

Mr.Wolf: a 1 GFLOP/s Energy-Proportional Parallel Ultra Low Power SoC for IoT Edge Processing

Antonio Pullini¹, Davide Rossi²,
Igor Loi³, Alfio Di Mauro¹, Luca Benini^{1,2}

¹ETH Zürich ²University of Bologna

³Greenwaves Technologies

Content slide

- Introduction
- Mr. Wolf Overview
- Energy Proportional Architecture
- Efficient I/O subsystem
- Power Management
- Conclusion

Which Computing Platform?

SENSOR BANDWIDTH

COMPUTATIONAL DEMAND

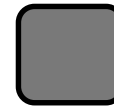
COMPUTING PLATFORM



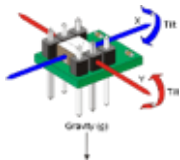
~ 1 bps



<1 MOPS



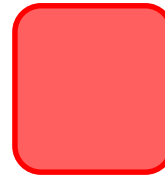
e.g. CortexM0



~1 Kbps



~10 MOPS



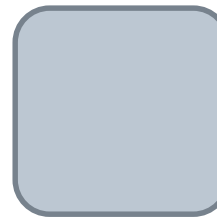
e.g. CortexM3



~100 Kbps



~ 100 MOPS



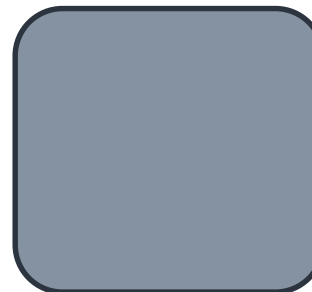
e.g. CortexM4/M4F



~1000 Kbps



> 1000 MOPS

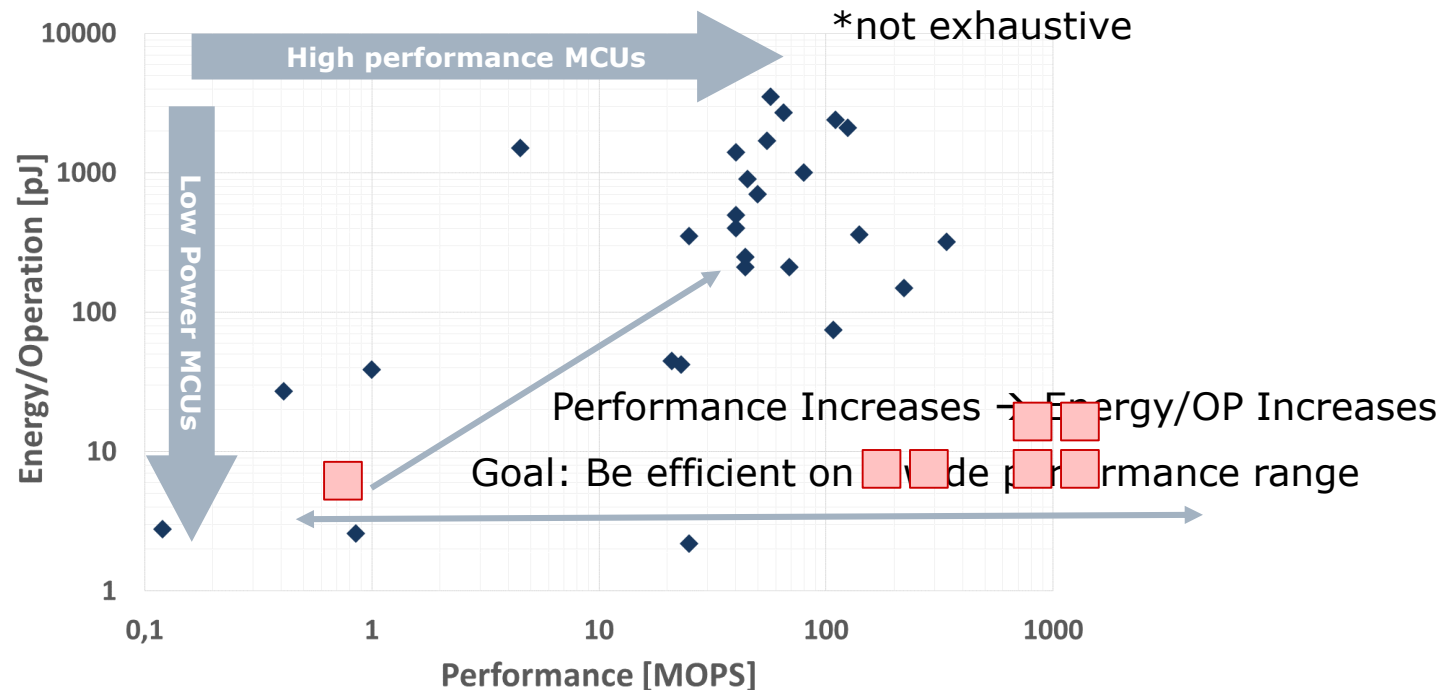


e.g. CortexM7/A

Antonio Pullini

Mr.Wolf: a 1 GFLOP/s Energy-Proportional Parallel
Ultra Low Power SoC for IoT Edge Processing

