Title: Hardware-software coupling in AI accelerator platforms: patent analysis of software-layer shocks and hardware innovation

Authors: Noah Fehr, Anurag Panda* (anurag.panda@gess.ethz.ch), Tobias Schmidt
*Presenter

Affiliation: Energy Technology and Policy Group, ETH Zürich, Zürich, Switzerland

Abstract:

Artificial intelligence (AI) accelerators, including graphics processing units (GPUs), tensor processing units (TPUs), and related domain-specific architectures, are core infrastructure for modern economies and national security. The geographic concentration of production and excludability of these chips has motivated industrial policy, including the U.S. CHIPS and Science Act, the EU Chips Act, and China's National Integrated Circuit Industry Investment Fund (Sastry et al., 2024; OECD, 2023; NIST, 2024). While these interventions primarily target fabrication, the development of accelerator capability requires not only manufacturing but also a complex design process shaped by interactions between hardware architecture and low-level software stacks.

In response to the underspecification of policy interventions targeting chip design, this paper conceptualizes AI accelerator design as platforms where innovation emerges through tight coupling between hardware architecture and low-level software systems. While existing platform innovation literature emphasizes modular interfaces and complementor ecosystems (Gawer & Cusumano, 2014; Jacobides et al., 2018), innovation in AI accelerators occurs through cross-layer co-design, rather than modular development across a stable boundary. Due to this coupling, shocks in one layer can fundamentally alter the direction of innovation in another layer. This distinction has critical implications for industrial policy and the political economy of technological leadership.

This paper asks: are discrete software-layer shocks in AI accelerator platforms associated with systematic changes in hardware innovation within and across platforms? We characterize software shocks as within-platform (e.g., new capabilities in CUDA), or ecosystem-wide (e.g., the public release of TensorFlow). Although innovation across layers is inherently bidirectional, we treat software-layer changes as observational shocks to examine corresponding shifts in hardware design trajectories. To measure these trajectories, we construct a dataset of AI accelerator hardware design using U.S. patent data via keyword and classification code search (OECD, 2025), followed by a large language model-based classification pipeline (Asirvatham et al., 2026). To measure the direction of innovation, we apply Latent Dirichlet Allocation topic modeling to patent claims, allowing documents to be assigned to multiple topics (Blei et al., 2003; Kelly et al., 2018). The analysis tracks changes in topic composition, entry, and similarity across firms and over time.

Preliminary results indicate cross-layer influence. Software shocks originating within a platform are associated with the strongest shifts in hardware innovation within that platform, with weaker but occasionally detectable spillovers to competitors. Higher-level software shocks are associated with broader changes in hardware innovation across platforms, suggesting a broader reorientation of hardware

design. We also document heterogeneity in topic composition across platforms, indicating differentiated technological trajectories. These findings provide empirical evidence of the tight coupling between low-level software and hardware architecture, supporting the view of AI accelerator design as platforms.

For policymakers, the results highlight tradeoffs among performance gains from tightly coupled innovation, cross-platform portability for competition, and the preservation of innovation diversity. Heterogeneity in technological trajectories suggests that coupling not only improves performance but also supports differentiated innovation paths. Enabling cross-platform portability, rather than enforcing cross-layer compatibility, may strengthen competition without forcing convergence towards a single design. However, policies promoting openness may accelerate capability diffusion across borders. Overall, this study provides a more granular foundation for industrial policy targeting AI accelerator design.

Bibliography

Asirvatham, H., Mokski, E., & Shleifer, A. (2026). GPT as a measurement tool. NBER Working Paper No. 34834. National Bureau of Economic Research, Cambridge, MA. https://www.nber.org/papers/w34834

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. https://jmlr.csail.mit.edu/papers/v3/blei03a.html

Gawer, A., & Cusumano, M. A. (2014). Industry platforms and ecosystem innovation. Journal of Product Innovation Management, 31(3), 417–433. https://doi.org/10.1111/jpim.12105

Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems. Strategic Management Journal, 39(8), 2255–2276. https://doi.org/10.1002/smj.2904

Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2018). Measuring technological innovation over the long run. NBER Working Paper No. 25266. National Bureau of Economic Research, Cambridge, MA. https://www.nber.org/papers/w25266

National Institute of Standards and Technology (NIST). (2024). Chips for America: Federal programs supporting the U.S. semiconductor supply chain and workforce. https://www.nist.gov/system/files/documents/2024/04/22/Fact%20Sheet%20-%20Federal%20Incentives-updated-508C.pdf

OECD. (2023). A blueprint for building national compute capacity for artificial intelligence (OECD Digital Economy Papers No. 350). OECD Publishing. https://www.oecd.org/en/publications/a-blueprint-for-building-national-compute-capacity-for-artificial-intelligence_876367e3-en.html

OECD. (2025). *Identifying emerging AI technologies using patent data: A semi-automated approach*. OECD Publishing. https://doi.org/10.1787/d17e9a1a-en

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., & Coyle, D. (2024). Computing power and the governance of artificial intelligence. Centre for the Governance of AI. https://arxiv.org/pdf/2402.08797