

Escaping the Matrix: AI and Submarine Cables

Alan Mauldin

Submarine Networks World 2024

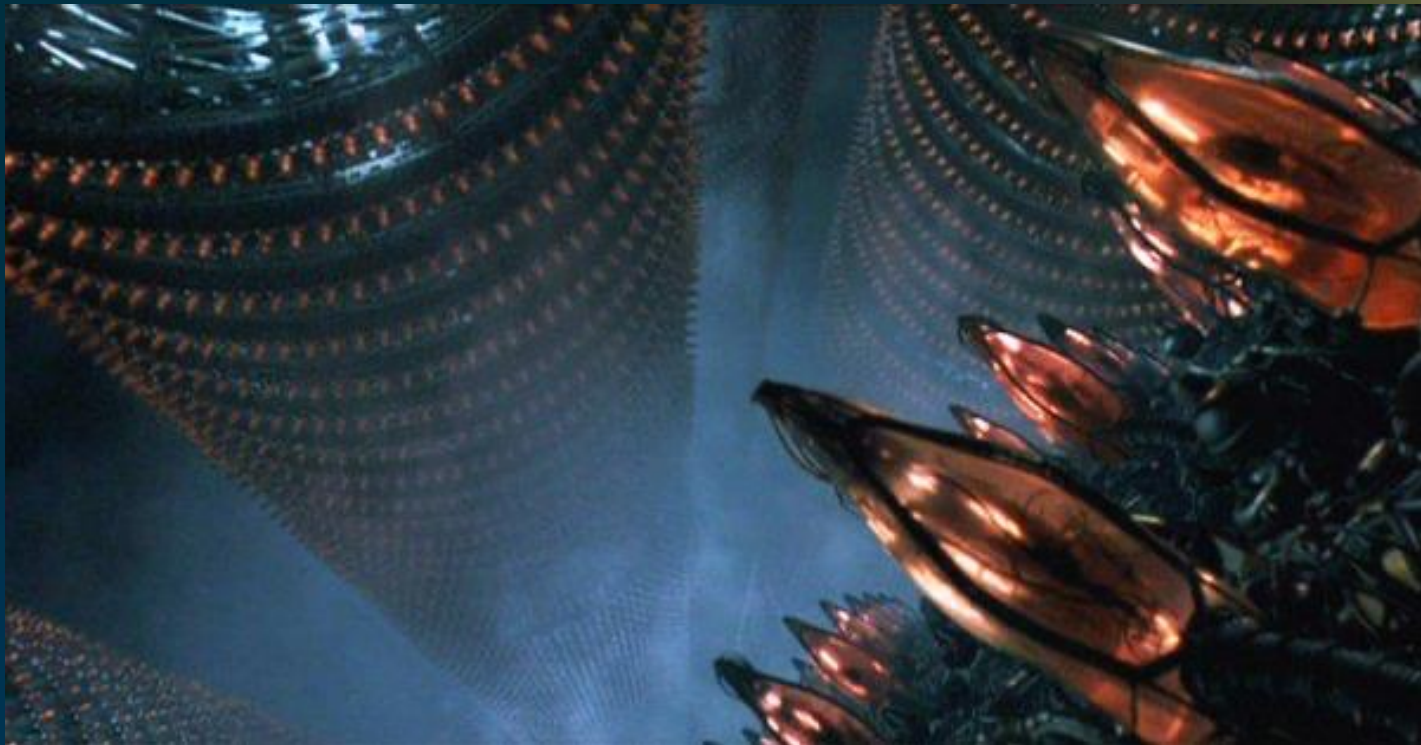
September 26, 2024

Free your mind: What is real?

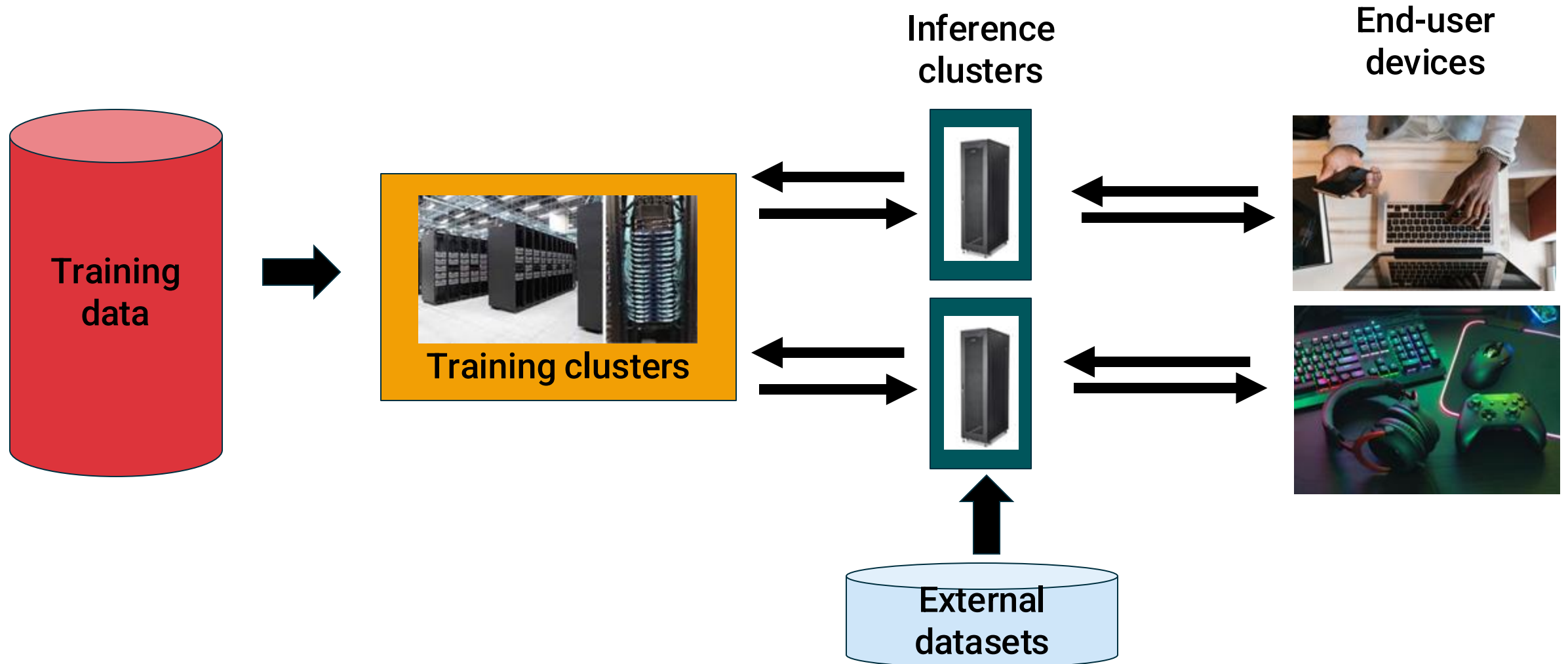


- How big are AI training data sets and models?
- How might AI impact demand for submarine cable bandwidth?
- Is there an AI-fueled bubble in new submarine cables?

How Does AI Work



(Roughly) how does AI work?



Training Data and AI Models



How big are AI training data sets?

