Hardware Accelerated Similarity Search

George Williams
Who Am I?

Director, GSI Technology

Previously, Chief Data Scientist
Senior Data Scientist
AI Research Scientist
Software Engineer
Recent Headlines

- A new ARM-based server processor challenges for the data center
- Facebook is looking for engineers to build its own chips
- JUST HOW LARGE CAN NVIDIA’S DATACENTER BUSINESS GROW?
Convergence and Integration

That was then...
This is Now: Technology Disintegration

Amazon Web Services™
EC2
Google Cloud Platform
Intel AI
Facebook
Qualcomm
Habana
ARM
Movidius
an Intel company
Microsoft
Microsoft Azure
HoloLens
NVIDIA
MYTHIC
Baidu
More Innovation Around The Corner

High Performance Memories & Associative Computing
GSI’s Similarity Search Accelerator

High Performance Memories & Associative Computing
Agenda

- Chip Explosion
- GSI Technology
- What is Vector Similarity Search?
- GSI’s Similarity Search Accelerator
- Integration Case Studies: Bio, Database
- Early Adopters Program
Who Is GSI Technology?

High Performance Memories & Associative Computing
What We Do

High Performance SRAM and DRAM

Aerospace, Government, R&D

GSI Vector Similarity Search Accelerator Chip

High Performance Memories & Associative Computing
What Is Vector Similarity Search?
What is Vector Similarity Search?

☑️ Numeric Representation

Bit-vector → 0110000100
Coordinates → (2.3, 5.6)
What is Vector Similarity Search?

- Numeric Representation
- Simple “Distance” Function

\[ d = \text{Func}(a, b) \]
What is Vector Similarity Search?

- Numeric Representation
- Simple “Distance” Function
- K Nearest Neighbor (Top-K)
What is Vector Similarity Search?

- Numeric Representation
- Simple “Distance” Functions
- K Nearest Neighbor (Top-K)
- Search is Computational
- E-Commerce, Bioinformatics
Visual Search

- Binary Codes, Continuous Embeddings
- Euclidean, L1, Hamming, Cosine
- >1 Billion Images
Visual Search: Embedding Space
Bioinformatics: Molecule Similarity

- Fingerprints
- Tanimoto
- Many Large DBs
- 100s GB
Molecule Similarity: Tanimoto

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>A^B</th>
<th>A^C</th>
<th>B^C</th>
<th>Tanimoto Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.6 or 60%</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

A = Number of bits set in both = 3  
B = Number of bits set in (1), but not in (2) = 2  
C = Number of bits set in (2), but not in (1) = 0  

Tanimoto Coefficient = \( \frac{A}{A + B + C} \)  
= \( \frac{3}{3 + 2 + 0} \) = 0.6 or 60%

Similarity Searching

- ✔️ Jaccard
- ✔️ Intersection / Union
Bioinformatics: Molecule Similarity

- Drug Discovery of Novel Molecules
- Virtual Screening
- Activity (Toxicity) Prediction
CREATE TABLE features (  
id bigint(11) NOT NULL AUTO_INCREMENT,  
feature_vector binary(4096) DEFAULT NULL,  
KEY id (id) USING CLUSTERED COLUMNSTORE)

SELECT  
id  
FROM  
features  
WHERE  
DOT_PRODUCT(feature_vector, <Input>) > 0.9

- Idiomatic SQL  
- Integrate Into Data Pipelines  
- Leverage Skills of Data Eng & Scientists
Many Domains and Applications

- E-Commerce / Recommendations
- Bioinformatics / Genomics
- Healthcare / Medical Records
- Cybersecurity / Malware Detection
- Computer Vision / Video Surveillance
GSI's Similarity Search Accelerator
Computational Memory

High Performance Memories & Associative Computing
“In-Place” Associative Processing

Bit Logic
- Programmable
- 2 million

Vector A

Vector B

C = f(A, B)
Consumer Board Solution

- PCIe Card
- 16GB Memory
- 2 Chips Per Board
- On Board DDR4 Main Memory
- SRAM Cache Per Chip

High Performance Memories & Associative Computing
1 Chassis (4U) Solution

4 Boards Per Chassis

Chassis

- PCIe Card
  - 16GB Memory
  - 128Mb
  - 128Mb

High Performance Memories & Associative Computing
Multiple Chassis Solution

One Chassis Is The Master
Network Attached Storage

RDMA support (NAS As Data Source)

High Performance Memories & Associative Computing
Segmentation by Clustering

- Offline Clustering / K-Means
- Avoids Full DB “Scan”
- Faster Performance
Availability

- Q4, 2018: Chip
- Q1, 2019: Demo boards
- Q2, 2019: Mass production

PCIe Card

- 16GB Memory
- 128Mb
- 128Mb
“GSI’s [Accelerator] can dramatically reduce the time required to search our small molecules database...”

