Topic 4: AI-Infrastructure

Team: Adam, Dev, Divyanshi, Ferdian, Malik, Melissa, Randall & Wentao

CORNELL TECH

JACOBS INSTITUTE

Training vs. Inferencing

Resource Intensity: Training large AI models requires massive computational power and memory, often outstripping the resources needed for inferencing.

Scalability Requirements: Organizations must scale up infrastructure for training phases, but can often scale down for production inferencing workloads.

Cost Considerations: Training costs can be significant due to extended GPU usage or specialized hardware, while inferencing is typically more cost-efficient per request.

Optimization Trade-Offs: Training focuses on maximizing accuracy, while inferencing prioritizes efficiency. Techniques like pruning and quantization simplify trained models, reducing size and speeding up inferencing without significant accuracy loss, as seen in LLaMA 3.1.

Hardware Specialization: Specialized hardware like TPUs and ASICs is tailored for AI tasks. TPUs excel in high-throughput training, while GPUs like Nvidia H100 are designed for low-latency inferencing.

Cloud Business Model - Basic Services

Foundational infrastructure to develop and run applications on the cloud.

Business Model	Description	Use Cases	Examples
Infrastructure as a Service (IaaS)	Provides raw computing resources like VMs, storage, database, and networking.	Hosting applications, data processing, backups.	- Google Compute Engine - AWS EC2 - Microsoft Azure VMs - Alibaba & Tencent Cloud
Platform as a Service (PaaS)	Provides a platform for developers to build, test, and deploy applications.	Web app development, microservices, API management.	- Google App Engine - Azure App Service - Alibaba & Tencent Cloud
Software as a Service (SaaS)	Delivers fully functional software applications over the internet.	Email, CRM, collaboration tools, accounting software.	- Google Workspace - Microsoft 365 - Salesforce - Tencent's WeChat Work
Multi Cloud Solutions	Tools to manage and integrate resources across multiple cloud providers.	Hybrid cloud management, multi cloud orchestration, migration.	- Google Anthos - Azure Arc - AWS Outposts - Alibaba & Tencent Cloud

Cloud Business Model - AI Specialized Services

Provide the capability to build, train, and deploy AI models on cloud.

Business Model	Description	Use Cases	Examples
AI Computing Power	High-performance GPUs or TPUs for AI training/inference.	Deep learning model training, large-scale inference.	- Google Cloud A2 / TPUs - AWS EC2 P/T Series - Azure NC/ND Series
Machine Learning Platforms	End-to-end AI/ML tools for building, training, deploying, and monitoring models.	Automated ML workflows, MLOps, deployment.	- Google Vertex AI - AWS SageMaker - Azure ML Studio
Pre-trained AI Models	Ready-to-use APIs for vision, language, and speech AI tasks.	Image recognition, NLP, translation, text-to-speech.	- Google Vision AI - AWS Rekognition, Polly - Azure Cognitive Services
Distributed Training	Services for training large AI models across multiple GPUs or nodes.	Training large language models or multi-node systems.	- Google Cloud TPU Pods - AWS Elastic Fabric Adapter - Azure Distributed GPU Clusters
Strategic Partnership	Collaborations with leading chip providers or AI tools to get access to cutting edge AI supercomputing power	High-performance AI infrastructure to train large AI models and to develop state-of-the-art solutions	- Google Cloud: Enabling enterprises to leverage NVIDIA's DGX Cloud Infra via GCP's infrastructure

CORNELL TECH

JACOBS INSTITUTE

Cloud Computing

High-Performance Infrastructure: Provides on-demand CPU/GPU/TPU clusters optimized for deep learning, big data analytics, and other AI-heavy tasks.

Global Footprint: Data centers across multiple geographic regions enable low-latency access and compliance with local data regulations.

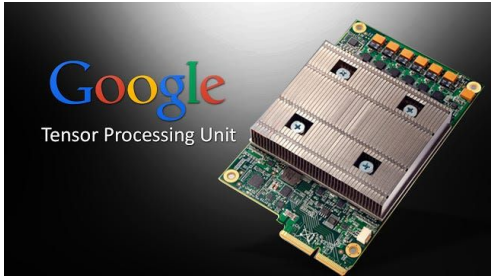
Managed Services: Offerings like container orchestration (Kubernetes) and serverless computing simplify deployment of AI workloads at scale.

