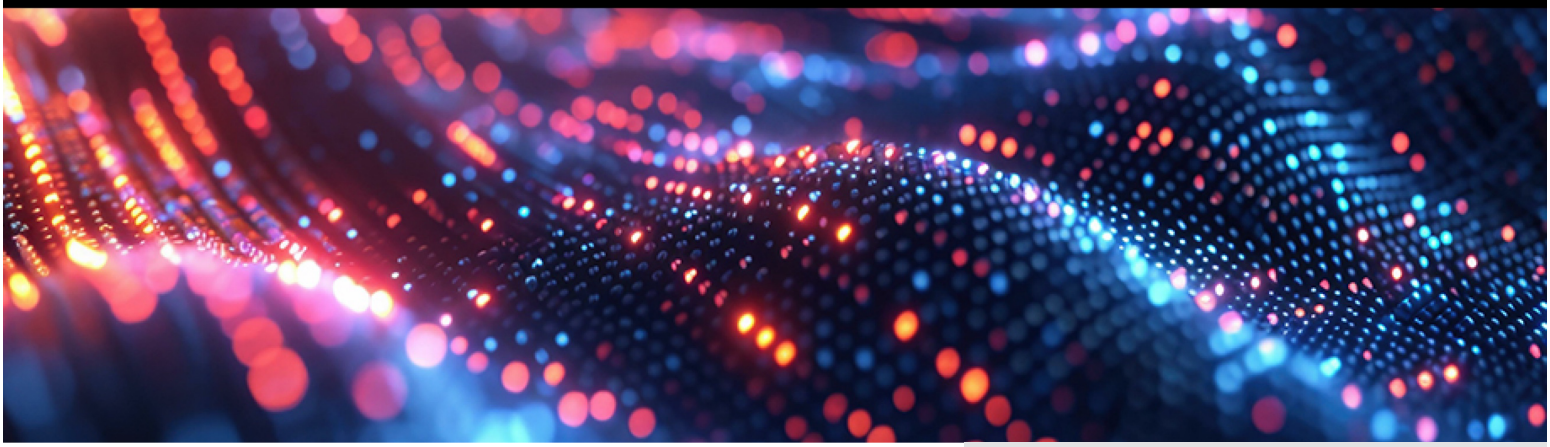

Accelerating large language model inference with HBM3E

The rapid growth of large language models (LLMs) like GPT and Llama in recent months has placed greater demands on hardware. These models are powerful AI systems designed to process and generate humanlike text, and they require significant memory and compute resources to effectively manage vast amounts of data, especially for inference. And as AI models evolve, becoming more capable and intelligent, the demands placed on memory and compute will only grow. The recently released high-bandwidth memory 3 enhanced (HBM3E), the latest generation of high-performance memory technology to provide ample bandwidth and capacity for AI tasks, offers many benefits that enhance the performance of LLMs. Specifically, HBM3E at the cube level provides[2]:

- **>1.2 TB/s** memory bandwidth, a 1.4 times increase over the prior HBM generation
- **16GB to 24GB** capacity**,** a nearly 50% increase over the prior generation

The faster data transfer speeds and expanded memory capacity help realize the full potential of GPUs for AI inference tasks. For AI researchers, developers and technology leaders looking to understand the latest advancements in GPU memory performance, this technical brief provides a comprehensive analysis of LLM inference performance comparing NVIDIA HGX H100 [1] with HBM3 and HGX H200 [2] with HBM3E.

Our results show that the expanded capacity and higher bandwidth of HBM3E can deliver up to a **1.8** times performance improvement for LLM inference, offering enormous potential for more efficient and powerful AI computing.

1. Llama 2 70B in FP16 precision requires 140GB of HBM3 capacity to hold the model weights. H200 has six HBM sites while H100 has five HBM sites. At the cube level, there is a 50% increase in capacity and up to 80% capacity increase at the system level.
2. HBM3E bandwidth and capacity numbers calculated per a single H200 GPU.

Key takeaways

1.8x

The higher memory bandwidth and capacity of HBM3E in an NVIDIA H200 system (4.8 TB/s) increases inference performance of Llama 2 70B (INT4 quantized model execution) by 1.8 times.

