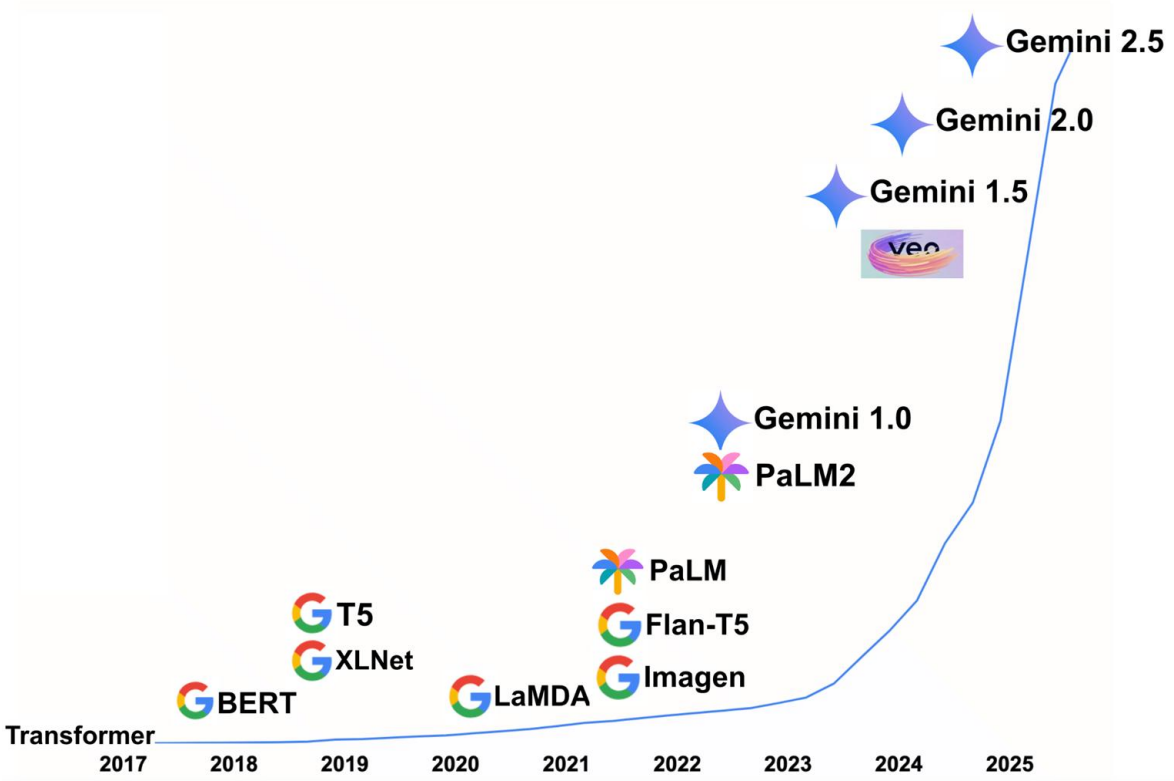
Towards Ultra Scale Energy Efficiency

Google

Compute Energy Nexus Workshop, Dec 2025

Exponential ML Compute Demand

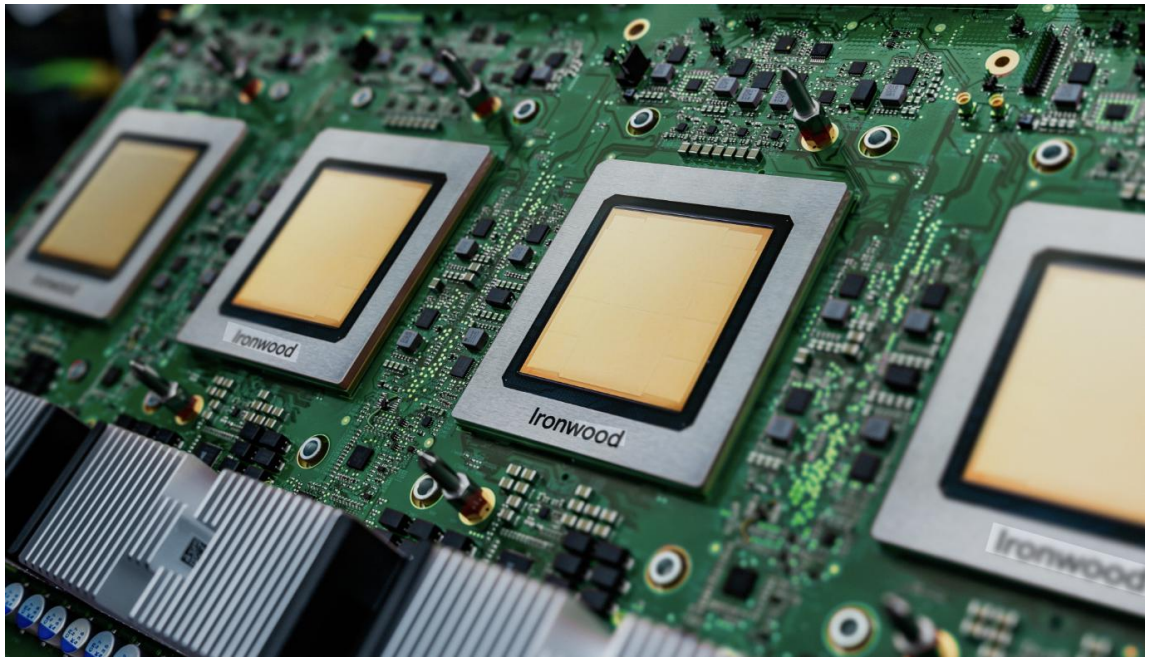


How do we get 1000x more efficient?

- Scale up -> Scale out → Sustainable specialization
