

RISC-V на международных суперкомпьютерных конференциях: доклады, новинки, тренды

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





Валерия Пузикова

К.ф.-м.н., эксперт по разработке ПО,
руководитель команды разработки
математических библиотек, 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.

HPC на RISC-V: почему уже пора?

HPC в мировых трендах экосистемы RISC-V

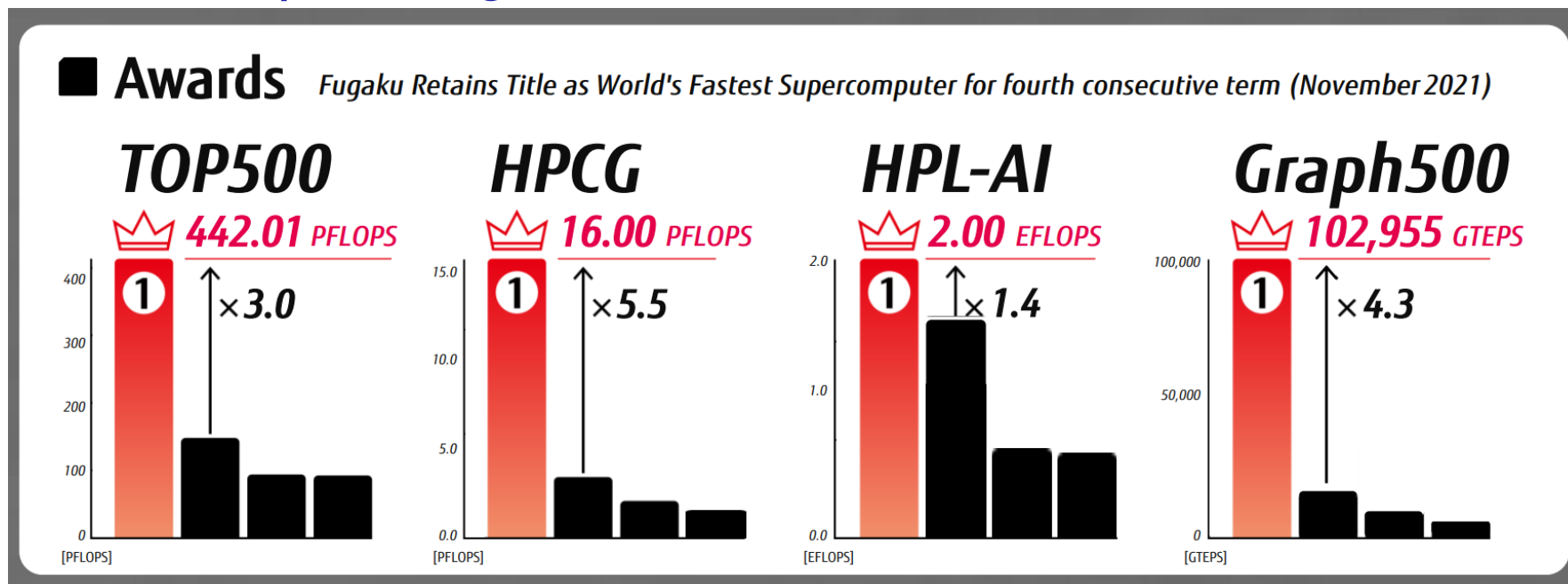
Примеры докладов с HPC RISC-V воркшопов 2024

RISC-V HW: на чем тестировать HPC SW уже сейчас и чего ждать

Выводы и полезные ссылки



HPC на ARM: Fujitsu Fugaku №1 – с 2020 года



- **Впервые** суперкомпьютер на ARM (причем гомогенный) стал №1 в Top500.
- **Единственный в истории** стал №1 **во всех** основных суперкомпьютерных рейтингах
- **До сих пор №1** в HPCG, HPL-AI, Graph500.
- **Мировой рекорд:** число ядер увеличили на 4,5%, а производительность на Linpack выросла на 6,4%, на HPCG – в 5,4 раза.
- **На 45% превосходит** производительность всех остальных суперкомпьютеров из **Top10 HPCG**.

* Источник: http://www.storagenews.ru/76/Fujitsu_Fugaku-2.pdf



Fujitsu готовила SW экосистему с 2014 года

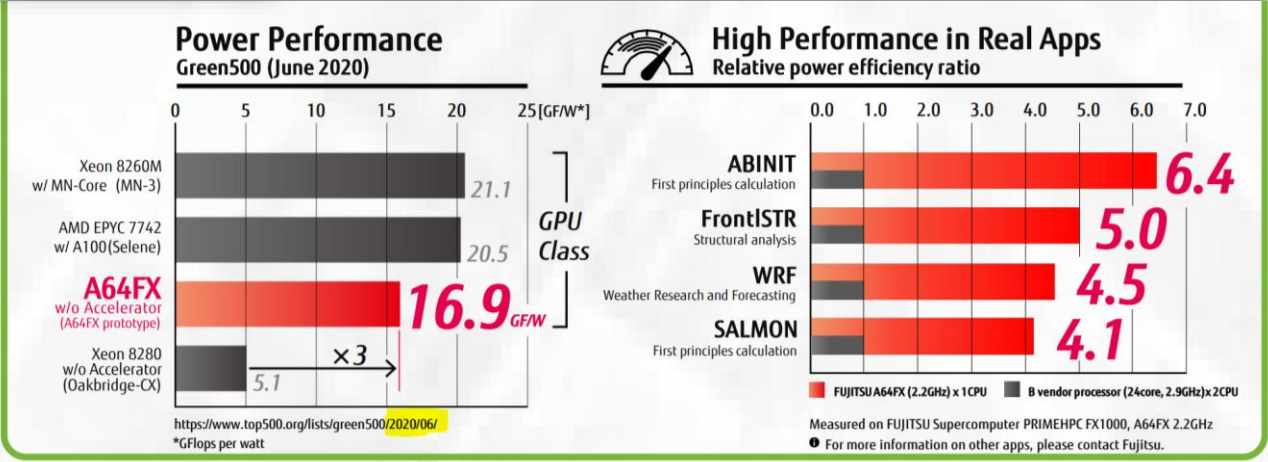
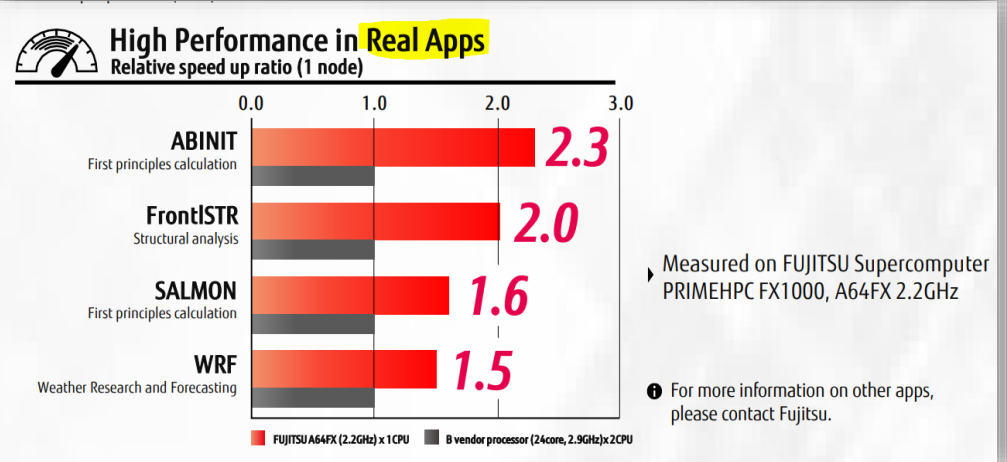
Мораль: развитие SW экосистемы HW – дело важное и долгое!

LS-DYNA (by Ansys, Inc.)	ADVENTURECluster (by Allied Engineering Co.)	Altair Radioss™ (by Altair Engineering, Inc.)	Ansys Fluent (by Ansys, Inc.)
Poynting (by Fujitsu Limited)	CONVERGE CFD SOFTWARE (by Convergent Science)	HELIX™ (by ENGYS Ltd. & VINAS Co., Ltd.)	JMAG™ Provided by J SOL Corporation (by J SOL Corporation)
Chemistry*	Marc (by MSC Software Ltd.)	scFLOW (by Software Cradle Co., Ltd.)	Simcenter STAR-CCM+ (by Siemens Industry Software Inc.)
Amber	VASP	VPS (PAM-CRASH) (by ESI Group)	
Gaussian16 (by Gaussian, Inc.)	*Collaboration with Australian National University		

**All application names used in this slide are trademarks or registered trademarks of their respective vendors.

Рис. 8. Совместные разработки Fujitsu с независимыми поставщиками прикладного ПО, которые могут выполняться на FX1000, FX700 и Fugaku.

Academia <ul style="list-style-type: none"> Nano-science Particle physics 	Government <ul style="list-style-type: none"> Long-range forecasting Disaster prevention
Oil and Gas <ul style="list-style-type: none"> Exploration and production Seismic analysis 	Manufacturing <ul style="list-style-type: none"> Structural Analysis Aerodynamics Computational fluid dynamics Crash test simulations



