Synadia

Powering AI at the edge with NATS.io

By David Gee

V1.0

Introduction

Edge computing and services are providing businesses across the globe a competitive advantage. The edge provides means of connecting both information technology (IT), operational technology (OT) and manufacturing processes to systems that provide artificial intelligence-enabled functionality, predominantly machine learning and deep learning-centric use cases. In turn these systems provide enhancements, optimisations and new insights, resulting in higher revenues for businesses adopting this wave of technology. As AI technology becomes pervasive, the businesses that leverage the vast amounts of data generated by modern business operations, will be the most impactful and profitable.

This document explores the state of the edge in 2023, presents our view of the problem space and describes how to solve the requirements with technology from Synadia leveraging NATS.io. So why is there so much interest in the edge today?

> **83% believe that edge computing will be essential to remaining competitive in the future.**
>
> *- Ramalingam.R, Tung.Teresa. (2023). Leading with Edge Computing.* ***Accenture***

This paper also discusses how the edge technologies can be made simpler and cheaper to operate at scale, all through adopting a unified approach to application communication. The edge does not require a standalone set of technologies, and to prevent edge projects having lackluster outcomes, this document describes a landscape where communicating with datacenter services is the same as communicating with a manufacturing machine or a street sensor. It's time to unify disparate application communications and realize the power of the edge.

Resurgence of AI

Every decade or two, artificial intelligence (AI) raises its head and we all get excited about what the future might bring. Unlike other periods in technology in which AI has been popular, AI is here to stay, even after the excitement dies down because of the powerful outcomes for businesses and governments across the globe.

As hardware gets cheaper and faster, AI will become even more pervasive and personalized, becoming a staple part of every task. You already witness these technologies at work in facial recognition systems on your smartphone, in your banking app, on your streaming entertainment platform and even dare I say, your kitchen appliances. Enjoy your self-driving Tesla? Chances are you've already been chauffeured by it. The insights and information that AI systems can yield help organizations maximize their profits across the globe.

The next generation of AI is destined to power smart cities, power grids, vehicles and communications on a scale never before seen. With the correct utilization of AI, the data gathered from every sensor and process can provide unparalleled insight and monetization opportunities.

An Argument for AI at the Edge

Industries have already bought into the idea of software eating their world; everything from raw goods and supply chain management, to work planning, communication network optimisation and defect detection, is software driven.

As an order is placed by a seller, a factory can predict how long the manufacturing process will take and plan appropriate slots in the end-to-end process. Each component is tracked from beginning to assembled product and many of these aspects use AI for everything from cleaning up data, prediction, plant optimisation and even generative design to augment or replace human designed components.

All of this activity requires compute, storage and network connectivity, and the machinery must be interfaced to them. Placing AI systems in a distant public cloud isn't going to satisfy near real-time use and access patterns.

In an ideal world, a manufacturer will have their own on-site data center to cope with the demands, but there are different approaches to the architectures including on-premises, public-cloud and smaller edge-based data centers. Each of these modes presents a set of challenges to connect the hardware and software infrastructure to the models.

AI training systems use copious amounts of computing and storage resources which translates to significant spend on electrical power. In addition, high quality models require lots of high quality data which is generated from additional human and computing processes. This is a time and space problem, and we must make it easy to move data between different spatial locations, with the lowest latencies possible.

Locations for the functional components might be dictated by ownership, which is true in the case of on-premise data centers and cloud, available electrical power and also the work process, which could be executed in an office, factory floor, race track, vehicle or even space. We are in the era of the edge with some specific problems to solve; moving data to the training system, distributing the trained models to edge locations, and reliably handling both inputs to AI systems and signals from their usage.

Defining the edge(s)

From our perspective the edge is more of a concept vs. a specific location. So as we look outside of the current public cloud we see three primary buckets where edge innovation is occurring, inspired by SUSE definition of edge[1].

In general, there are two dimensions these edges include: quantity of compute and storage and latency from a client to the edge location.

Near Edge

This is the first hop away from the corporate data center or public cloud with typically tens to a few hundred devices. Physically, this network location might not be on-site, but to be of use, will be closer in proximity. Devices typically are servers and maybe GPUs for targeted training use cases.

