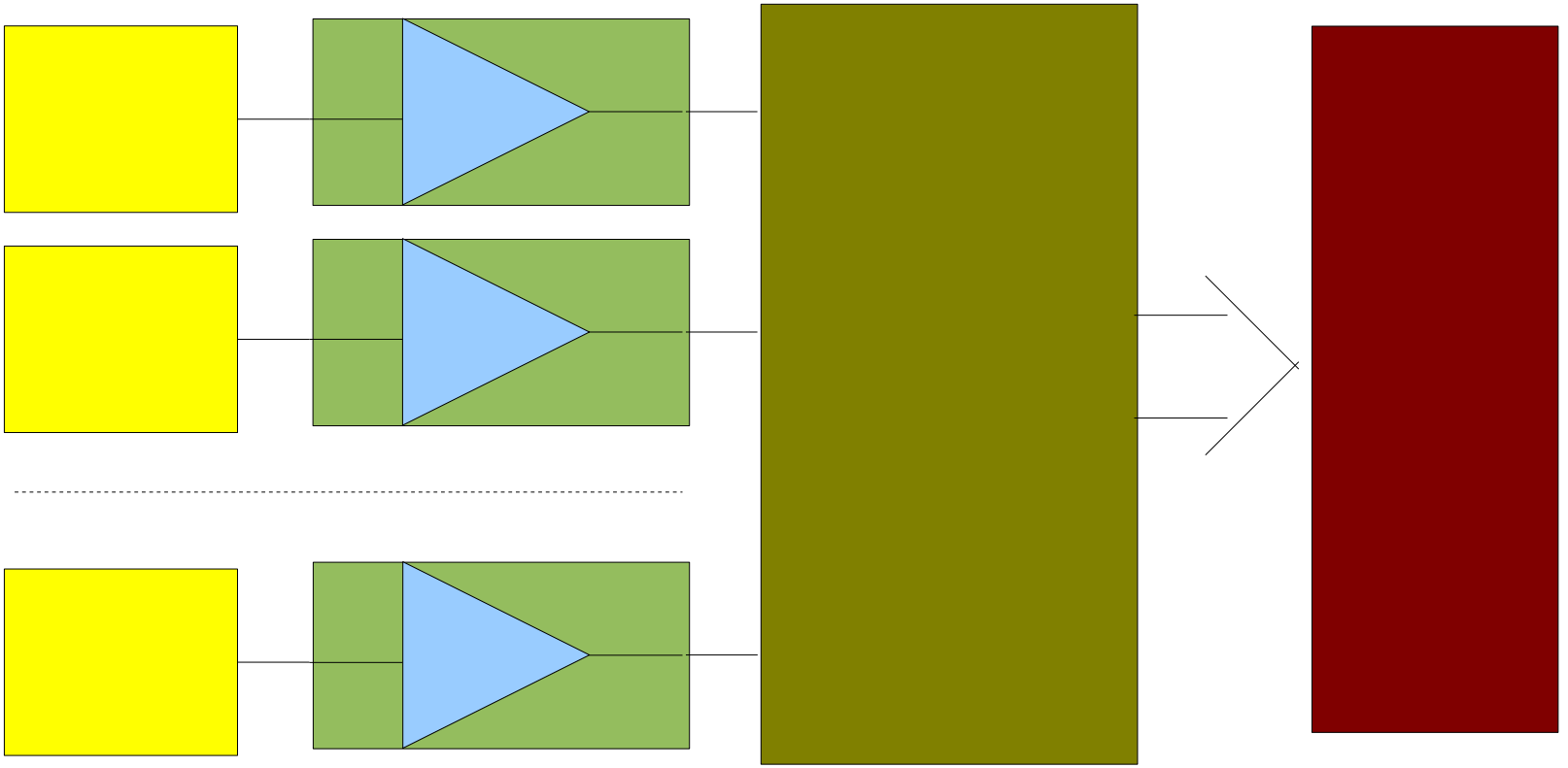
Technology trends and architectures for data driven front-end

Angelo Rivetti
INFN-Sezione di Torino, Italy

Acquisition chain



- The closer one is to the sensor, the higher is the degree of customization needed

Measurements of interest

- **Locate the event in space: sensor granularity.**
- **Locate the event in time: time resolution.**
- **Measured the deposited energy: measuring the charge released by the particle in the sensor.**

- **Not all measurements can be simultaneously required to the same system.**

Specification range

- Sensor **capacitance**: from **fF** to **nF**. (10^6)
- Signal **charge**: from **aC** to **pC** (10^4 -10^5).
- Signal **collection time**: from < **1ns** to > **1 µs.**
- Event **rate** per channel: from **Hz** to **MHz**.
- **Power consumption** per channel: from **uW** to **mW.**
- The current source contains the information on the **signal shape.**
- Required **integration density**: from a **few channels** per chip to > **10.000.**

- No single design can fulfill all possible requirements.
- One architecture can serve multiple applications.
- Different architectures have been often used to address similar requirements.

Triggered systems

- In triggered front-end, the data are stored locally in buffers, till a decision on the event is taken.
- The choice is made on the basis of the physics interest of the event.
- If the event is accepted a selection signal (trigger) is issued to the front-end and data are transmitted to the DAQ.
- Multiple selection levels can be required.
- The trigger system is made of dedicated detectors, computing nodes and software procedures.
- Optimized for particular physics channels.

Self-triggered systems

- **All data are transmitted to the DAQ (in principle…)**
- **The event selection is based on reconfigurable hardware (FPGA, GPU).**
- **Changing the trigger strategy is a firmware upgrade: no need of changing the hardware.**
- **Trigger criteria can be optimized for the physics channel under-study.**
- **More general purpose detectors.**
- **Appealing especially for systems looking at rare events.**

Triggered vs self-triggered FE

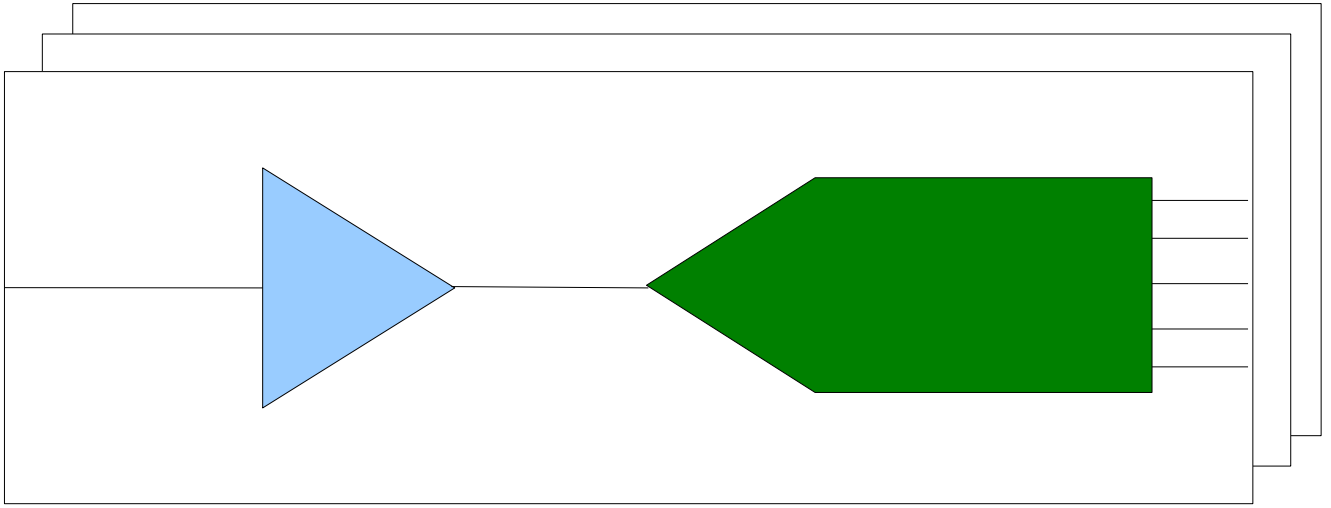
Triggered

- Making a trigger decision a propagating it back to the front-end requires time: latency.
- The front-end electronics needs to have enough data buffering to keep the data for the latency time.
- A trigger matching logic is needed (but this is usually simple).
- Only a few data are transmitted off-chip: output bandwidth is modest.

