White Paper

Zeblok's Ai-MicroCloud™ Enables Rapid Scaling of AI at the Edge With OpenVINO™ Toolkit Optimization

Author

Mouli Narayanan

March 2022

Executive Summary

An estimated one trillion edge devices by 2035[1] will require millions of Multi-Access Edge Computing data centers (MECs). This is a paradigm shift, which requires substantial edge data center autonomous operations investments to match the expectations from decades of investments in automation in traditional data centers. Furthermore, it requires efficient packaging/software distribution on disparate hardware platforms at the edge.

Challenges in delivering edge AI applications can be summarized as one of ecosystem, workflows, automation, and perhaps most importantly, scale, requiring end-to-end lifecycle management and future proofed solutions.

Ecosystem: Edge AI application delivery involves multiple different hardware SKUs, cloud service providers, and independent AI software vendors (ISVs).

Workflow: Application developers, especially those building deep learning inferencing, need integrated workflows to automate AI inference deployments at scale. Workflows provide consistency in AI digital asset curation. Furthermore, operators need tools to move data from the edge to upstream hybrid-cloud data centers to refine AI-driven insights further.

Automation: ML DevOps, continuous integration & deployment, handling multiple ML pipelines into edge data centers is made difficult due to the lack of tools typically taken for granted in the public cloud environments such as compute, storage, and network virtualization.

Scale: Low latency inferencing demands are driving deployments of AI inference engines to thousands of MECs across multiple geographic regions.

End-to-End Lifecycle: Customer success requires a comprehensive solution starting with topological design patterns that enable hybrid-cloud, multi-cloud, and edge-cloud architectures to integrate with users' help desk and support apparatus for edge AI applications.

Future Proofing: Numerous edge AI use cases require several AI algorithms, but there is no consistent way to curate AI digital assets across different AI algorithm classes, such as computer vision, machine learning, natural language processing, nor are there corresponding methods for interacting with these capabilities to deliver an edge AI solution. The requirements underlying edge AI solutions constantly evolve, and the solution must be flexible to rapidly enable bringing in the multiple AI ISVs needed.

Our solution, the Ai-MicroCloud™, addresses these challenges:

- **Portability:** The Ai-MicroCloud™ can be installed on-premises, in the public cloud, and on MECs, addressing the automation and tooling gap, particularly at edge data centers (MECs)
- **Flexibility:** Ai-MicroCloud™ supports heterogeneous environments – developers using NVIDIA GPUs and CUDA software for model training can bring pre-trained models into OpenVINO™ toolkit to optimize them and model serve on a variety of Intel® architectures (XPUs), enabling the deployment of numerous AI inferences as AI APIs within a single edge server
- **Scalability:** Proprietary Ai-API Engine delivers inferences to the edge via innovative model serving, easily addressing the problem of scalability
- **Resource Depth:** Ai-MicroCloud™ enables the creation of Ai-AppStore, which solves the problem of AI algorithm aggregation – AI ISVs develop specialized algorithms, and each is different
- **End-to-End Lifecycle Management:** Ai-MicroCloud™ enables software-enabled design for any topology, infrastructure deployment, AI model development/training, AI inference optimization, and deployment of AI software, all on one cohesive platform, with AI-driven integration hooks for monitoring and help desk support
- **Versatility:** Ai-MicroCloud™ has composable foundational components and cloud-native architecture that allows for efficient and customized delivery of workflows for a variety of enterprise roles such as developer, ML Ops, product, and support
- **Speed:** Ai-MicroCloud™ provides high-performance computing (HPC) orchestration out of the box for computationally intensive model training activities that require crossing server boundary

Table of Contents

Zeblok Computational: Ai-MicroCloud™ Software-Defined AI Ecosystem for Digital Transformation 3.0

Zeblok's Ai-MicroCloud™ provides a cloud-to-edge ML DevOps platform that assists AI ISVs and hardware vendors in delivering edge AI applications at scale while supporting an entire deployment lifecycle.

