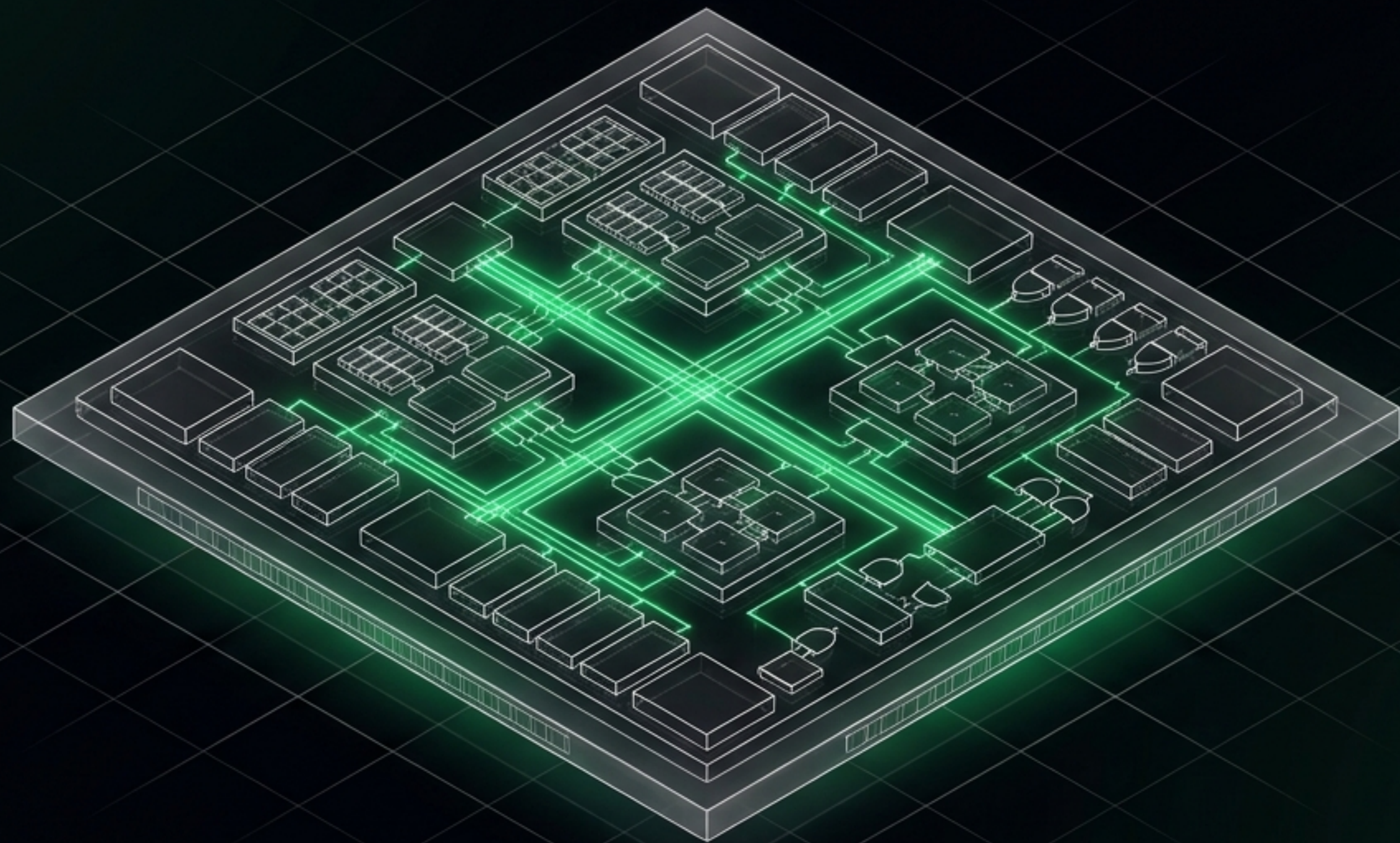


# NVIDIA「RTX Spark」が拓く知財実務のローカルLLM革新

スーパーチップの構造特性と特許業務への実装シナリオ



