

Toward Among-Device AI from On-Device AI with Stream Pipelines

MyungJoo Ham*
myungjoo.ham@samsung.com

Sangjung Woo
sangjung.woo@samsung.com

Jaeyun Jung
jy1210.jung@samsung.com

Wook Song
wook16.song@samsung.com

Gichan Jang
gichan2.jang@samsung.com

Yongjoo Ahn
yongjoo1.ahn@samsung.com

Hyoung Joo Ahn
hello.ahn@samsung.com

Samsung Research, Samsung Electronics
Seoul, Republic of Korea

ABSTRACT

Modern consumer electronic devices often provide intelligence services with deep neural networks. We have started migrating the computing locations of intelligence services from cloud servers (traditional AI systems) to the corresponding devices (on-device AI systems). On-device AI systems generally have the advantages of preserving privacy, removing network latency, and saving cloud costs. With the emergent of on-device AI systems having relatively low computing power, the inconsistent and varying hardware resources and capabilities pose difficulties. Authors' affiliation has started applying a stream pipeline framework, NNStreamer, for on-device AI systems, saving developmental costs and hardware resources and improving performance. We want to expand the types of devices and applications with on-device AI services products of both the affiliation and second/third parties. We also want to make each AI service atomic, re-deployable, and shared among connected devices of arbitrary vendors; we now have yet another requirement introduced as it always has been. The new requirement of "among-device AI" includes connectivity between AI pipelines so that they may share computing resources and hardware capabilities across a wide range of devices regardless of vendors and manufacturers. We propose extensions of the stream pipeline framework, NNStreamer, for on-device AI so that NNStreamer may provide among-device AI capability. This work is a Linux Foundation (LF AI & Data) open source project accepting contributions from the general public.

ACM Reference Format:

MyungJoo Ham, Sangjung Woo, Jaeyun Jung, Wook Song, Gichan Jang, Yongjoo Ahn, and Hyoung Joo Ahn. 2022. Toward Among-Device AI from On-Device AI with Stream Pipelines. In *Proceedings of The 44th International*

*The corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

ICSE 2022, May 21–29, 2022, Pittsburgh, PA, USA

© 2022 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

Conference on Software Engineering (ICSE 2022). ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 INTRODUCTION

Since the rise of deep neural networks, we have witnessed a lot of AI applications affecting our daily lives. Such AI applications have been expanding and have started to run on devices: on-device AI [14]. Mobile phones run photo enhancements, video stream filtering, automatic speech recognition, noise-canceling, and object detection with deep neural networks on devices. Televisions run video stream filtering, audio enhancements, and contents analysis. Robotic vacuums run object detection, localization, and mapping. Even consumer electronics products that do not appear to be "smart", ovens, refrigerators, and air conditioners, start to run deep neural networks on devices. Running them on devices can preserve privacy; consumers do not want to send live video streams from robotic vacuums to servers. It reduces network costs, especially for high-bandwidth live data such as camera streams. It also reduces the cloud service costs; we have millions of new devices deployed each month.

The previous work [8] is an on-device AI pipeline framework enabling developers to describe and execute AI systems as stream pipelines. We have deployed it to recent devices of the affiliation and a few software platforms. Implementing on-device AI systems as stream pipelines with NNStreamer has satisfied the related requirements mentioned in [8], has improved the overall performance, and has saved the developmental costs. After its initial deployments, we have observed new types of on-device AI applications pursued: distributed on-device AI systems where inputs, outputs, and inferences might happen on different devices. Developers have implemented prototypes of such systems with off-the-shelf GStreamer plugins of TCP, UDP, and RTSP, allowing interconnecting independent pipelines.

We envision an open IoT ecosystem where any vendors may deploy their devices with AI services and connect to AI services of nearby devices within a house, office, or building. Such an IoT ecosystem should be able to avoid exposing data to clouds so that AI services may fully utilize personal data safely. Intelligence services may be executed on multiple devices in such systems, sharing hardware resources and data: "among-device AI". Among-device

AI may improve user experiences. For example, devices without high computational power may provide intelligence services by connecting to nearby high-end devices, not cloud servers. In other words, we can provide AI services with consistent quality regardless of the computation capability of the interfacing devices; thus, we can expand the interfaces and chances to provide AI services. We may save average device costs by sharing computational power because only a few devices need high computational power.

For instance, a television without computational power for real-time pose estimation may provide an AI-based exercise trainer by connecting to a mobile phone or a home IoT hub that has enough computational power. Conventional speakers and cameras with WiFi connections may become additional user interfaces or sensors for intelligence services. An oven or washer/dryer unit without powerful processors may provide intelligence services of automatic speech recognition and context-aware natural language understanding by connecting to nearby refrigerators or mobile phones.

Raw network connection plugins mentioned above have been satisfactory for prototypes of the above examples. However, they have critical issues for actual deployment, and AI application developers have expressed the need for improvement. More critically, application developers want to connect pipelines by expressing capabilities, not IP addresses. With initial prototypes satisfying such needs, we could have acquired further requirements from application developers and iterated a few times with a few more prototypes. We can list up the requirements briefly as follows:

- R1. Each AI or input/output service exists atomically and is deployed independently. Such a service includes an AI inference processor, a shared input stream, or a user interface.
- R2. Transmission between such entities may be schemaless, or updating schema should not require recompiling or redeploying software. It should be able to compress and synchronize transmitted data frames from different devices.
- R3. Discovering such entities and connecting to them should not require the knowledge of specific addresses (e.g., IP address). Besides, there may be multiple entities discovered for a given connection request, where one of them will be connected.
- R4. If a connected entity becomes unavailable, it can be automatically connected to an alternative entity.
- R5. Any vendors and application developers may freely use, alter, or redistribute the given software framework for any purpose without charges. Besides, it is desirable if anyone may participate in its development to reduce fragmentation and a bias for an affiliation.
- R6. It should be extensible for different platforms. In other words, it should be able to be connected with different AI pipeline frameworks (e.g., MediaPipe [12] and DeepStream [15]) or services with different operating systems, especially those for microcontrollers (a.k.a. RTOS).
- R7. Each AI service is an on-device AI application that has requirements mentioned in [8].

Our approach extends the on-device AI pipeline platform, NNStreamer, to the among-device AI pipeline platform. The on-device AI capability inherited from [8] R7 is preserved while new requirements (R1 to R6) are satisfied by this work. To satisfy the new requirements, we first extend the AI stream data type (“other/tensors”

MIME) to support tensors with flexible dimensions and schemaless data streams. Then, in addition to off-the-shelf networking plugins including TCP, UDP, and RTSP, we propose to use MQTT [4] for service discovery and connections, which is applied to both publish/subscribe and workload offloading, along with synchronization methods. We also provide a lightweight library with minimal dependencies, “NNStreamer-Edge”, which allows implementing non-NNStreamer software packages compatible with the proposed among-device AI pipeline connections.

