Bringing ‘Intelligence’ to Enterprise Storage Drives

Neil Werdmuller
Director Storage Solutions
Arm
Who am I?

- 28 years’ experience in embedded
- Lead the storage solutions team
- Work closely with the industry’s top storage suppliers
- Previously in wireless at Texas Instruments
- BSc in Computer Science from Portsmouth University (UK)
- I enjoy brewing beer at home!
What will we cover today?

- What benefit does in-storage compute bring
- What is needed for in-storage compute
- Ecosystem support available
- Machine Learning in-storage
Arm computing is everywhere

#1 shipping processor in storage devices

21Bn Arm-based chips shipped in 2017

> 5Bn people using Arm-based mobile phones

120Bn Arm-based chips to date

Arm computing is everywhere
Why computation is moving to storage

Bandwidth  Power  Cost  Latency  Reliability  Security
1. Compute waits for data
   • Takes time to move data across fabric

2. Adds latency
   • Multiple layers of interface and protocols
   • Data copied many times
   • Bottlenecks often exist

3. Consumes bandwidth and power
   • Moving data is expensive
   • Data copies increase system DRAM

Moving data to compute:

1. Request data from storage
2. Move data to compute
3. Compute
4. Move results to storage
In-storage compute

1. Compute happens on the data
   • Moved from flash to in-drive DRAM and processed

2. Lowest possible latency
   • No additional protocols – just flash to DRAM

3. Minimum bandwidth and power
   • Data remains on the drive – only results delivered

4. Data centric processing
   • Workloads specific to the computation deployed to the drive

5. Security
   • Unencrypted data does not leave the drive

1. Request operation
2. Compute
3. Return result
Compute:  
- **Frontend**: Host I/F + Flash Translation Layer  
  - Cortex-R or Cortex-A series  
- **Backend**: Flash management  
  - Cortex-R or Cortex-M series  
- **Accelerators**:  
  - Encryption, LDPC,…  
  - Arm NEON, ML, FPGA,…

Memory: DRAM ~1GB for each 1TB of flash  
Storage: 256GB to 64TB… flash storage  
Interfaces: PCIe/SATA/SAS…
What is needed for in-storage compute?

Application processor to run a HLOS
- Runs high-level OS through a memory management unit
- Linux for Open Source software stacks
  - All major Linux distribution run on Arm
- Networking protocol stacks: Ethernet, TCP/IP, RDMA...
- Linux workloads:
  - NVMe-oF, databases, file-systems, SDS, custom applications,...
  - Containerization for workload deployment and portability

Accelerators for specific workloads or for Machine Learning (ML)
- Potentially combined with additional accelerators: Custom hardware, ML, FPGA, GPU, DSP...

Custom workloads can be run without apps processor, but complex to develop/deploy
In-storage compute evolution

Separate Cortex-A series processor
- Enables any SSD (or HDD) to run Linux
- Wide performance range from Cortex-A5...Cortex-A76

Single SoC for cost/latency reduction
- Lower latency by removing internal (PCIe) interface
- Separation of apps processor and the SSD processing
- Shared DRAM and other SoC resources

Combined into frontend/apps processor
- Hypervisor provides SSD frontend separation from Linux
- Lowest cost and tightest integration
- Lowest possible latency
- Highest internal bandwidth
The benefits of in-storage compute

Scalability of compute
• From a single, low-power core to multiple clusters of high-performance cores

Flexibility
• One SSD SoC that is suitable for:
  • In-storage compute, Edge SSD, NVMe-oF,...

Security
• TrustZone isolates Linux and SSD functionality
• Processing of data is all done on the drive
• Decrypted data remains on the drive
In-storage compute

1. Compute happens on the data
   • Moved from flash to in-drive DRAM and processed

2. Lowest possible latency
   • No additional protocols – just flash to DRAM

3. Minimum bandwidth and power
   • Data remains on the drive – only results delivered

4. Data centric processing
   • Workloads specific to the computation deployed to the drive

5. Security
   • Unencrypted data does not leave the drive
Linux ecosystem on Arm

www.linaro.org
A few ‘Works on Arm’ partners

- FreeBSD
- Ubuntu
- Debian
- OpenStack
- Kubernetes
- Docker

www.worksonarm.com
Flexible, Scalable ML Solutions

Only Arm can enable ML everywhere

- Mali GPUs
- Arm NPUs
- Cortex-M/A CPUs

ML performance (ops/second)

Typical ML hardware choice
- Autonomous driving
- Image enhancement
- Voice & image recognition
- Smart cameras
- Pattern training
- Keyword detection

ML capabilities

Data center

Deliver use cases with multiple hardware solutions
Choose best balance of ML performance versus capabilities per use case
For each piece of data used to train the model, millions of model parameters are adjusted. The process is repeated many times until the model delivers satisfactory performance.
When new data is presented to the trained model, large numbers of multiply-add operations are performed using the new data and the model parameters. The process is performed once.
Project Trillium: Arm’s ML computing platform
Arm Compute Library

Optimized low-level functions for CPU and GPU

- Most popular CV and ML functions
- Supports common ML frameworks
- Over 80 functions in all
- Quarterly releases
- CMSIS-NN separately targets Cortex-M

Enable faster deployment of CV and ML

- Targeting CPU (NEON) and GPU (OpenCL)
- Significant performance uplift compared to OSS alternatives (up to 15x)

Publicly available now (no fee, MIT license)
devolver.arm.com/technologies/compute-library

Key Function Categories

<table>
<thead>
<tr>
<th>Neural network</th>
<th>Convolutions</th>
<th>Colour manipulation</th>
<th>Feature detection</th>
<th>Basic arithmetic</th>
<th>GEMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyramids</td>
<td>Filters</td>
<td>Image reshaping</td>
<td>Mathematical functions</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Application

TensorFlow / Caffe etc.

Arm NN

Compute Library

Cortex-A

Mali

Arm ML Processor
Enterprise SSD already has considerable compute performance
  - Cortex-A series already adopted by some Arm partners

In-storage compute delivers with low-cost, low-power and lowest-latency

Machine Learning use cases growing rapidly

In-storage compute and Edge SSD opens up many possibilities
  - Please download this presentation
  - COMP-301-1: “Bringing Intelligence to Enterprise Storage Drives”
  - If you missed my first talk on Tuesday please download the presentation
  - ARCH-102-1: “Transforming an SSD into a Cost-Effective Edge Server”
To learn more...

For more information, visit storage.arm.com.

neil.werdmuller@arm.com
linkedin.com/nwerdmuller
Thank You!
Danke!
Merci!
谢谢!
ありがとう!
Gracias!
Kiitos!
감사합니다
धन्यवाद