CPB-2026-002

LLM çıkarımı için GPU belleği ve bant genişliği 2026: dokuz açık modelde H100, H200 ve L40S

H100, H200 ve L40S'te dokuz Qwen, Llama, DeepSeek ve MiniMax modeli için model sığması, KV önbellek kapasitesi ve decode bant genişliği tavanlarının türetilmiş analizi.

SÜRÜM
1.0

YAYIMLANDI
6 Kasım 2026

GÜNCELLENDİ
6 Kasım 2026

ÇEVRİMİÇİ SÜRÜM
https://www.cloudproviderbenchmarks.com/tr/reports/llm-cikarimi-gpu-bellek-bant-genisligi-2026/

İçindekiler

Yönetici özeti

1. Amaç ve yöntem

2. GPU’lar

3. Model sığması ve KV bütçesi

4. Bağlam uzunluğuna göre eşzamanlılık

5. Decode tavanları

6. Erişilebilirlik ve fiyatlar

7. Yorum

8. Sonraki sürüm

Sınırlamalar

Sürüm geçmişi

Kaynaklar

KISA YANIT

FP8 ile 7B–8B modeller üç GPU'nun herhangi birine sığar; 32B yoğun modeller tek H100 veya H200'e sığar ancak L40S'te KV için az yer bırakır; 70B–72B modeller tek H200 veya iki H100 gerektirir; DeepSeek-V3 ise 8 GPU'luk tam bir H200 sunucusu gerektirir. Decode tavanları bant genişliğiyle ölçeklenir: H200 ≈ 1,43× H100 SXM ≈ 5,6× L40S.

YÖNETİCİ ÖZETİ

Bu rapor, yayımlanmış bir bellek modelini dokuz açık ağırlıklı LLM'e — Qwen2.5 7B ve 72B, Qwen3 32B ve 30B-A3B, Llama 3.1 8B ve 3.3 70B, DeepSeek-R1-Distill-Qwen-32B, DeepSeek-V3 ve MiniMax-M1 — NVIDIA H100, H200 ve L40S üzerinde uygular. Her model ve hassasiyet için gereken en az GPU sayısını, KV önbellek bütçesini, bağlam uzunluğuna göre en fazla eşzamanlı dizi sayısını ve bant genişliğine bağlı decode tavanını raporlar. Tüm değerler teknik özelliklerden türetilmiştir ve bu şekilde etiketlenmiştir.

TEMEL BULGULAR

01 FP8 ile Llama 3.3 70B, yaklaşık 49 GB KV bütçesiyle (yaklaşık 18 eşzamanlı 8K token'lık dizi) tek H200'e, yaklaşık 65 GB ile iki H100'e sığar.

02 FP8 ile Qwen3 32B, tek H100'de yaklaşık 16, tek H200'de yaklaşık 40 eşzamanlı 8K dizi barındırır.

03 Qwen3 30B-A3B'nin FP8 decode tavanı H100 SXM'de yaklaşık 1.015 token/s'dir, çünkü token başına yalnızca ~3,3B parametre etkindir.

04 L40S'te 32B FP8 model yaklaşık 8 GB KV bütçesi — üç adet 8K dizi — bırakır; bu da kullanımını küçük modellerle veya çok kartlı kurulumlarla sınırlar.

05 Ölçülmüş sonuç yer almamaktadır; ilgili benchmark sayfaları sonuçları yalnızca doğrulamadan sonra yayımlar.

1. Amaç ve yöntem

LLM çıkarımı için GPU yapılandırması seçmek bellekle başlar: ağırlıklar sığıyor mu ve eşzamanlı istekler için ne kadar KV önbellek kalıyor? Bu rapor, GPU ve LLM çıkarım metodolojisini kullanarak bu soruyu dokuz model ve üç GPU için yanıtlar:

- ağırlıklar = parametre sayısı × parametre başına bayt;

- token başına KV = 2 × katman × KV head × head boyutu × 2 bayt;
- kullanılabilir bellek = cihaz belleğinin %85'i (%90 kullanım − %5 pay);
- decode tavanı = bellek bant genişliği ÷ etkin ağırlık baytları.