- Factors that influence AI training data size
 - Model complexity: more parameters $\square$ more training data
 - Model goal: training data need to be refined/curated depending on the model's goal.
 - Data quality: higher quality training data $\square$ more accurate models
- Sample public source data sets used for model training
 - The Pile – 886 GB
 - OpenImages – 18 TB
 - YouTube 8m – 32.5 TB
 - Common Crawl – 327 TB (August '24 crawl only)
- Models are trained on multiple data sets
 - Google's Gemini training data:
 - Infiniset (Google proprietary)
 - Common Crawl
 - Wikipedia
 - Scientific Papers and more

August 18, 2024

August 2024 Crawl Archive Now Available

The crawl archive for August 2024 is now available. The data was crawled between August 3rd and August 16th, and contains 2.3 billion web pages (or 327.4 TiB of uncompressed content).

How big are AI training data sets?

- Training data set size is soaring
 - For example, in 2022 Meta claimed they have exabytes of training data.
 - The company has reported a 1.75-2x increase in the amount of data they trained models on between 2020-22, and 3-4x increase in the data ingestion throughput.

Data ingestion pipeline overview

We have exabytes of training data powering our models, and the amount is growing rapidly. We have a wide variety of models that train on terabyte scale data, but we do not have the storage capacity at that scale to train



Growth Trends

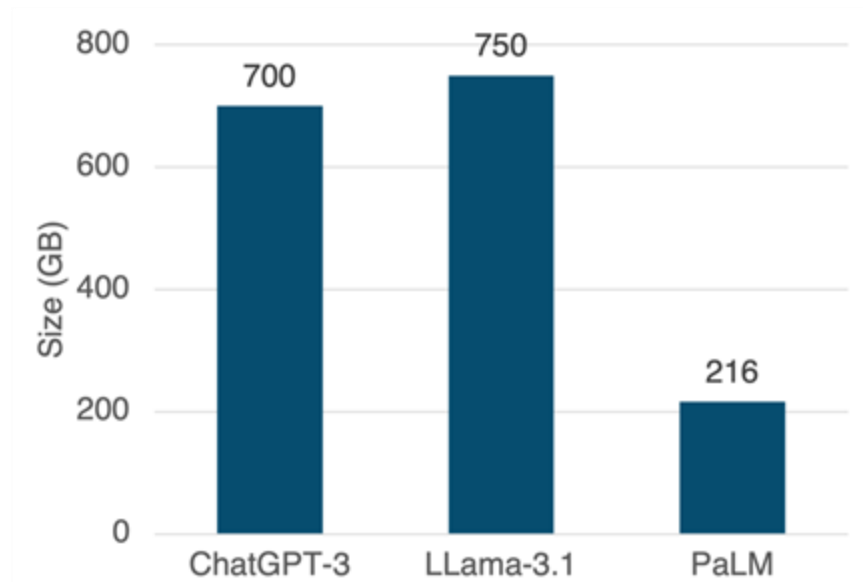
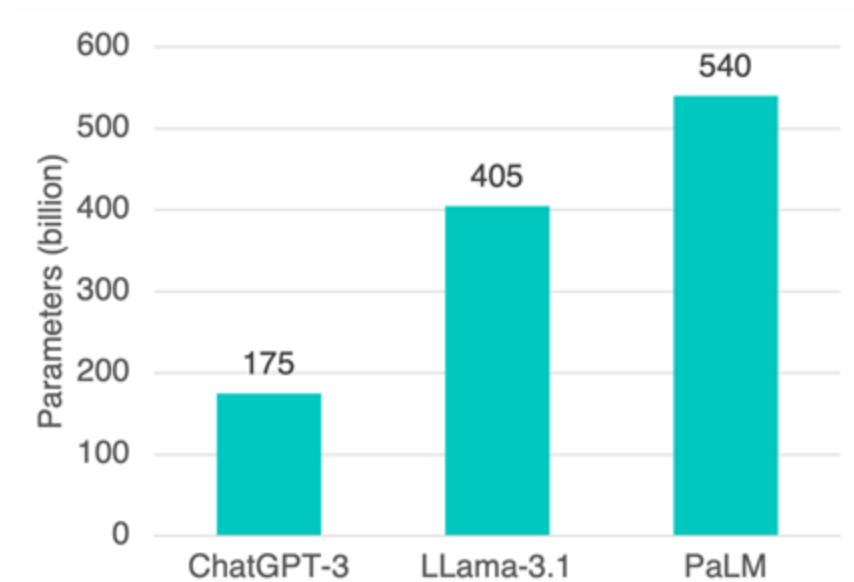


Source: Meta, <https://engineering.fb.com/2022/09/19/ml-applications/data-ingestion-machine-learning-training-meta/>

How big are AI models?

- Factors that influence AI model size
 - Number of parameters, model architecture, volume and type of training data, compression techniques
- High parameter count does not always lead to larger model size

Number of Parameters versus Model Size



Source: Public sources

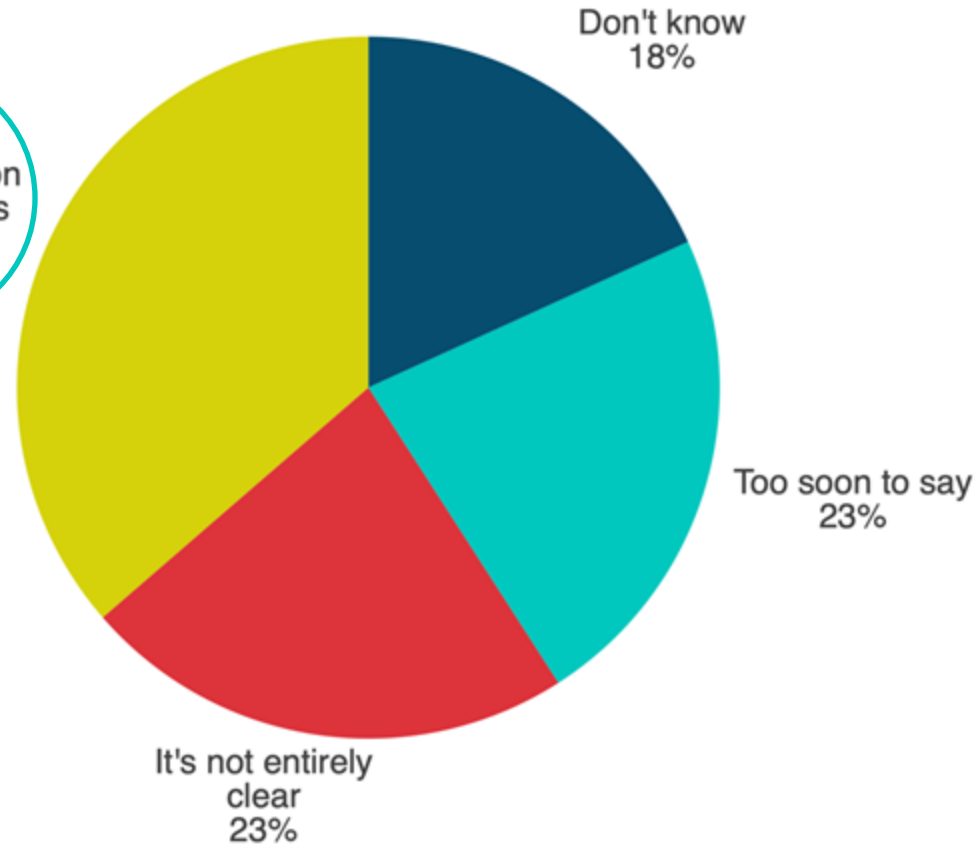
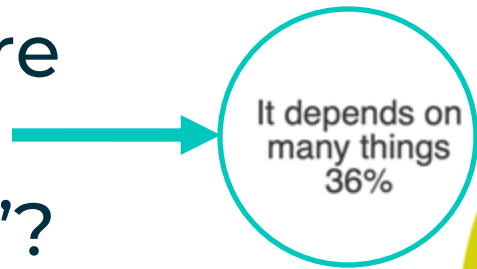
AI and Submarine Cable Bandwidth



