together.ai






Building an Eco-system for Open Foundation Models, Together

Ce Zhang

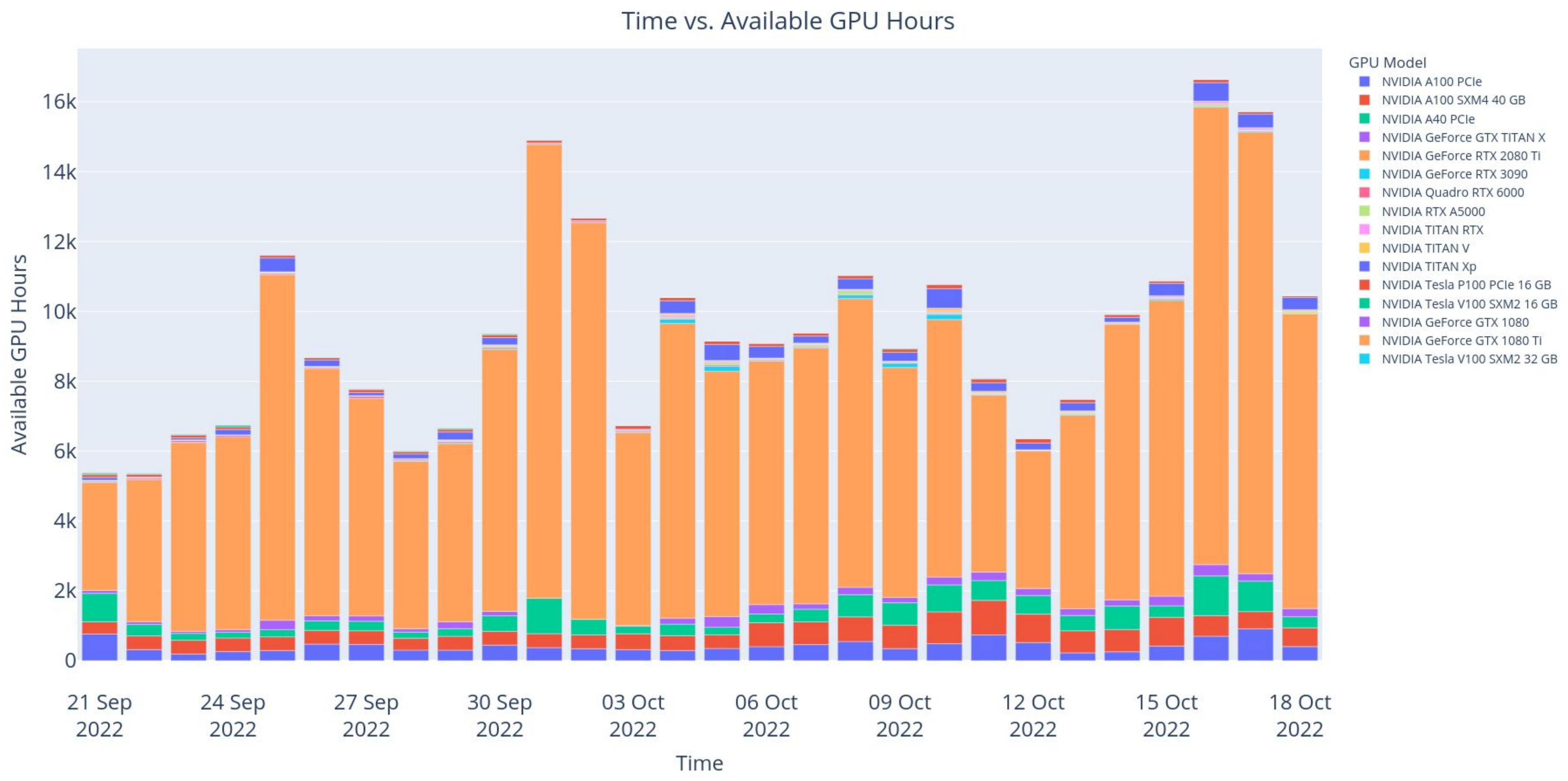
Once upon a time, there was no llama.

The first version of HELM

BigScience	BLOOM	176B	July 2022
	T0pp	11B	October 2021
EleutherAI	GPT-J	6B	July 2021
	GPT-NeoX	20B	February 2022
清華大學 Tsinghua University	GLM	130B	August 2022
Google Research	UL2	20B	October 2022
	T5	11B	February 2020
Meta AI	OPT	175B	June 2022
	OPT	66B	June 2022
Yandex	YaLM	100B	June 2022

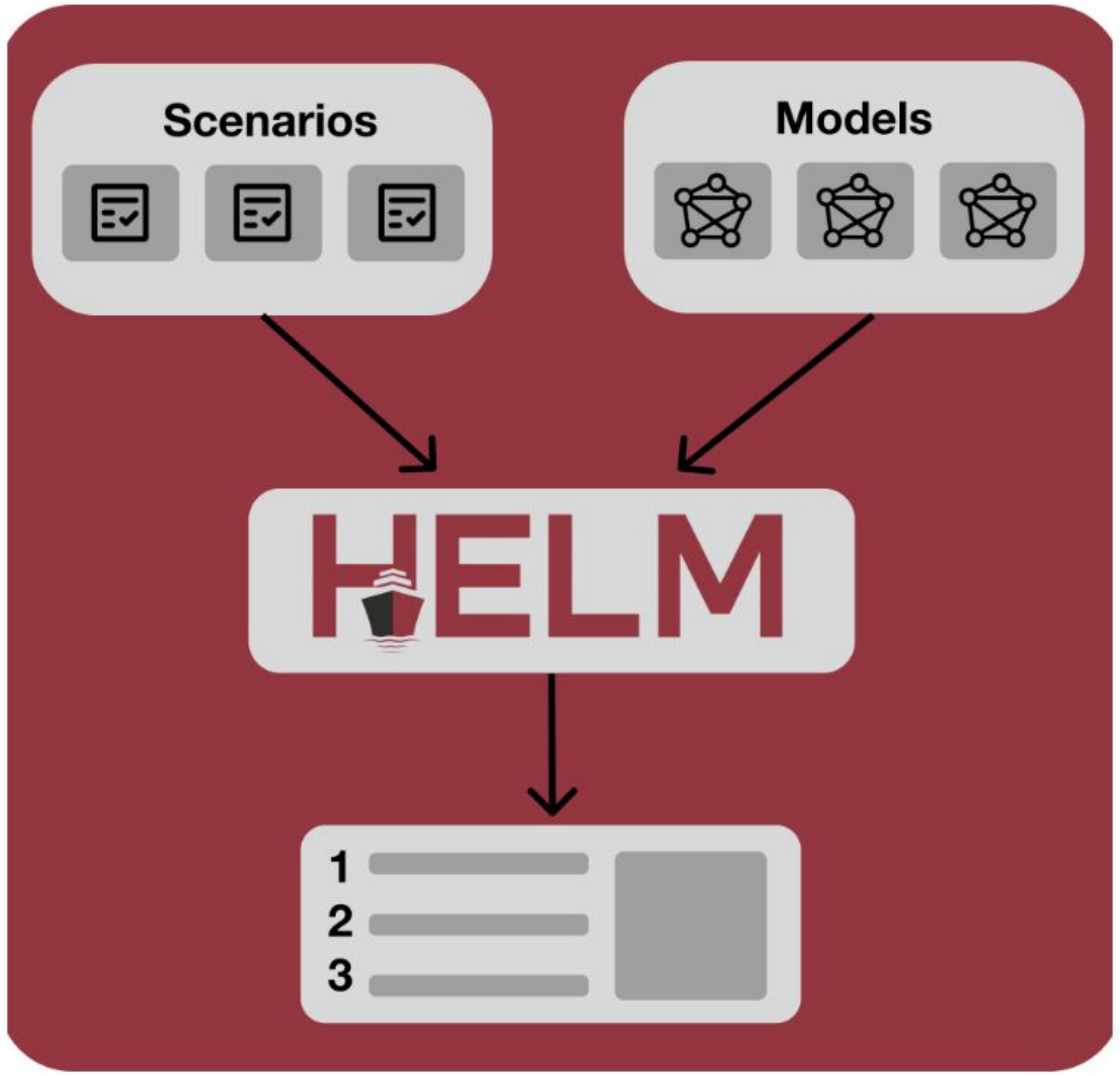
The first version of HELM

BigScience	BLOOM	176B	July 2022
	T0pp	11B	October 2021
EleutherAI	GPT-J	6B	July 2021
	GPT-NeoX	20B	February 2022
清華大學 Tsinghua University	GLM	130B	August 2022
Google Research	UL2	20B	October 2022
	T5	11B	February 2020
Meta AI	OPT	175B	June 2022
	OPT	66B	June 2022
Yandex	YaLM	100B	June 2022



The Community is Getting There

Center for Research on Foundation Models HELM Lite Leaderboard Models Scenarios Predictions GitHub Release: v1.0.0



Model	Mean win rate
GPT-4 (0613)	0.962
GPT-4 Turbo (1106 preview)	0.834
Palmyra X V3 (72B)	0.821
Palmyra X V2 (33B)	0.783
PaLM-2 (Unicorn)	0.776
Yi (34B)	0.772

SEE MORE

There are also *challenges* that we are facing

Data

Great Datasets Exist (Pile, C4, The Stack, etc.), but…

<u>Data Recipe is Missing</u> — *How can we map "raw" data to a high-quality core to maximize model quality? How can we clean them? How can we mix them? How can we reproduce a recipe given a new domain?*

<u>Diversity of Data is Lacking</u> — *How can we go beyond Internet / Code / Synthetic data? Is there a way to open up the channel for the open source community to access a more diverse pool of data?*

Compute

Community manages to get computation resources that made several amazing projects possible, but…

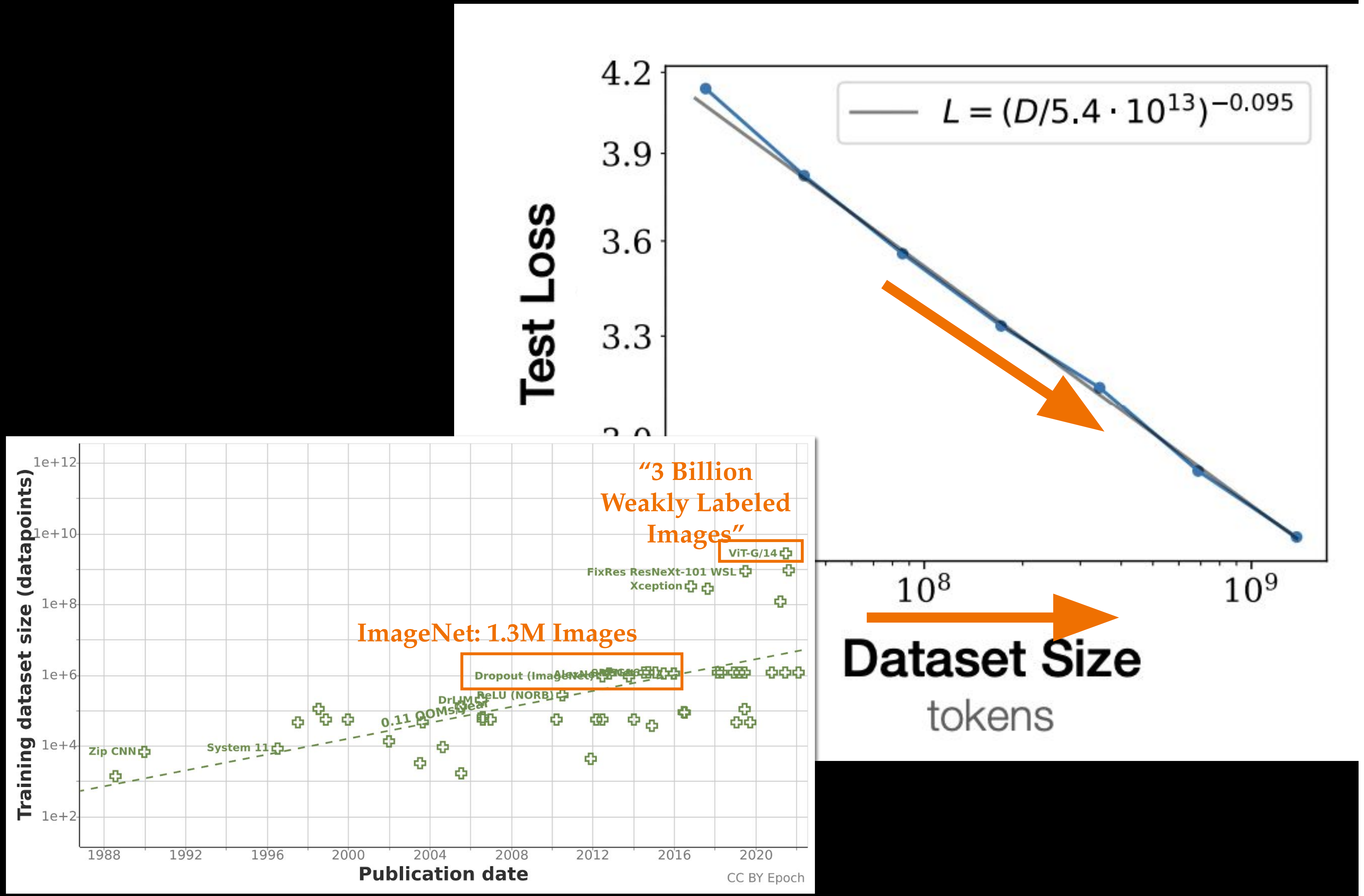
<u>Availability of Compute is Still Hindering Progress</u> — *Low availability and high price. And it is amazing how much impact that even a small, sustainable amount (e.g., 8x A100) compute can make on open source and research projects.*

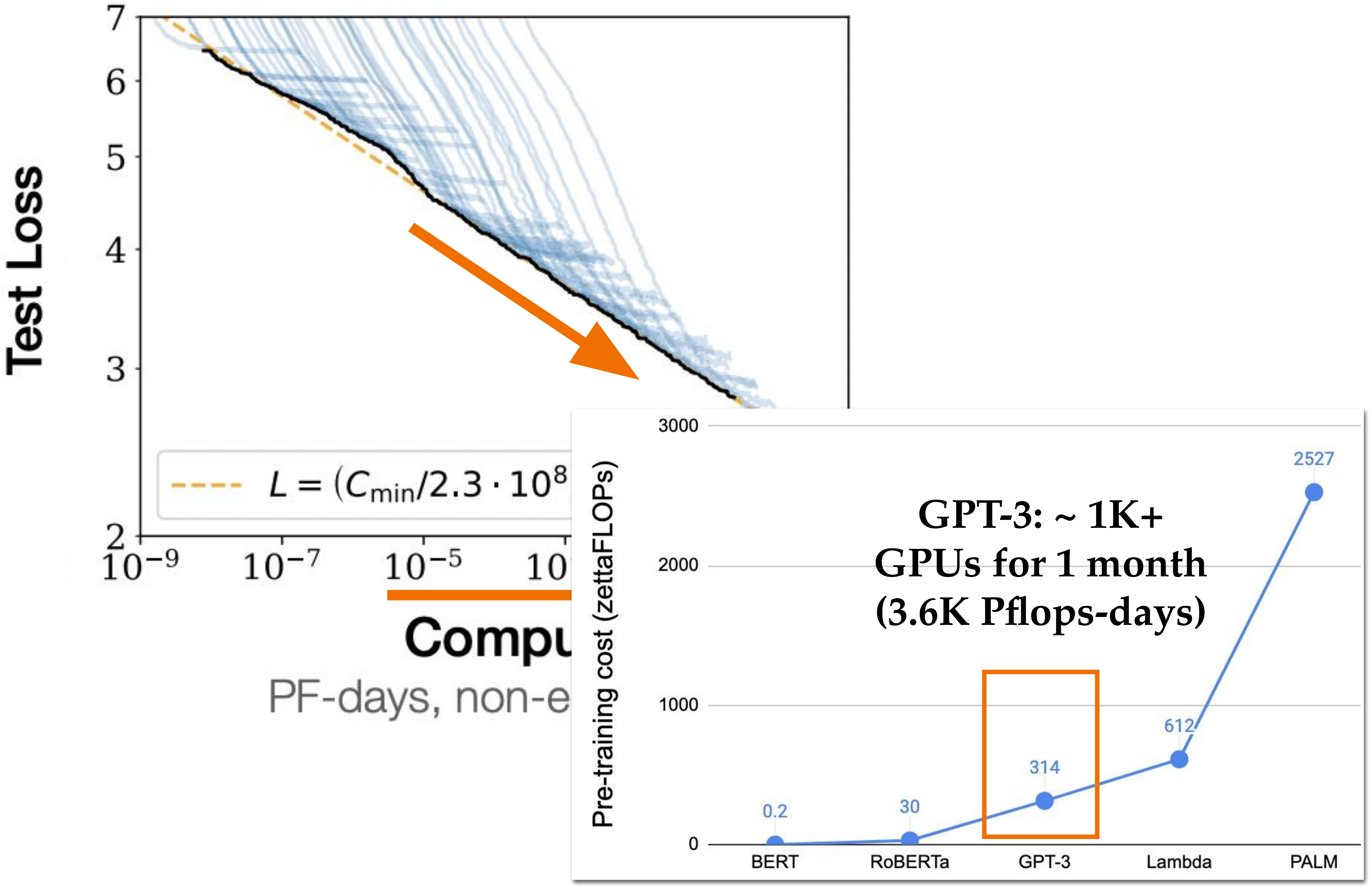
<u>Strong requirements on connectivity and locality of computes</u> — 200Gbps-3.2Tbps connections and growing.

<u>Reuse is lacking</u> — *As a community, we have so many 7B base models for 1-2T tokens in 2023; They don't build on each other, and they don't from a systematic path of exploration to maximize our understanding.*

These questions need the whole community to work out together; and we hope to help stimulate it in our own way.

