EE141-Spring 2012 Digital Integrated Circuits

Lecture 23
Clocks + Power

Administrativia

- Project Phase 2 due Today.
- Project Phase 3 to be launched today
- Assignment 9 posted today
  - One more assignmnet (#10) – will not be graded

Class Material

- Last lecture
  - Registers + Timing
- Today's lecture
  - Timing + Clocks
- Reading (Ch 10)

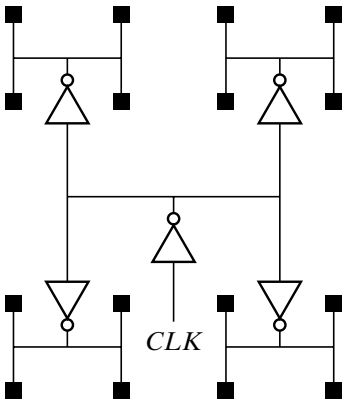
Clock Distribution

Clock Distribution

- Single clock generally used to synchronize all logic on the same chip
  - Need to distribute clock over the entire die
  - While maintaining low skew/jitter
  - (And without burning too much power)

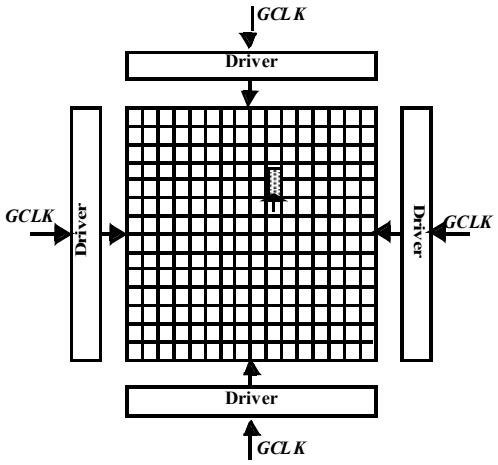
What matters is to minimize clock skew – not clock delay

H-Tree



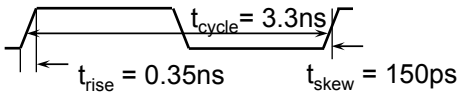
Equal wire length/number of buffers to get to every location

Clock Grid

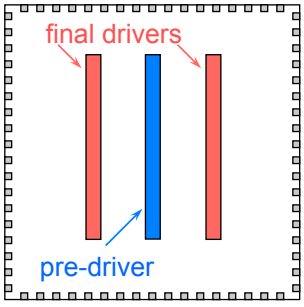


- No RC matching
- But huge power

Example: DEC Alpha 21164 (1995)

