Why should you care?

Benchmarking vs. Benchmarketing

Why should you care?

Felix Baum
AI Software, Director of Product Management
Qualcomm Technologies, Inc.
September, 2020
AI will drive transformation across industries

Powering the factory of the future

Shaping the future of transportation
AI Performance Continuum

- **0.5-1+ TOPS**
  - Qualcomm QCS400
  - Qualcomm QCS603
  - Qualcomm QCS605

- **3-5 TOPS**
  - Snapdragon 427
  - Snapdragon 429
  - Snapdragon 439
  - Snapdragon 450

- **3rd Gen Snapdragon Automotive Cockpit Platforms**
  - Snapdragon 845
  - Snapdragon 855
  - Snapdragon 8cx
  - Snapdragon 765
  - Snapdragon 665
  - Snapdragon 7c

- **7-15 TOPS**
  - Snapdragon 730/G
  - Snapdragon 765
  - Snapdragon 665
  - Snapdragon 630/36

- **5-10 TOPS**
  - Snapdragon 8cx
  - Snapdragon 7c

- **15+ TOPS**
  - Snapdragon 865
AI Performance Continuum

Qualcomm Ride Platform
Automotive ADAS

Qualcomm Cloud AI 100
350 TOPS

2x Qualcomm Cloud AI 100
700 TOPS

0.5-1+ TOPS

>15 TOPS
Still FPS
~5 TOPs required
~4 Neural Nets

Photo Capture with AI
Segmentation
De-noise
Classification
Detection
60 FPS

>10 TOPs required

~8

Neural Nets

4K HDR video capture with AI

Segmentation  De-noise  Audio noise suppression  Audio echo cancellation
Classification  Detection  Enhancement  Semantic Filtering
XR

- 60 FPS
- 2 Displays
- >15 TOPs required
- ~8+ Neural Nets

**Features:**
- Head tracking
- Plane detection
- Audio noise suppression
- Context Awareness
- Hand tracking
- Classification
- Controllers
- Super Resolution
- Semantic Segmentation
- Facial expression
How to compare AI capabilities of hardware: apples vs. oranges

AI is changing fast and there are factors to consider when using AI benchmarks

Not all TOPS are created equal

Arbitrary
Not actual performance
Incentives vaporware

Intel's new AI chip performs
Operations per second
Intel introduced its new AI-embedded chipset, code-named Keem Bay – The burn-in

10 trillion

MediaTek Dimensity 1000 5G marks its comeback into the high-end chip market...
the new chip boasts of
Operations per second
– Gizmochina

4.5 TOPS (trillion)

MediaTek Dimensity 800 vs Snapdragon 765G: What’s the Difference?
This allows the processor to perform a total of approximately
Operations per second
– DealNTech

2.5 TOPS (trillion operations)
How to compare AI capabilities of hardware: apples vs. apples (slightly better)

AI is changing fast and there are factors to consider when using AI benchmarks

Public well known repeatable workflow → Inception v3 or other → Measure performance

Caveats

- Not indicative to the best performance that hardware can deliver
- Some models may be irrelevant to commercial use cases
- Hard to use this to compare hardware from various vendors (some vendors don’t support the model)
How to compare AI capabilities of hardware: better apples (fragmented)

AI is changing fast and there are factors to consider when using AI benchmarks

<table>
<thead>
<tr>
<th>Commercial AI Benchmarks Challenges</th>
<th>Considerations</th>
<th>What is the impact?</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI Benchmarks using legacy networks vs, state of the art</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New networks and operators continue to emerge so older AI benchmarks may not reflect these changes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New networks &amp; operators require new software to accelerate processing that AI benchmarks don’t support</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factors</th>
<th>Considerations</th>
<th>What is the impact?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Type</td>
<td>Integer vs Floating Point</td>
<td>Affects AI use case performance &amp; power</td>
</tr>
<tr>
<td>Runtime Framework</td>
<td>Vendor SDK vs NNAPI</td>
<td>OEMs mostly using Vendor SDKs</td>
</tr>
<tr>
<td>Score Weighting</td>
<td>Bias scoring toward a test case or category</td>
<td>Can be misleading AI score</td>
</tr>
</tbody>
</table>
AI benchmark Score Bias: (real example)

Benchmarks should be checked for fair weighting to reflect real-world usage (e.g., between networks, floating point versus integer data types, CPU/GPU/accelerator cores).

Scores composition must be transparent and using the latest networks & use cases (demonstrate no biases for chipsets, nor obsolete networks/use cases).

Be sure to look at multiple AI workloads/benchmarks.

How to check AI benchmark score weightage?

Benchmarks should be checked for fair weighting to reflect real-world usage (e.g., between networks, floating point versus integer data types, CPU/GPU/accelerator cores).

Scores composition must be transparent and using the latest networks & use cases (demonstrate no biases for chipsets, nor obsolete networks/use cases).

Be sure to look at multiple AI workloads/benchmarks.

Raw Data (Example)

Device A

A is 2x better than B in InceptionV3

Device B

B is only 3 inf/s better than A in VGG16

Who should win?

