Navigating the LLM Deployment Dilemma

On Rising GPU Costs and Innovative Solutions

codesphere

Everyone wants AI but how much will that cost

Corporations around the globe have understood that large language models have vast potential. The surge in demand for GPUs has driven prices for training and operating models to unprecedented heights, presenting a formidable challenge for organizations aiming to harness the power of LLMs. When deciding how to operate LLMs there are two predominant options, each laden with its own set of challenges.

- **Option 1: Providers Cloud APIs (i.e. OpenAI's APIs)**

  It's a convenient and initially cheap choice. However costs easily spiral out of control as traffic increases. This approach, while offering scalability, introduces concerns related to data privacy and vendor lock-in. On demand usage based pricing has advantages for testing but it is rarely a sensible choice for bigger projects.

- **Option 2: Self-Hosted LLMs (i.e. run open source models on own infrastructure)**

  Companies like Meta and others released open source versions of their models that can be used commercially. While offering more granular control over data and pricing, this avenue comes with its own set of challenges. Acquiring and maintaining dedicated GPU resources for self-hosting needs experts & time and without proper provisioning and deployment solutions costs are still inefficiently high. Since individual requests require high amounts of computing resources, overprovisioning is more costly than for other software.

Self-hosting is better - Despite complex configuration and operations concerns

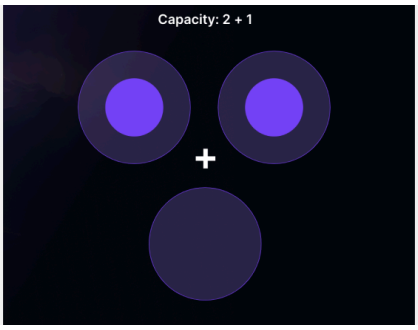
Configuring LLM infrastructure

ML operations is a relatively new & complex field and most companies do not have dedicated teams in place that can provision the needed infrastructure. Even if such teams exist they typically are an expensive bottleneck. This slows down LLM adoption and costs time & money.

How much computing resources an application needs fluctuates with the amount of users using the application at the same time. LLMs have very high computing consumption per user, making it more crucial than elsewhere that the available resources cover the load on the system. Typical cloud providers are not fast enough to provision additional resources in real time - therefore companies typically need to overprovision by a factor 2x. You pay for double the resources you need at any point in time to cope with fluctuating load. With a single GPU server costing between 1000 and 5000 Euro per month that's a lot of money and energy wasted. Luckily both the complexity issue and the operations inefficiencies can be solved.

Reactive Inference: 1-click GPU infrastructure that scales in real-time

Codesphere addresses these challenges with its groundbreaking solution, Reactive Inference (patent pending). Reactive Inference allows for the rapid activation of additional GPU & CPU resources within 10-20 milliseconds, enabling seamless scalability based on demand in real time without disrupting your users.



Key Features of Reactive Inference

1. **Instant Cold Start Activation:**
    - Rapid activation of turned off (GPU) resources, ensures that additional computational power is available almost instantly when needed. For low traffic applications (i.e. internal expert applications) a scale down to 0 can save up to 90% as the application only consumes compute when used.
2. **Scaling resources in real-time:**
    - The deployment setup can efficiently utilize additional GPU resources only when necessary, avoiding overprovisioning and optimizing cost-effectiveness. For applications with fluctuating traffic this can save up to 35% on infrastructure.
3. **Zero Configuration Infrastructure Setup:**
    - Codesphere's Reactive Inference can be set up by any developer with a few clicks. It doesn't require expensive experts for setup & maintenance. Deploying an LLM app follows the same streamlined best practice workflow like any other application that your developers are already familiar with.

On prem, (private) cloud or hybrid infrastructure

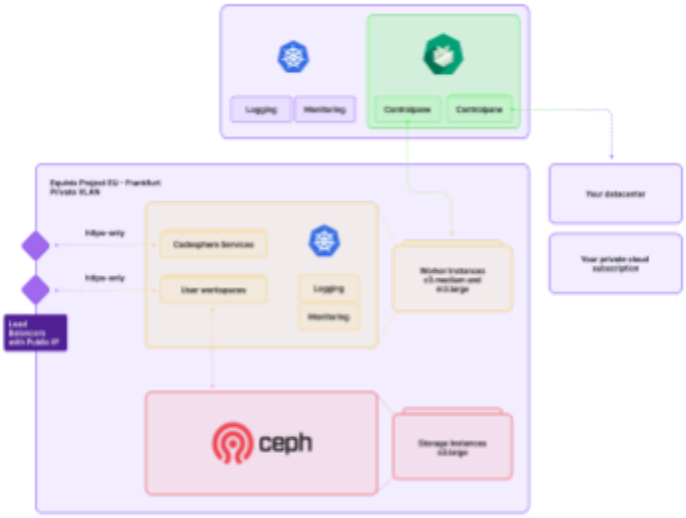
Codesphere's solution is versatile and can be deployed in various scenarios, depending on your data security and scalability requirements.

Cloud Deployment:

- Utilize Codesphere's Reactive Inference on popular cloud platforms, hyperscalers or private clouds for flexibility and scalability.
- Pay only what you need without overprovisioning for fluctuating traffic

On-Premises:

- Implement the solution within on-premises infrastructure, ensuring highest data security and compliance - ideal for industries with stricter requirements
- Increase the density of applications you can fit on any given size of infrastructure by more efficient balancing traffic between your applications



Codesphere's Multi Cloud Infrastructure

Hybrid Model:

- Utilize cheaper (per compute unit) bare metal on premise machines for the base load & flexibly add cloud based computing on demand to handle traffic spikes
- Combine cloud and on-premises deployment to achieve a balanced and tailored solution.

Technology that makes self hosting viable

Codesphere's Reactive Inference emerges as a game-changer in the deployment of large language models, offering not only unprecedented cost savings but also a flexible and scalable solution. It eliminates the complexity of setting up infrastructure for self hosted AI models. As the industry grapples with the challenges of LLM deployment costs, Codesphere's innovation stands out as a beacon of efficiency, addressing the current limitations and reshaping the landscape of large language model utilization.

Your contacts

Oliver Winkelmann

Head of Sales,
Karlsruhe

Timo Bräutigam

Partnerships,
Aachen

Simon Pfeiffer

Head of Product,
Bonn