Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

IXPUG
INTEL EXTREME PERFORMANCE USERS GROUP

Annual Fall Conference

Optane PMem as an Enabler for Large DNN Models with Homomorphic Encryption

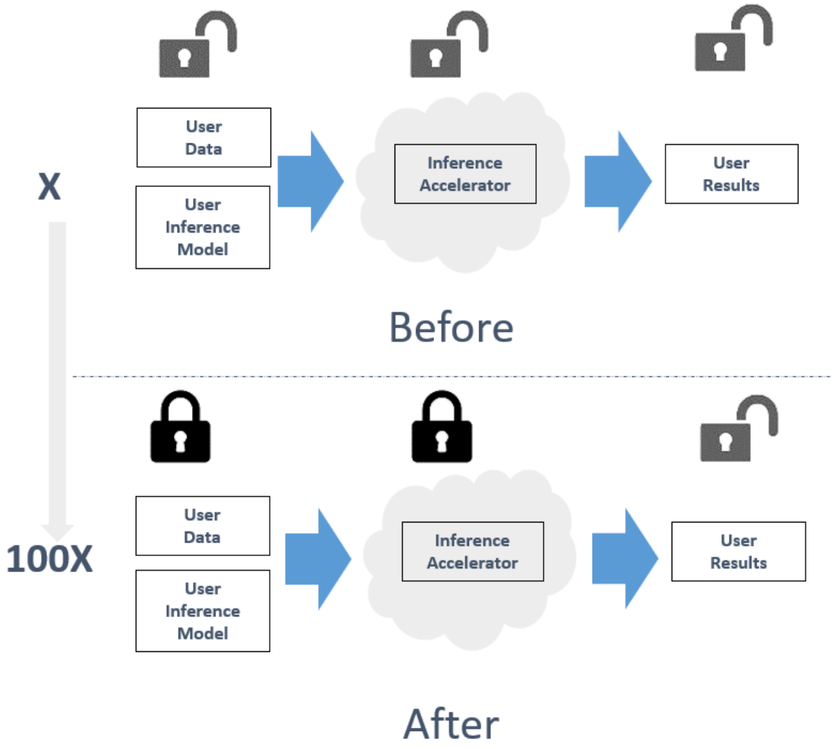
October 13, 2020

Guillermo Lloret-Talavera (BSC), Marc Jorda (BSC), Harald Servat (Intel), Fabian Boemer (Intel), Chetan Chauhan (Intel), Shigeki Tomishima (Intel), Nilesh N. Shah (Intel) and Antonio J. Peña (BSC)

Introduction

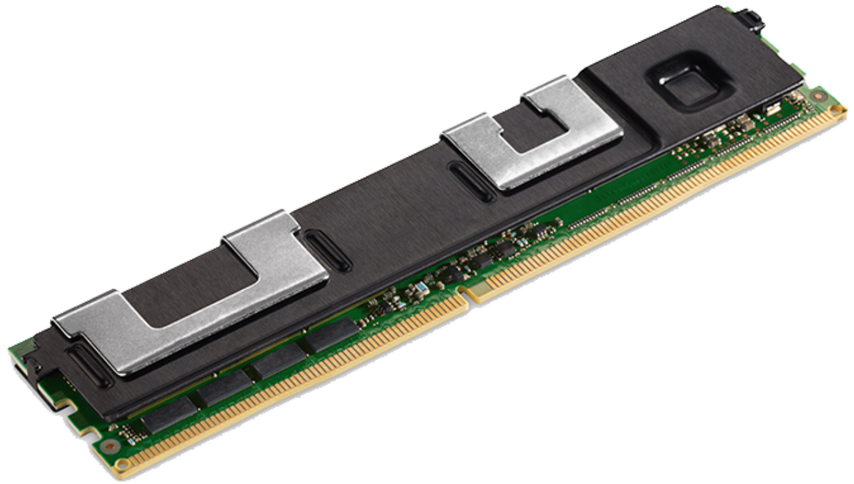
Homomorphic Encryption (HE)

- HE is a type of encryption that enables computation of the ciphertext, without use of the secret key
- HE can be used to process confidential data in an untrusted environment, for example for Privacy Preserving Machine Learning (PPML)
- 100x-10,000x memory and runtime overhead
- Intel Optane can be used by cloud service providers to accommodate the memory requirements



Intel Optane Persistent Memory

- Non-volatile byte-addressable memory DIMMS
- Compatible with 2nd Generation Intel Xeon Scalable
- Allows for large capacity (128 GB, 256 GB and 512 GB)
- Power consumption per byte ~10 times lower than DRAM
- Read/Write bandwidth ~3/~9 times lower than DRAM
- Read/Write latency up to ~6/~30 times higher than DRAM



intel OPTANE™ DC PERSISTENT MEMORY

Objectives

- Analyze the memory requirements for two of the most popular deep neural networks (MobileNetV2 & ResNet50)
- Compare the performance of DRAM only systems versus hybrid systems (DRAM + Optane)
- Analyze the behavior and access pattern of the different neural network operations
- Evaluate whether this particular application is suitable for Intel Optane's Memory Mode.

