

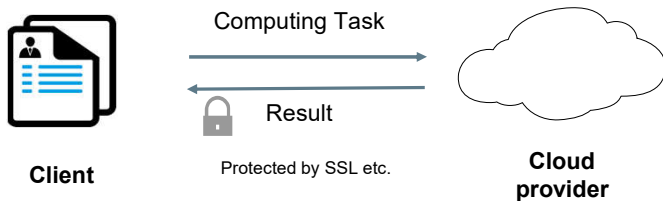
# Enabling Rack-scale Confidential Computing using Heterogeneous Trusted Execution Environment

王文浩  
2020年7月

# • Our TEE-related projects

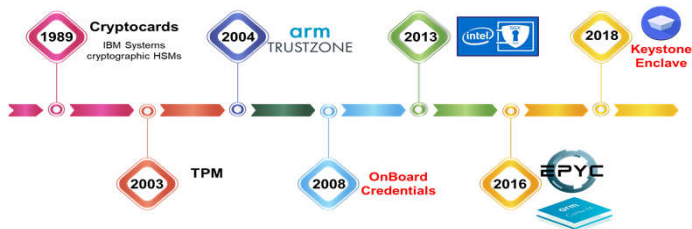
- Intel SGX side channel attacks and defenses
  - Page-level (access-bit based) attack, Cache-DRAM attack
  - 2-level directional branch prediction attack
  - Hyperrace: protection against AEX and HT-based attacks
- Enhancing existing TEEs (implemented on SGX)
  - Multi-users in-enclave service – protection against memory corruption attacks
  - Confidential attestation – attestation without undermining code privacy
- Others
  - TEE assisted HE/MPC/ML/FL
  - Customized TEEs
    - **Heterogeneous TEE**
    - Hardware in-enclave isolation
    - Hardware protection against side channels

- **Trusting the cloud provider is difficult**
  - Bugs in the software stack, such as hypervisor
  - The cloud provider may steal user data for its own interest
  - Malicious insider threat (from cloud administrators)



# • Introducing (hardware) TEEs to isolate computation

- TEEs cannot be bypassed by software
  - Hardware root of trust
  - Protection against privileged software attacks (such as corrupted hypervisor, operating system, SMM, BIOS etc.)
  - (Optional) Protection against certain hardware attacks
  - (Optional) Remote trust establishment with remote attestation
  - Intel SGX, ARM TrustZone, AMD SEV etc.



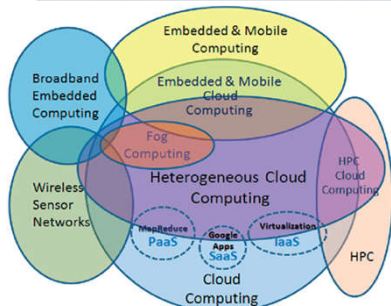
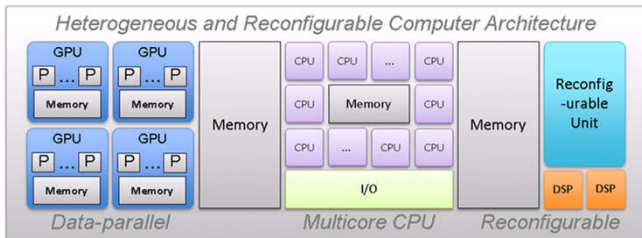
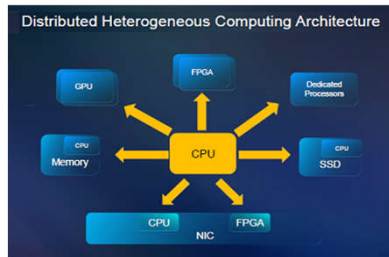
# • Heterogeneous (Cloud) Computing

