PHEP: Paillier Homomorphic Encryption Processors for Privacy-Preserving Applications in Cloud Computing

Guiming Shi[1], Yi Li[2], Xueqiang Wang[2], Zhanhong Tan[1], Dapeng Cao[3],
Jingwei Cai[1], Yuchen Wei[1], Zehua Li[3], Yifu Wu[4], Wuke Zhang[4],
Wei Xu[1*], and Kaisheng Ma[1*]

[1]Tsinghua University
[2]HuaKong TsingJiao
[3]Xi'an JiaoTong University
[4]Polar Bear Tech

*Corresponding Authors

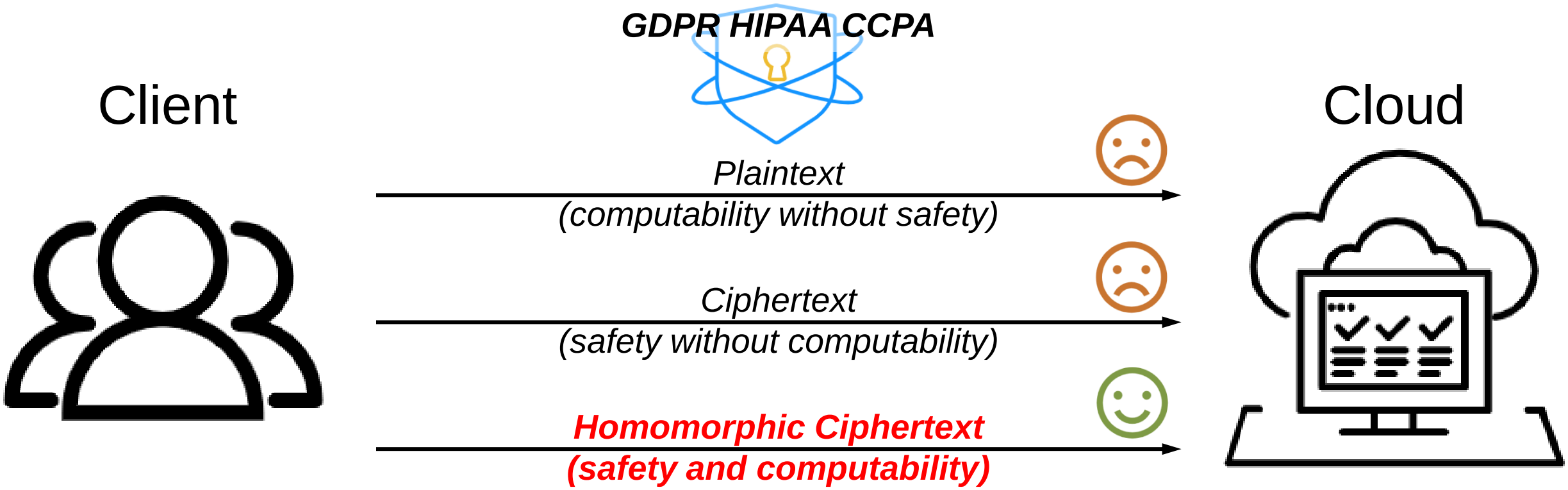
HOT CHIPS

Extended Abstract

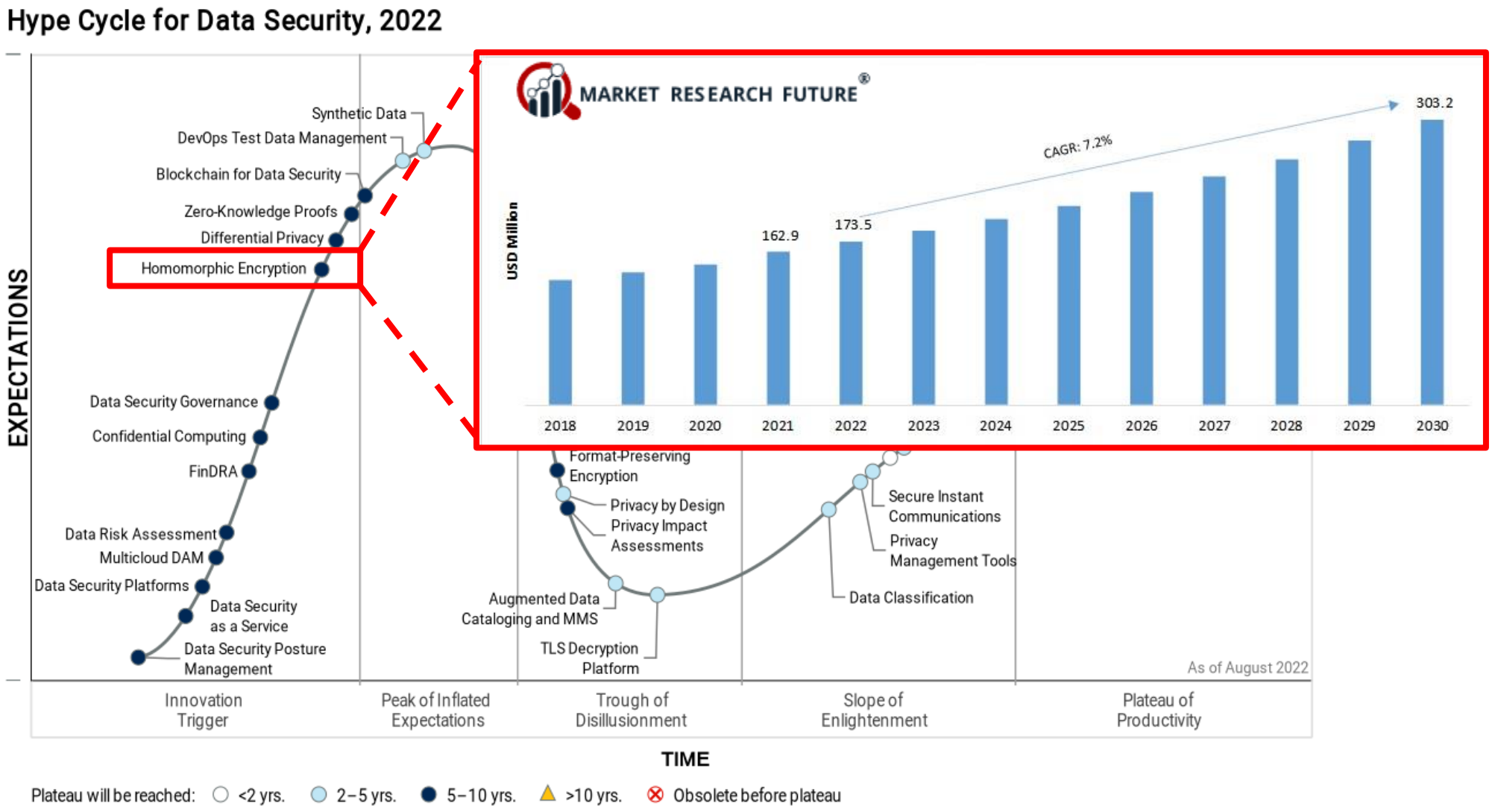
- Cloud computing has evolved into the key infrastructure of emerging applications, storing massive amounts of data. Yet, how to safely handle this sensitive data in a shared cloud is a major concern. **Paillier homomorphic encryption is an important privacy protection approach** that permits arithmetic operations on ciphertext without first decrypting it, offering a viable solution to the privacy dilemma.
- The Paillier approach has a significant computational overhead compared to plaintext computation because computing in the **ciphertext domain requires expensive large integer modular operations that are inefficient for CPUs**. As a result, it is preferable to create **domain-specific processors** for Paillier. **Paillier computing patterns are divided into two types, both of which are extensively employed in Paillier applications: independent vector operations and multiply-and-accumulate (MAC) operations.** The former is primarily employed in applications such as private information retrieval and on the client side for privacy-preserving AI. In contrast, the latter is required for cloud-side AI inference, particularly computing convolution in neural networks.
- **We introduce PHEP: Paillier Homomorphic Encryption Processors for cloud-based privacy-preserving applications. PHEP is built on two Paillier acceleration chips: Paillier engine-1 and Paillier engine-2, both produced on the same wafer. Paillier engine-1 focuses on vector operations** and attempts to increase computation as much as feasible. It contains 80 processing elements (PE) and can provide 480 TOPS (INT8) for a 16-chip Full-Height-Full-Length (FHFL) PCIe card. **Paillier engine-2 is designed for MAC operations** and has 16 high-performance bit-serial sparse PEs. It only has 192 TOPS (INT8) for an 8-chip FHFL PCIe board. However, it is specialized for matrix operations like convolutions. Both engine chips have the same hardware interface, allowing them to use the same PCB board, FPGA scheduler, and software framework design. The PHEP accelerator card also contains a host FPGA. The host FPGA schedules both data transfers and computation among these engine chips. To manage these engines, we use a complex software stack. The software stack includes an offline compiler and an online task scheduler for automatically balancing compute workload across multiple cards on the same server and even across multiple servers. The findings of the end-to-end evaluation reveal that PHEP can perform Paillier-based machine learning workloads 1-2 orders of magnitude faster than state-of-the-art CPUs (Intel Xeon Platinum 8260M with 192 cores), making these privacy-preserving applications practical.

