Design for Manufacturability
Design of Experiment

Total Yield

$$Y_{total} = Y_{line} \times Y_{batch}$$

Here Y_{line} denotes line yield or wafer yield which is the fraction of wafers which survive through the manufacturing line.

Y_{batch} is the fraction of integrated circuits which on each wafer fully functional at the end of the line.

Y_{batch} can be further classified based on either type of defect or of failure. Failure-type taxonomy is as follows.

Catastrophic Yield Loss. **These are functional failures such as open or short circuits which cause the part to not work at all. Extra or missing material particle defects are the primary causes for such failures.**

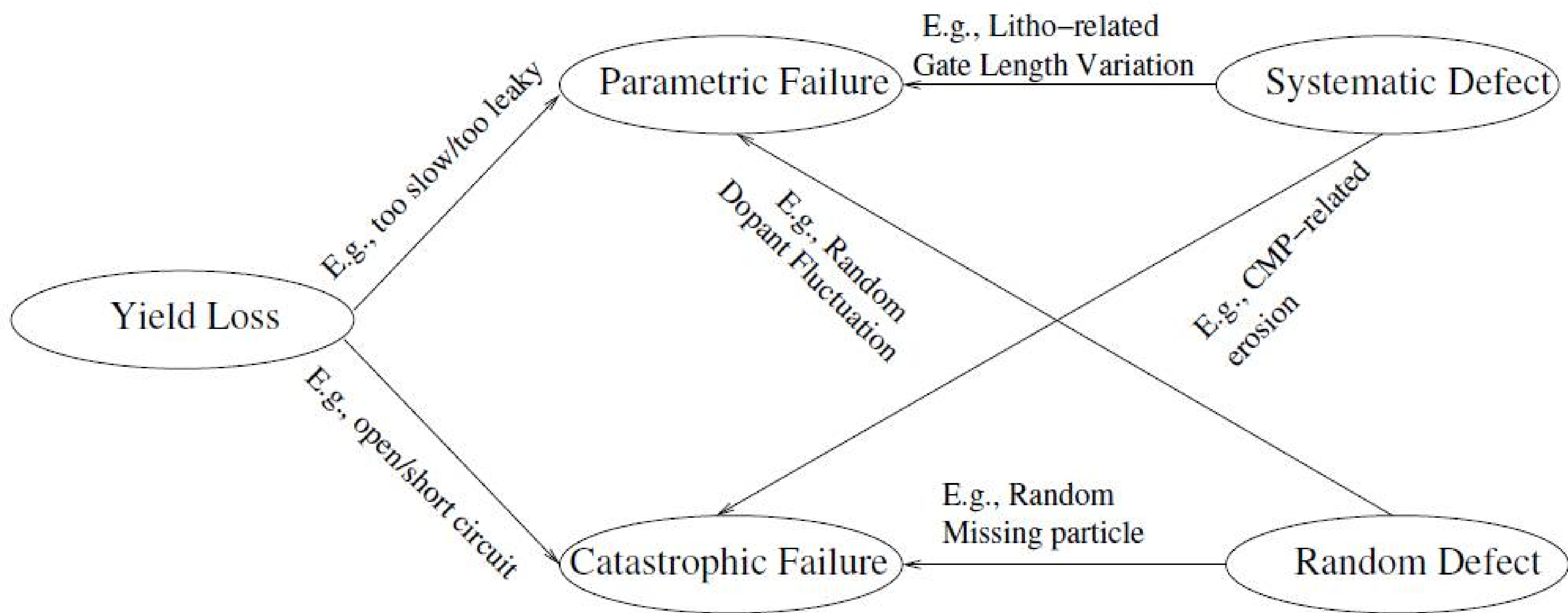
Parametric Yield Loss. **Here the chip is functionally correct but it fails to meet some power or performance criteria. Parametric failures are caused by variation in one or set of circuit parameters, such that their specific distribution in a design makes it fall out of specifications.**

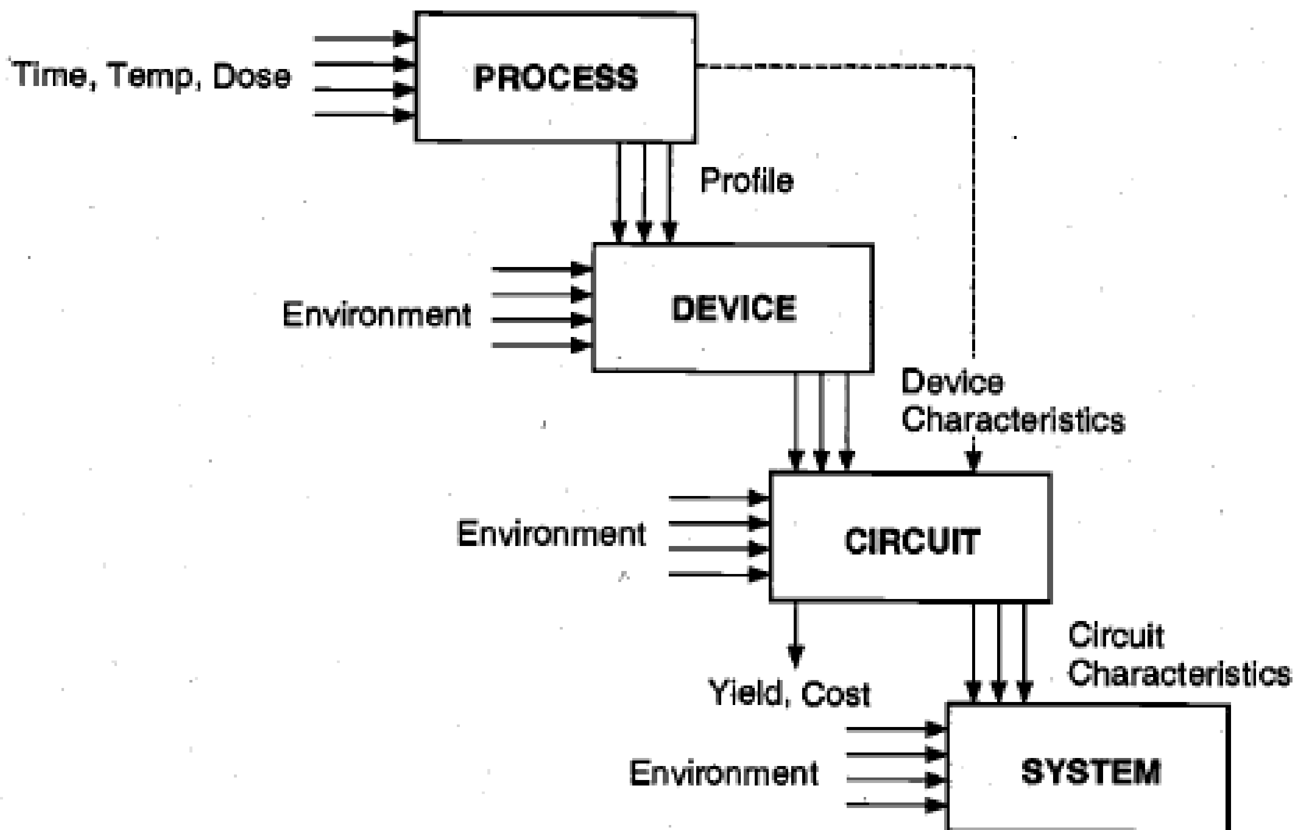
Defect

Defect types can be classified as follows [1].

- *Random Defects.* These are randomly distributed faults such as particle contamination.
- *Systematic Defects.* These kind of defects are predictable. Example sources include CMP (Chemical Mechanical Polishing) and photoresist pattern collapse.

It is important to understand that both random and systematic defects can cause parametric or catastrophic yield loss. For example, lithographic variation which is typically systematic and pattern dependent can cause catastrophic line-end shortening leading gate (polysilicon over diffusion) not forming and hence a functional failure.


Yield Loss
Parametric Failure
Catastrophic Failure
Systematic Defect
Random Defect
E.g., too slow/too leaky
E.g., open/short circuit
E.g., Litho–related
Gate Length Variation
E.g., Random
Dopant Fluctuation
E.g., CMP–related
erosion
E.g., Random
Missing particle



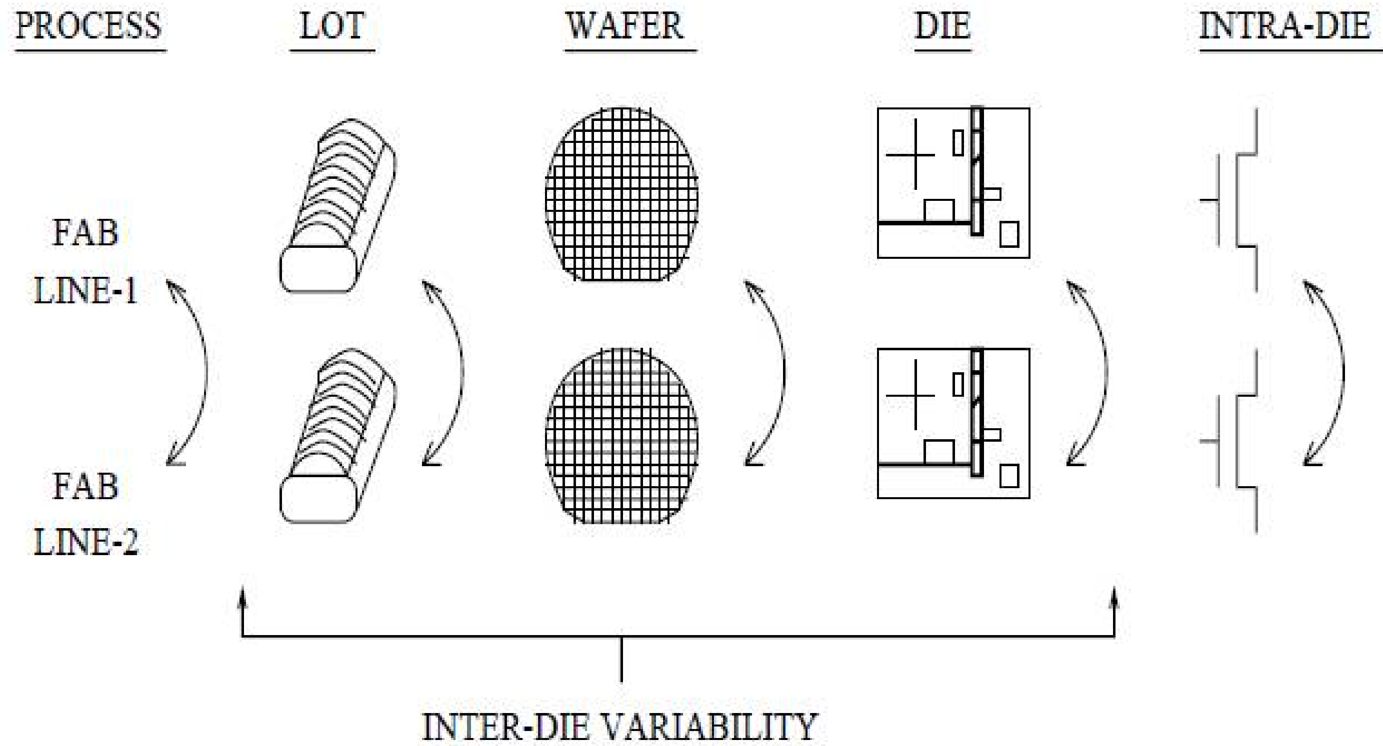
It is, therefore, important that the effect of these inevitable statistical variations in the processing and environmental conditions be considered early on during the design of the circuit. The circuit performance should be made least sensitive to these variations and should have enough margins that a large fraction of the manufactured circuits pass the acceptability criteria. **This is the essential motivation for *design for manufacturability* (DFM).**

- Process vs. Environmental. Variation occurring during circuit operation (e.g. temperature, power supply, etc) are environmental in nature while those occurring during the manufacturing process (e.g. mask misalignment, stepper focus, etc) are physical. We will focus only on process variations.

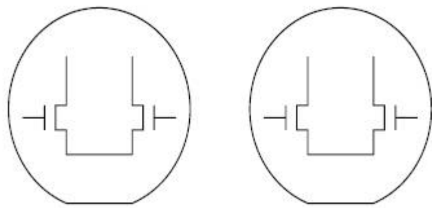
- Systematic vs. Random. As discussed earlier systematic variations (e.g. metal dishing, lithographic proximity effects, etc) can be modeled and predicted while random variations (e.g. material variations, dopant fluctuations, etc) are inherently unpredictable.

- Inter-die vs. Intra-die. Depending on the spatial scale of the variation, it can be classified as die-to-die (e.g. material variations) or within-die (e.g. layout pattern dependent lithographic variation). Inter-die variations correspond to variation of a parameter value across nominally identical die. Such variations may be die-to-die, wafer-to-wafer or even lot-to-lot. Inter-die variations are typically accounted for in design, by shift in the mean of a parameter value. Intra-die variations on the other hand correspond to parameter fluctuations across nominally identical circuit elements such as transistors. Intra-die perturbations are usually accounted in design by guardbanding

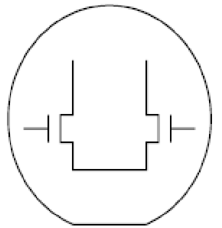
Building a statistical model requires knowledge of transistor parameter variance

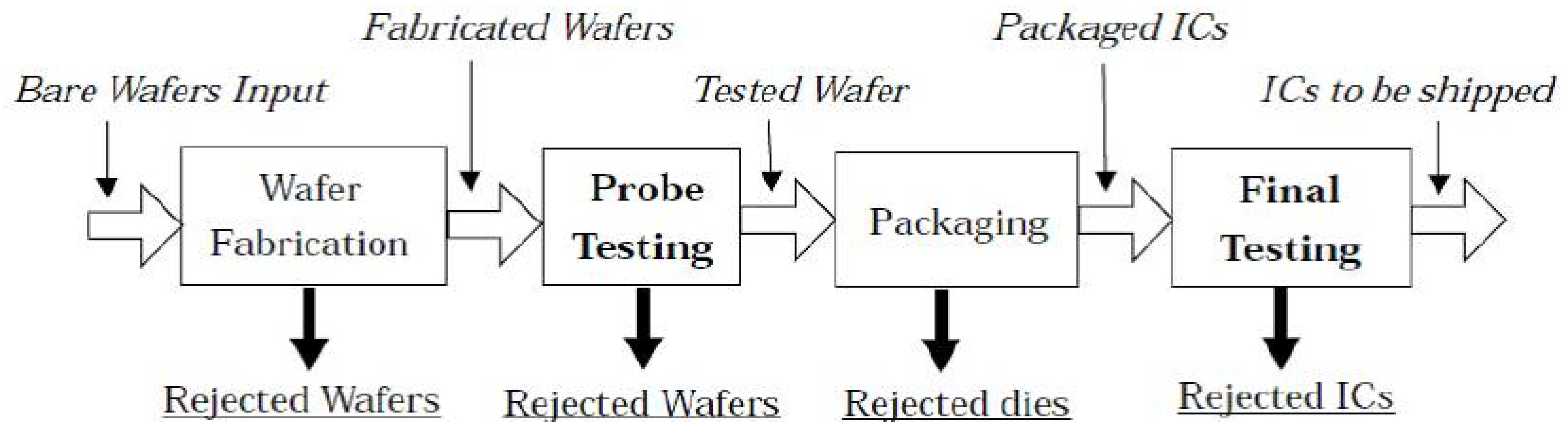


Inter-die parameter variation

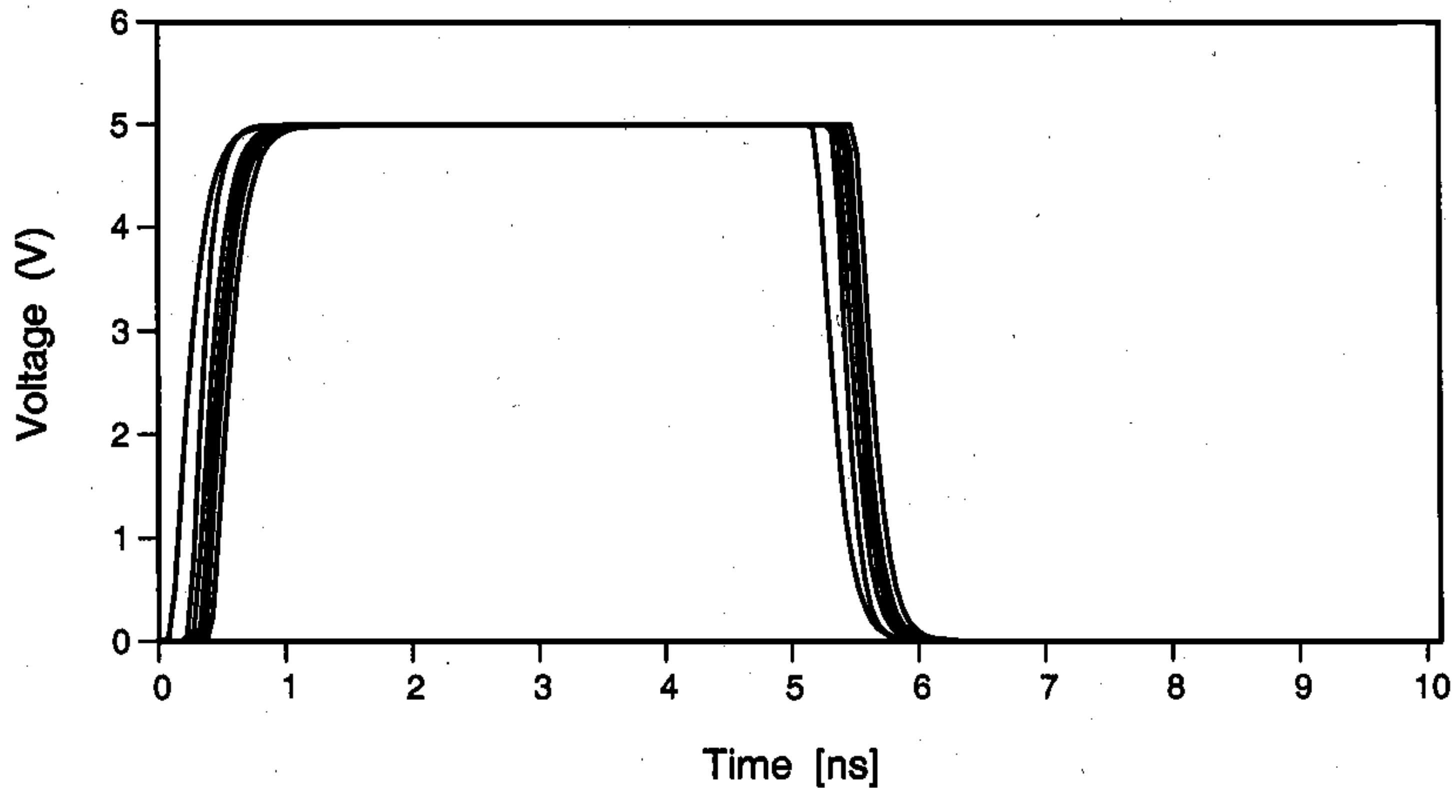


Intra-die parameter variation





- In the wafer fabrication process the structure of integrated circuits is sketched on the wafers and each of them is tested with the help of a probe in the probe testing stage.
- Once tested, the wafers are then cut (diced) into many pieces, with each piece containing a copy of a fully functional IC, these individual pieces are called a die.
- The dies that pass the test stage are packaged and sent for a final yield test before shipping.
- Faults or processing issues that may occur during any of these stages can cause some or all of the ICs on the wafers to malfunction.
- Such failures in ICs are detected at any of the two testing stages, probe testing or final testing.



