Intelligent Hybrid Flash Management

Jérôme Gaysse

Senior Technology&Market Analyst

jerome.gaysse@silinnov-consulting.com

Flash Memory Summit

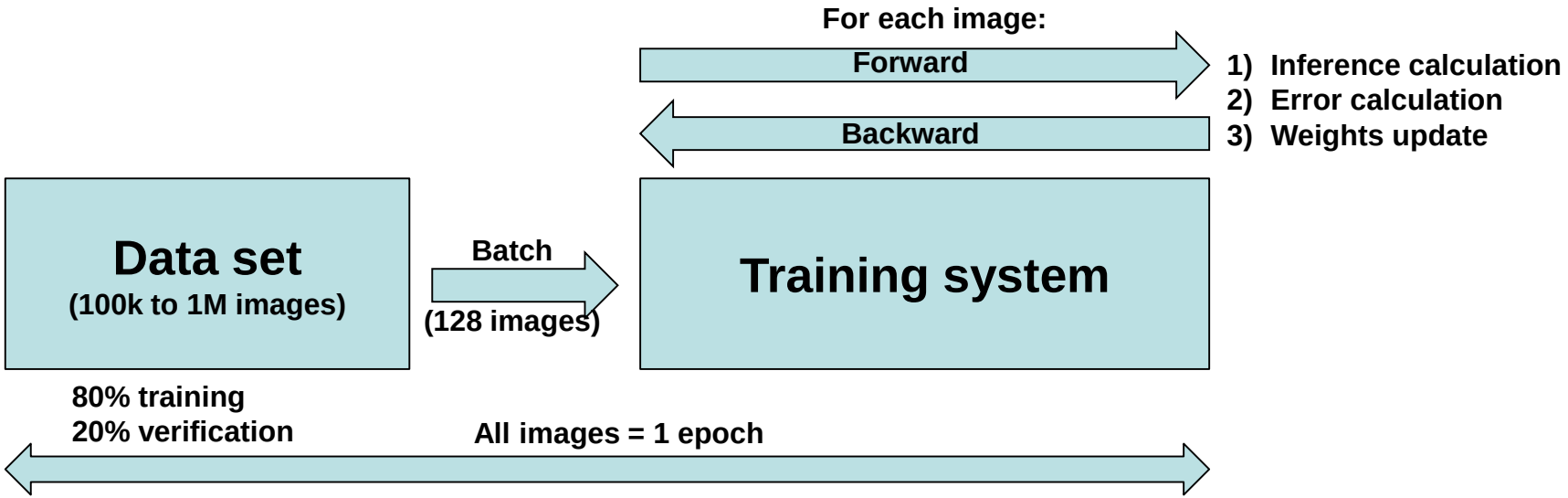
Research context

- Analysis of system & application
- Performance modeling
- Emerging technologies analysis
- New architecture definition
- Performance simulation

Deep learning

- Inference and training
  - Training value = weights value
- Training problem
  - Take a long time: time to market impact
  - Expansive hardware resources : TCO impact

