PREPARING THE FOUNDATION FOR THE AI OF TOMORROW

ABIresearch
THE TECH INTELLIGENCE EXPERTS™

Lian Jye Su, Research Director
Malik Saadi, Vice President, Strategic Technologies

CONTENTS

EVERY AI INVESTMENT HAS TO BE A FORWARD-LOOKING BUSINESS DECISION

The 2020s are shaping up as the decade of Artificial Intelligence (AI). With the right investments, the technology is poised to be widely deployed in various Information Technology (IT) and Operational Technology (OT) use cases. PwC's Global AI Study: Exploiting the AI Revolution estimates that AI will contribute US$15.7 trillion to the global economy by 2030.

Not surprisingly, multiple AI adoption studies conducted by major AI players reveal that many businesses have increased their AI budget quite significantly compared to previous years. However, most investment decisions are driven by the broad promise of AI without focusing on specific implementation hurdles in large organizations with legacy footprints. As a result, many businesses risk creating a siloed, loosely integrated, and proprietary system.

This whitepaper aims to unpack the different facets of AI and their respective computational requirements, showing that AI investments must be based on long-term business outcomes and values.

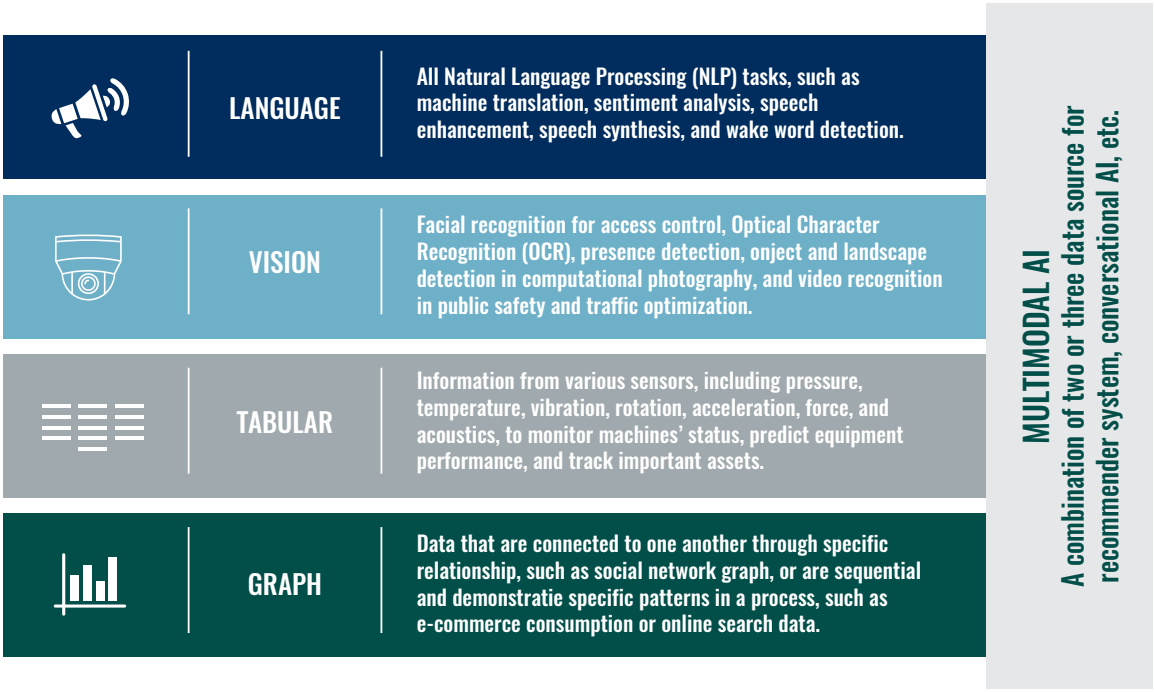
Commissioned by Intel Corporation.

MULTIMODALITY IS THE FUTURE OF AI

Before diving into the infrastructure requirements, it is worthwhile having an overview of today's AI. As implemented today, AI generally focuses on four major applications, as described in Figure 1.

Figure 1: Major Classes of AI of Today

(Source: ABI Research)



AI applications, such as Robotics Process Automation (RPA) and Optical Character Recognition (OCR), are already automating mundane and repetitive workflows, helping human employees perform better at their current tasks and assisting businesses with staying compliant with legal requirements. However, the AI models in these applications are focused on a highly structured workload using one of the four data types mentioned above: language, vision, tabular, or graph. Therefore, a significant portion of tomorrow's AI models will be more versatile, leveraging most, if not all, of the four data types.

These AI models are designed for a variety of tasks. They can improve learning, decision-making, and experiences by relying on training from various data sources, records and archives, and personal information with the individual's explicit consent. However, before this vision becomes a reality, the industry needs to develop more robust multimodal learning models to process various data in real time. These models will also need to be supported by high-performance heterogeneous computing architecture to enable a wide range of tasks and functions, from data gathering and structuring, which consumes a significant share of processing resources to real-time inference and other mission critical tasks that require ultra-low latency and high accuracy.

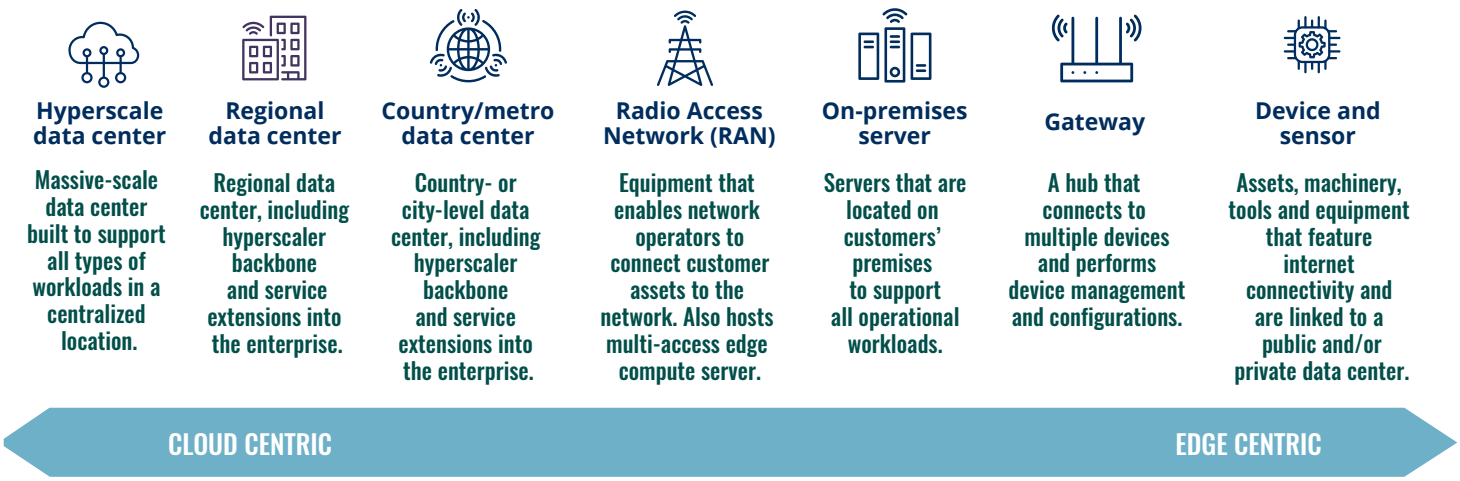
DIVERSITY IN AI IMPLEMENTATION ENVIRONMENTS

AI technology suppliers and developers are also taking advantage of new AI techniques and edge computing technology to deploy AI across all devices from the cloud to the edge. On one hand, Cloud Service Providers (CSPs) are leveraging scalable and hyperscale data centers to develop and deploy high-performance vision, language, and graph models that are constantly increasing in size. On the other hand, businesses are looking for AI models embedded in devices and gateways to improve latency, protect privacy, and reduce reliance on cloud infrastructure. Tiny Machine Learning (TinyML) pushes the boundary further by introducing ultra-low power AI inference in sensors and battery-powered devices. Software capabilities like Neural Architectural Search (NAS) and new model compression technologies like knowledge distillation, pruning, and quantization will enable AI developers and implementers to create the most optimized model for their target environment.

