SwiftInference for Telcos: Monetizing Edge AI at the Tower

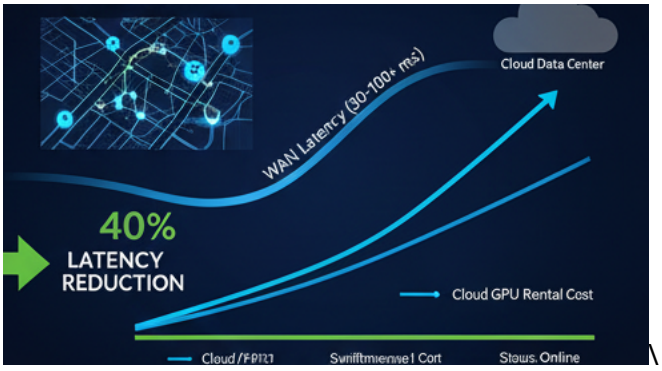
Introduction & Product Overview

Telecom operators are uniquely positioned to deliver ultra-low-latency AI services by leveraging their cell towers as edge compute sites. A user device is often **<1 millisecond** away from a nearby tower, offering unparalleled proximity for real-time processing. **SwiftInference** is an edge AI inference platform purpose-built for this opportunity. It packages data center-class AI performance into a compact, telecom-ready appliance. Built on NVIDIA's **GB10 Grace-Blackwell** Superchip (as used in DGX Spark) and a high-efficiency **BES-1** AI accelerator, SwiftInference delivers up to **1 PetaFLOP** of AI compute in a shoebox-sized form factor. The unit comes preloaded with a full AI software stack (NVIDIA DGX OS and containers) for out-of-the-box deployment at the network edge. Multi-tenant virtualization and strong security are integrated, allowing operators to host models from different enterprise customers in isolation. In short, SwiftInference lets telcos turn each tower into a mini cloud for AI – with **tower-friendly power, cooling, and remote management** design.

Figure: NVIDIA's DGX Spark (top) exemplifies the compact hardware behind SwiftInference's edge platform – delivering 1 PFLOP of AI performance in a 150×150×50 mm chassis. Such small, power-efficient devices can be installed in base station cabinets with minimal impact on space or cooling.

Edge Deployment Readiness

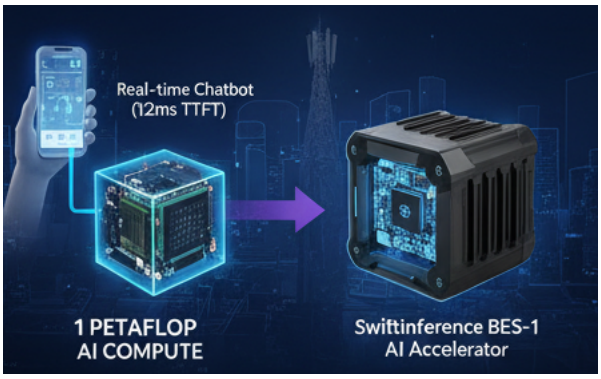
SwiftInference appliances are engineered for **hassle-free tower deployment**. Each unit measures about *15 cm (6”)* on a side and weighs ~1.2 kg, easily fitting alongside existing baseband and MEC equipment. Power draw is modest – roughly **200–250 W** under load – comparable to a small cell radio, and well within available power budgets[1]. The fanless chassis uses an all-metal heat-sink design for efficient cooling, meaning it can operate in outdoor cabinet conditions without specialized HVAC. For ruggedness, the hardware tolerates wide temperature ranges and vibration typically found at tower sites. SwiftInference nodes network seamlessly with the telco’s backhaul via 10 GbE or fiber, and can optionally leverage 5G fronthaul for direct RAN integration. Deployment is plug-and-play: once mounted and powered, the node auto-configures and registers with a central orchestration system. **Tenant control** is enabled through this orchestration – operators (or their enterprise clients) can remotely upload models, allocate inference slots, and set usage policies for each tenant, all while the platform ensures strong isolation between workloads. This gives telcos fine-grained control to meet SLAs for each customer or application on a shared tower resource. Finally, SwiftInference’s compact and efficient design ensures **minimal impact on tower power/cooling**. A typical unit dissipates under 1,000 BTU/hr, easily handled by existing cooling at many sites. In summary, the platform is “tower-ready” – both physically and in remote manageability – allowing rapid rollout across tens of thousands of cell sites.



