Choisir une racine de confiance RoT combinant sécurité IoT maximale et coût abordable

Que vous soyez spécialiste de la sécurité IT ou IoT chez un équipementier, un fabricant de semi-conducteurs ou un intégrateur système, il est essentiel de connaître le rôle d'une « racine de confiance » (RoT, Root-of-Trust) – qui est la source de l'identité d'un objet connecté – et de comprendre comment différentes technologies apportent des niveaux distincts de sécurité. Cet article recense et compare les technologies RoT actuelles.

AUTEUR

Chris Jones, spécialiste de la sécurité IoT chez Crypto Quantique.

Des milliards d'objets connectés (ou IoT pour Internet of Things) sont déployés chaque année, et ils embarquent tous une certaine forme d'intelligence. On trouve en général dans chacun de ces objets connectés un microcontrôleur, une puce-système (SoC), un circuit ASIC ou d'autres types de semi-conducteurs. Ces produits sont majoritairement situés à la périphérie d'un réseau IoT et il s'agit le plus souvent de capteurs collectant des informations sur leur environnement, typiquement des données liées à la température, à l'humidité, aux vibrations, aux mouvements et à la localisation. Ces dispositifs se connectent à des réseaux locaux qui procèdent au transfert des données vers des applications et des services hébergés dans le cloud ou sur des serveurs installés sur site. Parmi les fournisseurs de services cloud, on peut citer AWS, Azure, Google, etc.

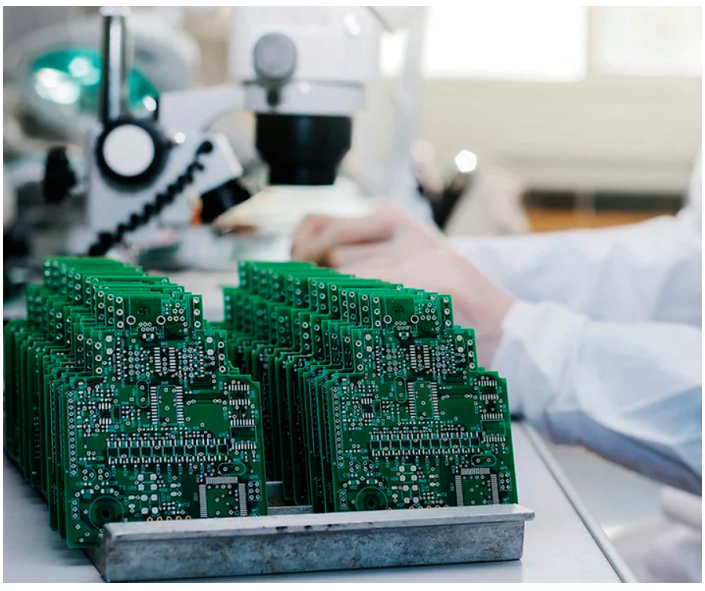
Le défi de la cybersécurité IoT

Chacun de ces équipements de périphérie (edge) offre aux pirates un accès potentiel aux réseaux.
La sécurité des objets connectés s'appuie sur celle instillée dans le microcontrôleur. De nombreux garde-fous doivent être intégrés dans ce dernier, tels qu'un démarrage sécurisé inaltérable, une résistance à l'altération des données (tamper) et des contremesures face aux attaques par canal auxiliaire (side-channel). Afin de communiquer en toute sécurité avec un appareil connecté, l'authentification dudit appareil, c'est-à-dire la preuve de son identité, constitue un facteur déterminant : son identité doit être unique. Une fois cette identité établie et prouvée, un lien de communication sécurisé peut être mis en œuvre. Cette liaison est chiffrée au moyen de clés cryptographiques.
L'identité de l'appareil et ses clés de chiffrement sont en pratique dérivées de nombres aléatoires appelées « graines d'entropie » (seeds). Et c'est donc un microcontrôleur sécurisé, une identité unique et des clés de chiffrement qui constituent une racine de confiance RoT (Root-of-Trust). Cette RoT, c'est la fondation de la sécurité au sein d'un réseau IoT.

L'identité de l'appareil et ses clés de chiffrement sont en pratique dérivées de nombres aléatoires appelées « graines d'entropie » (seeds). Et c'est donc un microcontrôleur sécurisé, une identité unique et des clés de chiffrement qui constituent une racine de confiance RoT (Root-of-Trust).

Comment un circuit intégré peut-il embarquer une RoT ?

