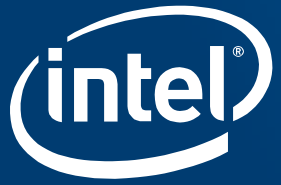


# Accelerating Data Analytics on Intel FPGAs

Mike Strickland, Director, Data Center Solution Architect
Intel Programmable Solutions Group
July 2017

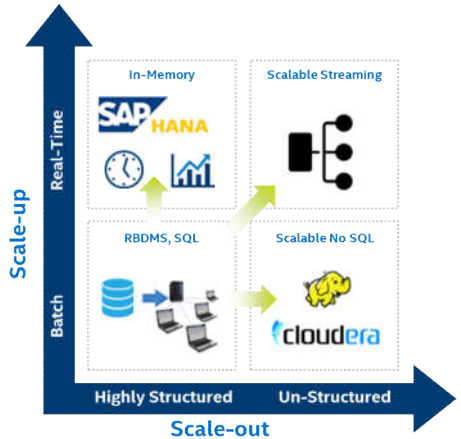
# Accelerate Big Data Analytics with Intel Frameworks and Libraries with FPGA's

1. **Intel Big Data Analytics Frameworks & API Libraries**
   Accelerate innovation in Big Data Analytics with frameworks built on Software Defined Infrastructure with open standard building blocks.

2. **Intel Frameworks & Libraries integrated with FPGAs**
   Run unmodified customer applications, use runtime orchestration with both Xeon ® and FPGA support, and leverage end to end virtualization & security.

3. **Accelerate Relational, NoSQL, and Un-Structured**

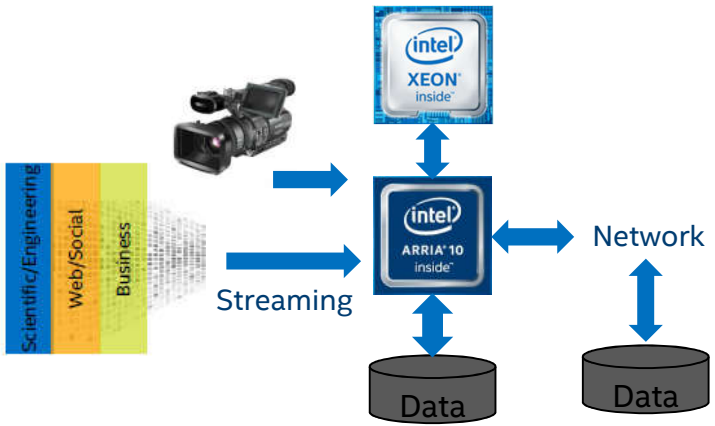
   FPGA data access, networking, and algorithm acceleration options with a single FPGA for highly structured, semi-structured, and un-structured data for better TCO, flexibility, and future proofing.




**Analytics** Landscape and Scaling

# FPGAs Offer Unique Value for Analytics/Streaming

## Single Multi-function Accelerator



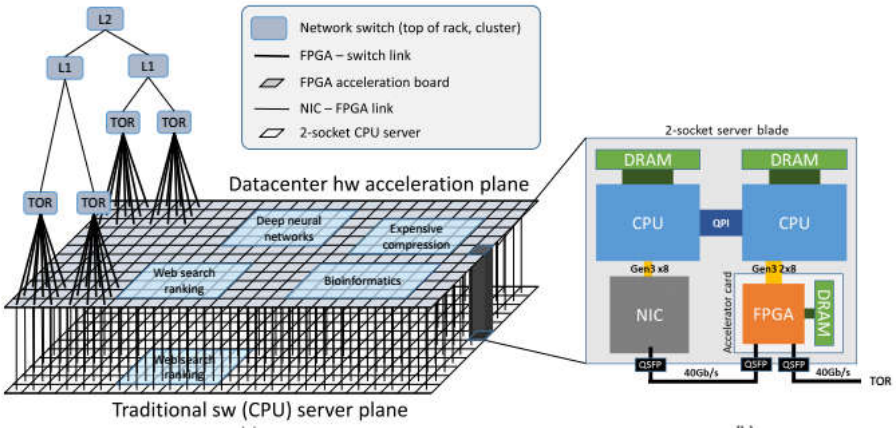

## Integrate to Intel Frameworks & API's

- Run unmodified customer applications
- Orchestration run time advantage: Intel® Xeon® processors or Intel® FPGAs
- End to End Security & Virtualization framework