Experimental Setup

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Hardware configuration

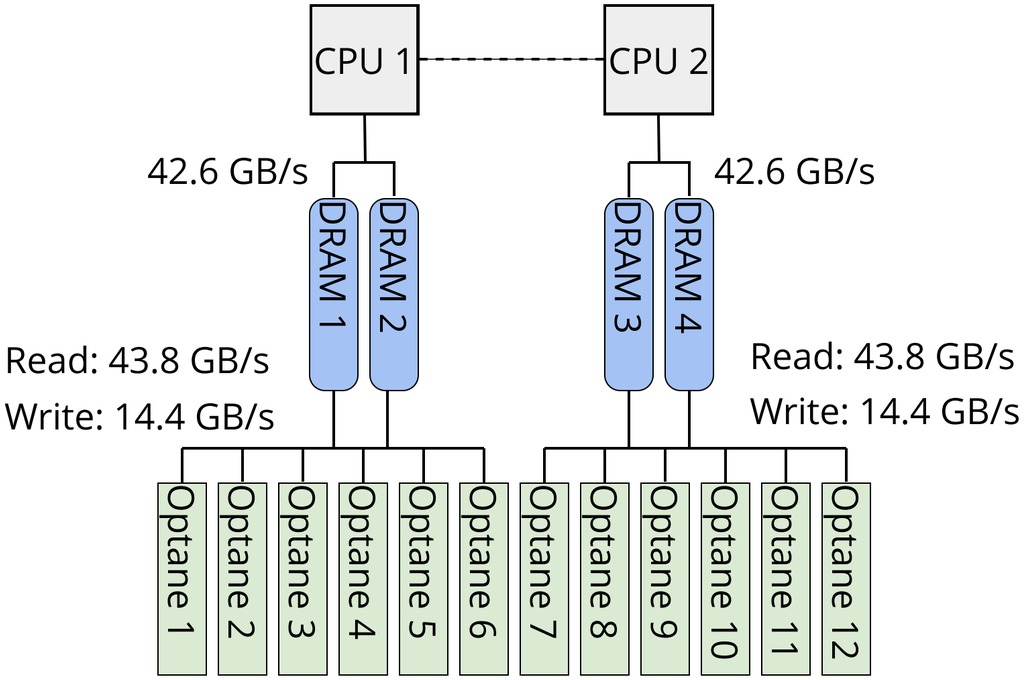
- 2x Intel Xeon Platinum 8260L
  - 48 cores (2x 24 cores)

- Memory configuration
  - Memory Mode (MM)
  - DRAM Only (DO)

Hardware configuration

- 2x Intel Xeon Platinum 8260L
  - 48 cores (2x 24 cores)

- Memory configuration
  - **Memory Mode (MM)**
    - DRAM: 32 GB (4x 8 GB)
    - Optane: 6 TB (12x 512 GB)
  - DRAM Only (DO)

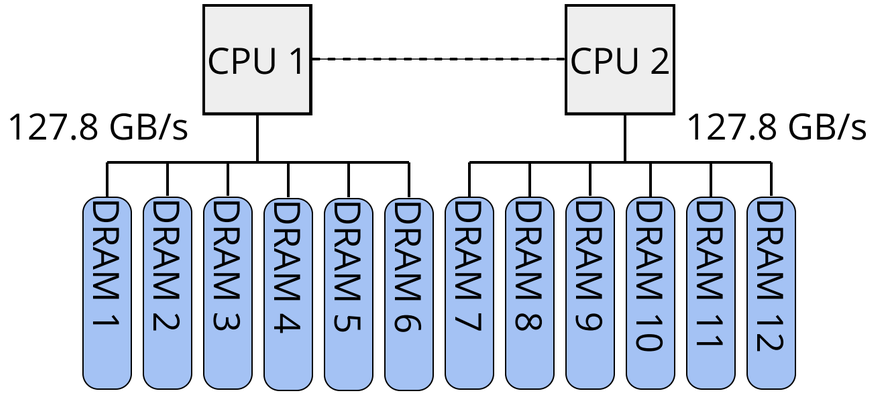


DRAM act as a cache (invisible to the user) for Intel Optane

Hardware configuration

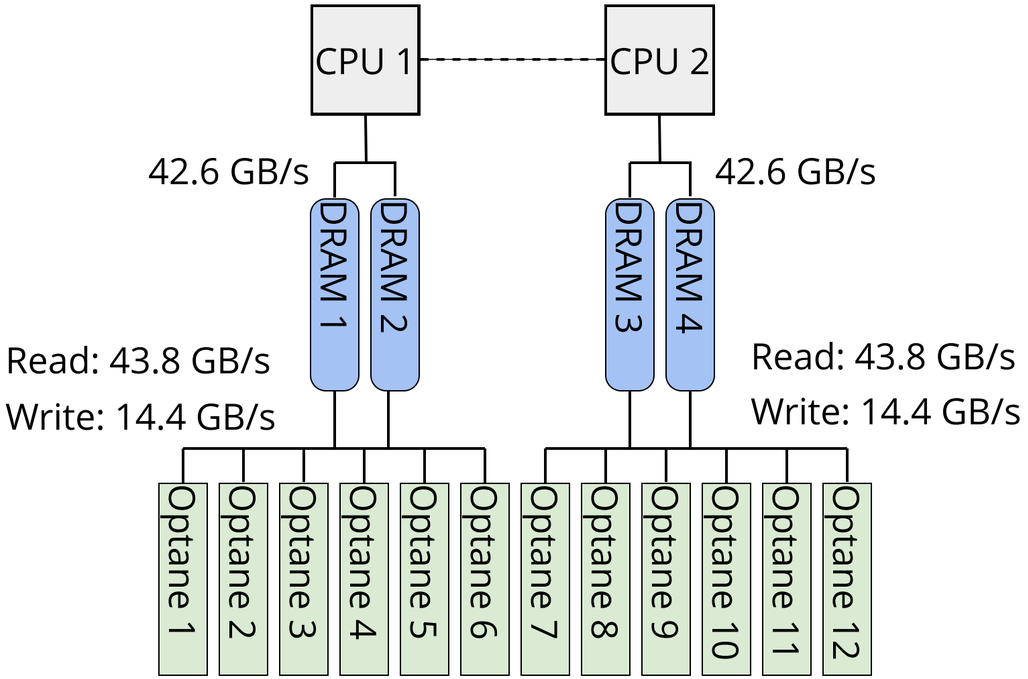
- 2x Intel Xeon Platinum 8260L
  - 48 cores (2x 24 cores)

