Assessment 1: Smart Life and Edge Cloud Technologies

Watson Ndethi

2025-02-09

Table of contents

1 Introduction

1.1 Background on smart life technologies and edge cloud computing

1.1.1 Smart Life

Smart life is enabled by the intersection of the web of smart sensor embedded devices that are connected with each other and the internet **(Internet of Things) powering greater operational conveniences, increased energy efficiency, improved safety at the home.

Take for instance a smart security light that will automatically dim itself at dawn, or send a notification to a home owner when its motion sensor is triggered by movement at night.

1.1.2 Edge Cloud Computing

As the term edge suggests, this paradigm involves placing cloud resources in close proximity to where data is generated and where end users are located. The main motivation for this architecture being reducted data latency (as data is transmitted via a shorter route) and maintaining high availability of compute resources.

This is especially intsturmental in processing sensor data for IoT and Smart Life applications.

2 Exercise 1: Smart Life Technologies at LifeX

2.1 Investing in Smart Life Technologies

2.1.1 Raising Kids Smart, An Opporuntity for LifeX

Central to LifeX' mission is augmenting life through smart technologies. The X in LifeX points to the multiplicative potential of technology to enhance quality of life. and nowhere is this more needed than in the increasingly busy lives of middle class parents raising toddlers.

Now more than ever, owing to dire economic times both parents are working full time jobs and with more corporations rethinking their work from home policies - increasingly pushing for a return to full time office work, young middle class parents are turning to full time nannies to take care of their toddlers.

While the human touch of a good nanny is irreplaceable, nannies may not effectively juggle playtime and keeping track of a baby's vitals by the minute. Thus parents could do with a technological intervention to actively monitor their kid's wellbeing while they are away. Enter the LifeX Smart AI Baby Monitor.

2.1.2 Lifex Smart AI Baby Monitor

Lifex Smart AI Baby Monitor is a dual device system comprising of a smart camera and a smart wristband. The smart camera mountable on a standard baby crib records footage and audio of the baby while the wearable, monitors , in the similar fashion to smart watches, the baby's vitals - heart rate, temperature, oxygen levels,etc. The two devices stay interconnected via a low power bluetooth connection. They send a stream of multimodal data(audio, visual, text) to a hosted ai model that runs analytics on the data and presents metrics to on a dashboard accessible via an app on the parents smartphone. It can also integrate with existing smart assistants present in the home to send local notifications to the nanny in case of vitals metrics threshold breaches. For instance it can trigger dimming of the smart bulb in the childs room, play lullabies on the Amazon Alexa when the baby is snoozing or crying. The parent can also talk to While the

2.1.3 Finacial Model

The Lifex Smart AI Baby Monitor starter hardware(camera and wristband) will be sold as a bundle for a one time fee of $200. Henceforth users will pay a $30 a month subscription to use the app and AI service. Partnerships with health facilities will allow LifeX to subsidize the cost for new parents with the caveat that they will allow potential partners to access anonymized data for research purposes.

2.2 Virtualization and Cloud Computing - Foundational Pillars of Smart Life Technologies

2.2.1 Virtualization

By using software to create simulated hardware or software components, we can ensure more efficient utilization of resources used my smart life tech. Using LifeX' on prem servers as an example, using virtualization locally through VMWare would allow LifeX to run multiple

instances of the Smart Baby AI monitor across different Operating Systems without having to purchase extra hardware.

2.2.2 Cloud Computing

Cloud Computing provides access to a shared pool of compute resources. By leveraging the five key characteristics of cloud computing as defined by NIST: On-demand self-service cuts out proxies between LifeX and the compute resources provider - bolstering unfettered and immediate access Broad network access means connected IOT devices powering smart life tech anywhere in the world can acccess compute resources as long as they are so configured. Through resource pooling, economies of scale enter the equation allowing smaller players like LifeX to essentially afford compute. Rapid elasticity is key to flexible and instant scaling of compute needs depending on demand of resources - offloading the on prem scaling timing headache. That cloud resources are a measured service, means services consumed are billed based on utilization of resources which supports pragmatic decision making around budgeting and pricing of products and services LifeX offfers. In addition, As Big Tech corps retain the ascendancy of the most powerful compute hardware (e.g NVidia has dominated the GPU market for a while) , cloud computing offers smaller players in the market more affordable access to the power of High Performance Computing which they would otherwise forfeit because of the high acquisition costs.