~2.5x

HBM3E enables 2.5 times more batch sizes (inference requests) than the previous HBM generation, supporting more concurrent clients for a single GPU.

~80%

An expanded system memory capacity with HBM3E at 144GB (24GB per cube) enables higher precision (FP16) processing, overcoming memory limitations of previous GPU generations. This result allows the Llama 2 70B model to run at full precision.[1]

Authors: Felippe Vieira Zacarias, Evelyn Grevelink, Sudharshan Vazhkudai, and Kiran Palli

Milliseconds matter

In the data center, this performance boost may allow providers to serve more users by responding to more inference requests. Consider a cloud service provider (CSP) that can now handle 2.5 times more concurrent queries without adding more servers. This means CSPs could serve more leading AI-powered chatbot requests (Anthropic's Claude Sonnet 3.5, OpenAI's ChatGPT, etc.), including translation services or complex data analysis requests, without additional hardware having to be purchased. The financial implications are substantial — instead of investing in new server infrastructure, these providers can use advanced memory technologies like HBM3E to scale their services more efficiently.

Beyond the economic benefits, these improvements to performance fundamentally reshape user interaction with technology. Consider a person using a translation app and having to wait several seconds for the text to convert. That delay can be frustrating, and that's why these performance gains matter so much. In the case of real-time language translation, every millisecond counts. The difference between a continuous conversation and an interrupted one can be just a fraction of a second. Faster inference means a translation is provided almost instantly, creating a user experience that feels as natural as talking to someone.

Memory requirements for LLMs

The memory requirements for LLMs like Llama 2 [3] are enormous, in gigabytes (GB) and terabytes (TB). The memory needed depends on many things like **sequence length** and **batch size** in addition to the model definition, also known as hyper-parameters. Table 1 shows the increasing approximate memory needs as the parameters increase into the billions and the possible solutions for varying model sizes.

Approximate Model memory requirements

Model	Parameters	Approximate model memory requirement[3] (inference) FP16	Cost-effective solution
Llama 2-7B	7B	14GB	HBM3, HBM3E
Llama 2-13B	13B	26GB	HBM3, HBM3E
Llama 2-70B	70B	140GB	HBM3E

Table 1: Approximate memory needed by inference for various Llama model parameters

3. Additional memory required for KV cache during inference.

Hardware and software specifications

To analyze the performance of Llama 2 70B with DeepSpeed[4] ZeRO-Inference [4], we tested and validated a single NVIDIA HGX H100/H200 box manufactured by Supermicro. Tables 2 and 3 provide the key configuration details of the test system, outlining the GPU and host specifications, respectively. Note that "HGX" is used as shorthand in the sections that follow. Also, the input sequence length (ISL) and output sequence length (OSL) used during the tests are defined for each result.

Both H100 and H200 are based on the Hopper architecture. However, the NVIDIA HGX H200 is an upgraded version of H100 that includes key improvements to memory capacity and bandwidth, making it more effective at handling AI workloads.

Note: The test systems in Figures 1 and 2 represent the memory perspective, meaning they are simplified and may not accurately depict the entire configuration.

4. DeepSpeed ZeRO-Inference is part of the ZeRO collection, known for their powerful memory and parallelism optimizations that enable efficient large-scale model training and inference across multiple GPUs [5][6][7]. For Llama 2 70B (70 billion parameters), ZeRO-Inference enables efficient inference by splitting the model across available GPUs, reducing per-device memory requirements and allowing complex models to run on hardware with limited memory.

