Kneron & AI SoC (KDP520 NPU)

Sep. 2018
Kneron, established in San Diego in 2015, is a leading provider of edge AI solutions. We focus on designing integrated software and hardware to enable edge side AI solutions for smart home, smart surveillance, smartphone, robot, drone, and IoT devices.
Founder and Core Team

Founder & CEO, Albert Liu
Albert Liu studied at UC Berkeley/UCLA/UCSD joint research master and doctoral degree, and got UCLA EE doctoral degree. After graduation, he worked in world class enterprises including Qualcomm, Samsung, MStar, and Wireless Info.

Team
World-class R&D team, who graduated from UCLA, MIT, Cornell, NTHU, NTU and so on, and worked in technology leading companies including Google, Microsoft, Bell Laboratories, IBM, and Intel. This team accumulated more than 10 years knowledge in AI, computer vision, and image processing.
## Academic Contribution

- **Patent:** Original AI algorithm, neural network software, hardware core design, applied ~20 patents in 4 countries
- **Excellent research team:** Published research result in multiple international top ranking conference, including ISSCC in which only 19 China papers are approved

<table>
<thead>
<tr>
<th>Published Papers</th>
<th>Publication</th>
<th>Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 16-Gb/s 14.7-mW Tri-Band Cognitive Serial Link Transmitter With Forwarded Clock to Enable PAM-16/256-QAM and Channel Response Detection</td>
<td>IEEE Journal of Solid State Circuits</td>
<td>1</td>
</tr>
<tr>
<td>A 2.3mW 11cm Range Bootstrapped and Correlated Double Sampling (BCDS) 3D Touch Sensor for Mobile Devices</td>
<td>IEEE International Solid-State Circuits Conference</td>
<td>2</td>
</tr>
<tr>
<td>A Reconfigurable Streaming Deep Convolutional Neural Network Accelerator for Internet of Things</td>
<td>IEEE Transactions on Circuits and Systems I</td>
<td>4</td>
</tr>
<tr>
<td>Invited — A 2.2 GHz SRAM with high temperature variation immunity for deep learning application under 28nm</td>
<td>IEEE/ACM Design Automation Conference, June, 2016</td>
<td>5</td>
</tr>
<tr>
<td>A Novel Fully Synthesizable All-Digital RF Transmitter for IoT Applications</td>
<td>IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems</td>
<td>10</td>
</tr>
</tbody>
</table>
Partner and Customer
Core team: top engineers from Qualcomm, Intel, Bell Labs and more

Albert Liu
PhD, Founder & CEO
UCLA EE PhD • IEEE senior member
Previously worked in Samsung on Galaxy
development, in Qualcomm on AI product, in Bell Lab
and NASA JPL; Numerous publications and patents
in machine learning.

Roger Liu
PhD, COO
Stony Brook University EE PhD
7 patents, 20 years experience, previously
worked in Sentelic Corp as VP of engineering, in
Milkywaysilicon Technology as COO / CTO

HsiangTsun Li
PhD, CSO
Stony Brook University EE PhD
70+ patents, 20+ years experience, previously
worked in Qualcomm as Director on MM system,
arquitectura y algoritmo design, in Spreadtrum,
HuaWei, Vivo as AVP of MM system engineering.
Edge AI
Benefits of edge AI

Consumers:
- Maintain privacy
- Response in real-time

Chip & device makers:
- Save bandwidth
- Increase stability
- Scalable business model
- Predictable long-term cost
NPU enables AI on edge devices
Make possible for power consumption, throughput, and cost

Source: National Science Foundation (NSF)
Core technology
Kneron AI Turkey Solution
One stop service aiming for fastest time to market

- Mainstream CNN networks
- 2D /3D Vision applications
- Audio applications

Commercialized CNN Model

Model Compression Tech
- Lossless / lossy compression
- Jointly work with hardware
- Reduce model size and computation cost

CNN Accelerator (NPU) IP
- Smart Phone  KDP3xx series
- Smart Home  KDP5xx series
- Surveillance KDP7xx series
AI SoC (KDP520 NPU)
AI SoC high level block

- KDP520 NPU (400MHz)
- Dual Cortex M4 (200MHz for sys & 250MHz for AI)
- 512KB SRAM
- 32MB / 64MB DRAM
- Process: 40nm
- Die size: 4000 x 4000
- Power: 200mW
<table>
<thead>
<tr>
<th>Specification</th>
<th>IP Version</th>
<th>KDP520</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-bit MAC</td>
<td></td>
<td>144</td>
</tr>
<tr>
<td>8-bit MAC</td>
<td></td>
<td>576</td>
</tr>
<tr>
<td>Arithmetic Precision</td>
<td></td>
<td>8/16-bit fixed-point</td>
</tr>
<tr>
<td>Support CNN layer kernel size</td>
<td></td>
<td>1 to 9</td>
</tr>
<tr>
<td>Support Channel/Feature Number</td>
<td></td>
<td>up to 1024</td>
</tr>
<tr>
<td>Support Max pooling/Average pooling Function with kernel size</td>
<td></td>
<td>2,3</td>
</tr>
<tr>
<td>SD &amp; Frame Buffer</td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td>Process Node</td>
<td></td>
<td>UMC 40nm LP</td>
</tr>
<tr>
<td>Maximum Clock Frequency (MHz)</td>
<td></td>
<td>400</td>
</tr>
<tr>
<td>NPU SRAM Size</td>
<td></td>
<td>512K</td>
</tr>
<tr>
<td>Operating Logic Power (mW)</td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Operating SRAM Power (mW)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Computing MAC Efficiency

