A Digitally-Controlled Ring Oscillator in 28 nm CMOS technology

Stefano Capra*, Francesco Crescioli†, Luca Frontini*, Maroua Garci† and Valentino Liberali*
*Università degli Studi di Milano and INFN - Sezione di Milano, Milano, Italy †LPNHE, Paris, France

***Abstract*—A digitally-controlled ring oscillator for phase locked loops designed in a commercial 28 nm CMOS technology is presented. Its operating frequency ranges from 2 GHz to 3.2 GHz. A much wider frequency range available in typical case compensates for frequency limitations induced by process variability. The circuit is based on a ring oscillator in which the switching speed of the inverters is controlled by a stream of digital bits. This oscillator is part of a digital PLL, in which the frequency of this oscillator is divided by 8 and used to track an incoming clock signal between 250 MHz and 400 MHz. This circuit is used to obtain eight different clock signals, synchronized with the reference one, to be distributed in massive parallel computation VLSI devices in order to spread the total power consumption of the chip across the reference clock cycle. Post-layout simulations demonstrate that the device is fully compliant in every process corner with the requirements of the future associative memory chip for the track trigger of the ATLAS detector at CERN.**

I. Introduction

Massive parallel computation VSLI devices are modern solutions for the growing real-time pattern-recognition tasks. The field of applicability ranges from scientific calculations to image processing for intelligent autonomous devices to image reconstruction for electro-medical apparatuses.

After the long shut-down from 2019 to 2021, the Large Hadron Collider (LHC) at CERN will reach the unprecedented luminosity of $3 \times 10^{34}\,\mathrm{cm^{-2}\,s^{-1}}$. A level-based trigger scheme enables to reduce the effective amount of data to be processed to an average event rate of $60\,\mathrm{kHz}$ ($100\,\mathrm{kHz}$ maximum). In particular the first hardware level-1 trigger system will be followed by the Fast-TracKer (FTK), responsible for a global track reconstruction after every level-1 trigger and that will provide tracking information to the level-2 trigger system. FTK will process the data coming from the pixel and SemiConductor Tracker (SCT) and from the Insertable B-Layer (IBL) pixel detector.

The FTK system is based on dedicated processing units implemented in FPGA that work in combination with VLSI Associative Memories (AM) [1], that provide highly-parallelized and fast pattern recognition.

An AM chip is a synchronous circuit: the data is loaded in real-time and processed at each clock cycle. Given the massive parallel structure of the device, a lot of power is drawn from the power rails synchronously with the incoming clock. This can lead to unwanted voltage drops at each clock rising edge. Such effects can be clearly noticed during the operation of AM06 [2], which is a large associative memory chip previously realized in $65\,\mathrm{nm}$ technology. AM06 contains 128 k-patterns and its working frequency is $100\,\mathrm{MHz}$.

When the AM06 operation switches between the 'idle' mode and 'compare' mode the current rises from $0.1\,\mathrm{A}$ to $2.2\,\mathrm{A}$ in about $0.1\,\mathrm{ns}$ and current peak are synchronous with the $100\,\mathrm{MHz}$ clock. This current consumption generates a ripple on the supply voltage of the chip that can affect the correct functionality of the whole device if not limited [2].

Several workarounds were implemented at board-level to solve this problem and ensure full functionality of the AM06 chip [3].

The AM chip that will be included in the production version of the future processing units in 2020 is the AM09 chip. This chip, designed in $28\,\mathrm{nm}$ CMOS technology, will contain 3×128 k-patterns. The comparison rate is determined by the input clock, which can vary from $250\,\mathrm{MHz}$ up to $400\,\mathrm{MHz}$.

For the AM09 chip, in which the total cell count is considerably higher respect to the previous prototypes, this power consumption peak must be mitigated also at chip-level.

The proposed solution is to divide the total number of memory cells into eight groups and to spread the activation time of these groups along the reference clock cycle. In order to do this, eight phase-shifted replicas of the original clock are needed. A digital PLL is used to produce a clock signal locked in phase with the incoming one, but with a frequency that is 8 times higher.

