Efficient AI Inference at the Edge with Imagination IP

April 2018
Imagination: A Global Technology Leader

A technology powerhouse for multimedia, communications and AI IP

An industry leader since 1985

Developing innovative IP

- Leader in embedded Graphics and GPU compute IP
- Leading the advance in dedicated Vision and AI IP
- Leader in RPU communications IP

Delivering exceptional service

- Enabling very fast time to market
- Enabling customers to leverage IP to maximise differentiation

Driving major markets

- Helping our partners to create successful solutions
- Influencing new and emerging opportunities
- Showcasing and proving our technology with real products
Business Model

Our royalty based business model means we only succeed if our customers succeed
Licensees and Partners

Key Licensees

Strategic Partners

© Imagination Technologies
Neural Network Landscape & Growth Opportunity

Large variety of Edge AI Applications across many different Markets

- **IoT**
  - Natural speech synthesis
  - Speech understanding
  - Gesture Analysis
  - Face recognition
  - Age/gender recognition
  - Medical diagnostics

- **Mobile**
  - Natural speech synthesis
  - Speech understanding
  - Gesture Analysis
  - Face recognition
  - Age/gender recognition
  - Medical diagnostics

- **AR/VR**
  - Feature description
  - Movement Analysis
  - Feature detection
  - Gesture Analysis
  - Eye Tracking
  - Scene Understanding

- **Drone**
  - Collision avoidance
  - Subject tracking
  - Camera optimisation
  - Gesture recognition

- **Industrial**
  - Defect detection
  - Object detection
  - Context awareness

- **Automotive**
  - Driver alertness monitoring
  - Driver gaze tracking
  - Seat Occupancy
  - Road Sign Detection
  - Drivable Path Analysis
  - Road User Detection
  - Driver Recognition

- **Security**
  - Face detection/recognition
  - People counting
  - Doorway surveillance
  - Crowd/loitering detection
  - Hazard detection
  - Multi-camera tracking

© Imagination Technologies
Imagination Product Portfolio

Imagination
The best solution for embedded graphics, vision, AI and communications

PowerVR GPU
Leading graphics IP cores for embedded devices

PowerVR Vision & AI
Dedicated AI and Computer Vision IP Products

Ensigna
Connectivity and broadcast communications
High performance, low power

XE/XM GPU
Focused Features
Fillrate/mm² & Performance/mm²

XT GPU
Feature rich
Performance/mW

PowerVR 2NX
Neural Network Accelerator
Performance/mm² Performance/mW

PowerVR ISP
Low area, high quality & highly power efficient
Multi camera capable ISP

RF
Wi-Fi, Bluetooth

Connectivity
Wi-Fi, Bluetooth IEEE 802.15.4 GNSS

Broadcast
TV, Digital Radio
## Edge Neural Network Processing Resources

### Why GPU and Neural Network Accelerators are Best

<table>
<thead>
<tr>
<th>Resource</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>Fully Flexible</td>
</tr>
<tr>
<td></td>
<td>BUT inefficient and slow for high compute workloads</td>
</tr>
<tr>
<td>DSP</td>
<td>Fully Flexible</td>
</tr>
<tr>
<td></td>
<td>BUT hard to program – no standardisation, INT focussed</td>
</tr>
<tr>
<td>GPU</td>
<td>Fully Flexible</td>
</tr>
<tr>
<td></td>
<td>Standardised APIs for Compute, Float and INT support</td>
</tr>
<tr>
<td>Neural Network</td>
<td>Configurable</td>
</tr>
<tr>
<td>Accelerator</td>
<td>Lowest power with domain specific flexibility</td>
</tr>
<tr>
<td>Fixed Function</td>
<td>Single usage case</td>
</tr>
<tr>
<td></td>
<td>Lowest power BUT zero flexibility</td>
</tr>
</tbody>
</table>

Efficiency sweetspot
PowerVR NNA Leading Position

*Only a dedicated hardware solution offers required PPA for long battery life*

- PowerVR 2NX NNA
- Other solutions

Area of circles represents area – smaller = better

- PowerVR NNA
- PowerVR NN

Flexibility Costs

PPA Efficiency

CPU/GPU/DSP Efficiency

Accelerators

Specific Accelerator

Use with existing programmable cores if needed

© Imagination Technologies
PowerVR 2NX NNA – Architecture and Features

IP core to enable efficient inferencing of neural networks in SoCs at the “edge”

