HW/SW Programmable Engine:

Domain Specific Architecture for Project Everest

Juanjo Noguera, Goran Bilski, Jan Langer, Baris Ozgul, Tim Tuan, David Clarke, Peter McColgan, Sneha Date, Zachary Dickman, Pedro Duarte, Stephan Munz, Jose Marques, Sridhar Subramanian, Saurabh Mathur, Malav P. Shah, Chris Dick, Paul Newson, Kaushik Barman, Ramon Uribe, Helen Tarn, Engin Tunali, Richard Walke, Vinod Kathail, Shail Aditya Gupta, Akella Sastry, Mukund Sivaraman, Rishi Surendran, Abraham Lee, Chia-Jui Hsu, Kumud Bhandari, Jack Lo, Kristof Denolf, Phil James-Roxby, Samuel Bayliss, Kees Vissers, Ralph Wittig, Gaurav Singh

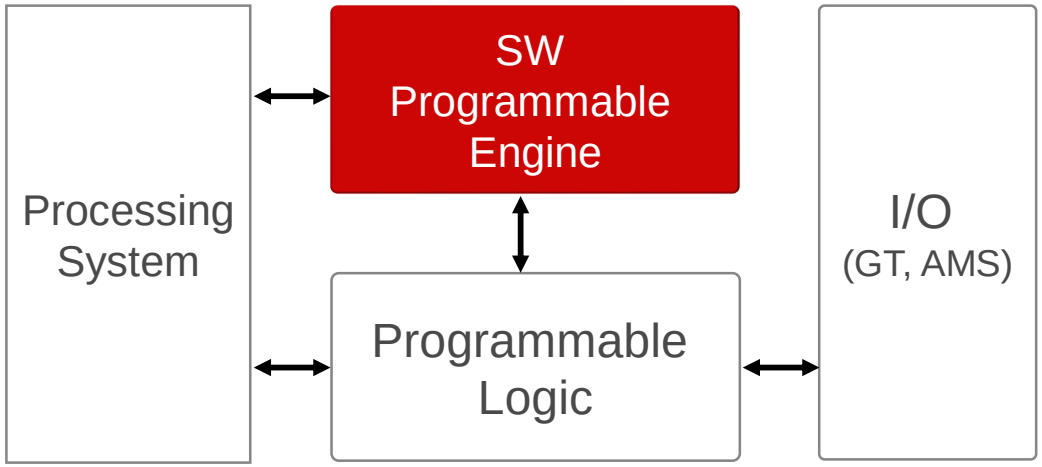
21st August 2018

XILINX.

© Copyright 2018 Xilinx

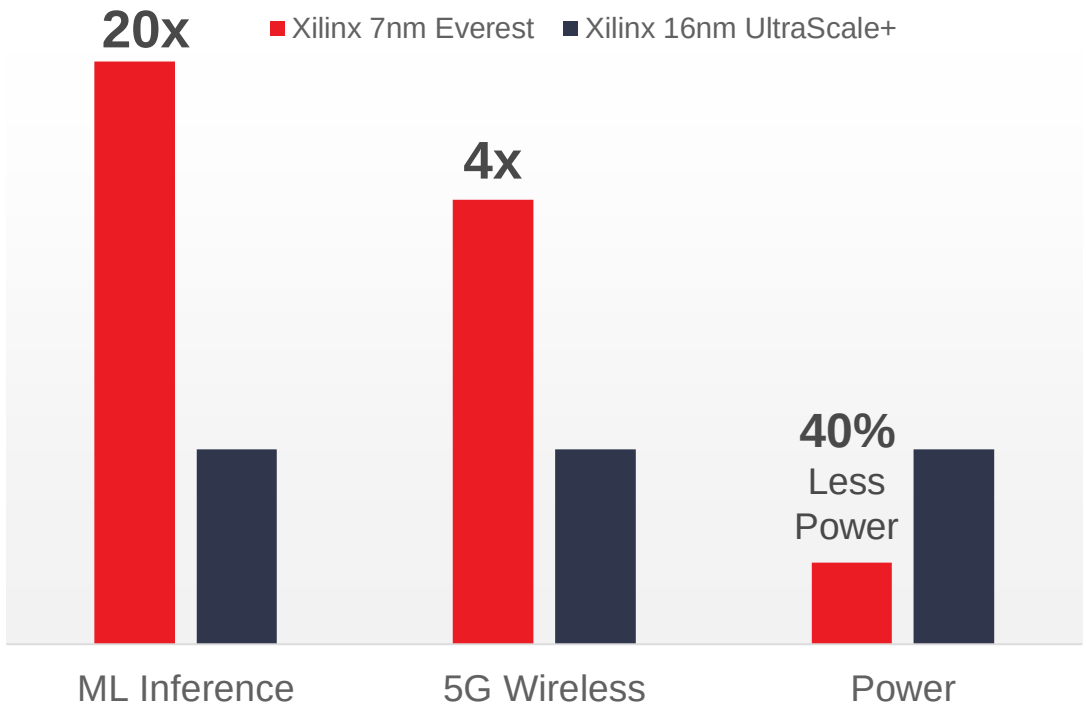
Heterogenous Computing Architecture

Everest Heterogeneous Architecture



- Compute Efficiency
- Reduced Power
- Software Programmable

Application-Level Performance Enabled by SW Programmable Engine



7nm Everest is the First Product to Integrate this New Architecture

Tape Out in 2018

© Copyright 2018 Xilinx

XILINX

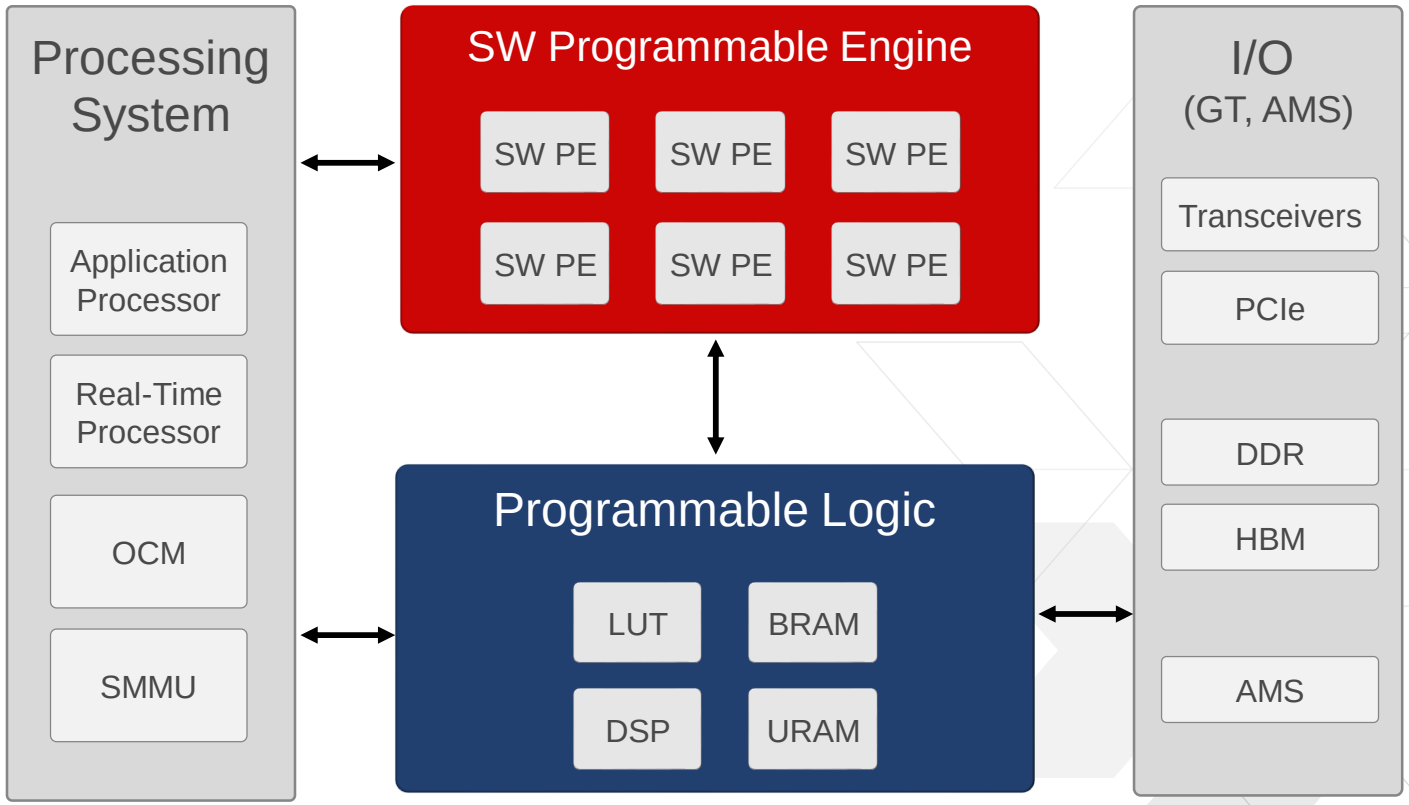
Everest Architecture Block Diagram

SW Programmable Engine

- Domain Specific Architecture
- Hardened 7nm technology
- Throughput-oriented, low-latency

