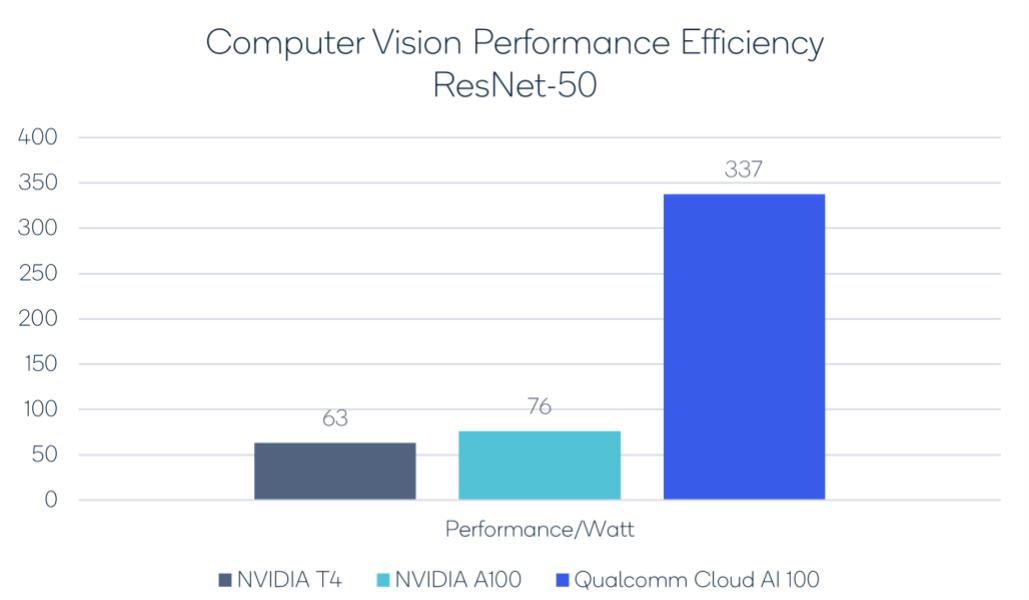
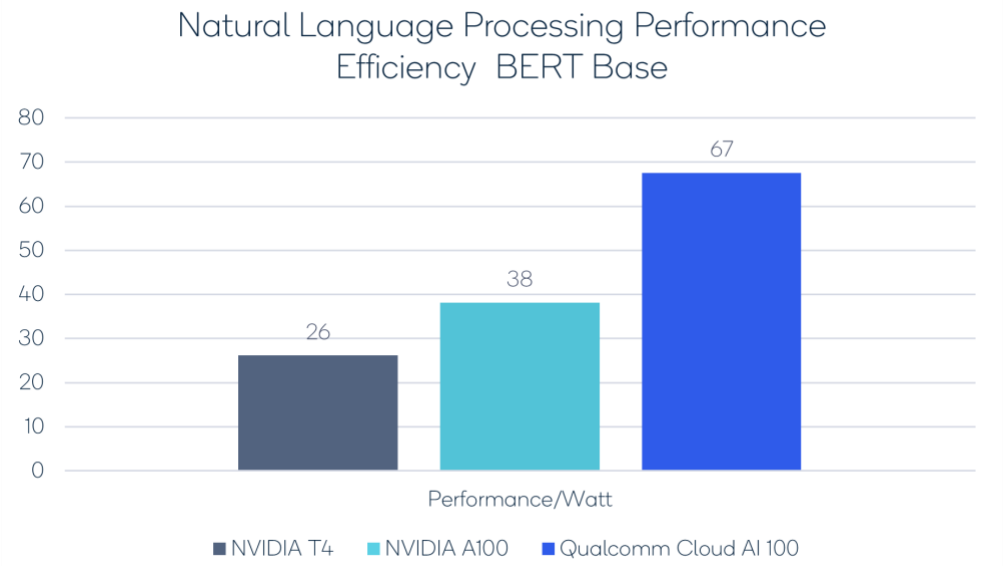
THE QUALCOMM CLOUD AI 100 DELIVERS SIGNIFICANT TCO SAVINGS FOR INFERENCE PROCESSING

Qualcomm Cloud AI 100 is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

INTRODUCTION

As AI (Artificial Intelligence) becomes ubiquitous, inference will surpass training processing in terms of IT (Information Technology) spending. Solution providers are addressing the growing demand for large AI inference workloads. Qualcomm Technologies, Inc. has been designing and producing AI hardware and software for over a decade and is expanding from mobile processors to the data center market with the Qualcomm® Cloud AI 100 platform, a purpose-built solution for accelerating inference workloads in cloud and edge infrastructure. This AI accelerator was adopted by HPE for inclusion in the company's server products. With the most recent MLPerf™ 2.1 benchmark results, Qualcomm Technologies sets a higher bar for power-efficient inference processing with the Qualcomm Cloud AI 100, delivering the highest Performance/Watt. These advantages stem from Qualcomm Technologies' superior performance at the low 75-watt power envelope.