* Источник: http://www.storage-news.ru/76/Fujitsu_Fugaku-2.pdf



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

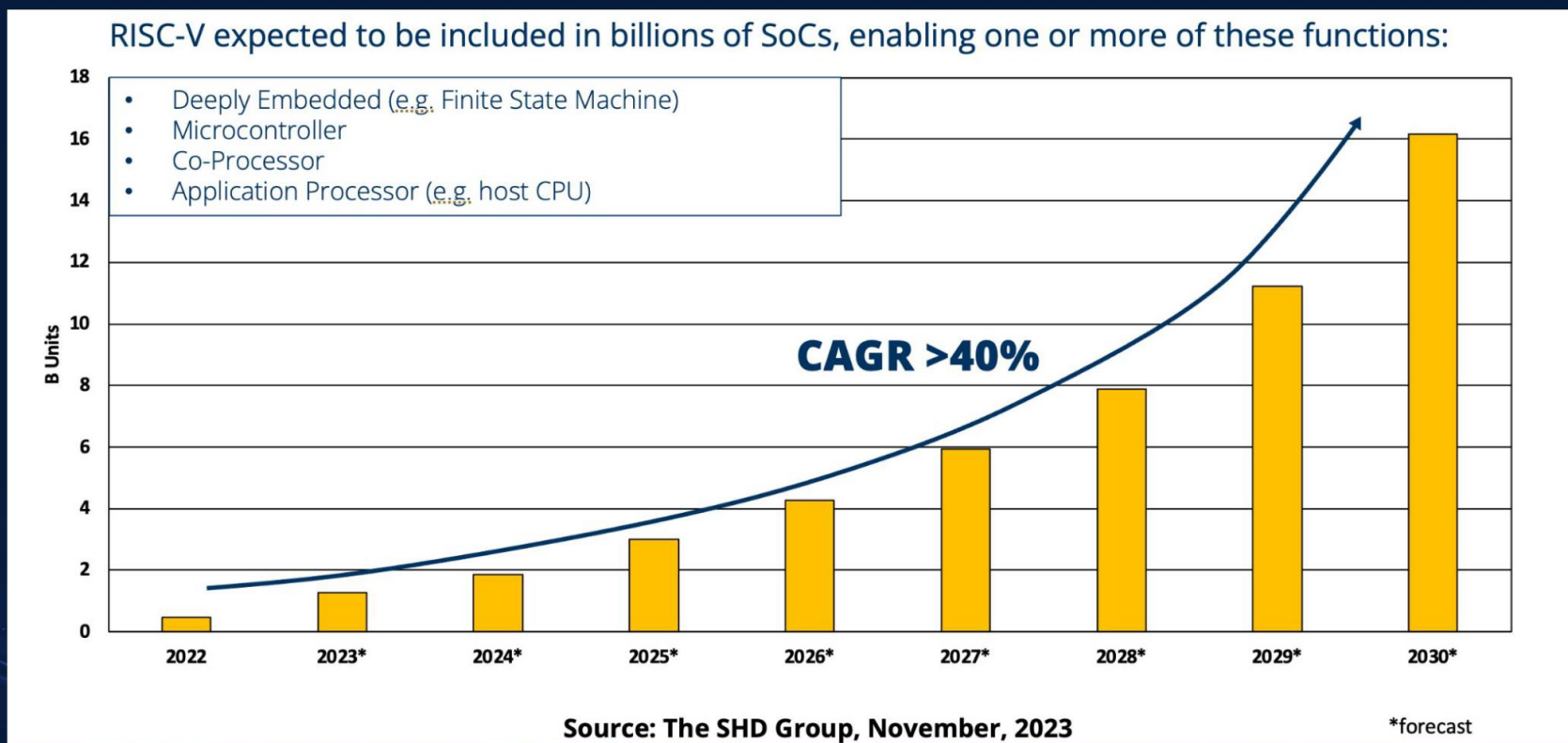
RISC-V membership up 28% in 2023





Прогнозируемый рост впечатляет

RISC-V will be in more than 16 billion SoCs by 2030

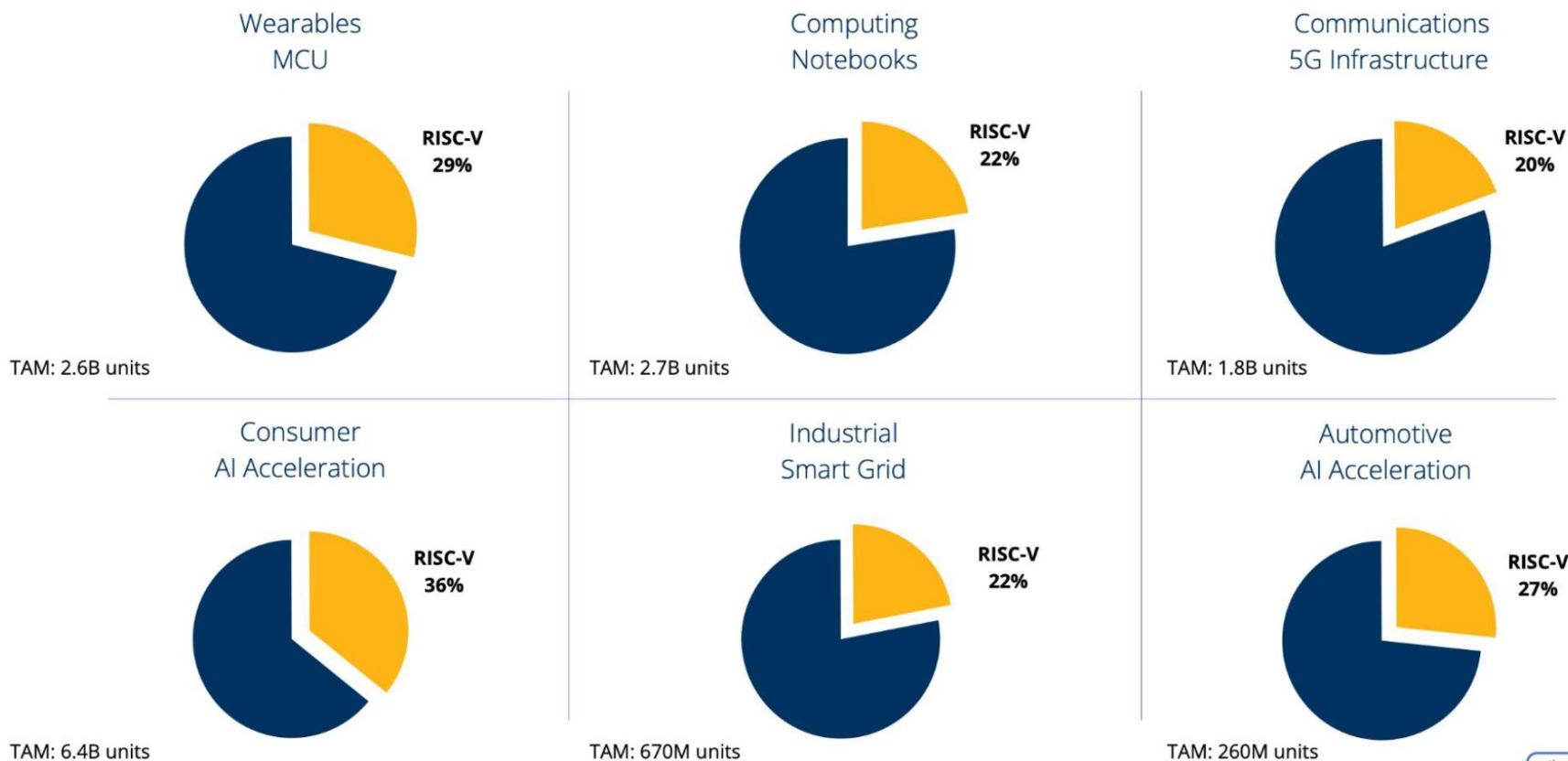


* **CAGR** – совокупный среднегодовой темп роста.



К 2030 году RISC-V займет не менее пятой части рынков

Selected Market Share Projections for RISC-V in 2030



Based on projected SoC volumes

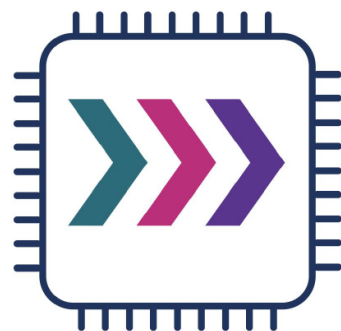
Source: The SHD Group, November, 2023



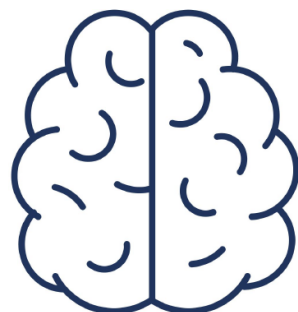


Ближайшие прогнозы по высокопроизводительным ядрам

Industry outlook: Datacenter & Cloud



RISC-V offers unique opportunity for accelerators



Custom computing for AI and other emerging workloads



Achieve your performance and power targets

RISC-V CPU core market will grow 115% CAGR, capturing >14% of all CPU cores by 2025

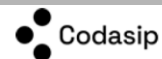
Semico Research, December 2021



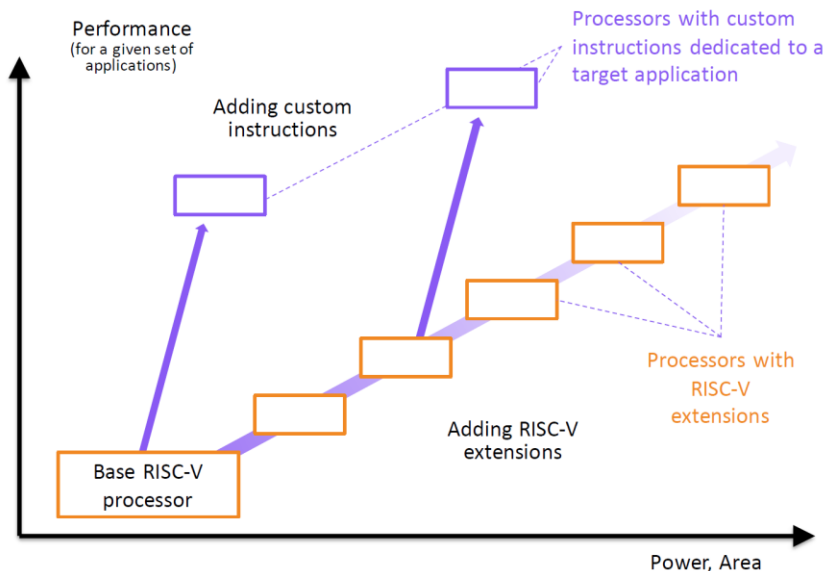


Плюсы RISC-V: модульная архитектура

→ RISC-V extensions give flexibility



→ RISC-V Custom Instructions enable efficiency

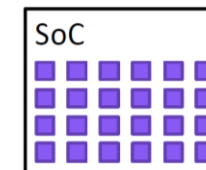


- RISC-V allows custom instructions
- Optimally designed for target application

→ Custom Compute

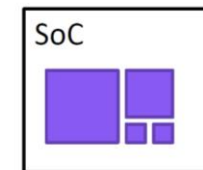
- RISC-V is modular
- Base is minimal
- Extensions target specific application

Specialized cores



- AI
- Systolic array engines
- In memory compute
- ...

Heterogeneous compute



- Optimized main CPU
- Vector/tensor engines
- DSP, VLIW
- GPU
- ...



