



華控清交
HUA KONG
TSINGJIAO



Polar Bear Tech
北极雄芯

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

Guiming Shi¹, Yi Li², Xueqiang Wang², Zhanhong Tan¹, Dapeng Cao³,
Jingwei Cai¹, Yuchen Wei¹, Zehua Li³, Wuke Zhang⁴, Yifu Wu⁴,
Wei Xu^{1*}, and Kaisheng Ma^{1*}

¹Tsinghua University

²HuaKong TsingJiao

³Xi'an JiaoTong University

⁴Polar Bear Tech

*Corresponding Authors

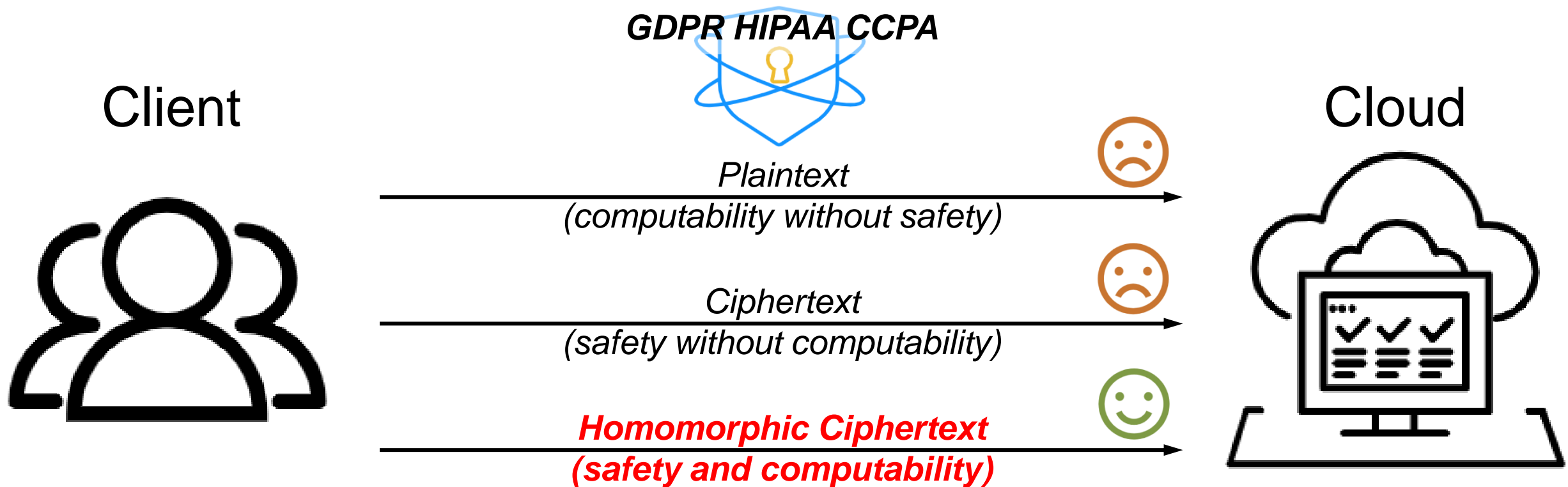


Extended Abstract

- 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

Hype Cycle for Data Security, 2022



Plateau will be reached: ○ <2 yrs. ● 2-5 yrs. ● 5-10 yrs. ▲ >10 yrs. ⊗ Obsolete before plateau