## 圧倒的な効率化



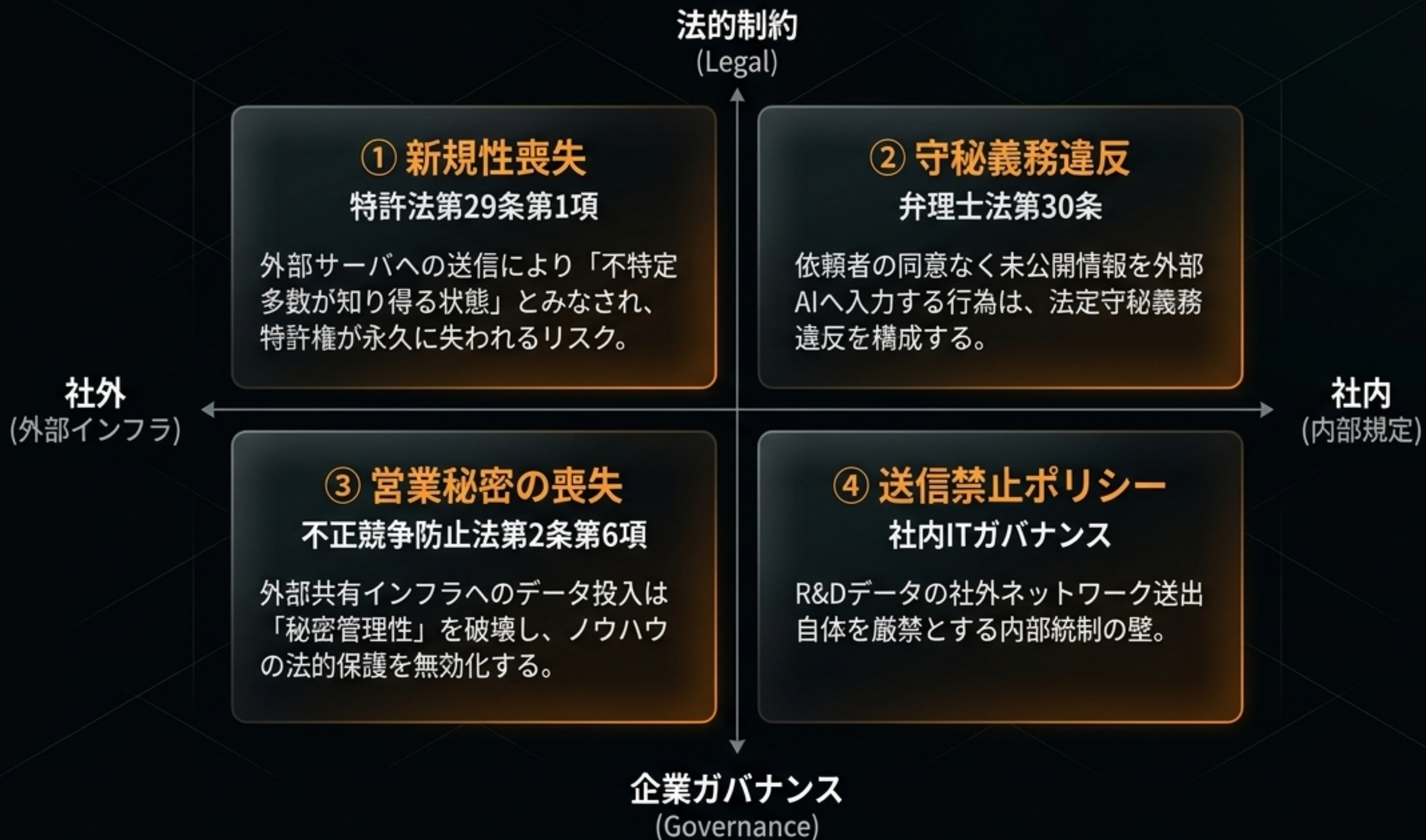
知財分野における  
生成AI導入のジレンマ：  
クラウドの境界線を  
越えた瞬間、発明の  
権利は消滅する。

