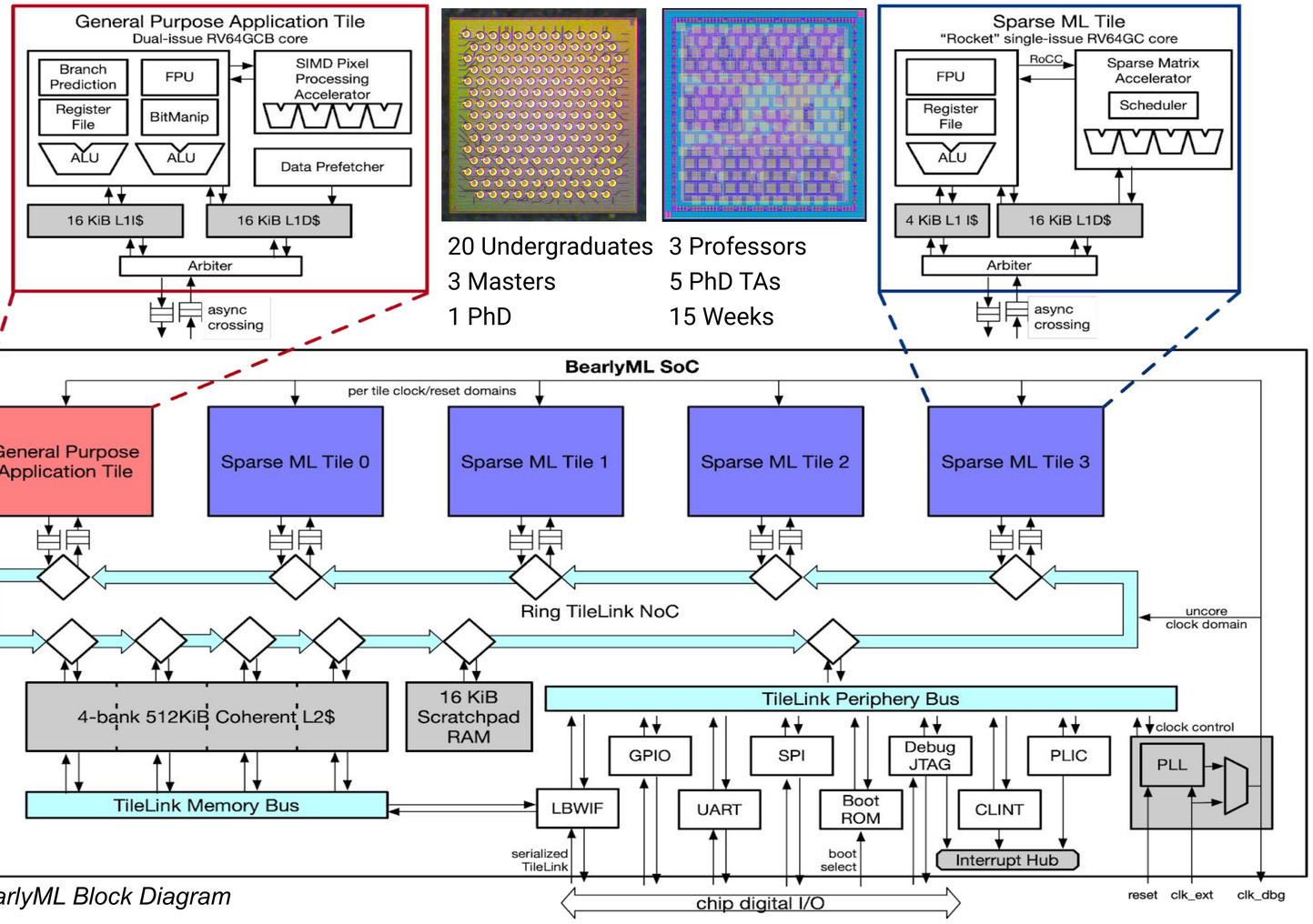
A Heterogeneous RISC-V SoC for ML Applications in Intel 16 Technology

Yufeng Chi, Franklin Huang, Raghav Gupta, Ella Schwarz, Jennifer Zhou, Reza Sajadiany, Animesh Agrawal, Max Banister, Michelle Boulos, Jason Chandran, Jessica Dowdall, Leena Elzeiny, Claire Gantan, Anthony Han, Roger Hsiao, Chadwick Leung, Edwin Lim, Jose Rodriguez, Tushar Sondhi, Mitchell Twu, Rongyi Wang, Mike Xiao, Ruohan Yan, Paul Kwon, Zhaokai Liu, Jerry Zhao, Bob Zhou, Ali Niknejad, Kristofer Pister, Borivoje Nikolić