Homomorphic Encryption in Cloud Computing

- Data privacy is a critical problem in Cloud Computing.
- Paillier Homomorphic Encryption can protect the privacy of the data and enable computing on the ciphertext without decryption first.



Hype Cycle for Data Security: Developing Markets for Homomorphic Encryption



Gartner

Paillier has Vast Applications in Different Sectors

Federated Learning

FATE

PFL
Paddle Federated Learning

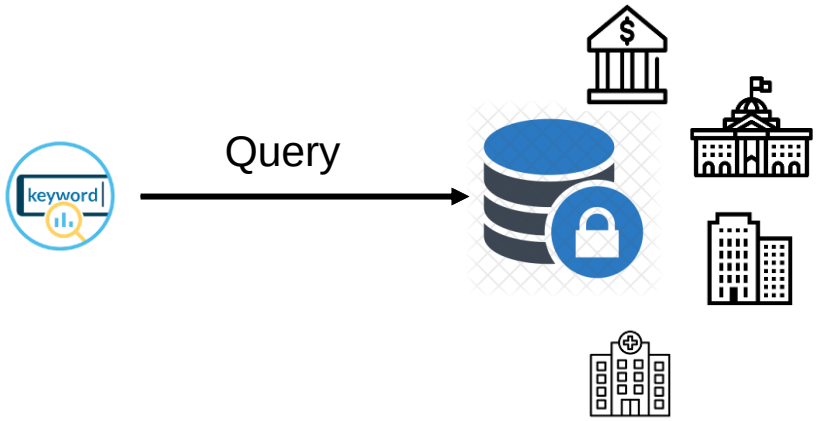
OpenFed

FedML

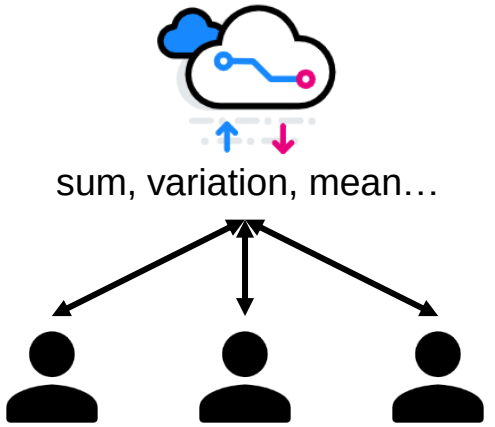
LEAF

Homomorphic Commitment

Privacy-Preserving Query



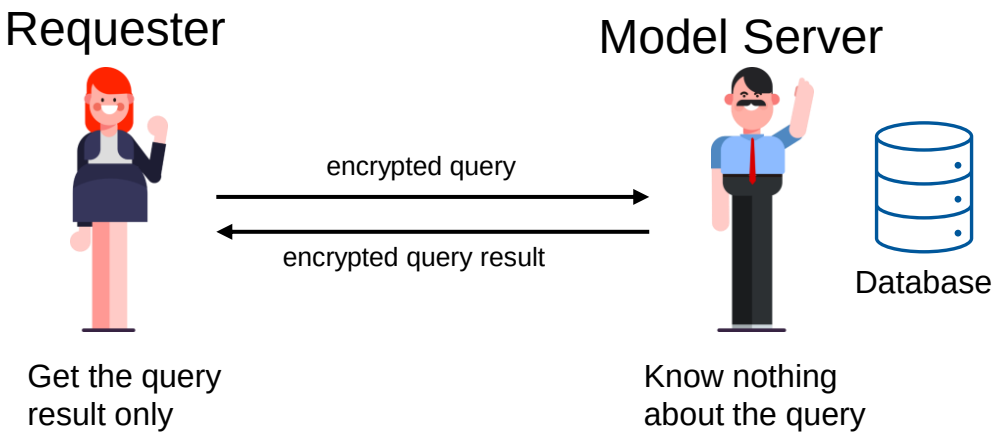
Collaborative Statistics



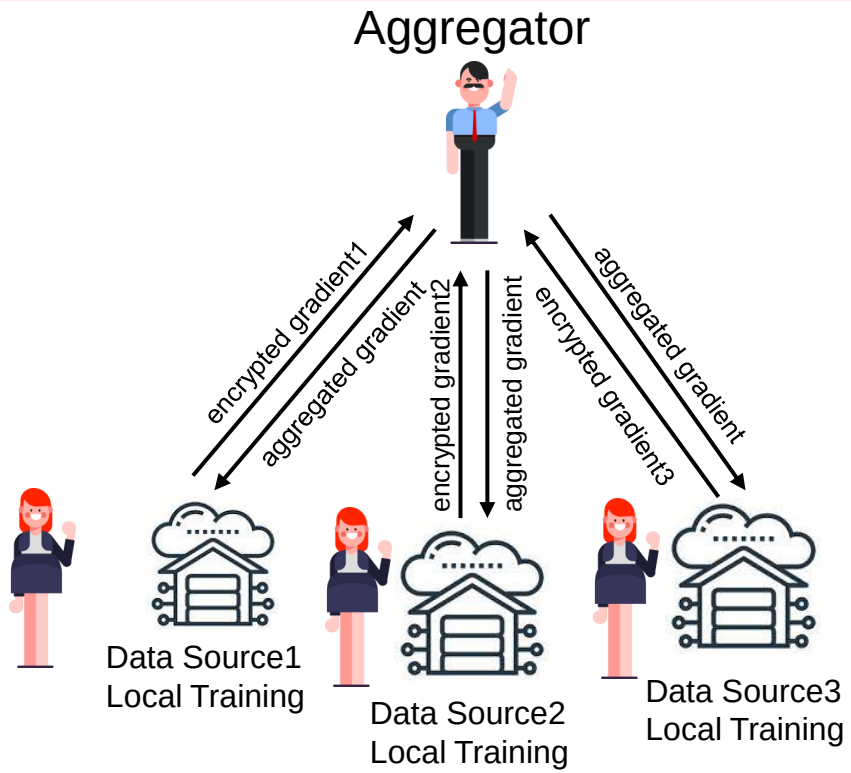
Electronic Signature

