



Freedom Revolution

A Customizable RISC-V AI SoC Platform

Krste Asanovic

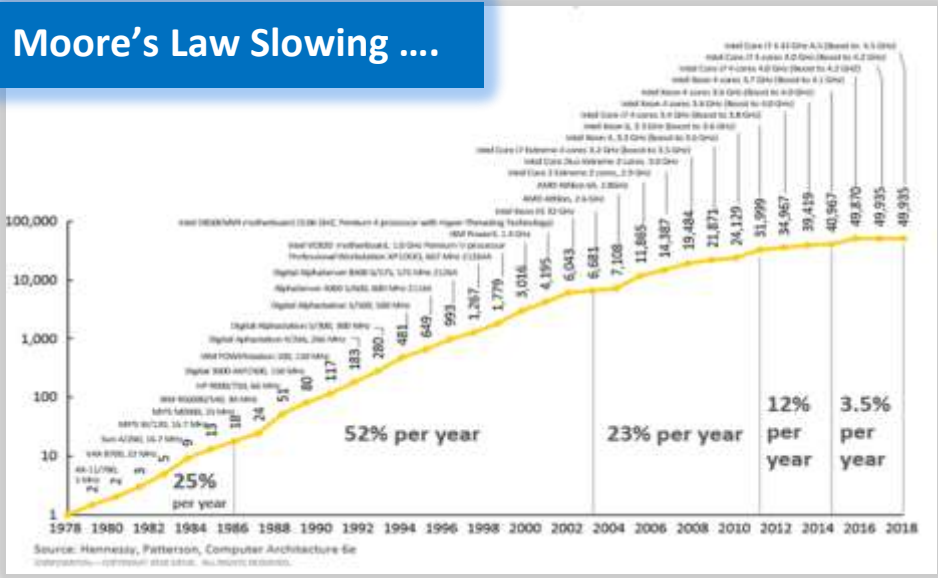
Co-Founder and Chief Architect



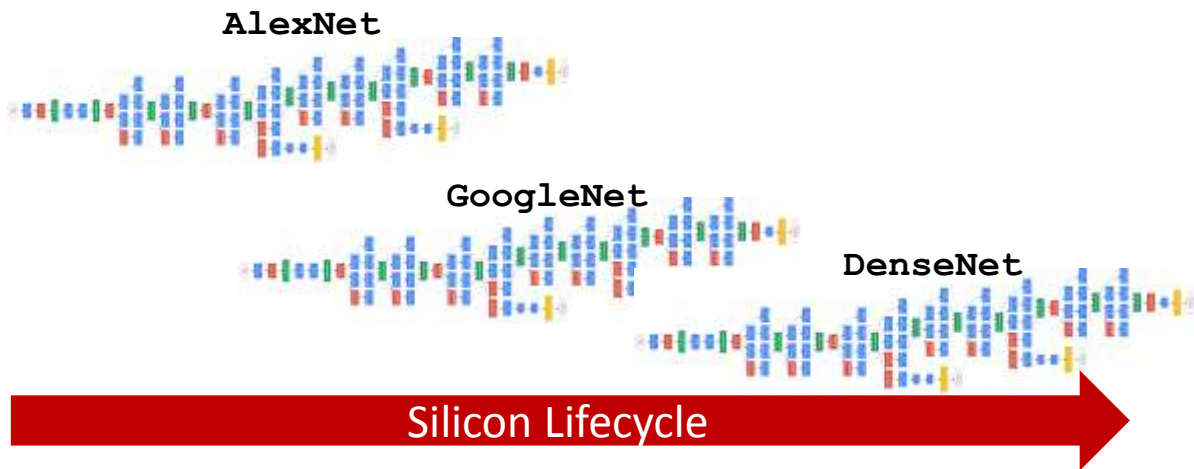


AI Silicon Challenges

Moore's Law Slowing



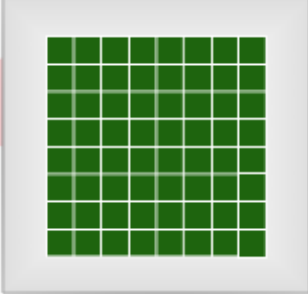
AI Models Evolving Quickly



Need a New Si Approach



CISC-> RISC-> Multi-core



Fixed H/W Accelerators
GPU, ASSP, ASIC



Domain Specific Architectures

Cost

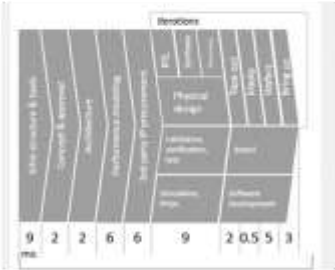
Chip Development Too Costly



\$500M+ for 7nm

Time

Development Cycle Too Long



2 - 4 years

Expertise

Too Many Experts Needed



14+ Disciplines



AI Accelerator Design Metrics

Inference at edge: cost/performance/W matter most

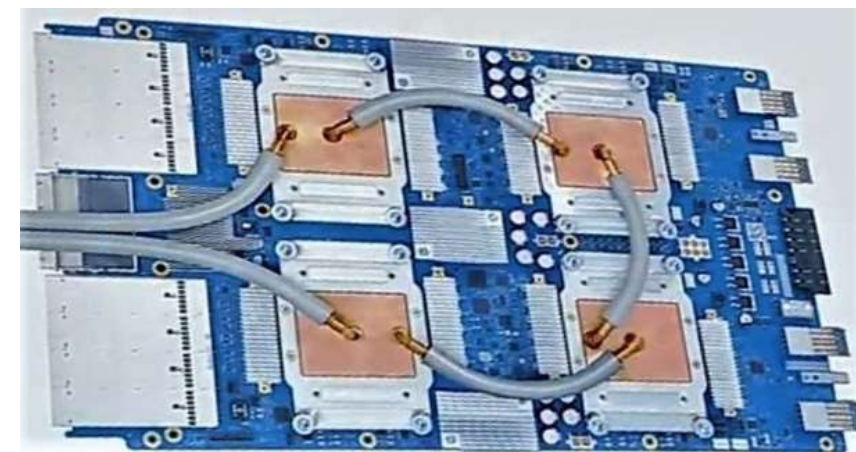
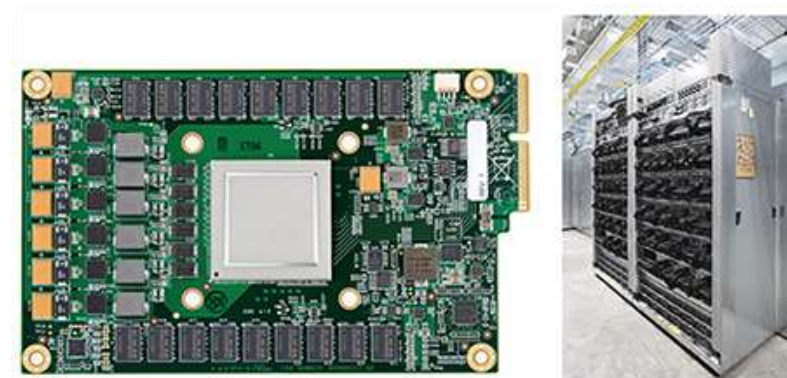
- Want highly compressed models to reduce system cost and power
- Trade performance for cost

Inference in cloud: latency, throughput/cost

- Value in serving user requests quickly and predictably
- Many users, so need to keep cost/user down

Training in cloud: performance, performance, performance


- Value is in data models created, justifies high cost/power
- Single training run can take weeks
- Optimize scarce resource of skilled human developers
- More similar to traditional HPC than conventional cloud
- Extreme technology, most advanced node, interconnects, cooling





Leveraging Cloud & Templates to Accelerate Chips to Market!