Self-triggered

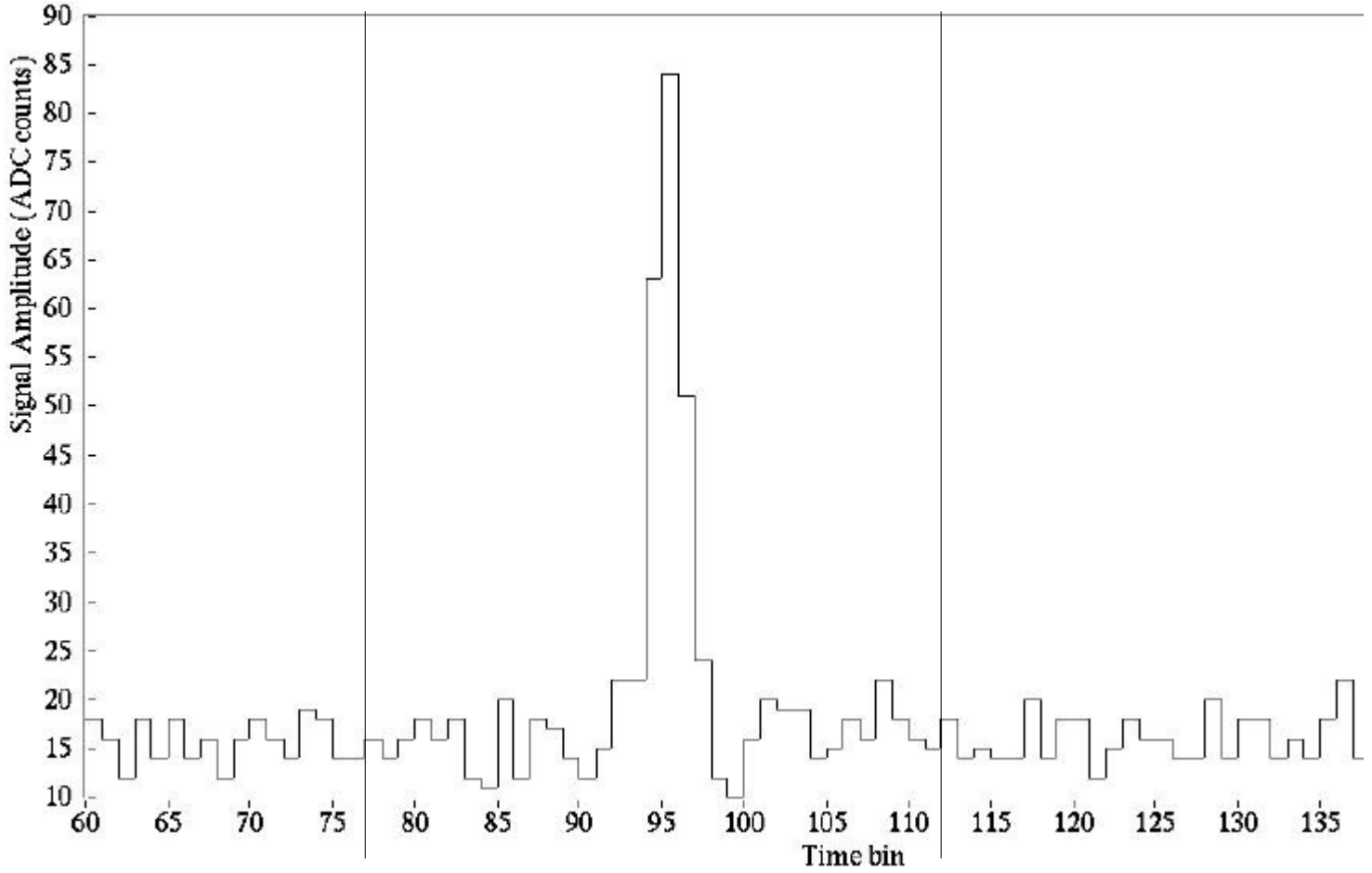
- As much data as possible transmitted of chip.
- Large output bandwidth needed.
- Data need to be buffered while queuing for readout.

A first system

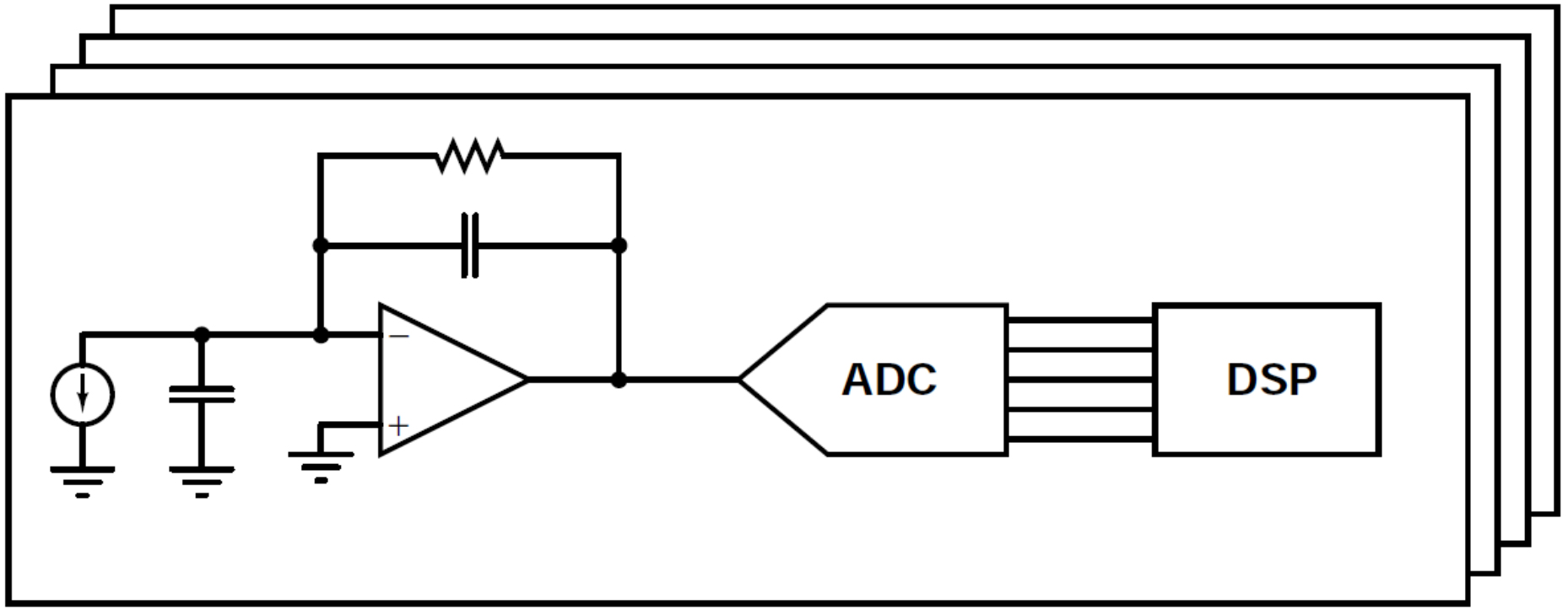


- The front-end output is immediately digitized by a fast ADC. All samples transmitted off-line. Good idea?
- Assuming 10 bit ADC@50 MS/s: 500 Mbit/sec of transmission bandwidth per channel. Assuming a 64 channel chip, 32 Gbit/sec/chip.
- A lot of power to transmit noise out of the front-end!!
- Need to put at least a cut on noise.

A digitized pulse

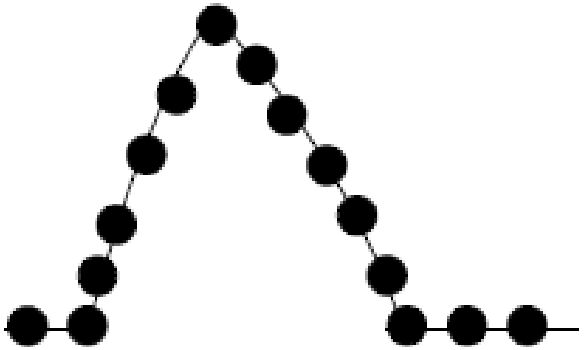


A better system



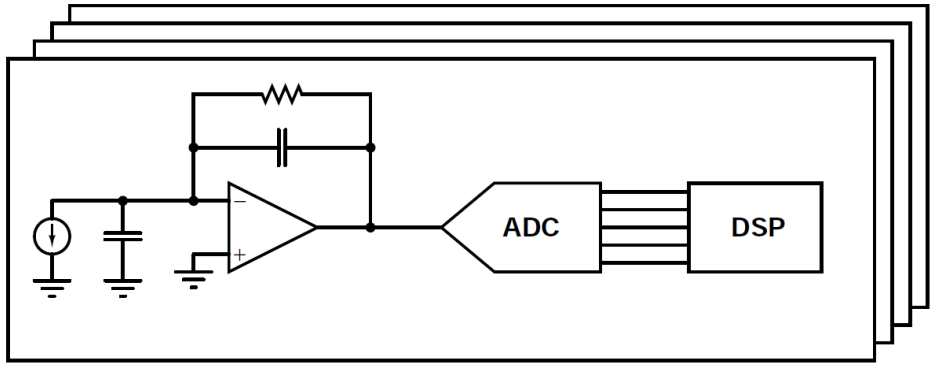
- Output of front-end amplifier **directly digitized** by the ADC.
- Only samples **over noise** or in the **interesting region** are **kept**.
- Possibility of applying **corrections** before event selection:very interesting to reject **common mode** noise.
- Accuracy of **feature extraction** does **not** depend on the accuracy of the **analog components**.

Operation on a digitized pulse



- Each sample is ordered in time attaching to it a time stamp. Provided by a counter.
- Energy can be measured by identifying the peak
- Timing information can also be extracted by waveform sampling.
- The clock counter allows a rough measurement.
- The amplitude information is used for interpolating.
- The fine time can be extracted by different methods (extrapolating towards zero, center of gravity, digital constant fraction discriminators).
- Rule of thumb: accuracy about 5% of the waveform rise time.
- Only a few critical operations can be done on chip, demanding the most complex feature extraction to FPGA.