Energy-Proportional MCU

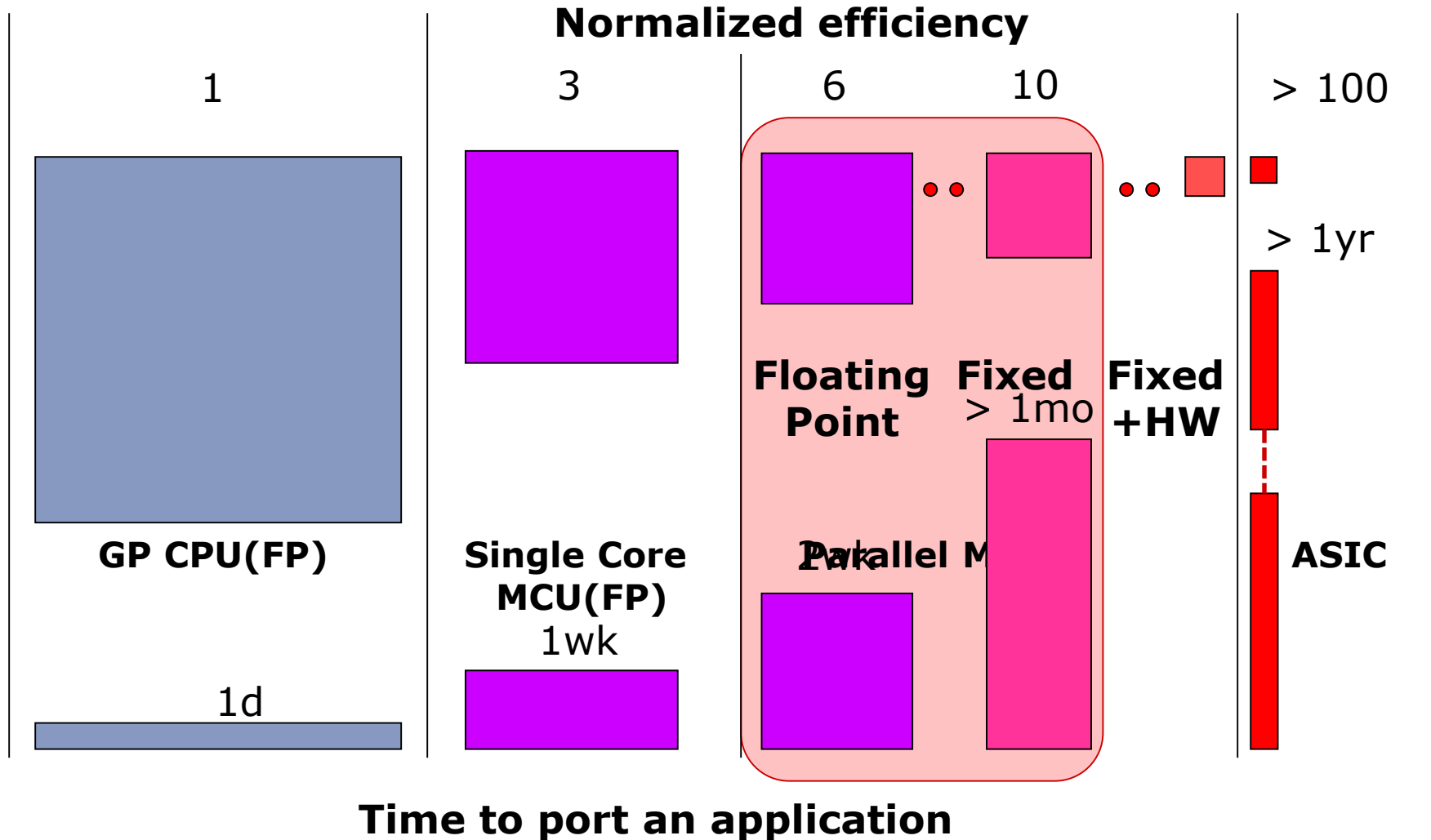


- ❑ Low Voltage Operation → Efficiency
- ❑ Parallelism → Performance
- ❑ Aggressive power management of unused resources
- ❑ Reduce the cost of parallelization (sharing + efficient synchronization)
- ❑ Duty Cycling → State Retentive Deep Sleep

Antonio Pullini

Mr. Wolf: a 1 GFLOP/s Energy-Proportional Parallel
Ultra Low Power SoC for IoT Edge Processing

Ease of use



Antonio Pullini

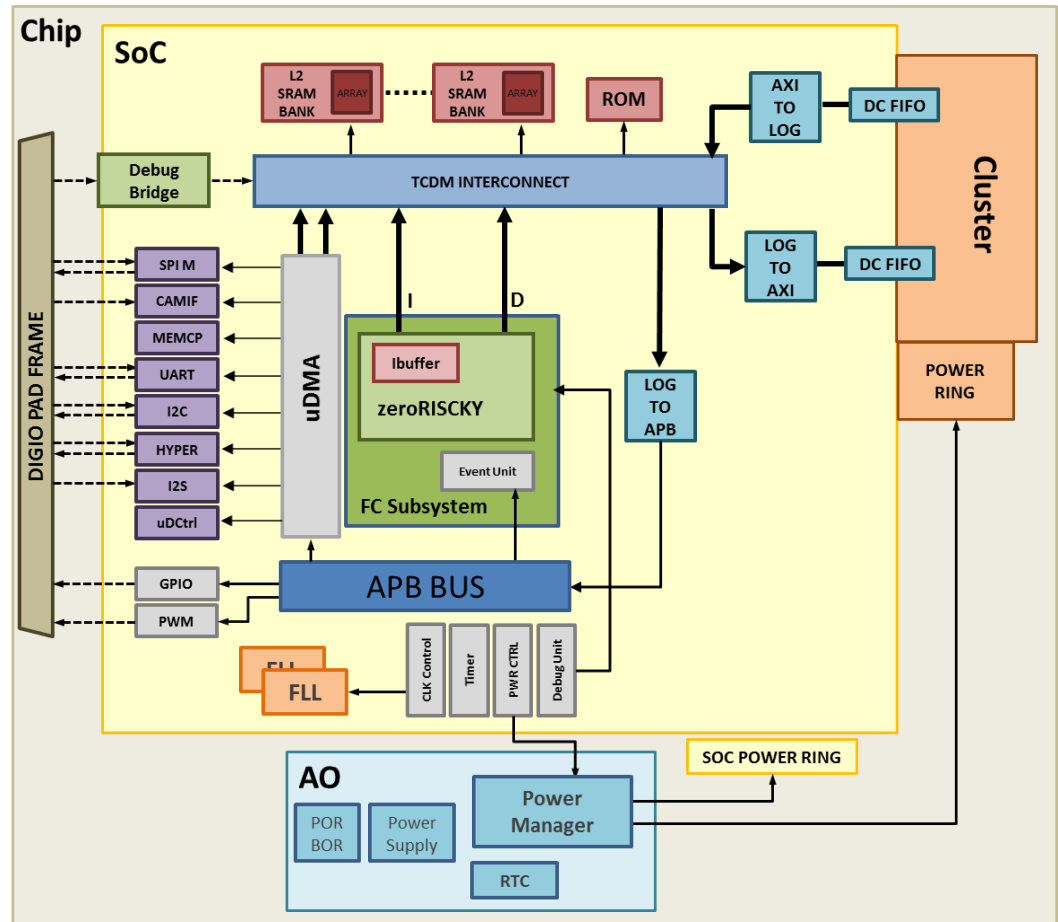
Mr.Wolf: a 1 GFLOP/s Energy-Proportional Parallel
Ultra Low Power SoC for IoT Edge Processing