Customer IP (Chip)



Core/Block/Chip Configurators/Generators



Chip Templates



Chisel



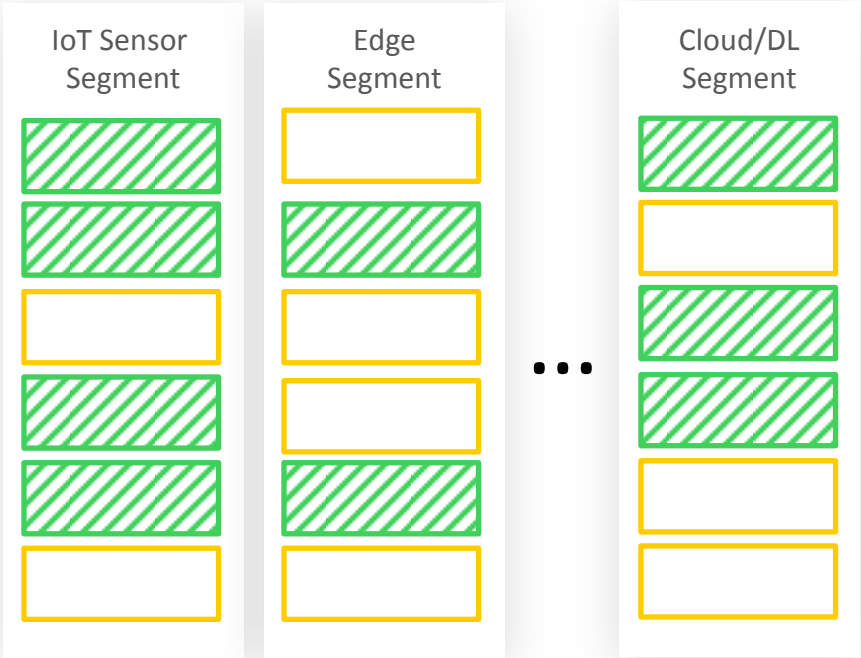
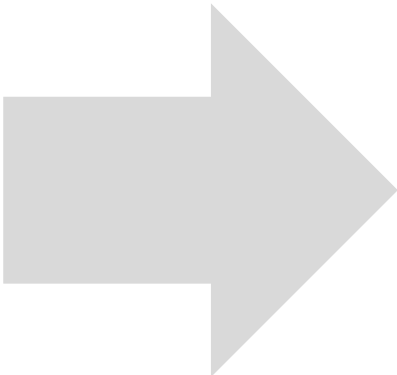
EDA Tools



Hardware



« Innovators focus on high-level work



28nm TSMC
1.5GHz 5x RISC-V Cores
DDR4, GE, Peripherals



DesignShare Partners
20+ in the Pipeline....

