Guaranteeable Memory: An HBM-Based Chiplet for Verifiable AI Workloads

James Petrie

Abstract

Potential risks from large-scale AI experiments motivate the development of hardware-level guarantees for computational workloads. We propose Guaranteeable Memory, which integrates a "Guarantee Chiplet" beneath High Bandwidth Memory (HBM) stacks within AI accelerators. This chiplet directly observes memory traffic, enabling attestations about executed workloads, including memory snapshots, GPU instructions, or probabilistic checks of the correctness of claimed workloads by verifying random subsets of computations and data transfers. Key advantages of this HBM-based design include direct access to system memory, compatibility with multiple leading accelerators via the HBM standard, and inherent physical tamper-resistance due to its integration geometry. This approach aims to provide trustworthy workload guarantees without relying on the integrity of other system components, including the main accelerator die.

1. Introduction

The development of frontier AI systems could pose large-scale risks (Bengio et al.). Hardware-level mechanisms capable of reliably logging workload details could provide a foundation for auditing to check that AI compute is being used responsibly. Furthermore, automated checks performed directly by the hardware could potentially verify compliance without exposing sensitive intellectual property (IP) to external auditors (Aarne et al.; Kulp et al.).

Nvidia's implementation of confidential computing (Dhanuskodi et al.) is the most developed solution towards the goal of hardware-backed guarantees about AI workloads, though it has limitations. Most significantly, Confidential Computing has a large attack surface that includes CPUs that may be difficult to trust for high-stakes verification.

Systems that could provide "Flexible Hardware-Enabled Guarantees" (flexHEG) (Petrie et al.) have been proposed to address these gaps, by prioritising security and open-source transparency. While flexHEG could be implemented in various ways (e.g., as an IP block on the main accelerator, an external microcontroller, or a modified peripheral like a Network Interface Card (NIC)), this paper focuses on a specific promising design: an open-source Guarantee Chiplet positioned directly beneath the HBM stack.

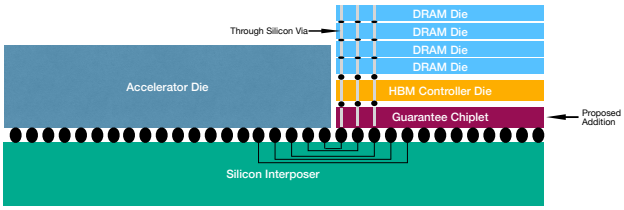

Figure 1. Side view of the accelerator die and HBM, with the addition of a Guarantee Chiplet sitting on the Silicon Interposer underneath the HBM stack.

This location grants the chiplet privileged access to the primary data pathway for AI accelerators. It can observe data read from and written to HBM, perform limited computations for verification, cryptographically attest to data contents or check results, and potentially intervene (e.g., by blocking memory access) to enforce predefined policies. A significant advantage of this architecture is its potential to establish a root of trust largely independent of other system hardware. Because HBM is essential for AI workloads and all HBM traffic must pass through this chiplet, guarantees can be established even if the GPU, CPU, and other components are considered untrusted.

2. Comparison with Alternative Designs

The HBM-based Guarantee Chiplet has advantages over other potential flexHEG implementations, and also some limitations:

- **Centrality:** HBM is the critical high-bandwidth memory interface for virtually all modern AI accelerators. Monitoring this interface provides comprehensive visibility into the primary data movements, without having to trust other hardware.

. Correspondence to: James Petrie <james.petrie@protonmail.com>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

- **Integration Feasibility:** HBM already uses stacked chiplets; adding another layer, while challenging, is conceptually aligned with existing manufacturing processes. Compatibility with JEDEC HBM standards (JEDEC) could allow integration into diverse accelerator designs from different vendors (Nvidia, AMD, Google TPU, Huawei, etc.) relatively late in their design cycles (e.g. Nvidia HBM4 qualification 3-6 months before mass production (Trendforce)).

- **Development Time:** Chiplet design and integration is faster than integrating an IP block into an accelerator die, but slower than software or firmware updates.

- **Security complexity:** only the guarantee chiplet needs to be trusted, which is a much smaller attack surface than the CPU and GPU hardware and software needed for confidential computing.

- **Tamper Resistance:** Microbumps above and below the Guarantee Chiplet are critical for system functioning and delicate, making tampering difficult by default. Optionally, a "secure enclosure" (Petrie et al.) could be added around the system with additional tamper sensors and tamper response mechanisms.