NVIDIA H100

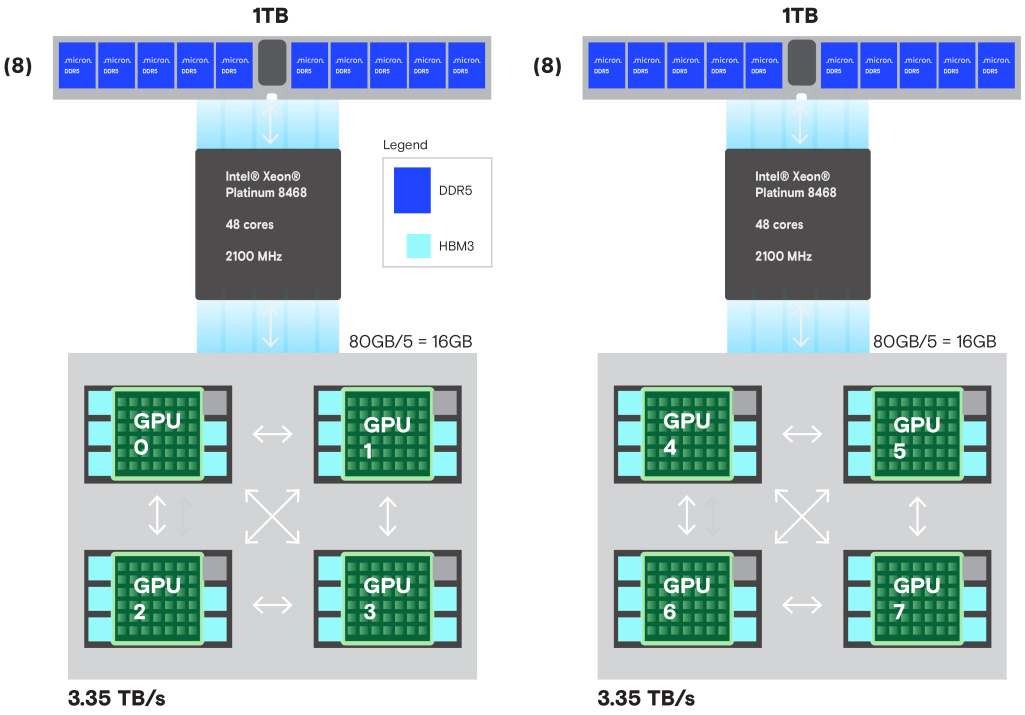

Figure 1. NVIDIA HGX H100 (baseline system)

