The HDF Group: Thought Leader in Exascale Computing

Takes on Real-Time PCAP Ingestion, Storage, and Analytics

June 5, 2017

David Pearah, CEO, The HDF Group

The HDF Group

Who is the HDF Group?

HDF Group has developed open source solutions for Big Data challenges for nearly 30 years

Small company (~ 40 employees) with focus on High Performance Computing and Scientific Data

Offices in Champaign, IL + Boulder, CO

Our flagship platform – HDF5 – is at the heart of our open source ecosystem.

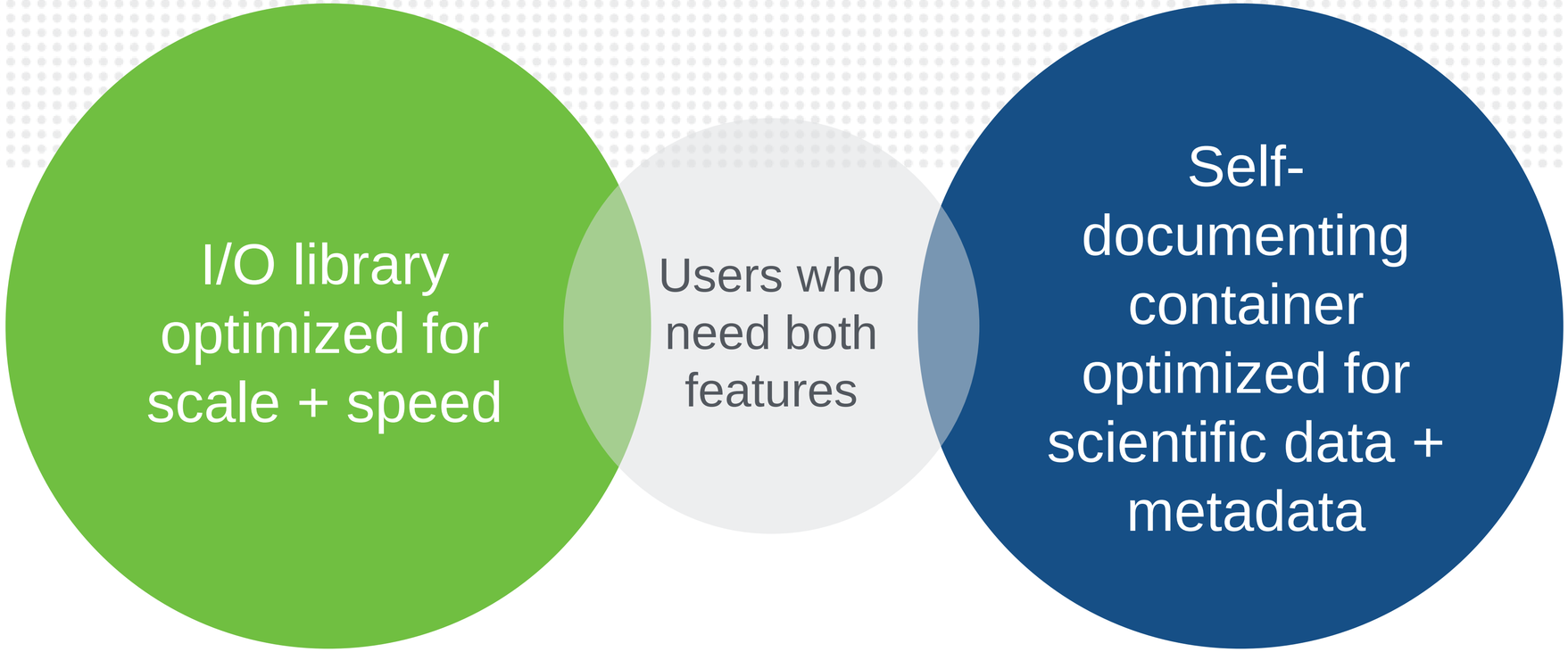
Tens of thousands use HDF5 every day, as well as build their own solutions (~~600~~ ~~700~~ 800+ projects on Github)

“De-facto standard for scientific computing” and integrated into every major analytics + visualization tool

What does the HDF Group do?

Products	• HDF Capture: Software solution for PCAP Ingest + Storage (Beta) • HDF5 Library • Connectors: ODBC + Cloud (Beta) • Add-Ons: compression + encryption
Support	• HDF Support Packages (Basic + Pro + Premier) • Support for h5py + PyTables + pandas (NEW) • Training
Consulting	• HDF: new functionality + performance tuning for specific platforms • General HPC software engineering with fintech expertise (ex. MPI implementation for back testing) • Metadata science and expert services

Why Use HDF5?


