Simulation of digital pixel readout chip architectures with the RD53 SystemVerilog-UVM verification environment using Monte Carlo physics data

E. Conti, a,1 S. Marconi, a,b,c J. Christiansen, a P. Placidi b,cand T. Hemperek d

a CERN, 1211 Geneva, Switzerland
b Department of Engineering, University of Perugia, Via G. Duranti 93, I-06125 Perugia, Italy
c INFN — Sezione of Perugia, Via Pascoli, I-06123 Perugia, Italy
d Physikalisches Institut, Universität Bonn, Nussallee 12, 531 15 Bonn, Germany

E-mail: elir.conti@cern.ch

ABSTRACT: The simulation and verification framework developed by the RD53 collaboration is a powerful tool for global architecture optimization and design verification of next generation hybrid pixel readout chips. In this paper the framework is used for studying digital pixel chip architectures at behavioral level. This is carried out by simulating a dedicated, highly parameterized pixel chip description, which makes it possible to investigate different grouping strategies between pixels and different latency buffering and arbitration schemes. The pixel hit information used as simulation input can be either generated internally in the framework or imported from external Monte Carlo detector simulation data. The latter have been provided by both the CMS and ATLAS experiments, featuring HL-LHC operating conditions and the specifications related to the Phase 2 upgrade. Pixel regions and double columns were simulated using such Monte Carlo data as inputs: the performance of different latency buffering architectures was compared and the compliance of different link speeds with the expected column data rate was verified.

KEYWORDS: Simulation methods and programs; Pixelated detectors and associated VLSI electronics; Front-end electronics for detector readout

1Corresponding author.
1 Introduction

A flexible simulation and verification platform is being developed within the RD53 Collaboration [1] using the SystemVerilog hardware description and verification language and the Universal Verification Methodology (UVM) library. Such an environment, called VEPIX53 (Verification Environment for RD53 PIXEL chips), is a powerful development tool to be used for the next generation hybrid pixel readout chips [2]. A high-level approach adopted by multiple designers for performing global architecture optimization can address the main design challenges of complex systems like the ATLAS and CMS Phase 2 pixel upgrades in the High Luminosity - Large Hadron Collider (HL-LHC): improved resolution, very high hit rate (up to 3 GHz/cm^2), increased trigger latency time and rate (from 6 to 20 µs and 1 MHz, respectively), extremely hostile environment with radiation levels up to 1 Grad, very high output bandwidth and low power consumption [3, 4]. Furthermore, high-level design, simulation and verification techniques are not new to the High Energy Physics (HEP) community, as shown by their recent use for different applications (e.g. [5, 6]).

A block diagram of the VEPIX53 environment is reported in figure 1. The testbench represents the core of the framework, as it contains the UVM Verification Components (UVCs) and constitutes a reusable and configurable block. The user can then identify a specific test scenario by building a dedicated test in the library, where a particular configuration of the testbench UVCs can be defined. This level of re-usability and flexibility, made possible by the UVM standard classes, makes the chosen methodology highly valuable for the purpose. The connection to the Design Under Test (DUT), wrapped by the top module, is achieved through a set of SystemVerilog interfaces, which was defined to meet the environment requirements: the hit interface (hit_if in figure 1) includes the charge signals generated in the pixel sensor matrix due to particles crossing the detector; the trigger interface (trigger_if) is in charge of the trigger signal; the output data interface (output_data_if)
Different categories of input stimuli can be injected to the DUT through the hit interface. Realistic-looking clusters of hit pixels can be generated internally using a set of pre-defined classes of hits [2]. On the other hand, new functionalities have recently been implemented for importing physics data produced by Monte Carlo simulations of pixel detectors.

In this paper the VEPIX53 environment is used for an explorative study of digital pixel readout chip architectures that expands the results presented in previous works [2, 7], where simulations were run using internally generated hits with the constraints of the Phase 2 operating conditions above described. For this work the architectures have been described at behavioral level with a parameterized pixel chip model and they have been simulated using Monte Carlo physics data related to the CMS and ATLAS pixel detectors. The paper is organized as follows: in section 2 the behavioral parameterized pixel chip model and the architectures under study are presented; section 3 describes the Monte Carlo data used for the simulations; the most relevant simulation results are then reported in section 4, while the discussion and summary can be found in section 5 with future outlooks.

