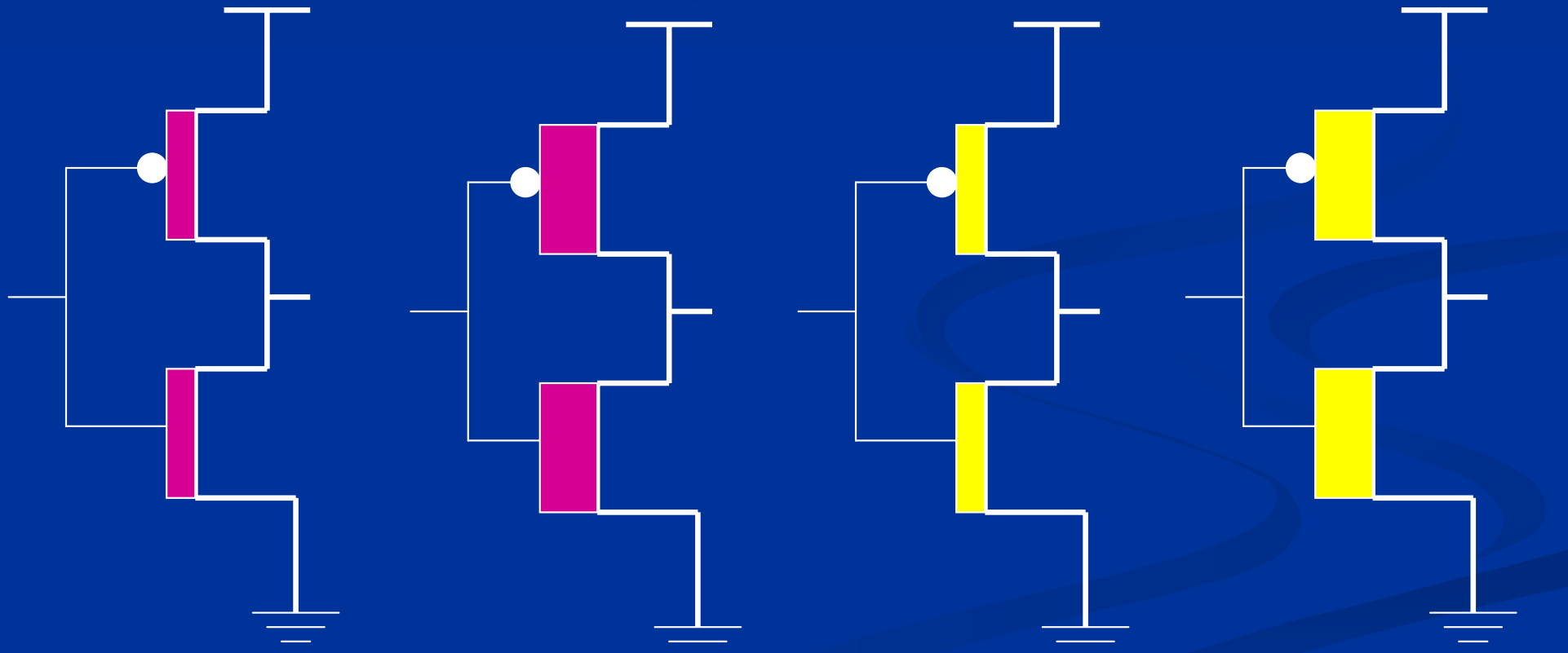
Tunneling Current ↓

Delay ↑

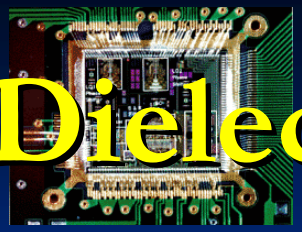


# Why Dual-K and Dual-T ? (Example: Four Types of Inverter)

**Assumption**: all transistors of a logic gate are of same $K_{gate}$ and equal $T_{gate}$.

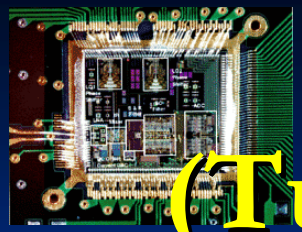


(1) $K_1T_1$ (2) $K_1T_2$ (3) $K_2T_1$ (4) $K_2T_2$



# Dielectrics for Replacement of $SiO_2$

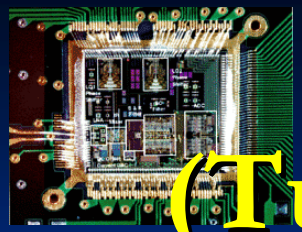
- **Silicon Oxynitride ($SiO_xN_y$) (K=5.7 for SiON)**
- Silicon Nitride ($Si_3N_4$) (K=7)
- Oxides of :
  - Aluminum (Al), Titanium (Ti), Zirconium (Zr), Hafnium (Hf), Lanthanum (La), Yttrium (Y), Praseodymium (Pr),
  - their mixed oxides with $SiO_2$ and $Al_2O_3$
- **NOTE**: $I_{gate}$ is still dependent on $T_{gate}$ irrespective of dielectric material.



# Related Works (Tunneling Current Reduction)

- **Inukai et. al. in CICC2000:** Boosted Gate MOS (BGMOS) device using dual $T_{ox}$ and dual $V_{Th}$ for both gate and subthreshold standby leakage reduction.

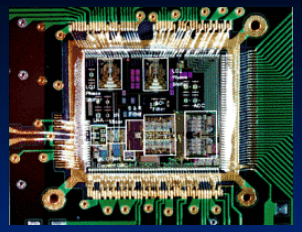
- **Rao et. al. in ESSCIRC2003:** Sleep state assignment for MTCMOS circuits for reduction of both gate and subthreshold leakage.



# Related Works (Tunneling Current Reduction)

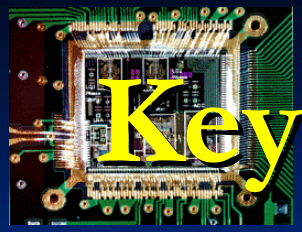
- **Lee et. al. in DAC2003 and TVLSI2004Feb :** Pin reordering to minimize gate leakage during standby positions of NOR and NAND gates.

- **Sultania, et. al. in DAC2004 and ICCD2004:** Heuristic for dual $T_{ox}$ assignment for tunneling current and delay tradeoff.



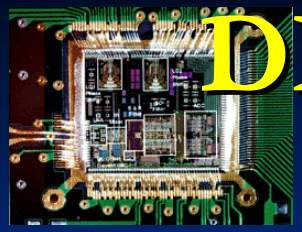
# Related Works

- Developed methods that use oxide of different thicknesses for tunneling reduction.
- Do not handle emerging dielectrics that will replace $SiO_2$ to reduce the tunneling current.
- Either consider ON or OFF state, but do not account both.
- Degradation in performance due to dual thickness approach.



# Key Contributions of this Work

- Introduces a new approach called dual dielectric assignment for tunneling current reduction.
- Considers dual thickness approach for both of the dielectrics.
- Explores a combined approach called DKDT (Dual-K of Dual Thickness) and proposes an assignment algorithm.
- Accounts the tunneling current for both ON and OFF state.
- Presents a methodology for logic gates characterization for worst-case tunneling considering **non-$SiO_2$** dielectrics for low end nano-technology.



# DKDT Based Logic Synthesis

Input Circuit

Technology Independent Optimization

Intermediate Circuit

Technology Mapping

Mapped Circuit

DKDT Assignment and Optimization

Tunneling Optimized Circuit

Placement and Routing

Placement Legalization and ECO Routing

Final Circuit Layout

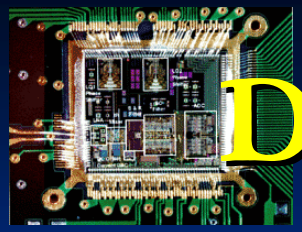
Cell Library (4 Types)

$K_1T_1$

$K_1T_2$

$K_2T_1$

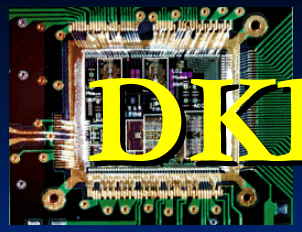
$K_2T_2$



# DKDT Assignment : Basis

- **Observation**: Tunneling current of logic gates increases and propagation delay decreases in the order $K_2T_2$, $K_2T_1$, $K_1T_2$, and $K_1T_1$ (where, $K_1 < K_2$ and $T_1 < T_2$).

- **Strategy**: Assign a higher order K and T to a logic gate under consideration
  - To reduce tunneling current
  - Provided increase in path-delay does not violate the target delay



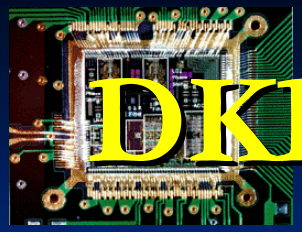
# DKDT Assignment : Algorithm

**Step 1:** Represent the network as a directed acyclic graph G(V, E).

**Step 2:** Initialize each vertex $v \in G(V, E)$ with the values of tunneling current and delay for $K_1T_1$ assignment.

**Step 3:** Find the set of all paths $P\{\Pi_{in}\}$ for all vertex in the set of primary inputs $(\Pi_{in})$, leading to the primary outputs $\Pi_{out}$ .

**Step 4:** Compute the delay $D_P$ for each path $p \in P\{\Pi_{in}\}$.



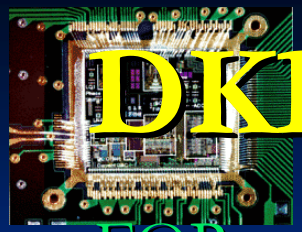
# DKDT Assignment : Algorithm

**Step 5:** Find the critical path delay $D_{CP}$ for $K_1T_1$ assignment.

**Step 6:** Mark the critical path(s) $P_{CP}$, where $P_{CP}$ is subset $P\{\Pi_{in}\}$.

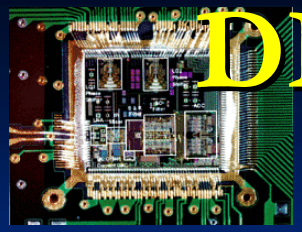
**Step 7:** Assign target delay $D_T = D_{CP}$.

**Step 8:** Traverse each node in the network and attempt to assign K-T in the order $K_2T_2$, $K_2T_1$, $K_1T_2$, and $K_1T_1$ to reduce tunneling while maintaining performance.