- Age of Information → Age of Insight

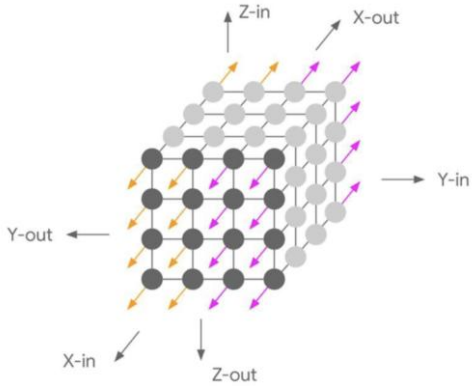
Hardware Specialization: TPU

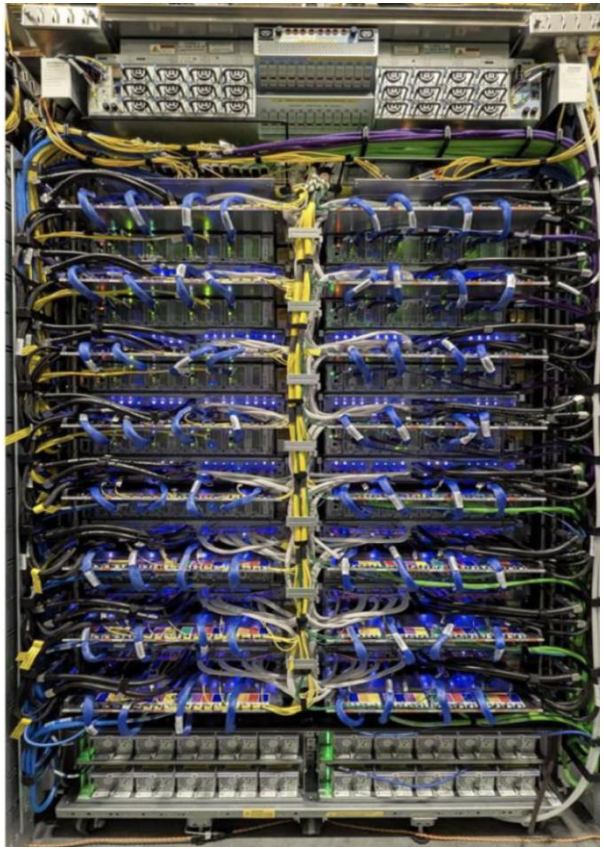
- Large Systolic arrays - dense compute
- Scatter/Gather engine - sparse operations
- Custom ICI - scale up
- Gen over gen micro-architectural improvements for energy efficiency