All these advancements mean AI models will be present in every node of the computing continuum, ranging from hyperscale data centers to regional data centers, on-premises servers, edge computing gateways, devices, and sensors. All these locations enable businesses to deploy AI at the most optimal location in terms of computing power, latency, connectivity, and regulation. In addition, the arrival of next-generation telecommunication technologies, such as 5G and Wi-Fi 6, also allows the transfer of a large amount of data for training and inference. While such infrastructure is not currently available in every location, this reality will change as cloud AI giants, telecommunication service providers, industrial companies, and edge computing companies continue to build the relevant infrastructure in the next few years. As a result, AI developers will widely introduce AI models based on new techniques, heterogeneous hardware, and low latency connectivity.

Figure 2: Edge-to-Cloud Computing Continuum

(Source: ABI Research)



PRIVACY AND SECURITY-ENHANCED AI

Aside from classical AI and some Machine Learning (ML) models, most of today's Deep Learning (DL) models are black boxes that lack transparency. AI developers do not have complete knowledge of all the individual neurons, layers, and parameters in a DL model that work together to produce final output. The role of all these components and their influence over each other remains largely unexplained.

Moving forward, AI developers will make AI models more transparent. In most cases, a data and AI development platform will be designed to explain to users the limitations of training and testing data, the logic behind all AI training and inference processes, and potential bias, drift, and other gaps. In some cases, AI models may be designed to explain themselves to the end users. The transparency and explainability will enable AI models to be used in high-risk environments, as they can withstand scrutiny and evaluation.

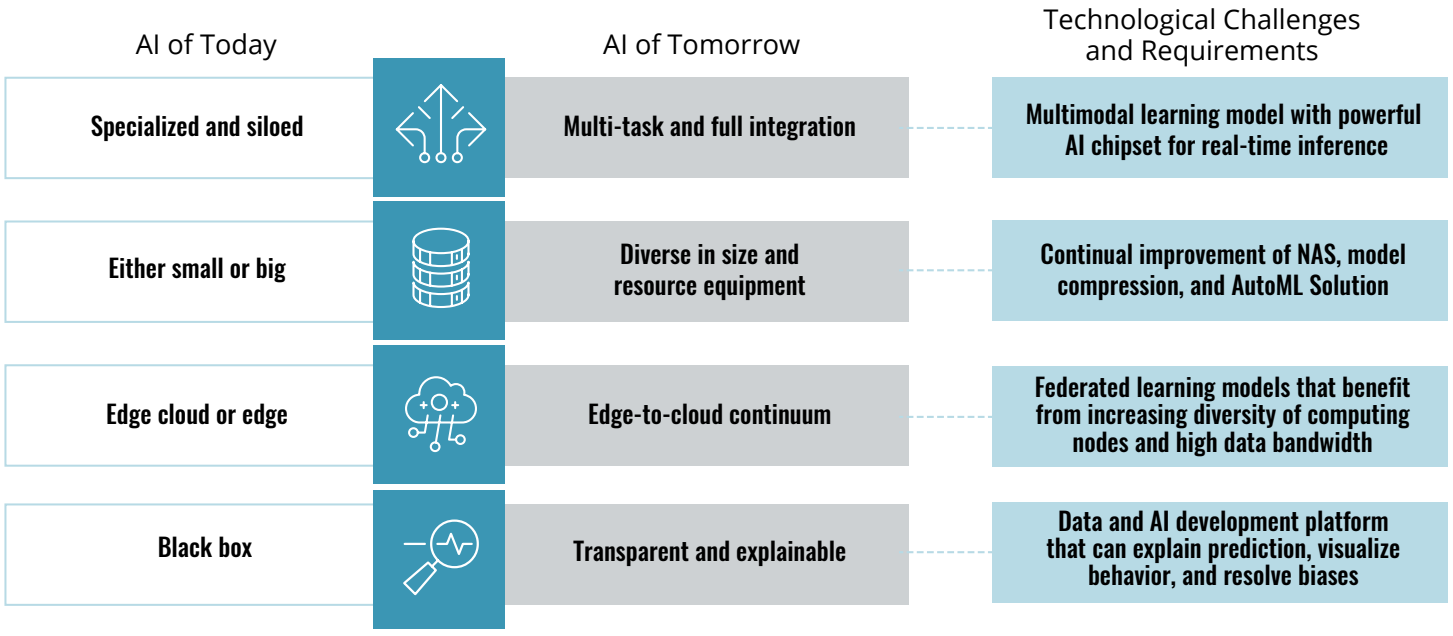
Furthermore, AI developers can enhance security in an AI system by limiting data transfer to the cloud and deploying the AI model at the edge. By keeping and processing raw data at the edge, end users do not need to worry about their data being hijacked by malicious actors. At the same time, consumers concerned about storing Personal Identifiable Information (PII) in the cloud will no longer need to worry, as such data will be processed in edge devices and local servers.

PREPARING FOR THE AI OF TOMORROW

When we compare the state of AI today with the vision for tomorrow's AI, a few distinct characteristics jump out, as shown in Figure 4.

Figure 3: AI Evolution and Technological Challenges and Requirements

(Source: ABI Research)



To truly make the AI of tomorrow a reality, it is apparent that businesses will need to continue their investment in the right AI infrastructure and approach.

New Learning Techniques

The evolution of AI will be shaped by the emergence of new AI techniques, such as federated learning, multimodal learning, graph DL, reinforcement learning, and meta-learning. These AI techniques use a fully distributed, highly customized computing landscape to learn and handle multiple tasks. The ability to interpret contextual information, understand the linkage between different factors, and work and learn from other AI models creates immense opportunities for AI to change how businesses operate. They will be ready to handle more complicated tasks than human employees can handle.