It can be seen that the output varies significantly due to variations in device parameters caused by process fluctuations

$$I_d = \mu C_{ox} \frac{W}{L} f(V_{DS}, V_{GS}, V_T)$$

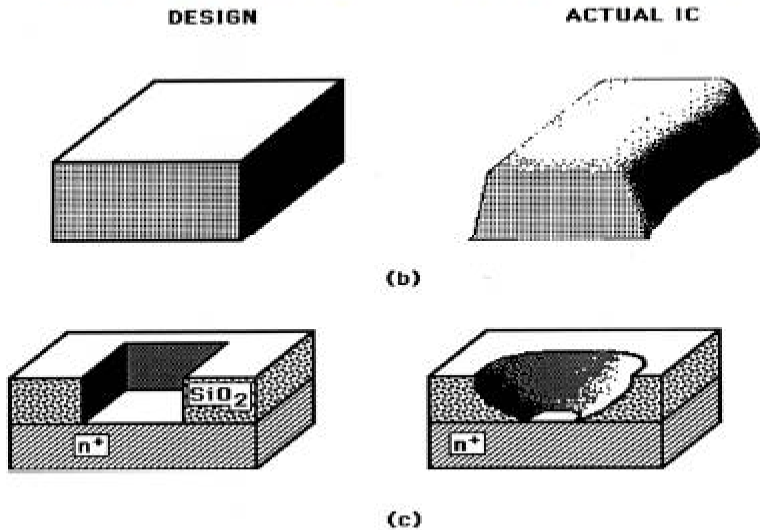
μ is the mobility of electrons in nMOS (holes in pMOS),
$C_{ox} = \varepsilon_{ox}/t_{ox}$ is the gate oxide capacitance per unit area,
W/L is the ratio of the channel width to channel length, and
V_T is the threshold voltage of the transistor.

Process Variability Causes Deformations

- Geometrical
  - Lateral
  - Vertical
  - Spot defects
- Electrical
  - Global
  - Local

Deformations have *deterministic* and *random* components, are *global* and/or *local*, can be *independent* or can *interact*.

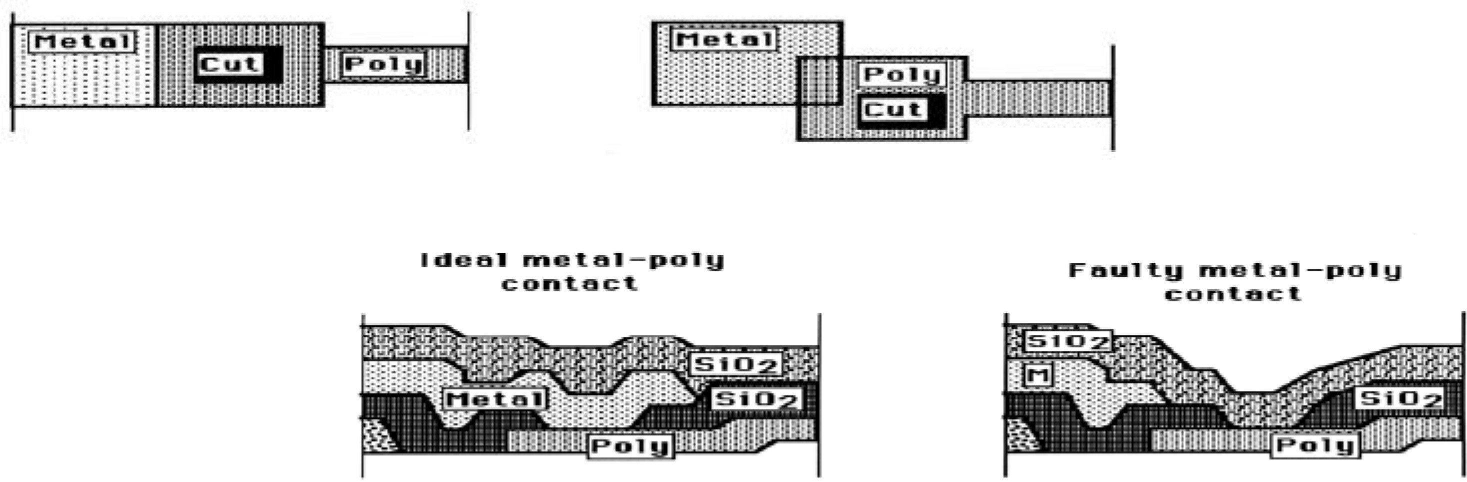
Deformations of Ideal Design



What limits Functional Yield?

- Gross Misalignments
- Particles
- Mask Defects

Mask Misalignment



Circuit Parameters

Due to random variations in manufacturing processes and operating conditions, it is expected that the actual values of circuit parameters will be different from their nominal or target values. For instance, the actual channel width W of a MOS transistor can be decomposed into a statistically varying component ΔW and a nominal component W^o, i.e., $W = W^o + \Delta W$. In general, any circuit parameter can be considered to have a nominal component and an uncontrollable statistically varying component as shown in Table 15.1.

	Actual	=	Nominal	+	Random
Geometrical Parameters					
MOS channel width	W	=	W^o	+	ΔW
MOS channel length	L	=	L^o	+	ΔL
Device Model Parameters					
Threshold voltage	V_T	=	V_T^o	+	ΔV_T
Gate oxide thickness	t_{ox}	=	t_{ox}^o	+	Δt_{ox}
Mobility	μ	=	μ^o	+	$\Delta\mu$
Operating Conditions					
Power supply voltage	V_{DD}	=	V_{DD}^o	+	ΔV_{DD}
Temperature	T	=	T^o	+	ΔT

In this table, the geometrical parameters have a nominal component which can be set to particular values by the circuit designer. Such a nominal component is said to be *designable* or *controllable*, e.g., W^o and L^o. The statistically varying component of the geometrical parameters is called the *noise* component, and it represents the uncontrollable fluctuation of a circuit parameter about its designable component, e.g.,

$$x_i = d_i + s_i$$

where d_i is the designable component and s_i is the random noise component. For circuit parameters which do not have a designable component, d_i is set to zero. Similarly, for circuit parameters which are completely controllable, s_i is set to zero.

For the device model parameters and operating conditions, the nominal component is not under the control of the designer and is set by nominal processing and operating conditions. For these parameters, the nominal and random components are together called the noise component, e.g., *VT* and *VDD.*

Other commonly used terminology classifies the noise parameters based on whether their variation is related to fluctuations in manufacturing processes or operating conditions. The former are referred to as *internal noise parameters* and the latter as *external noise parameters*. For instance, V_T is considered to be an internal noise parameter, while V_{DD} is an external noise parameter.

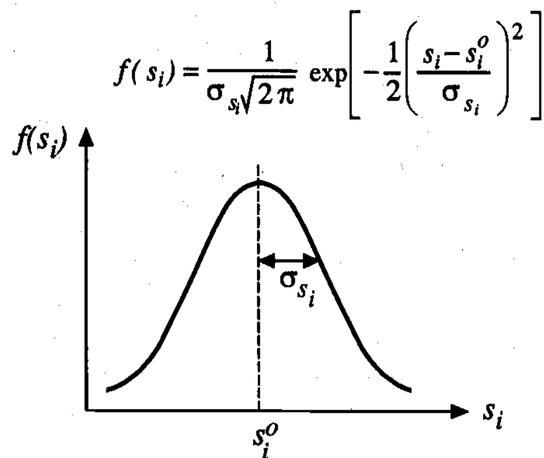
Noise Parameter Distributions

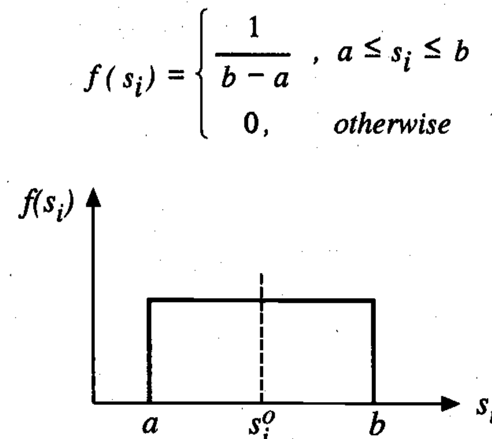
The statistical distributions of the internal noise parameters can be obtained via test structure measurements and parameter extraction.

To simplify the analysis, a commonly used assumption is that the **internal noise parameters have a Gaussian distribution**, while the **external noise parameters are uniformly distributed random variables**

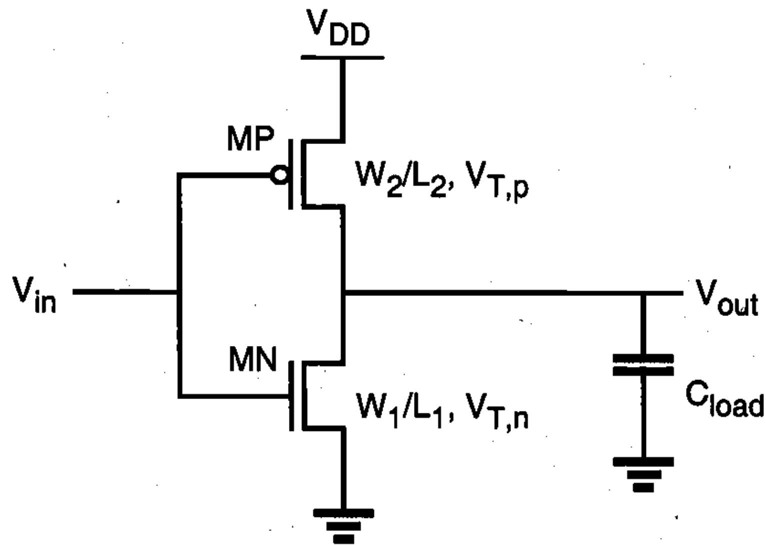
The internal noise parameters are *statistically correlated* due to the sequential nature of the processing steps.

The external noise parameters are, however, *statistically independent* random variables.





A Typical example



following assumptions about the circuit parameters

(i) The widths and lengths of the MOS transistors are fixed and are not subject to statistical variations, i.e., W_1, L_1, W_2, and L_2 are designable parameters.

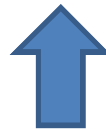
(ii) The internal noise parameters are the threshold voltages of MN and MP, $V_{T,n}$ and $V_{T,p}$, and the common gate oxide thickness t_{ox}. We assume for simplicity that these random variables are uncorrelated (independent) and Gaussian.

(iii) The external noise parameters are the power supply voltage V_{DD}, which is uniformly distributed in the range [4.8 V, 5.2 V], and the operating temperature T, which is uniformly distributed in the range [30 °C, 90 °C]. Moreover, V_{DD} and T are considered to be independent random variables.

Now Performance measure : propagation delay

$$\tau_{PHL} = \frac{C_{load}}{k_n(V_{DD} - V_{T,n})}\left[\frac{2V_{T,n}}{V_{DD} - V_{T,n}} + \ln\left(\frac{4(V_{DD} - V_{T,n})}{V_{DD}} - 1\right)\right]$$

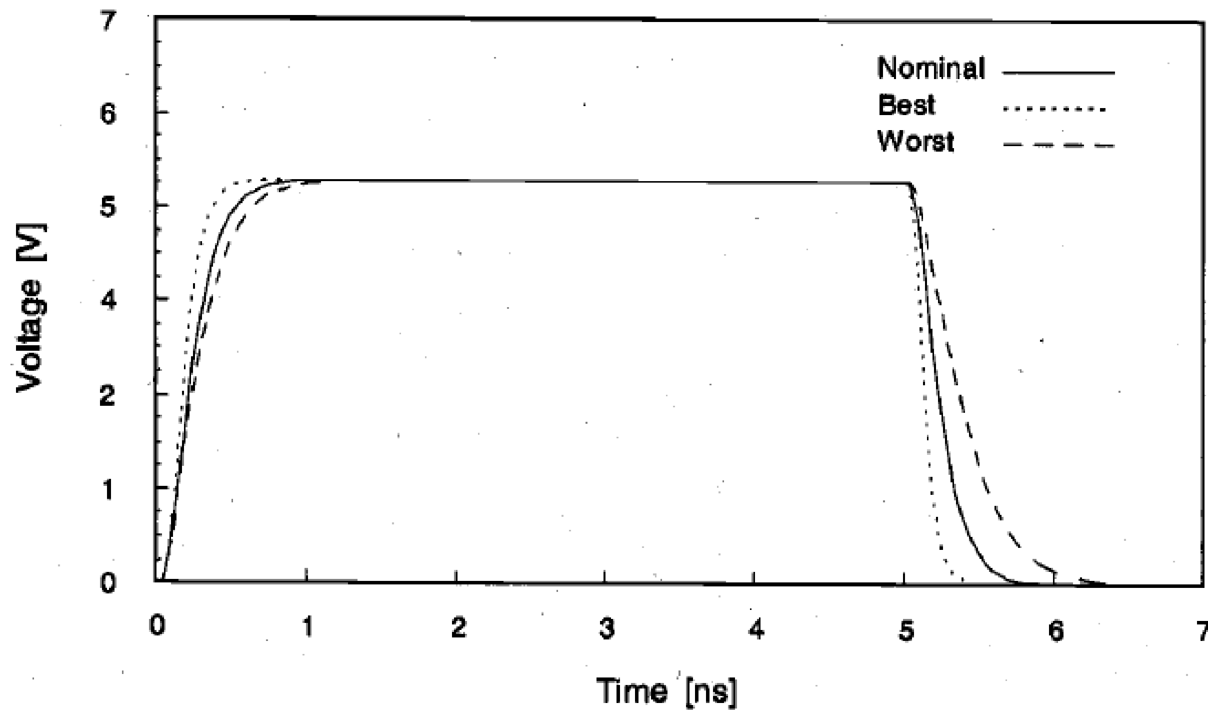
$$\tau_{PLH} = \frac{C_{load}}{k_p\left(V_{DD} - |V_{T,p}|\right)}\left[\frac{2|V_{T,p}|}{V_{DD} - |V_{T,p}|} + \ln\left(\frac{4\left(V_{DD} - |V_{T,p}|\right)}{V_{DD}} - 1\right)\right]$$

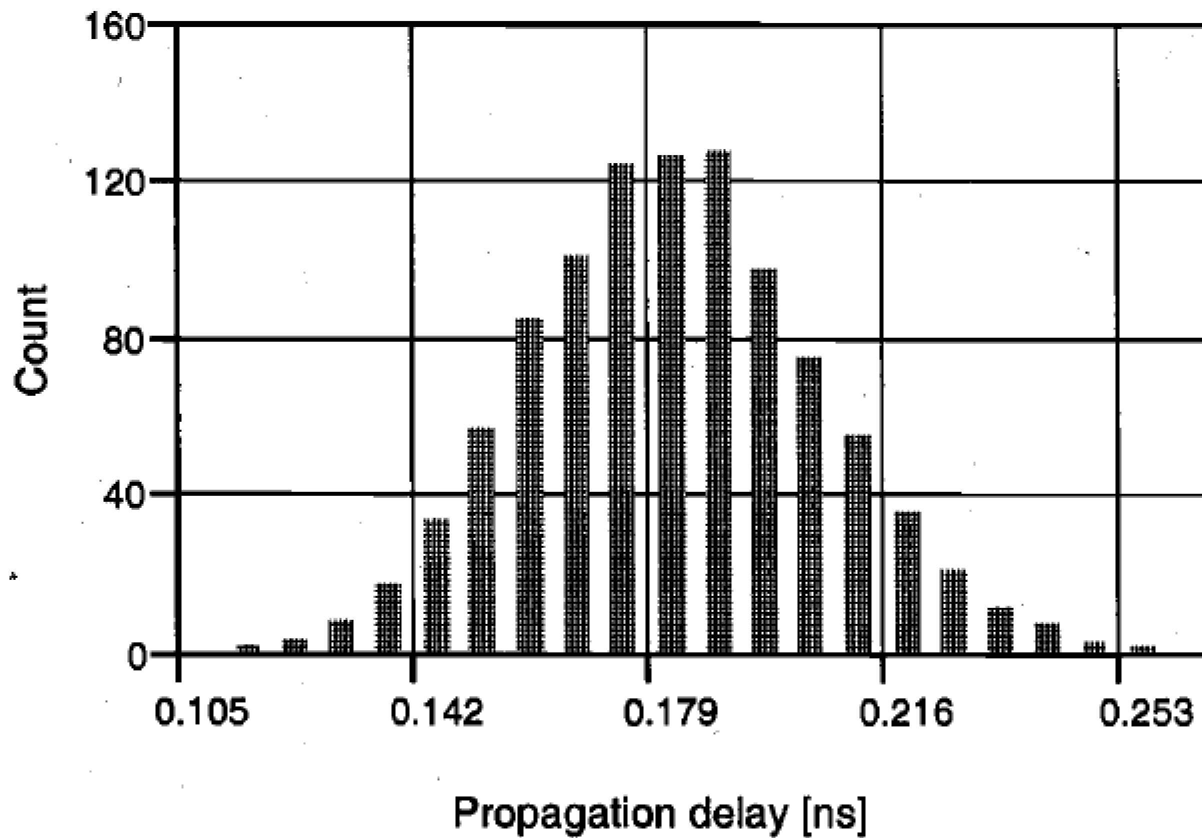


$$r = r(\mathbf{d} + \mathbf{s}) = r(\mathbf{x})$$

When the internal and external noise parameters are assumed to be fixed at their mean values, the corresponding circuit performance value is called the *nominal value*. Since it depends on the designable parameters alone, the nominal value of a performance r is denoted by

$$r^o(\mathbf{d}) = r(\mathbf{d} + \mathbf{s}^o)$$





Histogram showing variation of τ_p with $V_{T,n}$, $V_{T,p}$, and t_{ox}.