BearlyML



20 Undergraduates
3 Masters
1 PhD

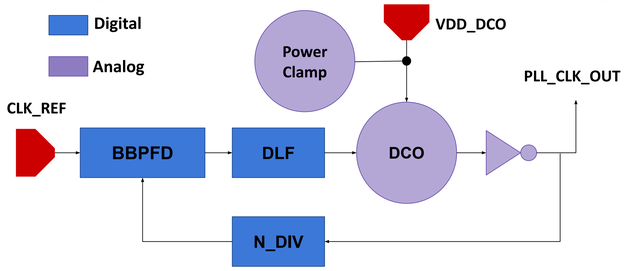
3 Professors
5 PhD TAs
15 Weeks

BearlyML Block Diagram

- **2mm x 2mm RISC-V SoC in Intel 16 finFET process**
- Network-on-Chip system bus with unidirectional torus topology
  - First tapeout of Chipyard Constellation
- 1 experimental RV64GCB dual-issue core with DSP extensions
  - Prefetcher, Bit-manipulation, Pixel SIMD RoCC (Peak 0.6 GOPS)
- 4 specialized RV64GC Rocket cores
- 512 KiB 8-way-associative second-level cache system (4 banks)
- 16 KiB on-chip scratchpad memory
- Tile-specific clock domain with configurable MMIO clock tree

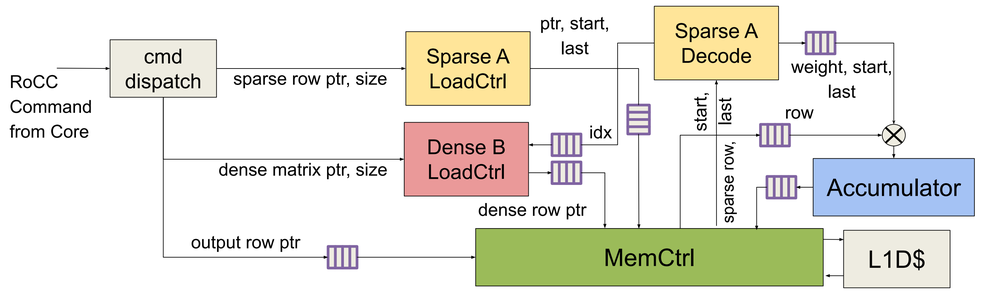
Clocking

- Off-chip 100MHz clock
- On-chip PLL (0.1~1.2 GHz) with DCO designed by Sean Huang and generated with the Berkeley Analog Generator (BAG)
- Configurable via MMIO and exposed on debug pin



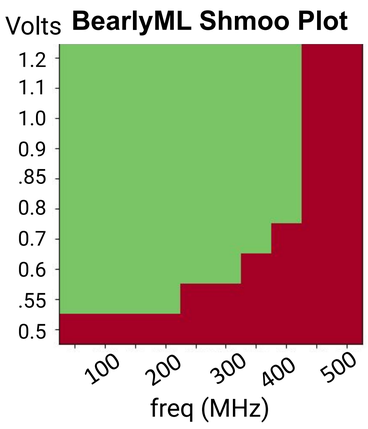
ML Inference Tile

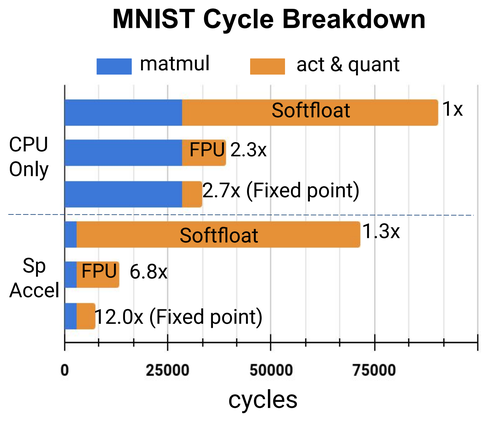
- Rocket Core: 5-stage in-order core, 16KiB L1 D-cache, 4KiB L1 I-cache
- Custom Sparse-Dense Matrix Multiplication Accelerator
  - Uses Rocket Custom Coprocessor Interface (RoCC)
- Peak Theoretical/Achievable Throughput: **4 GOPS / 2.6 GOPS**



Bringup & Demo

- Test PCB are designed by students in the following bringup class
  - Tested max operating frequency: 400 MHz core clock
- MNIST Demo
  - 32MiB off-chip Flash and 16KiB of on-chip scratchpad
  - 91.23% and 95.42% accurate MNIST, 196x32x10 FC int8 neural net, up to 12x acceleration





Tools and Acknowledgements

We would like to acknowledge and thank:
Apple for supporting integrated circuit engineering classes at UC Berkeley EECS department through Apple's New Silicon Initiative program; Intel University Shuttle Program for donating fabrication and packaging; NSF CCRI ENS Chipyard Award #2016662 and Scalable Asymmetric Lifecycle Engagement (SCALE).