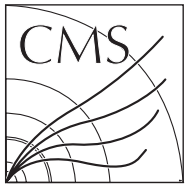
Available on CMS information server

CMS CR -2019/009



The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

31 January 2019 (v5, 24 June 2019)

Serial powering and high hit rate efficiency measurement for the Phase 2 Upgrade of the CMS Pixel Detector.

D. Ruini for the CMS Tracker Group

Abstract

A serially powered pixel detector is the baseline choice for the High Luminosity upgrade of the inner tracker of the CMS experiment. A serial power distribution scheme, compared to parallel powering, requires less cable mass, offers higher power efficiency and is less susceptible to voltage transients. A prototype pixel readout chip has been designed for serial powering in 65 nm CMOS technology by the RD53 collaboration. Performance results from testing the prototype chip, called RD53A, are reported. The performance of RD53A operating in a chain consisting of four chips powered in series is compared with the performance under a conventional powering scheme. Additionally, the readout efficiency of RD53A in a high hit rate environment in different operation modes is presented for the first time. The results indicate that serial powering is a robust and reliable power distribution scheme.

Presented at *PIXEL2018 International Workshop on Semiconductor Pixel Detectors for Particles and Imaging 2018*

Prepared for submission to JINST

9TH International workshop on Semiconductor Pixel Detectors for Particles and Imaging
December 10–14, 2018
Academia Sinica, Taipei

Serial powering and high hit rate efficiency measurement for the Phase 2 Upgrade of the CMS Pixel Detector.

D. Ruini on behalf of the CMS Tracker Group

Institute for Particle Physics and Astrophysics, ETH Zürich

E-mail: daniele.ruini@cern.ch

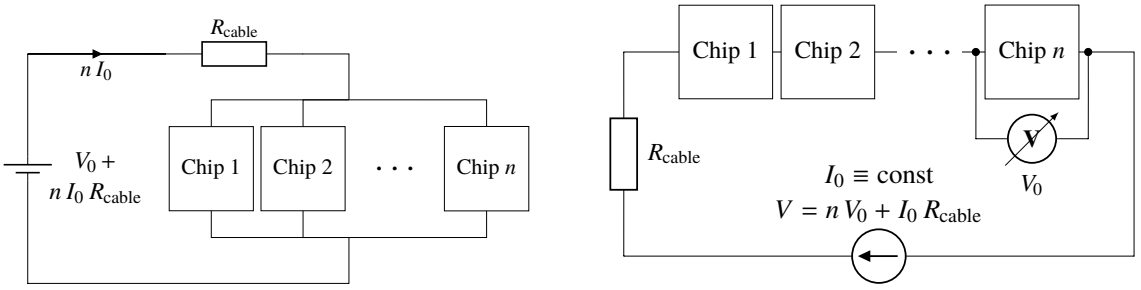
Abstract: A serially powered pixel detector is the baseline choice for the High Luminosity upgrade of the inner tracker of the CMS experiment. A serial power distribution scheme, compared to parallel powering, requires less cable mass, offers higher power efficiency and is less susceptible to voltage transients. A prototype pixel readout chip has been designed for serial powering in 65 nm CMOS technology by the RD53 collaboration. Performance results from testing the prototype chip, called RD53A, are reported. The performance of RD53A operating in a chain consisting of four chips powered in series is compared with the performance under a conventional powering scheme. Additionally, the readout efficiency of RD53A in a high hit rate environment in different operation modes is presented for the first time. The results indicate that serial powering is a robust and reliable power distribution scheme.

