A New DSP Approach for 5G and AI

Albert Camilleri
VP Business Development North America
VSORA Inc.
Company Background

• Company founded in 2015
• Headquarters: France
• Each founder has more than 10 years experience in Digital Signal Processor (DSP) design, working in global consumer markets
• Previous founders’ designs widely used in successful consumer, automotive and industrial high volume products
Reinventing ‘Digital Signal Processing’ (DSP)

5G Wireless Communications
- mmWave, MiMo, Beamforming, Carrier Aggregation
- Enhanced 1Gbps+ Mobile Broadband
- Massive Machine Type Comms, Smart Home / Cities
- Ultra reliable low latency comms (< 1ms), IoT
- New Short Range Wireless, 802.11af, ay, bb (LiFi)
- Both terminals and infrastructure

Artificial Intelligence (Terminals / Edge)
- Neural Networks
- Image / video
- Speech recognition / Audio
- Language Translation
Traditional Architecture Limits Flexibility

- Single threaded processors falling further behind 1 Gbps+ demand
- Bespoke, fixed algorithm, co-processors increase the well known ASIC problems
- Inflexible, hard to mature quickly, inappropriate in the new world of rapid standards evolutions
The Memory Bottleneck Problem

Signal Memory bottleneck will stall and limit the promise of 5G and AI

- Need for ever greater symbol word length and depth
- Signal Memory (Cache) I/O bandwidth explosion
- 5G modems and Massively Parallel Neural Network Processors are predominantly built on the same DSP type architectures today
Introducing the Matrix Processor Unit (MPU)

- Completely configurable:
  - Number of ALUs
  - Memory size
  - Quantization (IEEE754 like), i.e. number of exponent/mantissa bits

- Liberates the “Bottleneck”
  - Signal (cache) memory more tightly coupled
  - Signals manager pre-configures signal data

- DSP is tightly controlled by the host processor
Single-core / Multi-core Architecture

- MPUs are programmed at an algorithm level in C++ with a MATLAB like API

- High-level simulation methodology provides performance/power/area trade-off data
  - Can be modified and iterated at the algorithmic level to attempt 100% DSP utilization

- Algorithm code compiled directly to DSP via modified LLVM compiler
  - No low level code required
  - Engineering productivity enhancer

- Completely configurable in terms of:
  - The number of cores (single/multi-core)
  - The number of DMAs/core

Ability to map complex systems onto multiple cores, and dimension optimal solutions.
AI Supported Frameworks

VSORA AI Framework Load Tool
Graph Optimization
Model Quantization

VSORA Library

Compiler

VSORA AI-DSP

High BW Signal Memory

Signals Manager

The information contained in this document is confidential and shall not be disclosed to third parties without a written consent from Vsora.
VSORA AI Solution

• Fully programmable Solution
  • TensorFlow, PyTorch, …, supported frameworks
• Configurable:
  • Number of MACs: 256, 1024, 2304, 4096, 6400, 9216, 12544, 16384, …, 65536
  • IEEE754 Quantization: number of bits (sign/exponent/mantissa)
  • Number of DMAs
• High MPU processing efficiency
  • Does not suffer memory bandwidth bottleneck to load large numbers of MACs

The information contained in this document is confidential and shall not be disclosed to third parties without a written consent from Vsora
The information contained in this document is confidential and shall not be disclosed to third parties without a written consent from Vsora.

**Reinvented Development Flow**

**Benefits**
- Reduced personnel
- Fast algorithm definition and DSP dimensioning
- Easy integration of Signal Processing & Embedded SW code

**Drawbacks**
- Four different, large engineering teams
- Very slow process, exceedingly expensive
Summary

Highly configurable “tiled” solution
• “Unlimited” number of Cores
• Scalable memory/DMA bandwidth avoids bottlenecks

Eliminates need for inflexible co-processors
• Flexible coding: mix signal processing and link-layer/neural-processing SW

Implementation independent, high-level programmability
• Supports design flexibility to facilitate market evolution

Tiered simulation platforms
• MATLAB/Tensorflow level, FPGA (Cloud) platform, IP/RTL simulation

Compiler technology empowers 100% DSP utilization
• Optimizes engineering efficiency
• Facilitates performance/area/power tradeoffs
Thank You