Programmable Logic

- ‘Soft’ logic
- Flexibility
- Custom memory hierarchy



Chips with Different Implementations of the Architecture will be Announced Shortly

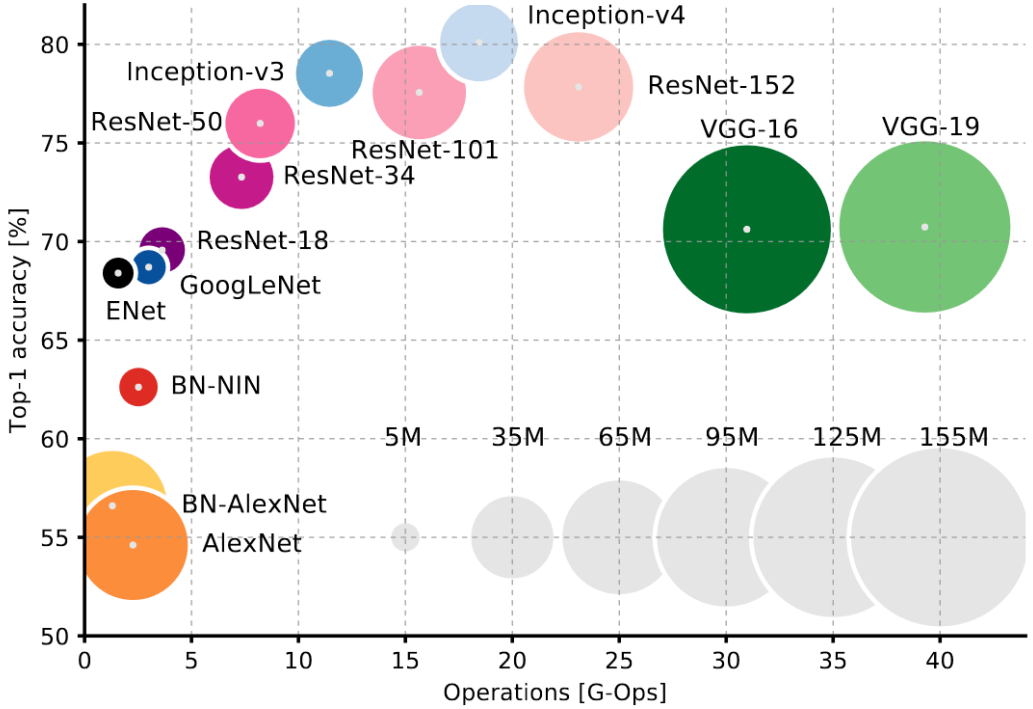
© Copyright 2018 Xilinx

XILINX.

Market Requirements and Trends: Data Center

Increasing Compute

Lower Latency Requirements



Source: Canziani et. al, "An Analysis of Deep Neural Network Models for Practical Applications"
https://arxiv.org/abs/1605.07678

Heterogeneous Workloads

ML Becoming an Essential Part of Data Processing

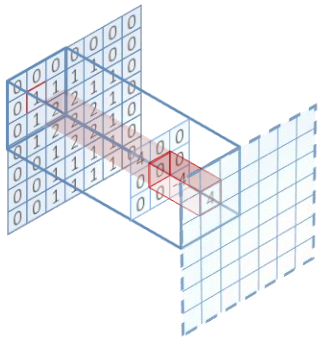
- Video + ML
- Genomics + ML
- Risk Modelling + ML
- Database + ML
- Network IPS + ML
- Storage + ML

