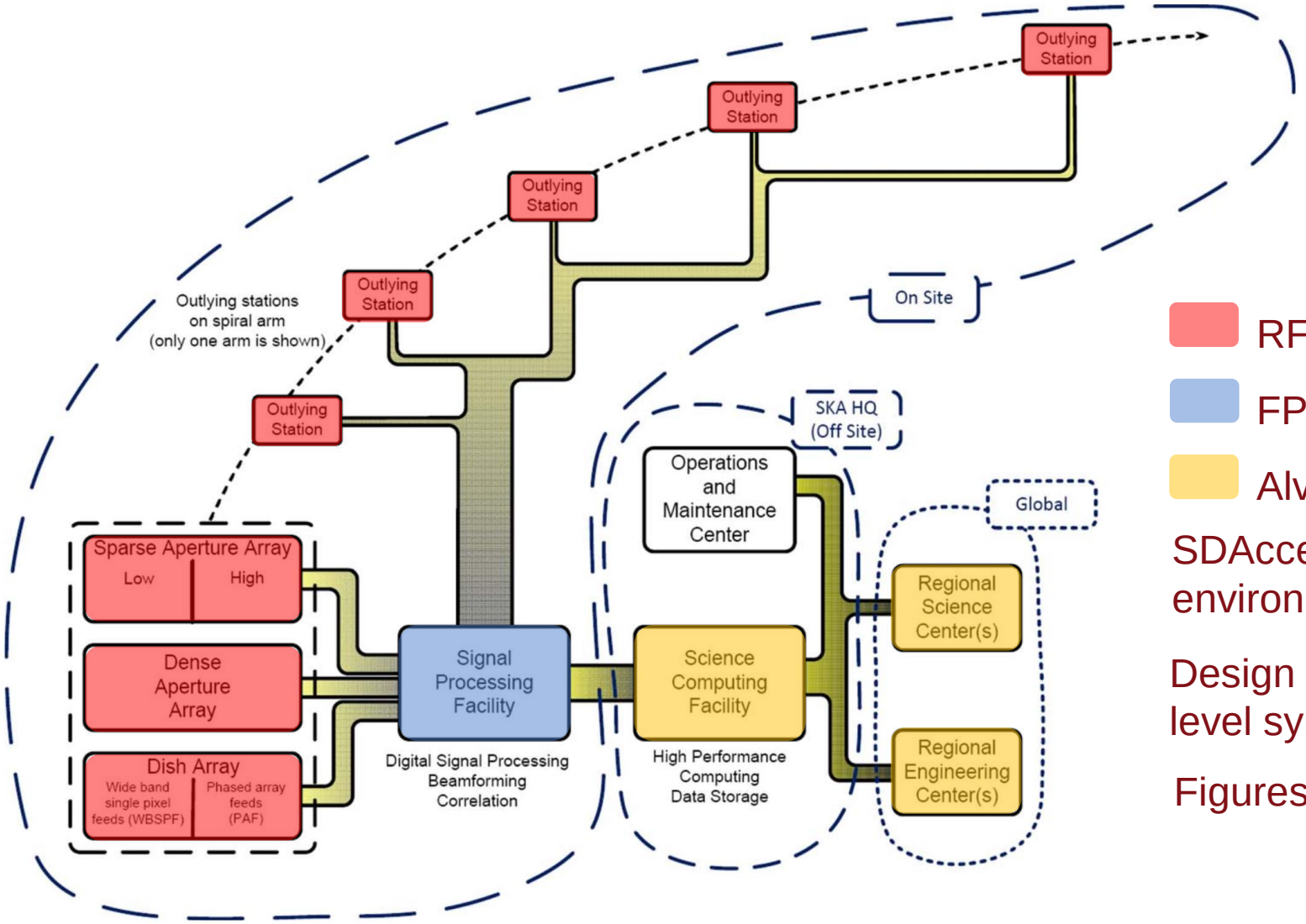
Xilinx Solutions for Radio Telescope Arrays

Name: Michael Reznik

Date: 15 February 2019

XILINX.

What does Xilinx offer?



Source: The Square Kilometer Array

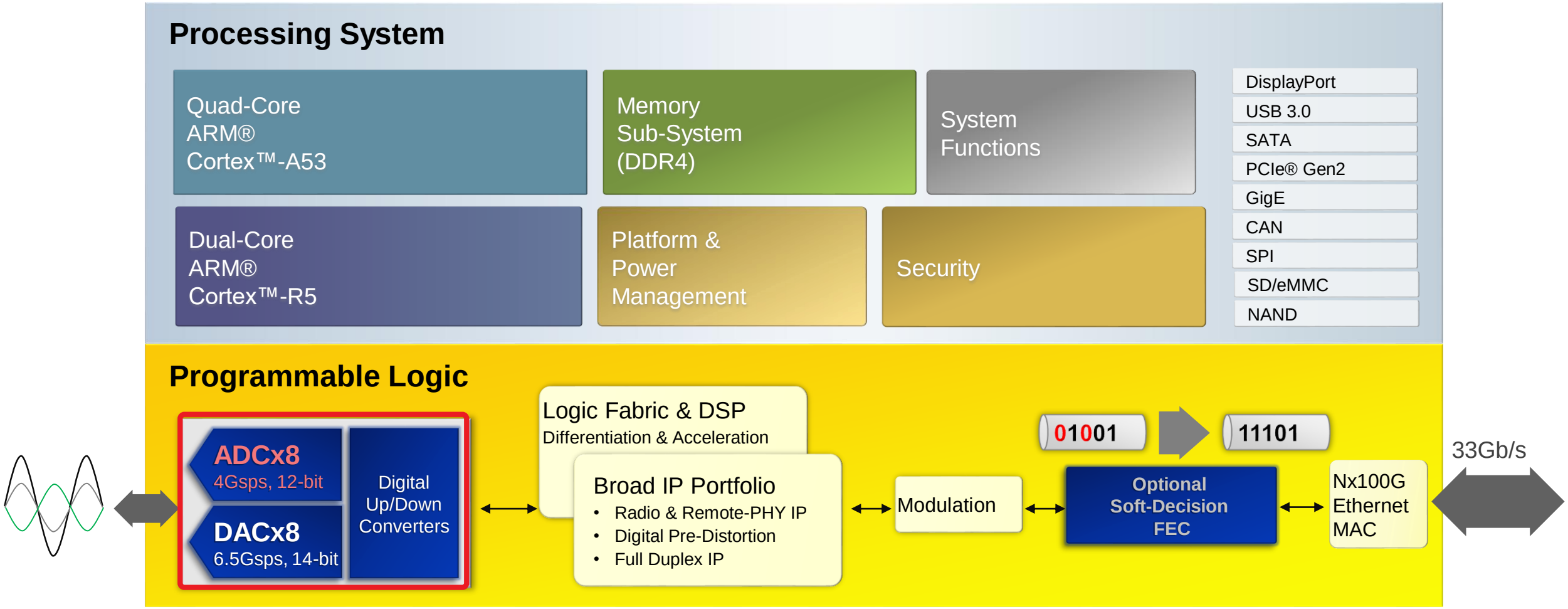
- RFSoC and super sampling blocks
- FPGAs (with on-chip HBM available)
- Alveo acceleration cards

SDAccel: Software development environment for Alveo

Design considerations for writing high-level synthesis code for an FX correlator

Figures of merit for next-gen ACAP

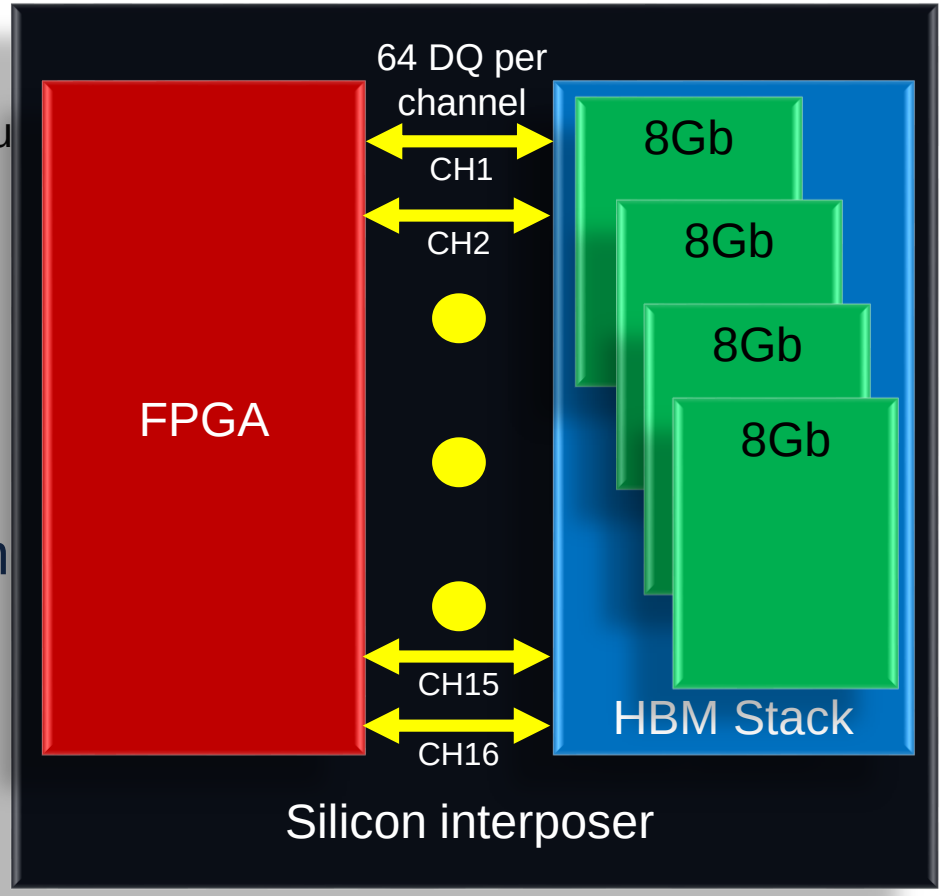
RFSoC Block Diagram



XILINX.