2. GPU'lar

GPU spesifikasyonları (NVIDIA veri sayfaları)

GPU	MİMARİ	BELLEK (GB)	BELLEK TİPİ	BANT GEN. (GB/S)	FP8 (TFLOPS)	BF16 (TFLOPS)	FP32 (TFLOPS)
H100 SXM5 SXM5	Hopper	80	HBM3	3.350	1.979	989	67
H200 SXM SXM5	Hopper	141	HBM3e	4.800	1.979	989	67
L40S PCIe dual-slot	Ada Lovelace	48	GDDR6 (ECC)	864	733	362	91,6

Tensor değerleri yoğun (seyrek olmayan) tepe değerleridir; NVIDIA seyrekli değerleri bunun iki katı olarak yayımlar.

3. Model sığması ve KV bütçesi

Model bellek gereksinimleri ve minimum GPU sayısı (FP8 ağırlıklar)

MODEL	PARAMETRE (B)	BF16 (GB)	FP8 (GB)	INT4 (GB)	KV / TOKEN (KİB)	KV @32K (GB)	MİN. H100 SXM5
Llama 3.1 8B Instruct Yoğun · GQA	8	16	8	4	128	4,3	1×
Llama 3.3 70B Instruct Yoğun · GQA	70,6	141	71	35	320	10,7	2×
Qwen2.5 7B Instruct Yoğun · GQA	7,6	15	8	4	56	1,9	1×
Qwen3 32B Yoğun · GQA	32,8	66	33	16	256	8,6	1×
Qwen3 30B-A3B MoE · 3,3B aktif · GQA	30,5	61	31	15	96	3,2	1×
Qwen2.5 72B Instruct Yoğun · GQA	72,7	145	73	36	320	10,7	2×
DeepSeek-R1-Distill-Qwen-32B Yoğun · GQA	32,8	66	33	16	256	8,6	1×
DeepSeek-V3 MoE · 37B aktif · MLA	671	1.342	671	336	69	2,3	>8
MiniMax-M1-80k MoE · 45,9B aktif · HYBRID-LINEAR	456	912	456	228	–	–	8×

Ağırlık boyutu = parametre × hassasiyet baytı. KV = 2 × katman × KV başlığı × başlık boyutu × 2 bayt (MLA için latent boyutu). Minimum GPU: %85 kullanılabilir bellek ve GPU başına en az 4 GB KV payı. Ölçüm değil, kapasite modelidir.

4. Bağlam uzunluğuna göre eşzamanlılık

Aşağıdaki matrisler, temsili yapılandırmalar için her bağlam uzunluğunda hangi eşzamanlılık düzeylerinin sığdığını gösterir.

Llama 3.1 8B Instruct · FP8 · 1× L40S

Çalıştırılabilir senaryo matrisi (KV önbelleği kapasite modeli) · KV budget ≈ 32.8 GB

BAĞLAM	1	5	10	25	50	100	MAKS. EŞZ.
8K	✓	✓	✓	✓	–	–	30
16K	✓	✓	✓	–	–	–	15
32K	✓	✓	–	–	–	–	7
64K	✓	–	–	–	–	–	3
128K	✓	–	–	–	–	–	1

✓ senaryo bellek modeline sığar · — KV önbelleği aşılır (çalıştırılmaz) · ✕ ağırlıklar sığmaz · n/a model bu bağlamı desteklemez. Model: %90 bellek kullanımı, %5 çalışma zamanı rezervi, 16-bit KV önbelleği.

Qwen3 32B · FP8 · 1× H100 SXM5

Çalıştırılabilir senaryo matrisi (KV önbelleği kapasite modeli) · KV budget ≈ 35.2 GB

BAĞLAM	1	5	10	25	50	100	MAKS. EŞZ.
8K	✓	✓	✓	–	–	–	16
16K	✓	✓	–	–	–	–	8
32K	✓	–	–	–	–	–	4
64K	✓	–	–	–	–	–	2
128K	✓	–	–	–	–	–	1

✓ senaryo bellek modeline sığar · — KV önbelleği aşılır (çalıştırılmaz) · ✕ ağırlıklar sığmaz · n/a model bu bağlamı desteklemez. Model: %90 bellek kullanımı, %5 çalışma zamanı rezervi, 16-bit KV önbelleği.