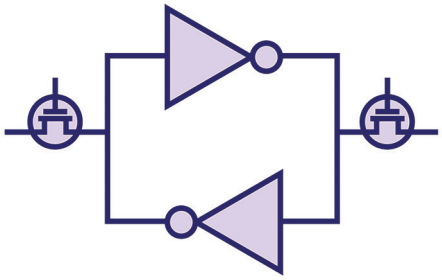
Il existe deux méthodes principales pour créer une RoT dans un microcontrôleur. La plus commune utilise un calculateur sécurisé externe appelé HSM (Hardware Security Module). C'est un calculateur conçu spécifiquement pour la génération de nombres aléatoires et de clés de chiffrement, ainsi qu'à la gestion de ces clés. Ces dernières sont donc créées en dehors du microcontrôleur et de l'objet connecté, puis programmées en son sein – on parle d'injection de clés – au travers d'une interface de programmation. Il existe potentiellement des problèmes de sécurité avec l'injection de clés du fait que cette interface n'est pas forcément chiffrée.



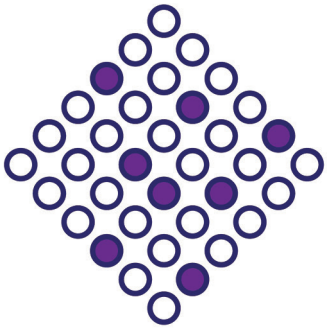
Avec l'autre méthode, le circuit génère lui-même des valeurs uniques et les convertit en clés de chiffrement. Classiquement, le microcontrôleur exploite des variations physiques aléatoires induites par le processus de fabrication de la puce pour générer des graines aléatoires. Ces variations de processus sont appelées fonctions physiques non clonables ou PUF (Physical Unclonable Functions). Les PUF génèrent des graines aléatoires qui peuvent être

1 PRINCIPE DE LA PUF DE TYPE SRAM

Lorsque les puces intégrant de la mémoire SRAM sont mises sous tension, chaque cellule SRAM prend un état correspondant à « zéro » ou « un », en fonction de petites variations physiques au niveau de la plaquette de silicium. Ces variations sont utilisées pour créer les graines qui permettent de générer des clés de chiffrement.



SRAM Cell



PUF of random 0s and 1s

converties en identités et clés de chiffrement par un générateur de clés, une fonction périphérique souvent intégrée dans le microcontrôleur.

L'injection de clés en détail

L'injection de clés peut s'avérer relativement coûteuse car elle nécessite un équipement de programmation spécifique. En général, celui-ci est fourni par une société de programmation spécialisée employant des programmateurs étroitement liés aux HSM. Mais, comme il s'agit d'une société tierce partie, cette option introduit une faille potentielle de sécurité, et cette approche va à l'encontre des dernières recommandations des experts en sécurité qui préconisent d'adopter une approche à risque zéro en évitant toute implication de tierce partie.

Les clés injectées doivent être stockées en mémoire à l'intérieur du composant. Elles sont généralement conservées dans une mémoire non volatile, puis protégées par une technologie matérielle de sécurisation au sein du microcontrôleur. Par exemple, la technologie TrustZone d'Arm scinde l'environnement d'exécution en mémoires, périphériques et fonctions, sécurisés et non sécurisés. Mais, même avec de telles mesures en place, les clés demeurent vulnérables et peuvent être lues par des individus mal intentionnés puisqu'elles sont juste stockées dans la mémoire flash standard du circuit.

Autre vulnérabilité : les clés sont souvent transférées vers les circuits au moyen d'une liaison non chiffrée, en l'occurrence la liaison entre le HSM et le programmateur ou, plus couramment encore, entre le programmateur et le circuit lui-même, ce qui les exposent à des attaques.

Comment les PUF éliminent une partie des risques de sécurité

Observons les PUF plus en détail. La PUF SRAM est un bon exemple de technologie PUF de première génération. Une mémoire SRAM est intégrée dans la plupart des microcontrôleurs et microprocesseurs. Quand ces puces sont mises sous tension, chaque cellule SRAM prend un état correspondant à « zéro » ou « un », en fonction de petites variations physiques au niveau de la plaquette de silicium (figure 1). Ces variations aléatoires sont utilisées pour créer les graines qui permettent de générer des clés de chiffrement. La mémoire SRAM du circuit devient ainsi son empreinte digitale et lui octroie une identité unique. Comme les PUF SRAM exploitent une technologie mémoire déjà présente dans le microcontrôleur, il suffit de routines logicielles pour les gérer.