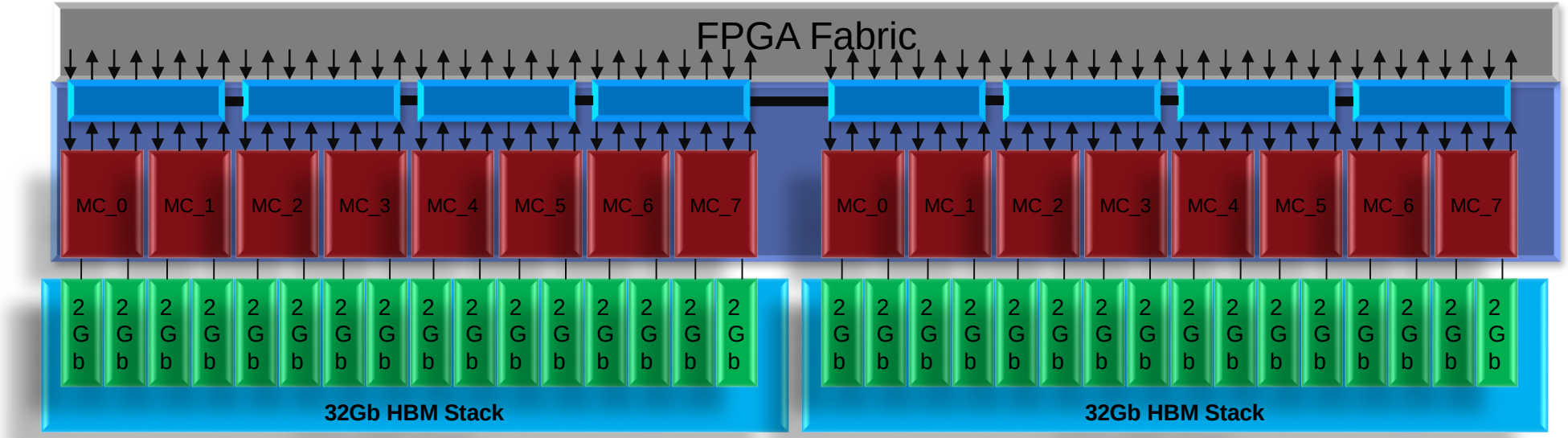
HBM: Terabit/s memory bandwidth by the numbers

- HBM memory organization
  - 64 DQ (Bidirectional Data) signals per channel each ru
  - 16 Channels per HBM stack
  - Up to 2 HBM stacks per FPGA
  - Up to 3.68Tbps bandwidth HBM
    - 64*1800*16*2 = 3.686Tb/s
    - 64*1800*16*2/8 = 460GB/s
- Xilinx used 4 high HBM 3D stacked mem
  - Up to 64Gb of memory per FPGA
    - 4H*8Gb*2HBM stacks=64Gb
    - 4H*8Gb*2HBM stacks/8=8GB



HBM Architecture + Xilinx innovation

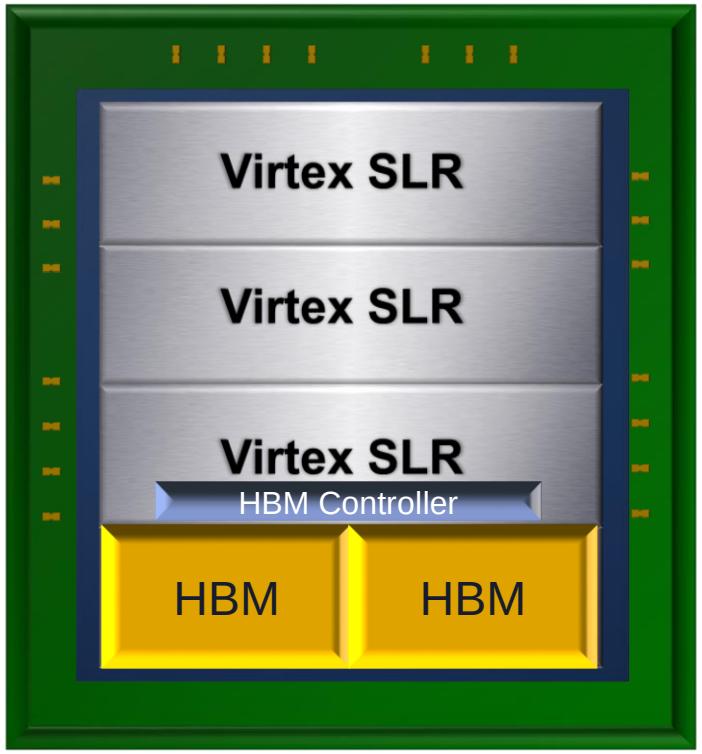
- Standard HBM architecture
  - 16 Pseudo channels per HBM Stack, accessing a discrete 2Gb memory
  - 8 Memory controller per HBM stack
  - 16 512 bit AXI RX/TX ports per HBM stack
  - Each AXI port can address a corresponding 2Gb section of memory
- Xilinx innovations
  - Added flexible addressing that creates a unified memory map
    any port can access any memory address
  - Extend AXI ports into fabric to ease timing

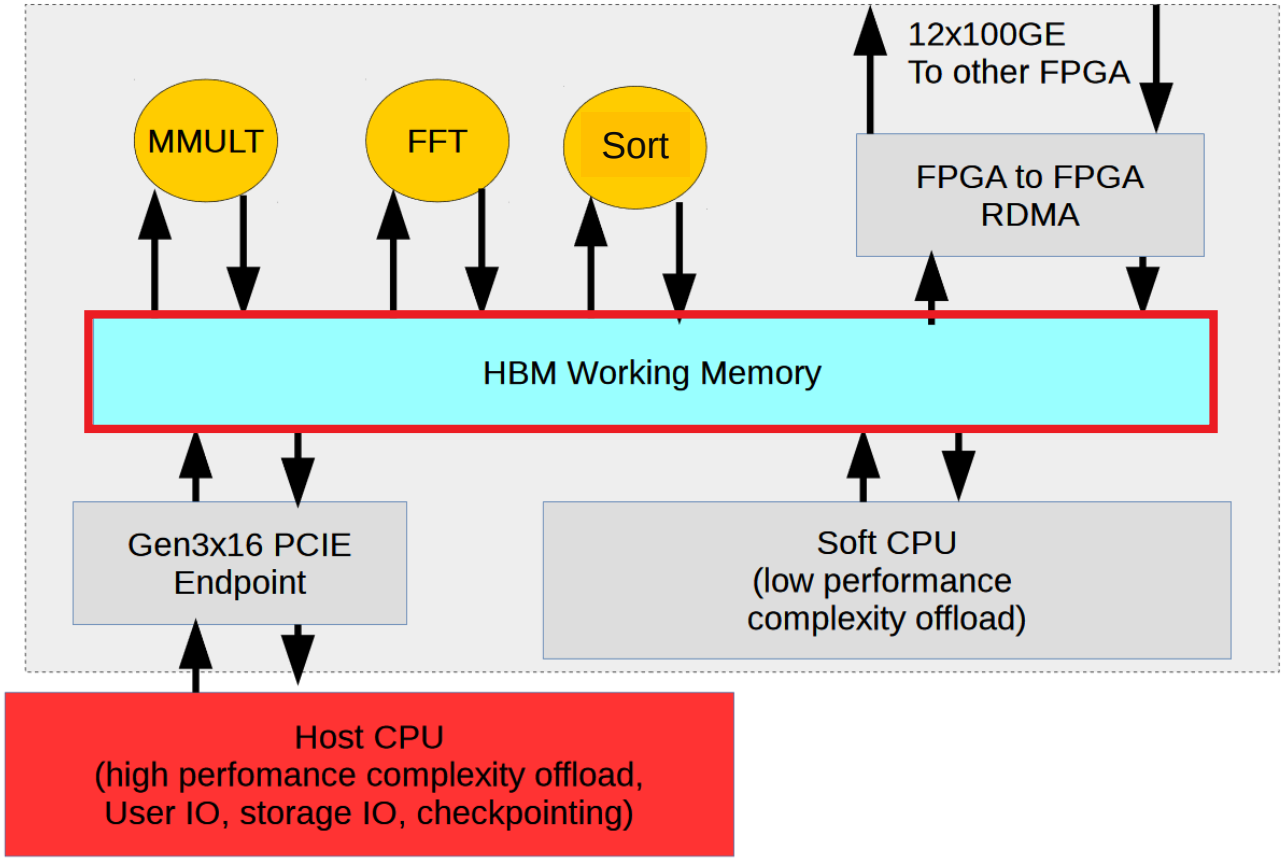


XILINX

FPGA with On-chip High-Bandwidth Memory (HBM)