## 不可逆的な開示リスク

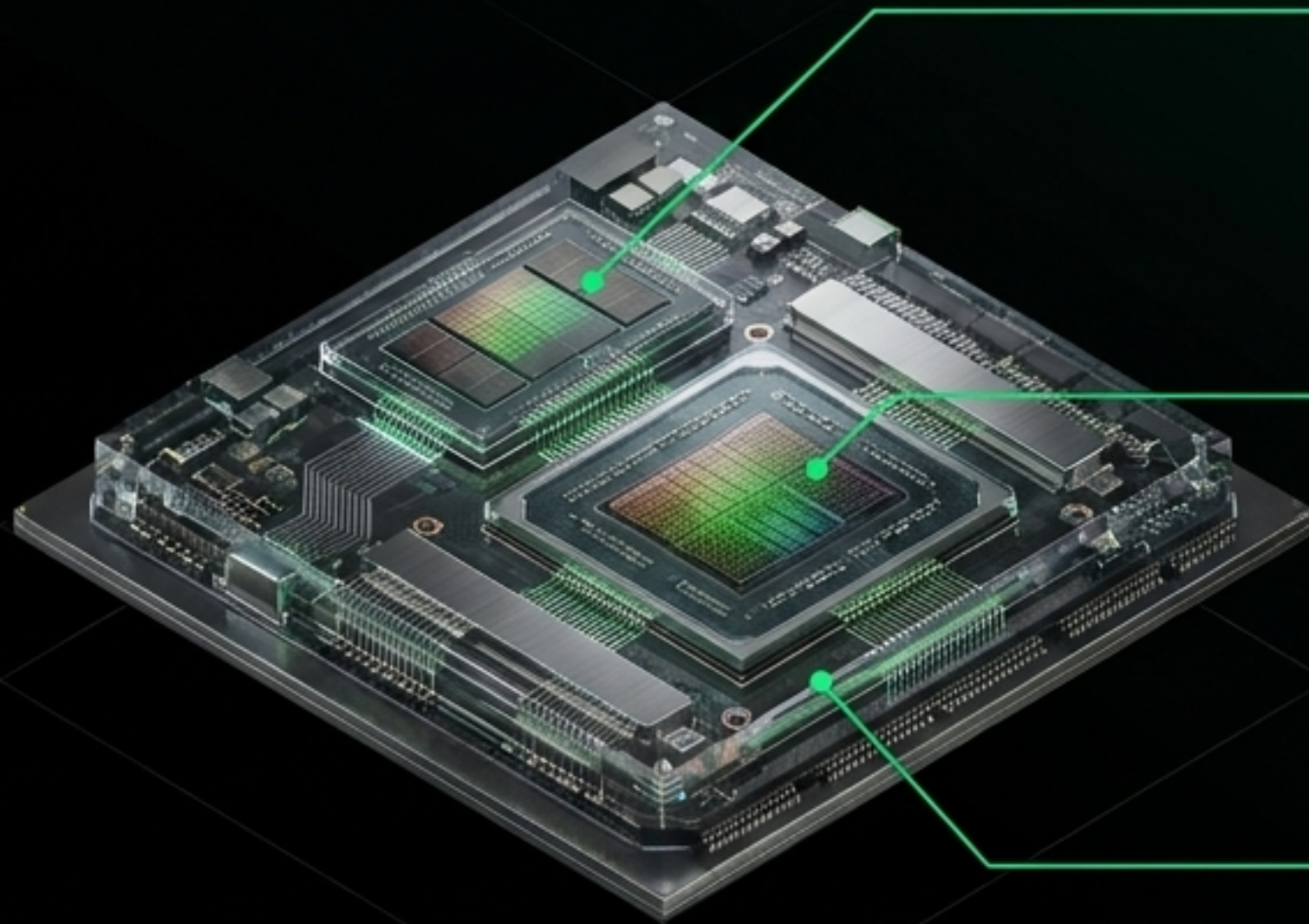


マルチテナント型クラウドLLMの利用は、  
エンタープライズ契約やオプトアウト設定  
を施しても、**構造的な情報漏洩・権利喪失  
リスクを排除**できない。

# クラウド型LLMの構造的・法的リスク（脅威マトリクス）



# 「完全閉域・オンデバイスAI」の具現化



## CPU

最大20コア  
Armベース「NVIDIA Grace」

## GPU

Blackwellアーキテクチャ  
RTX GPUダイ  
(6,144 CUDA / 第5世代Tensor)

