OCP FUTURE TECHNOLOGIES SYMPOSIUM

OCP Global Summit

November 8, 2021 | San Jose, CA

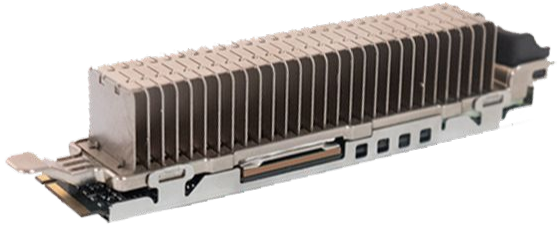
SEAL: Accelerating DLRM with an OCP M.2 Accelerator

Mike Ashby

Tom Lagatta

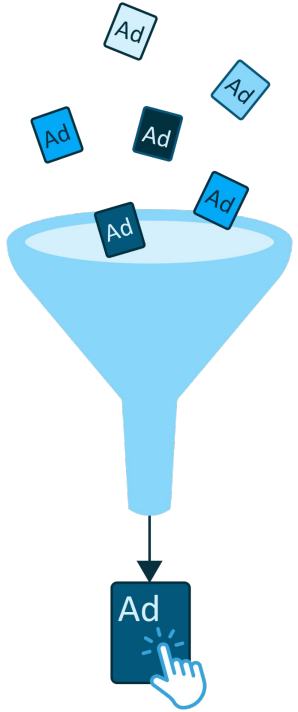
The Challenge

- HW/SW co-design is a rich source of large TCO savings for OCP infrastructure
- Programmable OCP M.2 Accelerator Modules are ideal for realizing these TCO savings
- $3.4B of OCP infrastructure was purchased to run recommendation models in 2019
- Energy requirements are unsustainable
- Recommendation models such as DLRM have unique compute and memory challenges



Recommendation Models

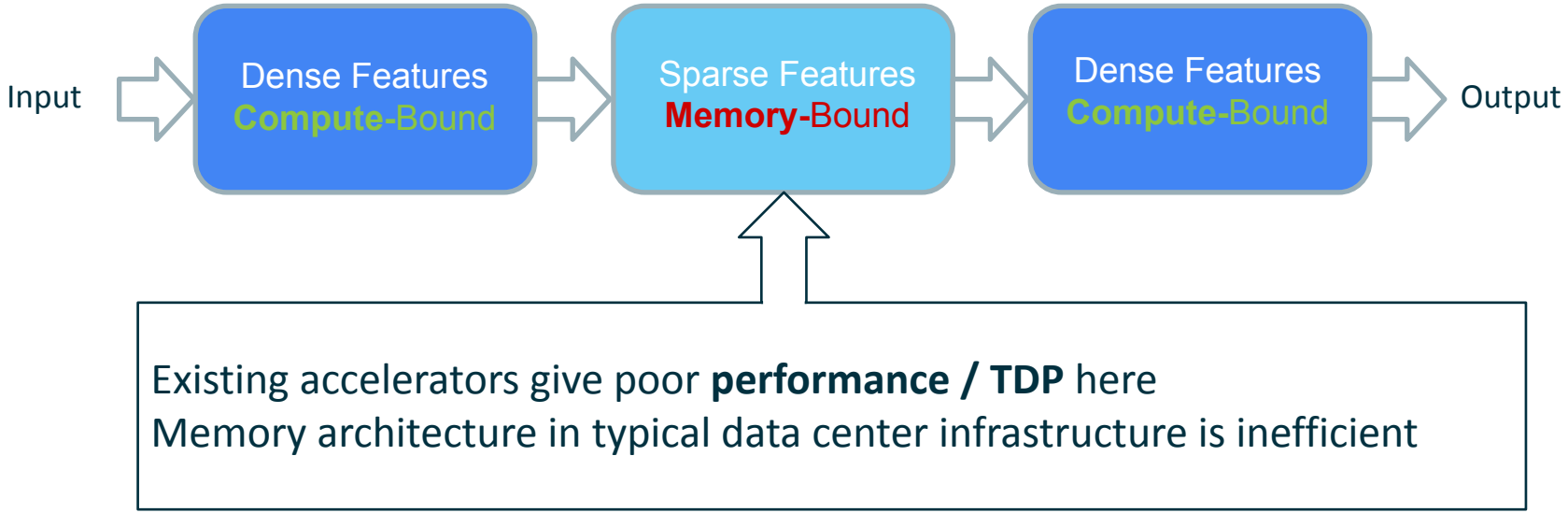
- One of the most common data center workloads
- 79% of AI cycles in some hyperscaler data centers[1]
- Used for search, adverts, feeds, entertainment and personalization



1. U. Gupta, et al. , “The architectural implications of facebook’s dnn-based personalized recommendation,” in 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA) . IEEE, 2020, pp. 488–501.