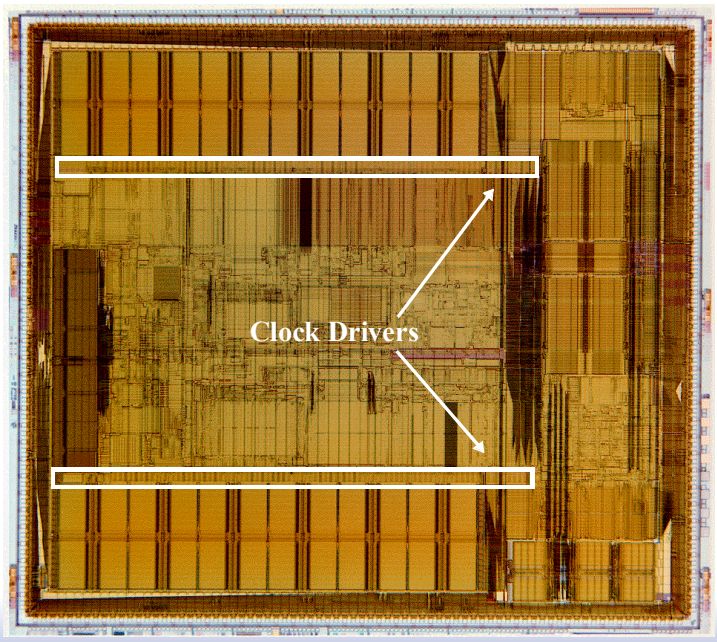


Clock waveform



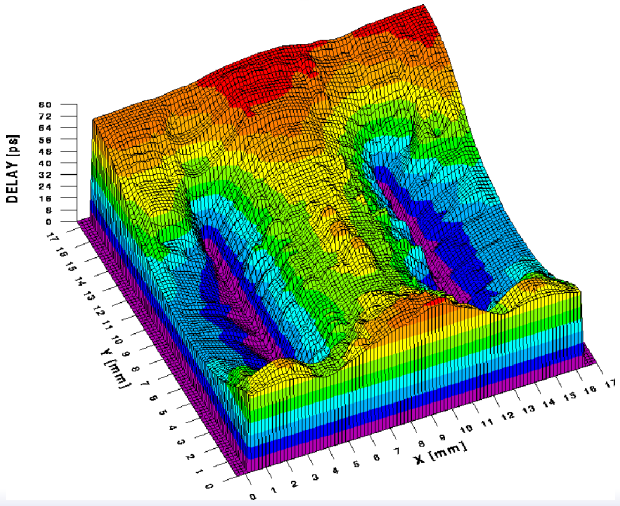
Location of clock driver on die

- 2 phase single wire clock, distributed globally
- 2 distributed driver channels
  - Reduced RC delay/skew
  - Improved thermal distribution
  - 3.75nF clock load, 20W power
  - 58 cm final driver width
- Local inverters for latching
- Conditional clocks in caches to reduce power
- More complex race checking
- Device variation

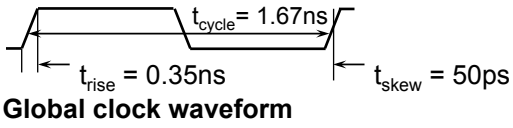


[9]

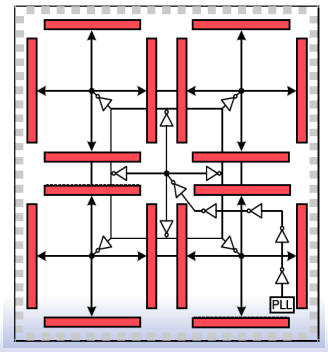
Clock Skew in Alpha Processor



EV6 (Alpha 21264) Clocking
600 MHz – 0.35 micron CMOS

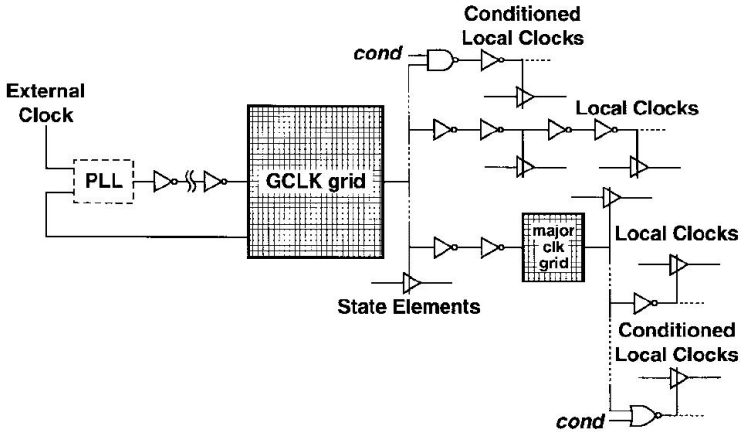


Global clock waveform

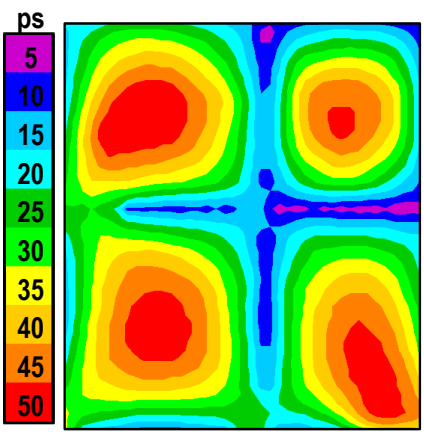


- 2 Phase, with multiple conditional buffered clocks
  - 2.8 nF clock load
  - 40 cm final driver width
- Local clocks can be gated "off" to save power
- Reduced load/skew
- Reduced thermal issues
- Multiple clocks complicate race checking

