Unlocking organizational productivity through the power of Hybrid AI

Executive Summary

As organizations look for new ways to drive greater operational efficiency and scale while maintaining high standards of data security, Hybrid AI[1] is emerging as a new lever for enhancing organizational productivity. But why are organizations, many of which have already begun their AI journeys, now interested in adopting Hybrid AI?

Hybrid AI enables organizations to unlock real-time data insights at the edge while leveraging the computational power and scalability of the cloud by integrating cloud computing with edge AI. This dual approach reduces costs, decreases latency, optimizes bandwidth, and enhances data security—helping overcome the common limitations of traditional, cloud-only AI systems[2]. These capabilities can prove highly valuable for decision makers who are constantly tasked with driving efficiencies.

Hybrid AI allows organizations to distribute AI workloads to meet the specific needs of different business units. By improving resource utilization and creating faster iteration cycles, organizations can further enhance productivity. From customer service, where AI-powered agents can deliver real-time personalized support, to predictive maintenance in manufacturing and logistics, Hybrid AI drives faster, smarter decision-making while lowering costs. The range of potential use cases is broad. Further, for those who have started their AI journeys but have yet to fully maximize the ROI of these investments, Hybrid AI can offer a robust framework for scaling and optimizing AI initiatives while increasing cost efficiencies.

This white paper defines Hybrid AI and provides various frameworks for how IT decision-makers (ITDMs) can approach adopting, transitioning to, and creating value with Hybrid AI. It outlines the essential components of a Hybrid AI tech stack, offers tips for efficient implementation, provides a framework for evaluating the Hybrid AI opportunity and deploying Hybrid AI, and explores the latest technical advancements (like NPUs and energy-efficient infrastructure) that will continue making Hybrid AI a compelling productivity and value enhancing opportunity for businesses[3].

Defining Hybrid AI

Hybrid AI is the strategic combination of public / private cloud and on-device processing, designed to optimize AI performance across diverse environments. Cloud-only AI systems are constrained by high costs, latency, bandwidth limitations, and security risks. Hybrid AI can solve these challenges by dividing tasks: the cloud is utilized for compute intensive tasks (cloud can often manage large computations), while real-time inference and decision-making can happen locally on devices or at the edge.

This approach is critical in industries requiring real-time data analysis and where fast response times are essential, for example predictive maintenance and customer service. With models running locally, organizations can act instantly on data without the delays or costs of sending data to the cloud. On the other hand, the cloud supports scalability for tasks such as inferencing on large scale models that require high computational power.

By merging cloud scalability with edge responsiveness, Hybrid AI enables real-time data analysis, eliminates bottlenecks, and improves overall productivity in distributed environments like remote sites[4]. With Hybrid AI, CIOs, Chief Data Officers, and AI Architects can customize their AI infrastructures to better align with their business needs and to balance compute-intensive tasks and real-time processes more effectively.

Hybrid AI unlocks the following productivity outcomes:

1. **Increased agility:** Developers can easily iterate on and test AI models across cloud and edge environments, significantly speeding up AI deployments and fostering rapid innovation.
2. **Improved time to insights:** Insights can be generated faster by leveraging domain-specific models and real-time data on devices, facilitating quick decision-making and more personalized customer experiences.
3. **Automation of routine tasks and optimization of workflows:** Hybrid AI-enabled automation streamlines routine tasks, freeing up employees for higher-value work. For example, in the logistics industry, edge devices can track and manage inventory in real-time, while cloud-based AI models optimize delivery routes and schedules, reducing operational costs while improving delivery times.

Hybrid AI also delivers the following benefits:

- **Cost savings:** Reduces expenses associated with data transfer, storage, and processing and optimizing resources through smaller, task-specific models at the edge.
- **Data privacy:** Enables organizations to process sensitive information locally, reducing the risk of exposure during data transfer[5].
- **Energy efficiency:** Processes data locally and minimizes energy consumption (i.e., liquid cooled GPUs used for cloud energy consumption), making operations more sustainable,
- **Scalability:** Expands to meet demand, ensuring consistent performance across mobile devices, private data centers, and edge environments[6].

Overall, as businesses grapple with more data, rising costs, privacy concerns resulting from a growing amount of sensitive information, and more work being done off-prem, a Hybrid AI approach becomes a compelling path forward.

A Technical Framework for Hybrid AI

The Hybrid AI tech stack integrates multiple layers, each optimized to manage specific tasks across cloud, edge, and secure connection environments. This framework orchestrates workloads based on where they are best executed, ensuring that AI systems deliver low-latency insights, secure data transfer, and high-performance computing across distributed environments. The breakdown below introduces the interconnected layers and their core components, which work together to support seamless AI-driven operations.

1. **Edge Layer:** The edge Layer is the foundation of Hybrid AI, capable of handling real-time AI inference and localized data processing close to the data source. Diverse types of edge devices—such as IoT sensors, edge servers, and AI clients —are optimized to perform distinct roles within this layer. While AI PCs and AI workstations are not typically classified as edge devices, they can provide localized AI processing for specific use cases, especially in distributed or remote environments.

2. **Secure Connection Layer:** The Secure Connection Layer ensures secure and efficient data flow between the cloud and edge environments. It manages encryption, data synchronization, and network performance to maintain data integrity. Utilizing secure infrastructures like VPN and SD-WAN, it establishes encrypted communication channels, while data encryption tools protect transmitted data. Network management optimizes bandwidth and connection stability for smooth data transfer. This layer ensures sensitive data is securely transferred, maintaining privacy, regulatory compliance, and data integrity across environments.

3. **Cloud Layer:** At the top of the Hybrid AI framework, the cloud Layer manages resource-intensive tasks such as AI model training, advanced analytics, large scale model inferencing, and large-scale data storage. GPUs and TPUs provide the computational power for inferencing when required performance and accuracy are not possible at the edge. Centralized data lakes aggregate data from the edge for long-term analytics, while APIs ensure seamless communication and data exchange between cloud and edge environments. By offloading complex computations to the cloud, this layer ensures real-time inferencing capabilities can meet any organizational requirement.

“Different types of edge devices—such as IoT sensors, edge servers, and AI clients —are optimized to perform distinct roles within the edge layer.”

By deploying AI models to these varied edge devices, **the system can perform real-time analysis and generate actionable insights tailored to specific use cases**. By minimizing cloud reliance, the Edge Layer reduces latency, enhances operational responsiveness, and enables faster decision making.

- **IoT sensors** typically **gather and process small-scale, high-frequency data directly at the source**, allowing for real-time monitoring and control in environments like manufacturing or logistics.

