分组密码算法抗故障攻击能力度量方法研究

欧庆于 罗 芳* 叶伟伟 周学广
(海军工程大学信息安全系 武汉 430033)

摘 要：该文从算法层面对分组密码固有的故障泄露特点进行了分析，提出一种可用于刻画其故障传播特性的传播轨迹框架，并以此为基础构建了适用于单次和多次故障注入场景的抗故障攻击能力度量方法。实验表明，该度量方法能够有效刻画不同故障注入场景下密钥空间的变化规律，进而揭示其算法层面的抗故障攻击能力。

关键词：分组密码；故障攻击；度量
中图分类号： TP309.1 **文献标识码**： A **文章编号**：1009-5896(2017)05-1266-05
DOI: 10.11999/JEIT160548

Metric for Defences Against Fault Attacks of Block Ciphers

OU Qingyu LUO Fang YE Weiwei ZHOU Xueguang
(*Department of Information Security, Naval University of Engineering, Wuhan* 430033, *China*)

Abstract: A detailed analysis of the fault features for the block cipher is performed, and an analysis framework for propagation of faults is proposed. Furthermore, a security evaluation methodology with single fault injection or multi fault injection is presented. The experiment results show that the change of the key space for the block cipher, using different fault attacks, can be charactered effectively and the ability of the fault-resistant can be presented well.

Key words: Block ciphers; Fault attacks; Metric

1 引言

在诸多基于物理泄露信息的攻击手段中，故障攻击由于能够被攻击者主动策划，为成功猜测秘密信息提供了更多的选择和更高的可能性，对密码应用安全造成了严重威胁[1−6]。为了应对故障类攻击对密码应用安全造成的威胁，多种防御方案被提出[7−10]。然而，正如已经被指出的，“设计出一个没有旁路安全脆弱性的通用密码应用框架是不可行的”[11]。这意味着，故障攻击的安全威胁是密码应用领域需要长期面对的问题。

本文深入分析了分组密码算法层面故障攻击特点，提出了一种可用于刻画其故障传播特性的传播轨迹框架，并以此为基础构建了适用于单次和多次故障注入场景的抗故障攻击能力度量方法。实验表明，该方法能够有效刻画分组密码在不同故障攻击场景中密钥空间的变化规律，并进而揭示其抗故障攻击的能力。

2 算法层面故障攻击特点分析

当前，有一类故障攻击方法有赖于对分组密码正常和故障数据流的分析(如 CFA[12], IFA[13], DFA[14], SEA[15]等)，进而达到猜测密钥的目的。由于在算法、明文、密钥及故障注入场景确定的前提下，正常数据流与故障数据流不因密码算法的具体实现方式而改变，而仅与算法结构相关，故称该类故障攻击方法为算法层面的攻击。算法层面攻击的实施基础是密码算法的固有特性，该特性在算法结构被确定后即被固化。

在 DFA 攻击中，攻击者根据故障密文 $\widetilde{C}$ 与正常密文 C 之间的差异 ΔC，对正确密钥 K 进行猜测。其中，密钥 K 由各轮的轮密钥组成。ΔC 与密钥 K 之间的关系为

$$\Delta C = \mathrm{Diff}\left(C,\widetilde{C}\right) = \mathrm{Diff}\left(E_{\mathrm{pr}}\left(I,K_{\mathrm{pr}}\right),E_{\mathrm{pr}}\left(\tilde{I},K_{\mathrm{pr}}\right)\right) \quad (1)$$

其中，$\mathrm{Diff}(\cdot,\cdot)$ 表示两个变量之间的差分值，E_{pr} 为故障注入点之后的局部加密操作，K_{pr} 为参与局部加密操作的密钥、I 为正常中间数据、$\tilde{I}$ 为被故障注入干扰的中间数据。从式(1)可以看出，DFA 所利用的泄露信息主要是由故障注入点之后的故障局部数据流与正常局部数据流之间的差异造成的，而这种数据流差异外在表现为输出密文之间的差异。

