ETH zürich

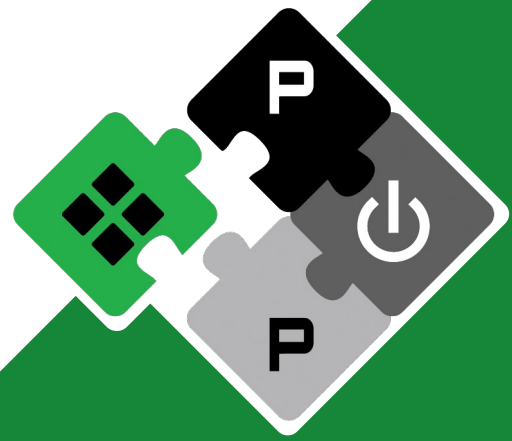
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ICECS 2023
2023 30th IEEE International Conference on Electronics, Circuits and Systems (ICECS)
4-7 December 2023 – Hilton Maslak İstanbul / Türkiye
TECHNOSAPIENS FOR SAVING HUMANITY

MinPool:
A 16-core NUMA-L1 Memory RISC-V Processor Cluster for Always-on Image Processing in 65nm CMOS

Integrated Systems Laboratory (ETH Zürich)

Samuel Riedel, Matheus Cavalcante,
Manos Frouzakis, Domenic Wüthrich,
Enis Mustafa, Arlind Billa, Luca Benini



PULP Platform
Open Source Hardware, the way it should be!

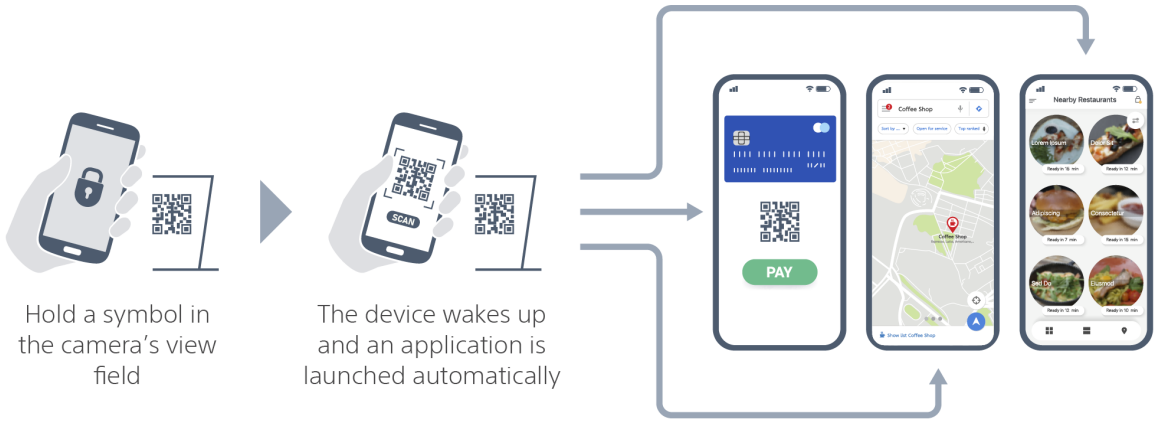
@pulp_platform
pulp-platform.org
youtube.com/pulp_platform

Always-on Image Processing

The processor is idle while the device is in the sleep mode, until a human presence is detected

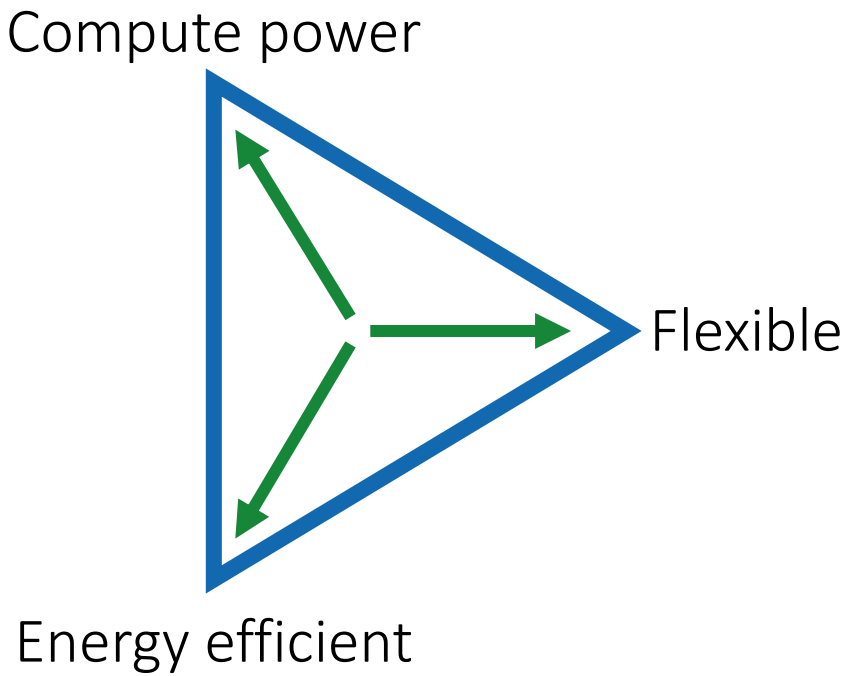
The screen stays on while the user's face is directed toward the screen