Table 1: Key AI Trends and Their Business Implications

(Source: ABI Research)

New AI Techniques	Descriptions	Strengths	Technology Platform Requirements
Federated learning	Federated learning is a distributed ML approach in which multiple users collaboratively train a model without moving data to a single server or data center. Instead, each compute node will execute the same model, train such a model on the local data, and thus compute and store a local version of the model in each node.	Federated learning provides edge devices with quality ML models without centralizing the data. Deployed in various environments, including smartphones, healthcare, and finance, the technique allows the access of datasets from different users, institutions, or databases, while helping to comply with required privacy and confidentiality laws.	Stable and ubiquitous connectivity backbone, with AI compute taking place in local and cloud environments. Frequent exchange of data and sync between the cloud and different edge nodes to ensure the model is up to date.
Multimodal learning	Multimodal learning can simultaneously process various data types (image, text, speech, numerical data) using multiple algorithms. Multimodal AI can interpret such multimodal signals together and make decisions based on contextual understanding.	AI-based on multimodal learning can mimic human decision-making by ingesting different data sources. As a result, it often outperforms single-modal AI in many real-world problems, such as customer services, client engagement, and patient care.	Databases that can ingest various data sources and AI models that process different data modalities and perform inference in real-time.
Reinforcement learning	Reinforcement learning is an ML training method that rewards the learning agent when making desired behaviors and punishes it when making undesired ones. Generally, a reinforcement learning agent can perceive and interpret its environment, take actions, and learn the associations between stimuli, activities, and the outcomes of its actions through trial and error.	Reinforcement learning has been widely adopted in simulation to train and retrain behaviors of autonomous vehicles and robots for traffic management, material handling, route optimization, and space management. Aside from the physical system, the software can also be trained using reinforcement learning for a data-driven product or process optimization, such as supply chain optimization, prototyping, and generative design.	Powerful AI compute platform in the cloud with precise and realistic rendering of the real-world environment. Alternatively, a highly optimized, unsupervised self-learning model in end devices.
Graph Neural Network (GNN)	GNNs are DL neural networks designed to perform inference on data stored in graph databases. Graph databases connect specific datapoints (nodes) and create relationships (edges) in the form of graphs that the user can then pull with queries. The AI can understand the interdependency between each datapoint and provides relevant predictions.	GNNs are ideal for analyzing a specific issue that involves numerous factors. For example, credit risk for credit card customers requires understanding current credit scores, credit history, employment, income, and other socio-economic factors. Other use cases include recommendation systems, molecular cell structure study, reading comprehension, and social influence prediction.	Hardware and software are optimized for GNNs, as GNNs take much longer to train.
Meta-learning	Meta-learning refers to ML algorithms that learn from the output of other AI algorithms. These models can learn across a suite of related prediction tasks through an adaptive process, allowing them to speed up their learning process while learning multiple functions simultaneously.	Meta-learning is still embryonic, so it is rather difficult to predict how influential and impactful the technology will become. Nonetheless, if the current prediction is accurate, meta-learning models can learn quickly, requiring less training time and fewer resources to design and develop. This will save enormous amounts of time and accelerate time to market.	Ultra-high-performance hardware and software, as meta-learning is an ensemble of the most advanced ML techniques, such as reinforcement learning and transfer learning.

More importantly, they allow businesses to scale out and scale up depending on business needs without needing to expand and train their workforce. Some of these techniques, such as federated learning and reinforcement learning, have already been adopted by cloud AI giants and large corporations to design advanced AI models in large-scale recommender systems, fraud detection, and virtual assistance. Meta-learning and GNNs, on the other hand, are slowly maturing and appearing in some interesting use cases, such as drug design, robotics training, and disease diagnostics.

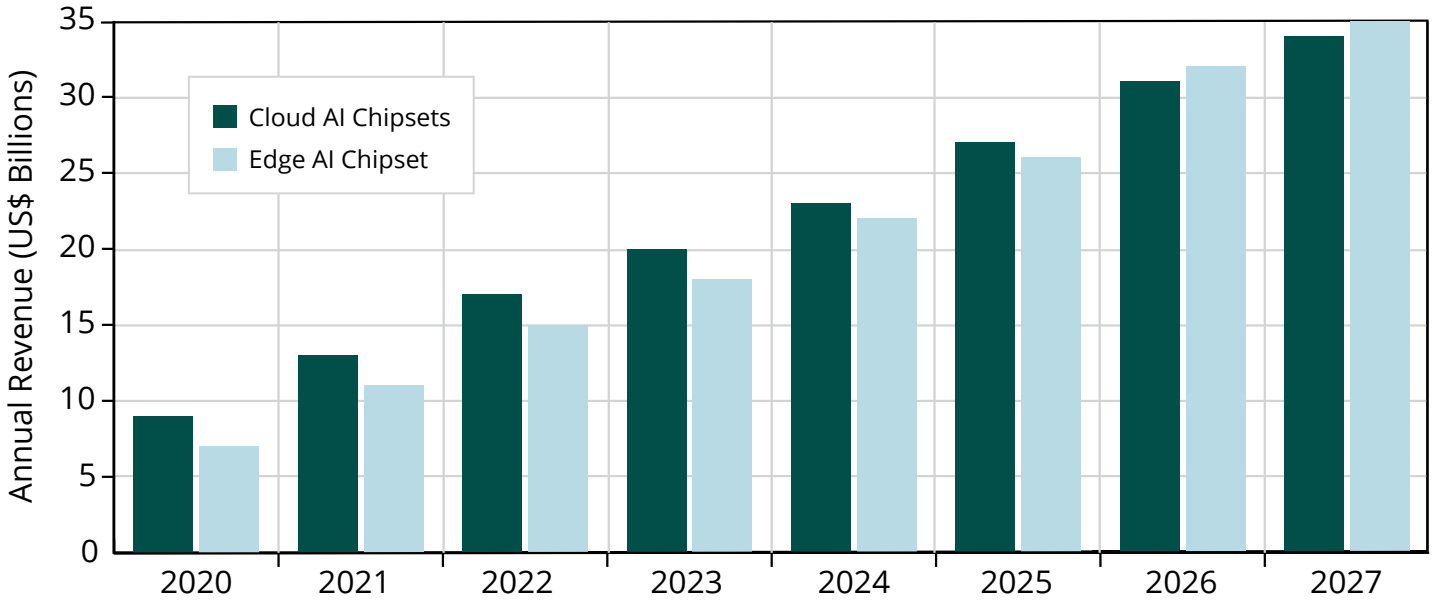
Optimized AI Infrastructure

In recent years, AI technology providers have progressively reduced the barriers to entry by actively launching innovative products and services. Graphic Processing Units (GPUs) and Application-Specific Integrated Circuits (ASICs) are being adopted for AI training and inference. Nowadays, more and more general-purpose Central Processing Units (CPUs) can support AI inference and training. The introduction of pre-training language models allows developers to build complex applications, such as speech recognition and machine translation, without training a model from scratch. Auto Machine Learning (AutoML) provides methods, tools, and techniques to make the development process easier for non-AI experts by automating AI workflow.

Not surprisingly, all these advancements have led to a huge demand for AI chipsets. According to ABI Research's Artificial Intelligence and Machine Learning market data (MD-AIML-109), the global AI chipset market is estimated to be US$32.3 billion in 2022. This includes the sales of AI training and inference chipsets, including the CPU, GPU, Field Programmable Gated Array (FPGA), Neural Processing Unit (NPU), AI accelerator, microcontroller, and neuromorphic chipset, in all data centers and end devices. Furthermore, the democratization of AI will lead to AI being deployed across a wide range of physical sites and compute nodes. As a result, this market is expected to grow to US$68.8 billion in 2027, with a Compound Annual Growth Rate (CAGR) of 26%.

Chart 1: Total Revenue from AI Chipset Sales World Markets: 2020 to 2027

(Source: ABI Research)



KEY PRINCIPLES OF AI INFRASTRUCTURE INVESTMENT

Today's AI is narrowly focused, requires a wide range of expertise, and exists in a silo. In contrast, the AI of tomorrow requires enormous amounts of resources and deep technological knowledge, which remains out of reach for most businesses. Therefore, businesses must start early, identify the business outcomes that cannot be easily achieved without AI, actively build internal capabilities, and roll out these advanced AI techniques widely across the entire organization. In summary, below are four key pillars when considering AI infrastructure:

- **AI Infrastructure Must Be Driven by Business Outcomes:** The vision of AI infrastructure must be based on the intended business outcome of AI deployment. Businesses must first understand the short- and long-term values AI brings to their operation before designing the most suitable AI models. When an AI project has a clear business outcome, it has actual financial values that senior management can recognize.

