

Развитие экосистемы RISC-V

Валерия Пузикова,
к.ф.-м.н., эксперт по разработке ПО





Валерия Пузикова

К.ф.-м.н., эксперт по разработке ПО,
руководитель команды разработки библиотек
вычислений и обработки данных, YADRO

- С 2010 года разрабатываю и реализую на C/C++ с CUDA/MPI/OpenMP численные методы для решения задач линейной алгебры, вычислительной аэрогидродинамики, AR/VR
- Работала в Huawei, Fortum, ИСП РАН им. В.П. Иванникова, МГТУ им. Н.Э. Баумана и др.

Проясним «базу»

Преимущества, тренды, прогнозы

Программная экосистема

Выводы и полезные ссылки



CISC vs RISC

Complex Instruction Set Computer	vs	Reduced Instruction Set Computer
x86: Intel, AMD	Где?	IBM Power PC; Arm (Apple M1, NVIDIA, Fujitsu A64FX и т.д.)
Несколько (например, есть операции вида Read-Modify-Write)	Действий за инструкцию	Одно (load/store четко отделены от остальных инструкций — load-and-store-архитектура)
Зависит от инструкции	Длина инструкции	Фиксированная



RISC: откуда плюсы?

Инструкции проще, в них нет большого количества разных режимов адресации.

В командах остается больше бит, чтобы указать номер **регистра** — регистров может быть больше. Например, если для указания регистра есть только 3 бита, то может быть закодировано $2^3 = 8$ регистров; а если для указания регистра остается 5 бит, то может быть закодировано уже $2^5 = 32$ регистра.



RISC: откуда плюсы?

Инструкции проще, в них нет большого количества разных режимов адресации.

В командах остается больше бит, чтобы указать номер **регистра** — регистров может быть больше. Например, если для указания регистра есть только 3 бита, то может быть закодировано $2^3 = 8$ регистров; а если для указания регистра остается 5 бит, то может быть закодировано уже $2^5 = 32$ регистра.

Процессоры проще, для их реализации требуется меньше транзисторов.

- Это позволяет сильно снизить **энергопотребление**.
- Это освобождает больше кремниевой площади, например, для **кешей**.



RISC: откуда плюсы?

Инструкции проще, в них нет большого количества разных режимов адресации.

В командах остается больше бит, чтобы указать номер **регистра** — регистров может быть больше. Например, если для указания регистра есть только 3 бита, то может быть закодировано $2^3 = 8$ регистров; а если для указания регистра остается 5 бит, то может быть закодировано уже $2^5 = 32$ регистра.

Процессоры проще, для их реализации требуется меньше транзисторов.

- Это позволяет сильно снизить **энергопотребление**.
- Это освобождает больше кремниевой площади, например, для **кешей**.

Инструкции состоят из секций и имеют фиксированную одинаковую длину.

Этапы fetch — decode — execute — memory — write чётко разделены, fetch знает, где будет следующая инструкция без декодирования. Это упрощает **конвейеризацию** инструкций.



2010 г.: 5-е поколение — открытая и свободная модель (open-source)

Нет лицензионных сборов — любой может использовать её для RnD.

- «Твори, выдумывай, пробуй!»
- Стремительно растёт сообщество и техподдержка.
- Снижаются расходы на RnD, срок вывода дизайна ip на рынок.

Немного статистики из мира RISC-V



More than 4,100 RISC-V Members across 70 Countries



Dec 2023 update

121 Chip

SoC, IP, FPGA

4 Systems

ODM, OEM

3 I/O

Memory, network, storage

14 Industry

Cloud, mobile, HPC, ML, automotive

23 Services

Fab, design services

165 Research

Universities, Labs, other alliances

59 Software

Dev tools, firmware, OS

3k+ Individuals

RISC-V engineers and advocates

RISC-V membership up 28% in 2023



Основные цели Консорциума RISC-V в 2025 году

AI/ML второй год подряд входит в число главных приоритетов работы Консорциума.

- Нарастает темп разработки матричных расширений.
- Развивается векторное расширение.
- Добавляется поддержка типов данных пониженной точности.

Основные цели Консорциума RISC-V в 2025 году

AI/ML второй год подряд входит в число главных приоритетов работы Консорциума.

- Нарастает темп разработки матричных расширений.
- Развивается векторное расширение.
- Добавляется поддержка типов данных повышенной точности.

Ускорение основных хот-спотов AI/ML:

- умножения плотных матриц;
- умножения разреженной матрицы на плотную;
- матрично-векторные произведения
- и т.д.



Основные цели Консорциума RISC-V в 2025 году

AI/ML второй год подряд входит в число главных приоритетов работы Консорциума.

- Нарастает темп разработки матричных расширений.
- Развивается векторное расширение.
- Добавляется поддержка типов данных пониженной точности

Аппартная поддержка квантизации
моделей ИИ

Проясним «базу»

Преимущества, тренды, прогнозы

Программная экосистема

Выводы и полезные ссылки

Как снизить затраты на AI/ML?



Ноу-хау в архитектуре нейронной сети

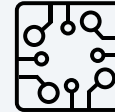
- Глубокие знания
- Время
- Инвестиции в RnD

Как снизить затраты на AI/ML?



Ноу-хау в архитектуре нейронной сети

- Глубокие знания
- Время
- Инвестиции в RnD



Энергоэффективные вычислительные системы

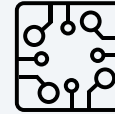
CPU + NPU / GPU

Как снизить затраты на AI/ML?



Ноу-хау в архитектуре нейронной сети

- Глубокие знания
- Время
- Инвестиции в RnD



Энергоэффективные вычислительные системы

CPU + NPU / GPU



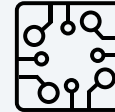
упрёмся в пропускную
способность шины

Как снизить затраты на AI/ML?



