

Machine Learning Engineering

by Stas Bekman



Machine Learning Engineering Open Book

This is a PDF version of [Machine Learning Engineering Open Book by Stas Bekman](#).

As this book is an early work in progress that gets updated frequently, if you downloaded it as a pdf file, chances are that it's already outdated - make sure to check the latest version at <https://github.com/stas00/ml-engineering>.

This PDF was generated on 2025-01-22.

Machine Learning Engineering Open Book

This is an open collection of methodologies, tools and step by step instructions to help with successful training and fine-tuning of large language models and multi-modal models and their inference.

This is a technical material suitable for LLM/VLM training engineers and operators. That is the content here contains lots of scripts and copy-n-paste commands to enable you to quickly address your needs.

This repo is an ongoing brain dump of my experiences training Large Language Models (LLM) (and VLMs); a lot of the know-how I acquired while training the open-source [BLOOM-176B](#) model in 2022 and [IDEFICS-80B](#) multi-modal model in 2023, and RAG models at [Contextual.AI](#) in 2024.

I've been compiling this information mostly for myself so that I could quickly find solutions I have already researched in the past and which have worked, but as usual I'm happy to share these notes with the wider ML community.

Table of Contents

Part 1. Insights

1. [The AI Battlefield Engineering](#) - what you need to know in order to succeed.
2. [How to Choose a Cloud Provider](#) - these questions will empower you to have a successful compute cloud experience.

Part 2. Hardware

1. [Compute](#) - accelerators, CPUs, CPU memory.
2. [Storage](#) - local, distributed and shared file systems.
3. [Network](#) - intra- and inter-node networking.

Part 3. Orchestration

1. [Orchestration Systems](#) - managing containers and resources
2. [SLURM](#) - Simple Linux Utility for Resource Management

Part 4. Training

1. [Training](#) - model training-related guides

Part 5. Inference

1. [Inference](#) - model inference insights

Part 6. Development

1. [Debugging and Troubleshooting](#) - how to debug easy and difficult issues
2. [And more debugging](#)
3. [Testing](#) - numerous tips and tools to make test writing enjoyable

Part 7. Miscellaneous

1. [Resources](#) - LLM/VLM chronicles

Updates

I announce any significant updates on my twitter channel <https://twitter.com/StasBekman>.

PDF version

Download the [PDF](#) version of the book.

I will try to rebuild it once a week or so, but if you want the latest, the instructions for building are [here](#).

Thanks to HuggingFace for giving me permission to host my book's PDF at the [HF hub](#).

Discussions

If you want to discuss something related to ML engineering this repo has the [community discussions](#) available - so please don't hesitate to share your experience or start a new discussion about something you're passionate about.

Key comparison tables

High end accelerators:

- [Theoretical accelerator TFLOPS](#)
- [Accelerator memory size and speed](#)

Networks:

- [Theoretical inter-node speed](#)
- [Theoretical intra-node speed](#)

Shortcuts

Things that you are likely to need to find quickly and often.

Tools:

- [all_reduce_bench.py](#) - a much easier way to benchmark network throughput than nccl-tests.
- [torch-distributed-gpu-test.py](#) - a tool to quickly test your inter-node connectivity
- [mamf-finder.py](#) - what is the actual TFLOPS measurement you can get from your accelerator.

Guides:

- [debugging pytorch applications](#) - quick copy-n-paste solutions to resolve hanging or breaking pytorch applications
- [slurm for users](#) - a slurm cheatsheet and tricks
- [make tiny models/datasets/tokenizers](#)
- [LLM/VLM chronicles collection](#)

Gratitude

None of this would have been possible without me being entrusted with doing the specific LLM/VLM trainings I have learned the initial know-how from. This is a privilege that only a few enjoy due to the prohibitively expensive cost of renting huge ML compute clusters. So hopefully the rest of the ML community will vicariously learn from these notes.

Special thanks go to [Thom Wolf](#) who proposed that I lead the BLOOM-176B training back when I didn't know anything about large scale training. This was the project that catapulted me into the intense learning process. And, of course, HuggingFace for giving me the opportunity to work full time on BLOOM-176B and later on IDEFICS-80B trainings.

Recently, I continued expanding my knowledge and experience while training models and building scalable training/inference systems at [Contextual.AI](#) and I'm grateful to that opportunity to Aman and Douwe.

I'd also like to say thanks to the numerous [contributors](#) who have been making this text awesome and error-free.

Contributing

If you found a bug, typo or would like to propose an improvement please don't hesitate to open an [Issue](#) or contribute a PR.

License

The content of this site is distributed under [Attribution-ShareAlike 4.0 International](#).

Citation

```
@misc{bekman2024mlengineering,
  author = {Bekman, Stas},
  title = {Machine Learning Engineering Open Book},
  year = {2023-2024},
  publisher = {Stasosphere Online Inc.},
  journal = {GitHub repository},
  url = {https://github.com/stas00/ml-engineering}
}
```

My repositories map

- ✓ Machine Learning: [ML Engineering Open Book](#) | [ML ways](#) | [Porting](#)
- ✓ Guides: [The Art of Debugging](#)
- ✓ Applications: [ipyexperiments](#)
- ✓ Tools and Cheatsheets: [bash](#) | [conda](#) | [git](#) | [jupyter-notebook](#) | [make](#) | [python](#) | [tensorboard](#) | [unix](#)

The AI Battlefield Engineering - What You Need To Know

This chapter is one person's opinionated overview of the ML/AI Engineering reality, which may or may not be another person's reality. The intention is to help you start asking the right questions and get your ML Engineering needs met.

Basics

What's important in the AI race?

Training:

1. How fast one can train a better model (first to market advantage)
2. How much \$\$ was spent (do we still have money left to pay salaries to talent after training?)

Inference:

1. Fast latency (users are used to msec response times and will leave if the response takes seconds)
2. Fast throughput (how many concurrent queries can be processed)
3. How much \$\$ is being spent per user (can we rent more GPUs to acquire more users and/or improve (1) and (2)?)

What are the needs of LLM training?

1. Fast compute massively dominated by matrix multiplications
2. Fast enough memory, IO, network and CPU to feed the compute

Corollary: If when you buy or rent hardware you invest in the fastest accelerators, but cheap out on any of the other components you wasted \$\$ and you might not win the race as it'll take longer to train.

What are the workhorses of ML?

- An accelerator or a processing unit is what does most of the work.
- Since ML does a lot of parallel processing ([SIMD](#)) GPUs were used at the beginning, but now you additionally have TPUs, IPUs, FPGAs, HPUs, QPUs, RDUs, etc. Recent CPUs are becoming used as accelerators as well, especially for inference.

[More details.](#)

AI driving entities

- AI companies - train models/build products around self-trained or trained-by-others' models, in-house research.
- Academia - does massive research and write papers. Lots of new ideas are generated.
- AI enthusiasts - lots of good will available, some pull resources/talents together to train open access models, with donated compute by HPCs and an occasional cloud, or a university cluster.
- Entrepreneurs - lots of low hanging fruit to pick - creative reselling of services, making ML-driven apps, and using various ingenious combinations of available resources to create amazing outcomes.

Information sharing

- It's very surprising that almost everybody involved in the domain of AI shares a lot of the discoveries with the community.
- Surely, companies don't disclose all of their IP, but a lot of it does get shared in the form of knowledge or model

- weights
- Companies that publish a lot of IP and models tend to attract higher quality talent.
 - Twitter seems to be the central platform where one must be to follow what's going on

The AI bubble

- The [Dot-com bubble](#) occurred during 1995-2000. And a very similar situation is happening right now in the AI space.
- There is a lot of money available to create new startups or boost the existing companies. It's relatively easy to raise millions of dollars.
- As we are in the wild-wild-west stage of the AI industry it's very difficult to predict the future, and so pretty much anything goes as far as startup ideas go, as long as it sounds reasonable.
- What distinguishes the AI bubble from the Dot-com bubble, is that one didn't actually need much money to operate a Dot-com company - most of the raised money went to marketing and some to staff, barely any to compute. AI companies need millions of dollars because training LLMs requires an insane amount of compute, and that compute is very expensive. e.g. 1x NVIDIA H100 costs ~\$30k and a company may need 512 of those, which is \$15M (not counting the other hardware components and related costs)!

ML Engineer's heaven and hell

This is my personal LLM/VLM trainings-based heaven and hell. YMMV.

ML Engineer's heaven

1. A well built HPC, or a full service cloud based cluster, where someone diligently and timely takes care of the hardware and the systems.
I just need to bring my training software and do the training, which is already an insanely complicated job requiring special skills.
2. Lots of nodes available for exclusive unlimited use
3. Fast inter-node connectivity that doesn't bottleneck the accelerators and which isn't shared with other users
4. Huge local super-fast NVME based shared filesystem that can fit datasets and checkpoints
5. Barebones Linux w/ SLURM and minimal software to be able to launch training jobs
6. sudoer access to ease the work with a team of people

ML Engineer's hell

1. A cloud or in-house cluster, where you have to do everything - sysadmining, replacing hardware, dealing with outages, etc. And to do the training on top of that.
2. A smallish slow shared filesystem (NFS?), with cloud to draw data from and checkpoint to
3. Slow inter-node leading to low accelerator utilization
4. Inter-node shared with other users which make the network erratic and unpredictable
5. Super-complicated cloud console with gazillion of screens and steps to set even simple things up
6. Not being able to swap out failing hardware fast
7. Needing to timeshare the nodes - with wait times between training jobs
8. Having other concurrent users who might use up the whole disk, leading to trainings crashing

9. Not being able to kill jobs others on the team started and went to sleep

Getting compute

There are 3 main choices to where one gets compute:

- Rent on the cloud
- Get a timeshare on an HPC
- Buy it

Renting on the cloud

This is currently the prevalent way of getting compute.

Pros:

- Easy to expand or contract the size of the cluster
- Easy to upgrade from the old hardware generation to the new one in a few years
- Cluster management could be easily outsourced

Cons:

- Expensive, unless you negotiate a long term (1-3 year) contract for hundreds of accelerators
- You will be tempted to buy many tools and services that you may or may not need
- You always get charged whether you use your cluster fully or not

Using HPC

There aren't that many HPCs out there and so the amount of available resources is limited.

Pros:

- Managed for you - all you need is your software to do the training and a bit of [SLURM](#) know-how to launch jobs
- Often sponsored by the local government/university - probably could get the job done for less \$\$ or even free (e.g. we trained [BLOOM-176B](#) for free on [JeanZay HPC!](#))

Cons:

- needing to time share compute with other teams == short job times with possible long wait times in between - could be difficult to finish training quickly
- The inter-node network is likely to be unstable as it'll be used by other teams
- Have to abide by the HPC's rules (e.g. no sudo access and various other rules to follow)
- In a way the HPC cluster will be what it'll be - you can't make the network faster and often even getting some software installed can be tricky.

Buying hardware

It's mainly universities that buy and build their own clusters, and some big companies do that too.

Pros:

- If you can deploy the hardware 24/7 for more than a few years the total cost will be cheaper than renting
- Easy to provide fast local storage - a good NVME raid would be much cheaper and faster than online storage

Cons:

- You're stuck with the outdated hardware just a few years after it was purchased - might be able to resell
- Must buy more than needed - Hardware tends to break, especially when it's used 24/7, RMA could take weeks
- Have to hire talent to manage the in-house solution
- Have to figure out cooling, electric costs, insurance, etc.

Managing compute

- Unless you use a fully managed HPC compute you absolutely need to hire a sysadmin. It may feel that your ML engineers can swing that between their training jobs, but they will be losing a lot of time to managing disk space, dealing with problematic nodes, asking users to behave, etc.

The needs of technology

Can you feed the furnace fast enough?

Imagine a steam locomotive - the engine is great, but if the [fireman](#) isn't fast enough to shovel the coal in, the train won't move fast.



[source](#)

This is the current state of ML hardware: The bottleneck is in moving bits and not the compute.

- Accelerators get ~2x faster every 2 years ([Moore's law](#))
- Network and memory are not! Already now both are compute bottlenecks
- IO can be another bottleneck if your DataLoader has to pull data from the cloud
- CPU is fine as long as it has enough cpu-cores for DataLoader workers, and main processes

Corollary: research the whole machine and not just its engine.

a crazy idea: the older GPUs might do fine if you can actually feed them as fast as they can compute. And if you can get 3x of them at the same cost as the next generation GPU you might finish training sooner and a lower cost.

TFLOPS

- Once you choose the architecture and the size of the model and how many tokens you want to train the model for you immediately know how much compute will be required to accomplish this goal. Specifically you can now calculate [how many floating point operations will be needed](#).
- All that is missing is comparing different compute providers to how many floating point operations their hardware can compute per secs (TFLOPS) and their cost per unit and now you can tell the total approximate cost of the training.
 - Calculate the time needed to train given the TFLOPS of the considered solution: `total_tflops_required / tflops_of_this_compute_unit = time_in_seconds` Let's say it came to be 604800 secs or 7 days.
 - Look at the cost of using this compute solution for 7 days and now you know the total \$\$ to train this model.
 - Look at other proposals and calculate the same - chose the best option.
- As mentioned earlier, time is of a huge importance, so you might still choose a more expensive solution if finishing the training sooner is important because you want to be first to market.

Unfortunately, this math is only partially correct because the advertised peak TFLOPS are typically unachievable. The MFU section delves into it.

Model FLOPS Utilization (MFU)

As mentioned in the previous section, some (most?) vendors publish unrealistic peak performance TFLOPS - they aren't possible to achieve.

Model FLOPS Utilization (MFU) is the metric that tells us how well the accelerator is utilized. Here is how it is calculated:

- Measure the actual TFLOPS by calculating how many floating point operations a single training iteration takes and dividing that number by the number of seconds this iteration took.
- Divide the actual TFLOPS by advertised TFLOPS to get the MFU

Example: Let's say you're training in BFLOAT16 precision:

- If a single iteration requires 624 Tera floating point operations and it took 4 secs to run then we know that we get: $624/4=156$ actual TFLOPS
- now BF16@A100 is [advertised as 312TFLOPS](#) so $156/312=0.5$ gives us 50% MFU.

Practically:

- with NVIDIA GPUs if you're above 50% MFU on a multi-node setup with a large model you're already doing fantastic
- recent advancements in more efficient scalability solutions keep on increasing MFU
- slow networks and inefficient frameworks or untuned configuration lower MFU

Therefore once you know the MFU you can now adjust the cost estimate from the previous section. In the example there we said it'll take 7 days to train, but if MFU is 50%, it means it'll take 14 days to train.

Moving bits

Why can't the advertised TFLOPS be achieved? It's because it takes time to move data between accelerator memory and compute and additionally it takes even more time to move data from disk and other GPUs to the accelerator's memory.

- There is not much that can be done about the accelerator memory since its bandwidth is what it is - one can only write more efficient software to make data move faster to/from the accelerator - hint: fused and custom written kernels (like [torch.compile](#) and [flash attention](#))
- If you only have a single GPU and the model fits its memory, you don't need to worry about the network -

accelerator memory is the only bottleneck. But if you have [to shard the model across multiple GPUs](#) network becomes the bottleneck.

- Intra-node Network - is very fast, but difficult to take advantage of for large models - [Tensor parallelism](#) and [sequence parallelism](#) address part of this problem. ([more](#)).
- Inter-node Network - typically is too slow on most server setups - thus this is the key component to research! Efficient frameworks succeed to partially hide the comms overhead by overlapping compute and comms. But if comms take longer than compute, the comms are still the bottleneck. [more](#).
- Storage IO is important primarily for feeding the DataLoader workers and saving the checkpoints. [more](#).
 1. Typically with enough DL workers the DataLoader adds very little overhead.
 2. While checkpoints are being saved the accelerators idle unless some async saving solution is used, so fast IO is crucial here

Key hardware components

Accelerators

As of this writing here are the most common accelerators that can be used for training, finetuning and inference ML models:

Widely available:

- NVIDIA H200s are gradually replacing A100s and H100s. H200s have more of and a more efficient HBM and thus make them more cost-effective than H100s.

Available, but locks you in:

- Google TPUs - fast! but the cost is a lock-in into a single vendor and cloud

Emerging to general availability:

- NVIDIA H200 - faster HBM and more memory than H100 - Q4-2024 on select clouds (not all big clouds are planning to stock on these).
- NVIDIA B200 and GB200 - are starting to emerge.
- AMD MI300X ~= H100 - Tier 2 clouds have those since Q2-2024 - you need to use the latest ROCm and activate many optimizations to get the high TFLOPs here. AMD MI325X > H200 will be available shortly.
- Intel Gaudi3 > H200 - is available on Intel's cloud
- GraphCore IPU - very difficult to find if at all, was shortly available on paperspace but no more.
- Cerebras WaferScale Engine - available on Cerebras' cloud

For the full list and more recently announced accelerators see [Accelerators](#).

Accelerator Interoperability

In general most (all?) accelerators are supported by major frameworks like PyTorch or TensorFlow and the same code should run everywhere with small modifications as long as it doesn't use any accelerator-specific functionality.

For example, if your PyTorch application calls `torch.mm` - it should work everywhere, but if it includes custom CUDA kernels it'll only work on NVIDIA GPUs and may be on the recent AMD MI-series.

- NVIDIA GPUs: all based on [CUDA](#), which most training frameworks support. You can easily move between different NVIDIA GPUs and most things would work the same.

- AMD MI250/MI300X: with PyTorch using [ROCm](#) you can run most CUDA-based software as is. This is really the only inter-operable accelerator with the NVIDIA stack.
- Intel Gaudi2/Gaudi3: if you use HF Transformers/Diffusers you can use [optimum-habana](#). If you use HF Trainer with NVIDIA GPUs it should be relatively easy to switch to train/infer on Gaudi2.
- GraphCore IPU: can also be run via PyTorch via [poptorch](#)
- Cerebras: is also working on PyTorch support via [Cerebras Software Platform \(CSoft\) via XLA](#).

Also in general most ML code could be compiled into cross-platform formats like [Open Neural Network Exchange \(ONNX\)](#) which can be run on a variety of accelerators. This approach is typically used more often for inference workloads.

Network

- If you want to train a large model that doesn't fit onto a single accelerator's memory you have to rely on the intra- and inter-node networks to synchronize multiple accelerators.
- The biggest issue right now is that compute hardware advancements move faster than networking hardware, e.g. for NVIDIA NVLink intra-node (unidirectional bandwidth):

GPU	Computefp16 TFLOPS	Compute speedup	Intra-node GBps	Intra-node speedup
V100	125	1	150	1
A100	312	2.5	300	2
H100	989	8	450	3
B200	2250	18	900	6

- You can see that A100 was 2.5 faster than V100, and H100 is ~3x faster than A100. But the intra-node speed of NVLink has only increased by 150GBps each generation. NVLink 5.0 doubled the speed over NVLink 4.0 so it catches up a little bit with the compute speed ups. But the speed up is still insufficient.
- Moreover, the first 4 generations of NVLink use identical NICs of the same 25GBps unidirectional bandwidth. They have just doubled and tripled the number of links to speed things up. So there was 0 progress in that technology.
- The inter-node situation isn't any better with most NICs there doing 100 or 200Gbps, and some 400Gbps are starting to emerge. (correspondingly in GBps: 12.5, 25 and 50). It's the same story here, some solutions provide dozens of NICs to get to higher speeds.
- Also typically with LLMs the payload is so large that network latency is often negligible for training. It's still quite important for inference.

Intra-node Network

- Pay attention to bytes vs bits. 1Byte = 8bits. 1GBps = 8Gbps.
- If you need to reduce bits (e.g. gradients) across multiple nodes, it's the slowest link (Inter-node) that defines the overall throughput, so intra-node speed doesn't matter then
- [Tensor parallelism](#) and [sequence parallelism](#) have to remain within the node to be efficient - only makes sense with fast intra-node speed

NVIDIA:

- NVIDIA-based compute nodes come with 50GBps duplex NVLink
- Some have a lot of NVLinks, others less but typically plenty w/ at least 450GBps (3.6Tbps) unidirectional bandwidth for H100, 300GBps for A100 nodes

Intel Gaudi2:

- 8 x 21 NICs of 100GbE RoCE v2 ROMA for a total of 2.1TBps

[More details](#)

Inter-node Network

- An order of magnitude slower than Intra-node
- You will see a wide range of speeds from 50Gbps to 3200 Gbps
- You need to reduce gradients and other bits faster than compute to avoid idling accelerators
- You typically get at most 80% of advertised speed. e.g., if you are told you get 800Gbps, expect ~640Gbps.
- If moving to fp8 H100 is 18x faster than V100
- We are yet to see if 3200Gbps for H100s will be enough to keep high MFU.
- Practically less than 3x but it's a good estimate

[More details](#).

Storage

There are 3 distinct Storage IO needs in the ML workload:

1. You need to be able to feed the DataLoader fast - (super fast read, don't care about fast write) - requires sustainable load for hours and days
 2. You need to be able to write checkpoints fast - (super fast write, fastish read as you will be resuming a few times) - requires burst writing - you want super fast to not block the training for long (unless you use some sort of cpu offloading to quickly unblock the training)
 3. You need to be able to load and maintain your codebase - (medium speed for both reading and writing) - this also needs to be shared since you want all nodes to see the same codebase - as it happens only during the start or resume it'll happen infrequently
- Most of the time you're being sold 80% of what you paid. If you want a reliable 100TBs you need to rent 125TBs or your application may fail to write long before the disk is full.
 - Shared Distributed Filesystem:
 1. non-parallel shared file systems can be extremely slow if you have a lot of small files (=Python!)
 2. You want Parallel FS like GPFS (IBM Spectrum Scale) or Lustre (Open Source)

[More details](#).

CPU Memory

You need enough memory for:

- 2-3 possibly DL workers per Accelerator (so 16-24 processes with 8 accelerators per node)
- Even more memory for DL workers if you pull data from the cloud
- Enough memory to load the model if you can't load to accelerator directly
- Often used for accelerator memory offloading - extends accelerator's memory by swapping out the currently unused

layers - if that's the target use, then the more cpu memory is available - the better!

CPU

This is probably the least worrisome component.

- Most clouds provide beefy CPUs with plenty of cpu cores
- You need to have enough cores to run 2-3 DL workers +1 per gpu - so at least 30 cores
- Even more cores for DL workers if you have complex and/or slow DL transforms (CV)
- Most of the compute happens on GPUs

Impress others with your ML instant math

Tell how many GPUs do you need in 5 secs

- Training in half mixed-precision: `model_size_in_B * 18 * 1.25 / gpu_size_in_GB`
- Inference in half precision: `model_size_in_B * 2 * 1.25 / gpu_size_in_GB`

That's the minimum, more to have a bigger batch size and longer sequence length.

Here is the breakdown:

- Training: 8 bytes for AdamW states, 4 bytes for grads, 4+2 bytes for weights
- Inference: 2 bytes for weights (1 byte if you use quantization)
- 1.25 is 25% for activations (very very approximate)

For example: Let's take an 80B param model and 80GB GPUs and calculate how many of them we will need for:

- Training: at least 23 GPUs $80*18*1.25/80$
- Inference: at least 3 GPUs $80*2*1.25/80$

[More details.](#)

Traps to be aware of

As you navigate this very complex AI industry here are some thing to be aware of:

Say no to "will make a reasonable effort to ..." contracts

- If you contract doesn't have clear deliverables (time and performance) don't be surprised if you paid for something you won't receive in time you need it or not at all
- Be very careful before you sign a contract that includes clauses that start with "we will make a reasonable effort to ...".

When was the last time you went to the bread section of the supermarket and found a lump of half-baked dough with a note "we made a reasonable effort to bake this bread, but alas, what you see is what you get"?

But for whatever reason it's acceptable to create a legal contract where the provider provides neither delivery dates nor performance metrics and doesn't provide stipulations for what will they do in recompense when those promises aren't fulfilled.

Beware of hardware and software lock-in scenarios

- Some cloud providers will make you use very proprietary tools or hardware that will make it very difficult for you to leave down the road because you will have to retool everything if you leave
- Consider what would be the cost of moving to a different provider should this provider prove to be not satisfactory or if they don't have a capacity to fulfill your growing needs.
- If you rent a cluster with a generic Linux box with generic open source tools it should be trivial to move from one provider to another as almost everything would work out of the box
- Obviously if you choose compute that requires custom software that works for that hardware only and you can't rent this hardware anywhere else you're setting yourself up for a lock-in

Don't buy what you don't really need

- The cloud providers have mostly the same generic hardware, which leads to a very slim \$\$ margin and so in order to make big \$\$ they invent products and then try to convince you that you need to buy them. Sometimes you actually need those products, but very often not. See also the previous section on lock-in, since proprietary products usually mean a partial lock-in.
- Often it's easy to observe the 3 step marketing technique for solutions that seek a problem to solve:
 1. Convince a couple of well respected customers to use the provider's proprietary products by giving them huge discounts or even pay them to use them
 2. Use those in step 1 as the social approval lever to reel in more converts
 3. Then scoop the rest of the strugglers by telling them that 80% of your customers (1+2) use these amazing products

When marketing these products it's important:

- to mention how well they work with a dozen of other products, since now you're not buying into a single product but into a whole proprietary product-sphere.
- to use really nice looking complicated diagrams of how things plug into each other, and move really fast to the next slide before someone asks a difficult question.

HPCs are probably a good group of compute providers to learn from - they have no funds to create new products and so they creatively address all their needs using mostly generic open source tools with some custom written software added when absolutely needed.

Unsolicited advice

To conclude I thought I'd share some insights to how one could slightly improve their daily AI battlefield experience.

FOMO and avoiding depression

If you read Twitter and other similar ML-related feeds you're guaranteed to feel the fear of missing out, since there is probably at least one new great model getting released weekly and multiple papers are getting published daily and your peers will publish their cool achievements every few minutes.

We are dealing with **very complex** technology and there is a small handful of people who can absorb that much new material and understand / integrate it.

This can be extremely depressing and discouraging.

I deal with it by looking at twitter about once or twice a week. I mostly use Twitter in broadcast mode - that is if I have something to share I post it and only watch for possible follow up questions.

Usually all the important news reach me through other people.

Don't try to know everything

The pace of innovation in the field of AI is insane. It's not possible to know all-things-AI. I'd dare to say it's not possible to know even 10% of it for most of us.

I realized this very early one and I stopped paying attention to most announcements, tutorials, keynotes, etc. Whenever I have a new need I research it and I discover what I need and I have to be careful not to try to learn other things not pertinent to the goal at hand.

So I actually know very little, but what I have researched in depth I know quite well for some time and later I forget even that (that's why I write these notes - so that I can easily find what I have already researched).

So if you ask me something, chances are that I don't know it, but the saving grace for me is that if you give me time I can figure it out and give the answer or develop a solution.

Don't beat yourself up when using half-baked software

Because the ML field is in a huge race, a lot of the open source software is half-baked, badly documented, badly tested, at times poorly supported. So if you think you can save time by re-using software written by others expect spending hours to weeks trying to figure out how to make it work. And then keeping it working when the updates break it.

The next problem is that most of this software depends on other software which often can be just as bad. It's not uncommon where I start fixing some integration problem, just to discover a problem in a dependent package, which in its turn has another problem from another package. This can be extremely frustrating and discouraging. Once expects to save time by reuse, but ends up spending a long time figuring out how to make it work. At least if I write my own software I have fun and it's a creative process, trying to make other people's software work is not.

So at the end of the day we are still better off re-using other people's software, except it comes at an emotional price and exhaustion.

So first of all, try to find a way not to beat yourself up if the software you didn't write doesn't work. If you think about it, those problems aren't of your creation.

Learning how to [debug efficiently](#) should also make this process much less painful.

How to Choose a Cloud Provider

Having used multiple compute clouds over long and short terms, and participating in many "discovery" calls, I've learned that it's absolutely crucial to approach the cloud choosing process with an utmost care and dedication. Especially for the long term contracts - you may end up in a 3-year lock-in where you pay millions of dollars and end up having a terrible experience and no way to get out of the contract.

To give you a perspective - a 64-node cluster may easily cost USD\$20-50M over a 3 year period. This is often more than what startups pay for the salaries.

I can't stress this enough that choosing a bad 3-year contract may prevent your startup from succeeding.

In this article I'm not going to tell which clouds to avoid, but instead try to empower you to avoid having a bad experience and to have at least a decent one, that will give your company a chance to succeed.

These notes assume you already know what compute you want for your specific workloads. If you don't please skim through the [Accelerator](#), [Storage](#) and [Network](#) chapters to know what's available out there. Most of the time you want the latest the clouds have to offer.

Glossary

- CSP: Cloud Service Provider
- SLA: Service-level_agreement
- SLO: Service Level Objective

Contracts

If you're paying per hour, you don't need to worry about contracts. But this method isn't good long term because you will be paying many times more and you won't have a steady reliable accelerator foundation. A long term contract at times and with a good negotiator can lead to a 10x in TCO savings (and time)!

Free Trials

Most cloud service providers (CSPs) have trial programs where you can "kick the tires" for a few days/weeks on a few nodes for free.

Granted, it won't give you an indication of how well the bigger cluster would scale, but it should be sufficient to be able to run quite a few benchmarks and experiments.

It also will give you a good opportunity to check how the provider's customer support works (if any support is included in the free package that is).

Half-baked solutions

Since a new generation of accelerators happens roughly every 12-18 months and the customer wants those latest accelerators "yesterday" to have a business advantage over their competitors - this gives CSPs barely any time to integrate the new generation of the hardware, test it, adapt their software stack and burn those components in.

So if you want the latest generation as soon as it becomes available you're almost guaranteed to have a bad experience because, well, time is needed to get things right - we are talking about months of waiting. But customers rule - so the CSPs give them what they want, often not quite telling that what the customer gets is not quite ready.

I'm not sure if CSPs are to blame, because often they get the hardware delivery months after it was promised by the

manufacturers and, of course, by now they can't keep their promises to the customers, so they just go ahead and deliver...

Then some CSPs develop their own hardware (e.g. network stack) in order to have better margins and then they fail to complete those custom solutions in time, the latest accelerators are there, but the whole system is limping. It's much safer when off-the-shelf components are offered, since those are most likely to be well-tested working components (expect it's likely to cost more).

I think it's OK if the customer wants the hardware early, there should just be an honest disclosure as in: *"look we need some 3 more months to make things solid, if you want the nodes now you can have them but we can't guarantee anything."*

We-will-do-our-best clause

A lot of the long-term cloud contracts are likely to include a lot of "we will do our best" clauses.

Yet:

1. The customer is not allowed to "do their best" to pay, they are legally obliged to pay the amount they agreed to pay and on time.
2. The customer is not allowed to break a contact before its term runs its course.

In my experience "we will do our best" is demonstrated by Tier-1 clouds by sending 10+ people to the meetings with the customers. Some of them will be clueless and will be just sitting there making the company look resourceful: *"look, we are allocating 10+ people to the problem you're experiencing. You have nothing to worry about"*. Except, most of the time those people can't solve your problem.

What you need is just 2 cloud support people on the call - one product manager and one engineer directly responsible for solving the problem at hand. And in my experience this sort of meeting could take weeks to months to manifest or not at all. Usually one needs to have good connections to be able to escalate the issue to "top brass".

For every critical component of the package you're purchasing you need a quantifiable delivery. For example, if the network you were sold is supposed to run at X GBps at that many nodes doing all-reduce, and you measured it to be significantly lower, there should be a stipulation of what the CSP will do when this happens. How long do they have to fix the problem and whether you can break a contract should this not happen within the agreed by both sides time.

Same goes for storage, accelerators and any other critical component that you plan to rely on.

Of course, it's up to you to negotiate the specific repercussions, but probably the best one is that you stop paying until the problem is fixed. That way there is a huge incentive for the problem to be fixed.

Alas, not paying helps, but not being able to use the compute is still a huge problem. And breaking the contract and migrating to another provider is a huge undertaking not to be taken lightly. But at least there is something you could do if you don't get what you need.

I must also say that it's almost never the problem of the engineers, very often they are amazing experienced people - most of the time it's the issue of management and resource allocation. So please be as gentle as possible with the people you interact with, while firmly demanding a resolution. I know it's a difficult one - more than once I was at the end of the rope, and I couldn't always keep it cool.

Service Level Agreement

As a continuation of a previous section, a [Service Level Agreement \(SLA\)](#) is an agreement between a service providers and a customer that define various guarantees and expectations with regards to service quality and availability, and various responsibilities.

The other term is Service Level Objective (SLO) where SLA is quantified. For example, an SLO may define a Monthly Uptime Percentage to 99.5%, if the uptime is less than 99.5% the provider credits the customer to a certain percentage of the \$\$ spent. For example, 10% if the uptime is 99-99.5%, 25% for 95-99%, etc. Here a [GCP SLA](#).

The main category one should care for when renting ML clusters is failing accelerators and/whole nodes. If you paid for 64

nodes but were able to use only 60 you should be reimbursed/credited for those nodes you couldn't use. Your SLA should define the duration of downtime after which the provider starts paying you back and how much.

Same goes for network and storage, albeit those typically fail a lot less often than accelerators, but they do fail.

In general any critical part of the service should have an SLO and clearly defined repercussions if the SLOs aren't met.

Most Tier 1 companies should already include their standard SLAs in the contract. In theory the customer should be able to negotiate those to adapt to their needs, though it might not always be possible. Sometimes offering to pay more may allow for a better than standard SLO.

Discuss a contract breaking clause

Both sides should be empowered to experience a mutually beneficial business experience.

Therefore it's critical that you should be able to legally exit the contract should your business experience not be beneficial because the other side is failing to meet the agreed upon expectations.

This, of course, implies not to have a legal battle which can be very costly and Tier-1 clouds have a lot of money to hire the best lawyers, so it might be a losing battle.

It's up to you to negotiate under which circumstances the contract can be cleanly exited before its term runs out.

Must have paid support included

In one of the companies I worked at our cloud contract didn't include the paid support service and the only support we had was via a customer chat. The paid support was skipped to save costs, but boy did we end up losing days of compute because of that.

Do not try to save here - you will end up losing a lot of money, developer time and hair. Make sure you have a way to submit tickets with priority labels and a defined in the contract expectation to how quickly they will be dealt with.

When you try to use customer chat to solve an urgent problem, there is zero obligation for them to do anything, or at least to do it in a timely manner.

If you're dealing with PMs, you need to know how quickly you could talk directly to the end-point engineer, while removing the middle-man.

Support during off-hours

Do you get human support for emergencies on weekends/holidays/nights? e.g. On one of the HPCs I used the human support was only available Mon-Fri 9-5.

If this is not available, at the very least ensure that your team can perform cluster resuscitation themselves - and do a drill to ensure this is actually doable. This means you need to have an API to perform all those things without the provider's support.

Next generation accelerator migration

On average a new generation of accelerators comes out every 12-18 months, but a typical contract is for 3 years. Which means that for about half of that time you will end up using an inferior product.

Nobody wants to use a 2-5x slower accelerator when a much faster version is available, but most customers now are stuck with the old accelerators for the full 3 year duration.

You need to negotiate the ability to move to the new generation before the end of the term, which would obviously require some additional money paid for this to happen.

Accelerators

This group of questions/issues is specific to accelerators.

Accelerators need to be burned in

When a new batch of components arrives the provider has to "burn them in" before handing them to customers. This is a process of running an extensive stress testing to detect any accelerators and other system components that are faulty.

If this is not done, the customer ends up discovering the "bad apples" the hard way, while running their workloads. This leads to lost compute and developer time. If the workload uses a few nodes, one failing accelerator isn't a big problem most of the time, but if the workload uses dozens or hundreds of nodes the cost is huge.

It shouldn't be the responsibility of the customer to discover bad accelerators. And while there is no guarantee that the accelerator will not fail after it has been stress tested - it should happen rarely.

Otherwise, a new batch of accelerators often has a 3-10% failure rate, which is huge and very costly to the customer!

So ask your provider how long did they burn in your accelerators/systems for, if at all.

I'm yet to find a golden reference point, but, for example, [SemiAnalysis](#) suggests that OEM provider performs a 3-4 weeks burn-in, and then the CSP conducts another 2-3 day long burn-in/acceptance test. So if that's the case you want to ensure that the systems were stress-tested for at least 2-3 days.

Dealing with accelerator failures

In my experience, while other compute components do fail occasionally, 95% of the time it's the accelerators that fail.

Therefore you need to have a very clear and quick path to an accelerator replacement.

Ideally this process needs to be automated. So you need to ask if there is an API to release a broken node and get a replacement. If you have to ask a human to do that, it usually doesn't work too well. The more automated things are, the more efficient the experience.

How many accelerators do you have in the provider-side back up pool available to you? They will usually commit to a certain number of fast replacement per month.

That's said if time is of an essence to your workflows, as most of the time you won't be able to get instant replacements you should always pay for about 10% more nodes than you need. The extra nodes can be used for development and if you have failing nodes during training you can instantly use your own extra nodes.

Ensure all your nodes are on the same network spine

Unless you're renting 10k gpus, most smaller clusters can easily be co-located on the same network spine - so that it takes the same time to perform inter-node network traffic from any node to any other node.

Ensure that any back up nodes that you're not paying for, but are there to deal with failing accelerators, reside on the same network spine as the nodes you're paying for. If they don't, you are going to have a big problem if you do multi-node training - since that one replacement node will be further away from all other nodes and will slow the ensemble down (the weakest link in the chain).

Ensure you keep your good accelerators on reboot

You want your cluster to have a fixed allocation. Which means that if you need to re-deploy nodes, and especially if you're planning a downtime, other customers aren't going to grab those nodes!

Once you spent weeks filtering out the bad nodes from the good nodes, it's crucial to keep those nodes to yourself and not start the painful and costly filtering again.

Do you think you will need to expand?

This is a difficult one, because it's hard to know ahead of time if the amount of nodes you're asking for will need to grow in the future.

Ideally you'd want to discuss this with your provider in case they could plan for your imminent expansion.

Because otherwise, say, you want to double the number of your nodes, but in order to get more nodes, they could only be allocated on another network spine - this is going to be a problem, as it'd impact the training speed.

Chances are that you will have to drop your current allocation and move to another bigger allocation - possibly even in a different region if they don't have local capacity. And moving to a different region can be a very slow and costly experience because you have to move your storage to where your new cluster is. Based on a personal experience - don't treat this lightly.

Storage

Large and fast storage is very important for both - good developer experience and fast training/fine-tuning/inference workloads - in particular with regards to loading/saving checkpoints.

Guaranteed maximum capacity

Ask how much of the storage you will be paying for is guaranteed.

For example, if the Lustre filesystem is used the customer needs to know that they have to over-provision by 25% to get the actual storage capacity they need, because Lustre can fail to write at 80% total storage capacity, because of bad disk balancing design. And the onus of paying for the extra 25% is on the customer!

Most other filesystems I had an experience with typically reach 100% capacity without failing, but it's always good to ask for the specific filesystem you plan to use.

Know your storage IO requirements

At one of the clouds we used a non-parallel distributed filesystem and the developer experience was absolutely terrible. While dealing with large files was acceptable, the small files experience was extremely slow - it'd take 30 minutes to install a basic Conda environment and 2 minutes to run `python -c "import torch"`. This is because Python has tens of thousands of 4-16kb files and if the file system isn't optimized to handle those and the meta-data servers are weak, it'd be a very frustrating experience.

In general a typical Python shop needs a filesystem that can deal with:

- tens of thousands of tiny files
- few huge files

But, of course, only you know what your workloads' specific requirements are. Also consider the relationship between local storage and remote (shared) storage, as some providers will reduce the size and performance of local drives to save money. In many cases, developers will read data from a shared filesystem that can be cached locally (code libraries, models, datasets). Teaching people how to use `rsync` with local NVMe can improve the developer experience, and reduce I/O on the shared filesystem.

Please refer to the notes and guidance in the [\[..\]/storage/\]\(https://github.com/stas00/ml-engineering/blob/master/insights/Storage chapter\)](#) to know the nuances of storage requirements and their benchmarking.

What happens when storage fails

With advanced expensive distributed filesystems the chance of failure is relatively small, but it's quite big with cheaper storage solutions.

But it may still happen with any system.

You need to know:

- Who is in charge of fixing the problem?
- How long will it take to recover?
- Who pays for the downtime?
- What are the users to do while there is the problem?

If the resolution will take a long time often one needs to add another temporary filesystem partition to enable people to do their work. And, of course, you will have to pay for it.

Region migration

A cluster may be forced to migrate to a different region when upgrading to a next generation accelerators or expanding the capacity, if the region you're in doesn't have what you need. The storage has to be in the same region as the accelerators for the workflows to be fast.

The migration event triggers a sometimes very painful storage migration experience.

Here are some critical questions you need to ask long before the migration starts.

- Is the provider responsible for moving your data or is it your responsibility?
- Have you checked that the provided tooling is good enough to move TBs of data in a few hours, or will it take many days to move? For example, using a storage cloud to migrate will typically drop all file metadata, which can be a huge problem. If you have 5 million tiny files, it could take forever to copy. Unless you use `tar`, but which may take many hours to create and do you have the 2x storage to have 2 copies of your data?
- Are you supposed to pay for the storage and the compute for both overlapping clusters?
- What happens to the files being edited and created while the filesystem is on the move - do you send everybody home while the migration is happening and freeze the filesystem?

Backup and Archive

Many CSPs only have one tier of file storage available at one price point. However, organisations can have needs for multiple tiers of storage. For example, you might want to archive old model checkpoints or finetuning datasets to cheap, cold storage such as S3 object on HDD.

Having the flexibility to expand your total storage capacity, and keep the "hot" (local NVMe), "warm" (shared NVMe), "cold" (shared HDD), and "archive" (tape) in sync can help improve the resiliency of systems, save money, and allow for easier migration or expansion over time.

Network

This segment is mostly relevant to those planning to do training and finetuning. If you need to rent accelerators either for inference via large deployments of microservices or for small, on-demand, interactive work (i.e. notebooks) you can safely ignore this information. The only exception is when you plan on inferencing very big models that require more than one node for a single replica.

In general you want to ensure that the offered [intra-node](#) and [inter-node](#) network speeds match the promise and your expectations.

Ask for the actual performance numbers

Compute theory never matches reality, and the reality may dramatically vary from provider to provider even if they all use the same components, as it depends on the quality of all involved components and how well the racks were designed and put together.

The easiest ask is to request an `all-reduce` benchmark plot over 4-8-16-32-64 nodes (or more if your cluster is more than 64 nodes). You'd expect the bandwidth to gradually become worse with more participating nodes, but not dramatically so. Some networks become very inefficient at higher number of nodes.

Please refer to [Real network throughput](#) for more details.

Ideally you want to benchmark at least a few payloads - the ones that are of a particular interest to you because you know that this is the collective payload you will be using in your workloads. I usually just start by asking for a plot of a big payload of about 4-16GB (16GB would get the best bandwidth on the latest fastest inter-node networks), if the performance drops below 80% of the theoretical GBps, then I know we have a problem.

Does the network steal from the accelerator memory?

One surprise I experienced on one of the clouds is that when I started using the GPUs I discovered that 5GB of each was already used by the networking software - we managed to reduce it to a lower value, but still we were sold GPUs with less than their memory size and nobody told us about that before we signed the contract.

As accelerators become much bigger this will probably become unimportant, but when you get 75GB of usable memory instead of 80GB on H100 - that's a huge amount of memory lost per GPU.

Infiniband or Ethernet?

In general, CSPs follow NVIDIA's [DGX SuperPOD Reference Architecture](#) which provides a lot of detail on how to build a rail-optimized InfiniBand network. Rail-optimized basically means that each GPU in an 8-way system connects to its own leaf switch. Everything else is a standard fat-tree.

However, many of the largest GPU clusters in the world now run RoCEv2 instead of Infiniband. Meta has [proven](#) that you can train frontier-class Llama models on a RoCEv2 network. Semianalysis/Fabricated Knowledge show a [significant drop-off](#) in NVIDIA's networking attach rate for their GPUs.

Since multi-node training depends on network collectives (i.e. NCCL or RCCL), the type of network can significantly impact performance and user experience.

Miscellaneous

Tier 1 vs Tier 2 clouds

I don't yet have a clear recommendation for whether Tier 1 clouds (AWS, GCP, Azure, etc.) vs emerging smaller Tier 2 clouds are better. My intuition is that Tier 2 clouds are likely to provide a better and more personal support as they have to work harder to secure customers.

Price-wise, Tier 2 clouds in general are cheaper because otherwise they won't be able to compete with Tier 1 clouds. However, it's obvious that their "margin" will be much smaller, because Tier 2 clouds don't have the volume buying power of Tier 1 clouds.

Tier 2 clouds are more likely to be more flexible, have non-mainstream accelerators (e.g., AMD and Intel) and probably are more likely to lend hand at tuning things up at no to little cost.

Orchestration

A well-oiled node orchestration is critical for successfully using multi-node clusters.

Make sure you know which one you need - usually [SLURM](#), Kubernetes or a combination of the two and make sure it's well supported. Some clouds would only support one of them, or provide a very limited support for another type. These days SLURM is mostly used for training/finetuning and Kubernetes for inference. And there are other [emerging orchestration platforms out there](#).

Same as with hardware, depending on whether you're planning to administrate your own cluster you need to know who will deal with any problems. This is a very crucial component of your stack, since if the orchestration is broken, nobody can use the cluster and you lose time/money.

Up-to-date software/OS versions

Make sure to ask that the provider isn't going to force you into some old versions of the software and an operating system. I have had experiences where we were forced to use some very old Ubuntu versions because the provider's software stack which we had to use wasn't supporting more recent and up-to-date OS.

System administration

These days it can be difficult to find a good system administrator that understands the specific needs of the ML workloads, so it's a good idea to ask if some of that work could be offloaded to the CSP. Tier-1 CSPs sub-contract service companies that can provide various degrees of system administration. Smaller clouds are likely to offer their own direct services. They usually have a good grasp of what ML workloads need.

You won't be able to succeed without someone experienced taking care of your cluster. Using your ML engineers to also deal with system administration work can be very counter-productive, since it can be a very time-demanding and interrupting work.

Either hire a system administrator or hire a service company that will do it for you.

Conclusion

These notes are based on my direct experience and clearly I haven't been exposed to all possible things that may go wrong and wreck havoc with your cluster or make your whole team burn out and lose a lot of their hair. But this should be a good foundation to start thinking about.

Add your own questions, by thinking what's important for you, what failures may prevent you from accomplishing your compute goals.

If you have a particular CSP that you're casing out ask the community about them, especially what pitfalls to avoid with that cloud.

The key message of this article is for you to choose a cloud where your choice hasn't been taken away and that you don't get stuck with a service your developers hate, which is likely to lead to people leaving your company.

If you feel that these notes are overwhelming for you, I occasionally consult helping with due diligence and joining discovery calls. You can contact me at stas@stason.org.

Compute

1. [**Accelerator**](#) - the work horses of ML - GPUs, TPUs, IPUs, FPGAs, HPUs, QPUs, RDUs (WIP)
2. [**CPU**](#) - cpus, affinities (WIP)
3. [**CPU Memory**](#) - how much CPU memory is enough - the shortest chapter ever.

Accelerators

Compute accelerators are the workhorses of the ML training. At the beginning there were just GPUs. But now there are also TPUs, IPUs, FPGAs, HPUs, QPUs, RDUs and more are being invented.

There exist two main ML workloads - training and inference. There is also the finetuning workload which is usually the same as training, unless a much lighter [LORA-style](#) finetuning is performed. The latter requires significantly fewer resources and time than normal finetuning.

In language models during inference the generation is performed in a sequence - one token at a time. So it has to repeat the same `forward` call thousands of times one smallish `matmul` (matrix multiplication or GEMM) at a time. And this can be done on either an accelerator, like GPU, or some of the most recent CPUs, that can handle inference quite efficiently.

During training the whole sequence length is processed in one huge `matmul` operation. So if the sequence length is 4k long, the training of the same model will require a compute unit that can handle 4k times more operations than inference and do it fast. Accelerators excel at this task. In fact the larger the matrices they have to multiply, the more efficient the compute.

The other computational difference is that while both training and inference have to perform the same total amount of `matmuls` in the `forward` pass, in the `backward` pass, which is only done for training, an additional 2x times of `matmuls` is done to calculate the gradients with regards to inputs and weights. And an additional `forward` is performed if activations recomputation is used. Therefore the training process requires at 3-4x more `matmuls` than inference.

Subsections

General:

- [Benchmarks](#)

NVIDIA:

- [Troubleshooting NVIDIA GPUs](#)

AMD:

- [Troubleshooting AMD GPUs](#)
- [AMD GPUs Performance](#)

Bird's eye view on the high end accelerator reality

While this might be changing in the future, unlike the consumer GPU market, as of this writing there aren't that many high end accelerators, and if you rent on the cloud, most providers will have more or less the same few accelerators to offer.

GPUs:

- As of today, ML clouds/HPCs started transitioning from NVIDIA H100s to H200s and this is going to take some months due to the usual shortage of NVIDIA GPUs. B200, GB200 were announced in Q1-2024, but it'll probably take till mid-2025 before we will be able to use those, because of the delays in production. B300 were announced on 2024-12!
- AMD's MI300X is now widely available on Tier 2 cloud providers. MI325X is supposed to become available in early 2025. MI355X should be there towards the end of 2025. MI400X hopefully in 2026.

HPU:

- Intel's Gaudi2 is available at Intel's cloud. It's also available on-premises implementations via Supermicro, WiWynn,

and soon others.

- Gaudi3 is available since late 2024.
- Falcon Shores is to replace Gaudi in 2025
- Jaguar Shores is to replace Falcon Shores in 2026

IPU:

- Graphcore with their IPU offering was briefly available at Paperspace, but it's gone now. I'm not sure if anybody offers those.

TPU:

- Google's TPUs are, of course, available but they aren't the most desirable accelerators because you can only rent them, and the software isn't quite easily convertible between GPUs and TPUs, and so many (most?) developers remain in the GPU land, since they don't want to be locked into a hardware which is a Google monopoly.

On Pods and racks:

- Cerebras' WaferScale Engine (WSE)
- SambaNova's DataScale
- dozens of different pod and rack configs that compose the aforementioned GPUs with super-fast interconnects.

That's about it as Q4-2024.

As most of us rent the compute, and we never see what it looks like, here is a how an 8xH100 node looks like physically (this is the GPU tray of the Dell PowerEdge XE9680 Rack Server):

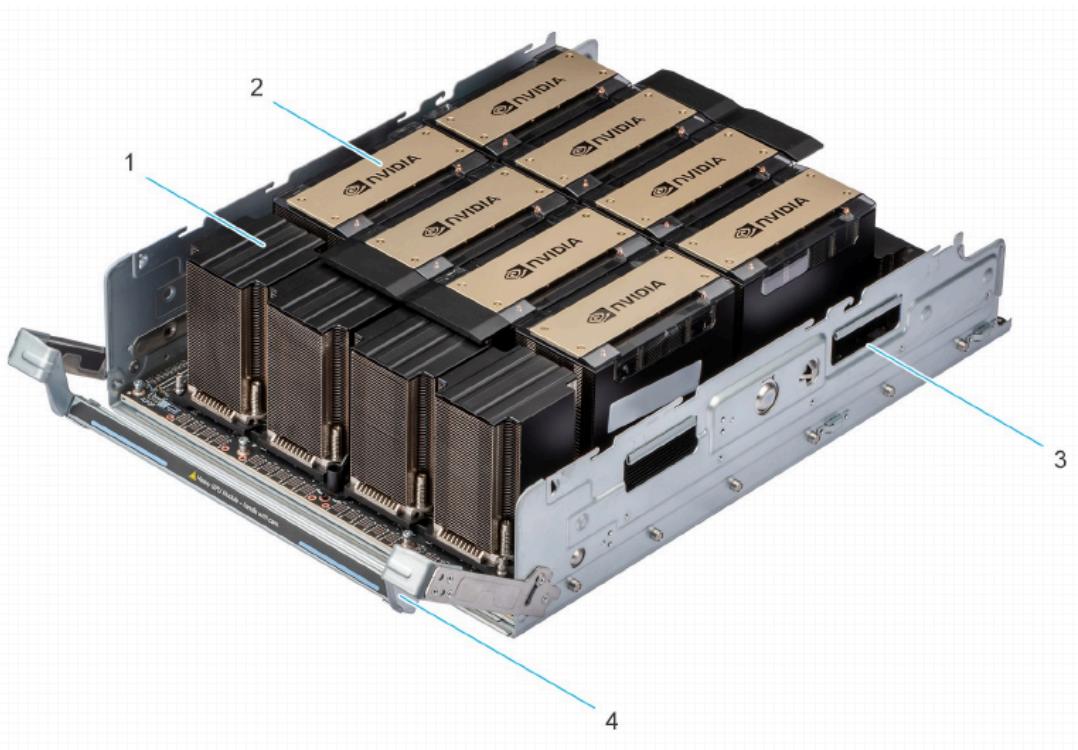


Figure 8. GPU H100/H800 Inside view

- | | |
|----------------------|-----------------|
| 1. NVSwitch Heatsink | 2. GPU Heatsink |
| 3. Chassis | 4. Handle |

Glossary

- CPU: Central Processing Unit
- FMA: Fused Multiply Add
- FPGA: Field Programmable Gate Arrays
- GCD: Graphics Compute Die
- GPU: Graphics Processing Unit
- HBM: High Bandwidth Memory
- HPC: High-performance Computing
- HPU: Habana Gaudi AI Processor Unit
- IPU: Intelligence Processing Unit
- MAMF: Maximum Achievable Matmul FLOPS
- MME: Matrix Multiplication Engine
- QPU: Quantum Processing Unit
- RDU: Reconfigurable Dataflow Unit
- TBP: Total Board Power
- TDP: Thermal Design Power or Thermal Design Parameter
- TGP: Total Graphics Power
- TPU: Tensor Processing Unit

[Additional glossary @ Modal](#)

The most important thing to understand

I will make the following statement multiple times in this book - and that it's not enough to buy/rent the most expensive accelerators and expect a high return on investment (ROI).

The two metrics for a high ROI for ML training are:

1. the speed at which the training will finish, because if the training takes 2-3x longer than planned, your model could become irrelevant before it was released - time is everything in the current super-competitive ML market.
2. the total \$\$ spent to train the model, because if the training takes 2-3x longer than planned, you will end up spending 2-3x times more.

Unless the rest of the purchased/rented hardware isn't chosen carefully to match the required workload chances are very high that the accelerators will idle a lot and both time and \$\$ will be lost. The most critical component is [network](#), then [storage](#), and the least critical ones are [CPU](#) and [CPU memory](#) (at least for a typical training workload where any CPU limitations are compensated with multiple [DataLoader](#) workers).

If the compute is rented one usually doesn't have the freedom to choose - the hardware is either set in stone or some components might be replaceable but with not too many choices. Thus there are times when the chosen cloud provider doesn't provide a sufficiently well matched hardware, in which case it's best to seek out a different provider.

If you purchase your servers then I recommend to perform a very indepth due diligence before buying.

Besides hardware, you, of course, need software that can efficiently deploy the hardware.

We will discuss both the hardware and the software aspects in various chapters of this book. You may want to start [here](#) and [here](#).

What Accelerator characteristics do we care for

Let's use the NVIDIA A100 spec as a reference point in the following sections.

	A100 80GB PCIe	A100 80GB SXM
FP64	9.7 TFLOPS	
FP64 Tensor Core	19.5 TFLOPS	
FP32	19.5 TFLOPS	
Tensor Float 32 (TF32)	156 TFLOPS 312 TFLOPS*	
BFLOAT16 Tensor Core	312 TFLOPS 624 TFLOPS*	
FP16 Tensor Core	312 TFLOPS 624 TFLOPS*	
INT8 Tensor Core	624 TOPS 1248 TOPS*	
GPU Memory	80GB HBM2e	80GB HBM2e
GPU Memory Bandwidth	1,935 GB/s	2,039 GB/s
Max Thermal Design Power (TDP)	300W	400W ***
Multi-Instance GPU	Up to 7 MIGs @ 10GB	Up to 7 MIGs @ 10GB
Form Factor	PCIe Dual-slot air-cooled or single-slot liquid-cooled	SXM
Interconnect	NVIDIA® NVLink® Bridge for 2 GPUs: 600 GB/s ** PCIe Gen4: 64 GB/s	NVLink: 600 GB/s PCIe Gen4: 64 GB/s
Server Options	Partner and NVIDIA-Certified Systems™ with 1-8 GPUs	NVIDIA HGX™ A100-Partner and NVIDIA-Certified Systems with 4,8, or 16 GPUs NVIDIA DGX™ A100 with 8 GPUs

[source](#)

TFLOPS

As mentioned earlier most of the work that ML training and inference do is matrix multiplication. If you remember your algebra matrix multiplication is made of many multiplications followed by summation. Each of these computations can be counted and define how many of these operations can be performed by the chip in a single seconds.

This is one of the key characteristics that the accelerators are judged by. The term TFLOPS defines how many trillions of FloatingPointOperations the chip can perform in a second. The more the better. There is a different definition for different data types. For example, here are a few entries from the theoretical peak TFLOPS from [A100 spec](#):

Data type \ TFLOPS	w/o Sparsity	w/ Sparsity
FP32	19.5	n/a
Tensor Float 32 (TF32)	156	312
BFLOAT16 Tensor Core	312	624
FP16 Tensor Core	312	624
FP8 Tensor Core	624	1248
INT8 Tensor Core	624	1248

Notes:

- INT8 is measured in TeraOperations as it's not a floating operation.
- the term FLOPS could mean either the total number of FloatingPointOperations, e.g. when counting how many FLOPS a single Transformer iteration takes, and it could also mean FloatingPointOperations per second - so watch out for the context. When you read an accelerator spec it's almost always a per second definition. When model architectures are discussed it's usually just the total number of FloatingPointOperations.

So you can see that int8 is 2x faster than bf16 which in turn is 2x faster than tf32.

Moreover, the TFLOPs depend on the matrices size as can be seen from this table:

Mixed Precision Matrix Multiply on A100

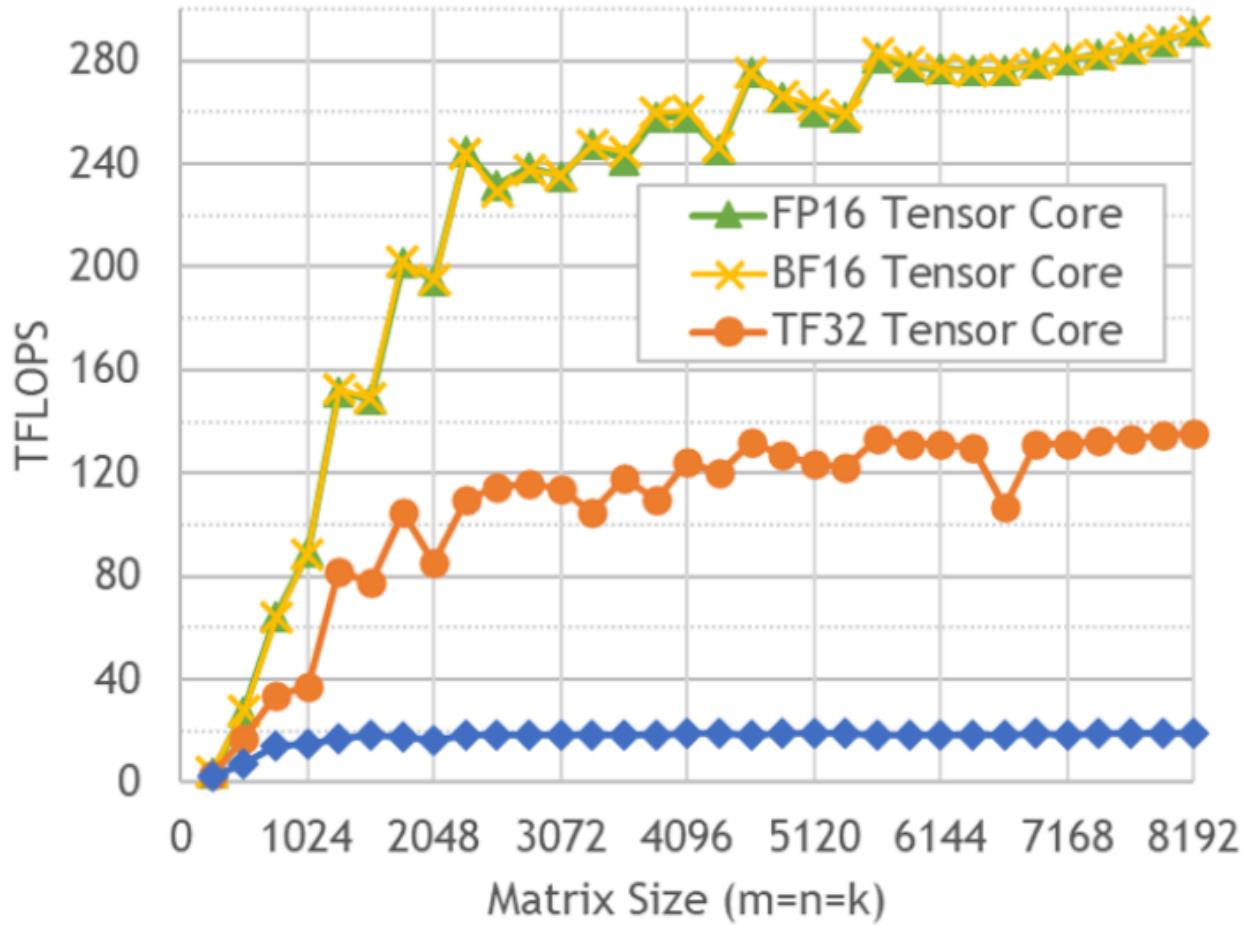


Figure 6. Mixed-precision matrix multiply on A100 with cuBLAS.

[source](#)

As you can see the difference in performance is non-linear due to [the tile and wave quantization effects](#).

How To Calculate Theoretical TFLOPS

Theoretical peak FLOPS is what gets published on the accelerator's spec. And it's calculated as:

```
Theoretical FLOPS = compute_unit_clock_speed * FLOPs_per_clock_cycle_per_compute_unit * num_compute_units
```

where:

- `compute_unit_clock_speed` - how many times the compute unit clock ticks per second in Hz
- `flops_per_clock_cycle_per_compute_unit` - the number of floating point operations the compute unit can execute per clock cycle.
- `num_compute_units` - how many units there is in the device

FLOPs per clock cycle per compute unit is usually not published, but what one often finds is the FMAs per clock cycle per compute unit specs. FMA is Fused Multiply Add. And since 1 FMA is made of 2 FLOPs, we can expand the above formula to:

```
Theoretical TFLOPS = compute_unit_clock_speed * FMAs_per_clock_cycle_per_compute_unit * 2 * num_compute_units
```

Let's validate that this formula checks out. Let's compute some BF16 (half precision) TFLOPS and compare to the published specs.

First, let's extract the necessary accelerator specs from [wiki](#).

The tricky part was to find the FMAs ops per CUDA core per clock cycle for BF16 (half precision). I found them [here](#). Most are coming from the [A100 whitepaper](#) (search the pdf for "FMA" and then choose the ones listed for the target precision you're after). The [H100 whitepaper](#) omitted a lot of specific FMA numbers, but included the multipliers wrt FMAs listed in the A100 whitepaper).

For NVIDIA @ BF16:

For NVIDIA BF16 operations a compute unit is a CUDA core.

Accelerator	Boost Clock	FMAs ops per CUDA Core per clock cycle	CUDA Cores	Spec TFLOPS
H100 SXM	1980Mhz	512	528	989
A100 SXM	1410MHz	256	432	312

Now let's do the math, by inserting the numbers from the table above into the last FMA-based formula:

- $1980*10^{12} * 512 * 2 * 528 / 10^{12} = 1070.530$ TFLOPS
- $1410*10^{12} * 256 * 2 * 432 / 10^{12} = 311.87$ TFLOPS

The calculated A100 SXM TFLOPS number matches the published 312 TFLOPS, but H100 SXM is slightly off (some 80 points higher than the spec) - most likely when its theoretical specs were calculated a lower boost clock speed was used. We can reverse engineer what it was using the spec TFLOPS: $989 / (512 * 2 * 528 / 10^{12}) / 10^{12} = 1829.20$. Indeed some Internet articles publish 1830Mhz as the actual boost clock speed of H100 SXM.

For AMD @ BF16:

Accelerator	Boost Clock	FMAs ops per Tensor Core per clock cycle	Tensor Cores	Spec TFLOPS
MI300X	2100Mhz	256	1216	1307

Let's calculate ourselves as before:

- $2100*10^{12} * 256 * 2 * 1216 / 10^{12} = 1307.4$ TFLOPS - matches the published spec, even though most of the time you will see the rounded down 1300 TFLOPS in the literature.

For Intel @ BF16:

Intel Gaudi uses MMEs to do BF16 `matmul`

Accelerator	Boost Clock	FMAs ops per MME per clock cycle	MMEs	Spec TFLOPS
Gaudi 2	1650Mhz	256*256	2	432
Gaudi 3	1600Mhz	256*256	8	1677

Let's calculate ourselves as before:

- Gaudi 2: $1650*10^{**6} * 256*256 * 2 * 2 / 10^{**12} = 432.5$ TFLOPS - matches the published spec
- Gaudi 3: $1600*10^{**6} * 256*256 * 2 * 8 / 10^{**12} = 1677$ TFLOPS - note that this doesn't match the published spec in the whitepaper (1835 TFLOPS), because in order to have 1835 TFLOPS the clock has to be 1750Mhz. i.e. the current incarnation of Gaudi3 is running at 1600Mhz.

It should become obvious now that if your accelerator runs at a lower boost clock than the spec (e.g. overheating that leads to accelerator throttling) the expected TFLOPS will be lower than advertised.

To check the actual clock speed when your accelerator is under load:

- NVIDIA: `nvidia-settings -q GPUCurrentClockFreqs`
- AMD: `rocm-smi -g` for actual and `amd-smi metric --clock` for theoretical
- Intel: `hl-smi -display CLOCK`

TFLOPS comparison table

Let's look at the supported [dtypes](#) and the corresponding theoretical peak TFLOPS specs across the high end accelerators (w/o sparsity). Sorted by the bf16 column.

Accelerator \ TFLOPS	fp32	tf32	fp16	bf16	fp8	int8	fp6	fp4	Notes
NVIDIA GB200 SXM	??	1250.0	2500	2500	5000	5000	5000	10000	2
AMD MI555X	??	??	2300	2300	4600	4600	9200	9200	
NVIDIA B200 SXM	??	1125.0	2250	2250	4500	4500	4500	9000	
Intel Gaudi3	229	459	459	1677	1677	V	X	X	1,8
AMD MI325X	163.4	653.7	1300	1300	2600	2600	X	X	7
AMD MI300X	163.4	653.7	1300	1300	2600	2600	X	X	
NVIDIA H200 SXM	67.0	494.5	989	989	1979	1979	X	X	4
NVIDIA H100 SXM	67.0	494.5	989	989	1979	1979	X	X	3
NVIDIA GH200 SXM	67.0	494.5	989	989	1979	1979	X	X	6
NVIDIA H100 PCIe	51.0	378.0	756	756	1513	1513	X	X	
Google TPU v5p	X	X	X	459	X	918	X	X	
Intel Gaudi2	V	V	V	432	865	V	X	X	1
AMD MI250X	47.9	X	383	383	X	383	X	X	
NVIDIA L40S	91.6	183.0	362	362	733	733	X	X	
AMD MI250	45.3	X	362	362	X	362	X	X	
NVIDIA A100 SXM	19.5	156.0	312	312	X	624	X	X	
NVIDIA A100 PCIe	19.5	156.0	312	312	X	624	X	X	5
Google TPU v4	X	X	X	275	X	275	X	X	
Google TPU v5e	X	X	X	197	X	394	X	X	
NVIDIA B300 SXM	??								

Row-specific notes:

1. Intel Gaudi2 and 3 only have partial TFLOPS [specs](#) published, but it does support FP32, TF32, BF16, FP16 & FP8, INT8 and INT16. These numbers are for MME (Matrix) compute.
2. Since GB200 is 2x B200 chips the table includes TFLOPS per chip for a fair comparison - you'd 2x it for the real GB200 - it also seems to run the B200 chips a bit faster so higher specs than standalone B200. This also means that instead of your typical 8-GPU node, with GB200 you will get a 4-GPU node instead (but it'd be the equivalent of 8x B200 w/ an additional ~10% faster compute).
3. I didn't include **NVIDIA H100 dual NVL** as it's, well, 2x GPUs - so it won't be fair - it's the same FLOPS as H100 but 2x everything, plus it has a bit more memory (94GB per chip, as compared to 80GB H100) and the memory is a bit faster.
4. H200 is the same as H100 but has 141GB vs 80GB of HBM memory, and its memory is faster, HBMe@4.8TBps vs HBM@3.35TBps - so basically H200 solves the compute efficiency issues of H100.
5. Oddly NVIDIA A100 PCIe and SXM revisions [spec](#) are reported to have the same TFLOPS, which is odd considering the SXM version uses 30% more power and uses a 5% faster HBM.
6. GH200 - same note as GB200 - this is 2 chips, so the table includes specs per chip w/o sparsity.
7. MI325X is the same compute as MI300X, but has more memory and more power (more efficient compute).
8. Gaudi3 as of this writing is running at 1600Mhz (MME) and not the planned 1750Mhz, therefore its BF16 TFLOPS are 1677 and not 1835 as per whitepaper spec. Same goes for fp8 which runs at the same TFLOPS as BF16.

General notes:

- int8 is measured in TeraOperations as it's not a floating operation.
- if you find numbers that are double of the above - it usually means with sparsity (which at the moment almost nobody can benefit from as our matrices are dense).
- when looking at specs be very careful at which numbers you're reading - many vendors often publish TFLOPS with sparsity, as they are ~2x bigger, but if they even indicate this they often do it in small print. I had to ask NVIDIA to add a note to their H100 spec that those numbers were w/ sparsity as they originally didn't mention this important technical fact. And 99% of the time as of this writing you will be not using sparsity and thus the actual theoretical TFLOPs that you care for most of the time are w/o sparsity (i.e. the table above).
- also beware that if accelerator A publishes a higher TFLOPS than accelerator B, it doesn't mean A is faster. These are theoretical numbers which not only can never be achieved in practice - the actual TFLOPS efficiency (HFU) can vary a lot from vendor to vendor or even for the same vendor's different accelerator architectures.

Maximum Achievable FLOPS

The problem with the advertised theoretical peak FLOPS is that they are **very** theoretical and can't be achieved in practice even if all the perfect conditions have been provided. Each accelerator has its own realistic FLOPS which is not advertised and there are anecdotal community reports that do their best to find the actual best value, but I'm yet to find any official reports.

If you find solid reports (papers?) showing the actual TFLOPS one can expect from one or more of the high end accelerators discussed in this chapter please kindly submit a PR with this information. The key is to have a reference to a source that the reader can validate the proposed information with.

To provide a numerical sense to what I'm talking about let's take an A100 with its 312 TFLOPS bf16 peak performance in the specs of this card. Until the invent of FlashAttention it was known that 150TFLOPS was close to the highest one could get for fp16/bf16 mixed precision training regime. And with FlashAttention it's around 180+TFLOPS. This is, of course, measured for training LLMs where the network and IO are involved which create additional overheads. So here the

maximum achievable peak performance probably lays somewhere between 200 and 300 TFLOPS.

You could measure the the actual achievable peak TFLOPS by doing a perfectly aligned max-size matrices `matmul` measured on a single accelerator. You can use [Maximum Achievable Matmul FLOPS Finder](#) to reproduce the results. But, of course, this will only tell you how well your given accelerator and its software stack do `matmul` - depending on the workload this might be all you need to know, or not.

MAMF stands for [Maximum Achievable Matmul FLOPS](#), which is a term coined by yours truly. It is very practical for those who do performance optimization work.

Maximum Achievable Matmul FLOPS comparison table

The following measurements are for `matmul` with BF16 and FP8 inputs (no sparsity) TFLOPS (see above for what MAMF means). Using a mean of 100 iterations after 50 warmup iterations for each shape. Sorted by accelerator efficiency:

BF16:

Accelerator	MAMF	Theory	Efficiency	Best Shape MxNxK	torch ver	Notes
NVIDIA A100 SXM	271.2	312	86.9%	1024x10240x5120	2.6.0+cu126	
NVIDIA GH200 SXM	828.6	989	83.6%	1024x15360x4096	2.6.0+cu126	900W 141GB HBM3e version
NVIDIA A100 PCIe	252.9	312	81.1%	2048x5120x6144	2.5.1+cu124	
NVIDIA H100 SXM	794.5	989	80.3%	2048x2048x13312	2.7.0+cu126	
AMD MI325X	784.9	1300	60.4%	13312x10240x8192	2.6.0+6.2.4	1000W, PYTORCH_TUNABLEOP_ENABLED=1
AMD MI300X	668.4	1300	51.4%	10240x15360x8192	2.5.1+6.3.42131	PYTORCH_TUNABLEOP_ENABLED=1
Intel Gaudi 2		432				
Intel Gaudi 3		1677				

FP8 (float8_e4m3fn):

Accelerator	MAMF	Theory	Efficiency	Best Shape MxNxK	torch ver	Notes
NVIDIA GH200 SXM	1535.0	1979	77.6%	1024x14336x14336	2.6.0+cu126	900W 141GB HBM3e version
NVIDIA H100 SXM	1402.6	1979	70.9%	1024x9216x14336	2.7.0+cu126	
Intel Gaudi 2		865				
Intel Gaudi 3		1677				

Accelerator	MAMF	Theory	Efficiency	Best Shape MxNxK	torch ver	Notes
AMD MI300X		2600				

The following is the older v1 version table that didn't reset the cache during the benchmark and in theory should have given higher scores - It will get removed once I re-populate the v2 tables.

Accelerator	MAMF	Theory	Efficiency	Best Shape	Notes
Intel Gaudi 2	429.3	432.0	99.4%	20224x19968x11520	Gaudi 1.15
NVIDIA A100 SXM	267.9	312.0	85.9%	6912x2048x16384	CUDA-12.1
NVIDIA GH200 SXM	821.0	989.0	83.0%	11264x1536x19712	CUDA-12.5
NVIDIA A100 PCIe	256.4	312.0	82.2%	2304x1536x5120	CUDA-12.1
NVIDIA H100 SXM	792.1	989.0	80.1%	6144x2816x17920	CUDA-12.1
Intel Gaudi 3	1288.8	1677.0	76.8%	22272x12288x7936	Gaudi 1.19
AMD MI250X	147.0	191.5	76.7%	1024x19968x14080	ROCM-6.2 / 1 GCD
AMD MI300X	781.9	1300.0	60.1%	4096x4864x10240	ROCM-6.2

Caveat emptor: these numbers were achieved by a brute-force search of a non-exhaustive sub-space of various shapes performing `matmul`. See: [Maximum Achievable Matmul TFLOPS Finder](#) using the software components available at the time of taking the measurement, so I highly recommend you re-run `mamf-finder.py` on your particular setup to get the true to your setup numbers. The numbers in this table are a rough estimation and shouldn't be used as absolute. As the software improves these numbers will improve coming closer to the theoretical spec. So ideally they ought to be re-run every 6 months or so.

Notes:

- For the full set of theoretical ones see [Theoretical accelerator TFLOPS](#)
- Efficiency is MAMF/Theory*100
- While `mean` is probably what most users are interested in, the script reports `max`, `median` and `mean` - should you want the other numbers.
- Best shape is the one detected by the script, but there could be many others with similar performance - it's listed for reproducibility
- If you get a much lower performance than the numbers in this table, check that the target hardware has an adequate cooling, if the accelerator is overheated it'd usually throttle its performance down. And, of course, the assumption here is that the power supply matches the spec. The latter is rarely a problem in data centers, but bad cooling is not unheard of.
- Which software you use can make a huge difference - e.g., with MI300X I clocked 450TFLOPS using ROCm-6.1, but as you can see there was a dramatic improvement in ROCm-6.2 where it jumped a whooping additional 300 TFLOPS up. BLAS library type/version may have a big impact as well.
- Then there are various system optimizations - e.g. in the case of MI300X disabling `numa_balancing` in the kernel settings is a must.
- AMD MI250X has 2 GCDs - so the theoretical TFLOPS needs to be halved, as a single `matmul` uses only 1 of them and 383 TFLOPS is reported for 2 GCDs.

Also it's important to understand that knowing the Maximum Achievable Matmul TFLOPS at some particular shape like $4352 \times 3840 \times 13568$ doesn't mean you can expect to get the same performance in your real application because chances are close to 0 that you will ever hit that exact shape. Instead, to know your system well, you'd run the MAMF Finder with the actual shapes your model is using during its training. This is really the key intention of this tool. Once you have that TFLOPS measurement you will have a good sense of where you can stop optimizing when you measure the actual TFLOPS reported by your training.

And to conclude this section I'd like to repeat again that the intention here is not to point fingers at which accelerator is more efficient than another, but to give a sense of what's what and how to navigate those theoretical specs and to help you understand when you need to continue optimizing your system and when to stop. So start with these notes and numbers as a starting point, then measure your own use case and use that latter measurement to gain the best outcome.

Not all accelerators are created equal

While measuring how well an accelerator performs, you need to be aware that while it gives you the ballpark performance numbers, other accelerators are likely to perform slightly differently. I have seen 5% and higher differences on an 8-gpu node.

This partially has to do with manufacturing processes, how well each accelerator is installed and much more about how equally each accelerator is cooled. For example, when air cooling is used it's very likely that the accelerators closer to the source of cooling will perform better than those further away, especially since now the hot air dissipated from one row gets blown into the next row of accelerators. Things should be better with liquid cooling.

Therefore, you want to measure the performance of all accelerators on the node and do it at the same time. For example, on NVIDIA nodes, if each benchmark measures a single accelerator, you could do:

```
CUDA_VISIBLE_DEVICES=0 ./some-benchmark
CUDA_VISIBLE_DEVICES=2 ./some-benchmark
...
CUDA_VISIBLE_DEVICES=7 ./some-benchmark
```

Now here what you want is the slowest performance as when used in an ensemble that slowest accelerator (struggler) will set the speed for all other accelerators.

If you do multi-node training then, of course, you'd want to measure them all.

So if you decide to calculate your achievable [MFU](#) (rather than theoretical one) you'd want to measure the achievable FLOPS across all participating accelerators and pick the value of the slowest accelerator. (If it really is an outlier you might want to consider replacing it as well).

Accelerator memory size and speed

The accelerators use [High Bandwidth Memory](#) (HBM) which is a 3D version of SDRAM memory. For example, A100-SXM comes with HBM2 at 1.6TBps, and H100-SXM comes with HBM3 at 3.35TBps (see the full table per accelerator below).

Here are the specs:

Type	Max data rate speed per pin (Gbps)	Stack Height	Bits per Channel	Number of dies per stack	Die capacity per stack (GBs)	Max capacity per stack (GBs)	Max data rate per stack (GBps)
HBM1	1.0	8	128	4	1	4	128
HBM2	2.4	8	128	8	1	8	307

Type	Max data rate speed per pin (Gbps)	Stack Height	Bits per Channel	Number of dies per stack	Die capacity per stack (GBs)	Max capacity per stack (GBs)	Max data rate per stack (GBps)
HBM2e	3.6	8	128	12	2	24	461
HBM3	6.4	16	64	12	2	24	819
HBM3e	9.8	16	64	16	3	48	1229
HBM4	6.4	32	64	16	4	64	1638

Notes:

- While I was researching this table I found a wide variation of the above numbers. I think it's because either there were different implementations or the specs changed several times and different publications caught different specs. The table above comes from [wikipedia](#).
- Since HBM is a stack of multiple DRAM chips, the *Stack Height* specifies how many chips are per device.

Typically the more on-device memory the accelerator has the better. At any given time usually most of the model weights aren't being used as they wait for their turn to be processed and thus large memory allows more of the model to be on the accelerator memory and immediately available for access and update. When there is not enough memory, sometimes the model has to be split across multiple accelerators, or offloaded to CPU and/or disk.

Here are the memory specs for the recent high end accelerators (some aren't GA yet), sorted by memory size, then bandwidth:

Accelerator	Memory (GBs)	Type	Peak Bandwidth (TBps)
NVIDIA B300 SXM	288	HBM3e	8.00
AMD MI355X	288	HBM3e	8.00
AMD MI325X	256	HBM3e	6.00
NVIDIA B200 SXM	192	HBM3e	8.00
AMD MI300X	192	HBM3	5.30
NVIDIA GH200 SXM (2)	141	HBM3e	4.80
NVIDIA H200 SXM	141	HBM3e	4.80
Intel Gaudi3	128	HBM2e	3.70
AMD MI250	128	HBM2e	3.28
AMD MI250X	128	HBM2e	3.28
NVIDIA GH200 SXM (1)	96	HBM3	4.00
Intel Gaudi2	96	HBM2e	2.46
Google TPU v5p	95	HBM2e	4.80
NVIDIA H100 SXM	80	HBM3	3.35

Accelerator	Memory (GBs)	Type	Peak Bandwidth (TBps)
NVIDIA A100 SXM	80	HBM2e	2.00
NVIDIA H100 PCIe	80	HBM3	2.00
NVIDIA A100 PCIe	80	HBM2e	1.94
NVIDIA L40S	48	GDDR6	0.86
Google TPU v4	32	HBM2	1.20
Google TPU v5e	16	HBM2	1.60

Notes:

- I didn't include NVIDIA H100 dual NVL as it's 2x H100 GPUs with 14GB memory extra per chip and slightly faster memory (3.9TBps vs 3.35TBps) - but it would have an unfair advantage in the above table as everything else is per-chip. (I guess AMD250 is also 2 GCDs, but they aren't very competitive anyway and will soon be displaced from this table by newer offerings)

Memory speed (bandwidth) is, of course, very important since if it's not fast enough than the compute ends up idling waiting for the data to be moved to and from the memory.

Caches

High performance cache is used for storing frequently used instructions and data. L1 is usually the smallest and fastest, then L2 is a bit larger and a bit slower and there can be an L3 cache which is even bigger and slower. All these caches massively reduce trips to HBM.

The cache size is often important for when running benchmarks - as one needs to reset the cache between each experiment.

It's somewhat difficult to show a comparison of caches on different accelerators because they use different approaches.

Columns:

- The L3 column is for optional additional caches: Some accelerators have only L1/L2 caches, yet others have additional caches - e.g. MI300X has 256MB of AMD Infinity cache which they also call Last Level Cache (LLC), and Gaudi3 can have its L2 cache used as L3 cache.
- Units can be different things in different accelerators, e.g. in AMD those would be Accelerator Complex Dies (XCD) or compute dies, for NVIDIA this is usually the SMs, for Intel these are DCOREs (Deep Learning Core).

Sorting by L2 Total, as it seems to be the cache that is in all accelerators listed here.

Accelerator	L1/Unit	L2/Unit	Units	L1 Total	L2 Total	L3 Total	Notes
Intel Gaudi3		24MB	4		96MB		2,4
NVIDIA GH100 SXM	256KB		132	33.00MB	60MB		
NVIDIA GH200 SXM	256KB		132	33.00MB	60MB		
NVIDIA H200 SXM	192KB		132	24.75MB	50MB		

Accelerator	L1/Unit	L2/Unit	Units	L1 Total	L2 Total	L3 Total	Notes
NVIDIA H100 SXM	192KB		132	24.75MB	50MB		
Intel Gaudi2					48MB		2,3
NVIDIA A100 SXM	128KB		108	20.25MB	40MB		
NVIDIA A100 PCIe	128KB		108	20.25MB	40MB		
AMD MI300X	32KB	4MB	8	0.25MB	32MB	256MB	1
AMD MI325X	32KB	4MB	8	0.25MB	32MB	256MB	1
AMD MI355X	???						
NVIDIA B200 SXM	???						
NVIDIA B300 SXM	???						

Notes:

1. AMD provides L3 AMD Infinity Cache which it also calls Last Level Cache (LLC) in the specs
2. Gaudi has a different architecture than a GPU. In Gaudi's case, the MME and TPC have private buffer that perform some of the functions of an L1 cache, called Suspension Buffers. The main function that these buffers provide is data reuse from the buffer (instead of reading the same data multiple times from L2/L3/HBM). Both Gaudi2 and Gaudi3 have the same these buffers for the TPC and MME.
3. Gaudi2 doesn't have a cache. It has scratchpad SRAM instead of a cache, meaning that software determines what goes in or out of the SRAM at any moment. There are dedicated DMA engines that software needs to program to perform all the data movement between SRAM and HBM.
4. The 96MB cache can be configured by software to be either a single L3 cache or 4 slices of 24MB L2 cache (this is at tensor-level granularity). L2 configuration is 2x faster than L3.

Clock speed

Also known as [clock rate](#) this spec tells us at which frequency the card runs. As hardware becomes faster newer generations will typically increase the clock speed.

When you read specs you're likely to see one or two specifications:

- Base clock is the minimal clock at idling accelerator
- Boost clock (or Peak Engine clock) is the guaranteed clock at heavy load - but it might be surpassed.

Often just the boost clock is specified.

These numbers are useful if you need to [calculate theoretical TFLOPS](#).

I've observed that the same accelerator may have different clock rates published in different specs, probably because not "final" versions are created equal. So always double check your specific accelerator for its actual specs.

Clock speed is in Mhz

Accelerator	Boost Clock	Notes
NVIDIA H200 SXM	1830	

Accelerator	Boost Clock	Notes
NVIDIA H100 SXM	1830	
NVIDIA A100 SXM	1410	
NVIDIA A100 PCIe	1410	
AMD MI300X	2100	
AMD MI325X	2100	
Intel Gaudi2	1650	MME=1650, TPC=1800
Intel Gaudi3	1600	MME=1600, TPC=1600
NVIDIA B200 SXM	?	
NVIDIA B300 SXM	?	
AMD MI355X	?	

Power consumption

There are three different definitions, whose only difference is which parts of the accelerator card is included in the measurement:

1. **Thermal Design Power (TDP)** is the maximum power that a subsystem is allowed to draw and also the maximum amount of heat an accelerator can generate. This measurement is just for the accelerator chip.
2. **Total Graphics Power (TGP)** is the same as TDP, but additionally includes the PCB's power, yet without cooling and LEDS (if any).
3. **Total Board Power (TBP)** is the same as TGP, but additionally includes cooling and LEDS (if any).

As typically high-end accelerators require external cooling and have no LEDS, TGP and TBP usually imply the same.

The actual power consumption in Watts will vary, depending on whether the accelerator is idle or used to compute something.

If you're a cloud compute user you normally don't care for these values because you're not paying for power consumption directly, as it's already included in your package. For those who host their own hardware these values are important because they tell you how much power and cooling you'd need to keep the hardware running without getting throttled or melting down.

These numbers are also important for knowing how much closer one can get to the theoretical TFLOPS published, the higher the TDP the more efficient the compute will be. For example, while AMD MI325X has the same theoretical compute specs as its MI300X predecessor, the former is more efficient at effective compute because its TDP is 250W higher. In other words, given 2 accelerators with the same or very similar [theoretical compute specs](#) - the one with the higher TDP will be better at sustainable compute.

Some specs report TDP, others TGP/TBP so the table has different columns depending on which measurement has been published. All measurements are in Watts:

Accelerator	TGP/TBP	TDP	Notes
NVIDIA GB300 SXM		1400	
NVIDIA B300 SXM		1300	

Accelerator	TGP/TBP	TDP	Notes
NVIDIA GB200 SXM		1200	
NVIDIA B200 SXM		1000	
AMD MI325X	1000		
Intel Gaudi3		900	
AMD MI300X	750		
NVIDIA H200 SXM		700	
NVIDIA H100 SXM		700	
Intel Gaudi2		600	
NVIDIA H200 NVL		600	
AMD MI250X		560	
NVIDIA H100 NVL		400	
NVIDIA A100 SXM		400	1
NVIDIA A100 PCIe		300	
AMD MI355X	??		

1. HGX A100-80GB custom thermal solution (CTS) SKU can support TDPs up to 500W

Additional notes:

1. Google doesn't publish power consumption specs for recent TPUs, the older ones can be found [here](#)

Cooling

This is of interest when you buy your own hardware, when you rent on the cloud the provider hopefully takes care of adequate cooling.

The only important practical understanding for cooling is that if the accelerators aren't kept cool they will throttle their compute clock and slow everything down and could even crash sometimes, albeit throttling is supposed to prevent that.

For NVIDIA GPUs to check if your GPU gets throttled down, run `nvidia-smi -q -d PERFORMANCE` - if SW Thermal Slowdown or some other entries are Active - then you are not getting the full performance of your GPU and you need to investigate better cooling.

High end accelerators for LLM/VLM workloads

Cloud accelerators

Most common accelerators that can be either rented on compute clouds or purchased:

NVIDIA:

- B200 - no official spec yet - only can be derived from the DGX spec: <https://www.nvidia.com/en-us/data-center/hgx/> (XXX: update when official specs are released)
- [H200](#) - mainly the same as H100, but with more and faster memory! Supposed to become available some time

mid-2024.

- [H100](#) - 2-3x faster than A100 (half precision), 6x faster for fp8, has been available on all Tier-1 compute clouds since Q4-2023.
- [GH200](#) - 2 chips on one card - (1) H100 w/ 96GB HBM3 or 144GB HBM3e + (2) Grace CPU w/ 624GB RAM - first units have been reported to become available. Do not confuse with H200, which is a different card.
- [L40S](#) - a powerful card that is supposed to be more than 2x cheaper than H100, and it's more powerful than A100.
- [A100](#) - huge availability, but already getting outdated. But given the much lower cost than H100 this is still a great GPU.

AMD:

- [MI250](#) ~= A100 - very few clouds have them
- [MI300X](#) ~= H100 - available mainly on Tier-2 clouds (lots of new startups)
- [MI325X](#) ~= H200 - just starting to emerge, mainly on Tier-2 clouds

Intel:

- [Gaudi2](#) somewhere between A100 and H100 theoretical TFLOPS-wise [spec](#) - available on Intel cloud. AWS has the older Gaudi1 via [DL1 instances](#). It's also available on-premises implementations via Supermicro and WiWynn.
- [Gaudi3](#), somewhat below B200 theoretical TFLOPS-wise - already available on Intel cloud - [spec](#)

Graphcore:

- [IPU](#) - available via [Paperspace](#). the latest product MK2 (C600) has only 0.9GB SRAM per card, so it's not clear how this card can do anything ML-wise - even inference of a small model won't fit its model weights - but there is something new at works at Graphcore, which I'm told we should discover soon. Here is a good explanation [of how IPU works](#).

SambaNova:

- [DataScale SN30](#)

On-premises accelerator clusters

Cerebras:

- [clusters](#)
- [systems](#) based on WaferScale Engine (WSE).

Cloud-only solutions

These can be only used via clouds:

Google

- [TPUs, specs](#) - lock-in, can't switch to another vendor like NVIDIA -> AMD

Cerebras:

- [Cloud](#)

New hardware startups

These are possible future competitors to the big boys.

They typically target inference.

- [TensTorrent, n150s/n300s specs](#)
- [d-Matrix, specs](#)

How to get the best price

Remember that the advertised prices are almost always open to negotiations as long as you're willing to buy/rent in bulk

or if renting for a 1-3 years. What you will discover is that the actual price that you end up paying could be many times less than the advertised "public" price. Some cloud providers already include the discount as you choose a longer commitment on their website, but it's always the best to negotiate directly with their sales team. In addition or instead of a \$\$-discount you could be offered some useful features/upgrades for free.

If your company has venture capital investors - it could help a lot to mention that, as then the cloud provider knows you are likely to buy more compute down the road and more likely to discount more.

Tier 2 clouds are likely to give better prices than Tier 1. Tier 1 as of this writing is AWS, OCI, Azure and GCP.

For the baseline prices it should be easy to find a few good sites that provide an up-to-date public price comparisons across clouds - just search for something like [cloud gpu pricing comparison](#). Some good starting points: [vast.ai](#) and specifically for clusters [gpulist.ai](#).

When shopping for a solution please remember that it's not enough to rent the most powerful accelerator. You also need fast [intra-node](#) and [inter-node](#) connectivity and sufficiently fast [storage](#) - without which the expensive accelerators will idle waiting for data to arrive and you could be wasting a lot money and losing time.

Accelerators in detail

NVIDIA

Abbreviations:

- CUDA: Compute Unified Device Architecture (proprietary to NVIDIA)

NVIDIA-specific key GPU characteristics:

- CUDA Cores - similar to CPU cores, but unlike CPUs that typically have 10-100 powerful cores, CUDA Cores are weaker and come in thousands and allow to perform massive general purpose computations (parallelization). Like CPU cores CUDA Cores perform a single operation in each clock cycle.
- Tensor Cores - special compute units that are designed specifically to perform fast multiplication and addition operations like matrix multiplication. These perform multiple operations in each clock cycle. They can execute extremely fast computations on low or mixed precision data types with some loss (fp16, bf16, tf32, fp8, etc.). These cores are specifically designed for ML workloads.
- Streaming Multiprocessors (SM) are clusters of CUDA Cores, Tensor Cores and other components.

For example, A100-80GB has:

- 6912 CUDA Cores
- 432 Tensor Cores (Gen 3)
- 108 Streaming Multiprocessors (SM)

H100 has:

- 16896 FP32 CUDA Cores
- 528 Tensor Cores (Gen 4)
- 132 Streaming Multiprocessors (SM)

AMD

AMD-specific key GPU characteristics:

- Stream Processors - are similar in functionality to CUDA Cores - that is these are the parallel computation units. But they aren't the same, so one can't compare 2 gpus by just comparing the number of CUDA Cores vs the number of Stream Processors.
- Compute Units - are clusters of Stream Processors and other components

for example, AMD MI250 has:

- 13,312 Stream Processors
- 208 Compute Units

Intel Gaudi

[Architecture](#)

- 24x 100 Gigabit Ethernet (RoCEv2) integrated on chip - 21 of which are used for intra-node and 3 for inter-node (so $21*8=168$ cards for intra-node (262.5GBps per GPU), and $3*8=24$ cards for inter-node (2.4Tbps between nodes))
- 96GB HBM2E memory on board w/2.45 TBps bandwidth per chip, for a total of 768GB per node

A server/node is built from 8 GPUs, which can then be expanded with racks of those servers.

There are no official TFLOPS information published (and from talking to an Intel representative they have no intention to publish any.) They publish the [following benchmarks](#) but I'm not sure how these can be used to compare this compute to other providers.

Comparison: supposedly Gaudi2 competes with NVIDIA H100

API

Which software is needed to deploy the high end GPUs?

NVIDIA

NVIDIA GPUs run on [CUDA](#)

AMD

AMD GPUs run on [ROCM](#) - note that PyTorch you can use CUDA-based software on ROCm-based GPUs! So it should be trivial to switch to the recent AMD MI250, MI300X, and other emerging ones.

Intel Gaudi

The API is via [Habana SynapseAI® SDK](#) which supports PyTorch and TensorFlow.

Useful integrations:

- [HF Optimum Habana](#) which also includes - [DeepSpeed](#) integration.

Apples-to-apples Comparison

It's very difficult to compare specs of different offerings since marketing tricks get deployed pretty much by all competitors so that one can't compare 2 sets of specs and know the actual difference.

- [MLPerf via MLCommons](#) publishes various hardware benchmarks that measure training, inference, storage and other tasks' performance. For example, here is the most recent as of this writing [training v3.0](#) and [inference v3.1](#) results.