The screen goes to stand-by when no person/face is detected



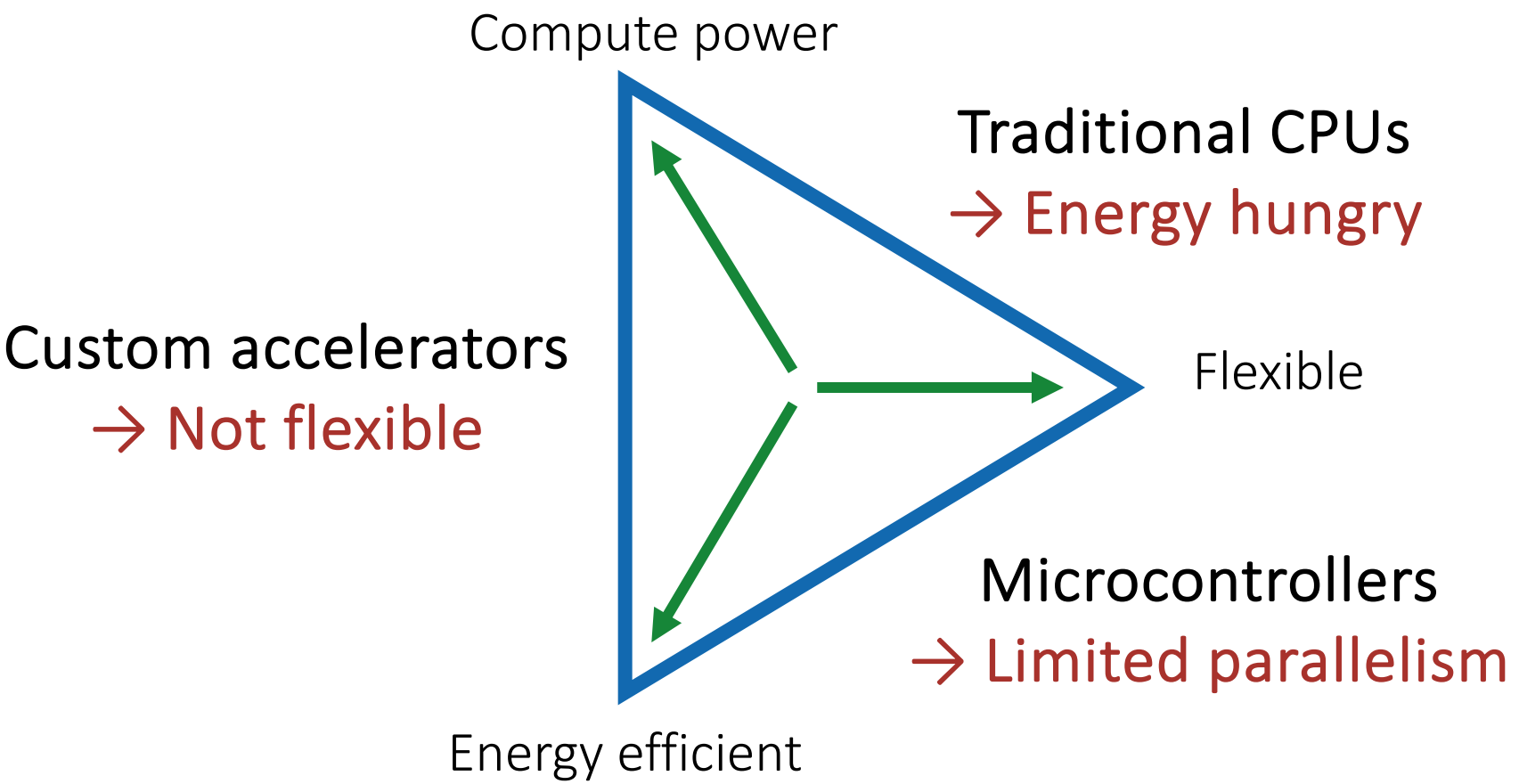
Hold a symbol in the camera's view field

The device wakes up and an application is launched automatically



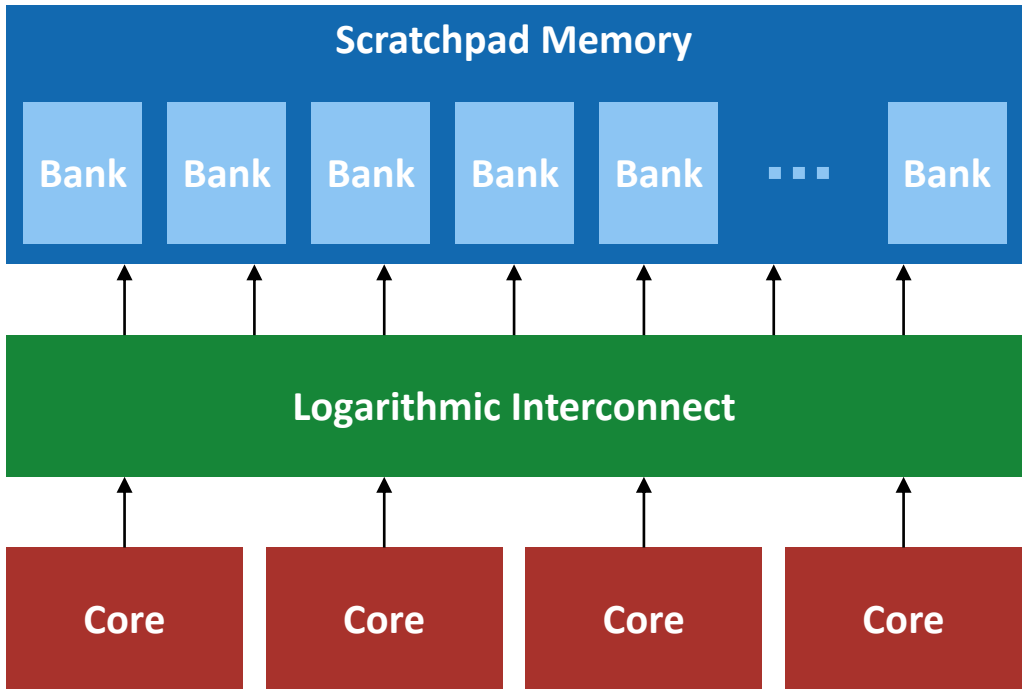
Source: https://www.sony-semicon.com/en/technology/mobile/always-on.html

Common Architectures



The Compute Cluster

- Widely used building block
  - GPU, AI, IoT
- Low latency access SPM
  - Multi-banked architecture
  - Fast logarithmic interconnect
- Efficient cores
  - Hide SPM “residual” latency
  - Expressive, domain-specific ISA extensions
- Powerful, efficient, flexible

