NG-IoT Workshop on Standardization

VEDLIoT Overview and Standardization activities

Jens Hagemeyer
Bielefeld University
**Very Efficient Deep Learning for IoT – VEDLIoT**

### Platform
- Hardware: Scalable, heterogeneous, distributed
- Accelerators: Efficiency boost by FPGA and ASIC technology
- Toolchain: Optimizing Deep Learning for IoT

### Use cases
- Industrial IoT
- Automotive
- Smart Home

### Open call
- 10 projects covering a wide range of AIoT applications
- Early use and evaluation of VEDLIoT technology

- **Call:** H2020-ICT2020-1
- **Topic:** ICT-56-2020 Next Generation Internet of Things
- **Duration:** 1. November 2020 – 31. Oktober 2023
- **Coordinator:** Bielefeld University (Germany)
- **Overall budget:** 7 996 646.25 €
- **Consortium:** 12 partners from 4 EU countries (Germany, Poland, Portugal and Sweden) and one associated country (Switzerland).

More info:
- [https://www.vedliot.eu/](https://www.vedliot.eu/)
- [https://twitter.com/VEDLIoT](https://twitter.com/VEDLIoT)
- [https://www.linkedin.com/company/vedliot/](https://www.linkedin.com/company/vedliot/)
VEDLIoT Hardware Platform

- Heterogeneous, modular, scalable microserver system
- Supporting the full spectrum of IoT from embedded over the edge towards the cloud
- Different technology concepts for improving
  - Performance
  - Cost-effectiveness
  - Maintainability
  - Reliability
  - Energy-Efficiency
  - Safety
RECS Architecture – RECS|BOX

**High-Performance Carrier**
(up to 3 microservers)

**Low-Power Carrier**
(up to 16 microservers)

---

**RECS Server Backplane** (up to 15 Carriers)

- **Microserver (High Performance)**
  - #1
  - #2
  - #3

- **Microserver (Low Power)**
  - #1
  - #2
  - #3
  - ... (up to 16)

- **Compute Network** (up to 40 GbE)
- **Management Network** (KVM, Monitoring, ...)
- **High-Speed Low-Latency Network** (PCIe, High-Speed Serial)

---

**High-Performance Microserver (COM Express)**
- x86
- ARM v8
- FPGA SoC

**Low-Power Microserver**
(Apalis/Jetson)
- FPGA SoC
- ARM Soc
- GPU SoC

---

**Ext. Connectors**
- QSFP+
- HDMI/USB
- RJ45
- iPass+ HD
RECS Architecture – t.RECS

t.RECS Edge Server
- Optimized platform for local / edge applications
- Provide interfaces for
  - Video
  - Camera
  - Peripheral input (USB)
- Combine FPGA and GPU acceleration
- Compact dimensions
  1 RU, E-ATX form factor
  (2 RU/ 3 RU for special cases)
RECS Architecture – u.RECS

u.RECS AIoT Server

- Supports ML acceleration
  - FPGA
  - ASIC
- Communication interfaces
  - Wired (CAN, Ethernet, CSI)
  - Wireless (WLAN, LoRa, 5G)
- Sensors
  - Camera
  - Environment (Temp./Hum.)
  - Housekeeping
- Embedded Device
  (~ 20x20x6 cm)
## Microserver overview

<table>
<thead>
<tr>
<th>CPU</th>
<th>COM Express Intel Core i7 8th Gen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Apalis Exynos (2xARM Cortex-A15)</td>
</tr>
<tr>
<td></td>
<td>COM Express ARM v8 Server SoC Hi1616</td>
</tr>
<tr>
<td></td>
<td>COM Express AMD EPYC 3451</td>
</tr>
<tr>
<td></td>
<td>COM Express AMD Ryzen V1807B</td>
</tr>
<tr>
<td></td>
<td>COM-HPC client size A Intel Core i7 11th Gen</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FPGA SoC</th>
<th>COM Express Xilinx Zynq 7045</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Apalis Xilinx Zynq 7020</td>
</tr>
<tr>
<td></td>
<td>COM Express Intel Stratix 10</td>
</tr>
<tr>
<td></td>
<td>COM-HPC client size B Xilinx Zynq UltraScale+</td>
</tr>
<tr>
<td></td>
<td>COM-HPC server size D Intel Agilex</td>
</tr>
<tr>
<td></td>
<td>COM-HPC client size B Xilinx Versal</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GPU SoC</th>
<th>Jetson TX2 NVIDIA Tegra X2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jetson nano NVIDIA Xavier NX</td>
</tr>
<tr>
<td></td>
<td>Jetson AGX NVIDIA Xavier</td>
</tr>
<tr>
<td></td>
<td>COM-HPC Size B NVIDIA Xavier AGX</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ML SoC</th>
<th>SMARC Coherent Logix HX40416</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SMARC Coral SoM</td>
</tr>
<tr>
<td></td>
<td>SMARC Xilinx Zynq UltraScale</td>
</tr>
</tbody>
</table>