- Memory configuration
  - Memory Mode (MM)
  - **DRAM Only (DO)**
    - DRAM: 192 GB (12x 16 GB)

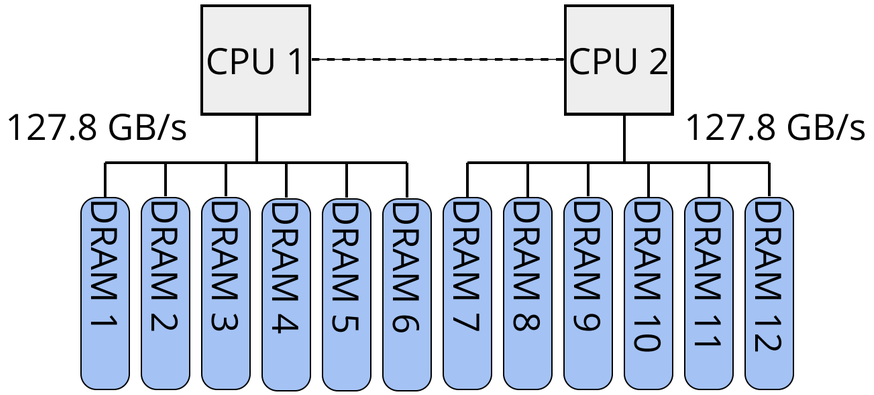


Hardware configuration

Memory Mode (MM)



DRAM Only (DO)



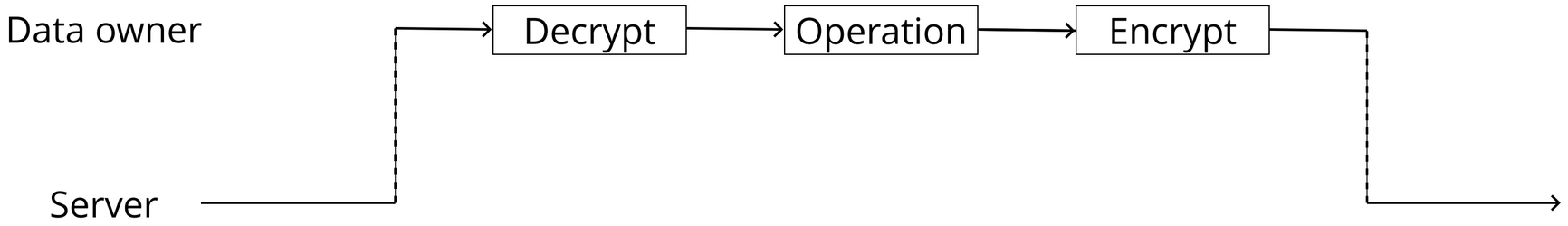
Intel nGraph Compiler

- Open-source graph compiler and runtime for artificial neural networks.
- Compatible with popular frameworks like Tensorflow and ONNX
- Support for different hardware platforms (CPU, Nvidia/AMD GPU…)
- **Intel HE-Transformer**
  - Backend to Intel nGraph Compiler that support Homomorphic Encryption on CPU
  - Relies on Microsoft Simple Encrypted Arithmetic Library (SEAL)
  - Currently support the CKKS encryption scheme
    - Approximate arithmetic for real numbers
    - Encrypted addition and multiplication

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Unsupported operations

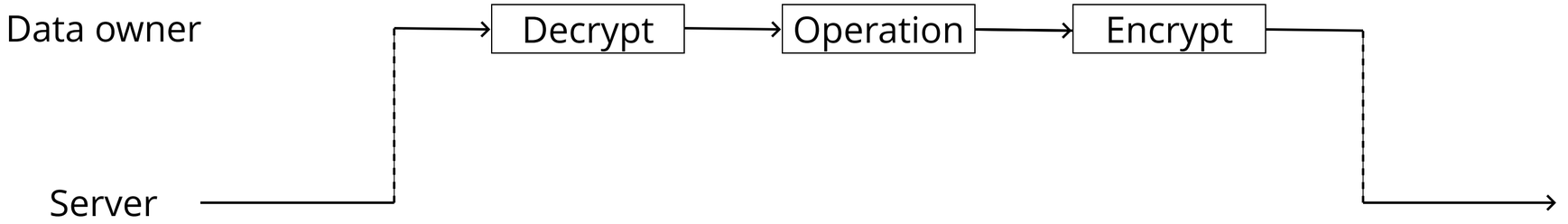
(Operations that cannot be expressed in terms of addition and multiplication)



Unsupported operations

(Operations that cannot be expressed in terms of addition and multiplication)

Real configuration



Experiment configuration

Deep Neural Network Models

MobileNetV2

- Mobile neural network for visual recognition (classification, object detection and semantic segmentation)
- About 9 times less computations than comparable deep neural networks
- Two adjustable parameters that directly affect the size of the model (only smallest settings were tested using HE before)
  - Width multiplier (0.35 - 1.4)
  - Input resolution (96x96 px - 224x224 px)

ResNet50

- Deep neural network used as a backbone for many computer vision tasks
- 50 layers deep
- Input images resolution: 224x224 px
- Never performed inference before with this model using HE

We have used the pre-trained models of both networks.

Our experiments only take into account inference.

