

## Design Margin

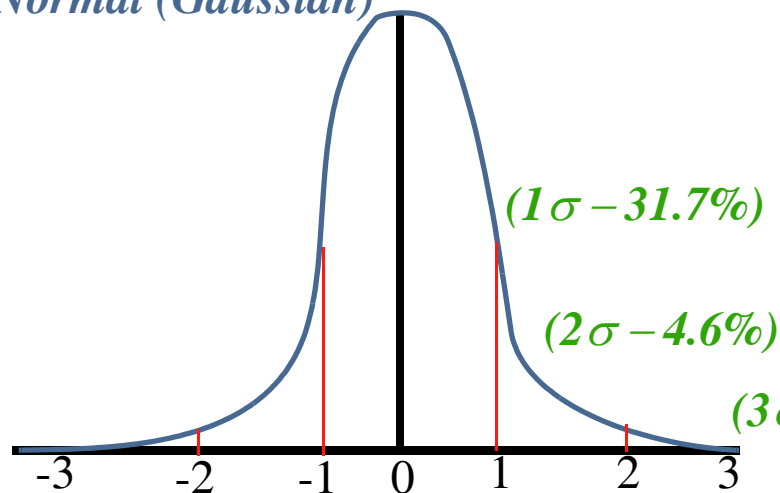
Design Margin required as there are three sources of variation- two environmental and one manufacturing:

- Supply Voltage
- Operating temperature
- Process variation

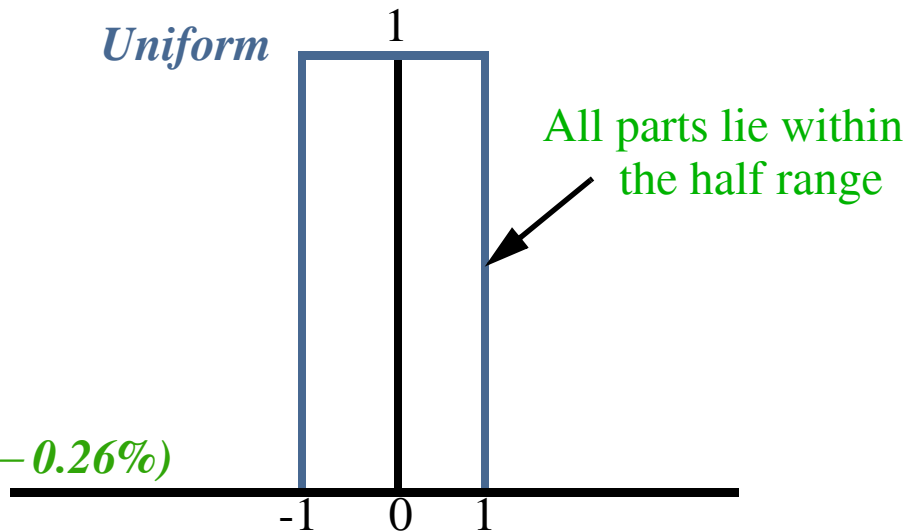
Aim is to design the circuit that will reliably operate over all extremes of these three variables.

Variations can be modeled as *uniform* or *normal (Gaussian)* statistical distributions.

*Normal (Gaussian)*



*Uniform*



## Design Margin

### Supply Voltage

Supply voltage may vary due to tolerance of voltage regulators, IR drop along the supply rail and di/dt noise.

Typically the supply is specified as +/- 10% around nominal (uniform distribution)

Speed is roughly proportional to  $V_{DD}$ , also noise budgets are affected.

### Temperature

Parts must operate over a range of temperatures.

<i>Standard</i>	<i>Minimum</i>	<i>Maximum</i>
<i>Commercial</i>	0°C	70°C
<i>Industrial</i>	-40°C	85°C
<i>Military</i>	-55°C	125°C

## Design Margin

### Process Variation

Devices have variations in film thicknesses, lateral dimensions, doping concentrations etc.