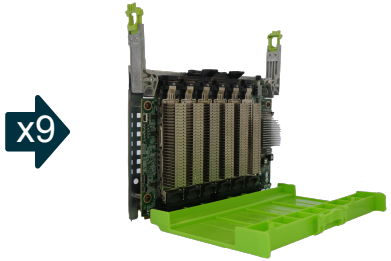
DLRM

- Benchmark Deep Learning Recommendation Model at MLPerf.org



DUT: OCP M.2 Accelerator Module for DLRM

- Compatible with Glacier Point V2 carrier cards, for Yosemite V2 and V2.5 chassis



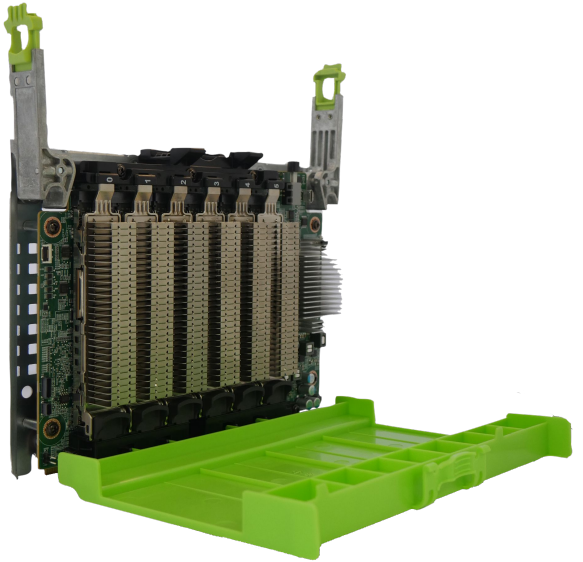
- Standard width
- 14.85 W TDP
- Contains DDR4 Memory (32 GB or 16 GB) and programmable silicon (FPGA)
- Programmable silicon enables support for new operations, data formats and workloads

Adaptable Acceleration

- Programmable silicon in the data center, alongside custom ASICs
- Invaluable now. Invaluable in 5 years time. For multiple workloads

- Quickly explore the hardware-software design space
- Realize $100Ms TCO savings

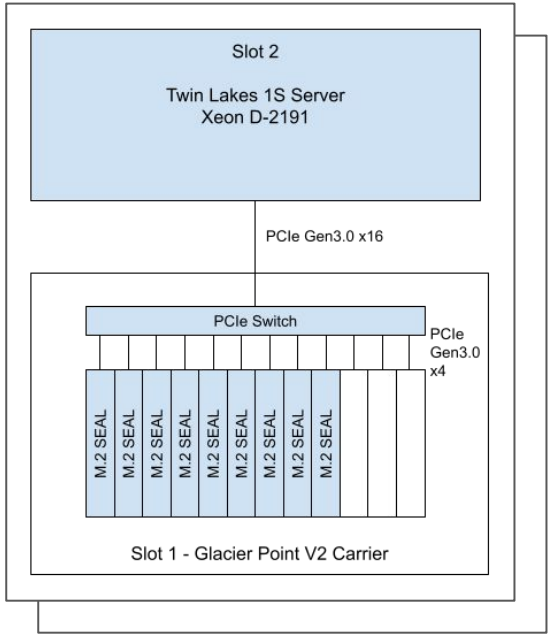
- Reprogrammable via software
- Accelerates Recommendation Models and other memory-bound workloads e.g. k-NN



