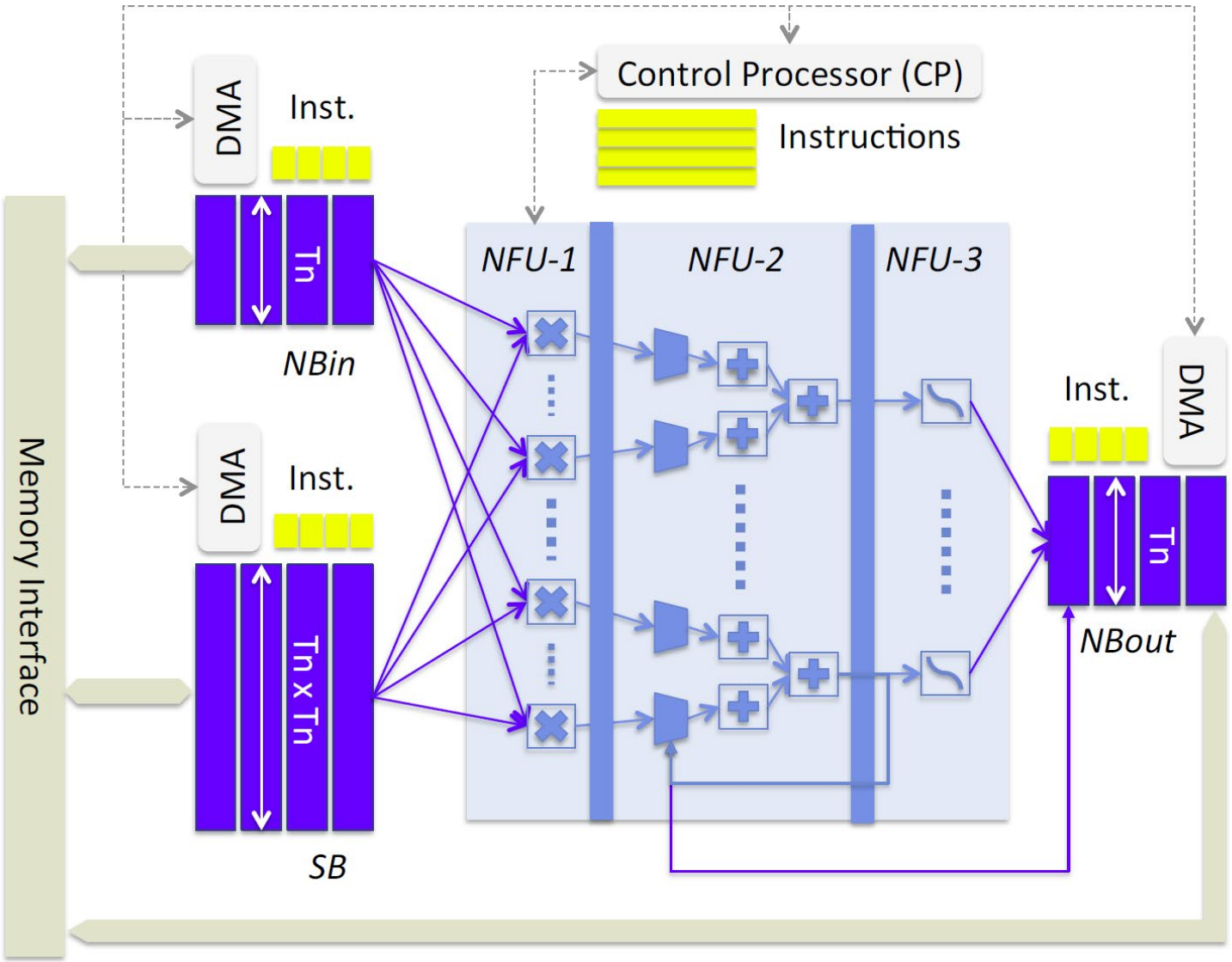
Lecture: DianNao and DaDianNao

- Topics: Diannao wrap-up and DaDianNao

DianNao


Control Processor (CP)
Instructions
DMA
Inst.
Tn
NBin
DMA
Inst.
Tn x Tn
SB
NFU-1
NFU-2
NFU-3
Inst.
DMA
Tn
NBout
Memory Interface