## • Flexibility

- CPU > GPU > FPGA > ASIC

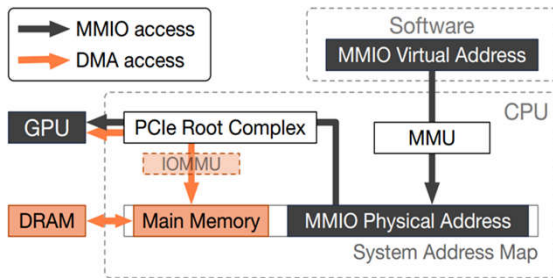
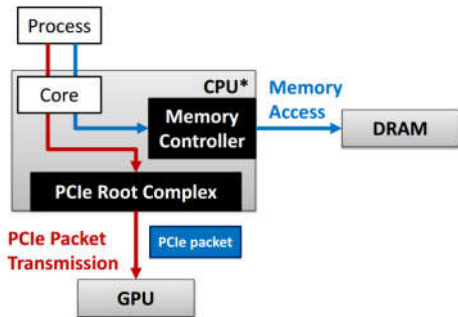
## • Efficiency

- CPU < GPU < FPGA < ASIC



- **Problem**
- TEE support for heterogeneous computing units
  - GPU, FPGA, (AI) accelerators
  - Scalable confidential compute or data intensive (CDI) computing
- Existing research proposals for GPU TEEs
  - *Graviton, HIX*
  - Hardware changes to CPU (adopted by *HIX*) or GPU (adopted by *Graviton*)
  - Performance critical hardware (GPU cores) unchanged

- **How are CPU and GPU connected?**

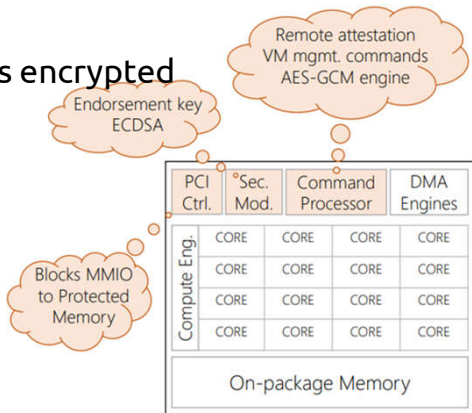
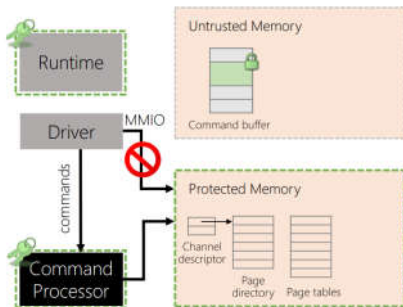


I/O path in PCI Express system architecture

# Existing GPU TEEs

- Graviton

- CPU's MMIO accesses to protected memory are blocked by the GPU hardware
- DMA buffer in untrusted memory is encrypted

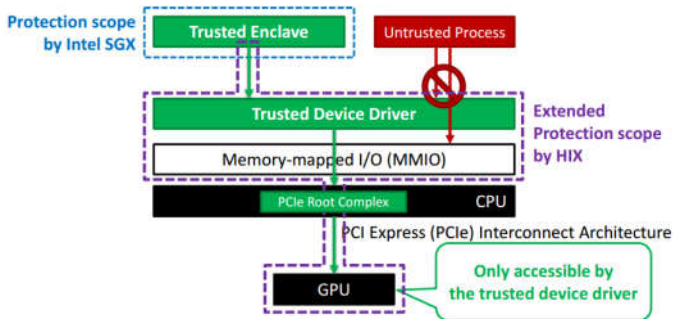




# Existing GPU TEEs

- HIX

- Extend TEE to I/O path (from SGX enclave to the device)
- Modify CPU to support GPU enclave (trusted GPU driver), which has exclusive access to GPU MMIO region
- DMA buffer in untrusted memory is encrypted