Far Edge

This could be connected to the corporate data center and the near-edge location, it's normally on-site and the farthest point that is connected to the data center but services the needs of co-located services and equipment. This can be hundreds or thousands of devices, including gateway aggregation devices that serve the tiny edge.

Tiny Edge

Operational Technology (OT)[2] like PLC devices and SCADA systems, single board compute systems and custom embedded electronics for machinery make up the tiny-edge. These are sensors, actuators and process control machinery and could also be called the Internet-of-Things (IoT).

[1] SUSE Edge Computing eBook:
https://more.suse.com/rs/937-DCH-261/images/FY21-SUSE-Guide-to-Edge-Computing.pdf

[2] Definition of Operational Technology (OT):
https://www.gartner.com/en/information-technology/glossary/operational-technology-ot

Many of these systems lack a native way to communicate over TCP/IP, but in the last decade have gained WiFi, Bluetooth and low baud rate communication capabilities like LoRaWAN, and may rely on other devices in the tiny edge for data transmission and security.

Data entering this domain is converted to what the tiny-edge system understands, like MQTT, ModBus, OPC-UA, ONViF, CAN, I2C, SPI and more. Data leaving this domain is converted into a common data structure, typically schema-driven in JSON, XML or Protobuf format.

Devices and systems operating in the tiny edge operate at real-time or near real-time speeds and require single millisecond or less latency on the tiny edge network. Whilst the other edge locations demand low latency for responsiveness, this edge is the most critical regarding immediate response times.

AI Edge Constraints

Global power usage for data centers has recently grown from 1% to 3% and is projected to be as high as 4% as we hit 2030[3]. Coincidently as the industry continues to invest in AI systems, the connected device count is set to double from 15 billion devices in 2023 to over 30 billion by 2030[4] and AI systems will increasingly sink data generated by these devices.

As businesses embrace AI for personalized customer experiences and enhanced operations, getting the data from where it’s generated to where it’s needed is one problem with the second being powering the AI training infrastructure. The latter is a simple problem of deploying the technology to the nearest location with power and most basic network connectivity available such as high speed IP routing or dark fiber. Therefore, when we design AI Edge solutions, it is critical to think in two dimensions, which are: Resource Availability and Communication Latency.

Resource Availability

This is the challenge of ensuring resources such as AI training infrastructure and compute systems that run inference systems, are available at the right location to enable the correct user experience.

Physical proximity and networking rarely correlate. Just because things are physically close to each other, it never means they're actually near each other on any network. As networks get faster, we will get closer to the physical limits of working with light and the speed of it. As the use of OT rises, the network is suddenly important to reduce communication latency. That could mean multiple networks, such as short range radio, private dark fiber, satellite or direct internet connectivity from an ISP. To enable the vision of pervasive AI at the edge, the lowest latency path between the data source or sink and network service is of critical importance. Vehicles such as self-driving cars and UAVs can generate important events onboard which are of interest to other systems and as the availability of connectivity varies whilst the vehicle travels, the system must be able to store those events and catch up remote systems when connectivity permits. Whilst the availability of compute resources on these vehicles supports the vehicles primary functions, there are requirements such as fine tuning natural language interfaces in the form of Large Language Models (LLMs) and resource planning, which are all off vehicle.

[3] Data Center power consumption:
https://datacentremagazine.com/articles/efficiency-to-loom-large-for-data-centre-industry-in-2023 and https://www.projectfinance.law/publications/2020/october/powering-data-centers/

[4] Device count for IoT devices:
https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

Communication Latency

Solving network latency and proximity is only one aspect and the very nature of how these networks function end-to-end must be rethought. Today’s networks are fragile, configuration-driven and lack the coherency and dynamism required for successful AI at the Edge deployments. In summary, the systems that reside on the network change faster than the humans that design, operate and troubleshoot the networks.

By unifying the underlying heterogeneity of the transport networks and moving the network intelligence to a layer above, network communications can be transformed into a cooperating system with automatic lowest latency service usage and tightly coupled on-net data systems such as streams, key-value and object storage. The ultimate result is a self-organizing network that provides the lowest latency paths between connected assets that show interest in each other.