NVIDIA H200

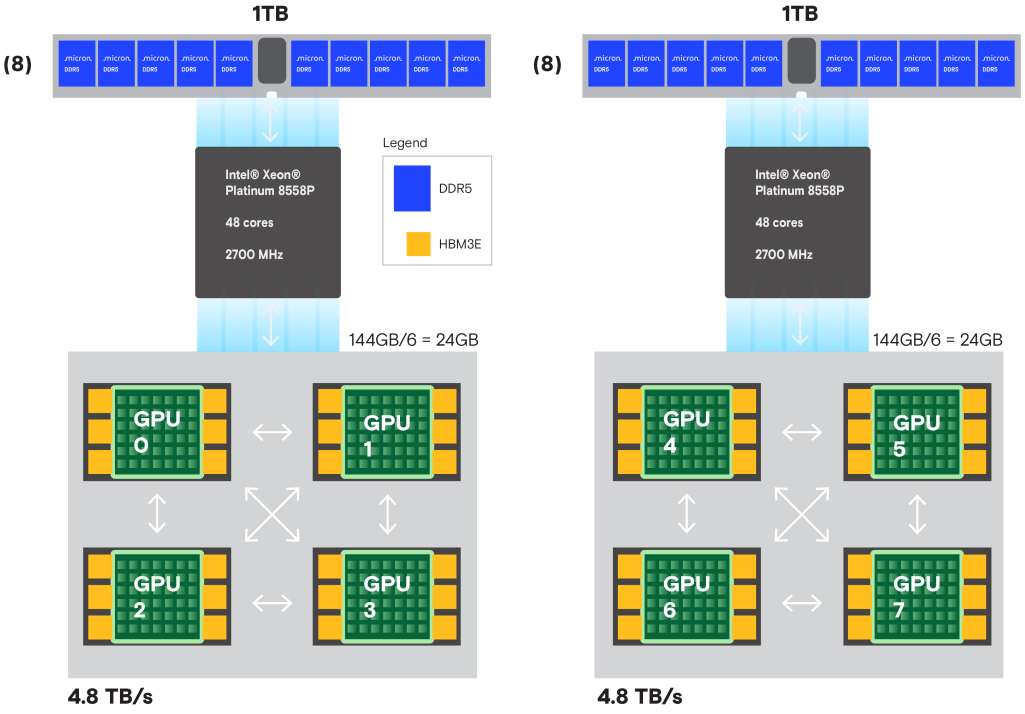

Figure 2. NVIDIA HGX H200 system

SUT host configuration		
Model	**HGX H100**	**HGX H200**
Model	Intel® Xeon® Platinum 8468	Intel® Xeon® Platinum 8558P
Core(s) per socket	48	48
Socket(s)	2	2
CPU max speed	2100 MHz	2700 MHz
L3 cache	105 MB	260 MB
Memory type	32x Micron 64GB DDR5 RDIMMs	32x Micron 64GB DDR5 RDIMMs
Total memory capacity	2TB	2TB
Memory speed	4400 MT/s	4400 MT/s
DIMMs/channel	2 DIMMs per channel / 16 channels	2 DIMMs per channel / 16 channels
Operating system	Ubuntu 20.04	Ubuntu 20.04

Table 2. Host CPU configuration of the system under test

Key differences between H100 and H200 used in our experiment			
	H100	**H200**	**Note**
Architecture	Hopper	Hopper	Both use same architecture
Memory type	HBM3	HBM3E	HBM3E is faster and more efficient
HBM placements	5	6	
Memory capacity (GPU)	80GB	**144GB** (24GB per cube)	76% increase in capacity
Memory bandwidth (GPU)	3.35 TB/s	**4.8 TB/s** (800 GB/s per cube based on system specifications)	43% higher bandwidth. HBM3E can reach >1.2 TB/s per cube.
		Same compute power	

Table 3. Comparison of the NVIDIA HGX H100 and H200 systems

Improving performance with faster bandwidth

Fast data transfer rates are important for LLMs like Llama 2, which must rapidly access and process enormous amounts of data during both training and inference. For example, a model must quickly fetch and process model weights and sequences at rates of hundreds of gigabytes per second and then generate and output the requested text just as quickly. Slower data transfer speeds can affect user experience, limiting a model's ability to operate at peak performance and deliver responses within quality of service (QoS) requirements.

The H200 system uses HBM3E, which has 43% higher bandwidth than the H100 system. The higher bandwidth of HBM3E in H200 (4.8 TB/s) allows for faster data transfer and therefore higher throughput inference. This improved bandwidth is particularly important for real-time applications, like chatbots or virtual assistants, where the model needs to provide instant responses to user queries. The higher memory bandwidth allows the GPU to more rapidly fetch and process the vast amounts of data required by LLMs, reducing latency and improving overall throughput. In practical terms, this means many things for LLMs — faster retrieval of model parameters, reduced waiting times when generating responses and more seamless handling of compute tasks. Consider the customer service chatbot mentioned earlier. With higher memory bandwidth, the chatbot can now understand context more quickly and provide more rapid responses.

Inference throughput (bandwidth only)

Throughput (tokens per second) with same batch size. Higher is better.

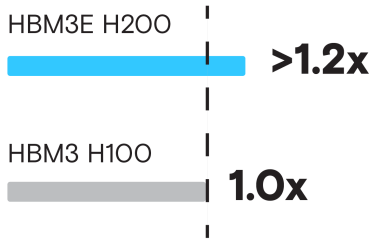

Figure 3. Throughput increase over previous generation of HBM (HBM3 with H100) using the same batch size[5]

Using HBM3E, inference throughput performance increases by **1.2 times** due to the higher bandwidth of HBM3E in H200 over the previous generation HBM3.

Note: System bandwidth for HBM3E varies based on system design.

5. Llama 2 70B: ISL 512, OSL 768 | H100 SXM 1x GPU HBM3 BS 1 | H200 SXM 1x GPU HBM3E BS 1. INT4 quantization. Memory utilization efficiency and, as a result, performance varies with type of workloads.

Further enhancing inference performance with capacity

While the initial performance boost of 1.2x (as shown in the figure above) demonstrated the critical role of bandwidth in improving LLM inference, the true value of HBM3E emerges when we couple higher **bandwidth** with expanded memory **capacity**. The combination of these two further enhances inference throughput. A 1.2x performance gain from higher bandwidth is just the foundation—the real breakthrough comes when this faster data transfer is combined with the ability to store and instantly access larger model parameters and context-aware data structures. The result is a remarkable **1.8x** performance improvement that further establishes the value of HBM3E.