The PLL is based on a Digitally-Controlled Oscillator driven by a fairly simple state-machine. This circuit solution was chosen instead of typical Voltage-Controlled Oscillators. In fact VCO schematics are based on operational amplifiers and passive components. Such scaled technologies are mainly intended for digital circuits and generally don't provide modules for passive components that can be used successfully in the design of analog circuits. Moreover, the value of these analog components suffer from great process variability.

For these reasons the solution adopted is a digital ring oscillator with full-digital frequency control, in which the analog part of the circuit is minimized and realized only with active devices.

The main requirements for the DCO are: a frequency from 2 GHz to 3.2 GHz, a power consumption of less than 10 mW (the whole AM09 will consume approximately 4 W), a phase noise not exceeding 1/8 of the clock period. Linearity is not required, while monotonicity is a strong requirement.

The paper is organized as follows. Section II shows the circuit structure, Section III shows the layout design, Section IV describe the simulation results and Section V concludes the paper.

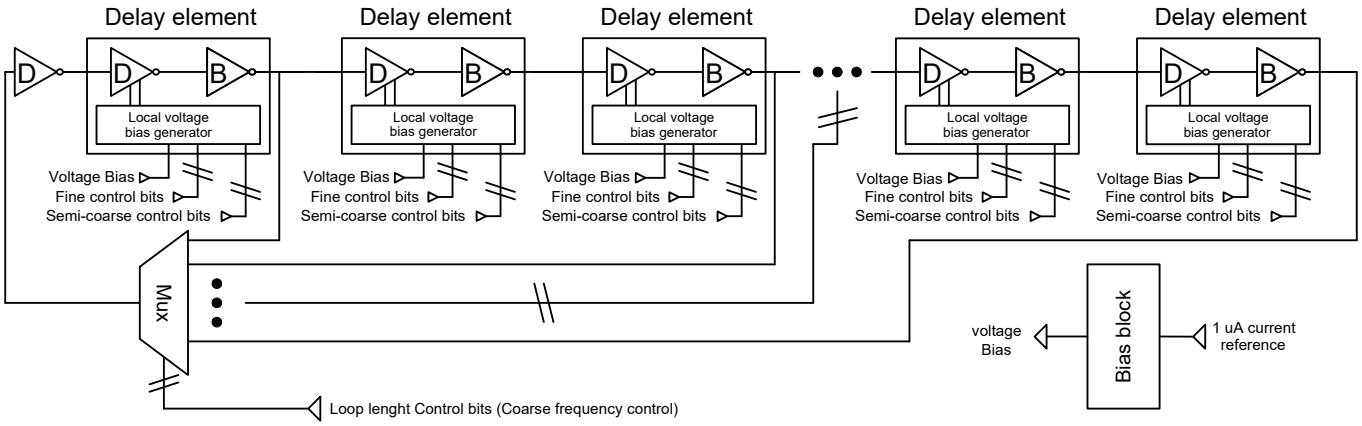

Fig. 1. Block diagram of the ring oscillator. For clarity, only 5 delay elements of 25 are shown. The bias block in the bottom-right corner produces a voltage bias used by the delay elements. The multiplexer has 13 inputs (only 4 are shown).

II. Circuit Structure

The DCO is based on a ring oscillator with a frequency range from 2 GHz to 3.2 GHz in typical case. The oscillation frequency is tuned by acting on three separate thermometric controls. The first one, namely the "coarse frequency control" (12 bits), selects the length of the delay chain in the ring. The fastest configuration is made of just an inverter and one delay element, while the slowest is made of an inverter and 25 delay elements. The two other frequency controls, the "semi-coarse control" (6 bits) and the "fine control" (63 bits) are used to adjust the delay introduced by each element.

