<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Speaker</th>
<th>Time</th>
<th>Topic</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00-9:40</td>
<td>AIBench Overview</td>
<td>Jianfeng Zhan</td>
<td>13:00-13:15</td>
<td>Hands-on demos on how to use AIBench Scenario</td>
<td>Fei Tang</td>
</tr>
<tr>
<td>9:40-10:00</td>
<td>AIBench Scenario: Scenario-distilling AI Benchmarking</td>
<td>Tianshu Hao</td>
<td>13:15-13:30</td>
<td>Hands-on demos on how to use Edge AIBench</td>
<td>Chuanxin Lan</td>
</tr>
<tr>
<td>10:00-10:20</td>
<td>Edge AIBench: towards Comprehensive End-to-end Edge Computing Benchmarking</td>
<td>Tianshu Hao</td>
<td>13:30-13:45</td>
<td>Hands-on demos on how to use AIBench Training</td>
<td>Chuanxin Lan</td>
</tr>
<tr>
<td>10:20-10:40</td>
<td>AIBench training, subsets, and its rankings</td>
<td>Fei Tang</td>
<td>13:45-14:00</td>
<td>Hands-on Demos on How to Use HPC AI500</td>
<td>Zihan Jiang</td>
</tr>
<tr>
<td>10:40-11:00</td>
<td>HPC AI500: A Benchmark Suite and HPC AI Ranking for HPC AI Systems</td>
<td>Zihan Jiang</td>
<td>14:00-14:10</td>
<td>Hands-on Demos on How to Use AIBench Micro</td>
<td>Fei Tang</td>
</tr>
<tr>
<td>11:00-11:20</td>
<td>AIBench Micro and Synthetic Benchmarks</td>
<td>Fei Tang</td>
<td>14:10-14:25</td>
<td>Hands-on Demos on How to Use AloTBench</td>
<td>Chunjie Luo</td>
</tr>
<tr>
<td>11:20-11:40</td>
<td>AloTBench for Benchmarking Mobile and Embedded Device Intelligence and Its Rankings</td>
<td>Chunjie Luo</td>
<td>14:25-14:35</td>
<td>Hands-on Demos on How to Use AIBench Inference</td>
<td>Chuanxin Lan</td>
</tr>
<tr>
<td>11:40-12:00</td>
<td>AIBench Inference Benchmarks and Whole-picture Workload Characterization (WPC)</td>
<td>Chuanxin Lan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00-13:00</td>
<td>Lunch Break</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
AIBench: AI Benchmarking Challenges, Methodology and Progress

Prof. Dr. Jianfeng Zhan
zhanjianfeng@ict.ac.cn or jianfengzhan.benchcouncil@gmail.com
https://www.benchcouncil.org
ICT, Chinese Academy of Sciences & BenchCouncil
AIBench Tutorial @ ISCA 2021
Learning dynamics are not well understood

- High dimension non-convex optimization problem
  - A slight change leads to a different optimization path
  - Heavily dependent on the experience for parameter tuning
AI Benchmarking Challenges

- Prohibitive cost
- Conflicting requirements in different stages
- Short shelf-life
- Scalability
- Repeatability

Prohibitive Cost Challenge

- Running an entire training session is mandatory!
  - Some optimizations improve the throughput but finally hurt the model quality.

- Take several weeks to run a complete training session on a small-scale system
  - Simulators with slowdowns 10 to 1,000 times exacerbate the challenge

- A microbenchmark like HPL-AI cannot model the learning dynamics of deep learning

[1] HPL-AI Mixed-Precision Benchmark — HPL-AI 0.0.2 documentation. https://icl.bitbucket.io/hpl-ai/
AI Benchmarking Challenges

- Prohibitive cost
- Conflicting requirements in different stages
- Short shelf-life
- Scalability
- Repeatability

[1] AIBench Training: Balanced Industry-Standard AI Training Benchmarking.[PDF]
Conflicting-requirement Challenge

