AIST
Create the Future, Collaborate Together

CloudEdge Fusion

# Building Frameworks with Edge-side Data Processing and Federated Learning on The Computing Continuum

In the CloudEdge Fusion (CEF) project, we are developing foundational technologies that make it easy and efficient to aquire, process, and utilize a wide variety and massive amount of data from the real world, by maximizing the "high-speed, large-capacity" and "massive simultaneous connections" enabled by 5G and post-5G networks. ”Dexpot” and “REFL” frameworks represent our latest research outcomes.

## Dexpot: Common Data Processing with Accelerators at Edge Servers

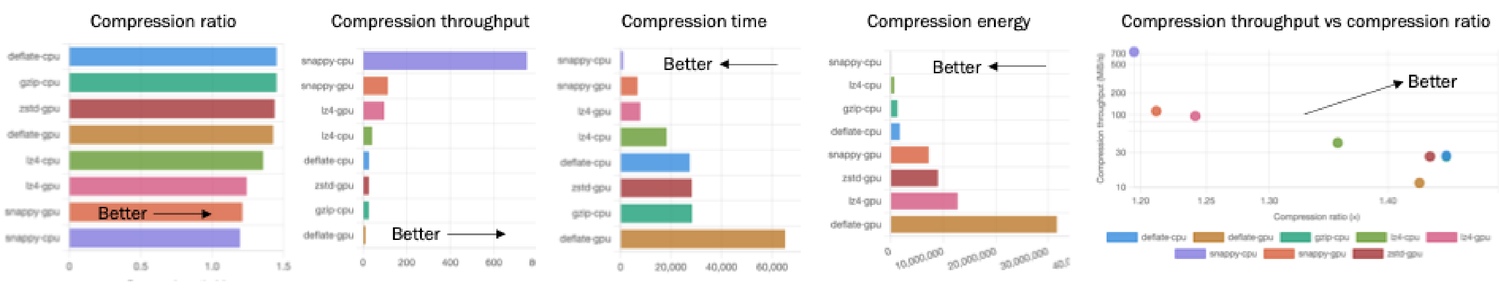
Dexpot intermediates data transfer from devices, and allows users to properly reduce data amount sent to clouds.

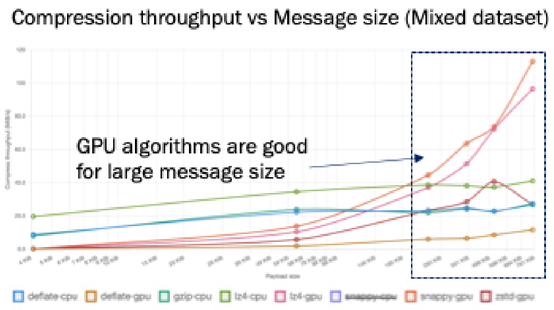
- Pre-defined data reduction by using accelerators is supported. Currently, CPU/GPU/FPGA based compression is available.
- Both SQL-like and storage-level user-defined data reduction with accelerators can be implemented as a data processing module.
- Monitoring system allows users to choose optimal algorithms for different deployment scenarios—balancing data reduction efficiency, processing speed, energy consumption, and memory usage on edge servers.

### Compression performance analysis (CPU vs GPU)

- Datasets: Numerical, images, audio, mixed datasets (message size: 4KB - 768KB)
- Metrics: Compression ratio; Compression time ($\mu s$); Throughput (MB/s); Energy ($\mu J$)

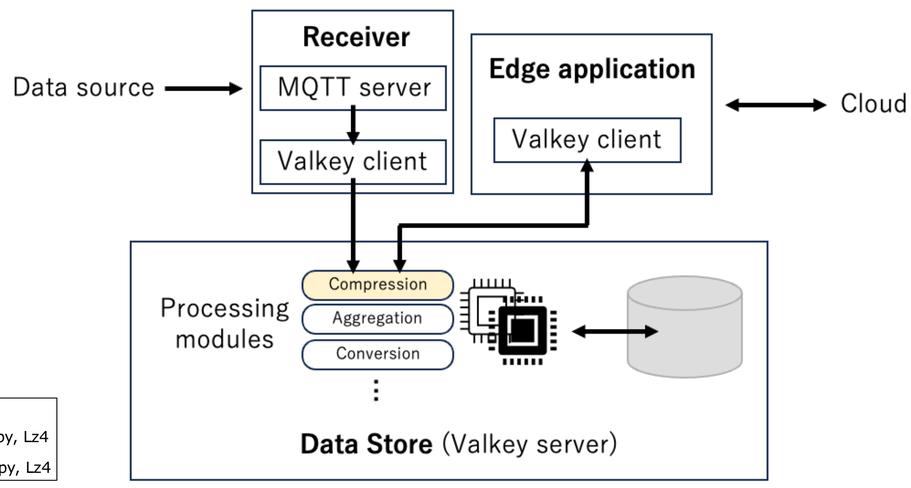
**Implementation**: C, Nvidia nvCOMP v5
**CPU compresison**: Deflate, Gzip, Snappy, Lz4
**GPU compression**: Deflate, Zstd, Snappy, Lz4




| Scenario | Best Choice | Win | Cost |
|---|---|---|---|
| Best compression ratio | Deflate/Gzip-CPU | Highest compression | Slow & energy hungry |
| Best throughput | Snappy-CPU | ~600+ MB/s | Poor ratio |
| Best energy | Snappy/LZ4-CPU | Lowest power | Moderate ratio |
| Small msgs | Snappy/LZ4-CPU | Fast, efficient | Moderate ratio |
| Large msgs | Zstd/Snappy-GPU | Throughput scales | High energy, bad at small sizes |
| Balanced ratio + speed | LZ4-CPU / Zstd-GPU | Middle ground | Not best at either |