The length of the chain should be adjusted only at power-up because is intended mainly to select between different operating frequencies (e.g., 2 GHz or 3 GHz) and to compensate for frequency variations due to the fabrication process. During operation, frequency and phase tracking is performed only with the "fine" and "semi-coarse" controls. The multiplexer can be used also activate-deactivate the oscillator and to provide a start-up stimulus to avoid possible equilibrium states. The bias circuit takes a 1 µA current reference, turns it into a voltage bias and distributes it to the delay elements. Fig. 1 shows a block diagram of the ring oscillator.

Each delay element is made of two inverters and a Local Voltage Bias Generator (LVBG).

The first inverter, marked with "D" in Fig. 1, is the main source of delay, while the second inverter (marked with a "B" letter in Fig. 1) works like an inverting buffer. Fig. 2 illustrates the schematic diagram of the main delay element. The two voltage bias BIAS1 and BIAS2 are propagated to the degeneration transistors. The speed of the inverter made of transistors $\mathrm{MP_{inv}}$ and $\mathrm{MN_{inv}}$ is determined both by the number of active degeneration transistors and by their conductance. The number of active degeneration transistors is the frequency "semi-coarse" control, while the analog bias BIAS1 and BIAS2 depend on the "fine" control. The source of both the N- and P-MOS transistors is degenerated with an array of transistors. These arrays are responsible for a reduction in the current flow from the power rails to the load connected to the

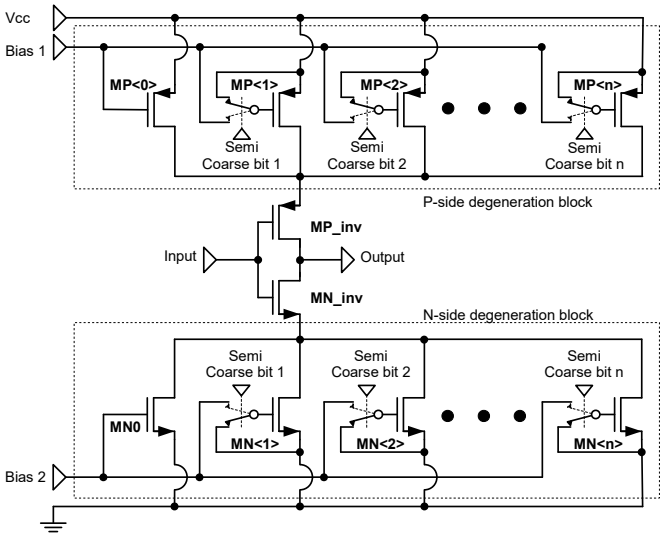

Fig. 2. Schematic of the first inverter inside the delay element.

output of this stage (and vice-versa) during the switching phase of the NOT port. This translates directly into a delay between the incoming signal and the one produced at the output. The second inverter decouples the first from any possible capacitive load connected to the output of the cell and squares up the signal.

Previous works [4], [5] implement similar circuits but with the gate of the degeneration transistors connected either to VDD or VSS. This means that in similar solutions the degeneration transistors contribute to the switching current flow with their full conductance. In the same work was pointed out how the charge stocked inside the parasitic capacitances of the degeneration transistors gives a relevant contribution to the current flow during the switching phase. With this circuit solution the frequency of the oscillator is heavily dependent on the threshold voltage, which, in turn, has a relevant process variation. Our solution moves forward from the one in liter-

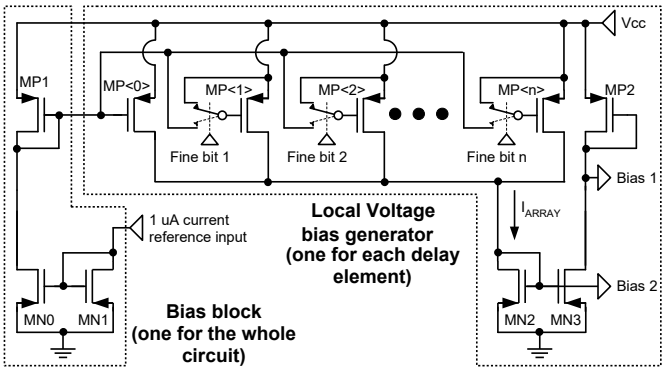

Fig. 3. Schematic of the Local voltage bias generator inside each delay element. For better clarity, the schematic of the bias block (one for the whole circuit) is reported on the left.