- Overall
  - component benchmarks can not run on simulators.
  - Micro benchmarks are affordable but can not model learning dynamic.
- Earlier-stage evaluations of a new architecture or system
  - Affordable
  - Portability (Micro benchmarks)
  - Simplicity
- Later-stage evaluations or purchasing off-the-shelf systems
  - Comprehensiveness and representativeness
  - Overall system performance in reality
AI Benchmarking Challenges

- Prohibitive cost
- Conflicting requirements
- Short shelf-life
- Scalability
- Repeatability

[1] AIBench Training: Balanced Industry-Standard AI Training Benchmarking.[PDF]
Short Shelf-life Challenge

- AI model evolutions and changes outpace the AI benchmarks.
  - One year to walk through benchmark design, implementation, community adoption, and large-scale testing.

- Synthetic benchmarks like ParaDNN [1] can traverse many networks, but cannot model learning dynamics.

AI Benchmarking Challenges

- Prohibitive cost
- Conflicting requirements
- Short shelf-life
- Scalability
- Repeatability

## Scalability Challenge

- An AI task’s problem scale is fixed, not scalable

<table>
<thead>
<tr>
<th>Rank</th>
<th>Source</th>
<th>Time(min)</th>
<th>AI Accelerators</th>
<th>Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fujitsu[1]</td>
<td>1.2</td>
<td>Tesla V100*2048</td>
<td>MXNeT</td>
</tr>
<tr>
<td>2</td>
<td>Google[2]</td>
<td>2.2</td>
<td>TPU V3*1024</td>
<td>TensorFlow</td>
</tr>
<tr>
<td>3</td>
<td>Sony[3]</td>
<td>3.7</td>
<td>Tesla V100*2176</td>
<td>NNL</td>
</tr>
</tbody>
</table>

*Picture from HPC AI500 Ranking, Image Classification*

HPL-AI [1] is scalable, but it cannot model the learning dynamics, and fail to consider the model quality.

[1] HPL-AI Mixed-Precision Benchmark — HPL-AI 0.0.2 documentation. [https://icl.bitbucket.io/hpl-ai/](https://icl.bitbucket.io/hpl-ai/)
AI Benchmarking Challenges

- Prohibitive cost
- Conflicting requirements
- Short shelf-life
- Scalability

- **Repeatability**

The benchmark mandates being repeatable, while training deep networks is stochastic.

Factors of randomness:
- Model initialization
- Data augment
- Data shuffle
- Dropout
- Etc.

Run-to-run Variation of AIBench Training
The epochs to achieve target quality vary significantly under different random seeds.

Example: Randomness

- Run Image-to-Text from AI-Bench Training four times using different random seeds.

Overview

- Challenges
- Related Work
- AIBench Methodology
- AIBench
  - AIBench Scenario
    - Edge AIBench
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion
AI Benchmarks

**Fathom**
- arXiv 2016
- IISWC 2016
- Eight workloads
- Training and inference
- No quality metric

**DeepBench, Baidu Github 2017**
- AI basic operators, containing gemm, convolution, recurrent layer and all reduce
- Only has micro benchmarks

**DAWNBench, NIPS 2017**
- Image classification and question answer
- Use time-to-accuracy as metric

**AIBench, Bench’18, IISWC’18, PACT’18, arXiv 2019, 2020, ISPASS’21**
- First proposing scenario benchmarks
- 19 tasks, 19 workloads, 3 subsets

**DNNMark**
- GitHub 2016
- GPGPU 2017
- Eight micro benchmarks

**BenchIP**
- arXiv 2017
- JCST 2018
- 10 microbenchmarks
- 11 neural network models

**TBD Suite**
- GitHub 2018
- IISWC 2018
- Eight workloads, six domains

**MLPerf, 2018**
- GitHub 2019
- SysML 2020
- Five domains seven workloads

**AIIA-DNN**
- GitHub 2019
- Designed to support training and inference, but only provides inference implements now
HPC AI Benchmarking

**HPC AI500, Bench’18, arXiv 2020**
*Bench’18:*
Cover 3 representative application of scientific deep learning.
*arXiv 2020:*
Hierarchical benchmarking methodology; 3 benchmarking levels and rules; Use Valid FLOPS as the metric; Two representative and repeatable AI workloads (Business + Scientific).