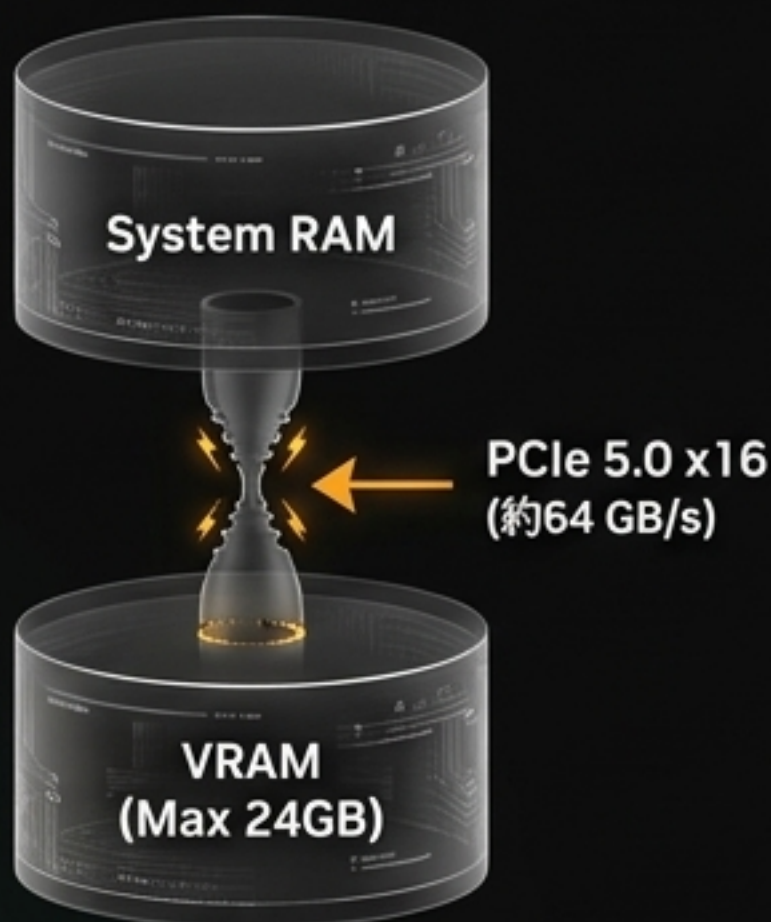
## AI Performance

最大 1 PFLOPS (FP4) 演算性能

サーバー級の大規模AI演算性能を、外部クラウドと完全に遮断された約2kgの薄型ノートPC内に凝縮。

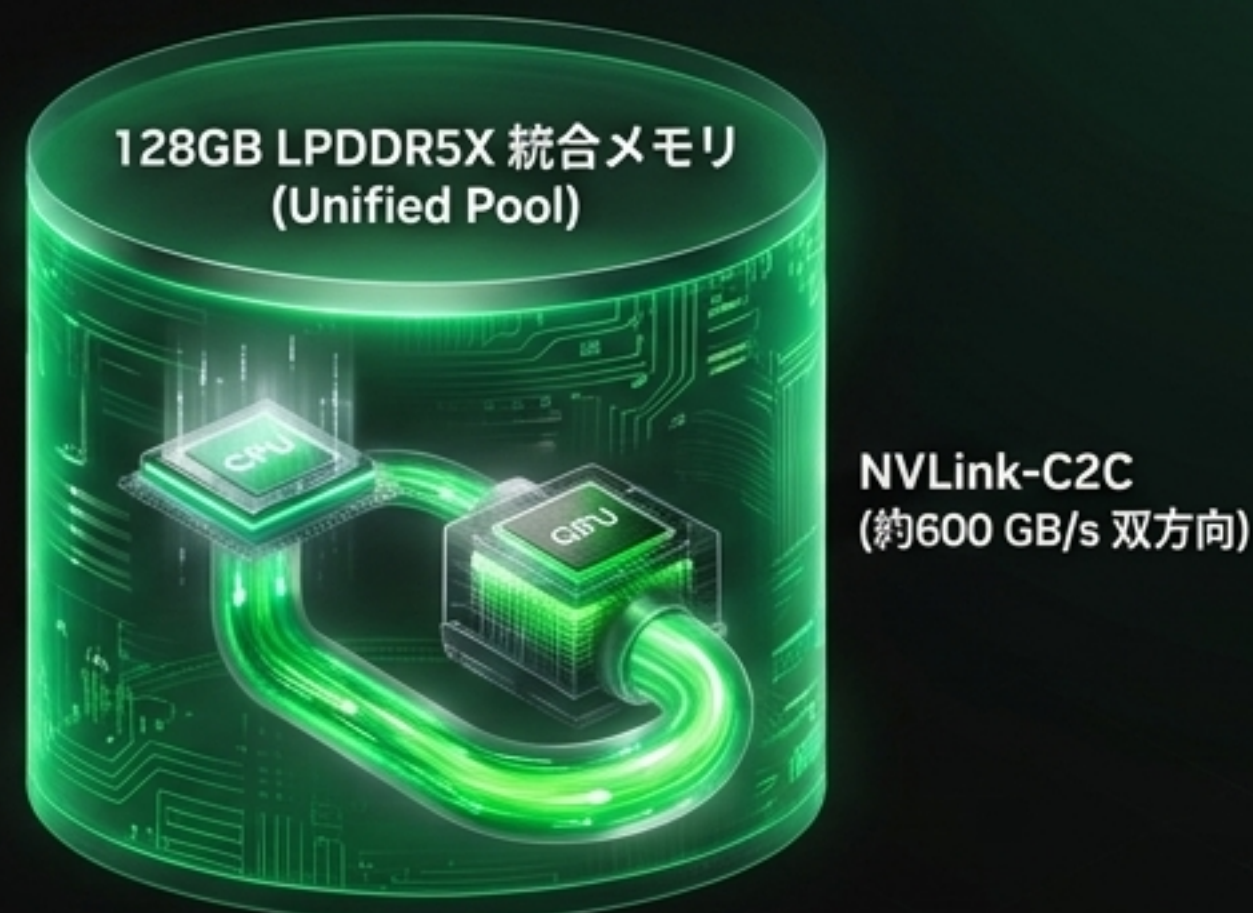
# メモリ・アーキテクチャの革命的変遷 (Architectural Memory Revolution)