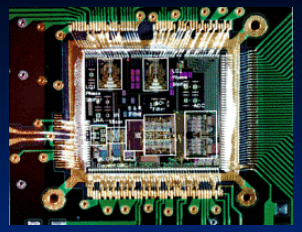
# DKDT Assignment : Algorithm

(1) FOR each vertex $v \in G(V, E)$
(2) {
    (1) Determine all paths $P_v$ to which node $v$ belongs ;
    (2) Assign $K_2T_2$ to $v$ ;
    (3) Calculate new critical delay $D_{CP}$ ;
    (4) Calculate slack in delay as $\Delta D = D_T - D_{CP}$ ;
    (5) IF ( $\Delta D < 0$) then
    (6) {
        (1) Assign $K_2T_1$ to $v$ ; Calculate $D_{CP}$ ; Calculate $\Delta D$ ;
        (2) IF ($\Delta D < 0$) then
        (3) {
            (1) Assign $K_1T_2$ to $v$ ; Calculate $D_{CP}$ ; Calculate $\Delta D$ ;
            (2) IF ($\Delta D < 0$) then
                (1) reassign $K_1T_1$ to $v$ ;
        (4) } // end IF
    (7) } // end IF
(3) // end FOR

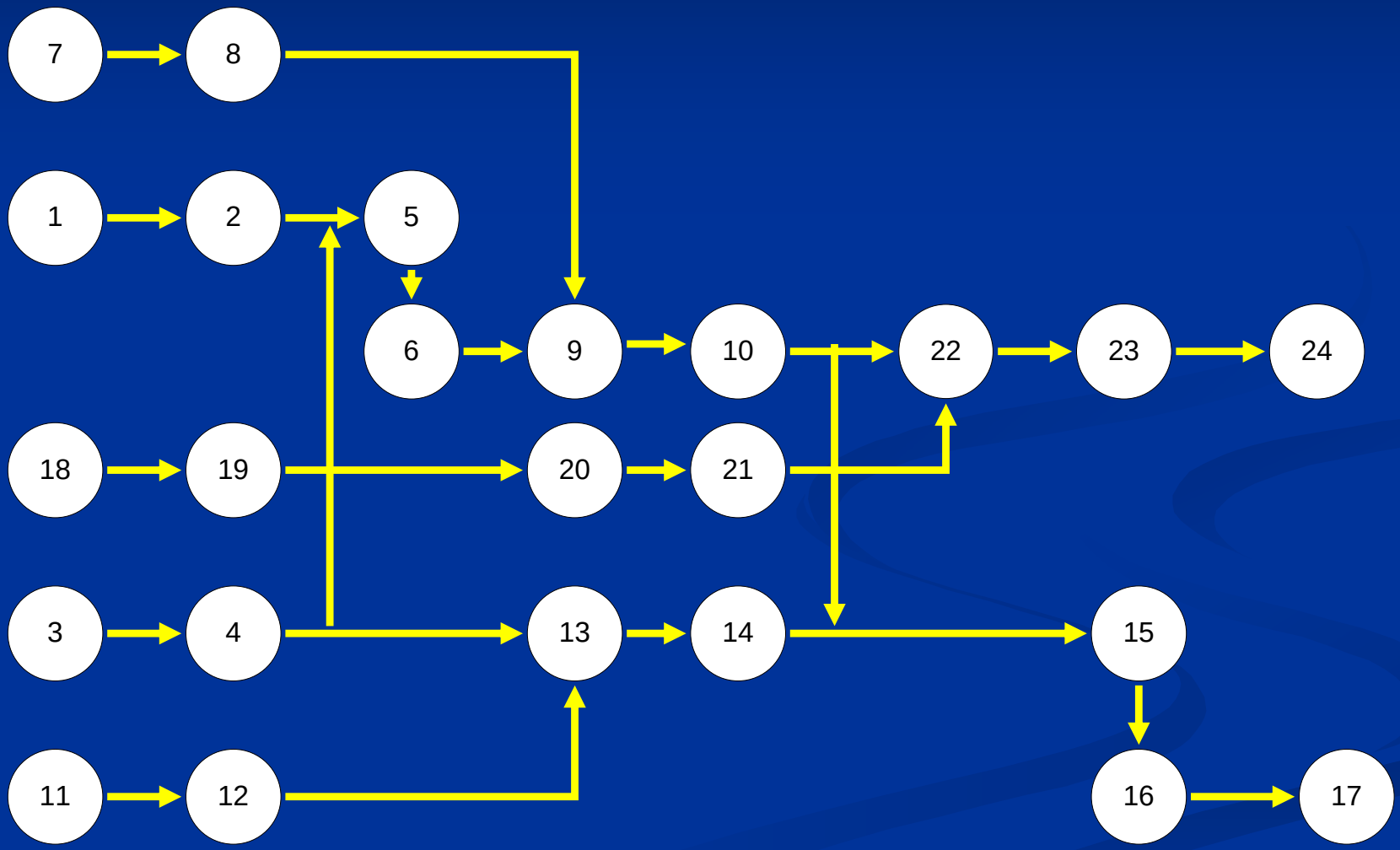


# DKDT Assignment Algorithm (Time Complexity)

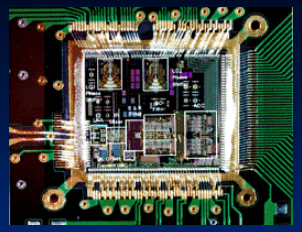
- Assume that there are n number of gates in the original network representing any circuit.
- The statements from Step-01 to Step-03 take $\Theta(n^2)$ time in worst case.
- The run time for statements from Step-04 to Step-07 are of $\Theta(n^2)$ complexity.
- The heuristic loop which assigns DKDT takes $\Theta(n^3)$ time.
- Thus, the overall worst case time complexity of the DKDT assignment algorithm is $\Theta(n^3)$.