Paillier has Vast Applications in Different Sectors

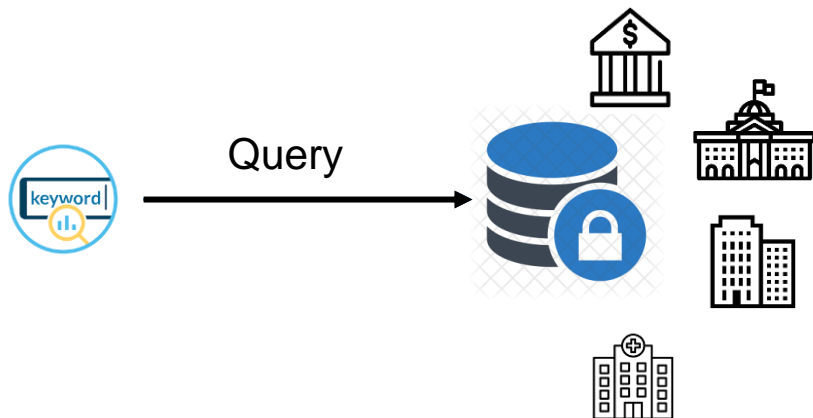
Federated Learning



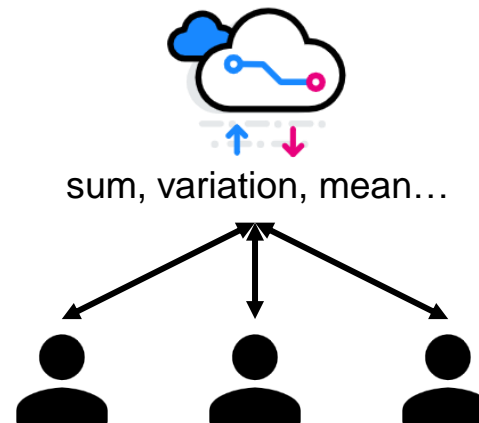
Homomorphic Commitment



Privacy-Preserving Query



Collaborative Statistics



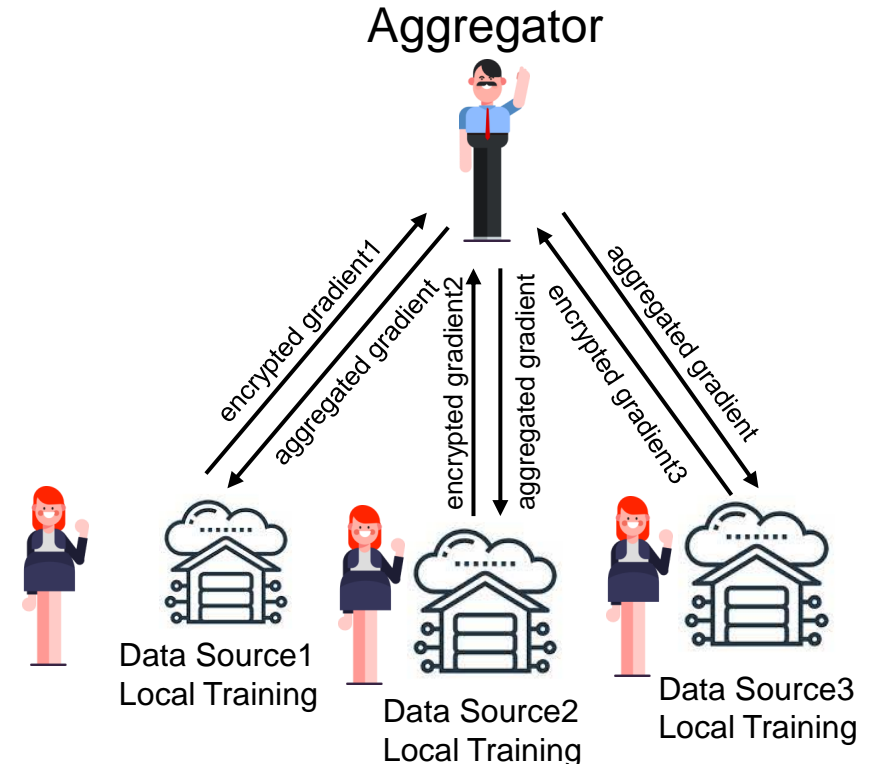
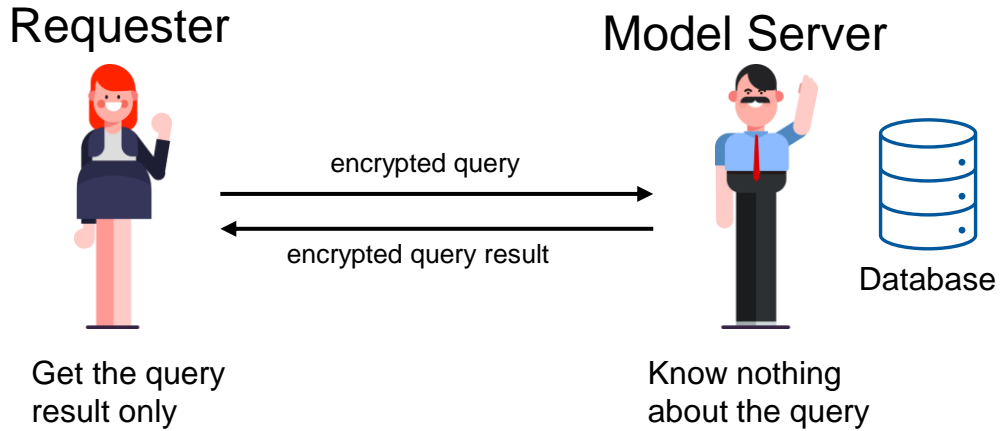
Electronic Signature



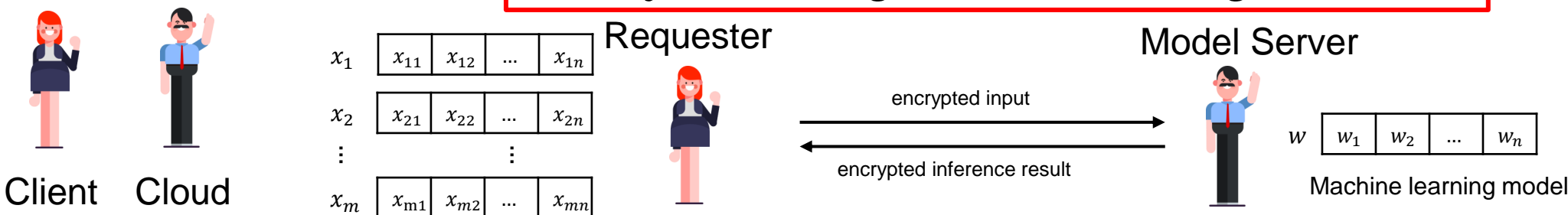
Typical Application Building Blocks for Paillier