Векторное расширение RISC-V RVV

Vector computing in HPC

ISA-defined max VL

intel	SSE	AVX2	AVX512
	128b	256b	512b

arm

NEON
128b

Implementation-defined max VL

NEC	Runtime Variable VL	RISC-V	arm
NEC-VE	Variable VL	RVV	SVE
16384b	[128b → *]	[128b → *]	[128b → 2048b]

64

Scalar processor

256

SIMD (e.g., AVX2)

256

128

Variable VL (e.g., RVV)
"Vector length agnostic"

HIPEAC Conference 2024, München, 17 Jan 2024

Vector Length Agnostic Code

- VL is loaded prior to executing the vector instruction with a special instruction
- No need to handle "loop tails"
- Makes the code "vector length agnostic"

```

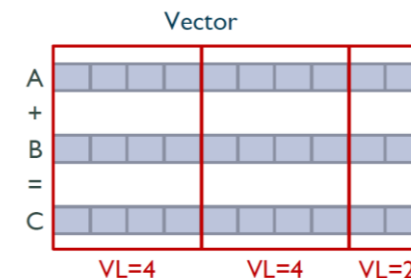
7 void axpy(double a, double *dx, double *dy, int n) {
8   int i;
9
10  long gvl = __builtin_eabi_vsetvl(n, __eabi_e64, __eabi_m1);
11  __eabi_lxf64 v_a = __MM_SET_f64(a, gvl);
12
13  for (i = 0; i < n; i += gvl) {
14    gvl = __builtin_eabi_vsetvl(n - i, __eabi_e64, __eabi_m1);
15    __eabi_lxf64 v_dx = __MM_LOAD_f64(&dx[i], gvl);
16    __eabi_lxf64 v_dy = __MM_LOAD_f64(&dy[i], gvl);
17    __eabi_lxf64 v_res = __MM_MACC_f64(v_dy, v_a, v_dx, gvl);
18    __MM_STORE_f64(&dy[i], v_res, gvl);
19  }
    
```

Short VL

- As many Functional Units as VL.
- Vector instructions executed in 1 cycle

Long VL


- Cannot afford (area, power, cost) hundreds of Functional Units
- Vector instructions are executed on multiple cycles





Разработка матричных расширений RISC-V

<https://habr.com/ru/companies/yadro/articles/827430/>

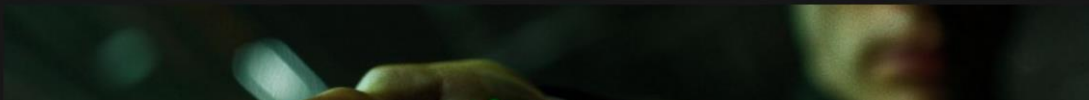
 **valeriaP**
9 июл в 17:17

Панорама матричных расширений: от x86 до RISC-V


15 мин 8.5K

Блог компании YADRO, Высокая производительность*, Алгоритмы*, Математика*, Машинное обучение*

Обзор




<https://habr.com/ru/companies/yadro/articles/827434/>

 **valeriaP**
30 июл в 17:32


Погружение в матрицу: расширение RISC-V от T-Head

17 мин 2.2K

Блог компании YADRO, Высокая производительность*, Алгоритмы*, Математика*, Машинное обучение*




<https://habr.com/ru/companies/yadro/articles/827432/>

 **valeriaP**
16 июл в 15:33


Заглянем в хрустальный шар: как продвигается разработка стандартных матричных расширений RISC-V

10 мин 3.8K

Блог компании YADRO, Высокая производительность*, Алгоритмы*, Математика*, Машинное обучение*




<https://habr.com/ru/companies/yadro/articles/833948/>

 **Andy31**
7 авг в 13:35

Математика матричных расширений: как происходит умножение матриц на примере T-Head Matrix Extension

13 мин 7.9K

Блог компании YADRO, Высокая производительность*, Алгоритмы*, Математика*, Машинное обучение*



HPC на RISC-V: почему уже пора?

HPC в мировых трендах экосистемы RISC-V

Примеры докладов с HPC RISC-V воркшопов 2024

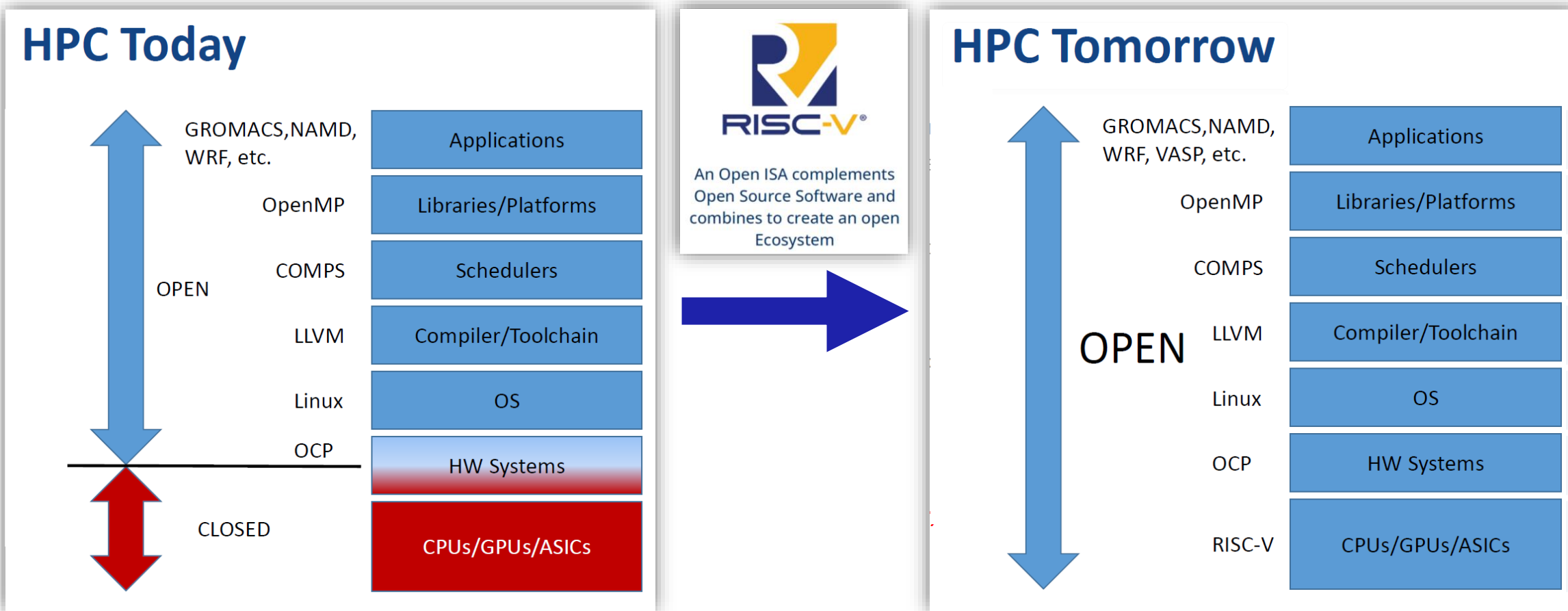
RISC-V HW: на чем тестировать HPC SW уже сейчас и чего ждать

Выводы и полезные ссылки



Новый мировой тренд – парадигма Open HPC

Не только стек HPC SW должен быть open-source, но и HPC HW



* Источник: <https://www.european-processor-initiative.eu/wp-content/uploads/2022/03/EPI-@-HPC-User-Forum.pdf>



RISC-V HPC инициативы

- Европейская инициатива по развитию собственной технологической независимости **EPI (European Processor Initiative)** предполагает разработку процессоров и ускорителей на базе RISC-V: решения на ARM не признаются ее частью.



Академическое сообщество

 **Государственная поддержка**

RISC-V HPC Research



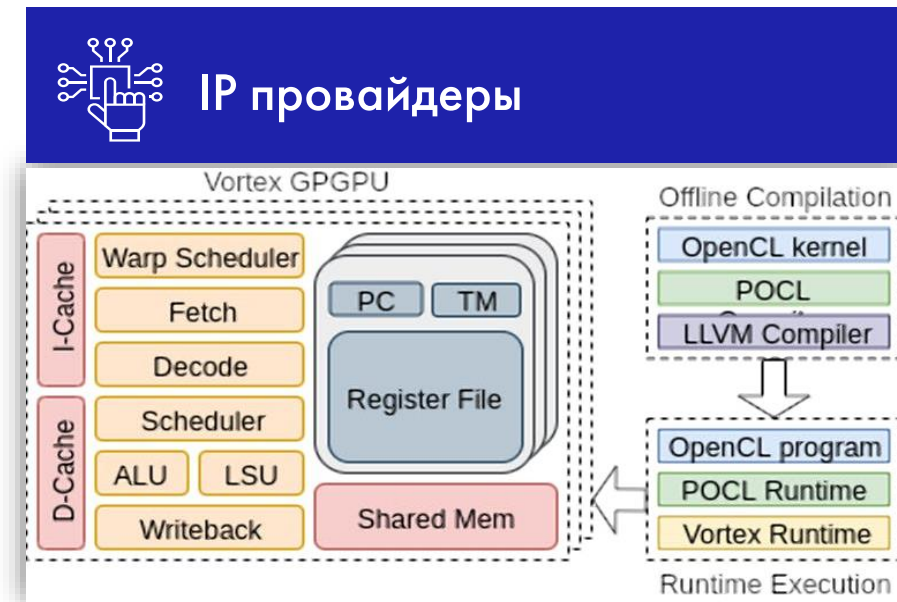
Logos shown include: European Processor Initiative (EPI), EuroHPC Joint Undertaking, eProcessor, THE EUPILOT, MEEP (MareNostrum Experimental Exascale Platform), TRISTAN RISC-V, and ISOLDE RISC-V.

- Крупнейшие европейские HPC центры – **BSC (Barcelona Supercomputing Center)** и **EPCC (Edinburgh Parallel Computing Center)** – активно развивают центры компетенции RISC-V в рамках грантовой поддержки **Euro HPC** (правительственная инициатива).



HPC тренды в развитии RISC-V HW и SW

- Появляются **высокопроизводительные ядра**: несколько IP провайдеров разрабатывают процессоры для дата-центров.
- На базе архитектуры RISC-V разрабатываются не только CPU, но и **ускорители** (GPU, AI).



Founding Member of RISE to Ensure RISC-V Software Readiness

RISE
RISC-V Software Ecosystem

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

Крупный мировой бизнес

- **Первая RISC-V HPC система** ожидается в 2025-2026 гг. в рамках 6-й реинкарнации европейского суперкомпьютера MareNostrum (BSC).
- **Консорциум RISE (RISC-V Software Ecosystem)** 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>



<https://riscv-optimization-guide.riseproject.dev/>



Тематика HPC RISC-V воркшопов

- Примеры использования и тематические исследования с RISC-V
- **Уроки, извлеченные из использования RISC-V в HPC**
- Отраслевые документы, посвященные изучению использования RISC-V
- **Перенос кода на RISC-V**
- **Новое оборудование и ускорители на основе RISC-V**
- Инструменты и методы, помогающие использовать RISC-V для HPC
- **Наработки в библиотеках HPC для их переноса на RISC-V**
- **Расширения RISC-V, ускоряющие HPC приложения**
- Компилятор и поддержка среды выполнения для RISC-V
- Экосистема RISC-V
- Взгляд в будущее: как RISC-V может развить сообщество HPC
- И все, что связано с RISC-V и HPC!





HPC SIG (Special Interest Group)

- Организация [семинаров](#) по HPC на RISC-V на профильных международных конференциях
- **Цель –** популяризация RISC-V в HPC (**способствовать портированию HPC SW на RISC-V и т.д.**)



HPC RISC-V воркшопы: 2024



Strategic EU-level perspective on RISC-V

RISC-V: the cornerstone ISA for the next generation of HPC infrastructures

17th January 2024 | **Alexandra Kourfali** | Munich, DE

HiPEAC 2024

- RISC-V Workshop: *RISC-V: the cornerstone ISA for the next generation of HPC infrastructures*
 - Organizers: E4 and BSC
 - **Accepted**
- Full Day workshop
- Munich, Germany
- January 17-19, 2024
- More details next meeting...

