WHITE PAPER

FASTDATA IO

Plasma Engine™

DATA PROCESSING AT THE SPEED OF THOUGHT

Introduction

Time and information are the world's most precious resources.

The emergence of real-time analytics, deep learning, artificial intelligence (AI), machine learning (ML), natural-language processing, autonomous vehicles, robotics, and Internet of Things (IoT) has given companies the tools to react much faster to ever-changing customer demands, to learn from their customers in almost-real-time, and to predict their behavior with surprising precision.

This incredibly powerful, rapidly evolving ability to collect massive amounts of data in real-time has given businesses almost unabated access to customers' thought process and communications and has thus fundamentally changed the way modern businesses operate.

Data, and more accurately, the valuable insight a person or company can glean from large amounts of data, has become a key commodity in every organization, regardless of industry. The speed and efficiency by which data is acquired and processed has become an arms race for modern business.

Countless software technologies have emerged as a result of unyielding industry demand, and most of them have been open sourced. Some of the most important and widely used include Apache Hadoop, MongoDB, Cassandra, Apache Spark, Apache Flink, and Apache Kafka. The common thread of these technologies? They've all been developed for data processing and machine learning, yet none of them were truly developed for performance and efficiency.

Simply put, data processing technology has not kept up. In 2015, humans collectively generated eight zettabytes (eight billion terabytes) of data, but less than one percent of that data was processed. By 2020, we're expected to generate more than 32 zettabytes of data. In the current software landscape, we only have one solution: build more and more datacenters, use more and more power. At the current trajectory, we have no chance of keeping up with all the data generated.

Enter the NVIDIA GPU (Graphics Processing Unit) + CUDA (Compute Unified Device Architecture):

We want computers to process data the way humans do, in streaming or real-time fashion.

As data streams in, the brain processes the most important and relevant elements first. The extraneous data is then efficiently and subconsciously "thrown away." Why shouldn't computers process data in the same way: in a streaming fashion with speed and efficiency?

With the amount of data that's being generated and needs processing, Moore's Law failed us a long time ago, and the traditional CPU is an obsolete technology for the world's massive, real-time computing problem.

The solution is to extend beyond Moore's Law, the limited speeds of CPU, and into the realm of massively parallel processing (MPP). If data processing could be parallelized and we could process thousands, or millions, of pieces of data at the same time, MPP would finally give us the ability to process data in streaming fashion and move from the batch processing realm of collect > store > process to the stream processing realm of collect > process > store, the way our brains process data, in real-time.

Enter the GPU, the hardware technology that allows us to write efficient software and take advantage of MPP's incredible power. Originally designed to render video games, the GPU was designed and developed specifically for real-time graphics processing, a necessary alternative to CPUs, which are batch processors in essence and inefficient at executing real-time workloads.

In 2008, NVIDIA Corporation, the leader in GPU hardware, recognized the unique real-time power of general purpose computing on GPUs (GPGPU), and released CUDA (Compute Unified Device Architecture).

