Design of an On-Chip Random Number Generator using Metastability

D. J. Kinniment,
University of Newcastle, UK
David.Kinniment@ncl.ac.uk

E.G.Chester
University of Newcastle, UK
Graeme.Chester @ncl.ac.uk

Abstract

This paper shows that the internal noise in a bistable exhibits a Gaussian distribution, and is close to the value expected from thermal agitation. We describe a random number generator based on this property that is capable of on chip integration, and is a primary source of high entropy data at 100MHz. The device is held close to metastability by a feedback loop, and is therefore relatively insensitive to circuit asymmetries and drift. Measurements of post-processed data from this source also show a relatively high bit rate and sequences of 2^{23} bits are shown to pass stringent tests for randomness.

1. Introduction

Secure encryption is dependant on sources of truly random numbers for generating keys, and there is a need for an on chip random number generator to achieve adequate security, for example, on a smart card. Pseudo random deterministic processes are vulnerable to attack, because the entropy of the key can never exceed that of the seed state, and the space that needs to be searched is reduced as a result. Here the entropy of a *k* bit key is given by: $H = -\sum_{i-1}^{n} p_i \log_2 p_i$, where p_i is the probability of state *i* out of *n* states. Thus if all 2^k possible states are equally likely, the information cannot be represented by a sequence shorter than k bits, and $H = k$ [1].

Most true hardware random number generators depend primarily on a source of thermal noise, which is then post-processed to reduce the effects of deterministic internal and external influences such as power supply variations, DC bias, and electromagnetic fields [2]. While full integration on to silicon reduces vulnerability to attack, attack methods can include the use of physical, or high power RF interference to bias the distribution of numbers, thus random number generators also require on-chip hardware to continuously monitor the quality of the output and fast enough production of random bits to permit analysis in real time.

Any on-chip source of entropy, such as random motion of charge carriers must also be largely dependent on unpredictable effects, and compatible with i.c. technology. Mechanisms such as nuclear decay are not appropriate, and deterministic internal effects such as bias, and bit-to-bit correlations may reduce the effective entropy, requiring post-processing of large numbers of bits. It is therefore desirable to start with a high bit rate output device.

A method that is both simple to integrate, and high bandwidth, is the use of a bistable device in metastability. The idea that choice between two equally desirable outcomes may be determined by random processes has been known for centuries [3]. More recently the final state of bistable circuits set close to the metastable point has been shown to be affected by noise.[4],[5]. If the origin of the noise is thermal motion, then its random nature suggests that repeatedly clocking a bistable device forced into metastability will produce a succession of binary bits with little correlation between any pair in the sequence. There are several ways in which this effect could be used to produce random bits.

- Hold the bistable close to metastability, and then allow it to resolve, producing a high or a low level according to the polarity of the internal noise at the time of release [8]. The problem here is how to ensure that offset of the initial bias point is much smaller than the noise level.

- Use two independent oscillators, for clock and data, and record only outputs for which metastability lasts much longer than the normal response time (typically 10-15τ, where τ is the time constant of the device) This ensures that the start point bias was much less than noise, but reduces the output rate to only a few bits per second. This rate can be slightly improved by:

- Measuring the time between clock and resolved output, again with two independent oscillator inputs. The probability of the bistable remaining in metastability decreases exponentially with observed output time, so

the variation in time value bits will not be large, leading to low added entropy. There are also difficulties in that the two oscillators are in close proximity, and may become coupled.

For these reasons we have chosen to investigate the first method. Thermal noise, even in sub-micron devices, is relatively small, typically represented by RMS voltage levels of 1mV or less. To ensure that the output is random, any deterministic signals contributing to the input, such as the effects of the previous output, or drift, must be reduced to a level much lower than 1mV. Bellido, [8], relies only on layout symmetry to achieve low offset, but our measurements suggest this is insufficient to achieve an unbiased output stream.

In this paper, Section 2 presents a fast bistable comparator which we call an R-Flop, and gives the results of internal noise and offset measurement, Section 3. shows how feedback is used to overcome offset and drift. and presents the characteristics of the resulting random bit stream. Simple techniques for removing any remaining lack of randomness in the output to a level where it passes the most stringent tests, and monitoring imperfections are given in Section 4.

