三星电子推出zHBM技术，该内存可3D直接堆叠在AI芯片上方。在AI智能体时代，三星提出愿景：将现有AI系统的响应速度提升10倍以上。

当地时间9月16日，三星美洲半导体事业部（DSA）内存产品企划负责人Kim Indong，在美国加州圣克拉拉会展中心举办的AI基础设施峰会发表主旨演讲，阐述AI存储芯片领域的范式变革。

Kim Indong常务表示：“当前对话式AI系统，单用户响应速度上限约为每秒100 token。面向即将到来的AI智能体时代，我们目标实现跨越式突破，将速度提升至每秒1000 token。”

三星表示，想要实现这一巨大性能提升，核心突破口就是zHBM。

除此之外，三星电子还公布了面向端侧AI市场、基于NAND闪存的3D存储方案zNAND-O的技术路线图。

Kim Indong称：“到2030年，不只是大型服务器，在AI工作站等设备本地运行万亿参数超大模型将会普及。如果完全依靠DRAM搭建适配万亿参数模型的内存，成本极高；而采用zNAND-O方案，仅需六分之一的成本就能实现。”

三星计划自2028年起，向客户大规模提供zNAND-O样品。Kim Indong补充道：“AI芯片竞争的核心在于产品上市速度，我们有信心在此方面为客户提供独一无二的价值。”