Scalable AI server

Q SynQuacer

Mar. 2018
S3 project

socio\text{next}
Where is the data for AI inference?

Object → CNN → Feature weight → Matching → Result

Where do we place data?
What we do on demo

Original Data set

VGG16 model with leaned weight

Object

❓

CNN

Feature weight

❓

Apple
In real world

Learning

Inference

Sensors

Learned data
For large size model data
Technical requirement

Object & Weight

Search matching data

Sensors

CNN part

Search part

CNN computing performance
# of channel for sensor connection

Scalability
Low latency
High throughput
AI specific hardware

2x performance
1/5x power

SPR280S AI module: 22TOPS x 16 CPU card = 352 TOPS
All CPU cards are connected via Socionext DDT
Logical Ethernet Communication on PCI Express

<table>
<thead>
<tr>
<th>Item</th>
<th>Performance on Socionext DDT (Direct Data Transaction) Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency</td>
<td>1.5us/path</td>
</tr>
<tr>
<td></td>
<td>Same as PCIe Transaction latency</td>
</tr>
<tr>
<td></td>
<td>Fixed Datapath and Direct Data transfer from Node to Node</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>20-40G Ethernet class</td>
</tr>
<tr>
<td></td>
<td>Performs Maximum TCP Window Speed from Beginning</td>
</tr>
</tbody>
</table>
Extended version

2x chassis connected by PCI express optical cable
PEC: Processor Element Card

Server node = PEC
CPU on PEC

Safe

CPU: Cortex A53

24 cores/1GHz
No speculation algorithm
No OOO execution
CPU-CPU connection

μSEC latency

Socionext DDT
Low latency
High throughput
Server-server connection

112x servers

DDT-R
Low latency
High throughput
100x efficiency

- Matrix Neural Processor
  High performance
  @ low clock frequency
- Outstanding power efficiency
  ➢ 7000 Gops/W
  • Google TPU: 613
  • Tesla P100: 60

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock (MHz)</td>
<td>100</td>
</tr>
<tr>
<td>Throughput (Top/s)</td>
<td>5.6</td>
</tr>
<tr>
<td>Power (W)</td>
<td>0.8</td>
</tr>
</tbody>
</table>
AI accelerator chip on 96 board

Movie