Rack Specialization

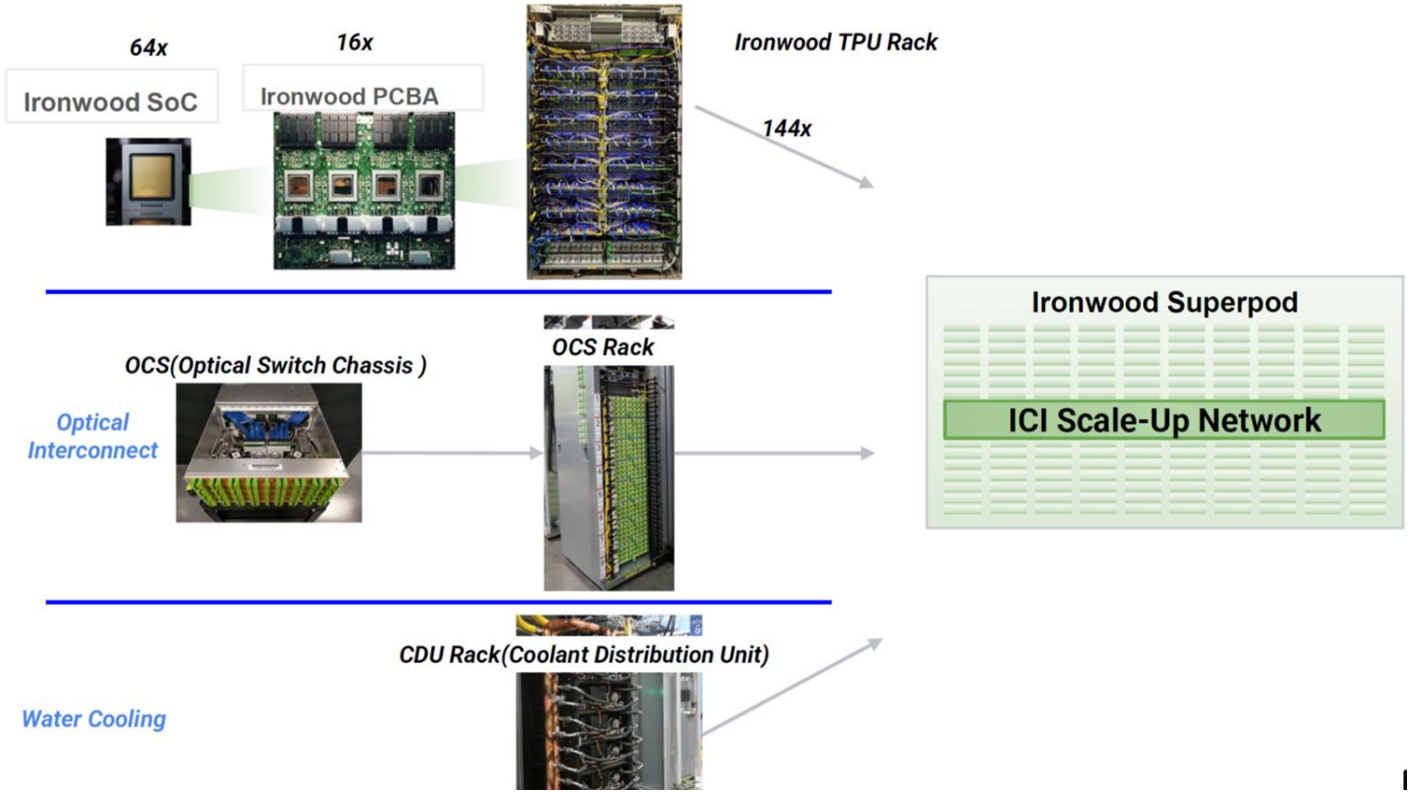
- Enables liquid cooling to support maximum density
  - Reduces latency, improve energy efficiency
- A 4X4X4 building block
  - 9216 c ıper pod





