Design the Right Storage Systems to Accelerate Deep Learning

Jérôme Gaysse, Silinnov Consulting

Senior Technical & Market Analyst

jerome.gaysse@silinnov-consulting.com

Why improving DL processing?

- Time to market: may need weeks and months to design and/or train a DNN
- Edge computing
  - Not possible to send data to the cloud, and limited local resources.
- More complex DNN

Flash Memory Summit

Technology evolution

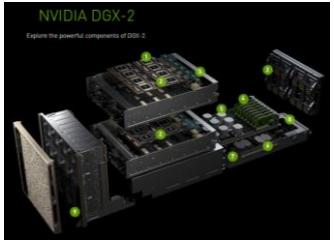
July 1909

April 2005

July 2019

?

2018

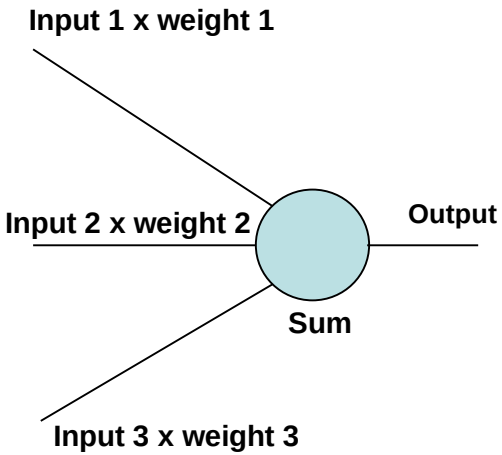


?

Flash Memory Summit

Deel Learning computing

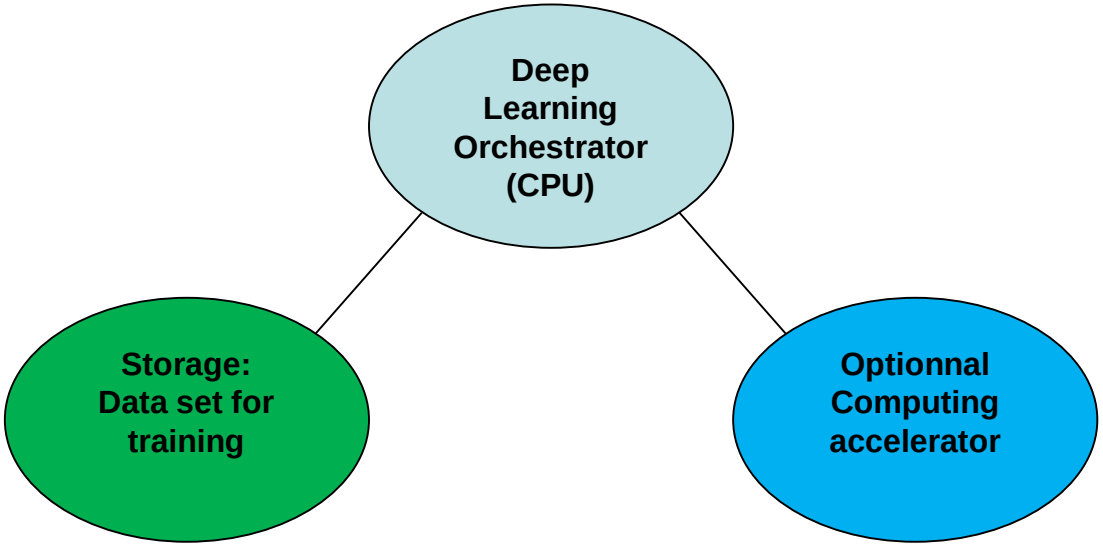
- DNN needs memory, compute and storage
- Restnet50 example:
  - Compute : 3.9GMAc
  - Memory : 25M
  - Storage : training dataset