The industry is facing an inflection point and a pressing requirement to change.

The Need to Move up the Network Stack

Network conversations today are point-to-point, point-to-multipoint and some multipoint-to-multipoint. They take place over IPv4 and IPv6, over both public and private network infrastructure. Network conversations may be ephemeral, exactly like traditional Ethernet and IP, and some conversations may require durability, such as placing payloads in an ordered stream to act as inputs to other systems at a different time.

Enterprises across the globe utilize VPN technologies, MPLS, private 4G and 5G, short and long range radio networking, dark fiber and satellite connectivity, all of which across IPv4 and IPv6 families.

Networking is complex for fixed systems, but once we introduce real-time transport and routing of data to and from moving vehicles or remote locations out of the range of urban communication technologies like 5G, it becomes extremely difficult and fragile due to the amount of network state, which must be managed properly.

Devices that connect to networks historically were either client or server; some systems needed data and other systems provided it. These systems were designed for high uptime connectivity and made exceptions for periods of sporadic downtime.

DNS records referenced the server IP addresses, so that the accessing systems knew where to find them. As servers were deployed infrequently, DNS was the perfect glue. With modern edge systems, systems provide both data and services, increasingly over dynamic IP addresses private IP addresses that are behind layers of IP address exhaustion prevention technologies like carrier grade NAT (CG-NAT).

Modern edge application developers should not be burdened with connectivity requirements and data synchronization. Instead of utilizing raw sockets, developers can embrace intelligent libraries that act as middleware, in turn, brokering connectivity and providing data synchronization in periods of no connectivity using a store and forward pattern.

OT like smart vehicles and smart devices follow this pattern. Some of these OT systems come from a factory and are not subject to continuous delivery (CD) practices, making regular Over The Air (OTA) software updates infeasible or entirely impossible.

Introducing a “subjects not sockets” paradigm for network applications, an application requires the use of only one socket to gain access to every enterprise network service and data store it is authorized for. This shifts the design to not be concerned with the **where**, but rather, the **what**.

Irrelevant of physical location, IP address family and network type, subject-based networking offers homogenous access and lowest latency routing to services and data over the top of an heterogeneous network. This paradigm shift reimagines how we connect data to services, enabling connectivity from data centers to the edge, to moving vehicles, satellites and to your wrist watch, paving the way for AI services across an entire enterprise. By moving up the stack as hinted at in this section, developers can spend more time working on AI models and perfecting their insight delivering capabilities and less time on infrastructure.

Unifying the Developer Experience

As the requirements placed on those designing these systems skew out of control, it's expected and normal that the number of technologies to choose from and utilize, increases. With ever-increasing choices for developers, architects, and operators who must choose, learn, integrate, and ship, the operational and maintenance complexity grows exponentially.

Software development costs increase in an O(N * Fx) fashion, where N is the number of platforms, and F (a function), returns a measure of effort based on the time to get started, maintain and solve common problems.

Organizations that apply DevOps, continuous delivery, or lean practices optimize the continuous feedback loop for creating, delivering, and iterating on software. However, by simplifying the technology stack and the development process, organizations stand to gain a common business language, a set of patterns, and deep understanding of software systems.

These gains reduce the total cost of ownership and cognition overhead. In addition, operations teams have a minimized surface area to maintain, secure, troubleshoot and operate.

Enabling Simplicity

As organizations require AI to be “on-net” and ready to be used, connecting these systems organizationally requires exceptional planning, project management and deployment. IT projects come with notorious expectations and introducing AI en-masse will push skills to the limit.

A system at the edge should be able to make a request against a system in a data center as easily and as simply as accessing another service at an edge location, which may even be a website on a smartphone or tablet, across a homogeneous and reliable network.

As the inverse relationship between where we can make use of trained data and where the data is generated becomes a critical business problem, AI systems will move from centralized to distributed and for smaller more focused applications, the AI might even be trained at the near or far edge location, close to where the data is generated. The model and weights will then be transmitted over the network for backup, testing, validation and re-use purposes, and other systems will interact with the freshly trained AI system where it was trained and hosted as an inference engine. Edge systems like connected vehicles will also be able to download new models and weights to update onboard AI systems for the same system and reliably distribute signals detected by those models to other interested systems. We are about to enter the era of distributed system engineering.

