NSTDA Supercomputer Center
ThaiSC

An AI Journey from Research to Operation:

A case from supercomputer to cloud

Putt Sakdhnagool

Supercomputing Infrastructure and Operation Team

NSTDA Supercomputer Center (ThaiSC)

putt.sakdhnagool@nectec.or.th

NSTDA Supercomputer Center (ThaiSC)

National Science and Technology Development Agency (NSTDA)
missions include the development of S&T infrastructure to support national STI development in Thailand

National S&T Infrastructure (NSTI)

NSTDA Supercomputer Center (ThaiSC): commissioned in 2019 to provide a world-class supercomputer facility for

1. Supporting Thailand's R&D needs for computational power
2. Addressing important and urgent national agendas requiring advanced computing resources
3. Promoting high-tech industries through advanced AI & computing.

Supercomputer at a Glance

LANTA SUPERCOMPUTER

*Peak performance at 8.15 PFLOPS

- 346 nodes Heterogeneous HPE Cray EX cluster
  - 176 GPU nodes with 704 NVIDIA A100 GPUs
  - 160 CPU nodes with 20,480 CPU-cores
  - 10 High-memory nodes, each contains 4TB of memory
- 10 PB of high-performance parallel storage
- High-performance interconnect using 200 Gbps

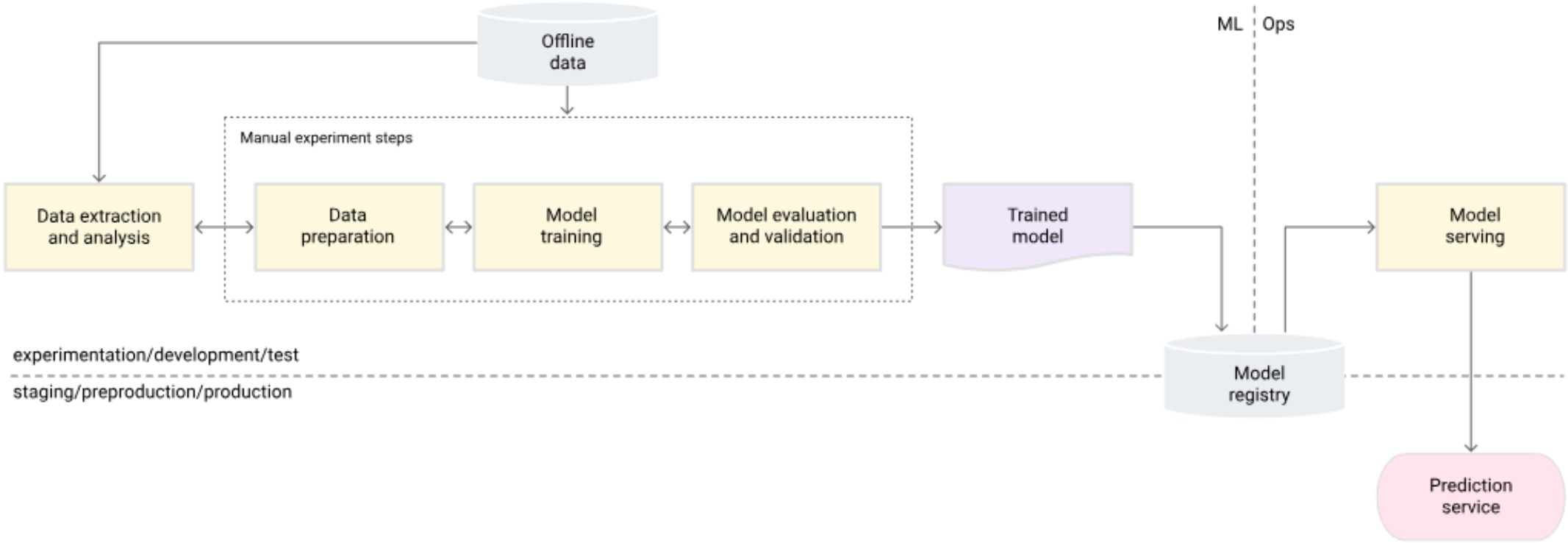
TOP 500 The List.

