Helix is a system designed to overcome the challenges associated with serving LLMs on heterogeneous GPU clusters. It does so by focusing on three core areas: model placement, request scheduling, and communication optimization.

**1. Model Placement**

Helix approaches model placement as a max-flow problem on a directed, weighted graph, where:

* **Nodes** represent GPUs with varying capabilities.
* **Edges** represent the communication links between these GPUs, weighted by their bandwidth and latency.

Helix uses Mixed Integer Linear Programming (MILP) to solve this problem, determining the optimal distribution of the model's layers across the GPUs. This ensures that each GPU is utilized according to its strengths, while also minimizing the communication overhead between them.

**2. Request Scheduling**

In Helix, request scheduling is handled dynamically, with each inference request being processed through a custom pipeline. This differs from traditional fixed pipelines, which can lead to inefficiencies in heterogeneous environments. Helix's per-request pipeline approach allows it to adapt to the current state of the cluster, rerouting requests to avoid bottlenecks and underutilized GPUs.

**3. Communication Optimization**

Given the distributed nature of the GPU clusters, Helix places significant emphasis on optimizing communication between GPUs. By taking into account the network conditions, Helix ensures that data transfers are minimized, and when necessary, are carried out in the most efficient manner possible. This reduces the overall latency, ensuring quicker response times even when multiple GPUs are involved.

**4. Evaluation and Results**

Helix was evaluated on various heterogeneous GPU cluster setups, demonstrating its effectiveness across different configurations. The system showed a remarkable improvement in both throughput and latency:

* **Throughput:** Helix achieved up to a 2.7× increase in the number of inference requests processed per unit time, compared to existing methods.
* **Latency:** The system reduced prompting latency by up to 2.8× and decoding latency by 1.3×, ensuring faster and more responsive LLM serving.

These results highlight Helix's ability to adapt to the complexities of heterogeneous GPU environments, making it a highly efficient solution for large-scale LLM deployment.