- 8 GB of HBM
- Up to 460GB/s of memory bandwidth between HBM and programmable logic fabric





Source: [Breaking Memory Bandwidth Barriers Using High Bandwidth Memory FPGA]

Alveo Accelerator Cards

Alveo U200

- **18.6** Peak INT8 TOPs
- **77GB/s** DDR Memory Bandwidth
- **31TB/s** Internal SRAM Bandwidth
- **892,000** LUTs

Alveo U250

- **33.3** Peak INT8 TOPs
- **77GB/s** DDR Memory Bandwidth
- **38TB/s** Internal SRAM Bandwidth
- **1,341,000** LUTs

Alveo U280

- **24.5** Peak INT8 TOPs
- **460GB/s** HBM2 Memory Bandwidth
- **30TB/s** Internal SRAM Bandwidth
- **1,079,000** LUTs

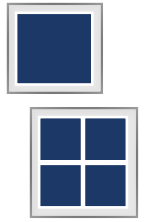
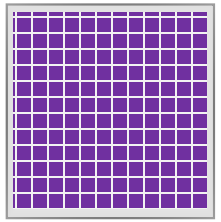
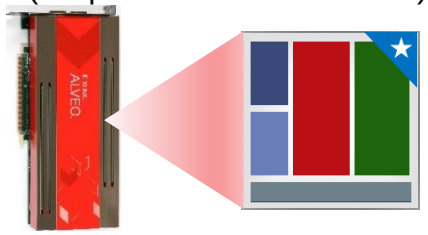
> **PCIe interface: Gen3x16 (U200 & U250), Gen4x8 w/ CCIX (U280)**

> **Network connectivity: 2x QSFP28**

> **Power: 100W (typ)**

XILINX.

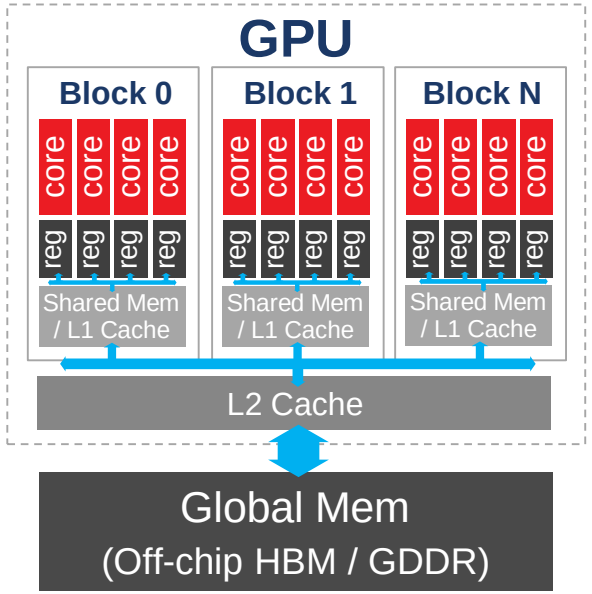
GPU vs. Alveo Competitive Overview

	CPU (Sequential)	GPU (Parallel)	Alveo (Sequential + Parallel + Mix)
			
SW Programmable	✓	✓	✓
HW Adaptable	–	–	✓
Workload Flexibility	✓	–	✓
App / Workload Perf	–	✓	✓
Device / Power Efficiency	–	✓	✓
Network Access	✓	–	✓

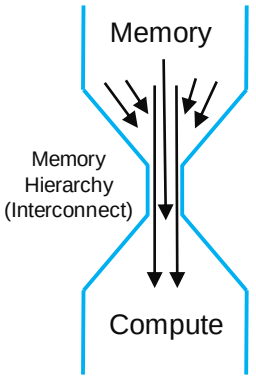
Leverage Alveo's HW adaptability to deliver highest application performance & efficiency

Key Adaptable Advantage vs. GPGPU - Memory Hierarchy

Latency	Power
1X	1X
2X	10X
80X	100X

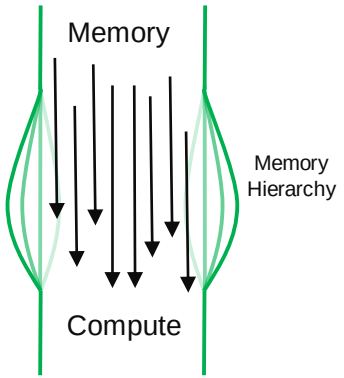


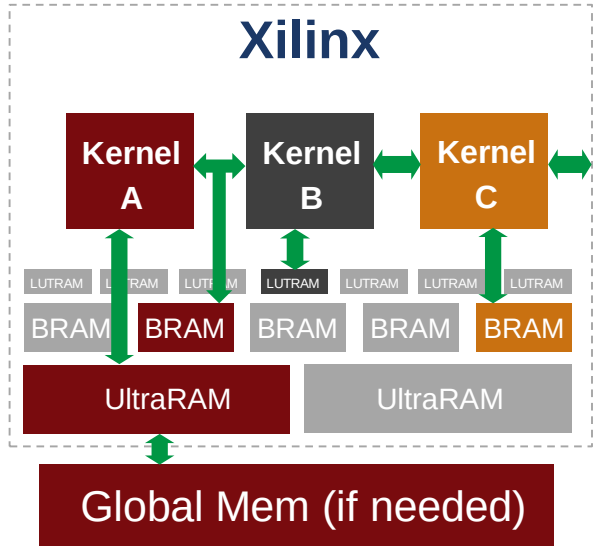
Fixed



- Rigid memory hierarchy & SW defined dataflow
- High “data locality” required for workload efficiency
- “Batching” improves efficiency at expense of latency

Adaptable





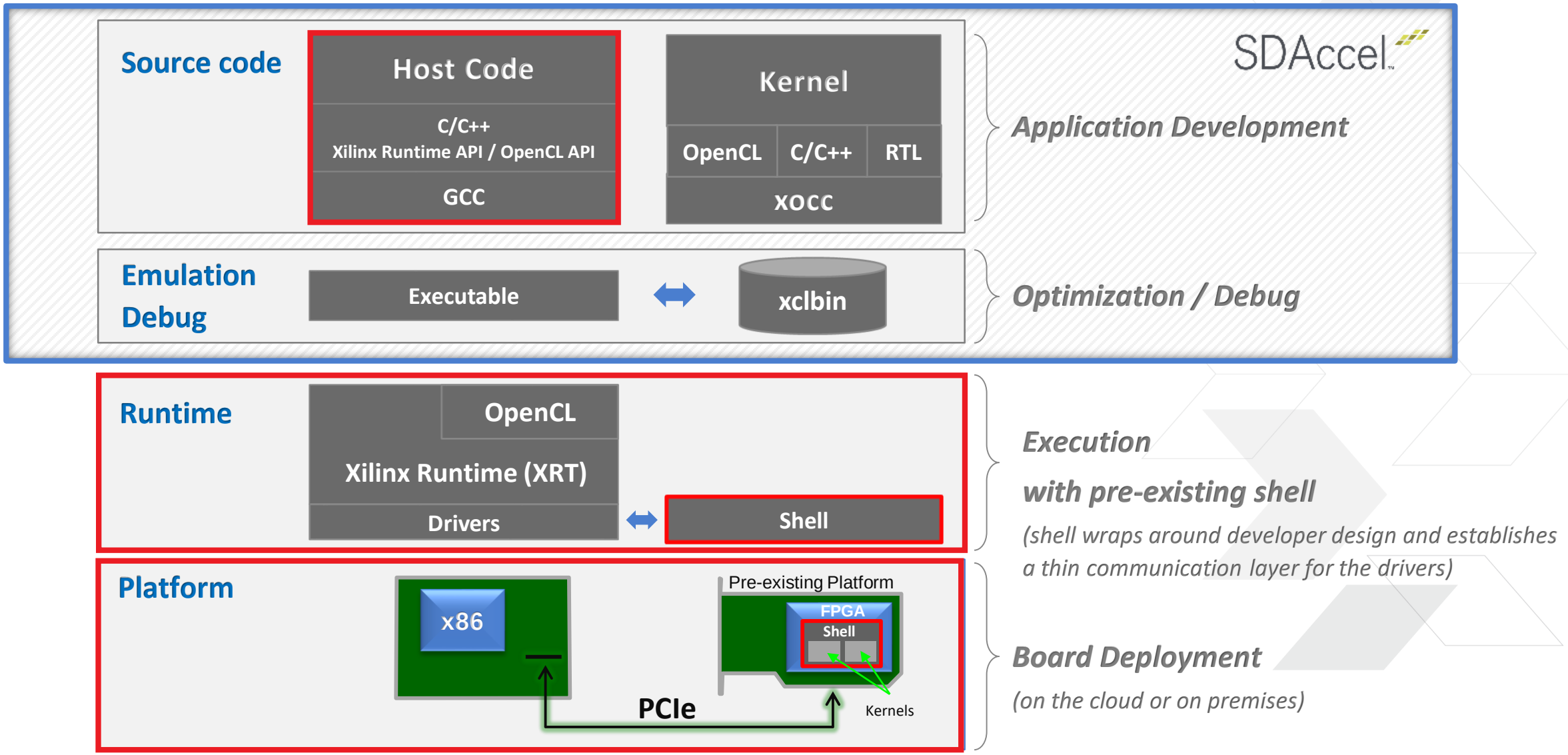
- Adaptable memory hierarchy & datapath
- ~5X more on-chip memory
- Max throughput, min latency – no batching required

