ENIAD: A Reconfigurable Near-data Processing Architecture for Web-Scale AI-enriched Big Data Service

Jialiang Zhang

Jing(Jane) Li

Department of Electrical and Systems Engineering

University of Pennsylvania

Hot Chips 33, August 22-24,2021

Abstract

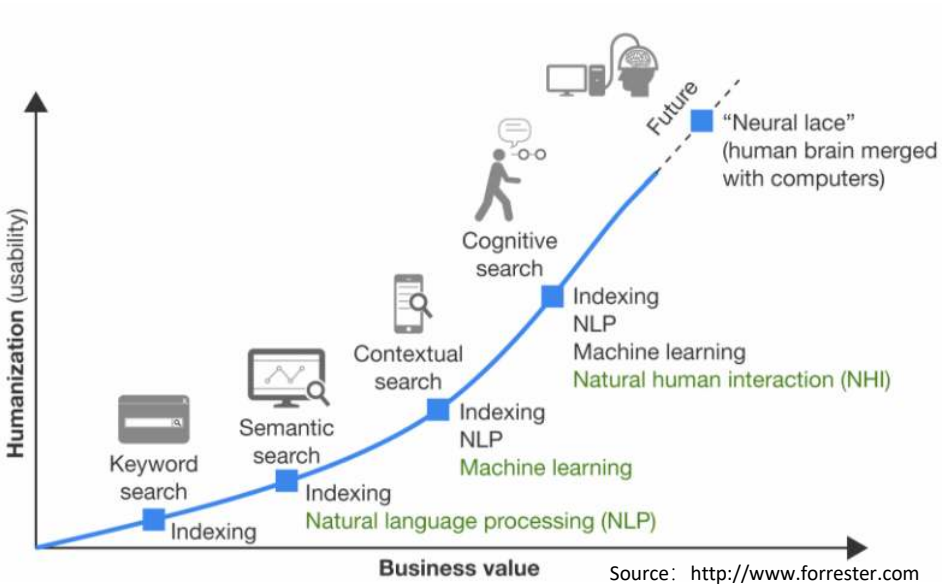
To meet the surging demands required by AI-enriched Big Data services, cloud vendors are turning toward domain specific accelerators for improved efficiency, scalability and performance.

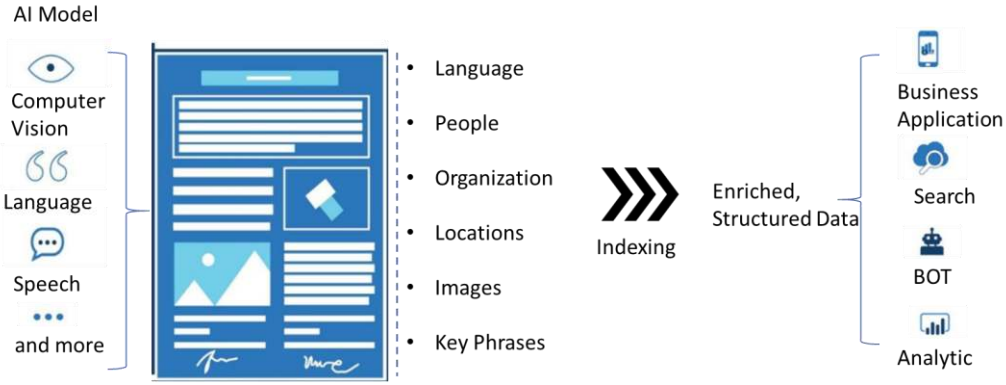
ENIAD, the first end-to-end infrastructure for AI-enriched Big Data serving in real time, accelerates both deep neural network inferencing and billion-scale indexing at the data-center scale. Exploiting near-data computation, reconfigurable computing and rapid/agile hardware deployment flow, ENIAD serves state-of-the-art, online built indexing service with high efficiency at low batch sizes.

A high-performance, index (data)-adaptable FPGA soft processor is at the heart of the system and able to serve 10x larger index size with 14x lower latency compared to state-of-the-art CPU and GPU architectures.