- **Scalable architecture**
  - 2048 8-bit MACs/clock equals 4Tops/core
  - Multi-core scalable achieves beyond performance
  - Architected for creation of future cores with different performance points and feature sets to address multiple markets and applications

- **Flexible bit-depth data type support**
  - 16, 12, 10, 8, 7, 6, 5, 4-bit support – covering different markets taking advantage of the benefits of lower precision
  - Per layer adjustment for both weights and activations
  - Maximum performance at minimum power and bandwidth

- **Variable precision internal data formats**
  - High precision, where required, inside the accumulator
  - Quantisation for following layers for efficient power/processing efficiency
  - Ensures optimum accuracy for results
Research on Low bit Benefit

*PowerVR 2NX – precision flexibility for optimised performance, power and bandwidth*

<table>
<thead>
<tr>
<th>Only fully connected weights (bits)</th>
<th>Relative inference/s</th>
<th>Relative bandwidth</th>
<th>Relative energy/ inference</th>
<th>Relative accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>4</td>
<td>160%</td>
<td>54%</td>
<td>69%</td>
<td>99%</td>
</tr>
</tbody>
</table>

PowerVR precision flexibility enables 60% performance increase, at 54% of the bandwidth, 69% of the power with less than 1% drop in accuracy.

PowerVR precision flexibility means requiring as little as 25% of the bandwidth compared with competing solutions.
Flexibility on Low-bit

Maximises performance, maintains accuracy, minimises power and bandwidth

- PowerVR 2NX NNA supports variable (including low) precisions for data and weights
- High Internal precision maintains network accuracy
- Allows higher performance at lower bandwidth and power
- Configurable output format enables CPU/GPU/DSP compatibility
- PowerVR 2NX is the only solution on the market with this level of flexibility – unique benefit
Easy Going on PVR Platform

It’s as easy as 1 .. 2 .. 3 - PowerVR 2NX workflow

1. Network Training
   - Network Model
   - Large Training Data
   - Tensorflow Caffe
   - Trained Model

2. (Optional) Network Optimisation
   - Subset of Training Data
   - PowerVR Optimised Framework
   - Tuned Model

3. NNA mapping and Implementation
   - PowerVR 2NX Accelerator Configuration
   - PowerVR NX Mapping Tool
   - Control/Data for PowerVR NX Driver
Performance & Bandwidth (compiler)

It's as easy as 1 .. 2 .. 3 - PowerVR 2NX workflow

1. Training
   - Network Model
   - Large Training Data
   - Floating point 32-bits

2. Tuning (Optional)
   - Trained Model
   - Subset of Training Data
   - Quantisation 4 to 16 bits

3. Mapping
   - (Tuned) Model
   - PowerVR Configuration
   - Optimal usage of Hardware (Bandwidth, Performance Tuning)
Cross Platform API

Training, network types and models, frameworks, inference engines, APIs …
Complete Imagination AI System

PowerVR GPU, PowerVR NNA (Neural Network Accelerator)

On Chip Memory
- Optional
- Reduces bandwidth of Neural Network Processing
- Can be used by other system components and other usage scenarios e.g. GPU Graphics

PowerVR GPU
- Reuse existing hardware (if applicable)
- Fast performance
- No extra silicon area (vs. graphics)
- Can run all layer types
- Low power

PowerVR Neural Network accelerator
- Dedicated hardware
- Fastest performance
- Smallest silicon area
- Lowest power and bandwidth
- Works with PowerVR GPU or 3rd Party IPs
PowerVR the Best for Business

PowerVR 2NX designed for mobile and Android

- PowerVR 2NX is the only IP solution in the market that can deliver against all the requirements for a deployable mobile solution
- Low area for PowerVR 2NX combined with the low area of the PowerVR 9XE GPU provides a GPU+NNA solution in the same footprint as a competing GPU alone

- Requirements met with PowerVR 2NX
  - Low power – full hardware ensures lowest power/inference
  - Low area – most efficient solution in terms of inferences/mm²
  - MMU – easy integration in SoCs supporting high level OSs
  - Android support - PowerVR has long history of Android support
Thank you