Motivation behind developing a compiler for DataFrame

Jan. 8, 2025

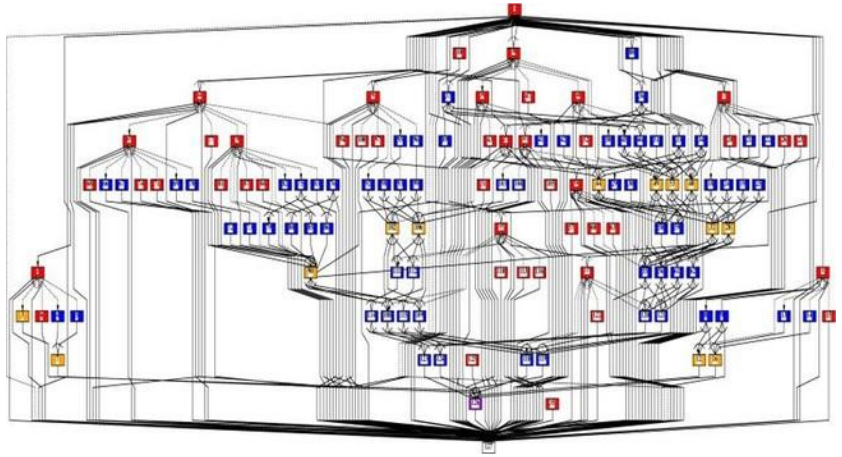
Kazuhisa Ishizaka, NEC

Kazuhisa Ishizaka

Primary author of FireDucks

Background:

- Automatic parallelizing compiler (Ph.D)
- Parallel processing for manycore processor
- Software for vector supercomputer
  - TensorFlow-VE
  - LLVM-VE compiler