Except I have no idea how to make use of it - it's close to impossible to make sense of or control the view. This is a great intention lost in over-engineering and not thinking about how the user will benefit from it, IMHO. For example, I don't care about CV data, I only want to quickly see the LLM rows, but I can't do it. And then the comparisons are still not apples to apples so how can you possibly make sense of which hardware is better I don't know.

Power and Cooling

It is most likely that you're renting your accelerator nodes and someone else is responsible for ensuring they function properly, but if you own the accelerators you do need to know how to supply a sufficient power and adequate cooling.

Power

Some high end consumer GPU cards have 2 and sometimes 3 PCI-E 8-Pin power sockets. Make sure you have as many independent 12V PCI-E 8-Pin cables plugged into the card as there are sockets. Do not use the 2 splits at one end of the same cable (also known as pigtail cable). That is if you have 2 sockets on the GPU, you want 2 PCI-E 8-Pin cables going from your PSU to the card and not one that has 2 PCI-E 8-Pin connectors at the end! You won't get the full performance out of your card otherwise.

Each PCI-E 8-Pin power cable needs to be plugged into a 12V rail on the PSU side and can supply up to 150W of power.

Some other cards may use a PCI-E 12-Pin connectors, and these can deliver up to 500-600W of power.

Low end cards may use 6-Pin connectors, which supply up to 75W of power.

Additionally you want the high-end PSU that has stable voltage. Some lower quality ones may not give the card the stable voltage it needs to function at its peak.

And of course the PSU needs to have enough unused Watts to power the card.

Cooling

When a GPU gets overheated it will start throttling down and will not deliver full performance and it can even shutdown if it gets too hot.

It's hard to tell the exact best temperature to strive for when a GPU is heavily loaded, but probably anything under +80C is good, but lower is better - perhaps 70-75C is an excellent range to be in. The throttling down is likely to start at around 84-90C. But other than throttling performance a prolonged very high temperature is likely to reduce the lifespan of a GPU.

Accelerator Benchmarks

Maximum Achievable Matmul FLOPS Finder

Maximum Achievable Matmul FLOPS (MAMF) Benchmark: [mamf-finder.py](#)

For a detailed discussion and the numbers for various accelerators see [Maximum Achievable FLOPS](#).

While some accelerator manufacturers publish the theoretical TFLOPS these usually can't be reached. As a result of this when we try to optimize our software we have no realistic performance bar to compare ourselves to. The Model FLOPS Utilization (MFU) metric measures TFLOPS achieved against theoretical TFLOPS. Usually when one scores around 50% MFU it's considered a win. But this gives us no indication how far are we from the real achievable throughput.

This benchmark scans various large shapes of matmul and reports the highest achievable TFLOPS it registered. As transformers training and partially inference workloads are dominated by large matmul operations it's safe to use the best matmul TFLOPS one can measure on each accelerator as a rough estimation that this is the Maximum Achievable Matmul FLOPS (MAMF). Now instead of the previously used MFU, one can use Model Achievable Matmul FLOPS Utilization (MAMFU).

Therefore now you can compare the TFLOPS you measured for your training or inference against a realistic number. As you will now be much closer to 100% it'll be much easier to know when to stop optimizing.

Currently supported high end architectures:

- NVIDIA: V100, A100, H100, ...
- AMD: MI250, MI300X, MI325X, ...
- Intel Gaudi2/3

Fairness notes:

- if you can find a better and more efficient way to detect the best matmul TFLOPS by approaching each new accelerator as a black box, please kindly send a PR with the improvement including the generated log file.
- also if you know that this benchmark should be run under special conditions to show the best results, such as some kernel settings or similar, please submit a PR to add such special instructions. For example, for AMD MI300X I'm being told disabling the numa_balancing is supposed to help.

Architecture specific notes:

Follow the special setup instructions before running the benchmark to achieve the best results:

MI300x, MI325X, etc.:

1. Turn numa_balancing off for better performance:

```
sudo sh -c 'echo 0 > /proc/sys/kernel/numa_balancing'
```

2. Enable:

```
export PYTORCH_TUNABLEOP_ENABLED=1
```

This will make the first iteration very slow, while it's searching for the best GEMM algorithm in the BLAS libraries for each `matmul` shape it encounters, but subsequent operations are likely to be significantly faster than the baseline. See

[Accelerating models on ROCm using PyTorch TunableOp](#) (requires `torch>=2.3`) doc.

Examples of usage

In the ranges below K is the reduction dimension so that $(M \times K) \times (K \times N) = (M \times N)$ and we print the $M \times K \times N$ shape for the best measured TFLOPS.

Also by default we use 50 warmup and 100 measured iterations for each shape and then fastest result is picked (not the average). You can change the number of iterations via the args `--num_warmup_iterations` and `--num_iterations` correspondingly.

You can specify the data type via `--dtype` argument, it has to be one of the valid `torch` dtypes - e.g., `float8_e4m3fn`, `float16`, `bfloat16`, `float32`, etc. If not specified `bfloat16` is used.

Here we do `torch.mm(MxK, KxN) -> MxN`

1. A quick run (under 1min) - should give around 80-90% of the maximum achievable result - good for a quick try out, but not enough to get a high measurement.

```
./mamf-finder.py --m_range 0 20480 256 --n 4096 --k 4096 --output_file=$(date +"%Y-%m-%d-%H:%M:%S").txt
```

2. A more exhaustive search (15-30min) - but you can Ctrl-C it when it run long enough and get the best result so far:

```
./mamf-finder.py --m_range 0 16384 1024 --n_range 0 16384 1024 --k_range 0 16384 1024  
--output_file=$(date +"%Y-%m-%d-%H:%M:%S").txt
```

Feel free to make the steps smaller from 1024 to 512 or 256 - but it'd 8x or 64x the run time correspondingly. 1k steps should cover the different shape ranges well and fast.

3. A super long exhaustive search (may take many hours/days) - but you can Ctrl-C it when it run long enough and get the best result so far:

```
./mamf-finder.py --m_range 0 20480 256 --n_range 0 20480 256 --k_range 0 20480 256 --output_file=$(date  
+"%Y-%m-%d-%H:%M:%S").txt
```

4. If you want to measure a specific shape that is used by your training, use the exact shape, instead of the range, so let's say you wanted to measure 1024x1024x1024 - you'd run:

```
./mamf-finder.py --m 1024 --n 1024 --k 1024 --output_file=$(date +"%Y-%m-%d-%H:%M:%S").txt
```

5. Accelerator specific range seeking suggestions

But then it appears that different accelerators have different ranges of shapes that lead to best TFLOPS, thus it's difficult to suggest a range that will work well for all of them - instead here are some suggestions based on experiments and suggestions from contributors:

- A100 + MI300X

```
./mamf-finder.py --m_range 0 5376 256 --n_range 0 5376 256 --k_range 0 5376 256 --output_file=$(date  
+"%Y-%m-%d-%H:%M:%S").txt
```

- H100

```
./mamf-finder.py --m_range 0 20480 256 --n_range 0 20480 256 --k_range 0 20480 256 --output_file=$(date +"%Y-%m-%d-%H:%M:%S").txt
```

To understand better which shapes give the highest matmul FLOPS for a particular accelerator, see [Vector and matrix size divisibility](#).

Results

The measurements that I have gathered so far can be found at [Maximum Achievable Matmul FLOPS comparison table](#). When I had access to a particular accelerator I run the benchmarks myself, when I didn't it was the kind contributors who invested their time to get these numbers. So I'm very grateful to [those](#).

How to benchmark accelerators

CUDA benchmarks

There are a few excellent detailed write ups on how to perform CUDA benchmarks:

1. [How to Accurately Time CUDA Kernels in Pytorch](#)
2. [How to Benchmark Code on CUDA Devices?](#) - this one is different from (1) in that it suggests to set both GPU and Memory clocks, whereas (1) only locks the GPU clock.

You can see these instructions applied in [mamf-finder.py](#) (other than clock locking)

Troubleshooting NVIDIA GPUs

Glossary

- DBE: Double Bit ECC Error
- DCGM: (NVIDIA) Data Center GPU Manager
- ECC: Error-Correcting Code
- FB: Frame Buffer
- SBE: Single Bit ECC Error
- SDC: Silent Data Corruption

Xid Errors

No hardware is perfect, sometimes due to the manufacturing problems or due to tear and wear (especially because of exposure to high heat), GPUs are likely to encounter various hardware issues. A lot of these issues get corrected automatically without needing to really understand what's going on. If the application continues running usually there is nothing to worry about. If the application crashes due to a hardware issue it's important to understand why this is so and how to act on it.

A normal user who uses a handful of GPUs is likely to never need to understand GPU-related hardware issues, but if you come anywhere close to massive ML training where you are likely to use hundreds to thousands of GPUs it's certain that you'd want to understand about different hardware issues.

In your system logs you are likely to see occasionally Xid Errors like:

```
NVRM: Xid (PCI:0000:10:1c): 63, pid=1896, Row Remapper: New row marked for remapping, reset gpu to activate.
```

To get those logs one of the following ways should work:

```
sudo grep Xid /var/log/syslog
sudo dmesg -T | grep Xid
```

Typically, as long as the training doesn't crash, these errors often indicate issues that automatically get corrected by the hardware.

The full list of Xid Errors and their interpretation can be found [here](#).

You can run `nvidia-smi -q` and see if there are any error counts reported. For example, in this case of Xid 63, you will see something like:

```
Timestamp : Wed Jun  7 19:32:16 2023
Driver Version : 510.73.08
CUDA Version : 11.6

Attached GPUs : 8
GPU 00000000:10:1C.0
```

```

Product Name          : NVIDIA A100-SXM4-80GB
[...]
ECC Errors
  Volatile
    SRAM Correctable      : 0
    SRAM Uncorrectable    : 0
    DRAM Correctable      : 177
    DRAM Uncorrectable    : 0
  Aggregate
    SRAM Correctable      : 0
    SRAM Uncorrectable    : 0
    DRAM Correctable      : 177
    DRAM Uncorrectable    : 0
  Retired Pages
    Single Bit ECC        : N/A
    Double Bit ECC         : N/A
    Pending Page Blacklist : N/A
  Remapped Rows
    Correctable Error      : 1
    Uncorrectable Error    : 0
    Pending                 : Yes
    Remapping Failure Occurred : No
  Bank Remap Availability Histogram
    Max                     : 639 bank(s)
    High                    : 1 bank(s)
    Partial                 : 0 bank(s)
    Low                     : 0 bank(s)
    None                    : 0 bank(s)
[...]

```

Here we can see that Xid 63 corresponds to:

```
ECC page retirement or row remapping recording event
```

which may have 3 causes: HW Error / Driver Error / FrameBuffer (FB) Corruption

This error means that one of the memory rows is malfunctioning and that upon either reboot and/or a gpu reset one of the 640 spare memory rows (in A100) will be used to replace the bad row. Therefore we see in the report above that only 639 banks remain (out of 640).

The Volatile section of the `ECC Errors` report above refers to the errors recorded since last reboot/GPU reset. The Aggregate section records the same error since the GPU was first used.

Now, there are 2 types of errors - Correctable and Uncorrectable. The correctable one is a Single Bit ECC Error (SBE) where despite memory being faulty the driver can still recover the correct value. The uncorrectable one is where more than one bit is faulty and it's called Double Bit ECC Error (DBE). Typically, the driver will retire whole memory pages if 1 DBE or 2 SBE errors occur at the same memory address. For full information see [this document](#)

A correctable error will not impact the application, a non-correctable one will crash the application. The memory page containing the uncorrectable ECC error will be blacklisted and not accessible until the GPU is reset.

If there are pages scheduled to be retired you will see something like this in the output of `nvidia-smi -q`:

```
Retired pages
  Single Bit ECC      : 2
  Double Bit ECC      : 0
  Pending Page Blacklist : Yes
```

Each retired page decreases the total memory available to applications. But the maximum amount of pages retired amounts to only 4MB in total, so it doesn't reduce the total available GPU memory by much.

To dive even deeper into the GPU debugging, please refer to [this document](#) - it includes a useful triage chart which helps to determine when to RMA GPUs. This document has additional information about Xid 63-like errors

For example it suggests:

If associated with XID 94, the application that encountered the error needs to be restarted. All other applications on the system can keep running as is until there is a convenient time to reboot for row remapping to activate. See below for guidelines on when to RMA GPUs based on row remapping failures.

If after a reboot the same condition occur for the same memory address, it means that memory remapping has failed and Xid 64 will be emitted again. If this continues it means you have a hardware issue that can't be auto-corrected and the GPU needs to RMA'ed.

At other times you may get Xid 63 or 64 and the application will crash. Which usually will generate additional Xid errors, but most of the time it means that the error was uncorrectable (i.e. it was a DBE sort of an error and then it'll be Xid 48).

As mentioned earlier to reset a GPU you can either simply reboot the machine, or run:

```
nvidia-smi -r -i gpu_id
```

where `gpu_id` is the sequential number of the gpu you want to reset, e.g. `0` for the first GPU. Without `-i` all GPUs will be reset.

uncorrectable ECC error encountered

If you get an error:

```
CUDA error: uncorrectable ECC error encountered
```

as in the previous section, checking the output of `nvidia-smi -q` this time for `ECC Errors` entries will tell which GPU is the problematic one. But if you need to do a quick check in order to recycle a node if it has at least one GPU with this issue, you can just do this:

```
$ nvidia-smi -q | grep -i correctable | grep -v 0
  SRAM Uncorrectable      : 1
  SRAM Uncorrectable      : 5
```

On a good node, this should return nothing, as all counters should be 0. But in the example above we had one broken GPU - there were two entries because the full record was:

```
ECC Errors
  Volatile
    SRAM Correctable      : 0
    SRAM Uncorrectable    : 1
    DRAM Correctable      : 0
    DRAM Uncorrectable    : 0
  Aggregate
    SRAM Correctable      : 0
    SRAM Uncorrectable    : 5
    DRAM Correctable      : 0
    DRAM Uncorrectable    : 0
```

The first entry is for `Volatile` (errors counted since the last time the GPU driver reload) and the second is for `Aggregate` (total errors counter for the whole life time of the GPU). In this example we see a `Volatile` counter for SRAM Uncorrectable errors to be 1 and for the life-time counter it's 5 - that is this is not the first time the GPU runs into this problem.

This typically would correspond to Xid 94 error (see: [Xid Errors](#), most likely w/o Xid 48).

To overcome this issue as in the previous section, reset the problematic GPU:

```
nvidia-smi -r -i gpu_id
```

Rebooting the machine will have the same effect.

Now when it comes to Aggregate SRAM Uncorrectable errors, if you have more than 4, that's usually a reason to RMA that GPU.

Running diagnostics

If you suspect one or more NVIDIA GPUs are broken on a given node, `dcgmi` is a great tool to quickly find any bad GPUs.

NVIDIA® Data Center GPU Manager (DCGM) is documented [here](#) and can be downloaded from [here](#).

Here is an example slurm script that will run very in-depth diagnostics (-r 3), which will take about 10 minutes to complete on an 8-GPU node:

```
$ cat dcgmi-1n.slurm
#!/bin/bash
#SBATCH --job-name=dcgmi-1n
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=96
#SBATCH --gres=gpu:8
#SBATCH --exclusive
#SBATCH --output=%x-%j.out

set -x -e
```

```
echo "START TIME: $(date)"  
srun --output=%x-%j-%N.out dcgmi diag -r 3  
echo "END TIME: $(date)"
```

Now to run it on specific nodes of choice:

```
sbatch --nodelist=node-115 dcgmi-1n.slurm  
sbatch --nodelist=node-151 dcgmi-1n.slurm  
sbatch --nodelist=node-170 dcgmi-1n.slurm
```

edit the nodelist argument to point to the node name to run.

If the node is drained or downed and you can't launch a slurm job using this node, just `ssh` into the node and run the command directly on the node:

```
dcgmi diag -r 3
```

If the diagnostics didn't find any issue, but the application still fails to work, re-run the diagnostics with level 4, which will now take more than 1 hour to complete:

```
dcgmi diag -r 4
```

footnote: apparently silent data corruptions (SDC) can only be detected with `dcgmi diag -r 4` and even then some might be missed. This problem happens occasionally and you may not even be aware that your GPU is messing up the `matmul` at times. I'm pretty sure we had this happen to us, as we were getting weird glitches during training and I spent many days with the NVIDIA team diagnosing the problem, but we failed to do so - eventually the problem disappeared probably because the bad GPU(s) got replaced due to reported failures.

For example, if you run into a repeating Xid 64 error it's likely that the diagnostics report will include:

Diagnostic	Result
Deployment	GPU 3 has uncorrectable memory errors and row remappings are pending

so you now know to RMA that problematic GPU, if remapping fails.

But, actually, I found that most of the time `-r 2` already detects faulty GPUs. And it takes just a few minutes to complete. Here is an example of the `-r 2` output on a faulty node:

GPU Memory	Pass - GPUs: 1, 2, 3, 4, 5, 6, 7
	Fail - GPU: 0
Warning	GPU 0 Thermal violations totaling 13.3 second

```

|           | s started at 9.7 seconds into the test for GP  |
|           | U 0 Verify that the cooling on this machine is  |
|           | s functional, including external, thermal mat  |
|           | erial interface, fans, and any other component  |
|           | ts.

```

The `dcmi` tool contains various other levels of diagnostics, some of which complete in a matter of a few minutes and can be run as a quick diagnostic in the epilogue of SLURM jobs to ensure that the node is ready to work for the next SLURM job, rather than discovering that after the user started their job and it crashed.

When filing an RMA report you will be asked to run `nvidia-bug-report` script, the output of which you will need to submit with the RMA request.

I usually save the log as well for posterity using one of:

```

dcmi diag -r 2 | tee -a dcmi-r2-`hostname`.txt
dcmi diag -r 3 | tee -a dcmi-r3-`hostname`.txt
dcmi diag -r 4 | tee -a dcmi-r4-`hostname`.txt

```

How to get the VBIOS info

GPU VBIOS version might be important when researching issues. Let's add the name and bus id to the query, we get:

```

$ nvidia-smi --query-gpu=gpu_name,gpu_bus_id,vbios_version --format=csv
name, pci.bus_id, vbios_version
NVIDIA H100 80GB HBM3, 00000000:04:00.0, 96.00.89.00.01
[...]
NVIDIA H100 80GB HBM3, 00000000:8B:00.0, 96.00.89.00.01

```

Hint: to query for dozens of other things, run:

```
nvidia-smi --help-query-gpu
```

How to check if your GPU's PCIe generation is supported

Check the PCIe bandwidth reports from the system's boot messages:

```

$ sudo dmesg | grep -i 'limited by'
[ 10.735323] pci 0000:04:00.0: 252.048 Gb/s available PCIe bandwidth, limited by 16.0 GT/s PCIe x16 link
at 0000:01:00.0 (capable of 504.112 Gb/s with 32.0 GT/s PCIe x16 link)
[...]
[ 13.301989] pci 0000:8b:00.0: 252.048 Gb/s available PCIe bandwidth, limited by 16.0 GT/s PCIe x16 link
at 0000:87:00.0 (capable of 504.112 Gb/s with 32.0 GT/s PCIe x16 link)

```

In this example, as PCIe 5 spec is 504Gbps, you can see that on this node only half of the possible bandwidth is usable,

because the PCIe switch is gen4. For PCIe specs see [this](#).

Since most likely you have [NVLink](#) connecting the GPUs to each other, this shouldn't matter for GPU to GPU comms, but it'd slow down any data movement between the GPU and the host, as the data speed is limited by the speed of the slowest link.

How to check error counters of NVLink links

If you're concerned your NVLink malfunctions you can check its error counters:

```
$ nvidia-smi nvlink -e
GPU 0: NVIDIA H100 80GB HBM3 (UUID: GPU-abcdefab-cdef-abdc-abcd-ababababab)
    Link 0: Replay Errors: 0
    Link 0: Recovery Errors: 0
    Link 0: CRC Errors: 0

    Link 1: Replay Errors: 0
    Link 1: Recovery Errors: 0
    Link 1: CRC Errors: 0

    [...]

    Link 17: Replay Errors: 0
    Link 17: Recovery Errors: 0
    Link 17: CRC Errors: 0
```

Another useful command is:

```
$ nvidia-smi nvlink --status
GPU 0: NVIDIA H100 80GB HBM3 (UUID: GPU-abcdefab-cdef-abdc-abcd-ababababab)
    Link 0: 26.562 GB/s
    [...]
    Link 17: 26.562 GB/s
```

this one tells you the current speed of each link

Run `nvidia-smi nvlink -h` to discover more features (reporting, resetting counters, etc.).

How to detect if a node is missing GPUs

If you got a new VM, there are odd cases where there is less than expected number of GPUs. Here is how you can quickly test you have got 8 of them:

```
cat << 'EOT' >> test-gpu-count.sh
#!/bin/bash

set -e
```

```
# test the node has 8 gpus
test $(nvidia-smi -q | grep UUID | wc -l) != 8 && echo "broken node: less than 8 gpus" && false
EOT
```

and then:

```
bash test-gpu-count.sh
```

How to detect if you get the same broken node again and again

This is mostly relevant to cloud users who rent GPU nodes.

So you launched a new virtual machine and discovered it has one or more broken NVIDIA GPUs. You discarded it and launched a new and the GPUs are broken again.

Chances are that you're getting the same node with the same broken GPUs. Here is how you can know that.

Before discarding the current node, run and log:

```
$ nvidia-smi -q | grep UUID
GPU UUID : GPU-2b416d09-4537-ecc1-54fd-c6c83a764be9
GPU UUID : GPU-0309d0d1-8620-43a3-83d2-95074e75ec9e
GPU UUID : GPU-4fa60d47-b408-6119-cf63-a1f12c6f7673
GPU UUID : GPU-fc069a82-26d4-4b9b-d826-018bc040c5a2
GPU UUID : GPU-187e8e75-34d1-f8c7-1708-4feb35482ae0
GPU UUID : GPU-43bfd251-aad8-6e5e-ee31-308e4292bef3
GPU UUID : GPU-213fa750-652a-6cf6-5295-26b38cb139fb
GPU UUID : GPU-52c408aa-3982-baa3-f83d-27d047dd7653
```

These UUIDs are unique to each GPU.

When you then re-created your VM, run this command again - if the UUIDs are the same - you know you have the same broken GPUs.

To automate this process so that you always have this data as it'd be too late if you already rebooted the VM, add somewhere in your startup process this:

```
nvidia-smi -q | grep UUID > nvidia-uuids.$(hostname).$(date '+%Y-%m-%d-%H:%M').txt
```

You'd want to save the log file on some persistent filesystem for it to survive reboot. If you do not have one make it local and immediately copy to the cloud. That way it'll always be there when you need it.

Sometimes just rebooting the node will get new hardware. In some situations you get new hardware on almost every reboot, in other situations this doesn't happen. And this behavior may change from one provider to another.

If you keep on getting the same broken node - one trick to overcoming this is allocating a new VM, while holding the broken VM running and when the new VM is running - discarding the broken one. That way you will surely get new GPUs - except there is no guarantee they won't be broken as well. If the use case fits consider getting a static cluster where it's much easier to keep the good hardware.

This method is extra-crucial for when GPUs don't fail right away but after some use so it is non-trivial to see that there is a problem. Even if you reported this node to the cloud provider the technician may not notice the problem right away and put the bad node back into circulation. So if you're not using a static cluster and tend to get random VMs on demand you may want to keep a log of bad UUIDs and know you have got a lemon immediately and not 10 hours into the node's use.

Cloud providers usually have a mechanism of reporting bad nodes. Therefore other than discarding a bad node, it'd help yourself and other users to report bad nodes. Since most of the time users just discard the bad nodes, the next user is going to get them. I have seen users getting a very high percentage of bad nodes in some situations.

How to get the real GPU utilization metrics

As explained [here](#) the `Volatile GPU-Util` column in the `nvidia-smi` output isn't really telling you the GPU Utilization. What it's telling you is the percentage of time during which one or more kernels were executing on the GPU. It's not telling you whether a single SM is being used or all of them. So even if you run a tiny `matmul` all the time, you may get a very high `gpu util`, while most of the GPU isn't doing anything.

footnote: I have seen GPU util column showing 100% on all gpus when one GPU would stop responding and then whole machinery was blocked waiting for that gpu to respond. Which is how I discovered that it couldn't be showing the real GPU utilization in the first place.

What you want to measure instead is GPU's utilization of the available capacity, otherwise known as "saturation". Alas, this information isn't provided by `nvidia-smi`. In order to get this information you need to install [dcgm-exporter](#) (which in turn currently requires a recent golang and DCGM (datacenter-gpu-manager) and a root access).

Please note that this tool works only high-end data center NVIDIA GPUs, so if you have a consumer level GPU it won't work.

After installing the prerequisites I built the tool:

```
git clone https://github.com/NVIDIA/dcgm-exporter.git
cd dcgm-exporter
make binary
```

And then I was able to get the "real" utilization metrics described in the article with this `dcgm-exporter` config file:

```
$ cat << EOT > dcp-metrics-custom.csv
DCGM_FI_PROF_SM_OCCUPANCY, gauge, The ratio of number of warps resident on an SM.
DCGM_FI_PROF_PIPE_TENSOR_ACTIVE, gauge, Ratio of cycles the tensor (HMMA) pipe is active.
DCGM_FI_PROF_PIPE_FP16_ACTIVE, gauge, Ratio of cycles the fp16 pipes are active.
DCGM_FI_PROF_PIPE_FP32_ACTIVE, gauge, Ratio of cycles the fp32 pipes are active.
EOT
```

Then I launched the daemon (root is required):

```
$ sudo cmd/dcgm-exporter/dcgm-exporter -c 500 -f dcp-metrics-custom.csv
[...]
INFO[0000] Starting webserver
INFO[0000] Listening on address="[:]:9400"
```

```
-c 500 refreshes every 0.5sec
```

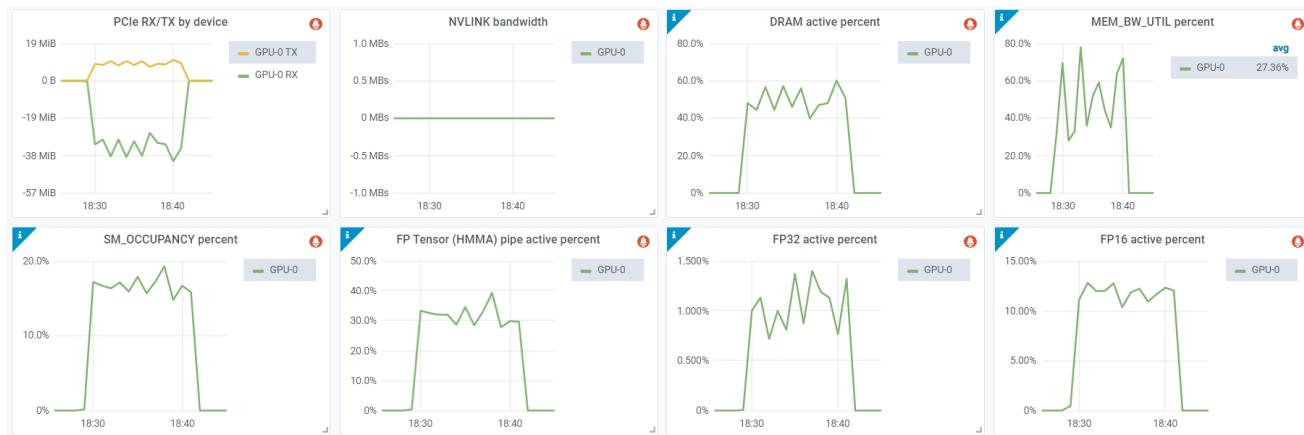
and now I was able poll it via:

```
watch -n 0.5 "curl http://localhost:9400/metrics"
```

by running it in one console, and launching a GPU workload in another console. The last column of the output is the utilization of these metrics (where $1.0 == 100\%$).

`etc/dcp-metrics-included.csv` from the repo contains all the available metrics, so you can add more metrics.

This is a quick way of doing that, but the intention is to use it with [Prometheus](#) which will give you nice charts. E.g. the article included an example where you can see the SM occupancy, Tensor core, FP16 and FP32 Core utilization in the second row of the charts:



[\(source\)](#)

For completion here is an example from the same article showing a 100% gpu util with a CUDA kernel that is doing absolutely nothing compute-wise other than occupying a single Streaming Multiprocessor (SM):

```
$ cat << EOT > 1_sm_kernel.cu
__global__ void simple_kernel() {
    while (true) {}
}

int main() {
    simple_kernel<<<1, 1>>>();
    cudaDeviceSynchronize();
}
EOT
```

Let's compile it:

```
nvcc 1_sm_kernel.cu -o 1_sm_kernel
```

And now run it in console A:

```
$ ./1_sm_kernel
```

and in console B:

```
$ nvidia-smi
Tue Oct  8 09:49:34 2024
+-----+-----+-----+
| NVIDIA-SMI 550.90.12      Driver Version: 550.90.12      CUDA Version: 12.4      |
+-----+-----+-----+
| GPU  Name                  Persistence-M | Bus-Id      Disp.A | Volatile Uncorr. ECC | | |
| Fan  Temp     Perf          Pwr:Usage/Cap |          Memory-Usage | GPU-Util  Compute M. |
|          |                                         |              |          |          MIG M. |
+=====+=====+=====+=====+=====+=====+=====+
|  0  NVIDIA A100 80GB PCIe      Off | 00000000:01:00.0 Off |          0 | |
| N/A  32C     P0      69W / 300W | 437MiB / 81920MiB | 100%     Default |
|          |                                         |          |          Disabled |
+-----+-----+-----+
```

You can see the 100% GPU-Util. So here 1 SM is used whereas A100-80GB PCIe has 132 SMs! And it's not even doing any compute as it just runs an infinite loop of doing nothing.

Troubleshooting AMD GPUs

XXX: this is very early - collecting various tools/notes

As most of us are well familiar with NVIDIA tools, I will try to provide the mapping where possible to the familiar tools.

Tools

ROCR_VISIBLE_DEVICES

To select a specific gpu (CUDA_VISIBLE_DEVICES equivalent):

```
ROCR_VISIBLE_DEVICES=0,1 python my-program.py
```

rocm-smi

rocm-smi (nvidia-smi equivalent) shows a condensed state of all the ROCm accelerators.

For example here is an 8xMI300X node:

```
$ rocm-smi
=====
          ROCm System Management Interface
=====
          Concise Info
=====
Device  [Model : Revision]      Temp      Power      Partitions      SCLK      MCLK      Fan      Perf      PwrCap
VRAM%  GPU%
Name (20 chars)      (Junction)  (Socket)  (Mem, Compute)
=====
0      [0x74a1 : 0x00]      45.0°C      173.0W      NPS1, SPX      132Mhz  900Mhz  0%  auto  750.0W
0%  0%
      AMD Instinct MI300X
1      [0x74a1 : 0x00]      41.0°C      179.0W      NPS1, SPX      132Mhz  900Mhz  0%  auto  750.0W
0%  0%
      AMD Instinct MI300X
2      [0x74a1 : 0x00]      47.0°C      180.0W      NPS1, SPX      131Mhz  900Mhz  0%  auto  750.0W
0%  0%
      AMD Instinct MI300X
3      [0x74a1 : 0x00]      45.0°C      178.0W      NPS1, SPX      131Mhz  900Mhz  0%  auto  750.0W
17%  0%
      AMD Instinct MI300X
4      [0x74a1 : 0x00]      45.0°C      175.0W      NPS1, SPX      132Mhz  900Mhz  0%  auto  750.0W
0%  0%
      AMD Instinct MI300X
5      [0x74a1 : 0x00]      43.0°C      175.0W      NPS1, SPX      132Mhz  900Mhz  0%  auto  750.0W
0%  0%
      AMD Instinct MI300X
6      [0x74a1 : 0x00]      45.0°C      175.0W      NPS1, SPX      132Mhz  900Mhz  0%  auto  750.0W
```

```

0% 0%
    AMD Instinct MI300X
7      [0x74a1 : 0x00]      43.0°C      176.0W      NPS1, SPX      132Mhz  900Mhz  0%  auto  750.0W
0% 0%
    AMD Instinct MI300X
=====
===== End of ROCm SMI Log
=====
```

Oddly it shows no real memory usage - only the percentage, which isn't very practical.

A handy alias to watch updates in real time:

```
alias wr='watch -n 1 rocm-smi'
```

rocminfo

`rocminfo` (nvidia-smi -q equivalent) shows the detailed information about each accelerator.

This one shows both the CPU and the GPU information

Here is a snippet for cpu0 and gpu0 (note it starts counting the cpus as nodes 0..1, and then GPUs as nodes 2..9):

```

$ rocminfo
ROCK module is loaded
=====
HSA System Attributes
=====
Runtime Version:      1.1
System Timestamp Freq.: 1000.000000MHz
Sig. Max Wait Duration: 18446744073709551615 (0xFFFFFFFFFFFFFF) (timestamp count)
Machine Model:        LARGE
System Endianness:    LITTLE
Mwaitx:               DISABLED
DMAbuf Support:       YES

=====
HSA Agents
=====
*****
Agent 1
*****
Name:          AMD EPYC 9534 64-Core Processor
Uuid:          CPU-XX
Marketing Name: AMD EPYC 9534 64-Core Processor
Vendor Name:    CPU
Feature:        None specified
Profile:        FULL_PROFILE
Float Round Mode: NEAR
Max Queue Number: 0(0x0)
```

```

Queue Min Size:          0(0x0)
Queue Max Size:          0(0x0)
Queue Type:              MULTI
Node:                   0
Device Type:             CPU
Cache Info:
  L1:                   32768(0x8000) KB
  Chip ID:              0(0x0)
  ASIC Revision:        0(0x0)
  Cacheline Size:       64(0x40)
  Max Clock Freq. (MHz): 2450
  BDFID:                0
  Internal Node ID:    0
  Compute Unit:         128
  SIMDs per CU:         0
  Shader Engines:       0
  Shader Arrs. per Eng.: 0
  WatchPts on Addr. Ranges: 1
  Features:              None
Pool Info:
  Pool 1
    Segment:             GLOBAL; FLAGS: FINE GRAINED
    Size:                792303268(0x2f3996a4) KB
    Allocatable:          TRUE
    Alloc Granule:        4KB
    Alloc Alignment:      4KB
    Accessible by all:   TRUE
  Pool 2
    Segment:             GLOBAL; FLAGS: KERNARG, FINE GRAINED
    Size:                792303268(0x2f3996a4) KB
    Allocatable:          TRUE
    Alloc Granule:        4KB
    Alloc Alignment:      4KB
    Accessible by all:   TRUE
  Pool 3
    Segment:             GLOBAL; FLAGS: COARSE GRAINED
    Size:                792303268(0x2f3996a4) KB
    Allocatable:          TRUE
    Alloc Granule:        4KB
    Alloc Alignment:      4KB
    Accessible by all:   TRUE
ISA Info:
[...]
Name:                  gfx942
Uuid:                  GPU-ababaeffecddc50
Marketing Name:        AMD Instinct MI300X
Vendor Name:           AMD
Feature:               KERNEL_DISPATCH
Profile:               BASE_PROFILE
Float Round Mode:     NEAR

```

```

Max Queue Number: 128(0x80)
Queue Min Size: 64(0x40)
Queue Max Size: 131072(0x20000)
Queue Type: MULTI
Node: 2
Device Type: GPU
Cache Info:
  L1: 16(0x10) KB
  L2: 8192(0x2000) KB
  Chip ID: 29857(0x74a1)
  ASIC Revision: 1(0x1)
  Cacheline Size: 64(0x40)
  Max Clock Freq. (MHz): 2100
  BDFID: 50688
  Internal Node ID: 7
  Compute Unit: 304
  SIMDs per CU: 4
  Shader Engines: 32
  Shader Arrs. per Eng.: 1
  WatchPts on Addr. Ranges: 4
  Coherent Host Access: FALSE
  Features: KERNEL_DISPATCH
  Fast F16 Operation: TRUE
  Wavefront Size: 64(0x40)
  Workgroup Max Size: 1024(0x400)
  Workgroup Max Size per Dimension:
    x 1024(0x400)
    y 1024(0x400)
    z 1024(0x400)
  Max Waves Per CU: 32(0x20)
  Max Work-item Per CU: 2048(0x800)
  Grid Max Size: 4294967295(0xffffffff)
  Grid Max Size per Dimension:
    x 4294967295(0xffffffff)
    y 4294967295(0xffffffff)
    z 4294967295(0xffffffff)
  Max fbarriers/Workgrp: 32
  Packet Processor uCode:: 132
  SDMA engine uCode:: 19
  IOMMU Support:: None
  Pool Info:
    Pool 1
      Segment: GLOBAL; FLAGS: COARSE GRAINED
      Size: 201310208(0xbfffc000) KB
      Allocatable: TRUE
      Alloc Granule: 4KB
      Alloc Alignment: 4KB
      Accessible by all: FALSE
    Pool 2
      Segment: GLOBAL; FLAGS: EXTENDED FINE GRAINED
      Size: 201310208(0xbfffc000) KB

```

```

Allocatable:          TRUE
Alloc Granule:       4KB
Alloc Alignment:     4KB
Accessible by all:  FALSE

Pool 3
Segment:            GLOBAL; FLAGS: FINE GRAINED
Size:                201310208(0xbfffc000) KB
Allocatable:          TRUE
Alloc Granule:       4KB
Alloc Alignment:     4KB
Accessible by all:  FALSE

Pool 4
Segment:            GROUP
Size:                64(0x40) KB
Allocatable:          FALSE
Alloc Granule:       0KB
Alloc Alignment:     0KB
Accessible by all:  FALSE

ISA Info:
ISA 1
Name:                amdgcn-amd-amdhsa--gfx942:sramecc+:xnack-
Machine Models:      HSA_MACHINE_MODEL_LARGE
Profiles:            HSA_PROFILE_BASE
Default Rounding Mode: NEAR
Default Rounding Mode: NEAR
Fast f16:             TRUE
Workgroup Max Size:  1024(0x400)
Workgroup Max Size per Dimension:
  x                  1024(0x400)
  y                  1024(0x400)
  z                  1024(0x400)
Grid Max Size:        4294967295(0xffffffff)
Grid Max Size per Dimension:
  x                  4294967295(0xffffffff)
  y                  4294967295(0xffffffff)
  z                  4294967295(0xffffffff)
FBarrier Max Size:   32

```

AMD GPUs Performance

As I haven't had a chance to do any serious work with AMD GPUs, just sharing links for now.

- [AMD Instinct MI300X system optimization](#)
- [AMD Instinct MI300X workload optimization](#)

Profilers

[omniprof](#) - Advanced Profiling and Analytics for AMD Hardware - e.g. can plot a roofline performance of your AMD accelerator and many other things.

CPU

As of this writing Machine learning workloads don't use much CPU so there aren't too many things to tell in this chapter. As CPUs evolve to become more like GPUs this is likely to change, so I'm expecting this chapter to evolve along the evolution of the CPUs.

How many cpu cores do you need

Per 1 accelerator you need:

- 1 cpu core per process that is tied to the accelerator
- 1 cpu core for each `DataLoader` worker process - and typically you need 2-4 workers.

2 workers is usually plenty for LMs, especially if the data is already preprocessed.

If you need to do dynamic transforms, which is often the case with computer vision models or VLMs, you may need 3-4 and sometimes more workers.

The goal is to be able to pull from the `DataLoader` instantly, and not block the accelerator's compute, which means that you need to pre-process a bunch of samples for the next iteration, while the current iteration is running. In other words your next batch needs to take no longer than a single iteration accelerator compute of the batch of the same size.

Besides preprocessing if you're pulling dynamically from the cloud instead of local storage you also need to make sure that the data is pre-fetched fast enough to feed the workers that feed the accelerator furnace.

Multiply that by the number of accelerators, add a few cores for the Operation system (let's say 4).

If the node has 8 accelerators, and you have `n_workers`, then you need $8 * (\text{num_workers} + 1) + 4$. If you're doing NLP, it'd be usually about 2 workers per accelerator, so $8 * (2 + 1) + 4 \Rightarrow 28$ cpu cores. If you do CV training, and, say, you need 4 workers per accelerator, then it'd be $8 * (4 + 1) + 4 \Rightarrow 44$ cpu cores.

What happens if you have more very active processes than the total number of cpu cores? Some processes will get preempted (put in the queue for when cpu cores become available) and you absolutely want to avoid any context switching.

But modern cloud offerings typically have 50-100+ cpu-cores so usually there is no problem to have enough cores to go around.

See also [Asynchronous DataLoader](#).

CPU offload

Some frameworks, like [Deepspeed](#) can offload some compute work to CPU without creating a bottleneck. In which case you'd want additional cpu-cores.

NUMA affinity

See [NUMA affinity](#).

Hyperthreads

[Hyper-Threads](#) double the cpu cores number, by virtualizing each physical core into 2 virtual ones, allowing 2 threads to use the same cpu core at the same time. Depending on the type of workload this feature may or may not increase

the overall performance. Intel, the inventor of this technology, suggests a possible 30% performance increase in some situations.

See also [To enable Hyper-Threads or not.](#)

CPU memory

This is a tiny chapter, since usually there are very few nuances one needs to know about CPU memory - which is a good thing!

Most of the ML workload compute happens on GPUs, but typically there should be at least as much CPU memory on each node as there is on the GPUs. So, for example, if you're on a H100 node with 8x 80GB GPUs, you have 640GB of GPU memory. Thus you want at least as much of CPU memory. But most recent high end cloud packages usually come with 1-2TBs of CPU memory.

What CPU memory is needed for in ML workloads

- Loading the model weights, unless they are loaded directly onto the GPUs - this is usually a transitory memory usage that goes back to zero once the model has been moved to GPUs.
- Saving the model weights. In some situations each GPU writes its own checkpoint directly to the disk, in other cases the model is recomposed on the CPU before it's written to disk - this too is a transitory memory usage.
- Possible parameter and optimizer state offloading when using frameworks like [Deepspeed](#). In which case quite a lot of CPU memory might be needed.
- Activations calculated in the `forward` pass, and which need to be available for the `backward` path can also be offloaded to CPU, rather than discarded and then recomputed during the backward pass to save the unnecessary overhead
- `DataLoader` is usually one of the main users of CPU memory and at times it may consume very large amounts of memory. Typically there are at least 2x 8 DL workers running on each node, so you need enough memory to support at least 16 processes each holding some data. For example, in the case of streaming data from the cloud, if the data shards are large, these processes could easily eat up hundreds of GBs of CPU memory.
- The software itself and its dependent libraries uses a bit of CPU memory, but this amount is usually negligible.

Things to know

- If the `DataLoader` uses HF datasets in `mmap` mode the Resident memory usage may appear to be using a huge amount of CPU memory as it'll try to map out the whole datasets to the memory. Except this is misleading, since if the memory is needed elsewhere the OS will page out any unneeded mmap'ed pages back to the system. You can read more about it [here](#). This awareness, of course, applies to any dataset using `mmap`, I was using HF datasets as an example since it's very widely used.

Storage: File Systems and IO

3 Machine Learning IO needs

There are 3 distinct IO needs in the ML workload:

1. You need to be able to feed the DataLoader fast - (super fast read, don't care about fast write) - requires sustainable load for hours and days
2. You need to be able to write checkpoints fast - (super fast write, fastish read as you will be resuming a few times) - requires burst writing - you want super fast to not block the training for long (unless you use some sort of cpu offloading to quickly unblock the training)
3. You need to be able to load and maintain your codebase - (medium speed for both reading and writing) - this also needs to be shared since you want all nodes to see the same codebase - as it happens only during the start or resume it'll happen infrequently

As you can see these 3 have very different requirements both on speed and sustainable load, and thus ideally you'd have 3 different filesystems, each optimized for the required use case.

If you have infinite funds, of course, get a single super-fast read, super-fast write, that can do that for days non-stop. But for most of us, this is not possible so getting 2 or 3 different types of partitions where you end up paying much less is a wiser choice.

Glossary

- NAS: Network Attached Storage
- SAN: Storage Area Network
- DAS: Direct-Attached storage
- NSD: Network Shared Disk
- OSS: Object storage server
- MDS: Metadata server
- MGS: Management server

Which file system to choose

Distributed Parallel File Systems are the fastest solutions

Distributed parallel file systems dramatically improve performance where hundreds to thousands of clients can access the shared storage simultaneously. They also help a lot with reducing hotspots (where some data pockets are accessed much more often than others).

The 2 excellent performing parallel file systems that I had experience with are:

- [Lustre FS](#) (Open Source) ([Wiki](#))
- [GPFS](#) (IBM), recently renamed to IBM Storage Scale, and before that it was called IBM Spectrum Scale.

Both solutions have been around for 2+ decades. Both are POSIX-compliant. These are also not trivial to create - you have to setup a whole other cluster with multiple cpu-only VMs dedicated exclusively for those filesystems - only then you can mount those. As compared to weaker cloud-provided "built-in" solutions which take only a few screens of questions to answer in order to activate. And when creating the storage cluster there is a whole science to which VMs to choose for which functionality. For example, here is a [Lustre guide on GCP](#).

case study: At JeanZay HPC (France) we were saving 2.3TB checkpoint in parallel on 384 processes in 40 secs! This is insanely fast - and it was GPFS over NVME drives.

NASA's cluster has [a long long list of gotchas around using Lustre](#).

Some very useful pros of GPFS:

- If you have a lot of small files, you can easily run out of inodes (`df -i` to check). GPFS 5.x never runs out of inodes, it dynamically creates more as needed
- GPFS doesn't have the issue Lustre has where you can run out of disk space at 80% if one of the sub-disks got full and wasn't re-balanced in time - you can reliably use all 100% of the allocated storage.
- GPFS doesn't use a central metadata server (or a cluster of those) which often becomes a bottleneck when dealing with small files. Just like data, metadata is handled by each node in the storage cluster.
- GPFS comes with a native NSD client which is superior to the generic NFS client, but either can be used with it.

Other parallel file systems I don't yet have direct experience with:

- [BeeGFS](#)
- [WekaIO](#)
- [DAOS](#) (Distributed Asynchronous Object Storage) (Intel)
- [NetApp](#)

Most clouds provide at least one implementation of these, but not all. If your cloud provider doesn't provide at least one of these and they don't have a fast enough alternative to meet your needs you should reconsider.

OK'ish solutions

There are many OK'ish solutions offered by [various cloud providers](#). Benchmark those seriously before you commit to any. Those are usually quite decent for handling large files and not so much for small files.

case study: As of this writing with GCP's Zonal FileStore over NFS solution `python -c "import torch"` takes 20 secs to execute, which is extremely slow! Once the files are cached it then takes ~2 secs. Installing a conda environment with a handful of prebuilt python packages can easily take 20-30 min! This solution we started with had been very painful and counter-productive to our work. This would impact anybody who has a lot of python packages and conda environments. But, of course, GCP provides much faster solutions as well.

Remote File System Clients

You will need to choose which client to use to connect the file system to your VM with.

The most common choice is: [NFS](#) - which has been around for 4 decades. It introduces an additional overhead and slows things down. So if there is a native client supported by your VM, you'd have an overall faster performance using it over NFS. For example, GPFS comes with an [NSD](#) client which is superior to NFS.

File Block size

If the file system you use uses a block size of 16mb, but the average size of your files is 16k, you will be using 1,000 times more disk space than the actual use. For example, you will see 100TB of disk space used when the actual disk space will be just 100MB.

footnote: On Linux the native file systems typically use a block size of 4k.

So often you might have 2 very different needs and require 2 different partitions optimized for different needs.

1. thousands to millions of tiny files - 4-8k block size
2. few large files - 2-16mb block size

case study: Python is so bad at having tens of thousand of tiny files that if you have many conda environments you are likely to run out of inodes in some situations. At JeanZay HPC we had to ask for a special dedicated partition where we would install all conda environments because we kept running out of inodes on normal GPFS partitions. I think the problem is that those GPFS partitions were configured with 16MB block sizes, so this was not a suitable partition for 4KB-large files.

The good news is that modern solutions are starting to introduce a dynamic block size. For example, the most recent GPFS supports sub-blocks. So, for example, it's possible to configure GPFS with a block size of 2mb, with a sub-block of 8k, and then the tiny files get packed together as sub-blocks, thus not wasting too much disk space.

Distributed storage servers proximity to clients

The cluster that uses a shared distributed storage should have the storage servers places close to the cluster that uses those servers. If the VMs running the storage servers are located many hops (switches) away, the IO latency can be high and the interactive use of the storage can be frustratingly slow. Think any interactions with metadata servers as an example, when you try to run `du` and other tools that access metadata of many files.

So if you have control ask the cloud provider to give you the cpu-only storage servers VMs allocated as close as possible to your accelerator VMs network-distance-wise.

Cloud shared storage solutions

Here are shared file system storage solutions made available by various cloud providers:

- [GCP](#)
- [Azure](#)
- [AWS](#)

Local storage beats cloud storage

While cloud storage is cheaper the whole idea of fetching and processing your training data stream dynamically at training time is very problematic with a huge number of potential issues around it.

Same goes for dynamic offloading of checkpoints to the cloud.

It's so much better to have enough disk space locally for data loading.

For checkpointing there should be enough local disk space for saving a checkpoint in a fast and reliable way and then having a crontab job or a slurm job to offload it to the cloud. Always keep the last few checkpoints locally for a quick resume, should your job crash, as it'd be very expensive to wait to fetch the checkpoint from the cloud for a resume.

case study: we didn't have a choice and had to use cloud storage for dataloading during IDEFICS-80B training as we had barely any local storage and since it was multimodal data it was many TBs of data. We spent many weeks trying to make this solution robust and it sucked at the end. The biggest issue was that it was very difficult at the time to keep track of RNG state for the DataSampler because the solution we used, well, didn't bother to take care of it. So a lot of data that took a lot of time to create was wasted (not used) and a lot of data was repeated, so we didn't have a single epoch of unique data.

In some situations people find good solutions for working with cloud-based datasets, I personally haven't had a smooth experience yet and that's why I advocate local storage. If you found a good streaming solution that can properly resume without losing data and repeating the same data, doesn't require huge local workers then it might work OK.

Beware that you're often being sold only 80% of the storage you pay for

There is a subtle problem with distributed shared storage used on compute nodes. Since most physical disks used to build the large file systems are only 0.3-2TB large, any of these physical disks can get full before the combined storage gets full. And thus they require constant rebalancing so that there will be no situation where one disk is 99% full and others are only 50% full. Since rebalancing is a costly operation, like most programming languages' garbage collection, it happens infrequently. And so if you run `df` and it reports 90% full, it's very likely that any of the programs can fail at any given time.

From talking to IO engineers, the accepted reality (that for some reason is not being communicated to customers) is that only about 80% of distributed large storage is reliable.

Which means that if you want to have 100TB of reliable cloud storage you actually need to buy 125TB of storage, since 80% of that will be 100TB. So you need to plan to pay 25% more than what you provisioned for your actual needs. I'm not sure why the customer should pay for the technology deficiency but that's how it is.

For example, GCP states that only [89%](#) can be used reliably, albeit more than once the storage failed already at 83% for me there. Kudos to Google to even disclosing this as a known issue, albeit not at the point of where a person buys the storage. As in - we recommend you buy 12% more storage than you actually plan to use, since we can only reliably deliver 89% of it.

I also talked to [Sycomp](#) engineers who provide managed IBM Storage Scale (GPFS) solutions, and according to them GPFS doesn't have this issue and the whole 100% can be reliably used.

Also on some setups if you do backups via the cloud provider API (not directly on the filesystem), they might end up using the same partition, and, of course, consume the disk space, but when you run `df` it will not show the real disk usage - it may show usage not including the backups. So if your backups consume 50% of the partition.

Whatever storage solution you pick, ask the provider how much of the storage can be reliably used, so that there will be no surprises later.

Beware that on some cloud providers backups use the same partition they backup

This makes no sense to me but with some providers when you make a back up of a partition using their tools, the back up will use space on that same partition. And on some of those providers you won't even know this happened until you run out of disk space when you really used 30% of the partition you allocated. On those providers running `df` is pointless because it'll tell you the free disk space, but it won't include any back ups in it. So you have no idea what's going on.

If you start making a backup and suddenly everything fails because all processes fail to write but `df` reports 30% usage, you will now know why this happened. Snapshots too use the same partition.

So say you paid for a 100TB partition and you used up 95TB and now you want to back it up - well, you can't - where would it put 95TB of data if it has 5TB of data left even if it compresses it.

As I discover specific solution that have this unintuitive behavior I will add pointers to how you can see the actual disk usage:

- [GCP FileStore](#) (but it doesn't work for Basic Tier)

Don't forget the checksums

When you sync data to and from the cloud make sure to research whether the tool you use checks the checksums, otherwise you may end up with corrupt during transmission data. Some tools do it automatically, others you have to enable this feature (since it usually comes at additional compute cost and transmission slowdown). Better slow, but safe.

These are typically MD5 and SHA256 checksums. Usually MD5 is sufficient if your environment is safe, but if you want the additional security do SHA256 checksums.

Concepts

Here are a few key storage-related concepts that you likely need to be familiar with:

Queue Depth

Queue depth (or **IO depth**) is the number of IO requests that can be queued at one time on a storage device controller. If more IO requests than the controller can queue are being sent the OS will usually put those into its own queue.

On Linux the local block devices' queue depth is usually pre-configured by the kernel. For example, if you want to check the max queue depth set for `/dev/sda` you can `cat /sys/block/sda/queue/nr_requests`. To see the current queue depth of a local device run `iostat -x` and watch for `aqo-sz` column. (`apt install sysstat` to get `iostat`.)

Typically the more IO requests get buffered the bigger the latency will be, and the better the throughput will be. This is because if a request can't be acted upon immediately it'll prolong the response time as it has to wait before being served. But having multiple requests awaiting to be served in a device's queue would typically speed up the total throughput as there is less waiting time between issuing individual requests.

Direct vs Buffered IO

Direct IO refers to IO that bypasses the operating system's caching buffers. This corresponds to `O_DIRECT` flag in `open(2)` system call.

The opposite is the **buffered** IO, which is usually the default way most applications do IO since caching typically makes things faster.

When we run an IO benchmark it's critical to turn the caching/buffering off, because otherwise the benchmark's results will most likely be invalid. You normally won't be reading or writing the same file hundreds of times in a row. Hence most likely you'd want to turn the direct mode on in the benchmark's flags if it provides such.

In certain situations opening files with `O_DIRECT` may actually help to overcome delays. For example, if the training program logs to a log file (especially on a slow shared file system), you might not be able to see the logs for many seconds if both the application and the file system buffering are in the way. Opening the log file with `O_DIRECT` by the writer typically helps to get the reader see the logged lines much sooner.

Synchronous vs asynchronous IO

In **synchronous** IO the client submits an IO request and waits for it to be finished before submitting the next IO request to the same target device.

In **asynchronous** IO the client may submit multiple IO requests one after another without waiting for any to finish first. This requires that the target device can [queue up multiple IO requests](#).

Sequential vs Random access IO

Sequential access IO is when you read blocks of data one by one sequentially (think a movie). Here are some examples:

- reading or writing a model's checkpoint file all at once
- loading a python program
- installing a package

Random access IO is when you're accessing part of a file at random. Here are some examples:

- database querying
- reading samples from a pre-processed dataset in a random fashion
- moving around a file using `seek`

Benchmarks

Time is money both in terms of a developer's time and model's training time, so it's crucial that storage IO isn't a bottleneck in your human and compute workflows.

In the following sections we will discuss various approaches to figuring out whether the proposed storage solution satisfies your work needs.

Metrics

The three main storage IO metrics one typically cares for are:

1. [Throughput](#) or Bandwidth (bytes per second - can be MBps, GBps, etc.)
 2. [IOPS](#) (Input/output operations per second that a system can perform)
 3. [Latency](#) (msecs or usecs)
 - *IOPS* measures how many input and/or output operations a given storage device or a cluster can perform per second. Typically read and write IOPS won't be the same. And for many systems it'll also depend on whether the operation is sequential or random. So a storage system will have 4 different IOPS rates:
 - 1. IOPS of random reads
 - 2. IOPS of random writes
 - 3. IOPS of sequential reads
 - 4. IOPS of sequential writes
- *Throughput* refers to how much data can be processed per second.

IOPS vs. Throughput

- when you deal with small files high IOPS is important.
- when you deal with large files high throughput is important.

IOPS correlates to Throughput via block size: `Throughput = IOPS * block_size`

Thus given a fixed IOPS - the larger the block size that the system can read or write the bigger the throughput will be.

And since there are 4 IOPS categories, correspondingly there are 4 throughput values to match.

Latency: is the delay between the moment the instruction to transfer data is issued and when the response to that instruction arrives.

Typically the more distance (switches, relays, actual distance) the packet has to travel the bigger the latency will be.

So if you have a local NVME drive your read or write latency will be much shorter as compared to reading or writing to a storage device that is located on another continent.

fio

[fio - Flexible I/O tester](#) is a commonly used IO benchmarking tool, which is relatively easy to operate. It has many options which allow you to emulate pretty much any type of a load and it provides a very detailed performance report.

First install `fio` with `apt install fio` or however your package manager does it.

Here is an example of a read benchmark:

```
base_path=/path/to/partition/
fio --ioengine=libaio --filesize=16k --ramp_time=2s --time_based --runtime=3m --numjobs=16 \
--direct=1 --verify=0 --randrepeat=0 --group_reporting --unlink=1 --directory=$base_path \
--name=read-test --blocksize=4k --iodepth=64 --readwrite=read
```

Here 16 concurrent read threads will run for 3 minutes. The benchmark uses a block size of 4k (typical for most OSes) with the file size of 16k (a common size of most Python files) in a sequential reading style using [non-buffered IO](#). So this particular set of flags will create a good benchmark to show how fast you can import Python modules on 16 concurrent

processes.

case study: on one NFS setup we had `python -c "import torch"` taking 20 seconds the first time it was run, which is about 20x slower than the same test on a normal NVME drive. Granted once the files were cached the loading was much faster but it made for a very painful development process since everything was slow.

good read: [Fio Output Explained](#) - it's an oldie but is still a goodie - if you have a more up-to-date write up please send me a link or a PR.

Important: if you don't use the `--unlink=1` flag make sure to delete `fio`'s work files between different benchmarks - not doing so can lead to seriously wrong reports as `fio` will reuse files it prepared for a different benchmark which must not be re-used if the benchmark parameters have changed. Apparently this reuse is an `fio` feature, but to me it's a bug since I didn't know this nuance and got a whole lot of invalid reports because of it and it took awhile to realize they were wrong.

Going back to the benchmark - the parameters will need to change to fit the type of the IO operation you care to be fast - is it doing a lot of pip installs or writing a checkpoint on 512 processes, or doing a random read from a parquet file - each benchmark will have to be adapted to measure the right thing.

At the beginning I was manually fishing out the bits I was after, so I automated it resulting in `fio-scan` benchmark that will run a pair of read/write benchmarks on 16KB, 1MB and 1GB file sizes each using a fixed 4k block size (6 benchmarks in total). It uses a helper `fio-json-extract.py` to parse the log files and pull out the average latency, bandwidth and iops and report them in a nicely formatted markdown table.

Here is how to run it:

```
git clone https://github.com/stas00/ml-engineering/
cd ml-engineering
cd storage

path_to_test=/path/to/partition/to/test
./fio-scan $path_to_test
```

Adapt `path_to_test` to point to the partition path you want to benchmark.

note: the log parser uses python3. if `fio-scan` fails it's most likely because you run it on a system with python2 installed by default. It expects `python --version` to be some python 3.x version. You can edit `fio-scan` to point to the right python.

Here is an example of this IO scan on my Samsung SSD 980 PRO 2TB NVME drive ([summary](#)):

- filesize=16k read

lat msec	bw MBps	IOPS	jobs
4.0	1006.3	257614	16

- filesize=16k write

lat msec	bw MBps	IOPS	jobs
3.2	1239.1	317200	16

- filesize=1m read

lat msec	bw MBps	IOPS	jobs
1.7	2400.1	614419	16

- filesize=1m write

lat msec	bw MBps	IOPS	jobs
2.1	1940.5	496765	16

- filesize=1g read

lat msec	bw MBps	IOPS	jobs
1.4	2762.0	707062	16

- filesize=1g write

lat msec	bw MBps	IOPS	jobs
2.1	1943.9	497638	16

As you can see as of this writing this is a pretty fast NVMe drive if you want to use it as a base-line against, say, a network shared file system.

Usability perception IO benchmarks

Besides properly designed performance benchmarks which give you some numbers that you may or may not be able to appreciate there is a perception benchmark, and that is how does a certain functionality or a service feel. For example, when going to a website, does it feel like it's taking too long to load a webpage? or when going to a video service, does it take too long for the video to start playing and does it stop every few seconds to buffer the stream?

So with file system the questions are very simple - does it feel that it takes too long to install or launch a program? Since a lot of us live in the Python world, python is known to have thousands of tiny files which are usually installed into a virtual environment, with [conda](#) being the choice of many as of this writing.

In one of the environments we have noticed that our developers' productivity was really bad on a shared filesystem because it was taking up to 30min to install a conda environment with various packages needed for using a certain ML-training framework, and we also noticed that `python -c "import torch"` could take more than 20 seconds. This is about 5-10x slower than a fast local NVME-based filesystem would deliver. Obviously, this is bad. So I devised a perception test using `time` to measure the common activities. That way we could quickly tell if the proposed shared file system solution that we contemplated to switch to were significantly better. We didn't want a solution that was 2x faster, we wanted a solution that was 10x better, because having an expensive developer wait for proverbial paint to dry is not a good thing for a business.

So here is the poor man's benchmark that we used, so this is just an example. Surely if you think about the workflow of your developers you would quickly identify where things are slow and devise yours best fitting your needs.

note: To have a baseline to compare to do these timing tests on a recently manufactured local NVME. This way you know what the ceiling is, but with beware that many shared file systems won't be able to match that.

Step 1. Install conda onto the shared file system you want to test if it's not there already.

```

export target_partition_path=/mnt/weka # edit me!!!
mkdir -p $target_partition_path/miniconda3
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh -O $target_partition_path/miniconda3/miniconda.sh
bash $target_partition_path/miniconda3/miniconda.sh -b -u -p $target_partition_path/miniconda3
rm -rf $target_partition_path/miniconda3/miniconda.sh
$target_partition_path/miniconda3/bin/conda init bash
bash

```

notes:

- adapt `target_partition_path` and the miniconda download link if you aren't on the x86 platform.
- at the end we launch a new `bash` shell for conda setup to take an effect, you might need to tweak things further if you're not a `bash` user - I trust you will know what to do.

Step 2. Measure conda install time (write test)

Time the creation of a new conda environment:

```
time conda create -y -n install-test python=3.9
```

```

real    0m29.657s
user    0m9.141s
sys     0m2.861s

```

Time the installation of some heavy pip packages:

```

conda deactivate
conda activate install-test
time pip install torch torchvision torchaudio

```

```

real    2m10.355s
user    0m50.547s
sys     0m12.144s

```

Please note that this test is somewhat skewed since it also includes the packages download in it and depending on your incoming network speed it could be super fast or super slow and could impact the outcome. But once the downloaded packages are cached, in the case of conda they are also untarred, so if you try to install the packages the 2nd time the benchmark will no longer be fair as on a slow shared file system the untarring could be very slow and we want to catch that.

I don't worry about it because usually when the file system is very slow usually you can tell it's very slow even if the downloads are slow, you just watch the progress and you can just tell.

If you do want to make this benchmark precise, you probably could keep the pre-downloaded conda packages and just deleting their untar'ed dirs:

```
find $target_partition_path/miniconda3/pkgs -mindepth 1 -type d -exec rm -rf {} +
```

in the case of `pip` it doesn't untar anything, but just caches the wheels it downloaded, so the `time pip install` benchmark can definitely be more precise if you run it the 2nd time (the first time it's downloaded, cached and installed, the second time it's installed from cache). So you could do:

```
conda create -y -n install-test python=3.9
conda activate install-test
pip install torch torchvision torchaudio
conda create -y -n install-test2 python=3.9
conda activate install-test2
time pip install torch torchvision torchaudio
```

As you can see here we time only the 2nd time we install the `pip` packages.

Step 3. Measure loading time after flushing the memory and file system caches (read test)

```
sudo sync
echo 3 | sudo tee /proc/sys/vm/drop_caches
time python -c "import torch"
```

As you can see before we do the measurement we have to tell the OS to flush its memory and file system caches.

If you don't have `sudo` access you can skip the command involving `sudo`, also sometimes the system is setup to work w/o `sudo`. If you can't run the syncing and flushing of the file system caches you will just get incorrect results as the benchmark will be measuring the time to load already cached file system objects. To overcome this either ask your sysadmin to do it for you or simply come back in the morning while hopefully your file system caches other things and evicts the python packages, and then repeat the python one liner then with the hope those files are no longer in the cache.

Here is how to see the caching effect:

```
$ time python -c "import torch"

real    0m5.404s
user    0m1.761s
sys     0m0.751s

$ time python -c "import torch"

real    0m1.977s
user    0m1.623s
sys     0m0.519s

$ sudo sync
$ echo 3 | sudo tee /proc/sys/vm/drop_caches
$ time python -c "import torch"

real    0m5.698s
```

```
user  0m1.712s
sys   0m0.734s
```

You can see that the first time it wasn't cached and took ~3x longer, then when I run it the second time. And then I told the system to flush memory and file system caches and you can see it was 3x longer again.

I think it might be a good idea to do the memory and file system caching in the write tests again, since even there caching will make the benchmark appear faster than what it would be like in the real world where a new package is installed for the first time.

Another time I noticed that `git status` was taking multiple seconds. I use [bash-git-prompt](#) and it runs `git status` before every return of the prompt when inside a git repo clone, and it was becoming super sluggish and difficult to work. So I benchmarked `git status`:

```
git clone https://github.com/pytorch/pytorch
cd pytorch
time git status
```

and it was taking 3.7secs on this slow file system and needed to be fixed (it was taking 0.02 secs on a local SSD). The good thing this actual perception benchmark was easy to pass to a sysadmin and them reproducing the problem instantly and then working on fixing it, while re-using this benchmark as a reference.

Yet, another time I noticed, `pytest` was taking forever to start, so I measured its collection and it indeed was very slow:

```
time pytest --disable-warnings --collect-only -q
```

So now you have a plethora of examples to choose from and I trust you will find your own use cases which are easy to reliably reproduce and use as a reference point for what feels good and what doesn't and which need to be fixed.

other tools

- [HPC IO Benchmark Repository](#) (`mdtest` has been merged into `ior` in 2017)
- [DLIO](#)

XXX: expand on how these are used when I get a chance to try those

Published benchmarks

Here are some published IO benchmarks:

- [MLPerf via MLCommons](#) publishes various hardware benchmarks that measure training, inference, storage and other tasks' performance. For example, here is the most recent as of this writing [storage v0.5](#) results. Though I find the results are very difficult to make sense of - too many columns and no control whatsoever by the user, and each test uses different parameters - so how do you compare things.

Then various benchmarks that you can run yourself:

Why pay for more storage when you can easily clean it up instead

Talking to a few storage providers I understood that many companies don't bother cleaning up and just keep on buying

more and more storage. If you're not that company and want to keep things tidy in the following sections I will share how to easily prune various caches that many of us in the Python/Pytorch ecosphere use (and a lot of those will apply to other ecospheres).

HuggingFace Hub caches

The very popular HuggingFace Hub makes it super easy to download models and datasets and cache them locally. What you might not be aware of is that whenever a new revision of the model or a dataset is released, the old revisions remain on your disk - so over time you are likely to have a lot of dead weight.

The cached files are usually found at `~/.cache/huggingface` but it's possible to override those with `HF_HOME` environment variable and place them elsewhere if your `/home/` doesn't have space for huge files. (and in the past those were `HUGGINGFACE_HUB_CACHE` and `TRANSFORMERS_CACHE` and some others).

The other solution that requires no mucking with environment variables, which requires you to remember to set them, is to symlink your cache to another partition. You could do it for all of your caches:

```
mkdir -p ~/.cache
mv ~/.cache /some/path/
ln -s /some/path/.cache ~/.cache
```

or just for HF hub caches:

```
mkdir -p ~/.cache/huggingface
mv ~/.cache/huggingface /some/path/
ln -s /some/path/cache/huggingface ~/.cache/cache/huggingface
```

The `mkdir` calls are there in case you have haven't used the caches yet, so they weren't there and they ensure the above code won't fail.

Now that you know where the caches are, you could, of course, nuke the whole cache every so often, but if these are huge models and datasets, and especially if there was some preprocessing done for the latter - you really won't want to repeat those time consuming tasks again and again. So I will teach you how to use special tools provided by HuggingFace to do the cleanup.

The way revisions work on the HF hub is by pointing `main` to the latest revision of the files while keeping the old revisions around should anyone want to use the older revision for some reason. Chance are very high you always want the latest revision, and so here is how to delete all old revisions and only keeping `main` in a few quick steps without tedious manual editing.

In terminal A:

```
$ pip install huggingface_hub["cli"] -U
$ huggingface-cli delete-cache --disable-tui
File to edit: /tmp/tmpundr71ky.txt
0 revisions selected counting for 0.0. Continue ? (y/N)
```

Do not answer the prompt and proceed with my instructions.

(note your tmp file will have a different path, so adjust it below)

In terminal B:

```
$ cp /tmp/tmpedbzb00ox.txt cache.txt
$ perl -pi -e 's|^#(.*)|(detached).*)|$1|' cache.txt
$ cat cache.txt >> /tmp/tmpundr7lky.txt
```

The perl one-liner uncommented out all lines that had `(detached)` in it - so can be wiped out. And then we pasted it back into the tmp file `huggingface-cli` expects to be edited.

Now go back to terminal A and hit: N, Y, Y, so it looks like:

```
0 revisions selected counting for 0.0. Continue ? (y/N) n
89 revisions selected counting for 211.7G. Continue ? (y/N) y
89 revisions selected counting for 211.7G. Confirm deletion ? (Y/n) y
```

Done.

If you messed up with the prompt answering you still have `cache.txt` file which you can feed again to the new tmp file it'll create when you run `huggingface-cli delete-cache --disable-tui` again.

attached as a snapshot as well as it's easier to read on twitter, but use the message to copy-n-paste from.

Please note that you can also use this tool to choose which models or datasets to delete completely. You just need to open `cache.txt` in your editor and remove the `#` in front of lines that contain `main` in it for models/datasets you want to be deleted for you. and then repeat the process explained above minus the `perl` one liner which you'd replace with manual editing.

Additionally you will find that HF datasets have a `~/.cache/huggingface/datasets/downloads` dir which often will contain a ton of leftovers from datasets downloads and their preprocessing, including various lock files. On one setup I found literally a few millions of files there. So here is how I clean those up:

```
sudo find ~/.cache/huggingface/datasets/downloads -type f -mtime +3 -exec rm {} \+
sudo find ~/.cache/huggingface/datasets/downloads -type d -empty -delete
```

The first command leaves files that are younger than 3 days in place, in case someone is in the process of download/processing things and we don't want to swipe the carpet from under their feet.

As usual you may need to adjust the paths if you placed your caches elsewhere.

note: if your team uses `HF_HOME` to share the HF hub models/datasets/etc - the `$HF_HOME/token` will get shared as well, which works fine as long as ungated models are used. But if you want to access gated models you might run into problems there. Therefore you most likely want to not share the access token. You can fix that by adding something like:

```
export HF_TOKEN_PATH=~/cache/hf_hub_token
```

(then put it into `~/.bashrc` to always work)

and now how each user run once:

```
huggingface-cli login
```

which will ask them to add their access token from <https://huggingface.co/settings/tokens> - it'll save it under `~/.cache/hf_hub_token`.

Now each member of your team will have their unique token and the gated models approved for their HF hub user will now be accessible by them.

Python package manager cleanups

conda and pip will pile up more and more files on your system over time. conda is the worst because it keeps the untarred files which consume an insane amount of inodes and make backups and scans slow. pip at least caches just the wheels (tarred files).

So you can safely nuke these dirs:

```
rm -rf ~/.cache/pip
rm -rf ~/anaconda3/pkggs/
```

Make sure edit the last command if your conda is installed elsewhere.

Share caches in group environments

If you have more than 2 people working on the same system, you really want to avoid each person having their own cache of pip, conda, HF models, datasets and possibly other things. It is very easy to get each user's setup to point to a shared cache.

For example, let's say you make pip and conda caches under `/data/cache` like so:

```
mkdir /data/cache/conda
mkdir /data/cache/pip
chmod a+rwx /data/cache/conda
chmod a+rwx /data/cache/pip
```

now you just need to symlink from each user's local cache to this shared cache:

```
mkdir -p ~/.cache

rm -rf ~/.cache/pip
ln -s /data/cache/pip ~/.cache/pip

rm -rf ~/.conda/pkggs
ln -s /data/cache/conda/pkggs ~/.conda/pkggs
```

note that we wiped out the existing caches, but you could also move them to the shared cache instead - whatever works, you will want to periodically nuke those anyway.

So now when pip or conda will try to reach the user caches they will get redirected to the shared cache. If you have 20

people in the group that's 20x less files - and this is very important because conda pkg files are untarred and take up a huge amount of inodes on the disk.

So the only issue with this approach is file permissions. If user A installs some packages, user B might not be able to read or write them.

If this is an isolated cluster where there are no malicious users you can simply ask everybody to use `umask 000` in their `~/.bashrc` or even configuring this setting system-wide via `/etc/profile` or `/etc/bash.bashrc` and different other shell config files if `bash` isn't your shell of choice.

Once `umask 000` is run, most files will be created with read/write perms so that all users can read/write each others files.

Of course, if you are using a sort of HPC, where many unrelated groups use the same cluster this won't work and then you would either use groups instead of making files read/write by all, with possibly `setgid` bit preset or using `ACL`. In any such environments there are always sysadmins so you can ask them how to setup a shared cache for your team and they will know what to do.

Additionally, recently some of these applications added tools to do the cleanup, e.g. for `conda` and `pip`:

```
conda clean --all -f -y
pip cache purge
```

General disk usage

Of course, sooner or later, your partition will get bigger and bigger, and you will probably want to understand where data is leaking. Typically you will need to find the users who contribute to the most of data consumption and ask them to do some cleanups.

So for example to find which users consume the most disk run:

```
sudo du -ahd1 /home/* | sort -rh
```

it will sort the data by the worst offenders. If you want to help them out you could go into their dirs and analyse the data a level deeper:

```
sudo du -ahd1 /home/*/* | sort -rh
```

or for a specific user `foo`:

```
sudo du -ahd1 /home/foo/* | sort -rh
```

You could also set disk usage quotas but usually this doesn't work too well, because depending on the workflows of your company some users need to generate a lot more data then others, so they shouldn't be punished for that with inability to do their work and have their job crash - which could have been run for many hours and all that work will be lost - so at the end of the day the company will be paying for the lost time.

Getting users to be aware of them using too much disk space can be a very difficult task.

Partition inodes limit

Also beware of inode usage, on some shared partitions on HPCs I have seen more than once cases where a job crashed not because there was no disk space left, but because the job used up the last inodes and the whole thing crashed.

To see inode usage, use `df -i`:

```
$ /bin/df -hi
Filesystem  Inodes  IUsed  IFree  IUse%  Mounted on
tmpfs        16M    1.9K    16M     1%    /run
/dev/sda1    59M    4.1M    55M     7%    /
```

`-h` formats huge numbers into human-readable strings.

So here you can see the the `/` partition is using 7% of the total possible inodes.

Depending on the type of filesystem in some cases it's possible to add more inodes whereas in other cases it's not possible.

So as part of your monitoring of disk space you also need to monitor inode usage as a critical resource.

`/tmp` on compute nodes

Normally compute nodes will use `/tmp/` for temp files. The problem is on most set ups `/tmp` resides on the tiny `/` filesystem of each node (often <100GB) and since `/tmp/` only gets reset on reboot, this doesn't get cleaned up between SLURM jobs and this leads to `/tmp` running out of space and so when you try to run something that let's say untars a file you're likely to run into:

```
OSERROR: [Errno 28] No space left on device
```

The solution is to set in your SLURM launcher script.

```
export TMPDIR=/scratch
```

Now, the slurm job will use a much larger `/scratch` instead of `/tmp`, so plenty of temp space to write too.

footnote: while `/scratch` is quite common - the mounted local SSD disk mount point could be named anything, e.g. `/localssd` - it should be easy to see the right path by running `df` on one of the compute nodes.

You can also arrange for the SLURM setup to automatically clean up such folders on job's termination.

How to find users whose checkpoints consume a lot of disk space

Do you have a problem when your team trains models and you constantly have to buy more storage because huge model checkpoints aren't being offloaded to bucket storage fast enough?

Here is a one-liner that will recursively analyze a path of your choice, find all the checkpoints, sum up their sizes and print the totals sorted by the biggest user, so that you could tell them to clean up their act :) Just edit `/mypath` to the actual path

```
find /mypath/ -type f -regextype posix-egrep -regex ".*\.(pt|pth|ckpt|safetensors)$" | \
perl -nle 'chomp; ($uid,$size)=(stat($_))[4,7]; $x{$uid}+=$size;
END { map { printf qq[%-10s: %7.1fTB\n], (getpwuid($_))[0], $x{$_}/2**40 }
```

```
sort { $x{$b} <=> $x{$a} } keys %x ',
```

gives:

```
user_a    :    2.5TB
user_c    :    1.6TB
user_b    :    1.2TB
```

Of course, you can change the regex to match other patterns or you can remove it altogether to measure all files:

```
find /mypath/ -type f | \
perl -nle 'chomp; ($uid,$size)=(stat($_))[4,7]; $x{$uid}+=$size;
END { map { printf qq[%-10s: %7.1fTB\n], (getpwuid($_))[0], $x{$_}/2**40 }
sort { $x{$b} <=> $x{$a} } keys %x }'
```

If you want to exclude some sub-dirs efficiently:

```
find /mypath/ -regextype posix-egrep \
-type d -regex "/mypath/(exclude_a|exclude_b|exclude_c)/*" -prune -o \
-type f -regex ".*\.(pt|pth|ckpt|safetensors)$" | \
perl -nle 'chomp; ($uid,$size)=(stat($_))[4,7]; $x{$uid}+=$size;
END { map { printf qq[%-10s: %7.1fTB\n], (getpwuid($_))[0], $x{$_}/2**40 }
sort { $x{$b} <=> $x{$a} } keys %x }'
```

hint: the second line tells `find` to skip folders matching the `/mypath/(exclude_a|exclude_b|exclude_c)/*` regex. Adapt to your use case as needed.

How to automatically delete old checkpoints

Continuing the item from above, if you want to automatically delete old checkpoints instead (e.g. those older than 30 days).

First try to ensure the candidates are indeed good to delete:

```
find /mypath/ -regextype posix-egrep -regex ".*\.(pt|pth|ckpt|safetensors)$" -mtime +30
```

and when you feel it's safe to delete, only then add `rm`

```
find /mypath/ -regextype posix-egrep -regex ".*\.(pt|pth|ckpt|safetensors)$" -mtime +30 -exec rm {} +
```

fio benchmark results for hope on 2023-12-20-14:37:02

partition /mnt/nvme0/fio/fio-test

- filesize=16k read

lat msec	bw MBps	IOPS	jobs
4.0	1006.3	257614	16

- filesize=16k write

lat msec	bw MBps	IOPS	jobs
3.2	1239.1	317200	16

- filesize=1m read

lat msec	bw MBps	IOPS	jobs
1.7	2400.1	614419	16

- filesize=1m write

lat msec	bw MBps	IOPS	jobs
2.1	1940.5	496765	16

- filesize=1g read

lat msec	bw MBps	IOPS	jobs
1.4	2762.0	707062	16

- filesize=1g write

lat msec	bw MBps	IOPS	jobs
2.1	1943.9	497638	16

Inter-node and Intra-Node Networking Hardware

Subsections:

- [Communication Patterns](#)
- [Network Debug](#)
- [Network Benchmarks](#)

Introduction

It's not enough to buy/rent expensive accelerators to train and infer models fast. You need to ensure that your [storage IO](#), [CPU](#) and network are fast enough to "feed the accelerator furnace". If this is not ensured then the expensive accelerators will be underutilized leading to lost \$\$, slower training time and inference throughput. While it can be any other of the mentioned components, the network is often the bottleneck during the training (assume your DataLoader is fast).

If your model fits on a single accelerator, you have little to worry about. But nowadays most models require several accelerators to load and LLM/VLM models require multiple compute nodes for training and some even for inference.

Most compute nodes contain 8 accelerators, some 4, others 16, and even more accelerators and recently there are some that have one super-accelerator per node.

When the model spans several accelerators and doesn't leave a single node all you need to worry about is fast [Intra-node networking](#). As soon as the model requires several nodes, which is often the case for training as one can use multiple replicas to parallelize and speed up the training, then fast [Inter-node networking](#) becomes the key.

This article covers both types of networking hardware, reports their theoretical and effective bandwidths and explains how they inter-play with each other.

Glossary and concepts

You can safely ignore the many concepts and abbreviations listed here until you need them and then return here.

- ALU: Arithmetic Logic Units
- AR: Adaptive Routing (but also could mean Aggregation Router)
- DMA: Direct Memory Access
- EFA: Elastic Fabric Adapter
- HCA: Host Channel Adapter
- IB: Infiniband
- MFU: Model Flops Utilization (e.g. $mfu=0.5$ at half-precision on A100 comes from getting 156TFLOPs, because peak half-precision spec is 312TFLOPS, and thus $156/312=0.5$)
- NIC: Network Interface Card
- OPA: Omni-Path Architecture
- OPX: Omni-Path Express
- OSFP: Octal Small Form Factor Pluggable (transceiver)
- RDMA: Remote Direct Memory Access
- RoCE: RDMA over Converged Ethernet
- RoE: RDMA over Ethernet
- SHARP: Scalable Hierarchical Aggregation Reduction Protocol
- VPI: Virtual Protocol Interconnect
- xGMI: Socket to Socket Global Memory Interface

Speed-related:

- Unidirectional: a transmission from one point to another in one direction A -> B
- Bi-directional, Duplex: a transmission from one point to another in both directions A <-> B, typically 2x speed of unidirectional
- GBps, GB/s: Gigabytes per secs (1GBps = 8Gbps) transferred in a channel
- GT/s: GigaTransfers per second - the number of operations transferring data that occur in each second.
- Gbps, Gb/s: Gigabits per secs (1Gbps = 1/8GBps) transferred in a channel
- Bisection Width: minimum number of links cut to divide the network into two parts (not necessarily equal). The bandwidth of those links is known as Bisection Bandwidth - which is often used as a metric for real network bandwidth). Sometimes it's referred to as the worst-case network capacity. Here is a [good answer](#) that explains this and related concepts, but it's unlikely you need to understand this other than knowing what is being meant, as chances are your cluster's topology has already been done by the provider.
- Adaptive Routing improves Static routing to enable out of order packets on the network. Packets are load balanced at each switch to better distribute the network workload.
- [Remote Direct Memory Access](#)

footnote: In the following sections pay close attention that 1GBps = 8Gbps.

Unidirectional vs Bidirectional (Duplex)

Most benchmarking / bandwidth measurement tools will report a unidirectional bandwidth. So be careful when you look at unidirectional vs. bidirectional (duplex) speeds. Typically the latter is ~2x faster.

If you measure the bandwidth on your setup and it's about 40% of the advertised speed, carefully check if the advertised speed said duplex and if so half that and then your measured bandwidth should now be about 80% which is expected.

case study: for a while I couldn't understand why when I run the nccl-tests all_reduce benchmark on an A100 node with advertised 600GBps intra-node speed I was getting only 235GBps (40%) until Horace He kindly pointed out that I should be looking at unidirectional speed which is 300GBps, and then I get 80% of the theoretical spec which checks out.

Cluster networks

Each node of the cluster has 3 networks, each running at a very different speed from each other.

1. [Frontend](#)
2. [Backend](#)
3. [Out-of-band](#)

Frontend networking

Frontend networking is typically for the Internet connection (e.g. downloading python packages and offloading to the cloud storage), distributed network storage (e.g. checkpoints and datasets) and orchestration (e.g. SLURM and Kubernetes). As of this writing a typical node is likely to have a single 100-400Gbps connection.

footnote: not all clusters will have external Internet connection available, e.g. many HPC environments only provide external access via special cpu-only nodes.

Backend networking

Backend networking is to perform GPU-to-GPU connectivity which allows training and inference to scale to multiple accelerators (e.g. all-reduce, all-gather and other collective comms). This is the most important part of the AI cluster. Typically this would be either an [Infiniband](#) or [RoCEv2 Ethernet](#). It then breaks down into [intra-node networking](#) and [inter-node networking](#) - the GPUs on the same node typically can communicate with each other at faster speed than with GPUs on other nodes. Here the typical top speeds as of this writing would be around 5600Gbps for intra-node and 3200Gbps per node for inter-node networking. There will be at least one backend connection per accelerator and at times there can be multiple connections per accelerator, especially if low bandwidth NICs are used.

footnote: not all providers will match the industry's standard networking speeds - on some the inter-node networking speed could be up to 10x slower. So always check what you get.

Out-Of-Band networking

Out-Of-Band (OOB) networking is used for bootstrapping backend networking, monitoring node's health, remote re-imaging of the nodes, etc. It typically uses a single slow 1Gbps ethernet connection.

RDMA networking

Remote Direct Memory Access is like DMA (Direct Memory Access) on the node, but across nodes. It allows data exchange between nodes w/o the overhead using the local processor, OS kernel and caches, which is what TCP/IP uses. The 3 main implementations are:

1. Infiniband
2. RDMA over Converged Ethernet (RoCE) (IB or UDP-based RDMA)
3. iWARP (TCP-based RDMA)

Here is a [good overview article](#).

Intra-node networking

There are multiple platforms/solutions out there that provide intra-node networking:

1. Generic: [PCIe](#)
2. NVIDIA: [NVLink](#) and [NVSwitch](#)
3. AMD: [Infinity Fabric](#)
4. Intel: [Gaudi2](#), [Gaudi3](#)

All-to-all bandwidth

Here is intra-node unidirectional theoretical all-to-all peak bandwidth cross-comparison for current solutions sorted by bandwidth:

Interconnect	Accelerator	GBps
NVIDIA NVLink 5	B200, B*	900.0
Intel	Gaudi3	600.0
NVIDIA NVLink 4	H100, H*	450.0
AMD XGMI	MI325X	448.0
AMD XGMI	MI300X	448.0
AMD XGMI	MI250X	350.0
NVIDIA NVLink 3	A100	300.0
Intel	Gaudi2	300.0
PCIe 5		63.0
PCIe 4		31.0

Notes:

- NVSwitch operates at the same speed as NVLink of that generation. See [NVSwitch](#).
- Pay close attention to when the spec says unidirectional vs bidirectional (duplex) speeds - if you read an online spec and it doesn't explicitly declare the directionality - look for an answer. I had to research many docs to figure it out in some of the tables below as some vendors omit this crucial information in the published specs. I even had to edit a few wiki pages to add the missing information. Remember that for the vendors the bigger, the better so almost always they will use the duplex number, which is typically 2x bigger than the unidirectional one.

Peer-to-peer bandwidth

Some vendors have their all-to-all and peer-to-peer (GPU-to-GPU) bandwidth the same, while others don't. For example, AMD MI3* are 64GBps GPU-to-GPU (peer-to-peer), but 448GBps in total on a board of 8 accelerators, since $64*7=448$.

Here is the intra-node unidirectional theoretical peer-to-peer peak bandwidth cross-comparison for current solutions sorted by bandwidth:

Interconnect	Accelerator	GBps
NVIDIA NVLink 5	B200, B*	900.0
Intel	Gaudi3	600.0
NVIDIA NVLink 4	H100, H*	450.0
NVIDIA NVLink 3	A100	300.0
Intel	Gaudi2	300.0
AMD XGMI	MI325X	64.0
AMD XGMI	MI300X	64.0
AMD XGMI	MI250X	50.0

When peer-to-peer bandwidth is much lower than all-to-all it means that if you don't use all of the accelerators on the node by the same application, you will end up with a much lower bandwidth and your application will have a performance impact if the accelerators have to communicate between each others.

To validate this the [all_reduce_bench.py](#) was run on a 8x GPU AMD MI300X node with a 4GB payload and the `busbw` measurements were:

- 2 GPUs: 47.671 GBps
- 8 GPUs: 312.912 GBps

i.e. 2 GPUs performed 6.5x slower than 8.

So if you have to deploy TP=2, TP=4, or ZeRO-DP/FSDP over 2 or 4 GPUs, be it training or inference, the network will become a bottleneck. If you use TP=1 or TP=8 or ZeRO-DP/FSDP over 8 GPUs, or DP over 1-GPU replicas there is no problem. (If you're not sure what TP/ZeRO-DP/DP mean please see [model-parallelism](#).)

You will find the details analysis of each technology in the following sections.

PCIe

[PCIe](#) is a high-speed serial computer expansion bus standard that can be found even on the cheapest computer desktop.

Interconnect	Lane/Direction	Lanes	Unidirection	Duplex
PCIe 4	~2.0 GBps	16	31 GBps	62 GBps
PCIe 5	~4.0 GBps	16	63 GBps	126 GBps
PCIe 6	~7.5 GBps	16	121 GBps	242 GBps
PCIe 7	~15.0 GBps	16	242 GBps	484 GBps

If one compares the latest generations of different intra-node networking technologies (see the following sections) PCIe is usually an order of magnitude behind.

NVLink

- [NVLink](#) is a wire-based serial multi-lane near-range communications link developed by Nvidia. Here is the [What Is NVLink](#) blog post with more background on it.

I found the NVLink wiki page to be quite difficult to follow, so I will try to help bring clarity into this. And I'm pretty sure as of this writing some of the numbers on that wiki page are bogus and it doesn't look like NVIDIA maintains that page.

Effective payload rate of intra-node GPU-to-GPU communication hardware:

Interconnect	Lane/Direction	Lanes	Links	Unidirection	Duplex	GPU
NVLink 1	2.50 GBps	8	4	80 GBps	160 GBps	P100
NVLink 2	3.125 GBps	8	6	150 GBps	300 GBps	V100
NVLink 3	6.25 GBps	4	12	300 GBps	600 GBps	A100
NVLink 4	12.50 GBps	2	18	450 GBps	900 GBps	H100, H200, GH200
NVLink 5	25.00 GBps	2	18	900 GBps	1800 GBps	B200, B*, GB*

There is a good overview of evolution of NVLink (1 to 4) [here](#).

The largest PCIe 16x slot has 16 lanes. Smaller slots have less lanes, 1x == 1 lane.

NVIDIA Hopper nodes typically come equipped with PCIe 5 and NVLink 4. So there NVLink is 7x faster than PCIe.

NVIDIA Blackwell nodes will be equipped with PCIe 5 and NVLink 5. So there NVLink will be 14x faster than PCIe.

Let's look at several examples of nodes and correlate the theory with reality.

If you use multiple GPUs the way cards are inter-connected can have a huge impact on the total training time. If the GPUs are on the same physical node, you can run:

```
nvidia-smi topo -m
```

and it will tell you how the GPUs are inter-connected.

On a machine with dual-GPU and which are connected with NVLink, you will most likely see something like:

	GPU0	GPU1	CPU Affinity	NUMA Affinity
GPU0	X	NV2	0-23	N/A
GPU1	NV2	X	0-23	N/A

on a different machine w/o NVLink you may see:

	GPU0	GPU1	CPU Affinity	NUMA Affinity
GPU0	X	PHB	0-11	N/A
GPU1	PHB	X	0-11	N/A

The report includes this legend:

```

X = Self
SYS = Connection traversing PCIe as well as the SMP interconnect between NUMA nodes (e.g., QPI/UPI)
NODE = Connection traversing PCIe as well as the interconnect between PCIe Host Bridges within a NUMA
node
PHB = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
PXB = Connection traversing multiple PCIe bridges (without traversing the PCIe Host Bridge)
PIX = Connection traversing at most a single PCIe bridge
NV# = Connection traversing a bonded set of # NVLinks

```

So the first report **NV2** tells us the GPUs are interconnected with 2 NVLinks, and the second report **PHB** we have a typical consumer-level PCIe+Bridge setup.

Check what type of connectivity you have on your setup. Some of these will make the communication between cards faster (e.g. NVLink), others slower (e.g. PHB).

Depending on the type of scalability solution used, the connectivity speed could have a major or a minor impact. If the GPUs need to sync rarely, as in DDP, the impact of a slower connection will be less significant. If the GPUs need to send messages to each other often, as in ZeRO-DP, then faster connectivity becomes super important to achieve faster training.

Now, let's look at the topology of the A100 and H100 nodes:

- A100 topology:

\$ nvidia-smi topo -m										
	GPU0	GPU1	GPU2	GPU3	GPU4	GPU5	GPU6	GPU7	CPU Affinity	NUMA Affinity
GPU0	X	NV12	0-23	0						
GPU1	NV12	X	NV12	NV12	NV12	NV12	NV12	NV12	0-23	0
GPU2	NV12	NV12	X	NV12	NV12	NV12	NV12	NV12	0-23	0
GPU3	NV12	NV12	NV12	X	NV12	NV12	NV12	NV12	0-23	0
GPU4	NV12	NV12	NV12	NV12	X	NV12	NV12	NV12	24-47	1
GPU5	NV12	NV12	NV12	NV12	NV12	X	NV12	NV12	24-47	1
GPU6	NV12	NV12	NV12	NV12	NV12	NV12	X	NV12	24-47	1
GPU7	NV12	X	24-47	1						

You can see there are 12 NVLinks and 2 NUMA Groups (2 CPUs w/ 24 cores each)

- H100 topology:

```
$ nvidia-smi topo -m
  GPU0  GPU1  GPU2  GPU3  GPU4  GPU5  GPU6  GPU7  CPU Affinity  NUMA Affinity
GPU0   X    NV18  NV18  NV18  NV18  NV18  NV18  0-51      0
GPU1  NV18   X    NV18  NV18  NV18  NV18  NV18  0-51      0
GPU2  NV18  NV18   X    NV18  NV18  NV18  NV18  0-51      0
GPU3  NV18  NV18  NV18   X    NV18  NV18  NV18  0-51      0
GPU4  NV18  NV18  NV18  NV18   X    NV18  NV18  52-103    1
GPU5  NV18  NV18  NV18  NV18  NV18   X    NV18  52-103    1
GPU6  NV18  NV18  NV18  NV18  NV18  NV18   X    NV18  52-103    1
GPU7  NV18  NV18  NV18  NV18  NV18  NV18  NV18   X    52-103    1
```

You can see there are 18 NVLinks and 2 NUMA Groups (2 CPUs w/ 52 cores each)

Of course, other A100 and H100s node reports may vary, e.g. the number of cpu cores is likely to be different.

NVSwitch

[NVSwitch](#) can connect more than 8 GPUs at the speed of [NVLink](#). It's advertised to connect up to 256 GPUs in the future generations of the switch.

The benefit of connecting more than 8 GPUs at the speed of NVLink is that it allows all-to-all GPU communications at a much faster speed than any intra-node hardware can provide. And with ever increasing compute speeds the network is the likely bottleneck leading to underutilized super-expensive GPUs.