World's 1st Cloud Tape-out with Microsoft



TMACS/W/mm² are Only Part of the Story

Hardware

Most complexity in an AI enabled SoC is not AI specific

Memory interfaces

I/O interfaces

On-chip interconnect

Software

AI specific

Compiling models to run on custom compute engines

Not AI specific

Supporting devices/interfaces on an SoC



Embedding Intelligence from the Edge to the Cloud



U Cores

**64-bit Application
Processors**



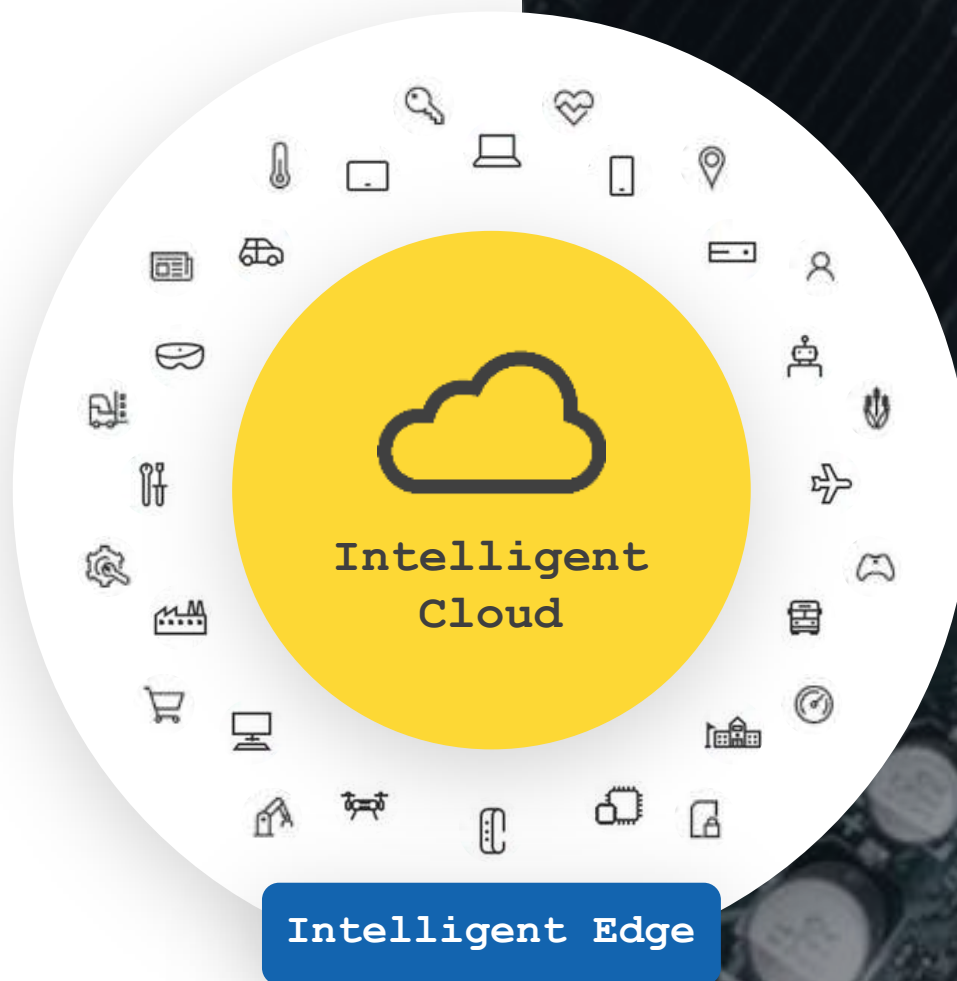
S Cores

**64-bit Embedded
Processors**



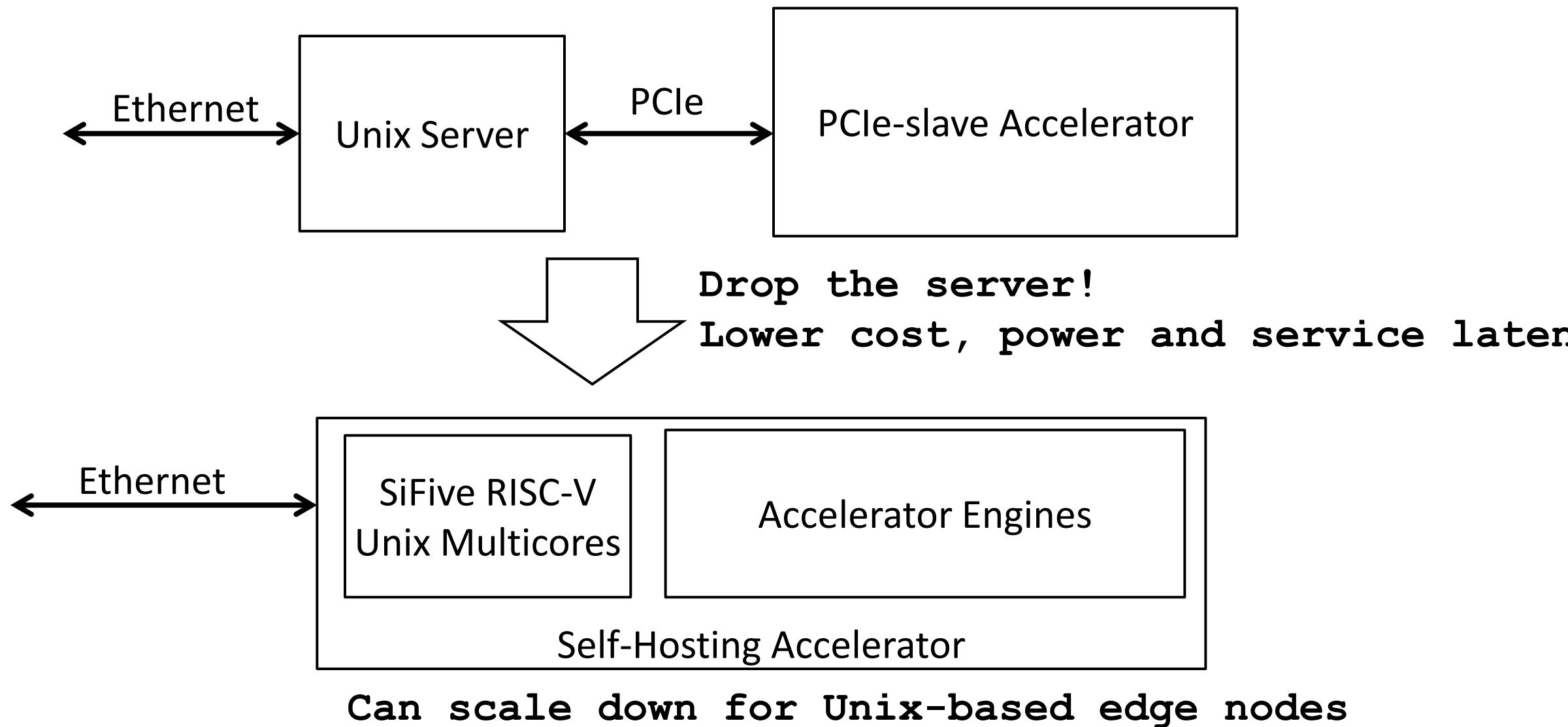
E Cores

**32-bit Embedded
Processors**



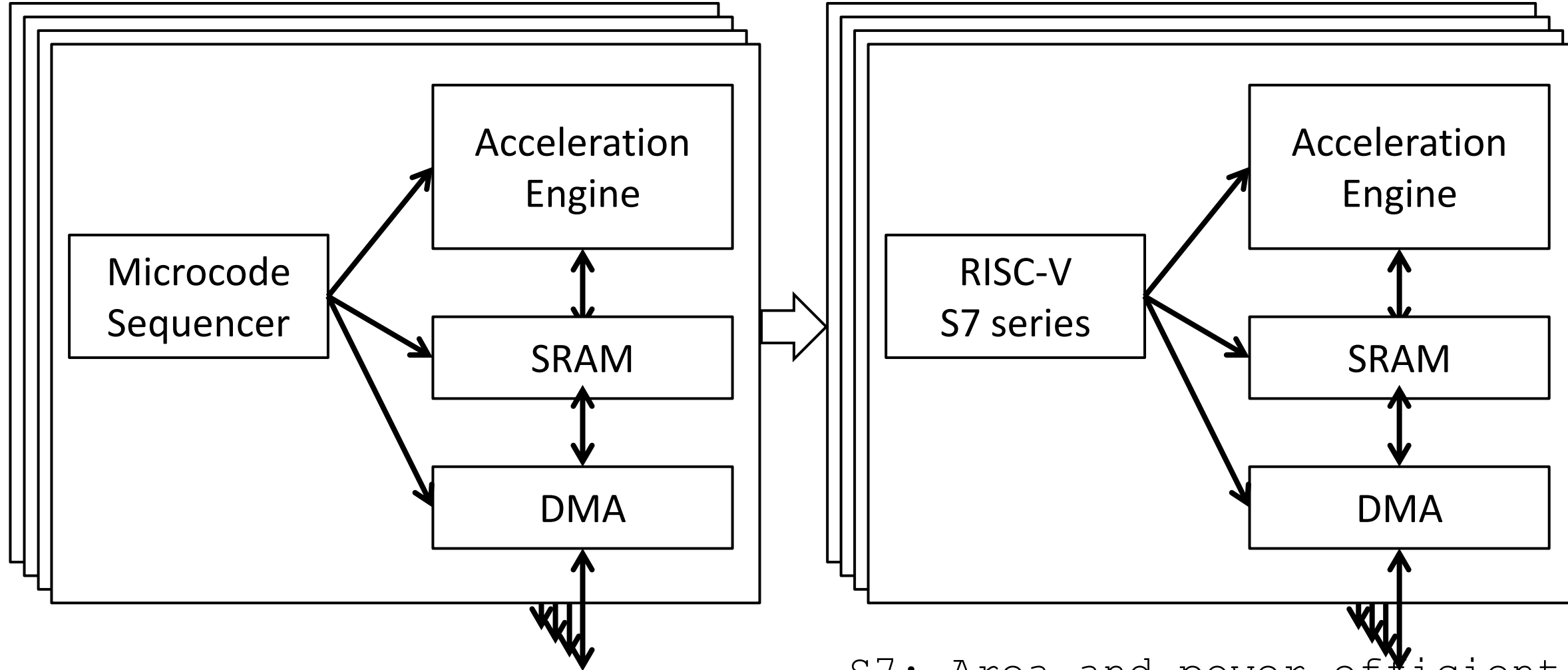


SiFive U-cores for Self-Hosted Accelerators





SiFive S-series for Management Cores



S7: Area and power-efficient
dual-issue 64-bit embedded
cores



RISC-V Foundation Vector Extension Overview

vl

Vector length CSR sets number of elements active in each instruction

vtype

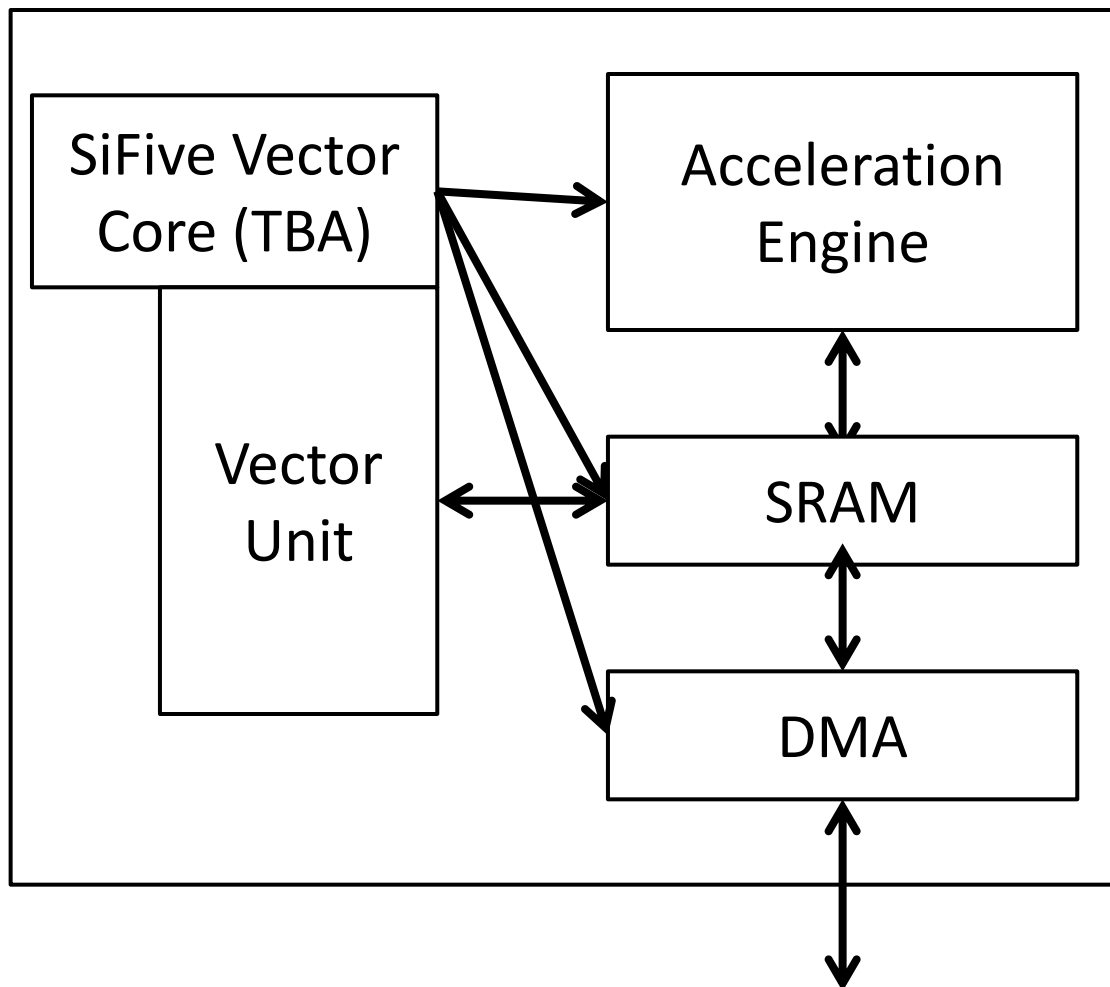