Our main contribution includes:

- Based on users’ feedback, including AI algorithm researchers, application and OS developers, and device vendors, we identify requirements, propose a corresponding design, and deploy the implementation.
- Extending the previous work [8], we provide mechanisms reflecting the lessons learned by deploying products.
- We propose and deploy stream transmission protocols for among-device AI systems tested for products. The implementation is open sourced and maintained in public so that anyone may deploy products with the proposed mechanisms.

2 RELATED WORK

NVIDIA has proposed a DeepStream AI pipeline framework based on GStreamer for NVIDIA hardware customers [10]. Recent releases of DeepStream [15] propose edge-to-cloud AI pipelines. In addition to the issues of not treating tensors as 1st class citizens of stream data [8], it is proprietary software dedicated to NVIDIA hardware, which cannot meet R5, R6, and R7.

Google has proposed MediaPipe AI pipeline framework [12], which is now targeting on-device AI applications of IoT devices in addition to cloud AI services. Although they may have among-device AI capabilities and inter-device connectivity internally, they are not publishing documents or releasing codes of such features.

GStreamer [7] has already provided various connectivity plugins so that multimedia pipelines may listen to media servers or broadcast to media players: TCP, UDP, RTSP, and others. It also provides data serialization plugins, GStreamer Data Protocol (GDP-PAY [22]), to connect remote pipelines. Extending such capability, TRAMP [24] has proposed a distributed multimedia system with GStreamer pipelines partially satisfying R1, R2, and R3 for general multimedia applications, which, in turn, has given hints when we were requested to develop a software framework for edge-AI or among-device AI systems.

The previous work [8] is an AI pipeline framework satisfying requirements for on-device AI systems of various consumer electronics. We extend NNStreamer further with new requirements identified for among-device AI systems.

3 DESIGN

As in the previous work [8], the first principle is to re-use well-known and battle-proven open source software components. GStreamer [7] and its off-the-shelf plugins are still the basis of this work. For capability-based publish/subscribe stream connections, we adopt MQTT [4] via “paho.mqtt.c” library [3]. We have started initial prototypes with TCP and RTSP plugins. Then, we have applied ZeroMQ [9], which, fortunately, has an off-the-shelf and open sourced

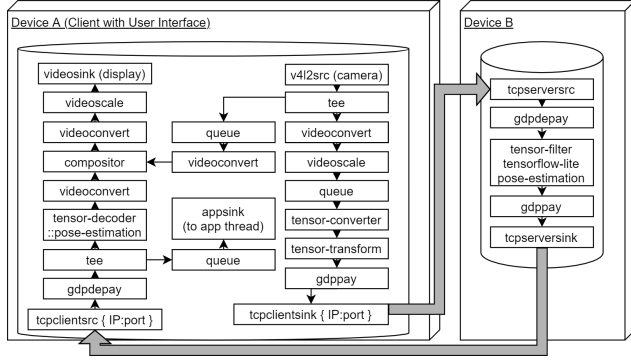


Figure 1: An offloading example with the previous work.

GStreamer plugin for second prototypes. Then, we have switched to MQTT [4] for the last prototypes. We choose MQTT because, ultimately, we want to contribute to IoT standards including Matter [1] and SmartThings [20], which include MQTT. Inter-pipeline connectivity for among-device AI is deployed as GStreamer plugins making it compatible with general GStreamer pipelines. As a result, the provided capability is not restricted to AI applications but is available to general stream pipelines.

Initially, for earlier prototypes for internal clients, we have assumed that data serializations with GDPPAY [22], Protocol Buffers, and FlatBuffers for static tensor streams would suffice along with off-the-shelf network transport GStreamer plugins of TCP, UDP, and RTSP. However, we have found requirements not satisfied by such design by iterating prototypes with clients. R1, R2, R3, and R4 enforce applications to implement sophisticated network handling, which can be avoided by providing such features with NNSStreamer. Besides, implementing such features in applications inevitably leads to fragmentation and compatibility issues: every application may implement protocols differently.

For example, a simple inference workload offloading (query) application with TCP plugins in Figure 1 has initially appeared satisfactory. However, having multiple clients—another additional user requirement—over-complicates pipelines. Even with a single client, separating query (*tcpclientsink*) and answer (*tcpclientsrc*) filters of a client complicates pipelines with synchronization and topology issues. Thus, this approach does not satisfy R1. Besides, it requires clients to specify IP addresses and port numbers of servers, neglecting R3 and making it challenging to satisfy R4.

For R1 and R4, we propose query capability as pipeline filters, shown in Figure 2. Client pipelines become trivial and we can easily support multiple clients and multiple servers for an instance of service. We integrate network protocols and serialization mechanisms in the client/server filters for the simplicity of pipelines. For the robustness of services, it is desirable to allow multiple servers (Device B) compatible with given service requests from clients (Device A).

Another demonstrative among-device AI user scenario is a system with multiple input devices and independent processing and output devices. The example application for this scenario has two input devices with USB cameras, one processing device (demonstrated with a Google Coral USB accelerator), and one output device

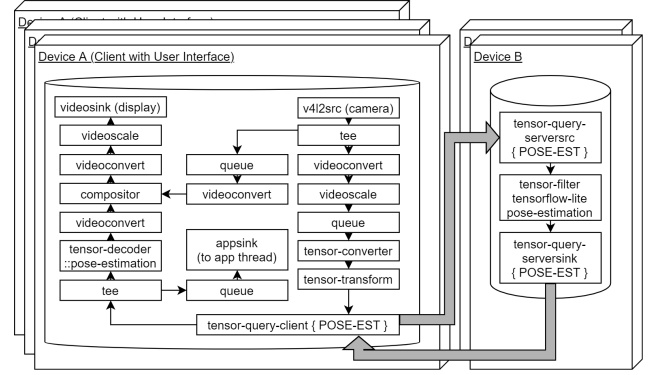


Figure 2: An offloading example with the proposed work.

(demonstrated with an LCD). This scenario stands for home IoT systems with lightweight input devices (sensors, cameras, and microphones) and output devices (speakers and displays). Vendors may expose their AI services via such lightweight devices (e.g., refrigerators, washers, and inexpensive display devices) while executing AI services at home for privacy protection, network latency reduction, and cloud cost reduction. Such AI services can be consistently (for both latency and quality) provided across different user interfacing devices regardless of their processing power if they satisfy R1, R2, and R3. In other words, such services can achieve the above if devices can discover available resources and services, connect to the discovered ones efficiently, and provide their resources to other devices. If a vendor wants to expand their device ecosystem by inviting other vendors' devices, R5 and R6 are also required.

