SNIA PERSISTENT MEMORY + SUMMIT 2022 COMPUTATIONAL STORAGE

VIRTUAL EVENT • MAY 24-25, 2022

An NVMe-based Query Engine for accelerating Big-Data Applications

Presented by Stephen Bates, CTO, Eideticom

Content

- NVMe-based Computational Storage
- NoLoad®: The Eideticom platform for Computational Storage
- Query Offload: What is it and why do it?
- NoLoad® Query Offload Engine.
- Conclusions

NVMe-Based Computational Storage

- Computational Storage is all about building more powerful and efficient systems by pushing compute to the storage layer.
- We need an open standard that defines how a host can push compute tasks to a storage target.
- NVM Express is an excellent choice for this.

nvm EXPRESS

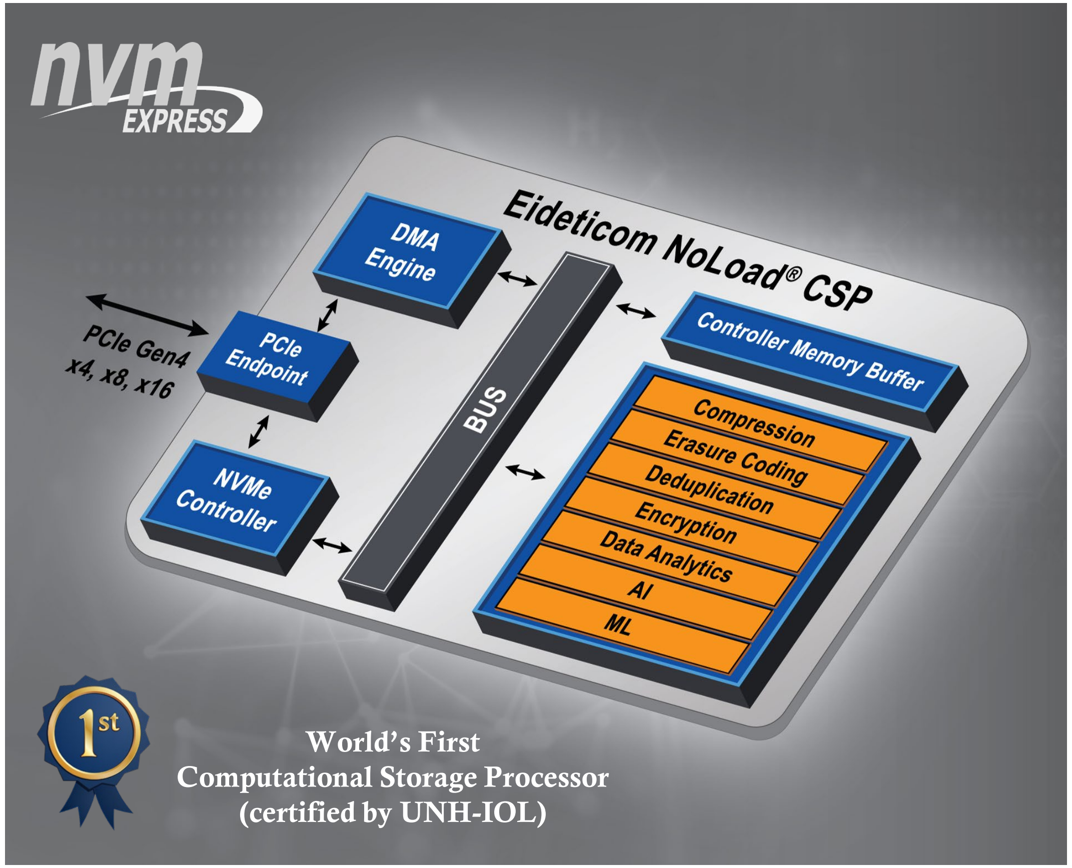
- Performant.
- Pervasive.
- Flexible.
- Works inside a server and across a data-center.
- Scalable.
- Well defined management.