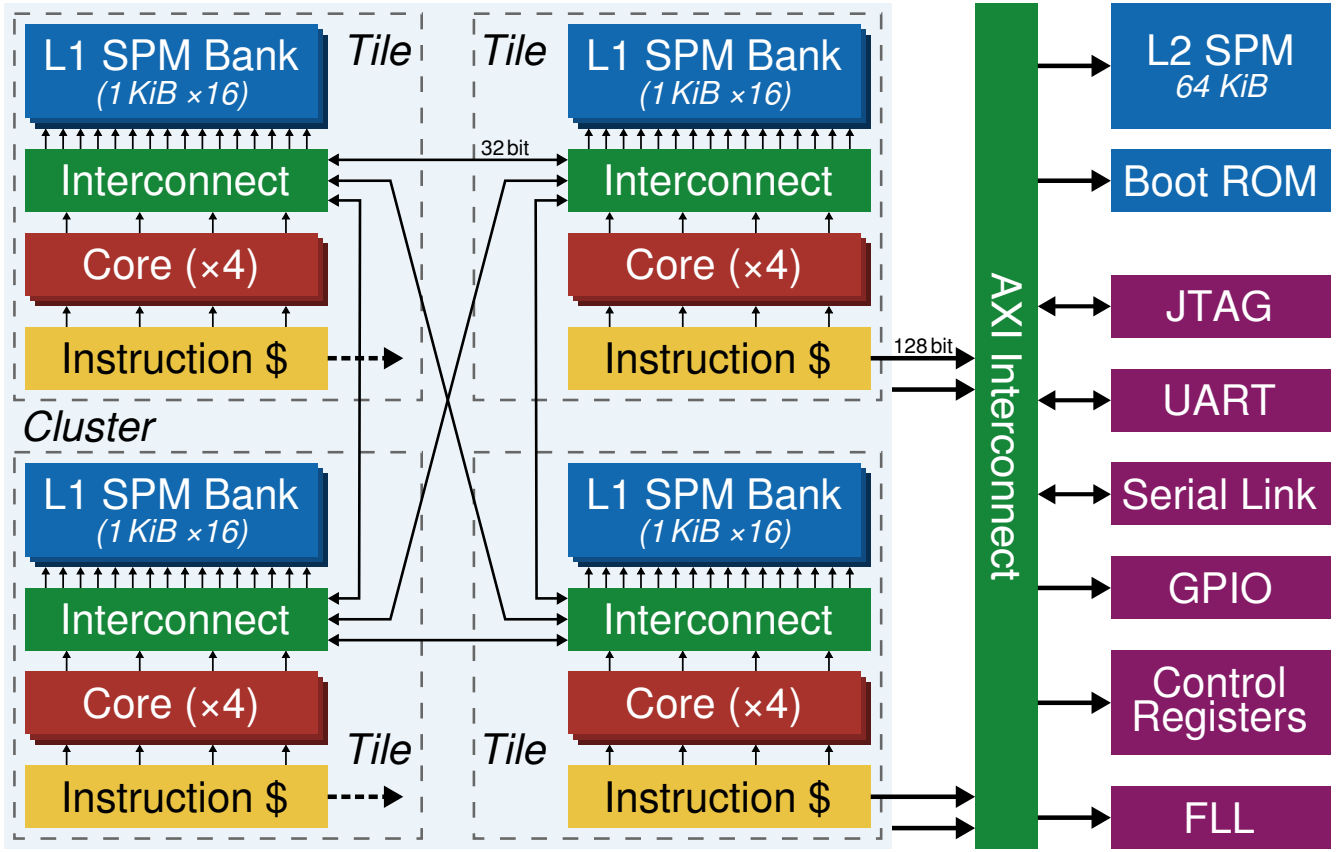


The Chip: MinPool

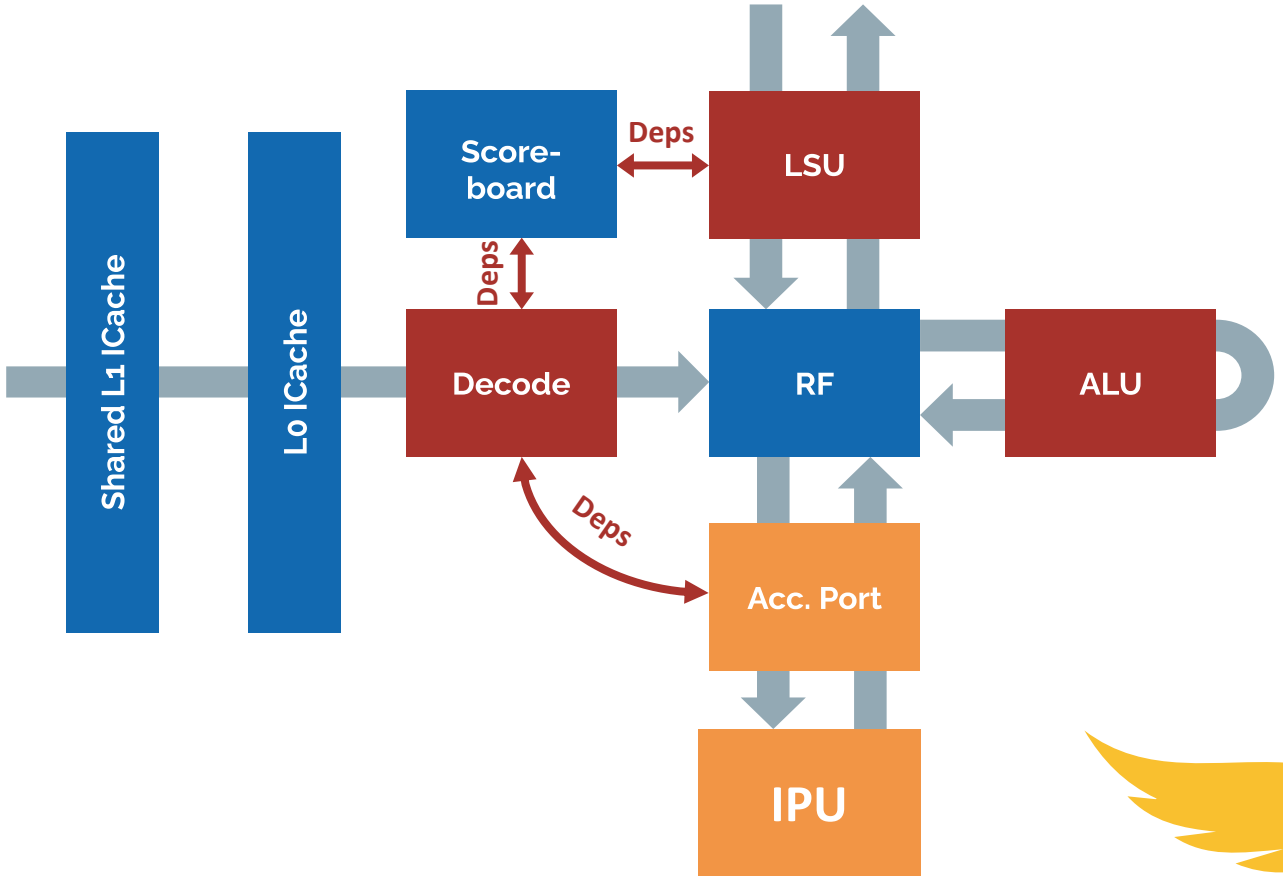
- 16 32-bit RISC-V cores
  - Custom DSP ISA extension
  - → Compute power
  - Latency tolerant
- 64 KiB shared L1 memory
  - NUMA
  - Accessible in less than 3 cycles
  - → Flexible
- TSMC's 65nm process
  - Low-leakage
  - → Efficient