2. Noise

In order to measure the internal noise in a bistable circuit we designed the R-Flop shown in Figure 1.

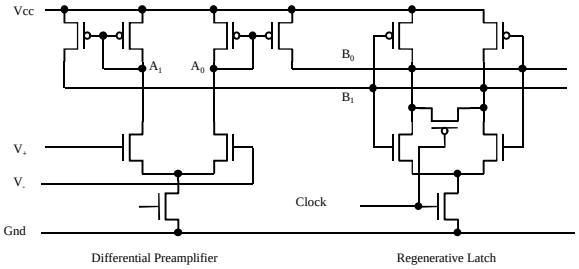

Figure 1 R-flop circuit

This circuit has an analog input stage that drives the bistable with a small current difference, so that we can observe the proportion of outputs that go high or low when the bistable is clocked, with very small initial offset levels. Using DC inputs to give very small inputs to the latch avoids the jitter problems of maintaining the very precise timing between data and clock required to measure noise effects, since input voltage variations can be held to well within 1mV at the low frequencies used to provide the offset. Our circuit has a differential preamplifier input, followed by a bistable latch, and an SEM scan of the device shown in Figure 2.

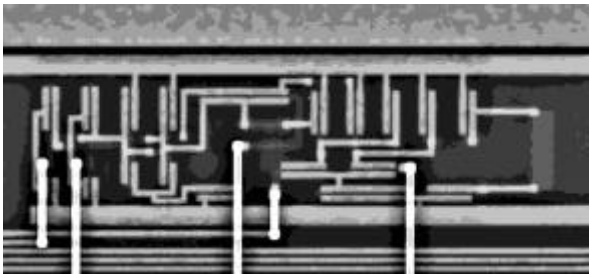

Figure 2 SEM of circuit

The device was fabricated using an AMS 0.6µ process, and we measured the noise in the bistable when it was metastable by continuously clocking the device, while observing the proportion of outputs with V_1 high as the input was varied. When $V+$ is very close to $V-$ the bistable output is determined mainly by thermal noise on the B nodes, since the RMS noise voltage on these nodes is greater than the offset due to the input. Under these circumstances the random nature of the output can be clearly observed, and as the input voltage changes from negative through zero to positive the proportion of V_1 high outputs goes from zero to 100%. Plotting the change in this proportion for a given input change against the actual input gives the graph of Figure 3, where the points measured are compared with a Gaussian curve with an equivalent RMS noise value of 1.7mV at the input. We measured the effective gain of the preamplifier by using a SPICE model to determine the input ($V+$ - $V-$) that would be needed to overcome an initial offset of 1mV in the bistable latch. The results of these measurements show that 3 mV at the input is equivalent to 4.55 mV between nodes A_0 and A_1, and 1 mV between nodes B_1 and B_0

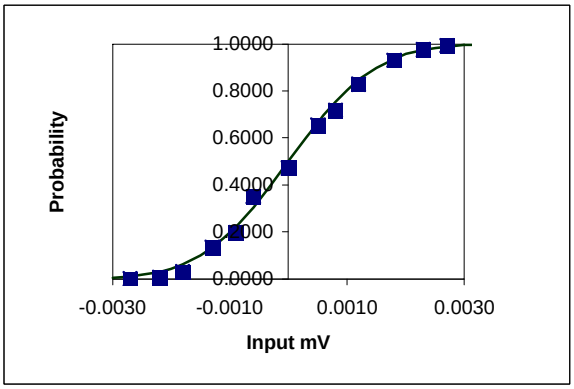

Figure 3 Probability of V_1 high

Thermal fluctuations in an FET give rise to drain current noise and gate noise. Van der Ziel [7], gives drain current noise as $i_{nd}^{\ 2} = 4kTgg_{d0}\Delta f$, where g_{d0} is the drain-source conductance, and the parameter γ has

a value typically between 2 and 3. The equivalent gate noise voltage is approximately $e_{ng}{}^2 = \frac{4kTd\Delta f}{5g_{d0}}$ where δ is typically 5-6. Noise at nodes B_1 and B_0 has a bandwidth limited by the capacitance, C, to $\frac{g_d}{4C}$, and is in the range 0.4 mV to 0.5 mV for our process, depending on the values of γ and δ. Between nodes B_1 and B_0 this is $\sqrt{2}$ greater, or between 0.55 mV and 0.7 mV. Similarly the noise between A_1 and A_0 is about 1 mV, but this only makes a small contribution towards the noise at B, because the gain from A to B was 1/(4.55) in the circuits we measured. Total thermal noise should be therefore equivalent to between 1.65mV and 2.1mVat the input. Our measurement of approximately 1.7mV RMS corresponds to about 0.6mV total between B_1 and B_0, and is in the expected range. Consequently we are confident that the noise is largely thermal in origin.