## Significant Acceleration

- PCIe lookaside acceleration
- Networking + streaming + data access
  - Inline acceleration and protocol acceleration
- Compression, filtering, encryption
- Fast lookups/hashing

# Multi-Function FPGA: Algorithm + Networking + Data Access Acceleration



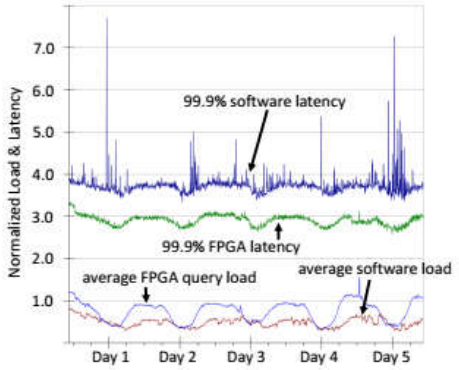




Fig. 7. Five day query throughput and latency of ranking service queries running in production, with and without FPGAs enabled.

## Microsoft Scale Out FPGA Multi-Function Accelerator

- "Diversity of cloud workloads and … rapid … change" (weekly or monthly)
  - Search, SmartNIC, Machine Learning, Encrypt, Compress, Big Data Analytics,…
- Lower & predictable latency for ranking vs. software
- FPGA CNN perf. to Peak TFLOPs ratio similar to gpu
  - MSFT Hot Chips presentation (Sept 2015)

Source: Microsoft

# The power of deep learning on FPGA
## from Microsoft Build 2017 Conference

**Performance**

- Tens to hundreds of TOPS of effective inference throughput at low batch sizes
- Ultra-low latency serving on modern DNNs
- >10X better than CPUs and GPUs
- Scale to many FPGAs in single DNN service

**Flexibility**

- FPGAs ideal for adapting to rapidly evolving ML
- CNNs, LSTMs, MLPs, reinforcement learning, feature extraction, decision trees, etc.
- Inference-optimized numerical precision
- Custom binarized, ternarized, tiny precision nets
- Sparsity, deep compression for larger, faster models

**Scale**

- Microsoft has the world's largest cloud investment in FPGAs
- Multiple Exa-Ops of aggregate AI capacity
- We have built powerful DNN serving platform on our FPGA fabric