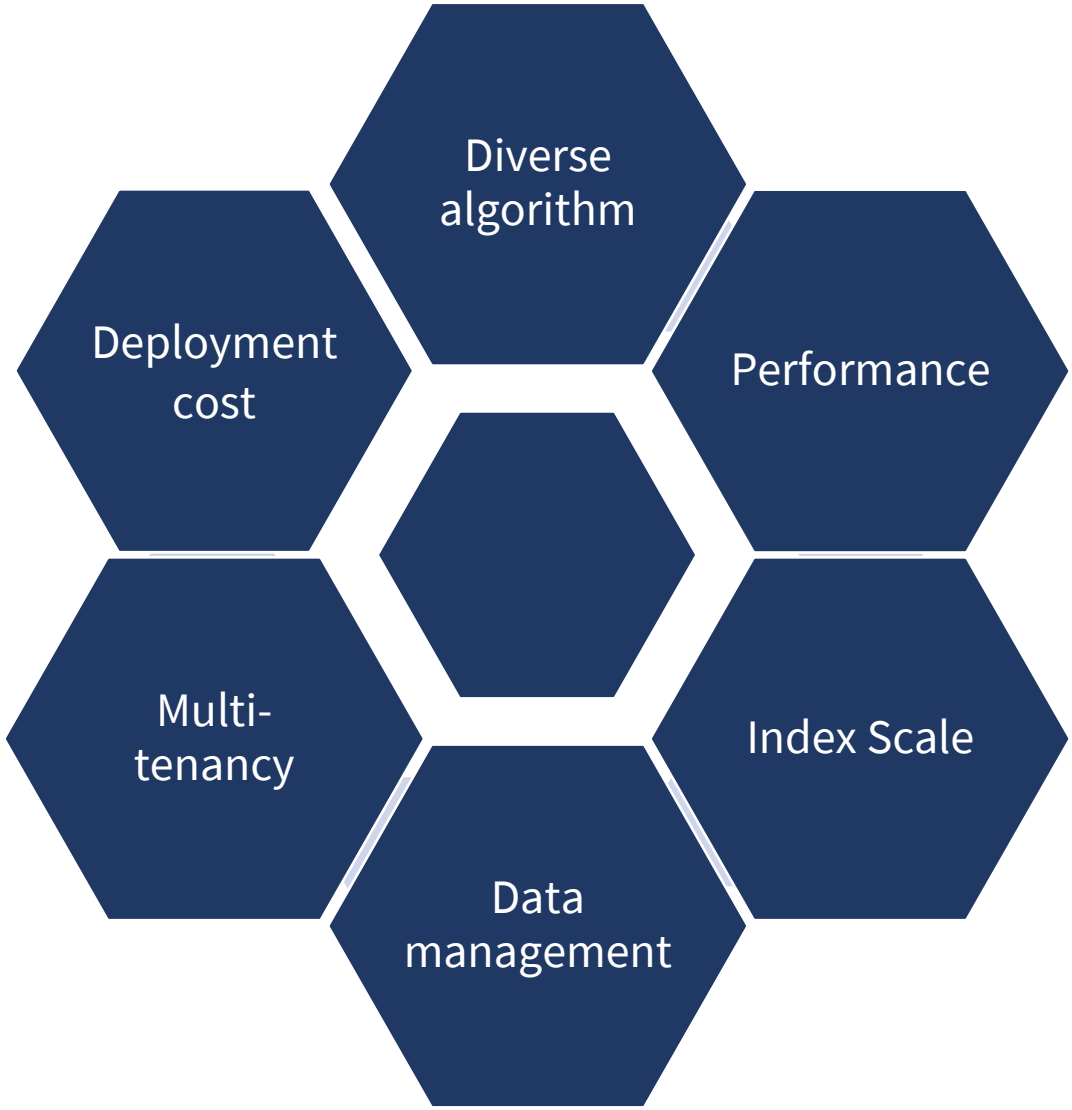
The Rise of Cognitive Search

- **Core of the next-generation intelligent data analytic service**
  - Full text search
  - Business analytics
  - Content-based e-commerce site search
  - Video indexing
  - Knowledge mining for data science
  - And more…





Challenges of Serving Cognitive Search at datacenter-scale



It is extremely challenging to design **specialized hardware accelerator** to meet all constraints at data-center scale.