Fixed Point Arithmetic

- DianNao uses 16b fixed-point arithmetic; much more efficient than 32b floating-point arithmetic, and little impact on accuracy

Type	Area (μm^2)	Power (μW)
16-bit truncated fixed-point multiplier	1309.32	576.90
32-bit floating-point multiplier	7997.76	4229.60

Table 2. *Characteristics of multipliers.*

Fixed Point Error Rates

Type	Error Rate
32-bit floating-point	0.0311
16-bit fixed-point	0.0337

Table 1. *32-bit floating-point vs. 16-bit fixed-point accuracy for MNIST (metric: error rate).*

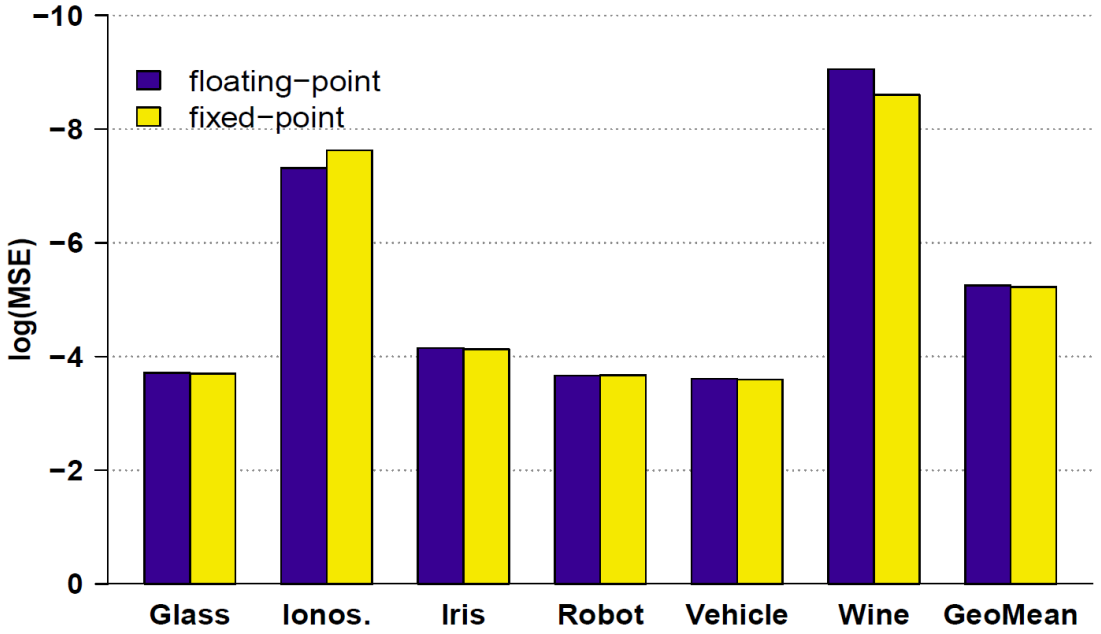
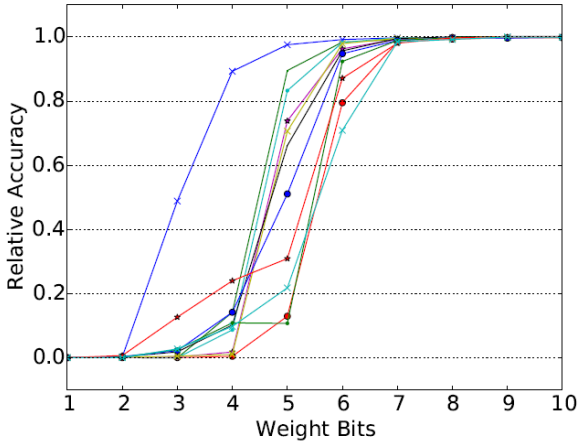


Figure 12. *32-bit floating-point vs. 16-bit fixed-point accuracy for UCI data sets (metric: log(*Mean Squared Error*)).*

Precision Analysis for DNNs

Network	Data (Per Layer)	Weights (Uniform)
LeNet[12]	2,4,3,3	7
Convnet[13]	8,7,7,5,5	9
AlexNet[14]	10,8,8,8,8,8,6,4	10
NiN[15]	10,10,9,12,12,11,11,11,10,10,10,9	10
GoogLeNet[4]	14,10,12,12,12,12,11,11,11,10,9	9

TABLE I: Minimum precision, in bits, for data and weights for a set of neural networks.