Source: Microsoft

# Ecosystem Partner Solutions & POCs using Intel® FPGAs

## SQL over Relational open databases

- Traditional Data Warehousing Acceleration
- Real Time Analytics Acceleration

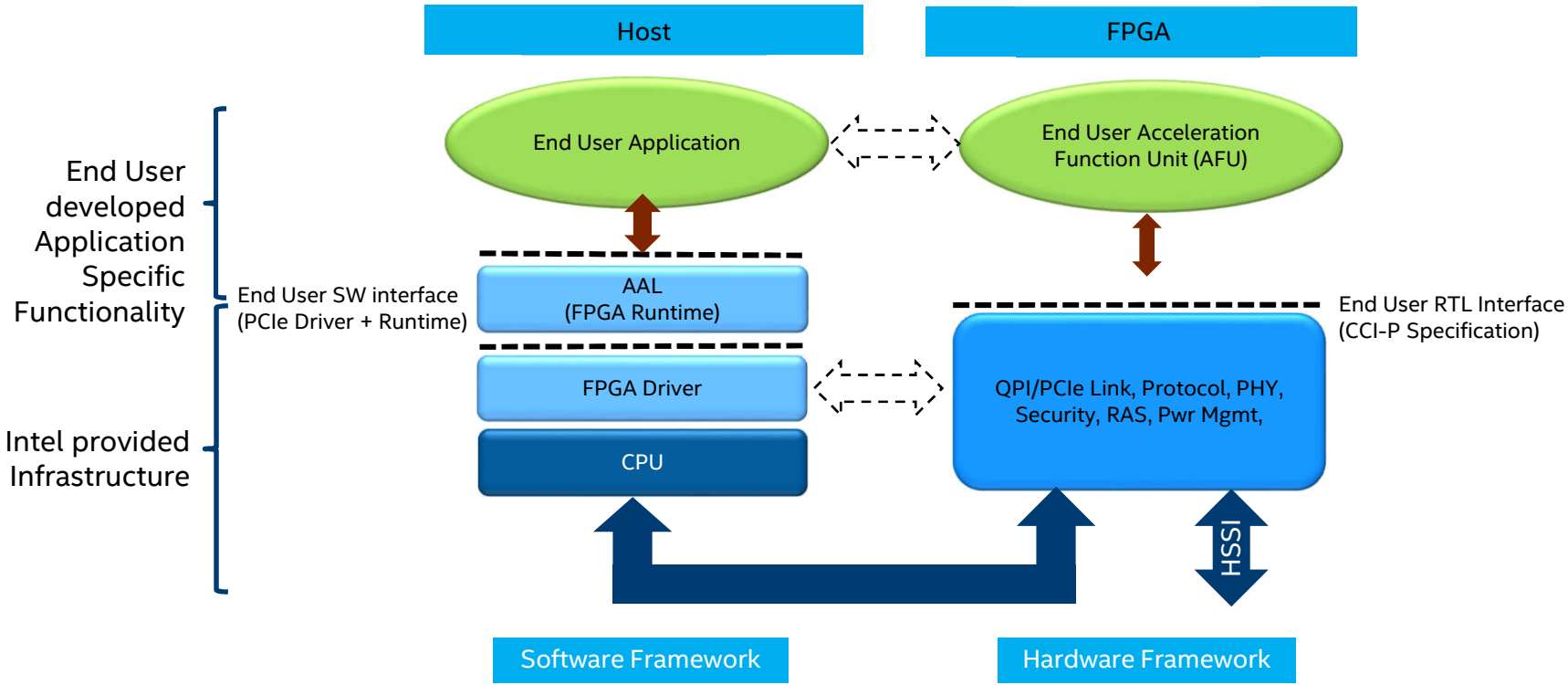
## NoSQL

- Memcached/KVS, Cassandra

## HADOOP, SPARK

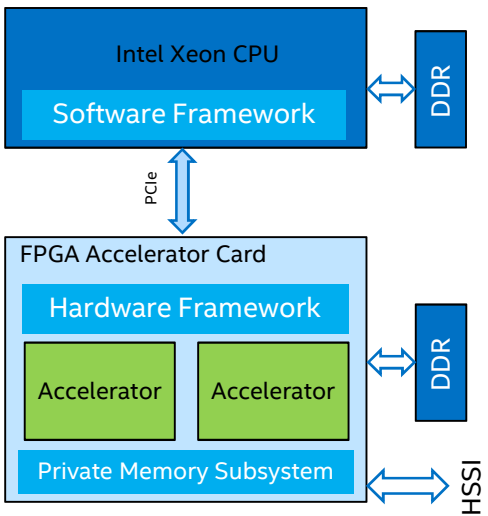
- Accelerate aggregation or “shuffle” phase with better compression
- Data Ingest: FPGA does “inline” offload of Extract, Transform, and Load (ETL)
- SPARK MLlib unstructured machine learning api’s – e.g. K-means, SVM, …
- SQL over SPARK

# End User Programming Interfaces



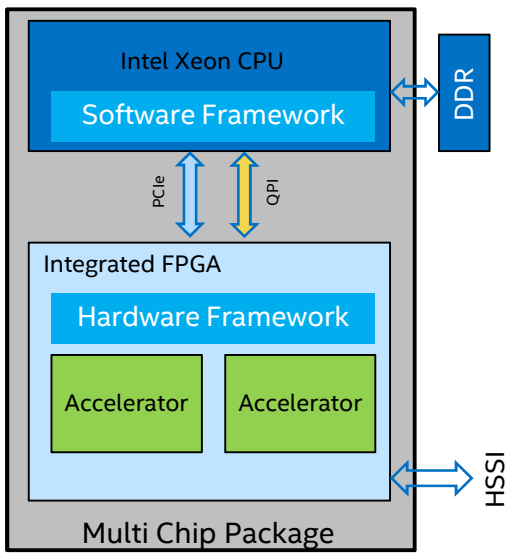
# FPGA Accelerator Card and Integrated FPGA Platforms

