BANG Cost Analysis

1 Cost Analysis

Creating ANNS systems with high QPS and recall is crucial, yet the associated hardware and operational costs are equally significant in practical applications. Therefore, when evaluating different ANNS algorithms, it is essential to consider the cost factor alongside QPS and recall metrics.

We adhere to the cost analysis methodology specified in the NeurIPS 2021 Challenge Leaderboard [1], which incorporates cost metrics alongside traditional QPS and recall metrics. The model normalizes the baseline by estimating the hardware needed to scale ANNS implementations to achieve a minimum recall QPS of 0.9 at 100,000 QPS. The cost analysis entails two metrics: CapEx, representing the total cost to construct the ANNS system operating at the base recall-QPS, and 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 formulas are outlined in the NeurIPS 2021 Challenge leaderboard [1]. For collecting the power metrics, the leaderboard provides framework tools. Since we are running into configuration issues with the tools, we used alternative methods to collate power usage data. The 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 in those values in the formula indicated in [1].

Table 1 shows the CapEx and OpEx metrics and their values. The CapEx and OpEx costs for BANG on the SIFT1B dataset are determined to be **18,994** USD and **124** USD, respectively. As a result, BANG is ranked in the top three on the Cost Ranking Leaderboard.

2 Detailed Cost Computation

		Cost/Units		Notes
CapEx	Intel 2-Socket Server	5,62,000	Rs	Cost incurred to build the basic system (CPU, RAM, Chassis etc) and minus 18% GST.
	Nvidia A100 GPU card	8,00,000	Rs	MSRP is approx 10000$
	Additional RAMs	1,57,500	Rs	64 GB RAM, 9 Nos, Each costing Rs 17,500
	Grand Total	15,19,500	Rs	This is the cost to build one system
	Conversion to dollars	18993.75	$	
	Baseline recall@10	0.9		
	Our algorithm's QPS at baseline recall	1,20,000		
	No of systems needed to scale to 100,000 QPS	1		
	(i) Final CapEx (per leaderboard formula)	**18,994**	$	
OpEx				
	Ws/q	0.000354167		500 W used for 85 ms to run 120000 queries
	kWh/query	9.83796E-11		convert from Ws/q
	Scale to 4 years	124.1	$	Multiply above kWh by $(4 \times 365 \times 24 \times 60 \times 60 \times 100000) \times 0.10$/kWh
	(ii) Final OpEx (per leaderboard formula)	**124**	$	
Total Cost	(i) + (ii)	**19,118**	$	

Table 1: BANG: CapEx and OpEx Computation

References

[1] E. Bernhardsson, "T3 track public dataset leaderboards," https://github.com/harsha-simhadri/big-ann-benchmarks/tree/main/neurips21/t3#cost_leaderboard, 2023, accessed: Dec 30, 2023.