White Paper for AI-Native Deployable 5G Systems and Autonomous Edge Orchestration

1. Reference Web-Link & Product Summary

- **Reference Web-Link:** NextServer AI 5G Fly-Away Kits - NextComputing
- **Product Summary:** The NextServer AI 5G Fly-Away Kit (FAK) is a portable "Data Center in a Suitcase" engineered for teams requiring high-performance computing in the most remote environments on Earth. Integrating a high-density AmpereOne® 192-core arm64-bit CPU with carrier-grade 5G software (SRS/LF OCUDU) and Edge AI capabilities, the FAK functions as a secure, globally mobile "Tactical Cloud." The team is comprised of NextComputing, Software Radio Systems (SRS), DeepSIG, Canonical or Rancher RGS, Ampere Computing (AARCH 64 and opensource AI enablement tools and models) and the OCUDU Opensource RAN project https://ocudu.org/ and ARM

- **Key Capabilities from Web Summary:**
  - **Performance:** Features up to 1TB RAM across 8 channels of DDR5 memory, massive storage capacity (up to 1 PB NVMe in JBOD or RAID), and optional NVIDIA RTX 4000 or L4 GPUs providing acceleration for high-demand inference.
  - **Connectivity:** Equipped with a built-in 5G radio unit (RU), portable cellular antennas, and satellite backhaul support (e.g., Starlink Mini) to ensure private, secure 5G deployments where data never leaves the local network.
  - **Form Factor:** TSA/IATA-compliant airline carry-on (22” x 14” x 9”).
  - **Ecosystem:** Optimized for Ubuntu 24.04 LTS and Canonical Kubernetes or Rancher RGS Kubernetes.
  - **Mission Applications:** Enables real-time wildlife cataloging (animal detection, species classification, bear facial recognition), autonomous site operations in disaster zones, cybersecurity threat hunting, and the coordination of agricultural robot swarms.
  - **Easy Management:** AI applications are quickly deployed and managed by Kasm Workspaces

NextServer AI 5G Fly-Away Kit]

The exterior of the NextServer AI 5G is ruggedized airline carry-on suitcase, designed for global travel. Internally, the 2U modular chassis is visible, featuring cooling fans, removable NVMe drive bays, and connectors for portable cellular and backhaul antennas.

---

2. Executive Summary and Strategic Value

The contemporary landscape of tactical communications and remote sensing is increasingly defined by the necessity for high-density computational resources at the network edge.

The emergence of the NextServer AI 5G Fly-Away Kit (FAK) represents a significant technological leap, consolidating a carrier-grade 5G base station, a high-density AI compute cluster, and a multi-access edge computing (MEC) into a single, TSA-compliant portable unit. This architectural synthesis enables a "Bring Your Own 5G and AI MEC" capability, providing a localized "Tactical AI Cloud" that facilitates autonomous drone swarm orchestration, real-time RF threat hunting, and high-fidelity edge analytics in locations where traditional terrestrial

infrastructure is either destroyed or economically impractical.

Figure 1: Single HotSpot

---

3. Physical and Mechanical Design for Global Mobility

The architectural integrity of a deployable system is fundamentally limited by its portability and resilience to diverse environmental conditions. The NextServer AI 5G system is integrated into a custom-engineered, tool-less modular hard case that adheres to TSA and IATA airline carry-on dimensions of 22 inches by 14 inches by 9 inches. This form factor is a critical design requirement, allowing field teams to transport a data-center-class base station as personal luggage, thereby bypassing the logistics delays and security risks associated with traditional air cargo shipping and checked baggage.

The mechanical design prioritizes rapid field servicing, utilizing removable mid-plane drive bays and an internal 2U configuration that facilitates the integration of full-height or half-height PCI Express expansion cards without the need for specialized tools. Environmental tolerance is a primary consideration for the system's operational envelope. Power delivery is SWaP optimized.

Figure 3 - NextComputing FlyAway Kit]

A user easily transports the NextComputing FlyAway Kit as rolling luggage. The accompanying hardware specifications highlight its capabilities: an Ampereone CPU (191-core, 2.6 GHz), NVIDIA RTX 4000 GPU, and Ubuntu 24.04 OS packed into the rugged enclosure.