Zeblok Computational serves value-added distributors, CSPs, MSPs, telecommunications, and their customers, as well as real estate operators and municipalities to create proprietary AI digital assets, aggregate multiple ISVs' algorithms in their own Ai-AppStore, optimize completed AI/ML models for heterogeneous architectures and deploy AI inferences to thousands of edge locations via its proprietary Ai-API Engine on the Ai-MicroCloud™.

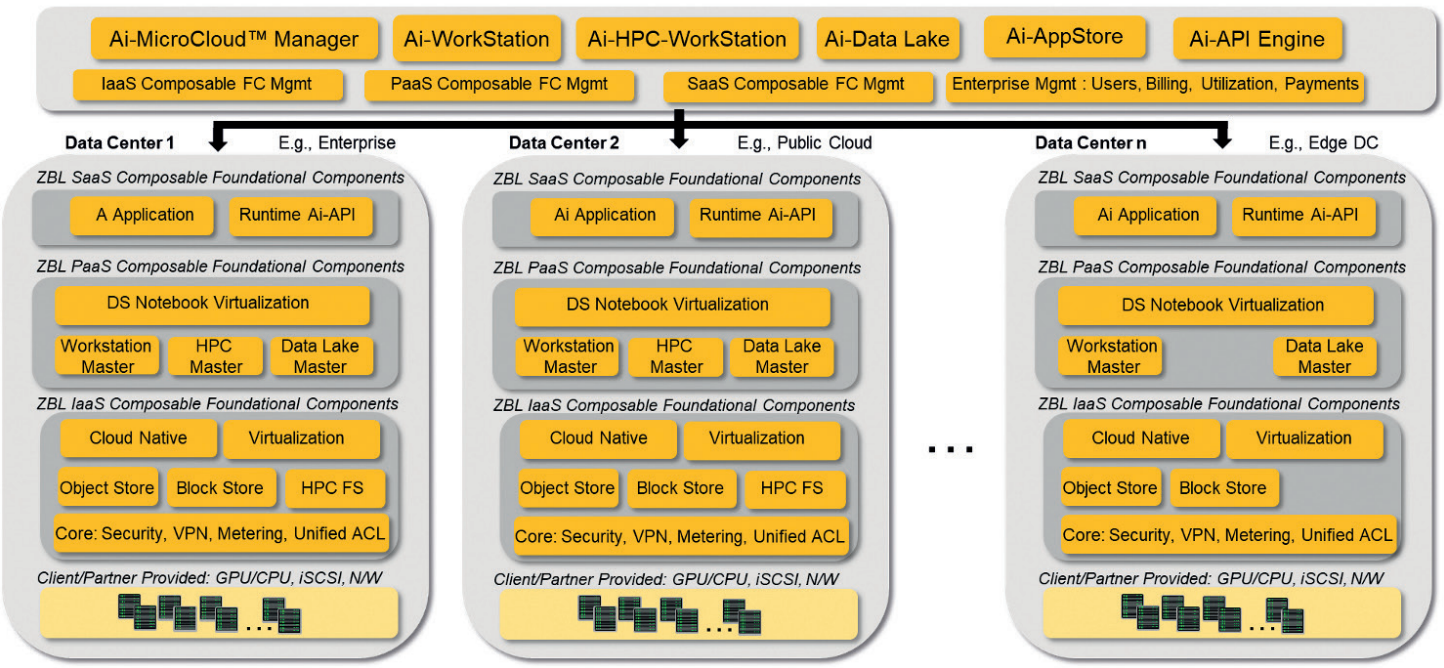
Zeblok's Ai-MicroCloud™ delivers end-to-end edge AI solutions comprising disparate technologies and hardware vendors via a software-defined AI ecosystem. Implementation is designed to fit any topology, enabling the deployment of the Ai-MicroCloud™ anywhere. Zeblok Computational's Ai-AppStore provides an express route to cost and time-efficient AI inferencing. Zeblok's Ai-MicroCloud™ allows packaging and promoting different AI/ML pipelines as AI APIs, while securely solving the software distribution challenge from cloud to edge. For those who do not have an ML DevOps platform, the Ai-MicroCloud™ also provides a fully featured AI platform for model development and training.