I/O library optimized for scale + speed
Users who need both features
Self-documenting container optimized for scientific data + metadata

Why did we create HDF Capture?

Approached by PCAP appliance vendor (confidential) with Big Data challenge

- Many of their clients want to store, index, and query all the data that has ever passed through the appliance: data lake of every single transaction
- The appliance vendor supplies an API with connectors to popular solutions: Splunk, Kafka, Hadoop, etc.

However, their clients kept running into challenges

- Forced to sample or select subset of data: database or log solutions couldn't keep up
- Transformation of data from PCAP into other format meant that regulatory compliance couldn't be met
- Inability to "replay" the market: load testing of software, backtesting, etc.

Appliance vendor familiar with our work in particle physics, and asked HDF Group to create a turnkey solution

- Lots of data at once with perfect fidelity of content and long-term storage
- Challenge: build a solution that could ingest + store + index 500K messages per second

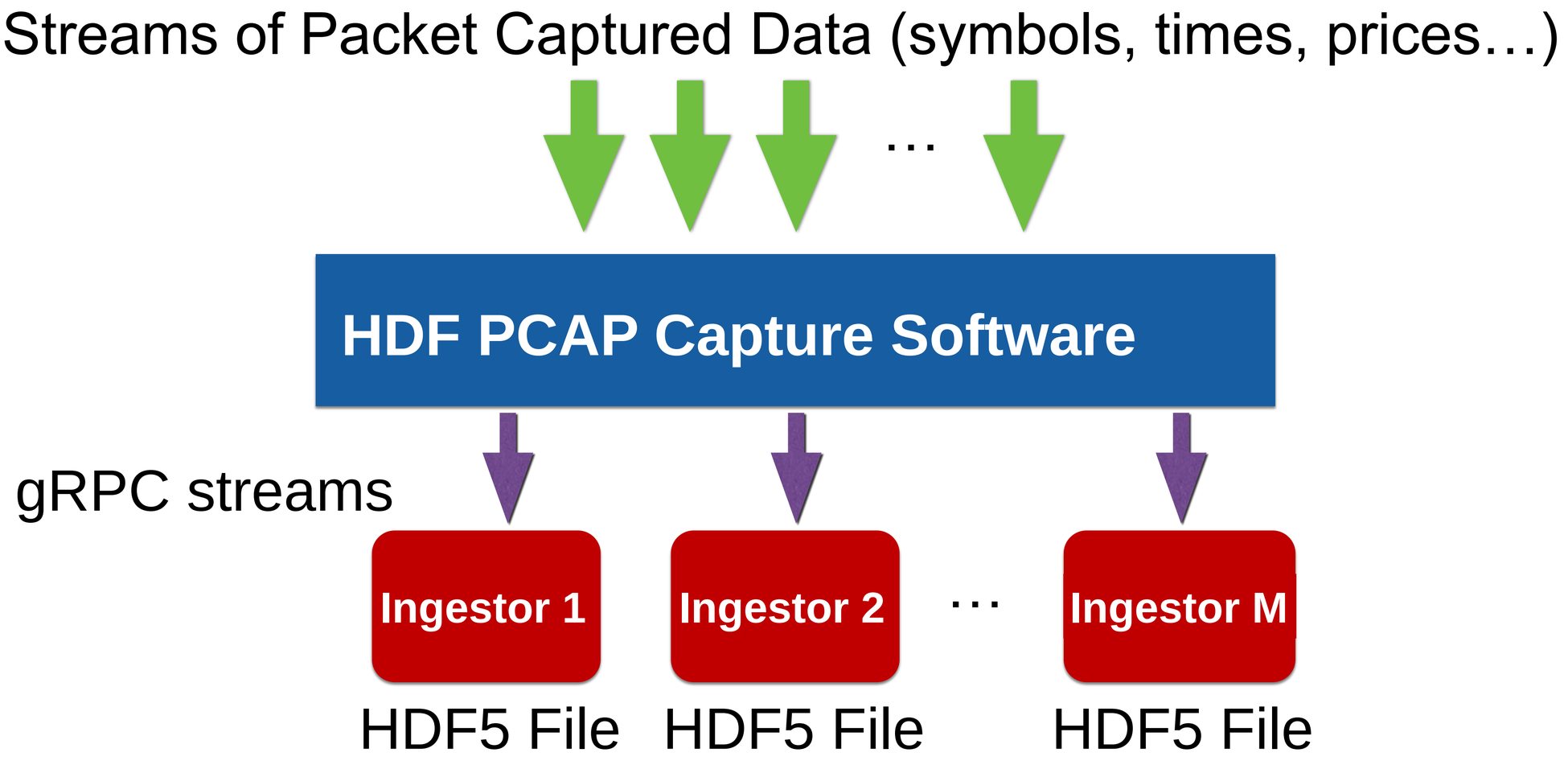
HDF5 PCAP Capture Software

- A client subscribed to ~10 streams full of financial PCAP data only needs a pipe with < 100 Mbps for ingesting all the data.