Inception lead is reduced to 1.5x

Overweighting a subtest biases benchmark to skew results

VGG16 weighted >50% of overall score

Device A

Device B

Inception

VGG16

other

other

VGG16 weighted >50% of overall score
AI benchmark Score Bias: (another example)

Benchmark over weighing FP16 results in a misleading AI performance score

Almost every FP16 test weights 2x than its INT8 counter part

Device A leads Device B on every category except FP16 (Device A leads on FP32)

Normalized score break down for each section

- CPU INT8
- CPU FP16
- INT8
- FP16
- FP32
- FP16 PARALLEL
- Combined Score
AI benchmarks overview: Almark

Almark has a variety of common AI models and use cases

Description:
Developed by Ludashi Holdings software company
- Ludashi offers variety of software products for PCs and smartphones

Almark considerations:
- Common models which most chipset vendors use for acceleration
- Using Qualcomm® Neural Processing SDK (Qualcomm NPE) allows concurrent use of Qualcomm’s Hexagon Vector eXtensions + Hexagon Tensor Accelerator that shows Qualcomm’s AI performance capability
- Most vendors and ISVs use each chipset vendor’s SDK, so Aimark results are more representative
- Uses INT8 models

Use Cases | Models
--- | ---
Classification | InceptionV3
Classification | Resnet34
Object Detection | MobilenetV1_SS
Segmentation | DeeplabV3

Access to ranking chart
Aimark benchmark ranking can be accessed within the app
By clicking “Benchmark Chart” benchmark ranking will be displayed
AI benchmarks overview: Altutu

Altutu benchmark focuses on Classification & Object Detection use cases

Description:
Developed by Antutu
- Antutu benchmark developer well known for Antutu system benchmark
- Altutu utilizes vendor’s SDK

Altutu considerations:
- Includes a breakdown of scores for accuracy
- Leverage both of Qualcomm’s Hexagon Vector eXtensions + Hexagon Tensor Accelerator
- Uses INT8 precision

<table>
<thead>
<tr>
<th>Use Cases</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>InceptionV3</td>
</tr>
<tr>
<td>Object Detection</td>
<td>MobilenetV2_SSD</td>
</tr>
</tbody>
</table>

Access to ranking chart

Altutu (Antutu AI Benchmark) app available at Google Play and Antutu Website
Aitutu benchmark rankings available in English and Chinese
Reference SDK based AI benchmark comparisons

Snapdragon 865 and Snapdragon 765 significantly lead on AI performance

- Reflective of actual users experience on Snapdragon 865 and 765 smartphones

AI benchmarks summary
AI benchmarks overview: ETH Zurich AI Benchmark

ETH Zurich AI Benchmark evaluates variety of use cases on the various processors

**Description:**
Developed by graduate student in Computer Vision Lab at ETH Zurich University
- First benchmark using TF lite
- Latest version v4 had just been released

**ETH AI Benchmark considerations:**
- Uses a range of AI models & use cases
- Includes a breakdown of scores
- 14 out of 43 tests utilize Hexagon Vector eXtension
  - Relevant to actual user’s experience on smartphone
- 29 out of 43 tests are FP16 or utilize only CPU
  - Helpful reference to developers who are evaluating the performance of not yet commercialized and unoptimized models during development

### ETH AI Benchmark v4

<table>
<thead>
<tr>
<th>Use Cases</th>
<th>Models</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>MobilenetV2</td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td>InceptionV3</td>
<td></td>
</tr>
<tr>
<td>Face Recognition</td>
<td>MobilenetV3</td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td>MobilenetV2</td>
<td></td>
</tr>
<tr>
<td>OCR</td>
<td>CRNN</td>
<td>FP only</td>
</tr>
<tr>
<td>Deblur</td>
<td>PyNet-Mini</td>
<td></td>
</tr>
<tr>
<td>Super Resolution</td>
<td>VGG19</td>
<td></td>
</tr>
<tr>
<td>Super Resolution</td>
<td>SRGAN</td>
<td></td>
</tr>
<tr>
<td>Bokeh</td>
<td>U-Net</td>
<td></td>
</tr>
<tr>
<td>Segmentation</td>
<td>DeeplabV3+</td>
<td></td>
</tr>
<tr>
<td>Enhancement</td>
<td>DPED-Resnet</td>
<td></td>
</tr>
<tr>
<td>Text Completion</td>
<td>LSTM</td>
<td>FP only</td>
</tr>
<tr>
<td>Memory Limits</td>
<td>SRCNN</td>
<td>Memory test</td>
</tr>
</tbody>
</table>
AI benchmarks overview: MLPerf

MLPerf benchmark

**Description:**
Open industry forum group establishing benchmark standard for ML hardware and software

- Represents the consensus of the industry due to large list of the participating companies (see https://mlperf.org/)
- New rounds of result submission and first version of the mobile AI benchmark app is expected to be available in Oct 2020

**MLPerf considerations:**
- Active participation and feedback from vendors such as Google, Intel, Nvidia, Qualcomm, etc.
- Benchmark implementation and decision process is transparent
There are good and bad benchmarks – Here is why

AI is changing fast and there are factors to consider when using AI benchmarks

<table>
<thead>
<tr>
<th>TOPS</th>
<th>Public well-known repeatable workflows</th>
<th>Well Known Benchmarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arbitrary</td>
<td>Not indicative to the best performance that hardware can deliver</td>
<td>Can be biased on certain types of models</td>
</tr>
<tr>
<td>Not actual performance</td>
<td>Some models may be irrelevant to commercial use cases</td>
<td>Tend to use legacy models</td>
</tr>
<tr>
<td>Incentives vaporware</td>
<td>Hard to use this to compare hardware from various vendors</td>
<td>Vendor results might not be accelerated</td>
</tr>
</tbody>
</table>

**Bottom Line:** There is no perfect solution to comparing AI hardware but certain benchmarks are more representative of actual performance
Thank you
Resource Slide

- Altutu benchmark ranking at: http://www.antutu.com/ranking/rank3.htm
- ML Perf benchmark at: https://mlperf.org/get-involved#join-the-forum

2020 Embedded Vision Summit

- “Advancing Embedded Vision for an Autonomous World”
  - Tuesday, September 15, 11:30 – 12:00 AM PT
- Qualcomm AI: Leading the Way with Distributed Intelligence
  - Thursday, September 24 11:00 – 11:30 AM PT