## Intel® Xeon®+ FPGA Accelerator Card Platform



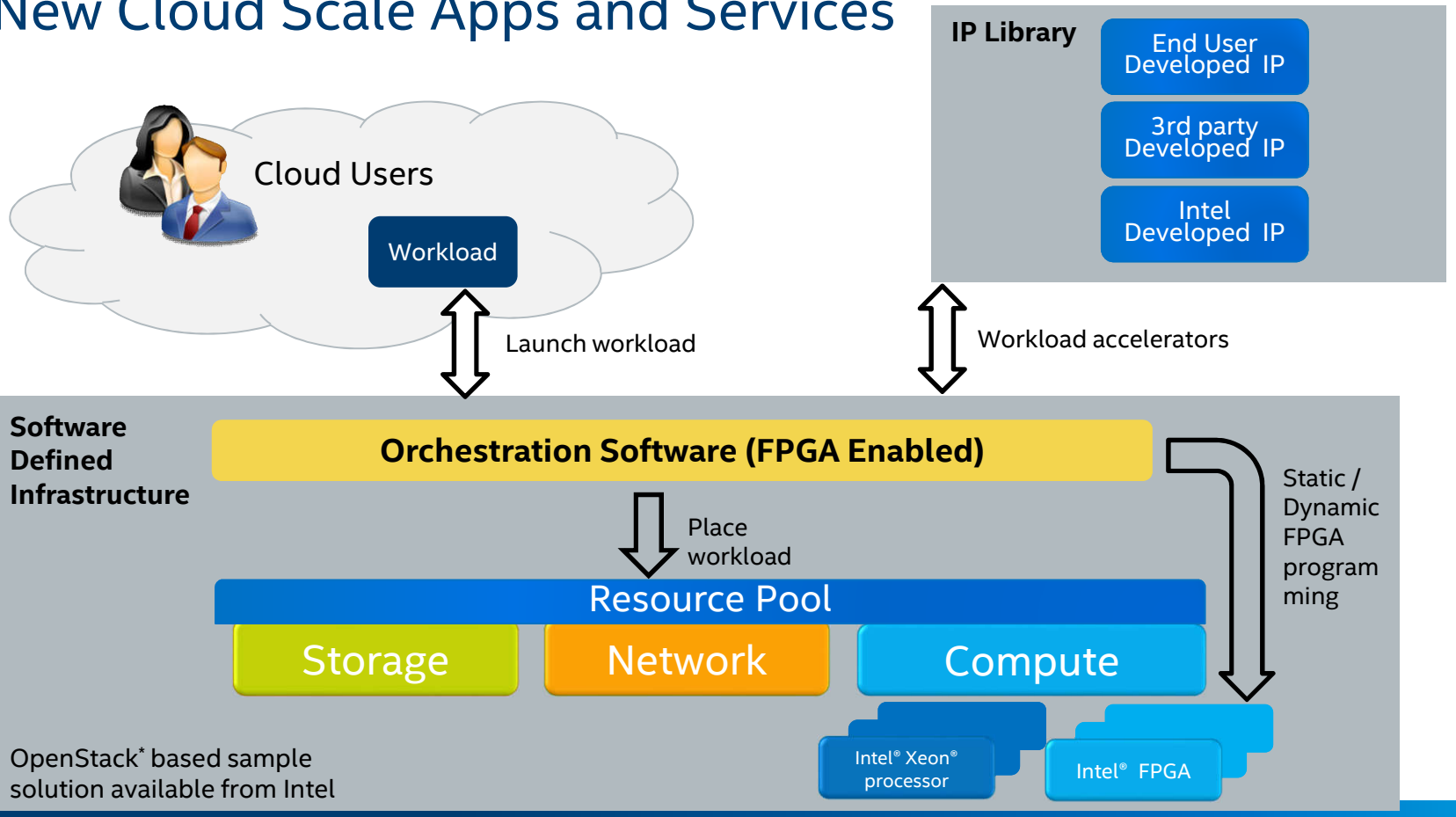

Today: Intel® Arria® 10 PCIe Available

## Intel® Xeon® processors+Intel ® FPGA Integrated Platform

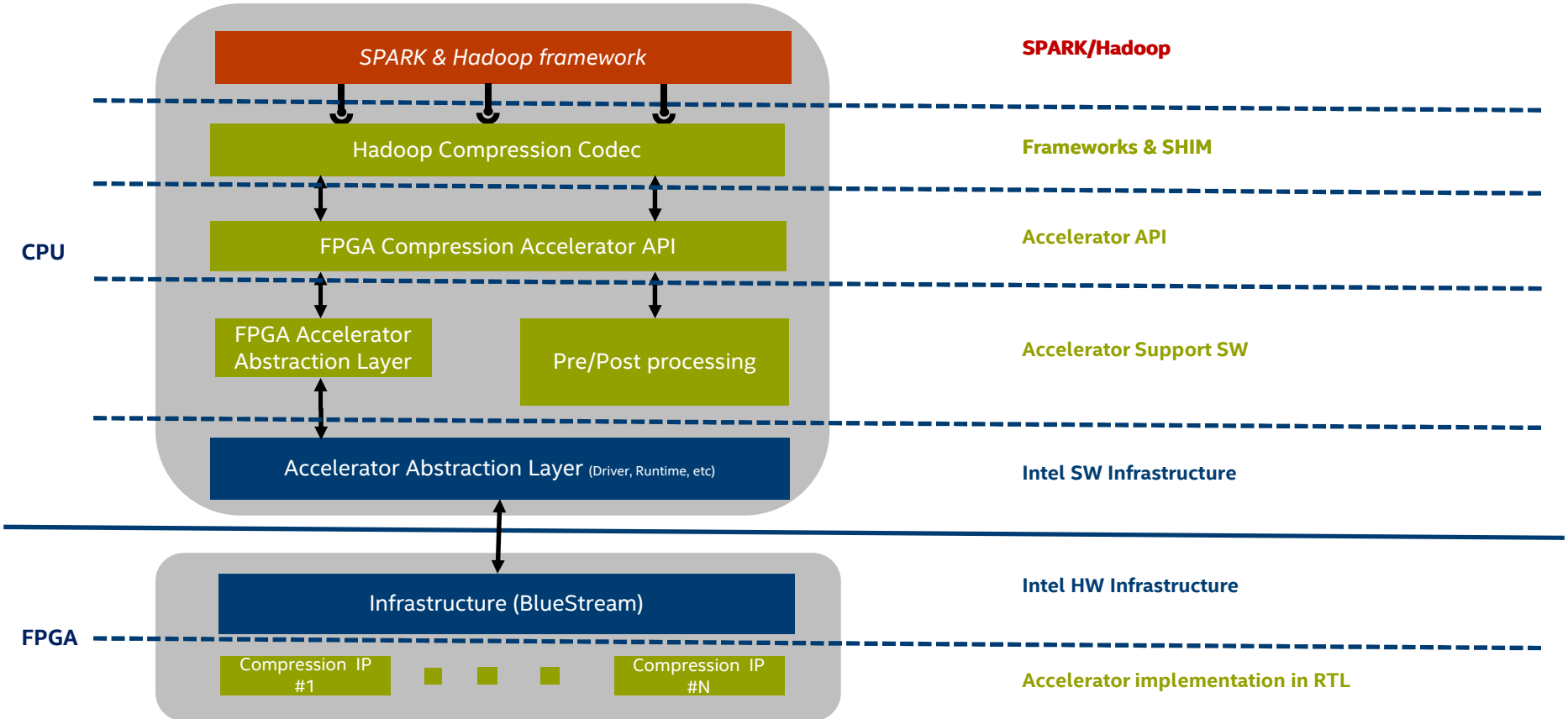



Today: BDX+FPGA MCP Pilot

# New Cloud Scale Apps and Services

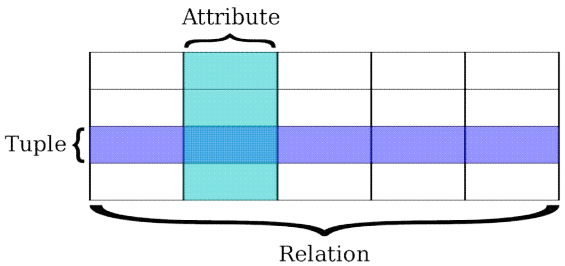


# Intel® Xeon® processor +Intel® FPGA Software Stack for compression acceleration SPARK/Hadoop (Integrated & Accelerator Card Reference Design)



# Different Data Store Approaches

**Structured Data/ Relational**



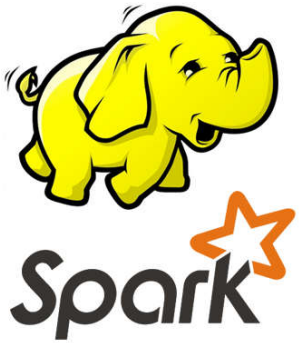

**Semi-structured Data/NoSQL**

cassandra

mongoDB.

| Key | Value |
|---|---|
| K1 | AAA,BBB,CCC |
| K2 | AAA,BBB |
| K3 | AAA,DDD |
| K4 | AAA,2,01/01/2015 |
| K5 | 3,ZZZ,5623 |