System Specialization

- Flexible job sizes
- Resilient routing
- Energy efficient scale up network



Specialized Data Representation

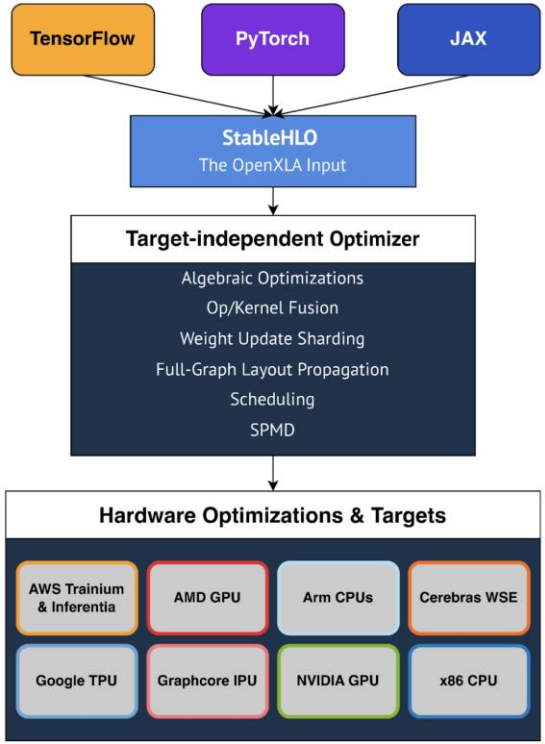
- TPUs have supported int4, int8, and bfloat16. Supporting fp8 from Ironwood.
- Reduced precision a key tool in the low-power tool kit, needs accompanying recipes to be quality neu

Integer	
Add	
8 bit	0.03pJ
32 bit	0.1pJ
Mult	
8 bit	0.2pJ
32 bit	3.1pJ

FP	
FAdd	
16 bit	0.4pJ
32 bit	0.9pJ
FMult	
16 bit	1.1pJ
32 bit	3.7pJ

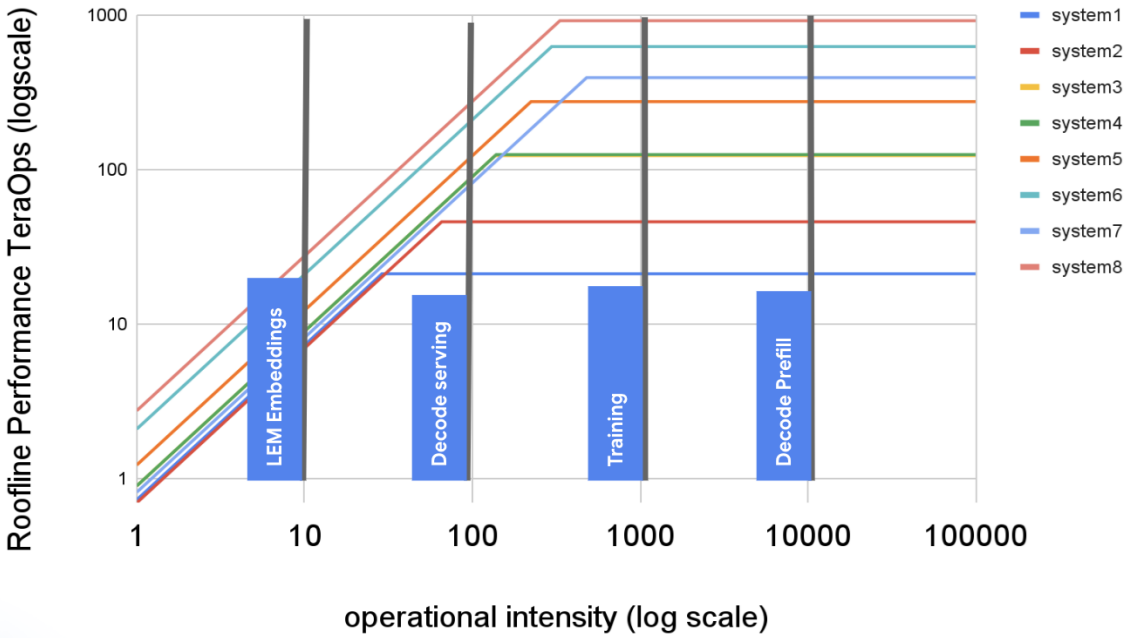
Source: Computing's energy problem (and what we can do about it) - Horowitz et al, 2014

Compiler Optimizations



- TPU optimizations:
  - Memory Layout Assignment
  - Fleet efficiency tune-up
  - Scaling - Pod collectives, ICI+DCN
  - Co-design with HW
- https://github.com/openxla

