Zeblok
COMPUTATIONAL

Why Bring AI Stack to Data?

A NARRATIVE FOR DATA CENTER OPERATORS

BACKGROUND

In about a decade we have gone from $500M cluster to $1B cluster to right now hearing about $1T cluster for AGI, with companies hurrying to lock in power contracts to run these massive GPU clusters.
0.5 Order of Magnitude (OOM) improvement per year is the projection leading to projections such as PhD level intelligence by 2027[source: situational-awareness.ai]. The quality of open-source models are dramatically improving. The evolution here is largely focused on consumers adopting and utilizing AI in daily life, and adoption of such models by the largest players in delivering consumer experience. i.e. B2C

Enterprises i.e. B2B have a different problem. They want the innovation but there are serious challenges in protecting data. They operate in more controlled regimes and have a different set of challenges. Their challenge areas are:

RISK MANAGEMENT

1. Data Leakage Risk

- **Avoided**: Sensitive or proprietary enterprise data being sent to third-party servers (e.g., OpenAI, Anthropic, etc.).
- **Why it matters**: Enterprises often handle confidential customer data, trade secrets, and regulated information that must not leave internal infrastructure.

2. Regulatory and Compliance Risks

- **Avoided**: Violations of data residency laws, industry regulations (e.g., HIPAA, GDPR, PCI-DSS).
- ·**Why it matters**: On-premise deployment allows full control over data flow and auditability, helping meet strict compliance standards.

3. Third-Party Dependency and SLA Risks

- **Avoided**: Downtime or latency caused by external service providers or API outages.
- ·**Why it matters**: Critical enterprise operations may depend on the model's availability, and relying on third-party uptime could introduce vulnerabilities.

4. Security Risks from Multi-Tenancy

- **Avoided**: Potential attack surface from shared cloud environments (e.g., data leakage between tenants or abuse of shared infrastructure).
- **Why it matters:** Enterprises may be wary of co-tenancy with unknown or untrusted entities.

Product Family

Ai-MicroCloud®

Ai-Domain Copilot

Ai-Rover

Ai-Quantum Tunnel®

Important Use Cases

- Hybrid AI Cloud across multiple data centers
- Hub to Edge AI across multiple edge locations all under a singe pane of glass
- AI Side Car ecosystem that aggregates a variety of ISV AI assets and open sources software for rapid integration and assimilation into entperise applications.

© Zeblok Computational, Inc. 2025. All rights reserved. All third-party logos and trademarks are the property of their respective owners.

Zeblok
COMPUTATIONAL

Unique combination of Risk Management and Deployment Management creates a need for AI Stack closer to Data

5. Vendor Lock-in
- **Avoided**: Being tied to a specific LLM provider's pricing, APIs, or ecosystem.
- **Why it matters**: Running models internally gives enterprises flexibility to switch models, customize, or build on top of open-source solutions.

6. Telemetry and Usage Data Exposure
- **Avoided**: Model usage data (prompts, metadata, behavior patterns) being collected by external vendors.
- **Why it matters**: This data could unintentionally reveal internal processes, business priorities, or security postures.

7. Custom Model and Data Integration Challenges
- **Avoided**: Complexity or restrictions in integrating with internal data sources or systems (e.g., ERPs, knowledge bases).
- **Why it matters**: On-prem deployments allow tighter, more secure integrations with existing infrastructure.

DEPLOYMENT MANAGEMENT

1. Infrastructure & Scalability
- **Compute Requirements**: AI models, especially large language models (LLMs), are resource-intensive. Procuring and managing GPUs or high-performance infrastructure is costly and complex. Running these GPU system is a power challenge.
- **Scalability**: Ensuring the system can handle increased usage or real-time inference without degrading performance.
- **Latency & Availability**: Meeting enterprise-grade SLAs for responsiveness and uptime can be hard, especially for on-prem or hybrid setups or perhaps even Edge setups.

2. Cost & TCO & ROI Uncertainty
- **High Upfront Investment**: Cost of infrastructure, talent, and tooling can be high.
- **Unclear ROI**: It can be hard to quantify value early on, especially in R&D or experimental phases.
- **Unclear TCO**: Monitoring, maintenance, and model retraining can increase total cost of ownership, especially when paying with arcane Token based models.

3. Integration with Business Systems
- **Legacy Systems:** Integrating AI with older platforms (like SAP, Oracle, SharePoint, etc.) can require significant engineering
- **APIs & Data Pipelines**: delivering new AI capabilties from a variety of internal teams and external partners consistently as an API or pipeline so it can be integrated into exising applications.

© Zeblok Computational, Inc. 2023. All rights reserved. All third-party logos and trademarks are the property of their respective owners.