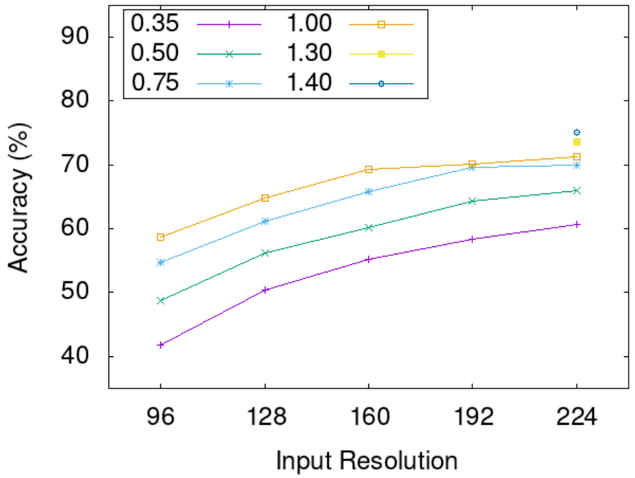
MobileNetV2

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

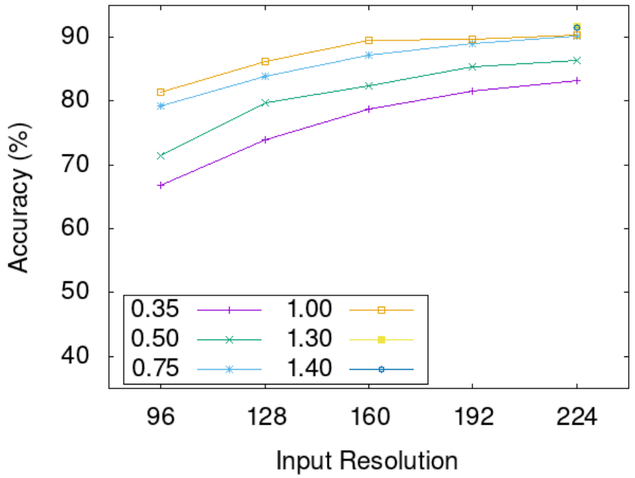
Initial analisis

(Batch size: 2048)

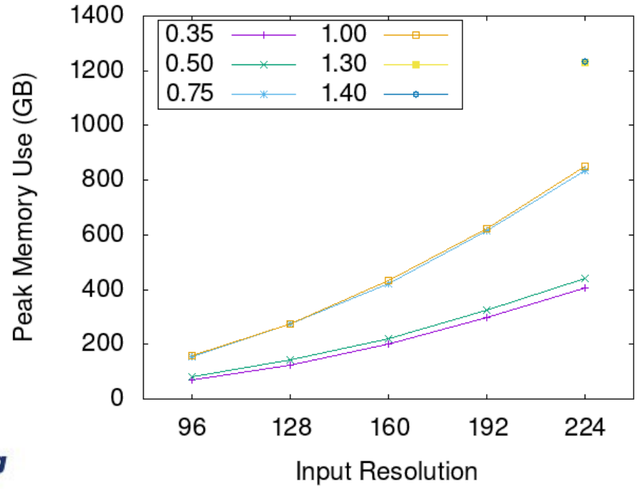
Top 1 accuracy



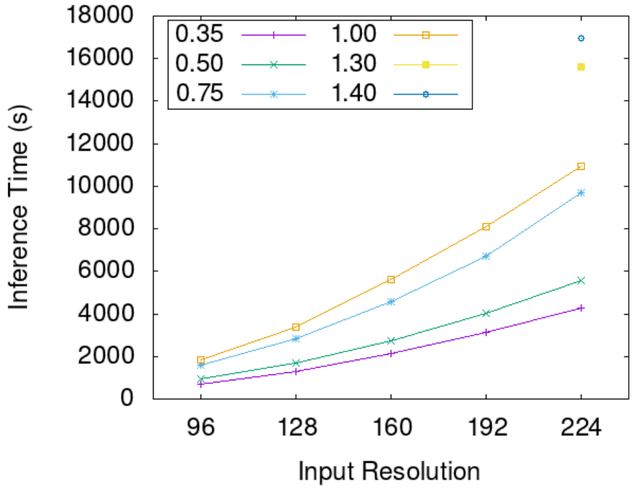
Top 5 accuracy



Peak Memory Use



Inference time (MM)



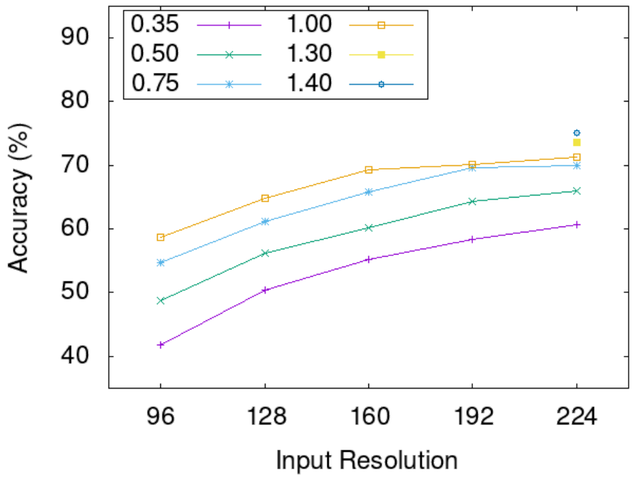
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

For the width multipliers 1.3 & 1.4 only the resolution 224x224 px is available

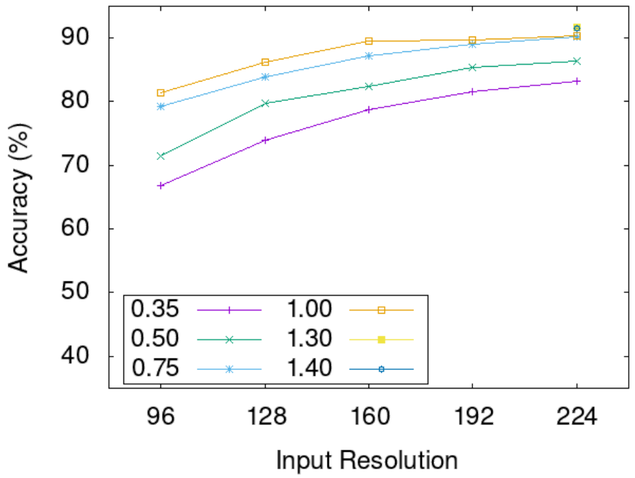
Initial analisis

(Batch size: 2048)