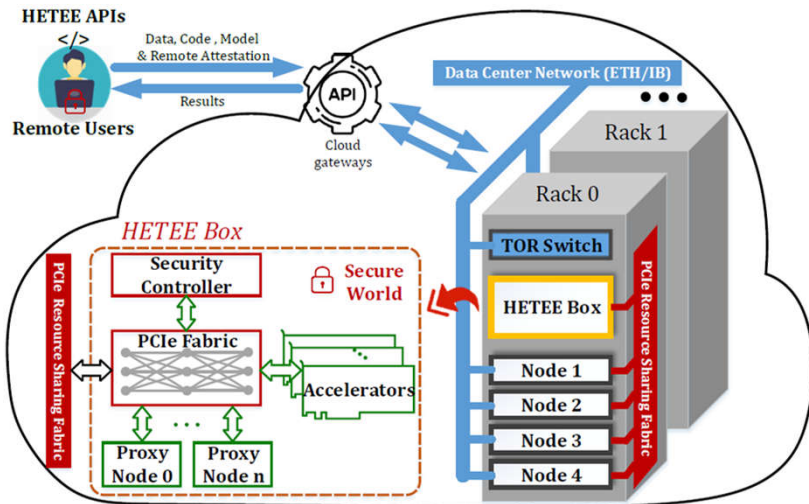
## • Existing GPU TEEs

- Hardware modifications leave large volumes of **legacy** GPUs unprotected
- GPU programming paradigm features **frequent communication**
  - Communication overhead for frequent encryption/decryption
- Communication patterns lead to **side channel leakages**
  - For example, GPU kernel execution time
  - Ref: Using timing information to recover image classification in ImageNet [1]

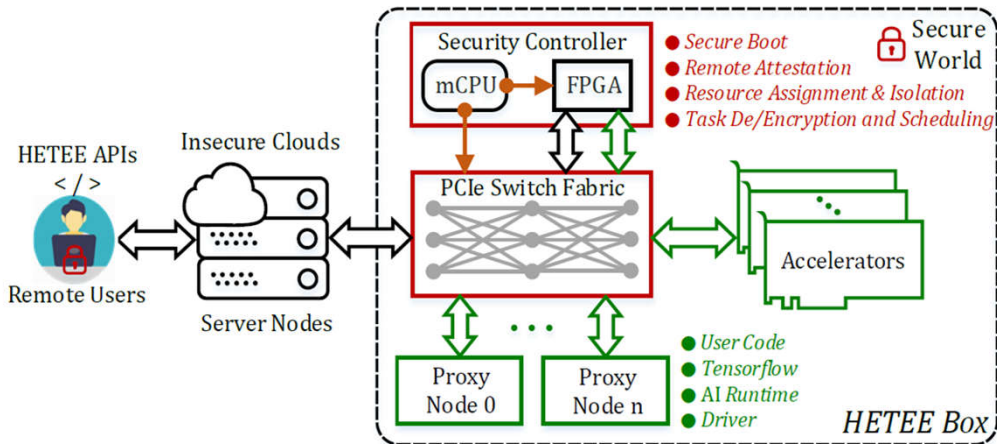
[1] Hunt et al. Telekine: Secure Computing with Cloud GPUs (NSDI'2020)

- **Design goals**
  - Low (no) hardware changes to existing GPUs or accelerators
  - Small trusted computing base (TCB)
  - Limited (side channel) attack surface
  - Low performance overhead
  - Threat Model
    - Software attacks
    - Physical attacks
    - Firmware
    - PaaS Model

- # Rack-scale Application Scenario

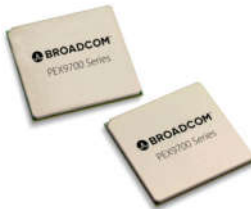


# • HETEE Overview



- # HETEE Overview

## PCI-e switch fabric

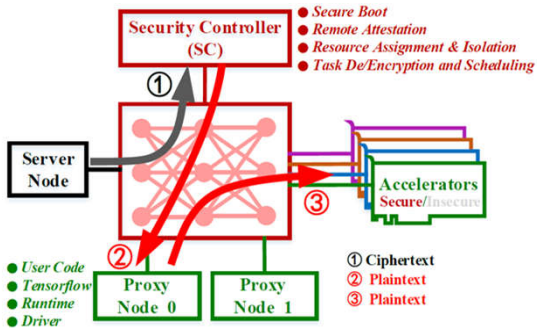


The PEX9797 family offers fully **non-blocking** and low-latency PCI Express Gen 3 managed switches (96 to 12 lanes) for **high-performance, low-latency, scalable, and cost-effective** PCIe-based Flash JBODs, NVMe HBAs and Rackscale Fabrics. This Broadcom technology provides **enterprise and cloud data center equipment designers the ability to share pools of I/Os and compute resources** and to enable multiple hosts to reside on a single PCIe-based network topology using standard PCIe enumeration – a capability not previously available in PCIe. The hosts communicate through Tunneling Window Connection (TWC), Ethernet-like DMA, and do so using standard hosts, end-points. Broadcom offers complete turn-key solutions for various applications that include switch silicon and software that allow customers to rapidly release their products to the market.