Some internal clients have specified a data serialization method for inter-pipeline connections: schemaless FlexBuffers of FlatBuffers. We do not recommend using schemaless FlexBuffers for connecting stream pipelines; we recommend dynamic (flexible) schema instead. Schemaless protocol usually incurs more overheads and run-time issues; we cannot verify data types at launch. Moreover, being schemaless is meaningless in many among-device AI systems; i.e., it transfers the responsibility of interpreting the output of a neural network in a sender from the sender to the receiver. In other words, except for not generating errors for data type mismatches (we do not recommend this but have failed to persuade the clients), it burdens the receiver to be more aware of the specifications of the sender. However, we have accepted it with the condition of using FlexBuffers only for research prototypes and using “other/tensors” MIME types for products. Note that having different data types between prototypes and products may incur unnecessary technical debts.

Writing pipelines with raw network protocol plugins (TCP, UDP, and RTSP) over-complicates pipelines, as we have experienced with the previous example. For example, we may have multiple pipelines sharing the output of a single pipeline. We may also have a pipeline that needs to connect to any compatible pipeline available or does not want to specify network addresses. Thus, after the initial prototype with ZeroMQ [9], we have implemented MQTT [4] stream pub/sub plugins compatible with general GStreamer data streams. With MQTT pub/sub plugins, we have implemented prototype

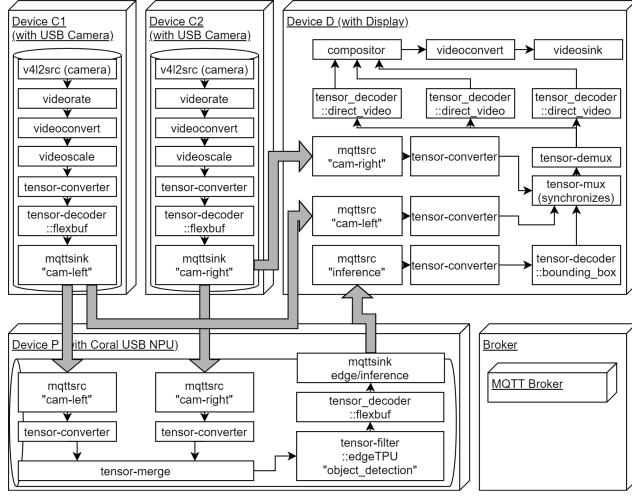


Figure 3: Stream pub/sub connections for distributed IoT AI application example.

pipelines described in Figure 3, showing that NNStreamer meets the requirements of the given application scenario. Because MQTT requires a broker for relaying messages and discovering published services, users need to deploy an MQTT broker service. In the figure, we have omitted flows of messages redirected by the broker.

With this prototype instance, the client’s initial requirements (AI application developers) have been satisfied. However, as it has always been, the client has found additional requirements after seeing the demonstration. The client has requested another demonstration of pipelines that minimizes the difference between timestamp values of the two camera input frames when Device P merges frames from the two. In this system, Device C1 and C2 mark timestamps when the data frame is created in the pipeline or given by the camera hardware. Another additional requirement is the capability to compress data transmitted between devices. The last additional requirement is the robustness of connections where alternative resources can be connected if one fails in run-time.

The first additional requirement is satisfied with the same pipeline architecture by updating MQTT plugins (mqttpsink and mqttpsrc) with additional features, including the NTP [13] protocol, which we discuss in the next section. The second is about performance optimization, which is more appropriate with products than with prototypes. By abandoning FlexBuffers from the pipelines, as mentioned above for products, we can easily apply compression mechanisms (zlib-gst [2], JPEG for each frame, or MPEG real-time compression). We further provide performance optimization for inter-pipeline transmissions with an additional protocol for MQTT, which we discuss in the next section. The nature of MQTT and the way we use MQTT for inter-pipeline connections satisfy the last additional requirement: connect with the capability of pipelines, not with the address of devices. Note that some clients have explicitly requested sparse tensor streams to compress streams for language and speech models.

4 IMPLEMENTATION

This section describes how we have upgraded NNStreamer for among-device AI based on the design described in the previous section: stream data types, protocols for inter-pipeline transmissions, and “NNStreamer-Edge” library for further compatibility with different vendors and operating systems. Then, we describe how the requirements, R1 to R7, are met with the given implementation. Lastly, we describe additional features assisting among-device AI features and updates from the previous work.

4.1 Data types

R2 does not only imply using FlexBuffers [5] for inter-pipeline transmissions. Data streams in a pipeline are often required to be schemaless (or dynamic schema). For example, a pipeline may pre-process a live video stream with an object detection neural network to crop the video, making a video stream filled with a detected human body. Then, the pre-processed tensor stream representing the cropped video may have varying dimensions per frame, which another neural network (e.g., pose estimation) may use as its input stream. Thus, even while schemaless data exchanges between pipelines might be useless, there is a need for the dynamic schema.

We update the tensor stream data type, “other/tensors” (MIME for GStreamer capability), so that users may specify the format of *static*, *dynamic*, and *sparse*. The conventional tensor stream with schema is *static*, and it is the default format. With *dynamic* format, the dimension and type may vary for each frame in a stream, which can serve schemaless data from remote pipelines and stream with dynamic schema. Unlike *static* format, which does not have format information in the buffers of a data frame, the *dynamic* format has a header in each data frame buffer that specifies dimension and type for each data frame. The other format, *sparse*, allows expressing tensors with the coordinate list format (COO) [19].

Filters for tensor streams, *tensor_** elements, are updated to handle *static* and *dynamic* formats. However, *tensor_** elements do not directly handle *sparse* formats because its binary representation of tensor data is not compatible with the other two formats. Thus, we provide converting filters, *tensor_sparse_enc* and *tensor_sparse_dec*.

In addition to the Protocol Buffers and FlatBuffers support of the previous work, we add FlexBuffers of FlatBuffers for schemaless data transmissions as sub-plugins of *tensor_converter* (FlexBuffers to “other/tensors”) and *tensor_decoder* (“other/tensors” to FlexBuffers).

4.2 Protocols

There are various raw network protocol stream filters as off-the-shelf shared libraries of GStreamer. According to GStreamer’s reference page at <https://gstreamer.freedesktop.org>, they include TCP, UDP, RTSP, HTTP, HTTPS, FTP, HLS, and many others with different serialization mechanisms. However, as explained in the previous section, they are not enough for among-device AI applications while they appear almost complete for multimedia applications and services.

4.2.1 Pub/Sub Protocol. Publish-subscribe architecture has been widely accepted by the robotics community, including ROS [16]. AI application and service developers have requested a similar capability to publish services and subscribe to the services. After

iterations of different implementations from ROS, ZeroMQ [9], and MQTT [4], we have decided to use MQTT as the basis. ROS is not chosen because it is over-complicated for the need, requires recompiling software for updated schema, and excessively requires software packages for its dependencies. On the other hand, ZeroMQ and MQTT are lightweight and do not require additional software packages for their dependencies. ZeroMQ is the most lightweight among these; however, we have chosen MQTT because major home IoT standards (Matter [1] and SmartThings [20]) use MQTT already. We want to make the among-device AI capability of NNStreamer compatible with such home IoT standards.