(m) GoogLeNet: Weights

Min bits required per layer for accuracy within 1%
Source: Proteus, Judd et al., WAPCO'16

Implementation Details/Results

- Cycle time of 1.02ns, area of 3mm^2, 485mW power
- The NFU is composed of 8 pipeline stages
- Peak activity is nearly 500 GOP/s
- 44KB of RAM capacity
- Buffers are about 60% of area/power, while NFU is ~30%
- Energy is 21x better than a SIMD baseline; this is limited because of the high cost of memory accesses
- Big performance boosts as well: higher computational density, tiling, prefetching

DianNao Conclusions

- Tiling to reduce memory traffic
- Efficient NFU and buffers to reduce energy/op and prefetch
- Even with these innovations, memory is the bottleneck, especially in a small accelerator with few pins
- For example, each classifier layer step needs 256 new synaptic weights (512 bytes), while 4 memory channels can only bring in 64 bytes per cycle
- Energy consumed by the DianNao pipeline per cycle = 500pJ
- Energy per cycle for fetching 64 bytes from memory = 35nJ

  (70 pJ/b for DDR3 at 100% utilization, Malladi et al., ISCA'12)

It's all about the memory bandwidth/energy !!!

The GPU Option

- Modest (but adequate) memory capacities (a few giga-bytes)
- High memory bandwidth, e.g., 208 GB/s (NVIDIA K20M)
- But, high compute-to-cache ratio on the chip
- Therefore, average GPU power of ~75 W, plus expensive memory accesses ➔ GPU card TDP of ~225W
- A GPU out-performs DianNao by ~2X

DaDianNao Philosophy

- Need giga-bytes of storage for weights and accessing these weights is the clear bottleneck
- Can't store giga-bytes on 1 chip, but can store giga-bytes on many chips on a board
- Surround a DianNao circuit with a large eDRAM (dense) structure to replace main memory (high storage-to-compute)
- Every operation is spread across several "tiles" to maximize parallelism

DaDianNao Philosophy II

- Synapses stay in place and neuron values move around (since synapses are so much larger)
- Use eDRAM instead of SRAM (about 2.85x higher density)
- Can get high internal bandwidth by having many banks for eDRAM storage (they use 4)
- Implement many tiles on a chip – each tile has 4 eDRAM banks for weights, and all tiles share 2 eDRAM banks for input/output
- They allow 32-bit operations because it is useful for training

