CoreWeave

Capacity Plans for Real-World AI Workloads

A unified framework for managing capacity for your AI workloads

Explore Capacity Plans

Talk with an expert to see how CoreWeave Capacity Plans can support your AI workloads.

Contact us

CoreWeave Capacity Plans offer a more flexible way to run AI that match demand patterns, anchored by Flex Reservations and Spot. With these offerings, CoreWeave sets the standard for running the AI cloud: guaranteed capacity when it counts, flexible pricing when you need it.

Innovation thrives when capacity is available

When capacity can't be trusted, teams stop building toward the upside. They over-protect early ideas with always-on commitments, or they keep projects small because scaling feels like a risk. Either way, the cloud stops being the place where new ideas can grow without friction.

You need a cloud built for demand variability

Being ready for AI isn't just raw performance. You need the ability to handle real demand variability without changing how you build or forcing every workload into the same cost-and-capacity tradeoff. CoreWeave Capacity Plans separate guaranteed access from 24/7 economics to enable innovation without limitation.

21% budget lost	Nearly a quarter of enterprise cloud spend is drained on underutilized resources
$44.5B cloud waste	That's how much organizations lost in 2025 from waste on cloud usage
55% guesswork	Over half of developers say that purchasing commitments are based on guesswork

Source
Harness Press Release, February 2025

Match workload demands to the right pricing model

AI in production moves in waves. Demand rises and falls over time, which makes it harder to predict than steady baseline workloads. Your cloud capacity plans should support both.

CoreWeave Capacity Plans enable teams to run the full range of AI workloads with flexible, usage-based pricing models tailored to their real-world needs.

	Flex Reservations	Reservations	Spot	On-Demand
Capacity	Fully guaranteed	Fully guaranteed	Best-effort	Best-effort
Interruption	No	No	Yes (7 min. notice)	No
Pay-as-you-go	Usage only (fixed holding rate)	No	Yes	Yes
Commitment	Fixed term	Fixed term	None	None
Ideal for	Spikes and mission-critical dev	Steady, 24/7 production	Interruptible workloads	Short-term experiments

Capacity plan options aligned with demand

CoreWeave's flexible Capacity Plans let you buy capacity that matches the shape of your workloads—baseline, peaks, and tolerant work—with clear tradeoffs between cost and certainty.

7 min. | Spot preemption notice on CoreWeave

Up to 3.5x more time to checkpoint and recover than notice windows commonly offered by general-purpose cloud providers.

Reservations

Always-on, guaranteed capacity for steady, continuous workloads

Reservations are designed for baseline needs when you know capacity is running 24/7, such as production inference, core training environments, or foundational infrastructure. Reservations requires that you commit to a defined amount of capacity and receive predictable economics in return.

Best for: steady-state, always-on workloads

Tradeoff: highest commitment in exchange for maximum predictability

Flex Reservations

Guaranteed capacity for workloads that rise and fall

With Flex, customers select a ceiling and CoreWeave guarantees access up to that level. Unlike traditional reservations, you don't pay as if they are running at peak all the time to keep that guarantee in place.

Flex introduces a holding fee to maintain guaranteed access, plus usage charges when instances are actually running. This makes it a better fit for environments where utilization is unpredictable, adoption curves are uncertain, and growth can accelerate quickly.

Flex offers guaranteed capacity without requiring you to pre-schedule exact windows or overbuy full-time reservations just to innovate.

Best for: predictable but non-constant peaks

Tradeoff: modest holding cost in exchange for guaranteed burst capacity

On-Demand

Immediate, best-effort capacity

With a CoreWeave Reservations subscription, you can use On-Demand to pay only while instances are running. There's no reservation or holding requirement for the capacity you launch. Capacity is not guaranteed and depends on current availability at the time of request.

On-Demand is useful for incremental top-up, urgent needs, or highly variable workloads where flexibility matters more than certainty.

Best for: short-term or unpredictable needs

Tradeoff: no commitment, but no guarantee of availability

Spot

Lower-cost capacity for workloads that can tolerate interruption

Spot represents a lower cost option where capacity may be preempted, but interruption is explicit and signaled in advance. With this notice, teams can shift work, checkpoint progress, or gracefully terminate workloads. It is designed for experiments, backfills, batch processing, overflow, and other tolerant workloads.

Delivered as a node pool type in CoreWeave Kubernetes Service (CKS), Spot includes clear preemption signaling and operational guardrails. Spot allows teams to intentionally trade interruption tolerance for lower cost.