© Copyright 2018 Xilinx

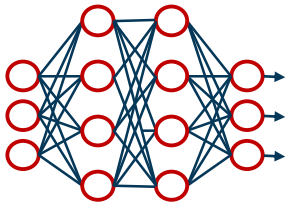
XILINX

ML Inference on Heterogenous Architecture

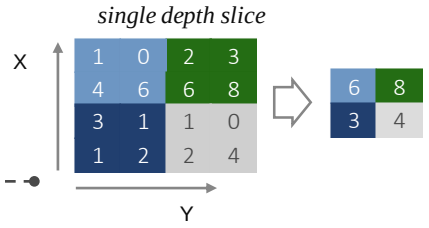
Convolutions



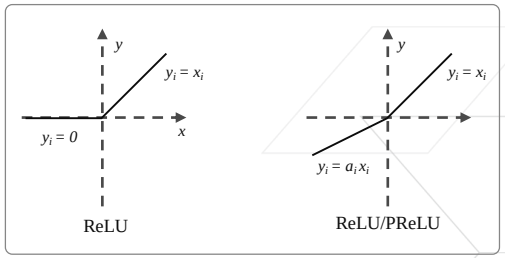
Fully Connected Layers

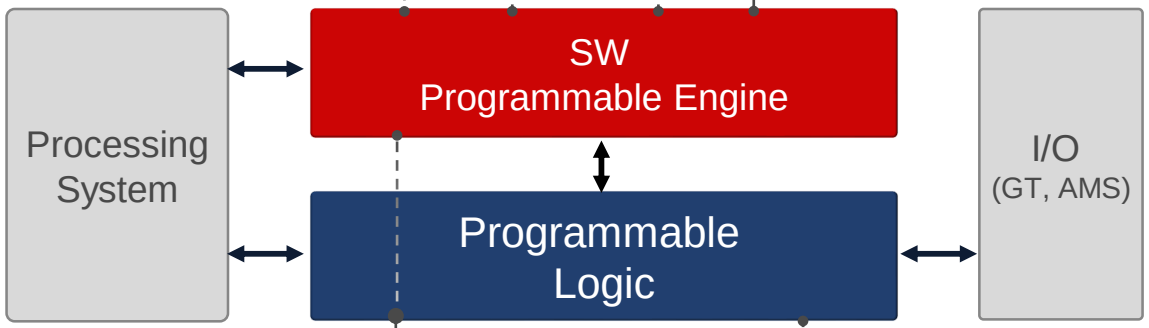


Pooling

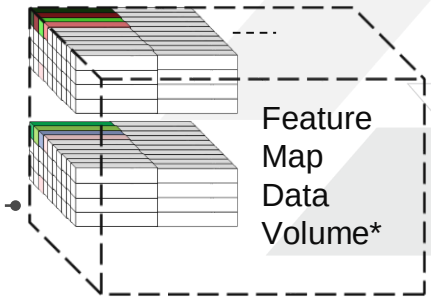


Activations





- Video
- Genomics
- Storage
- Database
- Network IPS
- Risk modeling



Custom Memory Hierarchy

*Figure credit: https://en.wikipedia.org/wiki/Convolutional_neural_network

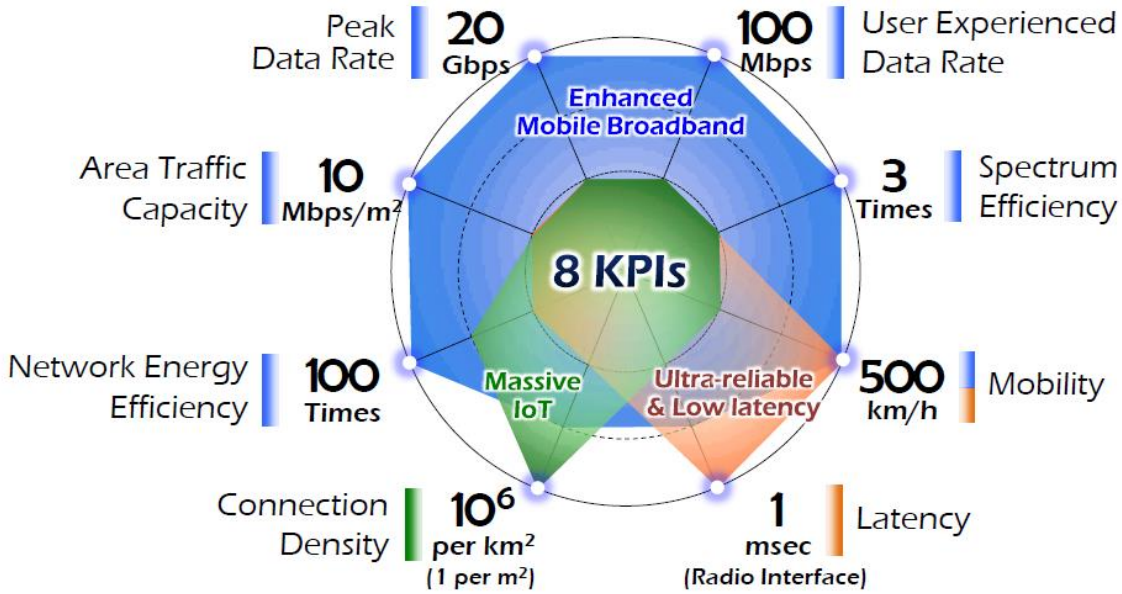
© Copyright 2018 Xilinx

XILINX

Market Requirements and Trends: Wireless 5G

5G Complexity is 100X that of 4G

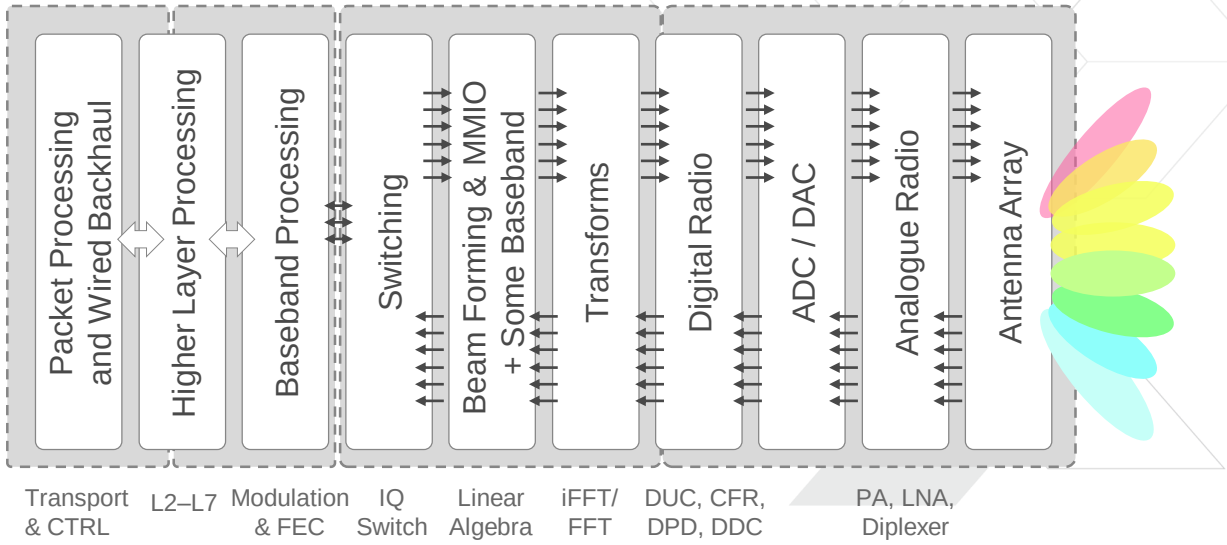
Still Evolving Standard



ETRI RWS-150029,
5G Vision and Enabling Technologies: ETRI Perspective 3GPP RAN Workshop
Phoenix, Dec. 2015
http://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_70/Docs

New Technologies in 5G

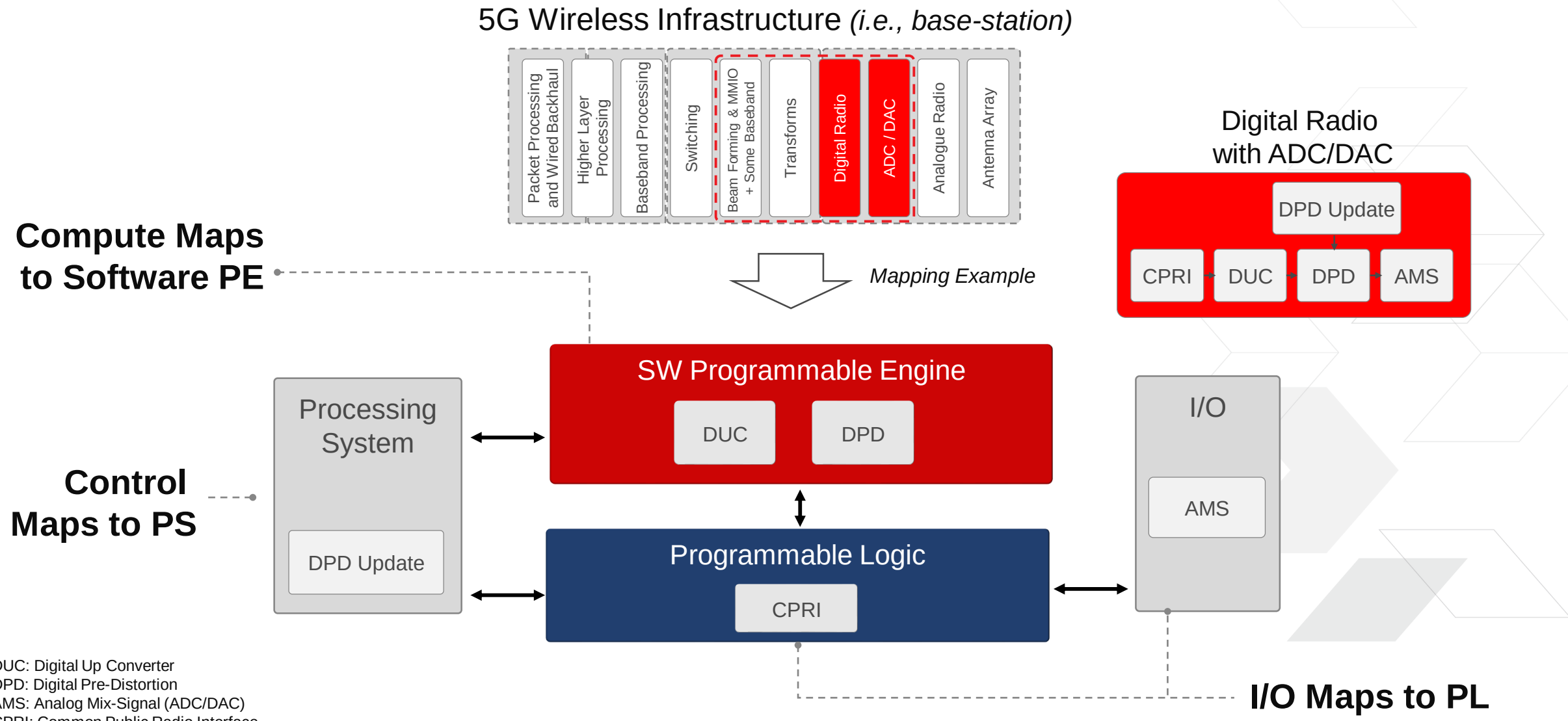
- Massive MIMO
- Multiple antenna, frequency bands
- Changing functional partitioning



© Copyright 2018 Xilinx

XILINX.

5G Wireless on Heterogenous Architecture



1: DUC: Digital Up Converter
2: DPD: Digital Pre-Distortion
3: AMS: Analog Mix-Signal (ADC/DAC)
4: CPRI: Common Public Radio Interface

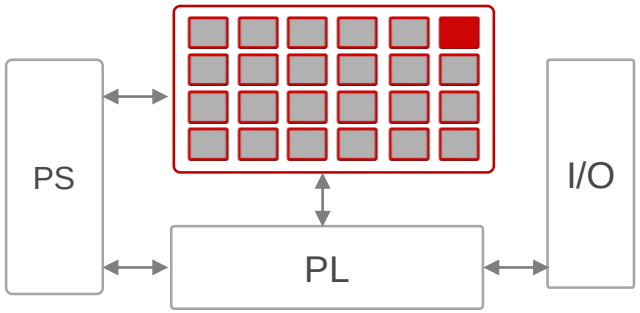
© Copyright 2018 Xilinx

XILINX.