For pub/sub capability (Figure 3), we provide two GStreamer plugins: *mqttpsink* and *mqttpsrc*. With *mqttpsink*, a pipeline may publish a stream (output of the pipeline) declared with a topic string. With *mqttpsrc*, a pipeline may subscribe to a published stream and fetch an input stream for the pipeline from an output stream of another pipeline discovered by the given topic string. The topic string is independent of GStreamer capability (GSTCAP), representing the type information of a stream between two pipeline elements. A schemaless stream may have a GSTCAP of “other/flexbuf”, and a receiving pipeline should interpret as a properly structured stream type. A dynamic-schema stream may have a GSTCAP of “other/tensors,format=flexible”, and a receiving element (a neural network or a tensor processing element) can directly handle without additional conversion or interpretation.

We use “Eclipse Paho MQTT C client library” [3] for MQTT implementation. It is open source software with high coverage of MQTT features and high portability (an independent C/C++ library), which has a port for lightweight RTOS devices (e.g., FreeRTOS). Via MQTT connections, among-device AI transport plugins transmit data and metadata, including the corresponding GSTCAP and data size.

4.2.2 Query Protocol. We provide three GStreamer plugins for query capability (inference workload offloading in Figure 2): *tensor_query_client*, *tensor_query_serversrc*, and *tensor_query_serversink*. In a pipeline, *tensor_query_client* behaves equivalently to *tensor_filter* representing a neural network model. Thus, for other parts of the pipeline or the application running the pipeline (the client), *tensor_query_client* hides all the details for inference task offloading transparently and may be switched with local on-device AI elements. It sends queries (input stream for a neural network) to *tensor_query_serversrc* and receives inference results from *tensor_query_serversink*. In a server-side pipeline, the two elements, *tensor_query_serversrc* (input for the server) and *tensor_query_serversink* (output for the server), are paired and share information of client connections. In case there are multiple clients for a server-side pipeline, *tensor_query_serversrc* tags a client ID to the stream’s metadata, which *tensor_query_serversink* accesses to choose the proper client connections.

Query elements implement two different transport protocols that users may choose: TCP-raw and MQTT-hybrid. With TCP-raw, the connection between clients and servers are raw TCP connections, which does not provide the flexibility and robustness required by R3 and R4. MQTT-hybrid transmits the connection and control information via MQTT connections, which easily satisfies R3 and R4 by allowing multiple server pipelines compatible with a given topic

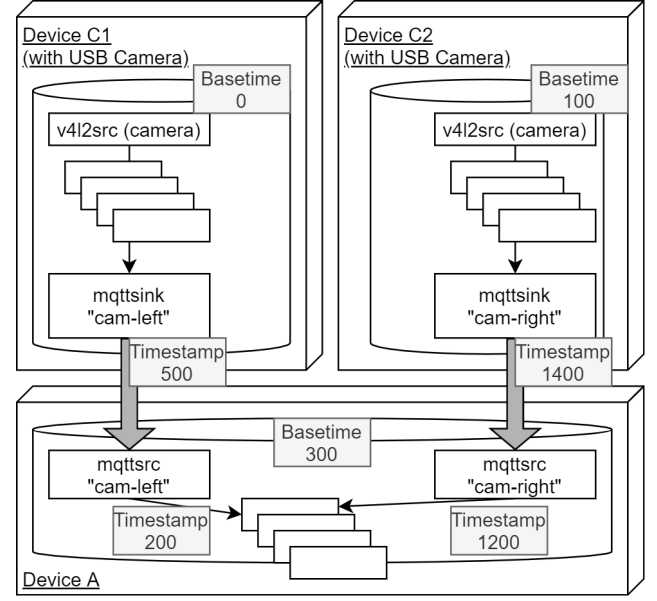


Figure 4: Timestamps from multiple source pipelines [21]

requested by a client. To allow multiple servers for a given topic of a client pipeline, we use wildcards and topic filters for MQTT topics, which subscribers (clients) use to choose publishers (servers) dynamically. For example, with servers of “/objdetect/mobilev3” and “/objdetect/yolov2”, a client may choose either of them by subscribing to “/objdetect/#”.

MQTT-hybrid transmits data transmission via direct TCP connections without brokers for higher throughput and lower broker overheads. Note that MQTT connections are not suitable for high-bandwidth streaming because of excessive overheads of a broker; thus, per clients’ requests, we have designed the MQTT-Hybrid protocol for queries. The clients of pub/sub have not yet requested higher bandwidth; however, we will provide MQTT-hybrid along with pure MQTT for pub/sub with the subsequent releases of NN-Streamer.

4.2.3 Timestamp Synchronization. An internal client has requested to minimize temporal differences between multiple input sources. For such needs, we add an inter-pipeline synchronization mechanism for pub/sub streams [21]. Note that “query” does not suffer from such synchronization issues because client pipelines do not depend on the timestamp values of servers. This timestamp synchronization mechanism makes stream publishers send base-time values of the publishing pipeline converted to universal time and relative timestamp values of buffers. Then, receivers (subscribers) correct timestamp values of incoming buffers with their base-time. The mechanism relies on clock synchronization between connected pipelines implemented by NTP protocol synchronizing relative clocks of inter-connected pipeline elements (*mqttpsink* and *mqttpsrc*). We can test this by injecting latency to a publisher in Figure 3 by inserting the *queue2* GStreamer plugin, which holds data until the specified size or duration.

4.3 NNStreamer-Edge Library

NNStreamer-Edge is a lightweight and portable library implementing the proposed protocols to connect with NNStreamer pipelines without adopting NNStreamer. NNStreamer-Edge is an open source software package independent from NNStreamer and its basis, GStreamer: <https://github.com/nnstreamer/nnstreamer-edge>. It does not depend on NNStreamer or GStreamer so that devices that cannot afford GStreamer or heavy operating systems may easily use NNStreamer-Edge. NNStreamer-Edge has minimal library dependency, Paho MQTT C client library. This library also has minimal library dependency; if SSL is disabled, it only requires essential tools (e.g., GCC and libc).

NNStreamer-Edge of October 2021 supports tensor stream publish for remote sensors and cameras: “edge_sensor” module. This module behaves like an “mqttpsink” NNStreamer element with “other/tensors” streams to connect with “mqttpsrc” NNStreamer elements. We have designed other modules of NNStreamer-Edge, “edge_output” and “edge_query_client”, but have not yet released them.