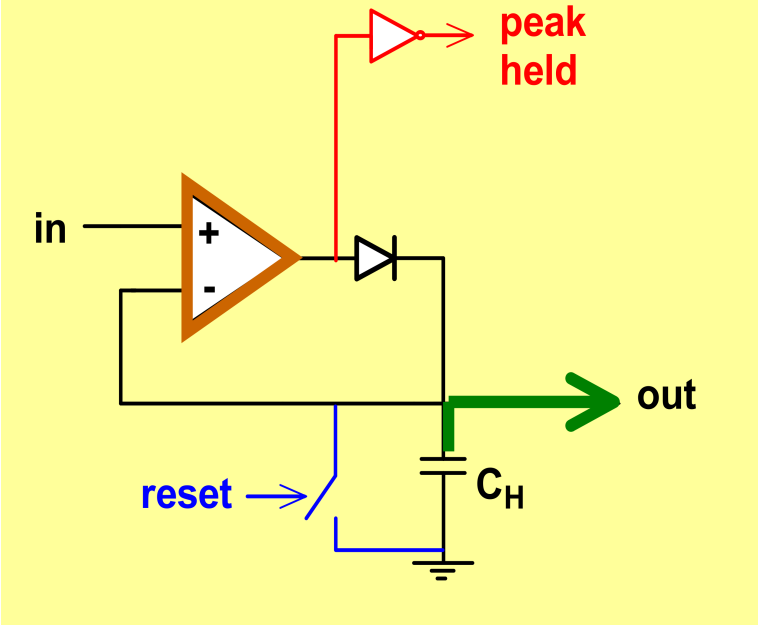
Why they are not so popular?



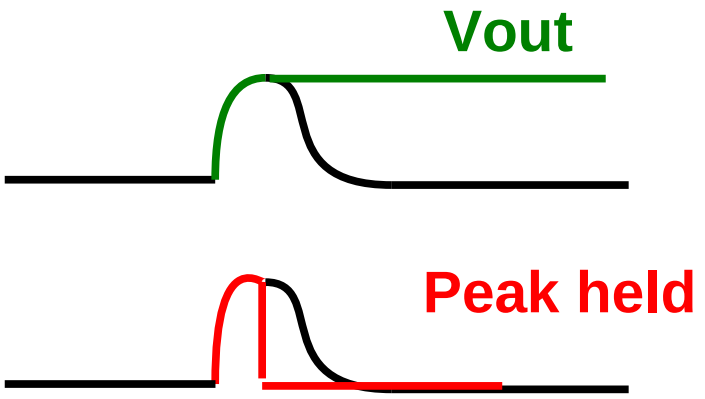
- Key elements:
  - Front-end amplifier
  - ADC
  - Digital processing unit

- A full-sampling system requires a fast low-power ADC.
- ADC used to be bulky and power-hungry components.
- Not suitable for most applications with radiation detectors, which require low-power, unless few bits were used.
- In old technologies, also digital was bulky and power hungry.
- Fear of having sensitive analog and complex digital on the same chip.
- But things are changing....

Analogue alternatives (1)



- Principle: retain only the basic features.
- Preserve only the peak amplitude and the time of occurrence.
- The system must be able to self-detect the peak.



- Method: introduce a unidirectional element in the feed-back loop of an opamp.
- The voltage on C_H follows the input till to the peak and then stays constant.
- Design is not straightforward, but very good results in term of accuracy and event rate capabilities have been achieved.

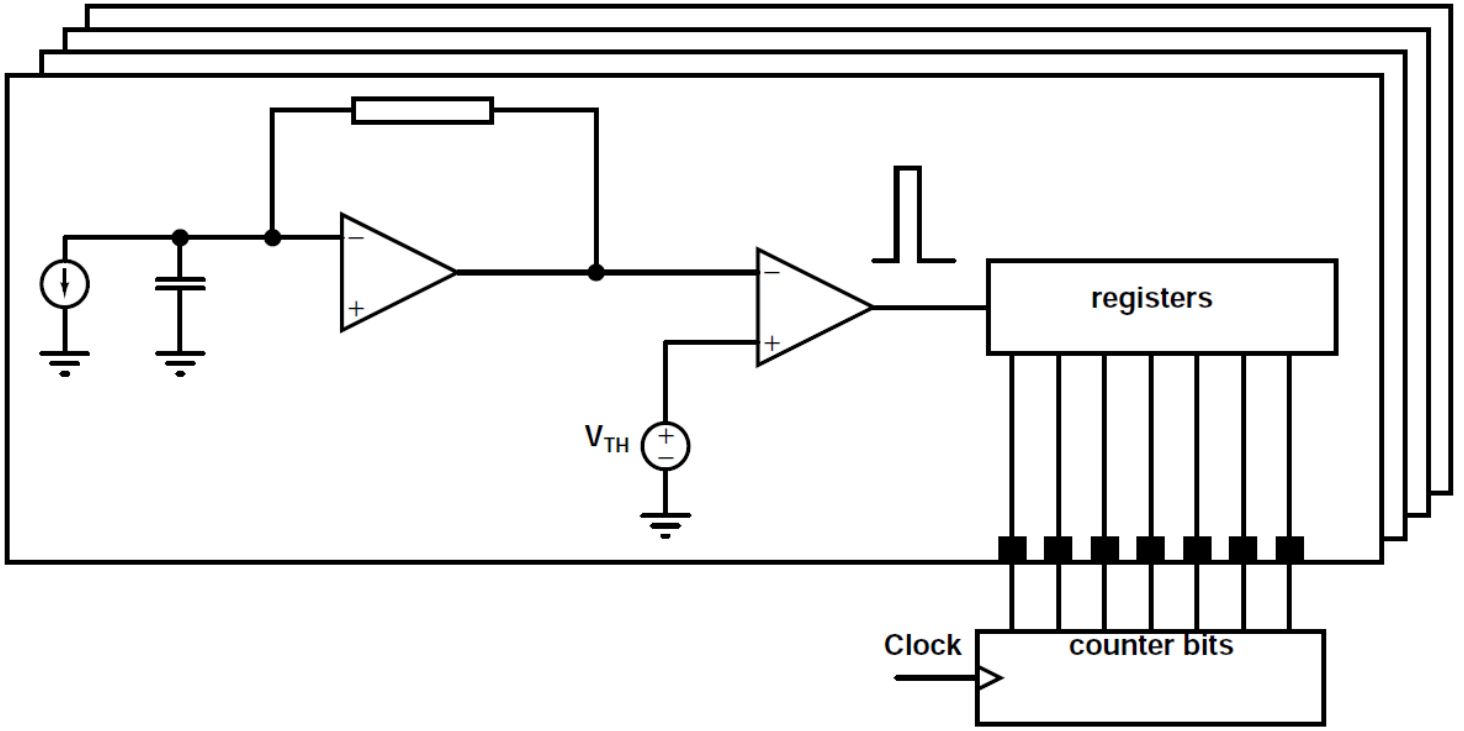
Analogue alternatives (2)

G. De Geronimo, P. O'Connor, A. Kandasamy:
“Analog CMOS peak detect and hold circuits.
Part I. Analysis of the classical configuration”
Nucl. Instr. Methods A 484 (2002) pp. 533-543.

“Analog CMOS peak detect and hold circuits.
Part II. The two phase offset free and derandomizing configuration”.
Nucl. Instr.Methods A 484 8 (2002) pp. 544-556.

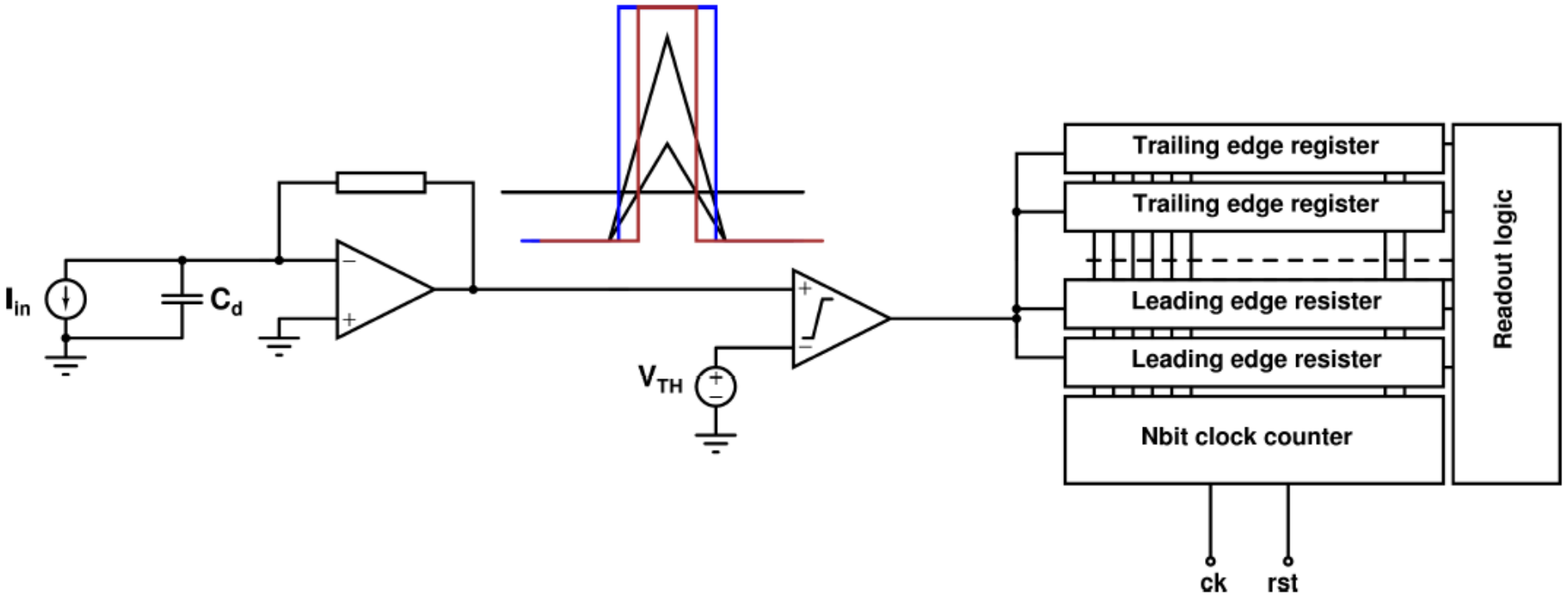
“Analog peak detectors and derandomizer for high rate spectroscopy”.
IEEE Trans. Nucl. Sci. vol. 49, n. 4, August 2002, pp. 1769-1773.