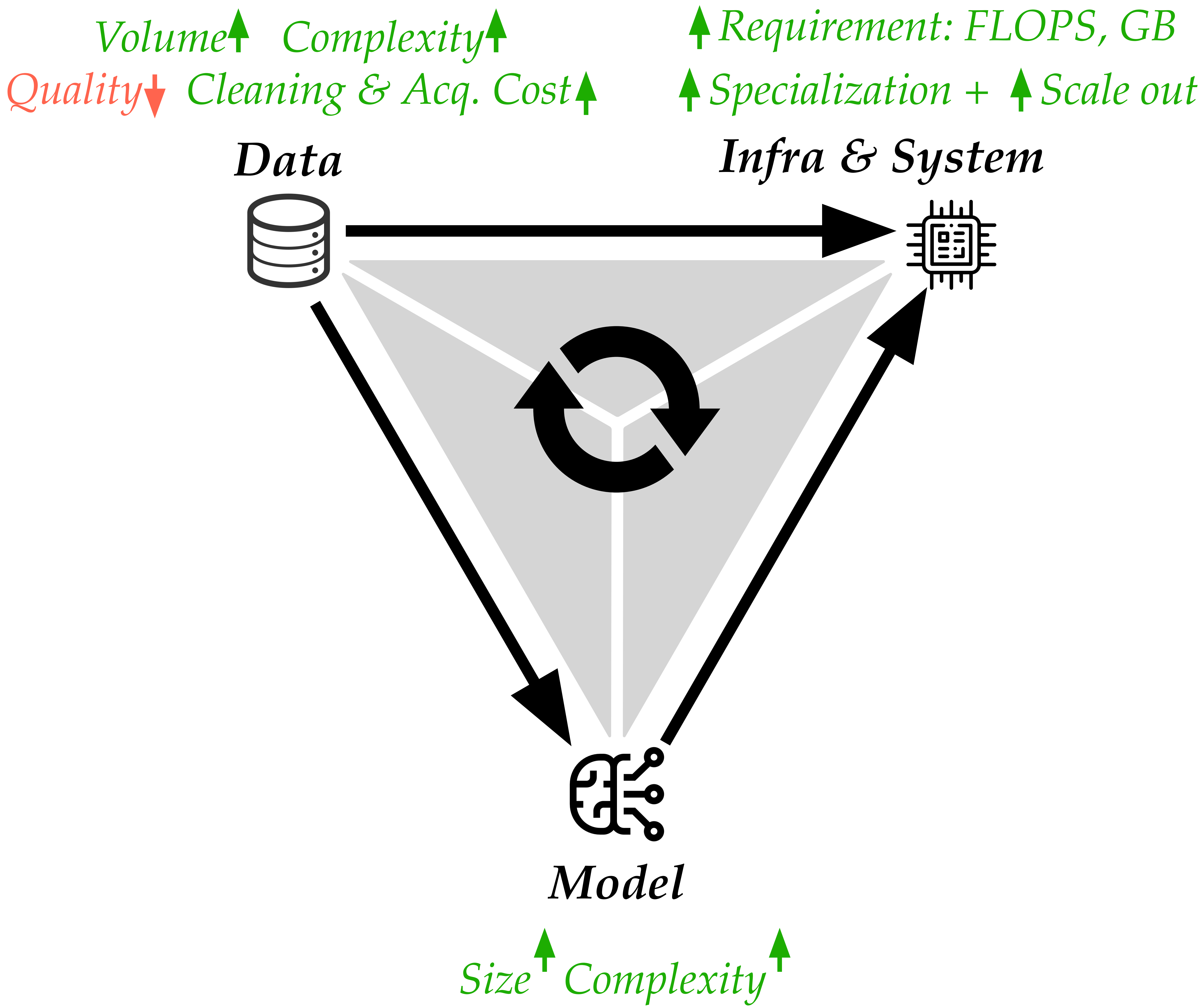
Data ⟵ Driving Forces ⟶ Compute





Source: (1) Scaling Laws for Neural Language Models; (2) https://www.lesswrong.com/posts/asqDCb9XzXnLjSfgL/trends-in-training-dataset-sizes

Tensions



Volume↑ Complexity↑
Quality↓ Cleaning & Acq. Cost↑
Data
↑Requirement: FLOPS, GB
↑Specialization + ↑Scale out
Infra & System
Model
Size↑ Complexity↑

Our Belief

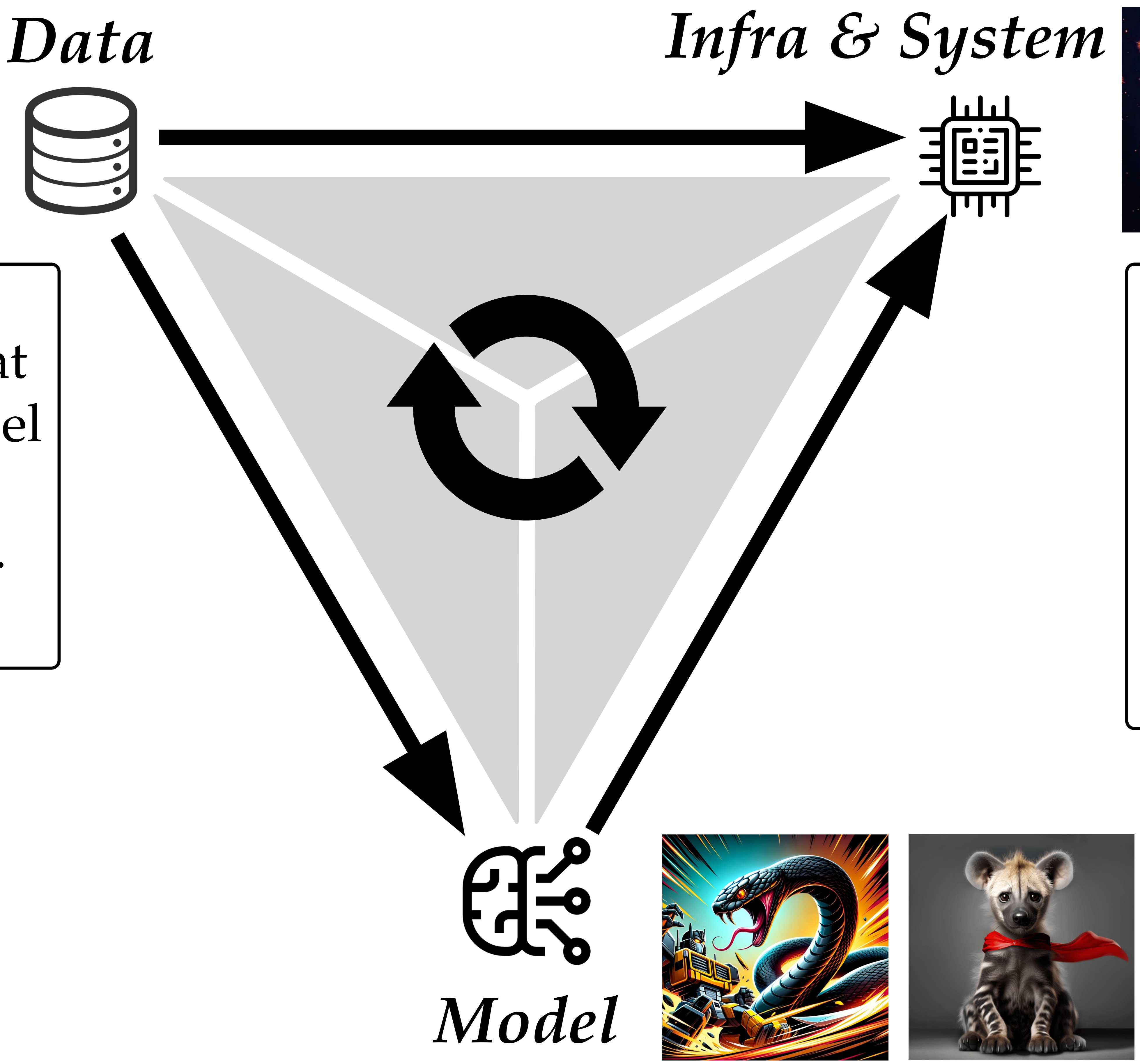
- **Infrastructure** (and its trend) decides the fundamental cost of compute and data movement
- **Model architecture (and algorithm)** defines the fundamental limit of utilization of the infrastructure, and (potential) tradeoff on quality
- **Data** enables capacity and decides requirements on compute and data movement

- These three dimensions will to come together and we will find the right balance as a community.

Our Efforts

Data and Data Recipe that maps data quality to model quality in a *principled*, *reproducible*, *scalable* way.

RedPajama-v1; RedPajama-v2



FlashAttention; FlashDecoding; FlexGen; CocktailSGD

Improving the *economics* — A joint optimization between System (kernel, communication, architecture), Algorithm, Modeling, & Hardware.

Mamba; StripedHyena

Bring *Beyond*-Transformer Architecture from Research to Production.

RedPajama v2: 30T Tokens and a Modular View of Data Quality

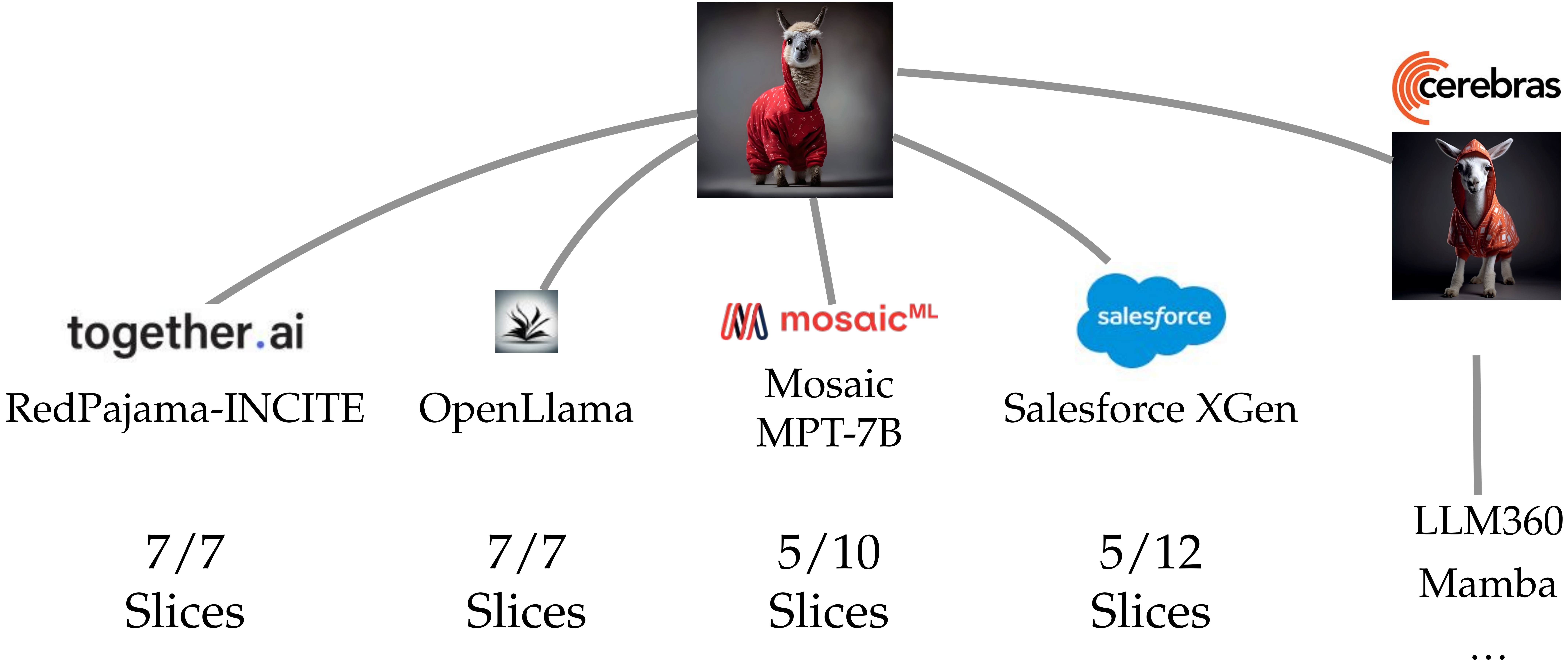
What Llama-1 brought for the community was hope.

RedPajama-v1: A Best Effort Llama Reproduction

- **CommonCrawl**: Five dumps of CommonCrawl, processed using the CCNet pipeline, and filtered via several quality filters including a linear classifier that selects for Wikipedia-like pages.
- **C4**: Standard C4 dataset
- **GitHub**: GitHub data, filtered by licenses and quality
- **arXiv**: Scientific articles removing boilerplate
- **Books**: A corpus of open books, deduplicated by content similarity
- **Wikipedia**: A subset of Wikipedia pages, removing boilerplate
- **StackExchange**: A subset of popular websites under StackExchange, removing boilerplate

	RedPajama	LLaMA*
CommonCrawl	878 billion	852 billion
C4	175 billion	190 billion
Github	59 billion	100 billion
Books	26 billion	25 billion
ArXiv	28 billion	33 billion
Wikipedia	24 billion	25 billion
StackExchange	20 billion	27 billion
Total	1.2 trillion	1.25 trillion

Fueling an Exciting Generation of Open Models and Data



Data, and Measures of Quality

- **<u>Lessons from RedPajama-v1</u>**
  - Being able to flexibly twist data recipe is important — degree of deduplication, different filtering criteria, different combinations of filters, etc.
  - Systematically mapping data quality to model quality is an open problem; it might be more useful to build up the framework to help the community to explore data quality than building yet another model with a fixed, “magical” dataset.
- **<u>RedPajama-v2</u>**
  - All CommonCrawl dumps passing through ccNet => 100T Raw Tokens, 5 Languages
  - Very basic, minimum filtering (exact dedup) => 30T Tokens
    - Keeping both filtered and unfiltered (so you can study the impact of deduplication)
  - Overlay tokens with 40+ quality signals:
    - ML Heuristics: e.g., DSIR (Xie et al.): *Given a bag of {1,2}-wordgram model trained on Wikipedia articles p, and a model trained on the source domain q, this is the logarithm of the ratio p(doc)/q(doc).*
    - All Quality rules from C4, Pretrainer’s Guide, RefinedWeb, Gopher: E.g., *The number of occurrences of the word "javascript" in each line; The ratio between the number of uppercase letters and total number of characters in each line.*
    - Fuzzy Deduplication Signals: E.g., *Banded minhash signature of the document, for fuzzy deduplication at*

1 CommonCrawl Dump, 40 Ways of Measuring Quality

RedPajama v2: A Modular View of Data Quality

```python
def gopher_rules_pass(sample) -> bool:
    """ function returns True if the sample complies with Gopher rules
"""
    signals = json.loads(sample["quality_signals"])

    # rule 1: number of words between 50 and 10'000
    word_count = signals["rps_doc_word_count"][0][2]
    if word_count < 50 or word_count > 10_000:
        return False

    # rule 2: mean word length between 3 and 10
    mean_word_length = signals["rps_doc_mean_word_
    if mean_word_length < 3 or mean_word_length >
        return False

    # rule 2: symbol to word ratio below 0.1
    symbol_word_ratio = signals["rps_doc_symbol_
    if  symbol_word_ratio > 0.1:
        return False

    # rule 3: 90% of lines need to start without
    n_lines = signals["ccnet_nlines"][0][2]
    n_lines_bulletpoint_start = sum(map(lambda l
signals["rps_lines_start_with_bulletpoint"]))
    if n_lines_bulletpoint_start / n_lines > 0.9
        return False

    # rule 4: the ratio between characters in the most frequent 2-gram
and the total number
    # of characters must be below 0.2
    top_2_gram_frac = signals["rps_doc_frac_chars_top_2gram"][0][2]
    if top_2_gram_frac > 0.2:
        return False

    # rule 5: ...


    return True

ds = load_dataset("togethercomputer/RedPajama-Data-V2", name="sample")
filtered_dataset = list(filter(gopher_rules_pass, ds["train"]))
```

```python
def rpv1_rules_pass(sample) -> bool:
    """ function returns True if the sample complies with the filtering
rules used in RP-V1 """
    signals = json.loads(sample["quality_signals"])

    # rule 1: the wikipedia reference classifier score must be higher
than 0.25
    wikiref_score = signals["rps_doc_ml_wikiref_score"][0][2]
    if wikiref_score < 0.25:
        return False
```

And potentially, learn to combine all the filters, potentially dynamically at different stages of training.

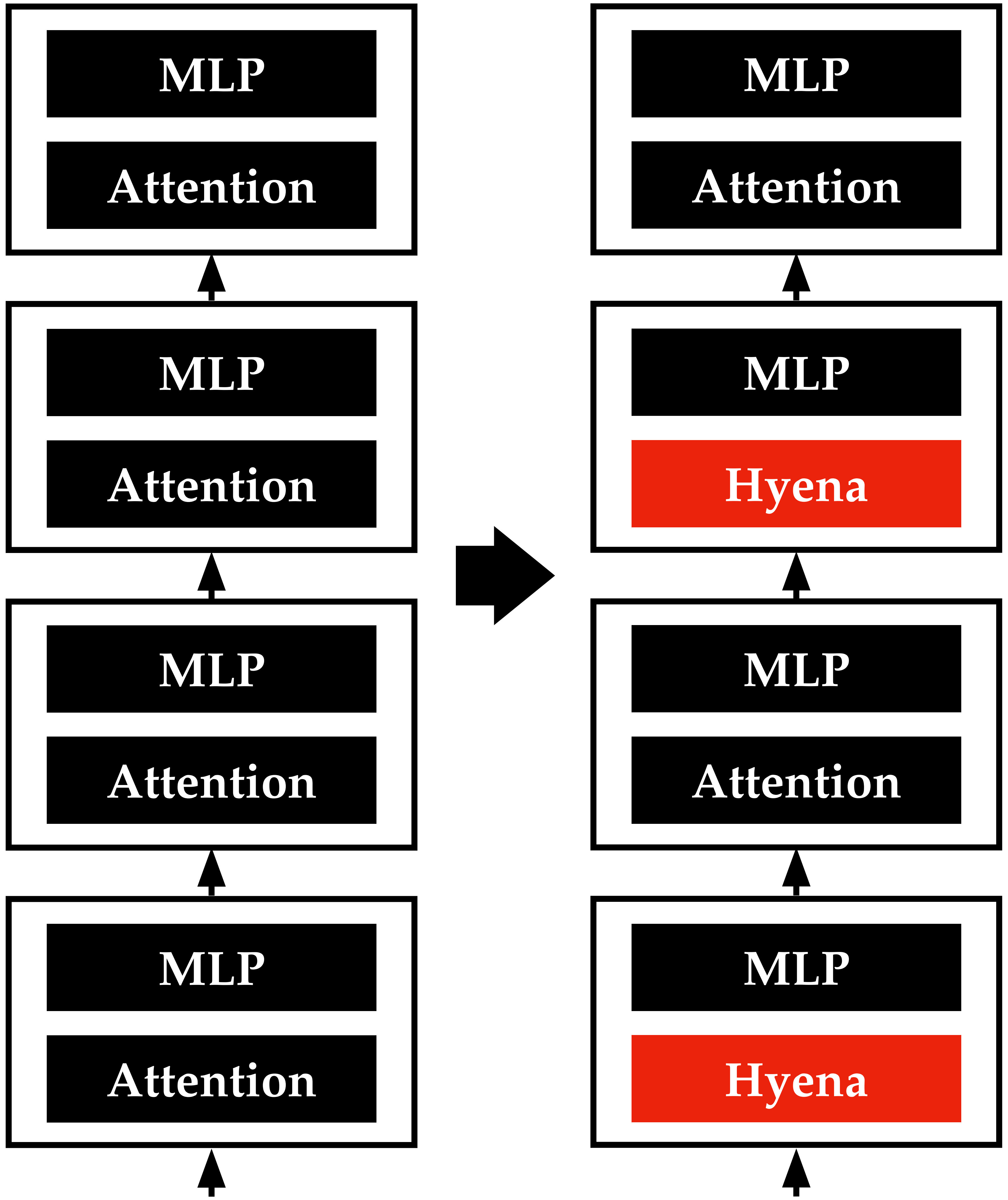
```python
ple) -> bool:
rns True if the sample complies with the filtering

ads(sample["quality_signals"])

 3 sentences
ignals["rps_doc_num_sentences"][0][2]
< 3:

    # rule 2: page may not contain bad words
    n_bad_words = signals["rps_doc_ldnoobw_words"][0][2]
    if n_bad_words > 0:
        return False

    # rule 3: page may not contain placeholder "lorem ipsum" text
    lorem_ipsum = signals["rps_doc_lorem_ipsum"][0][2]
    if lorem_ipsum > 0:
        return False

    # rule 4: ...

    return True
```