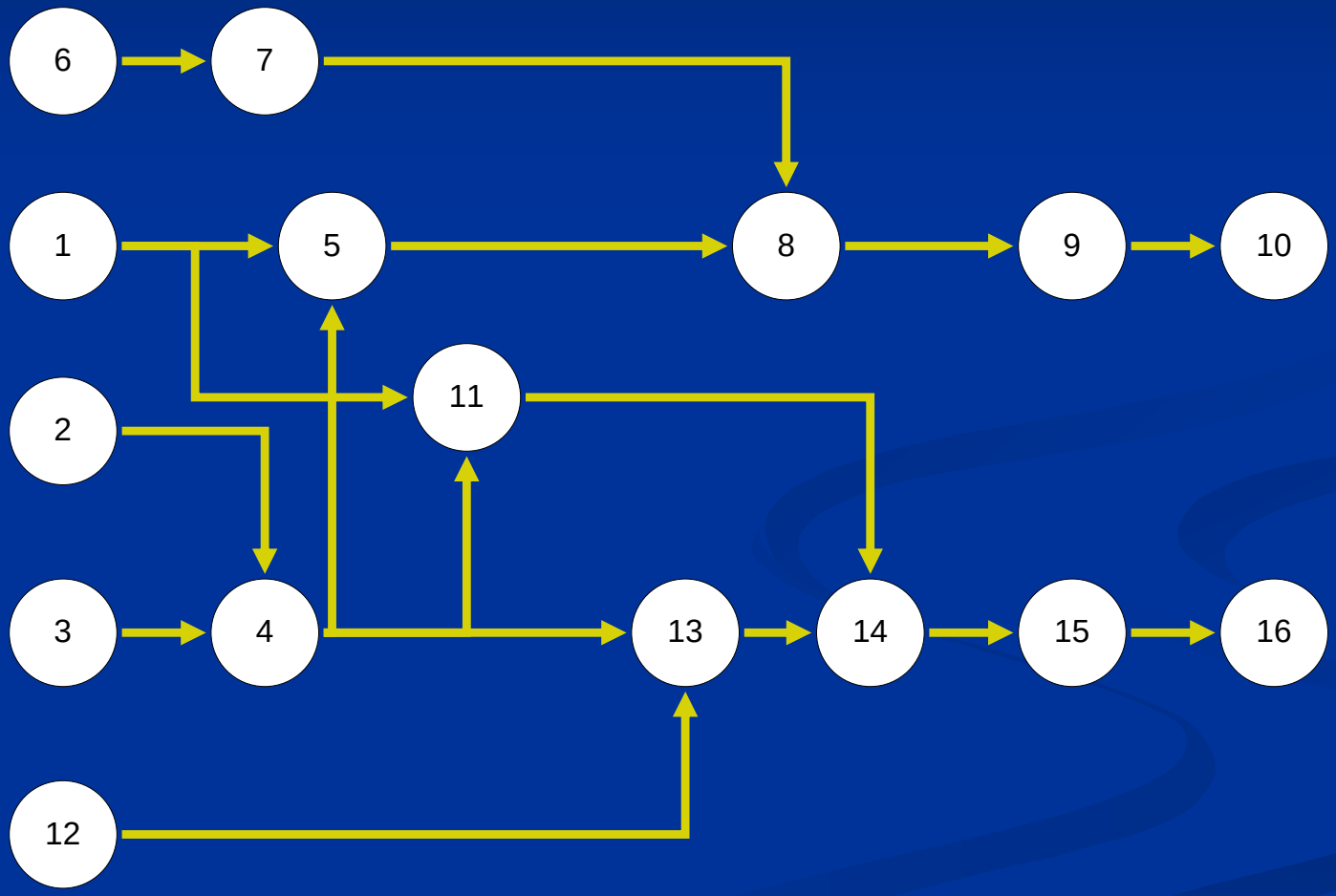
# DKDT Algorithm : Demo



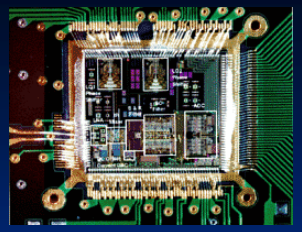

Original Network



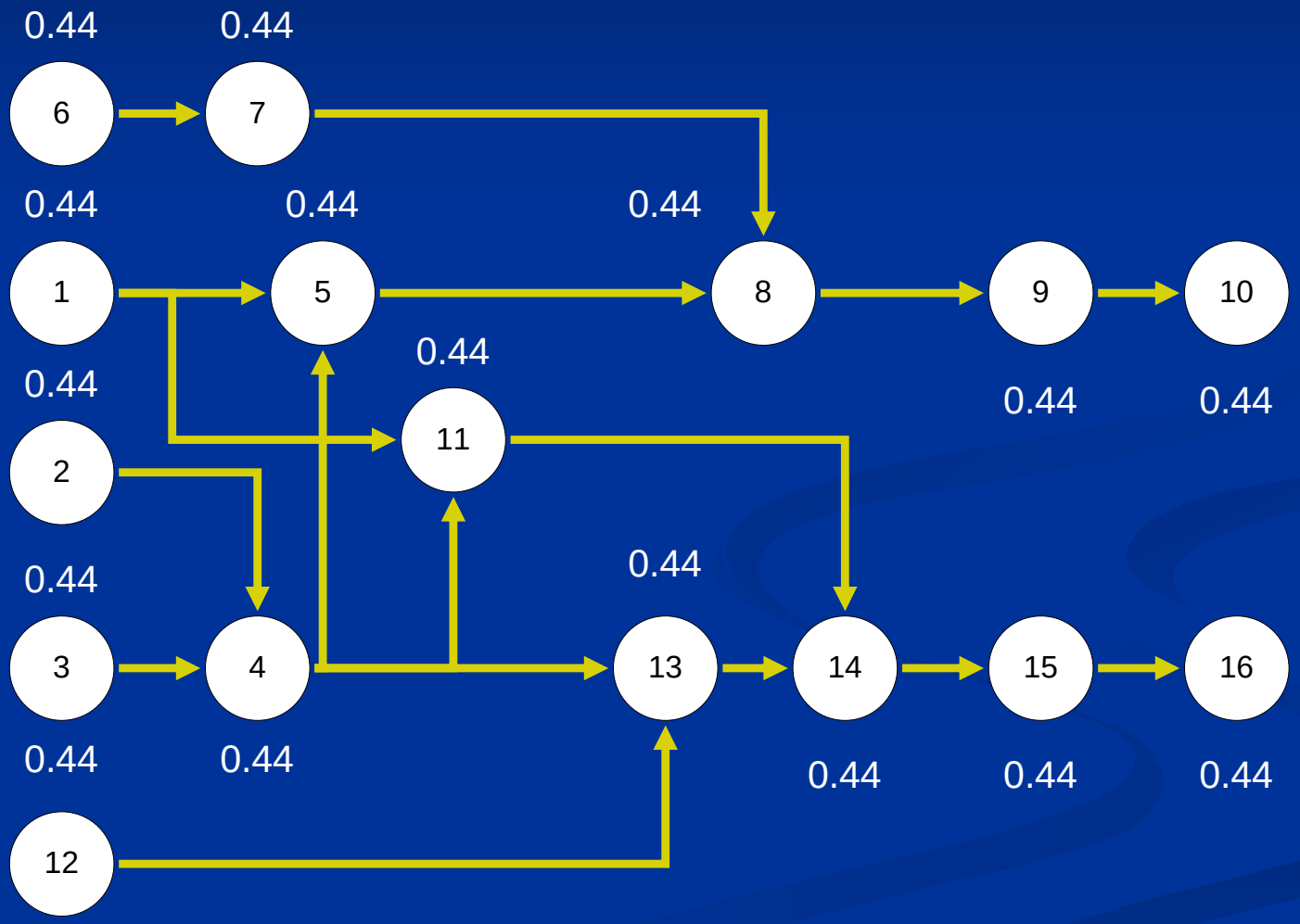
# DKDT Algorithm : Demo



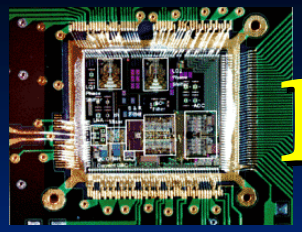

NAND Network



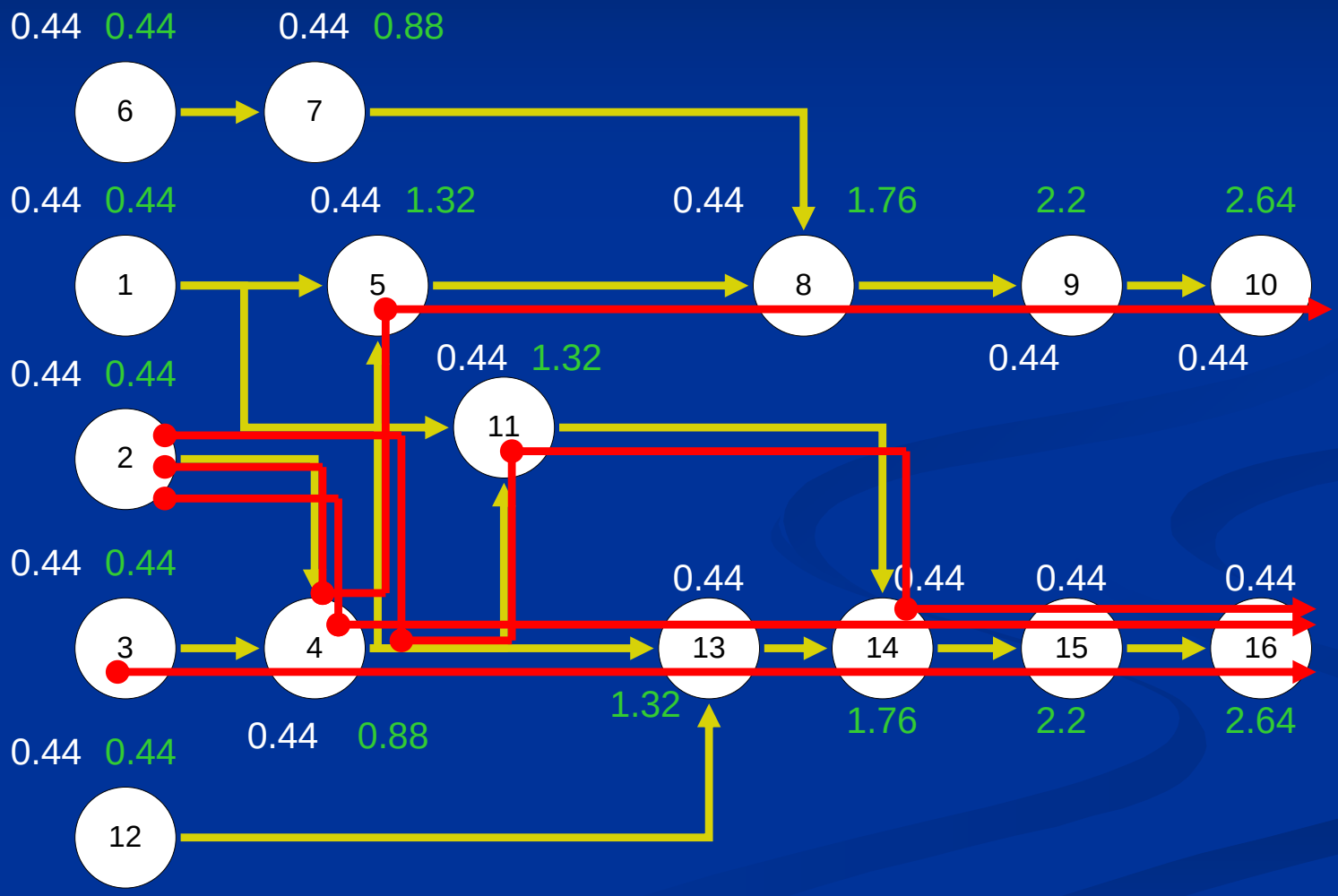
# DKDT Algorithm : Demo



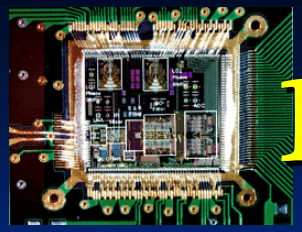

Network with Node Delays



# DKDT Algorithm : Demo

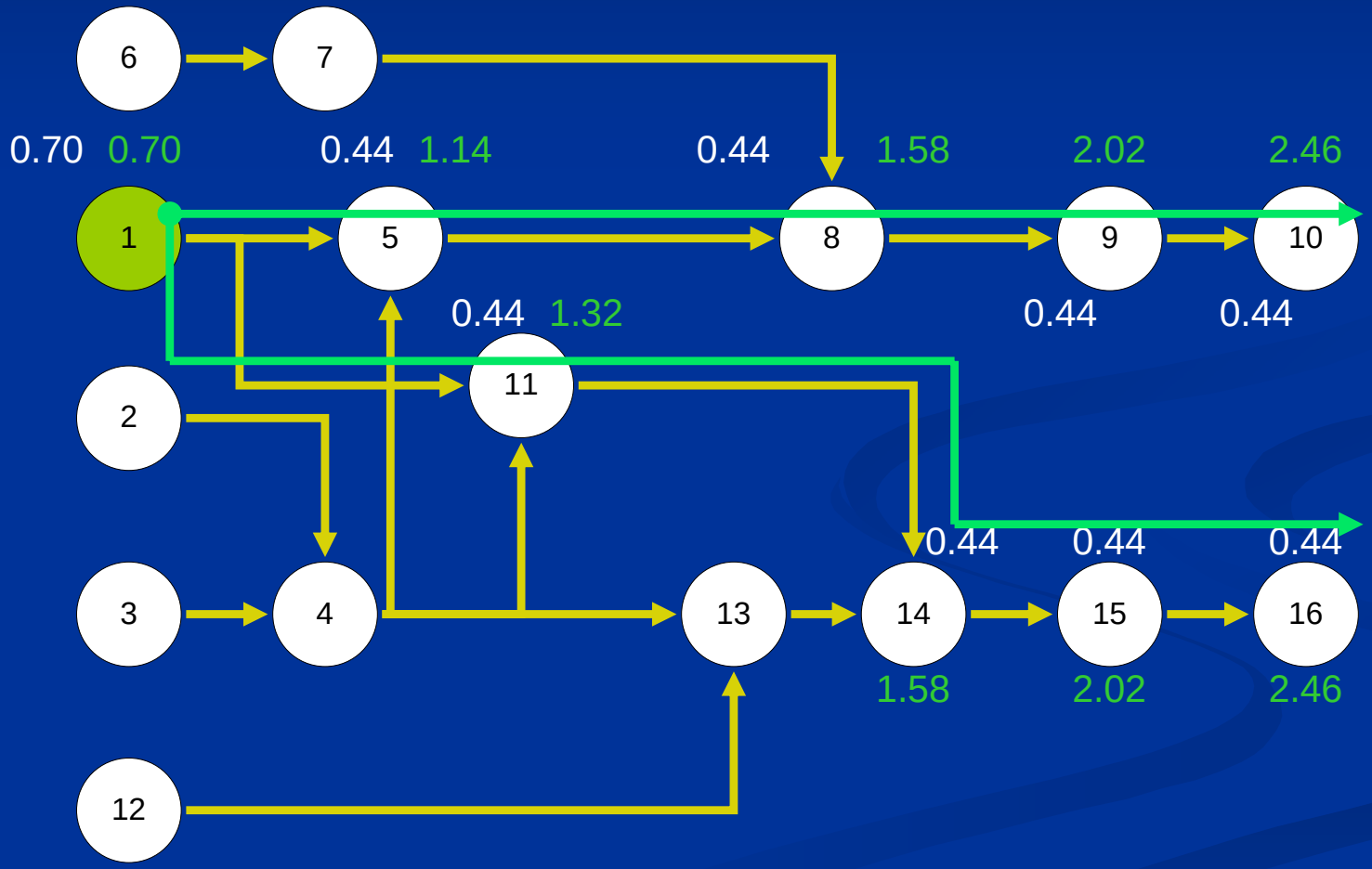



**Network with Path Delays (Fix, $D_T$= 2.64)**

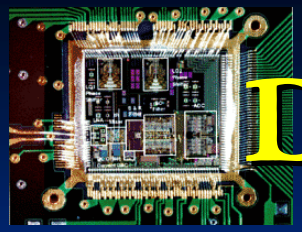


# DKDT Algorithm : Demo

Node1 : $K_2T_2$ ?

