An IOWN Global Forum PoC report
Sensor Data Aggregation and Ingestion in the disaggregated edge computing by RDMA over IOWN All-Photonics Network

MAY 1ST, 2024

Hyde Sugiyama
Chief Technology Strategist, Red Hat
Blog
Who am I?

Hidetsugu (Hyde) SUGIYAMA has been with Red Hat for ten years. He is Global Chief Technology Strategist where he is now focused on Telecom, Media Entertainment and Edge segment. During his 35 years, he has worked in the telecommunications industry on distributed systems, multi-layer networking, programmable network, SDN/NFV, virtualization, and heterogeneous computing. He also serves as the co-lead of PoC consultation Task Force and an alternate director of IOWN Global Forum.

Current his majors are;

a) Data Centric Infrastructure (DCI) for AI-native disaggregated infra over All-Photonics Network
b) Open Programmable Infrastructure in IOWN DCI/Composable Disaggregated Infra
c) Mobile Network for Beyond 5G toward 6G
d) Reference Implementation Model for IOWN Cyber Physical System/AI integrated Communication use cases
e) Post Quantum Security
The Holistic Approach of IOWN (Innovative Optical and Wireless Network)

The following vision is to achieve by 2030.
Lower Power consumption by 100x  Higher transmission capacity by 125x  Lower end-to-end latency by 200x

1. Low delay
2. High quality and high capacity
3. Lower power consumption
4. Data-centric for AI (App/computing-aware FDN (Function Dedicated Networks))
5. Optimization for Computing and Networking scale
6. Location aware/agnostic
7. User’s applications
8. Memory access innovation

The following vision is to achieve by 2030.
Lower Power consumption by 100x  Higher transmission capacity by 125x  Lower end-to-end latency by 200x
IOWN and the Linux Foundation - Open Source Collaboration for Driving the Future of Communication Infrastructure
by Arpit Joshipura,

The Linux Foundation’s goal is to create the greatest shared technology investment in history by enabling open collaboration across companies, developers and users.

We are the nonprofit organization of choice to accelerate open source technology development and commercial adoption on a global scale.
MOU between IOWN Global Forum and the Linux Foundation

1) IOWN GF RIM Task Force collaboration opportunity
   • Exploring LF Edge Akrino new blueprint by Disaggregated Edge Computing with RDMA over IOWN All-Photonics (ex IOWN CPS use case RIM SDAI PoC)

2) IOWN GF DCI Task Force collaboration opportunity
   • CNCF Kubernetes Dynamic Resource Allocation in IOWN Data-Centric Infrastructure/Composable Disaggregated Infrastructure

3) IOWN GF INS/DCI Task Force collaboration opportunity
   • Open Programmable Infrastructure with DPU/IPU for IOWN function dedicated network (ex Distributed UPF, RDMA, etc)

4) IOWN GF OAA Task Force collaboration opportunity
   • LF Networking OpenDaylight SDN Transport PCE for IOWN All-Photonics Network Controller Network


Contact: H.SUGIYAMA (hsugiyam@redhat.com)
IOWN use case RIM PoC (WIP) and LF Projects

- Needs Dynamic Resource Allocation for GPU/FPGA/IPU/DPU in Composable Disaggregated Infrastructure
- NIC(DPU/IPU) needs more intelligence
- Exploring for new edge connecting with IOWN infrastructure
- Needs Multi-layer SDN controller(s) + APN controller for deterministic networking and computing

Contact: H.SUGIYAMA (hsugiyma@redhat.com)
Joint Press Release

NTT and Red Hat Fuel AI Analysis at the Edge with IOWN

---

Red Hat blog

PR at NTT

PR at Red Hat

---

(A) Dashboard of Kepler

(B) Power Consumption of a Container

KEPLER

An Energy Efficiency project in IOWN GF(WIP)

40% - 60% Power consumption reduction

---

Ref: IOWN GF reference implementation model for CPS Area Management use case

60% Latency reduction for AI analysis

Contact: H.SUGIYAMA (hsugiyma@redhat.com)
Overview of IOWN SDAI PoC

• Demonstration system

• Demonstration results (with single GPU)
  • Latency required to aggregate sensor data for AI analysis was reduced by 60%
  • Power consumption of AI analysis for each camera was reduced by 40%
  • You can find the details in the recognized PoC report in IOWN GF.
    • https://iowngf.org/recognized-pocs/
    • “Sensor Data Aggregation and Ingestion”

NOTE: To confirm the effectiveness of our accelerated data pipeline, APN is consistently used through all the evaluation, including evaluation of conventional technologies.
Overview of CPS Use Cases in IOWN Global Forum

Continuously analyze a large number of sensor data streams (e.g., video images and LiDAR data) at remote computing sites (or Regional Edge Cloud) to enable prompt actions and automation.

- **Monitored Area** (city, factory, highway, etc.):
  - ~1,000 video cameras
  - ~1,000 LiDAR sensors

- **Local Aggregation**

- **Regional Edge Cloud** (Remote GPU service):
  - ~ several hundreds of Km
  - about 70 Gbps of bandwidth

- **Real-time AI Analysis**

- **Aggregated Hardware Resources**

---

Contact: H.SUGIYAMA (hsugiym@redhat.com)
Accelerated AI Workloads

- **Accelerated data pipeline for AI inference:**
  - RDMA over APN for reducing the protocol-handling overheads in the conventional network
  - Full-offloading of AI workloads to GPU for improving the power efficiency

- **Large-scale AI data analysis in real time:**
  - Red Hat OpenShift, supporting Kubernetes operators, for enabling improved flexibility and easier deployment across disaggregated sites, including remote data centers

---

**Diagram:**

- **RDMA over APN** to directly fetch large-scale sensor data from local sites to the memory in an accelerator in a remote data center.
- **Monitored Area** (Aggregation of sensor data)
- **Customer Edge**
- **Regional Edge Cloud** (AI analysis of sensor data)

---

**Key Points:**

- **IOWN APN** (~a few hundreds of Km)
- **IOWN DCI's full-offloading** to completes AI analysis within accelerators (i.e., SmartNIC and GPU), powered by NVIDIA's libraries for data pipeline acceleration
- **Containers and k8s operators** to hide the complexity of implementing hardware-based accelerators
Summary

- RDMA base Remote GPU service by Innovative Optical and Wireless Network is new approach for AI infrastructure biz.
  - Successfully demonstrated 60% AI latency reduction, 40% power save.

- RDMA based multi-vendor Customer Edge solution helps to accelerate the remote GPU service use case supporting Energy Efficiency project in IOWN Global Forum.
  - New common customer edge will be needed!
Thank you

linkedin.com/company/Red-Hat

facebook.com/RedHatInc

youtube.com/user/RedHatVideos

twitter.com/RedHat

https://iowngf.org/