The primary objective is to extend connectivity for devices of arbitrary vendors; thus, the proposed mechanisms should be compatible with or included in alliances such as Matter [1] and SmartThings [20]. Because NNStreamer-Edge has chosen Apache 2.0 license and does not have extra dependencies, anyone may implement their proprietary software with NNStreamer-Edge. For example, third-party developers may implement a proprietary MediaPipe plugin with NNStreamer-Edge so that arbitrary MediaPipe pipelines may communicate with NNStreamer pipelines. Note that DeepStream pipelines are GStreamer pipelines; thus, they can trivially connect to NNStreamer pipelines, which are also GStreamer pipelines.

4.4 How the requirements are met

This section describes how the proposed implementation satisfies each requirement mentioned in the first section.

- R1. As long as we implement AI services and shared input/output streams as NNStreamer pipelines, we can deploy them atomically as pipeline instances or pipeline descriptions. With *tensor_query_client/server*, we can offload inference tasks to another pipeline. With *mqttpsrc/sink*, a pipeline may serve as a data stream publisher or subscriber.
- R2. If a stream is typed as “other/flexbuf” using “Flexbuf” sub-plugins, it is schemaless. A NNStreamer-native stream can be schemaless if it is typed as “other/tensor,format=flexible”. The former is for compatibility with third-party software, and the latter is for in-pipeline or inter-pipeline streams.
- R3. The adoption of MQTT for connections has satisfied this requirement. Multiple subscribers or clients can connect with a server with topic names. Multiple publishers or servers may be available for a given client or subscriber to choose with topic filters and wildcards. The timestamp synchronization mechanism described in Section 4.2.3 supports synchronization. Sparse tensors and *gst-gz* [2] support compressed transmissions.
- R4. MQTT client libraries and MQTT brokers handle this.
- R5. NNStreamer is LGPL 2.1 with all source codes included as default plugins to satisfy this requirement. For systems without NNStreamer and GStreamer, but with their proprietary middleware, we release NNStreamer-Edge with Apache 2.0 so that users may distribute their software without the condition of LGPL 2.1. Anyone may participate in designing and developing both packages in open space, Github.com.
- R6. NNStreamer is cross-platform and deployed for different operating systems officially: Tizen, Android, Ubuntu, Yocto, and macOS. In addition, users have reported using it in Debian and OpenSUSE as well. It would not be too difficult to port it for Windows or iOS, but we have not tried it yet. The inter-pipeline connectivity library, NNStreamer-Edge, is designed to make the among-device AI capability compatible with different operating systems and different pipeline frameworks.
- R7. We satisfy this by allowing developers to implement AI services with conventional NNStreamer plugins and adding inter-pipeline transmission plugins. For example, in Tizen, adding among-device AI capability (Tizen 6.5 M2) does not break the backward compatibility of the machine learning API set.

5 USAGE EXAMPLE AND EVALUATION

This section describes exemplar among-device AI systems with implementation details and experimental results.

5.1 Workload Offloading with Query

Listing 1: The code implementing workload offloading with query elements, depicted in Figure 2.

```
# Device A code
v4l2src ! tee name=ts
ts. videoconvert ! videoscale !
  video/x-raw,width=300,height=300,format=RGB !
  queue leaky=2 ! tensor_converter !
  tensor_transform ${TROPT}$ !
  tensor_query_client operation=${SVCNAME}$ !
  tee name=tc
ts. queue leaky=2 ! videoconvert ! mix.sink_1
tc. queue leaky=2 ! appsink name=appthread
tc. tensor_decoder mode=${DECMODE} ${DECOPTS} !
  videoconvert ! mix.sink_0
compositor name=mix sink_0::zorder=2 sink_1::
  zorder=1 ! videoconvert ! videoscale !
  video/x-raw,width=640,height=480 ! ximagesink

# Device B code
tensor_query_serversrc operation=${SVCNAME} !
tensor_filter framework=tensorflow-lite model=${
  MODELPATH}$ ! tensor_query_serversink
```

Listing 1 shows a script code representing GStreamer pipelines in Figure 2. GStreamer API or its wrapper, NNStreamer API (Tizen ML API), can execute such a script. We can also execute the script directly on a shell with “gst-launch” for prototyping and testing. We can apply such pipelines to home IoT systems consisting of a

device with comfortable user interfaces without high computing power (e.g., an inexpensive TV) and a device with less comfortable user interfaces with high computing power (e.g., a home IoT hub or a mobile phone). In other words, if a user has an inexpensive but large display (Device A) and a high-end mobile phone (Device B). The user may run home fitness and training applications by offloading a pose-estimation neural network to Device B while using the camera and screen of Device A.

Please note the simplicity of server (Device B) code; declaring the service name (*\$SVCNAME*) is all developers need to do. If users want more operations in servers—e.g., pre-processing or data collecting—, they may add corresponding filters to the pipeline. The client-side pipeline is straightforward as well. The only changes from its on-device AI equivalent pipeline are replacing *tensor_filter* with *tensor_query_client* and declaring the service name. Then, we may have multiple clients and multiple servers for the given service name for resource sharing and robustness.

The string variables in Listing 1 depend on neural network models and their input and output formats. For example, if it is a Mobilenet V2 object detection model [18] from TensorFlow Hub with COCO 2017 dataset [23], the variables are:

```

TROPT = "mode=arithmetic_option=typecast:float32,
        add:-127.5,div:127.5"
SVCNAME = "objectdetection/ssdv2"
DECMODE = "bounding_boxes"
DECOPTS = "option1=mobilenet-ssd_option2=/PATH/
          coco_labels_list.txt_option3=/PATH/box_priors.
          txt_option4=640:480_option5=300:300"
MODELPATH = /PATH/ssd_mobilenet_v2_coco.tflite

```

Configurations and behaviors of queues and merging points (compositor in this example) are crucial for the efficiency of parallelism. With the *leaky=2* option, a *queue* drops older buffers if it becomes full. Users may alter options, including the size of the queue and leaky modes, for further optimization.

5.2 Stream Pub/Sub

Listing 2: The code implementing remote sensors with pub/sub in Figure 3 along with timestamps mechanisms.

```

# Device C1 / C2
v4l2src do-timestamp=true ! videoconvert !
videorate ! video/x-raw,width=${W},height=${H},
        format=RGB,framerate=10/1 !
tensor_converter ! tensor_decoder mode=flexbuf !
queue ${OPT} ! mqttsink pub-topic=${CAM} sync=true

# Device P
tensor_merge name=m mode=linear option=1 !
tensor_decoder mode=direct_video ! videoscale !
video/x-raw,width=300,height=300,format=RGB !
tensor_converter ! tensor_filter framework=edgetpu
        , model=${MODELPATH} !
tensor_decoder mode=flexbuf !
mqttsink pub-topic=edge/inference

mqttsrc sub-topic=camleft is-live=false !

