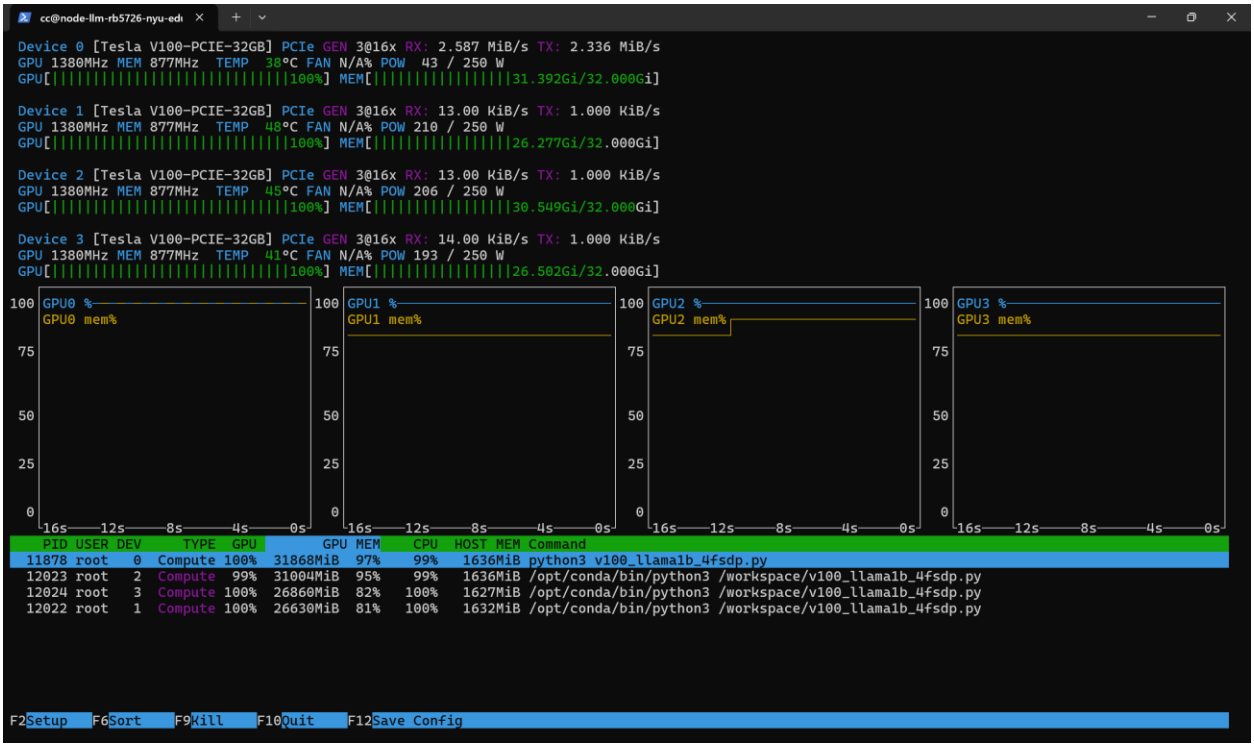
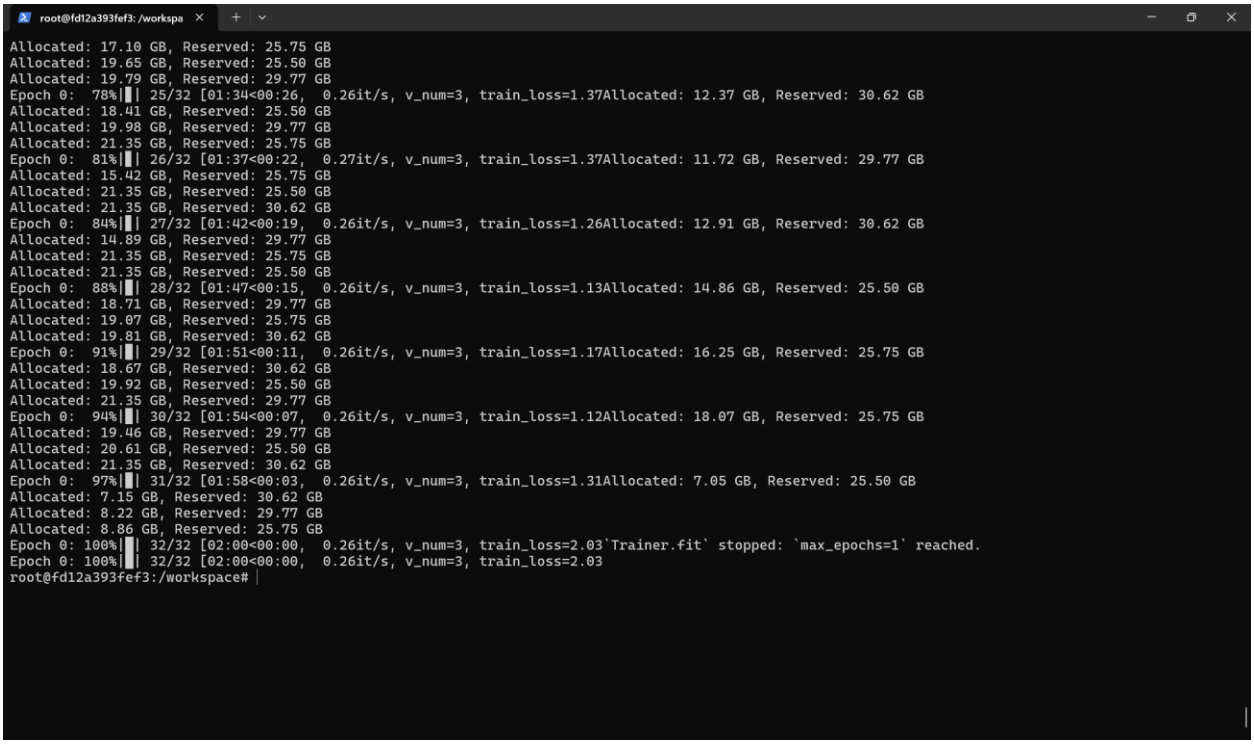
