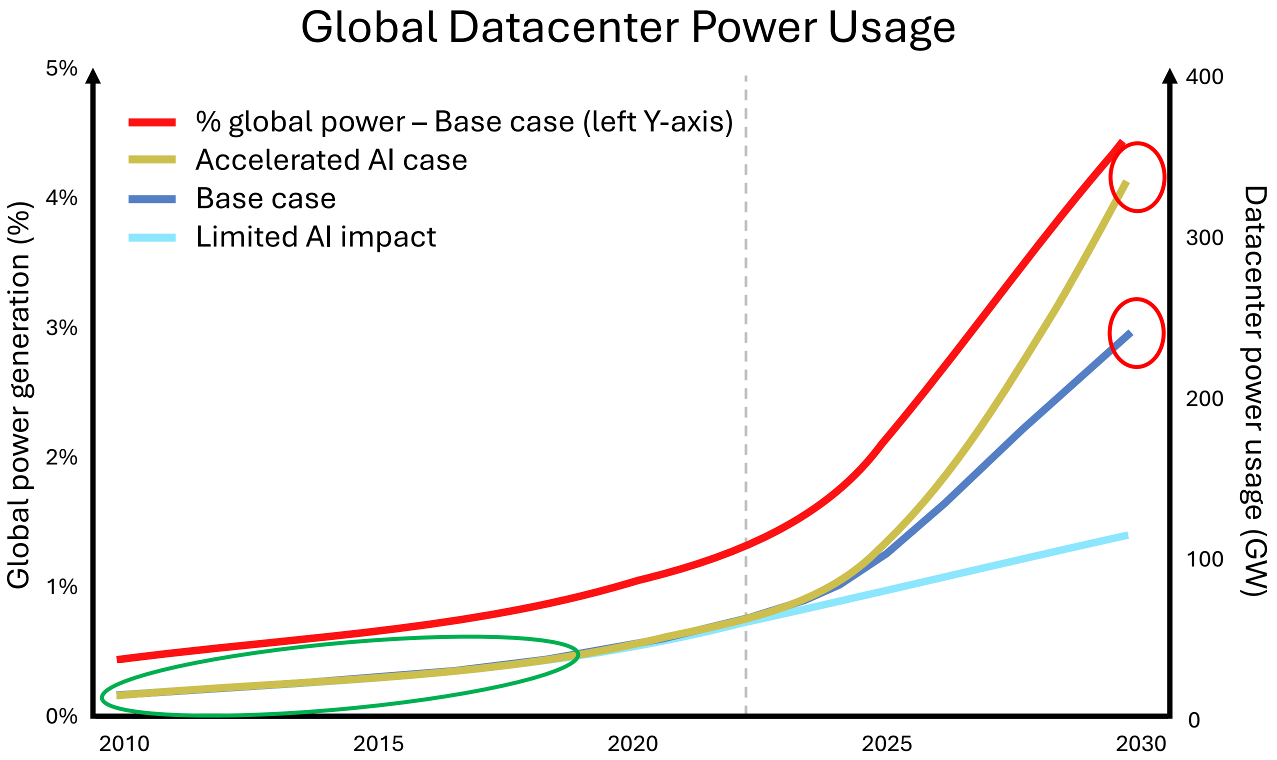
Some datacenter power and energy challenges in the AI era

Ricardo Bianchini

Microsoft Azure

AI datacenter electricity demand



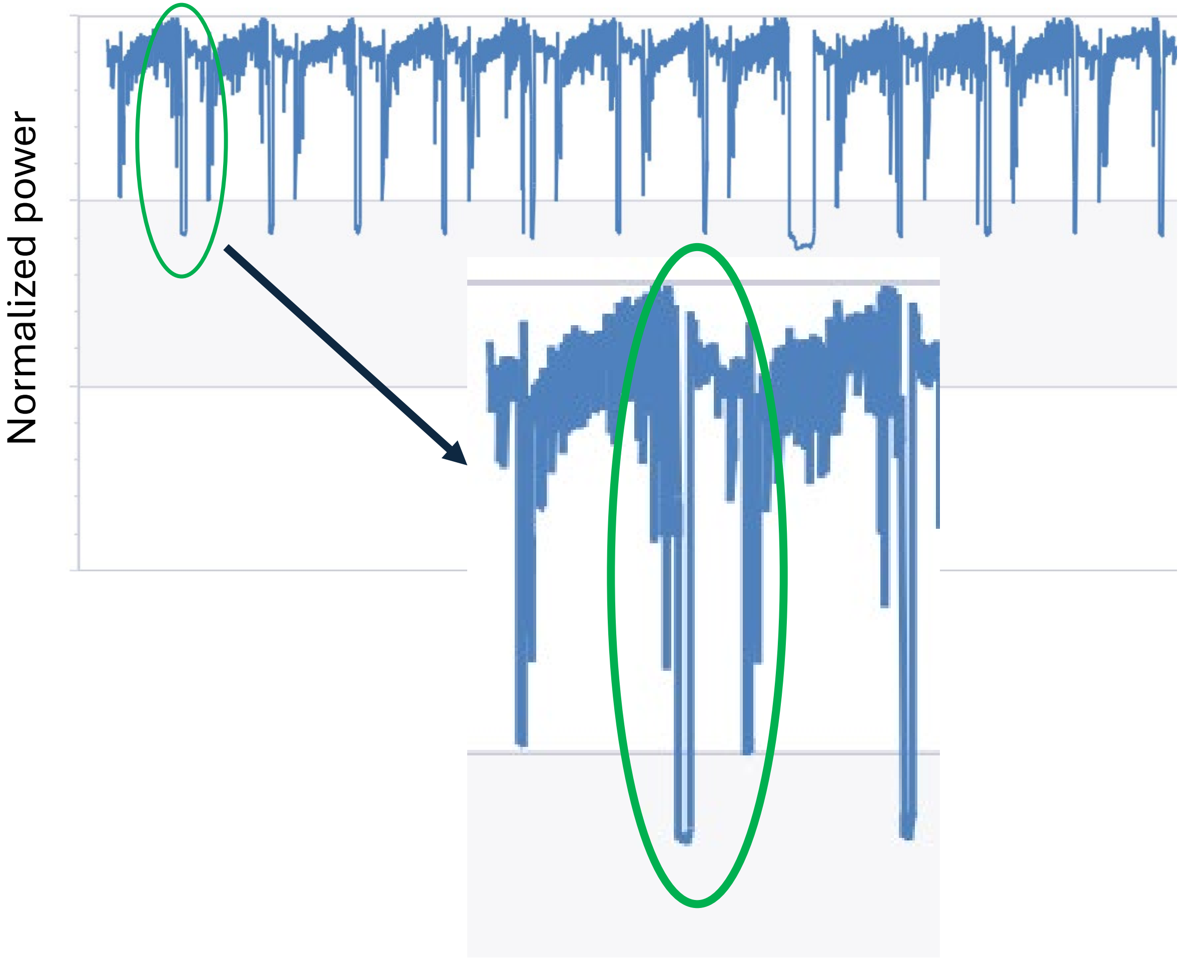
Base case suggests 171 GW in 2030 (3x today!)