---
Peak Performance of DL Accelerators

- Peak performance values of specialized accelerators, provided by the vendors (precisions varying from INT8 to FP32)

Average efficiency at 1000 GOPS /W
Yolo v4 accelerator performance

- Performance of Yolo v4 for different hardware platform has been evaluated
- Performance measurement for other networks (Resnet, EfficientNet) available as well
Microserver Standardization – COM-HPC

- Large, open consortium
- Specification final and released
- Driven by industry requirements

<table>
<thead>
<tr>
<th>COM-HPC Client</th>
<th>COM-HPC Server</th>
</tr>
</thead>
<tbody>
<tr>
<td>49x PCIe</td>
<td>65x PCIe</td>
</tr>
<tr>
<td>2x MIPI-CSI</td>
<td>8x 25GbE KR</td>
</tr>
<tr>
<td>2x 25GbE KR</td>
<td>BaseT (up to 10 Gb)</td>
</tr>
<tr>
<td>3x DDI</td>
<td>2x USB4</td>
</tr>
<tr>
<td>2x BaseT (up to 10 Gb)</td>
<td>2x USB3.2</td>
</tr>
<tr>
<td>2x SoundWire, I²S</td>
<td>4x USB2.0</td>
</tr>
<tr>
<td>4x USB4</td>
<td>2x SATA</td>
</tr>
<tr>
<td>4x USB2.0</td>
<td>eSPI, 2x SPI, SMB</td>
</tr>
<tr>
<td>2x SATA</td>
<td>2x I²C, 2x UART</td>
</tr>
<tr>
<td>eSPI, 2x SPI, SMB</td>
<td>2x I²C, 2x UART</td>
</tr>
<tr>
<td>2x I²C, 2x UART</td>
<td>12x GPIO</td>
</tr>
<tr>
<td>12x GPIO</td>
<td>12x GPIO</td>
</tr>
</tbody>
</table>
Microserver Standardization – COM-HPC

- Large, open consortium
- Specification final and released
- Driven by industry requirements

<table>
<thead>
<tr>
<th>COM-HPC Client</th>
<th>COM-HPC Server</th>
</tr>
</thead>
<tbody>
<tr>
<td>49x PCIe</td>
<td>65x PCIe</td>
</tr>
<tr>
<td>2x MIPI-CSI</td>
<td>8x 25GbE KR</td>
</tr>
<tr>
<td>2x 25GbE KR</td>
<td>BaseT (up to 10 Gb)</td>
</tr>
<tr>
<td>3x DDI</td>
<td>2x USB4</td>
</tr>
<tr>
<td>2x BaseT (up to 10 Gb)</td>
<td>2x USB3.2</td>
</tr>
<tr>
<td>2x SoundWire, lPS</td>
<td>4x USB2.0</td>
</tr>
<tr>
<td>4x USB4</td>
<td>2x SATA</td>
</tr>
<tr>
<td>4x USB2.0</td>
<td>eSPI, 2x SPI, SMB</td>
</tr>
<tr>
<td>2x SATA</td>
<td>2x I2C, 2x UART</td>
</tr>
<tr>
<td>eSPI, 2x SPI, SMB</td>
<td>2x GPIO</td>
</tr>
<tr>
<td>2x I2C, 2x UART</td>
<td>12x GPIO</td>
</tr>
<tr>
<td>12x GPIO</td>
<td></td>
</tr>
</tbody>
</table>
Reconfigurable DL accelerators

- VEDLioT accelerators support a large variety of reconfigurable architectures
  - From small embedded FPGAs to large ACAPs

- Large design space for FPGA-based accelerators

- Dynamic hardware reconfiguration
  - Adapt to changing requirements at run-time
  - Change characteristics of DL-accelerator
  - Trade-off between power and performance, power and accuracy, etc.