The Architecture

- Based on MemPool
  - Open source
    https://github.com/pulp-platform/mempool
- Hierarchical
- Tiles
  - Compute cluster
  - 4 cores and 16 memory banks
- Cluster
  - 4 tiles
- Peripherals to interface camera



Hiding Latencies with Snitch



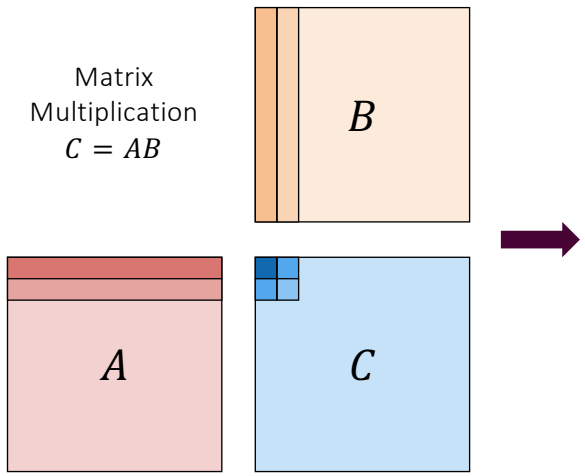
- Latency-tolerant → Scoreboard
  - Extended LSU for out-of-order retiring
- Extensible → IPU
  - Supports the *Xpulpimg* extension

The Specifications

- Architecture
  - 16 RISC-V cores with IPUs
  - 64KiB of L1 memory
  - 64KiB of L2 memory
  - 8KiB of Instruction cache
- TSCM 65 nm
  - 9 metal tracks
  - 5.8 mm^2
  - 0.8-1.5V
  - Up to 400 MHz

MinPool's Compilers

- Support both LLVM and GCC
  - Custom extensions fully supported
- Instruction scheduling
  - Hide load latency
  - Hide latency of accelerator




```
lw    a1,0(t2)
lw    t0,0(s0)
lw    a3,0(s1)
lw    a5,4(s1)
lw    t3,4(t2)
lw    a2,4(s0)
lw    a0,0(s2)
lw    a7,4(s2)
mul   t6,a1,a3
mul   a1,a1,a5
mul   a3,t0,a3
mul   a5,t0,a5
add   t4,t6,t4
add   t5,a1,t5
mul   t6,t3,a0
add   t1,a3,t1
mul   a0,a2,a0
mul   t3,t3,a7
add   a6,a5,a6
mul   a2,a2,a7
addi  s3,s3,2
add   t4,t6,t4
add   t5,t3,t5
add   t1,a0,t1
add   a6,a2,a6
...
```

Programming Models

- **Bare-metal C runtime**
  - All cores execute the same code (SPMD)
  - Option to go to Assembly

  \+ Full control
  \+ High performance
  – High effort

- **OpenMP**
  - Single-threaded master core
  - Fork-join behavior

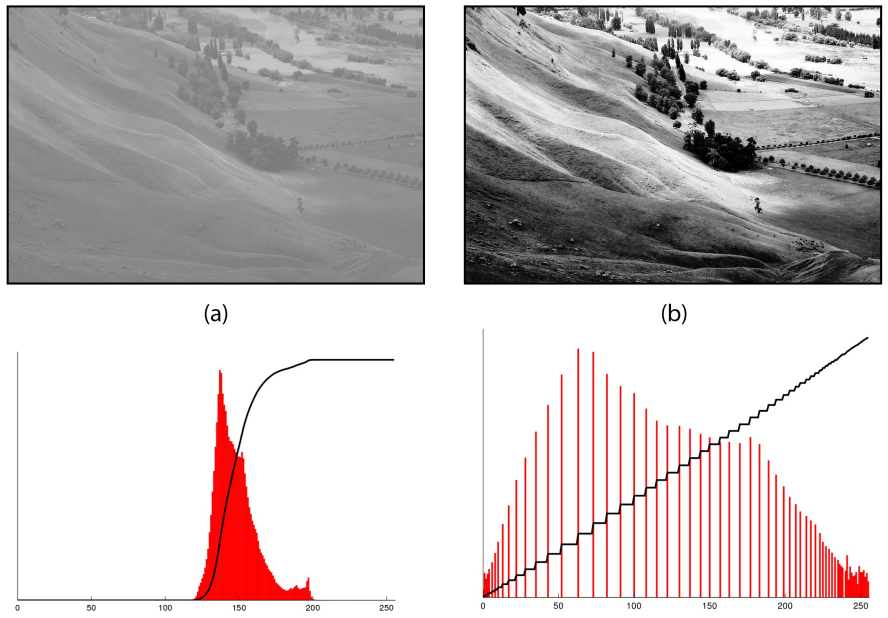
  \+ Medium effort
  \+ Well-known framework
  – Runtime overhead
  – Less room for optimization

- **Halide**
  - DSL extending C++
  - Functional traits
    - Decouple function & execution

  \+ Low effort
  \+ Can optimize across pipeline stages
  – Hard to control
  – Runtime overhead
  – More challenging to support than OpenMP