- **AI Infrastructure Must Be Heterogenous and Flexible:** To unlock the actual value of AI and yield maximum benefits, scale-up and scale-out of AI applications are critical. Building an AI infrastructure that offers the proper foundation to support different facets of AI model design, development, and deployment across different computing platforms goes a long way to protect and future-proof current investments. A heterogeneous compute platform will offer the best performance across all AI tasks. AI developers can use the CPU for data gathering and preparation, before switching to the GPU and ASIC for model training, and finally using either the GPU, ASIC, or CPU for AI inference workload.

- **AI Infrastructure Must Be Backward Compatible:** All AI infrastructure must be able to work with existing enterprise solutions. Therefore, setting a versatile, robust, and interoperable foundation with all existing solutions is a must. Incompatibility risks creating many silos in the business operation, leading to poorly optimized IT/OT infrastructure and processes.

- **AI Infrastructure Must Be Open and Secure:** Businesses always want to avoid vendor lock-in. An AI infrastructure consisting of open hardware and software that can interoperate with other solutions is significant in ensuring smooth IT/OT processes. At the same time, openness should not lead to a compromise in security. The AI foundation must feature state-of-the-art cybersecurity and data protection mechanisms to prevent hacking, protect user data, and comply with legal requirements.

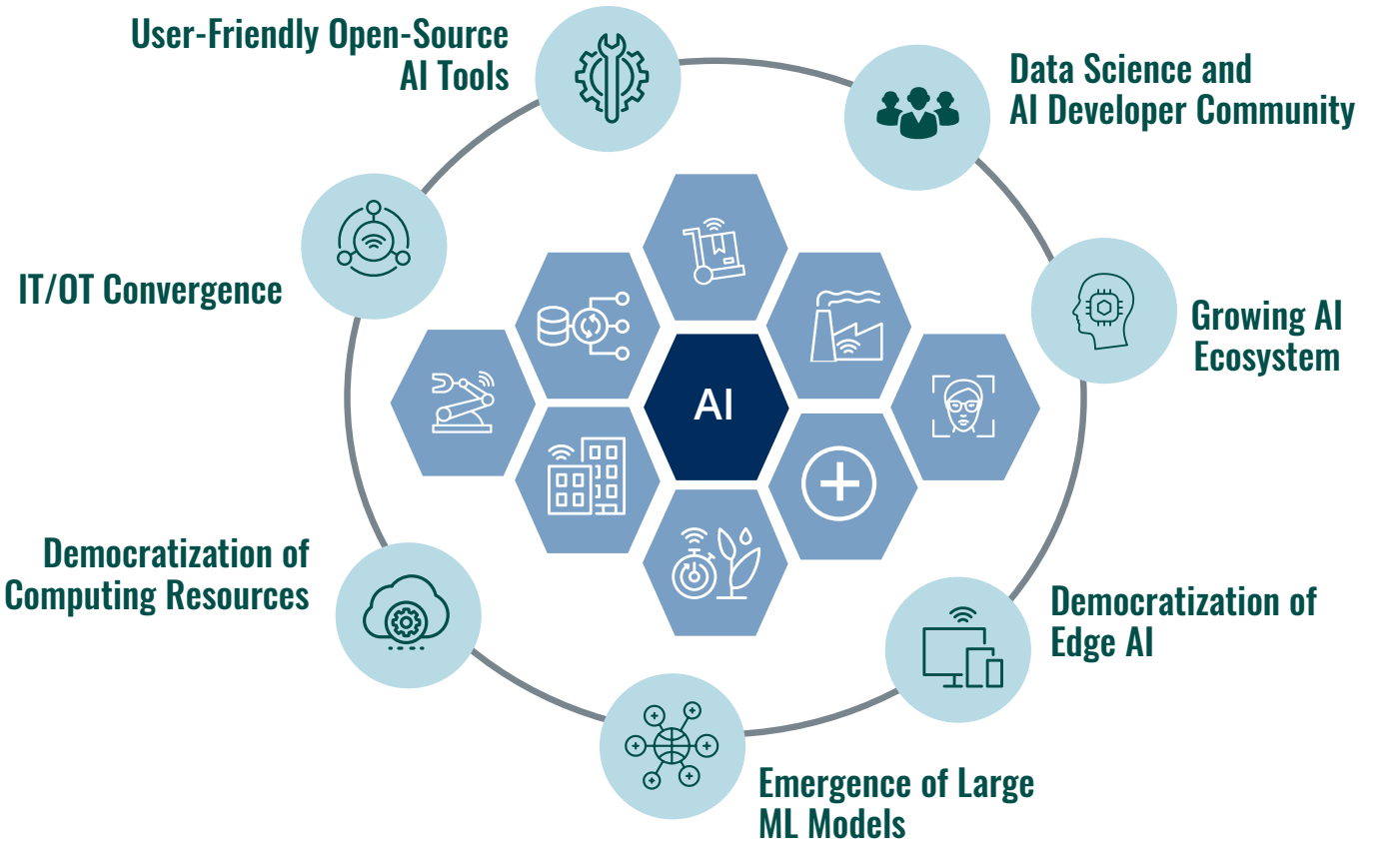
It is clear that AI is still in its infancy, and building the proper foundation for it is critical for its future success. Instead of looking at AI from today's lens, all businesses must have a clear long-term plan. This vision will help them navigate the challenges and technology requirements for AI, helping them make the right decision in investing in the most optimal and future-proof AI infrastructure. The following section discusses various approaches businesses can take to deploy AI. In addition, it highlights the key features and characteristics businesses must pay attention to when selecting their AI technologies.

CHARACTERISTICS OF A FUTURE-PROOFED AI INFRASTRUCTURE

Enabling AI broadly across enterprises requires a different mindset. Businesses must understand that building AI for business is a continuous process involving many building blocks. As shown in Figure 5, several key recent advancements have allowed AI to become a reality.

Figure 4: Key AI Technology Trends

(Source: ABI Research)



All of them have a significant impact on how businesses should deploy their AI.