Source: Power Efficiency Data from Respective websites Sept 2022
Nvidia: https://developer.nvidia.com/deep-learning-performance-training-inference
Qualcomm: https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/Cloud-AI-100-Inference-performance.pdf

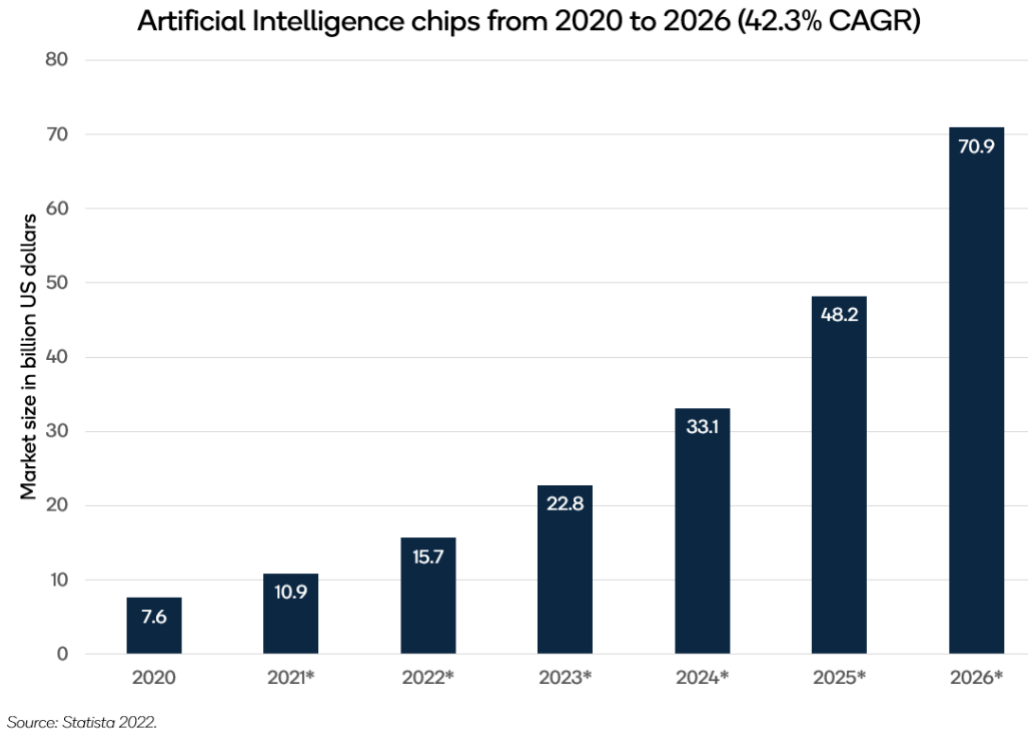


Figure 1: Per © Statista 2022, the artificial intelligence chip market is expected to reach 70.9 billion dollars by 2026 with a compound annual growth rate (CAGR) of approximately 42.3 percent. With highest Performance/Watt Qualcomm Cloud AI 100 continues to be highest stake holder to this growth.

The Total Cost of Ownership (TCO) is also driven by the cost of energy. We analyzed the power efficiency across benchmarks and accelerators. The Qualcomm Cloud AI 100 is significantly more power efficient than any other AI inference accelerators currently available to deploy. Qualcomm Technologies led the entire field for data center power efficiency which triggered our interest in trying to assess TCO resulting in potential cost and power savings.

Given the criticality of AI to a plethora of industries, the added power consumed by AI processing impacts cloud edge, enterprise AI and large data centers alike, where power costs will be critical for wide adoption of the technology. This research paper assesses the potential TCO savings of the Qualcomm Cloud AI 100 for inference processing. In addition to reducing operating expenses, our analysis highlights how the Qualcomm Cloud AI 100 also benefits companies aiming to reduce their carbon footprint.

EFFICIENCY COMPARISONS

Calculating and comparing energy consumption can be a complex endeavor. Should one measure the power consumed at the chip, the server, the rack, the row, or even the entire datacenter? Finding quality power consumption and performance metrics can be a challenge, although published MLPerf™ benchmarks and vendors' websites can help. For this study, we

made comparisons at the accelerator, server, and rack level. Moreover, we normalized the equipment config variability by setting targets to deliver equivalent number of inferences per second at each level. Let us start by looking at the hardware and the benchmarks that were used in our assessment.

BENCHMARKS AND SERVERS USED IN THIS ANALYSIS

We used ResNet-50, BERT Base and Large benchmarks to capture two of the most used AI applications: computer vision and natural language processing. We compared the widely deployed NVIDIA T4 GPU and a recent NVIDIA A100 GPU versus Qualcomm Cloud AI 100 in high-volume standard servers.

All performance and power data were leveraged and sited from the respective company websites in September 2022. We then modeled the number of accelerator cards, servers, and racks needed to deliver equivalent inference processing capacity at scale and calculated the potential savings in power and infrastructure. For this research paper, we modeled the infrastructure needed to process one hundred million inferences per second (IPS) for Natural Language Processing and one billion inferences per second for Computer Vision. This helps to scale the analysis from accelerators cards to servers to rack level. Before arriving to conclusions, several factors were made customizable such as the type of servers used, the number of required inferences per second, electricity rates, etc.

A note on the GPU's we selected is in order. The NVIDIA T4 is an N-1 generation GPU, but its affordability and performance have made it a popular choice still to this day for demanding inference applications with significant market traction. The TDP (Thermal Design Power) Power of the T4 (70W) has a similar power range of Qualcomm Cloud AI 100 (75W). The T4 and the Qualcomm Cloud AI 100 both fit in 2U height servers, and thus we use the same servers as a base for a fair comparison. The only changes are performance and power consumed by each accelerator to deliver the same number of inferences at accelerator card, server, and rack level.

In addition, we also chose to compare Qualcomm Cloud AI 100 with a current generation high performing NVIDIA A100 which is shipping today. Although the TDP Power of NVIDIA A100 GPU is higher, it also delivers cutting edge high performance needed for the newest AI use cases. For the NVIDIA A100, we used 4U height Gigabyte servers used in MLPerf™ submissions.

To keep the comparison focused on power consumed by AI accelerators for AI compute, we used the same 300W theoretical power consumed by the host for both NVIDIA and Qualcomm Cloud AI 100 powered servers.

