ETH Zürich
Energy and Technology Policy Group

Co-Evolution of Hardware and Software in AI Accelerator Platforms

Noah Fehr
Anurag Panda
EU-SPRI 2026

Outline

1. Motivation and Research Question

2. Methods

3. Results

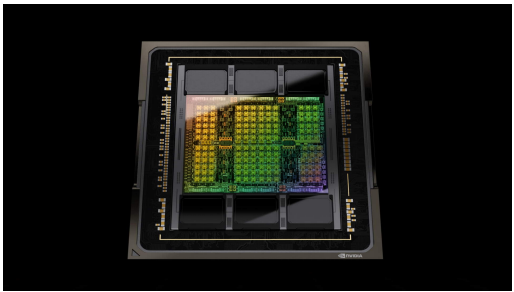
4. Discussion and Policy Implications

Why AI Accelerator Governance Matters

AI accelerators are specialized hardware used for the training and inference of AI models.

- Examples: GPUs, NPUs, and TPUs.
- Governance of AI accelerator innovation is critical for economic competitiveness and national security (Chips Act 2.0; U.S. controls on NVIDIA H20 exports to China).

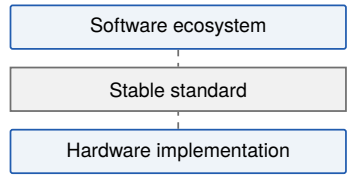
European Commission (2026); NVIDIA (2025).



Current Standards and Platform Literature

Stable standards allow independent development.

- Software can evolve above a stable standard while hardware improves below it.
- This is the modularity logic behind platform theory.
- Standardization literature emphasizes strategic choices about the targeted layer and its governance.



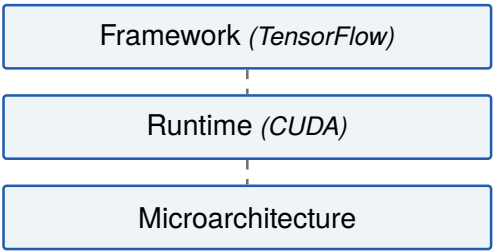
Canonical Wintel logic: Windows software and Intel hardware develop independently around a stable standard

Blind (2004); Gawer & Cusumano (2002); Bresnahan & Greenstein (1999).

Where AI Accelerator Innovation Happens

Accelerator innovation happens across the stack.

- Frameworks expose new workloads and execution patterns.
- Runtimes translate software abstractions into hardware demands.
- Microarchitectures adapt around scheduling, memory, and compute primitives.



Research Questions

To better understand the nature of innovation in AI accelerators, we ask the following:

1. Are platform-specific software shocks associated with directional changes in hardware innovation within the same platform?

2. Are ecosystem-level software developments associated with systematic changes in hardware innovation across accelerator platforms?

Outline

1. Motivation and Research Question

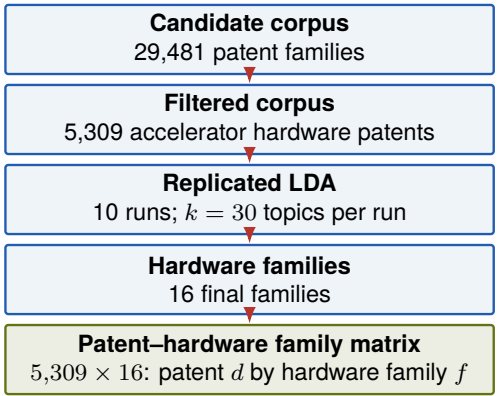
2. Methods

3. Results

4. Discussion and Policy Implications

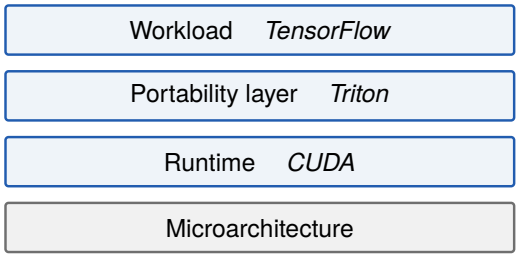
Method 1: Building the Patent–Hardware Family Matrix

- **Patent claims corpus:** U.S. patents, 2006–2026; CPC codes and keyword search (OECD, 2025).
- **Relevance filter:** retain accelerator hardware design patents.
- **Preprocessing.**
- **Replicated LDA runs:** repeated estimation to address model stochasticity.
- **Topic aggregation:** group recurring topics into stable hardware families (Hoyle et al., 2022).



