AIBench Scenario: Scenario-distilling AI Benchmarking

Tianshu Hao
Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
AIBench Tutorial at ASPLOS 2021
Apr. 14, 2021
The Landscape of Modern Workloads

- Big Data, Machine learning (AI), Internet services
- Industry-scale Application Scenarios
  - Microservice-based Architectures
    - Complex execution paths & A diversity of AI and non-AI modules
  - The whole process from receiving user request to feeding back results
  - Online service + Offline training (Real-time updating of AI models)

**Industry-scale application scenarios**

Query requests

Feedback

Online Service

Incremental data

Model updating

Offline Training
Modern Workloads vs. Traditional Workloads (1)

- **Code size**
  - **One hundred thousand lines** of code for traditional desktop workloads
    - data compression, image manipulation
  - **Hundreds of thousands of lines** of code for web server workloads
    - Nginx server
  - **Millions of lines** of code for modern workloads
    - Complex runtime environment stacks, e.g., TensorFlow
Modern Workloads vs. Traditional Workloads (2)

- Deployment scale
  - Single node for traditional desktop workloads
  - A small scale cluster for web server workloads
  - A large-scale cluster for modern workloads
    - Tens of thousands of nodes
**Execution Path**
- Modern Internet service adopts microservice based Architecture
  - A collection of loosely coupled services
    - Various modules and complex execution path
    - Massive scale and complex hierarchy of infrastructure
Modern workloads dwarf the traditional one in terms of code size, deployment scale, and execution path!

Challenges to Traditional Benchmarking Methodology
Is Micro Benchmark Sufficient?

- AI workloads need to consider both computational efficiency and model quality
  - FLOPS is no longer the only metric
- Mixed-precision training significantly improve FLOPS, however, it may deteriorate the model quality

FLOPS comparison of ResNet50 model and operators

The ResNet50 quality comparison
Is Micro Benchmark Sufficient?

- No quality information
- Impact of data input for different layers
- No single kernel
Example: No Single Kernel

- The kernels’ runtime breakdown of 17 AI workloads
  - Some micro benchmarks may occupy a little percentage
Is Component Benchmark Sufficient?

Benchmarking with a Single Component

Putting the Component into a realistic scenario:

Online Translation Intelligence

Only a single component may need to error-prone conclusions

Single Component vs. Realistic Application

- **E-commerce Search Intelligence**
- The overall system tail latency deteriorates even 100X comparing to a single component tail latency
  
  - **2.2X** comparing to recommendation component
  - **180X** comparing to text classification component

Benchmarking a single component cannot reflect the overall system’s effects

For *E-commerce Search Intelligence*

- Model accuracy improvement **1.5%** => overall system 99th percentile latency deteriorates by **9.7X**
  - Replace ResNet50 with ResNet152 for image classification
    - Overall system 99th percentile latency
      - 1136.79 millisecond => 10985.49 millisecond

Benchmarking a single component cannot reflect the tradeoff between model accuracy and QoS

Statistical Model + Component Benchmarks?

- Whether a statistical model can predict the overall system tail latency, through profiling many components’ tail latency performance?
  - **NO!**

- Simple queueing model
  - **E-commerce Search Intelligence Scenario**
    - 8.6X between the actual average latency and the theoretical one
    - 3.3X between the actual 99th percentile latency and the theoretical one

- Sophisticated queueing network model
  - **E-commerce Search Intelligence Scenario**
    - **4.9X** between the actual average latency and the theoretical one
    - **Difficult** for tail latency predicting: non-superposition property

---

Why the Statistical Model Failed?

The M/M/1 model is far away from the realistic situation

- Industry-scale application scenario is complex and uncertain
  - the service rate doesn’t follow the exponential distribution.
- The more generalized model (such as G/G/1 model) is difficult to be used to calculate the tail latency.

The queuing network model is much infeasible

- The 99th percentile latency is hard to be predicted accurately
  - the non-superposition property of the overall system tail latency: being not equal to the sum of each component
Scenario Benchmark is needed!

- A proxy of a realistic application scenario
  - The real one is treated as first-class confidential issues

- Capturing the critical path and primary modules
  - The **permutations** of a series of AI and non-AI components
Content

- AIBench Scenario Methodology

- AIBench Scenario Benchmarks

- Experimental Results

- Conclusion
Why Scenario Benchmarks?