### Dexpot architecture




### Execution example:

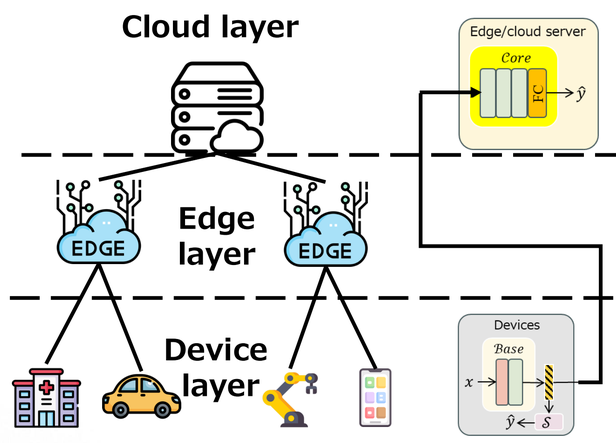
(Enable GPU-based LZ4 compression and store data into Dexpot by setting timestamp as a key)
$ valkey-cli DEXPOT.SYS.INITONCE
$ valkey-cli DEXPOT.ADM.ASSIGNDB 4 by gpu
$ valkey-cli -x DEXPOT.ADM.REGISTERGRP < group_conf.json
$ valkey-cli **dexpot.push** group1 topic1 1777447573.00 ”sample data”

(Load all compressed data for the specified topic from Dexpot, without decompressing)
$ valkey-cli **dexpot.pull** group1 topic1 ”*” -1 -1 0 ASIS

## REFL: Resource-Aware Hybrid Federated Learning with Coordinated Edge–Cloud Workload Reduction

### Challenges (Training at IoT devices)

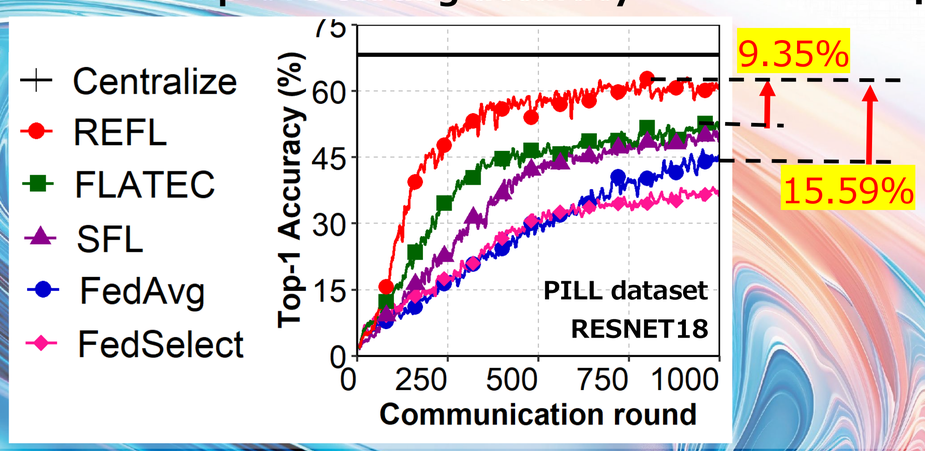
- Communication bottleneck at cloud server.
- Large communication overhead comes from frequent model updates.
- Resource-limited IoT devices make full-model training difficult.
- Low accuracy caused by heterogeneous data distributions (non-IID data).



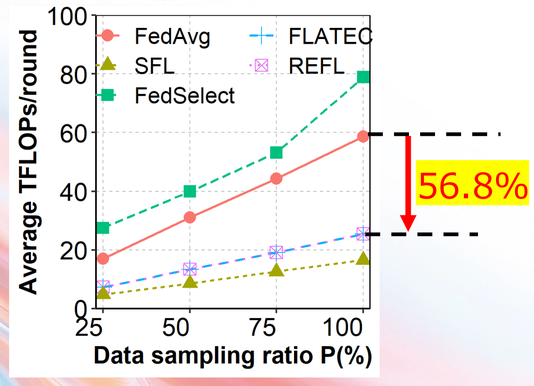

### Our Approach (Hybrid Training)

- Centralized training computationally expensive Core model using feature data.
- Temporary store feature data from devices
- Intermediate feature data selection.
- Training feature extraction layers (light-weight base model) using FL.
- Upload p% of extracted feature data to edge servers.

### Improve testing accuracy




### Reduced computation required at edge devices




### Reduce the cloud communication to achieve target accuracy (54%)

| | Comm. Round | Cloud Traffic (GB) |
|---|---|---|
| **FLATEC** | 969 | 325.6 |
| **REFL** | 306 | 46.0 |
| **Reduction** | 68.4% | 85.9% |

**AIST SC Website**

This work is based on results obtained from the project ”Research and Development Project of the Enhanced Infrastructures for Post-5G Information and Communication System” (JPNP20017), commissioned by the New Energy and Industrial Technology Development Organization (NEDO).