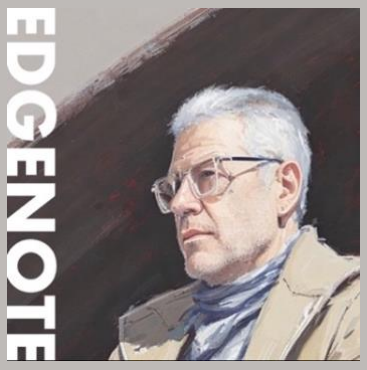



# Fasten Your Seatbelts - Generative AI is Accelerating the Edge

**Pete Bernard – EDGECELSIOR**

***more EDGENOTES are available at edgecelsior.com/edgenotes***

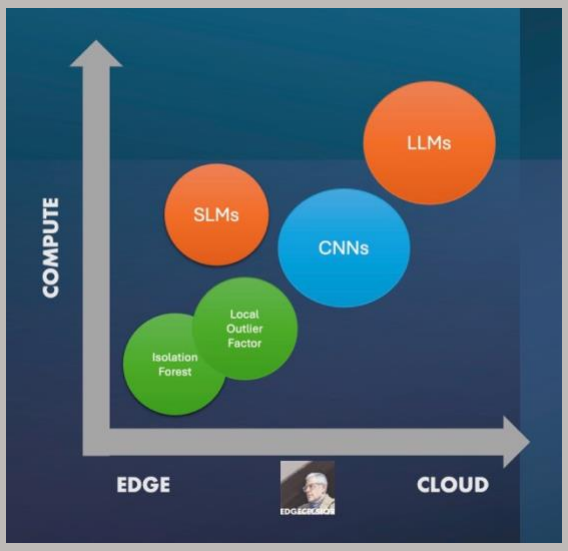
***KEY THEME – GENERATIVE AI WILL ACCELERATE AND ENHANCE THE DEVELOPMENT AND DEPLOYMENT OF EDGE AI SCENARIOS; COMPANIES SHOULD DEVELOP STORIES THAT COMBINE AI TYPES TO CREATE COMPETITIVE ADVANTAGE . . .***

This EDGENOTE is designed to educate, inform, and hopefully inspire - from an unbiased perspective – investments in new forms of AI that are cascading over us like waves at high tide. By taking a story -based perspective that puts the user at the center, we can begin to understand the power of Generative AI beyond the headlines and how it can combine with other "traditional" AI model types to create powerful new capabilities from cloud to edge and across a range of industries.

## AI Model Types and Where They Run

Artificial Intelligence has been around for quite some time. In fact, the term was first coined at a conference at Dartmouth College in 1956. The study of creating "neural networks" and ways of modeling human intelligence have been around for decades.

But, as we know, back in the 50s, 60s and frankly up until the last decade, the ability for semiconductors, networks, and our compute infrastructure to keep up with the massive amounts of training and model execution needed for AI simply didn't exist. And once the compute power did cross that threshold – BOOM – the flood gates were open.




We also need to understand that there are different types of AI models and different places where those models can run – from cloud to edge.

One well understood AI model architecture is called the CNN, or convolutional neural network. This type of architecture is a type of deep learning algorithm that is

particularly well-suited for image recognition and processing tasks. CNNs can automatically learn hierarchical feature representations from raw input images, which makes them more efficient and accurate than traditional machine learning algorithms for computer vision tasks. This is the kind of model used for license plate recognition at toll booths, image analysis in healthcare for x-rays and MRIs, or face recognition, people counting and other surveillance scenarios.

A CNN based model is typically run on edge equipment like cameras, in a vehicle, or on a gateway that is being fed images by multiple streams. However, you could run this model in the cloud against a large database of images or videos after the fact on a batch of video data or escalate a workload to the cloud from edge if it's deemed too complex (such as a license plate recognition that does not give a confident result).

Another category of AI models is for Anomaly Detection, such as Isolation Forest, Local Outlier factor or other options. In these scenarios data is observed for differences that could indicate an impending mechanical failure, or an incident where glass is broken, or gunshots detected, etc.


These models are almost exclusively run on edge equipment like sensors or lightweight compute boxes, although you could run them in a batch processing style on the cloud over a large amount of data that has been collected over a long period of time.

The "hot" AI model category of Generative AI that everyone (including AI!) is talking about actually consists of several possible model types that include Large Language Models, or LLMs that include model frameworks such as Transformers, and General Adversarial Networks (GANs).