Table 2: Key AI Technology Trends and Their Business Implications

(Source: ABI Research)

Key Trends	Descriptions	Business Implications
IT/OT convergence	IT and OT have traditionally been developed separately, with no ability to exploit operations and production data to make more informed decisions for optimized workflow and well-planned production and maintenance processes. COVID-19 has sped up the digital transformation process in many businesses, leading to emerging technologies, such as the Internet of Things (IoT) and robotics automation. As a result, businesses also start to collect more operational data that are very useful for planning, optimization, and upgrades.	This convergence will enable businesses to gain insight and make data-driven decisions. They can also optimize their existing workflows without needing to scale up rapidly.
Democratization of computing resources	The democratization of computing resources for AI has been achieved by the wide availability of public cloud computing. Indeed, AI developers may want to leverage the centralized processing and storage offered by CSPs instead of deploying their own AI hardware, which is not economical enough, mainly for executing dense AI networks. The uniformity and scalability of CSPs' compute and storage architecture enable them to handle compute-intensive DL models on an on-demand basis, significantly lowering the barrier-to-entry for AI developers without the ability to build and maintain their own AI infrastructure. As AI models are becoming more complex, the access provided by CSPs is essential for the democratization of AI.	Businesses must invest in the right AI infrastructure by buying from established CSPs or building their private cloud infrastructure. While public cloud solutions are scalable, they can be costly when compared to being well-planned for private infrastructure based on long-terms goals. Therefore, businesses should also consider leveraging the best of both worlds with hybrid cloud deployments to get the best price-performance advantage and flexibility.
Emergence of large DL models	Another primary reason behind the gain in accuracy and performance is the growth in DL models, precisely the number of parameters and hyperparameters. AI models have scaled significantly in the past years. Depending on the application, large models provide fundamentally unique advantages. For example, OpenAI's GPT-3, widely considered the most advanced Natural Language Processing (NLP) model of 2021, has 125 million to 175 billion parameters and can handle advanced applications, such as generative emails or document summaries. Newer models like BLOOM from BigScience, which has 176 billion parameters, support multiple human languages and programming languages.	Businesses need to consider the cost of AI training and implementation. For context, the cloud computing cost for the training of BLOOM, which is around 330 Gigabytes (GB) in size is estimated to be in the multi-million-dollar range. Aside from the proper hardware infrastructure, businesses must also identify the suitable applications and use cases they want to deploy.
Democratization of edge AI	Highly optimized and miniaturized AI models are currently embedded in smart sensors, devices, and gateways. These carefully crafted smaller models can also perform narrowly-focused applications, specifically in always-on computer vision and time-series data analysis. Solution providers have introduced power-efficient AI processors (ASIC, NPU, neuromorphic chipset), DL model optimization techniques (knowledge distillation, pruning, and quantization), developer-friendly tools and services, and more intelligent resource allocation.	Edge AI is a way to minimize latency, privacy risk, and connectivity costs. Businesses must look into the long-term benefits of edge AI and develop a strategy to deploy in their operation.
Growing AI ecosystem	The AI ecosystem continues to grow at a rapid rate. AI startups are offering a wide range of solutions. The most well-known startups work on facial recognition, Advanced Driver-Assistance Systems (ADAS), and NLP. COVID-19 has become a catalyst behind the rapid adoption of AI-based enterprise automation, such as RPA, AI-aided speech recognition, transcription and translation, and sales and marketing enablement tools. Those looking to build their own custom AI will lean toward startups that offer comprehensive data science and AI development platforms. AI chipset startups are also gaining prominence in recent years, providing the ideal computing solutions.	There is no better time for businesses that prefer to outsource their AI expertise. There is a wide range of solution providers in AI chipsets, software, and services. They can also choose to form industrial partnerships with innovative startups.
User-friendly open-source tools	Open-source algorithms and frameworks enable the AI community to embark on new research without heavy software investment, while being supported with future releases and updates. All leading AI frameworks available today are open source, with TensorFlow and PyTorch leading the pack. In addition, the involvement of commercial entities sponsoring the development of AI frameworks has led to solid governance, improved capabilities (especially in edge inference), familiarity, hardware, and commercial support.	Open-source solutions come with their own set of strengths and weaknesses. Therefore, it is essential to identify vendors that provide the proper support for open-source solutions.
Data science and AI developer community	The data science industry has come a long way since the beginning of big data. The expertise and capability of data science have become even more critical with the emergence of ML. Key data processes, such as data analytics, management, stewardship, and compliance, form the foundation for AI models' training, testing, and validation. In addition, the openness of the data science discipline has created a robust community, which later includes AI developers, engineers, and implementers. These communities are often inclusive, innovative, and collaborative, serving as valuable knowledge sources.	Open communities provide invaluable insights into the latest technological trends in data science and AI. Businesses can take advantage of these insights and improve their internal data and AI operations. They should also contribute back to the communities.

Knowing these key trends, what should a business do? In general, businesses can consider the approaches outlined in Table 3 when selecting their AI option.

Table 3: AI Solution Options and Their Business Implications

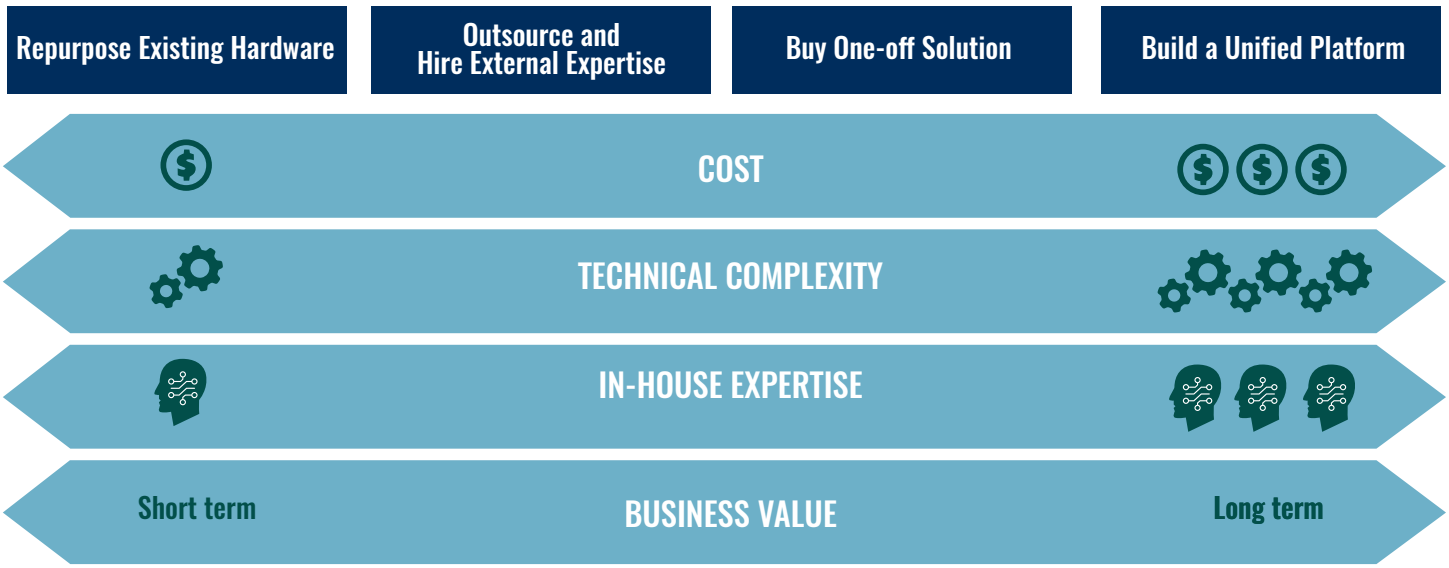
(Source: ABI Research)