The circuit performance measure, which is a function of the random circuit parameters, is also a random variable. Therefore, a performance measure will have a mean value and a standard deviation.

Device parameter	Relevant process step
μ	Ion implantation, diffusion , annealing, stress
C_{ox}	Gate oxide formation
W,L	Etching, lithography
V_{th}	Ion implantation, gate oxidisation, annealing, etching, lithography

Mobility refers to the ability of the carriers (electrons or holes) to travel through the channel of a MOSFET in response to an applied electric field. It can be mathematically expressed as in Equation 2.6.

$$\mu = \frac{q\tau}{m^*}$$

where q is the electron charge, τ is the mean free time between carrier collisions and m is the effective mass of carriers (electron and hole). The mobility of carriers in the channel of a MOSFET device is also given as a function of the doping concentration

Gate Oxide Capacitance (C_{ox}) is the capacitance that is formed by the silicon oxide between the polysilicon in the gate stack and the channel of the MOSFET. Equation shows that the oxide capacitance is determined only of the oxide thickness (t_{ox}) and the dielectric constant of silicon dioxide or other gate insulator.

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$$

The formation of gate oxide using thermal growth of silicon dioxide or silicon nitride is a relatively well-controlled process step during device fabrication.

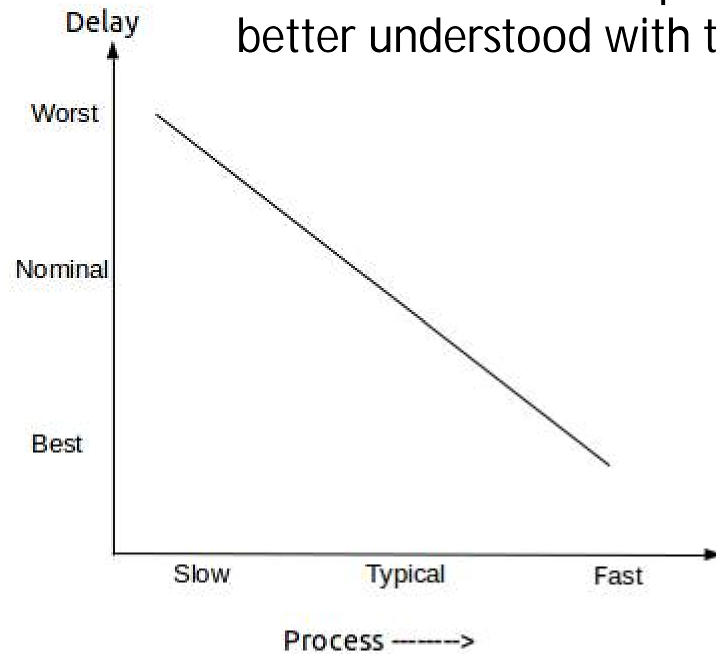
Threshold Voltage

The threshold voltage of a MOSFET is defined as the gate-to-source voltage (V_{GS}) that is required to form an inversion layer in the channel at the interface between the gate oxide and the silicon surface under the gate, thus allowing a current to flow from the source to drain terminals of the transistor. The threshold voltage is one of the key device parameters in CMOS technology, since it allows transistors to act like switches and hence is a suitable device to perform logic operations.

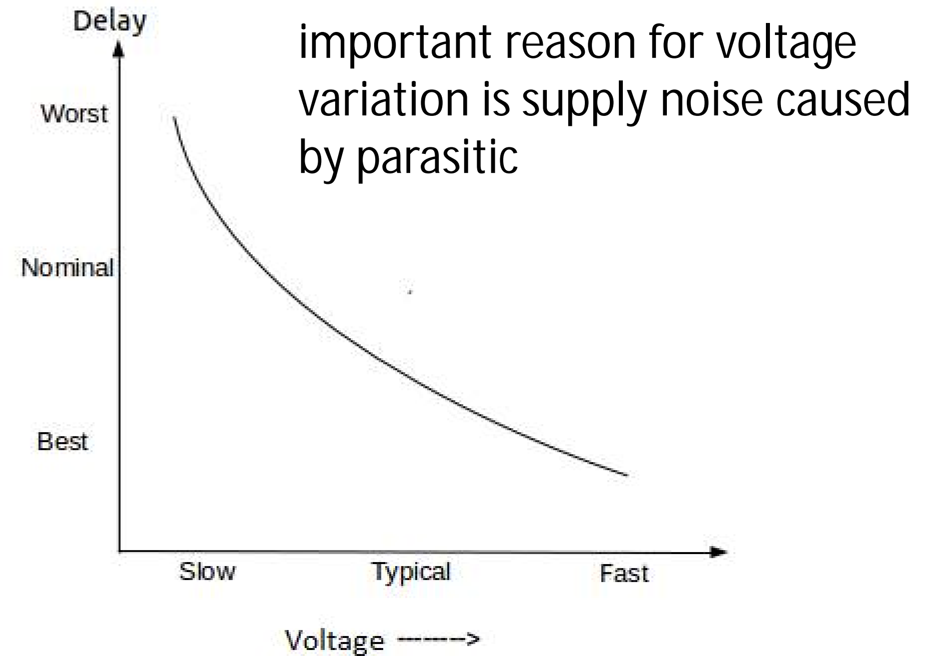
PVT

- PVT is abbreviation for Process, Voltage and Temperature. In order to make our chip to work in all possible conditions, like it should work in -40°C and also at 60°C, we simulate it at different corners of process, voltage and temperature which IC may face after fabrication. These conditions are called as corners. All these three parameters affect the delay of the cell.
- **Process:**
- Process variation is the deviation in attributes of transistor during the fabrication.
- **Voltage:**
- Now a days, supply voltage for a chip is very less. Lets say chip is operating at 1V. So there are chances that at certain instance of time this voltage may vary. It can go to 1.1V or 0.9V. To take care of this scenario, we consider voltage variation.
- **Temperature:**
- The temperature variation is with respect to junction and not ambient temperature.

the relation between process and delay can be better understood with the following curve



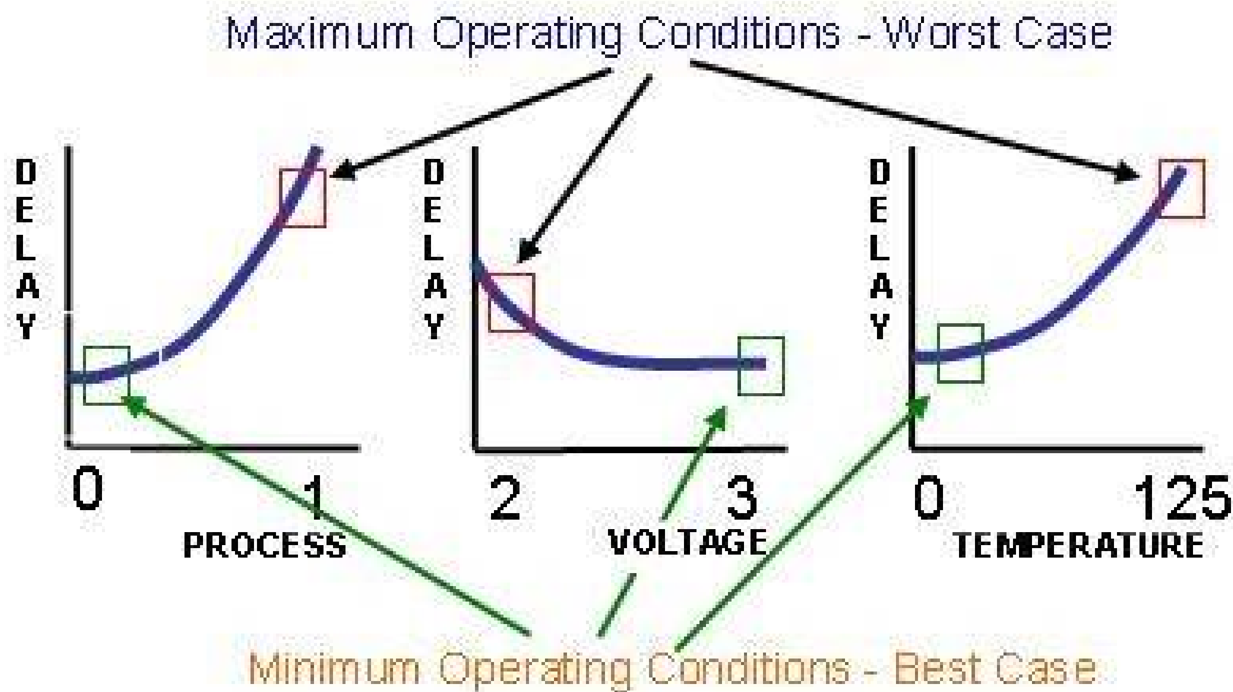
important reason for voltage variation is supply noise caused by parasitic



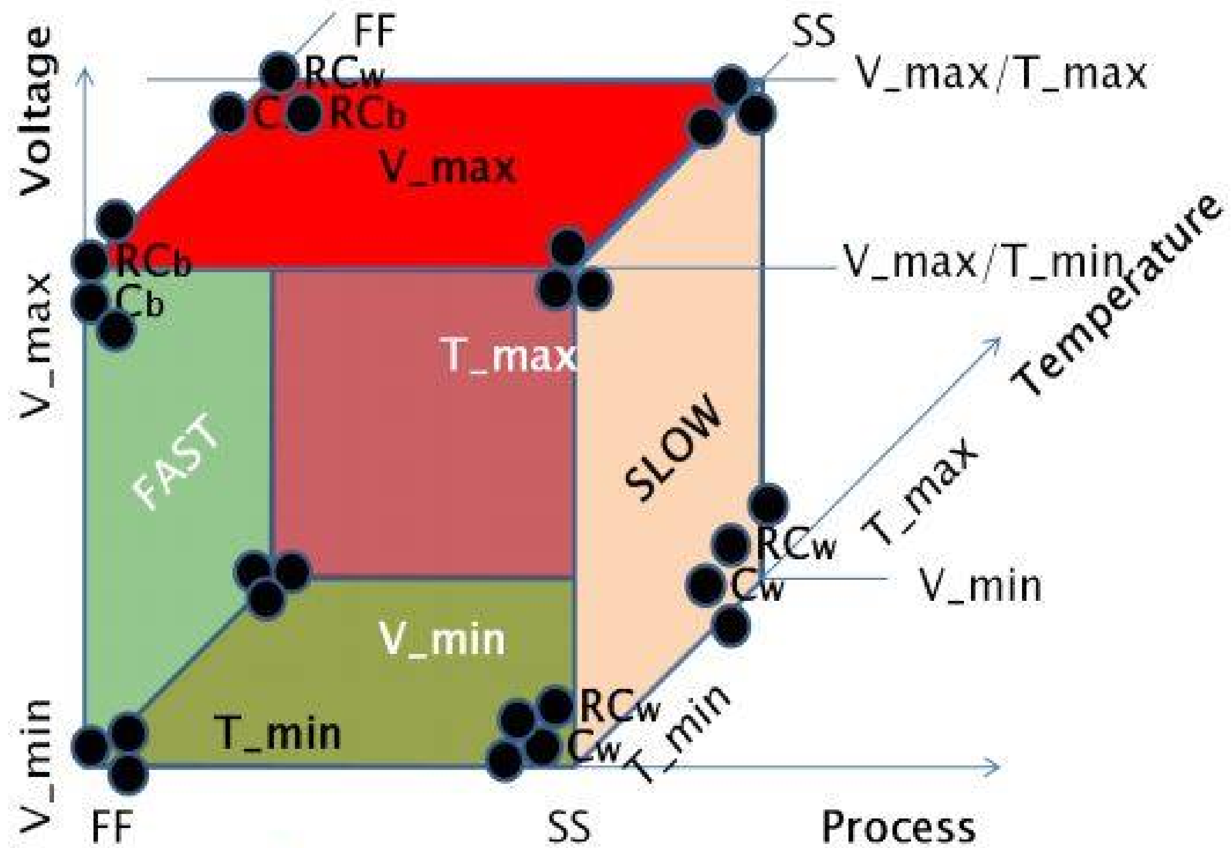
<u>On Chip Variations (OCV):</u>

Variations are of two types:

1. <u>Global variations</u>
2. <u>Local Variations</u>



Corners: A corner is a *point* in variation space. For example, a traditional PVT corner had a modelset value, a voltage value, and a temperature value, such as $\{modelset = FF,\ v_{dd} = 1.3\ V,\ T = 15\ °C\}$.



Voltage
V_max
V_min
FF
SS
RCw
C
RCb
Cb
V_max/T_max
V_max/T_min
V_max
T_max
FAST
SLOW
V_min
T_min
Temperature
T_max
RCw
Cw
V_min
T_min
Process
Examples of PVT/RC Corners
Via corners are not shown to make illustration simpler to view

Process Corner

- From the designer's point of view, the collective effects of process and environmental variation can be lumped into their effect on transistors: *typical* (also called *nominal*), *fast*, or *slow*.
- In CMOS, there are two types of transistors with somewhat independent characteristics, so the speed of each can be characterized. Moreover, interconnect speed may vary independently of devices. When these processing variations are combined with the environmental variations, we define *design* or *process corners*.

Manufacturing Stack

- FEOL* (Front End of Line)
  - First portion of IC fabrication
  - CMOS fabrication steps to form transistors:
    - Selecting type of wafer*
    - Cleaning of the wafer
    - Well formation
    - Gate, Drain and Source module formation
- BEOL* (Back End of Line)
  - Second portion of IC fabrication
  - Devices get interconnected with wiring
  - BEOL includes
    - Vias
    - Insulating layers (dielectrics)
    - Metal layers
    - Bonding sites (chip2package connections)

Timing Corners

FEOL	BEOL	Voltage	Temp
SS	C_w	Min	Min
TT	C_b	Typ	Typ
FF	RC_w	Max	Max
SF	RC_b		
FS	RC_{typ}		

1. **FEOL** (Front end of the line) - these refer to the fabrication of the active and passive elements of the circuit. These are the resistors (or conductors), capacitors, diodes, and transistors that make up the various elements of the IC.

Types of Variations

- Parameters of transistors vary from:
  - Lot to lot (Inter process variation*)
  - Wafer to wafer (Inter process variation)
  - Die to die (Intra process variation)
  - Transistor to transistor (On Chip Variation)

2. **BEOL** (Back end of the line) - these are the metallic layers that are used to make the interconnections between the various components fabricated in FEOL and also to the connections for the external devices.