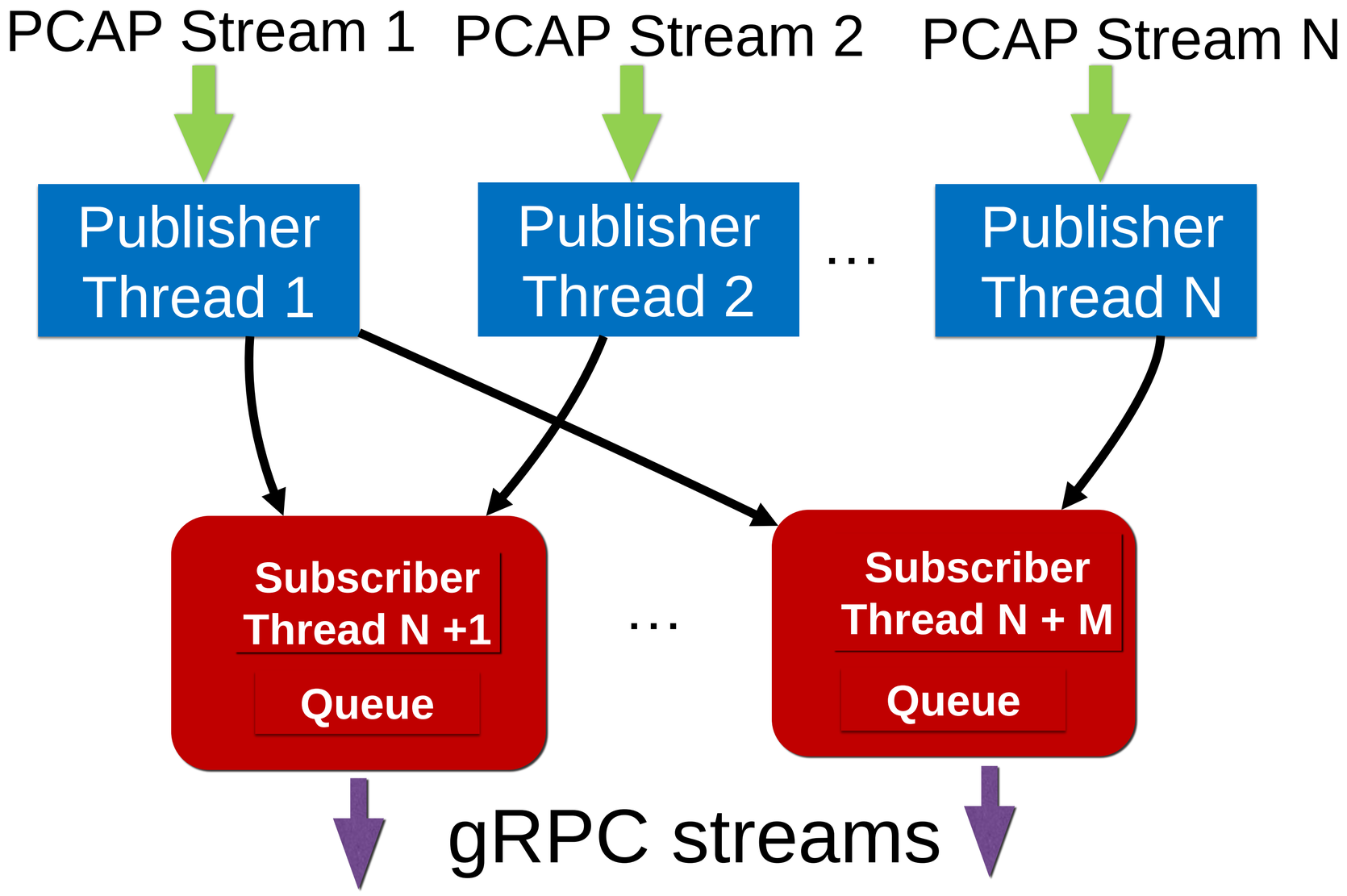
- Using compression allowed a reduction of ~5x, allowing to decrease the mean length of messages from ~100 bytes to ~20 bytes per message.

