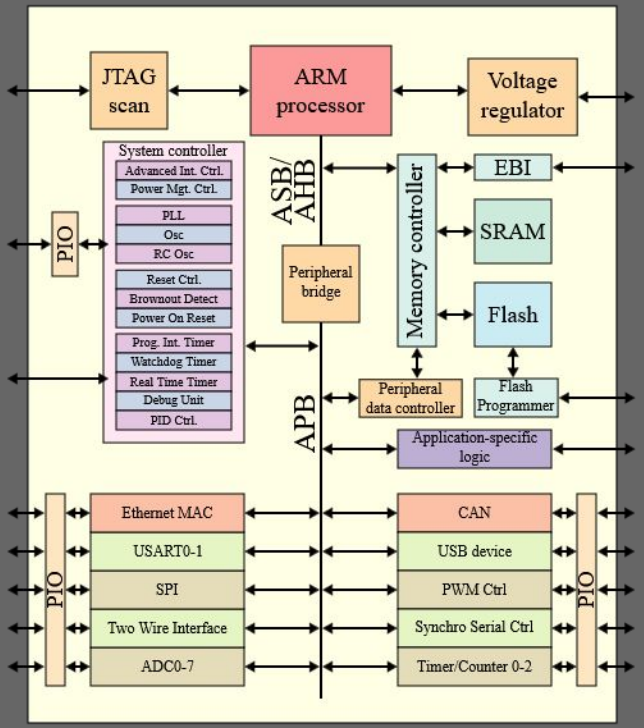
CPE 470 - SoC Design



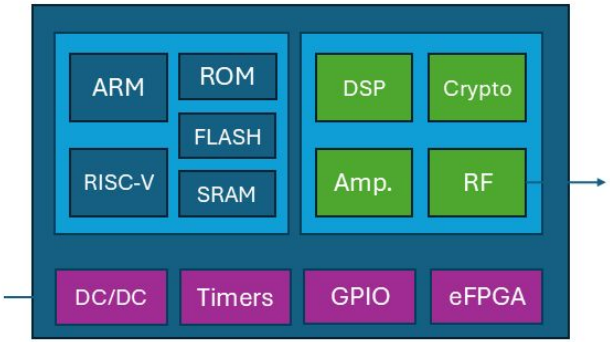
SoC ASIC Design

Glossary

SoC: System on Chip

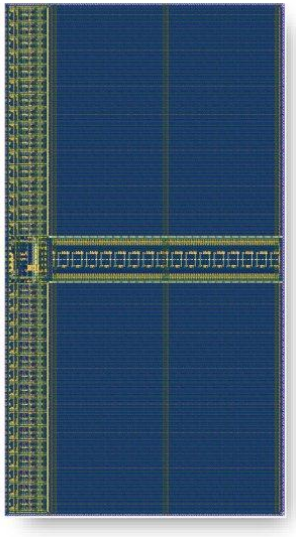
IP: Intellectual Property

- **SoC**: both a category and a methodology
  - **Category**: any system that integrates a processor with its peripherals
    - memory, IO, storage, flash, DSP, accelerators, etc.
    - Examples: Apple A-Series, Qualcomm Snapdragon
    - As opposed to a chip that is just a CPU
  - **Methodology**: a way of designing chips with a focus on integration, design reuse, and **IP** blocks
    - License or use pre-made and tested blocks
    - Spend development time on high performing interconnects
    - Don't reinvent the wheel of a processor



Intellectual Property

- **Hard IP:** fully laid out design
  - Analog (DACs/ADCs) and RAM will always be hard IP
  - Specific to one PDK, not transferable
  - User has to trust vendor
- **Soft IP:** RTL code for a design
  - User has to synthesize, PNR, etc
  - Vendor has to carefully control access, could get leaked
- **Firm IP:** post-synthesis netlist of design
  - User still has to PNR, but can expect similar performance
  - Less risky for vendor
- Many companies exist primarily on licensing IP
  - Synopsys and Cadence of course
  - Processors → ARM
  - DRAM → Lattice, Altera

