BANG: Cost Analysis

In order to choose the right Approximate Nearest Neighbour Search (ANNS) algorithm for a downstream task in a particular domain, the following metrics are taken into consideration: (1) Throughput (QPS) (2) Recall (3) Cost of the machine required to run the ANN search algorithm (4) Power requirement. In this document, we evaluate BANG to determine how it compares to state-of-the-art techniques based on these metrics.

1 Cost Analysis

ANN search methods vary in their techniques, algorithms, scale, implementation, and target hardware which makes it challenging to compare and select the top implementation. The `Billion-Scale Approximate Nearest Neighbor Search Challenge` [1] introduced a leaderboard strategy for objectively comparing ANNS algorithms based on cost factors, with a detailed template available for evaluation [2]. We draw inspiration from this *Cost Leaderboard* to evaluate BANG's cost-effectiveness.

The cost leaderboard computes the overall cost by considering cost, power, throughput and recall. The overall cost is determined by two factors: (1) capital cost and (2) operational cost.

1. The capital cost (CapEx) is the one-time investment cost needed to set up the necessary hardware to implement the respective ANNS technique that can achieve or exceed a predefined baseline recall value (i.e. 0.9 10-recall@10) and deliver a target QPS of 100,000.

2. The operational cost (OpEx) is the running cost to keep the ANNS implementation running for 4 years by continuously delivering the baseline recall and target QPS. The OpEx is computed by estimating the power consumption of the previously described ANNS system and multiplying it with a fixed power tariff ($0.10 / kilowatt-hour).

Thus, the cost leaderboard provides a comprehensive framework for evaluating and comparing ANN search implementations using a normalised scale.

The results of the competition are published on the competition webpage [3], wherein individual scores and corresponding ranks are listed for each submission. Utilizing the evaluation outcomes of BANG on our setup, we calculate BANG's scores and rank it according to the cost leaderboard formula. The resulting cost leaderboard is shown in Table 1. Note: We present the detailed cost computation for BANG on each dataset separately in Section 2. Based on the values shown in Table 1, our technique, BANG, achieves rank **#2**, showcasing its exceptional cost-efficiency and high performance in ANN search implementation. The only implementation outperforming BANG is *optanne_graphann*, which uses Intel Optane DC Persistent Memory Module (PMM). However, this hardware is no longer in production. Consequently, **BANG emerges as the superior** implementation among other state-of-the-art alternatives which are readily integrable into production systems.

For completeness, we also show BANG's rank in the other competition leaderboards *viz.* throughput and power in Table 2 and Table 3 respectively.

Cost Rankings

Rank	Submission	Team	Hardware	Score*	DEEP1B	BigANN	SPACEV1B
1	optanne_graphann	Intel	2x Xeon 6330N + Optane	$-1,971,552.00	$16,086.82	$15,439.92	$16,382.81
2	**BANG**	**-**	**1x NVIDIA A100 GPU**	**$-1,945,809.00**	**$31,863**	**$20,507**	**$21,283**
3	cuanns_multigpu	NVidia	8x NVIDIA A100 GPUs	$-1,566,812.00	$151,009.85	$150,824.13	$150,816.00
4	cuanns_ivfpq	NVidia	1x NVIDIA A100 GPU	$-1,258,211.00	$303,929.39	$304,166.48	$153,155.12
5	gemini	GSI Technology(org)	LedaE APU	$-483,030.00	$569,058.09	$569,210.35	$398,163.18
6	faiss_t3	Facebook Research(org)	1x NVIDIA V100 GPU	0 (baseline)	$545,633.16	$737,886.17	$735,942.66

Table 1: Cost Leaderboard (*smaller values are better)

Throughput Rankings

Rank	Submission	Team	Hardware	Score*	DEEP1B	BigANN	SPACEV1B
1	cuanns_multigpu	Nvidia	NVidia GPU	2,377,864	801,694	747,421	839,749
2	optanne_graphann	Intel	Intel Optane	679,365	196,546	335,991	157,828
3	**BANG**	**-**	**Nvidia GPU**	**299,986**	**86,956**	**117,647**	**106,383**
4	cuanns_ivfpq	NVidia	NVidia GPU	269,112	91,701	80,109	108,302
5	gemini	GSI Technology(org)	LedaE APU	26,798	10,704	10,672	16,422
6	diskann	Microsoft Research India(org)	Dell PowerEdge	27,524	12,927	19,094	6,503
7	faiss_t3	Facebook Research(org)	NVidia GPU	0 (baseline)	4,464	3,271	3,265