ENIAD: A Scalable FPGA-powered Platform for serving Cognitive Search

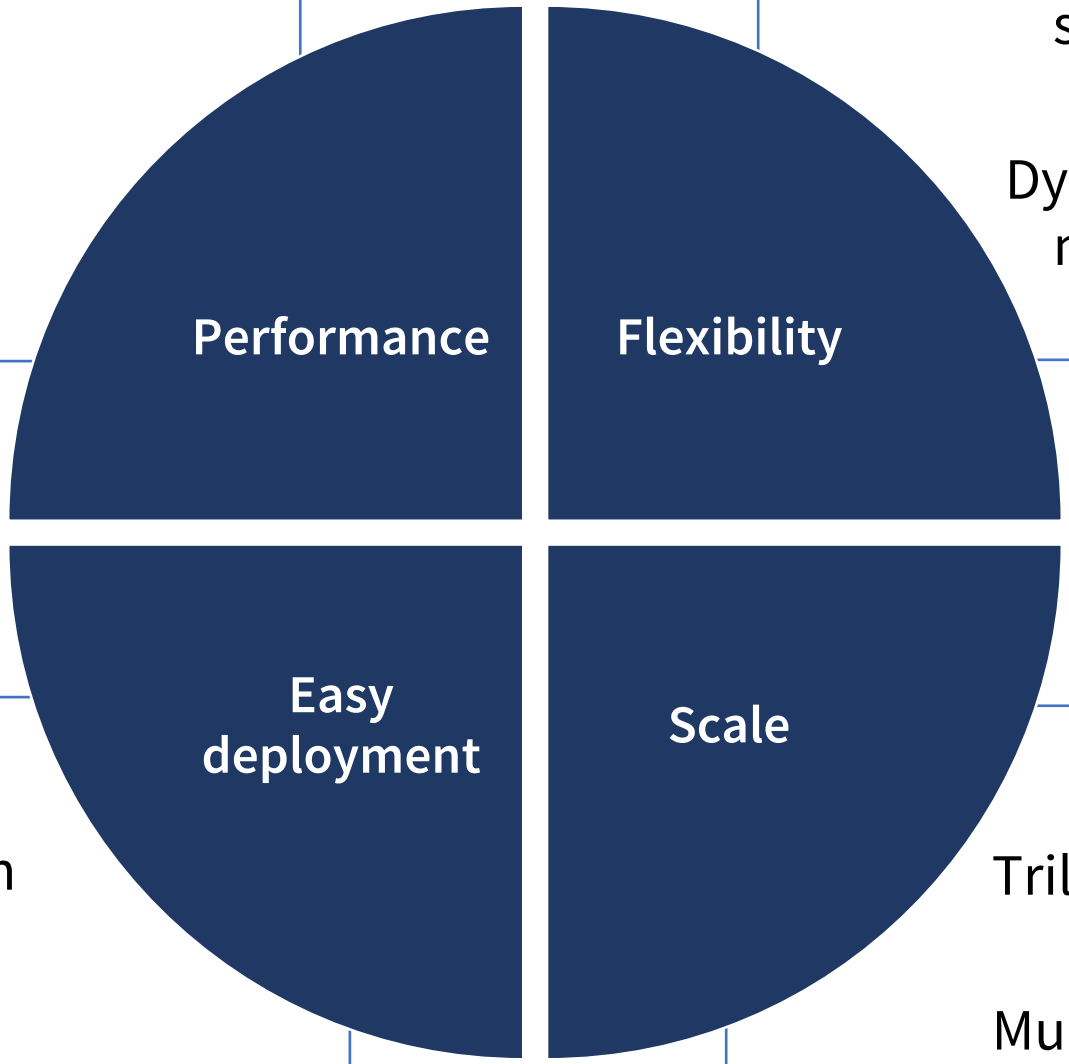
Ultra low latency

High throughput

Low batch size

Diverse indexing scheme

Dynamic data management



Rapid Compilation

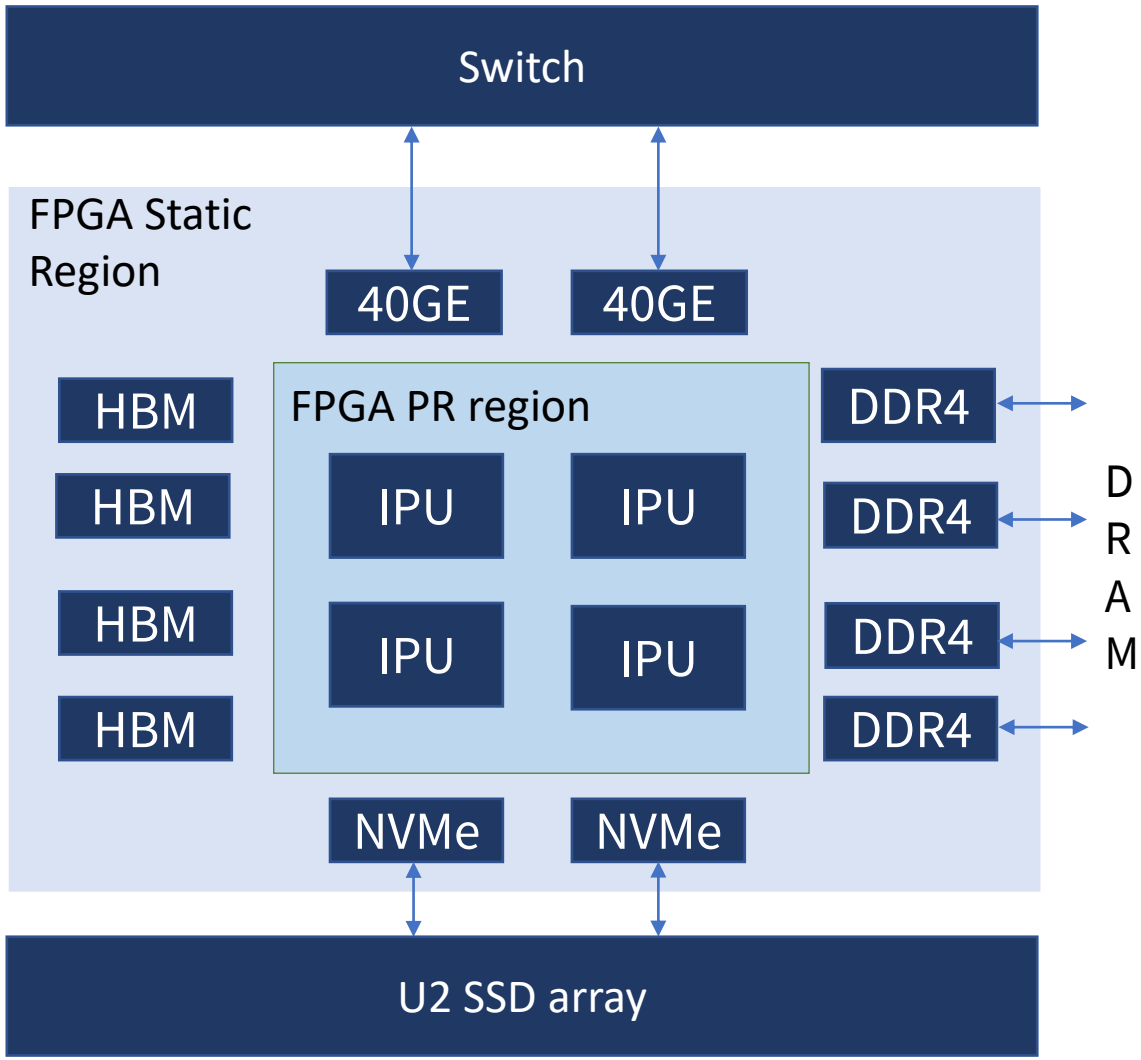
Turnkey development

Trillion-scale index

Multi-tenancy

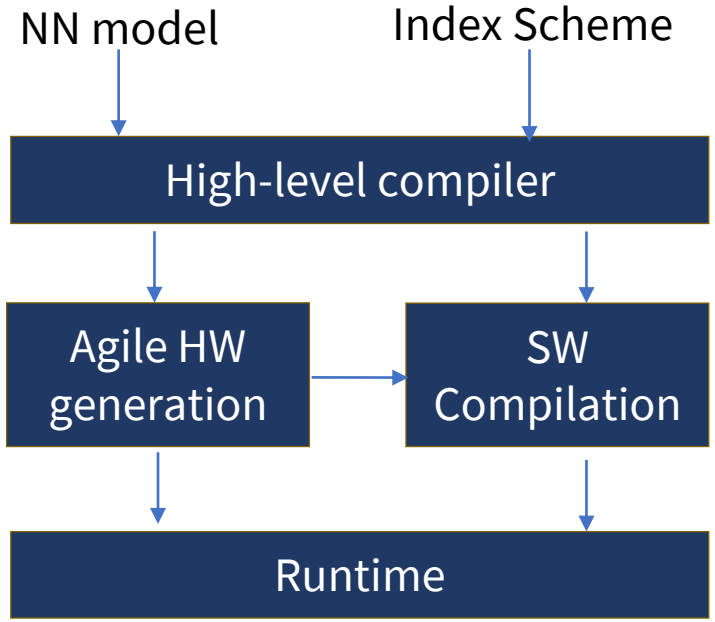
ENIAD Hardware

- **Field-configurable IPU**
  - Highly customized for each indexing scheme
  - Fast deployment using partial reconfiguration
  - High performance :
    - 20 TFLOPS tensor op/s
    - 3T hashing op/s
    - 811 GTEPS graph performance
    - 4T table lookup/s
- **Near mem/storage computation**
  - 160 GB/s SSD bandwidth
  - 2TB/s memory bandwidth



Framework Integration + Development

- **Software API**
  - Seamless integration with popular frameworks: PyTorch, Milvus, etc.
- **Agile hardware generation**
  - Fast generation within minutes
- **Runtime:**
  - Orchestration
  - FPGA Partial Reconfiguration
  - Memory and storage management



End-to-end Performance

NLP Index: MS GEN Encoder + HNSW (Graph Index)				
	1 CPU	16CPU	1 ENIAD Node	Improvement
Index Size	100M	1B	10B	ENIAD serves 10× larger index at 14× lower latency
E2E latency Per batch 1 request @ 95%	29ms	9.8ms	0.71ms	
Index Build Time	23 mins	4hrs	1 hrs	

Image index: Deep1B + IVFPQ (Inverted File + Quantization Index)				
	1 GPU	16 GPU	1/4 ENIAD Node	Improvement
Index Size	1B	1B	10B	ENIAD serves the same index size with 4× fewer nodes at 68× lower latency
E2E latency Per batch 1 request @ 95%	198ms	89ms	1.3ms	
Index Build Time	1 mins	11 mins	18 mins	