Zeblok COMPUTATIONAL

There is a critical gap between enterprise customers and data center partners offering

CUSTOMER PERSONA CHALLENGES WITHIN ENTERPRISE

- Compliance department doesn't want the data to be moved into hyperscaler or feed data into models in the public domain due to risks identified in the background, but the project team needs to deliver AI based capabilities into existing applications.
- They may have procured GPUs but it is complex to run and manage the AI Stack that effectively utilizes and scales models of different types across GPUs
- It is NOT just about scaling GPUs, the data has to be feed to it. Very soon it becomes an object store problem and high performance computing problem and many more.
- They may have built some prototypes but scaling for larger number of users is a problem
- Data science is integral to their business insights but the AI engineering for the AI Stack takes 80% of the time and it is not core to their business
- Budget is allocated for GenAI projects but have a hard time forecasting the expenses when AI models in hyperscaler are utilized because the Token based costs on hyperscalers is not easy to comprehend. Summary, chat, search type use cases utilize a lot of tokens and it is very hard to predict expenses. Customer wants a predictable cost for the project.
- Running larger footprint of GPUs in their data center is a problem because of the power problem
- Customer potentially has credits with hyperscalers as well but are NOT able to utilize it because AI Stack is different. It is back to drawing board from engineering point of view.
- Sometimes the computer vision inferences have to be deployed in Edge environments, now AI stack is needed at the Edge
- Sometimes disconnected Edge environments are needed, this adds more complexity to AI stack at the Edge

DATA CENTER PARTNER PERSONA CHALLENGES

- They've already made investment into GPU infrastructure
- They don't look at themselves as just a real-estate company.
- They are interested in establishing Customer private AI Cloud in their data center.
- They've not figured out how to offer a hybrid model where customers may have their own GPU servers on premise and customer wants to use both capacities
- Their customers need EdgeAI solution as well because just centralized model doesn't meet the customer needs e.g. computer vision.
- What starts as computer vision easily becomes other AI use cases, like GenAI that may sell more GPUs, so the data center partner wants to win that Edge AI business.
- So the hub has to run in a primary DC environment and satellite in Edge locations where inference and applications have to be delivered for lower latency
- The moment EdgeAI comes into picture then Multi-Access Edge Compute (MEC) management/orchestration is important, and potentially 5G mobility may come into the picture e.g. 5G Private Mobile Network or 5G Mobile Adhoc network.
- Now the DC Partner has focused on GPUs all along, but it needs to expand offering to support the deployment of vision inferences at the Edges and without that solution they cannot monetize the GPU investments effectively.
- Their product team struggles to define the product offering to monetize
- They need hybrid private AI cloud solution
- They need hub-to-edgeAI solution and potentially need 5G mobility solution for the Edge

© Zeblok Computational, Inc. 2023. All rights reserved. All third-party logos and trademarks are the property of their respective owners.

Zeblok
COMPUTATIONAL

There is a critical gap between enterprise customers and data center partners offering

OUR SOLUTION

- A new offering for the DC operators through Zeblok
- Allows them to stay focused on their core infrastructure-as-a-service business while leapfrogging into AI-Cloud business with minimal investment
- Customer comes into the DC marketplace and triggers the creation of Ai-MicroCloud® (Pls see data sheet on **Ai-MicroCloud®** - Bringing AI Stack to Data)
- Customer will find workflows in aggregating a variety of AI assets into Ai-MicroCloud® and workflows, SDK, CLI for power users for automating lifecycle management of AI application. E.g. ML Dev Sec Ops, Containerization and deployment of a variety of models, pipelines, apis, and microservices/ applications.
- Customer will find Ai-MicroCloud® to be an AI Side Car for aggregating and releasing AI apis, pipelines and microservices that can be readily integrated into existing enterprise applications.

Product Family

Ai-MicroCloud®

Ai-Domain Copilot

Ai-Rover

Ai-Quantum Tunnel®

- Customer, if needed, could attach GPUs in their premises along with an extension of their data center in the data center partner premises, while choosing where to keep the data.
- Customer, if needed, could attach resources within their hyperscaler VPC
- Customer gets to decide where the data needs to be and AI Stack is available near data
- Optionally, if customer wants to deploy inferences and applications in the Edge locations then they can select **Ai-MicroCloud® multi-site edition** from the DC market place. Edge Servers can be attached and a single pane of management glass available with operational workflows.
- Optionally, when highest level of encryption is needed by the sites, i.e. hybrid site or edgeAI, then **Ai-QuantumTunnel** can be utilized to encrypt data with Post Quantum Cryptography on top of regular VPN channel.
- Optionally, if the Edge Servers are to work in private mobile networks or adhoc mobile networks then **Ai-Harmony** product can be added from the DC market place.
- Optionally, where the primary focus of the customer is GenAI, then **Ai-DomainCopilot** simplifes running LLMs. Chat, summary, knowledge distillation, RAG etc could be done at a predictable cost because it all runs fully within Ai-MicroCloud®. Customers can event pull down models from hugging face and run it locally. Customers can simply fine tune domain specific smaller models and they own the models created. Smaller models are more power efficient and less expensive to operate.
-

© Zeblok Computational, Inc. 2023. All rights reserved. All third-party logos and trademarks are the property of their respective owners.