OCP M.2 Accelerator Modules, each containing 32GB DDR and an FPGA

Benchmarking Setup

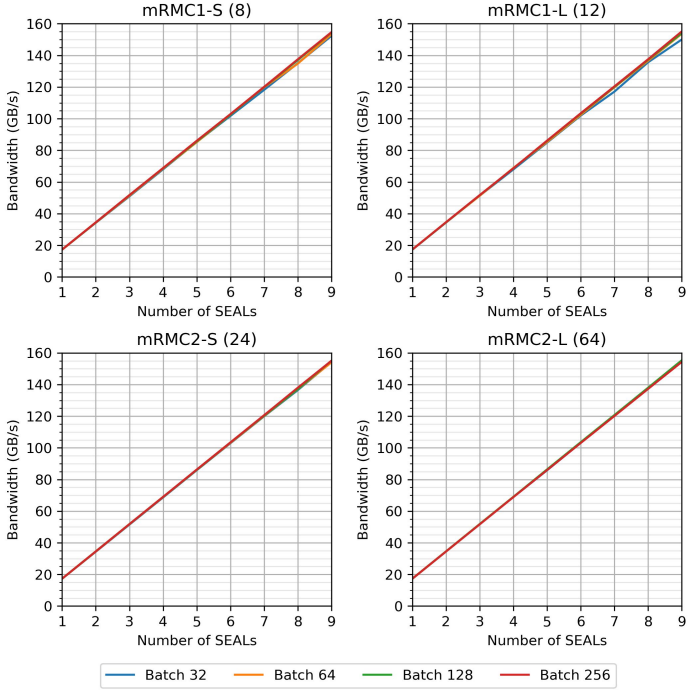
- Characterised on Yosemite V2 and Yosemite V2.5
- SEAL Accelerator™ modules in Glacier Point V2 carrier card
- Up to 9 SEAL Accelerator modules per card
- Adds 576 GB memory capacity per Yosemite
- Adds 162 GB/s memory bandwidth per GPv2



x 2 per Yosemite

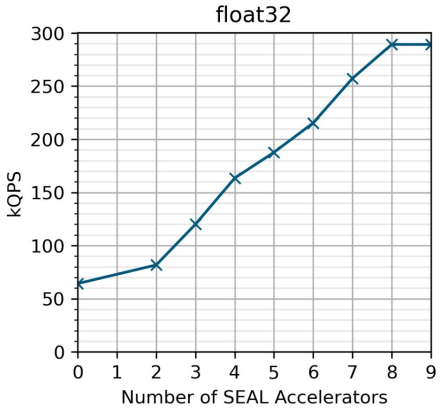
Benchmark results

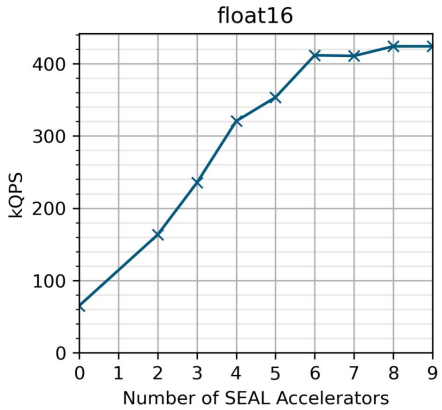
- Random embedding table lookups
- Near perfect linear scaling of achieved bandwidth as workload is scaled up
- On all bit-depths, a wide variety of batch sizes and between 8 - 64 embedding tables per model
- Efficiency of each accelerator was >95%
- **155 GB/s** achieved for random PyTorch EmbeddingBag operations for multiple floating point formats using 9 SEAL Accelerators in GPv2



DLRM benchmark results

- Evaluated using LoadGen traffic generator
- 10 ms P99 latency bounded throughput
- **4.5x** improvement in latency-bound throughput compared with a single precision CPU baseline
- **6.5x** improvement to a float16 CPU baseline
- **2.5x** increase in compute density in the rack, within power budget





SEAL Accelerator

- OCP form factor and production status enable immediate impact
- **2.5x** compute density in an OCP rack
- Up to **59%** reduction in energy use
- **$250M** Capex and **$100M** Opex saving per 100k servers per year
- Datasheet and performance guide available at myrtle.ai


OCP
TOTAL SOLUTIONS

OCP FUTURE TECHNOLOGIES SYMPOSIUM

2021 OCP Global Summit | November 8, 2021, San Jose, CA