Ноу-хау в архитектуре нейронной сети

- Глубокие знания
- Время
- Инвестиции в RnD



Энергоэффективные вычислительные системы

CPU + NPU / GPU



упрёмся в пропускную
способность шины

Как уйти от гетерогенности?

RISC-V: модульная архитектура

Легковесная базовая ISA

(RV32I / RV64I)



Расширения

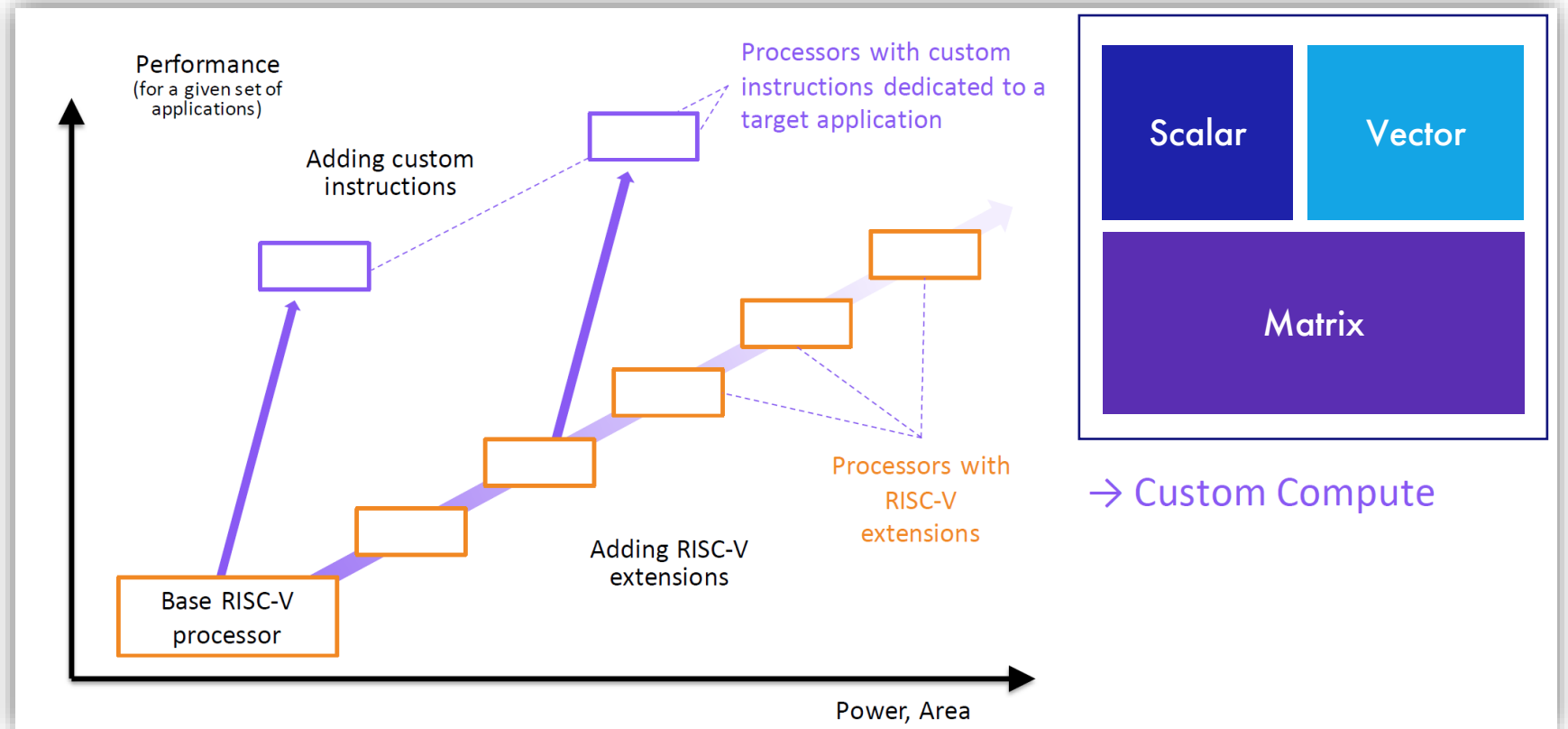
(M — mult., F — float, V — vector, B — bit, A — atomic и т. д.)

- Система команд имеет зарезервированные в спецификации биты для кодирования расширений.
- Расширения могут быть стандартизированные или кастомные (пользовательские).



Плюсы модульной архитектуры RISC-V

- Можно создать гетерогенную вычислительную среду с ядрами, оптимизированными для различных типов вычислений (CPU + Vector/Matrix engines + DSP, VLIW + GPU + ...).
- От встроенных систем до суперкомпьютеров.



Пример: Semidynamics для Big Data & AI/ML

Our vision: Fusing CPU, GPU, and NPU

Each Vector Unit : from 4 to 32 FMAC units

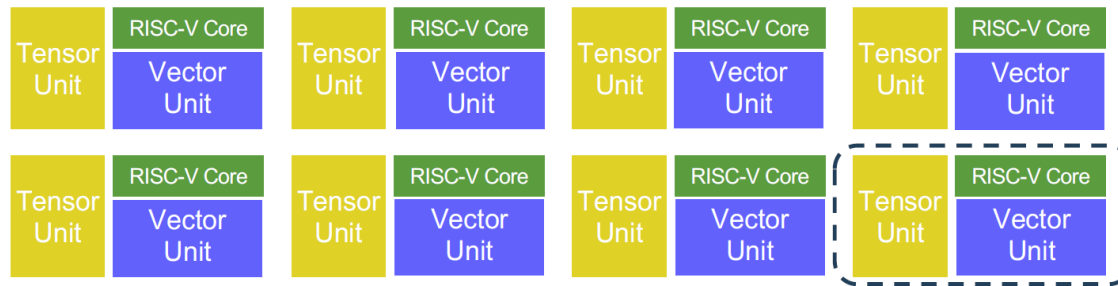
Each Tensor Unit : from $\frac{1}{4}$ TOPS₈ to 2 TOPS₈