Vetype sets width of element in each vector register (e.g., 32-bit, 16-bit)

32 vector registers

v31[0]	v31[1]		v31[VLMAX-1]
v1[0]	v1[1]		v1[VLMAX-1]
v0[0]	v0[1]		v0[VLMAX-1]

- Unit-stride, strided, scatter-gather, structure load/store instructions
- Rich set of integer, fixed-point, and floating-point instructions
- Vector-vector, vector-scalar, and vector-immediate instructions
- Multiple vector registers can be combined to form longer vectors to reduce instruction bandwidth or support mixed-precision operations (e.g., 16b*16b->32b multiply-accumulate)
- Designed for extension with custom datatypes and widths

Maximum vector length (VLMAX) depends on implementation, number of vector registers used, and type of each element.



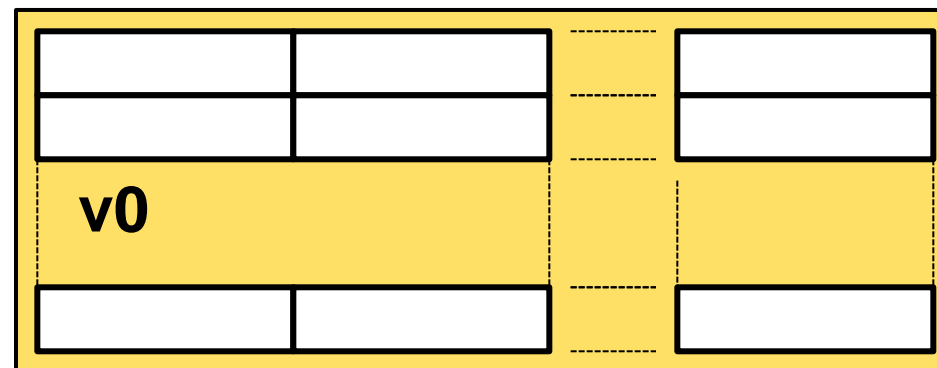
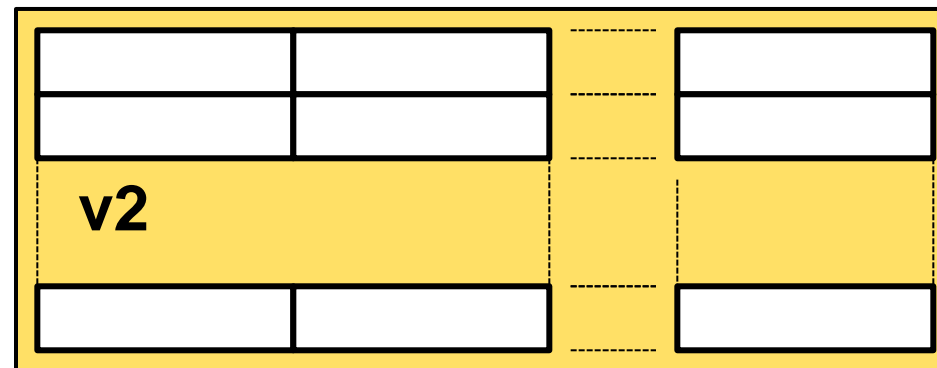
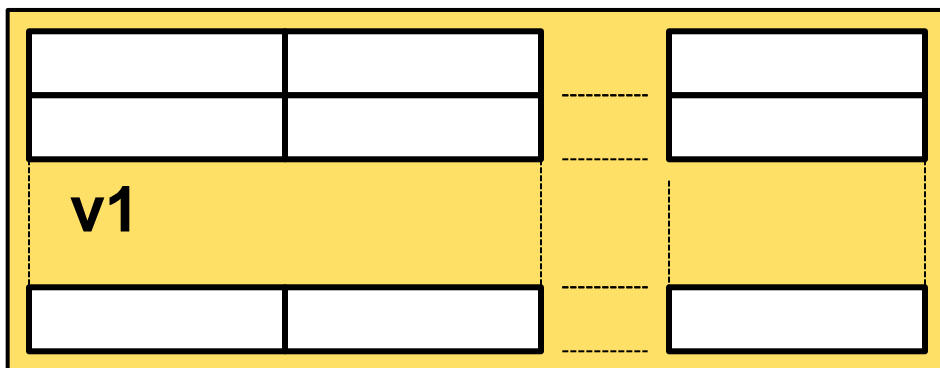
Vector unit can provide programmable support to fixed acceleration engine