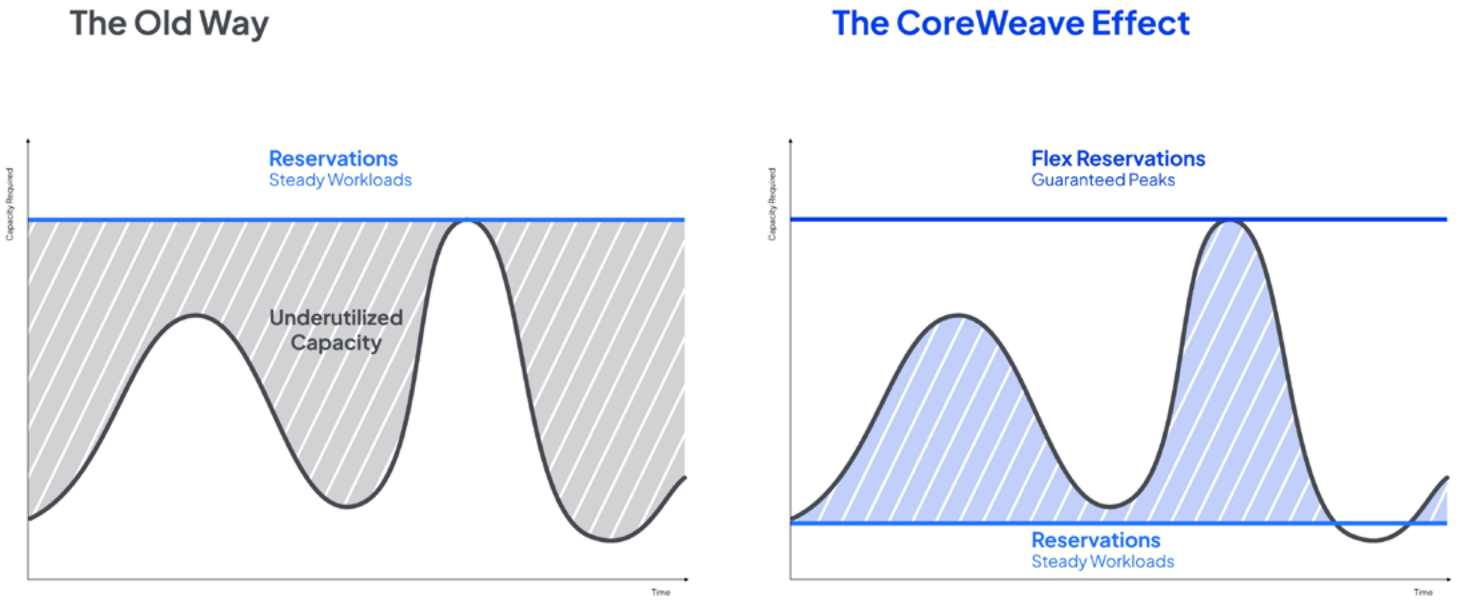
Best for: batch, experiments, backfills, overflow

Tradeoff: lower cost in exchange for potential interruption

Unlock true cloud elasticity with CoreWeave Cloud

Most clouds still force a blunt tradeoff. You either overbuy on always-on capacity to protect availability, or you rely on best-effort scaling when demand spikes. With CoreWeave, you get a clean, controllable way to match cost and certainty to how your AI actually runs. Innovation isn't gated by capacity tradeoffs that don't fit the workload.

In practice, many teams use a mix of these plans, since most AI environments are portfolios. The right approach won't force every demand pattern into one rigid shape. We give teams a simple framework: reserve what is steady, use Flex when peaks must be protected, and use Spot or On-Demand when work can move or when best-effort is acceptable.



Prove AI in production

CoreWeave ARENA: Validate real workloads before you scale. CoreWeave ARENA gives teams a controlled way to run end-to-end workloads under production-like conditions. Test performance, understand cost behavior, and evaluate scaling dynamics before committing to long-term capacity. It's designed for the moment when experimentation becomes serious and decisions need real signal.

Weights & Biases Inference: Deploy and iterate in production with workflow continuity. Weights & Biases Inference provides a tooling-integrated path to deploy and manage inference in production, built on CoreWeave. Teams can iterate existing ML workflows while preserving explicit GPU selection, infrastructure-aligned economics, and visibility into performance and cost as demand evolves.

Manage AI capacity with clarity and control

No more one-size commitments or best-effort risk. No more guesswork. CoreWeave Flexible Capacity Plans restore the elasticity AI teams need to match each workload to the right capacity model, maintain consistent economics, and enable predictable scaling. The impact: new ideas can grow, successful workloads can scale, and production systems can be operated with control instead of guesswork.

From ground to cloud

Capacity plans aren't the only thing CoreWeave purpose-built for AI. Our AI-native platform of technology, tools, and teams is fully integrated to power your most complex workloads.

CoreWeave ARENA

Want to test your Spot workloads before buying? Assess workload performance and investment cost on CoreWeave ARENA before you commit to production.

Real-world AI runs here

CoreWeave offers the most flexible way to run AI with guaranteed peaks without pre-scheduling or overbuying.

Contact us

Flex and Spot, explained

How Flex Reservations and Spot work, how they're priced, and when to use each.

Read more →

Capacity planning made easy

See how our Capacity Plans are built for how you actually run AI.

Watch now →

AI Inference on CoreWeave

Run low-latency inference on high-performance compute.

Learn more →