COMPF4:
AI Hardware Summary

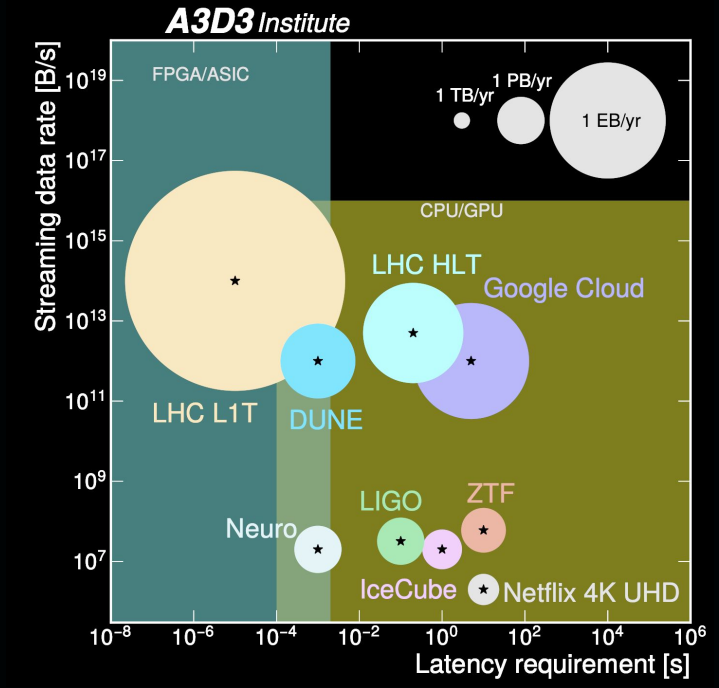
Javier Duarte (jduarte@ucsd.edu),
Nhan Tran (ntran@fnal.gov)

UC San Diego

Fermilab

Challenges

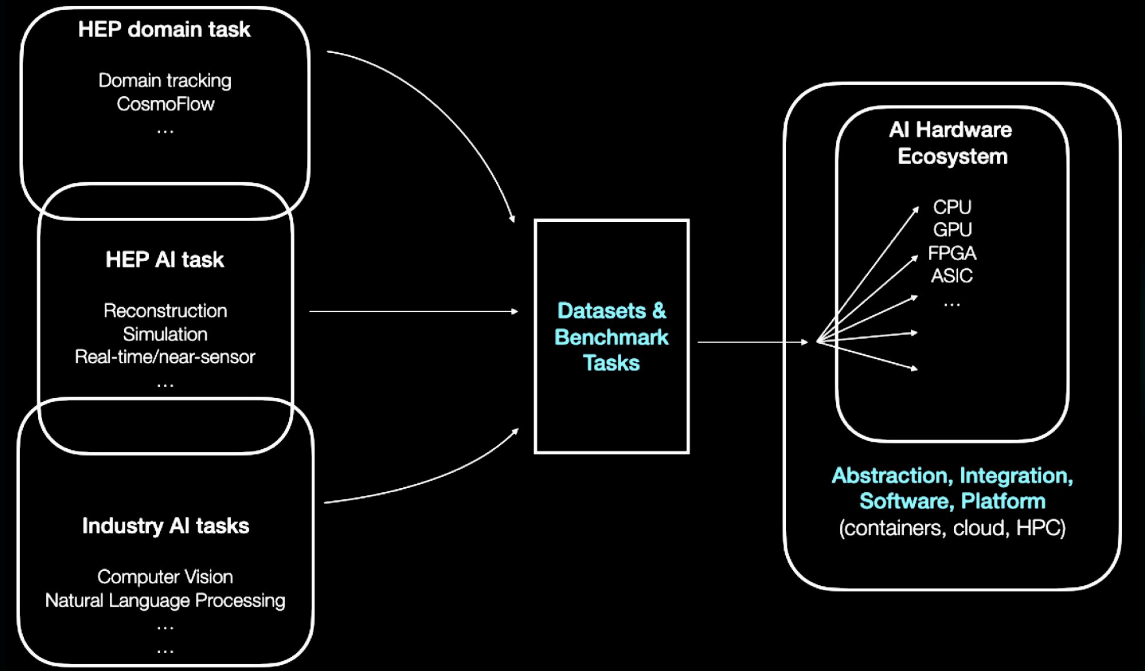
- HEP computing challenges go beyond traditional industry workloads in data rates, latency/throughput requirements, data volumes, and data representations