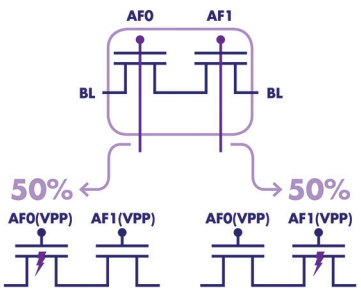
La mémoire flash constitue une autre sorte de PUF. La mémoire flash est elle aussi disponible sur la majeure partie des microcontrôleurs. Les cellules flash sont « programmées » en les stressant jusqu'à causer des ruptures dans la couche d'isolation en oxyde de silicium des grilles des transistors de chaque cellule mémoire. En raison des différences intrinsèques entre les deux transistors présents dans chaque cellule flash, cela donne soit un « zéro », soit un « un » (figure 2). Dans cette technologie, la rupture de l'oxyde de grille nécessite une tension élevée et donc une phase de programmation initiale ; mais, une fois ce programme terminé, chaque cellule flash contient une donnée aléatoire, qui peut être lue. Tout cela requiert donc une certaine préparation, mais exploite, comme avec la SRAM, une technologie qui est déjà présente dans le microcontrôleur ou le microprocesseur.

Avantages et inconvénients des technologies PUF de 1re génération

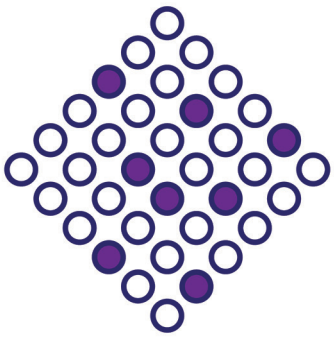
Prenons d'abord le cas de la variante SRAM. L'un de ses principaux avantages réside dans le fait qu'il n'est pas nécessaire d'injecter des clés dans le microcontrôleur. Les graines à l'origine des clés sont créées par la mémoire SRAM elle-même, qui est déjà présente dans la puce. Ces clés ne sont pas stockées en mémoire

2 PRINCIPE DE LA PUF DE TYPE FLASH

Les cellules flash sont « programmées » en les stressant jusqu'à causer des ruptures dans la couche d'isolation en oxyde de silicium des grilles des transistors de chaque cellule mémoire. En raison des différences intrinsèques entre les deux transistors présents dans chaque cellule flash, cela donne soit un « zéro », soit un « un ».



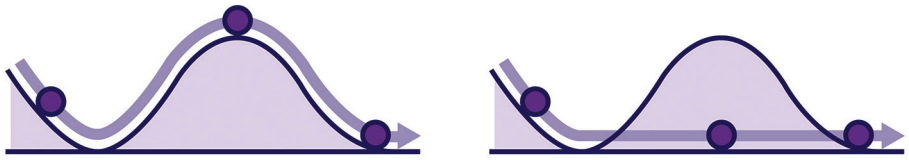
Flash Memory Cell



PUF of random 0s and 1s

3 QDID, UNE TECHNOLOGIE PUF QUI UTILISE L'EFFET TUNNEL QUANTIQUE

La technologie QDID de la société Crypto Quantique s'appuie sur les positions atomiques et les imperfections des nanostructures de la couche d'oxyde de silicium des transistors CMOS d'un réseau de 64 x 64 cellules intégré dans un composant CMOS. Le bloc de traitement QDID associé mesure le courant d'effet tunnel quantique à travers la couche d'oxyde de chaque paire de transistors.



mais dans la structure physique des cellules SRAM : cela rend la puce difficile à pirater.

Plusieurs fabricants de semi-conducteurs comme Intel, Microsemi (désormais intégré dans Microchip), NXP et Xilinx ont recours à la technologie PUF SRAM. Cette technologie a toutefois ses limites, l'une d'entre elles étant qu'elle ne peut généralement créer qu'une seule graine. Si l'on désire obtenir plusieurs clés de chiffrement, il faut les générer à partir de cette graine commune : elles seront de ce fait mathématiquement corrélées, et donc intrinsèquement moins sécurisées que si ce lien entre elles n'existait pas.

Autre faille potentielle : les cellules ne démarrent pas toujours dans le même état. Cela implique qu'une correction d'erreurs est nécessaire pour s'assurer que la graine créée à partir des cellules est stable et reproductible. Le niveau de répétabilité dépend de la fabrication de cette mémoire spécifique, et parfois l'entropie (c'est-à-dire la propension à être aléatoire) d'une PUF de type SRAM peut être assez faible. Comme l'identité figure à l'intérieur de la cellule SRAM, la résistance aux attaques peut être remise en cause : en observant les flux de courant électrique ou d'autres phénomènes électriques, une attaque physique par canal auxiliaire peut être mise en œuvre pour lire l'état de chaque cellule.