Gazzillion Technology to enable sustained DRAM access beyond 50 bytes/cycle

Other circuits/
Interface subsystem

PCIe

Optimized Cache

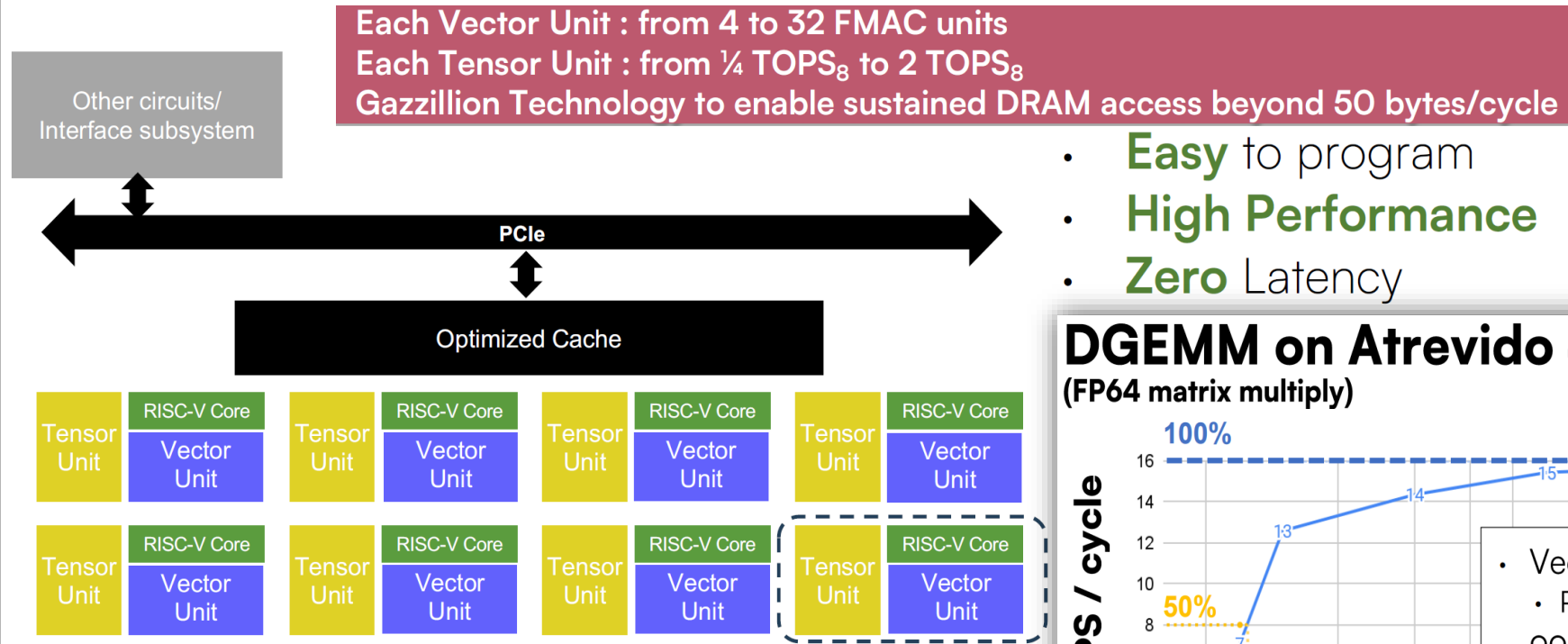


- **Easy** to program
- **High Performance**
- **Zero** Latency
- Better **PPA**
- **Resilient** to new AI
- **Unified** programming