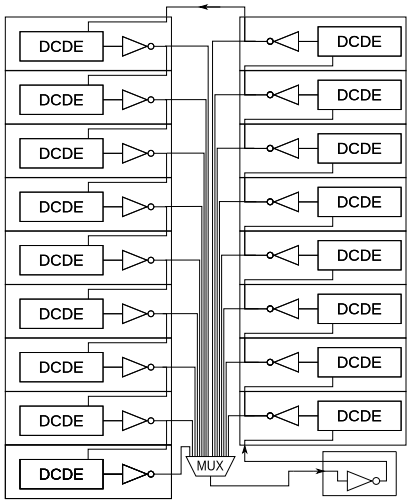

Fig. 4. Block diagram of the delay elements placement in the layout.

ature because it implements a double control of the behavior of the degeneration blocks. In order to limit the frequency dependence on the threshold voltage, the conductance of the degeneration transistors is controlled with a mirror-like circuit and derived from a current reference.

Each degeneration transistor can be either switched "ON" or "OFF"; however, when in the "ON" state, the gate voltage is not fixed to VDD or VSS, but it can span across the available voltage dynamic range. In this way, the conductance of each degeneration transistor can be modulated continuously. The digital control that activates or deactivates each transistor is the "semi-coarse control" while the analog voltage that biases the gate of the degeneration transistors when in "ON" state is determined by the "fine control". The activation of the degeneration transistors is performed following a thermometric code to ensure monotonicity of the control. Since there are two degeneration blocks, one on the N-side and one on the P-side, the "fine control" is translated into two analog voltages, called BIAS1 and BIAS2, that set the conductance of both the N-type degeneration transistors and the P-type ones accordingly. The switches connected to the gate of the degeneration transistors are designed with minimum capacitance, ensuring fast frequency response after switching and minimum voltage bounce on the bias networks.

The voltage biases BIAS1 and BIAS2 are provided by a local voltage bias generator built inside each delay element (see Fig. 3). BIAS1 and BIAS2 are produced by mirroring the variable current I_{ARRAY} on a cascade of two trans-diodes. The I_{ARRAY} current is produced by an array of P-type transistors. These transistors can be switched "ON" and "OFF" with the "fine control". When in the "ON" condition, the gate voltage is fixed to a voltage bias derived from the main $1\,\mu\mathrm{A}$ current reference. The array of digital signals that drives this array of transistors must follow a thermometric code in order to guarantee monotonicity.

In fact, monotonicity is critical to ensure the stability of the digital feedback loop of the PLL in which the DCO is inserted. The size of transistors in the array are not uniform because the frequency response of the DCO to the current I_{ARRAY} is not linear. These transistors have been sized in order to maximize the linearity of the code-frequency response and to achieve the best distribution of available frequency. The choice of replicating the local voltage bias generator block inside each delay element, instead of having just one that generates the references for all the delay elements, is dictated by the requirements in response speed of the DCO to frequency control changes. Since BIAS1 and BIAS2 are analog signals, their bandwidth is limited on one side by the parasitic capacitance connected to the line and on the other side by the transconductance and the bias current of MP2 and MN2.

Simulations demonstrated that, keeping a single central voltage bias generator connected to all the delay elements, the frequency variation of the DCO induced by a code change could not be faster than $30\,\mathrm{ns}$. Embedding a local voltage bias generator inside each delay element, the length of the BIAS1 and BIAS2 paths was considerably reduced together with the associated parasitic capacitance. Moreover, with this circuital solution each local voltage bias generator drives just one array of degeneration transistors, whose gate capacitance is dominant in determining the time constant of the frequency changes. This solution reduces the frequency response time from $30\,\mathrm{ns}$ to $10\,\mathrm{ns}$.