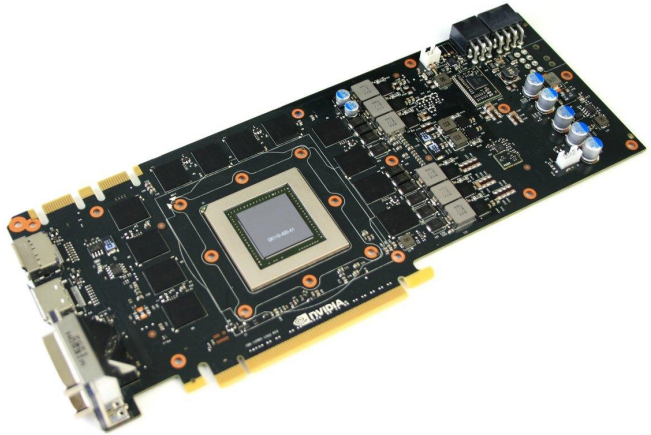


SRAM Hard IP

A Pre-SoC era

- Entire Die area of a chip is devoted to CPU (+ cache and others)
  - Requires external short and long term memory
    - DRAM for short term
    - Drives for long term
  - Has to expose entire main memory bus to motherboard
    - 1700 + pins
  - Cannot function on its own → is a piece of a system
- Not everything has to be an SoC!
  - Sacrifices die area that could be used for compute
    - Ex: graphics cards need tons of die area
    - Still might use SoC design principles
  - No upgradability

Not SoCs!

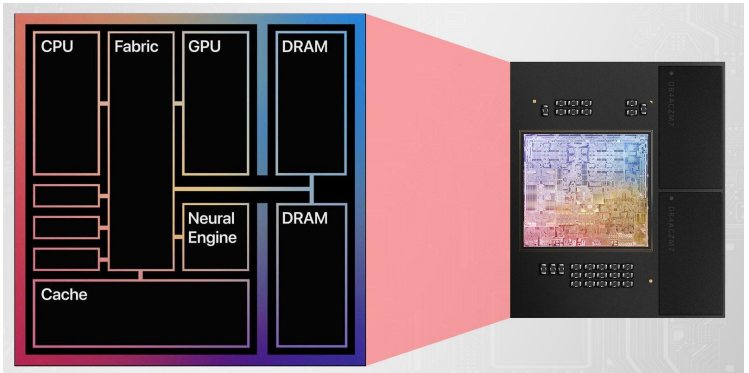


Why SoC?

- Physical Size Constraints
  - Integrate all necessary components in one package, less space
- Unit Cost
  - Long run unit cost is quite low despite high up-front cost
  - Cheaper than buying each individual components and manufacturing
- Performance
  - Cache Misses rarely leave the chip, memory is local
  - As opposed to:
    - Bus ⟶ IO Pin ⟶ Wirebond ⟶ Package Pin ⟶ PCB ⟶ RAM Pin ⟶ Ram Wirebond ⟶ RAM
    - And back

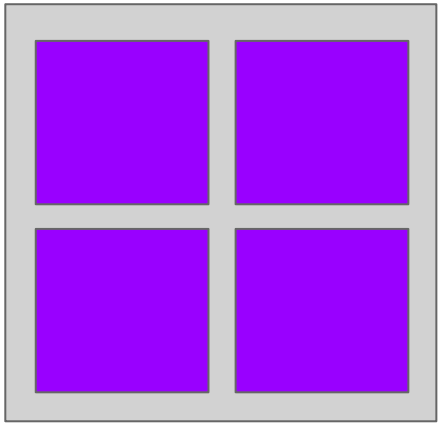
Size Constraints: ESP32 Pico SoC

Performance Benefits: Apple M Series

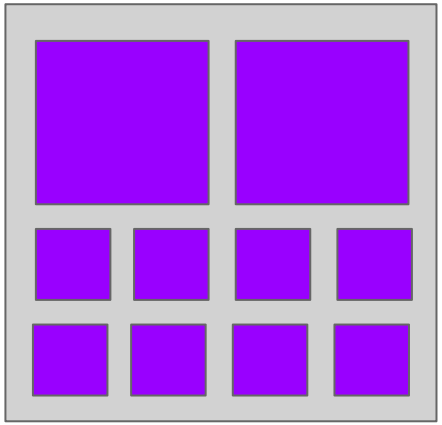


Homogeneous vs Heterogeneous SoCs

- **Homogeneous:** composed of many of the same processor
  - Simpler multithreading as all cores share ISA