Furthermore, while our analysis is specific to NVIDIA, we believe the conclusions would not substantially change if other industry platforms were incorporated including offerings from Intel Habana and Google TPU for example. Most have yet to publish the data needed for this analysis, and the Qualcomm Cloud AI 100 power efficiency is demonstrably better than any vendor submitting into MLPerf™ benchmarks. To be clear, NVIDIA T4 and A100 are quite power-efficient compared to most alternatives, except for the Qualcomm Cloud AI 100. This should not surprise anyone since the Qualcomm Cloud AI 100 architecture share ~~a~~ legacy with the Snapdragon® processors designed for mobile handsets. However, the magnitude of potential savings when deployed at scale may surprise many.

THE METHODOLOGY

We modelled the number of accelerators, servers and racks needed to achieve one hundred million inference per second for natural language processing and one billion inference per second for image processing based on company website published performance and efficiency. Once we knew how many accelerators, server or racks were used to deliver requested inference per second, we then calculated the infrastructure and power costs amortized to annual costs considering equipment refresh cycle of 4 years.

THE RESULTS

Based on our analysis, the cost of equipment and power usage incurred by deploying Qualcomm Cloud AI 100 across a cloud or edge data center for the same amount of inference work is substantially lower than alternatives considered.

COMPUTER VISION RESNET-50 - QUALCOMM CLOUD AI 100 VERSUS NVIDIA A100 AND T4 GPUS

- Qualcomm Cloud AI 100 Total Cost of Ownership (TCO) savings versus the NVIDIA A100 and NVIDIA T4 runs over 80% at Accelerator, Server, and Rack level for ResNet-50 inferences. This translates TCO Efficiency > 6x for accelerator card and > 5x for server and rack level

- The AI inference solutions powered by Qualcomm Cloud AI 100 saves up to 101 Giga Watt Hours of electricity to deliver one billion computer vision ResNet-50 inferences per second at the rack level versus the NVIDIA A100 GPU AI solution

- The AI inference solutions powered by Qualcomm Cloud AI 100 saves up to 154 Giga Watt Hours of electricity to deliver one billion computer vision ResNet-50 inferences per second at the rack level versus the NVIDIA T4 GPU AI solution

- Put into US dollar terms, for one billion inference per second these **savings go** greater than **$198M for A100 and $193M for T4** to run computer vision ResNet-50 inferences at

Rack Level annually in infrastructure and energy costs. Below you can see the excerpts for the rack-level analyses for Qualcomm Cloud AI 100 vs the NVIDIA A100 for Computer Vision ResNet-50 Network.

TCO Comparison for ML Inference - Datasource Respective AI Accelerator Website

Competetive Configuration	TCO Analysis Summary AI Rack Level	Qualcomm Configuration
G492-ID0-Intel-Xeon 8380-40core, 8x NVIDIA A100	ResNet50V1.5-INT8	Gigabyte G292-Z43, 16x Qualcomm Cloud AI 100
NVIDIA A100		Qualcomm Cloud AI 100
$248 Million	Total Cost per year for 1,000 Million inf/Sec	$49 Million

Power and TCO Cost Savings

Power Saving Per year	101 GigaWh	TCO Cost saving per year	$199 Million
TCO Cost Efficiency	5.05 X	Percent Savings	80%

TCO Analysis Details

	TCO Inputs	Competetive Config		Base Config	
		NVIDIA A100		Qualcomm Cloud AI 100	
	14.4KW 42U rack	G292-Z43-AMD-Milan-DP-7713(64-cores)		Gigabyte G292-Z43, 16x Qualcomm Cloud AI 100	
Single Server Performance					
Use Drop down below to pick network		*Throughput*	*Actual Power*	*Throughput*	*Actual Power*
ResNet50v1.5-INT8	<-Select Network	235,952 inf/Sec	3,405 W	356,304 inf/Sec	1,356 W
Rack Level TCO Summary					
Datacenter scenario	Lease	Retail lease electricity cost based on 100% rack utilization at TDP			
Performance requirement	1,000,000,000 inf/sec	Enter Target Performance to be achieved in terms of Total Inference per Sec for given Network			
Acquisition Cost					
Number of servers			4239		2807
Number of racks			2120		468
Equipment refresh cycle	4 yrs				
Amortized equipment acquisition cost per year			$190,398,055		$35,758,556
OpEx Power Cost					
Electric utility rate	$0.10 /KWh				
PUE	1.2				
Server utilization	90%				
Power consumption per year			136,540,472 KWh		36,010,567 KWh
Electricity cost per year			$13,654,047		$3,601,057
OpEx DC Cost					
Cost of power to rack	$120.00 /KW/mo		$43,960,320		$9,704,448
Cost of rack space	$1,700.00 (30sf)		$0		$0
DC cost per year			$43,960,320		$9,704,448
Total Power cost per year			$57,614,367		$13,305,505
Total Cost per year			$248,012,422		$49,064,060

Figure 2: TCO Analysis for image processing, Qualcomm Cloud AI 100 vs NVIDIA A100.

