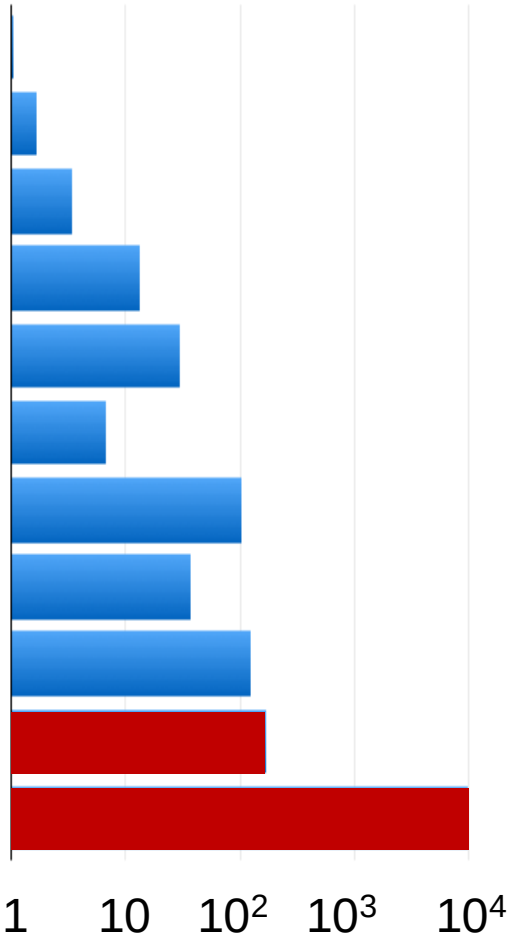
Data Movement Dominates Energy Consumption

Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Multiply	0.2
32b Multiply	3.1
16b FP Multiply	1.1
32b FP Multiply	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

[**Horowitz**, *ISSCC* 2014]



Data movement more energy than computation

→ *Reduce **amount** of data movement*
- exploit data reuse
- compute in memory

Farther and larger memories use more energy

→ *Reduce **energy** of data movement*
- chiplets and 3D
- optical and superconductors

Challenges include manufacturing cost/yield, robustness, scaling, and overhead energy

Co-Design Models and Hardware

Reduce both the ***energy*** per compute and the ***amount*** of compute

Reduce Precision
(bits per weight and inputs)

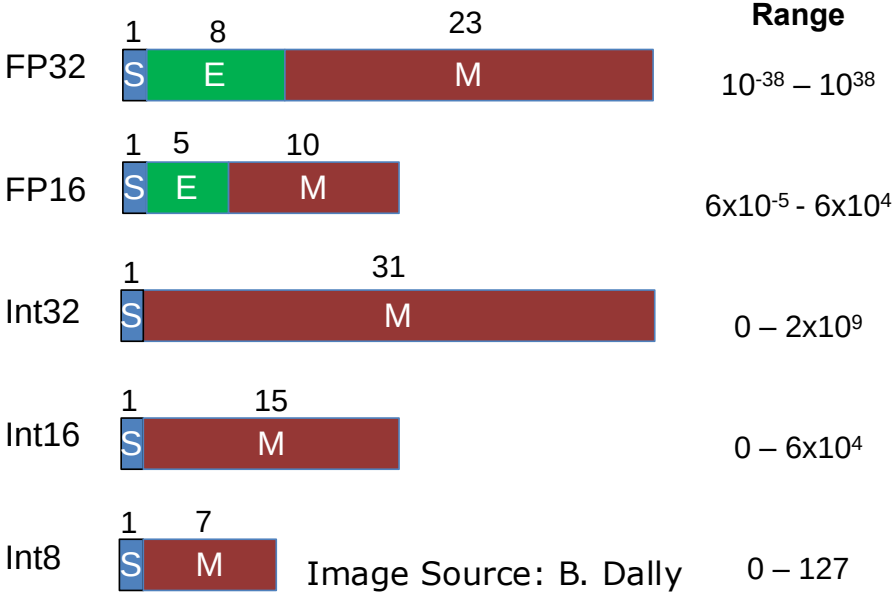
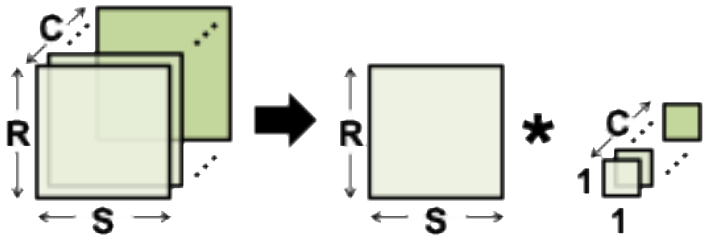