NoLoad®

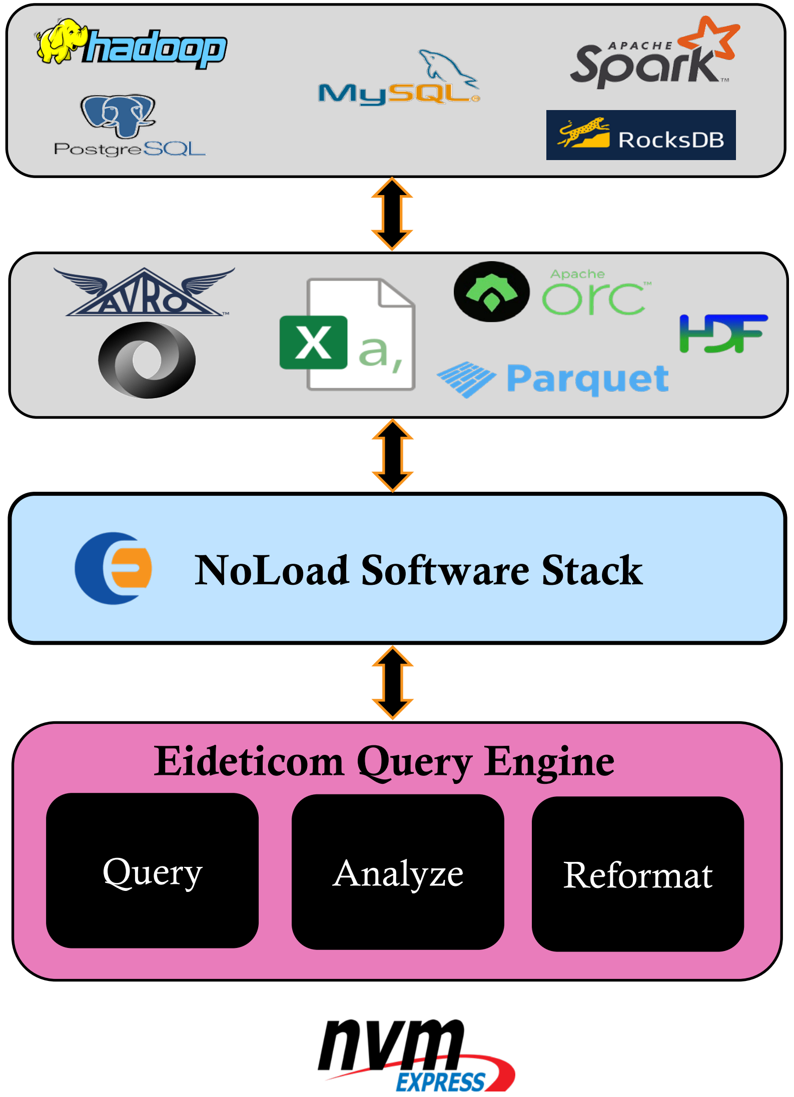
Eideticom's NoLoad®

Purpose built for acceleration of storage and compute-intensive workloads

1) **NoLoad Software Stack**
   - End-to-end computational storage solution providing transparent computational offload
   - Complete Software and IP core stack
2) **NoLoad NVMe Front End**
   - NVMe compliant, standards-based interface
   - High performance interface tuned for computation
3) **NoLoad Computational Accelerators**
   - **Storage Accelerators:** Compression, Encryption, Erasure Coding, Deduplication
   - **Compute Accelerators:** Query Analytics



NoLoad® Query Offload Engine



Data from user space applications is stored using many different formats

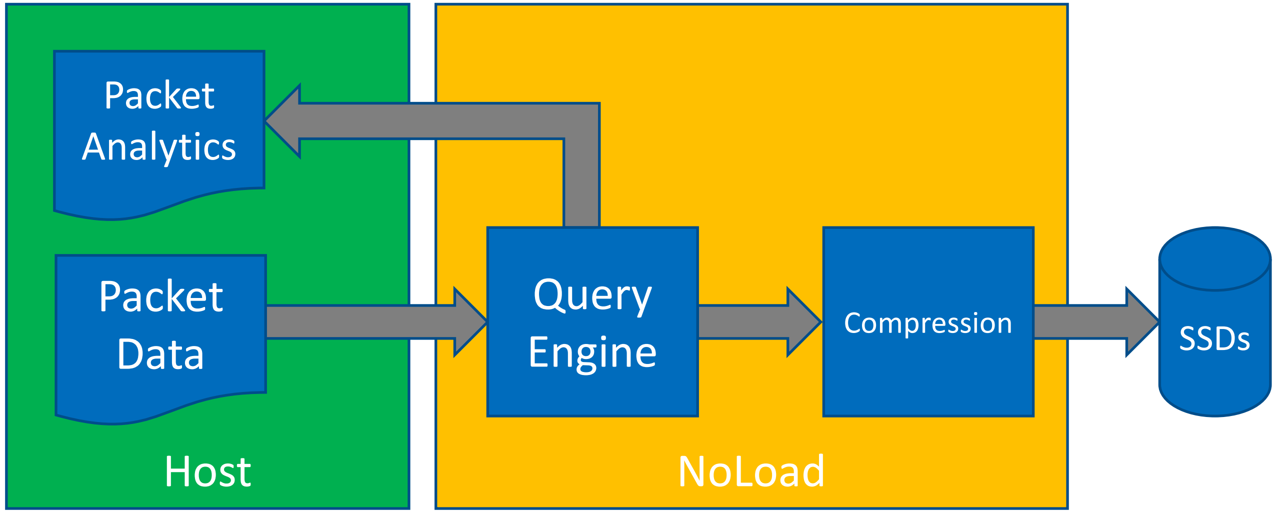
NoLoad SW Stack connects NoLoad Accelerators to end-user applications