Source: Bibliography - Analysis Data source 1, 2, 4

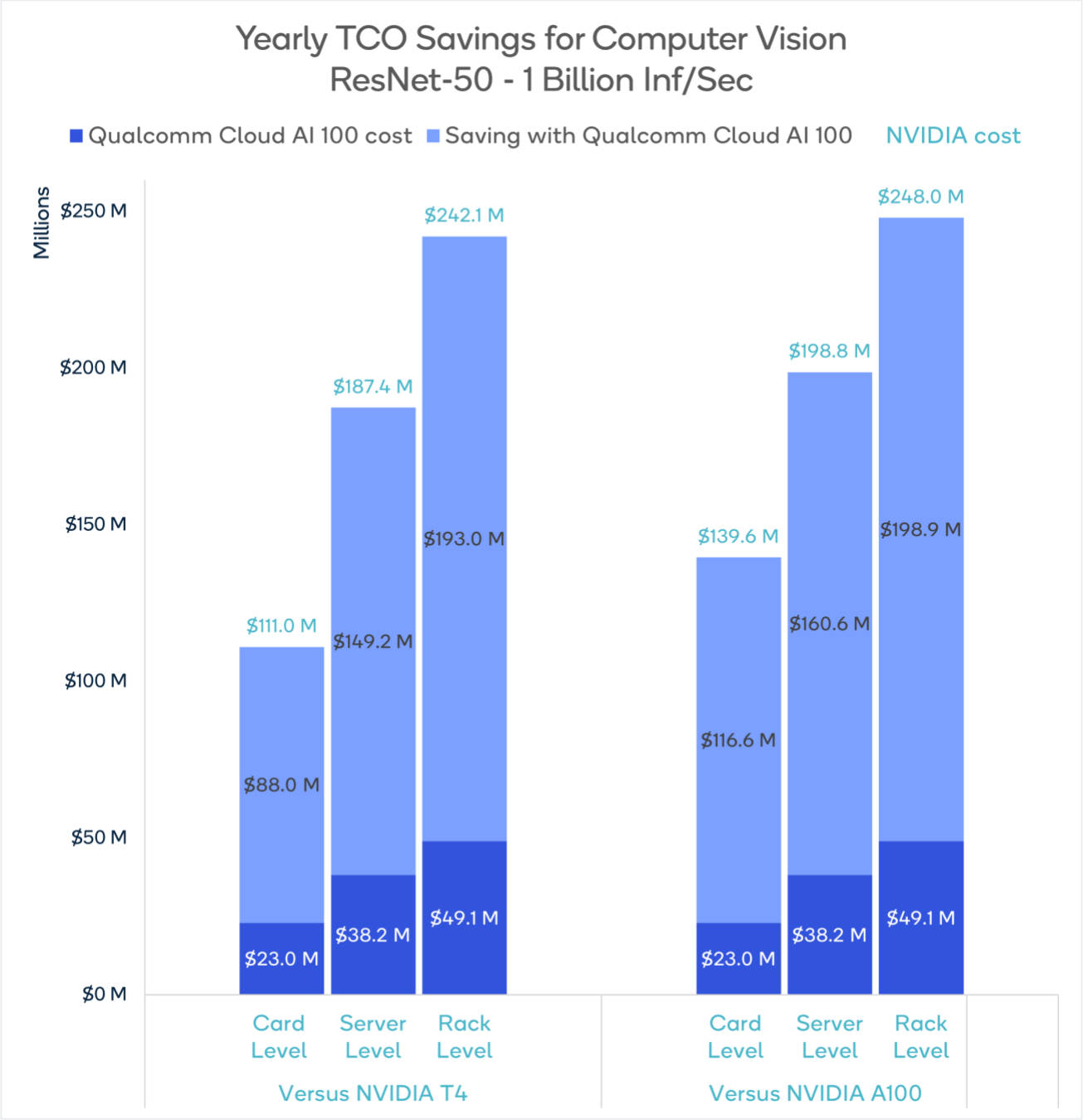


Figure 3: TCO Cost savings with Qualcomm Cloud AI 100 vs NVIDIA T4 and NVIDIA A100 to deliver one billion Inference per second at accelerator, server, and rack level for Computer Vision Networks

Source: Bibliography - Analysis Data source 1, 2, 3, 4

NATURAL LANGUAGE PROCESSING BERT LARGE - QUALCOMM CLOUD AI 100 VERSUS NVIDIA A100 AND NVIDIA T4 GPU

- Qualcomm Cloud AI 100 Total Cost of Ownership (TCO) savings versus the NVIDIA A100 and NVIDIA T4 runs over 60% at Accelerator, Server, and Rack level for BERT Large - 128 inferences. This translates TCO Efficiency > 2.5x for NVIDIA T4 and 1.9x for NVIDIA A100 at accelerator card, server, and rack level

- The AI inference solutions powered by Qualcomm Cloud AI 100 saves up to 23 Giga Watt Hours of electricity to deliver one hundred million natural language BERT Large 128 inferences per second at the rack level versus the NVIDIA A100 GPU AI solution

- The AI inference solutions powered by Qualcomm Cloud AI 100 saves up to 102 Giga Watt Hours of electricity to deliver one hundred million natural language BERT Large 128 inferences per second at the rack level versus the NVIDIA T4 GPU AI solution

- Put into US dollar terms, for one hundred million inference per second these **savings** are greater than **$121M for T4 and $67M for A100** to run Natural Language Processing BERT-Large 128 inferences at Rack Level annually for infrastructure and energy costs. Below you can see the summary of analyses for Qualcomm Cloud AI 100 vs the NVIDIA A100 and NVIDIA T4 for Natural Language Processing BERT Large