All-in-One
AI IP Element

Пример: Semidynamics для Big Data & AI/ML

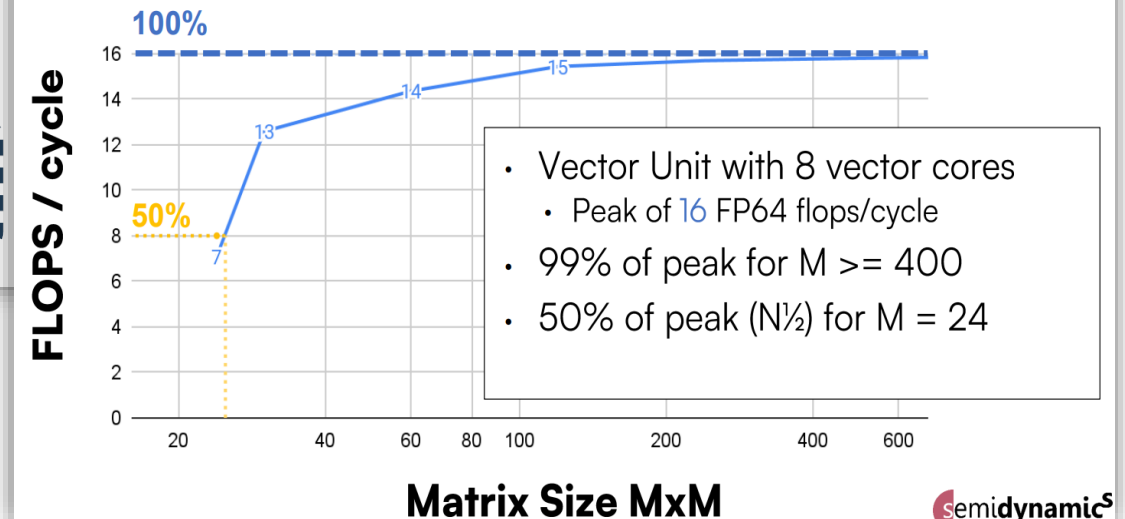
Our vision: Fusing CPU, GPU, and NPU



- **Easy** to program
- **High Performance**
- **Zero** Latency

DGEMM on Atrevido 423 + V8

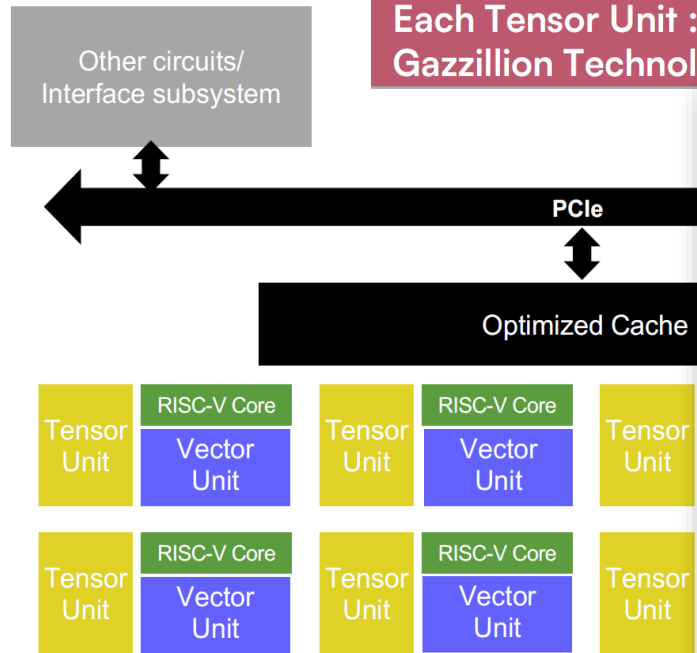
(FP64 matrix multiply)



Пример: Semidynamics для Big Data & AI/ML

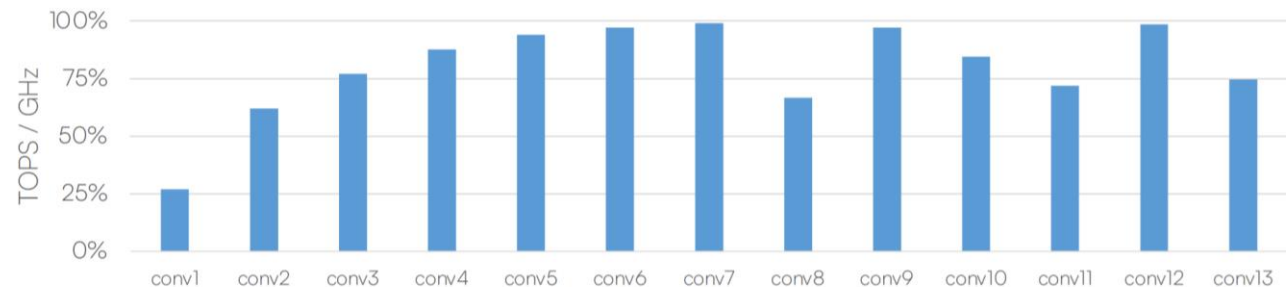
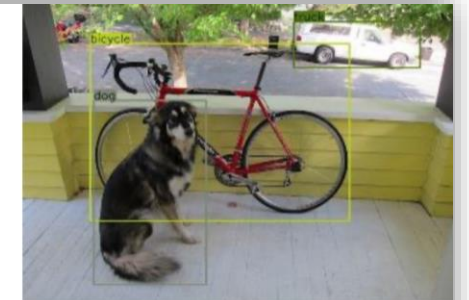
Our vision: Fusing CPU, GPU, and NPU

Each Vector Unit : from 4 to 32 FMAC units
 Each Tensor Unit : from $\frac{1}{4}$ TOPS₈ to 2 TOPS₈
 Gazzillion Technology to enable sustained DRAM access beyond 50 bytes/cycle



YOLO on our fused IP: 33 FPS

- Performance at 1GHz
 - ATV4+Vector Unit + Tensor Unit (bf16): 33.03 FPS
 - Real-time performance with one Tensor Unit