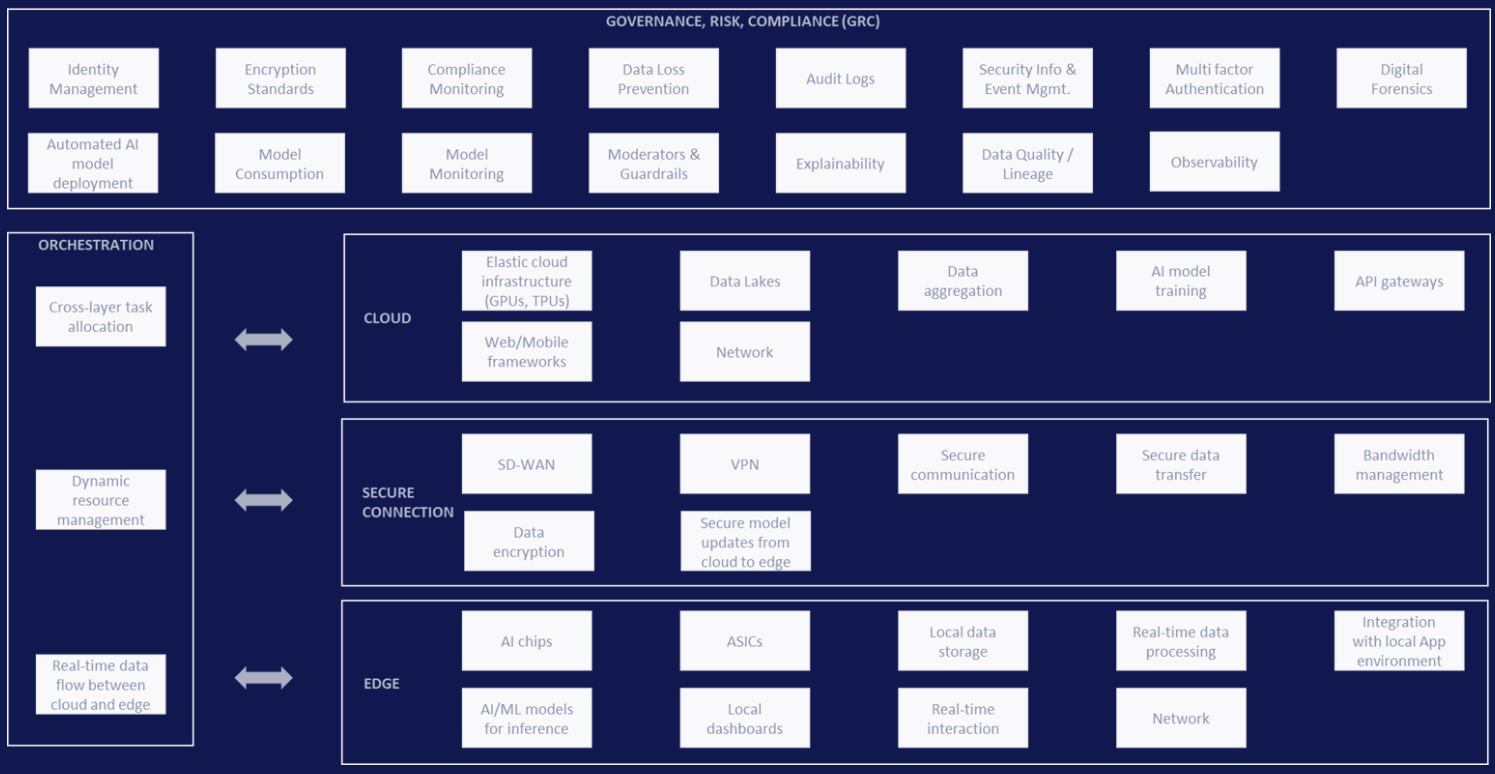
- **Edge servers and AI clients** provide scalable processing power and real-time AI inference capabilities close to the data source, **enabling efficient processing without the latency and dependency on the cloud**. Edge servers can also aggregate multiple data streams for more comprehensive analysis, while AI clients execute less computationally intensive tasks independently.

- **AI PCs and AI workstations** provide advanced on-device AI processing capabilities and manage more complex inferencing tasks when required. Although not traditional edge devices, they are **ideal for executing AI tasks autonomously in remote locations or environments with limited connectivity to the cloud**, such as distributed work environments, research labs, or advanced manufacturing setups

4. **Orchestration:** Orchestration is the core management layer that integrates cloud, edge, and secure connection environments. This automates resource allocation, task distribution, and data synchronization. It makes real-time decisions to compute at the edge or cloud layers, referencing the required model, network capabilities, and cross layer performance metrics. The orchestrator dynamically allocates tasks based on device type and workload demands, choosing lightweight IoT devices for routine monitoring and prioritizing more complex tasks for AI PCs, workstations, or cloud. By dynamically managing resources and maintaining data consistency, orchestration optimizes system performance and ensures compliance. This automation drives productivity by reducing latency, preventing resource bottlenecks, and enabling more efficient decision-making.

5. **Governance, Risk, Compliance (GRC):** The GRC layer ensures the Hybrid AI system operates securely and complies with regulatory standards. It governs data privacy, access controls, and risk management across cloud, edge, and secure connection environments. It enforces strict encryption, implements Zero Trust architecture, and continuously audits compliance with global regulations. Key functions include protecting sensitive data, limiting access with MFA and RBAC, and managing risks by identifying vulnerabilities and enforcing real-time mitigation. By safeguarding data integrity and ensuring proper lifecycle management, the GRC layer enhances productivity by reducing compliance risks and minimizing operational disruptions.

Figure 1: Hybrid AI Technology Stack



The Hybrid AI technical framework integrates critical components across the edge, secure connection, and cloud layers to create a scalable, efficient AI system. Each component operates across layers, optimized for performance, real-time processing, and seamless data flow.

The key components integrated across layers include:

1. **Infrastructure:** The edge infrastructure is designed for real-time AI inference and low-latency data processing, supporting both vertical and horizontal scaling to manage dynamic workloads. IoT devices, such as sensors, manage lightweight, high-frequency data at the source, while AI PCs offer on-device AI processing for autonomous tasks, reducing cloud dependency. Edge servers and AI clients provide scalable processing power and real-time AI inference capabilities close to the data source, enabling efficient processing without the latency or reliance on cloud infrastructure. Edge servers aggregate multiple data streams for more comprehensive analysis, managing moderate computational tasks while reducing data transfer needs. Meanwhile, AI clients execute less computationally intensive tasks independently, ensuring distributed AI processing with minimal latency.

> "AI workstations manage more complex inferencing tasks in environment required greater computational capacity"

AI workstations manage more complex inferencing tasks in environments requiring greater computational capacity, such as advanced manufacturing or research labs. The cloud infrastructure supports larger AI tasks that exceed the edge's capacity, including large-scale model inferencing, model training, and data-intensive analytics. It provides elastic compute and storage resources, ensuring scalability to meet varying AI demands. This tiered infrastructure supports a hybrid AI ecosystem, balancing edge, and cloud resources to optimize performance and minimize latency across different AI workloads.

2. **Data:** At the edge, data is processed locally for immediate insights, with real-time classification and compression to optimize storage and reduce transmission to the cloud. While edge environments can support localized "mini-data lakes" to store and process data close to the source, these edge data stores are designed for short-term, high-frequency data handling rather than large-scale analytics. This localized storage enables faster decision-making without constant cloud interaction, but it relies on real-time synchronization with cloud systems for broader, long-term data analysis. In the cloud, aggregated data from the edge is stored in centralized data lakes, enabling deeper analytics, AI model training, and long-term storage. Ensuring data consistency across environments guarantees reliable analytics and processing.