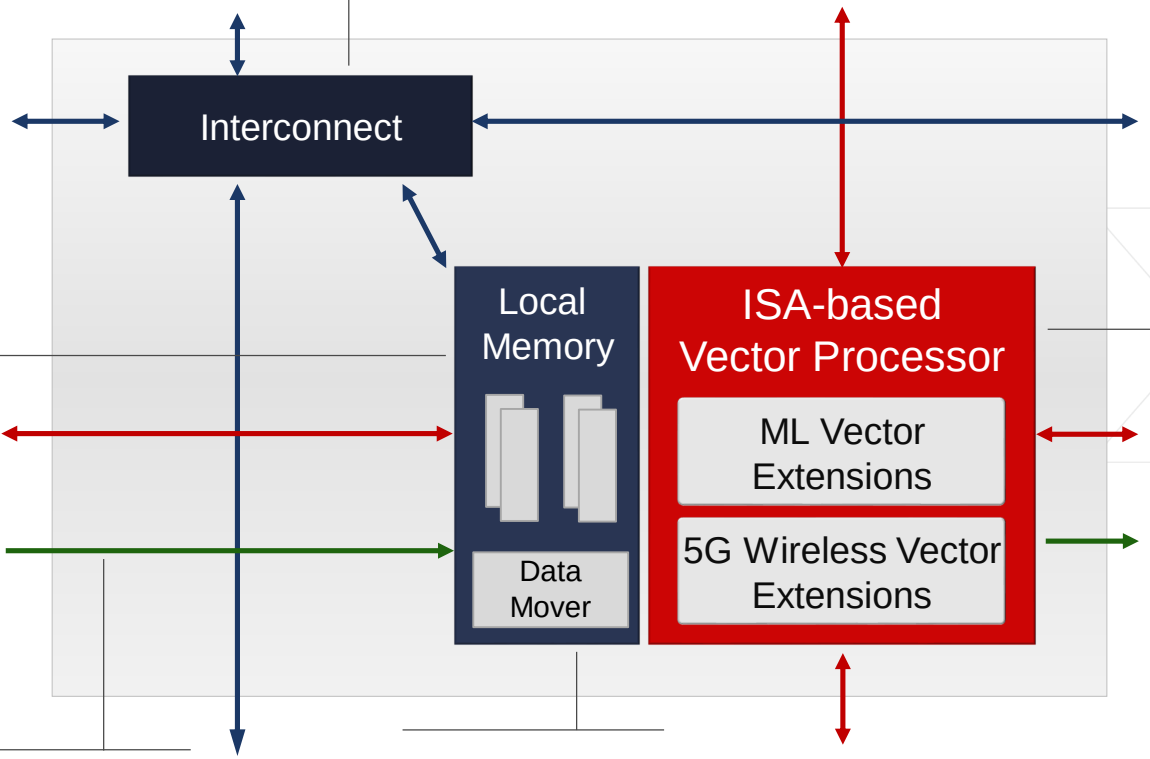
Architecture Overview

© Copyright 2018 Xilinx

Tile-Based Architecture



Non-Blocking Interconnect
Up to 200+ GB/s bandwidth per tile



Local Memory
Multi-bank implementation
Shared across neighbor cores

ISA-based Vector Processor
Software Programmable (e.g., C/C++)

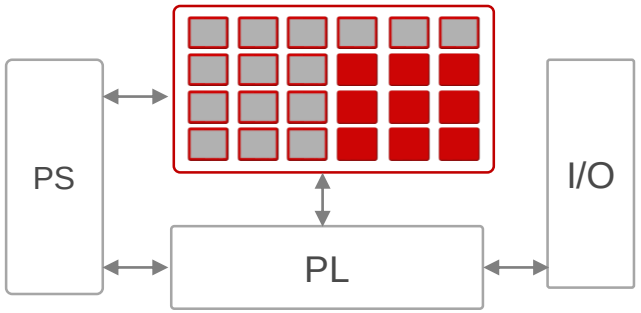
Cascade Interface
Partial results to next core

Data Mover
Non-neighbor data communication
Integrated synchronization primitives

© Copyright 2018 Xilinx

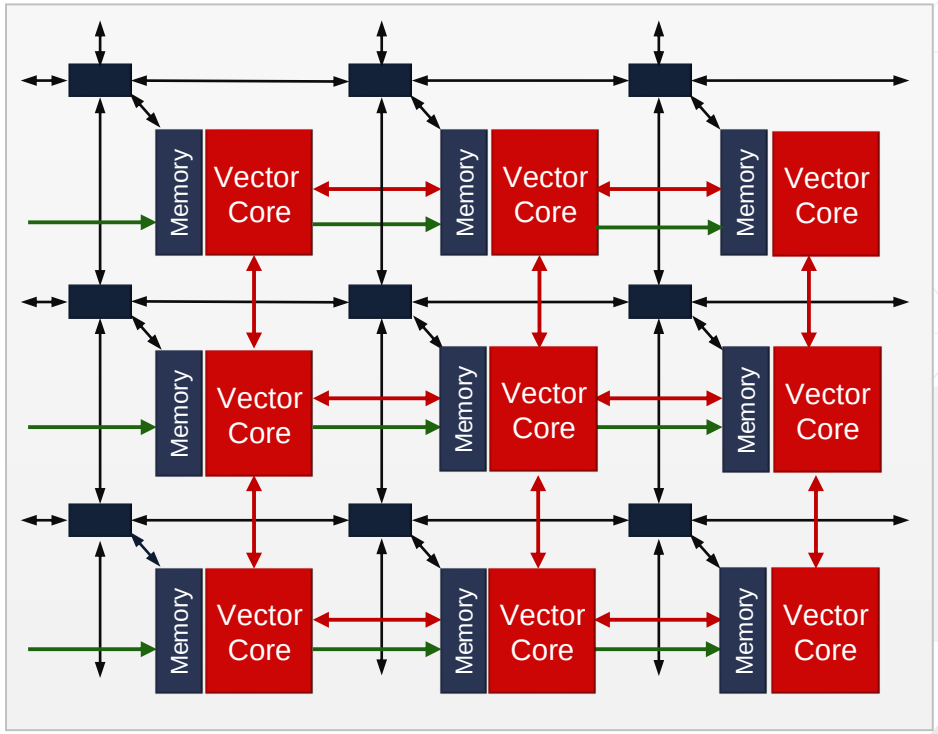
XILINX

Tile-Based Architecture (Continued)



Massive multi-core engines

- Increase in compute, memory and communication bandwidth



Modular and scalable architecture

- Instantiate multiple tiles for more compute
- 10's-100's Instances*

*Device Specific

Distributed memory hierarchy
Maximize memory bandwidth

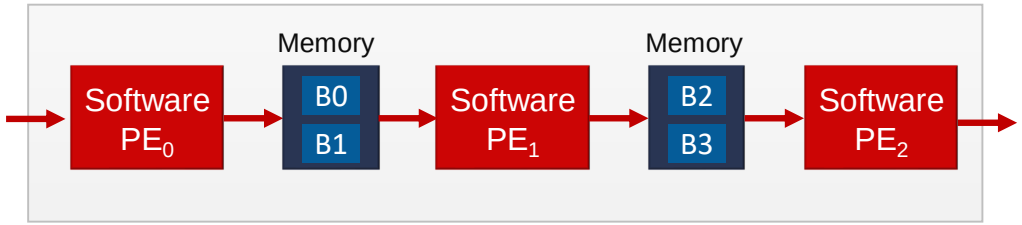
Architectural Focus: Deterministic Performance & Low Latency

© Copyright 2018 Xilinx

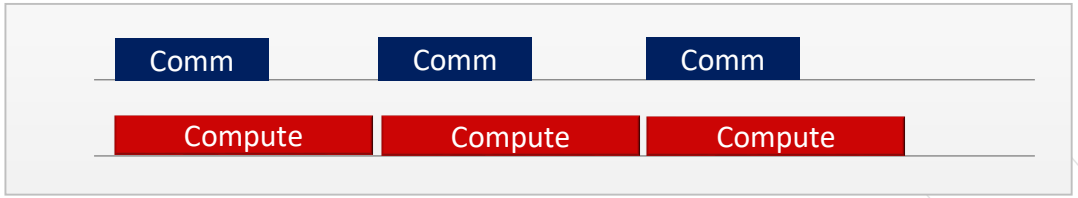
XILINX

Data Movement Architecture: Examples (1/2)

