Библиотека векторных математических функций RVVMF в задачах анализа рыночных рисков

Андрей Николаев

Поддержка: Иосиф Мееров и Валентин Волокитин (ННГУ), Виктория Патокова (Сбер)

Данная информация носит ознакомительный характер и не предназначена для коммерческого, некоммерческого и иного использования, в том числе использования в общественных целях без письменного согласования с ПАО «Сбербанк»

Повестка

Что такое STAC

STAC-A2 бенчмарк

Реализация STAC-A2 бенчмарка

Сравнительный анализ реализаций RISC-V и Intel64

Выводы

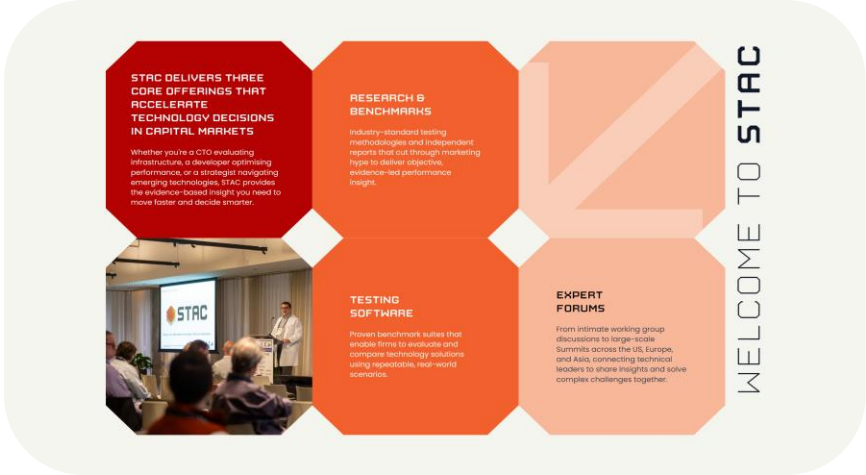
Что такое STAC

- Securities Technology Analysis Center
- Основан в 2006 by Peter Lankford

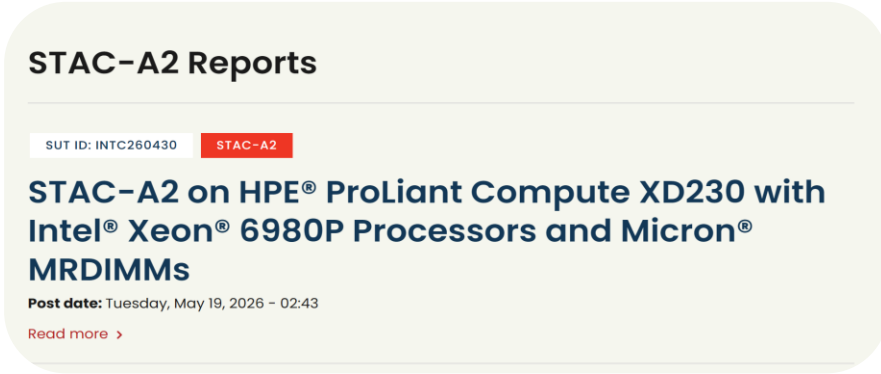
- STAC Benchmark Council
- 500+ финансовых компаний (банки, фонды, биржи)
- 70+ поставщиков технологических решений

Технологические исследования, саммиты, рабочие группы

Бенчмарки для тестирования производительности технологий для финансовых рынков



https://stacresearch.com/about/



https://docs.stacresearch.com/a2

Что такое STAC

Большой набор STAC стандартов и тестов для оценки эффективности оборудования и ПО

FAST WORKLOADS

- **M1:** low-latency direct-feed or "ticker plant" solutions
- **M2:** Messaging middleware under market data distribution workloads
- **N2:** Network IO between data centers and cloud instances
- **M1:** Tick-to-trade latency measurements
- **N1/T0:** Network IO of network stacks in a market data or tick-to-trade environment
- **TS:** Time synchronization and tracebility