Upcoming RISC-V HPC Events



- HPC Asia RISC-V Workshop
 - <https://riscv.epcc.ed.ac.uk/community/hpcasia24-workshop/>
 - 25th Jan 2024



- Fourth International workshop on RISC-V for HPC
 - <https://riscv.epcc.ed.ac.uk/community/isc24-workshop/>
 - 16th May 2024



- RISC-V Summit Europe
 - <https://riscv.epcc.ed.ac.uk/community/isc24-workshop/>
 - 24 - 27th June 2024

Workshop at HPC-Asia

- Workshop HPC-Asia (Nagoya, Japan) : **Third International Workshop on RISC-V for HPC (RVHPC)**
 - Michael Wong, Nick Brown, and John Davis submitted a workshop proposal
 - Conference, end of January, 2024 about 500 people
 - Accepted
 - ½ Day (morning)

* Источник: <https://www.hipeac.net/2024/munich/#/program/sessions/8088>

HPC на RISC-V: почему уже пора?

HPC в мировых трендах экосистемы RISC-V

Примеры докладов с HPC RISC-V воркшопов 2024

RISC-V HW: на чем тестировать HPC SW уже сейчас и чего ждать

Выводы и полезные ссылки

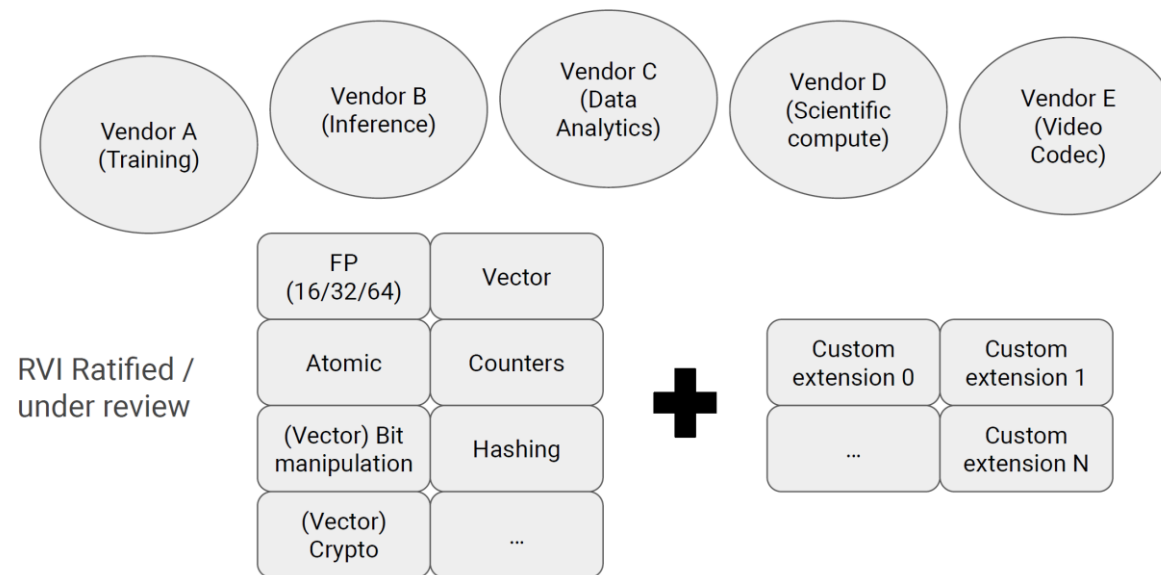


Challenges of Building an Open Source Ecosystem (1 / 2)

Problem Statement

- Silicon companies want to release hardware for which optimised OSS stacks are already present
- OSS community wants to support all platforms which users would like to run on
- Many different micro-architectures to bring to market
- Representative hardware doesn't exist yet (although RISC-V vector CPUs do exist)
- Avoid leaking of proprietary information before hardware is released
- Limited resources in:
 - Silicon companies - can't port everything themselves
 - OSS community - need to prioritise work in terms of impact

Server Class RISC-V: Designed for Fragmentation





Challenges of Building an Open Source Ecosystem (2/2)

Call to Action

- Participate in consortia / standardisation bodies
 - [RISE](#), [RVI](#), [UXL](#)
- Contribution to frameworks and tools used in multiple projects
 - Frameworks / APIs such as oneAPI, xsimd, OpenMP runtimes
 - Compilers are ubiquitous. Improvements in toolchain needed for one project helps many others
 - Raise issues for [GCC](#) and [LLVM](#)
- Improvements in OS packages, such as SIMD for frequently used operations (e.g. zlib (de)compression)
- RISE put out [RFPs](#) for prioritised development work
- Contribute to your favourite project

Open source software very open to pull requests, e.g.

- OS bring up (Fedora, Ubuntu, Android)
- OpenBLAS has had branch `risc-v` since August 2022
- Linux kernel supports hardware interfaces for with new non-ISA specs (e.g. hardware probe)
- 60/100 last commits in QEMU have a 'riscv' tag

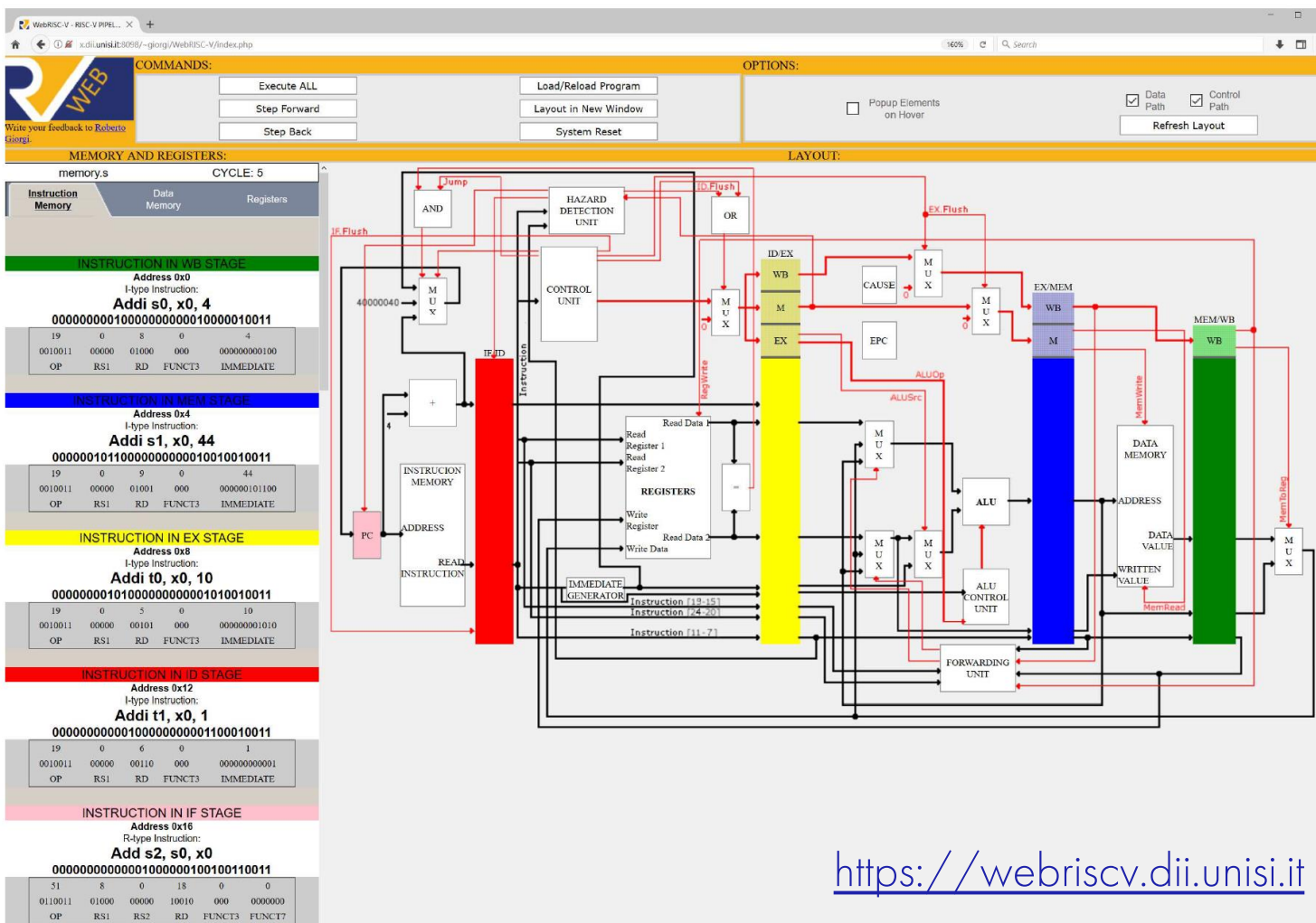
How to Optimise Code

- What are good / bad practices?
 - Use up-to-date toolchain. Most recently released, or better close to tip of tree
 - LLVM has autovec, GCC in the works
 - RISE optimisation guide imminently available (RV64GCV)
- Micro-architecture agnostic
 - Can double check codegen on [Compiler Explorer](#)
 - Proxy for performance via dynamic instruction counts in emulator
- Optimised for 'generic' target
 - Stick to intrinsics, and v1.0 vector spec
 - What LMUL, what ILP, what order? Don't over optimise. Strip-mined loops are good.
 - Don't optimise for order - OoO doesn't care, and different in-order may prefer different ordering
 - Performance estimation via LLVM MCA, for example



WebRISC-V: a web-based educational simulator

Providing a simple, easy accessible educational tool to test RISC-V programs on a pipelined processor.



<https://webriscv.dii.unisi.it>

Program-entry box

- Both "free" assembly (parsed for errors) and
- Predefined examples (simple calculator, factorial, ...)

Empty Text Box [1] Insert

2] Load the following program 3] Analyze pipeline

Automatic pipeline diagrams

- Capability of squashing loops, by marking them on the diagram

SQUASHED LOOPS

Instruction	CPU Cycles																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
addi a2, x0, 2	F	D	X	M	W													
addi a0, gp, 8	F	D	X	M	W													
beq a0, gp, 48		F	-	D	X	M	W											
lw t0, 100(a0)				F	D	X	M	W										
add t0, t0, a2					F	-	D	X	M	W								
sw t0, 200(a0)						F	D	X	M	W								
jal x0, -32							F	D	X	M	W							
addi a0, a0, -4								F	D	X	M	W						
beq a0, gp, 48									F	-	D	X	M	W				
lw t0, 100(a0)										F	D	X	M	W				
addi x0, x0, 0											F	D	X	M	W			

LOOP #0 - 'beq a0, gp, 48' TO 'addi a0, a0, -4': 8 cycles 2 times



SW stack for future HPC machines based on RISC-V 128 bit

Opportunities and challenges for RISC-V and 128 bit

Heterogeneity

ISA extension is about **managing heterogeneity in an homogenous way**:

- Base RISC-V ISA on all clusters
- Various set of ISA extension on different clusters

Operating System

A 128 bit address space can provide a **unified view of the 100 M cores**:

- **Single system image** of the machine
- A 128 bit **process spans the whole machine with a single virtual address space**

Issues:

- **Distributed system issues**. Like what is the status of a 128 bit process?
- **Threads migration** across clusters?
- Need for **transactions**?

Starting point:

- PGAS and its variants

Langages & Tools

A single address space for a process means:

- The **compiler could work on the full application at once**
- Room for a **generalised OpenMP-like programming model**
- Potentially **replace MPI** by VM operations

Opportunity:

- MLIR-based DSL.