Options	Descriptions	Business Implications
Repurpose existing hardware	Large businesses have investments in private data centers for internal processes. They can repurpose existing servers to test and trial AI models that are less compute-intensive, but most, if not all, of the existing hardware was not initially designed with AI workloads in mind. While this approach can lower the barrier to entry for AI, it poses several challenges in the long run, including a lack of scalability, poor optimization, and an infrastructure silo. Implementers should consider hardware infrastructure that is already ubiquitous and optimized to handle AI workloads.	Repurposing existing solutions is a short-term option for AI solutions that are limited in capabilities. They may result in silos when businesses need to run multiple AI models, lack scalability and flexibility for future use cases, and have high costs and slow performance due to poor optimization.
Outsource and hire external expertise	Businesses use hardware and software from various solution providers on a pay-per-use basis. For example, cloud service providers provide cloud computing infrastructure and AI design and development platforms, while specialized AI vendors provide dedicated business automation solutions. However, for this approach to be successful, businesses must have a long-term vision to combine everything and create a hyper-automated workflow.	While businesses can benefit from third-party skills and expertise, businesses will find it challenging to build in-house AI capabilities. In addition, businesses may risk being tied into existing relationships and have limited customization options.
Buy one-off solutions	One-off solutions are fully configurated hardware and software offered by third-party solution providers or system integrators for purpose-specific functions. These hardware and software are generally ubiquitous and widely available. Hence, instead of subscribing to AI-as-a-Service, businesses own the entire solution and can configure it to the maximum efficiency and value. This is ideal for today's AI, which focuses on a specific task.	Businesses have complete control over their AI solution. They can build their team for maintenance and upgrade, retaining valuable skill sets. However, the one-off solution will result in silos, difficulties for use-case scalability, or even performance upgrades.
Build a unified platform	Businesses with more experience in AI usually seek to build a broad platform that can host different AI applications on a single unifying architecture. This approach enables businesses to rapidly scale their AI applications when needed, as all their AI applications are hosted on the same infrastructure. However, significant downsides of this approach include human resources and opportunity costs due to the need to maintain and upgrade a fully proprietary unified platform.	Building a unified platform takes much longer than buying a one-off solution, but this can be mitigated by focusing on commercially available hardware and software highly optimized for AI computing. This not only provides uniformity, scalability, and maximum configurability, but also helps to reduce long-term maintenance costs.

Figure 5 summarizes the pros and cons of each AI infrastructure approach.

Figure 5: Pros and Cons of Different AI Infrastructure Approaches

(Source: ABI Research)



It is a common consensus that building or buying is never the end-all-be-all solution. Businesses should consider buying AI solutions and building whenever they must. However, if businesses are serious about developing internal capabilities in AI, they must think long term and look to build or leverage a unified AI platform. While it is true that building a unified platform is the most capital-intensive and complicated route and demands high technical expertise, businesses must revisit and be mindful of the key pillars of AI infrastructure mentioned above:

- AI infrastructure must be driven by business outcomes.
- AI infrastructure must be heterogenous and flexible.
- AI infrastructure must be backward compatible.
- AI infrastructure must be open and secure.

Based on these pillars, ABI Research believes that all businesses must consider the following features and characteristics when selecting their AI infrastructure suppliers.

COMPREHENSIVE AND HETEROGENOUS INFRASTRUCTURE

AI inference and training workloads rely more and more on parallel accelerated computing capabilities. The explosive demand for GPUs and AI accelerators is a clear sign of the critical role of accelerated computing in the age of AI. However, AI is more than just peak computing performance or a high degree of optimization for specific applications. AI hardware with a high degree of specialization is excellent in handling specific AI models, but it could be overwhelmed when handling AI models not optimized for the specific hardware.

Therefore, businesses must understand their AI needs and invest in a more heterogenous AI infrastructure with a combination of AI computing chipsets that meet specific application needs. The heterogeneity is essential in future-proofing existing infrastructure when there is a need for broad AI models. In addition, such infrastructure can handle a wide range of AI model training and inference workloads through optimized bandwidth, speed, and latency:

- **Computing:** Most AI training and inference today are based on GPU, thanks to their parallel processing ability and graph processing DNA, which is considered a great asset for processing AI workloads. Not surprisingly, cloud AI giants and enterprises are expected to continue investing in GPU-optimized systems for AI training and inference workloads in the cloud. In recent years, the emergence of high-performing and power-efficient ASICs has played a crucial role in further accelerating AI inference workloads due to their massive parallel computing performance and credit to their ability to process some complex functions in the hardware domain.

That said, the CPU is essential in processing neural networks, regardless of using an AI processor for inference and training computation. All AI infrastructure requires a powerful CPU to perform several critical tasks, including computing task allocation, management orchestration, and executing instructions kept in the computer memory, delivering ultra-high Input/Output (I/O) bandwidth, and supporting large amounts of systems memory. The CPU is also very good at processing preliminary tasks necessary for AI training, including data cleaning, classification, filtering, labeling, and structuring. In addition, with CPUs now evolving toward multicore and parallel processing architectures, they will increasingly have the ability to rival GPUs and hardware accelerators on handling more mission-critical AI workloads and models, including vector, matrix, and spatial models. Furthermore, as more AI models emerge, heterogenous computing is key to ensuring AI infrastructure's flexibility, availability, and versatility. For example, the CPU has demonstrated its capabilities in supporting classical ML algorithms that are difficult to parallelize for GPUs, and models using large-size data samples, such as Three-Dimensional (3D) data, for training and inference.

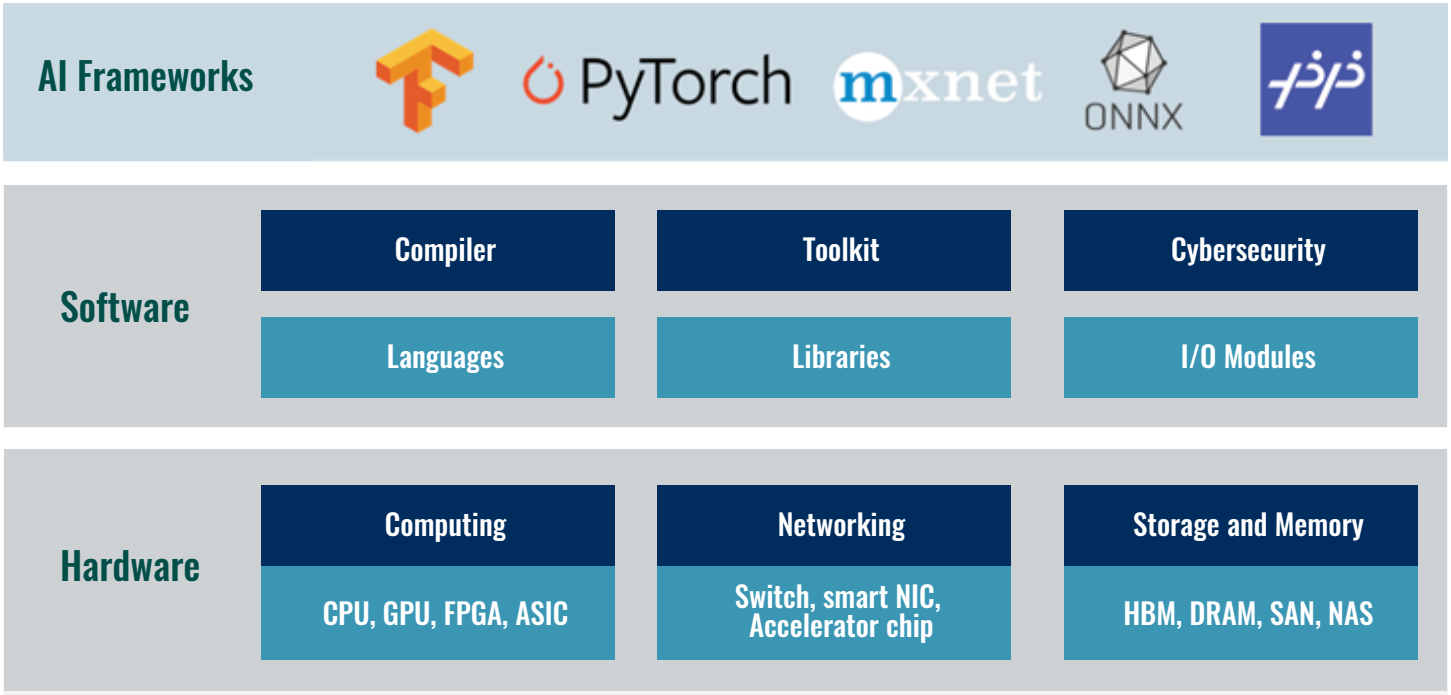
- **Storage:** Data are the most critical assets in AI training. To achieve high performance, the storage system for AI infrastructure needs to be smart at managing data flow, priority, and access. Key attributes include high read and re-read performance, good write I/O, multi-tier storage hierarchy, public cloud access, multi-protocol support, security, and extensible metadata to facilitate data classification. For businesses dealing with ultra-large datasets, purpose-built architecture is more cost-effective.