Assuming the "accelerated AI" case, 242 GW!

Efficiency kept demand increasing slowly

Source: D. Patel, D. Nishball, and J. E. Ontiveros. "AI Datacenter Energy Dilemma – Race for AI Datacenter Space". March'24.
Image: Semianalysis

Software: Training

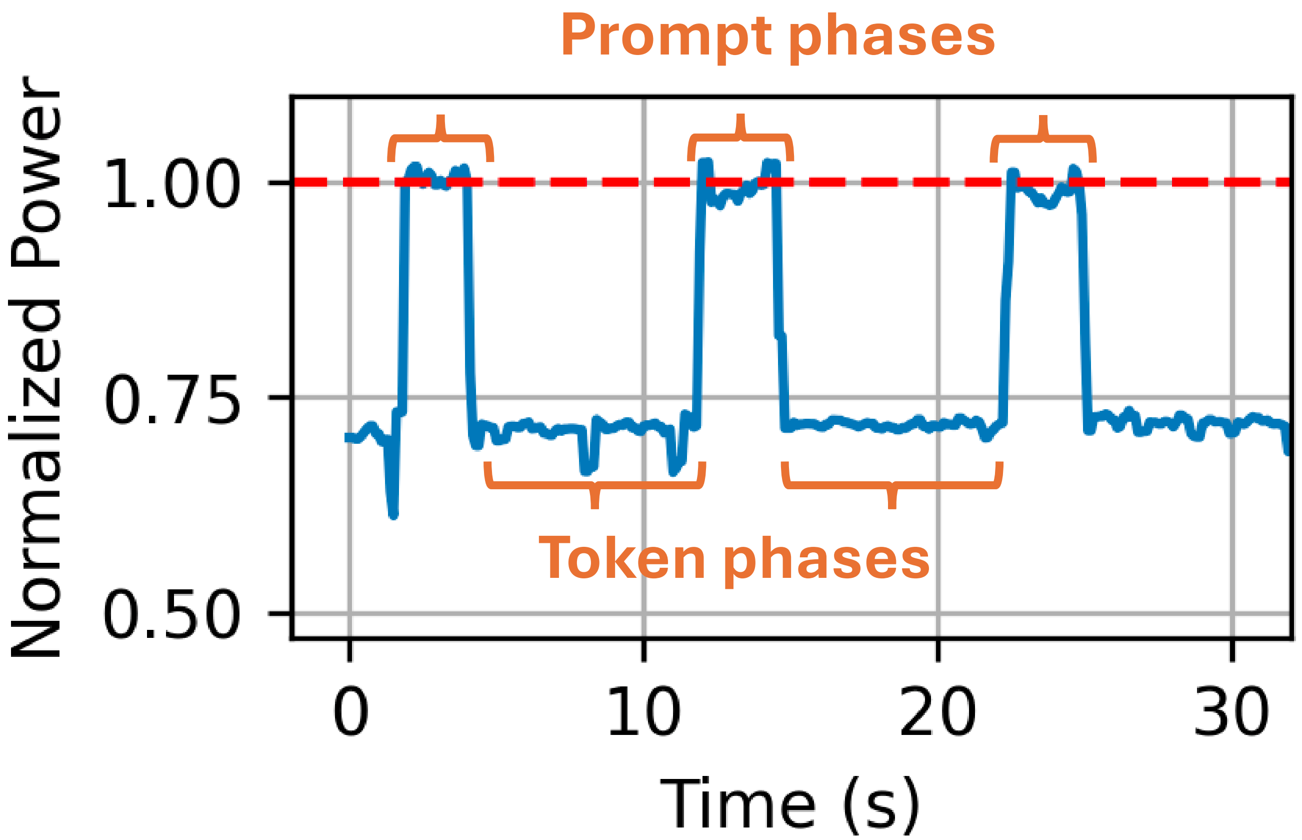


Batch job with high power utilization
Massive synchronous computation

Challenges:

- Power swings can de-stabilize the grid
- Decreasing effective utilization due to failures

Software: Inference



3x BLOOM-176B inference requests on 8 GPUs

Interactive service with lower utilization
Phases behave very differently

Challenges:

- Optimizing for such different behaviors
- Maximizing power oversubscription