**HPL-AI, 2019**
Micro benchmark based on LU decomposition; Scalable but can not reflect model quality.

**MLPerf, SysML 2020**
7 workloads covering 5 domains; 2 benchmarking levels and rules; Use time-to-train as the metric.

**SciML, GitHub 2021**
Benchmarking AI for Science domain, including material, life, and earth sciences, particle physics and astronomy.

**Deep500, IPDPS 2019**
A framework covering 4 level benchmarking; No concrete reference implementation.
AI Benchmarks for Edge Computing

**Edge AI Bench, Bench 2018, arXiv 2019**
- Scenario benchmarking
- ICU patient monitoring, camera monitoring, smart home, and automatic driving
- Integrated federal learning

**EdgeBench, UCC Companion 2018**
- Speech recognition, and image classification

**EEMBC MLMark, 2019**
- Image classification, object detection, translation, and speech recognition
- Closed source
AI Benchmarks for IoT

AIOTBench, Bench 2018
- Vision, audio, and NLP domain
- Supports Android and Raspberry Pie
- TensorFlow Lite, Caffe 2
- End-to-end, microbenchmarks

ETH Zurich AI Benchmark, ECCV 2018
- Only supports vision domain
- Only supports Android and TensorFlow Lite
- End-to-end
Summary of Related Work

- DawnBench (2017): the first benchmark that proposes time-to-accuracy as the primary metric.

- AIBench (2018): the first benchmark modeling the critical paths of a real-world application scenario.

- ParaDNN (2020) is the first synthetic AI benchmark.

- HPL-AI (2019) is a micro benchmark that uses mixed-precision LU decomposition to achieve upper bound FLOPS performance.
  - scalable, but LU decomposition not relevant to most of the AI workloads.

- Both ParaDNN and HPL-AI can not model the learning dynamics, and also fail to consider the model quality.
AIBench and MLPerf are two systematic AI benchmarking projects. They are concurrent and complemental.

The AIBench suites are by far the most comprehensive AI benchmark suites tackling Challenges #1-5.
- prohibitive cost
- conflicting requirements in different stages
- short shelf-life and fast evolution of AI models
- scalability challenge due to the fixed problem scale
- repeatability challenge due to the stochastic nature of AI.

MLPerf (2019) includes seven benchmarks for training and five benchmarks for inference.
MLPerf performs the most large-scale testing, but fails to present a benchmarking methodology to justify the choice for and update AI tasks, models, and data sets.
MLPerf fails to consider the conflicting requirements, shelf-life, scalability challenges (Challenges #2-4).
Overview

- AI Benchmarking Challenges
- Related Work
- **AIBench Methodology**
- **AIBench**
  - AIBench Scenario
    - Edge AIBench
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic
- Conclusion
International Open Benchmark Council (BenchCouncil)
- **https://www.benchcouncil.org**
- a non-profit international organization
  - aims to promote standardizing and incubating Big Data, AI, Chip and other emerging technology.

BenchCouncil AIBench
- **https://www.benchcouncil.org/aibench**
<table>
<thead>
<tr>
<th>Scenario benchmark</th>
<th>AIBench Scenario</th>
<th>Edge AIBench</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>AIBench Training</td>
<td>AIBench Subset</td>
</tr>
<tr>
<td>Inference</td>
<td>AIBench Inference</td>
<td>AIoTBench</td>
</tr>
<tr>
<td>Micro</td>
<td>Conv</td>
<td>FC</td>
</tr>
<tr>
<td>Synthetic</td>
<td>AIBench Synthetic</td>
<td></td>
</tr>
</tbody>
</table>
As a joint work with increasing industry partners, AIBench is a comprehensive AI benchmark project focusing on methodology, frameworks, and continuous improvements.