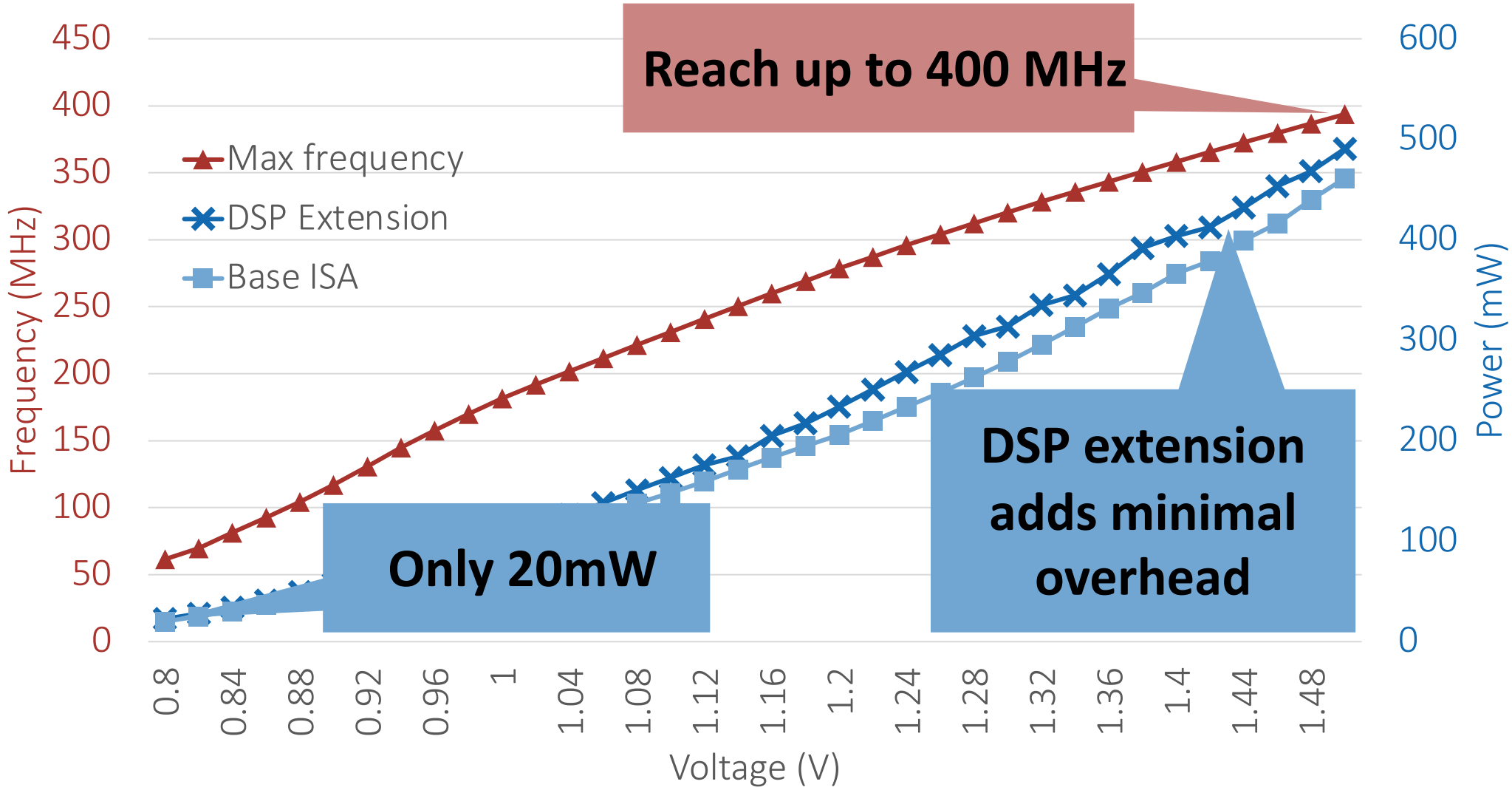
MinPool's Kernels

- DSP Kernels
  - *matmul*: Matrix Matrix multiplication → Image transformation
  - *2dconv*: 2D Convolution → Image filter
  - *dct*: Discrete Cosine Transformation → JPEG encoding
  - *axpy*: αx+y → Image transformation
  - *dotp*: Dot product → Ray-tracing
- Image processing pipelines
  - Histogram equalization
  - Bayer pattern demosaicing