Contents

1 The High Luminosity Large Hadron Collider and the need for serial powering

The Large Hadron Collider (LHC) [1], to date the world's largest and most powerful particle accelerator, will undergo a profound upgrade during a long shutdown starting in 2024. At the restart of operations in 2026 the instantaneous luminosity[1] will increase by a factor of about seven with respect to the nominal design value, reaching up to $7.5 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The goal of the upgraded accelerator, called High Luminosity LHC (HL-LHC) [2], is to increase the recorded integrated luminosity by a factor ten over the course of ten to twelve years and reach $4000\,\mathrm{fb}^{-1}$ by the late 2030s. While greatly increasing the physics potential of the LHC, the increase in luminosity poses severe challenges to the experiments in terms of data rate and radiation damage, requiring a replacement of large parts of the detectors in order to guarantee high quality data taking throughout the high luminosity era. The CMS experiment foresees, among other upgrades [3], to install a completely new pixel detector [4]. The power needs of the upgraded detector are driven in particular by the high granularity (about one billion individual pixels of $2500\,\mu\mathrm{m}^2$) and the trigger latency of $12.5\,\mu\mathrm{s}$, which requires large amounts of on-chip memory buffers. The RD53 collaboration [5, 6] has developed a prototype pixel chip addressing these challenges in 65 nm CMOS technology [7]. The prototype, called RD53A, operates at low voltage (1.2 V) and high current (750 mA). The final chip, which will be about double in size, will consume up to 2 A. As the pixel detector will consist of 13488 chips, a parallel voltage based power distribution scheme would result in an unacceptably high cable mass and/or power loss (Figure 1a). A serial power distribution scheme [8–10] works instead with a constant current powering several loads, as shown in Figure 1b. This makes it inherently more power efficient and immune to over-voltages as there can be no sudden current drop on the power lines. These properties mean that the cable mass can be reduced significantly and that a lightweight system can be built. Serial powering is therefore the baseline choice for the new CMS pixel detector, at the cost of increased complexity: a dedicated

[1] The instantaneous luminosity is the proton-proton collision rate per unit area, usually expressed in $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$.



(a) Parallel powering scheme.

(b) Serial powering scheme.

Figure 1: In a system with n independent loads, each requiring voltage V_0 and current I_0, the power lost in the cables (RI^2) is proportional to n^2 in a parallel powering scheme (a) and is constant in a serially powered system (b).

on-chip regulator is required to generate a constant voltage and shunt the excess current, and the failure of one element of the serially powered chain might compromise the entire system. These aspects are treated in the next two sections.

2 Serial powering and the shunt-LDO regulator

The dynamic current variations of the CMOS circuitry are too fast to be regulated by an external power supply, especially as a serially powered loop has a large inductance (the power cables of the pixel detector have a length in the order of 100 m). It is therefore necessary that a constant current, enough to satisfy the power needs of the chips at any moment, circulates in the system. This includes the maximal possible load current plus a headroom to ensure that the fast dynamical current variations are not visible at the input of the system. The optimal amount of headroom current, enough to guarantee reliable operation while optimizing power efficiency, is being investigated. For the studies presented here it is set conservatively to $\sim$25% of the maximal chip consumption. The current needs to be converted to a stable local voltage for the chip and the excess current must be shunted. These tasks are accomplished by a specialized regulator integrated on chip, which combines the functionalities of a current shunt and of a low-dropout (LDO) regulator, and is typically referred to as shunt-LDO or SLDO [11]. The circuit diagram is shown in Figure 2. The voltage regulation is performed by the LDO in the left part of the diagram, marked in blue: amplifier $A1$ compares a stable reference voltage V_{ref} to the voltage between $R1$ and $R2$, regulating the voltage drop over transistor $M1$ such that they are equal. This ensures that the load supply voltage V_{DD} is constant, independent of the input voltage V_{in}. An additional path for the excess current (i.e. not needed by the load) is through the shunt transistor $M4$, regulated by the control loop in the right part of the diagram, marked in brown: a constant reference current, generated directly from the input by resistance $R3$ ($A4$ and $M7$ provide an adjustable voltage offset for power optimization), is compared to the current flowing through the LDO input transistor $M1$. $M4$ is regulated such that the total current consumption is constant: $I_{M1} + I_{M4} \equiv \text{const}$. Overall the SLDO behaves like a resistance with a voltage offset: this allows to operate multiple SLDOs in parallel with a well defined current sharing determined by the configurable effective resistance $R3$.