```

```

other/flexbuf ! tensor_converter ! queue !
m.sink_0
mqttsrc sub-topic=camright is-live=false !
other/flexbuf ! tensor_converter ! queue !
m.sink_1

# Device D
compositor name=mix sink_0::xpos=1 sink_0::ypos=0
        sink_0::zorder=0 sink_1::xpos=${W} sink_1::
        ypos=0 sink_1::zorder=0 sink_2::xpos=1 sink_2
        ::ypos=0 sink_2::zorder=1 !
videoconvert ! ximagesink

tensor_mux name=mux ! queue !
tensor_demux name=dmux

dmux.src_0 ! tensor_decoder mode=direct_video !
queue ! mix.sink_0
dmux.src_1 ! tensor_decoder mode=direct_video !
queue ! mix.sink_1
dmux.src_2 !
tensor_decoder mode=direct_video option1=RGBA !
queue ! mix.sink_2

mqttsrc sub-topic=camleft is-live=false !
other/flexbuf ! tensor_converter ! queue !
mux.sink_0
mqttsrc sub-topic=camright is-live=false !
other/flexbuf ! tensor_converter ! queue !
mux.sink_1

mqttsrc sub-topic=edge/inference is-live=false !
queue ! other/flexbuf ! tensor_converter !
other/tensors,num_tensors=4,dimensions="
        4:20:1:1,20:1:1:1,20:1:1:1,1:1:1:1",types="
        float32,float32,float32,float32" !
option1=tf-ssd option2=${LABELPATH}
option3=0:1:2:3,40 option4=${W}:${H}
option5=300:300 !
tensor_converter ! queue ! mux.sink_2

```

Listing 2 shows a GStreamer pipeline description of the application example shown in Figure 3, where FlexBuffers [5] serializes published streams. When the *v4l2src* creates a video frame buffer, fetching video streams from a USB camera attached to a Raspberry Pi 4 board in our experiments, the *v4l2src* provides a timestamp value with the option of *do-timestamp=true* to enable the mechanism in Figure 4. Note that this code does not require the camera to be a USB camera or the system to be a Raspberry Pi 4. The same code works for an embedded camera of a mobile phone or a typical Linux desktop PC without modification. If multiple cameras are attached to the system, users may specify it with additional options; otherwise, it uses */dev/video0*.

The pipeline topology of the target application in Figure 3 has become more complex than the application in Figure 2. It would require a considerable amount of time and effort to implement such systems without pipeline frameworks as we have experimented

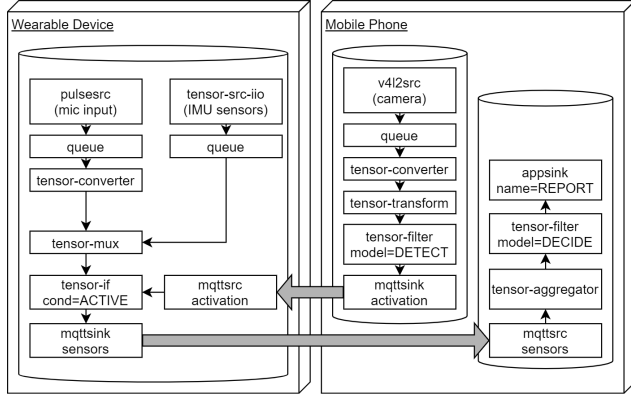


Figure 5: Augmented worker application with a wearable device and a mobile device.

in [8]; i.e., it would probably require well over thousands of lines of codes. With the off-the-shelf GStreamer/NNStreamer plugins and effortlessly applied pipe-and-filter architecture, users can write such an among-device AI system within 100 lines of codes, which is easy to extend and port. We can execute the same pipeline descriptions in Ubuntu PC, Yocto devices, Tizen devices (TVs, home appliances, robots, and wearable devices), or Android mobile phones. With the NNStreamer-Edge library, developers can interconnect pipelines with more varying devices, including MediaPipe/DeepStream pipelines running in clouds or workstations and lightweight devices (microcontrollers) running RTOS.

Note that FlexBuffers, FlatBuffers, or Protocol Buffers are not required to connect NNStreamer pipelines and NNStreamer-Edge instances only. For other independent non-NNStreamer processes subscribing to the published streams, we can apply FlatBuffers, FlexBuffers, or Protocol Buffers, popular data serialization mechanisms.

5.3 Multi-device and Multi-modal

Another example, Figure 5, shows an augmented worker application, which is both multi-device and multi-modal. Such a system may assist workers in manufacturing plants by detecting and notifying events requiring attention. For example, if a worker assembles parts incorrectly, the system sends an alarm to the worker.

In the left-hand side pipeline of the mobile device, the “DETECT” model detects if an action of assembling parts is starting and let the wearable device know. Then, the wearable device will start streaming related data from the microphone and IMU sensors back to the mobile device. Then, the right-hand side pipeline of the mobile device decides whether the assembling activity is correct or incorrect based on the data from the wearable device and reports to the application logic. To further optimize power consumption in the wearable device, we may turn on and off the sensors based on the “activation” signal from the mobile device. As we demonstrate with previous examples, the pipeline description code of this example incurs short lines of codes as well; i.e., usually a single line per block with exceptions of *tensor_if*, where we need to describe the condition with a script or a C function.

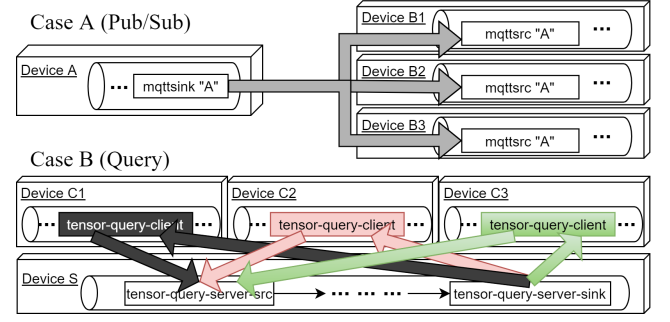


Figure 6: Pipelines for performance evaluation.

5.4 Performance Evaluation

We evaluate the proposed among-device AI transport mechanisms: MQTT pub/sub and MQTT-hybrid query along with their lighter and faster counterparts: ZeroMQ pub/sub and TCP query. We have experimented with Raspberry Pi 4 boards connected via Ethernet and NNStreamer 2.1.0-unstable. Figure 6 shows the evaluated among-device AI pipelines. Case A evaluates the stream pub/sub of MQTT and ZeroMQ. Case B evaluates the query client/server streams of MQTT-hybrid and TCP direct connections. We measure the throughput, CPU usage, and peak memory consumption to evaluate both performance and overhead. We experiment with three different bandwidths of input streams from Device A and Device C: high, mid, and low bandwidths. Each bandwidth corresponds to a Full-HD video stream, a VGA (640x480) video stream, and a QQVGA (160x120) video stream with a 60 Hz framerate.