2 Behavioral parameterized pixel chip model

An extensive architecture study of pixel readout chip requires an investigation of each building block that could take into account different operating modes and configurations. At the level of a single Pixel Unit Cell (PUC) different digitization schemes can be evaluated, e.g. Time over Threshold (ToT) versus ADC. PUCs can then be grouped in so-called Pixel Regions (PRs) in order to share digital logic, especially the logic dedicated to trigger latency buffering. Therefore several configurations with different size and shape can be taken into account, as well as latency buffering schemes and derandomizers, which are the memories which store trigger selected data waiting to be transferred. A higher order of grouping can be introduced for handling the communication between the PRs and the pixel chip periphery: this is usually achieved through a single or double column, but also a more generic structure called pixel core can be imagined [8] with a given shape. Different
Figure 2. Block diagrams of latency buffering architectures for pixel regions: (a) zero-suppressed FIFO; (b) distributed latency counters (memory elements are highlighted in yellow) [2].

 arbitration schemes between the PRs of a core and different types of links can be considered. Finally, at the periphery (End of Column/Core, EoC) data compression and merging from different links can be investigated, as well as the readout port.

In order to support some of the various features above described, the DUT simulated with the VEPIX53 framework has been described at behavioral level with a set of parameters related to pixel regions and cores: further details on these groups of pixels are described in the following subsections.

2.1 Pixel region: latency buffering architectures

For pixel regions it is possible to set the size and shape in terms of PUCs. Moreover, two different latency buffering architectures can be chosen (figure 2) and they have been described in detail in [2]: i) a fully shared architecture (called zero-suppressed FIFO) featuring a single shared hit packet buffer; ii) a distributed architecture (called distributed latency counters) featuring a shared hit time buffer containing latency counters and independent ToT buffers in each pixel unit cell. The number of locations of latency and derandomizing buffers are parameterized as well.

The architecture performance for pixel regions at behavioral level is evaluated by monitoring i) hit loss and ii) buffer occupancy through the VEPIX53 analysis UVC. The former, at this stage, is due to two main sources: dead time of the PUC/PR and latency buffer overflow. The latter, on the other hand, is used for building the occupancy distribution, from which it is possible to carry out the corresponding buffer overflow probability. An additional parameter is defined for keeping or neglecting the dead time in the PUCs associated to the conversion of the hit charge into a discriminator output pulse.

2.2 Pixel core arbitration scheme

Similarly to the case of the pixel region, the size and shape of the pixel core are parameterized. A dedicated SystemVerilog interface is introduced for describing in an abstract fashion the link between the pixel regions of the core: this makes it possible to describe different arbitration schemes. Moreover, the link speed can be changed by introducing transfer delays.

The arbitration that has currently been defined between the PRs of the core is a token passing scheme with fast skipping and is represented in figure 3. This scheme is similar to that implemented in the ATLAS FE-I4 pixel chip [9]. A token buffer is defined inside the PR link interface in order to generate tokens associated to triggers and forwarded to each region of the core. Their generation is regulated by a daisy-chained request signal that comes out of each pixel region as the logic OR...
Figure 3. Block diagram of the arbitration scheme implemented for the pixel core of the behavioral parameterized pixel model.

of the request coming from the previous region and its internal one (associated to the presence of hit packets in the derandomizing buffer). This introduces a priority in the arbitration, as the pixel regions at the top of the core output their hit packets first. Furthermore, no clock cycles are wasted if a pixel region has no data to output.

The architecture performance for pixel cores at behavioral level is assessed by monitoring derandomizer occupancy and hit packet latency for each pixel region through the VEPIX53 analysis UVC. This is done in order to verify the compliance with the available bandwidth for the link.

3 Description of Monte Carlo input stimuli

Several sets of Monte Carlo simulation data were provided by both the CMS and ATLAS experiments, featuring different parameters and operating conditions related to HL-LHC and the specifications related to the Phase 2 upgrade.

The CMS data, produced by a workflow based on the CMS data analysis framework (CMSSW), were provided both in ROOT and ASCII text format. Data sets contain events related to layer 0 of the pixel detector with different pixel sizes (50×50 or 25×100 µm²), sensor thickness of 150 µm, pileup of 140 and 1500 e⁻ as digitizer threshold. The ATLAS data, on the other hand, were extracted from Analysis Object Data (xAOD) generated with the ATLAS simulation chain and are related to all the four layers of the detector, with a pixel size of 50×50 µm², sensor thickness of 150 µm and digitizer threshold of 500 e⁻. Pileup could not be simulated for these data sets, so they have been manipulated in order to obtain an increased hit rate by integrating the hit patterns over the modules along the φ direction. For both the CMS and ATLAS data, subsets have been extracted related to modules at the center and edges of the barrel, respectively.