E.g., feature extraction/normalization, activation functions, convergence statistics



RISC-V 2-D Vector Extensions

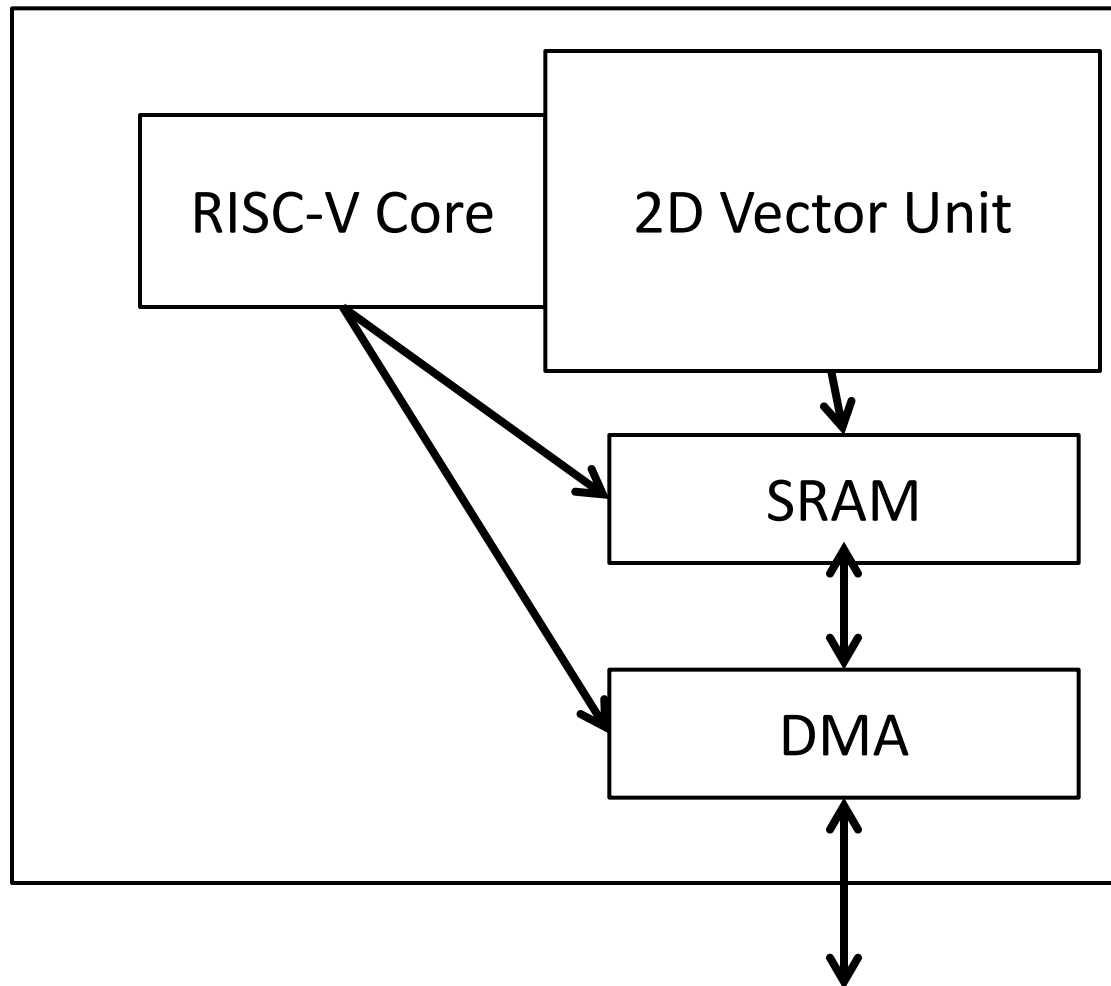
Matrix multiply $v0 += v1 * v2$
Vfmmadd v0,v1,v2



- Vector registers configured as 2D matrices
- Single instructions for matrix multiply and convolutions



RISC-V Vector Unit as Edge Inference Engine



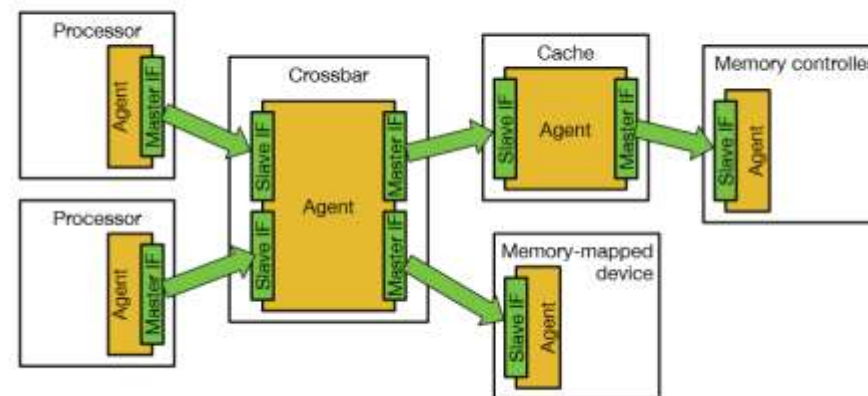
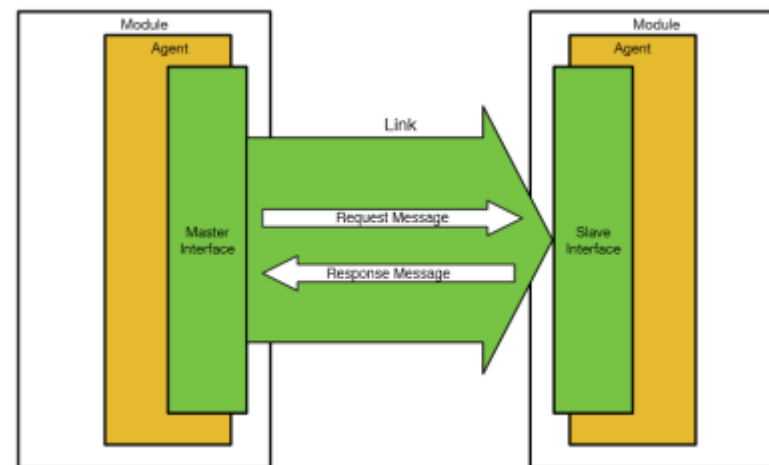
2D Vector Unit can provide efficient execution of all parts of algorithm

