Google Discussions

Exam Professional Machine Learning Engineer All Questions

View all questions & answers for the Professional Machine Learning Engineer exam

Go to Exam

EXAM PROFESSIONAL MACHINE LEARNING ENGINEER TOPIC 1 QUESTION 209 DISCUSSI...

Actual exam question from Google's Professional Machine Learning Engineer

Question #: 209

Topic #: 1

[All Professional Machine Learning Engineer Questions]

You are training a custom language model for your company using a large dataset. You plan to use the Reduction Server strategy on Vertex AI. You need to configure the worker pools of the distributed training job. What should you do?

A. Configure the machines of the first two worker pools to have GPUs, and to use a container image where your training code runs. Configure the third worker pool to have GPUs, and use the reductionserver container image.

B. Configure the machines of the first two worker pools to have GPUs and to use a container image where your training code runs. Configure the third worker pool to use the reductionserver container image without accelerators, and choose a machine type that prioritizes bandwidth.

C. Configure the machines of the first two worker pools to have TPUs and to use a container image where your training code runs. Configure the third worker pool without accelerators, and use the reductionserver container image without accelerators, and choose a machine type that prioritizes bandwidth.

D. Configure the machines of the first two pools to have TPUs, and to use a container image where your training code runs. Configure the third pool to have TPUs, and use the reductionserver container image.

Show Suggested Answer

by pikachu007 at *Jan. 13, 2024, 5:32 a.m.*

Comments

Type your comment...

Submit

⊟ **wences** 1 month ago

Selected Answer: B

The real reason for answer B is the custom model, which means it was not suited well for TPU

👍 ↩ ⚑ upvoted 1 times

⊟ **fitri001** 6 months ago

Selected Answer: B

GPUs for Training: Configure the first two worker pools with GPUs to leverage the hardware acceleration capabilities for your custom language model training code.
Reduction Server without GPUs: The third worker pool should use the reductionserver container image. This image is pre-configured for Reduction Server functionality and doesn't require GPUs.
High-Bandwidth CPU: Choose a machine type with high bandwidth for the third pool since Reduction Server focuses on communication and gradient reduction.

👍 ↩ ⚑ upvoted 2 times

⊟ **fitri001** 6 months ago

A. GPUs for Reduction Server: Reduction Server itself doesn't require or benefit from GPUs. It focuses on communication and reduction of gradients. It's better to use a CPU-based machine type for the third pool.
C. TPUs instead of GPUs: While TPUs can be used for training some language models, Reduction Server specifically works with GPUs using the NCCL library. Configure your first two pools with GPUs for your training code.
D. TPUs in Reduction Server pool: Similar to option A, Reduction Server doesn't benefit from TPUs. It's best to use a CPU with high bandwidth for the third pool.

👍 ↩ ⚑ upvoted 1 times

⊟ **pinimichele01** 6 months, 2 weeks ago

Selected Answer: B

https://cloud.google.com/blog/topics/developers-practitioners/optimize-training-performance-reduction-server-vertex-ai

In this article, we introduce Reduction Server, a new Vertex AI feature that optimizes bandwidth and latency of multi-node distributed training on NVIDIA GPUs for synchronous data parallel algorithms.

👍 ↩ ⚑ upvoted 1 times

⊟ **shadz10** 9 months, 1 week ago

Selected Answer: B

TPUs are not supported for reductionserver so B

👍 ↩ ⚑ upvoted 3 times

⊟ **winston9** 9 months, 1 week ago

Selected Answer: B

bandwidth is important for the reduction server

👍 ↩ ⚑ upvoted 2 times

⊟ **pikachu007** 9 months, 1 week ago

Selected Answer: B

Worker Pools 1 and 2:
These pools are responsible for the actual model training tasks.
They require GPUs (or TPUs, if applicable to your model) to accelerate model computations.
They run the container image containing your training code.
Worker Pool 3:
This pool is dedicated to the reduction server.
It doesn't require accelerators (GPUs or TPUs) for gradient aggregation.
Prioritize machines with high network bandwidth to optimize gradient exchange.
Use the specific reductionserver

👍 ↩ ⚑ upvoted 2 times

Social Media

Facebook , Twitter
YouTube , Reddit
Pinterest

We are the biggest and most updated IT certification exam material website.

Using our own resources, we strive to strengthen the IT professionals community for free.

ExamTopics doesn't offer Real Microsoft Exam Questions. ExamTopics doesn't offer Real Amazon Exam Questions. ExamTopics Materials do not contain actual questions and answers from Cisco's Certification Exams.

CFA Institute does not endorse, promote or warrant the accuracy or quality of ExamTopics. CFA® and Chartered Financial Analyst® are registered trademarks owned by CFA Institute.