## 旧型 / Legacy - The Bottleneck



VRAMの壁と転送のボトルネック

## 新型 / RTX Spark - The Unified Pool



物理的境界の消失と単一アドレス空間

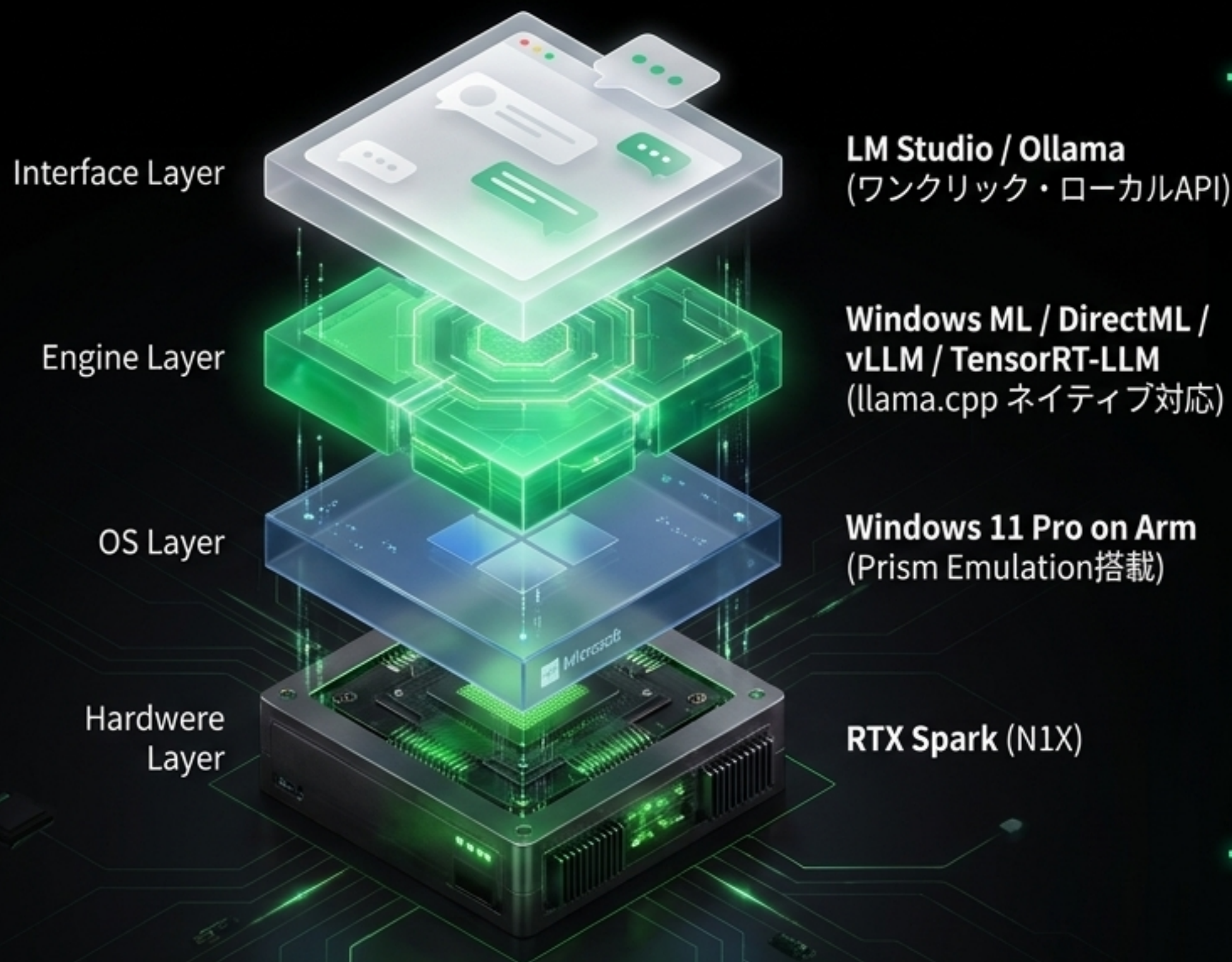
キャッシュコヒーレンスの確立により、テンソル転送待ち時間を根絶。  
数十億パラメータの超巨大モデルを単一のメモリ空間に直接展開。

