Framework for creating performance model of AI algorithms for early architecture exploration

Amit Dudeja, Sr R & D Engineer, Synopsys India Pvt. Ltd.
Amit Tara, Sr R & D Engineer, Synopsys India Pvt. Ltd.
Amit Garg, Principal Engineer, Synopsys India Pvt. Ltd.
Tushar Jain, R & D Engineer, Synopsys India Pvt. Ltd.
Agenda

• AI Market and Technology Trends
• AI SoC Design Challenges
• Shift Left Architecture Exploration and Optimization
• AI Exploration Framework
• NN driven Deep Learning Architecture Optimization
  – Case Study Resnet18 with NVDLA
AI Market and Technology Trends

New Neural Network algorithms
• Higher accuracy, lower size and less processing
• But: less data re-use, less cycles per byte

Neural Network Compiler optimizations
• Loop-tiling, -unrolling, and –parallelization
• Splitting and fusing of Neural Network layers
• Memory layout optimization across layers
• Optimized code generation to utilize available hardware accelerators

Deep Learning Accelerator optimizations
• Schedule workload on parallel hardware engines
• Optimize/reduce data transfers to & from memory
AI SoC Design Challenges

• Choosing the right algorithm and architecture: CPU, GPU, FPGA, vector DSP, ASIP
  – CNN graphs are evolving fast, need short time to market and cannot optimize for one single graph
  – Joint design of algorithm, compiler, and target architecture
  – Joint optimization of power, performance, accuracy, and cost

• Highly parallel compute drives memory requirements
  – High on-chip and chip to chip bandwidth at low latency
  – High memory bandwidth requirements for parameters and layer to layer communication

• Performance analysis requires realistic workloads to consider dynamic effects
  – Scheduling of AI operators on parallel processing elements
  – Unpredictable interconnect and memory access latencies

Shift Left Architecture Exploration and Optimization

Differentiation by Joint Algorithm and Architecture Optimization
Shift Left Architecture Exploration and Optimization
Objective

• Systematic creation of AI performance model in an automated manner.
  – Model AI operations like convolution, batchNorm etc
  – Generate complete topology of AI algorithm.

• Construction of generic and configurable AI Centric Hardware Subsystem
  – Model scalable AI engines for compute and memory load.
  – Build hierarchical subsystems for complex designs.
PA Ultra AI Exploration Framework

- AI Operator Library
  - Workload models for different operators (layers) for AI CNN algorithm development

- Automated generation of workload model for the Neural Networks

- AI centric HW architecture Engine
Once the Convolution Operator block gets triggered:
- Input frame and coefficient are read in parallel.
- Once the necessary input data is read, block process the convolutions and write backs the resulting feature maps.
AI Operator Library

- Library of Operators
- Design Canvas
- User Parameters
- Exploration Parameters
- S/W Optimization Parameters
PA Ultra AI Exploration Framework

• AI Operator Library
  – Workload models for different operators (layers)
    for AI CNN algorithm development

Automated generation of workload model for Neural Networks

• AI centric HW architecture Engine
Creation of performance model of AI algorithms
PA Ultra AI Exploration Framework

• AI Operator Library
  – Workload models for different operators (layers) for AI CNN algorithm development

• Automated generation of workload model for Neural Networks

AI centric HW architecture Engine
PA Ultra Hardware IP Models Library

**Traffic, Processors, RTL**
- Task-based and trace-based workload models
- Cycle accurate processor for ARM, ARC, Tensilica, CEVA
- RTL Co-simulation/emulation

**Interconnect Models**
Generic:
- SBL-TLM2-FT (AXI)
- SBL-GCCI (ACE, CHI)

IP Specific:
- Arteris FlexNoC & Ncore
- Arm AHB/APB
- Arm PL300
- Arm SBL-301
- Arm SBL-400
- Synopsys DW AXI

**Memory Subsystems**
- Generic multiport memory controller (GMPMC)
- DesignWare uMCTL2 memory controller
- DesignWare LPDDR5 memory controller
- Co-simulate with RTL
Creation of AI Centric Hardware Engine

- AI-centric hardware engine caters to the compute and memory requirements of AI operations.
- Can be constructed hierarchically by encapsulating n-sub engines underneath.
- Can be characterized with performance related attributes like operations-per-cycle, stochastic cache, branch prediction, pipeline depth etc.
Platform Architect Ultra

Capture Workload Model

Capture Architecture Model

Map Workload to Architecture Model

Analyze Power & Performance

© Accellera Systems Initiative
Case Study: Resnet-18 with NVDLA

Goals:
① 5 ms latency, ② minimize power, ③ minimize energy

Optimize Hardware configuration:
– SIMD width
– Burst size, outstanding transactions
– speed of DDR memory and of data path
ResNet-18 Workload model generated with AI-XP
NVDLA Initial Configuration

– Burst size: 16
– Max Outstanding transactions: 8
– Clock frequency of data path: 1GHz
– SIMD width: 16 operations per cycle
Performance limited by processing, use wider SIMD data path
Impact of SIMD Width on Performance

Resource Utilization of CONV Datapath (yellow), CONV DMA (red) and other components