Flash Memory Summit

DL processing architecture

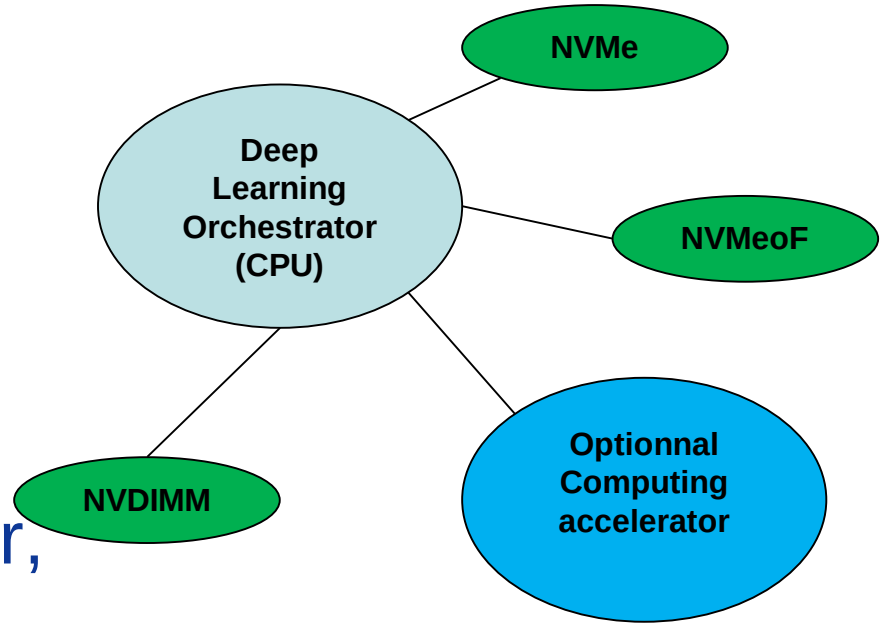
- Mainly based on GPU



Flash Memory Summit

3 storage interface options

- How to reduce data transfer between storage and computing?
- How to reduce power, space and cost?