- Inference and training on FPGA
  - Supports quantization from int8 to float32
  - DL and Deep Reinforcement Learning
DL accelerator co-design

Monolithic design
- One engine computes all the core layers
- E.g. TPU

SEML
- One engine computes all layers of the same type
- PW engine, DW engine

SESL
- One engine per layer
- E.g. FINN

FiBHA
- SESL + SEML

*FiBHA: Fixed Budget Hybrid CNN Accelerator*, Fareed Qararyah, Muhammad Waqar Azhar, Pedro Trancoso, IEEE 34th International Symposium on Computer Architecture and High-Performance Computing (SBAC-PAD 2022), Bordeaux, France, November 2–5 2022
Security

- Common environment for running distributed applications
  - WebAssembly runtime + Trusted Execution Environment
  - Security for edge (and cloud) devices
- Advances on attestation
  - Better support for edge devices
  - Distributed (Byzantine fault-tolerant) attestation and configuration service
- Secure IoT Gateway
Safety and Robustness

Simulation platform for ML accelerators

- RISC-V SoCs and Custom Function Units
- Improve test and verification
- Co-simulate Verilog blocks
- Used in Google’s CFU Playground
- Continuous integration based in Gitlab and Google Cloud Platform
A compositional architecture framework for AIoT

Knowledge creation (e.g. definition of safety goals).

Concept design (e.g. introduction of redundancy to fulfil safety goals).

Final design (e.g. assigning functions to independent processors to guarantee redundancy).

Monitoring concept definition (e.g. monitoring fulfilment of safety goals at run-time).
Use case: Automotive

- Focus on collision detection/avoidance scenario
- Improve performance/cost ratio – AI processing hardware distributed over the entire chain
Use case: Industrial IoT – drive condition classification

- Control applications need DL-based condition classification
  - On the edge device for low power consumption
  - Suggestions for control and maintenance

- DL methods on all communication layers
  - DL in a distributed architecture
  - Dynamically configured systems

- Sensored testbench with 2 motors
  - Acceleration, Magnetic field, Temperature, IR-Cam (temperature), Current-Sensors, Torque

Challenge: Low-power / Efficiency

- On / Off detection without motor current or voltage
- Cooling fault detection
- Bearing fault detection
Use case: Industrial IoT – Arc detection

- AI based pattern recognition for different local sensor data
  - current, magnetic field, vibration, temperature, low resolution infrared picture

- Safety critical nature
  - response time should be <10ms
  - AI based or AI supported decision made by the sensor node itself or by a local part of the sensor network
Use case: Smart Mirror – Neural Networks

- Face recognition
  - Mobilenet SSD trained on WIDERFACE dataset
- Object detection
  - YoloV3, Efficient-Net, yoloV4-tiny
- Gesture detection
  - YoloV4-tiny with 3 Yolo layers (usually: 2 layers)
- Speech recognition
  - Mozilla DeepSpeech
- AI Art: Style-Gan trained on works of arts
- Collect usage data in situation memory

Challenge: Data privacy, Efficiency
Use case: Open calls

- **AI_RIDE**
  - Driving school

- **BEAM_IDL**
  - Laser welding

- **FLAIR**
  - 5G federated learning

- **MushR**
  - Mushroom harvesting

- **DUNE RCO**
  - Indoor localization

- **FLEDGED**
  - AI for Wearables

- **Power Edge RL**
  - AI for power electronics

- **AccBD**
  - Biomarker discovery

- **Edge4iwelli**
  - Smart mirror

- **Honey.AI**
  - Pollen analysis
Summary – Standardization in VEDLIoT

- Hardware/microserver form factors
  - Active contribution to PICMG Standards COM-HPC and COM Express ([https://www.picmg.org/openstandards/com-hpc](https://www.picmg.org/openstandards/com-hpc))

- Several Open Source contributions to large projects ([https://vedliot.eu/open-source-software](https://vedliot.eu/open-source-software))
  - Renode + Kenning – Emulator and Simulator for distributed IoT, Verilator support
  - Memory Protection for RISC-V: RISC-V PMP
  - TEEs support for WebAssembly: Integration for Trustzone (ARM) and SGX (Intel) into WebAssembly

- Recommendations: Design framework IoT and AI
  - Compositional architecture framework for AIoT developed within VEDLIoT
  - Can help system design to comply with regulatory constraints (e.g. EU AI Act)
Thank you for your attention.

Contact

Jens Hagemeyer, Carola Haumann
Bielefeld University, Germany
chaumann@cor-lab.uni-bielefeld.de
jhagemey@cit-ec.uni-bielefeld.de