一种真随机数发生器的后处理方法

欧海文[1] 赵 静[2] 李启瑞[2]
(北京电子科技学院 北京 100070)[1] (西安电子科技大学通信工程学院 西安 710071)[2]

摘 要 从理论上分析了几种随机数发生器后处理方法,提出了一种新的针对真随机数发生器的后处理方法,经过后处理的随机序列满足均匀性、独立性以及提高每比特熵的要求。对实际带有偏差的数字化噪声序列使用这种方法进行处理后得到的内部随机序列进行随机性检测。检测结果表明,这是一种简单有效的后处理方法,它满足实现面积小、功耗低的要求,可以用于智能卡芯片中。

关键词 真随机数,后处理,偏差,随机性检测
中图法分类号 TN402 **文献标识码** A

Post-processing Method in Truly Random Number Generator

OU Hai-wen[1] ZHAO Jing[2] LI Qi-rui[2]
(Beijing Electronic Science and Technology Institute, Beijing 100070, China)[1]
(Institute of Telecommunication Engineering, Xidian University, Xi'an 710071, China)[2]

Abstract Several post-processing methods were theoretically analyzed in this thesis and a new post-processing method for true random number generator was proposed. The random sequence after post-processing meets the requirements of uniformity, independence, improve of entropy per bit. In case of arrays with bias digital noise, the above method is used. Moreover, randomness tests are performed on the derived inner random arrays. The evaluate results show that this is a simple and effective post-processing method that meets the requirements of small area and low power. The conclusion can be used in smart card.

Keywords Truly random number, Post-processing, Bias, Randomness test

1 引言

随着计算机网络和通信技术的发展,密码学的应用也随之迅速地发展起来,作为密码重要组成内容之一的真随机数如今已被广泛应用在密钥生成、随机填位、密码协议以及为伪随机数发生器提供种子等方面。真随机数是通过提取自然界物理现象中的随机特性而产生的,具有随机性、不可预测性、不可重复性。目前产生真随机数的方法有:直接放大热噪声;基于振荡器采样;基于亚稳态电路[1];利用量子力学基本量的完全随机性产生随机数——如放射性元素的衰变和机关斑纹图案空间分布的随机性等[2]。

在理想情况下,真随机数发生器输出的随机位出现'0'和'1'的概率应相等。但实际上,由于电路内部其它非高斯型噪声影响、电路外部环境温度变化对系统稳定性影响、以及实现模拟电路的精确性等因素影响[3],使得其所产生的随机位不可能是严格等概率的,这就必将影响所产生随机序列分布的均匀性,为此必须采取所谓数字后处理方法对随机位的概率偏差加以修正。

本文分析和研究了冯-诺伊曼、异或链、线性反馈移位寄存器及异或周期序列等数字后处理方法,并在此基础上提出了一种针对真随机数发生器需求的后处理方法,在调节序列的均匀行、独立性以及每比特熵等方面有很好的效果,增强了序列的随机性。

2 常见的数字后处理方法

根据德国联邦信息安全办公室(BSI)发布的用于检验真随机数随机性能的AIS31标准的要求,随机数产生的流程如图1所示[4]。

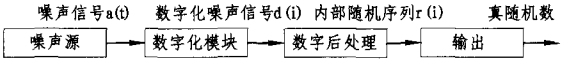

图1 随机数生成流程

噪声源产生模拟的噪声信号 $a(t)$ 经过数字化后成为数字化噪声信号 $d(i)$,再经过数字后处理称为内部随机数 $r(i)$。数字后处理必须满足两条性质:第一,必须能够调整数字化噪声信号 $d(i)$ 的概率分布,使其能够克服由非确定性噪声源或数字化过程带来的统计缺点,从而使内部随机数 $r(i)$ 的分布与数字化噪声信号 $d(i)$ 相比更接近均匀分布;第二,数字后处理不能降低平均每比特的熵,最好能够增加每比特的平均熵(例如使用压缩函数,以降低输出序列的输出速率为代价)。由此可见,数字后处理的目的就是在不降低输出序列熵值的

欧海文(1963-),男,博士,教授,硕士生导师,主要研究方向为密码编码与密码应用技术等;**赵 静**(1986-),女,硕士生,主要研究方向为信息安全、真随机数发生器及后处理等,E-mail:zhaojing596@sina.com;**李启瑞**(1985-),男,硕士生,主要研究方向为信息安全、侧信道攻击与防御技术等。

前提下，纠正序列概率分布及相关性等统计弱点，使其可以通过随机性检测。常见的数字后处理技术有冯-诺伊曼校正、异或链、线性反馈移位寄存器等，下面将对典型的后处理方法进行详细的分析。