Content slide

- Introduction
- Mr. Wolf Overview
- Energy Proportional Architecture
- Efficient I/O subsystem
- Power Management
- Conclusion

Mr. Wolf Overview

- **The SoC is an advanced microcontroller based on an ultra-low-power 32-bit RISC-V processor**
 - 512kB of L2 Memory
- **Rich set of peripherals:**
 - QSPI, SPI, I2C, I2S
 - HyperRam + HyperFlash
 - Camera Interface
 - JTAG (Debug), GPIOs, PWM, ROM
 - RTC, Interrupt controller
- **Autonomous IO DMA Subsystem**
- **Advanced power management**
 - 2 Switchable Power Domains
 - 2 low-power FLLs (IO, SoC, Cluster)
 - On-chip DCDC and LDO
 - State-Retentive L2 Memory
- **Parallel Programmable Accelerator:**
 - 8x DSP enhanced RISC-V processors
 - 64kB of Shared L1 Memory
 - 2x Shared Floating Point Units
 - Shared latch-based Instruction Cache
 - High performance DMA(L2<->L1)
 - Event Unit supporting fast synchronization among cores



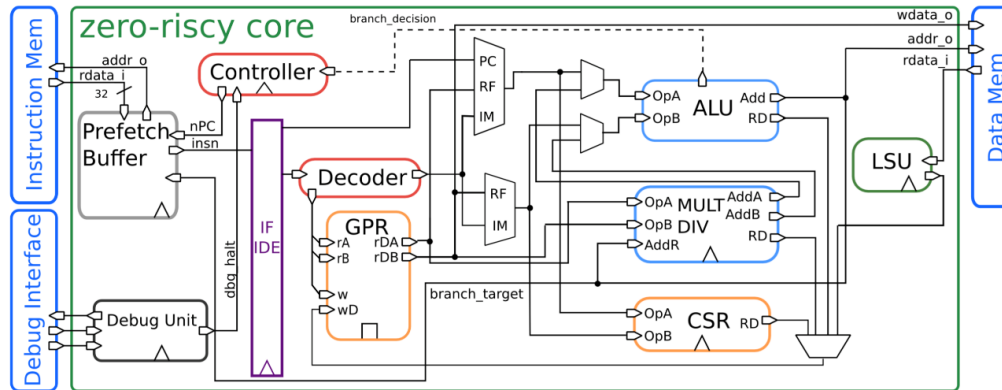
Antonio Pullini

Mr. Wolf: a 1 GFLOP/s Energy-Proportional Parallel
Ultra Low Power SoC for IoT Edge Processing

Content slide

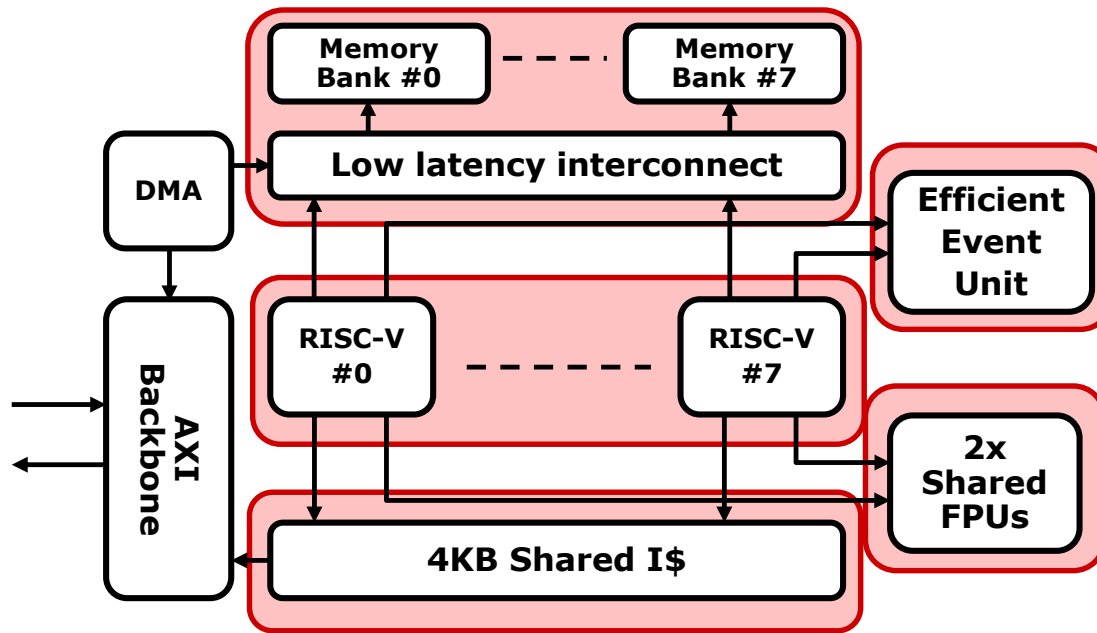
- Introduction
- Mr. Wolf Overview
- Energy Proportional Architecture
- Efficient I/O subsystem
- Power Management
- Conclusion

Zero-riscy SoC controller



- 2-stage pipeline
 - RV32IMC RISC-V w/Integer, Multiply, Compressed
 - **15 kGE**
 - 30 logic levels of critical path
 - Coremark/MHz 2.44
 - Optimized for area/leakage
 - High resource sharing
 - 3/4 cycles multiplier
 - 1bit iterative divider