Exhaustive survey results

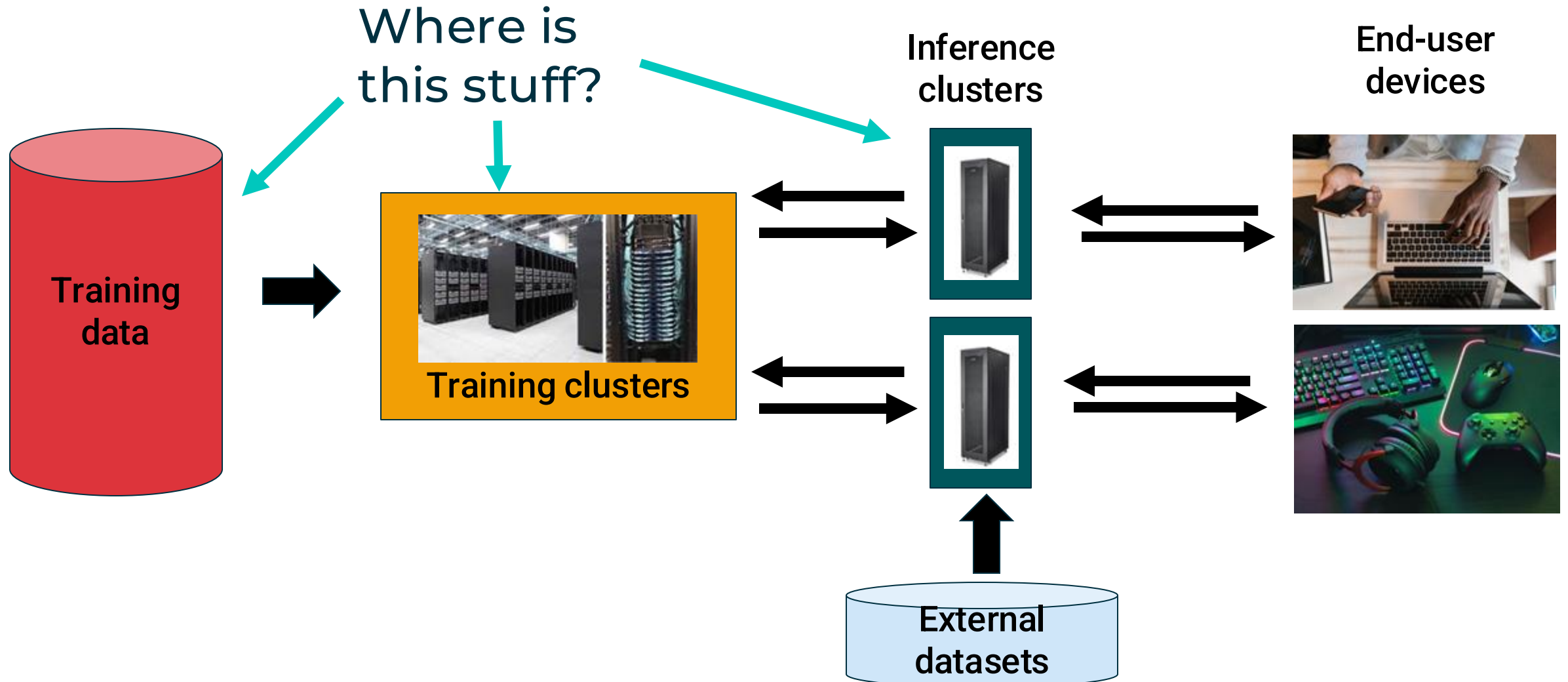
How will AI impact demand for submarine cable bandwidth?

What are these “things”?



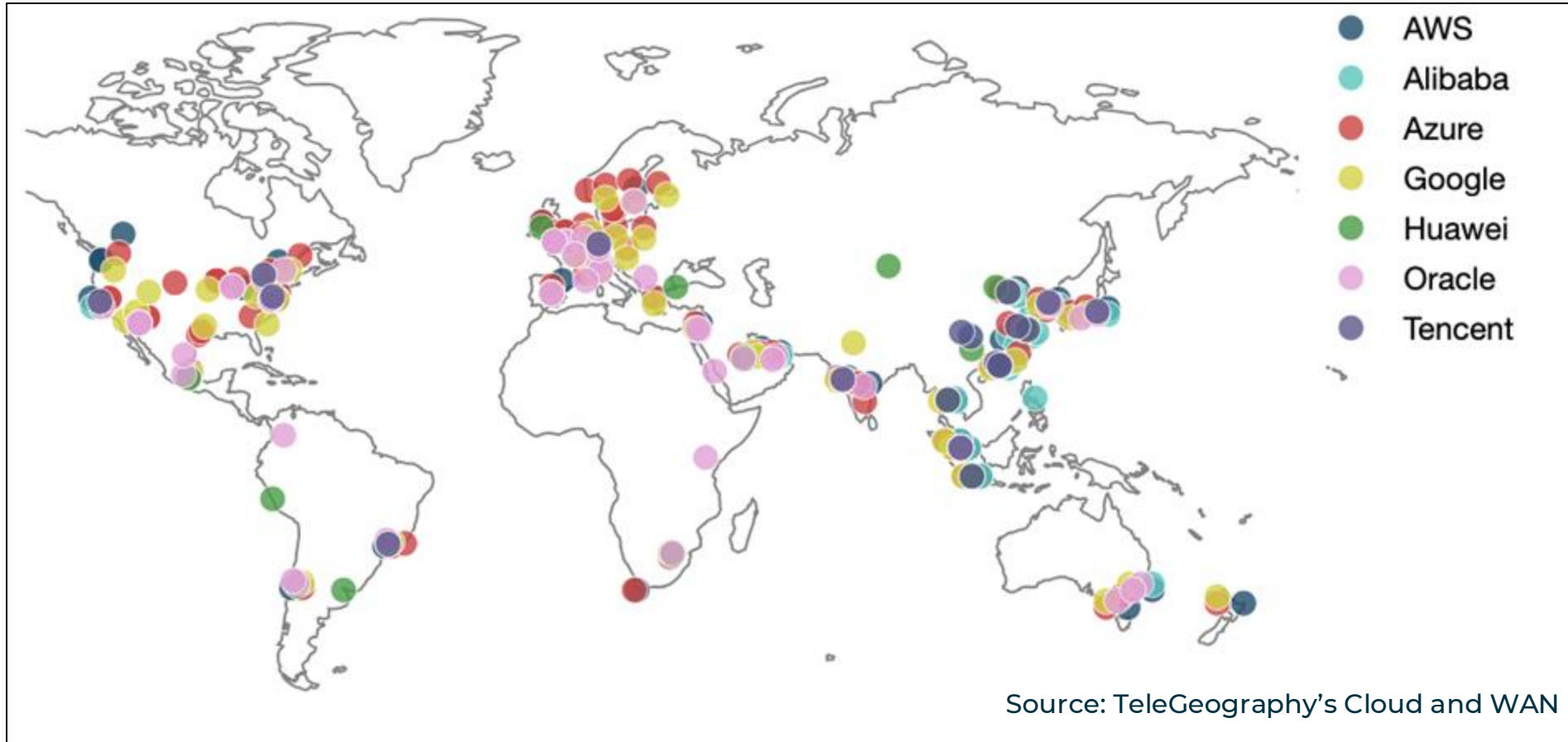
Source: Non-scientific fake survey, 2024