3. Random number generator

In order to produce a random bit stream, it is necessary to ensure that the bistable device is initially held at the metastable point, and also that variations in bias are much less than the 0.5mV noise voltage. Our samples showed typical input offset voltages of around 10mV, and this could be as much as 50mV, an order of magnitude greater than the effective noise level at the input. Between samples, and with power supply change, this offset value varies considerably, and therefore the random number generator must have some means of adjusting the bias. We provided this with a negative feedback loop, which averages the output voltage to ensure that the output approximates to 50% of high and 50% of low values. This is achieved by means of the switched capacitor network in Figure 4, in which the output bit stream charges a 0.01pF capacitance on each half of the clock cycle, and shares charge with a 100pF capacitance on the other half.

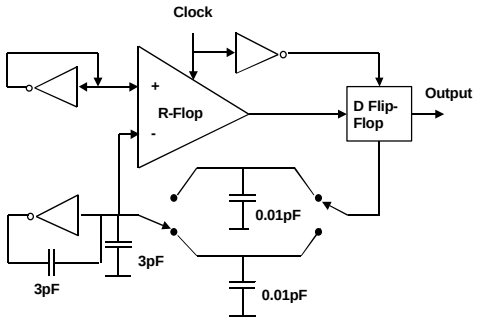

Figure 4 Random number generator

An effective capacitance of 100pF is produced by using the Miller effect to multiply a 3pF capacitor by the gain of an inverter. Here each 3.3V output high causes an increase in voltage of 0.17mV, and each output low of 0V causes a similar decrease This change is still quite large when compared with 1.7mV, so we lower the gain of the circuit shown in Figure 1, by splitting the input stage tail current transistor and inserting a resistor. This increases effective noise range at the input to about 12mV, so that the feedback step size of 0.17mV is less than 2% of the noise. The feedback causes some colouring of the output, since a high output causes the input to be slightly biased in favour of following lows, and a low output biases in favour of following highs. Other causes of reduction in the entropy of the output are DC bias and parasitic feedback of the output to the inputs via power supplies or other coupling. We measured the correlation of highs and lows in the output by feeding it into a shift register and XORing the output of each stage in the shift register with the first. Averaged over a sufficiently long output sequence, each XOR output should produce a value close to 50% high and 50% low. If the ratio of clock frequency to integration time constant is too low, the first few XOR outputs show inverse correlation, and in the case of a 1000:1 capacitor ratio we observed that 61% of outputs were the opposite of the previous value rather than 50%.

4. Noise quality

The XOR circuits are effectively performing an autocorrelation on the output bit stream in real time, and can be used to monitor the quality of the output. If the XOR outputs are each counted over a suitable number of clock periods, it is possible to determine whether the number of high values is significantly different from 50%, and by implication whether the output is sufficiently random over that number of bits. At a clock rate of 100MHz, a sample of 1000 bits can be checked every 10 μs, and if the quality is insufficient, perhaps because of attack, this can be indicated.

The randomness in the output of our circuit is affected by residual error in the feedback loop, which can give a small DC bias to the output, and by the correlation between successive bits discussed in the previous section. Inevitably other systematic noise may also affect the sensitive input of the R-Flop, but because the bit rate of our device can be very high, it is possible to reduce these effects by XORing two strings, from similar circuits, and using the resulting string as the output, or alternatively taking the parity of a pair of successive bits. If the two strings are independent, his 'whitening' process has the effect of reducing a probability of (0.5 + x) in the correlation between any two bits to ($0.5 + 2x^2$). Thus 61% of inverse correlation

becomes 52.4%. N inputs to the XOR process give a $0.5(1 + (2x)^N)$. If the bits are from only one string the whitening process relies on the fact that correlation reduces as the bits are more widely spaced, and larger values of N must be used to get the same effect, but it is trivially implemented by counting the number of 1's in the output over N clock periods. While this method reduces the bit rate in order to increase the entropy, it is superior to methods which maintain the bit rate by using a linear feedback shift register (LFSR) [8]. For these methods, because a knowledge of the feedback paths and the output enables the input to be reconstructed, entropy is not increased in the output.