在故障条件下基于明文 P_0 输出故障密文 $\widetilde{C}_0$，如能找到碰撞对 $\left\{\widetilde{C}_0,C_1\right\}$ (C_1 为基于明文 P_1 产生的正常密文)，且 $\widetilde{C}_0 = C_1$，则可实施 CFA 攻击。碰撞

收稿日期：2016-05-28；改回日期：2017-04-17；网络出版：2017-04-19
*通信作者：罗芳 ouqingyv@163.com
基金项目：国家自然科学基金(61202338)
Foundation Item: The National Natural Science Foundation of China (61202338)

对$\{P_0, P_1\}$与密钥 K 之间的关系为

$$\begin{aligned}\text{Collision} &= \text{Equal}\left(C_1, \widetilde{C}_0\right)\\ &= \text{Equal}\left(E_{\text{pr}}\left(I_1, K_{\text{pr}}\right), E_{\text{pr}}\left(\tilde{I}_0, K_{\text{pr}}\right)\right)\end{aligned} \quad (2)$$

其中，Collision 表示碰撞攻击，Equal(·,·)表示两个变量相等，I_1为故障注入点位置处，与明文 P_1相对应的正常中间数据；$\tilde{I}_0$为与明文 P_0相对应的被干扰中间数据。从式(2)可看出，CFA 利用的泄露信息主要由局部数据流相同的明文对(P_0, P_1)造成，其根源在于明文与密钥对算法数据流的共同影响。从 C_1与 $\widetilde{C}_0$ 的差异角度，式(2)可表述为

$$\text{Diff}\left(C_1, \widetilde{C}_0\right) = \text{Diff}\left(E_{\text{pr}}\left(I_1, K_{\text{pr}}\right), E_{\text{pr}}\left(\tilde{I}_0, K_{\text{pr}}\right)\right) = 0 \quad (3)$$

由于密码算法可逆，从解密的角度，式(3)又可重新定义为

$$\begin{aligned}\Delta P &= \text{Diff}\left(P_0, \widetilde{P}_0\right)\\ &= \text{Diff}\left(D_{\text{pr}}\left(I_1, K_{\text{pr}}\right), D_{\text{pr}}\left(\tilde{I}_0, K_{\text{pr}}\right)\right)\end{aligned} \quad (4)$$

其中，D_{pr} 为故障注入点之后的局部解密操作，K_{pr}为参与局部解密操作的密钥，$\widetilde{P}_0 = P_1$，$\tilde{I}_0 = I_1$。因此，从式(1)和式(4)可发现，DFA 和 CFA 两种故障攻击方法均是基于差分故障泄露信息实施的。

在 IFA 和 SEA 攻击过程中，攻击者通过寻找不受特定故障注入影响的明文 P，对密钥 K 进行猜测。此时，无效攻击 Ineffecive 与密钥 K 之间的关系为

$$\begin{aligned}\text{Ineffective} &= \text{Equal}\left(C, \widetilde{C}\right)\\ &= \text{Equal}\left(E_{\text{pr}}\left(\tilde{I}, K\right), E_{\text{pr}}(I, K)\right)\end{aligned} \quad (5)$$

从式(2)，式(5)可看出，IFA, SEA 与 CFA 在本质上是相同的。

综上所述，DFA, CFA, IFA, SEA 等算法层面故障攻击方法的成功实施，均有赖于密码算法数据流在故障条件下的差分故障信息泄露。

3　分组密码抗故障攻击能力度量方法研究

3.1 传播轨迹

从映射的角度，分组密码实现了明文空间 $\boldsymbol{P}$ 与密文空间 $\boldsymbol{C}$ 在密钥 K=$(k_0,k_1,\cdots,k_n)$控制下经过一系列中间值空间的映射，其中 k_i 为轮密钥，具体的映射关系序列构成明文在特定算法和特定密钥下的传播轨迹。