StripedHyena: Towards State-of-the-art Beyond-Transformer Models

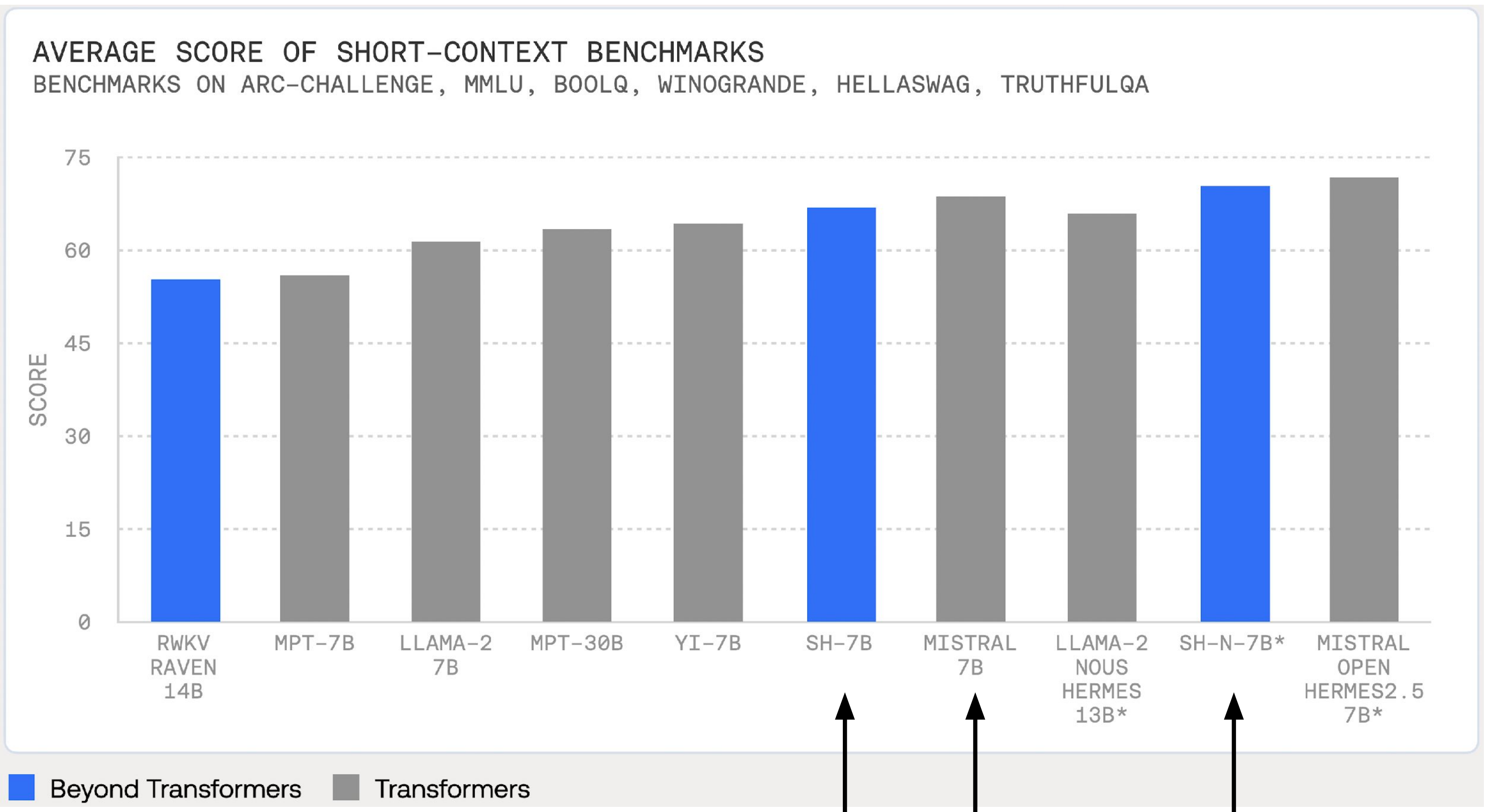
Can alternative architecture matches top modern transformers in quality?

An Transformer / Hyena Hybrid

StripedHyena-7B



Short Context



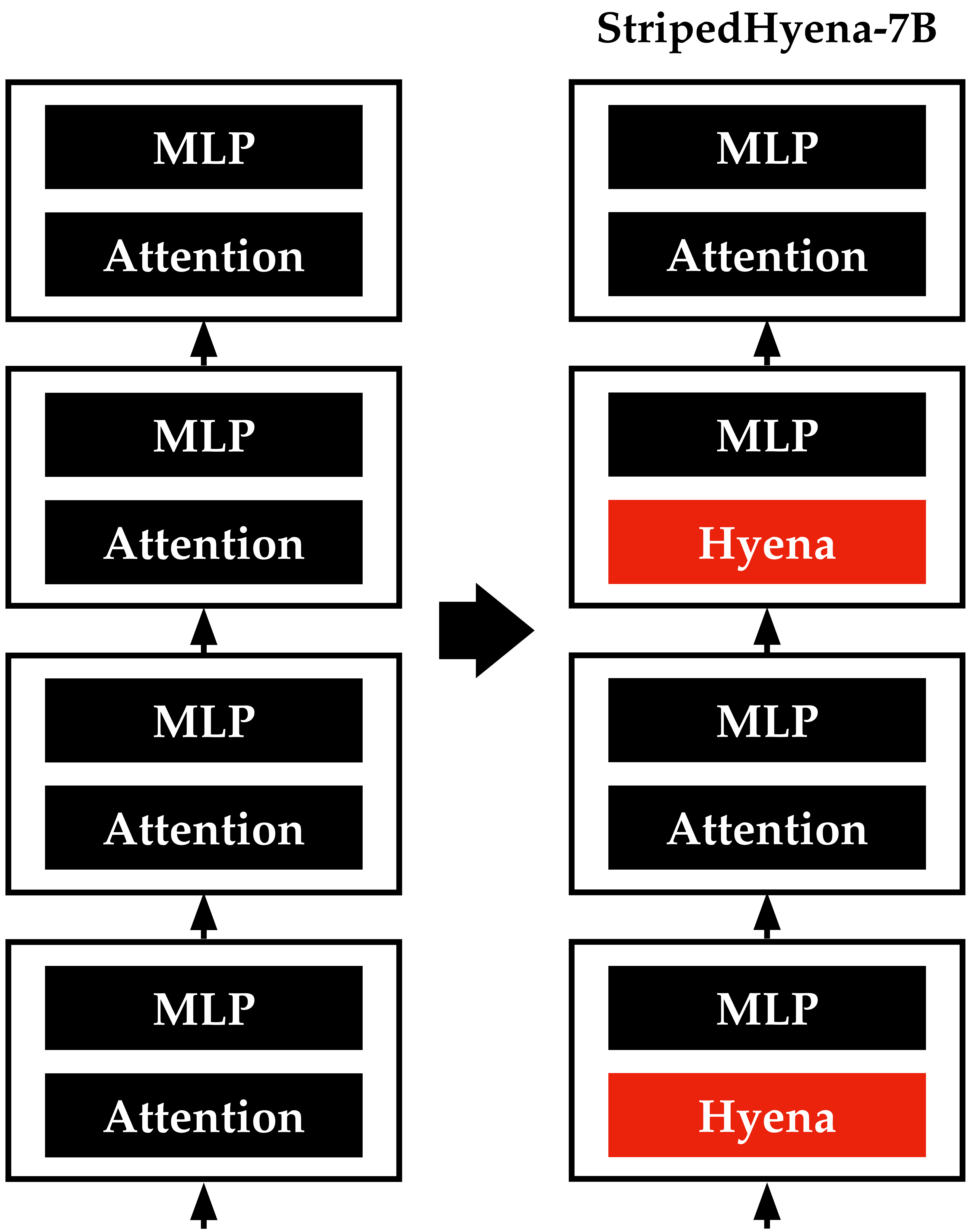
Base: Only 2-3 points behind Mistral

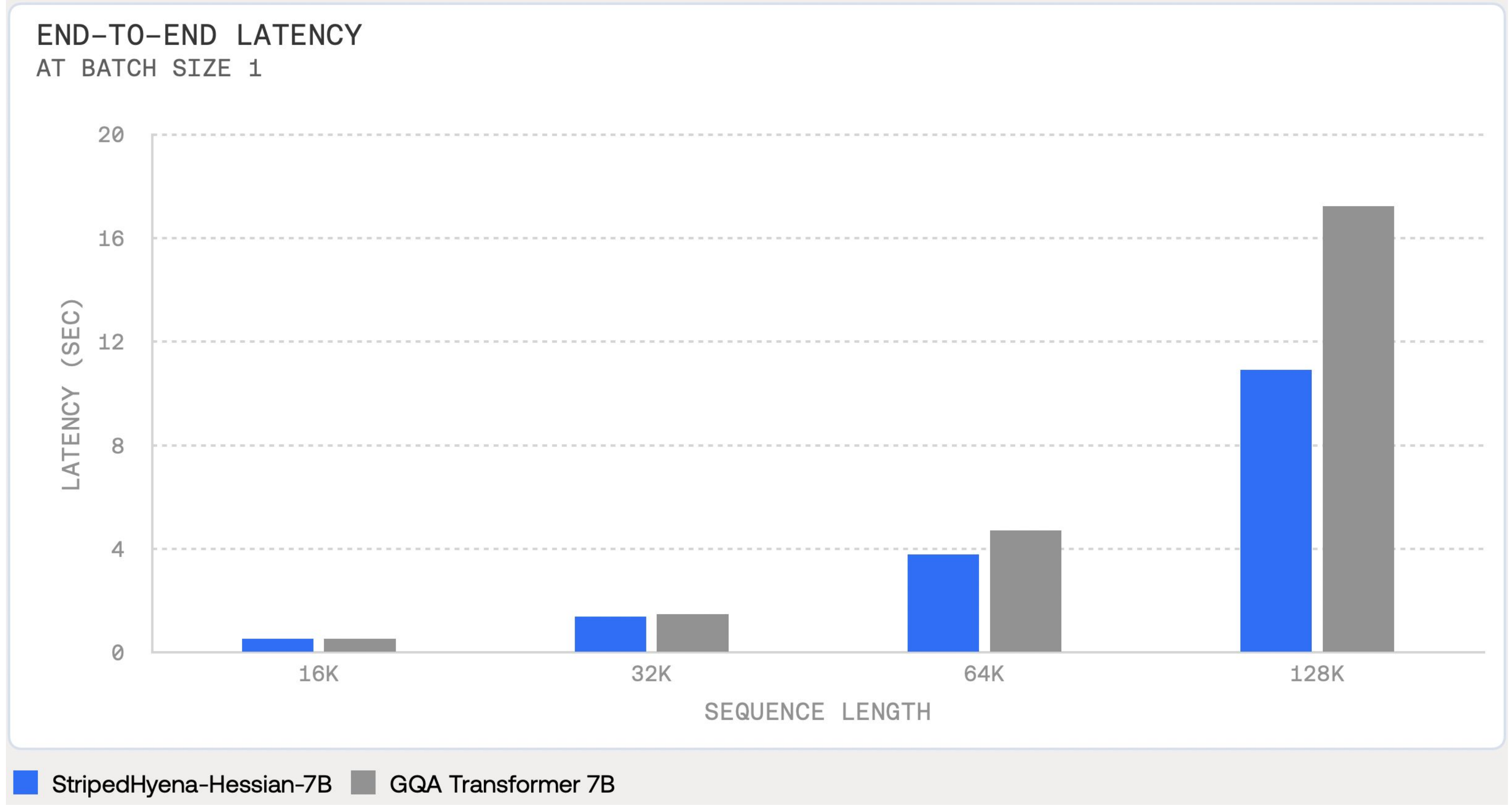
Further Instruction-tuning helps

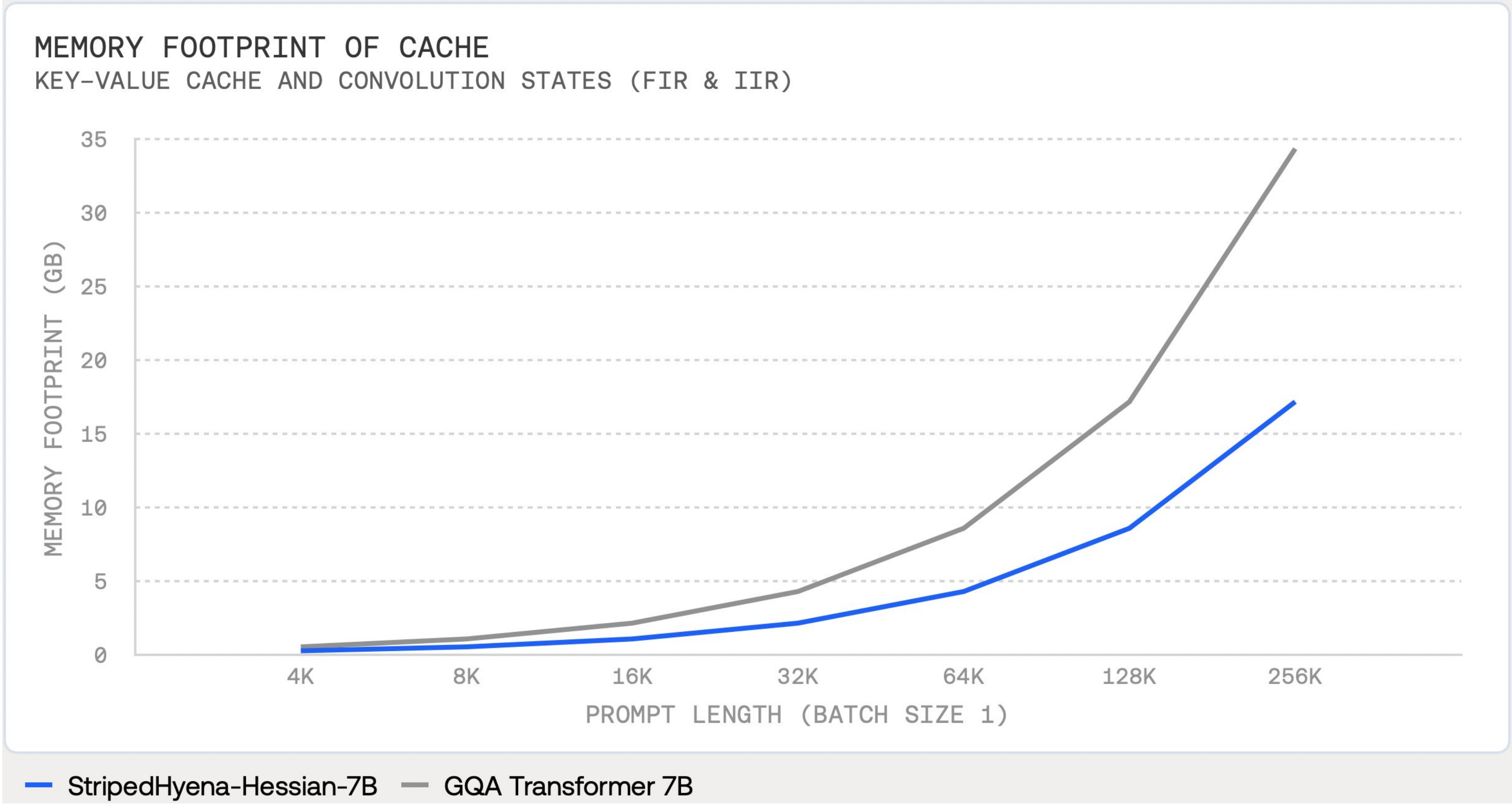
Long Context

Benchmark (shot)	SH 7B	Mistral 7B
GovReport, F1 (0)	27.9	17.5
NarrativeQA, F1 (0)	25.8	24.7
Qasper, F1 (0)	28.8	30.3
Average	27	24