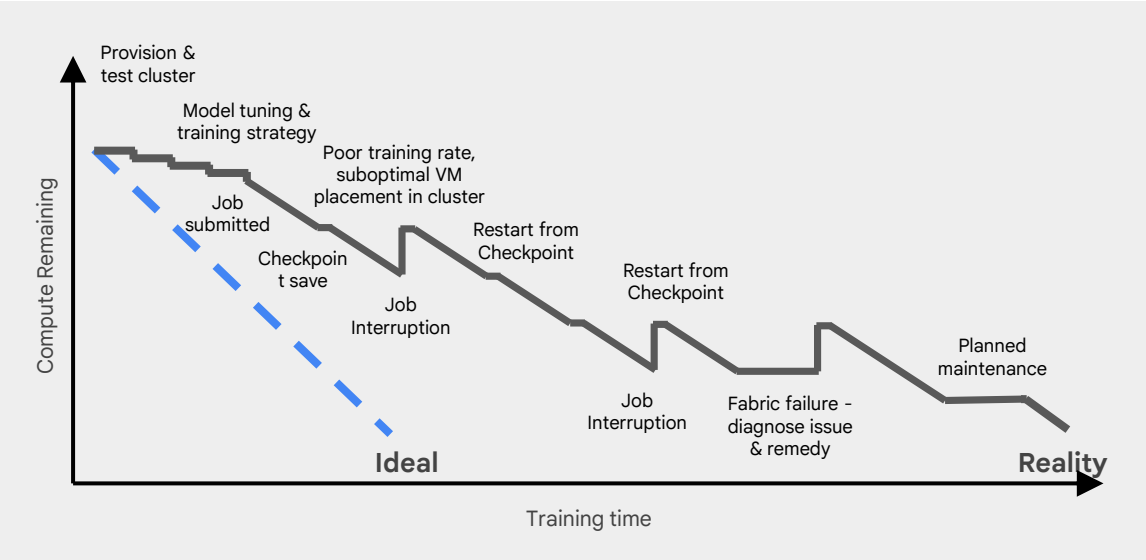
Workload Profiling



- ML workloads come in all sizes and shapes
  - Compute bound
  - Memory bound
  - Network bound
- Single server jobs to multi-clusters
- Disaggregated systems