Resultant Workload Speed-Up vs. GPU

DB - SQL	DB - RegEx	ML Infer	Genomics	Video
4X	3X	3X	6X	6X

XILINX

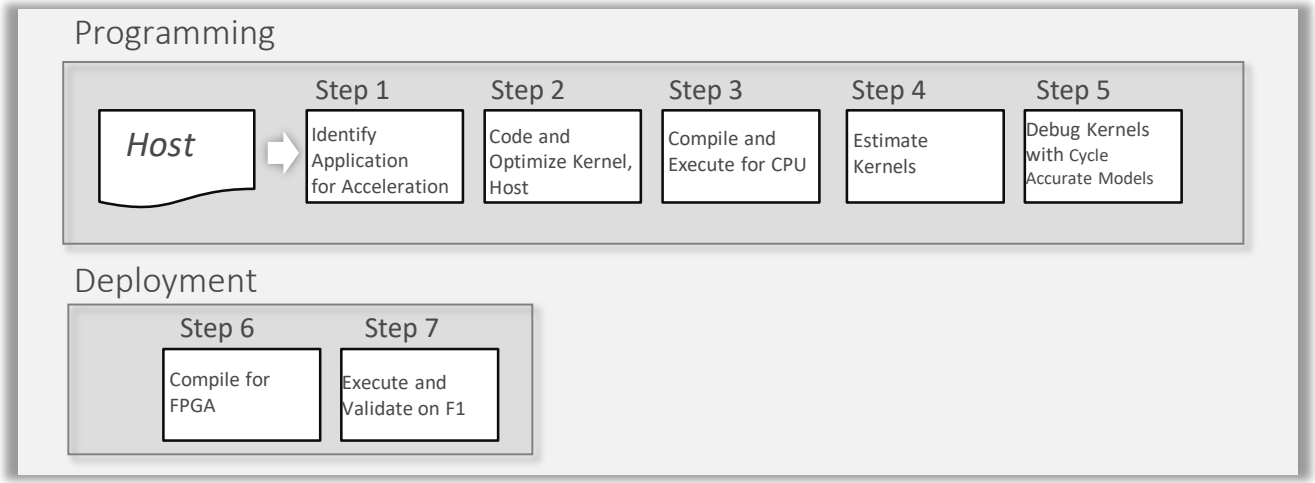
SDAccel, Runtime and Platform



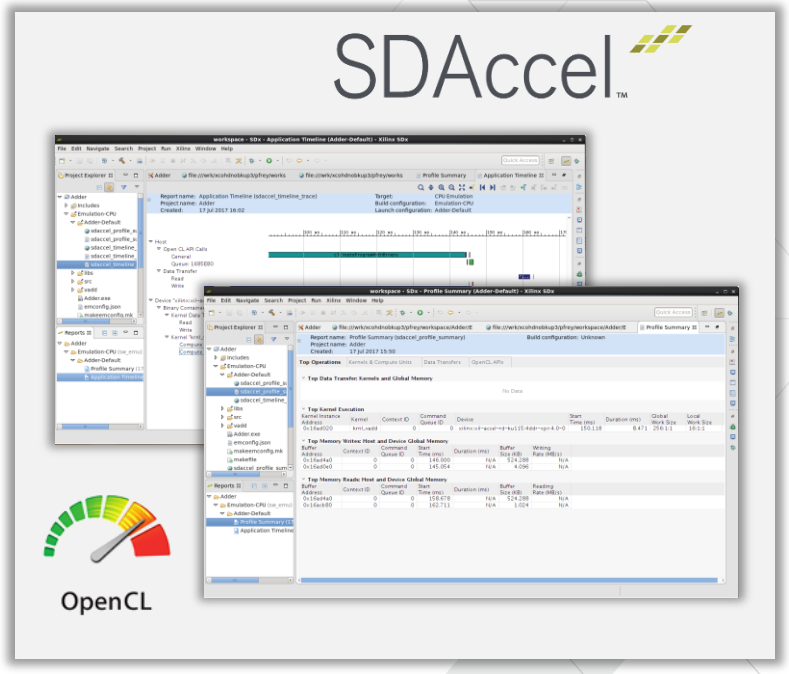
XILINX

Xilinx SDAccel

- **Develop, profile and deploy OpenCL applications**
  - OpenCL uses standard APIs (code is portable)
- **F1 platform aware**
- **Flexible kernels development**
  - C / C++ / OpenCL / RTL



Programming Steps

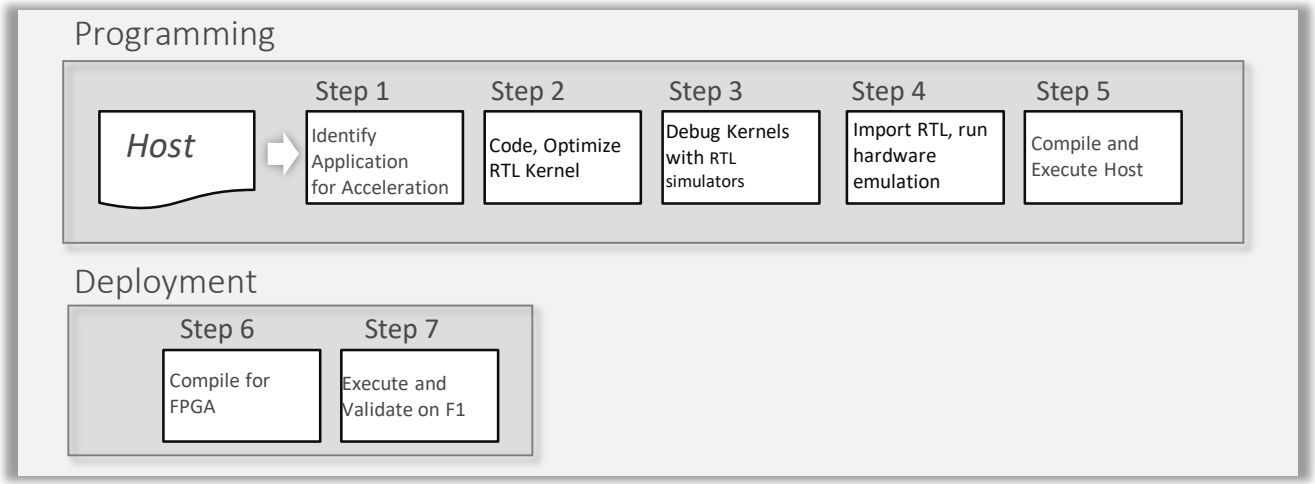


Comprehensive debug and profiling environment

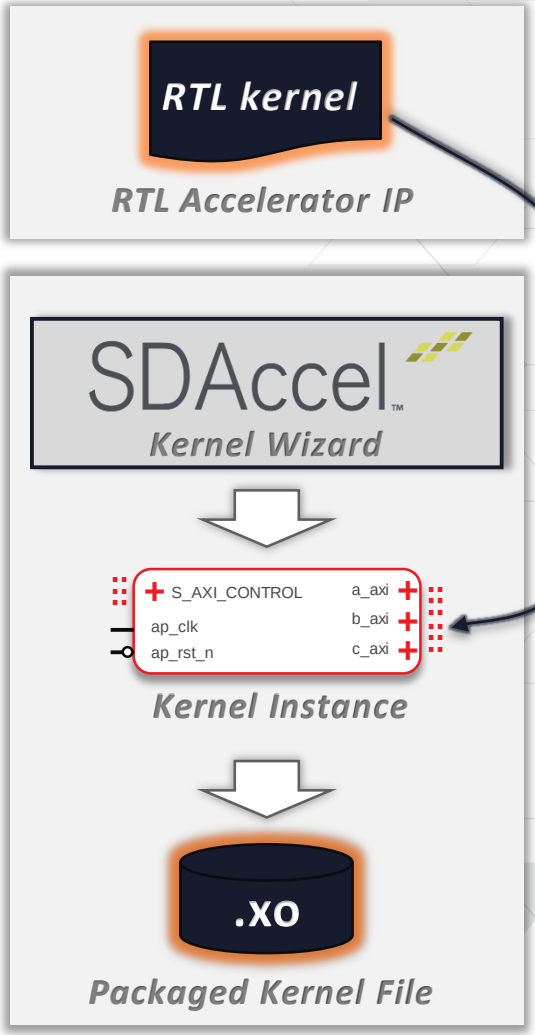
XILINX

Xilinx SDAccel with RTL Kernels

- **RTL import through kernel wizard in SDAccel**
  - Top level needs to match interface requirements
- **Leveraging existing RTL IP**
  - RTL is resource efficient and high performance
- **Hardware emulation mode for simulation**