These have been running almost exclusively in the cloud due to their extreme requirements of compute horsepower (running math models) and RAM/memory requirements. A good area to keep an eye on here are SLMs, or Small Language Models, like Google Gemini, Meta's LLAMA 2 or Microsoft Phi, which have LLM capabilities but are designed to run in more resource constrained environments, like on Phones, PCs or commercial edge devices.

## Real-world Generative AI + Edge AI Scenarios

Let's get to some real-world scenarios where **Edge AI combines with GenAI**.

**Jim's Washing Machine**

Jim sees a notification on his phone - his washing machine is trying to reach him. A high current draw pattern has been detected – this typically means that the washer has been overloaded repeatedly by Jim and the heavy usage is wearing out the motor (whoops!). This anomaly detection will help Jim do some pre-emptive maintenance, but what's the next step?

In the notification there is a link to start a chat session with a washing machine representative that is powered by Generative AI. The "representative" can converse with Jim and set up an appointment for a motor swap – at a substantial discount and with barely any downtime in Jim's busy household laundry cycle!

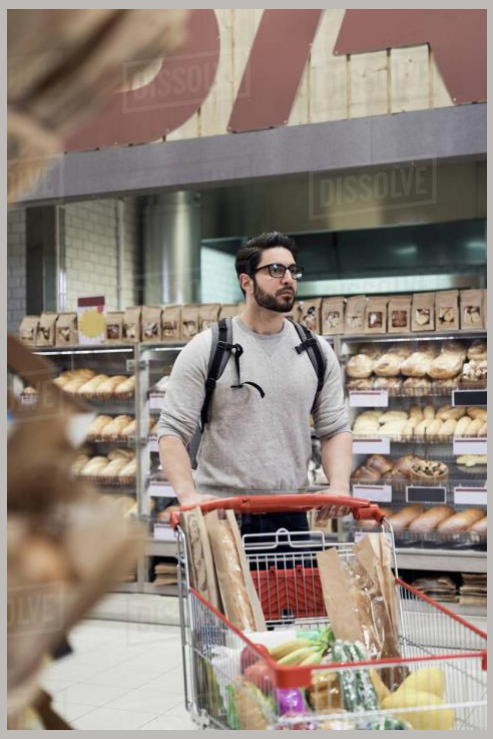
**Kathy's Coat**

Kathy has been dreaming about a long winter coat. Sort of dark grey, with big puffy buttons and large collar. She heads to the web and uses GenAI to create the coat of her dreams – she even whips up a few images of her wearing the coat!

She loads these images into her Poshmark app. Poshmark has tens of thousands of clothing items for sale – used, vintage or just resold. Poshmark uses an AI model to match Kathy's dream coat with several real coats for sale on Poshmark – in her size!

**Paul's Family Dinner**

*(editor's note: this is lifted directly from Art Miller's excellent blog about Qualcomm at NRF 2024 – read the while piece here).*



Imagine, if you will, getting off a busy workday — you haven't even begun to think about what's for dinner. Instead of contemplating over what to make and defaulting to a drive-thru, generative AI helps you decide, becoming your personal shopping assistant. Knowing that Italian is your favorite cuisine, for example, your AI assistant recommends a quick and easy bolognaise recipe, provides a custom shopping list that includes exactly what items you'll need, and even tells you where in the store to find them.

When you're in the store, you might be asking for alternative ingredients or other complex questions. In-store interactive AI customer service kiosks provide customers with the answers they're looking for, while helping retailers learn more about their customers to better serve them in the future.

**Laura's AI PC (or AI Phone?)**

*(editor's note: Although today's AI PCs are starting with cloud based Co-Pilot and apps that leverage DirectML – there is a future where this platform actually becomes a "pro-active" AI PC. Samsung is the closest to this with the announcement in January 2024 of their S24 phones and platform – and I'm excited to see where they go with it.)*

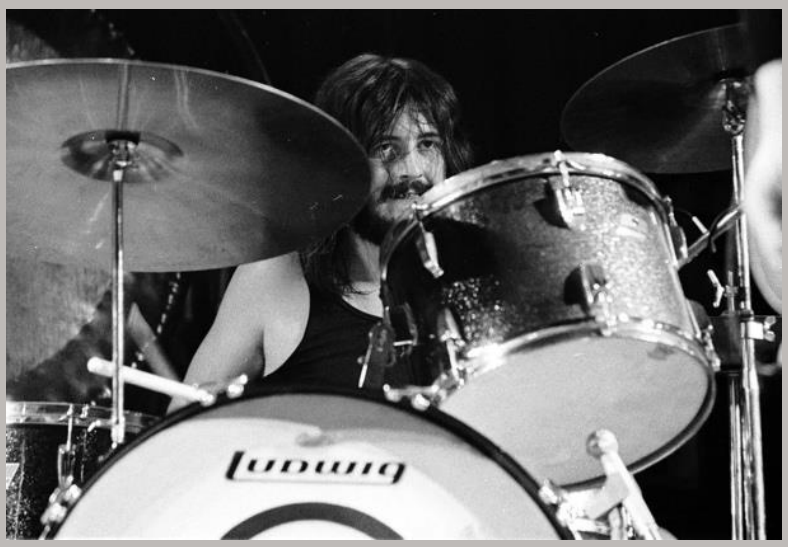
Laura is in crunch time getting this paper done – their boss has been breathing down their neck all week.