Convolutional layer in YOLOv3-tiny



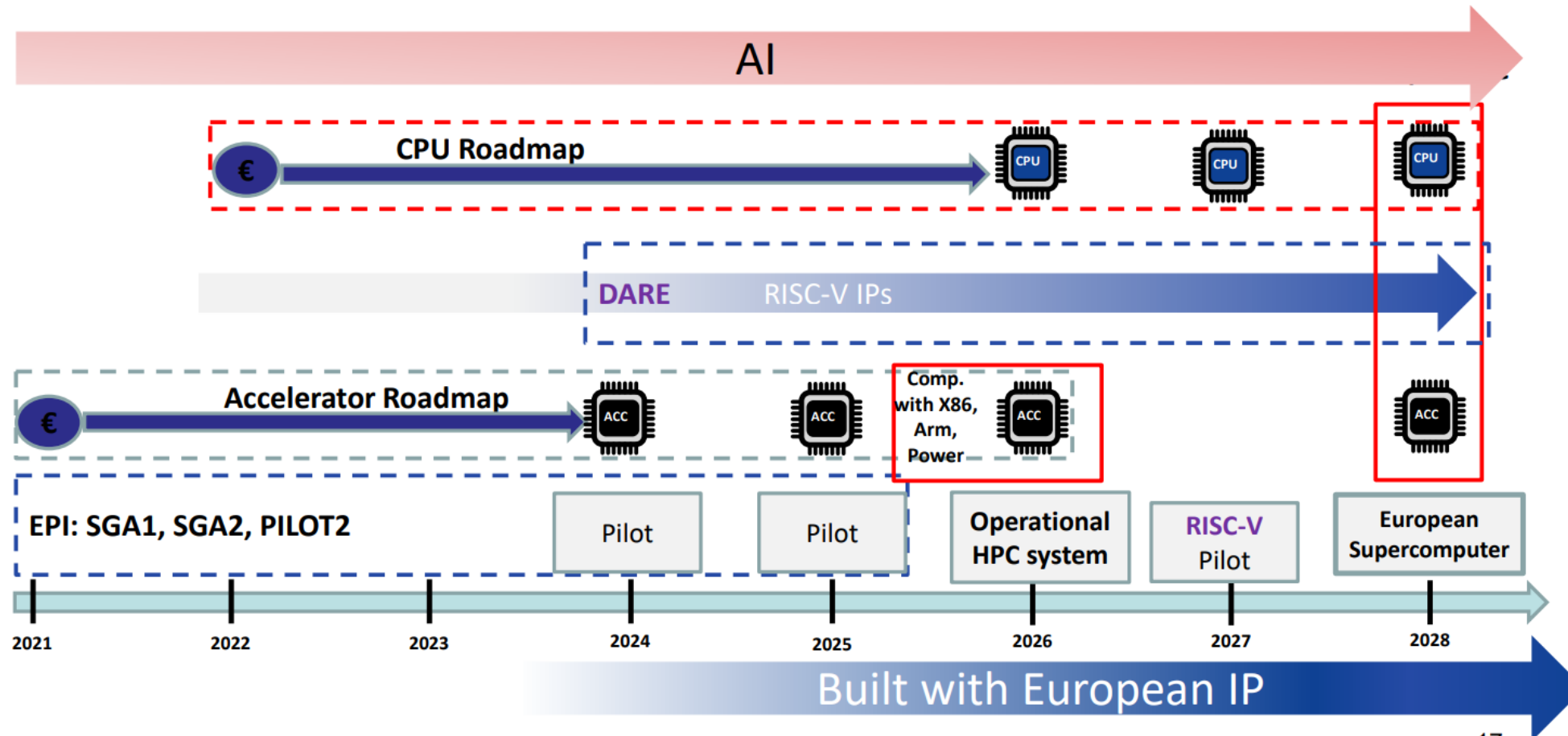
Тренды в развитии RISC-V HardWare

- Появляются **высокопроизводительные ядра** (процессоры для дата-центров):
 - ET-SoC-1: Esperanto's RISC-V Supercomputer-on-a-Chip (1088 ядер, 158 ResNet50 Int8 Inf/Sec/Watt)
 - Thunderbird RISC-V Supercomputer Cluster-on-a-Chip (1536 ядер, FP64 24 TFLOPS, 75 GFLOPS/Watt)
 - Ventana Veyron, Milk-V Pioneer
- На базе архитектуры RISC-V разрабатываются не только CPU, но и **ускорители**.
 - **AI:** Axelera AI Metis, Andes AX25-V100 (+Andes AIRE AI Software Stack)
 - **Vector:** EPAC-VEC VPU "Vitrius" (256 FP64/регистр)
 - **GPU:** OpenCL Compatible RISC-V GPGPU Vortex