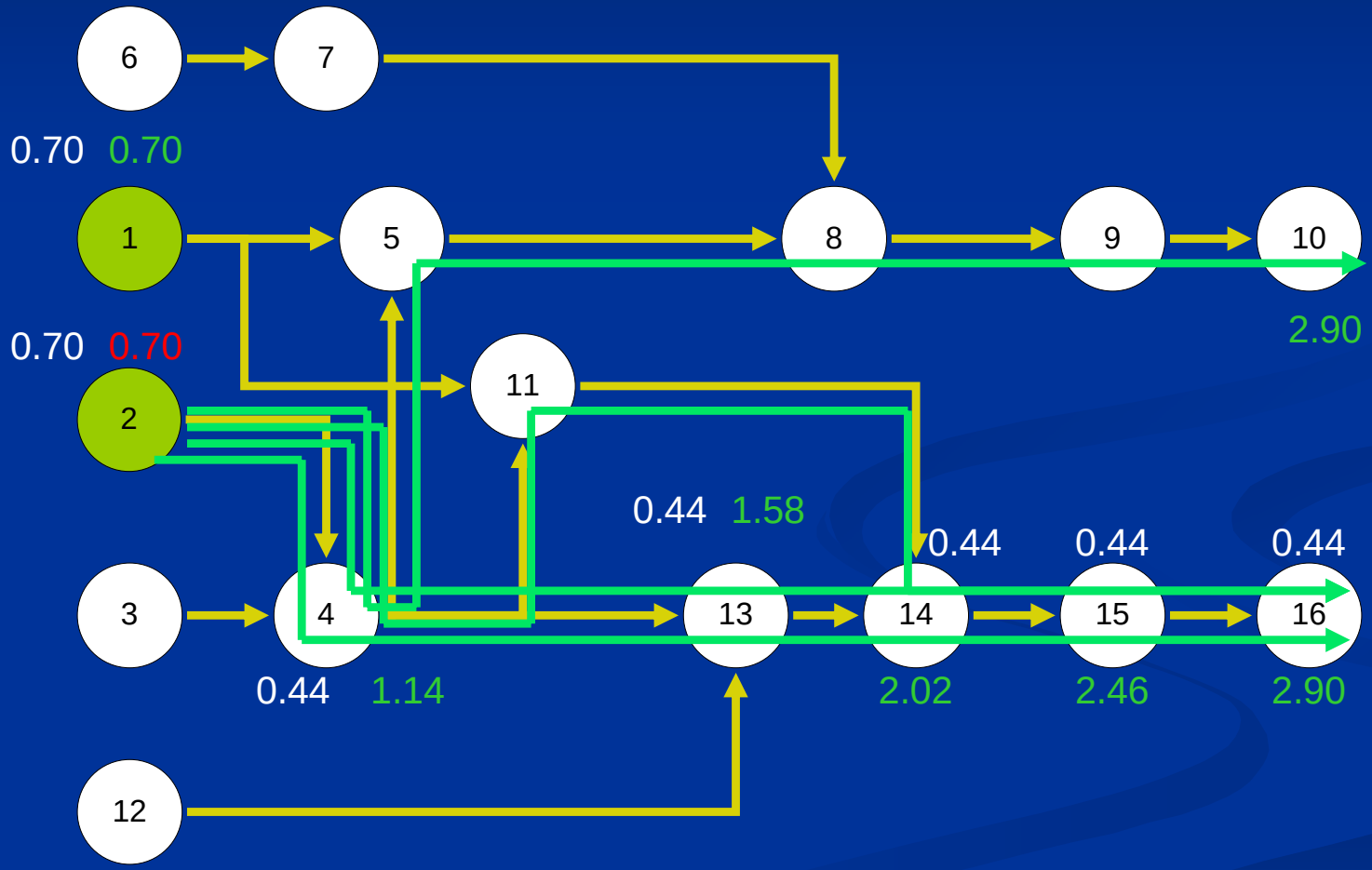



$D_{CP} < D_T$ → Yes, $K_2T_2$ for Node1

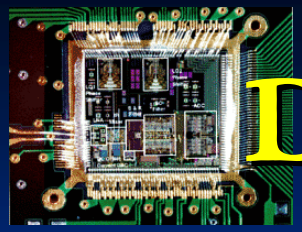


# DKDT Algorithm : Demo

## Node2 : $K_2T_2$ ?

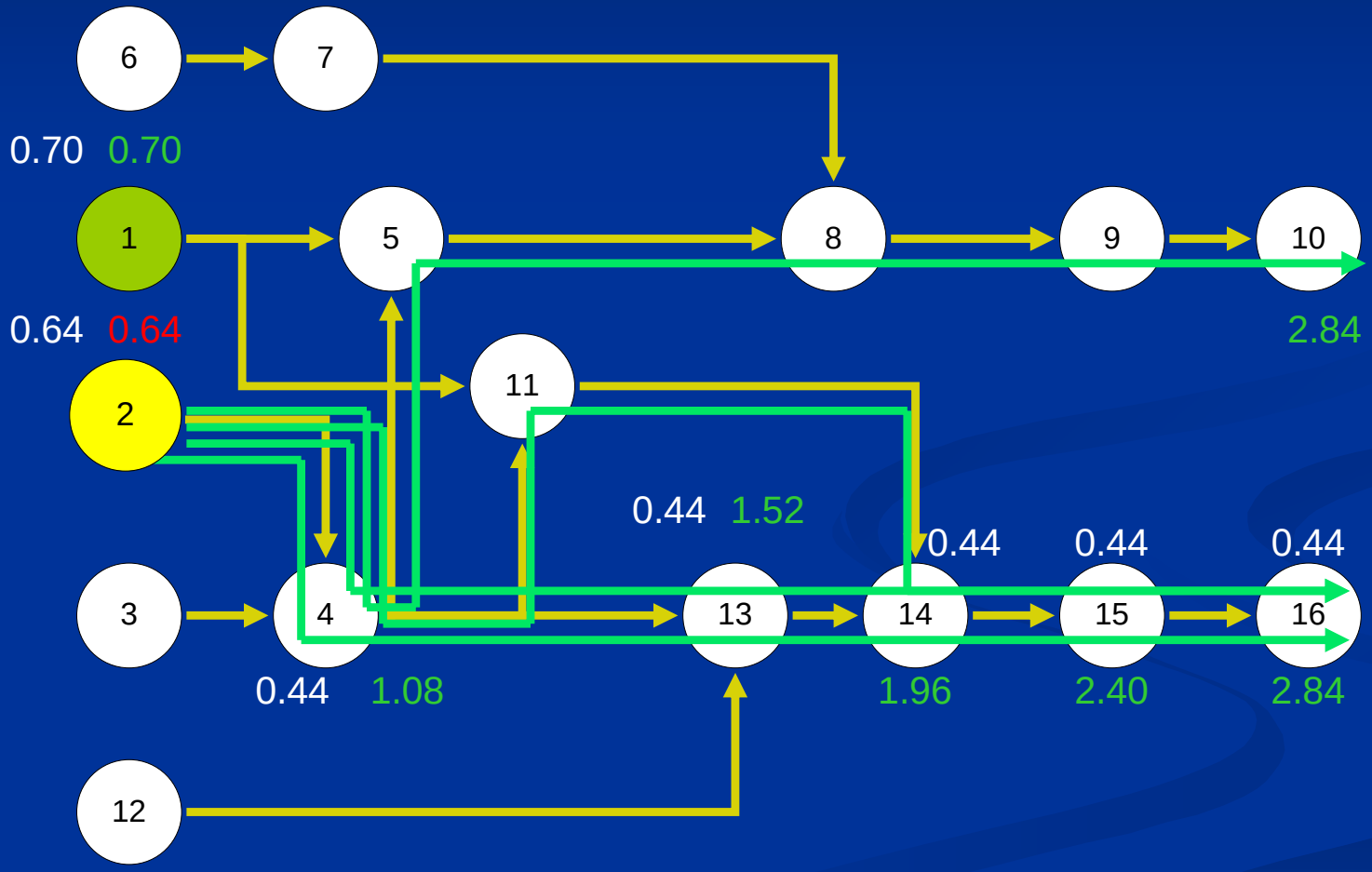



$D_{CP} > D_T \rightarrow$ No for $K_2T_2$

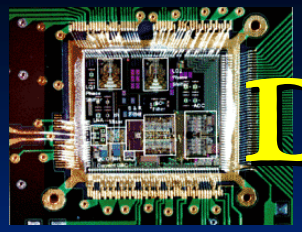


# DKDT Algorithm : Demo

Node2 : $K_2T_1$ ?

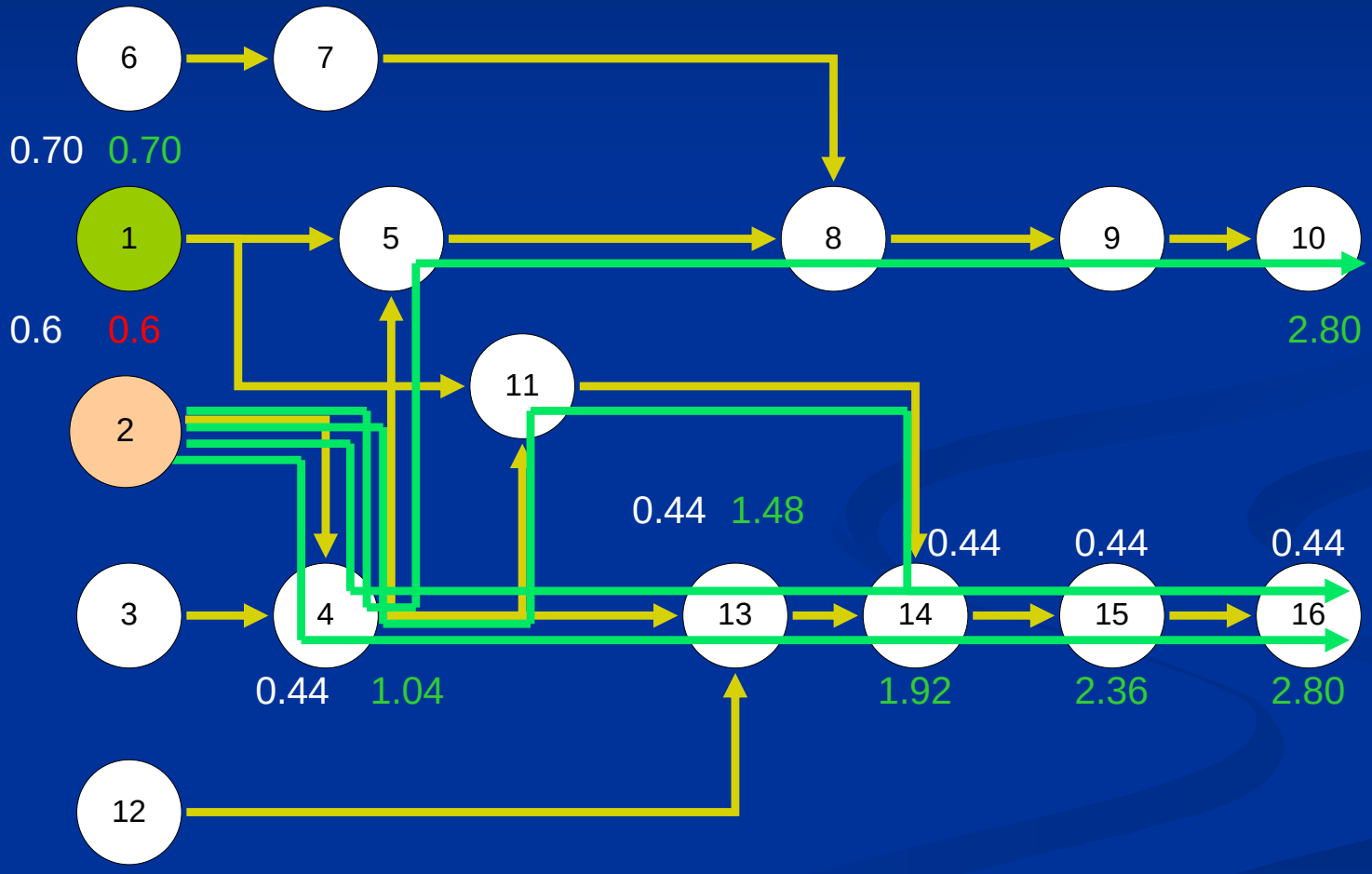



$D_{CP} > D_T \rightarrow$ No for $K_2T_1$

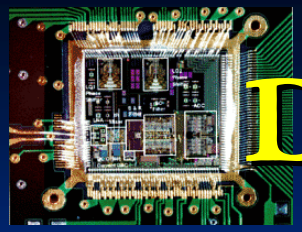


# DKDT Algorithm : Demo

Node2 : $K_1T_2$ ?