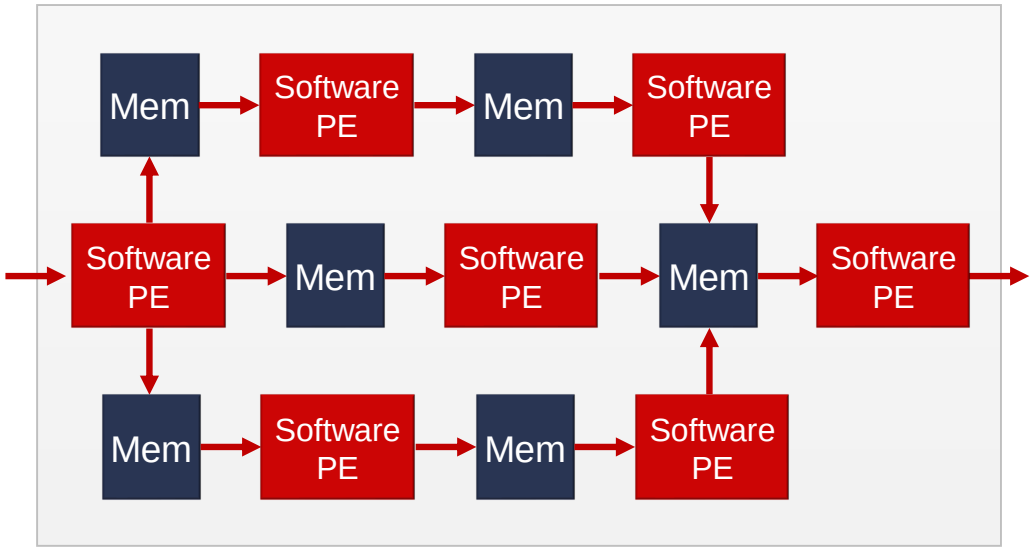
1 Dataflow Pipeline



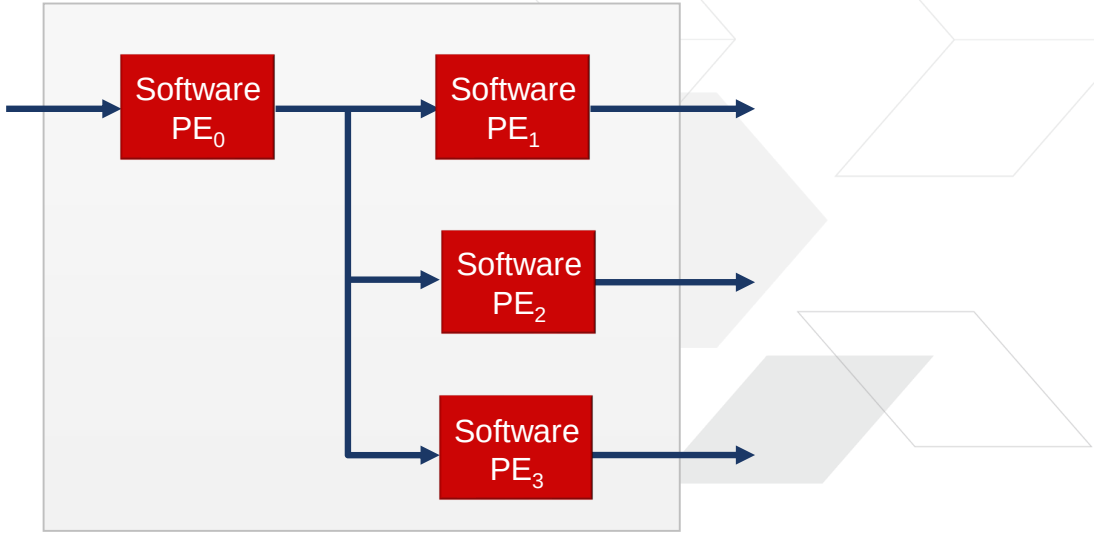
Overlap Compute and Communication



2 Dataflow Graph



3 Streaming Multicast

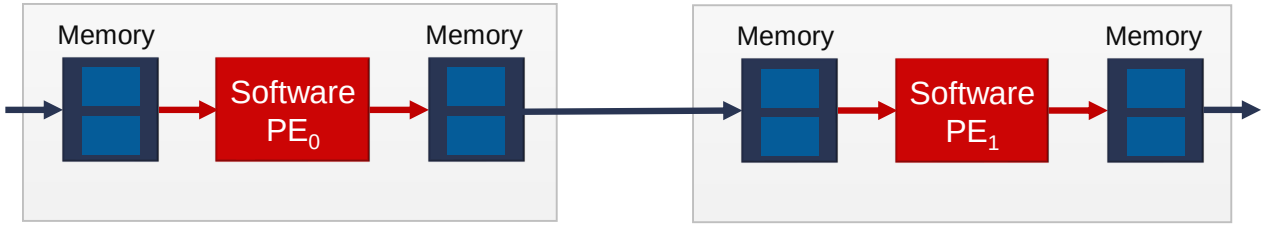


© Copyright 2018 Xilinx

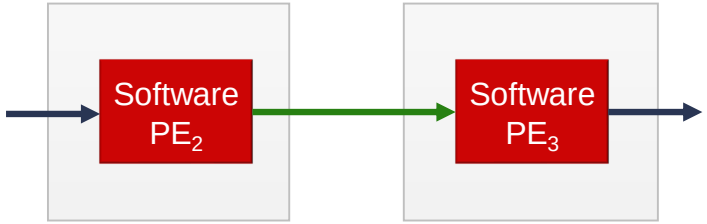
XILINX.

Data Movement Architecture: Examples (2/2)