# メモリ容量と帯域幅の戦略的比較 (Strategic Comparison of Memory Capacity & Bandwidth)

|          | NVIDIA RTX Spark         | GeForce RTX 5090      | Apple M5 Max | DGX Spark                |
|----------|--------------------------|-----------------------|--------------|--------------------------|
| 最大メモリ容量  | 128GB<br>統合              | 32GB<br>専用VRAM        | 128GB<br>統合  | 128GB<br>統合              |
| システム帯域幅  | ~273-300+ GB/s           | ~1,792 GB/s           | ~800+ GB/s   | ~273 GB/s                |
| インターコネクト | NVLink-C2C<br>(600 GB/s) | PCIe 5.0<br>(64 GB/s) | オンダイ・ファブリック  | NVLink-C2C<br>(600 GB/s) |
| 価格帯      | 約100万円~                  | 約1999ドル~<br>(GPUのみ)   | 約82万円~       | 約4,699ドル~                |

インサイト：知財実務の最適解は「最高のスループット(5090)」ではなく、数十万トークンを保持できる「大容量メモリと広コンテキスト耐性(Spark)」にある。

# Windowsエンドポイントへのシームレス統合



## 脱Linux環境

オープンソースLLMエコシステムが、  
ビジネス標準のWindowsエンドポイン  
トへ完全統合。複雑なコンパイル不要  
で即時展開可能。

# メモリフットプリントの戦略的活用: 128GB 統合メモリの内訳

128GB

OS/System  
(8-16GB)

モデル重み (70B FP8): 約70GB

KV Cache (コンテキスト保持領域):  
40GB+

## インメモリ特許RAGの実現



256K~1Mトークンの長文コンテキスト耐性により以下を同時保持可能:

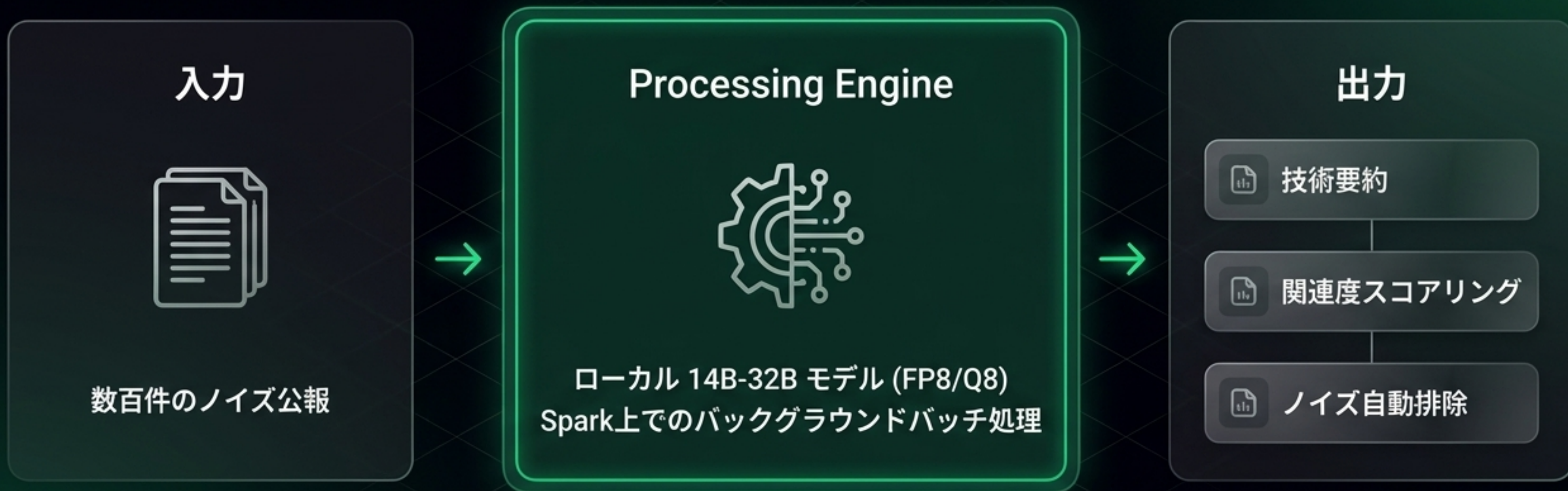
対象特許明細書全文

先行技術文献 (引例) PDF 3~4件

審査官の拒絶理由通知書

従来の24GB VRAMでは物理的に不可能な  
『全量一括照合』をローカルで完遂。

# Workflow 1: 調査・スクリーニング工程の変革



ROI / セキュリティ統制効果

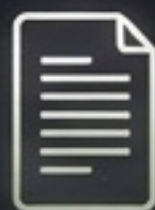
完全オフライン

従量課金ゼロ

未公開検索クエリの完全秘匿

# Workflow 2: 明細書作成・中間処理の変革

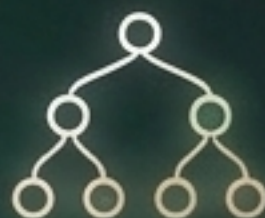
## 明細書作成 (Patent Drafting)



発明開示書 (アイデアシート)



70B-120B ローカルモデル  
(INT4-FP8)



独立・従属クレームの展開と  
実施形態起草



特許法第29条・弁理士法第30条  
完全準拠 (社内完結)

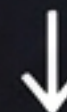
## 中間処理 (OA Response)



拒絶理由通知 + 複数引例PDF



128K+ 超長文コンテキスト空間



微視的差異の全量一括照合



細切れ要約では見落とす「段落単位  
の矛盾」を長文横断で直接対比

# 機密分離型「二段階ハイブリッドワークフロー」

Firewall / Boundary

ローカル（完全閉域）

Phase 1: 生データ読込

生データ読込



サニタイズ処理

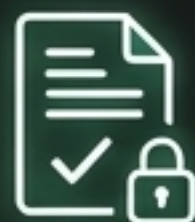
固有名詞・数値を抽象化し、  
論理骨子(A+B+C)のみを抽出。



置換マッピング表:

|               |                    |
|---------------|--------------------|
| <COMPONENT_X> | <cm000kz0c066...>  |
| <PARAM_Y>     | <Encrypt000000...> |
| <PARAM_Y>     | <Encrypt000000...> |
| <PARAM_Y>     | <Encrypt000000...> |