The parameters of individual transistors vary from:

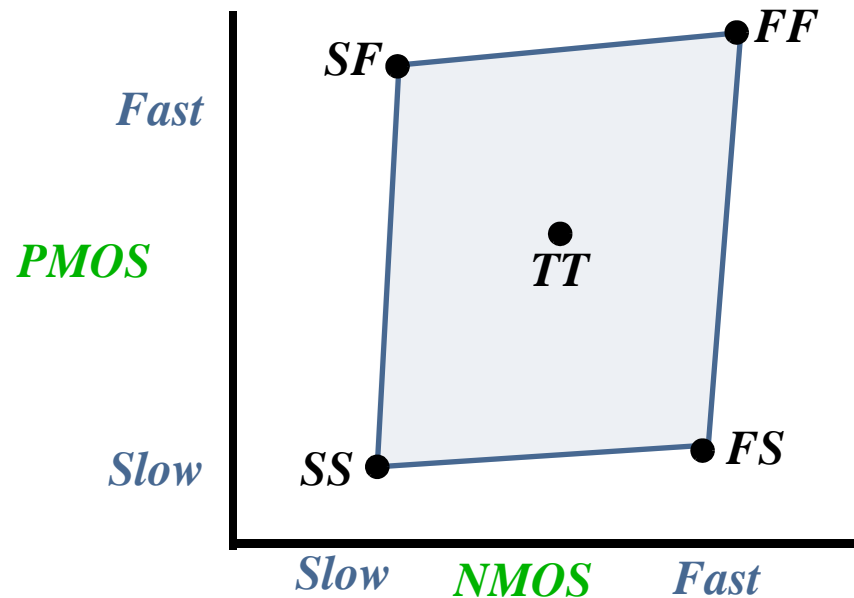
- *Lot to lot* (interprocess variation)
- *Wafer to wafer* (interprocess variation)
- *Die to die* (intraprocess variation)

### Design Corners

From the designer's point of view, the collective effects of process and environmental variations can be lumped into their effect on transistors:

- *typical (nominal)*
- *fast*
- *slow*

Speed of each type of transistors, interconnect speed variations and environmental variations are used to define *design* or *process corners*.

*Design Margin**Design Corners (contd.)**Environmental corners (1.8V process)*

<i>Corner</i>	<i>Voltage</i>	<i>Temperature</i>
<i>Fast (F)</i>	1.98	0°C
<i>Typical (T)</i>	1.8	70°C
<i>Slow (S)</i>	1.62	125°C

*Design Margin*

Design corner checks

<i>Corner</i>					<i>Purpose</i>
<i>NMOS</i>	<i>PMOS</i>	<i>Wire</i>	<i>V<sub>DD</sub></i>	<i>Temp</i>	
T	T	T	S	S	timing specifications (binned parts)
T	S	S	S	S	timing specifications (conservative)
F	F	F	F	F	DC power dissipation, race conditions, hold time constraints, pulse collapse, noise
F	F	F	F	S	subthreshold leakage noise, overall noise analysis
S	S	F	S	S	races of gates against wires
F	F	S	F	F	races of wires against gates
S	F	T	F	F	pseudo-NMOS & ratioed circuits noise margins, memory read/write, race of PMOS against NMOS
F	S	T	F	F	ratioed circuits, memory read/write, race of NMOS against PMOS

## Reliability

Reliability problems cause integrated circuits to fail permanently, including:

- Electromigration
- Self-heating
- Hot Carriers
- Latchup
- Overvoltage failure

### *Mean Time Between Failures (MTBF)*

# devices \* hours of operation / # failures

### *Failures in Time (FIT)*

The number of failures that would occur every thousand hours per million devices.

e.g. 1000 FIT is one failure in  $10^6$  hours = 114 years. (good for a single chip !!!)

