CACHEOMS: Practical Exploitation of **Cache** Side Channel **O**n **M**ultiprocessor **S**ystems

Aravind Machiry & Varun Kulkarni Somashekhar

Abstract— **Cache side channel is well known attack on cryptographic implementations[?][?][?]. This is primarily based on observation that infrequently used information incurs a large timing penalty, thus revealing some information about the frequency of use of the memory blocks. This attack is straight forward on a single processor system with single level of cache, but current systems (both desktop and mobile) are multiprocessor where each processor has multilevel caches also there are some cache levels which are common to a subset of processors. Moreover, latest ARM processors have TrustZone support[?] which ensures System-On-Chip (SOC) wide security by avoiding shared cache access. In addition to this, there has been lot of work done both on Hardware and Software side to mitigate cache side channel[?][?]. It is high time we revisit and evaluate cache side channel, its bandwidth and implications[?].**

We intend to discuss the attack and defense strategies involved in a cache side channel, specifically following are our goals:

- **Investigate, understand and report cache side channel with examples. We aim to make the documentation simple enough to be understood by a computer science graduate without much knowledge of cryptography or computer architecture.**
- **Research on the recent improvements and state of art on cache side channel and clearly report the findings.**
- **Evaluate the applicability of these attacks on well known protocols on current multiprocessor systems, both on x86 and ARM.**

I. Introduction

The basic arithmetic operations (i.e. addition, multiplication, and inversion) in prime and binary extension fields, $GF(p)$ and $GF(2^n)$, have several applications in cryptography, such as decipherment operation of RSA algorithm [**?**], Diffie-Hellman key exchange algorithm [**?**], the Government Digital Signature Standard [**?**] and also elliptic curve cryptography [**?**], [**?**]. Recently, speeding up inversion operation in both fields has been gaining some attention since inversion is the most time consuming operation in elliptic curve cryptographic algorithms when affine coordinates are selected [**?**], [**?**], [**?**], [**?**], [**?**].

In this paper, we will give and analyze multiplicative inversion algorithms for $GF(p)$ and $GF(2^n)$ which allow very fast and area-efficient, unified and scalable hardware implementations. The algorithms are based on the Montgomery inverse algorithms given in [**?**].

II. Background

In this section we give a brief overview of the background needed to understand the attack presented in this work. After summarizing the design of cache architecture, a short explaination about different types of cache attacks are provided.

A. Cache Design and Operation

A cache is a small memory placed between the CPU and RAM to reduce the big latency added by retreival of data. Modern processors usually have more than one level of cache to improve the efficiency of memory accesss.Most modern processes offer 3 levels of cache.with the first level (L1) being the closest to the CPU registers. Generally, L1 and L2 caches are each split into instruction caches and separate data caches. L3 usually stores both instructions and data.The cache size is much smaller than the number of directly addressable bytes in main memory. Hence a mapping strategy needs to be adopted.The cache associativity determines how the main memory blocks map into blocks of the cache.

When a memory reference is made by the CPU, the tag address of the main memory block is compared with all the tags in the corresponding set. If the tag is found then this reference qualifies as a cache hit. Meaning that there is no need to retrieve the data from main memory since it is already located in the cache, and the data can be immediately provided to the CPU. Conversely, if the tag is not found in the corresponding set then the memory reference qualifies as a cache miss.The caching mechanism operation is based on the principals of spatial and temporal locality, which help to minimize the number of cache misses. Temporal locality states that the same data blocks will likely be requested repeatedly during the execution of a process. Spacial locality states that data blocks from nearby addresses are likely to be subsequently accessed. Even though the number of cache misses is reduced by these principles, they are not eliminated.

Obviously, data in the cache can be accessed much faster than data present only in memory.This is also true for multi-level caches where data accessed from the L1 cache will experience lower latencies, than data accessed from subsequent cache levels. These time differences are used to decide whether a specific portion of memory resides in the cache - implying that the corresponding data has been accessed recently. This resulting information leakage stemming from microarchitectural time differences when data is retrieved from cache rather than memory forms the basis of cache side channel attacks.

Authors are with the Department of Computer Science, University of California, Santa Barbara, CA 93106. E-mail: {you,me}@cs.ucsb.edu

This work is supported by Motorola.