定义 1　分组密码加密过程中，明文 P 及其在每轮加密变换中的值 I 所组成的序列 $\{P, I^1, I^2, \cdots, I^r\}$ 称为传播轨迹。

在密码算法、明文及密钥确定的前提下，假设故障在第 $r-1$ 轮被成功注入，并篡改第 $r-1$ 轮中间值，使得原来在轮密钥 k_{r-1} 作用下至中间值 I_i^{r-1} 的映射，被篡改为中间值 I_k^{r-1} $(I_k^{r-1} \neq I_i^{r-1})$，并进而造成后续的在轮密钥 k_r 作用下的第 $r-1$ 轮中间值空间至密文空间的映射被篡改，如图 1 所示。

设正常及故障情况下，明文在密码处理流程中分别形成传播轨迹 $T = (t_0, t_r, \cdots, t_n)$，$\widetilde{T} = (\tilde{t}_0, \tilde{t}_1, \cdots, \tilde{t}_n)$。其中，$t_i$,$\tilde{t}_i$ 分别表示第 i 个中间值空间中的取值。经分析可发现，在明文确定的前提下，各中间值空间中的取值仅受算法结构和密钥两方面的影响。因此，从理论上可将算法结构或密钥对传播轨迹的影响单独提取出来进行分析。

定义 2　在仅考虑算法结构影响情况下，明文 P 及其在每轮加密变换中的值 I_{base} 所组成的序列 $\left\{P, I_{\text{base}}^1, I_{\text{base}}^2, \cdots, I_{\text{base}}^r\right\}$ 称为传播轨迹基。

定义 3　在仅考虑密钥影响情况下，明文 P 及其在每轮加密变换中的值 I_K 所组成的序列 $\left\{P, I_K^1, I_K^2, \cdots, I_K^r\right\}$ 称为传播轨迹偏移。

根据以上定义，可将传播轨迹 T 和 $\widetilde{T}$ 分解为 $T = \left\langle T_{\text{base}}, T_K \right\rangle$，$\widetilde{T} = \left\langle \widetilde{T}_{\text{base}}, \widetilde{T}_K \right\rangle$。其中，$T_{\text{base}}$，$\widetilde{T}_{\text{base}}$ 表示传播轨迹基；T_K，$\widetilde{T}_K$ 表示传播轨迹偏移。传播轨迹基与传播轨迹偏移间采用特定算子 $*$ 结合：

$$T = T_{\text{base}} * T_K,\ \widetilde{T} = \widetilde{T}_{\text{base}} * \widetilde{T}_K \quad (6)$$

3.2 算子的选择

为了便于实施泄露分析，结合算子 $*$ 需满足可分离性、分离的唯一性和包含性。此外，传播轨迹 T 作为反映特定明文在特定密码算法和特定密钥共同作用下的映射序列关系表征，传播轨迹基与传播轨迹偏移之间的结合关系应与密码算法中密钥对数据流的影响方式相容。

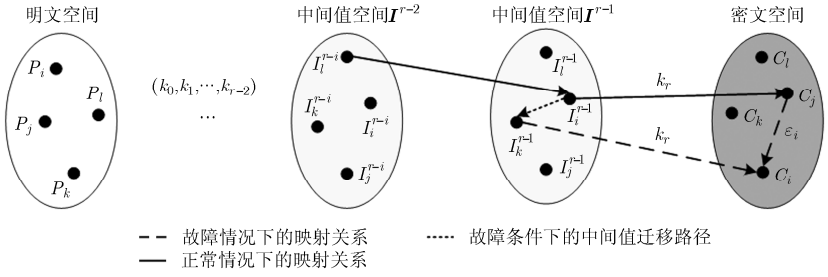


图 1 故障注入情况下的明文传播轨迹示意图

在当前大多数分组密码算法中，轮密钥均基于异或算子参与密码运算。而在如 IDEA 等密码算法中，轮密钥虽然不直接基于异或算子参与密码运算，但每一轮的各路输出仍依赖异或算子产生，从而仍可等价地认为轮密钥基于异或算子参与密码运算。异或算子显然满足可分离性和分离的唯一性；另一方面，由于二元布尔函数的异或的支持空间是这两个布尔函数的支持空间和的子空间，故异或算子满足包含性[16]。