For example, in the universe of Tensor Parallelism (Megatron), one doesn't use TP degree of more than 8, because TP is only efficient at NVLink speed. ZeRO-DP (Deepspeed/FSDP) would also run much faster if the whole cluster uses NVLink speed and involves no slow inter-node connections.

NVSwitch is used for intra-node connectivity.

NVSwitch gen 1 came out with V100, gen 2 with A100, gen 3 with H100, and gen 4 with B200 - the speed corresponds to the NVLink version of the same technology.

The [NVIDIA DGX H100](#) has a 3.6 TBps of full-duplex NVLink Network bandwidth provided by 72 NVLinks (NVLink 4). The normal NVLink 4 has 18 NVLinks (0.9 TBps duplex). So this setup has 4 switches ($18*4=72$) and therefore $0.9*4=3.6$ TBps. Note, that this server has 8 GPUs, so here we get a much faster intra-node communications as compared to the standard NVLink 4.0 which provides only 0.9 TBps all-to-all connectivity for 8 GPUs.

NVIDIA DGX A100 has 6 switches of 12 NVLinks for a total of 72.

[DGX H100 SuperPOD](#) combines 32 DGX H100 servers, for a total of 256 GPUs. It looks like here they use only half the NVLinks they used for a single DGX H100, so only 1.8 TBps per node, for a total of 57.6 TBps in total.

Additionally, NVSwitch gen3 and higher comes with [NVIDIA Scalable Hierarchical Aggregation Reduction Protocol \(SHARP\)](#) which can boost both the intra- and inter-node speeds. For example, NCCL is working on NCCL_ALGO=NVL5 which already boosts the intra-node bandwidth above the normal spec and as of this writing work is being done to boost inter-node bandwidth as well.

Recently [GB200 NVL72](#) has been introduced, which uses NVSwitch to put 72 Blackwell GPUs into a single node all inter-connected at NVLink 5 900GBps unidirectional speed. So instead of having a 8-gpu node, now we have a 72-gpu node (even though physically they don't all reside on the same board).

Infinity Fabric / xGMI

AMD MI* Accelerators Intra-node communication is performed by AMD Infinity Fabric, which is also known as xGMI (Socket to Socket Global Memory Interface).

This is AMD's answer to [NVLink](#).

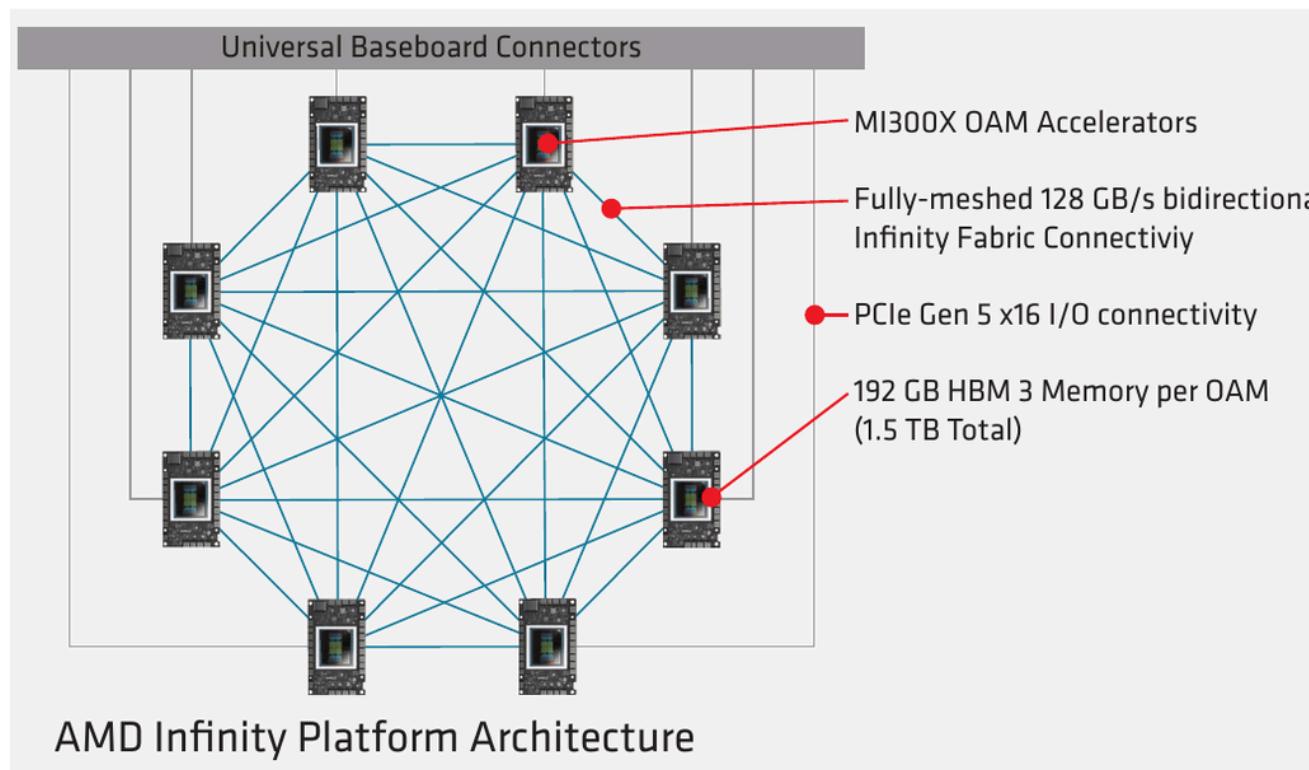
The following is the all-to-all bandwidth.

| Interconnect | Link/Direction | Links | Unidirection | Duplex |

	peer-to-peer	all-to-all	all-to-all
MI325X	64 GBps	7	448 GBps
MI300X	64 GBps	7	448 GBps
MI250X	50 GBps	7	350 GBps
MI355X	??		

The peer-to-peer bandwidth is just that of a single link/direction (the 2nd column). This means that unless you use the whole 8-GPU node in a single process group you will have a 7x slower comms performance. See [Peer-to-peer bandwidth](#) for details.

Other intra-node solutions typically have the same all-to-all and peer-to-peer intra-node bandwidth, so Infinity Fabric appears to be dramatically slower. I suppose that is because these were created mainly for inference, as these slow speeds would dramatically slow down LLM training.



Platform specs:

- [MI250X](#)
- [MI300X](#)
- [MI325X](#)
- MI355X ??

Gaudi2

According to [Gaudi2 spec](#), these nodes provide the same 100GbE RoCE v2 RDMA hardware for inter- and intra-node connectivity (24x 100Gbps per card).

- intra-node: 8x 7x3 NICs - 300Gbps card to card
- inter-node: 8x 1x3 NICS - for a total of 2.4Tbps (300GBps)

Gaudi3

According to [Gaudi3 spec](#), these nodes provide the same setup as Gaudi2 except the cards are 2x faster using 200GbE RoCE v2 RDMA for inter- and intra-node connectivity (24x 200Gbps per card).

- intra-node: 8x 7x3 NICs - 600Gbps card to card
- inter-node: 8x 1x3 NICS - for a total of 4.8Tbps (600GBps)

Inter-node networking

As inter-node hardware used to be about of an order of magnitude slower than intra-node hardware in this universe Gbps are used instead of GBps. (1 GBps = 8 Gbps) (Though as of recent inter-node speeds are almost as fast as [intra-node](#))

When it comes to inter-node networking hardware, there are the well established [InfiniBand](#) from NVIDIA and a few other players, various NVLink-based NVIDIA products and there are many new comers that mainly are coming from compute cloud providers who can't compete on the slim margin renting out someone else's hardware so they build their own (AWS EFA, Google GPUDirect-TCPX), and there are also HPE and Cornelis Networks with recently updated products.

Here is inter-node unidirectional theoretical peak bandwidth cross-comparison for current technologies sorted by total bandwidth of common node setups:

Interconnect	NICs x Gbps	Total GBps	Notes
Intel Gaudi3	24x200	600	
NVIDIA Quantum-2 IB	8x400	400	H100
AWS EFA v2	32x100	400	H100
Intel Gaudi2	24x100	300	
InfiniBand XDR1600	8x200	200	
Intel GPUDirect-TCPX	4x200	100	
HPE Slingshot	4x200	100	
Omni-Path CN100	8x100	100	
AWS EFA v1	4x100	50	
InfiniBand NDR400	4x100	50	
in the future:			
Omni-Path CN5000	8x400	400	Q2-2025
InfiniBand GDR3200	8x400	400	2025
Omni-Path CN6000	8x800	800	2026

Notes:

- these are common/popular node setups - some custom nodes may have a different configuration more often with less NICs and rarely with more NICs. And, yes, AWS EFA v2 puts 32 NICs on each node - that must be a lot of wires.
- Note how the once order-of-magnitude difference between inter- and [intra-node bandwidth](#) is starting to disappear - I have recently rescaled the speeds here from Gbps to GBps.

You will find the details analysis of each technology in the following sections.

InfiniBand

[InfiniBand](#) (IB) has been around for a few decades so there are many available configurations that can be found out there. So that if someone says they have InfiniBand that is insufficient information. What you need to know is the signaling rate and the number of IB links.

InfiniBand is a complete network protocol that implements RDMA (bypasses TCP/IP).

Here are the most recent signaling rates which you are likely to see in the current hardware offerings:

Signaling rate of uni-directional links in Gbps:

Links	EDR	HDR	NDR	XDR	GDR	LDR
1	25	50	100	200	400	800
4	100	200	400	800	1600	3200
8	200	400	800	1600	3200	4800
12	300	600	1200	2400	4800	9600

Notes:

- the GDR is planned in 2025 and LDRs some years later

Latency in usecs:

EDR	HDR	NDR	XDR	GDR	LDR
0.5	0.6	??	??	??	??

?? = NDR and later didn't publish latency data

InfiniBand provides [RDMA](#).

Here are some examples of NVIDIA devices with the fastest IB:

- One configuration of NVIDIA DGX H100 comes with 8x NVIDIA ConnectX-7 (CX7) Ethernet/InfiniBand ports each of 200Gbps, for a total of 1.6 Gbps to connect with other DGX servers.
- For DGX H100 SuperPOD the ConnectX-7s across all 32 DGX servers and associated InfiniBand switches provide 25.6 TBps of full duplex bandwidth for use within the pod or for scaling out the multiple SuperPODs - that is an equivalent of 0.8 TBps per node (6.4Tbps!).
- NVIDIA GB200-based solutions will come with 400Gbps or 800Gbps NDR via Quantum-2 InfiniBand 800G switches (2x400G NDR interfaces)

According to wikipedia while [InfiniBand](#) used to have multiple manufacturers - at the moment it's just Intel (purchased QLogic) and NVIDIA (purchased Mellanox). Also see [InfiniBand Trade Association](#).

Practical links:

- [InfiniBand Utilities](#) (the link could be outdated as it's versioned) - these are useful when debugging an IB setup.

NVIDIA Quantum-2 InfiniBand

[NVIDIA Quantum-2 InfiniBand Platform](#) supports 400Gbps bandwidth per link, provides RDMA, includes in-network computing with [SHARP](#), supports PCIe-5.

The switches can connect 64 devices at 400Gbps.

EFA

[Elastic Fabric Adapter \(EFA\)](#) is a recent inter-node networking technology created by AWS.

- EFA v1 0.4 Tbps (effective 340 Gbps for all_reduce tests) (P4 AWS instances)
- EFA v2 3.2 Tbps (since Q3-2023, P5 AWS instances - 32 NICs!)

Gaudi2 (inter-node)

According to [Gaudi2 spec](#), these nodes provide $3*8=24$ NICs of 100GbE RoCE v2 RDMA for a total of 2.4Tbps of inter-node connectivity with other Gaudi2 nodes.

Gaudi2 (inter-node)

According to [Gaudi3 spec](#), these nodes provide $3*8=24$ NICs of 200GbE RoCE v2 RDMA for a total of 4.8Tbps of inter-node connectivity with other Gaudi2 nodes.

According to [Gaudi2 spec](#), these nodes provide $3*8=24$ NICs of 100GbE RoCE v2 RDMA for a total of 2.4Tbps of inter-node connectivity with other Gaudi2 nodes.

HPE Slingshot interconnect

[HPE Slingshot interconnect](#) seems to be used by HPCs. As of this writing it provides 200Gbps per link. Some HPCs use 4 of those links to build 800Gbps interconnects, and, of course, with more links will deliver a higher overall bandwidth.

GPUDirect-TCPX

GPUDirect-TCPX is a new hardware/software networking stack introduced in A3 instances of GCP. The docs are scarce, but here is [some information](#).

This technology didn't catch on and will be phased out to be replaced with RoCE starting with Blackwell instances at GCP.

Omni-Path

[Omni-Path Architecture](#) (OPA). Originally by Intel, the technology got sold to Cornelis Networks. It's also known as Omni-Path Express (OPX).

case study: I used this technology at JeanZay HPC in France in 2022. It was only 135Gbps and while the vendor tried to fix it a year later it was still the same speed. Hopefully the issue has been resolved and the speed is much faster nowadays. Because it was so slow we had to use [Megatron-Deepspeed](#) for training BLOOM-176B instead of the much easier to use DeepSpeed ZeRO).

As of this writing I see that the product comes with either 100 or 200Gbps bandwidth. So it's unlikely you will see anybody offering this solution for ML workloads, unless they manage to install many NICs perhaps?

[CN-100](<https://github.com/stas00/ml-engineering/blob/master/network/Cornelis%20Omni-Path%20Accelerated%20Host%20Fabric%20Adapter%20CN-100HFA>) 100Gbps NICs have been around for many years now.

[CN5000](#) 400Gbps NICs will be launched by Cornelis Networks in Q2-2025. One upcoming MI300X setup uses 8x of these for 3200Gbps of total unidirectional inter-node bandwidth.

Omni-Path provides [RDMA](#).

Ultra Accelerator Link (UALink)

The [UALink initiative](#) is an attempt to create an open standard to compete with [NVLink](#). Supposedly it'll be based on AMD's [Infinity Fabric](#). As of this writing there is no actual hardware to speak of.

Other essential network technologies

SHARP

NVIDIA [Scalable Hierarchical Aggregation and Reduction Protocol \(SHARP\)](#) - allows performing data reductions and aggregations on the network itself (in-network computing). This is very useful if you do a lot of MPI, NCCL and other network collectives that support SHARP, as those should get their latencies much improved.

To understand the importance of this technology - for all-reduce operations, instead of $2N$ sends, it will only need $N+1$ sends - so for a large N - it almost doubles the effective all-reduce throughput. (N is the number of communicating ranks/gpus). For details see [all-reduce operation compatibility](#) (you'd have to scroll down to get to that section).

Recent NCCL versions will automatically use this technology if it is available.

The SHARP hardware that is part of the NVSwitch or Infiniband switches includes arithmetic logic units (ALU) that perform the compute directly rather than using GPUs. It's said that it can perform math in FP64, FP32, FP16 and BF16 dtypes.

case study: I discovered SHARP accidentally when an H100 intra-node NVLink 4.0 [all-reduce](#) benchmark reported 480GBps for a 4GB payload when the theoretical spec was only 450GBps! We figured out it's because NCCL turned on the new [NVLS](#) algo as it detected Infiniband SHARP. I still don't understand how it clocked speed faster than what the physical medium allows. I'm pretty sure that `busbw` calculation algorithm needs to be adjusted there from $2N$ to $N+1$ to get the real speed. There is a detailed discussion about this [here](#). Bottom line: `busbw` may or may not be giving you the real bandwidth number depending on the `algo` NCCL chose to use, where only when `Ring` algo is used the `busbw` is correct.

Understanding why inter-node network speed is of a huge importance

This is probably one of the most important multi-segment section that you really want to understand well. While it seeks out to show how important the inter-node speed is, to build up the case it'll teach on the way many important training-related concepts.

The basics

First, let's get a bit of a feeling what all those Gbps/GBps practically mean.

If your model is 80B parameter large, and you need to transmit every parameter or a gradient on the network even once in float32 (fp32) format, which requires 4 bytes per parameter, so you need to send $80*4$ 320GB of data, or 2560Gb (*8). If your network's bandwidth is 200Gbps it will take 12.8 seconds ($2560/200$) to transmit. And if you had 1600Gbps network then it'd take only 1.6 seconds. Why does it matter?

1-GPU training

Let's start with a much smaller model of say 2B params, to train it you'd need at least [18 bytes per parameter](#) in mixed

half precision. So 18×2 36GB of memory just for model weights, optimizer states and gradients. Plus you need additional memory for activations and it'll depend on the batch size and sequence length. But with 80GB A100 GPU we can definitely train this model on a single GPU.

We then assume for the moment that the DataLoader is fast enough to be negligible in duration compared to the compute time. And thus we get a close to a perfect MFU (Model FLOPs Utilization):

```
[DL][  compute  ][DL][  compute  ][DL][  compute  ]  
-----> time  
|<--iteration-->||<--iteration-->||<--iteration-->|
```

which means that the GPU just needs to do many matmuls and it'd do it amazing fast. In this situation you get the highest ROI (Return on Investment).

Single node training

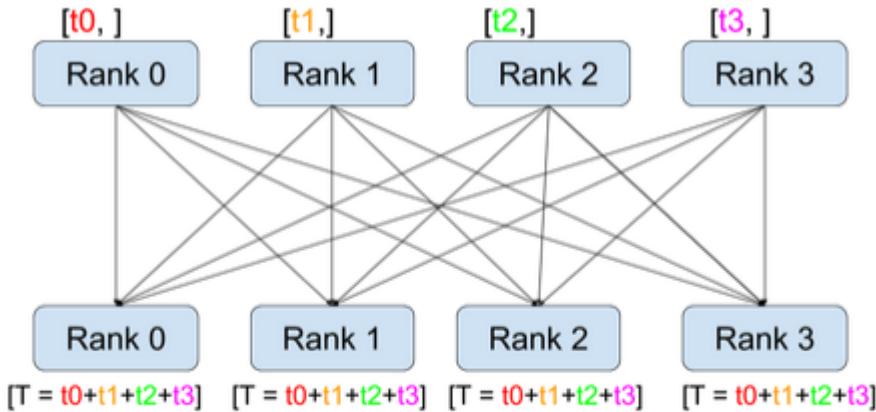
The previous situation was fantastic due to the close to perfect MFU, but you realize that the training on a single GPU is going to take quite some time, since we are in AI race you'd probably want to finish the training sooner than later. So you'd ask - can I train the model on 8 GPUs instead, and the answer would be - yes, of course. With one caveat - at the end of each iteration you'd need to sync the gradients between the 8 processes (each process for a single GPU), so that each participating process of the training can benefit from what the other 7 have learned during the last iteration.

footnote: You could, of course, use less than 8 GPUs, it is just that most NVIDIA GPU-based compute nodes these days have 8 GPUs so why not get the best return on investment.

footnote: in the ideal world the training on 1 GPU for 8 durations of time, should cost the same as training on 8 GPUs for 1 duration of time. That's one would expect to spend the same \$\$ and to finish 8 times faster. But because of data synchronization requirements, this is not the case.

If the experimental model still contains 2B params like in the previous section and grads are in fp32 then the training program needs to send 8GB ($2B * 4B$) of data on every iteration. Moreover, since syncing the gradients requires an [all_reduce_collective](#) collective - it needs to transmit the data twice - the first time sending the gradient data by each GPU, computing the sum of gradients and send this value back to each participating GPU so that each training process will benefit from the learning advancements each of its peers made in the last iteration.

Here is the all-reduce collective visualized:



All-Reduce

([source](#))

So we need to send 8GB twice, which means we need to send 16GB of data.

footnote: and to be exact the 2x comms volume for all-reduce is really $2*(n-1)/n$ where n is the number of participating GPUs. So if n=2, the coefficient is just 1 since $2*(2-1)/2=1$ and 1.75 for n=8 since $2*(8-1)/8=1.75$ and it becomes already very close to 2 at n=64.

footnote: there is also the important issue of latency of the network - which is multiplied several times due to how data is gathered from all participating GPUs. But, given that here we are moving a very large payload the latency contributes a very small overhead and for simplicity can be ignored.

How long will it take to send 16GB of data?

- A100 @ 300GBps: $16/300 = 0.053$ secs
- H100 @ 450GBps: $16/450 = 0.035$ secs

which is incredibly fast!

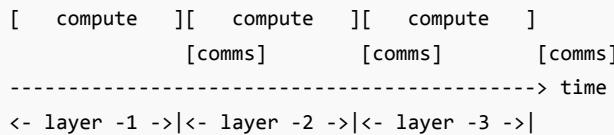
And here is how our timeline will look like:

```
[DL][  compute ][comms][DL][  compute ][comms][DL][  compute ][comms]
-----> time
|<---- iteration ---->||<---- iteration ---->||<---- iteration ---->|
```

oh and this whole synchronization protocol is called DDP ([DistributedDataParallel](#)) in the PyTorch lingo.

Comms and compute overlap

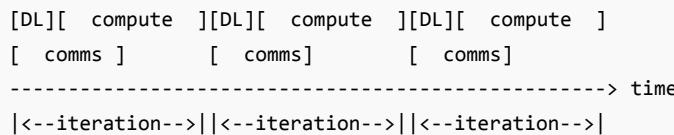
Even with this really fast comms the network still creates a bottleneck and leads to a short idling of the GPUs. To solve this issue the advanced algorithms implement an overlap of comms and compute. Until now we approached the problem as one single transmission, but in reality each model is made of many layers and each layer can transmit the gradients it has computed, while the next layer is computing its gradients. So if you look at the level of the model, what happens in the backward path is:



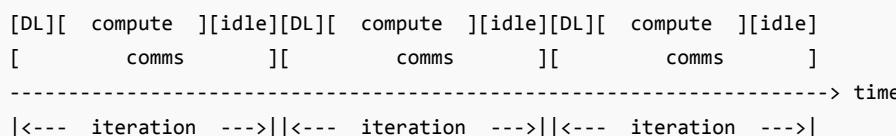
so once the last layer (-1) computed its gradients it all-reduces them while the 2nd to last layer performs its backward, and so on, until the first layer finished with gradients and it finally sends its gradients out.

So now you understand how overlapping works, So we can now update our bigger picture diagram to be:

Now our timing diagram becomes very similar to the diagram we had for a single GPU:



and we hope that comms are faster than DL+compute, since if they aren't faster than we have the following GPU idling gaps:



When comms take longer than compute, the comms part that doesn't overlap is called "exposed communication". Here the compute is blocked waiting for the arrival of the data it needs to continue.

Calculating TFLOPS

Calculating TFLOPS answers the question of how long will it take to perform a compute.

There is a bit of nomenclature confusion here as TFLOPS as the final s sometimes means sec and at other times just ops.

For example, when you read, the [A100 spec](#) the TFLOPS there means TeraFloatingPointOperations per second.

So let's define these abbreviations exactly:

- TFLOPS - TeraFLoatingpointOPerations per Second (another way is TFLOP/s)
- TFLOP - TeraFLoatingpointOPerations (or TFLOPs - lower case s but it's already confusing)

Also see the [wiki page](#) for more clarifications.

For GPT-family of decoder transformers models we can use the math described in this [BLOOM-176 docs](#):

Here is how many TFLOP are processed per second:

```
tflops = model_size_in_B * 4 * 2 * seqlen * global_batch_size / (time_in_sec_per_interation * total_gpus * 1e3)
```

This formula assume one uses [activation recomputation](#) which saves GPU memory while introducing a smallish overhead.

If one doesn't use it then replace 4 with 3 as the model has to do only 1x compute per `forward` and 2x per `backward` (since the grads are calculated twice - once for inputs and once for weights). With activation recomputation the `forward` is done twice and thus you have an additional path which leads to a multiplier of 4 instead of 3

footnote: activation recomputation and gradient checkpointing both refer to the same technique.

so let's remove the time component, which will give us the total TFLOP

```
tflop = model_size_in_B * 4 * 2 * seqlen * global_batch_size / (total_gpus * 1e3)
```

So let's say we have:

- `seqlen=2048` (sequence length)
- `global_batch_size=16`

and we already defined:

- `total_gpus=8`
- `model_size_in_B=2`

This gives us:

```
tflops = 2 * 4 * 2 * 2048 * 16 / (8 * 1e3) = 65.536 TFLOP
```

So if we do a mixed half-precision training and most of the operations are done in half-precision then we can roughly say that we do [312 TFLOPS on A100](#) and usually a well optimized framework on a well-tuned hardware will do at least 50% MFU - that is it'll be able to compute at about 1/2 peak performance.

footnote: It's a ~3x [989 TFLOPS on H100](#) (scroll to the end) and also it shows a misleading 2x numbers for sparsity so you have to mentally divide it by 2.

So continuing this train of thought it means that the setup will have about 156TFLOPS - and so it'll take 0.42 secs to process a single iteration (2x `forward` and 2x `backward` compute) if we ignore the overhead of the DataLoader (which we hope is close to instant).

Earlier we said that a typical A100 node has an intra-node NVLink connection of 300GBps, and thus we said that to send 16GB of grads will take $16/300 = 0.053$ secs.

And we measured our compute to be 0.42 secs, so here the network isn't a bottleneck as $0.42 > 0.053$ so the compute will be slower than communication.

You can now do several thought experiments - for example if you halve the batch size or the sequence length you will halve the compute time.

footnote: this is a very rough suggestion since GPUs work the fastest when the matrices they multiple are huge. But this is good enough for a simplified thought experiment we are having here. In reality halving the dimension will not halve the compute time.

OK, but hopefully at this point it's quite clear that if you remain at the boundaries of a single node, you don't need to worry about your GPUs idling.

But what if you want to speed up the training even more and throw say 4x 8-GPU nodes at it. (and of course you don't have a choice but to use multiple nodes if you have a much larger model). Suddenly, the comms can become an even bigger bottleneck.

Multiple node training

So here we are continuing with the idea of 2B param model and we will now use 32 GPUs across 4 nodes to speed up the training even more.

While each group of 8 GPUs is still connected with super-fast NVLink technology, the inter-node connections are usually in an order of magnitude slower.

Let's say you have a 200Gbps connection. Let's repeat the math from the previous section of how long it'll take to reduce 16GB of gradients.

16GB is 128Gb, and so at 200Gbps this will take 0.64 seconds.

And if stick to the compute taking 0.42 seconds, here we end up with comms taking longer than compute since $0.64 > 0.42$.

Let's bring both use cases together:

nodes	comms	compute	comms is a bottleneck
1	0.027	0.42	no
4	0.64	0.42	yes

on this 200Gbps inter-node setup the comms are 23x slower than the same performed on an intra-node NVLink connections.

In this case even though we still have the much faster NVLink connection, we don't really benefit from it, since the whole ensemble communicates at the speed of the slowest link. And that slowest link is the inter-node connection.

So in this particular situation if you were able to get a 400Gbps inter-node the speed would double and the comms will finish in 0.32 secs and thus will be faster than that 0.42 secs the compute would take.

footnote: you will never be able to get the advertised speed fully on the application level, so if it's advertised as 400Gbps in the best case expect to get 320Gbps (about 80%). So make sure to take this into the account as well. Moreover, depending on the payload of each collective - the smaller the payload the smaller the actual network throughput will be.

And remember this was all handling a pretty tiny as considered these days 2B param model.

Now do the same math with 20B and 200B parameter model and you will see that you need to have a much much faster inter-node connectivity to efficiently scale.

Large model training

Of course, when we train large models we don't use DDP, because we simply can't fit the whole model on a single GPU so various other techniques are used. The details are discussed in a dedicated chapter on [Model Parallelism](#), but the only important thing to understand immediately is that all scalability techniques incur a much larger comms overhead, because they all need to communicate a lot more than just gradients. and therefore the amount of traffic on the network can easily grow 3x and more as compared to the DDP protocol overhead we have been exploring so far.

It can be difficult to do even approximate math as we did in this chapter, because the actual compute time depends on the efficiency of the chosen framework, how well it was tuned, how fast the DataLoader can feed the batches and many other things, therefore there is no standard MFU that one can use in the math and you will discover your MFU when you configure and run the first few steps of the large model training. and then you will read the [Performance chapters](#) and improve your MFU even more.

As I have shown in these sections it should be possible to be able to do a back-of-envelope calculations once you understand the specific scalability technique and its networking costs, so that you could know ahead of time which

Inter-node network speed you need to require from your acquisition manager. Of course, you also need to understand the particular model architecture and calculate how many TFLOP it will take to do a single iteration.

Important nuances

Real network throughput

The network throughput in the advertised spec and the actual throughput will never be the same. In the best case you can expect about 80-90% of the advertised spec.

Then the network throughput will depend on the size of payload being sent during each communication. The higher the payload the higher the throughput will be.

Let's demonstrate this using [nccl-tests](#) on a single A100 node

```
$ ./build/all_reduce_perf -b 32k -e 16G -f 2 -g 8 -n 50
[...]
      size   time   algbw   busbw
      (B)   (us)   (GB/s)   (GB/s)
  32_768   43.83   0.75   1.31
  65_536   46.80   1.40   2.45
 131_072   51.76   2.53   4.43
 262_144   61.38   4.27   7.47
 524_288   80.40   6.52  11.41
 1048_576  101.9   10.29  18.00
 2097_152  101.4   20.68  36.18
 4_194_304  101.5   41.33  72.33
 8_388_608  133.5   62.82 109.93
 16_777_216  276.6   60.66 106.16
 33_554_432  424.0   79.14 138.49
 67_108_864  684.6   98.02 171.54
 134_217_728 1327.6  101.10 176.92
 268_435_456 2420.6  110.90 194.07
 536_870_912 4218.4  127.27 222.72
 1_073_741_824 8203.9  130.88 229.04
 2_147_483_648 16240   132.23 231.41
 4_294_967_296 32136   133.65 233.88
 8_589_934_592 64074   134.06 234.61
 17_179_869_184 127997  134.22 234.89
```

footnote: I massaged the output to remove unwanted columns and made the size more human readable

This benchmark run an `all_reduce` collective for various payload sizes from 32KB to 16GB. The value that we care about is the `busbw` - this column tells us the real network throughput as explained [here](#).

As you can see for payloads smaller than 8MB the throughput is very low - and it starts saturating around payload size of 536MB. It's mostly because of latency. Reducing a single 4GB payload is much faster than 1000x 4MB payloads.

Here is a benchmark that demonstrates that: [all_reduce_latency_comp.py](#). Let's run it on the same A100 node:

```
$ python -u -m torch.distributed.run --nproc_per_node=8 all_reduce_latency_comp.py
```

```
----- 1x 4.0GB -----
busbw: 1257.165 Gbps

----- 1000x 0.004GB -----
busbw: 374.391 Gbps
```

It's easy to see that it's about 3x slower in this particular case to send the same payload but in 1000 smaller chunks.

So when you calculate how long does it take to `all_reduce` a given payload size, you need to use the corresponding `busbw` entry (after of course you have run this benchmark on your particular hardware/environment).

Figuring out the payload can be tricky since it'd depend on the implementation of the framework. Some implementations will reduce each weight's gradient alone which obvious would lead to a very small payload and the network will be very slow. Other implementations bucket multiple gradients together before reducing those, increasing the payload and minimizing the latency impact.

But let's go back to the benchmark results table. This test was done on an A100 node that runs NVLink advertised as uni-directional 300GBs so we get about 78% of the theoretical speed with 17GB payload and more than that the benchmark crashes. It can be seen from the last few rows of the table that not much more can be squeezed.

We can also run [p2pBandwidthLatencyTest](#) which performs a low-level p2p benchmark:

```
./p2pBandwidthLatencyTest
[...]
Unidirectional P2P=Enabled Bandwidth (P2P Writes) Matrix (GB/s)
D\D   0     1     2     3     4     5     6     7
0 1581.48 274.55 275.92 272.02 275.35 275.28 273.62 273.20
1 274.70 1581.48 275.33 272.83 275.38 273.70 273.45 273.70
2 274.81 276.90 1594.39 272.66 275.39 275.79 273.97 273.94
3 273.25 274.87 272.12 1545.50 274.38 274.37 274.22 274.38
4 274.24 275.15 273.44 271.57 1584.69 275.76 275.04 273.49
5 274.37 275.77 273.53 270.84 274.59 1583.08 276.04 273.74
6 275.61 274.86 275.47 273.19 272.58 275.69 1586.29 274.76
7 275.26 275.46 275.49 273.61 275.50 273.28 272.24 1591.14
[...]
```

As you can see in the Unidirectional section of the report we do get 274 GBps out of the advertised 300GBps (~91%).

Please note that when I re-run this same test on H100s (NVLink 4.0) I got a much worse efficiency:

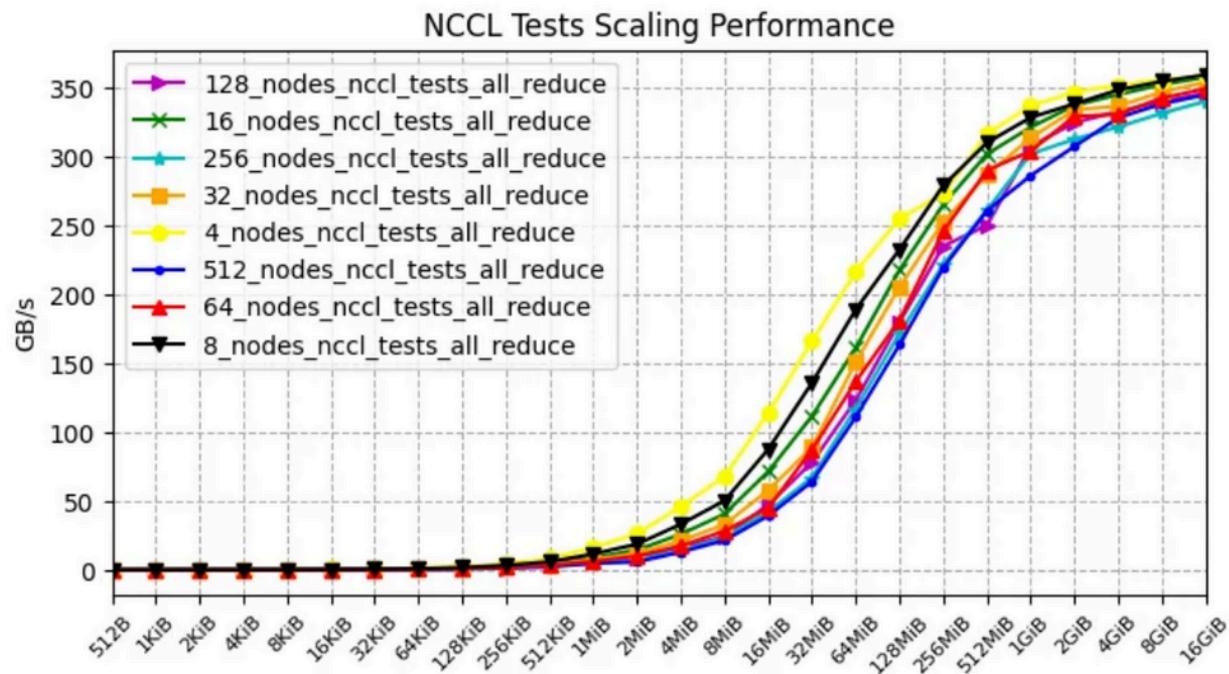
```
Unidirectional P2P=Enabled Bandwidth (P2P Writes) Matrix (GB/s)
D\D   0     1     2     3     4     5     6     7
0 2494.51 364.13 375.99 378.03 376.77 376.71 374.85 375.66
1 375.18 2533.95 376.08 374.98 376.21 375.96 375.76 375.12
2 363.43 393.28 2532.67 376.35 377.14 376.47 375.76 375.48
3 369.90 375.92 393.63 2525.38 376.58 375.88 376.13 377.01
4 376.20 376.28 375.20 393.52 2526.02 375.82 375.05 376.10
5 376.26 376.60 375.54 375.52 376.81 2521.18 376.37 376.60
6 374.31 376.19 376.80 376.32 376.83 376.44 2529.85 376.39
7 376.17 376.49 376.53 374.95 376.30 376.82 375.71 2519.78
```

So 376GBps out of 450GBps is 83% (not very good).

Bottom line - in this particular setup:

1. if you have huge payloads you will be able to use about 80% of the advertised 300GBps
2. if the payload of each communication is smallish it could be far far lower.

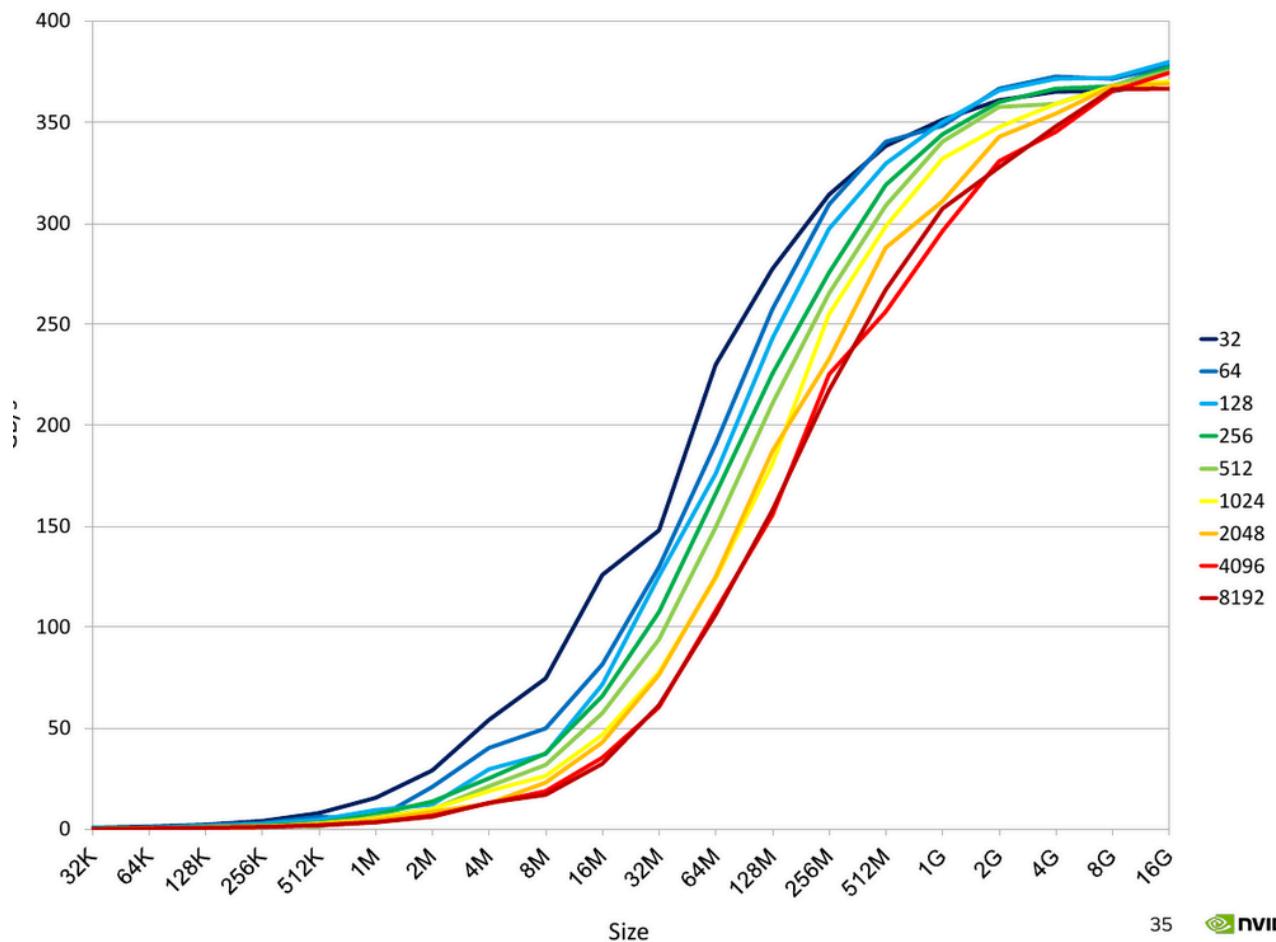
The following plot demonstrates how the actual bandwidth changes for all-reduce with the size of the message and the number of participating nodes (4 to 512 nodes):



[\(source\)](#)

And here is a similar plot, but using NVLSTree algo, which helps to reach an even better performance on H100s (4 to 1024 nodes):

NCCL Allreduce BusBW
ALGO=NVLSTree, 32 to 8K GPUs



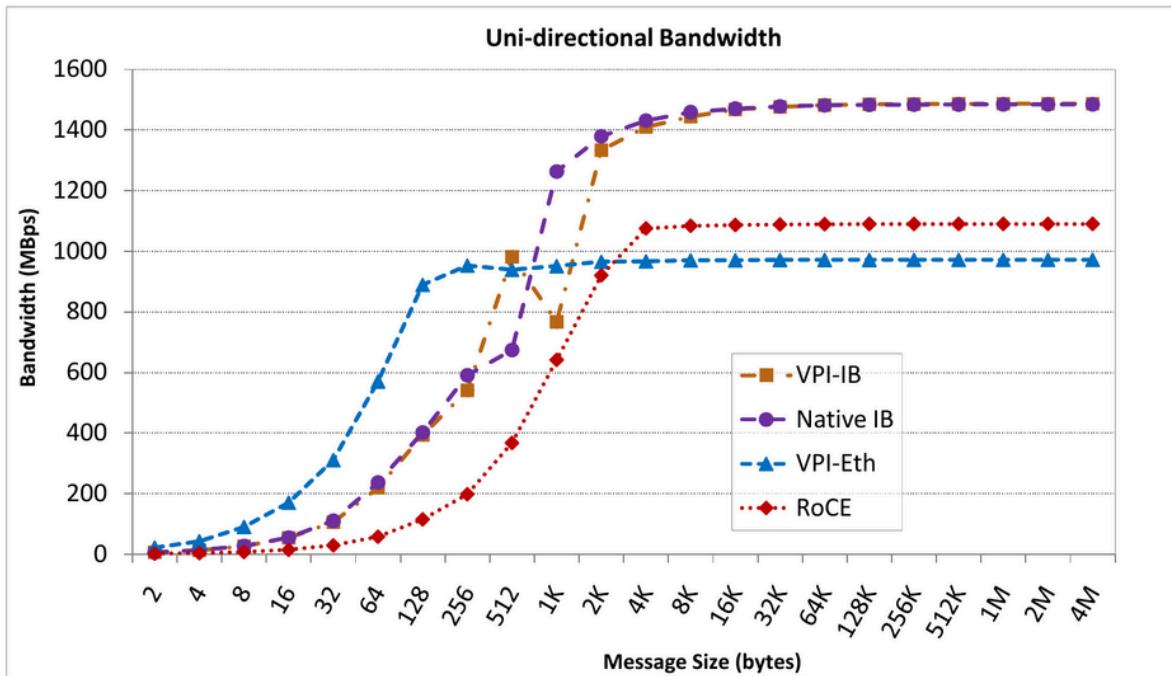
35

 NVIDIA

[source](#)

Here is another similar plot but it compares the message sizes and several networks:

Low-level Uni-directional Bandwidth Measurements



ConnectX-DDR: 2.4 GHz Quad-core (Nehalem) Intel with IB and 10GE switches

CCGrid '11

113

[\(source\)](#)

That last plot is from 2011, and the former ones are from 2024 - comparing these you can appreciate how much faster the networks have become and how much bigger messages are being sent.

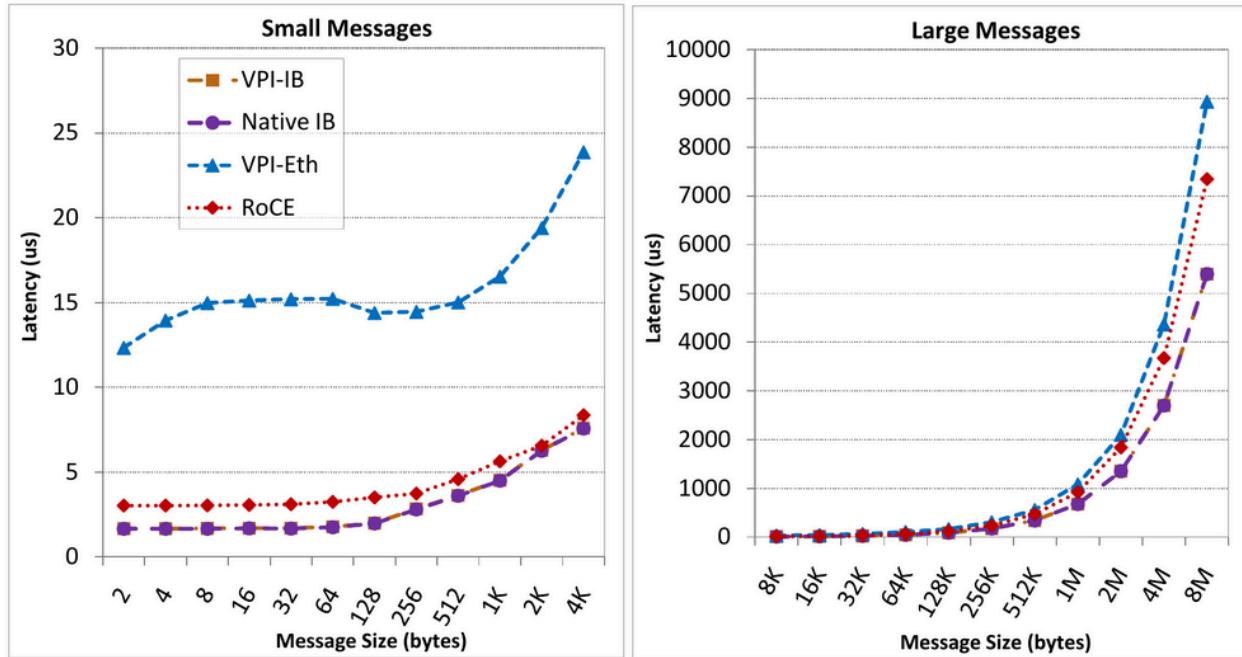
Another tool for bandwidth measurements on NVIDIA GPUs is [NVIDIA/nvbandwidth](#).

Latency

Latency tells us how long it takes to send or receive a message. It has an inverse relationship with throughput - the faster is the throughput the lower is the latency.

Here is an old but good plot demonstrating how the latencies change with message size and the type of the network:

Low-level Latency Measurements



ConnectX-DDR: 2.4 GHz Quad-core (Nehalem) Intel with IB and 10GE switches

RoCE has a slight overhead compared to native IB because it operates at a slower clock rate (required to support only a 10Gbps link for Ethernet, as compared to a 32Gbps link for IB)

CCGrid '11

112

[\(source\)](#)

Typically the more "hops" the message has to travel, the bigger the latency. 2 accelerators residing on the same node and connected directly to each other (e.g., NVLink) will have the least amount of latency. If their communication path traverses a PCIe switch the latency will be bigger. 2 accelerators residing on 2 different nodes sharing a single switch will have a bigger latency because there is a switch to traverse. The further they get away from each other, the more switches the message has to travel through, the bigger the latency.

Proprietary network hardware and NCCL

Proprietary network hardware vendors like AWS (EFA) don't disclose their secrets and therefore the public libraries like [nccl](#) cannot support those out of the box. These vendors have to supply their own versions of the network collective libraries to be used by users of their hardware.

Originally proprietary hardware vendors used the trick of telling the users to use `LD_LIBRARY_PATH` and/or `LD_PRELOAD` to dynamically overload `libncc1.so` to get their custom version loaded into PyTorch or another framework. But recently NCCL developed a [NCCL Net Plugin](#) which should be used now instead. This feature was added in NCCL v2.12.

Now, when NCCL is initialized, it will look for a `libncc1-net.so` library and dynamically load it, then look for symbols inside the library. That's where proprietary hardware vendors should now put their custom APIs. This library, of course, should still be either in `LD_LIBRARY_PATH` or the `/etc/ld.so.conf` config.

For more information about dynamic library loading see [this section](#).

Node Proximity

If you get 2 random nodes from the cloud they may not reside on the same subnet and there will be an additional latency incurred for all transmissions.

You want to make sure that the nodes used for a single training all reside on the same subnet/spine so they are all one hop away from each other.

When you plan to eventually have a large cluster but starting small make sure that your provider can expand the cluster while keeping all the nodes close to each other.

Here are the cloud-specific ways of accomplishing node proximity:

- Azure: [availability set](#)
- GCP: [compact placement policies](#)

Depending on the type of package you have or what type of machines you rent - you may or may not be able to use those.

Shared internode network

If you use a shared HPC environment, or even if you have your own cluster but sharing it with your colleagues expect the network bandwidth to be unreliable and fluctuate at different times of the day.

This situation unfortunately makes it extremely difficult to finetune the performance of your training setup. Since every time you run a test the TFLOPs will vary, so how do you do the optimization? This is at least the situation with SLURM-based clusters. Apparently when Kubernetes is used, one can use cluster namespaces to segregate the network.

case study: we had this issue at JeanZay HPC when we were doing preliminary experiments before we started training BLOOM-176B. As that HPC has many users it was pretty much impossible to do speed optimizations, as even running the exact same setup again and again gave different throughput results. Luckily just before we launched BLOOM-176B training we were given an exclusive access to the new at that time A100 partition so we were the only users and we were able to greatly optimize the throughput.

Parallelism network collectives

See [Parallelism network collectives](#).

Communication Patterns

The intention of this chapter is not to show code examples and explain APIs for which there are many tutorials, but to have excellent visuals that explain how the various types of communication patterns work.

Point-to-point communications

Point-to-point communications are the simplest type of communication where there is always a single sender and a single receiver.

For example, [Pipeline Parallelism](#) performs a point-to-point communication where the activations from the current vertical stage is sent to the next stage. So the current gpu performs `send` and the gpu holding the next stage performs `recv`.

PyTorch has `send` and `recv` for blocking, `isend` and `irecv` for non-blocking p2p comms. [more](#).

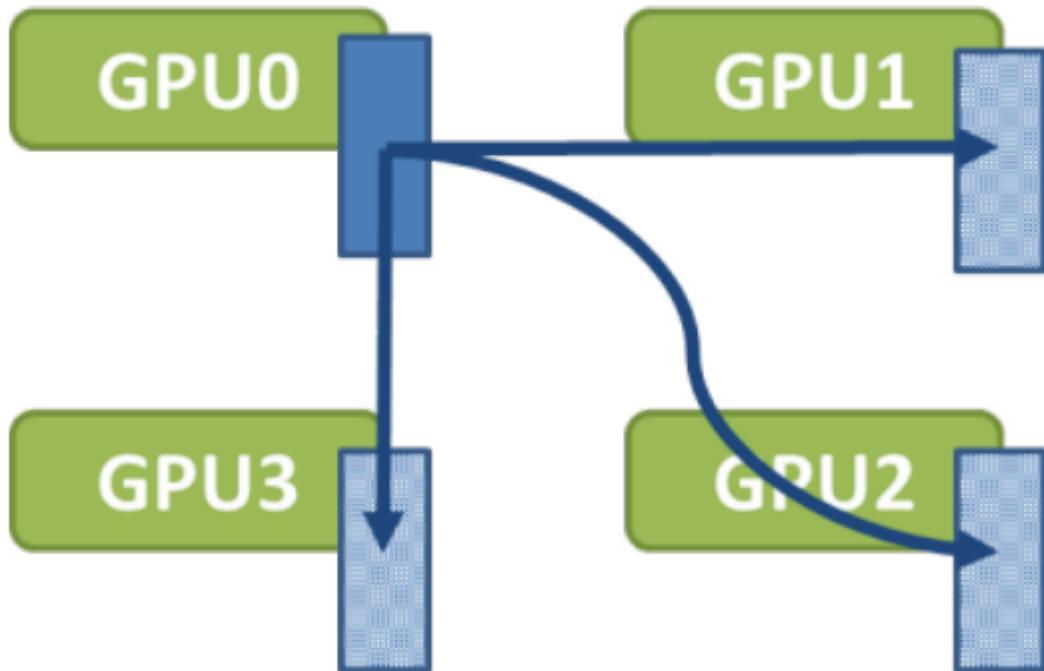
Collective communications

Collective communications include either multiple senders and a single receiver, a single sender and multiple receivers or multiple senders and multiple receivers.

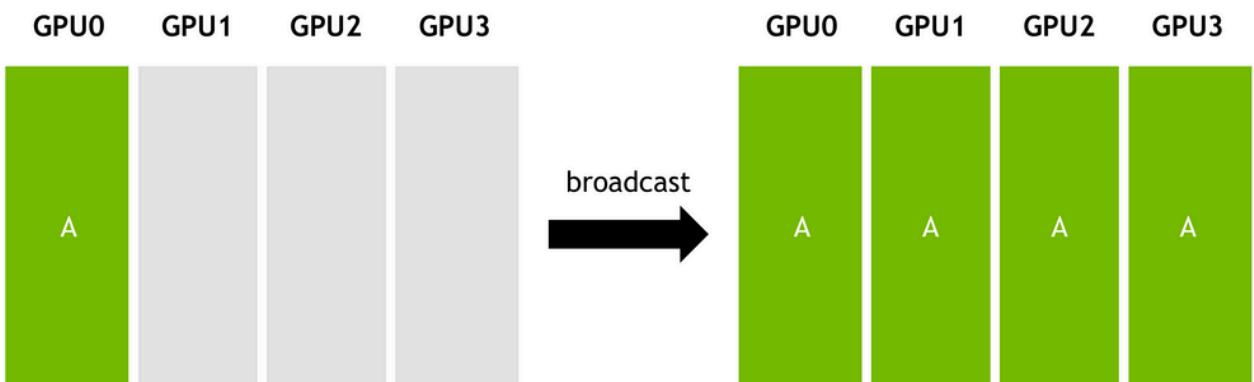
In the world of PyTorch typically each process is tied to a single accelerator, and thus accelerators perform collective communications via process groups. The same process may belong to multiple process groups.

Broadcast

Broadcast



[source](#)

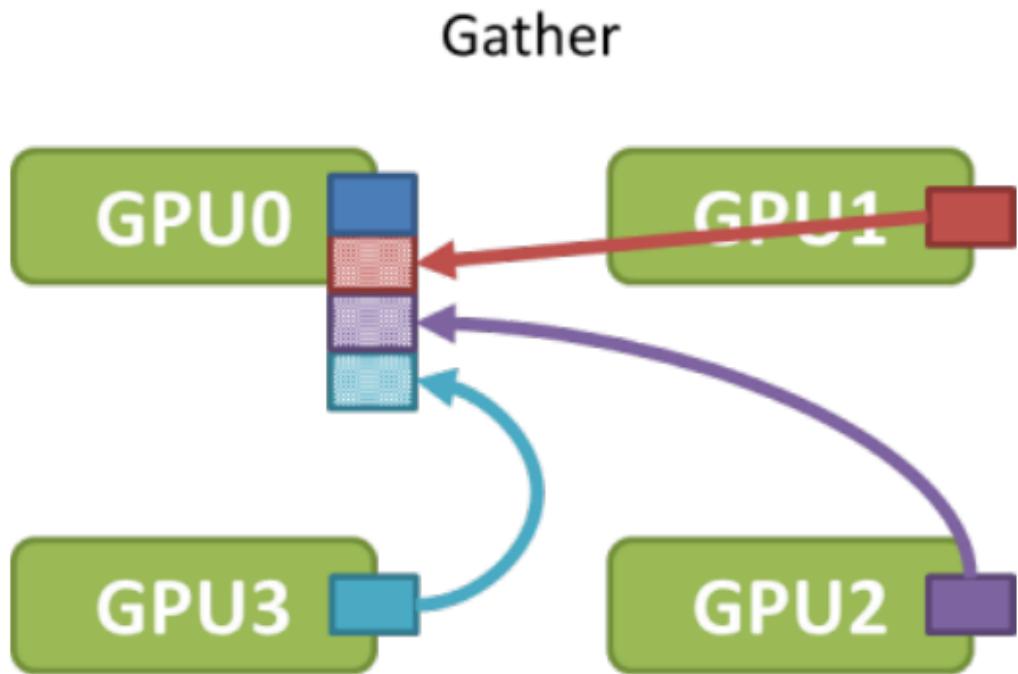


[source](#)

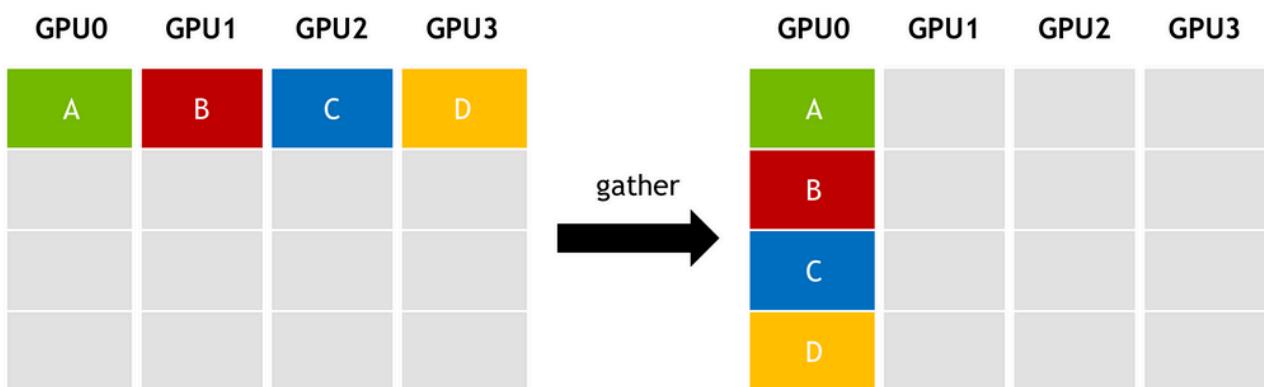
PyTorch API example:

`dist.broadcast(tensor, src, group):` Copies tensor from src to all other processes. [doc](#).

Gather



[source](#)



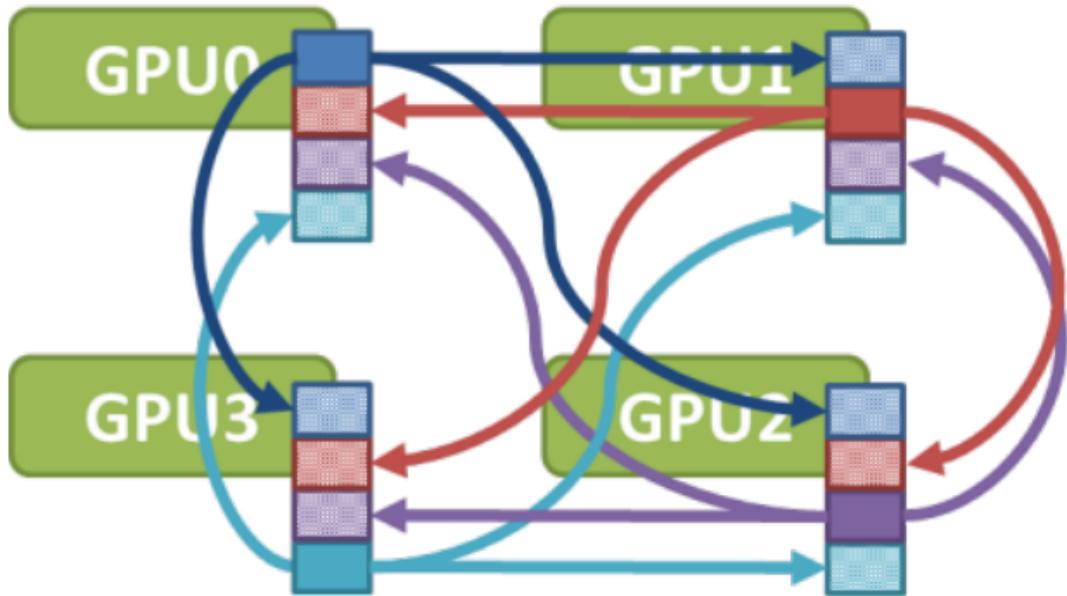
[source](#)

PyTorch API example:

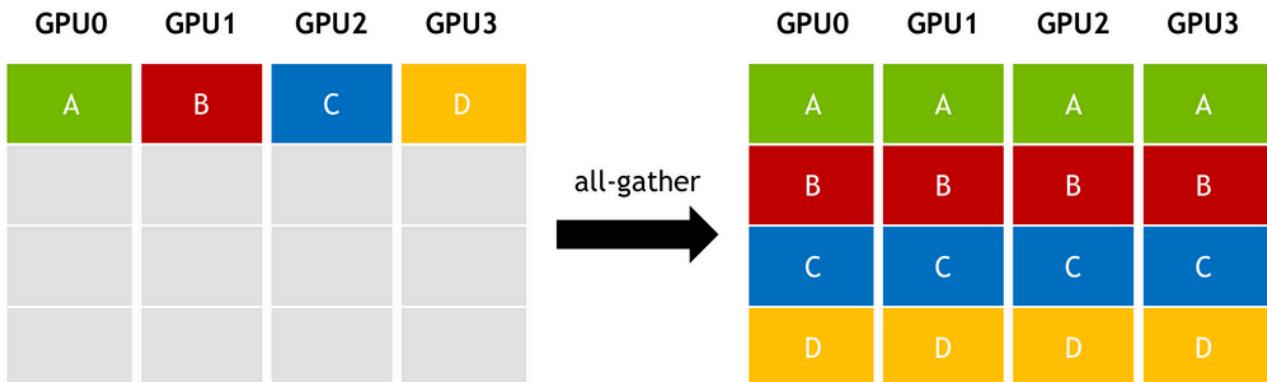
`dist.gather(tensor, gather_list, dst, group):` Copies tensor from all processes in dst. [doc](#)

All-gather

All-Gather



[source](#)



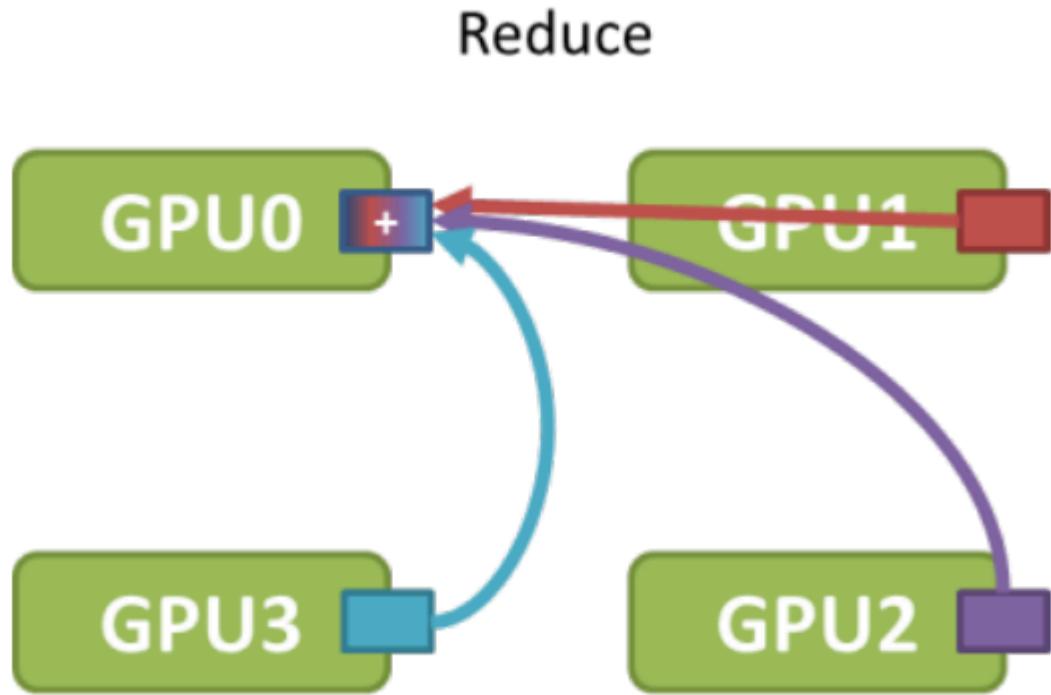
[source](#)

For example, this collective is used in [ZeRO](#) (Deepspeed and FSDP) to gather the sharded model weights before forward and backward calls.

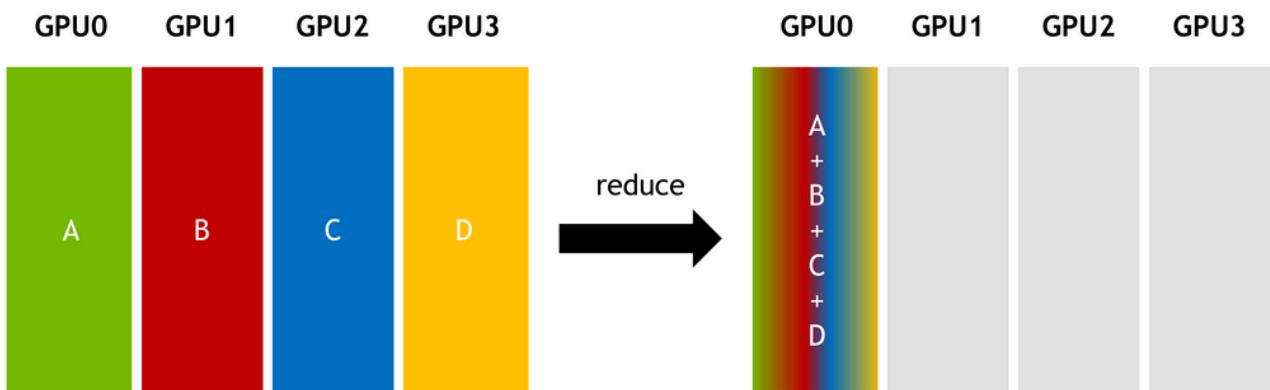
PyTorch API example:

`dist.all_gather(tensor_list, tensor, group):` Copies `tensor` from all processes to `tensor_list`, on all processes. [doc](#)

Reduce



[source](#)



[source](#)

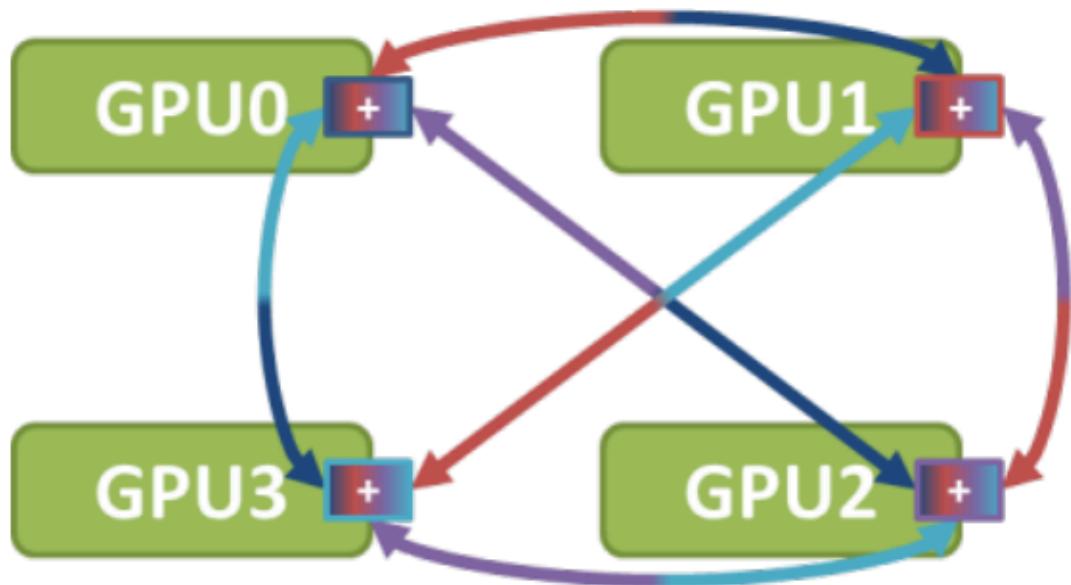
PyTorch API example:

`dist.reduce(tensor, dst, op, group):` Applies op to every tensor and stores the result in dst. [doc](#)

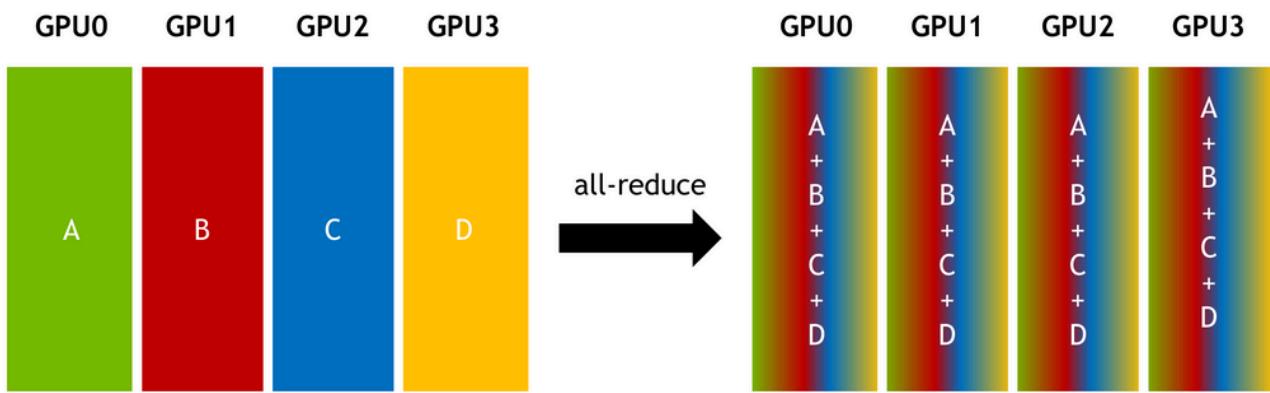
PyTorch supports multiple reduction operations like: `avg`, `sum`, `product`, `min`, `max`, `band`, `bor`, `bxor`, and others - [full list](#).

All-reduce

All-Reduce



[source](#)



[source](#)

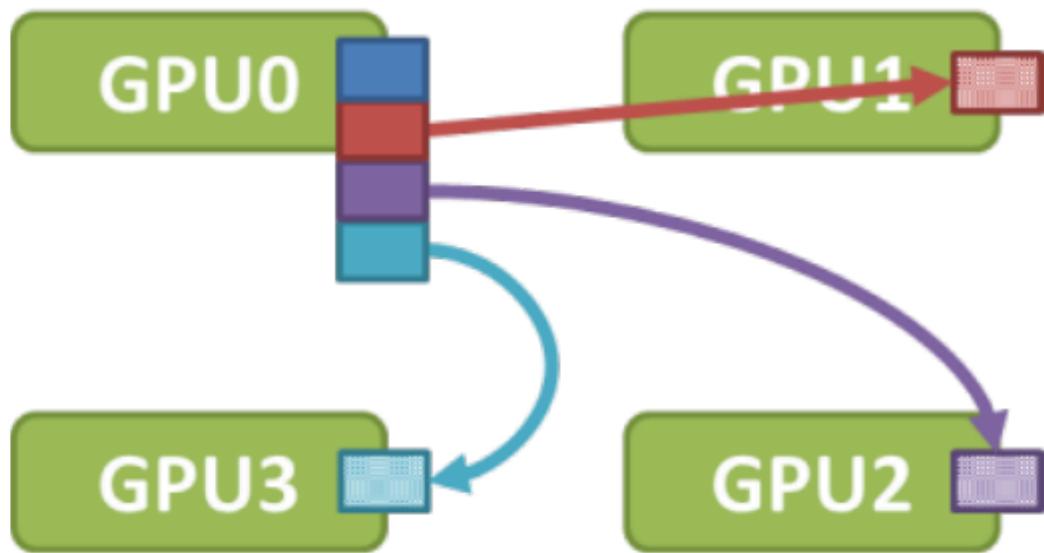
For example, this collective is used in [DDP](#) to reduce gradients between all participating ranks.

PyTorch API example:

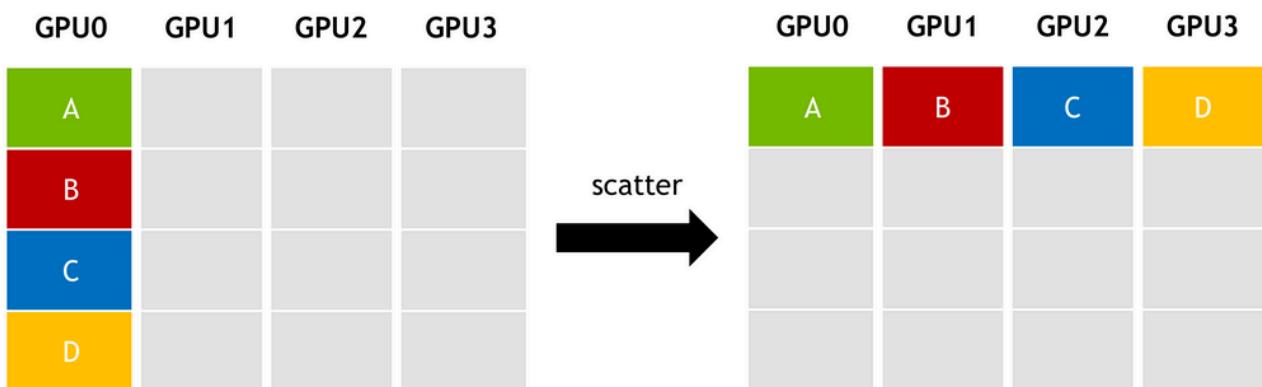
`dist.all_reduce(tensor, op, group):` Same as reduce, but the result is stored in all processes. [doc](#)

Scatter

Scatter



[source](#)

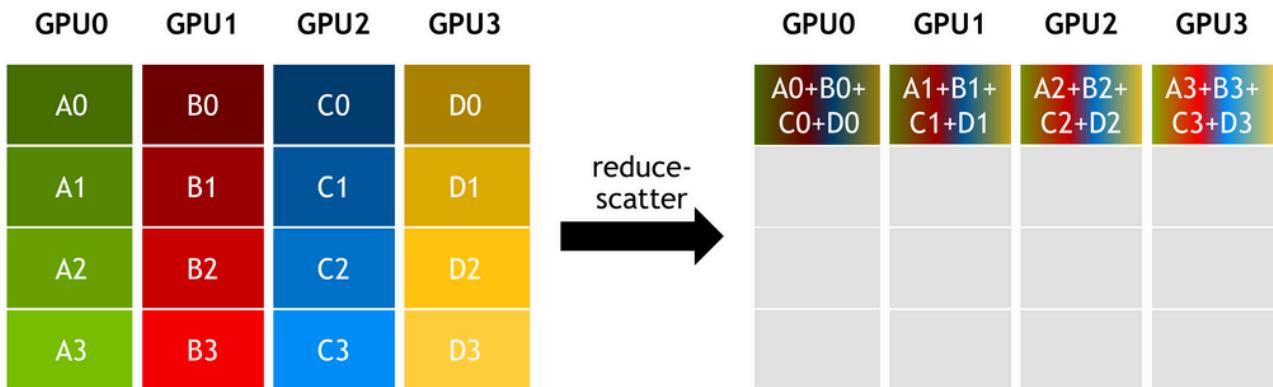


[source](#)

PyTorch API example:

`dist.scatter(tensor, scatter_list, src, group):` Copies the i -th tensor `scatter_list[i]` to the i -th process. [doc](#)

Reduce-Scatter



[source](#)

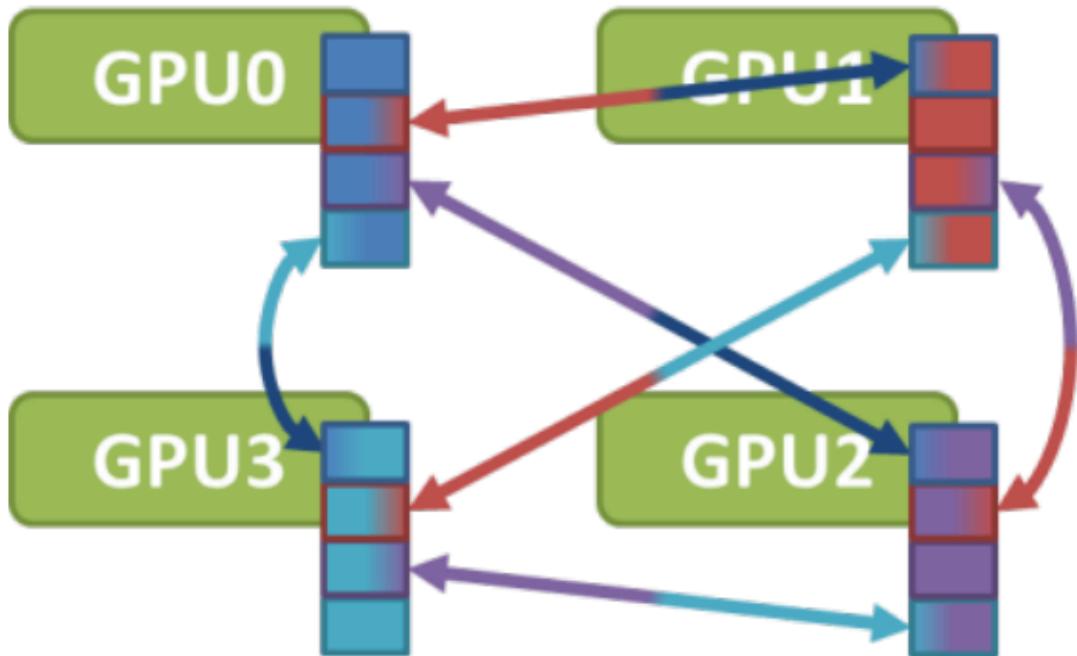
For example, this collective is used in [ZeRO](#) (Deepspeed and FSDP) to efficiently reduce gradients across all participating ranks. This is 2x more efficient than [all-reduce](#).

PyTorch API example:

`reduce_scatter(output, input_list, op, group, async_op)`: Reduces, then scatters a list of tensors to all processes in a group. [doc](#)

All-to-all

All-to-All



[source](#)

GPU0	GPU1	GPU2	GPU3
A0	B0	C0	D0
A1	B1	C1	D1
A2	B2	C2	D2
A3	B3	C3	D3

all-to-all

GPU0	GPU1	GPU2	GPU3
A0	A1	A2	A3
B0	B1	B2	B3
C0	C1	C2	C3
D0	D1	D2	D3

[source](#)

For example, this collective is used in [Deepspeed Sequence Parallelism](#) for attention computation.

PyTorch API example:

`dist.all_to_all(output_tensor_list, input_tensor_list, group):` Scatters list of input tensors to all processes in a group and return gathered list of tensors in output list. [doc](#)

Algorithms

The collective communications may have a variety of different implementations, and comm libraries like `ncc1` may switch between different algorithms depending on internal heuristics, unless overridden by users.

Ring

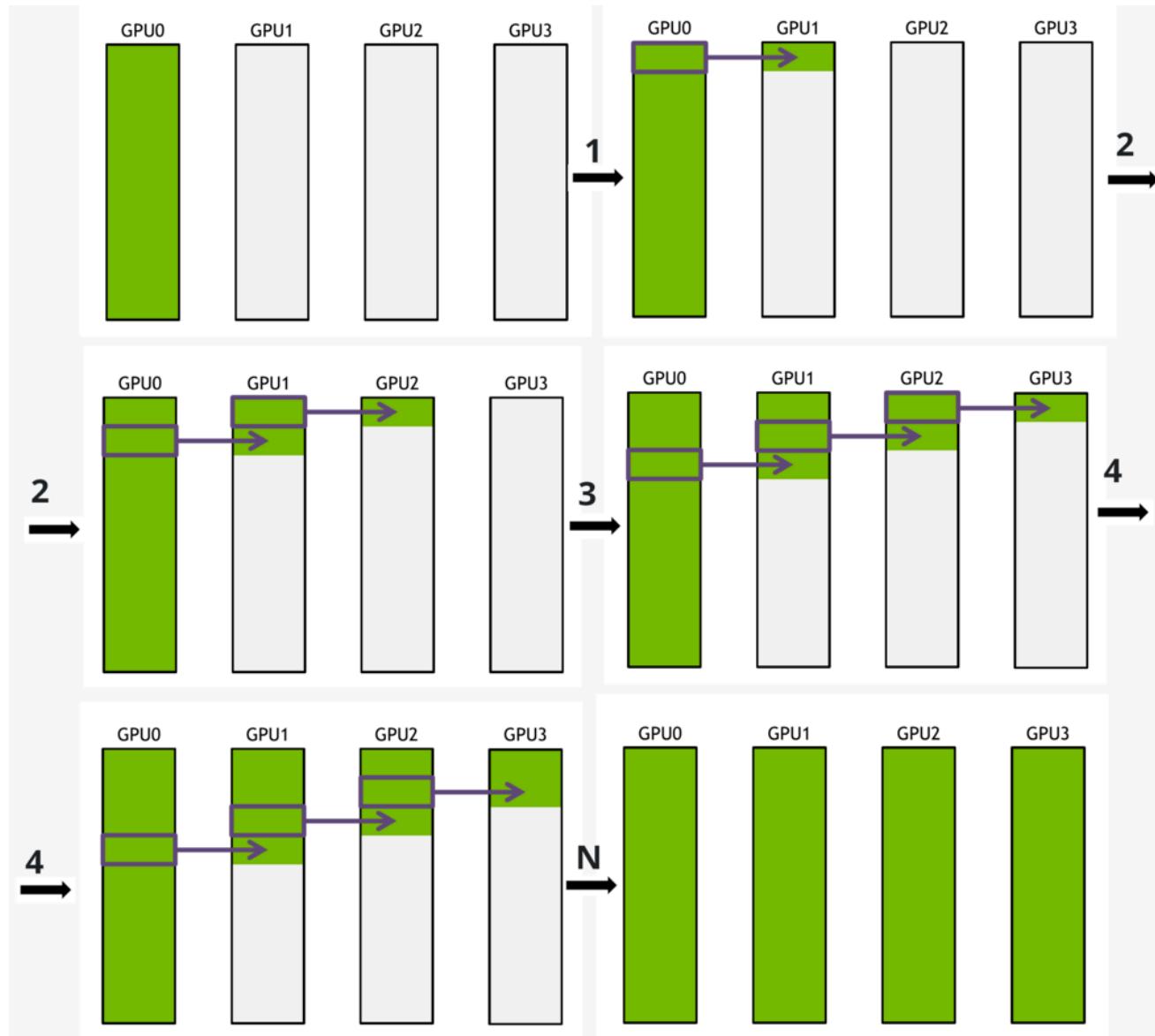
Broadcast with unidirectional ring

Given:

- N : bytes to broadcast
- B : bandwidth of each link
- k : number of GPUs

A naive broadcast will send N/B at each step. The total time to broadcast to k GPUs will take: $(k-1)*N/B$

Here is an example of how a ring-based broadcast is performed:



[source](#)

This algorithm splits N into S messages

At each step $N/(S*B)$ is sent, which is S times less than the naive algorithm sends per step.

The total time to broadcast N bytes to k GPUs will take:

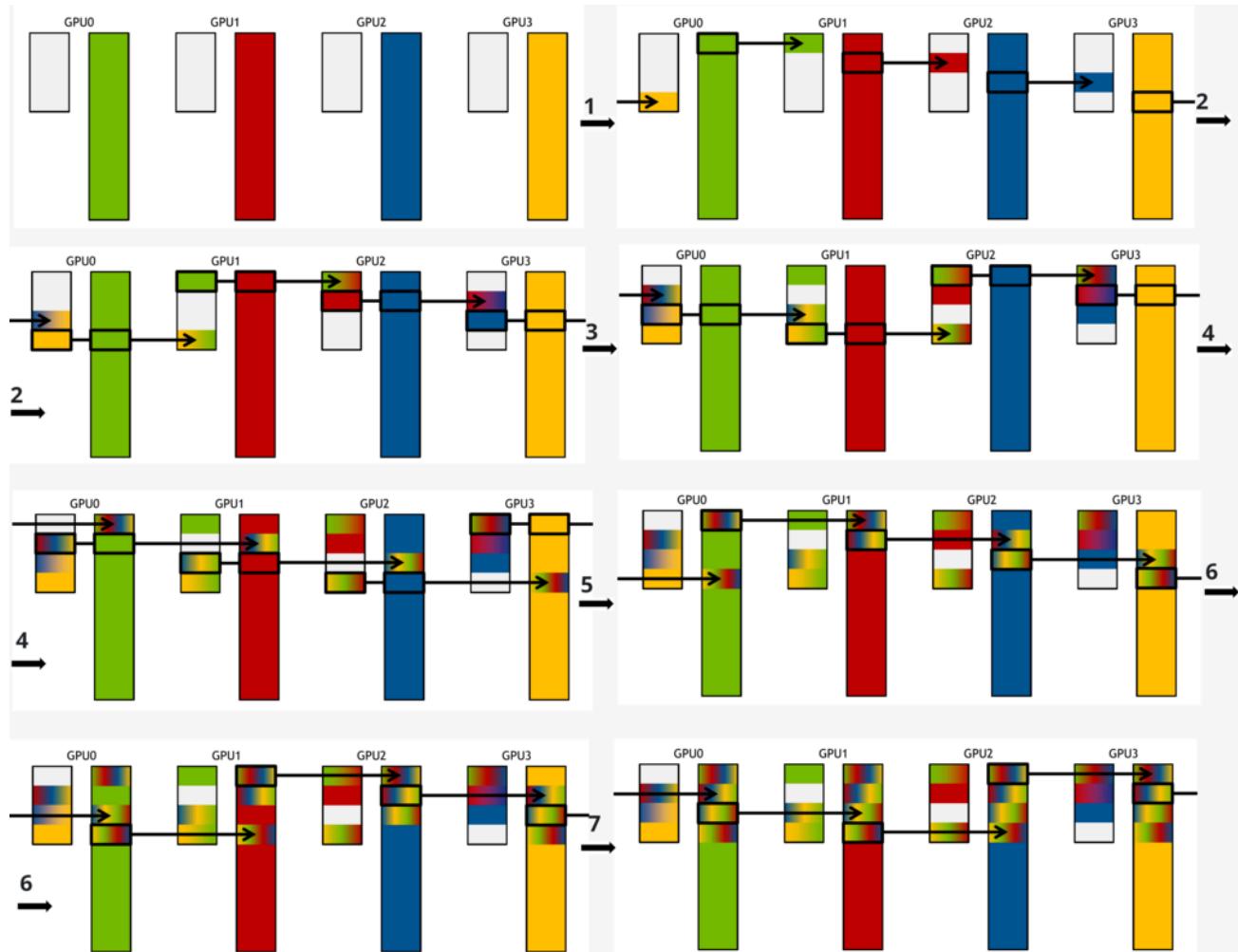
$$S*N/(S*B) + (k - 2)*N*(SB) = N*(S + k - 2)/(S*B)$$

and if split messages are very small so that $S \gg k$: $S + k - 2 \approx S$ and then the total time is about N/B .

All-reduce with unidirectional ring

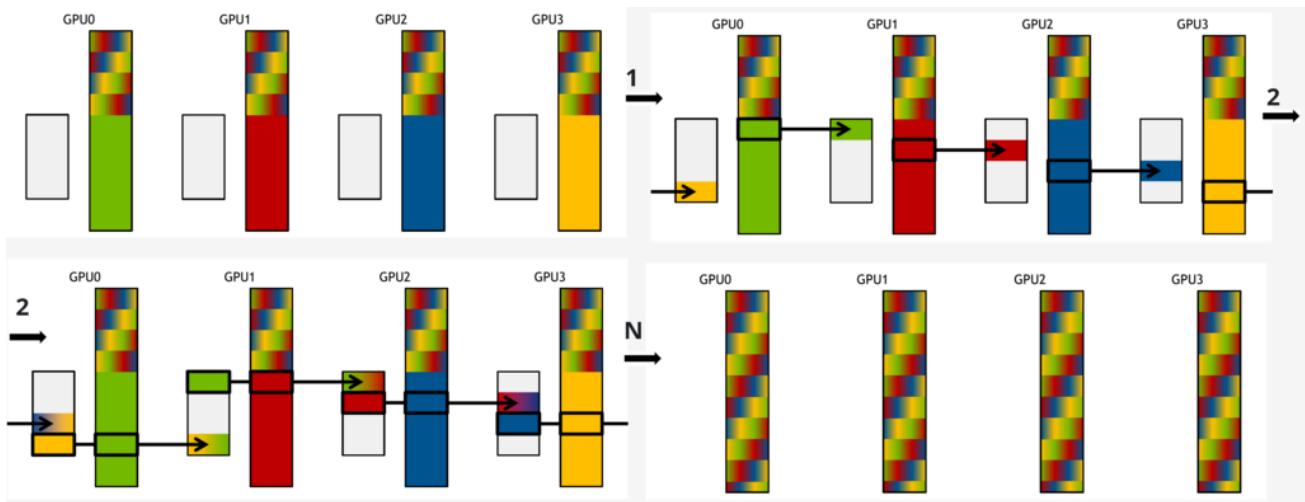
Ring-based `a11-reduce` is done similarly to [broadcast](#). The message is split into many small messages and each GPU sends a small message to the next GPU in parallel with other GPUs. `a11-reduce` has to perform 2x steps than `broadcast`, because it performs a reduction - so the size of the message needs to be sent twice over the wire.

Moreover, the whole message can be first split into chunks, to make the process even more efficient. Here is the reduction of the first chunk:



[source](#)

Then the next chunk is done, until all smaller messages are reduced:



[source](#)

More guides

Here are some additional guides with good visuals:

- [UvA Deep Learning Tutorials](#)

Network Debug

Often you don't need to be a network engineer to figure out networking issues. Some of the common issues can be resolved by reading the following notes.

Glossary

- OOB: Out-of-Band (typically a slower ethernet NIC)
- Bonding: using multiple NICs together for faster speed or as a back up
- IB: InfiniBand (Originally by Mellanox, acquired by NVIDIA)
- NIC: Network Interface Card

How to diagnose NCCL multi-gpu and multi-node connectivity issues

This section is definitely non-exhaustive and is meant to cover some of the most common setup issues that I have often encountered. For more complex problems please research the [NCCL repo Issues](#) or file a new Issue if you can't find one matching your situation. NCCL also includes a brief [troubleshooting section](#) but usually one learns a lot more from reading [Issues](#).

For the network diagnostics work, instead of using a full application which may take a long time to launch and have unrelated issue, I recommend using this specially developed design test script: [torch-distributed-gpu-test.py](#).

First, run the nccl-based program after setting:

```
export NCCL_DEBUG=INFO
```

which will print a lot of debug info about the NCCL setup and its network traffic.

For example if you're using the aforementioned debug script, for a single node with 8 GPUs, you might do:

```
NCCL_DEBUG=INFO python -m torch.distributed.run --nproc_per_node 8 --nnodes 1
torch-distributed-gpu-test.py
```

To launch it on multiple nodes, you'd have to either use some orchestration software like SLURM or Kubernetes, or manually launch it on each node (`pdsh` would be of a huge help) - see the instructions inside [torch-distributed-gpu-test.py](#) for details. But to understand how things work I recommend starting with just 1 node and then progressing to 2, and later to more nodes.

Now, inspect the output of the program and look for a line that starts with:

```
NCCL INFO NET/
```

and then inspect which protocol and which interfaces it is using.

For example, this output:

```
NCCL INFO NET/FastSocket : Using [0]ibs108:10.0.19.12<0> [1]ibs109:10.0.19.13<0> [2]ibs110:10.0.19.14<0>
[3]ibs111:10.0.19.15<0> [4]ibs112:10.0.19.16<0> [5]ibs113:10.0.19.17<0> [6]ibs114:10.0.19.18<0>
[7]ibs115:10.0.19.19<0>
```

tells us that [nccl-fastsocket](#) transport layer plugin is used and it discovered 8 `ibs*` network interfaces (NIC cards). If you're using Google Cloud this is correct, and your NCCL is likely setup correctly. But if you're using InfiniBand (IB) and you're getting the above output, you're likely to clock a very low internode speed, because this means that you activated the wrong plugin.

In the case of IB, what you want to see is `NET/IB` and its IB interfaces:

```
NCCL INFO NET/IB : Using [0]mlx5_0:1/IB [1]mlx5_1:1/IB [2]mlx5_2:1/IB [3]mlx5_3:1/IB [4]mlx5_4:1/IB
[5]mlx5_5:1/IB [6]mlx5_6:1/IB [7]mlx5_7:1/IB [RO]; OOB eno1:101.262.0.9<0>
```

Here, you can see that IB is used with 8 `mlx5_*` interfaces for collective comms, and one OOB, which stands for Out-Of-Band, and is used for doing bootstrapping the connections and is usually using a slower Ethernet NIC (at times [several NICs bonded into one](#) - in case you're wondering what does `bond` in the interface name stand for).

To know which TCP/IP interfaces your node has you run `ifconfig` on one of the nodes (typically all similar nodes will have the same interface names, but not always).

If your collective comms network is IB, instead of `ifconfig` you'd run `ibstat`. The last example of `NCCL INFO NET` would correspond to the following output:

```
$ ibstat | grep mlx5
CA 'mlx5_0'
CA 'mlx5_1'
CA 'mlx5_2'
CA 'mlx5_3'
CA 'mlx5_4'
CA 'mlx5_5'
CA 'mlx5_6'
CA 'mlx5_7'
```

Since besides the fast inter-node connectivity NICs, you're also likely to have a slow management Ethernet NIC (or even several of those), that is there to be able to configure the node, use a shared file system, access the Internet, it's almost certain that `ifconfig` will also include additional NICs. Also you are likely to have a docker network interface, `lo` loopback and some others. For example on my desktop I may get the following output:

```
$ ifconfig
docker0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 172.99.0.1  netmask 255.255.0.0  broadcast 172.99.255.255
        inet6 f330::42:fe33:f335:7c94  prefixlen 64  scopeid 0x20<link>
          ether 02:42:fe:15:1c:94  txqueuelen 0  (Ethernet)
            RX packets 219909  bytes 650966314 (650.9 MB)
            RX errors 0  dropped 0  overruns 0  frame 0
            TX packets 262998  bytes 20750134 (20.7 MB)
            TX errors 0  dropped 0  overruns 0  carrier 0  collisions 0
```

```

lo: flags=73<UP,LOOPBACK,RUNNING> mtu 65536
    inet 127.0.0.1 netmask 255.0.0.0
    inet6 ::1 prefixlen 128 scopeid 0x10<host>
        loop txqueuelen 1000 (Local Loopback)
        RX packets 1147283113 bytes 138463231270 (138.4 GB)
        RX errors 0 dropped 0 overruns 0 frame 0
        TX packets 1147283113 bytes 138463231270 (138.4 GB)
        TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500
    inet 10.0.0.23 netmask 255.255.255.0 broadcast 10.0.0.255
    inet6 2601:3108:1c71:600:4224:7e4b:13e4:7b54 prefixlen 64 scopeid 0x0<global>
        ether 04:41:1a:16:17:bd txqueuelen 1000 (Ethernet)
        RX packets 304675330 bytes 388788486256 (388.7 GB)
        RX errors 0 dropped 0 overruns 0 frame 0
        TX packets 74956770 bytes 28501279127 (28.5 GB)
        TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0
    device memory 0xa3b00000-a3bfffff

```

The reason I mention all these is that the critical part is to ensure that NCCL reports only the correct interfaces in its `Using` debug line. If any interfaces like `docker0` or `lo` or `eth0` end up being reported, e.g.:

```
NCCL INFO NET/Socket : Using [0]eth0:10.0.0.23<0>
```

it's most likely not what you want if you have faster network interfaces available. But, of course, in some situations the Ethernet NIC is all you have, in which case the above is just fine - it'll be just very slow.