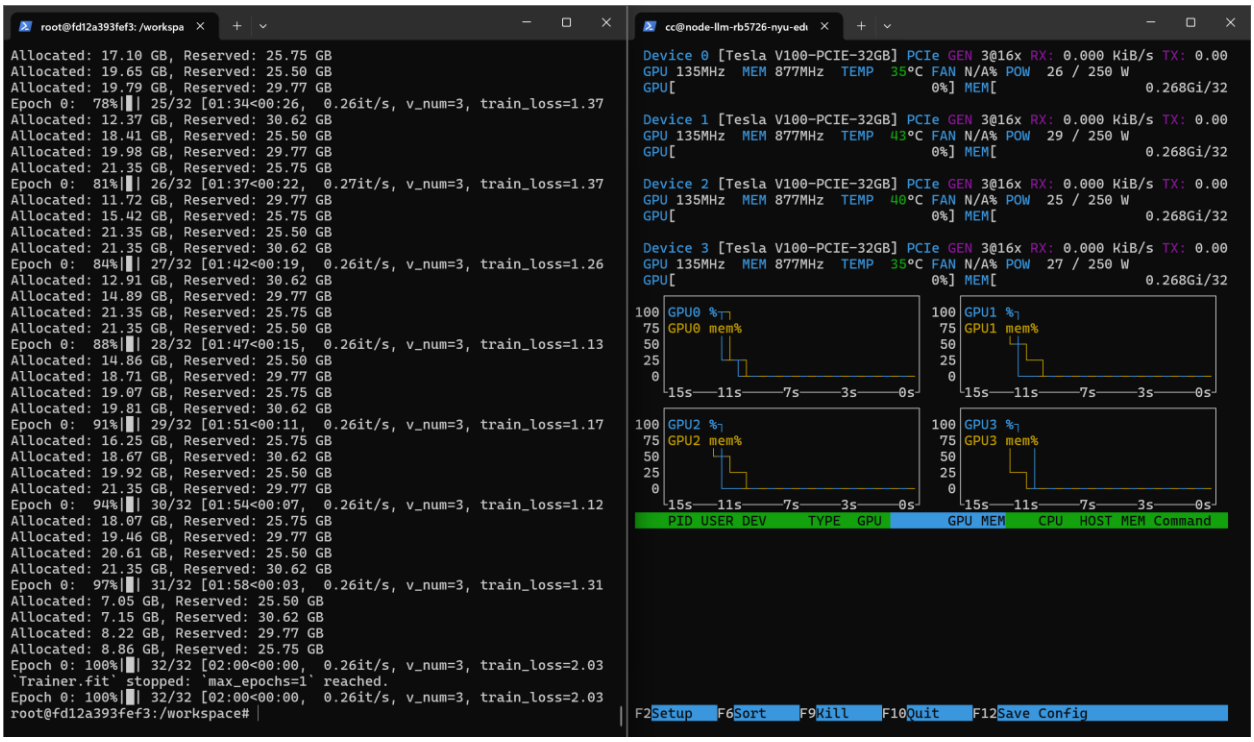
TinyLlama 1.1B model on 4x V100 32GB with FSDP