Privacy-Preserving Machine Learning Training

Private Information Retrieval

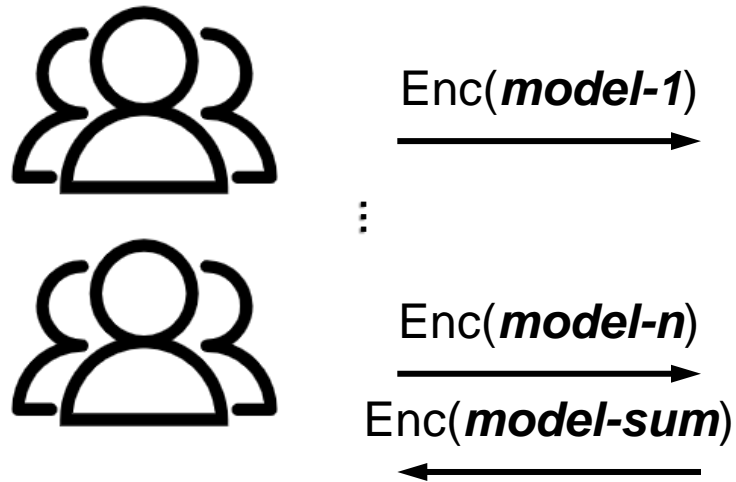


Privacy-Preserving Machine Learning Inference

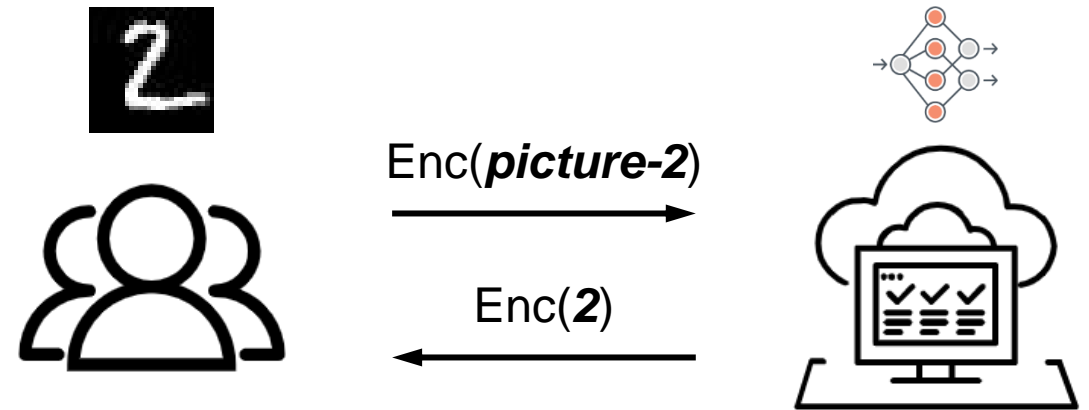


Bottleneck in Training and Inference Applications are Different

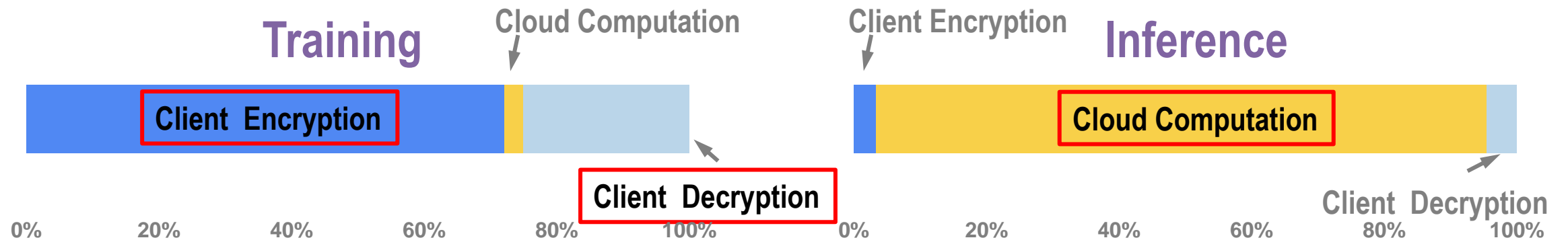
*Application-Training, Bottleneck:
Client Encryption and Decryption*



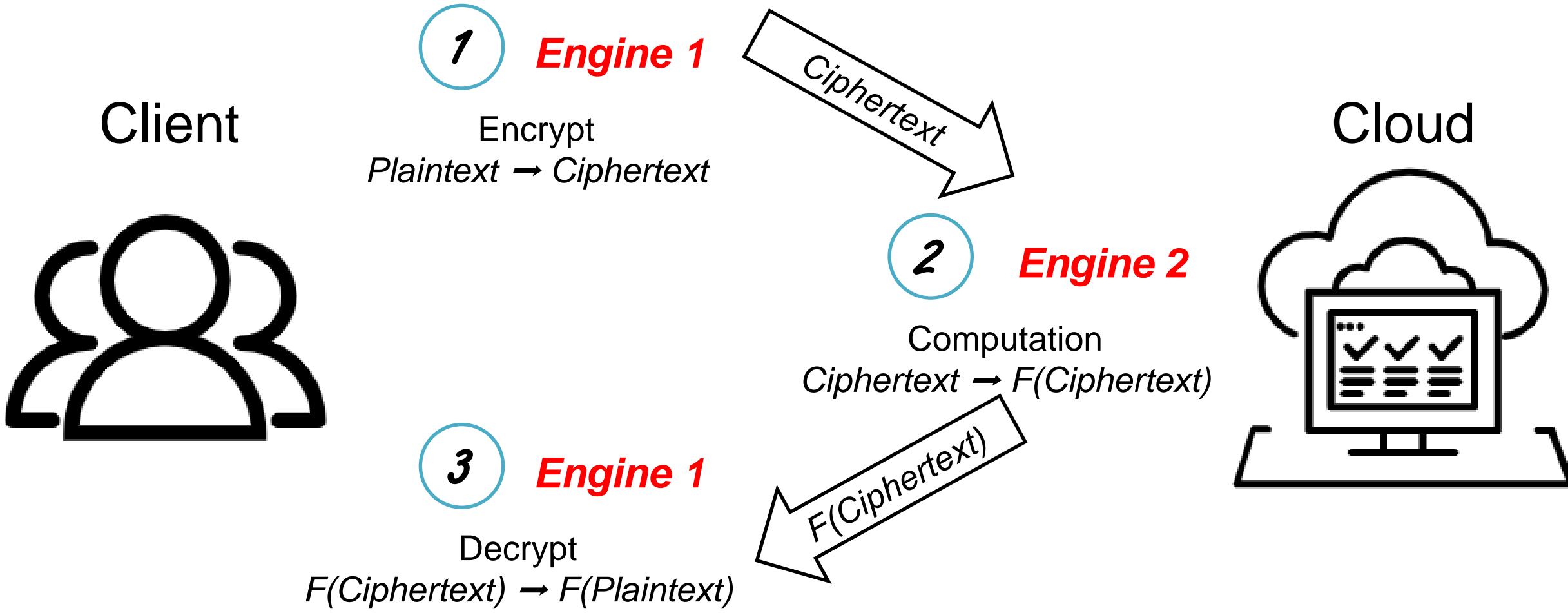
*Application-Inference, Bottleneck:
Cloud Computation*