Sometimes, if the wrong interface ends up being used, the application might just hang.

If you have all the correct interfaces, plus some incorrect interfaces NCCL might work but at the slower speed.

If it's a cloud environment, typically your cloud provider should give you instructions on how to set things up correctly. If they didn't then you need to at least ask them which network interfaces you need to use to setup NCCL.

While NCCL tries hard to auto-discover which interfaces it should use, if it fails to do so correctly you can then help it by telling it which interfaces to use or not to use:

- `NCCL_SOCKET_IFNAME` can be used to specify which `ifconfig` interfaces to include or exclude when not using Infiniband. Here are some examples:

```

export NCCL_SOCKET_IFNAME=eth:           Use all interfaces starting with eth, e.g. eth0, eth1, ...
export NCCL_SOCKET_IFNAME==eth0:         Use only interface eth0
export NCCL_SOCKET_IFNAME==eth0,eth1:   Use only interfaces eth0 and eth1
export NCCL_SOCKET_IFNAME=^docker:     Do not use any interface starting with docker
export NCCL_SOCKET_IFNAME=^=docker0:   Do not use interface docker0.

```

The full doc is [here](#).

- When using IB RDMA (IB Verbs interfaces), instead of `NCCL_SOCKET_IFNAME` use `NCCL_IB_HCA` env var which selects the interfaces for the collective communications. Examples:

```
export NCCL_IB_HCA=mlx5 : Use all ports of all cards starting with mlx5
export NCCL_IB_HCA==mlx5_0:1,mlx5_1:1 : Use ports 1 of cards mlx5_0 and mlx5_1.
export NCCL_IB_HCA=^=mlx5_1,mlx5_4 : Do not use cards mlx5_1 and mlx5_4.
```

The full doc is [here](#).

For example, often with IB, there will be additional interfaces like `mlx5_bond_0` which you don't want to be included in the NCCL comms. For example, this report would indicate that the wrong `[8]mlx5_bond_0:1/RoCE` interface was included and this would almost certainly lead to a low bandwidth:

```
NCCL INFO NET/IB : Using [0]mlx5_0:1/IB [1]mlx5_1:1/IB [2]mlx5_2:1/IB [3]mlx5_3:1/IB [4]mlx5_4:1/IB
[5]mlx5_5:1/IB [6]mlx5_6:1/IB [7]mlx5_7:1/I [8]mlx5_bond_0:1/RoCE [RO]; OOB ibp25s0:10.0.12.82<0>
```

There you'd exclude it with:

```
export NCCL_IB_HCA=^mlx5_bond_0:1
```

or alternatively you could list explicitly the interfaces you want, e.g.:

```
export NCCL_IB_HCA==mlx5_0,mlx5_1,mlx5_2,mlx5_3,mlx5_4,mlx5_5,mlx5_6,mlx5_7
```

As mentioned earlier using `ibstat` on one of the nodes interconnected with IB will show you the available IB interfaces.

Since NCCL tries to automatically choose the best network interfaces, you only need to do the above if NCCL doesn't work or it's slow. In normal circumstances NCCL should work out of the box, without the user needing to do anything special.

Additionally, depending on which cloud is used, it's very likely that the provider may give you a slew of environment variables to set. If you set some of them incorrectly, NCCL might work slowly or not work at all.

Another typical problem users run into is when they try to reuse their NCCL setup they had working on cloud A on a cloud B. Often things don't translate and one has to carefully remove any previously set environment variables and set them correctly anew for the new cloud. This issue is likely to occur even if you're using the same cloud, but different types of instances, as some network setups are very specific for a given instance and won't work elsewhere.

Once you think you have set up the NCCL correctly, the next thing is to benchmark your connectivity and ensure that it matches the advertised speed (well, ~80% of it). Proceed to the [benchmark chapter](#).

NCCL with docker containers

- Give enough resources by adding to the docker `run` these additional args: `-shm-size=1g -ulimit memlock=-1` ([more details](#))
- Privileged access: sometimes you need to add `--privileged` to the docker `run` args.
- Having the docker image include the right packages, e.g. if using IB you'd want at least to install `libibverbs1` `librdmacm1`

How to check if P2P is supported

Sometimes you need to know if the GPUs on your compute node support P2P access (Peer2Peer). Disabling P2P will

typically lead to a slow intra-node connectivity.

You can see that on this particular 8x NVIDIA H100 node the P2P is supported:

```
$ nvidia-smi topo -p2p r
      GPU0    GPU1    GPU2    GPU3    GPU4    GPU5    GPU6    GPU7
GPU0  X      OK      OK      OK      OK      OK      OK      OK
GPU1  OK      X      OK      OK      OK      OK      OK      OK
GPU2  OK      OK      X      OK      OK      OK      OK      OK
GPU3  OK      OK      OK      X      OK      OK      OK      OK
GPU4  OK      OK      OK      OK      X      OK      OK      OK
GPU5  OK      OK      OK      OK      OK      X      OK      OK
GPU6  OK      OK      OK      OK      OK      OK      X      OK
GPU7  OK      OK      OK      OK      OK      OK      OK      X
```

Legend:

```
X      = Self
OK     = Status Ok
CNS    = Chipset not supported
GNS    = GPU not supported
TNS    = Topology not supported
NS     = Not supported
U     = Unknown
```

On the other hand with this particular 2x NVIDIA L4 the P2P is not supported:

```
$ nvidia-smi topo -p2p r
      GPU0    GPU1
GPU0  X      CNS
GPU1  CNS    X
```

As you can see from the Legend, CNS signifies that "Chipset is not supported".

If you're using a high-end datacenter GPUs this is very unlikely to happen. Though some low-end datacenter GPUs may not support P2P like the example of L4 above.

For consumer-level GPUs there could be a variety of reasons for your GPU not being supported, often it's the IOMMU and/or ACS features being enabled. At other times it's just the driver version. And if you spend some time searching you might find someone hacking drivers to enable P2P in GPUs that shouldn't support P2P, like this [4090 P2P support repo](#).

To check if PCI Access Control Services (ACS) are enabled and to disable those follow [this guide](#).

IOMMU can be disabled in the BIOS.

You can also check P2P support between specific GPUs using torch - here are we checking for GPUs 0 and 1:

```
python -c "import torch; print(torch.cuda.can_device_access_peer(torch.device('cuda:0'),
torch.device('cuda:1')))"
```

If there is no P2P support, the above would print `False`.

How to count NCCL calls

Enable NCCL debug logging for subsystems - collectives:

```
export NCCL_DEBUG=INFO
export NCCL_DEBUG_SUBSYS=COLL
```

if you're working in a slurm environment with many nodes you probably want to perform this only on rank 0, like so:

```
if [[ $SLURM_PROCID == "0" ]]; then
  export NCCL_DEBUG=INFO
  export NCCL_DEBUG_SUBSYS=COLL
fi
```

Assuming your logs were all sent to `main_log.txt`, you can then count how many of each collective call were performed with:

```
grep -a "NCCL INFO Broadcast" main_log.txt | wc -l
2590
grep -a "NCCL INFO AllReduce" main_log.txt | wc -l
5207
grep -a "NCCL INFO AllGather" main_log.txt | wc -l
1849749
grep -a "NCCL INFO ReduceScatter" main_log.txt | wc -l
82850
```

It might be a good idea to first isolate a specific stage of the training, as loading and saving will have a very different pattern from training iterations.

So I typically first slice out one iteration. e.g. if each iteration log starts with: `iteration: ...` then I'd first do:

```
csplit main_log.txt '/iteration: /' "{}"
```

and then analyse one of the resulting files that correspond to the iterations. By default it will be named something like `xx02`.

Useful NCCL Debug Environment Variables

The following env vars are most useful during debugging NCCL-related issues such as hanging and crashing. The full list of those can be found [here](#).

NCCL_DEBUG

This is the most commonly used env var to debug networking issues.

Values:

- **VERSION** - Prints the NCCL version at the start of the program.
- **WARN** - Prints an explicit error message whenever any NCCL call errors out.
- **INFO** - Prints debug information
- **TRACE** - Prints replayable trace information on every call.

For example:

```
NCCL_DEBUG=INFO python -m torch.distributed.run --nproc_per_node 2 --nnodes 1
torch-distributed-gpu-test.py
```

This will dump a lot of NCCL-related debug information, which you can then search online if you find that some problems are reported.

And **NCCL_DEBUG_FILE** should be very useful when using **NCCL_DEBUG** as the information is copious especially if using many nodes.

NCCL_DEBUG_FILE

When using **NCCL_DEBUG** env var, redirect all NCCL debug logging output to a file.

The default is **stdout**. When using many GPUs it can be very useful to save each process' debug info into its own log file, which can be done like so:

```
NCCL_DEBUG_FILE=/path/to/nccl-log.%h.%p.txt
```

- **%h** is replaced with the hostname
- **%p** is replaced with the process PID.

If you then need to analyse hundreds of these at once, here are some useful shortcuts:

- grep for a specific match and also print the file and line number where it was found:

```
grep -n "Init COMPLETE" nccl-log*
```

- show **tail -1** of all nccl log files followed by the name of each file

```
find . -name "nccl*" -exec sh -c 'echo "$(tail -1 "$1") ($1)"' _ {} \;
```

NCCL_DEBUG_SUBSYS

NCCL_DEBUG_SUBSYS used in combination with **NCCL_DEBUG** tells the latter which subsystems to show. Normally you don't have to specify this variable, but sometimes the developers helping you may ask to limit the output to only some sub-systems, for example:

```
NCCL_DEBUG_SUBSYS=INIT,GRAPH,ENV,TUNING
```

NCCL_P2P_DISABLE

Disables P2P comms - e.g. NVLink won't be used if there is one and the performance will be much slower as a result of that. Normally you don't want this but in a pinch sometimes this can be useful during debug.

NCCL_SOCKET_IFNAME

This one is very useful if you have multiple network interfaces and you want to choose a specific one to be used.

By default NCCL will try to use the fastest type of an interface, which is typically `ib` (InfiniBand).

But say you want to use an Ethernet interface instead then you can override with:

```
NCCL_SOCKET_IFNAME=eth
```

This env var can be used at times to debug connectivity issues, if say one of the interfaces is firewalled, and perhaps the others aren't and can be tried instead. Or if you are not sure whether some problem is related to the network interface or something else, so it helps to test other interfaces to invalidate that the issue comes from network.

Networking Benchmarks

Tools

all_reduce benchmark

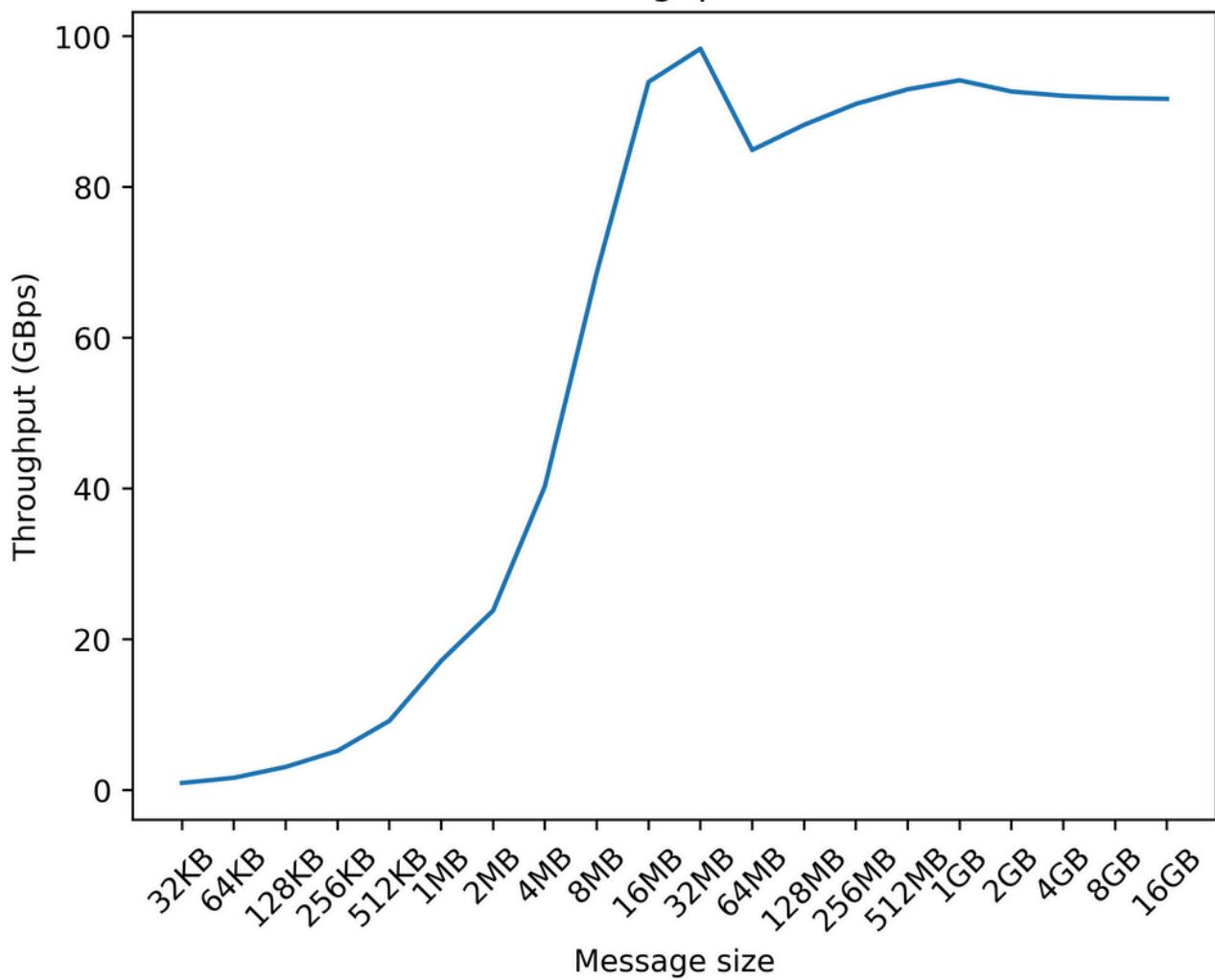
[all_reduce_bench.py](#) - a tool to benchmark the real network bandwidth while performing `all_reduce` on a largish amount of data. This is useful for finding out what one gets in reality as compared to the advertised spec. Somewhat similar to `nccl-tests`, but requires just PyTorch to run.

It generates output like this:

payload	busbw	a1gbw
32KB	0.92GBps	0.48GBps
64KB	1.61GBps	0.83GBps
128KB	3.05GBps	1.58GBps
256KB	5.18GBps	2.67GBps
512KB	9.17GBps	4.73GBps
1MB	17.13GBps	8.84GBps
2MB	23.79GBps	12.28GBps
4MB	40.30GBps	20.80GBps
8MB	68.62GBps	35.42GBps
16MB	93.93GBps	48.48GBps
32MB	98.34GBps	50.76GBps
64MB	84.90GBps	43.82GBps
128MB	88.23GBps	45.54GBps
256MB	91.01GBps	46.97GBps
512MB	92.95GBps	47.98GBps
1GB	94.15GBps	48.59GBps
2GB	92.66GBps	47.83GBps
4GB	92.09GBps	47.53GBps
8GB	91.80GBps	47.38GBps
16GB	91.69GBps	47.32GBps

And it also creates a plot:

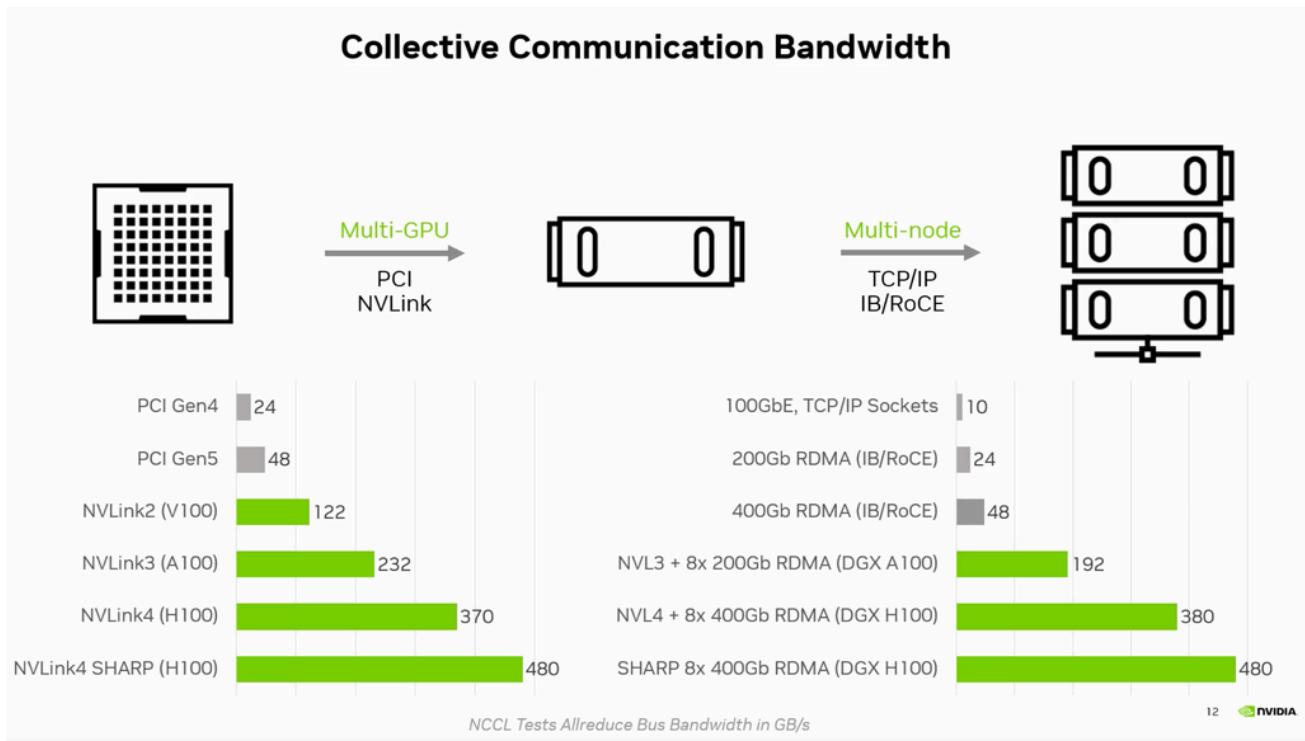
Bandwidth Throughput for ranks=32



```
_CudaDeviceProperties(name='NVIDIA H100 80GB HBM3', major=9,  
minor=0, total_memory=81109MB, multi_processor_count=132,  
uuid=ebb32a99-dba0-f05a-ee6f-f0ace21e552e,  
L2_cache_size=50MB)
```

For launching examples and notes please see the top of [all_reduce_bench.py](#).

This table should give a good sense for what scores you should expect for all-reduce collective on a well-tuned network (left is intra-node and right is inter-node):



[source](#)

If you're benchmarking a different collective the expected bandwidth can be very different from the above all-reduce results. [This presentation](#) also gives point-to-point communication bandwidth expectations.

all_gather_object vs all_reduce

[all_gather_object_vs_all_reduce.py](#) - a quick benchmark showing 23x speed up when moving from `all_gather_object` to `all_reduce` when collecting completion status from the process group. e.g. when implementing some sort of all-processes-are-done flag. This technique is usually used for synchronizing gpus when they may complete at different number of iterations - which one needs for inference over multiple DP channels, or when one wants to sync a `StopIteration` event in `DataLoader`. See also [all_gather_object_vs_all_gather.py](#).

all_reduce latency comparison

[all_reduce_latency_comp.py](#) - exemplifies how 1x 4GB reduction is much faster than 1000x 4MB reductions.

Crucial reproducibility requirements

The most important requirements for a series of successful experiments is to be able to reproduce the experiment environment again and again while changing only one or a few setup variables.

Therefore when you try to figure out whether some change will improve performance or make it worse, you must figure out how to keep things stable.

For example, you need to find a way to prevent the network usage from fluctuations. When we were doing performance optimizations for [108B pre-BLOOM experiments](#) it was close to impossible to perform, since we were on a shared internode network and the exact same setup would yield different throughput depending on how many other users used the network. It was not working. During BLOOM-176B we were given a dedicated SLURM partition with an isolated network where the only traffic was ours. Doing the performance optimization in such environment was just perfect.

Network throughput

It's critical to understand your particular model size and framework requirements with regard to network bandwidth, throughput and latency. If you underpay for network you will end up having idle gpus and thus you wasted money and time. If you overpay for very fast network, but your gpus are slow, then again you wasted money and time.

If your network is very slow, your training is likely to be network-bound and many improvements in the training setup will not help with the improving performance.

Note: The [EAI cookbook](#) contains a set of [communication benchmarks](#) for each collective that you can use to quickly measure the throughput of your internode or intranode network.

Here is a simple all-reduce benchmark that you can use to quickly measure the throughput of your internode network:

[all_reduce_bench.py](#)

Usually benchmarking at least 4 nodes is recommended, but, of course, if you already have access to all the nodes you will be using during the training, benchmark using all of the nodes.

To run it on 4 nodes:

```
GPUS_PER_NODE=8
NNODES=4
MASTER_ADDR=$(scontrol show hostnames $SLURM_JOB_NODELIST | head -n 1)
MASTER_PORT=6000
python -u -m torch.distributed.run \
    --nproc_per_node $GPUS_PER_NODE \
    --nnodes $NNODES \
    --rdzv_endpoint $MASTER_ADDR:$MASTER_PORT \
    --rdzv_backend c10d \
    --max_restarts 0 \
    --role `hostname -s`:
    --tee 3 \
    all_reduce_bench.py
```

Notes:

- adapt `MASTER_ADDR` to rank 0 hostname if it's not a SLURM environment where it's derived automatically.

Here is how to run launch it in a SLURM env with 4 nodes:

```
salloc --partition=mpartition --nodes=4 --ntasks-per-node=1 --cpus-per-task=48 --gres=gpu:8
--time=1:00:00 bash
srun --gres=gpu:8 --nodes=4 --tasks-per-node=1 python -u -m torch.distributed.run --nproc_per_node=8
--nnodes 4 --rdzv_endpoint $(scontrol show hostnames $SLURM_JOB_NODELIST | head -n 1):6000 --rdzv_backend
c10d all_reduce_bench.py
```

Notes:

- You are likely to need to adapt `--cpus-per-task` and `--partition` arguments there.
- You do `salloc` once and then can repeat `srun` multiple times on the same allocation.

You may get results anywhere between 5Gbps and 1600Gbps (as of this writing). The minimal speed to prevent being network bound will depend on your particular training framework, but typically you'd want at least 400Gbps or higher.

Though we trained BLOOM on 50Gbps.

Frameworks that shard weights and optim stages like [DeepSpeed](#) w/ ZeRO Stage-3 do a lot more traffic than frameworks like [Megatron-DeepSpeed](#) which do tensor and pipeline parallelism in addition to data parallelism. The latter ones only send activations across and thus don't need as much bandwidth. But they are much more complicated to set up and run.

Of course, an efficient framework will overlap communications and compute, so that while one stage is fetching data, the other stage in parallel runs computations. So as long as the communication overhead is smaller than compute the network requirements are satisfied and don't have to be super fantastic.

To get reasonable GPU throughput when training at scale (64+GPUs) with DeepSpeed ZeRO Stage 3 with V100s

1. 100Gbps is not enough
2. 200-400 Gbps is ok
3. 800-1000 Gbps is ideal

[full details](#)

Of course, the requirements are higher for A100 gpu nodes and even higher for H100s (but no such benchmark information has been shared yet).

Extrapolating benchmark results from several nodes to many

As it's often not easy to benchmark hundreds of nodes, often we try to benchmark interconnect performance using, say, 4 nodes. I wasn't sure whether this would give the correct indication for when 40 or 400 nodes will be used so I asked about it [here](#) and the answer was:

Extrapolating at scale is not that hard for ring and tree (we have a function in `tuning.cc` predicting it, based on the ring linear latency and the tree log latency with reduced BW). Now as you scale, there are many factors which may cause your real performance to be very far off the prediction, like routing. Also note on an IB network you'll be able to use SHARP; that way your latency stays mostly constant as you scale, your bandwidth doesn't degrade much either, and you're always better than both ring and tree.

Disable Access Control Services

PCI Access Control Services (ACS) used for IO virtualization (also known as VT-d or IOMMU) force P2P PCIe transactions to go up through the PCIe Root Complex, which does not enable GDS to bypass the CPU on paths between a network adapter or NVMe and the GPU in systems that include a PCIe switch.

For the optimal GDS performance, disable ACS by following these instructions [here](#). Here are some [additional notes](#)

Please note that if you're using Virtual machines you can't disable ACS as it's a required feature. To run with maximum performance inside virtual machines, Address Translation Service (ATS) needs to be enabled in network adapters.

Performance-Oriented NCCL Environment Variables

While NCCL is excellent at automatically figuring out the best performance for any given network, sometimes it needs some help, in which case the following NCCL env vars are used to tune up performance. Let's look at a few common ones you might want to be aware of, and the full list of those can be found [here](#).

Note that some `NCCL_IB_*` env vars apply to RoCEv2 networks as well.

`NCCL_ALGO`

This one defines which algorithms NCCL will use. Typically it's one of:

1. Tree
2. Ring
3. CollnetDirect and CollnetChain (IB SHARP)
4. NVLS (NVLink SHARP)

I was asking questions about how a user can do the optimization and was told at [this NCCL Issue](#) that basically the user shouldn't try to optimize anything as NCCL has a ton of smart algorithms inside that will try to automatically switch from one algorithm to another depending on a concrete situation.

Sylvain Jeaugey shared:

There used to be a static threshold, but it's been replaced by a more complex tuning system. The new system builds a model of the latency and bandwidth of each algorithm/protocol combination (that's many, many combinations) and decides which one should perform best depending on the size. So there is no longer an env var and a static value, which is good because the performance of each algorithm depends on the number of nodes and number of GPUs per node and therefore we need to navigate a 2D space of algo/protocols which isn't easy. You can always force one algorithm with `NCCL_ALGO=TREE` and `NCCL_ALGO=RING` and see what performance you get and whether NCCL switches at the right point. I know it's hard to understand, but it's also the best solution we found to have the best performance across all platforms and users without users having to manually tune the switch points. Downside is, if you want to manually tune things, you can't.

If you use `NCCL_ALGO` you need to list the algorithms to consider, but otherwise you have no control over it. So, really, this is only useful if you want to make sure that one of the algorithms isn't used.

When asking about which algorithm is better, I received:

Roughly speaking, ring is superior in terms of peak bandwidth (except on 2 nodes), tree is superior in terms of base latency (especially as we scale). `Bandwidth = Size / Time`, so whether you look at the time or the bandwidth for a given size, it will be a combination of both the peak bandwidth and the base latency. For a fixed size, as you scale, the base latency of ring will become prevalent and tree will be better.

There is also a new algo, named `NVLS`, which if NVLink SHARP is available will run faster than NVLink itself, e.g. with NVLink 4.0 (450GBps) one can clock 480GBps doing all-reduce benchmarks. They are working on the inter-node version of that which [requires IB or RoCE](#) - this new algo is not documented anywhere as of this writing.

And finally, if you would like to know which algo is being used - you can't - see [this answer](#). So if you want to know which algo gives which throughput you will have to try them all explicitly by setting `NCCL_ALGO` env var and then you'd know which one was chosen. Or you can edit and recompile NCCL as suggested in that same answer, but you won't want this in production.

NCCL_CROSS_NIC

The `NCCL_CROSS_NIC` variable controls whether NCCL should allow rings/trees to use different NICs, causing inter-node communication to use different NICs on different nodes.

To maximize inter-node communication performance when using multiple NICs, NCCL tries to communicate between same NICs between nodes, to allow for network design where each NIC from each node connects to a different network switch (network rail), and avoid any risk of traffic flow interference. The `NCCL_CROSS_NIC` setting is therefore dependent on the network topology, and in particular depending on whether the network fabric is rail-optimized or not.

This has no effect on systems with only one NIC.

Values accepted:

- 0: Always use the same NIC for the same ring/tree, to avoid crossing network rails. Suited for networks with per NIC switches (rails), with a slow inter-rail connection. Note there are corner cases for which NCCL may still cause cross-rail communication, so rails still need to be connected at the top.
- 1: Do not attempt to use the same NIC for the same ring/tree. This is suited for networks where all NICs from a node are connected to the same switch, hence trying to communicate across the same NICs does not help avoiding flow collisions.
- 2: (Default) Try to use the same NIC for the same ring/tree, but still allow for it if it would result in better performance.

NCCL_IB_QPS_PER_CONNECTION

This is relevant if you're on a multi-layer Infiniband or RoCEv2 network.

NCCL_IB_QPS_PER_CONNECTION defines the number of IB queue pairs to use for each connection between two ranks. This can be useful on multi-level fabrics which need multiple queue pairs to have good routing entropy. In other words, when your jobs are crossing spine or super-spine switches.

By default it is set to 1, but having a higher number might benefit throughput.

Depends on the size of the network. you could start with something like 4 for any cluster over 64 GPUs (i.e. any cluster that's bigger than the radix (number of ports) of its IB switch (e.g. the IB NDR switch radix is 64.)

Ideally you'd ask your cloud provider if they have already researched the best value, but if they didn't you can do it yourself, albeit it might be use-case specific.

The other gotcha is that when the value is higher than 1 an additional GPU memory will be consumed.

NCCL_MIN_CTAS and NCCL_MAX_CTAS

Cooperative Thread Array (CTA) implements CUDA thread blocks - You can read about it [here](#).

The CTA value is derived algorithmically by NCCL, but the default behavior can be overridden by setting the lower and upper limits via the env vars: `NCCL_MIN_CTAS` and `NCCL_MAX_CTAS`. And then NCCL's tuner will be limited to choose the best value in the user-imposed range. The same can be accomplished from the program using `pg_options` in `torch.distributed.init_process_group` via `ncclConfig_t`'s `minCTAs` and `maxCTAs` (other process group creation functions have `pg_options` as well). The latter approach allows you to set different CTA settings to different process groups, whereas the env vars will apply globally to all process groups.

Here is an example that directly sets both values to 32 per process group:

```
import torch
nccl_options = torch.distributed.ProcessGroupNCCL.Options()
nccl_options.config.min_ctas = 32
nccl_options.config.max_ctas = 32
torch.distributed.init_process_group(..., pg_options=nccl_options)
```

In order to find the best performance to experiment with different values against a specific benchmark of choice, that emulates the intended workload, you could set both config options to the same value and then bisect on a range of 1 to 64 or similar.

Infiniband

Infiniband adaptive routing

Make sure your cloud provider enables IB adaptive routing which could greatly improve the performance.

For nuances see this paper: [Adaptive Routing in InfiniBand Hardware](#).

Network Benchmarks Results

- [Disabling NVLink](#)

Disabling NVLink Benchmark

Let's compare the training of a gpt2 language model training over a small sample of wikitext.

The results are:

NVlink	Time
Y	101s
N	131s

You can see that NVLink completes the training ~23% faster. In the second benchmark we use `NCCL_P2P_DISABLE=1` to tell the GPUs not to use NVLink, which will use PCIe instead.

We will use [HF Transformers examples](#).

Here is the full benchmark code and outputs:

```
# DDP w/ NVLink

rm -r /tmp/test-clm; CUDA_VISIBLE_DEVICES=0,1 python -m torch.distributed.launch \
--nproc_per_node 2 examples/pytorch/language-modeling/run_clm.py --model_name_or_path gpt2 \
--dataset_name wikitext --dataset_config_name wikitext-2-raw-v1 --do_train \
--output_dir /tmp/test-clm --per_device_train_batch_size 4 --max_steps 200

{'train_runtime': 101.9003, 'train_samples_per_second': 1.963, 'epoch': 0.69}

# DDP w/o NVLink

rm -r /tmp/test-clm; CUDA_VISIBLE_DEVICES=0,1 NCCL_P2P_DISABLE=1 python -m torch.distributed.launch \
--nproc_per_node 2 examples/pytorch/language-modeling/run_clm.py --model_name_or_path gpt2 \
--dataset_name wikitext --dataset_config_name wikitext-2-raw-v1 --do_train
--output_dir /tmp/test-clm --per_device_train_batch_size 4 --max_steps 200

{'train_runtime': 131.4367, 'train_samples_per_second': 1.522, 'epoch': 0.69}
```

Hardware: 2x TITAN RTX 24GB each + NVlink with 2 NVLinks (`nv2 in nvidia-smi topo -m`) Software: `pytorch-1.8-to-be + cuda-11.0 / transformers==4.3.0.dev0`

Orchestration

There are many container/accelerator orchestration solutions - many of which are open source.

So far I have been working with SLURM:

- [SLURM](#) - Simple Linux Utility for Resource Management, which you're guaranteed to find on most HPC environments and typically it's supported by most cloud providers. It has been around for more than 2 decades

The other most popular orchestrator is Kubernetes:

- [Kubernetes](#) - also known as K8s, is an open source system for automating deployment, scaling, and management of containerized applications. Here is a good [comparison between SLURM and K8s](#).

Here are various other less popular, but still very mighty orchestration solutions:

- [dstack](#) is a lightweight, open-source alternative to Kubernetes & Slurm, simplifying AI container orchestration with multi-cloud & on-prem support. It natively supports NVIDIA, AMD, & TPU.
- [SkyPilot](#) is a framework for running AI and batch workloads on any infra, offering unified execution, high cost savings, and high GPU availability.
- [OpenHPC](#) provides a variety of common, pre-built ingredients required to deploy and manage an HPC Linux cluster including provisioning tools, resource management, I/O clients, runtimes, development tools, containers, and a variety of scientific libraries.
- [run.ai](#) - got acquired by NVIDIA and is planned to be open sourced soon.
- [Docker Swarm](#) is a container orchestration tool.
- [IBM Platform Load Sharing Facility \(LSF\)](#) Suites is a workload management platform and job scheduler for distributed high performance computing (HPC).

Working in SLURM Environment

Unless you're lucky and you have a dedicated cluster that is completely under your control chances are that you will have to use SLURM to timeshare the GPUs with others. But, often, if you train at HPC, and you're given a dedicated partition you still will have to use SLURM.

The SLURM abbreviation stands for: **Simple Linux Utility for Resource Management** - though now it's called The Slurm Workload Manager. It is a free and open-source job scheduler for Linux and Unix-like kernels, used by many of the world's supercomputers and computer clusters.

These chapters will not try to exhaustively teach you SLURM as there are many manuals out there, but will cover some specific nuances that are useful to help in the training process.

- [SLURM For Users](#) - everything you need to know to do your training in the SLURM environment.
- [SLURM Administration](#) - if you're unlucky to need to also manage the SLURM cluster besides using it, there is a growing list of recipes in this document to get things done faster for you.
- [Performance](#) - SLURM performance nuances.
- [Launcher scripts](#) - how to launch with `torchrun`, `accelerate`, `pytorch-lightning`, etc. in the SLURM environment

SLURM Administration

Run a command on multiple nodes

1. to avoid being prompted with:

```
Are you sure you want to continue connecting (yes/no/[fingerprint])?
```

for every new node you haven't logged into yet, you can disable this check with:

```
echo "Host *" >> ~/.ssh/config
echo " StrictHostKeyChecking no" >> ~/.ssh/config
```

Of course, check if that's secure enough for your needs. I'm making an assumption that you're already on the SLURM cluster and you're not ssh'ing outside of your cluster. You can choose not to set this and then you will have to manually approve each new node.

2. Install pdsh

You can now run the wanted command on multiple nodes.

For example, let's run date:

```
$ PDSH_RCMD_TYPE=ssh pdsh -w node-[21,23-26] date
node-25: Sat Oct 14 02:10:01 UTC 2023
node-21: Sat Oct 14 02:10:02 UTC 2023
node-23: Sat Oct 14 02:10:02 UTC 2023
node-24: Sat Oct 14 02:10:02 UTC 2023
node-26: Sat Oct 14 02:10:02 UTC 2023
```

Let's do something more useful and complex. Let's kill all GPU-tied processes that didn't exit when the SLURM job was cancelled:

First, this command will give us all process ids that tie up the GPUs:

```
nvidia-smi --query-compute-apps=pid --format=csv,noheader | sort | uniq
```

So we can now kill all those processes in one swoop:

```
PDSH_RCMD_TYPE=ssh pdsh -w node-[21,23-26] "nvidia-smi --query-compute-apps=pid --format=csv,noheader | sort | uniq | xargs -n1 sudo kill -9"
```

Slurm settings

Show the slurm settings:

```
sudo scontrol show config
```

The config file is `/etc/slurm/slurm.conf` on the slurm controller node.

Once `slurm.conf` was updated to reload the config run:

```
sudo scontrol reconfigure
```

from the controller node.

Auto-reboot

If the nodes need to be rebooted safely (e.g. if the image has been updated), adapt the list of the node and run:

```
scontrol reboot ASAP node-[1-64]
```

For each of the non-idle nodes this command will wait till the current job ends, then reboot the node and bring it back up to `idle`.

Note that you need to have:

```
RebootProgram = "/sbin/reboot"
```

set in `/etc/slurm/slurm.conf` on the controller node for this to work (and reconfigure the SLURM daemon if you have just added this entry to the config file).

Changing the state of the node

The change is performed by `scontrol update`

Examples:

To undrain a node that is ready to be used:

```
scontrol update nodename=node-5 state=idle
```

To remove a node from the SLURM's pool:

```
scontrol update nodename=node-5 state=drain
```

Undrain nodes killed due to slow process exit

Sometimes processes are slow to exit when a job has been cancelled. If the SLURM was configured not to wait forever it'll automatically drain such nodes. But there is no reason for those nodes to not be available to the users.

So here is how to automate it.

The keys is to get the list of nodes that are drained due to "Kill task failed", which is retrieved with:

```
sinfo -R | grep "Kill task failed"
```

now extract and expand the list of nodes, check that the nodes are indeed user-process free (or try to kill them first) and then undrain them.

Earlier you learned how to [run a command on multiple nodes](#) which we will use in this script.

Here is the script that does all that work for you: [undrain-good-nodes.sh](#)

Now you can just run this script and any nodes that are basically ready to serve but are currently drained will be switched to `idle` state and become available for the users to be used.

Modify a job's timelimit

To set a new timelimit on a job, e.g., 2 days:

```
scontrol update JobID=$SLURM_JOB_ID TimeLimit=2-00:00:00
```

To add additional time to the previous setting, e.g. 3 more hours.

```
scontrol update JobID=$SLURM_JOB_ID TimeLimit=+10:00:00
```

When something goes wrong with SLURM

Analyze the events log in the SLURM's log file:

```
sudo cat /var/log/slurm/slurmctld.log
```

This, for example, can help to understand why a certain node got its jobs cancelled before time or the node got removed completely.

SLURM for users

Quick start

Simply copy this [example.slurm](#) and adapt it to your needs.

SLURM partitions

In this doc we will use an example setup with these 2 cluster names:

- dev
- prod

To find out the hostname of the nodes and their availability, use:

```
sinfo -p dev
sinfo -p prod
```

Slurm configuration is at `/opt/slurm/etc/slurm.conf`.

To see the configuration of all partitions:

```
scontrol show partition
```

Wait time for resource granting

```
squeue -u `whoami` --start
```

will show when any pending jobs are scheduled to start.

They may start sooner if others cancel their reservations before the end of the reservation.

Request allocation via dependency

To schedule a new job when one or more of the currently scheduled job ends (regardless of whether it still running or not started yet), use the dependency mechanism, by telling `sbatch` to start the new job once the currently running job succeeds, using:

```
sbatch --dependency=CURRENTLY_RUNNING_JOB_ID tr1-13B-round1.slurm
```

Using `--dependency` may lead to shorter wait times than using `--begin`, since if the time passed to `--begin` allows even for a few minutes of delay since the stopping of the last job, the scheduler may already start some other jobs even if their priority is lower than our job. That's because the scheduler ignores any jobs with `--begin` until the specified time arrives.

Make allocations at a scheduled time

To postpone making the allocation for a given time, use:

```
salloc --begin HH:MM MM/DD/YY
```

Same for `sbatch`.

It will simply put the job into the queue at the requested time, as if you were to execute this command at this time. If resources are available at that time, the allocation will be given right away. Otherwise it'll be queued up.

Sometimes the relative begin time is useful. And other formats can be used. Examples:

```
--begin now+2hours
--begin=16:00
--begin=now+1hour
--begin=now+60  # seconds by default
--begin=2010-01-20T12:34:00
```

the time-units can be `seconds` (default), `minutes`, `hours`, `days`, or `weeks`:

Preallocated node without time 60min limit

This is very useful for running repetitive interactive experiments - so one doesn't need to wait for an allocation to progress. so the strategy is to allocate the resources once for an extended period of time and then running interactive `srun` jobs using this allocation.

set `--time` to the desired window (e.g. 6h):

```
salloc --partition=dev --nodes=1 --ntasks-per-node=1 --cpus-per-task=96 --gres=gpu:8 --time=6:00:00 bash
salloc: Pending job allocation 1732778
salloc: job 1732778 queued and waiting for resources
salloc: job 1732778 has been allocated resources
salloc: Granted job allocation 1732778
```

now use this reserved node to run a job multiple times, by passing the job id of `salloc`:

```
srun --jobid $SLURM_JOBID --pty bash
```

if run from inside `bash` started via `salloc`. But it can be started from another shell, but then explicitly set `--jobid`.

if this `srun` job timed out or manually exited, you can re-start it again in this same reserved node.

`srun` can, of course, call the real training command directly and not just `bash`.

Important: when allocating a single node, the allocated shell is not on the node (it never is). You have to find out the hostname of the node (reports when giving the allocation or via `squeue` and `ssh` to it).

When finished, to release the resources, either exit the shell started in `salloc` or `scancel JOBID`.

This reserved node will be counted towards hours usage the whole time it's allocated, so release as soon as done with it. Actually, if this is just one node, then it's even easier to not use `salloc` but to use `srun` in the first place, which will both allocate and give you the shell to use:

```
srun --pty --partition=dev --nodes=1 --ntasks=1 --cpus-per-task=96 --gres=gpu:8 --time=60 bash
```

Hyper-Threads

By default, if the cpu has [Hyper-Threads](#) (HT) enabled, SLURM will use it. If you don't want to use HT you have to specify `--hint=nomultithread`.

footnote: HT is Intel-specific naming, the general concept is simultaneous multithreading (SMT)

For example for a cluster with 2 cpus per node with 24 cores and 2 hyper-threads each, there is a total of 96 hyper-threads or 48 cpu-cores available. Therefore to utilize the node fully you'd need to configure either:

```
#SBATCH --cpus-per-task=96
```

or if you don't want HT:

```
#SBATCH --cpus-per-task=48
#SBATCH --hint=nomultithread
```

This last approach will allocate one thread per core and in this mode there are only 48 cpu cores to use.

Note that depending on your application there can be quite a performance difference between these 2 modes. Therefore try both and see which one gives you a better outcome.

On some setups like AWS the all-reduce throughput degrades dramatically when `--hint=nomultithread` is used! Whereas on some other setups the opposite is true - the throughput is worse without HT!

To check if your instances has HT enabled, run:

```
$ lscpu | grep Thread
Thread(s) per core: 2
```

If it's 2 then it is HT-enabled, if it's 1 then it isn't.

Reuse allocation

e.g. when wanting to run various jobs on identical node allocation.

In one shell:

```
salloc --partition=prod --nodes=16 --ntasks=16 --cpus-per-task=96 --gres=gpu:8 --time=3:00:00 bash
echo $SLURM_JOBID
```

In another shell:

```
export SLURM_JOBID=<JOB ID FROM ABOVE>
srun --jobid=$SLURM_JOBID ...
```

You may need to set `--gres=gpu:0` to run some diagnostics job on the nodes. For example, let's check shared memory of all the hosts:

```
srun --jobid 631078 --gres=gpu:0 bash -c 'echo $(hostname) $(df -h | grep shm)'
```

Specific nodes selection

To exclude specific nodes (useful when you know some nodes are broken, but are still in IDLE state):

```
sbatch --exclude nodeA,nodeB
```

or via: `#SBATCH --exclude ...`

To use specific nodes:

```
sbatch --nodelist= nodeA,nodeB
```

can also use the short `-w` instead of `--nodelist`

The administrator could also define a `feature=example` in `slurm.conf` and then a user could ask for that subset of nodes via `--constraint=example`

Signal the running jobs to finish

Since each SLURM run has a limited time span, it can be configured to send a signal of choice to the program a desired amount of time before the end of the allocated time.

```
--signal=[[R][B]:]<sig_num>[@<sig_time>]
```

TODO: need to experiment with this to help training finish gracefully and not start a new cycle after saving the last checkpoint.

Detailed job info

While most useful information is preset in various `SLURM_*` env vars, sometimes some information is missing. In such cases use:

```
scontrol show -d job $SLURM_JOB_ID
```

and then parse out what's needed.

For a job that finished its run use:

```
sacct -j JOBID
```

This command is also useful to discover if you have any `srun` jobs already running on that allocation (including those that were finished or cancelled). For example, you could kill some run-away `srun` step via `scancel <jobid>.<step-id>` and you'd find that `<step-id>` via the above command. The main job will continue running if it's an interactive job even if you cancelled all step jobs.

To see more details:

```
sacct -ojobid,start,end,state,exitcode --format nodelist%300 -j JOBID
sacct -j JOBID --long
```

Or to see all jobs with their sub-steps while limiting the listing to a specific partition and only for your own user:

```
sacct -u `whoami` --partition=dev -ojobid,start,end,state,exitcode --format nodelist%300
sacct -u `whoami` --partition=prod -ojobid,start,end,state,exitcode --format nodelist%300
```

To see how a particular job was launched and all of its `srun` sub-step command lines:

```
sacct -j JOBID -o submitline -P
```

show jobs

Show only my jobs:

```
squeue -u `whoami`
```

Show jobs by job id:

```
squeue -j JOBID
```

Show jobs of a specific partition:

```
squeue --partition=dev
```

Aliases

Handy aliases

```
alias myjobs='squeue -u `whoami` -o "%.16i %9P %26j %.8T %.10M %.81 %.6D %.20S %R"
alias groupjobs='squeue -u foo,bar,tar -o "%.16i %u %9P %26j %.8T %.10M %.81 %.6D %.20S %R"
alias myjobs-pending="squeue -u `whoami` --start"
alias idle-nodes="sinfo -p prod -o '%A'"
```

Zombies

If there are any zombies left behind across nodes, send one command to kill them all.

```
srun pkill python
```

Detailed Access to SLURM Accounting

`sacct` displays accounting data for all jobs and job steps in the Slurm job accounting log or Slurm database.

So this is a great tool for analysing past events.

For example, to see which nodes were used to run recent gpu jobs:

```
sacct -u `whoami` --partition=dev -o jobid,start,end,state,exitcode --format nodelist%300
```

`%300` here tells it to use a 300 char width for the output, so that it's not truncated.

See `man sacct` for more fields and info fields.

Queue

Cancel job

To cancel a job:

```
scancel [jobid]
```

To cancel all of your jobs:

```
scancel -u <userid>
```

To cancel all of your jobs on a specific partition:

```
scancel -u <userid> -p <partition>
```

Tips

- if you see that `salloc`'ed interactive job is scheduled to run much later than you need, try to cancel the job and ask for shorter period - often there might be a closer window for a shorter time allocation.

Logging

If we need to separate logs to different log files per node add `%N` (for short hostname) so that we have:

```
#SBATCH --output=%x-%j-%N.out
```

That way we can tell if a specific node misbehaves - e.g. has a corrupt GPU. This is because currently pytorch doesn't log which node / gpu rank triggered an exception.

Hoping it'll be a built-in feature of pytorch <https://github.com/pytorch/pytorch/issues/63174> and then one won't need to make things complicated on the logging side.

Show the state of nodes

```
sinfo -p PARTITION
```

Very useful command is:

```
sinfo -s
```

and look for the main stat, e.g.:

```
NODES(A/I/O/T) "allocated/idle/other/total".  
597/0/15/612
```

So here 597 out of 612 nodes are allocated. 0 idle and 15 are not available for whatever other reasons.

```
sinfo -p gpu_p1 -o "%A"
```

gives:

```
NODES(A/I)  
236/24
```

so you can see if any nodes are available on the 4x v100-32g partition (`gpu_p1`)

To check a specific partition:

```
sinfo -p gpu_p1 -o "%A"
```

See the table at the top of this document for which partition is which.

sinfo states

- idle: no jobs running
- alloc: nodes are allocated to jobs that are currently executing
- mix: the nodes have some of the CPUs allocated, while others are idle
- drain: the node is unavailable due to an administrative reason
- drng: the node is running a job, but will after completion not be available due to an administrative reason

Node state codes

The node state could be followed by a single character which has a special meaning. It is one of:

- *: The node is presently not responding and will not be allocated any new work. If the node remains non-responsive, it will be placed in the DOWN state (except in the case of COMPLETING, DRAINED, DRAINING, FAIL, FAILING nodes).
- ~: The node is presently in powered off.
- #: The node is presently being powered up or configured.
- !: The node is pending power down.
- %: The node is presently being powered down.
- \$: The node is currently in a reservation with a flag value of "maintenance".
- @: The node is pending reboot.
- ^: The node reboot was issued.
- -: The node is planned by the backfill scheduler for a higher priority job.

Job state codes

- CD | Completed: The job has completed successfully.
- CG | Completing: The job is finishing but some processes are still active.
- F | Failed: The job terminated with a non-zero exit code and failed to execute.
- PD | Pending: The job is waiting for resource allocation. It will eventually run.
- PR | Preempted: The job was terminated because of preemption by another job.
- R | Running: The job currently is allocated to a node and is running.
- S | Suspended: A running job has been stopped with its cores released to other jobs.
- ST | Stopped: A running job has been stopped with its cores retained.

drained nodes

To see all drained nodes and the reason for drainage (edit %50E to make the reason field longer/shorter)

```
% sinfo -R -o "%50E %12U %19H %6t %N"
```

or just -R if you want it short:

```
% sinfo -R
```

Job arrays

To run a sequence of jobs, so that the next slurm job is scheduled as soon as the currently running one is over in 20h we use a job array.

Let's start with just 10 such jobs:

```
sbatch --array=1-10%1 array-test.slurm
```

%1 limits the number of simultaneously running tasks from this job array to 1. Without it it will try to run all the jobs at once, which we may want sometimes (in which case remove %1), but when training we need one job at a time.

Alternatively, as always this param can be part of the script:

```
#SBATCH --array=1-10%1
```

Here is toy slurm script, which can be used to see how it works:

```
#!/bin/bash
#SBATCH --job-name=array-test
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1          # crucial - only 1 task per dist per node!
#SBATCH --cpus-per-task=1           # number of cores per tasks
#SBATCH --time 00:02:00             # maximum execution time (HH:MM:SS)
#SBATCH --output=%x-%j.out          # output file name
#SBATCH --error=%x-%j.out           # error file name (same to watch just one file)
#SBATCH --partition=dev

echo $SLURM_JOB_ID
echo "I am job ${SLURM_ARRAY_JOB_ID}_${SLURM_ARRAY_TASK_ID}"
date
sleep 10
date
```

Note \$SLURM_ARRAY_JOB_ID is the same as \$SLURM_JOB_ID, and \$SLURM_ARRAY_TASK_ID is the index of the job.

To see the jobs running:

```
$ squeue -u `whoami` -o "%10i %9P %26j %.8T %.10M %.6D %.20S %R"
  JOBID PARTITION          NAME      STATE      TIME  NODES      START_TIME
NODELIST(REASON)
591970_[2-  dev      array-test  PENDING      0:00      1 2021-07-28T20:01:06
(JobArrayTaskLimit)
```

now job 2 is running.

To cancel the whole array, cancel the job id as normal (the number before _):

```
scancel 591970
```

To cancel a specific job:

```
scancel 591970_2
```

If it's important to have the log-file contain the array id, add `%A_%a`:

```
#SBATCH --output=%x-%j.%A_%a.log
```

More details https://slurm.schedmd.com/job_array.html

Job array trains and their suspend and release

In this recipe we accomplish 2 things:

1. Allow modification to the next job's slurm script
2. Allow suspending and resuming job arrays w/o losing the place in the queue when not being ready to continue running a job

SLURM is a very unforgiving environment where a small mistake can cost days of waiting time. But there are strategies to mitigate some of this harshness.

SLURM jobs have a concept of "age" in the queue which besides project priority governs when a job gets scheduled to run. If your have just scheduled a new job it has no "age" and will normally be put to run last compared to jobs that have entered the queue earlier. Unless of course this new job comes from a high priority project in which case it'll progress faster.

So here is how one can keep the "age" and not lose it when needing to fix something in the running script or for example to switch over to another script.

The idea is this:

1. `sbatch` a long job array, e.g., `-array=1-50%`
2. inside the slurm script don't have any code other than `source another-script.slurm` - so now you can modify the target script or symlink to another script before the next job starts
3. if you need to stop the job array train - don't cancel it, but suspend it without losing your place in a queue
4. when ready to continue - unsuspend the job array - only the time while it was suspended is not counted towards its age, but all the previous age is retained.

The number of nodes, time and hardware and partition of a running job cannot be modified, but you can change pending jobs in the job array by `scontrol update jobid=<desired_job_id> numnodes=<new number> partition=<new partition>`.

If you do have `sudo` access then you can change the job time of the current job as well.

Here is an example:

Create a job script:

```
$ cat train-64n.slurm
#!/bin/bash
#SBATCH --job-name=tr8-104B
#SBATCH --nodes=64
#SBATCH --ntasks-per-node=1          # crucial - only 1 task per dist per node!
#SBATCH --cpus-per-task=96          # number of cores per tasks
#SBATCH --gres=gpu:8                # number of gpus
#SBATCH --time 20:00:00              # maximum execution time (HH:MM:SS)
```

```
#SBATCH --output=%x-%j.out          # output file name
#SBATCH --partition=dev

source tr8-104B-64.slurm
```

Start it as:

```
sbatch --array=1-50%1 train-64.slurm
```

Now you can easily edit `tr8-104B-64.slurm` before the next job run and either let the current job finish if it's desired or if you need to abort it, just kill the currently running job, e.g. `1557903_5` (not job array `1557903`) and have the train pick up where it left, but with the edited script.

The nice thing is that this requires no changes to the original script (`tr8-104B-64.slurm` in this example), and the latter can still be started on its own.

Now, what if something is wrong and you need 10min or 10h to fix something. In this case we suspend the train using:

```
scontrol hold <jobid>
```

with `being` either a "normal" job, the id of a job array or the id for a job array step

and then when ready to continue release the job:

```
scontrol release <jobid>
```

How to keep an salloc allocation alive while exiting its shell

If you run allocated a node like so:

```
salloc --partition=dev --nodes=1 --ntasks-per-node=1 --time=1:00:00 bash
```

and you exited the shell, or your ssh connection got dropped, the allocation will be lost.

If you want to open an allocation that should survive exiting the shell, use `--no-shell` and `no bash` like so:

```
salloc --no-shell --partition=dev --nodes=1 --ntasks-per-node=1 --time=1:00:00
```

and now if you need to join the session see [How to rejoin the allocated node interactively](#).

But beware, that if you `ssh` to the allocated node and launch something normally and then close the connection that job will be lost as the connecting shell will send `SIGHUP` to its child processes. To avoid that and to keep the job running use `nohup` while putting the program into a background process. Example:

```
nohup my-program &
```

nohup will ignore `SIGHUP` and will redirect stderr to stdout and append stdout to a special file `nohup.out`. If you want to control where the std streams should be written use normal stdout redirect `>` or `>>`, e.g.:

```
nohup my-program >> some-file.txt &
```

As mentioned earlier the program is also sent into the background with `&`.

Now you can safely disconnect and the program will continue running when you come back.

This solution will prevent the program from exiting, but you won't be able to interact with it normally when you connect again as the std streams will be redirected. You can of course still kill the program via its pid, change its `nice` state, etc., like you'd do with any other process.

But if you want to use something where you can disconnect and reconnect and continue using the program normally you'd have to use a [terminal multiplexer](#) program like [tmux](#) or [GNU screen](#) which run a daemon on the node and allow you to regain the normal control over the program on reconnection. There are also [mosh](#) and other similar tools which further aid this process.

How to rejoin the allocated node interactively

To have multiple interactive shells into the same job `--overlap` should be used.

For example, in console A, let's allocate a single node:

```
$ salloc --partition=dev --nodes=1 --ntasks-per-node=1 --cpus-per-task=26 --gres=gpu:1 --time=2:00:00 bash
salloc: Granted job allocation 1916
salloc: Nodes my-node-1 are ready for job
```

In console B:

```
$ srun --overlap --pty --jobid 101 bash
```

and the above can be repeated in as many consoles as wanted.

If it's the first pseudo terminal shell you don't even need `--overlap`, but you need it for the additional shells.

It works the same if you initially allocated the node via `srun --pty`

```
srun --pty -p dev --gpus 8 --time=2:00:00 bash
```

You can, of course, also access the node via `ssh` but if your SLURM has been setup to do all kinds of virtualizations (e.g. give only a few GPUs to each user, or virtualize `/tmp/` or `/scratch` with auto-cleanup on exit), the view from `ssh` won't be the same. For example, if a job allocated 2 GPUs, the `ssh` shell will show all of the GPUs and not just the 2 - so if you're sharing the node with others this won't work well.

This works for multi-node allocations and by default you will get an interactive shell on the first node of the allocation. If

you want to enter a specific node use `-w` to specify it. For example, say you got `node-[1-4]` allocated and you want to enter `node-3`, then specify:

```
srun --pty -p dev --gpus 8 --time=2:00:00 -w node-3 bash
```

and if it fails with:

```
srun: error: Unable to create step for job 1930: Invalid generic resource (gres) specification
```

add back the `--gres=gpu:8` setting. You won't need to do it if your original allocation command used this flag already.

Troubleshooting

SLURM_PROCID early interpolation

When using SLURM with multi-node setup it's crucial that this is set correctly:

```
--machine_rank \$SLURM_PROCID"
```

it must not be interpolated before time, since if this is set as `--machine_rank \$SLURM_PROCID` the launcher will hang.

It's best to isolate the launcher from the program like so:

```
export MASTER_ADDR=$(scontrol show hostnames \$SLURM_JOB_NODELIST | head -n 1)
export MASTER_PORT=3333
ACCELERATE_CONFIG_FILE=path/to/accelerate.config.yaml # edit me
LAUNCHER="python -u -m accelerate.commands.launch \
    --rdzv_conf "rdzv_backend=c10d,rdzv_endpoint=\$MASTER_ADDR:\$MASTER_PORT" \
    --config_file \$ACCELERATE_CONFIG_FILE \
    --main_process_ip \$MASTER_ADDR \
    --main_process_port \$MASTER_PORT \
    --machine_rank \$SLURM_PROCID \
    --role \$(hostname -s|tr -dc '0-9'): --tee 3 \
"
PROGRAM="myprogram.py"

CMD="\$LAUNCHER \$PROGRAM"

SRUN_ARGS=" \
    --wait=60 \
    --kill-on-bad-exit=1 \
    --unbuffered \
    --jobid \$SLURM_JOBID \
"
srun \$SRUN_ARGS bash -c "\$CMD" 2>&1 | tee -a main_log.txt
```

Now the launcher will always work and the users will only need to tweak the `PROGRAM` variable.

With `torchrun`:

```
export $GPUS_PER_NODE=8
export MASTER_ADDR=$(scontrol show hostnames $SLURM_JOB_NODELIST | head -n 1)
export MASTER_PORT=3333
LAUNCHER="python -u -m torch.distributed.run \
--nproc_per_node $GPUS_PER_NODE \
--nnodes $NNODES \
--node_rank \$SLURM_PROCID \
--rdzv_endpoint $MASTER_ADDR:$MASTER_PORT \
--rdzv_backend c10d \
--max_restarts 0 \
--role `hostname -s`::--tee 3 \
"
"
```

See [Single and Multi-node Launchers with SLURM](#) for complete working examples.

Mismatching nodes number

If the pytorch launcher fails it often means that the number of SLURM nodes and the launcher nodes are mismatching, e.g.:

```
grep -ir nodes= tr123-test.slurm
#SBATCH --nodes=40
NNODES=64
```

This won't work. They have to match.

You can add a sanity check to your script:

```
#!/bin/bash
#SBATCH --job-name=test-mismatch
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=1          # crucial - only 1 task per dist per node!
#SBATCH --cpus-per-task=96          # number of cores per tasks
#SBATCH --gres=gpu:8                # number of gpus
#SBATCH --time 0:05:00              # maximum execution time (HH:MM:SS)
#SBATCH --output=%x-%j.out          # output file name
#SBATCH --partition=prod

[...]

NNODES=2

# sanity check for having NNODES and `#SBATCH --nodes` match, assuming you use NNODES variable
if [ "$NNODES" != "$SLURM_NNODES" ]; then
    echo "Misconfigured script: NNODES=$NNODES != SLURM_NNODES=$SLURM_NNODES"
```

```
    exit 1
fi
[...]
```

or you could just do:

```
#SBATCH --nodes=2
[...]
NNODES=$SLURM_NNODES
```

and then it will always be correct

Find faulty nodes and exclude them

Sometimes a node is broken, which prevents one from training, especially since restarting the job often hits the same set of nodes. So one needs to be able to isolate the bad node(s) and exclude it from `sbatch`.

To find a faulty node, write a small script that reports back the status of the desired check.

For example to test if cuda is available on all nodes:

```
python -c 'import torch, socket; print(f"{{socket.gethostname()}}: {{torch.cuda.is_available()}}")'
```

and to only report the nodes that fail:

```
python -c 'import torch, socket; torch.cuda.is_available() or print(f"Broken node:
{{socket.gethostname()}}")'
```

Of course, the issue could be different - e.g. gpu can't allocate memory, so change the test script to do a small allocation on cuda. Here is one way:

```
python -c "import torch; torch.ones(1000,1000).cuda()"
```

But since we need to run the test script on all nodes and not just the first node, the slurm script needs to run it via `srun`. So our first diagnostics script can be written as:

```
srun --jobid $SLURM_JOBID bash -c 'python -c "import torch, socket; print(socket.gethostname(),
torch.cuda.is_available())"'
```

I slightly changed it, due to an issue with quotes.

You can always convert the one liner into a real script and then there is no issue with quotes.

```
$ cat << EOT >> test-nodes.py
#!/usr/bin/env python
import torch, socket
print(socket.gethostname(), torch.cuda.is_available())
EOT
$ chmod a+x ./test-nodes.py
```

Now let's create a driver slurm script. Use a few minutes time for this test so that SLURM yields it faster:

```
#!/bin/bash
#SBATCH --job-name=test-nodes
#SBATCH --nodes=4
#SBATCH --ntasks-per-node=1          # crucial - only 1 task per dist per node!
#SBATCH --cpus-per-task=96          # number of cores per tasks
#SBATCH --gres=gpu:8                # number of gpus
#SBATCH --time 0:05:00              # maximum execution time (HH:MM:SS)
#SBATCH --output=%x-%j.out          # output file name
#SBATCH --partition=prod

source $six_ALL_CCFRWORK/start-prod
srun --jobid ${SLURM_JOBID} ./test-nodes.py
```

Once it runs check the logs to see if any reported `False`, those are the nodes you want to exclude.

Now once the faulty node(s) is found, feed it to `sbatch`:

```
sbatch --exclude=hostname1,hostname2 ...
```

and `sbatch` will exclude the bad nodes from the allocation.

Additionally please report the faulty nodes to `#science-support` so that they get replaced

Here are a few more situations and how to find the bad nodes in those cases:

Broken NCCL

If you're testing something that requires distributed setup, it's a bit more complex. Here is a slurm script that tests that NCCL works. It sets up NCCL and checks that barrier works:

```
#!/bin/bash
#SBATCH --job-name=test-nodes-nccl
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=1          # crucial - only 1 task per dist per node!
#SBATCH --cpus-per-task=96          # number of cores per tasks
#SBATCH --gres=gpu:8                # number of gpus
#SBATCH --time 0:05:00              # maximum execution time (HH:MM:SS)
#SBATCH --output=%x-%j.out          # output file name
#SBATCH --partition=prod
```

```

source $six_ALL_CCFRWORK/start-prod

NNODES=2

GPUS_PER_NODE=4
MASTER_ADDR=$(scontrol show hostnames $SLURM_JOB_NODELIST | head -n 1)
MASTER_PORT=6000

export LAUNCHER="python -u -m torch.distributed.launch \
--nproc_per_node $GPUS_PER_NODE \
--nnodes $NNODES \
--master_addr $MASTER_ADDR \
--master_port $MASTER_PORT \
"

export SCRIPT=test-nodes-nccl.py

cat << EOT > $SCRIPT
#!/usr/bin/env python
import torch.distributed as dist
import torch
import socket
import os
import fcntl

def printflock(*msgs):
    """ print """
    with open(__file__, "r") as fh:
        fcntl.flock(fh, fcntl.LOCK_EX)
        try:
            print(*msgs)
        finally:
            fcntl.flock(fh, fcntl.LOCK_UN)

local_rank = int(os.environ["LOCAL_RANK"])
torch.cuda.set_device(local_rank)
dist.init_process_group("nccl")
header = f"{socket.gethostname()}-{local_rank}"
try:
    dist.barrier()
    printflock(f"{header}: NCCL {torch.cuda.nccl.version()} is OK")
except:
    printflock(f"{header}: NCCL {torch.cuda.nccl.version()} is broken")
    raise
EOT

echo $LAUNCHER --node_rank $SLURM_PROCID $SCRIPT

srun --jobid $SLURM_JOBID bash -c '$LAUNCHER --node_rank $SLURM_PROCID $SCRIPT'

```

The script uses `printflock` to solve the interleaved print outputs issue.

GPU Memory Check

This tests if each GPU on the allocated nodes can successfully allocate 77Gb (e.g. to test 80GB A100s) (have to subtract a few GBs for cuda kernels).

```
import torch, os
import time
import socket
hostname = socket.gethostname()

local_rank = int(os.environ["LOCAL_RANK"]);

gbs = 77
try:
    torch.ones((gbs*2**28)).cuda(local_rank).contiguous() # alloc on cpu, then move to gpu
    print(f"{local_rank} {hostname} is OK")
except:
    print(f"{local_rank} {hostname} failed to allocate {gbs}GB DRAM")
    pass

time.sleep(5)
```

Broken Network

Yet another issue with a node is when its network is broken and other nodes fail to connect to it.

You're likely to experience it with an error similar to:

```
work = default_pg.barrier(opts=opts)
RuntimeError: NCCL error in: /opt/conda/conda-bld/pytorch_1616554793803/work/torch/lib/c10d/
ProcessGroupNCCL.cpp:825, unhandled system error, NCCL version 2.7.8
ncclSystemError: System call (socket, malloc, munmap, etc) failed.
```

Here is how to debug this issue:

1. Add:

```
export NCCL_DEBUG=INFO
```

before the `srun` command and re-run your slurm script.

2. Now study the logs. If you find:

```
r11i6n2:486514:486651 [1] include/socket.h:403 NCCL WARN Connect to 10.148.3.247<56821> failed :
Connection refused
```

Let's see which node refuses to accept connections. We get the IP address from the error above and reverse resolve it to its name:

```
nslookup 10.148.3.247
247.3.148.10.in-addr.arpa      name = r10i6n5.ib0.xa.idris.fr.
```

Add `--exclude=r10i6n5` to your `sbatch` command and report it to JZ admins.

Run py-spy or any other monitor program across all nodes

When dealing with hanging, here is how to automatically log py-spy traces for each process.

Of course, this same process can be used to run some command for all nodes of a given job. i.e. it can be used to run something during the normal run - e.g. dump all the memory usage in each process via `nvidia-smi` or whatever other program is needed to be run.

```
cd ~/prod/code/tr8b-104B/bigscience/train/tr11-200B-ml/
salloc --partition=prod --nodes=40 --ntasks-per-node=1 --cpus-per-task=96 --gres=gpu:8 --time 20:00:00
bash 200B-n40-bf16-mono.slurm
```

In another shell get the JOBID for the above `salloc`:

```
squeue -u `whoami` -o "%16i %9P %26j %.8T %.10M %.81 %.6D %.20S %R"
```

adjust jobid per above and the nodes count (XXX: probably can remove `--nodes=40` altogether and rely on `salloc` config):

```
srun --jobid=2180718 --gres=gpu:0 --nodes=40 --tasks-per-node=1 --output=trace-%N.out sh -c 'ps aux | grep
python | egrep -v "grep|srun" | grep `whoami` | awk "{print \$2}" | xargs -I {} py-spy dump --native --pid
{}' || echo "failed"
```

now all py-spy traces go into the `trace-$nodename.out` files under `cwd`.

The key is to use `--gres=gpu:0` or otherwise the 2nd `srun` will block waiting for the first one to release the gpus.

Also the assumption is that some conda env that has py-spy installed got activated in `~/.bashrc`. If yours doesn't already do that, add the instruction to load the env to the above command, before the py-spy command - it'll fail to find it otherwise.

Don't forget to manually release the allocation when this process is done.

Convert SLURM_JOB_NODELIST into a hostfile

Some multi-node launchers require a `hostfile` - here is how to generate one:

```
# autogenerate the hostfile for deepspeed
```

```

# 1. deals with: SLURM_JOB_NODELIST in either of 2 formats:
# r10i1n8,r10i2n0
# r10i1n[7-8]
# 2. and relies on SLURM_STEP_GPUS=0,1,2... to get how many gpu slots per node
#
# usage:
# makehostfile > hostfile
function makehostfile() {
perl -le '$slots=split //, $ENV{"SLURM_STEP_GPUS"}; $_=$ENV{"SLURM_JOB_NODELIST"}; if
(/^(.*?)\[(\d+)-(\d+)\]\)/ { print map { "$1$_ slots=$slots\n" } $2..$3} elsif (/,/) { print map { "$1$_
slots=$slots\n" } split /,/ }
}

```

Environment variables

You can always do:

```
export SOMEKEY=value
```

from the slurm script to get a desired environment variable passed to the program launched from it.

And you can also add to the top of the slurm script:

```
#SBATCH --export=ALL
```

The launched program will see all the environment variables visible in the shell where it was launched from.

Crontab Emulation

One of the most important Unix tools is the crontab, which is essential for being able to schedule various jobs. It however usually is absent from SLURM environment. Therefore one must emulate it. Here is how.

For this presentation we are going to use `$WORK/cron` as the base directory. And that you have an exported environment variable `WORK` pointing to some location on your filesystem - if you use Bash you can set it up in your `~/.bash_profile` or if a different shell is used use whatever startup equivalent file is.

1. A self-perpetuating scheduler job

We will use `$WORK/cron/scheduler` dir for scheduler jobs, `$WORK/cron/cron.daily` for daily jobs and `$WORK/cron/cron.hourly` for hourly jobs:

```

$ mkdir -p $WORK/cron/scheduler
$ mkdir -p $WORK/cron/cron.daily
$ mkdir -p $WORK/cron/cron.hourly

```

Now copy these two slurm script in `$WORK/cron/scheduler`:

- [cron-daily.slurm](#)

- [cron-hourly.slurm](#)

after editing those to fit your specific environment's account and partition information.

Now you can launch the crontab scheduler jobs:

```
$ cd $WORK/cron/scheduler
$ sbatch cron-hourly.slurm
$ sbatch cron-daily.slurm
```

This is it, these jobs will now self-perpetuate and usually you don't need to think about it again unless there is an even that makes SLURM lose all its jobs.

2. Daily and Hourly Cronjobs

Now whenever you want some job to run once a day, you simply create a slurm job and put it into the `$WORK/cron/cron.daily` dir.

Here is an example job that runs daily to update the `mlocate` file index:

```
$ cat $WORK/cron/cron.daily/mlocate-update.slurm
#!/bin/bash

#SBATCH --job-name=mlocate-update      # job name
#SBATCH --ntasks=1                     # number of MP tasks
#SBATCH --nodes=1
#SBATCH --hint=nomultithread          # we get physical cores not logical
#SBATCH --time=1:00:00                  # maximum execution time (HH:MM:SS)
#SBATCH --output=%x-%j.out             # output file name
#SBATCH --partition=PARTITION        # edit me
#SBATCH --account=GROUP@PARTITION # edit me

set -e
date
echo "updating mlocate db"
/usr/bin/updatedb -o $WORK/lib/mlocate/work.db -U $WORK --require-visibility 0
```

This builds an index of the files under `$WORK` which you can then quickly query with:

```
/usr/bin/locate -d $WORK/lib/mlocate/work.db pattern
```

To stop running this job, just move it out of the `$WORK/cron/cron.daily` dir.

The same principle applies to jobs placed into the `$WORK/cron/cron.hourly` dir. These are useful for running something every hour.

Please note that this crontab implementation is approximate timing-wise, due to various delays in SLURM scheduling they will run approximately every hour and every day. You can recode these to ask SLURM to start something at a more precise time if you have to, but most of the time the just presented method works fine.

Additionally, you can code your own variations to meet specific needs of your project, e.g., every-30min or every-12h jobs.

3. Cleanup

Finally, since every cron launcher job will leave behind a log file (which is useful if for some reason things don't work), you want to create a cronjob to clean up these logs. Otherwise you may run out of inodes - these logs files are tiny, but there could be tens of thousands of those.

You could use something like this in a daily job.

```
find $WORK/cron -name "*.out" -mtime +7 -exec rm -f {} +
```

Please note that it's set to only delete files that are older than 7 days, in case you need the latest logs for diagnostics.

Nuances

The scheduler runs with Unix permissions of the person who launched the SLURM cron scheduler job and so all other SLURM scripts launched by that cron job.

Self-perpetuating SLURM jobs

The same approach used in [building a scheduler](#) can be used for creating stand-alone self-perpetuating jobs.

For example:

```
#!/bin/bash
#SBATCH --job-name=watchdog          # job name
#SBATCH --ntasks=1                   # number of MP tasks
#SBATCH --nodes=1
#SBATCH --time=0:30:00                # maximum execution time (HH:MM:SS)
#SBATCH --output=%x-%j.out           # output file name
#SBATCH --partition=PARTITION        # edit me

# ensure to restart self first 1h from now
RUN_FREQUENCY_IN_HOURS=1
sbatch --begin=now+${RUN_FREQUENCY_IN_HOURS}hour watchdog.slurm

... do the watchdog work here ...
```

and you launch it once with:

```
sbatch watchdog.slurm
```

This then will immediately schedule itself to be run 1 hour from the launch time and then the normal job work will be done. Regardless of whether the rest of the job will succeed or fail, this job will continue relaunching itself approximately once an hour. This is imprecise due to scheduler job starting overhead and node availability issues. But if there is at least one spare node available and the job itself is quick to finish the requirement to run at an approximate frequency should be sufficient.

As the majority of SLURM environment in addition to the expensive GPU nodes also provide much cheaper CPU-only nodes, you should choose a CPU-only SLURM partition for any jobs that don't require GPUs to run.

Getting information about the job

From within the slurm file one can access information about the current job's allocations.

Getting allocated hostnames and useful derivations based on that:

```
export HOSTNAMES=$(scontrol show hostnames "$SLURM_JOB_NODELIST")
export NUM_NODES=$(scontrol show hostnames "$SLURM_JOB_NODELIST" | wc -l)
export MASTER_ADDR=$(scontrol show hostnames "$SLURM_JOB_NODELIST" | head -n 1)
```

Convert compact node list to expanded node list

Sometimes you get SLURM tools give you a string like: `node-[42,49-51]` which will require some coding to expand it into `node-42, node-49, node-50, node-51`, but there is a special tool to deal with that:

```
$ scontrol show hostnames node-[42,49-51]
node-42
node-49
node-50
node-51
```

Voila!

case study: this is for example useful if you want get a list of nodes that were drained because the job was too slow to exit, but really there is no real problem with the nodes. So this one-liner will give you the list of such nodes in an expanded format which you can then script to loop over this list to undrain these nodes after perhaps checking that the processes have died by this time:

```
sinfo -R | grep "Kill task failed" | perl -lne '/(node-.*[\d\w]+)/ && print $1' | xargs -n1 scontrol show hostnames
```

Overcoming the lack of group SLURM job ownership

SLURM runs on Unix, but surprisingly its designers haven't adopted the concept of group ownership with regards to SLURM jobs. So if a member of your team started an array of 10 jobs 20h each, and went on vacation - unless you have `sudo` access you now can't do anything to stop those jobs if something is wrong.

I'm yet to find why this is so, but so far we have been using a kill switch workaround. You have to code it in your framework. For example, see how it was implemented in [Megatron-Deepspeed](#) (Meg-DS). The program polls for a pre-configured at start up path on the filesystem and if it finds a file there, it exits.

So if we start Meg-DS with `--kill-switch-path $WORK/tmp/training17-kill-switch` and then at any point we need to kill the SLURM job, we simply do:

```
touch $WORK/tmp/training17-kill-switch
```

and the next time the program gets to check for this file it'll detect the event and will exit voluntarily. If you have a job

array, well, you will have to wait until each job starts, detects the kill switch and exits.

Of course, don't forget to remove it when you're done stopping the jobs.

```
rm $WORK/tmp/training17-kill-switch
```

Now, this doesn't always work. If the job is hanging, it'll never come to the point of checking for kill-switch and the only solution here is to contact the sysadmins to kill the job for you. Sometimes if the hanging is a simple case pytorch's distributed setup will typically auto-exit after 30min of preset timeout time, but it doesn't always work.

How to gracefully exit on SLURM job preemption

There are several ways to gracefully handle time- and QoS-based SLURM pre-emption which are covered indepth in this section: [Dealing with forced job preemption](#).

How many GPUs a job uses

To figure out how many gpus are used by an already running job, parse the `JOB_GRES=gpu:` entry in `show job -d` output. For example, if the job was started with:

```
srun --pty --partition=dev --nodes=2 --ntasks-per-node=1 --gres=gpu:8 --time=8:00:00 bash
```

that is we allocated 16 GPUs, we can now get that number back programmatically via:

```
$ TOTAL_JOB_GPUS=$(scontrol show job -d $SLURM_JOBID | perl -ne 'm|JOB_GRES=gpu:(\d+)| && print $1')
$ echo $TOTAL_JOB_GPUS
16
```

Replace `$SLURM_JOBID` with the SLURM job id if it's not already set in the shell you run the command from ([squeue](#)).

How long did the job take to run

While normally `squeue` will show you the duration of the currently running job, in order to see how long a job run for when it finished, you need to know the job id and then you can query it like so:

```
$ sacct -j 22171 --format=JobID,JobName,State,Elapsed
JobID      JobName      State      Elapsed
-----
22171      example      COMPLETED  00:01:49
```

so we know the job finished running in under 2min.

FairShare

Many SLURM clusters use the FairShare system where the more someone uses the cluster the less of the priority they get to run jobs or if there is a pre-emption in place they are more likely to get pre-empted

To see your FairShare scores run:

```
sshare
```

Example:

Account	User	RawShares	NormShares	RawUsage	EffectvUsage	FairShare
root			0.000000	711506073	1.000000	
all		1	0.500000	711506073	1.000000	
all	stas	1	0.022727	14106989	0.019827	0.288889

If your FairShare score is more than 0.5 that means you have been using the cluster less than what you have been allocated, if it's less than 0.5 it means you have been using more than what was allocated.

As the time passes this score gets decayed so if you were having a very low score and have been using the cluster much less then your score will raise over time.

To see the score of a specific user:

```
sshare -u username
```

To see everybody's scores, sorted by FairShare:

```
sshare --all | sort -nk7 -r
```

This is the most important output, since it doesn't really matter what your score is alone. What matters is your score relative to all other users. Everybody who has a higher score than you will have a higher chance at getting their job yielded first and a lower chance of getting their job preempted.

Besides FairShare the priorities are typically configured based on a combination of multiple metrics, usually including the length of time a job has been waiting in the queue, job size, Quality of Service (QOS) setting, partition specifics, etc. The specifics will depend on how the slurm has been configured by your sysadmin.

SLURM Performance

Here you will find discussions of SLURM-specific settings that impact performance.

srun's `--cpus-per-task` may need to be explicit

You need to make sure that the launched by `srun` program receives as many cpu-cores as intended. For example, in a typical case of a ML training program, each gpu needs at least one cpu-core for the process driving it plus a few more cores for the `DataLoader`. You need multiple cores so that each task can be performed in parallel. If you have 8 gpus and 2 `DataLoader` workers per gpu, you need at least $3*8=24$ cpu-cores per node.

The number of cpus per task is defined by `--cpus-per-task`, which is passed to `sbatch` or `salloc` and originally `srun` would inherit this setting. However, recently this behavior has changed:

A quote from the `sbatch` manpage:

NOTE: Beginning with 22.05, `srun` will not inherit the `--cpus-per-task` value requested by `salloc` or `sbatch`. It must be requested again with the call to `srun` or set with the `SRUN_CPUS_PER_TASK` environment variable if desired for the task(s).

Which means that if in the past your SLURM script could have been:

```
#SBATCH --cpus-per-task=48
[...]
srun myprogram
```

and the program launched by `srun` would have received 48 cpu-cores because `srun` used to inherit the `--cpus-per-task=48` settings from `sbatch` or `salloc` settings, according to the quoted documentation since SLURM 22.05 this behavior is no longer true.

footnote: I tested with SLURM@22.05.09 and the old behavior was still true, but this is definitely the case with 23.x series. So the change might have happened in the later 22.05 series.

So if you leave things as is, now the program will receive just 1 cpu-core (unless the `srun` default has been modified).

You can easily test if your SLURM setup is affected, using `os.sched_getaffinity(0)`, as it shows which cpu-cores are eligible to be used by the current process. So it should be easy to count those with `len(os.sched_getaffinity(0))`.

Here is how you can test if you're affected:

```
$ cat test.slurm
#!/bin/bash
#SBATCH --job-name=test-cpu-cores-per-task
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=48  # adapt to your env if you have less than 48 cpu cores
#SBATCH --time=0:10:00
```

```
#SBATCH --partition=x      # adapt to your env to the right partition name
#SBATCH --output=%x-%j.out

srun python -c 'import os; print(f"visible cpu cores: {len(os.sched_getaffinity(0))}")'
```

If you get

```
visible cpu cores: 48
```

then you don't need to do anything, if however you get:

```
visible cpu cores: 1
```

or another value smaller than 48 then you're affected.

To fix that you need to change your SLURM script to either:

```
#SBATCH --cpus-per-task=48
[...]

srun --cpus-per-task=48 myprogram
```

or:

```
#SBATCH --cpus-per-task=48
[...]

SRUN_CPUS_PER_TASK=48
srun myprogram
```

or automate it with write-once-and-forget:

```
#SBATCH --cpus-per-task=48
[...]

SRUN_CPUS_PER_TASK=$SLURM_CPUS_PER_TASK
srun myprogram
```

To enable Hyper-Threads or not

As explained in the [Hyper-Threads](#) section you should be able to double the number of available cpu-cores if your CPUs support hyper-threading and for some workloads this may lead to an overall faster performance.

However, you should test the performance w/ and w/o HT, compare the results and choose the setting that gives the best

outcome.

case study: on AWS p4 nodes I discovered that enabling HT made the network throughput 4x slower. Since then we were careful to have HT disabled on that particular setup.

Single and Multi-node Launchers with SLURM

The following are complete SLURM scripts that demonstrate how to integrate various launchers with software that uses `torch.distributed` (but should be easily adaptable to other distributed environments).

- `torchrun` - to be used with [PyTorch distributed](#).
- `accelerate` - to be used with [HF Accelerate](#).
- `lightning` - to be used with [Lightning](#) (“PyTorch Lightning” and “Lightning Fabric”).
- `srun` - to be used with the native SLURM launcher - here we have to manually preset env vars that `torch.distributed` expects.

All of these scripts use [torch-distributed-gpu-test.py](#) as the demo script, which you can copy here with just:

```
cp ../../debug/torch-distributed-gpu-test.py .
```

assuming you cloned this repo. But you can replace it with anything else you need.

Training

Subsections:

- [Model parallelism](#)
- [Performance](#)
- [Fault Tolerance](#)
- [Reproducibility](#)
- [Instabilities](#)
- [Checkpoints](#)
- [Training hyper-parameters and model initializations](#)
- [Tensor precision / Data types](#)
- [Emulate a multi-node setup using just a single node](#) - instructions on how to emulate a multi-node setup using just a single node - we use the `deepspeed` launcher here.
- [Re-train HF hub models from scratch using finetuning examples](#)

Tools:

- [printflock.py](#) - a tiny library that makes your `print` calls non-interleaved in a multi-gpu environment.
- [multi-gpu-non-interleaved-print.py](#) - a `flock`-based wrapper around `print` that prevents messages from getting interleaved when multiple processes print at the same time - which is the case with `torch.distributed` used with multiple-gpus.

Model Parallelism

Parallelism overview

In the modern machine learning the various approaches to parallelism are used to:

1. Overcome GPU memory limitations. Examples:
 - fit very large models - e.g., t5-11b is 45GB in just model params
 - fit very long sequences - e.g.,
2. significantly speed up training - finish training that would take a year in hours

We will first discuss in depth various 1D parallelism techniques and their pros and cons and then look at how they can be combined into 2D and 3D parallelism to enable an even faster training and to support even bigger models. Various other powerful alternative approaches will be presented.

While the main concepts most likely will apply to any other framework, this article is focused on PyTorch-based implementations.

Two main approaches are used to enable training and inferring models that are bigger than the accelerator's memory:

1. 3D parallelism - very network efficient, but can be very invasive into the modeling code and require a lot more work to make it work correctly
2. ZeRO parallelism - not very network efficient, but requires close to zero changes to the modeling code and very easy to make to work.

Scalability concepts

The following is the brief description of the main concepts that will be described later in depth in this document.

1. [Data Parallelism](#) (DP) - the same setup is replicated multiple times, and each being fed a slice of the data. The processing is done in parallel and all setups are synchronized at the end of each training step.
2. [TensorParallelism](#) (TP) - each tensor is split up into multiple chunks, so instead of having the whole tensor reside on a single gpu, each shard of the tensor resides on its designated gpu. During processing each shard gets processed separately and in parallel on different GPUs and the results are synced at the end of the step. This is what one may call horizontal parallelism, as the splitting happens on horizontal level.
3. [PipelineParallelism](#) (PP) - the model is split up vertically (layer-level) across multiple GPUs, so that only one or several layers of the model are places on a single gpu. Each gpu processes in parallel different stages of the pipeline and working on a small chunk of the batch.
4. [Zero Redundancy Optimizer](#) (ZeRO) - Also performs sharding of the tensors somewhat similar to TP, except the whole tensor gets reconstructed in time for a forward or backward computation, therefore the model doesn't need to be modified. It also supports various offloading techniques to compensate for limited GPU memory. Sharded DDP is another name for the foundational ZeRO concept as used by various other implementations of ZeRO.
5. [Sequence Parallelism](#) - training on long input sequences requires huge amounts of GPU memory. This technique splits the processing of a single sequence across multiple GPUs.
6. [Expert Parallelism](#) - Mixture-Of-Experts (MoE) can be partitioned so that each expert has a dedicated GPU (or several of them).

The introduction sections of this paper is probably one of the best explanations I have found on most common parallelism techniques [Breadth-First Pipeline Parallelism](#).

Data Parallelism

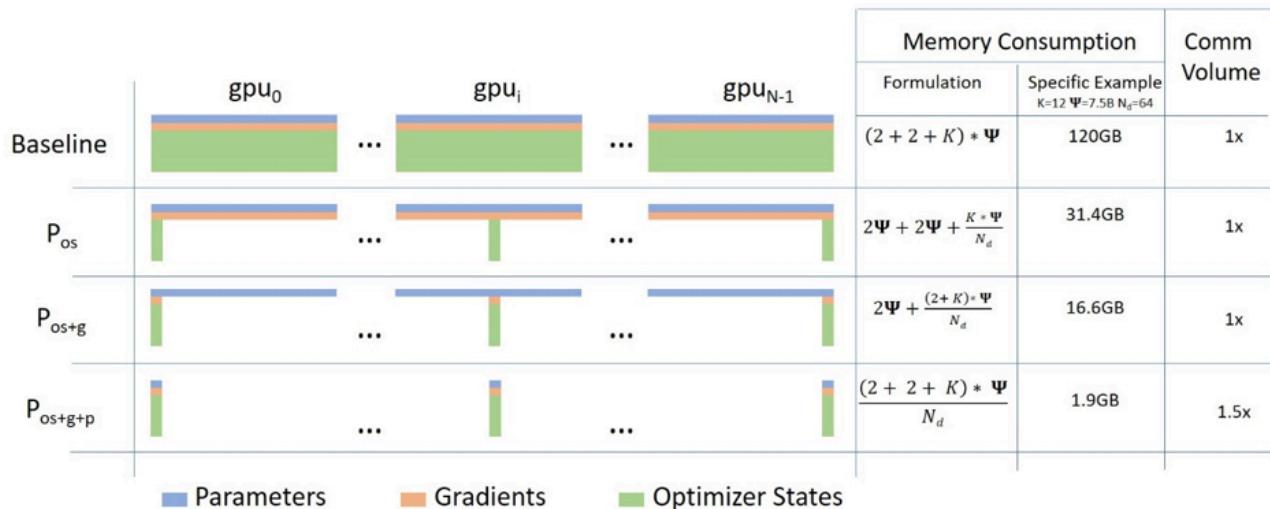
DDP

Most users with just 2 GPUs already enjoy the increased training speed up thanks to `DataParallel` (DP) and `DistributedDataParallel` (DDP) that are almost trivial to use. This is a built-in feature of Pytorch.

For details see [DistributedDataParallel](#)

ZeRO Data Parallelism

ZeRO-powered data parallelism (ZeRO-DP) is described on the following diagram from this [blog post](#)



It can be difficult to wrap one's head around it, but in reality the concept is quite simple. This is just the usual `DataParallel` (DP), except, instead of replicating the full model params, gradients and optimizer states, each GPU stores only a slice of it. And then at run-time when the full layer params are needed just for the given layer, all GPUs synchronize to give each other parts that they miss - this is it.

Consider this simple model with 3 layers, where each layer has 3 params:

```
La | Lb | Lc
---|---|---
a0 | b0 | c0
a1 | b1 | c1
a2 | b2 | c2
```

Layer La has weights a0, a1 and a2.

If we have 3 GPUs, the Sharded DDP (= Zero-DP) splits the model onto 3 GPUs like so:

```
GPU0:
La | Lb | Lc
---|---|---
a0 | b0 | c0
```

```
GPU1:
```

```
La | Lb | Lc
---|---|---
a1 | b1 | c1
```

```
GPU2:
La | Lb | Lc
---|---|---
a2 | b2 | c2
```

In a way this is the same horizontal slicing, as tensor parallelism, if you imagine the typical DNN diagram. Vertical slicing is where one puts whole layer-groups on different GPUs. But it's just the starting point.

Now each of these GPUs will get the usual mini-batch as it works in DP:

```
x0 => GPU0
x1 => GPU1
x2 => GPU2
```

The inputs are unmodified - they think they are going to be processed by the normal model.

First, the inputs hit the layer La.

Let's focus just on GPU0: x0 needs a0, a1, a2 params to do its forward path, but GPU0 has only a0 - it gets sent a1 from GPU1 and a2 from GPU2, bringing all pieces of the model together.

In parallel, GPU1 gets mini-batch x1 and it only has a1, but needs a0 and a2 params, so it gets those from GPU0 and GPU2.

Same happens to GPU2 that gets input x2. It gets a0 and a1 from GPU0 and GPU1, and with its a2 it reconstructs the full tensor.

All 3 GPUs get the full tensors reconstructed and a forward happens.

As soon as the calculation is done, the data that is no longer needed gets dropped - it's only used during the calculation. The reconstruction is done efficiently via a pre-fetch.

And the whole process is repeated for layer Lb, then Lc forward-wise, and then backward Lc -> Lb -> La.

To me this sounds like an efficient group backpacking weight distribution strategy:

1. person A carries the tent
2. person B carries the stove
3. person C carries the axe

Now each night they all share what they have with others and get from others what they don't have, and in the morning they pack up their allocated type of gear and continue on their way. This is Sharded DDP / Zero DP.

Compare this strategy to the simple one where each person has to carry their own tent, stove and axe, which would be far more inefficient. This is DataParallel (DP and DDP) in Pytorch.

While reading the literature on this topic you may encounter the following synonyms: Sharded, Partitioned.

If you pay close attention the way ZeRO partitions the model's weights - it looks very similar to tensor parallelism which will be discussed later. This is because it partitions/shards each layer's weights, unlike vertical model parallelism which is discussed next.

Implementations of ZeRO-DP stages 1+2+3:

- [DeepSpeed](#)

- [PyTorch](#) (originally it was implemented in [FairScale](#) and later it was upstreamed into the PyTorch core)
- [torchtitan](#)

Deepspeed ZeRO Integration:

- [HF Trainer integration](#)
- [Accelerate](#)
- [PyTorch Lightning](#)
- [Determined.AI](#)

FSDP Integration:

- [HF Trainer integration](#)
- [Accelerate](#)
- [PyTorch Lightning](#)
- [torchtitan](#)

Important papers:

Deepspeed and ZeRO in general:

- [ZeRO: Memory Optimizations Toward Training Trillion Parameter Models](#)
- [ZeRO-Offload: Democratizing Billion-Scale Model Training](#)
- [ZeRO-Infinity: Breaking the GPU Memory Wall for Extreme Scale Deep Learning](#)
- [ZeRO++: Extremely Efficient Collective Communication for Giant Model Training](#)
- [DeepSpeed Ulysses: System Optimizations for Enabling Training of Extreme Long Sequence Transformer Models](#)
- [AMSP: Reducing Communication Overhead of ZeRO for Efficient LLM Training](#)

PyTorch:

- [PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel](#)

Main DeepSpeed ZeRO Resources:

- [Project's github](#)
- [Usage docs](#)
- [API docs](#)
- [Blog posts](#)

Overcoming the huge global batch size issue

If you use, say, 1024 accelerators, you'll have tiny shards per accelerator and a ton of free memory for micro-batch-size (MBS), let's say you can fit MBS=32 - you end up with GBS=32k - most likely not what you want.

So you either need to deploy [Tensor Parallelism](#) which is non-trivial to implement, or often it's much simpler to deploy [Sequence Parallelism](#). I'm yet to try it in action, but so far what I gathered is for:

- Deepspeed ZeRO use [Deepspeed-Ulysses](#)
- FSDP use [Paged Ring Attention \(paper\)](#)

Please note that most likely it won't be as efficient as [Tensor Parallelism](#) - but I don't yet know of the actual additional overhead.

ZeRO with multiple replicas

By default ZeRO uses all GPUs to create a single model replica - that's the model is spread out across all gpus. Which leads to various limitations such as:

1. the global batch size is inflexible - it's always a function of `total_gpus*micro_batch_size` - which on large clusters could lead to a huge global batch size which might be detrimental for efficient convergence. Granted one could use a tiny micro batch size to keep the global batch size in check, but this leads to smaller matrices on each GPU which

results in less efficient compute

2. the much faster intra-node networking is not being benefited from since the slower inter-node network defines the overall speed of communications.