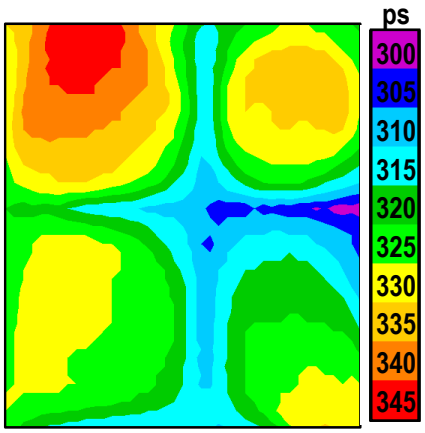
21264 Clocking



EV6 Clock Results



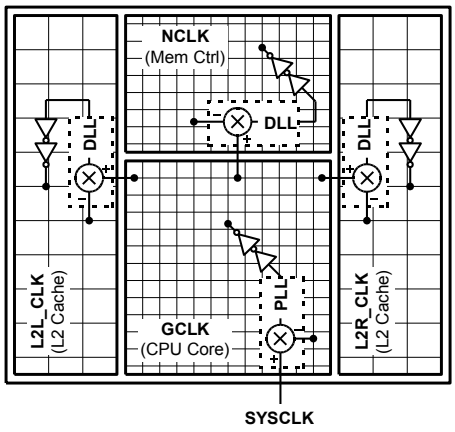
GCLK Skew
(at Vdd/2 Crossings)



GCLK Rise Times
(20% to 80% Extrapolated to 0% to 100%)

EV7 Clock Hierarchy (2002)

Active Skew Management and Multiple Clock Domains

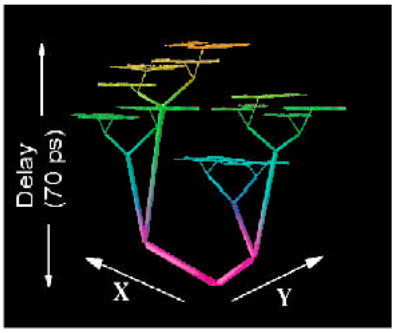


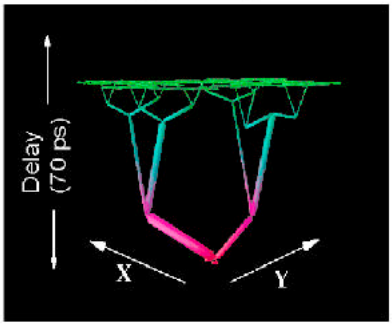
+ widely dispersed drivers
+ DLLs compensate static and low-frequency variation
+ divides design and verification effort
– DLL design and verification is added work
+ tailored clocks

Clock Animations

- By Phillip Restle (IBM)

http://www.research.ibm.com/people/r/restle/resonate.html



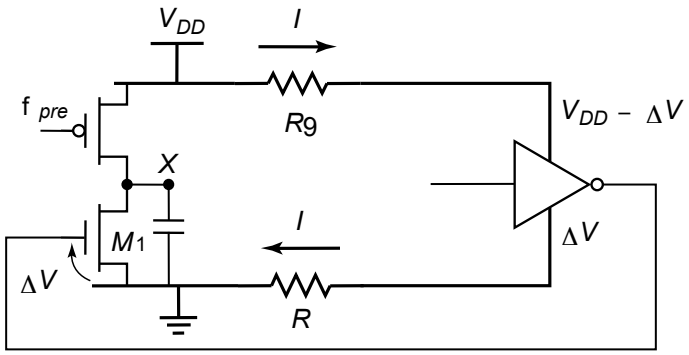


Power Distribution

Impact of Resistance

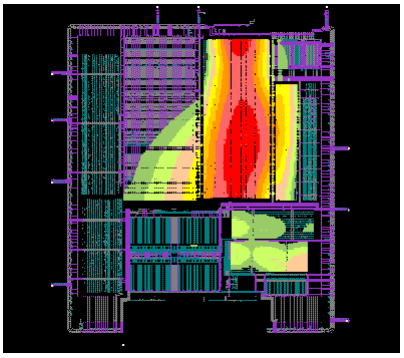
- We have already learned how to drive RC interconnect
- Impact of resistance is commonly seen in power supply distribution:
  - IR drop
  - Voltage variations
- Power supply is distributed to minimize the IR drop and the change in current due to switching of gates