Latency Breakdown On CPU

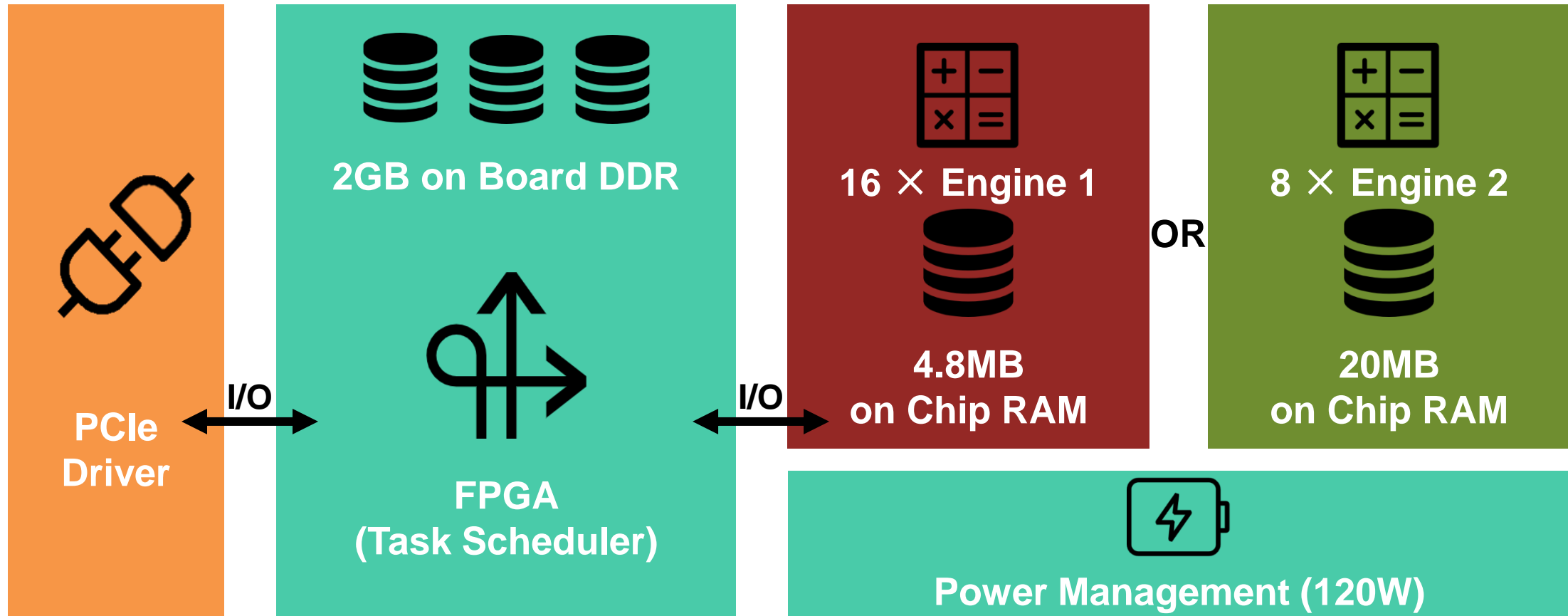


Our Solution: **Build 2 Chips with the Same Hardware Interface** to Meet the Requirement of Different Scenes