2.3 Centralized AI vs. Edge AI

2.3.1 Centralized AI

The centralized AI idea involves sending data from smart devices to a central location for analysis and processing. While network intensive and bearing higher latencies, processing in the cloud has the undeniable benefit where thin smart devices with little processing power can leverage the power of more capable cloud servers. In the case of Lifex' AI Baby Monitor, sending its data to a centralized ai would allow the AI model to also collect data and learn from multiple monitors, resulting in a richer dataset. In terms of design it would allow a simpler design for the two part monitor hardware as it would not need to be as powerful as it would be if it were to process the data locally.

2.3.2 Edge AI

Edge AI in contrast involves processing data on device or closer to where the data is generated as much as possible. In the case of our baby monitor, the wristband could be designed to send data to the monitor hardware via a low powered bluetooth connection or some form of wifi connection. The chip in the monitor would then process the data locally - with new advancements in small language models and distillation techniques (where a smaller model

can learn utilizing the outputs of a larger model) it is possible to run a 1-2B parameter model on device) and send only the more critical data to the cloud for further processing. This would reduce the amount of data sent to the cloud saving on bandwidth and reduce the general latency of the system. Additionally containing the data on device enhances the overall security of a system so vitally in need of privacy as a baby monitor.

2.3.3 A Hybrid Approach towards Sustainable Smart Life Technologies

In reality, an optimal approach would be a combination of both centralized and edge AI working in tandem. Edge AI through a small language model deployed on device would process the data locally and send only the most critical data to the centralized ai in the cloud. A feedback mechanism or a distillation process would then allow the edge ai to learn from the outputs of the centralized ai model - ideally a larger billion parameter model. This would continuously improve the edge model and the centralized model. This hybrid approach not only optimizes resource utilization (minimizing round trip comms with the cloud, saves bandwidth and reduces api costs) and reduces latency (especially between device and app - increased customer satisfaction , leading to more positive reviews)but also fosters long-term growth by enabling continuous learning , ensuring the system and models remain up to date. Leadership in the baby smart tech and ai product space will lean heavily on first mover advantage on the dataset front - the more useful data the system collects the more accurate the models become and the more value the system will provide to the parents through the app.

2.4 Recommendations for LifeX Product Portfolio Growth

While the Smart AI Baby Monitor is a great first stub at a new emerging market (the burgeoning smart baby tech market), LifeX should not rest on its laurels for the legacy markets it has already captured for example the general smart home tech market. I would advocate for the below strategies to push growth in the ai age for the Smart AI Baby Monitor and the entire LifeX product portfolio:

2.4.1 A Tiered Subscription Model

Cloud is not cheap - the more api calls you make the more you pay. To sustainably counter the costs of running the hybrid ai model (costs largely will be incurred on the cloud side) while recovering the costs of hardware production and maintenance, LifeX ought to consider a tiered subscription model for their SaaS offering. A basic tier, a premium tier and an enterprise tier would allow LifeX to cater to different customer segments with increasing needs and budgets respectively. The differentiator across the tiers would be the inference needs of the customer(the more advanced their analytics and ai assistance needs the more they would pay. This Pay For What You Use model would effectively hedge LifeX against the risk of overprovisioning resources(bleed cash) for customers who do not need them.

3 Exercise 2: Edge Cloud Computing

3.1 Cloud vs Edge

3.1.1 Defining Characteristics and Key Differences

3.1.1.1 Cloud Computing

Cloud computing involves utilizing a web of servers remotely hosted on the internet to store, manage and process data. These servers typically will be installed in data centers in remote locations far from end-users they serve.(as opposed to in edge computing) These servers provide storage,compute and networking capabilities to end users on demand.

One common use case of cloud computing is in the provision of Software as a Service (SaaS) products. For instance, Google Docs, a cloud based word processor, allows users to create, edit and share documents online [@google-docs]. The processing of the document is done on Google's servers and the user only needs a web browser to access the service.

3.1.1.2 Edge Computing

On the contrary, edge computing brings cloud resourcees in close proximity to where the data is being generated and where end users are located rather than trasmitting it to a distant cloud server. This is particularly useful in situations where data latency is a concern. For example, in the case of a self driving vehicle, the data from the vehicle's myriad sensors and cameras needs to be processed in real time to facilitate split second turn by turn navigation of the vehicle. To enable this, for instance on a Tesla sensor data is processed on the onboard computer system on chip (SoC) rather than being sent to a remote cloud server for processing. This reduces the latency of the system and ensures the vehicle can make decisions in real time.