Table 2: Throughput Leaderboard (*larger values are better)

Power Rankings

Rank	Submission	Team	Hardware	Score*	DEEP1B	BigANN	SPACEV1B
1	cuanns_multigpu	NVidia	NVidia GPU	-0.4137	0.0029	0.0024	0.0023
2	optanne_graphann	Intel	Intel Optane	-0.4101	0.0041	0.0022	0.0049
3	**BANG**	**-**	**Nvidia GPU**	**-0.4023**	**0.0081**	**0.0043**	**0.0065**
4	cuanns_ivfpq	NVidia	NVidia GPU	-0.3892	0.0112	0.0119	0.009
5	gemini	GSI Technology(org)	LedaE APU	-0.3305	0.0337	0.0341	0.023
6	faiss_t3	Facebook Research(org)	NVidia GPU	0 (baseline)	0.1117	0.1576	0.152

Table 3: Power Leaderboard (*smaller values are better)

2 Detailed Cost Computation

The overall cost computation comprises two metrics: (i) CapEx, representing the total cost to construct the ANN search system operating at the base recall-QPS, and (ii) OpEx, reflecting the cost of sustained operation over an extended period (4 years). OpEx is computed based on the system's power usage, and the detailed steps and formulae are outlined in the NeurIPS 2021 Challenge [2].

Sections 2.1, 2.2 and 2.3 present the detailed cost computation for the three datasets separately. For collecting the power metrics, the leaderboard provides framework tools [4] (using IPMI). However, we could not configure this tool to run on our system and thus we used an alternative method (explained futher) exposed by the platform to collate power usage data. Our hardware platform exposes overall system power usage metrics via the system management GUI. The GUI collects these metrics from the IPMI sensors in the backend. So, we used the power metrics reported from the GUI and plugged those values in the formula indicated in [2].

2.1 SIFT1B Dataset

Sl No	Item	Value	Units	Notes
i	Server with 2x Intel Xeon Gold 6326	5,62,000	Rs	Cost incurred to build the basic system (CPU, RAM, Chassis etc) and minus 18% GST (tax)
ii	Nvidia A100 GPU card	8,00,000	Rs	Single card
iii	Additional RAMs	1,57,500	Rs	64 GB RAM, 9 Nos, Each costing Rs 17,500
iv	Grand Total (cost to build one system)	15,19,500	Rs	(i) + (ii) + (iii)
v	Conversion to dollars	18,993.75	$	using 1$ = 80Rs
vi	Baseline recall	0.9	10-recall@10	
vii	Our algorithm's QPS at baseline recall	117,647	QPS	
viii	Additional GPU cards needed to scale to 100,000 QPS	0	-	
ix	Final CapEx (per leaderboard formula)	**18,994**	$	(v) + (viii) × (ii) / 80

Table 4: BANG: CapEx Computation for SIFT1B dataset. (QPS = Queries Per Second)

Sl No	Item	Value	Units	Notes
x	watt-seconds per query	4.31E-3	Ws	508 W used for 85 ms to run 10,000 queries
xi	kilowatt-hour per query	1.19E-9	kWh	(x) × 0.001/3600
xii	Scale to 4 years (delivering the baseline recall and target QPS)	1,513.02	$	Multiply (xi) by (4 × 365 × 24 × 60 × 60 × 100000) × 0.10$/kW-h
xiii	Final OpEx (per leaderboard formula)	**1,513**	$	

Table 5: BANG: OpEx Computation for SIFT1B dataset. (Ws = watt-seconds, kWh = kilowatt-hour)