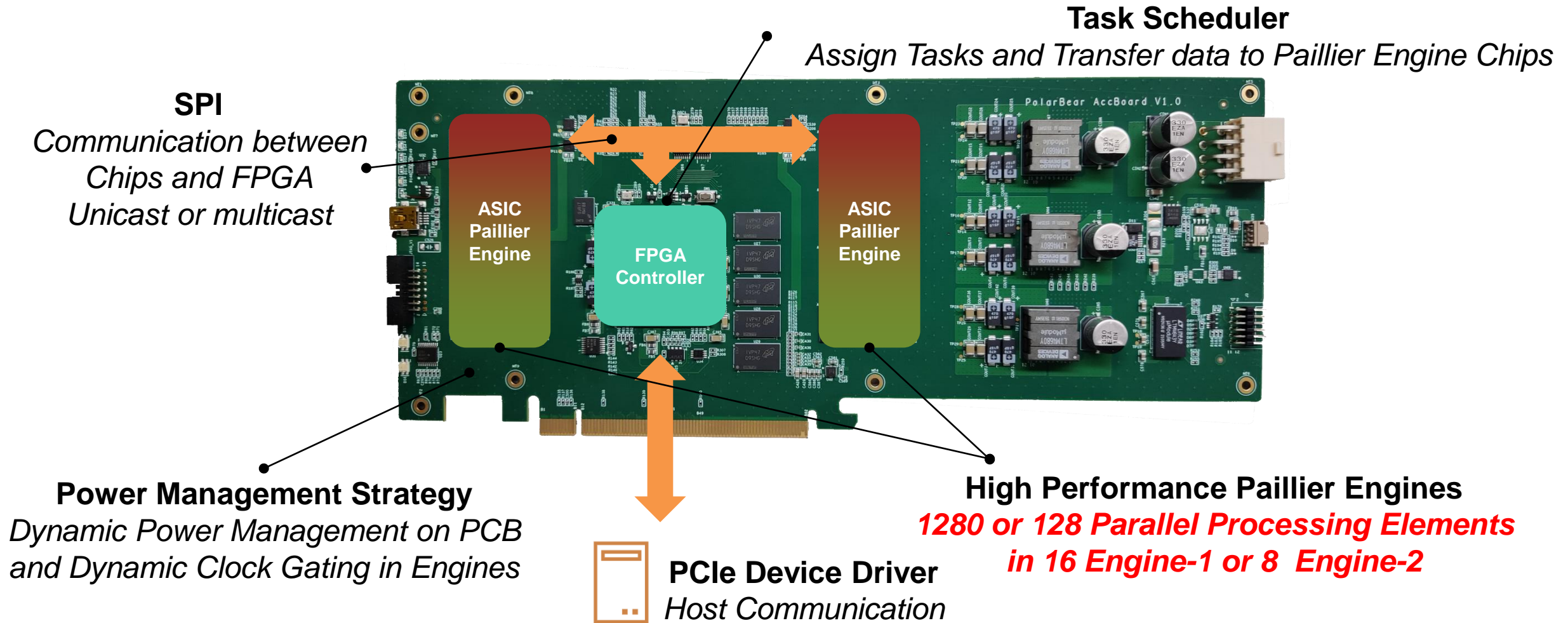


The PHEP Hardware

PHEP Accelerate Board: Supporting both Engine Chips

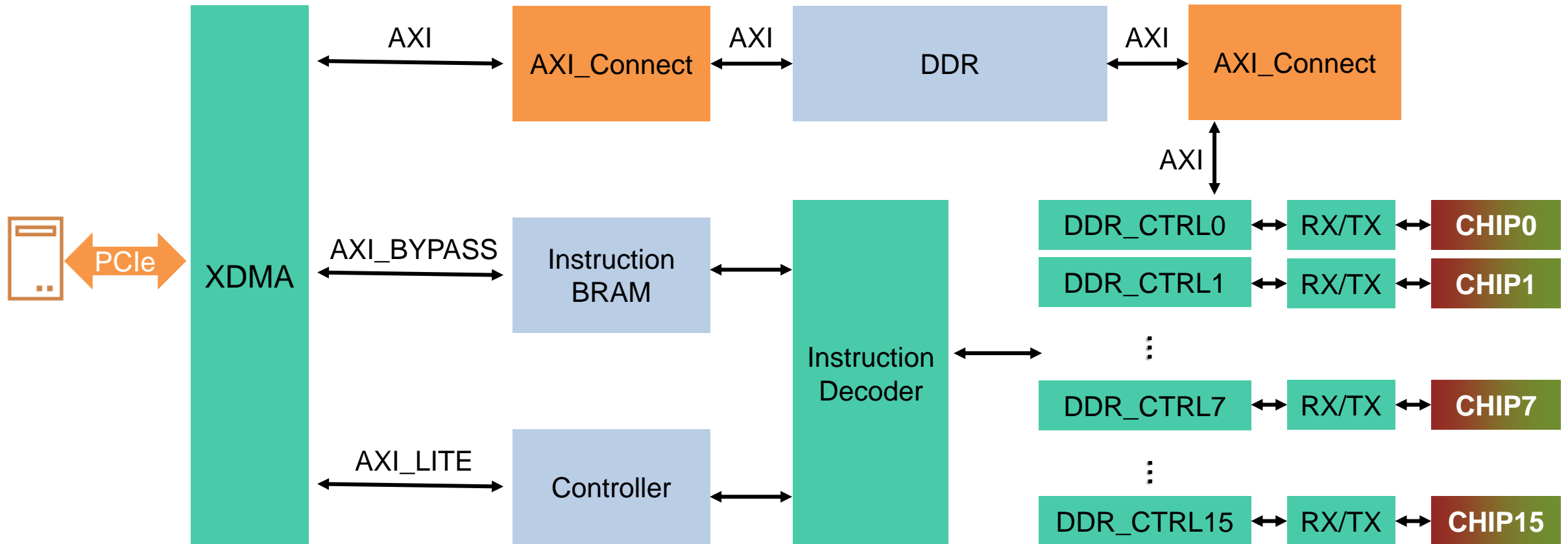


PHEP Accelerate Board Overview



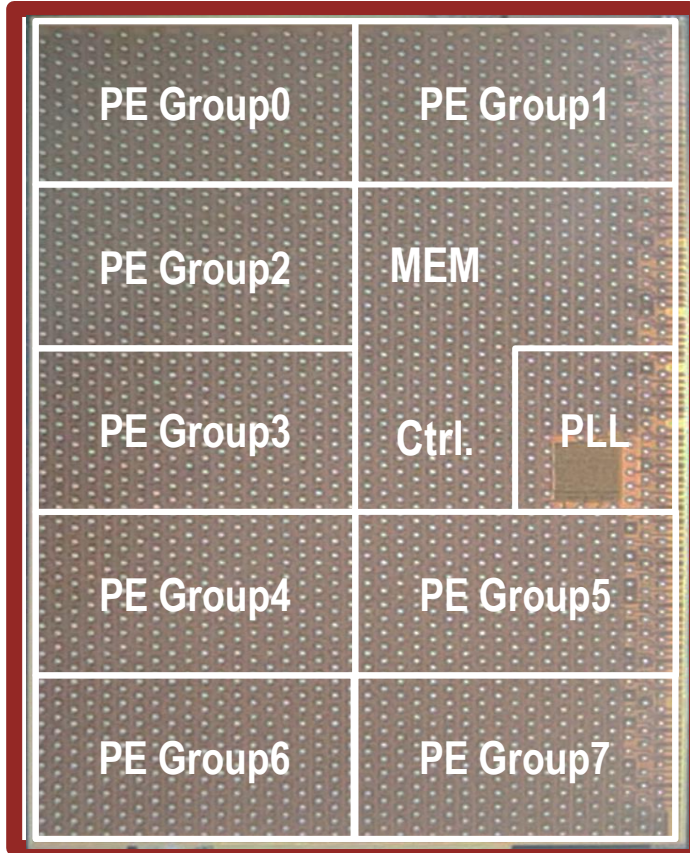
FPGA-Based Scheduler and I/O Controller

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

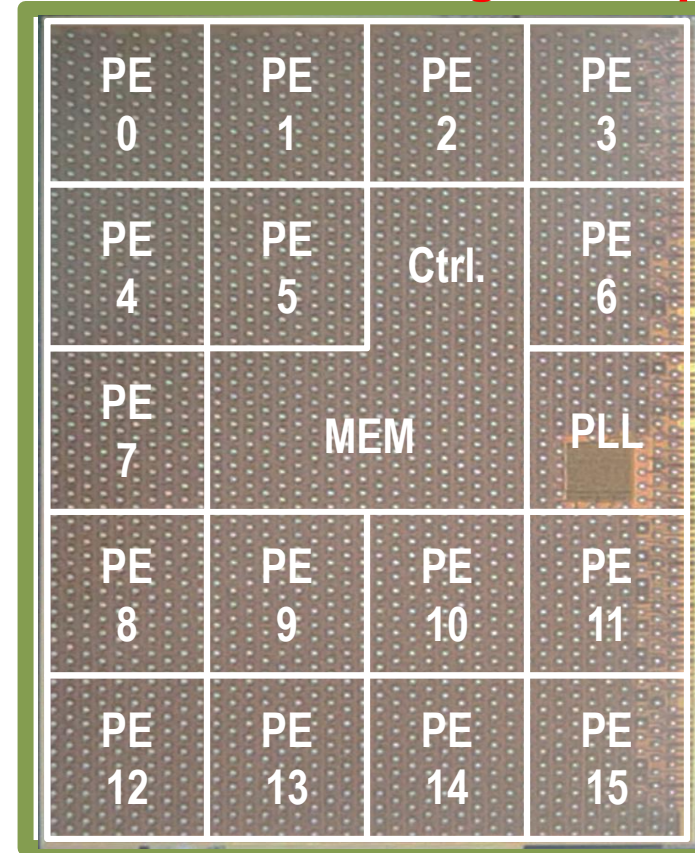