Minimalistic approach



- **Key elements:**
  - **Front-end amplifier**
  - **Discriminator (comparator)**
  - **Counter and registers**
- **Provides:**
  - **Geographical location.**
  - **Coarse timing information.**
  - **Power at high ck speed.**

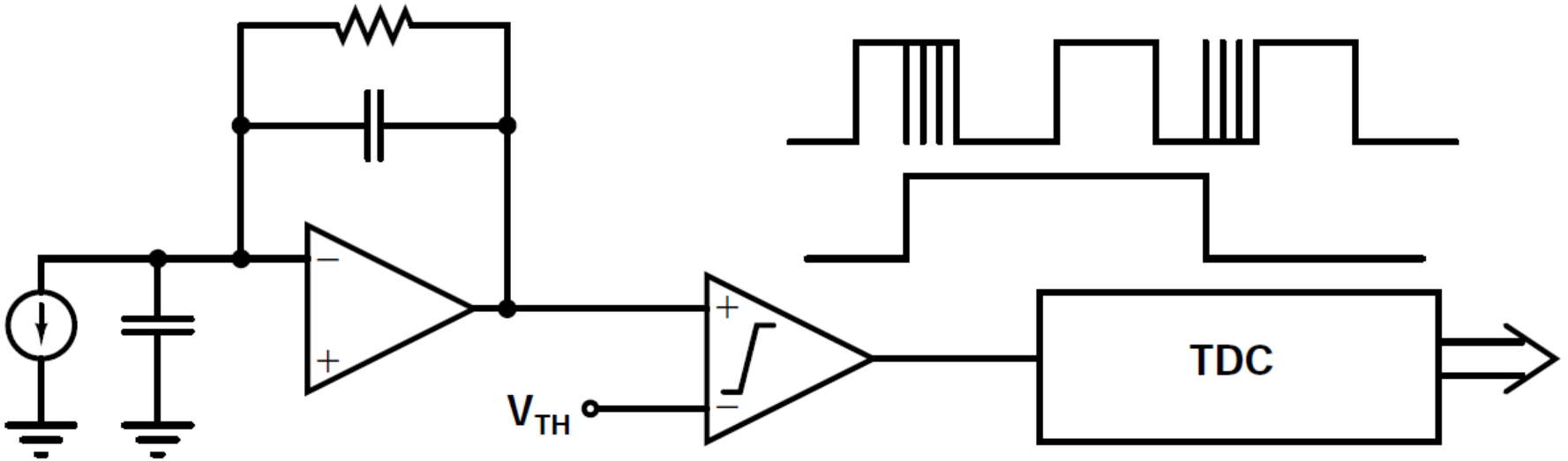
Time over Threshold (ToT)



- The **duration** of the comparator pulse is in general **proportional** to the signal amplitude.
- With suitable techniques, a **linear relationship** can be achieved.
- The time of the **leading and trailing edges** are measured.
- The **difference** provides the ToT and **hence** the charge information.

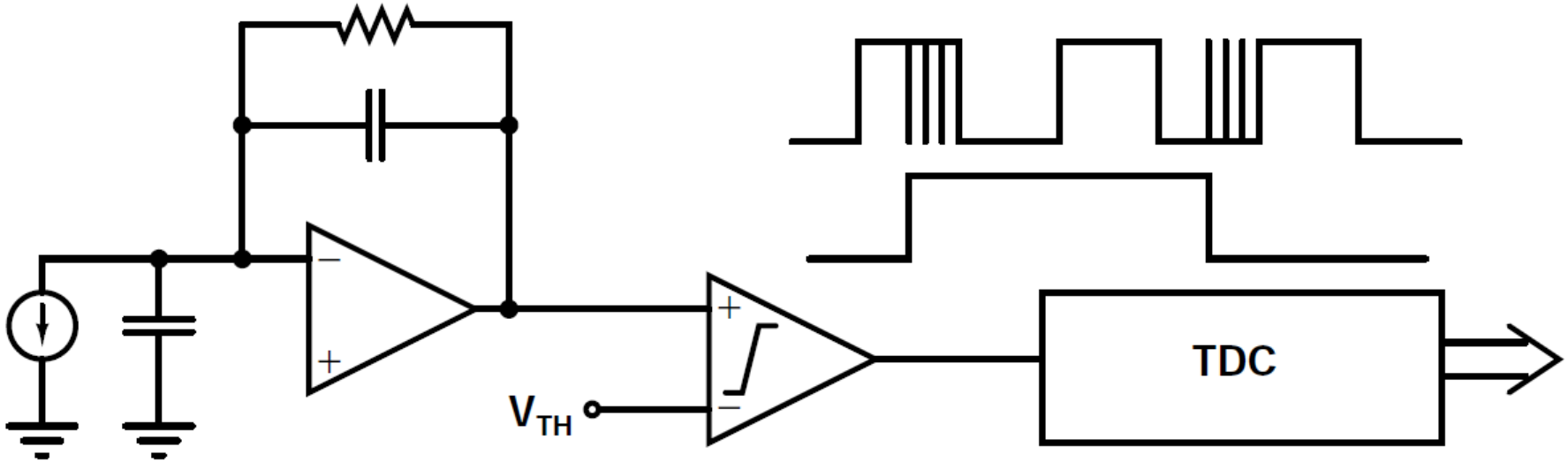
Increasing the time resolution

- Clock counting is effective for **low to moderate** time resolution (O(ns)).
- Modern microprocessors run at **3 GHz**. Using a 3 GHz clock yields in principle **300 ps** of time binning and **100 ps** of rms resolution.
- Issue: power consumption associate to the clock/time stamp **distribution network.**
- In several applications a **better time resolution** is mandatory.
- A better approach: measure the time elapsing between the signal and one clock edge → **Time to Digital Converter (TDC)**.



ToT with TDC

- Time to digital converters can be used to capture both the leading and the trailing edges with high resolution.
- Possibility to measure the duration of fast pulses with high accuracy.
- Fast time-based readout systems with dead-time limited by the front-end.



Key ingredients

- Low noise front end amplifiers.
- High performance comparator.
- Low Power Analog to Digital Converters.
- Low Power Time to Digital Converters
- Fast data transmission systems.
- FPGA.

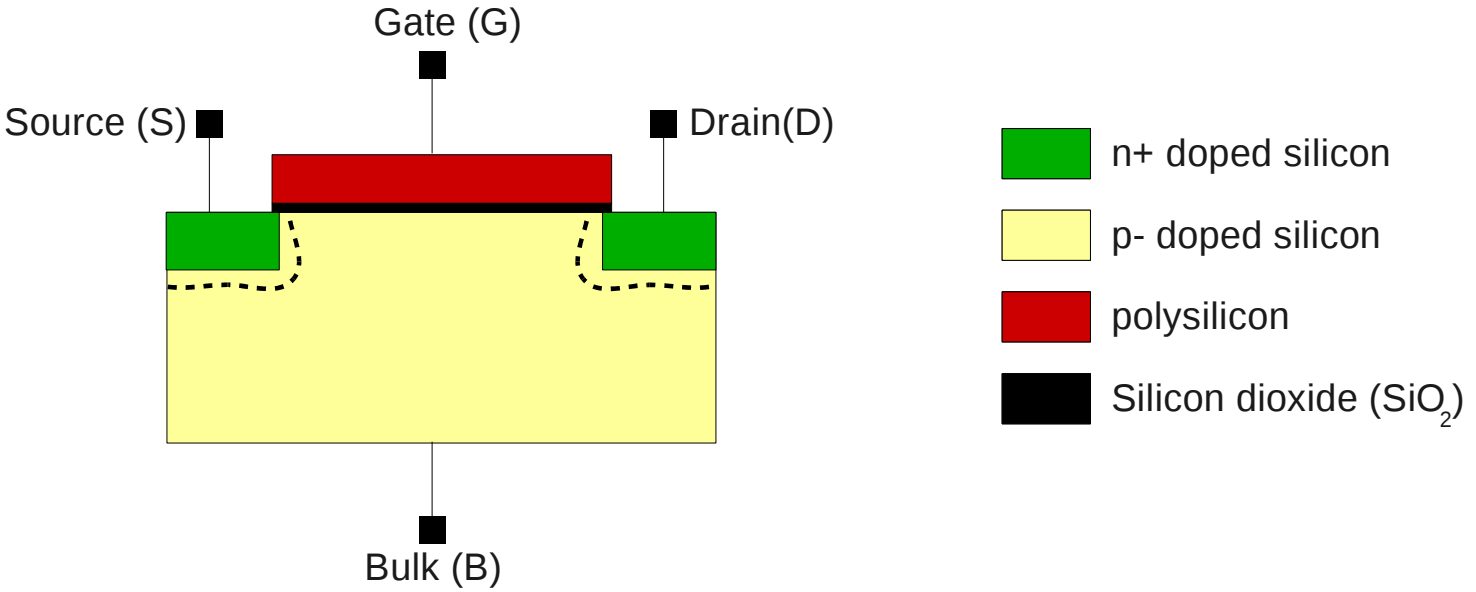
Which technology? CMOS!