Typical Application Building Blocks for Paillier

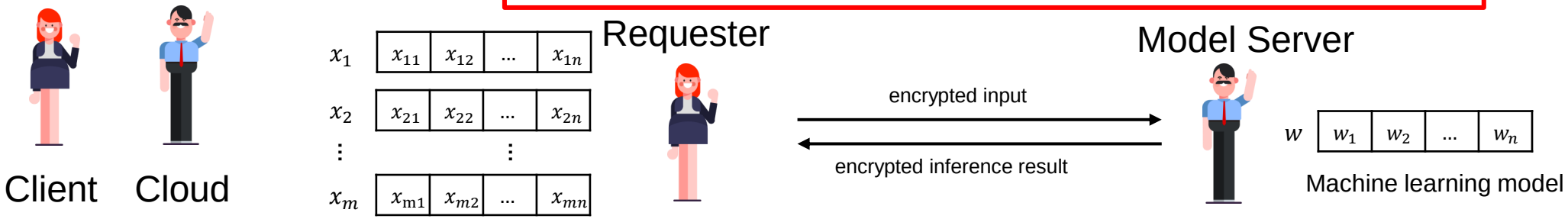
Private Information Retrieval



Privacy-Preserving Machine Learning Training

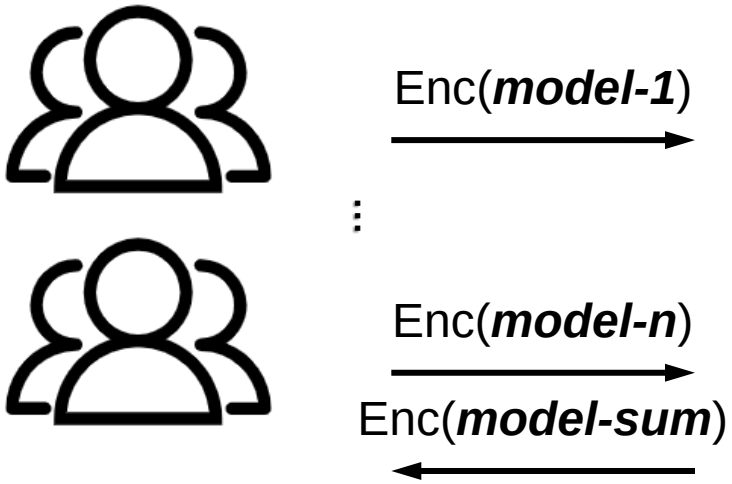


Privacy-Preserving Machine Learning Inference

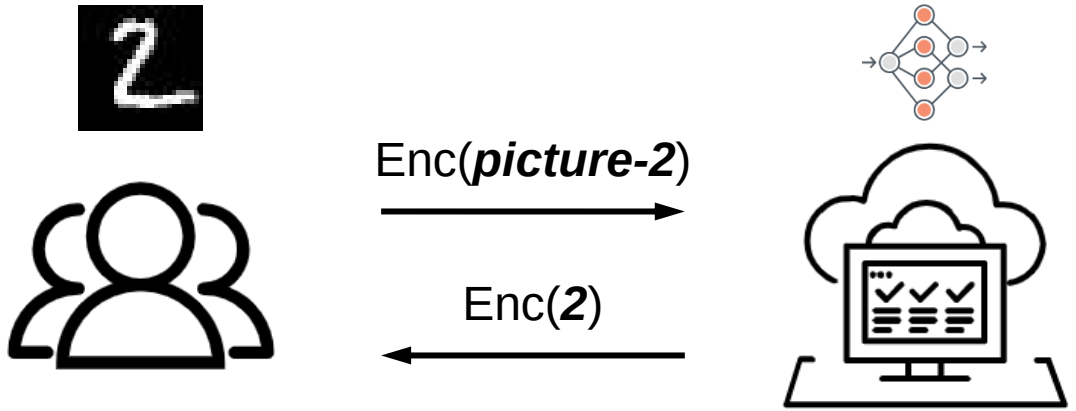


Bottleneck in Training and Inference Applications are Different

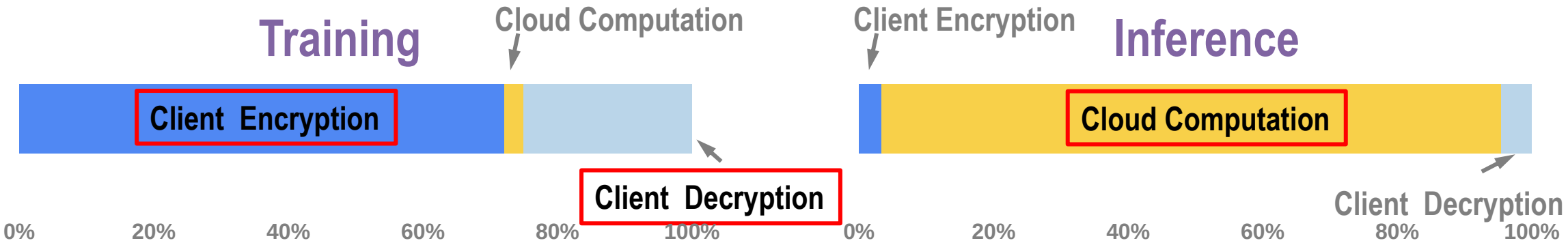
Application-Training, Bottleneck:
Client Encryption and Decryption



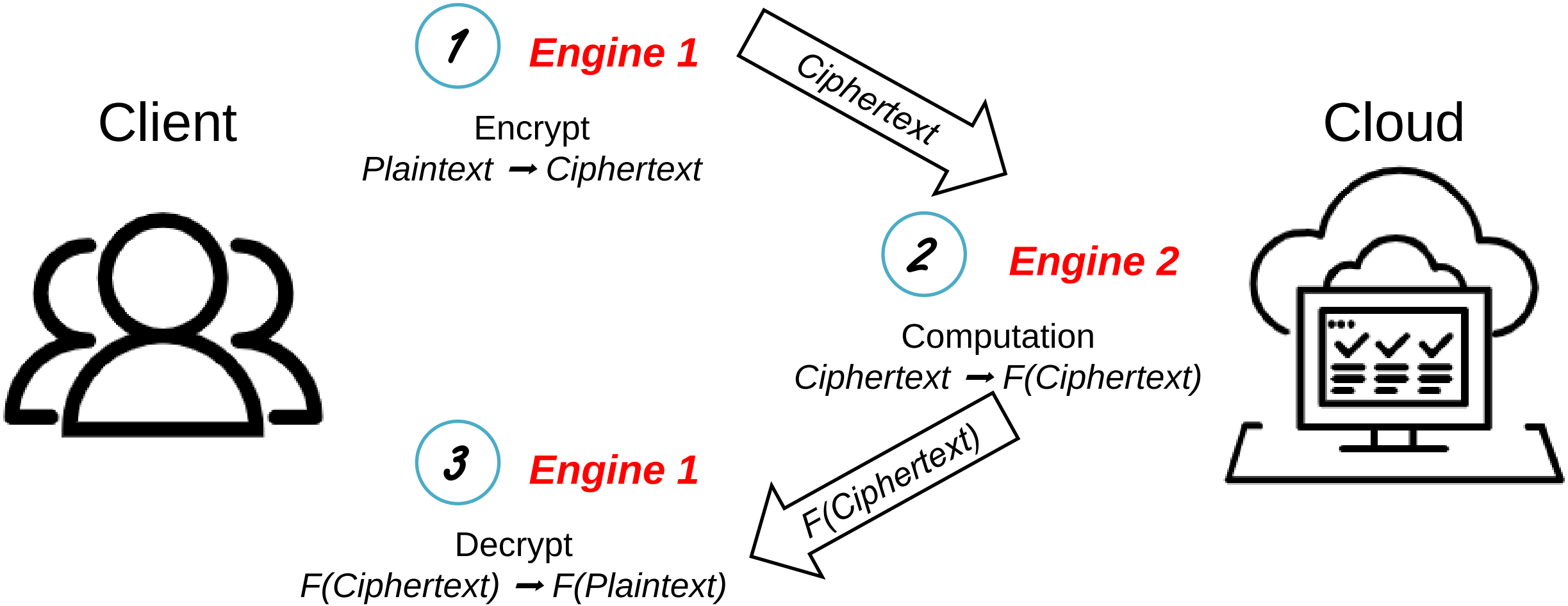
Application-Inference, Bottleneck:
Cloud Computation



Latency Breakdown On CPU

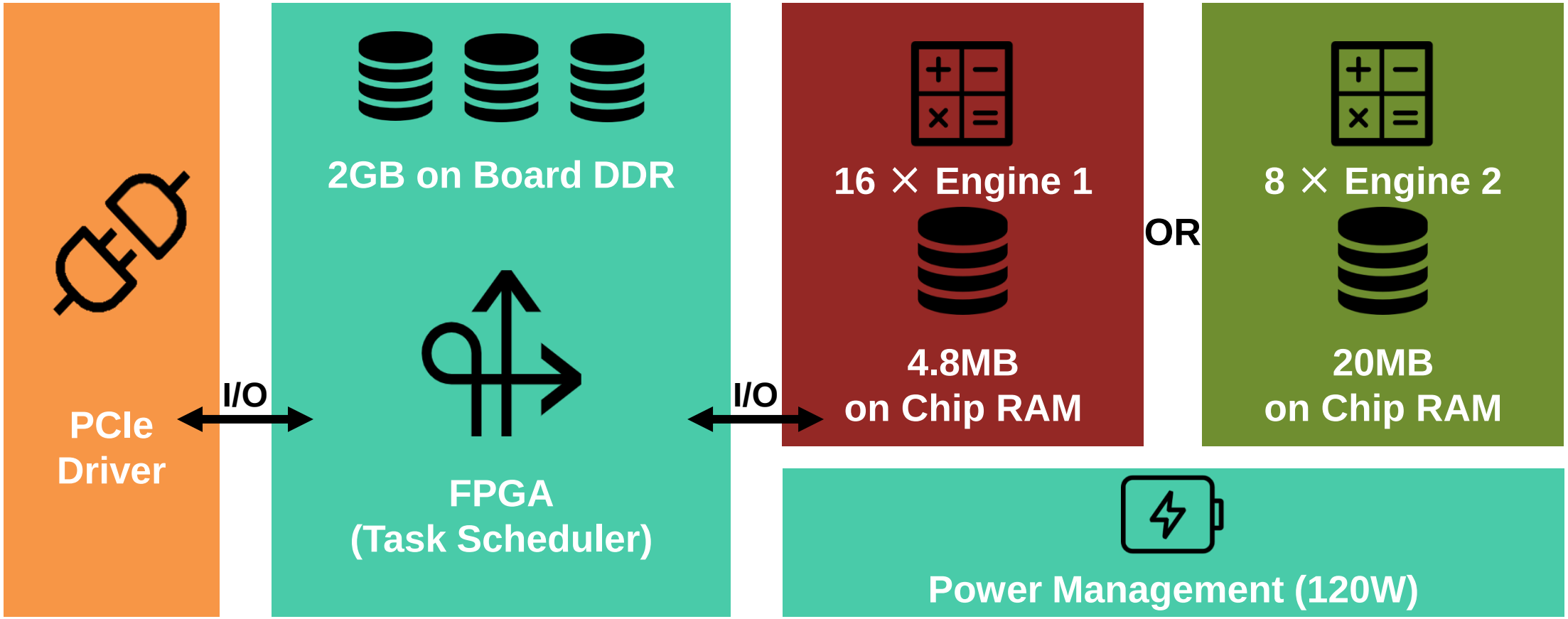


Our Solution: Build 2 Chips with the Same Hardware Interface to Meet the Requirement of Different Scenes