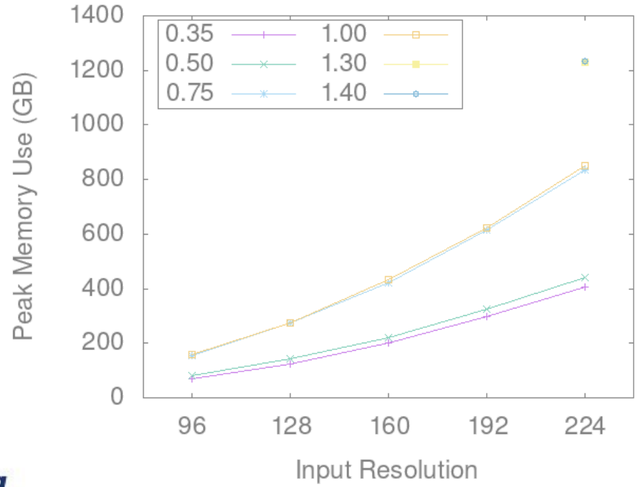
Top 1 accuracy



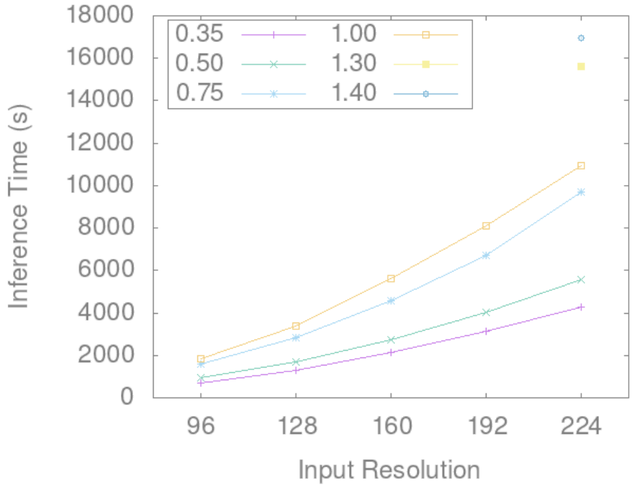
Top 5 accuracy



Peak Memory Use



Inference time (MM)

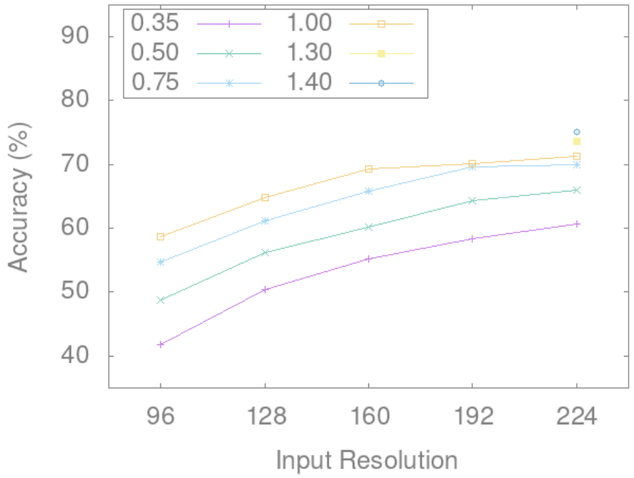


For the width multipliers 1.3 & 1.4 only the resolution 224x224 px is available

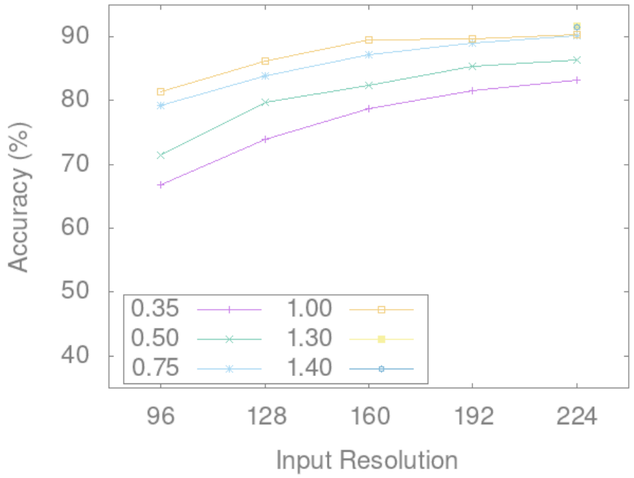
Initial analisis

(Batch size: 2048)