2.1 冯-诺伊曼校正器

冯-诺伊曼校正器针对的对象是0、1出现概率固定的真随机数发生器的输出，且输出的数字化噪声信号是不相关的。校正的具体方法是：对发生器输出的数字化噪声序列分组，每相邻的两位为一组，对每个分组进行判断，如表1所列，如果是'00'和'11'则丢弃，如果是'01'则输出'1'，如果是'10'则输出'0'。

表1 冯-诺伊曼校正判断准则

分组	输出
00	无
01	1
10	0
11	无

假设真随机数发生器输出'1'的概率设为 $p=0.5+e$，则输出'0'的概率设为 $q=0.5-e$。很容易得到"00"发生概率为 $(0.5-e)^2$，"11"发生概率为 $(0.5+e)^2$，"01"和"10"发生概率均为 $(0.5-e)*(0.5+e)$。因此，冯-诺伊曼校正器可以完全去除偏差，但是要以丢弃一半以上的输入数据为代价，降低了产生真随机数的速率。需要被抛弃的分组发生的概率是 $0.5+2e^2$，那么若需 X 比特输出，则需要向校正器输入 $X/(0.25-e^2)$ 比特。

冯-诺伊曼后处理在纠正序列的均匀分布和相关性方面效果很好，但是其缺点也十分明显，即造成内部随机序列输出速率的不定倍数的下降，因此，这种后处理方法不适于对真随机数产生的速率有明确要求的应用。

2.2 异或链

简单的级联异或链相当于一个奇偶校验，可以解决序列相关性与均匀分布的问题，但是真随机序列的产生速率会下降。简单级联异或链的结构如图2所示，其是由4个D触发器组成的4级异或链，可以根据需要通过改变触发器的数量来改变异或链的级数。

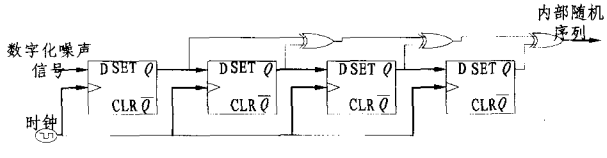

图2 4级异或链

当数字化噪声信号的占空比为 P，即产生1的概率为 P 时，产生0的概率就是 $1-P$。由于实际产生的数字化噪声信号存在偏差，因此 P 不等于0.5。将相邻的两个采样值进行异或，输出1的概率是 $2P(1-P)$，输出零的概率是 $P^2+(1-P)^2$。由数学归纳法，如果采用 n 级异或链，那么输出1的概率是 $0.5-2^{n-1}(P-0.5)^n$，输出0的概率是 $0.5+2^{n-1}(P-0.5)^n$。当 n 趋近无穷大时，输出0和1的概率都趋近于0.5。

实验发现，应用中需要至少8级以上的异或链才能有效清除随机序列中存在的偏差，当真随机数发生器产生的数字化噪声信号偏差较大时，需要的异或链的级数就更多，因此大大降低了随机系列的输出速率，这种后处理方法不适于对真随机数产生的速率有很高要求的应用。

2.3 线性反馈移位寄存器

这种数字后处理是由长度为 k 的线性反馈移位寄存器实现的，具体结构如图3所示，其采用的本原反馈多项式可以有多种选择。数字化噪声信号的输入与移位寄存器的循环移位同步，反馈的位与数字化噪声信号当前的位进行异或运算后反馈到移位寄存器的最低位，移位寄存器的低8位作为输出。

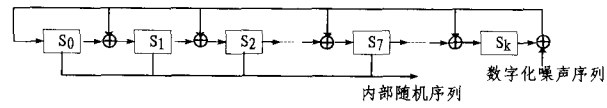

图3 线性反馈移位寄存器做后处理

对于每一个初始配置，这类数字后处理可以看作是数字化噪声序列集合到内部随机序列集合的一个双向映射。每输入 n 比特数字化噪声信号，输出8比特的内部随机数，这就实现了一个压缩率为 $n/8$ 的压缩函数。若输入的数字化噪声信号和寄存器的内部状态分别是独立的，那么就可以保证输出的内部随机序列是独立的[5]。

移位寄存器方式的压缩率对随机性直接影响，输入的数字化噪声信号的独立性对输出序列的独立性也有直接的影响，这种后处理方法不适合使用在产生的数字化噪声信号独立性不好的真随机数发生器中。

2.4 异或周期序列

均匀分布并且不具有相关性的周期序列与均匀分布的非周期序列做异或运算，可以产生一个均匀分布并且不具有相关性的非周期序列[6]。均匀分布并且不具有相关性的周期序列可通过多种方式实现。

