Линейная алгебра как путь к высокопроизводительному анализу данных: программные и аппаратные средства, их проблемы и возможности платформы RISC-V

Семён Григорьев

Санкт-Петербургский Государственный Университет

20 сентября 2024г.

Семён Григорьев

- Область интересов: алгоритмы анализа графов, высокопроизводительная разреженная линейная алгебра, теория формальных языков и её применение для анализа данных

- Доцент кафедры системного программирования СПбГУ
- Почта: rsdpisuy@gmail.com
- ТГ: @rsdpisuy

Разреженная линейная алгебра

- Линейная алгебра: матрицы, вектора, операции над ними
  - Операции естественным образом распараллеливаются по данным: эффективные реализации для многоядерных систем, GPGPU, и т.д.
  - Абстракция по операциям: (полу)кольца, моноиды, ...
- **Разреженная** линейная алгебра: в матрице или векторе много одинаковых элементов
  - Часто говорят что в матрице (векторе) много «нейтральных элементов», «нулей» или что-то подобное, но это не всегда так
- Хотим не хранить одинаковые элементы
  - Специальные структуры для хранения матриц и векторов[1]
  - Специальные алгоритмы для выполнения операций[2]

[1] COO, CSR, CSC, DCSR, Quad-Tree, ...

[2] Не забываем про параллельность

Области применения

- Машинное обучение
  - Разреженное внимание (sparse attention)
  - Графовые нейронные сети
- Робототехника
  - Задачи навигации
  - ...
- Численные методы
  - Разреженные системы уравнений
  - ...
- ...

- Анализ графов
  - Графовые базы данных
  - Анализ социальных, банковских и других сетей
  - Статический анализ кода
  - Биоинформатика

GraphBLAS

Стандарт C API, определяющий набор примитивов обобщённой разреженной линейной алгебры и операций над ними

- Домашняя страница: https://graphblas.org/
- Коллекция полезных ссылок по теме (GraphBLAS-Pointers): https://github.com/GraphBLAS/GraphBLAS-Pointers
- Эталонная реализация на C (SuiteSparse:GraphBLAS): https://github.com/DrTimothyAldenDavis/GraphBLAS
- Обёртка для Python (python-graphblas): https://github.com/python-graphblas/python-graphblas
- Для распределённой обработки (CombBLAS): https://github.com/PASSIONLab/CombBLAS
- Графовая база данных (FalkorDB): https://github.com/FalkorDB/falkordb

Специализированные решения для разреженной линейной алгебры

- Dedicated Hardware Accelerators for Processing of Sparse Matrices and Vectors: A Survey, 2024 год
- A Survey of Accelerating Parallel Sparse Linear Algebra, 2023 год
- A Systematic Literature Survey of Sparse Matrix-Vector Multiplication, 2024 год
- GraphLily (подмножество GraphBLAS на FPGA): https://github.com/cornell-zhang/GraphLily

А что про RISC-V?

- Идёт работа над расширениями[3]
  - IndexMAC: A Custom RISC-V Vector Instruction to Accelerate Structured-Sparse Matrix Multiplications, 2024 год
  - Optimizations for Very Long and Sparse Vector Operations on a RISC-V VPU: A Work-in-Progress, 2023 год
- Vortex
  - RISC-V GPGPU: https://github.com/vortexgpgpu/vortex
  - Молодой, перспективный, ...

[3] Оставим в покое RVV, Integrated Matrix Extension, XuanTie Matrix Extension

Заключение

- Высокопроизводительная разреженная линейная алгебра ⇒ высокопроизводительные приложения
  - Машинное обучение
  - Графовые базы данных
  - Анализ социальных, банковских и других сетей
  - Анализ кода
  - ...
- Сделать **разреженную** линейную алгебру высокопроизводительной **сложно**
  - Нерегулярный доступ к данным
  - Хорошая алгебра — обобщённая алгебра
  - Сложности с балансировкой нагрузки
  - ...
- Но люди пытаются
  - Даже институты для этого создают[4]

[4] Sparsitute: A mathematical Institute for Sparse Computations in Science and Engineering, https://sparsitute.lbl.gov/