Implementing GPU Slicing on EKS

To enable GPU slicing on EKS clusters, follow these steps:

1. **Select Appropriate GPU Instances:**
    - **For MIG: Utilize NVIDIA A100 GPUs, such as the p4d or p4de instance types, which support MIG.**
    - **For Time-Slicing: Choose compatible NVIDIA GPUs like the V100 or T4, which support time-slicing.**
2. **Deploy the NVIDIA GPU Operator:**
    - **The NVIDIA GPU Operator automates the management of GPU drivers and Kubernetes plugins. It facilitates the deployment and configuration of GPU slicing features.**
    - **Installation Steps:**
        - **Create a Dedicated Node Group: Set up an EKS node group with GPU-enabled instances. Ensure the nodes run a supported operating system, such as Ubuntu, as the GPU Operator may not support Amazon Linux 2.**
        - **Install NVIDIA GPU Operator: Deploy the operator using Helm or Kubernetes manifests. This will handle the installation of necessary components, including GPU drivers and the device plugin.**
3. **Configure GPU Slicing:**
    - **For MIG:**
        - **After deploying the GPU Operator, configure MIG by specifying the desired GPU partitions. This setup allows multiple pods to utilize separate GPU instances on the same physical GPU.**
    - **For Time-Slicing:**
        - **Ensure the NVIDIA device plugin is configured to enable time-slicing, allowing multiple pods to share the GPU resources efficiently.**

Integrating GPU Slicing with Karpenter Autoscaler

Karpenter is a flexible, high-performance Kubernetes cluster autoscaler that can help manage dynamic workloads. To leverage GPU slicing with Karpenter:

1. **Define Node Templates:**
    - **Create Karpenter node templates specifying GPU-enabled instance types and the necessary labels or taints. This ensures that GPU workloads are scheduled appropriately.**
2. **Configure Provisioners:**

- **Set up Karpenter provisioners with resource requirements that match your GPU-sliced nodes. This configuration allows Karpenter to scale nodes based on the GPU workloads' demands.**

3. **Pod Specification:**
    - **In your pod definitions, request GPU resources as needed. For MIG, specify the GPU instance type; for time-slicing, request fractional GPU resources if supported.**