Issues:

- Need for **transactions**?

[1] A. Waterman and K. Asanović, “Chapter 6, RV128I Base Integer Instruction Set, Version 1.7,” in The RISC-V Instruction Set Manual - Volume I: Unprivileged ISA, 20191213, The RISC-V Foundation, 2019. Available online at <https://riscv.org/technical/specifications/>

Name	Registers	Register width	Address width
RV32	32	32 bit	
RV32E	16	32 bit	
RV64	32	64 bit	
RV128	32	128 bit	

- Addresses and integers are 128 bit wide.
- Still base ISA + ISA extensions, just like 32 and 64 bit :
 - Integer multiply & divide
 - 32 bit floating-point
 - 64 bit floating-point
 - 128 bit floating-point
 - etc.

The challenge is how to take advantage of RISC-V and 128 bit to improve

- The heterogeneity (↗) of the machine
- The operating system stack
- Programming languages and tools



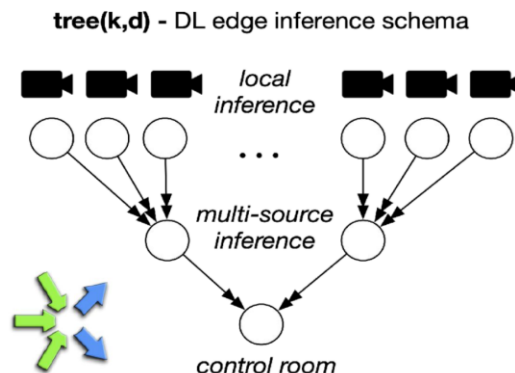
RISC-V for AI: enabling modern workloads on modern HW

Use case: Man Down Detection

Goal: real-time distributed AI surveillance at a large scale

Target: detect people in lying on the ground in distress

Design: leverage YOLO-V5 on multiple RISC-V-based edge nodes in a tree structure connected via *FastFlow*



Challenges: limited RISC-V ecosystem, need to port:

- FastFlow
- AI library (e.g., PyTorch)

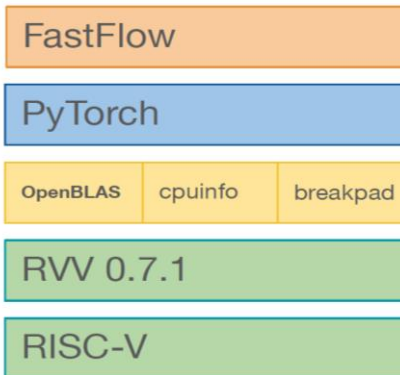
Outcome: implemented with the Fast Federated Learning (FFL) framework:

- based on C/C++ for performance (libtorch + FastFlow)
- supports both *federated training* and *distributed inference*

Accelerated PyTorch WiP: preliminary results

System	Cores	Total [s]	[ms]/image
k230	1	254.11	79.41
Milk-V (OpenBLAS)	1	254.91	79.66
Milk-V	64	137.91	43.09
Milk-V (OpenBLAS)	64	25.88	8.08
Intel	1	11.76	3.67
Intel	64	1.95	0.61

4-layer (2 convolutional + 2 fully connected) DNN performance on 100 batches of 32 MNIST images



Pro: highly parallel: 64 cores supporting RWV 0.7.1

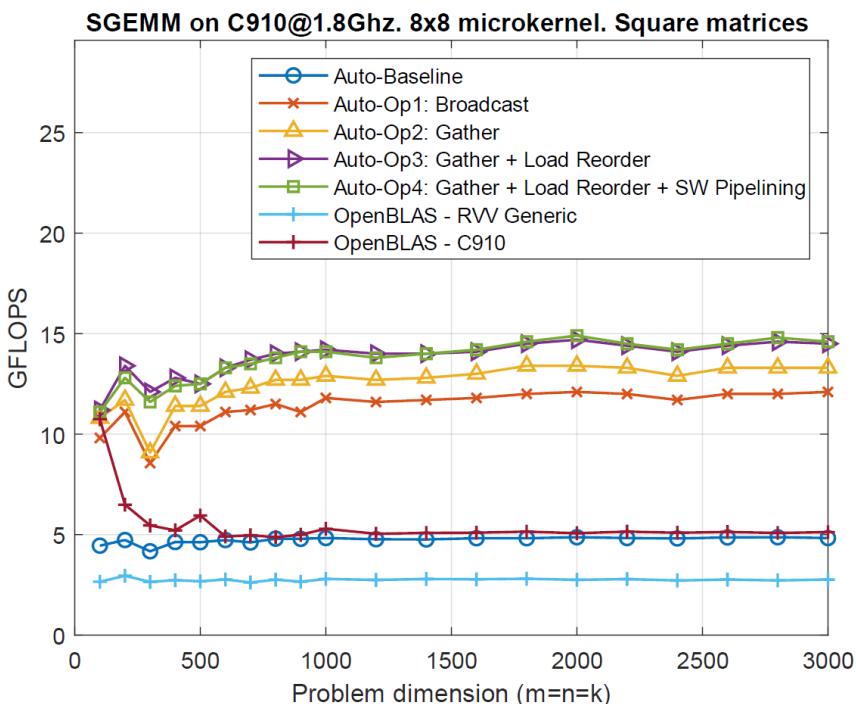


Performance analysis (& optimization) of BERT on RISC-V

We focus on BERT + inference

- Useful across several NLP tasks
- Illustrative of the potential of architectures and space for optimization in transformers
- Inference typically deployed on low-power CPUs, typically with SIMD

Results - C910, 8x8 microkernel, square matrices



- **Auto-Baseline vs. OpenBLAS**
 - 1.72x improvement vs. OpenBLAS RVV Generic
 - Similar performance than OpenBLAS C910
- **Auto-Op1 (*bcast*)**
 - 2.38x improvement vs. Auto-Baseline
- **Auto-Op2 (*gather*)**
 - 2.62x improvement vs. Auto-Baseline
- **Auto-Op3 (*load reorder*)**
 - **2.90x improvement vs. Auto-Baseline**
 - **2.59x improvement vs. C910 OpenBLAS**
- **Auto-Op4 (*SW pipelining*)**
 - 2.88x improvement vs. Auto-Baseline

HPC на RISC-V: почему уже пора?

HPC в мировых трендах экосистемы RISC-V

Примеры докладов с HPC RISC-V воркшопов 2024

RISC-V HW: на чем тестировать HPC SW уже сейчас и чего ждать

Выводы и полезные ссылки



Кластеры

Barcelona Supercomputer Center ([BSC](#))

Board	OS	Details
PolarFire	Fedora	4 cores w/ 2 GB
BeagleV	Fedora	2 cores w/ 8 GB
Unmatched	Fedora/Ubuntu	4 cores w/ 16 GB
Allwinner D1 (Vector extension)	Fedora	1 core w/ 2 GB

Edinburgh Parallel Computing Center ([EPCC](#))

Board	Processor (SoC)	# Cores	DRAM (GB)	Qty
NezhaSTU	C906 (D1)	1	0.5	4
MangoPi MQ-Pro	C906 (D1)	1	1	2
HiFive Unmatched	U74 (FU740)	4	16	1
StarFive VisionFive V1	U74(JH7100)	2	8	3
StarFive VisionFive V2	U74(JH7110)	4	8	15
Lichee Pi 4A (on order)	C910 (TH1520)	4	16	2

* Источник: <https://excalibur.ac.uk/excalibur-events-isc-23/>

E4: Monte Cimone (V1)

4x E4 RV007 1U Custom Server Blades:

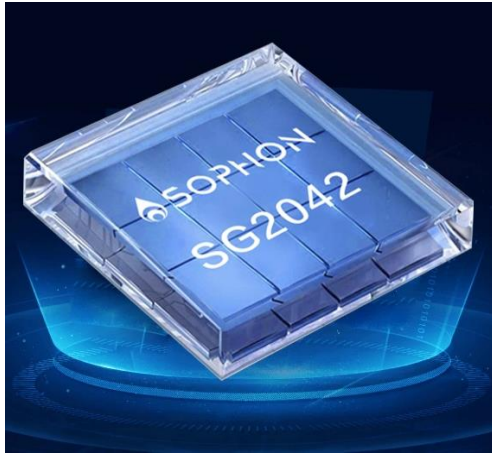
- 2x SiFive U740 SoC with 4x U74 RV64GCB cores
- 16GB of DDR4
- 1TB node-local NVME storage
- PCIe expansion card w/InfiniBand HCAs
- Ethernet + IB parallel networks

E4: Monte Cimone (V2 = V1 + SG 2042)





Milk-V Pioneer – IP для датацентров и AI/ML



Pioneer Box

- 1X SG2042 CPU
- 1x Developer Board
- 250W ATX Power supply
- Intel AX210 WiFi 6E / BT5.2 card
- Dual 10G SFP Network Card
- Graphice Card AMD Radeon RX550 4GB
- Nice and compact enclousre with carrying handle
- 1TB Nvme SSD
- 2x 16G DDR4
- Powerful RGB CPU cooler

Milk-V Vega – первый в мире RISC-V коммутатор стандарта 10GbE компании Shenzhen MilkV Technology (Milk-V).



Предназначен для:

- сетей широкополосного доступа,
- платформ видеонаблюдения и аудиовизуальных сервисов,
- систем умных городов и пр.

> 1 TFLOPS(FP64)

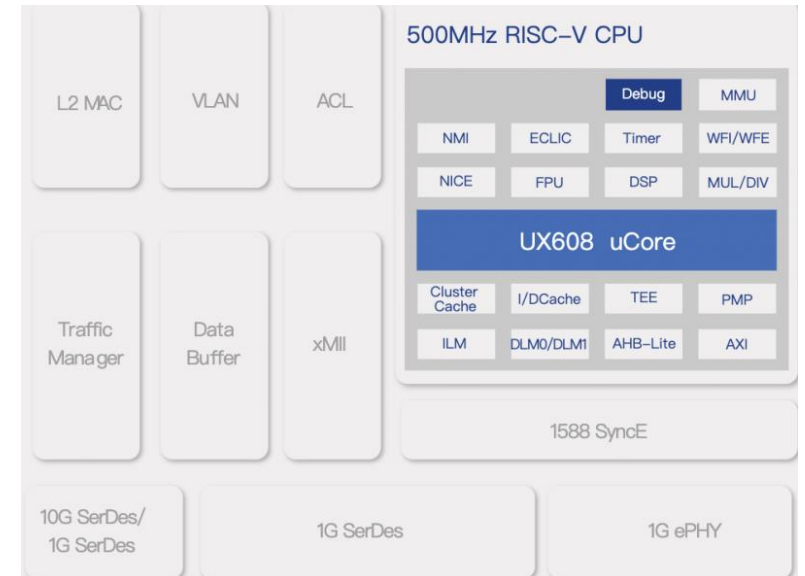
- 64 Cores
- 2 GHz
- 120 W TDP
- 3200 MHz (Max DIMM Frequency)

- 1 Gbit Ethernet
- 1 LPC



- Up to 256 GB RAM
- 4 MB L1 Cache
- 16 MB L2 Cach2
- 64 MB L3 Cache

- 2 SPI Flash Interface
- 2 General SPI Controller



* Источник: <https://servernews.ru/1081875>

Banana Pi BPI-F3: 8-ядерный процессор SpacemiT K1

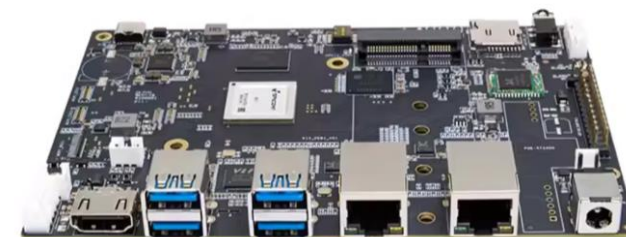


https://docs.banana-pi.org/en/BPI-F3/BananaPi_BPI-F3

- Ядра **SpacemiT X60** (4 ядра из 8 с **Integrated Matrix Extension**).
 - 256-битные векторные регистры.
 - 1.3x Arm Cortex A55.
- Бенчмарки: 2.0 TOPs AI.
- Спецификация: <https://github.com/space-mit/riscv-ime-extension-spec>



BPI-F3 SpacemiT K1
Octa-core RISC-V



Доступны
для заказа.