It is possible to extract basic statistical information on the Monte Carlo data sets with the VEPIX53 framework, such as the monitored hit rate on the full matrix and the hit amplitude distribution per pixel, an example of which is shown in figure 4. It is planned to expand this part in order to provide useful data validation checks.
4 Simulation results

The architecture study reported in this work is focused at the level of single pixel region and single pixel core. In order to evaluate the worst case conditions, the presented simulations have run using Monte Carlo data sets related to the innermost layer of the detector at the edges of the barrel, featuring a pixel size of 50×50 µm² and a pileup of 140. For these data the corresponding monitored hit rate is 2.7 GHz/cm².

4.1 Single pixel region simulation

The fully shared and distributed latency buffering architectures were simulated for relevant pixel region configurations 1×1, 2×2 and 4×4 pixels. Simulations were run with 10 µs trigger latency for 484000 bunch crossing clock cycles (~12 ms, average simulation time: 2 hours), in order to collect sufficient statistics on the pixel region performance using the available Monte Carlo data.

The hit loss rate due to dead time for each architecture and configuration is reported in figure 5 (a). These results are compatible with those produced using internally generated hits [2] and show an increasing dead time for the zero-suppressed FIFO architecture as the region gets bigger: this is due to the fact that, in this simple and non-optimized behavioral description, during the dead time of a single pixel all the other pixels of the region are unable to accept later hits. In the distributed latency counters architecture, on the other hand, the hit loss rate is constant with respect to the PR size and it has also been proven that it is comparable with the hit loss rate that is calculated analytically using the average ToT of the pixel hits [9].

The latency buffer occupancy was monitored (examples of histograms are shown in figure 5 (b)) by simulating DUTs where the PUC dead time is neglected, in order to collect statistics more extensively, and the latency buffers are oversized, in order to carry out the buffer overflow probability as a function of the number of locations. From these it is possible to determine the required number of locations that keep such a probability below a certain design value (e.g. 1% or 0.1%). Also in this case, the results agree with those obtained by simulating internally generated hits.

Using the suggested number of locations related to an overflow probability below 0.1%, further double check simulations have been run with fixed size buffers: as reported in table 1, the monitored hit loss due to buffer overflow is in most cases below ~0.1%.
Table 1. Hit loss rate due to buffer overflow.

| Pixel region (z × ϕ) | Buffer locations | Hit loss rate |
|----------------------|------------------|--------------|
|                      |                  | Zero-suppressed FIFO | Distributed latency counters |
| 2×2                  | 8                | 0.030%        | 0.129%         |
| 4×4                  | 12               | 0.002%        | 0.032%         |

4.2 Single core simulation

For the single core simulation a double column was chosen with the arbitration scheme described in section 2, made of 2×64 pixel regions featuring a configuration of 2×2 pixels and a distributed latency buffering architecture. The corresponding pixel region hit packet format is composed of a 7-bit address of the region in the double column, plus a 4-bit ToT per pixel: this results in an approximately 3-byte wide packet. Simulations were run for 660000 bunch crossing clock cycles (∼16.5 ms, average simulation time: 2.5 hours) for different trigger rates, with the constraint of random generation of independent trigger pulses, and different link speeds: a full width parallel bus, which is able to transfer the 3-byte packet in a single clock cycle, and an 8-bit bus which requires 3 clock cycles. VEPIX53 simulation time was of the order of 2 hours for low trigger rates.

It is possible to verify the priority introduced in the double column by the token passing scheme by comparing the latency histograms for the different regions of the core. Examples are reported in figure 6 for the full width parallel bus at 1 MHz trigger rate and for the 8-bit bus at 10 MHz trigger rate. It can be noticed how the average latency is lower for the hit packets produced by the pixel regions at the top of the double column.

The compliance with the available bandwidth for the link speeds taken into account was verified as well. This was initially done with the comparison of each link rate with the expected data rate coming out from the double column, calculated analytically (it is given by the pixel region rate multiplied by trigger rate and hit packet width); then it was validated with VEPIX53 simulations by evaluating the average occupancy of the pixel region derandomizing buffers.