这种后处理方法的优点是可以很好地调整数字化噪声性信号的随机性，并且不会降低数据的输出速率。缺点是对于均匀性不好的数字化噪声信号，这种方法不适用；由于没有增加随序列的每比特熵，因此对于每比特的熵没有达到要求的真随机数发生器不适用。

3 后处理方法的设计

为了使经过后处理的随机序列独立且均匀分布，并且每比特有足够的熵，我们在异或周期序列的后处理方法上进一步改进、设计的后处理方法的基本结构如图4所示。

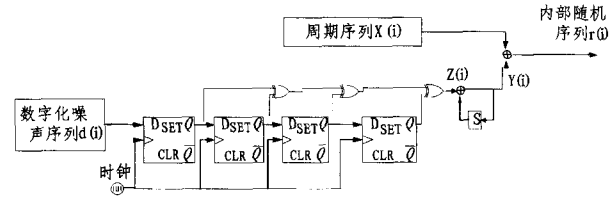

图4 改进的后处理方法

数字化噪声序列 $d(i)$ 是非周期序列，通过四级异或链的处理，初步调整了数字化噪声信号的均匀分布和独立性。由于异或链实现了对数字化噪声信号压缩率为4的压缩，因此也提高了随机序列的平均每比特熵，将产生的随机序列记为 $Z(i)$。

S 可存储1比特的数据，其初始值可以是0或1。S 中的值与 $Z(i)$ 异或后记为 $Y(i)$，将 $Y(i)$ 一方面存于 S 中用于下一位的计算，另一方面输出与均匀分布并且不具有相关性的周期序列 $X(i)$ 做异或运算，产生内部随机序列 $r(i)$。均匀分

布并且不具有相关性的周期序列可由多种方式实现，这里使用32位线性反馈移位寄存器产生的小 m 序列是均匀分布的并且有极好的周期的自相关特性[7]。

由异或运算的性质可知：

$a \oplus b = a+b-2ab, (a,b \in \{0,1\})$

我们可以得到

$E[Y(i+1)]=E[Y(i)]+E[Z(i)]-2E[Y(i)]E[Z(i)]$

显然 $E[Y(i+1)]=E[Y(i)]$，所以当 $E[Z(i)]\neq 0$ 时，$E[Y(i)]=0.5$，即通过异或使得 $Y(i)$ 是均匀分布的，最终使得产生内部随机序列 $r(i)$ 是均匀分布并且不具有相关性。

经过分析可知这种后处理方法对输入的数字化噪声信号没有特别的要求，可以解决随机序列均匀分布与独立性的问题，压缩率小，并能有效地提高每比特熵，符合真随机数发生器的要求。

4 随机性检测及结果分析

根据 AIS31 标准的要求，内部随机序列必须能够通过 T0—T5 这几项检测，并且必须保证后处理过程没有降低每比特的熵，每项测试的名称及通过的条件如表 2 所列。我们选取100组数据进行测试，每组数据为包含 7200000 比特的二进制序列。

表 2 AIS31 标准的随机性检测项

编号	测试项	Pass low	Pass high
T0	disjointness test	——	——
T1	monobit test	9654	10346
T2	poker test	1.03	57.4
T3	runs 1	2267	2733
	runs 2	1079	1421
	runs 3	502	748
	runs 4	223	402
	runs 5	90	223
	runs 6+	90	223
T4	long run test	1	33
T5	autocorrelation test	2326	2674

用于做随机测试的数字化噪声信号来源于基于振荡器采样的真随机数发生器。由于受电路内部以及外界环境中非理想因素的影响，使得产生的数字化噪声序列带有偏差。我们对100组数字化噪声信号进行检测，结果发现数字化噪声信号并不能完全通过随机性检测，根据 AIS31 测试标准对这种发生器产生的数字化噪声信号进行熵值测试得到的每比特熵值<7.976，需要增加每比特的熵。

使用 C 语言对上述的几种后处理方案进行实现，然后分别对100组数字化噪声信号进行后处理。由于数字化噪声信号的每比特的熵不足，因此不能采用异或周期序列的后处理方法。

异或链方式采用的是4级异或链；线性反馈移位寄存器方式采用的压缩率是8，即每输入64比特数据则输出8比特数据。得到的4类内部随机序列经过随机性检测的结果如表3所列。

实际上，模拟电路上产生数字化噪声信号的速率往往不大，如果采用压缩率较大的数字后处理方法，那么产生真随机数的速率就会更小，达不到实际应用的需求，因此数字后处理的压缩率不能过大。