3. Use Cases and Guarantee Mechanisms

The primary function of the Guarantee Chiplet is to provide trustworthy evidence about AI workloads for external verification or automated checks. Given the evolving nature of AI safety and governance, the system is designed to be flexible, allowing secure updates to its verification logic if cryptographically authorized by designated stakeholders.

3.1. Random Verification of Computation and Data Transfer

If the main accelerator's execution cannot be fully trusted, the Guarantee Chiplet can implement probabilistic checks:

- **Data Transfer Verification:** Data transfers between HBM stacks could be verified by comparing a random fraction of the sent and received data afterwards. Guarantee Chiplets could coordinate on which "random" fraction of data to log by choosing based on a secret that was shared beforehand.

- **Computation Verification:** Accelerators could provide "claims" about intended computations. Guarantee Chiplets could then randomly select a small fraction of the claimed output results, securely log the necessary input data before it is overwritten, and later recompute the selected outputs to verify the claim.

3.2. Automated Verification of Guarantees

Guarantee Chiplets could potentially automatically parse workload logs locally to check for compliance with guarantees (e.g., by constructing a compute graph from the workload logs and parsing it to check for properties like whether gradient descent is occurring). Communication between chiplets to coordinate on these checks could be facilitated via the host CPU, using a small, dedicated portion of the HBM address space. Messages exchanged between chiplets would be encrypted and authenticated using shared keys established during a secure initialization phase.

3.3. Guarantee Enforcement

Beyond monitoring and verification, the Guarantee Chiplet could potentially enforce compliance. By temporarily blocking memory access to the HBM stack, it could prevent the accelerator from being used in ways that violate predefined guarantees. This enforcement action could be configured to trigger automatically upon the failure of an internal verification check or if a required, cryptographically signed usage certificate has not been received within a specified number of clock cycles.

3.4. Trusted Logging

Reliable logs of accelerator state play a fundamental role in several AI verification proposals (Shavit; Harack et al.). The Guarantee Chiplet can provide these in several ways:

- **Attested Memory Snapshots:** At random intervals or specific triggers, the chiplet could pause memory access, read the entire contents of the HBM stack, compute a cryptographic hash (e.g., HMAC using a secure key), and either store this attestation locally in secure memory or transmit it externally. This aligns with proposals requiring verifiable memory states for auditing (Shavit).

- **Attested Randomized Memory Operations:** Continuously hashing all HBM traffic (potentially hundreds of GB/s or TB/s) may be computationally infeasible for a resource-constrained chiplet. As an alternative, the chiplet could attest to a randomly selected fraction of all memory read/write operations. While not providing a complete log, this statistically representative sample could be used to verify claims about the overall workload when correlated with prover-supplied snapshots or logs.

- **Attested Memory Region Logging:** Certain memory regions may be particularly critical. For instance, if GPU kernel instructions must reside in HBM to be executed, the chiplet could log and attest to all data written to the specific memory region(s) containing

these instructions. This relatively small data stream could offer significant insights into the nature of the computation.

3.5. Flexibility to Support Future Governance Needs

Recognizing that governance requirements and safety protocols for AI are likely to evolve, the Guarantee Chiplet is designed for flexibility. Its guarantee logic and verification mechanisms can be securely updated post-deployment. The system can be configured to accept firmware or configuration updates only if they are cryptographically signed by a predefined set of authorized stakeholders, potentially requiring a quorum for approval. This ensures the chiplet can adapt to future governance needs without compromising its security foundation.

4. Guarantee Chiplet Design

The Guarantee Chiplet is envisioned as a small, secure processing unit integrated into the HBM interface.

4.1. Physical Integration and HBM Compatibility

- **Location:** Positioned directly beneath the stack of HBM memory dies, with Through-Silicon Vias (TSVs) that connect the HBM stack to the interposer, as shown in Figure 1.
- **HBM Standard:** The JEDEC HBM4 standard (JEDEC) defines the interposer-HBM microbump interface. By designing the Guarantee Chiplet to act as a repeater that conforms to this standard, it would be compatible with all leading accelerators and HBM. If the industry moves towards custom HBM interfaces (), the guarantee chiplet could potentially be integrated into the base die or placed above the base die.
- **Thermal Management:** Adding an active chiplet within the HBM stack introduces an additional heat source. The design must operate within the thermal budget of the HBM assembly, and not significantly disrupt thermal dissipation out the bottom of the HBM stack.
- **Signal Integrity:** Acting as a repeater, the chiplet must maintain high-speed signal integrity for HBM communication.
- **Process Node:** Using a mature process node (e.g., 28nm or older) could enhance auditability and reduce design costs (Kilbourn et al.).