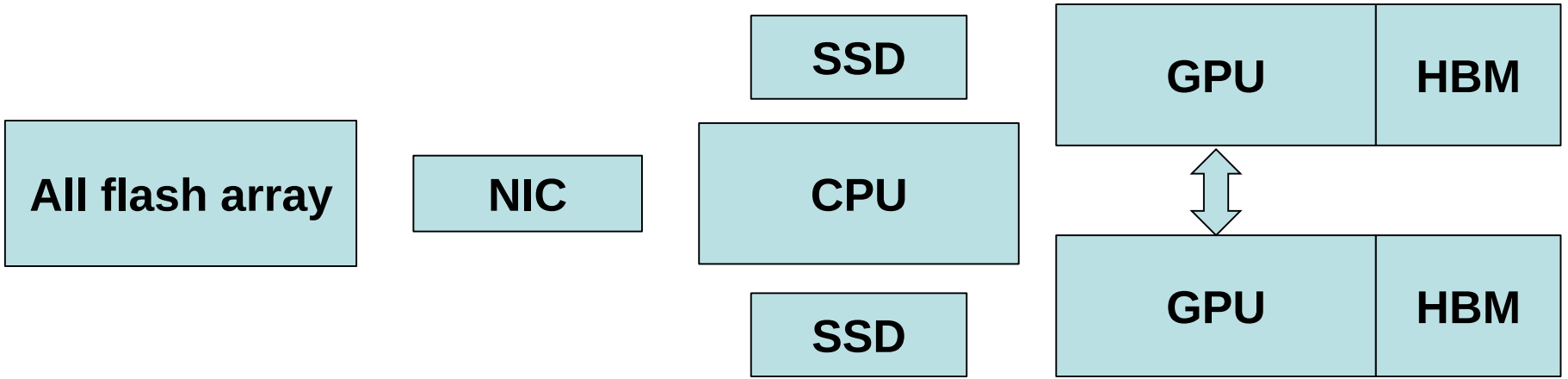
Training process



Training = multiple epochs

Deep learning training

- System architecture



Performance analysis

- Focus on ResNet50 neural network
  - 50 layers
  - 25M parameters
  - 3.9GMACs
- Many benchmarks available
  - Nvidia, HPE, Dell…

Flash Memory Summit

Performance analysis

- Throughput : 400 images per second/GPU
  - FP32 resolution
  - 25M parameters => 100MB model size
- Checkpointing (about every 400 images)
  - =>100MB/s write
- Data set : 100kB images @400fps
  - => 40MB/s read

No IO storage bottleneck

Flash Memory Summit

DL improvements

- From FP32 to FP16 to INT8
  - Less computing requirements
  - Less memory bandwidth requirements
- Pruning (less connexions between neurons)
  - Less computing requirements
  - Less memory bandwidth requirements

Training throughput to increase

Flash Memory Summit

Training performance increase

- With DL optimization,
- and new Deep Learning Processor developement

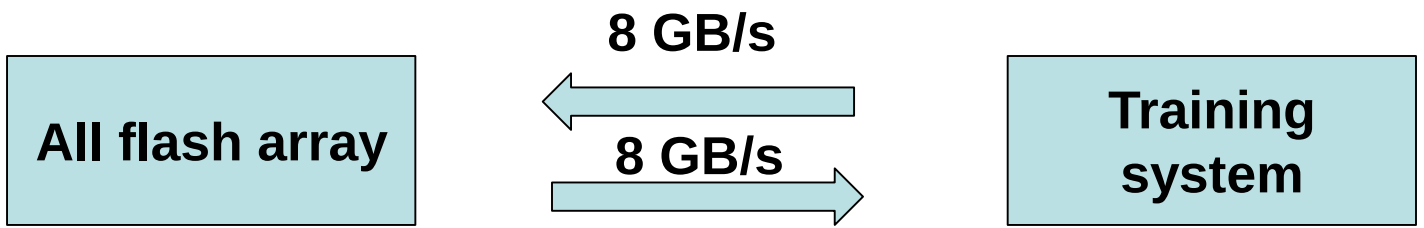
- From 400 FPS to 10,000FPS (estimation)
  - x25

Flash Memory Summit

IO impact

- If throughtput x25
  - Read => 1GB/s (x25)
  - Write => 1GB/s (x10)
    - x25 accesses but less data to write