EuroHPC Chips Roadmap



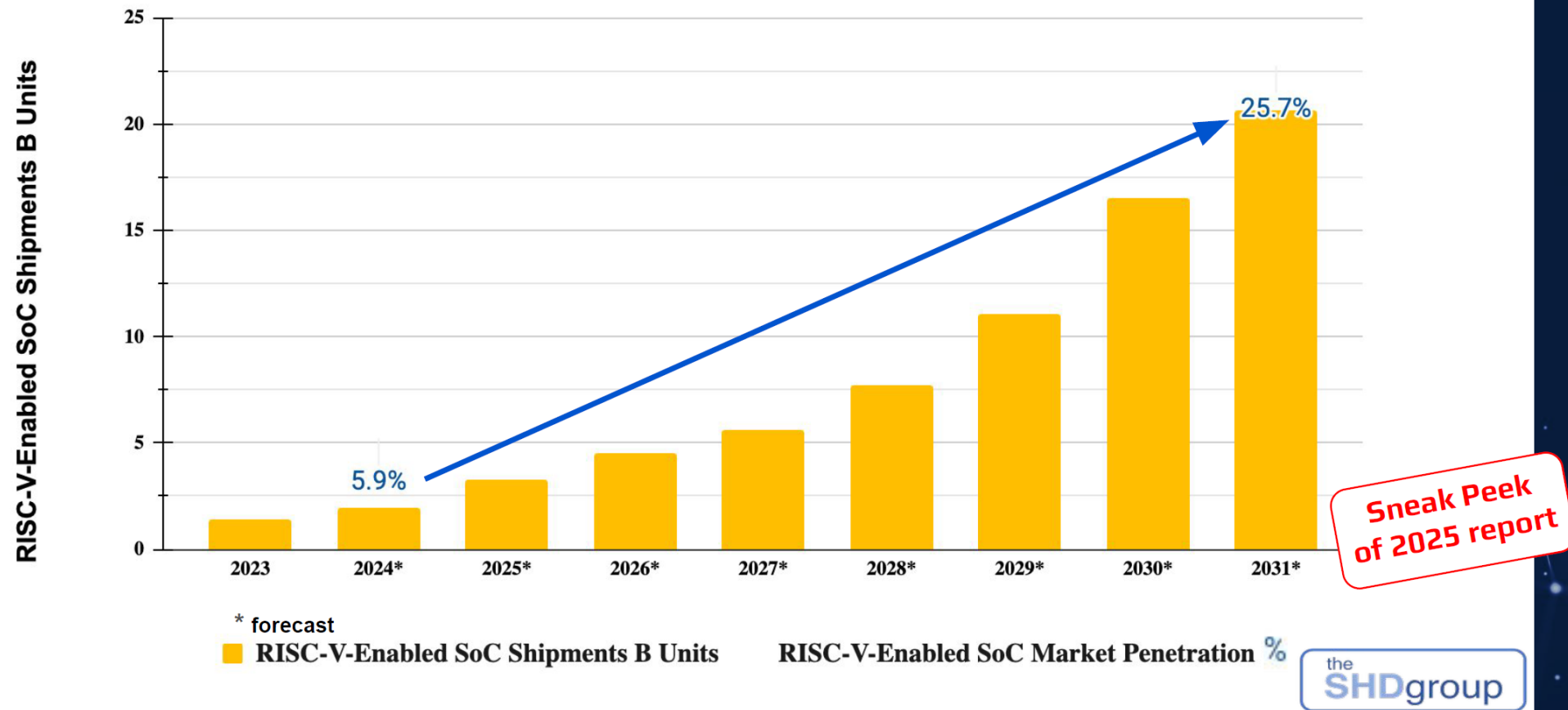
EuroHPC
Joint Undertaking



* Источник: <https://www.hipeac.net/2025/barcelona/#/program/sessions/8179/>

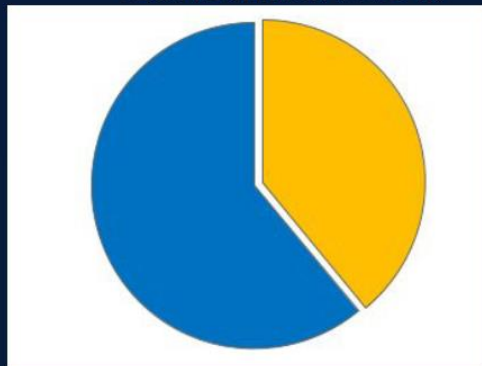
Прогнозируемый рост числа RISC-V-процессоров

20B RISC-V SoCs, to Surpass 25% of Market

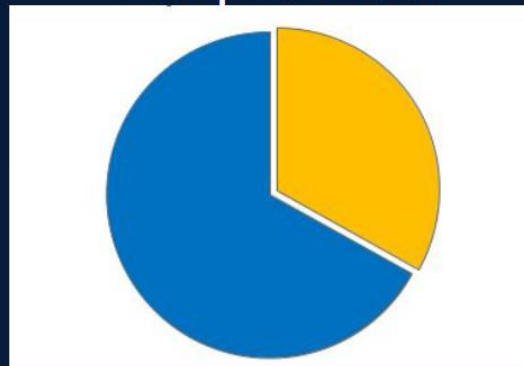


Top Markets for RISC-V 2031

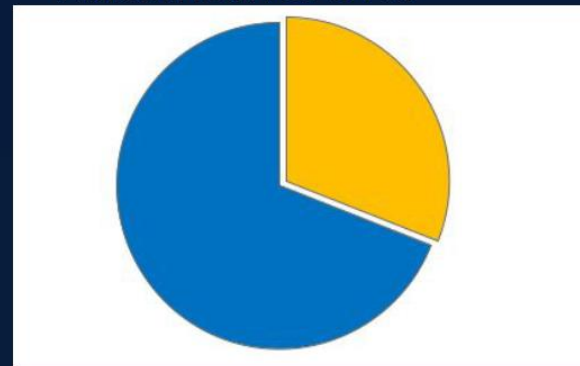
Consumer: 39%



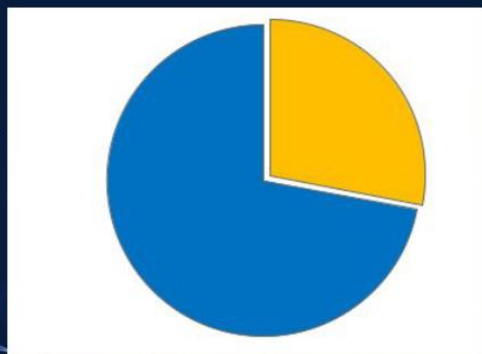
Computer: 33%



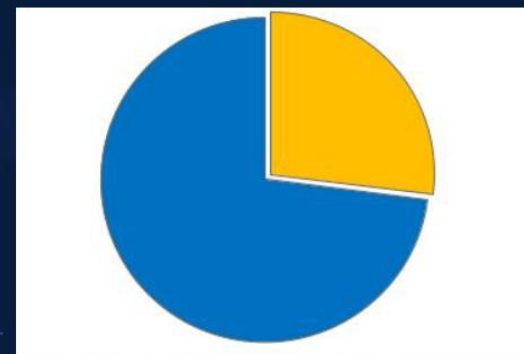
Automotive: 31%



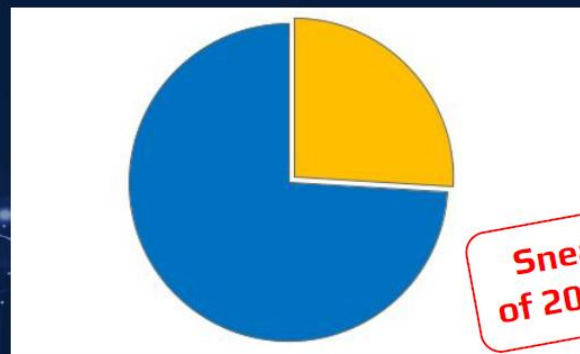
Data Center: 28%



Industrial: 27%



Networking: 26%



Sneak Peek
of 2025 report

Source: The SHD Group October 2024

■ RISC-V

the
SHDgroup

Проясним «базу»

Преимущества, тренды, прогнозы

Программная экосистема

Выводы и полезные ссылки

К чему мы привыкли?

Applications	Computer Vision	Speech	Natural Language Processing	Autonomous Systems	Recommendations	Finance
Models	ResNet VGGNet YOLO	HMM LSTM	GPT BERT	SLAM ControlNet	Content Filter Gradient Boosted	ARIMA Monte Carlo
Frameworks	 TensorFlow Lite		 TensorFlow	 PyTorch	 ONNX	

И все это быстро работает....

К чему мы привыкли?

Applications	Computer Vision	Speech	Natural Language Processing	Autonomous Systems	Recommendations	Finance
Models	ResNet VGGNet YOLO	HMM LSTM	GPT BERT	SLAM ControlNet	Content Filter Gradient Boosted	ARIMA Monte Carlo
Frameworks	 TensorFlow Lite		 TensorFlow	 PyTorch		 ONNX

И все это быстро работает.... На чем?

Platform	
----------	---

К чему мы привыкли?