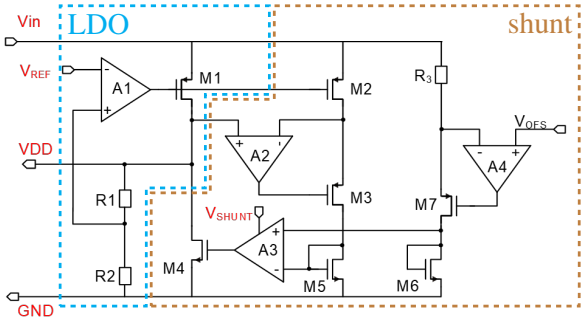


Figure 2: Simplified circuit diagram of the SLDO integrated on the RD53A chip. The LDO provides a constant voltage VDD to the load, while the shunt ensures a constant total voltage consumption. Adapted from [7].

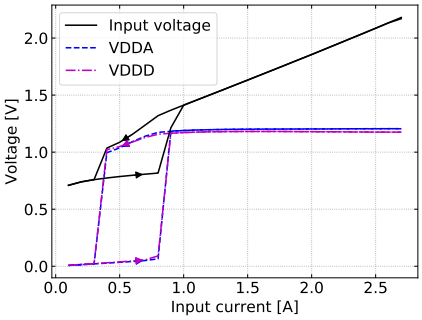


Figure 3: Input and output voltage as function of the input current of the two parallel SLDOs on a RD53A chip. The hysteresis between switch-on and switch-off procedure is shown.

SLDOs on the RD53A chip. The RD53A chip integrates two SLDOs, one per power domain (analog and digital), connected in parallel. Each SLDO can carry up to 2.0 A, allowing for a significant current margin during operation of the chip. This is important in case of failure of a chip, as described in section 3. The typical *VI* curve (both input and output voltages) of the SLDOs on a RD53A chip is shown in Figure 3 for stepwise increase and decrease of the common input current. In the plot, VDDA(D) is the voltage powering the analog (digital) part of the chip. Note that the minimal current needed for proper regulation is higher during ramp up than during ramp down. This hysteresis is caused by the biasing scheme of the SLDOs generating the reference and offset voltages based on bandgap references. Due to significant production variations, some chips require high currents close to 2 A to start up. This is undesirable as it requires injecting a high current at the beginning to start regulation and decreasing it afterwards. A new version of the SLDO circuit with an improved biasing scheme, unaffected by this issue, is being prototyped and is expected to be integrated in the final pixel chip.

3 Serial powering in CMS

CMS plans to build serially powered chains of up to 11 modules as sketched in Figure 4. Each module will consist of two or four chips, depending on the location of the chain within the detector. On each module the chips are powered in parallel: should a chip fail, resulting in an open circuit as shown in Figure 5, the current would still flow through the remaining chips without compromising the functionality of the rest of the serially powered chain. This is possible as the chips are able to withstand an increased current by design, provided sufficient cooling. For a four-chip module, the failure of one chip would increase the current through the other chips by 33% and, due to the resistive behaviour of the SLDOs shown in Figure 3, the input voltage would also increase. The combined increase of voltage and current would result in an increase in power consumption of 45%, which can be accommodated by the cooling system. If instead one chip would fail on a two chip module, the power consumption would increase by 150%, which would be difficult to manage. For this case, the option of connecting two modules (i.e. four chips) in parallel is being investigated.

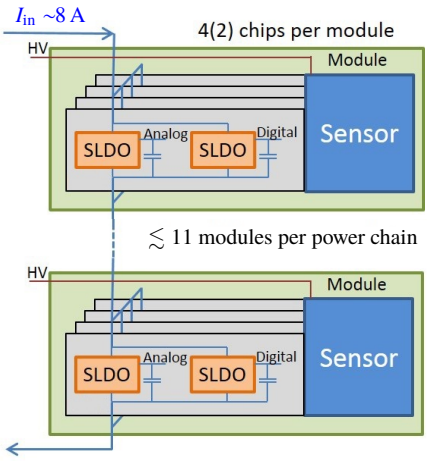


Figure 4: Serial powering implementation in CMS. HV is the high voltage bias for the sensor. Adapted from [4].