RI Introduced Noise

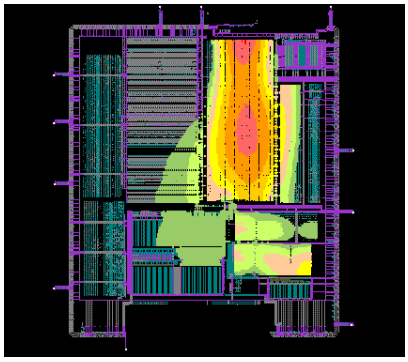


Resistance and the Power Distribution Problem

Before



After



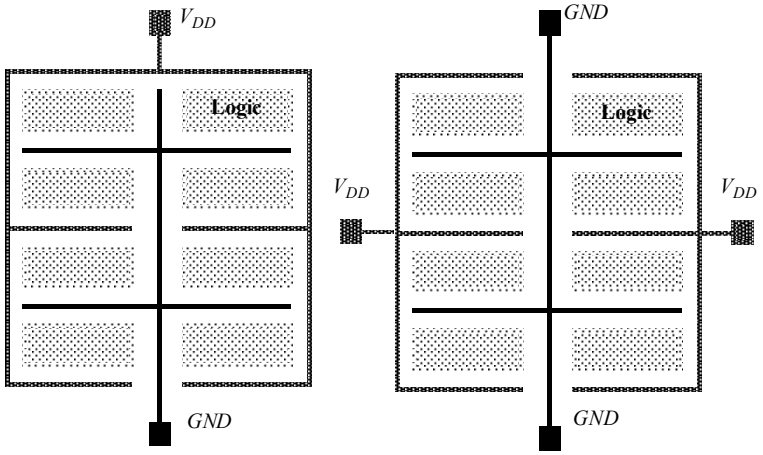
- **Requires fast and accurate peak current prediction**
- **Heavily influenced by packaging technology**

Source: Cadence

Power Distribution

- Low-level distribution is in Metal 1
- Power has to be 'strapped' in higher layers of metal.
- The spacing is set by IR drop, electromigration, inductive effects
- Always use multiple contacts on straps

Power and Ground Distribution



(a) Finger-shaped network (b) Network with multiple supply pins

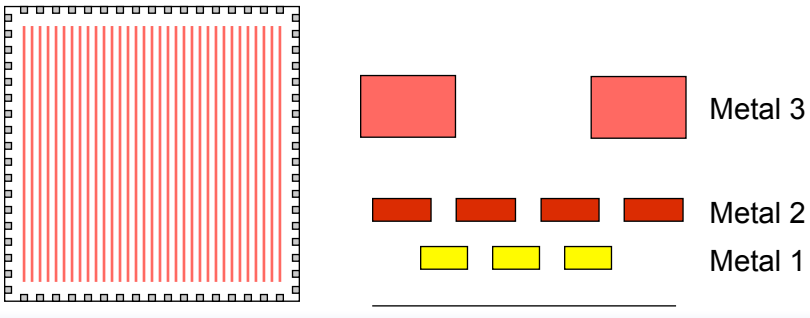
3 Metal Layer Approach (EV4)

3rd "coarse and thick" metal layer added to the technology for EV4 design

Power supplied from two sides of the die via 3rd metal layer

2nd metal layer used to form power grid

90% of 3rd metal layer used for power/clock routing



Courtesy HP

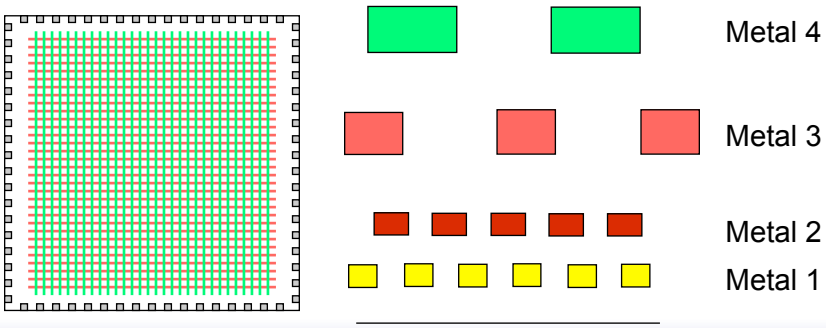
4 Metal Layers Approach (EV5)

