Best Practices in Designing and Deploying End-to-End HPC and AI Solutions

Kevin Tubbs Ph.D.,
Sr. Director, Advanced Solutions Group

GTC 2018

About Penguin Computing

- 20-year-old, U.S.-based global provider of high-performance computing (HPC) and data center solutions with more than 2,500 customers in 40 countries, across 8 major vertical markets

- Comprehensive portfolio of Linux servers and software, integrated, turn-key clusters, enterprise-grade storage, and bare metal HPC on cloud, as well as expert HPC and AI services, financing, and top-rated support

Growth Has Resulted in Gaps in AI Design

Explosion of data has resulted in huge growth in AI systems but many designs fail to address the most critical challenges, failing to:

- Balance researchers' workflow and technical compute needs
- Sizing AI infrastructure to optimize utilization
- Balancing data workflows and performance
- Deploying AI infrastructure efficiently

Ways to Address Gaps

- Flexible hardware and software required to meet research and workload needs
- Well-designed, balanced systems are required to ensure data and compute achieves the overall production needs
- Fast and optimized storage necessary to deliver large amounts of data required to improve accuracy
- Deploying AI infrastructure efficiently requires balancing power, cooling and performance

Essentials on How to Achieve These Objectives

Designs and product must take into account these essential components of AI system design:

- Managing Software Ecosystem, focusing on Orchestration and Workload portability
- Accurately Sizing AI Infrastructure
- Balancing Overall Data workflows and performance
- Deploying and Scaling AI Infrastructure Efficiently

Managing AI Software Ecosystem

Managing AI software ecosystem for infrastructure designs must balance researchers' workflow and technical compute needs by focusing on:

- Managing System and Application Orchestration and Workload Portability
- Curating Deep Learning Frameworks such as TensorFlow, PyTorch, Caffe2, MXNet
- Providing Libraries and SDK's optimized for hardware

Accurately Sizing AI Infrastructure

Ensuring optimal use of compute resources can be challenging, particularly in heterogeneous architectures.

- CPU Based architectures (x86 , Power, Arm64)
- Specialty compute architectures (GPUs, FPGAs, TPUs and emerging ASICs).

Selecting the proper heterogeneous compute options based time to market/production, TCO and performance: **Nvidia GPUs**

Optimizing Training and Inference Workloads on Nvidia GPUs

- Intranode Communication: GPU Topology and NVLink
- Internode Communication: Infiniband and Ethernet Networks

Balancing Overall Data Workflows and Performance

The need to effectively manage large amounts of data, is at the center of a modern deep learning solution.

- Larger amounts of Data have varying ingest, processing and data life cycle requirements
- Storage Performance requirements dictated by data type, size, workflow stage, etc.
- Delivering storage to GPU clusters at high IOPs, Bandwidth and low latency is crucial

Fast and optimized storage is critical to deliver the large amounts of data required to improve accuracy

Deploying and Scaling AI Infrastructure

Designing, Deploying, and Scaling AI infrastructure efficiently involves optimizing at Rack, Datacenter and multiple Datacenter Scales.

Optimizing AI infrastructure at Rack and Datacenter level requires balancing power, cooling, and performance

Optimizing AI Infrastructure at multiple datacenter scales presents a challenge may include a shift from on-premise to hybrid to private/public cloud

Value of Moving Towards Open Technologies

PENGUIN COMPUTING

Open Architectures for Artificial Intelligence /Deep Learning

- Flexible compute architectures (x86,Power, ARM64, **Nvidia GPUs**, FPGAs, ASICs, …)
- Flexible rack scale platforms (EIA,OCP)
- Flexible Software Defined networking solutions
- Flexible software defined storage solutions

OPEN Compute Project

A Scalable Software-defined Artificial Intelligence / Deep Learning Environment

Proven AI Deployments Based on Essentials

Unnamed AI Cluster

Client deploying clustered, dense GPU nodes at scale, including

- NVIDIA DGX1 at scale
- Penguin Computing Relion 4118GTS at scale
- Strong storage component