- **Networking:** As AI requires large amounts of data for training and testing, the system needs a network topology that can handle a high volume of both north/south (server to storage) traffic and east/west (server to server) traffic. Dedicated chipsets have emerged in recent years capable of orchestrating data movement from different storage tiers, moving data from the edge to the primary storage and data lake, then into the data prep staging tier, and finally into the training cluster.

- **Software:** The final component of sound AI infrastructure is an AI development platform that supports major AI frameworks and all the reference models, kernel libraries, containers, firmware, drivers, and tools. The platform must perform infrastructure provisioning and management, orchestration, and job scheduling during AI training and inference. Once AI is tested and validated, it must be deployed in the field. AI workflow management tools help ensure data management and monitoring, bias and model drift, model retraining, and upgrade. In other words, developers should not need to manage their AI workloads manually. Instead, the software should orchestrate the entire process, enabling developers to do what they are best at—developing innovative AI models that address key business pain points.

Figure 6: AI Technology Stack

(Source: ABI Research)



EDGE-TO-CLOUD VISION

An edge-to-cloud strategy should include processing data, and training, deploying, monitoring, and maintaining models at every node of the distributed computing architecture. Understanding the nature of each AI application allows businesses to deploy AI models at the most optimal compute node.

For example, if an AI model requires large amounts of resources and ultra-high precision, but has higher latency tolerance, it can be deployed in a hybrid or private cloud. The cloud infrastructure allows businesses to flexibly scale up the AI model without resource constraints. However, for AI models that are mission critical and require ultra-low latency, the AI model needs to be deployed at the edge. Businesses must also understand how their AI models integrate with emerging decentralized technologies, such as blockchain and edge computing.

Therefore, businesses must understand the types of AI workloads they want to run and test the deployment thoroughly to ensure that it meets their needs.

OPENNESS

AI models are often built using popular open software and an open-source AI framework, such as TensorFlow and PyTorch, which leads to a growing open AI community. This open community works collaboratively, continuously updating the various databases and data ingestion protocols, open Operating Systems (OSs), and open-source AI libraries for public usage. This openness creates a horizontal playing field where innovation thrives.

Therefore, sound AI infrastructure must be able to support all major open software, providing AI developers with all required libraries and drives. These features must be fully optimized, tested, and supported. These should be packaged in a rich Software Development Kit (SDK) that is easily downloaded, installed, and ready to use. The system should also be able to support data scientists bringing their data and AI models. These features enable developers to focus on developing and launching innovative apps, rather than spending time resolving coding and interoperability challenges.

After all, the true value of AI is in the business impacts enabled by all AI applications. Solution suppliers should position themselves as innovation enablers, rather than controlling system designs using proprietary hardware, software, tools, and services. They should allow all their partners, including the Original Equipment Manufacturers (OEMs), AI developers, and system integrators, to build reference designs, procure adequate components, and implement AI applications as a function of market demand.

SECURITY INFUSED AT EVERY LAYER

Currently, most AI frameworks, toolkits, and applications available do not implement security, relegating them to disconnected experiments and lab implementations. As data are the critical ingredients of AI, lax data security opens up unnecessary risks. Businesses will face challenges in ensuring data integrity, privacy, and compliance. With the rapid evolution of physical and cybersecurity threats, organizations are expected to leverage third-party solutions or services to safeguard their data.

Therefore, AI infrastructure suppliers must demonstrate capabilities to define, classify, secure, and protect data and AI models according to priority and sensitivity. They should also allow AI developers to partition their applications into a physically isolated environment, such as a secure enclave in a CPU or a security chip like a Trusted Platform Module (TPM) to help increase application security. These solutions employ the latest security protocols end-to-end through authentication, authorization, impersonation, and encryption. They must also be subjected to extensive security scanning and penetration testing, from data acquisition and preparation to training and inference. Lastly, clear documentation must be created to define all procedures for safeguarding information.

BACKWARD COMPATIBILITY

The AI of tomorrow will be tightly integrated with new applications, such as digital twins, simulations, the metaverse, edge computing, and blockchain. They may even be supported by futuristic computing infrastructure consisting of neuromorphic chipsets and quantum computers. As a result, AI will become more pervasive and integral than ever before.

However, before that vision is reached, today's businesses must ensure that whatever AI technology they deploy is compatible with existing business software and applications. Large enterprises run various hardware and software to support custom applications, including

Computer-Aided Design (CAD), Manufacturing Execution System (MES), Enterprise Resource Planning (ERP), Supply Chain Management (SCM), Warehouse Management System (WMS), Customer Relationship Management (CRM), Quality Assurance (QA), Environment, Health, and Safety (EHS), Human Resources (HR), sales enablement, and marketing software. According to McAfee, the largest enterprises run nearly 800 custom applications, many supported by specific and proprietary hardware.

Businesses need AI to support all of this hardware and software for AI to achieve its maximum value. Nowadays, AI solutions can be backward compatible with these systems through data sharing and exchange. However, this is far from ideal, as businesses will continue to expand existing infrastructure or add new infrastructure. Therefore, the perfect approach would be running both legacy and new enterprise software in virtual machines or containers on a broad and ubiquitous platform that can scale quickly and optimize for AI. This approach will help protect existing investment and optimizes productivity across IT and OT systems.

KEY TAKEAWAYS AND RECOMMENDATIONS FOR END USERS

Aside from the key pillars, there are some steps implementers should take if they want to introduce AI into their businesses.

TAKEAWAY 1: DEVELOP A CLEAR INTERNAL AI ROADMAP BASED ON BUSINESS OUTCOMES

The first move of implementers is to formulate a clear AI roadmap, namely a set of guiding principles to establish the right technology, people, and processes that lead to the creation of AI for businesses.

The roadmap must start with figuring out organizational capabilities in AI. A thorough evaluation of human resources, enterprise data management, IT/OT hardware, and software should reveal the level of AI maturity and readiness. As data have become increasingly embedded with daily decision-making, organizations must pay close attention to any employee that collects, processes, and uses data. Naturally, the primary concern is hiring the right data scientists, AI implementers, and DevOps engineers. Beyond human resources, organizations must appoint the right personnel to take charge of deploying the correct data and AI infrastructure, managing day-to-day AI operations, overseeing AI security and governance policies, and monitoring and mitigating regulatory, legal, and ethical risks.