Custom extensions for new compressed data types



TileLink – High-Performance Scalable Cache-Coherent Fabric

- Open standard designed for RISC-V
- In production since Rocket chip (open source SoC)
- Can be bridged to existing AMBA designs
- Clean slate: avoid prior pitfalls
- Decouple message protocol from wire protocol
- Designing for RISC-V requirements = simpler design
 - Reduced Message Protocol; RISC => RMP
 - Assume all connected hardware is trusted
 - Only power-of-2 block transfers



Example of a TileLink Network Topology



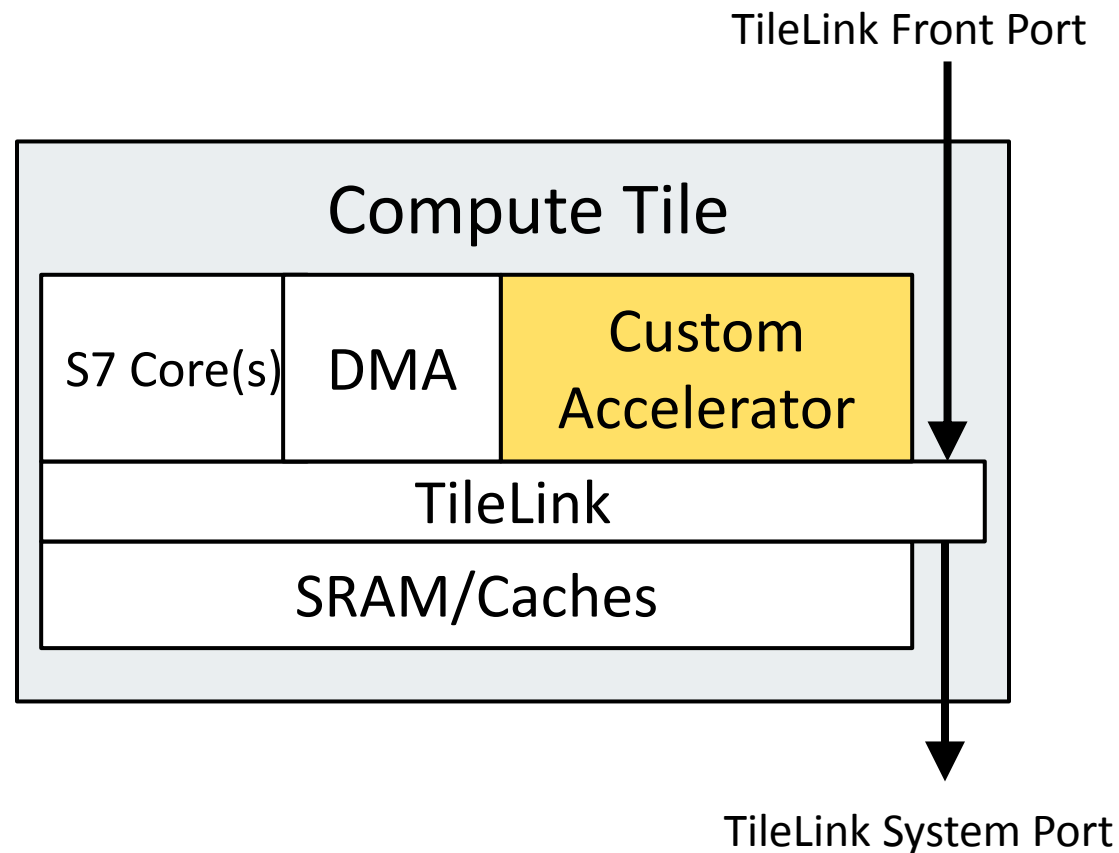
SiFive Accelerator Bays

Accelerator bay provides system plumbing to connect custom accelerator into system

- Management core or multicore
- Local scratchpad memory or caches
- Memory interconnect
- Interrupts
- DMA

Accelerator types:

- SiFive-supplied IP
- Customer-supplied IP
- Open-source (e.g., NVDLA)
- HLS-generated



Not just for AI, many other use cases.



SiFive Freedom Chip/ IP Validation Platforms

Freedom Everywhere
TSMC 180nm



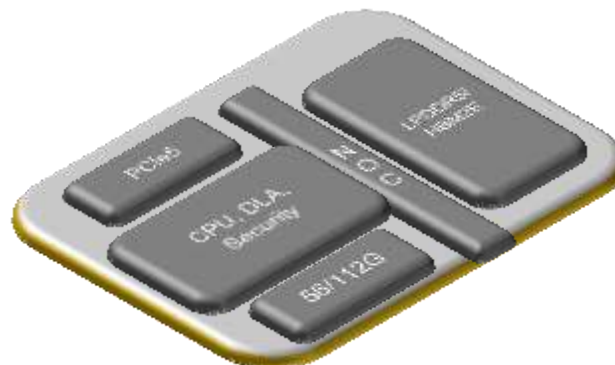
Freedom Unleashed
TSMC 28nm



Freedom Revolution
TSMC 16nm
HBM2 IP Validation



Freedom Revolution
TSMC 7nm



High-Speed Interface (TSMC 7nm)

- 2GHz TCAM
- HBM2E 3.2Gbps Interface
- LVDS Interface
- PLL, DLL, LDO
- PVT Sensor

☒ Tapeout Q1 2019

HBM2E 2.5D SiP (TSMC 7nm)

- E27 Core
- TileLink
- High-Performance Caches
- HBM2E 3.2Gbps PHY & Controller
- LL-HBM Support
- DMA, Peripherals

☐ Q3 2019

Freedom Revolution (TSMC 7nm)