- Chip area increasing (greater integration)
- Device dimensions scaled down
- Scaled wires are
  - Longer (chip area scaling)
  - Thinner (minimum dimension scaling)
  - Taller (Do not increase resistance)

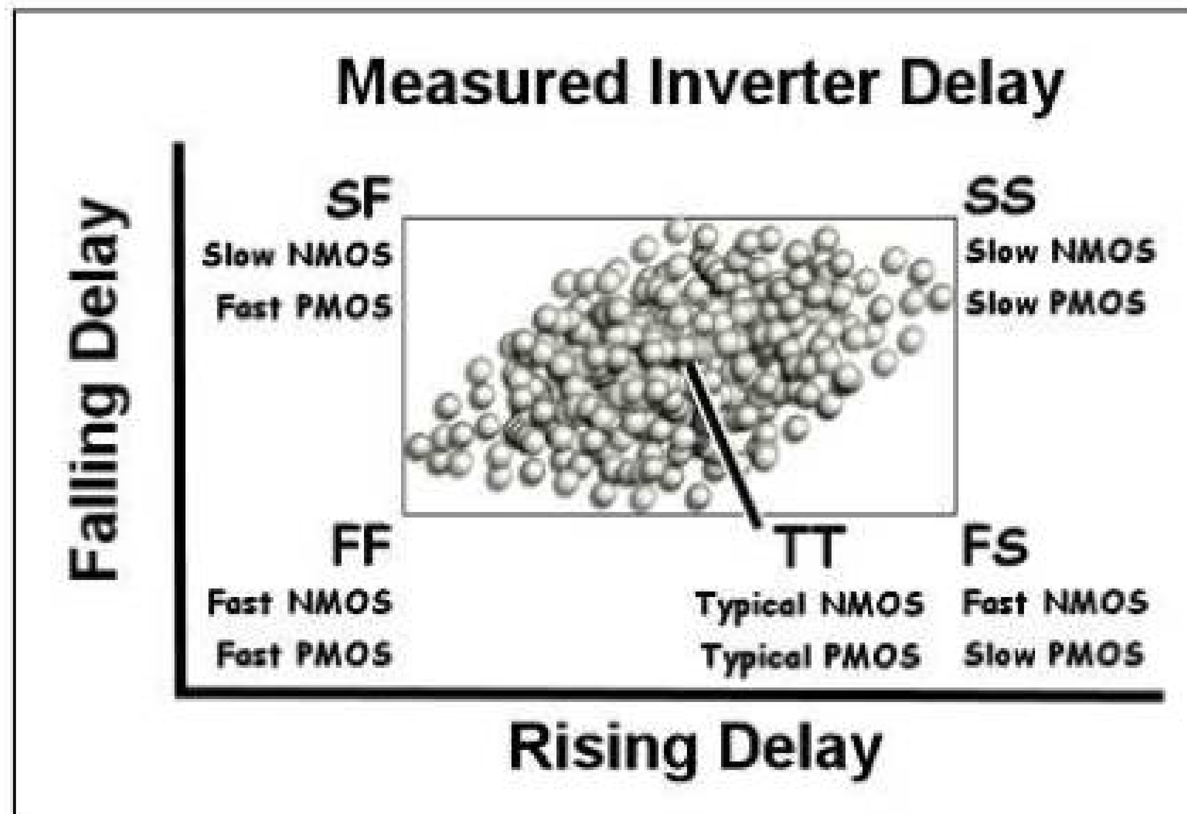
- Wire delay* limiting performance

Process convention

Industry uses two-letters to describe corners:

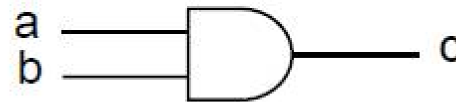
- 1st letter → NMOS device
- 2nd letter → PMOS device

- 5 classic corners:
  - FF (fast fast)
  - SF (slow fast)
  - SS (slow slow)
  - FS (fast slow)
  - TT (typical typical)

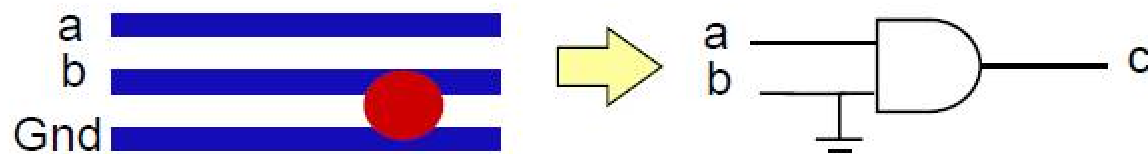


Defects

Consider one two-input AND gate



Defect: a short to ground

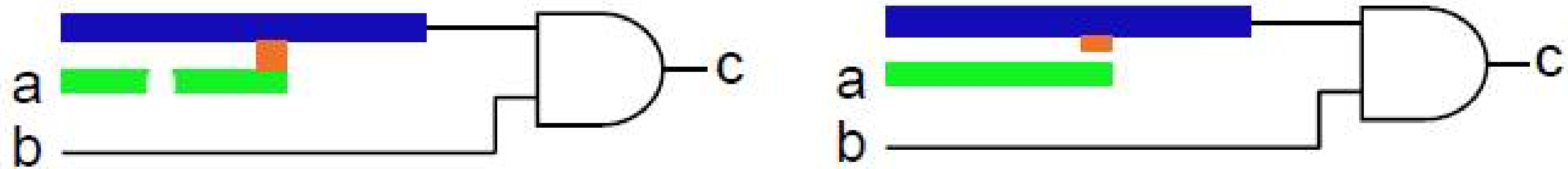


Fault: signal b stuck at logic 0

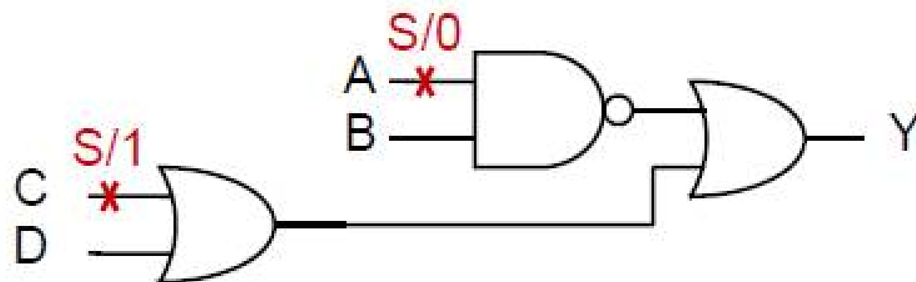
Error: a=1, b=1, c=0 (correct output c=1)

Note that the error is not permanent. As long as at least one input is 0, there is no error in the output

- Different types of defects may cause the same fault



- Different types of faults may cause the same error
  - E.g., A stuck-at-0, Y=1; C stuck-at-1, Y=1

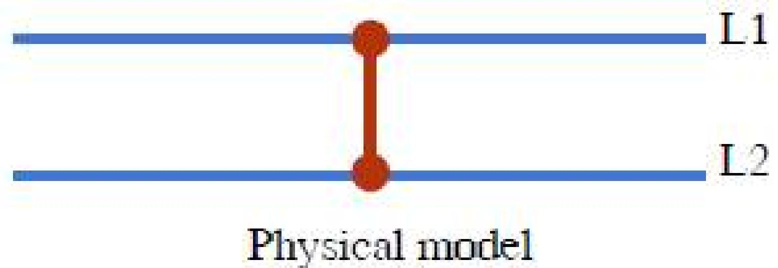


Failures in Integrated Circuits

Failure mode is used in reference to the manifestation of a ***defect*** at the electrical level.

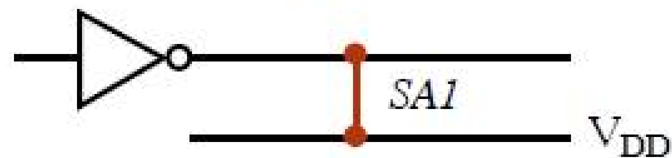
Failure modes are modeled as ***faults*** at logic or behavioral level of abstraction

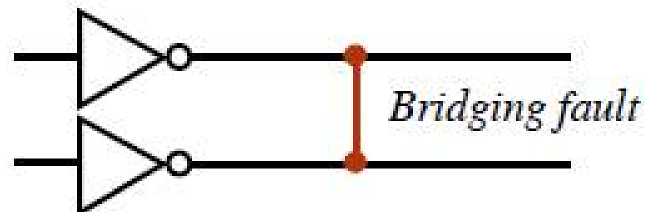
Physical defect



Physical model

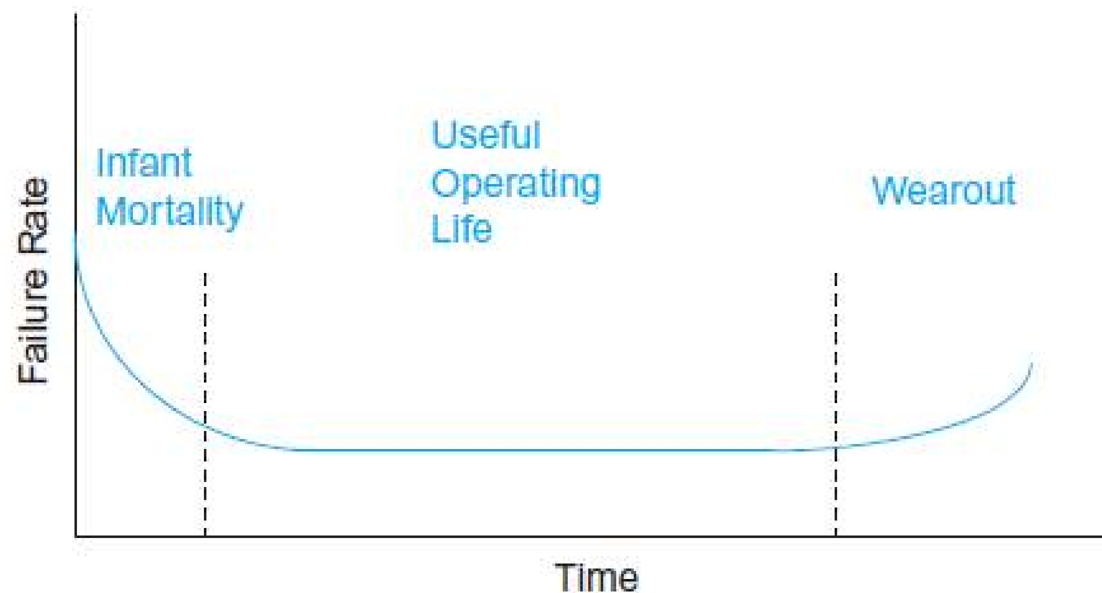
At the logic level, failure mode can be interpreted in different ways.





Reliability

- Infant mortality : Early failure in life :: Most IC product follow some initial failure as a function of its versions. Failure rate decreases rapidly to a low value with more iteration.
- Bathtub curve : Failure will gain increases with end of life. End of product lifetime --- failure is high.
- Overall reliability of any IC chip can be measured by using the mean time between failure (MTBF) for a repairable product or mean time to failure (MTTF) for a fatal failure.
- Also uses failure in time (FIT) where 1 FIT equals a single failure in 10^9 hours.



- MTBF is the *mean time between failures*: (number of devices × hours of operation) / number of failures.
- FIT is the *failures in time*, the number of failures that would occur every thousand hours per million devices

Fault Model

- Mapping from Physical fault to logical fault.
- Distinguish among different logical fault.
- Three types of logical fault
  - Degradation fault
  - Open circuit fault
  - Short circuit fault.
- Degradation fault of two types
  - Parametric fault
  - Delay fault.

Cont.

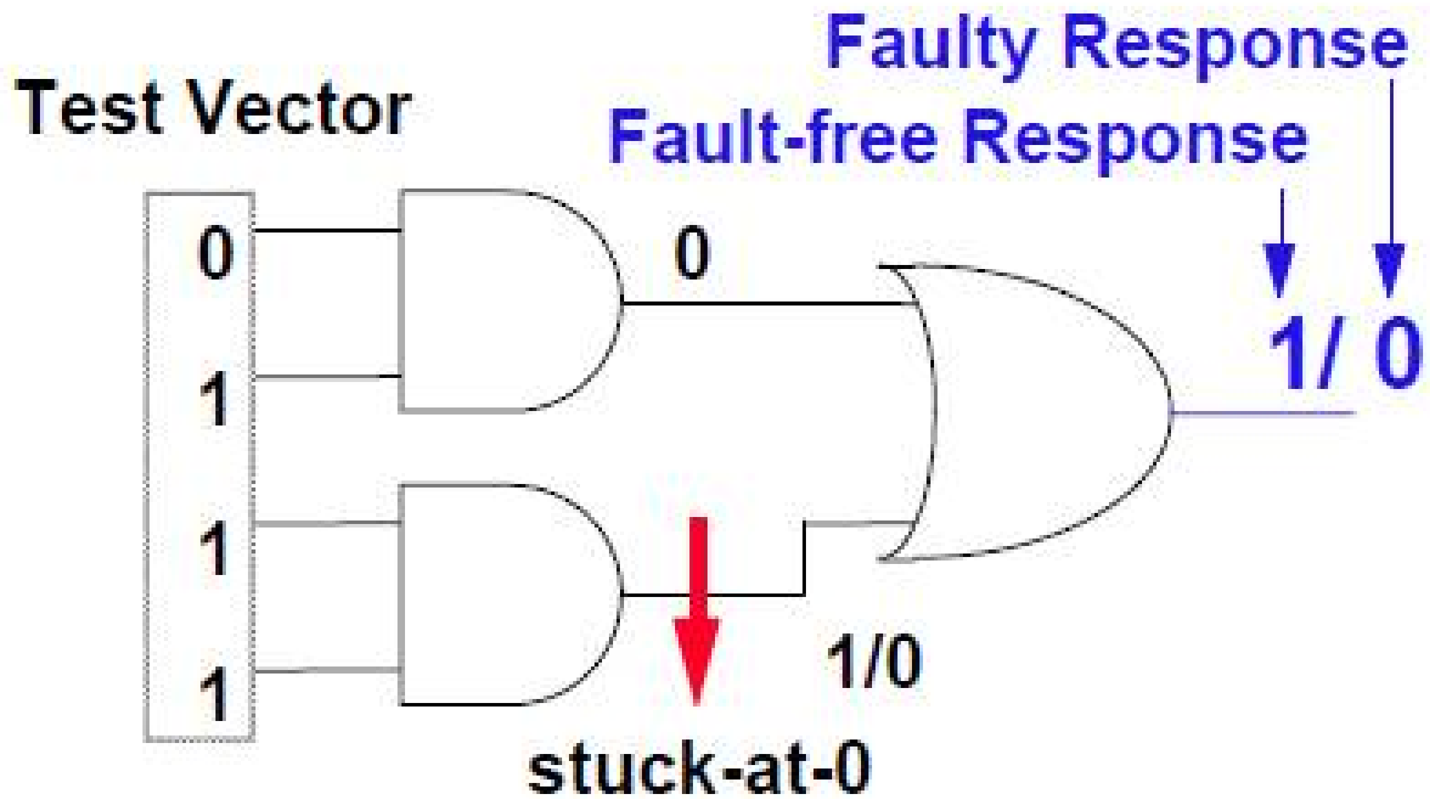
- Parametric fault leads to incorrect switching value/ threshold value
- Delay fault leads to critical path
- Open circuit or short circuit fault provides wrong logic level transfer from one end to other.
- Most short circuit fault occurs due to bad interconnections – often called as bridging fault.
- Bridging fault can be feedback type or non-feedback type.

Single stuck at fault

- It is a one fault model.
- Can be used as separate entity in multiple stuck at fault model.
- There are some eqv model like stuck on fault like stuck open / stuck off fault model.

Why Single Stuck-At Fault Model?

- **Complexity is greatly reduced.**
  Many different physical defects may be modeled by the same logical single stuck-at fault.
- **Single stuck-at fault is technology independent.**
  Can be applied to TTL, ECL, CMOS, etc.
- **Single stuck-at fault is design style independent.**
  Gate Arrays, Standard Cell, Custom VLSI



Faulty Response
Test Vector
Fault-free Response
0
1
1
1
0
1/ 0
1/0
stuck-at-0

Cont.

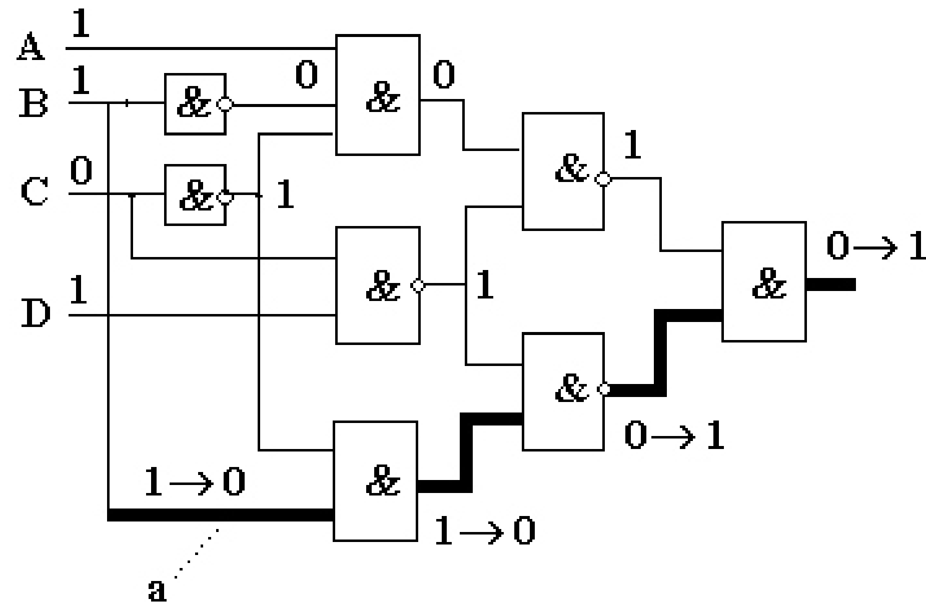
- Net fault : forces all the logic cell inputs either to logic 0 or logic 1.
- Input fault : attached to a logic cell input forces the input to logic 1 or 0. But it is confined to a particular logic cell and may not affects other.
- Output Fault: If it is a rail strength fault--- all the logic cell output connected to supply rail will force to either 0->1 or 1->0.