[ZeRO++](#) solves the 2nd limitation by introducing Hierarchical Weight Partition for ZeRO (hpZ). In this approach instead of spreading whole model weights across all the gpus, each model replica is restricted to a single node. This increases the memory usage by the total number of nodes, but now the 2x `all_gather` calls to gather the sharded components are performed over a much faster intra-node connection. Only the `reduce_scatter` to aggregate and redistribute gradients is performed over the slower inter-node network.

The first limitation doesn't exactly get fixed since the overall global batch size remains the same, but since each replica is more efficient and because the additional memory pressure is likely to limit the possible micro batch size on each gpu, this overall should improve the throughput of the system.

PyTorch FSDP has this feature implemented in [shardingStrategy.HYBRID_SHARD](#)

Papers:

- [ZeRO++: Extremely Efficient Collective Communication for Giant Model Training](#)
- [PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel](#)

ZeRO variations

Published papers that propose modifications to the ZeRO protocol:

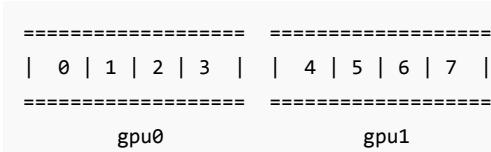
- [MiCS: Near-linear Scaling for Training Gigantic Model on Public Cloud](#) (2022)
- [AMSP: Super-Scaling LLM Training via Advanced Model States Partitioning](#) (2023)

Pipeline Parallelism methods

Naive Model Parallelism (Vertical)

Naive Model Parallelism (MP) is where one spreads groups of model layers across multiple GPUs. The mechanism is relatively simple - switch the desired layers `.to()` the desired devices and now whenever the data goes in and out those layers switch the data to the same device as the layer and leave the rest unmodified.

We refer to it as Vertical MP, because if you remember how most models are drawn, we slice the layers vertically. For example, if the following diagram shows an 8-layer model:



we just sliced it in 2 vertically, placing layers 0-3 onto GPU0 and 4-7 to GPU1.

Now while data travels from layer 0 to 1, 1 to 2 and 2 to 3 this is just the normal model. But when data needs to pass from layer 3 to layer 4 it needs to travel from GPU0 to GPU1 which introduces a communication overhead. If the participating GPUs are on the same compute node (e.g. same physical machine) this copying is pretty fast, but if the GPUs are located on different compute nodes (e.g. multiple machines) the communication overhead could be significantly larger.

Then layers 4 to 5 to 6 to 7 are as a normal model would have and when the 7th layer completes we often need to send the data back to layer 0 where the labels are (or alternatively send the labels to the last layer). Now the loss can be computed and the optimizer can do its work.

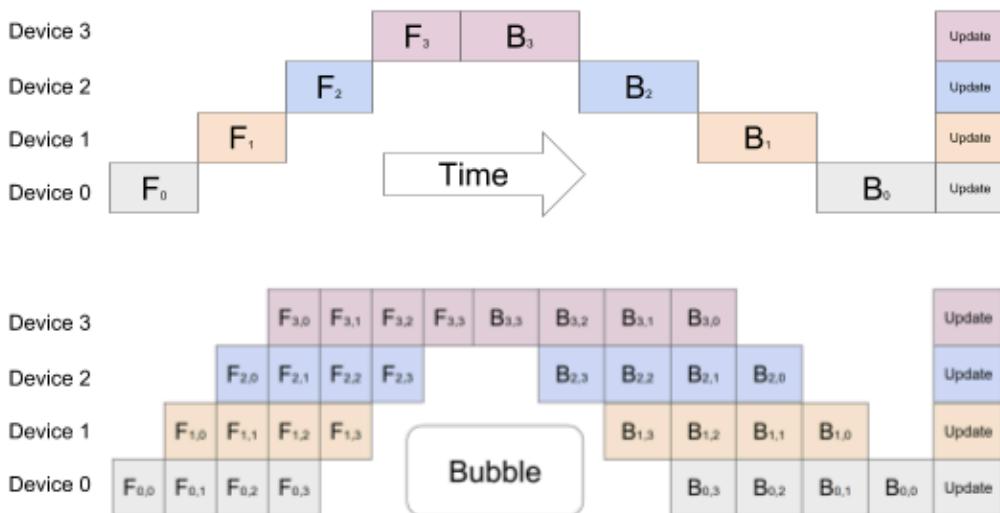
Problems:

- the main deficiency and why this one is called "naive" MP, is that all but one GPU is idle at any given moment. So if 4 GPUs are used, it's almost identical to quadrupling the amount of memory of a single GPU, and ignoring the rest of the hardware. Plus there is the overhead of copying the data between devices. So 4x 6GB cards will be able to accommodate the same size as 1x 24GB card using naive MP, except the latter will complete the training faster, since it doesn't have the data copying overhead. But, say, if you have 40GB cards and need to fit a 45GB model you can with 4x 40GB cards (but barely because of the gradient and optimizer states)
- shared embeddings may need to get copied back and forth between GPUs.

Pipeline Parallelism

Pipeline Parallelism (PP) is almost identical to a naive MP, but it solves the GPU idling problem, by chunking the incoming batch into micro-batches and artificially creating a pipeline, which allows different GPUs to concurrently participate in the computation process.

The following illustration from the [GPipe paper](#) shows the naive MP on the top, and PP on the bottom:



Top: The naive model parallelism strategy leads to severe underutilization due to the sequential nature of the network. Only one accelerator is active at a time. Bottom: GPipe divides the input mini-batch into smaller micro-batches, enabling different accelerators to work on separate micro-batches at the same time.

It's easy to see from the bottom diagram how PP has less dead zones, where GPUs are idle. The idle parts are referred to as the "bubble".

Both parts of the diagram show a parallelism that is of degree 4. That is 4 GPUs are participating in the pipeline. So there is the forward path of 4 pipe stages F_0 , F_1 , F_2 and F_3 and then the return reverse order backward path of B_3 , B_2 , B_1 and B_0 .

PP introduces a new hyper-parameter to tune and it's `chunks` which defines how many chunks of data are sent in a sequence through the same pipe stage. For example, in the bottom diagram you can see that `chunks=4`. GPU0 performs the same forward path on chunk 0, 1, 2 and 3 ($F_{0,0}$, $F_{0,1}$, $F_{0,2}$, $F_{0,3}$) and then it waits for other GPUs to do their work and only when their work is starting to be complete, GPU0 starts to work again doing the backward path for chunks 3, 2, 1 and 0 ($B_{0,3}$, $B_{0,2}$, $B_{0,1}$, $B_{0,0}$).

Note that conceptually this is the same concept as gradient accumulation steps (GAS). Pytorch uses `chunks`, whereas

DeepSpeed refers to the same hyper-parameter as GAS.

Because of the chunks, PP introduces the concept of micro-batches (MBS). DP splits the global data batch size into mini-batches, so if you have a DP degree of 4, a global batch size of 1024 gets split up into 4 mini-batches of 256 each (1024/4). And if the number of chunks (or GAS) is 32 we end up with a micro-batch size of 8 (256/32). Each Pipeline stage works with a single micro-batch at a time.

To calculate the global batch size of the DP + PP setup we then do: `mbs*chunks*dp_degree (8*32*4=1024)`.

Let's go back to the diagram.

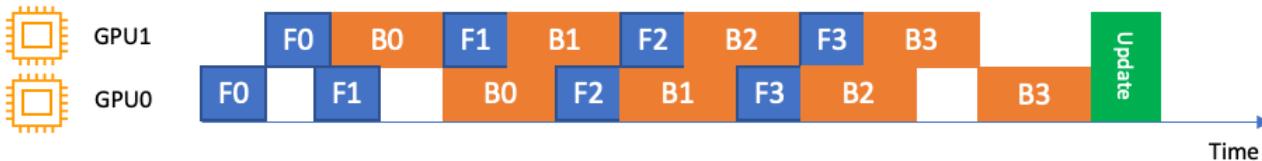
With `chunks=1` you end up with the naive MP, which is very inefficient. With a very large `chunks` value you end up with tiny micro-batch sizes which could be not every efficient either. So one has to experiment to find the value that leads to the highest efficient utilization of the gpus.

While the diagram shows that there is a bubble of "dead" time that can't be parallelized because the last `forward` stage has to wait for `backward` to complete the pipeline, the purpose of finding the best value for `chunks` is to enable a high concurrent GPU utilization across all participating GPUs which translates to minimizing the size of the bubble.

The choice of the schedule is critical to the efficient performance, with the most common schedules being in the order of invention:

- sequential [Gpipe: Efficient training of giant neural networks using pipeline parallelism](#)
- interleaved 1F1B [Pipedream: Fast and efficient pipeline parallel dnn training](#)
- looped, depth-first [Efficient large-scale language model training on gpu clusters using Megatron-LM](#)
- breadth-first [Breadth-First Pipeline Parallelism](#)
- Llama 3 training used a combination of depth and breadth first for best performance and also allowed them to progressively modify the global batch size as the training progressed, which is typically very difficult to accomplish with PP. See [<https://arxiv.org/abs/2407.21783>](The Llama 3 Herd of Models) section 3.3.2 Parallelism for Model Scaling.

Here is for example an interleaved pipeline:



Here the bubble (idle time) is further minimized by prioritizing backward passes.

It's used by DeepSpeed, Varuna and SageMaker to name a few.

Varuna further tries to improve the schedule by using simulations to discover the most efficient scheduling.

[DeepSeek v3](#) introduced an even more efficient PP via DualPipe that reduces the bubble size and succeeds at a better compute/comms overlap. See section 3.2.1 of the paper for the specific details.



([source](#))

There are 2 groups of PP solutions - the traditional Pipeline API and the more modern solutions that make things much

easier for the end user by helping to partially or fully automate the process:

1. Traditional Pipeline API solutions:

- Megatron-LM
- DeepSpeed
- PyTorch

2. Modern solutions:

- PiPPy
- Varuna
- Sagemaker
- DeepSeek

Problems with traditional Pipeline API solutions:

- have to modify the model quite heavily, because Pipeline requires one to rewrite the normal flow of modules into a `nn.Sequential` sequence of the same, which may require changes to the design of the model.
- currently the Pipeline API is very restricted. If you had a bunch of python variables being passed in the very first stage of the Pipeline, you will have to find a way around it. Currently, the pipeline interface requires either a single Tensor or a tuple of Tensors as the only input and output. These tensors must have a batch size as the very first dimension, since pipeline is going to chunk the mini batch into micro-batches. Possible improvements are being discussed here <https://github.com/pytorch/pytorch/pull/50693>
- conditional control flow at the level of pipe stages is not possible - e.g., Encoder-Decoder models like T5 require special workarounds to handle a conditional encoder stage.
- have to arrange each layer so that the output of one model becomes an input to the other model.
- The first stage contains a heavy embedding which can be quite huge if the vocabulary is large - and this may require a custom splicing so that the first stage will contain less transformer blocks than other stages.

I'm yet to try to experiment with Varuna and SageMaker but their papers report that they have overcome the list of problems mentioned above and that they require much smaller changes to the user's model.

Implementations:

- [Pytorch](#) (initial support in pytorch-1.8, and progressively getting improved in 1.9 and more so in 1.10). Some [examples](#)
- [FairScale](#)
- [DeepSpeed](#)
- [Megatron-LM](#) has an internal implementation - no API.
- [Varuna](#)
- [SageMaker](#) - this is a proprietary solution that can only be used on AWS.
- [OSLO](#) - this is implemented based on the Hugging Face Transformers.
- [PiPPy: Pipeline Parallelism for PyTorch](#) - automatic PP via `torch.fx`
- [nanotron](#)
- [torchtitan](#)

Related reading

- [Pipeline-Parallelism: Distributed Training via Model Partitioning](#)

Tensor Parallelism

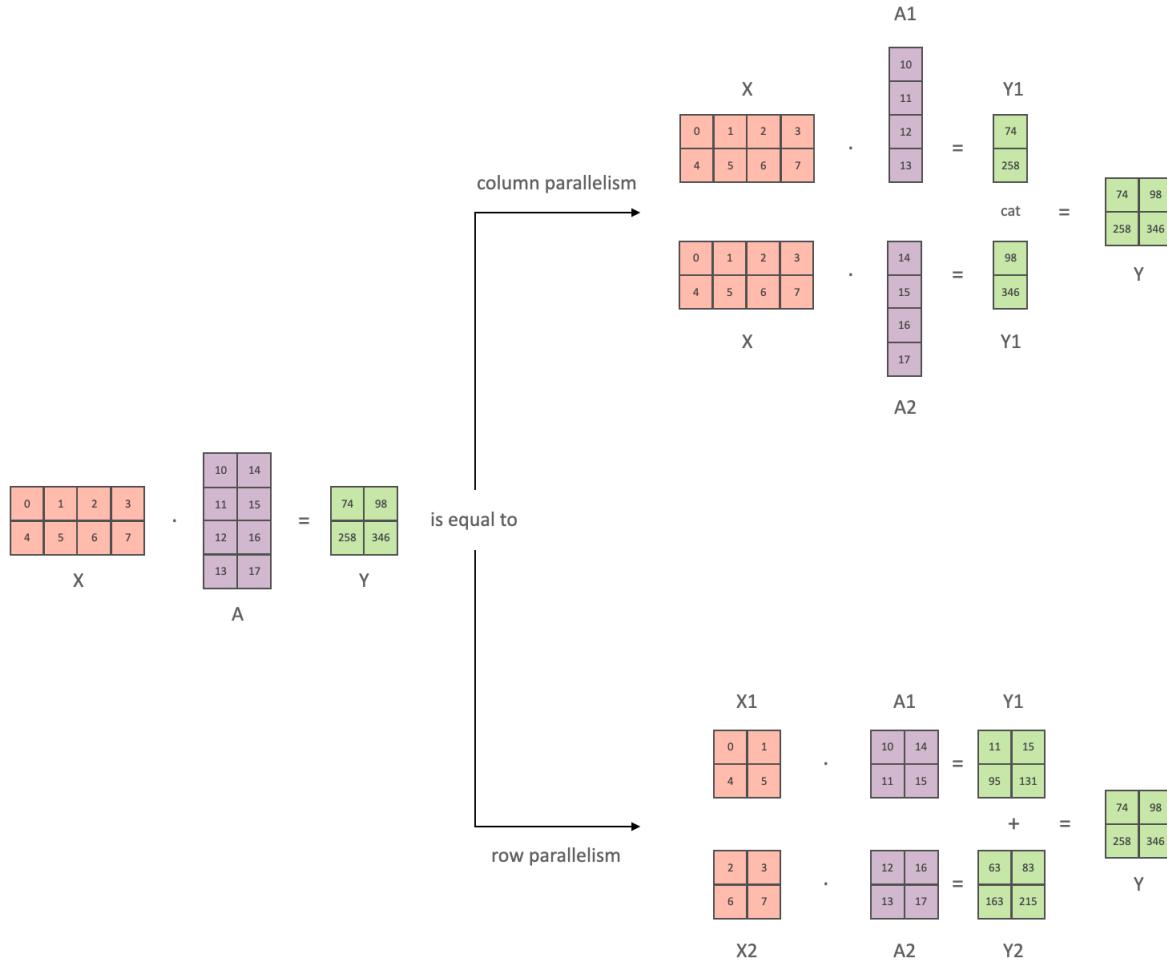
In Tensor Parallelism each GPU processes only a slice of a tensor and only aggregates the full tensor for operations that require the whole thing.

In this section we use concepts and diagrams from the [Megatron-LM](#) paper: [Efficient Large-Scale Language Model Training on GPU Clusters](#).

The main building block of any transformer is a fully connected `nn.Linear` followed by a nonlinear activation `GeLU`.

Following the Megatron's paper notation, we can write the dot-product part of it as $\mathbf{y} = \text{GeLU}(\mathbf{x}\mathbf{a})$, where \mathbf{x} and \mathbf{y} are the input and output vectors, and \mathbf{a} is the weight matrix.

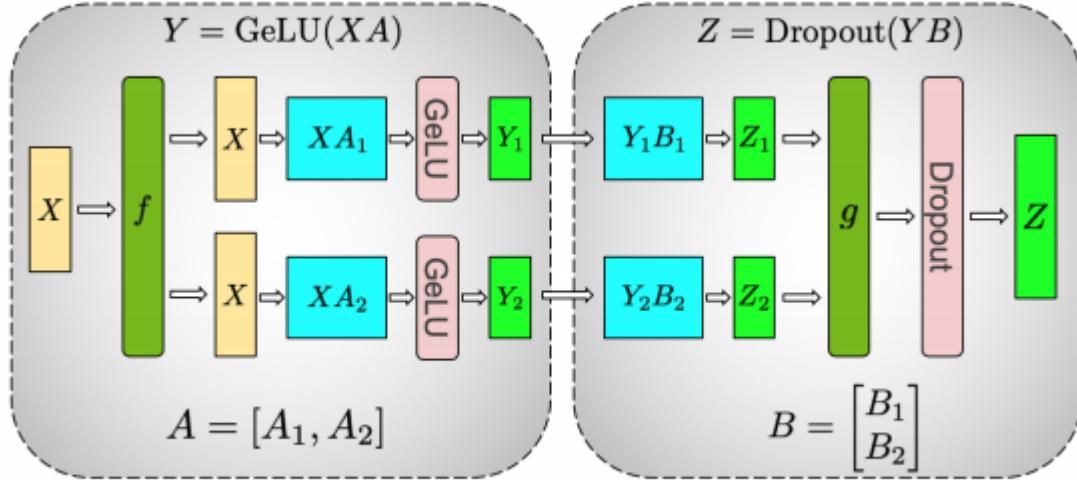
If we look at the computation in matrix form, it's easy to see how the matrix multiplication can be split between multiple GPUs:



If we split the weight matrix \mathbf{A} column-wise across N GPUs and perform matrix multiplications $\mathbf{x}\mathbf{a}_1$ through $\mathbf{x}\mathbf{a}_n$ in parallel, then we will end up with N output vectors $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ which can be fed into `GeLU` independently:

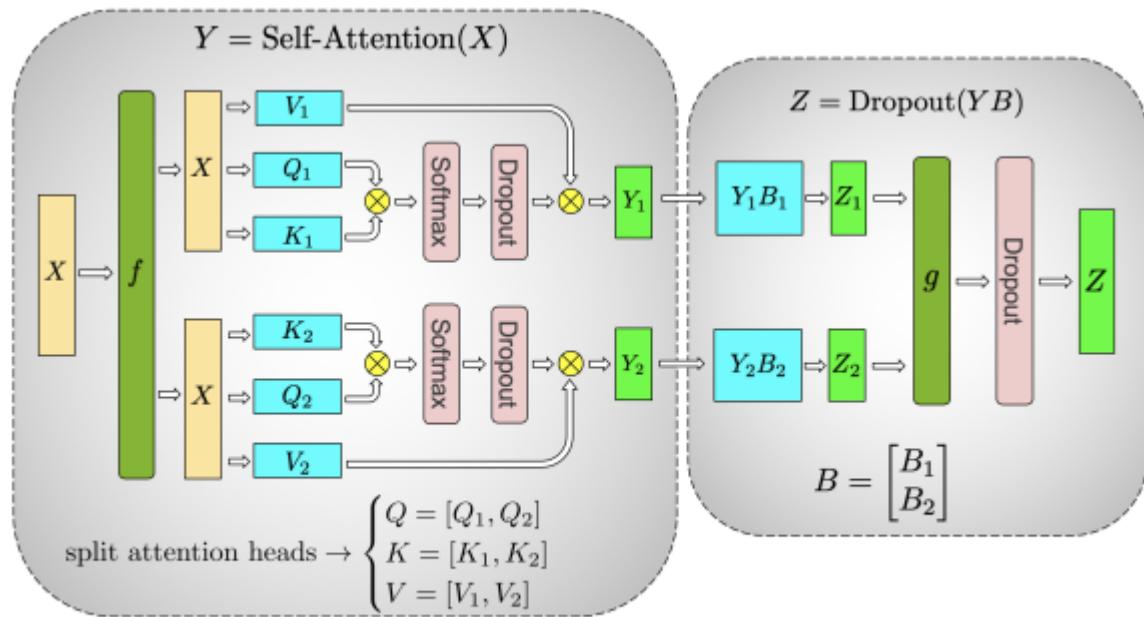
$$[\mathbf{Y}_1, \mathbf{Y}_2] = [\text{GeLU}(\mathbf{X}\mathbf{A}_1), \text{GeLU}(\mathbf{X}\mathbf{A}_2)]$$

Using this principle, we can update an MLP of arbitrary depth, without the need for any synchronization between GPUs until the very end, where we need to reconstruct the output vector from shards. The Megatron-LM paper authors provide a helpful illustration for that:



(a) MLP

Parallelizing the multi-headed attention layers is even simpler, since they are already inherently parallel, due to having multiple independent heads!



(b) Self-Attention

Important: TP requires very fast network, and therefore since typically intra-node networks are much faster than inter-node networks it's not advisable to do TP across nodes. Practically, if a node has 4 GPUs, the highest TP degree is therefore 4. If you need a TP degree of 8, you need to use nodes that have at least 8 GPUs.

TP can be combined with other parallelization methods.

Alternative names:

- DeepSpeed calls it [tensor slicing](#)

Implementations:

- [Megatron-LM](#) has an internal implementation, as it's very model-specific

- [PyTorch](#)
- [SageMaker](#) - this is a proprietary solution that can only be used on AWS.
- [OSLO](#) has the tensor parallelism implementation based on the Transformers.
- [nanotron](#)
- [parallelformers](#) (only inference at the moment)
- [torchtitan](#)

Async TP

One of the deficiencies of TP is that it's difficult to overlap its comms with compute. PyTorch is proposing to overcome this with [Async-TP](#) which decomposes the dependent sequence of all-gather + matmul into series of cudaMemcpyAsync calls and smaller partial matmuls - and it does it automatically for you using `torch.compile`!

- [Megatron-LM](#) has it implemented as well via `--tp-comm-overlap`.

Related reading

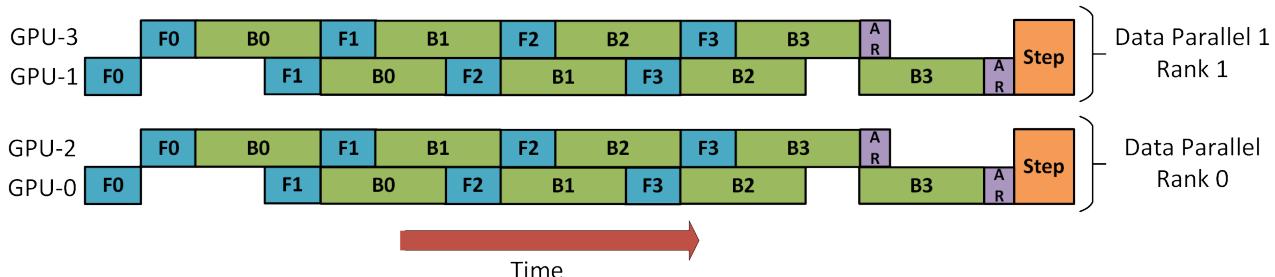
- [Tensor Parallelism and Sequence Parallelism: Detailed Analysis](#)

TP+SP

TP can be combined with SP in the same process group to minimize communication costs as explained in [Reducing Activation Recomputation in Large Transformer Models](#). For example in LLMs, TP is used for embedding, attention and linear layers and when dropout and layer norm are reached SP is used instead.

DP+PP

The following diagram from the DeepSpeed [pipeline tutorial](#) demonstrates how one combines DP with PP.



Here it's important to see how DP rank 0 doesn't see GPU2 and DP rank 1 doesn't see GPU3. To DP there is just GPUs 0 and 1 where it feeds data as if there were just 2 GPUs. GPU0 "secretly" offloads some of its load to GPU2 using PP. And GPU1 does the same by enlisting GPU3 to its aid.

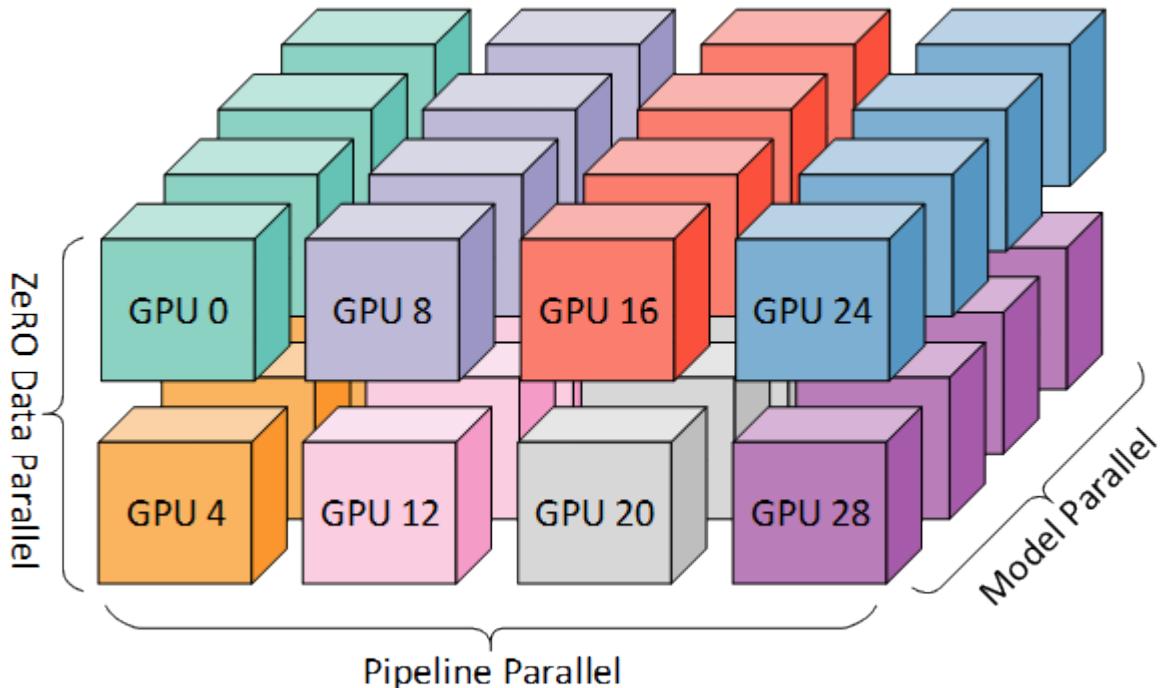
Since each dimension requires at least 2 GPUs, here you'd need at least 4 GPUs.

Implementations:

- [DeepSpeed](#)
- [Megatron-LM](#)
- [Varuna](#)
- [SageMaker](#)
- [OSLO](#)
- [nanotron](#)
- [torchtitan](#)

DP+PP+TP

To get an even more efficient training a 3D parallelism is used where PP is combined with TP and DP. This can be seen in the following diagram.



This diagram is from a blog post [3D parallelism: Scaling to trillion-parameter models](#), which is a good read as well.

Since each dimension requires at least 2 GPUs, here you'd need at least 8 GPUs.

Implementations:

- [DeepSpeed](#) - DeepSpeed also includes an even more efficient DP, which they call ZeRO-DP.
- [Megatron-LM](#)
- [Varuna](#)
- [SageMaker](#)
- [QSLQ](#)
- [nanotron](#)
- [torchtitan](#)

ZeRO DP+PP+TP

One of the main features of DeepSpeed is ZeRO, which is a super-scalable extension of DP. It has already been discussed in [ZeRO Data Parallelism](#). Normally it's a standalone feature that doesn't require PP or TP. But it can be combined with PP and TP.

When ZeRO-DP is combined with PP (and optionally TP) it typically enables only ZeRO stage 1 (optimizer sharding).

While it's theoretically possible to use ZeRO stage 2 (gradient sharding) with Pipeline Parallelism, it will have bad performance impacts. There would need to be an additional reduce-scatter collective for every micro-batch to aggregate the gradients before sharding, which adds a potentially significant communication overhead. By nature of Pipeline Parallelism, small micro-batches are used and instead the focus is on trying to balance arithmetic intensity (micro-batch size) with minimizing the Pipeline bubble (number of micro-batches). Therefore those communication costs are going to hurt.

In addition, there are already fewer layers than normal due to PP and so the memory savings won't be huge. PP already reduces gradient size by $1/PP$, and so gradient sharding savings on top of that are less significant than pure DP.

ZeRO stage 3 is not a good choice either for the same reason - more inter-node communications required.

And since we have ZeRO, the other benefit is ZeRO-Offload. Since this is stage 1 optimizer states can be offloaded to CPU.

Implementations:

- [Megatron-DeepSpeed](#) and [Megatron-Deepspeed from BigScience](#), which is the fork of the former repo.
- [OSLO](#)
- [torchtitan](#)

Important papers:

- [Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model](#)

Sequence Parallelism

ML tasks, such as DNA sequencing, may require training with very long sequence lengths (e.g. 256K), and even normal LLMs could be trained on sequences of 10k and longer.

Self-Attention, which is the key component of Transformers, suffers from quadratic memory requirements with respect to the sequence length, therefore when sequence length gets to a certain length, even a batch size of 1 might not be able to fit onto a single GPU and require additional partitioning along the sequence dimension. And once this is done, the sequence can be of any length.

As this type of parallelism is orthogonal to the other parallelization types described in this document, it can be combined with any of them, leading to 4D, ZeRO-DP+SP and other combinations.

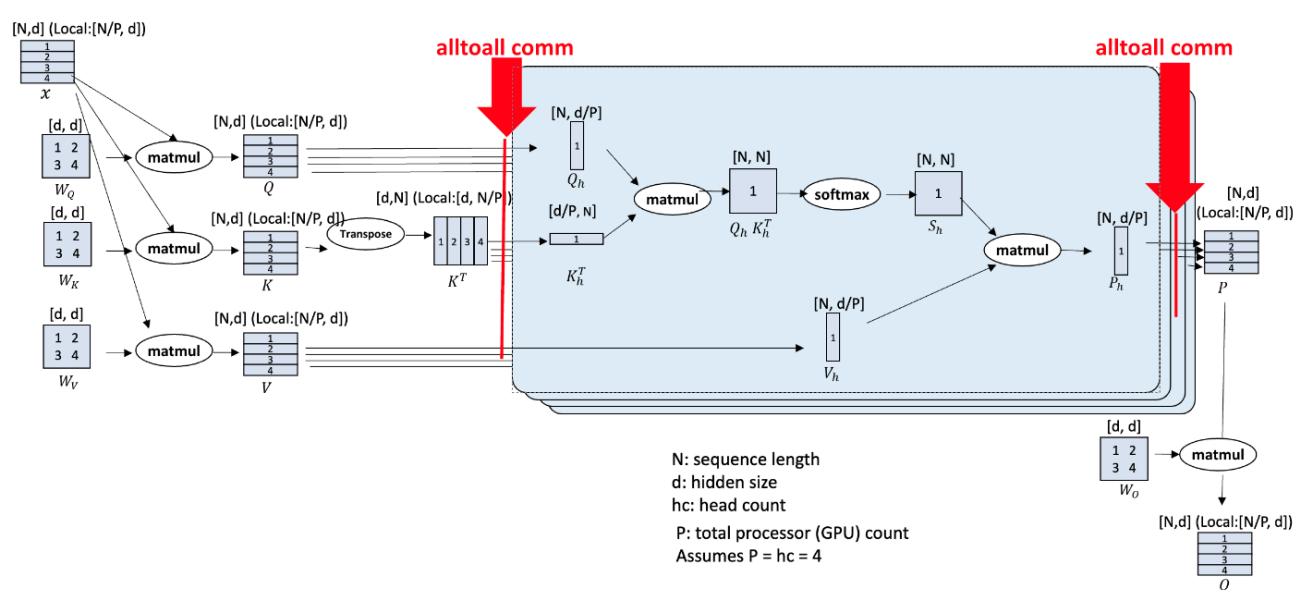
Deepspeed-Ulysses SP

Paper: [DeepSpeed Ulysses: System Optimizations for Enabling Training of Extreme Long Sequence Transformer Models](#)

In this implementation 2 elements are sharded:

1. The multiple-head attention weights are split across the participating GPUs so that each GPU has a few sub-heads only. This is done when the model is created/loaded. This is somewhat similar to [Tensor Parallelism](#).
2. During training each input sequence is partitioned into chunks and each chunk is sent to one of the GPUs, which reminds us of ZeRO-3 sharding, except instead of weights the inputs are sharded.

During compute each sequence chunk is projected onto QKV and then gathered to the full sequence QKV on each device, computed on each device only for the subheads it owns and then gathered again into the full attention output for the MLP block.



$$\text{Output context} = \text{Softmax} \left(\frac{QK^T}{\sqrt{d}} \right) V$$

5. At the end another all-to-all collective transforms output context tensor of attention computation to sequence (N/P) parallel for subsequent operators (MLP MatMul, layer norm, etc.) in the remaining modules of transformer layer block.

Example: Let's consider seqlen=8K, num_heads=128 and a single node of num_gpus=8

1. each GPU gets a 1K-long chunk of the original sequence ($8K/8$)
2. each GPU gets assigned 16 sub-heads ($128/8$)
3. a. on gpu0 before forward the original sequence is gathered back into 8K tokens b. the attention computation is done on the first 16 sub-heads the same logic is performed on the remaining 7 GPUs, each computing 8k attention over its 16 sub-heads

You can read the specifics of the very efficient comms [here](#).

DeepSpeed-Ulysses keeps communication volume consistent by increasing GPUs proportional to message size or sequence length.

Colossal-AI's SP

Paper: [Sequence parallelism: Long sequence training from system perspective](#)

Colossal-AI's SP implementation uses ring self-attention, a ring-like communication collective in which query projections

are local whereas key and values projections are transmitted in a ring-style to compute global attention, resulting in communication complexity linear in message size, M .

Megatron-LM's SP

Paper: [Reducing Activation Recomputation in Large Transformer Models](#)

Megatron-LM's SP is tightly integrated with its TP. Megatron-LM partitions sequence along sequence dimensions and applies allgather and reduce scatter collective to aggregate QKV projections for attention computation. Its communication volume increases linearly with message size (M) regardless of number of compute devices.

Ring Attention with Blockwise Transformers

Paper: [Ring Attention with Blockwise Transformers for Near-Infinite Context](#)

1. Tensors are sharded along the sequence dimension throughout: `(seq_len // N, d_model)`-shaped
2. In the attention layers, every GPU starts by computing the part of the attention scores they are able to w/ their available shards.
3. Simultaneously, the keys and values from other sequence chunks are communicated around.
4. Once the keys/values from another chunk are available, each GPU continues on with their attention computation using the key/value tensors from this new segment of the sequence
5. Continue until attention computation is complete.

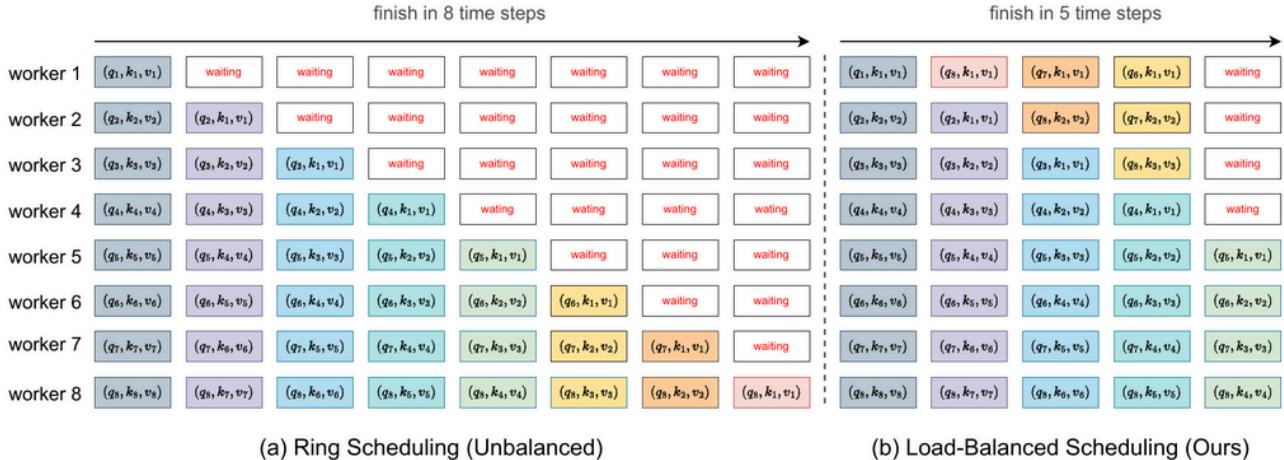
SP Implementations:

- [Megatron-LM](#)
- [DeepSpeed](#)
- [Colossal-AI](#)
- [torchtitan](#)

PyTorch is also working on this feature and calling it Context Parallel (CP).

DistFlashAttn

[DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training](#) is reported to be many times faster than Ring Self-Attention, because it load balances the KVQ per token computation between the workers while performing Sequence Parallelism.



Related reading

- [Tensor Parallelism and Sequence Parallelism: Detailed Analysis](#)

Expert Parallelism

When Mixture-Of-Experts (MoE) is used (in particular during inference) one could give each expert its own accelerator (or a few if one isn't enough). This adds another dimension for parallelization and can significantly speed things up for large batches that are likely to hit all of the experts.

For detailed explanations please see:

- [DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale](#)
- [Mixture of Experts Explained](#)

FlexFlow

[FlexFlow](#) also solves the parallelization problem in a slightly different approach.

Paper: ["Beyond Data and Model Parallelism for Deep Neural Networks" by Zhihao Jia, Matei Zaharia, Alex Aiken](#)

It performs a sort of 4D Parallelism over Sample-Operator-Attribute-Parameter.

1. Sample = Data Parallelism (sample-wise parallel)
2. Operator = Parallelize a single operation into several sub-operations
3. Attribute = Data Parallelism (length-wise parallel)
4. Parameter = Model Parallelism (regardless of dimension - horizontal or vertical)

Examples:

- Sample

Let's take 10 batches of sequence length 512. If we parallelize them by sample dimension into 2 devices, we get 10×512 which becomes $5 \times 2 \times 512$.

- Operator

If we perform layer normalization, we compute std first and mean second, and then we can normalize data. Operator parallelism allows computing std and mean in parallel. So if we parallelize them by operator dimension into 2 devices (cuda:0, cuda:1), first we copy input data into both devices, and cuda:0 computes std, cuda:1 computes mean at the same time.

- Attribute

We have 10 batches of 512 length. If we parallelize them by attribute dimension into 2 devices, 10×512 will be $10 \times 2 \times 256$.

- Parameter

It is similar with tensor model parallelism or naive layer-wise model parallelism.

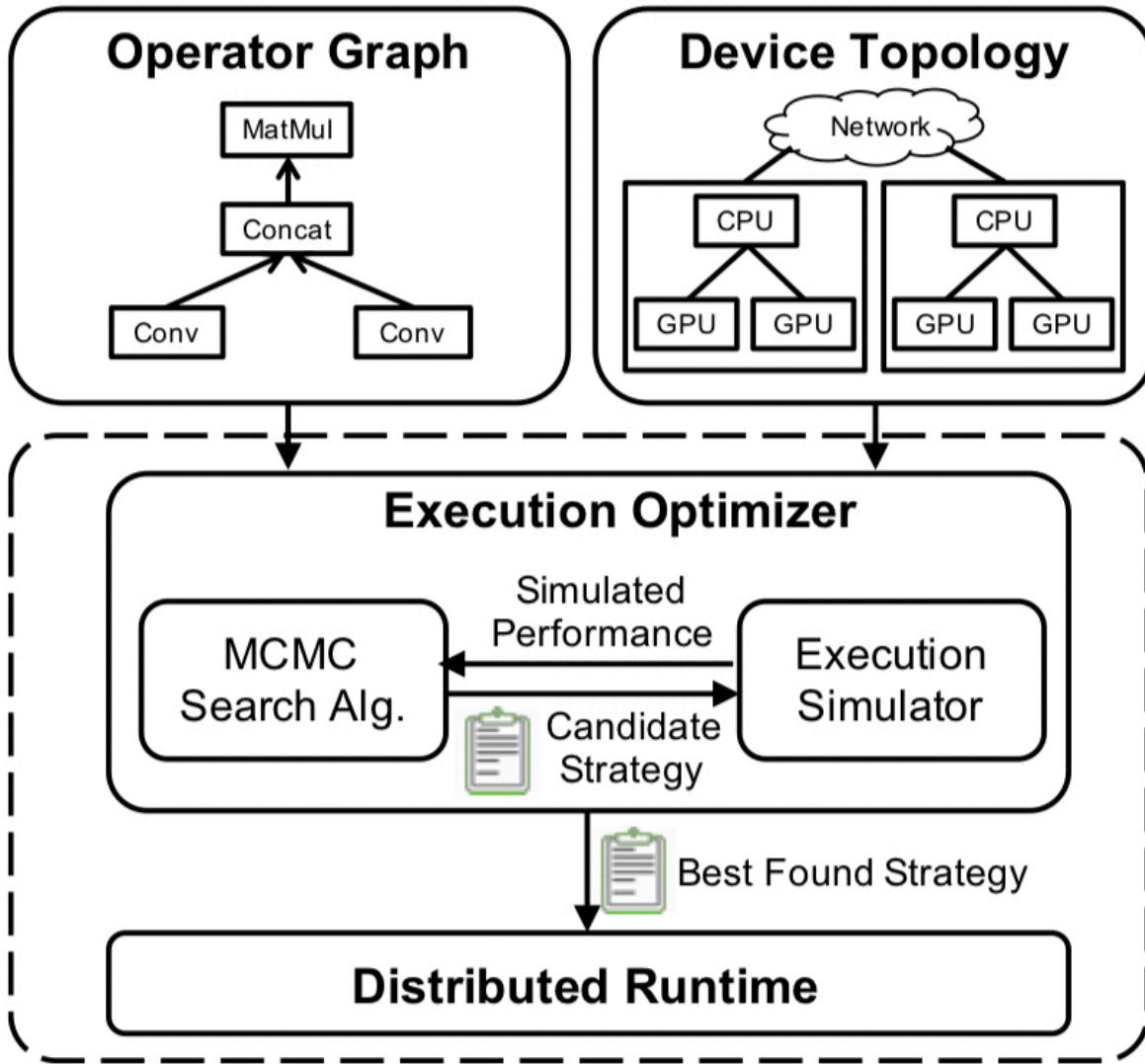


Figure 1. FlexFlow overview.

The significance of this framework is that it takes resources like (1) GPU/TPU/CPU vs. (2) RAM/DRAM vs. (3) fast-intra-connect/slow-inter-connect and it automatically optimizes all these algorithmically deciding which parallelisation to use where.

One very important aspect is that FlexFlow is designed for optimizing DNN parallelizations for models with static and fixed workloads, since models with dynamic behavior may prefer different parallelization strategies across iterations.

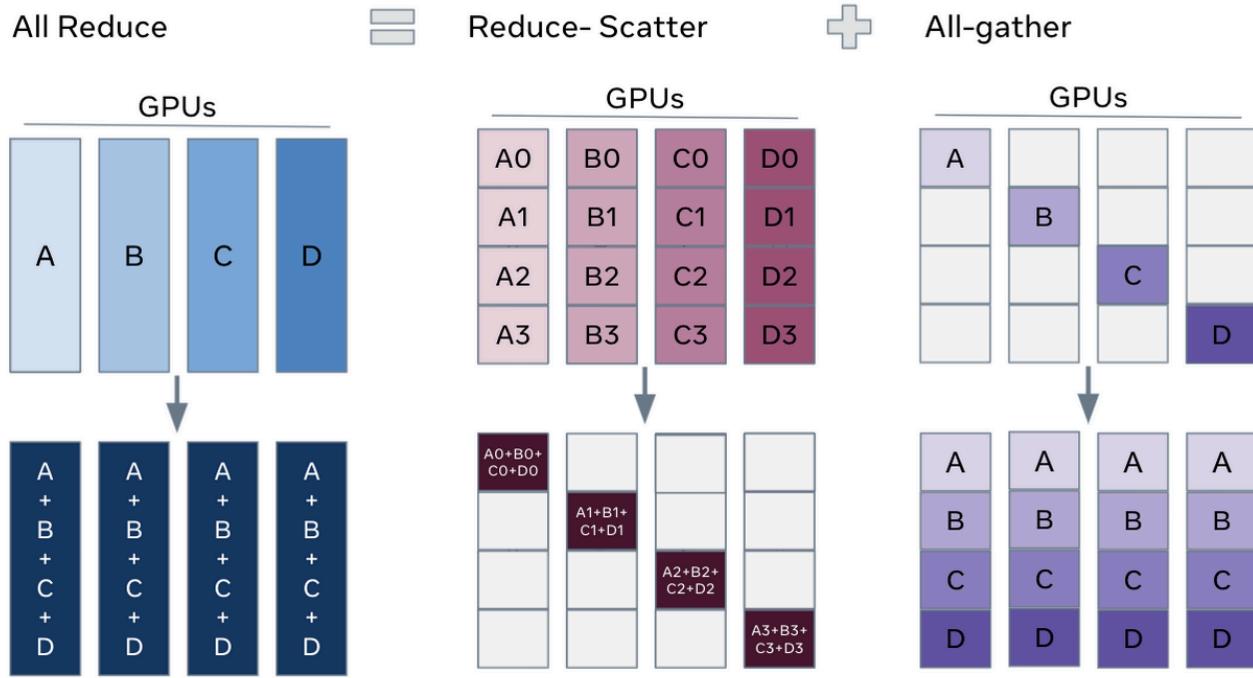
So the promise is very attractive - it runs a 30min simulation on the cluster of choice and it comes up with the best strategy to utilise this specific environment. If you add/remove/replace any parts it'll run and re-optimize the plan for that. And then you can train. A different setup will have its own custom optimization.

Parallelism network collectives

As intra- and inter-node speeds typically have a 10x difference, it's crucial to choose different parallelization techniques for intra- and inter-node crossing. e.g. TP must always remain within the node because of its massive synchronization requirements. Moreover, some accelerators, like the recent AMD MI3** series have a very slow gpu-to-gpu connectivity

which again impacts how parallelism is done for the best performance.

Here is a useful tidbit: the all-reduce collective can be decomposed into two separate phases: reduce-scatter and all-gather.



[\(source\)](#)

Here is the breakdown of which collectives are used for which parallelization strategies:

- DDP: 1x all-reduce for the gradients - ideally overlapping with compute - total volume: 2x model params comms
- ZeRO-DP ZeRO-1/ZeRO-2: 1x all-gather for optimizer states plus 1x reduce-scatter for gradients - total volume: 2x model params comms
- ZeRO-DP ZeRO-3: 2x all-gather for weights (before forward + before backward) plus 1x reduce-scatter for gradients - total volume: 3x model params comms (1.5x more than DDP and ZeRO-1/ZeRO-2)
- TP: 2x all-gather and 2x reduce-scatter
- PP: 2x send + 2x recv - overlapping with compute in the steady phase
- SP: depends on implementation: with hidden size h, sequence length of N, and parallelism degree of P
 - Megatron-LM: 2x all-gather and 2x reduce-scatter with volume $4*N*h$ per transformers layer [section 3.2 in paper](#)
 - DeepSpeed Ulysses: 2x all-to-all with volume $4*N*h/P$ per transformers layer [section 3.2 in paper](#)

It's possible that you will find different implementations that may use different communication patterns.

Inter-node speed requirements to use ZeRO

The ZeRO scalability protocol, be it DeepSpeed ZeRO or PyTorch FSDP, requires a lot more inter-node traffic than TP+PP+DP solutions, and sometimes it can't take advantage of the faster intra-node connectivity, and therefore if your inter-node network is slow your expensive GPUs might be massively bottlenecked by the comms.

The ZeRO protocol partially overlaps comms with compute, so ideally you want to get close to `comms_time <= compute_time`. The overlap is not perfect, so there will be always some network bottleneck, but we want to make sure that `comms_time` is not much larger than `compute_time`.

In ZeRO-3, we have `all_gather` on weights in `forward`, then `all_gather` on weights in `backward`, last is `reduce_scatter` on gradients in `backward`. In total there are 3 global collective calls each sending a model size multiplied by how many bytes

per parameter are used. e.g. a 10B param model in bf16 under ZeRO-3 will need to send $10*2*3 = 60\text{GB}$ of data.

In comparison [DistributedDataParallel](#) (DDP) uses a single `all_reduce` call, but which requires 2x data transmission, and so a 10B param model in bf16 under DDP will need to send $10*2*2 = 40\text{GB}$ of data.

ZeRO-1 which only shards the optimiser states, like DDP, will too need to transmit 40GB of data (one `all_gather` and one `reduce_scatter`.)

Here is how to calculate time in seconds for communication and compute:

- `comms_time = n_transmissions * n_bytes * model_size_in_B / inter-node-throughput_in_GBps`
- `compute_time = n_passes * n_bytes * model_size_in_B * seqlen * global_batch_size / (total_gpus * 1e3 * tflops_wo_comms)`

The compute time formula is a rough estimate which works for any Transformer-block based model. It ignores any small computations and includes only the massive `matmuls`.

As an experiment let's use the data points from [IDEFICS-80B](#) training.

When we trained IDEFICS-80B with a 340GBs EFA we were getting only 90TFLOPs w/ Deepspeed ZeRO-3 on A100s as compared to 150+TFLOPs one was getting with Megatron's TP+PP+DP. and moreover a big chunk of the model was frozen as were building a new models based on one language and one vision model. So our multiplier was less than 3. On the other hand we were using activation recomputation to save memory, so this is an additional transmission of all model weights and to top it all off since nccl wasn't supporting proper half-precision reduction we used fp32 for gradient reductions, so really our multiplier wasn't 3 but more like 4.5.

Values used for IDEFICS-80B training:

- `model_size_in_B = 80`
- `n_bytes = 2` in case of bf16 which is 2 bytes
- `n_transmissions = 3` in the case of ZeRO-3/FSDP (1x `reduce_scatter` + 2x `all_gather` (fwd + bwd)) and 2 in case of ZeRO-1 (1x `reduce_scatter` + 1x `all_gather`),
- additionally, in the case of IDEFICS-80B we decided to reduce grads in fp32 to minimize NCCL accumulation loss, so we actually had `n_transmissions*n_bytes=3*2+2=4*2` for the additional 2 bytes but since half the model was frozen only about half of gradients were sent, so we still have the multiplier of 3.
- `n_passes = 4` with activation recomputation, or 3 w/o it. The model has to do only 1x compute per `forward` and 2x per `backward` (since the grads are calculated twice - once wrt inputs and once wrt weights). And with activation recomputation one more `forward` is done.
- `total_gpus = 512`
- `global_batch_size = 3584`
- `seqlen = 1024`
- `inter-node-throughput_in_GBps = 42.5` (340Gbps) (AWS EFA v1) -`tflops_wo_comms` is the tflops w/o the communication overhead. Not theoretical peak as that is unachievable, but perhaps 75% in the case of A100@BF16 - so $312*0.75=234$ TFLOPS

We derived 340Gbps inter-node network throughput using [all_reduce_bench.py](#) which by default uses a payload of 4GB. In the case of IDEFICS-80B we had 80 layers, so approximately each layer was 1B params large. Which means that each layer was sending 2GB of data for bf16 tensors and 4GB of data with fp32 tensors, which matches the network benchmark. If you were to have a much smaller layer size, I'd recommend adapting the benchmark to that size. For example, if your layer size was only 100M param large, then your payload would be 0.2GB for bf16 tensors. As this is an order of magnitude smaller, the network is likely to give you a lower bandwidth, and you should use that in your calculations.

footnote: if parts of your model are frozen, then there will be less data sent in syncing the gradients. in IDEFICS we had more than half of the model frozen, so when grads were reduced we only had about half the traffic.

Which gives us:

- `comms = 3 * 2 * 80 / 42.5 = 11 sec`

- $\text{compute} = 4 * 2 * 80 * 1024 * 3584 / (512 * 1e3 * 250) = 18 \text{ sec}$

If we check against our IDEFICS-80B logs, which had each iteration at about 49 seconds.

So the good news is that the math checks out as comms + compute are in the ballpark of the measured time, except

We can do another sanity check by feeding the compute formulae 90 TFLOPS that we logged, in which case:

- $\text{compute} = 4 * 2 * 80 * 1024 * 3584 / (512 * 1e3 * 90) = 51 \text{ sec}$

and so 49 and 51 secs are pretty close. Except this tells us nothing since the logged TFLOPS were calculated using this formula, so, of course, it should match.

What I'd expect in the best case is where I have used close to theoretical peak TFLOPS in the formula and received the compute estimate to be about the same as the actual compute time measured on the system. Remember that since comms are interleaved with compute, when we measure `forward+backward` wallclock time it includes comms in it.

What's the conclusion? I'd say more investigation is needed as clearly there are additional hidden bottlenecks here. I no longer have access to this setup to investigate, so I will repeat this exercise afresh when I train another largish model and share the updated math with you. But this workout should give you a feeling for what's going on behind the scenes and how all these numbers work together.

Also this discussion didn't include into the math gradient accumulation steps (GAS). In the case of IDEFICS-80B it wasn't used. If $\text{GAS}>1$ the theoretical compute time doesn't change, but comms time instead of $3*2*M/\text{GBps}$ would become $\text{GAS}*3*2*M/\text{GBps}$. The weights gathering via `all_gather` for `forward` and `backward` would transpire as many times as there are gradient accumulation steps. In theory for grads it'd need to happen only once, but since there is no place to store intermediary grads of the gathered weight on each GPU it'll have to be reduced GAS times as well. This is for ZeRO-2 and ZeRO-3. For ZeRO-1 GAS>1 requires no additional comms.

We also didn't discuss the `DataLoader` as a potential bottleneck here, but we tested that it was under 1 sec, i.e. a very small overhead.

Going back to comms math, we also didn't take into account various hardware latencies, but when dealing with a large payloads they shouldn't add up a significant additional overhead.

And now you know how long it'll take to transmit that many GBs over the network of your system. For example, if the network were to be 5x slower than the one we used for IDEFICS-80B training, that is 8.5Gbps (68Gbps) then:

- $\text{comms} = 3 * 2 * 80 / 8.5 = 56 \text{ sec}$

which would definitely be a huge bottleneck compared to the faster compute.

If the network were to be 5x faster, that is 212GBs (1700Gbps) then:

- $\text{comms} = 3 * 2 * 80 / 212 = 2 \text{ sec}$

which would be insignificant comparatively to the compute time, especially if some of it is successfully overlapped with the commute.

Also the Deepspeed team empirically [benchmarked a 176B model](#) on 384 V100 GPUs (24 DGX-2 nodes) and found that:

1. With 100 Gbps IB, we only have <20 TFLOPs per GPU (bad)
2. With 200-400 Gbps IB, we achieve reasonable TFLOPs around 30-40 per GPU (ok)
3. For 800 Gbps IB, we reach 40+ TFLOPs per GPU (excellent)

To remind the peak TFLOPS for NVIDIA V100 at fp16 is [125 TFLOPS](#).

But be careful here - this benchmark is for V100s! Which is about 2-3x slower than A100, and 4-8x slower than H100 for half-precision. So the comms have to be at least 4-8x faster for H100 nodes to match the above table at half precision. We need more benchmarks with more recent hardware.

footnote: the 2-3x range is because the official specs claim 3x TFLOPS increase for V100->A100, and A100->H100 each, but users benchmarking the difference report at most 2.5x improvements.

They also noticed that when training at scale, the communication overhead is more pronounced with small micro-batch size per GPU. And we may not be able to increase micro-batch size since global-batch size is often fixed to achieve good model convergence rate. This is solved by the recently introduced [ZeRO++](#).

Finally, when doing the math above you need to know the actual bandwidth you get on your setup - which changes with payload size - the larger the payload the better the bandwidth. To get this information you need to look at your `reduce_bucket_size` and `prefetch_bucket_size` settings in the DeepSpeed configuration file for reduction and prefetch correspondingly. The default is 0.5B params, which is 1GB in half-precision (0.5B x 2 bytes), or 2GB (0.5B x 4 bytes) if you use fp32 precision. So in order to measure the actual throughput you need to run an `a11_reduce` benchmark with that payload and see what bandwidth gets reported. Then you can feed it to the calculations above.

Which Strategy To Use When

Here is a very rough outline at which parallelism strategy to use when. The first on each list is typically faster.

▷ Single GPU

- Model fits onto a single GPU:
 1. Normal use
- Model doesn't fit onto a single GPU:
 1. ZeRO + Offload CPU and optionally NVMe
 2. as above plus Memory Centric Tiling (see below for details) if the largest layer can't fit into a single GPU
- Largest Layer not fitting into a single GPU:
 1. ZeRO - Enable [Memory Centric Tiling](#) (MCT). It allows you to run arbitrarily large layers by automatically splitting them and executing them sequentially. MCT reduces the number of parameters that are live on a GPU, but it does not affect the activation memory. As this need is very rare as of this writing a manual override of `torch.nn.Linear` needs to be done by the user.

▷ Single Node / Multi-GPU

- Model fits onto a single GPU:
 1. DDP - Distributed DP
 2. ZeRO - may or may not be faster depending on the situation and configuration used
- Model doesn't fit onto a single GPU:
 1. PP
 2. ZeRO
 3. TP

With very fast intra-node connectivity of NVLINK or NVSwitch all three should be mostly on par, without these PP will be faster than TP or ZeRO. The degree of TP may also make a difference. Best to experiment to find the winner on your particular setup.

TP is almost always used within a single node. That is TP size \leq gpus per node.

- Largest Layer not fitting into a single GPU:
 1. If not using ZeRO - must use TP, as PP alone won't be able to fit.
 2. With ZeRO see the same entry for "Single GPU" above

▷ Multi-Node / Multi-GPU

- If the model fits into a single node first try [ZeRO with multiple replicas](#), because then you will be doing ZeRO over

the faster intra-node connectivity, and DDP over slower inter-node

- When you have fast inter-node connectivity:
 1. ZeRO - as it requires close to no modifications to the model
 2. PP+TP+DP - less communications, but requires massive changes to the model
- when you have slow inter-node connectivity and still low on GPU memory:
 1. DP+PP+TP+ZeRO-1

Software Tune Up For The Best Performance

The faster you can make your model to train the sooner the model will finish training, which is important not only to being first to publish something, but also potentially saving a lot of money.

In general maximizing throughput is all about running many experiments and measuring the outcome and choosing the one that is superior.

In certain situations your modeling team may ask you to choose some hyper parameters that will be detrimental to throughput but overall beneficial for the overall model's success.

Glossary and concepts

- HFU: Hardware FLOPS Utilization
- MFU: Model FLOPS Utilization

MACs vs FLOP vs FLOPS vs FLOP/s

This section is here to try to disambiguate the common performance metric definitions and their relationship to each other.

MAC vs FLOP:

- 1 FLOP (FLoating point OPeration) can be one of addition, subtraction, multiplication, or division operation.
- 1 MAC (Multiply-ACCumulate) operation is a multiplication followed by an addition, that is: $a * b + c$

Thus 1 MAC = 2 FLOPs. It's also quite common for modern hardware to perform 1 MAC in a single clock cycle.

Please note that to calculate the number of MACs in relationship to FLOPs the reverse logic applies, that is MACs = 0.5 FLOPs - it's somewhat confusing since we have just said that 1 MAC = 2 FLOPs, but it checks out - observe: 100 FLOPs = 50 MACs - because there are 2 FLOPs in each MAC.

Moreover, while 1 MAC = 2 FLOPs, the reverse isn't necessarily true. That is 2 FLOPs isn't necessarily equal to 1 MAC. For example, if you did $.5 * .6$ 100 times it'd be 100 FLOPs, which here would equal to 100 MACs, because here only the multiply part of the MAC is executed.

FLOP vs FLOPS vs FLOP/s

- 1 FLOP (FLoating point OPeration) is any floating point addition, subtraction, multiplication, or division operation.
- 1 FLOPS (FLoating point OPeration per Second) is how many floating point operations were performed in 1 second - see [FLOPS](#)

Further you will find the following abbreviations: GFLOPS = Giga FLOPS, TFLOPS = Tera FLOPS, etc., since it's much easier to quickly grasp 150TFLOPS rather than 150000000000000FLOPS.

There is an ambiguity when FLOPS is used in writing - sometimes people use it to indicate the total quantity of operations, at other times it refers to operations per second. The latter is the most common usage and that is the definition used in this book.

In scientific writing FLOP/s is often used to clearly tell the reader that it's operations per second. Though this particular approach is hard to convert to a variable name since it still becomes `flops` when illegal characters are removed.

In some places you might also see FLOPs, which again could mean either, since it's too easy to flip lower and upper case `s`.

If the definition is ambiguous try to search for context which should help to derive what is meant:

- If it's a math equation and there is a division by time you know it's operations per second.
- If speed or performance is being discussed it usually refers to operations per second.
- If it talks about the amount of compute required to do something it refers to the total amount of operations.

TFLOPS as a performance metric

Before you start optimizing the performance of your training setup you need a metric that you can use to see whether the throughput is improving or not. You can measure seconds per iteration, or iterations per second, or some other such timing, but there is a more useful metric that measures TFLOPS.

Measuring TFLOPS is superior because without it you don't know whether you are close to the best performance that can be achieved or not. This measurement gives you an indication of how far you're from the peak performance reported by the hardware manufacturer.

In this section I will use BLOOM's training for the exemplification. We used 80GB A100 NVIDIA GPUs and we trained in mixed bf16 regime. So let's look at the [A100 spec](#) which tells us:

BFLOAT16 Tensor Core 312 TFLOPS

Therefore we now know that if we were to only run `matmul` on huge bf16 matrices of very specific dimensions without copying to and from the device we should get around 312 TFLOPS max.

Practically though, due to disk IO, communications and copying data from the GPU's memory to its computing unit overhead and because we can't do everything in bf16 and at times we have to do math in fp32 (or tf32) we can really expect much less than that. The realistic value will vary from accelerator to accelerator, but for A100 in 2022 getting above 50% (155 TFLOPS) was an amazing sustainable throughput for a complex 384 GPUs training setup.

footnote: in 2023 the invention of flash attention and other techniques have pushed the bar to more than 50%.

When we first started tuning things up we were at <100 TFLOPS and a few weeks later when we launched the training we managed to get 150 TFLOPS.

The important thing to notice here is that we knew that we can't push it further by much and we knew that there was no more point to try and optimize it even more.

So a general rule of thumb for when you prepare for a massive model training - ask around what's the top TFLOPS one can expect to get with a given accelerator on a multi-node setup with the specified precision - and optimize until you get close to that. Once you did stop optimizing and start training.

footnote: For 80GB A100s in 2022 that was 155, in 2023 it has been pushed to about 180 TFLOPS.

footnote: When calculating TFLOPS it's important to remember that the math is different if [Gradient checkpointing](#) are enabled, since when it's activated more compute is used and it needs to be taken into an account. Usually the cost is of an additional forward path, but recently better methods have been found that saves some of that recomputation.

For decoder transformer models the following is an estimation formula which slightly under-reports the real TFLOPS:

TFLOPS: `model_size_in_B * 4 * 2 * seqlen * global_batch_size / (time_in_sec_per_interation * total_gpus * 1e3)`

The factor of 4 is used with activation/gradient checkpointing, otherwise it will be 3. For 100B+ models, activation checkpointing will almost always be on.

So the $3*2$ is often called "model FLOPs" and $4*2$ - "hardware FLOPs", correlating to MFU and HFU (model and hardware FLOPS per second divided by the accelerator's theoretical peak FLOPS)

```
perl -le '$ng=64; $ms=52; $gbs=1024; $sp=127; $seqlen=2048; print $ms*4*2*$seqlen*$gbs / ( $sp * $ng * 1e3)'
```

(ng = total gpus, ms = model size in B, gbs = global batch size, sp = throughput in seconds)

Here is the same formula using bash env vars and which breaks down GBS into MBS*DP*GAS (GAS in this case corresponded to pp_chunks which was the number of chunks in the pipeline, but normally GAS just stands for Gradient Accumulation Steps):

```
echo "($MSIZE*4*2*SEQLEN*$MICRO_BATCH_SIZE*$DP_SIZE*$GAS)/($THROUGHPUT*$NNODES*4*1000)" | bc -l
```

The exact formula is in Equation 3 of Section 5.1 of the [Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM](#) paper. You can see the code [here](#).

footnote: For Inference only it'd be: $24Bsh^2 + 4Bs^2h$ floating point operations per layer.

Automating FLOP calculation

Until recently we had to rely on manual FLOP calculations as explained in the previous section - many of those formulas have mistakes in them, and many models behave differently depending on various configuration settings. So it can be tricky to get the FLOP count correctly (and across many different model architectures). But fear not, the awesome PyTorch team developed an automatic way of measuring FLOPs.

```
from torch.utils.flop_counter import FlopCounterMode

flop_counter = FlopCounterMode(mods=model, display=False, depth=None)
with flop_counter:
    model(**input).sum().backward()
total_flops = flop_counter.get_total_flops()
```

Voila, you have the FLOPs counted for you!

In my code I run it only on a 2nd iteration (as the first iteration is likely to have some additional compute that is run once). You don't need to repeat it again, you can just cache its value (well, unless you have a situation where iterations aren't the same for some reason).

So all that remains is measuring the time it took each specific iteration to run and dividing FLOPs by time in seconds and $1e12$ to get the performance TFLOPS.

```
tflops = total_flops / time / 1e12
```

This will give you a slightly different value on each iteration.

MFU vs HFU

Model FLOPS Utilization (MFU) and Hardware FLOPS Utilization (HFU) estimate how well the hardware is being utilized during forward and backward passes of the model (including any syncing networking overhead and possibly DataLoader IO).

```
MFU = Estimated_Achieved_FLOPS / Theoretical_FLOPS
HFU = Actual_Achieved_FLOPS / Theoretical_FLOPS
```

HFU measures the actual FLOPS. For example, the technique of [Gradient checkpointing/Activation Recomputation](#) repeats all or some parts of the `forward` pass a second time, so factually more FLOPs (Floating point Operations) are used. Whereas MFU ignores implementation details and accounts only for the theoretical needs of the computation and thus less accurate.

[Reducing Activation Recomputation in Large Transformer Models](#) is a good paper to read about these concepts.

Theoretical_FLOPS is what you see on the official accelerator specs. You can find the table of these values for high end accelerators [here](#). So let's use H100 as an example. Its BF16 theoretical TFLOPS is 989 TFLOPS.

Now, say, you measured your actual training loop's performance and it was 400 TFLOPS as actual achieved FLOPS. Then your MFU is:

$$\text{HFU} = 400/989 = 0.40\%$$

If you didn't use activation recomputation feature (not repeating `forward`) your HFU and MFU would be the same. If you did use it, your calculation will lead to less FLOPs and thus lower FLOPS and thus MFU will be lower than HFU.

For example [Megatron-LM](#) published the following stats for A100-80GB:

Model Size	Model FLOPs Utilization	Hardware FLOPs Utilization
22B	41.5%	43.7%
175B	51.4%	52.8%
530B	56.0%	57.0%
1T	56.3%	57.0%

As you can see, since Megatron-LM was using activation recomputation for these trainings, MFU < HFU.

More recent H100+A100 MFU/HFU numbers have been published [here](#).

Now, whenever you see MFU or HFU numbers published you have to be careful comparing those numbers to any other similar numbers, until you know that the same way was used to calculate FLOPS. Since `HFU=Actual_Achieved_FLOPS/Theoretical_FLOPS` and `Theoretical_FLOPS` are fixed, the only variable here is the `Actual_Achieved_FLOPS` and since most of the time an estimated value is calculated based on parameter shapes, there are many versions of calculations out there, some of which are slightly imprecise whereas others are very imprecise. Compilers may also impact the effective FLOPS, by optimizing some operations away. Moreover you don't know how iteration time was measured.

To recall `TFLOPS = FLOPs / iteration_duration`. So, in order to do a fair comparison the 2 main questions to ask are:

1. Was the total used floating point operations calculated in the same way?
2. Was the time component calculated back-to-back of each iteration, including the `DataLoader` and logging, vs only `fwd+bwd` parts.

If either one or both of these mismatch then you can't make a fair comparison.

Unfortunately, most of the time papers and blog posts just report the MFU number w/o a link to how it was calculated.

But, do not fear, if you have trouble comparing your results with competing results, remember the measurement artifacts described above. These artifacts do not improve the bottom-line throughput, thus, as long as you consistently use whatever way you choose to calculate TFLOPS, you will immediately see when your application's performance has improved or degraded, as relative numbers are most important for you.

MFU is a very rough approximation

Most of the training/finetuning regimes use a mixed precision, yet when MFU/HFU are calculated the fastest format's compute is used. For example, for a BF16-mixed precision training some parts of the compute are done in BF16, yet others in FP32! but we measure the FLOPS as if everything was done in BF16, which, of course, leads to a very imprecise measurement. Ideally, each segment of the compute will be measured separately to account to when which format was used. The reason it sort of works is because the smaller format compute usually dominates those mixed precision training regimes.

Moreover, depending on the cluster setup - in particular storage IO and network IO are heavily involved, the same software may deliver different MFUs, because not all clusters are created equal. Therefore it's OK to compare your particular setup's MFU before and after the optimization, but it's very difficult to compare your setup's MFU to another team's cluster's MFU.

How To Improve Speed and Save Memory

The more GPU memory you have for your batch size (BS) the more efficient the GPUs will be at performing compute, and the faster you will complete your task since you will be able to go through data faster.

Of course, this section is crucial for when you get GPU OOM with even BS=1 and you don't want to rent/buy more hardware.

Here is an overview of what features can help to either improve speed or save memory

Method	Speed	Memory
Gradient accumulation	Yes	Yes
Gradient checkpointing	No*	Yes
Mixed precision training	Yes	No
Batch size	Yes	Yes
Optimizer choice	Yes	Yes
DataLoader	Yes	No
DeepSpeed Zero	No	Yes
Flash Attention	Yes	Yes

* Gradient checkpointing slows things down for the given batch size, but since it frees up a lot of memory, enabling a much larger BS, it actually improves the overall speed.

Anatomy of Model's Operations

Transformers architecture includes 3 main groups of operations grouped below by compute-intensity.

1. Tensor Contractions

Linear layers and components of Multi-Head Attention all do batched **matrix-matrix multiplications**. These operations are the most compute-intensive part of training a transformer.

2. Statistical Normalizations

Softmax and layer normalization are less compute-intensive than tensor contractions, and involve one or more **reduction operations**, the result of which is then applied via a map.

3. Element-wise Operators

These are the remaining operators: **biases**, **dropout**, **activations**, and **residual connections**. These are the least compute-intensive operations.

This knowledge can be helpful to know when analyzing performance bottlenecks.

This summary is derived from [Data Movement Is All You Need: A Case Study on Optimizing Transformers 2020](#)

Anatomy of Model's Memory Usage

We've seen that training the model uses much more memory than just putting the model on the GPU. This is because there are many components during training that use GPU memory. The components on GPU memory are the following:

1. model weights
2. optimizer states
3. gradients
4. forward activations saved for gradient computation
5. temporary buffers
6. functionality-specific memory

A typical model trained in mixed precision with AdamW requires 18 bytes per model parameter plus activation memory and temp memory.

Let's look at the details.

Model Weights:

- 4 bytes * number of parameters for fp32 training
- 6 bytes * number of parameters for mixed precision training (maintains a model in fp32 and one in fp16/bf16 in memory)

Optimizer States:

- 8 bytes * number of parameters for normal AdamW (maintains 2 states)
- 4 bytes * number of parameters for AdamW running at bf16. See [this work](#) that uses [AnyPrecisionAdamW](#).
- 4 bytes * number of parameters for optimizers like SGD with momentum (maintains only 1 state) or LION, or Adafactor (and others) (Adafactor uses some additional memory beside 4 bytes)
- 2 bytes * number of parameters for 8-bit AdamW optimizers like [bitsandbytes](#)

Gradients

- 4 bytes * number of parameters for either fp32 precision and in some frameworks with mixed half-precision precision training.
- 2 bytes * number of parameters for non-mixed half-precision and in some frameworks with mixed half-precision precision training.

Forward Activations

- size depends on many factors, the key ones being sequence length, hidden size and batch size.

There are the input and output that are being passed and returned by the forward and the backward functions and the forward activations saved for gradient computation.

Temporary Memory

Additionally there are all kinds of temporary variables which get released once the calculation is done, but in the moment these could require additional memory and could push to OOM. Therefore when coding it's crucial to think strategically about such temporary variables and sometimes to explicitly free those as soon as they are no longer needed.

Functionality-specific memory

Then your software could have special memory needs. For example, when generating text using beam search, the software needs to maintain multiple copies of inputs and outputs.

For inference, the math is very similar to training, minus optimizer states and gradients. And for model weights there is just a single multiplier of the number of parameters:

- 6 bytes in mixed precision (4+2)
- 4 bytes in fp32
- 2 bytes in half precision
- 1 byte in quantized int8 precision

Another excellent resource that takes you through the memory needs and other requirements is [Transformer Math 101](#).

The [EAI cookbook](#) contains a set of [calculation scripts](#) that can output the theoretical memory overhead for a given training or inference calculation run based on your configuration and setup.

There is a very handy [GPU VRAM Estimator](#) from Alexander Smirnov, and [the notes to how it works](#).

Additional GPU memory usage

In addition to the memory usage described in the previous section, there are other consumers of the GPU memory - so you never get the full memory for your model's use.

Preloaded CUDA kernels memory usage

When PyTorch uses CUDA for the first time, it may use up 0.5-2GB of GPU memory, reducing the GPU's total available memory.

The size of allocated memory for cuda kernels varies between different GPUs, and also it can be different between pytorch versions. Let's allocate a 4-byte tensor on cuda and check how much GPU memory is used up upfront.

With `pytorch==1.10.2`:

```
$ CUDA_MODULE_LOADING=EAGER python -c "import torch; x=torch.ones(1).cuda(); free, total = map(lambda x: x/2**30, torch.cuda.mem_get_info()); \nused=total-free; print(f'pt={torch.__version__}: {used=:0.2f}GB, {free=:0.2f}GB, {total=:0.2f}GB')"\npt=1.10.2: used=1.78GB, free=77.43GB, total=79.21GB
```

With `pytorch==1.13.1`:

```
$ CUDA_MODULE_LOADING=EAGER python -c "import torch; x=torch.ones(1).cuda(); free, total = map(lambda x: x/2**30, torch.cuda.mem_get_info()); \nused=total-free; print(f'pt={torch.__version__}: {used=:0.2f}GB, {free=:0.2f}GB, {total=:0.2f}GB')"\npt=1.13.1: used=0.90GB, free=78.31GB, total=79.21GB
```

The older pytorch "wasted" 1.78GB of A100, the newer only 0.9GB, thus saving a whooping 0.9GB, which can be the saving grace for the OOM situations.

`CUDA_MODULE_LOADING=EAGER` is needed in the recent pytorch version if we want to force cuda kernels pre-loading, which are otherwise lazy-loaded on demand. But do not use this setting in production since it's likely to use more memory than needed. The whole point of lazy-loading is to load only the kernels that are needed.

With `pytorch==2.1.1`:

```
$ CUDA_MODULE_LOADING=EAGER python -c "import torch; x=torch.ones(1).cuda(); free, total = map(lambda x: x/2**30, torch.cuda.mem_get_info()); \n used=total-free; print(f'pt={torch.__version__}: {used=:0.2f}GB, {free=:0.2f}GB, {total=:0.2f}GB')"\n pt=2.1.1+cu121: used=0.92GB, free=78.23GB, total=79.15GB
```

As compared to the lazy mode:

```
$ python -c "import torch; x=torch.ones(1).cuda(); free, total = map(lambda x: x/2**30, \n torch.cuda.mem_get_info()); \n used=total-free; print(f'pt={torch.__version__}: {used=:0.2f}GB, {free=:0.2f}GB, {total=:0.2f}GB')"\n pt=2.1.1+cu121: used=0.47GB, free=78.68GB, total=79.15GB
```

There is a 450MB difference, but here we only loaded kernels to do `torch.ones` - the actual memory allocated at run time with other code using torch API will be somewhere between 0.47 and 0.92GB.

Memory fragmentation

As the model allocates and frees tensors, the memory could fragment. That is there could be enough free memory to allocate, say, 1GB of contiguous memory, but it could be available in 100s of small segments spread out through the memory and thus even though the memory is available it can't be used unless very small allocations are made.

Environment variable `PYTORCH_CUDA_ALLOC_CONF` comes to help and allows you to replace the default memory allocation mechanisms with more efficient ones. For more information see [Memory management](#).

Batch sizes

First, there are usually two batch sizes:

1. micro batch size (MBS), also known as batch size per gpu - this is how many samples a single gpu consumes during a model's single `forward` call.
2. global batch size (GBS) - this is the total amount of samples consumed between two optimizer steps across all participating GPUs.

Model replica is how many gpus are needed to fit the full model.

- If the model fits into a single GPU a model replica takes 1 GPU. Usually then one can use multiple GPUs to perform [Data Parallelism](#)
- If the model doesn't fit into a single GPU, it'd usually require some sort of sharding technique - it can be [Tensor Parallelism](#) (TP), [Pipeline Parallelism](#) (PP), or [ZeRO Data Parallelism](#) (ZeRO-DP).

You can have as many data streams as there are replicas. Which is the same as the value of DP.

- So in the simple case of a model fitting into a single GPU. There are as many data streams as there are GPUs. $DP=N_GPUS$
- when the model doesn't fit onto a single GPU, then $DP=N_GPUs/(TP*PP)$ in the case of 3D parallelism and $DP=ZeRO-DP$ in the case of ZeRO parallelism.

Going back to our global batch size (GBS) it's calculated as:

```
GBS = MBS*DP
```

So if you have 8 gpus (N_GPUS=8) and your MBS=4 and you do DP you end up with having GBS=32 because:

$$GBS = MBS * DP = 4 * 8 = 32$$

If you use TP with a degree of 2 (TP=2) and PP with a degree of 2 (PP=2) this means each model replica takes 4 gpus (TP*PP), and thus with N_GPUS=8

$$DP = N_GPUS / (TP * PP) = 8 / (2 * 2) = 2$$

and now GBS becomes:

$$GBS = MBS * DP = 4 * 2 = 8$$

If your training setup requires [Gradient Accumulation](#), one usually defines the interval of how many steps to wait before performing a gradient accumulation. The term is usually Gradient Accumulation Steps (GAS). If GAS=4 (i.e. sync grads every 4 steps) and TP=1, PP=1 and DP=8:

$$DP = N_GPUS / (TP * PP) = 8 / (1 * 1) = 8$$
$$GBS = MBS * DP * GAS = 4 * 8 * 4 = 128$$

Typically you want to make the micro batch size as large as possible so that the GPU memory is close to being full, but not too full.

With large models usually there is not much free GPU memory left to have a large micro batch size, therefore every additional sample you can fit is important.

While it's super important that sequence length and hidden size and various other hyper parameters are high multiples of 2 (64, 128 and higher) to achieve the highest performance, because in most models the batch dimension is flattened with the sequence length dimension during the compute the micro batch size alignment usually has little to no impact on performance.

Therefore if you tried to fit a micro batch size of 8 and it OOM'ed, but 7 fits - use the latter rather than 4. The higher the batch size the more samples you will be able to fit into a single step.

Of course, when using hundreds of GPUs your global batch size may become very large. In that case you might use a smaller micro batch size or use less GPUs or switch to a different form of data parallelism so that the GPUs work more efficiently.

Gradient Accumulation

The idea behind gradient accumulation is to instead of calculating the gradients for the whole batch at once to do it in smaller steps. The way we do that is to calculate the gradients iteratively in smaller batches by doing a forward and backward pass through the model and accumulating the gradients in the process. When enough gradients are accumulated we run the model's optimization step. This way we can easily increase the overall batch size to numbers that would never fit into the GPU's memory. In turn, however, the added forward and backward passes can slow down the training a bit.

Gradient Accumulation Steps (GAS) is the definition of how many steps are done w/o updating the model weights.

When using Pipeline parallelism a very large Gradient Accumulation is a must to keep the [pipeline's bubble to the minimum](#).

Since the optimizer step isn't performed as often with gradient accumulation there is an additional speed up here as well. The following benchmarks demonstrate how increasing the gradient accumulation steps improves the overall throughput (20-30% speedup):

- [RTX-3090](#)
- [A100](#)

When [data parallelism](#) is used gradient accumulation further improves the training throughput because it reduces the number of gradient reduction calls, which is typically done via the `all_reduce` collective which costs a 2x size of gradients to be reduced. So for example, if one goes from GAS=1 to GAS=8 in [DistributedDataParallelism](#) (DDP) the network overhead is reduced by 8x times, which on a slow inter-node network can lead to a noticeable improvement in the training throughput.

Gradient Checkpointing

[Gradient Checkpointing](#) is also known as [Activation Recomputation](#) and [Activation Checkpointing](#).

This methodology is only relevant for training, and not during inference.

Enabling gradient checkpointing allows one to trade training throughput for accelerator memory. When this feature is activated instead of remembering the outputs of, say, transformer blocks until the `backward` pass is done, these outputs are dropped. This frees up huge amounts of accelerator memory. But, of course, a `backward` pass is not possible without having the outputs of `forward` pass, and thus they have to be recalculated.

This, of course, can vary from model to model, but typically one pays with about 20-25% decrease in throughput, but since a huge amount of gpu memory is liberated, one can now increase the batch size per gpu and thus overall improve the effective throughput of the system. In some cases this allows you to double or quadruple the batch size if you were already able to do a small batch size w/o OOM. (Recent papers report as high as 30-40% additional overhead.)

Activation checkpointing and gradient checkpointing are 2 terms for the same methodology.

For example, in HF Transformers models you do `model.gradient_checkpointing_enable()` to activate it in your custom Trainer or if you use the HF Trainer then you'd activate it with `--gradient_checkpointing 1`.

XXX: expand on new tech from the paper: [Reducing Activation Recomputation in Large Transformer Models](#) which found a way to avoid most activation recomputations and thus save both memory and compute.

Memory-efficient optimizers

The most common optimizer is Adam. It and its derivatives all use 8 bytes per param (2x fp32 tensors - one for each momentum), which account for almost half the memory allocation for the model, optimizer and gradients. So at times using other optimizers may save the day, if they successfully train that is. Not all optimizers are suitable for all training tasks.

4-byte optimizers:

- There are optimizers like Adafactor that need only 4 bytes. Same goes for the recently invented [LION optimizer](#).
- AnyPrecisionAdamW. Some courageous souls try to do the whole training in BF16 (not mixed precision!), including the optimizer and thus need only 4 bytes per parameter for optim states. See [this work](#). Hint: this optimizer requires Kahan summation and/or stochastic rounding, see [Revisiting BFloat16 Training \(2020\)](#). You need only 8 bytes per parameter for weights, optim states and gradients here! Instead of 18!

2-byte optimizers:

- There are quantized solutions like `bnb.optim.Adam8bit` which uses only 2 bytes instead of 8 (1 byte per momentum). You can get it from [here](#). Once installed, if you're using HF Trainer, you can enable it on with just passing `--optim adamw_bnb_8bit!`

For speed comparisons see [this benchmark](#) Speed-wise: apex's `apex.optimizers.FusedAdam` optimizer is so far the fastest implementation of Adam. Since pytorch-2.0 `torch.optim.AdamW` added support for `fused=True` option, which brings it almost on par with `apex.optimizers.FusedAdam`.

Model execution speed

forward vs backward Execution Speed

For convolutions and linear layers there are 2x flops in the backward compared to the forward, which generally translates into ~2x slower (sometimes more, because sizes in the backward tend to be more awkward). Activations are usually bandwidth-limited, and it's typical for an activation to have to read more data in the backward than in the forward (e.g. activation forward reads once, writes once, activation backward reads twice, `gradOutput` and output of the forward, and writes once, `gradInput`).

Memory profiler tools

In this chapter we discussed the theoretical math of how much this or that feature should consume in MBs of memory. But often in reality things aren't exactly the same. So you plan for a certain model size and batch sizes but when you come to use it suddenly there is not enough memory. So you need to work with your actual code and model and see which part takes how much memory and where things got either miscalculated or some additional missed overhead hasn't been accounted for.

You'd want to use some sort of memory profiler for that purpose. There are various memory profilers out there.

One useful tool that I developed for quick and easy profiling of each line or block of code is [IPyExperiments](#). You just need to load your code into a jupyter notebook and it'll automatically tell you how much CPU/GPU memory each block allocates/frees. So e.g. if you want to see how much memory loading a model took, and then how much extra memory a single inference step took - including peak memory reporting.

Vector and matrix size divisibility

The paper, [The Case for Co-Designing Model Architectures with Hardware](#) investigates the effects of transformer sizing on the underlying hardware. The [associated scripts](#) allow you to run the benchmarks yourself if you're running on hardware besides NVIDIA V100/A100.

One gets the most efficient performance when batch sizes and input/output neuron counts are divisible by a certain number, which typically starts at 8, but can be much higher as well. That number varies a lot depending on the specific hardware being used and the dtype of the model.

For fully connected layers (which correspond to GEMMs), NVIDIA provides recommendations for [input/output neuron counts](#) and [batch size](#).

[Tensor Core Requirements](#) define the multiplier based on the dtype and the hardware. For example, for fp16 a multiple of 8 is recommended, but on A100 it's 64!

For parameters that are small, there is also [Dimension Quantization Effects](#) to consider, this is where tiling happens and the right multiplier can have a significant speedup.

[The Case for Co-Designing Model Architectures with Hardware](#) provides much greater detail on tile/wave quantization and the number of attention heads, but the highlights are:

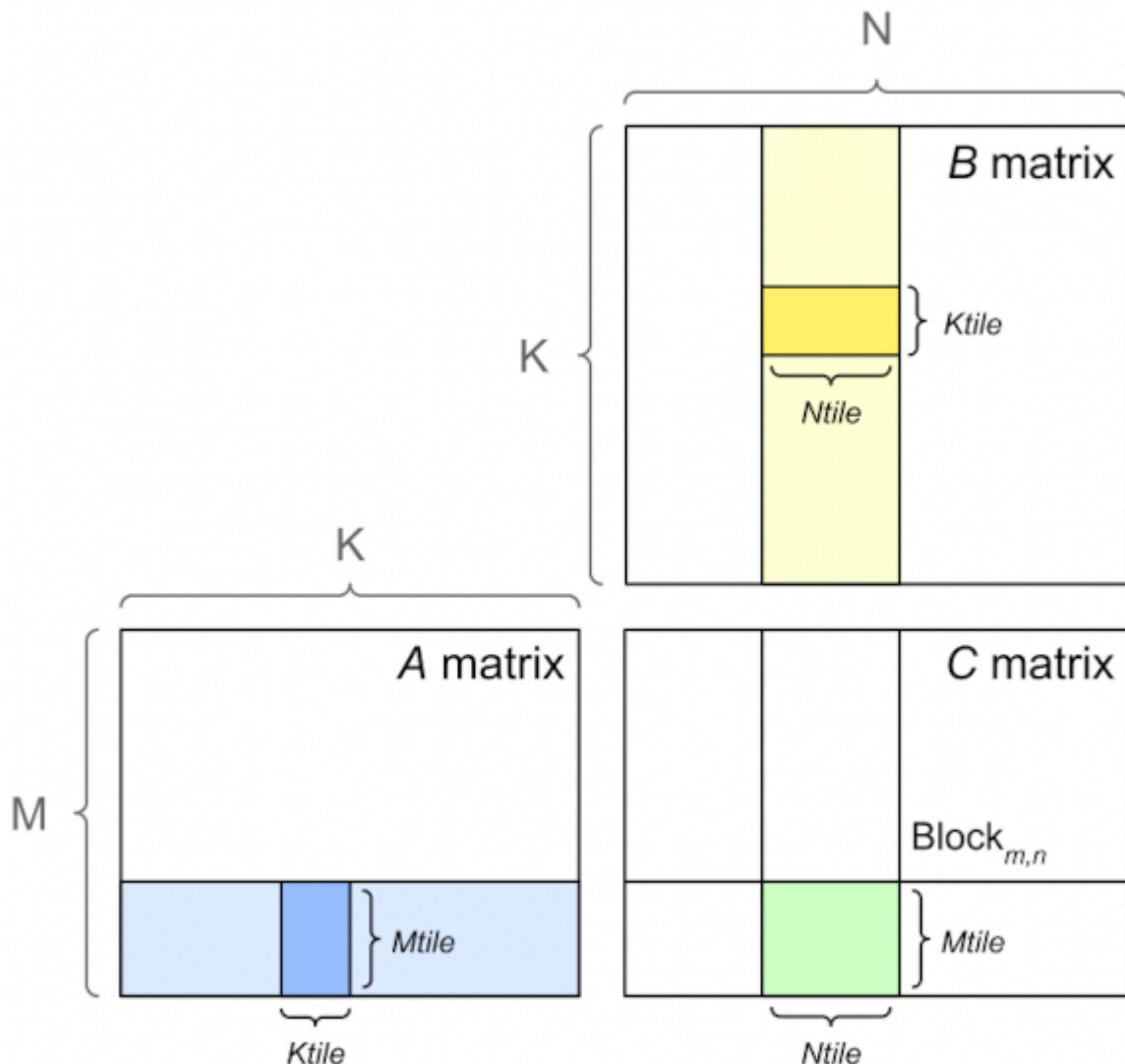
Tile and wave quantization

Notation:

- a: Number of attention heads
- h: Hidden dimension size
- s: Sequence length
- b: Microbatch size
- t: Tensor-parallel size

First, some background.

NVIDIA GPUs divide the output matrix into regions or tiles as shown in the below figure and schedule them to one of the available streaming multiprocessors (SM) on the GPU (e.g., A100 GPUs have 108 SMs). Each tile or thread block is processed in a Tensor Core, which NVIDIA introduced for fast tensor operations. NVIDIA Tensor Cores are only available for GEMMs with appropriate dimensions. Tensor Cores can be fully utilized when GEMM dimensions m , k , and n are multiples of 16 bytes and 128 bytes for V100 and A100 GPUs, respectively. Since a FP16 element is 2 bytes, this corresponds to dimension sizes that are multiples of 8 and 64 elements, respectively. If these dimension sizes are not possible, Tensor Cores perform better with larger multiples of 2 bytes.



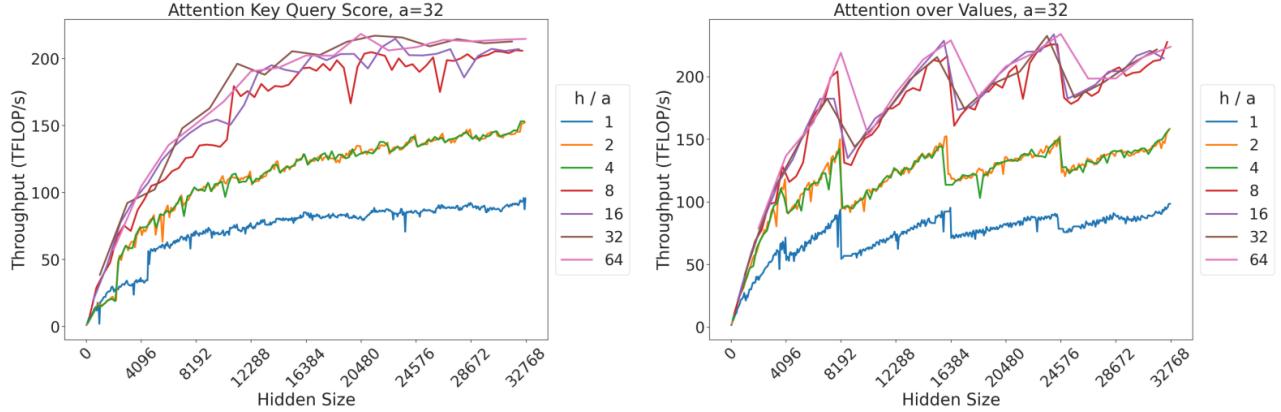
There are multiple tile sizes that the kernel can choose from. If the GEMM size does not divide evenly into the tile size, there will be wasted compute, where the thread block must execute fully on the SM, but only part of the output is necessary. This is called the **tile quantization** effect, as the output is quantized into discrete tiles.

Another quantization effect is called **wave quantization**. As the thread blocks are scheduled to SMs, only 108 thread blocks

at a time may be scheduled. If, for example, 109 thread blocks must be scheduled, two rounds, or waves, of thread blocks must be scheduled to GPU. The first wave will have 108 thread blocks, and the second wave will have 1. The second wave will have almost the same latency as the first, but with a small fraction of the useful compute. As the matrix size increases, the last or tail wave grows. The throughput will increase, until a new wave is required. Then, the throughput will drop.

What this means for transformers, is that for a given ratio of h/a , one needs to ensure they're on the crest of a wave. If you're using NVIDIA V100/A100 GPUs, we've already done this work for you in <https://arxiv.org/pdf/2401.14489.pdf>

An example of this for 32 attention heads:



(a) Attention key-query score GEMM throughput for 32 attention heads. (b) Attention over value GEMM throughput for 32 attention heads.

Fig. 7: Attention GEMM performance on A100 GPUs. Each plot is a single series (i.e. if we didn't split, there would be three regions with spikes), but split by the largest power of two that divides h/a to demonstrate that more powers of two leads to better performance up to $h/a = 64$.

More powers of 2 in h/a helps!