Cluster Architecture

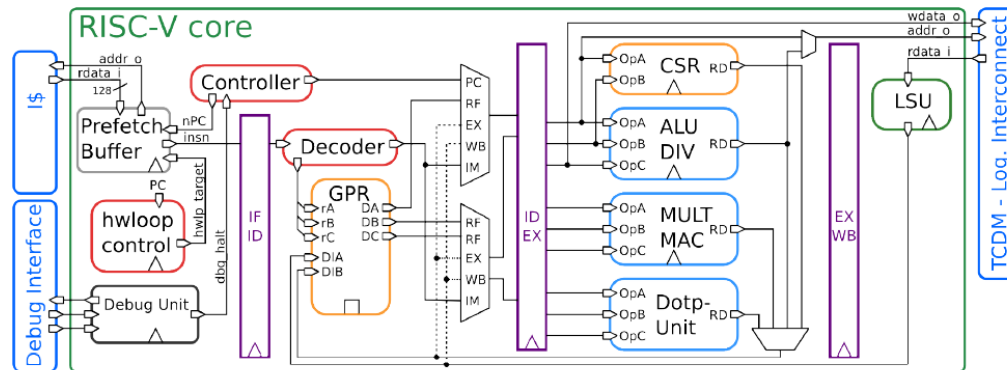


- ❑ 8 CPU
- ❑ Data shared through a low latency logarithmic interconnect(1 cycle access, <10% penalty w.c.)
- ❑ Ultra fast synchronization(2 cycles barriers)
- ❑ I\$ shared and implemented with SCM
- ❑ FPU sharing -> 4x Area reduction w/ <10% perf. Penalty
 - Supports FADD,FMUL,SQRT,DIV,FMA,ABS
 - Fully pipelined for high throughput except SQRT,DIV

Antonio Pullini

Mr.Wolf: a 1 GFLOP/s Energy-Proportional Parallel
Ultra Low Power SoC for IoT Edge Processing

Cluster Processing Element

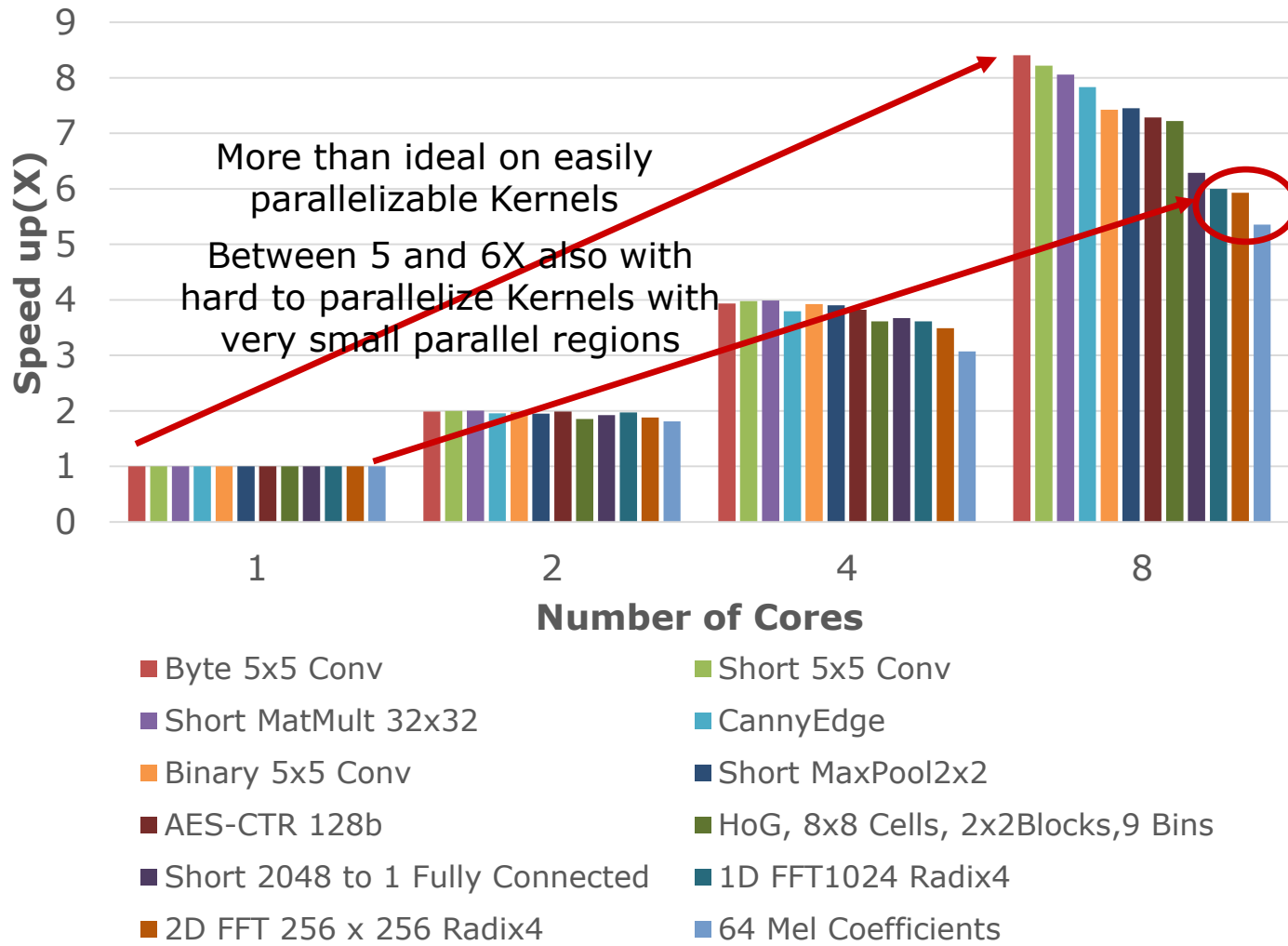


- ❑ 4-stage pipeline
 - RV32IM[F]C RISC-V w/Integer, Multiply, Compressed and Float
 - 60 kGE (vs. 15 kGE Zero-Risky)
 - 30 logic levels of critical path
 - Coremark/MHz 3.19 (2.44 Zero-Risky)
- ❑ Includes various DSP extensions
 - Packed SIMD
 - Fixed point
 - Bit manipulations
 - HW loops