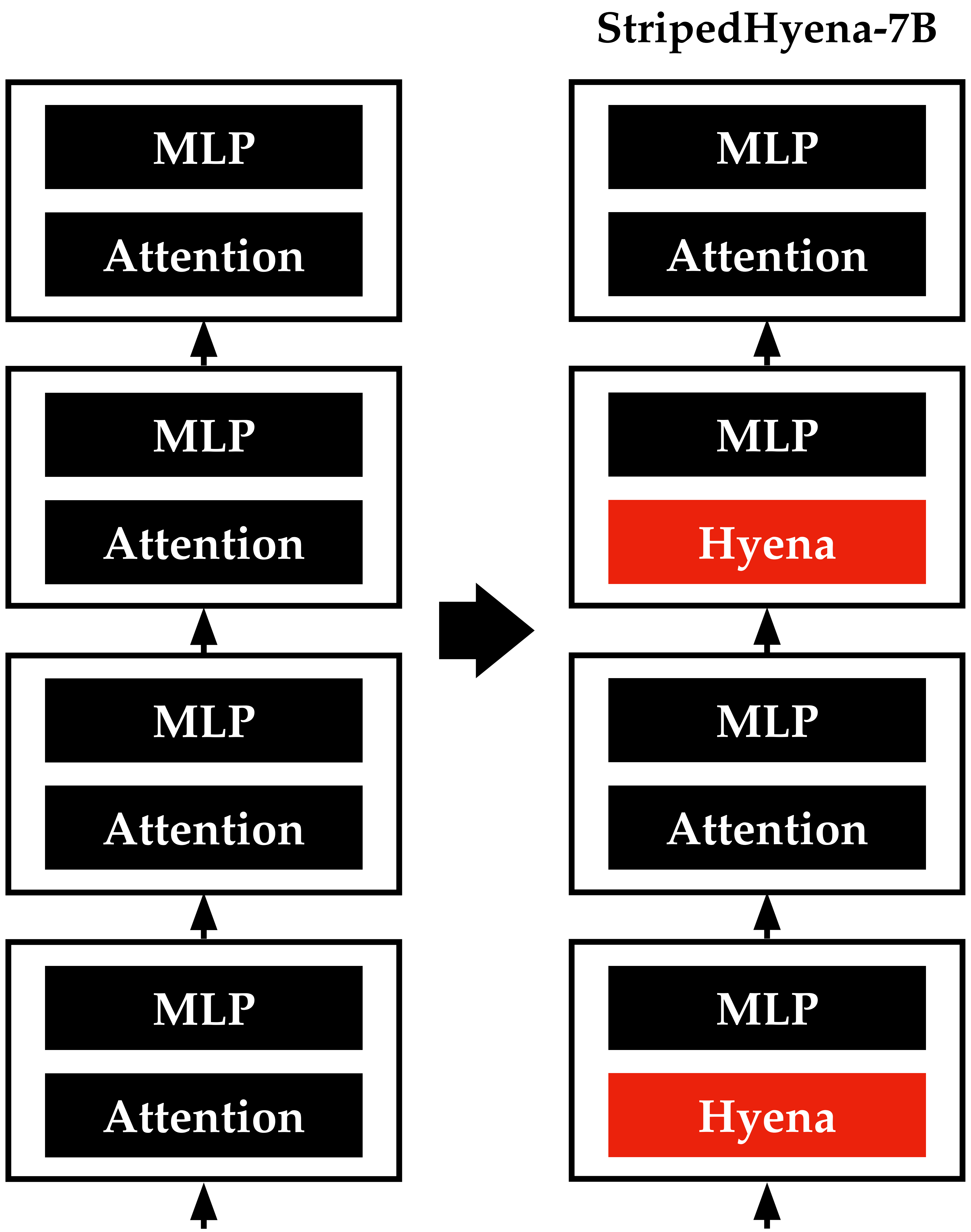
An Transformer / Hyena Hybrid





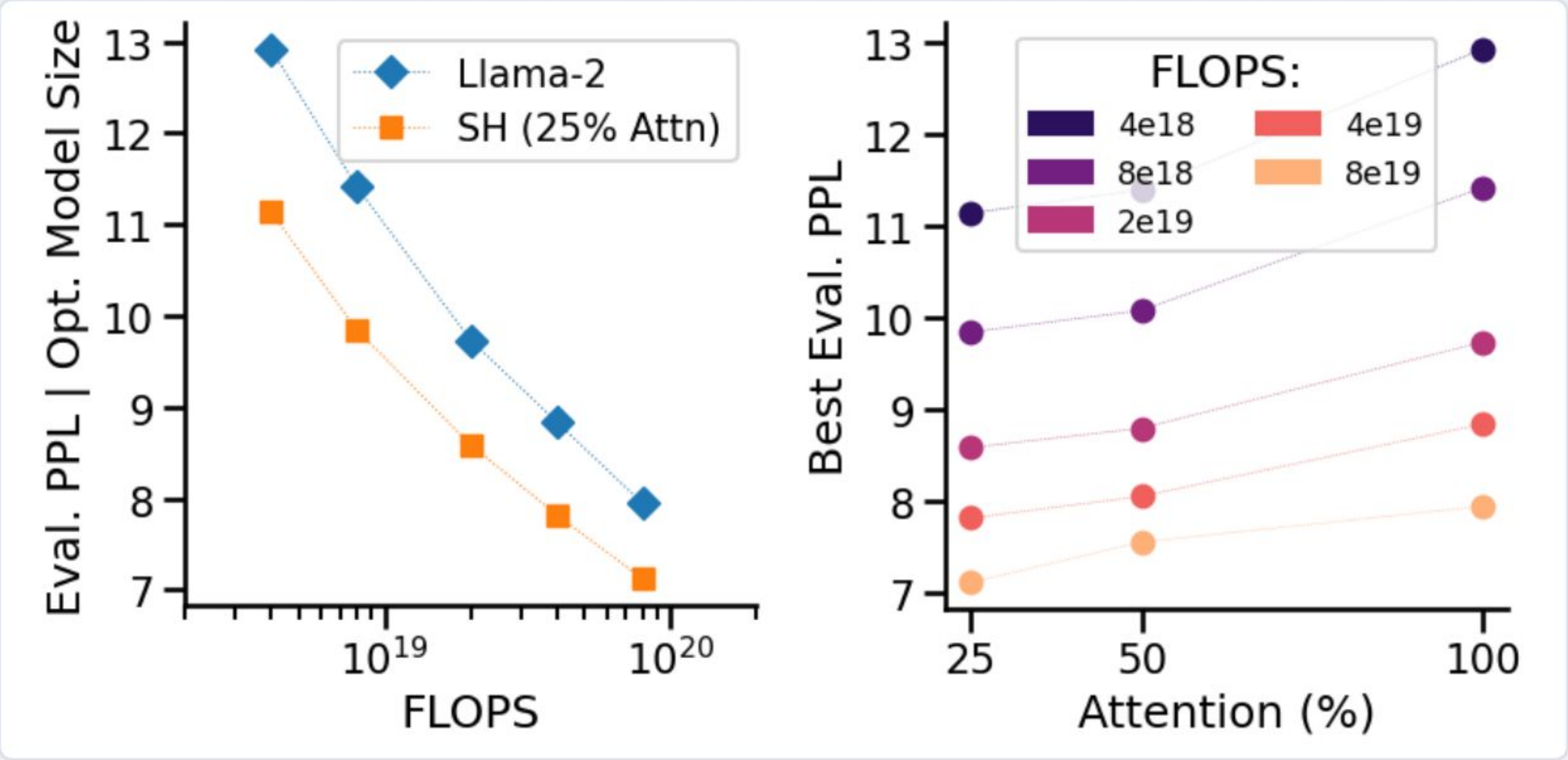


Early Lessons

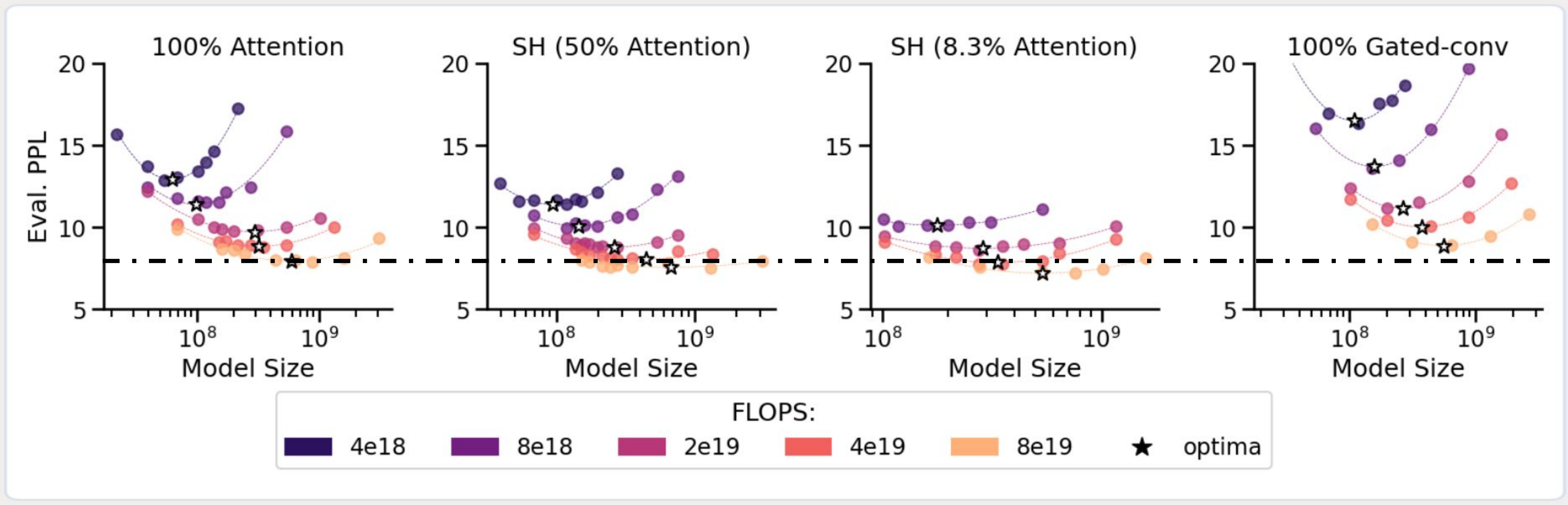


- Alternative Architectures can achieve SOTA against the strongest Transformer baselines.
- Alternative Architectures can provide benefit against the strongest Transformer baselines.
- There are so much more todo, but there is hope.

Towards Architecture-aware Scaling Law



Towards Architecture-aware Scaling Law



Pulsar: Economics of Serving and Training

Cost of Inference

together.ai

CHAT, LANGUAGE, AND CODE MODELS

MODEL SIZE	PRICE 1M TOKENS
Up to 4B	$0.1
4.1B - 8B	$0.2
8.1B - 21B	$0.3
21.1B - 41B	$0.8
41B - 70B	$0.9

The day will come when people count tokens in Billions not Millions.

Cost of Inference

together.ai

CHAT, LANGUAGE, AND CODE MODELS

MODEL SIZE	PRICE 1M TOKENS
Up to 4B	$0.1
4.1B - 8B	$0.2
8.1B - 21B	$0.3
21.1B - 41B	$0.8
41B - 70B	$0.9

Our Dream

The day will come when people count tokens in Billions not Millions.

Cost of Inference

together.ai

CHAT, LANGUAGE, AND CODE MODELS

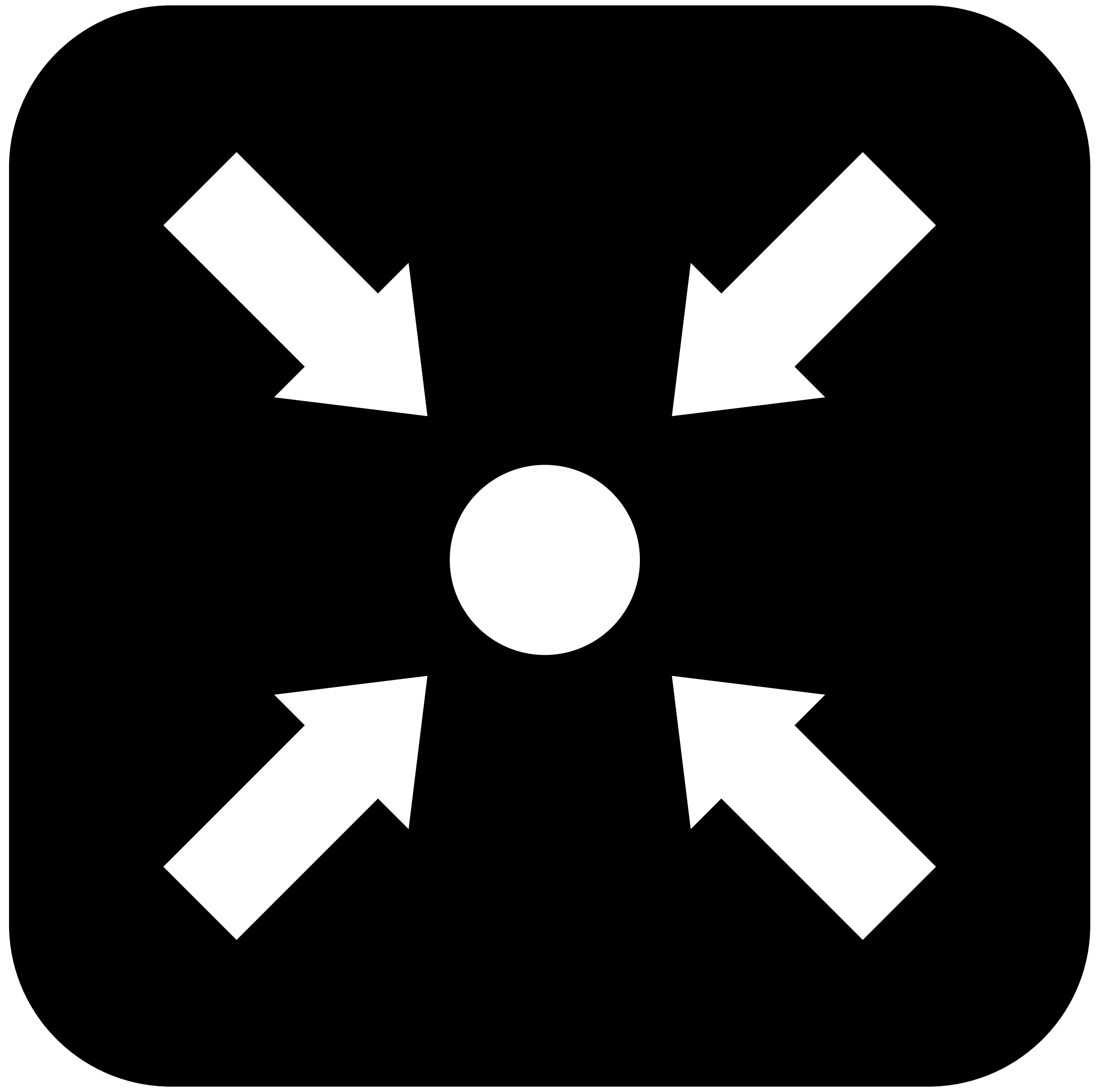
MODEL SIZE	PRICE 1M TOKENS
Up to 4B	$0.1
4.1B - 8B	$0.2
8.1B - 21B	$0.3
21.1B - 41B	$0.8
41B - 70B	$0.9

Our Dream

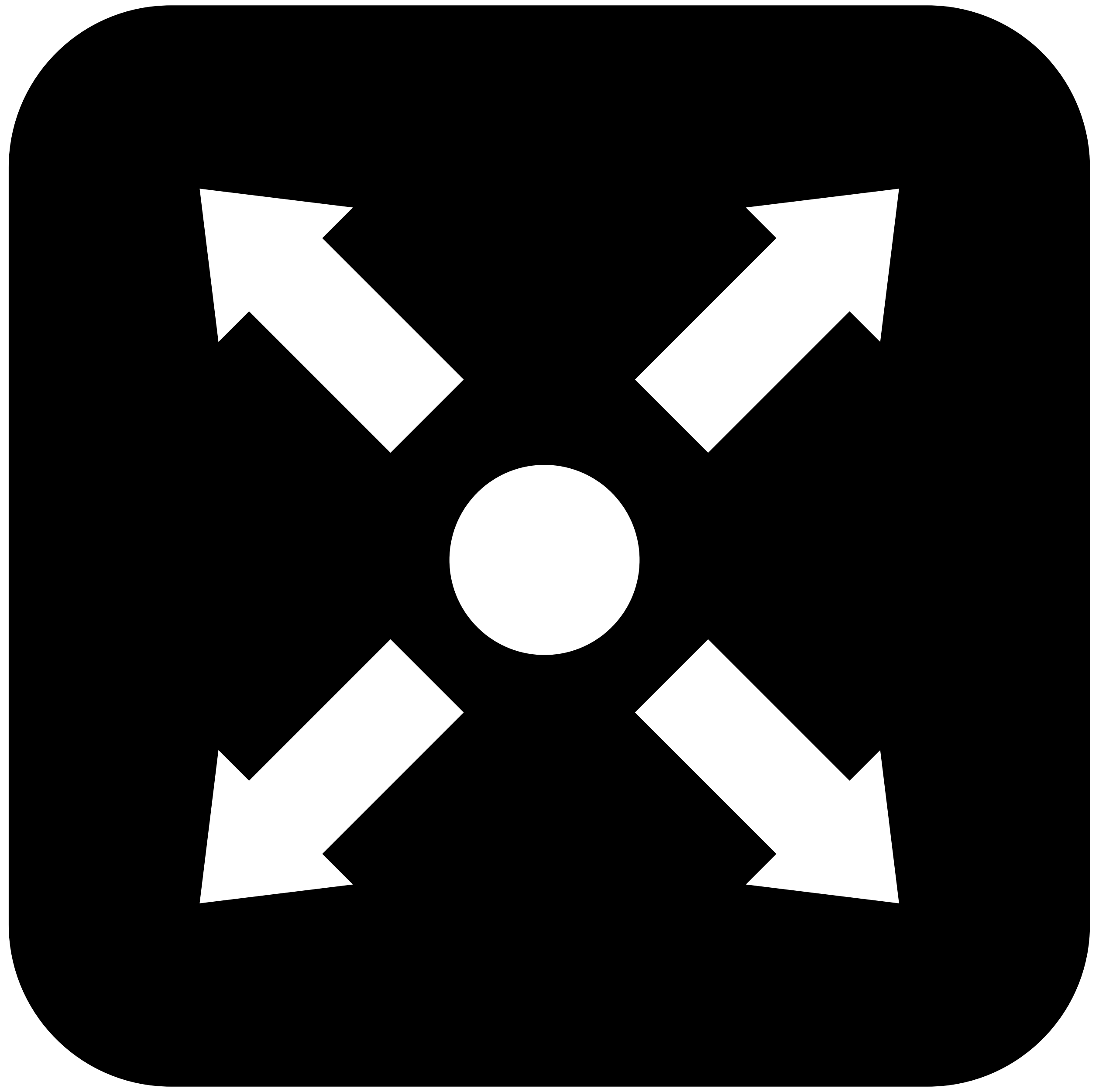
The day will come when people count tokens in Billions not Millions.