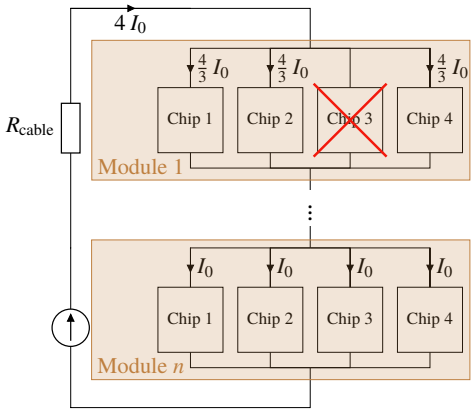


Figure 5: Possible failure scenario in a chain of n four-chip modules: if one chip breaks, the current is carried by the remaining chips.

4 Operation of a serially powered chain of RD53A chips

Four RD53A chips have been tested, both in single chip mode and in a serially powered chain, to validate the serial powering concept. As an example of electrical tests, the voltage drop across each chip in the chain, i.e. the local input voltage seen by the SLDOs on that chip, in shown in Figure 6a as function of the input current. Also shown is the total input voltage of the chain, divided by 4 for a better comparison. As can be seen, above 1 A, which is the typical current needed for operation, all chips have by default very similar input voltages within approximately 100 mV. This difference could, if needed, be reduced by optimizing the offset voltages on each chip.

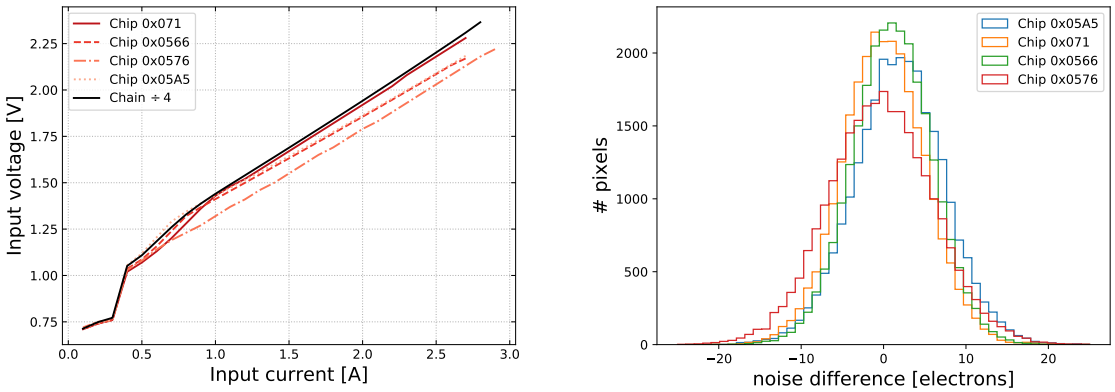
Communication with the chips is tested by injecting calibration charges into each pixel[2] and counting how many can be read out by the chip. For all chips, both when operated alone and in the chain, 100% of the calibration signals can be read out.

Another aspect relevant for the operation of the chips is the noise. It is first measured separately for single chip and chain operation, then the difference between the two is computed for each pixel. The resulting distributions are shown in Figure 6b: for all chips the noise difference is centered at 0, indicating that the noise is unaffected by the different powering modes. The typical noise level of a RD53A chip without sensor is around 70 electrons [12].

5 Readout efficiency in high hit rate environment

This section describes the measurement of the readout efficiency at high hit rates of the linear frontend implemented in RD53A. The maximal expected hit rate in the pixel detector during the high luminosity era is 3 GHz cm^{-2} in the region closest to the interaction point. The design goal is to have a detector able to correctly read out 99% of the hits at this rate [4]. Data loss can happen

[2]This and the following tests are performed only on the linear frontend. The RD53A contains two more frontend technologies, differential and synchronous, for testing purposes.



(a) Voltage drop (i.e. local input voltage) across each of the four RD53A chips powered in series and the total input voltage of the chain, divided by 4.

(b) Pixel-by-pixel noise difference between single chip and chain, for all four chips. Operating the chips in chain has no influence on the noise.