Programming Steps

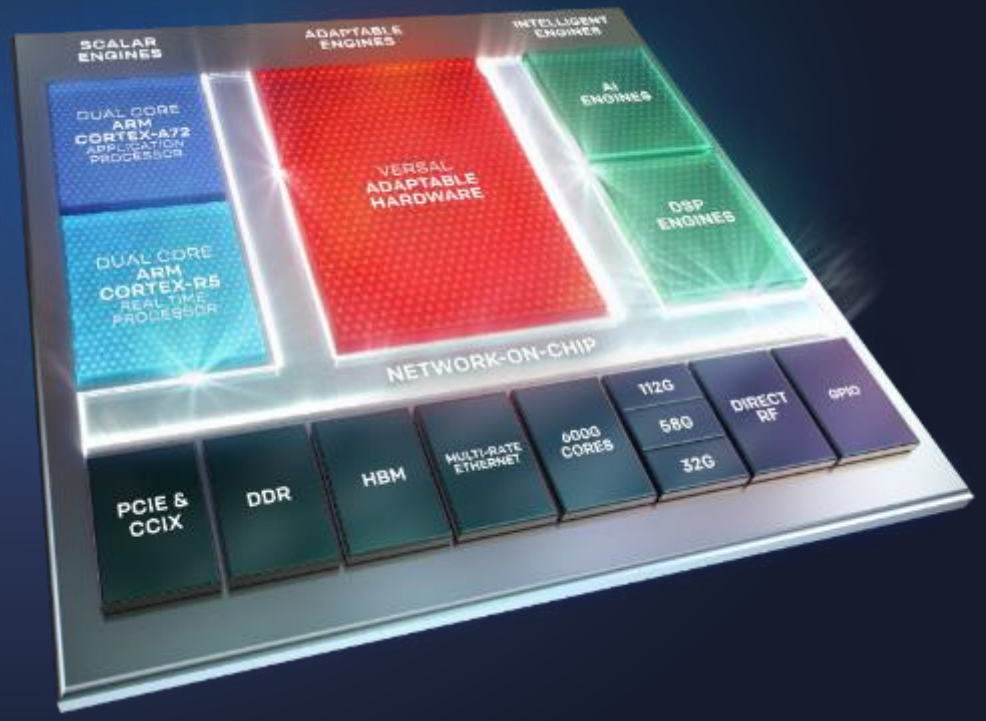


RTL Kernel Import from SDAccel

XILINX

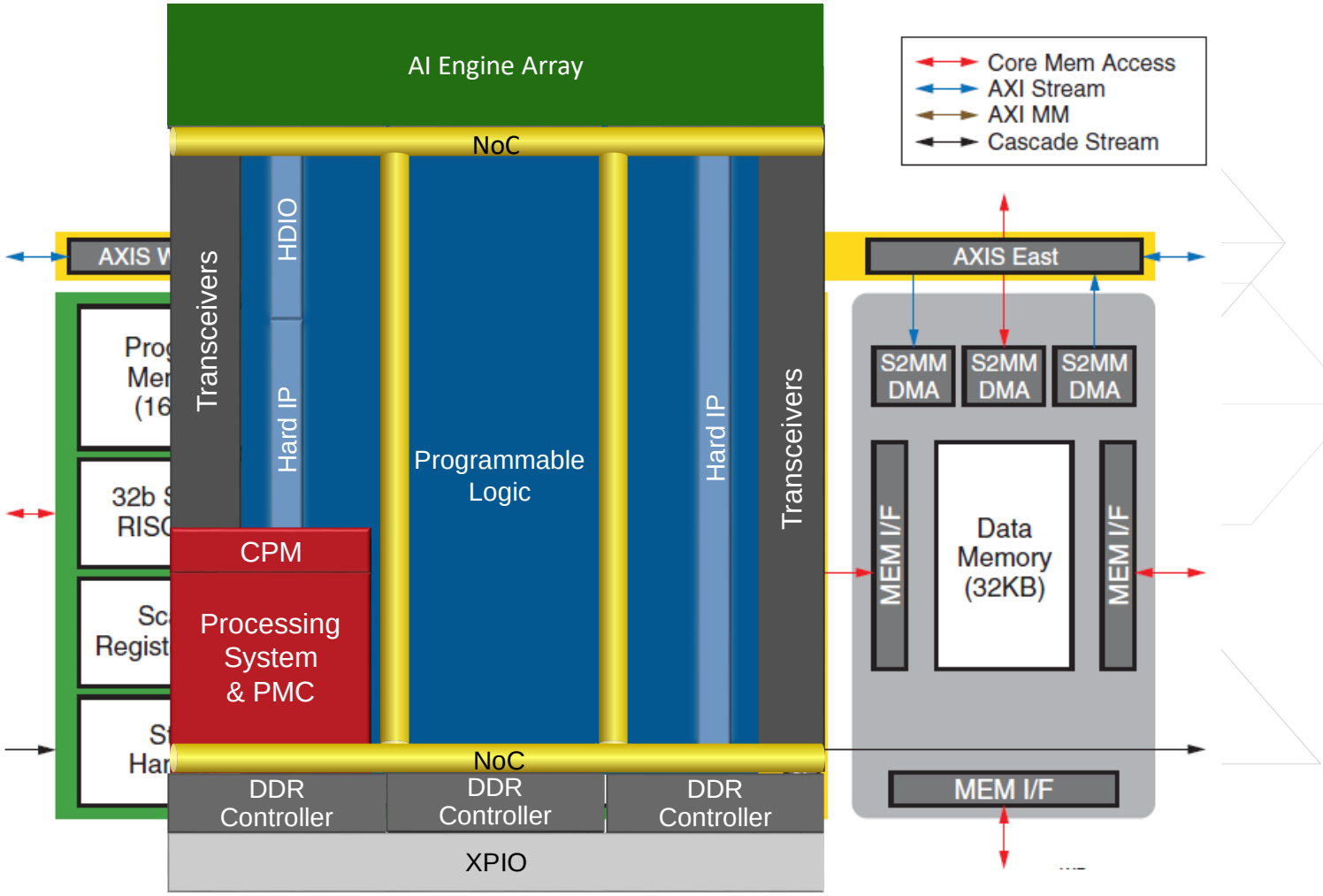
Versal ACAP Technology Tour

- Scalar Processing Engines
- Adaptable Hardware Engines
- Intelligent Engines
  SW Programmable, HW Adaptable
- Breakout Integration of Advanced Protocol Engines

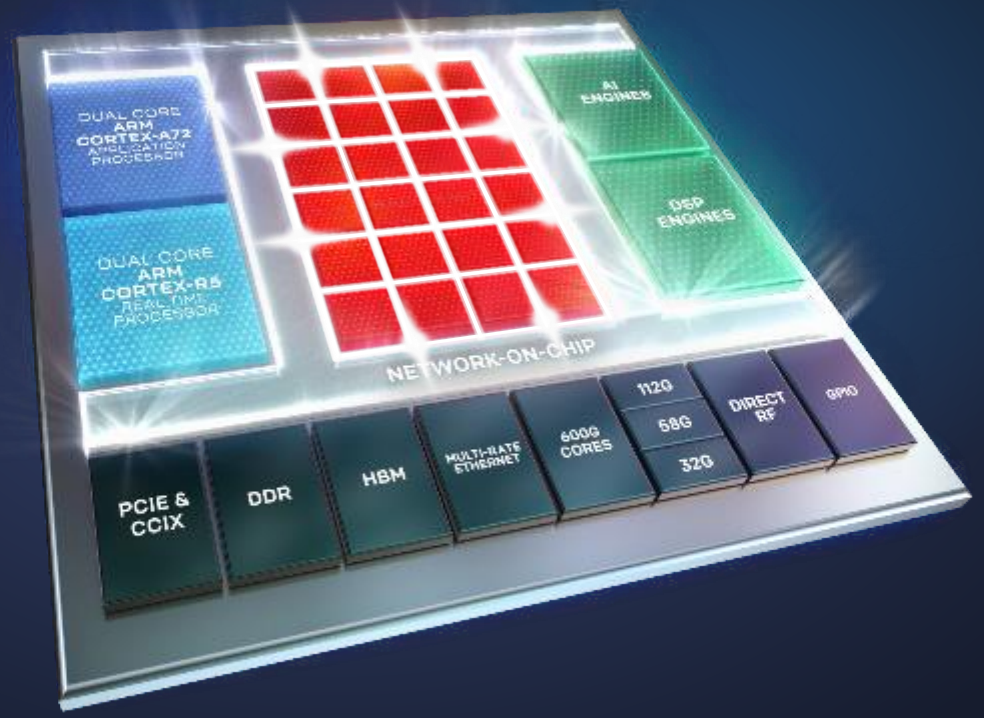


7nm Adaptive Compute Acceleration Platform (ACAP)

- **AI engine peak performance**
  - int8: 133 TOPS
  - int16: 33 TOPS
  - fp32: 8 TFLOPS
- **DSP engine peak performance**
  - int8: 13.6 TOPS
  - int24: 4.5 TOPS
  - fp32: 3.2 TFLOPS
- **Memory bandwidth**
  - Block RAM: 118 Tb/s
  - Ultra RAM: 49 Tb/s
  - DDR4: 816 Gb/s
  - LPDDR4: 1.096 Tb/s
  - Network-on-Chip: 2.5 Tb/s



XILINX.



