micron

Hey AI:
Say bye to lag

10 considerations for a winning cloud-to-edge AI strategy

Winning the edge AI race starts now

Data centers and IT organizations like yours have been tasked with implementing AI to improve business and the bottom line. Now.

Success means implementing a cloud-to-edge AI strategy that moves business-critical AI workloads to where they can most effectively and efficiently harness the power of your data. Moving select workloads to the edge can meet and exceed your leadership's expectations of what AI can do for your business.

AI and the edge fit together naturally, since moving AI workloads to the edge can provide real-time insights, reduce data transport costs, and lower power consumption.

However, the question arises: Which AI workloads should you move to the edge?

Choose use cases that optimize GPU usage, data egress, and power consumption

Smart retail

- Analyze customer behavior, manage inventory, and personalize shopping
- Deliver insights instantly, driving better decision-making on the sales floor
- Optimize operations, improve customer experience, and increase sales

Computer vision

- Gain real-time processing and low latency for computer vision workloads
- Reduce the time to act on visual data, leading to faster response times
- Lessen the need for bandwidth-intensive data transfers to the cloud

Predictive maintenance

- Monitor devices to help prevent equipment failures and minimize downtime
- Access real-time analysis of sensor data and quickly identify many potential issues
- Reduce the latency and data egress costs associated with cloud-based operations

NLP

- Enhance interactions between humans and machines with real-time inferencing
- Optimize solutions such as voice-activated assistants and language translation
- Achieve smoother, more responsive interactions, and more efficient data inferencing

micron

Finding balance between cloud and edge

When moving AI workloads to the edge, you need to balance performance, budget, and power consumption.

Latency

For some workloads, moving to the edge can reduce latency, which in turn can improve customer experiences, make safer work environments, decrease downtime, and provide real-time insights. Other workloads don't rely as heavily on low-latency performance, making them more suitable for the cloud.

Data transport

Cloud bills can skyrocket if the volume of data transport gets too high. Edge AI can reduce the strain by processing most of the data locally, and only transferring the essentials to the cloud.

Resource efficiency

Lightweight workloads can often be moved to the edge to run more efficiently. At the same time, deploying edge AI devices can be costly, leading to compromises about how to balance performance and efficiency.

Security

Cloud systems can provide suitable security for a range of workloads. However, there are some situations where edge servers provide a necessary extra layer of security to comply with security regulations.

Edge computing is the epicenter for your AI success

Winning the AI race requires a cloud-to-edge strategy that distributes key workloads to where data can be most effectively collected, analyzed, and utilized.

Edge strategies can:

- Optimize GPU utilization
- Improve data security
- Reduce costs and power consumption

Moving inferencing and model retraining and tuning to edge servers allows organizations to harness the power of AI at the epicenter of where data is created.

micron

10 considerations for a winning cloud-to-edge AI strategy

1. What is the primary goal of the workloads?

Real-time processing with low latency → EDGE

Large-scale data processing and complex model training → CLOUD

2. How sensitive is the workload to latency?

Very sensitive → EDGE

Not sensitive → CLOUD

3. How much data needs to be transported?

High data volumes

Can data be processed locally to reduce transport costs?

Yes → EDGE

No → CLOUD

Low data volumes → CLOUD

4. How important is data privacy and security?

Highly sensitive data

Can data be processed locally to ensure privacy?

Yes → EDGE

No → CLOUD

Less sensitive data → CLOUD

5. How frequently does the model need retraining?

Frequent retraining → EDGE

Infrequent retraining → CLOUD

6. What are the compute and power requirements?

High compute power and continuous power supply → EDGE

Limited power or compute → CLOUD

7. Are there regulatory or compliance requirements?

Strict regulations on data location → EDGE

Flexible regulations → CLOUD

8. What is the cost structure of the solution?

High cost for cloud transport/storage → EDGE

High cost for deploying edge devices → CLOUD

9. Does the workload require continuous connectivity to the cloud?

Yes → EDGE

No → CLOUD

10. What are the retooling requirements?

Minimal retooling or reconfiguration needed for edge deployment → EDGE

Significant retooling required → CLOUD

Overcome the competition

AI and edge are powerful, but they can be hard for businesses to implement. Your competitors are struggling with a lack of in-house AI experience, and many are deploying AI solutions that fail to meet leadership's expectations.

Micron can give you an AI advantage that propels you ahead of your competitors. Micron's AI experts rigorously test and optimize memory and storage to optimize GPU utilization and improve AI workload performance. We can work with you to better understand AI's potential (and limitations), and deal with unreasonable project expectations that can be a barrier to success.

Learn more at micronpg.com/edgeAI