综上所述，选择异或算子 $\oplus$ 作为结合算子，并可通过将密钥 K 置为 0 获得传播轨迹基 T_{base} 和 $\widetilde{T}_{\text{base}}$：

$$T_{\text{base}} = T\left|K=0,\ \widetilde{T}_{\text{base}} = \widetilde{T}\right|K=0 \tag{7}$$

3.3 单次故障注入情况下的泄露度量

在单次故障注入场景下，设正常传播轨迹为 T，故障传播轨迹为 $\widetilde{T}$，则差分传播轨迹定义为

$$T_{\text{diff}} = T \oplus \widetilde{T} \tag{8}$$

其对应的差分传播轨迹基定义为

$$T_{\text{diff}|\text{base}} = T_{\text{base}} \oplus \widetilde{T}_{\text{base}} \tag{9}$$

又设在密码算法及明文保持不变的情况下，基于猜测密钥 $\widehat{K}$ 所产生的正常传播轨迹偏移为 $T_{\widehat{K}}$、故障传播轨迹偏移为 $\widetilde{T}_{\widehat{K}}$，则差分故障分析攻击可基于传播轨迹描述为

$$\begin{aligned}\text{DFA}\left(T_{\text{diff}},\widehat{K}\right) &= \left(T_{\text{base}} \oplus \widetilde{T}_{\text{base}}\right) \oplus \left(T_K \oplus T_{\widehat{K}}\right) \\ &\quad \oplus \left(\widetilde{T}_K \oplus \widetilde{T}_{\widehat{K}}\right) = T_{\text{diff}|\text{base}} \oplus \omega_1 \oplus \omega_2\end{aligned} \tag{10}$$

其中，$\omega_1 \oplus \omega_2$ 为在猜测密钥 $\widehat{K}$ 下获得的传播轨迹偏移差异，表示为 $\Delta T_{\widehat{K}}$。由于算法层面故障攻击的实施，依赖于算法数据流在故障条件下的差分故障信息泄露，故式(10)同样可应用于其他故障攻击方法中。

定理 1 对于确定的传播轨迹基 T_{base}，密钥 K 唯一确定传播轨迹偏移 T_K。

证明 假设在确定传播轨迹基 T_{base} 的前提下，密钥 K 可确定两条传播轨迹偏移 T_K^1 和 T_K^2，且 $T_K^1 \neq T_K^2$，则 $T_{\text{base}} \oplus T_K^1 \neq T_{\text{base}} \oplus T_K^2$。从而可得出在密钥 K 的作用下，同一明文被映射为不同密文的结论，与密码算法中明密文对在密钥作用下的唯一性相矛盾。

推论 1 当 $\widehat{K} = K$ 时，必有 $\omega_1 = \omega_2 = 0$，即 $T_{\text{diff}|\widehat{K}} = 0$。

定义 4 在确定传播轨迹基差异 $T_{\text{diff}|\text{base}}$ 的前提下，将猜测密钥空间 $\boldsymbol{K}$ 根据各猜测密钥 $\widehat{K}_i$ 获得的 $T_{\text{diff}|\widehat{K}_i}$ 进行划分，形成子空间集合 $U = \{u_0, u_1, \cdots, u_m\}$。其中，$u_i = \left\{\widehat{K}_j \mid T_{\text{diff}|\widehat{K}_j} = i\right\}(i = 0,1,\cdots,m)$。称该猜测密钥空间划分为差分传播轨迹偏差划分。

定理 2 对于密钥空间差分传播轨迹偏差划分中的任意两个子空间 u_i 和 u_j，必存在 $u_i \cap u_j = \varnothing$。

证明 根据定理 1 即得证。

定义 5 对于 $u_0 = \left\{\widehat{K}_j \mid T_{\text{diff}|\widehat{K}_j} = 0\right\}$，称其为差分传播轨迹偏差划分的特征空间。