Location matters



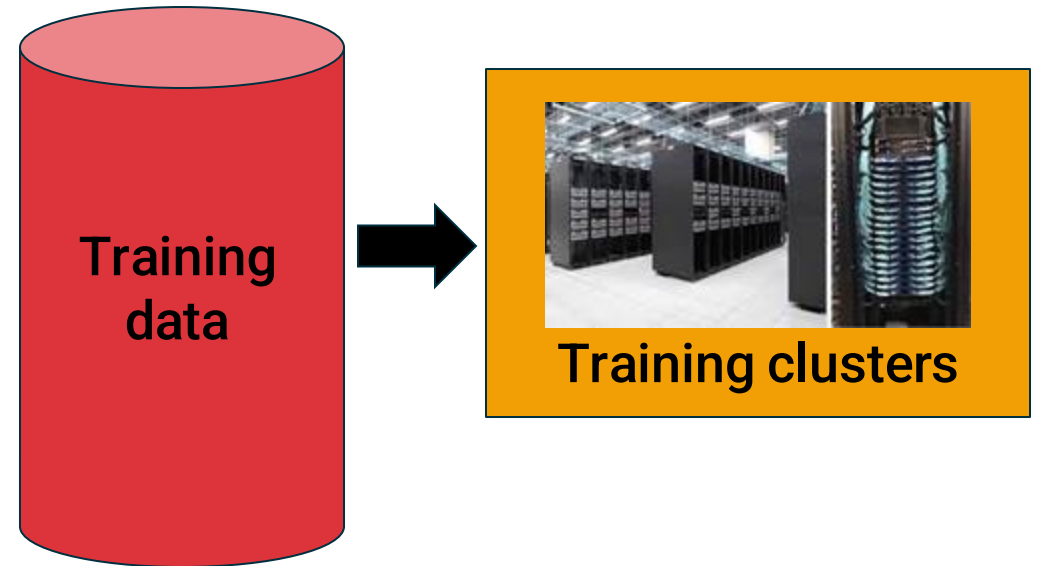
Focused heavily in the cloud

Existing and Planned Cloud Data Centers



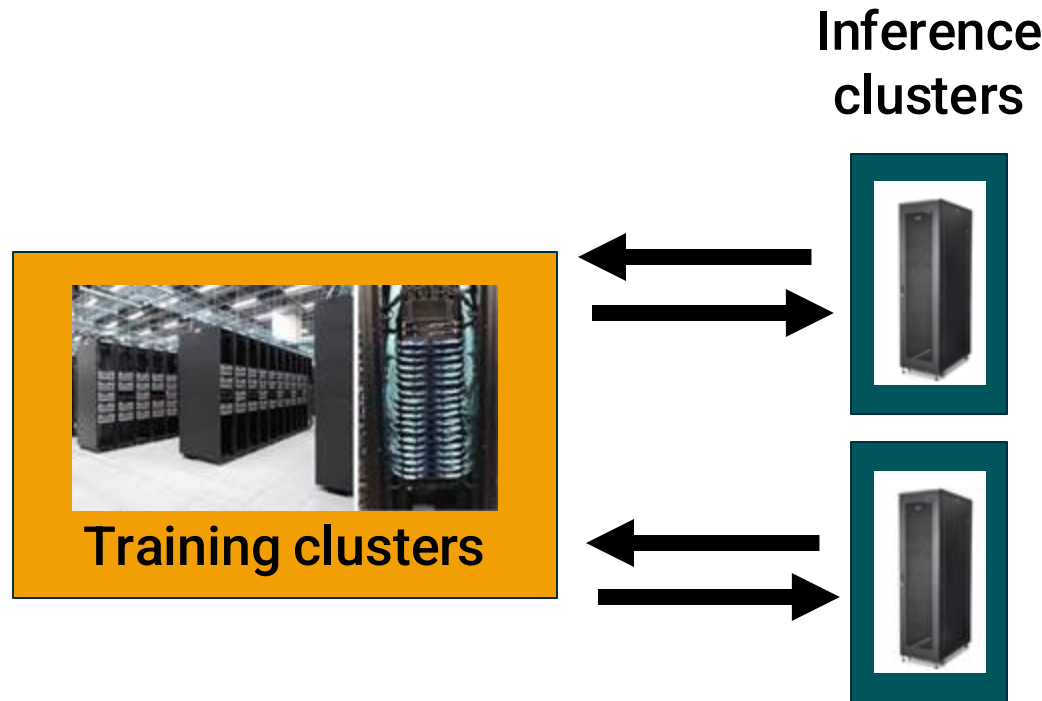
Training

- Model training data -> Training clusters
 - Collecting the training data can lead to large demand spikes when it is transported into training clusters.
 - Frequency of training
 - Role for cables: Training may be conducted in locations far way from training data storage.
- Model training
 - Role for cables: not much. In training clusters, the GPUs and the training data generally need to be in close proximity.



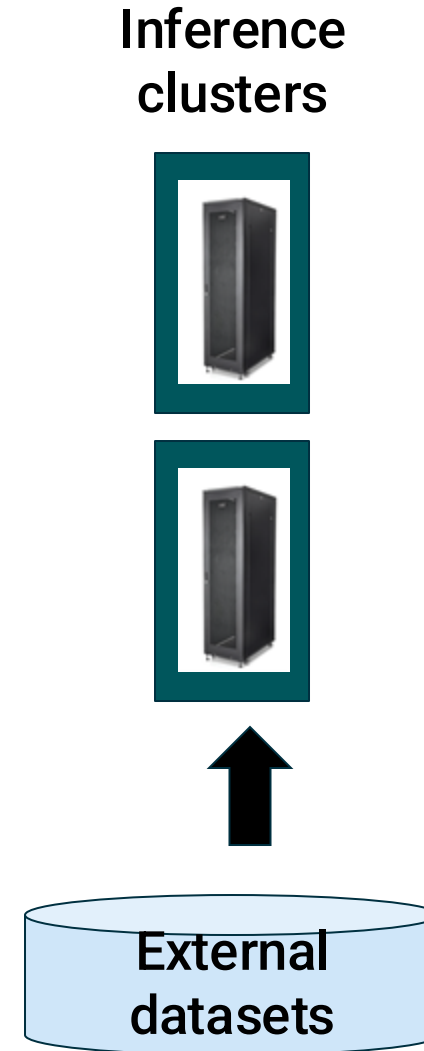
Training

- Training clusters -> Inference clusters
 - Frequency of updates matters.
 - Role for cables: need to distribute trained models to inference clusters.



Inference

- Inference clusters <-> External data sets
 - Role for cables: Possibly some external data accessed during inference may be located in a different country.



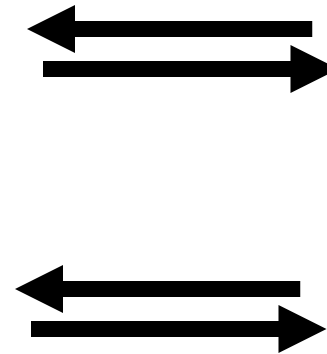
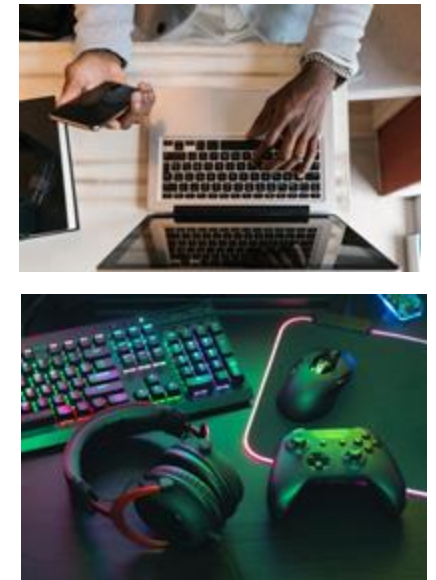
Inference

- Inference clusters <-> End-users
 - Need for lower latency will limit distance between these two.
 - Role for cables: possible that some end-users will be in a different country from an inference cluster.