Figure 4 - Key RAN Components]

Mechanical and Environmental Specifications

| Category | Specification Details | Operational Significance |

| **Form Factor** | TSA/IATA Compliant Carry-on (22” x 14” x 9”) | Enables rapid global deployment via commercial air. |

| **Chassis Depth** | 21.75” W x 13.875” D x 9” H | Optimized for suitcase-level transportability. |

| **Redundant Power** | 1+1 Hot-Swap 600W / Single 850W 80 Plus Platinum | Ensures uptime during PSU failure or power source shift. |

| **Operating Temp** | 0°C to 40°C (32°F to 104°F) / Storage: -20°C to 70°C| Supports deployment in varied climatic conditions. |

| **Storage Capacity** | Up to 1PB NVMe high performance SSDs | Supports massive local data recording at the edge. |

| **Compliance** | Meets FCC Class A, CE, TUV, ROHS, TAA Compliant | Meets federal security and environmental standards. |

---

4. Computational Architecture: The AmpereOne 192-Core Processor

The transition from traditional x86-based architectures to high-density ARM64-bit platforms is driven by the need for superior performance-per-watt and massive parallelism at the edge. Maximizing the compute for the give size, weight and power (SWaP). The NextServer AI 5G is

powered by the AmpereOne A192-32X processor, featuring 192 custom-designed ARM cores operating at 2.6GHz. This architecture is specifically engineered for cloud-native workloads, providing a dedicated core for every network function and AI task, which significantly reduces the latency overhead introduced by the core-sharing and context-switching common in multi-threaded x86 environments.

The processor’s 192 cores are arranged in a high-speed mesh interconnect that facilitates low-latency data movement between cores and memory. Each A1 core is equipped with 2MB of private L2 cache, totaling 384MB of L2 cache across the socket, supplemented by a 64MB system-level cache. This cache hierarchy is essential for managing the high throughput required for 5G baseband processing and real-time video analytics simultaneously. Furthermore, the move to eight channels of DDR5 memory provides a substantial boost in bandwidth, which is critical for the memory-intensive nature of edge AI inference pipelines.

Performance and Power Metrics The performance benchmarks indicate that the AmpereOne part provides exceptional value in terms of cores-per-dollar and performance. For "cloud-native" deployment scenarios—where the system is constantly under load from 5G baseband tasks and AI analytics—the performance-per-watt advantage is highly compelling.

---

5. Software-Defined Infrastructure and O-RAN Integration

The system’s flexibility is derived from its "Linux of RAN" approach, utilizing an open-source, neutrally-governed software stack developed by Software Radio Systems (SRS) in collaboration with the Linux Foundation. The LF OCUDU (Open Centralized Unit / Distributed Unit) implementation provides a carrier-grade 5G RAN that eliminates vendor lock-in and the associated high recurring licensing fees of proprietary Tier 1 vendors. This hardware-agnostic software architecture allows for the rapid integration of diverse Radio Units (RUs) and the deployment of network functions on standard ARM-based compute nodes.

Layered above Canonical's Ubuntu 24.04 LTS operating system on this hardware are all components of our end-to-end processing pipeline: (a) timing-critical O-RAN software; (b) the 5G control plane; (c) the mesh network management software; (d) the autonomous drone swarm pipeline; and (e) containerized AI microservices. Orchestration via Canonical Kubernetes creates a stable environment for dynamic management of mission-specific applications, such as the Android Team Awareness Kit (TAK) or push-to-talk (PTT) services, directly at the edge.

O-RAN Disaggregation and Functional Split

The system supports a disaggregated O-RAN architecture, which is fundamental to its ability to scale across different deployment environments. By separating the Centralized Unit (CU),

Distributed Unit (DU), and Radio Unit (RU), operators can optimize resource allocation based on the mission's latency and bandwidth requirements.