BIG WORKLOADS

Тема этого доклада

- **M3:** Enterprise Tick analytics
- **A3:** Strategy backtesting
- **A2:** Risk computations

AI AND ML

- **ML:** Machine Learning workloads in finance
- **AI:** Artificial Intelligence technologies in finance including LLM and RAG

STAC-A2 бенчмарк

STAC-A2 – индустриальный стандарт тестирования технологических стеков в задачах вычисления цен финансовых активов и управления рисками

Опцион – контракт, представляющий покупателю право купить (call) или продать (put) акцию по заранее определенной цене в фиксированный момент времени в будущем или до него

Греки – метрики чувствительности стоимости опциона к изменению параметров рынка (например, к начальной стоимости цены акции, банковской ставке)

Слой бенчмарка	Оценка качества реализации слоя
Вычисление exp(), log(), sqrt(), IEEE-754 DP	Unit in the last place (ULP)
Генерация независимых нормальных величин	Андерсон Дарлинг, критерий Гильберта Шмидта
Генерация коррелированных нормальных величин	Андерсон Дарлинг, среднеквадратическая ошибка оценки корреляций
Генерация цен одной акции (модель Хестона, метод Андерсена)	Среднеквадратическая ошибка среднего и дисперсии цены, ошибка оценки call и put опционов
Генерация цен пакета акций (метод Андерсена)	Среднеквадратическая ошибка оценки корреляций
Вычисление цены опциона Американского типа на пакет акций (метод Лонгштаффа Шварца)	Сравнение с ценой опциона на базе биномиальной модели Блэка Шоулса для отдельных акций
Вычисление греков	Сравнение с ценой опциона на базе формул Хестона/Маграбэ для отдельных/двух акций

STAC-A2 бенчмарк

STAC-A2 – индустриальный стандарт тестирования технологических стеков в задачах вычисления цен финансовых активов и управления рисками

Производительность заданной конфигурации
5 акций, 252 шагов, 25 000 Монте-Карло траекторий

Масштабирование нагрузки по

- количеству акций: определение максимального пакета акций, для которого возможно вычислить греки в течение 10 мин
- количеству Монте-Карло траекторий цен акций

Масштабирование технологического стека, определяется технологией

- например, по кол-ву процессоров, сокетов, ядер

Оценка энергопотребления сервера

- во время вычисления греков для максимально возможного пакета акций

Реализация STAC-A2 бенчмарка

Вычислительная схема



Строительные блоки

Численное дифференцирование

Delta = $\frac{\partial P}{\partial S}$

Gamma = $\frac{\partial^2 P}{\partial S^2}$

Ro = $\frac{\partial P}{\partial r}$

Exp()
Линейная регрессия
Матричное умножение

Генератор мультивариантного нормального распределения
Log()
Sqrt()

Размеры данных

- IEEE 754 DP
- Path[2 x 252 x 25 000] – цена/волатильность одной акции
- Exp (V), V[1 x 252 x 25 000]
- A x = B, A[n x 8], B[n]
- A * B, A[8 x n], B[n x 8]
- n=0,...25 000, n зависит от значений цен

Реализация STAC-A2 бенчмарка

Алгоритмические оптимизации

Размеры данных

- Цены / волатильности 5 акций - **480 МБ (RAM)**
- Функция выплаты опциона – **96 МБ (RAM)**
- Функция выплаты опциона в точке – **400 КБ (L2)**