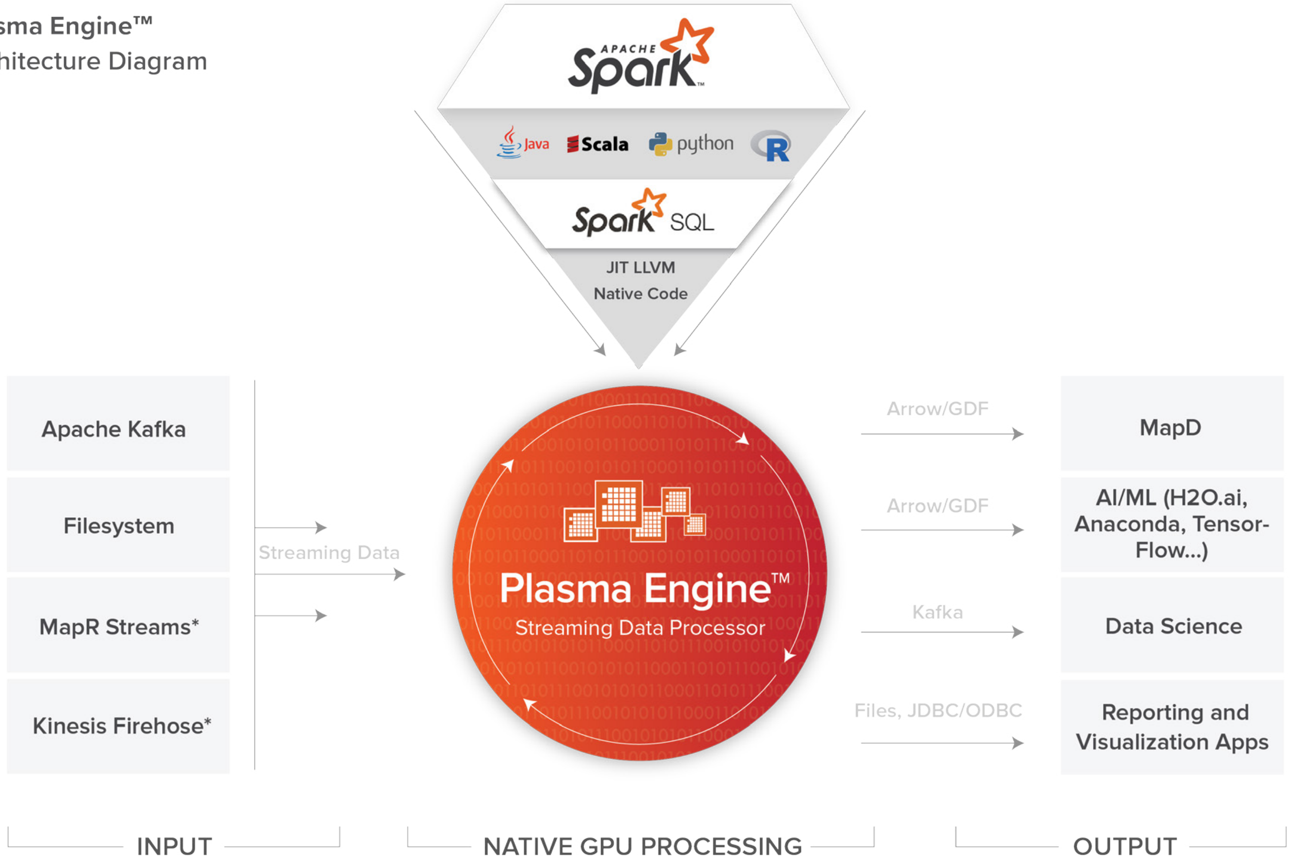
NVIDIA's CUDA API has revolutionized the data processing space and has enabled the creation of advanced artificial intelligence and machine learning software frameworks and engines. Over the past few years, this space has exploded with new technologies and new ways to harness the power of data.

Processing and extracting a meaningful answer or valuable intel from data has gone from days, months, or years to minutes, seconds, or even nanoseconds with GPUs on the job.

Introducing the Plasma Engine

We at FASTDATA.io have recognized the power of this repurposed and emerging technology and the implications for stream processing of big data in motion. By exploiting the power of the GPU and NVIDIA's platform, we have created the fastest streaming data engine in the world today.

Plasma Engine™
Architecture Diagram



Our team at FASTDATA.io has designed and developed the Plasma Engine from the ground up to take full advantage of the extreme efficiency and MPP capability of the GPU, pushing the theoretical limits of data processing. We have also taken extreme care to make sure our Plasma Engine is fully compatible with any existing algorithms and programs already written for Apache Spark.

Plasma Engine is the first GPU-based streaming data engine which customers can run either stand-alone or as a part of a current Apache Spark infrastructure. Since our engine has been natively written and designed to take advantage of the GPU's processing power, current and newly written Spark programs will perform up to 1000x faster than the most optimized Apache Spark engine.

We have designed our engine in a manner that CTOs, CPOs, CDOs, data scientists, engineers, architects, product managers, and software developers can take full advantage of this massive performance and efficiency boost. This newfound ability to process data as it's generated is available without any additional work or recoding.

Plasma Engine comes with full support of all current APIs supported on Apache Spark. Those include Scala API, Java API, Python API and R API bindings to its streaming SQL. It also has fully featured, natively supported SQL processing on continuous data streams. Plasma Engine is the first and only data processing engine with over 80% of its processing done solely on GPUs.

The streaming data problem is much different than batch processing data that is loaded from static and finite data sources like HD, SSD, or Filesystem. In a world of streaming data, all data is new or "hot data."