Inference clusters



End-user devices



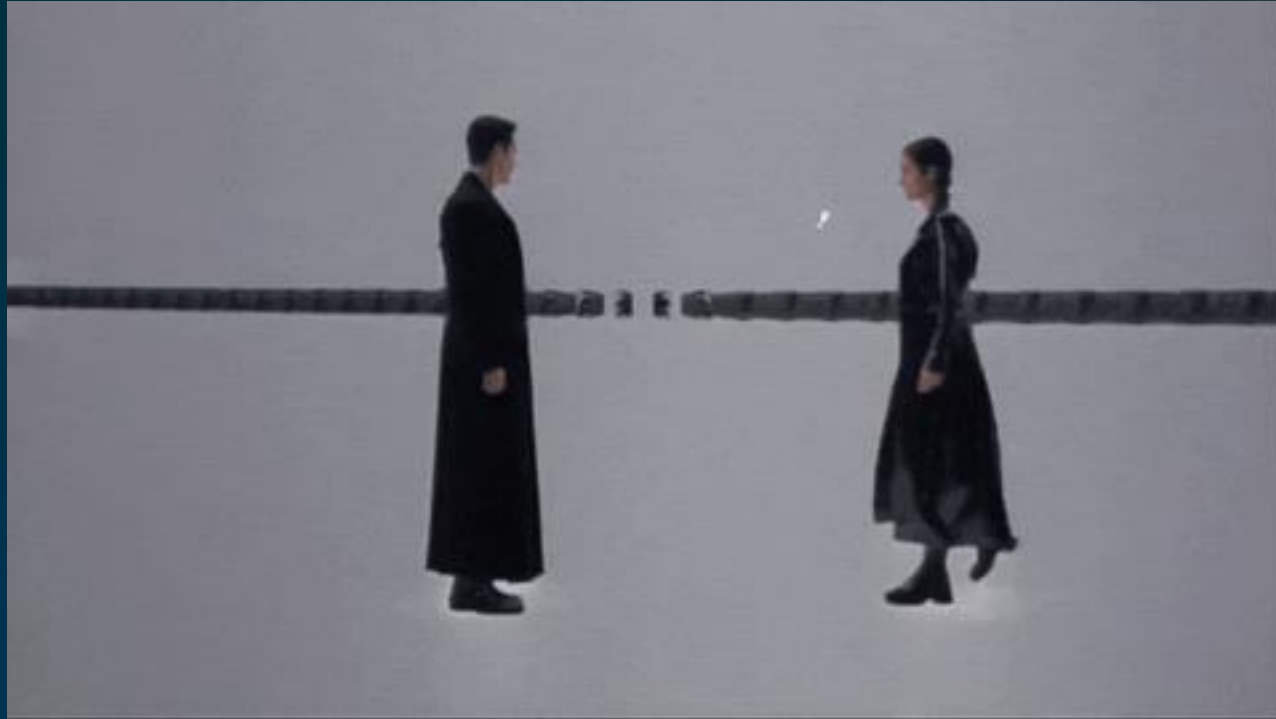
Factors boosting long-haul demand from AI

- Distributed training and federated learning
 - Use of multiple locations to train models which could be in different countries
- Spatial temporal load shifting
 - Moving workloads among data centers to optimize for carbon neutral and/or more affordable power
- Power availability and prices
 - Model training may be done in remote locations outside of traditional cloud hubs

Factors limiting long-haul demand from AI

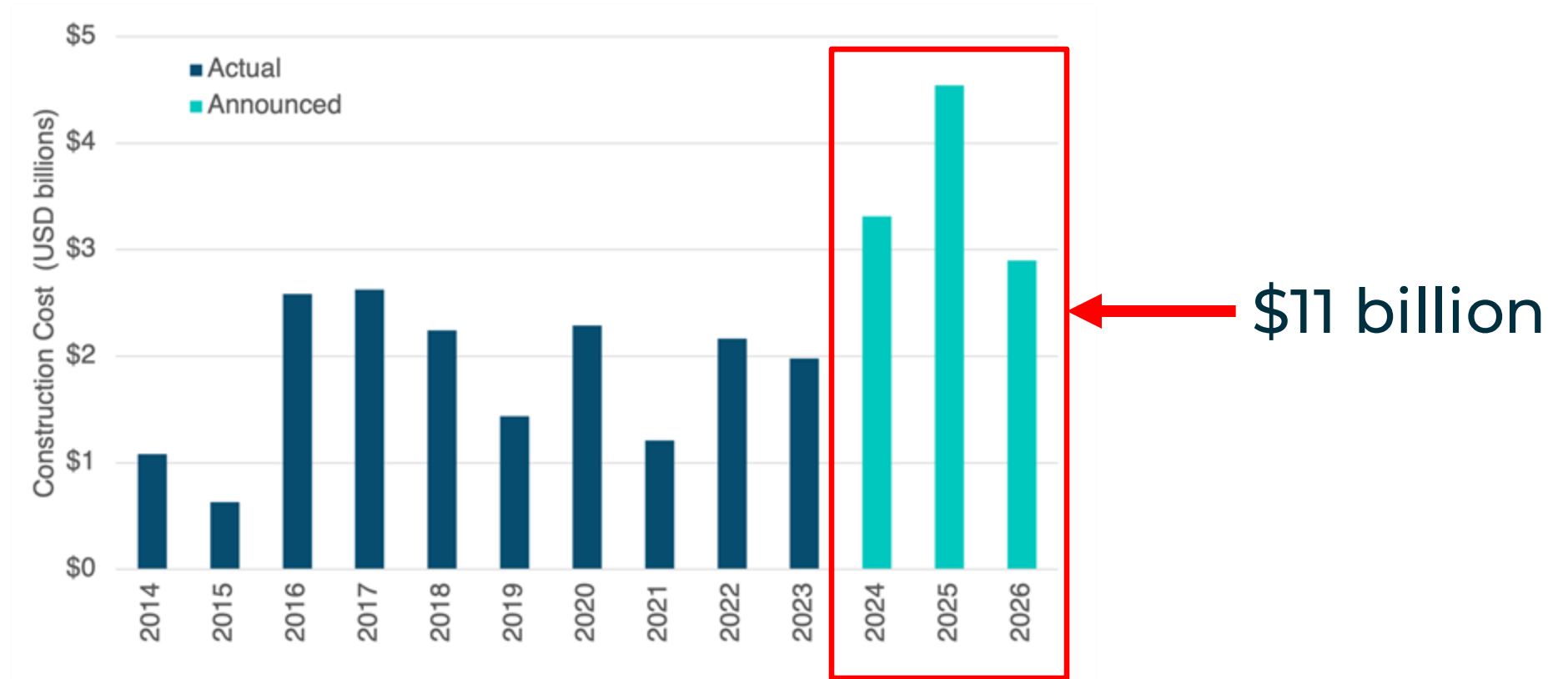
- Data sovereignty laws
 - Over 100 countries have laws governing where data can be stored and processed
- Inference latency requirements
 - Limiting distance between users and inference clusters is key for performance
- On-device inference
 - Can reduce data sent to inference clusters in a data center
 - Limited computational resource and power on devices are challenges

AI-Fueled Submarine Cable Investment Bubble?



