# EECS 452 – Lecture 6

| | |
|---|---|
| Today: | Rounding and quantization<br>Analog to digital conversion |
| Announcements: | Lab 3 starts next week<br>Hw3 due on tuesday<br>Project teaming meeting: today 7-9PM, Dow 3150<br>My new office hours: M 2:30-4PM, Fri 10:30-12 |

Seminar: D. Jeon, "Energy-efficient Digital Signal Processing Hardware Design"
Mon Sept 22, 9:30-11:30am in 3316 EECS

References: Please see last slide.

Last one out closes the lab door!!!! Please keep the lab clean and organized.

Nothing is more difficult, and therefore more precious, than to be able to decide.
— Napoleon I.

# Teaming meeting mechanics

Meet in Dow 3150

6:45PM Pizza arrives

7:00PM Overview of lab resources for projects (K. Metzger)

7:15PM Project pitches: Students pitch their favorites (5 min each)

8:00PM Project assignments: Students sign up for projects

9:00PM Adjourn

Notes:

- Review project ideas in PPI booklet before meeting
- If you have a strong interest in a project idea prepare a 5min pitch
- Group size: target team size is 3.
- Only 8 or 9 projects can be accomodated so be flexible!
- We will only assign people to teams if necessary.

# How to pick a project

- Choose something that
  - Relates to the class and lab.
  - That you will enjoy doing.
  - Has a reasonable degree of challenge.
- **Be brave**
  - Take chances and be prepared to make mistakes.
  - You will need to get *something* working.
  - Aim high: combined forces of 3 people can do a lot!
- This will be a lot of work, but...
  - It's your MDE.
  - You will be working on your final product that you chose.
  - So you should get some enjoyment out of it!

# What to do after the tonight's meeting?

Schedule

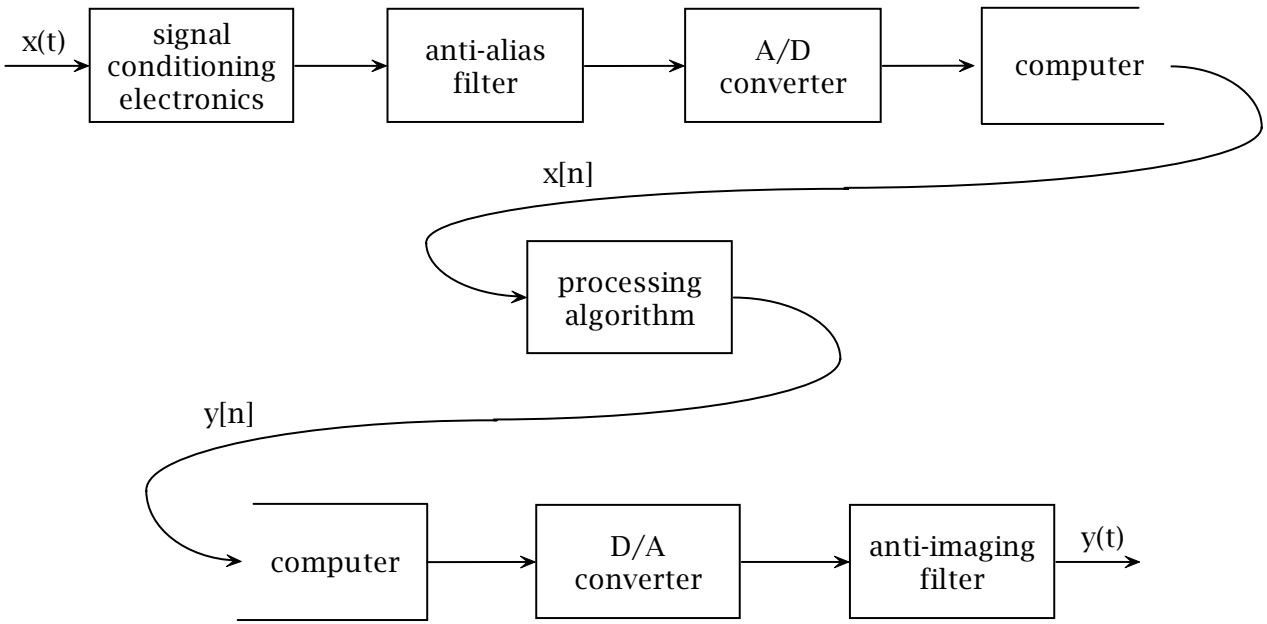
- Start preparing your proposal (due Fri Sept 26 by email to `hero`).
- Your team should meet face-to-face at least once before proposal is due.
- At this meeting you should designate a *team spokesperson*
- The spokesperson will be the team's POC and be responsible for
  - communicating proposal and final report to `hero` before due date
  - Signing up for 30 min proposal presentation slot on Sept 29, 6-10PM
  - Signing up for milestone I and II meetings
  - Registering the team for Design Expo

# Preparation of project proposal

The project proposal must use the template available on the EE452 webpage (Homeworks/Projects)

1. Introduction and overview
2. Description of project
   - 2.1 System concept, feasibility of project
   - 2.2 Describe system architecture with detailed block diagram including DSP, FPGA, peripherals as applicable.
   - 2.3 Prodict what can go wrong and your contingency plan.
   - 2.4 Provide a preliminary parts list including devices available the lab and those that you wish to purchase (parts numbers, cost and links to webpages). Give a total cost projection.
3. Milestones
   - 3.1 Milestone 1 (Th Nov 6)
   - 3.2 Milestone 2 (Tu Nov 25)
4. Contributions of each member of the team.
5. References and citations

# ADC and DAC - Recall basic DSP paradigm (From Lecture 1)




**Physical signal $\rightarrow$ Digital signal $\rightarrow$ Physical signal**

# Digitization errors and distortions(Lecture 1)

**Shannon's data processing theorem**: Digitization (sampling/quantization) of signals entails loss of information

- Distortion due to sampling
  - Aliasing distortion → increase sample rate
  - Non-ideal sampling distortion → increase sampling bandwidth
  - Non-ideal reconstruction → increase sample rate, improve interpolation
  - Clock jitter → stabilize clock rates.
- Distortion due to quantization
  - Round-off error → increase resolution (# bits)
  - Saturation/overload error → proper scaling/companding
  - Roundoff error propagation → careful balancing of arithmetic opns
- Other types of errors: thermal noise, non-linear transducers, latency, drift.

In addition to ADC/DAC implementation, we will cover red issues today.

# ADC and DAC

ADC (analog to digital converter) and DAC (digital to analog converter) are critical elements in DSP.