Scope/definition

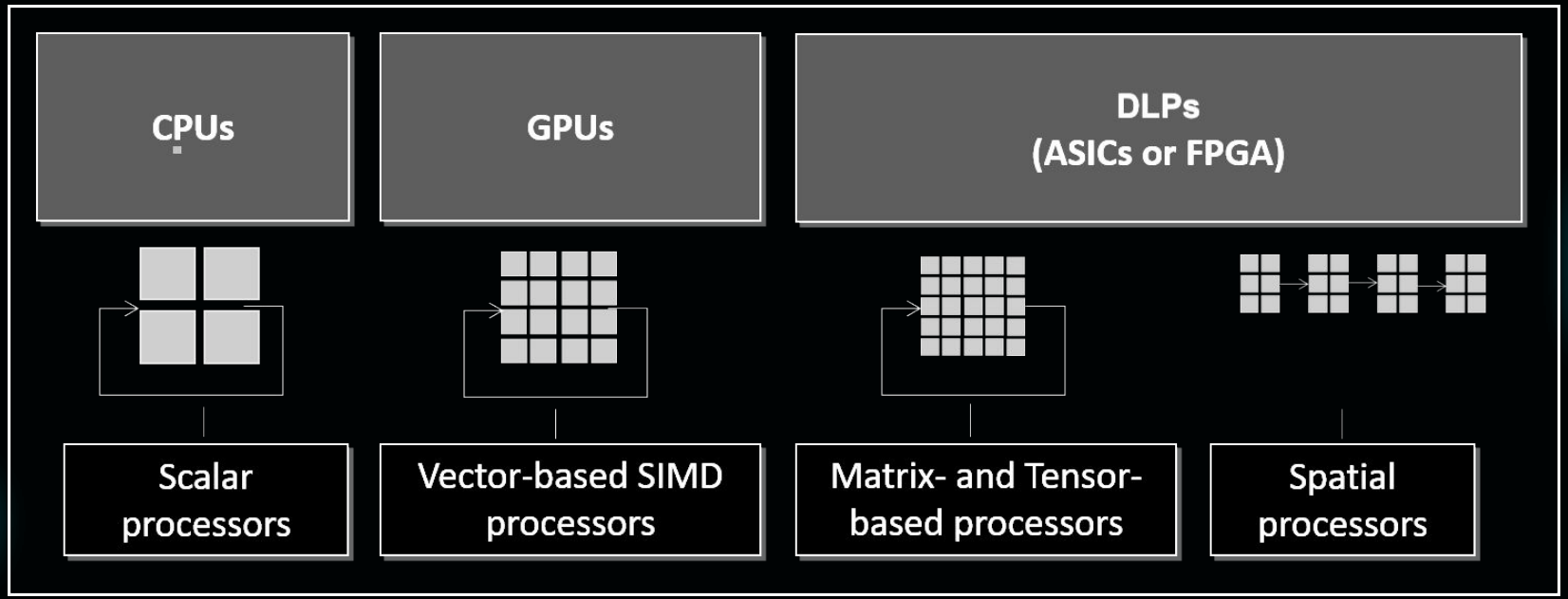
- Application of novel **AI hardware** for accelerating offline data processing
- Closely related focus areas: CompF03 (ML), IF04 (instrumentation trigger and data acquisition) (IF04), and IF07 (electronics/ASICS)

AI hardware ecosystem: connection with HEP challenges including areas for development, benchmarking and abstraction



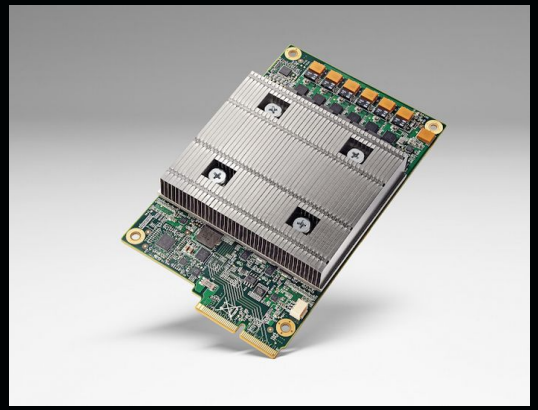
Hardware Taxonomy

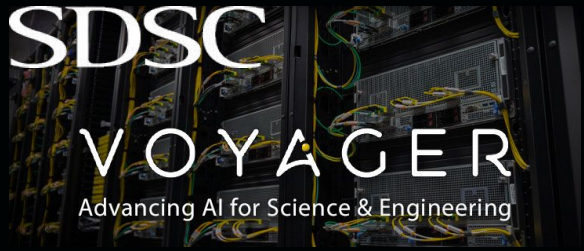
- Scalar processors (CPUs), vector-based processors (GPUs), and deep learning processors (DLPs)
- DLPs are specialized for this application domain: often implemented with ASICs or FPGAs