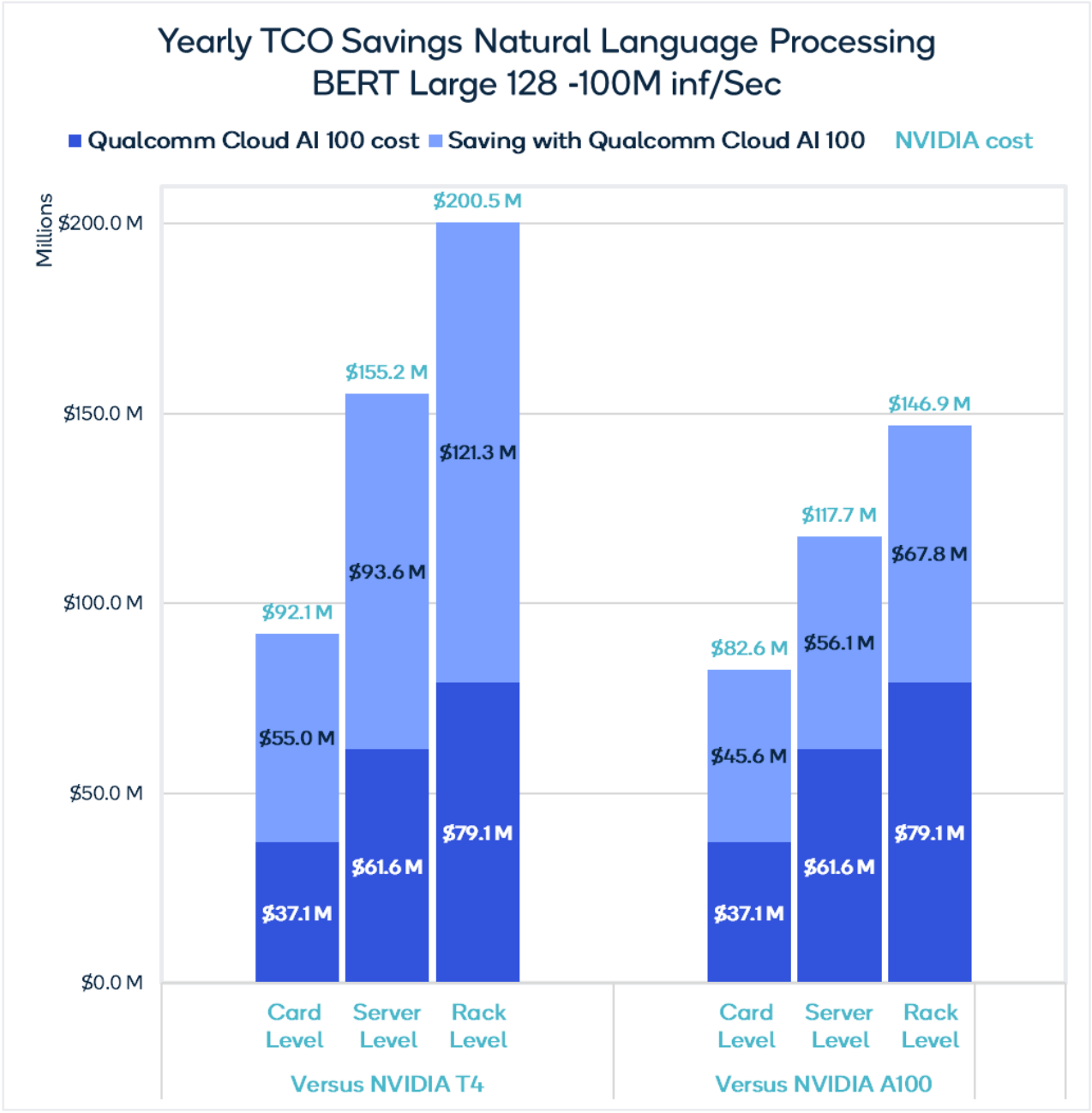


Figure 4: TCO Cost savings with Qualcomm Cloud AI 100 vs NVIDIA A100 and NVIDIA T4 to deliver one hundred million Inference per second at Accelerator, Server, and Rack level for Natural Language Processing BERT Large Seq Len 128.

Source: Bibliography - Analysis Data source 1, 2, 3, 4

NATURAL LANGUAGE PROCESSING BERT BASE - QUALCOMM CLOUD AI 100 VERSUS NVIDIA A100 AND NVIDIA T4 GPU

- Qualcomm Cloud AI 100 Total Cost of Ownership (TCO) savings versus the NVIDIA T4 runs over 57% and versus NVIDIA A100 50% at Accelerator, Server, and Rack level for BERT Base-128 inferences. This translates TCO Efficiency > 2.3x for NVIDIA T4 and 1.9x for NVIDIA A100 at accelerator card, server, and rack level

- The AI inference solutions powered by Qualcomm Cloud AI 100 saves up to 9 Giga Watt Hours of electricity to deliver one hundred million natural language BERT Base 128 inferences per second at the rack level versus the NVIDIA A100 GPU AI solution

- The AI inference solutions powered by Qualcomm Cloud AI 100 saves up to 28 Giga Watt Hours of electricity to deliver one hundred million natural language BERT Base 128 inferences per second at the rack level versus the NVIDIA T4 GPU AI solution

- Put into US dollar terms, for one hundred million inference per second these **savings** are more than **$34M for T4 and $23M for A100** to run Natural Language Processing BERT Base-128 inferences at Rack Level annually for infrastructure and energy costs. Below you can see the summary of analyses for Qualcomm Cloud AI 100 vs the NVIDIA A100 and NVIDIA T4 for Natural Language Processing BERT Base –128.