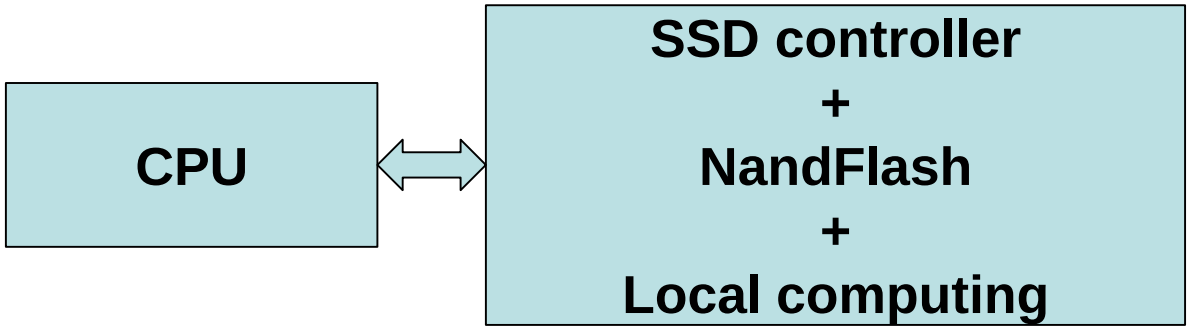
Need new architecture



- Huge data movement between training system and storage array
  - => High power consumption
  - => High silicon cost (AFA controller, NIC…)

Computational storage concept

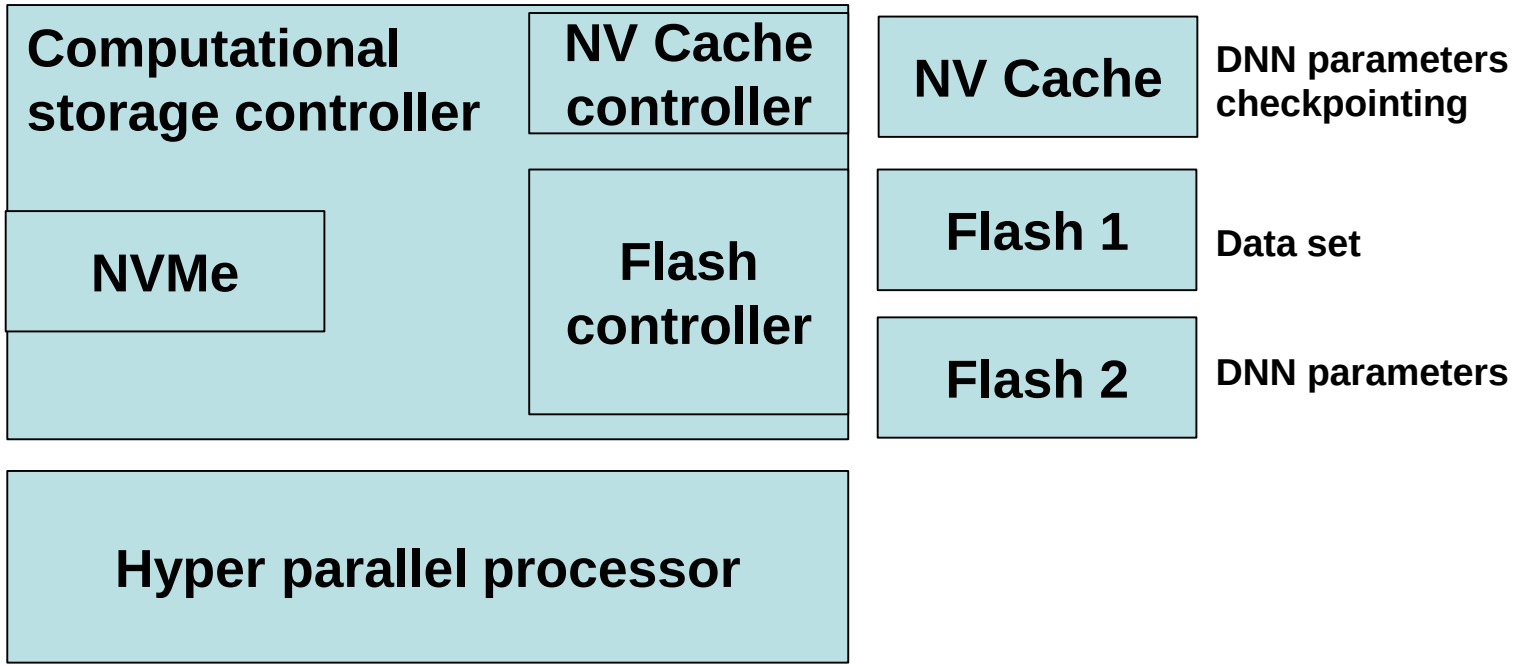
- Computational storage = computing capabilities in the SSD