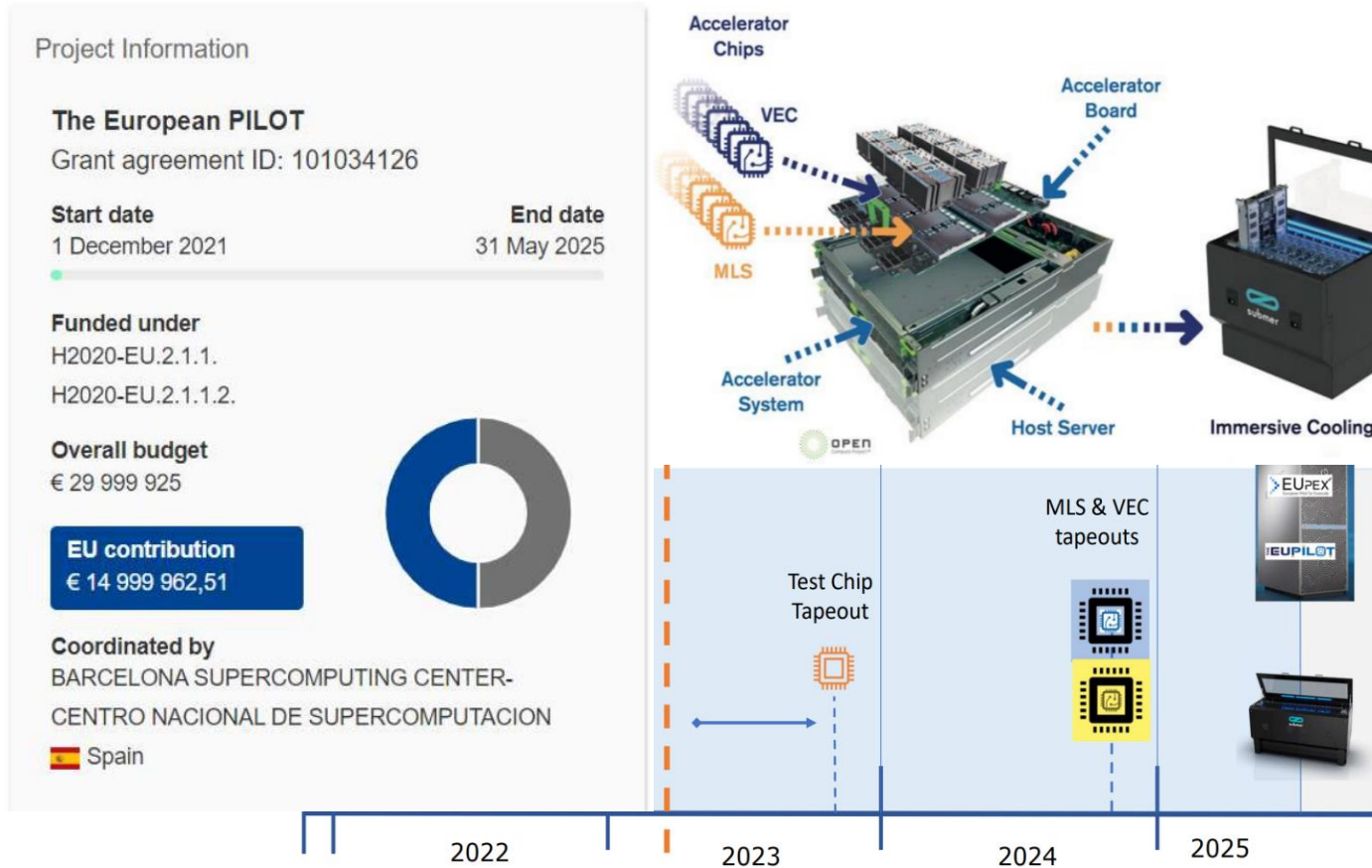
7 794 ₽

Цена за 1 лот •





EUPILOT: разработка RISC-V ускорителей для HPC и AI



- Создание европейской платформы для HPC и AI.
- Достижение европейского цифрового суверенитета в HPC.

Target: Chips → Deployments

Hardware	Chips → Modules → Boards
Systems	Boards → Systems → Liquid Immersion Deployments
Software	Drivers → OS → Compilers → Frameworks → Apps

- Основа для европейских систем Exascale.
- Расширение экосистемы RISC-V на домены HPC и HPDA.

* Источник: <https://open-src-soc.org/2022-05/media/slides/RISC-V-International-Day-2022-05-05-15h45-John-Davis.pdf>



Occamy от ETH (Zurich)

A 432-core, Multi-TFLOPs RISC-V-Based 2.5D Chiplet System for Ultra-Efficient (Mini-)Floating-Point Computation

Our latest design Occamy: 0.75 TFLOP/s, 400+ cores

Dual Chiplet System Occamy:

- 216+1 RISC-V Cores
- 0.75 TFLOP/s
- GF12LPP
- Area: 73mm²

2x 16GBByte HBM2e DRAMs Micron

2.5D Integration

Silicon Interposer Hedwig:

- Technology: 65nm, passive (only BEOL)
- Area: 26.3mm x 23.05mm

Carrier PCB:

- RO4350B (Low-CTE, high stability)
- 52.5mm x 45mm

PULP
Parallel Ultra Low Power

Preliminary measured results:

- Dense Kernels:
 - GEMMS: ≥ 80% FPU utilization (also for SIMD MiniFloat)
 - Conv2d: ≥ 75% PFU utilization (also for SIMD MiniFloat)
- Stencil Kernels: 80%+ FPU utilization
- Sparse Kernels: 50%+ FPU utilization

➔ The power of ISA extensions!

Peak System perf. @1GHz:

FP64:	768 GFLOp/s
FP32:	1.536 TFLOp/s
FP16:	3.072 TFLOp/s
FP8:	6.144 TFLOp/s

There is much more to come in Q3-2023 ...

<http://pulp-platform.org> @pulp_platform

- Initial discussions 20th of October 2020
- Started on 20th of April 2021
- Taped out Chiplet on 1st of July 2022
- Taped out Interposer on 15th of October 2022
- Currently being assembled

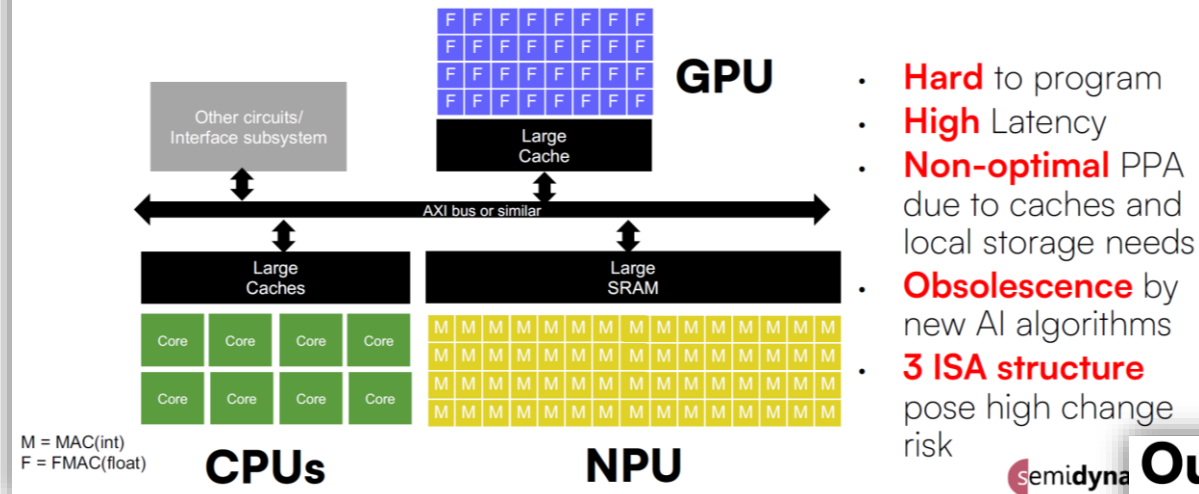


* Источник: <https://pulp-platform.org/occamy/>



Разработки Semidynamics для Big Data & AI/ML

Typical AI-focused Subsystem SOC to date...



Our IP



Atrevido 勇敢な
无所畏惧 Fearless 용감한
64b out-of-order CPU
RISC-V
AXI and CHI



Avispado 賢い
机智 Smart 똑똑한
64b in-order CPU
RISC-V
AXI and CHI

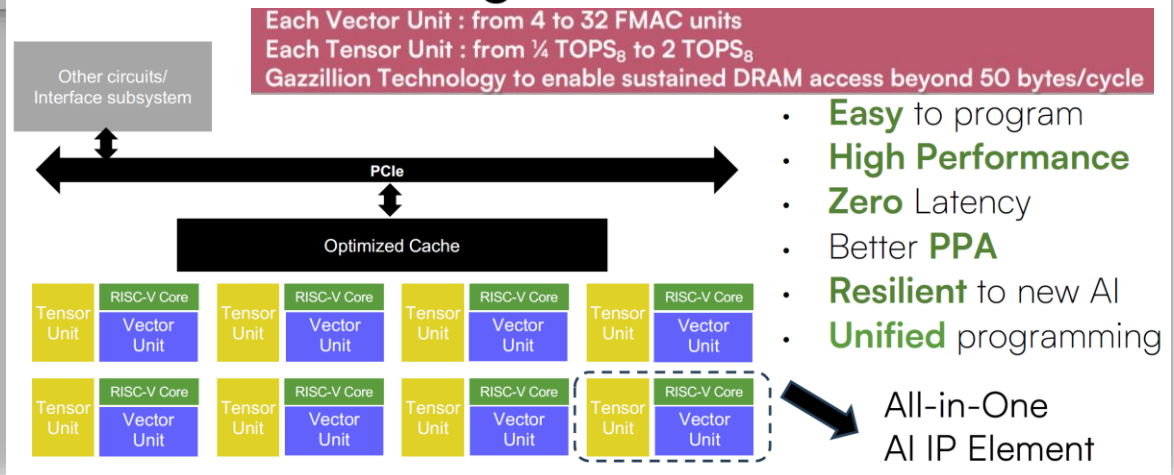


Vector Unit
RVV1.0
Out-of-order



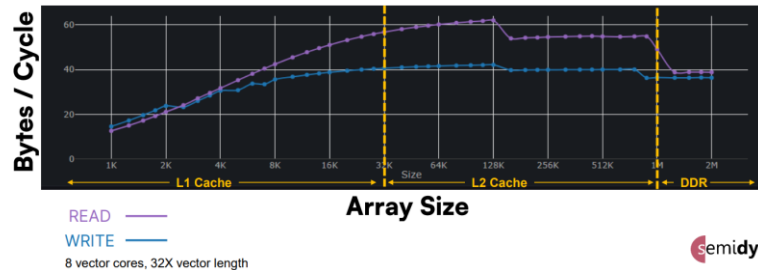
Tensor Unit
BF16, FP16, INT8

Our vision: Fusing CPU, GPU, and NPU



Vector + Gazzillion: A bandwidth rocket!