Specifically, LLM models require substantial memory *capacity* to store their heavy-duty parameter sets, which can reach into the billions or even trillions of parameters. In addition to the higher bandwidth of HBM3E, an expanded capacity of **144GB** (24GB per cube) allows the GPU to store larger model parameters, associated data structures and key-value caches, enabling the successful deployment of more complex and capable LLMs. Thus, the inference performance is further improved. To fully realize the enormous potential of LLMs and deliver higher throughput at the lowest total cost of ownership (TCO), continued advancements in memory technologies are necessary to address these computational challenges. The H200 system uses HBM3E, which has 76% higher GPU memory capacity in addition to the 43% higher bandwidth (discussed earlier) than the H100 system.

Batch size and inference requests

Increase in inference requests due to increase in batch size. Higher is better.

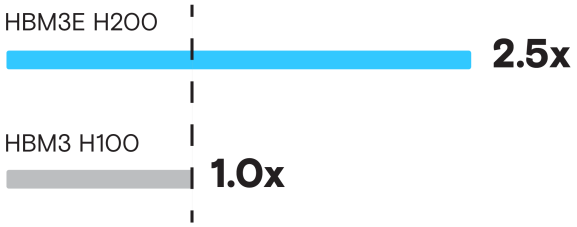

Figure 4. Increase in inference requests over previous generation of HBM[6]

Using HBM3E, we can process 2.5 times more inference requests (which technically means larger batch size) than the previous generation of HBM (Figure 4). This means the GPU is now able to compute more data in parallel, effectively doubling the number of users supported at the same time.

Think of it like upgrading from a small storage facility to a large warehouse, moving more "goods" in the form of data in the same amount of time. The practical result is simple — more clients can be served faster on the same hardware. By enabling the server to increase the number of concurrent clients per GPU, we can significantly reduce the overall cost of running AI services at scale.

Inference throughput (bandwidth and capacity)

Increase in throughput (tokens per second) due to increase in batch size. Higher is better.

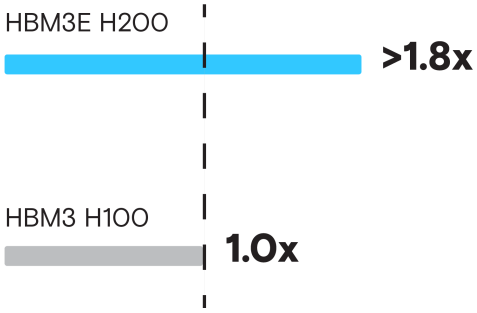

Figure 5. Throughput increase over previous generation of HBM[6]

The combination of increased memory capacity and higher memory bandwidth from HBM3E (H200) provides a **1.8x** increase in inference throughput over the previous generation HBM3 (H100) (Figure 5).

Note: Bandwidth alone (without factoring in capacity) provides 1.2x increase in throughput as shown in our earlier experiment.

An 80% increase in system memory capacity for HBM3E also lets the GPUs run larger models at higher precision, like FP16, using fewer systems. This capability was not possible with the previous generation of HBM.

Note: In our experiment, we used systems with HBM3 16GB and HBM3E 24GB capacities.

6. Llama 2 70B: ISL 512, OSL 512 | H100 SXM 1x GPU HBM3 BS 48 | H200 SXM 1x GPU HBM3E BS 120. INT4 quantization. Throughput is defined as total tokens per second (tokens/s).

Improving TCO from inference at lower power

HBM is a powerful solution for enabling data centers to effectively face the computational demands of running AI inference at scale. Traditional server memory architectures require significant power infrastructure and cooling systems to maintain performance. HBM3E changes this paradigm by reducing power requirements while maintaining — and often improving — computational throughput.