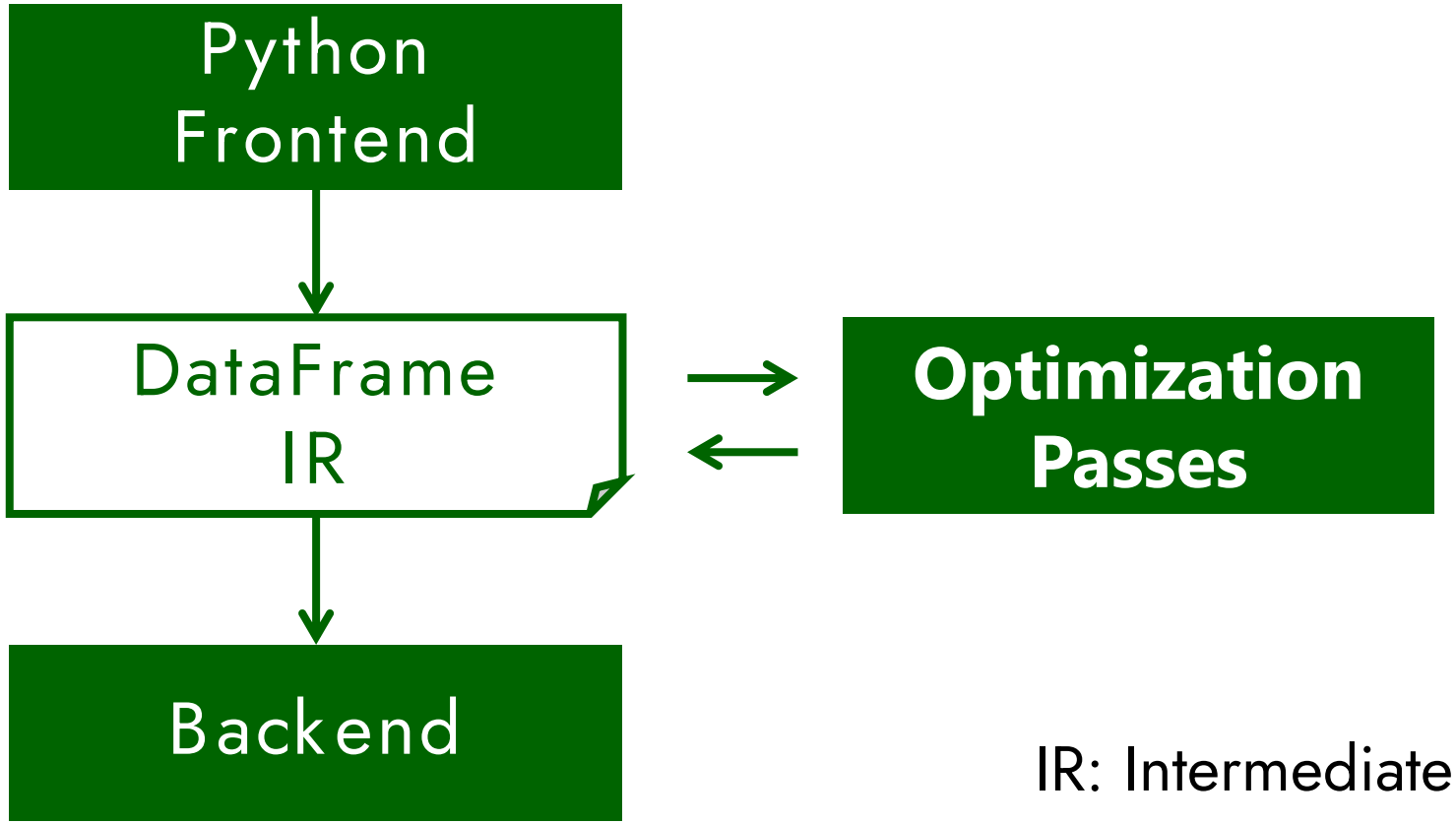


OSCAR Parallelizing Compiler

NEC SX-Aurora TSUBASA
(Supercomputer)

https://prtimes.jp/main/html/rd/p/000000036.000027784.html
https://jpn.nec.com/hpc/sxauroratsubasa/specification/index.html

FireDucks: DataFrame Compiler



IR: Intermediate Representation

Architecture of FireDucks

Background in 2021

Needs for Speed
in Data Science

Evolution of
Compiler Technology

* Development of FireDucks started in 2021

Background in 2021

Needs for Speed
in Data Science

Evolution of
Compiler Technology

Data preparation in Data Science

https://x.com/BigDataBorat/status/306596352991830016

Beyond pandas

Wes McKinney

Apache Arrow and the "10 Things I Hate About pandas"

PANDAS APACHE ARROW

AUTHOR
Wes McKinney

Sep. 21, 2017

1. Internals too far from "the metal"
2. No support for memory-mapped datasets
3. Poor performance in database and file ingest / export
4. Warty missing data support
5. Lack of transparency into memory use, RAM management
6. Weak support for categorical data
7. Complex groupby operations awkward and slow
8. Appending data to a DataFrame tedious and very costly
9. Limited, non-extensible type metadata
10. Eager evaluation model, no query planning
11. "Slow", limited multicore algorithms for large datasets

https://wesmckinney.com/blog/apache-arrow-pandas-internals/

Beyond pandas

Wes McKinney

Sep. 21

Apache Arrow

https://arrow.apache.org/

- Core library for high performance data science
  - Columnar memory format and operations implemented in C++
- PyArrow: python binding, but different API from pandas

