ETH Zürich
Energy and Technology Policy Group

Co-Evolution of Hardware and Software in AI Accelerator Platforms

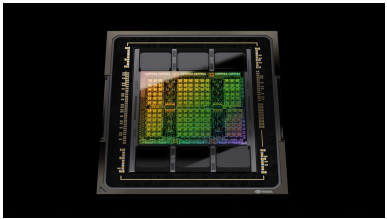
Noah Fehr
Anurag Panda
ISS 2026

Outline

Why AI Accelerator Governance Matters

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

- GPUs, NPUs, and TPUs.
- Governance of AI accelerator innovation is critical for economic competitiveness and national security.
- Growing appetite in the U.S. and Europe to intervene not only in chip supply, but in the structure of AI accelerator ecosystems.
  - Chips Act 2.0; NVIDIA H20 export controls.

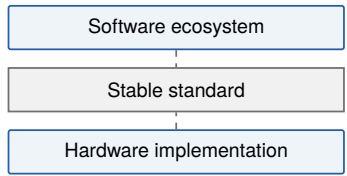


Zúñiga et al. (2024); European Commission (2026); NVIDIA (2025).

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.
- Standards function as technical infrastructure, shaping innovation incentives and search.

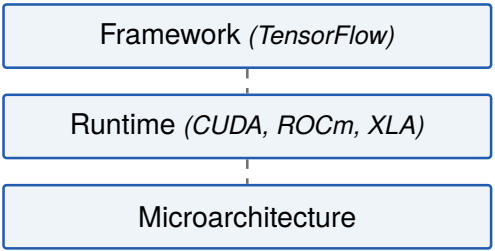


Canonical Wintel logic: Windows software and Intel hardware develop independently around a stable standard

Blind (2004); Gawer & Cusumano (2002); Bresnahan & Greenstein (1999); Tassey (2000).

AI 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 execution primitives.
- Runtime ecosystems remain horizontally incompatible, limiting portability across platforms.



Simcoe & Watson (2019).

Research Questions

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

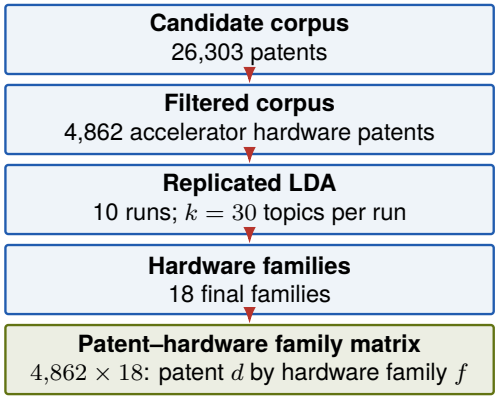
1. Are platform-specific runtime changes associated with directional changes in hardware innovation within the same platform?

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

Outline

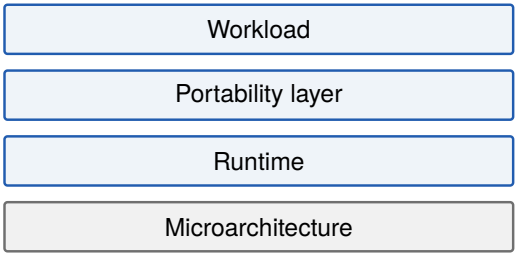
Method 1: Building the Patent–Hardware Family Matrix

- **Patent claims corpus:** U.S. patent documents, 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 Changes and Hardware Alignment

- **Software change events:** framework, runtime, and workload changes that alter what hardware must efficiently support.
- Observational: directional alignment, not causal identification.
  - Software may embody new hardware capabilities, or hardware may respond to new software demands.

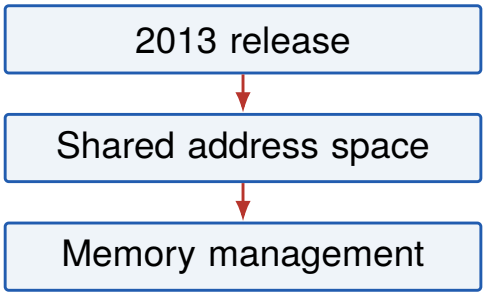


Outline

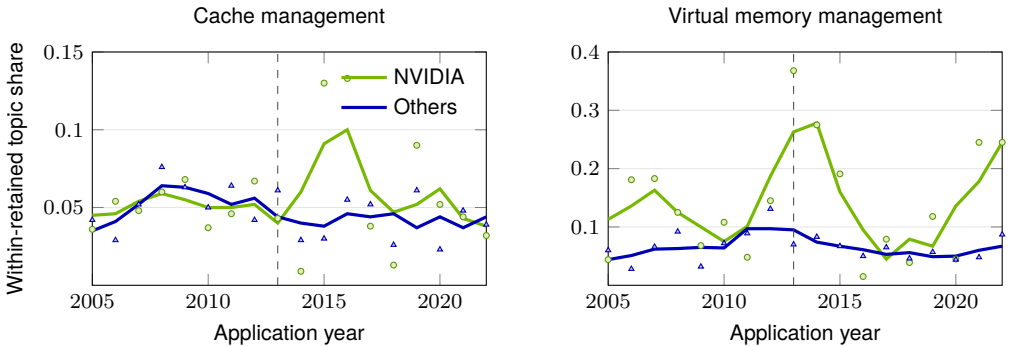
RQ1: CUDA Unified Memory

Mechanism. CUDA Unified Memory exposed a shared address space across CPU and GPU memory.