DaDianNao Layouts

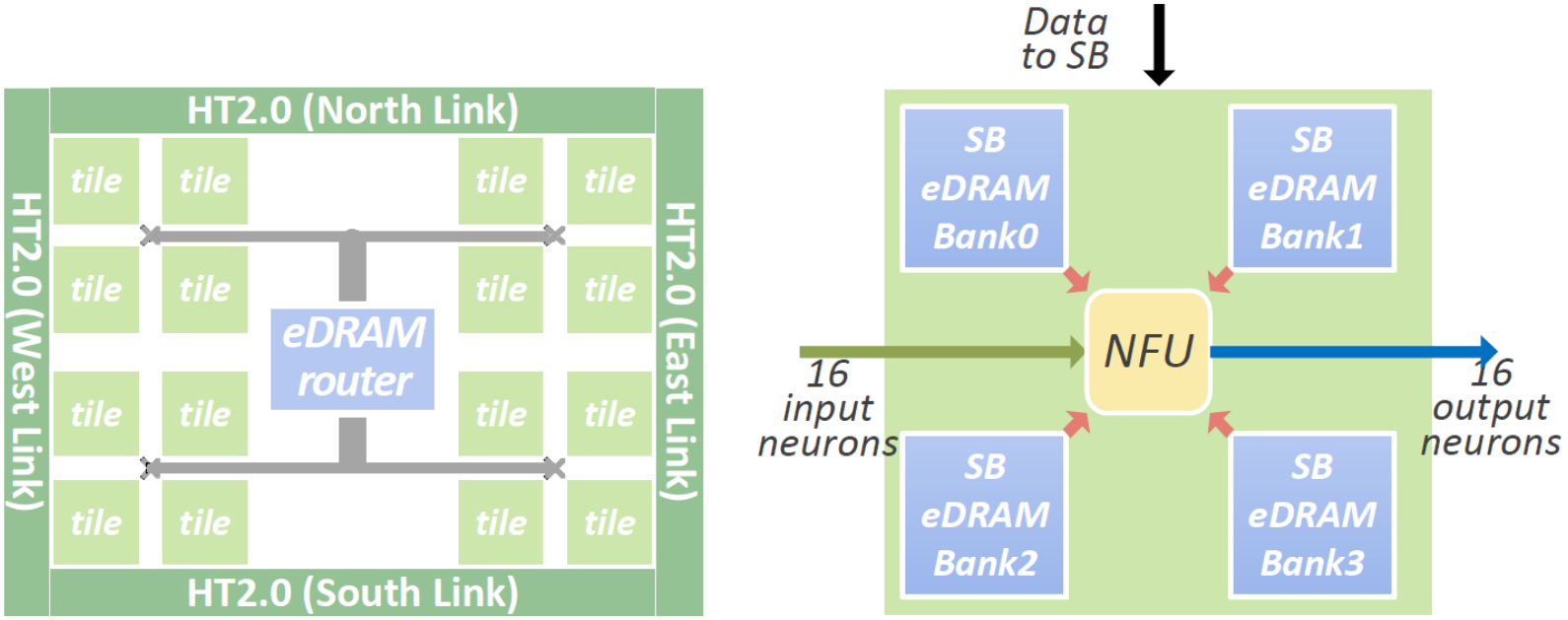


Figure 5: *Tile-based organization of a node (left) and tile architecture (right). A node contains 16 tiles, two central eDRAM banks and fat tree interconnect; a tile has an NFU, four eDRAM banks and input/output interfaces to/from the central eDRAM banks.*

Each eDRAM bank size is 512 KB (3 cyc); central eDRAM bank is 2MB (10 cyc); total node storage is 36 MB; HT bw is 6.4 x 4 GB/s (80ns).

Other Details

- 606 MHz clock (because of the eDRAM)
- 5.58 Tera ops/second
- 16 W node (chip), 68 mm^2
- Area breakdown: 45% tiles; 26% HT; 12% central eDRAM; 9% central wiring
- Half the chip is eDRAM storage
- Power breakdown: 39% in tiles, 50% HT; eDRAM power is 38%; combinational circuits are 38%, 19% is registers