- This is how the DSP converses with the outside physical world.
- You will get to practice ADC and DAC in Lab 4.

Analog to digital conversion (ADC):

- Discretize it in time: *sampling* or *time quantization*.
- Discretize it in value/amplitude: *quantization* or *amplitude quantization*.

ADC yields binary number that represents quantized value of analog input.
Digital to analog conversion (DAC):

- Reverse quantization: convert binary number to real valued number
- Reverse sampling: interpolate over discrete time gaps to obtain continuous time signal.

DAC yields real valued number that represents the binary output of DSP.

# Rounding in finite precision arithmetic

Quantization in an ADC can be thought of as an infinite resolution extension of rounding of finite precision numbers

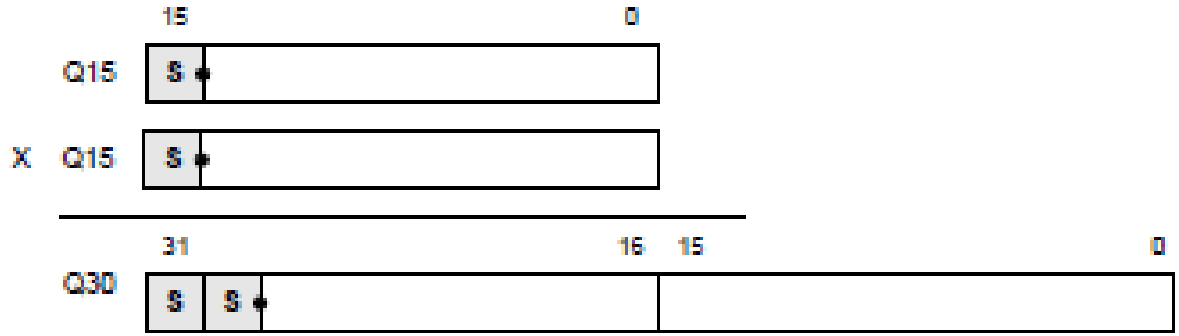
Q. How to convert a signed 16-bit Q(15) number to 8 bit Q(7)?
A. Truncation method. Shift Q(15) number right by 8. The 8 lsb's give the Q(7) conversion.

This method "rounds down' the Q(15) number to the nearest representable Q(7) number.

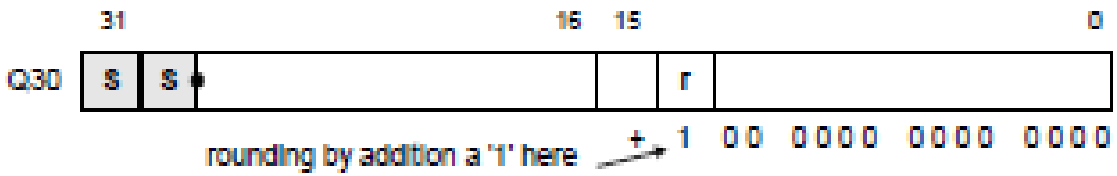
Alternative: regular rounding: round to nearest representable Q(7) number

# Illustration for multiplication in Q(15)

Multiplying two Q15 numbers gives a Q30 resultant




Add $2^{-16}$ (0x4000) and shift right by 15.




Extract the 16 lsb's bits to obtain the converted Q15 number

# Rounding in finite precision arithmetic

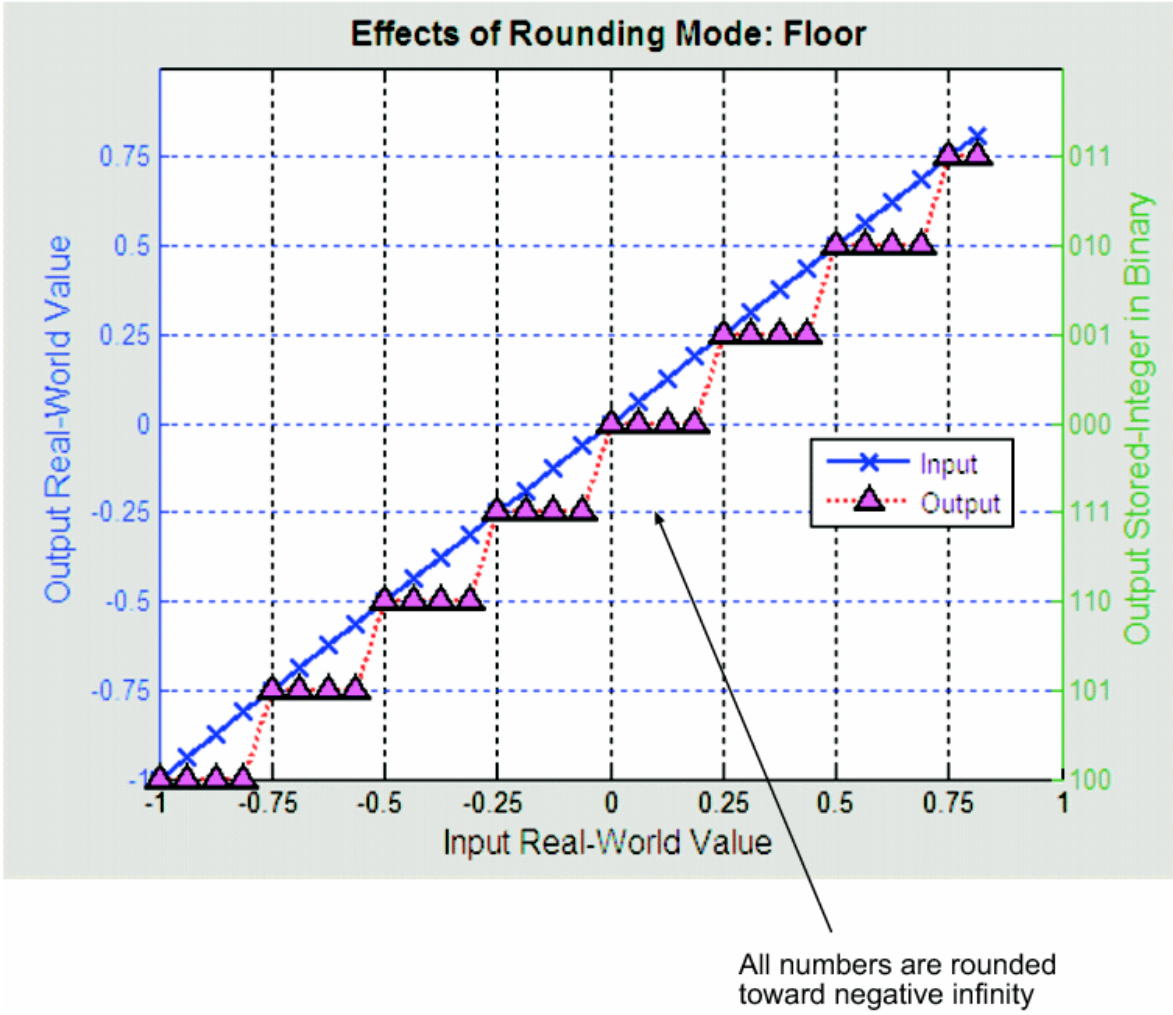
Q. Can we do better? Nearest integer rounding, or other rounding?
Conversion methods for 2's complement Q(15) to Q(7)