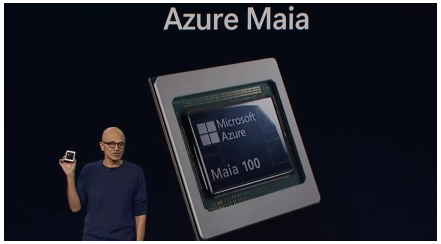
Security & Compliance: Robust encryption, compliance certifications, and identity management services are essential for sensitive AI data processing.

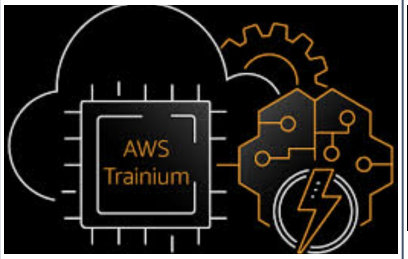
Cost Management Tools: Cloud platforms offer monitoring, alerts, and budgeting features to help optimize spending on AI workloads.

CORNELL TECH
JACOBS INSTITUTE

Google Cloud	Microsoft Azure	aws	ORACLE Cloud Infrastructure
Gemini	Google DeepMind	OpenAI	inworld		perplexity	NVIDIA.	xAI GROK
character.ai	AI21labs	MISTRAL AI_	Inflection	ANTHROP\C	Hugging Face	Meta	OpenAI
cohere	Hugging Face	perplexity	Palantir	alexa	crypto.com	Uber	Palantir
UNREAL ENGINE	MISTRAL AI_	ANTHROP\C	BMW	Adobe	abridge	NETFLIX	cohere	Microsoft Copilot







The AI Energy Challenge

Surge in Power Demand

- U.S. data center electricity consumption is projected to increase from 3.7% of total demand in 2023 to 11.7% by 2030
- AI workloads are expected to represent about 19% of data center power demand by 2028
- The energy intensity of AI is stark: a single ChatGPT query consumes 2.9 watt-hours, compared to 0.3 watt-hours for a Google search

Environmental Concerns

- Global data center energy consumption reached 460 TWh in 2022 and could hit 1,000 TWh by 2026
- In the U.S., data center load may grow from 19 GW in 2023 to 35 GW by 2030
- This AI-driven growth could potentially double the tech sector's current carbon footprint

Innovations and Regulations: Balancing Growth with Sustainability

Efficiency and Innovation

- Data centers are adopting advanced cooling technologies:
  - Liquid cooling systems can significantly reduce cooling energy consumption
  - AI-driven optimization of data center operations is increasing overall efficiency
- Hardware innovations are crucial:
  - New GPU, CPU, and ASIC designs are improving energy efficiency
  - Average power densities in data centers doubled from 8 kW to 17 kW per rack in just two years, allowing more computing power in less space

Regulatory and Corporate Sustainability Landscape

- The U.S. Federal Data Center Energy Practitioner program mandates efficiency evaluations every four years
- The EU's revised Energy Efficiency Directive introduces mandatory reporting on PUE and other efficiency metrics from 2024

Corporate Commitments:

- Google: Net-zero emissions by 2030, targeting 24/7 carbon-free energy
- Microsoft: Committed to being carbon negative by 2050

AI Infrastructure

Data Warehouses & Repositories

Data Warehouse vs. Repositories

A **Data warehouse** is a centralized system aggregates data from multiple sources into a consistent repository for data mining, AI, and analysis.
A **data repository** is a storage space for specific data management tasks like archiving.

Integration with Machine Learning

A data warehouse stores vast volumes of data from various sources, utilized for training AI and ML models. Paramount for enhancing ML models through extensive data application. It enables complex analyses and queries, crucial for developing AI-driven insights within an organization.

Data Running Out & Role in AI

Epoch AI predicts that by 2032, tech companies will deplete public text data for AI training, potentially stalling advancements and forcing reliance on private or synthetic data.

Critical Challenges in AI Network Infrastructure

Modern AI systems generate and process unprecedented volumes of data across distributed networks. Traditional cloud-centric architectures face fundamental challenges when handling real-time AI workloads, necessitating new approaches to network infrastructure.

Latency Issues
Network latency between data generation and processing impact system performance and reliability

Bandwidth Constraints
The sheer volume of data generated by AI systems strains network capacity

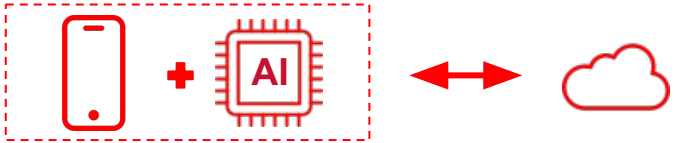
Privacy Concerns
Transmitting sensitive data across networks introduces security challenges

These fundamental challenges require rethinking how we design AI infrastructure, shifting from centralized to more distributed architectures that balance performance, cost, and security.