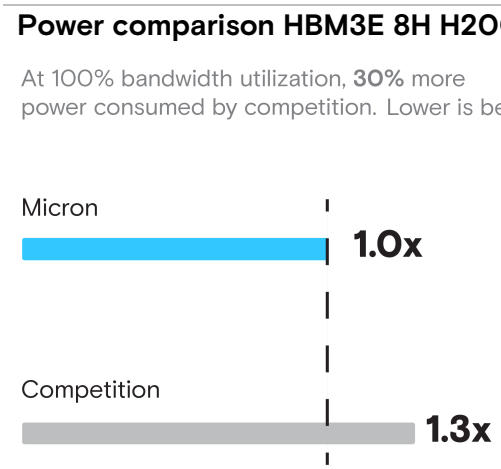

Figure 6. Lower power use of Micron HBM3E 8H

To understand the power savings in a real-world scenario, we ran inference on Llama 2 70B, which demonstrated 1.8x better throughput[7]. In this earlier experiment we were able to achieve 50% memory bandwidth utilization, which resulted in the competitor HBM3E consuming 20% more power than Micron's solution.

To further stress the memory bandwidth usage (beyond what was achieved in the earlier experiment), we use BabelStream [8], a microbenchmark which shows us the highest memory bandwidth usage possible. Our results using BabelStream show up to **30%** power consumption at 100% utilization.

Power savings become more pronounced as bandwidth utilization approaches 100%, reinforcing the value of higher bandwidth for HBM3E.

7. Llama 2 70B: ISL 512, OSL 512 | H100 SXM 1x GPU HBM3 BS 48 | H200 SXM 1x GPU HBM3E BS 120. INT4 quantization.

These power savings also translate directly to operational benefits, as shown in the table below. Micron HBM3E contributes to a 4% reduction in data center total cost of ownership (TCO), primarily through operational expense (OpEx) savings. Importantly, the amount of savings scales proportionally with the number of accelerators and HBM placements per accelerator.

Metric	Competition HBM3E 8H	Micron HBM3E 8H
HBM cube power (W)	1.3X	1.0X
Rack power (KW)	1.00X	0.96X
Data center OpEx (TCO) savings[8]	-	4%

Table 4. OpEx (TCO) savings with Micron HBM3E 8H

8. TCO calculation methodology: Rack power of 40 kW total, with 40 accelerators per rack and 6 HBM placements per accelerator. Comparison based on differential HBM power consumption between Micron and competitor HBM3E solutions.

Conclusion

The benefits of HBM3E, as we've shown in our empirical data, are tangible and substantial, setting a higher standard for enhanced user experience. Moreover, the ability to significantly improve inference performance (tokens per second), support larger models at higher precision like FP16, and deliver substantial total cost of ownership reductions positions HBM3E as a desirable solution for GPU acceleration in data centers. Looking ahead, organizations choosing HBM3E are not just upgrading hardware; they are smartly investing in their AI infrastructure. As AI models will become increasingly sophisticated and data-intensive, the performance, scalability and efficiency of memory technologies like HBM3E will be paramount. Data centers that embrace these technologies will be better positioned to deliver faster, more sustainable, user-focused, and more cost-effective AI services, ultimately advancing progress across industries.

Learn more

For more information on high-bandwidth memory (HBM) technology, check out our HBM3E product page.

References

[1] NVIDIA. (2025). NVIDIA H100 Tensor Core GPU. https://www.nvidia.com/en-us/data-center/h100/

[2] NVIDIA. (2025). NVIDIA H200 Tensor Core GPU. https://www.nvidia.com/en-us/data-center/h200/

[3] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, A., Bashlykov, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv. https://doi.org/10.48550/arXiv.2307.09288

[4] DeepSpeed. (2022, September 9). ZeRO-Inference: Enabling efficient inference of large language models at unprecedented scale. https://www.deepspeed.ai/2022/09/09/zero-inference.html

[5] Aminabadi, R. Y., Rajbhandari, S., Awan, A. A., Li, C., Li, D., Zheng, E., Ruwase, O., et al. (2022). DeepSpeed-inference: Enabling efficient inference of transformer models at unprecedented scale. In SC22: International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-15). IEEE.

[6] Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2020). Zero: Memory optimizations toward training trillion parameter models. In SC20: International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-16). IEEE.

[7] Rasley, J., Rajbhandari, S., Ruwase, O., & He, Y. (2020). DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 3505-3506).

[8] Deakin T, Price J, Martineau M, McIntosh-Smith S. Evaluating attainable memory bandwidth of parallel programming models via BabelStream. International Journal of Computational Science and Engineering. Special issue. Vol. 17, No. 3, pp. 247–262. 2018. DOI: 10.1504/IJCSE.2018.095847

micron.com/HBM3E