- Shift right and round in direction of $\infty$ (`ceil`)
- Shift right and round in direction of $-\infty$ (`floor`)
- Shift right and round in direction of nearest integer (`nearest`)
- Shift right and use convergent rounding (`convergent`)

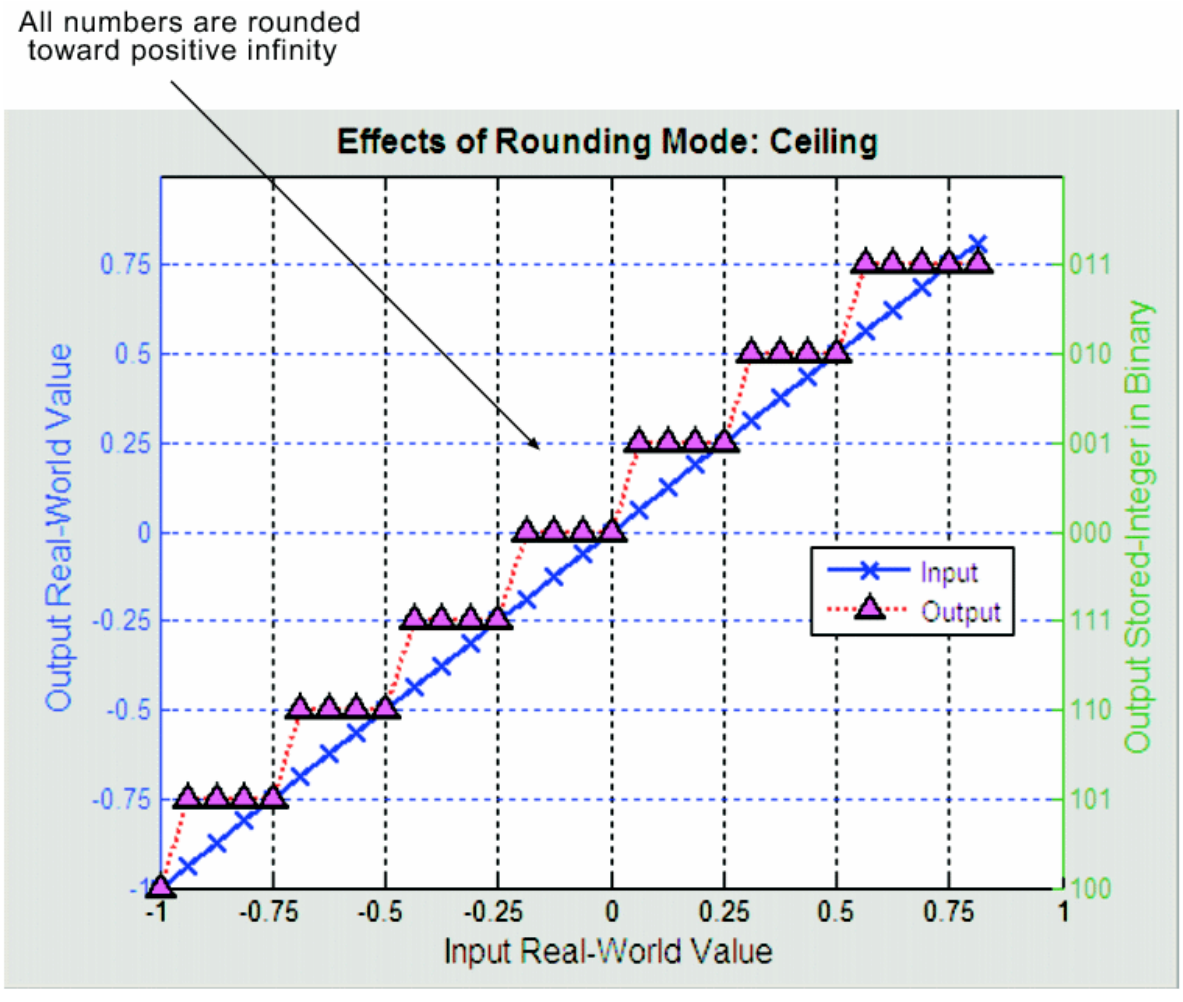
Convergent rounding has the lowest bias and rms error

Illustration: convert 2's complement 5 bit Q(4) number to a 3 bit Q(2) number.