Takeway: Opportunity to optimize perf/W by workload type

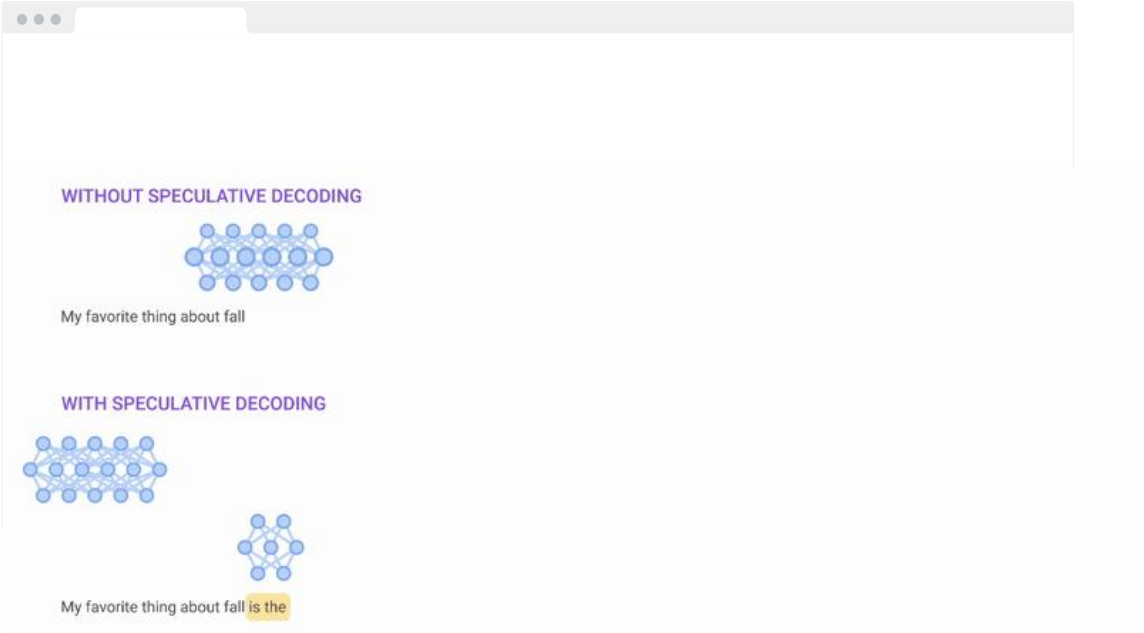
Distributed Training at Scale



Challenges

- Provisioning speed, VM proximity
- Model-aware cluster tuning
- Troubleshooting depth and speed
- Checkpointing limitations
- Degraded operations

AI inference



Challenges

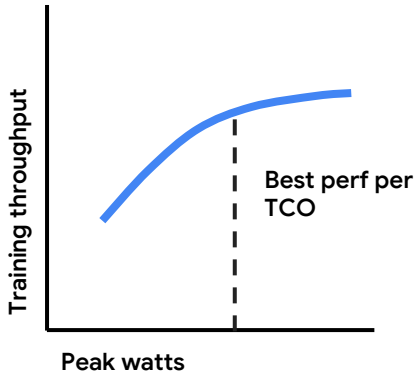
- Computational demands
- Cost vs. interactivity trade-offs
- Rapidly evolving state-of-the-art
- Operational complexity

Operating region and Provisioning Power

Get more out of each Watt

Optimize the power parameters for each job

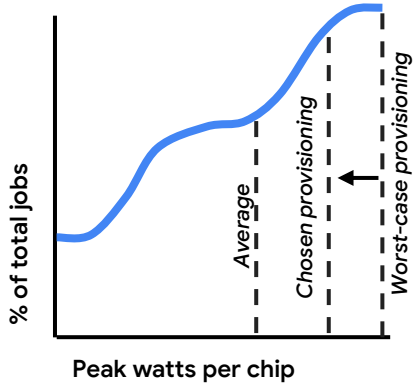
ML job



Provision the right amount of Watts

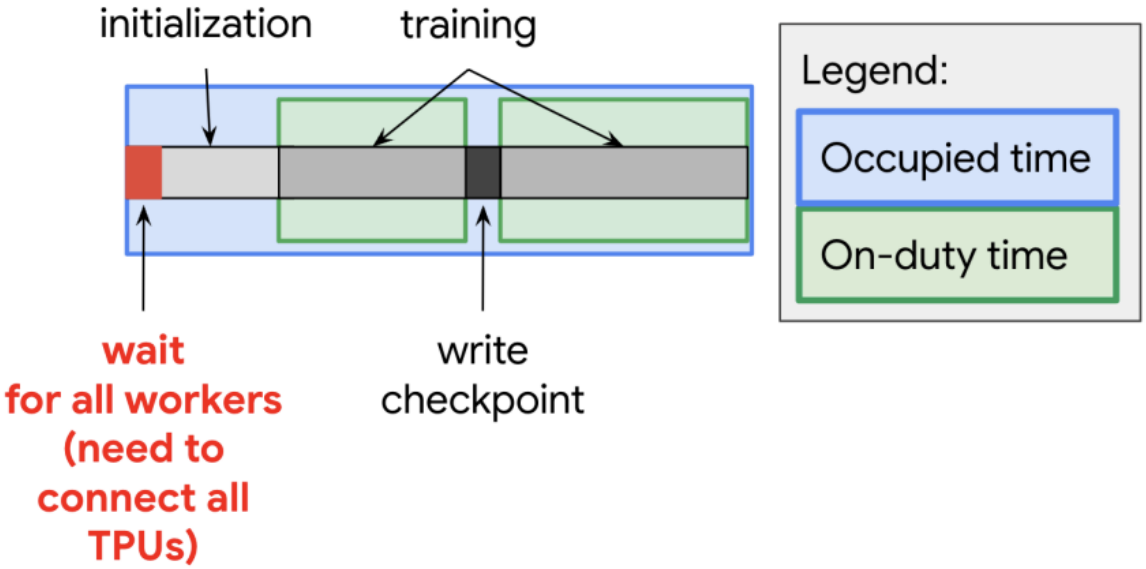
Load balance jobs to avoid worst-case, concentrated power peaks

Power demand CDF



Barroso and Holzle, “The Case for Energy Proportional Computing.” IEEE Computer, 2007.