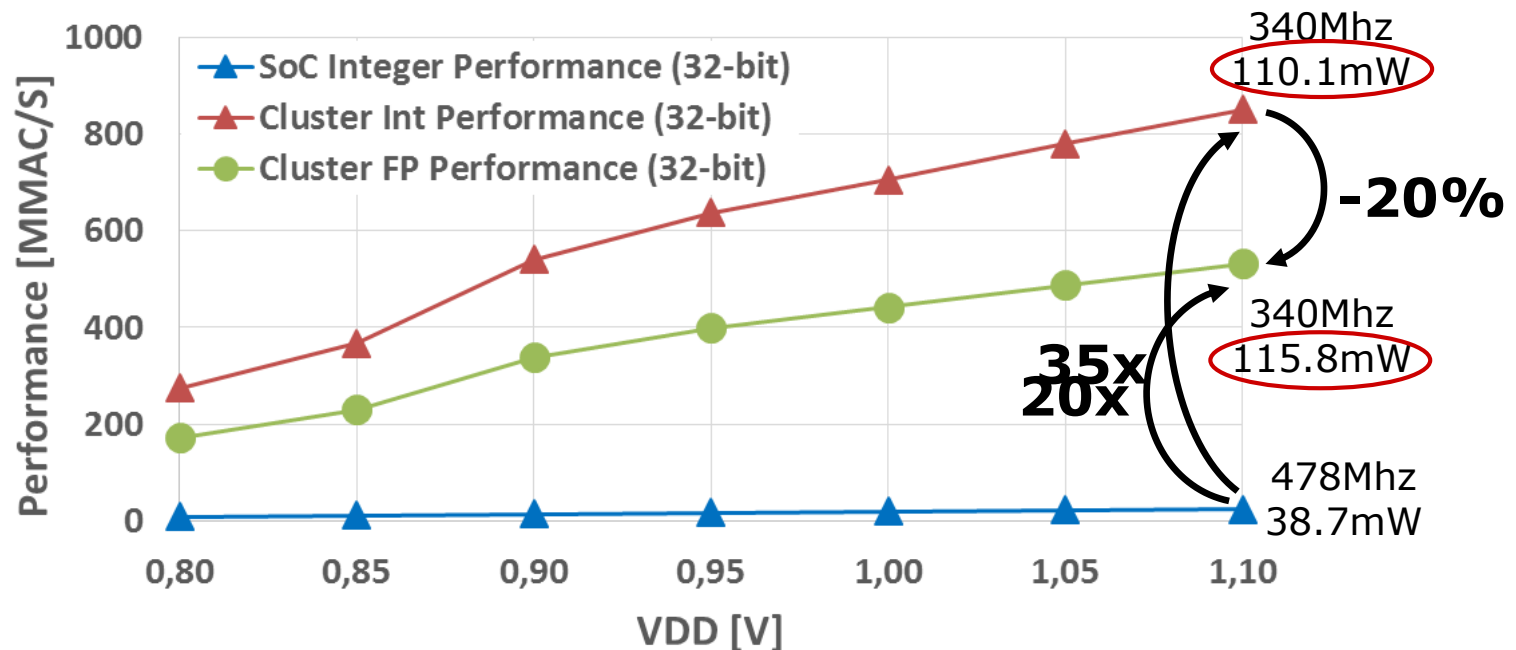


2.6x better than pure RISC-V and 5x better than Zero-Risky

Parallel speed-up

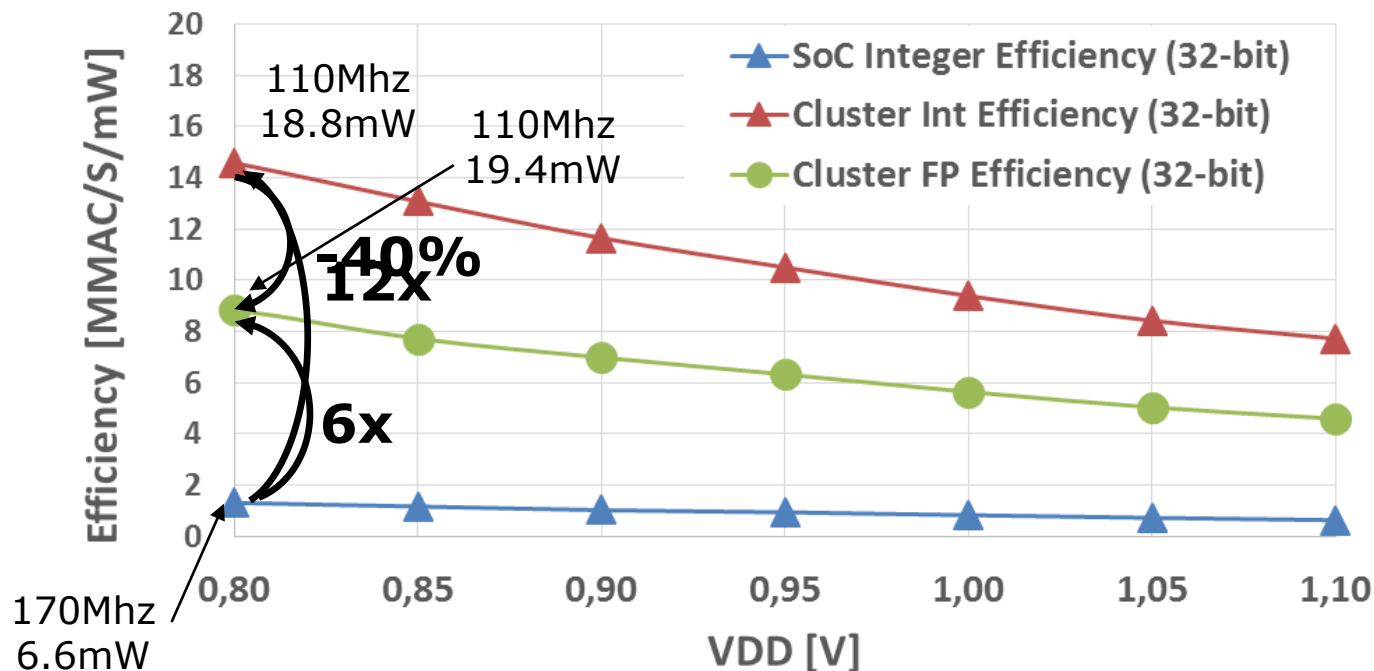


Performances on DSP kernels



- Peak integer performances 850MMAC/s @ 320MHz, 1.1V
 - DSP optimized CPUs + 8xCPU
- Peak FP performances 500FMMAC/s @ 320MHz, 1.1V
 - Loss vs Integer due to sharing of FPU resources
- Similar peak power consumption when using FP

