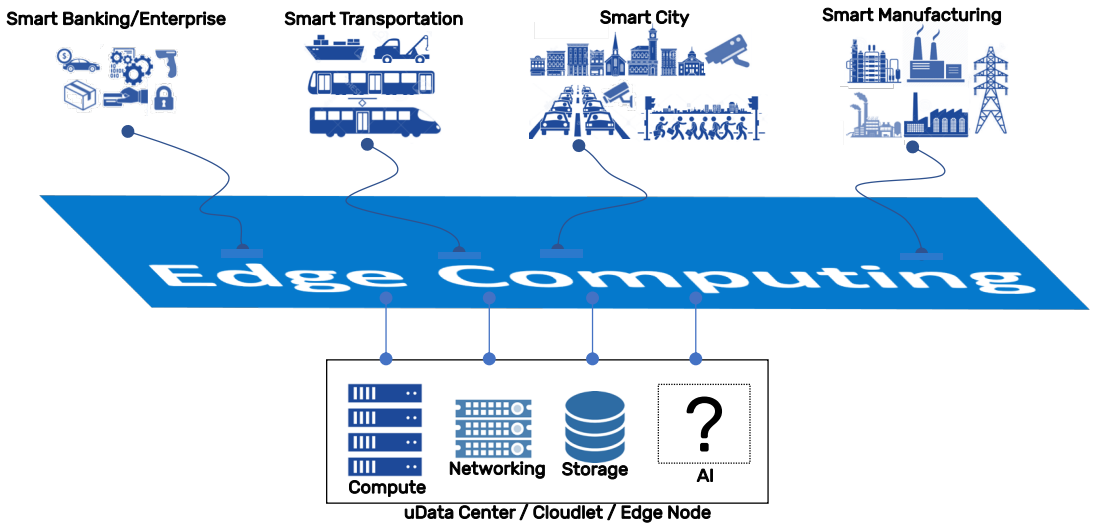
Challenge – Driving AI to the Edge

The proliferation of IoT devices, estimated to reach 20.4 billion in 2020, creates a new data indigestion and processing model. Many of the IoT sensors and actuators need to react in real-time, mandating that computing and actions driven by the data need to be in close proximity to the physical IoT world. Hence the concept of 'Edge Computing' was created. There is no precise definition what is actually 'Edge'. For the purpose of this paper, we will assume that the 'Edge' is a physical location which hosts the minimal storage, compute and networking in the vicinity of the IoT sensors and actuators. An example would be a micro datacenter or cloudlet, physically located in a communication room in a skyscraper, a factory, a macro base station, or a shopping center hub.

The uniqueness of Edge deployment is that it needs to satisfy stringent cost, power, and size metrics, while providing multi-tenancy for applications serving tens of thousands of IoT devices (or more), all of which can vary in traffic, storage and processing loads. Unlike the cloud, where statistics of large numbers create the ability to aggregate demand and sectionalize it, the edge must cater to more spike-like behavior and heterogeneous loads.

The advent of AI and Machine Learning (ML) have created the potential for new business models and application spaces based on extracting value from data – particularly from inferences running in real time, deciphering business intelligence from IoT device outputs (e.g. images, texts, audio and other signals). The need for real-time behavior and data governance will dictate the move of AI and ML inference to the edge. It is also likely that training on smaller sets of data set will be done at the edge. The challenge is in the implementation. GPUs and other types of AI ASICs, such as tensor processing units (TPUs), were designed for the mass scale of the hyperscale data center. The concepts of batching, scheduling and partitioning are not well-aligned with edge requirements.

Bitfusion's Elastic GPUs and Elastic Partial GPUs can offer an economical and feasible implementation for pushing AI to the Edge.



Why AI at the Edge?

While many AI workloads today run in the public cloud, there are clear economic benefits in the ability to run machine learning inference algorithms, and potentially limited training, right at the edge.

1) The case for edge inference is easily understood. Take for example a smart city in which hundreds of surveillance cameras need inferences decoded out of images in real time. Applying the ML model and running inference at the edge will create a true real-time response. It also will not subject the data (e.g. images) to data governance compromise – the inference is done completely at the edge with no need for the data to be shared with a regional or public cloud.

2) Similar benefits apply to voice recognition and for multiple sensor outputs, for example in a smart factory where sensor data such as material fatigue signals, temperature variations, etc. can be acted upon in real-time without the latency caused by a round trip to the public cloud. Companies cannot afford to wait to react to such mission-critical data.

3) Although training typically assumes large data sets needing public cloud deployment, there are many instances of localized information

(secure or very unique, such as factory data sets) which are better hosted and located at the edge. Re-training in this case, to fine tune the model, can happen in the cloud based on functionality demand.

Much like storage, compute and networking – which are driven to be hosted at the edge – the next wave of evolution will provide ML hardware capability at the edge when needed. Of course, there will not be a wholesale shift of AI from public cloud to the edge, but there will be numerous cases where AI should be executed at the data source and in proximity to the IoT devices and other sensors – whether for security, data governance, mission criticality or other reasons.