3. **Models:** At the edge, AI models are deployed for real-time inference and optimized for low-power environments to ensure rapid decision-making with minimal resource usage. In the cloud, AI models are trained on large datasets using distributed training frameworks, which enables continuous optimization. Updated models are deployed back to the edge for real-time inference.

4. **APIs:** APIs facilitate real-time, low-latency data exchanges between the edge and cloud, ensuring seamless communication and integration. These APIs send the necessary data and configurations to the cloud layer based on orchestrator decisions; then can receive generated output for on device presentation. Simultaneously, APIs will interact with cybersecurity platforms to ensure secure data transfer.

Technology Tech Stack Considerations

To optimize performance across the Hybrid AI tech stack, IT Decision Makers (ITDMs) must take a strategic approach to managing the Edge Layer, Secure Connection Layer, and Cloud Layer while ensuring robust Governance, Risk, and Compliance (GRC). Each of these layers must be optimized to manage dynamic workloads, real-time processing, and secure data flows. Below are some technical considerations for each layer, and how they impact productivity.

EDGE LAYER

Technical Considerations

- Deploy AI-optimized hardware that supports scalable real-time inference and processing
- Use containerized AI models to enable decision-making directly on edge devices, reducing cloud dependence
- Apply compression and deduplication to optimize local data storage and minimize transfer volume to the cloud
- Implement edge data lakes for short-term, localized data storage and processing

Productivity Impact

- Localized processing reduces latency, enabling immediate responses and boosting operational efficiency
- Minimizing cloud interaction lowers bandwidth usage and cloud costs, improving overall efficiency
- Edge data lakes improve operational speed by storing and processing data close to the source, allowing rapid data access and insights without cloud delays.

SECURE LAYER

Technical Considerations

- Prioritize SD-WAN and traffic-shaping for low-latency, high-bandwidth communication
- Ensure all data transfers between edge and cloud are protected with TLS 1.3, VPN, and Zero Trust security
- Dynamically allocate bandwidth to prioritize critical data transfers

Productivity Impact

- Optimized network and bandwidth management enable real-time data synchronization, enhancing decision-making speed
- Secure data transfers reduce security risks, ensuring smooth operations and high productivity

CLOUD LAYER

Technical Considerations

- Use Kubernetes, GPU/TPU autoscaling, and IaC for responsive scaling based on workload demands
- Implement distributed frameworks to accelerate training and model deployment to edge devices
- Ensure real-time synchronization and use distributed storage for long-term processing and analytics

Productivity Impact

- Elastic infrastructure reduces training times, enabling faster deployment and real-time insights
- Autoscaling ensures performance under varying data loads, preventing downtime and optimizing operational efficiency

GOVERNANCE, RISK, AND COMPLIANCE (GRC) LAYER

Technical Considerations

- Implement least-privilege access, MFA, and continuous monitoring across all layer
- Use AI-driven tools for real-time monitoring and continuous auditing
- Enforce encryption and data governance policies to maintain compliance with global regulations

Productivity Impact

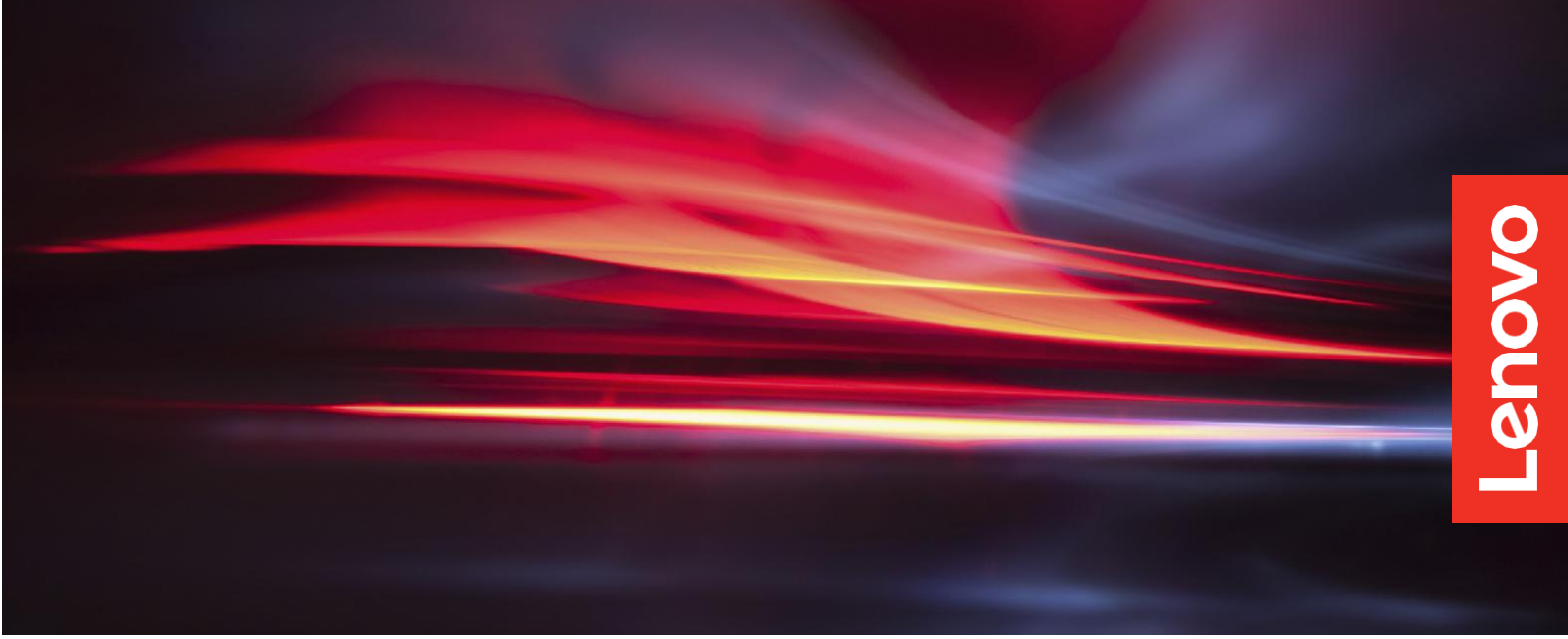
- Real-time auditing reduces manual effort, allowing teams to focus on critical tasks
- Continuous compliance minimizes disruptions, ensuring uninterrupted productivity

Figure 2: Key technical considerations

Even after optimizing the Hybrid AI tech stack using the considerations outlined in Figure 2, organizations may still encounter barriers to success. A key challenge many organizations face is imbalanced investment across the technology stack. Overinvesting in certain layers (e.g., cloud infrastructure) while neglecting others (e.g., edge optimization or secure data transfer) can create performance bottlenecks and operational inefficiencies. This fragmented approach can lead to:

1. **Resource Overload or Underutilization:** If resources such as GPUs or TPUs are concentrated only in the cloud layer, real-time decision-making at the edge may suffer from latency and inefficient resource usage. Similarly, deploying models that are not optimized for production at the edge can result in latency or performance degradation, limiting the value of real-time AI inference.
2. **Data Synchronization Issues:** Without seamless synchronization between the edge and cloud layers, organizations face data consistency issues. Frequent causes of data sync issues include network interruptions, inconsistent data formatting between edge and cloud, or latency in edge storage updates. This could result in incomplete or delayed insights, compromising the accuracy of AI models and operational responses.
3. **Network Bottlenecks:** Inadequate investment in network optimization (e.g., lack of SD-WAN or traffic-shaping) can lead to high-latency, low-bandwidth data transfers, creating significant delays in real-time data flow between cloud and edge environments.
4. **Compliance and Security Gaps:** Failing to implement a comprehensive Zero Trust architecture and real-time compliance auditing can leave the system vulnerable to security breaches and non-compliance with regulations, leading to potential downtime or data loss.

