CS550 Project Abstract

Julien de Castelnau
julien.decastelnau@epfl.ch

Cemalettin Cem Belentepe
cemalettin.belentepe@epfl.ch

November 2023

Systolic arrays are digital circuits composed of many simple interconnected processing elements, which compute individual operations using data from their neighbors to form a larger result. As compared to traditional architectures (CPU, GPU, etc.), systolic arrays minimize memory traffic by passing intermediate results within the device instead of using a main memory. As a result, they have been highly successful in accelerating parallel digital signal processing algorithms such as matrix multiplication and Fast Fourier Transform (FFT) [4]. Some notable designs are Google's TPU [3], and Nvidia's Tensor Cores [1], which both use systolic architectures to accelerate matrix multiplication for deep learning applications.

However, despite the increasing complexity of these designs - with Google's latest TPU reaching 22 billion transistors [2], similar to that of a high end CPU - little research has been devoted to the verification of systolic arrays. In our case study, we intend to build a simple but extensible hardware accelerator based on a systolic array, and apply the inductive techniques outlined in [5] to prove its functionality. We will use the Stainless framework to formally describe the semantics of the hardware, and create tools to compile our Stainless specification to synthesizable Verilog HDL.

Paper of choice: [5]

References

[1] Jack Choquette, Wishwesh Gandhi, Olivier Giroux, Nick Stam, and Ronny Krashinsky. Nvidia a100 tensor core gpu: Performance and innovation. *IEEE Micro*, 41(2):29–35, 2021.

[2] Norman P. Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and David Patterson. Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings, 2023.

[3] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, and Jonathan Ross. In-datacenter performance analysis of a tensor processing unit. 2017.

[4] Hyesook Lim and E.E. Swartzlander. Efficient systolic arrays for fft algorithms. In *Conference Record of The Twenty-Ninth Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 141–145 vol.1, 1995.

[5] N. Ling, T. Shih, and J. Huang. Inductive techniques for formal verification of systolic array designs in dsp applications. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 573–576 vol.5, 1992.