AI Infrastructure - Network

Federated Learning: Optimizing Network Architecture

Transforming AI Processing with Federated Computation & Learning Architecture



Real-time Processing
Reduces response time for time-sensitive decisions

Smart Data Management
Processes majority of data locally, sending only model parameters to cloud

Enhanced Privacy
Sensitive data stays local, with only model parameters transmitted

Federated Learning enables real-time event detection and response, while maintaining secure network connections for system updates and continuous learning.

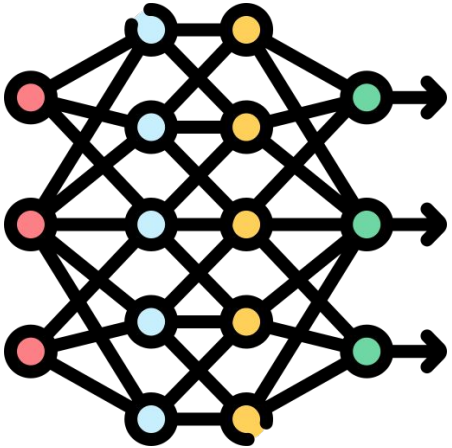
What are Deep Learning Frameworks?

High-level APIs:

1. Automatic differentiation
2. Visualization tools
3. ...

Basis of Deep Learning:

1. Develop models
2. Training
3. Deployment

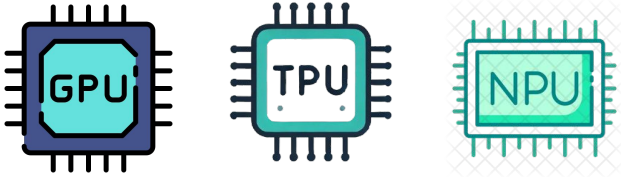


Frameworks and AI Infrastructure

Accelerators Support

1. GPU, TPU, NPU …
2. Operator-level optimizations
3. Example:

   DIOPI(Device-Independent Operator Interface) support Ascend, Cambricon, Cuda etc.



Distributed Training

1. Data Parallelism (Zero), Model Parallelism, Tensor Parallelism
2. Example:

   1024 NPU to train LLama 2 70B

DeepSpeed

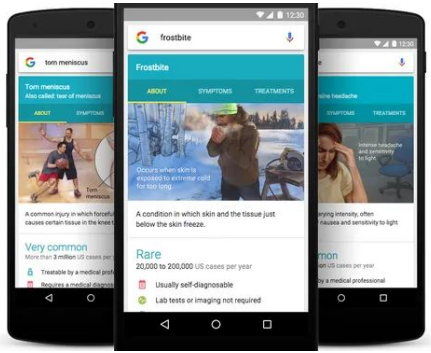
LLAMA 2

Prem Ramaswami

Head of Data Commons at Google

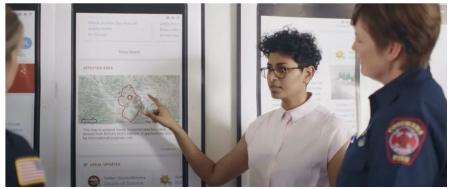
DC Google

Data Commons



Google Health Search

SIDEWALK LABS



Google Crisis Response

Carnegie Mellon University

Harvard Business School

CORNELL TECH
JACOBS INSTITUTE

Appendix

1) https://www.statista.com/statistics/1537014/data-center-power-demand-us/
2) https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand
3) https://prismecs.com/blog/innovative-technologies-to-optimize-data-center-energy-consumption
4) https://www.utilitydive.com/news/artificial-intelligence-doubles-data-center-demand-2030-EPRI/717467/
5) https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand
6) https://hwglaw.com/2024/09/09/data-centers-artificial-intelligence-spurs-need-for-more-energy-efficiency-2/
7) https://www.mhp.com/en/insights/blog/post/the-energy-efficiency-act-for-data-centers
8) https://cloud.google.com/compute/all-pricing
9) https://azure.microsoft.com/en-us/pricing/details/azure-arc/core-control-plane/
10) https://aws.amazon.com/sagemaker/