Top 500 and Green 500

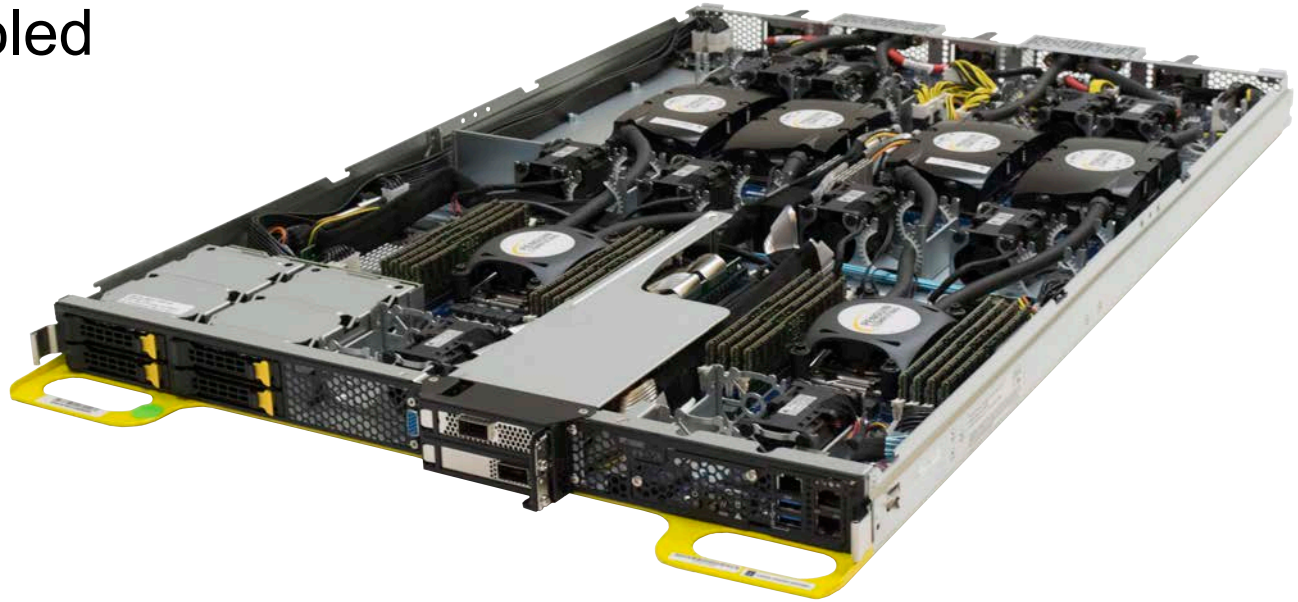
Proven AI Deployments Based on Essentials

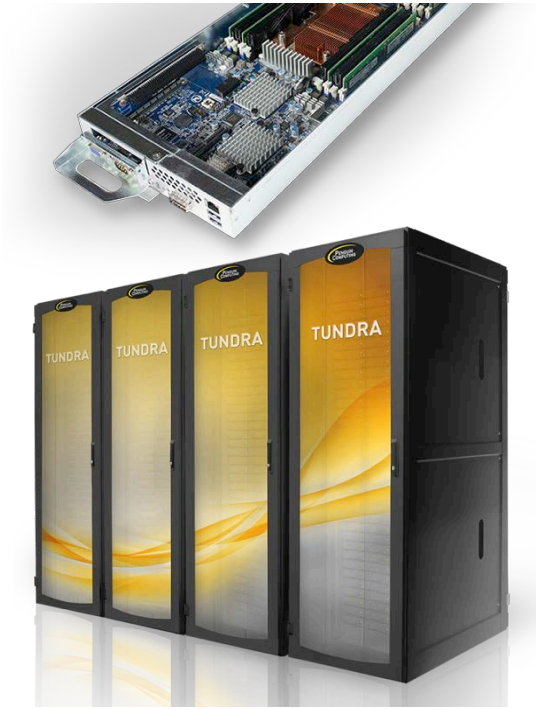
PENGUIN COMPUTING

Unnamed Hybrid AI and HPC Cluster

Client deploying clustered, OCP-compliant Relion X1904GTS servers

- Tundra® Extreme Scale HPC system
- Air Cooled
- Liquid Cooled





TOP 500 The List.

The GREEN 500

Summary of Best Practices

- Manage AI software ecosystem to balance workflow and compute needs by focusing on **orchestration and workload portability**.
- Select the proper heterogeneous compute options based **time to market**/ production, **TCO** and **performance**: **Nvidia GPUs**
- **Fast and optimized storage** required to deliver the **large amounts of data** required to improve accuracy
- **Deploying and Scaling AI infrastructure efficiently** involves optimizing power and cooling at Rack, Datacenter and multiple Datacenter Scales.
- Consider **open technologies** for a **scalable software-defined artificial intelligence / deep learning environment**

Question & Answer

For more information, visit www.penguincomputing.com/ai