Query Engine Value Prop:

- ✓ User Programmable (C/C++)
- ✓ High Throughput
- ✓ Low Latency
- ✓ CPU Offload

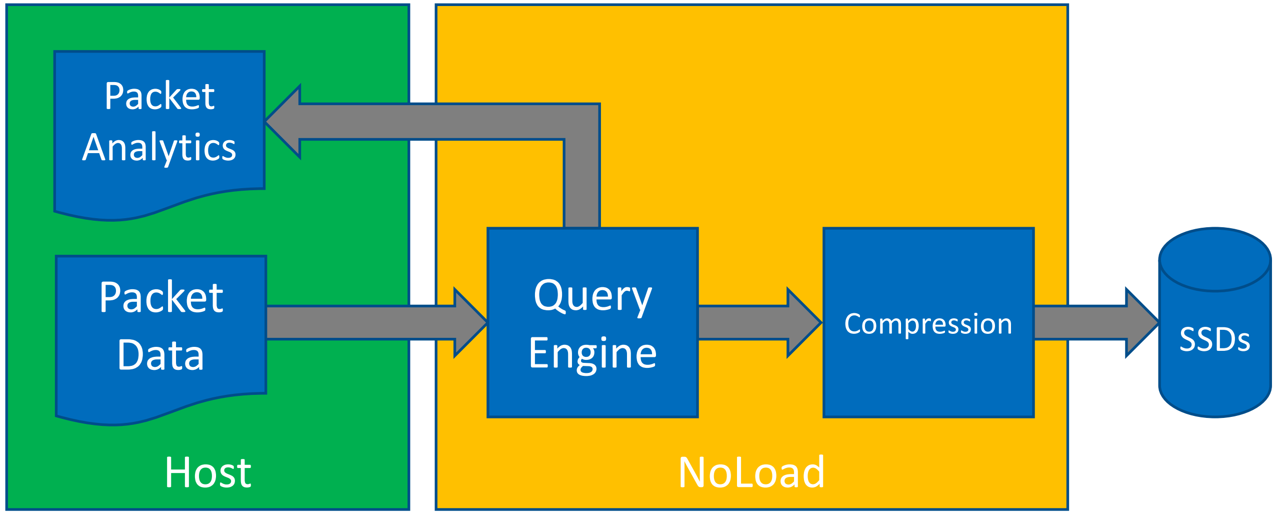
Use Case 1: PCAP Data Analysis

- Packet capture allows Fintech companies to monitor and analyze network traffic for market data
- First level is to take standard packet capture headers and pull out analytics data:
  - IP addresses
  - TCP/UDP header information
  - Packet lengths
  - Packet rates
  - Store analytics data as CSV
- Set real-time alarms on interesting packet events



pcap is a well-defined data format for packet capture and is well suited to analysis by the NoLoad® query engine.