- Dr. Efrat Ben-Zeev, Computational Chemist
Weizmann Institute Case Study

Molecule Similarity Search
✓ Biovia Pipeline Pilot Application
✓ Query of 34M Molecule DB Takes 10 Minutes!

Using GSI Accelerator (estimated)
✓ Query Latency Reduced To 300ms
✓ 400 Queries In 1 Second
In-Memory Database Integration

- Database
- C Library
- DRIVERS
- GSI Accelerator

High Performance Memories & Associative Computing
## In-Memory: Expected Performance

<table>
<thead>
<tr>
<th></th>
<th>Memory Speed</th>
<th>Vector Size</th>
<th>Throughput (imgs/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MemSQL</td>
<td>~50GB / sec</td>
<td>4 KB</td>
<td>12.5 million images/sec</td>
</tr>
<tr>
<td>GSI</td>
<td>~100GB / sec</td>
<td>4 KB</td>
<td>25 million images/sec</td>
</tr>
</tbody>
</table>
Early Adopters Program

- Consult With Our Hardware and AI Experts
- Co-Development and App Integration
- Access to simulator and test hardware
- Co-Marketing Opportunity
GSI Upcoming Events

- Nov, Open Data Science Panel, Visual Search
- Nov, PyData (Washington DC)
- Dec, GSI Similarity Search Accelerator Workshop
- Coming Soon, GSI’s Tech Meetup
- 2019, First Chips and Boards Available
Contact Us

Twitter: @gsitechnology
@cgeorgewilliams

Blogs: gsitechnology.com
medium.com/gsitechnology

Email: associativecomputing@gsitechnology.com
The End. Thanks!
Query

Option 1: INFRINGEMENT IN PRODUCTION
- Image2Vec (VGG, Resnet)
- NLP
- LSTN

Option 2: Query Vector done by External Application
- Fingerprint Query
- In Memory Vector (Cosine)

Option 3: 3rd Party Inference Vector
Single Board

- PCle Card
- 16GB Memory
- 128Mb
- 128Mb

- Small Database
- Fit All Data Into Cache For Lowest Latency
- If Larger, Paging Occurs To Memory
- Cluster Techniques
Large Database

Chassis

PCIe Card

16GB Memory

128Mb 128Mb

PCIe Card

16GB Memory

128Mb 128Mb

PCIe Card

16GB Memory

128Mb 128Mb

PCIe Card

16GB Memory

128Mb 128Mb

Large Database (<1TB, Flat)

Pharma, Drug Search, Weizmann Molecule Search, In-Memory

High Performance Memories & Associative Computing
Multi Board Solutions

For Huge Databases (~1TB)

16GB Memory

PCIe Card

128Mb 128Mb

For Throughput: Batch Queries Split Across Boards

Chassis Host Master Merges The Results

High Performance Memories & Associative Computing
Offline Data Preparation

Training → Inference → Optimize For Cache and Memory
Clustering For Large Databases

- Offline Clustering
- Centroids List <16GB
- Reduces Storage
- Only Centroids Are Kept Local
- For Real-Time Performance
Huge Database

Chassis

- PCIe Card
  - 16GB Memory
  - 128Mb 128Mb

- PCIe Card
  - 16GB Memory
  - 128Mb 128Mb

- PCIe Card
  - 16GB Memory
  - 128Mb 128Mb

- PCIe Card
  - 16GB Memory
  - 128Mb 128Mb

Approx For Large Vectors (Quantization)

Exact Nearest Neighbors For Small Vectors

Large Scale Sim Search, FAISS

High Performance Memories & Associative Computing
Biovia Application Integration

Biovia Application Used By Thousands of Bio-Tech Companies
Weizmann: Load A Database

Load database

High Performance Memories & Associative Computing
Weizmann: 3rd Party Search

High Performance Memories & Associative Computing
Weizmann: Select Search Method

Step 1: Select Search Method
Step 2: Search for similar fingerprints
Weizmann: Define Parameters

GSI's Python code embedded here.

Step 3: Search parameters defined here, e.g. similarity threshold, K
Weizmann: Run Protocol