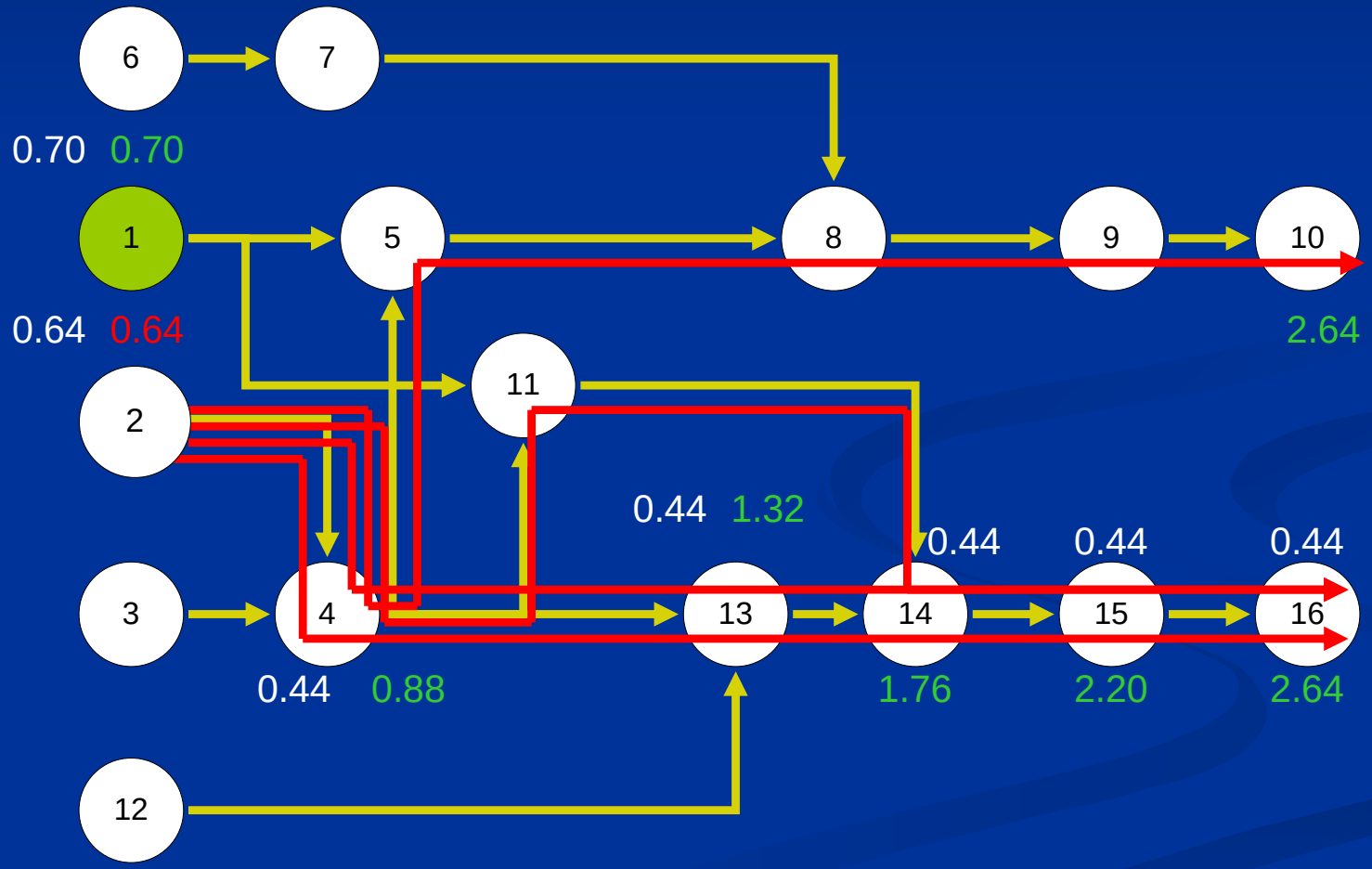


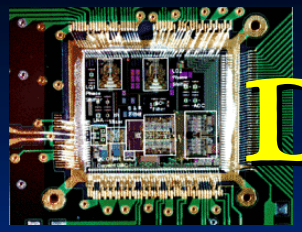

$D_{CP} > D_T \rightarrow$ No for $K_1T_2$



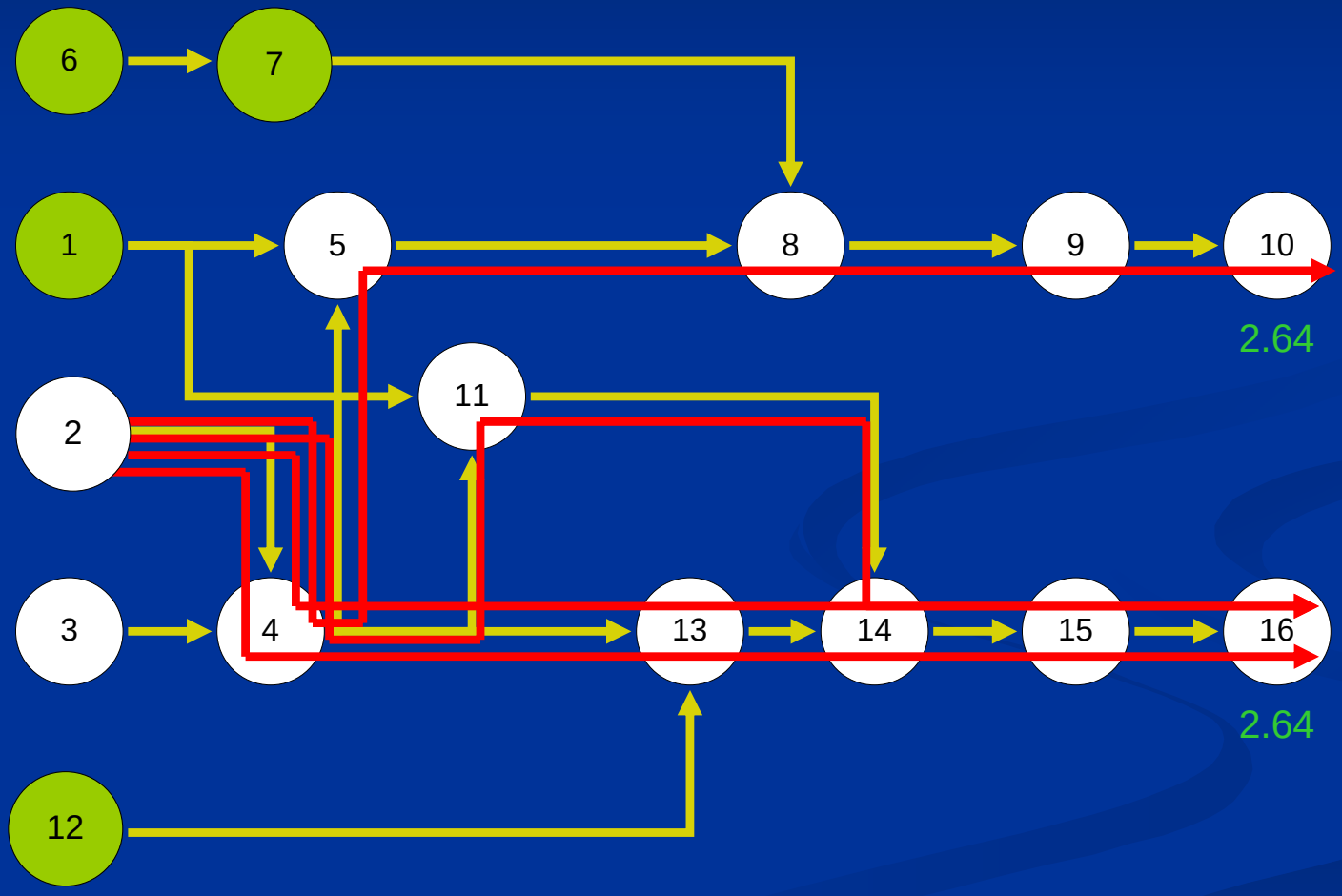
# DKDT Algorithm : Demo

## Node2 : Reassign $K_1T_1$

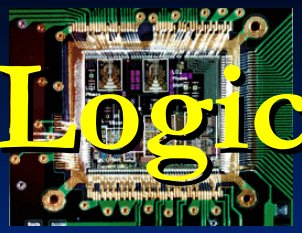




# DKDT Algorithm : Demo

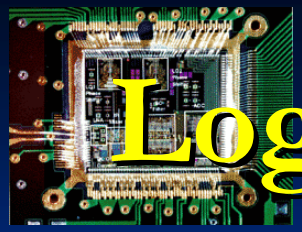



Final Assignment



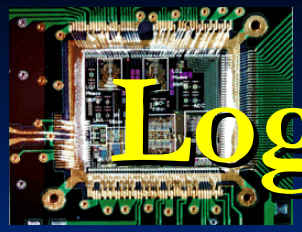
# Logic Cell Characterization : Load

- The Berkeley Predictive Technology Model (BPTM) has been used.
- The first step in the characterization was the selection of an appropriate capacitive load ($C_{Load}$ = 10 * $C_{ggPMOS}$ used).
- The supply voltage is held at $V_{DD}$ = 0.7V.
- We define the delay as the time difference between the 50% level of input and output.



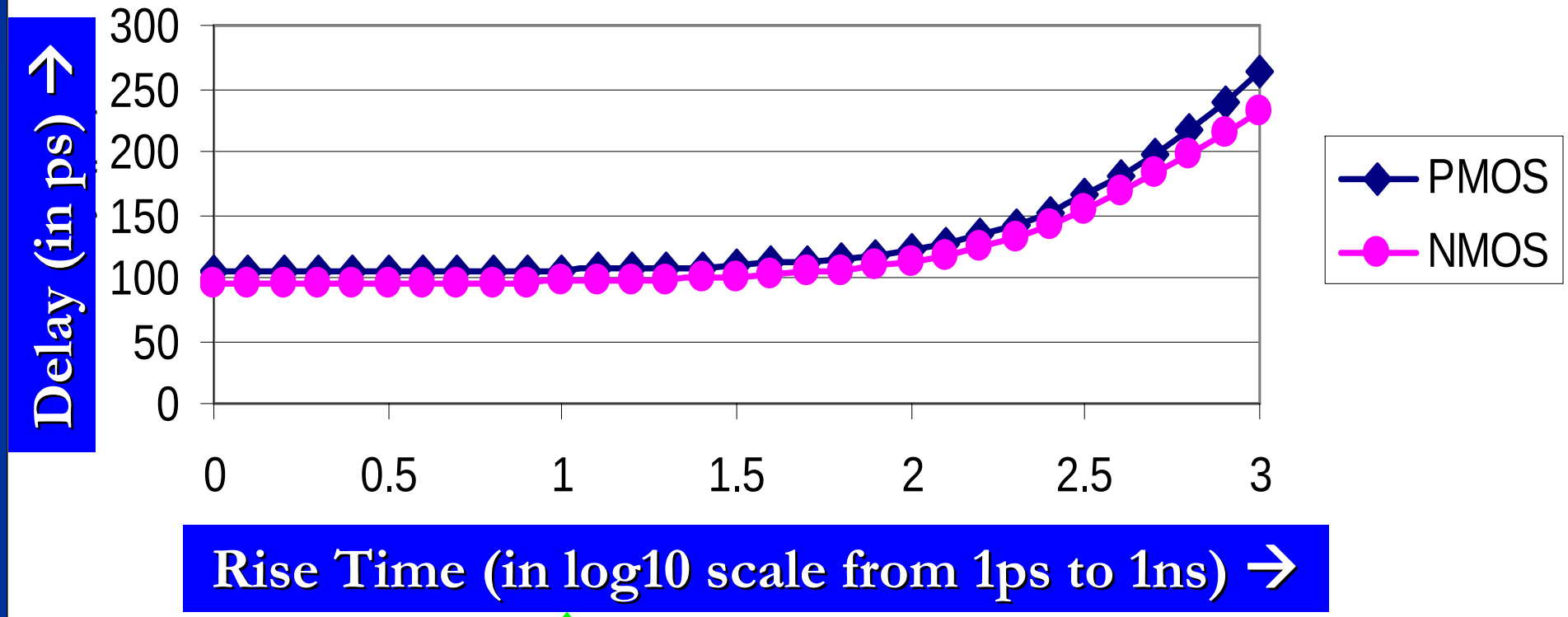
# Logic Cell Characterization : $t_r$

- For worst-case scenarios in the development of the algorithm, we chose the maximum delay time [ i. e. maximum ($t_{pdr}$, $t_{pdf}$) ].
- The effect of switching pulse rise time $t_r$ was initially examined on the delay characteristics.
- To eliminate an explicit dependence of the algorithm results on $t_r$ , we chose a value that is realistic yet does not affect the delay significantly.



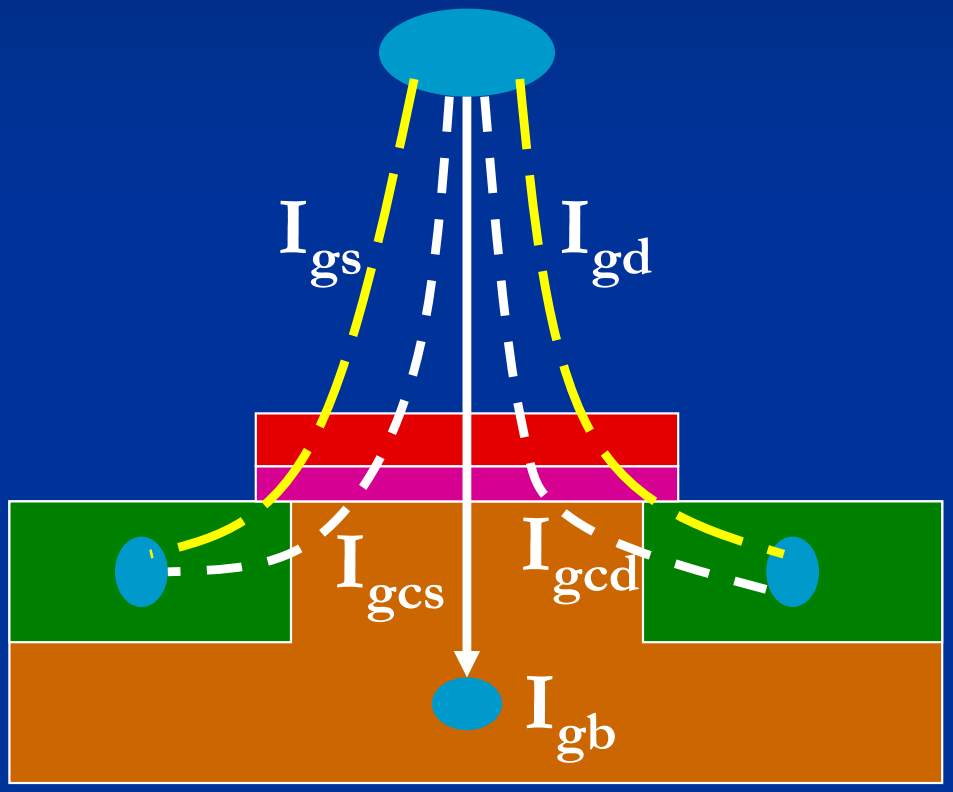
# Logic Cell Characterization : $t_r$

## Delay Versus Rise Time



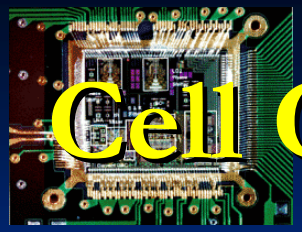

Selected, $t_r$ = 10ps

# Logic Cell Characterization : $I_{gate}$




BSIM4 Model

- Direct tunneling current is calculated by evaluating both the source and drain components.
- For the logic gate, $I_{gate} = \Sigma_{\forall MOS} (|I_{gs}| + |I_{gd}|)$.
- This accounts for tunneling current contributions from devices in both the ON and OFF state.