4th "coarse and thick" metal layer added to the technology for EV5 design

Power supplied from four sides of the die

Grid strapping done all in coarse metal

90% of 3rd and 4th metals used for power/clock routing



Courtesy HP

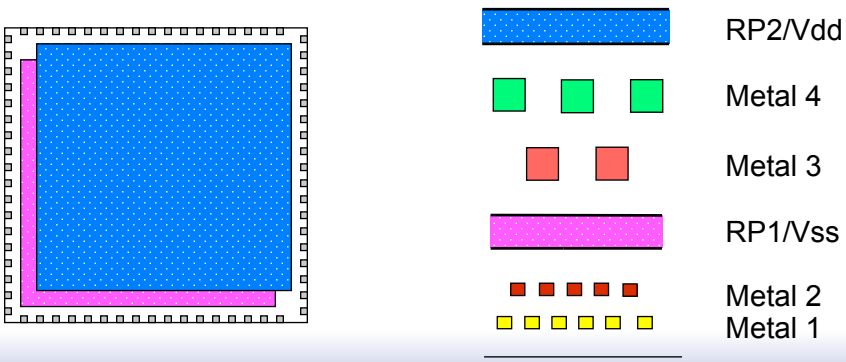
6 Metal Layer Approach – EV6

2 reference plane metal layers added to the technology for EV6 design

Solid planes dedicated to Vdd/Vss

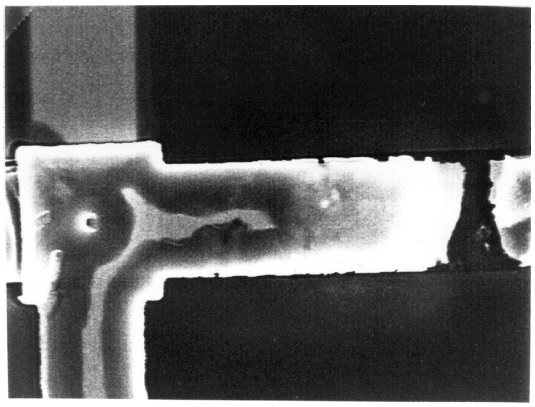
Significantly lowers resistance of grid

Lowers on-chip inductance



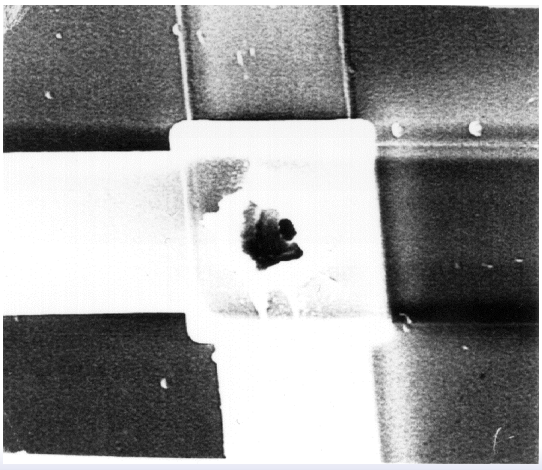
Courtesy HP

Electromigration (1)



Limits dc-current to 1 mA/μm

Electromigration (2)



CMOS Transistor Scaling

Goals of Technology Scaling

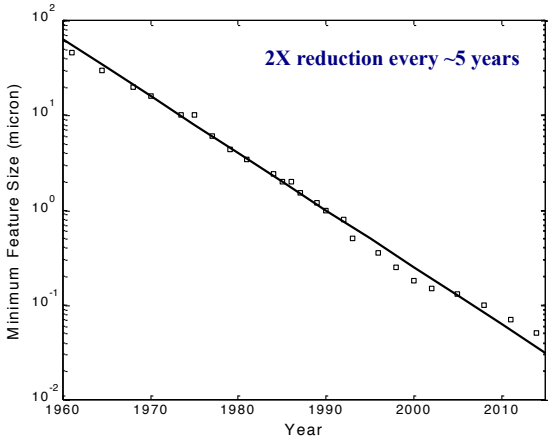
- Make things cheaper:
  - Want to sell more functions (transistors) per chip for the same money
  - Or build same products cheaper
  - Price of a transistor has to be reduced
- But also want to be faster, smaller, lower power…