ent

y

11.“Slow”, limited multicore algorithms for large datasets

https://wesmckinney.com/blog/apache-arrow-pandas-internals/

Beyond pandas

Wes McKinney

Sep. 21

ent

y

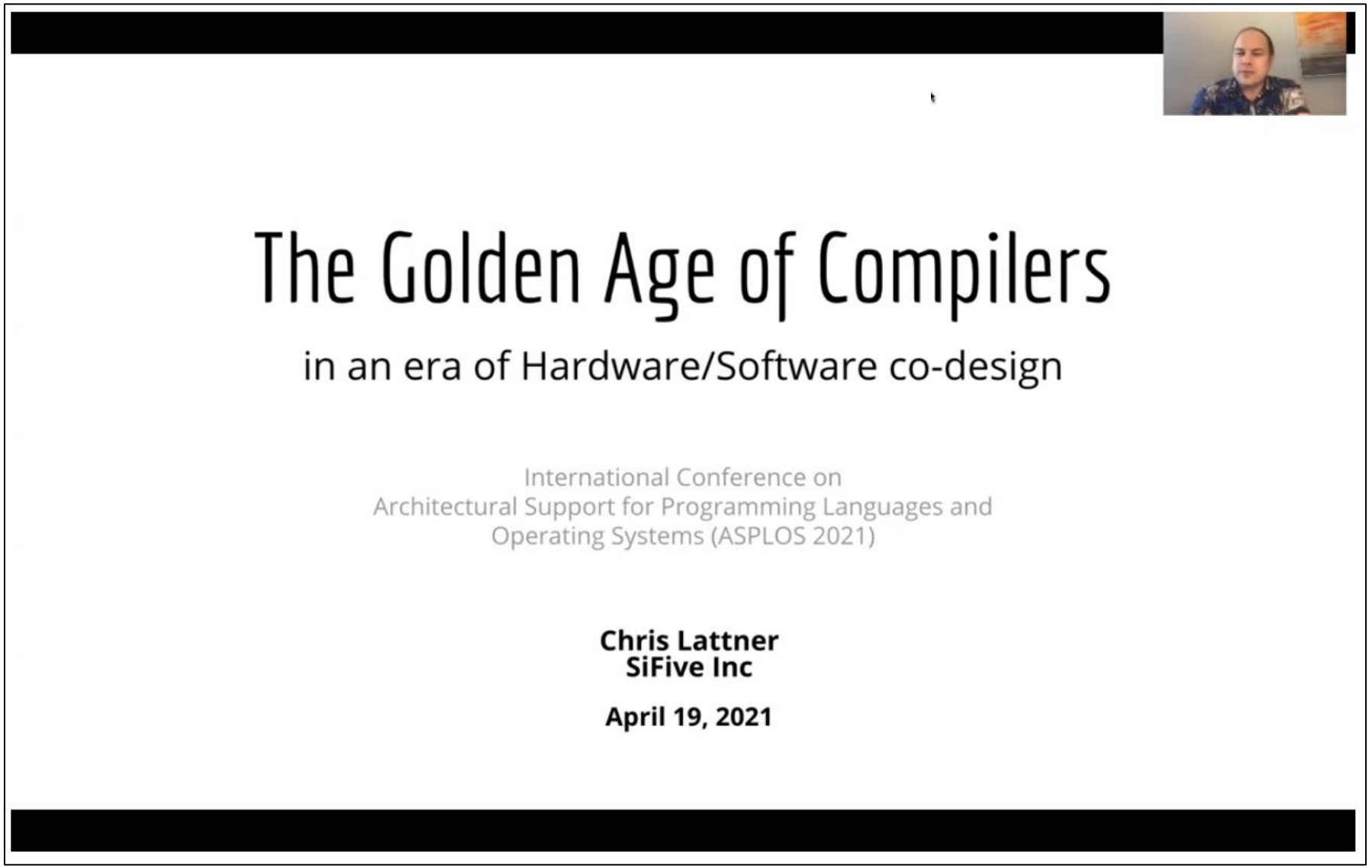
11.“Slow”, limited multicore algorithms for large datasets