Network-on-Chip (NoC)

Ease of Use

Inherently software programmable
Available at boot, no place-and-route required

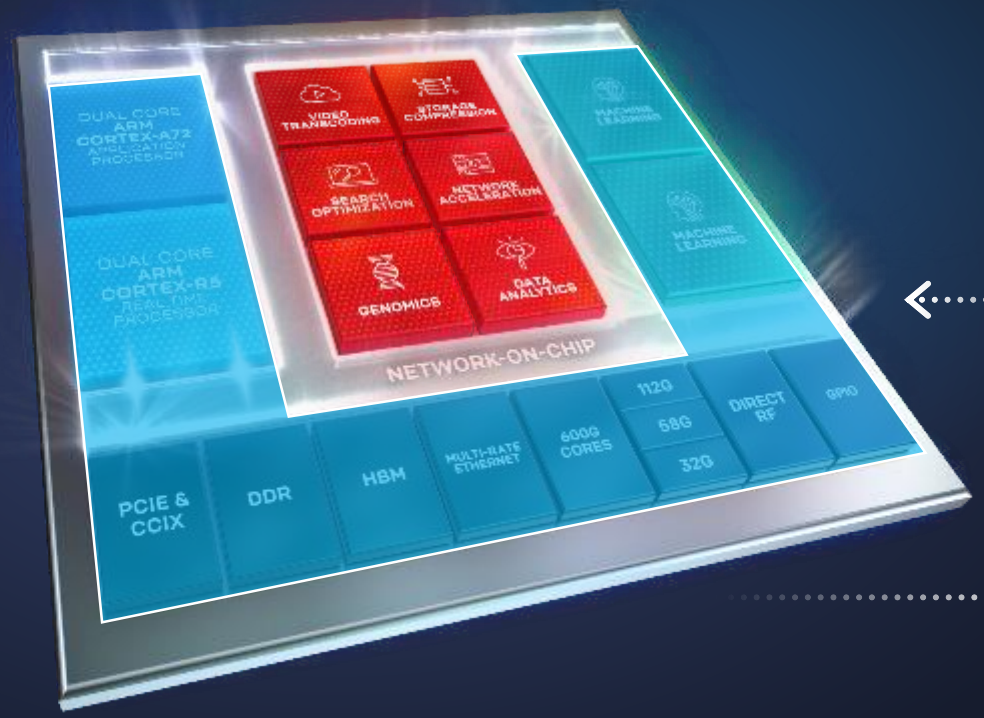
High Bandwidth and Low Latency

Multi-terabit/sec throughput
Guaranteed QoS

Power Efficiency

8X power efficiency vs. soft implementations
Arbitration across heterogeneous engines

NoC Enables Software Programmability



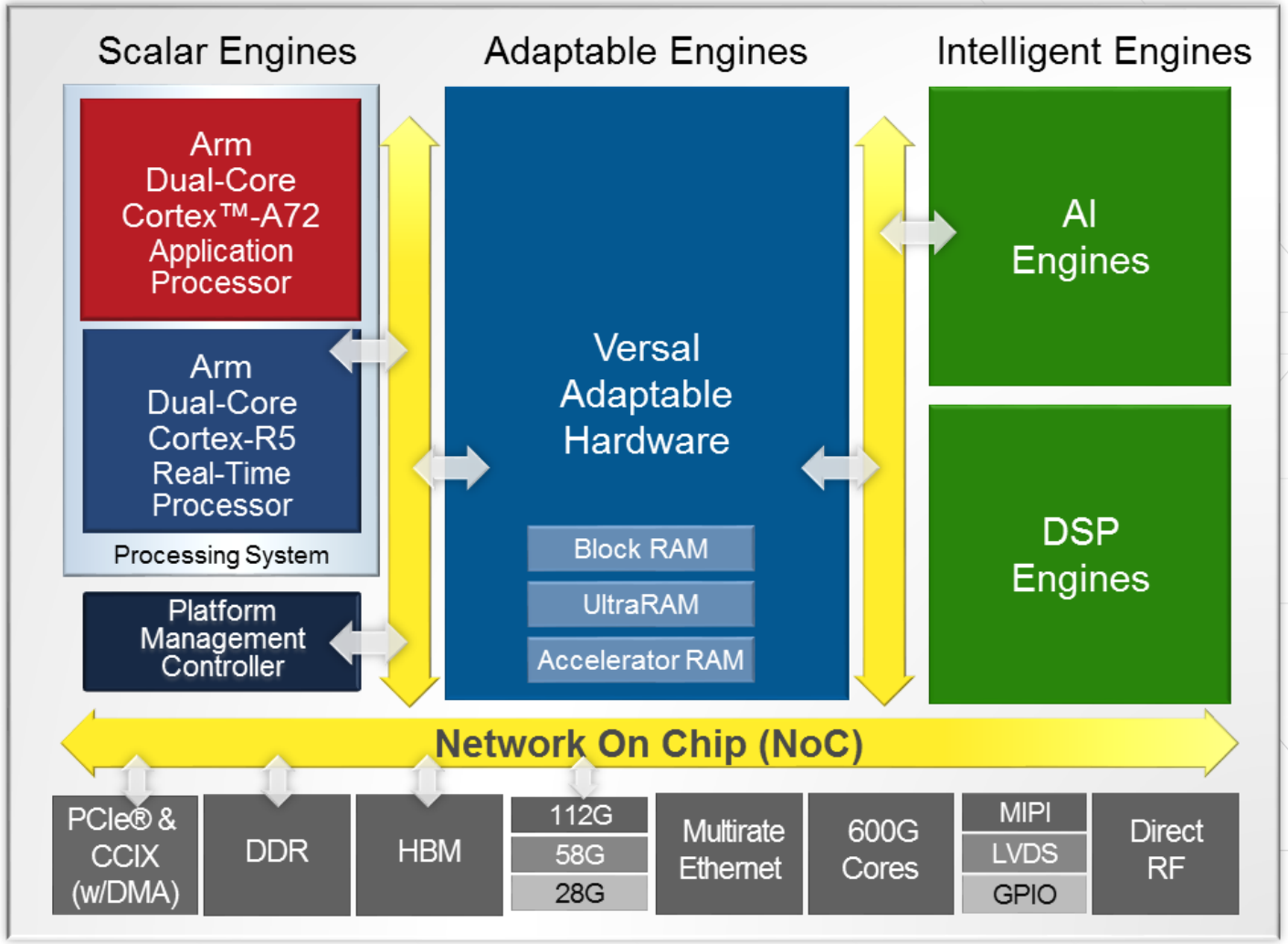
Data Transfer between Engines and Memory

Seamless Integration

XILINX

Adaptable Architecture Connected Via NoC

- **Scalar Engines**
  - Arm® Cortex™-A72 APU
  - Arm Cortex-R5 RPU
- **Adaptable Engines**
  - CLBs
  - Internal Memory
- **Intelligent Engines**
  - AI Engine
  - DSP Engine
- **Connectivity**
  - PCIe w/CCIX
  - Ethernet
  - DDR Memory Controllers
  - Transceivers
  - I/O
- **Platform Resources**
  - Network-On-Chip
  - Platform Management Controller