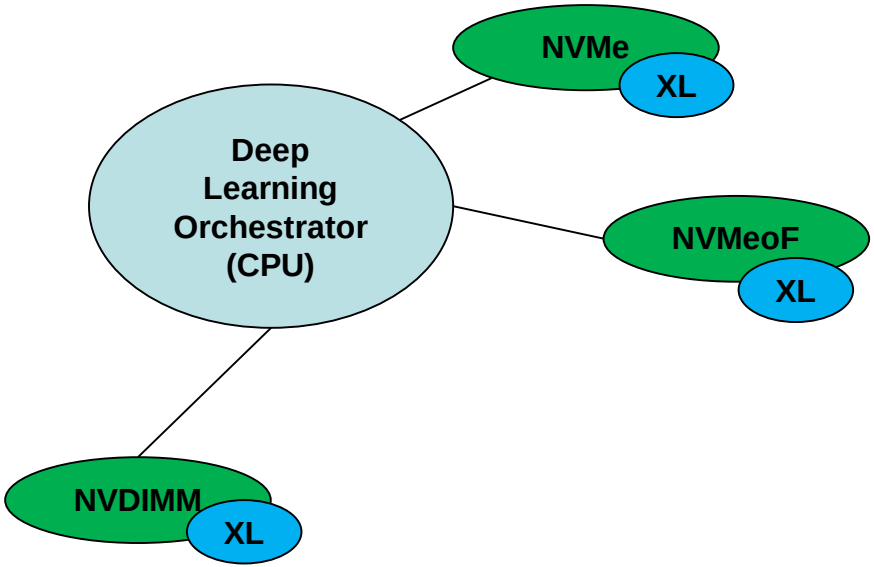
Flash Memory Summit

Let's move compute to storage

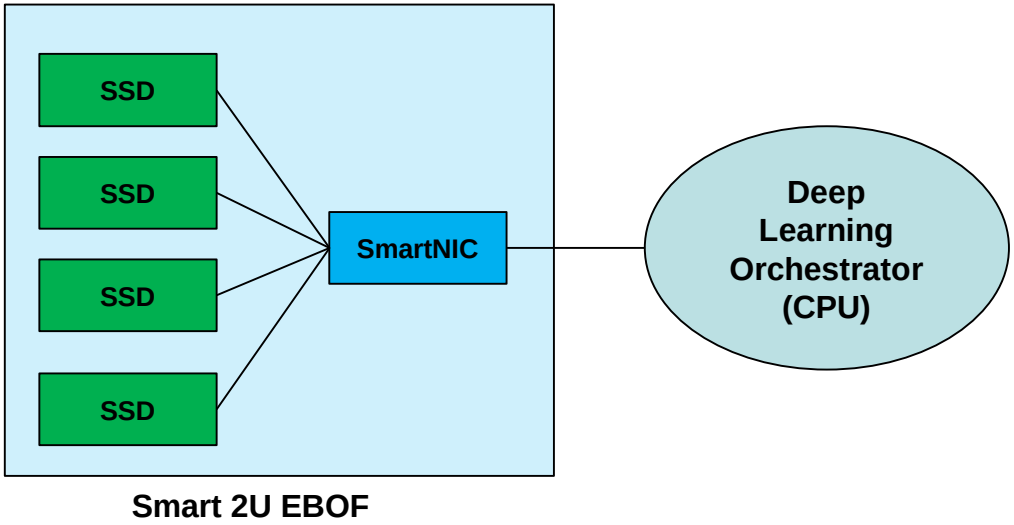
- Smaller accelerator

but

- Massively distributed
  - Increased bandwidth between storage and compute

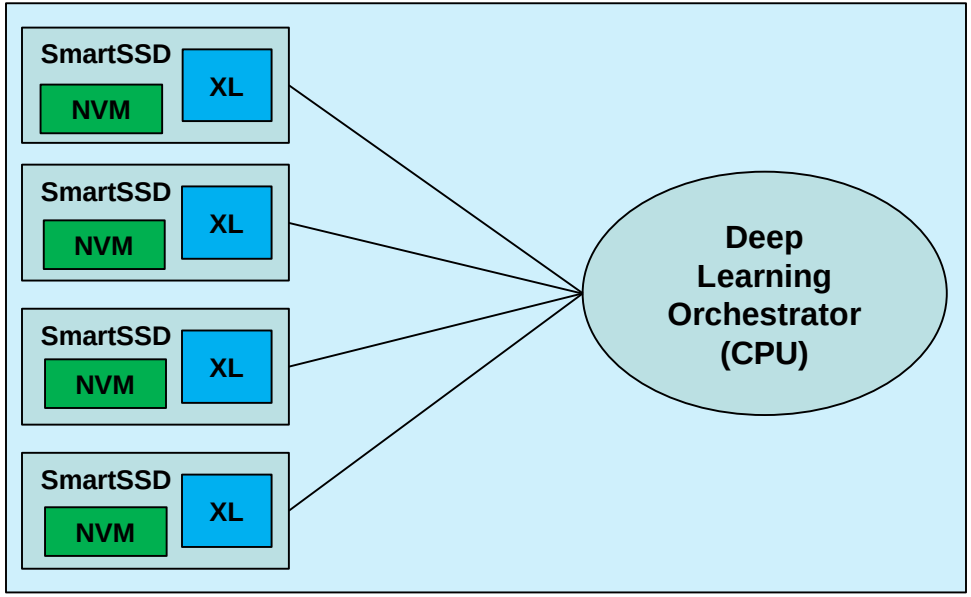


NVMeoF based



- Up to 38GB/s BW between storage and compute in a 2U systems

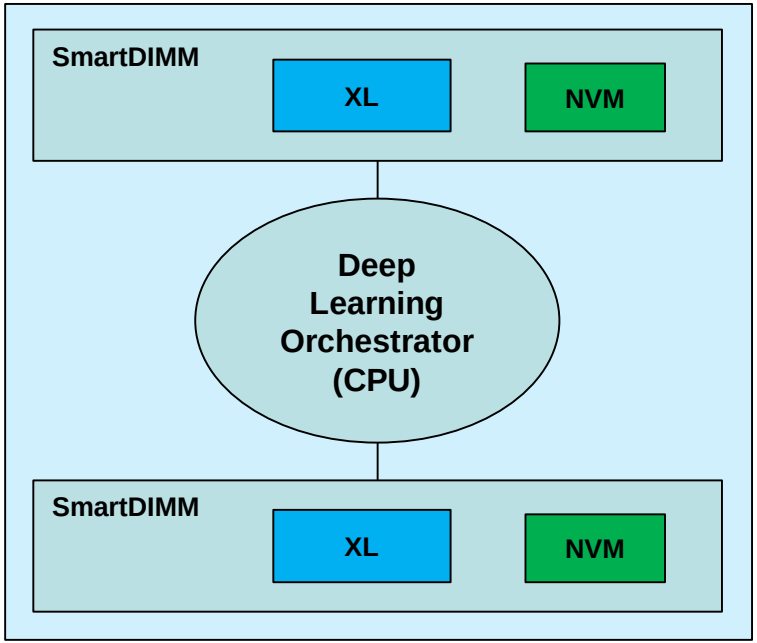
NVMe based



Smart 2U DL Appliance

- Up to 150GB/s BW between storage and compute in a 2U systems

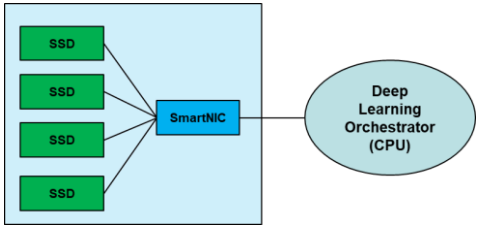
NVDIMM based



