Snowflake deep neural network accelerator
Performance

SnowFlake is a co-processor IP designed for SoC and verified on FPGA

Advantages:

> 95% efficiency (theoretical/actual ops/s)

100x better than GPUs (performance/watt)

low memory bandwidth
new architecture
Accelerator Hardware

- **Functional units**
  - Multiply-accumulate (MAC)
  - Comparators (maxpool)

- **On-chip memory**
  - Cache / scratchpad for maps and weights

- **Configuration logic**
  - Instruct on-chip memory to stream to MACs
  - Instruct MACs to write-back results
Traces

Require N instructions
1 instruction per cycle

Kernel Trace
Map Trace

Output activation

MAC R0, R1, R2
kop mop

MAC R0, R1, R2, imm
Kernel trace start address
map trace start address

Require 1 instruction
1 instruction per trace
Require start address, length
Data Organization

Kernel Trace
- Subdivided into M vectors

Map Trace
- Subdivided into M vectors

KW * N

Vec 1  Vec 2  ...  Vec M

KW

N

KH
Data Organization

Output activation

Kernel Trace 1

Kernel Trace kH

Map Trace 1

Map Trace kH

Mult-Acc
Vector Multiply-Accumulate (vMAC)
Scaling Up with Compute Units

Compute unit
- 4 vMACs per CU
- 16 MACs per vMAC
- 1 KB weights buffer per MAC
- 64 KB maps (double) buffer per CU
Maxpooling

- ~1% of network computation
- Hide latency of MAX behind MAC
Scaling Further with Clusters

Memory Interface

Data Distribution Network

Compute Cluster

Control Core
Compute Unit
Compute Unit
Compute Unit

Data Distribution Network

Data Distribution Network

Data Distribution Network

Compute Unit
Compute Unit
Compute Unit
...

Control Core
Compute Unit
Compute Unit
...

Reg. File

Instrs. from memory

Fetch
Decode
Dispatch

ALU

To compute units
Comparison of Perf. and B/W

AlexNet

Layer-wise comparison of performance and bandwidth

GoogLeNet

Layer-wise comparison of performance and bandwidth

ResNet-50

Layer-wise comparison of performance and bandwidth
Demo: object categorization
Demo: face identification
Thank you