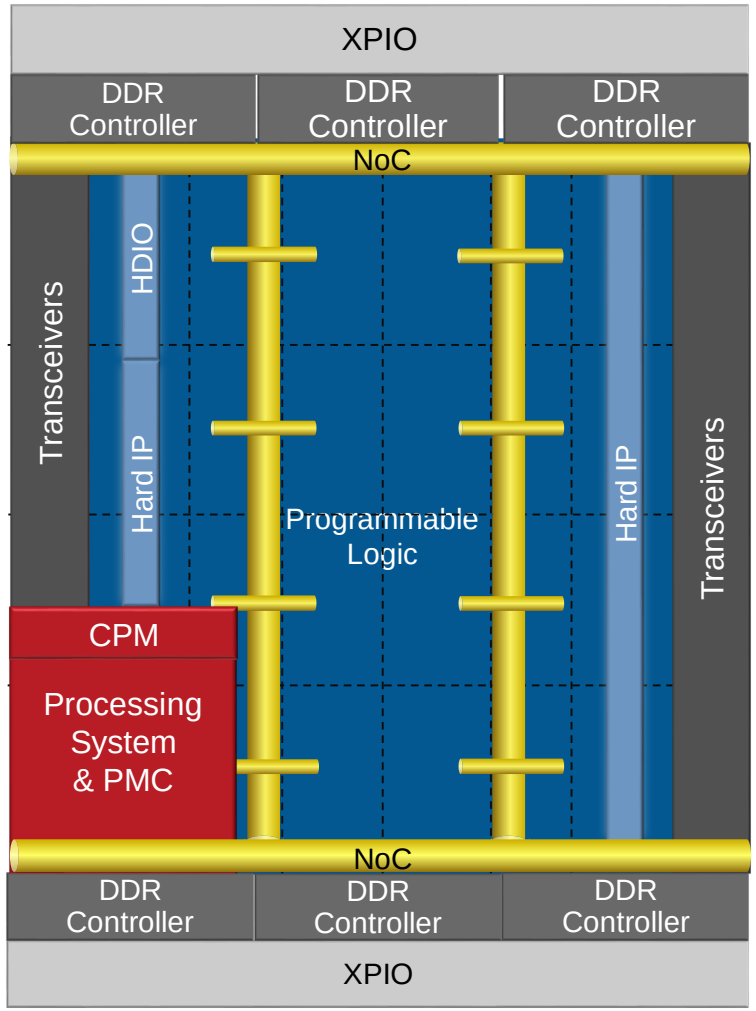
XILINX.

Versal™ Network-on-Chip

- **High bandwidth terabit network-on-chip**
  - Memory mapped access to all resources
  - Built-in arbitration between engines and memory
  - AXI4 based structure spanning full device (height and width)

- **High bandwidth, low latency, low power**
  - Guaranteed QoS
  - 8X power efficiency vs. FPGA implementations
  - Support AXI4 MM and AXI4 Stream

- **Adaptable kernel placement**
  - Every PL region has master and slave interface
  - Easily swap kernels at NoC port boundaries
  - Simplifies connectivity between kernels



XILINX

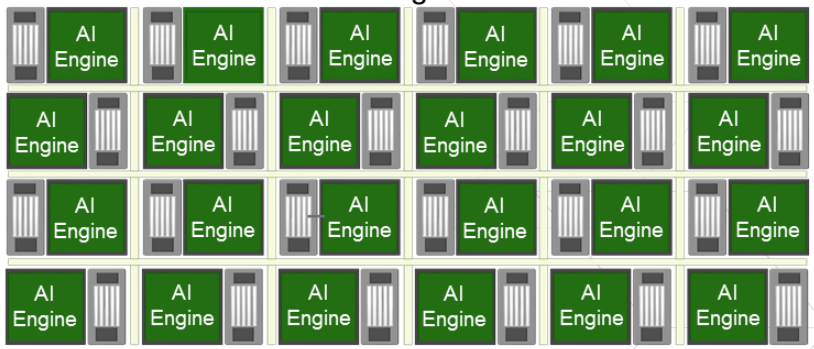
Digital Signal Processing Capability

Function	DSP48E2	DSP58
DSP Tile/Slice Type	DSP48E2	DSP58
Multiplier and MACC	27x18	27x24
32b/16b Single Precision Floating Point Multiply-Add	Soft	√
Complex 18b x Complex 18b	N/A	2 x DSP58
3 x Int8 Dot Product	N/A	√

AI Engine 2D Array

VLIW and SIMD Architecture

C/C++ Programmable

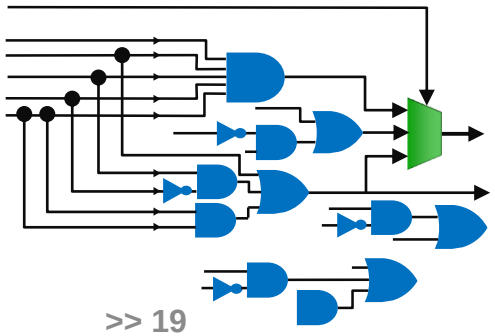


Capability

FPGA Fabric DSP

LUT and Memory

RTL Entry



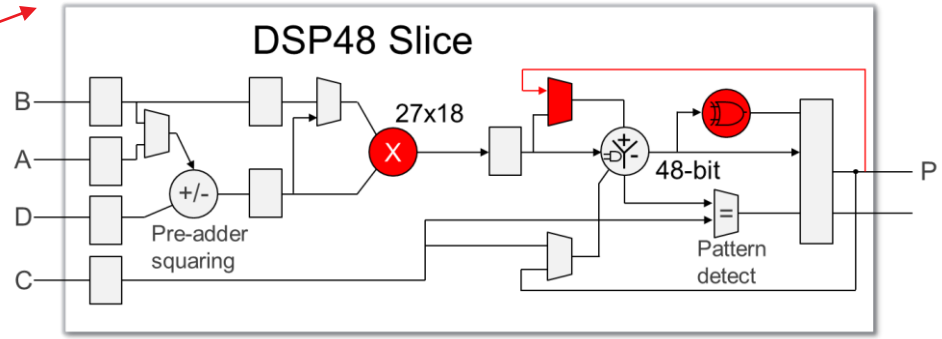
Capability

DSP48E2 Slice

Hardened MULT & ADDERS

ACC = ACC + (A × B)