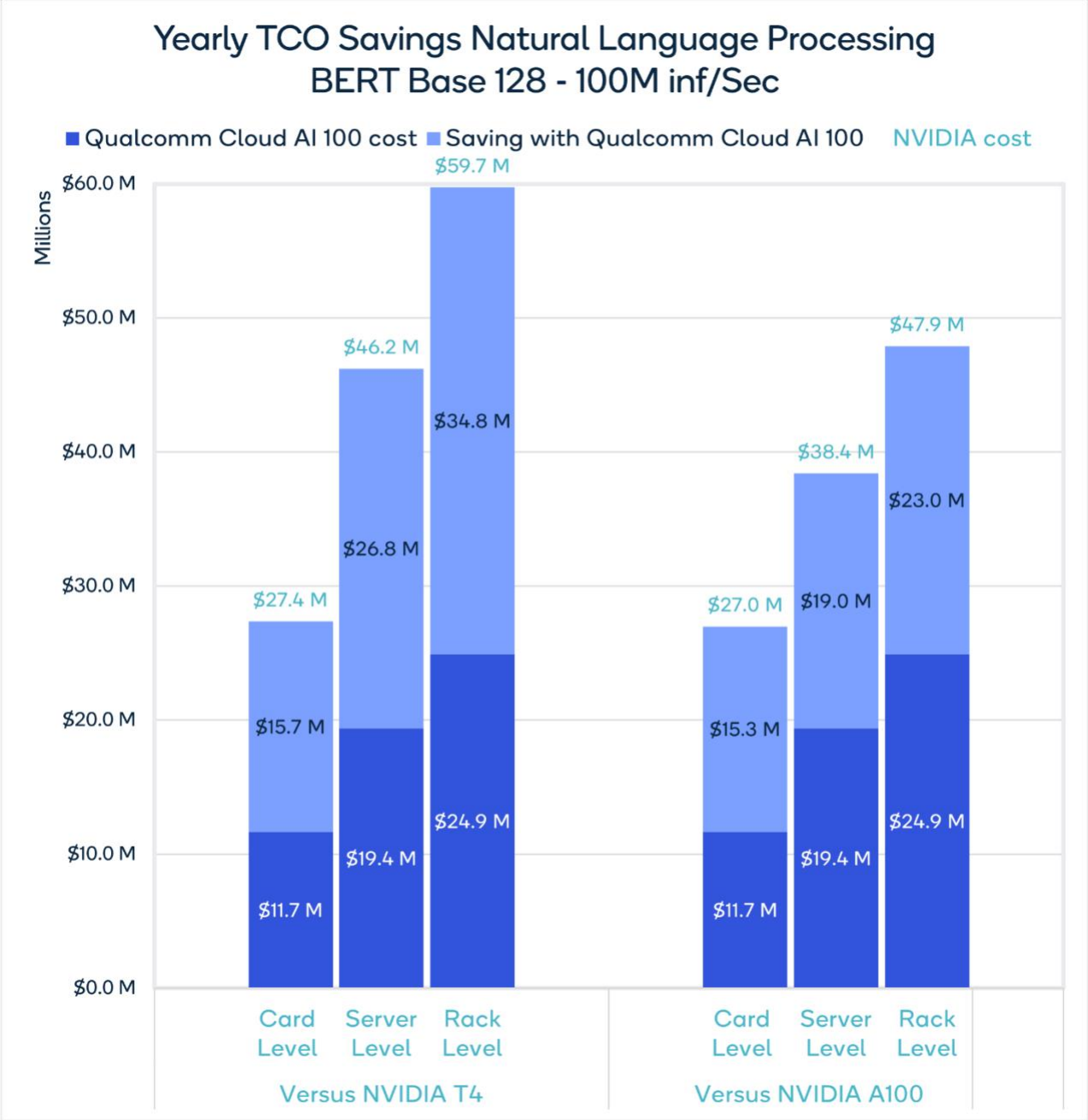


Figure 5: TCO Cost savings with Qualcomm Cloud AI 100 vs NVIDIA A100 and NVIDIA T4 to deliver one hundred million Inference per second at accelerator, server, and rack level for Natural Language Processing BERT Base Seq Len 128.

Source: Bibliography - Analysis Data source 1, 2, 3, 4

Below is the summary of Qualcomm Cloud AI 100 TCO cost versus NVIDIA A100 and NVIDIA T4 GPUS at accelerator level.