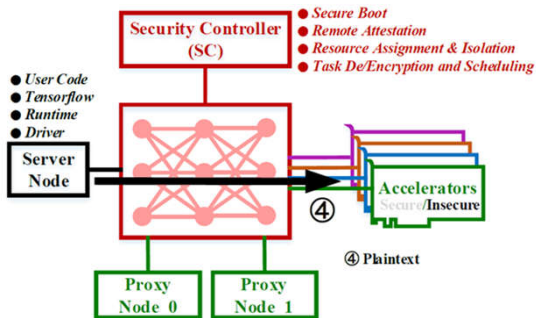
# • Elastic Resources Allocation and Isolation

PCI-e switch fabric: Software-defined fabric

- High performance
- Flexible topology



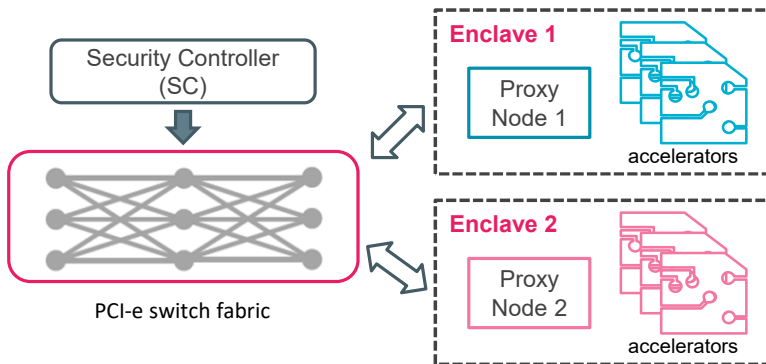
(a) secure mode



(b) insecure mode

- **HETEE enclaves**

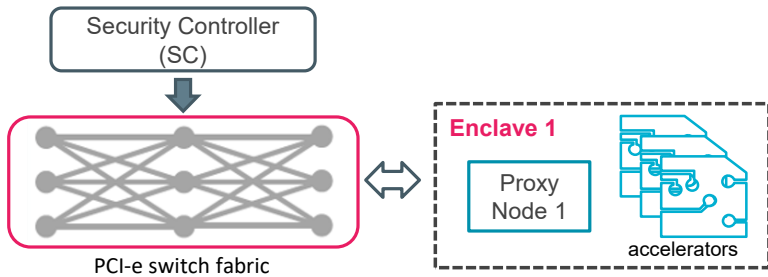
- Physical isolation for *concurrent enclaves*





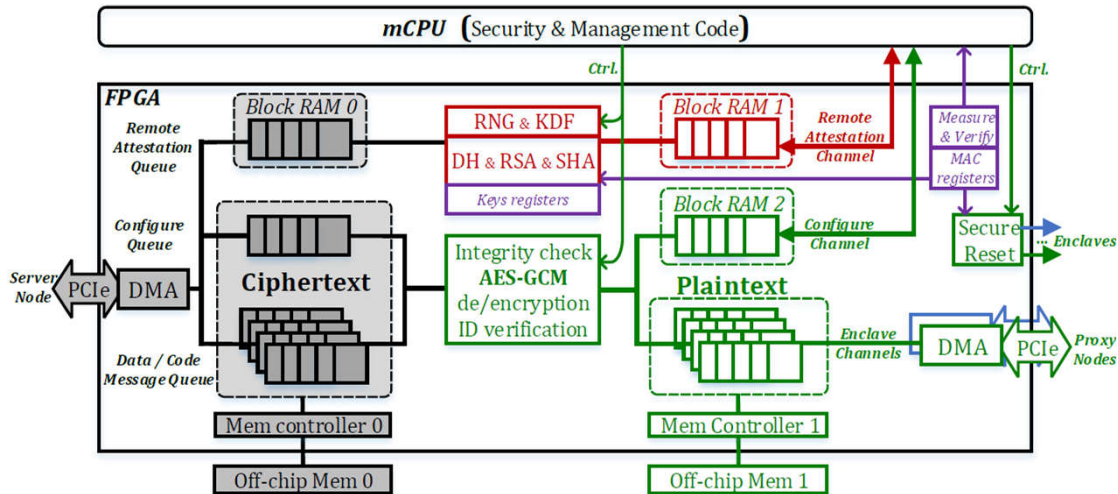
# • HETEE enclaves

- Physical isolation for *sequential enclaves*
  - protected by secure reset mechanism
  - how to securely reset a GPU/accelerator/proxy node?
  - ◆ `nvidia-smi --gpu-reset`
  - assumptions: accelerator firmwares are protected