Reduce data movement:
Less power, higher performance

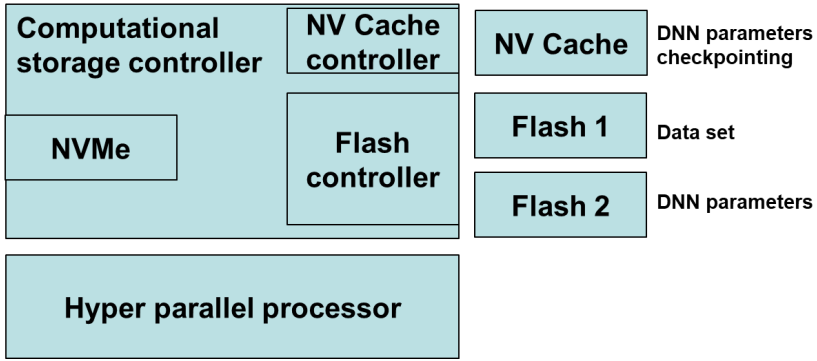
Flash Memory Summit

Computational storage architecture



Theory of operation

- Namespaces
  - Weights
  - Data set
- Vendor specific commands
  - Start training



DNN weights storage, why?

- NV Cache could be enough
- Only few dozens of MB: small storage
- Another reason?
  - Yes, for weight convergence analysis

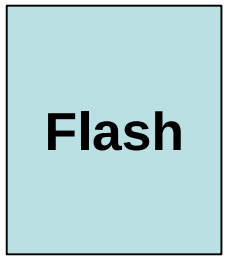
Multi Flash requirements

- Dataset storage
  - Capacity: 100G-1TB range
  - Read only : 1GB/s
- Weight storage
  - Capacity : 5TB
  - Write only: 1GB/s