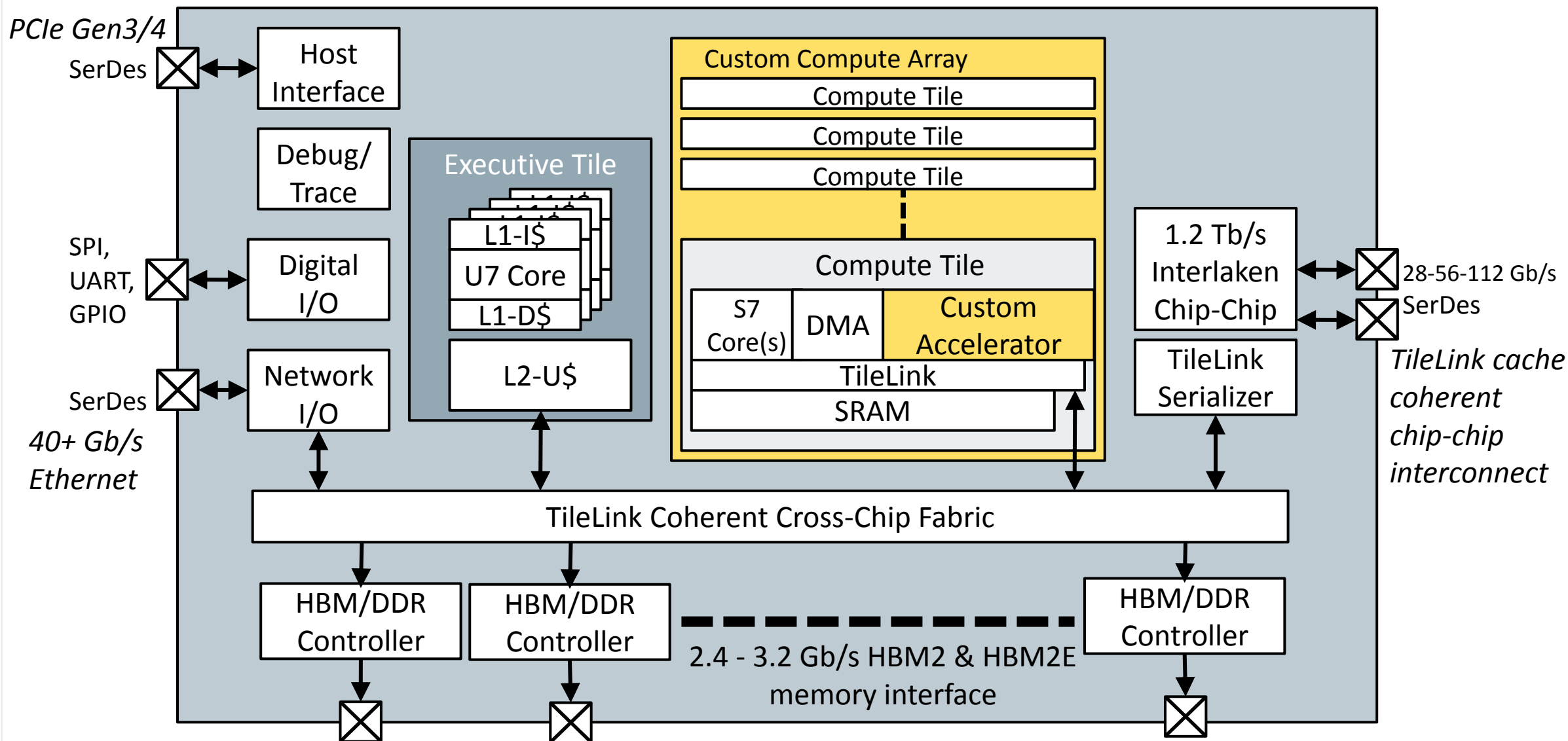
- U7 & S7 Series MC
- TileLink
- Accelerator Bays & Security
- LPDDR5/HBM2E PHY & Controller
- PCIe5 SerDes
- 56/112G SerDes

☐ Q4 2019

**includes some partner IP*



Freedom Revolution Platform for AI applications in TSMC 16nm/7nm





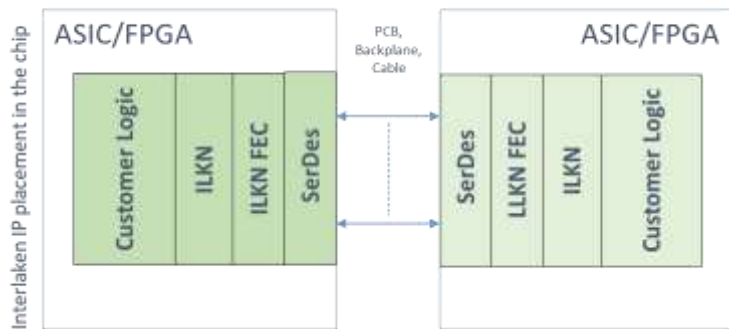
High-Speed Interface IP Subsystems

Interlaken IP Subsystem

ILKN & ILKN FEC

- Works with up to 48 parallel physical SerDes lanes 3.125 Gbps to 56 Gbps speeds
- Supports bandwidth of up to 1.2 Tbps
- Configurable user interface of 128/256 bit width

Interlaken IP 75+ licensees to date
10+ years of Tier-1 customer engagements

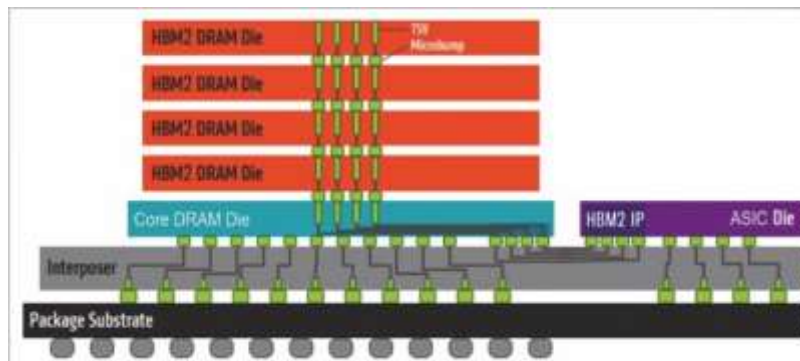
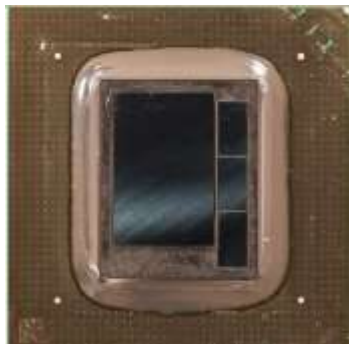


Ethernet IP Subsystem

- **MAC IP** supports 400/200/100/50/25/10GE client port
- **FlexE IP** is fully compliant to OIF Flex Ethernet Standard v1.1
- **PCS IP** supports IEEE standard 802.3 for 10G/25G/40G/50G/100G/200G/400G data rates
- **MCMR FEC** (Forward Error Correction) IP supports Ethernet standard with up to 400G data rates



Memory Interface IP Subsystem - HBM2 PHY + Controller



HBM2 ASIC SiP Validation Board

- Silicon-Proven in TSMC 16nm
- TSMC's OIP Ecosystem Forum 2017 Customers' Choice Award for best paper "HBM2"
- Total bandwidth: >300GB/s
- Data transfer: 1.6-2.4 Gb/s
- SiP TSMC CoWoS (2.5D)
- Interposer TSMC 65nm
- Interposer trace length < 5mm
- Supports all HBM stack vendors
- *IP available now*
- *Several licensees to date*



Next-Generation HBM2E IP Subsystem

AXI-based or TileLink-based HBM2E IP subsystem development

Targeting 3.2 Gbps per-pin data rates, and beyond, in TSMC's latest 7nm FinFET technologies

Supports up to 3.2 Gbps/pin data rates and beyond

Supports up to 8 channels (16 pseudo channels)