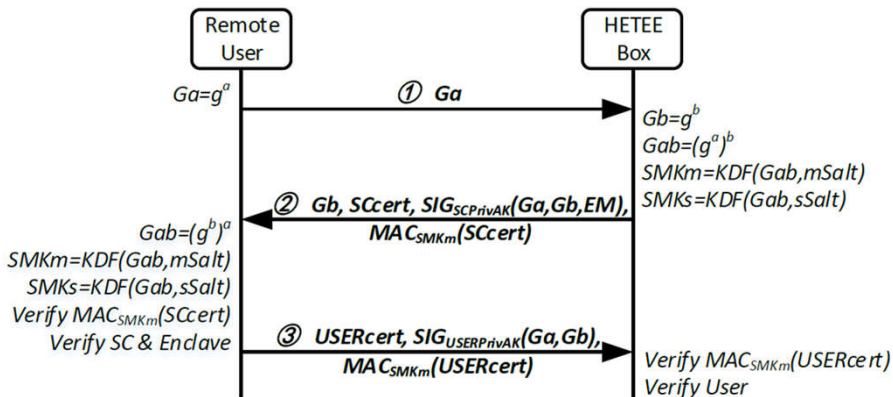


Q. Why the proxy node software (OS/TensorFlow framework/GPU driver etc.) are outside the TCB?

- SC modules

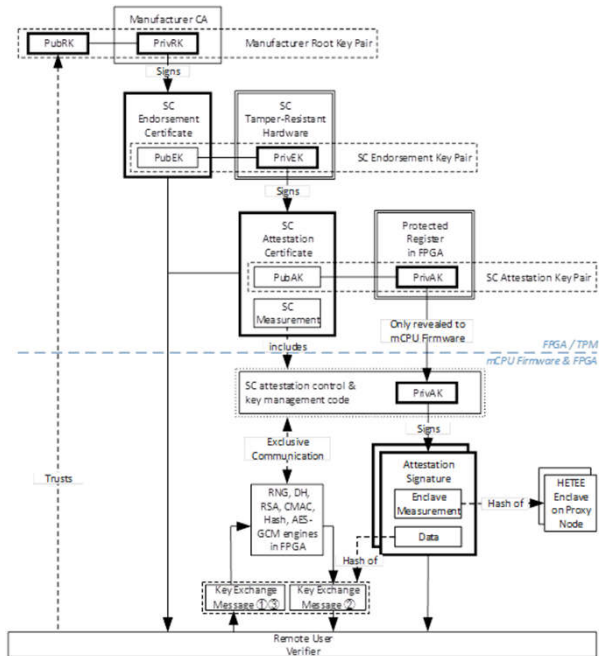


# • Remote Attestation and Symmetric Key Negotiation



Each HETEE platform includes two sets of public key pairs, Endorsement Key (EK) and Attestation Key (AK). The SC\_cert contains the **SC measurement**, certificate chain and device ID, and is signed using the EK private key. The enclave measurement (EM) is signed with the AK private key.

# • Certificate Chain



- **Security Analysis**
- Physical attack protection
  - A microcontroller (MCU) system and a set of sensors (e.g., pressure, vibration and temperature etc.)
- Protection against software attacks
  - Host software can be fully controlled by the attacker
  - Proxy software can have vulnerabilities (under the PaaS model)
- Trust chain
  - RoT: endorsement key, stored in encrypted FPGA bitstream
  - FPGA (bitstream) -> SC firmware code -> proxy cpu firmware
- Side Channels
  - Physical isolation to prevent shared resources

- **Security Analysis**
- TCB analysis
  - SC: **FPGA encrypted bitstream, and mCPU firmware**
  - **GPU firmware**
  - **MCU for protection against physical attacks**
  - Proxy CPU firmware: verified by SC and **not** included
- Validity for assumptions
  - Is FPGA bitstream secure <sup>[2]</sup>? Bitstream encryption for authenticity and confidentiality
  - Is the mCPU firmware secure (secure boot support)?
  - GPU vendors adopt firmware signature checking?
  - Physical attack protection using the MCU?