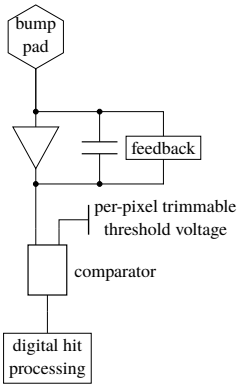
Figure 6: Input voltage of each chip in the serially powered chain (a) and noise difference between chips operated alone and in the chain (b).

essentially at two stages during the hit processing: within the analog part of the readout chain and within the memory buffer architecture. The analog readout chain is sketched in Figure 7a: the signal, reaching the frontend through the bump pad, is fed to a charge sensitive amplifier with Krummenacher feedback. The signal is then compared to a threshold (which is tunable for each pixel) and digitised by counting the time over threshold (ToT) with a 40 MHz clock. This mechanism becomes inefficient when a pixel receives two very close hits, as sketched in Figure 7b: the frontend is then not able to resolve the two hits and instead records only one with a long ToT. Keeping these inefficiencies under control is the main reason for having small pixels of $50 \times 50\,\mu m^2$ on the readout chip, as the analog inefficiencies scale linearly with the pixel area.

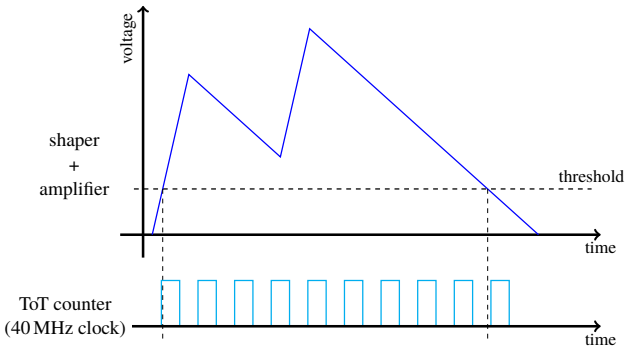
The memory buffers are organized according to the so called distributed buffer architecture[3] (DBA) shown in Figure 8: the pixel matrix is divided into *regions* of four pixels each, spanning four columns and one row. The ToT information is stored locally in each pixel (there are eight ToT registers per pixel), while there is a shared timestamp buffer for the region which can store up to eight timers. A timer starts every time a pixel (or more than one, if in the same bunch crossing) in the region is hit. Sharing the memory between neighbouring pixels is motivated by the fact that charged particles originating from collisions typically cause hits in several neighbouring pixels: these hits occur within the same bunch crossing and sharing the timestamp reduces the memory costs. The average size of the hit cluster varies approximately from 4 to 15 pixels, depending on the position within the detector. Inefficiencies in this architecture occur when the hit rate is too high and the buffers overflow, or if the hits are not clustered, thus not profiting from the shared memory.

Measurement of the readout efficiency. The readout efficiency is measured with the X-ray setup in Figure 9: a RD53A chip equipped with a sensor manufactured by HLL [13] with pixels of $25 \times 100\,\mu m^2$ is placed in the direct X-ray beam. This setup represents the worst case scenario for

[3]This readout architecture is shared by the linear and differential frontends. The synchronous frontend implements a different architecture, see [7] for details.



(a) Analog hit processing of the linear frontend.



(b) Source of analog inefficiency: when a second hit arrives while the output of the comparator is still high only one hit with large ToT is digitised.

Figure 7: Analog readout chain (a) and sources of inefficiency (b).

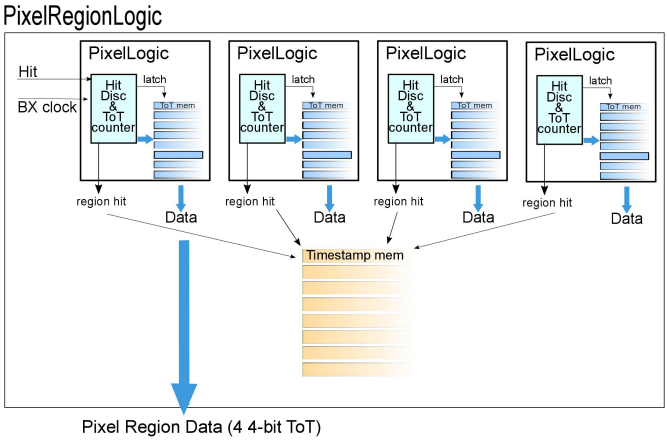


Figure 8: Region of four pixels with eight ToT memories each and eight shared timestamp memories. This buffer architecture is optimized for clusters of several neighbouring pixel hits. Adapted from [7].