Cont.

- Generally Stuck at fault injected at input or output pins of logic cell.
- We do not injected this to the internal nodes like internal pins of flip flop etc.
- We call the earlier fault as pin fault model/ structural fault model/ gate level model.
- If internal fault is applied then it is transistor / switch level model.
- Fault effects travel through the circuit to other logic cells causing other constitutive faults – This is called fault propagation.
- Fault collapsing : same output due to multiple fault. Can be minimized to a single model.

Fault Detection in Combinational Circuits

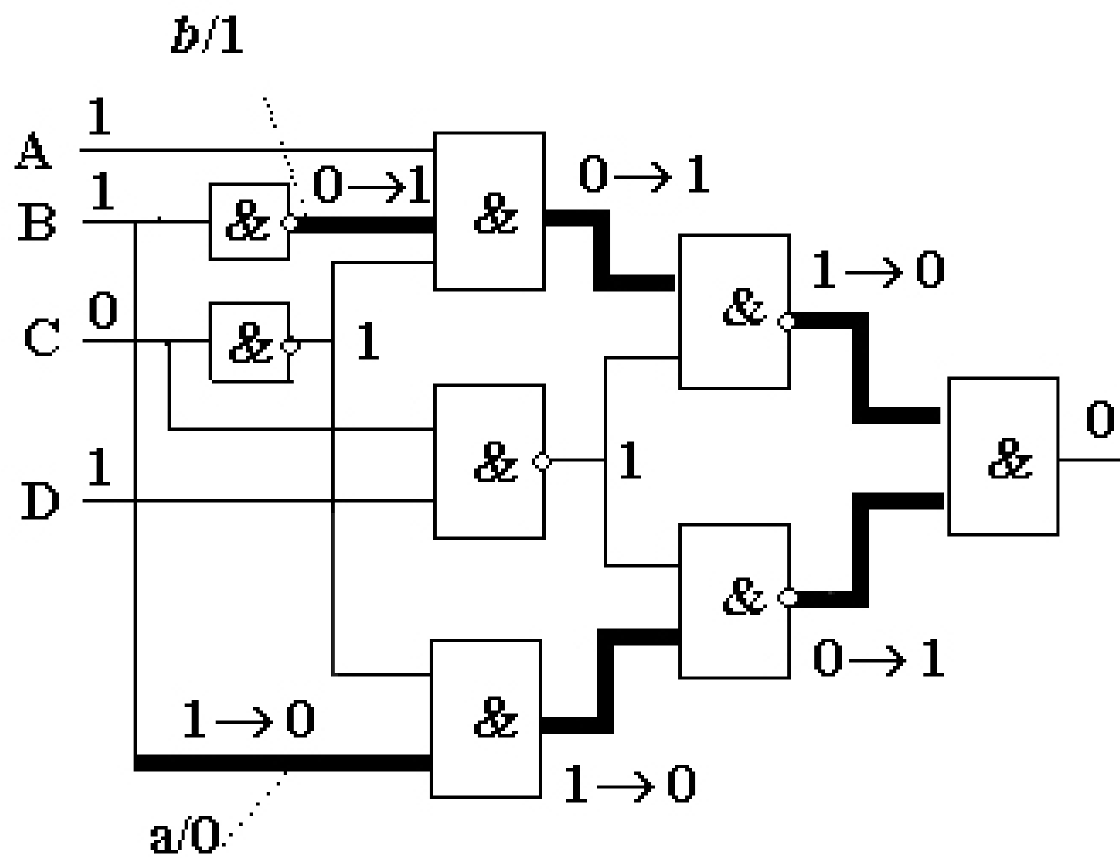
- Circuit: Let *Z(x)* be the logic function of a circuit *N*, where x represents an arbitrary input vector and *Z(x)* denotes the mapping realized by *N circuit*.
- Faulty circuit: The presence of a fault *f* transforms *N* into a new circuit N_f with a new function $Z_f(x)$.
- Test: Denote by *t*, a specific input vector (**test vector**), and by *Z(t)* the response of *N*. Let us call a sequence of test vectors by **test**: $T=\{t_1\ t_2 \ldots t_n\}$.
- The circuit is tested by applying a test *T* and by comparing the output response with the expected output response of *N* *Z(t1)*...
- Fault detection: A test vector *t* **detects** a fault *f* if Z(t) not equals to $Z_f(t)$



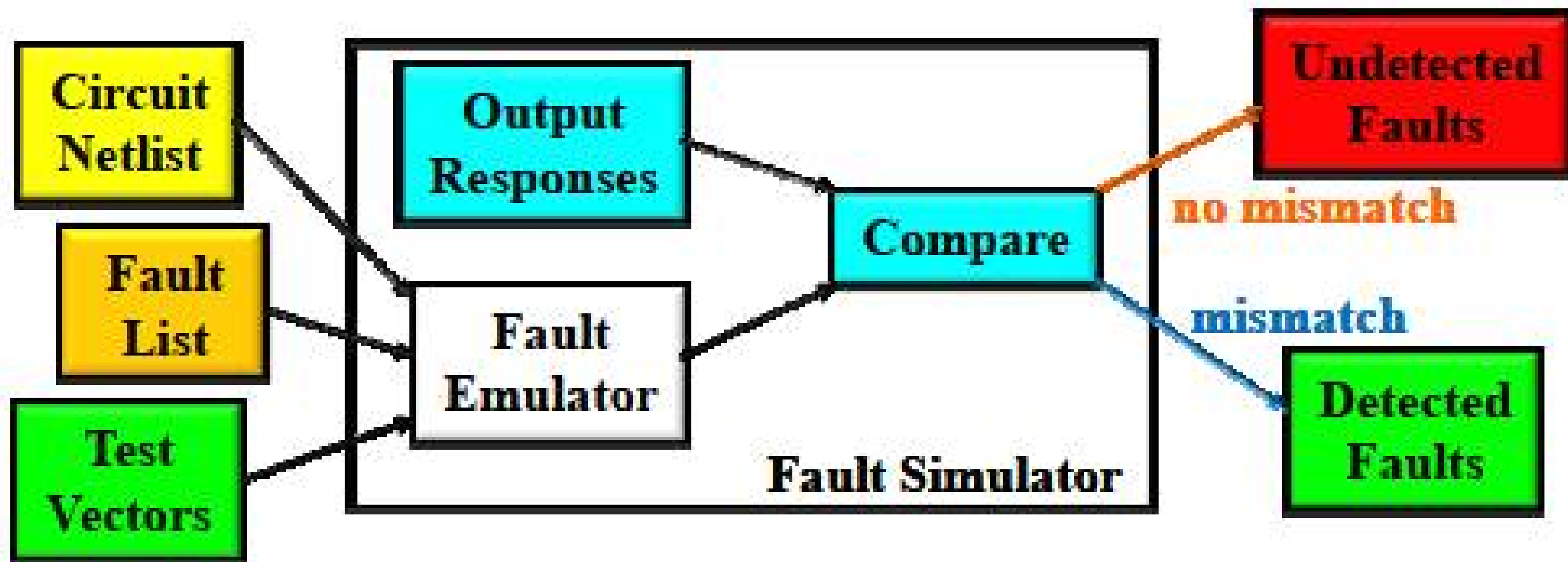
- In Figure **a fault a/0 is sensitized** by the value 1 on a line a.
- A test t = 1101 is simulated, both without and with the fault a/0. The results of the simulation are different in the two cases, shown in a form v->vf where v and vf are corresponding signal values in the fault-free and in the faulty circuit. The fault is detected since the output values in the two cases are different. A path from the faulty line *a* is sensitized (bold lines) to the primary output of the circuit.

Detectability of Faults

- A fault *f* is **detectable** if there exists a test *t* that detects *f* , otherwise, *f* is an **undetectable** fault.
- The presence of an undetectable fault *f* may prevent the detection of another fault *g.*
- In Figure, the fault *b*/1 is undetectable. As we saw in the previous example, the test *t* = 1101 detects the fault *a*/0. However, in the presents of *b*/1, the test *t* is not any more able to detect the fault *a*/0.

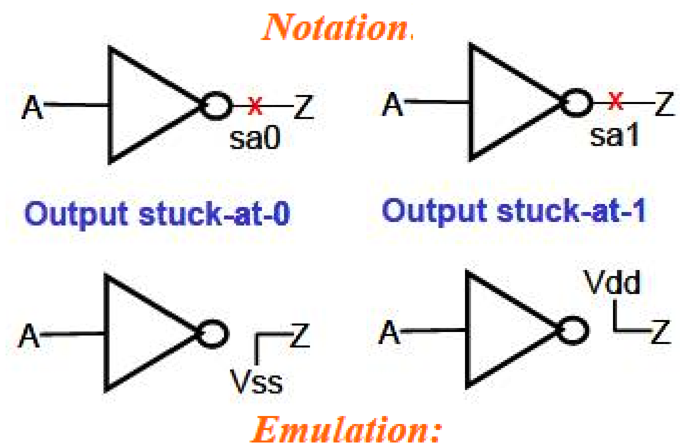


b/1
A
B
C
D
1
1
0
1
0→1
0→1
1
1→0
1
0
0→1
1→0
1→0
a/0

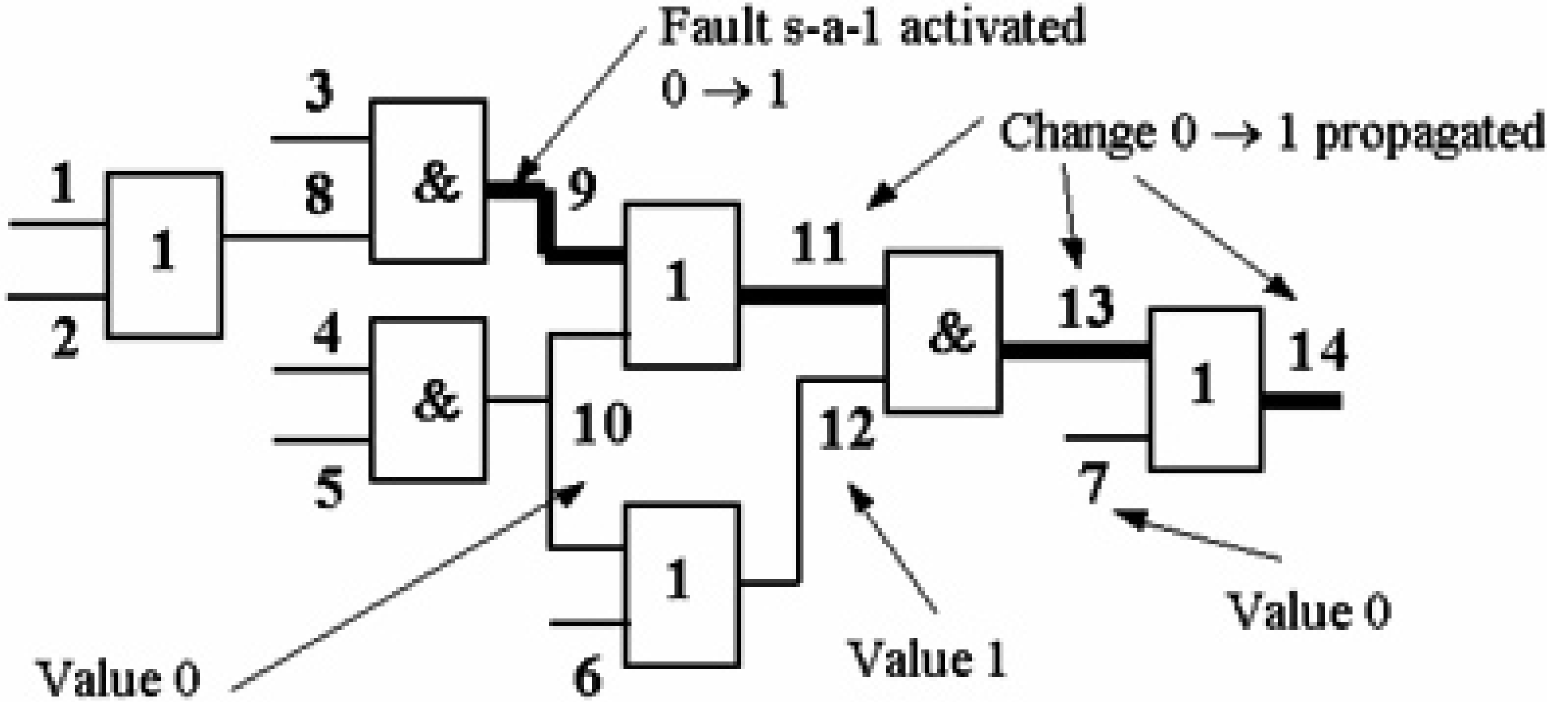


Fault simulators:

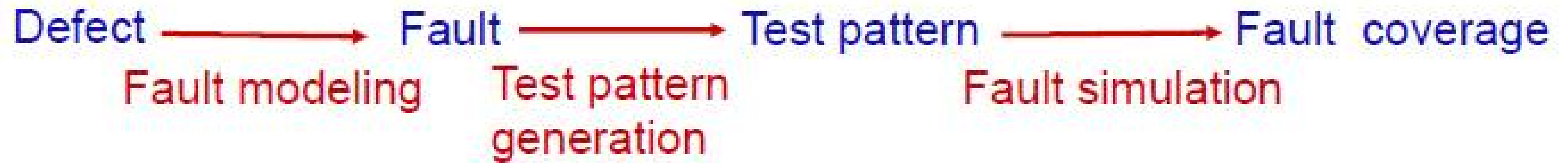
- emulate faults
- compare output response to known good circuit output response
  - For a given set of input test patterns
    - At least one mismatch ⇒ fault is detected
    - No mismatches ⇒ fault is not detected

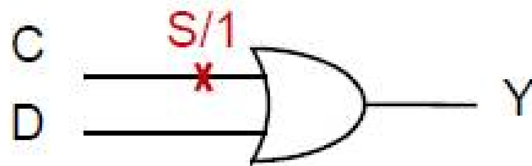


Fault Propagation



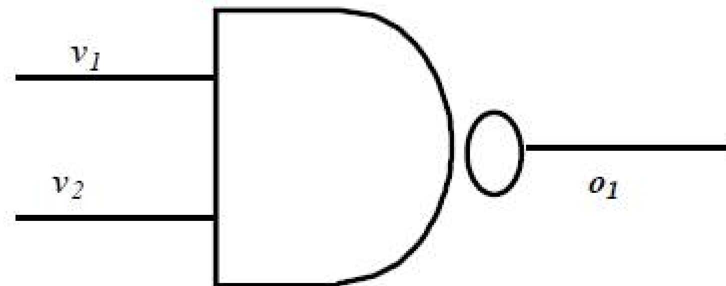
Test Problem





C	D	Y	Y(C is S/1)
0	0	0	1
0	1	1	1
1	0	1	1
1	1	1	1

Example: NAND Gate



Input		Output
v_1	v_2	o_1
0	0	1
0	1	1
1	0	1
1	1	0

This test for the NAND gate is just the starting point

Detailed tests for the NAND gate

- Digital Functionality
  - Verify input/output of Table 1
- Delay Test
  - 0 to 1: time taken by the gate to rise from 0 to 1.
    - *v1=1, v2=1* changed to *v1=1, v2=0;* After this change in input, time taken by o_1 to change from 0 to 1.
    - *v1=1, v2=1* changed to *v1=0, v2=1;* After this change in input, time taken by o_1 to change from 0 to 1.
    - *v1=1, v2=1* changed to *v1=0, v2=0;* After this change in input, time taken by o_1 to change from 0 to 1.

Detailed tests for the NAND gate

- 1 to 0: time taken by the gate to fall from 1 to 0.
  - *v1=0, v2=0* changed to *v1=1, v2=1;* After this change in input, time taken by o_1 to change from 1 to 0.
  - *v1=1, v2=0* changed to *v1=1, v2=1;* After this change in input, time taken by o_1 to change from 1 to 0.
  - *v1=0, v2=1* changed to *v1=1, v2=1;* After this change in input, time taken by o_1 to change from 1 to 0.

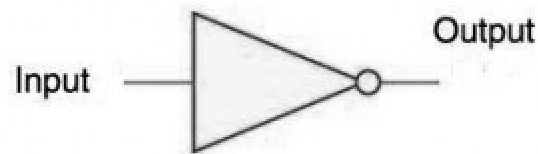
- Fan-out capability:
  - Number of gates connected at o_1 which can be driven by the NAND gate.

Detailed tests for the NAND gate

- Power consumption of the gate
  - Static power: measurement of power when the output of the gate is not switching.
  - Dynamic power: measurement of power when the output of the gate switches from 0 to 1 and from 1 to 0.
- Threshold Level
  - Minimum voltage at input considered at logic 1
  - Maximum voltage at input considered at logic 0
  - Voltage at output for logic 1
  - Voltage at output for logic 0
- Test at extreme conditions
  - Performing the tests at temperatures (Low and High Extremes) as claimed in the specification document.