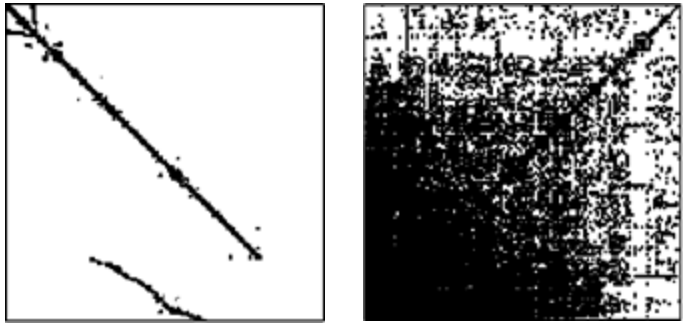
Image Source: B. Dally

Efficient Models
(number of weights and inputs)



Example: MobileNet

Sparsity
(% of zeros in weights and inputs)



Zero Values Nonzero Values

Note: Savings often require hardware support

Challenges include accuracy and overhead energy

Hardware Flexibility vs Efficiency

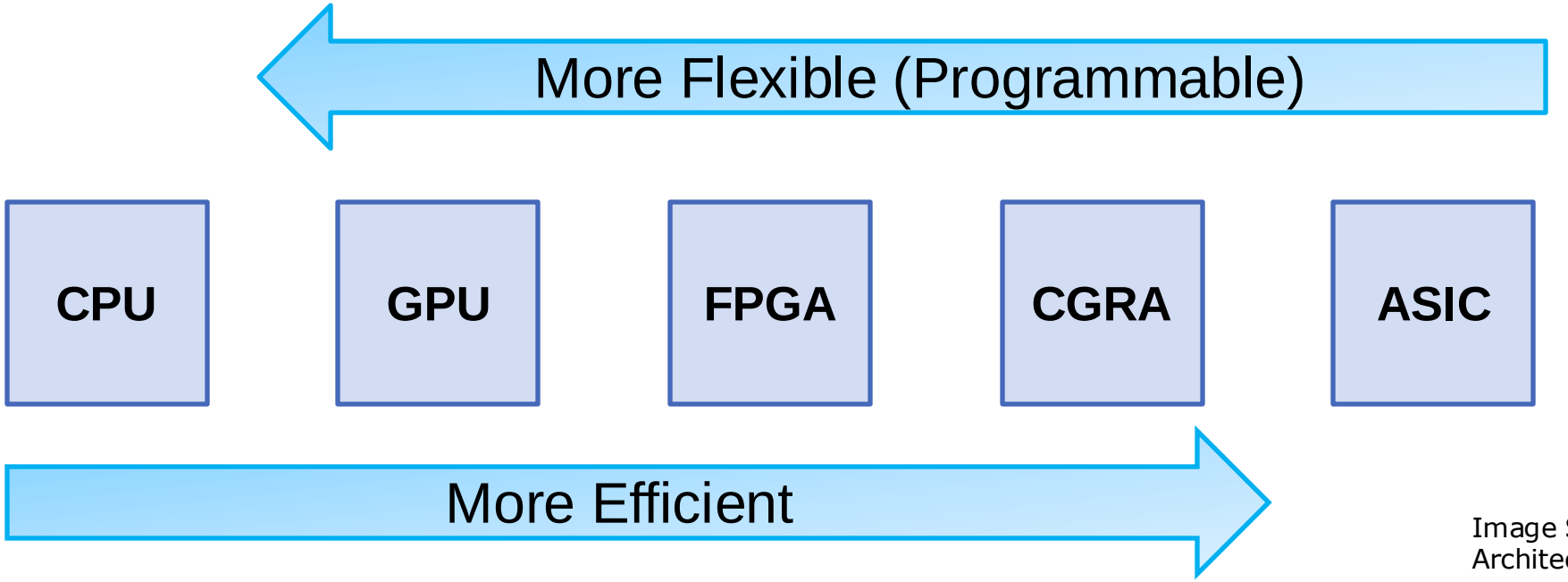


Image Source: 6.5930 Hardware Architectures For Deep Learning

FPGA: Field programmable gate array
CGRA: Coarse-grained reconfigurable array
ASIC: Application-specific integrated circuit

Efficiency = Throughput / Power Consumption

Many forms of specialization: configurability/programmability (instructions, granularity), allocation of resources (memory/compute), dataflow/mapping, etc.

Challenges include how much to specialize, utilization, future proof, and limitations on innovation