アナリストの眼

AI と GPU

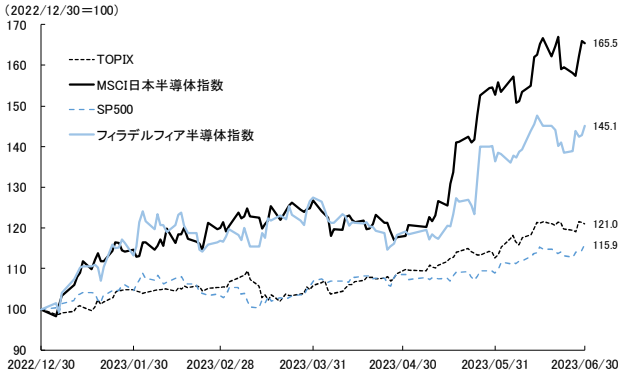
【ポイント】

1. 生成 AI が半導体関連銘柄の株価上昇を促す材料になっており、中でも契機となった決算を発表した GPU メーカーは非常に高いパフォーマンスを示している。
2. GPU 自体の性能のみならず、それを使用する環境作りなども相まって、圧倒的な競争力を保持、製品価格は非常に高価で、2 番手メーカーやハイパースケーラーの製品投入などもみられるが、勢力図が変わる可能性は低いだろう。
3. マイナス面として、高額な GPU 含む AI への投資が他分野への投資抑制につながっていること、また GPU においても特定プロセスの生産能力不足がボトルネックになる可能性が懸念される。

１．市場参加者予想の上限を上回るサプライズ決算

2023 年 5 月 24 日、ゲーミング PC やデータセンターなどに用いられる GPU（Graphics Processing Unit）を開発・販売する半導体ファブレス（製品製造のための自社工場を持たない企業）メーカー（以下、当該メーカー）の決算が公表され、今後の売上高予想は市場参加者の事前予想の上限をも上回る強気な内容であった。中でも生成 AI（人工知能）向けの GPU が寄与するとされ、日本市場でも生成 AI の恩恵を受けると考えられる銘柄への買いが見られた。日米の半導体関連指数は大きく上昇し、年初来で市場全体の指数を上回る状況が続いている（図表 1）。特に当該メーカーの年初来パフォーマンスは 7 月 13 日時点で＋200％を超え、米国の半導体指数であるフィラデルフィア半導体指数の構成銘柄 30 銘柄のうち、最も高いパフォーマンスを示している。GPU について、PC を含む市場全体のシェアで見ると、他の半導体メーカーのシェアが大きいものの、AI 含むサーバーのみに限定すると、当該メーカーのシェアが他を圧倒している。この圧倒的シェアを背景に、当該メーカーの GPU は非常に高額で販売されている。ここでは AI における GPU の競争環境について整理し、その継続性、高額な GPU が他分野へ及ぼす影響、GPU 市場の懸念点について見ていきたい。

図表１．日米半導体株式市場の動き



（資料）各種資料より富国生命投資顧問作成

２．なぜ GPU が必要なのか

AI において GPU が必須とされているのは、AI の処理速度が GPU に左右されるためである。GPU は処理速度を高めるアクセラレータという位置づけで使用される。そもそも GPU 自体は、ノートパソコンなどに CPU（Central Processing Unit、中央処理装置）と統合されて標準装備されるものであり、それ自体が珍しいものではない。一部ゲーミング PC などの高性能な画像処理を必要とするものには、単体の GPU を使用している

ものもある。更に AI サーバーなど極めて高い処理速度を要するものには、ゲーミング PC 向けとは別の製品を用いる。CPU と GPU の違いは、CPU が複雑な計算を分割して順番に処理する連続性において優位であるのに対し、GPU は単純な計算を並列で同時に処理できる点に利がある。その GPU の特性は、画像処理だけでなく、機械学習やディープラーニング、仮想通貨のマイニングなどに適しており、画像処理以外に用いられる GPU を、GPGPU（General Purpose GPU）と呼ぶ。現在、生成 AI の中でも最も有名な対話型 AI の公開時の言語モデルは、パラメータ[1]数が 1,750 億個と非常に大規模であり、CPU だけが高性能でも、処理速度は実用レベルに達せず、GPU が不可欠である。

３．競争力の源泉

現在、AI を含むデータセンター向け GPU の各社のシェアについて正確な数字は取れないものの、当該メーカーが 70%程度を占めると言われている。この高シェアは、AI 等のソフトウェア開発環境における地位と、その開発環境下で自社の GPU の性能がより活きる、という双方向からの強みによりもたらされている。当該メーカーは、2007 年に自社の GPU を利用して AI 等のソフトウェアを開発する環境の提供を始めた。同環境は複雑なグラフィックス言語を使わず、汎用的な C 言語や Python を使用し、より平易に開発できる環境が整備されており、AI 等の開発環境の業界標準となっている。当該メーカーの GPU はこの開発環境で最適な動作をするよう作られており、AI 等の開発者が当該メーカーの GPU を使おうという動きは必然である。

エンジニアリング・コンソーシアム（共同事業体）である MLcommons が公表する、機械学習用の性能ベンチマークテスト「MLperf」において、当該メーカーの存在感は圧倒的である。このテストには、画像処理や、生成 AI に使用される LLM（Large Language Model）[2]学習などの訓練項目を『クラスタ[3]がどのくらいの時間で処理できるのか』が示されており、全クラスタのうち、携帯デバイス（Mobile）や小型組込みデバイス（Tiny）の推論を除く分野では、当該メーカーの GPU を使用したクラスタが 7 割超、更にエッジ推論を除けば 80%半ばを占めた（図表 2）。また、上記 LLM の項目で性能結果を得られたクラスタは、当該メーカーともう 1 社のものしかない。クラスタ毎の GPU の数が異なるため単純な比較はできないものの、当該メーカーのクラスタは 9～10 倍の数の GPU を用いて、他社の 30 倍の速度で処理するとされる。