First, a trigger was randomly generated within the testbench with a constrained trigger rate of 1 MHz and simulated; the monitored value was 0.72 MHz due to randomization of the trigger.
pulses. This corresponds to an expected core data rate of 7.46 Mbits/s, which is 0.77% of a full width parallel bus (which has an associated link rate of 960 Mbits/s) and 2.33% of a 8-bit bus (associated link rate: 320 Mbits/s). The VEPIX53 simulations then confirmed that both the links can well support such a data rate, as the overflow probability of the derandomizing buffer related to a single memory location, reported in table 2, is significantly below 1% for the nominal trigger rate of 1 MHz. Further simulations were run with higher trigger rate in order to assess whether or not the links can operate in worse conditions. The expected core data rate associated to a 10 MHz trigger rate (actual monitored rate: 9.073 MHz) is 94.07 Mbits/s and corresponds to the 9.80% of the full width parallel bus and the 29.39% of the 8-bit bus; as shown in table 2, the derandomizing buffer overflow probability carried out from simulation results was still less than 1% for the former link and slightly higher that 1% for the latter. Finally, an extreme case was taken into account of a simulation with 40 MHz trigger rate (actual monitored rate: 36.362 MHz; expected core data rate: 414.6 Mbits/s), which resembles a close to non-triggered operation of the pixel chip. The full width parallel bus is the only link of the two that can support such a high data rate with an overflow probability of the derandomizing buffer around 1% for a single memory location.

5 Conclusions

A simulation framework using physics Monte Carlo data is crucial for optimization in view of the CMS and ATLAS Phase 2 challenges of pixel chip design. The latest additions to the VEPIX53 environment have shown that simulations with Monte Carlo data are compatible with previous results found using internally generated hits or analytically. Double column simulations have highlighted
that the derandomizing buffers can be small at 1 MHz trigger rate, so the derandomization stage can conveniently take place on the same memory as trigger latency buffer, as happens in already existing pixel chips such as the ATLAS FE-I4. Simulations also indicated that a full width parallel bus, for the double column under investigation with a fast skipping arbitration scheme, can support both triggered and non-triggered operation, which can be related to test modes of the pixel chip even though they feature a considerably smaller hit rate.

Further additions and investigations will have to be done for proceeding with the extensive architecture study. It is very important to introduce data merging and compression schemes, based on clustering, between several pixel cores as the bottleneck for data rate is introduced by the readout. Other architectures could be considered as well for attempting at maximizing the data rate. A comprehensive validation of the injected Monte Carlo data will be implemented, also in the perspective of simulating combinations of externally provided hit patterns with internally generated extreme events. Finally the same framework will be used for extensive design verification at gate level including radiation damage effects.

Acknowledgments

The authors would like to thank E. Migliore (INFN Turin, Italy) for providing CMS data and R. Carney (LBNL, California) for providing ATLAS data.

References

[1] J. Christiansen and M. Garcia-Sciveres, RD Collaboration proposal: development of pixel readout integrated circuits for extreme rate and radiation, LHCC-P-006 (2013).

[2] S. Marconi et al., The RD53 collaboration’s SystemVerilog-UVM simulation framework and its general applicability to design of advanced pixel readout chips, 2014 JINST 9 P10005.

[3] ATLAS collaboration, ATLAS Letter of intent phase-II upgrade, LHCC-I-023 (2012).

[4] CMS collaboration, Technical proposal for the upgrade of the CMS detector through 2020, CMS-UG-TP-1 (2011) [LHCC-P-004].

[5] T. Poikela et al., VeloPix: the pixel ASIC for the LHCb upgrade, 2015 JINST 10 C01057.

[6] A. Fiergolski, M. Quinto, F. Cafagna and E. Radicioni, Upgrade of the TOTEM DAQ using the Scalable Readout System (SRS), in proceedings of IEEE Nuclear Science Symposium and Medical Imaging Conference ( NSS/MIC), October 27–November 2, Seoul, Korea (2013).

[7] S. Marconi, E. Conti, J. Christiansen and P. Placidi, Reusable SystemVerilog-UVM design framework with constrained stimuli modeling for High Energy Physics applications, in proceedings of IEEE International Symposium on Systems Engineering (ISSE), September 28–30, Rome, Italy (2015).

[8] M. Garcia-Sciveres, A. Mekkaoui and D. Gnani, Towards third generation pixel readout chips, Nucl. Instrum. Meth. A 731 (2013) 83.

[9] D. Arutinov et al., Digital architecture and interface of the new ATLAS pixel front-end IC for upgraded LHC luminosity, IEEE Trans. Nucl. Sci. 56 (2009) 388.