Comparison of the Two Engines: Specification

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

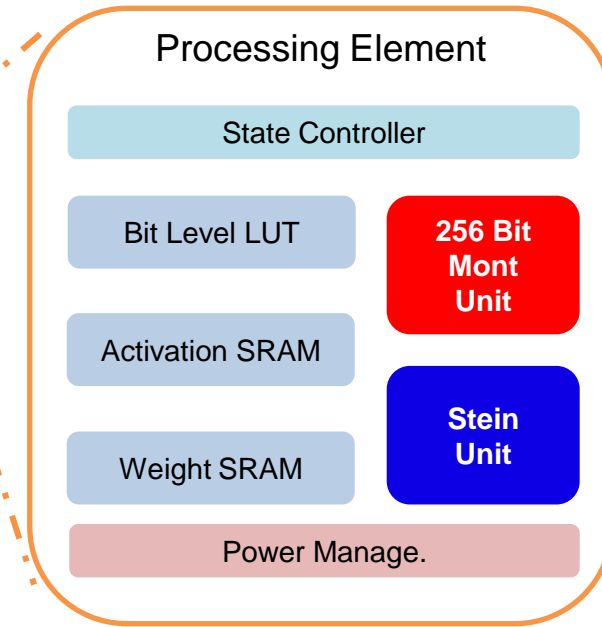
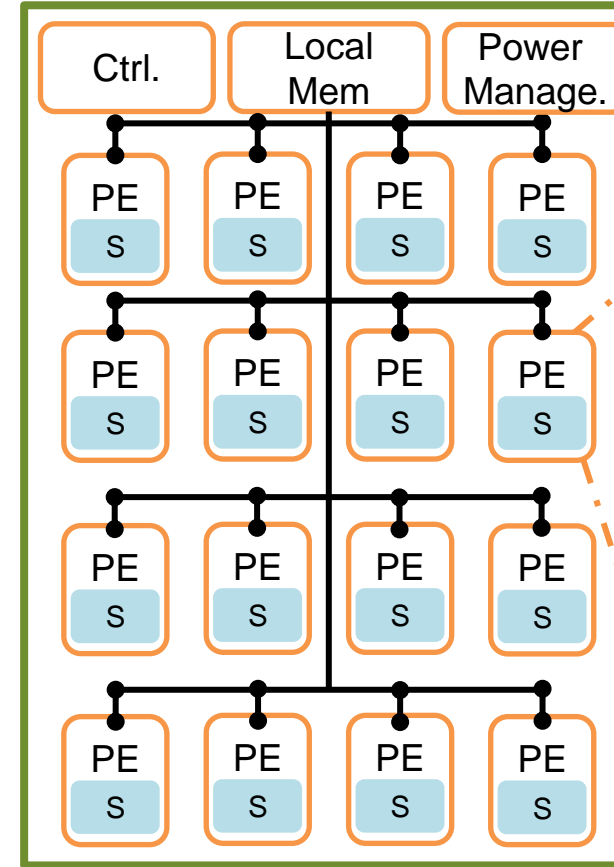
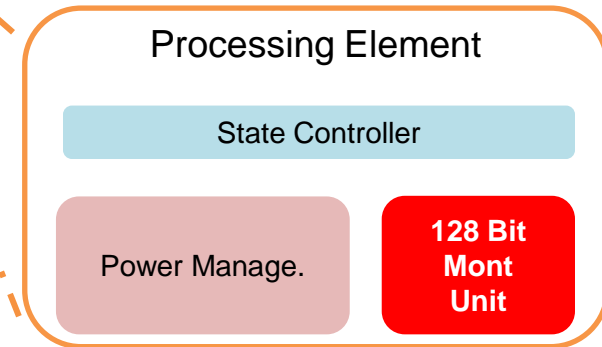
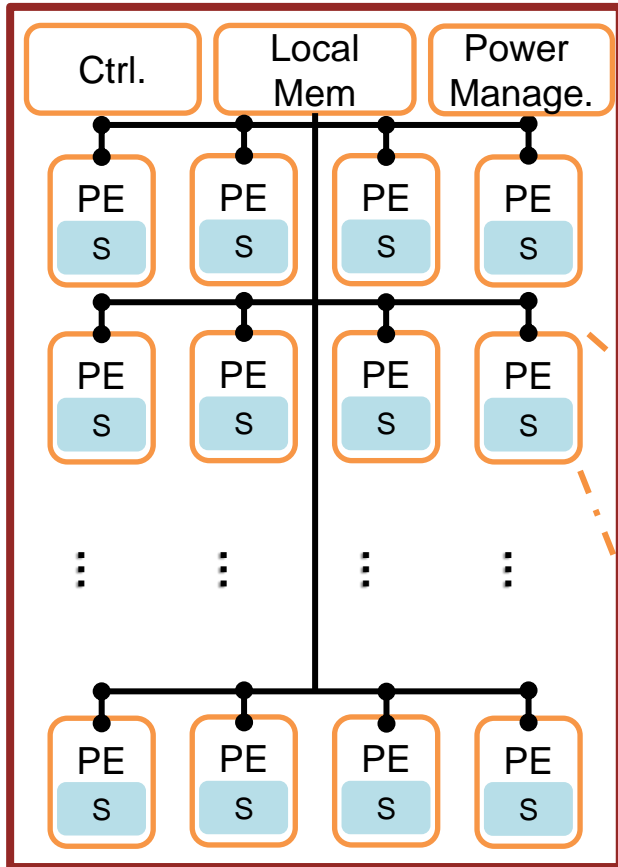
- Optimized for Performance**

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!!!