Total Cost = (ix) + (xiii) = **$20,507**

2.2 DEEP1B Dataset

Sl No	Item	Value	Units	Notes
i	Server with 2x Intel Xeon Gold 6326	5,62,000	Rs	Cost incurred to build the basic system (CPU, RAM, Chassis etc) and minus 18% GST (tax)
ii	Nvidia A100 GPU card	8,00,000	Rs	MSRP is approx 10000$
iii	Additional RAMs	1,57,500	Rs	64 GB RAM, 9 Nos, Each costing Rs 17,500
iv	Grand Total (cost to build one system)	15,19,500	Rs	(i) + (ii) + (iii)
v	Conversion to dollars	18,993.75	$	
vi	Baseline recall	0.9	10-recall@10	
vii	Our algorithm's QPS at baseline recall	86,956	QPS	
viii	Additional GPU cards needed to scale to 100,000 QPS	1	-	(v) + (ii)
ix	Final CapEx (per leaderboard formula)	**28,994**	$	(v) + (viii) × (ii) / 80

Table 6: BANG: CapEx Computation for DEEP1B dataset. (QPS = Queries Per Second)

Sl No	Item	Value	Units	Notes
x	watt-seconds per query	8.18E-3	Ws	712 W used for 115 ms to run 10,000 queries
xi	kilowatt-hour per query Scale to 4 years	2.27E-9	kWh	(x) × 0.001/3600
xii	(delivering the baseline recall and target QPS)	2,869.07	$	Multiply (xi) by $(4 \times 365 \times 24 \times 60 \times 60 \times 100000) \times 0.10$/kW-h
xiii	Final OpEx (per leaderboard formula)	**2,869**	$	

Table 7: BANG: OpEx Computation for DEEP1B dataset. (Ws = watt-seconds, kWh = kilowatt-hour)

Total Cost = (ix) + (xiii) = **$31,863**

2.3 SPACEV1B Dataset

Sl No	Item	Value	Units	Notes
i	Server with 2x Intel Xeon Gold 6326	5,62,000	Rs	Cost incurred to build the basic system (CPU, RAM, Chassis etc) and minus 18% GST (tax)
ii	Nvidia A100 GPU card	8,00,000	Rs	
iii	Additional RAMs	1,57,500	Rs	64 GB RAM, 9 Nos, Each costing Rs 17,500
iv	Grand Total cost to build one system)	15,19,500	Rs	
v	Conversion to dollars	18,993.75	$	
vi	Baseline recall	0.9	10-recall@10	
vii	Our algorithm's QPS at baseline recall	106,383	QPS	
viii	Additional GPU cards needed to scale to 100,000 QPS	0	-	
ix	Final CapEx (per leaderboard formula)	**18,994**	$	(v) + (viii) × (ii) / 80

Table 8: BANG: CapEx Computation for SPACEV1B dataset. (QPS = Queries Per Second)

Sl No	Item	Value	Units	Notes
x	watt-seconds per query	6.53E-3	Ws	695 W used for 94 ms to run 10,000 queries
xi	kilowatt-hour per query	1.81E-9	kWh	(x) × 0.001/3600
xii	Scale to 4 years (delivering the baseline recall and target QPS)	2,289.16	$	Multiply (xi) by (4 × 365 × 24 × 60 × 60 × 100000) × 0.10$/kW-h
xiii	Final OpEx (per leaderboard formula)	**2,289**	$	

Table 9: BANG: OpEx Computation for SPACEV1B dataset. (Ws = watt-seconds, kWh = kilowatt-hour)

Total Cost = (ix) + (xiii) = **$21,283**

References

[1] H. V. Simhadri, "Billion-scale approximate nearest neighbor search challenge," https://big-ann-benchmarks.com/neurips21.html, 2024, accessed: July 18, 2024.

[2] E. Bernhardsson, "Neurips 2021 competition: Billion-scale ann - t3 track cost leaderboard," https://github.com/harsha-simhadri/big-ann-benchmarks/tree/main/neurips21/t3#cost_leaderboard, 2023, accessed: Dec 30, 2023.

[3] ——, "Neurips 2021 competition: Billion-scale ann - t3 track cost rankings," https://github.com/harsha-simhadri/big-ann-benchmarks/blob/main/neurips21/t3/LEADERBOARDS_PUBLIC.md, 2023, accessed: Dec 30, 2023.

[4] C. G. Williams, "Ipmicap : Power monitoring tool," https://github.com/fractalsproject/ipmicap, 2023, accessed: Dec 30, 2023.