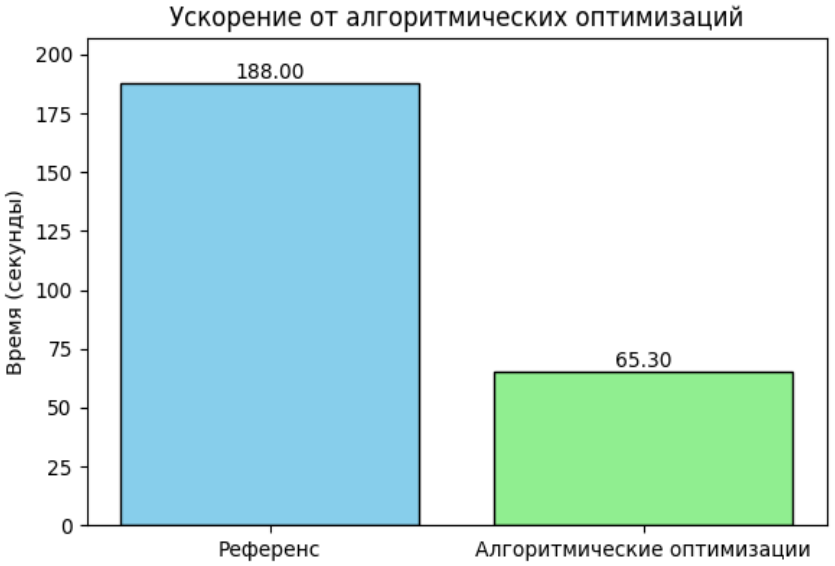
Оптимизация использования памяти

- Достаточно 2 минимальных цен для вычисления функции выплаты опциона – **192 МБ**, уменьшение требуемого объема памяти - **2.5x**

Оптимизация вычислений греков

- Изменение начальной цены не требует повторного вычисления Монте-Карло траектории цены/волатильности
- Изменение процентной ставки не требует повторного вычисления Монте-Карло траектории волатильности

Ускорение относительно референса на RISC-V– 2.89x



RISC-V: 64 ядра, 2.6 ГГц, L1: 128 КБ, L2: 2 МБ, L3: 64 МБ, RAM 135 ГБ,
clang 17.0.6, openEuler 24.03
clang -O3 -march=rv64gcv_zvl128b -ffast-math

Реализация STAC-A2 бенчмарка

Векторизация

RVVMF – векторная библиотека математических функций для RISC-V Vector Extension 1.0

Точность:
0.501 ULP

Производительность:
до 2x быстрее векторных альтернатив для выбранных функций

Расширяющийся список функций:
exp, log, cos, tan, atan, pow...

Тестовая система:
проверка требований стандартов IEEE-754, ISO/IEC 10967-2, производительности (пропускная способность и задержка)

В анализе ниже использована экспериментальная версия RVVMF, обеспечивающая точность 3.5 ULP

Реализация STAC-A2 бенчмарка

Векторизация

24 из 29 циклов в бенчмарке автовекторизованы
Инкрементальное ускорение из-за

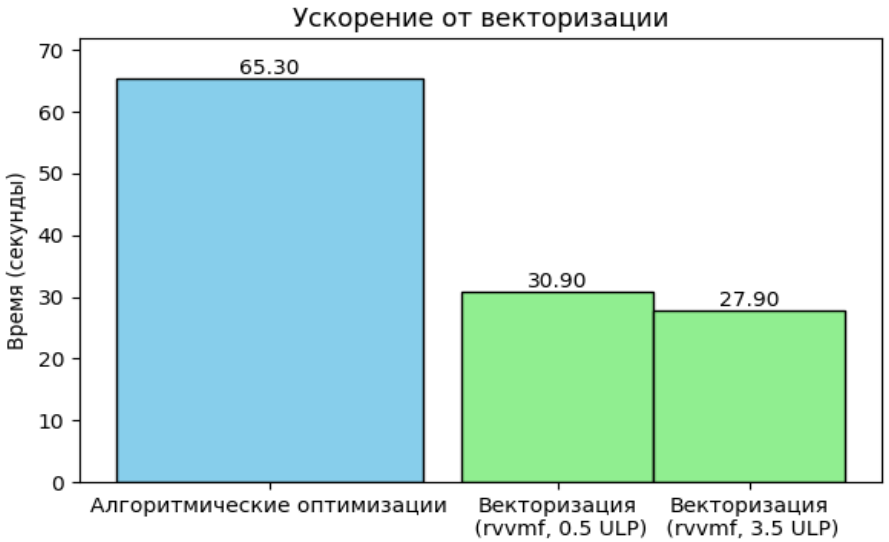
RVVMF (3.5 ULP exp(), log()), intrinsic sqrt(): 1.22x