Qwen3 32B · FP8 · 1× H200 SXM

Çalıştırılabilir senaryo matrisi (KV önbelleği kapasite modeli) · KV budget ≈ 87.0 GB

BAĞLAM	1	5	10	25	50	100	MAKS. EŞZ.
8K	✓	✓	✓	✓	–	–	40
16K	✓	✓	✓	–	–	–	20
32K	✓	✓	✓	–	–	–	10
64K	✓	✓	–	–	–	–	5
128K	✓	–	–	–	–	–	2

✓ senaryo bellek modeline sığar · — KV önbelleği aşılır (çalıştırılmaz) · ✕ ağırlıklar sığmaz · n/a model bu bağlamı desteklemez. Model: %90 bellek kullanımı, %5 çalışma zamanı rezervi, 16-bit KV önbelleği.

Llama 3.3 70B Instruct · FP8 · 1× H200 SXM

Çalıştırılabilir senaryo matrisi (KV önbelleği kapasite modeli) · KV budget ≈ 49.3 GB

BAĞLAM	1	5	10	25	50	100	MAKS. EŞZ.
8K	✓	✓	✓	–	–	–	18
16K	✓	✓	–	–	–	–	9
32K	✓	–	–	–	–	–	4
64K	✓	–	–	–	–	–	2
128K	✓	–	–	–	–	–	1

✓ senaryo bellek modeline sığar · — KV önbelleği aşılır (çalıştırılmaz) · ✕ ağırlıklar sığmaz · n/a model bu bağlamı desteklemez. Model: %90 bellek kullanımı, %5 çalışma zamanı rezervi, 16-bit KV önbelleği.

Llama 3.3 70B Instruct · FP8 · 2× H100 SXM5

Çalıştırılabilir senaryo matrisi (KV önbelleği kapasite modeli) · KV budget ≈ 65.4 GB

BAĞLAM	1	5	10	25	50	100	MAKS. EŞZ.
8K	✓	✓	✓	–	–	–	24
16K	✓	✓	✓	–	–	–	12
32K	✓	✓	–	–	–	–	6
64K	✓	–	–	–	–	–	3
128K	✓	–	–	–	–	–	1

✓ senaryo bellek modeline sığar · — KV önbelleği aşılır (çalıştırılmaz) · ✕ ağırlıklar sığmaz · n/a model bu bağlamı desteklemez. Model: %90 bellek kullanımı, %5 çalışma zamanı rezervi, 16-bit KV önbelleği.

5. Decode tavanları

Tek akışlı decode için bellek bant genişliği tavanı (FP8, tok/s)
Teorik üst sınır — ölçüm değil

MODEL	H100 SXM5	H200 SXM	L40S
Llama 3.1 8B Instruct	417	598	108
Llama 3.3 70B Instruct	95 2× TP	68	24 2× TP
Qwen2.5 7B Instruct	440	631	114
Qwen3 32B	102	146	26
Qwen3 30B-A3B	1.015	1.455	262
Qwen2.5 72B Instruct	92 2× TP	66	24 2× TP
DeepSeek-R1-Distill-Qwen-32B	102	146	26
DeepSeek-V3	sığmaz	1.038 8× TP	sığmaz

Tavan = toplam bellek bant genişliği ÷ token başına okunan aktif ağırlık baytı (batch 1, KV okumaları ve iletişim yükü hariç, mükemmel tensor parallelism varsayımı). Gerçek motorlar bu sınırın bir kısmına ulaşır.

Tek akışlı decode için bellek bant genişliği tavanı (BF16, tok/s)

Teorik üst sınır — ölçüm değil

MODEL	H100 SXM5	H200 SXM	L40S
Llama 3.1 8B Instruct	209	299	54
Llama 3.3 70B Instruct	95 4× TP	68 2× TP	24 4× TP
Qwen2.5 7B Instruct	220	315	57
Qwen3 32B	102 2× TP	73	26 2× TP
Qwen3 30B-A3B	508	727	262 2× TP
Qwen2.5 72B Instruct	92 4× TP	66 2× TP	24 4× TP
DeepSeek-R1-Distill-Qwen-32B	102 2× TP	73	26 2× TP
DeepSeek-V3	sığmaz	sığmaz	sığmaz

Tavan = toplam bellek bant genişliği ÷ token başına okunan aktif ağırlık baytı (batch 1, KV okumaları ve iletişim yükü hariç, mükemmel tensor parallelism varsayımı). Gerçek motorlar bu sınırın bir kısmına ulaşır.