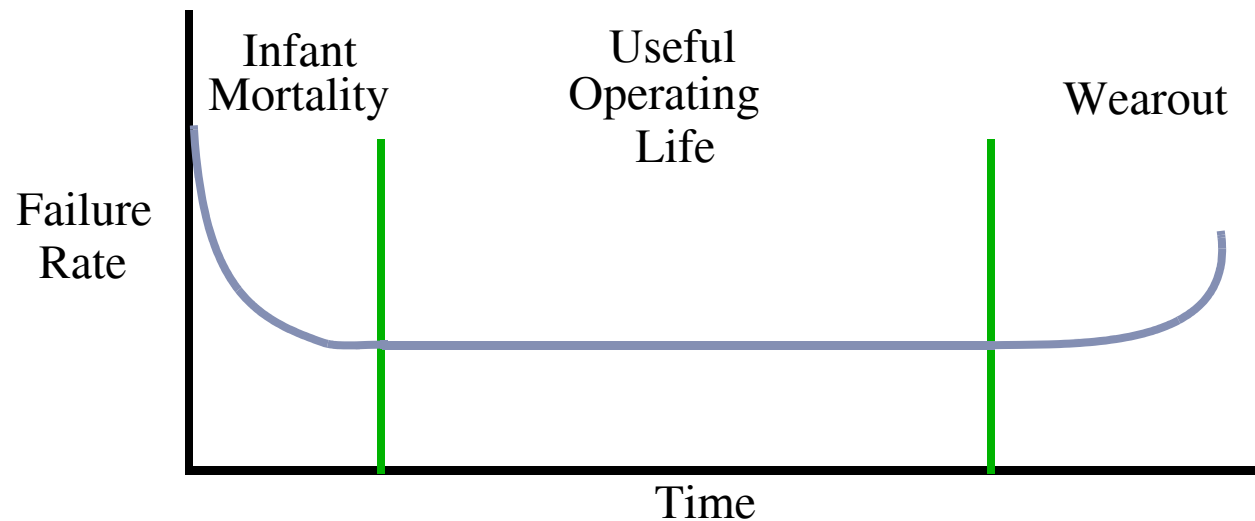
System with 100 chips each rated at 1000 FIT and you have 10 systems,

failure rate is  $100 * 1000 * 10 = 10^6$  FIT, or one failure every 1000 hours (42 days).

Need to target 100 FIT !!!

## Reliability

Most systems exhibit the bathtub curve.



Important to age systems past infant mortality (*burn-in*) before shipping products

## Electromigration

Causes wearout of metal interconnect through the formation of voids

High current densities lead to an 'electron wind' that causes metal atoms to migrate over time.

## *Reliability*

### *Electromigration (contd.)*

Depends on the current density  $J$ . Current limits are usually expressed as a maximum  $J_{dc}$

More likely to occur for wires carrying DC currents

Contact cuts have lower current density than metal lines

### *Self-heating*

Bidirectional wires are less prone to electromigration, their current density is limited by self-heating

High current dissipate power, raising in temperature and thus resistance and delay

Limited using reasonable values of  $J_{rms}$

In summary, electromigration is primarily a problem in power and ground lines, self-heating limits the RMS current density in bidirectional signal lines. Significant current flows through wire contacting NMOS and PMOS transistors and therefore needs consideration.



## *Reliability*

### *Hot Carriers*

As transistors switch, some high energy (hot) carriers may be injected into the gate oxide and become trapped there

Damaged oxides change I-V: increases current in PMOS and decreases current in NMOS

Hot carriers cause circuit wearout as NMOS transistors become too slow

Wear is limited by setting maximum values on input rise-time and stage electrical effort  
The maximum values depend on process and operating voltage