[2] Ender et al. The Unpatchable Silicon: A Full Break of the Bitstream Encryption of Xilinx 7-Series FPGAs. Usenix Security 2020.

# • Discussion

- Sealing

- System vendor or authorized third party

- Maintenance

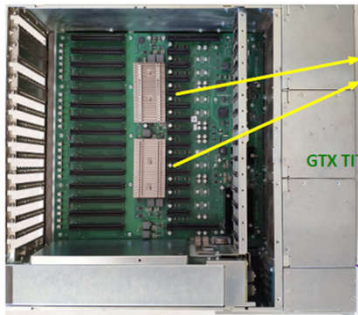
- Reopened by the system vendor or authorized third party
  - Proxy node software can be upgraded without opening the classic

- Cooling

- Cooling mechanisms, such as air channel design, front/rear thermal vent, fans with speed control, air or water cooling can be integrated

# • Prototype System

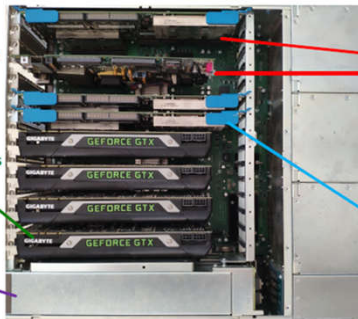
Component	Hardware	Software
Security Controller	Intel Xeon-E3 1220V6 DDR4 16GB 2400MHz Xilinx Zynq FPGA	Tailored coreboot 4.10 with security management code, binary size is < 300KB
Proxy Node	Intel Xeon-E3 1220V6 DDR4 16GB 2400MHz	TensorFlow 1.11.0 CUDA 9.0 Nvidia Driver 396.54 CentOS 7.2
GPU	Nvidia GTX TITAN	
PCIe Fabric	PEX9797	



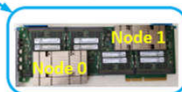
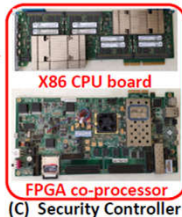
(a) PCIe ExpressFabric backplane