The PHEP Hardware

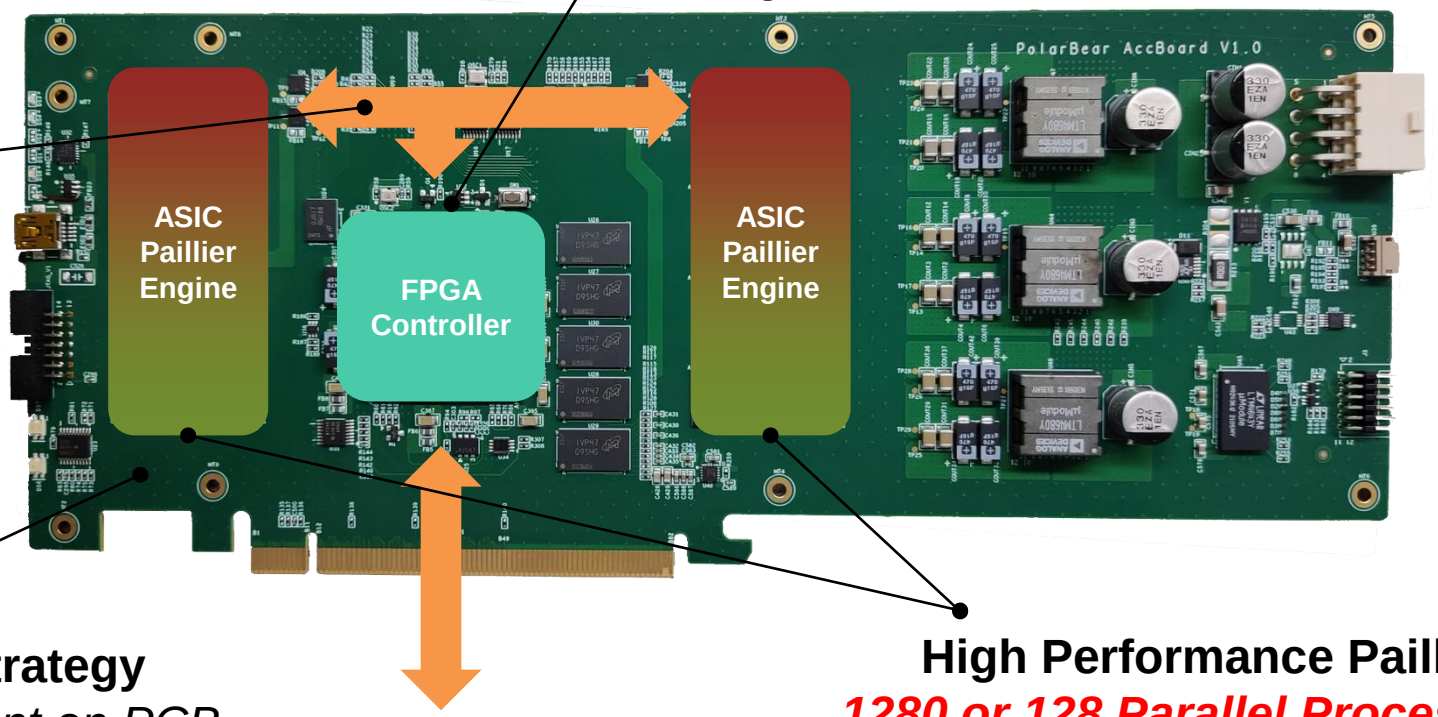
PHEP Accelerate Board: Supporting both Engine Chips



PHEP Accelerate Board Overview

Task Scheduler
Assign Tasks and Transfer data to Paillier Engine Chips

SPI
Communication between Chips and FPGA
Unicast or multicast



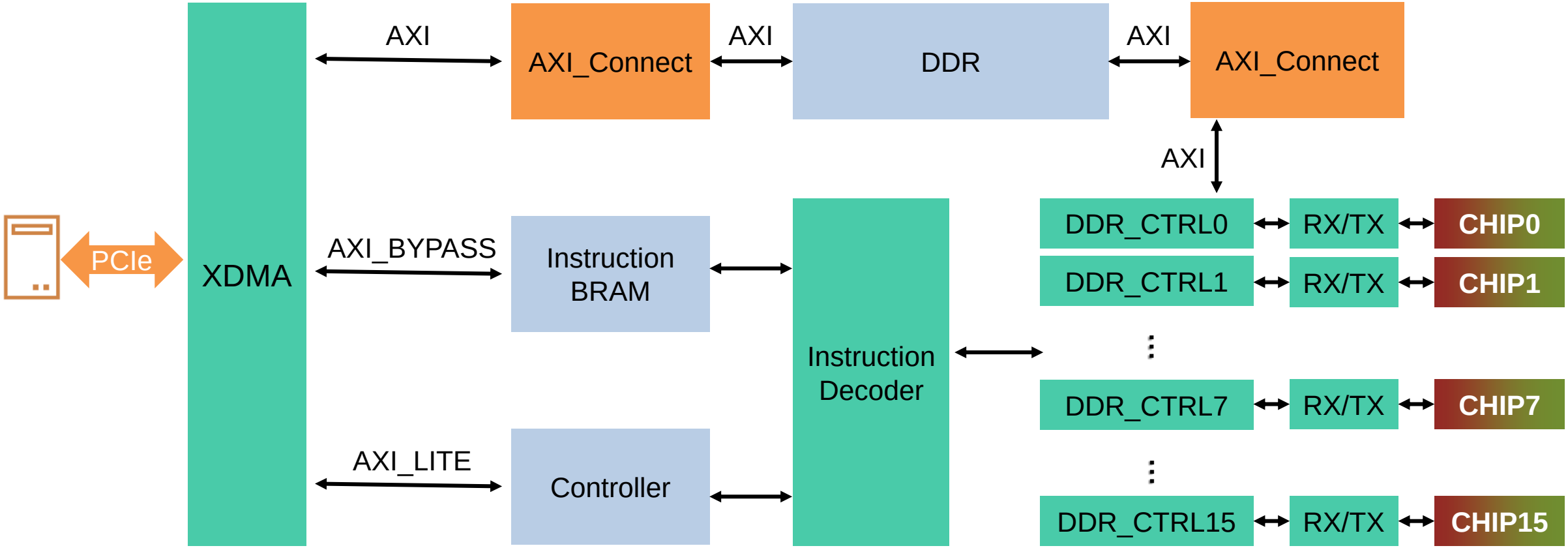
Power Management Strategy
Dynamic Power Management on PCB and Dynamic Clock Gating in Engines

High Performance Paillier Engines
1280 or 128 Parallel Processing Elements in 16 Engine-1 or 8 Engine-2