ROI and Economics per Tower

By deploying SwiftInference at towers, telcos can unlock new revenue streams in the emerging edge AI market. With mobile subscriber growth slowing, operators are eyeing AI inference services as a major revenue opportunity. **Monetization per tower** can be substantial: McKinsey estimates telco-provided GPU-as-a-Service could reach **$35–70 billion** globally by 2030. SwiftInference enables carriers to tap into this by renting out on-site AI processing to enterprise tenants, content providers, and application developers. The **hardware cost** of a SwiftInference node is modest (on the order of only a few thousand

dollars each[2]), yet each node can serve many customers or AI tasks concurrently thanks to virtualization and efficient scheduling. This utilization model means high ROI: a single tower unit running near capacity (for example, handling local AR/VR streams, language model queries, etc.) can generate enough monthly revenue to pay back its cost in **well under a year**. Even at lower utilization, payback periods of roughly **12–18 months** are achievable, after which the revenue is largely profit. Importantly, operators avoid the ongoing cloud GPU rental costs they would otherwise incur – for instance, a continuous GPU inference instance in the cloud can cost $500–$2,000 per month. Owning the inference hardware outright lets telcos **capture that margin**. Additionally, **tenant-specific pricing** can be applied: e.g. charging premium rates for strict latency SLAs or guaranteed capacity, and lower rates for best-effort bulk processing. The result is a flexible economic model that turns edge infrastructure into income. Beyond direct revenue, there are **cost savings** too – processing data on-site reduces backhaul bandwidth usage (and costs) by keeping heavy AI data local, and it can lower regulatory compliance costs by keeping sensitive data (e.g. video feeds, user data) on-premises rather than sending it to cloud. Overall, SwiftInference offers a compelling business case: **low capital cost, high utilization, and quick payback**, positioning each tower as an earning asset in the AI economy.



Performance and Latency SLAs

SwiftInference is designed to meet the **strict latency requirements** of real-time applications at the network edge. By hosting inference on-site at the tower, it minimizes the network hop – eliminating the 30–100+ ms of WAN latency that cloud data centers would introduce. This translates to **blazing-fast response times**: in fact, tests show placing AI inference closer to users (e.g. in AWS Local Zones) yields a consistent **40%+ latency reduction** across all percentiles (P50 through **P99**) compared to a distant region. For example, a user in Los Angeles saw P99 latency drop from ~197 ms in a cloud region to ~141 ms at a local edge site. In Honolulu, P99 improved from an unacceptable 472 ms down to 273 ms by using a nearby edge node – a dramatic cut in worst-case delay. SwiftInference leverages this proximity to help telcos **commit to strict latency SLAs**, such as <50 ms end-to-end response for interactive applications. Moreover, the platform's architecture provides **tail latency protection**: it uses adaptive scheduling and load

management to prevent occasional slow requests from impacting others. By running models in parallel "slots" and dynamically allocating extra compute to backlogged requests, SwiftInference keeps the **99th-percentile latency low** even under bursty loads. The system is also **streaming-first** in its approach to inference. This means that for workloads like generative AI text or video analytics, SwiftInference begins returning output incrementally as soon as inference starts, rather than waiting for completion. This streaming capability, combined with near-zero network transit time to the user, yields an almost instantaneous feel – the **Time-to-First-Token (TTFT)** for a chatbot or first bytes of an object detection video stream can be well under 100 ms in many cases. (By comparison, cloud-based generative AI often struggles to keep TTFT below 200 ms due to network overhead[3][4].) In short, telcos can confidently offer **ultra-low latency** inference services with SLAs backed by SwiftInference's edge deployment and intelligent scheduling.

Key Use Cases at the Edge

SwiftInference unlocks a variety of high-value AI use cases that benefit from **edge execution**. Some scenarios especially relevant to telcos include:

- **Large Language Model Services (OpenAI-style LLMs):** Telcos can host sizable GPT-style models at tower MEC sites to power local generative AI services. For instance, an enterprise tenant could deploy a customer-support chatbot or real-time translation service on a SwiftInference node at a regional 5G hub. The large unified memory (128 GB) allows running models with **up to 200 billion parameters** right on the edge. This yields fast, locality-aware responses for end-users and keeps sensitive data within local jurisdiction (a key advantage for data sovereignty and compliance). Instead of each user's query traveling to a distant cloud (introducing latency and potential privacy issues), the tower processes it immediately and streams back a response. This can enable telcos to offer **"LLM-as-a-Service"** at the edge, opening new B2B revenue opportunities.