THE GREEN 500

NSTDA Supercomputer Center
ThaiSC

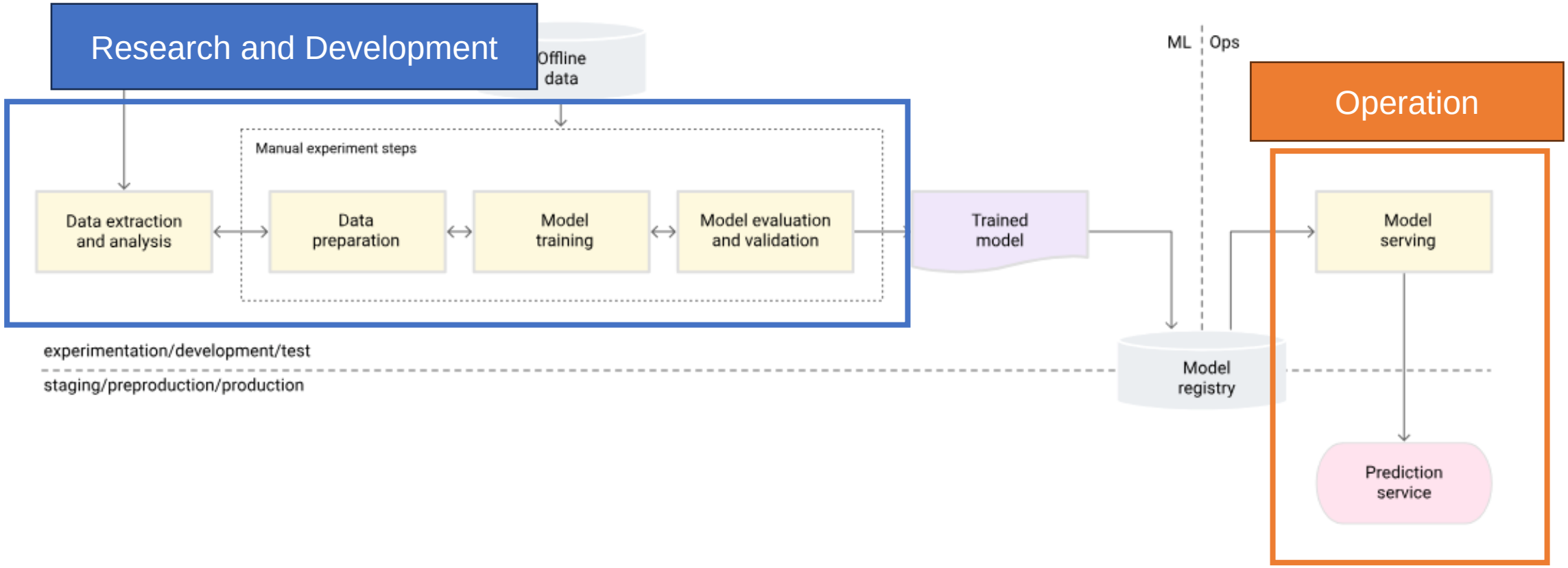
An AI Journey from Research to Operation