Reflect the overall system’s effects
- Critical Path
- Primary Modules

Measure the tradeoffs
- Model quality vs. QoS

Different level performance data
- Only component benchmarking may lead to error-prone conclusions
Evaluation Simulator

Scenario Benchmarks: overall system performance
AI Component Benchmarks: model performance & quality
AI Micro Benchmarks: code design and optimization

Formalization

Real-world Application Scenarios

DAG-based Tasks Modelling

Distill

Reusing Framework

Implementation

AI Component Benchmark
Models & Datasets
Metrics

Proxy

Scenario Benchmark Specification
Permutation of essential AI and non-AI tasks

Profiling

AI Component Benchmarks

Reference Implementation

Scenario Benchmark Specification

Scenario Benchmarks

Feedback & Optimization
(System/architecture/algorithm Innovations)

Physical Machine

Sampling traces

Evaluation

Distill

Scenario Benchmark Specification

Scenario Benchmarks

AI Component Specification

AI Micro Benchmarks

Real-world Application Scenarios

DAG-based Tasks Modelling

Distill

Reusing Framework

Implementation

AI Component Benchmark
Models & Datasets
Metrics

Proxy

Scenario Benchmark Specification
Permutation of essential AI and non-AI tasks

Profiling

AI Component Benchmarks

Reference Implementation

Scenario Benchmarks: overall system performance
AI Component Benchmarks: model performance & quality
AI Micro Benchmarks: code design and optimization

Feedback & Optimization
(System/architecture/algorithm Innovations)
Nine Typical Scenarios from 17 Industry Partners

- Search engine, e-commerce, social network, news feed, video and etc.
  - Primary AI and non-AI tasks

<table>
<thead>
<tr>
<th>Company</th>
<th>AI Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alibaba</td>
<td>E-commerce Search Intelligence, Online Translation Intelligence</td>
</tr>
<tr>
<td>Tencent</td>
<td>Content-based Image Retrieval, Web Searching</td>
</tr>
<tr>
<td>Baidu</td>
<td>Facial Authentication and Payment, News Feed</td>
</tr>
<tr>
<td>Netease</td>
<td></td>
</tr>
<tr>
<td>JD</td>
<td></td>
</tr>
<tr>
<td>TikTok</td>
<td>Live Streaming, Video Services, Online Gaming</td>
</tr>
<tr>
<td>PayPal</td>
<td></td>
</tr>
</tbody>
</table>
E-commerce Search Intelligence

- **AI tasks**
  - Classification, learning to rank, recommendation
- **Non-AI tasks**
  - Query parsing, database operation, indexing
- **Data**
  - User data, product data, query data
- **Metrics**
  - Precision, recall, latency
- **Model update frequency**
  - High

*Figure source: Amazon website*
Online Translation Intelligence

- AI tasks
  - Text-to-Text translation, speech recognition
- Non-AI tasks
  - Query parsing
- Data
  - Text, speech
- Metrics
  - Accuracy, latency
- Model update frequency
  - Low

Figure source: Baidu Translator
Content-based Image Retrieval

- **AI tasks**
  - Object detection, classification, spatial transformer, image-to-text
- **Non-AI tasks**
  - Query parsing, indexing, sort
- **Data**
  - Image
- **Metrics**
  - Precision, recall, latency
- **Model update frequency**
  - High

*Figure source: Alibaba AliTech101_Algorithms*
Web Searching

- **AI tasks**
  - Text summarization, learning to rank, recommendation

- **Non-AI tasks**
  - Query parsing, indexing, crawler, sort, hash

- **Data**
  - Product data, query data

- **Metrics**
  - Precision, recall, latency

- **Model update frequency**
  - High

*Figure source: baidu.com*
Facial Authentication and Payment

- **AI tasks**
  - Face embedding, 3D face recognition

- **Non-AI tasks**
  - Encryption

- **Data**
  - Face image

- **Metrics**
  - Accuracy, latency

- **Model update frequency**
  - Low

*Figure source: [https://www.sohu.com/a/281752161_670911](https://www.sohu.com/a/281752161_670911)*
**News Feed**

- **AI tasks**
  - Recommendation
- **Non-AI tasks**
  - Database operation, sort, basic statistics, filter
- **Data**
  - Text
- **Metrics**
  - Precision, recall
- **Model update frequency**
  - Low

*Figure source: baidu.com*
Live Streaming

- AI tasks
  - Image generation, image-to-image
- Non-AI tasks
  - Video codec, video capture
- Data
  - Image
- Metrics
  - Latency
- Model update frequency
  - Low

Figure source: https://www.tech-recipes.com/rx/73926/how-to-stream-live-on-tiktok-a-step-by-step-guide/
Video Services

- AI tasks
  - Image compression, video prediction
- Non-AI tasks
  - Video codec
- Data
  - Video
- Metrics
  - Accuracy, latency
- Model update frequency
  - Low