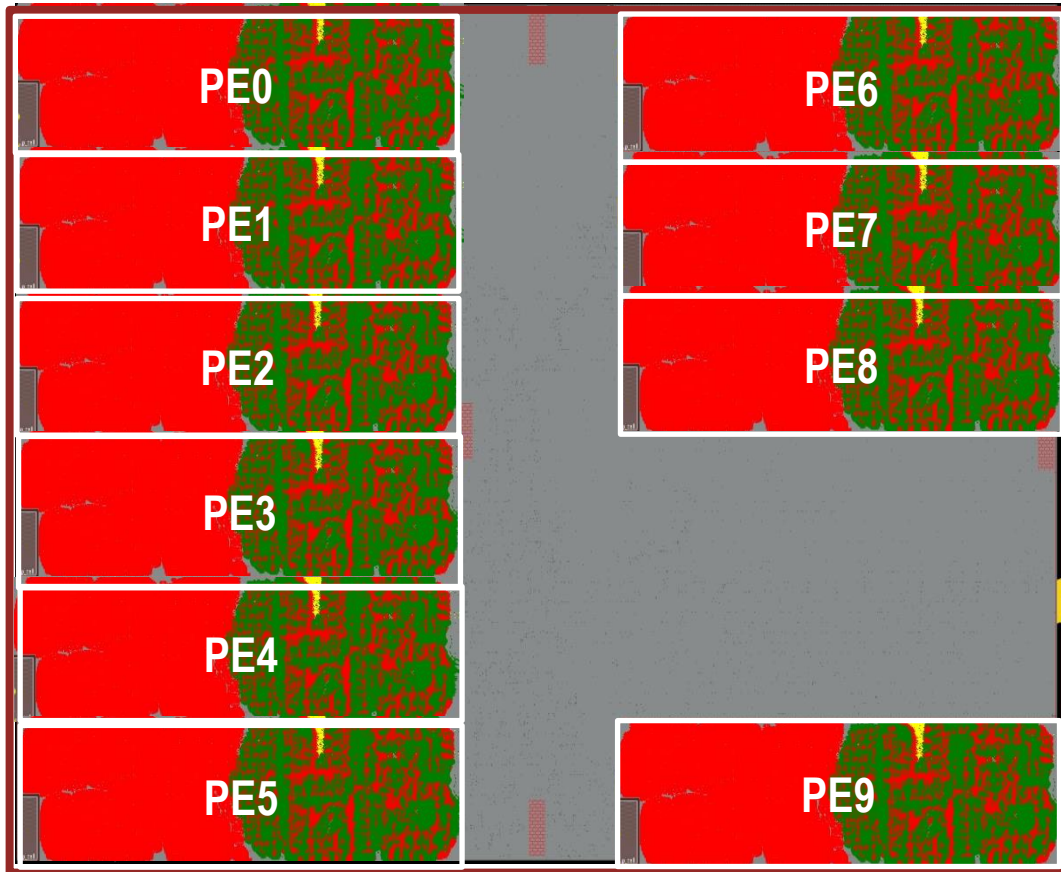


80 Parallel Processing Elements
400KB SRAM

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



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

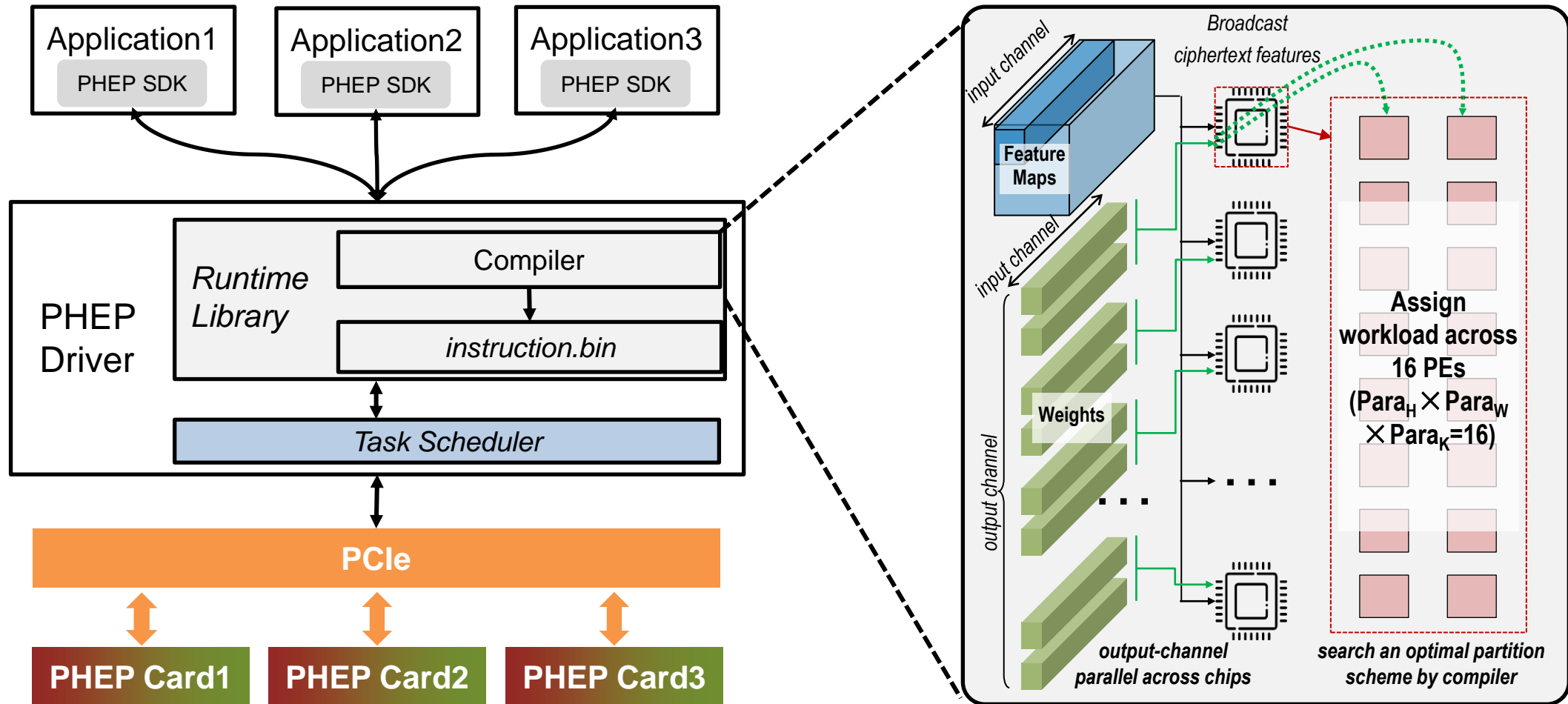


- Montgomery Unit
- Stein Modular Inversion Unit
- Register

The PHEP Software and Performance

Software Stack

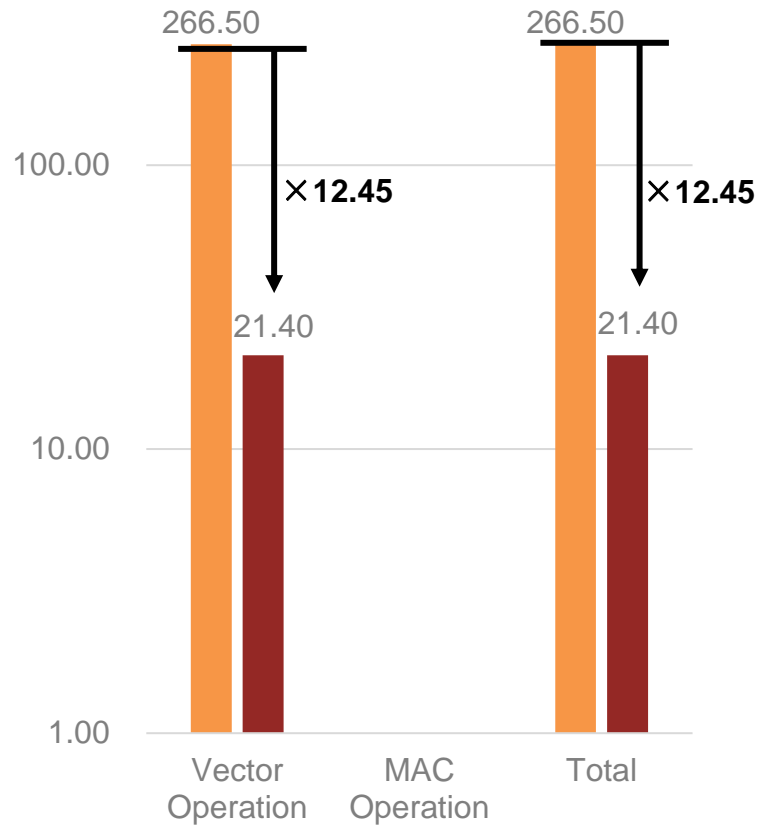
Paillier-Enabled Software Stack with PHEP Driver