They fire up their Co-Pilot agent and while they are hacking away at the document, they asks it to go out and develop some content they can integrate that compares various wireless connectivity standards and their trend of adoption over the next five years, where the company should place its bets

and why…meanwhile, Laura has their Co-Pilot agent read what they have written so far to give them editorial feedback and provide some local illustrations on key points.

Laura finishes up on a paper with great points for their strategy as well as great footnotes to the source data.

**Pat's Gig**




Pat has finally landed a gig at the Yardarm, a run-down watering hole with a great stage, a great sound system and a great reputation.
Unfortunately, the drummer has bailed (again) and so it's going to be up to Pat and the rest of the band – the show must go on!

They take the stage and call out the next tune – "The Rover" – the Generative AI model running in the edge drum machine thumps out the John Bonham intro and they're off to the races!

Ok, that last one was a cheat. We WILL see Generative AI models running on edge equipment, especially some of the "smaller" (relative) models coming out of research like Microsoft Phi, Google Gemini and Meta's LLAMA 2. Stay tuned.

## AI on the Edge – Where is the Friction?

AI models are running on edge equipment all around you – at your healthcare provider, at your favorite retailer, in your washing machine and dryer and probably in your car.

Despite their increasing use, the development, deployment, and maintenance of AI models on edge equipment is challenging.

- **Models need to be trained with a large amount of very specific data that may be difficult to obtain**
- **Each semiconductor partner has toolchains to optimize models and their own model "zoos"**
- **The operations or "MLOps" to deploy and update models on edge devices is usually very bespoke and customized**
- **There is little coherency between AI models running on the edge and running in the cloud**

Companies like NVIDIA, Intel and AMD have an advantage here as they provide semiconductor AI acceleration at both ends – the data center AND edge equipment. However, the lighter edge of compute is dominated by devices in the billions of units based on NXP, ST, Renesas, Qualcomm and others that have virtually no footprint inside of hyper-scalar data centers.

A number of companies are attacking these issues from different vectors, like Edge Impulse in their "low code no code" model training and deployment across heterogeneous hardware end points, or NVIDIA's Tao effort to harmonize cloud and edge AI models, or the TinyML Foundation to develop new ways to reuse components to speed development and deployment of increasingly sophisticated AI models on tinier and tinier edge equipment.

## By Products of the Edge AI + GenAI Combination

Now that we understand that that there are a wide variety of AI model types, and that these models can run somewhere on the edge or on the cloud (or somewhere in between) and we recognize the challenges creating friction in the development and deployment of edge AI, what are three key areas to focus on to leverage this new energy and enthusiasm for Generative AI to benefit Edge AI?

### 1) Synthetic Training

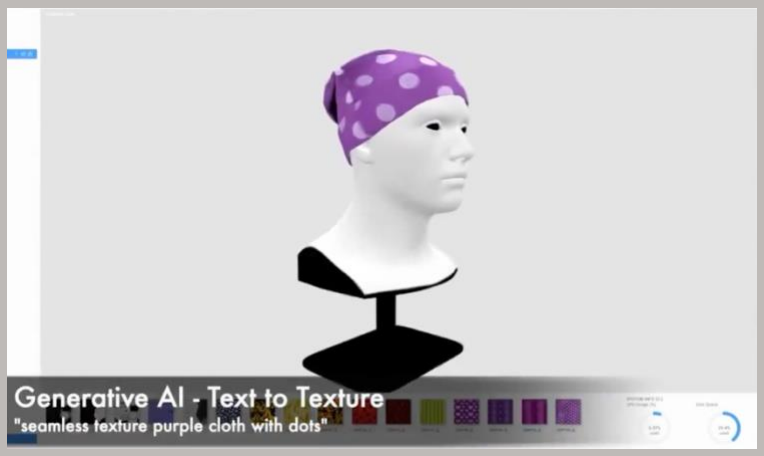



*Figure 1 - Image courtesy of SyntheticAIData*

Although transfer learning and base models are becoming more readily available to give companies a head start in tuning and deploying AI models, training AI models for the edge, such as a CNN, requires a large number of images. For example, to train a camera at self-checkout to recognize a head of broccoli at reasonable accuracy could take north of 15,000 images in the training set and 3,000 images in the validation and test sets.[1] That's a lot of broccoli.