Figure source: https://tech.qq.com/a/20170414/023987.htm
Online Gaming

- **AI tasks**
  - 3D object reconstruction, image generation, image-to-image

- **Non-AI tasks**
  - Rendering

- **Data**
  - Image

- **Metrics**
  - Latency

- **Model update frequency**
  - Low

Figure source: https://www.hindustantimes.com/health/how-youngsters-get-addicted-to-gaming-and-lose-control-of-lives/story-HRHdLxPmfB4GAOmIgrsymM.html
Distilling rules

- R1: Pruning DAG branches with similar processing logic
- R2: Pruning the DAG branches with smallest fraction
- R3: Pruning auxiliary functions
  - E.g., Monitoring system
- R4: Using a state-of-the-art AI model
- R5: Combining the components occurred in successive steps
- R6: Removing unnecessary components after performing the above five rules
Distilling Example: E-commerce Search

R1: Similar processing logic
Distilling Example: E-commerce Search

R2: Smallest fraction
Distilling Example: E-commerce Search

R3: Auxiliary functions

OLAP Analyzer
- User Behavior Analyzer
- Daily Transaction Analyzer
- Log Analyzer

Indexer
- Product Index
- Item-to-Item Index
- User-to-Item Index
- User-to-Category Index

R3: Auxiliary functions
R4: Provide one state-of-the-art model
Distilling Example: E-commerce Search

R5: Combine classifier

OLAP Analyzer

Job Scheduler

AI Offline Trainer

Search Planner

Searcher

Indexer

Offline Analyzer

Online Server

Supporting

Offline Analyzer

Online Server

Supporting
Distilling Example: E-commerce Search

Recommender
- Query Preprocessing
  - Object Detection
  - Text Classification
  - Audio Classification
- Entity Identification
- Personalized Recommendation
  - By user profile
  - By user behavior
- Product Recommendation
  - By popularity
  - By price
  - By credibility
  - By purchase rate
- Advertising (AD)
  - By click-through rate

Online Server
- Monitor Module
- Tuning Module
- Logging Module
- Ops Management

Job Scheduler
- Batch Processing
- Streaming-like

AI Offline Trainer
- Image-to-Text
- Image classification
- Face embedding
- Speech recognition
- Image-to-image
- Spatial transformer
- Learning to rank
- Image generation
- Object detection
- Recommendation

OLAP Analyzer
- User Behavior Analyzer
- Daily Transaction Analyzer
- Log Analyzer

Search Planner
- Query Preprocessing
  - Image/Text
  - Category classifier
  - Image Category Classification
  - Text Category Classification
  - Personalized Recommendation
    - By user profile
    - By user behavior

Indexer
- Product Index
- Item-to-index
- Popular product index
- Product Attribute Index
- User Index
- User-to-item Index
- User-to-category Index
- Multiple AI Models (LR, GBDT, WDL, DNN, CNN, RNN, MAB, RL)

Offline Analyzer
- User Info
- Product Info

Monitor Module
- Tuning Module
- Logging Module
- Ops Management

Supporting

Bench Council

ICT
Institute of Computing Technology, Chinese Academy of Sciences
**The First** Reusing Framework for easily constructing scenario benchmarks

- A highly extensible, configurable, and flexible benchmark framework
- AI-related and non AI-related Library
- Support critical paths and primary modules modelling
- Multiple loosely coupled modules
  - Individually
    - Micro/Component benchmarks
  - Collectively
    - Scenario benchmarks
Guidelines to Construct A Scenario Benchmark

- Choose the essential AI and non-AI component benchmarks from the reusing framework

- Specify the permutation of component benchmarks

- Modify the reusing framework’s deployment template
  - Cluster configurations, execution parameters, etc

- Train the AI models and transfer the trained models to the online inference module
Content

- AIBench Scenario Methodology
- AIBench Scenario Benchmarks
- Experimental Results
- Conclusion
Query generator
- simulate concurrent users and send query requests

Online module
- personalized searching and recommendations

Offline module
- a training stage to generate a learning model

Data storage module
- data storage, e.g., user database, product database
Scenario Benchmark: Online Translation Intelligence

Online Translation Intelligence Implementation

- **Image Preprocessing**
  - Image query

- **Search Planner**
  - Text query

- **Audio Preprocessing**
  - Audio query

- **Online Server**
  - Text query
  - Text input

  - **Text Translator**
    - Text output
    - Text-to-Text Translation

  - **Image-to-Text**
    - OCR

  - **Speech Recognition**
    - DeepSpeech2

- **Data Storage**

- **Job Scheduler**
  - Batch Processing
  - Streaming-like

- **AI Offline Trainer**
  - Text-to-Text Translation
  - Speech Recognition
  - Image-to-Text