nvtop

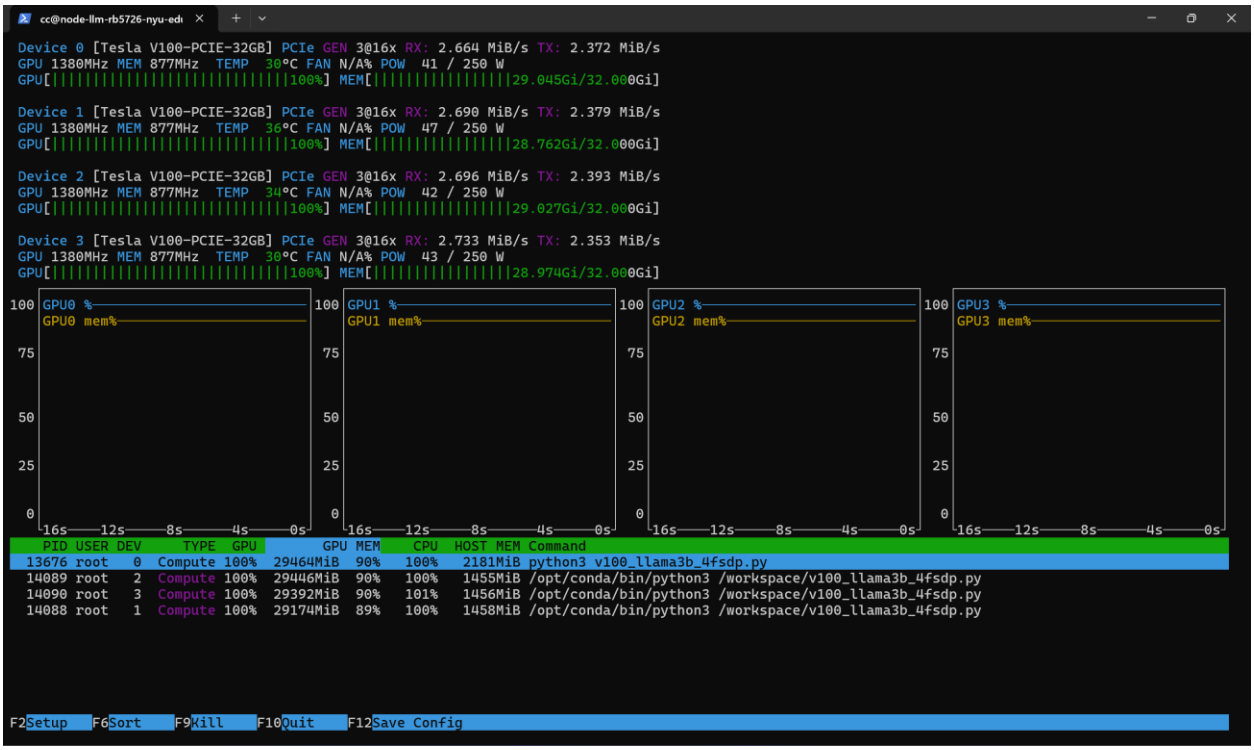


Training time

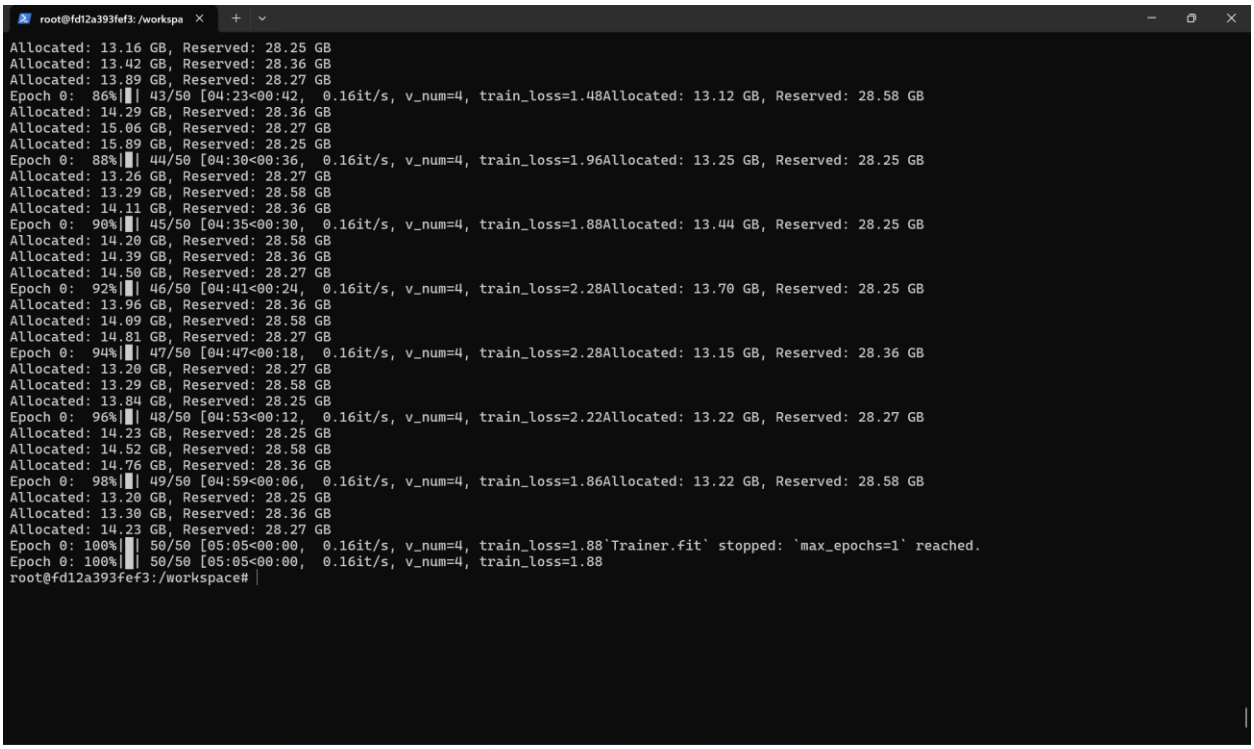


After completion of