Applications	Computer Vision	Speech	Natural Language Processing	Autonomous Systems	Recommendations	Finance
Models	ResNet VGGNet YOLO	HMM LSTM	GPT BERT	SLAM ControlNet	Content Filter Gradient Boosted	ARIMA Monte Carlo
Frameworks						

И все это быстро работает.... На чем?

X86 – Server CPU – Intel,

Arm – Server GPU – NVIDIA

Arm – Mobile – Apple/{Android}

Platform	
----------	---

К чему мы привыкли?

Applications	Computer Vision	Speech	Natural Language Processing	Autonomous Systems	Recommendations	Finance
Models	ResNet VGGNet YOLO	HMM LSTM	GPT BERT	SLAM ControlNet	Content Filter Gradient Boosted	ARIMA Monte Carlo
Frameworks	 TensorFlow Lite		 TensorFlow	 PyTorch		 ONNX

И все это быстро работает.... На чем?

Почему мы к этому привыкли?

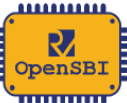
X86 – Server CPU – Intel,

Arm – Server GPU – NVIDIA

Arm – Mobile – Apple/{Android}

Platform	
----------	---

Как выглядит наш программный стек?

Applications	Computer Vision	Speech	Natural Language Processing	Autonomous Systems	Recommendations	Finance	Tools
Models	ResNet VGGNet YOLO	HMM LSTM	GPT BERT	SLAM ControlNet	Content Filter Gradient Boosted	ARIMA Monte Carlo	
Frameworks	 TensorFlow Lite	 TensorFlow	 PyTorch		 ONNX		
Runtimes	TFRT		 GLOW		 ONNX RUNTIME		
Math Backend	Eigen	OpenBLAS	XNNPACK	AMD ROCm	 NVIDIA CUDA		
Hypervisors Containers	 KVM		 docker		 kubernetes		
Operating System	 Linux		 ubuntu®		 fedora		
Firmware Early Boot, BIOS	 OpenSBI		 tianocore 		ACPI		
Platform							

 python


 LLVM
COMPILED INFRASTRUCTURE
 MLIR
 glibc
 GDB
The GNU Project
Debugger
 TRACE32
 OpenOCD

Как выглядит наш программный стек?

Applications	Computer Vision	Speech	Natural Language Processing	Autonomous Systems	Recommendations	Finance	Tools	
Models	ResNet VGGNet YOLO	HMM LSTM	GPT BERT	SLAM ControlNet	Content Filter Gradient Boosted	ARIMA Monte Carlo		
Frameworks								
Runtimes	TFRT							
Math Backend	Eigen	OpenBLAS	XNNPACK	AMD ROCm	NVIDIA CUDA			
Hypervisors Containers								
Operating System								
Firmware Early Boot, BIOS								
Platform								

А если «на дне» что-то непривычное?

Applications	Computer Vision	Speech	Natural Language Processing	Autonomous Systems	Recommendations	Finance
Models	ResNet VGGNet YOLO	HMM LSTM	GPT BERT	SLAM ControlNet	Content Filter Gradient Boosted	ARIMA Monte Carlo
Frameworks	 TensorFlow Lite		 TensorFlow	 PyTorch		 ONNX

RISC-V – CPU / GPU / NPU – да, они уже появляются

Platform	
----------	---

А если «на дне» что-то непривычное?

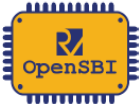
Applications	Computer Vision	Speech	Natural Language Processing	Autonomous Systems	Recommendations	Finance
Models	ResNet VGGNet YOLO	HMM LSTM	GPT BERT	SLAM ControlNet	Content Filter Gradient Boosted	ARIMA Monte Carlo
Frameworks						

1. Оно вообще запустится без переделок?
2. И что же будет с производительностью?

RISC-V – CPU / GPU / NPU – да, они уже появляются

Platform	
----------	---

Программную экосистему портируют уже сейчас

Applications	Computer Vision	Speech	Natural Language Processing	Autonomous Systems	Recommendations	Finance
Models	ResNet VGGNet YOLO	HMM LSTM	GPT BERT	SLAM ControlNet	Content Filter Gradient Boosted	ARIMA Monte Carlo
Frameworks	 TensorFlow Lite	 TensorFlow	 PyTorch		 ONNX	
Runtimes	TFRT		 GLOW		 ONNX RUNTIME	
Math Backend	Eigen	OpenBLAS	XNNPACK	AMD ROCm	 NVIDIA CUDA	
Hypervisors Containers	 KVM		 docker		 kubernetes	
Operating System	 Linux		 ubuntu®		 fedora	
Firmware Early Boot, BIOS	 OpenSBI		 tianocore		 UEFI ACPI	
Platform	RISC-V 					

Tools

 python
 C++
 GCC
 LLVM
 MLIR
 glibc
 GDB
 TRACE32
 OpenOCD

Консорциум RISE (RISC-V Software Ecosystem)



Founding Member of RISE to Ensure RISC-V Software Readiness



Mission

- Accelerate the development of open source software for RISC-V
- Raise the quality of RISC-V Platform software implementations
- Push the RISC-V Software ecosystem forward and align partners' efforts
- Ensure RISC-V is a tier 1 platform for all tools and libraries
- Accelerate RISC-V adoption for Client and above segments



Фокусируется на адаптации ключевого стека ПО для RISC-V, а также ускорении разработки СПО для RISC-V.

RISE RISC-V Optimization Guide

Vendor agnostic porting and optimization guide

- Does not cover CPU specific microarchitecture

Best practices for high performance RISC-V cores

- Including assembly code examples

Zero can be folded into any instruction with a register operand. There's no need to initialize a temporary register with 0 for the sole purpose of using that register in a subsequent instruction. The following table identifies cases where a temporary register can be eliminated by prudent use of x0.

