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…

Source: http://www.forrester.com
Challenges of Serving Cognitive Search at datacenter-scale

- Deployment cost
- Diverse algorithm
- Performance
- Index Scale
- Data management
- Multi-tenancy

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
- Rapid Compilation
- Turnkey development
- Diverse indexing scheme
- Dynamic data management
- Trillion-scale index
- Multi-tenancy
- Performance
- Flexibility
- Easy deployment
- Scale

Performance

Flexibility

Easy deployment

Scale
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)

<table>
<thead>
<tr>
<th></th>
<th>1 CPU</th>
<th>16 CPU</th>
<th>1 ENIAD Node</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index Size</td>
<td>100M</td>
<td>1B</td>
<td>10B</td>
<td>ENIAD serves 10× larger index at 14× lower latency</td>
</tr>
<tr>
<td>E2E latency Per batch 1 request @ 95%</td>
<td>29ms</td>
<td>9.8ms</td>
<td>0.71ms</td>
<td></td>
</tr>
<tr>
<td>Index Build Time</td>
<td>23 mins</td>
<td>4hrs</td>
<td>1 hrs</td>
<td></td>
</tr>
</tbody>
</table>

### Image index: Deep1B + IVFPQ (Inverted File + Quantization Index)

<table>
<thead>
<tr>
<th></th>
<th>1 GPU</th>
<th>16 GPU</th>
<th>1/4 ENIAD Node</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index Size</td>
<td>1B</td>
<td>1B</td>
<td>10B</td>
<td>ENIAD serves the same index size with 4× fewer nodes at 68× lower latency</td>
</tr>
<tr>
<td>E2E latency Per batch 1 request @ 95%</td>
<td>198ms</td>
<td>89ms</td>
<td>1.3ms</td>
<td></td>
</tr>
<tr>
<td>Index Build Time</td>
<td>1 mins</td>
<td>11 mins</td>
<td>18 mins</td>
<td></td>
</tr>
</tbody>
</table>