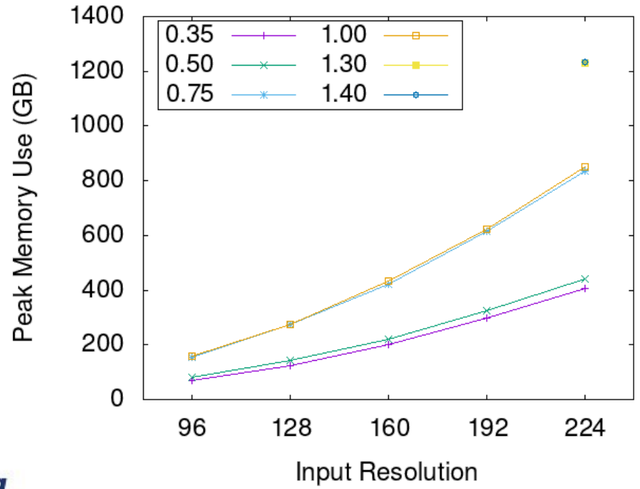
Top 1 accuracy



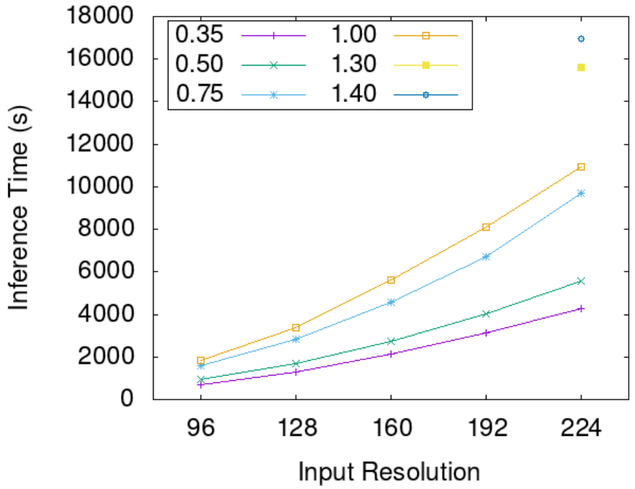
Top 5 accuracy



Peak Memory Use



Inference time (MM)



For the width multipliers 1.3 & 1.4 only the resolution 224x224 px is available

Memory Mode vs DRAM only

Width Multiplier	Input Resolution	Memory Usage (GB)	Time (s) Memory Mode	Time (s) DRAM Only
0.35	96	71	718	665
0.35	128	125	1,311	1,164
0.50	96	80	925	841
0.50	128	144	1,691	1,524
0.75	96	153	1,572	1,432
1.00	96	157	1,830	1,645

Only models that fit on DRAM (192 GB) are shown in this table

Memory Mode vs DRAM only

Width Multiplier	Input Resolution	Memory Usage (GB)	Time (s) Memory Mode	Time (s) DRAM Only
0.35	96	71	718	665
0.35	128	125	1,311	1,164
0.50	96	80	925	841
0.50	128	144	1,691	1,524
0.75	96	153	1,572	1,432
1.00	96	157	1,830	1,645

Memory Mode only ~10% slower than DRAM Only

Only models that fit on DRAM (192 GB) are shown in this table

Extrae & Paraver

- Extrae and Paraver are tools developed by BSC to analyze HPC applications
- Extrae is used to generate traces files that are later visualised and analyzed in Paraver
- Extrae automatically instrument the OpenMP runtime but we have also added custom events to mark the start and the end of the different functions during the inference
- We have capture the following performance hardware counters:
  - MEM_LOAD_RETIRED.LOCAL_PMM
  - MEM_LOAD_L3_MISS_RETIRED.REMOTE_PMM
  - MEM_LOAD_L3_MISS_RETIRED.LOCAL_DRAM
  - MEM_LOAD_L3_MISS_RETIRED.REMOTE_DRAM

Extrae & Paraver results

(Width multiplier: 0.75 | Resolution: 96x96 px)

Function	Time (s)		DRAM (K-loads)		Optane (K-loads)		Ratio
Add	26,605	0.78%	6,333	2.23%	1,345	6.02%	4.75
AvgPool	1,077	0.03%	80	0.03%	10	0.05%	7.65
BoundedRelu*	125,401	3.70%	88,402	30.90%	17,198	76.92%	5.14
Concat	10,578	0.31%	115	0.04%	5	0.03%	20.29
Constant	24,898	0.73%	664	0.23%	32	0.15%	20.46
Convolution	3,113,888	91.76%	187,003	56.37%	3,152	14.10%	59.32
Multiply	9,652	0.28%	1,841	0.64%	489	2.19%	3.76
Reshape	51,041	1.50%	870	0.30%	103	0.46%	8.39
Result	40	0.00%	117	0.04%	0.5	0.00%	207.42
Slice	30,338	0.89%	576	0.20%	19	0.09%	30.11
Total	3,393,517		286,006		22,358		12.79

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

*Not supported by encryption scheme

Extrae & Paraver results

(Width multiplier: 0.75 | Resolution: 96x96 px)

Function	Time (s)		DRAM (K-loads)		Optane (K-loads)		Ratio
Add	26,605	0.78%	6,333	2.23%	1,345	6.02%	4.75
AvgPool	1,077	0.03%	80	0.03%	10	0.05%	7.65
BoundedRelu*	125,401	3.70%	88,402	30.90%	17,198	76.92%	5.14
Concat	10,578	0.31%	115	0.04%	5	0.03%	20.29
Constant	24,898	0.73%	664	0.23%	32	0.15%	20.46
Convolution	3,113,888	91.76%	187,003	56.37%	3,152	14.10%	59.32
Multiply	9,652	0.28%	1,841	0.64%	489	2.19%	3.76
Reshape	51,041	1.50%	870	0.30%	103	0.46%	8.39
Result	40	0.00%	117	0.04%	0.5	0.00%	207.42
Slice	30,338	0.89%	576	0.20%	19	0.09%	30.11
Total	3,393,517		286,006		22,358		12.79

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

*Not supported by encryption scheme

Extrae & Paraver results

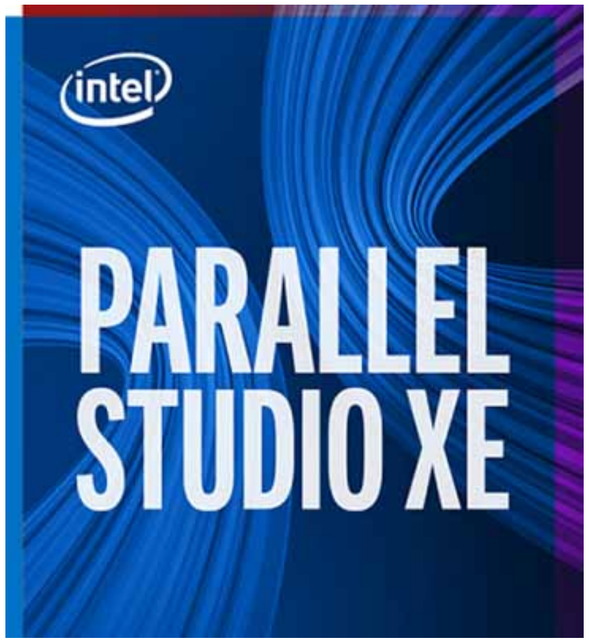
(Width multiplier: 0.75 | Resolution: 96x96 px)