Smart 2U DL Appliance

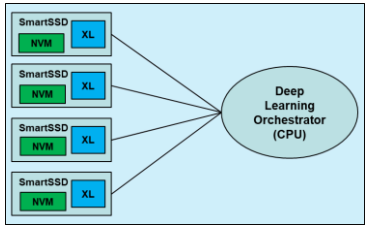
- Up to 19GB/s BW between storage and compute in a 2U systems

Synthesis



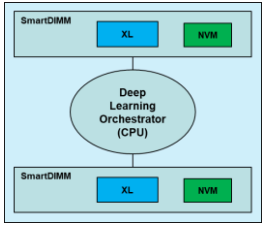
Shared Smart EBOF with mutltiple orcherstrators

Cloud



High BW between compute and storage

Cloud/on-prem



Better integration between XL and CPU thanks to LD/ST access

Cloud

Technology evolution

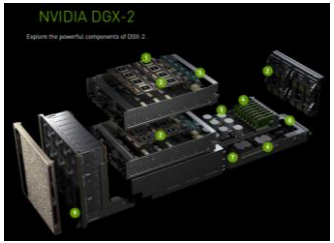
July 1909

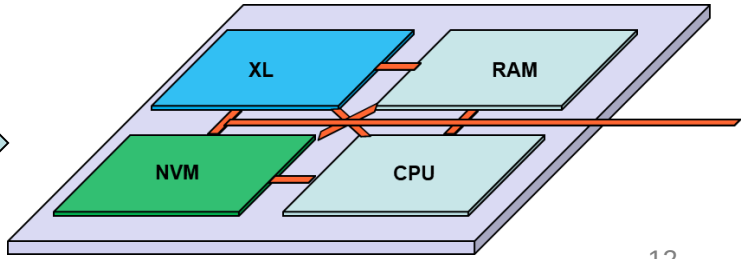
April 2005

July 2019

?

2018





100GB/s between NVM and XL….in a chip!