- CMOS technologies dominate today the market of integrated circuits:
  - Ability to integrated on the same substrate PMOS and NMOS devices.
  - CMOS transistors are fairly easy to squeeze.
- Integration density doubles roughly every 24 months (Moore's Law).
- Scaling makes transistors faster: CMOS has progressively replaced bipolar transistor in a growing number of applications, including RF.

Why CMOS in front-end electronics

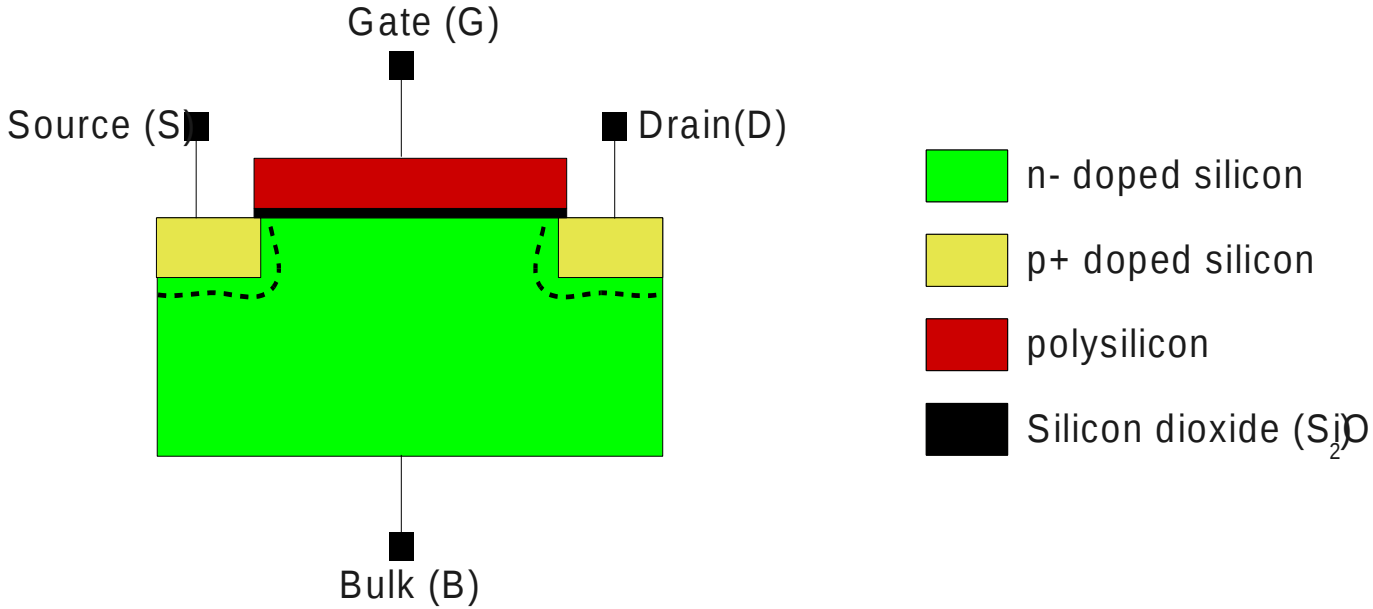
- CMOS was always the preferred choice to implement front-end for radiation detectors:
  - Easily available and cheap.
  - Allows the fabrication of good sampling circuits.
- CMOS became the solution for implementing front-end electronics in radiation-sensitive environments:
- Reason: thin gate oxide makes transistors less sensitive to radiation damage.
- Deep sub-micron front-end:
  - First generation: CMOS 0.25 µm. Now in the LHC detectors. Radiation hard with enclosed layout.
  - Second generation: CMOS 130 nm. Radiation hard even with standard layout. Increased functionality, smaller area.
  - Next step: 65 nm. Much increased functionality.

The NMOS transistor



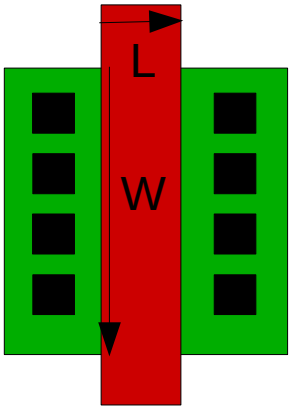
- The NMOS transistor is built over a p- doped substrate.
- Two n+ doped electrodes (source and drain) are implanted in the substrate.
- The electrodes form with the substrate np junctions, that must always be reversed biased.
- A depletion region is associated with the junction.
- A thin oxide layer is grown between source and drain.
- A polysilicon layer (gate) is grown on top of the oxide.

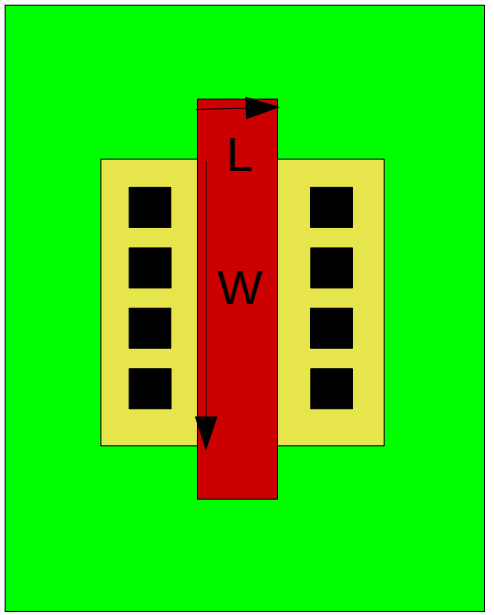
The PMOS transistor



- The PMOS transistor is built over a n- doped substrate.
- Two p+ doped electrodes (source and drain) are implanted in the substrate.
- The electrodes form with the substrate pn junctions, that must always be reversed biased.

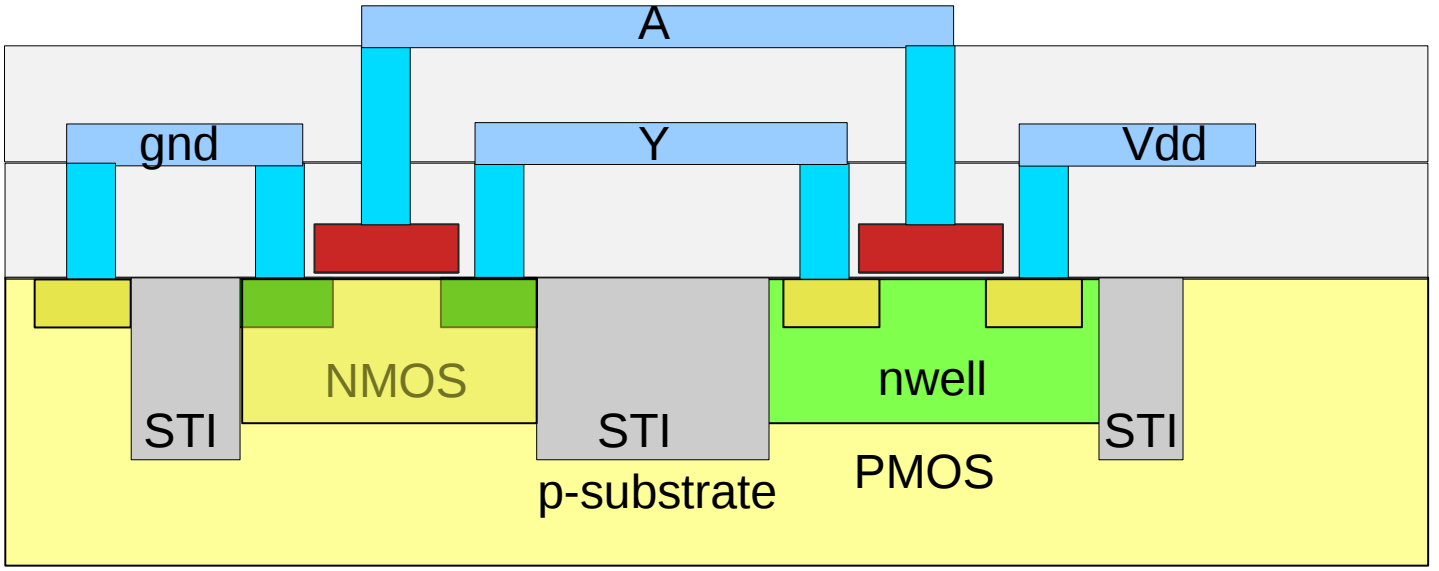
CMOS technologies





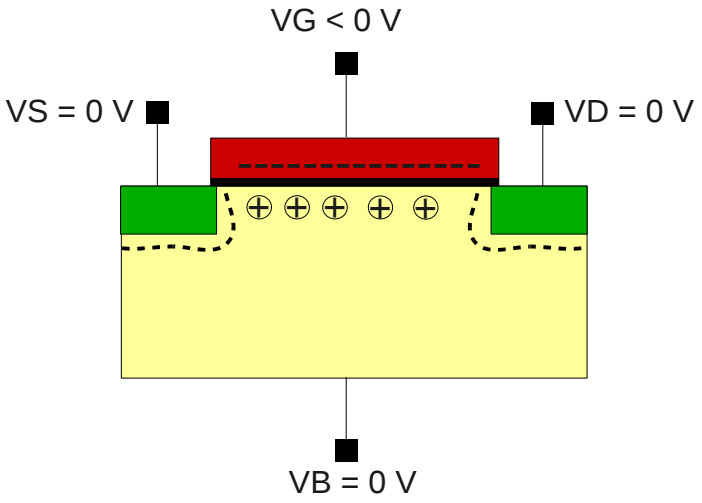
- CMOS technologies allow simultaneous fabrication of both NMOS and PMOS transistors on the same silicon wafer.
- The wafers substrate is typically p-
- Selected zones of the wafers are counter-doped to become n-type a provide the substrate for PMOS devices.
- Channel=gate AND diffusion
- Minimum possible L is use to define a specific technology node