(Llama-guard is already there)

Working Hypothesis on Infrastructure



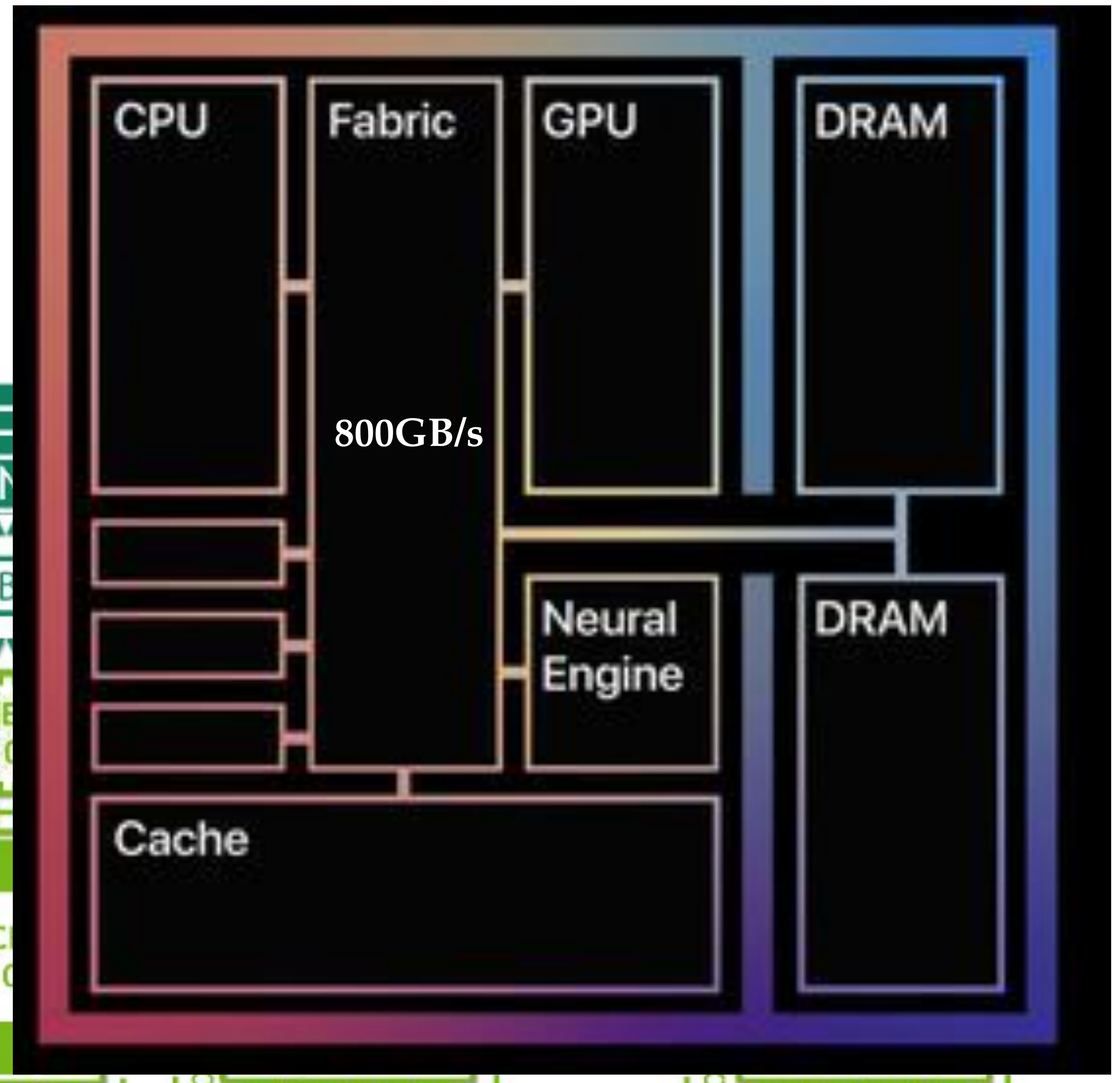
Aggregation



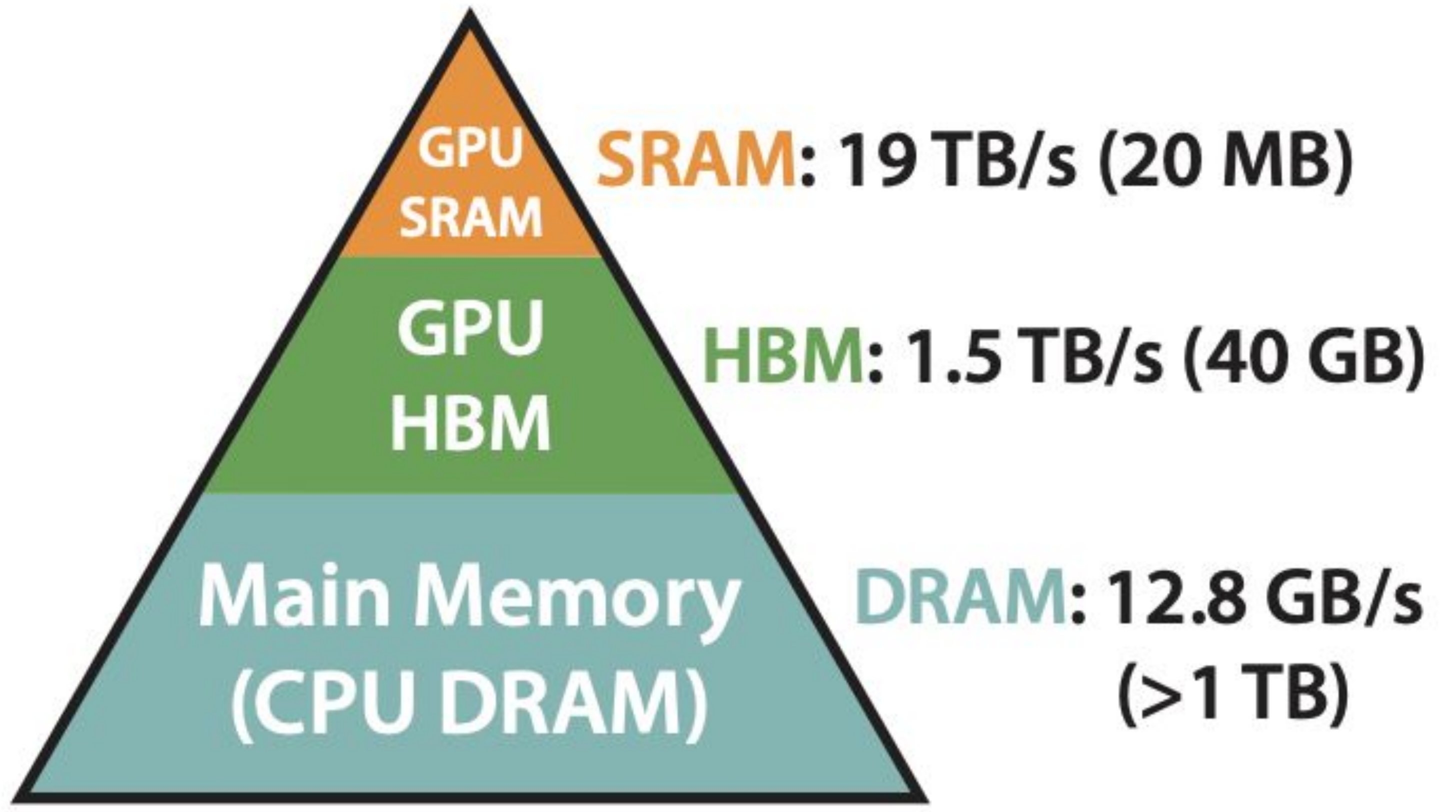
Disaggregation

Aggregation

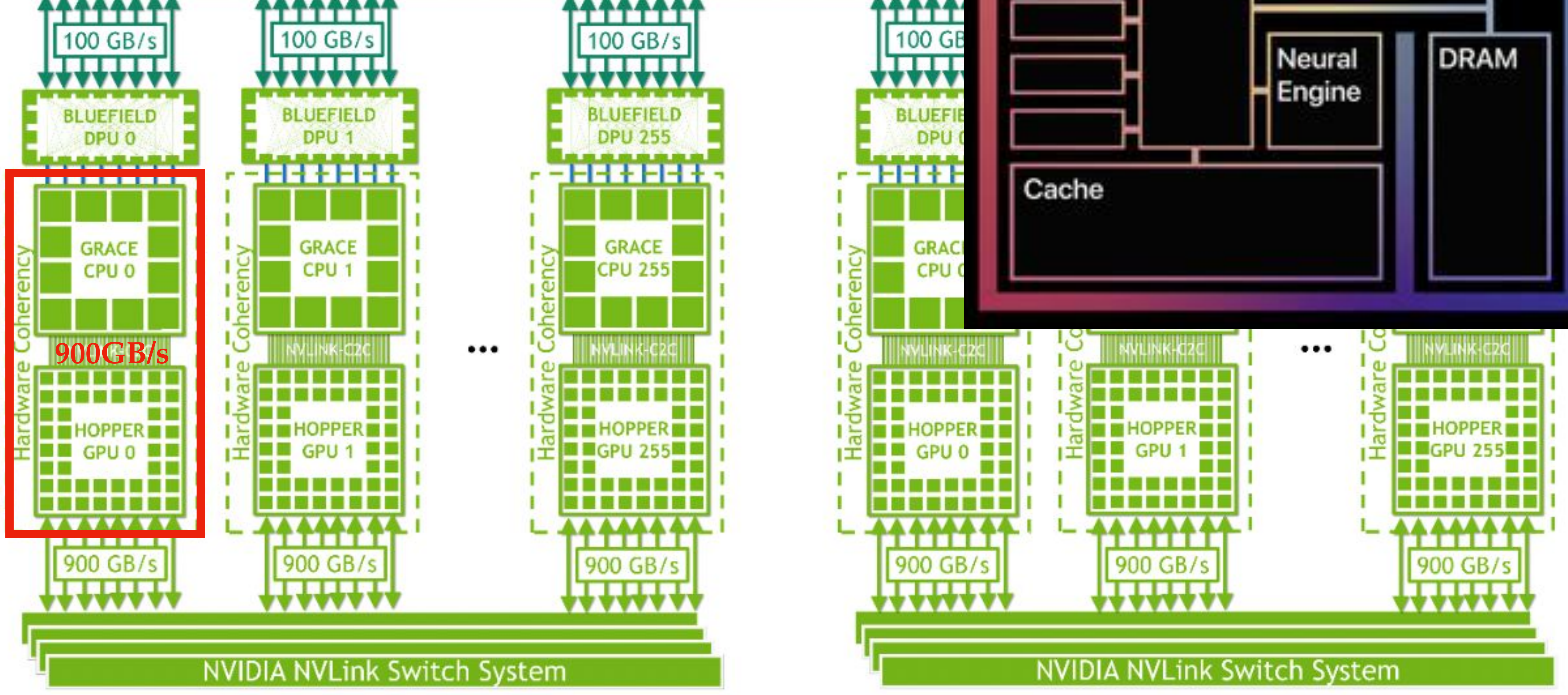
M2 Ultra



<u>**Working Hypothesis**</u>: Continued aggregation of Compute of Hardware via Memory System



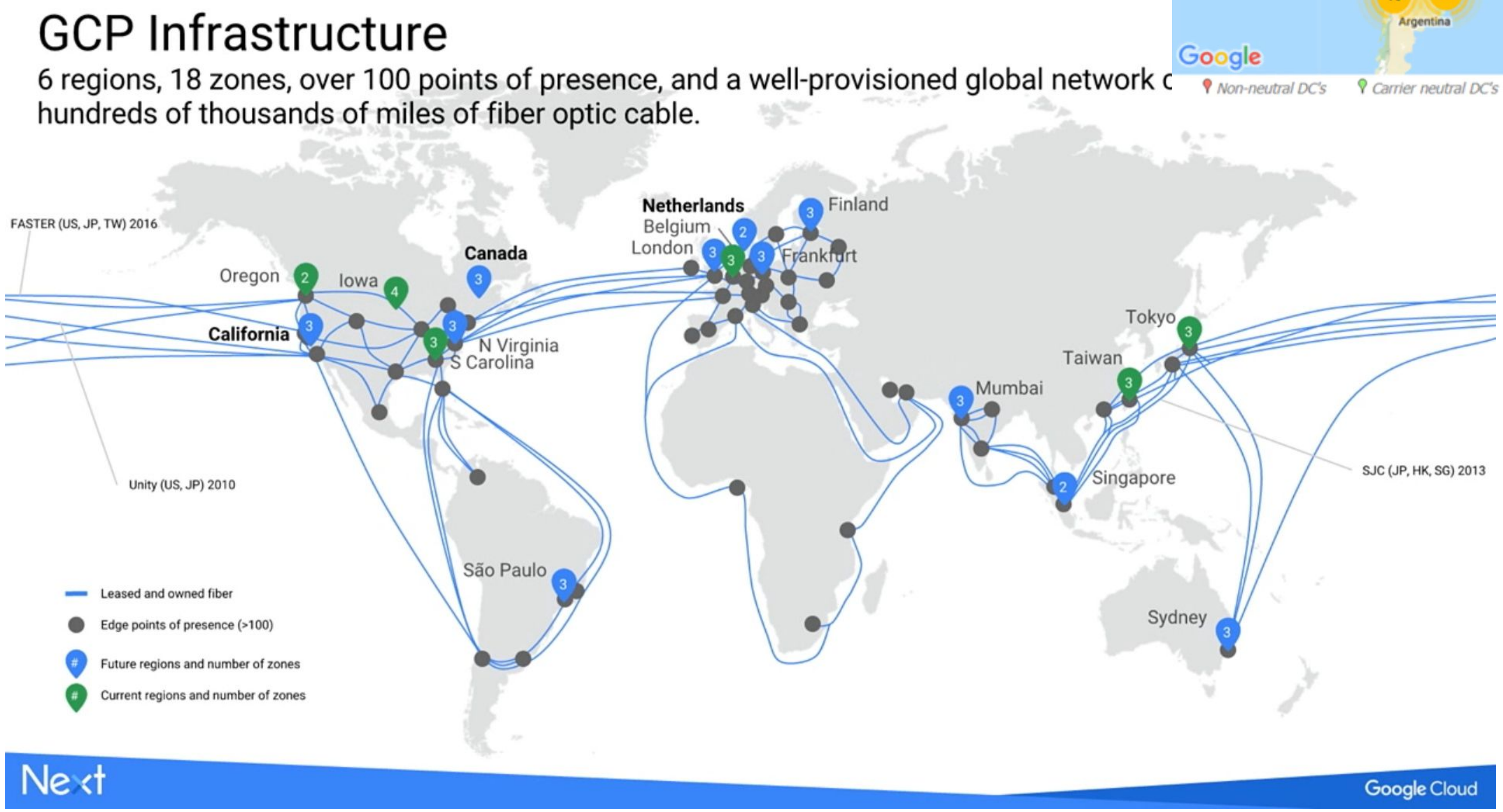
A100 GPU



GH200

Disaggregation

Working Hypothesis: Continued disaggregation of Compute across the world



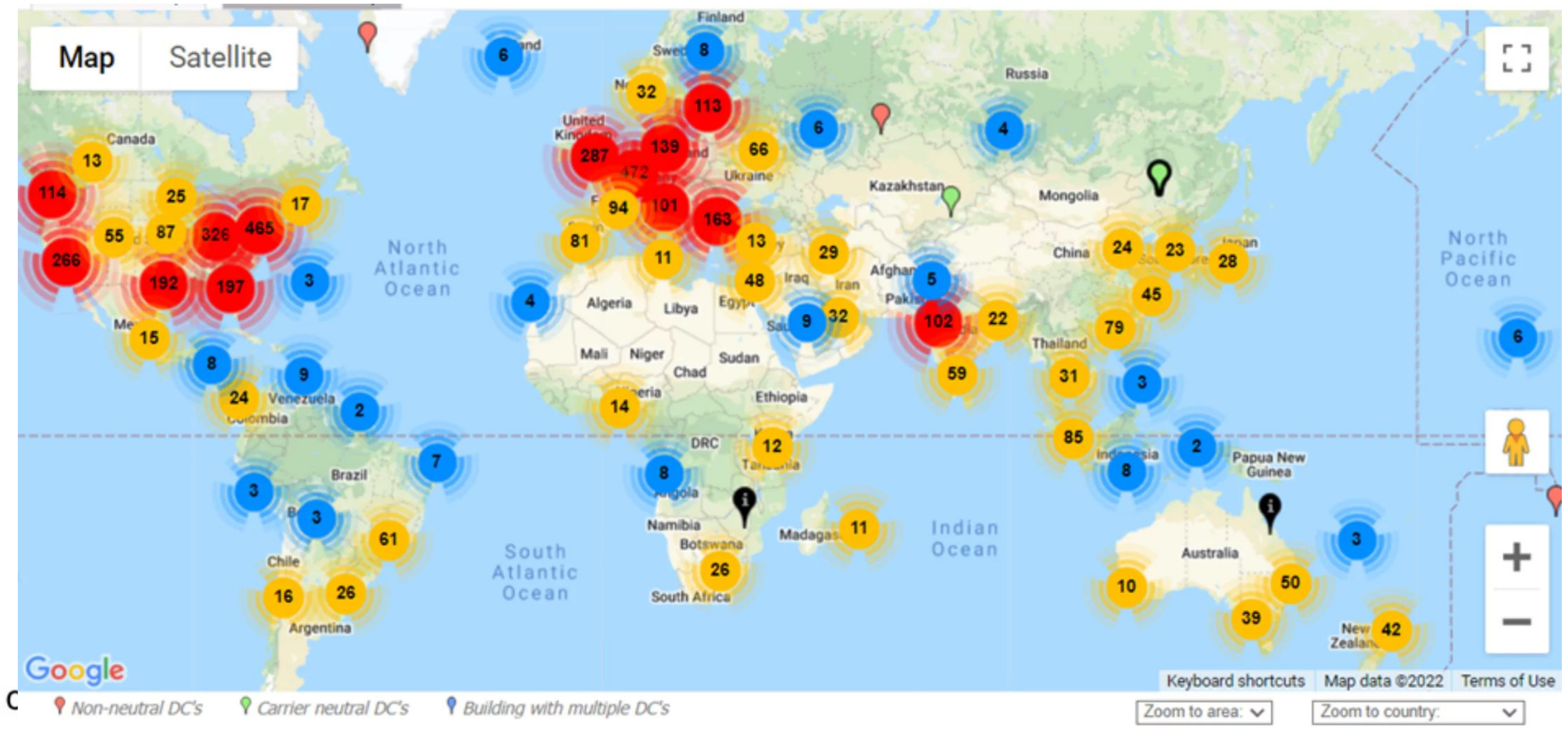


Table 3. Per component power usage

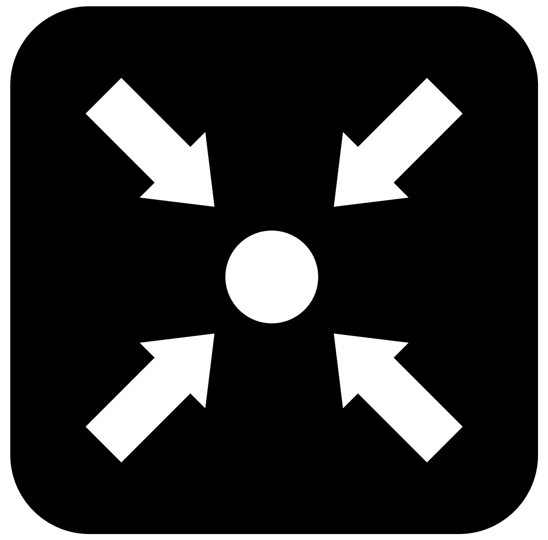
Equipment	Maximum Power
DGX A100 system	6.50 kW
Management nodes	0.60 kW
NVIDIA Quantum QM8790 switch	0.65 kW
NVIDIA SN34600C switch	0.50 kW
NVIDIA AS4610 switch	0.10 kW

Each SU requires 139 kW. The maximum power draw for a single rack is 26 kW. The total power required for the full DGX SuperPOD including storage (assumed at 20 kW) is approximately 1 MW. The rack layouts can be altered to match the power distribution and per-rack cooling requirements for a specific data center.

But maybe even more for GPU clouds:

- Power consumption
- Environmental factors
- Cost
- Limited Supply
- Easier to scale individual components
- …

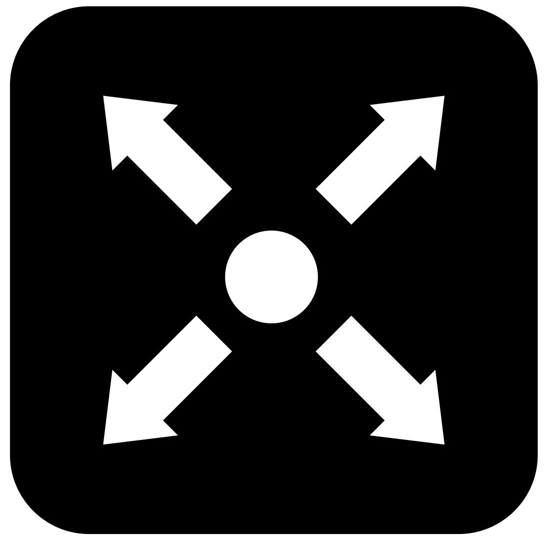
Working Hypothesis: What we see today might be fundamental in some way.