PEX9797  
Chips

GTX TITAN X GPUs



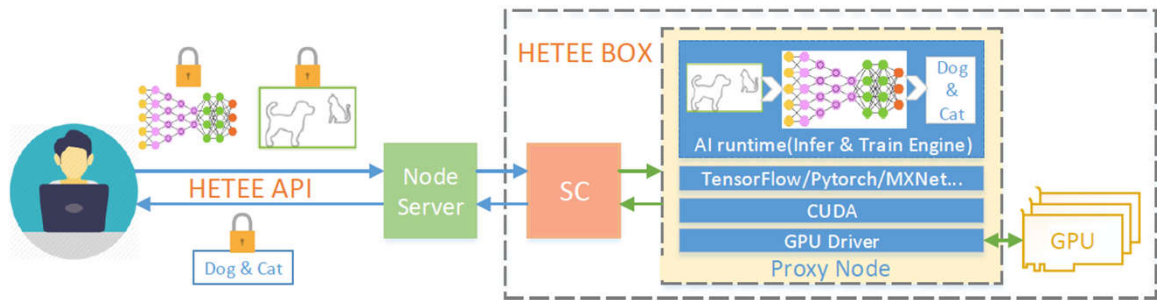
(b) HETEE Box



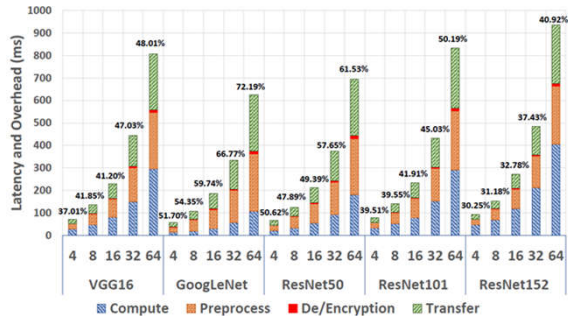
(d) Proxy Nodes



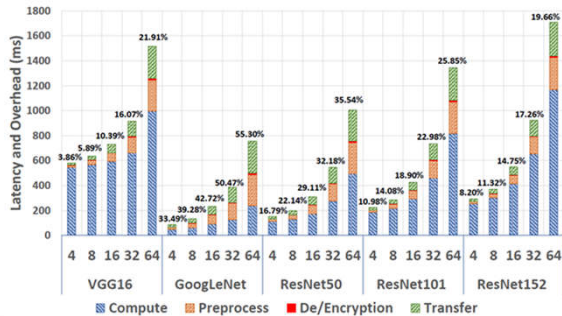
- **Confidential AI service**



# • Performance Evaluation



(a) Inference

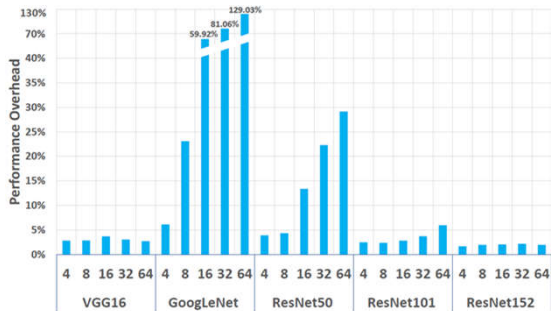


(a) Training

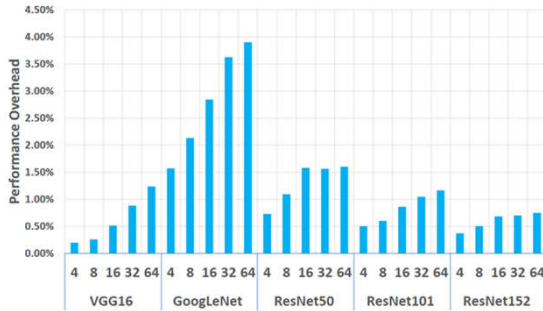
HETEE latency overhead on single GPU with different batch sizes.

# • Performance Evaluation

Model	Model size	Parameters	Layers	Image size	Type
VGG16 [54]	500 MiB	138 M	16	224x224x3	1K
GoogLeNet [55]	28 MiB	5 M	22	224x224x3	1K
ResNet50 [56]	100 MiB	25 M	50	224x224x3	1K
ResNet101 [56]	150 MiB	44 M	101	224x224x3	1K
ResNet152 [56]	200 MiB	60 M	152	224x224x3	1K



(a) Inference



(a) Training

HETEE throughput overhead on single GPU with different batch sizes.

# • Performance Evaluation

HETEE inference throughput scalability evaluation (normalized to the baseline)

Model	Batch size	4			8			16		
	Number of GPU	1 GPU	2 GPUs	4 GPUs	1 GPU	2 GPUs	4 GPUs	1 GPU	2 GPUs	4 GPUs
VGG16	Baseline	1.00	1.75	2.52	1.00	1.58	2.74	1.00	1.84	3.02
	HETEE	0.97	1.65	2.48	0.97	1.53	2.56	0.96	1.67	2.68
GoogLeNet	Baseline	1.00	1.37	1.60	1.00	1.47	1.76	1.00	1.29	1.72
	HETEE	0.94	1.33	1.45	0.81	1.38	1.49	0.63	1.02	1.24
ResNet50	Baseline	1.00	1.66	2.73	1.00	1.53	2.39	1.00	1.69	2.73
	HETEE	0.96	1.61	2.46	0.96	1.54	2.34	0.90	1.62	2.39
ResNet101	Baseline	1.00	1.73	2.89	1.00	1.65	2.71	1.00	1.64	2.76
	HETEE	0.98	1.63	2.82	0.98	1.59	2.58	0.97	1.56	2.70
ResNet152	Baseline	1.00	1.72	3.26	1.00	1.79	3.10	1.00	1.67	3.35
	HETEE	0.98	1.57	2.99	0.98	1.61	2.82	0.98	1.61	3.28
Average	Baseline	1.00	1.65	2.60	1.00	1.60	2.54	1.00	1.63	2.72
	HETEE	0.97	1.56	2.44	0.94	1.53	2.36	0.89	1.50	2.46

Thanks