EQUINIX

Improving Inferential performance - latency. Data and storage protection at scale.

Datacenters older than 5 years – H100+pedagogy(transformers)

GB200 DGX – 30times faster than GH100 with inferential workloads – OVX versus DGX

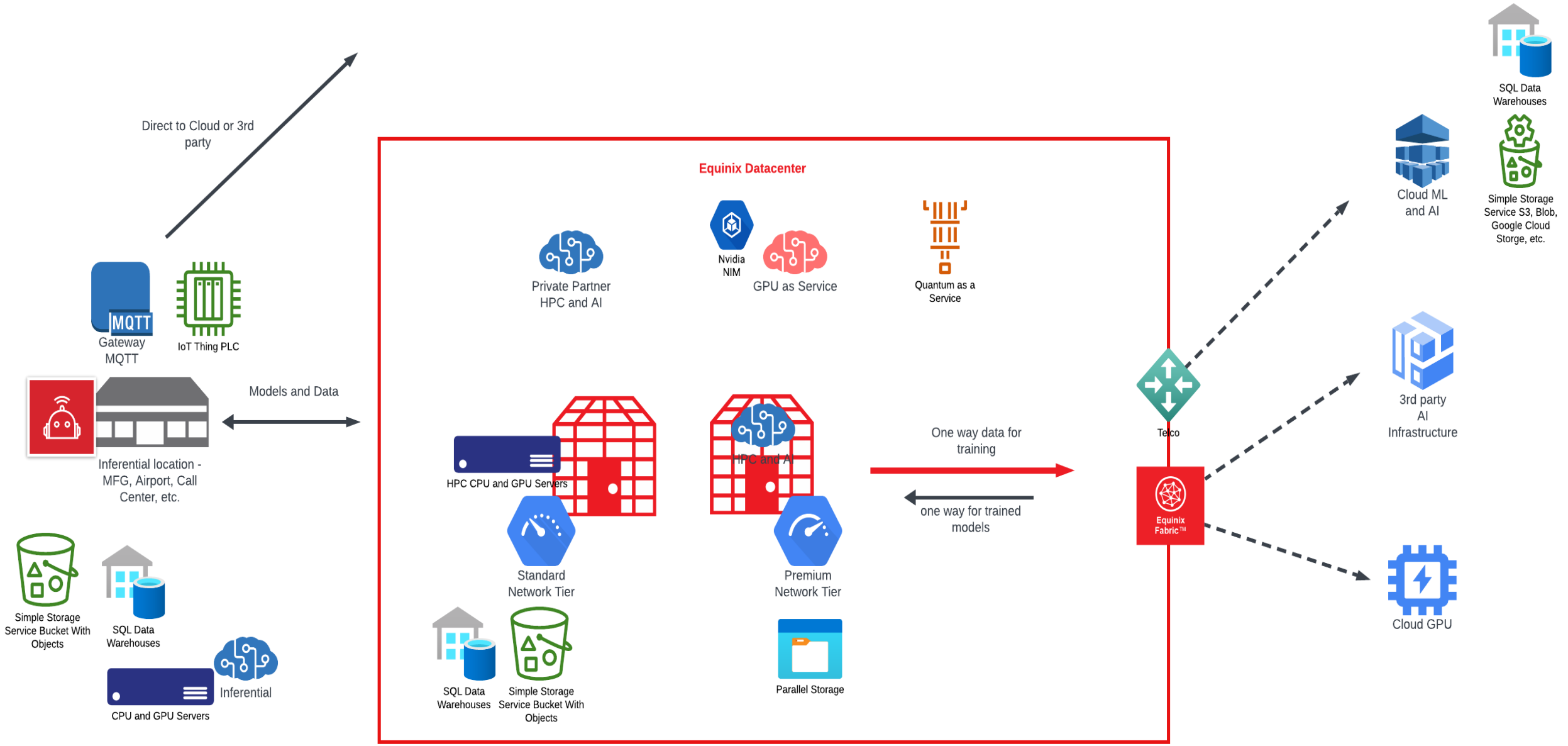
- Power Density per rack and data closets at the edge
  - From 20KW per rack to 120KW per rack – A and B power circuits, A/B/C circuits
- Cooling demands at the edge datacenters
  - Air cooling to Air and Water cooling – closed or open loop
- Resilience and Redundancy for operations
- Competing resources – Power for robots/ Lidar and 5G/6G workloads, Controller with new protocols – IoTs, PLC/MQTT, and SAP transactional servers
- Parallel Infrastructure needs – Weka, Infiniband/NVlink/RoCE (RDMA), GPUs.
  - Facial recognition and tracking
  - Changing manufacturing process for batches
  - Near live trading data – hunting for malfeasance and out of compliance