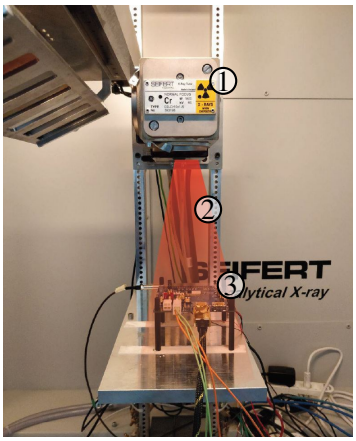
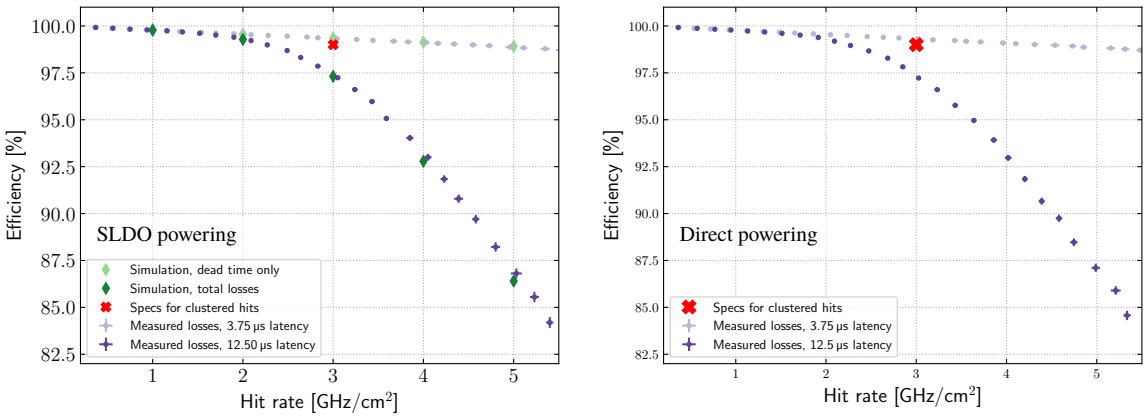


Figure 9: X-ray setup for the high hit rate efficiency measurement. The source (1) produces a beam (2) which directly illuminates the chip (3).

the chip: hits from photons do not produce clusters but randomly scattered single hits, not profiting from the shared memory architecture. The measurement procedure is as follows: first the threshold of all pixels is tuned to a uniform value. Then the chip is placed in the X-ray beam and a fixed number of test signals $n_t = 4 \times 10^5$ is injected into 27 arbitrarily chosen pixels distributed across the matrix[4]. Finally the whole matrix is read out. The injection and readout is done at a low frequency of 100 kHz in order to avoid limitations in the readout system. Furthermore only data from the exact bunch crossing where the test signal was injected is read out: this guarantees that the hits recorded in the selected pixels are really from injections and not from photons. The readout efficiency and the X-ray hit rate are then computed for each of the selected pixels and the corresponding region as

[4]Signals were injected starting from the pixel in row 7 and column 130 and then in every 7th row and 5th column.



(a) (b)

Figure 10: Readout efficiency as a function of hit rate measured with the chip powered with constant current through the SLDOs and simulation (a) and the same measurement with the chip powered directly with constant voltage bypassing the SLDOs (b).

follows:

$$\text{Single pixel efficiency} = \frac{\text{Number of hits recorded by the pixel out of } n_t \text{ signals}}{n_t},$$

$$\text{Region hit rate} = \frac{\text{Average occupancy of pixels in the same region}}{\text{pixel size} \cdot \Delta t \cdot \text{single pixel efficiency}},$$

with $\Delta t = n_t \cdot 25$ ns. Efficiencies and hit rate are then averaged over the pixels into which the signals were injected.