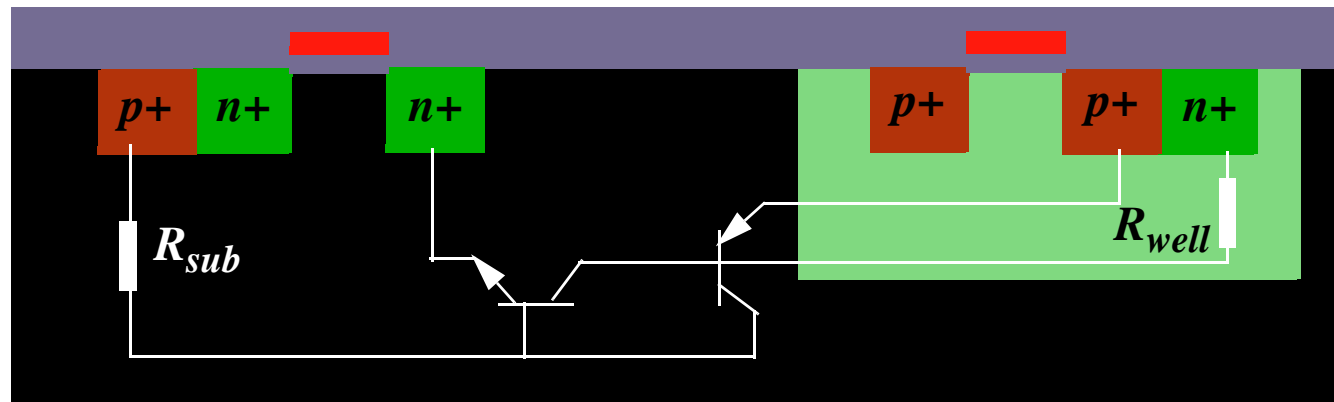
### *Latchup*

Parasitic bipolar transistors are formed by substrate, well and diffusion

If these transistors turn ON, it develops a low-resistance path between  $V_{DD}$  and GND, causing catastrophic meltdown, called latchup.

## Reliability

### Latchup (contd)



The cross-coupled transistors form a bistable silicon-controlled rectifier (SCR)

Ordinarily both transistors are off, but latchup can be triggered by transient current during normal chip power-up or external voltages outside the normal operating range

Latchup can be prevented by minimizing the two resistance values.

Can be accomplished by putting one tap (contact) per well, connecting substrate and well taps to the supply using metal lines, placing a tap per 5 transistors and clustering NMOS near GND and PMOS near  $V_{DD}$

## Reliability

### Overvoltage failures

Transistors can be easily damaged by overvoltage reliability problems due to:

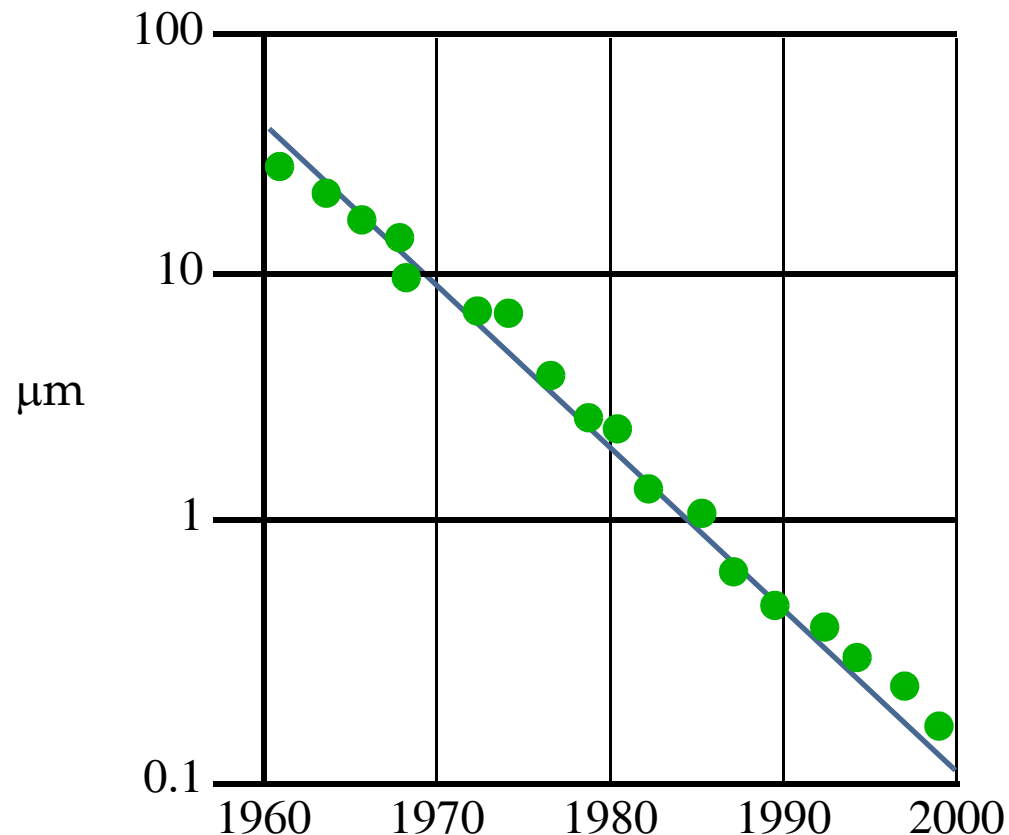
- **Electrostatic Discharge (ESD)**: Static electricity entering I/O pads can cause very large voltage and current transients
- **Breakdown and Arcing**: Undesired voltages applied to the gate can cause oxide breakdown, destroying the device
- **Punchthrough**: Higher than normal voltages applied between source and drain, can cause the source/drain depletion regions touch
- **Time-dependent Dielectric Breakdown (TDDB)**: Gate oxides wear out with time as tunneling currents cause irreversible damage to the oxide