Within a Machine

Continued *aggregation* of compute of hardware (via memory system)

Optimizing data movement across the whole stack
(Offloading among different compute units; eg FlexGen)



Across Data Centers

Continued *disaggregation* of compute globally

Compressing data movement across weak communication links

Computes are there, but scattered

GPT-3

175B Parameters

3.14E+23

Floating Point Ops.

FOLDING @HOME

Could be 35 Hours

2.43 exaFLOPS (April 2020)

If the community decides to build up an Open Model at GPT-3 Scale…

we do not lack compute!

… we have even been successful in designing incentives for people to contribute computes

GPT-3

175B Parameters

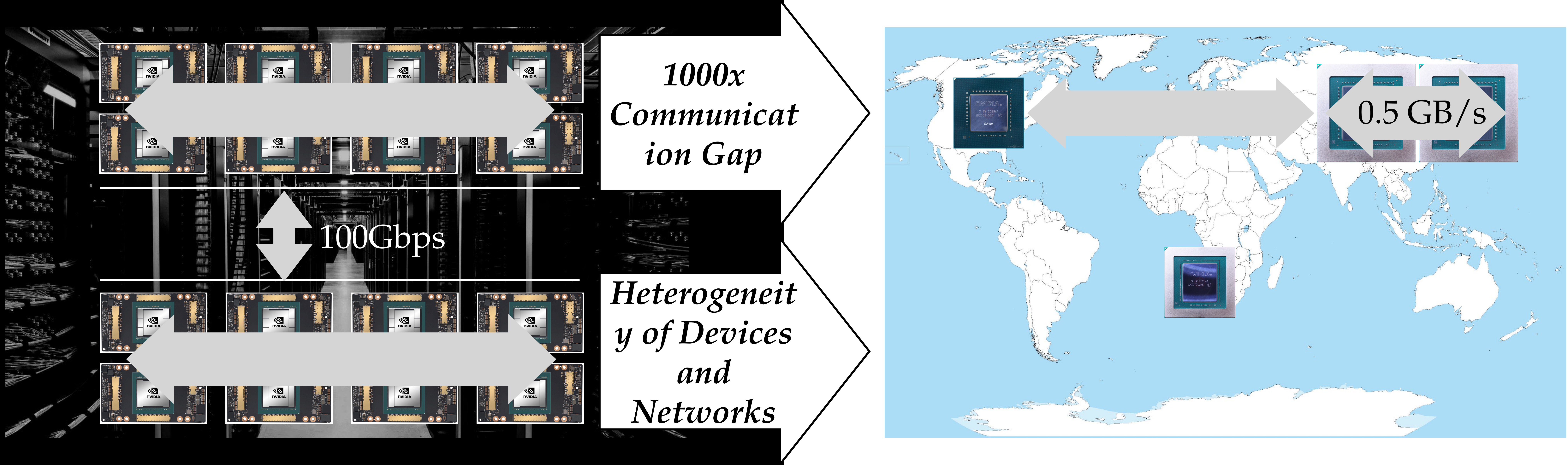
3.14E+23

Floating Point Ops.

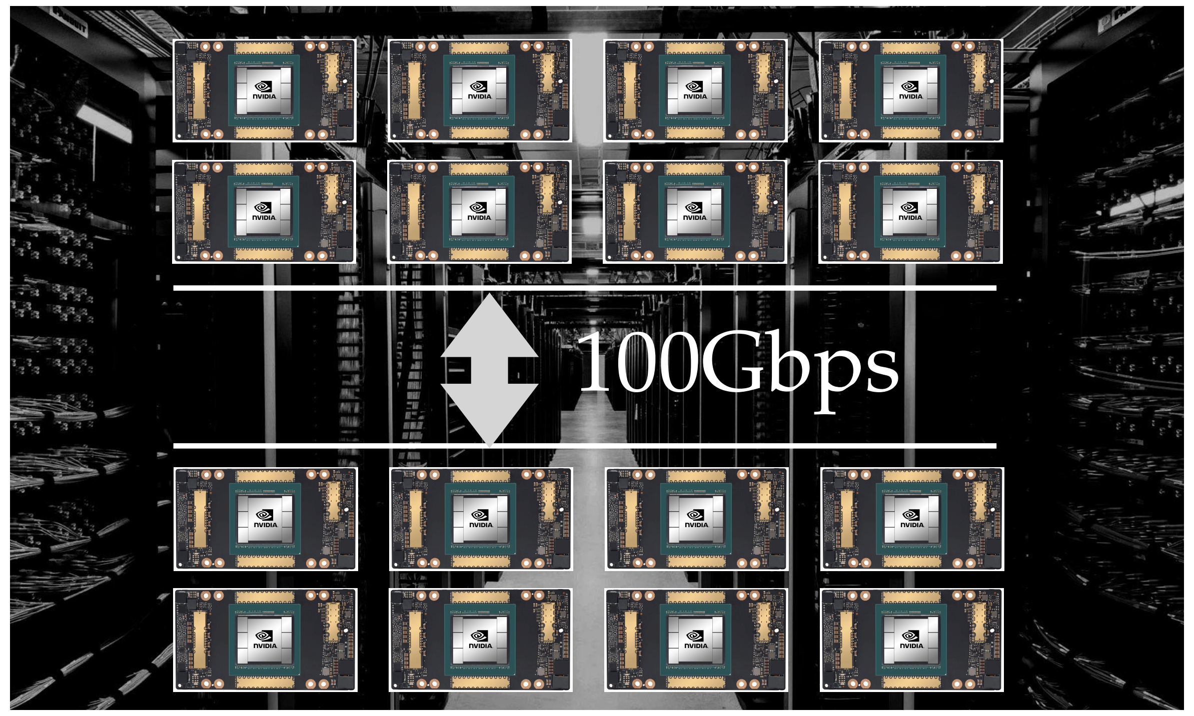
FOLDING @HOME

Could be 35 Hours

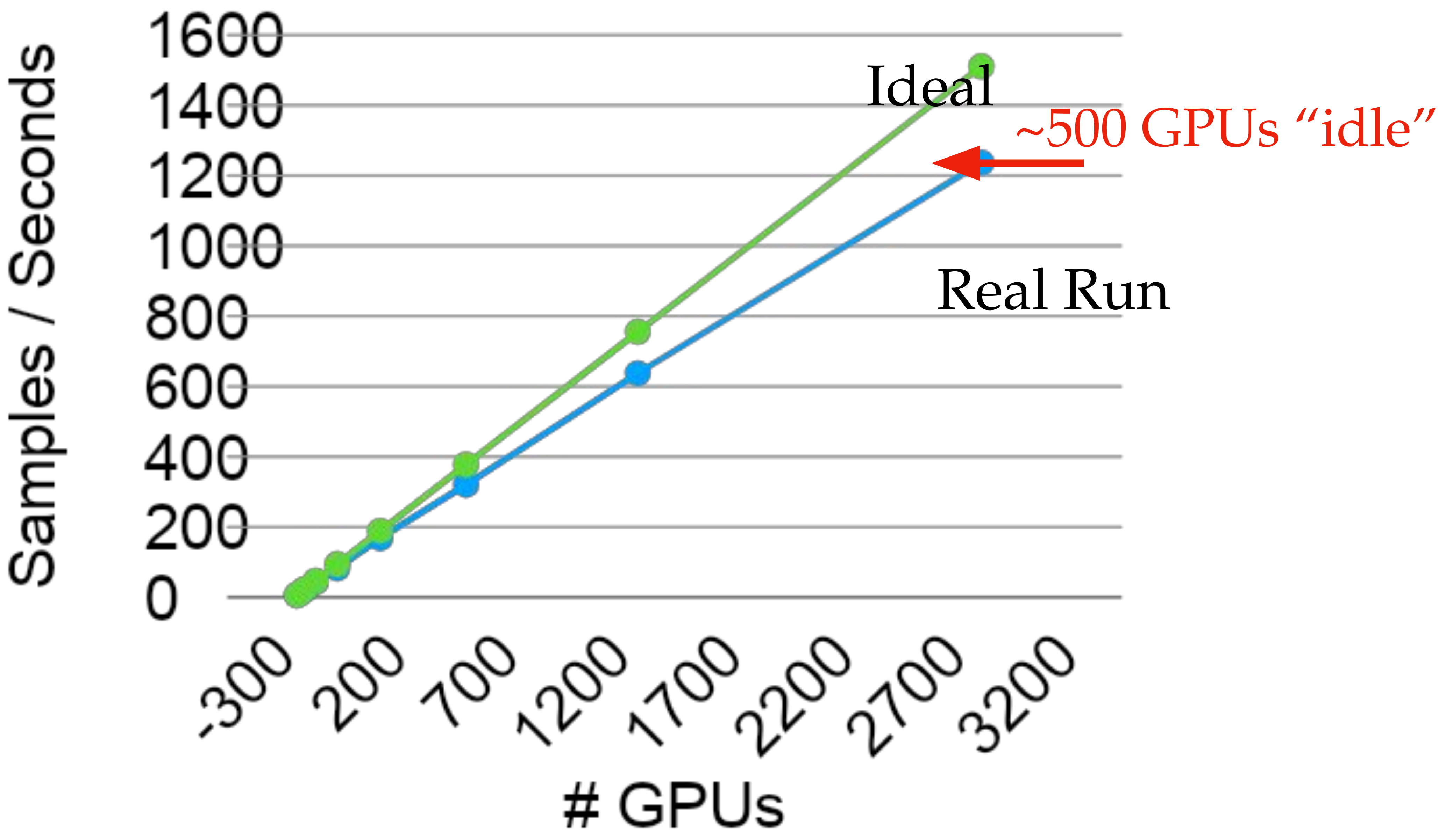
2.43 exaFLOPS (April 2020)



Communication Bottlenecks across Infrastructure

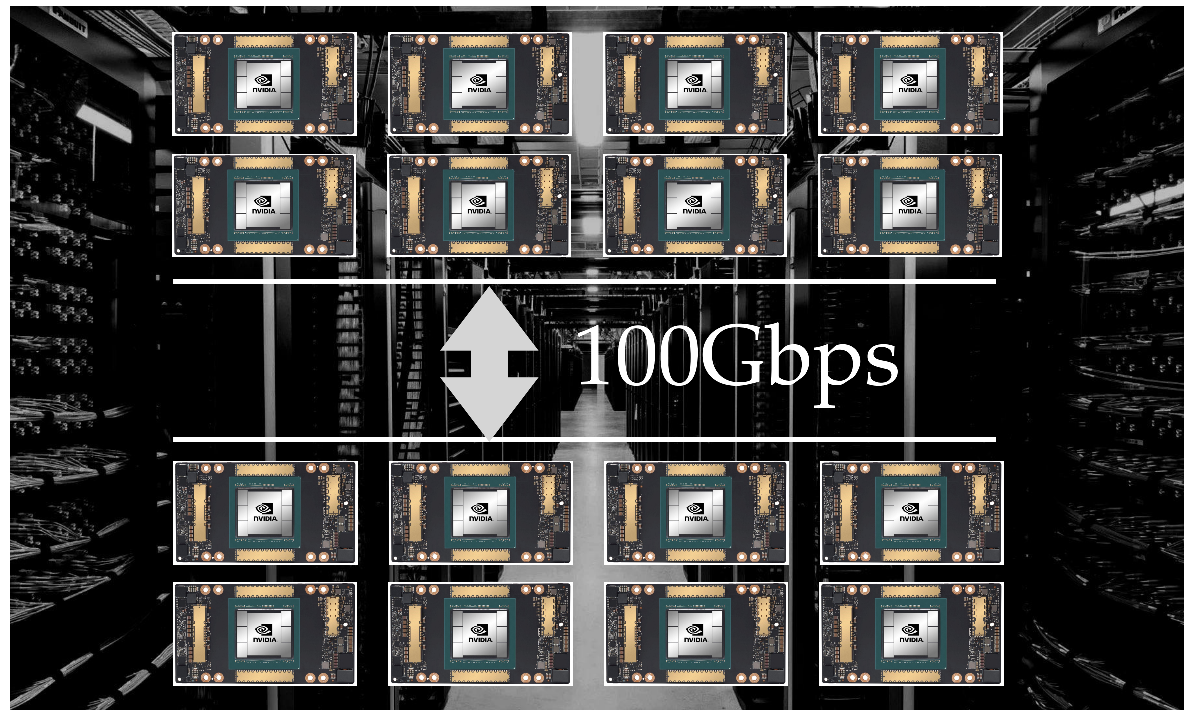


Data Center



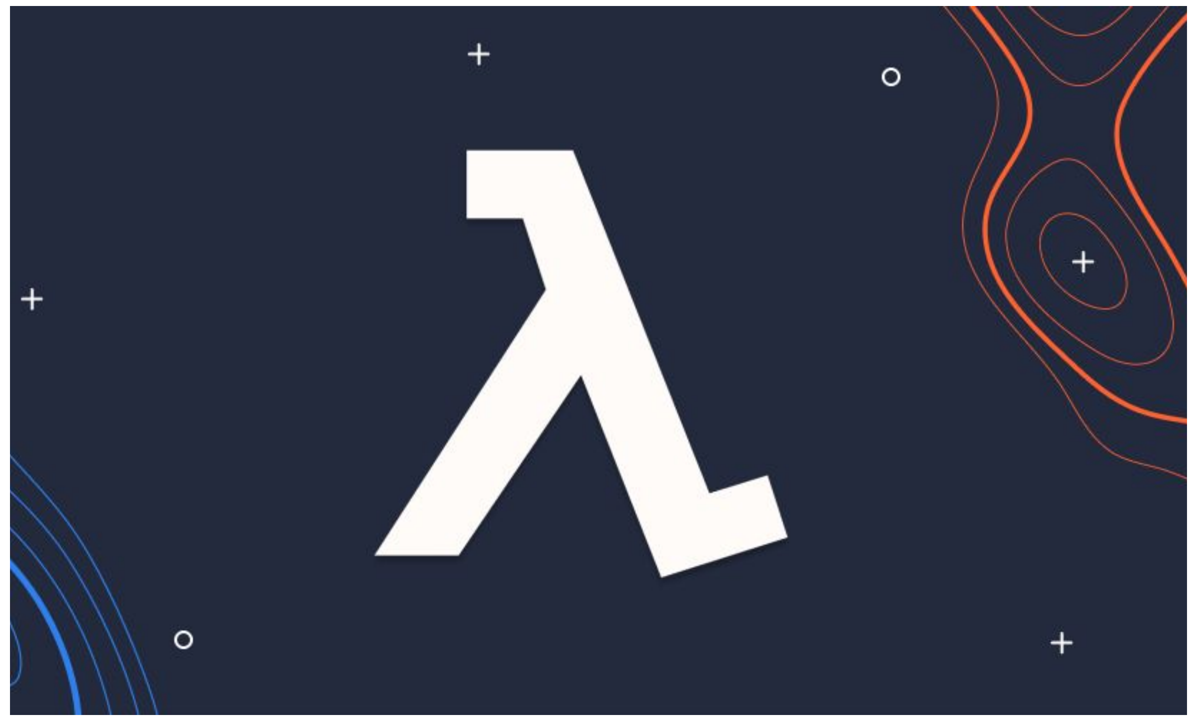
Communication Bottlenecks across Infrastructure

communication becomes slower, open up more choices (and some can be cheaper)