O-RAN Component	Software/Hardware Implementation	Primary Responsibility
CU (Centralized Unit)	srsRAN Enterprise 5G (L3/RRC)	Session management, mobility, and high-level control.
DU (Distributed Unit)	LF OCUDU (L1/L2 High)	Timing-critical MAC/RLC and scheduling.
RU (Radio Unit)	Band n48/n77 COTS RU	RF transmission, amplification, and DAC/ADC.
Backhaul Interface	Starlink Mini / 100G QSFP28	Global reach-back and high-speed data transport.

This modularity ensures that the NextServer AI 5G can operate as a standalone "all-in-one" unit or serve as a high-density compute node within a wider 5G mesh network.

---

6. Mission-Ready AI Applications: Wildlife and Border Surveillance

The Technology Readiness Level (TRL) of the NextServer AI 5G ecosystem is validated through an automated surveillance and individual identification in remote regions. The "Wildlife Conservation AI 5G Hotspot" demo serves as a universal blueprint for missions requiring real-time detection and classification in areas where traditional internet connectivity is absent.

The Edge AI Inference Pipeline

The integration of the Ampere-optimized AI library allows the system to process high-resolution video streams from 5G-capable cameras and drones with minimal latency. The inference pipeline typically involves several concurrent models:

1. **Animal/Object Detection:** A YOLO-based MegaDetector isolates subjects of interest from complex environmental backgrounds.

2. **Species/Target Classification:** SpeciesNet and similar classification architectures identify the specific type of subject detected.

3. **Individual Identification:** The BearID model provides facial recognition capabilities for tracking specific individuals (animals or persons) across a coverage area.

The system produces geolocated and timestamped datasets in standardized formats like Camtrap DP, ensuring interoperability with global biodiversity or security databases. This capability is immediately transferable to border security missions, where the identification of unauthorized movements and the classification of vehicles or person-carried equipment are critical for situational awareness.

Real-World Performance Benchmarks

The NextServer AI 5G has demonstrated the ability to serve a massive number of users and sensors while maintaining the low latencies required for edge AI.

Metric	Validated Performance	Operational Significance
Simultaneous Users	2,000 with multiple cell antennas	Supports large-scale sensor networks and field teams.
Aggregate Downlink	14.3 Gb/s	Enables multiple 4K/8K video streams to the edge.
Mean Network RTT	10-15 ms	Essential for real-time control and cognitive aids.
Frame Latency (TerraSLAM)	16.7 ms per frame	Allows for high-speed GPS-denied navigation.
Detection Speed (DeepSig)	100x Faster than traditional DSP	Enables immediate reaction to RF threats.

These benchmarks confirm that the system can function as a centralized "brain" for complex missions, offloading the thermal and power constraints of edge devices like drones and wearables to the FAK’s high-density compute core.

7. Autonomous Drone Swarm Orchestration: The SteelEagle Platform

Drone-agnostic, Mission-Centric, and Open Source: SteelEagle is an open-source platform developed at CMU for MEC-based AI on COTS drones. Through a cellular network, it transforms commercial off-the-shelf (COTS) drones into fully autonomous entities conducting a single collaborative swarm mission, or independent missions. The use of cellular networks overcomes the range limitations of single-hop wireless technologies, and enables BVLOS ("Beyond Visual Line of Sight") operations. Version 2.1 of SteelEagle was released in July 2025. SteelEagle cleanly separates drone-specific proprietary advances from drone-agnostic mission specifications and MEC AI. By using COTS drones, our work avoids drone hardware development and allows us to focus on networking.

Overcoming Drone SWaP Constraints via Offloading to AI MEC

Small, lightweight drones (nano-drones) are favored for tactical missions due to simplified regulatory approval (e.g., FAA rules for drones under 250g) and their ability to operate in confined spaces. However, these drones lack the onboard compute power required for complex mapping and trajectory planning. SteelEagle addresses this by transmitting real-time video to the NextServer AI 5G platform, which performs the intensive computer vision and pathfinding tasks before sending actuation commands back to the drone.

DroneHub Bridge and BVLOS Operations

A key challenge with COTS drones today is that most do not use cellular networks. Rather, they use WiFi, Lightbridge, or MicroHard. SteelEagle utilizes "DroneHub" modules—ranging from 26g to 40g—to bridge drone telemetry to standard 5G cellular networks. Our experiments indicate that a safe heuristic is a DroneHub weight limit of 10% of Maximum Takeoff Weight (MTW).