Fleet Metrics: Occupancy & Duty Cycle



Fleet Metrics: Goodput

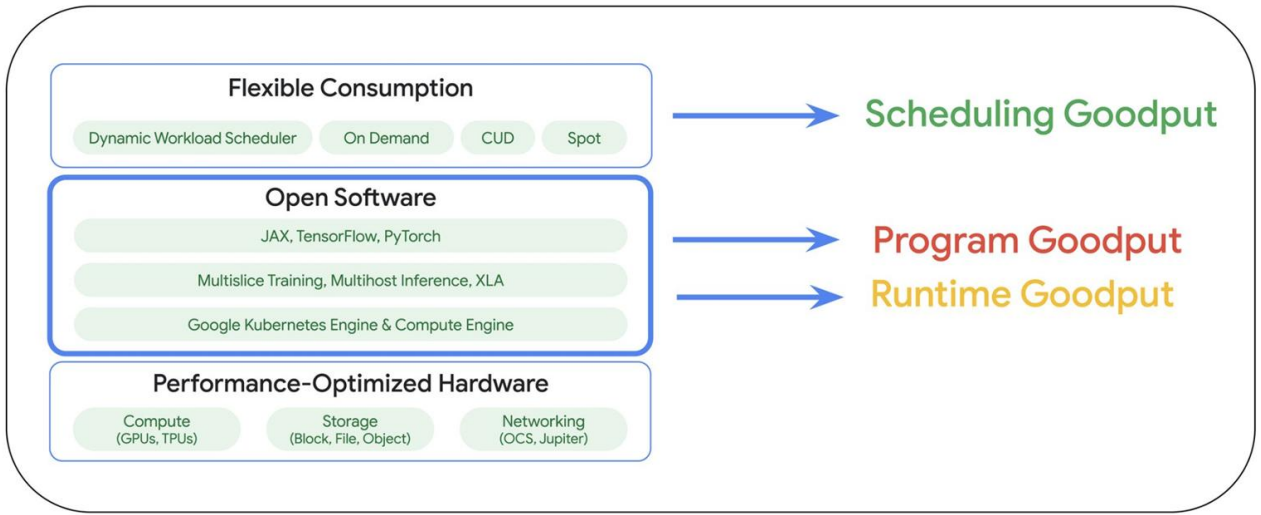
ML Productivity Goodput

Scheduling Goodput x **Runtime Goodput** x **Program Goodput**

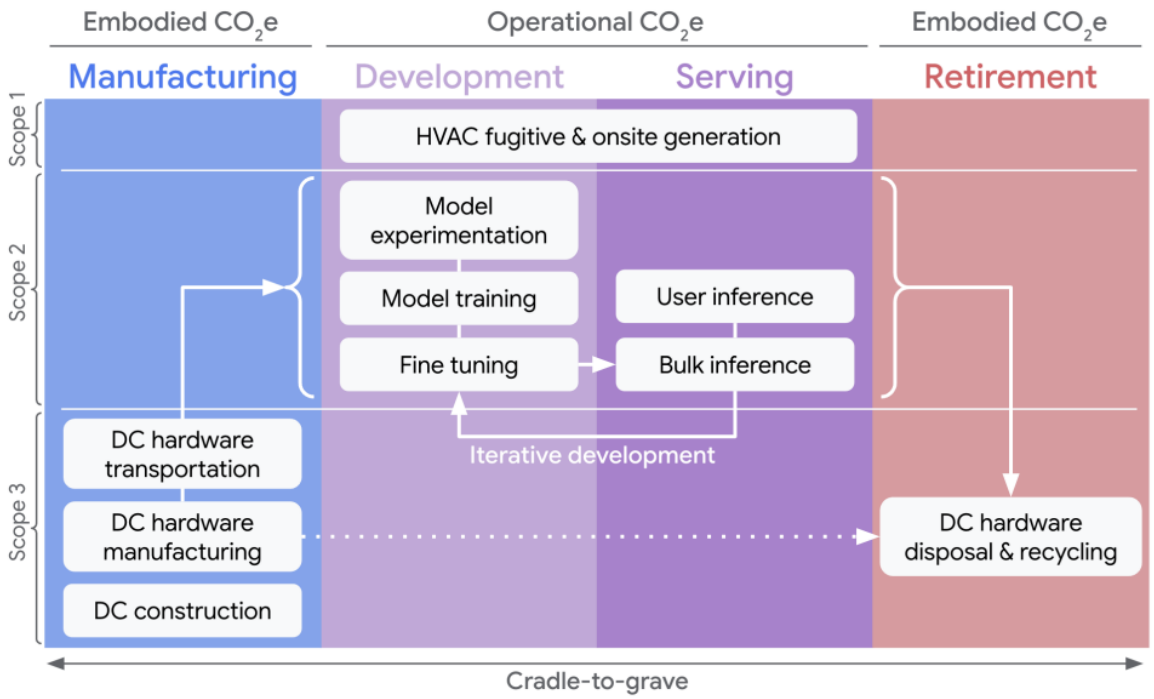
Fraction of time ALL required resource are available