### Soft Errors

Some errors are spontaneous and occur in random fashion. Known as soft errors and are mainly attributed to alpha particles (from decaying uranium, thorium etc used in the package). Refer to the book for more details.

## Scaling

Technology scaling rate is approximately 13%/year, halving every 5 years.



The size of the circuits also continues to increase.

Besides increasing the number of devices, scaling has had a profound impact on both speed and power.

## *Scaling*

### *Constant Field Scaling*

Critical parameters are scaled by a factor of  $S$ :

- All dimensions (in the  $x, y$  and  $z$  dimensions)
- Device voltages
- Doping concentration densities

### *Lateral Scaling*

Only the gate length is scaled. Commonly called a *gate shrink*.

Offers quadratic improvement according to first order model, but is close to linear improvement due to velocity saturation effects

### *Constant Voltage Scaling*

Feature size is shrunk keeping the supply voltage constant, providing quadratic improvement in delay as well as cost reduction. Worked for  $6\mu\text{m}$  to  $1\mu\text{m}$ .

*Scaling*

<i>Parameter</i>	<i>Sensitivity</i>	<i>Constant Field</i>	<i>Lateral</i>
Length: $L$ and Width: $W$		$1/S$ and $1/S$	$1/S$ and $1$
Gate oxide thickness: $t_{ox}$		$1/S$	$1$
Supply voltage: $V$ -and threshold voltage: $V_t$		$1/S$	$1$
Substrate doping $N_A$		$S$	$1$
$\beta$	$W/L * 1/t_{ox}$	$S$	$S$
Current: $I_{ds}$	$\beta(V_{DD}-V_t)^2$	$1/S$	$S$
Gate Delay: $\tau$	$RC$	$1 * 1/S = 1/S$	$1/S * 1/S = 1/S^2$
Clock frequency: $f$	$1/\tau$	$S$	$S^2$
Dynamic power dissipation (per gate): $P$	$CV^2f$	$1/S^2$	$S$
Chip area: $A$		$1/S^2$	$1$
Power density	$P/A$	$1$	$S$
Current density	$I_{ds}/A$	$S$	$S$

### *International Technology Roadmap for Semiconductors (ITRS)*

Developed by the *Semiconductor Industry Association (SIA)* to guide research efforts and predict future needs. Predictions from 2002 ITRS (high performance microprocessors):

