Right Sizing AI/ML for Small and Medium Size Deployments

Presenter: Glenn Fuller, Senior Director of SW Engineering at Viking Enterprise Solutions

FMS
the Future of Memory and Storage

Overview

- Benefits of AI/ML for Small and Medium Size Deployments
- Challenges and Barriers to Entry
- Alternative HW Solutions
- SW Solution and Learning Models
- Path Forward

Benefits of AI/ML

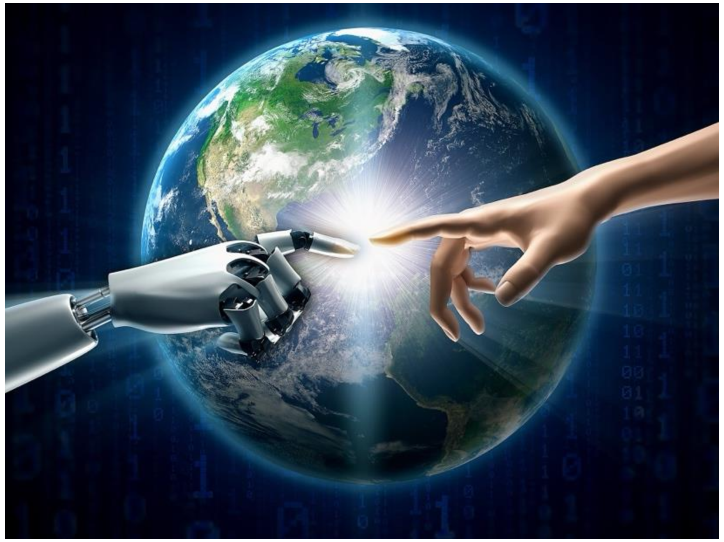
- Increased efficiency of data analysis resulting in more focused product development
- Improved customer experience through better analysis and identification of trends and issues
- Improved data management enabling better decision making and analytics
- Improved staff efficiency through more effective data analysis and management
- Improved risk management

Challenges and Barriers to Entry

- At least 50% of all AI deployments do not reach production
- Development of learning models
- Knowing what data and how to manipulate data for models
- Acquisition and operations costs
  - Access to high end systems on a pay per use basis can be expensive
  - Acquisition cost of high end systems is very high
  - Power and cooling requirements of high end systems is prohibitive
  - Existing IT infra-structure is based on traditional air cooled solutions
  - Technical capability and “know-how:” associated with AI

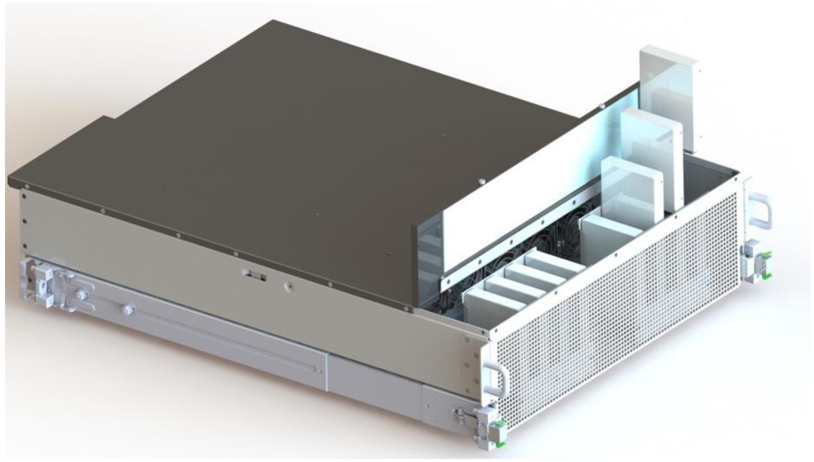
Alternative HW Solutions

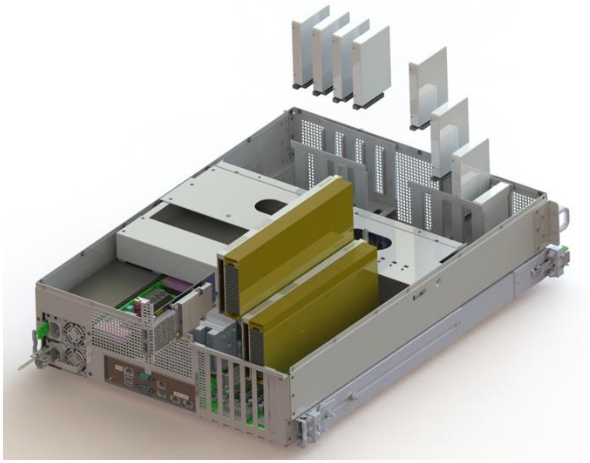
- Consolidated solutions that combine CPU, SSDs and lower power GPUs in a single, air cooled package
  - Leverages existing IT deployment model without moving to liquid cooling
  - Lower power fits within existing IT footprints
  - Lower processing power than large scale solutions, but at a fraction of the cost
- Lower total cost of acquisition and ownership for solution
- Targeted solution to AI/ML applications for small and medium deployments without causing major disruption to existing operations