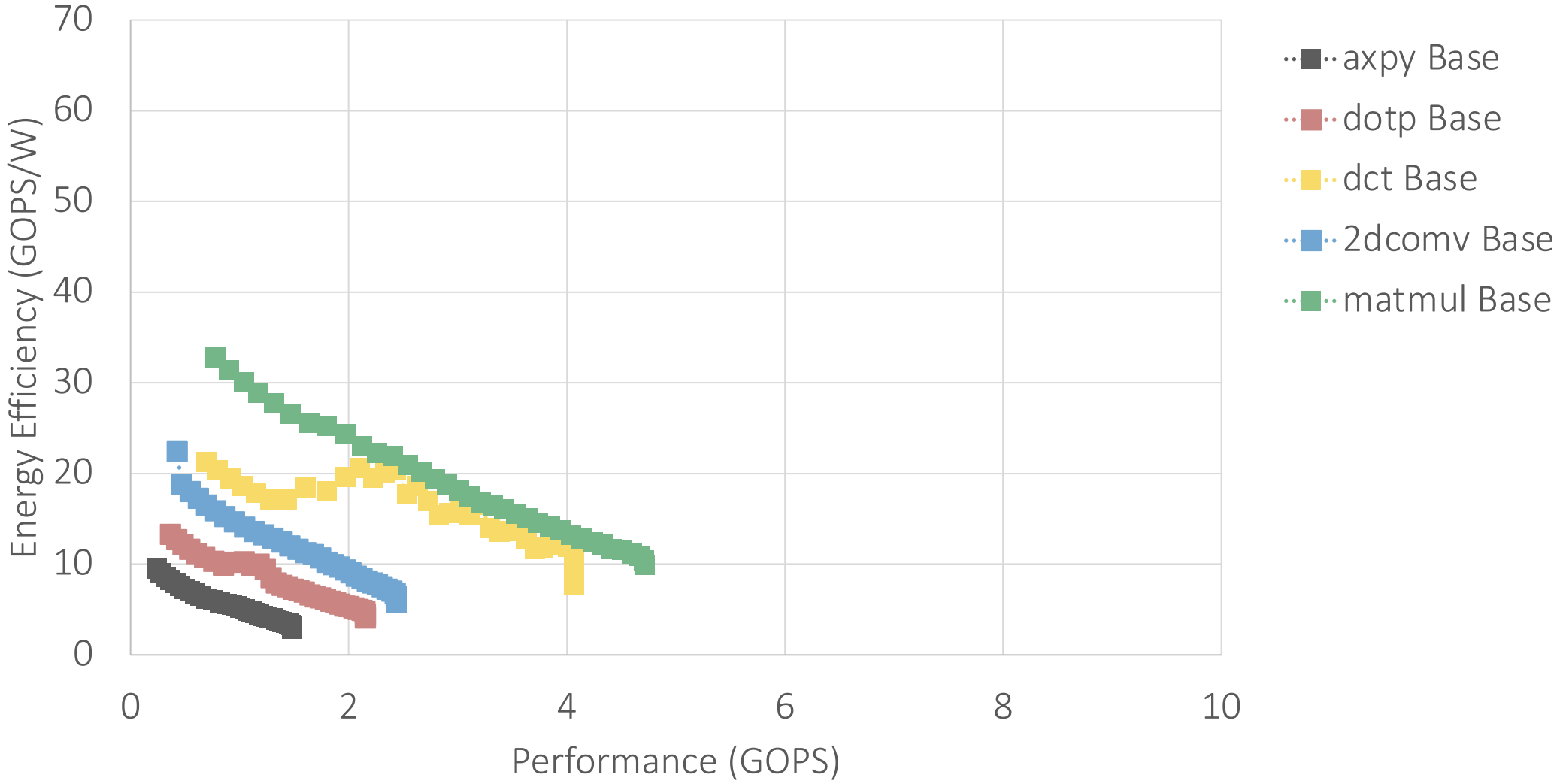


Source: https://commons.wikimedia.org/wiki/File:Histogram_equalization.png

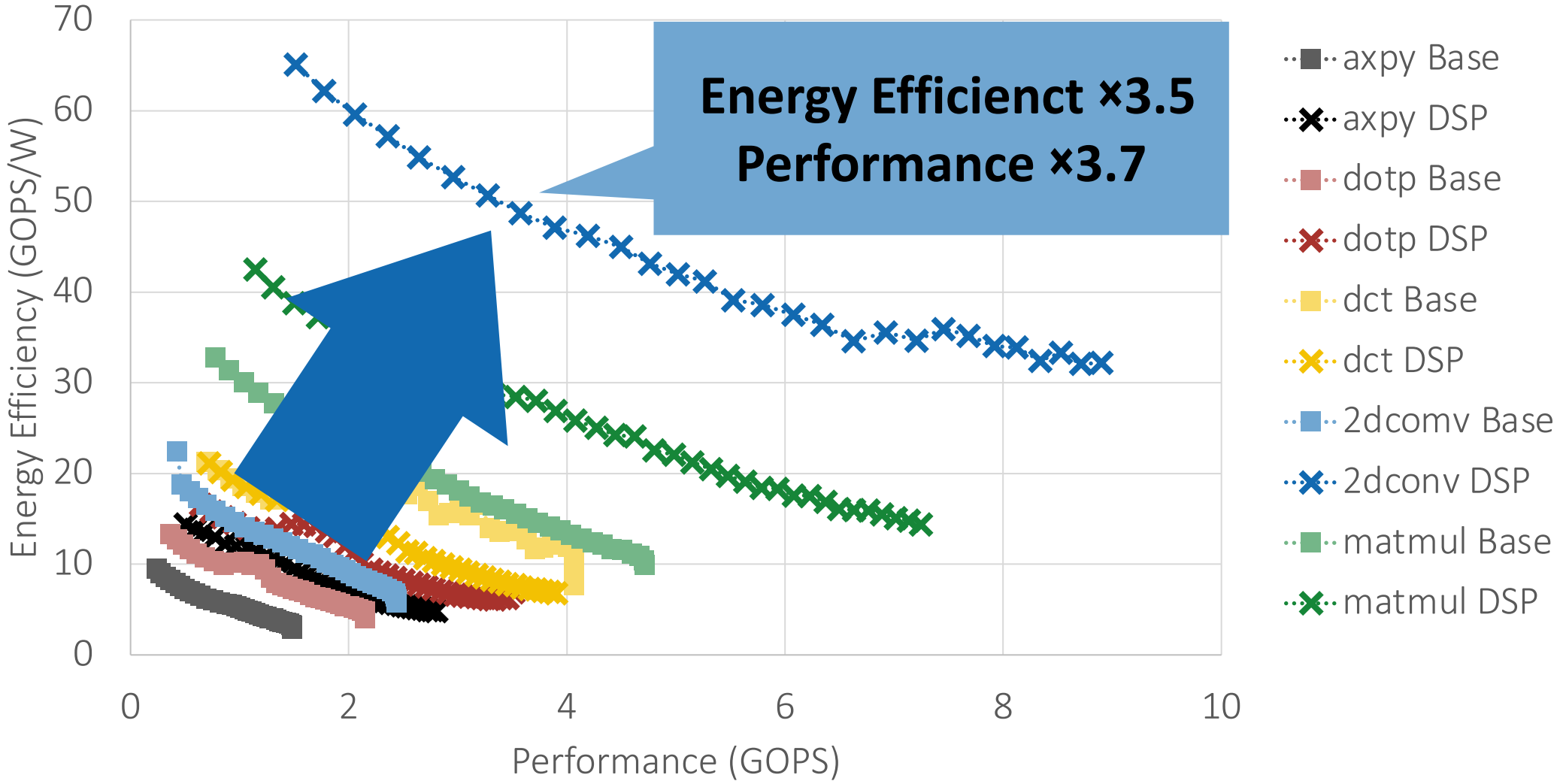
Power and Frequency



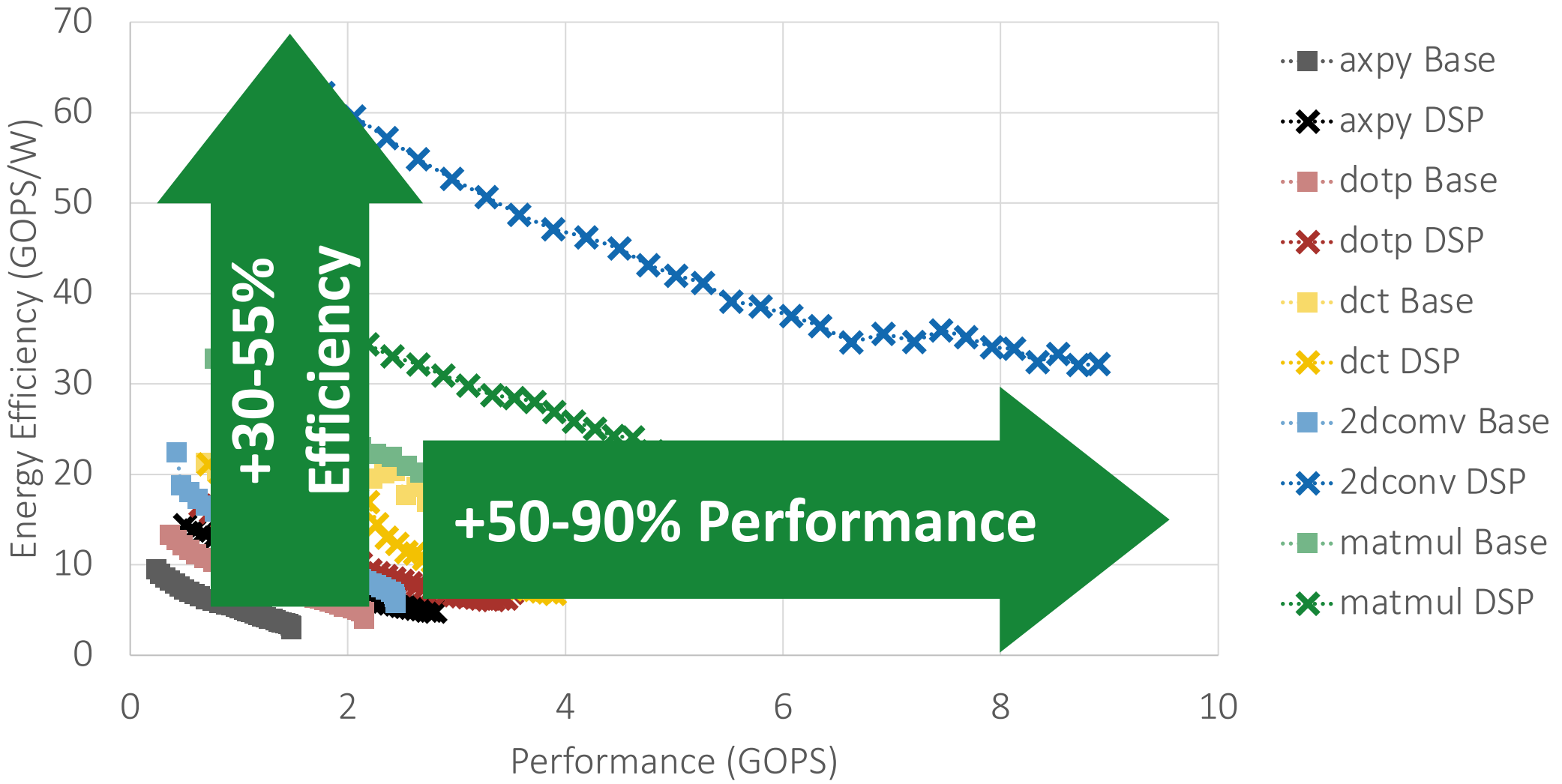
Energy Efficiency Base ISA



Energy Efficiency DSP extension

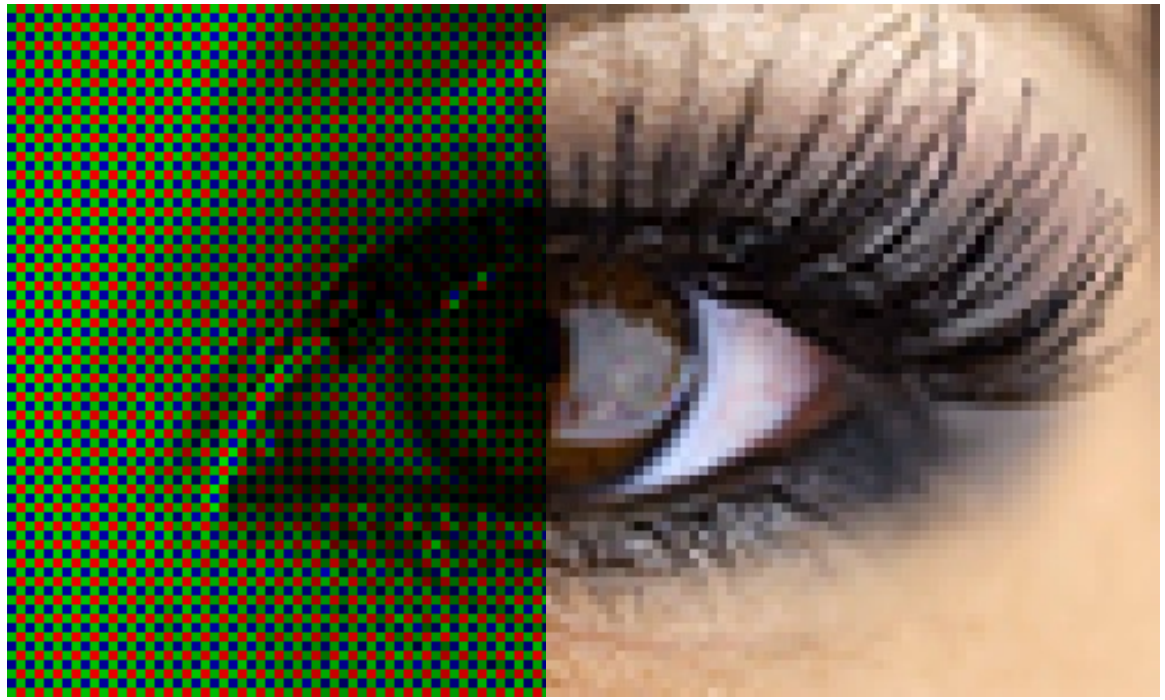


Energy Efficiency DSP extension



Bayer Pattern Demosaicing

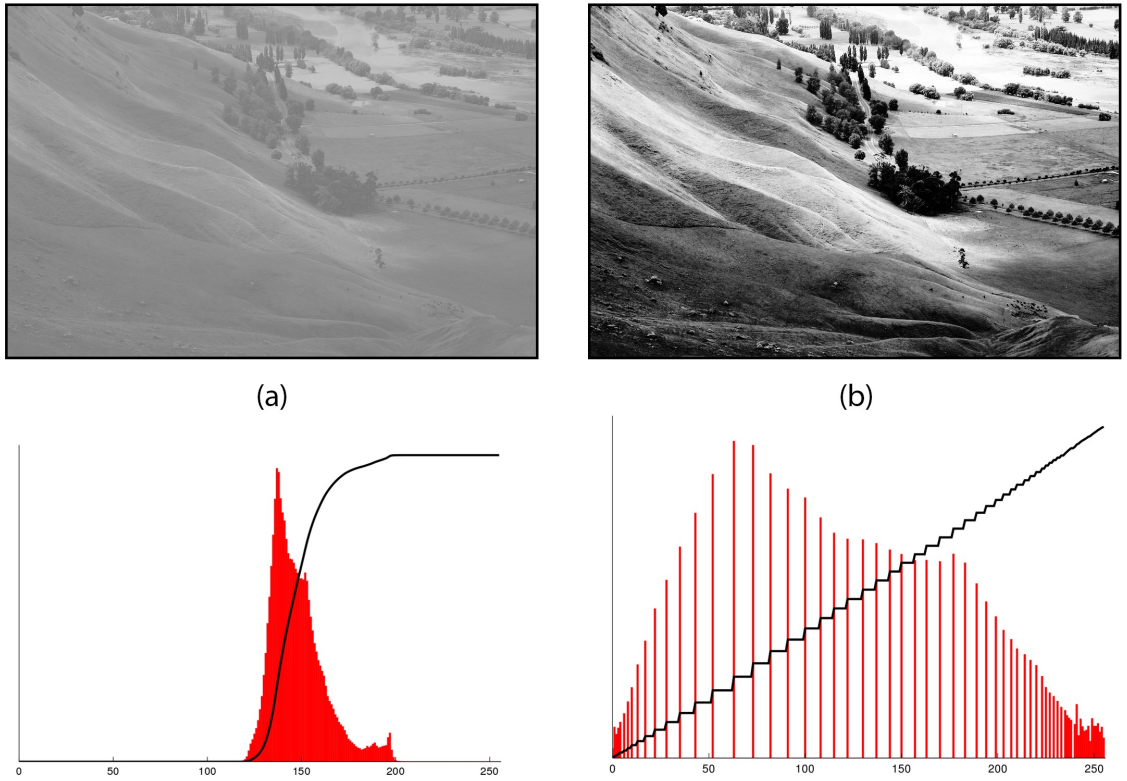
- Per Pixel operation
  - Bilinear interpolation
  - Requires surrounding pixel information
- Peak performance:
  380 Mpixel/s at 400 MHz
- Peak energy efficiency:
  2.7 Gpixel/s/W at 60 MHz



Source: https://www.red.com/red-101/bayer-sensor-strategy

Histogram Equalization

- Implemented in **Halide**
- Three key phases
  - Compute histogram
  - Compute transformation function
  - Transform image
- Peak performance:
  262 Mpixel/s at 400 MHz
- Peak energy efficiency:
  2 Gpixel/s/W at 60 MHz



Source: https://commons.wikimedia.org/wiki/File:Histogram_equalization.png