Method 2: Software Shocks and Event Windows

- Software shocks at multiple layers.
  - CUDA: platform runtime.
  - TensorFlow: ecosystem-level framework.
- Observational: shocks are difficult to isolate in the dense 2016–present software window.
- Interpreted through mechanism, not as clean causal identification.



Outline

1. Motivation and Research Question

2. Methods

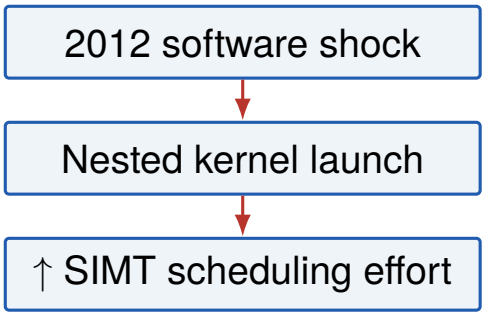
3. Results

4. Discussion and Policy Implications

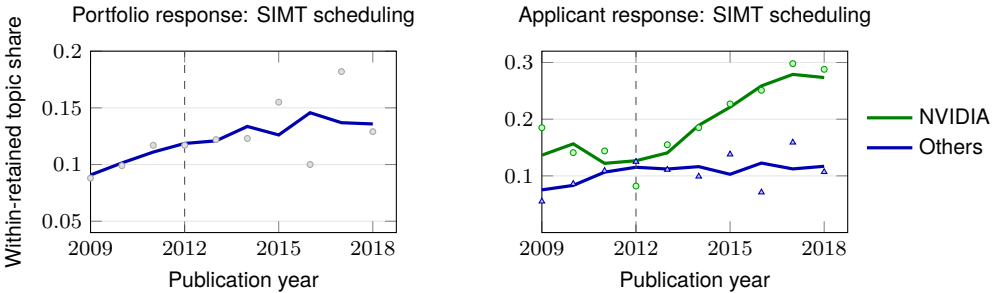
RQ1: CUDA Dynamic Parallelism

Mechanism. CUDA Dynamic Parallelism (2012) lets a running kernel launch follow-on kernels without returning to the host program. Nested launch complicates scheduling and synchronization.

- **SIMT** means single instruction, multiple threads.
- NVIDIA: ↑ SIMT scheduling effort.
- Firm-specific same-platform response.



RQ1 Result: Same-Platform Coupling

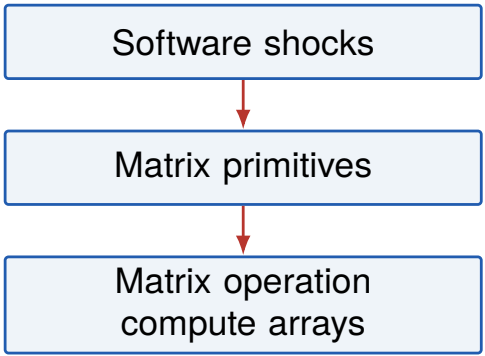


Within-platform coupling: CUDA Dynamic Parallelism is followed by increased *SIMT scheduling* effort at NVIDIA.

