**GDDR7: The Ideal Memory Solution in AI Inference**

The generative AI market is experiencing rapid growth, driven by the increasing parameter size of Large Language Models (LLMs). This growth is pushing the boundaries of performance requirements for training hardware within datacenters. For an in-depth look at this, consider the insights provided in "HBM3E: All About Bandwidth. Once trained, these models are deployed across a diverse range of applications. They are transforming sectors such as finance, meteorology, image and voice recognition, healthcare, augmented reality, high-speed trading, and industrial, to name just a few.

The critical process that utilizes these trained models is called AI inference. Inference is the capability of processing real-time data through a trained model to swiftly and effectively generate predictions that yield actionable outcomes. While the AI market has primarily focused on the requirements of training infrastructure, there is an anticipated shift towards prioritizing inference as these models are deployed.

The computational power and memory bandwidth required for inference are significantly lower than those needed for training. Inference engines typically need between 300-700GB/s of memory bandwidth, compared to 1-3TB/s for training. Additionally, the cost of inference needs to be lower, as these systems will be widely deployed not only in datacenters, but also at the network's edge (e.g. 5G) and in end-user equipment like security cameras, cell phones, and automobiles.

When designing an AI inference engine, there are several memory options to consider, including DDR, LPDDR, GDDR, and HBM. The choice depends on the specific application, bandwidth, and cost requirements. DDR and LPDDR offer good memory density, HBM provides the highest bandwidth but requires 2.5D packaging, and GDDR offers high bandwidth using standard packaging and PCB technology.

The GDDR7 standard, announced by JEDEC in March of this year, features a data rate of up to 192GB/s per device, a chip density of 32Gb, and the latest data integrity features. The high data rate is achieved by using PAM3 (Pulse Amplitude Modulation) with 3 levels (+1, 0, -1) to transmit 3 bits over 2 cycles, whereas the current GDDR6 generation uses NRZ (non-return-to-zero) to transmit 2 bits over 2 cycles.

GDDR7 offers many advantages for AI Inference having the best balance of bandwidth and cost. For example, an AI Inference system requiring 500GB/s memory bandwidth will need only 4 GDDR7 DRAM running at 32Gbp/s (32 data bits x 32Gbp/s per pin = 1024Gb/s per DRAM). The same system would use 13 LPDDR5X PHYs running at 9.6Gbp/s, which is currently the highest data rate available (32 data bits x 9.6Gb/s = 307Gb/s per DRAM).

Cadence stands at the forefront of AI inference hardware support, being the first IP company to roll out GDDR7 PHYs capable of impressive speeds up to 36Gb/s across various process nodes. This milestone builds on Cadence's established leadership in GDDR6 PHY IP, which has been available since 2019. The company caters to a diverse client base spanning AI inference, graphics, automotive, and networking equipment.



While GDDR7 continues to utilize standard PCB board technology, the increased signal speeds seen in GDDR6 (20Gbp/s) and now GDDR7 (36Gb/s) calls for careful attention with the physical design to ensure optimized system performance. In addition to providing the PHY, Cadence also offers comprehensive PCB and package reference design, which are essential in helping customers achieve optimal signal and power integrity (SI/PI) for their systems.

Cadence is dedicated to ensuring customer success beyond just providing hardware. They provide expert support in SI/PI, collaborating closely with customers throughout the design process. This approach ensures that customers can benefit from Cadence's expertise in navigating the complexities of high-speed design and achieving optimal performance in their AI inference systems.

As the AI market continues to advance, Cadence remains at the forefront by offering a comprehensive memory IP portfolio tailored for every segment of this dynamic market. From DDR5 and HBM3E, which cater to the intensive demands of training in servers and high-performance computing (HPC), to LPDDR5X designed for low-end inference at the network edge and in consumer devices, Cadence's offerings cover a wide range of applications.

Looking to the future, Cadence is dedicated to innovating at the forefront of memory system performance, ensuring that the evolving needs of AI training and inference are met with the highest standards of excellence. Whether it's pushing the boundaries with GDDR7 or exploring new technologies, Cadence is dedicated to driving the AI revolution forward, one breakthrough at a time.