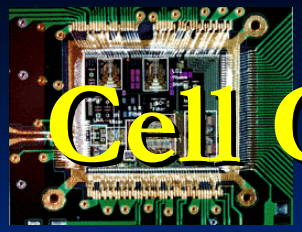


# Cell Characterization : $K_{gate}$ Modeling

- The effect of varying dielectric material was modeled by calculating an equivalent oxide thickness ($T^*_{ox}$) according to the formula:

$$T^*_{ox} = (K_{gate} / K_{ox})\, T_{gate}$$

- Here, $K_{gate}$ is the dielectric constant of the gate dielectric material other than $SiO_2$, (of thickness $T_{gate}$), while $K_{ox}$ is the dielectric constant of $SiO_2$.



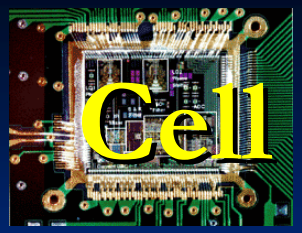
# Cell Characterization : $T_{gate}$ Modeling

- The effect of varying oxide thickness $T_{ox}$ was incorporated by varying TOXE in SPICE model.
- Length of the device is proportionately changed to minimize the impact of higher dielectric thickness on the device performance :

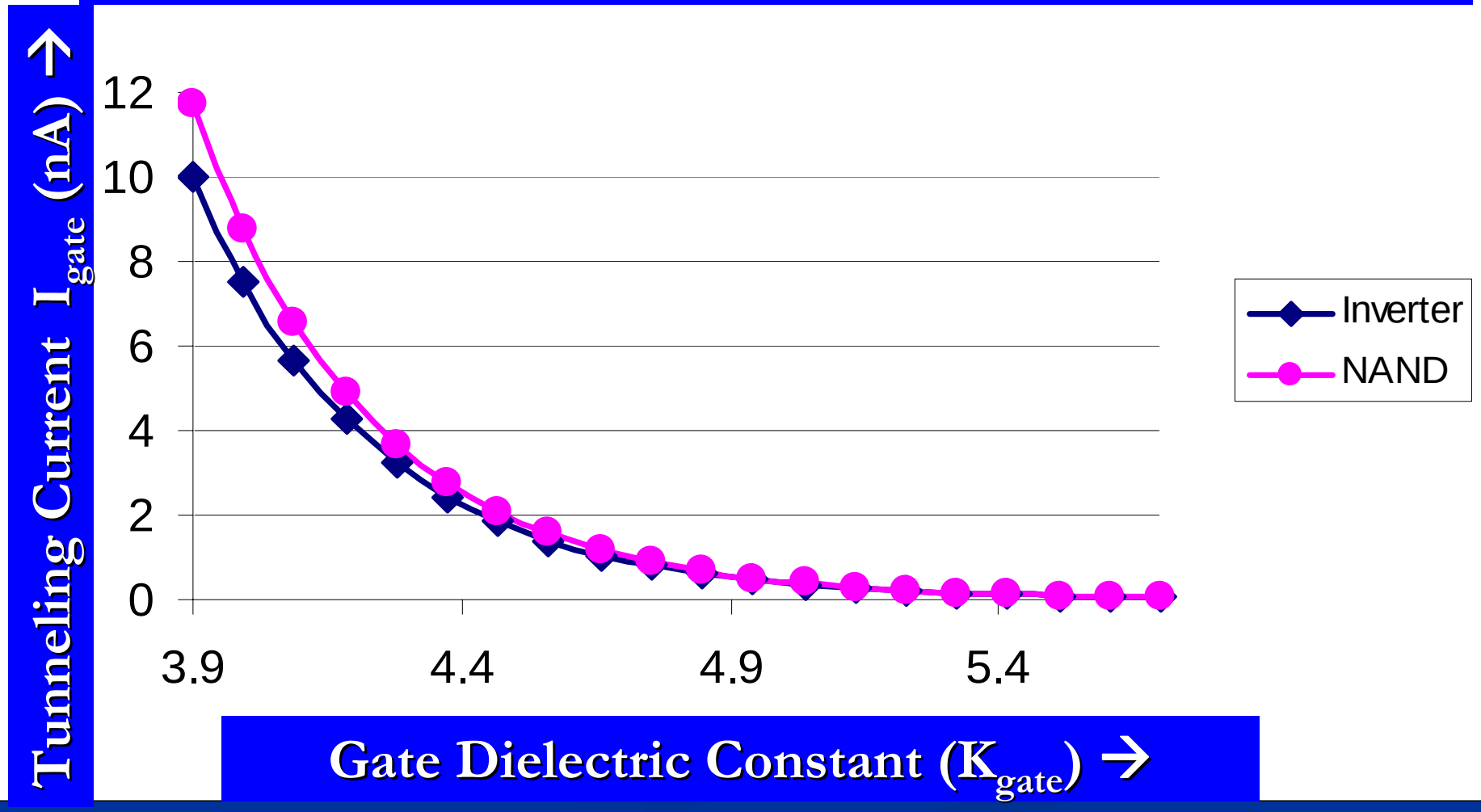
$$L^* = (T^*_{ox} / T_{ox})\, L$$

- Length and width of the transistors are chosen to maintain (W:L) ratio of (4:1) for NMOS and (8:1) for PMOS.

# Cell Characterization : $I_{gate}$ Vs $K_{gate}$

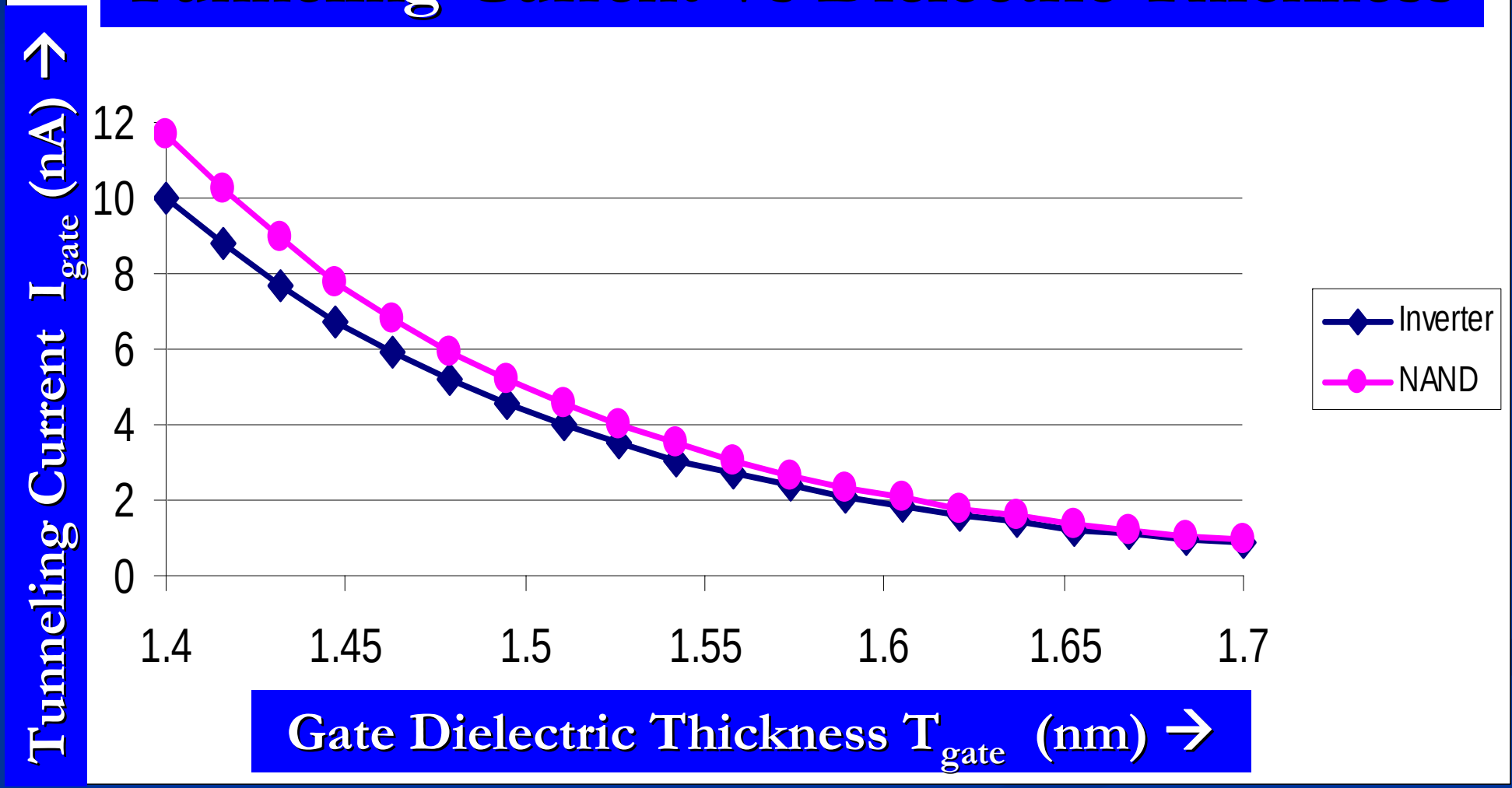


## Tunneling Current Vs Dielectric Constant



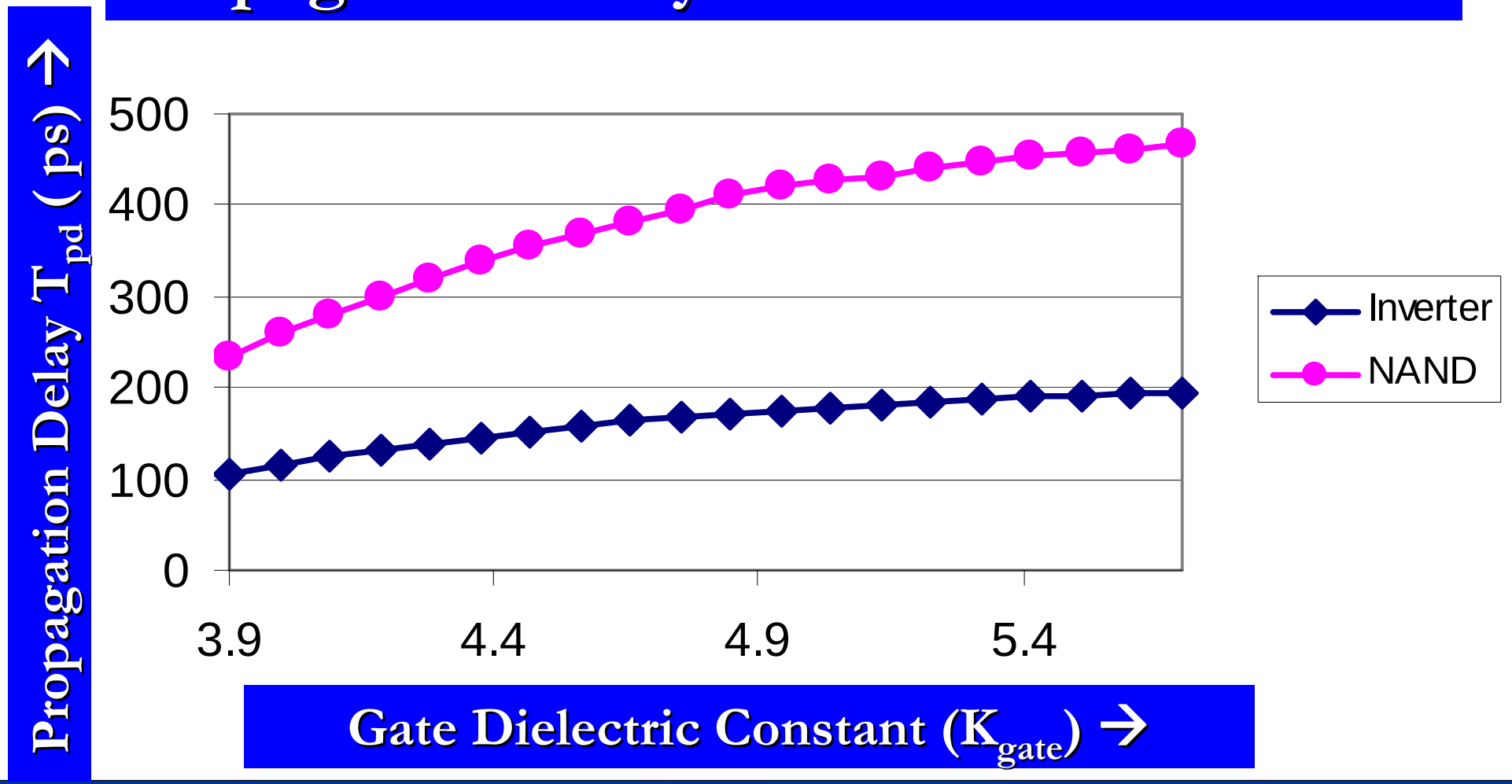
# Cell Characterization : $I_{gate}$ Vs $T_{gate}$