- **Thick Client DroneHub (e.g., Samsung SmartWatch - 26g):** Includes primitive onboard AI for mission persistence during brief signal interruptions. However, it suffers from thermal issues, sustaining only 0.7 FPS over 4G LTE before triggering thermal shutdown to prevent skin burns.
- **Thin Client DroneHub (e.g., Onion Omega - 40g):** Performs packet forwarding without onboard AI. Optimized for high frame rates (30 FPS) over cellular without thermal shutdown, operating as a strict thin client to the HotSpot.

Mission-Centric Benchmarks

To test a specific drone's performance in a HotSpot mesh setting requires standardized benchmarks to quantitatively assess factors like network latency, throughput, wireless handover, and MEC state handover based on actual drone flights.

- **Obstacle Avoidance Benchmark:** Focuses on obstacle avoidance vital for flights in building-dense spaces. Drones navigate a precisely-defined slalom pattern of racing flags. The difficulty is controlled by a spacing parameter (w). Adding network latency degrades the benchmark score, with more severe degradation at higher latency and close obstacle

spacing.

- **Agility Benchmark:** Assesses the agility of visual object tracking, where a drone follows an unpredictable moving target. The benchmark utilizes a DJI Robomaster S1 ground robot.

  (Obstacle Avoidance): (a) The Drone Camera View approaching the slalom course of black-and-white flags. (b) The Plan View detailing the flag arrangement and spacing parameter (w).

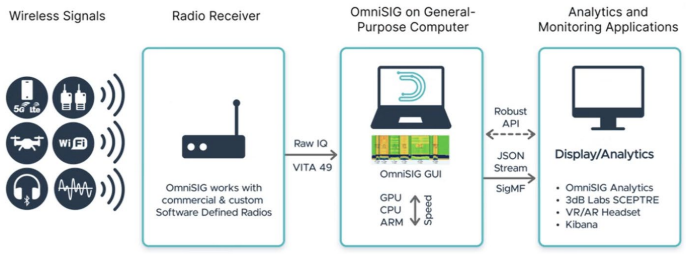
  (Network Latency): A bar chart demonstrating that adding artificial network latency (+250ms, +500ms) significantly reduces the drone's ability to successfully navigate the course across varying distances.

  (Target for Agility Benchmark): An image of the DJI Robomaster S1. This agile, unpredictable vehicle serves as the tracking target for the drone swarm's visual object tracking algorithms.

---

8. RF Spectrum Awareness and AI-Native Wireless Sensing

Deployable networks are highly susceptible to interference and jamming in contested environments. The NextServer AI 5G addresses this through the integration of DeepSig OmniSIG, which provides native, AI-driven spectrum sensing and signal classification.



DeepSig OmniSIG Architecture

The DeepSig OmniSIG architecture illustrates the end-to-end signal processing flow. On the left, **Wireless Signals** (5G/LTE, Walkie-Talkies, Drones, Wi-Fi, Bluetooth, AM/FM) are captured by a **Radio Receiver** (working with commercial & custom Software Defined Radios). The receiver outputs Raw IQ (VITA 49) data to **OmniSIG on a General-Purpose Computer**, which leverages GPU, CPU, and ARM hardware on the HotSpot for high-speed inference. Finally, the categorized data is sent via a Robust API (JSON Stream, SigMF) to **Analytics and Monitoring Applications** such as OmniSIG Analytics, 3dB Labs SCEPTRE, VR/AR Headsets, and Kibana Dashboard for display.

Passive Signal Categorization and Line of Bearing

Unlike traditional spectrum analyzers that require manual signal demodulation, OmniSIG utilizes deep learning to recognize and categorize signal types across instantaneous bandwidths of up to 500MHz. This allows the system to identify fast-frequency hoppers, UAV controllers, IoT emissions, and unauthorized cellular devices in minutes.

- **Passive Categorization:** Recognizes signal signatures without decrypting or demodulating the data stream.
- **Anomaly Detection:** Baselines the local RF environment to detect unusual spectral behavior that may indicate an impending cyber-attack or signal intrusion.
- **Line of Bearing (LoB):** When paired with multi-channel radios, OmniSIG provides azimuth and elevation data, enabling the physical pinpointing of unauthorized transmitters.