Our approach

Solving by compiler technologies
without changing well-used pandas API

https://wesmckinney.com/blog/apache-arrow-pandas-internals/

The Golden Age of Compilers

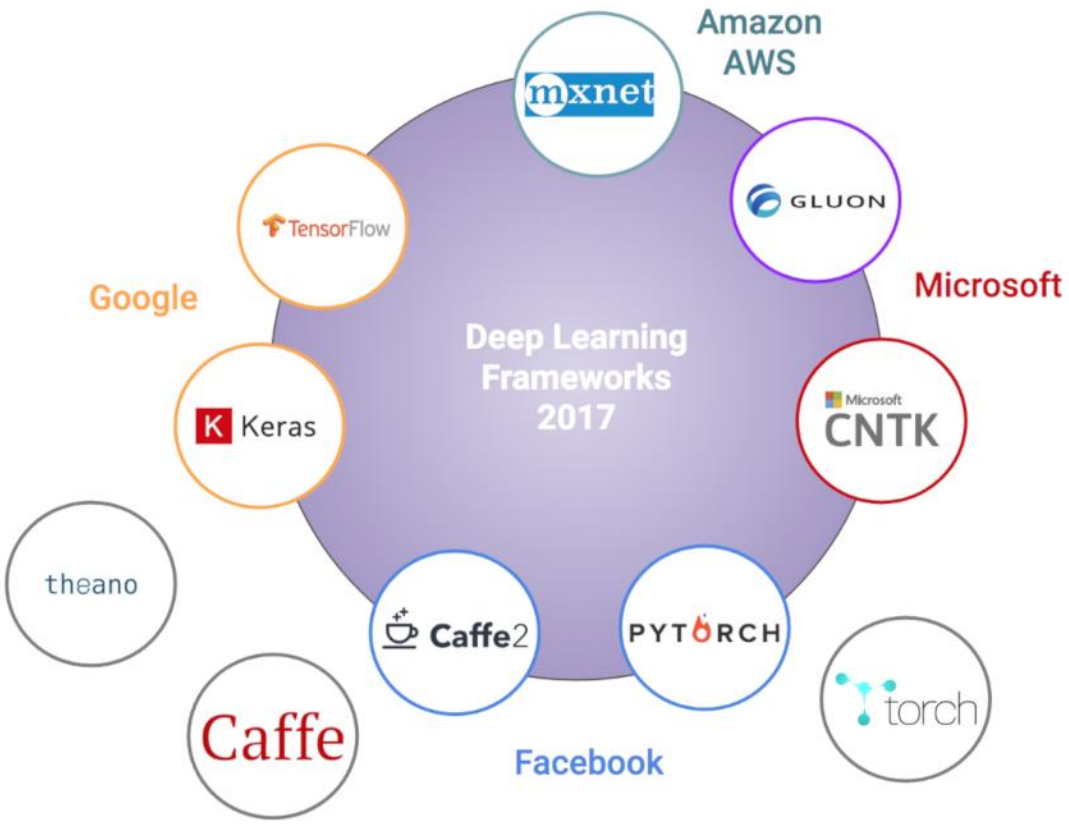


Chris Lattner, ASPLOS 2021 Keynote

https://www.youtube.com/watch?v=4HgShra-KnY

Acceleration of Deep Learning

Deep Learning Frameworks



Deep Learning HW

GPU TPU

CPU(SIMD)

Accelerators(xPU)