## Tunneling Current Vs Dielectric Thickness



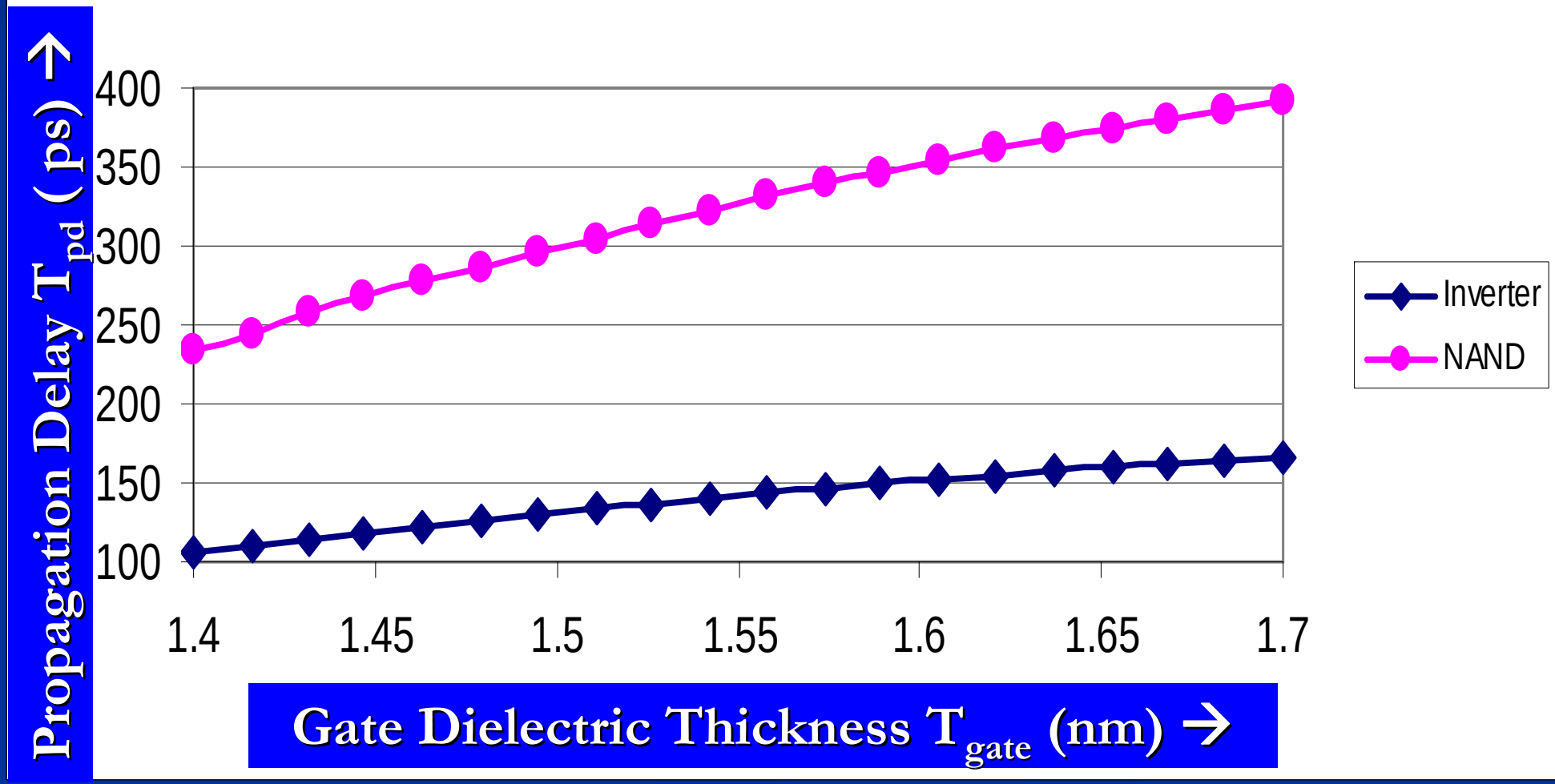
# Cell Characterization : $T_{pd}$ Vs $K_{gate}$

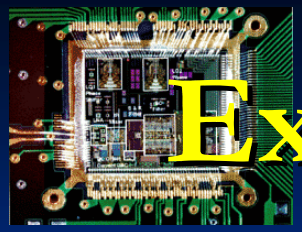
## Propagation Delay Vs Dielectric Constant



# Cell Characterization : $T_{pd}$ Vs $T_{gate}$

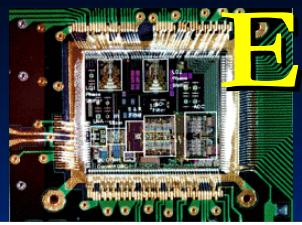
## Propagation Delay Vs Dielectric Thickness





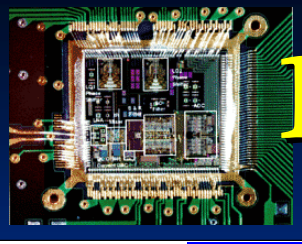
# Experimental Results: Setup

- DKDT algorithm was implemented in C and used along with SIS, and tested on the ISCAS'85 benchmarks.
- The library of gates consisting of four types of NAND and four types of inverters was characterized for the 45 nm technology using SPECTRE tool.
- We used $K_1 = 3.9$ (for $SiO_2$), $K_2 = 5.7$ (for SiON), $T_1 = 1.4$ nm, and $T_2 = 1.7$ nm to perform our experiments.
- The value of $T_1$ is chosen as the default value from the BSIM4.4.0 model card and value of $T_2$ is intuitively chosen based on the characterization process.



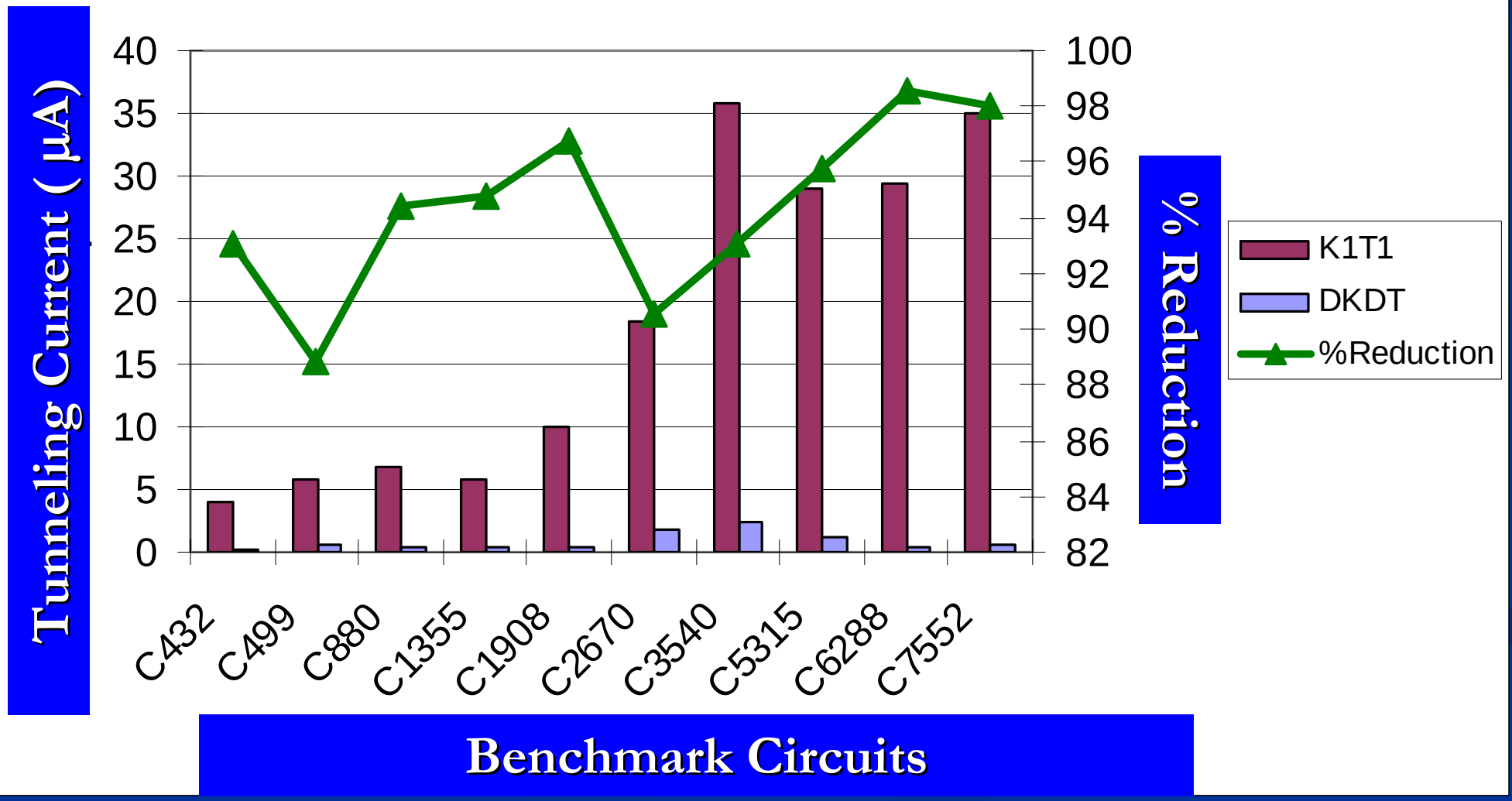
# Experimental Results : Table

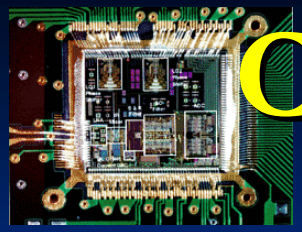
| CKTs | Gates | Critical Delay (ps) | Current for $K_1T_1$ (nA) | Current for DKDT (nA) | %Reduction |
|---|---|---|---|---|---|
| C17 | 24 | 2.22 | 187.2 | 93.66 | 49.96 |
| C432 | 160 | 6.67 | 4071.6 | 281.4 | 93.08 |
| C499 | 202 | 3.56 | 5885.1 | 656.06 | 88.85 |
| C880 | 383 | 10.68 | 6739.2 | 375.38 | 94.42 |
| C1355 | 546 | 3.56 | 5885.1 | 305.16 | 94.81 |
| C1908 | 880 | 11.57 | 10015.2 | 319.69 | 96.80 |
| C2670 | 1193 | 42.71 | 18415.8 | 1734.08 | 90.85 |
| C3540 | 1669 | 31.59 | 35708.4 | 2461.93 | 93.10 |
| C5315 | 2406 | 40.04 | 29027.7 | 1220.89 | 95.79 |
| C6288 | 2406 | 43.15 | 29355.3 | 413.96 | 98.59 |
| C7552 | 3512 | 45.82 | 34947.9 | 695.38 | 98.01 |



