**Gpu Matrix Required for Deep learning:**

|  |  |  |
| --- | --- | --- |
| **GPUs** | :Tesla V100 (16 GB or 32 GB) | Widely used for deep learning due to high compute power and memory. |

|  |  |  |
| --- | --- | --- |
| **GPU Memory (GB)** | :16 GB – 32 GB per GPU | Allows larger batch sizes and models in memory for faster training. |

|  |  |  |
| --- | --- | --- |
| **vCPUs** | :8 – 16 vCPUs per GPU | Supports efficient scheduling; TensorFlow offloads tasks to CPUs. |

|  |  |  |
| --- | --- | --- |
| **Memory (RAM):** | 64 GB – 128 GB | Ensures smooth data pre-processing and model training. |

|  |  |  |
| --- | --- | --- |
| **Network Bandwidth** | :10 Gbps – 25 Gbps | Required for transferring large datasets, especially in distributed setups. |

|  |  |  |
| --- | --- | --- |
| **EBS Bandwidth** | :1 Gbps – 2 Gbps (burstable up to 4 Gbps) | Affects read and write speed from cloud storage (EBS). |

**P2 Instance Details**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | GPUs | vCPUs | RAM (GiB) | Network Bandwidth | Price/Hour\* | RI Price / Hour\*\* |
| p2.xlarge | 1 | 4 | 61 | High | $0.900 | $0.425 |
| p2.8xlarge | 8 | 32 | 488 | 10 Gbps | $7.200 | $3.400 |
| p2.16xlarge | 16 | 64 | 732 | 20 Gbps | $14.400 | $6.800 |

*\* Pricing for US East (N. Virginia) and US West (Oregon)*

**P3 Instance Details**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Instance Size | GPUs - Tesla V100 | GPU Peer to Peer (reduce letency) | GPU Memory (GB) | vCPUs | Memory (GB) | Network Bandwidth | EBS Bandwidth | On-Demand Price/hr\* | 1-yr Reserved Instance Effective Hourly\* | 3-yr Reserved Instance Effective Hourly\* |
| p3.2xlarge | 1 | N/A | 16 | 8 | 61 | Up to 10 Gbps | 1.5 Gbps | $3.06 | $1.99 | $1.05 |
| p3.8xlarge | 4 | NVLink | 64 | 32 | 244 | 10 Gbps | 7 Gbps | $12.24 | $7.96 | $4.19 |
| p3.16xlarge | 8 | NVLink | 128 | 64 | 488 | 25 Gbps | 14 Gbps | $24.48 | $15.91 | $8.39 |
| p3dn.24xlarge | 8 | NVLink | 256 | 96 | 768 | 100 Gbps | 19 Gbps | $31.218 | $18.30 | $9.64 |

\* - Prices shown are for Linux/Unix in the US East (Northern Virginia) AWS Region and rounded to the nearest cent.

**P4 Instance Details**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Instance Size** | **vCPUs** | **Instance Memory (GiB)** | **GPU – A100** | **GPU memory** | **Network Bandwidth (Gbps)** | **GPUDirect RDMA** | **GPU Peer to Peer** | **Instance Storage (GB)** | **EBS Bandwidth (Gbps)** | **On-demand Price/hr** | **1-yr Reserved Instance Effective Hourly \*** | **3-yr Reserved Instance Effective Hourly \*** |
| p4d.24xlarge | 96 | 1152 | 8 | 320 GB HBM2 | 400 ENA and EFA | Yes | 600 GB/s NVSwitch | 8 x 1000 NVMe SSD | 19 | $32.77 | $19.22 | $11.57 |
| p4de.24xlarge (preview) | 96 | 1152 | 8 | 640 GB HBM2e | 400 ENA and EFA | Yes | 600 GB/s NVSwitch | 8 x 1000 NVMe SSD | 19 | $40.96 | $24.01 | $14.46 |

[Amazon EC2 P3 – Ideal for Machine Learning and HPC - AWS](https://aws.amazon.com/ec2/instance-types/p3/)

Customers can purchase P3 instances as On-Demand Instances, Reserved Instances, Spot Instances, and Dedicated Hosts.

**Open source Model**

[paraphrase-multilingual](https://ollama.com/library/paraphrase-multilingual)

<https://ollama.com/library/qwen>

<https://ollama.com/library/nomic-embed-text>

<https://huggingface.co/blog/manu/colpali>

 :