- Tackle the challenges #1-5 mentioned above in a systematic manner.
AIBench Scenario models the critical paths of a real-world application scenario as a permutation of the AI and non-AI modules.

Edge AIBench is an instance of the scenario benchmark suites, modeling end-to-end performance across IoT, edge, and Datacenter.
AIBench Training and AIBench Inference cover nineteen representative AI tasks (will update) with state-of-the-art models to guarantee diversity and representativeness.
AIBench Training provides two subsets for repeatable performance ranking (RPR) subset and workload characterization (WC) subset to improve affordability.
To evaluate large-scale HPC AI systems, HPC AI500 is derived from the AIBench Training RPR subset.

To evaluate various IoT and embedded devices, AIoTBench is derived from AIBench Inference.
- AIBench Micro provides the intensively-used hotspot functions, profiled from AIBench Training and Inference, for simulation-based architecture researches.
As complementary to real-world benchmarks, AIBench Synthetic provides several synthetic benchmarks with scalable problem sizes that can model learning dynamics.
Overview

- AI Benchmarking Challenges
- Related Work
- AIBench Methodology

**AIBench**
- **AIBench Scenario**
  - Edge AIBench
- AIBench Training
  - AIBench Subset
  - HPC AI500
- AIBench Inference
  - AIoTBench
- Micro Benchmarks
- AIBench Synthetic

- Conclusion
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.
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 (A scenario benchmark from AIBench)

OCR Model Pruning

Tail latency shortens by 51%

Accuracy decreases 1%

OCR Model Pruning

Optimization Technique

Inconsistent conclusion

Tail latency has no improvement
Accuracy decreases 1%

Only a single component may need to error-prone conclusions

Single Component vs. Realistic Application

- *E-commerce Search Intelligence* (*A scenario benchmark from AIBench*)

- 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 with a single component cannot reflect the overall system’s effects
For E-commerce Search Intelligence (*a scenario benchmark from AIBench*)

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

Benchmarking with a single component cannot reflect the tradeoff between model accuracy and QoS
Does a statistical model predict the overall system tail latency through profiling many components’ tail latency performance?

- NO!

A simple queueing model
- E-commerce Search Intelligence (a scenario benchmark from AIBench)
  - 8.6X between the actual average latency and the theoretical one
  - 3.3X between the actual 99th percentile latency and the theoretical one

A sophisticated queueing network model
- E-commerce Search Intelligence (a scenario benchmark from AIBench)
  - 4.9X between the actual average latency and the theoretical one
  - Difficult for tail latency predicting: non-superposition property
AIBench Scenario 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
Scenario Benchmark: E-commerce Search Intelligence

- **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

Search Planner
- Text query
- Image query
- Audio query

Text Translator
- Text input
- Text output
- Text-to-Text Translation

Speech Recognition
- DeepSpeech2
- Audio query

Image-to-Text
- OCR
- Image query

Data Storage

Job Scheduler
- Batch Processing
- Streaming-like

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

Offline Analyzer

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
- Classification
- Face embedding
- Image compression
- Text-to-Text translation
- Learning to rank
- Image-to-Image
- Image-to-Text

AI Bench Framework

AI-as-a-Service

AI for training
AI Benchmarking Challenges
Related Work
AIBench Methodology

**AIBench**
- AIBench Scenario
  - Edge AIBench
- AIBench Training
  - AIBench Subset
  - HPC AI500
- AIBench Inference
  - AIoTBench
- Micro Benchmarks
- AIBench Synthetic

Conclusion
An instance of the scenario benchmark suites

- modeling end-to-end performance across IoT, edge, and Datacenter

**Publication**

- *Edge AIBench: towards comprehensive end-to-end edge computing benchmarking. Bench’18*
- *Edge AIBench Specification*
  - [https://www.benchcouncil.org/file/EdgeAIBench_Specification.pdf](https://www.benchcouncil.org/file/EdgeAIBench_Specification.pdf)
Four Typical Edge AI Scenarios