PCIe Device Driver
Host Communication

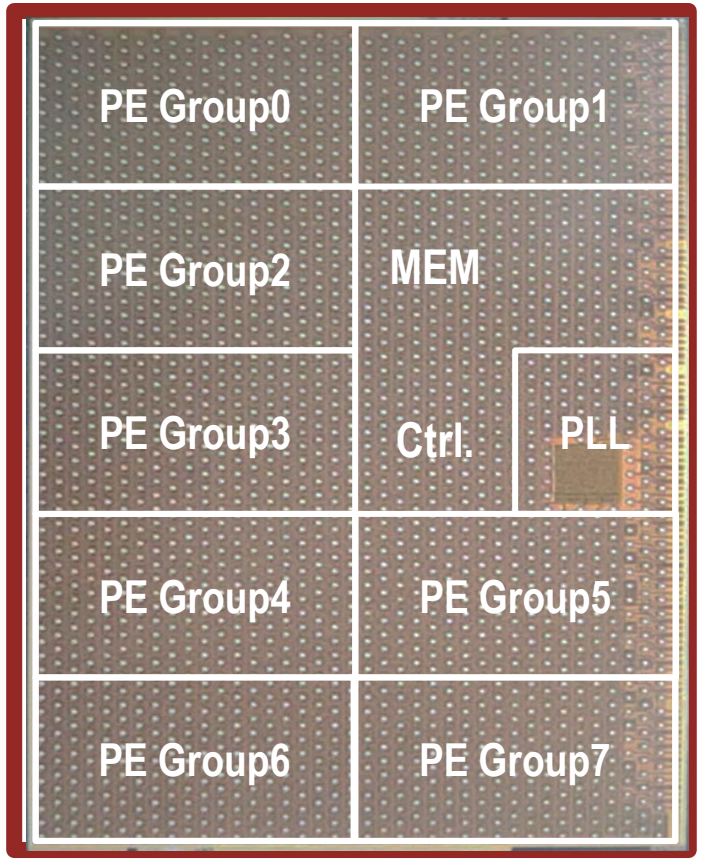
FPGA-Based Scheduler and I/O Controller

Scheduling 16 or 8 Chips and Transferring the Data between Chips and DDR

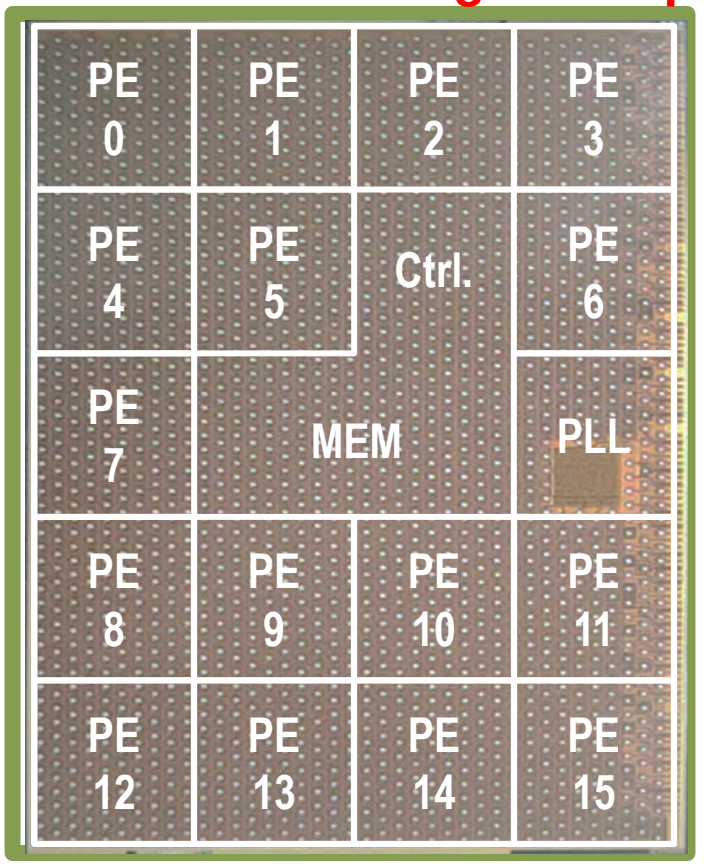


Comparison of the Two Engines: Specification

Fabricated on the Same Wafer: Significantly Reduces NRE of the Engine Chips



Same Area：43mm²
@ UMC 28nm HPC+



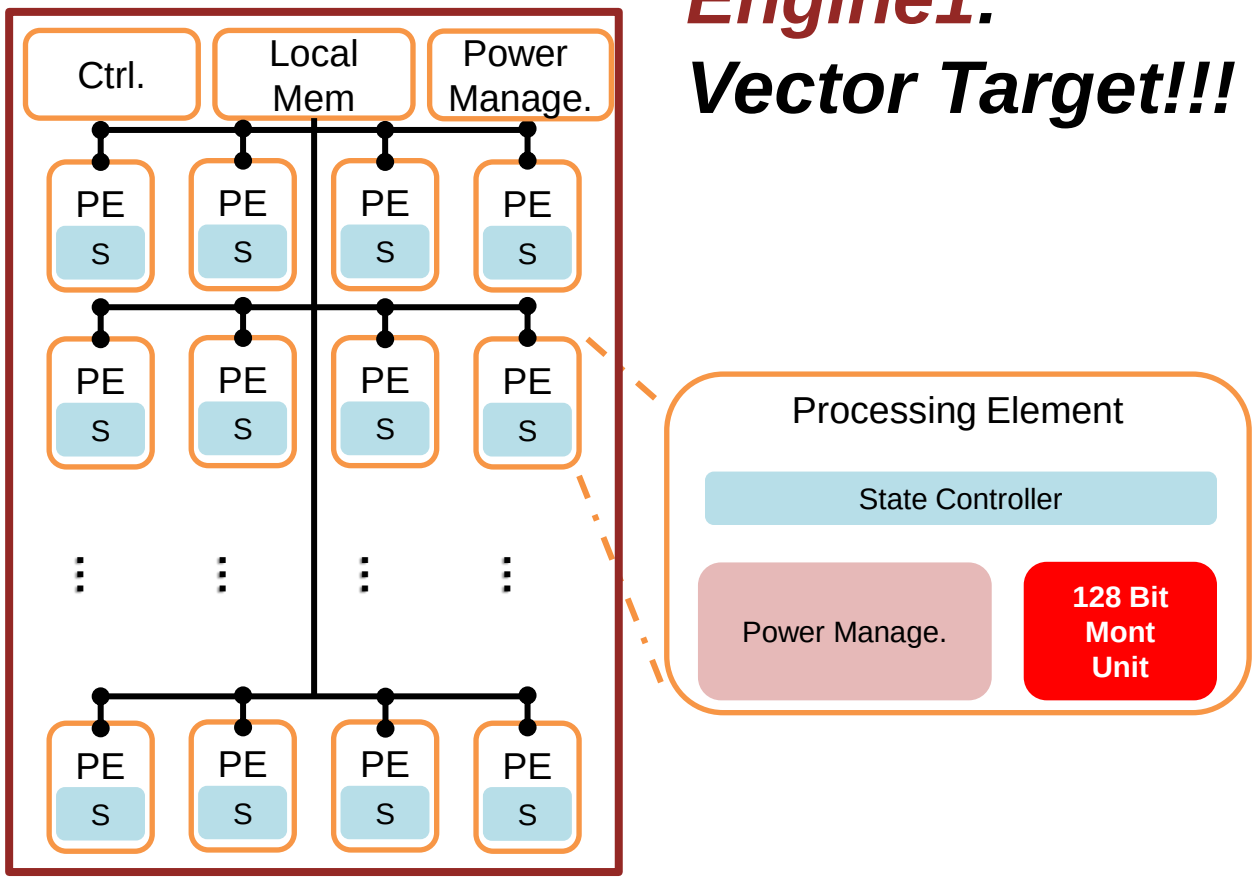
- **Optimized for Parallelism**

Items/PE	Montgomery Unit
Algorithm	Montgomery Multiplication
Arithmetic	3*128 Bit Multiplier

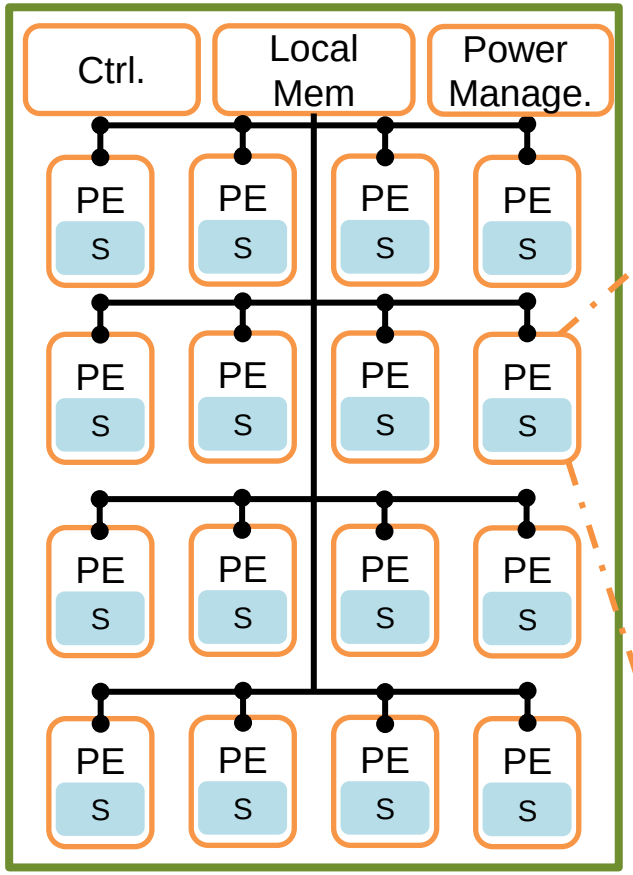
- **Optimized for Performance**

Items/PE	Montgomery Unit	Stein Unit
Algorithm	Montgomery Multiplication	Stein Modular Inversion
Arithmetic	3*256 Bit Multiplier	3*4102 Bit Adder