- The computing MAC efficiency is evaluated as data bandwidth is high enough and SRAM size is large enough
- ResNet50 (~3x gain)
  - Ours: 73%
  - Others: 23.16%
- GoogLeNet (~1.7x gain)
  - Ours: 74%
  - Others: 43.19%
- Reach 70-90% compute efficiency across most networks

<table>
<thead>
<tr>
<th>Model</th>
<th>Input Size</th>
<th>MAC # (x $10^3$) in Model</th>
<th>Total MAC # (x $10^3$) in IP (including idle MAC)</th>
<th>Computing MAC Efficiency (Avg. vs. Peak Throughput)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG16</td>
<td>224x224</td>
<td>15.5</td>
<td>16.2</td>
<td>96%</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>224x224</td>
<td>1.59</td>
<td>2.13</td>
<td>74%</td>
</tr>
<tr>
<td>ResNet 34</td>
<td>224x224</td>
<td>3.67</td>
<td>4.55</td>
<td>81%</td>
</tr>
<tr>
<td>ResNet 50</td>
<td>224x224</td>
<td>3.87</td>
<td>5.28</td>
<td>73%</td>
</tr>
<tr>
<td>ResNet 101</td>
<td>224x224</td>
<td>7.59</td>
<td>9.78</td>
<td>78%</td>
</tr>
<tr>
<td>ResNet 152</td>
<td>224x224</td>
<td>11.3</td>
<td>14.3</td>
<td>79%</td>
</tr>
<tr>
<td>ResNext</td>
<td>224x224</td>
<td>691</td>
<td>849</td>
<td>81%</td>
</tr>
<tr>
<td>MobileNet V1</td>
<td>224x224</td>
<td>0.57</td>
<td>0.73</td>
<td>77%</td>
</tr>
<tr>
<td>MobileNet V2</td>
<td>224x224</td>
<td>0.28</td>
<td>0.37</td>
<td>76%</td>
</tr>
<tr>
<td>MobileNet + SSD</td>
<td>640x480</td>
<td>3.85</td>
<td>4.85</td>
<td>79%</td>
</tr>
<tr>
<td>Yolo V2</td>
<td>640x480</td>
<td>26.2</td>
<td>27.8</td>
<td>94%</td>
</tr>
<tr>
<td>Tiny Yolo</td>
<td>416x416</td>
<td>3.49</td>
<td>4.15</td>
<td>84%</td>
</tr>
<tr>
<td>DenseNet</td>
<td>224x224</td>
<td>13.9</td>
<td>14.2</td>
<td>98%</td>
</tr>
</tbody>
</table>
Applications
Applications

01 3D sensing & 3D recognition (with 3rd party 3D camera module)

02 Low cost 3D face recognition (with regular RGB + NIR camera)

03 Image & voice recognition

04 AI companion chip
1-1. 3D sensing
(with 3rd party 3D camera module)

Features
- Body / object detection
- Distance / depth detection

Scenarios
- Robot, drone, AR, Door lock
1-2. 3D recognition
(with 3rd party 3D camera module)

Features
- Face detection
- 3D face recognition
- Liveness detection

Scenarios
- Smartphone, door lock, IoT devices
2. Low cost 3D face recognition (with regular camera module)

**Features**
- Face detection
- 3D face recognition
- Liveness detection

**Scenarios**
- Door lock
- Door bell
- Smartphone
3. Image & voice recognition

Features

- 2D Face detection / recognition
- Landmark / body / gesture / object detection
- Voice wakeup
- Keyword recognition
- Voiceprint recognition

Scenarios

- Web camera / Smart speaker / Home appliance
4. Compiler flow for AI companion chip

- Translate models of various framework to Kneron IP’s corresponding instructions, weight file, and data flow controls
- Front-end converter
- Compiler
  - Input
    - ONNX model
  - Output
    - command.bin - command for input model
    - setup.bin - firmware management variable
    - weight.bin - weight from input model
  - Model inference in hardware based on compiler outputs
SoC Roadmap
AI surveillance on edge
- KDP 720
  - Heterogeneous structure
  - HW de-compression
  - Engineer sample in Q2 2019

Image & Voice AI sensing
- KDP 520
  - Image & voice recog
  - Low power
  - Engineer sample in Q2 2019

Edge AI server hub
- KDP 730
  - Cascade structure
  - Ultimate computing power

Integrated AI sensing hub
- KDP 530
  - Video processing & recog (ISP, video encoder)
  - Audio processing & recog (mic array processing)

AI sensing for wearable
- KDP 330
  - Small die size
  - Ultra low power (5~10mW)
Thank You

AI Everywhere

www.kneron.com