Speedup over GPU

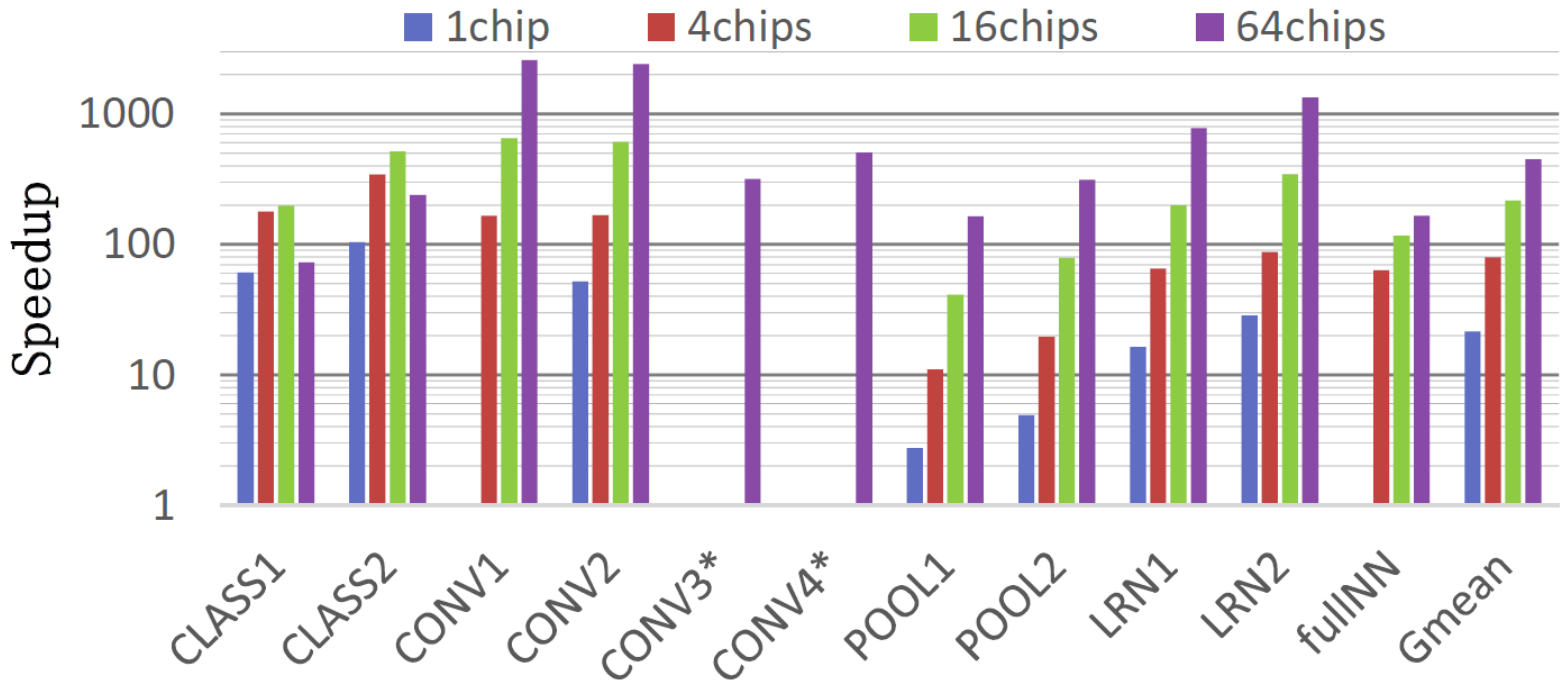


Figure 10: *Speedup w.r.t. the GPU baseline (inference). Note that CONV1 and the full NN need a 4-node system, while CONV3* and CONV4* even need a 36-node system.*