- SIMD-16
- SIMD-32
- SIMD-64
- SIMD-128

- CONV PE load
- CONV DMA load
- memory bandwidth bound
- processing bound

Diminishing performance gains
Resnet 18 Example Sweep

Goal: 5ms latency, minimize power & energy

<table>
<thead>
<tr>
<th>Name</th>
<th>timing_us</th>
<th>...</th>
<th>sz_in_bytes</th>
<th>speed_bin</th>
<th>Clk_period</th>
<th>opc</th>
</tr>
</thead>
<tbody>
<tr>
<td>run_b02_opc84_c64s_DDR4_1866M_os4</td>
<td>4418.966</td>
<td>4</td>
<td>32 DDR4-1866</td>
<td>0.5</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>run_b02_opc84_c64s_DDR4_1866M_os8</td>
<td>4323.059</td>
<td>8</td>
<td>32 DDR4-1866</td>
<td>0.5</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_2400_os4</td>
<td>2560.240</td>
<td>4</td>
<td>16 DDR4-2400</td>
<td>1</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_2400_os8</td>
<td>5457.129</td>
<td>8</td>
<td>16 DDR4-2400</td>
<td>1</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_1866M_os4</td>
<td>6964.125</td>
<td>4</td>
<td>16 DDR4-1866</td>
<td>1</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_1866M_os8</td>
<td>6676.125</td>
<td>8</td>
<td>16 DDR4-1866</td>
<td>1</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_2400_os4</td>
<td>5509.487</td>
<td>4</td>
<td>16 DDR4-2400</td>
<td>0.75</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_2400_os8</td>
<td>5180.488</td>
<td>8</td>
<td>16 DDR4-2400</td>
<td>0.75</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_1866M_os4</td>
<td>6495.257</td>
<td>4</td>
<td>16 DDR4-1866</td>
<td>0.75</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_1866M_os8</td>
<td>6207.377</td>
<td>8</td>
<td>16 DDR4-1866</td>
<td>0.75</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_2400_os4</td>
<td>5141.269</td>
<td>4</td>
<td>16 DDR4-2400</td>
<td>0.5</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_2400_os8</td>
<td>4707.354</td>
<td>8</td>
<td>16 DDR4-2400</td>
<td>0.5</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_1866M_os4</td>
<td>6402.813</td>
<td>4</td>
<td>16 DDR4-1866</td>
<td>0.5</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b16_opc128_ziklo_DDR4_1866M_os8</td>
<td>6008.660</td>
<td>8</td>
<td>16 DDR4-1866</td>
<td>0.5</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_2400_os4</td>
<td>4226.884</td>
<td>4</td>
<td>32 DDR4-2400</td>
<td>1</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_2400_os8</td>
<td>4086.654</td>
<td>8</td>
<td>32 DDR4-2400</td>
<td>1</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_1866M_os4</td>
<td>4444.511</td>
<td>4</td>
<td>32 DDR4-1866</td>
<td>1</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_1866M_os8</td>
<td>4236.462</td>
<td>8</td>
<td>32 DDR4-1866</td>
<td>1</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_2400_os4</td>
<td>2344.214</td>
<td>4</td>
<td>32 DDR4-2400</td>
<td>0.75</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_2400_os8</td>
<td>3261.565</td>
<td>8</td>
<td>32 DDR4-2400</td>
<td>0.75</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_1866M_os4</td>
<td>3663.721</td>
<td>4</td>
<td>32 DDR4-1866</td>
<td>0.75</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_1866M_os8</td>
<td>3769.144</td>
<td>8</td>
<td>32 DDR4-1866</td>
<td>0.75</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_2400_os4</td>
<td>2981.385</td>
<td>4</td>
<td>32 DDR4-2400</td>
<td>0.5</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_2400_os8</td>
<td>2305.655</td>
<td>8</td>
<td>32 DDR4-2400</td>
<td>0.5</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_1866M_os4</td>
<td>3483.364</td>
<td>4</td>
<td>32 DDR4-1866</td>
<td>0.5</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>run_b32_opc128_ziklo_DDR4_1866M_os8</td>
<td>3391.767</td>
<td>8</td>
<td>32 DDR4-1866</td>
<td>0.5</td>
<td>128</td>
<td></td>
</tr>
</tbody>
</table>

Sweep parameters
- Burst size: 16, 32
- Outstanding transactions: 4, 8
- DDR memory speed: DDR4-1866, DDR4-2400
- Clock frequency of data path: 1, 1.33, 2GHz
- SIMD width: 64, 128 operations per cycle
Sweep Over Hardware Parameters, Latency in µs

<table>
<thead>
<tr>
<th>DDR4 Speed</th>
<th>Burst Size</th>
<th>Outstanding Transactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 GHz</td>
<td>64</td>
<td>5000</td>
</tr>
<tr>
<td>2 GHz</td>
<td>128</td>
<td>5000</td>
</tr>
<tr>
<td>1.33 GHz</td>
<td>16</td>
<td>6000</td>
</tr>
<tr>
<td>2 GHz</td>
<td>32</td>
<td>6000</td>
</tr>
</tbody>
</table>

© Accellera Systems Initiative
Power/Performance/Energy Trade-off Analysis

Optimal solution

Datapath GHz

DDR

SIMD

Burst size

outstanding

© Accellera Systems Initiative
Example: Resnet-18 with NVDLA

**Goal:**
- 5 ms latency, minimize energy

**Optimize Hardware configuration:**
- SIMD width: 128 operations per cycle
- Burst size: 32 bytes
- outstanding transactions: 4
- speed of DDR memory: DDR4-1866
- speed of data path: 1GHz
Summary

• Demonstrated a framework to perform exploration studies in the field of AI.
• Ease of automatically creating a workload model out of an AI algorithm.
• Goal directed power and performance study to achieve an inference latency of 5ms with optimized power consumption.
Questions