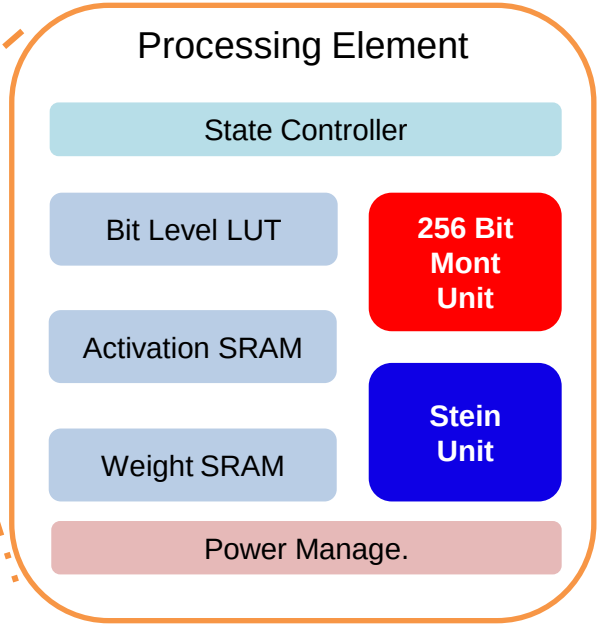
Comparison of the Two Engines: Architecture



Engine1:
Vector Target!!!



Engine2:
MAC Target!!!

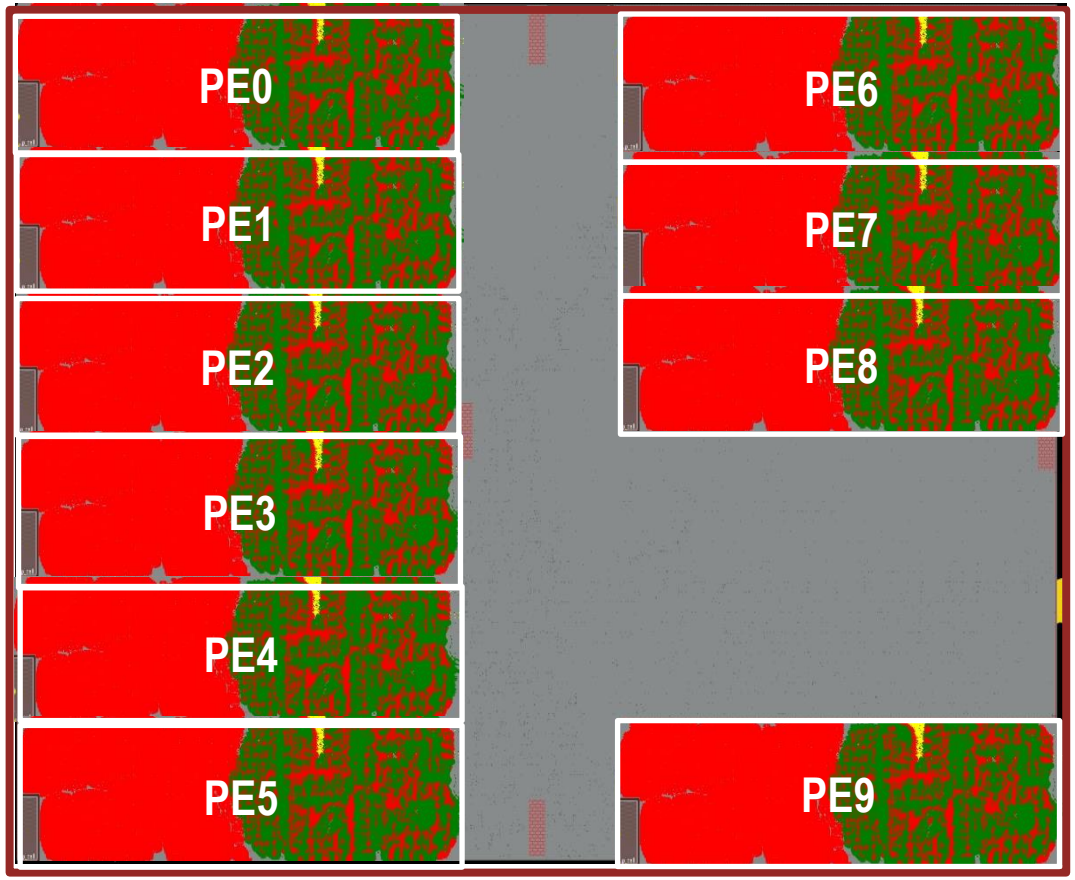


***80** Parallel Processing Elements*
400KB SRAM

***16** High-Performance Processing Elements*
2.5MB SRAM

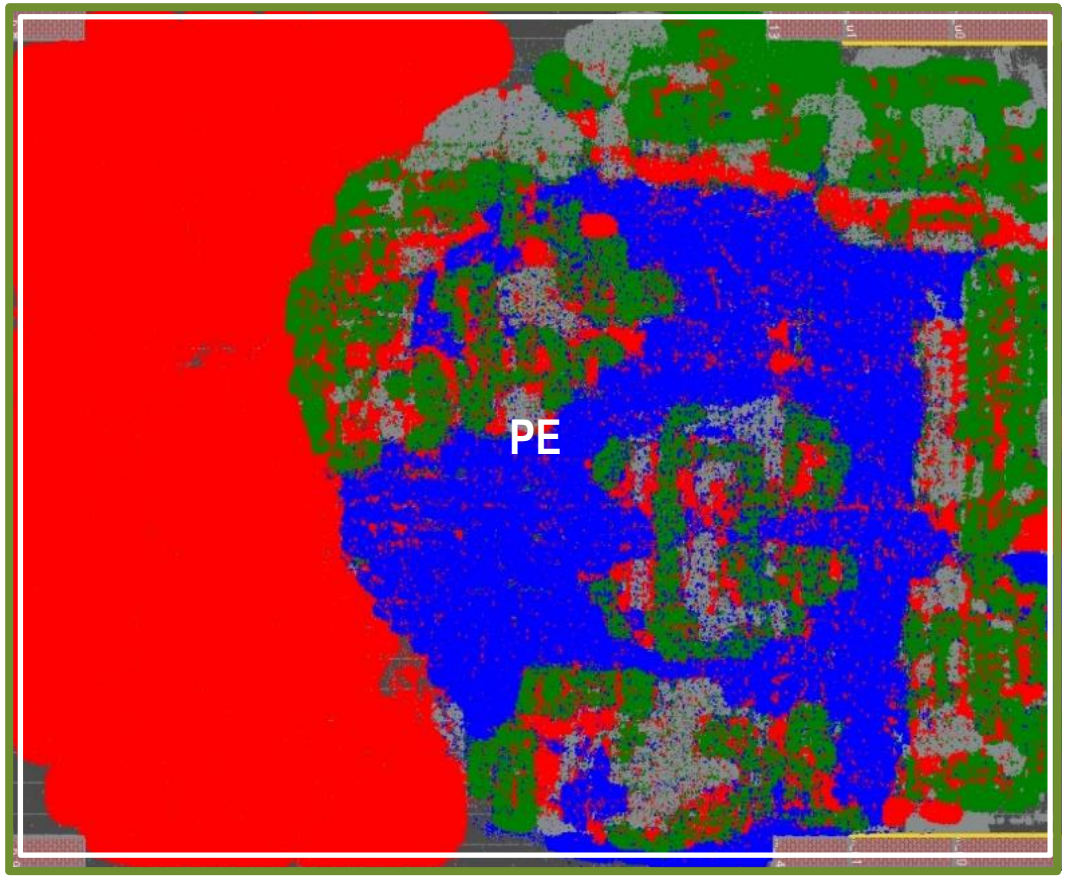
Comparison of the Two Engines: Physical Design

***30 Parallel** 128 Bit-Multiplier*
in one Harden Block @ 500MHz, 0.9V



Montgomery Unit
Stein Modular Inversion Unit
Register

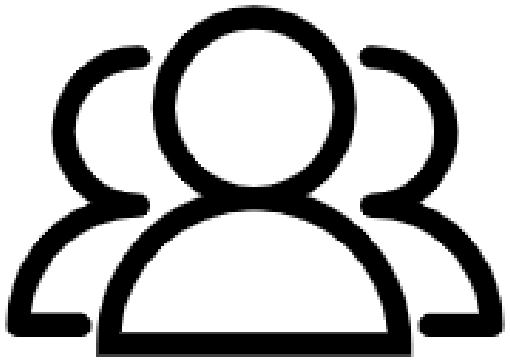
***3 Parallel** 256 Bit-Multiplier and 4102 Bit Adder*
in one Harden Block @ 500MHz, 0.99V



Paillier Engine-1 Vector Computation Routine

- Dataflow optimized for the **Client** and the Cloud sides.