図表２．MLPerf における GPU 使用比率

カテゴリー	認証されたシステム数	うち、当該メーカーのチップが使われている割合
訓練（Training）	90	91%
訓練（Training）：HPC	9	89%
推論（Inference）：Datacenter	77	79%
推論（Inference）：Edge	52	46%
推論（Inference）：Mobile	2	0%
推論（Inference）：Tiny	32	0%

（資料）MLCommonsより富国生命投資顧問作成

このような環境下で、当該メーカーが提供する、AI サーバー向け GPU は、最新のもので 3～4 万ドル、1 つ前の世代のものでも 1 万ドル以上の値付けがされている。台湾の調査会社によると、2023 年の AI サーバー出荷台数は約 120 万台と見込まれており、1 台に GPU が 1 つと単純に計算しても、相当な金額が必要となる。ユーザーからすれば、価格交渉の観点からも、BCP（事業継続計画）の観点からも、シェアの過度な偏りは望ましくない。2 番手とされる GPU メーカーは、高性能なサーバー向け GPU の発表や、

[1] 機械学習モデルの性能に影響する変数、多いほど性能が高くなりやすい。
[2] 大規模言語モデル。大量のテキストデータを使用し訓練された自然言語処理のモデルで、文章要約、テキスト生成、質問への応答などが可能。
[3] 複数のPCやサーバーを1台のPCやサーバーとみなして稼働させるシステム。

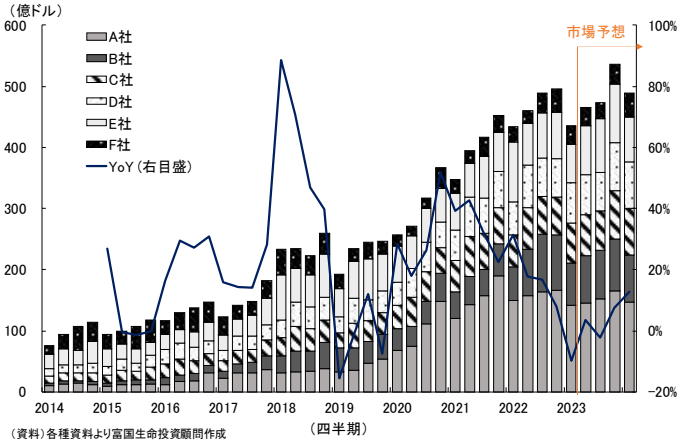
開発環境の広報活動を行うなど攻勢を続けており、競合としての地位が高まれば価格低下への期待も高まろう。ただし、この GPU の量産は今年の 10－12 月期になる予定であり、既に使用している GPU からの切り替えへの抵抗感や現状のベンチマークにおけるパフォーマンスの差などを鑑みると、当面大勢に影響を与える可能性は低いと思われる。

チップを開発するのは、既存のメーカーだけではない。AI への投資を積極的に行っている「ハイパースケーラー[4]」と呼ばれる企業群も、高額な GPU 価格への対応として、自社開発の動きを加速させている。既に複数社が独自のチップを実用化しており、その動きに追随する企業も見えている。ある企業は、自社が開発した最初のアクセラレータが、複雑さが低いモデルであれば GPU を超える性能を発揮するとの発表をしている。ただ、こちらも実用化の目途はまだ立っていない。

４．AI 以外への投資への影響と GPU のボトルネック

AI を巡る競争に勝つために関連投資を続ける必要がある企業にとっては、高額な GPU は大きな負担となる。ハイパースケーラーはその代表例だが、ネット広告の低迷などにより厳しい業績が続いており、設備投資額は低調である。市場予想によると、主なハイパースケーラーの投資動向は、2023 年 1～3 月期に前年比で大きく減少したのち、反転増加するものの伸び率は低位にとどまると見込まれている（図表 3）。トータルの投資意欲と、AI への投資意欲のギャップの分だけ、他分野への投資は低調になると考えられる。現在、その煽りを受けているのが AI 以外のサーバー向け投資であり、証左としてデータセンター関連企業は厳しい業績が続いている。

図表３．ハイパースケーラーの投資動向



（資料）各種資料より富国生命投資顧問作成

また、順調に見える AI 向け GPU も、短期的には供給が不足する可能性がある。現在最も懸念されているのはパッケージングプロセスの能力不足である。先端 GPU は、インターポーザー（中間基板）を用いた 2.5 次元パッケージ技術を用いて製造されており、急激な需要増加に対応するために生産能力の増強を急いで進めているが、主なサプライヤである台湾のファウンドリ（実際に半導体チップなどの製造を行う企業）が半導体製造装置の購入を手控えていた時期があったことが需要急増と重なり、ひっ迫感が高まっている。同ファウンドリは、2023 年中に関連の生産能力を倍増、2024 年は更にそこから倍増させることを計画している。中期的には需給はバランスしていくと思われるが、短期的な不足への対応として、当該メーカーは他のメーカーからの調達も検討しているようで、動向が注目される。

５．まとめ

以上、GPU を巡る現状について論じた。当該メーカーは、しばらく現在の優位性を保持しよう。GPU を利用する側からすれば、早期に 2 番手が技術・環境面で追いつくことを期待しつつも、自社開発含め打てる手立てを継続しよう。短期的には、GPU 生産のボトルネックに注意したい。

（富国生命投資顧問（株）　アナリスト　伊藤 浩士）

[4] 100万台規模のサーバーリソースを保有する企業。