Is there overinvestment in cables?

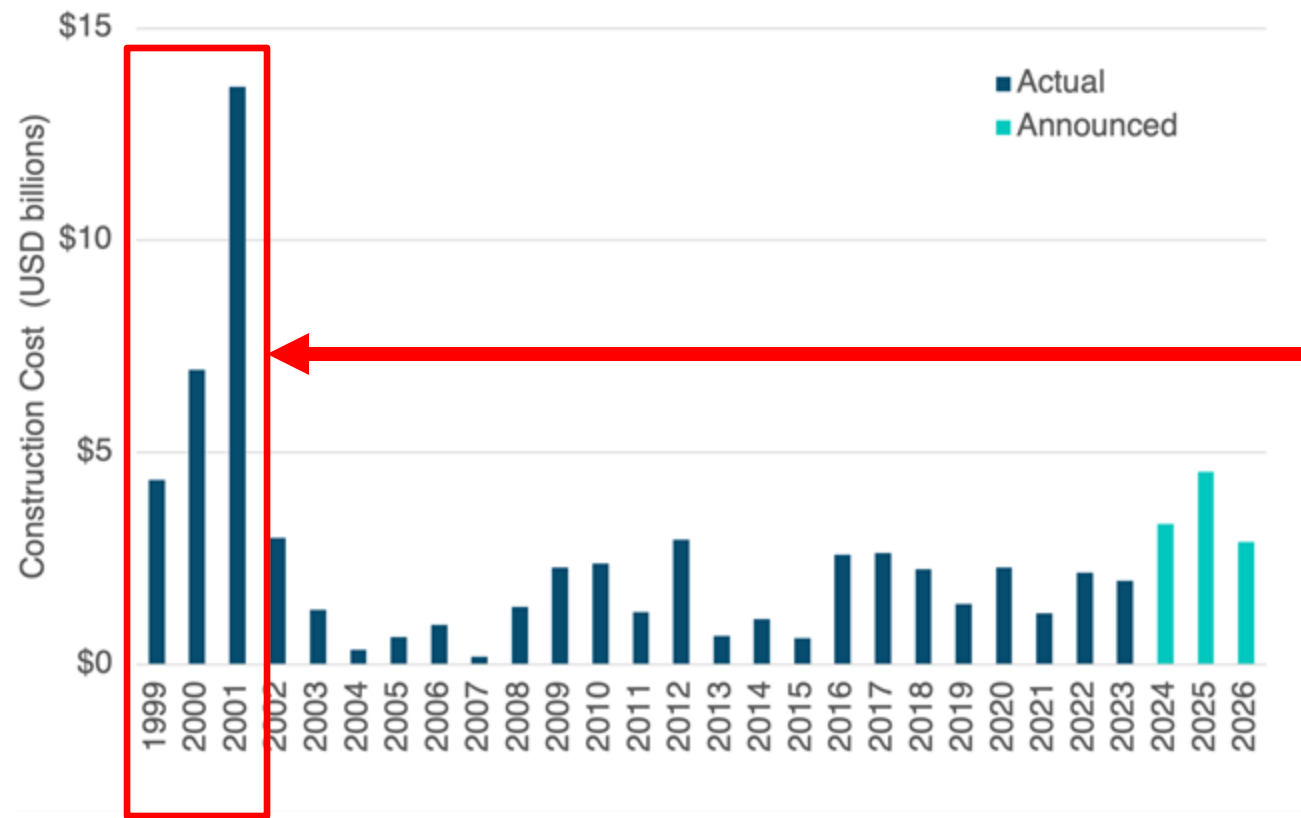
Construction Cost of New Submarine Cables by RFS



Source: TeleGeography's Transport Networks

Not partying like it's 1999

Construction Cost of New Submarine Cables by RFS

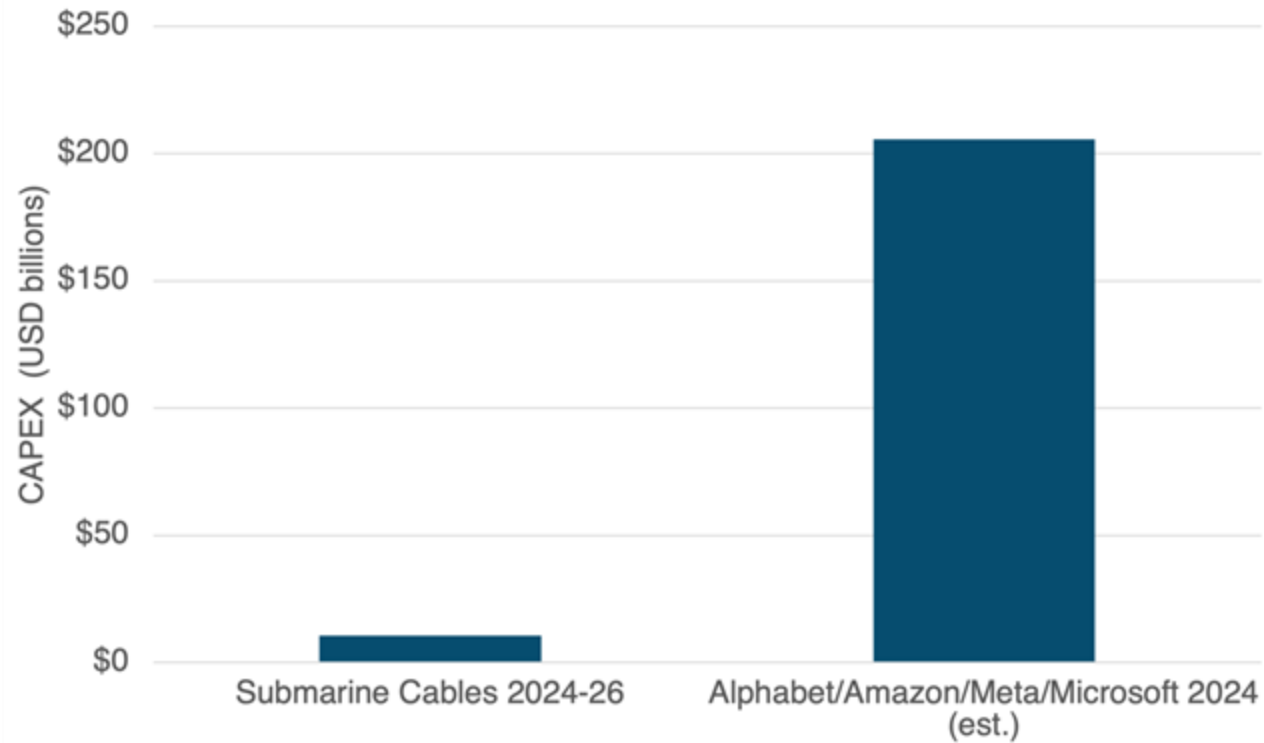


\$25 billion
(or \$43
billion in
current
dollars)