Application tolerance – not Model latency

Rough order of magnitude in a trading matching engine, manufacturing plant, airport

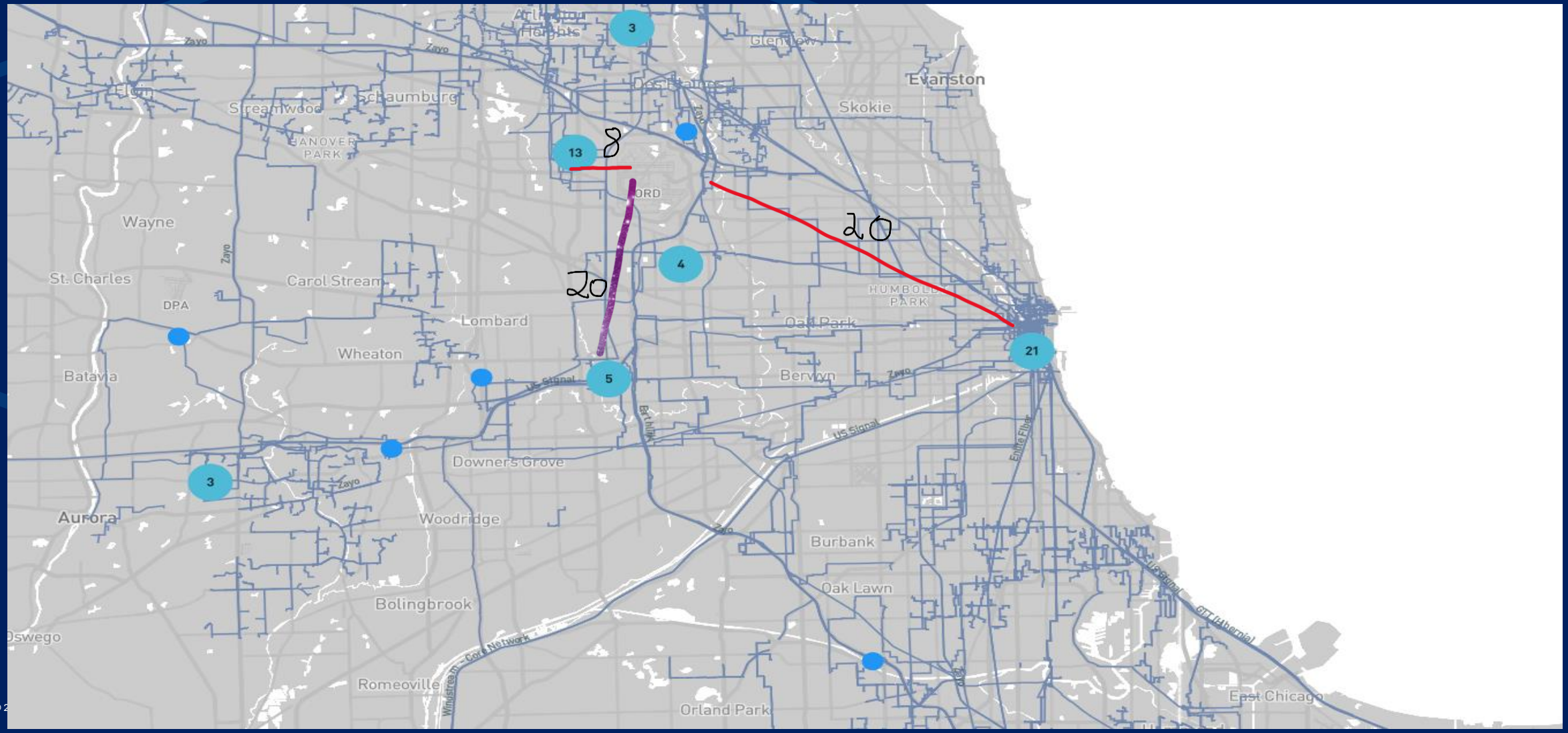
- 50 to 100 nano seconds can beat the others – ingestions at nano scale for trades, quality, security, accuracy, and reliability (25ms from CH to LD - Raft)
- PLCs – task is uninterruptable by all other tasks, and has the smallest period possible (ideally, 1ms or less)
- Airports:
  - Stationary facial recognition application – 2 to 3 seconds
  - Threat and Asset movements - real time from cameras to AI to application – 16ms effects of steering and pointing non-linear tasks