- (1) Autonomous Vehicle
  - High-accuracy, Latency-sensitive
  - Device Mobility
- (2) ICU Patient Monitor
  - Latency-sensitive (msec level)
  - May tolerate some errors
  - Parallel, massive patients
- (3) Surveillance Camera
  - Enormous Data
- (4) Smart Home
  - Heterogenous devices and data
### Nine Typical Edge AI Tasks

<table>
<thead>
<tr>
<th>Task Name</th>
<th>Edge AI Scenarios</th>
<th>Models</th>
<th>Datasets</th>
<th>Implementations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lane Detection</td>
<td>Autonomous Vehicle</td>
<td>LaneNet</td>
<td>Tusimple/CULane</td>
<td>Pytorch/Caffe</td>
</tr>
<tr>
<td>Traffic Sign Detection</td>
<td>Autonomous Vehicle</td>
<td>Capsule Network</td>
<td>German Traffic Sign Recognition Benchmark</td>
<td>Keras</td>
</tr>
<tr>
<td>Heart Failure Prediction</td>
<td>ICU Patient Monitor</td>
<td>LSTM</td>
<td>MIMIC-III</td>
<td>Tensorflow/Keras</td>
</tr>
<tr>
<td>Decompensation Prediction</td>
<td>ICU Patient Monitor</td>
<td>LSTM</td>
<td>MIMIC-III</td>
<td>Tensorflow/Keras</td>
</tr>
<tr>
<td>Death Prediction</td>
<td>ICU Patient Monitor</td>
<td>LSTM</td>
<td>MIMIC-III</td>
<td>Tensorflow/Keras</td>
</tr>
<tr>
<td>Person Re-identification</td>
<td>Surveillance Camera</td>
<td>DG-Net</td>
<td>Market-1501</td>
<td>Pytorch</td>
</tr>
<tr>
<td>Action Detection</td>
<td>Surveillance Camera</td>
<td>ResNet18</td>
<td>UCF101</td>
<td>Pytorch/Caffe</td>
</tr>
<tr>
<td>Face Recognition</td>
<td>Smart Home</td>
<td>Facenet/Sphere network</td>
<td>LibriSpeech/CASIA-Webface</td>
<td>Tensorflow/Caffe</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>Smart Home</td>
<td>DeepSpeech2</td>
<td>LibriSpeech</td>
<td>Tensorflow</td>
</tr>
</tbody>
</table>
Overview

- AI Benchmarking Challenges
- Related Work
- AIBench Methodology
  - **AIBench**
    - AIBench Scenario
      - Edge AIBench
    - **AIBench Training**
      - AIBench Subset
      - HPC AI500
    - AIBench Inference
      - AloTBench
    - Micro Benchmarks
    - AIBench Synthetic
- Conclusion
Consider Conflicting Benchmarking Requirements

- Benchmarking at different stages
  - Earlier-stage evaluations of a new architecture or system:
    - Portability (Micro benchmarks)
    - Simplicity
  - Later-stage evaluations or purchasing off-the-shelf systems:
    - Comprehensiveness/Representativeness
    - Performance in reality (Component or scenario benchmarks)
Diverse behaviors for workload characterization

Micro-architecture level
- FLOPs, memory access pattern, computation pattern, and I/O pattern

System level
- Time to quality, run-to-run variation (the ratio of the standard deviation to the mean of epochs to quality), convergence rate (epochs to quality), and number of hot functions

Algorithm level
- Model architectures and parameters
Coverage of diverse network architectures (CNN, ResNet, LSTM, GRU, Attention, etc.)

