Managing Massive Input Data in Flash for AI and Deep Learning Applications

Dejan Kocic
Netapp
The World Is Changing Fundamentally
Why Now?

Neural Networks date back decades, so why the resurgence?

1. Big Data
   - Larger Datasets
   - Easier Collection & Storage

2. Hardware
   - Graphics Processing Units (GPUs)
   - Massively Parallelizable

3. Software
   - Improved Techniques
   - New Models
   - Toolboxes

- Stochastic Gradient Descent
- Perceptron
  - Learnable Weights
- Backpropagation
  - Multi-Layer Perceptron
- Deep Convolutional NN
  - Digit Recognition
RISE OF GPU COMPUTING

40 Years of CPU Trend Data

Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten. New plot and data collected for 2010-2015 by K. Rupp
What’s happening in the world of AI?

AI is becoming a disruptive force impacting nearly every industry

40% of digital transformation initiatives will use AI services in 2019

50% of enterprise infrastructure will employ artificial intelligence by 2021

87% of global business leaders expect AI to bring better customer experiences within 3 years

AI is all about data

Data is distributed
Generated and consumed from multiple clouds and on-premises

Data is dynamic
Constantly changing and increasingly cloud-streamed

Data is diverse
Data comes in many forms: video, audio, images, quantitative, logs etc.
Architectural models come and go but data is eternal
Data is critical to AI, but presents significant challenges
(Source IDG research)

- 51% Data silos
- 37% Technology complexity
- 35% Data access
- 35% Data preparation
AI Capacity Growth Worldwide

WW Storage for Cognitive/AI Workloads Capacity (EB)

Source: IDC WW Storage for Cognitive/AI Workloads Forecast, 2017-2022
Edge, data comes from various places.
Data Collectors Come in Many Form Factors
Edge to Core to Cloud

**Edge**
- Ingest
  - Data collection
  - Edge-level AI

**Core**
- Data prep.
  - Unified data lake
- Training cluster
  - Training sets
  - Test
- Deployment
  - IM1
  - IM2
  - IM3
  - Repo

**Cloud**
- Analysis / Tiering
  - AWS
  - Microsoft Azure
  - Google Cloud Platform
  - Cloud AI (GPU instances)
  - Data tiering

---

Flash Memory Summit
Data Has Transformative Value

**ENABLE**
new customer touchpoints

**CREATE**
innovative business opportunities

**OPTIMIZE**
operations
AI is Uniquely Enabling a Range of Use Cases

- Social, Media, Internet and Cloud
- Cyber Security
- Life Sciences
- Defense Intelligence
- Internet of Things
- Financial Markets
- Autonomous Machines/Vehicles
Deployment Choices

Using cloud-based GPUs

GPUaaS
Cloud Shares
Training Sets

EDGE

CORE

Cloud

01010
10101
01010
10110

1010
END-TO-END SYSTEM FOR AV

COLLECT DATA

TRAIN MODELS

SIMULATE

RE-SIMULATE

MAPPING

Sources: NVIDIA, RAND Corporation
DATA COLLECTION AND LABELING FOR AI

- **100’s of petabytes of data from test vehicles**
- **10’s of billions of total images from test vehicles**
- **20% to 50% of data may not be useful**
- **1,500 workers label up to 1M images per month**
- **10+ DNNs for self-driving vehicles**

Source: Data from test fleets of 50-100 cars
## Data Generation from One Survey Car

<table>
<thead>
<tr>
<th>Category</th>
<th>Data Collected</th>
<th>Total Images</th>
<th>Labeled Images</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 petabytes per car / year</td>
<td>1 billion images / year</td>
<td>3 million images / year</td>
</tr>
</tbody>
</table>
AI FOR SELF-DRIVING WORKFLOW

1. Get Data
   - Labeled Data

2. Train & Test
   - Trained Model

3. Adjust
   - Fine Tune Model

4. Deploy
   - Export Model

5. Test & Validate
   - Inference at Edge

---

DNN Development
- Exploration
- Development
- Model Selection

---

Simulate

Re-Simulate
## How Flash Storage compares to other media

<table>
<thead>
<tr>
<th>Technology</th>
<th>DRAM</th>
<th>3D Xpoint</th>
<th>Flash</th>
<th>HDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access time</td>
<td>10 ns</td>
<td>7 µs</td>
<td>150-200 µs</td>
<td>6-12 ms</td>
</tr>
<tr>
<td>Scale</td>
<td>Baseline in ns</td>
<td>138 times slower than DRAM</td>
<td>20-30 times slower than 3DXP, over 2940 times slower than DRAM</td>
<td>Over 40 times slower than Flash, over 850 times slower than 3DXP, over 120K times slower than DRAM</td>
</tr>
<tr>
<td>Bandwidth (seq R/W)</td>
<td>13GB/s / 13GB/s</td>
<td>2.6GB/s / 2GB/s (M.2/NVMe, DIMM FF expected 6GB/s)</td>
<td>3GB/s / 2.6GB/s (M.2/NVMe)</td>
<td>112MB/s / 45MB/s</td>
</tr>
</tbody>
</table>
Where and how flash can help

- Ingesting data from the edge:
  - often lots of small files
  - lots of writes
- Lots of small files create random workload
- In some cases data from edge can be aggregated to reduce the number of IOPS
- Data ingest from Edge to Core can use flash as a landing space at Core to be able to accept huge amounts of data coming from Edge into Data Lakes at Core which then can be tiered to cheaper storage at Core
Where and how flash can help

- When using data from Data Lakes for training, throughput is important and flash can be used to help ingest data faster and make training process run faster.

- Speed and low latency are critical for inference, especially when used for real-time, mission critical applications like autonomous vehicles, voice/video recognition, security… and this is the area where flash can also help accelerate data transfer from data collectors to AI inference systems.
Questions?