Figure 7 shows the evaluation results of MQTT pub/sub and MQTT-hybrid query, normalized by their counterparts, ZeroMQ and TCP, respectively. Bars in the figure are normalized standard deviations, and values average five experimental runs of 30 seconds. Case M and H have failed to transmit 60 Hz, implying that the network bottlenecks the throughput. MQTT suffers from lower throughput and higher client memory overhead with high bandwidth streams (M and H). MQTT-hybrid successfully eliminates the performance overheads of MQTT while keeping the rich features of MQTT. For example, with the MQTT-hybrid query, if multiple servers are compatible with the client’s topic, the client pipeline is automatically switched to another server when a connected server becomes unavailable. Moreover, server pipelines may declare additional specifications for clients to choose, e.g., “server workload status” and “neural network model and version”. Note that the query plugins have both TCP direct and MQTT-hybrid implementations switchable at run-time. For higher performance, we plan to add MQTT-hybrid for pub/sub plugins, as well.

6 CONCLUSION

We have extended the on-device AI pipeline framework [8] for among-device AI systems (often marketed as edge AI). The proposed framework, NNStreamer, is developed and released at the GitHub organization owned by Linux Foundation, <https://github.com/nnstreamer>, where anyone may participate in discussions and

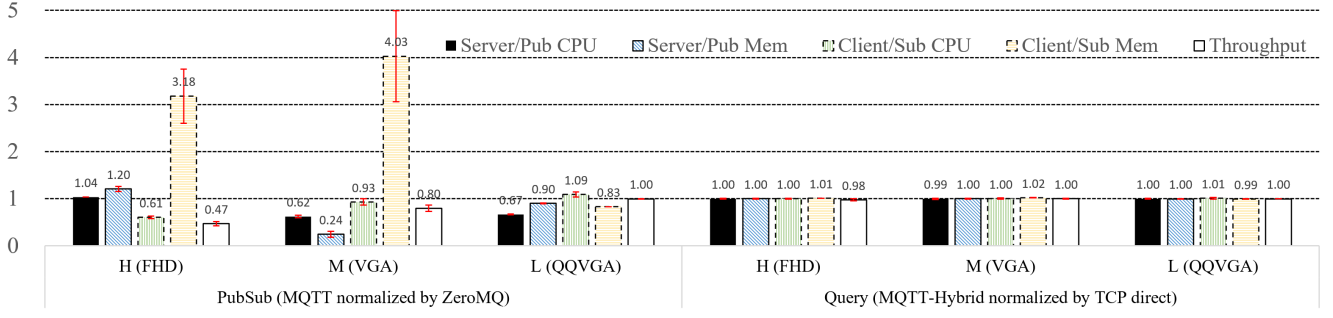


Figure 7: Evaluation results of Pub/Sub (MQTT normalized by ZeroMQ) and Query (MQTT-Hybrid normalized by TCP).

code contributions. It is deployed to the mentioned operating systems daily or regularly: Tizen (as default machine learning framework), Android (for Android Studio via public repositories), Yocto (via meta-neural-network layer), Ubuntu (Launchpad PPA), macOS (homebrew repository). We are deploying the among-device AI capabilities introduced in this paper via the Tizen 6.5 M2 release in October 2021; note that a few features (NNStreamer-Edge and MQTT-hybrid) are not yet approved by internal clients and are omitted in Tizen 6.5 releases but available in GitHub. A few home appliances of 2022, including TVs, will be using Tizen 6.5 M2; thus, the mentioned features will be available for products in 2022, and Tizen Studio users writing applications for such products.

6.1 Lessons Learned

After wide deployment of the previous work [8] across many prototypes and products, including mobile phones, wearable devices, TVs, and home appliances, we have worked on prototypes of various internal clients based on the proposed among-device AI methods. As a result, we have observed the following lessons and future work.

Many users appear to feel barriers against adopting pipe-and-filter architecture. It is easy to show that pipelines with NNStreamer work appropriately and significantly better than the conventional implementation for both developmental costs and run-time performance. However, we could have observed unexpectedly steep learning curves to adopt pipeline concepts and to describe pipeline topology. For the former issue, we are preparing to write more diverse pipeline examples for users. For the latter issue, we are implementing a WYSIWYG pipeline editor with a converter translating between GStreamer scripts and MediaPipe scripts (pbtxt) to re-use MediaPipe’s pipeline editor.

Analyzing and profiling pipeline performance becomes more complicated with among-device AI pipelines. We have an AI pipeline profiling tool, nnshark, which forks GstShark [6], created by a group of undergraduate students as an open source software project. We have ported nnshark for Tizen and Android devices and deployed it to users. However, with among-device AI capability, users are not satisfied with nnshark, and request profiling capability for the whole system consisting of multiple pipelines simultaneously.

We have been implementing initial prototypes or practicing pair programming for internal clients to address learning curves and

developer relations. For external users of open source communities, we try to catch up with technical questions in various channels (Slack, Github issues, mailing lists, and direct contacts) in addition to documents and sample applications. However, with frequent requests from internal clients and our roadmap of new features, writing documents and sample applications have always had lower priority and been behind schedule. Besides, with a few core contributors focusing on code contributions, it is challenging to motivate writing documents. It is also difficult to justify having a dedicated document writer while we directly support internal clients so that they do not even bother to read the documents.

NNStreamer has external users that have contacted the authors: NXP Semiconductors, Collabora, ODKMedia, and finders.ai. Feedback and contributions from them are constructive for identifying issues and requirements. Motivating external users to communicate with maintainers is essential; they are often too shy to do so. We recommend keeping such communication in public channels to motivate other users to communicate with maintainers and the community.

Users often write pipelines incorrectly or inappropriately, and it is usually too late when we find it out. For example, an NNStreamer application [11] drawing bounding boxes of detected objects with live video streams has an inappropriate pipeline design. Creating video streams from neural network output is supposed to be handled by *tensor_decoder*. Its authors [11] are aware of it; however, for neural network output formats not supported by default sub-plugins of *tensor_decoder*, they have abandoned *tensor_decoder* and made their application thread directly decode outputs and push decoded data into the pipeline. It is supposed to write a *tensor_decoder* sub-plugin for such a case, which incurs lower overhead, higher throughput, and less implementation effort. Fortunately, although inappropriately implemented, their NNStreamer implementation is shown to beat Open CV implementation with higher throughput. Promoting communication with the NNStreamer community may address this issue; however, we have observed similar cases with internal users: the users have sometimes released inappropriate pipelines before addressing them. Such releases are especially troubling considering that we are still at an early stage of deploying the pipeline paradigm to AI developers in the affiliation. Besides, the standard practice of software verification and testing is not ready for such issues. We may need another verification and testing

process for newly introduced frameworks with a new concept, allowing framework developers to audit. Such processes will provide more usage cases for framework developers and more chances to improve their frameworks.