3.2 Private Cloud vs Public Cloud

To be able to determine which of the two architectures better fits the needs of EdgeX, we need to first understand the defining characteristics of each architecture and then evaluate how these characteristics align with the needs of EdgeX.

3.2.1 Public Cloud Characteristics and Considerations

A public cloud architecture involves the use of cloud resources that are delivered over the internet and shared across multiple organizations and users. Providers of public cloud resources include Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). Some of the key defining characteristics of public cloud include: - Shared resources: Resources are shared amongst many users, leading to reduced costs - Low upfront costs: Users pay only for what they use, making it cost-effective - Low maintenance: The cloud provider takes care of maintenance and updates

3.2.2 Private Cloud Characteristics and Considerations

On the contrary a private cloud architecture entails the use of cloud resources dedicated to a single organisation. The cloud resources are not shared with other organisations. Some of the key defining characteristics of private cloud include: - Dedicated resources: Cloud resources are dedicated to a single organisation - High upfront costs: Setting up a private cloud can be expensive - More secure: Data is not shared with other organisations, leading to increased security

3.2.3 EdgeX Fit Evaluation

Seeing that the EdgeX brief expressly states that the Cloud Infrastructure Engineer should design towards reduced data latency and high availability, it is abundantly clear a public cloud architecture on the edge would be the better option. To achieve high availability requires a robust network of servers that can be accessed globally from anywhere. Public cloud providers e.g AWS have data centres in multiple regions globally, ensuring one - unprecedented redundancy and two - low latency because of proximity to regional users considering the spread of the data centers. This recommendation is made with the assumption that EdgeX is a small to medium sized enterprise with a limited budget for compute and seeing that their utmost priorities can be met by a public cloud architecure,spending on a private cloud might be difficult to justify at this juncture.

3.3 Real-time IoT Analytics

3.3.1 IoT and Edge Computing - A Match Made in Heaven

First, a definition of the terms is in order. The Internet of Things(IoT) can be referred to as a web of interconnected smart devices that communicate with each other and the internet. Edge computing as earlier explained involves placing cloud resources(compute, storage networking etc) in closer proximity to where the data is being generated with an aim to reduce data latency. To effectively process the data generated by IoT devices in real time, we need to leverage the

power of the compute resources available on the cloud, but closer to the IoT devices (so as to stay realtime) - hence the need for edge computing. This is especially important for EdgeX as they are in the business of providing real time analytics to their customers. Services like AWS IoT Greengrass [@aws-greengrass] and AWS Local Zones [@aws-local-zones] could be instrumental in achieving this real-time analytics for EdgeX.

3.4 Recommendation for EdgeX

Jensen Huang, the CEO of Nvidia, recently said that the next wave of AI is physical AI - AI that interacts with the physical world [@huang-physical-ai]. T That Jensen is a visionary figure in the AI domain is undeniable, the success Nvidia has had in the AI hardware space and the up and up trajectory of the Nvidia stock is testament to his foresight to a fair degree. As this robotics future unfolds, we can bet that big tech and highly funded startups will continue to dominate the hardware aspects of the space. A bigger playing ground for smaller players like EdgeX will be in the robotics software space. First mover advantage will be key in cementing EdgeX position as a leader in the robotics software space. To this end, I would recommend that EdgeX set aside an R&D budget to develop a proprietary realtime analytics platform that will allow smart home tech devices to communicate with humanoids and other robots. Imagine if EdgeX could provide a platform that would allow a humanoid to act as a local hub for all smart home devices in a home, effectively acting as an edge server for all the smart devices in the home but now with agency to act on realtime data from the smart home tech. This would be a game changer in the smart home tech space and would position EdgeX as a leader in the robotics software space , attracting more investment and partnerships with big tech and other robotics startups.

4 References

1. Google Docs. (n.d.). Retrieved from https://www.google.com/docs/about/
2. AWS IoT Greengrass. (n.d.). Retrieved from https://aws.amazon.com/greengrass/
3. AWS Local Zones. (n.d.). Retrieved from https://aws.amazon.com/local-zones/
4. Huang, J. (2022). Physical AI. Nvidia. Retrieved from https://www.nvidia.com/en-us/ai-physical-ai/