CMOS technologies

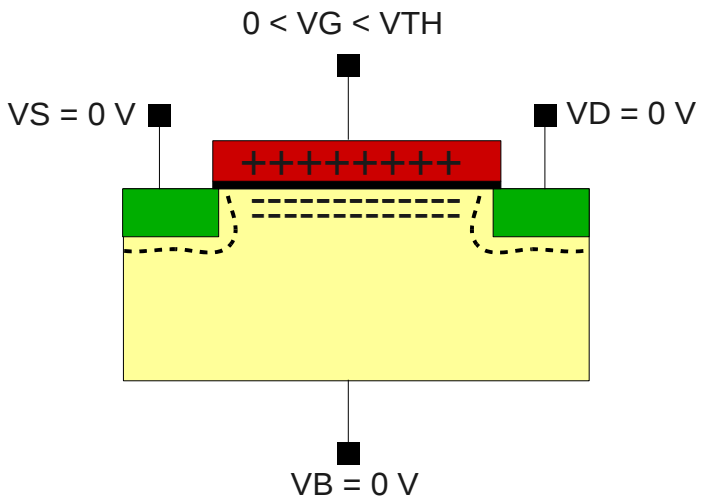


- In modern technologies devices are insulated with shallow trenches (STI).
- To form circuits, device need to be interconnected with metals.
- Modern technologies offer up to 8-9 layers of metals.
- Metal can be either Al or Cu.
- Different metal layers are interconnected where needed through vias in the silicon dioxide.
- Planarization of the Inter Level Dielectric (ILD) is a critical issue.

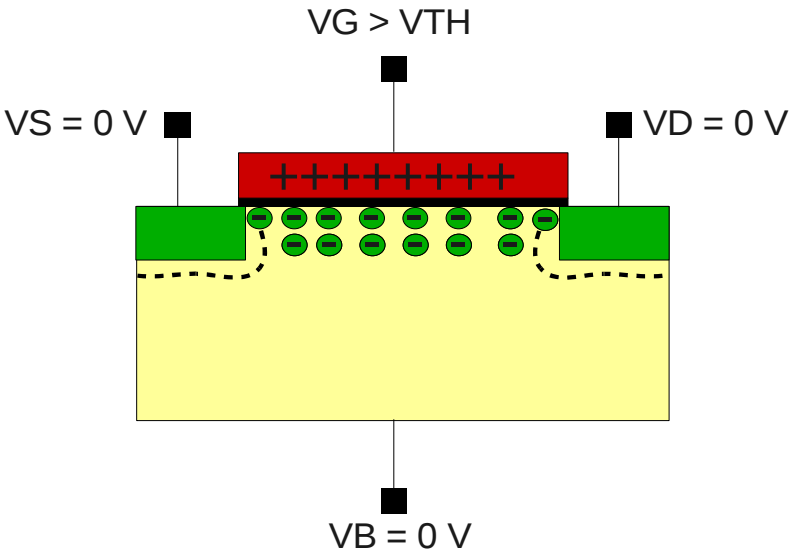
The capacitor analogy



Accumulation



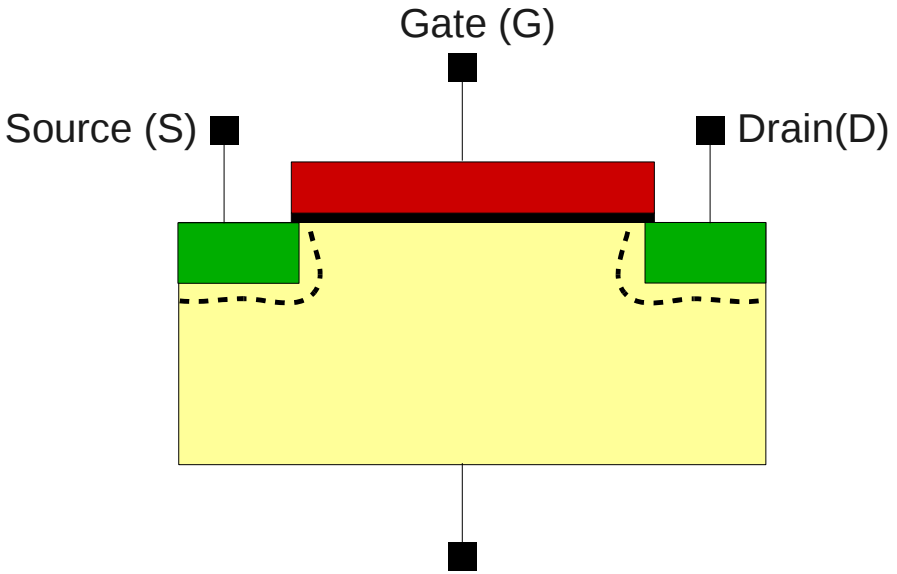
Depletion

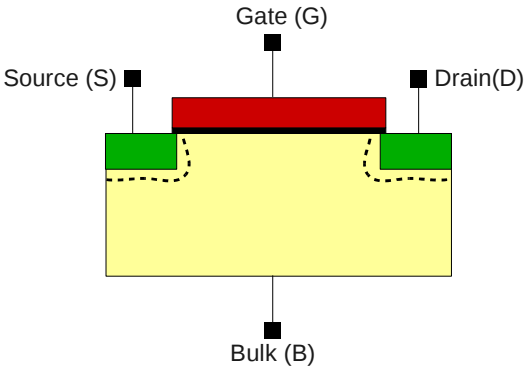


Inversion

The higher the bulk doping, the greater the device threshold.

Scaling CMOS transistors





α>1

Device width and length	→	1/α
Depletion region	→	1/α
Oxide thickness	→	1/α
Threshold voltage	→	1/α
Power supply	→	1/α
Cox	→	α
Device density	→	α^2

$$x_d = \sqrt{\frac{2\epsilon_s}{qN_A}(\phi_B + V)}$$

$$C_{OX} = \frac{\epsilon_{OX}}{t_{OX}}$$

Consequences of scaling

- For **digital circuits**: things can only **get better**! $P=f * C * V^2$
  - Scaling is optimized for digital circuits.
  - Digital gates becomes **smaller**, **faster** and lower power.
  - Note: while **power/gate decreases, power/chip increases** to the increased number of gate and **augmented functionality.**
- For **analog** circuits situation is **more complex**.
- Analog circuits sensitive to a number of effects arising in the deep sub-micron regime.
- For analog circuits, key parameters are:
  - Transconductance g_m: $\Delta I_{DS} = \frac{\partial I_{DS}}{\partial V_{GS}} \Delta V_{GS} = g_m \Delta V_{GS}$
  - Output conductance g_{ds}: $g_{ds} = \frac{\partial I_{DS}}{\partial V_{DS}} \approx \lambda I_{DS}$
  - Noise and matching...
- The g_m/g_{ds} ratio is a metric of a gain that can be achieved by CMOS amplifiers.

Transconductance

Strong inversion

$$I_{DS} = \frac{1}{2}\mu_n C_{OX}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \qquad g_m = \sqrt{2\mu C_{ox}\frac{W}{L}I_{DS}}$$

Weak inversion

$$I_{DS} = I_0\frac{W}{L}e^{\frac{V_{GS}}{nU_T}}(1 - e^{-\frac{V_{DS}}{nU_T}}) \qquad U_T = \frac{kT}{q} \qquad g_m = \frac{\partial I_{DS}}{\partial VGS} = \frac{I_{DS}}{nU_T}$$

$$I_C = \frac{I_{DS}}{2n\mu C_{ox}\frac{W}{L}U_T^2 I_{DS}}$$

I_C < 0.1: W.I.

In between: moderate inversion!

I_C> 10: S.I.