Function	Time (s)		DRAM (K-loads)		Optane (K-loads)		Ratio
Add	26,605	0.78%	6,333	2.23%	1,345	6.02%	4.75
AvgPool	1,077	0.03%	80	0.03%	10	0.05%	7.65
BoundedRelu*	125,401	3.70%	88,402	30.90%	17,198	76.92%	5.14
Concat	10,578	0.31%	115	0.04%	5	0.03%	20.29
Constant	24,898	0.73%	664	0.23%	32	0.15%	20.46
Convolution	3,113,888	91.76%	187,003	56.37%	3,152	14.10%	59.32
Multiply	9,652	0.28%	1,841	0.64%	489	2.19%	3.76
Reshape	51,041	1.50%	870	0.30%	103	0.46%	8.39
Result	40	0.00%	117	0.04%	0.5	0.00%	207.42
Slice	30,338	0.89%	576	0.20%	19	0.09%	30.11
Total	3,393,517		286,006		22,358		12.79

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

*Not supported by encryption scheme

Intel VTune Platform Profiler

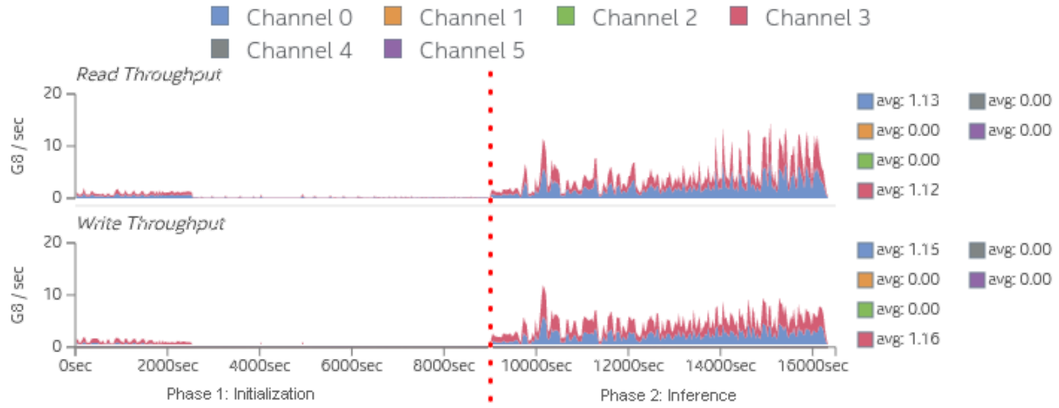


- System-level profiling tool (does not monitor specific functions)
- Periodically collect performance metrics (sampling)
- Smaller trace files allow us to study longer executions

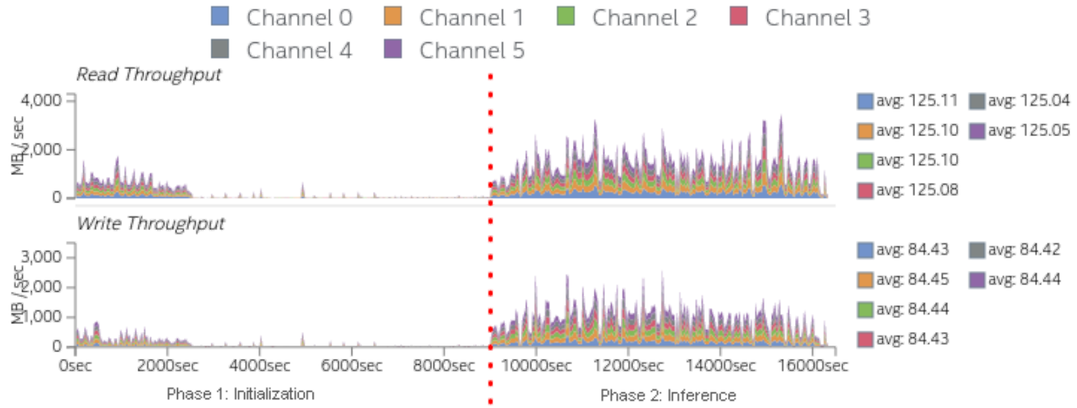
Intel VTune Platform Profiler results

(Width multiplier: 0.75 | Resolution: 96x96 px)

DRAM traffic



Optane traffic



Bandwidth:

- Total: 42.6 GB/s

Bandwidth:

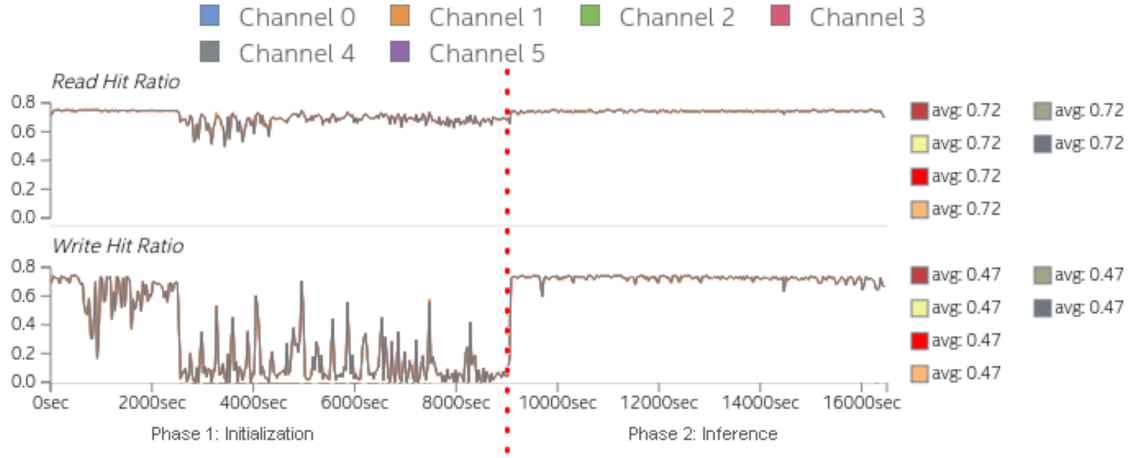
- Read: 42.6 GB/s
- Write: 14.4 GB/s

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Intel VTune Platform Profiler results