Digital Circuits and the Stuck at Fault Model

With a stuck at fault model you are applying a structural test approach. Instead of testing all combination of 1's and 0's to a VLSI device, you will test with a reduced set of test vectors. Stuck at Fault Models operate at the logic model of digital circuits. An input or an output can be Stuck at Zero (S@0) or Stuck at One (S@1)



Inverter Logic Table

A	D
1	0
0	1

Inverter Stuck faults list: A S@0, A S@1, D S@0, DS@1

Testing With Stuck at Fault Model

With a stuck at fault you apply a pattern (set of 1's and 0's) to the inputs of the logic gate such that you get a faulty response. So let's start with the inverter. Suppose A is S@1. Easy test, you need to apply a 0. Now lets' look at the output D stuck at 0, you need to apply a 0 to A. Table below tabulates the test pattern per S@ fault.

Inverter S@ Fault and Test Table

S@ Fault	Test A	Pass D	Failing D
A S@1	0	1	0
A S@0	1	0	1
DS@1	1	0	1
D S@0	0	1	0

Now while there exist 4 faults to test you only need 2 tests as shown in the table below. This is commonly called *fault collapsing*.

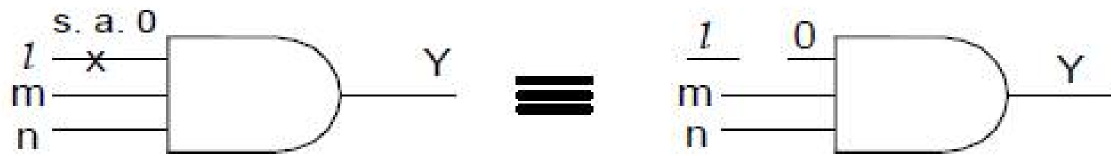
Inverter S@ Fault Coverage of Tests

A	Stuck @ Faults Detected
1	A-S@0, D-S@1
0	A-S@1, D-S@0

NAND S@ Fault and Tests

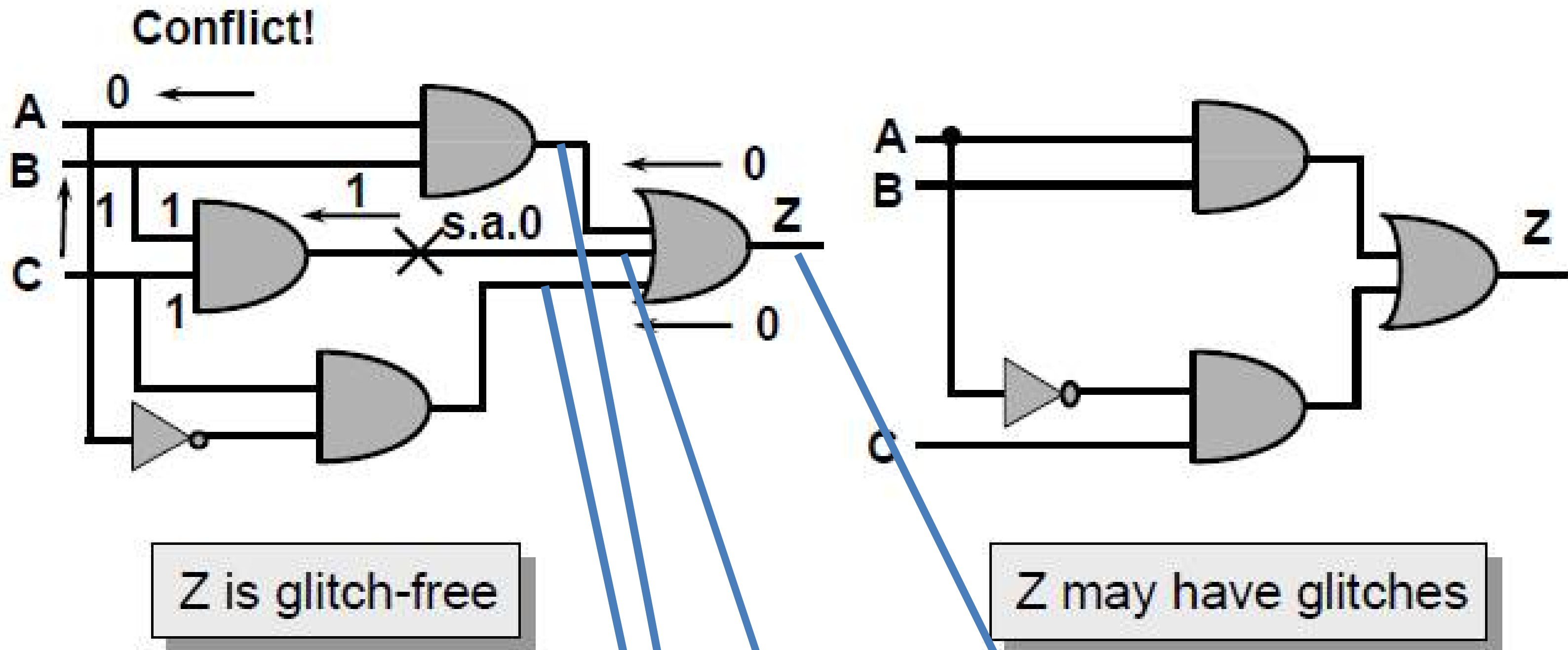
S@ Fault	Test A	Test B	Pass Y	Fail Y
A S@1	0	1	1	0
A S@0	1	1	0	1
B S@1	1	0	1	0
B S@0	1	1	0	1
Y S@1	1	1	0	1
Y S@0	1	0	1	0
Y S@ 0	0	1	1	0
CYS@ 0	0	0	1	0

Redundancy



- If l s.a.0 is undetectable, the entire gate can be removed & replaced by a constant 0 wire

Example: Redundancy Removal



If error happens then
1+ stuck at 0 + 0 = 1

1 x 1 = 1 (if no fault)

0 x 1 = 0

1 x 1 =1

1+0+1= 1

Design of Experiments (DOE)

Design of experiments (commonly referred to as DOE) is a data-driven technique for robust design. In the early 1900's, DOE was used by agricultural engineers to improve crop yields. Today circuit and system designers are applying the method as a means to the same end-yield improvement.

A typical DOE includes three primary steps:

- Plan the experiment:
  - Assess the experimental resource budget.
  - Identify the input and response variables.
  - Assign levels (values) to input variables.
- Perform the experiment and collect response data.
- Analyze the data using statistical methods.

Sequential application of this methodology can be used to improve the statistical performance of a given circuit or system. Because of an inherent compromise between statistical performance prediction accuracy and the number of input variables, a *screening* experiment is used to identify variables that contribute significantly to performance variation. Next a *refining* experiment can be used to *improve* on the target statistical response.

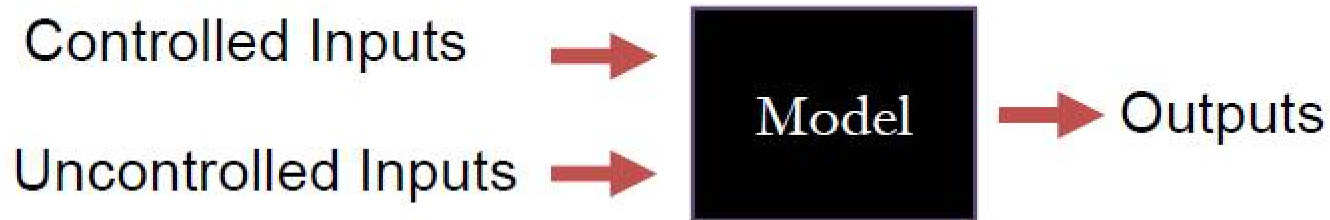
DOE and Computer Simulation

In a general application, DOE methods are designed to accommodate errors of the type found in any experiment. But because circuit and system simulators provide identical results for any analysis having the same input values, complexity in setting up, performing, and analyzing experiments is reduced.

Since the computer is being used to perform the experiment, a more complete characterization of input/output relationships can be realized. Finally, since the computer handles the tedious tasks of bookkeeping during the experiment, there is a further reduction in the possibility of human error.

The primary purpose of DOE is to characterize an unknown process. In circuit or system simulation, the unknown process is predicting the response of the design under test (DUT). A simple technique for characterizing a DUT is to perturb each input variable (*factor*) in turn, and to record the resulting output response. However this approach breaks down if the response due to a change in one factor depends on the value of a different factor.

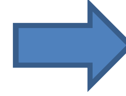
Testing A Model



- Model inputs are **independent** variables
  - Controlled inputs are usually called ***factors***
  - Uncontrolled inputs are often called ***blocking variables***, ***covariates*** or ***nuisance variables***
  - All inputs may be called ***predictor variables***, e.g. **X**
- Model outputs are **dependent** variables
  - Outputs are also called ***response variables***, e.g. **Y**

Linear Models

One-way Analysis of Variance (ANOVA) and simple linear regression are two familiar kinds of linear models



- ANOVA methods are used to compare mean response levels among groups
- Regression explores linear relationships between predictor and response variables

- Response variables determine the appropriate statistical model for the experiment
  - Most continuous responses use linear models (e.g. yield)
  - All categorical responses use generalized linear models

ANOVA

- The total variation present in a set of observable quantities may, **under certain circumstances**, be partitioned into a number of components associated with the nature of classification of the data
- The systematic procedure of achieving this is called analysis of variance (ANOVA).
- The purpose of ANOVA is to test for significant difference between means.

- The name is derived from the fact that in order to test for statistical significance between means , **we are actually comparing (analyzing) variances**.
- **Assumptions of ANOVA :**

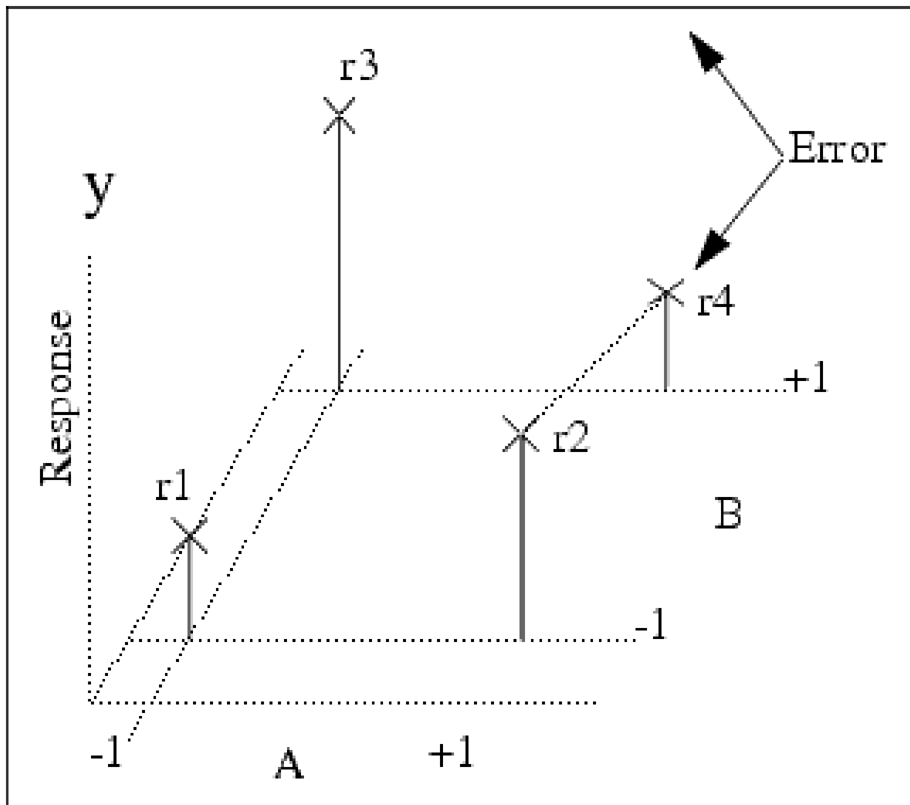
(i) Subjects are chosen via a simple random sample.

(ii) Within each group/population, the response variable is normally distributed.

(iii) While the population means may be different from one group to the next, the population standard deviation is the same for all groups.

DOE Concepts

The following figure shows two factors, A and B, and the associated response at various values (*levels*) of the factors.



Notice that the factors have two levels: one low (−1) and one high (+1). The ±1 notation indicates the factor values are in *design units* , and are obtained from physical values using the following equation:

$$\frac{X - X_{mid}}{(X_{hi} - X_{lo})/2}$$

where X is the minimum (maximum) physical value of the variable, and X_{lo} , X_{hi} , and X_{mid} are the minimum, middle, and maximum physical values. For example, a capacitor value might be 100pF ± 10%, leading to low, mid, and high values of 90, 100, and 110pF respectively.

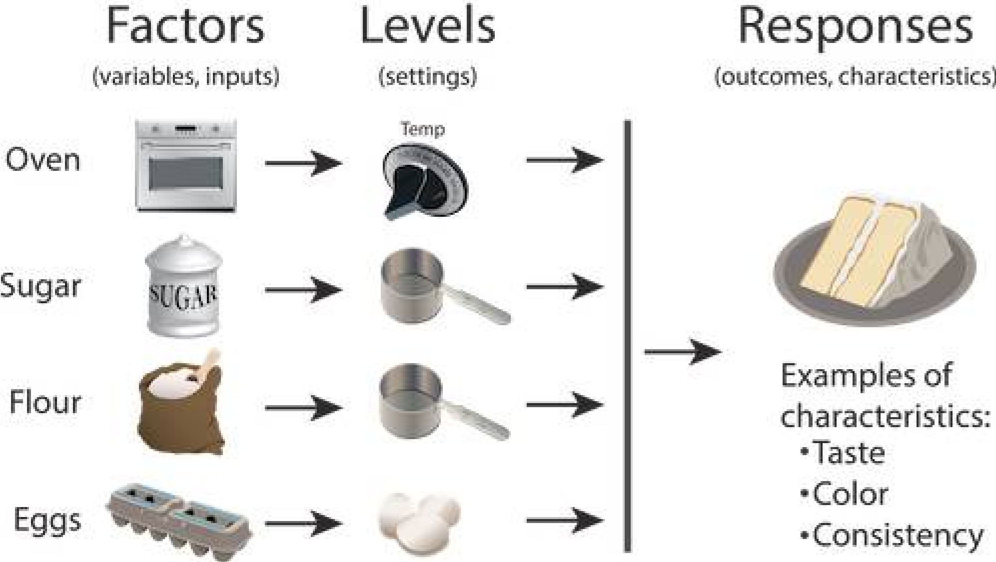
If we were to note the change in response due to a change in factor A (from low to high), we would be led to believe that increasing A causes an increase in the response- the same would be observed for factor B. A model from the three response points r1, r2, and r3 can be formulated as the plane surface.

Components of Experimental Design

Factors, or inputs to the process. Factors can be classified as either controllable or uncontrollable variables. In this case, the controllable factors are the ingredients for the cake and the oven that the cake is baked in. The controllable variables will be referred to throughout the material as factors. Note that the ingredients list was shortened for this example - there could be many other ingredients that have a significant bearing on the end result (oil, water, flavoring, etc). Likewise, there could be other types of factors, such as the mixing method or tools, the sequence of mixing, or even the people involved. People are generally considered a Noise Factor - an uncontrollable factor that causes variability under normal operating conditions, but we can control it during the experiment using blocking and randomization.

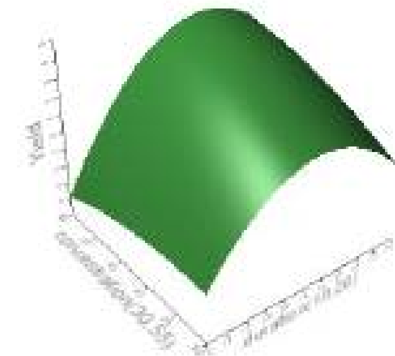
Levels, or settings of each factor in the study. Examples include the oven temperature setting and the particular amounts of sugar, flour, and eggs chosen for evaluation.

Response, or output of the experiment. In the case of cake baking, the taste, consistency, and appearance of the cake are measurable outcomes potentially influenced by the factors and their respective levels. Experimenters often desire to avoid optimizing the process for one response at the expense of another. For this reason, important outcomes are measured and analyzed to determine the factors and their settings that will provide the best overall outcome for the critical-to-quality characteristics - both measurable variables and assessable attributes.



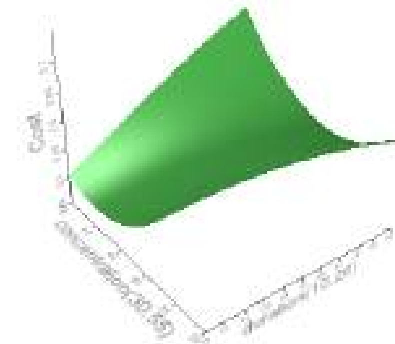
Why Response Surface?