6. Erişilebilirlik ve fiyatlar

Sağlayıcılara göre GPU erişilebilirliği ve yayımlanmış GPU-saat fiyatları

SAĞLAYICI	NVIDIA H100	NVIDIA H200	NVIDIA L40S	TÜRKİYE AL
PortvMind Cloud	Genel kullanımda gpu.h100s / gpu.h100m profiles ₺195,835 / GPU-h Sunucu başına 1, 2, 3 veya 4 GPU; GPU başına 80 GB (81.920 MB) VRAM; GPU başına 26 vCPU ve 496 GB RAM; yerel disk dahil.	Genel kullanımda GPU / AI Compute (H200 option) Mevcut GPU seçeneği olarak listelenir; konfigürasyon ve fiyatlar kamuya açık hesaplayıcı tablosunda yer almaz, konsol veya satış ekibi üzerinden sunulur.	Genel kullanımda GPU / AI Compute (L40S option) Mevcut GPU seçeneği olarak listelenir; konfigürasyon ve fiyatlar kamuya açık hesaplayıcı tablosunda yer almaz, konsol veya satış ekibi üzerinden sunulur.	Var
Amazon Web Services	Genel kullanımda EC2 P5 Tek GPU'lu p5.4xlarge ve 8 GPU'lu p5.48xlarge boyutları; erişilebilirlik bölgeye göre değişir.	Genel kullanımda EC2 P5e / P5en 8 GPU'lu sunucular; erişilebilirlik bölgeye göre değişir.	Genel kullanımda EC2 G6e Sunucu başına 1–8 L40S GPU.	Var
Microsoft Azure	Genel kullanımda NC H100 v5 / ND H100 v5 Bir veya iki H100 NVL GPU'lu NC H100 v5; sekiz H100 GPU'lu ND H100 v5.	Genel kullanımda ND H200 v5 VM başına sekiz H200 GPU.	Sunulmuyor L40S SKU'ları Azure Local / Azure Stack Hub (müşteri sahası donanımı) için belgelenmiştir.	Yok
Google Cloud	Genel kullanımda A3 High / A3 Mega VM başına 8'e kadar H100 GPU; zone erişilebilirliği değişir.	Genel kullanımda A3 Ultra VM başına 8 H200 GPU; zone erişilebilirliği değişir.	Sunulmuyor	Duyuruldu
DigitalOcean	Genel kullanımda GPU Droplet — NVIDIA HGX H100 $4,410 / GPU-h 1× veya 8× H100 80 GB; 1× planda 20 vCPU ve 240 GiB RAM.	Genel kullanımda GPU Droplet — NVIDIA HGX H200 $4,470 / GPU-h 1× veya 8× H200 141 GB; 1× planda 24 vCPU ve 240 GiB RAM.	Genel kullanımda GPU Droplet — NVIDIA L40S $1,570 / GPU-h 8 vCPU ve 64 GiB RAM ile 1× L40S 48 GB.	Yok
Vultr	Genel kullanımda Cloud GPU — NVIDIA HGX H100 $2,300 / GPU-h 8× H100 80 GB SXM (640 GB GPU RAM, 216 vCPU, 1.914 GB RAM).	Doğrulanmadı Not listed on the public pricing page Son inceleme tarihinde H200 Vultr fiyat sayfasında listelenmiyordu.	Genel kullanımda Cloud GPU — NVIDIA L40S $1,671 / GPU-h 1, 2, 4 veya 8× L40S 48 GB; 1× planda 16 vCPU ve 180 GB RAM.	Yok
Hetzner	Sunulmuyor	Sunulmuyor	Sunulmuyor	Yok

