WHITE PAPER

Artificial Intelligence
Intel® Xeon® processor
Intel AI® Builders

Intel® Xeon® Processor: A Workhorse for Enterprise AI

Author

Bhushan Desam, Ph.D
Artificial Intelligence Global Business Leader, Lenovo Data Center Group

Is your enterprise implementing artificial intelligence (AI) to be more efficient and to bring predictive capabilities for making better business decisions? Like many customers that we speak to, you may be considering various use cases that may involve tabular data, images, text, or sensor data from IoT and other sources. A common concern that many IT managers are now facing is the optimal data center infrastructure to support these varying and growing demands in organizations.

For developing AI algorithms, there are two fundamental architectural choices (i) scale-up nodes where computing is concentrated in a single node with densely-packed accelerators, typically purpose-built for neural network training or (ii) scale-out clusters in which computing resources are distributed but connected over a network. The latter is more common in enterprises like a cluster of Intel® Xeon® processor-based servers, running various workloads like databases, analytics, and even high-performance computing (HPC) over high-speed networks. How do you decide which one is the best option? To make such a choice, several factors need to be accounted for including life cycle of AI projects, a variety of use cases, and ultimately the best return on investment (ROI).

Scale-up nodes are ideal for quick experimentation or for large full scope training from scratch employing deep neural networks where time-to-train is critical. These systems require large amounts of clean data which many enterprises readily lack. In an end-to-end workflow, data preprocessing is a key task and may typically take 60-70% of the total time, which could be the limiting factor for continuous training operations. In such cases, the high acquisition cost that demands maximum system utilization to realize the investment is not justified.

A common scenario in enterprises, as AI projects progress, is that there will be multiple teams of data scientists and engineers preparing data, simultaneously developing algorithms and applications and deploying those in production. Further, the algorithms may encompass traditional analytical methods to classical machine learning to deep learning based on the available data and accuracy requirements. Such scenario demands flexible computing capacity to satisfy most tasks not just training deep neural networks (DNN). Even in that specific case, the computing requirements significantly vary from CNN to RNN, LSTM etc and whether

the training is from scratch or from a previous model such as in retraining, transfer learning, and hyperparameter tuning. To satisfy these diverse needs, you would need a system that provides flexibility while delivering the best ROI.

Meet the most familiar face in data centers, the Intel Xeon processor, which is the workhorse for many enterprise and HPC applications. By creating a cluster of servers powered by Intel Xeon processors over a high-speed ethernet or Intel® Omni-Path architecture network, it is possible to satisfy the needs of individual projects while allowing multiple users to take advantage of the cluster at the same time for various tasks including deep learning. Those workloads that demand the most computational resources like CNNs or training from scratch may be satisfied by adding few accelerators per node such that the TCO is optimal while meeting the performance demands. For this scale out approach to work well, enterprises need complementing software tools to manage the clusters and workloads efficiently as many are new to the area of distributed computing.

To help enterprises take advantage of the flexibility offered by scale out clusters and to simplify AI development, Lenovo launched LiCO* 5.1 (Lenovo intelligent Computing Orchestrator) optimized to run on its scalable infrastructure comprising SD530 and SR 650 powered by Intel Xeon processors. LiCO supports Intel® Distribution for Caffe*, Intel® Distribution for TensorFlow*, and the neon™ framework through containers and helps manage various frameworks along with different versions for multiple developers on the same cluster. Further, it simplifies launching training jobs even in distributed environments and help monitoring the training progress using objective functions. For users that prefer simplified workflows, LiCO offers end-to-end image classification in Intel Distribution for Caffe and transfer learning that eases training data and computing requirements. Currently, researchers at North Carolina State University (NCSU) are using LiCO to build deep learning models to improve accuracy of geospatial image analytics.

Enterprises can start implementing AI projects right away in their existing data centers with the flexibility that Intel Xeon processors offer and with powerful tools Lenovo provides to realize the benefits of AI.

Learn More

For more details about Lenovo, visit https://www3.lenovo.com/us/en/

For more details about Intel® Xeon® processors, visit www.intel.com/xeon