Client



Vector Operation

$$ct = Encrypt(pt) = g^{pt} \cdot r^n \bmod n^2$$

1 $temp_1 = g^{pt} \bmod n^2$ — *Computation Intensive*

2 $temp_2 = r^n \bmod n^2$ — *Computation Intensive*

3 $ct = temp_1 \cdot temp_2 \bmod n^2$ — *Communication Intensive*

PHEP Engine

$$pt = Decrypt(ct) = L\left(ct^{\lambda} \bmod n^2\right) \cdot \mu \bmod n$$

1 $temp_1 = ct^{\lambda} \bmod n^2$ — *Computation Intensive*

2 $temp_1 = L(temp_1)$ — *ASIC Inefficient Function*

3 $ct = temp_1 \cdot \mu \bmod n$ — *Communication Intensive*

PHEP Engine

Host CPU (Computation Time Less than 10% in Decryption Routine.)

pt: plaintext; ct: ciphertext; r: random; g, n: public key; λ, μ: private key.

Paillier Engine-2 MAC Computation Routine

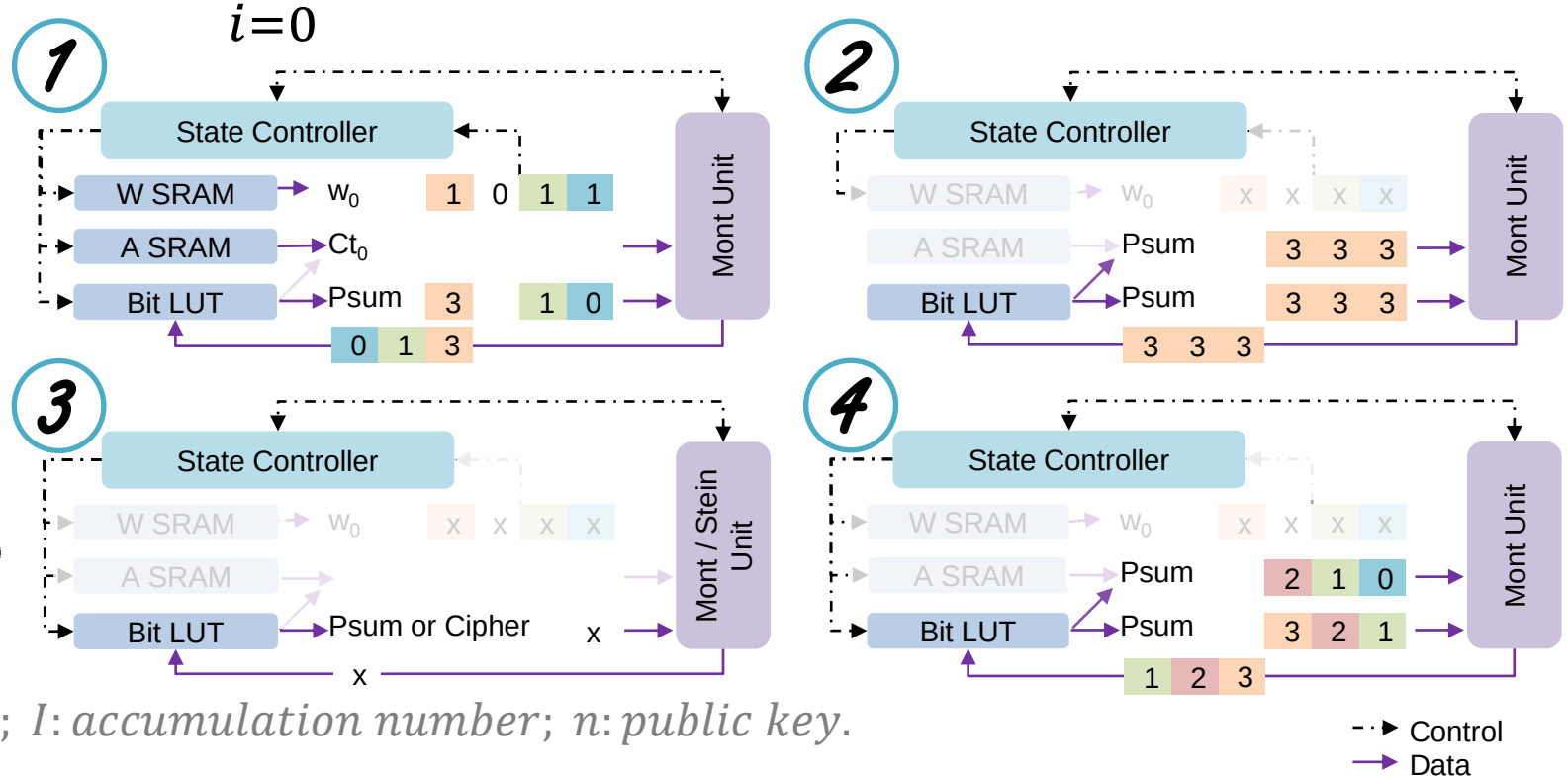
- Dataflow optimized for the Client and the **Cloud** sides.

$$ct' = \prod_{i=0}^{I} ct_i^{w_i} \bmod n^2$$

Cloud

MAC Operation

Bit-Serial Sparse MAC Dataflow inside PE

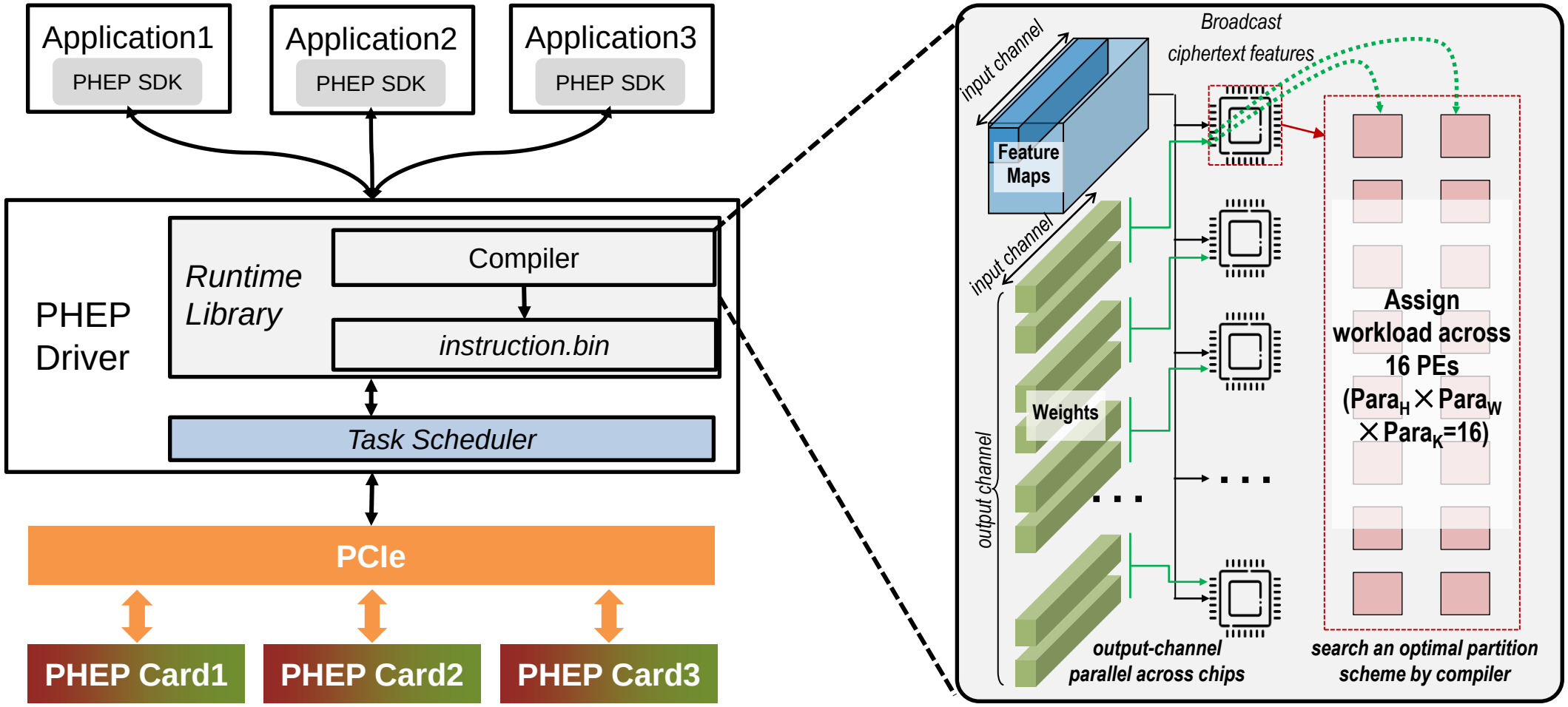


ct: $ciphertext$; w: $weight$; I: $accumulation\ number$; n: $public\ key$.