Then, implementers must develop and refine AI capabilities in accordance with the evaluation outcome. This includes establishing elaborate plans and processes to generate and analyze valuable data, identifying the right AI framework, hardware, and software for AI training and inference, and building DevOps and Machine Learning Operations (MLOps) processes. When the AI model is ready, implementers must seek approvals on risk, legal, and security before experimenting, troubleshooting, and validating against business outcomes. Continuous engineering and governance are critical to ensuring an AI model's smooth running.

Lastly, implementers should seek advice and increase their collaboration with ecosystem partners that have managed to gain experience in implementing AI and in developing adequate solutions tailored to the business needs of various enterprise players. They should avoid closed implementation and innovation approaches often driven by a single player and adopt a more open approach, involving best-of-breed solutions from a wide selection of ecosystem partners and system integrators.

TAKEAWAY 2: GET ORGANIZATIONAL BUY-IN

Once the business outcome and AI infrastructure roadmap have been determined, AI implementers need to share it with key stakeholders to get full buy-in. More often than not, this is a very challenging task. The organization must bring down the data silo, create cross-functional teams, and commit budgets to fully implement AI. Without all these, any attempt to design, develop, and deploy AI would fail.

Starting with high-value business outcomes that cannot be easily achieved without AI is essential. AI implementers must demonstrate the importance of AI in driving the desired result. Senior management will appreciate the importance of AI once they can validate the business outcome. Constant communication and bringing the direct and indirect teams to support the initiative throughout the process will be vital for success.

TAKEAWAY 3: FOCUS ON SOLUTION PROVIDERS AND CHIPSET SUPPLIERS THAT EMBRACE OPENNESS, FREEDOM OF CHOICE, TRUST, AND SECURITY

AI is a continuous process of integrating, monitoring, and upgrading. Investing in heterogenous computing architecture that includes the CPU, GPU, FPGA, and ASIC helps future-proof AI infrastructure, thereby protecting its long-term value. Therefore, AI implementers need to constantly identify ways to harmonize existing heterogeneous AI solutions through solutions from AI technology vendors. Furthermore, AI implementers need to pay attention to software support, integration with open standards, compliance with legal and ethical frameworks, and cybersecurity, while making hardware choices. Vendors that embrace openness, freedom of choice, trust, and security are in the best position to serve this need.

To ensure that vendors are not just paying lip service to these values, AI implementers must scrutinize all the claims, track records, and contributions to open-source communities and academia. Select only vendors with solid track records and wide installed bases for scale-out and scale-up of AI applications, while contributing continuously to the open-source community to lower entry barriers and drive innovation. In addition, having a trusted partner will enable AI implementers to start ahead of the curve.

TAKEAWAY 4: LEVERAGE SUPPORT FROM ECOSYSTEM PARTNERS

Lastly, the importance of partnerships and collaborations is paramount to being successful in AI. As mentioned earlier, a successful AI deployment requires an optimized data, hardware, and software strategy. By combining expertise from different vendors, AI implementers can leverage cutting-edge technology without the need to develop their solutions in every aspect of the AI technology, significantly reducing costs and time to market.

CASE STUDY: INTEL

Intel is a market leader in CPU and FPGA technology. Over the years, the company has been slowly building up its product portfolio in heterogeneous computing, as it understands the importance of heterogeneous computing for AI workloads. In recent years, Intel has launched its Intel® Data Center GPU Flex Series and Intel® Data Center GPU Max Series, while acquiring Habana Labs, an Israel-based data center AI ASIC vendor, for its Habana® Gaudi® and Habana® Greco™ chipsets.

At the same time, Intel continues to invest in open software solutions that are hardware agnostic. Intel's goal is to abstract the hardware complexity to developers, while enabling them to optimize and integrate their models with the best in-class hardware that is commercially available. For example, OpenVINO, Intel's machine vision DL inference framework, will support even DL AI inference chipsets. In addition, Intel intends to offer the most open cross-architecture programming model for AI chipsets for all major vendors through oneAPI. oneAPI is an open, cross-architecture programming model that frees developers to use a single code base across multiple architectures. The company announced an overhaul of the governance structure for oneAPI. Another key initiative is Developer Cloud, which provides access to Intel products from a few months up to a full year ahead of the product. Developers can use Intel Developer Cloud to create and optimize their AI solutions across various hardware configurations. Lastly, Intel has also introduced Geti, a low-to-zero code development platform for AI-based computer vision applications. The solution works with DL models in several formats, including Google's open-source TensorFlow framework and open-source OpenVINO toolkit files for Intel CPUs, GPUs, and Vision Processing Units (VPUs). By introducing these solutions, Intel hopes developers can appreciate the openness and ease of development when using Intel's AI software tools.

The commitment to AI hardware and software creates the foundation for Intel's ecosystem to develop several industry-leading AI models. In August 2020, Intel introduced VisualCOMET, a framework that enables common sense reasoning using images, enabling cognitive-level understanding beyond the scope of image classification, object detection, activity recognition, or image captioning. At the Conference and Workshop on Neural Information Processing Systems (NeurIPS) in 2021, the company's multimodal and multihop WebQA made it onto the winning list of the NeurIPS 2021 Competition. The model is a multimodal encoder and fusion-in-decoder architecture that enables incorporating multimodal sources into the language generation model. In August 2022, Intel released a video retrieval model based on multilingual knowledge transfer. The multimodal embeddings from the video were trained using Intel Gaudi accelerators, while the graph-based similarity search was performed on the Intel Xeon CPU.

Intel has also made a deliberate choice to highlight its expertise in hardware security. The goal is to reassure developers and ecosystem partners that Intel will pull out all the stops to help ensure the safety and security of innovative ideas, mainly through hardware-based memory encryption under its Software Guard Extensions (SGX) solution. The solution is a set of security-related instruction codes embedded in an Intel CPU, allowing developers to allocate specific parts of the memory for more secure computational workloads. This solution isolates AI codes and data in memory, helping prevent unauthorized access and deterring malicious behavior by bad actors that wish to prey on ideas from smaller startups and independent developers.

All these moves indicate Intel's intention to position itself as one of the most open and secure enabler for AI developers. These assets are designed to cater to the huge developer community around its AI solutions, which will likely spur innovation for all of Intel's Original Equipment Manufacturer (OEM) partners in the AI market. The company already has a comprehensive list of AI hardware and available software solutions. Still, Intel intends to provide AI developers with even more choices in the future through hardware-agnostic software, a wide range of industrial partnerships, optimization and integration tools to reduce the cloud computing costs to developers, and open software development solutions to free them from potential vendor lock-in with proprietary solutions. To assist with this, Intel has acquired Granulate, a data center workload optimization startup, and cnvrg.io, a heterogeneous computing platform for data science and AI workloads. Both offer hardware-agnostic solutions across multiple AI hardware vendors, helping to facilitate AI adoption and development, while making Intel a trusted partner in their AI development and commercialization process.

Published December, 2022
ABI Research
157 Columbus Avenue
New York, NY 10023
Tel: +1 516-624-2500
www.abiresearch.com

ABOUT ABI RESEARCH

ABI Research is a global technology intelligence firm delivering actionable research and strategic guidance to technology leaders, innovators, and decision makers around the world. Our research focuses on the transformative technologies that are dramatically reshaping industries, economies, and workforces today.