根据推论 1 可知，正确密钥 K 落入子空间 u_0 的概率 $\Pr(K \in u_0) = 1$。又根据定理 2 可知，$\overline{u}_0 = \{u_i \mid i \neq 0\}$，故 $\Pr(K \in u_i \mid i \neq 0) = 0$。因此，当猜测密钥 $\widehat{K}_j \in u_i \mid i \neq 0$，可排除猜测密钥 $\widehat{K}_j$ 作为正确密钥的可能性，这与 $\widehat{K}_j \notin u_0$ 是等价的。

定义 6 对于正确密钥 K，其可辨识度定义为

$$\text{distinguish} = 1\big/\left|u_0\right| \tag{11}$$

可辨识度越高，表示特征空间中的构成元素越少，则攻击者越容易确定正确的密钥 K。因此，在单次故障注入情况下，通过对密钥 K 可辨识度的计算，能够反映在特定参照系 $T_{\text{diff}|\text{base}}$ 下，故障泄露信息的程度。

3.4 多次故障注入情况下的泄露度量

设攻击者能够准确实施 w 次故障注入，产生差分轨迹序列 $T_{\text{diff}} = \left\langle T_{\text{diff}}^0, T_{\text{diff}}^1, \cdots, T_{\text{diff}}^w \right\rangle$，且该序列中存在 $m(m \leq w)$ 个不同元素，则可按式(7)和式(10)进行 m 轮分析，获得集列 $\boldsymbol{U} = \left\{U_0^1, U_0^2, \cdots, U_0^m\right\}$。其中，$U_0^i$ 为第 i 轮分析中获得的差分传播轨迹偏差划分的特征空间。当集列 $\boldsymbol{U}$ 中任意两个集合 $U_0^i \cap U_0^j \neq \varnothing$，由于正确密钥 K 必同时属于 U_0^i 和 U_0^j，攻击者可通过求 U_0^i 和 U_0^j 之间的交集，达到缩小特征空间规模的目的，从而增大可辨识度。因此，在多次故障注入情况下，特征空间收缩的速度对攻击的成功率产生直接的影响。

定义 7 在多次故障注入情况下，不同特征空间交集所造成的特征空间缩小现象称为特征空间衰减，ψ 为特征空间的衰减率

$$\psi = \frac{\max\left(\left|u_0^1\right|, \cdots, \left|u_0^m\right|\right) - \left(\left|u_0^1 \cap \cdots \cap u_0^m\right|\right)}{\max\left(\left|u_0^1\right|, \cdots, \left|u_0^m\right|\right) \times \text{num}} \tag{12}$$

其中，num 表示集列 $\boldsymbol{U}$ 中不同的元素数量。

在故障注入次数相同的情况下，特征空间衰减率越大，意味着特征空间缩小的程度越大，则攻击的成功率越高。因此，通过衰减率能够对多次故障注入情况下的信息泄露程度进行度量。

4 实验及分析

在密钥固定前提下，基于 500 组随机明文，分别对 DES, AES, IDEA 进行单次和多次单比特故障

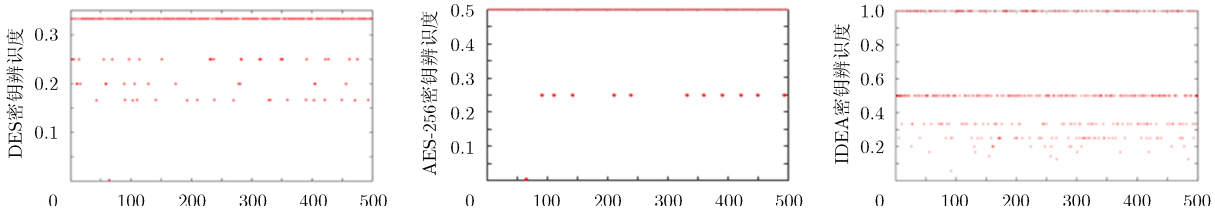
注入[13,17]，注入位置分别为 DES 第 15 轮的左半部分输入的单比特，AES 第 13 轮列混合操作前中间状态第 1 字节的单比特，IDEA 第 8 轮 Z_5 参与的模 $2^{16}+1$ 乘法操作。攻击目标分别为末轮轮密钥 k^{16}, k^{14} 和 Z_1^9。其密钥辨识度及特征空间衰减率的变化趋势如图 2，图 3 所示。