Gm in weak inversion

$$I_C = \frac{I_{DS}}{2n\mu C_{ox}\frac{W}{L}U_T^2 I_{DS}}$$

- Scaling the technology, C_{ox} increases.
- For the same bias current and aspect ratio, I_C decreases
- Transistors work more and more in weak inversion.
- g_m/I_d is maximized
- MOS resembles a bipolar transistor

Continuity across level of inversion

- **The EKV model was developed in an attempt to have a physical based model providing continuity of MOS characteristics across different level of inversion:**

MOS equation in saturation valid in all region of inversion

$$I_{DS} = 2n\mu C_{ox}\phi_T^2 \frac{W}{L}\left[\ln\left(1+e^{\frac{V_{GS}-V_{TH}}{2n\phi_T}}\right)\right]^2$$

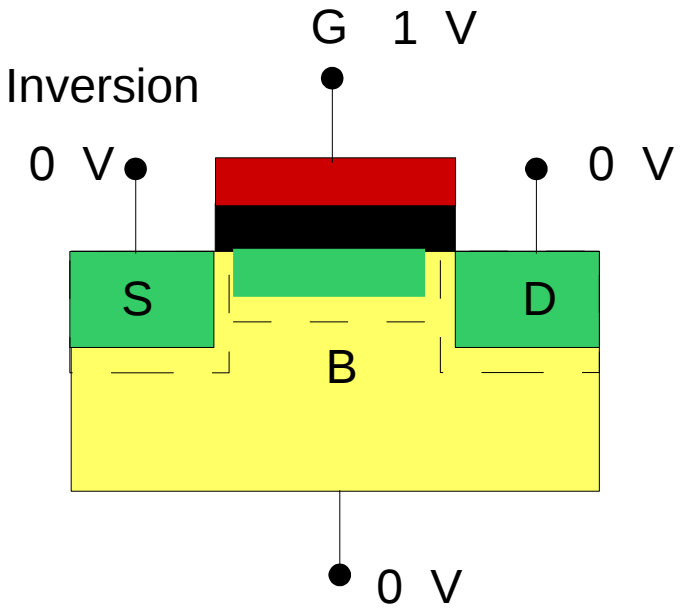
Trasconductance

$$g_m = \frac{I_D}{n\phi_T}\frac{1}{\sqrt{I_C}+0.5\sqrt{I_C}+1}$$

Key short channel effects

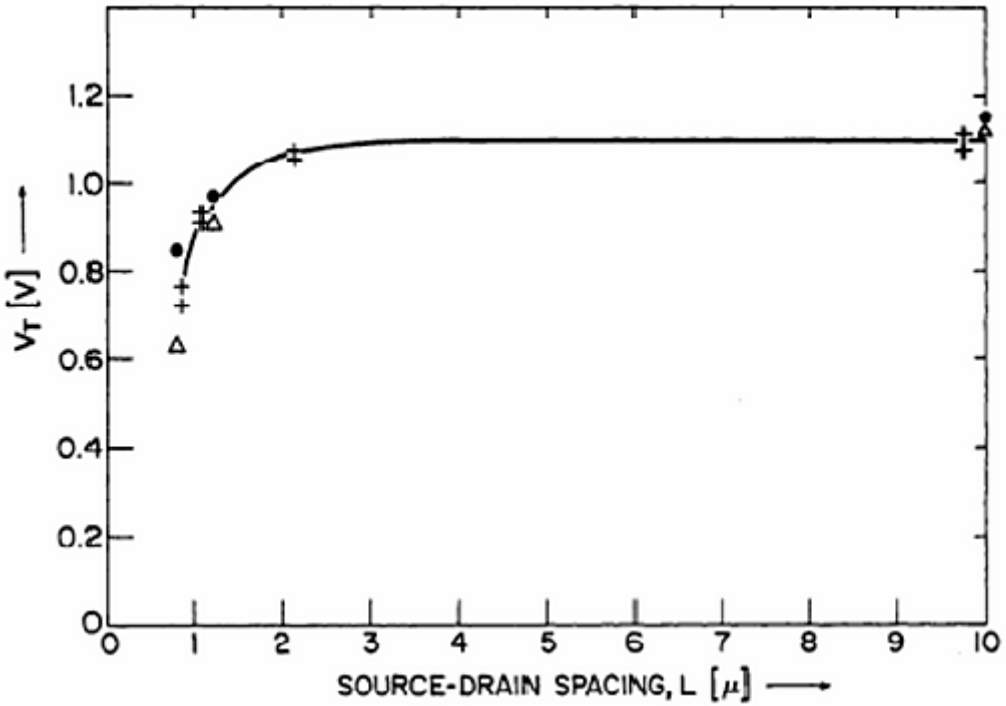
1. Drain-induced barrier lowering: **reduces g_{ds}**
2. Surface scattering: **reduces** mobility/**g_m**
3. Velocity saturation: **reduces g_m**
4. Impact ionization:substrate current, **reduces g_{ds}**
5. Hot electrons: long term **reliability**

Example: reverse short channel (1)

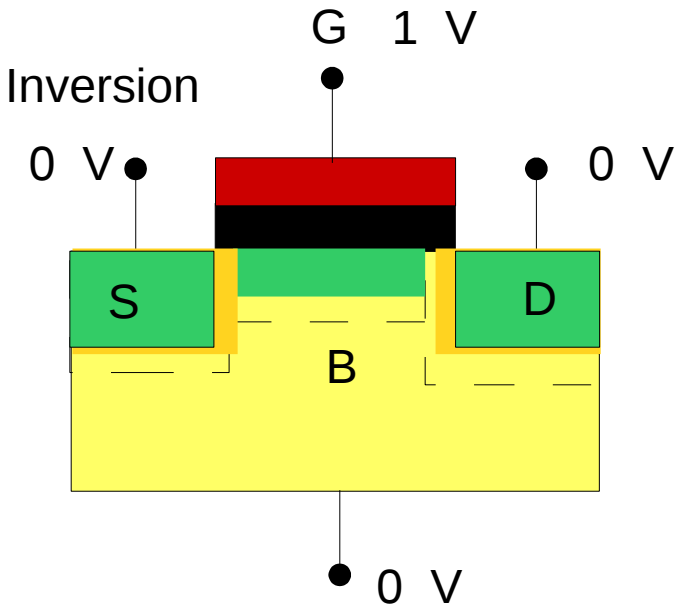


Example of traditional short channel effect: threshold is lower as channel length is decreased.

- The gate voltage needs to maintain also the depletion region.
- For short channel device source and drain depletion regions protrude significantly in the channel.
- For the same charge store on the gate, more carriers can be attracted in the channel, since the depletion region is partially supported by source/drain

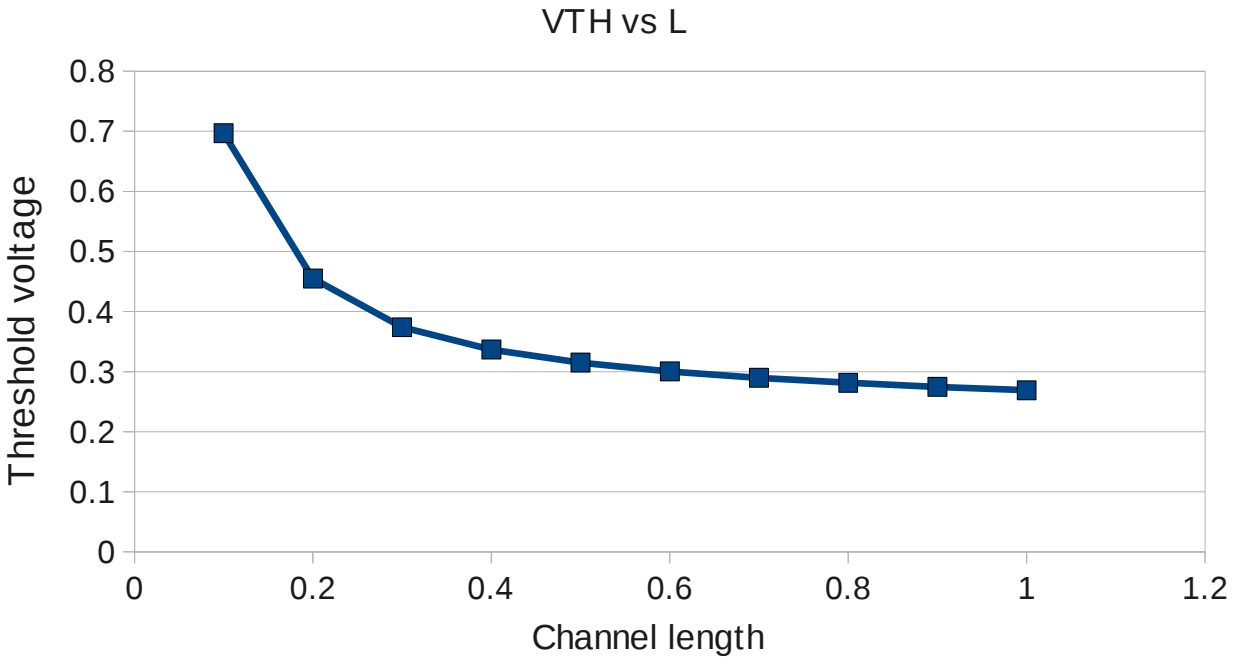


Example: reverse short channel (2)