To overcome these barriers, ITDMs must adopt an integrated approach that balances investment across all layers of the stack. This includes ensuring dynamic resource allocation at the edge and in the cloud, real-time data synchronization, secure and optimized data transfer, and a robust GRC framework that governs security and compliance across environments. By aligning the technology stack with strategic business needs, organizations can ensure their Hybrid AI infrastructure is agile, scalable, and resilient, capable of handling dynamic workloads and evolving AI demands.



Evaluating the Hybrid AI Opportunity

Before implementing Hybrid AI, organizations must start with clearly defined objectives. Leaders must first define their Hybrid AI goals—whether it is improving operational efficiency, reducing latency, enhancing real-time decision-making, or scaling AI-driven processes. Establishing specific tasks, performance metrics, and integration goals will help inform whether Hybrid AI is accretive to the broader business vision.

Once these objectives are identified, the next step assessing multiple business and technical parameters and their implications on Hybrid AI feasibility. The framework below can help decision-makers evaluate the Hybrid AI opportunity relative to their unique circumstances.

Business Parameters

Business parameters focus on the strategic and financial levers of adopting Hybrid AI, ensuring that the decision aligns with the organization's overall goals.

Decision Parameter	When to adopt Hybrid AI	When not to adopt Hybrid AI
Cost-Benefit Analysis	When it is feasible to finance the required initial CapEx, and long-term savings are expected through optimized resource allocation	When high upfront costs create financial strain and no plans to scale cloud consumption dilute ROI
Talent and Expertise	When skilled AI talent is available, or the organization can invest in acquiring expertise to manage Hybrid AI	When a steep learning curve or talent shortage creates significant barriers to implementation
Vendor and Ecosystem Alignment	When there is availability of vendors with Hybrid AI solutions that integrate well with the organization's existing systems and provide long-term support	When vendor lock-in or lack of interoperability between Hybrid AI components hinders scalability and flexibility

Figure 3a: Key considerations for whether to deploy Hybrid AI

Technical Parameters

Technical parameters address the operational and infrastructural considerations of adopting Hybrid AI, ensuring that the technical environment can support the new system.

Decision Parameter	When to adopt Hybrid AI	When not to adopt Hybrid AI
Data Security and Compliance	When managing sensitive data locally is essential for meeting security standards and regulatory requirements	When managing security across cloud and edge introduces excessive complexity or compliance risks
Workload Type	When there are varying computation workloads that can utilize real time decisions around which layer to do inferencing.	When workloads need to be exclusively computed in either a cloud or on-premises environment.
Model Requirements	When the AI models, regardless of size (small to large), are optimized for deployment across edge and cloud environments, leveraging techniques like model compression or quantization for real-time edge inference while maintaining full-scale models in the cloud for training	When the models require extremely high computational resources for real-time inference and cannot be sufficiently optimized for edge deployment without sacrificing performance
Maintenance and Support Complexity	When the organization has the capacity to manage complex hybrid systems with ongoing maintenance and updates across cloud and edge	When maintaining and supporting a hybrid environment adds excessive operational complexity and overheads
Scalability and Flexibility	When dynamic scaling across cloud and edge is critical for handling varying workloads and business needs[7]	When scaling requirements are minimal, and a single environment suffices for workload demands
Network and Connectivity Requirements	When network infrastructure cannot reliably support seamless connectivity, leading to bottlenecks in data transfers	Not Applicable

Figure 3b: Key considerations for whether to deploy Hybrid AI

Deploying Hybrid AI

Once an organization decides to deploy Hybrid AI, ITDMs need to devise an implementation roadmap. By adopting an agile approach for Hybrid AI, organizations can prototype, incubate, and learn iteratively. This methodology allows for continuous feedback and improvements, leading to a more effective and scalable Hybrid AI implementation.

Prototype Phase: Rapidly develop and test initial AI models and infrastructure components to validate feasibility and gather early feedback. This phase is crucial for early identification of potential issues and informs subsequent development.

Incubate Phase: Refine and expand the AI models and infrastructure based on feedback, ensuring scalability and robustness. This phase is crucial for transforming initial prototypes into scalable and reliable solutions.

Learn Phase: Continuously monitor, maintain, and optimize the AI system based on performance and evolving business needs. This phase ensures the AI system remains effective and relevant over time.

1 PROTOTYPE

Current State Analysis
- Assess existing IT infrastructure, data pipelines, and AI readiness and audit cloud resources, edge devices, security, and network capabilities to identify gaps

Target State Design
- Define cloud-edge architecture and assign responsibilities to each layer
- Specify where real-time processing and complex analytics will occur

Initial Infrastructure Setup
- Deploy minimal cloud and edge infrastructure to support initial AI model testing
- Set up basic data pipelines to facilitate data flow between cloud and edge environments

AI Model Prototyping and Testing
- Develop and deploy initial AI models focusing on core functionalities
- Conduct preliminary tests to validate model performance and integration with existing systems
- Gather feedback from a small group of testers and stakeholders and Identify areas for improvement

2 INCUBATE

Infrastructure Scaling
- Expand cloud and edge infrastructure to support increased data and model complexity
- Enhance data pipelines for more efficient data synchronization and processing

Model Refinement and Scaling
- Iterate on AI models based on feedback, improving accuracy and performance
- Implement secure APIs for seamless data exchange between cloud and edge
- Scale models to handle increased data volume and complexity, ensuring performance across different environments

Integration Testing
- Conduct extensive testing to ensure smooth interaction between cloud and edge systems
- Validate compliance with industry regulations and security standards

User Training and Documentation
- Provide training sessions for end-users and IT teams
- Develop comprehensive documentation for system usage and maintenance