How do we compare?

	Thestral[7]	Vega[8]	Mr. Wolf[9]	Darkside[10]	MinPool
Technology	22nm	22nm	40nm	65nm	**65nm**
Max Frequency (MHz)				290	**400**
Number of Cores				8	**16**

Up to 4.8 times better performance
4.3 times better efficiency
36 times better area efficiency

ount

Only Vega is more efficient due to technology advantage

All numbers are for an int32 matrix multiplication

[7] T. Benz, L. Bertaccini, F. Zaruba, F. Schuiki, F. K. Gurkaynak, and L. Benini, "A 10-core soc with 20 fine-grain power domains for energy- proportional data-parallel processing over a wide voltage and temperature range," *ESSCIRC 2021 - IEEE 47th Eur. Solid State Circuits Conf. Proc.*, pp. 263–266, Sep. 2021.
[8] D. Rossi *et al.*, "Vega: A ten-core soc for iot endnodes with dnn acceleration and cognitive wake-up from mram-based state-retentive sleep mode," *IEEE J. Solid-State Circuits*, vol. 57, no. 1, pp. 127–139, Jan. 2022.
[9] A. Pullini, D. Rossi, I. Loi, A. Di Mauro, and L. Benini, "Mr. wolf: A 1 gflop/s energy-proportional parallel ultra low power soc for iot edge processing," *ESSCIRC 2018 - IEEE 44th Eur. Solid State Circuits Conf.*, pp. 130–133, Oct. 2018.
[10] A. Garofalo *et al.*, "Darkside: 2.6gflops, 8.7mw heterogeneous risc-v cluster for extreme-edge on-chip dnn inference and training," *ESSCIRC 2022 - IEEE 48th Eur. Solid State Circuits Conf. Proc.*, pp. 273–276, 2022.

A quick summary

- **MinPool:** An energy-efficient image processor
  - 16 cores with DSP extensions
  - 64 KiB of shared L1 data memory
  - ≤ 3 cycles latency (without contention)
- Implemented in 65nm
  - Up to 400MHz at only 490mW
- Flexible and easy to program
  - Independent cores
  - Shared memory

Illustrates the benefit of combining small, latency-tolerant cores with shared memory

Improves performance by a factor of **3.5×**.

Up to **3.6×** more energy efficient

Allows running full image processing pipelines easily