RESEARCH NOTE

NVIDIA H100 CONFIDENTIAL COMPUTING

STEVE MCDOWELL, CHIEF ANALYST
APRIL 30, 2024

CONTEXT

This week, NVIDIA made its confidential computing capabilities for its flagship NVIDIA Hopper H100 GPU, previewed in August 2023, generally available. This makes NVIDIA's H100 the first GPU with these capabilities, which are critical for protecting data as it is being processed.

Let’s look at confidential computing and how it works on the NVIDIA H100 GPU.

BACKGROUND: WHAT IS CONFIDENTIAL COMPUTING?

Confidential computing is a security technology that protects data at the hardware level while that data is processed. It works by isolating data into a protected area of memory called a Trusted Execution Environment (TEE) or secure enclave.

By isolating sensitive computations in a protected enclave within the processor, confidential computing ensures that sensitive data, such as personal information, trade secrets, and other confidential information, is not exposed to different parts of the system or cloud service providers. This is critical for complying with privacy regulations like GDPR and HIPAA, which mandate stringent data protection measures.

Confidential computing addresses the vulnerability of data during computation, a period when it has traditionally been exposed to potential threats despite the robust protections typically provided for data at rest and in transit.

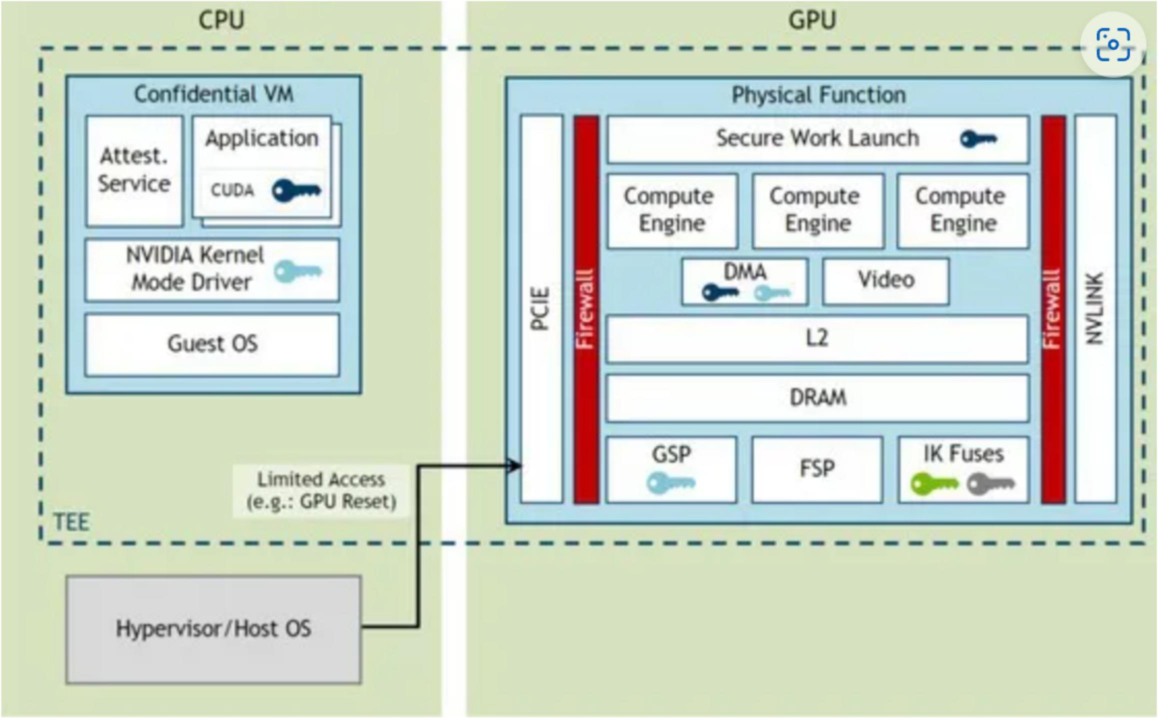
Key features of confidential computing include:

- **Trusted Execution Environment (TEE)**: A TEE provides a secure area within the processor where data can be processed and encrypted away from the typical operating environment. This isolation helps protect sensitive data from unauthorized access or modification by the operating system, hypervisor, or other applications, even if they are compromised.