Deep Learning Compilers

Model optimization and hardware adaption

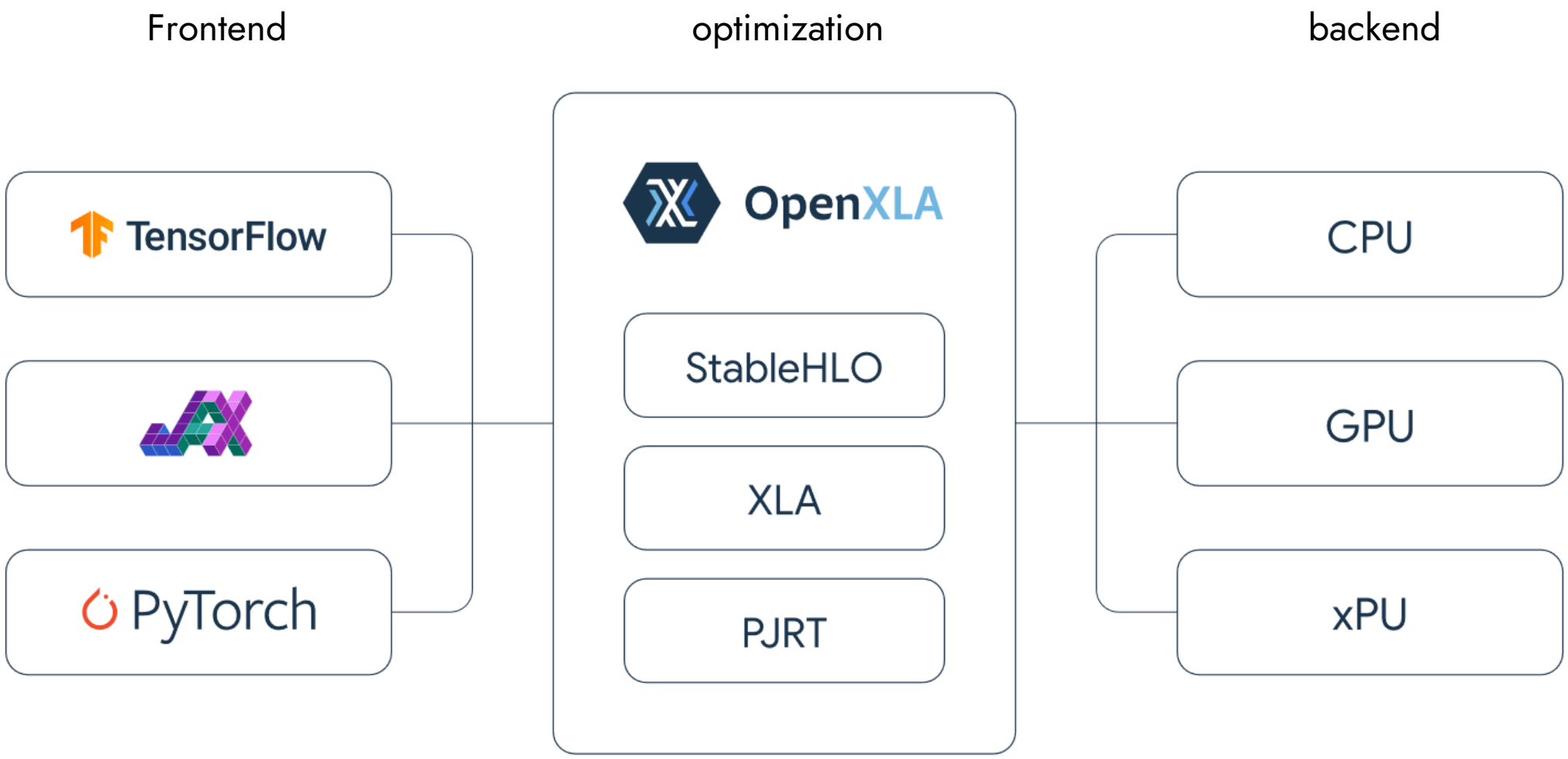
- Representing DL model as hardware-independent IR
- Applying DL-specific optimizations
- Executing optimized IR using deep learning kernels for a hardware

Table 1. The comparison of DL compilers, including TVM, nGraph, TC, Glow, and XLA.

		TVM	nGraph	TC	Glow	XLA
	Developer	Apache	Intel	Facebook	Facebook	Google
Frontend	Programm-ing	Python/C++ Lambda expression	Python/C++ Tensor expression	Python/C++ Einstein notation	Python/C++ Layer programming	Python/C++ Tensorflow interface
Frontend	ONNX support	✓ tvm.relay.frontend	✓ Use ngraph-onnx	×	✓ ONNXModelLoader	✓ Use tensorflow-onnx

M. Li et al., "The Deep Learning Compiler: A Comprehensive Survey," in IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 3, pp. 708-727, 1 March 2021, doi: 10.1109/TPDS.2020.3030548.