Number and size of attention heads

Generally, it's most computationally efficient to keep the ratio of h/a as large as possible without accuracy degradation. A good figure from [The Case for Co-Designing Model Architectures with Hardware](#) showing this effect is:

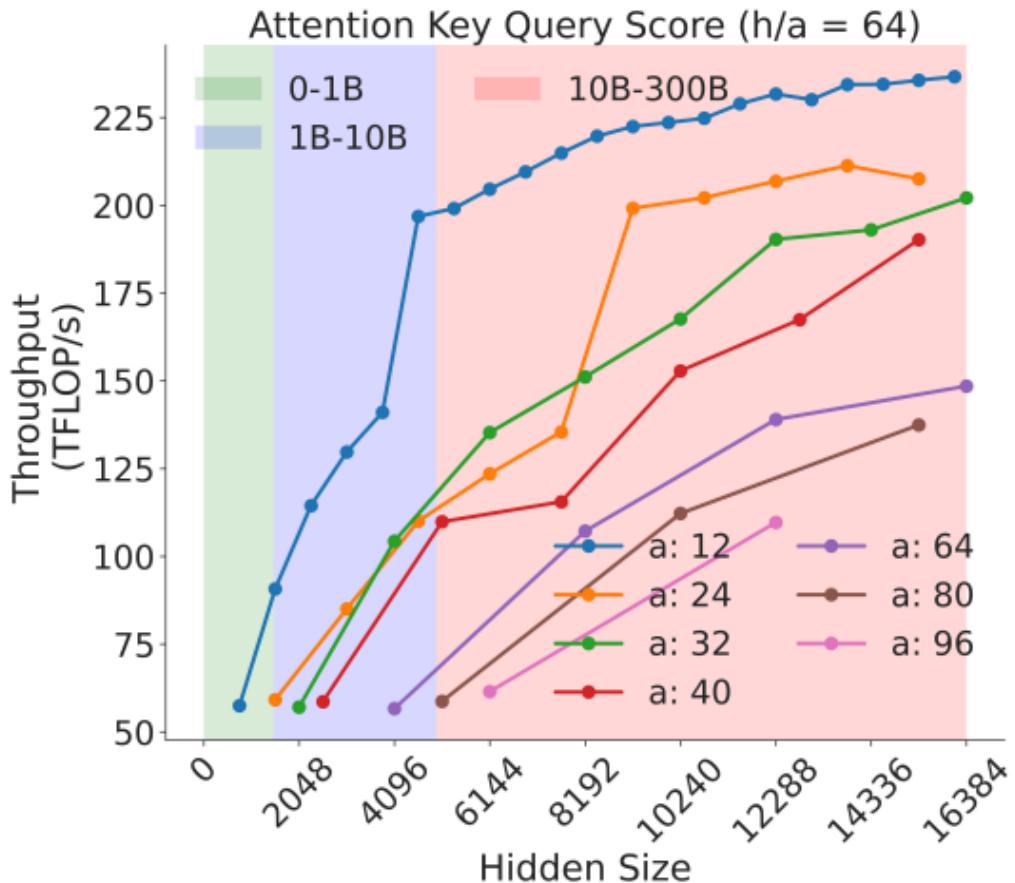


Fig. 8: Attention key-query score GEMM throughput assuming fixed ratio of $\frac{h}{a} = 64$ on A100 GPU

Flash attention

If you're using [Flash Attention](#), good news! These MHA sizing constraints are taken care of for you. Your only constraint is to have a large enough ratio of h/a to saturate your GPU cores:

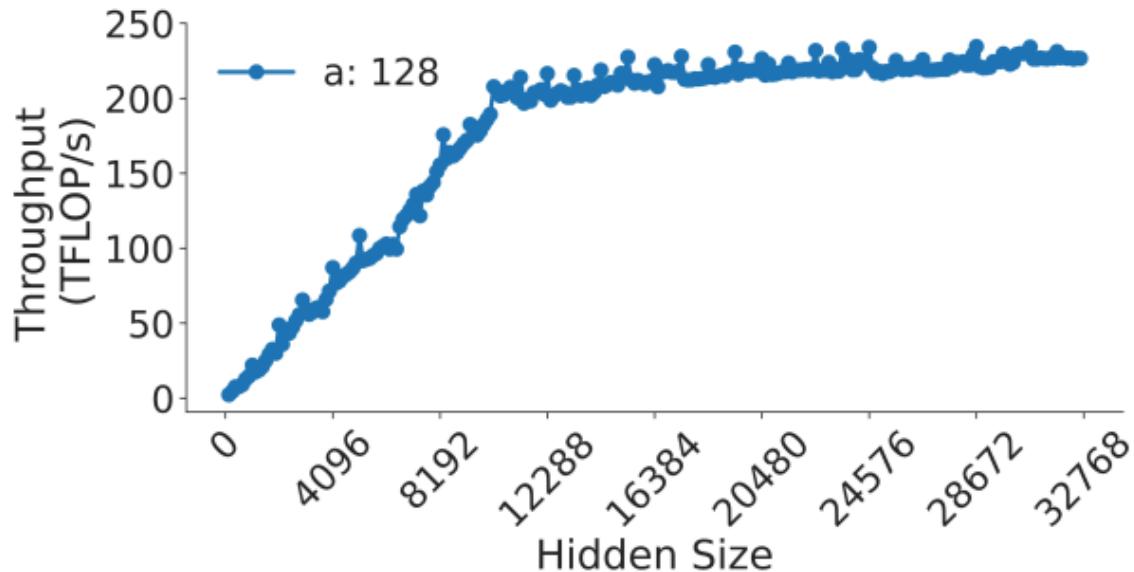


Fig. 12: Sweep over hidden dimension for FlashAttention (v2) [8] on NVIDIA A100 GPU.

SwiGLU-based MLP

Models such as PaLM, LLaMA, Mistral and others use the SwiGLU activation function in place of the more common GLU activation function.

The SwiGLU-based MLP contains an additional learned matrix in its activation function. There the MLP block contains 3 matrices instead of the original 2. To preserve the total number of parameters in the MLP block the paper that introduces [SwiGLU](#) proposes to use $\text{dim_mlp} = 8/3 \cdot \text{dim_attn}$ instead of the typical $\text{dim_mlp} = 4 \cdot \text{dim_attn}$. The [The Case for Co-Designing Model Architectures with Hardware](#) paper provides recommendations for finding the value of hidden dimension (h) that would lead to the best `matmul` performance, and if you used $8/3 \cdot h$ is likely to result in a much slower MLP block, because $8/3$ will break all the alignments.

In order to overcome this problem one only needs to realize that the $8/3$ coefficient is only a suggestion and thus it's possible to find other near by coefficients that would lead to better-shaped MLP matrices. In fact if you look at the publicly available Llama-2 models, its 7B variant uses $11008/4096 = 2.6875$ as a coefficient, which is quite close to $8/3 = 2.667$, and its 70B variant uses a much larger $28672/8192 = 3.5$ coefficient. Here the 70B variant ended up with an MLP block that contains significantly more parameters than a typical transformer block that doesn't use SwiGLU.

Now that we know the recommended coefficient isn't exact and since a good h has already been chosen, one can now search for a good nearby coefficient that still leads to high-performance GEMMs in the MLP. Running a brute-force search reveals that Llama-2-7B's intermediate size is indeed one of the best performing sizes in its range.

Here is [swiglu-maf-bench.py](#) that can be easily adapted to your use-case and once run on the target hardware the training will happen on, you will be able to find the best hidden size of the MLP.

Let's run it on H100 with $h = 4096$:

```
./swiglu-maf-bench.py
Wanted the closest to 10922 d_ff value that leads to the highest TFLOPS (d_hidden=4096)
```

```

Searching 50 steps in the range of 10822 .. 11022
Results: baseline, followed by near-by best performing d_ff results:

d_ff  tflops mlp_params
-----
10922  272.73 134209536
-----
10944  398.38 134479872
10848  395.62 133300224
10880  395.52 133693440
10912  395.16 134086656
11008  395.01 135266304

```

As it can be easily seen the $8/3*4096=10922$ leads to a rather slow performance. But **10944**, which is just 22 bigger than **10922**, gives a whooping 46% speedup for the `matmul`. The total corresponding MLP parameters is printed as well, should you want to choose slightly slower choices but with a different number of parameters.

Final recommendations for model sizing

The full recommendations are:

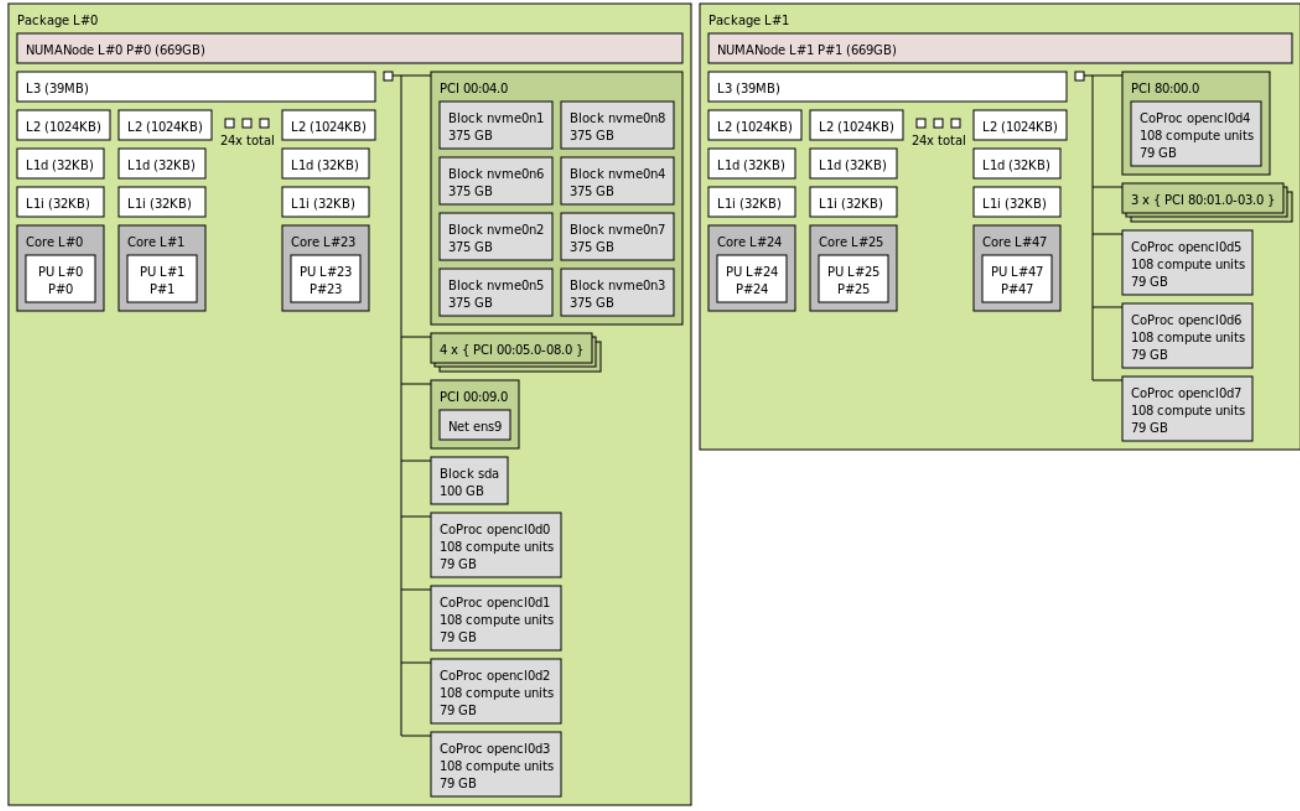
1. Vocab size divisible by 64
2. Microbatch size as large as possible
3. $b*s$, h/a , and h/t should be divisible by a power of 2
4. $(b*a)/t$ should be an integer
5. t should be small as possible
6. For SwiGLU search for the best performing hidden size close to $8/3*h$

NUMA affinity

[Non-uniform memory access \(NUMA\)](#) is a computer memory design used in multiprocessing, where the memory access time depends on the memory location relative to the processor. As modern servers have more than one CPU to get the best performance accelerators residing in the same NUMA node as the corresponding CPU should have the processes bound to that same NUMA node.

First, let's understand what do NUMA nodes signify.

Here is a typical A100 8x GPUs server NUMA nodes diagram:



As you can see it has 2 CPUs, each defining a NUMA block, and each such block contains a group of 4 GPUs. The GPUs are the grey blocks that say CoProc with 108 compute units (SMs) and 79GB of memory.

footnote: the diagram was generated by `lstopo a100.png` from [hwloc](#).

If you're using Hyper-Threads then you want to use `lstopo -1` to see the HT core count presented correctly. For example if you have 2 NUMA nodes with 8 accelerators and 104 physical cpu-cores and 208 logical cores - thus (208/8=26 HT-cores per GPU), then the HT cores will be for:

- gpu0..3 [0, 1, 2, 3, ..., 51, 104, 105, 106, ..., 129]
- gpu4..7 [52, 53, 54, ..., 103, 156, 157, 158, ..., 207]

You first get the physical cpu-core counts and then the remaining HT cores, hence the strange gap.

If this is an NVIDIA node, the other way to easily see the CPU affinity is to run:

\$ nvidia-smi topo -m										
	GPU0	GPU1	GPU2	GPU3	GPU4	GPU5	GPU6	GPU7	CPU Affinity	NUMA Affinity
GPU0	X	NV18	0-51,104-155	0						
GPU1	NV18	X	NV18	NV18	NV18	NV18	NV18	NV18	0-51,104-155	0
GPU2	NV18	NV18	X	NV18	NV18	NV18	NV18	NV18	0-51,104-155	0
GPU3	NV18	NV18	NV18	X	NV18	NV18	NV18	NV18	0-51,104-155	0
GPU4	NV18	NV18	NV18	NV18	X	NV18	NV18	NV18	52-103,156-207	1
GPU5	NV18	NV18	NV18	NV18	NV18	X	NV18	NV18	52-103,156-207	1
GPU6	NV18	NV18	NV18	NV18	NV18	NV18	X	NV18	52-103,156-207	1
GPU7	NV18	X	52-103,156-207	1						

On this H100 cluster you can see the CPU Affinity column which tells you which cores reside together with the first and

the second group of GPUs, and the **NUMA Affinity** column.

Now that it's clear that the various compute components are placed in 2 or more groups, to achieve the best performance we need to ensure that the components communicate within the group they belong to, and avoid any cross-talk. For example, if gpu0 belong to NUMA node 0, then the process that drives this GPU should only use cpu-cores from NUMA node 0.

The same should apply to networking or any other components that you may have control over.

Practically though in my experience so far if your workload is very light on CPU work this change will make very little difference to the overall performance, but can be quite impactful if a lot of CPU use is done. On the other hand if doing the most efficient thing is easy, even the tiniest improvement is likely to accumulate over long training jobs, so it's worth to implement, IMHO.

NUMA process binding

There are multiple ways to accomplish the binding of processes to the cpu-cores of the right NUMA node.

numactl

One of the most common tools to do that is using `numactl`, which sets the NUMA affinity as it launches a new process.

For example, let's see how it can be integrated with the `torchrun` launcher.

This launcher currently needs a helper util [numa-set.sh](#) to perform NUMA affinity settings, once you downloaded it and made it executable, you can now get the right NUMA affinity using:

```
torchrun --nproc_per_node=8 --role : --tee 3 --no-python ./numa-set.sh your-program.py
```

Note: you'd need `numactl` installed on your system for this util to work.

For example, here is how you can validate that the assignments are correct:

```
torchrun --nproc_per_node=8 --role : --tee 3 --no-python ./numa-set.sh python -c \
'import os; cores=os.sched_getaffinity(0); print(f"len(cores)} visible cpu cores: {cores}"}'
```

On a system with 208 HT cpu-cores, you will most likely see:

```
[:0]:104 visible cpu cores: {0, 1, 2, 3, 4, 5...
[:1]:104 visible cpu cores: {0, 1, 2, 3, 4, 5...
[:2]:104 visible cpu cores: {0, 1, 2, 3, 4, 5...
[:3]:104 visible cpu cores: {0, 1, 2, 3, 4, 5...
[:4]:104 visible cpu cores: {52, 53, 54, 55, ...
[:5]:104 visible cpu cores: {52, 53, 54, 55, ...
[:6]:104 visible cpu cores: {52, 53, 54, 55, ...
[:7]:104 visible cpu cores: {52, 53, 54, 55, ...
```

The first 4 accelerators use the first half of the cpu-cores and the other 4 the second half, which matches the earlier explanations of the right setting.

If you remove `./numa-set.sh`, you'd get:

```
torchrun --nproc_per_node=8 --role : --tee 3 --no-python python -c \
'import os; cores=os.sched_getaffinity(0); print(f"len(cores)} visible cpu cores: {cores}"}'
```

You will see that all 8 processes see all 208 cpu-cores:

```
[:0]:208 visible cpu cores: {0, 1, 2, 3, ...
```

so as each process has access to any cpu-core - a cross talk may occur, which may introduce a small performance overhead.

os.sched_setaffinity

You can, of course, change the NUMA affinity after the program was launched. You saw the use of `os.sched_getaffinity` to get the current settings, and the corresponding `os.sched_setaffinity` is used to change it.

```
import os
os.sched_setaffinity(0, [0, 1])
```

Here we told the system that the process running this script (0) can only use cpu-cores 0 and 1.

So now we just need to figure out how to programmatically get the right cpu sets for each accelerator's process. Here is how to do it with [pynvml](#).

pynvml

If you're using NVIDIA GPUs, `pynvml` (`pip install pynvml`) can be very helpful to get all sorts of information about the gpu and not needing to call `nvidia-smi` - in this situation we are going to use for it to tell us the correct affinity given a GPU index.

In [numa-set-pynvml.py](#) you will find a working helper function that you could call at the very top of your training loop like so:

```
local_rank = torch.distributed.get_rank()
set_numa_affinity(0, verbose=True)
```

call it before `DataLoader` is initialized to get the workers use the right cpu-cores!

Normally, the local process rank equals the gpu index, but if one uses `CUDA_VISIBLE_DEVICES` - this might not be true any longer - if you use it, you will need to remap the process rank to the actual index:

```
gpu_index = int(os.environ.get("LOCAL_RANK", 0))
if "CUDA_VISIBLE_DEVICES" in os.environ:
    ids = list(map(int, os.environ.get("CUDA_VISIBLE_DEVICES", "").split(",")))
    gpu_index = ids[gpu_index] # remap
```

The other gotcha can be `CUDA_DEVICE_ORDER` which typically defaults to `PCI_BUS_ID`, but one could also set it to `CUDA_DEVICE_ORDER=FASTEST_FIRST` if you have mixed GPUs, but it's very very unlikely that you will run into this in a high end server setup, so you can safely ignore this.

`srun`

If using SLURM and you're OK with using `srun` as the launcher, rather than `torchrun`, `accelerate`, etc., it'll do all the binding work for you automatically. See the full launcher [here](#).

To make it NUMA affinity-ready all you need to add is these 2 headers:

```
#SBATCH --gres-flags=enforce-binding
#SBATCH --ntasks-per-socket=4
```

`--ntasks-per-socket=4` assumes you have 2 cpu sockets with 8 accelerators - so $8/2=4$ accelerators per socket.

This is an even more precise solution, since it'd assign each process its own group of cpu-cores, rather than just give all the NUMA node 0 cpu-cores to the processes driving accelerators 0-3, and NUMA node 1 cpu-cores to the processes driving accelerators 4-7.

Specific launchers

Various launchers have support for NUMA affinity settings:

- [HF Accelerate](#) has a flag `--enable_cpu_affinity` that you add to the `accelerate` launch command and it'll do this for you. Available since `accelerate>0.28.0`.
- [torchrun](#) doesn't have it, but I showed how to do it in this [section](#).
- `srun` was covered [here](#).

DataLoader

Asynchronous DataLoader

The default setting is `num_workers=0`, which means whenever you call `next(iter(dataloader))` the data is actively fetched in real time - which means there will be an IO overhead and if there are any transforms those would be applied in real time as well - all blocking the accelerator's compute.

The solution is to use the asynchronous `DataLoader` by setting `num_workers > 0`. Typically, unless your `DataLoader` is extremely slow 2 workers is enough:

```
DataLoader(..., num_workers=2, ...)
```

Now when `next(iter(dataloader))` is called the data should be already in the CPU memory with all the transforms done. It still needs to be copied to the accelerator memory - to speed that up see [Pinned memory and non blocking device copy](#).

Here is a benchmark which emulates a slow data transform: [num-workers-bench.py](#)

```
num_workers=0: average time: 5.388
num_workers=1: average time: 3.900
num_workers=2: average time: 1.333
num_workers=3: average time: 0.839
num_workers=4: average time: 0.875
```

So you can see that in this particular case the speed was dramatically improving up to 3 workers. If the `DataLoader` is very light and does very little the difference will be much smaller, but 0 workers will always lead to the biggest overhead.

By measuring the performance of your workload you can finetune this number by trying lower and higher values. But remember that each one of the workers may consume a lot of CPU memory. So on a node of 8 accelerators with 2 workers, that would be 16 additional processes. Nowadays, the compute nodes have often hundreds of cpu cores and a TBs of CPU memory so there should be plenty of resources for many workers to be supported. In the past it was a different story.

Also note that because any data transforms are applied asynchronously and ahead of time, the CPU and memory speed don't matter much in this case. e.g. with 2 workers as long as the next iteration data preparation takes less than 2 compute iterations the `DataLoader` shouldn't be a bottleneck.

Beware that sometimes you may encounter problems using `num_workers > 0` - the pytorch Issues has a few related Issues that haven't been resolved in several years, where a worker would hang. In particular when having 2 `Dataloaders`. In fact we had this problem during BLOOM-176B training, where the training `Dataloader` worked fine with 2 workers, but once eval `Dataloader` was added it'd randomly hang - so we after failing to figure it out we resorted to a workaround of `num_workers=0` just for the eval and switch to doing it very rarely. Eventually, we stopped doing eval altogether and started doing lm-harness style async eval done to the saved intermediary checkpoints instead, which also sped up the training process.

case study: during IDEFICS-80B training we were using a streaming `Dataloader` which worked really bad and it was consuming huge amounts of memory per worker, and it'd spike often, and we had about 1TB of CPU memory and we couldn't spawn enough workers - so the `Dataloader` was a bottleneck. We didn't have time to find a better solution at that time so we finished training with it.

Pinned memory and non-blocking device copy

A combination of:

1. `DataLoader(pin_memory=True, ...)`
2. `batch.to(device="cuda", non_blocking=True)`

is likely to make the `DataLoader` less of a bottleneck.

1. Enabling pinned memory allows for a more efficient data transfer from CPU to accelerator memory.
2. non-blocking will further speed things up by allowing some overlap between compute and data movements

Here is a small benchmark demonstrating the difference: [pin-memory-non-block-bench.py](#). When I run it on an A100 80GB-PCIe, the output was:

```
pin_memory= True, non_blocking= True: average time: 0.459
pin_memory= True, non_blocking=False: average time: 0.522
pin_memory=False, non_blocking= True: average time: 0.658
pin_memory=False, non_blocking=False: average time: 0.646
```

so you can see that `pin_memory=True+non_blocking=True` is a worthy change.

For more background you can read [1](#) and [2](#).

Notes:

- Pinned memory is treated specially by the OS, by preventing the paging out when the total available memory is low, so it reduces the total amount of available memory to other programs. Thus use with care.
- I recall that a few times people reported problems when using pinned memory - I think it's mostly when the system doesn't have much CPU memory to start with or they used too much of pinned memory, so the OS can start swapping heavily.
- If you measure your [per iteration TFLOPS](#) you can compare the throughput with and w/o these changes and choose the one that works the best. It'd be even easier to see the impact if you measure the `DataLoader` overhead separately from the the forward/backwards and post-compute (usually logging, which can be surprisingly slow at times).

torch.compile

`torch.compile` will eventually speed things up amazingly for both training and inference. It's very difficult to make it work well on a random model, because the level of complexities to overcome is huge. There are some models that already work well with it, but many are still a long term work in progress.

If you tried it and thing don't work you:

1. may report it to the PyTorch team, ideally with a small reproducible example
2. can try to read this extensive [torch.compile, the missing manual](#) and you might be able to make some things work, and may still need to report some issues to PyTorch

One thing is certain is that you want to use the latest pytorch version, which most likely would be some recent nightly build, rather than the last released version (though you might start with the latter).

Automatic garbage collection

Python periodically performs an automatic garbage collection based on internal heuristics. In an LLM-training scenario with hundreds to thousands of accelerators used in synchronization - if different ranks follow even slightly different code paths the automatic garbage collection process could be triggered at different times for different ranks. Which means that one or more ranks could be slower than other ranks while performing this operation, and thus becoming stragglers, slowing down the whole ensemble.

Usually one can see this by studying [the MFU plot](#) where downward spikes can be observed.

If this happens to your training you can disable the automatic garbage collection with:

```
import gc
gc.disable()
```

at the beginning of your trainer and then manually perform garbage collection at the desired intervals. For example, calling this once in a training iteration:

```
import gc
gc.collect()
```

Refer to [gc's manpage](#) for more nuances.

Fault Tolerance

Regardless of whether you own the ML training hardware or renting it by the hour, in this ever speeding up domain of ML, finishing the training in a timely matter is important. As such if while you were asleep one of the GPUs failed or the checkpoint storage run out of space which led to your training crashing, you'd have discovered upon waking that many training hours were lost.

Due the prohibitively high cost of ML hardware, it'd be very difficult to provide redundancy fail-over solutions as it's done in Web-services. Nevertheless making your training fault-tolerant is achievable with just a few simple recipes.

As most serious training jobs are performed in a SLURM environment, it'll be mentioned a lot, but most of this chapter's insights are applicable to any other training environments.

Always plan to have more nodes than needed

The reality of the GPU devices is that they tend to fail. Sometimes they just overheat and shut down, but can recover, at other times they just break and require a replacement.

The situation tends to ameliorate as you use the same nodes for some weeks/months as the bad apples get gradually replaced, but if you are lucky to get a new shipment of GPUs and especially the early GPUs when the technology has just come out, expect a sizeable proportion of those to fail.

Therefore, if you need 64 nodes to do your training, make sure that you have a few spare nodes and study how quickly you can replace failing nodes should the spares that you have not be enough.

It's hard to predict what the exact redundancy percentage should be, but 5-10% shouldn't be unreasonable. The more you're in a crunch to complete the training on time, the higher the safety margin should be.

Once you have the spare nodes available, validate that your SLURM environment will automatically remove any problematic nodes from the pool of available nodes so that it can automatically replace the bad nodes with the good ones.

If you use a non-SLURM scheduler validate that it too can do unmanned bad node replacements.

You also need at least one additional node for running various preventative watchdogs (discussed later in this chapter), possibly offloading the checkpoints and doing cleanup jobs.

Queue up multiple training jobs

The next crucial step is to ensure that if the training crashed, there is a new job lined up to take place of the previous one.

Therefore, when a training is started, instead of using:

```
sbatch train.slurm
```

You'd want to replace that with:

```
sbatch --array=1-10%1 train.slurm
```

This tells SLURM to book a job array of 10 jobs, and if one of the job completes normally or it crashes, it'll immediately schedule the next one.

footnote: `%1` in `--array=1-10%1` tells SLURM to launch the job array serially - one job at a time.

If you have already started a training without this provision, it's easy to fix without aborting the current job by using the `--dependency` argument:

```
sbatch --array=1-10%1 --dependency=CURRENTLY_RUNNING_JOB_ID train.slurm
```

So if your launched job looked like this:

```
$ squeue -u `whoami` -o "%.10i %9P %20j %.8T %.10M %.81 %.6D %.20S %R"
  JOBID PARTITION NAME          STATE      TIME  TIME_LIM   NODES START_TIME
NODELIST(REASON)
  87      prod    my-training-10b  RUNNING 2-15:52:19 1-16:00:00    64  2023-10-07T01:26:28
node-[1-63]
```

You will note that the current's `JOBID=87` and now you can use it in:

```
sbatch --array=1-10%1 --dependency=87 train.slurm
```

and then the new status will appear as:

```
$ squeue -u `whoami` -o "%.10i %9P %20j %.8T %.10M %.81 %.6D %.20S %R"
  JOBID PARTITION NAME          STATE      TIME  TIME_LIM   NODES START_TIME
NODELIST(REASON)
  87      prod    my-training-10b  RUNNING 2-15:52:19 1-16:00:00    64  2023-10-07T01:26:28
node-[1-63]
  88_[10%1]  prod    my-training-10b  PENDING      0:00 1-16:00:00    64          N/A
(_dependency)
```

So you can see that an array of 10 jobs (`88_[10%1]`) was appended to be started immediately after the current job (87) completes or fails.

Granted that if the condition that lead to the crash is still there the subsequent job will fail as well. For example, if the storage device is full, no amount of restarts will allow the training to proceed. And we will discuss shortly how to avoid this situation.

But since the main reason for training crashes is failing GPUs, ensuring that faulty nodes are automatically removed and the new job starts with a new set of nodes makes for a smooth recovery from the crash.

In the SLURM lingo, the removed nodes are given a new status called `drained`. Here is an example of a hypothetical SLURM cluster:

```
$ sinfo
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
prod*      up    infinite      4  drain node-[0-3]
prod*      up    infinite     47  alloc node-[4-51]
```

```
prod*      up    infinite     23    idle node-[52-73]
```

Here we have 47 nodes being used (`alloc`), 23 available (`idle`) and 4 unavailable (`drained`).

The sysadmin is expected to periodically check the drained nodes, fix or replace them and then make them again available to be used by changing their state to `idle`.

The other approach is to daisy-chain jobs via `--dependency` as explained [here](#). Both of these approaches could also be combined.

How do you know when the job array or a daisy chain should not resume - well, normally the training loop will exit immediately if it knows the job is done. But you could also add features like `kill switch` which are even easier to use to prevent a job array from running.

Preferring fixed accelerator allocations to dynamic ones

Typically when getting a new set of accelerator nodes, especially when it's a new type of an accelerator that came out recently, many accelerators will fail, making LLM training quite problematic. There can be as large as 10% failure rate early on for new accelerators and still pretty high percentage of failures at later stages. Remember that if you have 8 accelerators, even one failing accelerator is like all 8 failing from the perspective of the training program.

If you use a fixed node allocation, after a few months, the bad accelerators will be weeded out and there should be very few accelerators failing. It'll still happen but it'll be a rare event.

Make sure your provider gives you new accelerators when they fail and doesn't return you the same accelerators after having them cool off (literally). For example, see how to track [the NVIDIA GPUs UUID](#). Those transient failures are likely to repeat when under heavy stress, so you want those to be replaced for real.

If you use a dynamic allocation, even a year after a new accelerator type has been released, expect lots of failing accelerators, since you'd be getting rejected nodes by other users. Surely, some clouds are better than others at diligently replacing bad hardware, the problem is that there are many accelerators that don't fail outright and when someone dropped a bad node, the technician looking at it may not see any problems with it when they try it out. And if the user just released the node without reporting it was broken, if the cloud provider isn't re-checking that a node is kosher before giving it to the next user the probability of getting a bad node is extremely high.

Frequent checkpoint saving

Whenever the training job fails, many hours of training can be lost. This problem is mitigated by a frequent checkpoint saving. When the training is resumed it'll continue from the last checkpoint saved. If the failure occurred 12 hours after the last checkpoint has been saved, 12 hours of training is lost and needs to be re-done. This can be very expensive if the training uses hundreds of GPUs.

In theory one could save a checkpoint every 10 minutes and only ever lose 10 minutes of training time, but this too would dramatically delay the reaching of the finish line because large models can't be saved quickly and if the saving time starts to create a bottleneck for the training this approach becomes counterproductive.

Depending on your checkpointing methodology and the speed of your IO storage partition the saving of a large model can take from dozens of seconds to several minutes. Therefore, the optimal approach to saving frequency lies somewhere in the middle.

The math is quite simple - measure the amount of time it takes to save the checkpoint, multiply it by how many times you'd want to save it and see how much of an additional delay the checkpoint saving will contribute to the total training time.

Use case: While training BLOOM-176B we had an incredibly fast GPFS over NVME filesystem and it took only 40 seconds to

save a 2.3TB checkpoint written concurrently on 384 processes. We saved a checkpoint approximately every 3 hours. As we trained for about 3 months, that means that we saved about 720 checkpoints ($90 \text{ days} * 24 \text{h} / 3\text{h}$) - that is an additional 8 hours was spent just saving the checkpoints ($720 \text{ times} * 40 \text{ secs} / 3600 \text{ secs}$) - or $\sim 0.37\%$ of the total training time ($8\text{h} / (90 \text{ days} * 24 \text{ hours})$). Now say if the IO were to be 5 times slower, which is not uncommon on the cloud unless one pays for premium IO, that would have become 2% of the training time, which would be quite significant.

footnote: If you don't have a large local storage and you have to offload the checkpoints to the cloud, make sure that the 2 most frequent checkpoints remain local to allow for a quick resume. The reason for 2 and not 1, is that it's possible that the very last checkpoint got corrupted or didn't finish saving if a crash occurred during its saving.

While this method introduces an overhead to the training, having training checkpoints is a hugely useful. Because these allow you to rollback many steps back should there be a divergence, are useful for analysis of various events and many trainings these day switch from in-training single loss measuring eval, which provide little useful signal to a full blown dataset-based evaluation on multiple benchmarks applied to each checkpoint during training. The latter can be done on additional nodes w/o slowing down the training for in-training evals.

Mutli-Replica-based fault tolerance

There is another approach to dealing with accelerator crashes which involves no checkpoint saving. This approach only works in situations where at least two model replicas are used during training.

Please review the various [model parallelism](#) techniques first to be able to follow along.

- If some variation of 3D model parallelism is used, that is you have either Tensor Parallelism (TP) and/or Pipeline Parallelism (PP) and/or Data Parallelism (DP), the number of replicas is equal to the DP degree.
- If Hybrid ZeRO-DP parallelism is used, then the number of replicas is equal to the degree of hybrid replicas.

For example, say you have a training setup that uses a 3D parallelism of TP=4, PP=2, DP=2 - so then you have 2 replicas, each using 8 gpus `node0` and `node1` (TP=4, PP=2 $\Rightarrow 4*2=8$) - practically, each replica uses a whole 8-GPU node.

Additionally you have a standby back up node `node2` with 8 GPUs idling but ready to be used at a moment's notice.

Now, say, during training `node0.gpu0` fails. Since you have a 2nd replica with intact data, you switch over to the standby 8GPU node, RDMA copy the data from the gpus of the 2nd replica and you can continue training where you left off. This is a very simplistic explanation since there are multiple nuances to figuring out such recovery depending at which stage of the iteration loop the failure occurred. In other words there is a complex algorithm to implement.

Of course, on a large scale training you're likely to have a hundred active nodes and a small handful of back up node.

This approach is superior to file system checkpointing saving because, you only ever lose one iteration, whereas with file system checkpointing this will hundreds of iterations lost.

I'm not aware of any open source implementations of this advanced fault tolerance method, but we know some of the big companies use this approach internally.

Kill switch

In many SLURM environments users have no sudo access and when one user started a training and went to sleep, and then a problem has been discovered, the other users can't easily stop the training and restart it again.

This was the situation during BLOOM-176B training and we implemented a kill-switch to handle that. The mechanism is very simple. The training loop polls for a specific file to appear before starting a new iteration and if the file is there the program saves the checkpoint and exits, allowing users other than the one who started the previous training to change things and restart it again. An additional poll was added at the very beginning of `main` so that if there was a long job array queued by the user who is asleep they could be "burned through" quickly by getting each job exit quickly on start.

This is also discussed [here](#).

This facility helps to minimize the amount of wasted training time.

Save switch

While mentioning the kill switch, it might be good to quickly mention its cousin, a save switch. Similarly to the kill switch the save switch is a variation of the former where instead of stopping the training, if the training loop discovers that a save-switch file appears - it will save a checkpoint, but will continue training. It'll also automatically remove the save-switch from the file-system, so that it won't accidentally start saving a checkpoint after every iteration.

This feature can be very useful for those who watch the training charts. If one sees an interesting or critical situation in the training loss or some other training metric one can quickly ask the training program to save the checkpoint of interest and be able to later reproduce the current situation at will.

The main use of this feature is around observing training loss spikes and divergences.

(note-to-self: better belongs to instabilities chapter)

Prevention

The easiest way to avoid losing training time is to prevent certain types of problems from happening. While one can't prevent a GPU from failing, other than ensuring that adequate cooling is provided, one can certainly ensure that there is enough of disk space remaining for the next few days of training. This is typically done by running scheduled watchdogs that monitor various resources and alert the operator of possible problems long before they occur.

Scheduled Watchdogs

Before we discuss the various watchdogs it's critical that you have a mechanism that allows you to run scheduled jobs. In the Unix world this is implemented by the [crontab facility](#).

Here is an example of how `~/bin/watch-fs.sh` can be launched every hour:

```
0 * * * * ~/bin/watch-fs.sh
```

The link above explains how to configure a crontab job to run at various other frequencies.

To setup a crontab, execute `crontab -e` and check which jobs are scheduled `crontab -l`.

The reason I don't go into many details is because many SLURM environments don't provide access to the `crontab` facility. And therefore one needs to use other approaches to scheduling jobs.

The section on [Crontab Emulation](#) discusses how to implement crontab-like SLURM emulation and also [Self-perpetuating SLURM jobs](#).

Notification facility

Then you need to have one or more notification facilities.

The simplest one is to use email to send alerts. To make this one work you need to ensure that you have a way to send an email from the SLURM job. If it isn't already available you can request this feature from your sysadmin or alternatively you might be able to use an external SMTP server provider.

In addition to email you could probably also setup other notifications, such as SMS alerting and/or if you use Slack to send slack-notifications to a channel of your choice.

Once you understand how to schedule watchdogs and you have a notification facility working let's next discuss the critical

watchdogs.

Is-job-running watchdog

The most obvious watchdog is one which checks that there is a training SLURM job running or more are scheduled to run.

Here is an example [slurm-status.py](#) that was used during BLOOM-176B training. This watchdog was sending an email if a job was detected to be neither running nor scheduled and it was also piping its check results into the main training's log file. As we used [Crontab Emulation](#), we simply needed to drop [slurm-status.slurm](#) into the `crontab/cron.hourly/` folder and the previously launched SLURM crontab emulating scheduler would launch this check approximately once an hour.

The key part of the SLURM job is:

```
tools/slurm-status.py --job-name $WATCH_SLURM_NAME 2>&1 | tee -a $MAIN_LOG_FILE
```

which tells the script which job name to watch for, and you can also see that it logs into a log file.

For example, if you launched the script with:

```
tools/slurm-status.py --job-name my-training-10b
```

and the current status report shows:

```
$ squeue -u `whoami` -o "%10i %9P %20j %8T %.10M %.81 %.6D %.20S %R"
  JOBID      PARTITION NAME          STATE      TIME      TIME_LIM      NODES      START_TIME
NODELIST(REASON)
  87      prod      my-training-10b  RUNNING  2-15:52:19  1-16:00:00   64  2023-10-07T01:26:28 node-[1-63]
```

then all is good. But if `my-training-10b` job doesn't show the alert will be sent.

You can now adapt these scripts to your needs with minimal changes of editing the path and email addresses. And if it wasn't you who launched the job then replace `whoami` with the name of the user who launched it. `whoami` only works if it was you who launched it.

Is-job-hanging watchdog

If the application is doing `torch.distributed` or alike and a hanging occurs during one of the collectives, it'll eventually timeout and throw an exception, which would restart the training and one could send an alert that the job got restarted.

However, if the hanging happens during another syscall which may have no timeout, e.g. reading from the disk, the application could easily hang there for hours and nobody will be the wiser.

Most applications do periodic logging, e.g., most training log the stats of the last N steps every few minutes. Then one could check if the log file has been updated during the expected time-frame - and if it didn't - send an alert. You could write your own, or use [io-watchdog](#) for that.

Low disk space alerts

The next biggest issue is running out of disk space. If your checkpoints are large and are saved frequently and aren't offloaded elsewhere it's easy to quickly run out of disk space. Moreover, typically multiple team members share the same cluster and it could be that your colleagues could quickly consume a lot of disk space. Ideally, you'd have a storage

partition that is dedicated to your training only, but often this is difficult to accomplish. In either case you need to know when disk space is low and space making action is to be performed.

Now what should be the threshold at which the alerts are triggered. They need to be made not too soon as users will start ignoring these alerts if you start sending those at say, 50% usage. But also the percentage isn't always applicable, because if you have a huge disk space shared with others, 5% of that disk space could translate to many TBs of free disk space. But on a small partition even 25% might be just a few TBs. Therefore really you should know how often you write your checkpoints and how many TBs of disk space you need daily and how much disk space is available.

Use case: During BLOOM training we wrote a 2.3TB checkpoint every 3 hours, therefore we were consuming 2.6TB a day!

Moreover, often there will be multiple partitions - faster IO partitions dedicated to checkpoint writing, and slower partitions dedicated to code and libraries, and possibly various other partitions that could be in use and all of those need to be monitored if their availability is required for the training not crashing.

Here is another caveat - when it comes to distributed file systems not all filesystems can reliably give you a 100% of disk space you acquired. In fact with some of those types you can only reliably use at most ~80% of the allocated storage space. The problem is that these systems use physical discs that they re-balance at the scheduled periods or triggered events, and thus any of these individual discs can reach 100% of their capacity and lead to a failed write, which would crash a training process, even though `df` would report only 80% space usage on the partition. We didn't have this problem while training BLOOM-176B, but we had it when we trained IDEFICS-80B - 80% there was the new 100%. How do you know if you have this issue - well, usually you discover it while you prepare for the training.

And this is not all. There is another issue of inodes availability and some storage partitions don't have very large inode quotas. Python packages are notorious for having hundreds to thousands of small files, which combined take very little total space, but which add up to tens of thousands of files in one's virtual environment and suddenly while one has TBs of free disk space available, but runs out of free inodes and discovering their training crashing.

Finally, many distributed partitions don't show you the disk usage stats in real time and could take up to 30min to update.

footnote: Use `df -ih` to see the inodes quota and the current usage.

footnote: Some filesystems use internal compression and so the reported disk usage can be less than reality if copied elsewhere, which can be confusing.

So here is [fs-watchdog.py](#) that was used during BLOOM-176B training. This watchdog was sending an email if any of the storage requirements thresholds hasn't been met and here is the corresponding [fs-watchdog.slurm](#) that was driving it.

If you study the watchdog code you can see that for each partition we were monitoring both the disk usage and inodes. We used special quota tools provided by the HPC to get instant stats for some partitions, but these tools didn't work for all partitions and there we had to fallback to using `df` and even a much slower `du`. As such it should be easy to adapt to your usecase.

Dealing with slow memory leaks

Some programs develop tiny memory leaks which can be very difficult to debug. Do not confuse those with the usage of MMAP where the program uses the CPU memory to read quickly read data from and where the memory usage could appear to grow over time, but this is not real as this memory gets freed when needed. You can read [A Deep Investigation into MMAP Not Leaking Memory](#) to understand why.

Of course, ideally one would analyze their software and fix the leak, but at times the leak could be coming from a 3rd party package or can be very difficult to diagnose and there isn't often the time to do that.

When it comes to GPU memory, there is the possible issue of memory fragmentation, where over time more and more tiny unused memory segments add up and make the GPU appear to have a good amount of free memory, but when the program tries to allocate a large tensor from this memory it fails with the OOM error like:

```
RuntimeError: CUDA out of memory. Tried to allocate 304.00 MiB (GPU 0; 8.00 GiB total capacity;
142.76 MiB already allocated; 6.32 GiB free; 158.00 MiB reserved in total by PyTorch)
```

In this example if there are 6.32GB free, how comes that 304MB couldn't be allocated.

One of the approaches my team developed during IDEFICS-80B training where there was some tiny CPU memory leak that would often take multiple days to lead to running out of CPU memory was to install a watchdog inside the training loop that would check the memory usage and if a threshold was reached it'd voluntarily exit the training loop. The next training job would then resume with all the CPU memory reclaimed.

footnote: The reality of machine learning trainings is that not all problems can be fixed with limited resources and often times a solid workaround provides for a quicker finish line, as compared to "stopping the presses" and potentially delaying the training for weeks, while trying to figure out where the problem is. For example we trained BLOOM-176B with `CUDA_LAUNCH_BLOCKING=1` because the training would hang without it and after multiple failed attempts to diagnose that we couldn't afford any more waiting and had to proceed as is. Luckily this environment variable that normally is used for debug purposes and which in theory should make some CUDA operations slower didn't actually make any difference to our throughput. But we have never figured out what the problem was and today it doesn't matter at all that we haven't, as we moved on with other projects which aren't impacted by that issue.

The idea is similar to the kill and save switches discussed earlier, but here instead of polling for a specific file appearance we simply watch how much resident memory is used. For example here is how you'd auto-exit if the OS shows only 5% of the virtual cpu memory remain:

```
import psutil
for batch in iterator:
    total_used_percent = psutil.virtual_memory().percent
    if total_used_percent > 0.95:
        print(f"Exiting early since the cpu memory is almost full: ({total_used_percent}%)")
        save_checkpoint()
        sys.exit()

    train_step(batch)
```

Similar heuristics could be used for setting a threshold for GPU memory usage, except one needs to be aware of cuda tensor caching and python garbage collection scheduling, so to get the actual memory usage you'd need to do first run the garbage collector then empty the cuda cache and only then you will get real memory usage stats and then gracefully exit the training if the GPU is too close to being full.

```
import gc
import torch

for batch in iterator:
    gc.collect()
    torch.cuda.empty_cache()

    # get mem usage in GBs and exit if less than 2GB of free GPU memory remain
    free, total = map(lambda x: x/2**30, torch.cuda.mem_get_info());
    if free < 2:
        print(f"Exiting early since the GPU memory is almost full: ({free}GB remain)")
```

```
save_checkpoint()  
sys.exit()  
  
train_step(batch)
```

footnote: don't do this unless you really have to, since caching makes things faster. Ideally figure out the fragmentation issue instead. For example, look up `max_split_size_mb` in the doc for [PYTORCH_CUDA_ALLOC_CONF](#) as it controls how memory is allocated. Some frameworks like [Deepspeed](#) solve this by pre-allocating tensors at start time and then reuse them again and again preventing the issue of fragmentation altogether.

footnote: this simplified example would work for a single node. For multiple nodes you'd need to gather the stats from all participating nodes and find the one that has the least amount of memory left and act upon that.

Dealing with forced job preemption

Earlier you have seen how the training can be gracefully stopped with a [kill switch solution](#) and it's useful when you need to stop or pause the training on demand.

On HPC clusters SLURM jobs have a maximum runtime. A typical one is 20 hours. This is because on HPCs resources are shared between multiple users/groups and so each is given a time slice to do compute and then the job is forcefully stopped, so that other jobs could use the shared resources.

footnote: this also means that you can't plan how long the training will take unless your jobs run with the highest priority on the cluster. If your priority is not the highest it's not uncommon to have to wait for hours and sometimes days before your job resumes.

One could, of course, let the job killed and hope that not many cycles were spent since [the last checkpoint was saved](#) and then let the job resume from this checkpoint, but that's quite wasteful and best avoided.

The efficient solution is to gracefully exit before the hard tile limit is hit and the job is killed by SLURM.

First, you need to figure out how much time your program needs to gracefully finish. This typically requires 2 durations:

1. how long does it take for a single iteration to finish if you have just started a new iteration
2. how long does it take to save the checkpoint

If, for example, the iteration takes 2 minutes at most and the checkpoint saving another 2 minutes, then you need at least 4 minutes of that grace time. To be safe I'd at least double it. There is no harm at exiting a bit earlier, as no resources are wasted.

So, for example, let's say your HPC allows 100 hour jobs, and then your slurm script will say:

```
#SBATCH --time=100:00:00
```

Approach A. Tell the program at launch time when it should start the exiting process:

```
srun ... torchrun ... --exit-duration-in-mins 5990
```

100h is 6000 minutes and so here we give the program 10 mins to gracefully exit.

And when you start the program you create a timer and then before every new iteration starts you check if the time limit is reached. If it is you save the checkpoint and exit.

case study: you can see how this was set [in the BLOOM training job](#) and then acted upon [here](#):

```
# Exiting based on duration
if args.exit_duration_in_mins:
    train_time = (time.time() - _TRAIN_START_TIME) / 60.0
    done_cuda = torch.cuda.IntTensor(
        [train_time > args.exit_duration_in_mins])
    torch.distributed.all_reduce(
        done_cuda, op=torch.distributed.ReduceOp.MAX)
    done = done_cuda.item()
    if done:
        if not saved_checkpoint:
            save_checkpoint_and_time(iteration, model, optimizer,
                                      lr_scheduler)
        print_datetime('exiting program after {} minutes'.format(train_time))
        sys.exit()
```

As you can see since the training is distributed we have to synchronize the exiting event across all ranks

You could also automate the derivation, by retrieving the `EndTime` for the running job:

```
$ scontrol show -d job $SLURM_JOB_ID | grep Time
RunTime=00:00:42 TimeLimit=00:11:00 TimeMin=N/A
SubmitTime=2023-10-26T15:18:01 EligibleTime=2023-10-26T15:18:01
AccrueTime=2023-10-26T15:18:01
StartTime=2023-10-26T15:18:01 EndTime=2023-10-26T15:18:43 Deadline=N/A
```

and then comparing with the current time in the program and instead setting the graceful exit period. There are other timestamps and durations that can be retrieved as it can be seen from the output.

Approach B.1. Sending a custom signal X minutes before the end

In your `sbatch` script you could set:

```
#SBATCH --signal=USR1@600
```

and then SLURM will send a `SIGUSR1` signal to your program 10min before job's end time.

footnote: normally SLURM schedulers send a `SIGCONT+SIGTERM` signal about 30-60 seconds before the job's time is up, and just as the time is up it will send a `SIGCONT+SIGTERM+SIGKILL` signal if the job is still running. `SIGTERM` can be caught and acted upon but 30 seconds is not enough time to gracefully exit a large model training program.

Let's demonstrate how the signal sending and trapping works. In terminal A, run:

```
python -c "
import time, os, signal

def sighandler(signum, frame):
```

```

print('Signal handler called with signal', signum)
exit(0)

signal.signal(signal.SIGUSR1, sighandler)
print(os.getpid())
time.sleep(1000)
"
```

it will print the pid of the process, e.g., 4034989 and will go to sleep (emulating real work). In terminal B now send `SIGUSR1` signal to the python program in terminal A with:

```
kill -s USR1 4034989
```

The program will trap this signal, call the `sighandler` which will now print that it was called and exit.

```
Signal handler called with signal 10
```

`10` is the numerical value of `SIGUSR1`.

So here is the same thing with the SLURM setup:

```

$ cat sigusr1.slurm
#SBATCH --job-name=sigusr1
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --time=0:03:00
#SBATCH --partition=partition
#SBATCH --output=%x-%j.out
#SBATCH --signal=USR1@170

srun python -c "
import time, os, signal

def sighandler(signum, frame):
    print('Signal handler called with signal', signum)
    exit(0)

signal.signal(signal.SIGUSR1, sighandler)
print(os.getpid())
time.sleep(1000)
"
```

In the SLURM script we told SLURM to send the program a signal 170 seconds before its end and the job itself was set to run for 180 secs (3 mins).

When this job has been scheduled:

```
sbatch sigusr1.slurm
```

10 seconds (180-170) after the job started, it will exit with the log:

```
58307
Signal handler called with signal 10
```

which means the job had a pid 58307 and it caught SIGUSR1 (10) and it exited.

Now that you understand how this machinery works, instead of immediate `exit(0)` you can set exit-asap flag, finish the currently run iteration, check that the flag is up, save the checkpoint and exit. This is very similar to the code shown in Approach A above.

Approach B.2. Choosing which process to send the signal to

Now what if your main program isn't the one launched with `srun` - if you were to use an intermediate launcher like `torchrun` or `accelerate` the above recipe won't work, because most likely `SIGUSR1` won't be propagated from the launcher to its children. In this case we need a slightly more complicated slurm script than

We have to replace:

```
#SBATCH --signal=USR1@600
```

with:

```
#SBATCH --signal=B:USR1@600
```

The added `B:` tells SLURM not to send the signal to the `srun` process (launcher) but to the `sbatch` shell.

And now we have to change the end of the SLURM script from a typical launcher-based code like:

```
CMD="python -u -m torch.distributed.run ... train.py ..." # real command here
LOG_FILE=/path/to/logs/main_log.txt
srun --jobid $SLURM_JOBID bash -c "$CMD" 2>&1 | tee -a $LOG_FILE
```

to this:

```
trap 'echo "SIGUSR1 received!"; \
pid=$(pgrep -f "^(python.*(accelerate|deepspeed|torchrun|distributed.run))"); \
pgrep -P $pid | xargs -r kill -USR1; \
wait;' SIGUSR1

CMD="python -u -m torch.distributed.run ... train.py ..." # real command here
LOG_FILE=/path/to/logs/main_log.txt
```

```

srun --jobid $SLURM_JOBID bash -c "$CMD" 2>&1 | tee -a $LOG_FILE &

wait

```

Since `--signal=B:USR1@600` earlier will now send the signal to the `sbatch` shell we can trap it and do something about it, and that's what the `trap` line does.

The magical code inside the signal handler passed to `trap` finds all processes that are immediate children of any of the launchers like `accelerate`, `deepspeed`, `torchrun` or `torch.distributed.run` and sends the `SIGUSR1` signal to them.

Finally the last change is that in order for `trap` to work we need to run `srun` in the background - so we added `&` at the end of the `srun` command and we needed to add `wait` so that the `sbatch` shell won't exit until `srun` finishes.

Your python code that catches the signal handler remains the same as in Approach B.1.

Here are the important parts of the SLURM script together:

```

$ cat launch.slurm
#!/bin/bash
[...]
#SBATCH --partition=dev
#SBATCH --signal=B:USR1 # Custom preemption signal
[...]

trap 'echo "SIGUSR1 received!"; \
pid=$(pgrep -f "^(python.*(accelerate|torchrun|deepspeed|distributed.run))"); \
pgrep -P $pid | xargs -r kill -USR1; wait;' SIGUSR1

CMD="python -u -m torch.distributed.run ... train.py ..." # real command here
LOG_FILE=/path/to/logs/main_log.txt
srun --jobid $SLURM_JOBID bash -c "$CMD" 2>&1 | tee -a $LOG_FILE &

wait

```

And your training loop that may have originally looked like this:

```

$ cat train.py

for batch in dl:
    train_iteration(batch)

```

Now it'll become:

```

$ cat train.py

import signal
import sys

```

```

pre_emption_activated = False
def activate_pre_emption(sig, frame):
    global pre_emption_activated
    print("SIGUSR1 received, saving checkpoint")
    pre_emption_activated = True

signal.signal(signal.SIGUSR1, activate_pre_emption)

for batch in dl:
    train_iteration(batch)

    if pre_emption_activated:
        save_checkpoint()
        sys.exit()

```

Of course, you will probably set a flag in the trainer object in the real software and not use a `global`, but for the sake of the short demo that's good enough.

If you want to test this solution, simply change your SLURM script header to:

```

#SBATCH --time=0:05:00
#SBATCH --signal=B:USR1@60

```

Here we tell SLURM to run the job for 5 minutes only (`--time=0:05:00`) and we ask it to send `SIGUSR1` to our `sbatch` script 60 seconds before 5 minutes expires, i.e. 4 minutes after the job started.

QoS-based SLURM preemption

We haven't discussed so far what happens when Quality of Service (QoS) is used, which may also forcefully preempt an existing job. The functionality is the same as job's-allocated-time-is-about-to-end sort of pre-emption, except it can happen any time and not X seconds before the end of the job.

Consider a SLURM setup where you have `--qos=high` which can preempt `--qos=low` jobs and the low priority job has grace time of 10 minutes to shut down:

```

$ sacctmgr show qos format=name,priority,preempt,MaxTRESPerUser,GraceTime,Preempt,Flags
  Name  Priority  MaxTRESPU  GraceTime  Preempt  Flags
  -----
  low      0        00:10:00
  high     0        00:00:00      low

```

This is very similar to the time-based pre-emption except here the grace time is hardcoded and can't be modified by the user.

If a job is launched with `--qos=high` and there aren't enough nodes, SLURM will kick out a few low priority jobs to make the nodes available for the high priority job.

By default `GraceTime` could be very short an insufficient for your program to wind down safely if it gets pre-empted - in which case ask your sysadmin to raise its duration to what will work for your needs.

Otherwise the same solutions described in Approaches B.1 and B.2 will work for this type of forced pre-emption.

Reproducibility

Achieve determinism in randomness based software

When debugging always set a fixed seed for all the used Random Number Generators (RNG) so that you get the same data / code path on each re-run.

Though with so many different systems it can be tricky to cover them all. Here is an attempt to cover a few:

```
import random, torch, numpy as np
def enforce_reproducibility(use_seed=None):
    seed = use_seed if use_seed is not None else random.randint(1, 1000000)
    print(f"Using seed: {seed}")

    random.seed(seed)      # python RNG
    np.random.seed(seed) # numpy RNG

    # pytorch RNGs
    torch.manual_seed(seed)          # cpu + cuda
    torch.cuda.manual_seed_all(seed) # multi-gpu - can be called without gpus
    if use_seed: # slower speed! https://pytorch.org/docs/stable/notes/randomness.html#cuda-convolution-benchmarking
        torch.backends.cudnn.deterministic = True
        torch.backends.cudnn.benchmark      = False

    return seed
```

a few possible others if you use those subsystems/frameworks instead:

```
torch.npu.manual_seed_all(seed)
torch.xpu.manual_seed_all(seed)
tf.random.set_seed(seed)
```

When you rerun the same code again and again to solve some problem set a specific seed at the beginning of your code with:

```
enforce_reproducibility(42)
```

But as it mentions above this is for debug only since it activates various torch flags that help with determinism but can slow things down so you don't want this in production.

However, you can call this instead to use in production:

```
enforce_reproducibility()
```

i.e. w/o the explicit seed. And then it'll pick a random seed and log it! So if something happens in production you can now reproduce the same RNGs the issue was observed in. And no performance penalty this time, as the `torch.backends.cudnn` flags are only set if you provided the seed explicitly. Say it logged:

```
Using seed: 1234
```

you then just need to change the code to:

```
enforce_reproducibility(1234)
```

and you will get the same RNGs setup.

As mentioned in the first paragraphs there could be many other RNGs involved in a system, for example, if you want the data to be fed in the same order for a `DataLoader` you need [to have its seed set as well](#).

Additional resources:

- [Reproducibility in pytorch](#)

Reproduce the software and system environment

This methodology is useful when discovering some discrepancy in outcomes - quality or a throughput for example.

The idea is to log the key components of the environment used to launch a training (or inference) so that if at a later stage it needs to be reproduced exactly as it was it can be done.

Since there is a huge variety of systems and components being used it's impossible to prescribe a way that will always work. So let's discuss one possible recipe and you can then adapt it to your particular environment.

This is added to your slurm launcher script (or whatever other way you use to launch the training) - this is Bash script:

```
SAVE_DIR=/tmp # edit to a real path
export REPRO_DIR=$SAVE_DIR/repro/$SLURM_JOB_ID
mkdir -p $REPRO_DIR
# 1. modules (writes to stderr)
module list 2> $REPRO_DIR/modules.txt
# 2. env
/usr/bin/printenv | sort > $REPRO_DIR/env.txt
# 3. pip (this includes devel installs SHA)
pip freeze > $REPRO_DIR/requirements.txt
# 4. uncommitted diff in git clones installed into conda
perl -nle 'm|"file://(.*/([^\/]+))"| && qx[cd $1; if [ ! -z "\$(git diff)" ]; then git diff >
\${REPRO_DIR}/\$2.diff; fi]' $CONDA_PREFIX/lib/python*/site-packages/*.dist-info/direct_url.json
```

As you can see this recipe is used in a SLURM environment, so every new training will dump the environment specific to the SLURM job.

1. We save which `modules` were loaded, e.g. in cloud cluster/HPC setups you're like to be loading the CUDA and cuDNN libraries using this

If you don't use `modules` then remove that entry

2. We dump the environment variables. This can be crucial since a single env var like `LD_PRELOAD` or `LD_LIBRARY_PATH` could make a huge impact on performance in some environments
3. We then dump the conda environment packages and their versions - this should work with any virtual python environment.
4. If you use a devel install with `pip install -e .` it doesn't know anything about the git clone repository it was installed from other than its git SHA. But the issue is that it's likely that you have modified the files locally and now `pip freeze` will miss those changes. So this part will go through all packages that are not installed into the conda environment (we find them by looking inside `site-packages/*.dist-info/direct_url.json`)

An additionally useful tool is [`conda-env-compare.pl`](#) which helps you to find out the exact differences 2 conda environments have.

Anecdotally, me and my colleague were getting very different training TFLOPs on a cloud cluster running the exact same code - literally launching the same slurm script from the same shared directory. We first compared our conda environments using [`conda-env-compare.pl`](#) and found some differences - I installed the exact packages she had to match her environment and it was still showing a huge performance difference. We then compared the output of `printenv` and discovered that I had `LD_PRELOAD` set up and she didn't - and that made a huge difference since this particular cloud provider required multiple env vars to be set to custom paths to get the most of their hardware.

Avoiding, Recovering From and Understanding Instabilities

Sub-sections:

- [Understanding Training Loss Patterns](#) - types of spikes, divergences, grokking moments, resumes, etc.

Learning from Training Logbooks

The best learning is to read [Publicly available training LLM/VLM logbooks](#) because there you can see exactly what happened and how the problem has been overcome.

STD Init

Correctly initializing the initial distribution of the tensors can have a tremendous impact on training's stability. The `std` value isn't fixed and depends on the hidden dimension size.

This proved to be a very crucial setting in our pre-BLOOM 104B experiments and we couldn't break past the first few thousands iterations until we figured out that the 0.02 default `--init-method-std` in Megatron-LM was a way too big for our model.

We referred to these two sources:

1. "Transformers without Tears" paper <https://arxiv.org/abs/1910.05895> prescribes: `sqrt(2/(NHIDDEN*5))`
2. The 530B training paper <https://arxiv.org/abs/2201.11990> they used an even smaller init formula:
`sqrt(1/(NHIDDEN*3))`

and decided to go with the 530B one as it leads to an even smaller init value.

To make it easier to compare the two formulas, they can be rewritten as:

1. `sqrt(0.4000/NHIDDEN)`
2. `sqrt(0.3333/NHIDDEN)`

Thus for `NHIDDEN=14336` the math was `sqrt(1/(14336*3)) = 0.00482` and that's what we used. It surely wasn't the only reason why we had no stability issues during BLOOM-176B training, but I think it was one of the crucial ones.

Numerical instabilities

Certain mathematical operations could be unstable when dealing with low precision numbers.

For example, please see this very interesting [PyTorch guide on numerical stability](#).

Now let's look at a specific example of this concept in action.

During 104B training experiments where fp16 mixed precision was used - the following improvement was proposed by [Corby Rosset](#) to make [self-attention more stable](#).

Specifically this [line](#) shows that the `norm_factor` may be multiplied after the Query * Key matrix multiplication. If the dim of Q and K are very large, the output may blow up and the `norm_factor` won't be able to save it.

Proposal: move the `norm_factor` inward, so Q and K are scaled down before matrix multiply:

```

matmul_result = torch.baddbmm(
    matmul_result,
    1.0/math.sqrt(self.norm_factor) * query_layer.transpose(0, 1),  # [b * np, sq, hn]
    1.0/math.sqrt(self.norm_factor) * key_layer.transpose(0, 1).transpose(1, 2),  # [b * np, hn,
sk]
    beta=0.0 if alibi is None else 1.0, alpha=1.0)

# change view to [b, np, sq, sk]
attention_scores = matmul_result.view(*output_size)

```

To make the operation mathematically equivalent, moving the norm factor inward requires taking `sqrt` again if `n` is a scalar, `A` and `B` matrices:

```
n * (A dot B) === (sqrt(n) * A) dot (sqrt(n) * B)
```

Now `A` and `B` dimensions can be significantly larger.

For CUDA kernel writers [CuBlas](#)'s `GemmStridedBatchedEx` at the time of this writing has a similar issue. It is defined as:

```
C+i*strideC=aop(A+i*strideA)op(B+i*strideB)+β(C+i*strideC), for i ∈ [0,batchCount-1]
```

The issue is that `alpha` is multiplied after the matrix-matrix multiplication is done so it can cause instability.

"Bad" combination of data batch and model parameter state

PaLM team observed dozens of loss spikes at "highly irregular intervals" when training larger models. While they were not able to track down the root cause, they mitigated the issue by restarting from an earlier checkpoint and skipping potentially problematic data batches. [Section 5.1 Training instability](#)

Time-domain correlation divergence in Adam

[A Theory on Adam Instability in Large-Scale Machine Learning](#) performs a rigorous study of divergence spikes while training LLMs at up to 546B parameters - and suggests that the time-domain correlation leads to divergence of Adam. This is triggered by the `epsilon` value not being small enough and gradient estimation components become similar to the `epsilon`.

In section 7.1 they propose practical suggestions, the most interesting one of them is setting `epsilon` to 0 and possibly dealing with division by zero condition.

Understanding Training Loss Patterns

Training loss plot is similar to the heart beat pattern - there is the good, the bad and you-should-worry one. After studying many training loss trajectories one develops an intuition to explain various loss behaviors during one's training and how to act on those.

I warn you that the "Understanding" in the title of this section is overloaded since very often we don't really understand why certain types of spikes happen. Here "understanding" refers to recognizing various patterns. We then usually have techniques to overcome the bad patterns and bring the training successfully to the finish line.

Thus you will find here a gallery of training loss patterns sometimes with real explanations, but more often than not educated guesses to what might be happening.

Please excuse the plot snapshots looking wildly different from each other as they have come from many sources over multiple years.

The good, the bad and the unexpected

Let's look at some good, bad and unusual patterns.

A very failed training

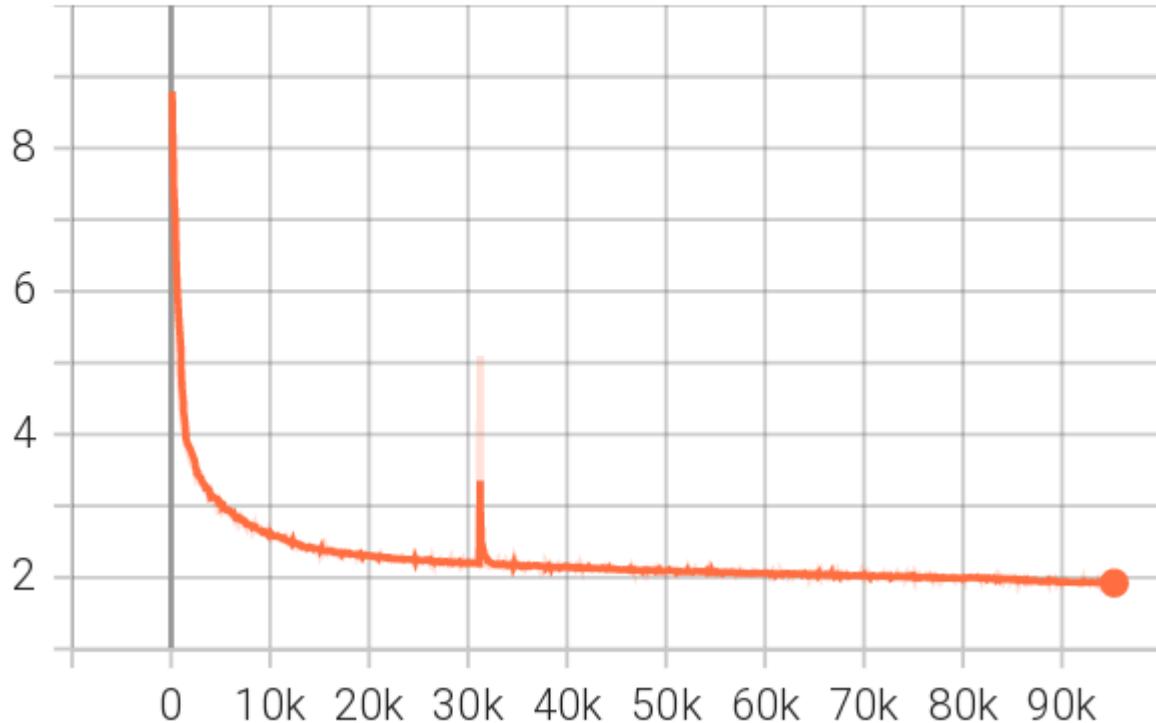
Prior to starting BLOOM-176B training we did multiple experiments with the [104B model](#). We failed to figure out how to not diverge very early on.



As you can see many attempts were made, many techniques were applied (see [chronicles](#)). We think the 2 main obstacles were using fp16 and data that had a lot of garbage in it. For BLOOM-176B we switched to bf16, used much cleaner data and also added an embedding layer-norm and that made all the difference.

An almost perfect training

lm-loss-training/lm loss tag: lm-loss-training/lm loss



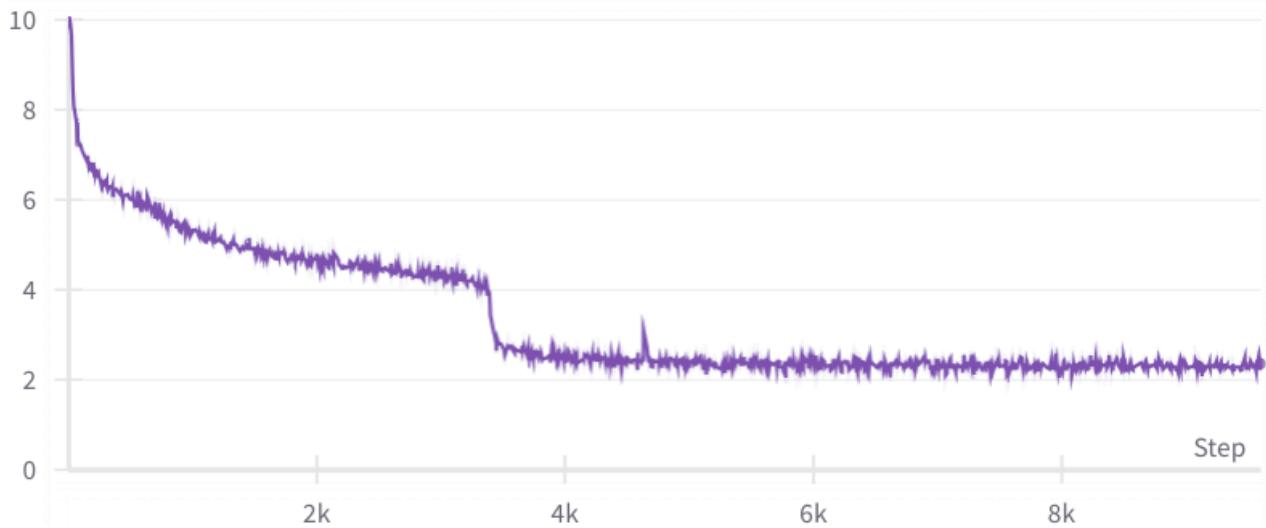
The [BLOOM-176B](#) training had a close to perfect training loss trajectory, with a single spike that has recovered in 200 steps.

You can inspect the [TB](#) to zoom in and check other plots.

This was the almost perfect training indeed. Lots of hard work was put into achieving this.

The grokking moment

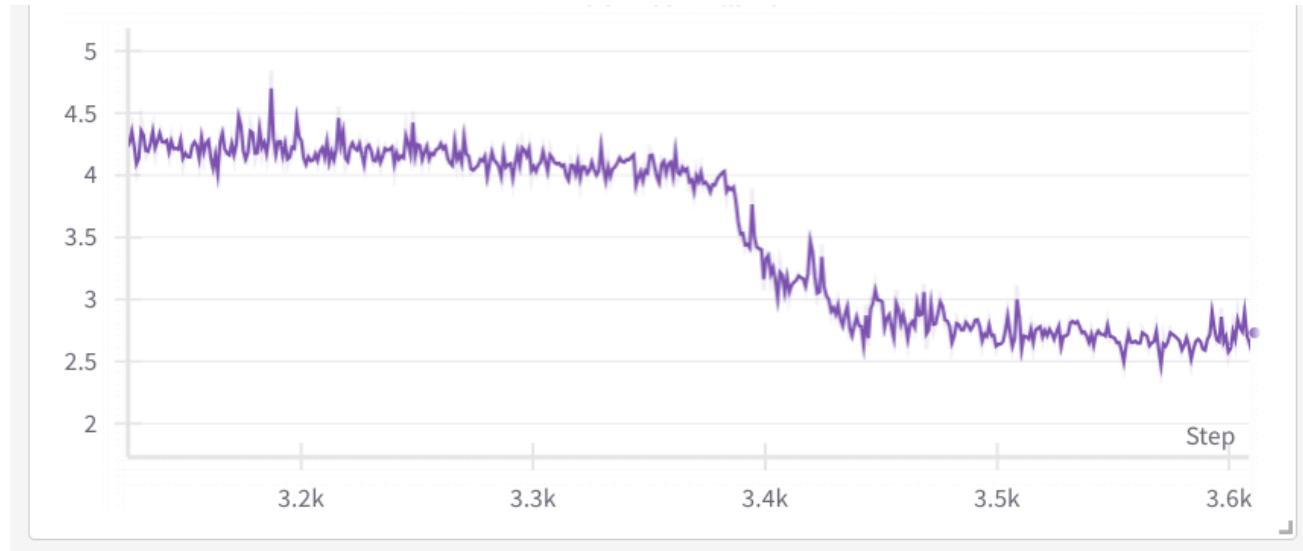
Recently I was doing some performance testing and run a tiny global batch size of 8 on 8x A100 nodes on llama-2-7b trained from scratch. (w/ Deepspeed ZeRO-3 DP using HF Transformers [Llama](#) implementation)



Here one can observe a rapid loss improvement from 4 to 2.5 in just 480 samples after a very steady much slower improvements. My colleague [Gautam Mittal](#) called it the [grokking](#) moment. In just a handful of steps the model suddenly generalized to much better predict the masked tokens.

Normally one doesn't see such a dramatic improvement when using a much larger batch size.

If we zoom in it took about 60 8-sample per iteration steps:



Main types of loss spikes

In general there are 3 types of loss spikes:

1. Fast recovering spikes
2. Slow recovering spikes
3. Not fully recovering spikes

The spikes usually happen because of a bad data pocket, either due to badly shuffled data or because it hasn't been cleaned from some garbage scraped from the websites.

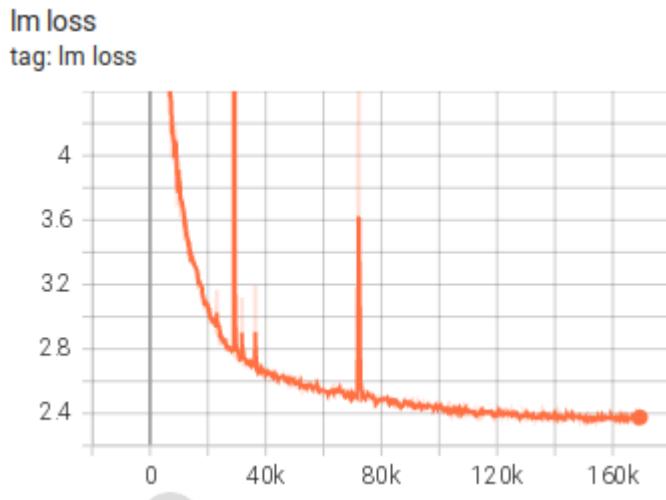
While one would suspect that the batch before the spike was the trigger, but if you were to study that batch's contents

you are likely to find nothing unusual - quite often the problem starts developing many steps before and then most of the sudden it happens. But also it might not be easy to study the batch, since it could amount to a size of a book when the global batch size and the sequence lengths are huge.

Fast recovering spikes

Loss spikes can happen often and as long as they quickly bounce back to where they left off the training usually continues as if nothing happened:

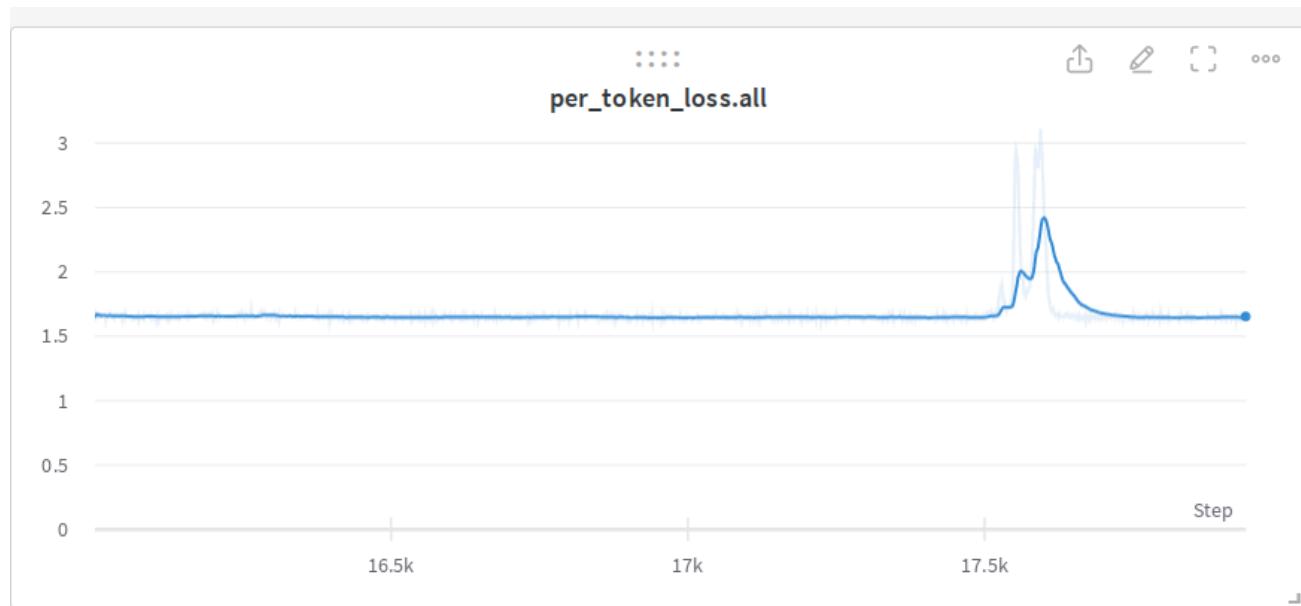
Here is an example of [the 13B pre-BLOOM training experiment](#):



As you can see there are many spikes, some of a huge magnitude but they have all quickly recovered.

Slow recovering spikes

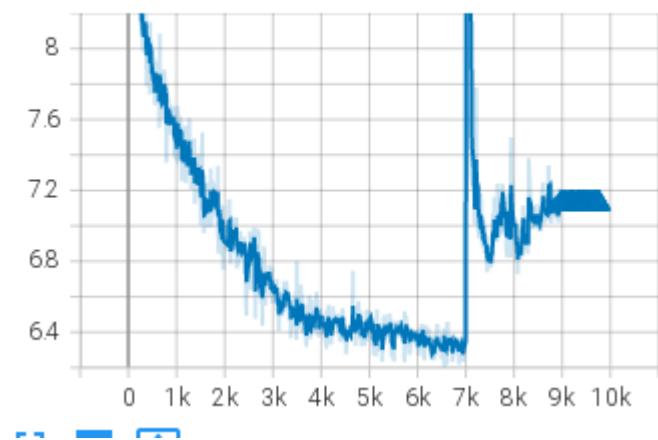
Here is a slow recovering spike from the [IDEFICS-80B](#) training:



Not fully recovering spikes

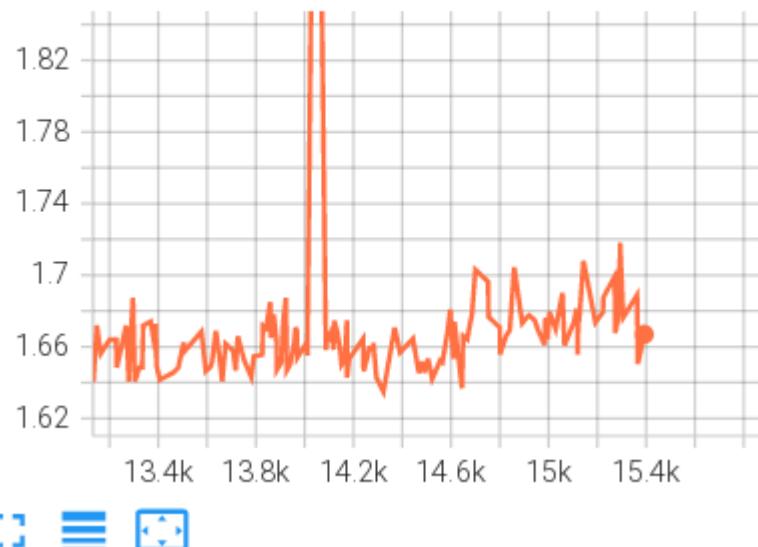
This [104B model attempt](#) spiked, started recovering but decided to not recover fully and instead started diverging

lm loss
tag: lm loss



Here is another example from the [IDEFICS-80B](#) training:

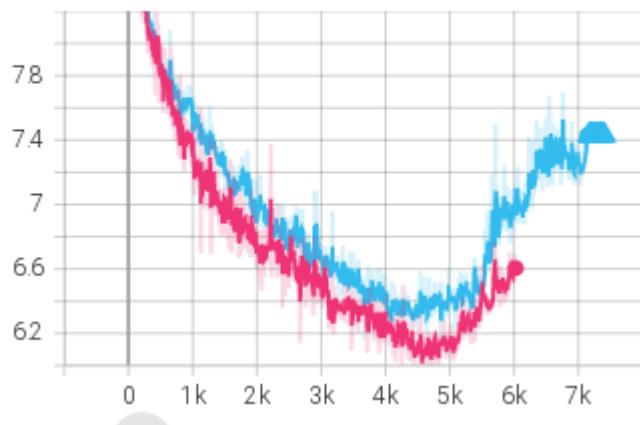
per_token_loss
tag: per_token_loss



Non-spike diverging

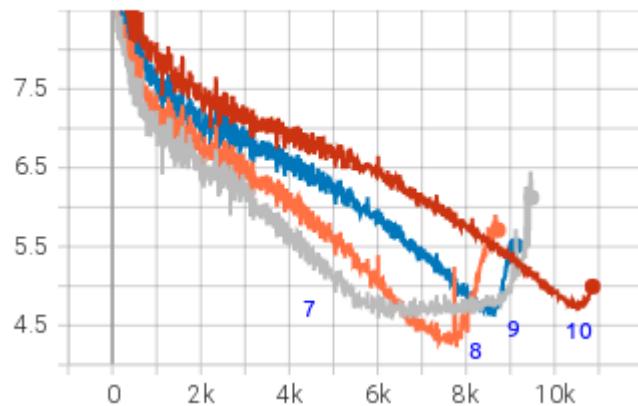
Here are a few examples of diverging that didn't go through a spike

lm loss
tag: lm loss



and here are a few more:

lm loss
tag: lm loss

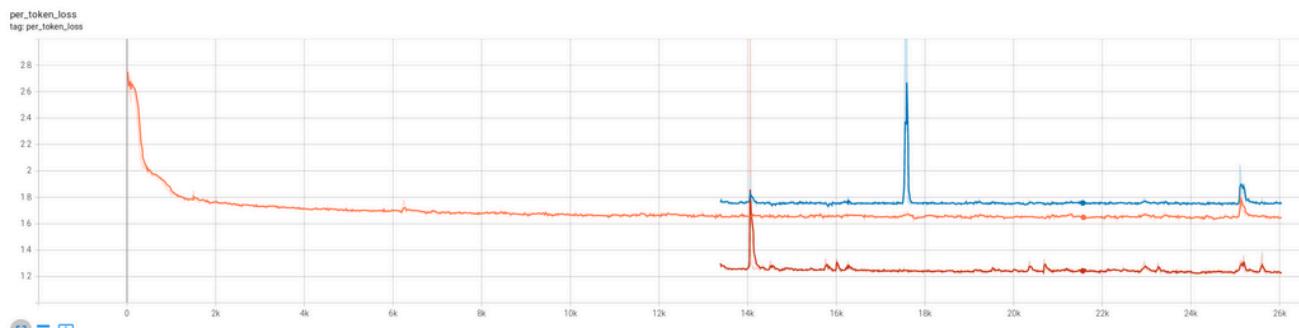


as you can see each restart makes a bit of progress and then the model diverges.

All these are from the [104B model attempts](#).

Multiple datasets spikes

During the [IDEFICS-80B](#) training we were using 2 different dataset types mixed together:



Legend: cm4 (high), average (mid) and pmid (low)

You can see that the loss spikes were sometimes happening simultaneously on both datasets and at other times only one

of the datasets loss would spike.

Here the model was learning two different data distributions and as you can see it was not reporting the same loss and the spike behaviors on both data distributions. The pmd datasets loss was much easier for the model than the cm4 one.

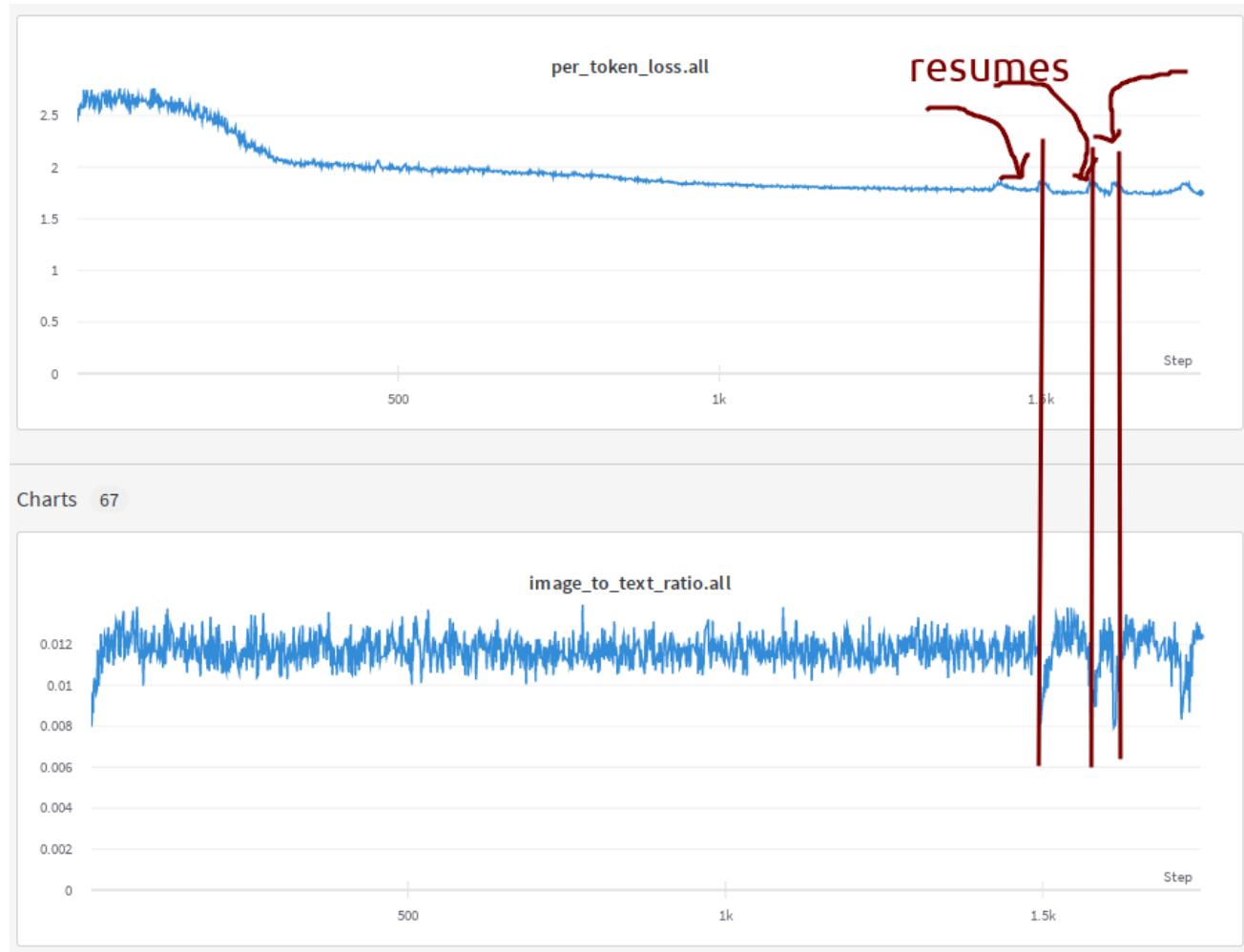
Resume-related spikes

Training resume due to a hardware crash or because a need to rollback to an earlier checkpoint due to encountering a divergence is pretty much guaranteed to happen. If your training software doesn't resume perfectly so that the model doesn't notice there was a resume various problems could be encountered.

The most complicated challenge of resume is restoring various RNGs, getting to the DataLoader index where the previous training was restored, and dealing with various other requirements if you use complex DataLoaders that are specific to your setup.

DataSampler related issues

During [IDEFICS-80B](#) training we had a very complicated DataLoader which was suffering from image to text ratio fluctuations when the DataLoader was getting restored on resume, so we ended up having a small spike on each resume which would then recover:



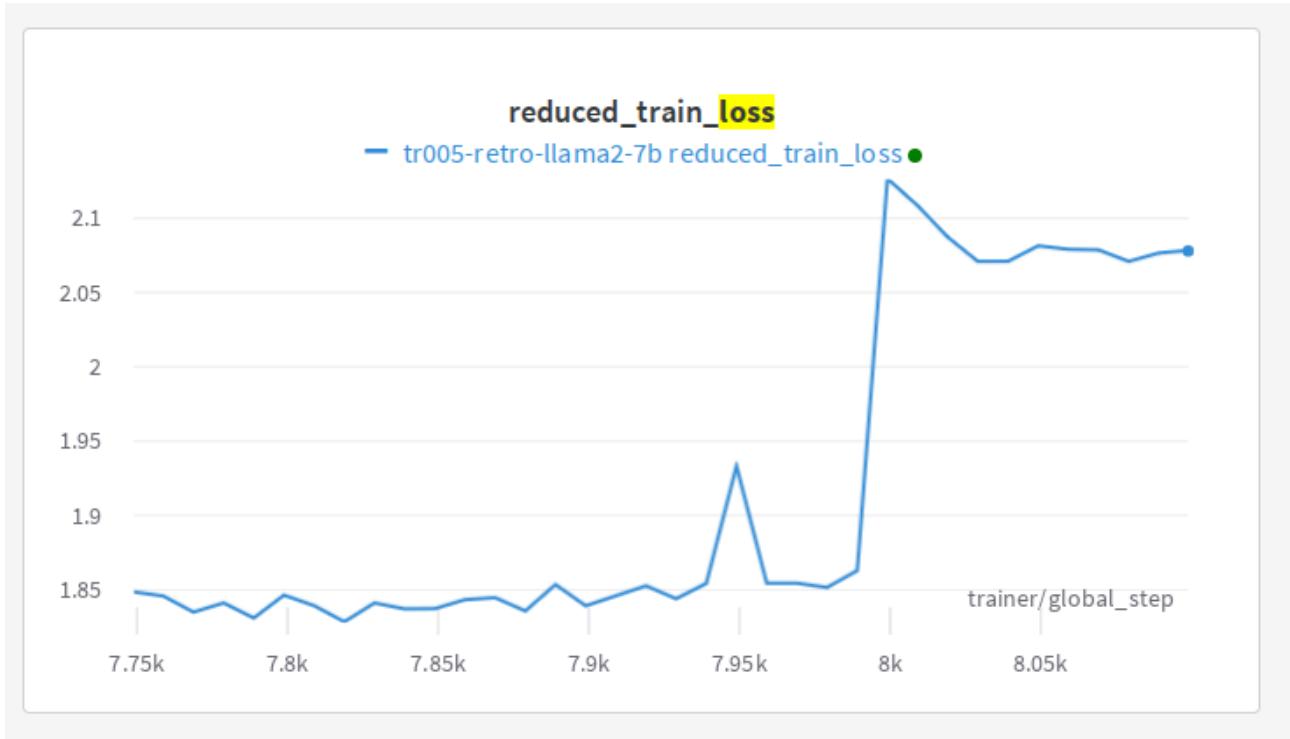
You can see the loss and ratio plots correlation here. As we had to resume about a dozen times we saw a lot of those spikes.

Impacts of repeat data

I was training a variation of Llama2 and saw this super unusual spike that didn't diverge or recover but which switched to a new higher loss level:

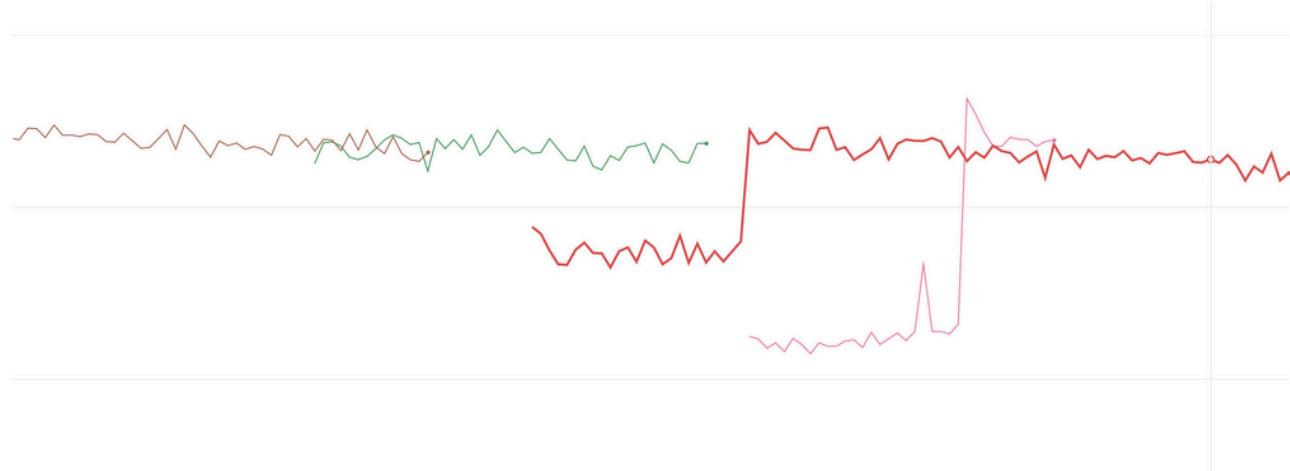


I rolled back to just before the weird behavior occurred and restarted. The loss training progressed at the same loss level for a bit and then again spiked and shifted to a higher loss.



I have never seen this type of divergence before. I was scratching my head for a while and then decided to look at the bigger picture.

As of this writing [Wandb](#) doesn't handle resume data plotting correctly if a rollback was performed, that is it ignores all new data after the rollback until the steps of the old data have been overcome. This forces us to start a new wandb plot for every resume with a rollback so that new data is shown. And if you need to see the whole plot you have to stitch them and which includes dead data points that are no longer true. So I did the stitching and saw this puzzle:



There was no real spike in the two earlier runs. The loss never went up in the first place. In both resumes it was under-reporting loss due to an exactly repeated data and then it reached data it hasn't seen before and started reporting correctly. In other words it was overfitting and reporting a false loss.

The cause of the problem is data repetition, and since it clearly memorised some of it it was reporting a better loss.

The problem comes from [pytorch-lightning](#) not handling resumes correctly wrt DataSampler automatically - basically every time you resume you start your data stream from scratch. This, of course, requires a user to somehow fix the situation. You could change the seed to somewhat ameliorate the situation and avoid the exact data sequence, but it

still leaves you with repeat data, which isn't what you want for any serious training (or ablation experiments, since your observation will be invalid, if they assume [IID data distribution](#)).

footnote: I discussed [this issue with the PTL developers](#) and they said that they tried hard to come up with a generic solution but it wasn't meant to be. So the user needs to figure it out.

Make sure to check your training framework documentation whether it handles the DataSampler resuming correctly. Make sure you didn't discover this problem after the training has finished and you ended up training 6x times the same 50B of tokens from the planned 300B tokens seen only once each.

Doing a couple of resumes early on before embarking on the real training should also expose if there is a problem. Albeit, if the data gets reshuffled on each resume you are unlikely to see it. It'll only be seen if the seed is the same.

Checkpoints

- [torch-checkpoint-convert-to-bf16](#) - converts an existing fp32 torch checkpoint to bf16. If `safetensors` are found those are converted as well. Should be easily adaptable to other similar use cases.
- [torch-checkpoint-shrink.py](#) - this script fixes checkpoints which for some reason stored tensors with storage larger than their view at the moment of saving. It clones the current view and re-saves them with just the storage of the current view.

Selecting Training Hyper-Parameters And Model Initializations

The easiest way to find a good hparam and model init starter set is to steal it from a similar training that you know has succeeded. Here is a [collection of public training LLM/VLM logbooks](#) to get you started. The other common source is papers if they disclose that information. You can also try to reach out to the authors and ask them for these details if they didn't publish it.

Glossary

Training jargon uses a multitude of abbreviations and terms, so here are some important for this chapter.