- **Offline Analyzer**

**AI Bench Framework**

**16 AI Component Benchmarks for representative AI Tasks**

- Speech recognition
- Text summarization
- Object detection
- Image generation
- 3D object reconstruction
- Video prediction
- Spatial transformer
- Recommendation
- 3D face recognition
- AI Units of Computation
- Text-to-Text translation
- Learning to rank
- Face embedding
- Image compression
- Image-to-Image
- Image-to-Text
Validation of a Scenario Benchmark

- Specification-level Validation
  - Much easier to validate whether the essential parts of a scenario are captured
    - reach a consensus with the industry partners’ feedback

- Implementation-level Validation
  - Compare the scenario benchmark with the real-world scenario
    - E.g., compare the overall system latency of the whole execution path under the same deployment.
Content

- AIBench Scenario Methodology
- AIBench Scenario Benchmarks
- Experimental Results
- Conclusion
Deployment

- E-commerce Search Intelligence Scenario Benchmark
  - 15-node CPU clusters

- Online Translation Intelligence Scenario Benchmark
  - 5-node CPU clusters
The overall system tail latency deteriorates dozens of times or even hundreds of times concerning a single component.

- E.g., E-commerce Intelligence
  - 2.2X comparing to recommendation component
  - 180X comparing to text classification component

(1) E-commerce Intelligence.

(2) Translation Intelligence.
AI vs. non-AI components

- E-commerce Intelligence
  - the average time spent on the AI and non-AI components is 137.12 and 58.16 milliseconds

- Translation Intelligence
  - nearly 99% execution time is spent on the AI components
Reproducibility of Scenario Benchmark

- Reproducibility is of great significance for a benchmark
  - We repeat the evaluation for five times

- Metric: the coefficient of variation (CV)
  - the standard deviation ratio to the mean value
    - For E-commerce Intelligence, the CV for the average, 90th percentile, and 99th percentile latency is 0.005, 0.007, and 0.014, respectively (Good reproducibility)
    - For Translation Intelligence, the values are 0.006, 0.018, and 0.024, respectively (Good reproducibility)
E-commerce Intelligence scenario benchmark vs. the real-world E-commerce scenario from our industry partner

**Overall system perspective:** *captures the overall system performance*

**Single component perspective:** *reflects its importance on the critical path and reflects similar workload behaviors*

- The most time-consuming component: Recommendation component
  - The average latency deviation is 6%
  - The CPU utilization and IPC deviations are 2% and 1.7%, respectively
Benchmarking Offline Training

- Why offline training matters for a scenario benchmark?
  - As witnessed by our many industry partners, when an AI model is used for online service, it has to be *updated in a real time manner*.
    - E.g., one E-commerce giant demands that “the updated model every one hour will bring in the award about 3% click-through rate and millions of profits.”

- There is a tradeoff among model update interval, accuracy improvement, and training overhead using offline training
We resize the input data volume to investigate how to balance the tradeoffs

- For Image Classifier
  - Resizing the volume from 60% to 100%, 51% additional training time → 4.96% accuracy improvement
  - From 80% to 100%, 35% additional training time → 1.9% accuracy improvement

- For Ranker
  - From 60% to 100%, 70% additional training time → 1.36% accuracy improvement
  - From 80% to 100%, 17% extra training time → 0.12% accuracy improvement

For different AI components, the tradeoffs have subtle differences.
Content

- AIBench Scenario Methodology
- AIBench Scenario Benchmarks
- Experimental Results
- Conclusion
Conclusion

- AIBench Scenario is needed for overall system performance
  - Scenario-distilling benchmarking methodology
    - Proxies to industry-scale application scenarios
  - A reusing framework
  - Two scenario benchmarks
    - E-commerce Search Intelligence
    - Online Translation Intelligence
References and Resources

  - *AIBench Scenario: Scenario-distilling AI Benchmarking*

- Download AIBench Scenario (Sign in/up BenchHub to get access !)
  - AIBench Scenario Benchmark
    - [http://www.benchcouncil.org/benchhub/AIBench/AIBench_Application_Benchmark/](http://www.benchcouncil.org/benchhub/AIBench/AIBench_Application_Benchmark/)
    - [http://www.benchcouncil.org/benchhub/AIBench/AIBench_DCMIX](http://www.benchcouncil.org/benchhub/AIBench/AIBench_DCMIX)
  - AIBench Framework
Thank you!