Can you find a core out there capable of streaming data at over 60 Bytes/cycle? And from main DDR memory (not from your cache)? We don't think so 😊





Atrevido 423-V8 от Semidynamics для Big Data & AI/ML

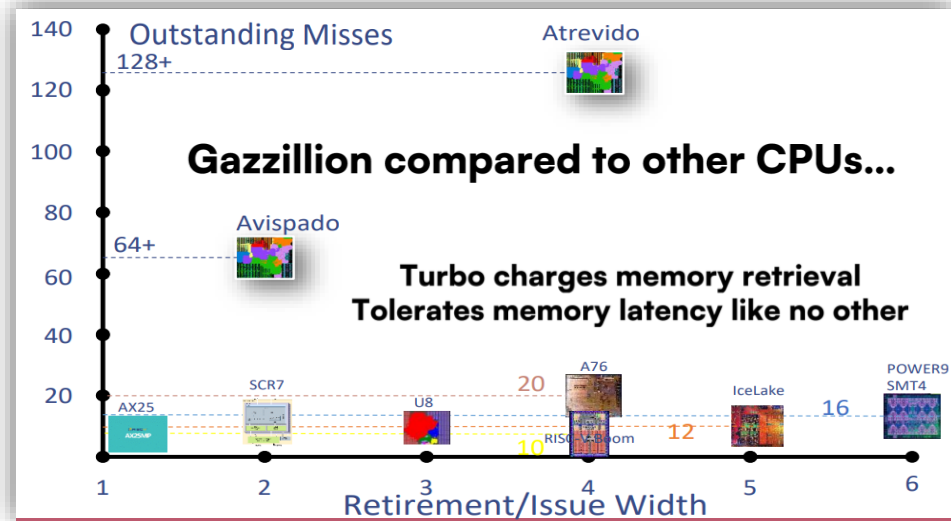
The Semidynamics Proposal

- Powerful **Out Of Order** based on Risc-V
- Combine **CPU** with **Vector** and **Tensor unit** to create powerful AI capable Compute building blocks
- Enable Hypervisor Support for Containerization
- Enable Crypto for Security / Privacy
- Easy to combine with custom logic / Unit — 3 custom instructions
- Use of **Gazzillion™ Technology** to efficiently manage large data sets



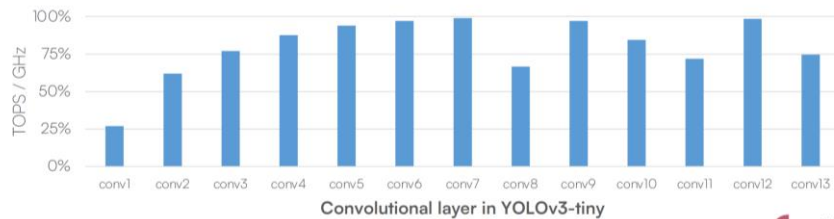
Benefits

- **Easy** to program
- **High Performance** for Parallel Codes
- **Zero** Communication Latency



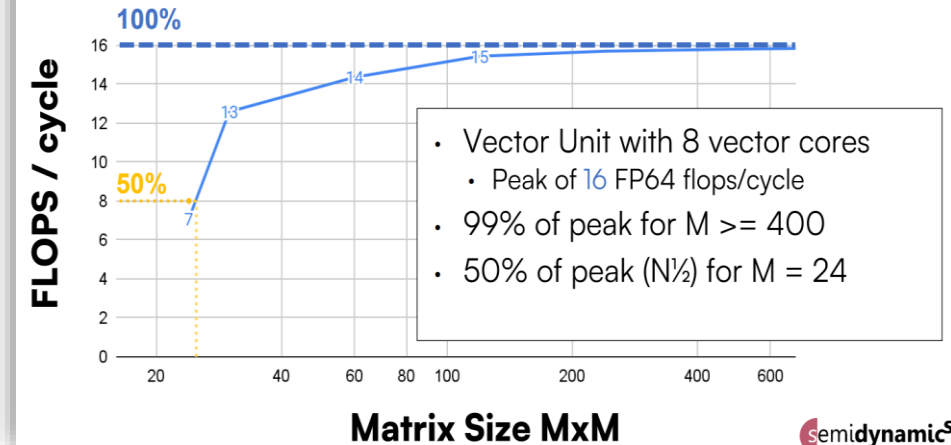
YOLO on our fused IP: 33 FPS

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



DGEMM on Atrevido 423 + V8

(FP64 matrix multiply)



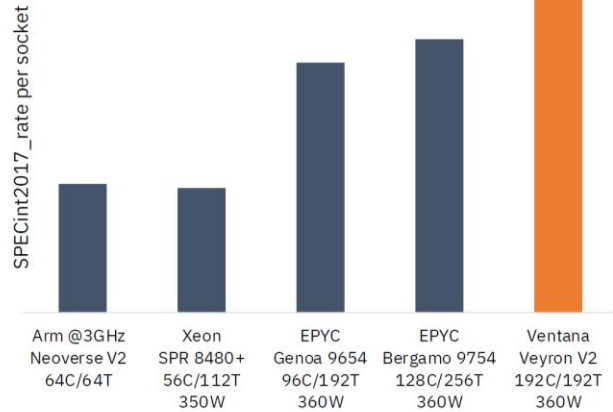
* Источник: <https://www.eejournal.com/article/want-tailormade-screamingly-high-performance-risc-v64-ip/>



Ventana Veyron (V2)

Veyron V2: Momentum to Mainstream with Complete Platform

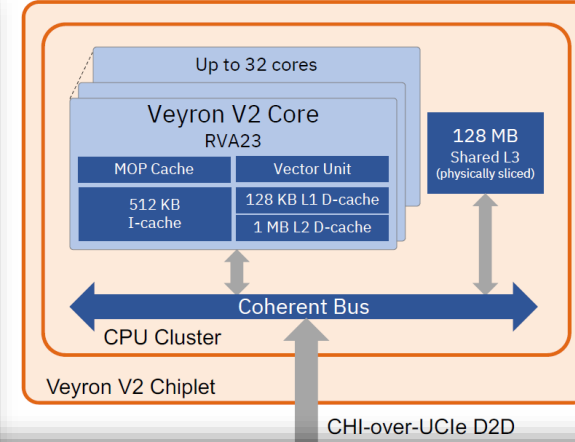
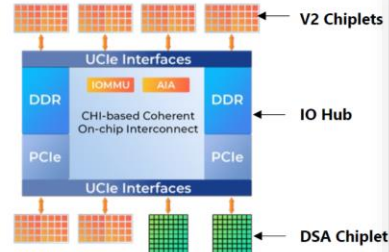
Highest Performance RISC-V Processor for Data Center, Automotive, Gen AI and Client



Highlights

- +40% performance, 32 cores per cluster, 4nm
- UClE chiplet
- RISC-V Vector Extension support
- Ventana AI Matrix Extensions
- Server-class IOMMU
- RISE support
- Domain Specific Acceleration

Available as UClE-Compatible Chiplets or IP



Vector Unit

- Full RVV1.0 “V” support plus new standard and custom RISC-V extensions:
 - Vector crypto
 - FP16 / BF16
 - Widening 8x8 int8 and BF16 matrix multiplies
- VLEN = DLEN = 512
 - 32 64B-wide vector registers
 - 64B-wide fully pipelined load, store, and register operations
 - No double pumping of datapaths
 - 64B load plus 64B store per cycle (with arbitrary alignments)
- Area and power efficient high-performance design
 - Separate vector register-operation scheduler, register file, and execution pipes from the “scalar” core
 - Five parallel execution pipes: Arithmetic, Mask, Permute, Load data, Store data
 - Out-of-order execution across execution pipes and within pipes without register renaming
 - LMUL chaining
 - Interleaving of LMUL>1 operations and complex operations within each pipe based on dependencies
 - No speculative register execution but full speculative load/store execution
 - No speculative execution recovery buffers



Ventana Veyron: планы

HPC Software Ecosystem

Applications Benchmarks	HPCG HPC Challenge Performance	NAMD LAMMPS GROMACS Molecular Dynamics	CP2K Quantum-Espresso NWChem Quantum Chemistry	OpenFOAM OpenSBLI NEMO CFD	WRF ROMS HPCW Climate Weather	QMCPACK RMG Manufacturing	Tools GCC LLVM GO python PAPI perf extrae GDB	
Middleware Frameworks, Libraries	VASP OpenMP	ALYA glibc OPEN MPI	MPI-IO boost LAPACK	GPFS OpenBLAS OpenCL	OneDNN FFTW HDF TensorFlow	slurm PyTorch		
Operating System	ubuntu®	fedora	Red Hat Enterprise Linux					
Hypervisors Containers	KVM	docker	kubernetes	SINGULARITYCE				
Firmware Early Boot, BIOS	OpenSBI	tianocore	ACPI					
Platform HPC Server								

AI/ML Software Ecosystem

Applications	Computer Vision	Speech	Natural Language Processing	Autonomous Systems	Recommendations	Finance	Tools python GCC LLVM MLIR glibc GDB OpenOCD
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			
HAL Libraries	oneAPI	OpenCL	OpenAI Triton	AMD ROCm	NVIDIA CUDA		
Hypervisors Containers	KVM	docker	kubernetes				
Operating System	Linux	ubuntu®	fedora				
Firmware Early Boot, BIOS	OpenSBI	tianocore	ACPI				
Platform	Ventana Veyron AI/ML Server 						



