SOC Design
SOC Introduction

Modern SOC Building Blocks and Main Features

Jiin Lai

Why Create a Custom System-on-chip (SoC)?

- Reduced cost
- Reduced PCB area and volume
- Increased performance and reduced power consumption
- Product differentiation
- SoCs integrate a range of IP types: processors, custom processors, accelerators, on-chip memories, peripherals and interfaces, Interconnect … etc.

parallel - seq

latency ↑

power ↑

A Modern Arm System-on-chip (SoC)

A core: big/little
R: real-time
M: MCU - embedded IOT

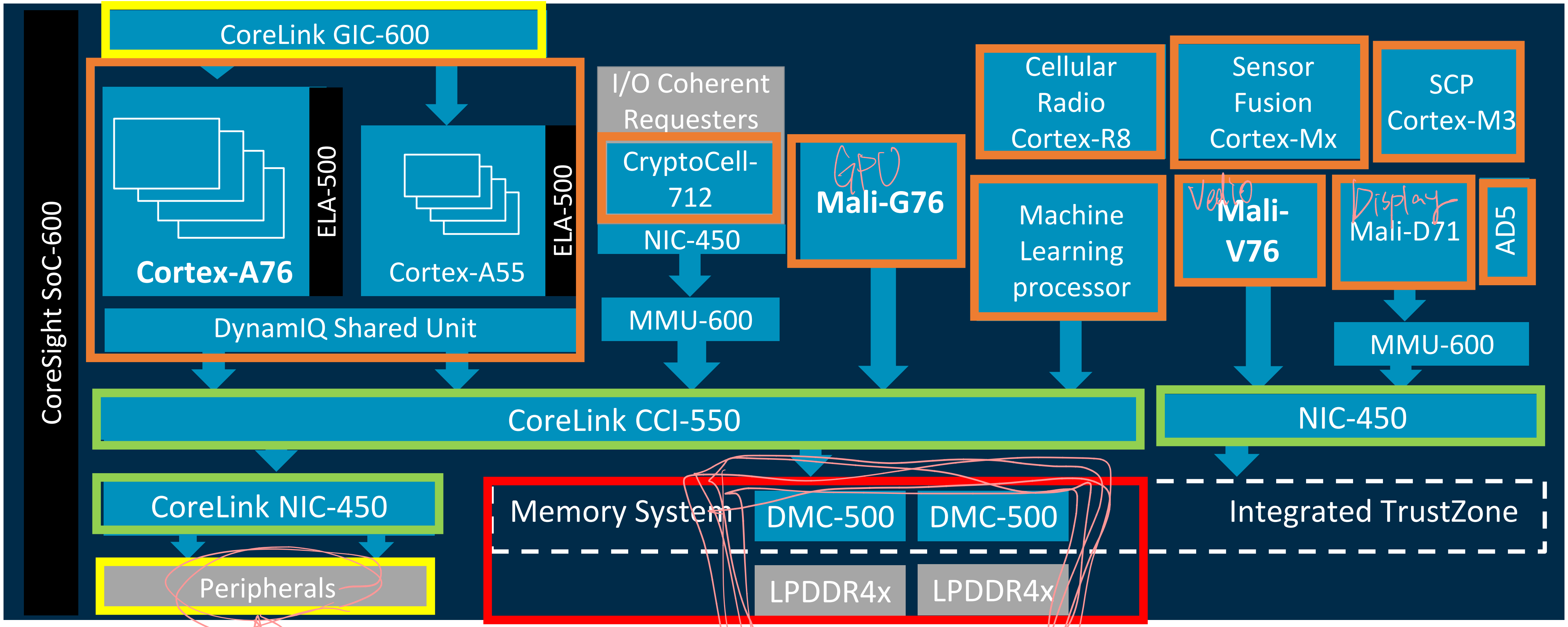
CPU

Computation

Peripheral

Interconnect

Memory



CoreSight SoC-600

CoreLink GIC-600

Cortex-A76

ELA-500

Cortex-A55

ELA-500

DynamIQ Shared Unit

I/O Coherent Requesters

CryptoCell-712

NIC-450

MMU-600

GPU

Mali-G76

Cellular Radio Cortex-R8

Sensor Fusion Cortex-Mx

SCP Cortex-M3

Machine Learning processor

Video

Mali-V76

Display

Mali-D71

AD5

MMU-600

CoreLink CCI-550

NIC-450

CoreLink NIC-450

Peripherals

Memory System

DMC-500

DMC-500

LPDDR4x

LPDDR4x

Integrated TrustZone

Heterogeneity

- Cores running the same ISA but with different microarchitectures
  - big.LITTLE (cortex-a76.CORTEX-A55) Cortex-A CPUs in one cluster with a shared coherent memory
  - Shared L3 caches and separate voltage/frequency domains within each cluster.
  - Improve performance and provide more opportunities for power saving.
- Cores extracting parallelism (data) (GPU)
  - The GPU (Mali-G76, Mali-V76) is specialized to efficiently exploit data-parallel parallelism.
  - Mali-G76 – 10 GPU providing 240 32-bit execution lanes (INT8)
  - Mali-V76 – Video processor (video encode and decode)
  - Mali-D71 – Display processor (scaling, rotation, composing layers, picture quality enhancement)
  - Assertive Display 5 (AD5) – HDR management, Ambient light adaptiity
- Cores specialized to specific tasks (Machine-Learning ML Processor)
  - Different instructions
  - Different programming model
  - Different pipeline structure
  - Different methods of communication