Efficiency on DSP kernels

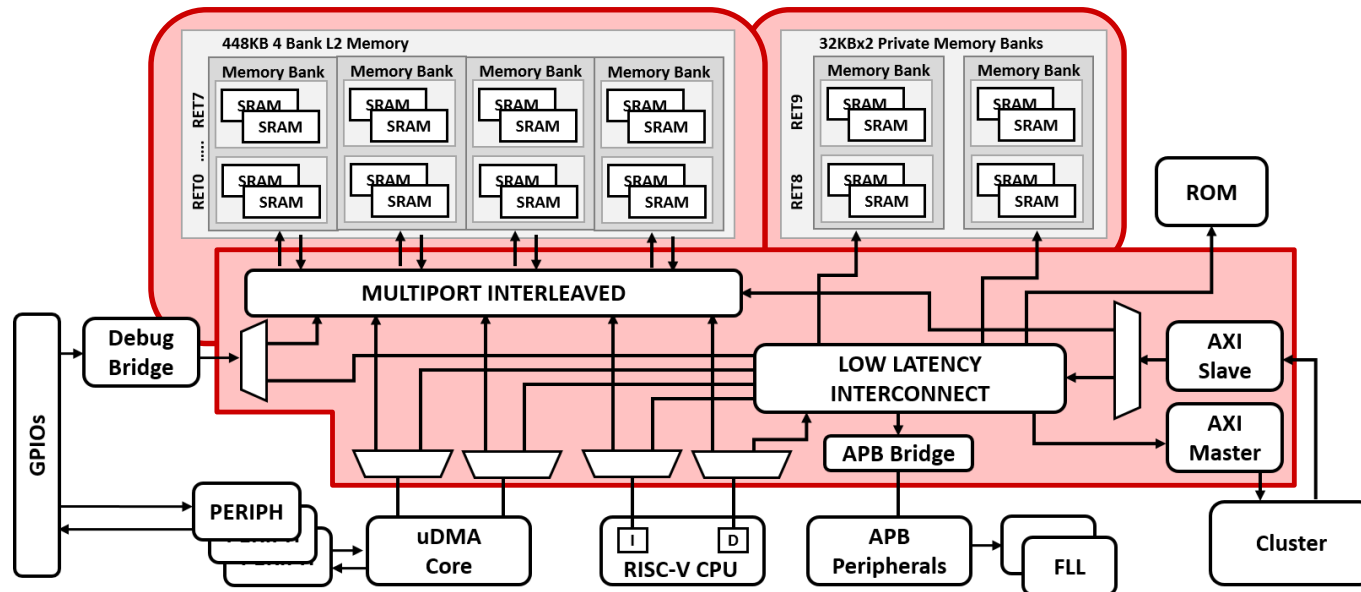


- Peak integer efficiency 15MMAC/s/mW @ 110MHz, 0.8V
 - 30MMAC/s/mW and 60MMAC/s/mW thanks to packed SIMD when executing 16bits or 8bits operations
- Peak FP performances 9 FPMAC/s/mW @ 110MHz, 0.8V

Content slide

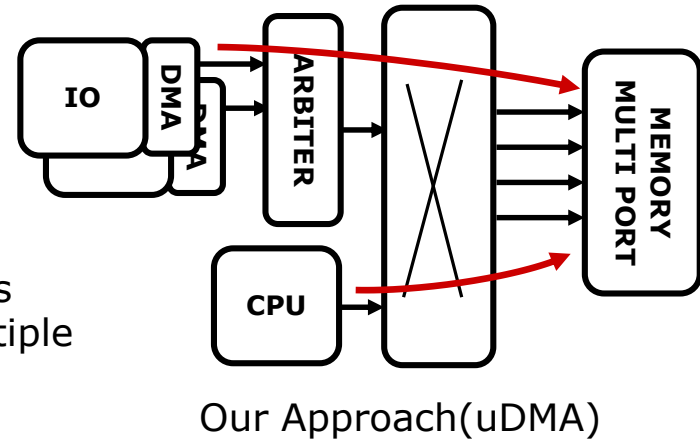
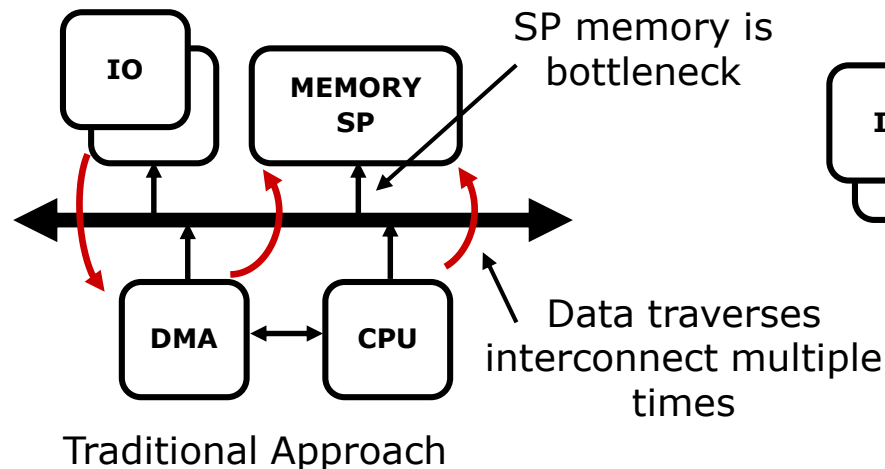
- Introduction
- Mr. Wolf Overview
- Energy Proportional Architecture
- Efficient I/O subsystem
- Power Management
- Conclusion