OpenBLAS (матричное умножение, решение системы линейных уравнений): 1.20x

генератора мультивариантного нормального распределения на базе MT19937: 1.30x

векторизации кода бенчмарка: 1.21x

Ускорение относительно алгоритмических оптимизаций на RISC-V – 2.34x



RISC-V: 64 ядра, 2.6 ГГц, L1: 128 КБ, L2: 2 МБ, L3: 64 МБ, RAM 135 ГБ,
clang 17.0.6, openEuler 24.03
clang -O3 -march=rv64gcv_zvl128b -ffast-math

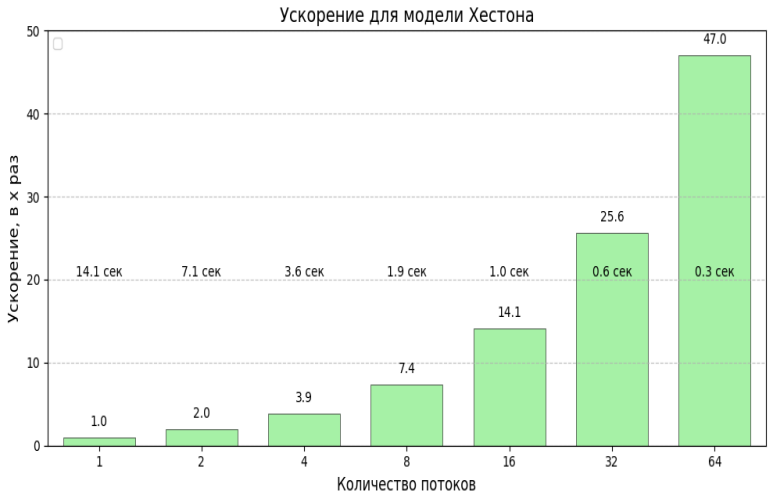
Дополнительные 7% ускорения из-за миграции на экспериментальную версию RVVMF пониженной точности (~0.5 -> ~3.5 ULP) без потери точности результатов бенчмарка

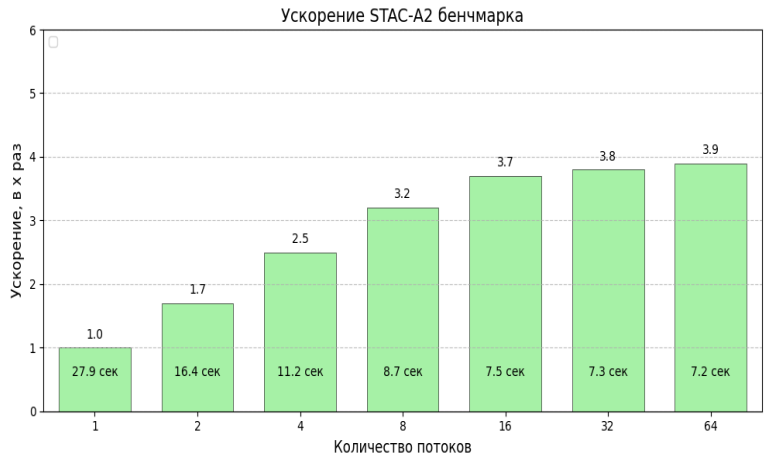
Реализация STAC-A2 бенчмарка

Параллелизация

Монте-Карло симуляции рынка акций (модель Хестона), использующие RVVMF и генератор мультивариантного нормального распределения

MT19937 сконфигурирован для поддержки параллельных вычислений в Монте-Карло симуляциях рынка акций