- **Data Encryption in Use**: Confidential computing ensures that data remains encrypted when stored or sent over a network and while it is being processed. This continuous encryption minimizes windows of vulnerability across the data lifecycle.

- **Attestation**: Attestation mechanisms validate the integrity and authenticity of the TEE. This process involves verifying that the TEE is correctly secured before processing sensitive data. Attestation ensures that the TEE hasn't been tampered with and is running the expected code, providing users with proof that their data is secure.

- **Hardware Root of Trust (RoT)**: Most confidential computing approaches are based on a hardware root of trust, an inherently trusted component that provides a foundation for security operations. The root of trust is responsible for launching and managing the TEE securely.



A trusted execution environment.

Confidential computing brings with it several key benefits:

- **Enhanced Security**: By isolating computation in a secure enclave, confidential computing protects sensitive data from unauthorized access and leaks, even if other parts of the system are compromised.

- **Privacy Assurance**: It enables organizations to process data while guaranteeing privacy, making it ideal for scenarios involving multiple parties or jurisdictions with strict data protection regulations.
- **Regulatory Compliance**: Confidential computing can help organizations meet stringent regulatory requirements for data protection, such as GDPR or HIPAA, by providing enhanced controls over data in use.
- **Secure Multi-Party Computation**: It facilitates scenarios where multiple parties compute on a shared dataset without exposing their data to each other, which is suitable for collaborative research or aggregated analytics.

Confidential computing helps mitigate various vulnerabilities, including side-channel attacks, where attackers exploit the hardware execution environment to access sensitive data. By securing data in use, confidential computing closes a critical gap in data protection that traditional security measures (focused only on data at rest and in transit) do not address.

H100 CONFIDENTIAL COMPUTING

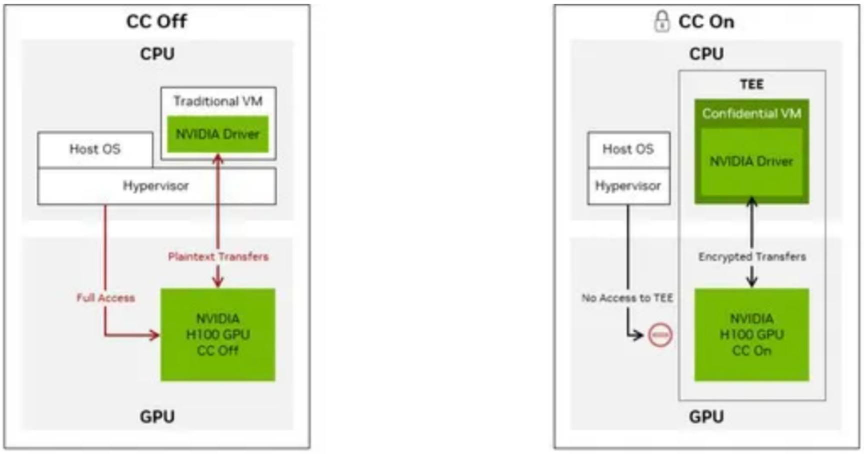
NVIDIA has introduced confidential computing to its NVIDIA H100 Tensor Core GPU, delivering a significant advancement in data security during processing. It also makes the NVIDIA H100 the first GPU with these capabilities.

The H100 GPU incorporates a hardware-based TEE anchored by an on-die hardware root of trust. This setup ensures that all computations are performed in an isolated, secure environment, shielding them from external threats and internal vulnerabilities.

The GPU also supports a secure and measured boot process. This process involves a sequence where the GPU boot is verified and authenticated, ensuring that the firmware and software initiating on the GPU are not tampered with or corrupted. The GPU also generates an attestation report, a cryptographically signed document providing evidence of the GPU's state and the integrity of its operational environment. The user can verify this report to ensure the system's security before processing data.