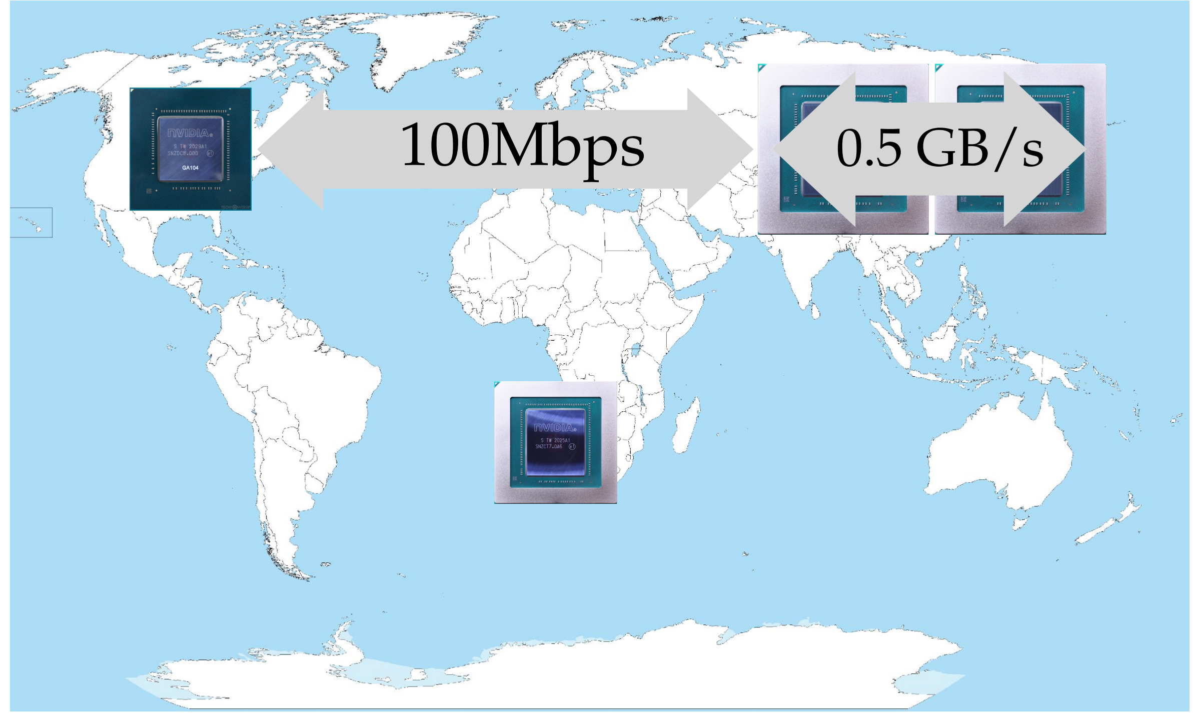


Data Center

(Multi-cloud) Spot Instances



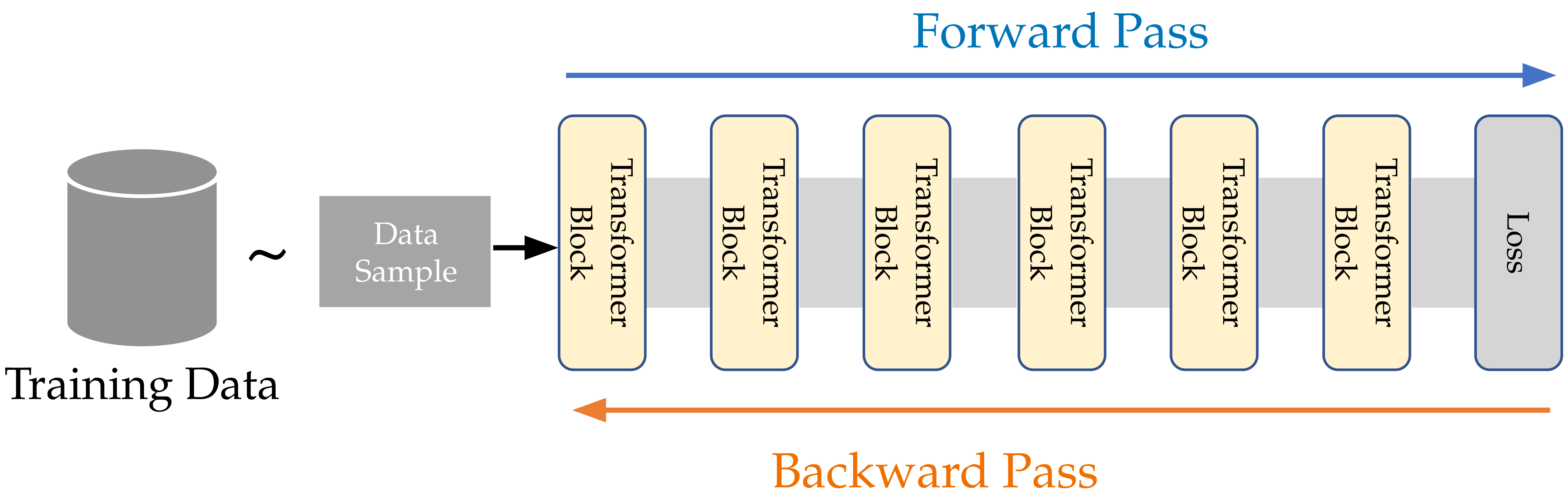
Serverless Environment



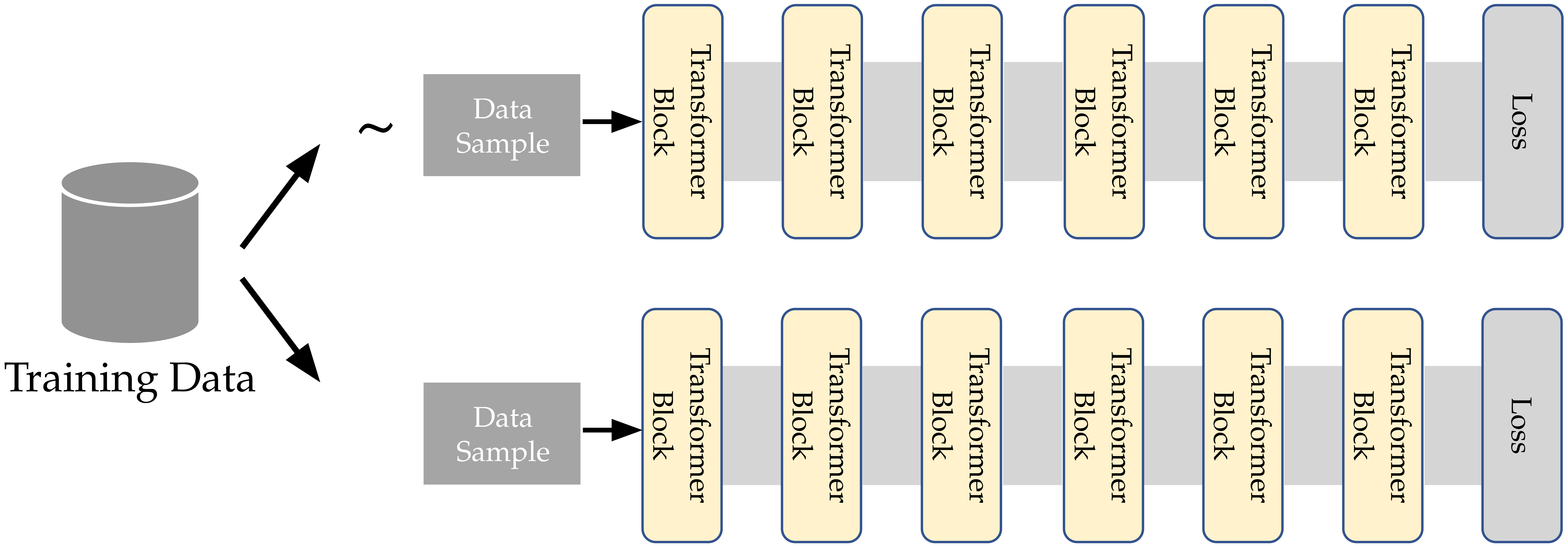
Decentralized Network

The more we can optimize communications, the more choices we have when building our infrastructure.

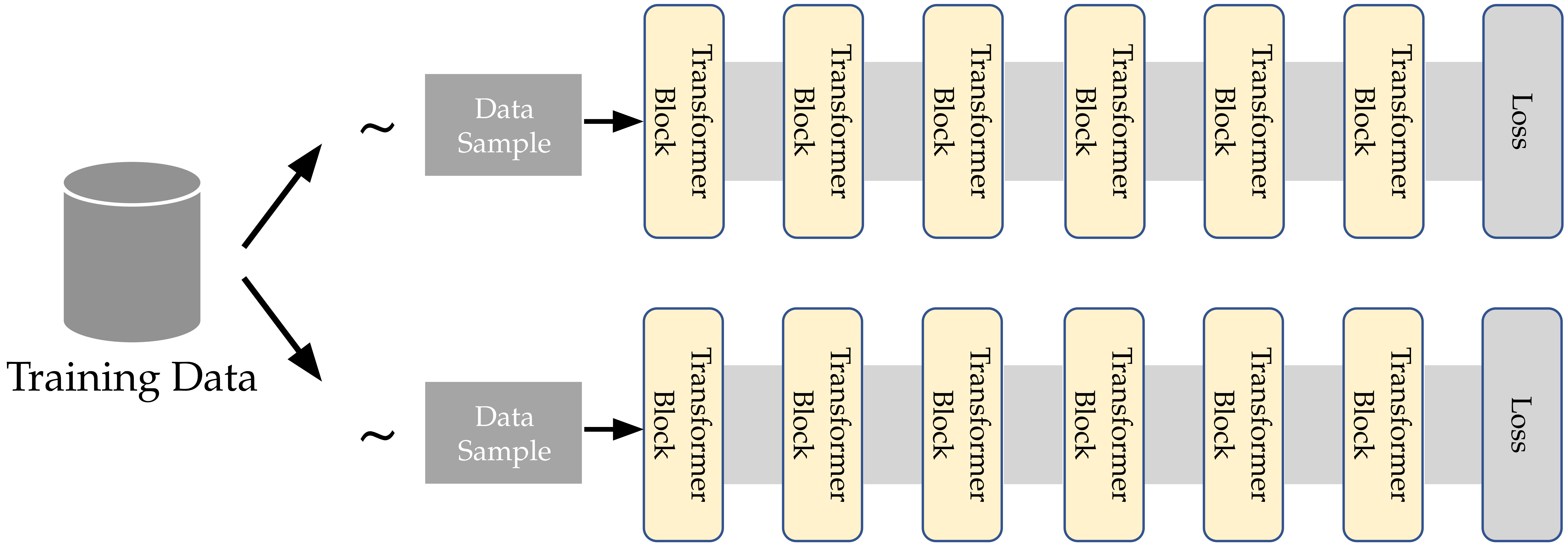
Training Neural Networks



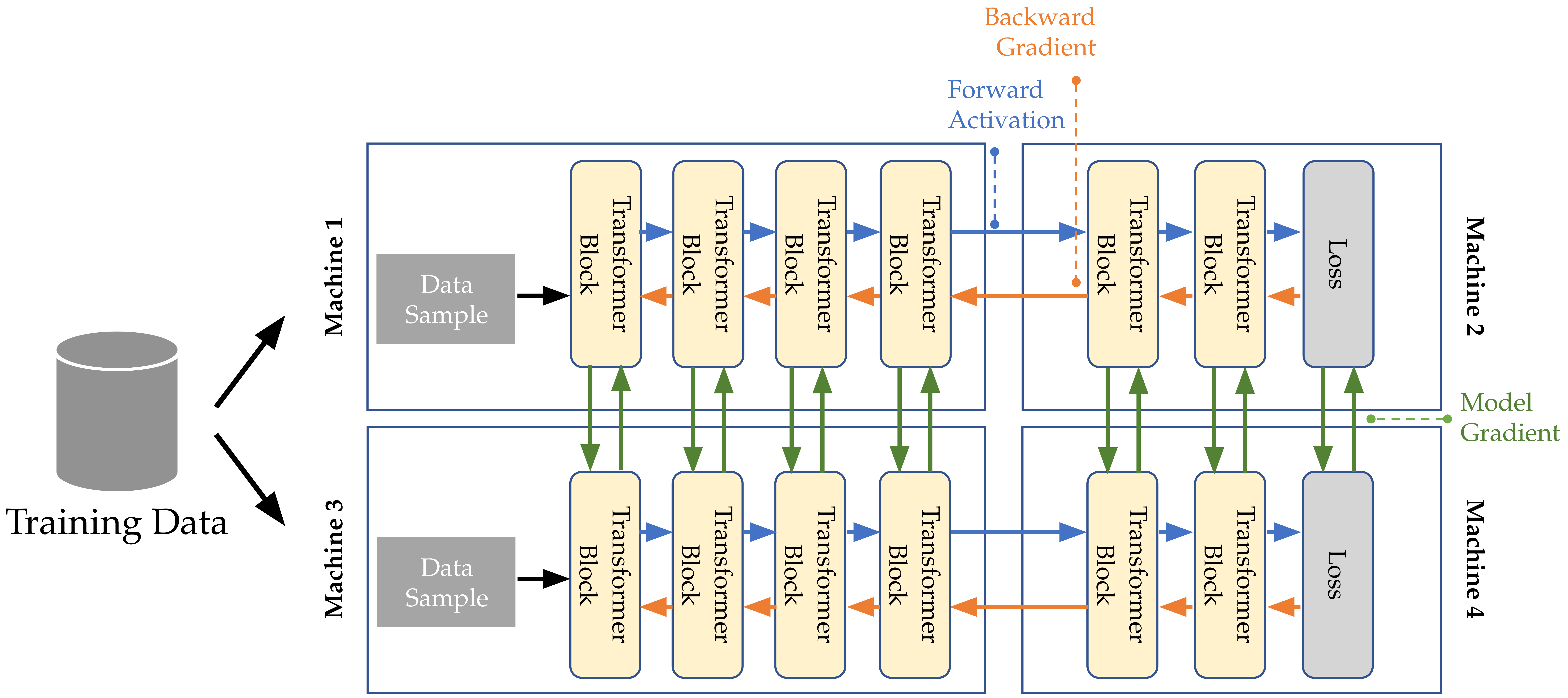
Training Neural Networks



Training Neural Networks

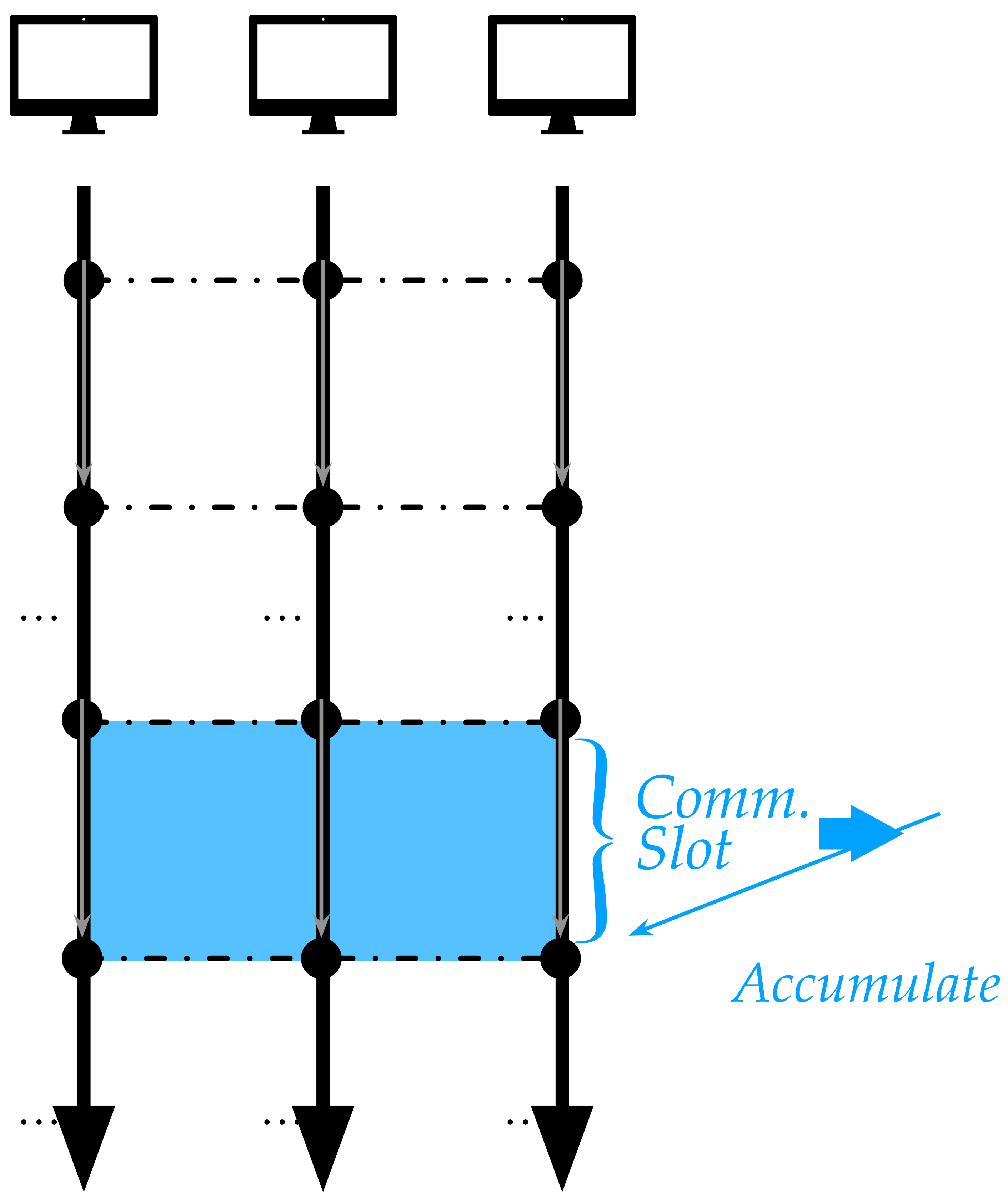


Takeaways



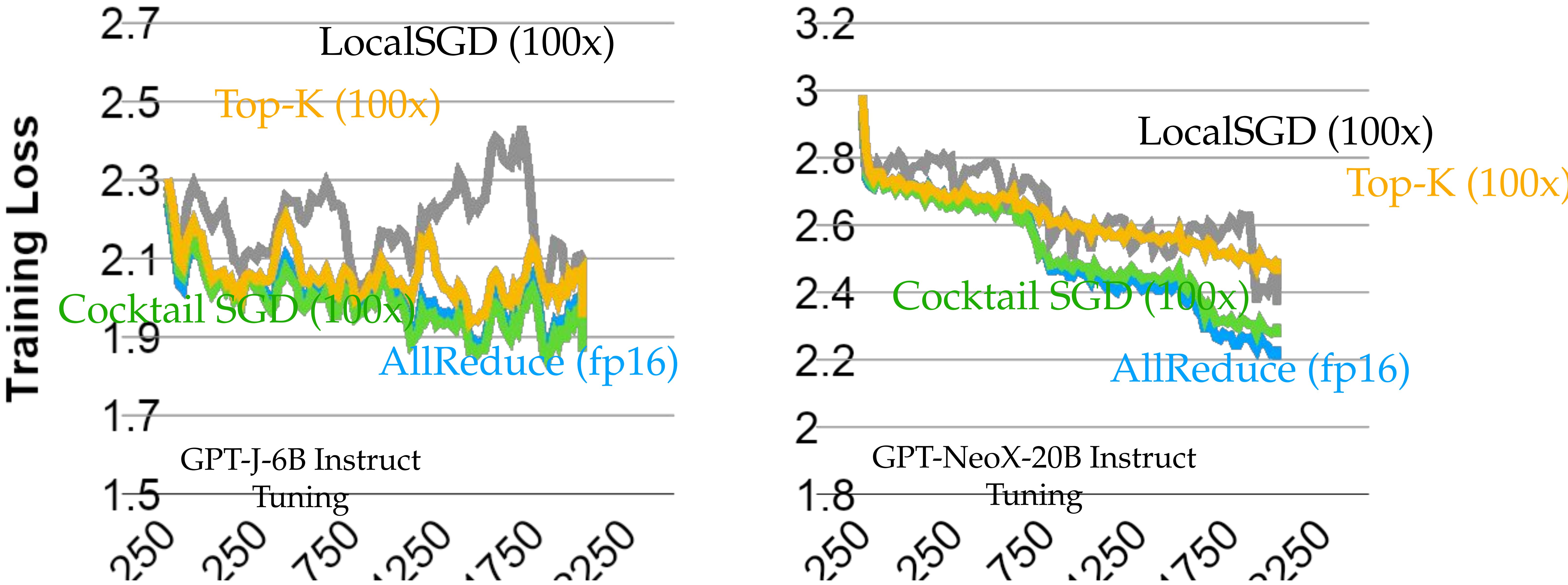
<u>Main Algorithmic Takeaway</u>: all <u>*three*</u> communication channels in the system can be compressed aggressively, without hurting the model quality

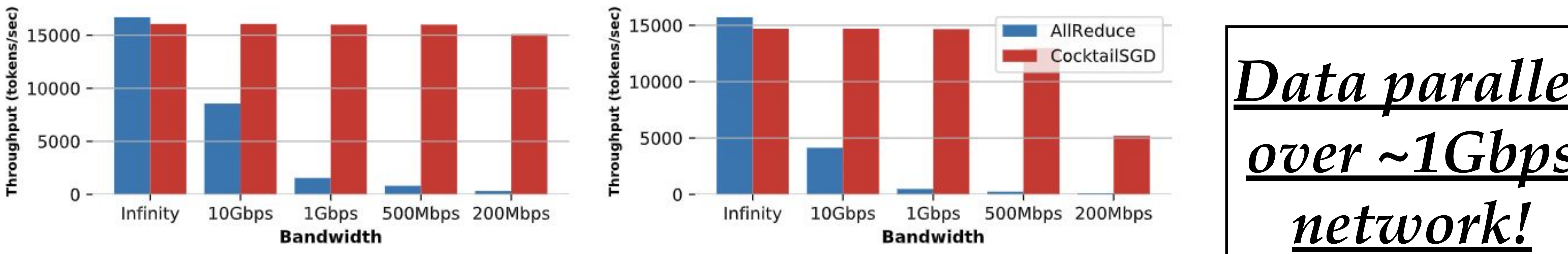
“Cocktail SGD”: Data Parallel over 1Gbps



As long as Communication fully fills the Comm. Slot, no slow down caused by communication.