**Unstructured Data**

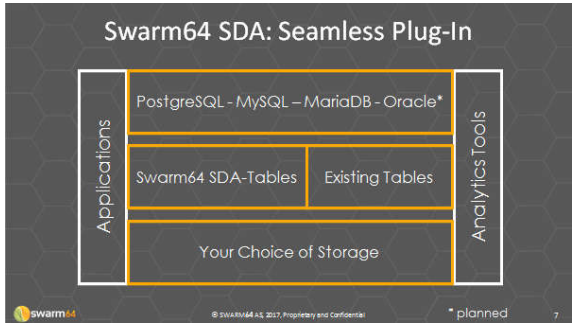


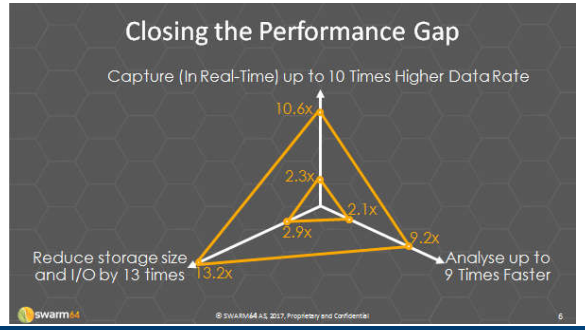

- Traditional Relational d/b for OLTP, Business Intelligence, and OLAP
- NoSQL databases run on multiple servers with no single point of failure
  - Key value store, column-oriented, field, range queries, regex, hybrid,…
- Hadoop designed to work on inexpensive hardware with redundancy
  - Hadoop usually w/unstructured HDFS flat files; SPARK RDD in memory is faster

# Swarm64 Relational Database Acceleration

## Two Workloads: Traditional Data Warehousing, Real Time Data Analytics

### Database accelerate with a plugin




### Acceleration Overview

- 5X+ updates/queries for real time data analytics
  - High Velocity Data, 1M updates/queries/s
- 2X+ queries for data warehousing
  - Using industry standard TPC-DS benchmark
- 3X+ storage compression
  - Data & tables managed by Swarm64

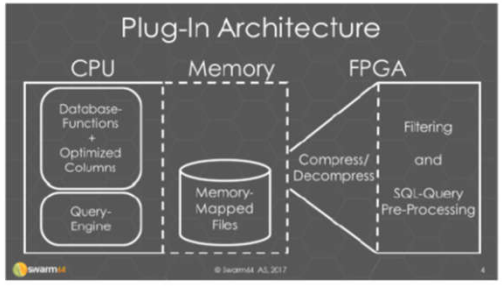
Note: this is SQL to relational d/b, not SQL to semi/unstructured data.

Source: Swarm64

# SWARM64 Relational Database Acceleration

## Scale Up Data Warehousing, Real Time Data Analytics, and storage compression

### Database accelerate with a plugin





### Overview

- No customer application change
  - Storage engine plugin: PostgreSQL, MySQL, …
- Query Engine accelerate INSERT, SELECT, …
- Optimized indexing
- More io bandwidth, mem depth from compress.

### Significant Acceleration

- ✓ Data access acceleration
- ✓ Compression, filtering, replication …
- ✓ Memory mapped acceleration, data cache
- ✓ “Optimized Columns” indexing

# SWARM64 Real-Time Analytics Use Cases

## Banking & Finance

- Fraud detection
- Trading & risk management

## Business Intelligence

- Real-time process monitoring

## Government

- Network monitoring and threat analysis

## Healthcare

- Remote patient monitoring
- Real-time asset monitoring

## Utilities

- Smart metering
- Service quality optimization

## IT & Telecommunications

- Network optimization & predictive maintenance
- Application log analysis (performance, optimization, data security)