The H100 GPU supports several modes of operation related to confidential computing:

- **CC-Off**: Standard operation where no confidential computing features are active.
- **CC-On**: All confidential computing features are fully activated. This includes all necessary security protocols, disabling performance counters to prevent side-channel attacks and activating all firewalls.
- **CC-DevTools**: A mode that allows developers to use performance counters for optimization and troubleshooting while maintaining a level of security similar to CC-On.



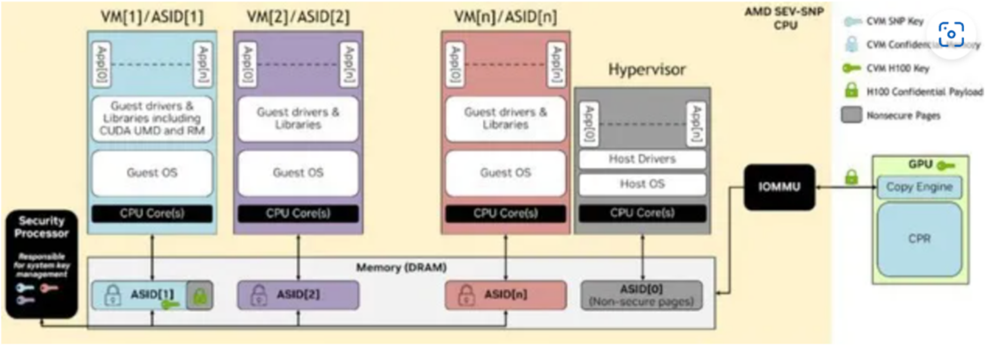
Without confidential computing enabled, the hypervisor has full access to all of system memory and all of GPU memory

With confidential computing guarantees, the hypervisor is blocked from accessing the Confidential VM in system memory and blocked from reading GPU memory

Protecting the GPU in confidential computing mode.

NVIDIA designed its confidential computing capabilities to work with CPUs supporting Confidential VMs (CVMs), ensuring that data remains secure across the CPU and GPU in virtualized and containerized environments. The interaction between the GPU and the CPU is managed through encrypted channels to maintain the data's confidentiality and integrity.

Data transfers between the CPU and the GPU and all command buffers and CUDA kernels are also encrypted and signed. This ensures that data remains secure while in transit and when being processed, preventing unauthorized access and tampering.



Confidential H100 GPU with an AMD SEV-SNP TEE.

The H100 utilizes AI technologies to enhance security operations. For example, AI algorithms help determine the last known clean state of the system, enabling users to restore operations to a secure checkpoint in case of an incident.

The H100's software stack, including the CUDA driver and GPU firmware, is specially designed to support confidential computing. This stack ensures that all GPU operations maintain the highest security standards, from initial boot to application execution.

ANALYSIS

The H100 GPU, with its robust hardware-based Trusted Execution Environment, takes a forward-thinking approach to protecting sensitive data during processing—a critical component often neglected in conventional security frameworks focused mainly on data-at-rest or in-transit.

NVIDIA's implementation of confidential computing on the H100 is based on a hardware root of trust, ensuring that computations are performed in an isolated and secure environment. This methodology enhances data integrity and confidentiality and broadens the GPU's utility across more security-sensitive applications, including those handling personally identifiable information (PII) or proprietary business intelligence.

The H100's support for multiple operational modes—ranging from full security (CC-On) to development-focused settings (CC-DevTools)—provides flexibility while maintaining rigorous security standards. This capability, combined with encrypted data transfers and comprehensive attestation reports, positions the H100 as a GPU with unparalleled security features to protect data regardless of location within a system.

The information contained within this document was believed by NAND Research to be reliable and is provided for informational purposes only. The content may contain technical inaccuracies, omissions, or typographical errors. This document reflects the opinions of NAND Research, which is subject to change. NAND Research does not warranty or otherwise guarantee the accuracy of the information contained within.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.