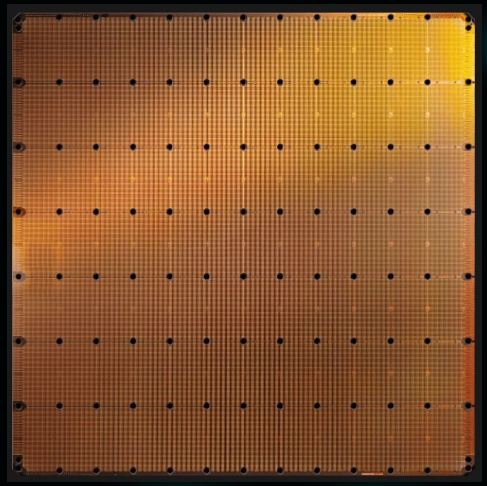


Examples

- Google TPU, Habana Goya, Cerebras WSE, …







Benchmarks

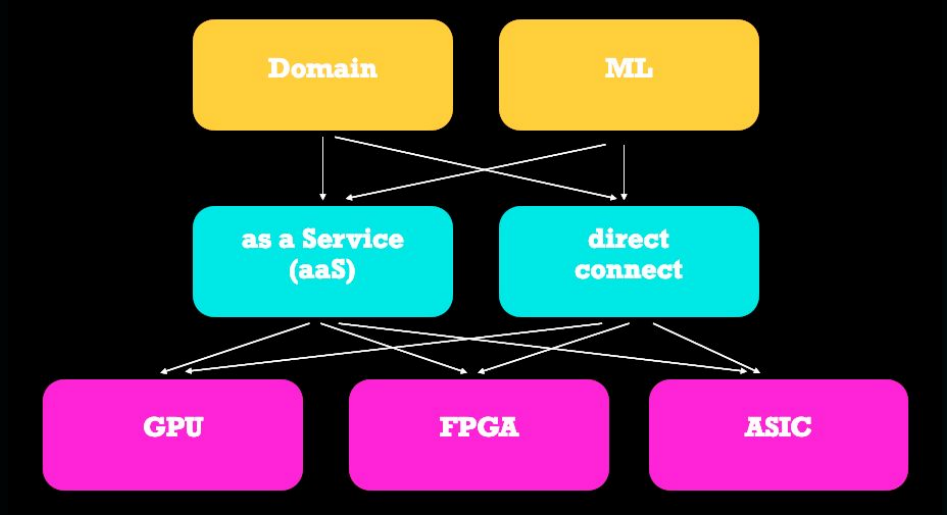
- Need for high-quality physics workload benchmarks to test and evaluate AI hardware
- E.g. MLCommons runs benchmarks like MLPerf Inference: Data Center
  - mlcommons.org/en/inference-datacenter-20
- FastML Science benchmark: arXiv:2207.07958

ML Commons

Area	Task	Model	Dataset	QSL Size	Quality	Server latency constraint
Vision	Image classification	Resnet50-v1.5	ImageNet (224x224)	1024	99% of FP32 (76.46%)	15 ms
Vision	Object detection (large)	SSD-ResNet34	COCO (1200x1200)	64	99% of FP32 (0.20 mAP)	100 ms
Vision	Medical image segmentation	3D UNET	KITS 2019 (602x512x512)	16	99% of FP32 and 99.9% of FP32 (0.86330 mean DICE score)	N/A
Speech	Speech-to-text	RNNT	Librispeech dev-clean (samples < 15 seconds)	2513	99% of FP32 (1 - WER, where WER=7.452253714852645%)	1000 ms
Language	Language processing	BERT-large	SQuAD v1.1 (max_seq_len=384)	10833	99% of FP32 and 99.9% of FP32 (f1_score=90.874%)	130 ms
Commerce	Recommendation	DLRM	1TB Click Logs	204800	99% of FP32 and 99.9% of FP32 (AUC=80.25%)	30 ms

Software abstraction & integration

- Different deployment tactics
  - As a service (aaS)
  - Direct connect
- Considerations:
  - Flexibility, cost-effectiveness, symbiosis, simplicity, containerization/orchestration, portability



Summary

- Recommendations include:
  - Developing physics workload benchmarks to evaluate AI hardware
  - Studying deployment strategies and synergy with existing infrastructure
  - Leveraging HPC testbeds like SDSC Voyager
  - Remaining nimble: fast-moving industry!
- Current draft of report: dropbox.com/s/9xpdo1vm31beo70/SnowmassBook_CompF4.pdf
  - Any feedback? Send to us!