如图 2 所示，单次故障注入时，DES 的密钥辨识度集中于 0.33 区域，即在大部分情况下，攻击者可通过单次故障注入获得 3 个候选密钥，其中包含正确密钥；AES 的密钥辨别度集中于 0.5 的区域，即在大部分情况下，攻击者可通过单次故障注入获得两个候选密钥，其中包含正确密钥；IDEA 的密钥辨识度大部分集中于 0.5 和 1.0 两个区域，即在大部分情况下，攻击者可通过单次故障注入获得两个或直接得出正确密钥。因此，在单次故障注入情况下，DES 的抗差分故障攻击能力最强，AES 次之，IDEA 再次之。

如图 3 所示，在多次故障注入时，DES 特征空间衰减率在第 2 次注入时达到峰值 0.2，即通过两次故障注入能够直接得出正确密钥。AES 和 IDEA 特征空间衰减率均在第 2 次故障时达到峰值 0.2，即通过两次故障注入能够直接得出正确密钥。然而，图 3(a)中，在约第 100 次注入时特征空间衰减率产生波动；图 3(b)中，在约第 70 次注入时特征空间衰减率产生波动。而在图 3(c)中，该波动的产生位置约位于第 10 次注入的时刻。由式(12)可知，产生衰减率波动的原因是由于 num 值的减小 —— 即集列 $\boldsymbol{U}$ 中不同元素数量的减小。这意味着，在连续实施多次差分故障攻击时，IDEA 候选密钥空间同质化的速度要快于 AES, AES 候选密钥空间同质化的速度要快于 DES。

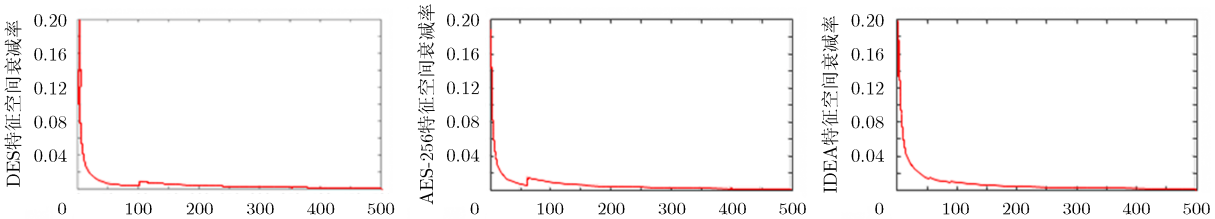
5　结论

围绕分组密码算法抗故障攻击能力度量问题，本文对算法层面的故障攻击特点进行了深入分析，基于密钥空间映射模型和传播轨迹理论提出了一种可用于刻画分组密码差分故障传播特性的技术框架。该技术框架能够较好刻画在故障攻击场景下分组密码算法密钥空间的变化规律，以该技术框架为基础构建的故障攻击能力度量方法，适用于分组密码算法单次故障注入场景和多次故障注入场景的算法层面抗故障攻击能力度量，对于评判分组密码算法的安全性，构建具备故障攻击免疫性的算法结构具有重要意义。



(a) DES-256密钥辨识度　(b) AES-256密钥辨识度　(c) IDEA密钥辨识度

图 2 密钥辨识度变化趋势



(a) DES特征空间衰减率　(b) AES-256特征空间衰减率　(c) IDEA特征空间衰减率

图 3 特征空间衰减率变化趋势

参 考 文 献

[1] DASSANCE F and VENELLI A. Combined fault and side-channel attacks on the AES key shedule[C]. Fault Diagnosis and Tolerance in Cryptography(FDTC), Leuven, Belgium, 2012: 63–71.