Plasma Engine was built as the first true GPU-based streaming data engine to tackle that exact problem. To get the most performance out of infinite streams of new data, most, if not all, of the data processing functionality should live on the GPU and its memory without any kind of CPU involvement. We have innovated deeply and creatively to make this a reality.

Going forward we are striving to achieve 95%+ pure GPU processing, and we're working tirelessly to reach that goal. Plasma Engine has absolutely no processing or software bottlenecks to impede its ability to process data much faster than its weakest link, the PCI-E Bus speed. It was designed and developed to be much faster than the PCI-E Bus, the current hardware bottleneck, where data is moved from the NIC to GPU memory. As the speed of the PCI-E Bus increases, so do the performance, power and efficiency of Plasma Engine.

Plasma Engine also has the first of its kind GPU-based data format converter that supports CSV, JSON, and soon to come Syslog data, converting them solely on GPU-to-GPU friendly, zero-copy reads Apache Arrow data format. Plasma Engine is an Apache Arrow data format-based engine, and we use it everywhere internally. Arrow format is a columnar in-memory based natively vectorized data format, which is extremely friendly to the GPU architecture.

This GPU-based data format converter will be open sourced by FASTDATA.io and will become a flagship project of the GOAI Initiative by NVIDIA and others, conveying our capabilities to the wider community and giving us the ability to collaborate with other GPU-based software developers. Plasma Engine has the ability to read data from multiple sources. Directly from socket, streaming data from Kafka, local or mounted filesystems, and JDBC/ODBC (HDFS coming soon).

Plasma Engine supports up to 16 GPUs per node/server, giving it the ability to run hundreds of thousands of operations at the same time. Plasma Engine features a first of its kind GPU-based streaming SQL with CLI front end. Our streaming SQL is fully compliant with industry standard SQL (filtering, aggregations, segmenting, and joining), but it's completely optimized to run on streaming data sets.

All the SQL-based operations run at native GPU speeds, giving it unprecedented performance. In order to achieve this level of performance in SQL, FASTDATA.io has implemented and optimized JIT (Just-In-Time) LLVM query compilation into NVIDIA GPU native code, and as a part of the optimizations, we have also implemented a very sophisticated compiled query caching system, which further aids in our extreme performance quest.

Since Plasma Engine was built to run specifically on GPUs, our JIT query compiling doesn't worry about legacy architectures. We do support x64 query compiling to run on CPUs, but it's not turned on by default. As stated above, almost all engine code is executed directly on GPUs. In addition to JIT query code compiling output, all Plasma Engine compiled code is highly vectorized and optimized for massively parallel operations on GPU.

All data caching by Plasma Engine is also done on GPU memory to minimize unnecessary movement of data between CPU and GPU. We designed and implemented our engine and all of its operations to have only a one-way data move, from NIC to CPU to GPU. This important architecture design goal frees up the precious PCI-E Bus in its entirety for new data being moved for processing to GPUs.

The implementation of this IP is significant. The querying of the data can also be done through JDBC and soon to come ODBC database communication protocols. At this time, Plasma Engine supports single-node multi-GPU configurations, with large-scale multi-node support coming in Q3 of 2018.