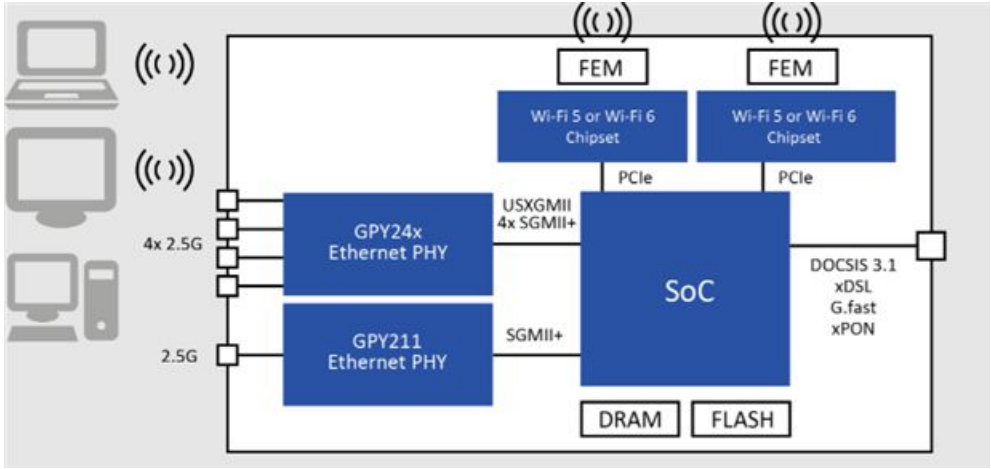


- **Heterogenous:** composed of different kinds of processors (same or different ISAs)
  - Can optimize for power by using smaller cores
  - Different ISAs complicates multithreading



SoC Applications: Networking

- High performance networking pushes to Terabits per second
  - How do you get 10^{12} bits to in and out of a system in a second?
- Packets have to go through many layers to get from wire to chip
  - Wire → Transceiver → PHY → MAC → CPU
- SoC Design for networking more tightly couples layers with the CPU
  - Bring MAC or even PHY into the SoC