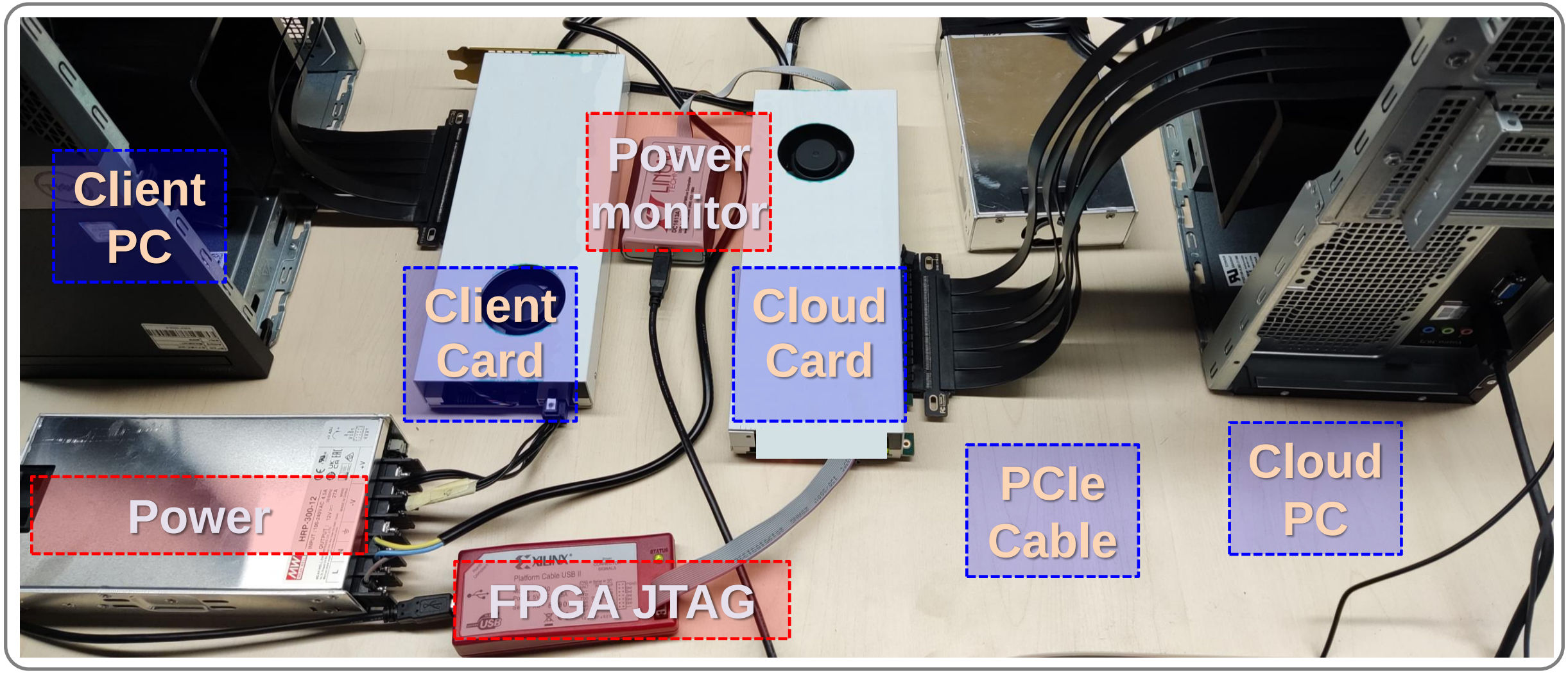
The PHEP Software and Performance

Software Stack

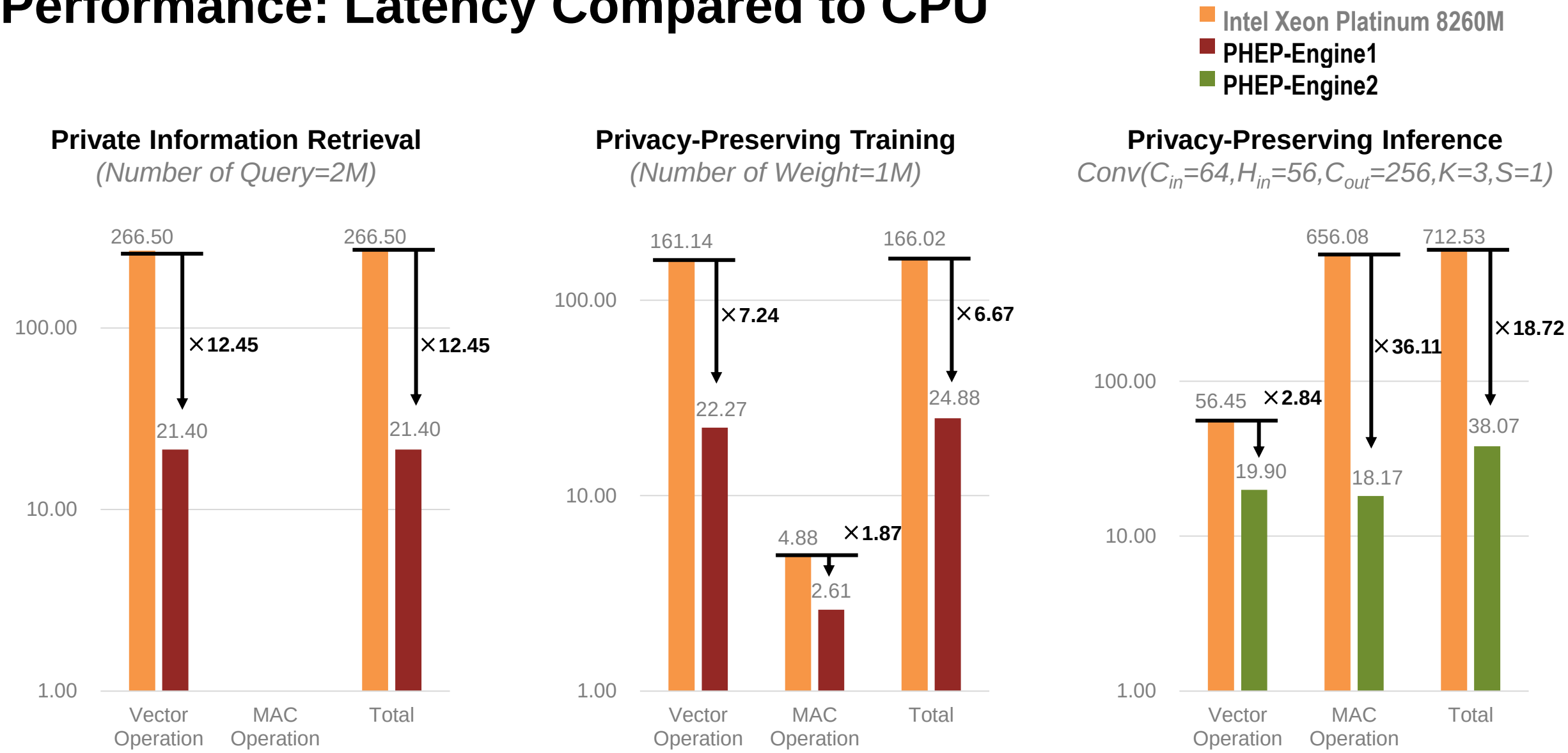
Paillier-Enabled Software Stack with PHEP Driver



System-Level Evaluation Platform



Performance: Latency Compared to CPU



$Bit\ width\ of\ ciphertext = 4096, Bit\ width\ of\ plaintext = 64, Bit\ width\ of\ weight\ in\ Conv = 8.$

High Performance Paillier Homomorphic Encryption Processors

PHEP Engine-1

- ***480 TOPS (INT8)***
- ***Client Encryption: 84KOPs***
- ***Cloud Computation: 402KOPs***
- ***Client Decryption: 106KOPs***

PHEP Engine-2

- ***192 TOPS (INT8)***
- ***Client Encryption: 52KOPs***
- ***Cloud Computation: 47MOPs***
- ***Client Decryption: 48KOPs***

$Bit\ width\ of\ ciphertext = 4096, Bit\ width\ of\ plaintext = 64, Bit\ width\ of\ weight\ in\ Conv = 8.$
$Maximum\ Performance\ in\ Optimized\ Applications.$

PHEP