Defining the Future of Edge Computing Using Micro Data Centers

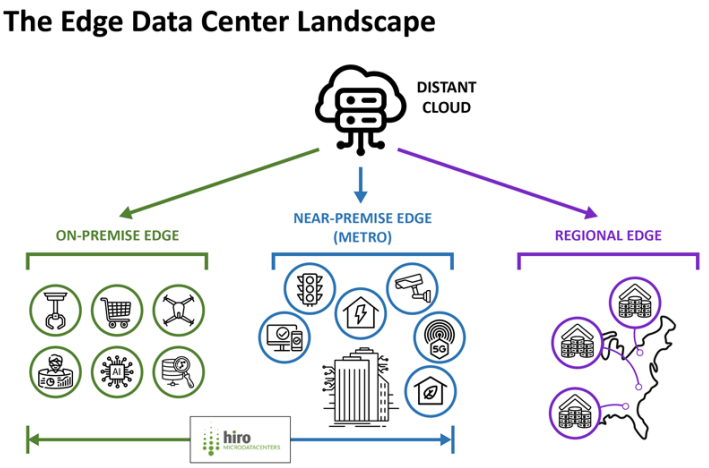
Author: Fred Buining, CEO of HIRO-MicroDataCenters
Date: 05/20/2024

Categories: AI (Artificial Intelligence)

Tag: #HIROMicroDataCenters #edgecomputing #powerelectronics

New breed of edge data centers must optimize power efficiency

Edge computing refers to the production and consumption of data at the edge of the internet. There, engineers are busy connecting groups of things, and even complete network environments, and equipping them with sensing, data and AI-processing capabilities. Data flows across the entire network landscape and Edge computing allows this data to be monetized, increasingly through AI-based services that are creating a new data economy.



Click image to enlarge

Figure 1. The edge data center landscape (HIRO 2023)

AI-based edge data center services can best be situated at locations where the majority of data is being processed and/or consumed. As a result, the predicted growth of edge computing was twice that of cloud computing through 2023, with a projected compound annual growth rate (CAGR) of 37.9%. Here are examples of the estimated divide between local edge device/edge-data center processing and distant cloud processing:

Autonomous Driving: 80-90% of data processing takes place in the car (object detection, lane tracking, collision avoidance) and at the roadside (safety, maintenance, emissions, zoning). The remainder involves data transmissions to and from distant locations for travel updates, maps, traffic information, media streams and centralized services such as parking and tolls.

Industrial IoT: 70-80% of processing happens at the edge in the form of real-time monitoring, control and immediate response to anomalies. The other 20-30% is addressed in the cloud and includes data transmission for centralized analysis, predictive maintenance and coordination across multiple sites.

Smart Grids: 60-70% of data processing occurs at the edge or end point for monitoring power consumption, voltage levels and grid stability, and for enabling real-time optimization and fault detection. Another 30-40% involves data transmission for centralized monitoring, analysis of grid performance and coordination of energy generation and consumption.

Healthcare Monitoring: 60-80% of processing is carried out locally with data from wearable devices or medical sensors for real-time health monitoring and anomaly detection. The rest involves data transmission to healthcare providers or cloud-based systems for long-term analysis, personalized treatment recommendations and population health management.

Why use micro data centers for your edge computing needs?

Edge data centers are built in different localities, each with its own requirements in terms of latency, security and types of service (Fig. 1). We are developing innovative Edge Micro Data Centers (EMDC's) that can be installed at the on-premise edge and network edge, and are optimized for the growing range of edge applications mentioned earlier and complement larger scale off-premise edge data centers. We call this infrastructure Powerful Edge as a Service (PEaaS) and is based on four technology pillars:

1. Highly compact, powerful, energy efficient, portable/mobile edge micro data centers (EMDCs) capable of locally processing big data and AI

2. Cloud services specifically designed for the edge with built-in intelligence for efficient data processing and security optimization

3. Data spaces that include governance and monetization services for enabling the training of data and AI models

4. Cognitive services platforms that train and deploy AI models and engender transparency and trust

Let's compare hyperscale and distant cloud providers with innovative edge cloud providers such as HIRO-MicroDataCenters. Distant cloud providers offer low threshold, low cost experimentation to customers, going through digital transformation and adopting microservices on virtualized IT equipment.

However, once a customer starts scaling up their applications and data volumes and implements increasingly more complex applications – or defines criteria such as availability, latency and security – costs tend to rise very quickly. At that point the customer has invested substantially in understanding and optimizing the proprietary service structure of this cloud provider, that migrating to another cloud provider has become less attractive. More recently, in order to capture the edge computing market and strengthen the customer lock-in, the large cloud service providers started offering on-premise solutions.

Hardware innovation and the omnipresence of the open-source software community enabled us to offer PEaaS as a cost efficient option for enterprises that require a fast turn-around time with large data volumes. PEaaS allows the customer to adopt an 'edge first strategy' meaning, cost-efficiently processes the majority of all the data and AI on-premise, and on top of that, where needed, freely migrate applications and data across off-premise cloud providers.