4.2. Chiplet Components

- **Private Key:** A unique, private cryptographic key is essential for signing attestations generated by the chiplet. Several methods could be employed for storing or deriving this key, including permanent storage in on-die antifuses, derivation from a Physical Unclonable Function (PUF), or a combination thereof. One promising PUF approach involves deriving the key based on the relative roundtrip times of external signal wires connected to the chiplet (Deric & Holcomb). This method offers the advantage that the private key is not fully constituted until the chiplet is integrated into the larger system and would likely be destroyed or altered if these signal wires are tampered with, enhancing physical security.
- **Secure Boot:** A secure boot process, potentially leveraging open projects like OpenTitan (lowRISC), ensures the integrity of the chiplet's firmware. The chiplet should have one-time programmable memory to prevent rollback attacks (Ban). Furthermore, the firmware update mechanism itself must be secure, potentially configured to require cryptographic authorization from a designated quorum of stakeholders before accepting new firmware versions.
- **SRAM:** Sufficient SRAM would be needed for caching inputs/outputs during random verification checks. Double the size of a Blackwell L1 cache (256KB x 2 = 512KB) per channel (of which there are 32 per DRAM stack) would likely be sufficient. This would be 16MB total, which would take up about $16mm^2$ with a 28nm process, or roughly 16% of the total chiplet area (approximately $100mm^2$). This SRAM should be possible to write at HBM line-rate (64GB/s per channel).
- **Processing Power:** Sufficient computational power to handle hashing (ideally 1+ GB/s SHA-3 for attested data streams), recomputing a fraction of computations, or parsing data structures like instruction logs. An SoC FPGA could offer a balance of processing power and reconfigurability, and be easier to scan for hardware trojans.
- **Dedicated Cryptographic Hardware:** Asymmetric cryptography hardware (e.g., ECDSA or RSA, or ideally a quantum resistant alternative) is necessary for signing attestation reports using the device private key. This hardware ideally incorporates additional protections against side-channel attacks that could reveal the private key.
- **(Optional) On-die Persistent Memory:** Embedded flash could be used for securely storing firmware, cryptographic keys, logs, and configuration. If on-die flash is not included, this data could potentially be stored off-die and loaded after cryptographic verification during boot.

4.3. Security

The Guarantee Chiplet is intended to be tamper evident (at scale) against nation-state actors attempting covert circumvention, and tamper proof (at scale) against less sophisticated actors seeking to overtly bypass guarantees.

- **Tamper Resistance/Evidence:** The fine-pitch microbump interface inherent to HBM provides a degree of tamper evidence by default. Microbump design choices (e.g solder material) should maximize the difficulty of physically bypassing or tampering with the chiplet without detection.
- **Fault Injection Resistance:** Defenses against attacks that induce faults to bypass security checks (e.g., signature verification). This includes redundant checks and potentially environmental monitoring.
- **Side-Channel Resistance:** Protection against attacks that infer secret information (keys, random numbers, checked operations) from power consumption, timing, or electromagnetic emissions.
- **Supply Chain Security:** Utilizing older, more mature process nodes could facilitate open-sourcing the design and enable more thorough physical inspection and verification, mitigating risks of malicious hardware modifications during manufacturing (Kilbourn et al.).

5. Conclusion

Guaranteeable Memory, implemented via an HBM-based Guarantee Chiplet, offers a potential hardware foundation for verifiable AI workloads. By positioning a secure, open-source chiplet at the critical HBM interface, we can gain trustworthy insights into accelerator activity, largely independent of other system components. This approach enables diverse guarantee mechanisms, from attested memory snapshots to probabilistic verification of computations, with the flexibility to adapt to future needs.

References

Aarne, O., Fist, T., and Withers, C. Secure, governable chips. using on-chip mechanisms to manage national security risks from AI & advanced computing. URL https://www.cnas.org/publications/reports/secure-governable-chips.