表 3 随机序列检测结果

Test \ Passed%		数字化噪声信号	冯-诺伊曼	异或链	线性反馈移位寄存器	设计的后处理
T0		100	100	100	100	100
T1		97	100	99	100	100
T2		98	100	100	100	100
T3	runs 1	98	100	100	100	100
	runs 2	98	100	100	100	100
	runs 3	99	100	100	100	100
	runs 4	99	100	100	100	100
	runs 5	99	100	100	100	100
	runs 6+	98	100	100	100	100
T4		100	100	100	100	100
T5		98	100	100	100	100

通过实验我们发现冯-诺伊曼后处理对不同组数据的压缩率从4到90不等。因此，尽管其硬件实现简单，但是这种后处理方法不适于对真随机数产生的速率有明确要求的应用。

4级异或链的压缩率为4，由表3可知，经过4级异或链后处理的内部随机序列，并不能全部通过随机测试，需要通过增加异或链的级数，也就是增大压缩率来提高内部随机序列的随机性。我们将压缩率增大到8，测试发现内部随机序列可以完全通过随机测试。

移位寄存器方式的压缩率对随机性有直接影响，通过测验得到，当压缩率大于1时，内部随机序列可以通过T0到T5的所有测试，但是当压缩率为1，也就是不对数字化噪声信号进行压缩时，T1项的通过率为99%，T2项的通过率降为99%。

所设计的后处理方式对序列进行压缩，不仅提高了每比特熵，并且可以很好地提高序列分布的均匀性和随机性，经其处理后的内部随机序列可以通过全部测试项，这种后处理方法硬件实现简单，所占面积小，能在一定程度上降低发生器的输出速率，综合来看是一种比较好的后处理方法，适用于智能卡芯片中。

结束语 本文分析了冯-诺伊曼、异或链、线性反馈移位寄存器及异或周期序列这几种数字后处理方法，设计了一种新的后处理方法，将这些后处理方法应用于带偏差的数字化噪声序列进行纠正后得到的内部随机序列应用 AIS31 标准进行随机性检测，比较检测结果并结合它们各自优缺点的分析，认为所设计的后处理方法电路实现简单面积小，对输入数据无特别要求，能有效提高随机序列的随机性，适于智能卡芯片等应用。

参 考 文 献

[1] Varchola M, Drutarovsky M. New High Entropy Element for FPGA Based True Random Number Generators[C]// Mangard S, Standaert F-X, eds. CHES 2010. LNCS 6225, Springer, Heidelberg, 2010: 351-365

[2] 吕玉祥，牛利兵，张建忠，等. 基于混沌激光的 500Mb_s 高速真随机数发生器[J]. 中国激光，2011，38(5)

[3] Bochard N, Bernard F, Fischer V. Observing the randomness in RO-based TRNG[C]// International Conference on Reconfigurable Computing and FPGAs. Cancun, Quintana Roo, Mexico, IEEE Computer Society, Los Alamitos, December 2009: 237-242

（下转第 14 页）

性变化的，而未改进的 QEMU 的响应时间随着程序大小的增加呈现抛物线的增长。

4.2 程序的危险系数

根据算法，对恶意程序的分析需要确定一个阈值，危险系数超过该阈值的都会报恶意程序。本次对网络上一些典型的恶意程序进行分析和计算（初始置权值和危险值为 1 的倍数），得出了总体的危险系数大于 90，又对网络上一些典型的白名单程序进行分析和计算，得出了总体白名单的危险系数都小于 30，故 $30<x<90$ 的值可以作为分析的阈值，权值和危险值都设置为 1 的倍数结果图如图 2 和图 3 所示。

