Intel's $949 GPU Has 32GB VRAM and 608 GB/s Bandwidth: What It Means for Local AI

March 25, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** The Intel Arc Pro B70 gives you 32GB VRAM for $949 — 8GB more than an RTX 3090 for roughly the same price. The catch: 608 GB/s bandwidth (vs the 3090's 936 GB/s) means slower token generation for models that already fit in 24GB. Buy it if you need more than 24GB and want a new card with a warranty. Stick with the 3090 if speed matters more than capacity.

📚 **More on this topic:** GPU Buying Guide · VRAM Requirements · What Can You Run on 24GB? · Used RTX 3090 Guide

Intel just did something nobody expected. The Arc Pro B70, launched today, puts 32GB of GDDR6 on a single card for $949. That's more VRAM than any consumer NVIDIA GPU under $2,000.

For anyone running local LLMs, 32GB opens a door that 24GB keeps shut. Models like Qwen 3.5 27B at Q6_K that barely squeeze into 24GB? They run comfortably with room for context. Llama 3.3 70B at aggressive quantization? Actually possible without a multi-GPU setup.

But there's a catch — and it's a big one. Intel's software ecosystem for AI inference is the weakest of the three GPU vendors. CUDA dominates. ROCm is catching up. Intel has OneAPI and SYCL, and whether those work well enough for your use case is the question this article tries to answer.

The specs

The Arc Pro B70 is a workstation GPU, not a gaming card. Intel built it on the full Battlemage die (BMG-G31) with 32 Xe2 cores, the same architecture as the gaming Arc B580 but with the full chip enabled.

Spec	Arc Pro B70
VRAM	32GB GDDR6

Spec	Arc Pro B70
Memory bandwidth	608 GB/s
Memory bus	256-bit
FP32 compute	22.94 TFLOPS
INT8 AI TOPS	367 TOPS
TDP	230W (Intel reference)
Power connector	1x 8-pin
PCIe	5.0 x16
ECC memory	Yes
Price	$949 (Intel branded)
Available from	March 25, 2026

Board partners (ARKN, ASRock, Gunnir, Maxsun, Sparkle) will sell their own versions with TDPs ranging from 160W to 290W. A lower-tier **Arc Pro B65** with the same 32GB is coming mid-April through partners only — pricing TBD.

How it compares to NVIDIA

Here's the comparison that matters for local AI:

	Arc Pro B70	RTX 3090 (used)	RTX 4090	RTX 5070
VRAM	32GB	24GB	24GB	12GB
Bandwidth	608 GB/s	936 GB/s	1,008 GB/s	~640 GB/s
Price	$949 new	$800-900 used	$1,600-1,800	$549
TDP	230W	350W	450W	250W
Software stack	OneAPI/SYCL	CUDA	CUDA	CUDA
Warranty	New, full warranty	Used, buyer beware	New, full warranty	New, full warranty

The bandwidth gap is the story. LLM inference is memory-bandwidth bound — the GPU spends most of its time reading model weights from VRAM, not computing. The RTX 3090's 936 GB/s

gives it roughly 54% more bandwidth than the B70's 608 GB/s. For any model that fits in 24GB, the 3090 will generate tokens faster.

But bandwidth only matters for models that fit. If you need 28GB to load a model, the 3090 can't do it at all. The B70 can.

What 32GB lets you run that 24GB can't

This is where the B70 gets interesting. Eight extra gigabytes opens up several models and configurations that choke on 24GB cards:

Model	Quantization	VRAM needed	Fits in 24GB?	Fits in 32GB?
Qwen 3.5 27B	Q6_K	~21GB	Tight, minimal context	Yes, with room
Qwen 3.5 27B	Q8_0	~28GB	No	Yes
Llama 3.3 70B	Q3_K_M	~30GB	No	Tight
Qwen 3.5 35B-A3B	Q4_K_M	~21GB	Yes	Yes, with huge context
DeepSeek R1 32B	Q6_K	~25GB	No	Yes
Any 14B model	FP16	~28GB	No	Yes

The sweet spot: models that need 25-30GB. That's where the B70 has no NVIDIA competition under $2,000. You'd need a 4090 with only 24GB, or jump to professional cards at $5,000+.

For context windows specifically, 32GB is a big deal. Qwen 3.5 27B at Q4_K_M fits in 24GB but spills to system RAM above 131K context on an RTX 3090. The B70 handles this without spillover.

The software problem

This is the elephant in the room. NVIDIA has CUDA, which everything supports out of the box. AMD has ROCm, which is rougher but works. Intel has OneAPI and SYCL, and the story is more complicated.

What works today

llama.cpp: Supported via the SYCL backend. You need Intel's oneAPI Base Toolkit (2025.0 or newer). Flash Attention support landed in March 2026, which reduces memory usage. It works, but setup is more involved than CUDA — you're installing a separate SDK and setting environment variables.

vLLM: Official Intel support exists via LLM-Scaler. The Arc Pro B-series is explicitly supported, including multi-GPU serving. This is probably the most mature option for serving models.