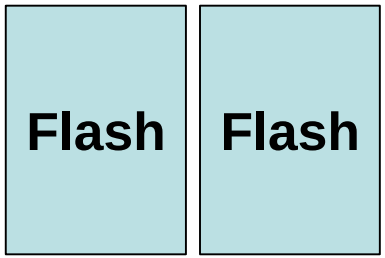
Flash options

- Criteria
  - Performance, cost, software simplicity

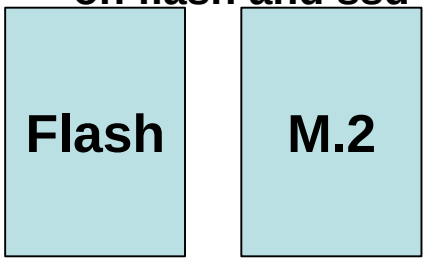
Data set and weights on the same media



Data set and weights on 2 media

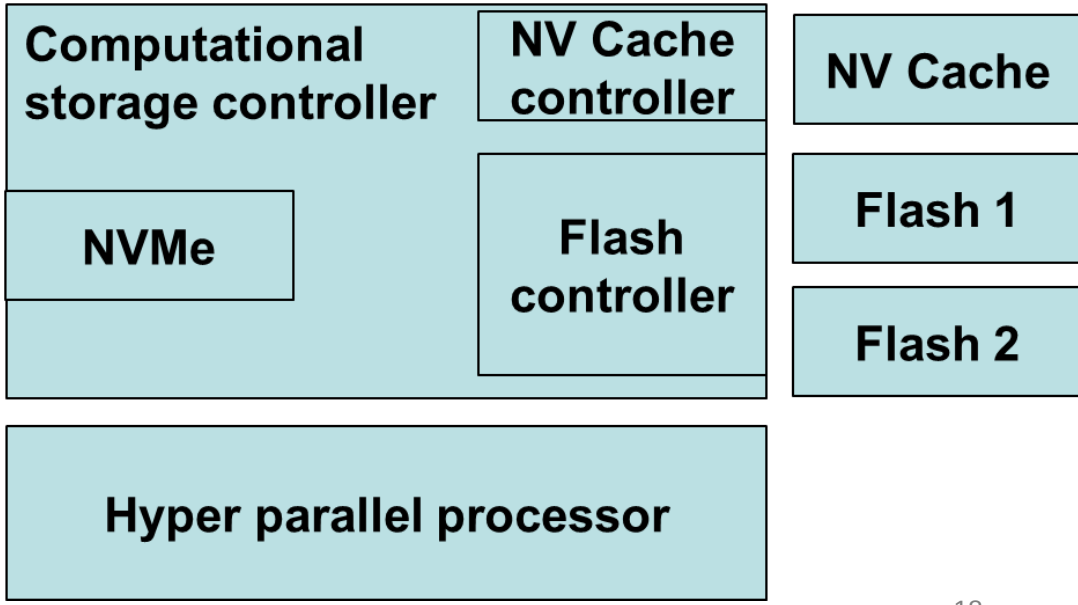


Data set and weights on flash and ssd

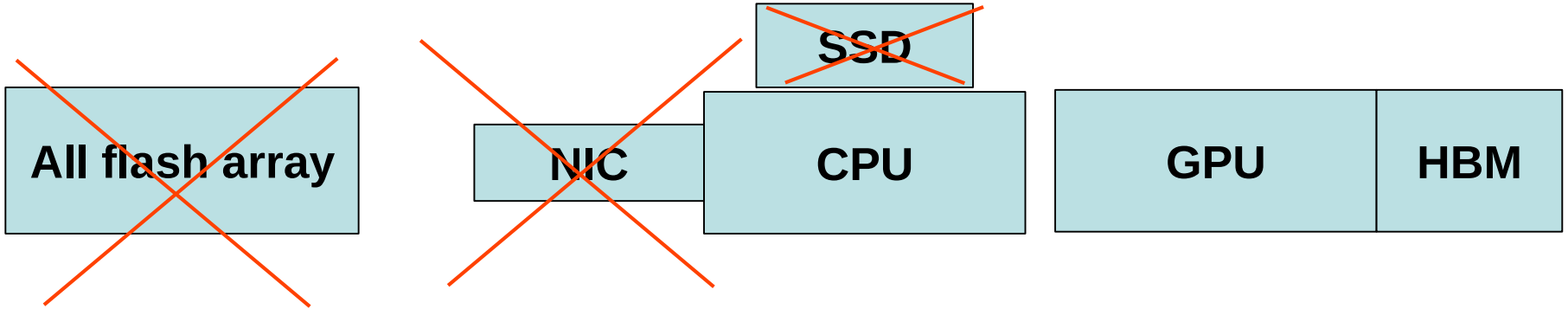


Embedded processing

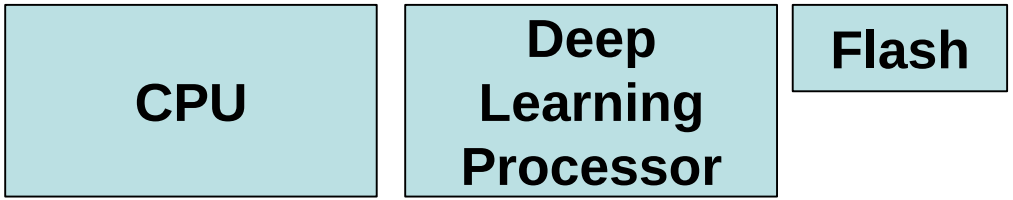
- KeyValueStore: for data set manegement
- Preprocessing
- DL



System benefit



- Removing hardware cost & power



Computational storage

Next research steps

- Simulation at system level
  - Detailed storage access (latency)
- Computational storage applied to NVDIMM?
- Benefits of new interconnects?
  - GenZ, OpenCapi, CCIX

Flash Memory Summit

Want to know more?

SDC 18
STORAGE DEVELOPER CONFERENCE

A Comparison of In-storage Processing Architectures and Technologies

Monday Sept 24, 10:35 AM - 11:25 AM