- The goal is to maximize, minimize or target a certain value of the response variable(s)

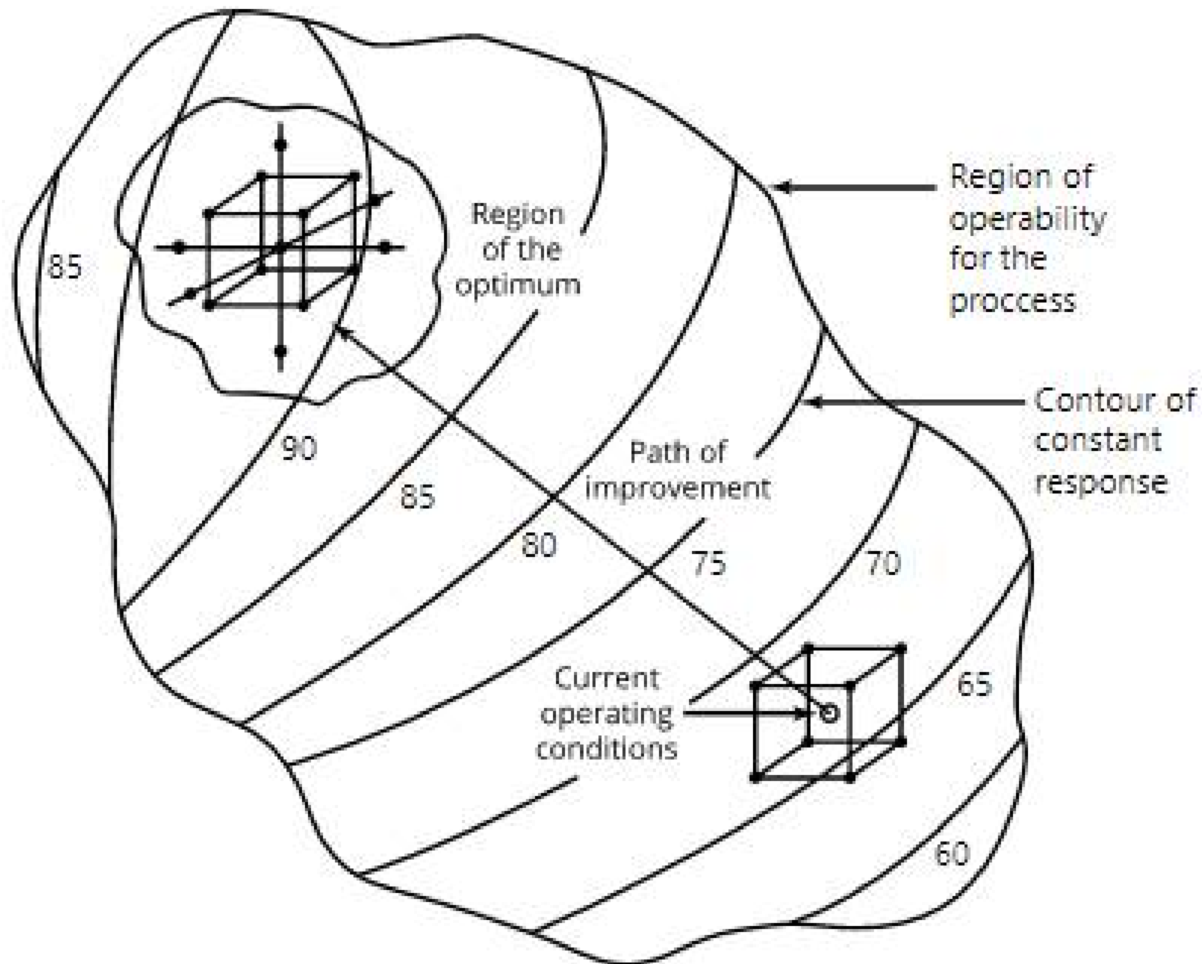


- We must accurately describe the complex multidimensional relationships between the predictors and responses to optimize results

- We use main effects, two-way interactions and 2nd order polynomials to describe these relationships, because they are easiest to interpret and cover most of the relationships



objective of Response Surface Methods (RSM) is optimization, finding the best set of factor levels to achieve some goal.

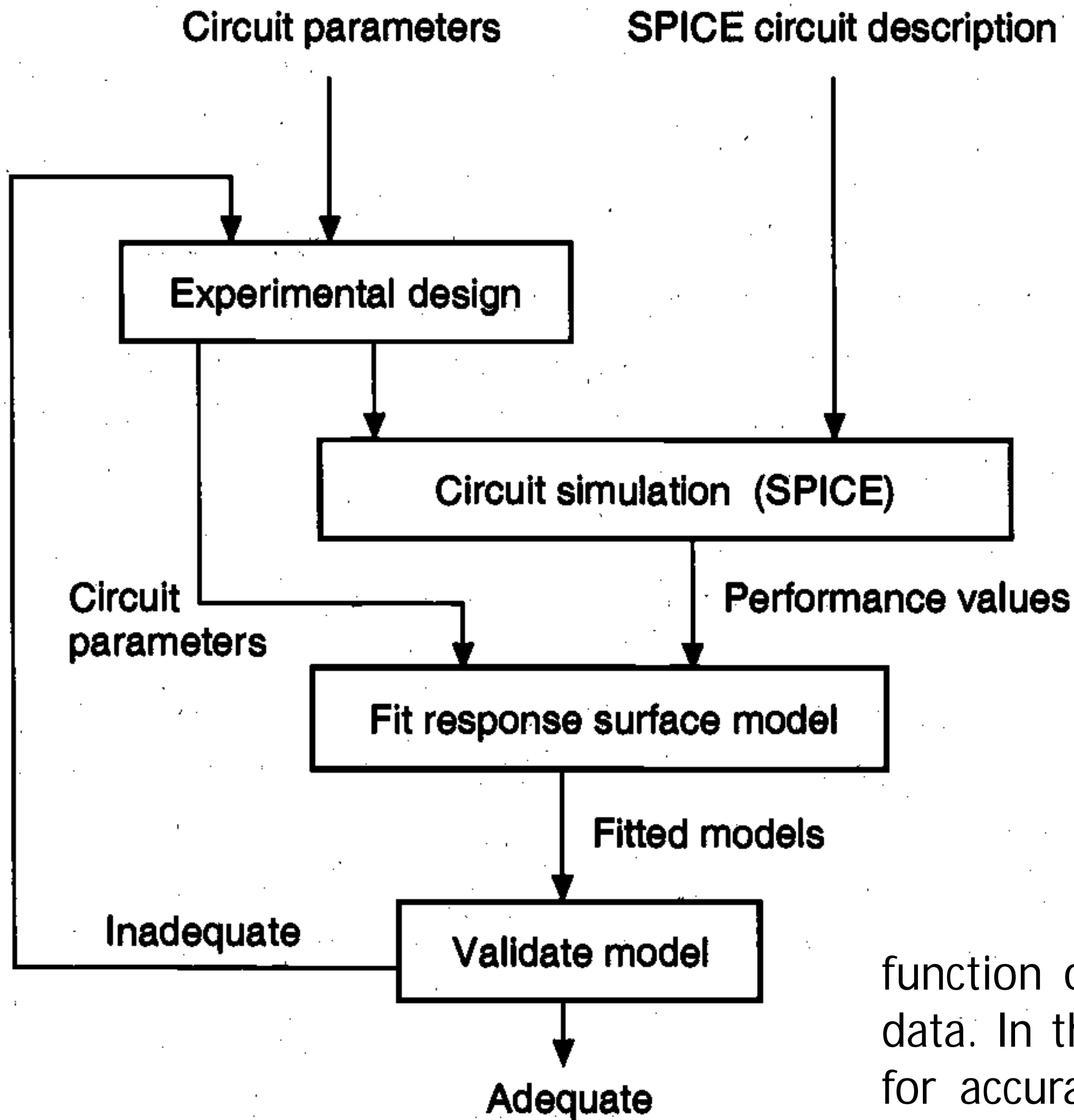


Suppose that there are n circuit parameters of interest denoted by $\mathbf{x} = (x_1, x_2, ..., x_n)$. These circuit parameters can be designable parameters or noise parameters. As noted earlier, a circuit performance r is a function $r(\mathbf{x})$ of these parameters. Usually, this function is not known explicitly, and for particular values of $\mathbf{x}$, r has to be evaluated using a circuit simulator such as SPICE. Circuit simulations are computationally expensive, especially if the circuit size is large and transient simulations are required. An attractive alternative would be to construct a compact model of the circuit performance in terms of the parameters $\mathbf{x}$ and then use the performance model instead of the circuit simulator to evaluate the performance. The utility of such an approach depends on two criteria. First, the model should be computationally efficient to construct and evaluate so that substantial computational savings can be achieved. Second, the model should be accurate. Clearly, these two features are conflicting requirements, and some optimal trade-offs should be made in order to develop a good model.

Since the model is to act as a surrogate to the circuit simulator, it should be built from values of *r* obtained by circuit simulations.

The model building procedure consists of four steps.

1. First, *m training points* are selected in the x-space. The ith training point is denoted by $\mathbf{x}_i = (\mathbf{x}_{1i}, x_{2i}, \ldots, x_{nj})$.
2. In the second step, the circuit is simulated at the *m* training points, and the values of the performance measure are obtained from the simulation results as $r(x_1),\ldots,r(x_m)$.



function of *r* in terms of x is "fitted" to the data. In the final step, the model is validated for accuracy. If the model is not sufficiently accurate, then the modeling procedure is repeated with a larger number of training points or with different models.

Sequential Nature of the Response Surface Methodology

Phase 0: At first some ideas are generated concerning which factors or variables are likely to be important in response surface study. It is usually called a **screening experiment**. The objective of factor screening is to reduce the list of candidate variables to a relatively few so that subsequent experiments will be more efficient and require fewer runs or tests. The purpose of this phase is the identification of the important independent variables.

Phase 1: The experimenter's objective is to determine if the current settings of the independent variables result in a value of the response that is near the optimum. If the current settings or levels of the independent variables are not consistent with optimum performance, then the experimenter must determine a set of adjustments to the process variables that will move the process toward the optimum. This phase of RSM makes considerable use of the first-order model and an optimization technique called the **method of steepest ascent (descent)**.

Phase 2: Phase 2 begins when the process is near the optimum. At this point the experimenter usually wants a model that will accurately approximate the true response function within a relatively small region around the optimum. Because the true response surface usually exhibits curvature near the optimum, a second-order model (or perhaps some higher-order polynomial) should be used. Once an appropriate approximating model has been obtained, this model may be analyzed to determine the optimum conditions for the process.

RSM resolve around the assumption that the response is a function of a set of independent(design) variables x1,x2,x3….xk and

function can be approximated in some region of polynomial model.

$$y = f\ (xi)$$

$$y = f\ (x1, x2 \ldots \ldots xk)$$

Here response variable is "y" that depend on the "k" independent variables.

- If the factors are given then directly estimate the effects and interaction of model.
- And if the factors are unknown then first calculate them by using the Screening method.

Estimate The Interaction effect using 1st order model.

$$y = \beta 0 + \beta 1 x 1 + \beta 2 x 2 + \varepsilon$$

If curvature is found then use the RSM. And 2nd order model will be used to approximate the response variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon$$

Make the graph and find the stationary point. Maximum response, Minimum response or saddle point by using the obtained values of $x1$, $x2$, $x3$ … . $x4$.

The computational cost of modeling depends on *m*, the number of training points, and on the procedure of fitting the model to the simulation data. The accuracy of the model is calibrated.by computing error measures which quantify the "goodness of fit." The accuracy of the derived model would be influenced greatly by the manner in which the training points are selected from the x-space.

$$r'(\mathbf{x}) = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_i + \sum_{i=1}^{n} \sum_{j=i}^{n} \alpha_{ij} x_i x_j$$

is the RSM used, where the coefficients α_0, α_i, and α_{ij} are the fitting parameters in the model. Note, however, that the discussion is valid for any other RSM as well.

Factorial Designs

Many experiments involve the study of the effects of two or more factors. *Factorial designs* are most efficient for this type of experiment.

In a factorial design, all possible combinations of the levels of the factors are investigated in each replication.

If there are *a* levels of factor A, and *b* levels of factor B, then each replicate contains all *ab* treatment combinations.

Main Effects

- The main effect of a factor is defined to be the change in response produced by a change in the level of a factor.
- The main effect of A is the difference between the average response at A_1 and A_2

		FACTOR B	
		B_1	B_2
FACTOR A	A_1	20	30
	A_2	40	52

Interaction

In some experiments we may find that the difference in response between the levels of one factor is not the same at all levels of the other factor. When this occurs, there is an *interaction* between the factors.

At B_1, the A effect is:

At B_2, the A effect is:

		FACTOR B	
		B_1	B_2
FACTOR A	A_1	20	40
	A_2	50	12

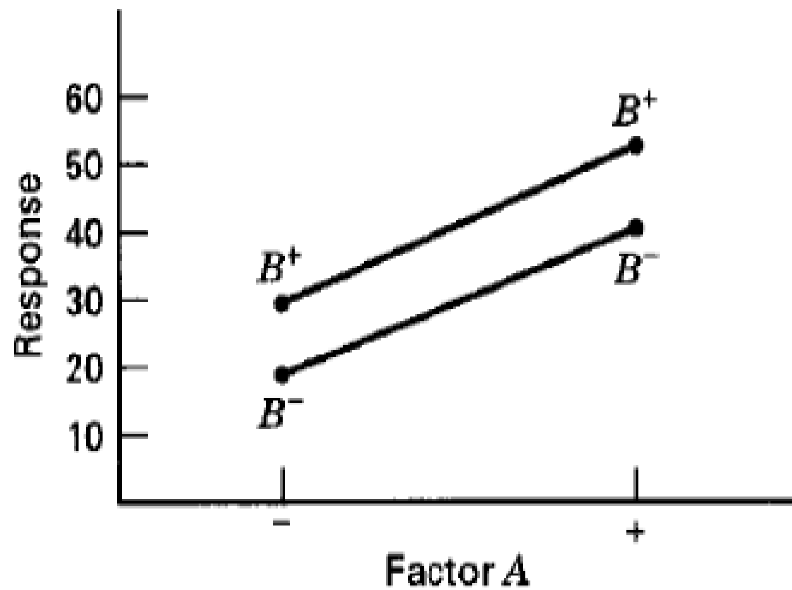


Figure 5-3 A factorial experiment without interaction.

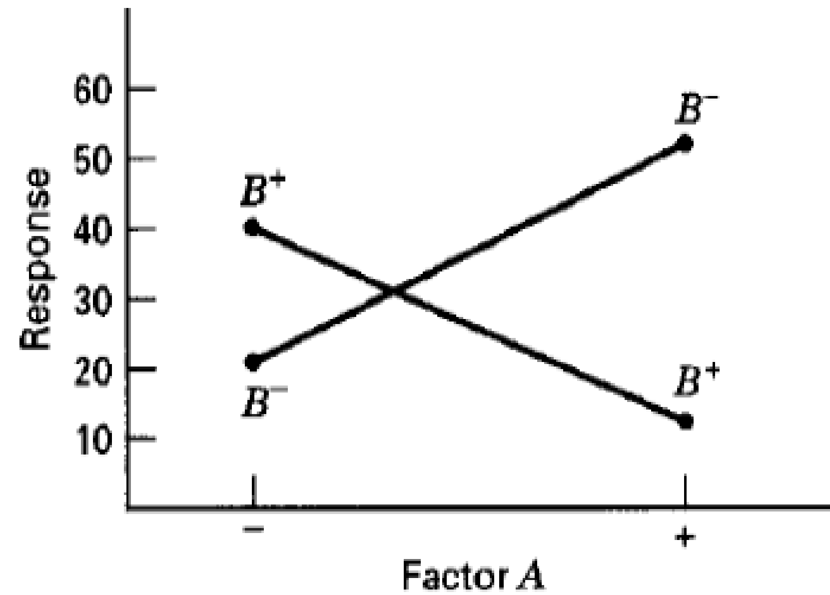


Figure 5-4 A factorial experiment with interaction.

A factorial design is necessary when interactions may be present to avoid misleading conclusions.

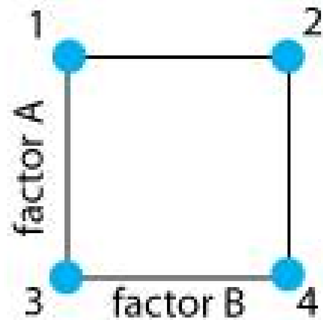
Factorial designs allow the effects of a factor to be estimated at several levels of the other factors, yielding conclusions that are valid over a range of experimental conditions.

For example, if we believe that factors A and B are independent and that each has only a first-order effect on the response, then the following equation is a suitable model.

$$R = \beta_0 + \beta_a A + \beta_b B$$

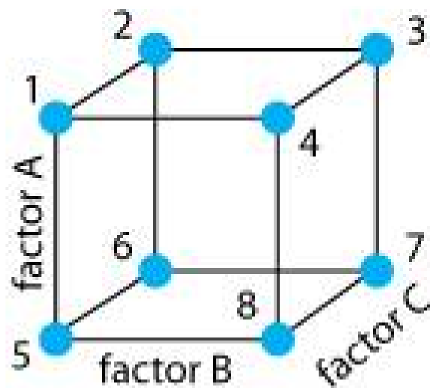
where R is the response, A and B are the factor levels, and β_0, β_a, and β_b are adjustable parameters whose values are determined by a linear regression analysis. We call this equation an empirical model of the response surface because it has no basis in a theoretical understanding of the relationship between the response and its factors.