- The HDF5 ingestion clients can achieve more than 500K mess/sec (~650K mess/sec by using a single machine with 16 physical cores).

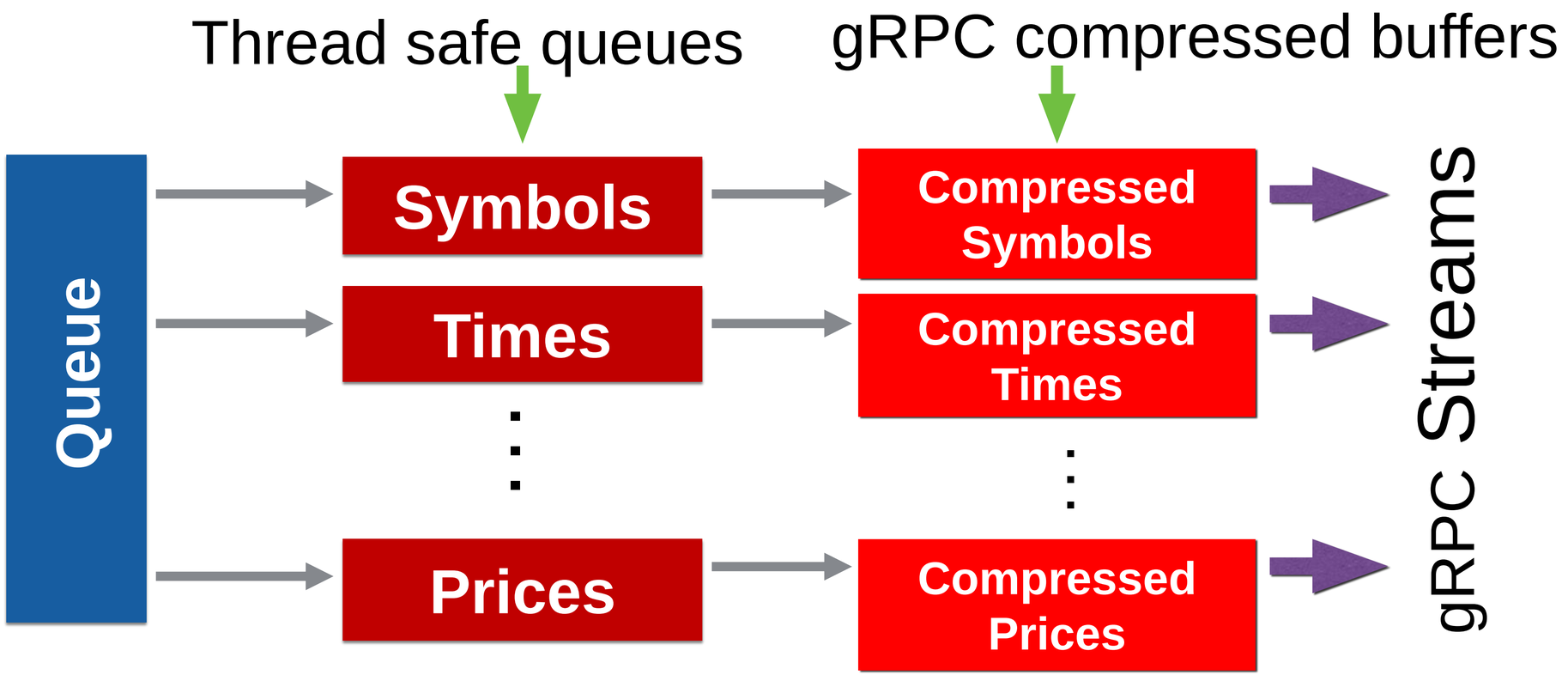
Global View


Streams of Packet Captured Data (symbols, times, prices…)
…
HDF PCAP Capture Software
gRPC streams
Ingestor 1
Ingestor 2
…
Ingestor M
HDF5 File
HDF5 File
HDF5 File

HDF5 Multi Stream Software



Detail of the Queue



- Every thread safe queue is compressed after being filled up for less memory consumption and faster transmission.

Questions? Comments?

www.hdfgroup.org

Dave Pearah
CEO
David.Pearah@hdfgroup.org

Dax Rodriguez
Director of Commercial Services and Solutions
Dax.Rodriguez@hdfgroup.org