TinyLlama 1.1B model on 4x V100 32GB with FSDP

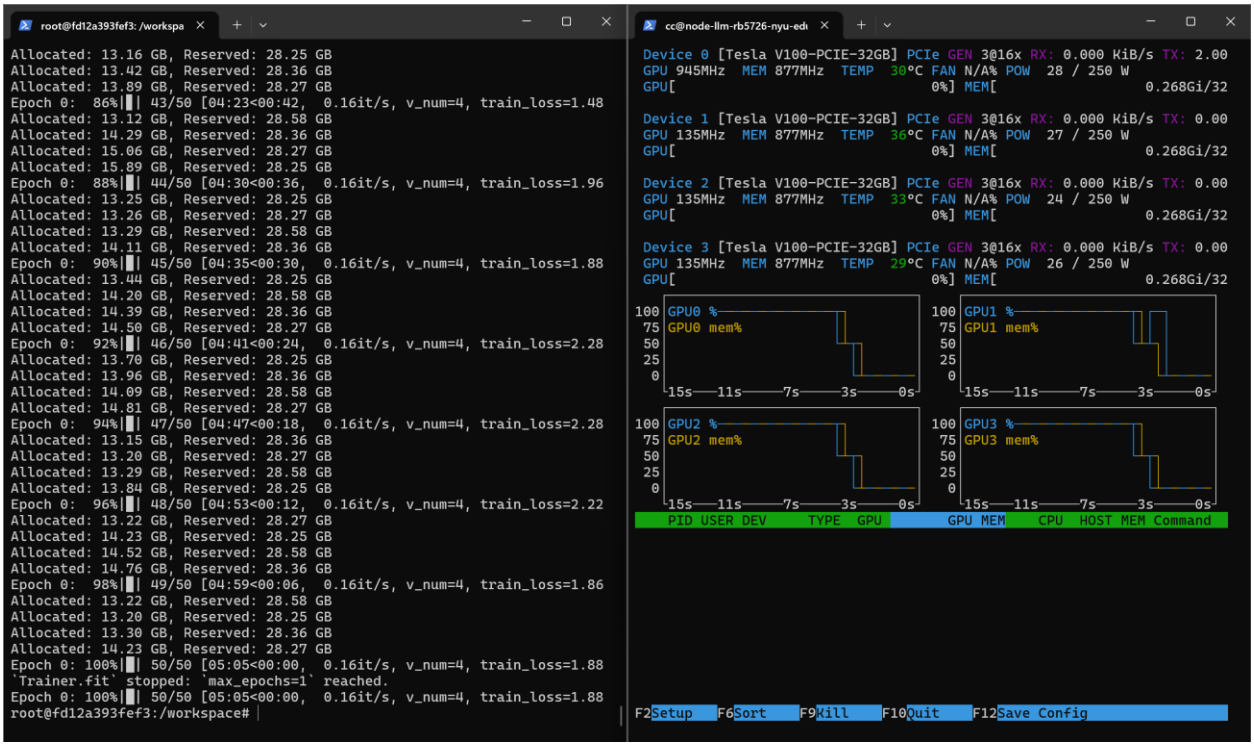
OpenLLaMA 3B model on 4x V100 32GB with FSDP