Ускорение относительно векторизации на RISC-V– 3.9x

RISC-V: 64 ядра, 2.6 ГГц, L1: 128 КБ, L2: 2 МБ, L3: 64 МБ, RAM 135 ГБ, clang 17.0.6, openEuler 24.03, OpenMP 5.1
clang -O3 -march=rv64gcv_zvl128b -ffast-math -fopenmp

Реализация STAC-A2 бенчмарка

Валидация

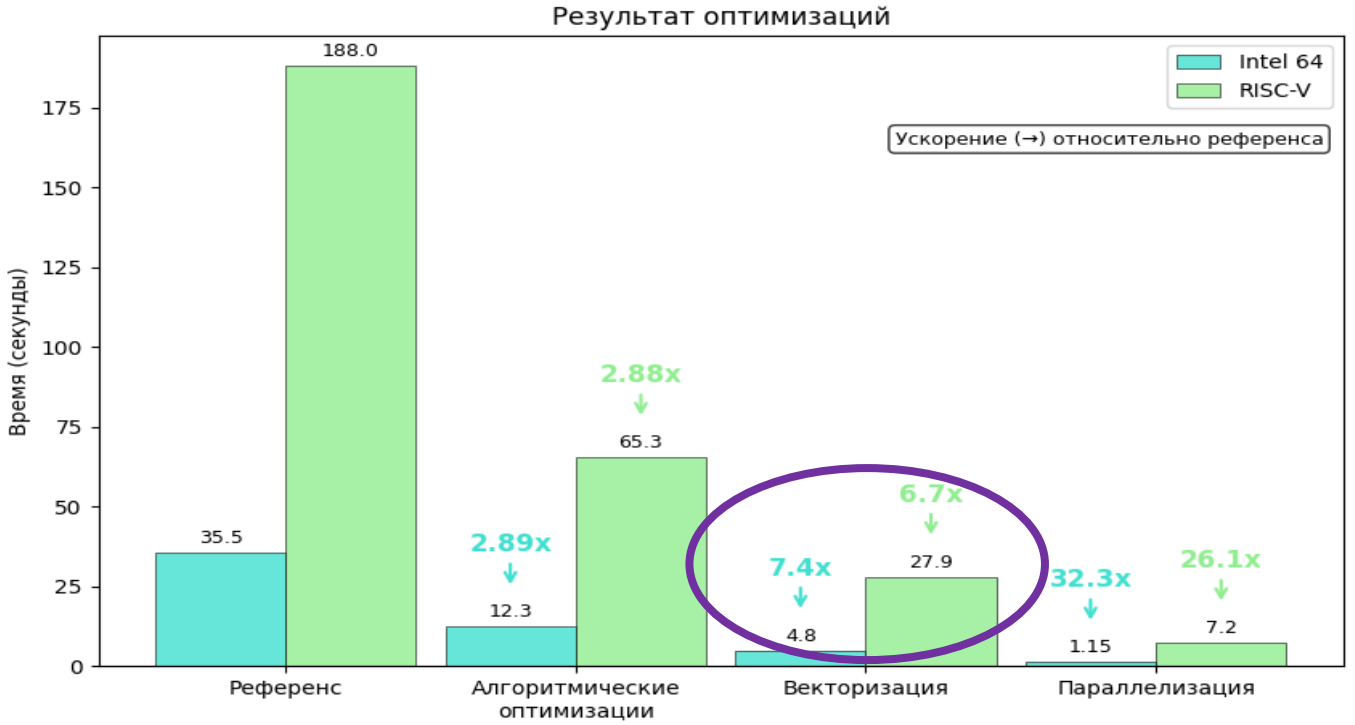
Этап валидации	Способ валидации
Вычисление греков	Сравнение с аналитическими значениями в модели Блэка Шоулса Построение доверительного интервала
Вычисление цены P put опциона Американского типа на самый дешевый актив на рынке акций	Построение доверительного интервала Опцион Европейского типа: сравнение с аналитической формулой Хестона Опцион Американского типа: сравнение с опубликованными эталонными в литературе
Монте-Карло симуляции рынка акций	Распределение волатильности