ET-SoC-1: Esperanto's RISC-V Supercomputer on a Chip

Over 1,000 64-bit RISC-V CPUs per Chip

- As low as 13W per SoC (workload dependent)
- High efficiency operation (inferences /sec / watt)
- Accelerates wide range of AI/ML workloads
 - Language Models (NLP/LLM)
 - Visual Models (Detection, Segmentation)
 - Recommendation Models (RecSys, DLRM)
- Allows flexible general-purpose computing
- Enables pre- and post-processing
- Highly efficient HPC workload through massive parallelism
- High bandwidth interconnect network
- Dynamic, tiered architecture of 160 MB on-die SRAM for caches
- Up to 32GB LPDDR4x DRAM
- 4 ET-Maxion high-performance OOO RISC-V cores
- TSMC 7nm

Esperanto's RISC-V hardware and software is here today

Eighty Thousand Esperanto RISC-V processors at work in photo at right

One standard rack can hold twenty Esperanto 2U servers with

- 320 Esperanto ET-SoC-1 accelerator chips
- 1088 64-bit RISC-V CPU with vector/tensor accelerators per ET-SoC-1
- 348,160 total RISC-V processors
- 24 PetaOps Int8 precision peak performance
- 6 PetaFlops FP16 precision peak performance
- 3 PetaFlops FP32 peak precision performance



Being used today for Machine Learning workloads

You can use it tomorrow for anything you want with General Purpose SDK

Get started with Esperanto's systems for ML and start preparing for HPC

Esperanto System Product Portfolio



Data Center Server



- Based on Gigabyte G292-G20 form factor
- Up to 16 ET-SoC-1 PCIe cards
- Over 16,000 RISC-V Cores

Enterprise Edge Server



- Based on Gigabyte E252 form factor
- Up to 6 ET-SoC-1 PCIe cards
- Over 6,000 RISC-V Cores

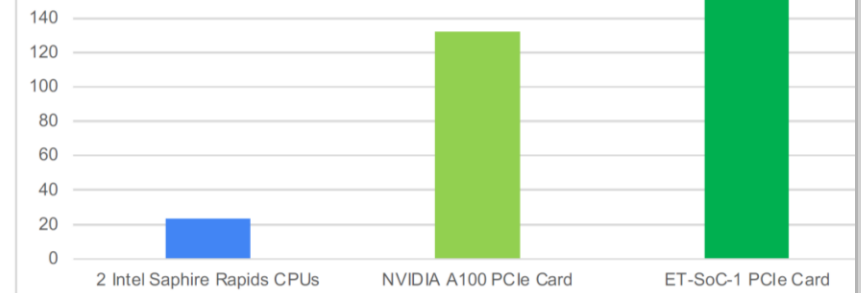
Rugged Edge Server



- Based on Stratus Edge ztC 250i form factor
- 1 ET-SoC-1 PCIe card
- Over 1,000 RISC-V Cores

ResNet50 Int8 Inferences/Sec/Watt

Esperanto RISC-V based ET-SoC-1 has excellent Performance per Watt



	2 Intel Sapphire Rapids CPUs	NVIDIA A100 PCIe Card	ET-SoC-1 PCIe Card
RN50 Inf/Sec	16178	39672	5059
Card or CPU Power	700	300	32
Inf/Sec/Watt	23	132	158



Esperanto ET-Minion

ET-Minion is an Energy-Efficient RISC-V CPU with a Vector/Tensor Unit

CPU is tailored for Massively Parallel ML Applications



ET-MINION IS A CUSTOM BUILT 64-BIT RISC-V PROCESSOR

- In-order pipeline with low gates/stage to improve MHz at low voltages
- Architecture and circuits optimized to enable low-voltage operation
- Two hardware threads of execution
- Software configurable L1 data-cache and/or scratchpad

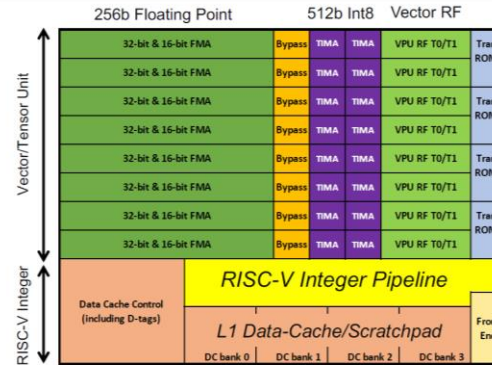
ML OPTIMIZED VECTOR/TENSOR UNIT

- 512-bit wide integer per cycle
- 128 8-bit integer operations per cycle, accumulates to 32-bit Int
- 256-bit wide floating point per cycle
- 16 32-bit single precision operations per cycle
- 32 16-bit half precision operations per cycle

New multi-cycle Tensor Instructions

- Can run for up to 512 cycles (up to 32K operations) with one tensor instruction
- Reduces instruction fetch bandwidth and reduces power
- RISC-V integer pipeline put to sleep during tensor instructions

Vector transcendental instructions



ET-Minion RISC-V Core and Tensor/Vector unit optimized for low-voltage operation to improve energy-efficiency

Optimized for energy-efficient ML operations. Each ET-Minion can deliver peak of 128 Int8 GOPS per GHz

RISC-V is the right choice for future merged ML/HPC Systems



RISC-V is not only the best choice, RISC-V is the only logical choice for future ML/HPC systems

Making systems easier to program with scalable set of processors with **one instruction set** should be the goal

- x86 and ARM processors too heavyweight to serve as both main CPU and accelerator
- GPU's too hard to program, can't be the main processor
- Only RISC-V has the ability for both:
 - High performance main cores: e.g. Tenstorrent, SemiDynamics, Ventana, Andes, RIVOS, ET-Maxion and others
 - Lightweight RVV vector cores: Esperanto's ET-Minions and likely many others

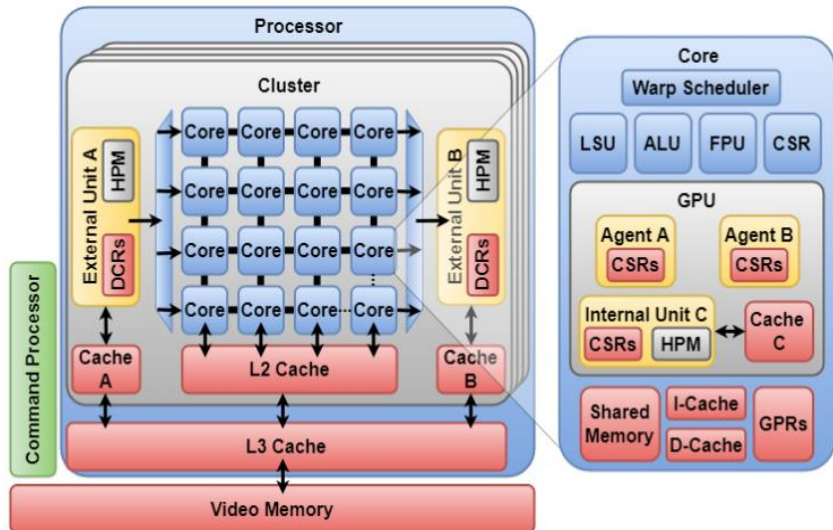
RISC-V is now mature and ready to start the revolution for future ML/HPC computing systems

Dave's prediction: RISC-V based system will win the Green500 in the next 5 years



Vortex: OpenCL Compatible RISC-V GPGPU

Работает на FPGA, есть конвейер для запуска программ на NVIDIA CUDA

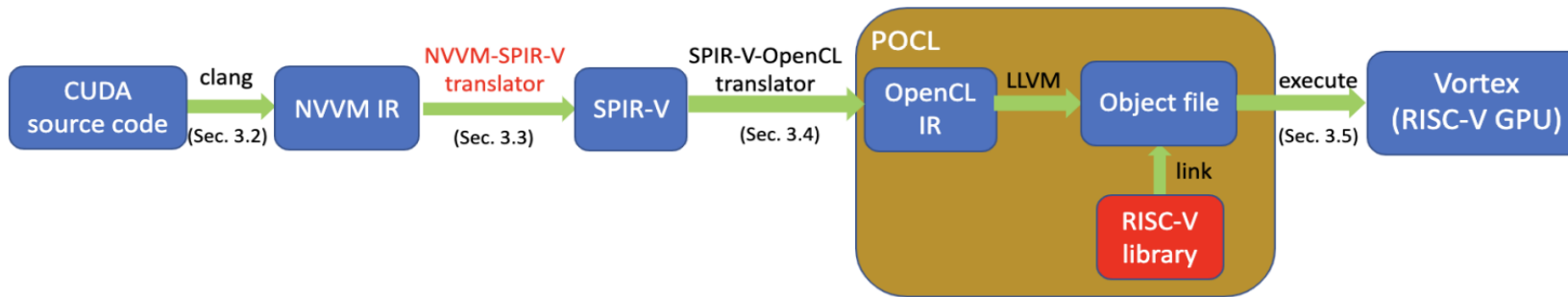


ISA Considerations

Operation Type	Considerations
Vertex/Frag Shaders	V Extension or Vec4 Custom
Number of Registers	Typically GPUs have more Vector Registers (e.g. 128) to avoid use of stack in a multithreaded environment
Data Types	Single Precision / Half Precision / fixed point (8 or + for HDR)
ISA Width	Often wide instructions 128-bit with embedded shuffle and write masks
Constant Register	GPUs have also a number of constant registers for uniforms
ABI	How to map Varyings / Uniforms / Attributes ?

- Translating applications in Rodinia benchmark
- Vortex(v0.2.2) NVPTX-SPIR-V translator(v0.1.0)

application	feature	support?
b+tree	-	yes
bfs	-	yes
cfd	double3 type	yes
huffman	atomic	yes
pathfinder	memory hierachy	yes
gaussian	-	yes
hotspot	-	yes
hotspot3D	-	yes
lud	memory hierachy	yes
nw	-	yes
streamcluster	-	yes
particlefilter	d2i	on going
backprop	__log2f	on going
lavaMD	d2i	on going
kmeans	texture	no
hybrid sort	texture	no
leukocyte	texture	no



- Vortex:
 - Support RISC-V RV32IMF ISA
 - Scalability: up to 64 cores with optional L2 and L3 caches
- Performance:
 - 1024 total threads running at 250 MHz
 - 128 Gflops of compute bandwidth
 - 16 GB/s of memory bandwidth
- Software: OpenCL 1.2 Support
 - Supported FPGAs:
 - Intel Arria 10
 - Intel Stratix 10

* Источник: <https://vortex.cc.gatech.edu/>

HPC на RISC-V: почему уже пора?

HPC в мировых трендах экосистемы RISC-V

Примеры докладов с HPC RISC-V воркшопов 2024

RISC-V HW: на чем тестировать HPC SW уже сейчас и чего ждать

Выводы и полезные ссылки



Выводы

- **RISC-V быстро развивается:** новые расширения ISA, процессоры серверного класса, ускорители, интерконнекты.
- **Перенос HPC кодов и реализация HPC алгоритмов на RISC-V – длительный процесс,** RISC-V HPC SIG призывает начинать его уже сейчас.
- **Актуальная задача – перенос SparseBLAS на RISC-V,** можно начать с библиотек Eigen, SuiteSparse, Kokkos.
- Для тестирования сейчас используются кластеры, собранные из существующих RISC-V плат, симуляторы, есть RISC-V GPU на FPGA.
- В мире работы идут более трех лет.

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