[2] THOMAS F, ELIANE J, VICTOR L, *et al*. Fault attacks on AES with faulty ciphertexts only[C]. Fault Diagnosis and Tolerance in Cryptography(FDTC), Santa Barbara, CA, 2013: 108–118.

[3] NAHID F G, BILGIDAY Y, MOSTAFA T, *et al*. Differential

fault intensity analysis[C]. Fault Diagnosis and Tolerance in Cryptography(FDTC), Busan, 2014: 71–78.

[4] RONAN L, GUILLAUME R, JEAN M D, *et al*. A DFA on AES based on the entropy of error distributions[C]. Fault Diagnosis and Tolerance in Cryptography(FDTC), Leuven, Belgium, 2012: 34–43.

[5] ABHISHEK C, BODHISATWA M, and DEBDEEP M. Combined side-channel and fault analysis attack on protected grain family of stream ciphers[OL]. http://eprint.iacr.org/2015/602.pdf, 2015.

[6] REN Y, WANG A, and WU L. Transient-steady effect attack on block ciphers[C]. Cryptographic Hardware and Embedded Systems(CHES), Saint Malo, France, 2015: 433–450.

[7] MA K, LIANG H, and WU K. Homomorphic property-based concurrent error detection of RSA: A countermeasure to fault attack[J]. *IEEE Transactions on Computers*, 2012, 61(4): 1040–1049.

[8] BRIAIS S, CIORANESCO J M, DANGER J L, *et al*. Random active shield[C]. Fault Diagnosis and Tolerance in Cryptography(FDTC), Leuven, Belgium, 2012: 103–114.

[9] SIKHAR P, ABHISHEK C, and Debdeep M. Fault tolerant infective countermeasure for AES[J]. *Security, Privacy and Applied Cryptography Engineering*, 2015, 935(4): 190–209.

[10] PEI L and YUNSI F. Faulty clock detection for crypto circuits against differential fault analysis attack[OL]. http://eprint.iacr.org.org/2014/883.pdf, 2014.

[11] 陈弘毅, 白国强, 徐秋亮, 等. 密码芯片和侧信道攻击发展研究[R]. 2009-2010 密码学学科发展报告, 2010: 126–149.
CHEN Hongyi, BAI Guoqiang, XU Qiuliang, *et al*. Advances in cryptographic integrated circuits and side-channel attacks[R]. 2009-2010 Report on Advances in Cryptology, 2010: 126–149.

[12] AMIEL F, CLAVIER C, and Tunstall M. Fault analysis of DPA-resistant algorithms[C]. Fault Diagnosis and Tolerance in Cryptography(FDTC), Yokohama, Japan, 2006: 223–236.

[13] BLOMER J and SEIFERT J P. Fault based cryptanalysis of the Advanced Encryption Standard (AES)[C]. Financial Cryptography, Heidelberg, 2003: 162–181.

[14] ROCHE T, LOMNE V, and KHALFALLAH K. Combined fault and side-channel attack on protected implementations of AES[C]. Smart Card Research and Advanced Applications, Leuven, Belgium, 2011: 65–83.

[15] JOYE M, QUISQUATER J J, Yen S M, *et al*. Observability analysis-detecting when improved cryptosystems fail[C]. Topics in Cryptology(CT-RSA), Heidelberg, 2002: 17–29.

[16] JOAN D and VINCENT R. The Design of Rijndael AES: The Advanced Encryption Standard[M]. Berlin, Heidelberg, Springer-Verlag, 2002: 123.

[17] CHRISTOPHE C, BENEDIKT G, and INGRID V. Fault analysis study of IDEA[OL]. https://securewww.esat.kuleuven.be/cosic/publications/article-1024, 2008.

欧庆于： 男，1978 年生，副教授，主要研究方向为密码芯片安全性分析.

罗 芳： 女，1983 年生，讲师，主要研究方向为密码编码理论.

叶伟伟： 男，1991 年生，硕士生，研究方向为密码芯片安全性分析.

周学广： 男，1966 年生，教授，博士生导师，主要研究方向为密码编码理论、内容安全.