Energy Reductions

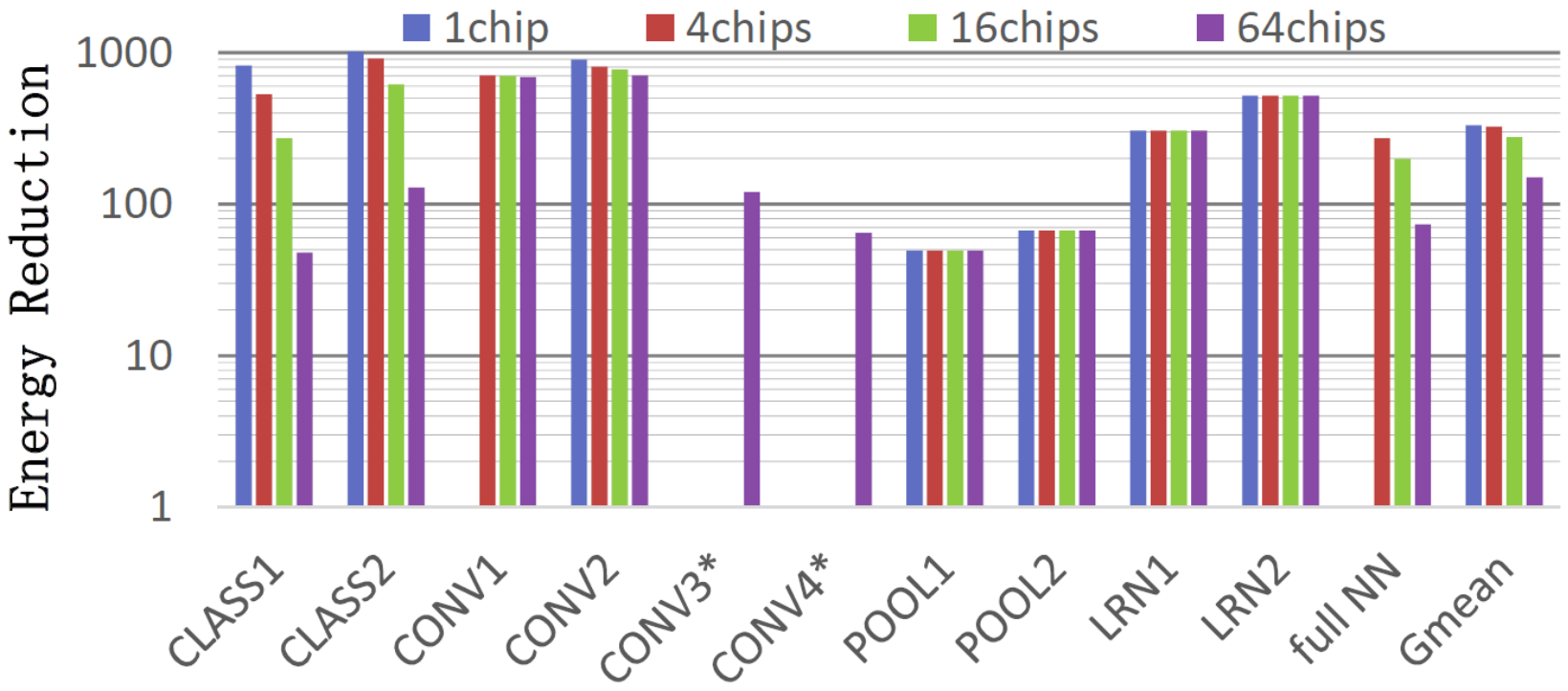


Figure 13: *Energy reduction w.r.t. the GPU baseline (inference).*

Benchmarks

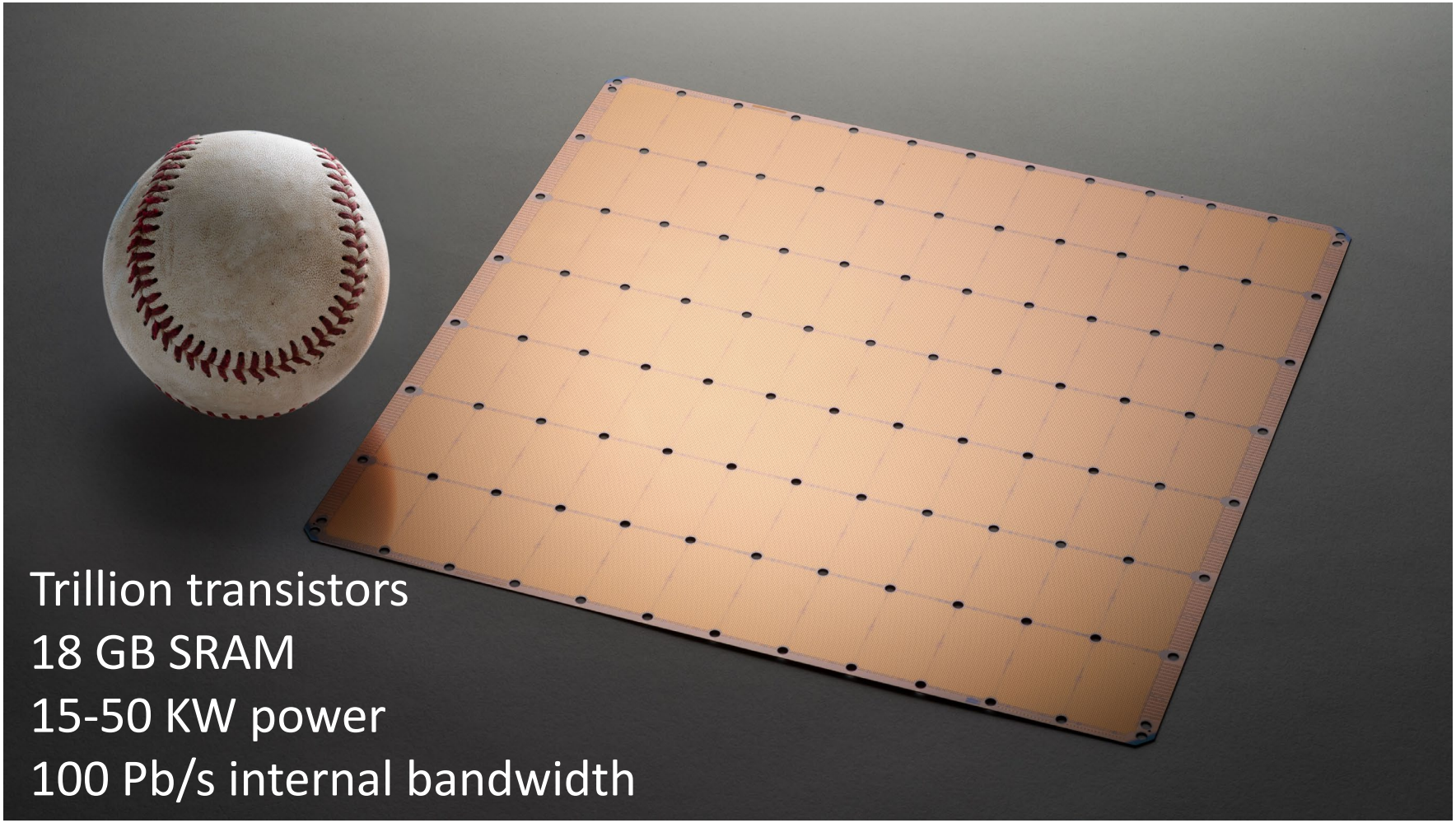
Layer	N_x	N_y	K_x	K_y	N_i or N_{if}	N_o or N_{of}	Synapses	Description
CLASS1	-	-	-	-	2560	2560	12.5MB	Object recognition and speech recognition tasks (DNN) [11].
CLASS2	-	-	-	-	4096	4096	32MB	Multi-Object recognition in natural images (DNN), winner 2012 ImageNet competition [32].
CONV1	256	256	11	11	256	384	22.69MB	
POOL2	256	256	2	2	256	256	-	
LRN1	55	55	-	-	96	96	-	
LRN2	27	27	-	-	256	256	-	
CONV2	500	375	9	9	32	48	0.24MB	Street scene parsing (CNN) (e.g., identifying building, vehicle, etc) [18].
POOL1	492	367	2	2	12	12	-	
CONV3*	200	200	18	18	8	8	1.29GB	Face Detection in YouTube videos (DNN), (Google) [34].
CONV4*	200	200	20	20	3	18	1.32GB	YouTube video object recognition, largest NN to date [8].

Table I: *Some of the largest known CNN or DNN layers (CONVx* indicates convolutional layers with private kernels).*

DaDianNao Summary

- Memory bandwidth is the key bottleneck, especially when handling the fully-connected classifier layers
- DaDianNao manages this by distributing weights across eDRAM banks in many chips
- A layer is executed in parallel across several NFUs/chips; outputs are moved around; the next layer then executes (no memory accesses for weights)

Cerebras Wafer Scale Integration



Source: Cerebras.net

References

- "DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine Learning", T. Chen et al., Proceedings of ASPLOS, 2014
- "DaDianNao: A Machine-Learning Supercomputer", Y. Chen et al., Proceedings of MICRO, 2014
- https://www.cerebras.net/wp-content/uploads/2019/08/Cerebras-Wafer-Scale-Engine-Whitepaper.pdf
- https://www.sigarch.org/the-first-trillion-transistor-chip-a-new-design-space