

Tencent「Hy4 preview」徹底分析：770BパラメータMoEモデルの実像

2026年8月発表。Tencentによる7700億パラメータの巨大MoEモデル。
Apache 2.0ライセンスで商用可能だが、極めて高い計算リソースを要求。

技術仕様とアーキテクチャ



100万トークン級の
超長文脈対応

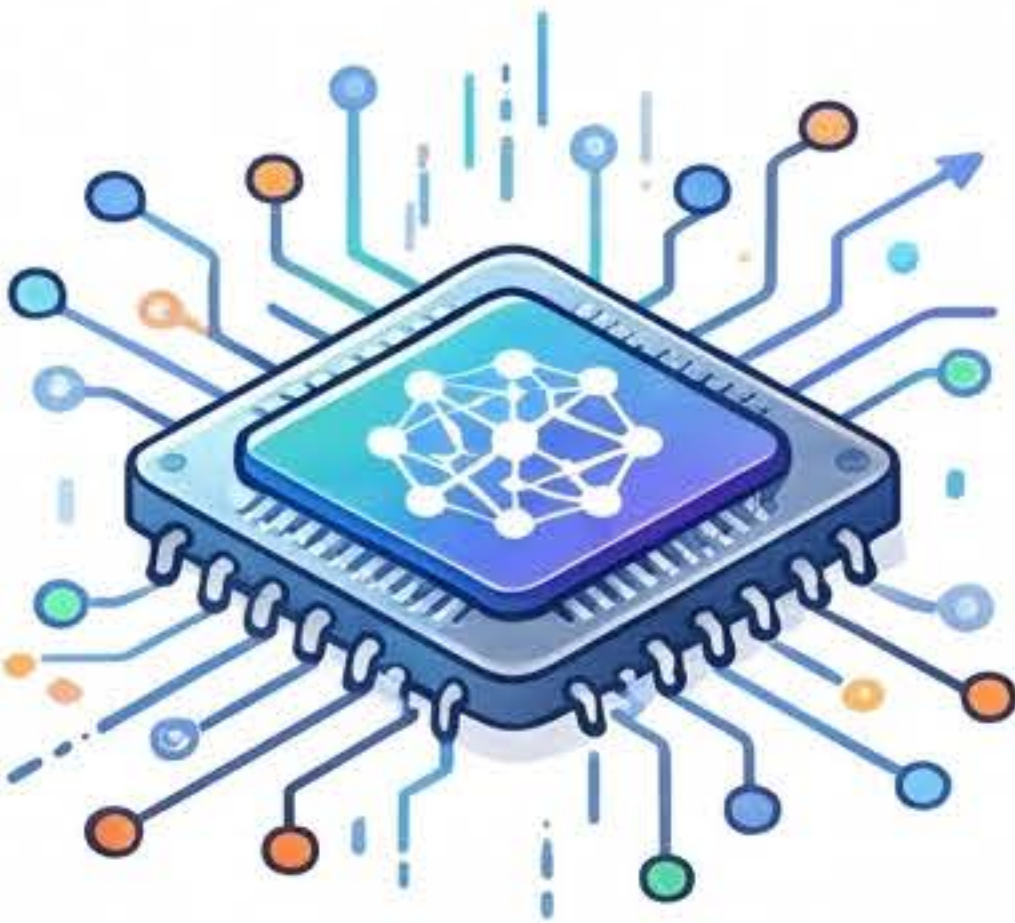
Gated DSAやIndexCacheにより、
長文推論の高速化とメモリ節約を実現。



MTPによる推論加速

独立した10BのMTP層が投機的デコード
を行い、推論スループットを向上。

1トークンあたり49Bのみを活性化し、
巨大な知識容量と計算効率を両立。



主要パラメータ構成

総層数：78層
(1層のDense FFN
+ 77層のMoE)

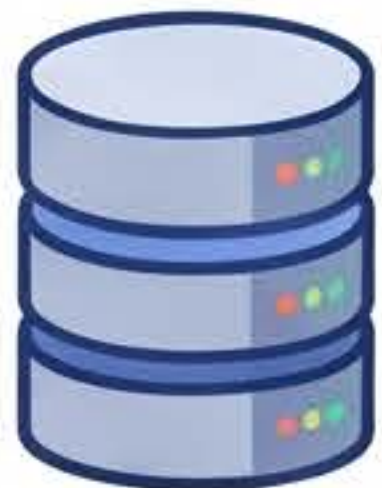
エキスパート数：
256 routed experts
(Top-8活性化)

コンテキスト長：
最大約1,000,000
トークン

実装環境とコスト



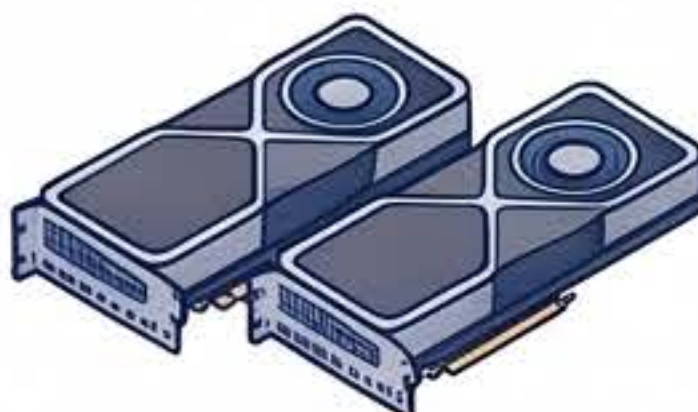
BF16ウェイト：
約1.5TB容量



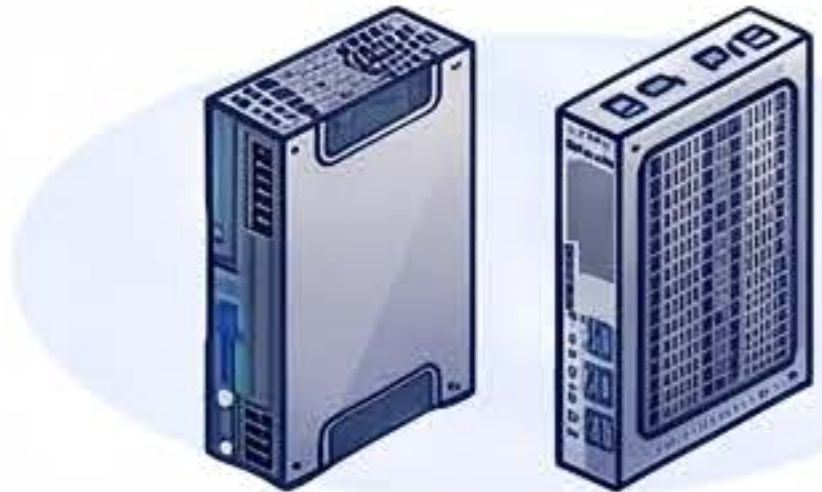
FP8ウェイト：
約760GB容量

膨大なビデオメモリ要求。ウェイトだけでこの容量が必要。

推奨ハードウェア構成



B200×8基
(TP8)



データセンター級構成を想定。



H200×16基
(TP16)



Apache 2.0ライセンス採用

商用利用可能で自由度が高い一方、学習
データの透明性や安全性評価には課題。