<i>Year</i>	<i>2004</i>	<i>2007</i>	<i>2010</i>	<i>2013</i>	<i>2016</i>
<i>Feature Size (nm)</i>	<i>90</i>	<i>65</i>	<i>45</i>	<i>32</i>	<i>22</i>
V <sub>DD</sub> (V)	1-1.2	0.7-0.11	0.6-1.0	0.5-0.9	0.4-0.9
Millions of transistors/die	385	773	1564	3092	6184
Wiring levels	9-13	10-14	10-14	11-15	11-15
Intermediate wire pitch (nm)	275	195	135	95	65
Interconnect dielectric constant	2.6-3.1	2.3-2.7	2.1	1.9	1.8
I/O signals	1024	1024	1280	1408	1472
Clock rate (MHz)	3990	6739	11511	19348	28751
FO4 delays/cycle	8.4	6.8	5.8	4.8	4.7
Maximum power (W)	160	190	218	251	288
DRAM capacity (Gbits)	1	4	8	32	64

## *Impacts on Design*

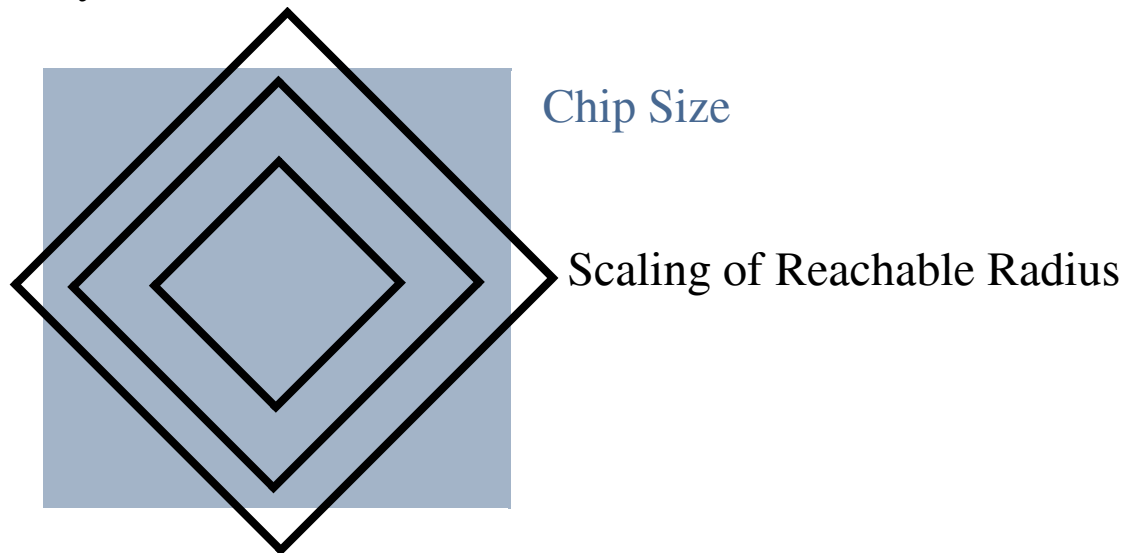
### *Improved Performance and Cost*

Positive impact of scaling: both are improving

When the transistor was invented, the prediction was that the price would eventually decrease to 50 cents a transistor. Today we can buy more than 100,000 for a penny !!!

### *Interconnect*

We have shifted to Cu and low-k dielectrics and design methodology has changed to specifically focus on interconnect delay





## *Impacts on Design*

### *Power*

Both dynamic and static power are predicted to increase.

Intel predictions of chip running with power density of a nuclear reactor in 2005, a rocket nozzle in 2010 and surface of sun in 2015 !!!

### *Productivity*

Number of transistors on chip is increasing faster than design productivity

Use EDA tools and make them more efficient while not degrading performance

Design teams approaching size of automotive and aerospace teams !!!

### *When will CMOS scaling end?*

Predictions (or fallacies):

1972: 0.25 $\mu$ m, 10-30 MHz.

1999: 100nm around 2004

2004: Most believe 2013, 35nm. **ANY BETS ???**