Example Small/Medium Solution

- Example solution for small and medium size enterprise
- Support for 3x dual width, FHFL GPU
- Air cooled and compatible with existing IT infra-structure





SW Solutions

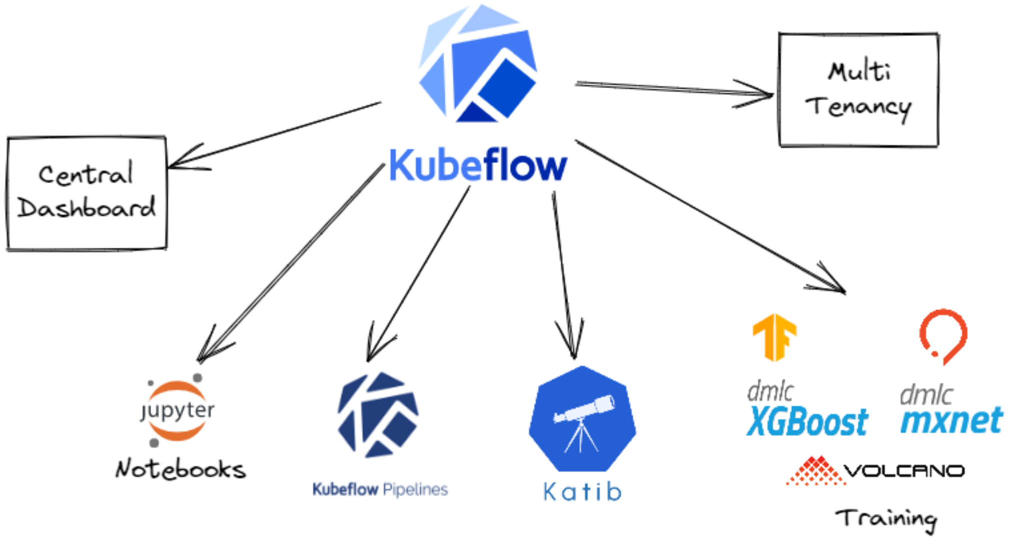
- Users need to become more efficient and cannot afford cloud based solutions
- Data is being collected but it’s not used to improve the business
- Cloud based solutions expect all data to be stored in their cloud
- All GPUs are rented by usage and costs are prohibitive
- On site hardware solutions separate the GPUs from the storage systems
- Networks become the bottleneck
- Integrating AI/ML software with existing hardware requires skilled developers

Cloud Native Orchestrator

- Kubernetes based storage and application orchestrator uses cloud technologies
- Using containers on bare metal provides the best performance
- Orchestrators can be used to provide both storage and applications in the same box
- Benefits are that storage is on the same bus as the GPUs
- Mitigates the network bottleneck caused by separate GPU and storage
- Users benefit from lower costs than the cloud
- Also gives customers control over the data and information
- Providing a complete AI/ML ecosystem allows customer to create their own models without the expense of cloud based systems

Cloud Native Orchestrator

- Pipelines can be hosted in the Kubernetes cluster for both inference and training pipelines
- ETL is performed at the edge on the incoming data stream
- Inference engines such as Triton and TensorRT are utilized in the cluster
- Pipelines can be built with Kubeflow using containers for specific purposes such as HiveMQ, Kafka, Flink, etc.
- Developers can load their own containers and also load subscription software
- Instead of developing in the cloud, perform training on the edge device
- Download the trained model to run in other edge systems



At The Edge

- IOT devices generate data that must be either stored locally or sent to the cloud
- Data in the cloud is kept there and is charged every month
- The solution is to store the data at the edge
- Perform inference and generative AI at the edge
- Data is processed as it’s collected so information can be acted on quickly
- Replication moves data that can be stored in a private or public cloud
- Models can be downloaded or new applications installed as needed
- Edge devices with high performance GPUs can be used for training
- Similarly, edge devices with lower performance GPUs can be used for inference

Path Forward

- Smaller and more focused HW deployments
  - Lower acquisition cost
  - Targeted to small and medium size deployments
  - Leverages existing IT infrastructure and deployment models
- Leverage commercially available models
  - Existing and proven models can be customized for application
  - Simple user interface requires less in-house SW expertise to deploy
- Implement cloud native systems
  - Keeps data secure on site
  - Provides cloud based technologies on premise
  - Lower TCO than public cloud services saves time (latency) and money

FMS