Source: TeleGeography's Transport Networks

Cables are a drop in the bucket

New Submarine Cables 2024-26 vs
4 Hyperscalers' CAPEX 2024

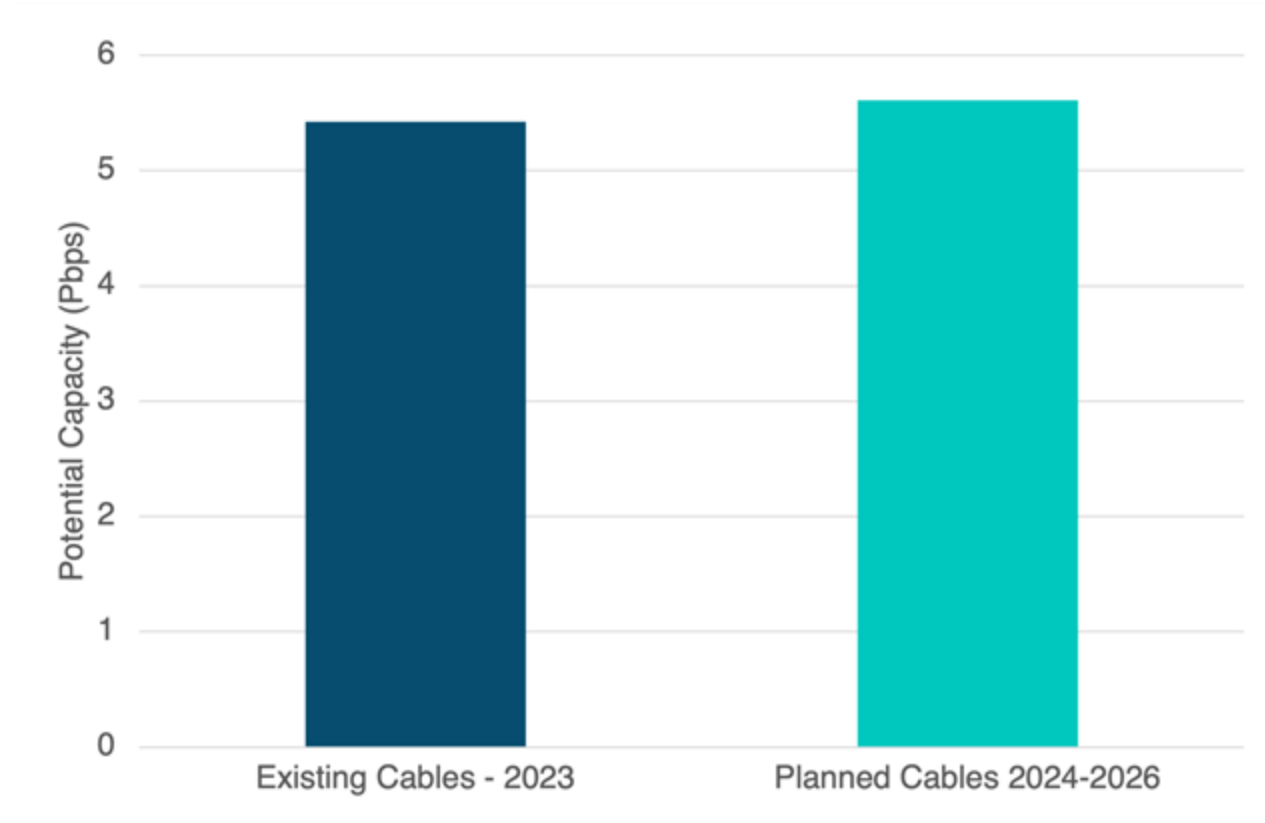


- “Even in scenarios where if it turns out we are overinvesting, these are infrastructure which are widely useful for us, they have long useful lives, and we can apply it across products. For us the risk of underinvestment far outweighs the risk of overinvestment.”

- Sundar Pichai, CEO Alphabet

We seem to be getting our money's worth

Existing and Planned Potential Cable Capacity on Major Routes



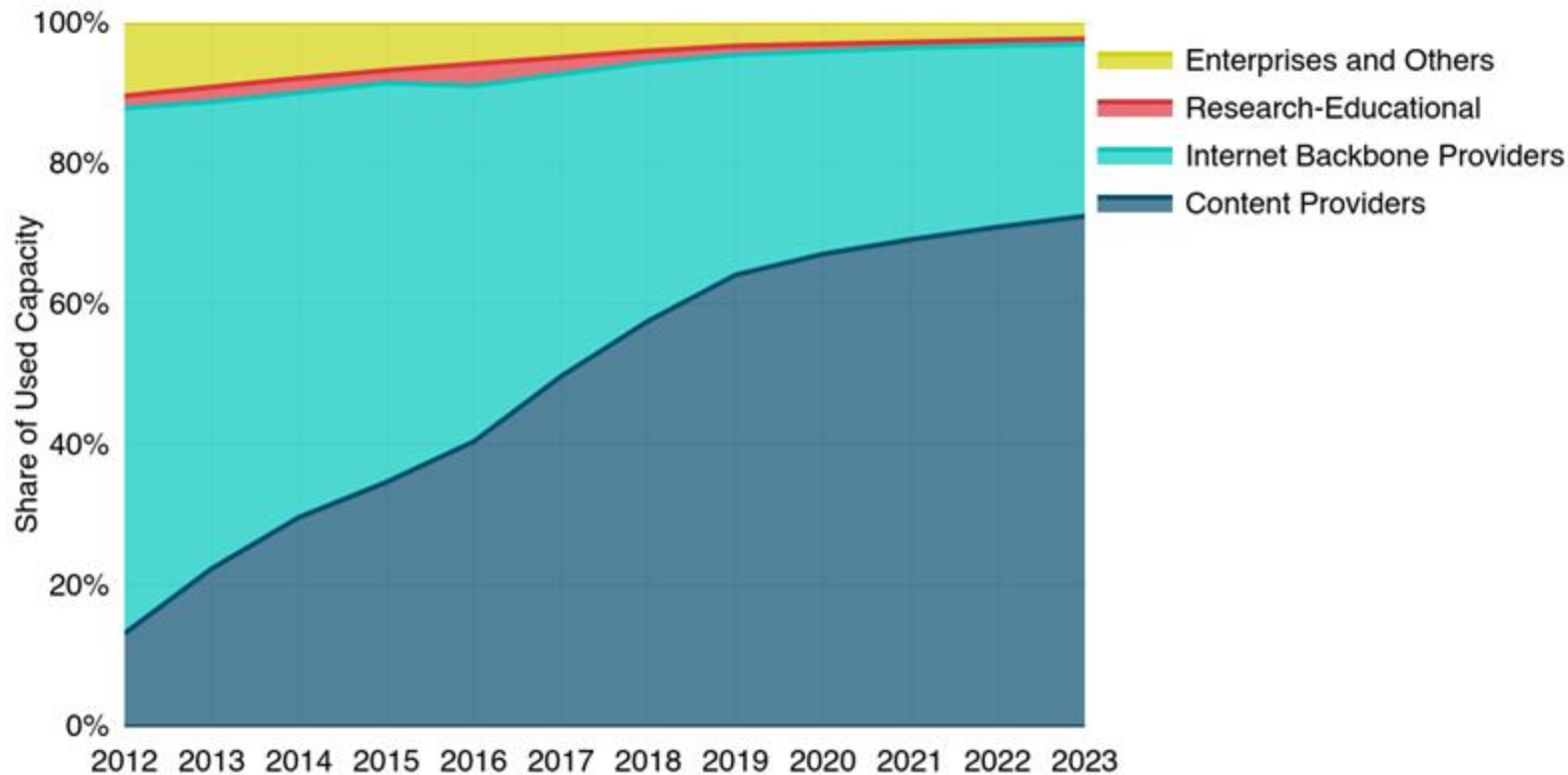
Source: TeleGeography's Transport Networks