- BS: Batch Size - here we mean batch size per gpu, often it is also referred to as MBS (micro-batch-size)
- GBS: Global Batch Size - total batch size per iteration - may include gradient accumulation
- GAS: Gradient Accumulation Steps - how many forward/backward cycles to perform before one full iteration is complete
- TFLOPs: Trillion FLOPs per second - [FLOPS](#)
- PP: Pipeline Parallelism

Global Batch Size Ramp Up

If you intend to train with a very large GBS, with say 1024, or 2048 samples and even higher, when you just start training, it's very wasteful to feed such large batch sizes to the model. At this point it's totally random and can't benefit from having too refined data. Therefore to save data and resources, one often ramps up the global batch size over some period of time.

It's also important to not start with GBS that is too small, since otherwise the progress won't be efficient. When there is too little data the compute (TFLOPS) is inefficient and will slow everything down. This is especially so when Pipeline Parallelism (PP) is used, since the most important thing about PP tuneup is a small GPU idleness bubble, and the smaller the GBS the larger the bubble is.

For example, for BLOOM-176B, where we did use PP, after doing throughput benchmarking we found that starting with GBS=16 was incredibly slow (8 TFLOPS), so we eventually started with GBS=192 (73 TFLOPS) and then we ramped up to GBS=2048 (150 TFLOPS) - we increased GBS by 16 every 9_765_625 samples.

STD Init

This hyper parameter is super-important and it requires math to get it right. For details see [STD Init](#).

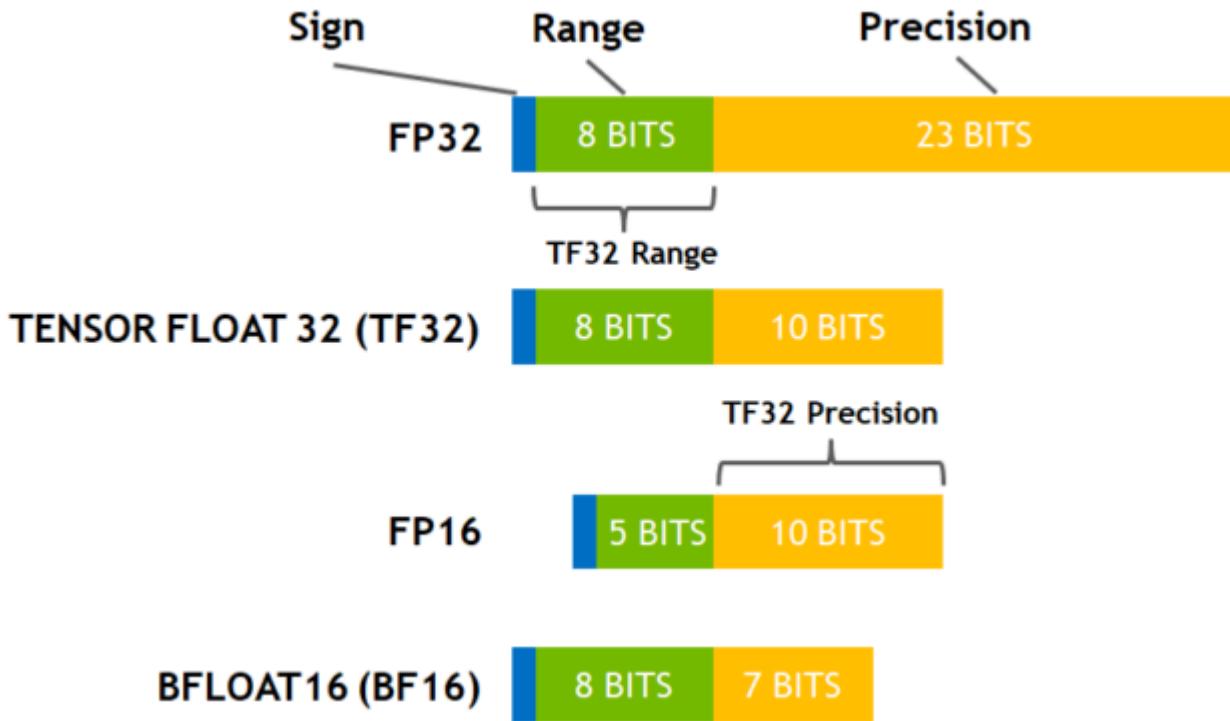
Tensor precision / Data types

These are the common datatypes that are used as of this writing in ML (usually referred to as `dtype`):

Floating point formats:

- fp32 - 32 bits
- tf32 - 19 bits (NVIDIA Ampere+)
- fp16 - 16 bits
- bf16 - 16 bits
- fp8 - 8 bits (E4M3 and E5M2 formats)
- fp6 - 6 bits
- fp4 - 4 bits

For visual comparison refer to this representations:



([source](#))

	sign	exponent							mantissa									
FP16	0	0	1	1	0	1	1	0	0	1	0	1	0	0	1	1	1	= 0.395264
BF16	0	0	1	1	1	1	1	0	1	1	0	0	1	0	1	1	0	= 0.394531
FP8 E4M3	0	0	1	0	1	1	0	1										
FP8 E5M2	0	0	1	1	0	1	1	0										

([source](#))

The new formats that are being adopted by new hardware are:

- fp4: float4_e2m1fn
- fp6: float6_e2m3fn and float6_e3m2fn
- fp8: float8_e3m4, float8_e4m3, float8_e4m3b11fnuz, float8_e4m3fn, float8_e4m3fnuz, float8_e5m2, float8_e5m2fnuz, float8_e8m0fnuz

There is an excellent explanation of each of these variations [here](#).

To decipher the letters followed by the numbers:

- The e indicates the length of exponent
- The m indicates the length of mantissa
- The b indicates the bias

To decipher the letters appearing after the numbers:

- The f indicates it is finite values only (no infinities).
- The n indicates it includes NaNs, but only at the outer range.
- The u stands for unsigned format.
- The uz stands for unsigned zero.

So for example: float8_e4m3b11fnuz stands for fp8 + 4-bit exponent + 3-bit mantissa + bias 11 + finite values only + includes NaNs, but only at the outer range + unsigned zero.

Integer formats used in quantization:

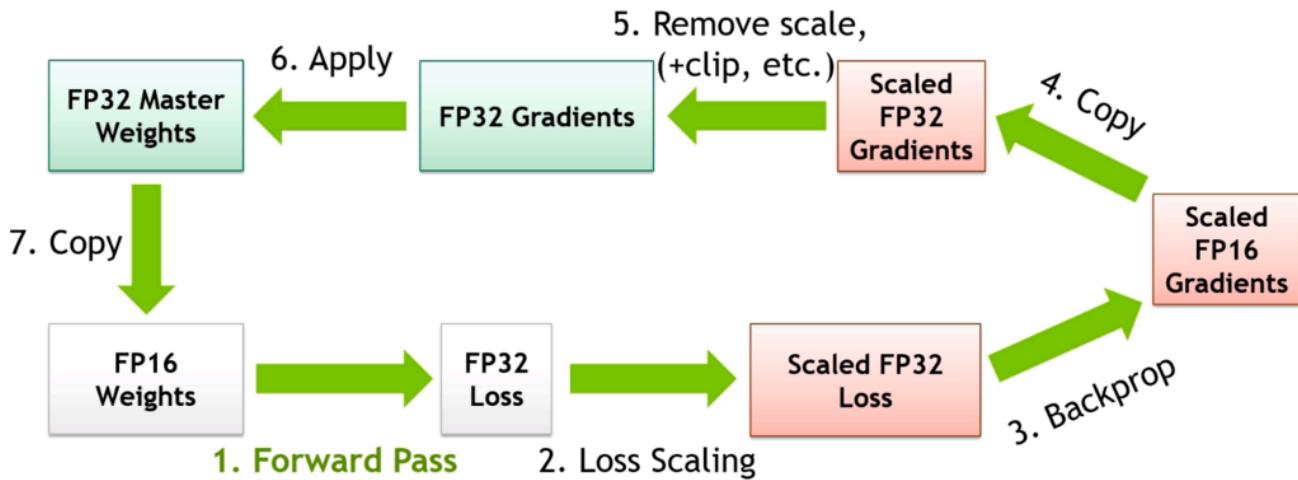
- int8 - 8 bits
- int4 - 4 bits
- int1 - 1 bits

ML dtype progression

Originally ML was using fp32, but it was very slow.

Next [mixed-precision was invented using a combination of fp16 and fp32](#) was invented which tremendously sped up the training speed.

MIXED PRECISION TRAINING



([source](#))

But fp16 proved to be not very stable and training LLM was extremely difficult.

Luckily bf16 came out and replaced fp16 using the same mixed precision protocol. This made the LLM training much more stable.

Then fp8 came and mixed precision has switched to [that](#) and which makes the training even faster. See the paper: [FP8 Formats for Deep Learning](#).

To appreciate the speed ups between the different formats have a look at this table for NVIDIA A100 TFLOPS spec (w/o sparsity):

Data type	TFLOPS
FP32	19.5
Tensor Float 32 (TF32)	156
BFLOAT16 Tensor Core	312
FP16 Tensor Core	312
FP8 Tensor Core	624
INT8 Tensor Core	624

Each next dtype is about 2x faster than the previous one (except fp32 which is much slower than the rest).

In parallel with the mixed training regime the ML community starting coming up with various quantization approaches. Probably one of the best examples is Tim Dettmers' [bitsandbytes](#) which provides many 4 and 8-bit quantization solutions.

The Deepspeed team also has some [interesting quantization solutions](#).

TF32

TF32 is a magical datatype that is available on NVIDIA GPUs since Ampere, and which allows fp32 `matmuls` performed at a much faster speed than normal fp32 `matmuls` with a small precision loss.

Here is an example of A100 TFLOPS (w/o sparsity):

Data type	TFLOPS
FP32	19.5
Tensor Float 32 (TF32)	156

As you can see TF32 is 8x faster than FP32!

It's disabled by default. To enable it add at the beginning of your program:

```
torch.backends.cuda.matmul.allow_tf32 = True
torch.backends.cudnn.allow_tf32 = True
```

For more information about the actual precision loss please see [this](#).

When to use fp32 accumulators

Whenever a low-precision dtype is used one has to be careful not to accumulate intermediary results in that dtype.

LayerNorm-like operations must not do their work in half-precision, or they may lose a lot of data. Therefore when these operations are implemented correctly they do efficient internal work in the dtype of the inputs, but using the fp32 accumulation registers and then their outputs are downcast to the precision of the inputs.

Generally it's just the accumulation that is done in fp32, since adding up many low-precision numbers is very lossy otherwise.

Here are some examples:

1. Reduction collectives

- fp16: ok to do in fp16 if loss scaling is in place
- bf16: only ok in fp32

2. Gradient accumulation

- best done in fp32 for fp16 and bf16, but definitely is a must for bf16

3. Optimizer step / Vanishing gradients

- when adding a tiny gradient to a large number, that addition is often nullified therefore typically fp32 master weights and fp32 optim states are used.
- f16 master weights and optim states can be used when using [Kahan Summation](#) or [Stochastic rounding](#) (introduced in [Revisiting BFloat16 Training](#)).

For an example of the latter see: [AnyPrecision optimizer](#) with the latest version found [here](#).

Changing precision post training

Sometimes it's OK to change precision after the model was trained.

- Using bf16-pretrained model in fp16 regime usually fails - due to overflows (the biggest number that can be represented in fp16 is 64k) for an indepth discussion and possible workaround see this [PR](#).
- Using fp16-pretrained model in bf16 regime usually works - it will lose some performance on conversion, but should work - best to finetune a bit before using it.

Emulate a multi-node setup using just a single node

The goal is to emulate a 2-node environment using a single node with 2 GPUs (for testing purposes). This, of course, can be further expanded to [larger set ups](#).

We use the `deepspeed` launcher here. There is no need to actually use any of the `deepspeed` code, it's just easier to use its more advanced capabilities. You will just need to install `pip install deepspeed`.

The full setup instructions follow:

1. Create a `hostfile`:

```
$ cat hostfile
worker-0 slots=1
worker-1 slots=1
```

2. Add a matching config to your ssh client

```
$ cat ~/.ssh/config
[...]

Host worker-0
  HostName localhost
  Port 22
Host worker-1
  HostName localhost
  Port 22
```

Adapt the port if it's not 22 and the hostname if `localhost` isn't it.

3. As your local setup is probably password protected ensure to add your public key to `~/.ssh/authorized_keys`

The `deepspeed` launcher explicitly uses no-password connection, e.g. on `worker0` it'd run: `ssh -o PasswordAuthentication=no worker-0 hostname`, so you can always debug ssh setup using:

```
$ ssh -vvv -o PasswordAuthentication=no worker-0 hostname
```

4. Create a test script to check both GPUs are used.

```
$ cat test1.py
import os
import time
import torch
import deepspeed
import torch.distributed as dist
```

```

# critical hack to use the 2nd gpu (otherwise both processes will use gpu0)
if os.environ["RANK"] == "1":
    os.environ["CUDA_VISIBLE_DEVICES"] = "1"

dist.init_process_group("nccl")
local_rank = int(os.environ.get("LOCAL_RANK"))
print(f'{dist.get_rank()=}, {local_rank=}')

x = torch.ones(2**30, device=f"cuda:{local_rank}")
time.sleep(100)

```

Run:

```

$ deepspeed -H hostfile test1.py
[2022-09-08 12:02:15,192] [INFO] [runner.py:415:main] Using IP address of 192.168.0.17 for node worker-0
[2022-09-08 12:02:15,192] [INFO] [multinode_runner.py:65:get_cmd] Running on the following workers:
worker-0,worker-1
[2022-09-08 12:02:15,192] [INFO] [runner.py:504:main] cmd = pdsh -S -f 1024 -w worker-0,worker-1 export
PYTHONPATH=/mnt/nvme0/code/huggingface/multi-node-emulate-ds; cd /mnt/nvme0/code/huggingface/
multi-node-emulate-ds; /home/stas/anaconda3/envs/py38-pt112/bin/python -u -m deepspeed.launcher.launch
--world_info=eyJ3b3JrZXItMCI6IFswXSwgIndvcmtlcioxIjogWzBdfQ== --node_rank=%n --master_addr=192.168.0.17
--master_port=29500 test1.py
worker-0: [2022-09-08 12:02:16,517] [INFO] [launch.py:136:main] WORLD INFO DICT: {'worker-0': [0],
'worker-1': [0]}
worker-0: [2022-09-08 12:02:16,517] [INFO] [launch.py:142:main] nnodes=2, num_local_procs=1, node_rank=0
worker-0: [2022-09-08 12:02:16,517] [INFO] [launch.py:155:main] global_rank_mapping=defaultdict(<class
'list'>, {'worker-0': [0], 'worker-1': [1]})
worker-0: [2022-09-08 12:02:16,517] [INFO] [launch.py:156:main] dist_world_size=2
worker-0: [2022-09-08 12:02:16,517] [INFO] [launch.py:158:main] Setting CUDA_VISIBLE_DEVICES=0
worker-1: [2022-09-08 12:02:16,518] [INFO] [launch.py:136:main] WORLD INFO DICT: {'worker-0': [0],
'worker-1': [0]}
worker-1: [2022-09-08 12:02:16,518] [INFO] [launch.py:142:main] nnodes=2, num_local_procs=1, node_rank=1
worker-1: [2022-09-08 12:02:16,518] [INFO] [launch.py:155:main] global_rank_mapping=defaultdict(<class
'list'>, {'worker-0': [0], 'worker-1': [1]})
worker-1: [2022-09-08 12:02:16,518] [INFO] [launch.py:156:main] dist_world_size=2
worker-1: [2022-09-08 12:02:16,518] [INFO] [launch.py:158:main] Setting CUDA_VISIBLE_DEVICES=0
worker-1: torch.distributed.get_rank()=1, local_rank=0
worker-0: torch.distributed.get_rank()=0, local_rank=0
worker-1: tensor([1., 1., 1., ..., 1., 1., 1.], device='cuda:0')
worker-0: tensor([1., 1., 1., ..., 1., 1., 1.], device='cuda:0')

```

If the ssh set up works you can run `nvidia-smi` in parallel and observe that both GPUs allocated ~4GB of memory from `torch.ones` call.

Note that the script hacks in `CUDA_VISIBLE_DEVICES` to tell the 2nd process to use gpu1, but it'll be seen as `local_rank==0` in both cases.

5. Finally, let's test that NCCL collectives work as well

Script adapted from [torch-distributed-gpu-test.py](#) to just tweak `os.environ["CUDA_VISIBLE_DEVICES"]`

```
$ cat test2.py
import deepspeed
import fcntl
import os
import socket
import time
import torch
import torch.distributed as dist

# a critical hack to use the 2nd GPU by the 2nd process (otherwise both processes will use gpu0)
if os.environ["RANK"] == "1":
    os.environ["CUDA_VISIBLE_DEVICES"] = "1"

def printflock(*msgs):
    """ solves multi-process interleaved print problem """
    with open(__file__, "r") as fh:
        fcntl.flock(fh, fcntl.LOCK_EX)
    try:
        print(*msgs)
    finally:
        fcntl.flock(fh, fcntl.LOCK_UN)

local_rank = int(os.environ["LOCAL_RANK"])
torch.cuda.set_device(local_rank)
device = torch.device("cuda", local_rank)
hostname = socket.gethostname()

gpu = f"[{hostname}-{local_rank}]"

try:
    # test distributed
    dist.init_process_group("nccl")
    dist.all_reduce(torch.ones(1).to(device), op=dist.ReduceOp.SUM)
    dist.barrier()
    print(f'{dist.get_rank()=}, {local_rank=}')

    # test cuda is available and can allocate memory
    torch.cuda.is_available()
    torch.ones(1).cuda(local_rank)

    # global rank
    rank = dist.get_rank()
    world_size = dist.get_world_size()

    printflock(f'{gpu} is OK (global rank: {rank}/{world_size})')

    dist.barrier()
    if rank == 0:
```

```

printflock(f"pt={torch.__version__}, cuda={torch.version.cuda}, nccl={torch.cuda.nccl.version()}")
printflock(f"device compute capabilities={torch.cuda.get_device_capability()}")
printflock(f"pytorch compute capabilities={torch.cuda.get_arch_list()}")


except Exception:
    printflock(f"{gpu} is broken")
    raise

```

Run:

```

$ deepspeed -H hostfile test2.py
[2022-09-08 12:07:09,336] [INFO] [runner.py:415:main] Using IP address of 192.168.0.17 for node worker-0
[2022-09-08 12:07:09,337] [INFO] [multinode_runner.py:65:get_cmd] Running on the following workers:
worker-0,worker-1
[2022-09-08 12:07:09,337] [INFO] [runner.py:504:main] cmd = pdsh -S -f 1024 -w worker-0,worker-1 export
PYTHONPATH=/mnt/nvme0/code/huggingface/multi-node-emulate-ds; cd /mnt/nvme0/code/huggingface/
multi-node-emulate-ds; /home/stas/anaconda3/envs/py38-pt112/bin/python -u -m deepspeed.launcher.launch
--world_info=eyJ3b3JrZXItMCI6IFswXSwgIndvcmtlci0xIjogWzBdfQ== --node_rank=%n --master_addr=192.168.0.17
--master_port=29500 test2.py
worker-0: [2022-09-08 12:07:10,635] [INFO] [launch.py:136:main] WORLD INFO DICT: {'worker-0': [0],
'worker-1': [0]}
worker-0: [2022-09-08 12:07:10,635] [INFO] [launch.py:142:main] nnodes=2, num_local_procs=1, node_rank=0
worker-0: [2022-09-08 12:07:10,635] [INFO] [launch.py:155:main] global_rank_mapping=defaultdict(<class
'list'>, {'worker-0': [0], 'worker-1': [1]})
worker-0: [2022-09-08 12:07:10,635] [INFO] [launch.py:156:main] dist_world_size=2
worker-0: [2022-09-08 12:07:10,635] [INFO] [launch.py:158:main] Setting CUDA_VISIBLE_DEVICES=0
worker-1: [2022-09-08 12:07:10,635] [INFO] [launch.py:136:main] WORLD INFO DICT: {'worker-0': [0],
'worker-1': [0]}
worker-1: [2022-09-08 12:07:10,635] [INFO] [launch.py:142:main] nnodes=2, num_local_procs=1, node_rank=1
worker-1: [2022-09-08 12:07:10,635] [INFO] [launch.py:155:main] global_rank_mapping=defaultdict(<class
'list'>, {'worker-0': [0], 'worker-1': [1]})
worker-1: [2022-09-08 12:07:10,635] [INFO] [launch.py:156:main] dist_world_size=2
worker-1: [2022-09-08 12:07:10,635] [INFO] [launch.py:158:main] Setting CUDA_VISIBLE_DEVICES=0
worker-0: dist.get_rank()=0, local_rank=0
worker-1: dist.get_rank()=1, local_rank=0
worker-0: [hope-0] is OK (global rank: 0/2)
worker-1: [hope-0] is OK (global rank: 1/2)
worker-0: pt=1.12.1+cu116, cuda=11.6, nccl=(2, 10, 3)
worker-0: device compute capabilities=(8, 0)
worker-0: pytorch compute capabilities=['sm_37', 'sm_50', 'sm_60', 'sm_70', 'sm_75', 'sm_80', 'sm_86']
worker-1: [2022-09-08 12:07:13,642] [INFO] [launch.py:318:main] Process 576485 exits successfully.
worker-0: [2022-09-08 12:07:13,642] [INFO] [launch.py:318:main] Process 576484 exits successfully.

```

Voila, mission accomplished.

We tested that the NCCL collectives work, but they use local NVLink/PCIe and not the IB/ETH connections like in real multi-node, so it may or may not be good enough for testing depending on what needs to be tested.

Larger set ups

Now, let's say you have 4 GPUs and you want to emulate 2x2 nodes. Then simply change the `hostfile` to be:

```
$ cat hostfile
worker-0 slots=2
worker-1 slots=2
```

and the `CUDA_VISIBLE_DEVICES` hack to:

```
if os.environ["RANK"] in ["2", "3"]:
    os.environ["CUDA_VISIBLE_DEVICES"] = "2,3"
```

Everything else should be the same.

Automating the process

If you want an automatic approach to handle any shape of topology, you could use something like this:

```
def set_cuda_visible_devices():
    """
    automatically assign the correct groups of gpus for each emulated node by tweaking the
    CUDA_VISIBLE_DEVICES env var
    """

    global_rank = int(os.environ["RANK"])
    world_size = int(os.environ["WORLD_SIZE"])
    emulated_node_size = int(os.environ["LOCAL_SIZE"])
    emulated_node_rank = int(global_rank // emulated_node_size)
    gpus = list(map(str, range(world_size)))
    emulated_node_gpus =
    ",".join(gpus[emulated_node_rank*emulated_node_size:(emulated_node_rank+1)*emulated_node_size])
    print(f"Setting CUDA_VISIBLE_DEVICES={emulated_node_gpus}")
    os.environ["CUDA_VISIBLE_DEVICES"] = emulated_node_gpus

set_cuda_visible_devices()
```

Emulating multiple GPUs with a single GPU

The following is an orthogonal need to the one discussed in this document, but it's related so I thought it'd be useful to share some insights here:

With NVIDIA A100 you can use [MIG](#) to emulate up to 7 instances of GPUs on just one real GPU, but alas you can't use those instances for anything but standalone use - e.g. you can't do DDP or any NCCL comms over those GPUs. I hoped I could use my A100 to emulate 7 instances and add one more real GPU and to have 8x GPUs to do development with - but nope it doesn't work. Asking NVIDIA engineers about it, there are no plans to have this use-case supported.

Acknowledgements

Many thanks to [Jeff Rasley](#) for helping me to set this up.

Re-train HF Hub Models From Scratch Using Finetuning Examples

HF Transformers has awesome finetuning examples <https://github.com/huggingface/transformers/tree/main/examples/pytorch>, that cover pretty much any modality and these examples work out of box.

But what if you wanted to re-train from scratch rather than finetune.

Here is a simple hack to accomplish that.

We will use `facebook/opt-1.3b` and we will plan to use `bf16` training regime as an example here:

```
cat << EOT > prep-bf16.py
from transformers import AutoConfig, AutoModel, AutoTokenizer
import torch

mname = "facebook/opt-1.3b"

config = AutoConfig.from_pretrained(mname)
model = AutoModel.from_config(config, torch_dtype=torch.bfloat16)
tokenizer = AutoTokenizer.from_pretrained(mname)

path = "opt-1.3b-bf16"

model.save_pretrained(path)
tokenizer.save_pretrained(path)
EOT
```

now run:

```
python prep-bf16.py
```

This will create a folder: `opt-1.3b-bf16` with everything you need to train the model from scratch. In other words you have a pretrained-like model, except it only had its initializations done and none of the training yet.

Adjust to script above to use `torch.float16` or `torch.float32` if that's what you plan to use instead.

Now you can proceed with finetuning this saved model as normal:

```
python -m torch.distributed.run \
--nproc_per_node=1 --nnode=1 --node_rank=0 \
--master_addr=127.0.0.1 --master_port=9901 \
examples/pytorch/language-modeling/run_clm.py --bf16 \
--seed 42 --model_name_or_path opt-1.3b-bf16 \
--dataset_name wikitext --dataset_config_name wikitext-103-raw-v1 \
--per_device_train_batch_size 12 --per_device_eval_batch_size 12 \
```

```
--gradient_accumulation_steps 1 --do_train --do_eval --logging_steps 10 \
--save_steps 1000 --eval_steps 100 --weight_decay 0.1 --num_train_epochs 1 \
--adam_beta1 0.9 --adam_beta2 0.95 --learning_rate 0.0002 --lr_scheduler_type \
linear --warmup_steps 500 --report_to tensorboard --output_dir save_dir
```

The key entry being:

```
--model_name_or_path opt-1.3b-bf16
```

where `opt-1.3b-bf16` is your local directory you have just generated in the previous step.

Sometimes it's possible to find the same dataset that the original model was trained on, sometimes you have to use an alternative dataset.

The rest of the hyper-parameters can often be found in the paper or documentation that came with the model.

To summarize, this recipe allows you to use finetuning examples to re-train whatever model you can find on [the HF hub](#).

Inference

XXX: this chapter is under construction - some sections are complete, some are still starting out, many are yet to be started, but there are already enough of useful sections completed to make it a good reading.

Glossary

- CLA: Cross-Layer Attention
- FHE: Fully Homomorphic Encryption
- GQA: Grouped-Query Attention
- ITL: Inter-Token Latency
- KV: Key Value
- LPU: Language Processing Unit™
- MHA: Multi-Head Attention
- MLA: Multi-Latent Attention
- MPC: Secure Multi-Party Computation
- MQA: Multi-Query Attention
- PPML: Privacy-Preserving Machine Learning
- QPS: Queries Per Second
- TPOT: Time Per Output Token
- TTFT: Time to First Token

See [Concepts](#) for more glossary-like entries.

Concepts

Prefill and Decode

When doing inference there are 2 stages:

Prefill

Prefill: as all tokens of the prompt are known - process the full prompt length at once (similar to training) and cache the intermediate states (KV cache). This stage contributes very little latency as even a 1k prompt can be processed really fast, given enough memory.

Decode

Decode: new tokens generation happens, one new token at a time (regressive approach) based on all the previous tokens (the prompt and any new tokens generated so far). Thus this stage contributes the most to the generation's latency as unlike prefill, decoding can't be parallelized.

Online vs Offline inference

When you have users that send queries in real time - this is Online inference also known, as Deployment or Interactive inference. Examples: chatbot, search engines, general REST APIs. In this case one always runs an inference server and there could be various clients querying it.

When you have a file with hundreds or thousands of prompts that you need to run inference on - this is Offline inference. Examples: benchmark evaluation and synthetic data generation. In this case the inference server is often not needed and the inference is run directly in the same program that sends the query (client and server in one application).

The 2 main use cases are often optimized for different performance metrics - the online inference use case requires a very low TTFT and low latency, whereas the offline inference requires high throughput.

Grounding

It's the process of giving the pre-trained model additional information that wasn't available during its training. For example [input-grounded tasks](#) give the model a lot of additional information in the prompt. Non zero-shot prompts ground the model in examples altering the default model behavior. Prompt-engineering is all about grounding the model to behave in a certain way during inference.

Retrieval Augmented Generation (RAG) is one of the main techniques for grounding models as it supplies the inference process with additional data that is relevant to the prompt. And the intention is that the model will give more significance to that information than the massive compressed information it was trained on.

Fine-tuning to a different knowledge domain is another grounding approach, we update the model to be grounded in a new dataset that could be quite distinct from the original domain of data the foundational model has been trained on.

Grounding can be thought of providing a context. As anybody can attest it's easier to answer a question when one understands the context of the question. The same applies with model generation. The better the context, the more relevant the generated output is.

In a multi-modal use case an image or a video supplied with the text prompt can be that grounding or a context.

Tasks

Input-grounded tasks

Input-grounded tasks are those where the generated response is derived mainly from the prompt, i.e. the main source of knowledge is contained in the prompt. These include:

- Translation
- Summarization
- Document QA
- Multi-turn chat
- Code editing
- Speech recognition (audio transcription)

Batching

Processing the decoding stage one token at a time is extremely accelerator-inefficient. Batching multiple queries together improved the accelerator utilization and enables processing multiple requests at once.

The maximum possible batch size depends on how much memory is left after loading the model weights and filling the KV-cache with intermediate states.

Static batching

This is the naive straightforward batching where the first N queries are batched together - the problem here is that if many queries have finished generating they will have to wait for the longest to generate query to complete before they can be returned to the caller - greatly increasing the latency.

Continuous Batching or In-flight batching

Continuous Batching or In-flight batching is a process where the generation engine removes completed results as soon as they are done and replacing them with new queries, without waiting for the whole batch to complete. So that a sequence in position 0 in the batch could be generating its 10th token, while a sequence in position 1 in the batch could be just starting its first token generation, and position 3 is producing its last token.

This improves the response time, since there is no need for a sequence that already finished not to be returned immediately and there is no need for a new prompt to wait for the next batch to become available. Of course, if all of the compute is fully busy, and there are no new openings in the batch, then some requests will have to wait before the compute will start processing those.

Paged Attention

Paged Attention is very popular with inference servers as it allows for a very efficient accelerator memory utilization, by the virtue of approaching the accelerator memory like the OS memory using paging, which allowed dynamic memory allocation and prevents memory fragmentation.

Decoding methods

The main decoding methods are: [Greedy decoding](#), [Beam search](#) and [Sampling](#).

Greedy decoding

Greedy decoding is when the model always chooses the token with the highest probability. This is the fastest decoding method but it doesn't always generate the best outcome, since it may choose a less ideal token path and miss out on a great future sequence of tokens.

One of the main issues with this method is creation of loops, where the same sentence is repeated again and again.

Beam search

Beam search overcomes the limitation of greedy decoding by generating multiple outputs at the same time, so instead of following the highest probability - with beam size of 3 it follows top 3 probabilities at each new token, then discards all but the 3 sub-paths out of 9 ($3*3$), that lead to the highest total probability of all tokens in the chain. Then at the end the path with the highest probability of all tokens is chosen.

This method is slower than greedy decoding because it has to generate n times more tokens and it requires n times more memory.

Sampling

Sampling-based decoding introduces randomness.

But, of course, choosing random words will not lead to a good result, so we still want greedy decoding like certainty but making it more interesting/alive by adding controlled randomness to it.

The most common sampling methods are:

- **Top-K Sampling** method chooses the top k tokens based on their logit probability and then randomly picks one of those tokens.
- **Top-p Sampling** (also known as **nucleus sampling**) is like Top-K Sampling, but the K varies for each next token and is calculated by adding up the top token probabilities till the threshold p is reached. So if there are predictions that the model is much more certain about only those will be considered.

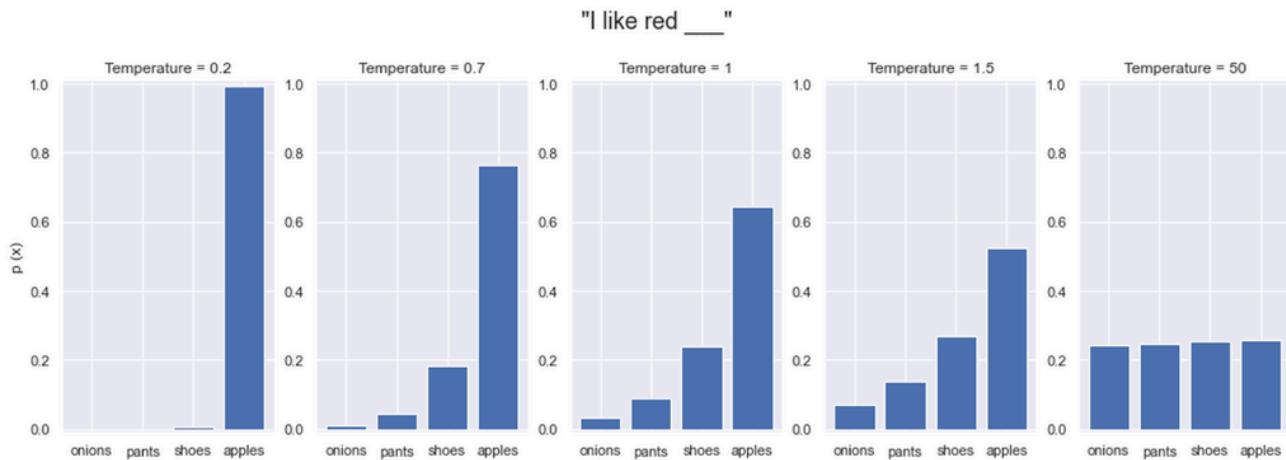
Temperature

Temperature is another component of [Top-p](#) sampling strategy which has the following impact depending on its value:

- $t==0.0$: ends up choosing the token with highest probability - no randomness here - same as greedy decoding - precise use cases.
- $0.0 < t < 1.0$: the probabilities are pushed further apart, so the closer to 0.0 the less randomness - somewhere between precise and balanced use cases.
- $t==1.0$: has no impact on sampling - the original training distribution is preserved here - balanced relevance and diversity use cases.

- $t > 1.0$: the probabilities are pushed closer together, creating a lot more randomness - creative use cases.

The following set of plots should make this crystal clear:



([source](#))

To really understand the impact, the temperature factor typically gets applied to the log probabilities just before or as part of the Softmax operation.

```
scaled_logits = logits / temperature
probs = softmax(scaled_logits)
```

The softmax operation turns logit differences into probability ratios - when we divide by $t < 1.0$, we make these differences larger, causing more extreme probability ratios and a more peaked distribution. When we divide by $t > 1.0$, we make these differences smaller, causing more similar probability ratios and a more uniform distribution. At $t=0$, this effectively makes the highest logit infinitely larger than the others (though division by zero is avoided in practice).

Temperature will have no impact on Greedy decoding, Beam search and Top-K sampling strategies, as it impacts the distance between logit probabilities and all of these strategies use the top probabilities based on their order and temperature doesn't change the order of probabilities. Whereas Top-p sampling allows more or less contenders to enter the sub-set the random sampling will be pulled from based on their total probability - so the closer the probabilities are (high temp) the more randomness is possible.

Other than $t==0.0$ and $t==\infty$ there are no hard prescribed values to copy from and you will have to experiment with each use case to find the values that work the best for your needs - though you surely will find people offering good baselines for different use cases if you search the Internet.

For more on decoding methods, see this [Huggingface blog](#).

Guided Text Generation

Also known as Structured Text Generation and Assisted generation.

If the model can return its generated output in a specific format, rather than unrestricted format, you don't want the model to hallucinate invalid formats. For example, if you want a model to return a JSON dict, it should do just that.

The way this is accomplished is by using guided text generation. Instead of choosing a generated token with highest probability, the technique uses the next best token with highest probability that fits the next expected token sub-set. To elucidate with an example: if you want the model to generate a JSON list of strings like `["apples", "oranges"]` thus we expect:

```
["string", "string", ..., "string"]
123...
```

The first generated token has to be [. If the model got ", for example, instead of [, as the highest probability, and [had a lower probability - we want to pick the one with lower probability so that it'll be [.

Then the next generated token has to be ". If it's not, search for tokens with lower probabilities until " is found and choose that.

The third token has to be a valid string (i.e. not [or ").

And so on.

Basically, for each next token we need to know a subset of tokens that is allowed and choose one with the highest probability from that subset.

This is a very cool technique. Instead of trying to repair the generated output which is not always possible to match the expected format, we get the model to generate the correct output in the first place.

This technique has several costs:

- it slows down the generation - the more complex the schema it has to adhere to the slower it'll be at generating tokens. From measuring generation speed I found some structured text generation libraries perform much faster than others.
- it may contribute to model hallucination.

There are multiple implementations of this technique, as of this writing the two popular libraries are:

- <https://github.com/outlines-dev/outlines>
- <https://github.com/noamgat/lm-format-enforcer>

You ideally want the implementations that have already been integrated into inference frameworks like vLLM and others.

Faster inference with guided generation

It's possible to use the schema to speed up inference as well. For example, consider this simple "profile" schema:

```
{
  "type": "object",
  "properties": {
    "name": { "type": "string" },
    "age": { "type": "integer" }
  },
  "required": ["name", "age"]
}
```

Since the schema has specific keys `name` and `age`, as soon as the model has predicted: {"n or {"a it doesn't need to perform an auto-regressive generation to come up with `{"name": and{"age": because both of these must lead to a specific unambiguous single outcome - so here it can perform a prefill instead of decoding and save a few slow steps at it knows 100% the next few tokens will be me": or ge": correspondingly. Clearly, this approach would be most beneficial when the schema has a lot of pre-determined keys and short generated values.

Speculative decoding

Also known as Speculative inference or Assisted generation.

Because it's very slow to generate tokens one at a time, sometimes it is possible to cheat and speed things up by using a much smaller and faster draft model. So for example, your normal inference uses Llama-70B which would be quite slow, but we could use Llama-7b as a draft model and then we could verify if the prediction is correct but doing it at once for all tokens.

Example: let's take a prompt `I'm turnin', turnin', turnin', turnin', turnin' around and all that I can see is just and now:`

1. use Llama-7b to predict another `lemon` tree auto-regressively, in 3 steps, but much faster than Llama-70b.
2. now use Llama-70b to run a batch of 3 prompts:

```
[...I can see is just]  
[...I can see is just another]  
[...I can see is just another lemon]
```

I shortened the full prompt for the sake of the demo with `...` - it should be there for real. And I'm pretending that each token is a full word here.

And now in a single step Llama-70B generates:

```
[...I can see is just] another  
[...I can see is just another] lemon  
[...I can see is just another lemon] tree
```

Now there could be multiple outcomes:

- if everything matches - in 3 short and 1 long step we generated the final result, instead of using 3 long steps.
- if only `another lemon` matched - we might still better off if it saved time.
- if nothing or little matched we wasted a bit of time.

Obviously, if instead of 3 tokens we had more tokens the savings are likely to be bigger.

Also, don't miss the fact that we did the same amount of compute here and then some, as compared to doing this generation with the large model normally, but the latency of this approach can be much better - so the user on average should get a better response time from your application using it - if the draft model is much smaller and still produces good predictions.

When there is a partial mismatch we can go back to the draft model and feed it all the matched tokens before the first mismatched token and the next good token predicted by the big model and get it to make a new fast prediction for the mismatching tail.

The draft model ideally should be trained on the same data (or least data from a similar distribution) and its tokenizer has to be the same as the large model.

Speculative decoding gives the highest return on [input-grounded tasks](#), such as translation, summarization, document QA, multi-turn chat because in those tasks the range of possible outputs is much smaller and the draft model is much more likely to match the big model.

For the same reason it works best in when used in [greedy decoding](#), as there is the least amount of possible variations during generation. If not using greedy decoding, you will want to have the value of [temperature](#) close to 0.

Here is a good indepth dive into this subject: [Assisted Generation: a new direction toward low-latency text generation](#).

One other much simpler solution for [input-grounded tasks](#), is to use [ngram prompt lookup decoding](#). In this approach there is no need for a draft model, instead the prompt is searched for matching strings to generate candidates. In some situations it's said to speed decoding up by 2x+.

Privacy-preserving inference

Most companies serving inference will run into a user privacy need. It should be safe for a user to submit a query w/o someone snooping on it. One solution would be an on-premise solution where the client runs the server themselves and then there is no privacy issue, but that most likely is going to expose provider's IP - model's weights and possibly code/algorithms. Therefore, there is a need for a fully encrypted generation - that is the computations are to be performed on client-encrypted data.

The solutions that address this need are called Privacy-Preserving Machine Learning (PPML).

One of the solutions is called Fully [Homomorphic Encryption](#) (FHE).

Have a look at one such implementation, [concrete-ml](#) that rewrites the model to be able to have the client run part of the model themselves, then the intermediary encrypted activations are sent to the server to perform the attention and then sent back to the client. Thus the provider retains part of their IP - and I suppose this part of IP prevents the client from stealing the full IP, since partial weights aren't enough to reconstruct the full model. [This article](#) goes into more details.

There are various other approaches, e.g. this paper: [LLMs Can Understand Encrypted Prompt: Towards Privacy-Computing Friendly Transformers](#) goes into a custom solution based on Secure Multi-Party Computation (MPC) and FHE and has a good reference list.

The problem with current solutions is the huge computational overhead - which greatly impacts the cost and latency. In the future ASIC solutions should address these issues.

Model parallelism

When a model can't fit onto a single accelerator or when it's more efficient to split the model across multiple accelerators even if it does fit but barely, the same [Model Parallelism techniques](#) from training apply to inference.

Tensor parallelism

Most of the time you are most likely to only run into [Tensor Parallelism](#) where the model weights are sharded across 2 to 8 accelerators. Ideally you want to try to fit the model into a single accelerator, because then it has the least amount of overhead during generation. But surprisingly you are likely to end up with higher decoding throughput if you use tensor parallelism - this is because it enables you to fit much larger batches and also because the `forward` call may be faster despite the additional comms between the accelerators. Of course, you will be getting this speed up at a cost of using more accelerators in some cases. So it's best to experiment, there will be use-cases where a higher tensor parallelism degree will give a better total throughput considering the same number of accelerators.

footnote: in my experiments TP=1 leads to the highest TTFT and lowest decoding throughput, as compared to TP>1. So if you're being requested to make the TTFT faster and the model fits, use smaller TP or TP=1. If you're being requested to make the decoding throughput faster, throw more accelerators at it with a higher TP degree.

Pipeline parallelism

Further, while tensor parallelism helps to lower latency, using [Pipeline Parallelism](#) could help increase the throughput. This is especially so for very large models where many accelerators have to be used anyway to even load the model's weights. If say you're using Llama 405B and TP=8 is used, then each accelerator has to all-reduce to 7 other accelerators, whereas with PP=8 each accelerator needs to communicate only with 2 other accelerators (`recv` the input from the previous stage and `send` the current output to the next stage), creating a much lower pressure on the networking layer and this can speed things up dramatically if the hardware supports it.

It's important to clarify here that PP can be superior to TP only if you use the full PP and not the naive PP. In the [naive PP](#) only one PP stage works at any given time so it'd perform worse than TP. To benefit from PP the inference framework needs to feeds all PP stages in parallel to perform [full PP](#).

The other important thing about PP inference is that unlike training, there is no `backward` pass, thus there is no need to solve the inactivity bubble problem. There will be only a tiny overhead of filling the PP stages in the first few

micro-batches.

And as with training you may find that some mix of TP and PP will lead to the best outcome (e.g. TP=4 + PP=4 for Llama 405B). So make sure to experiment and measure different configurations and pick the one that meets your needs.

Key inference performance metrics

There are two ways to look at performance metrics, the usual system metrics of latency and throughput, and the user-experience metrics: Time To First Token (TTFT) and Time Per Output Token (TPOT). Let's look at both pairs.

System performance metrics

Latency

Latency is the time it took to receive the complete response since a request was sent.

This includes the time to:

1. receive the request
2. pre-process the prompt (the prefill stage)
3. generate the new tokens of the response (the decoding stage)
4. send the response back to the client.

The time to receive the request and send the response is mostly the same with a small variation due to the differences in the length of the prompt and the generated response. These length variations should have a negligible impact to the total time.

The prefill stage processes all the prompt's tokens in parallel so here as well the variations in the length of the prompt shouldn't make too much of a difference, albeit longer prompts will consume more accelerator memory and impact the total throughput.

The decoding stage is the one most impacted by the length of the generated response since each new token is generated as a separate step. Here the longer the response the longer the decoding stage will be.

If the server doesn't have enough capacity to process all current requests at once and has to queue some of them, then the wait time in the queue extends the latency by that time.

footnote: if you think of car traffic on the road, latency is the time it takes one to drive from point A to point B (e.g. home to office), including the speed limitations due to traffic lights, jams and legal limits.

Throughput

Throughput measures the ability of an inference server to process many requests in parallel and batch requests efficiently.

The definition of throughput could be defined by how many requests can be served concurrently, but since some requests get served much faster than others, so that several short requests could be served during a single long request, it makes sense to count the total rate of tokens generated across the system.

Thus a more common definition of inference throughput is total tokens generated per second across the whole system.

footnote: if you think of car traffic on the road, throughput is how many cars can move through a given road at any given time. The more lanes the road has and the higher the speed limit the higher the throughput of that road. But clearly some vehicles are short and some are long, so some sort of normalization is needed. For example, ferries calculate how many meters or feet of vehicles they can fit it and thus long vehicles pay more than short ones.

User experience metrics

While there are many characteristics an inference server can be judged by - like power usage, efficiency and cost, one

could say that since the systems interface humans - the most important characteristics are all in the domain on having a smooth user experience. If the user experience is slow and choppy, the user will go to a competitor. Therefore the key needs are:

Time To First Token

Time To First Token (TTFT) is defined as the time that passed since the user hit the Submit button (or Enter) and the moment they have received a first word or a part of the word in return.

A very low Time To First Token (TTFT) is wanted. These days users are conditioned to expect from any application to start responding ideally faster than 1 second. Therefore the shorter the time the user has to wait before they start receiving the first tokens the better. This becomes even more important for chatbots which are expected to be interactive. The length of TTFT is impacted by many elements, the key ones being the computation of the [prefill stage](#) (pre-processing the prompt) and whether the request got its processing immediately upon user request received or whether it had to wait in the queue.

It's important to observe that TTFT w/o a load on a server can be very different from when a server is under a heavy load. If normally the server sends the first token in 1 sec, if the server is already busy processing all the requests it can handle at once and there is a queue, the effective TTFT other than for the first few requests, could easily be much much longer. So usually one should measure an average TTFT and report it together with the number of concurrent requests sent during the benchmark.

This is a non-trivial metric since depending on the prompt size the time will vary, so ideally you'd want to normalize it to the number of tokens in the prompt.

Time Per Output Token

Time Per Output Token (TPOT) is a per user metric. It measures how long does it take for a new token to be generated for a given user.

A relatively low Time Per Output Token (TPOT) is desired, but it doesn't have to be too high. This time ideally should be close to the reading speed of the human who sent the request. So for example if you serve first graders the TPOT can be quite low, but the more educated the person is the faster TPOT should be to achieve a smooth reading experience.

According to wiki there are [3 types of reading](#) and the reading speed is measured in words per minute (WPM).

The average tokens per word can vary from tokenizer to tokenizer, primarily depending on their vocab size and the language(s). Here let's consider an English tokenizer with about 1.5 tokens per word. Now we can convert words per minute (WPM) to tokens per minute (TPM).

And now we just need to divide by 60 to get Tokens Per Second (TPS) and invert to get time per output token (TPOT)

So $TPOT = 60 / (WPM * 1.5)$ in seconds

Reader	WPM	TPM	TPS	TPOT
Subvocal	250	375	6.25	0.16
Auditory	450	675	11.25	0.089
Visual	700	1050	18.75	0.057

Remember to change the 1.5 coefficient to the actual word to tokens average ratio of your tokenizer. For example, as of this writing OpenAI ChatGPT's with a 50k vocab is reported to be about 1.3 tokens per word, while many other LLMs have 30k vocabs, which lead to a higher tokens per words ratio.

As you can see TPOT is an awkward value to track and think of in one's head, so once you know your targeted TPOT it's better to convert it to Tokens Per Seconds (TPS) and track that instead.

Therefore in this example if your system can generate a sustainable 20 tokens per second per request your clients will be satisfied since that system will be able to keep up even with the super-fast readers at 700 words per minute.

And there, of course, will be users who would prefer to wait till the generation is complete before they would start reading the response. In which case faster is better.

Depending on the type of generation, the following is likely to apply:

1. Image - all-at-once
2. Text - as fast as user's reading speed or all-at-once if they prefer not to have moving parts before they start reading
3. Audio - as fast as user's listening speed
4. Video - as fast as user's watching speed

If this is an offline system that doesn't interface individual humans and there are just batches of requests processed these metrics make no difference, but latency and throughput are the key ones.

Simplified performance metrics

As you can tell the discussed above metrics have a lot of overlap in them. Practically we can reduce all of them to just these 2 metrics: Prefill throughput and Decode throughput - and probably how many parallel requests per second the system can handle.

Prefill throughput

This is how fast the system can pre-process the prompt - in tokens per second.

Assuming there is a negligible overhead of receiving and sending the request, in the absence of a queue where the incoming request gets immediately worked on **TTFT** is really the number of tokens in the prompt divided by the prefill tokens per seconds plus the time to generate the first token (which we can ignore as it'd be very fast).

If there is a queue then prefill throughput isn't enough, because then TTFT can be much longer as one has to add the time the request spent in the queue.

Decode throughput

This is how fast the system generates response tokens - in tokens per second.

This addresses, both the throughput and Time Per Output Token metrics.

The response latency then is the number of tokens in the prompt divided by the prefill throughput plus the number of generated tokens divided by the decode throughput.

More metric notes

Accelerator utilization

Accelerator utilization - either percentage or power measurement is a good indicator of whether your setup uses the accelerators efficiently. For example, if you use NVIDIA GPUs and you `watch -n 0.5 nvidia-smi` and you see you're at 10% "gpu util" while massively bombarding the inference server with many requests that usually means that either the inference server is very inefficient (e.g. spends a lot of time copying things back-and-forth) or it could be that the clients are inefficient at how they receive the data (i.e. too much IO blocking).

footnote: when I first wrote a simple benchmark using the openai client it worked fine at a low concurrency, but at a higher concurrency the inference server dropped its gpu util to 6-7%. After I replaced the client with `aiohttp` API it went up to 75%. Therefore beware that it's your benchmark that could be the culprit of bad performance reports and not the server.

This is somewhat of an equivalent of using [TFLOPS to measure training efficiency](#).

In the ideal case you want your accelerator utilization to be as high as possible. Beware that at least for NVIDIA GPUs `gpug util` [isn't what you might think it is](#), but if it reports a very low percentage it's good enough of a signal to know that there is definitely an inefficiency problem.

Percentiles

If you read benchmarks and run into things like p50, p75, p90, p95 and p99 percentiles - these are statistical filters that give you the results based on the percentage of results that fit under (or over) a certain threshold. Even the same request is likely to take a slightly different response time when it gets re-run multiple times. So, for example, if 95% of the time a throughput was higher than a certain value - that would be a p95 percentile. That also would mean that 5% of the time the throughput was lower than that same threshold value. The higher the number next to p, the more difficult it is to achieve.

For example, let's look at partial output of a system loading report generated by `k6` on an inference server:

```
http_req_duration..: avg=13.74s  min=12.54s  med=13.81s  max=13.83s  p(90)=13.79s  p(95)=13.83s
http_req_receiving.: avg=27.98μs  min=15.16μs  med=21.6μs  max=98.13μs  p(90)=44.98μs  p(95)=59.2μs
http_req_sending...: avg=133.8μs  min=20.47μs  med=75.39μs  max=598.04μs  p(90)=327.73μs  p(95)=449.65μs
```

If we look at the first line which reported the total generation time, if we look at the minimal recorded value of 12.54 seconds, we then know that 90% of responses took between 12.54 and 13.79 secs and 95% of responses took between 12.54 and 13.83 secs - and in this particular case the median reported value is between the p90 and p95 values.

The same interpretation applies to the other lines in the report, but the key exemplification here is that p90 values are lower than p95 values because time is being measured (the lower the better).

Percentiles are useful when outliers aren't important, so, for example, instead of looking at the slowest throughput measured you'd say ignore the worst 5% of outcomes and suddenly the system's performance looks much much better. But one has to be very careful with such discarding of bad outcomes when dealing with users, since it means that some of them will have a bad experience using your system. Also 5% translates to a whole lot of users if you have millions of them.

Please refer to [Percentile](#) for a much more indepth explanation.

Speeding up model loading time

When serving in production it might be OK to let the model takes its loading time since it happens once and then the server runs for days, so this overhead is amortized over many days. But when doing research, development and testing it's critical that the inference server starts serving really fast.

Sometimes the overhead is just loading to CPU and then moving the tensors to the accelerators, at other times there is an additional need to shard the tensors for multiple accelerators to perform [TP](#) and [PP](#).

Various approaches are used for that - most involve some sort of pre-sharding and caching, with a subsequent direct loading onto GPU.

For example:

- vLLM supports the `--load-format` flag, where one could choose options like `npformat` (numpy format caching) or `tensorizer` using CoreWeave's [Tensorizer](#). ([recipe](#) and, of course, if you use `TP>1` you want to [pre-shard the weights once](#)).
- TensorRT-LLM requires the user to build a model engine for each specific use-case and loads the pre-made shards at run time (unless you're using the simplified API which will build the model engine on the fly on every server start).

Benchmarks

You can write your own benchmark as explained in [key inference performance metrics](#) or use an existing one.

At the moment I use mainly the [prefill throughput](#) and [decode throughput](#) benchmarks. The first one just measures tokens per second from the moment the request was sent and the first generated token received, and the second one is the throughput between the first and the last generated tokens received. Here is the relevant snippet of such measurement using [openai client completions API](#):

```
[... create client, data, etc. ...]
prefill_tokens_len = len(prompt)
start_time = time.time()
decode_text = ""
decode_started = False
completion = client.completions.create(prompt=prompt, ...)
for chunk in completion:
    if chunk.choices:
        decode_text += text
        if not decode_started:
            decode_started_time = time.time()
            prefill_time = decode_started_time - start_time
            decode_started = True

    end_time = time.time()
    decode_time = end_time - decode_started_time
    decode_tokens = tokenizer.encode(decode_text)
    decode_tokens_len = len(decode_tokens)

    # tokens/per sec
    prefill_throughput = prefill_tokens_len / prefill_time
    decode_throughput = decode_tokens_len / decode_time
```

The `prefill_throughput` is not very precise here, since the client only know when it sent the request and received the first token, so a bit more went into this stage than pure prompt-preprocessing, but it should be close enough.

Of course, like any serious benchmark, you want to run this multiple times to get realistic numbers, as the variance between single runs can be quite large.

note: I've discovered that when I use the openAI client it doesn't scale well and with many concurrent requests the openAI client creates a bottleneck and doesn't measure the real server performance - I am yet to figure out if it's an issue in my code or the openAI client or how it interacts with vLLM server - I'm investigating here <https://github.com/vllm-project/vllm/issues/7935> - I found that [this version](#) of the client, rewritten to use `aiohttp` scales really well - so I switched to using it.

Here are some good starting points for load testing:

- https://github.com/vllm-project/vllm/blob/main/benchmarks/benchmark_throughput.py - my favorite tool so far
- <https://github.com/grafana/k6> - useful for load testing to simulate multiple concurrent clients - uses JavaScript clients.
- <https://github.com/bentoml/llm-bench> - benchmarks inference loads (not yet sure if it works only for BentoML)

What I'm missing right now is a tool to measure the highest concurrency the server can handle.

Anatomy of Model's Memory Usage

The inference memory usage is quite different from [training](#). Here we have:

1. Model weights
2. KV cache - crucial to not need to recalculate past tokens for each new generated token
3. Activation memory - this is the processing temporary memory which would depend on a batch size and a sequence length

Model Weights

- 4 bytes * number of parameters for fp32
- 2 bytes * number of parameters for fp16/bf16
- 1 byte * number of parameters for fp8/int8
- 0.5 bytes * number of parameters for int4

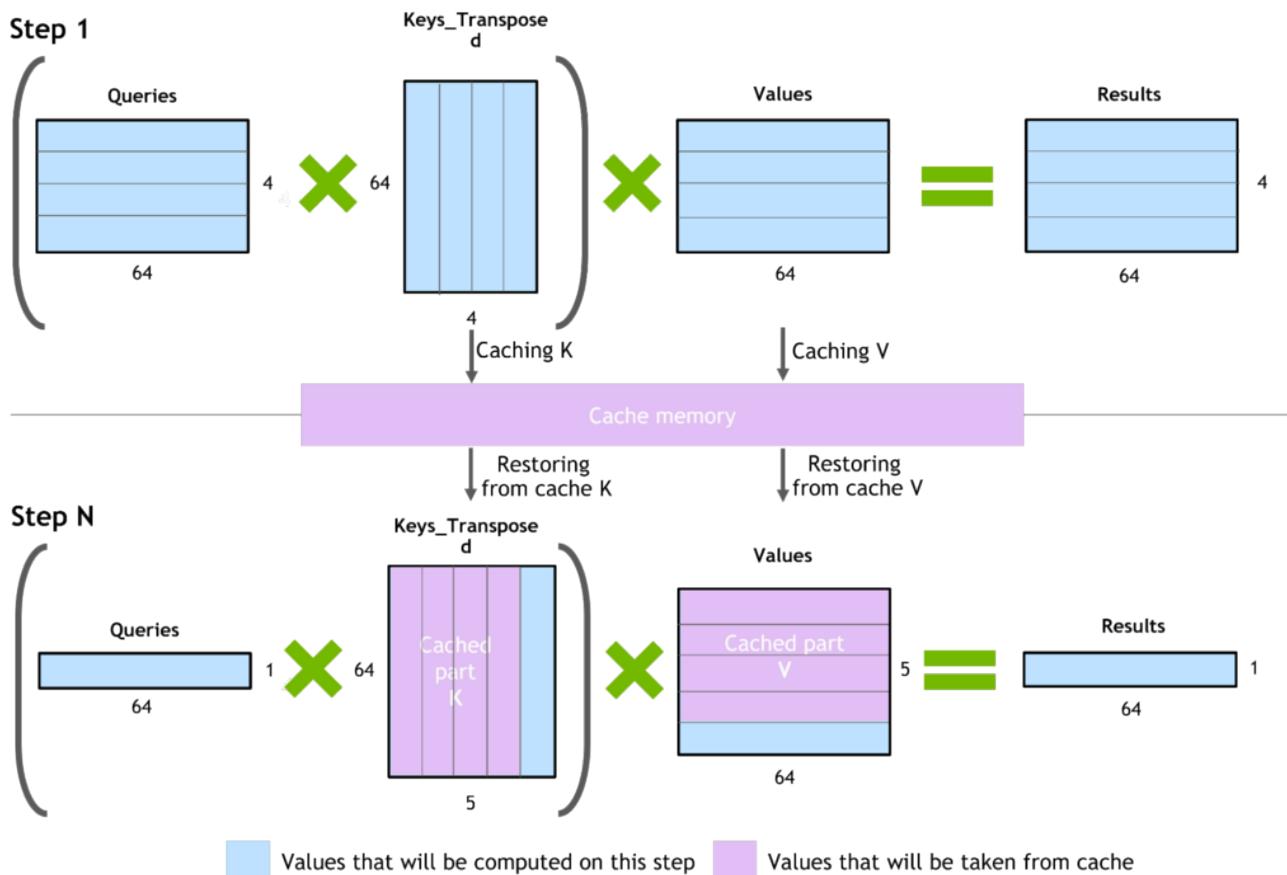
footnote: even more compact formats are being worked on as you read this, e.g. [microscaling format \(MX\)](#) also known as block floating point, where the exponent bits are shared between multiple elements of the tensor (MXFP6, MXFP4, etc.)

Example: Meta-Llama-3.1-8B in bf16 will need $2 \text{ (bf16 bytes)} * 8B \text{ (num of params)} = 16\text{GB}$ (approximately)

KV Caching

It'd be very expensive to recalculate all the previous KV (Key Value) values before each new token is generated and thus they are cached in accelerator's memory. Newly computed KV-values are appended to the existing cache.

$(Q * K^T) * V$ computation process with caching



([source](#))

KV cache size is directly proportional to the input sequence length and batch size. Past query values aren't used in the attention mechanism and thus don't need to be cached.

A KV cache of 1 token requires `dtype_bytes * 2 * num_hidden_layers * hidden_size * num_key_value_heads / num_attention_heads` bytes

notes:

- `dtype_bytes` is bytes per dtype: 4 bytes for fp32, 2 bytes for bf16/fp16, etc.
- 2 stands for keys + values as there are 2 of them.
- `num_key_value_heads / num_attention_heads` is the factor that will depend on whether multi-query (MQA), grouped-query (GQA) or multi-head attention (MHA) is used. for MHA it'll be 1, for MQA it'll be $1/\text{num_attention_heads}$ and for GQA it'll depend on how many queries are used per group, i.e. `num_key_value_heads / num_attention_heads` which is the general case for MHA and MQA.

You can get these dimensions from `config.json` inside the model's folder or from an equivalent file if it's different. e.g. [meta-llama/Meta-Llama-3.1-8B](#).

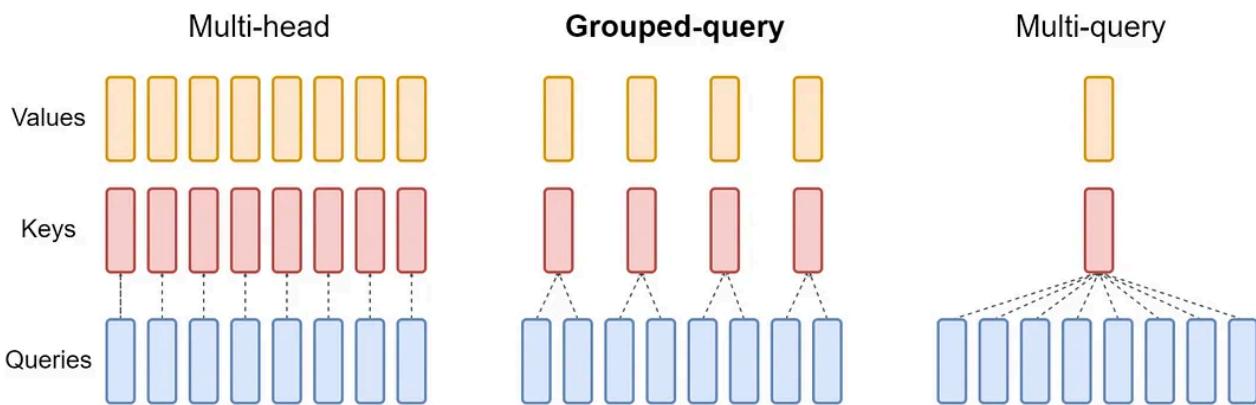
Examples:

1 token Meta-Llama-3.1-8B in bf16 will need: $2 \text{ (bf16 bytes)} * 2 \text{ (keys+values)} * 32 \text{ (num_hidden_layers)} * 4096 \text{ (hidden_size)} * 8 \text{ (num_key_value_heads)} / 32 \text{ (num_attention_heads)} / 10^{**6} = 0.131\text{MB}$. This model uses GQA so it uses 1/4th of the vanilla MHA.

A batch size of 1 of 1024 tokens will need $0.131 * 1024 = \sim 134\text{MB}$.

A batch size of 128 of 1024 tokens each will need $0.131 * 1024 * 128 / 10^{**3} = \sim 17.2\text{GB}$.

The KV cache for Meta-Llama-3.1-8B would have taken 4x more memory per token if it were to use MHA, 8x less memory if it were to use MQA. It's easy to see why from this diagram:



[source](#)

In this case the model has `num_key_value_heads=8` and `num_attention_heads=32`, hence MQA and GQA use 32x and 4x less memory than MHA, correspondingly.

[DeepSeek v3](#) introduced Multi-Latent Attention (MLA) which compresses the Key and Value into a latent vector, which further reduces the KV-cache size. See section 2.1.1 of the paper for the specific details.

[SwiftKV](#) was invented to deal with the common situation of 10:1 ratio of prefill vs decode use-cases, reducing inference computation during prompt processing rather than just compressing memory. By combining model rewiring and knowledge-preserving self-distillation, SwiftKV achieves substantial reductions in computational overhead during inference with minimal accuracy loss, leading to transformative improvements in throughput, latency and cost efficiency for enterprise LLM workloads by up to 2x.

KV cache while saving recomputation has a big negative impact on inference's performance. Here is a quote from [Dynamic Memory Compression: Retrofitting LLMs for Accelerated Inference](#):

2.3. Memory-Bound and Compute-Bound Operations Every operation performed with a GPU accelerator, such as General Matrix Multiply (GEMM), is either memory-bound or compute-bound. In the former case, the overall runtime is dominated by high bandwidth memory (HBM) access, while in the latter by the actual computations. Auto-regressive generation with Transformer LLMs, where the sequence length for every forward pass is $n = 1$, tends to be memory-bound rather than compute-bound. The vast majority of a forward pass is spent either processing linear layers (in MHSA, Feed-Forward, and output vocabulary projection) or calculating attention scores and outputs from Equation (4). For linear layers, the ratio of FLOPS to memory accesses improves as the batch size increases, and more FLOPS are performed with the set of layer weights retrieved from the HBM. Eventually, with a large enough batch size, linear layers become compute-bound. On the other hand, for the calculation of Equation (4) inside MHSA layers during auto-regressive inference, the ratio of FLOPS to input size remains constant, and MHSA layers are memory-bound regardless of the batch size. It follows that for those layers, latency scales linearly with the size of the KV cache.

- Equation (4) is the usual self-attention mechanism equation of $\text{softmax}(Q, K)V$

A smaller KV cache would lead to faster generation and higher GPU utilization. So various techniques like gisting, context distillation, key-value eviction policies, memory compression, multi-query attention, grouped-query attention, cross-layer attention, anchor-based self-attention, quantization and many others are used to accomplish that.

In the case of a small batch size you should check if disabling KV cache will not give a better overall performance.

Inference frameworks

There are many dozens of inference frameworks and more emerging every week, so it'd be very difficult to list them all. So this here you will find a starter list of a handful of inference frameworks that might be a good fit for your needs, but do check out other frameworks if the ones listed here don't satisfy your needs.

This section is trying hard to be neutral and not recommend any particular frameworks, since even if I was able to try them all out, there is no way I could possibly guess which framework will work best for which user/company.

vLLM

[vLLM](#)

DeepSpeed-FastGen

[DeepSpeed-FastGen](#) from [the DeepSpeed team](#).

TensorRT-LLM

[TensorRT-LLM](#) (also integrated what used to be [FasterTransformer](#))

Supports only NVIDIA gpus.

TGI

[TGI](#)

SGLang

[SGLang](#)

OpenPPL

[OpenPPL](#)

LightLLM

[LightLLM](#)

LMDeploy

[LMDeploy](#)

MLC-LLM

[MLC-LLM](#)

If your favourite inference framework isn't listed please make a PR and add it.

Accelerator-specific frameworks

Most inference framework obviously support NVIDIA CUDA. Some support AMD ROCm and Intel Gaudi.

But there are accelerator-specific frameworks:

Intel Gaudi, MAX, etc.

- <https://github.com/intel/intel-extension-for-transformers>

How to choose an inference framework

To choose the most suitable inference framework you need to answer at least the following questions:

1. Does the framework have the features that you need? Be careful here, some frameworks list that they support feature A, but when you try to use it it's not well integrated or works really slowly.
2. Does the framework have a permissive license that meets your current and future needs? In practice we have seen that frameworks with licenses that go against commercial use are likely to be rejected by the community. For example HF's TGI tried to charge for commercial use and it backfired - so its license got reverted to the original Apache 2.0 license and now they are trying to recover from being shunned by the community.
3. Does the framework have a thriving community of contributors? Go to the framework's github repo and check how many contributors it has - if it's very few I'd be concerned as thriving frameworks usually tend to invite contributions and that means that even if the core contributors don't have the time some feature, some contributors might do it for you.
4. Does the framework have a high adoption? github stars are often a good indication, but sometimes it can be hyped up via smart marketing moves. So seek out other signals - e.g. used by count on the framework's repo's main page on github - these are real numbers. Lots of PRs and Issues is another flag. Then search the web for how many articles are written about the given framework.
5. Are the framework maintainers responsive to Issues and PRs? Some frameworks will ignore many Issues and even PRs. Check the count of how many PRs and Issues not being addressed. A high outstanding open Issues is a difficult signal - from one side it means this is a popular project, from the other side it means the developer team and contributors can't cope with the needs of its users.
6. While the majority of ML inference frameworks are written in Python, with some sprinkling of C++ or Triton for fused kernels, some aren't written in Python. (e.g. NVIDIA's TensorRT-LLM is 99% C++, TGI's big chunk is written in

Rust). If something doesn't work the way you need it to and you filed an Issue and it's not being addressed, will you be able to get your hands dirty and modify the framework to do what you need?

7. The other issue you may run into is that some frameworks don't want your PRs where you implemented missing features or made improvements and then you will end up maintaining a fork, which can be extremely difficult if you want to continue syncing with the upstream and cause a lot of pain to your developers.
8. Run some sort of load [benchmarks](#) for the desired workloads to know if the performance is adequate.
9. Will you want to choose the [best cost-effective accelerator](#) down the road or are you OK being locked in into a specific vendor? For example, a framework from NVIDIA isn't likely to support any other accelerators besides NVIDIA's. Same goes for AMD and Intel.

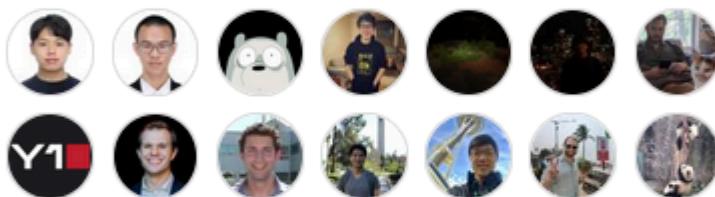
For example, here is a snapshot of [vLLM](#)'s stats as of 2024-08-24, which is one of the most popular inference frameworks as of this writing.

Used by 1.6k



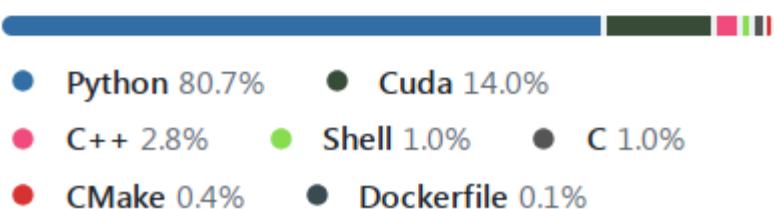
+ 1,551

Contributors 504



[+ 490 contributors](#)

Languages



You can see that it is used by many github repositories, it has a lot of contributors and that it's written mainly in Python. So it should be very easy to find this information about any inference framework you may consider. This was just an example and not an endorsement of vLLM.

Inference Chips

Besides general purpose accelerators some vendors have been working special ASICs that are designed to do Inference-only.

Groq

- [Groq](#)

Resources

- [A Survey on Efficient Inference for Large Language Models \(2024\)](#)

Debugging and Troubleshooting

Guides

- [Debugging PyTorch programs](#)
- [Diagnosing Hangings and Deadlocks in Multi-Node Multi-GPU Python Programs](#)
- [Network Debug](#)
- [Troubleshooting NVIDIA GPUs](#)
- [Underflow and Overflow Detection](#)

Tools

- [Debug Tools](#)
- [torch-distributed-gpu-test.py](#) - this a `torch.distributed` diagnostics script that checks that all GPUs in the cluster (one or many nodes) can talk to each other and allocate gpu memory.
- [NicerTrace](#) - this is an improved `trace` python module with multiple additional flags added to the constructor and more useful output.

Debugging PyTorch programs

Getting nodes to talk to each other

Once you need to use more than one node to scale your training, e.g., if you want to use DDP to train faster, you have to get the nodes to talk to each other, so that communication collectives could send data to each other. This is typically done via a comms library like [NCCL](#). And in our DDP example, at the end of training step all GPUs have to perform an `all_reduce` call to synchronize the gradients across all ranks.

In this section we will discuss a very simple case of just 2 nodes (with 8 GPUs each) talking to each other and which can then be easily extended to as many nodes as needed. Let's say that these nodes have the IP addresses 10.0.0.1 and 10.0.0.2.

Once we have the IP addresses we then need to choose a port for communications.

In Unix there are 64k ports. The first 1k are reserved for common services so that any computer on the Internet could connect to any other computer knowing ahead of time which port to connect to. For example, port 22 is reserved for SSH. So that whenever you do `ssh example.com` in fact the program open a connection to `example.com:22`.

As there are thousands of services out there, the reserved 1k ports is not enough, and so various services could use pretty much any port. But fear not, when you get your Linux box on the cloud or an HPC, you're unlikely to have many preinstalled services that could use a high number port, so most ports should be available.

Therefore let's choose port 6000.

Now we have: `10.0.0.1:6000` and `10.0.0.2:6000` that we want to be able to communicate with each other.

The first thing to do is to open port `6000` for incoming and outgoing connections on both nodes. It might be open already or you might have to read up the instructions of your particular setup on how to open a given port.

Here are multiple ways that you could use to test whether port 6000 is already open.

```
telnet localhost:6000
nmap -p 6000 localhost
nc -zv localhost 6000
curl -v telnet://localhost:6000
```

Most of these should be available via `apt install` or whatever your package manager uses.

Let's use `nmap` in this example. If I run:

```
$ nmap -p 22 localhost
[...]
PORT      STATE SERVICE
22/tcp    open  ssh
```

We can see the port is open and it tells us which protocol and service is allocated as a bonus.

Now let's run:

```
$ nmap -p 6000 localhost
```

```
[...]
```

```
PORT      STATE SERVICE
6000/tcp  closed X11
```

Here you can see port 6000 is closed.

Now that you understand how to test, you can proceed to test the `10.0.0.1:6000` and `10.0.0.2:6000`.

First ssh to the first node in terminal A and test if port 6000 is opened on the second node:

```
ssh 10.0.0.1
nmap -p 6000 10.0.0.2
```

if all is good, then in terminal B ssh to the second node and do the same check in reverse:

```
ssh 10.0.0.2
nmap -p 6000 10.0.0.1
```

If both ports are open you can now use this port. If either or both are closed you have to open these ports. Since most clouds use a proprietary solution, simply search the Internet for "open port" and the name of your cloud provider.

The next important thing to understand is that compute nodes will typically have multiple network interface cards (NICs). You discover those interfaces by running:

```
$ sudo ifconfig
```

One interface is typically used by users to connecting to nodes via ssh or for various other non-compute related services - e.g., sending an email or download some data. Often this interface is called `eth0`, with `eth` standing for Ethernet, but it can be called by other names.

Then there is the inter-node interface which can be Infiniband, EFA, OPA, HPE Slingshot, etc. ([more information](#)). There could be one or dozens of those interfaces.

Here are some examples of `ifconfig`'s output:

```
$ sudo ifconfig
enp5s0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
      inet 10.0.0.23  netmask 255.255.255.0  broadcast 10.0.0.255
      [...]
```

I removed most of the output showing only some of the info. Here the key information is the IP address that is listed after `inet`. In the example above it's `10.0.0.23`. This is the IP address of interface `enp5s0`.

If there is another node, it'll probably be `10.0.0.24` or `10.0.0.21` or something of sorts - the last segment will be the one with a different number.

Let's look at another example:

```
$ sudo ifconfig
ib0      Link encap:UNSPEC  HWaddr 00-00-00-00-00-00-00-00-00-00-00-00-00-00-00-00
          inet addr:172.0.0.50  Bcast: 172.0.0.255  Mask:255.255.255.0
            [...]
```

Here `ib` typically tells us it's an InfiniBand card, but really it can be any other vendor. I have seen [OmniPath](#) using `ib` for example. Again `inet` tells us the IP of this interface is `172.0.0.50`.

If you lost me, we want the IP addresses so that we could test if ip:port is open on each node in question.

Finally, going back to our pair of `10.0.0.1:6000` and `10.0.0.2:6000` let's do an `all_reduce` test using 2 terminals, where we choose `10.0.0.1` as the master host which will coordinate other nodes. For testing we will use this helper debug program [torch-distributed-gpu-test.py](#).

In terminal A:

```
$ ssh 10.0.0.1
$ python -m torch.distributed.run --role $(hostname -s): --tee 3 --nnodes 2 --nproc_per_node 8 \
--master_addr 10.0.0.1 --master_port 6000 torch-distributed-gpu-test.py
```

In terminal B:

```
$ ssh 10.0.0.2
$ python -m torch.distributed.run --role $(hostname -s): --tee 3 --nnodes 2 --nproc_per_node 8 \
--master_addr 10.0.0.1 --master_port 6000 torch-distributed-gpu-test.py
```

Note that I'm using the same `--master_addr 10.0.0.1 --master_port 6000` in both cases because we checked port 6000 is open and we use `10.0.0.1` as the coordinating host.

This approach of running things manually from each node is painful and so there are tools that automatically launch the same command on multiple nodes

pdsh

`pdsh` is one such solution - which is like `ssh` but will automatically run the same command on multiple nodes:

```
PDSH_RCMD_TYPE=ssh pdsh -w 10.0.0.1,10.0.0.2 \
"python -m torch.distributed.run --role $(hostname -s): --tee 3 --nnodes 2 --nproc_per_node 8 \
--master_addr 10.0.0.1 --master_port 6000 torch-distributed-gpu-test.py"
```

You can see how I folded the 2 sets of commands into 1. If you have more nodes, just add more nodes as `-w` argument.

SLURM

If you use SLURM, it's almost certain that whoever set things up already have all the ports opened for you, so it should just work. But if it doesn't the information in this section should help debug things.

Here is how you'd use this with SLURM.

```
#!/bin/bash
```

```

#SBATCH --job-name=test-nodes          # name
#SBATCH --nodes=2                     # nodes
#SBATCH --ntasks-per-node=1           # crucial - only 1 task per dist per node!
#SBATCH --cpus-per-task=10            # number of cores per tasks
#SBATCH --gres=gpu:8                 # number of gpus
#SBATCH --time 0:05:00                # maximum execution time (HH:MM:SS)
#SBATCH --output=%x-%j.out            # output file name
#
export GPUS_PER_NODE=8
export MASTER_ADDR=$(scontrol show hostnames $SLURM_JOB_NODELIST | head -n 1)
export MASTER_PORT=6000
#
srun --jobid $SLURM_JOBID bash -c 'python -m torch.distributed.run \
--nproc_per_node $GPUS_PER_NODE --nnodes $SLURM_NNODES --node_rank $SLURM_PROCID \
--master_addr $MASTER_ADDR --master_port $MASTER_PORT \
torch-distributed-gpu-test.py'

```

If you have more than 2 nodes you just need to change the number of nodes and the above script will automatically work for any number of them.

MPI:

Another popular way is to use [Message Passing Interface \(MPI\)](#). There are a few open source implementations of it available.

To use this tool you first create a `hostfile` that contains your target nodes and the number of processes that should be run on each host. In the example of this section, with 2 nodes and 8 gpus each it'd be:

```

$ cat hostfile
10.0.0.1:8
10.0.0.2:8

```

and to run, it's just:

```

$ mpirun --hostfile -np 16 -map-by ppr:8:node python my-program.py

```

Note that I used `my-program.py` here because `torch-distributed-gpu-test.py` was written to work with `torch.distributed.run` (also known as `torchrun`). With `mpirun` you will have to check your specific implementation to see which environment variable it uses to pass the rank of the program and replace `LOCAL_RANK` with it, the rest should be mostly the same.

Nuances:

- You might have to explicitly tell it which interface to use by adding `--mca btl_tcp_if_include 10.0.0.0/24` to match our example. If you have many network interfaces it might use one that isn't open or just the wrong interface.
- You can also do the reverse and exclude some interfaces. e.g. say you have `docker0` and `lo` interfaces - to exclude those add `--mca btl_tcp_if_exclude docker0,lo`.

`mpirun` has a gazillion of flags and I will recommend reading its manpage for more information. My intention was only to show you how you could use it. Also different `mpirun` implementations may use different CLI options.

Solving the Infiniband connection between multiple nodes

In one situation on Azure I got 2 nodes on a shared subnet and when I tried to run the 2 node NCCL test:

```
NCCL_DEBUG=INFO python -u -m torch.distributed.run --nproc_per_node=1 --nnodes 2 --rdzv_endpoint
10.2.0.4:6000 --rdzv_backend c10d torch-distributed-gpu-test.py
```

I saw in the debug messages that Infiniband interfaces got detected:

```
node-2:5776:5898 [0] NCCL INFO NET/IB : Using [0]ibP111p0s0:1/IB [1]rdmaP111p0s2:1/RoCE [RO]; OOB
eth0:10.2.0.4<0>
```

But the connection would then time out with the message:

```
node-2:5776:5902 [0] transport/net_ib.cc:1296 NCCL WARN NET/IB : Got completion from peer 10.2.0.5<33092>
with error 12, opcode 0, len
0, vendor err 129 (Recv)
node-2:5776:5902 [0] NCCL INFO transport/net.cc:1134 -> 6
node-2:5776:5902 [0] NCCL INFO proxy.cc:679 -> 6
node-2:5776:5902 [0] NCCL INFO proxy.cc:858 -> 6 [Proxy Thread]
```

and nothing works. So here the Ethernet connectivity between 2 nodes works but not the IB interface.

There could be a variety of reason for this failing, but of the most likely one is when you're on the cloud and the 2 nodes weren't provisioned so that their IB is connected. So your Ethernet inter-node connectivity works, but it's too slow. Chances are that you need to re-provision the nodes so that they are allocated together. For example, on Azure this means you have to allocate nodes within a special [availability set](#)

Going back to our case study, once the nodes were deleted and recreated within an availability set the test worked out of the box.

The individual nodes are often not meant for inter-node communication and often the clouds have the concept of clusters, which are designed for allocating multiple nodes as a group and are already preconfigured to work together.

Prefixing logs with node:rank, interleaved asserts

In this section we will use `torchrun` (`torch.distributed.run`) during the demonstration and at the end of this section similar solutions for other launchers will be listed.

When you have warnings and tracebacks (or debug prints), it helps a lot to prefix each log line with its `hostname:rank` prefix, which is done by adding `--role $(hostname -s): --tee 3` to `torchrun`:

```
python -m torch.distributed.run --role $(hostname -s): --tee 3 --nnodes 1 --nproc_per_node 2 \
torch-distributed-gpu-test.py
```

Now each log line will be prefixed with `[hostname:rank]`

Note that the colon is important.

If you're in a SLURM environment the above command line becomes:

```
srun --jobid $SLURM_JOBID bash -c 'python -m torch.distributed.run \
--nproc_per_node $GPUS_PER_NODE --nnodes $SLURM_NNODES --node_rank $SLURM_PROCID \
--master_addr $MASTER_ADDR --master_port $MASTER_PORT \
--role $(hostname -s): --tee 3 \
torch-distributed-gpu-test.py'
```

Of course adjust your environment variables to match, this was just an example.

Important! Note, that I'm using a single quoted string of commands passed to bash -c. This way hostname -s command is delayed until it's run on each of the nodes. If you'd use double quotes above, hostname -s will get executed on the starting node and then all nodes will get the same hostname as the prefix, which defeats the purpose of using these flags. So if you use double quotes you need to rewrite the above like so:

```
srun --jobid $SLURM_JOBID bash -c "python -m torch.distributed.run \
--nproc_per_node $GPUS_PER_NODE --nnodes $SLURM_NNODES --node_rank \$SLURM_PROCID \
--master_addr $MASTER_ADDR --master_port $MASTER_PORT \
--role \$$(hostname -s): --tee 3 \
torch-distributed-gpu-test.py"
```

\$SLURM_PROCID is escaped too as it needs to be specific to each node and it's unknown during the launch of the slurm job on the main node. So there are 2 \\$ escapes in this version of the command.

This prefixing functionality is also super-helpful when one gets the distributed program fail and which often results in interleaved tracebacks that are very difficult to interpret. So by grepping for one node:rank string of choice, it's now possible to reconstruct the real error message.

For example, if you get a traceback that looks like:

```
File "/path/to/training/dataset.py", line 785, in __init__
File "/path/to/training/dataset.py", line 785, in __init__
  if self.dataset_proba.sum() != 1:
AttributeError: 'list' object has no attribute 'sum'
  if self.dataset_proba.sum() != 1:
File "/path/to/training/dataset.py", line 785, in __init__
File "/path/to/training/dataset.py", line 785, in __init__
  if self.dataset_proba.sum() != 1:
  if self.dataset_proba.sum() != 1:
AttributeError: 'list' object has no attribute 'sum'
AttributeError: 'list' object has no attribute 'sum'
AttributeError: 'list' object has no attribute 'sum'
```

and when it's dozens of frames over 8 nodes it can't be made sense of, but the above -tee + --role addition will generate:

```
[host1:0]  File "/path/to/training/dataset.py", line 785, in __init__
[host1:1]  File "/path/to/training/dataset.py", line 785, in __init__
```

```
[host1:0]    if self.dataset_proba.sum() != 1:  
[host1:0]AttributeError: 'list' object has no attribute 'sum'  
[host1:1]    if self.dataset_proba.sum() != 1:  
[host1:2]  File "/path/to/training/dataset.py", line 785, in __init__  
[host1:3]  File "/path/to/training/dataset.py", line 785, in __init__  
[host1:3]    if self.dataset_proba.sum() != 1:  
[host1:2]    if self.dataset_proba.sum() != 1:  
[host1:1]AttributeError: 'list' object has no attribute 'sum'  
[host1:2]AttributeError: 'list' object has no attribute 'sum'  
[host1:3]AttributeError: 'list' object has no attribute 'sum'
```

and you can grep this output for just one host:rank prefix, which gives us:

```
$ grep "[host1:0]" log.txt  
[host1:0]  File "/path/to/training/dataset.py", line 785, in __init__  
[host1:0]    if self.dataset_proba.sum() != 1:  
[host1:0]AttributeError: 'list' object has no attribute 'sum'
```

and voila, you can now tell what really happened. And as I mentioned earlier there can be easily a hundred to thousands of interleaved traceback lines there.

Also, if you have just one node, you can just pass `-tee 3` and there is no need to pass `--role`.

If `hostname -s` is too long, but you have each host with its own sequence number like:

```
[really-really-really-long-hostname-5:0]  
[really-really-really-long-hostname-5:1]  
[really-really-really-long-hostname-5:2]
```

you can of course make it shorter by replacing `hostname -s` with `hostname -s | tr -dc '0-9'`, which would lead to much shorter prefixes:

```
[5:0]  
[5:1]  
[5:2]
```

And, of course, if you're doing debug prints, then to solve this exact issue you can use [printflock](#).

Here is how you accomplish the same feat with other launchers:

- `srun` in SLURM: add `--label`
- `openmpi`: add `--tag-output`
- `accelerate`: you can just pass the same `-tee + --role` flags as in `torchrun`

Dealing with Async CUDA bugs

When using CUDA, failing pytorch programs very often produce a python traceback that makes no sense or can't be acted upon. This is because due to CUDA's async nature - when a CUDA kernel is executed, the program has already moved on

and when the error happened the context of the program isn't there. The async functionality is there to make things faster, so that while the GPU is churning some `matmul` the program on CPU could already start doing something else.

At other times some parts of the system will actually tell you that they couldn't generate the correct traceback, as in this error:

```
[E ProcessGroupNCCL.cpp:414] Some NCCL operations have failed or timed out. Due to the
asynchronous nature of CUDA kernels, subsequent GPU operations might run on corrupted/
incomplete data. To avoid this inconsistency, we are taking the entire process down.
```

There are a few solutions.

If the failure is instant and can be reproduced on CPU (not all programs work on CPU), simply re-rerun it after hiding your GPUs. This is how you do it:

```
CUDA_VISIBLE_DEVICES="" python my-pytorch-program.py
```

The env var `CUDA_VISIBLE_DEVICES` is used to manually limit the visibility of GPUs to the executed program. So for example if you have 8 gpus and you want to run `program1.py` with first 4 gpus and `program2.py` with the remaining 2 gpus you can do:

```
CUDA_VISIBLE_DEVICES="0,1,2,3" python my-pytorch-program1.py
CUDA_VISIBLE_DEVICES="4,5,6,7" python my-pytorch-program2.py
```

and the second program won't be the wiser that it's not using GPUs 0-3.

But in the case of debug we are hiding all GPUs, by setting `CUDA_VISIBLE_DEVICES=""`.

Now the program runs on CPU and you will get a really nice traceback and will fix the problem in no time.

But, of course, if you your program requires multiple GPUs this won't work. And so here is another solution.

Rerun your program after setting this environment variable:

```
CUDA_LAUNCH_BLOCKING=1 python my-pytorch-program.py
```

This variable tells pytorch (or any other CUDA-based program) to turn its async nature off everywhere and now all operations will be synchronous. So when the program crashes you should now get a perfect traceback and you will know exactly what ails your program.

In theory enabling this variable should make everything run really slow, but in reality it really depends on your software. We did the whole of BLOOM-176B training using `CUDA_LAUNCH_BLOCKING=1` with [Megatron-Deepspeed](#) and had zero slowdown - we had to use it as pytorch was hanging without it and we had no time to figure the hanging out.

So, yes, when you switch from async to sync nature, often it can hide some subtle race conditions, so there are times that a hanging disappears as in the example I shared above. So measure your throughput with and without this flag and sometimes it might actual not only help with getting an in-context traceback but actually solve your problem altogether.

Note: [NCCL==2.14.3 coming with pytorch==1.13 hangs](#) when `CUDA_LAUNCH_BLOCKING=1` is used. So don't use it with that version of pytorch. The issue has been fixed in `nccl>=2.17` which should be included in `pytorch==2.0`.

segfaults and getting a backtrace from a core file

It's not uncommon for a complex pytorch program to segfault and drop a core file. Especially if you're using complex extensions like NCCL.

The corefile is what the program generates when it crashes on a low-level - e.g. when using a python extension - such as a CUDA kernel or really any library that is coded directly in some variant of C or another language and made accessible in python through some binding API. The most common cause of a segfault is when such software accesses memory it has not allocated. For example, a program may try to free memory it hasn't allocated. But there could be many other reasons.

When a segfault event happens Python can't do anything, as the proverbial carpet is pulled out from under its feet, so it can't generate an exception or even write anything to the output.

In these situation one must go and analyse the libC-level calls that lead to the segfault, which is luckily saved in the core file.

If your program crashed, you will often find a file that will look something like: `core-python-3097667-6`

Before we continue make sure you have `gdb` installed:

```
sudo apt-get install gdb
```

Now make sure you know the path to the python executable that was used to run the program that crashed. If you have multiple python environment you have to activate the right environment first. If you don't `gdb` may fail to unpack the core file.

So typically I'd go:

```
conda activate my-env
gdb python core-python-3097667-6
```

- adjust `my-env` to whatever env you use, or instead of conda use whatever way you use to activate your python environment - and perhaps you're using the system-wise python and then you don't need to activate anything.
- adjust the name of the core file to the file you have gotten - it's possible that there are many - pick the latest then.

Now `gdb` will churn for a bit and will give you a prompt where you type: `bt`. We will use an actual core file here:

```
(gdb) bt
#0 0x0000147539887a9f in raise () from /lib64/libc.so.6
#1 0x000014753985ae05 in abort () from /lib64/libc.so.6
#2 0x000014751b85a09b in __gnu_cxx::__verbose_terminate_handler() [clone .cold.1] () from /lib64/libstdc++.so.6
#3 0x000014751b86053c in __cxxabiv1::__terminate(void (*)()) () from /lib64/libstdc++.so.6
#4 0x000014751b860597 in std::terminate() () from /lib64/libstdc++.so.6
#5 0x000014751b86052e in std::rethrow_exception(std::__exception_ptr::__exception_ptr) () from /lib64/libstdc++.so.6
#6 0x000014750bb007ef in c10d::ProcessGroupNCCL::WorkNCCL::handleNCCLGuard() ()
  from .../python3.8/site-packages/torch/lib/libtorch_cuda_cpp.so
#7 0x000014750bb04c69 in c10d::ProcessGroupNCCL::workCleanupLoop() ()
  from .../python3.8/site-packages/torch/lib/libtorch_cuda_cpp.so
#8 0x000014751b88cba3 in execute_native_thread_routine () from /lib64/libstdc++.so.6
```

```
#9 0x000014753a3901cf in start_thread () from /lib64/libpthread.so.0
#10 0x0000147539872dd3 in clone () from /lib64/libc.so.6
```

and there you go. How do you make sense of it?

Well, you go from the bottom of the stack to the top. You can tell that a `clone` call was made in `libc` which then called `start_thread` in `libpthread` and then if you keep going there are a bunch of calls in the torch libraries and finally we can see that the program terminated itself, completing with `raise` from `libc` which told the Linux kernel to kill the program and create the core file.

This wasn't an easy to understand backtrace.

footnote: Yes, python calls it a *traceback* and elsewhere it's called a *backtrace* - it's confusing, but it's more or less the same thing.

Actually I had to ask pytorch devs for help and received:

- PyTorch `ProcessGroup` watchdog thread caught an asynchronous error from NCCL
- This error is an “unhandled system error” which in this particular case turned out to be an IB-OPA error
- The `ProcessGroup`’s `WorkCleanUp` thread rethrew the error so that the main process would crash and the user would get notified (otherwise this async error would not surface)

Trust me there are times when even if you’re inexperienced the backtrace can give you enough of a hint to where you should look for troubleshooting.

But fear not - most of the time you won’t need to understand the traceback. Ideally you’d just attach the core file to your filed Issue. But it can easily be 5GB large. So the developers that will be trying to help you will ask you to generate a `gdb` backtrace and now you know how to do that.

I didn’t promise it’ll be easy, I just showed you where to start.

Now another useful details is that many programs these days run multiple threads. And `bt` only shows the main thread of the process. But, often, it can be helpful to see where other threads in the process were when segfault has happened. For that you simply type 2 commands at the (`gdb`) prompt:

```
(gdb) thread apply all bt
(gdb) bt
```

and this time around you typically will get a massive report, one backtrace per thread.

py-spy

This is a super-useful tool for analysing hanging programs. For example, when you have a resource deadlock or there is an issue with a network connection.

You will find an exhaustive coverage of this tool [here](#).

strace

Similar to [py-spy](#), `strace` is a super-useful tool which traces any running application at the low-level system calls - e.g. `libc` and alike.

For example, run:

```
strace python -c "print('strace')"
```

and you will see everything that is done at the system call level as the above program runs.

But usually it's more useful when you have a stuck program that spins all CPU cores at 100% but nothing happens and you want to see what's it doing. In this situation you simply attach to the running program like so:

```
strace --pid PID
```

where you get the PID for example from the output of `top` or `ps`. Typically I just copy-n-paste the PID of the program that consumes the most CPU - `top` usually shows it at the very top of its listing.

Same as `py-spy` you may need `sudo` perms to attach to an already running process - it all depends on your system setup. But you can always start a program with `strace` as I have shown in the original example.

Let's look at a small sub-snippet of the output of `strace python -c "print('strace')"`

```
write(1, "strace\n", 7strace
) = 7
```

Here we can see that a `write` call was executed on filedescriptor 1, which almost always is `stdout` (`stdin` being 0, and `stderr` being 2).

If you're not sure what a filedescriptor is pointing to, normally you can tell from `strace`'s output itself. But you can also do:

```
ls -1 /proc/PID/fd
```

where PID is the pid of the currently running program you're trying to investigate.