# 2's C 5 bit Q(4) to 3 bit Q(2): down rounding

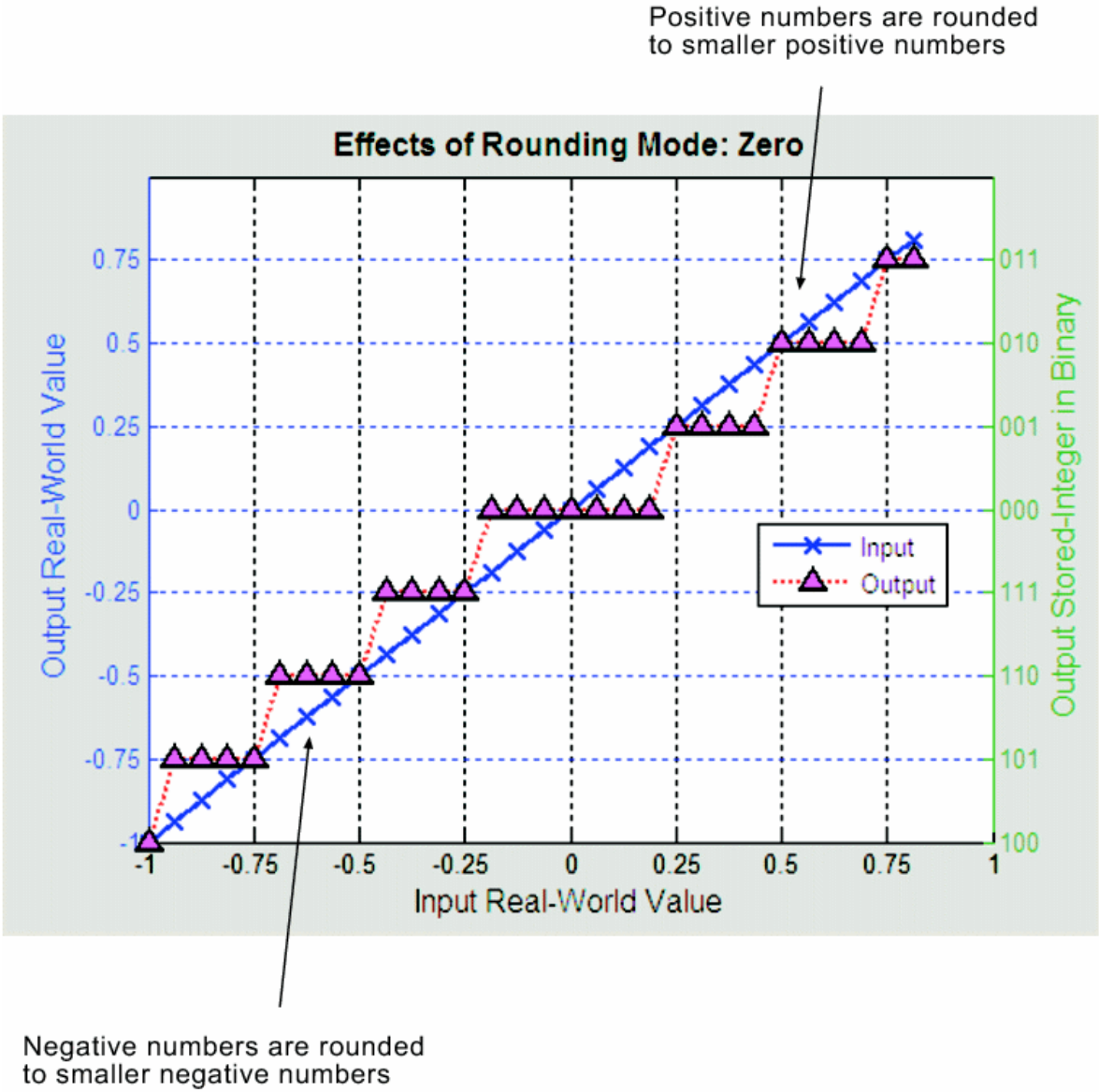


# 2's C 5 bit Q(4) to 3 bit Q(2): up rounding

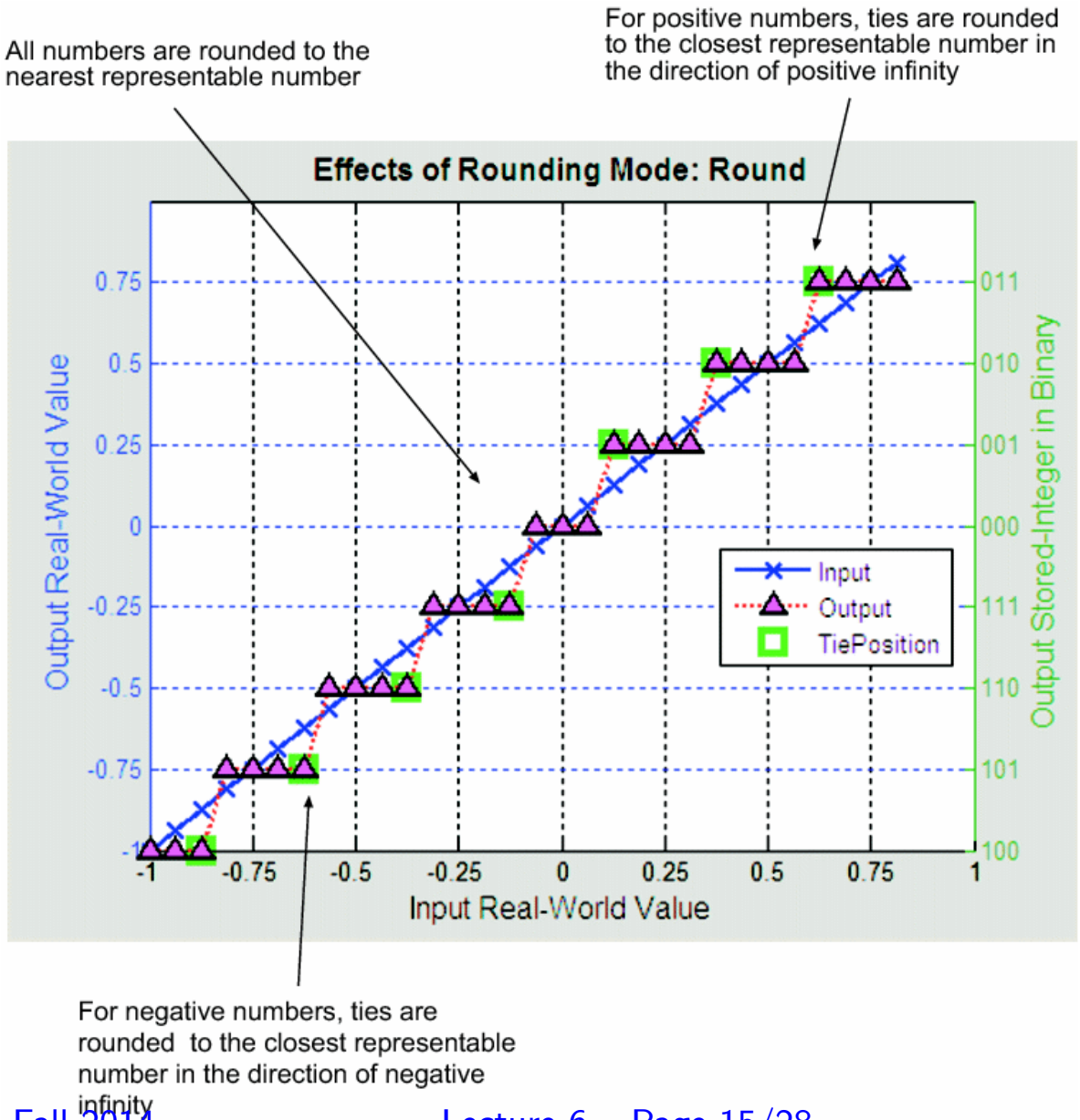