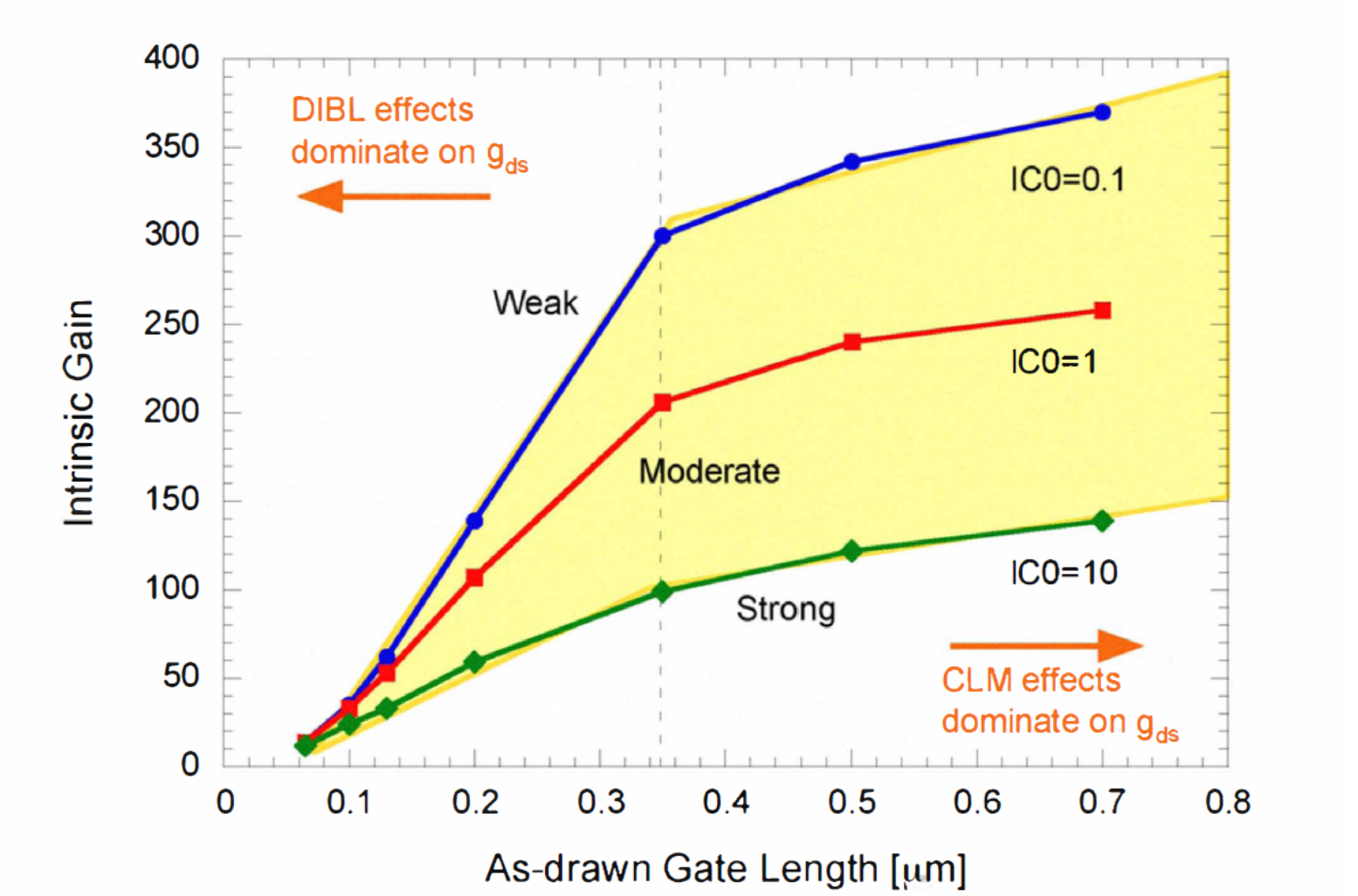


Reverse short channel effect: simulated threshold variation as a function of channel length in a 90 nm CMOS process.

- To prevent excessive extension of the source-drain depletion region into the regions around the electrodes receive a stronger substrate doping (halo doping).
- When the channel is very short the two regions tend to overlap, the local substrate doping in the channel region is increased and the threshold voltage increases.
- Another reason to keep away from minimum length devices in analog design!



Gain vs IC and channel length



M. Manghisoni et al, “Analog Design Criteria for High-granularity Detector Readout in the 65 nm CMOS Technology”, IEEE NSS-MIC Conference Records, N40-2, 2011.

Noise

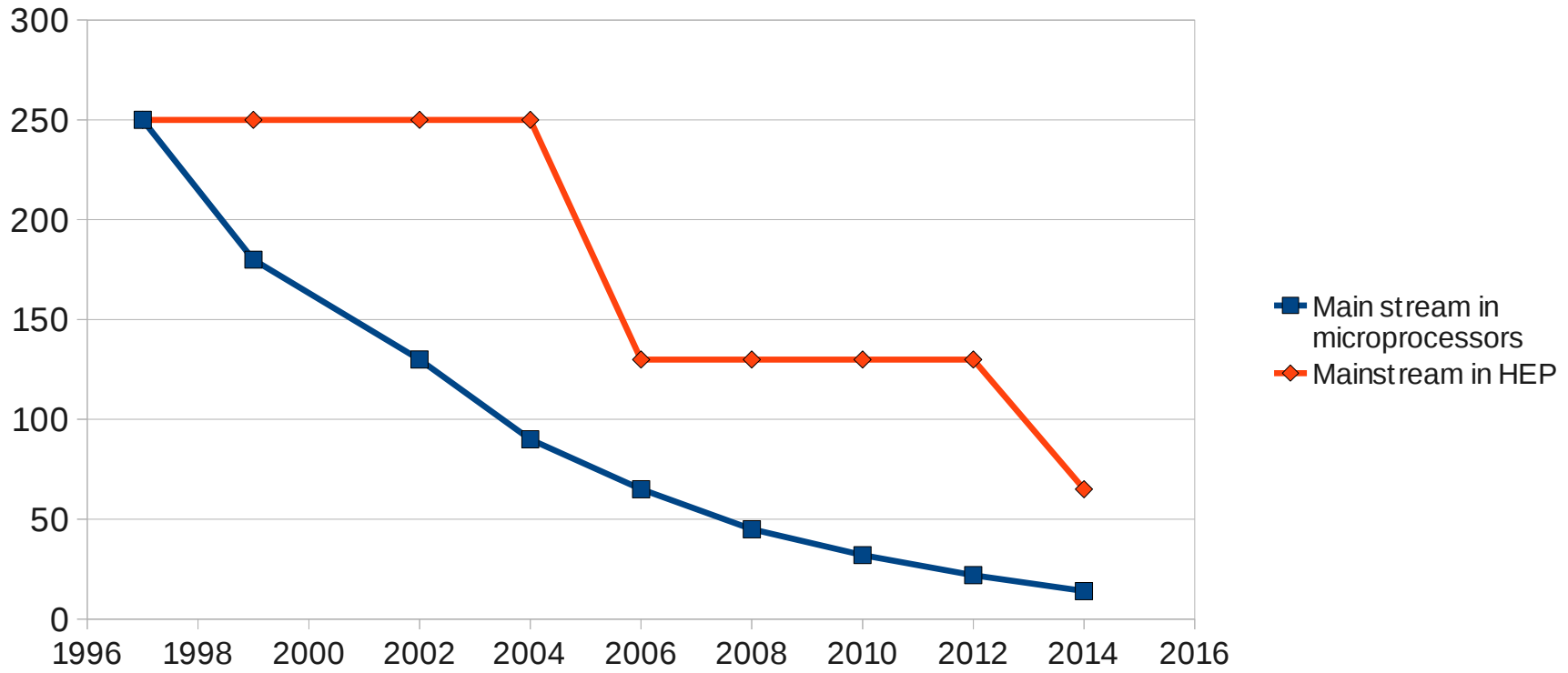
- **Key noise sources in MOS transistors:**
  - **Channel thermal noise**
  - **Channel flicker (1/f) noise**
- **No big surprises concerning thermal noise.**
- **Flicker noise:**

$$e_{n1/f}^2 = \frac{K_f(I_C, L)}{C_{ox} W L f^{\alpha_f}}$$

- **Kf may depend on biasing condition and channel length**
- **Deviation from pure 1/f behavior.**

Noise performance compatible with low noise design

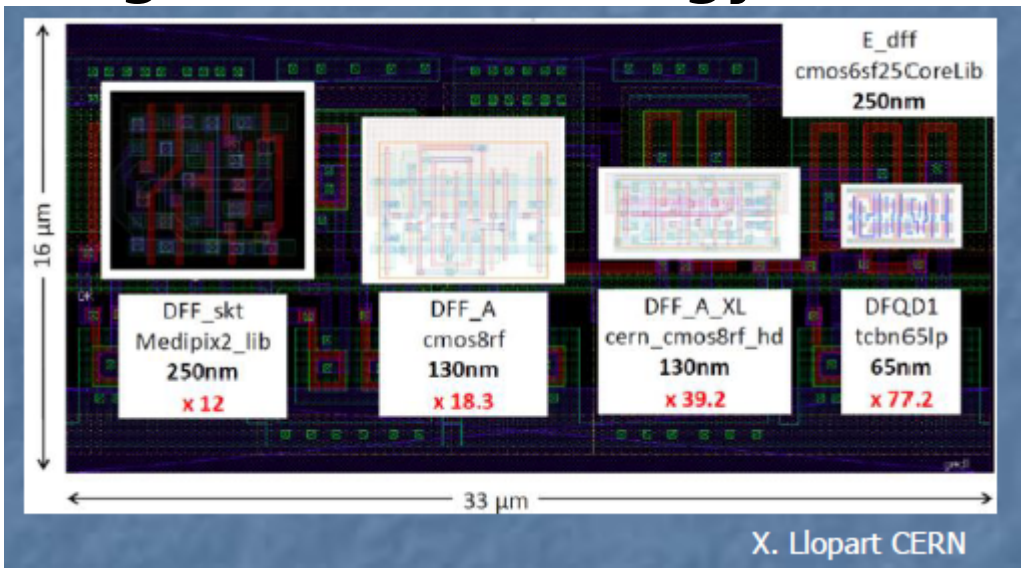
Where we are with technology?



- **Deep sub-micron technologies are however very complex**
- **A lot of reliability rules, especially for obtaining good analog performance and reasonable yield.**

UDSM usage

- **Direct scaling of the technology.....**



- **Reverse scaling of the design rules:**