As we know, the accuracy of self-checkout needs to be quite high to meet customer expectations. Multiply that head of broccoli by the range if vegetables at your local market and you quickly reach unmanageable numbers.

Although you shouldn't train your grocery store CNN on Generative AI data (hallucinating broccoli?), you CAN and SHOULD train it on synthetic data, and then use Generative AI to iterate on that synthetic data to create unlimited variations to increase your model accuracy. With use of realistic 3D models, you can generate an unlimited synthetic dataset. Synthetic data not only increases accuracy but also accelerates model training. You won't spend so much time on data collection and annotation. SyntheticAIData[2] is one company out of Denmark that is doing just that with a robust range of tools and services for not just vegetables but also clothing and other goods – here's a good demo on LinkedIn.

[2] https://syntheticaidata.com/

Extend this approach to agriculture for crop yields, healthcare for medical imaging, or smart cities for training on traffic patterns and the benefits of synthetic AI model training with generative AI can be profound.

**2) Container-based ML Ops to The Edge**

This one is a bit farther out, but we're seeing some of it in Cortex-A based systems and those running x86 architecture.

The idea borrows from a growing move to adopt Cloud DevOps across solutions, from cloud to edge. The ability to virtualize and containerize workloads is increasing incredible productivity to move workloads from server to server and even cloud to cloud.

Now, we are seeing several companies innovate to bring that AI workload orchestration to edge devices – the ones outside the data center that are either directly connected to the cloud or connected to interim gateways that then are cloud-connected.

A few examples:

**1. NVIDIA:** NVIDIA EGX[3] is a cloud-native, software-defined platform designed to make large-scale hybrid-cloud and edge operations possible and efficient. Within the platform is the EGX stack, which includes an NVIDIA driver, Kubernetes plug-in, NVIDIA container runtime and GPU monitoring tools, delivered through the NVIDIA GPU Operator


*Figure 2 - NVIDIA's EGX Stack*

**2. Cisco:** Cisco has developed a lightweight distribution of Kubernetes called K3s, which is optimized to run on ARM devices. This enables the deployment, scaling, and management of containerized AI inference applications on edge devices.

[3] NVIDIA EGX Simplifies AI Deployments with Enterprise Kubernetes .... https://blogs.nvidia.com/blog/ai-edge-deployments-kubernetes/.

**3. Microsoft:** Microsoft has integrated Azure IoT Edge with Kubernetes to enable the deployment and running of AI, Azure services, and custom logic directly on the cluster. Azure IoT Operations[4] is a good example of how this can be deployed.

**4. Google:** Google Kubernetes Engine (GKE) enables the running of containerized AI workloads at scale on the edge network.

**5. Zededa:** Zededa[5] is a cloud-native edge computing platform that enables the deployment and management of containerized applications and services on edge devices. Zededa provides a secure and scalable infrastructure for running AI workloads on the edge, which can help reduce latency and improve performance. Zededa has also developed an open-source orchestration engine called EVE-OS, which is designed to simplify the deployment and management of edge applications.

Keep in mind - the key point with AI workload orchestration is that the container and/or driver is written in such a way that that workload code can actually reach the hardware accelerated silicon "underneath" – the virtualization implementation needs to support this kind of hardware access.

## Now Is the Time to Focus On Holistic Stories

The past year or so has seen a resurgence of interest and investment AI as we have reached another threshold in compute capabilities and unleashed generative AI capabilities onto the public. There will be no slowing down as the models get larger, the systems get faster and our appetite grows. Hopefully, in all of this excitement, we as a community lay down some rational containment strategies to keep this powerful technology in check.

Now as the Generative AI part of the AI equation "comes to life" we are faced with incredible opportunities to think holistically about Synthetic Training, Cloud Dev MLOps and real-world valuable scenarios that leverage AI from the cloud to the edge. There is no better time for technology providers, solution providers and businesses to align their product strategies, partner strategies and marketing to harness the AI edge, the AI cloud and everything in between.

[4] Azure IoT Operations demo: https://youtu.be/cw9nhPFE_qE
[5] ZEDEDA Help Center. https://help.zededa.com/hc/en-us.