Technology Scaling

- Benefits of 30% "Dennard" scaling (1974):
  - Double transistor density
  - Reduce gate delay by 30% (increase operating frequency by 43%)
  - Reduce energy per transition by 65% (50% power savings @ 43% increase in frequency)
- Die size used to increase by 14% per generation (not any more)
- Technology generation spans 2-3 years

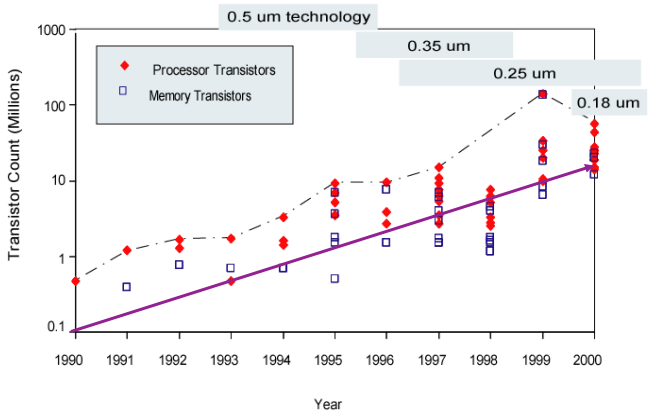
Moore was not always accurate

Technology Scaling (1)



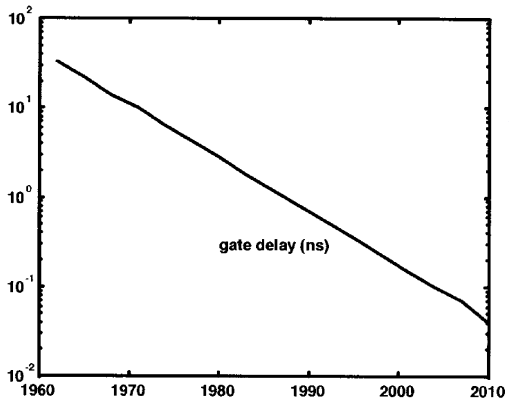
Minimum Feature Size

Technology Scaling (2)



Number of components per chip

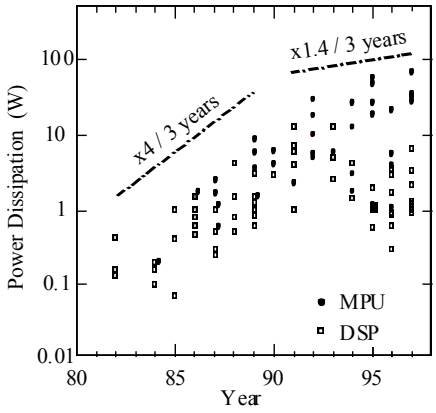
Technology Scaling (3)



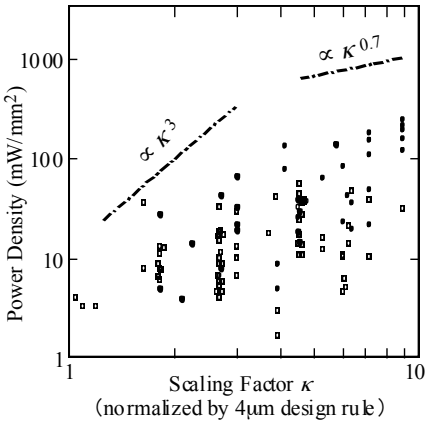
t_p decreases by 30%/year
f increases by 43%

Propagation Delay

Technology Scaling (4)



(a) Power dissipation vs. year.



(b) Power density vs. scaling factor.

From Kuroda

Technology Scaling Models

- **Full Scaling (Constant Electrical Field)**

  ideal model — dimensions and voltages scale together by the same factor *S*

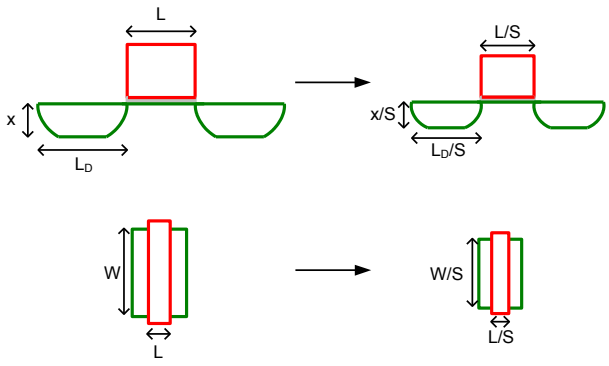
- **Fixed Voltage Scaling**

  most common model until 1990's
  only dimensions scale, voltages remain constant

- **General Scaling**

  most realistic for today's situation —
  voltages and dimensions scale with different factors

