Intelligent Storage Enables Image Similarity Search
Newport Platform Solution
August 2018
## AI Use Case: Image Similarity Search Problem Definition

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 M images</td>
<td>UCSB 2007</td>
</tr>
<tr>
<td>1 Billion images</td>
<td>Facebook 2017</td>
</tr>
<tr>
<td>1 Trillion images</td>
<td>2019</td>
</tr>
</tbody>
</table>

Facebook **AI** Similarity Search + Computational Storage

---

**Query**

- Image Similarity Search
- Computational Storage

- August 2018
- 2

NGD Systems, Inc. - FMS 2018
Real Time Image Similarity Search (ISS) model – Where RAM Fails

- **Indexing Process** – Google TensorFlow™
  - Image to vector conversion
  - Dataset creation
  - Training Index
  - Database load (I/O)
  - Add vectors to index

- **Searching Process** – Facebook AI Similarity Search (FAISS)
  - Query Management
  - Search
New Platforms add Lanes, Don’t solve mismatch

• With more lanes, we add maximum Storage

• Time to look at the strain on a high capacity platform with 16TB+ per SSD

• Fabrics help, but only to share loads, not solving who carries the weight of compute

Ouch!

Added CPU and still have a Bandwidth mismatch

media to compute bandwidth mismatch >60X

NGD Systems, Inc. - FMS 2018

August 2018
Move compute closer to data

- Reduce data movement across storage/network/memory/CPU for compute

Key Attributes:
- Maintain familiar methodology (no new learning)
- Use standard protocols and processes (no new commands)
- Minimize interface traffic (power and time savings)
Computational Storage Value Proposition

• **Simplify Design** by Maximizing CPU and Storage

• **Augment Computation** with In-Situ Storage saves CPU cycles

• **Scale up Storage** without CPU, Memory and Power costs

• **Align Bandwidth** through the use of intelligence in storage
Showcasing FAISS Use Model with and without In-Situ

**Top** - HOST and In-Situ Disabled NGD Drive

- Intel® Xeon(R) CPU E5-2630 v4 @ 2.20GHz × 20 – 64Bit

**Bottom** - HOST and In-Situ Enabled NGD Drive

- Intel® Xeon(R) CPU E5-2630 v4 @ 2.20GHz × 20 – 64Bit

Normal FAISS Process Flow:
- Load training set
- Index Data
- Load database
- Add data To Index

In-Situ Process Flow:
- Load, Train & Index with In-Situ

Time and Bandwidth Saved

How FAISS works:
NGD Systems In-Situ Processing Shows Real World Results

- Host Processing requires Data be pulled from storage into Memory
- In-Situ requires No data movement and returns stable results regardless of dataset size
- NGD Systems allows for dataset growth with predictable execution and response

Using main memory is unscalable due to Vector load times
In-Situ outperforms predictably at Scale

Log Scale of Time (s)

Using main memory is unscalable due to Vector load times
In-Situ outperforms predictably at Scale

Log Scale needed to see NGD In-Situ Processing!!
The Newport In-Situ Processing SSD Platform Controller

It’s an NVMe SSD at the core
• No impact on host read/write
• No impact on NVMe driver
• Standard protocols

But then there is MORE (Patented IP)
• Dedicated compute resources
• HW acceleration for data analytics
• Familiar programming model
• Extremely Scalable
Computational Storage Solutions

- In-Situ Software
- Patented Firmware
- Management Algorithms
- Optimized Hardware
- Silicon Architecture

**20 Granted Patents On SSD Technology Innovation**

**Creating a Solid Infrastructure**

- Starting with a Solid Foundation of ASIC Design
- SSD HW Designed for optimized Power and capacity

**Enabling an Optimized Ecosystem**

- 3D NAND Agnostic with Error Floor Free LDPC ECC

**Creating a Sustainable Product Platform**

- Optimized Drive FW to ensure consistent READ/WRITE Performance

**Creating a Solid Infrastructure**

- Starting with a Solid Foundation of ASIC Design
- SSD HW Designed for optimized Power and capacity

**Enabling an Optimized Ecosystem**

- 3D NAND Agnostic with Error Floor Free LDPC ECC

**Creating a Sustainable Product Platform**

- Optimized Drive FW to ensure consistent READ/WRITE Performance
New Benchmarks Coming to the Market

Setting NEW Industry Benchmarks

- Energy density (W/TB) 10x
- Cost reduction ($/GB) 4x
- Volumetric density (TB/in³) 5x

- Hardware acceleration
  - Quad-core 64-bit application processor
- Full Fledged on drive OS
- Light Virtualization
- In-Situ Software Defined

EDSFF/M. 2 Up to 16TB
U.2 15mm Gen3 x4 up to 32TB
FHTQL AIC Gen3 x4 up to 64TB

20 Granted Patents On SSD Technology Innovation

NGD Systems, Inc - Confidential
Managing Data Growth at the Edge

In-Situ Processing

Process Massive Amount of Data
Manage Limited bandwidth
Limit System Power Needs
Thank You