http://www.mathworks.com/help/toolbox/fixpoint/ug/f14935.html

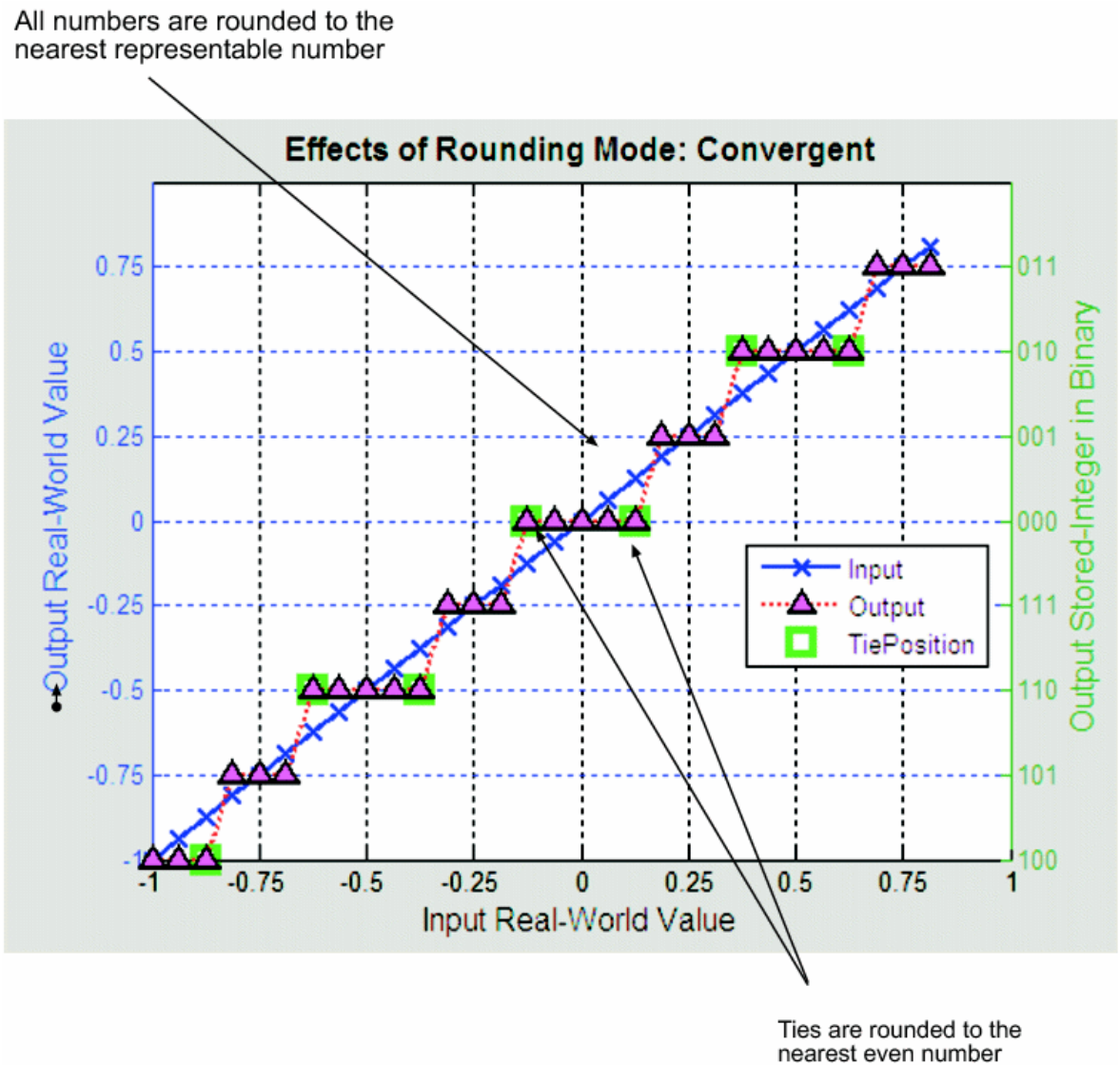
# 2's C 5 bit Q(4) to 3 bit Q(2): zero rounding