Scaling



Full Scaling (Dennard, Long-Channel)

- W, L, t_{ox}: 1/S
- V_{DD}, V_T: 1/S

- Area: WL
- C_{ox}: 1/tox
- C_L: C_{ox}WL
- I_D: $C_{ox}(W/L)(V_{DD}-V_T)^2$
- R_{eq}: V_{DD}/I_{DSAT}

Full Scaling (Dennard, Long-Channel)

- W, L, t_{ox}: 1/S
- V_{DD}, V_T: 1/S

- t_p: $R_{eq}C_L$
- P_{avg}: $C_L V_{DD}^2/t_p$
- P_{avg}/A: $C_{ox}V_{DD}^2/t_p$

Scaling Relationships for Long Channel Devices

Parameter	Relation	Full Scaling	General Scaling	Fixed Voltage Scaling
W, L, t_{ox}		1/S	1/S	1/S
V_{DD}, V_T		1/S	1/U	1
N_{SUB}	V/W_{depl}^2	S	S^2/U	S^2
Area/Device	WL	$1/S^2$	$1/S^2$	$1/S^2$
C_{ox}	$1/t_{ox}$	S	S	S
C_L	$C_{ox}WL$	1/S	1/S	1/S
k_n, k_p	$C_{ox}W/L$	S	S	S
I_{av}	$k_{n,p}V^2$	1/S	S/U^2	S
t_p (intrinsic)	C_LV / I_{av}	1/S	U/S^2	$1/S^2$
P_{av}	C_LV^2/t_p	$1/S^2$	S/U^3	S
PDP	C_LV^2	$1/S^3$	$1/SU^2$	1/S

Full Scaling (Dennard, Short-Channel)

- W, L, t_{ox}: 1/S
- V_{DD}, V_T: 1/S

- Area: WL
- C_{ox}: 1/tox
- C_L: $C_{ox}WL$
- I_D: $WC_{ox}v_{sat}(V_{DD}-V_T-V_{VSAT}/2)$
- R_{eq}: V_{DD}/I_{DSAT}

Full Scaling (Dennard, Short-Channel)

- W, L, t_{ox}: 1/S
- V_{DD}, V_T: 1/S

- t_p: $R_{eq}C_L$
- P_{avg}: $C_L V_{DD}^2/t_p$
- P_{avg}/A: $C_{ox}V_{DD}^2/t_p$

Transistor Scaling (Velocity-Saturated Devices)

Parameter	Relation	Full Scaling	General Scaling	Fixed-Voltage Scaling
W, L, t_{ox}		$1/S$	$1/S$	$1/S$
V_{DD}, V_T		$1/S$	$1/U$	1
N_{SUB}	V/W_{depl}^2	S	S^2/U	S^2
Area/Device	WL	$1/S^2$	$1/S^2$	$1/S^2$
C_{ox}	$1/t_{ox}$	S	S	S
C_{gate}	$C_{ox}WL$	$1/S$	$1/S$	$1/S$
k_n, k_p	$C_{ox}W/L$	S	S	S
I_{sat}	$C_{ox}WV$	$1/S$	$1/U$	1
Current Density	$I_{sat}/Area$	S	S^2/U	S^2
Ron	V/I_{sat}	1	1	1
Intrinsic Delay	$R_{on}C_{gate}$	$1/S$	$1/S$	$1/S$
P	$I_{sat}V$	$1/S^2$	$1/U^2$	1
Power Density	$P/Area$	1	S^2/U^2	S^2

An interesting question

- What ~~will~~ did cause this model to break?
  - Leakage set by kT/q
    - Temp. does not scale
    - V_T set to minimize power
  - Power actually increased
    - Leakage increased drastically
    - *f* increased faster than device speed
    - Hit cooling limit
  - Process Variation
    - Hard to build very small things accurately (less averaging)

Wire Scaling