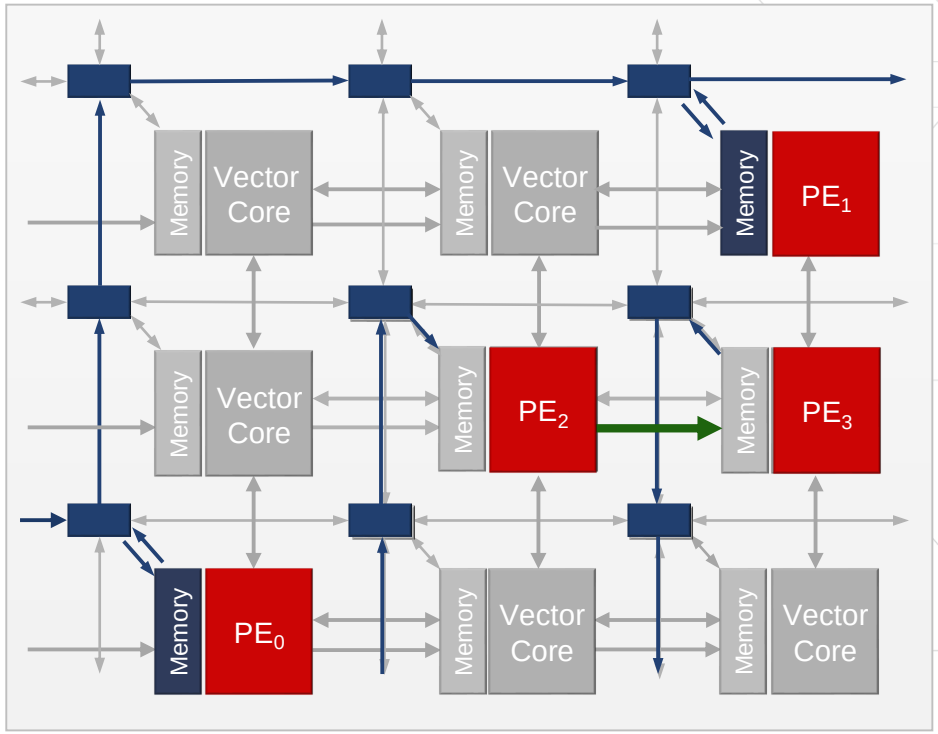
4 Non-neighbor communication



5 Cascade streaming



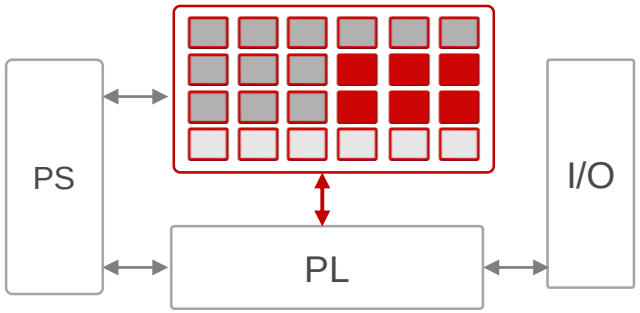
Physical Mapping

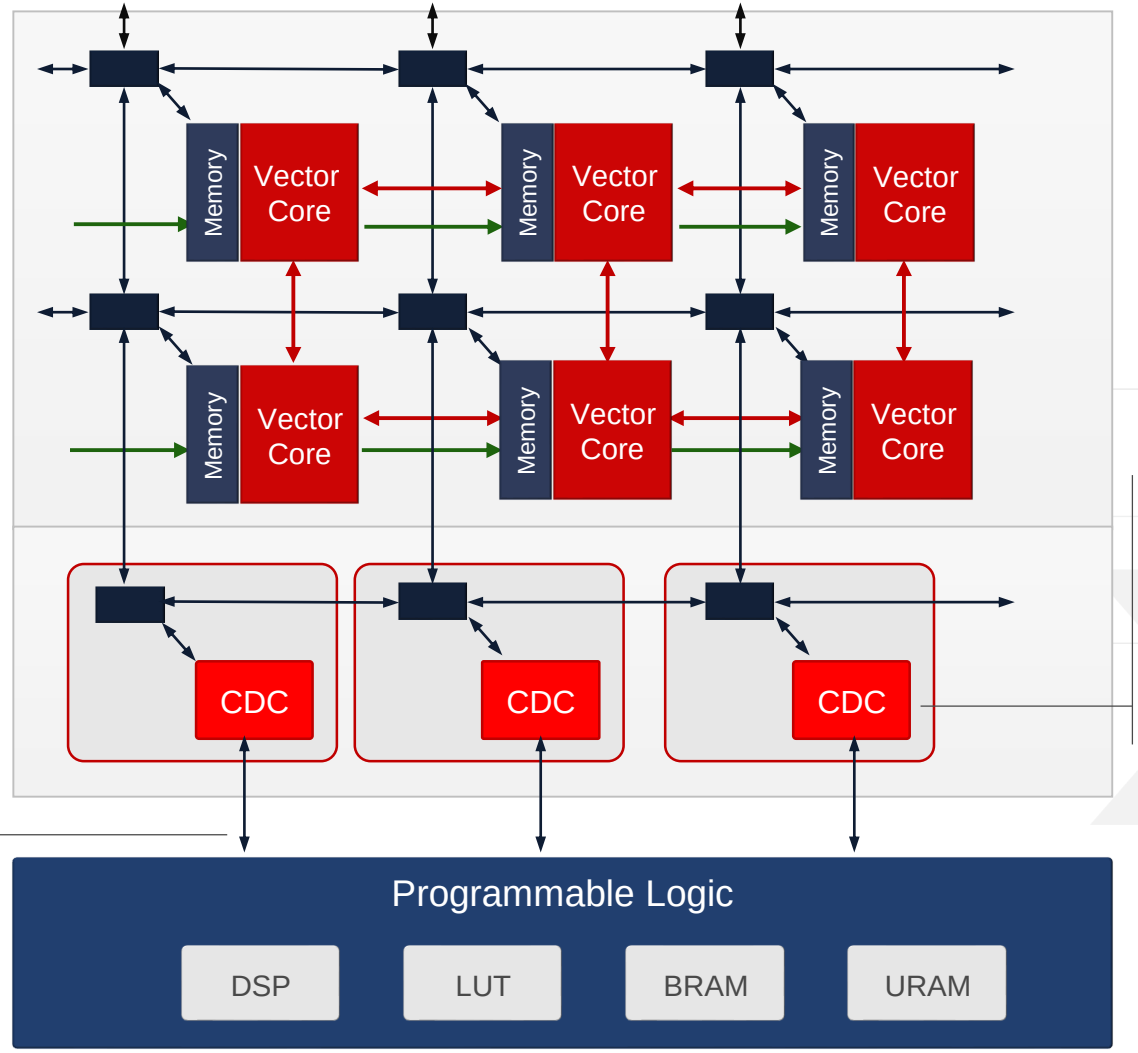


© Copyright 2018 Xilinx

XILINX.

Device Integration





Clock-domain crossing (CDC) between PL & Software PE clocks

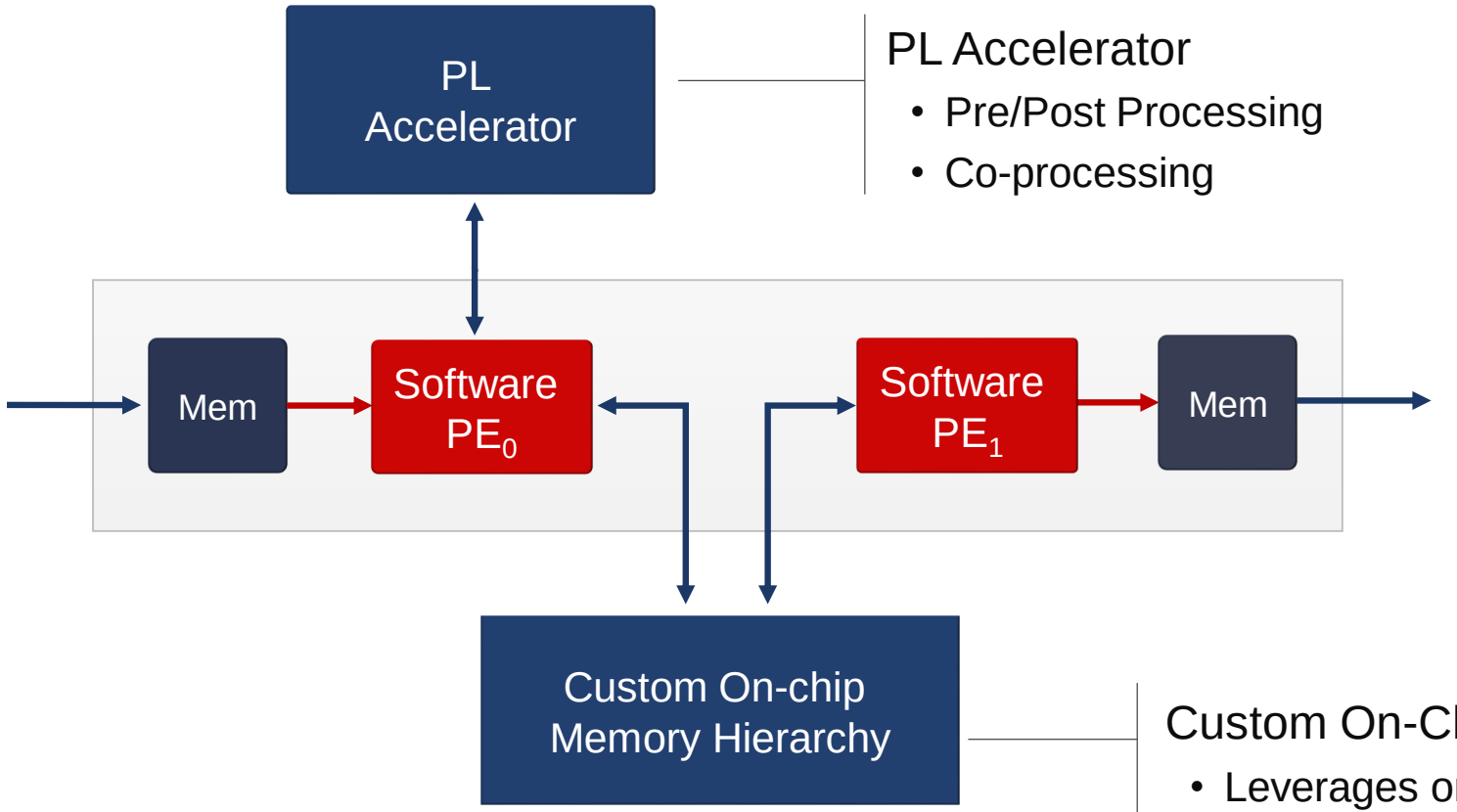
Up to TB/s of bandwidth between Software PE and PL*

*Device Specific

© Copyright 2018 Xilinx

XILINX.

Data Movement Architecture: Examples PL Integration



PL Accelerator

- Pre/Post Processing
- Co-processing

Custom On-Chip Memory Hierarchy

- Leverages on-chip memory resources in PL (e.g., BRAM, URAM)
- High on-chip memory bandwidth and lower memory access latency