Fraction of time all required resources are available and making forward progress

Fraction of peak available compute utilized by the program



Fleet Metrics: Life Cycle Analysis, perf/CO2e



Fleet Metrics: Power Usage Effectiveness

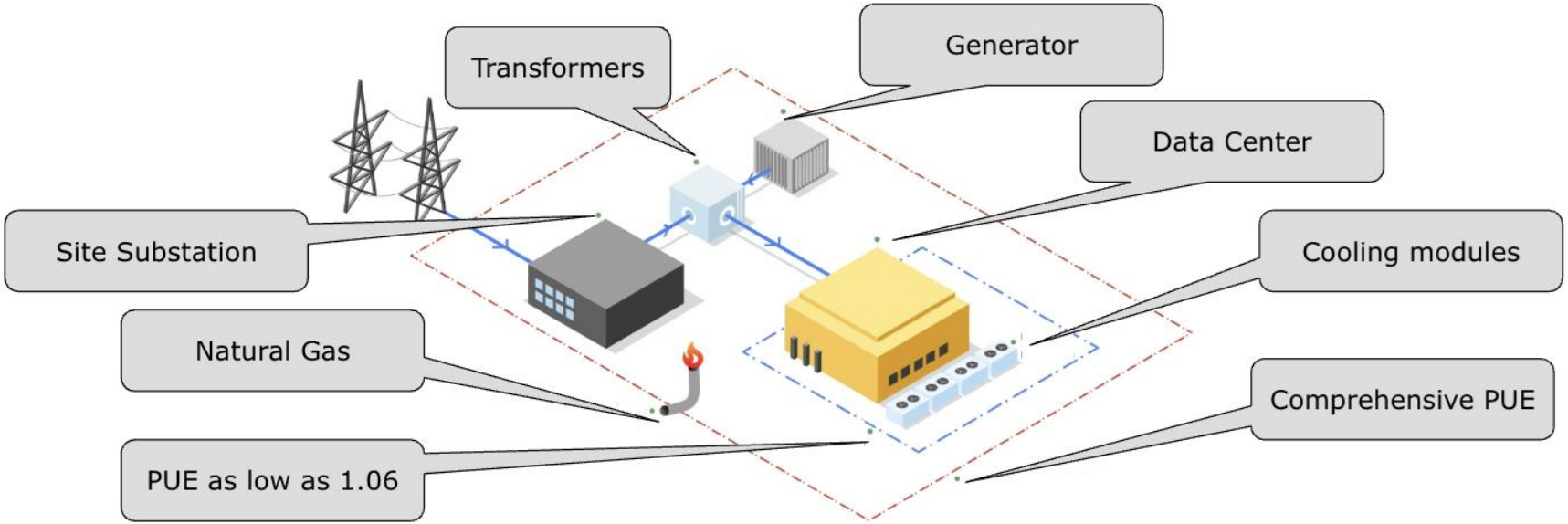


Figure 1: Google Data Center PUE measurement boundaries. The average PUE for all Google Data Centers is 1.10, although we could boast a PUE as low as 1.06 when using narrower boundaries.

PUE = Total Facility Energy / IT Equipment Energy

Conclusions