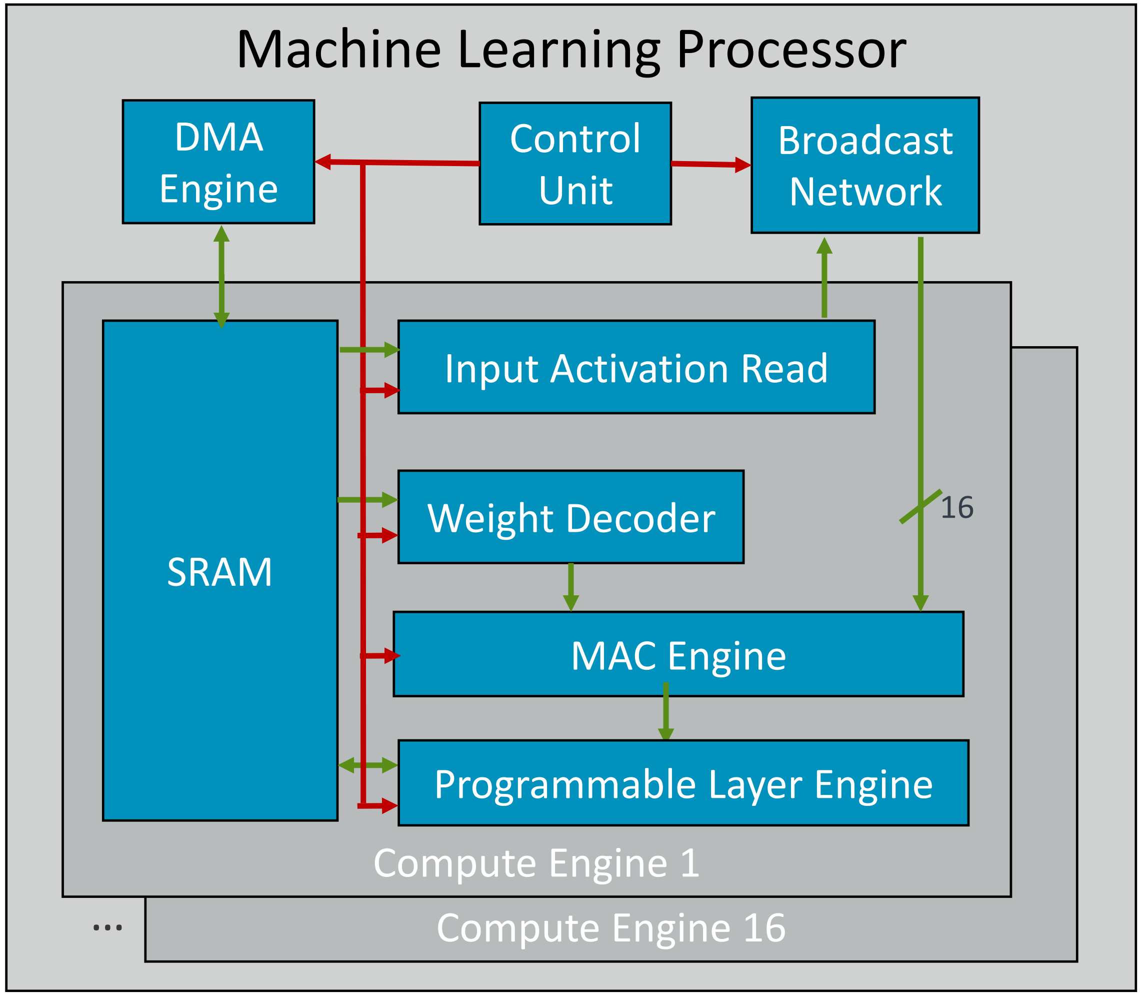
programmable

Network mapping

Key aim: Reduce power consumption but increase performance

Specialization – A Machine Learning Processor

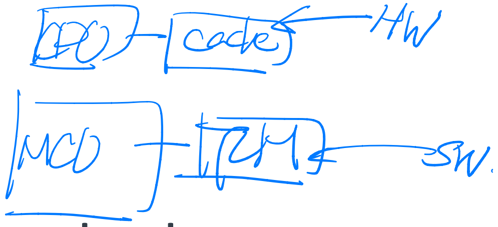
- Specialized for neural-network inference
- Example of Arm ML processor with 16 compute engines
  - 8-bit quantized integer support
  - No caches
- Convolutional neural networks statically mapped onto the compute engines
  - Output feature maps interleaved across engines
  - Weights held in that engine's SRAM
  - Input feature maps interleaved across all SRAMs



Real-time Processing

A-core archetecture 太大 reach data 不是 time

M-core



- The software requirements for the latest cellular communications standards are complex.
  - E.g., LTE Advanced Pro and 5G
- Requires real-time multi-core processor
  - In our SoC example, a quad-core coherent cluster of Cortex-R8 cores is used.
- Low-latency and hard real-time requirements
  - Large Tightly Coupled Memories (TCMs) in addition to traditional instruction/data caches
  - Simple Memory Protection Unit (MPU) rather than virtual memory to reduce memory latency
  - Extra interface ports to tightly couple the rest of the latency-sensitive modem system with cores
- Reliability
  - Improved error detection, correction, and containment schemes

Interconnect

lab4

ao report

AMBA Bus

1. APB - a lower speed peripheral bus; sort of like south bridge.
2. AHB - several versions (older north bridge).
3. AXI - a newer multi-CPU (master) high speed bus. Example NIC301.
4. ACE - an AXI extension.

Cache Coherent Interconnect (CCI-550） AXI

- Provides cache coherence between CPU clusters and the GPU, network interfaces, and the machine learning accelerator (snoop filter)
- A fully coherent GPU simplifies software development and improves performance.
  - Removing the need for software-managed cache maintenance

NIC-450

- Highly configurable topology with Network on Chip properties
- Up to 128 requesters and 64 completers in a combination of different AMBA protocols