AI Journey: The Common Story



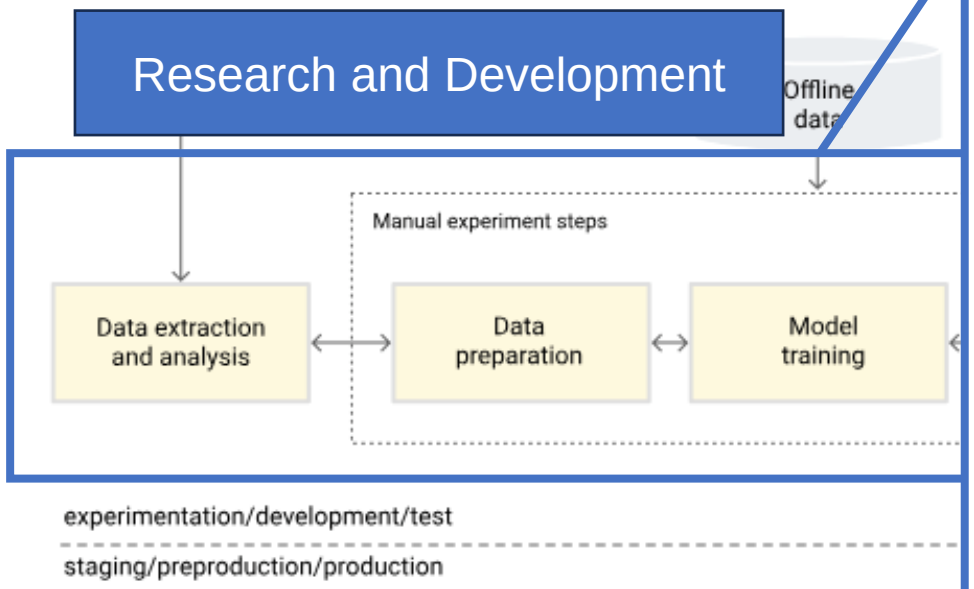
Source: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