RTL Entry

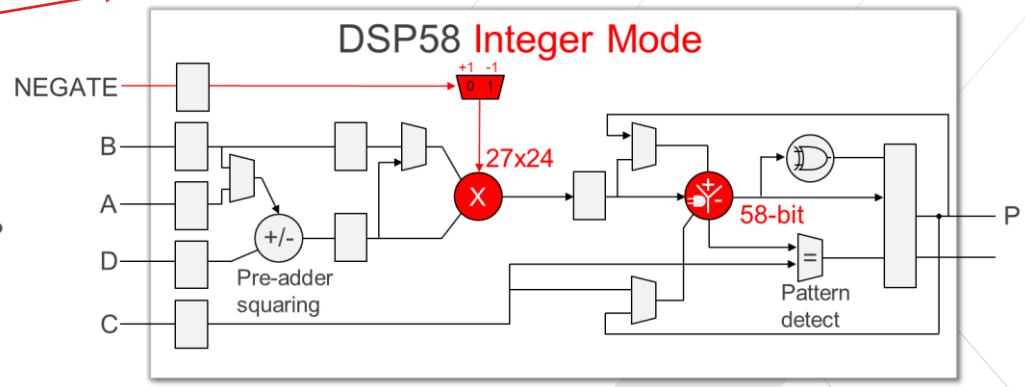


Capability

DSP58

Additional features

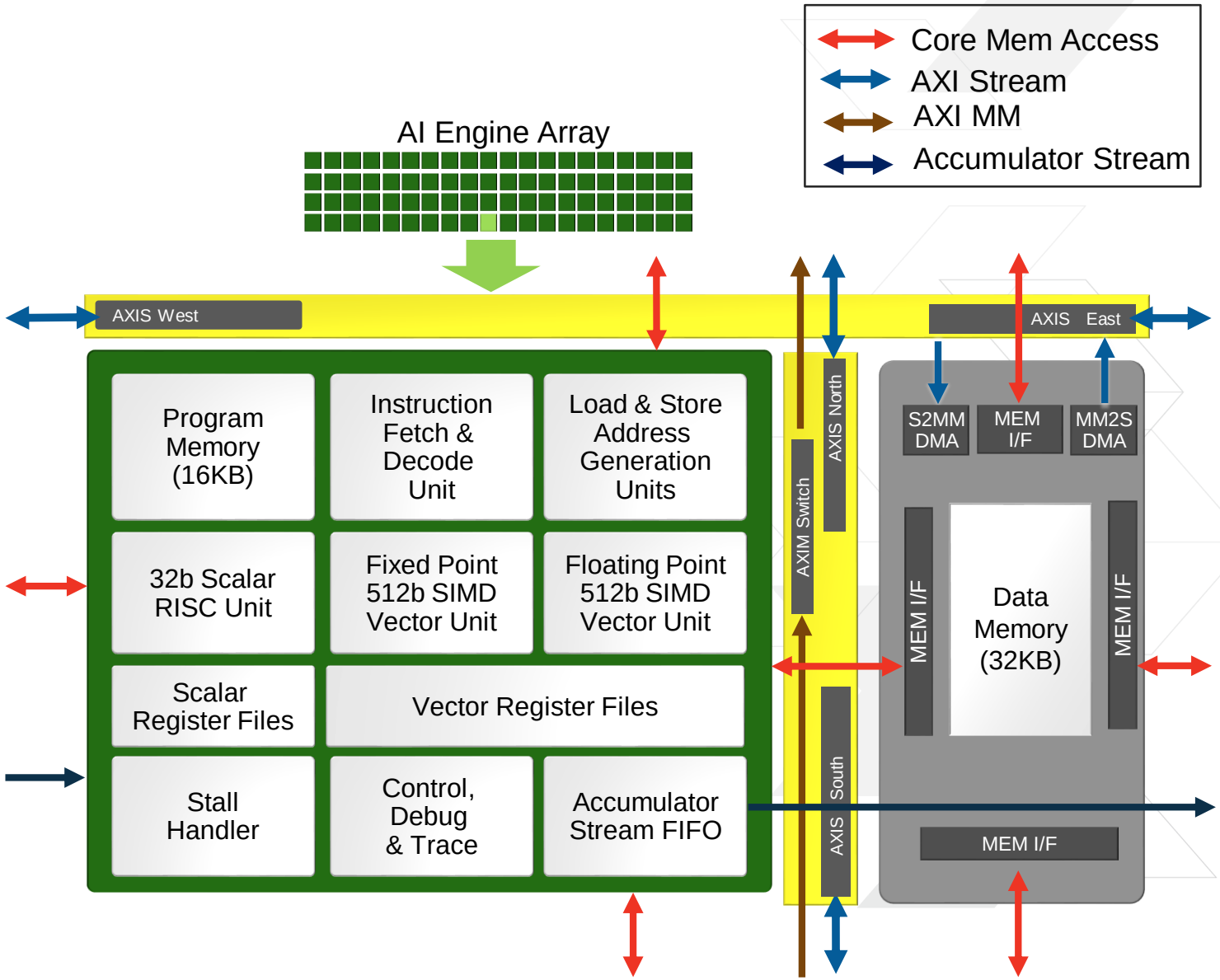
RTL Entry



XILINX.

AI Engine Architecture

- **AI Engine tile**
  - AI Engine, data memory, and interconnect
- **1+ GHz VLIW/SIMD AI Engine**
  - 32-bit Scalar RISC processor with fixed and floating point vector units
- **Each AI Engine can access 4 Memory Modules (N,E,S,W) as one contiguous memory**
- **AXI-MM Switch for configuration, control and debugging functionality**
- **AXI-Stream crossbar switch for routing N/E/S/W streams**



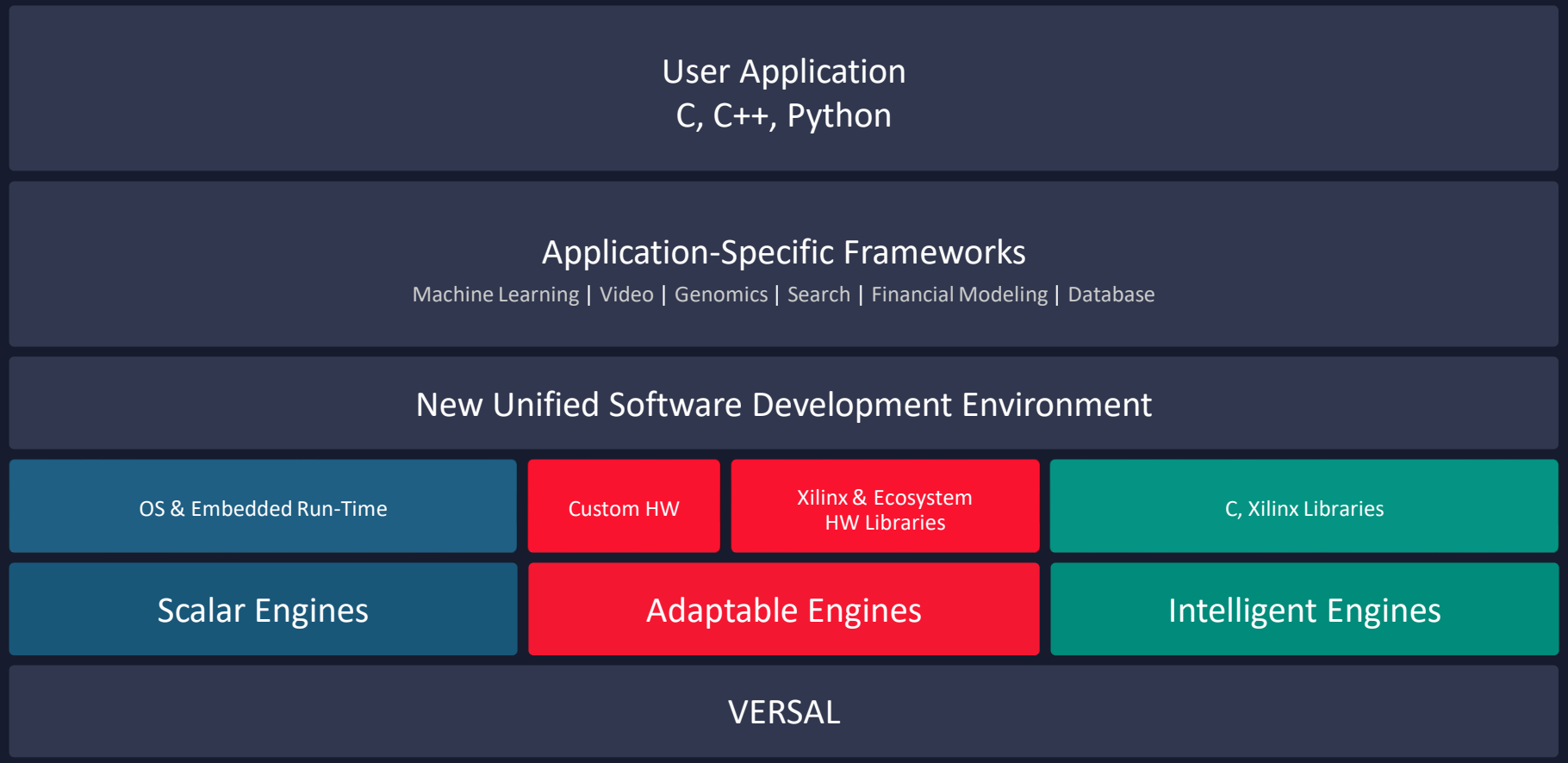
XILINX.

Versal Development Experience

Adaptable For Any Application

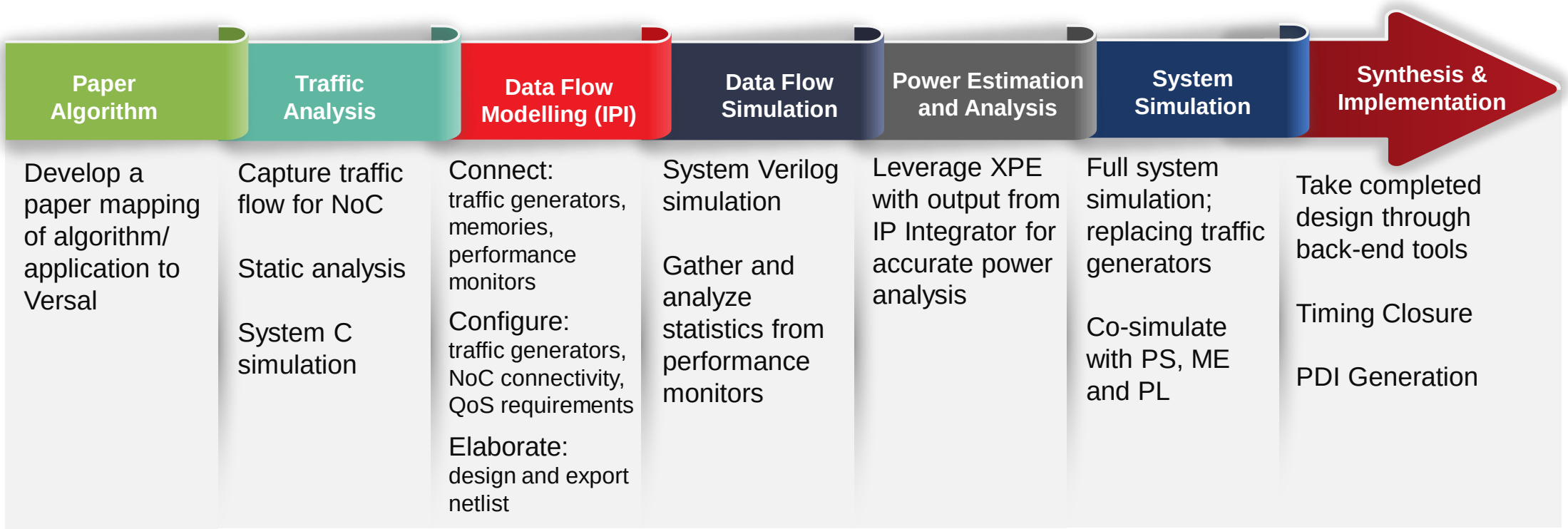
Software Programmable

Heterogeneous Platform



XILINX

System Design Methodology



Leverage these steps

XILINX.

Unified Tool Chain for Device Programming