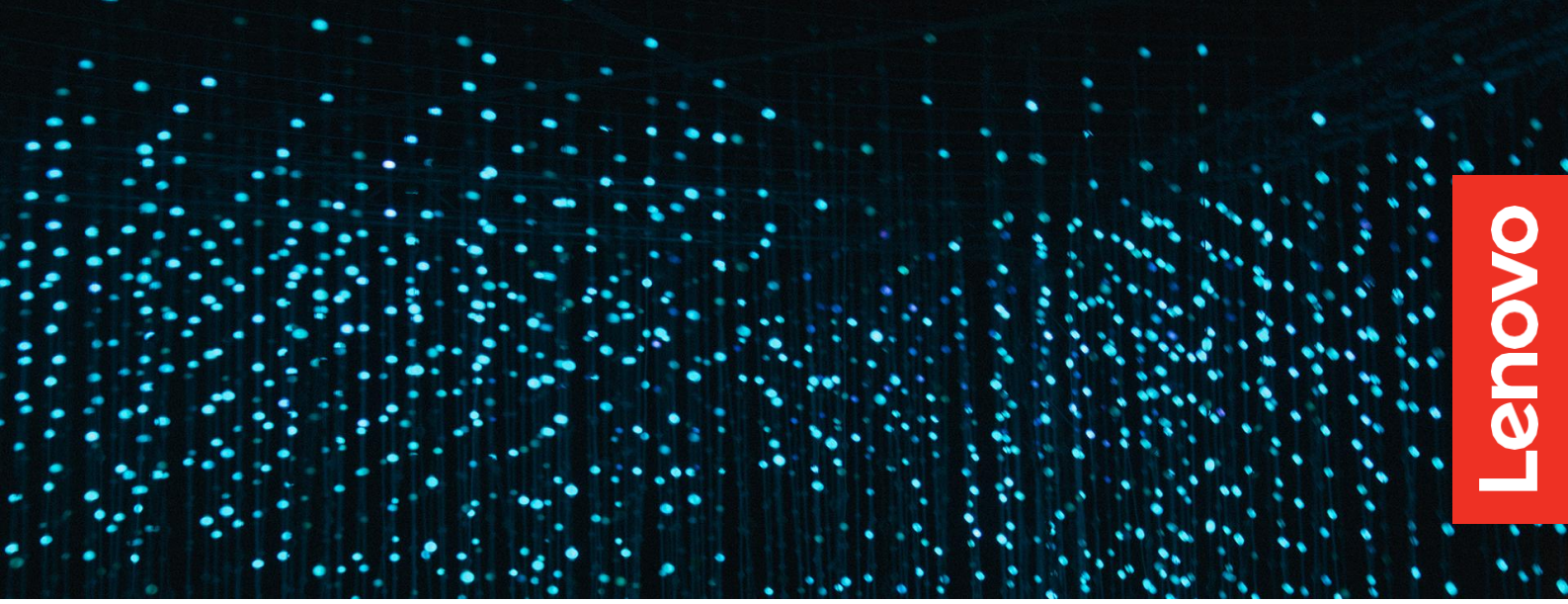
3 LEARN

Performance Monitoring
- Implement real-time dashboards to monitor AI model performance, latency, and data flow
- Use AI Ops to detect anomalies and optimize infrastructure performance

Continuous Improvement
- Regularly review system performance and gather user feedback
- Schedule updates and retrain models based on new data and insights
- Leverage LLM Ops to manage and optimize large language models, ensuring they remain up-to-date and perform optimally

Proactive Maintenance
- Implement break/fix strategies to quickly address and resolve issues
- Plan and execute maintenance activities to ensure system stability and minimize downtime

Figure 4: Key activities during the Hybrid AI deployment journey



Throughout the agile Hybrid AI deployment process there are critical technical strategies that can accelerate deployment and end user adoption. Here are some of these technical tips to help drive a more successful and efficient Hybrid AI deployment:

1 PROTOTYPE

Automated IT Audit
- Leverage automated discovery and assessment tools to audit infrastructure, data pipelines, and AI readiness, detecting gaps and bottlenecks early

Virtual Architecture Prototyping
- Utilize architecture simulation tools to design and validate cloud-edge architecture, optimizing workload distribution without deploying physical resources

IaC-Based Infrastructure Provisioning
- Use Infrastructure as Code (IaC) to automate the rapid deployment of minimal infrastructure, allowing for agile and scalable setups

AI Model Sandbox Deployment
- Deploy AI models in isolated, containerized environments to evaluate core functionalities and gather system integration feedback

Microservice Data Pipelines
- Implement modular, microservice-driven data pipelines to enable rapid iterations and testing of data flows between cloud and edge systems

2 INCUBATE

Elastic Infrastructure Scaling
- Integrate auto-scaling mechanisms for both cloud and edge environments, ensuring dynamic resource provisioning and load balancing based on real-time demand

Decentralized Orchestration
- Employ distributed orchestration platforms (e.g., Kubernetes) to manage AI workloads across cloud and edge, ensuring high availability and fault tolerance

Continuous API Validation
- Automate API schema validation and asynchronous API testing to ensure seamless, secure data exchange and minimize latency during system scaling

Concurrent Integration Testing
- Conduct parallelized integration and compliance testing across all systems to ensure adherence to regulatory standards and minimize integration delays

Scenario-Based Role Training
- Deploy scenario-driven, interactive training modules for both IT and end-users, providing hands-on experience with the AI systems and operational workflows

3 LEARN

AI Ops-Driven Dashboards
- Deploy AI Ops-enabled telemetry dashboards that provide real-time performance metrics, anomaly detection, and automated infrastructure optimization

Automated Model Retraining
- Implement pipelines for continuous LLM retraining, driven by performance metrics and new data inputs, to ensure optimal AI model performance

Self-Healing Infrastructure
- Use predictive analytics with self-healing mechanisms to autonomously detect and resolve issues, reducing downtime and enhancing system reliability

Continuous Real-Time Feedback
- Establish real-time feedback loops to dynamically adjust AI models, infrastructure configurations, and system parameters based on operational data

Figure 5: Technical tips across the Hybrid AI deployment process

In the prototype phase, organizations must ensure that a baseline infrastructure is in place to support early testing and validation. This includes establishing stable connectivity between cloud and edge environments for seamless data transfer, securing minimal compute resources at both the cloud (for basic model training) and edge (for lightweight inference on AI PCs or IoT devices), and setting up local storage at the edge alongside cloud storage for data processing. Basic security measures, such as encryption and access controls, are also necessary to protect data during communication. This foundational setup allows for rapid iteration and validation of AI models without the need to implement the full AI tech stack. In the incubate phase, auto scaling and decentralized orchestration allow for seamless expansion, enabling systems to manage growing complexity and data loads without manual intervention. This ensures faster iterations and reduces operational delays. In the learn phase, real-time monitoring through AI Ops dashboards and self-healing mechanisms proactively addresses performance issues, maintaining system stability and optimizing decision-making. Through continuous

Through continuous improvement, organizations can minimize disruptions and enhance long-term operational efficiency.

One of the primary challenges organization encounters is a lack of clarity between the roles of cloud and edge environments. Organizations may overestimate the capabilities of the edge or underutilize the cloud, leading to inefficiencies and potential overloading of resources. A comprehensive assessment of IT infrastructure and data needs must be conducted to balance workloads across cloud and edge environments. This ensures that workloads are distributed efficiently across cloud and edge environments, preventing resource overload or underutilization.