- **Real-Time Voice AI:** Voice-driven applications (personal assistants, interactive voice response, augmented reality audio, etc.) demand extremely low latency to feel natural. With SwiftInference at the cell site, telcos can support **real-time speech recognition, language translation, and audio analytics** for mobile calls or apps. For example, imagine a voice translator on a phone call: audio from the caller is sent to the tower, transcribed and translated by an AI model on SwiftInference, and the translated speech is sent to the other party – all within a few tens of milliseconds. This ultra-low round-trip delay enables seamless conversation. Similarly, telcos can deploy AI noise suppression or voice biometrics at the edge to enhance call quality and security. By processing voice data locally, they also reduce core network load and address privacy concerns (since raw voice needn't leave the area). **Latency SLAs** can be maintained even under load, ensuring voice interactions remain smooth.

- **Computer Vision and Video Analytics:** Many emerging services – from smart city surveillance to industrial IoT – rely on analyzing video or images in real time. SwiftInference provides the GPU acceleration to run **computer vision models (CNNs, object detectors, etc.)** directly at the tower where camera feeds or sensor data converge. For instance, a telco could offer an edge AI service to retailers: store CCTV feeds are streamed to the nearest cell site, where SwiftInference runs a vision model to count customers, detect shoplifting, or analyze inventory, and then only sends summarized insights to the cloud. The heavy image processing happens on-site, minimizing bandwidth usage. Another example is **smart traffic management**: cameras on intersections or vehicles send video to a 5G tower; SwiftInference detects pedestrians or hazards and sends instantaneous alerts to connected cars nearby. The local processing avoids the multi-hop delays of cloud processing – critical for safety. By enabling **real-time computer vision** at the edge, telcos can support applications like AR/VR (low-latency scene recognition), drone or robot control, and more, with new revenue models (e.g. charging per video stream analyzed).

- **V2X and Autonomous Driving Support:** Connected vehicle applications are among the most demanding in latency and reliability. SwiftInference nodes co-located with 5G base stations can function as **“ roadside AI servers ”** for connected cars. They

can aggregate data from vehicles, roadside sensors, and HD maps, and run intensive AI algorithms (e.g. 3D object detection, sensor fusion, predictive analytics) to assist with autonomous driving beyond the car's onboard capabilities. For example, a cluster of vehicles approaching an intersection could all feed live data to the SwiftInference at the cell tower; the edge AI system quickly computes a unified view of hidden obstacles or optimal traffic flow and broadcasts guidance back to the cars within milliseconds. This **cooperative perception** and decision-making greatly enhances safety and efficiency. Because the inference is happening one hop away on the cellular network, the **end-to-end latency** (vehicle-to-tower-to-vehicle) can be kept low enough for collision avoidance and real-time control loops – something not feasible if relying on distant cloud servers. Telcos can partner with automotive companies and cities to provide these V2X edge services, monetizing them per vehicle or intersection. The **tail latency guarantees** of SwiftInference (consistent performance even under peak traffic) are especially vital here to meet the strict deadlines of autonomous systems.

Competitive Comparison

Telcos evaluating SwiftInference will consider how it stacks up against alternatives like central cloud processing or fully on-device AI. Below we compare these approaches:

- **Versus Cloud-Based Inference (Central Data Center):** Cloud AI services (e.g. running models on AWS or other hyperscalers) offer scalability but suffer from inherent **network latency and less control**. As discussed, simply moving an inference endpoint from a distant region to a local edge site cut P99 latencies by **~30–50%** in practice. SwiftInference takes this to the extreme by placing inference **at the last mile**, often eliminating dozens of milliseconds of transit. This is crucial for meeting emerging application needs (AR, vehicle safety, real-time interactivity) that cloud cannot reliably support due to jitter and variable paths. Additionally, cloud inference incurs ongoing usage charges and can spike in cost as usage grows, whereas SwiftInference is **capex-based** – telcos make a one-time investment and then amortize it over many queries, often at a much lower per-query cost than cloud's pricing. There is also a strategic angle: running AI in your own network keeps you in control of the service quality and data. Many enterprises (and governments) prefer that sensitive AI data (e.g. live camera feeds, customer conversations) **stay within the telco's network** rather than traversing the public internet to third-party clouds. SwiftInference enables telcos to be that trusted, sovereign AI infrastructure provider. In summary, compared to cloud, SwiftInference offers **far lower and more predictable latency, lower long-term cost at scale, and greater data control**. The trade-off (managing hardware) is mitigated by SwiftInference's integrated solution and remote management tools, making it nearly as convenient as using cloud – but with performance that cloud can't match.