- **Text processing (7)**
  - Text-to-Text, Text summarization, Learning to Rank, Recommendation, Neural Architecture Search, Advertising, Nature Language Processing (NLP)

- **Image processing (8)**
  - Image Classification, Image Generation, Image-to-Text, Image-to-Image, Face Embedding, Object Detection, Image Compression, Spatial Transformer

- **Audio processing (1)**
  - Speech Recognition

- **Video processing (1)**
  - Video Prediction

- **3D data processing (2)**
  - 3D Face Recognition, 3D Object Reconstruction
Take E-commerce as an Example

<table>
<thead>
<tr>
<th>E-commerce</th>
<th>Search (text, image, and audio)</th>
<th>Advertising</th>
<th>Intelligent assistant</th>
<th>Products management</th>
<th>Virtual fitting</th>
<th>Facial authentication and payment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Image classification</td>
<td>Object detection</td>
<td>Spatial transformer</td>
<td>Translation</td>
<td>Recommendation</td>
<td>Learning-to-Rank</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speech recognition</td>
<td>Image compress</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Image-to-Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Text summarization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Image-to-Image</td>
<td></td>
<td></td>
<td></td>
<td>3D face recognition</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Face embedding</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Representativeness

<table>
<thead>
<tr>
<th>Methodology</th>
<th>AIBench Training v1.1</th>
<th>MLPerf Training v0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Balanced methodology considering conflicting requirements</td>
<td>According to commercial and research relevance</td>
</tr>
<tr>
<td>Task</td>
<td>Nineteen tasks and models</td>
<td>six tasks and eight models</td>
</tr>
<tr>
<td>Dataset</td>
<td>Text, image, 3D, audio, and video data</td>
<td>Text and image data</td>
</tr>
<tr>
<td><strong>Algorithm behavior</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computation</td>
<td>0.09 to 282830 MFLOPs</td>
<td>0.21 to 29000 MFLOPs</td>
</tr>
<tr>
<td>Complexity</td>
<td>0.03 to 110 million parameters</td>
<td>5.2 to 110 million parameters</td>
</tr>
<tr>
<td>Optimizer categories</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Loss function categories</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td><strong>System behavior</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hotspot functions</td>
<td>30</td>
<td>9</td>
</tr>
<tr>
<td>Convergence</td>
<td>6 to 96 epochs</td>
<td>3 to 49 epochs</td>
</tr>
<tr>
<td><strong>Micro-architecture behavior</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achieved occupancy</td>
<td>0.12 to 0.61</td>
<td>0.12 to 0.54</td>
</tr>
<tr>
<td>IPC efficiency</td>
<td>0.02 to 0.77</td>
<td>0.02 to 0.74</td>
</tr>
<tr>
<td>Gld efficiency</td>
<td>0.28 to 0.94</td>
<td>0.52 to 0.85</td>
</tr>
<tr>
<td>Gst efficiency</td>
<td>0.27 to 0.98</td>
<td>0.75 to 0.98</td>
</tr>
<tr>
<td>DRAM utilization</td>
<td>0.08 to 0.61</td>
<td>0.08 to 0.61</td>
</tr>
</tbody>
</table>
AI Benchmarking Challenges
Related Work
AIBench Methodology

**AIBench**
- AIBench Scenario
  - Edge AIBench
- AIBench Training
  - AIBench Subset
  - HPC AI500
- AIBench Inference
  - AIoTBench
- Micro Benchmarks
- AIBench Synthetic

Conclusion
Requirements for Ranking

- Performance Ranking
  - Fairness
  - Low cost
- Workload Characterization
  - Representativeness
  - Comprehensiveness
  - Diversity

Balanced by:
- Performance Ranking
- Workload Characterization
Two Subsets for Affordability

- Subset for repeatable performance ranking (RPR subset)
  - Image Classification, Object Detection, and Learning to Rank

- Subset for workload characterization (WC subset)
  - Spatial Transformer, Image-to-Text, and Speech-to-Text
    - the nearest to the centroid of three clusters, respectively
Overview

- AI Benchmarking Challenges
- Related Work
- AIBench Methodology

**AIBench**
- AIBench Scenario
  - Edge AIBench
- AIBench Training
  - AIBench Subset
  - HPC AI500
- AIBench Inference
  - AIoTBench
- Micro Benchmarks
- AIBench Synthetic

- Conclusion
HPC AI500 Benchmarking Methodology

- The criteria for choosing the workloads.
  - Representativeness and Affordability
  - Repeatability
  - Computation complexity
  - Tasks, Models, and datasets
  - Scalability

- AIBench RPR subset satisfies repeatability, representativeness, and affordability.
Scalability Requirement

- AIBench subset computation comparison (Single training batch).