We tested the output of our circuit with a discrete-component averaging network, storing 2^{23} bits of the post processed files on a PC where they could be analysed for randomness. The effectiveness of the parity method as measured by the variance of 8 bit messages is shown in Figure 5

As the number of bits, N, increases the variance reduces, and the inverse correlation between successive bits can be seen in the difference between odd and even values of N. The entropy ratio H/k for N=5, 25, and 100 is 0.9808, 0.99999 and 1.0 respectively. Both of these last two files pass overlapping m-tuple tests, known as the 'Monkey Test' [9], for up to 16 bits. These tests are more stringent than the empirical tests of Knuth. [10].

5. Conclusions

We have shown that it is possible to use the internal noise of a bistable device in metastability to generate a stream of bits with a high level of randomness, and a frequency of over 100MHz. In order to achieve high entropy in the output, it is essential to use feedback to overcome offset and drift

Because of internal correlations this bit stream has an entropy ratio of only about 0.85, but post processing that reduces the bit rate by a factor of 50, can improve it to the level where it is indistinguishable from a true random source. The techniques used are simpler than an LFSR, and increase entropy at the expense of bit rate. The LFSR method does not increase entropy unless the bit rate is also reduced. All the techniques described are relatively simple to implement of a single chip, and the high bit rate ensures that, if necessary, real time on-chip monitoring of the quality of the output could be carried out.

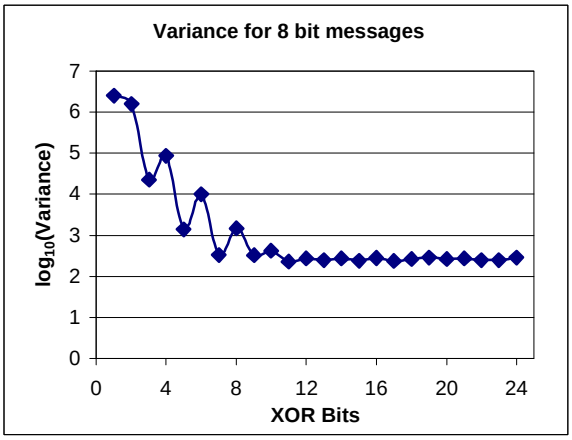

Figure 5 Variance against N

6. References

[1] C. E. Shannon, "A Mathematical Theory of Communication" The Bell System Technical Journal, Vol. 27, p379-423, July 1948

[2] B. Jun and P Kocher "The INTEL Random Number Generator", Cryptography Research Inc

[3] N Rescher, "Choice without preference: A study of the logic and the history of the problem of Buridan's Ass" Kant-Studien, 1959/60, No. 51, pp142-175.

[4] C Dike and E Burton "Miller and Noise Effects in a Synchronizing Flip-Flop" IEEE Journal of Solid State Circuits Vol. 34 No. 6, pp849-855, June 1999

[5] G. R. Couranz., and D.F. Wann, "The theoretical and experimental behaviour of synchronizers operating in the metastable region", IEEE Transactions on Computers C-24, (6) pp. 604-616 June 1975.

[6] M.J.Bellido, A.J.Acosta, M.Valencia, A.Barriga, and L.J.Huertas "Simple Binary Random Generator" Electronics Letters Vol 28 No.7 pp617-618

[7] A van der Ziel, "Thermal Noise in Field Effect Transistors", Proc. IEEE, August 1962, pp1801-12

[8] A.J.Acosta, M.J.Bellido, M.Valencia, A.Barriga, and L.J.Huertas "Fully Digital Redundant Random Number Generator in CMOS Technology" 19th ESSCIRC conf. 1993 pp198-201.

[9] G. Marsaglia, "A Current view of Random Number Generators" Keynote address, Proc. Computer Science and Statistics, 16th symposium on the Interface Elsevier 1985.

[10] D.E.Knuth "The art of Computer Programming" 3rd Edition, Addison Wesley, Reading MA, 1998, p75.