

半導体漫遊記

湯之上隆

(340)

2022年11月に米オプティンAIがChatGPTを公開したことをきっかけとして、生成AI(人工知能)が世界中に爆発的に普及している。生成AIは、AI半導体を搭載したサーバー上で動作するが、AI半導体の80%を独占している米NVIDIAのGPU(画像プロセッサ)の供給が足りず、米クラウドメーカーなどの間で引っぱりだこの状態である。その結果、NVIDIAのGPUの最上位機種「H100」が1個4万ドル(約600万円)ものべら棒な価格で取引されている。

SKハイニックス(50%)、サムスン電子(40%)、米マイクロン(10%)となっており、現在のところSKハイニックスがシェア1位となっている。しかし3社合わせても、HBMの供給が不足している。その理由は次の通りである。

HBMは、GPUとともに高性能化しなければならぬ。具体的に言うところデータを転送する帯域幅の拡大、メモリ容量の増大、動作速度の向上を同時に実現

NVIDIAのGPUがブレイクした23年にはHBM2が主力となり、今年24年から来年25年にかけてはDRAM積層数を8〜12層に増加させたHBM3がそれに取って代わる。さらに26〜27年にはHBM3Eが主流となり、27年以降はDRAM積層数を16層に増大したHBM4が主役になると予測されている。

このようにGPUに搭載されるHBMは約2年で世代交代する。その結果、H

GPUの不足続く

HBMも、量産へ熾烈な争い

NVIDIAのGPUが不足している理由が二つある。一つはTSMCのパッケージのキャパシティが不足している問題である。NVIDIAのGPUは、TSMCが前工程も後工程も行って製造している。その後工程では、12インチシリコ

ンウエハを角型にカットしたシリコンインターポーザの上にGPU、CPU、メモリなどを配置することによりCoWoS(Chip on Wafer on Substrate)と呼ばれるパッケージを製造している。この後工程のキャ

パシティーが全く足りていない。

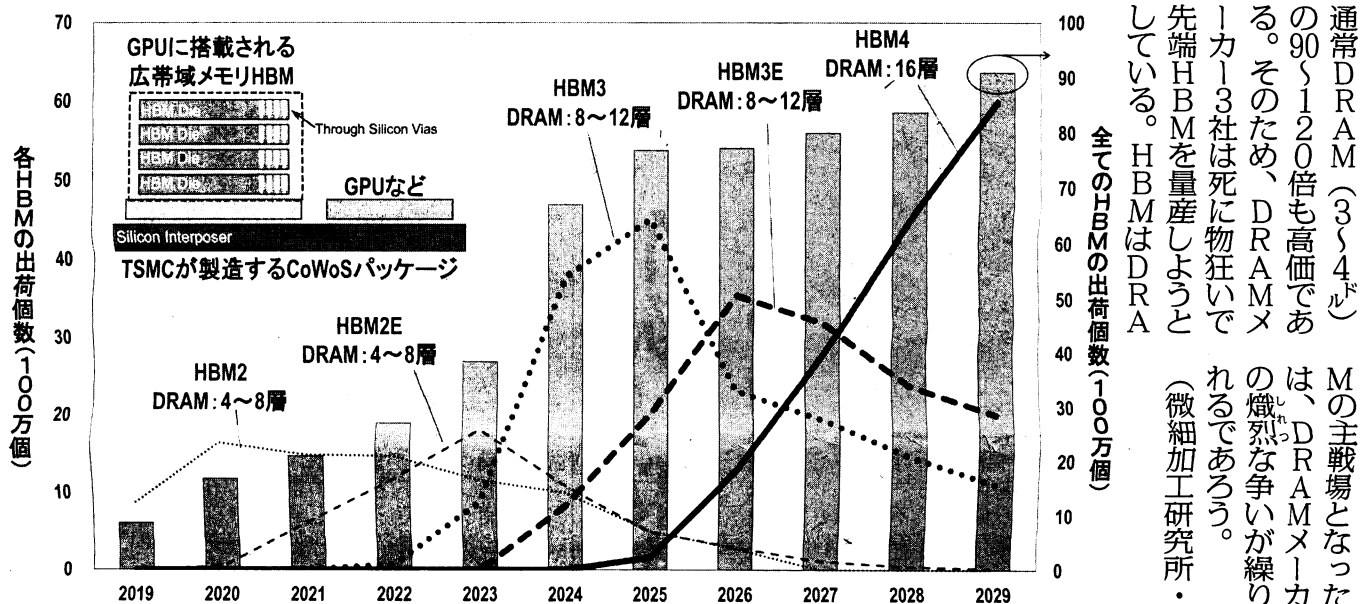
もう一つは、シリコンインターポーザ上にDRAMを積層した広帯域メモリ(High Bandwidth Memory、HBM)を多数搭載するが、このHBMも不足していることが分かってきた。

HBMを製造するメモリメーカー(と市場シェア)は、

しなればならない。そしてメモリ容量の増大のためにはHBM内に積層するDRAMチップ数を増やす必要があるし、動作速度の向上のためには微細化を推進しなければならぬ。

上記の事情を頭に入れて、各種HBMの出荷個数の推移を見てみよう。19〜22年はHBM2が最も多数つくられていた。ところが、

現在、先端のHBM3Eは361ドルもする。これは



各種HBMの出荷個数および全HBMの出荷個数

出所: Yole Intelligenceのデータを基に筆者作成

通常DRAM(3〜4層)の90〜120倍も高価である。そのため、DRAMメーカー3社は熾烈な争いが繰り広げられるであろう。(微細加工研究所・所長)