Generative AI deployment basic guide

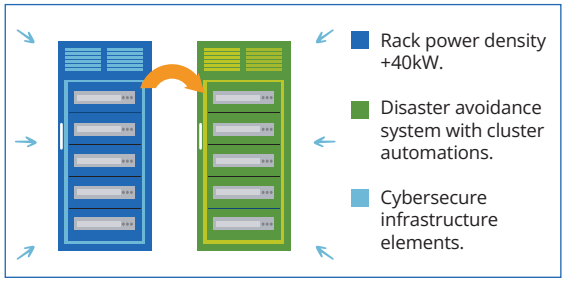
Power management your Generative AI infrastructure

Eaton's solutions for a proper Generative AI infrastructure deployment

Democratization of Generative AI applications and the evolution of the infrastructure

The future is now, and as a result of IT companies' endeavors, Gen AI applications have become more easily accessible than ever. Nevertheless, alongside the advancement of these applications, there must also be a corresponding evolution in the supporting infrastructure. Regardless of the unique traits found in different business sectors, they all share a common requirement for fundamental features like security, scalability and availability. The central challenge persists: how can we enhance the availability and resilience of our systems to enable efficient training and sustained inference.

And now think about verticals like healthcare, education or e-mobility. In all of these, their IT infrastructure will need to evolve in terms of rack power density, disaster avoidance systems and cybersecurity; and so these requirements are the base of the power management solutions Eaton has developed.



Explaining the Generative AI architecture

AI IT architecture encompasses the infrastructure and components necessary to support the deployment, training and inference of AI models. It involves two primary components: the training model and the inference model. Let us delve into each of these components.

Generative AI architecture

Training model

The training model refers to the phase where AI models are developed and refined using large datasets. This process involves feeding the model with labeled data, allowing it to learn patterns, relationships and make predictions. The training model architecture typically includes the following elements:

- **Training infrastructure module***
- **Data storage module***
- **Training algorithms**
- **Optimization and evolution modules**

Inference model

The inference model represents the phase where the trained AI models are deployed and utilized to make predictions or perform tasks in real-time. The inference model includes the following components:

- **Inference infrastructure module**
- **Input data processing**
- **Output presentation**
- **Scalability and performance***

*Requires Eaton next generation of power management solutions.

Overall, AI IT architecture for inference and training models requires careful consideration of infrastructure, computational resources, algorithm selection and optimization techniques to enable efficient and accurate AI deployments.

EATON
Powering Business Worldwide

TD SYNNEX

Eaton's new technologies for cost-effective, time-efficient and risk-reduced Generative AI deployment.

Reduce risk

Reduce cybersecurity risks with the first UPS network card to meet both UL 2900-1 and IEC 62443-4-2 cybersecurity standards.

Specialized fiber cable management that mitigates misconnection risks.

PredictPulse™

Remote monitoring service with predictive analytics will maintain your systems to run MTBF and avoid unscheduled downtime.

Save time

Self-contained deployment that includes in-row cooling and will protect your system in every environment and saves time with its inter-connection features.

Brightlayer Data Centers suite

With Brightlayer Data Center suites you can monitor the site, run intelligent automations and update firmware remotely saving a lot of on-site time.

Up to 46kW high-density power distribution units (PDU) with universal input making installation faster and with C39 outlets that let you connect C13 or C19 inputs.

Save money

High-efficiency UPS with up to 99% efficiency will lower your OpEx.

Scalable 3-phase system with redundancy capability for future CapEx savings.

10-year lifetime of lithium-ion batteries drastically reduces TCO.