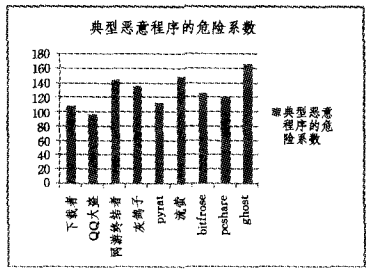

图 2 典型恶意程序的危险系数

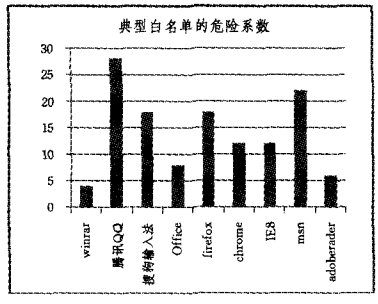

图 3 典型白名单的危险系数

5 恶意程序检测结果

采用两组独立恶意程序样本集进行完全独立的实验。实验一与实验二采用同样的模型，对样本全集进行 8 次不同的阈值测试后的实验结果如表 2 所列。

表 2 两次实验结果

阈值	实验一 $O_{2821} \cup O_{8223}$		实验二 $T_{1694} \cup T'_{7143}$	
	检出率	误报率	检出率	误报率
30	97.3%	11.3%	98.8%	19.3%
40	95.9%	9.1%	97.4%	7.8%
50	95.7%	8.9%	97.4%	8.3%
60	95.2%	8.5%	97.3%	9.1%
70	95.3%	6.8%	95.9%	11.2%
80	91.8%	8.9%	93.7%	12.0%
90	87.3%	9.7%	90.8%	13.4%

从表 2 可以看出：

(1)对于大量的恶意样本和少量的白样本，恶意样本中每个都有许多的恶意行为，对于不同的阈值，都出现了很高的检出率，但是在阈值较低的情况下，如阈值等于 30，有很高的误报率，这主要是由于大量的白样本的一些行为被当作恶意程序的恶意行为。同样在阈值很高的情况下，如阈值等于 100，检出率还是很高，并且还有很高的误报率，这主要是由于有很大一部分恶意程序的危险指数停留在 70～80 阶段，导致了这些恶意程序被当作白名单程序。大概在阈值等于 70 时达到了一个平衡，有着很高的检出率 95.3%，但是误报率只有 6.8%。

(2)对于少量的恶意样本和大量的白样本，总体出现了高的检出率，但是误报率很高。在阈值等于 30 的情况下，检出率有 98.8%，但是误报率高达 19.3%，这主要是由于大量的白样本的一些行为被当作恶意程序的恶意行为。而在阈值等于 40 的情况下，出现了高的检出率，却只有 7.8%的误报率。这主要是由于有大量的白样本的恶意指数在 30 和 40 之间，这些白样本在阈值的增长中被正确检出。随着阈值的增加，检出率总体减小，但是误报率还是呈增长趋势。

结束语 针对当前检测技术以及沙盒技术的不足提出了一种基于沙盒虚拟机动态分析技术，利用 QEMU 实现了一个进程仿真所需要的进程虚拟机，得到了应用程序行为序列流，对该行为序列流进行了形式化定义和分析，提出了行为分析方法。在实践中对上万个样本进行了分析和检测，检测结果表明，该系统可以有效地对恶意程序进行检测，与传统的分析方法相比，可以有效地检测出未知的恶意程序，与主流的沙盒系统相比，增加了人性化操作和易用性、可扩展性。

参 考 文 献

[1] 吴冰，云晓春，高琪. 基于网络的恶意代码检测技术[J]. 通信学报，2007，11

[2] Vimal K K. Securing communication using function extraction technology for malicious code behavior analysis[J]. Computers and Security，2009，28：77-84

[3] Shankarapani M K，et al. Malware detection using assembly and API call sequences[J]. Journal in Computer Virology，2011：107-119

[4] Sami A，Yadegari B，et al. Malware detection based on mining API calls [C]// Proceedings of the 2010 ACM Symposium on Applied Computing. Mar. 2010：1020-1025

[5] 曹跃，梁晓，李毅超，等. 基于差异分析的隐蔽恶意代码检测[J]. 计算机科学，2008

[6] 沙为超，谢荣传. 一种基于本地化特征的恶意代码检测系统设计[J]. 电脑知识与技术：学术交流，2007

[7] 刘艳萍. 恶意代码分析与检测研究现状[J]. 微电脑世界，2009

[8] 王海峰，夏洪雷，孙冰. 基于程序行为特征的病毒检测技术与应用[J]. 计算机系统应用，2006

[9] 史培培，吕林，张明威，等. 基于行为的恶意程序监测研究[J]. 计算机时代，2007

(上接第 11 页)

[4] Killmann W，Schindler W. AIS 31：Functionality classes and evaluation methodology for true (physical) random number generators，version 3.1，Bundesamt fur Sicherheit in der Informationstechnik (BSI). Bonn，2001

[5] Bucci M，Luzzi R. Digital Post-processing for Testable Random Bit Generators[M]. IEEE Circuit Theory and Design (ECCTD 2007). 2007

[6] Sugahara TT，Inoue T. Statistical Properties of Modulo-2 Added Binary Sequences[J]. IEICE Trans. Fundamentals，2004，E87-A (9)：2267-2273

[7] Sarwate D V，Pursley M B. Crosscorrelation properties of pseudorandom and relater sequencesproc[J]. IEEE，68(3)：593-617