A la mise sous tension, les PUF SRAM pâtissent en outre d'un délai de configuration relativement long, durant lequel le microcontrôleur et l'appareil IoT qu'il pilote sont susceptibles d'être attaqués.

Passons maintenant aux PUF à mémoire flash. Ici aussi, il n'est pas nécessaire d'injecter des clés dans le circuit, et les graines sont créées dans la mémoire flash déjà présente au sein du microcontrôleur. Lorsque les graines ont été programmées dans la mémoire, elles peuvent être extraites avec une faible latence au moyen d'une simple instruction de lecture. Ce procédé ne réclame pas de correction d'erreurs car l'état de chaque cellule flash ne varie pas après sa programmation.

Bien sûr, le stockage des graines en mémoire peut les rendre sujettes aux attaques. Autre inconvénient des PUF à mémoire flash, la pompe de charge, nécessaire pour obtenir les fortes tensions associées à la rupture de la couche d'oxyde du semi-conducteur, occupe une surface de silicium additionnelle.

A l'instar des PUF SRAM, les PUF flash n'ont pas recours à un bloc matériel de sécurité spécifique et indépendant. Et, quand la mémoire flash est réaffectée à la génération de clés, elle n'est pas disponible pour d'autres tâches. De plus, le délai de configuration est là encore assez long puisqu'il faut programmer la mémoire pour occasionner la rupture de la couche d'oxyde.

Comment les PUF de 2e génération améliorent la sécurité IoT

Les PUF de seconde génération sont des blocs IP silicium spécialement conçus pour la sécurisation des données et optimisés pour les processus de fabrications CMOS standard. Le bloc IP de Crypto Quantique baptisé QDID (pour Quantum-Driven IDentity) en est un exemple.

QDID[1] est un réseau de 64 x 64 cellules intégré dans un composant CMOS et d'où une empreinte digitale est extraite. Le caractère aléatoire de cette empreinte s'appuie sur les positions atomiques et les imperfections des nanostructures de la couche d'oxyde de silicium des transistors CMOS du réseau. Le bloc de traitement QDID mesure le courant d'effet tunnel quantique à travers la couche d'oxyde de chaque paire de transistors. De la nature aléatoire et probabiliste de ce courant découlent les états binaires de l'empreinte digitale (figure 3).

QDID mesure des courants de l'ordre du femtoampère. Les variations de ces courants génèrent des nombres aléatoires utilisés pour produire des identités de circuits uniques, inaltérables et non clonables, à partir desquelles peuvent également être créées des clés de chiffrement à la demande. Ces PUF de seconde génération apportent le plus haut niveau de sécurité possible. Tout d'abord, elles disposent d'une entropie élevée car elles proviennent de la mesure d'un effet quantique probabiliste. Ensuite, elles produisent de multiples clés non corrélées au moyen du réseau de 64 x 64 cellules, et ces clés sont créées à la demande et n'ont donc pas à être stockées en mémoire où elles seraient susceptibles de fuiter.

Des tests indépendants de QDID montrent que cette technologie est sécurisée contre toutes les méthodes d'attaque connues et qu'elle est conforme à la certification de sécurité EAL4[2]. Elle est également pré-certifiée PSA Level 2[3].

Comme leurs variantes SRAM et flash, les PUF de seconde génération éliminent bien évidemment le besoin d'injecter des clés et les risques de sécurité liés à ce procédé. QDID présente d'autres avantages, à savoir une surface de silicium réduite afin de limiter les coûts au minimum, et un faible taux d'erreur facilement compensé par un petit algorithme de correction d'erreurs ne mobilisant que peu de ressources du processeur. Des prototypes fonctionnels de test sont disponibles dès à présent, et plusieurs grands fabricants de semi-conducteurs s'apprêtent à annoncer dans les mois à venir qu'ils font de QDID leur technologie de RoT[4]. ■

(1) https://www.cryptoquantique.com/products/qdid/

(2) Https://www.cryptoquantique.com/press/quantum-tunnelling-semiconductor-ip-verified-as-secure-against-all-known-iot-attacks/

(3) https://www.psacertified.org/getting-certified/silicon-vendor/overview/level-2/

(4) L'auteur a créé une vidéo explicitant plus avant les RoT qui peut être visualisée sur YouTube : https://www.youtube.com/watch?v=vZn2FU5Mn8Y&t=17s