Migrating AI from the public cloud to the edge poses serious challenges in terms of technology, operations, and expenses. As the mainstream vehicles to execute ML workloads, high-end GPUs are designed to be hosted in the data center – with clusters, size, power, and budget of a hyperscale deployment. To create the fit of GPU hardware to the edge requires a new hypervisor technology platform, one such as Bitfusion FlexDirect.

Bitfusion FlexDirect Enables AI at the Edge

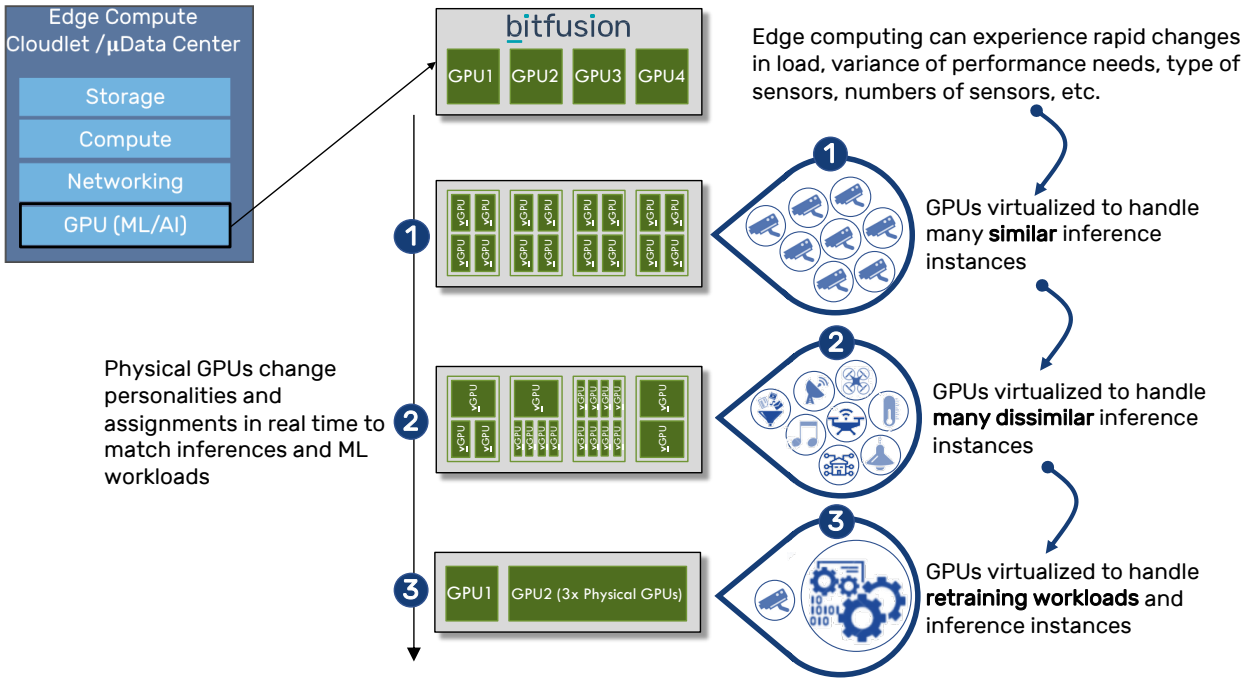
Bitfusion FlexDirect is based on an elastic and virtual GPU architecture that is an exceptional fit for edge computing:

4) FlexDirect can sub-divide a physical GPU which was originally designed for data center implementation into mini-elements of any size to fit the transient workloads of IoT edge devices.

5) FlexDirect can allocate virtual GPUs in real-time without interrupting the real-time operations of neighboring vGPUs which reside on the same physical GPU. This is extremely important at the edge since workloads and inferences are not in synch, and one set of IoT devices cannot interfere with others.

6) FlexDirect can change the profiles of vGPUs in real time to provide ultimate flexibility for the edge. A physical GPU can be sliced into 20 vGPUs in a one-time slice, then into 10 parts and then into slices of vGPUs of varying sizes as needed.

7) FlexDirect operates with any compute Hypervisor, such as VMware and KVM; hence it meshes well with compute virtual machines.



Edge Elasticity

The concepts of elasticity, virtual machines, storage and networking are critical for the edge. There can be sudden changes of load and numbers of tenants, as well as response time / processing / power tradeoffs. Hence each hardware resource at the edge must be elastic and virtual. Bitfusion delivers this capability with its core technology. With Bitfusion, the four different resources: compute, storage, networking, and GPUs are all connected together to track the bursty and unpredictable nature of the edge. In addition, it is important to leverage any type of software, hypervisor or container platform in the edge. Bitfusion FlexDirect as a user space platform natively supports the full range of containers, hypervisors and operating systems. Welcome to the AI Edge, where virtualization and elasticity are critical.

Bitfusion delivers real-time partial GPUs with **flexible** partitioning; each vGPU is isolated from the others