IPEX-LLM: Intel's own PyTorch extension for LLM inference. Supports INT4, FP4, INT8, FP8 quantization. On an Arc A770 (16GB), this hits ~70 tok/s on Mistral 7B — respectable, but that's a smaller model on a card with higher bandwidth-per-GB.

Ollama: A SYCL support PR exists, but it's unclear whether it's merged into mainline Ollama yet. Don't count on `ollama run` working out of the box on day one.

ComfyUI: Works on Intel Arc via `torch.xpu`. Docker images exist. Performance on an Arc A770 is roughly RTX 3070 Ti equivalent for Stable Diffusion workflows.

What this means in practice

If you're used to running `ollama run qwen3.5:27b` and having it just work, the B70 will frustrate you. You'll need to install oneAPI, build llama.cpp with SYCL support, and troubleshoot driver issues that CUDA users never think about.

If you're comfortable building from source and don't mind extra setup, the tools are there. llama.cpp and vLLM both have working Intel backends.

Expected inference speed

No independent B70 benchmarks exist yet (the card launched today). Here's what we can estimate based on existing Intel Arc performance and the bandwidth numbers.

LLM token generation speed is roughly proportional to memory bandwidth divided by model size. With 608 GB/s:

Model	Quantization	Estimated tok/s (B70)	RTX 3090 tok/s
Qwen 3.5 27B	Q4_K_M (~17GB)	~18-22	~25-35

Model	Quantization	Estimated tok/s (B70)	RTX 3090 tok/s
Llama 3.1 8B	Q4_K_M (~5GB)	~50-65	~80-100
Mistral 7B	Q4_K_M (~4.5GB)	~55-70	~90-110
DeepSeek R1 32B	Q4_K_M (~19GB)	~16-20	~22-28

These are estimates based on bandwidth ratios. Intel's own marketing claims "up to 85% higher multi-user throughput" versus the RTX Pro 4000 ($1,800) — but that's comparing against a different NVIDIA card at a higher price point, and multi-user throughput isn't the same as single-user tok/s.

The honest take: for models that fit in 24GB, expect the B70 to be roughly 35-45% slower than an RTX 3090 in token generation. The B70's advantage is running models that don't fit in 24GB at all.

Intel's own benchmarks (take with salt)

Intel published some comparisons against the NVIDIA RTX Pro 4000 ($1,800):

- **Context window**: 93K tokens with Llama 3.1 8B BF16, vs 42K on the RTX Pro 4000
- **Time to first token**: Up to 6.2x faster than the RTX Pro 4000
- **Multi-GPU KV cache**: 4x B70s can hold 304K context on Qwen3 32B FP8

These numbers compare against a more expensive NVIDIA workstation card. No direct RTX 3090 comparison was published. Treat these as marketing until independent reviewers run their own tests.

Who should buy this

Buy the Arc Pro B70 if:

- You need more than 24GB VRAM on a single card and don't want to spend $2,000+
- You're comfortable with non-CUDA software stacks (OneAPI, SYCL, building from source)
- You want a new card with a warranty instead of gambling on the used market
- You plan to run models in the 25-30GB range (70B at Q3, 27B at Q8, 14B at FP16)
- You're already in an Intel ecosystem or want ECC memory for reliability

Stick with a used RTX 3090 if:

- You want maximum tok/s for models that fit in 24GB
- You want everything to work out of the box with CUDA
- You use Ollama, LM Studio, or other tools that assume NVIDIA
- You value the massive CUDA ecosystem (ComfyUI, vLLM, text-generation-webui)
- You're budget-conscious ($800-900 used vs $949 new is close, but the 3090 is faster for most workloads)

Wait if:

- You do image generation primarily (Intel's SD/Flux performance needs independent benchmarks)
- You want to see independent LLM benchmarks before committing
- You're interested in the Arc Pro B65 (same 32GB, potentially cheaper, mid-April)

The bottom line

32GB of VRAM for $949 is unprecedented for a new GPU. Intel priced this aggressively, and the VRAM capacity alone makes the B70 interesting for anyone hitting the 24GB ceiling.

But I wouldn't preorder this blind. The software ecosystem is the risk. CUDA has a decade of momentum. OneAPI and SYCL work, but the setup friction is real, community support is thinner, and not every tool you rely on will have Intel support on day one.

My recommendation: wait two weeks. Let the r/LocalLLaMA community get their hands on it. Once we see real tok/s numbers from llama.cpp SYCL, real Ollama compatibility reports, and real ComfyUI benchmarks — then you'll know whether the VRAM advantage is worth the software tradeoff.

If independent benchmarks confirm 18-22 tok/s on Qwen 3.5 27B Q4_K_M with a working llama.cpp SYCL build, this card earns a spot next to the RTX 3090 in our recommendations. That's a genuine alternative for the first time.

Related guides

- GPU Buying Guide for Local AI

- How Much VRAM Do You Need for Local LLMs?
- What Can You Run on 24GB VRAM?
- Used RTX 3090 Buying Guide
- AMD vs NVIDIA for Local AI
- Intel Arc B580 for Local LLMs

Get notified when we publish new guides.

Subscribe — free, no spam