The efficiency is measured for different trigger latencies: the nominal 12.5 µs (500 bunch crossings) and 3.75 µs (150 bunch crossings). The results obtained with different powering modes of the chip (with constant current through the SLDOs and with constant voltage bypassing the SLDOs) are shown, together with simulation, in Figure 10. The simulation was performed for a four-pixel region, including both analog and buffer loss mechanisms, with Poisson-distributed single pixel hits. The trigger latency is set to 12.5 µs and the ToT has the same average value as the X-ray hits. The measured efficiency shows no significant difference between powering modes. For the short trigger latency, shown by the light circles, the efficiency shows a slow, linear decrease, due to analog inefficiencies only: the memory buffers are emptied frequently and they never fill up. This is confirmed by the simulation, as the measurement agrees well with the simulated analog losses shown by the light green diamonds. At the nominal trigger latency, shown by the dark circles, the losses due to buffer overflow are visible as a fast drop in efficiency at rates above $\sim 2\,\text{GHz}\,\text{cm}^{-2}$, in agreement with simulation. At the nominal hit rate of $3\,\text{GHz}\,\text{cm}^{-2}$ this results in an efficiency of 97.5%, which is below the target value for clustered hits of 99%. Results from simulations of clustered hits are instead in closer agreement with the design goal. An option under investigation for further improvement in the final chip is to count the ToT at 80 MHz, i.e. with both edges of the 40 MHz clock. This would allow a faster return to the amplifier baseline, reducing the analog inefficiencies without decreasing the charge resolution.

6 Conclusion

A serial power distribution scheme based on constant supply current is the baseline choice for the high luminosity upgrade of the CMS pixel detector. The prototype pixel readout chip RD53A integrates a specialized SLDO regulating circuit, which provides a constant supply voltage to the electronics and shunts the excess current. Preliminary system tests show that the chips work reliably when powered in series, with no influence on the readout noise. Serially powered chains in the pixel detector will consist of up to 11 modules, with up to four chips powered in parallel per module. Robustness against single chip failure is provided by the on-module parallel powering: the resulting current increase through the remaining chips is shunted by the SLDOs. The measured readout efficiency is in good agreement with the simulated data loss for unclustered hits and simulations of clustered hits indicate that the RD53A chip can fulfil the design goal.

References

[1] O. Brüning et al. *LHC Design Report*. CERN Yellow Reports: Monographs. CERN, Geneva, 2004.

[2] G. Apollinari et al. *High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1*. CERN Yellow Reports: Monographs. CERN, Geneva, 2017.

[3] CMS Collaboration. Technical Proposal for the Phase-II Upgrade of the CMS Detector. Technical Report CERN-LHCC-2015-010. LHCC-P-008. CMS-TDR-15-02, CERN, Geneva, Jun 2015.

[4] CMS Collaboration. The Phase-2 Upgrade of the CMS Tracker. Technical Report CERN-LHCC-2017-009. CMS-TDR-014, CERN, Geneva, Jun 2017.

[5] J. Chistiansen and M. Garcia-Sciveres. RD Collaboration Proposal: Development of pixel readout integrated circuits for extreme rate and radiation. Technical Report CERN-LHCC-2013-008. LHCC-P-006, CERN, Geneva, Jun 2013.

[6] F. Arteche Gonzalez et al. Extension of RD53. Technical Report CERN-LHCC-2018-028. LHCC-SR-008, CERN, Geneva, Sep 2018.

[7] M. Garcia-Sciveres. The RD53A Integrated Circuit. Technical Report CERN-RD53-PUB-17-001, CERN, Geneva, Oct 2017.

[8] T. Stockmanns et al. Serial powering of pixel modules. *Nucl. Instrum. Meth.*, A511:174–179, 2003.

[9] D.B. Ta et al. Serial powering: Proof of principle demonstration of a scheme for the operation of a large pixel detector at the LHC. *Nucl. Instrum. Meth.*, A557:445–459, 2006.

[10] L. Gonella et al. Performance evaluation of a serially powered pixel detector prototype for the HL-LHC. *Journal of Instrumentation*, 12(03):P03004–P03004, mar 2017.

[11] M. Karagounis et al. An integrated Shunt-LDO regulator for serial powered systems. In *Proceedings of ESSCIRC*, pages 276 – 279, Oct 2009.

[12] L. Gaioni. Test results and prospects for RD53A, a large scale 65 nm CMOS chip for pixel readout at the HL-LHC. *Nuclear Inst. and Methods in Physics Research*, 2018.

[13] A. Macchiolo et al. Characterization of RD53A compatible n-in-p planar sensors. In *Proceedings of PIXEL2018 (this issue)*, 2019.