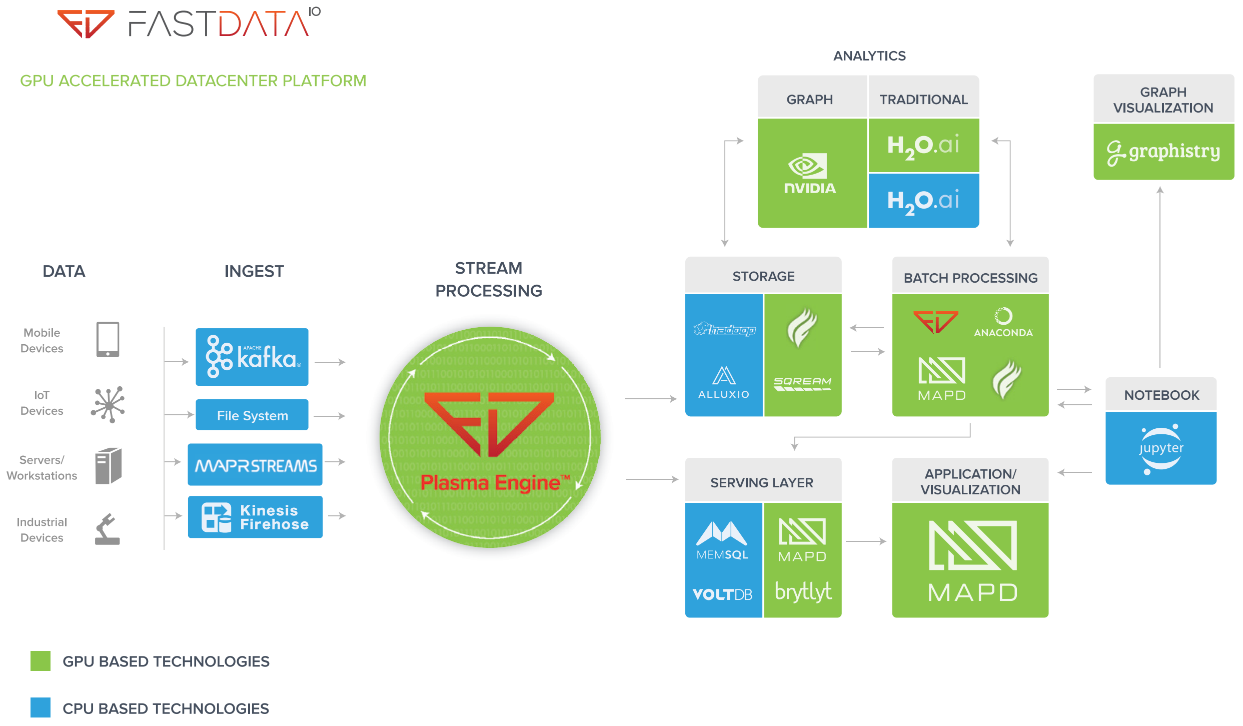
Plasma Engine also fully supports both NVIDIA NVLink and the new NVLink 2.0 technology developed for the new Volta architecture. This next generation interconnect technology is key to our efforts at minimizing and eventually removing all unnecessary data movements between CPU and GPU. NVLink gives us the ability to move data directly between GPUs at speeds of up to 4.5x (with NVLink 2.0 technology) compared to latest PCI-E speeds.

Plasma Engine will also be ported to IBM POWER architecture and will be able to run on IBM OpenPOWER servers, which have NVLink technology implemented between the POWER CPU and NVIDIA GPU. This means that Plasma Engine will perform up to 4.5x faster on POWER architecture due to significantly increased bandwidth between CPU and GPU. This port will also debut in 2018.

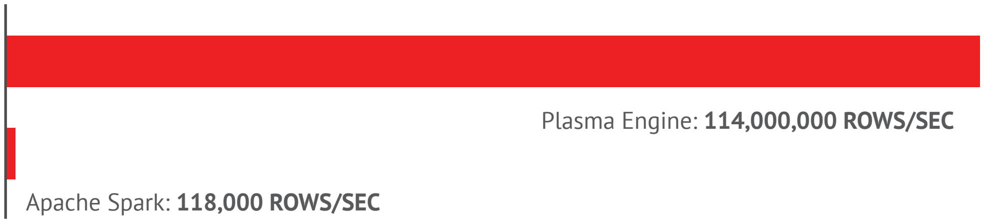
To achieve the highest GPU processing efficiency – in other words, to take full advantage of the massively parallel nature of a GPU architecture, together with all of our compiled query code vectorization – all internal data movement is vectorized and optimized for highly parallel operations, and almost all internal data movement within Plasma Engine is in the Apache Arrow data format. Plasma Engine is fully compliant with the new, NVIDIA-endorsed, GPU Open Analytics Initiative (GOAI). GOAI streamlines the data pipeline and allows developers to fully utilize the performance of GPUs. The project's GPU Data Frame (GDF) format is based on Apache Arrow. Plasma Engine is the first software of its kind to be built with this format in mind from the ground up.

To further our internal data movement and processing efficiency, Plasma Engine also has smarts built in to, when necessary, strip unnecessary parts of Arrow overhead for purely Arrow-compliant vectorized and optimized data movement and processing. Plasma Engine is also the first software to have a fully compliant native Arrow format data output, giving us the ability to output highly efficient and vectorized data into other GPU centric software (like MapD, BlazingDB, Anaconda, etc.), AI and ML projects and frameworks for further analytics processing.