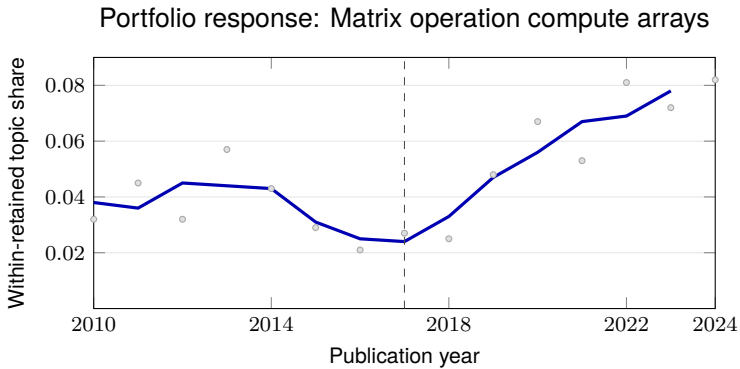
RQ2: Software-Exposed Matrix Operations

Mechanism. Several software shocks arrive close together and point in the same hardware direction.

- Examples: TensorFlow/XLA (2017) and CUDA 9 Tensor Cores/WMMA (2017).
- These shocks expose matrix operations more directly and increase demand for matrix operation compute arrays.



RQ2 Result: Ecosystem-Level Alignment

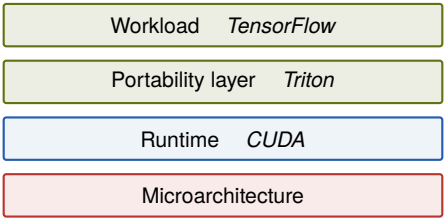


Ecosystem-level alignment: software shocks exposing matrix primitives coincide with rising *matrix operation compute array* emphasis.

Outline

1. Motivation and Research Question
2. Methods
3. Results
4. Discussion and Policy Implications

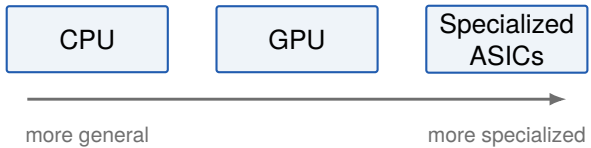
Standardization, but where?



Standards should promote competition via portability while enabling architectural experimentation below.

Specialized ASICs Raise The Stakes

- CPU → GPU → specialized ASICs.
- Co-development of hardware and supporting software enables innovation.
  - More specialization; more architectural freedom.
  - AMD chiplets vs. Cerebras wafer-scale design.



Policy Implications

- When hardware and software co-evolve, the layer at which standards are set shapes both competition and technical search.
- Low-level standardization may improve interoperability but can prematurely narrow architectural search, especially as accelerator architectures become more specialized.
- Compute policy should prioritize portability at higher layers while preserving architectural experimentation below.

Noah Fehr
Anurag Panda
noah.fehr@gess.ethz.ch

ETH Zürich
Energy and Technology Policy Group

https://epg.ethz.ch

Supporting References I

- Blind, Knut. 2004. *The Economics of Standards: Theory, Evidence, Policy*. Cheltenham, UK: Edward Elgar Publishing.
- Bresnahan, Timothy F., and Shane Greenstein. 1999. “Technological Competition and the Structure of the Computer Industry.” *Journal of Industrial Economics* 47(1): 1–40. https://doi.org/10.1111/1467-6451.00088
- European Commission. 2026. *Chips Act 2.0*. https://digital-strategy.ec.europa.eu/en/policies/chips-act-2
- Gawer, Annabelle, and Michael A. Cusumano. 2002. *Platform Leadership: How Intel, Microsoft, and Cisco Drive Industry Innovation*. Boston, MA: Harvard Business School Press.

Supporting References II

- Hennessy, John L., and David A. Patterson. 2019. “A New Golden Age for Computer Architecture.” *Communications of the ACM* 62(2): 48–60. https://doi.org/10.1145/3282307
- Hoyle, Alexander, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. “Are Neural Topic Models Broken?” *Findings of ACL: EMNLP 2022*, 5321–5344. https://aclanthology.org/2022.findings-emnlp.390/
- NVIDIA Corporation. 2025. “Current Report on Form 8-K.” Filed April 15, 2025. https://www.sec.gov/Archives/edgar/data/1045810/000104581025000082/nvda-20250409.htm
- OECD. 2025. *Identifying Emerging AI Technologies Using Patent Data: A Semi-Automated Approach*. OECD Publishing, Paris. https://doi.org/10.1787/d17e9a1a-en