Different communication compression techniques complement each other and compose well!



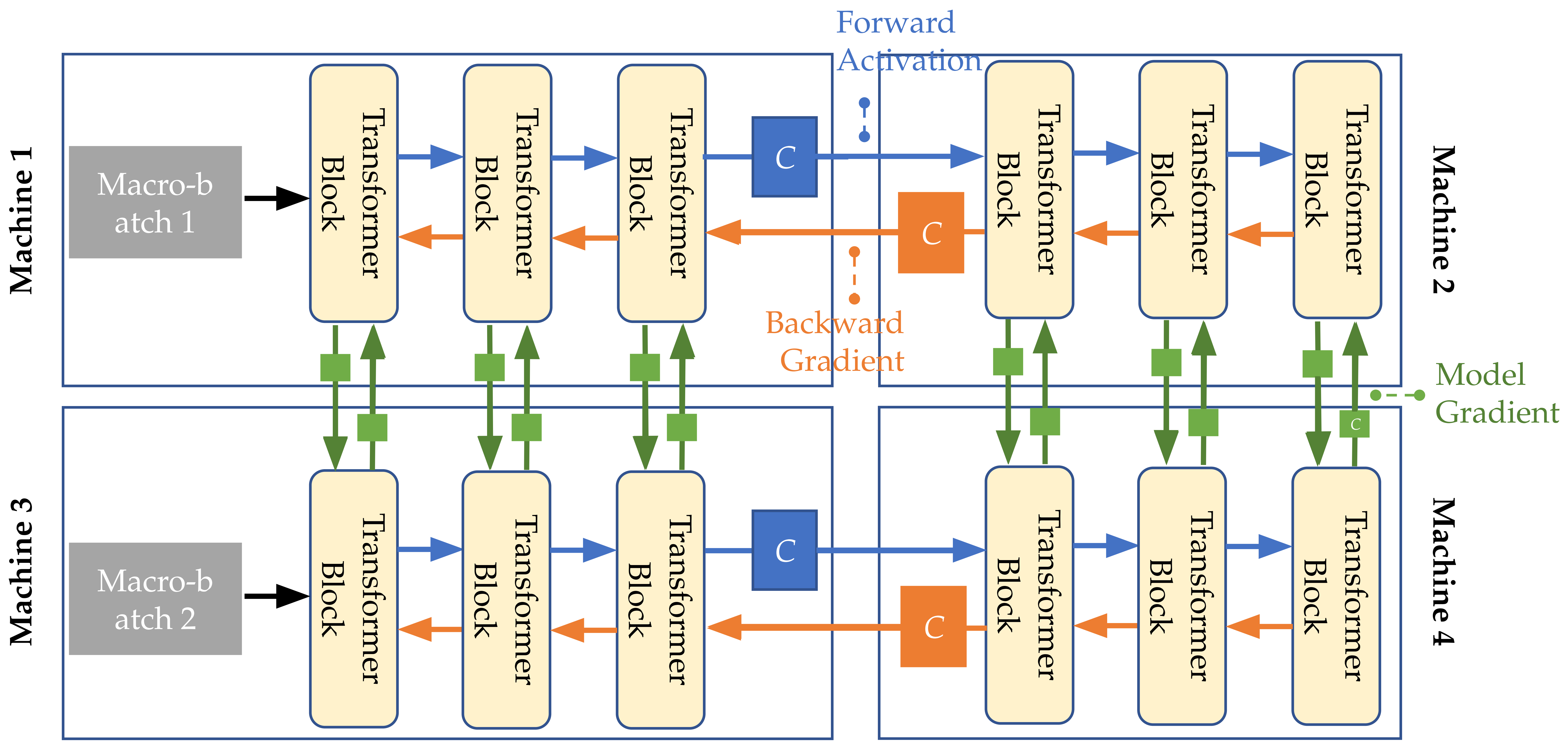


(b) GPT-J-6B

(c) GPT-NeoX-20B

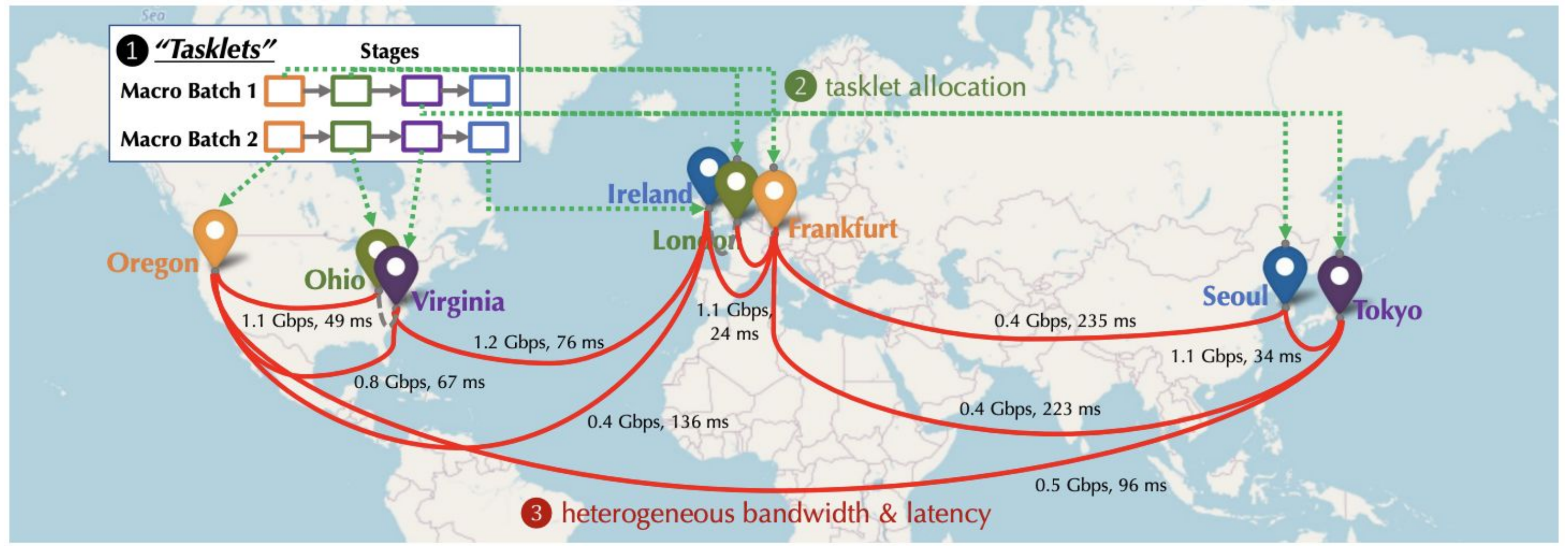
Data parallel over ~1Gbps network!

All Challenges Can Be Compressed

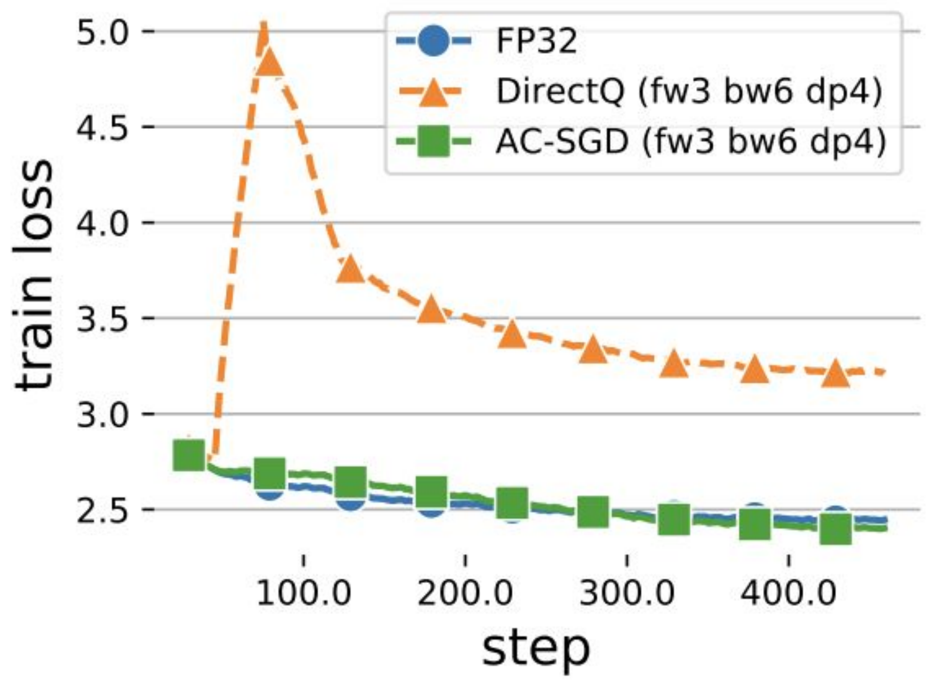


End-to-end Compression

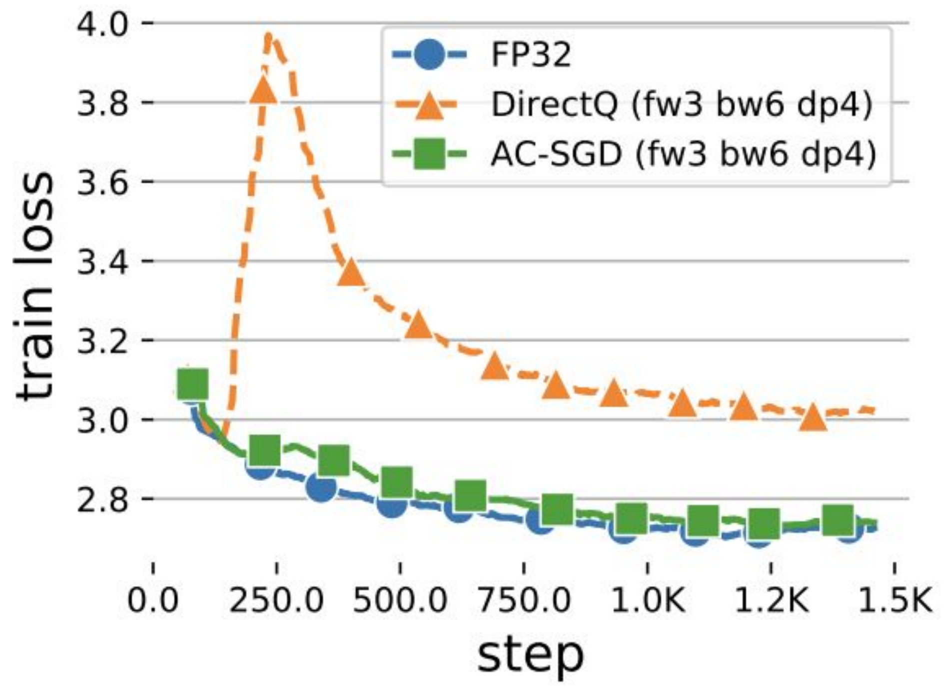
All Communication Channels are Compressed



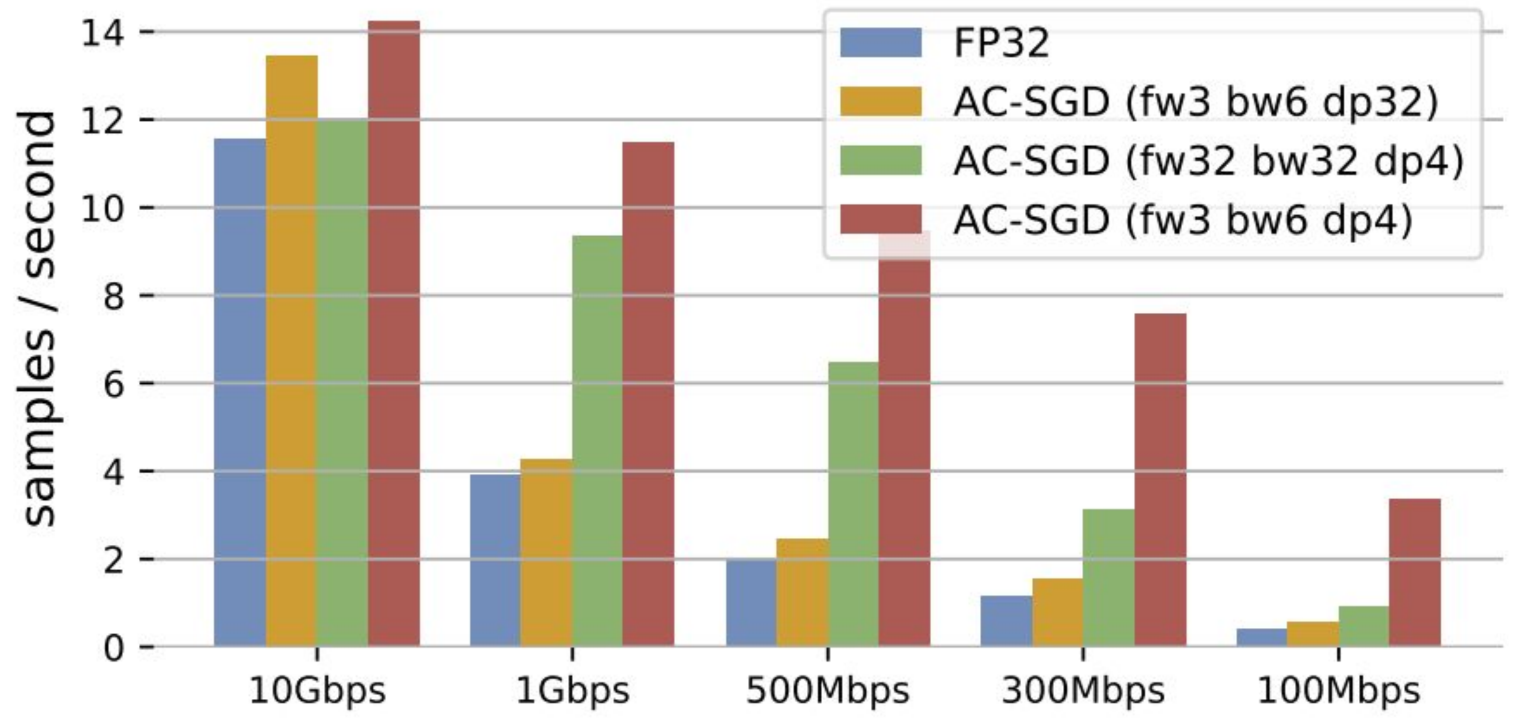
3 bits forward
6 bits backward
4 bits DP gradient
8 cuts



(a) WikiText2, GPT2-1.5B

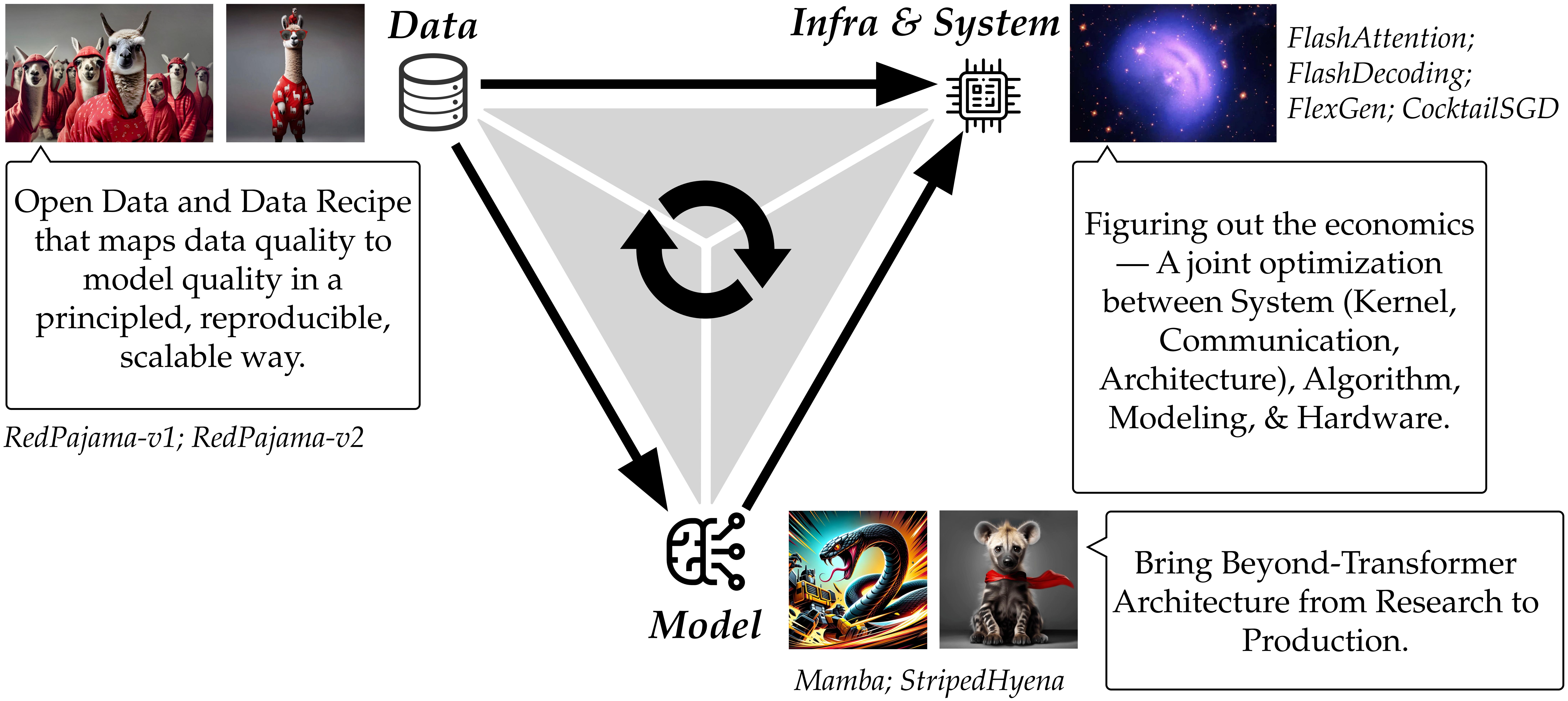


(b) arXiv, GPT2-1.5B



(c) Training Throughput

q


Data
Infra & System
Model
Open Data and Data Recipe that maps data quality to model quality in a principled, reproducible, scalable way.
RedPajama-v1; RedPajama-v2
FlashAttention; FlashDecoding; FlexGen; CocktailSGD
Figuring out the economics — A joint optimization between System (Kernel, Communication, Architecture), Algorithm, Modeling, & Hardware.
Mamba; StripedHyena
Bring Beyond-Transformer Architecture from Research to Production.

The Open Source Community, Thank you!