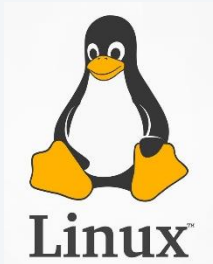
Do	Don't
<code>fmv.d.x f0,x0</code>	<code>li x5,0</code> <code>fmv.d.x f0,x5</code>
<code>amoswap.w.aqr1 a0,x0,(x10)</code>	<code>li x5,0</code> <code>amoswap.w.aqr1 x6,x5,(x10)</code>
<code>sb x0,0(x5)</code>	<code>li x6,0</code> <code>sb x6,0(x5)</code>
<code>bltu x0,x7,1f</code>	<code>li x5,0</code> <code>bltu x5,x7,1f</code>





Компоненты экосистемы ПО с RVV-оптимизациями

OS



Tools



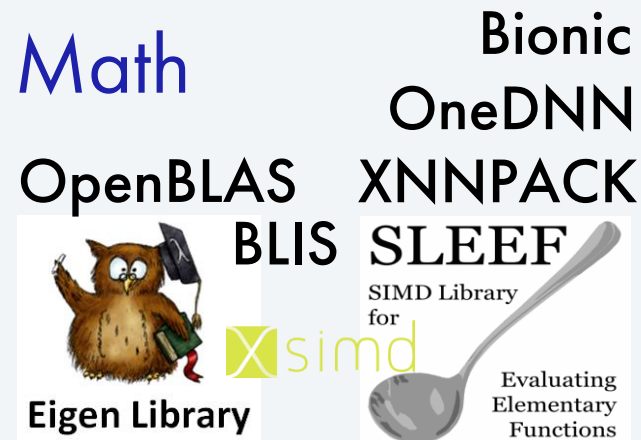
Web



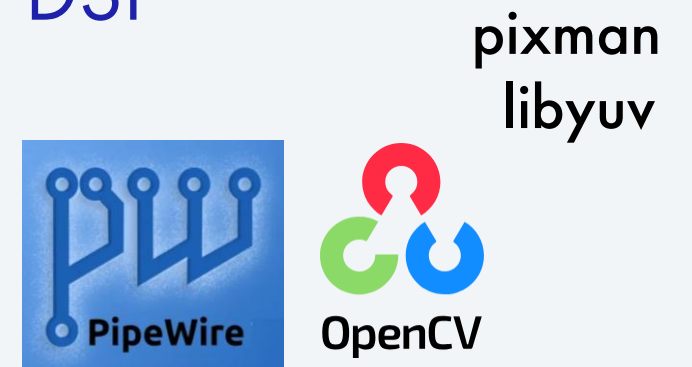
Data



Math



DSP



Проясним «базу»

Преимущества, тренды, прогнозы

Программная экосистема

Выводы и полезные ссылки

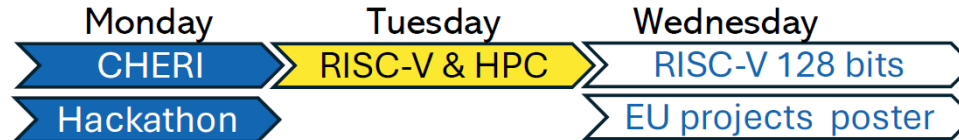
Как отслеживать текущее состояние?



С 2023 г.: HPC RISC-V-воркшопы



RISC-V is tangible!



Как отслеживать текущее состояние?



С 2023 г.: HPC RISC-V-воркшопы



- Портирование стека ПО: очередность, фокус, оптимизации.
- Советы/наработки по оптимизациям: что удалось в плане оптимизации основных вычислительных микрокernels, импакт на фреймворки, бенчмарки
- Новинки и анонсы процессоров, бенчмарки.



Выводы

- **RISC-V быстро развивается:** новые расширения ISA, процессоры серверного класса, ускорители, интерконнекты.
- **Гетерогенность в RISC-V реализуется в рамках одного процессора:** данные не надо пересылать на внешние ускорители и видеокарты.
- **Портирование программной экосистемы идет уже сейчас,** включая низкоуровневые оптимизации вычислительных ядер.

SIG-HPC Initiatives

- Guide and enable the community
 - Virtual Memory
 - SV57, SV57K, SV64, SV128
 - Accelerators
 - ISA Extensions
 - HPC Software Stack
 - Starting with HPC Libraries
 - HPC SW & HW ecosystem & roadmap

(!) Обновления на [Github](#):

- [HPC SIG](#)
- [AI/ML & Graphics SIG](#)
- [Vector SIG](#)





Выводы

- **RISC-V быстро развивается:** новые расширения ISA, процессоры серверного класса, ускорители, интерконнекты.
- **Гетерогенность в RISC-V реализуется в рамках одного процессора:** данные не надо пересылать на внешние ускорители и видеокарты.
- **Портирование программной экосистемы идет уже сейчас,** включая низкоуровневые оптимизации вычислительных ядер.
- **Эксперты прогнозируют, что в течение 5 лет системы на базе RISC-V возглавят суперкомпьютерный рейтинг Green500 и займут центральное место в сегменте AI/ML.**

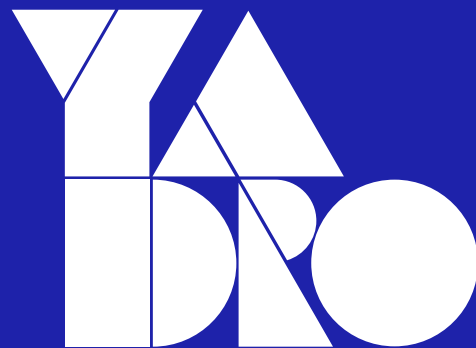
SIG-HPC Initiatives

- Guide and enable the community
 - Virtual Memory
 - SV57, SV57K, SV64, SV128
 - Accelerators
 - ISA Extensions
 - HPC Software Stack
 - Starting with HPC Libraries
 - HPC SW & HW ecosystem & roadmap

(!) Обновления на [Github](#):

- [HPC SIG](#)
- [AI/ML & Graphics SIG](#)
- [Vector SIG](#)





Москва,
ул. Рочдельская, 15, стр. 13
+7 800 777-06-11

yadro.com