AI Journey: The Common Story



Source: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

AI Journey: Research and Development

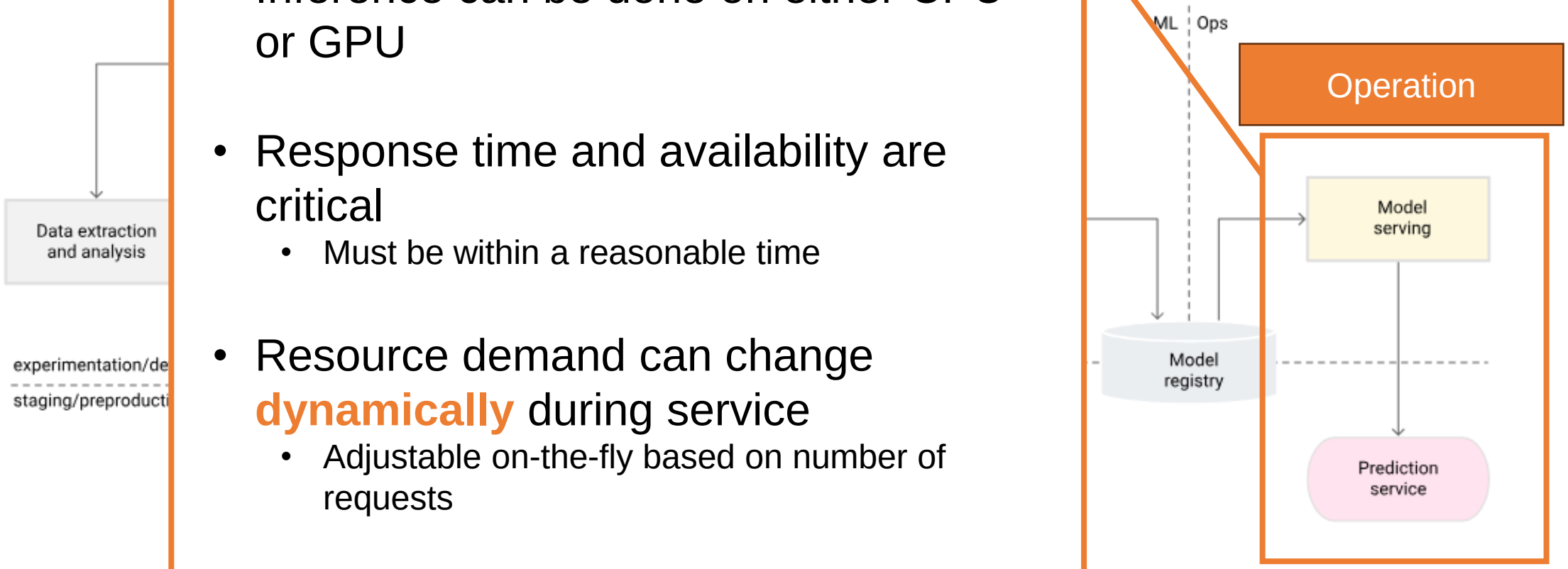


- Training large models requires a lot of resources
  - GPT-3 requires 1000s of GPUs for training
- Don’t expect immediate response
- Resource demand remains **constant** through the training
- Requires **HPC Cluster** or **Supercomputer** to train efficiently

Source: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

AI Journey: Operation

- Inference can be done on either CPU or GPU
- Response time and availability are critical
  - Must be within a reasonable time
- Resource demand can change **dynamically** during service
  - Adjustable on-the-fly based on number of requests
- **Cloud infrastructure** is more suitable

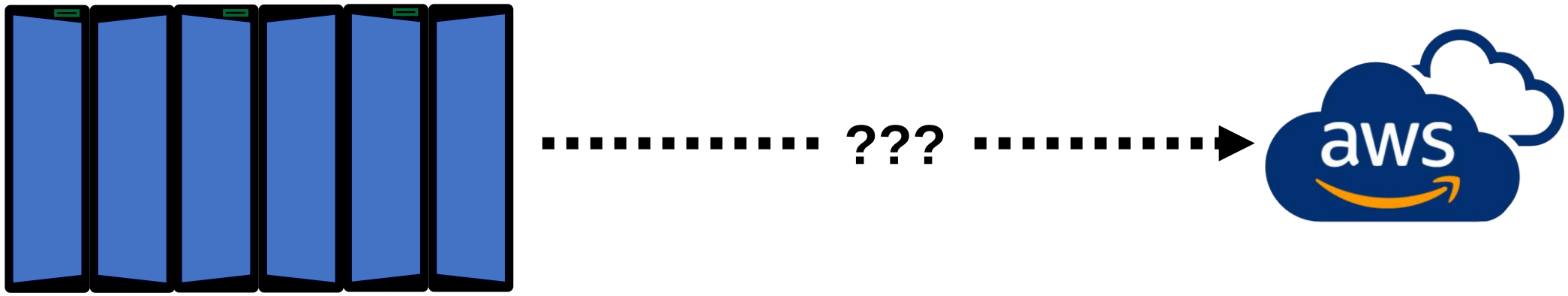


Source: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

AI Journey: The Missing Link

- We have supercomputer for training and cloud for operation

How to **quickly** deploy a trained model from supercomputer to operate on cloud?



AI Journey: Deploying AI Model

- AI model is only useful when deployed in production
  - **Concept Drift:** The model accuracy is at its best until you start using it

- The faster you can deploy, the earlier you can test and improve your model

AI Journey: Deploying AI Model

- Deploying model requires different skill set than developing it
  - Development (HPC cluster) and operation (cloud) are two completely different environment
- There are several approaches to deploy and maintain AI model in production
- Mastering skill is time-consuming

AI Journey: An On-Going Work

- Collaboration between ThaiSC, VISTEC, and AWS to building a tool for deploying AI model from LANTA supercomputer to AWS cloud

- **Next 2 Months**: Prototype deployment utility to deploying AI models from LANTA to SageMaker inference engine.

AI Journey: Meanwhile…

- Proposal submission is open for LANTA's Pilot Phase
  - Up to 4,800 GPU hours
- Submit your proposal at https://thaisc.io
- Or Contact us at thaisc@nstda.or.th

NSTDA Supercomputer Center
ThaiSC

Thank you

Contacts

Website: https://thaisc.io

Email: putt.sakdhnagool@nectec.or.th