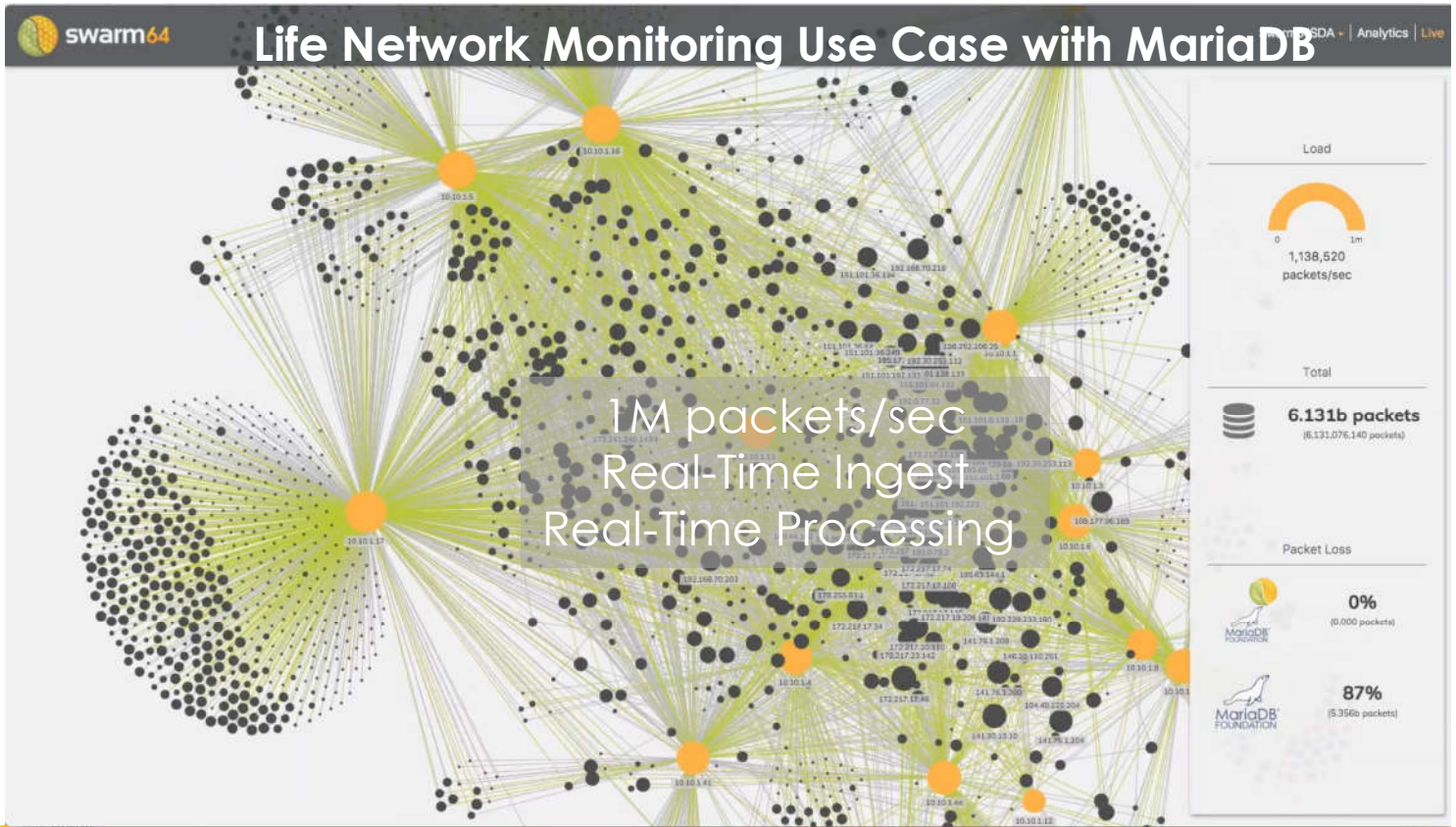
## Retail and e-commerce

- Multi-channel sales optimization
- Customer behaviour modelling
- Real-time recommendation engines

## Transportation

- Assets & fleet monitoring
- Fuel consumption optimization
- Traffic management

# 10x Faster on Ingests and Query Processing Example




Source: Swarm64

# NoSQL: Key Value Store w/Algo-Logic Systems (Intel Partner)
## Networking + Data Access Acceleration

### Compelling KVS Results

- 150M searchs/s.
- Sub-microsecond
  - Low jitter on the latency
- Better energy efficiency

### Comprehensive KVS White Paper

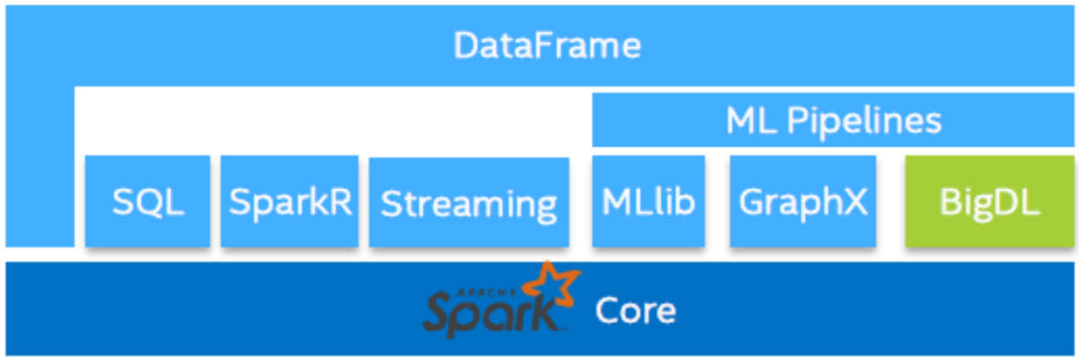
- Comparison to:
  - Cpu with sockets
  - Cpu with DPDK

| All Datapaths Summary | Latency (µseconds) | Tested Throughput (CSMs/sec) | Power (µJoules/CSM) |
|---|---|---|---|
| Sockets | 41.54 | 4.0 | 11 |
| DPDK | 6.434 | 16 | 6.6 |
| RTL | 0.467 | 15 | 0.52 |

| All Datapaths Summary | Latency (µseconds) | Maximum Throughput (CSMs/sec) | Power (µJoules/CSM) |
|---|---|---|---|
| GDN vs. Sockets | 88x less | 13x | 21x less |
| GDN vs. DPDK | 14x less | 3.2x | 13x less |