AI projects are fraught with complexity and high costs, and enterprise IT project statistics offer a grim insight. By using a traditional approach to solve new challenges, many organizations sadly will form a part of the numbers below.

- 52.7% of software projects will cost 189% of their original estimates.
- 75% of business and IT executives anticipate their software projects will fail.
- Only 16.2% of software projects are completed on time and on budget.

Zipdo, (2023). Software Project Failure Statistics: Slide Deck ***Zipdo.co***

By leveraging a simple but powerful data communications layer, any organization can embrace AI at the Edge requirements enabling the trends of tomorrow.

Acknowledging and Preparing for Trends

Despite huge advancements in AI, predicting the future is still not as reliable as any of us would like it to be. As new technologies become available, as a communications industry, we must provide a path for the un-expected. Interacting with data rich systems is a predictable future requirement and as the human race evolves alongside technology, explores the stars, and relies on self-driving cars, reliable, simple and unified communication systems will be front and center.

As continual improvements are made with silicon and chips become smaller and more power efficient, it is almost inevitable that we will see the network become a distributed AI communication system. It's not difficult to imagine AI systems programming and updating other systems as part of day to day routine operations, removing humans from operations.

> **Anything that can get connected will be - and, as a result, massive amounts of data will be (and already are) generated at the edge. The scale of this rapidly exceeds the available network bandwidth to upload all of this data to the cloud and it's exacerbated by the fact that today's networks are not optimized for uploading but rather for downloading.**
>
> *Ouissal.S, Bridgwater.A. (2023, August 21). Why AI Is On The 'Edge'. **Forbes***

Solving These Requirements with Synadia

The NATS.io project, created and maintained by **Synadia**, was initially designed to be aligned with the vision in this document. There are many applications of this technology, and in the scope of AI, it connects users and machines to AI systems, regardless of their physical location, including IP address with a real-time distributed data platform that scales across clouds & geographies.

It provides subject-based, conversation style transport at Layer 7 and for network transport, rides over the time proven TCP/IP. All of the access patterns mentioned in this document can be satisfied and any sensor can connect to any interested system.

Trained AI systems can run as a service at the edge and the outcome is a true distributed, uninhibited, data communications system.

For remote tiny edge locations, NATS also offers asynchronous catch-up, meaning data can be collected locally and copied to a remote location without loss in the event of sporadic connectivity.

It provides both ephemeral and durable data handling, offers point-to-point (request-reply) and point-to-multipoint (pub-sub) communication paths without using IP multicast and also offers functionality like data streams, key-value and object storage.

Features like encryption for data at rest and in-flight, and a myriad of authorization options, including a decentralized zero-trust model, makes NATS operationally ready today to cope with the demands of modern AI connected systems, from the furthest of edge to the most powerful data center.

Edge is a concept vs. a specific location or environment. You can find NATS deployed in vehicles, satellites, point-of-sale systems, and AI systems across the globe.

> **Thanks to the commercial maturation of neural networks, proliferation of IoT devices, advances in parallel computation and 5G, there is now robust infrastructure for generalized machine learning. This is allowing enterprises to capitalize on the colossal opportunity to bring AI into their places of business and act upon real-time insights, all while decreasing costs and increasing privacy.**
>
> *Yeung.T, (2022, February 17). What is Edge AI and How Does It Work?* ***Nvidia***

NATS provides all the functionality required for handling raw data and providing access to services like AI inference, all through a consistent platform and set of tooling.

Whilst modern infrastructure makes it easier to communicate over existing IP networks, it does not make it data centric and requires low level handling of the **how** and less about the **what**, whereas NATS utilizes the best of modern communication infrastructure and makes the **what** extremely simple. Simply put, NATS is carrier-grade networking for your application.

If you are interested in learning more about how Synadia can accelerate your edge AI application deployments our team is available to discuss your use case. **synadia.com/contact**.

Synadia