Things to keep an eye on



New int'l bandwidth consumers

Used International Bandwidth by Network Type

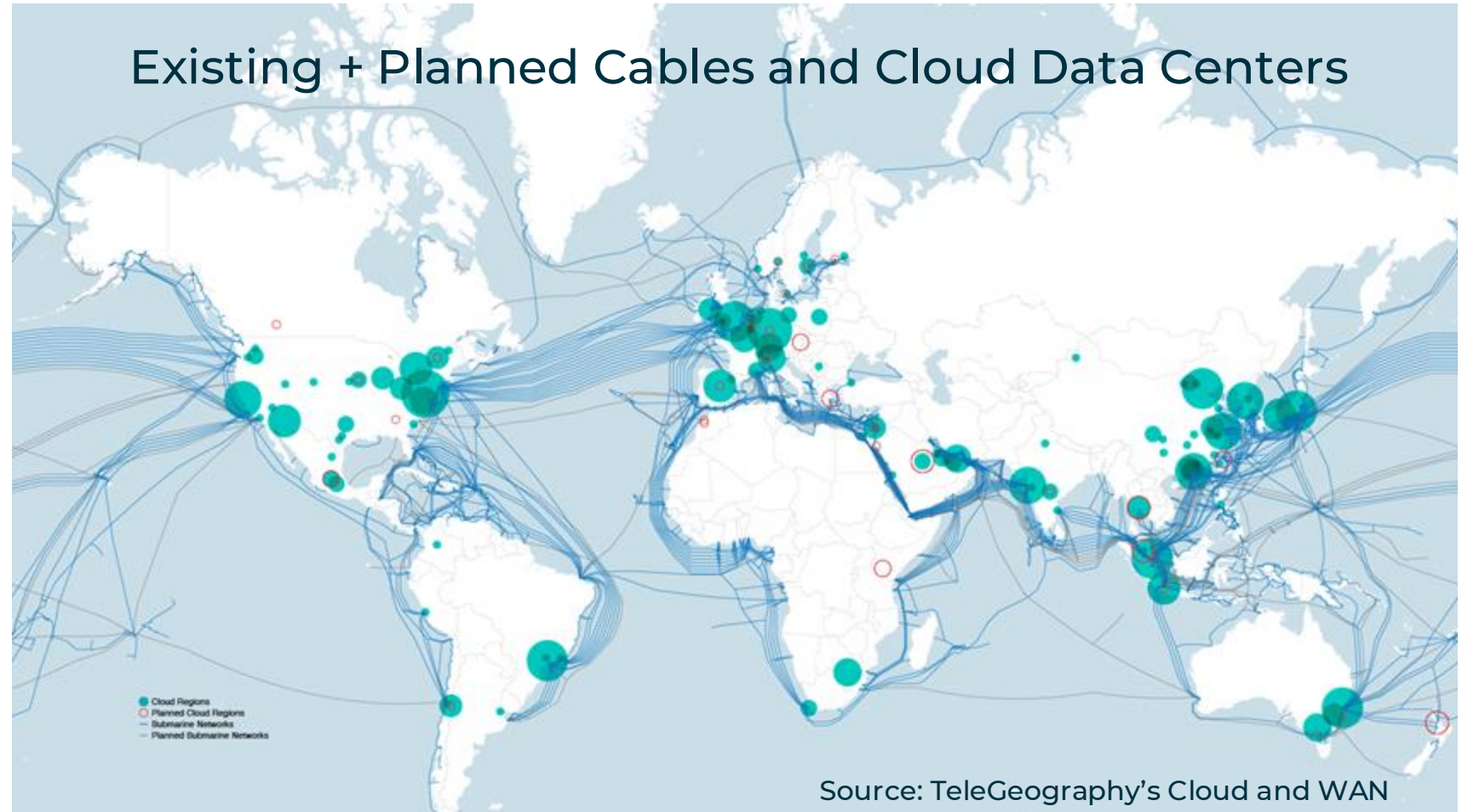


Source: TeleGeography's Transport Networks

- Hyperscalers' networks support their own AI models *and* third-party models too
- Some carriers starting to offer GPUaaS
- New AI-focused companies may eventually need subsea bandwidth or fiber (e.g. CoreWeave)

Connecting new locations

- Where will AI factories (AI-focused data centers) be built?
- New cables may be needed to locations currently not connected or underserved



Scaling chips, data centers, and power

- How rapidly can AI infrastructure expand to meet demand?

Data Centers Face Seven-Year Wait for Power Hookups in Virginia

Longer wait times are set to squeeze the development of the large-scale data center projects.



Bloomberg News
August 30, 2024

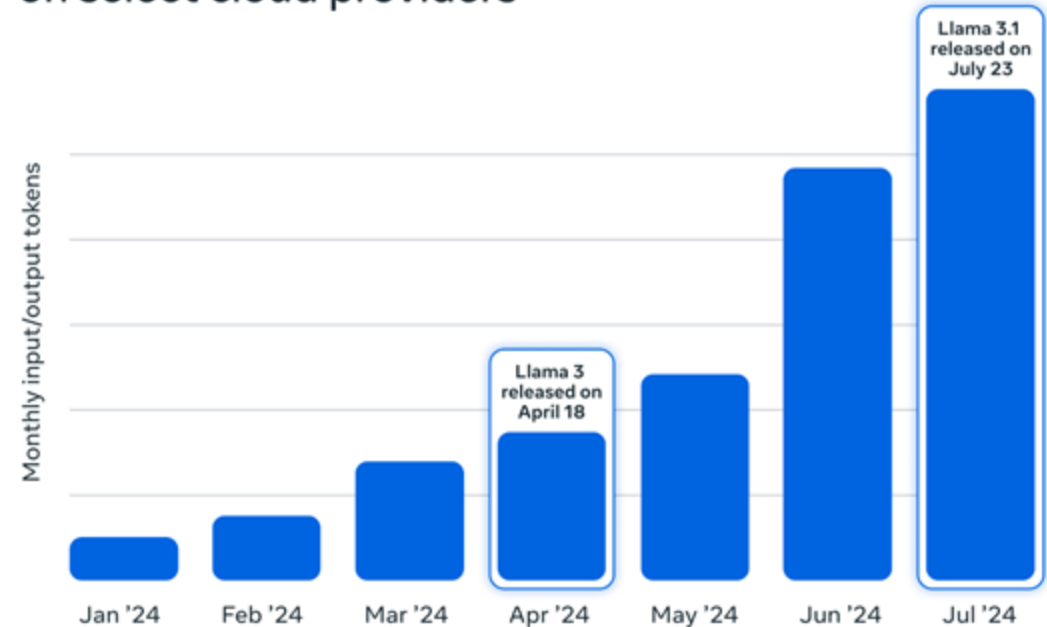
🕒 1 Min Read

Editor's Choice

Pace of AI adoption

- Thousands of AI models exist but how fast will usage rise and in which locations?
- Not many metrics exist for this
 - Meta reported usage of their Llama model increased 2x from May 2024-July 2024

Hosted API usage of Llama on select cloud providers



Source: Meta, <https://ai.meta.com/blog/llama-usage-doubled-may-through-july-2024/>

Use cases still emerging

How it begins —
Humans as AI model trainers

NVIDIA Accelerates Humanoid Robotics Development

Developers Gain Access to New NVIDIA NIM Microservices for Robotics Simulation in Isaac Lab and Isaac Sim, OSMO Robot Cloud Compute Orchestration Service, Teleoperated Data Capture Workflow and More

July 29, 2024



How it ends —
Humans as AI batteries



THANK YOU

Alan Mauldin
Research Director
amauldin@telegeography.com