Phase 3: ローカル復元



ローカル復元

置換マッピング表を逆適用し、正確なパラメータを注入。

クラウド（外部最先端モデル）

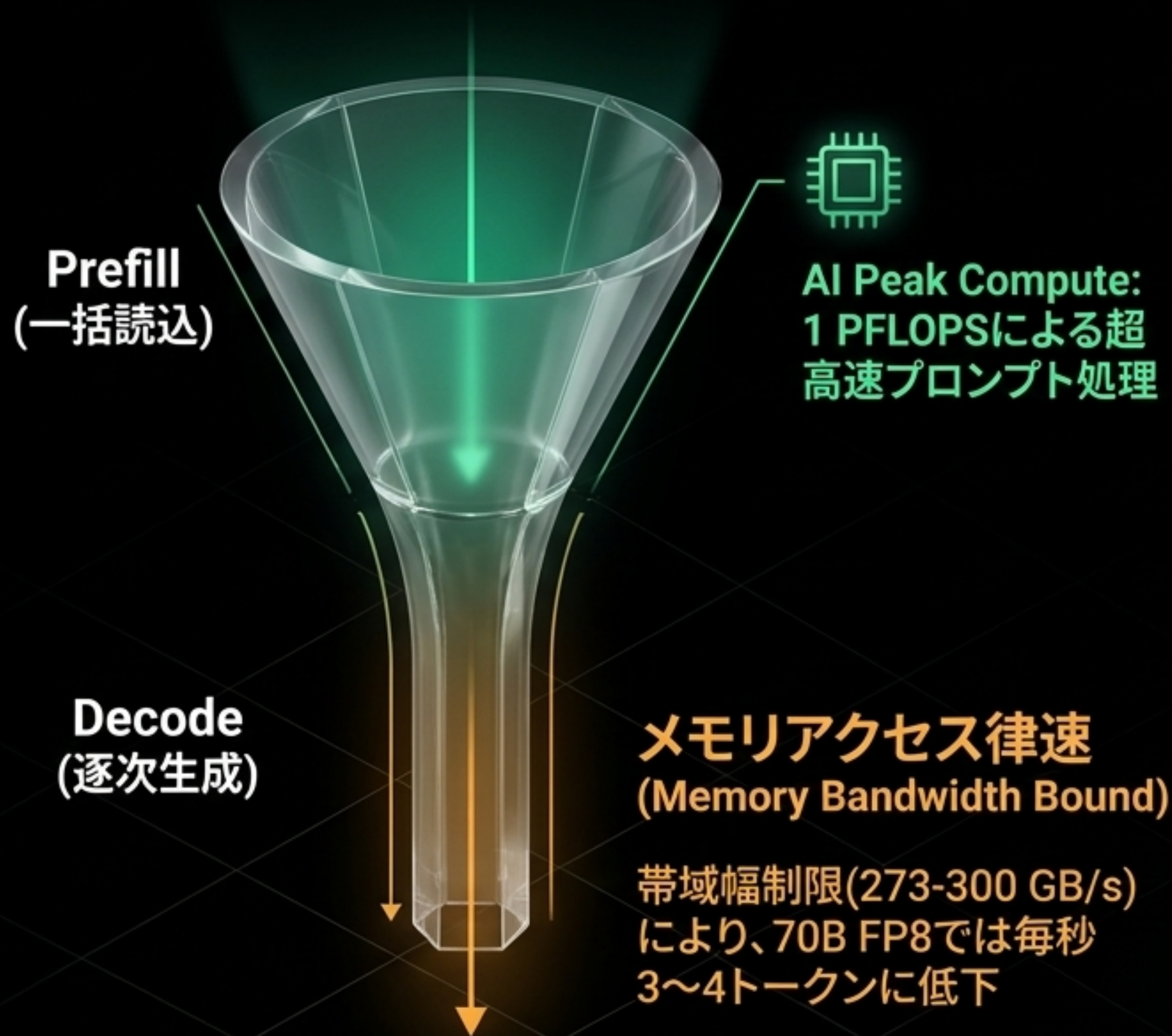
Phase 2: クラウド処理



Claude Opus / GPT-5.5

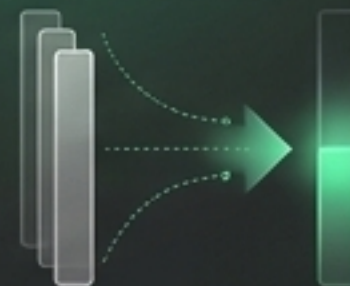
抽象化骨子の高度な法的推敲と  
英語表現の最適化。  
技術本質はマスキングされているため  
新規性喪失リスクゼロ。

# 推論速度の現実と最適化 (Technical Trade-off)



## 最適化アプローチ (Solutions)

### ① 量子化 (Quantization)



INT4精度への圧縮により、データ移動量を半減させスループットを向上。

### ② 投機的デコード (Speculative Decoding)



将来のトークンを予測生成し、実効スループットを毎秒15~25+トークンへ劇的に加速。

# ハードウェア導入ROIと調達戦略

## ハードウェア導入ROI (Financial Break-even Analysis)



## 推奨される調達・配備戦略 (Deployment Strategy)



### 共有高負荷解析端末

部門内の共通インフラとして、大量公報のバッチ処理や高負荷スクリーニング専用ステーションとして運用。



### 専任弁理士・主席知財 部員への重点配備

全社一律配布ではなく、明細書ドラフティングを主導するコア人材へ集中的に投資し、ROIを極大化。

# エンドポイントセキュリティの再構築



**BitLocker**  
厳格なストレージ  
暗号化



**EDR**  
ローカルマルウェアに  
対する振る舞い検知



**MFA**  
デバイスロック解除の多要  
素認証



**遠隔ワイプ  
(Remote Wipe)**

盗難・紛失時の即時データ  
抹消機構

ローカルAIのパラドックス：クラウド流出は防げるが、端末内部に「未公開発明と社内特許DB」の全量が丸ごと保存されるという新たな物理的脆弱性が発生する。

ローカルLLMワークステーションの稼働には、これらの物理的・論理的統制の完全実装が必須条件となる。

# RTX Sparkによる知財業務の再定義



RTX Sparkは単なるPCではない。業務効率化と法務コンプライアンス（機密保持）を完全両立させる、知財部門の「中核インフラ」である。