<table>
<thead>
<tr>
<th>Workloads</th>
<th>Computation (FLOPs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Classification</td>
<td>23 G</td>
</tr>
<tr>
<td>Object Detection</td>
<td>691 G</td>
</tr>
<tr>
<td>Learning to Rank</td>
<td>0.08 M</td>
</tr>
</tbody>
</table>

Image Classification and Object Detection meet the computation requirement and are chosen as two typical workloads for HPC AI benchmarking.
Problem Domain, Dataset, and Model of HPC AI500

- **Problem Domain**
  - Extreme Weather Analysis: detect the patterns of extreme weather, essentially Object Detection. The application that wins Gordon Bell Prize.
  - Image Classification: ResNet50/ImageNet is a de facto benchmark for optimizing HPC AI systems.

- **Dataset**
  - The extreme weather dataset: 16 channels, 768*1052, 2 TB
  - ImageNet 2012: 3 channels, 256*256, 136 GB

- **Model**
  - Faster-RCNN
  - ResNet-50 V1.5
Overview

- AI Benchmarking Challenges
- Related Work
- AIBench Methodology

**AIBench**

- AIBench Scenario
  - Edge AIBench
- AIBench Training
  - AIBench Subset
  - HPC AI500
- **AIBench Inference**
  - AIoTBench
- Micro Benchmarks
- AIBench Synthetic

- Conclusion
Why Comprehensive Inference Workloads?

- Comprehensive workloads are not a burden!
  - Inference time is much shorter

- Diversity of data types, AI models, AI frameworks

- Diversity of workload behaviors
  - Algorithm, System, Architecture
Coverage of diverse network architectures (CNN, ResNet, LSTM, GRU, Attention, etc.)

- **Text processing (7)**
  - Text-to-Text, Text summarization, Learning to Rank, Recommendation, Neural Architecture Search, Advertising, Nature Language Processing (NLP)

- **Image processing (8)**
  - Image Classification, Image Generation, Image-to-Text, Image-to-Image, Face Embedding, Object Detection, Image Compression, Spatial Transformer

- **Audio processing (1)**
  - Speech Recognition

- **Video processing (1)**
  - Video Prediction

- **3D data processing (2)**
  - 3D Face Recognition, 3D Object Reconstruction
Overview

- AI Benchmarking Challenges
- Related Work
- AIBench Methodology
  - **AIBench**
    - AIBench Scenario
      - Edge AIBench
    - AIBench Training
      - AIBench Subset
      - HPC AI500
    - AIBench Inference
      - AIoTBench
    - Micro Benchmarks
    - AIBench Synthetic
- Conclusion
AIoTBench aims to evaluate the AI models, frameworks, and hardware on mobile and embedded environments.

- Out-of-box benchmarks, more affordable
- Make comprehensive and apple-to-apple comparisons of the algorithm, software, and hardware.

AIoTBench covers representative and diverse models and frameworks. It has 60 off-the-shelf workload instances in total.

- Models
  - ResNet50, InceptionV3, DenseNet121, SqueezeNet MobileNetV2, MnasNet
- Frameworks
  - Tensorflow Lite, Caffe2, Pytorch Mobile
- For each model in Tensorflow Lite
  - three quantization versions: dynamic range quantization, full integer quantization, float16 quantization
  - CPU and NNAPI delegate
Different operators have different proportions of processing time in different models.

The convolutions have diverse distributions of processing time in different models.

The time breakdown of operators for different models.