**Source: Algo-Logic.** Intel core i7 4770k 3.4 GHz CPU and an Intel 82598 [Intel82598] 10 GE NIC running on the CentOS (RedHat-based) operating system, and implemented with the OCSMbased KVS in C. Used the software socket implementation as a baseline for the software compared to Nallatech P385 board [P385] with an Intel® Altera® Stratix® V A7 FPGA. **http://algo-logic.com/sites/default/files/Key_Value_Search.pdf**

# SPARK: Five Acceleration Areas

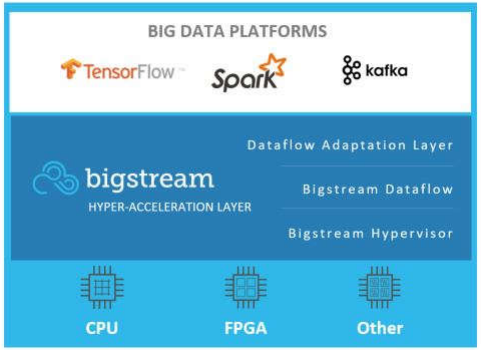




- Shuffle Phase: with high compression ratio (Intel POC)
- Ingest/Kafka: Extract, Transform, Load (ETL) and filtering (partner)
- BigDL: Deep Learning acceleration (under investigation)
- Machine Learning MLlib: ALS, other… (Intel POC, partner)
- SQL over SPARK (partner)

# Bigstream (startup partner) Hyper-acceleration

## The only true in-line acceleration of Big Data/ML using Intel FPGAs

- Frictionless acceleration: up to 10x using Intel® Arria® 10 and Intel® Stratix® 10 (Source: Bigstream)
  - Zero code changes
  - Cross platform: Spark, Kafka, TensorFlow
  - Cloud or on-prem
- Intelligent and adaptive
  - Automatic partitioning of computation
    - Between CPU and FPGA
  - Overlay dataflow execution on FPGA
- Spark SQL TPC-DS results
  - Up to 2.5X with Intel® Xeon ® processor (Source: Bigstream)
  - Up to 10X with Intel® Arria® 10 and Intel® Stratix® 10
    - (Source: Bigstream)
- Industry targets: FinServ/FinTech, AdTech, Healthcare
- Use cases: Spark SQL analytics, ingest/ETL, EDW




http://bigstream.co/resources/whitepaper

http://bigstream.co/resources/video-strata

Customer POC with Arria 10 late Summer 2017

# Summary

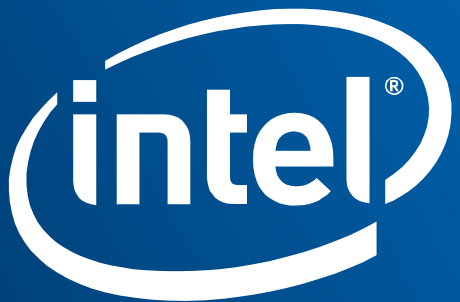
- Intel has comprehensive standards based frameworks and API's for data analytics
- Customers can run analytics workloads without change on these frameworks and Intel API's with FPGAs underneath
- A single FPGA per server can deliver significant acceleration for multiple workloads
- Broad and Developing Data Analytics Ecosystem Partner Solutions & POCs

# Legal Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.  Check with your system manufacturer or retailer or learn more at www.intel.com.

Tests measure performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.  For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

*Other names and brands may be claimed as the property of others.


intel

# Swarm64 Resources

- Public Landing Page (Product Sheet, Solution Sheet, Videos)
  - https://www.altera.com/solutions/industry/computer-and-storage/applications/data-analytics/solutions.html
- Overview Article
  - https://www.nextplatform.com/2017/04/18/fpgas-shake-stodgy-relational-databases/
- 451 Research has a report on Swarm64 for subscribers
  - https://451research.com/report-short?entityId=92588