Project Helix

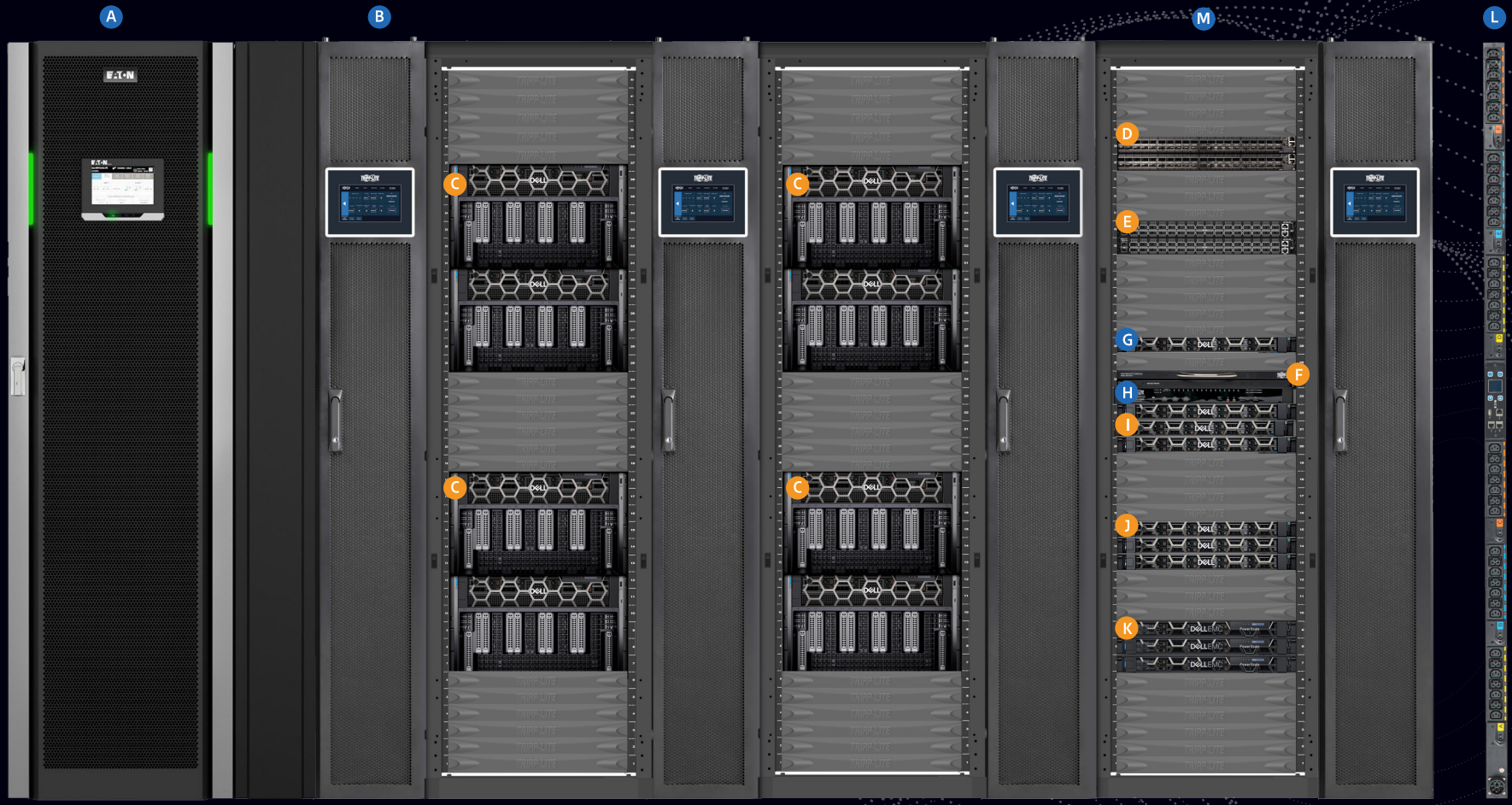
Dell Technologies and NVIDIA have announced a joint initiative called Project Helix to facilitate the development and utilization of generative AI models by businesses. The project aims to enable enterprises to deliver improved customer service, market intelligence, enterprise search and other capabilities securely and efficiently. Project Helix will provide a range of full-stack solutions built on Dell and NVIDIA infrastructure and software. It includes a blueprint that assists enterprises in leveraging their proprietary data and deploying generative AI models responsibly and accurately.

The initiative simplifies generative AI deployments by offering optimized hardware and software combinations from Dell. By utilizing Dell PowerEdge servers, NVIDIA GPUs and networking solutions, enterprises can convert their data into valuable insights while maintaining data privacy. All of these while security and privacy being foundational components of the project.

Overall, Project Helix enables businesses to deploy generative AI models, harness the value of their data and drive innovation in their respective industries quickly and securely. And so, this new generation of applications requires a specialized infrastructure that can handle it.

To learn more visit NvidiaNews.com.

Reference architecture: Large Model Training – minimum unit



- A 93PM 208V UPS scalable to 120kW w/ lithium-ion external batteries
- B 25kW In-row cooling system
- C Training module - PowerEdge XE9680 server
- D IB Module - QM9700 InfiniBand switches
- E Ethernet Module - Z9432F - ON Switches
- F Bright/Omnia Head R660 Server
- G Console with 19“ LCD
- H 16-Port Cat5 KVM over IP Switch
- I Management Module - R660 Servers
- J MLOPs/Data Prep Module - R660 Servers
- K Data Module - F600 PowerScale Storage
- L UPDU- 23kW universal PDU
- M 42U rack or self-contained deployment

Remote monitoring Service - PredictPulse

Intelligent power management - Brightlayer Data Centers suite

Reference architecture: Large Model Inferencing – minimum unit

- A 93PM 208V UPS scalable to 60kW w/lithium-ion external batteries
- B 25kW In-row cooling system
- C Inference module - PowerEdge XE9680 server
- D Ethernet Module - Z9432F - ON Switches
- E Bright/Omnia Head R660 Server
- F Console with 19" LCD
- G 16-Port Cat5 KVM over IP Switch
- H Management Module - R660 Servers
- I MLOPs/Data Prep Module - R660 Servers
- J Data Module - F600 PowerScale Storage
- K UPDU- 23kW universal PDU
- L 42U rack or self-contained deployment

Let your solution speak for you, as with this alliance you will achieve:

Grow	Expand	Develop
Sustainability High efficiency	Availability Scalability	Security Resilience
To learn more, click here	To learn more, click here	To learn more, click here

Remote monitoring Service - PredictPulse

Intelligent power management - Brightlayer Data Centers suite

For more information, visit **Eaton.com/Alliances**

Powering Business Worldwide

Eaton
1000 Eaton Boulevard
Cleveland, OH 44122
United States
Eaton.com

Publication No. BR152095EN / GG
July 2023

TD SYNNEX

Follow us on social media to get the latest product and support information.