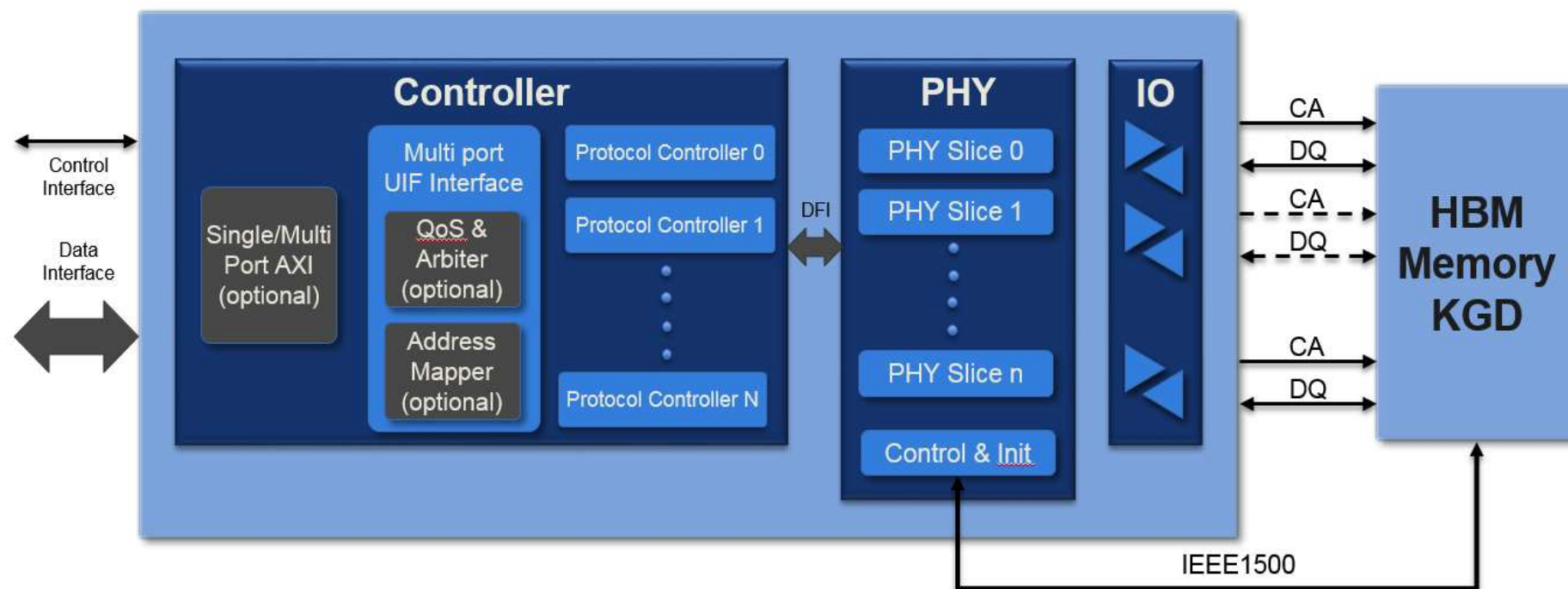
Supports up to > 400GBytes of total bandwidth

Supports full DFI4.0 compliant controller and PHY interface

Supports multi-port AXI interface or TileLink

Supports different schemes of arbitration and scheduling (QoS)

Supports different address mapping modes





SiFive for IP Components, or Entire Custom AI SoC

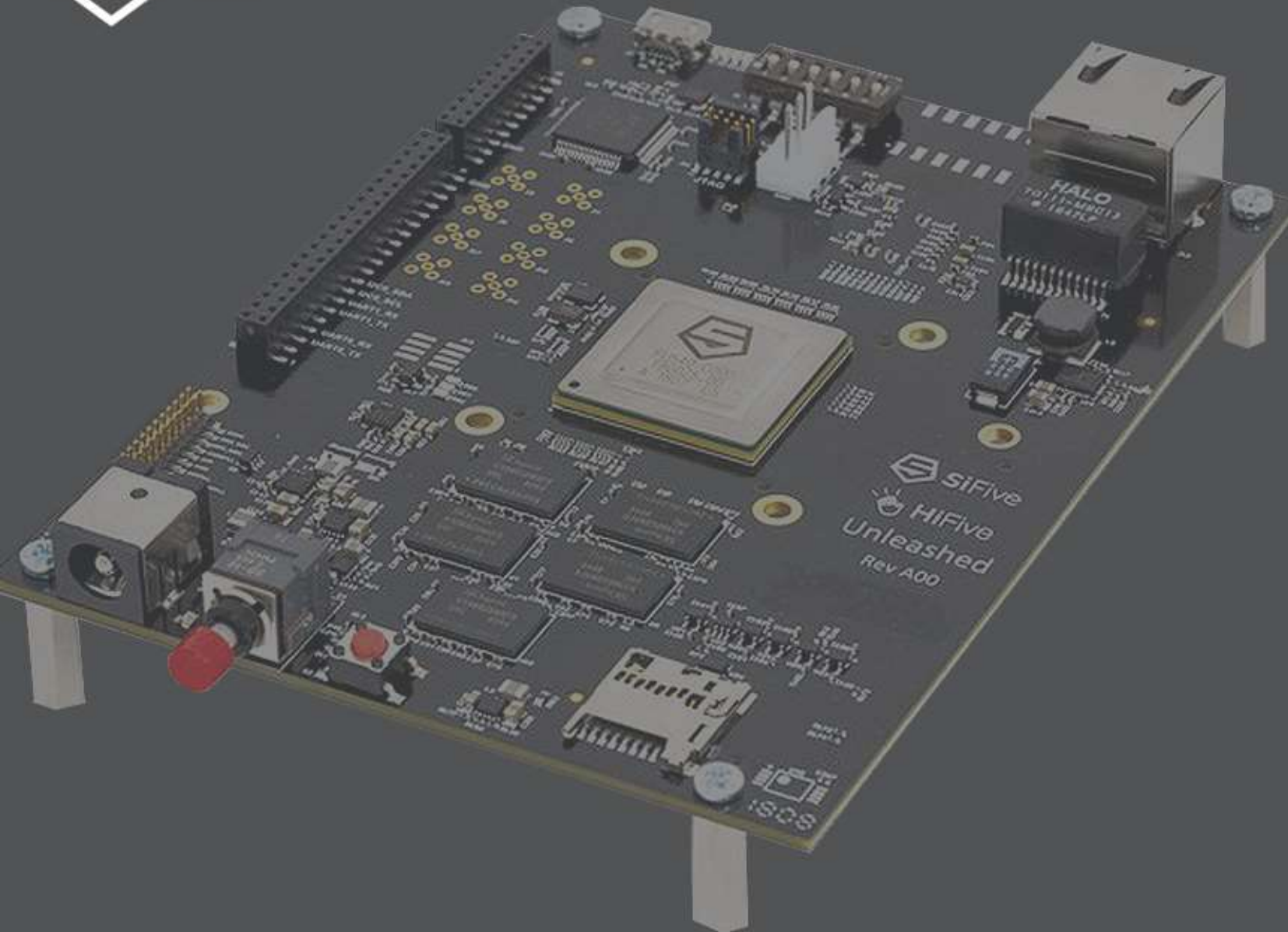
Available:

- RISC-V U7 and S7 series Core IP
- TileLink on-chip coherent fabric and peripherals
- TileLink chip-chip coherent serialization
- 2.4 Gb/s HBM2 IP Subsystem (phy and controller)
- 1.2 Tb/s Interlaken IP Subsystem
- 40+Gb/s Ethernet IP Subsystem

In development:

- E7/S7/U7 vector and custom extensions
- Accelerator Bays
- 3.2 Gb/s HBM2E IP Subsystem (phy and controller)
- Higher-performance RISC-V processors

Please come talk to us for your RISC-V AI chip needs!



Silicon verified. Market proven.

The most advanced configurable core IP and silicon solutions from the inventors of RISC-V.

Microcontrollers ■ Embedded ■ Linux ■ Multicore

■ Networking ■ Storage ■ Computing ■ AI
■ Industrial ■ IoT ■ Consumer ■ Automotive

www.sifive.com