© Copyright 2018 Xilinx

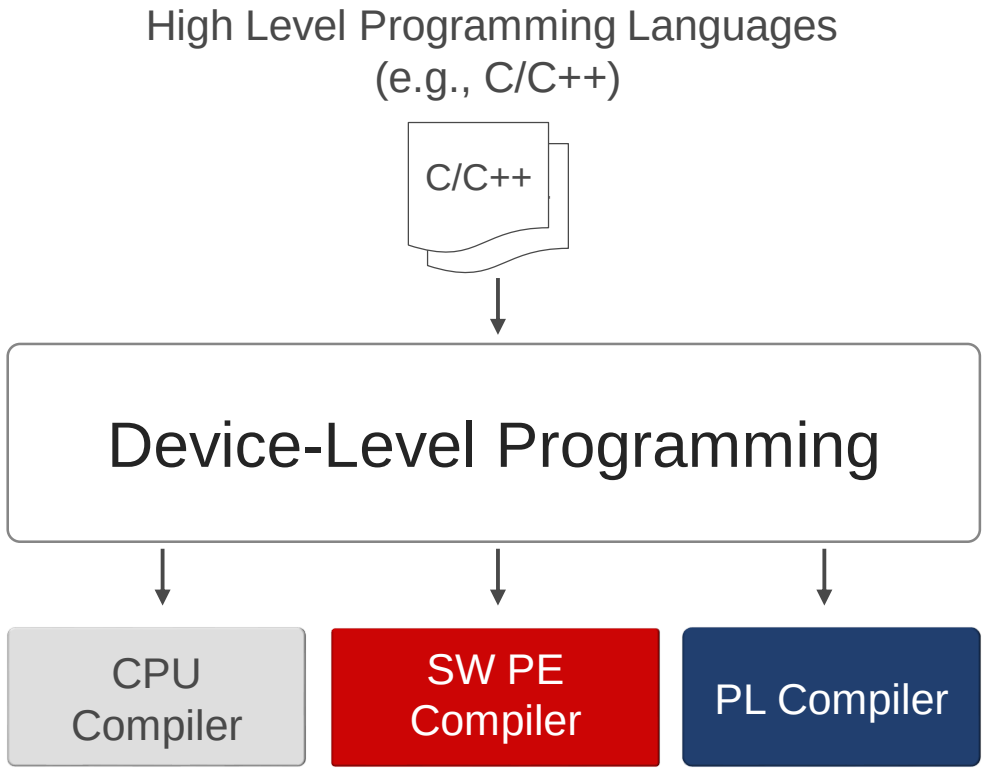
Programming Environment

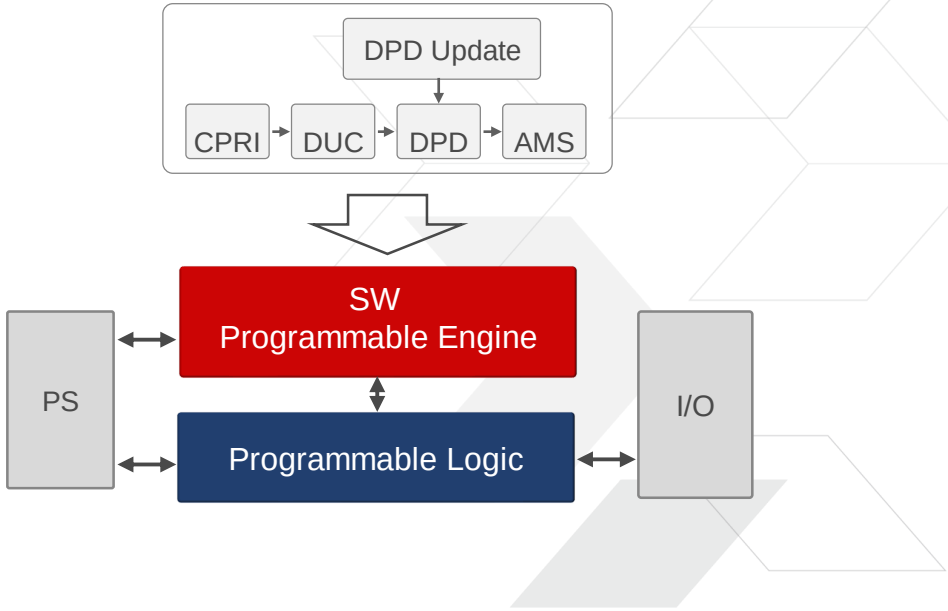
© Copyright 2018 Xilinx

Programming Environment

Full Device Programming Solution

Fully Integrated with Tools for Other Everest Fabrics

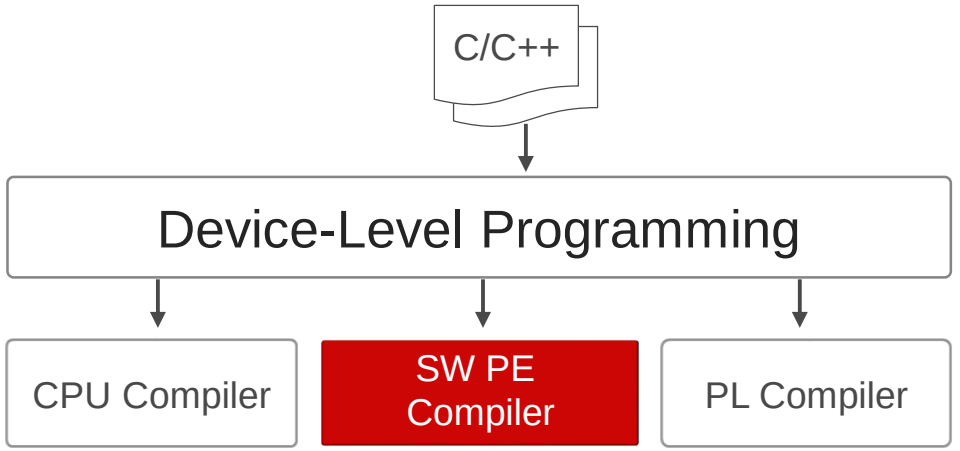




© Copyright 2018 Xilinx

XILINX.

Complete Software Development Stack



1 Full SW Programming Tool Chain
(Single-Core and Multi-Core)

- IDE
- Compiler
- Debugger
- Performance Analysis

2 Performance-Optimized Software Libraries
(Examples)

- Wireless 4G/5G Library
- ML (BLAS) Library
- Vision Library

3 Run-Time Software
(Examples)

- Error Management
- Memory Management
- Boot + Configuration
- Power/Thermal Management

© Copyright 2018 Xilinx

XILINX.

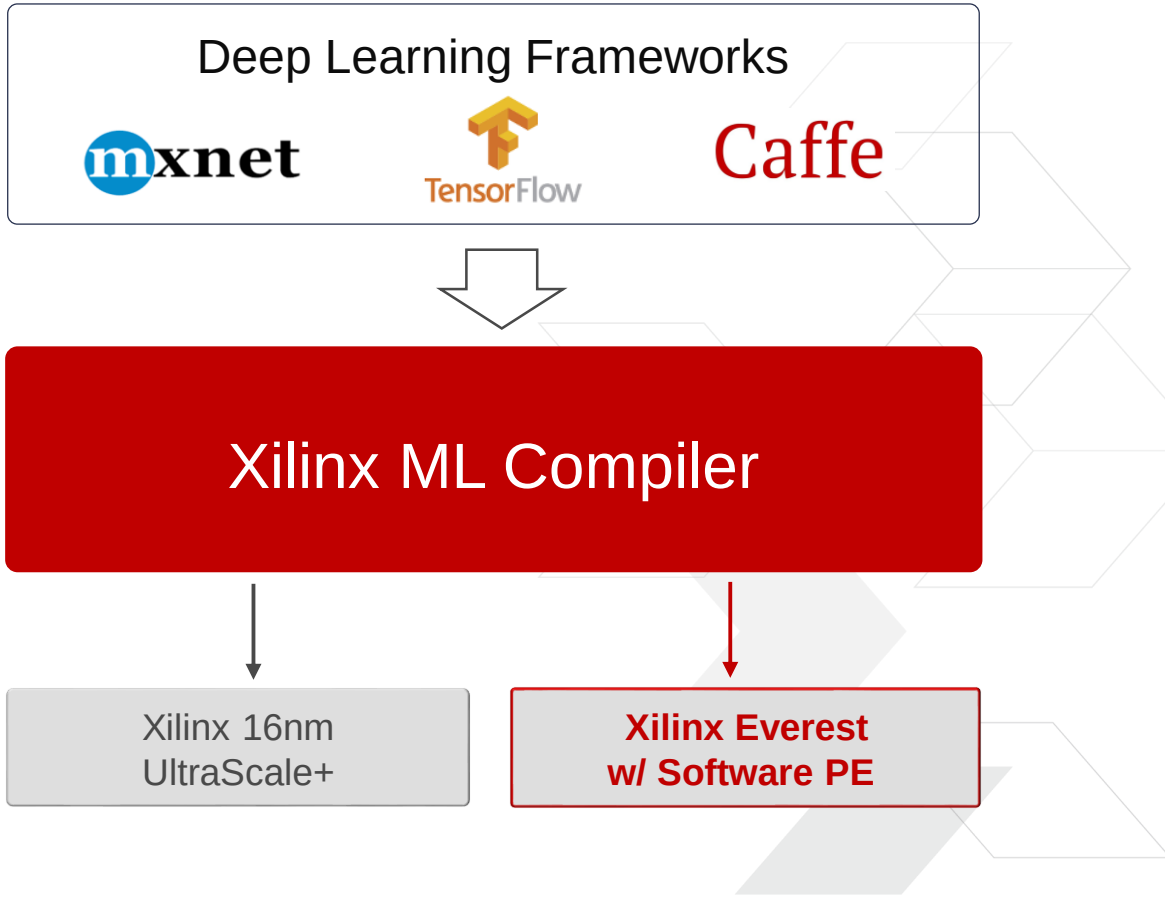
High-level Software Compiler for ML Inference