III. Layout Design

The layout of the device was conceived keeping in mind that the propagation delays introduced by metal connections and their associated parasitic capacitances are absolutely critical ans should be minimized, being almost out of control and process-dependent. The block diagram of the delay elements placement is reported in Fig. 4. The delay elements are stacked in two columns, leaving the inverters of the ring in the middle. The signal bus of all the possible ring routings runs in the middle and connects the delay elements to the multiplexer.

The shortest ring length is chosen when the chip belongs to the slow fabrication corner and the speed limitations of

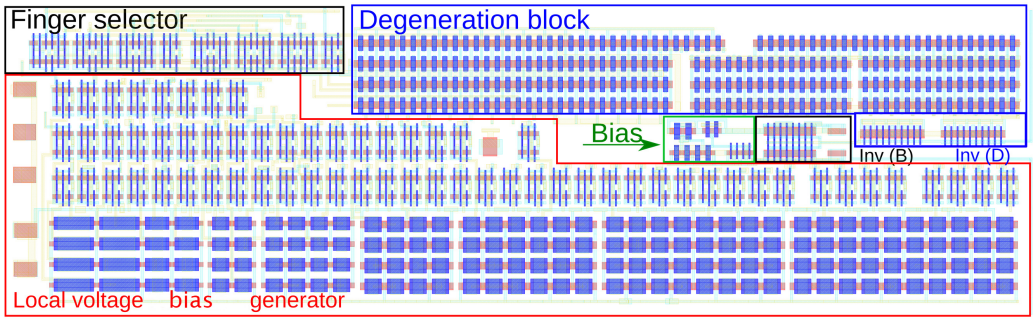

Fig. 5. Layout of a single delay element

TABLE I
SIZE AND RELATIVE AREA OCCUPATION OF THE DELAY ELEMENT'S BUILDING BLOCKS.

Description	Width [µm]	Length [µm]	Area [µm^2]	Percentage [%]
LVBG	21.6	3.7	79.3	55
Deg. block	14.5	2	29	20
Non-active area			24.9	17
MUX	6.8	1	6.8	5
Bias transistors	3.2	1.1	3.5	2
Inverters	1	0.9	0.9	1
Total	21.9	6.6	144.3	100

the DCO are potentially an issue. Thus the total signal path, especially for the shortest ring configuration, should be minimized and the inverters involved must be put very close to the multiplexer. The delay elements are packed in a way to keep the relative distances between the inverters of the ring as low as possible.

Each delay element has an area of $21.9\,\mu\text{m} \times 6.6\,\mu\text{m}$ and the signal path inside it should be minimized (see Fig. 5). For this reason, the critical transistors are packed on one side, leaving the static ones in the body of the block. The major part of the area is occupied by the local voltage bias generator. The dimensions of each building block are reported in Tab. I. The long interconnections between the delay elements and the MUX constitute a long parallel bus that travels across the whole DCO. Such signal lines are realized alternating metal-2 and metal-3 to reduce the parasitic capacitances between them. As common practice, local wirings are realized with low-level metals while the power supply lines are made with higher-level ones. The channel length of the transistors inside the LVBG are deliberately chosen not minimal to reduce the process variability of their electrical parameters.

IV. SIMULATIONS

In order to evaluate the performance of the DCO some post-layout simulations have been performed. These include both transient at circuit start-up and long-term simulations varying the frequency control code. The device demonstrated to start oscillating properly independently from the start-up slope of the power voltage lines. Transient simulations never showed the appearance of higher harmonics or meta-stable states.