---

10. Regulatory, Security, and Compliance Frameworks

The deployment of 5G systems for federal missions requires adherence to stringent cybersecurity and supply chain security standards. The NextServer AI 5G is designed as a TAA-compliant solution, manufactured in the United States to ensure a secure supply chain.

Cybersecurity and NIST AI Agent Standards

As AI agents become more autonomous in drone swarm missions, the need for standardized authentication and governance becomes paramount. The system is being aligned with the NIST AI Agent Standards Initiative, launched in early 2026 to address interoperability, identity, and security challenges associated with autonomous agents. This includes exploring technical approaches for agent authentication, permission scoping, and activity logging within the 5G network framework.

BlueUAS and NDAA Compliance

For drone-based missions, the system is validated for use with BlueUAS-approved platforms. Drones on the "Blue List," such as the Skydio X10D and Parrot ANAFI USA GOV, are cyber-secure and compliant with Section 848 of the NDAA for Fiscal Year 2020, which prohibits the use of UAS equipment from certain foreign entities. The NextServer AI 5G functions as the "brain" for these secure platforms, ensuring that the entire end-to-end data link remains within a trusted, US-governed ecosystem.

Compliance Category	Standard/Requirement	System Status
Manufacturing	TAA Compliant (United States Origin)	Fully Compliant.

Cybersecurity	Secure Boot, TPM 2.0, Boot Guard	Native Integration.
Drone Platforms	BlueUAS / NDAA Section 848	Support for all approved platforms.
O-RAN Interop	3GPP and O-RAN Alliance	Adheres to all open standards.
Network Security	TLS 1.3, OAuth 2.0, Zero Trust	Ongoing alignment (WG11).

---

11. Lifecycle Cost Estimates and Procurement Strategy

The "Bring Your Own 5G" model provides a significant reduction in the Total Cost of Ownership (TCO) compared to proprietary systems. By utilizing open-source software and general-purpose ARM compute, agencies can avoid the high per-user and per-cell licensing fees that traditionally characterize mobile infrastructure.

12. Conclusion: Strategic Value of the AI-Native 5G Edge

The NextServer AI 5G Fly-Away Kit represents the culmination of several technological convergences: the shift toward ARM-based high-density computing, the maturation of open-source 5G software, and the emergence of autonomous swarm intelligence. By providing a "Tactical AI Cloud" in a TSA-compliant suitcase, the system addresses the critical need for low-latency processing and resilient communications in environments where infrastructure is unavailable or compromised.

The system's "Linux of RAN" architecture, built on the LF OCUDU stack, ensures that DHS and other operational components can deploy state-of-the-art 5G technology without the risks of vendor lock-in or proprietary security vulnerabilities. When integrated with AI-native spectrum sensing from DeepSig and autonomous drone orchestration from CMU’s SteelEagle, the FAK becomes more than a communications hub—it becomes a strategic asset capable of detecting RF threats, navigating GPS-denied zones, and cataloging complex environmental data in real-time.

As the system moves through the VINES research roadmap, the focus will shift toward the "battle-hardening" of cross-layer APIs and ARM-optimized physical layer libraries, preparing the platform for the transition to 5G-Advanced and early 6G studies. This forward-looking architecture ensures that federal missions remain "Connected When it Matters Most," providing a stable and future-looking platform for the next generation of autonomous and intelligent

network systems.

Appendix A: Demonstration Readiness and Specifications Summary

Technical Specification and Strategic Roadmap for AI-Native Deployable 5G Systems and Autonomous Edge Orchestration

The NextServer AI 5G Fly-Away Kit (FAK) represents a paradigm shift in tactical communications, consolidating a carrier-grade 5G base station, a high-density AI compute cluster, and a multi-access edge computing (MEC) framework into a single, TSA-compliant portable unit. This architectural synthesis enables a "Bring Your Own 5G and MEC" capability, providing a localized "Tactical AI Cloud" that facilitates autonomous drone swarm orchestration, real-time RF threat hunting, and high-fidelity edge analytics in infrastructure-denied environments.