Zeblok Ai-MicroCloud™ Cloud-Native Scalable Software-Defined AI Ecosystem

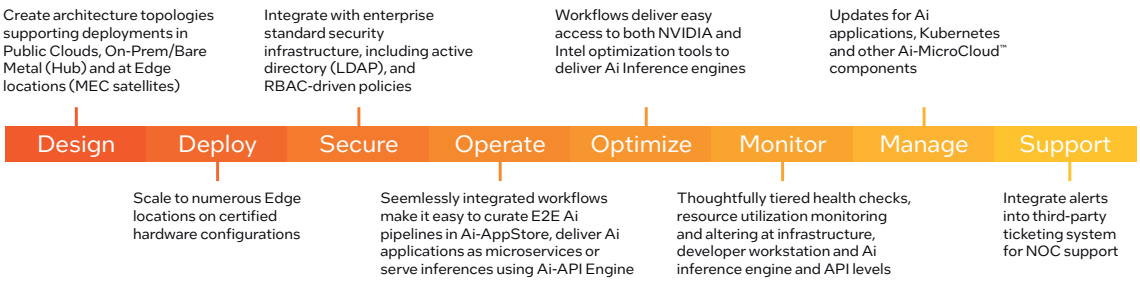
Zeblok Computational's Ai-MicroCloud™ uses cloud-native architecture, utilizing many popular open-source frameworks. In addition, simple Kubernetes (K8S) orchestrations can be extended using high-performance computing models. Key innovations include significant shrink-wrapping of cloud-based technologies as AI-middleware deployable in various environments such as kiosks, MEC Hubs, public CSPs, on-premises data centers, and different OEM manufacturer platforms.

Ai-MicroCloud™ aggregates several composable foundational components, which deliver ML DevOps, Ai-WorkStation, software distribution via AI API deployment, and workflows enabling socket-specific optimization of AI models. By integrating developer-friendly model training and distribution workflows with OpenVINO™ toolkit or oneAPI, the Ai-MicroCloud™ introduces Ai-Optimization-as-a-Service™ for enterprises. The Ai-MicroCloud™ is the only platform necessary to go from development, test, and beta to production delivery of runtime inference engines optimized for heterogeneous chipsets (CPUs, GPUs, and FPGAs) securely behind industry standard 256-bit AES encryption SSL-protected endpoint, as an L4-L7 network load-balanced service.

Absent the Ai-MicroCloud™, it can be challenging for AI ISVs and enterprise developers to access different versions of oneAPI or OpenVINO™ toolkits simultaneously. As a result, enterprise developers are left with staging these on multiple chip architectures and simultaneously reviewing the performance characteristics of inference engines in different versions of AI tools (e.g., TensorFlow, scikit-learn, etc.) or chip architecture (e.g., 3rd Gen Intel® Xeon® Scalable processors, Intel® Xeon® W processors, Intel® Movidius™ VPUs, etc.). Ai-MicroCloud™ allows developers to spin up environments within seconds to set up their AI experiments for post model training and model serving.



Ai-MicroCloud™ Enables End-to-End Edge AI Solutions – A Full Lifecycle Management Platform



Foundational Composable Components

Ai-WorkStation

Zeblok's Ai-WorkStation places all resources at the user's fingertips (see sample screenshot below). Ai-WorkStation makes it easy for data scientists to launch a Jupyter notebook with familiar frameworks such as PyTorch, R, Rapids, etc., and access proven, curated third-party algorithms in Zeblok's Intelligence Marketplace. It was built from the ground up to facilitate collaboration between developers. In addition, Zeblok's Ai-WorkStation scales seamlessly to HPCs as necessary.