Реализация STAC-A2 бенчмарка

Первые итоги оптимизаций для реализаций RISC-V и Intel 64

- Разработаны и оптимизированы бенчмарки для реализаций архитектур RISC-V и Intel 64
- Бенчмарки успешно прошли валидацию

- Потенциал дополнительного ускорения из-за генератора мультивариантного нормального распределения, замены MT19937 на другой, оптимизаций в бенчмарке



RISC-V: 64 ядра, 2.6 ГГц, L1: 64 КБ, L2: 2 МБ, L3: 64 МБ, clang 17.0.6, openEuler 24.03, OpenBLAS 0.3.28, OpenMP 5.1
clang: -O3 -march=rv64gcv_zvl128b -ffast-math -fopenmp

Intel 64: 2x16 ядер, HT=on, 2.9 ГГц, L1: 48 КБ, L2: 1,25 МБ, L3: 24 МБ, clang 16.0.6, Fedora 38, Intel MKL 2025.0, OpenMP 5.0
clang -O3 -march=native -ffast-math –fopenmp

Параллелизация показана для 64 потоков на RISC-V и для 32 потоков на Intel64
Нормировка на частоту ЦПУ отсутствует

Сравнительный анализ реализаций RISC-V и Intel64

Последовательная версия бенчмарка

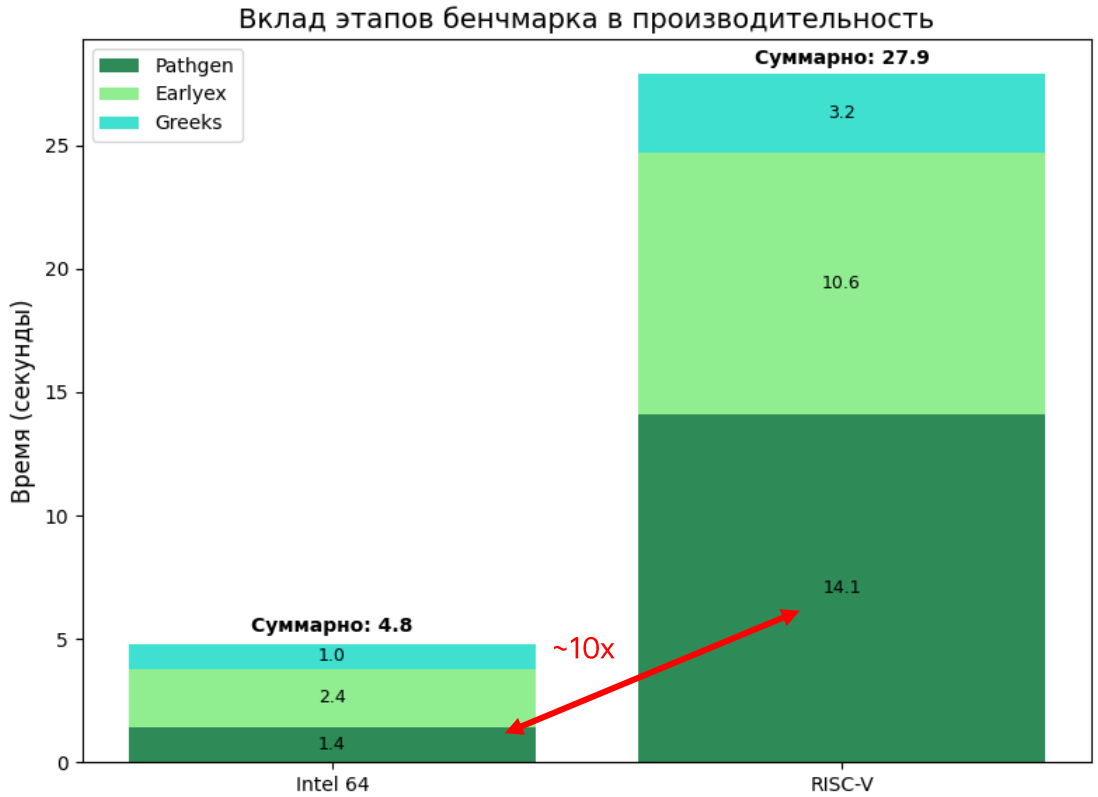
На базе модели ‘max (compute_t, data_transfer_t)’