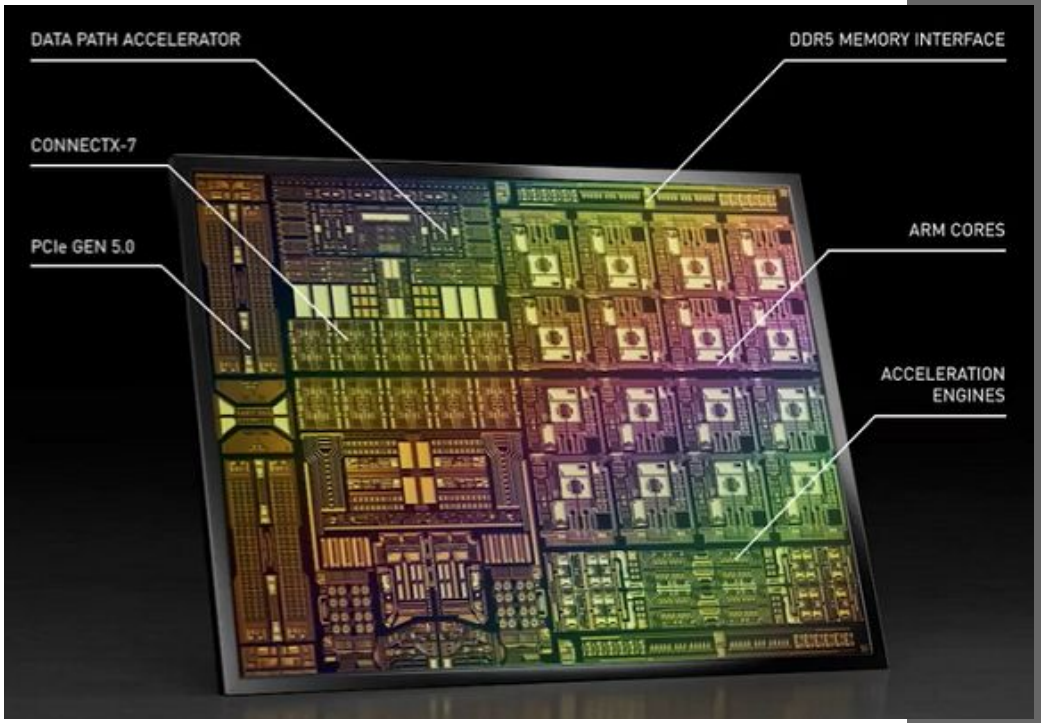


SoC Example: NVIDIA BlueField3

Glossary

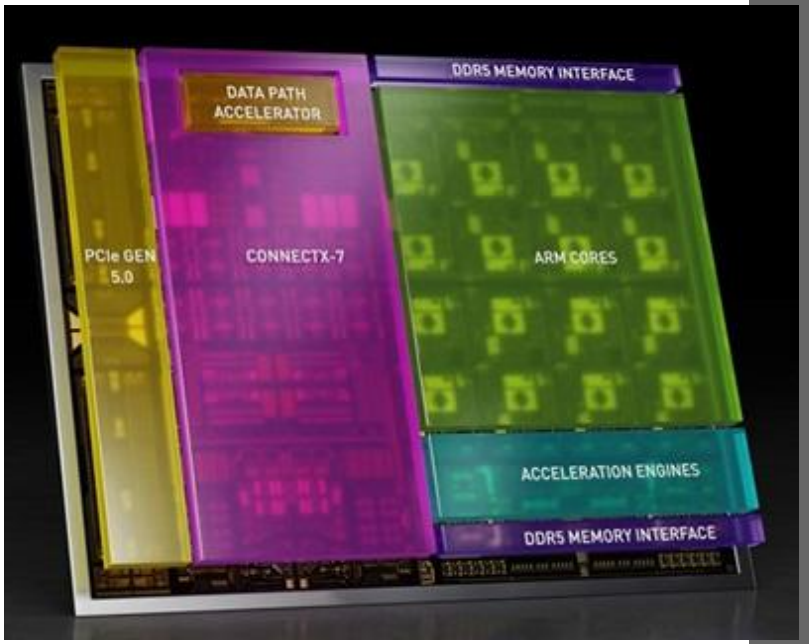
DPU: Data Processing Unit

- 400 Gbps **DPU**
  - High Speed Packet Processor
- Follows SoC Design Principles
  - Processor: Arm IP
  - ConnectX-7 Switch: Mellanox IP
    - NVIDIA bought Mellanox
  - Accelerators and connectivity: likely mix of own and purchased IP
- Accelerators
  - Cryptographic
    - AES, SHA, ECC, Diffie Hellman
  - Networking
    - IP, TCP, UDP Offload, etc.
- Heterogeneous Architecture
  - 16 Arm cores
  - ~200 Tiny Risc-V cores
- Connectivity
  - DDR5 & PCIE 5



BlueField3 Takeaways

- So NVIDIA bought/reused a bunch of IP. What is the big deal?
  - The work is in the interconnects!
  - Need extremely high throughput bus architecture to connect all components at 0.4 Tbps
- How do you provide data to an accelerator with 200 RISC-V-32 Cores
  - Cannot use classical techniques of having a 200-port memory
  - Have to move to more complex bus architectures and multi-level caching
- Integration is the primary focus of an SoC designer!