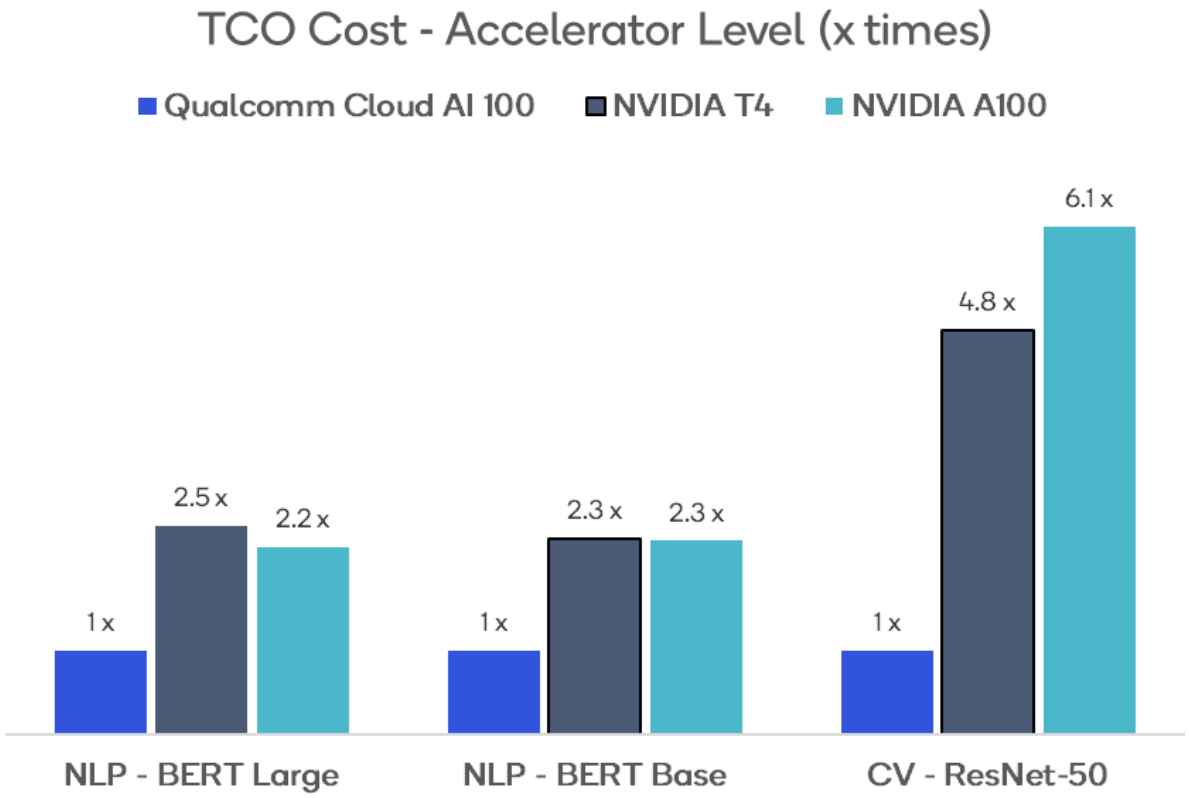


Figure 6: TCO Cost comparison with Qualcomm Cloud AI 100 vs NVIDIA A100 and NVIDIA T4 to deliver one hundred million Inference per second for Natural Language processing (BERT Large and BERT Base) and one billion inference per second for Computer Vision (ResNet-50) at accelerator level

Source: Bibliography - Analysis Data source 1, 2, 3, 4

We also assessed the savings across a wide range of GPU pricing, as firm GPU price data for volume purchases is not readily available and found that NVIDIA GPU and other equipment like servers and racks would have to be priced at 20% over the discounted retail website prices to deliver the same Perf/TCO as Qualcomm Cloud AI 100 solutions at list price.

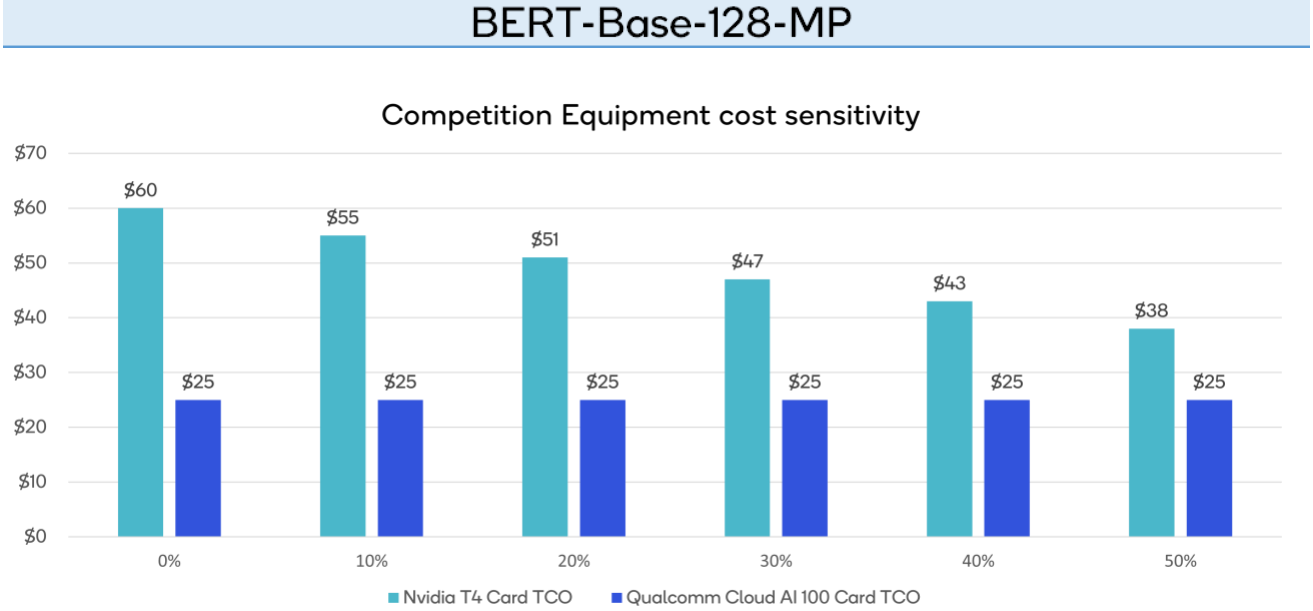


Figure 7: The TCO savings swamp the cost of the equipment, even at a 50% discount

A Few Important Caveats

Note that these analyses do not assess other alternatives that also may merit consideration for large-scale inference deployments. Significantly, the Intel Xeon Scalable Processors are already widely installed in data centers and new SKUs have AI-specific acceleration called DL-Boost. Unfortunately, we cannot think of a way to model the TCO of these CPUs for AI since they are also used simultaneously for other workloads. Also, we do not yet have any performance and power metrics for the NVIDIA Hopper (H100) platforms recently introduced, so the H100 has not been included. Finally, the AMD (Advanced Micro Devices) Instinct GPU line is not included as the company has yet to publish inference benchmark results as well.

We also note that many startups are beginning to market inference processors, however, for most of these public benchmarks are scarce.

CONCLUSIONS

Qualcomm Cloud AI 100 is setting a new standard for power efficiency for AI inference acceleration for cloud and edge deployments. Our analysis shows that the Qualcomm Cloud AI 100 platform offers significant energy and capital savings over current GPU-based accelerators for large-scale computer vision and natural language processing that can lead to annual savings of hundreds of millions of dollars. The resulting reduction in carbon footprint is also a significant step forward for organizations towards meeting their sustainability goals.