I. Immediate Demonstration Readiness

The Wildlife Conservation AI 5G Hotspot is a ready solution available for immediate 1-to-2-day demonstration as early as March 2026. This pilot showcases the integration of carrier-grade 5G and high-density Edge AI in a portable form factor, serving as a universal blueprint for remote surveillance and identification missions. This was demonstrated by SRS at the Mobile World Congress show in Barcelona, Spain March 2-5, 2026 and was demonstrated by ARM at the International Conservation Technology Conference February 20-24, 2026

Demonstration Overview and Operational Setup

- **Mission Objective:** Real-time wildlife detection, species classification, and individual identification in environments where traditional connectivity is non-existent.
- **Hardware Deployment:** The NextServer AI 5G FAK serves as the central hub, processing streams from 5G-capable trail cameras and autonomous drones.

- **Satellite Reach-Back:** Utilizes Starlink Mini for high-bandwidth internet connectivity, ensuring geolocated and timestamped datasets (Camtrap DP format) are synchronized with global databases in real-time.

- **Provisioned User Equipment (UE):** The demonstration package includes BlueUAS-approved drones with DroneHub modules, 5G smartphones for TAK/PTT applications, and remote 5G-capable sensors.

Capability	Status	Operational Significance
Wildlife AI Hotspot	Ready	Proven for immediate field

		deployment and evaluation.
High-Density Scaling	Validated	Supports 2,000 simultaneous users and 14.3 Gb/s downlink.
Edge AI Pipeline	Operational	Real-time YOLO detection and BearID facial recognition.
Satellite Backhaul	Integrated	Starlink Mini support for global connectivity anywhere.

II. Physical and Mechanical Design for Global Mobility

The system is integrated into a custom-engineered, tool-less modular hard case adhering to TSA/IATA airline carry-on dimensions (22” x 14” x 9”). This form factor allows field teams to transport a data-center-class base station as personal luggage, bypassing traditional air cargo logistics.

Mechanical and Environmental Specifications

- **Dimensions:** 9” H x 21.75” W x 13.875” D (operational case).
- **Environmental Tolerance:** Operating range of 0 to 40C; storage from -20C to 70C.
- **Power Stability:** 1+1 hot-swap redundant 600W or single 850W 80 Plus Platinum power supplies.
- **Storage Density:** Up to 1PB of high-endurance NVMe SSDs for massive local data recording.

III. Computational Architecture: AmpereOne 192-Core Processor

The NextServer AI 5G is powered by the AmpereOne A192-32X processor, featuring 192 custom-designed ARM cores operating at 2.6GHz. This architecture provides a dedicated core for every network function and AI task, significantly reducing latency overhead compared to traditional multi-threaded x86 environments.

Metric	**AmpereOne A192-32X**	**Operational Advantage**

Core Count	192 Dedicated ARM Cores	Massive parallelism for 5G RAN and AI concurrent tasks.
L2 Cache	384 MB (2MB per core)	High-speed data movement for 5G baseband processing.
Memory	8 Channels DDR5-5200	Up to 1TB RAM for memory-intensive inference pipelines.
Peak AC Power	431 Watts (Avg) / 696 Watts (Peak)	High-density compute within a portable thermal envelope.

IV. Software-Defined Infrastructure and O-RAN Integration

The system utilizes a "Linux of RAN" approach, featuring an open-source, neutrally-governed software stack (LF OCUDU) developed by Software Radio Systems (SRS). This hardware-agnostic architecture eliminates vendor lock-in and the high recurring fees of proprietary Tier 1 vendors.

Disaggregated O-RAN Architecture

- **O-DU (Distributed Unit):** SRS-developed LF OCUDU software managing timing-critical L1/L2 layers.
- **O-CU (Centralized Unit):** srsRAN Enterprise 5G managing session control and mobility.
- **O-RU (Radio Unit):** Supports multi-vendor O-RAN compliant RUs for bands n48 (CBRS) and n77.
- **Orchestration:** Built on Ubuntu 24.04 LTS and managed via Canonical Kubernetes for secure, containerized mission models.