Yayımlanmış GPU-saat liste fiyatları

SAĞLAYICI	GPU	GPU-SAAT	GB-SAAT	ESAS
PortvMind	NVIDIA H100	₺195,835	₺2,4479	gpu.h100s.xlarge (1× H100) saatlik liste fiyatı, tr-ist-01, vergiler hariç. Çoklu GPU profillerinde GPU-saat başına ₺141,12'den başlar (3× ve 4× profiller). · 6 Kasım 2026
DigitalOcean	NVIDIA H100	$4,410	$0,0551	İsteğe bağlı, 1× H100 GPU Droplet. · 6 Kasım 2026
DigitalOcean	NVIDIA H200	$4,470	$0,0317	İsteğe bağlı, 1× H200 GPU Droplet. · 6 Kasım 2026
DigitalOcean	NVIDIA L40S	$1,570	$0,0327	İsteğe bağlı, 1× L40S GPU Droplet. · 6 Kasım 2026
Vultr	NVIDIA H100	$2,300	$0,0287	İsteğe bağlı "başlangıç" GPU başı fiyatı, 8 GPU'lu HGX konfigürasyonu. · 6 Kasım 2026
Vultr	NVIDIA L40S	$1,671	$0,0348	İsteğe bağlı GPU başı fiyat, 1× L40S. · 6 Kasım 2026

7. Yorum

- **Yapılandırmayı bellek, hızı bant genişliği belirler.** H200'ün H100'e karşı avantajı tamamen bu iki değerdedir.
- **MoE modeller dengeyi değiştirir.** Qwen3 30B-A3B, 30B'lik bir modelin belleğine ihtiyaç duyar ancak 3B'lik bir model gibi decode eder.
- **Uzun bağlamlar pahalıdır.** 128K token'da, 16-bit KV önbellekle H100 üzerindeki 8B'lik bir model bile yalnızca üç eşzamanlı diziyi destekler.
- **L40S küçük modellere uygundur.** FP8 ile ~14B parametrenin üzerinde 48 GB belleği eşzamanlılığı ciddi biçimde kısıtlar.

8. Sonraki sürüm

Bu raporun 2.0 sürümü, GPU'ların mevcut olduğu takip edilen her sağlayıcıda doğrulanmış vLLM koşularından ölçülmüş TTFT, TPOT, verim ve milyon token başına maliyet değerlerini ekleyecektir.

Sınırlamalar

- Tüm değerler, belirtilen formüllerle teknik özelliklerden türetilmiştir; bunlar ölçüm değil, planlama sınırlarıdır.
- Önek paylaşımı olmayan 16-bit KV önbellek varsayılmıştır.
- MiniMax-M1'in hibrit attention yapısı KV formülü kapsamında değildir.
- H100 ve H200 için SXM varyantları modellenmiştir; PCIe ve NVL varyantları farklıdır.

Sürüm geçmişi

1.0	6 Kasım 2026	İlk baskı. Yalnızca türetilmiş analiz; ölçülmüş sonuçlar sonraki bir sürümde eklenecek.

Kaynaklar

[1] NVIDIA H100 Tensor Core GPU — NVIDIA Üretici veri sayfası
erişim: 6 Kasım 2026

[2] NVIDIA H200 Tensor Core GPU — NVIDIA Üretici veri sayfası
erişim: 6 Kasım 2026

[3] NVIDIA L40S GPU — NVIDIA Üretici veri sayfası
erişim: 6 Kasım 2026

[4] Llama 3.3 70B Instruct — Meta Resmî dokümantasyon
erişim: 6 Kasım 2026

[5] Llama 3.1 8B Instruct — Meta Resmî dokümantasyon
erişim: 6 Kasım 2026

[6] Qwen3 32B — Qwen Resmî dokümantasyon
erişim: 6 Kasım 2026

[7] Qwen3 30B-A3B — Qwen Resmî dokümantasyon
erişim: 6 Kasım 2026

[8] Qwen2.5 72B Instruct — Qwen Resmî dokümantasyon
erişim: 6 Kasım 2026

[9] Qwen2.5 7B Instruct — Qwen Resmî dokümantasyon
erişim: 6 Kasım 2026

[10] DeepSeek-V3 — DeepSeek Resmî dokümantasyon
erişim: 6 Kasım 2026

[11] DeepSeek-R1-Distill-Qwen-32B — DeepSeek Resmî dokümantasyon
erişim: 6 Kasım 2026

[12] MiniMax-M1-80k — MiniMax Resmî dokümantasyon
erişim: 6 Kasım 2026

[13] vLLM documentation — vLLM project Resmî dokümantasyon
erişim: 6 Kasım 2026