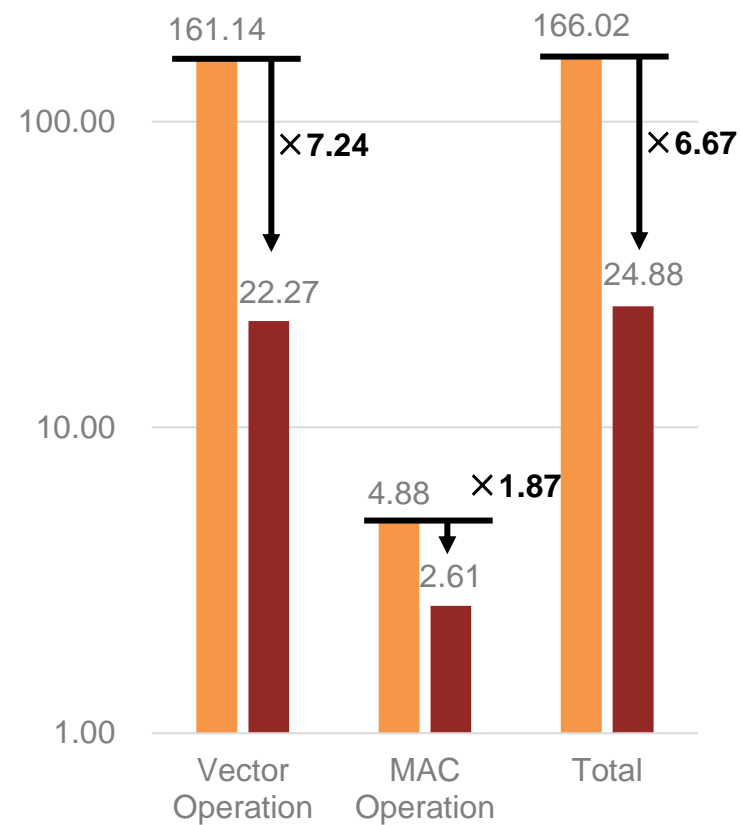
Performance: Latency Compared to CPU

- Intel Xeon Platinum 8260M
- PHEP-Engine1
- PHEP-Engine2

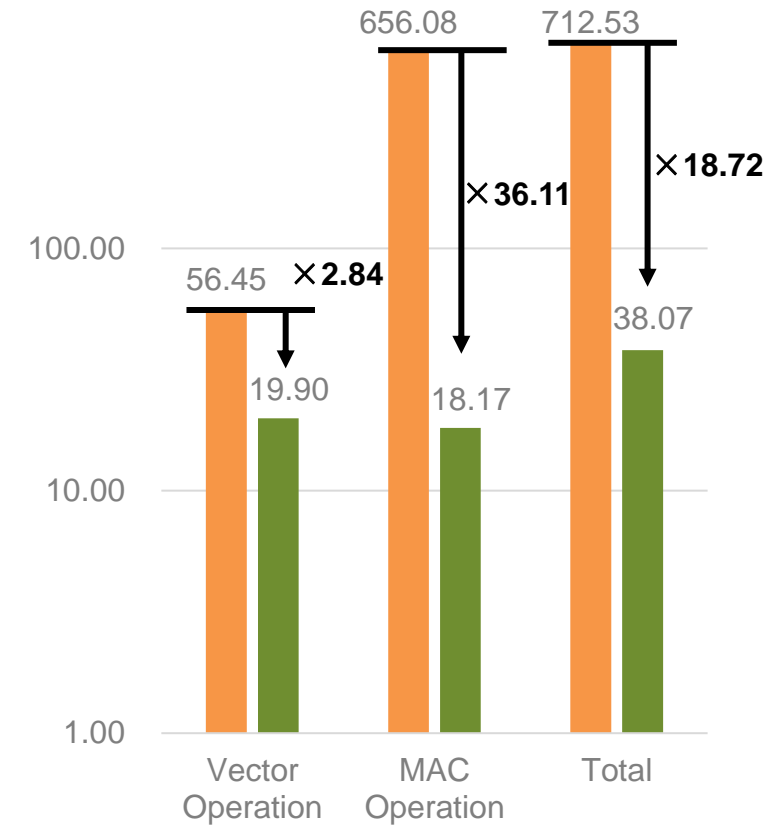
Private Information Retrieval
(Number of Query=2M)



Privacy-Preserving Training
(Number of Weight=1M)



Privacy-Preserving Inference
 $Conv(C_{in}=64, H_{in}=56, C_{out}=256, K=3, S=1)$



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.

Thank You

shigm21@mails.tsinghua.edu.cn