nvtop



Training time

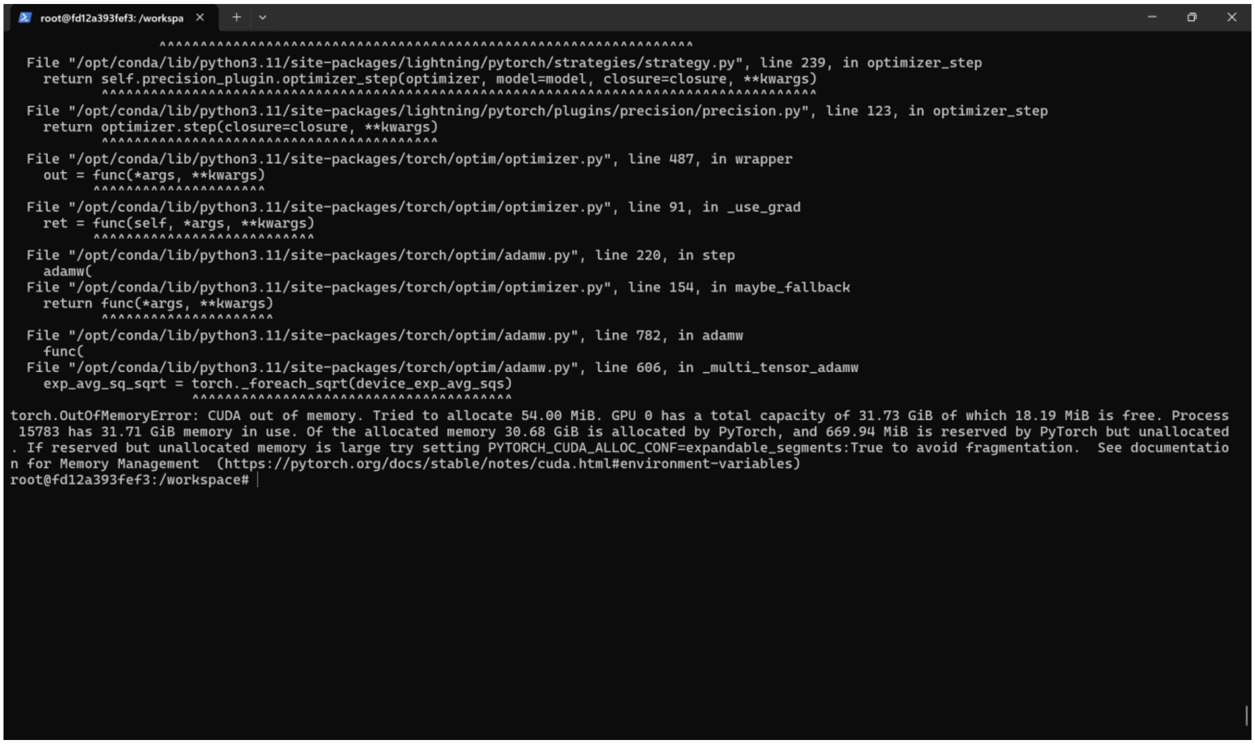


After completion of OpenLLaMA 3B model on 4x V100 32GB with FSDP

We cannot train this model on a single V100 without running out of memory - try

 # run inside pytorch container

python3 v100_llama3b_1device.py



```
root@fd12a393fef3: /workspa
                   ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/opt/conda/lib/python3.11/site-packages/lightning/pytorch/strategies/strategy.py", line 239, in optimizer_step
    return self.precision_plugin.optimizer_step(optimizer, model=model, closure=closure, **kwargs)
           ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/opt/conda/lib/python3.11/site-packages/lightning/pytorch/plugins/precision/precision.py", line 123, in optimizer_step
    return optimizer.step(closure=closure, **kwargs)
           ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/opt/conda/lib/python3.11/site-packages/torch/optim/optimizer.py", line 487, in wrapper
    out = func(*args, **kwargs)
          ^^^^^^^^^^^^^^^^^^^^^
  File "/opt/conda/lib/python3.11/site-packages/torch/optim/optimizer.py", line 91, in _use_grad
    ret = func(self, *args, **kwargs)