SoC Interconnect



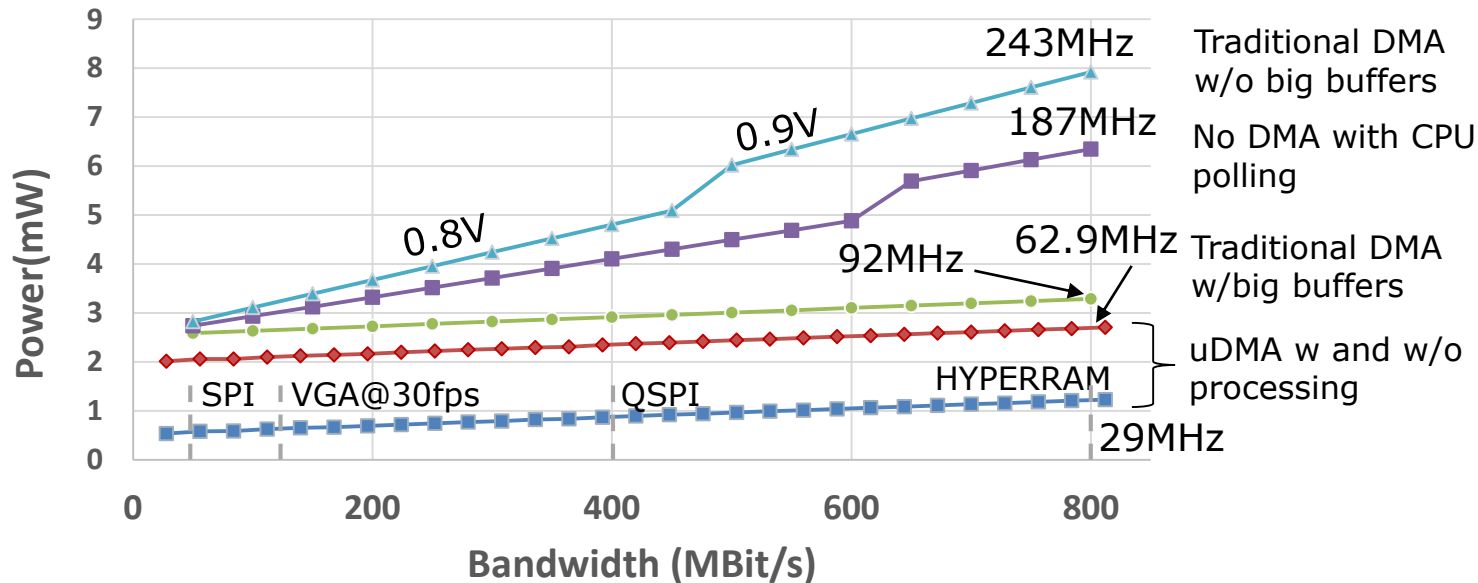
- ❑ Low latency interconnect with 1 cycle access when no contention
- ❑ 4-port Shared L2 w/interleaving support for efficient data sharing between Cluster, CPU and I/O
- ❑ Private regions for low latency access(code, stack)

uDMA Architecture



- Dedicated DMA channels for each peripheral
 - Peripherals themselves are controlling data transfer
 - Peripherals can be fully autonomous
- Fair arbitration between channels + dedicated ports to memory subsystem -> bounded latency to memory
 - Very small buffers needed
- Low contention on memory and on interconnect
 - I/O transfer do not interfere with CPU operation

uDMA results

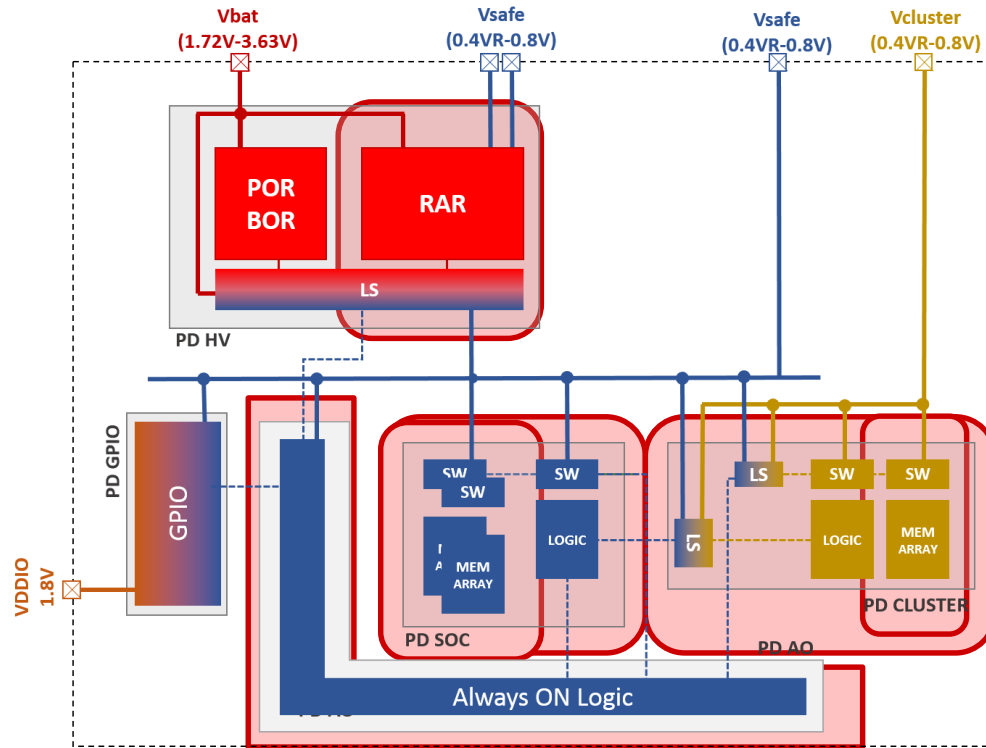


- Allows to operate the system at low frequency also with high I/O BW(800Mbit/s with Soc running as low as 29MHz)
- Similar performances to traditional systems with DMA and big(1KB) buffer dedicated to each peripherals
- I/O transfer happens in parallel to CPU operation
 - CPU@50MOPs -> 62.9MHz, I/O@800MBit/s -> 29MHz
 - CPU@50MOPs+I/O@800Mbit/s -> 62.9MHz(not true with traditional approach where operating frequency is added)