- **Versus Fully On-Device Inference (Mobile/UE Hardware):** At the opposite end, one could push AI models entirely onto user devices (smartphones, IoT devices) using chips like Apple's Neural Engine or Qualcomm's AI DSP. On-device AI has the benefit of zero network latency – computations happen in microseconds on the device – but **severe limitations in model size and updateability**. Mobile neural accelerators can only handle relatively small models (typically a few hundred million parameters at most) due to limited memory and power. They excel at tasks like filtering photos or short voice commands, but they cannot run the kind of large-scale models (multi-billion parameter transformers, advanced vision networks) that drive cutting-edge AI experiences. Developers are forced to compromise model complexity to fit on device, often sacrificing accuracy or capabilities. SwiftInference removes that limitation by giving devices access to **server-grade AI** just one wireless hop away. A phone or AR headset can offload a heavy model to the edge and get a result in, say, 20–50 ms – virtually indistinguishable from an on-device 5 ms processing when factoring network delay. This means **low-latency interactivity with high-complexity models**. Another consideration is **deployment and updates**: On-device models require app updates or firmware changes to improve, and not all users update promptly; in contrast, with SwiftInference, a telco or developer can update the model on the edge server and instantly improve service for all users in that cell coverage. Moreover, on-device AI doesn't generate revenue for telcos (it bypasses the network), whereas edge inference keeps the telco in the loop as a service provider. Finally, devices vary widely – some users have older phones without advanced AI chips, making on-device experience inconsistent. Edge inference standardizes the experience: every user connected to a SwiftInference-enabled tower gets **the same blazing-fast, powerful AI service**, regardless of their device's age. The combination of near-device speed and cloud-level model sophistication gives SwiftInference a **competitive sweet spot** that neither pure cloud nor pure on-device approaches can achieve.

In summary, SwiftInference allows telcos to offer **cloud-grade AI at on-device latencies** – a unique value prop. Competing solutions like public cloud inference struggle with latency and data locality, while on-device AI struggles with scale and manageability. By leveraging its **high-performance per watt hardware** (tailored AI silicon delivering 1 PFLOP in ~210 W[1]) and strategic edge placement, SwiftInference achieves industry-leading performance per dollar and per millisecond of latency, outpacing both traditional approaches. It effectively creates a new category of **carrier-edge AI service** that can attract customers from both ends (cloud users seeking lower latency, and app developers seeking more power than devices can provide).

Conclusion: Enabling Telco AI Innovation

SwiftInference empowers telecom operators to transform their infrastructure into an agile AI cloud – one that is **geographically distributed at the edge**, extremely performant, and under the telco's full control. With SwiftInference, each cell tower or central office can

become a revenue-generating **“AI hub”** serving the next generation of applications that demand real-time intelligence. The platform’s **tower-ready design**, strong ROI profile, and proven performance (with p99 latencies slashed and streaming inference capabilities) make it a compelling solution for telcos eager to capitalize on edge computing. Notably, leading operators worldwide have begun pilot deployments of AI inferencing services at the network edge – a clear signal that the industry is moving in this direction. By adopting SwiftInference early, telcos can gain a competitive edge in offering differentiated 5G services (like immersive media, smart city solutions, V2X safety services, etc.), all while cultivating new partnerships (with AI startups, enterprises, and cloud providers) who will be the tenants of this edge infrastructure. In a landscape where speed and intelligence define user experience, SwiftInference gives telcos the tools to meet the most demanding **latency SLAs** and to do so profitably. The message is clear: **AI at the edge is now viable and valuable**, and SwiftInference is the turnkey platform to make it a reality. Telecom operators should seize this opportunity to pilot SwiftInference in their networks – those who deliver ultra-low-latency AI services first will not only delight customers but also **unlock lucrative new revenue** in the AI era. SwiftInference for Telcos is where connectivity meets intelligence – right at the tower, where it can deliver the most value.