Based on the analyses we have done for Qualcomm Cloud AI 100 accelerators, Computer Vision workloads at the rack level can save up to **Two Hundred million** US dollars yearly, while saving more than **One Hundred and fifty Giga Watt hours** of power every year to deliver one billion inferences per second. Similarly, Natural Language Processing workloads at the rack level can save up to **One Hundred Twenty million** US dollars yearly, while saving more than **one hundred Giga Watt hours** of power every year to deliver one hundred million inferences per second.

BIBLIOGRAPHY - ANALYSIS DATA SOURCE

1. Nvidia Inference performance and Power Efficiency for T4 and A100 GPU: https://developer.nvidia.com/deep-learning-performance-training-inference > inference tab > Sub-Tab inference - Dated July 2022

2. Qualcomm Inference performance and Power efficiency for Cloud AI 100: https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/Cloud-AI-100-Inference-performance.pdf - Dated Nov 2022

3. Nvidia T4 GPU pricing: https://www.thinkmate.com/product/nvidia/900-2g183-0000-001 , https://www.newegg.com/p/1FT-0004-006X0 Dated July 2022

4. Gigabyte AI Server including 8x A100 GPU: G492-ID0 Pricing: https://www.itcreations.com/configurator/model/g492id0/722

5. AI Chip Market Size 2020 – 2026: https://www.statista.com/statistics/1283358/artificial-intelligence-chip-market-size/ - July 2022

APPENDIX A – TCO ANALYSIS SUMMARY VERSUS NVIDIA T4

TCO Comparison for ML Inference - Datasource Respective AI Accelerator Website				
Competetive Configuration		**TCO Analysis Summary AI Rack Level**		**Qualcomm Configuration**
G292-Z43-AMD-Milan-DP-7713(64-cores)		BERT-Large-128-MP		G292-Z43-AMD-Milan-DP-7713(64-cores)
Nvidia T4 x 16				Qualcomm Cloud AI 100 PCIe Pro x 16
$200 Million		Total Cost per year for 100 Million inf/Sec		$79 Million
Power and TCO Cost Savings				
Power Saving Per year	102 GigaWh		TCO Cost saving per year	$121 Million
TCO Cost Efficiency	2.53 X		Percent Savings	61%

TCO Comparison for ML Inference - Datasource Respective AI Accelerator Website				
Competetive Configuration		**TCO Analysis Summary AI Rack Level**		**Qualcomm Configuration**
G292-Z43-AMD-Milan-DP-7713(64-cores)		BERT-Large-128-MP		G292-Z43-AMD-Milan-DP-7713(64-cores)
Nvidia T4 x 16				Qualcomm Cloud AI 100 PCIe Pro x 16
$60 Million		Total Cost per year for 100 Million inf/Sec		$25 Million
Power and TCO Cost Savings				
Power Saving Per year	28 GigaWh		TCO Cost saving per year	$35 Million
TCO Cost Efficiency	2.40 X		Percent Savings	58%

TCO Comparison for ML Inference - Datasource Respective AI Accelerator Website				
Competetive Configuration		**TCO Analysis Summary AI Rack Level**		**Qualcomm Configuration**
G292-Z43-AMD-Milan-DP-7713(64-cores)		ResNet50V1.5-INT8		G292-Z43-AMD-Milan-DP-7713(64-cores)
Nvidia T4 x 16				Qualcomm Cloud AI 100 PCIe Pro x 16
$242 Million		Total Cost per year for 100 Million inf/Sec		$49 Million
Power and TCO Cost Savings				
Power Saving Per year	154 GigaWh		TCO Cost saving per year	$193 Million
TCO Cost Efficiency	4.93 X		Percent Savings	80%

APPENDIX B – TCO ANALYSIS SUMMARY VERSUS NVIDIA A100

TCO Comparison for ML Inference - Datasource Respective AI Accelerator Website				
Competetive Configuration		TCO Analysis Summary		Qualcomm Configuration
G492-ID0-Intel-Xeon 8380-40core, 8x NVIDIA A100		AI Rack Level		Gigabyte G292-Z43, 16x Qualcomm Cloud AI 100
NVIDIA A100		BERT-Base-128-Mixed		Qualcomm Cloud AI 100
$147 Million		Total Cost per year for 100 Million inf/Sec		$79 Million
Power and TCO Cost Savings				
Power Saving Per year	23 GigaWh		TCO Cost saving per year	$68 Million
TCO Cost Efficiency	1.86 X		Percent Savings	46%

TCO Comparison for ML Inference - Datasource Respective AI Accelerator Website				
Competetive Configuration		TCO Analysis Summary		Qualcomm Configuration
G492-ID0-Intel-Xeon 8380-40core, 8x NVIDIA A100		AI Rack Level		Gigabyte G292-Z43, 16x Qualcomm Cloud AI 100
NVIDIA A100		BERT-Base-128-Mixed		Qualcomm Cloud AI 100
$48 Million		Total Cost per year for 100 Million inf/Sec		$25 Million
Power and TCO Cost Savings				
Power Saving Per year	9 GigaWh		TCO Cost saving per year	$23 Million
TCO Cost Efficiency	1.92 X		Percent Savings	48%

TCO Comparison for ML Inference - Datasource Respective AI Accelerator Website				
Competetive Configuration		TCO Analysis Summary		Qualcomm Configuration
G492-ID0-Intel-Xeon 8380-40core, 8x NVIDIA A100		AI Rack Level		Gigabyte G292-Z43, 16x Qualcomm Cloud AI 100
NVIDIA A100		ResNet50V1.5-INT8		Qualcomm Cloud AI 100
$248 Million		Total Cost per year for 1,000 Million inf/Sec		$49 Million
Power and TCO Cost Savings				
Power Saving Per year	101 GigaWh		TCO Cost saving per year	$199 Million
TCO Cost Efficiency	5.05 X		Percent Savings	80%

IMPORTANT INFORMATION ABOUT THIS PAPER