The distribution of convolution processing time. Horizontal axis refers to time (millisecond), vertical axis refers to the frequency.
The same model has different performance on different framework.
Overview

- AI Benchmarking Challenges
- Related Work
- AIBench Methodology
  - **AIBench**
    - AIBench Scenario
      - Edge AIBench
    - AIBench Training
      - AIBench Subset
      - HPC AI500
    - AIBench Inference
      - AIoTBench
  - **Micro Benchmarks**
  - **AIBench Synthetic**
- Conclusion
Micro Benchmarks

- Tackle the prohibitive cost challenge
- Easily portable across different architectures
- Primitive AI operators and widely-used hotspot functions from AIBench Training and AIBench Inference
## Micro Benchmarks

<table>
<thead>
<tr>
<th>No.</th>
<th>Micro Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC-AI-M1</td>
<td>Convolution</td>
</tr>
<tr>
<td>DC-AI-M2</td>
<td>Fully Connected</td>
</tr>
<tr>
<td>DC-AI-M3</td>
<td>Relu</td>
</tr>
<tr>
<td>DC-AI-M4</td>
<td>Sigmoid</td>
</tr>
<tr>
<td>DC-AI-M5</td>
<td>Tanh</td>
</tr>
<tr>
<td>DC-AI-M6</td>
<td>MaxPooling</td>
</tr>
<tr>
<td>DC-AI-M7</td>
<td>AvgPooling</td>
</tr>
<tr>
<td>DC-AI-M8</td>
<td>CosineNorm</td>
</tr>
<tr>
<td>DC-AI-M9</td>
<td>BatchNorm</td>
</tr>
<tr>
<td>DC-AI-M10</td>
<td>Dropout</td>
</tr>
<tr>
<td>DC-AI-M11</td>
<td>Element-wise multiply</td>
</tr>
<tr>
<td>DC-AI-M12</td>
<td>Softmax</td>
</tr>
</tbody>
</table>

- **GEMM**: DeepBench microbenchmark
- **Convolution**: AIBench microbenchmark
- **RNN**: All Reduce
- **All Reduce**
Overview

- AI Benchmarking Challenges
- Related Work
- AIBench Methodology
  - AIBench
    - AIBench Scenario
      - Edge AIBench
    - AIBench Training
      - AIBench Subset
      - HPC AI500
    - AIBench Inference
      - AIoTBench
    - Micro Benchmarks
    - AIBench Synthetic
- Conclusion
Tackle the short shelf-life and scalability challenges

Design Spaces
- Representative building blocks from AIBench Training
  - Fully connected layer, residual block, etc.
  - Depth, width, and input sizes
  - Stride and padding sizes

Train the generated models to convergence to reflect the learning dynamics, which ParaDNN[1] cannot reflect.

Overview

- AI Benchmarking Challenges
- Related Work
- AIBench Methodology
- **AIBench**
  - AIBench Scenario
    - Edge AIBench
  - AIBench Training
    - AIBench Subset
    - HPC AI500
  - AIBench Inference
    - AIoTBench
  - Micro Benchmarks
  - AIBench Synthetic

- **Conclusion**
AIBench distills and abstracts real-world application scenarios across Datacenter, HPC, IoT, and Edge into the scenario, training, inference, micro, and synthetic benchmarks.

http://www.benchcouncil.org/aibench
\section*{CFPs}

- **Bench’21**
  - Broad benchmarks conference (architecture, HPC, DB, machine learning)
  - \url{https://www.benchcouncil.org/bench21/}
    - Abstracts: July 16, 2021
    - Full Papers: July 30, 2021

- BenchCouncil Transactions on Benchmarks, Standards and Evaluations
  - \url{https://www.editorialmanager.com/tbench/default.aspx}
BenchCouncil Achievement Award
- This award recognizes a senior member who has made long-term contributions to benchmarks, data, standards, evaluations, and optimizations. The winner is eligible for BenchCouncil Fellow. ($3000)

BenchCouncil Rising Star Awards
- This award recognizes a young researcher who demonstrates outstanding research and practice in benchmarks, data, standards, evaluations, and optimizations. The winner is eligible for BenchCouncil Senior Member. ($1000)

BenchCouncil Distinguished Doctoral Dissertation Award
- This award recognizes and encourages superior research and writing by doctoral candidates in the broad field of benchmarks, data, standards, evaluations, and optimizations community. ($1000)
Thank you!