6.2 Directions of Evolution

NNStreamer has started evolving for among-device AI systems. Besides the apparent performance optimization, we have suggestions for among-device AI systems and pipeline frameworks as future work.

With frameworks including NNStreamer [8], DeepStream [15], and MediaPipe [12], developers have started writing AI systems as pipelines. These frameworks provide various documents and sample pipelines to mitigate difficulties suffered by developers writing AI pipelines. However, as our users have complained, they are not enough.

We would like to suggest another method directly intervening pipeline development process. First, we need to provide common parts of pipelines (sub-pipelines) as libraries; that developers can invoke or insert sub-pipelines in their pipelines. There are various common parts in different AI applications: e.g., pre-processing video streams for object detection or audio streams for RNN-T [17]. This feature may often prevent inappropriately designing pipelines as well.

Then, we need a pipeline run-time repository where processes may register pre-defined pipelines, and other processes may invoke such pipelines. This feature enables operating systems or middleware to register pipelines for applications. It enables applications without particular AI features to invoke such pipelines without actually writing pipelines. Moreover, suppose a vendor wants to separate the application development and AI development divisions; this approach will cleanly separate code repositories for corresponding divisions as a client has wanted.

We have further future directions that require more time and effort; the above can be implemented and deployed within a year. We envision an IoT ecosystem where devices of any vendors may join, which is related to R5 and R6. Matter [1] along with SmartThings [20] is a promising candidate for such an IoT ecosystem; however, it lacks a data transmission protocol for inter-device AI services. In the future, we expect to propose such a protocol with among-device AI capabilities mentioned in this paper with extensions of interoperability with microcontrollers and other pipeline frameworks. With such protocols included in the IoT ecosystem, connected devices will be able to consistently provide AI services regardless of interface devices' computation power. This inclusiveness and consistency will promote the proliferation of various on-device and among-device AI services without the need for exposing privacy and private data to external computing nodes.

ACKNOWLEDGMENTS

We greatly appreciate contributions from the open source software community with various affiliations: not only Samsung but also NXP Semiconductors, Collabora, universities, and independent hobbyists. We also appreciate developers and users who gave us precious feedback and other core NNStreamer contributors, including Jijoong Moon, Parichay Kapoor, Dongju Chae, Jihoon Lee, and

Hyeonseok Lee. The discussions with and advice from the following Samsung Research members have especially been crucial for this project; Daehyun Kim, Gu-Yeon Wei, JeongHoon Park, Semun Lee, Jinmin Chung, Sunghyun Choi, Daniel Dongyuel Lee, and Sebastian Seung. This project is supported by Samsung Research of Samsung Electronics and LF AI & Data Foundation.

REFERENCES

- [1] Connectivity Standards Alliance. (accessed 1 Oct 2021). Matter. <https://buildwithmatter.com/>.
- [2] Alexandre Esse. 2017 (accessed 1 Oct 2021). zlib based GStreamer plugins. <https://github.com/Snec/gst-gz> (GitHub repository).
- [3] Eclipse Foundation. (accessed 1 Oct 2021). Eclipse Paho C Client Library for the MQTT Protocol. <https://github.com/eclipse/paho.mqtt.c> (GitHub repository).
- [4] Simeon Furrer, Wolfgang Schott, Hong Linh Truong, and Beat Weiss. 2006. The IBM wireless sensor networking testbed. In *2nd International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities, 2006. TRIDENTCOM 2006*. IEEE, 5–pp.
- [5] Google. (accessed 1 Oct 2021). FlexBuffers. <https://google.github.io/flatbuffers/flexbuffers.html>.
- [6] Michael Gruner. 2017. GstShark profiling: a real-life example - GStreamer. (Oct 2017). <https://gstreamer.freedesktop.org/conference/2017/talks-and-speakers.html> RidgeRun, GStreamer Conference 2017.
- [7] GStreamer. 1999 (accessed 1 Oct 2021). gstreamer open source mutlimedia framework. <https://gstreamer.freedesktop.org/>.
- [8] MyungJoo Ham, Jijoong Moon, Geunsik Lim, Jaeyun Jung, Hyoungjoo Ahn, Wook Song, Sangjung Woo, Parichay Kapoor, Dongju Chae, Gichan Jang, Yongjoo Ahn, and Jihoon Lee. 2021. NNStreamer: Efficient and Agile Development of On-Device AI Systems. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. 198–207. <https://doi.org/10.1109/ICSE-SEIP52600.2021.00029>
- [9] Pieter Hintjens. 2013. *ZeroMQ: messaging for many applications*. " O'Reilly Media, Inc."
- [10] Tushar Khinvasara. 2018. Introduction to DeepStreamSDK. (Oct 2018). <https://gstreamer.freedesktop.org/conference/2018/talks-and-speakers.html> Nvidia, GStreamer Conference 2018.
- [11] SeongJoo Lee, Yechan Choi, and Yongjun Park. 2020. Implementation of Object Detection System for Real-time Video with NNStreamer. In *2020 Korea Software Congress*. The Korean Institute of Information Scientists and Engineers, 1559–1561.
- [12] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. MediaPipe: A Framework for Building Perception Pipelines. *arXiv preprint arXiv:1906.08172* (2019).
- [13] David L. Mills. 1985. Network Time Protocol (NTP). RFC 958. <https://doi.org/10.17487/RFC0958>
- [14] MIT Technical Review. (accessed 1 Jun 2020). On-Device AI. <https://www.technologyreview.com/hub/ubiquitous-on-device-ai/>.
- [15] Nvidia. 2018 (accessed 7 Oct 2021). DeepStream SDK. <https://developer.nvidia.com/deepstream-sdk>.
- [16] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source Robot Operating System. In *ICRA Workshop on Open Source Software*, Vol. 3.
- [17] Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. 2017. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 193–199.
- [18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.
- [19] SciPy. 2008 (accessed 8 Oct 2021). SciPy documentation: scipy.sparse.coo_matrix. https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.coo_matrix.html.
- [20] SmartThings. (accessed 1 Oct 2021). SmartThings: one simple home system. A world of possibilities. <https://smartthings.com/>.
- [21] Wook Song. 2021 (accessed 1 Oct 2021). Synchronization in the MQTT elements. <https://github.com/nnstreamer/nnstreamer/blob/main/Documentation/synchronization-in-mqtt-elements.md>.
- [22] Thomas Vander Stichele. (accessed 1 Oct 2021). gdpay (GDP/Payload). <https://gstreamer.freedesktop.org/documentation/gdp/gdpay.html>.
- [23] TensorFlow. (accessed 1 Oct 2021). TensorFlow Hub: ssd_mobilenet_v2. https://tfhub.dev/tensorflow/ssd_mobilenet_v2/2.
- [24] Ján Vorčák. 2014. *Creating a GStreamer plugin for low latency distribution of multimedia content*. Master's thesis. University of Oslo.