Пропускная способность подсистем памяти*

векторные	L1, ГБ/с	L2, ГБ/с	L3, ГБ/с	RAM, ГБ/с
RISC-V load/store	41/22	17/9	8/4	5/3
Intel 64 load/store	189/189	61/61	26/26	13/13
Intel 64 / RISC-V load/store	4.6/8.6x	3.6/6.7x	3.25/6.5x	2.6/4.3x

Пропускная способность векторных инструкций**

векторные	RISC-V, ГОп/с	Intel 64, ГОп/с	Intel 64 / RISC-V
vfmadd	10.4	46.4	4.46x
vfmul	10.4	46.4	4.46x
vfsub	10.4	46.4	4.46x
vfdiv	0.26	1.45	5.57x



Разница ~5x (из 10x) в производительности Pathgen на реализациях RISC-V и Intel 64 – из-за субоптимальной производительности генератора мультивариантного нормального распределения.

* Модифицированная версия STREAM Copy

** Собственные бенчмарки для измерения производительности инструкций

Сравнительный анализ реализаций RISC-V и Intel64

Pathgen – compute bound

~34 вычислительные инструкции на 1 выгрузку результатов (без учета вычислений в генераторе и др.)

Промежуточные вычисления используют регистры

Пример типичных вычислений

```
vdouble v4_vec = vfmadd(v1_vec, v2_vec, v3_vec);
vdouble v5_vec = vfmul(v3_vec, v1_vec);
vdouble v6_vec = vfdiv(v1_vec, v2_vec);
```

Дополнительный паттерн вычислений в виде зависимых итераций и зависимых инструкций внутри итерации – в MT19937

Отношение скоростей вычислений на RISC-V и на Intel 64

~ 4.9x (характеристики инструкций – на предыдущем слайде, учтены частоты ЦПУ; для MT19937 – аналогичный анализ на базе задержек используемых в генераторе инструкций)

Разница 5x (из 10x) в производительности Pathgen – из-за скорости инструкций в реализациях архитектур RISC-V и Intel 64

EarlyEx и Greeks – memory-bound

Пример типичных вычислений

```
vdouble v1_vec = vload(&v1);
v1_vec         = vfmul(v1_vec, const);
v1             = vstore(v1_vec);
```

Отношение скоростей вычислений одной итерации EarlyEx (данные в L2) на RISC-V и на Intel 64 ~5.7x

Отношение скоростей вычислений Greeks (данные в RAM) на RISC-V и на Intel 64 ~3.8x

(характеристики подсистем памяти – на предыдущем слайде, учтены частоты ЦПУ)

Разница в производительности EarlyEx / Greeks – из-за скорости подсистем памяти в реализациях архитектур RISC-V и Intel 64

Выводы

- Доступна первая реализация STAC-A2 для RISC-V
- Доступна библиотека RVVMF - критический ингредиент ПО экосистемы RISC-V, обеспечивающий требования по скорости и точности. Предназначена для эффективного решения спектра задач, включая анализ рыночных рисков
- Сформирован достаточный набор функционала спецификаций RISC-V для задач анализа финансовых рисков

СЛЕДУЮЩИЕ ШАГИ

- Дополнительные оптимизации бенчмарка
- Уточнение сравнительного анализа реализаций архитектур для последовательной версии бенчмарка
- Сравнительный анализ реализаций архитектур для параллельной версии бенчмарка

Спасибо за внимание

Оцените, пожалуйста, доклад