For example, when I run the above while running a `pytest` test with `gpus`, I got (partial output):

```
1-wx----- 1 stas stas 64 Mar 1 17:22 5 -> /dev/null
1r-x----- 1 stas stas 64 Mar 1 17:22 6 -> /dev/urandom
1rwx----- 1 stas stas 64 Mar 1 17:22 7 -> /dev/nvidiactl
1rwx----- 1 stas stas 64 Mar 1 17:22 8 -> /dev/nvidia0
1r-x----- 1 stas stas 64 Mar 1 17:22 9 -> /dev/nvidia-caps/nvidia-cap2
```

so you can see that a device `/dev/null` is open as FD (file descriptor) 5, `/dev/urandom` as FD 6, etc.

Now let's go look at another snippet from our `strace` run.

```
access("/etc/ld.so.preload", R_OK) = -1 ENOENT (No such file or directory)
```

Here it tried to see if file `/etc/ld.so.preload` exists, but as we can see it doesn't - this can be useful if some shared library is missing - you can see where it's trying to load it from.

Let's try another one:

here we can see that it opens `/lib/x86_64-linux-gnu/libpthread.so.0` and assigns it FD 3, it then reads 832 chars from FD 3, (we can also see that the first chars are ELF - which stands for a shared library format), then memory maps it and closes that file.

In this following example, we see a python cached file is opened, its filepointer is moved to 0, and then it's read and closed.

It's important to notice that file descriptors are re-used, so we have seen the same FD 3 twice, but each time it was open to a different file.

If your program is for example trying to reach to the Internet, you can also tell these calls from `strace` as the program would be reading from a socket file descriptor.

So let's run an example on a program that downloads files from the HF hub:

```
strace python -c 'import sys; from transformers import AutoConfig; AutoConfig.from_pretrained(sys.argv[1])' t5-small
```

here is some relevant to this discussion snippet:

```
socket(AF_INET6, SOCK_STREAM|SOCK_CLOEXEC, IPPROTO_TCP) = 3
setsockopt(3, SOL_TCP, TCP_NODELAY, [1], 4) = 0
ioctl(3, FIONBIO, [1]) = 0
connect(3, {sa family=AF_INET6, sin6 port=htons(443), sin6 flowinfo=htonl(0)}, {inet pton(AF_INET6,
```

```

"2600:1f18:147f:e850:e203:c458:10cd:fc3c
", &sin6_addr), sin6_scope_id=0}, 28) = -1 EINPROGRESS (Operation now in progress)
poll([{fd=3, events=POLLOUT|POLLERR}], 1, 10000) = 1 ([{fd=3, revents=POLLOUT}])
getsockopt(3, SOL_SOCKET, SO_ERROR, [0], [4]) = 0
[...]
write(3, "\26\3\3\0F\20\0\0BA\4\373m\244\16\354\334\205\361j\225\356\202m*\305\332\275\251\17J"..., 126)
= 126
read(3, 0x2f05c13, 5) = -1 EAGAIN (Resource temporarily unavailable)
poll([{fd=3, events=POLLIN}], 1, 9903) = 1 ([{fd=3, revents=POLLIN}])
read(3, "\24\3\3\0\1", 5) = 5
read(3, "\1", 1) = 1
read(3, "\26\3\3\0(", 5) = 5
read(3, "\0\0\0\0\0\0\0\0\344\273\225\4\24m\234~\371\332%1\364\254\34\3472<\0356s\313"..., 40) = 40
ioctl(3, FIONBIO, [1]) = 0
poll([{fd=3, events=POLLOUT}], 1, 10000) = 1 ([{fd=3, revents=POLLOUT}])
write(3, "\27\3\3\1.\0\374\361\217\337\377\264g\215\364\345\256\260\211$\326pkR\345\276,\321\221`-..."..., 307) = 307
ioctl(3, FIONBIO, [1]) = 0
read(3, 0x2ef7283, 5) = -1 EAGAIN (Resource temporarily unavailable)
poll([{fd=3, events=POLLIN}], 1, 10000) = 1 ([{fd=3, revents=POLLIN}])

```

You can see where that again it uses FD 3 but this time it opens a INET6 socket instead of a file. You can see that it then connects to that socket, polls, reads and writes from it.

There are many other super useful understandings one can derive from using this tool.

BTW, if you don't want to scroll up-down, you can also save the output to a file:

```
strace -o strace.txt python -c "print('strace')"
```

Now, since you're might want to strace the program from the very beginning, for example to sort out some race condition on a distributed filesystem, you will want to tell it to follow any forked processes. This what the **-f** flag is for:

```
strace -o log.txt -f python -m torch.distributed.run --nproc_per_node=4 --nnodes=1 --tee 3 test.py
```

So here we launch 4 processes and will end up running strace on at least 5 of them - the launcher plus 4 processes (each of which may spawn further child processes).

It will conveniently prefix each line with the pid of the program so it should be easy to tell which system was made by which process.

But if you want separate logs per process, then use **-ff** instead of **-f**.

The strace manpage has a ton of other useful options.

Invoke pdb on a specific rank in multi-node training

Once pytorch 2.2 is released you will have a new handy debug feature:

```

import torch.distributed as dist
[...]

def mycode(...):

    dist.breakpoint(0)

```

This is the same as `ForkedPdb` (below) but will automatically break for you on the rank of your choice - `rank0` in the example above. Just make sure to call `up;;n` right away when the breakpoint hits to get into your normal code.

Here is what it does underneath:

```

import sys
import pdb

class ForkedPdb(pdb.Pdb):
    """
    PDB Subclass for debugging multi-processed code
    Suggested in: https://stackoverflow.com/questions/4716533/how-to-attach-debugger-to-a-python-subprocess
    """
    def interaction(self, *args, **kwargs):
        _stdin = sys.stdin
        try:
            sys.stdin = open('/dev/stdin')
            pdb.Pdb.interaction(self, *args, **kwargs)
        finally:
            sys.stdin = _stdin

def mycode():

    if dist.get_rank() == 0:
        ForkedPdb().set_trace()
    dist.barrier()

```

so you can code it yourself as well.

And you can use that `ForkedPdb` code for normal forked applications, minus the `dist` calls.

Floating point math discrepancies on different devices

It's important to understand that depending on which device the floating point math is performed on the outcomes can be different. For example doing the same floating point operation on a CPU and a GPU may lead to different outcomes, similarly when using 2 different GPU architectures, and even more so if these are 2 different types of accelerators (e.g. NVIDIA vs. AMD GPUs).

Here is an example of discrepancies I was able to get doing the same simple floating point math on an 11 Gen Intel i7 CPU

and an NVIDIA A100 80GB (PCIe) GPU:

```
import torch

def do_math(device):
    inv_freq = (10 ** (torch.arange(0, 10, device=device) / 10))
    print(f"{inv_freq[9]:.20f}")
    return inv_freq.cpu()

a = do_math(torch.device("cpu"))
b = do_math(torch.device("cuda"))

torch.testing.assert_close(a, b, rtol=0.0, atol=0.0)
```

when we run it we get 2 out of 10 elements mismatch:

```
7.94328212738037109375
7.94328308105468750000
[...]
AssertionError: Tensor-likes are not equal!

Mismatched elements: 2 / 10 (20.0%)
Greatest absolute difference: 9.5367431640625e-07 at index (9,)
Greatest relative difference: 1.200604771156577e-07 at index (9,)
```

This was a simple low-dimensional example, but in reality the tensors are much bigger and will typically end up having more mismatches.

Now you might say that the `1e-6` discrepancy can be safely ignored. And it's often so as long as this is a final result. If this tensor from the example above is now fed through a 100 layers of `matmuls`, this tiny discrepancy is going to compound and spread out to impact many other elements with the final outcome being quite different from the same action performed on another type of device.

For example, see this [discussion](#) - the users reported that when doing Llama-2-7b inference they were getting quite different logits depending on how the model was initialized. To clarify the initial discussion was about DeepSpeed potentially being the problem, but in later comments you can see that it was reduced to just which device the model's buffers were initialized on. The trained weights aren't an issue they are loaded from the checkpoint, but the buffers are recreated from scratch when the model is loaded, so that's where the problem emerges.

It's uncommon that small variations make much of a difference, but sometimes the difference can be clearly seen, as in this example where the same image is produced on a CPU and an MPS device.

To better illustrate the problem that motivated me to post this issue, I can present some output from CURL. This first image is from the CPU:



This is from the mps device:



This snapshot and the commentary come from this [PyTorch Issue thread](#).

If you're curious where I pulled this code from - this is a simplified reduction of this original code in [modeling_llama.py](#):

```
class LlamaRotaryEmbedding(nn.Module):
    def __init__(self, dim, max_position_embeddings=2048, base=10000, device=None):
        super().__init__()

        self.dim = dim
        self.max_position_embeddings = max_position_embeddings
        self.base = base
        inv_freq = 1.0 / (self.base ** (torch.arange(0, self.dim, 2).float().to(device) / self.dim))
        self.register_buffer("inv_freq", inv_freq, persistent=False)
```

Debug Tools

git-related tools

Useful aliases

Show a diff of all files modified in the current branch against HEAD:

```
alias brdiff="def_branch=\$(git symbolic-ref refs/remotes/origin/HEAD | sed 's@^refs/remotes/origin/@@'); git diff origin/\$def_branch..."
```

Same, but ignore white-space differences, adding `--ignore-space-at-eol` or `-w`:

```
alias brdiff-nows="def_branch=\$(git symbolic-ref refs/remotes/origin/HEAD | sed 's@^refs/remotes/origin/@@'); git diff -w origin/\$def_branch..."
```

List all the files that were added or modified in the current branch compared to HEAD:

```
alias brfiles="def_branch=\$(git symbolic-ref refs/remotes/origin/HEAD | sed 's@^refs/remotes/origin/@@'); git diff --name-only origin/\$def_branch..."
```

Once we have the list, we can now automatically open an editor to load just added and modified files:

```
alias bremacs="def_branch=\$(git symbolic-ref refs/remotes/origin/HEAD | sed 's@^refs/remotes/origin/@@'); emacs \$(git diff --name-only origin/\$def_branch...) &"
```

git-bisect

(note to self: this is a sync from `the-art-of-debugging/methodology.md` which is the true source)

The discussed next approach should work for any revision control system that supports bisecting. We will use `git bisect` in this discussion.

`git bisect` helps to quickly find the commit that caused a certain problem.

Use case: Say, you were using `transformers==4.33.0` and then you needed a more recent feature so you upgraded to the bleed-edge `transformers@main` and your code broke. There could have been hundreds of commits between the two versions and it'd be very difficult to find the right commit that lead to the breakage by going through all the commits. Here is how you can quickly find out which commit was the cause.

footnote: HuggingFace Transformers is actually pretty good at not breaking often, but given its complexity and enormous size it happens nevertheless and the problems are fixed very quickly once reported. Since it's a very popular Machine Learning library it makes for a good debugging use case.

Solution: Bisecting all the commits between the known good and bad commits to find the one commit that's to blame.

We are going to use 2 shell terminals: A and B. Terminal A will be used for `git bisect` and terminal B for testing your software. There is no technical reason why you couldn't get away with a single terminal but it's easier with 2.

1. In terminal A fetch the git repo and install it in devel mode (`pip install -e .`) into your Python environment.

```
git clone https://github.com/huggingface/transformers
cd transformers
pip install -e .
```

Now the code of this clone will be used automatically when you run your application, instead of the version you previously installed from PyPi or Conda or elsewhere.

Also for simplicity we assume that all the dependencies have already been installed.

2. next we launch the bisecting - In terminal A, run:

```
git bisect start
```

3. Discover the last known good and the first known bad commits

`git bisect` needs just 2 data points to do its work. It needs to know one earlier commit that is known to work (good) and one later commit that is known to break (bad). So if you look at the sequence of commits on a given branch it'd have 2 known points and many commits around these that are of an unknown quality:

```
..... orig_good ..... .... .... .... orig_bad ....
----->----->-----> time
```

So for example if you know that `transformers==4.33.0` was good and `transformers@main (HEAD)` is bad, find which commit is corresponding to the tag `4.33.0` by visiting [the releases page](#) and searching for `4.33.0`. We find that it was commit with SHA [5a4f340d](#).

footnote: typically the first 8 hex characters are enough to have a unique identifier for a given repo, but you can use the full 40 character string.

So now we specify which is the first known good commit:

```
git bisect good 5a4f340d
```

and as we said we will use `HEAD` (latest commit) as the bad one, in which case we can use `HEAD` instead finding out the corresponding SHA string:

```
git bisect bad HEAD
```

If however you know it broke in `4.34.0` you can find its latest commit as explained above and use that instead of `HEAD`.

We are now all set at finding out the commit that broke things for you.

And after you told `git bisect` the good and the bad commits it has already switched to a commit somewhere in the middle:

```
..... orig_good ..... .... current .... ..... orig_bad .....
```

You can run `git log` to see which commit it has switched to.

And to remind, we installed this repo as `pip install -e .` so the Python environment is instantly updated to the current commit's code version.

4. Good or bad

The next stage is telling `git bisect` if the current commit is good or bad:

To do so in terminal B run your program once.

Then in terminal A run:

```
git bisect bad
```

If it fails, or:

```
git bisect good
```

if it succeeds.

If, for example, if the result was bad, `git bisect` will internally flag the last commit as new bad and will half the commits again, switching to a new current commit:

```
..... orig_good ..... current .... new_bad .... ..... orig_bad ....
```

And, vice versa, if the result was good, then you will have:

```
..... orig_good ..... .... new_good .... current ..... orig_bad ....
```

5. Repeat until no more commits left

Keep repeating step 4 step until the problematic commit is found.

Once you finished bisecting, `git bisect` will tell you which commit was responsible for breaking things.

```
..... orig_good ..... .... last_good first_bad .... .. orig_bad ....
```

If you followed the little commit diagrams, it'd correspond for the `first_bad` commit.

You can then go to <https://github.com/huggingface/transformers/commit/> and append the commit SHA to that url which will take you to the commit, (e.g. <https://github.com/huggingface/transformers/commit/1234567890>)

57f44dc4288a3521bd700405ad41e90a4687abc0 and which will then link to the PR from which it originated. And then you can ask for help by following up in that PR.

If your program doesn't take too long to run even if there are thousands of commits to search, you are facing n bisecting steps from 2^{**n} so 1024 commits can be searched in 10 steps.

If your program is very slow, try to reduce it to something small - ideally a small reproduction program that shows the problem really fast. Often, commenting out huge chunks of code that you deem irrelevant to the problem at hand, can be all it takes.

If you want to see the progress, you can ask it to show the current range of remaining commits to check with:

```
git bisect visualize --oneline
```

6. Clean up

So now restore the git repo clone to the same state you started from (most likely `HEAD) with:

```
git bisect reset
```

and possibly reinstall the good version of the library while you report the issue to the maintainers.

Sometimes, the issue emerges from intentional backward compatibility breaking API changes, and you might just need to read the project's documentation to see what has changed. For example, if you switched from `transformers==2.0.0` to `transformers==3.0.0` it's almost guaranteed that your code will break, as major numbers difference are typically used to introduce major API changes.

7. Possible problems and their solutions:

a. skipping

If for some reason the current commit cannot be tested - it can be skipped with:

```
git bisect skip
```

and it `git bisect` will continue bisecting the remaining commits.

This is often helpful if some API has changed in the middle of the commit range and your program starts to fail for a totally different reason.

You might also try to make a variation of the program that adapts to the new API, and use it instead, but it's not always easy to do.

b. reversing the order

Normally git expects `bad` to be after `good`.

```
..... orig_good ..... .... .... .... orig_bad ....  
----->----->----->-----> time
```

Now, if `bad` happens before `good` revision order-wise and you want to find the first revision that fixed a previously existing problem - you can reverse the definitions of `good` and `bad` - it'd be confusing to work with overloaded logic states, so it's

recommended to use a new set of states instead - for example, `fixed` and `broken` - here is how you do that.

```
git bisect start --term-new=fixed --term-old=broken
git bisect fixed
git bisect broken 6c94774
```

and then use:

```
git fixed / git broken
```

instead of:

```
git good / git bad
```

c. complications

There are sometimes other complications, like when different revisions' dependencies aren't the same and for example one revision may require `numpy=1.25` and the other `numpy=1.26`. If the dependency package versions are backward compatible installing the newer version should do the trick. But that's not always the case. So sometimes one has to reinstall the right dependencies before re-testing the program.

Sometimes, it helps when there is a range of commits that are actually broken in a different way, you can either find a range of `good...bad` commits that isn't including the other bad range, or you can try to `git bisect skip` the other bad commits as explained earlier.

Diagnosing Hangings and Deadlocks in Multi-Node Multi-GPU Python Programs

While the methodologies found in this article were developed while working with multi-node multi-gpu pytorch-based training, they, of course, can help with any multi-process multi-node Python programs.

Helper tools

Try to use the following script [torch-distributed-gpu-test.py](#) to diagnose the situation.

This will help primarily with discovering network-related issues. And also to quickly understand how multi-gpu communications work.

For code-related issues read the rest of this document.

Approaches to diagnosing multi-gpu hanging / deadlocks

py-spy

First do `pip install py-spy`.

Now you can attach to each process with:

```
py-spy dump -n -p PID
```

and it will tell you where the process hangs (very often it's a nccl collective function or a `barrier`).

- `PID` is the process id of the hanging python process.
- `-n` is useful if you want to see stack traces from python extensions written in C, C++, etc., as the program may hang in one of the extensions
- you may need to add `sudo` before the command - for more details see [this note](#).

If you have no `sudo` access your sysadmin might be able to perform this for you:

```
sudo echo 0 > /proc/sys/kernel/yama/ptrace_scope
```

which will allow you running `py-spy` (and `strace`) without needing `sudo`. Beware of the possible [security implications](#) - but typically if your compute node is inaccessible from the Internet it's less likely to be a risk.

To make this change permanent edit `/etc/sysctl.d/10-ptrace.conf` and set:

```
kernel.yama.ptrace_scope = 0
```

Here is an example of `py-spy dump` python stack trace:

```
Thread 835995 (active): "MainThread"
  broadcast (torch/distributed/distributed_c10d.py:1191)
  _aggregate_total_loss (deepspeed/runtime/pipe/engine.py:540)
  train_batch (deepspeed/runtime/pipe/engine.py:330)
  train_step (megatron/training.py:436)
  train (megatron/training.py:851)
  pretrain (megatron/training.py:187)
  <module> (pretrain_gpt.py:239)
```

The very first line is where the program is stuck.

If the hanging happens inside a CPP extension, add `--native` `py-spy` and it'll show the non-python code if any.

multi-process `py-spy`

Now, how do you do it for multiple processes. Doing it one-by-one is too slow. So let's do it at once.

If the launch command was `python`, what you do is:

```
pgrep -P $(pgrep -o python) | xargs -I {} py-spy dump --pid {}
```

`if deepspeed:`

```
  pgrep -P $(pgrep -o deepspeed) | xargs -I {} py-spy dump --pid {}
```

`for accelerate:`

```
  pgrep -P $(pgrep -o accelerate) | xargs -I {} py-spy dump --pid {}
```

you get the idea.

This particular approach will only analyse the main processes and not various other sub-processes/threads spawned by these processes. So if you have 8 gpus and 8 processes, the above will generate 8 stack traces.

If you want all processes and their subprocesses, then you'd just run:

```
pgrep -f python | xargs -I {} py-spy dump --pid {}
```

(and as before replace `python` with the name of the launcher program if it's not `python`)

multi-node `py-spy` via `srun`

What if you have multiple nodes?

You can of course `ssh` to each node interactively and dump the stack traces.

If you're using the SLURM environment you can use `srun` to do it on all nodes for you.

Now in another console get the `SLURM_JOBID` (or get it from `salloc` log):

```
squeue -u `whoami` -o "%16i %9P %26j %.8T %.10M %.81 %.6D %.20S %R"
```

Now use the following `srun` command after adjusting jobid with `SLURM_JOBID` from the outcome of the command above this sentence:

```
srun --jobid=2180718 --gres=gpu:0 --nodes=40 --tasks-per-node=1 --output=trace-%N.out sh -c 'ps aux | grep python | egrep -v "grep|srun" | grep `whoami` | awk "{print \$2}" | xargs -I {} py-spy dump --native --pid {}' || echo "failed"
```

Notes:

- One must use `--gres=gpu:0` for the monitor `srun` or otherwise it will block until the main `srun` (the one running the training) exits.
- Each node will generate its unique log file named `trace-nodename.out` - so this would help to identify which node(s) are problematic. You can remove `--output=trace-%N.out` if you want it all being dumped to stdout
- In some SLURM versions you may also need to add `--overlap`
- In some SLURM versions the jobid might not match that of reported in `squeue`, so you have to get the correct `SLURM_JOB_ID` from the logs of the job you're trying to "attach" to - i.e. your `srun` job that allocated the GPUs.
- Sometimes `bash` doesn't work, but `sh` does. I think it has to do with what dot files get sourced
- You might need to also activate a custom python environment, which you can do like so:

```
srun --jobid=2180718 --gres=gpu:0 --nodes=40 --tasks-per-node=1 --output=trace-%N.out sh -c 'conda activate myenvname; ps auxc | ... ' || echo "failed"
```

or you can do it inside `~/.bashrc` or whatever shell's rc file you decide to use.

As mentioned before if you want just the main processes you'd use this instead:

```
srun --jobid=2180718 --gres=gpu:0 --nodes=40 --tasks-per-node=1 --output=trace-%N.out sh -c 'pgrep -P $(pgrep -o python) | xargs -I {} py-spy dump --pid {}' || echo "failed"
```

Adjust `python` if need be as explained in the multi-gpu section above.

The previous longer command will deliver traces for all python processes.

If you're not getting anything, start with the basic debug like:

```
srun --jobid=2180718 --gres=gpu:0 --nodes=40 --tasks-per-node=1 --output=trace-%N.out sh -c 'date'
```

once you know you're talking to all the nodes, then you can progressively unravel the depth of calls, as in:

```
srun --jobid=2180718 --gres=gpu:0 --nodes=40 --tasks-per-node=1 sh -c 'date'  
srun --jobid=2180718 --gres=gpu:0 --nodes=40 --tasks-per-node=1 sh -c 'pgrep -o python'  
srun --jobid=2180718 --gres=gpu:0 --nodes=40 --tasks-per-node=1 sh -c 'pgrep -P $(pgrep -o python)'  
srun --jobid=2180718 --gres=gpu:0 --nodes=40 --tasks-per-node=1 sh -c 'pgrep -P $(pgrep -o python) | xargs -I {} py-spy dump --pid {}'
```

and at each stage check that the output makes sense - e.g. the 2nd and 3rd call you should be getting the PIDs of the processes.

multi-node py-spy via pdsh

pdsh seems to be a good easy tool to use to accomplish remote work on multiple nodes. Say, you're running on 2 nodes with hostnames `nodename-5` and `nodename-8`, then you can quickly test that remote execution is working by getting the `date` on all of these hosts with just:

```
$ PDSH_RCMD_TYPE=ssh pdsh -w nodename-[5,8] "date"
nodename-5: Wed Oct 25 04:32:43 UTC 2023
nodename-8: Wed Oct 25 04:32:45 UTC 2023
```

footnote: `pdsh` should be available via a normal OS package installer

Once you tested that `date` works it's time to move to `py-spy`.

To do `py-spy` on all python processes that are sub-processes, it'd be:

```
PDSH_RCMD_TYPE=ssh pdsh -w nodename-[5,8] 'pgrep -P $(pgrep -o python) | xargs -I {} py-spy dump --pid {}'
```

but as you're likely to need to have the `~/.bashrc` run, you will need to clone it into `~/.pdshrc`, reduce that clone to what is needed to be run (e.g. modify `PATH`, `activate conda`) and then `source` it, like:

```
PDSH_RCMD_TYPE=ssh pdsh -w nodename-[5,8] 'source ~/.pdshrc; pgrep -P $(pgrep -o python) | xargs -I {} py-spy dump --pid {}'
```

The reason you need a startup script is because usually `~/.bashrc` starts with:

```
# If not running interactively, don't do anything
case $- in
  *i*) ;;
  *) return;;
esac
```

so when you run such non-interactive workflows Bash won't process its `~/.bashrc` normally (exit early) and thus anything relying on this startup script won't work. So you can either remove the non-interactive exiting code above or fork `~/.bashrc` into a startup file that only contains what's needed for the remote command to succeed.

footnote: there is nothing special about `~/.pdshrc` - any other name would do, since you're manually `source`ing it.

And if your system isn't setup to run `py-spy` w/o `sudo` as explained a few sections up, you'd need something like this:

```
PDSH_RCMD_TYPE=ssh pdsh -w nodename-[5,8] 'sudo bash -c "source ~/.pdshrc; pgrep -P $(pgrep -o python) | xargs -I {} py-spy dump --pid {}"'
```

Of course, you may need to edit the `pgrep` section to narrow down which processes you want to watch.

Additionally, to avoid being prompted with:

```
Are you sure you want to continue connecting (yes/no/[fingerprint])?
```

for every new node you haven't logged into yet, you can disable this check with:

```
echo "Host *" >> ~/.ssh/config
echo " StrictHostKeyChecking no" >> ~/.ssh/config
```

Here I assume you're on an isolated cluster so you don't need to worry about security issues and thus bypassing such check is most likely OK.

multi-node py-spy via `ds_ssh`

This is yet another way, but please make sure to read the `pdsh` section above first.

The following notes require `pip install deepspeed`.

In one SLURM environment I also attempted using `pdsh` via `ds_ssh`, but somehow I wasn't able to run `py-spy` remotely - the main issue was that remote `ssh` command wasn't giving the same env as when I was logged in interactively via `ssh`. But if you have `sudo` access on the compute nodes then you could do:

First prepare `hostfile`:

```
function makehostfile() {
perl -e '$slots=split /,/, $ENV{"SLURM_STEP_GPUS"};
$slots=8 if $slots==0; # workaround 8 gpu machines
@nodes = split /\n/, qx[scontrol show hostnames $ENV{"SLURM_JOB_NODELIST"}];
print map { "$_ slots=$slots\n" } @nodes'
}
makehostfile > hostfile
```

Adapt `$slots` to the number of gpus per node. You may have to adapt this script if your `scontrol` produces a different output.

Now run the `py-spy` extraction command over all participating nodes:

```
ds_ssh -f hostfile "source ~/.pdshrc; ps aux | grep python | grep -v grep | grep `whoami` | awk '{print
\$2}' | xargs -I {} sudo py-spy dump --pid {} "
```

Notes:

- Put inside `~/.pdshrc` whatever init code that you may need to run. If you don't need any you can remove `source ~/.pdshrc;` from the command line.
- If you don't have it already `ds_ssh` is installed when you do `pip install deepspeed`.
- you might need to `export PDSH_RCMD_TYPE=ssh` if you get `rcmd: socket: Permission denied` error

Network-level hanging

The hanging could be happening at the network level. `NCCL_DEBUG=INFO` can help here.

Run the script with `NCCL_DEBUG=INFO` env var and try to study the outcome for obvious errors. It will tell you which device it's using, e.g.:

```
DeepWhite:21288:21288 [0] NCCL INFO NET/Socket : Using [0]enp67s0:192.168.50.21<0>
```

So it's using interface `enp67s0` over `192.168.50.21`

Is your `192.168.50.21` firewalled? or is it somehow a misconfigured network device?

Does it work if you use a loopback device `127.0.0.1`?

```
NCCL_DEBUG=INFO NCCL_SOCKET_IFNAME=lo python -m torch.distributed.run --nproc_per_node 4 --nnodes 1
torch-distributed-gpu-test.py
```

if not, see what other local network devices you have via `ifconfig` - try that instead of `lo` if any.

It's currently using `enp67s0` in the above example.

Isolate problematic GPUs

You can also try to see if only some GPUs fail

For example, does it work if you use the first 2 or the last 2 gpus:

```
CUDA_VISIBLE_DEVICES=0,1 python -m torch.distributed.run --nproc_per_node 2 --nnodes 1
torch-distributed-gpu-test.py
```

then the 2nd pair:

```
CUDA_VISIBLE_DEVICES=2,3 python -m torch.distributed.run --nproc_per_node 2 --nnodes 1
torch-distributed-gpu-test.py
```

python trace

Now what happens when the training doesn't just hang, but the hanging process stops responding? e.g. this happens when there is a serious hardware issue. But what if it is recurrent and `py-spy` won't help here, since it won't be able to attach to a process that is not responding.

So next came the idea of tracing all calls like one does with `strace(1)`, I researched python calls tracing facilities and have discovered that python has a `trace` sub-system.

The following code will trace all python calls and log them to the console and into a dedicated per process log file, via a custom `Tee` module I added.

This then can help to understand where some processes stopped responding, since we will have the log of the last call and all the previous calls before it went unresponsive.

```

$ cat train.py
[...]

def main():
    # [...]
    train()

import re
class Tee:
    """
    A helper class to tee print's output into a file.
    Usage:
    sys.stdout = Tee(filename)
    """

    def __init__(self, filename):
        self.stdout = sys.stdout
        self.file = open(filename, "a")

    def __getattr__(self, attr):
        return getattr(self.stdout, attr)

    def write(self, msg):
        self.stdout.write(msg)
        self.file.write(msg)
        self.file.flush()

    def flush(self):
        self.stdout.flush()
        self.file.flush()

if __name__ == "__main__":
    import sys
    import trace
    import socket
    import os

    # enable the trace
    if 0:
        cwd = os.path.realpath('.')
        pid = os.getpid()
        hostname = socket.gethostname()
        local_rank = int(os.environ["LOCAL_RANK"])
        trace_output_file = f"{cwd}/trace-{hostname}-{local_rank}-{pid}.txt"

        # create a Trace object, telling it what to ignore, and whether to
        # do tracing or line-counting or both.
        tracer = trace.Trace(
            ignoredirs=[sys.prefix, sys.exec_prefix],
            trace=1,

```

```

        count=1,
        timing=True,
    )

# run the new command using the given tracer
sys.stdout = Tee(trace_output_file)
tracer.run('main()')
else:
    main()

```

This code doesn't require any special handing other than enabling the trace by changing `if 0` to `if 1`.

If you don't set `ignoredirs`, this will now dump all python calls. Which means expect a lot of GBs of data logged, especially if you have hundreds of GPUs.

Of course, you don't have to start tracing from `main` - if you suspect a specific are you can start tracing there instead and it'll be much faster and less data to save.

I wish I could tell `trace` which packages to follow, but alas it only supports dirs to ignore, which is much more difficult to set, and thus you end up with a lot more data than `needrf`. But still this is a super useful tool for debugging hanging processes.

Also, your code will now run much much slower and the more packages you trace the slower it will become.

NicerTrace

As `Trace` proved to provide very limited usability when debugging a complex multi-node multi-hour run crash, I have started on working on a better version of the `trace` python module.

You can find it here: [NicerTrace](#)

I added multiple additional flags to the constructor and made the output much more useful. You fill find a full working example in that same file, just run:

```
python trace/NicerTrace.py
```

and you should see:

```

trace/NicerTrace.py:1 <module>
0:00:00 <string>:    1:        trace/NicerTrace.py:185 main
0:00:00 NicerTrace.py:  186:        img = Image.new("RGB", (4, 4))
                           PIL.Image:2896 new
0:00:00 Image.py:  2912:        _check_size(size)
                           PIL.Image:2875 _check_size
0:00:00 Image.py:  2883:        if not isinstance(size, (list, tuple)):
0:00:00 Image.py:  2886:        if len(size) != 2:
0:00:00 Image.py:  2889:        if size[0] < 0 or size[1] < 0:

```

as you will see in the example I set:

```
packages_to_include=["PIL"],
```

so it'll trace `PIL` plus anything that is not under `site-packages`. If you need to trace another package, just add it to that list. This is a very fresh work-in-progress package, so it's evolving as we are trying to make it help us resolve a very complex crashing situation.

Working with generated trace files

When the per-node-rank trace files has been generated the following might be helpful to quickly analyse the situation:

- grep for a specific match and also print the file and line number where it was found:

```
grep -n "backward" trace*
```

- show `tail -1` of all trace files followed by the name of each file:

```
find . -name "trace*" -exec sh -c 'echo "$1: $(tail -3 \"$1\")"' _ {} \;
```

- or similar to the above, but print 5 last lines with the leading filename and some vertical white space for an easier reading:

```
find . -name "trace*" -exec sh -c 'echo; echo $1; echo "$(tail -5 \"$1\")"' _ {} \;
```

- count how many times grep matched a given pattern in each file and print the matched file (in this example matching the pattern `backward`):

```
find . -name "trace*" -exec sh -c 'echo "$1: $(grep "backward" $1 | wc -l)"' _ {} \;
```

good old `print`

Now once you discovered where the hanging happens to further understand why this is happening, a debugger would ideally be used, but more often than not debugging multi-process (multi-node) issues can be very difficult.

In such situations a good old `print` works. You just need to add some debug prints before the calls where things hang, things that would help understand what lead to the deadlock. For example, some `barrier` was missing and one or a few processes skipped some code and while the rest of processes are still blocking waiting for everybody to send some data (for example in NCCL collective functions like `gather` or `reduce`).

You of course, want to prefix each print with the rank of the process so that you could tell which is which. For example:

```
import torch.distributed as dist
print(f"{dist.get_rank()}: passed stage 0")
```

What you will quickly discover is that if you have multiple GPUs these prints will be badly interleaved and you will have a hard time making sense of the debug data. So let's fix this. We are going to override `print` with a custom version of the same, but which uses `flock` to ensure that only one process can write to `stdout` at the same time.

The helper module `printflock.py` is included [here](#). To activate it just run this at the top of the module you're debugging:

```
from printflock import printflock as print
```

and now all your `print` calls in that module will magically be non-iterleaved. You can of course, just use `printflock` directly:

```
from printflock import printflock
import torch.distributed as dist
printflock(f"{dist.get_rank()}: passed stage 0")
```

core files

If the hanging happens inside non-python code, and `py-spy --native` isn't enough for some reason you can make the hanging program dump a core file, which is done with one of these approaches:

```
gcore <pid>
kill -ABRT <pid>
```

and then you can introspect the core file as explained [here](#).

If you don't get the core file dumped you need to configure your system to allow so and also specify where the core files should be saved to.

To ensure the file is dumped in bash run (other shells may use a different command):

```
ulimit -c unlimited
```

To make this persistent run:

```
echo '* soft core unlimited' >> /etc/security/limits.conf
```

On some systems like Ubuntu the core files are hijacked by `apport`, check the contents of `/proc/sys/kernel/core_pattern` to see where they are sent. You can override where they are sent with:

```
sudo sysctl -w kernel.core_pattern=/tmp/core-%e.%p.%h.%t
```

Change the directory if you want to, but make sure that the user the program is running under can write to that directory. To make this change permanent edit `/etc/sysctl.conf` and add `kernel.core_pattern=/tmp/core-%e.%p.%h.%t` (or modify if it's already there).

footnote: see `man core` for all the different templates available

If on Ubuntu by default it sends core files to `apport`, which may save the core to `/var/lib/apport/coredump` or `/var/crash`. But you can change it explained above.

A quick way to test if your setup can generate a core file is:

```
sleep 10 &
killall -SIGSEGV sleep
```

Normally `SIGSEGV` isn't recommended for a real situation of diagnosing a hanging program, because `SIGSEGV` is likely to launch a sighandler, but for this test it's good enough.

Code loops

Code loops can be tricky to debug in hanging scenarios. If you have code like the following:

```
for i, d in enumerate(data):
    some_hanging_call(d)
```

it's possible that one process hangs in the first iteration, and another process in the second iteration, which makes things very confusing. But the stack trace won't give such indication, as the line numbers would be the same, even though the processes aren't in the same place code progression-wise.

In such situations unroll the loop to be:

```
d_iter = iter(data)
some_hanging_call(next(d_iter))
some_hanging_call(next(d_iter))
```

and now when you run `py-spy` the line numbers will be correct. The processes hanging in the first iteration will report the first `some_hanging_call` and those in the second iteration in the second call - as each now has its own line.

Hardware-specific issues

AMD/ROCm hangs or slow with IOMMU enabled

AMD Instinct users may need to either [Disable IOMMU](#) or set it to:

```
GRUB_CMDLINE_LINUX_DEFAULT="iommu=soft"
```

in `/etc/default/grub` (the grub config file could be elsewhere depending on the OS).

Disabling is `GRUB_CMDLINE_LINUX="amd_iommu=off"`

Underflow and Overflow Detection

For this section we are going to use the [underflow_overflow](#) library.

If you start getting `loss=NaN` or the model inhibits some other abnormal behavior due to `inf` or `nan` in activations or weights one needs to discover where the first underflow or overflow happens and what led to it. Luckily you can accomplish that easily by activating a special module that will do the detection automatically.

Let's use a `t5-large` model for this demonstration.

```
from .underflow_overflow import DebugUnderflowOverflow
from transformers import AutoModel

model = AutoModel.from_pretrained("t5-large")
debug_overflow = DebugUnderflowOverflow(model)
```

`[underflow_overflow.DebugUnderflowOverflow]` inserts hooks into the model that immediately after each forward call will test input and output variables and also the corresponding module's weights. As soon as `inf` or `nan` is detected in at least one element of the activations or weights, the program will assert and print a report like this (this was caught with `google/mt5-small` under `fp16` mixed precision):

```
Detected inf/nan during batch_number=0
Last 21 forward frames:
abs min abs max metadata
encoder.block.1.layer.1.DenseReluDense.dropout Dropout
0.00e+00 2.57e+02 input[0]
0.00e+00 2.85e+02 output
[...]
encoder.block.2.layer.0 T5LayerSelfAttention
6.78e-04 3.15e+03 input[0]
2.65e-04 3.42e+03 output[0]
None output[1]
2.25e-01 1.00e+04 output[2]
encoder.block.2.layer.1.layer_norm T5LayerNorm
8.69e-02 4.18e-01 weight
2.65e-04 3.42e+03 input[0]
1.79e-06 4.65e+00 output
encoder.block.2.layer.1.DenseReluDense.wi_0 Linear
2.17e-07 4.50e+00 weight
1.79e-06 4.65e+00 input[0]
2.68e-06 3.70e+01 output
encoder.block.2.layer.1.DenseReluDense.wi_1 Linear
8.08e-07 2.66e+01 weight
1.79e-06 4.65e+00 input[0]
1.27e-04 2.37e+02 output
encoder.block.2.layer.1.DenseReluDense.dropout Dropout
0.00e+00 8.76e+03 input[0]
```

```

0.00e+00 9.74e+03 output
    encoder.block.2.layer.1.DenseReluDense.wo Linear
1.01e-06 6.44e+00 weight
0.00e+00 9.74e+03 input[0]
3.18e-04 6.27e+04 output
    encoder.block.2.layer.1.DenseReluDense T5DenseGatedGeluDense
1.79e-06 4.65e+00 input[0]
3.18e-04 6.27e+04 output
    encoder.block.2.layer.1.dropout Dropout
3.18e-04 6.27e+04 input[0]
0.00e+00      inf output

```

The example output has been trimmed in the middle for brevity.

The second column shows the value of the absolute largest element, so if you have a closer look at the last few frames, the inputs and outputs were in the range of $1e4$. So when this training was done under fp16 mixed precision the very last step overflowed (since under fp16 the largest number before inf is $64e3$). To avoid overflows under fp16 the activations must remain way below $1e4$, because $1e4 * 1e4 = 1e8$ so any matrix multiplication with large activations is going to lead to a numerical overflow condition.

At the very start of the trace you can discover at which batch number the problem occurred (here `Detected inf/nan during batch_number=0` means the problem occurred on the first batch).

Each reported frame starts by declaring the fully qualified entry for the corresponding module this frame is reporting for. If we look just at this frame:

```

encoder.block.2.layer.1.layer_norm T5LayerNorm
8.69e-02 4.18e-01 weight
2.65e-04 3.42e+03 input[0]
1.79e-06 4.65e+00 output

```

Here, `encoder.block.2.layer.1.layer_norm` indicates that it was a layer norm for the first layer, of the second block of the encoder. And the specific calls of the forward is `T5LayerNorm`.

Let's look at the last few frames of that report:

```

Detected inf/nan during batch_number=0
Last 21 forward frames:
abs min  abs max  metadata
[...]
    encoder.block.2.layer.1.DenseReluDense.wi_0 Linear
2.17e-07 4.50e+00 weight
1.79e-06 4.65e+00 input[0]
2.68e-06 3.70e+01 output
    encoder.block.2.layer.1.DenseReluDense.wi_1 Linear
8.08e-07 2.66e+01 weight
1.79e-06 4.65e+00 input[0]
1.27e-04 2.37e+02 output
    encoder.block.2.layer.1.DenseReluDense.wo Linear
1.01e-06 6.44e+00 weight

```

```

0.00e+00 9.74e+03 input[0]
3.18e-04 6.27e+04 output
    encoder.block.2.layer.1.DenseReluDense T5DenseGatedGeluDense
1.79e-06 4.65e+00 input[0]
3.18e-04 6.27e+04 output
    encoder.block.2.layer.1.dropout Dropout
3.18e-04 6.27e+04 input[0]
0.00e+00      inf output

```

The last frame reports for `Dropout.forward` function with the first entry for the only input and the second for the only output. You can see that it was called from an attribute `dropout` inside `DenseReluDense` class. We can see that it happened during the first layer, of the 2nd block, during the very first batch. Finally, the absolute largest input elements was $6.27e+04$ and same for the output was `inf`.

You can see here, that `T5DenseGatedGeluDense.forward` resulted in output activations, whose absolute max value was around 62.7K, which is very close to fp16's top limit of 64K. In the next frame we have `Dropout` which renormalizes the weights, after it zeroed some of the elements, which pushes the absolute max value to more than 64K, and we get an overflow (`inf`).

As you can see it's the previous frames that we need to look into when the numbers start going into very large for fp16 numbers.

Let's match the report to the code from [models/t5/modeling_t5.py](#):

```

class T5DenseGatedGeluDense(nn.Module):
    def __init__(self, config):
        super().__init__()
        self.wi_0 = nn.Linear(config.d_model, config.d_ff, bias=False)
        self.wi_1 = nn.Linear(config.d_model, config.d_ff, bias=False)
        self.wo = nn.Linear(config.d_ff, config.d_model, bias=False)
        self.dropout = nn.Dropout(config.dropout_rate)
        self.gelu_act = ACT2FN["gelu_new"]

    def forward(self, hidden_states):
        hidden_gelu = self.gelu_act(self.wi_0(hidden_states))
        hidden_linear = self.wi_1(hidden_states)
        hidden_states = hidden_gelu * hidden_linear
        hidden_states = self.dropout(hidden_states)
        hidden_states = self.wo(hidden_states)
        return hidden_states

```

Now it's easy to see the `dropout` call, and all the previous calls as well.

Since the detection is happening in a `forward` hook, these reports are printed immediately after each `forward` returns.

Going back to the full report, to act on it and to fix the problem, we need to go a few frames up where the numbers started to go up and most likely switch to the `fp32` mode here, so that the numbers don't overflow when multiplied or summed up. Of course, there might be other solutions. For example, we could turn off `amp` temporarily if it's enabled, after moving the original `forward` into a helper wrapper, like so:

```
import torch
```

```

def _forward(self, hidden_states):
    hidden_gelu = self.gelu_act(self.wi_0(hidden_states))
    hidden_linear = self.wi_1(hidden_states)
    hidden_states = hidden_gelu * hidden_linear
    hidden_states = self.dropout(hidden_states)
    hidden_states = self.wo(hidden_states)
    return hidden_states

def forward(self, hidden_states):
    if torch.is_autocast_enabled():
        with torch.cuda.amp.autocast(enabled=False):
            return self._forward(hidden_states)
    else:
        return self._forward(hidden_states)

```

Since the automatic detector only reports on inputs and outputs of full frames, once you know where to look, you may want to analyse the intermediary stages of any specific `forward` function as well. In such a case you can use the `detect_overflow` helper function to inject the detector where you want it, for example:

```

from underflow_overflow import detect_overflow

class T5LayerFF(nn.Module):
    [...]

    def forward(self, hidden_states):
        forwarded_states = self.layer_norm(hidden_states)
        detect_overflow(forwarded_states, "after layer_norm")
        forwarded_states = self.DenseReluDense(forwarded_states)
        detect_overflow(forwarded_states, "after DenseReluDense")
        return hidden_states + self.dropout(forwarded_states)

```

You can see that we added 2 of these and now we track if `inf` or `nan` for `forwarded_states` was detected somewhere in between.

Actually, the detector already reports these because each of the calls in the example above is a `nn.Module`, but let's say if you had some local direct calculations this is how you'd do that.

Additionally, if you're instantiating the debugger in your own code, you can adjust the number of frames printed from its default, e.g.:

```

from .underflow_overflow import DebugUnderflowOverflow

debug_overflow = DebugUnderflowOverflow(model, max_frames_to_save=100)

```

Specific batch absolute mix and max value tracing

The same debugging class can be used for per-batch tracing with the underflow/overflow detection feature turned off.

Let's say you want to watch the absolute min and max values for all the ingredients of each `forward` call of a given batch, and only do that for batches 1 and 3. Then you instantiate this class as:

```
debug_overflow = DebugUnderflowOverflow(model, trace_batch_nums=[1, 3])
```

And now full batches 1 and 3 will be traced using the same format as the underflow/overflow detector does.

Batches are 0-indexed.

This is helpful if you know that the program starts misbehaving after a certain batch number, so you can fast-forward right to that area. Here is a sample truncated output for such configuration:

```
*** Starting batch number=1 ***
abs min  abs max  metadata
                  shared Embedding
1.01e-06 7.92e+02 weight
0.00e+00 2.47e+04 input[0]
5.36e-05 7.92e+02 output
[...]
                  decoder.dropout Dropout
1.60e-07 2.27e+01 input[0]
0.00e+00 2.52e+01 output
                  decoder T5Stack
  not a tensor output
    lm_head Linear
1.01e-06 7.92e+02 weight
0.00e+00 1.11e+00 input[0]
6.06e-02 8.39e+01 output
                  T5ForConditionalGeneration
  not a tensor output

*** Starting batch number=3 ***
abs min  abs max  metadata
                  shared Embedding
1.01e-06 7.92e+02 weight
0.00e+00 2.78e+04 input[0]
5.36e-05 7.92e+02 output
[...]
```

Here you will get a huge number of frames dumped - as many as there were forward calls in your model, so it may or may not what you want, but sometimes it can be easier to use for debugging purposes than a normal debugger. For example, if a problem starts happening at batch number 150. So you can dump traces for batches 149 and 150 and compare where numbers started to diverge.

You can also specify the batch number after which to stop the training, with:

```
debug_overflow = DebugUnderflowOverflow(model, trace_batch_nums=[1, 3], abort_after_batch_num=3)
```

Faster debug and development with tiny models, tokenizers and datasets

If you're debugging problems and develop with full sized models and tokenizers you're likely not working in a very efficient way. Not only it's much more difficult to solve problem, the amount of waiting to get the program to restart and to get to the desirable point can be huge - and cumulatively this can be a huge drain on one's motivation and productivity, not talking about the resolution taking much longer, if at all.

The solution is simple:

Unless you're testing the quality of a model, always use a tiny random model with potentially tiny tokenizer.

Moreover, large models often require massive resources, which are typically expensive and can also make a debugging process super complicated. For example any debugger can handle a single process, but if your model doesn't fit and require some sort of [parallelization](#) that requires multiple processes - most debuggers will either break or have issue giving you what you need. The ideal development environment is one process and a tiny model is guaranteed to fit on an even cheapest single smallest consumer GPU available. You could even use the free [Google Colab](#) to do development in a pinch if you have no GPUs around.

So the updated ML development mantra then becomes:

- the larger the model the better the final product generates
- the smaller the model the quicker the final product's training can be started

footnote: the recent research shows that larger isn't always better, but it's good enough to convey the importance of my communication.

Once your code is working, do switch to the real model to test the quality of your generation. But even in this case still try first the smallest model that produces a quality result. Only when you can see that the generation is mostly right use the largest model to validate if your work has been perfect.

Making a tiny model

Important: given their popularity and the well designed simple API I will be discussing HF [transformers](#) models. But the same principle can be applied to any other model.

TLDR: it's trivial to make a tiny HF [transformers](#) model:

1. Fetch the config object of a full size model
2. Shrink the hidden size and perhaps a few other parameters that contribute to the bulk of the model
3. Create a model from that shrunken config
4. Save this model. Done!

footnote: It's critical to remember that this will generate a random model, so don't expect any quality from its output.

footnote: These notes were written with HF Transformers models in mind. If you're using a different modeling library you may have to adapt some of these things.

Now let's go through the actual code and convert ["google/mt5-small"](#) into its tiny random counterpart.

```
from transformers import MT5Config, MT5ForConditionalGeneration

mname_from = "google/mt5-small"
```

```

mname_very_small = "mt5-tiny-random"

config = MT5Config.from_pretrained(mname_from)

config.update(dict(
    d_model=64,
    d_ff=256,
))
print("new config", config)

very_small_model = MT5ForConditionalGeneration(config)
print(f"num of params {very_small_model.num_parameters()}")

very_small_model.save_pretrained(mname_very_small)

```

As you can see it's trivial to do. And you can make it even smaller if you don't need the hidden size to be at least 64. For example try 8 - you just need to make sure that the number of attention heads isn't larger than hidden size.

Also please note that you don't need any GPUs to do that and you could do this even on a huge 176B parameter model like [BLOOM-176B](#). Since you never load the actual original model, except its config object.

Before modifying the config you can dump the original parameters and choose to shrinks more dimensions. For example, using less layers makes it even smaller and easier to debug. So here is what you can do instead:

```

config.update(dict(
    d_model=64,
    d_ff=256,
    d_kv=8,
    num_layers=8,
    num_decoder_layers=8,
    num_heads=4,
    relative_attention_num_buckets=32,
))

```

The original ["google/mt5-small"](#) model file was 1.2GB. With the above changes (and vocab shrinking as explained in the following sections) we got it down to 126MB.

If you're dealing with a multi-level nested config, you will have to update each sub-level's config object separately. For example in [IDEFICS](#) we have 1 main and 2 nested objects:

```

config
config.perceiver_config
config.vision_config

```

If you wanted to shrink this model you'd want to update `config` and `config.vision_config` with smaller values:

```

config.update(dict(
    hidden_size=64,

```

```

        intermediate_size=37,
        num_hidden_layers=5,
        num_attention_heads=4,
        max_position_embeddings=64,
        max_sequence_length=64,
    ))
# sub object needs to be updated directly
config.vision_config.update(dict(embed_dim=64))

```

See [idefics-make-tiny-model.py](#) for a fully working script (I didn't bother adding the vocab shrinking as I'm just demonstrating how to update nested config objects here).

We can then further halve our tiny model size by converting the model to fp16 or bf16 (depending on the goal) before saving it:

```

very_small_model.half() # convert to fp16
#very_small_model.bfloat16() # convert to bf16
very_small_model.save_pretrained(mname_very_small)

```

this takes us to 64M file.

So you could stop here and your program will start much much faster already.

And there is one more step you could do to make it truly tiny.

What we haven't shrunken so far is the vocabulary dimension so $64 \times 250k$ (hidden*vocab) is still huge. Granted this 250k vocab model is not typical - normally models' vocab is $\sim 30-50k$, but even 30k is a lot if we want the model to be truly tiny.

So next we will look into various techniques to shrinking the tokenizer, as it defines our vocab size.

Making a tiny tokenizer

This task varies between a relatively simple procedure and a much more complex workout depending on the underlying tokenizer.

The following recipes have come from a few awesome tokenizer experts at Hugging Face, which I then adapted to my needs.

You probably don't really need to understand how these work until you actually need them, therefore if you're reading this for the first time you can safely jump over these to [Making a tiny model with a tiny tokenizer](#).

Anthony Moi's version

[Anthony Moi](#)'s tokenizer shrinker:

```

import json
from transformers import AutoTokenizer
from tokenizers import Tokenizer

vocab_keep_items = 5000

```

```

mname = "microsoft/deberta-base"

tokenizer = AutoTokenizer.from_pretrained(mname, use_fast=True)
assert tokenizer.is_fast, "This only works for fast tokenizers."
tokenizer_json = json.loads(tokenizer._tokenizer.to_str())
vocab = tokenizer_json["model"]["vocab"]
if tokenizer_json["model"]["type"] == "BPE":
    new_vocab = { token: i for token, i in vocab.items() if i < vocab_keep_items }
    merges = tokenizer_json["model"]["merges"]
    new_merges = []
    for i in range(len(merges)):
        a, b = merges[i].split()
        new_token = "".join((a, b))
        if a in new_vocab and b in new_vocab and new_token in new_vocab:
            new_merges.append(merges[i])
    tokenizer_json["model"]["merges"] = new_merges
elif tokenizer_json["model"]["type"] == "Unigram":
    new_vocab = vocab[:vocab_keep_items]
elif tokenizer_json["model"]["type"] == "WordPiece" or tokenizer_json["model"]["type"] == "WordLevel":
    new_vocab = { token: i for token, i in vocab.items() if i < vocab_keep_items }
else:
    raise ValueError(f"don't know how to handle {tokenizer_json['model']['type']}")
tokenizer_json["model"]["vocab"] = new_vocab
tokenizer._tokenizer = Tokenizer.from_str(json.dumps(tokenizer_json))
tokenizer.save_pretrained(".")

```

I later discovered that gpt2 seems to have a special token "<|endoftext|>" stashed at the very end of the vocab, so it gets dropped and code breaks. So I hacked it back in with:

```

if "gpt2" in mname:
    new_vocab = { token: i for token, i in vocab.items() if i < vocab_keep_items-1 }
    new_vocab["<|endoftext|>"] = vocab_keep_items-1
else:
    new_vocab = { token: i for token, i in vocab.items() if i < vocab_keep_items }

```

Lysandre Debut's version

[Lysandre Debut](#)' shrinker using `train_new_from_iterator`:

```

from transformers import AutoTokenizer

mname = "microsoft/deberta-base" # or any checkpoint that has a fast tokenizer.
vocab_keep_items = 5000

tokenizer = AutoTokenizer.from_pretrained(mname)
assert tokenizer.is_fast, "This only works for fast tokenizers."
tokenizer.save_pretrained("big-tokenizer")
# Should be a generator of list of texts.

```

```

training_corpus = [
    ["This is the first sentence.", "This is the second one."],
    ["This sentence (contains #) over symbols and numbers 12 3.", "But not this one."],
]
new_tokenizer = tokenizer.train_new_from_iterator(training_corpus, vocab_size=vocab_keep_items)
new_tokenizer.save_pretrained("small-tokenizer")

```

but this one requires a training corpus, so I had an idea to cheat and train the new tokenizer on its own original vocab which gave me:

```

from transformers import AutoTokenizer

mname = "microsoft/deberta-base"
vocab_keep_items = 5000

tokenizer = AutoTokenizer.from_pretrained(mname)
assert tokenizer.is_fast, "This only works for fast tokenizers."
vocab = tokenizer.get_vocab()
training_corpus = [vocab.keys()] # Should be a generator of list of texts.
new_tokenizer = tokenizer.train_new_from_iterator(training_corpus, vocab_size=vocab_keep_items)
new_tokenizer.save_pretrained("small-tokenizer")

```

which is almost perfect, except it now doesn't have any information about the frequency for each word/char (that's how most tokenizers compute their vocab, which if you need this info you can fix by having each key appearing `len(vocab)` - 1 times, i.e.:

```
training_corpus = [ (k for i in range(vocab_len-v)) for k,v in vocab.items() ]
```

which will make the script much much longer to complete.

But for the needs of a tiny model (testing) the frequency doesn't matter at all.

Hack the tokenizer file approach

Some tokenizers can be be just manually truncated at the file level, e.g. let's shrink Llama2's tokenizer to 3k items:

```

# Shrink the orig vocab to keep things small (just enough to tokenize any word, so letters+symbols)
# ElectraTokenizerFast is fully defined by a tokenizer.json, which contains the vocab and the ids,
# so we just need to truncate it wisely
import subprocess
import shlex
from transformers import LlamaTokenizerFast

mname = "meta-llama/Llama-2-7b-hf"
vocab_keep_items = 3000

tokenizer_fast = LlamaTokenizerFast.from_pretrained(mname)

```

```

tmp_dir = f"/tmp/{mname}"
tokenizer_fast.save_pretrained(tmp_dir)
# resize tokenizer.json (vocab.txt will be automatically resized on save_pretrained)
# perl -0777 -pi -e 's|(2999).*$1|,"merges": []}|msg' tokenizer.json # 0-indexed, so
vocab_keep_items-1!
closing_pat = '},"merges": []}''
cmd = (f"perl -0777 -pi -e 's|({vocab_keep_items-1}).*|${1{closing_pat}}|msg' {tmp_dir}/tokenizer.json")
#print(f"Running:\n{cmd}")
result = subprocess.run(shlex.split(cmd), capture_output=True, text=True)
# reload with modified tokenizer
tokenizer_fast_tiny = LlamaTokenizerFast.from_pretrained(tmp_dir)
tokenizer_fast_tiny.save_pretrained(".")

```

Please remember that the outcome is only useful for functional testing - not quality work.

Here is the full version of [make_tiny_model.py](#) which includes both the model and the tokenizer shrinking.

SentencePiece vocab shrinking

First clone SentencePiece into a parent dir:

```
git clone https://github.com/google/sentencepiece
```

Now to the shrinking:

```

# workaround for fast tokenizer protobuf issue, and it's much faster too!
os.environ["PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION"] = "python"

from transformers import XLMRobertaTokenizerFast

mname = "xlm-roberta-base"

# Shrink the orig vocab to keep things small
vocab_keep_items = 5000
tmp_dir = f"/tmp/{mname}"
vocab_orig_path = f"{tmp_dir}/sentencepiece.bpe.model" # this name can be different
vocab_short_path = f"{tmp_dir}/spiece-short.model"
# HACK: need the sentencepiece source to get sentencepiece_model_pb2, as it doesn't get installed
sys.path.append("../sentencepiece/python/src/sentencepiece")
import sentencepiece_model_pb2 as model
tokenizer_orig = XLMRobertaTokenizerFast.from_pretrained(mname)
tokenizer_orig.save_pretrained(tmp_dir)
with open(vocab_orig_path, 'rb') as f: data = f.read()
# adapted from https://blog.ceshine.net/post/trim-down-sentencepiece-vocabulary/
m = model.ModelProto()
m.ParseFromString(data)
print(f"Shrinking vocab from original {len(m.pieces)} dict items")
for i in range(len(m.pieces) - vocab_keep_items): _ = m.pieces.pop()
print(f"new dict {len(m.pieces)}")

```

```

with open(vocab_short_path, 'wb') as f: f.write(m.SerializeToString())
m = None

tokenizer_fast_tiny = XLMRobertaTokenizerFast(vocab_file=vocab_short_path)
tokenizer_fast_tiny.save_pretrained(".")


```

Making a tiny model with a tiny tokenizer

So now you can shrink the vocab size to as small as the tokenizer allows, that is you need to have at least enough tokens to cover the target alphabet and special characters, and usually 3-5k tokens is more than enough. Sometimes you could make it even smaller, after all the original ASCII charset has only 128 characters.

If we continue the MT5 code from earlier in this chapter and add the tokenizer shrinking code from the previous section, we end up with this script [mt5-make-tiny-model.py](#) and when we run it - our end model file is truly tiny - 3.34 MB in size! As you can see the script also has code to validate that the model can actually work with the modified tokenizer. The results will be garbage, but the intention is to test that the new model and the tokenizer are functional.

Here is another example [fsmt-make-super-tiny-model.py](#) - here you can see I'm creating a totally new tiny vocab from scratch.

I also recommend to always store the building scripts with the model, so that you could quickly fix things or make similar versions of the model.

Also be aware that since HF `transformers` needs tiny models for their testing, you are very likely to already find one for each architecture available mostly from <https://huggingface.co/hf-internal-testing> (except they didn't include the code of how they were made, but you can now figure it out based on these notes).

Another hint: if you need a slightly different tiny model, you can also start with an already existing tiny model and adapt it instead. Since it's random it's really only about getting the right dimensions. For example if the tiny model you found has 2 layers but you need 8, just resave it with this larger dimension and you're done.

Making a tiny dataset

Similar to models and tokenizers it helps to have a handy tiny version of a dataset you work with a lot. As usual this won't help with quality testing, but it's perfect for launching your program really fast.

footnote: the impact of using a tiny dataset won't be as massive as using a tiny model, if you're using already pre-indexed Arrow file datasets, since those are already extremely fast. But say you want the iterator to finish an epoch in 10 steps. Instead of editing your code to truncate the dataset, you could just use a tiny dataset instead.

This process of making a tiny dataset is somewhat more difficult to explain because it'd depend on the builder of the original model, which can be quite different from each other, but perhaps you can correlate my recipes to your datasets.

But the concept is still very simple:

1. Clone the full dataset git repo
2. Replace its full data tarball with a tiny one that contains just a few samples
3. Save it - Done!

Here are some examples:

- [stas/oscar-en-10k](#)
- [stas/c4-en-10k](#)
- [stas/openwebtext-10k](#)

In all of these I took the original tarball, grabbed the first 10k records, tarred it back, used this smaller tarball and that was

that. The rest of the builder script remained mostly the same.

And here are some examples of synthetic datasets, where instead of just shrinking the original tarball, I untar'ed it, manually chose the representative examples and then wrote a script to build any size of desired dataset based on those few representative samples:

- [stas/general-pmd-synthetic-testing](#) and the [unpacker](#)
- [stas/cm4-synthetic-testing](#) - and the [unpacker](#)

These are also the complex examples where each sample is more than a text entry, but may have multiple text entries and images as well.

The unpacker is what expands each complex multi-record sample into its own sub-directory, so that now you can easily go and tweak it to your liking. You can add image, remove them, make text records smaller, etc.. You will also notice that I'm shrinking the large images into tiny 32x32 images, so again I'm applying the important principle of tiny across all dimensions that don't break the requirements of the target codebase.

And then the main script uses that structure to build a dataset of any desired length.

And here is for example the instructions of deploying these scripts for [stas/general-pmd-synthetic-testing](#):

```
# prep dataset repo
https://huggingface.co/new-dataset => stas/general-pmd-synthetic-testing
git clone https://huggingface.co/datasets/stas/general-pmd-synthetic-testing
cd general-pmd-synthetic-testing

# select a few seed records so there is some longer and shorter text, records with images and without,
# a few variations of each type
rm -rf data
python general-pmd-ds-unpack.py --dataset_name_or_path \
general_pmd/image/localized_narratives__ADE20k/train/00000-00002 --ids 1-10 --target_path data

cd data

# shrink to 32x32 max, keeping ratio
mogrify -format jpg -resize 32x32\> /*/*jpg

# adjust one record to have no image and no text
cd 1
rm image.jpg text.txt
touch image.null text.null
cd -

cd ..

# create tarball
tar -cvzf data.tar.gz data

# complete the dataset repo
echo "This dataset is designed to be used in testing. It's derived from general-pmd/
localized_narratives__ADE20k \
dataset" >> README.md

# test dataset
```

```
cd ..  
datasets-cli test general-pmd-synthetic-testing/general-pmd-synthetic-testing.py --all_configs
```

I also recommend to always store the building scripts with the dataset, so that you could quickly fix things or make similar versions of the dataset.

Similar to tiny models, you will find many tiny datasets under <https://huggingface.co/hf-internal-testing>.

Conclusion

While in the domain of ML we have the dataset, the model and the tokenizer - each of which can be made tiny and enable super-speed development with low resource requirements, if you're coming from a different industry you can adapt the ideas discussed in this chapter to your particular domain's artifacts/payloads.

Backup of all scripts in this chapter

Should the original scripts this chapter is pointing to disappear or the HF hub is down while you're reading this, here is [the local back up of all of them](#).

note-to-self: to make the latest backup of files linked to in this chapter run:

```
perl -lne 'while (/^(https.*?.py)\//g) { $x=$1; $x=~s/blob/raw/; print qq[wget $x] }' make-tiny-models.md
```

A Back up of scripts

This is a backup of scripts discussed in [Faster debug and development with tiny models, tokenizers and datasets](#).

- [c4-en-10k.py](#)
- [cm4-synthetic-testing.py](#)
- [fsmt-make-super-tiny-model.py](#)
- [general-pmd-ds-unpack.py](#)
- [general-pmd-synthetic-testing.py](#)
- [m4-ds-unpack.py](#)
- [mt5-make-tiny-model.py](#)
- [openwebtext-10k.py](#)
- [oscar-en-10k.py](#)

Writing and Running Tests

Note: a part of this document refers to functionality provided by the included [testing_utils.py](#), the bulk of which I have developed while I worked at HuggingFace.

This document covers both `pytest` and `unittest` functionalities and shows how both can be used together.

Running tests

Run all tests

```
pytest
```

I use the following alias:

```
alias pyt="pytest --disable-warnings --instafail -rA"
```

which tells `pytest` to:

- disable warning
- `--instafail` shows failures as they happen, and not at the end
- `-rA` generates a short test summary info

it requires you to install:

```
pip install pytest-instafail
```

Getting the list of all tests

Show all tests in the test suite:

```
pytest --collect-only -q
```

Show all tests in a given test file:

```
pytest tests/test_optimization.py --collect-only -q
```

I use the following alias:

```
alias pytc="pytest --disable-warnings --collect-only -q"
```

Run a specific test module

To run an individual test module:

```
pytest tests/utils/test_logging.py
```

Run specific tests

If `unittest` is used, to run specific subtests you need to know the name of the `unittest` class containing those tests. For example, it could be:

```
pytest tests/test_optimization.py::OptimizationTest::test_adam_w
```

Here:

- `tests/test_optimization.py` - the file with tests
- `OptimizationTest` - the name of the test class
- `test_adam_w` - the name of the specific test function

If the file contains multiple classes, you can choose to run only tests of a given class. For example:

```
pytest tests/test_optimization.py::OptimizationTest
```

will run all the tests inside that class.

As mentioned earlier you can see what tests are contained inside the `OptimizationTest` class by running:

```
pytest tests/test_optimization.py::OptimizationTest --collect-only -q
```

You can run tests by keyword expressions.

To run only tests whose name contains `adam`:

```
pytest -k adam tests/test_optimization.py
```

Logical `and` and `or` can be used to indicate whether all keywords should match or either. `not` can be used to negate.

To run all tests except those whose name contains `adam`:

```
pytest -k "not adam" tests/test_optimization.py
```

And you can combine the two patterns in one:

```
pytest -k "ada and not adam" tests/test_optimization.py
```

For example to run both `test_adafactor` and `test_adam_w` you can use:

```
pytest -k "test_adafactor or test_adam_w" tests/test_optimization.py
```

Note that we use `or` here, since we want either of the keywords to match to include both.

If you want to include only tests that include both patterns, `and` is to be used:

```
pytest -k "test and ada" tests/test_optimization.py
```

Run only modified tests

You can run the tests related to the unstaged files or the current branch (according to Git) by using [pytest-picked](#). This is a great way of quickly testing your changes didn't break anything, since it won't run the tests related to files you didn't touch.

```
pip install pytest-picked
```

```
pytest --picked
```

All tests will be run from files and folders which are modified, but not yet committed.

Automatically rerun failed tests on source modification

[pytest-xdist](#) provides a very useful feature of detecting all failed tests, and then waiting for you to modify files and continuously re-rerun those failing tests until they pass while you fix them. So that you don't need to re start pytest after you made the fix. This is repeated until all tests pass after which again a full run is performed.

```
pip install pytest-xdist
```

To enter the mode: `pytest -f` or `pytest --looponfail`

File changes are detected by looking at `looponfailroots` root directories and all of their contents (recursively). If the default for this value does not work for you, you can change it in your project by setting a configuration option in `setup.cfg`:

```
[tool:pytest]
looponfailroots = transformers tests
```

or `pytest.ini`/`tox.ini` files:

```
[pytest]
looponfailroots = transformers tests
```

This would lead to only looking for file changes in the respective directories, specified relatively to the ini-file's directory. [pytest-watch](#) is an alternative implementation of this functionality.

Skip a test module

If you want to run all test modules, except a few you can exclude them by giving an explicit list of tests to run. For example, to run all except `test_modeling_*.py` tests:

```
pytest $(ls -1 tests/*py | grep -v test_modeling)
```

Clearing state

CI builds and when isolation is important (against speed), cache should be cleared:

```
pytest --cache-clear tests
```

Running tests in parallel

As mentioned earlier `make test` runs tests in parallel via `pytest-xdist` plugin (`-n x` argument, e.g. `-n 2` to run 2 parallel jobs).

`pytest-xdist`'s `--dist=` option allows one to control how the tests are grouped. `--dist=loadfile` puts the tests located in one file onto the same process.

Since the order of executed tests is different and unpredictable, if running the test suite with `pytest-xdist` produces failures (meaning we have some undetected coupled tests), use [pytest-replay](#) to replay the tests in the same order, which should help with then somehow reducing that failing sequence to a minimum.

Test order and repetition

It's good to repeat the tests several times, in sequence, randomly, or in sets, to detect any potential inter-dependency and state-related bugs (tear down). And the straightforward multiple repetition is just good to detect some problems that get uncovered by randomness of DL.

Repeat tests

- [pytest-flakefinder](#):

```
pip install pytest-flakefinder
```

And then run every test multiple times (50 by default):

```
pytest --flake-finder --flake-runs=5 tests/test_failing_test.py
```

footnote: This plugin doesn't work with `-n` flag from `pytest-xdist`.

footnote: There is another plugin `pytest-repeat`, but it doesn't work with `unittest`.

Run tests in a random order

```
pip install pytest-random-order
```

Important: the presence of `pytest-random-order` will automatically randomize tests, no configuration change or command line options is required.

As explained earlier this allows detection of coupled tests - where one test's state affects the state of another. When `pytest-random-order` is installed it will print the random seed it used for that session, e.g.:

```
pytest tests
[...]
Using --random-order-bucket=module
Using --random-order-seed=573663
```

So that if the given particular sequence fails, you can reproduce it by adding that exact seed, e.g.:

```
pytest --random-order-seed=573663
[...]
Using --random-order-bucket=module
Using --random-order-seed=573663
```

It will only reproduce the exact order if you use the exact same list of tests (or no list at all). Once you start to manually narrowing down the list you can no longer rely on the seed, but have to list them manually in the exact order they failed and tell pytest to not randomize them instead using `--random-order-bucket=none`, e.g.:

```
pytest --random-order-bucket=none tests/test_a.py tests/test_c.py tests/test_b.py
```

To disable the shuffling for all tests:

```
pytest --random-order-bucket=none
```

By default `--random-order-bucket=module` is implied, which will shuffle the files on the module levels. It can also shuffle on `class`, `package`, `global` and `none` levels. For the complete details please see its [documentation](#).

Another randomization alternative is: [pytest-randomly](#). This module has a very similar functionality/interface, but it doesn't have the bucket modes available in `pytest-random-order`. It has the same problem of imposing itself once installed.

Look and feel variations

pytest-sugar

[pytest-sugar](#) is a plugin that improves the look-n-feel, adds a progressbar, and show tests that fail and the assert instantly. It gets activated automatically upon installation.

```
pip install pytest-sugar
```

To run tests without it, run:

```
pytest -p no:sugar
```

or uninstall it.

Report each sub-test name and its progress

For a single or a group of tests via `pytest` (after `pip install pytest-pspec`):

```
pytest --pspec tests/test_optimization.py
```

Instantly shows failed tests

[pytest-instafail](#) shows failures and errors instantly instead of waiting until the end of test session.

```
pip install pytest-instafail
```

```
pytest --instafail
```

To GPU or not to GPU

On a GPU-enabled setup, to test in CPU-only mode add `CUDA_VISIBLE_DEVICES=""`:

```
CUDA_VISIBLE_DEVICES="" pytest tests/utils/test_logging.py
```

or if you have multiple gpus, you can specify which one is to be used by `pytest`. For example, to use only the second gpu if you have gpus 0 and 1, you can run:

```
CUDA_VISIBLE_DEVICES="1" pytest tests/utils/test_logging.py
```

This is handy when you want to run different tasks on different GPUs.

Some tests must be run on CPU-only, others on either CPU or GPU or TPU, yet others on multiple-GPUs. The following skip decorators are used to set the requirements of tests CPU/GPU/TPU-wise:

- `require_torch` - this test will run only under torch
- `require_torch_gpu` - as `require_torch` plus requires at least 1 GPU
- `require_torch_multi_gpu` - as `require_torch` plus requires at least 2 GPUs
- `require_torch_non_multi_gpu` - as `require_torch` plus requires 0 or 1 GPUs
- `require_torch_up_to_2_gpus` - as `require_torch` plus requires 0 or 1 or 2 GPUs
- `require_torch_tpu` - as `require_torch` plus requires at least 1 TPU

Let's depict the GPU requirements in the following table:

n gpus	decorator
≥ 0	<code>@require_torch</code>
≥ 1	<code>@require_torch_gpu</code>
≥ 2	<code>@require_torch_multi_gpu</code>
< 2	<code>@require_torch_non_multi_gpu</code>
< 3	<code>@require_torch_up_to_2_gpus</code>

For example, here is a test that must be run only when there are 2 or more GPUs available and pytorch is installed:

```
from testing_utils import require_torch_multi_gpu

@require_torch_multi_gpu
def test_example_with_multi_gpu():
```

These decorators can be stacked:

```
from testing_utils import require_torch_gpu

@require_torch_gpu
@some_other_decorator
def test_example_slow_on_gpu():
```

Some decorators like `@parametrized` rewrite test names, therefore `@require_*` skip decorators have to be listed last for them to work correctly. Here is an example of the correct usage:

```
from testing_utils import require_torch_multi_gpu
from parameterized import parameterized

@parameterized.expand(...)
@require_torch_multi_gpu
def test_integration_foo():
```

This order problem doesn't exist with `@pytest.mark.parametrize`, you can put it first or last and it will still work. But it only works with non-unittests.

Inside tests:

- How many GPUs are available:

```
from testing_utils import get_gpu_count
```

```
n_gpu = get_gpu_count()
```

Distributed training

`pytest` can't deal with distributed training directly. If this is attempted - the sub-processes don't do the right thing and end up thinking they are `pytest` and start running the test suite in loops. It works, however, if one spawns a normal process that then spawns off multiple workers and manages the IO pipes.

Here are some tests that use it:

- [test_trainer_distributed.py](#)
- [test_deepspeed.py](#)

To jump right into the execution point, search for the `execute_subprocess_async` call in those tests, which you will find inside [testing_utils.py](#).

You will need at least 2 GPUs to see these tests in action:

```
CUDA_VISIBLE_DEVICES=0,1 RUN_SLOW=1 pytest -sv tests/test_trainer_distributed.py
```

(`RUN_SLOW` is a special decorator used by HF Transformers to normally skip heavy tests)

Output capture

During test execution any output sent to `stdout` and `stderr` is captured. If a test or a setup method fails, its according captured output will usually be shown along with the failure traceback.

To disable output capturing and to get the `stdout` and `stderr` normally, use `-s` or `--capture=no`:

```
pytest -s tests/utils/test_logging.py
```

To send test results to JUnit format output:

```
py.test tests --junitxml=result.xml
```

Color control

To have no color (e.g., yellow on white background is not readable):

```
pytest --color=no tests/utils/test_logging.py
```

Sending test report to online pastebin service

Creating a URL for each test failure:

```
pytest --pastebin=failed tests/utils/test_logging.py
```

This will submit test run information to a remote Paste service and provide a URL for each failure. You may select tests as usual or add for example `-x` if you only want to send one particular failure.

Creating a URL for a whole test session log:

```
pytest --pastebin=all tests/utils/test_logging.py
```

Writing tests

Most of the time if combining `pytest` and `unittest` in the same test suite works just fine. You can read [here](#) which features are supported when doing that, but the important thing to remember is that most `pytest` fixtures don't work. Neither parametrization, but we use the module `parameterized` that works in a similar way.

Parametrization

Often, there is a need to run the same test multiple times, but with different arguments. It could be done from within the test, but then there is no way of running that test for just one set of arguments.

```
# test_this1.py
import unittest
from parameterized import parameterized

class TestMathUnitTest(unittest.TestCase):
    @parameterized.expand(
        [
            ("negative", -1.5, -2.0),
            ("integer", 1, 1.0),
            ("large fraction", 1.6, 1),
        ]
    )
    def test_floor(self, name, input, expected):
        assert_equal(math.floor(input), expected)
```

Now, by default this test will be run 3 times, each time with the last 3 arguments of `test_floor` being assigned the corresponding arguments in the parameter list.

And you could run just the `negative` and `integer` sets of params with:

```
pytest -k "negative and integer" tests/test_mytest.py
```

or all but `negative` sub-tests, with:

```
pytest -k "not negative" tests/test_mytest.py
```

Besides using the `-k` filter that was just mentioned, you can find out the exact name of each sub-test and run any or all of them using their exact names.

```
pytest test_this1.py --collect-only -q
```

and it will list:

```
test_this1.py::TestMathUnitTest::test_floor_0_negative
test_this1.py::TestMathUnitTest::test_floor_1_integer
test_this1.py::TestMathUnitTest::test_floor_2_large_fraction
```

So now you can run just 2 specific sub-tests:

```
pytest test_this1.py::TestMathUnitTest::test_floor_0_negative
test_this1.py::TestMathUnitTest::test_floor_1_integer
```

The module [parameterized](#) works for both: `unittests` and `pytest` tests.

If, however, the test is not a `unittest`, you may use `pytest.mark.parametrize`.

Here is the same example, this time using `pytest`'s `parametrize` marker:

```
# test_this2.py
import pytest

@pytest.mark.parametrize(
    "name, input, expected",
    [
        ("negative", -1.5, -2.0),
        ("integer", 1, 1.0),
        ("large fraction", 1.6, 1),
    ],
)
def test_floor(name, input, expected):
    assert_equal(math.floor(input), expected)
```

Same as with `parameterized`, with `pytest.mark.parametrize` you can have a fine control over which sub-tests are run, if the `-k` filter doesn't do the job. Except, this parametrization function creates a slightly different set of names for the sub-tests. Here is what they look like:

```
pytest test_this2.py --collect-only -q
```

and it will list:

```
test_this2.py::test_floor[integer-1-1.0]
test_this2.py::test_floor[negative--1.5--2.0]
```

```
test_this2.py::test_floor[large fraction-1.6-1]
```

So now you can run just the specific test:

```
pytest test_this2.py::test_floor[negative--1.5--2.0] test_this2.py::test_floor[integer-1-1.0]
```

as in the previous example.

Files and directories

In tests often we need to know where things are relative to the current test file, and it's not trivial since the test could be invoked from more than one directory or could reside in sub-directories with different depths. A helper class `testing_utils.TestCasePlus` solves this problem by sorting out all the basic paths and provides easy accessors to them:

- `pathlib` objects (all fully resolved):
 - `test_file_path` - the current test file path, i.e. `__file__`
 - `test_file_dir` - the directory containing the current test file
 - `tests_dir` - the directory of the `tests` test suite
 - `examples_dir` - the directory of the `examples` test suite
 - `repo_root_dir` - the directory of the repository
 - `src_dir` - the directory of `src` (i.e. where the `transformers` sub-dir resides)
- stringified paths -- same as above but these return paths as strings, rather than `pathlib` objects:
 - `test_file_path_str`
 - `test_file_dir_str`
 - `tests_dir_str`
 - `examples_dir_str`
 - `repo_root_dir_str`
 - `src_dir_str`

To start using those all you need is to make sure that the test resides in a subclass of `testing_utils.TestCasePlus`. For example:

```
from testing_utils import TestCasePlus

class PathExampleTest(TestCasePlus):
    def test_something_involving_local_locations(self):
        data_dir = self.tests_dir / "fixtures/tests_samples/wmt_en_ro"
```

If you don't need to manipulate paths via `pathlib` or you just need a path as a string, you can always invoked `str()` on the `pathlib` object or use the accessors ending with `_str`. For example:

```
from testing_utils import TestCasePlus

class PathExampleTest(TestCasePlus):
```

```
def test_something_involving_stringified_locations(self):
    examples_dir = self.examples_dir_str
```

Temporary files and directories

Using unique temporary files and directories are essential for parallel test running, so that the tests won't overwrite each other's data. Also we want to get the temporary files and directories removed at the end of each test that created them. Therefore, using packages like `tempfile`, which address these needs is essential.

However, when debugging tests, you need to be able to see what goes into the temporary file or directory and you want to know it's exact path and not having it randomized on every test re-run.

A helper class `testing_utils.TestCasePlus` is best used for such purposes. It's a sub-class of `unittest.TestCase`, so we can easily inherit from it in the test modules.

Here is an example of its usage:

```
from testing_utils import TestCasePlus

class ExamplesTests(TestCasePlus):
    def test_whatever(self):
        tmp_dir = self.get_auto_remove_tmp_dir()
```

This code creates a unique temporary directory, and sets `tmp_dir` to its location.

- Create a unique temporary dir:

```
def test_whatever(self):
    tmp_dir = self.get_auto_remove_tmp_dir()
```

`tmp_dir` will contain the path to the created temporary dir. It will be automatically removed at the end of the test.

- Create a temporary dir of my choice, ensure it's empty before the test starts and don't empty it after the test.

```
def test_whatever(self):
    tmp_dir = self.get_auto_remove_tmp_dir("./xxx")
```

This is useful for debug when you want to monitor a specific directory and want to make sure the previous tests didn't leave any data in there.

- You can override the default behavior by directly overriding the `before` and `after` args, leading to one of the following behaviors:
 - `before=True`: the temporary dir will always be cleared at the beginning of the test.
 - `before=False`: if the temporary dir already existed, any existing files will remain there.
 - `after=True`: the temporary dir will always be deleted at the end of the test.
 - `after=False`: the temporary dir will always be left intact at the end of the test.

footnote: In order to run the equivalent of `rm -r` safely, only subdirs of the project repository checkout are allowed if an explicit `tmp_dir` is used, so that by mistake no `/tmp` or similar important part of the filesystem will get nuked. i.e. please

always pass paths that start with `./`.

footnote: Each test can register multiple temporary directories and they all will get auto-removed, unless requested otherwise.

Temporary `sys.path` override

If you need to temporary override `sys.path` to import from another test for example, you can use the `ExtendSysPath` context manager. Example:

```
import os
from testing_utils import ExtendSysPath

bindir = os.path.abspath(os.path.dirname(__file__))
with ExtendSysPath(f"{bindir}/.."):
    from test_trainer import TrainerIntegrationCommon  # noqa
```

Skipping tests

This is useful when a bug is found and a new test is written, yet the bug is not fixed yet. In order to be able to commit it to the main repository we need make sure it's skipped during `make test`.

Methods:

- A `skip` means that you expect your test to pass only if some conditions are met, otherwise pytest should skip running the test altogether. Common examples are skipping windows-only tests on non-windows platforms, or skipping tests that depend on an external resource which is not available at the moment (for example a database).
- A `xfail` means that you expect a test to fail for some reason. A common example is a test for a feature not yet implemented, or a bug not yet fixed. When a test passes despite being expected to fail (marked with `pytest.mark.xfail`), it's an `xpass` and will be reported in the test summary.

One of the important differences between the two is that `skip` doesn't run the test, and `xfail` does. So if the code that's buggy causes some bad state that will affect other tests, do not use `xfail`.

Implementation

- Here is how to skip whole test unconditionally:

```
@unittest.skip("this bug needs to be fixed")
def test_feature_x():
```

or via pytest:

```
@pytest.mark.skip(reason="this bug needs to be fixed")
```

or the `xfail` way:

```
@pytest.mark.xfail
def test_feature_x():
```

Here's how to skip a test based on internal checks within the test:

```
def test_feature_x():
    if not has_something():
        pytest.skip("unsupported configuration")
```

or the whole module:

```
import pytest

if not pytest.config.getvalue("--custom-flag"):
    pytest.skip("--custom-flag is missing, skipping tests", allow_module_level=True)
```

or the `xfail` way:

```
def test_feature_x():
    pytest.xfail("expected to fail until bug XYZ is fixed")
```

- Here is how to skip all tests in a module if some import is missing:

```
docutils = pytest.importorskip("docutils", minversion="0.3")
```

- Skip a test based on a condition:

```
@pytest.mark.skipif(sys.version_info < (3,6), reason="requires python3.6 or higher")
def test_feature_x():
```

or:

```
@unittest.skipIf(torch_device == "cpu", "Can't do half precision")
def test_feature_x():
```

or skip the whole module:

```
@pytest.mark.skipif(sys.platform == 'win32', reason="does not run on windows")
class TestClass():
    def test_feature_x(self):
```

More details, example and ways are [here](#).

Capturing outputs

Capturing the stdout/stderr output

In order to test functions that write to `stdout` and/or `stderr`, the test can access those streams using the `pytest`'s [capsys system](#). Here is how this is accomplished:

```
import sys

def print_to_stdout(s):
    print(s)

def print_to_stderr(s):
    sys.stderr.write(s)

def test_result_and_stdout(capsys):
    msg = "Hello"
    print_to_stdout(msg)
    print_to_stderr(msg)
    out, err = capsys.readouterr()  # consume the captured output streams
    # optional: if you want to replay the consumed streams:
    sys.stdout.write(out)
    sys.stderr.write(err)
    # test:
    assert msg in out
    assert msg in err
```

And, of course, most of the time, `stderr` will come as a part of an exception, so `try/except` has to be used in such a case:

```
def raise_exception(msg):
    raise ValueError(msg)

def test_something_exception():
    msg = "Not a good value"
    error = ""
    try:
        raise_exception(msg)
    except Exception as e:
        error = str(e)
    assert msg in error, f"{msg} is in the exception:\n{error}"
```

Another approach to capturing `stdout` is via `contextlib.redirect_stdout`:

```
from io import StringIO
```

```

from contextlib import redirect_stdout

def print_to_stdout(s):
    print(s)

def test_result_and_stdout():
    msg = "Hello"
    buffer = StringIO()
    with redirect_stdout(buffer):
        print_to_stdout(msg)
    out = buffer.getvalue()
    # optional: if you want to replay the consumed streams:
    sys.stdout.write(out)
    # test:
    assert msg in out

```

An important potential issue with capturing stdout is that it may contain \r characters that in normal print reset everything that has been printed so far. There is no problem with pytest, but with pytest -s these characters get included in the buffer, so to be able to have the test run with and without -s, you have to make an extra cleanup to the captured output, using `re.sub(r'~.*\r', '', buf, 0, re.M)`.

But, then we have a helper context manager wrapper to automatically take care of it all, regardless of whether it has some \r's in it or not, so it's a simple:

```

from testing_utils import CaptureStdout

with CaptureStdout() as cs:
    function_that_writes_to_stdout()
print(cs.out)

```

Here is a full test example:

```

from testing_utils import CaptureStdout

msg = "Secret message\r"
final = "Hello World"
with CaptureStdout() as cs:
    print(msg + final)
assert cs.out == final + "\n", f"captured: {cs.out}, expecting {final}"

```

If you'd like to capture stderr use the `CaptureStderr` class instead:

```

from testing_utils import CaptureStderr

with CaptureStderr() as cs:

```

```
function_that_writes_to_stderr()
print(cs.err)
```

If you need to capture both streams at once, use the parent `CaptureStd` class:

```
from testing_utils import CaptureStd

with CaptureStd() as cs:
    function_that_writes_to_stdout_and_stderr()
    print(cs.err, cs.out)
```

Also, to aid debugging test issues, by default these context managers automatically replay the captured streams on exit from the context.

Capturing logger stream

If you need to validate the output of a logger, you can use `CaptureLogger`:

```
from transformers import logging
from testing_utils import CaptureLogger

msg = "Testing 1, 2, 3"
logging.set_verbosity_info()
logger = logging.get_logger("transformers.models.bart.tokenization_bart")
with CaptureLogger(logger) as cl:
    logger.info(msg)
assert cl.out, msg + "\n"
```

Testing with environment variables

If you want to test the impact of environment variables for a specific test you can use a helper decorator `transformers.testing_utils.mockenv`

```
from testing_utils import mockenv

class HfArgumentParserTest(unittest.TestCase):
    @mockenv(TRANSFORMERS_VERTOSITY="error")
    def test_env_override(self):
        env_level_str = os.getenv("TRANSFORMERS_VERTOSITY", None)
```

At times an external program needs to be called, which requires setting `PYTHONPATH` in `os.environ` to include multiple local paths. A helper class `testing_utils.TestCasePlus` comes to help:

```
from testing_utils import TestCasePlus
```

```
class EnvExampleTest(TestCasePlus):
    def test_external_prog(self):
        env = self.get_env()
        # now call the external program, passing `env` to it
```

Depending on whether the test file was under the `tests` test suite or `examples` it'll correctly set up `env[PYTHONPATH]` to include one of these two directories, and also the `src` directory to ensure the testing is done against the current repo, and finally with whatever `env[PYTHONPATH]` was already set to before the test was called if anything.

This helper method creates a copy of the `os.environ` object, so the original remains intact.

Getting reproducible results

In some situations you may want to remove randomness for your tests. To get identical reproducible results set, you will need to fix the seed:

```
seed = 42

# python RNG
import random

random.seed(seed)

# pytorch RNGs
import torch

torch.manual_seed(seed)
torch.backends.cudnn.deterministic = True
if torch.cuda.is_available():
    torch.cuda.manual_seed_all(seed)

# numpy RNG
import numpy as np

np.random.seed(seed)

# tf RNG
tf.random.set_seed(seed)
```

Debugging tests

To start a debugger at the point of the warning, do this:

```
pytest tests/utils/test_logging.py -W error::UserWarning --pdb
```

A massive hack to create multiple pytest reports

Here is a massive `pytest` patching that I have done many years ago to aid with understanding CI reports better.

To activate it add to `tests/conftest.py` (or create it if you haven't already):

```
import pytest

def pytest_addoption(parser):
    from testing_utils import pytest_addoption_shared

    pytest_addoption_shared(parser)

def pytest_terminal_summary(terminalreporter):
    from testing_utils import pytest_terminal_summary_main

    make_reports = terminalreporter.config.getoption("--make-reports")
    if make_reports:
        pytest_terminal_summary_main(terminalreporter, id=make_reports)
```

and then when you run the test suite, add `--make-reports=mytests` like so:

```
pytest --make-reports=mytests tests
```

and it'll create 8 separate reports:

```
$ ls -1 reports/mytests/
durations.txt
errors.txt
failures_line.txt
failures_long.txt
failures_short.txt
stats.txt
summary_short.txt
warnings.txt
```

so now instead of having only a single output from `pytest` with everything together, you can now have each type of report saved into each own file.

This feature is most useful on CI, which makes it much easier to both introspect problems and also view and download individual reports.

Using a different value to `--make-reports=` for different groups of tests can have each group saved separately rather than clobbering each other.

All this functionality was already inside `pytest` but there was no way to extract it easily so I added the monkey-patching overrides [testing_utils.py](#). Well, I did ask if I can contribute this as a feature to `pytest` but my proposal wasn't welcome.

Resources

Useful compilations

- [@StellaAthena](#) created the [Common LLM Settings spreadsheet](#) which can be a super-useful resource when you're about to embark on a new LLM training - as it tells you how many known LLM trainings were created.
- A few years back I started compiling information on [which dtype the models were trained in](#) - it only contains a handful of models but if you're doing a research on dtypes it can still be useful. I was using this information to try and write [a model pretraining dtype auto-detection](#) and here is a related [float16 vs bfloat16 numerical properties comparison](#).

Publicly available training LLM/VLM logbooks

Logbooks and chronicles of training LLM/VLM are one of the best sources to learn from about dealing with training instabilities and choosing good hyper parameters.

If you know of a public LLM/VLM training logbook that is not on this list please kindly let me know or add it via a PR. Thank you!

The listing is in no particular order other than being grouped by the year.

2021

- BigScience pre-BLOOM 108B training experiments (2021): [chronicles](#) | [the full spec and discussions](#) (backup: [1](#) | [2](#))

2022

- BigScience BLOOM-176B (2022): [chronicles-prequel](#) | [chronicles](#) | [the full spec and discussions](#) (backup: [1](#) | [2](#) | [3](#))
- Meta OPT-175B (2022): [logbook](#) | [Video](#) (backup: [1](#))
- THUDM GLM-130B (2022): [en logbook](#) | [Mandarin version](#) (backup: [1](#) | [2](#))

2023

- HuggingFace IDEFICS-80B multimodal (Flamingo repro) (2023): [Learning log](#) | [Training Chronicles](#) (backup: [1](#) | [2](#))
- BloombergGPT 50B LLM - section C in [BloombergGPT: A Large Language Model for Finance](#)

2024

- [MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs](#) - the paper covers various training issues and their resolution - albeit on models that are proprietary yet just as instructional/useful.
- Imbue's [From bare metal to a 70B model: infrastructure set-up and scripts](#) very detailed technical post covers many training-related issues that they had to overcome while training a proprietary 70B-param model.

Hardware setup logbooks

- Imbue published a detailed log of how they have set up a 512-node IB-fat-tree cluster and made it to work: [From bare metal to a 70B model: infrastructure set-up and scripts](#), they also open-sourced the [cluster tooling](#) they created in the process.

- SemiAnalysis published a great detailed writeup about [what it takes to set up a Neocloud cluster](#).

Contributors

Multiple contributors kindly helped to improve these ever improving and expanding notes.

1. Some of them did it via PRs, and are thus listed automatically [here](#)
2. Others did it via various other ways so I'm listing them explicitly here:

- [Adam Moody](#)
- [Alex Rogozhnikov](#)
- [Bowei Liu](#)
- [Derrick Horton](#)
- [Elio VP](#)
- [Garrett Goon](#)
- [Horace He](#)
- [Ivan Yashchuk](#)
- [Jack Dent](#)
- [Jordan Nanos](#)
- [Mark Saroufim](#)
- [Olatunji Ruwase](#)
- [Oren Leung](#)
- [Quentin Anthony](#)
- [Ross Wightman](#)
- [Samyam Rajbhandari](#)
- [Shikib Mehri](#)
- [Siddharth Singh](#)
- [Stéphane Requena](#)
- [Zhiqi Tao](#)

If you contributed to this text and for some reason you're not on one of these 2 lists - let's fix it by adding your name with a github or similar link [here](#).

Book Building

Important: this is still a WIP - it mostly works, but stylesheets need some work to make the pdf really nice. Should be complete in a few weeks.

This document assumes you're working from the root of the repo.

Installation requirements

1. Install python packages used during book build

```
pip install -r build/requirements.txt
```

2. Download the free version of [Prince XML](#). It's used to build the pdf version of this book.

Build html

```
make html
```

Build pdf

```
make pdf
```

It will first build the html target and then will use it to build the pdf version.

Check links and anchors

To validate that all local links and anchored links are valid run:

```
make check-links-local
```

To additionally also check external links

```
make check-links-all
```

use the latter sparingly to avoid being banned for hammering servers.

Move md files/dirs and adjust relative links

e.g. `slurm => orchestration/slurm`

```
src=slurm
dst=orchestration/slurm

mkdir -p orchestration
git mv $src $dst
perl -pi -e "s|$src|$dst|" chapters-md.txt
python build/mdbook/mv-links.py $src $dst
git checkout $dst
make check-links-local
```

Resize images

When included images are too large, make them smaller a bit:

```
mogrify -format png -resize 1024x1024\> *png
```