Assessment Framework

Inventory Hardware and Software Resources
Document current edge devices (e.g., IoT sensors, AI PCs) and cloud infrastructure (e.g., GPUs, storage systems)
Identify AI-optimized hardware and software for both cloud and edge environments
Assess Current Workloads
Identify workloads suited for real-time edge processing (e.g., low-latency tasks) and those best handled in the cloud (e.g., model training, large-scale analytics)
Determine how workloads can be distributed across cloud and edge environments to optimize performance
Evaluate Resource Utilization
Measure utilization rates of cloud and edge compute resources to avoid over- or under-provisioning, ensuring efficient resource allocation
Analyze Network Performance
Assess network latency, bandwidth, and reliability to ensure smooth data flow between cloud and edge environments
Data Storage and Management
Map data flows and ensure storage solutions meet the demands of both environments
Integration Points and APIs
Identify integration points between cloud and edge environments and ensure APIs are optimized for low-latency data transfers
Security and Compliance
Verify compliance with security standards and data privacy regulations for both environments
Ensure encryption, access controls, and secure communication protocols are in place to protect data across edge and cloud environments

Figure 6: Infrastructure Assessment Framework

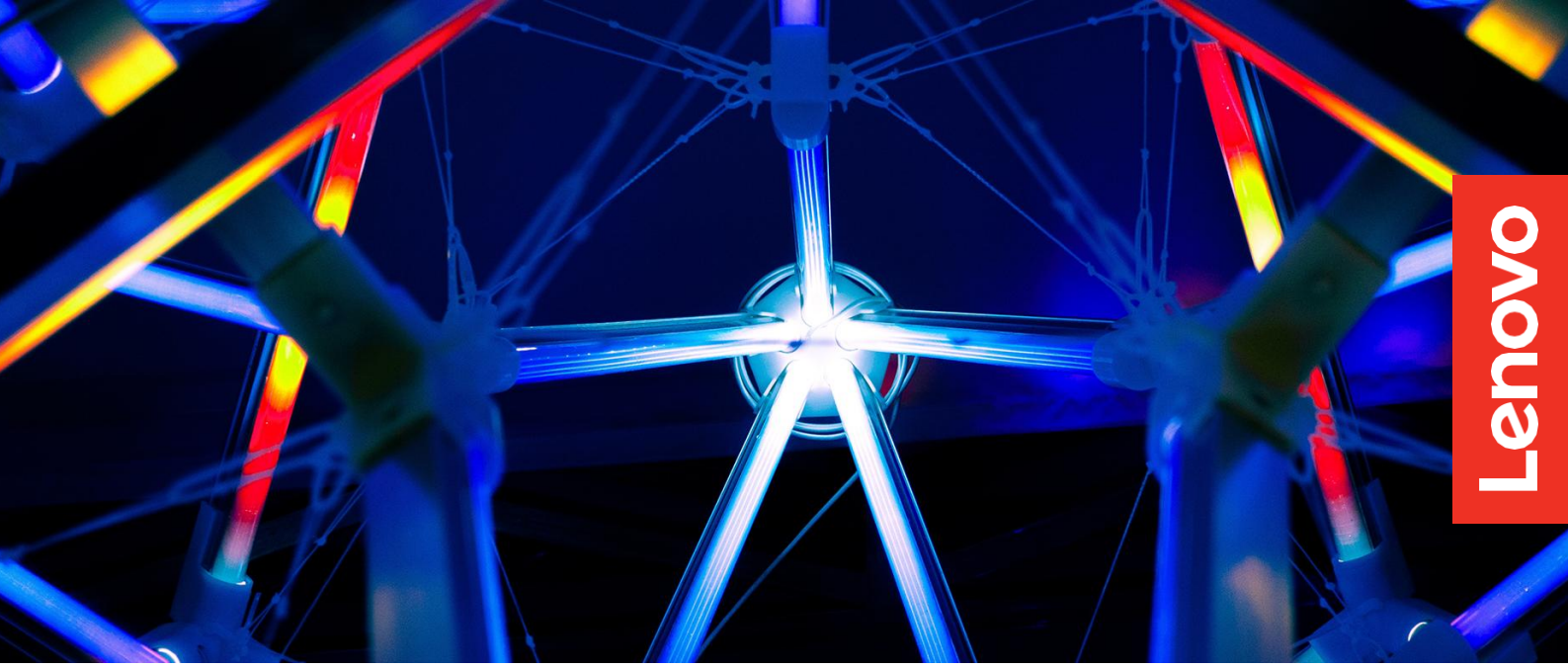
Further challenges may arise around integration complexities between cloud and edge environments. Poorly optimized APIs or inadequate orchestration mechanisms can lead to system bottlenecks, slowing down data flow and causing operational delays. Establishing asynchronous API integrations and decentralized orchestration early in the process can help mitigate these issues, ensuring smooth and efficient data exchanges between cloud and edge environments.

Lastly, not implementing predictive failure models or self-healing systems can result in unexpected downtime, affecting system performance. Not establishing processes around ongoing training, fine tuning, and deployment of AI models to the edge will result in reduced organizational agility to address data changes or new requirements. To prevent this, organizations should prioritize setting up advanced monitoring systems that detect anomalies in real-time, allowing teams to take corrective action before a system failure occurs. Regular performance reviews and predictive maintenance strategies are essential to maintaining an elevated level of system efficiency.

"Not establishing processes around ongoing training, fine tuning, and deployment of AI models to the edge will result in reduced organizational agility to address data changes or new requirements."

ITDMs should adopt the following best practices that streamline integration, minimize operational delays, and improve overall system efficiency:

1. **Leverage Pre-Configured Solutions:** Start with hybrid AI solutions that come with pre-configured infrastructure for both cloud and edge environments. Pre-configured edge devices and integrated systems designed for seamless cloud-edge interoperability simplify the setup process. These solutions reduce initial configuration complexity and allow organizations to quickly assess and validate hybrid AI models in real-world conditions, accelerating time-to-value and reducing operational delays.
2. **Implement Modular Infrastructure:** Adopt modular, scalable infrastructures that grow as AI workloads expand. Edge servers and cloud platforms should support dynamic resource allocation, enabling infrastructure to scale horizontally and vertically based on workload demands. This modular approach ensures that as data loads increase, the infrastructure can adjust seamlessly without major overhauls, offering flexibility and adaptability for diverse AI tasks.
3. **Leverage Automation and Orchestration:** Utilize automated orchestration tools to manage workload distribution between cloud and edge environments efficiently. Automation frameworks provide real-time monitoring and automatic workload adjustments, enabling predictive scaling and reducing the risk of system downtime. These tools optimize resource utilization dynamically, ensuring workloads are executed in the most suitable environment—whether at the edge for low-latency tasks or in the cloud for more resource-intensive operations.
4. **Establish Ongoing Monitoring and Maintenance:** Continuous system monitoring is crucial for maintaining high-performance hybrid AI systems. AI-powered monitoring tools offer real-time tracking of performance metrics, enabling early detection of anomalies and potential issues. By incorporating predictive maintenance, organizations can proactively address system vulnerabilities before they impact performance, ensuring sustained uptime and system reliability.