Ai-Data Lake

Zeblok's Ai-Data Lake is a high-performance data store that will enable you to import, filter, and instantly analyze objects. Our solution is designed for performance, scales up with your data, and provides industry-standard (256-bit AES encryption) SSL security.

Ai-MicroCloud™ Manager

Zeblok's Ai-MicroCloud™ Manager is the graphical user interface (GUI) that allows users to install their Ai-MicroCloud™. It provides a centralized dashboard for controlling the entire environment, allowing users to define roles and access control, spawn and stop notebooks, build new infrastructure plans, meter usage of notebook resources, and monitor the health of Ai-MicroCloud™ foundational components.

Ai-AppStore

Zeblok's Ai-AppStore is a collection of hardware SKUs, public cloud SKUs, and algorithms certified for their technical and operational compatibility with the Ai-MicroCloud™ so they can be cataloged for large scale market distribution.[2]

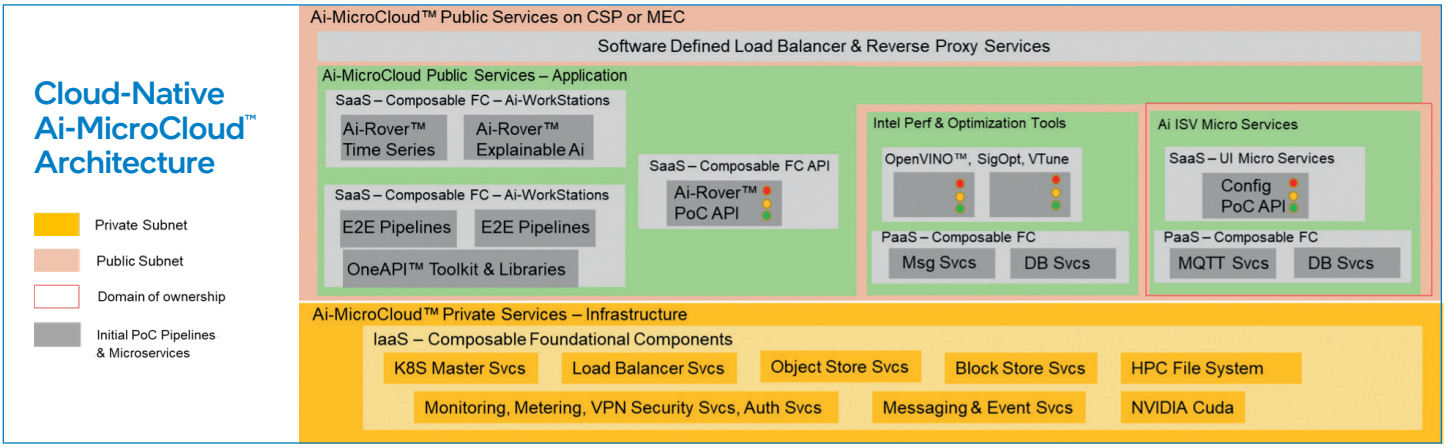
The certification program enables such distributions by ensuring that each partner is aligned with our SaaS in several key aspects:

- Technology Alignment—containerize the solution)
- Infrastructure Topology Alignment—cloud, on-prem, hybrid cloud, and edge
- Optimization Alignment—OpenVINO™ toolkit, Intel® oneAPI AI Analytics toolkit
- Orchestration Alignment—multiple algorithms able to run at a single edge location

Ai-API Engine

Zeblok's Ai-API Engine enables the deployment of a completed AI/ML inference engine as an AI API in production anywhere, supporting industry-standard protocols such as RESTful HTTPS services or Google RPC.

Cloud-Native Ai-MicroCloud™ Architecture



Load Balancer Services: Zeblok supports L4-L7 software-defined load balancers that are highly configurable and available as mesh services. Ai-MicroCloud™ consistently delivers reverse proxy and load balancer services across both internal Zeblok services with required ML DevOps and AI ISV applications. AI ISVs can create microservices of dependent components and deliver AI applications easily in a hub- and satellite-type compute infrastructure.