# DKDT Algorithm Results

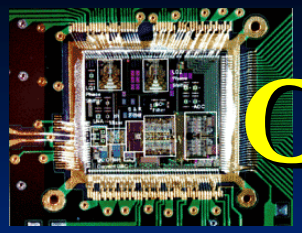
## Tunneling Current and % Reduction



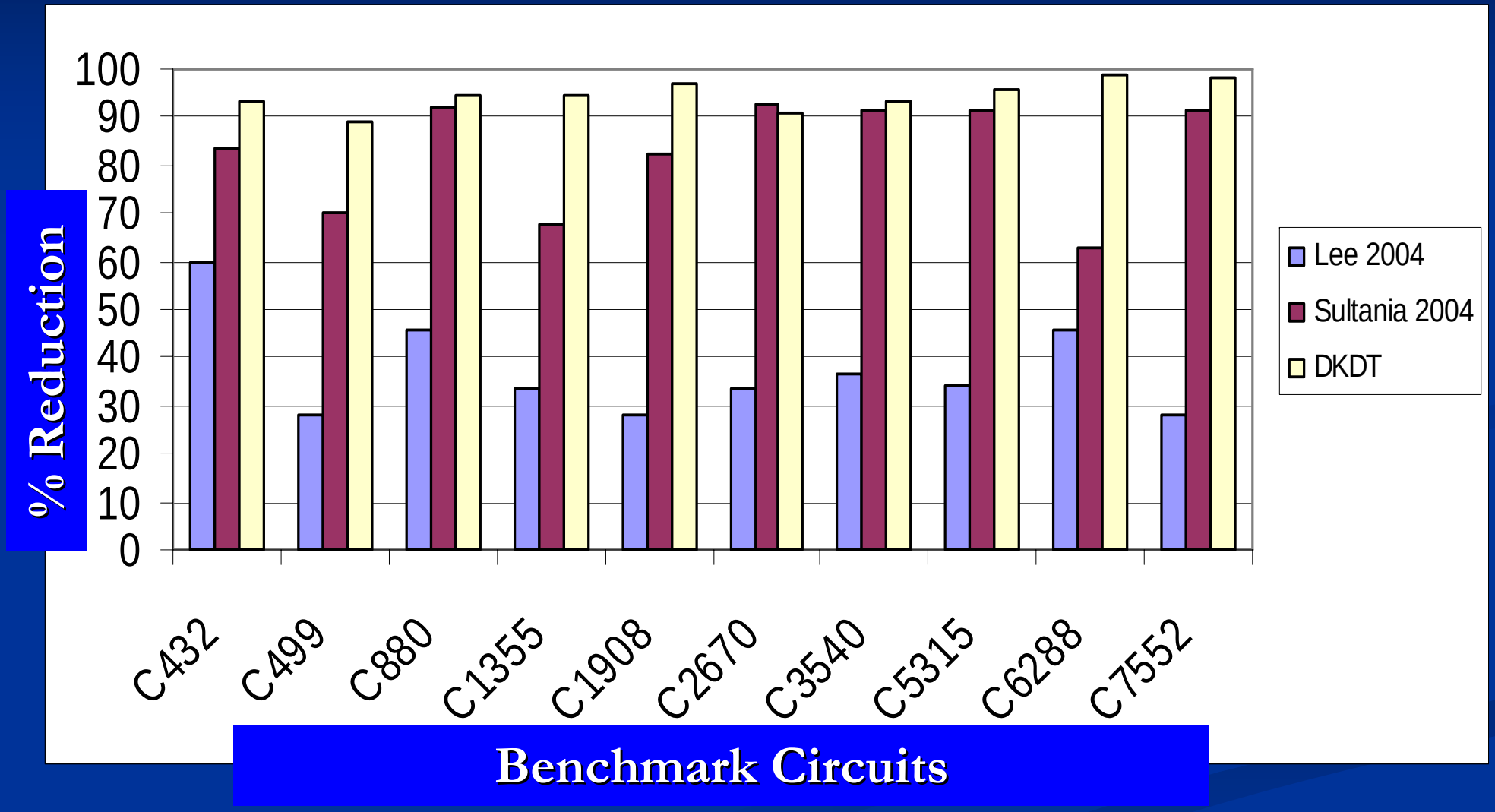


# Comparative View : Table

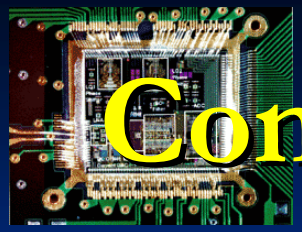
| Bench-mark Circuits | Sultania (100 nm) | | Lee (100 nm) | | DKDT (45 nm) | |
|---|---|---|---|---|---|---|
| | %Reduction | %Penalty | %Reduction | %Penalty | %Reduction | %Penalty |
| C432 | 83.8 | 24.6 | 59.52 | NA | 93.08 | 0 |
| C499 | 70.2 | 25.4 | 28.25 | NA | 88.85 | 0 |
| C880 | 92.3 | 25.5 | 45.43 | NA | 94.42 | 0 |
| C1355 | 67.9 | 25.0 | 33.50 | NA | 94.81 | 0 |
| C1908 | 82.3 | 25.2 | 27.76 | NA | 96.80 | 0 |
| C2670 | 92.6 | 25.3 | 33.80 | NA | 90.58 | 0 |
| C3540 | 91.4 | 25.1 | 36.40 | NA | 93.10 | 0 |
| C5315 | 91.7 | 26.1 | 34.34 | NA | 95.79 | 0 |
| C6288 | 62.7 | 25.7 | 45.86 | NA | 98.59 | 0 |
| C7552 | 91.6 | 25.3 | 28.10 | NA | 98.01 | 0 |



# Comparative View : Chart

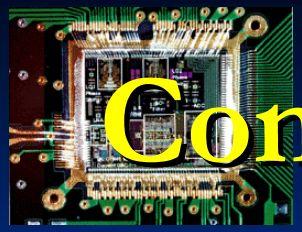



**NOTE**: DKDT has not time penalty.



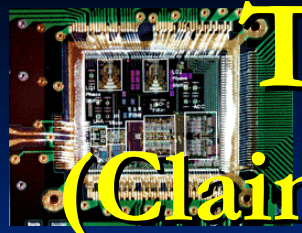
# Conclusions and Future Works

- New approach DKDT for tunneling current reduction accounting for both the ON and OFF states.
- Polynomial time complexity heuristic algorithm could carry out such DKDT assignment for benchmark circuits in reasonable amount of time.
- Experiments prove significant reductions in tunneling current without performance penalty.

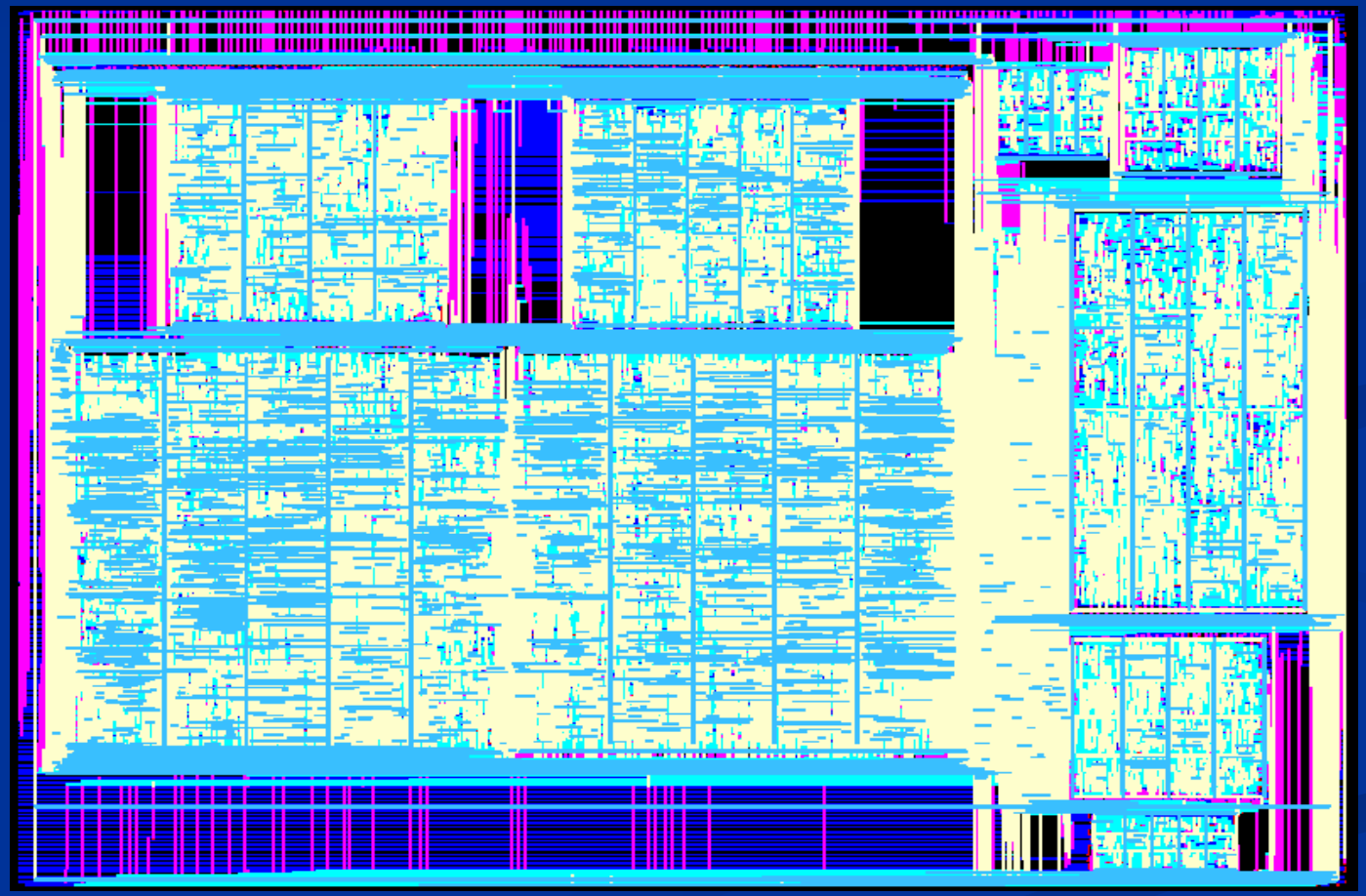


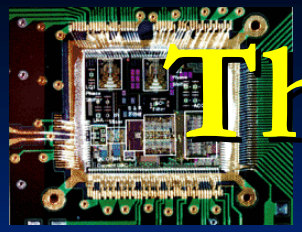
# Conclusions and Future Works

- Modeling of other high-K dielectrics is under progress.
- Development of optimal assignment algorithm can be considered.
- Tradeoff of tunneling, area and performance needs to be explored.
- DKDT based design may need more masks for the lithographic process during fabrication.



# The Latest Chip Designed (Claim: Lowest power consuming image watermarking chip available at present)





# The Latest Chip : Statistics

Technology : TSMC 0.25 μ

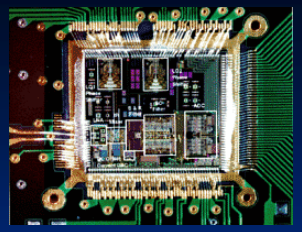
Total Area : 16.2 sq mm

Dual Clocks : 280 MHz and 70 MHz

Dual Voltages : 2.5V and 1.5V

No. of Transistors : 1.4 million

Avg. Power Consumption : 0.3 mW



# Thank You