Accelerating Value Creation with Hybrid AI

To ensure that the transition to Hybrid AI drives tangible value, executives should consider the following tactics:

1. **Capability Building and AI Education** - Transitioning to a Hybrid AI infrastructure poses significant challenges for organizations, particularly around capability building and AI education. Senior leadership plays a critical role in shepherding and sponsoring key initiatives to bridge skills gaps and foster AI readiness. Organizations often underestimate the complexity of AI education, resulting in a lack of readiness, stalled implementation momentum, and missed opportunities. Illustrative tactics to counter these challenges include:

    a. **Skills Gap Analysis:** Conduct a thorough review of current capabilities to identify gaps and develop targeted AI training programs to build internal talent

    b. **Talent Acquisition:** Partner with academia and research institutions to recruit emerging AI talent and expand internal AI expertise through strategic hires.

    c. AI Literacy: Launch AI literacy programs aimed at educating both leadership and employees on core AI concepts and the strategic value of Hybrid AI. Tailored workshops can align employees with the organization's AI goals and make AI adoption smoother.

    d. **Communications Campaigns:** Promote the benefits of Hybrid AI through success stories, internal campaigns, and case studies to create a culture of AI-driven innovation across the organization.

Capability building processes are traditionally owned by HR and project delivery teams—such as training and change management. IT plays a crucial role in informing technology requirements and ensuring that these needs are translated into educational resources. Aligning senior leadership, HR, IT, and project teams is essential for scaling AI education effectively and ensuring that training initiatives are tightly integrated with organizational goals and technology capabilities.

“Senior leadership plays a critical role in shepherding and sponsoring key initiatives to bridge skills gaps and foster AI readiness.”

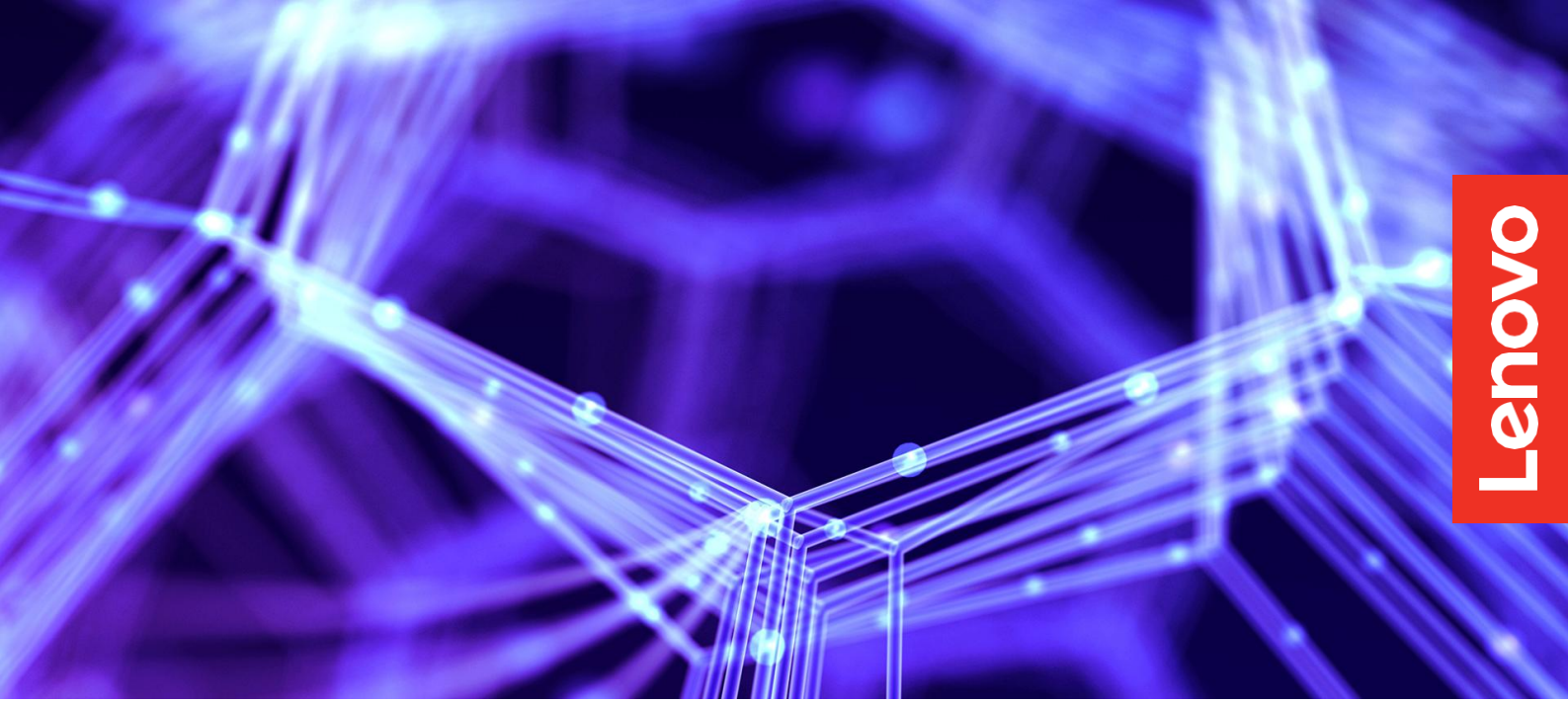
2. **Innovation** - Driving innovation with Hybrid AI can be a complex undertaking, especially for organizations that lack established frameworks for experimenting with modern technologies. Senior leadership plays a critical role in championing AI-driven innovation, fostering a culture of experimentation, and supporting initiatives that encourage creative problem-solving. Without strong leadership buy-in, innovation efforts can stagnate, leading to underutilization of AI capabilities and a failure to identify valuable new applications. Furthermore, the absence of structured processes to test and scale AI innovations can hinder progress. To overcome these challenges, leaders should prioritize the following innovation-driving tactics:

   a. **AI Labs and Pilots**: Establish dedicated AI innovation labs where cross-functional teams can experiment with Hybrid AI solutions in low-risk environments. Run pilot projects to assess proof-of-concept ideas, gather feedback, and iterate on AI models before full-scale deployment.

   b. **Innovation Incentives:** Introduce incentives, such as rewards for successful AI projects or competitions like hackathons, to motivate employees to develop creative AI solutions.

   c. **Tech Partnerships:** Collaborate with tech companies, startups, and industry groups to gain access to innovative AI innovations. For mature organizations, leveraging these partnerships can accelerate the adoption of advanced AI tools that may not be scalable internally.

R&D, product development, and IT teams traditionally own innovation processes, with senior leadership and technology executives playing a key role in sponsoring these initiatives. IT is critical in providing the infrastructure and technical support for pilot projects, while leadership ensures alignment with broader business goals.