- A shared address space creates address-translation and page-fault handling questions.
- CPU–GPU access raises cache-coherence, placement, and migration problems.
- Specific to the CUDA runtime and NVIDIA platform.



RQ1 Result: NVIDIA-Specific Hardware Alignment

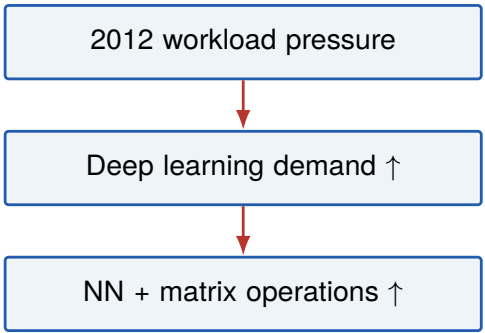


Within-platform codevelopment: after the 2013 CUDA Unified Memory release, NVIDIA deviates from other applicants in related memory-management families.

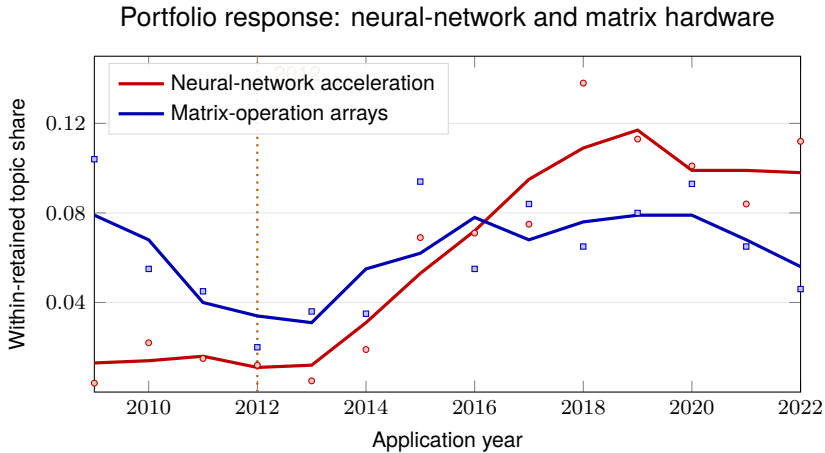
RQ2: Neural Network Workload Pressure

Mechanism. AlexNet/ImageNet result marked a visible inflection point in deep-learning workloads, changing expectations about the scale and type of computation accelerators would need to support.

- This is workload pressure, not a single platform API.
- Expected hardware focus: neural-network acceleration and matrix-operation arrays.



RQ2 Result: Workload-Layer Alignment



Workload-layer alignment: the 2012 deep-learning demand shift is followed by rising hardware attention to neural-network acceleration and matrix-operation arrays.

High-Level Results: Layer Matters

RQ1: Runtime changes matter when they expose bottlenecks

Platform-specific runtime shifts are followed by applicant-specific hardware deviation in the expected families.

RQ2: Hardware alignment is strongest at the workload layer

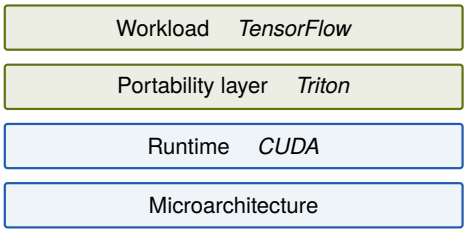
Workload shifts show broad hardware alignment; portability-layer changes show weaker, less systematic responses.

Outline

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

Policy Implication: Portability Above, Diversity Below

- Technology policy targeting AI accelerator capabilities should account for runtime–microarchitecture coupling and codevelopment.
- **Potential policy lever:** support innovation diversity and competition through portability higher in the stack, while avoiding premature lock-in of runtimes or microarchitectural choices.



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.
- Nylund, Petra A., and Alexander Brem. 2023. “Standardization in Innovation Ecosystems: The Promise and Peril of Dominant Platforms.” *Technological Forecasting and Social Change* 194: 122714. https://doi.org/10.1016/j.techfore.2023.122714

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

Supporting References III

- Simcoe, Timothy, and Jeremy Watson. 2019. “Forking, Fragmentation, and Splintering.” *Strategy Science* 4(4): 283–297. https://doi.org/10.1287/stsc.2019.0094
- Tassey, Gregory. 2000. “Standardization in Technology-Based Markets.” *Research Policy* 29(4–5): 587–602. https://doi.org/10.1016/S0048-7333(99)00091-8
- Zúñiga, Nicholas, Saheli Datta Burton, Filippo Blancato, and Madeline Carr. 2024. “The Geopolitics of Technology Standards: Historical Context for US, EU and Chinese Approaches.” *International Affairs* 100(4): 1635–1652. https://doi.org/10.1093/ia/iiae124

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.