As shown here, we build an empirical model by measuring the response for at least two levels for each factor—indicated by the plus and minus signs in the tables—and complete a simple regression analysis. This is known as a 2^k factorial design because it requires 2^k experiments where k is the number of factors.



factor levels

trial	A	B
1	+	–
2	+	+
3	–	–
4	–	+



factor levels

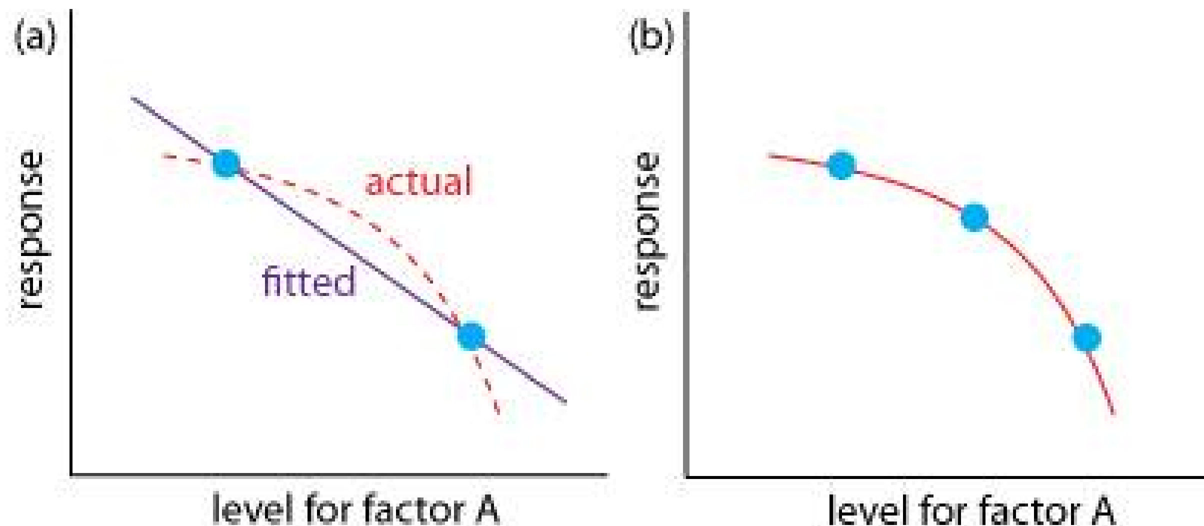
trial	A	B	C
1	+	–	–
2	+	–	+
3	+	+	+
4	+	+	–
5	–	–	–
6	–	–	+
7	–	+	+
8	–	+	–

A 2^k factorial design can model only a factor's first-order effect on the response. A 2^2 factorial design, for example, includes each factor's first-order effect (β_a and β_b).

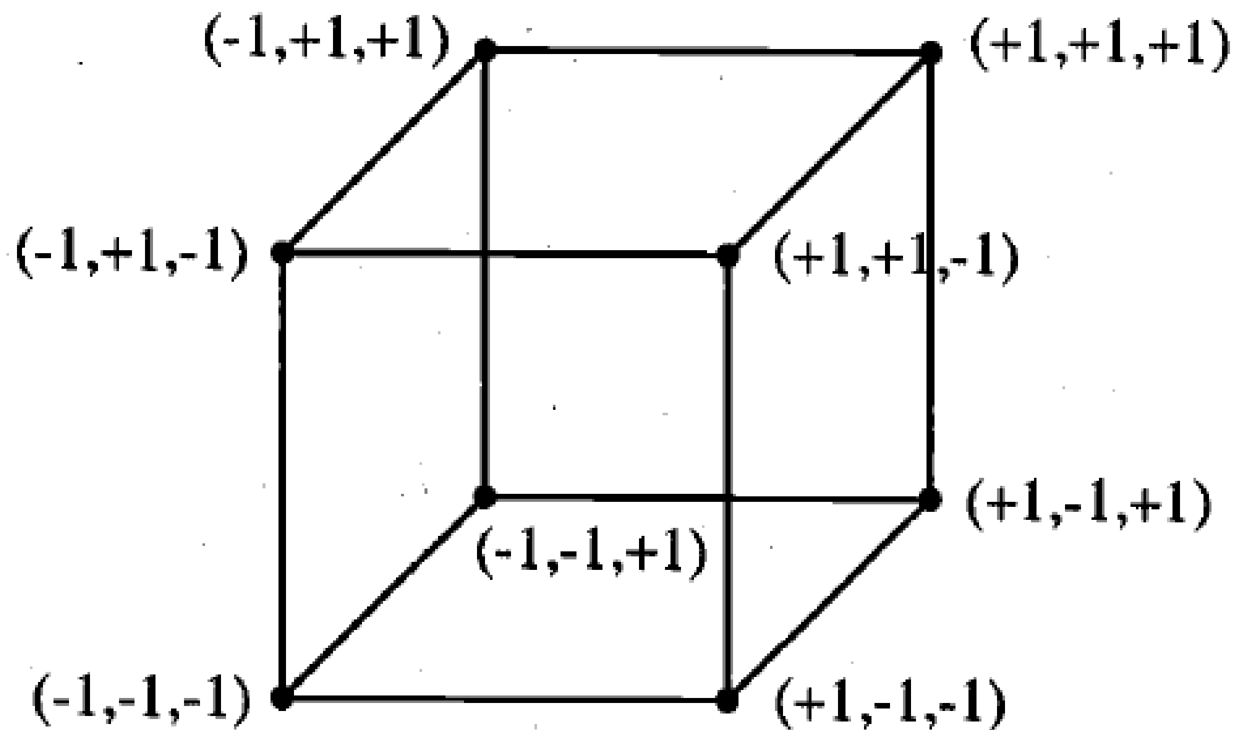
A first-order interaction between the factors (β_{ab}), and an intercept, (β_0); with four experiments we need enough information to calculate the four β values.

$$R = \beta_0 + \beta_a A + \beta_b B + \beta_{ab} AB$$

A 2^k factorial design cannot model higher-order effects because there is insufficient information. Here is simple example that illustrates the problem. Suppose we need to model a system in which the response is a function of a single factor. As illustrated here, a 2^1 factorial design has but two responses, which means we can fit only a straight line to the data. To see evidence of curvature we must measure the response for at least three levels for each factor, as shown in figure.



In this experimental design technique, each of the parameters $x_1, x_2, ..., x_n$ is quantized into two levels or settings (the smallest and largest values in its range). Without any loss of generality, we can assume that these values are −1 and +1 for each parameter after normalization. A full factorial design consists of all possible combinations of values for the n parameters. Therefore, a full factorial design for n parameters has 2^n training points or experimental runs. The design matrix for the case of $n = 3$ is shown in Table 15.2 and the design is depicted in Fig. 15.9. The perfomance data for the kth run are denoted by r_k. The full factorial design provides much information on the relationship between the parameters x_i and the performance r. For example, one can evaluate the *main* or *individual effect* of a parameter x_i, which quantifies how much the parameter affects the performance. The main effect is the difference between the average performance value when the parameter is at the high level (+1) and the average performance value when the parameter is at the low level (−1). The main effect of x_i is the coefficient of the x_i term in the RSM of (15.8). One can also determine the *interaction effects* of two or more parameters which quantify how those factors jointly affect the performance. Two factor interactions can be computed as the difference between the average performance value when both parameters are at the same level and the average performance value when the



Pictorial representation of full factorial design for $n = 3$.

parameters are at different levels. The two-factor interaction effect of parameters x_i and x_j is the coefficient of the $x_i x_j$ term in (15.8). Higher-order multifactor interactions can be computed in a recursive manner.

Run	Parameters			Interactions				
	x_1	x_2	x_3	$x_1 \times x_2$	$x_1 \times x_3$	$x_2 \times x_3$	$x_1 \times x_2 \times x_3$	r
1	−1	−1	−1	+1	+1	+1	−1	r_1
2	−1	−1	+1	+1	−1	−1	+1	r_2
3	−1	+1	−1	−1	+1	−1	+1	r_3
4	−1	+1	+1	−1	−1	+1	−1	r_4
5	+1	−1	−1	−1	−1	+1	+1	r_5
6	+1	−1	+1	−1	+1	−1	−1	r_6
7	+1	+1	−1	+1	−1	−1	−1	r_7
8	+1	+1	+1	+1	+1	+1	+1	r_8

Full factorial design for $n = 3$.

$2^k r$ Factorial Design with Replications

- Problem with 2^k factorial design is that it does not provide the estimation of experimental errors, since no repetitions
- Solution: Repeat an experiment r times ➡ replication
  - If each of the 2^k experiments is repeated r times ➡ $2^k r$ factorial design with replications
- Extended model

$$y = q_0 + q_A x_A + q_B x_B + q_{AB} x_A x_B + e$$

Consider the set up of complete factorial experiment, say 2^k. If there are four factors, then the total number of plots needed to conduct the experiment is $2^4 = 16$. When the number of factors increases to six, then the required number of plots to conduct the experiment becomes $2^6 = 64$ and so on. Moreover, the number of treatment combinations also become large when the number of factors increases. Sometimes, it is so large that it becomes practically difficult to organize such a huge experiment. Also, the quantity of experimental material needed, time, manpower etc. also increase and sometimes even it may not be possible to have so many resources to conduct a complete factorial experiment. The non-experimental type of errors also enters in the planning and conduct of the experiment. For example, there can be a slip in numbering the treatments or plots or they may be wrongly reported if they are too large in numbers.

About the degree of freedoms, in the 2^6 factorial experiment there are $2^6 - 1 = 63$ degrees of freedom which are divided as 6 for main effects, 15 for two-factor interactions and rest 42 for three or higher-order interactions. In case, the higher-order interactions are not of much use or much importance, then they can possibly be ignored. The information on main and lower-order interaction effects can then be obtained by conducting a fraction of complete factorial experiments. Such experiments are called as **fractional factorial experiments.**

One-half fraction of 2^3 factorial experiment

First, we consider the set up of 2^3 factorial experiment and consider its one-half fraction. This is a very simple set up to understand the basics, definitions, terminologies and concepts related to the fractional factorials.

Consider the setup of 2^3 factorial experiment consisting of three factors, each at two levels. There is a total of 8 treatment combinations involved. So 8 plots are needed to run the complete factorial experiment.

Suppose the material needed to conduct the complete factorial experiment in 8 plots is not available or the cost of total experimental material is too high. The experimenter has material or money which is sufficient only for four plots. So the experimenter decides to have only four runs, i.e., ½ fraction of 2^3 factorial experiment. Such an experiment contains a one-half fraction of a 2^3 experiment and is called 2^{3-1} factorial experiment. Similarly, $1/2^2$ fraction of 2^3 factorial experiment requires only 2 runs and contains $1/2^2$ fraction of 2^3 factorial experiment and is called as 2^{3-2} factorial experiment. In general, $1/2^p$ fraction of a 2^k factorial experiment requires only 2^{k-p} runs and is denoted as 2^{k-p} factorial experiment.

Thus, the full factorial design allows us to estimate all the first-order and cross-factor second-order coefficients in the RSM of (15.8). However, it does not allow us to estimate the coefficients of the pure quadratic terms x_i^2. Moreover, the number of experimental runs increases exponentially with the number of parameters. In most modeling situations, the information about the high-order multifactor interaction effects is often unnecessary and unimportant. It is possible to reduce the size of a full factorial design without compromising the accuracy of the main effects and the low-order interaction effects that are of primary interest. This is accomplished by considering only a fraction of the original full factorial design by systematically eliminating some of the runs. Such designs are called *fractional factorial designs*.

	Parameters			Interactions				
Run	x_1	x_2	x_3	$x_1 \times x_2$	$x_1 \times x_3$	$x_2 \times x_3$	$x_1 \times x_2 \times x_3$	r
1	−1	−1	+1	+1	−1	−1	+1	r_1
2	−1	+1	−1	−1	−1	−1	+1	r_2
3	+1	−1	−1	−1	+1	+1	+1	r_3
4	+1	+1	+1	+1	+1	+1	+1	r_4

Table 15.3. Half-fraction of full factorial design for $n = 3$.

It is not possible to distinguish between effects that correspond to identical columns. Such effects are said to be *confounded* or *aliased* with each other. In the fractional design of Table 15.3, we see that the main effect of x_3 is confounded with the interaction effect of x_1 and x_2. Moreover, the column $x_1 \times x_2 \times x_3$ is identical to a column of 1s. This implies that the three-factor interaction effect is confounded with the grand average of the performances. Confounding, however, is not really a problem since in most applications, high-order multifactor interactions are negligible and may be ignored. In the quadratic RSM in (15.8), only main effects and two-factor interaction effects are important, and it can be assumed that all higher-order interactions are absent. One of the most important attributes of fractional factorial designs is that these designs are orthogonal, which allows the model coefficients to be estimated with minimum errors.

	Parameters			Interactions				
Run	x_1	x_2	x_3	$x_1 \times x_2$	$x_1 \times x_3$	$x_2 \times x_3$	$x_1 \times x_2 \times x_3$	r
1	−1	−1	−1	+1	+1	+1	−1	r_1
2	−1	−1	+1	+1	−1	−1	+1	r_2
3	−1	+1	−1	−1	+1	−1	+1	r_3
4	−1	+1	+1	−1	−1	+1	−1	r_4
5	+1	−1	−1	−1	−1	+1	+1	r_5
6	+1	−1	+1	−1	+1	−1	−1	r_6
7	+1	+1	−1	+1	−1	−1	−1	r_7
8	+1	+1	+1	+1	+1	+1	+1	r_8

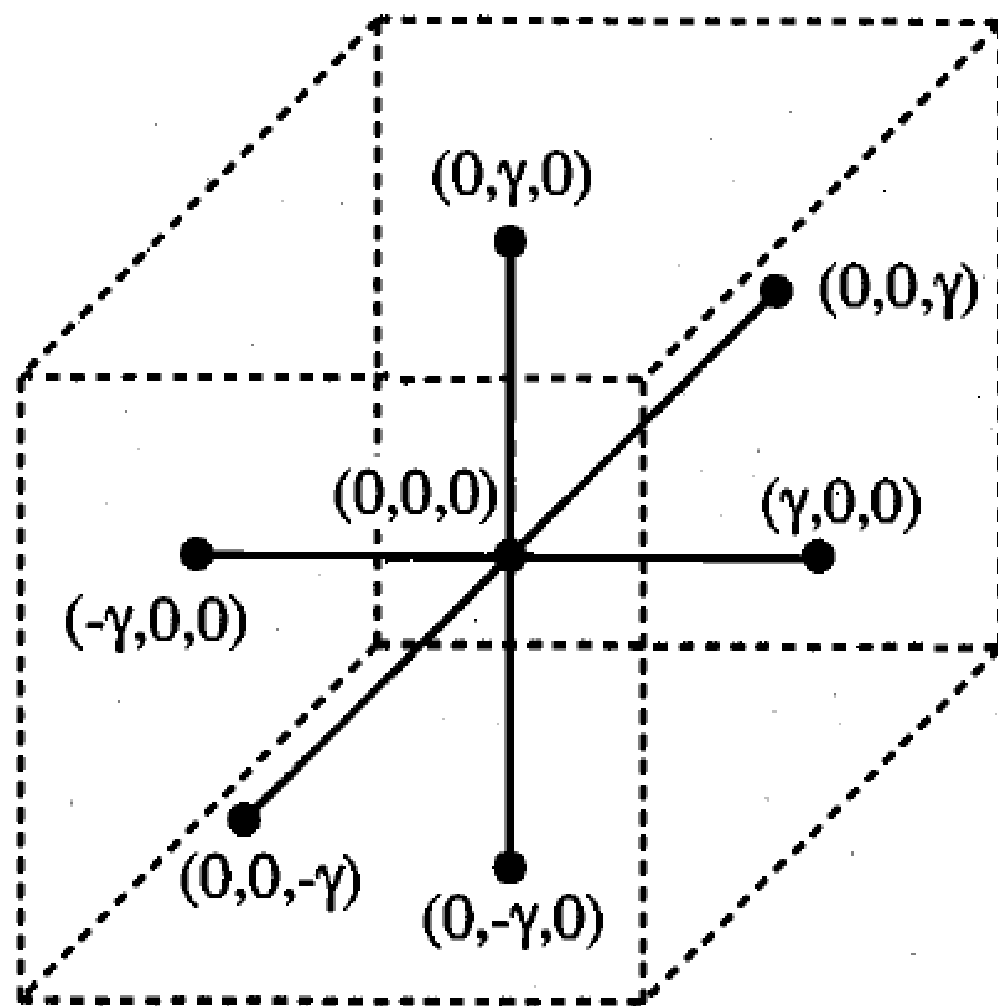
	Parameters			Interactions				
Run	x_1	x_2	x_3	$x_1 \times x_2$	$x_1 \times x_3$	$x_2 \times x_3$	$x_1 \times x_2 \times x_3$	r
1	−1	−1	+1	+1	−1	−1	+1	r_1
2	−1	+1	−1	−1	−1	−1	+1	r_2
3	+1	−1	−1	−1	+1	+1	+1	r_3
4	+1	+1	+1	+1	+1	+1	+1	r_4

Table 15.3. Half-fraction of full factorial design for $n = 3$.

Central Composite Designs

As mentioned above, one of the problems of factorial designs in regard to the RSM of (15.8) is that the coefficients of the pure quadratic terms cannot be estimated. This can be done with a *central composite design*, which is a combination of a factorial (full or fractional) and a "star" design. Figure 15.10 shows the central composite design for the case of $n = 3$. The factorial design is the "cube" design shown in dotted lines, while the star design is shown in solid lines. Each parameter in this design can take on five levels, 0, ± 1, $\pm\gamma$, where $0 < \gamma < 1$, and the star section of the design consists of $(2n + 1)$ runs, which includes

- one *center* point, where all parameters are set to 0, and
- $2n$ *axial* points, with the axial-pair values set to $-\gamma$ and $+\gamma$ while all other parameters are set to 0.


(0,γ,0)
(0,0,γ)
(0,0,0)
(γ,0,0)
(-γ,0,0)
(0,0,-γ)
(0,-γ,0)