3. **Change Management** - Implementing Hybrid AI requires a structured change management approach to ensure a smooth transition. Senior leadership plays a crucial role in driving this change, aligning teams, and fostering a culture that embraces AI adoption. Without a clear AI change management strategy and processes, organizations may face resistance, stakeholder misalignment, and slow progress in achieving AI-driven transformation. Leaders must ensure that all teams are aligned with the AI vision and equipped to manage the cultural and operational shifts that come with Hybrid AI adoption. To address these challenges, leadership should focus on the following change management tactics:

   a. **AI Strategy and Roadmap:** Develop a clear, phased AI strategy and roadmap that outlines the steps for AI adoption and integration. This roadmap should provide a structured approach to transitioning across technology, people, and processes, with milestones to track progress and ensure alignment with business goals.

   b. **Stakeholder Alignment:** Engage key stakeholders early in the AI journey to build support and alignment around AI initiatives. Regular check-ins and collaboration with cross functional teams helps ensure that every department is on board and working towards a common AI vision.

   c. **Culture of Innovation:** Empower AI champions to drive cultural change and promote AI innovation within their respective teams. These champions function as advocates for AI, fostering enthusiasm and providing support during the transition.

Change management processes are traditionally led by HR, project management, and organizational development teams, while senior leadership provides sponsorship and strategic oversight. IT plays a crucial role in supporting the roll out of change management strategies.

4. **Value Measurement:** Measuring the value and ROI of Hybrid AI deployments remains one of the more challenging aspects for organizations and many ITDMs express difficulty in quantifying resource optimization and efficiency gains. Moreover, tracking productivity improvements and tying them directly to AI automation requires a robust set of metrics that few companies have fully defined. Senior leadership plays a vital role in addressing this challenge by implementing clear measurement frameworks and continuously refining value assessment processes to align AI performance with business outcomes. To address these challenges, executives should lead the following value measurement initiatives:

   a. **Cost-Benefit Analysis:** While resource optimization and efficiency gains can be challenging to quantify, organizations should begin by capturing detailed before-and-after benchmarks around operational costs. Measurable factors include reduced cloud computing expenses, lower energy consumption from edge computing, and enhanced system performance. When direct measurement is difficult, scenario modeling—such as comparing actual versus projected project timelines—can help estimate savings. Comparing capital expenditures (CapEx) and operational expenditures (OpEx) with efficiency gains from resource optimization provides insight into financial benefits.

   b. **Performance Metrics:** Tracking productivity improvements driven by AI automation requires developing specific KPIs, such as time savings, system uptime, and operational throughput. Automation tools and dashboards can monitor workflows, helping organizations track reduced inefficiencies and time savings from AI-driven processes. These metrics offer a clear view of how AI automation impacts overall system performance and productivity.

   c. **Business Impact:** Beyond operational metrics, organizations should focus on quantifying AI's impact on strategic business outcomes. Key metrics include revenue growth, accelerated time-to-market, and enhanced customer experiences. By tracking the influence of AI solutions on reducing product development cycles or improving customer satisfaction, leadership can assess the broader business impact of Hybrid AI adoption.

   d. **Ongoing ROI Monitoring:** Continuous monitoring of ROI requires tools that provide real-time visibility into AI system health and performance. Dashboards that integrate with performance metrics can help assess system value over time. Implementing automated ROI reports that align financial outcomes with KPIs such as uptime, cost savings, and speed to market enables data-driven decision-making. This ongoing assessment helps leadership adjust strategies to maximize AI value.

The value measurement process should be a collaborative effort between finance, strategy, and IT teams, with leadership ensuring alignment with broader business objectives. IT teams provide the necessary infrastructure and real-time performance data, while finance and strategy teams establish frameworks for tracking ROI and aligning outcomes with key business goals.

Conclusion

ITDMs exploring Hybrid AI should focus on how this emerging opportunity can drive clear ROI for their organizations. Prioritizing use cases that deliver quick wins is crucial for demonstrating Hybrid AI impact and for garnering broader business alignment and buy-in. Below are some cross-functional use cases where Hybrid AI can help enhance organizational productivity:

Function	Hybrid AI Use Case
Customer Support	**Real-Time Support Automation:** Instant query handling via edge AI chatbots, with cloud-based analytics improving customer interactions[8]. **Sentiment Analysis:** Edge AI assesses customer emotions during calls, while the cloud analyzes long-term customer sentiment trends.
Supply Chain and Logistics	**Predictive Maintenance:** Real-time equipment health monitoring at the edge, with cloud-based optimization for maintenance schedules. **Dynamic Routing:** Edge AI adjusts delivery routes based on traffic, while cloud systems manage long-term logistics planning.
Manufacturing	**Quality Control:** Edge AI inspects products in real-time for defects, while cloud AI refines models to enhance quality control. **Demand Forecasting:** Real-time production adjusts at the edge, while cloud analytics predicts long-term demand and optimizes inventory.
Human Resources	**Talent Acquisition:** Edge AI automates resume screening and interview scheduling, with cloud-based analytics refining candidate selection. **Employee Engagement:** Real-time feedback is collected at the edge, with cloud systems assessing long-term engagement and satisfaction trends.
Finance	**Fraud Detection:** Instant detection of suspicious transactions by edge AI, with cloud systems improving fraud detection models over time. **Risk Management:** Real-time market risk analysis at the edge, with cloud AI fine-tuning risk models based on historical data.
IT Operations	**Network Performance Monitoring:** Edge AI monitors network performance in real time, while cloud analytics optimizes long-term network health and prevent outages. **AI-Driven Security:** Real-time threat detection at the edge, with cloud systems enhancing security protocols and improving threat models.

Figure 7: Illustrative cross functional Hybrid AI use cases

Hybrid AI

Hybrid AI does not have to be just a technical solution but can function as an inflection point in an organization's growth journey. As Hybrid AI continues to evolve with the integration of innovative innovations like NPUs and energy-efficient infrastructure, the potential for greater efficiency, reduced costs, and enhanced decision-making will continue to grow. For ITDMs, the path forward is clear: embrace Hybrid AI by prioritizing the right use cases and refining the technology stack over time. This will help organizations future-proof their AI strategies and maintain a competitive advantage in an increasingly complex and dynamic market environment.

References

[1] Lenovo: Achieve Better Economics and Performance Through Hybrid AI
[2] WSJ: The World of Hybrid AI
[3] Lenovo: Smarter AI for All
[4] Lenovo: Hybrid AI Innovation to Meet the Demands of the Most Compute Intensive Workloads
[5] TechTarget: What is Hybrid Cloud? The Ultimate Guide
[6] Medium: Edge Computing for AI
[7] IDC: The Increasing Significance of the Cloud Buyer in a Time of AI and Other Transformations
[8] IDC: Transform Your Business With AI-Powered Solutions

About Lenovo

Lenovo is a US$57 billion revenue global technology powerhouse, ranked #248 in the Fortune Global 500, and serving millions of customers every day in 180 markets. Focused on a bold vision to deliver Smarter Technology for All, Lenovo has built on its success as the world's largest PC company with a pocket-to cloud portfolio of AI-enabled, AI-ready, and AI-optimized devices (PCs, workstations, smartphones, tablets), infrastructure (server, storage, edge, high performance computing and software defined infrastructure), software, solutions, and services. Lenovo's continued investment in world-changing innovation is building a more equitable, trustworthy, and smarter future for everyone, everywhere. Lenovo is listed on the Hong Kong stock exchange under Lenovo Group Limited (HKSE: 992) (ADR: LNVGY).