Use Case 1: PCAP Data Analysis

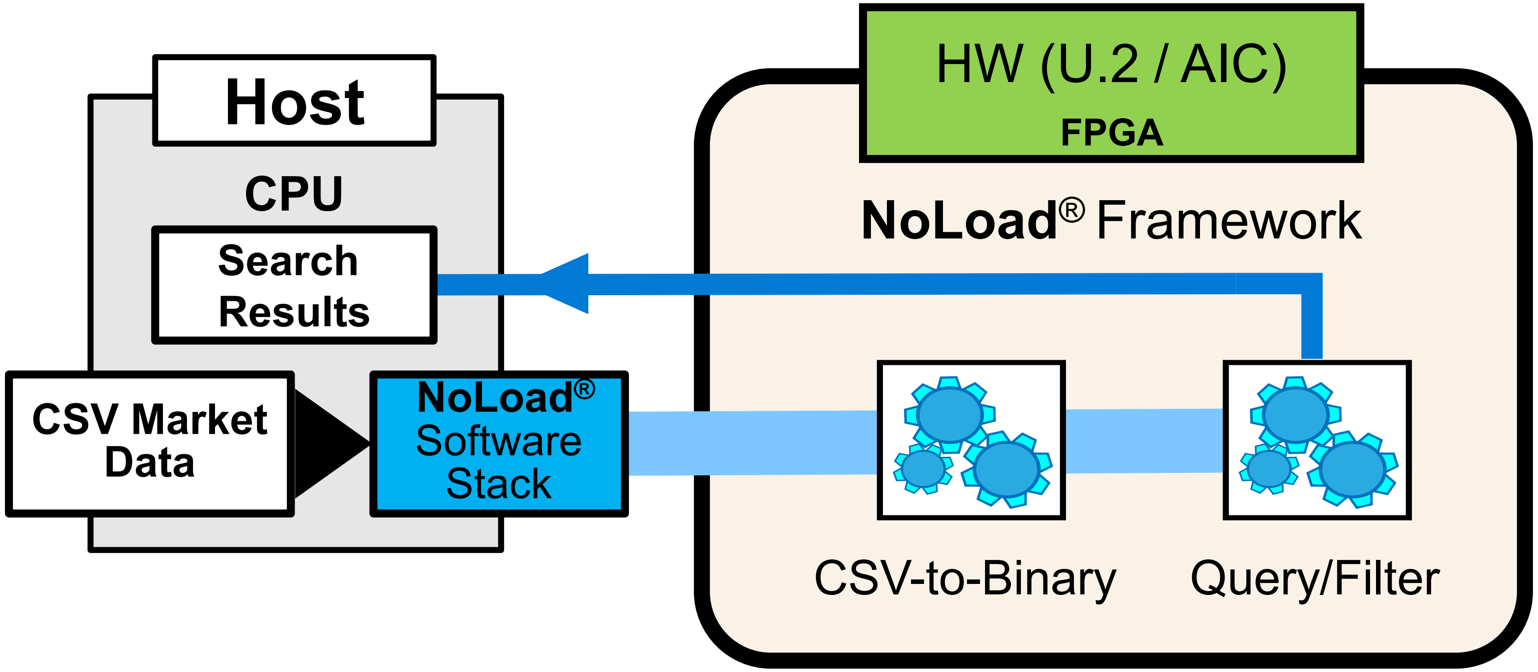


Average Packet Size	Software Throughput	Throughput / QE
256B	0.2GB/s	2.0GB/s
1024B	0.7GB/s	2.4GB/s
4096B	1.9GB/s	2.4GB/s

- Value Proposition:
  - Real-time PCAP header analytics
  - Real-time Compression
  - Achieve 100Gb/s in a Gen4x8 form factor
  - Allows low latency notification of interesting packets and packet statistics
  - Low Latency

Use Case 2: CSV-Based Fintech Data Analytics

Low Latency and **High Throughput** Data Analytics

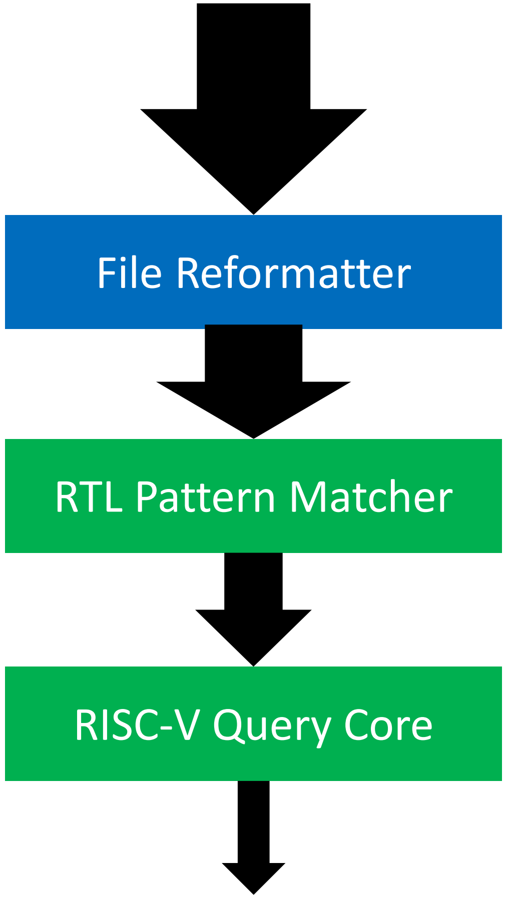


- Fintech companies can **Query, Analyze** and **Reformat** market data; also customize their workloads using our C/C++ **software programmable engines**

NoLoad® Query Engine Architecture

- RTL based data formatters can convert input data to new data formats. This is notoriously inefficient on a CPU.
- RTL based pattern-matching can filter input data reducing data passed to CPU engines.
- CPU engines are very flexible (aarch64 or riscv) but slower. Best working on filtered output.

- Data volume diminishes as we move through the blocks.
- All blocks are programmable by the host.



Conclusions

- NVM Express is a great protocol for computational storage.
- Querying data stored on NVMe subsystems is a very interesting use case for computational storage.
- In order to achieve good performance a mix of RTL-based and CPU-based blocks should be used.
- Filtering data via RTL-based blocks before passing those results into CPU-based blocks enables both good performance and flexability.
- NoLoad® Query Engine yields good performance versus host CPU for a range of use cases. More coming soon!

Please take a moment to rate this session.

- Your feedback is important to us.