Plasma Engine makes the whole processing pipeline more efficient, from data reading to data format converting, data loading and stream data processing, which ultimately leads to extremely efficient data output. Together with native Arrow/GDF, we support CSV, Apache Avro, and soon to come JSON, Protobuf, Syslog, XML, Parquet, and ORC output into Kafka, flat files and JDBC and ODBC sockets, giving Plasma Engine full compatibility with current big data pipes and workflows.



Performance Benchmark
Plasma Engine vs. Apache Spark Streaming



BENCHMARK TEST:

The test dataset was a 4TB block of telemetry data, and the query below computes the distance between two geo-coordinates and finds rows where two samples are further from each other than the certain distance:

SELECT lat1 FROM rows WHERE (asin(sqrt(sin((lat2-lat1) / 2D) * sin((lat2-lat1) / 2D) + cos((lat1))* cos((lat2)) * sin((lon2-lon1) / 2D) * sin((lon2-lon1) / 2D))) * 12742D) > 200D

TESTING PLATFORM:

One Amazon Web Services EC2 virtual machine (AWS g2.2xlarge GPU instance) with 8 vCPUs, 15GB of RAM and a 60 GB SSD with an NVIDIA Tesla GRID K520 GPU with 4GB of RAM (1536 CUDA/GPU cores)

THE RESULTS:

FASTDATA.io **Plasma Engine** on a single Amazon EC2 GPU instance **processed 114M rows/second**, while Databrick's Apache Spark engine **processed** the same query and dataset at **118K rows/second**.

BOTTOM LINE:

Plasma Engine supports Spark's Streaming Datasets & DataFrame APIs, but processes streaming data up to 1000x faster than Apache Spark when running on NVIDIA GPUs.

FASTDATA.io Plasma Engine Technology Delivery Frameworks

Plasma Engine is primarily delivered as a fully featured NV Docker container and is fully compliant to run on any Intel and NVIDIA-based hardware servers, as well as any of the following cloud providers:
Amazon AWS, Microsoft Azure, and NVIDIA GPU Cloud. We are in the process of certification for Google Cloud Platform, IBM Cloud, and NIMBIX and are expected to be certified in Q1 of 2018.

Conclusion

Up to this point, true real-time, big data processing abilities have been left to extremely sophisticated and multibillion-dollar black boxes developed by Wall Street quant funds solely for the high-frequency trading (HFT) of equities.

With the introduction of CUDA, NVIDIA opened up their incredibly powerful GPU hardware to massively parallel processing of generic data. This has given all businesses the opportunity of moving beyond Moore's Law and into the realm of data processing at the speed of thought.

What used to take weeks or days to process now has been reduced to minutes or even milliseconds. Processing infinite streaming data is now possible.

In order for everyone to take advantage of NVIDIA GPU hardware, an incredibly efficient software has to be written for it. FASTDATA.io has developed exactly that: an equally incredible, performant, and efficient software technology to transform data processing from the old batch world of "collect > store > process" to the new real-time streaming paradigm of "collect > process > store".

The rapidly-evolving power of GPU technology has allowed us, with our Plasma Engine, to provide businesses, developers, data scientists, architects, and product managers with the ability to leverage real-time data for business intelligence from the never-ending stream of live data coming from their customers and operations, to be agile like never before, and to open up new growth and opportunities in the future.

The future is here. Plasma Engine provides data processing performance approaching the speed of thought….and we are just getting started.

About FASTDATA.io, Inc.

FASTDATA.io has developed the world's fastest and most efficient stream processing software engine. Leveraging general-purpose computing on graphics processing units, FASTDATA.io has developed its Plasma Engine, a high-performance computing software engine, to exploit the massively parallel processing capability of NVIDIA's GPU platform. The Plasma Engine scales stream processing up to three orders of magnitude faster, while simultaneously cutting power and space requirements in the data center by more than 95%.

The company was founded in 2016 by Alen Capalik. FASTDATA.io is headquartered in Santa Monica, California with offices in Seattle, Washington. For more information, visit **fastdata.io.**