Users Working at ML Framework Level

Users don’t directly program Software PE

Seamless Migration

- Same experience as 16nm devices
- Current models continue to work

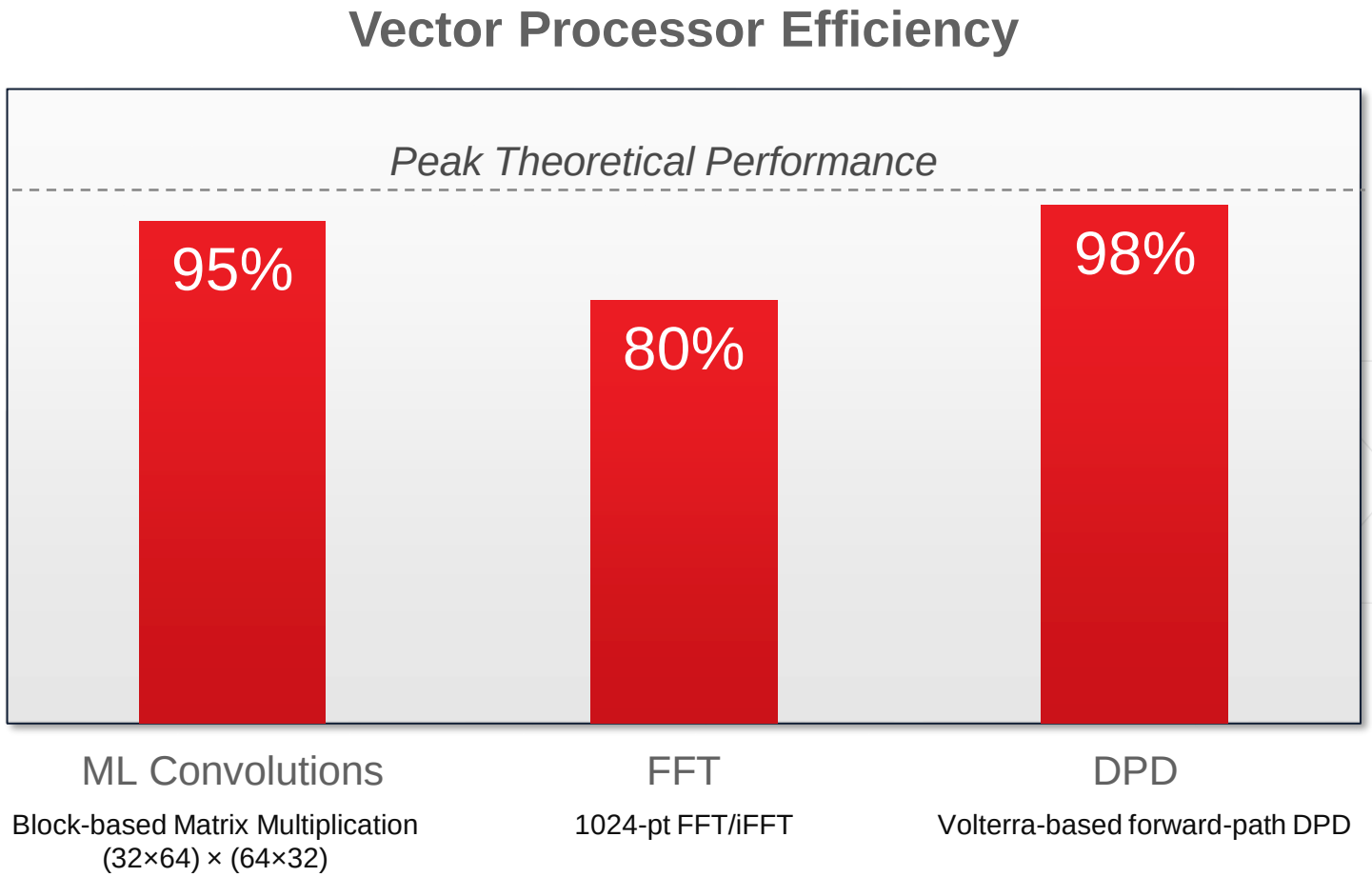


© Copyright 2018 Xilinx

Application Results

© Copyright 2018 Xilinx

Kernel-Level Performance: ML and Wireless 5G



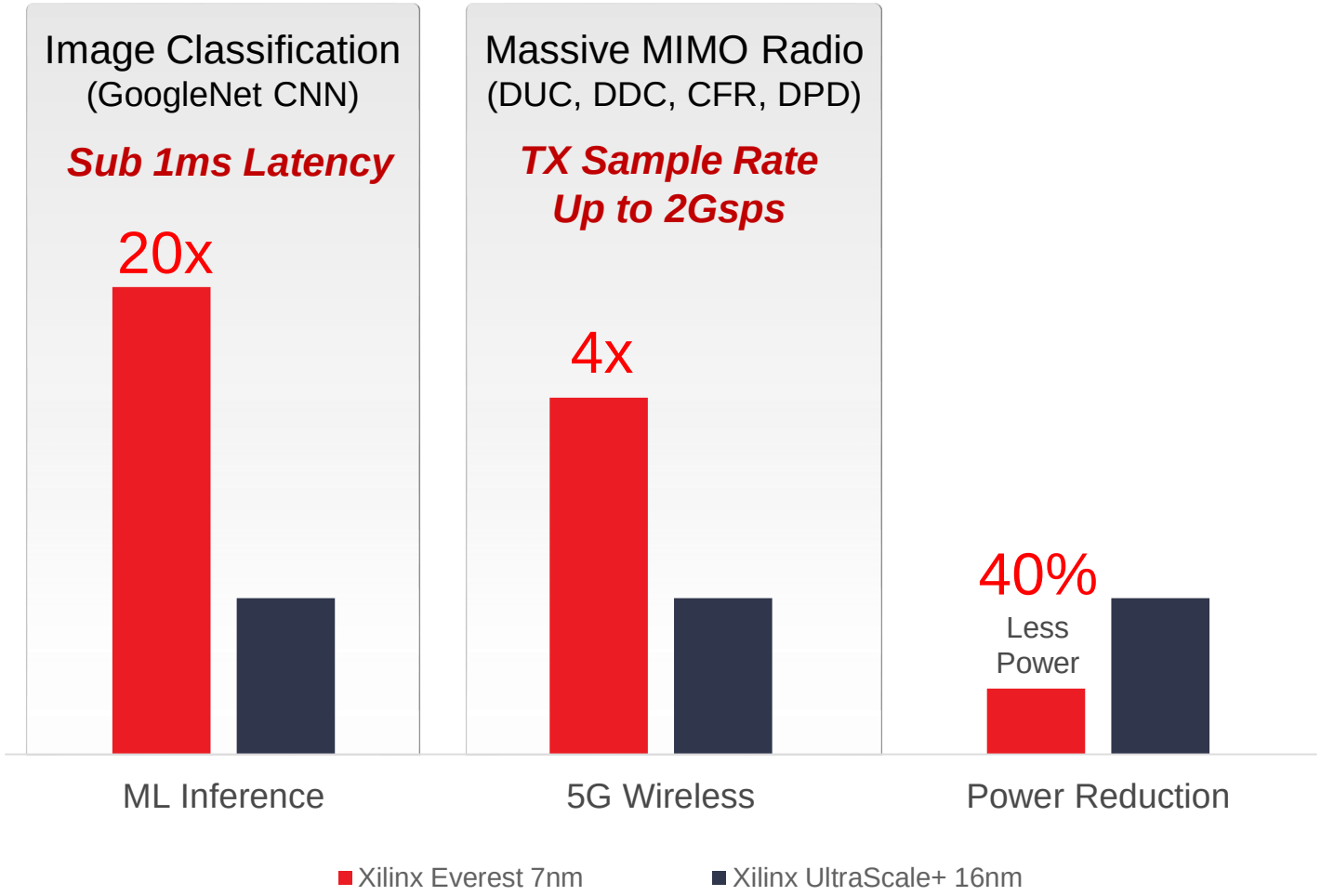
High Compute Efficiency for Key Functions in ML and Wireless 5G

© Copyright 2018 Xilinx

XILINX.

Software Programmable Engine: Summary

Application-level Performance Enabled by SW Programmable Engine



Compute Efficiency

- Domain specific architecture
- Increase in compute density
- Xilinx 7nm Everest

Heterogenous Architecture

- High-throughput, low-latency
- PL flexibility
- Custom memory hierarchy

Multiple Applications

- ML Inference for Cloud DC
- Wireless 5G: Radio, Baseband
- ADAS/AD embedded vision
- Wired: DOCSIS cable access

SW Programmable

- SW programmable (e.g., C/C++)
- Compile, execute, debug
- Optimized software libraries

© Copyright 2018 Xilinx

XILINX

HOT CHIPS

KEYNOTE

Adaptable Intelligence: The Next Computing Era

Victor Peng, Xilinx CEO

AUGUST 21ST @11:45 A.M.

XDF XILINX DEVELOPER FORUM

Connect • Learn • Share

XDF connects software developers and system designers to the deep expertise of Xilinx engineers, partners, and industry leaders.

Silicon Valley October 1-2

Learn More www.xilinx.com/xdf

© Copyright 2018 Xilinx

Adaptable.
Intelligent.

XILINX

© Copyright 2018 Xilinx