          ^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/opt/conda/lib/python3.11/site-packages/torch/optim/adamw.py", line 220, in step
    adamw(
  File "/opt/conda/lib/python3.11/site-packages/torch/optim/optimizer.py", line 154, in maybe_fallback
    return func(*args, **kwargs)
           ^^^^^^^^^^^^^^^^^^^^^
  File "/opt/conda/lib/python3.11/site-packages/torch/optim/adamw.py", line 782, in adamw
    func(
  File "/opt/conda/lib/python3.11/site-packages/torch/optim/adamw.py", line 606, in _multi_tensor_adamw
    exp_avg_sq_sqrt = torch._foreach_sqrt(device_exp_avg_sqs)
                      ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
torch.OutOfMemoryError: CUDA out of memory. Tried to allocate 54.00 MiB. GPU 0 has a total capacity of 31.73 GiB of which 18.19 MiB is free. Process
 15783 has 31.71 GiB memory in use. Of the allocated memory 30.68 GiB is allocated by PyTorch, and 669.94 MiB is reserved by PyTorch but unallocated
. If reserved but unallocated memory is large try setting PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True to avoid fragmentation.  See documentatio
n for Memory Management  (https://pytorch.org/docs/stable/notes/cuda.html#environment-variables)
root@fd12a393fef3:/workspace#
```