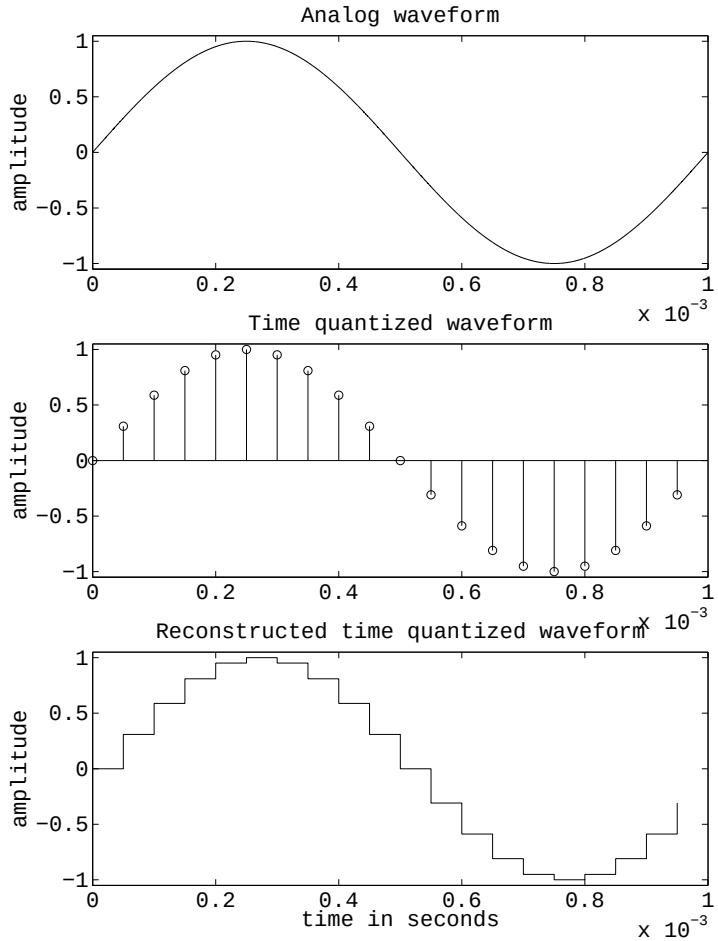
# 2's C Q(4) to Q(2): regular rounding



# 2's C Q(4) to Q(2): convergent rounding



# Visualize sampling & reconstruction

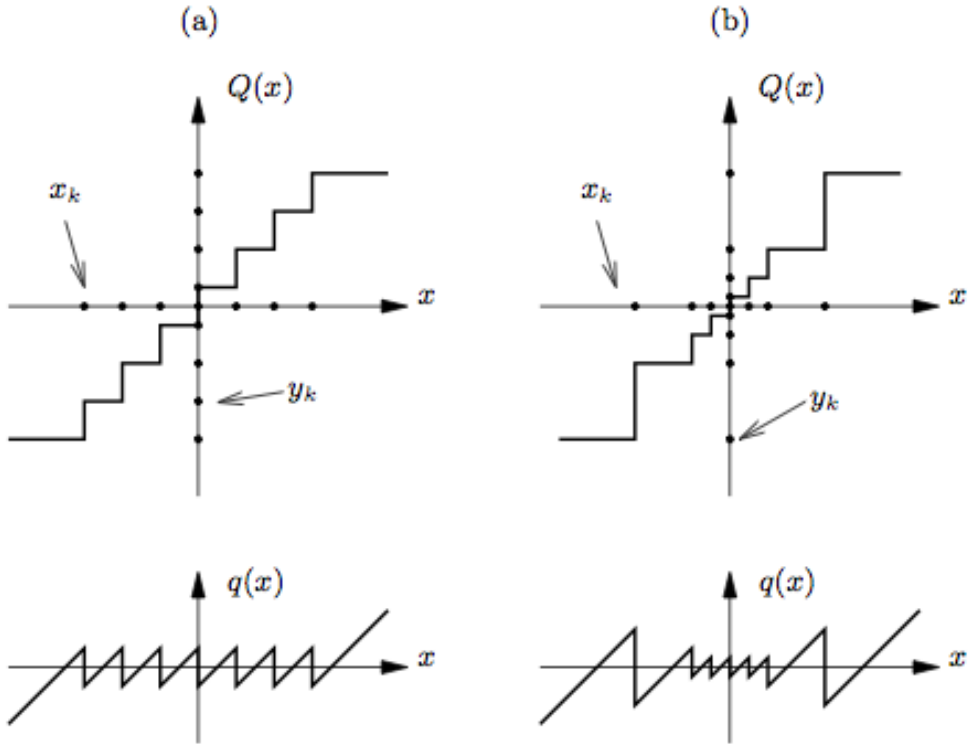



Analog waveform.

ADC discretizes amplitudes to B bits.

ADC discretizes in time to $f_s$ samples/sec

# Quantizer functions and their errors




Source: "Memoryless scalar quantization," Phil Schniter, Connexions module http://cnx.org/content/m32058/latest/.

- Top: uniform quantizer (left) non-uniform quantizer (right).
- Bottom: quantizer errors as a function of $x$.

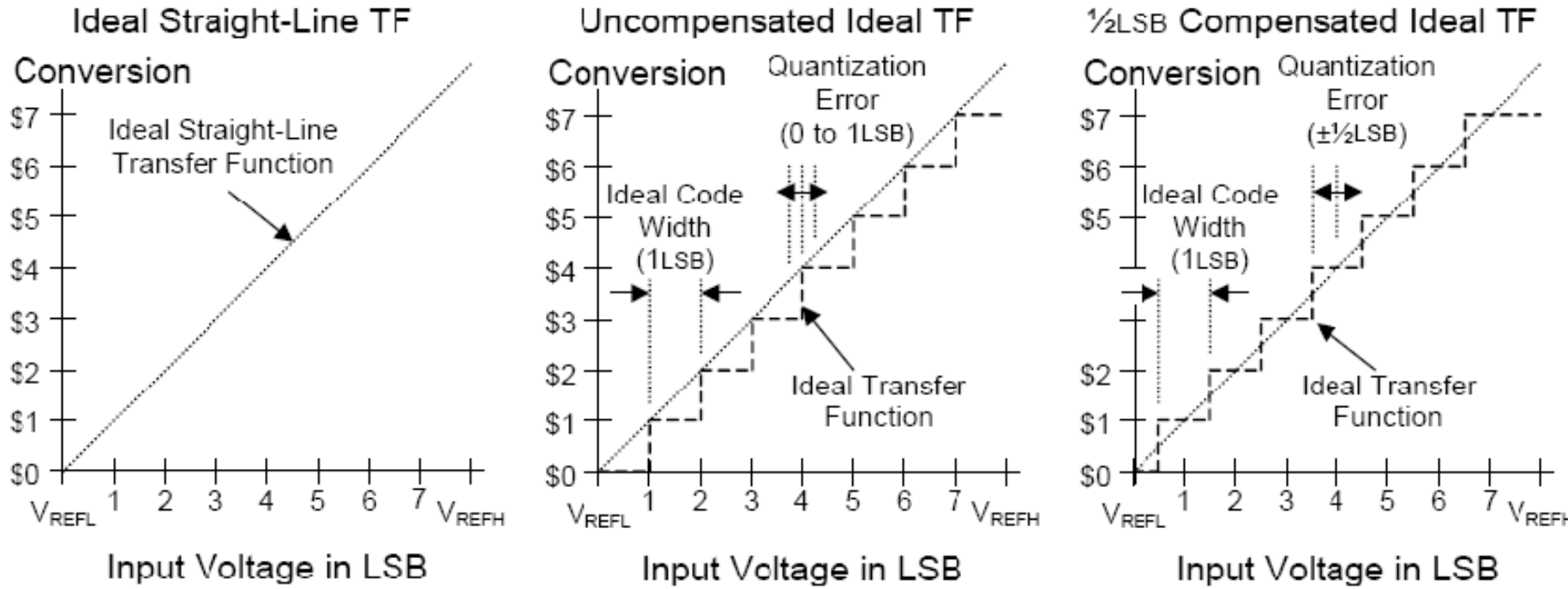
# Quantization as a mathematical function

- As shown on previous slide, quantizer is a function $Q$ mapping reals to reals.

$$Q(x) = \begin{cases} V_{REFH}, & x \geq V_{REFH} \\ y_k, & V_{REFL} + \Delta(k-1) < x \leq V_{REFL} + \Delta k \\ V_{REFL}, & x \leq V_{REFL} \end{cases}$$

- $y_k$ are quantizer levels
- $V_{REFH}, V_{REFL}$ are quantizer limits
- Length of each quantizer interval is $\Delta = (V_{REFH} - V_{REFL})2^{-B}$.
- The mapping $y = Q(x)$ is called the quantizer transfer function.