128x64 FIFOs

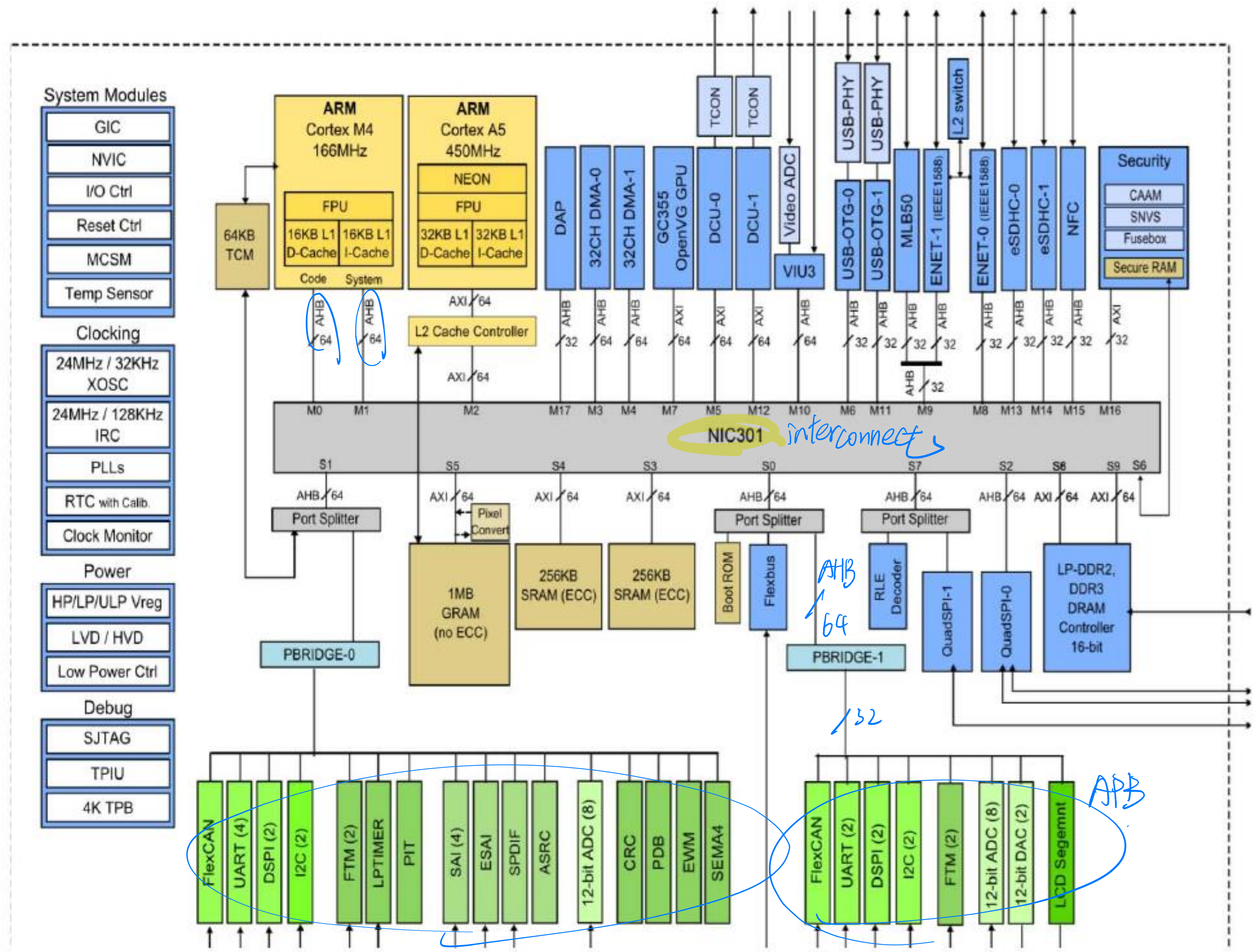


Figure 1. Vybrid internal architecture

① Freq
② data width
③ Protocol
④ FIFO
-Arbitration

Peripheral

Generic Interrupt Controller

- Performs interrupt management, prioritization, and routing
- Boosts processor efficiency and interrupt virtualization

General IO

- UART, I2C, SPI ….

timer

interrupt (resourse sharing)

Memory Controller (DMC-500)

MB/S bandwidth

Isochronous + priority
latency

- What does a memory controller have to do?
  - Convert system memory requests to the necessary series of commands to access the correct rank, bank, row and column in an external SDRAM
  - Buffer and reorder requests to optimize performance and meet QoS goals
  - Error checking and handling
  - Refresh control logic for SDRAM

utilization

DRAM controller

- What is it connected to? In our SoC example:
  - Each memory controller (Arm DMC-500) supports dual AXI4 (128-bit) system interfaces.
  - Connects to the actual DRAM PHY using a standard interface (called DFI).