HW/SW Programmable Engine:

Domain Specific Architecture for Project Everest


21st August 2018
Heterogenous Computing Architecture

Everest Heterogeneous Architecture

Processing System

SW Programmable Engine

Programmable Logic

I/O (GT, AMS)

> Compute Efficiency
> Reduced Power
> Software Programmable

Application-Level Performance Enabled by SW Programmable Engine

20x
> ML Inference

4x
> 5G Wireless

40%
> Power

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

Tape Out in 2018
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
Market Requirements and Trends: Data Center

Increasing Compute
Lower Latency Requirements

Heterogeneous Workloads
ML Becoming an Essential Part of Data Processing

Source: Canziani et. al, “An Analysis of Deep Neural Network Models for Practical Applications”
ML Inference on Heterogenous Architecture

Convolutions

Fully Connected Layers

Pooling

Activations

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

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

© Copyright 2018 Xilinx
Market Requirements and Trends: Wireless 5G

5G Complexity is 100X that of 4G
Still Evolving Standard

- Peak Data Rate: 20 Gbps
- User Experienced Data Rate: 100 Mbps
- Area Traffic Capacity: 10 Mbps/m²
- Spectrum Efficiency: 3 Times
- Network Energy Efficiency: 100 Times
- Connection Density: 10⁶ per km² (1 per m²)
- Mobility: 500 km/h
- Latency: 1 msec (Radio Interface)
- Mobility
- Massive MIMO
- Changing functional partitioning

New Technologies in 5G
- Massive MIMO
- Multiple antenna, frequency bands
- Changing functional partitioning

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

© Copyright 2018 Xilinx
5G Wireless on Heterogenous Architecture

5G Wireless Infrastructure (i.e., base-station)

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
Architecture Overview
Tile-Based Architecture

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

Local Memory
Multi-bank implementation
Shared across neighbor cores

Cascade Interface
Partial results to next core

Data Mover
Non-neighbor data communication
Integrated synchronization primitives

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

ML Vector Extensions
5G Wireless Vector Extensions

PS
I/O

PL
Tile-Based Architecture (Continued)

**Modular and scalable architecture**
- Instantiate multiple tiles for more compute
- 10’s-100’s Instances*

*Device Specific

**Distributed memory hierarchy**
Maximize memory bandwidth

**Massive multi-core engines**
- Increase in compute, memory and communication bandwidth

**Architectural Focus:** Deterministic Performance & Low Latency
Data Movement Architecture: Examples (1/2)

1. Dataflow Pipeline

```
Software PE0  Memory  B0  Software PE1  Memory  B2  Software PE2
```

2. Dataflow Graph

```
Mem  Software PE  Mem  Software PE  Mem  Software PE  Mem  Software PE
```

3. Streaming Multicast

```
Software PE0  Software PE1
Memory
Software PE2  Software PE3
```

Overlap Compute and Communication

```
Compute  Compute  Compute
```

© Copyright 2018 Xilinx
Data Movement Architecture: Examples (2/2)

4 Non-neighbor communication

5 Cascade streaming
Device Integration

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

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

*Device Specific
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
Programming Environment
Programming Environment

Full Device Programming Solution

_Fully Integrated with Tools for Other Everest Fabrics_
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
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
Application Results
Kernel-Level Performance: ML and Wireless 5G

Vector Processor Efficiency

Peak Theoretical Performance

- **ML Convolutions**: 95%
  - Block-based Matrix Multiplication
  - (32×64) × (64×32)

- **FFT**: 80%
  - 1024-pt FFT/iFFT

- **DPD**: 98%
  - Volterra-based forward-path DPD

High Compute Efficiency for Key Functions in ML and Wireless 5G
Software Programmable Engine: Summary

Application-level Performance Enabled by SW Programmable Engine

Image Classification (GoogleNet CNN)

- **Sub 1ms Latency**
- **20x**

Massive MIMO Radio (DUC, DDC, CFR, DPD)

- **TX Sample Rate Up to 2Gsps**
- **4x**
- **40% Less Power**

ML Inference

5G Wireless

Power Reduction

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
Adaptable Intelligence: The Next Computing Era
Victor Peng, Xilinx CEO

August 21st @11:45 a.m.

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
Adaptable.
Intelligent.