SSDs that Think

Intelligent SSDs Can Handle a Larger Computing Load at the Edge

Noam Mizrahi
Vice President, Technology and Architecture
CTO Office, Marvell
People have been mining forever

Data is the 21\textsuperscript{st} century mine.

Information is the gold nugget
DATA INFLATION

DATA GENERATED
44ZB
Global data grew x50 between 2010 to 2020

DATA STORED
1%
Of generated data is stored in Cloud storage infrastructure. Despite >5x growth in amount of data saved in Cloud since 2015

DATA TYPE
80%
Of stored data is Unstructured. Only <5% of it is actually analyzed. Rest remains “dark”
These technologies run over huge data sets, are all data intensive, and rely on Metadata to perform efficiently.

$210B/Y
Worldwide Big Data Analytics Market

Tools & Techs
Transforming Data to Knowledge (Info)
HD Video • IoT • M2M
Metadata is the key

- Tagging of data
- Makes unstructured data understandable

AI is a pivotal technology

- Generates tags over enormous sets of unstructured data
- Generates value out of it
Moving data around is expensive!

**Cloud**
- Network bottleneck
- Efficiency
- Heterogeneity

**Edge**
- Upload is expensive
- Latency
- Capacity
Location, Location, Location

Performing the “Right” task at the “Right” place is a key for efficient system
SSDs that *Think*

SSDs natively include all elements of a compute entity. These can be effectively used for tasks that are data related (e.g., Analytics) rather than function related (Disk operations).

Include additional entities (HW and/or FW/SW) to accelerate data processing at the storage edge for those use cases that make sense.

Efficiency at its most: Power • Performance • Cost
Usage Example 1

Host Co-Processor

Storage device is used as host offload engine

Offloads those host tasks in which proximity to the data is an advantage, while minimizing network traffic, power consumption and host utilization
**Data Pre-Processing Engine**

*Storage device is used to pre-process stored data*

Generates Metadata / Tags / Indexes of stored data, as pre-processing work prior to running analytics SW over the data (e.g. AI based analysis)

1. Data is stored in the Storage device

2. Storage device pre-process the data and generates Metadata / Tags / Indexes etc

3. Host retrieves Metadata upon need or Metadata / Tags are used for further Analytics

© 2018 Marvell Semiconductor, Inc. All rights reserved.
Efficiency

*Heterogeneous computation as a key for efficiency. Run the «Right» things at the «Right» place*

Not all workloads will benefit from running at the Storage Edge, however, those who will, would benefit much, and on many metrics.
Spot a terrorist in a crowd
- Requires very fast results
- Analysis of structured and tagged data base (face characteristics) that are compared with the current image taken characteristics

Locate a stolen car
- May be done over longer time w/more limited compute
- Analysis of unstructured and non-tagged data base
- With 1000s of video hours to go through, minimizing the amount of data moving is key for efficiency
Workload Examples

**Data Base**
- Key / Value acceleration
- Search
- Compare

**AI / ML Work**
- Vision Processing
- Video Analytics
- Text Processing

**Tagging & Indexing**
- Histogram
- Data tagging
- Metadata Management
DEMO

Create Metadata

Search in Database

Update the Database
STORAGE EDGE PROCESSING

DEMO RESULTS

HOST CPU UTILIZATION

- HOST PROCESSING: 85%
- STORAGE EDGE PROCESSING: <1%

LATENCY RATIO
(24*8TB SSDs Database Dual Socket Processor)

POWER RATIO
(24*8TB SSDs Database Dual Socket Processor)

NETWORK CAPACITY RATIO OVER TIME

1000 : 1
More Network Utilization for SW Processing*

* Per the demo database. May vary up or down with different databases

Demo Score - Scaled
- Host
- FPGA Demo
- Storage Edge

Up to >x100 lower latency

Scaled Estimation

© 2018 Marvell Semiconductor, Inc. All rights reserved.
Performance at Scale

Running the processing at the Storage Edge gains linear scale in performance, and is not bounded by the capabilities of a given CPU core.
SSDs with add-on logic are perfectly fit to generate key Metadata and Tags for effective data analytics processing.

While some analytics work will perform better on a host processor, others will perform better at the storage Edge.

**SUMMARY**

Background AI video/image/text processing is one example for superior efficiency if done at the storage edge.

Come see the demo at the Marvell booth!