Content slide

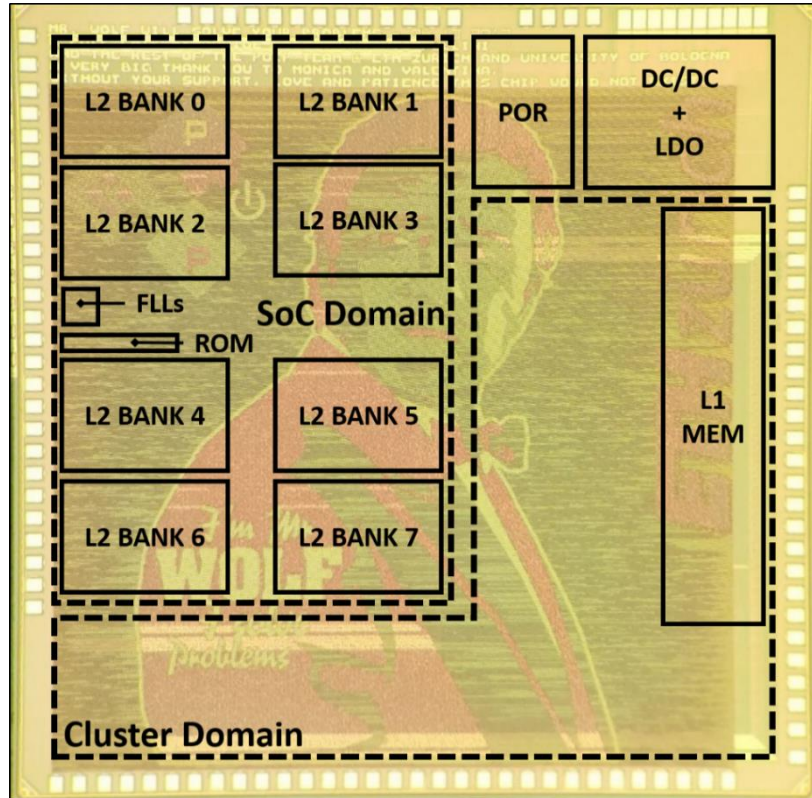
- Introduction
- Mr. Wolf Overview
- Energy Proportional Architecture
- Efficient I/O subsystem
- Power Management
- Conclusion

Power Management



- 2 switchable power domains, 1 Always ON
- Retention Alternating Regulator with high efficiency DC/DC + Low Quiescent Current LDO for retention provided by Dolphin Integration
- Low Leakage retentive Memory cuts
- 10 Levels of retention on L2 memory(72uW-108uW) 70nW/KB

Mr. Wolf Chip Results



Technology	CMOS 40nm LP
Chip area	10 mm ²
VDD range	0.8V - 1.1V
Memory Transistors	576 Kbytes
Logic Transistors	1.8 M gates
Frequency Range	32 kHz – 450 MHz
Power Range	72 μ W – 153 mW

Power Modes	VDD [V]	Freq.	Power
<i>Deep Sleep</i>	0.8	<i>n.a.</i>	72 μ W
<i>Ret. Deep Sleep</i>	0.8	<i>n.a.</i>	76.5 - 108 μ W
<i>SoC Active</i>	0.8 - 1.1	32 kHz - 450 MHz	0.97 - 38 mW
<i>Cluster Active</i>	0.8 - 1.1	32 kHz - 350 MHz	1.6 - 115 mW

Comparison with SOA

	LOW-POWER MCUs		WIDE PERF. RANGE DSPs		PULP SoCs		
	SLEEPWALKER	REISC	HEXAGON	FRISBEE	PULPv2	FULMINE	Mr.Wolf (this work)
Technology	CMOS 65nm LP GP	CMOS 65nm LP	CMOS 28nm LP	FD-SOI 28nm flip-well	FD-SOI 28nm flip-well	CMOS 65nm	CMOS 40nm LP
CPU	16-bit MSP430	32-bit	4x 32-bit VLIW	32-bit	32-bit OpenRISC	32-bit OpenRISC	32-bit RCVC32IMFX
FPU	no	no	no	no	no	No	2 shared
# of cores	1	1	1	1	4	4	1 + 8
I\$/D\$/L2	16kB(64b)/ 2kB/ n.a.	8kB(128b)/ 8kB(128b)/ n.a.	16kB/ 32kB/ 256kB	4kB/ 4kB/ n.a.	1kBx4/ 48kB/ 64kB	4K/ 64kB/ 192kB	4K/ 64kB/ 512kB
Power Management	DCDC + power gating	n.a.	LDO	n.a.	clock gating	off-chip DC/DC	DCDC + LDO + power gating
Sleep Power	1.5 μ W state ret.	n.a.	n.a.	n.a.	200 μ W	120 μ W	72 – 108 μW state ret.
IO Efficiency	Low	Low	Low	n.a.	Low	Mid	High
Voltage range (SRAMs)	0.4V (1.0V)	0.54V - 1.2V (0.4V - 1.2V)	0.6V - 1.05V	0.4V - 1.3V	0.32V – 1.15V (0.45V – 1.15V)	0.8V – 1.1V	0.8V – 1.1V
Max frequency	25 MHz	82.5 MHz	1.2 GHz	2.6 GHz	825 MHz	400 MHz	450 MHz
Best performance	25 MOPS	82.5 MOPS	3 GOPS	2.6 GOPS	3.3 GOPS	4.2 GOPS	7 GOPS
Best energy efficiency	64.5 MOPS/mW @ 25 MOPS	98 MOPS/mW @ 0.54 MOPS	43.1 MOPS/mW @ 230 MOPS	16.1 MOPS/mW @ 460 MOPS	193 MOPS/mW @ 162 MOPS	69 MOPS/mW @ 1.2 GOPS	120 MOPS/mW @2.2 GOPS

Conclusion

- **Mr. Wolf Application Processor for IoT**
 - Power/performance scalability with on-chip power management
 - Active: up to 7 GOPS pure RISC-V equivalent instructions
 - Idle: 108 μ W with 512 Kbytes of state-retentive memory
 - IO + Computing Efficiency
 - Up to 120 MOPS/mW ($>10\times$ w.r.t commercial micro-controllers)
- **Open-Source HW and SW approach**
 - Full Mr. Wolf RTL (except technology dependent IPs)
 - Compiler/software
 - Simulation tools/debug
- **Next step (Vincent Vega)**
 - 22nm FD-SOI Application processor
 - Deep near-threshold operation: temperature/process compensation with body biasing
 - Non-volatile on chip memory (4MB eMRAM)
 - More peripherals + floating point small formats