Edge, Cloud, and datacenter



Closing the latency gap – refractive index 1 for air 1.5 for glass

Nano to Micro to Milli-seconds – Round trip 1milli versus 300 Micro – 1 mile = 16 micro round trip.



Data = model

Models won’t share nightmares it learned unless you ask it – never mind the back propagation training

- Trillion parameters + million tokens + quality of data + design of the model
- Protecting 50 to 100 petabytes of warm and hot data
- Data lake house –Apache or Delta Lake - Dremio and Presto

Measures

- Airgap and immutability with multiple regions copies
- Bi-directional scanning
- Restore and integrity testing within layered timeline
- Cold storage of layered data –prior to timelines

Reference

- https://cac-llc.com/CAC/wp-content/uploads/2013/06/Low-Latency-Parcing-of-Network-Data-Using-Control-Logix-PLCs.pdf

Reserved Cloud GPUs




	GB200	B200	H200	H100
System	NVL 72-GPU rack	HGX B200 8-GPU	HGX H200 8-GPU	HGX H100 8-GPU
GPU Node Spec	36x NVIDIA Grace 72C CPUs 30TB system memory 550TB local NVMe cache	2x Intel Xeon 56C CPUs 3TB system memory 60TB local NVMe cache	2x Intel Xeon 56C CPUs 2.3TB system memory 30TB local NVMe cache	2x Intel Xeon 56C CPUs 2TB system memory 30TB local NVMe cache
GPU Memory	192GB (13.8TB tot.) HBM3e @ 8TB/s	180GB (1.4TB tot.) HBM3e @ 7.5TB/s	141GB (1.1TB tot.) HBM3e @ 4.8TB/s	80GB (0.6TB tot.) HBM3 @ 3.35TB/s
Interconnect	72x NVLink 1.8TB/s PCIe Gen 6 256GB/s	8x NVLink 1.8TB/s PCIe Gen 5 128GB/s	8x NVLink 900GB/s PCIe Gen 5 128GB/s	8x NVLink 900GB/s PCIe Gen 5 128GB/s
Networking	400G InfiniBand NDR 200G Spectrum Ethernet	400G InfiniBand NDR 200G Spectrum Ethernet	400G InfiniBand NDR 100G Ethernet	400G InfiniBand NDR 100G Ethernet
FP8 Tensor Core	10 PFLOPS	9.5 PFLOPS	4 PFLOPS	4 PFLOPS
Availability	Starting Q1'25	Starting Q1'25	Starting Q4'24	Now
Application	Training >400B	Training <400B	Available before B200	Available Now

NVIDIA GB200 NVL72

Best platform for large model training and inference

- Standout performance for 1T+ parameter models
- GB200 training up to +4x H100 and +30% B200
- GB200 inference up to +30x H100 and +17x B200

Rack-scale 72x GPU design with NVSwitch Interconnect

- NVSwitch with 1.8TB/s GPU bandwidth (14x PCIe 5)
- 72-GPU NVLink domain for 13.8TB of parallelism
- First platform with InfiniBand XDR 800G (2H 2025)

Direct to chip (D2C) Liquid cooling for optimal performance

- +11% FLOPS increase vs B200 at -7% lower power
- +150% FLOPS increase vs H100 at +31% higher power