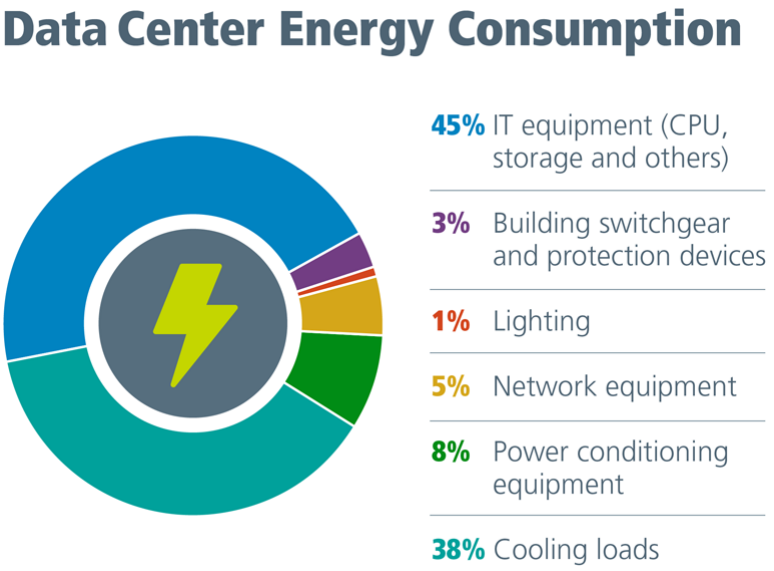
Similar to hyperscale cloud services, PEaaS also has a low financial entry threshold. The difference is that, as customers begin scaling their applications and data volume, the costs are transparent and can easily be extrapolated from the point of entry-level set-up. This means that early cloud adopters that scaled their cloud workloads, only to experience fast rising costs and the pain of vendor lock-in, are now able to better reassess their cloud strategies. Given the benefits, many have concluded that they want an edge-first alternative that will give them more predictability and control over their cloud costs, which can save millions on an annual basis.

How are PEaaS and EMDCs having an impact in today's edge environment?

The most challenging EMDC applications will have big data and AI feeding into decision support systems that are directly affecting human life. At HIRO, we currently are working on two healthcare demonstration projects. The first uses genomic and health data analytics to feed AI models used in decision support systems for cardiovascular healthcare, including an AI model that can be used to predict when a patient might experience a stroke. The second project uses MRI data and real-time ultrasound data to create augmented imaging for surgeons while they perform brain surgery. Both projects promise higher and faster accuracy in professional decision making in order to save human lives and reduce suffering.

EMDCs are closing the power gap over conventional cloud data centers

Let’s zoom out a bit to address data center power efficiency. The data center industry has pledged to become carbon neutral by 2030. This is necessary because global data center energy consumption of 460TWh accounted for two percent of all global electricity usage in 2022. By 2026, that is expected to rise to somewhere between 650TWh and 1,050TWh – an increase equivalent to the entire power consumption of Sweden at the lowest end of the scale, and Germany at the highest.



Click image to enlarge

Figure 2: Data Center Energy Consumption

The data center industry invests in improving their Power Utilization Efficiency (PUE) through (Fig. 2): -Raising utilization levels and efficiency of IT equipment, -More efficient cooling technologies, -Improved power conditioning. While this brings down the PUE of hyperscale data centers to PUE 1.2, the reality is that the efficiency improvements of the average data center over the past four to eight years have flattened out to PUE 1.58. Even worse today’s smaller edge data centers are very energy inefficient, with a PUE rating of 2.0 or more. Since in many cases, the energy footprint of a small on-premise data center is relatively small compared to the overall energy consumption of the enterprise, there is no incentive to improve edge data center PUE ratings.

With edge computing CAGR forecasted to grow twice as fast as cloud computing, HIRO has invested heavily to close the power differential between cloud and edge data centers by developing their EMDCs. To name a few of our innovations, we are using; - Small form factor industry standard PCB’s and state of the art energy conversion modules to minimize the power losses , -AI algorithms to optimize our services and hardware utilization levels, -Cooling technology with a 1.03 PUE rating that even far exceeds hyperscale data centers.

To achieve this milestone, HIRO partnered with leading researchers and technology providers to overcome the many limitations of current PEaaS-based hardware and software technologies.

HIRO EMDCs, for example, deliver highly compact and efficient power conversion using hardware that runs from $48V_{DC}$ power distribution, instead of $12V_{DC}$. Using high-density, high-efficiency power modules from Vicor Corporation, this higher voltage reduces I^2R losses across the power delivery network and enables solid-state, thermally adept, compact, energy-efficient EMDC designs. With their flexible cooling options, Vicor power modules also offer excellent volumetric power density, which lends itself to renewable energy opportunities.

The edge computing market is predicted to be bigger than the cloud computing market because it delivers a faster (middle layer) relay for data that enables mission-critical, real-time responses in devices and systems. By harnessing AI without incurring a significant power penalty, edge computing is supporting AI natively where

intelligent decisions are facilitated, instead of relying on the cloud.

HIRO-MicroDataCenters