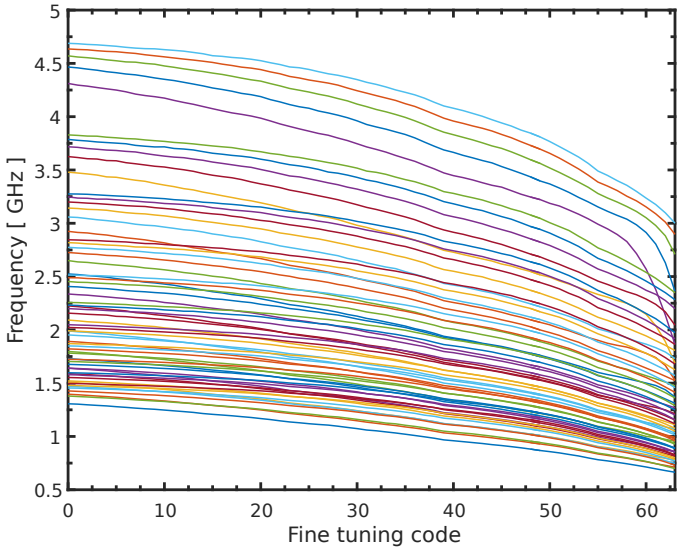

Fig. 6. DCO frequencies versus fine tuning code in typical case.

Since this circuit is strongly dependent on the fabrication speed corners, the DCO behavior has been simulated in every corner in order to ensure its compliance with the AM09 specifications.

Fig. 6 the oscillation frequency as a function of the fine tuning code, for different values of the coarse codes, in typical case. The achieved granularity is fine enough to perform the desired phase-locking with acceptable phase error.

The DC current consumption ranges from $1\,\text{mA}$ (2 GHz) to $5\,\text{mA}$ (3.2 GHz) in typical corner. The frequency variation due to the temperature variation is less than the variation due to corners. The influence of the temperature it's adjusted changing the control bits during time.

V. CONCLUSION

The simulations performed so far demonstrate the functionality of the proposed circuit. The power consumption, the frequency granularity and the speed of frequency change comply with the specifications of the AM09 associative memory chip. More simulations are to be performed in order to inspect the DCO behavior in every interesting condition. The digital feedback control for the PLL is being studied.

REFERENCES

[1] A. Andreani, A. Andreazza, A. Annovi, M. Beretta, V. Bevacqua, G. Blazey, M. Bogdan, E. Bossini, A. Boveia, V. Cavaliere, F. Canelli, F. Cervigni, Y. Cheng, M. Citterio, F. Crescioli, M. Dell'Orso, G. Drake, M. Dunford, P. Giannetti, F. Giorgi, J. Hoff, A. Kapliy, M. Kasten, Y. K. Kim, N. Kimura, A. Lanza, H. L. Li, V. Liberali, T. Liu, D. Magalotti, A. McCarn, C. Melachrinos, C. Meroni, A. Negri, M. Neubauer, J. Olsen, B. Penning, M. Piendibene, J. Proudfoot, M. Riva, C. Roda, F. Sabatini, I. Sacco, M. Shochet, A. Stabile, F. Tang, J. Tang, R. Tripiccione, J. Tuggle, V. Vercesi, M. Villa, R. A. Vitillo, G. Volpi, J. Webster, K. Yorita, and J. Zhang, "The FastTracker Real Time Processor and its impact on muon isolation, tau and b-jet online selections at ATLAS," *IEEE Transactions on Nuclear Science*, vol. 59, no. 2, pp. 348–357, April 2012.

[2] A. Annovi, M. M. Beretta, G. Calderini, F. Crescioli, L. Frontini, V. Liberali, S. Shojaii, and A. Stabile, "AM06: the associative memory chip for the Fast TracKer in the upgraded ATLAS detector," *Journal of Instrumentation*, vol. 12, no. 04, p. C04013, 2017. [Online]. Available: http://stacks.iop.org/1748-0221/12/i=04/a=C04013

[3] L. Frontini, A. Stabile, and V. Liberali, "Power distribution network optimization for associative memories," in *2017 6th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, May 2017, pp. 1–4.

[4] M. Maymandi-Nejad and M. Sachdev, "A digitally programmable delay element: design and analysis," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 5, pp. 871–878, Oct 2003.

[5] ——, "A monotonic digitally controlled delay element," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 11, pp. 2212–2219, Nov 2005.