Ban, T. Rollback protection in TF-m secure boot — trusted firmware-m unknown documentation. URL https://trustedfirmware-m.readthedocs.io/en/latest/design_docs/booting/secure_boot_rollback_protection.html.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., Newman, J., Ng, K. Y., Okolo, C. T., Raji, D., Sastry, G., Seger, E., Skeadas, T., South, T., Strubell, E., Tramèr, F., Velasco, L., Wheeler, N., Acemoglu, D., Adekanmbi, O., Dalrymple, D., Dietterich, T. G., Felten, E. W., Fung, P., Gourinchas, P.-O., Heintz, F., Hinton, G., Jennings, N., Krause, A., Leavy, S., Liang, P., Ludermir, T., Marda, V., Margetts, H., McDermid, J., Munga, J., Narayanan, A., Nelson, A., Neppel, C., Oh, A., Ramchurn, G., Russell, S., Schaake, M., Schölkopf, B., Song, D., Soto, A., Tiedrich, L., Varoquaux, G., Yao, A., Zhang, Y.-Q., Albalawi, F., Alserkal, M., Ajala, O., Avrin, G., Busch, C., Carvalho, A. C. P. d. L. F. d., Fox, B., Gill, A. S., Hatip, A. H., Heikkilä, J., Jolly, G., Katzir, Z., Kitano, H., Krüger, A., Johnson, C., Khan, S. M., Lee, K. M., Ligot, D. V., Molchanovskyi, O., Monti, A., Mwamanzi, N., Nemer, M., Oliver, N., Portillo, J. R. L., Ravindran, B., Rivera, R. P., Riza, H., Rugege, C., Seoighe, C., Sheehan, J., Sheikh, H., Wong, D., and Zeng, Y. International AI safety report. URL http://arxiv.org/abs/2501.17805.

Deric, A. and Holcomb, D. Know time to die – integrity checking for zero trust chiplet-based systems using between-die delay PUFs. pp. 391–412. ISSN 2569-2925. doi: 10.46586/tches.v2022.i3.391-412. URL https://tches.iacr.org/index.php/TCHES/article/view/9706.

Dhanuskodi, G., Guha, S., Krishnan, V., Manjunatha, A., O'Connor, M., Nertney, R., and Rogers, P. Creating the first confidential GPUs: The team at NVIDIA brings confidentiality and integrity to user code and data for accelerated computing. 21(4):68–93. ISSN 1542-7730, 1542-7749. doi: 10.1145/3623393.3623391. URL https://dl.acm.org/doi/10.1145/3623393.3623391.

Harack, B., Trager, R. F., Reuel, A., Manheim, D., Brundage, M., Aarne, O., Scher, A., Pan, Y., Xiao, J., Loke, K., Adan, S. N., Bas, G., Caputo, N. A., Morse, J. C., Ahuja, J., Duan, I., Egan, J., Bucknall, B., Rosen, B., Araujo, R., Boulanin, V., Lall, R., Barez, F., Alvira, S., Katzke, C., Atamli, A., and Awad, A. Verification for international AI governance.

JEDEC. High bandwidth memory (HBM4) DRAM | JEDEC. URL https://www.jedec.org/standards-documents/docs/jesd270-4.

Kilbourn, Q., Bellemare, S., Bunnie, and Gao, M. ZTEE - trustless supply chains | flashbots writ-

ings. URL https://writings.flashbots.net/ZTEE2-Supply-Chains.

Kulp, G., Gonzales, D., Smith, E., Heim, L., Puri, P., Vermeer, M. J. D., and Winkelman, Z. Hardware-enabled governance mechanisms: Developing technical solutions to exempt items otherwise classified under export control classification numbers 3a090 and 4a090. URL https://www.rand.org/pubs/working_papers/WRA3056-1.html.

lowRISC. lowRISC/opentitan. URL https://github.com/lowRISC/opentitan. original-date: 2019-08-26T16:30:16Z.

Petrie, J., Aarne, O., Ammann, N., and Dalrymple, D. Flexible hardware-enabled guarantees for AI compute. URL http://arxiv.org/abs/2506.15093.

Shavit, Y. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring. URL http://arxiv.org/abs/2303.11341.

Trendforce. [news] SK hynix reported to deliver HBM4 samples to NVIDIA in june, with mass production by q3 2025 | TrendForce news.