OpenXLA: Machine Learning Compiler



https://github.com/openxla/xla

LLVM and MLIR

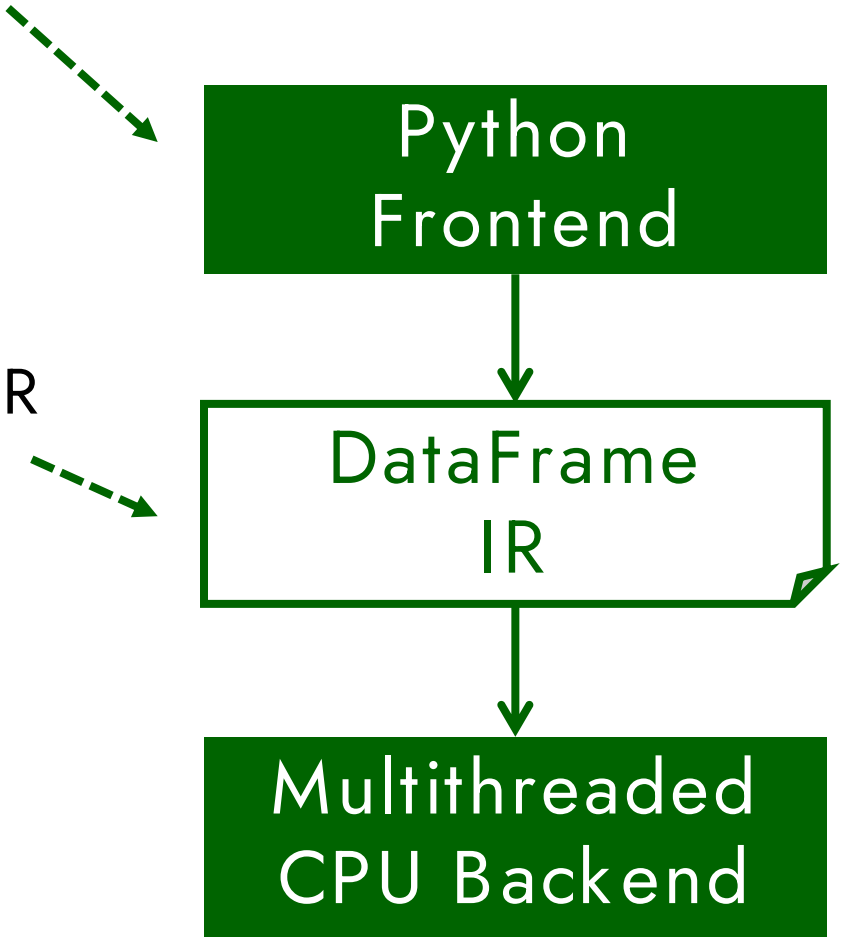
- De facto standard of compiler infrastructure
- Started as academic project at UIUC
- Used in many OSS and productions

MLIR

- Sub project of LLVM
- Used in OpenXLA
- Framework to define a compiler IR

Use of LLVM/MLIR in FireDucks

Frontend generates DataFrame IR from pandas API



Defined by MLIR

LLVM/MLIR Framework helps to implement DataFrame specific optimizations

Executing DataFrame IR by multithreaded kernels on top of Apache Arrow

Motivation behind FireDucks

Needs for Speed
in Data Science

Evolution of
Compiler Technology

Motivation for Future FireDucks

You

Use FireDucks in you projects and give us your feedback!

Apache Arrow

- Core library for high performance data science
- Columnar memory format and operations implemented in C++
- Used by many projects including Apache Spark, Dask, Polars, cudf, etc.

PyArrow

- Python binding of Apache Arrow
- Deferent API from pandas

APACHE ARROW

https://arrow.apache.org/