Example: Qualcomm DSP

Glossary

TCM: Tightly Coupled Memory, cache that is loaded manually by the user

- Qualcomm already had good *modem* chips…
- … but how can they break into the mobile market?
- Generalized the modem core to be multipurpose
  - Hexagon DSP:
  - With TCM you can have application/specific code at L2 Cache speeds!
    - aDSP: Application Specific
    - mDSP: Modem specific
    - … later NPU essentially just used this!

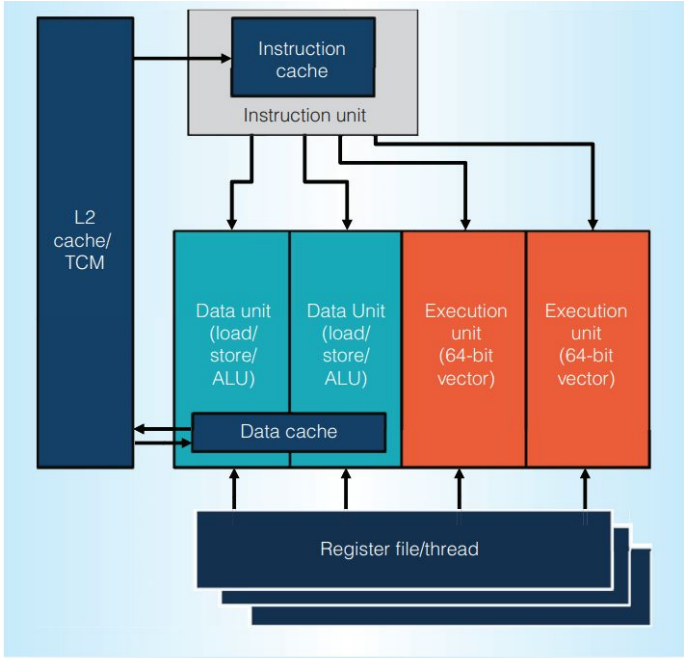


Figure 3. Hexagon block diagram. The architecture features a four-wide very long instruction word (VLIW) with dual load/store and dual single-instruction, multiple-data (SIMD) execution units and supports hardware multithreading.

Example: Hexagon NPU

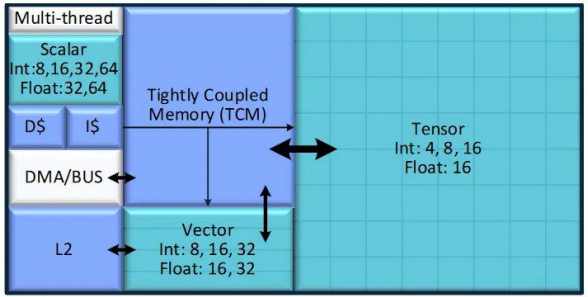
- Qualcomm wanted to start sniffing that AI Dough
  - How can they pull an EA and Ctrl+C Ctrl+V their existing stuff?
  - Ans: Just use the DSP (it's pretty close) but add SIMD and Vector/Tensor support
  - TCM acts as a way to cushion scatter/gather operations
- Conclusion: Use memory in related regions! The *memory* gave them performance.

Glossary

Gather/Scatter: Vectors are often transformed from data that is not contiguously in memory

Hexagon NPU

- Processor executing 3 instruction sets:
  - Scalar: For control flow and general purpose
  - Vector: General purpose data-parallel compute
  - Tensor: Matrix multiply and convolutional layer
  - Over multiple threads using shared memories (core local & cached DDR)
- DSP features:
  - VLIW, hardware looping
  - Targets DSP and compute-heavy workloads
- CPU-like features:
  - Virtual → Physical translation, security, caching
  - Branching (call/return/indirect), exceptions, interrupts
  - Conventional software tools (including LLVM)



- Maximized efficient single-core performance
  - Make the most of resources

References

- https://www.synopsys.com/blogs/chip-design/system-on-chip-design.html
- https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/documents/datasheet-nvidia-bluefield-3-dpu.pdf
- https://chipsandcheese.com/p/qualcomms-hexagon-dsp-and-now-npu