(Width multiplier: 0.75 | Resolution: 96x96 px)

- CPU request data to Optane in chunks of 64 bytes
- Each Optane DIMM controller integrates a media prefetch buffer of 256 bytes
- If the data is in the buffer the response time is similar to DRAM
- In a sequential access pattern 0.75 hit ratio is expected (MISS - HIT - HIT - HIT)



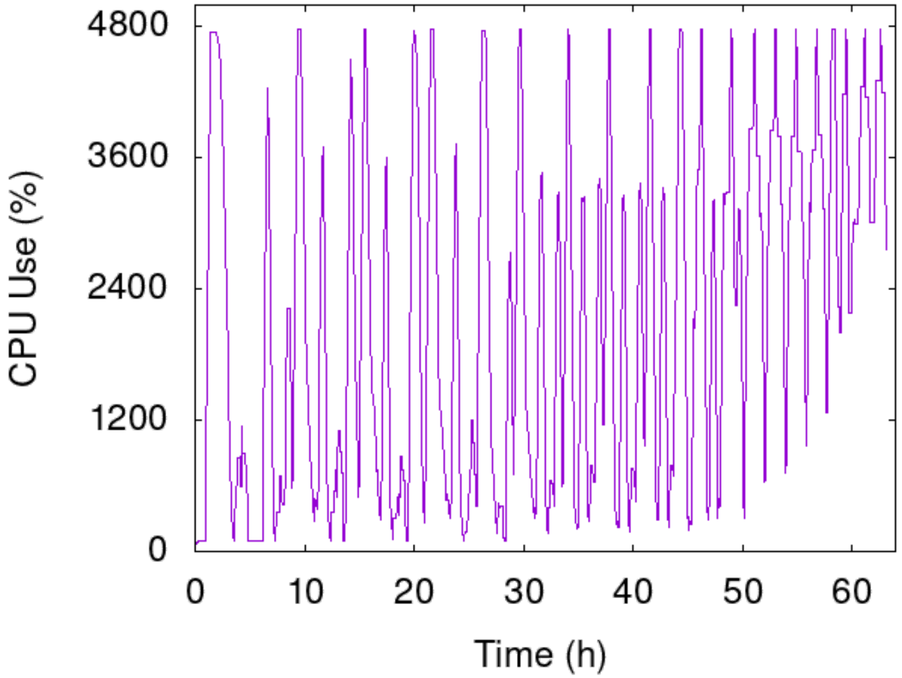
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

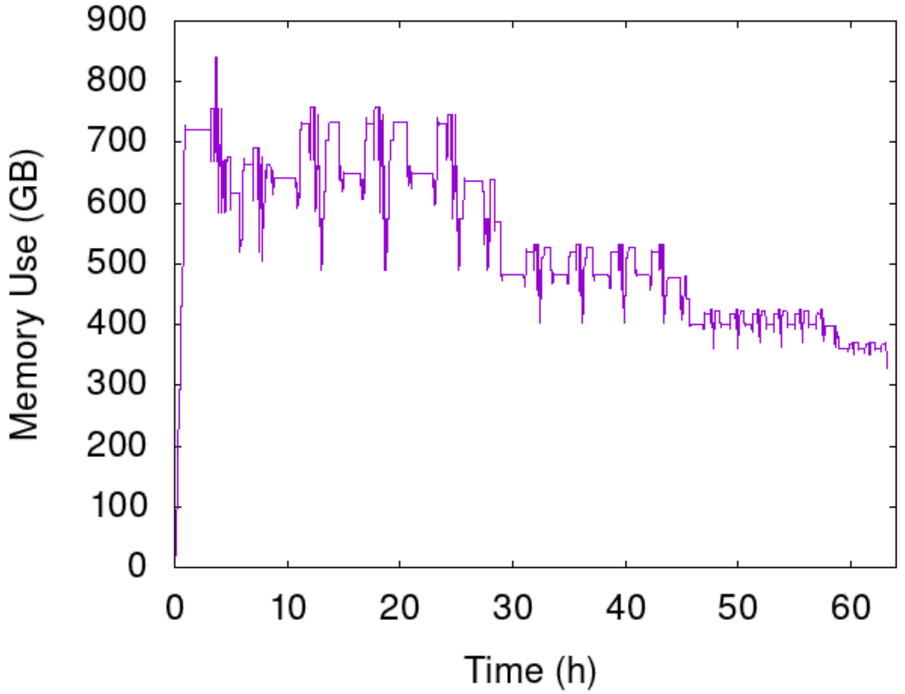
ResNet50

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Analysis

(Batch size: 2048)





Conclusions

- We have used, for the first time, persistent memory technology as a enabler for HE inference
- All of our analysis reveal that interference using HE (CKKS) feature access patterns that yield efficient use of Intel Optane in Memory Mode
- Sequential data accesses in the most common operation (convolution) enable the accessed data to be efficiently cached in DRAM

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you!

guillermo.lloret@bsc.es