GLOSSARY OF TECHNICAL TERMS

This glossary of technical terms contains explanations of certain technical terms used in this document. As such, these terms and their meanings may not correspond to standard industry meanings or usage of these terms.

"AI"	artificial intelligence, an area of computer science that focuses on simulating human intelligence by machines
"AI Computation Blocks"	specialized hardware units designed to efficiently execute core mathematical operations required for artificial intelligence workloads. These blocks are optimized for matrix multiplications, convolutions, attention mechanisms, and other tensor operations that dominate deep learning models
"AI inference"	the process of using a pre-trained model, whose parameters have already been derived from training, to execute the intended tasks by processing the new, unseen inputs through its learned patterns or knowledge, and generating outputs, such as predictions, classifications, or decisions
"AI inference chip"	a specialized integrated circuit designed to execute trained artificial intelligence models with high computational efficiency, minimal latency, and optimized power consumption, focusing exclusively on deriving outputs, such as predictions, classifications, or decisions, from the new, unseen input data
"AI model"	mathematical algorithms which can take unstructured data as input and transform them into informative outputs through its "intelligence," namely, the capability of perceiving the world, transcribing and organizing information, enhancing or generating contents, or making decisions
"AI training"	the process of teaching a machine learning model to recognize patterns or make predictions by exposing it to labeled or structured data. During training, the model iteratively adjusts its internal parameters (e.g., weights, biases) using algorithms like gradient descent to minimize the difference between its predicted outputs and the actual "correct" outputs (loss). This phase requires significant computational resources (e.g., GPUs/TPUs) and large datasets, and it produces a trained model ready for inference
"algorithm"	a procedure or formula for solving a problem, based on conducting a sequence of specific actions, especially by a computer
"algorithm-to-silicon"	a co-design methodology that directly maps computational algorithms to custom silicon architectures, optimizing hardware for specific workloads
"ASIP-LSF"	a specialized processor architecture that combines Application-Specific Instruction-Set Processors (ASIPs) with an optimized Load-Store Framework (LSF), which is designed to accelerate domain-specific workloads by tailoring the instruction set and memory access patterns to the target application

GLOSSARY OF TECHNICAL TERMS

"big data"	extremely large and complex datasets that exceed the processing capabilities of traditional data management tools. The field of big data analytics focuses on examining these large datasets to uncover patterns, correlations, and insights that can inform decision-making and strategic planning across various sectors
"Bluetooth"	a short-range wireless technology standard for data exchange between fixed and mobiles devices over short distances
"chip-to-chip mesh torus interconnect," or "C2C mesh torus interconnect"	a high-performance, scalable network topology used in multi-chip systems where processing units are connected in a torus or mesh configuration
"chiplet"	a small, modular integrated circuit that encapsulates a specific function, which is designed to be combined with other chiplets within a single package to form a complete system-on-chip (SoC), and to enable integration of multiple chiplets into one package, thereby scaling up the performance of the resulting chip
"cloud"	a network of remote servers hosted on the Internet/Intranet and used to store, manage, and process data in place of local servers or personal computers
"cloud computing"	the practice of storing computer data and programs on multiple servers that can be accessed through the internet
"cluster"	a group of interconnected computers or servers that work together as a unified system to improve performance, reliability, or scalability
"CNNs"	a specialized deep learning architecture designed for processing grid-structured data by automatically learning hierarchical spatial features through convolutional operations
"computing power"	the ability of a computer to perform an operation
"CPU"	central processing unit, a complex electronic circuitry assembly that executes a machine's operating system and applications, serving as the core for general-purpose computing
"deep learning"	a subset of AI and machine learning that mimics the working of biological neural systems such as human brains and uses multi-layered neural networks to deliver state-of-the-art accuracy in tasks such as object detection and recognition, speech recognition and natural language processing. Deep learning differs from traditional machine learning techniques in that it can automatically learn representations from data such as images, video or text, without introducing hand-coded rules or human domain knowledge. Its highly flexible architecture can learn directly from raw data and can increase its predictive accuracy when provided with some data
"Die"	an individual chip cut from a wafer before being packaged

GLOSSARY OF TECHNICAL TERMS

"die-to-die chiplet" or "D2D chiplet"	a high-bandwidth, low-latency communication technology that enables multiple chiplets to function as a single integrated system
"edge"	a distributed computing paradigm where data is aggregated through gateways near the data source and processed by local computer systems to provide immediate services. By eliminating the need for cloud transmission, this architecture meets critical industry requirements including real-time operations, application intelligence, and security/privacy protection. The infrastructure is typically deployed between end devices and cloud platforms
"edge computing"	a distributed computing paradigm that brings computation and data storage closer to the location where it is needed to improve response times and save bandwidth
"GPU"	graphic processing unit, a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images
"home-grown"	developed, built, or created entirely within an organization using its own resources, expertise, and infrastructure
"integrated circuit"	electronic circuit fabricated on semiconductor substrate
"IoT"	the Internet of Things, which refers to the network of physical devices embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data, allowing these devices to communicate with each other and with users
"ISA"	instruction set architecture, a standardized set of commands, operations, and rules that define how a processor interacts with software. It acts as the hardware-software interface, governing how programs control the processor, access memory, and execute computations
"KOL"	key opinion leader
"LLMs"	advanced AI models based on deep learning architectures trained on massive text datasets to understand, generate, and manipulate human-like language
"LSTMs"	a specialized type of Recurrent Neural Network (RNN) designed to model long-term dependencies in sequential data
"near-memory hyper-converged architecture"	an advanced computing paradigm that radically optimizes data processing efficiency by physically integrating computational units with high-bandwidth memory systems
"NPU"	neural processing unit, a microprocessor that specializes in the acceleration of machine learning algorithms, typically by operating on predictive AI models, generative AI models, and reasoning AI models
"operator"	prebuilt functions used to pre-process raw data

GLOSSARY OF TECHNICAL TERMS

"R&D"	research and development
"RNNs"	a type of neural network designed to process sequential data by maintaining a "memory" of previous inputs through hidden states
"SoC"	an integrated circuit that incorporates functionally diverse sub-modules within a single chip, forming a complete system tailored for specific application scenarios. Typically integrating multiple heterogeneous components, an SoC combines elements such as general-purpose processors, hardware-based codec units, baseband processors, and other specialized functional blocks into a unified silicon die
"tape-out"	the final stage of integrated circuit design before mass production, where the design is sent to a foundry for fabrication on a silicon wafer
"TOPS"	Tera Operations Per Second, a unit of measurement representing a processor's capability to perform one trillion (10^{12}) operations per second, commonly used to quantify the inference performance of AI accelerators, NPUs, or other deep learning hardware.
"terminal"	hardware devices that perform localized data processing, deliver real-time responses, and interact physically with users or environments, while offloading complex tasks from edge nodes or cloud platforms