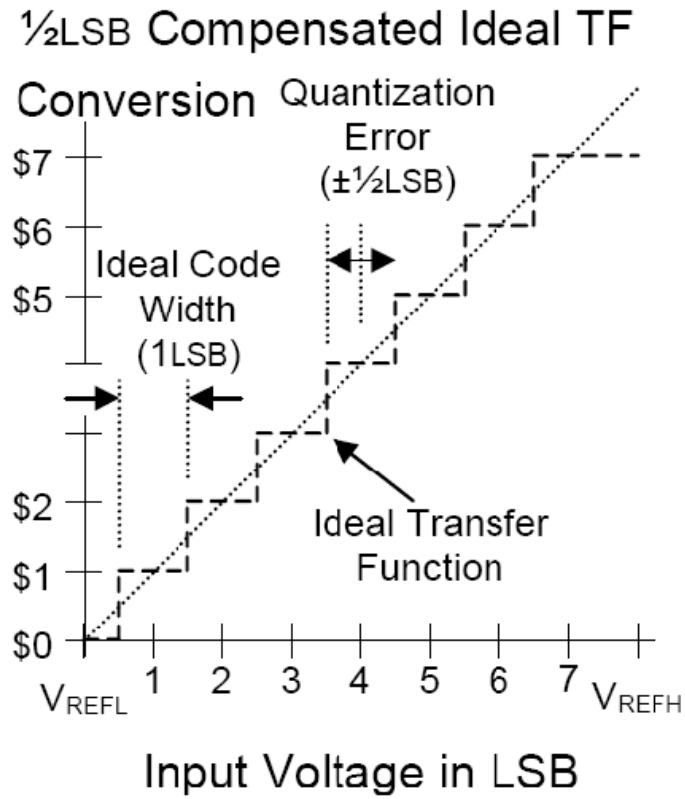
# Quantizer transfer function




Source: Freescale Semiconductor AN2438/D application notes (on course page).

- Note: Hexadecimal representation of $y$ on vertical axes.
- Ideal transfer function is linear. Cannot attain with finite B.
- Uncompensated TF corresponds to truncation conversion.
- Compensation reduces effect of quantization errors.

# Maximum QE and MSQE of uniform quantizer




- Quantization error is $e(x) = Q(x) - x$.
- Maximum quantization error[a]: $\Delta/2 = (V_{REFH} - V_{REFL})2^{-(B+1)}$
- If assume input values are uniformly distributed in $[V_{REFL}, V_{REFH}]$
  - Mean quantization error (bias) is 0.
  - Mean squared quantization error: $\Delta^2/12$ (We will show this later).

[a] Except near $V_{REFH}$ - safe to ignore for large $B$

# MSQE in dB

Quantized signal $y = Q(x)$ can be expressed as "signal plus noise"

$$y = x + Q(x) - x = x + e$$

Assume that $V_{REFL} = -V_p$ and $V_{REFH} = V_p$ (symmetric quantizer range)
Then $\Delta = 2V_p 2^{-B}$, and power of error signal $e$ is

$$P_e = \sigma_e^2 = \Delta^2/12 = V_p^2 2^{-2B}/3$$

Post-quantization signal-to-noise ratio:

$$\begin{aligned} SQNR &= \frac{P_x}{P_e} = P_x \left( \frac{2^{2B}3}{V_p^2} \right) \\ SQNR(dB) &= 10 \log_{10} SQNR = 10 \log_{10} \frac{3P_x}{V_p^2} + 10 \log_{10} 2^{2B} \\ &= \text{const. indep. of } B + 6B \end{aligned}$$

**We gain 6dB reduction of MSQE with each additional bit of resolution!**

# Bias and MSQE derivation

- Sufficient to consider first quantization cell: $0 \leq x \leq \Delta$.
- Quantization error in this cell is

$$e(x) = \begin{cases} x, & 0 \leq x \leq \Delta/2 \\ x-1, & \Delta/2 < x \leq \Delta \end{cases}$$

.

- Mean value of $e(x)$ if $x$ is uniformly distributed in cell

$$\mu_e = \frac{1}{\Delta}\int_0^{\Delta} e(x)dx = 0.$$

since $e(x)$ is antisymmetric about $x = \Delta/2$.

- Mean squared value of $e(x)$

$$\sigma_e^2 = \frac{1}{\Delta}\int_0^{\Delta} e^2(x)dx = \frac{2}{\Delta}\int_0^{\Delta/2} x^2 dx = \Delta^2/12.$$

since $e^2(x)$ is symmetric about $x = \Delta/2$

# Quantizer design issues

What can go wrong:

- Quantizer saturates (also called overload): the input $x$ exceeds $[V_{REFL}, V_{REFH}]$
- Quantizer scale and gain errors.
- Quantizer is not monotonic: an increase in the input produces decrease in output.
- Quantizer non-linearities.

What design constraints one needs to pay attention to:

- Speed (quantizer processing delays reduce digitization rate in sample/sec).
- Power (higher speed consumes more power)
- Accuracy (number of bits, linearity, saturation).
- Size (number of components).