Object Store Services: Object store services fully implement on-premises S3 compatible object storage, which is highly reliable and supports multiple data sharing configurations.

Block Store Services: Zeblok stores a full implementation of block store, which has multiple read/write configurations like ReadWriteOnce, ReadManyWriteOnce, and ReadWriteMany. The default configuration only enables ReadWriteOnce.

HPC (High Performance Computing): For enterprises or AI ISVs that need AI HPC for model training, Zeblok HPC is fully implementing an HPC cluster, which can be scaled horizontally to support hundreds of nodes. A parallel file system also accompanies HPC to enable workloads requiring MPI, Conda, Python, and multiple AI frameworks.

Kubernetes (K8S): Zeblok Ai-MicroCloud™ utilizes K8S services based on customer preference or masters its own K8S environment to deliver container and virtualization consistently. Ai-MicroCloud™ managed Kubernetes APIs to establish any spawned microservices remain consistent across various environments such as a public cloud, on-premises, or MECs.

Intel® oneAPI AI Analytics Toolkit and OpenVINO™ Toolkit Integration as Part of the Solution

How the Industry Typically Addresses This Issue

Today, the industry suffers from fragmentation when creating and deploying outcome-based solutions due to the diverse requirements of each AI application. Enterprises focused on delivering AI solutions at the edge must provide bespoke engineering to the available network, existing hardware infrastructure, and the unique demands of specific AI algorithms.

Furthermore, each algorithm is typically developed by an independent vendor, creating further variables to manage when a single entity seeks to deploy multiple solutions.

Since many AI ISVs implement their solutions in public cloud environments, using public cloud tools, and without the consistency introduced by standard APIs such as REST API or gRPC API, industries are stuck handling complex, time-consuming bespoke integrations. Zeblok's Ai-MicroCloud™ introduces some degree of standardization in orchestration and model serving to help solve this problem at scale.

Ai-Optimization-as-a-Service™ using OpenVINO™ Toolkit and Intel® oneAPI AI Analytics Toolkit

Zeblok's Ai-MicroCloud™ provides Ai-Optimization-as-a-Service™, enabling enterprises to leverage Intel® toolkits such as OpenVINO™ toolkit and Intel® oneAPI AI analytics toolkit to enable socket-specific optimized code running at the edge as an AI API.

Working Together for a Better Future

Developers can use Zeblok Ai-MicroCloud™ to:

- Certify once and deploy at scale
- Streamline the algorithm curation process as well as help certify hardware SKUs to create their customized Ai-AppStore
- Offer innovative business models with mix & match hardware and algorithms with a scalable delivery engine
- Instantly use existing infrastructure for AI workloads
- Portability of the Ai-MicroCloud™ enables no lock-in to a public cloud environment
- Maximize resource utilization by using OpenVINO™ toolkit, enabling optimization on any Intel® hardware

Conclusion

Zeblok's Ai-MicroCloud™ enables the creation of an Ai-AppStore, and its Ai-API Engine enables deployment of AI assets to the edge efficiently, thereby helping to scale. In addition, Zeblok's Ai-Optimization-as-a-Service™ allows developers to leverage tools such as OpenVINO™ and oneAPI to provide socket-specific optimization of code running at the edge as an AI API.

AI ISV capabilities and AI algorithms are the content of the AI API economy. For companies, content delivery requires curating AI digital assets and their distribution at scale to edge data centers to deliver low latency inferencing while supporting an entire lifecycle from design to support.

OpenVINO™ Toolkit

OpenVINO™ toolkit enables you to optimize, tune, and run comprehensive AI inference using the included model optimizer, runtime tools, and development tools.

OpenVINO™

1 https://www.iotworldtoday.com/2018/10/22/arms-iot-platform-for-a-1-trillion-device-world/

2 Hardware certification details: https://www.computational.zeblok.com/hardwarecertification
ISV certification details: https://www.computational.zeblok.com/isvcertification

Performance varies by use, configuration, and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software, or service activation.