

**Title: Driving Factors for Distributed Inference Sites (<1MW)**

**Sub-Title: Enabling more efficient, responsive, and secure deployment of AI applications across various industries and use cases.**

*(Flashing on booth screen as independent bullets under title and sub-title)*

* Latency Reduction: Distributed inference sites enable AI models to process data closer to the source, reducing latency. This is particularly important for applications requiring real-time responses, such as autonomous vehicles, industrial automation, and healthcare diagnostics.
* Bandwidth Savings: By performing inference locally, the need to transmit large volumes of data to a central server is reduced. This conserves bandwidth and can result in significant cost savings, especially in environments with limited or expensive network resources.
* Scalability: Distributed inference allows for more scalable deployment of AI models. Instead of relying on a single, centralized server to handle all inference requests, the workload can be distributed across multiple locations, balancing the load and improving performance.
* Privacy and Security: Local processing of sensitive data enhances privacy and security by minimizing the exposure of data to potential breaches during transmission. This is crucial in sectors such as healthcare, finance, and any industry handling sensitive personal information.
* Reliability and Fault Tolerance: Distributed systems can be more resilient to failures. If one inference site experiences issues, others can continue to operate, ensuring continuous availability and reducing the risk of a single point of failure.
* Edge Computing Growth: The rise of edge computing, where processing is done at the edge of the network, aligns with the need for distributed inference. Edge devices, such as IoT sensors and smart cameras, often have the capability to perform AI inference, facilitating the shift from centralized processing.
* Customization and Adaptability: Distributed inference allows for models to be tailored and optimized for specific locations or devices, leading to better performance and adaptability to local conditions and requirements.
* Cost Efficiency: Utilizing local hardware for inference can be more cost-effective than maintaining large centralized data centers. This is particularly true as edge devices become more powerful and cost-efficient.
* Regulatory Compliance: Certain regulations require data to remain within specific geographic boundaries. Distributed inference helps comply with such regulations by processing data locally, ensuring it does not leave the designated region.
* Technological Advances: Advances in hardware, such as GPUs, TPUs, and other specialized AI accelerators, have made it feasible to deploy powerful inference capabilities in distributed environments, including edge devices and smaller data centers.