# Scale and Offset Errors

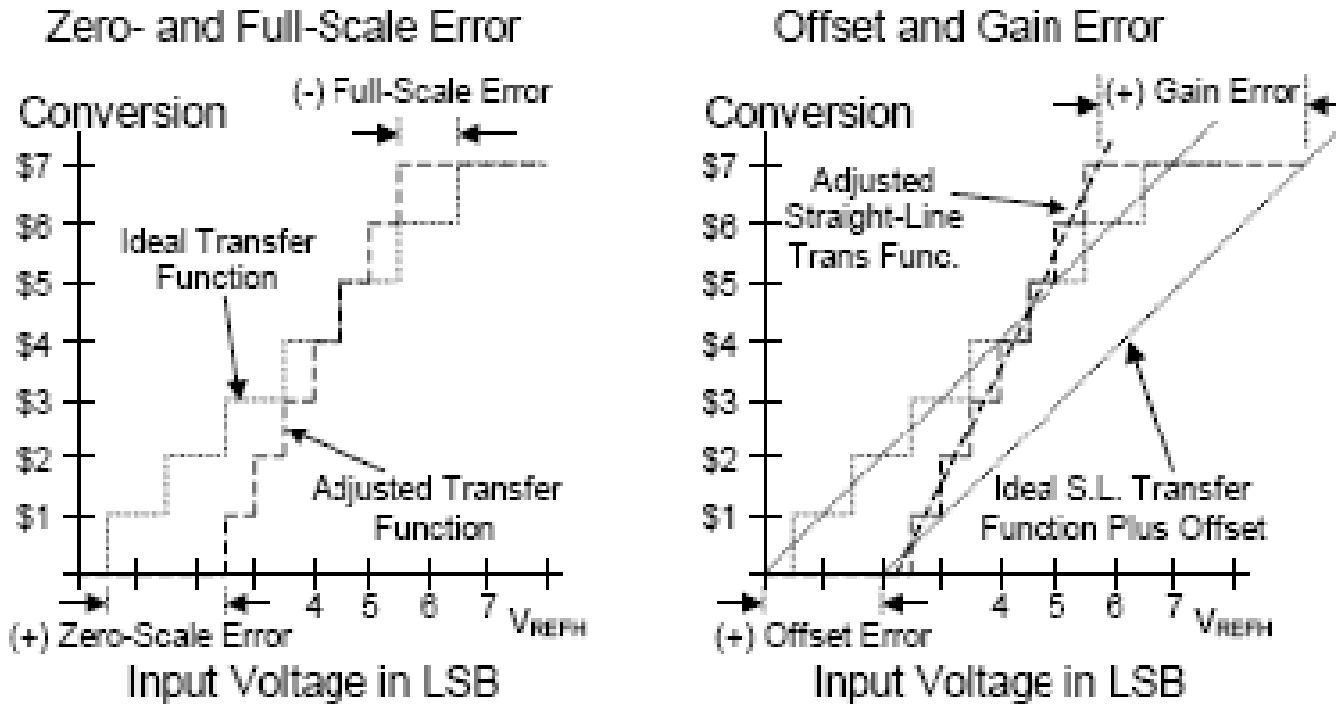
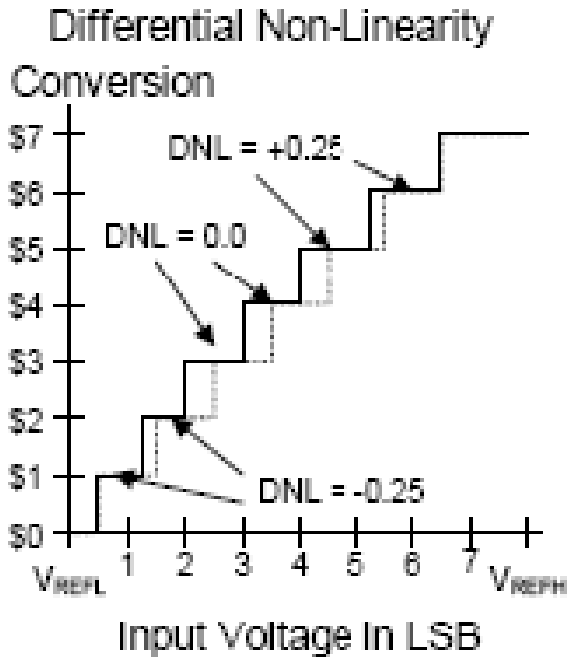
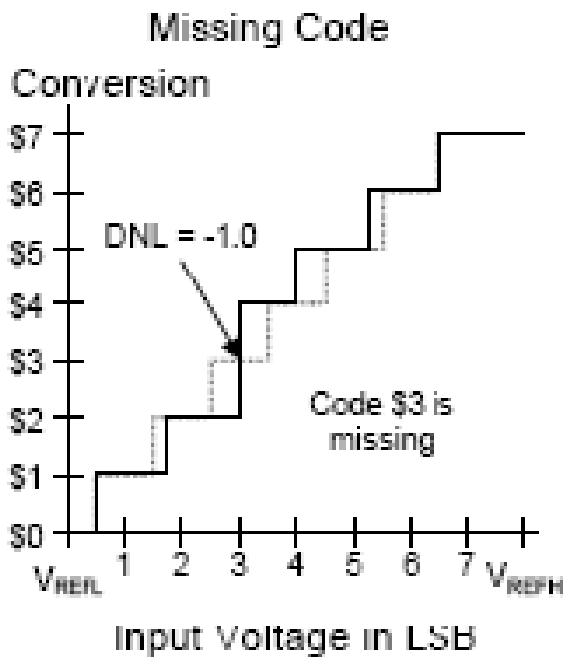


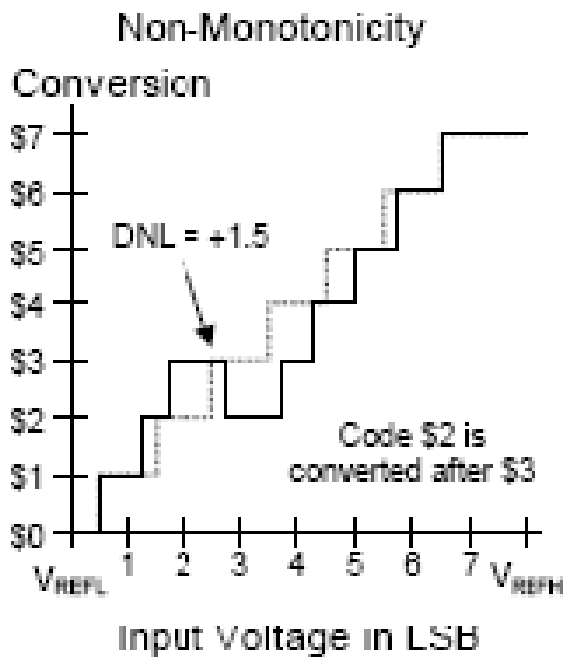

**Figure 2. Endpoints Error Graph**

- Zero scale error is due to low resolution for small input values.
- Full scale error is due to low resolution for large input values.
- Offset and gain errors causes quantizer dead zone and saturation.

# Non-linearity and non-monotonicity




- Differential non-linearity: TF deviates from the ideal diagonal line.
- Missing code: a stuck bit at output of quantizer.
- Non-monotonicity: TF fails change in concert with input.
- Note: in some cases non-linearity is intentional! ($\mu$-law companding)

# Summary of what we covered today

- Analog to digital conversion
- Converting binary numbers to lower precision: rounding errors
- ADC: quantization errors
- ADC: circuit implementations

# References

"ADC Definitions and specifications," Freescale Semiconductor AN2438/D application notes, 2003.
`http://www.freescale.com/files/microcontrollers/doc/app_note/`

"Equalizing Techniques Flatten DAC Frequency Response," MAXIM, Application Note 3853, 2006.
`http://www.maxim-ic.com/app-notes/index.mvp/id/3853`

"Understanding digital signal processing," R. Lyons, 2004.