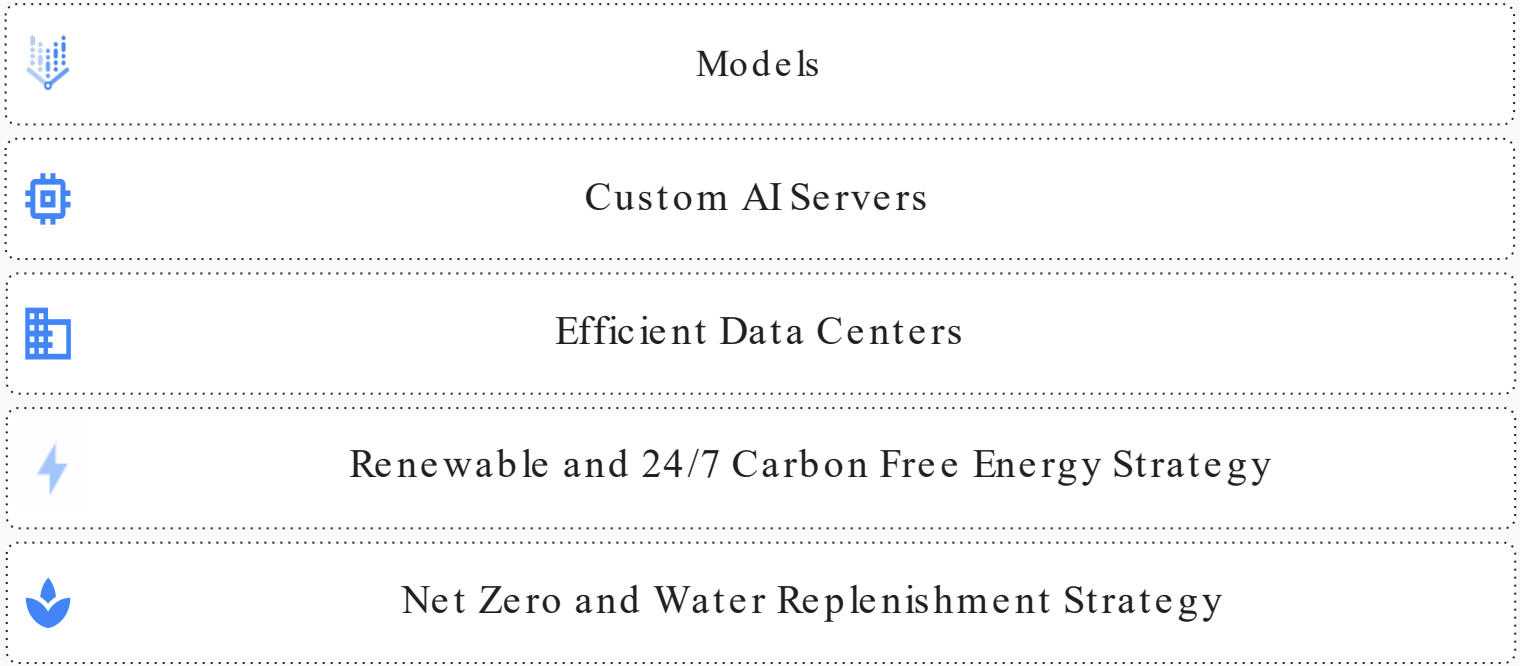
Google approach for more sustainable AI



Google stack for more sustainable AI

Net Zero & Water Replenishment

- How do you minimize emissions from hardware manufacturing, transportation, dc construction?
- Balance of water usage &energy
- Can we use AI to help? (DeepMind reduced our cooling bills by 40% back in 2016 → $ saving too [open sourced])

Renewable & 24/7 Carbon Free Energy

- How do you reduce Scope 2 emissions, how do you enable new clean energy sources to scale

- Target to be 24/7 CFE by 2030
- Also working with partners such as Fervo (Geothermal), Kairos (SMR) to scale new clean energy sources

Efficient Data Centers

- Utilization
- PUE
- Can you choose time/location your loads are placed - intelligent shifting
- Our current data centers deliver 6 x more computing power than 5 years ago

Custom AI Servers

- Industry is making rapid progress - our Ironwood TPU is nearly 30 times more efficient than first TPU from 2018
- NVIDIA estimates their Blackwell chip trains models using 75% less power than older GPUs for same task

Models

- Huge opportunity!
- What's the right model for the problem?
- What techniques can we use to optimize models - quantization (reducing precision), model pruning, knowledge distillation (student, teacher)
- Relatively short feedback time to implement - rapid changes
- As a rough proxy, look at performance vs. cost over the last few years

How do you optimize across each stage?